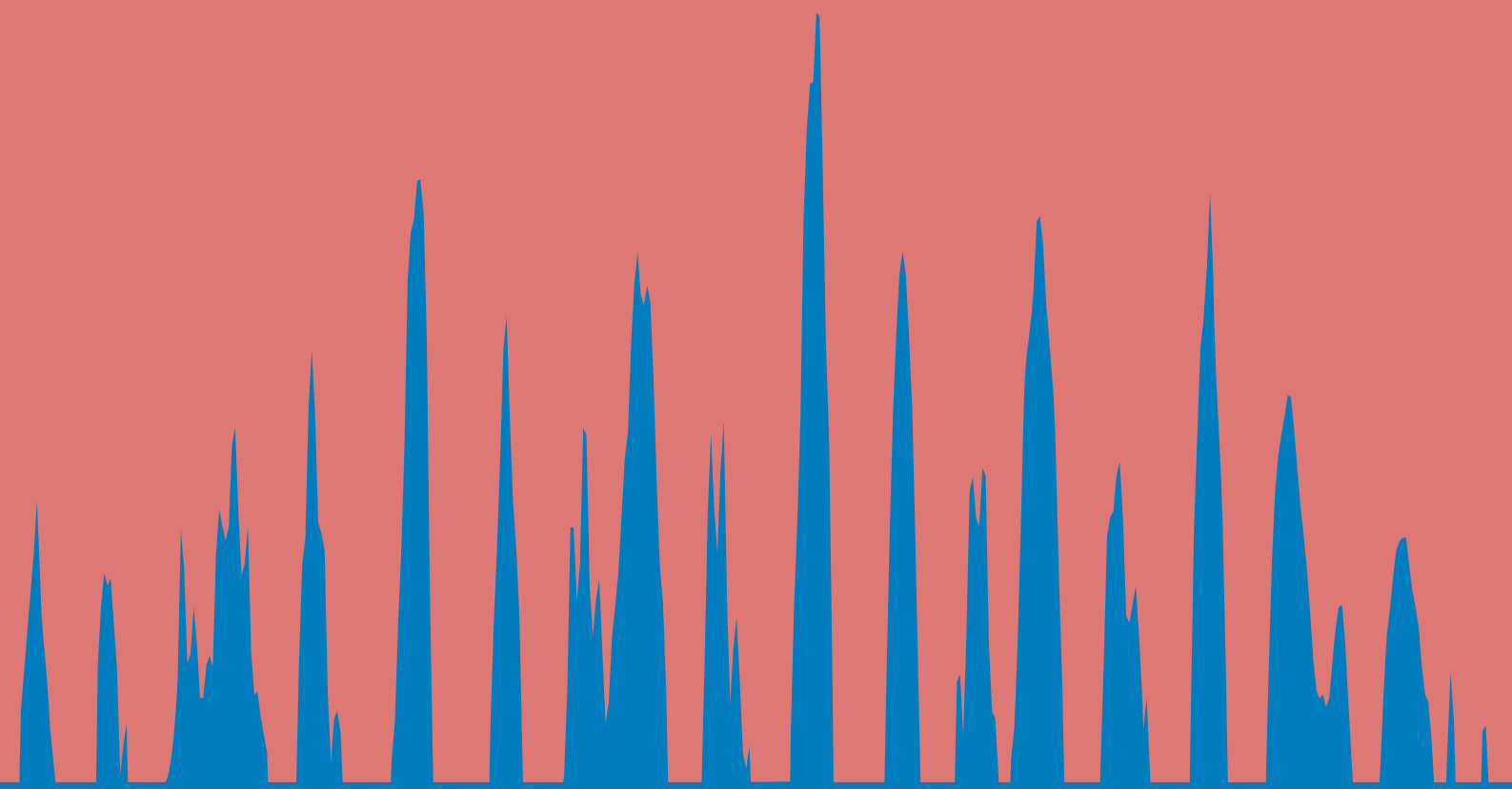
August 7–11, 2023 — Prague, Czech Republic

ICPhS 2023

CONGRESS PROCEEDINGS



ICPhS 2023
August 7–11, 2023
Prague Congress Center, Czech Republic

20th International Congress of Phonetic Sciences (ICPhS)

August 7–11, 2023
Prague Congress Center, Czech Republic





ELSEVIER





Title: Proceedings of the 20th International Congress of Phonetic Sciences - ICPhS 2023
Edited by Radek Skarnitzl and Jan Volín
Publication Date: 2023
Publisher: The International Phonetic Association
ISSN 2412-0669

Publication and Review Ethics statement:
https://www.internationalphoneticassociation.org/content/icphs/ethics_statement

Copyright:
Authors (This in an open access publication licensed by the authors of papers under the CC BY-NC-ND 4.0 license).

To cite papers in this volume:
Author Name(s) (2023). Paper title. In: R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023* (pp. XX–YY). International Phonetic Association.

WELCOME TO ICPhS 2023

Welcome from the Chair of the Organizing Committee for ICPhS 2023

Welcome everyone to the 20th International Congress of Phonetic Sciences. I am happy to welcome all of you here in Prague, in the heart of Europe, and I am truly honoured to host what some people call the "phonetic Olympics", the congress during which we discuss the latest developments in the broad field of speech sciences.

It seems like a lifetime away when, on the 9th of August 2019, the 19th ICPhS was concluded in Melbourne, Australia. Four years, nearly 1,500 days during which the world has changed beyond recognition. It was only a few months after the last ICPhS that the new strain of coronavirus started spreading, eventually throwing the whole world into chaos, a series of lockdowns, and the omnipresence of facemasks. The covid pandemic has also changed the way we – scientists and academics – operate. Indeed, the facemask has triggered an avalanche of research of various facets of speech communication – searching Google Scholar for the combination of the terms "speech" and "facemask" yields thousands of entries, and several presentations at this congress will also be dedicated to what, for many of us, was a new phenomenon in our lives. However, while the pandemic has brought a number of positive changes into our world, precious little positive can be found in other developments over the past four years – such as the senseless aggression which has been going on in Ukraine, merely several hundred kilometres from here for the past 18 months.

Phonetic sciences have a long tradition in the Czech Republic. The phonetic laboratory in Prague was established in 1916, making it one of the oldest laboratories in central Europe and the prime centre of phonetic sciences in the Central European area for several decades. The birth of the lab is closely linked to Josef Chlumský, an experimental phonetician who studied under Abbé Rousselot in Paris. True to his experimental upbringing, Chlumský dedicated a lot of effort to obtaining modern laboratory equipment – the most prized being a set of tuning forks, the tonometre, which was used to identify resonance frequencies of speech sounds. And it is my pleasure to invite those who would like to see our humble collection of historical phonetic instruments to visit the Faculty of Arts, in the very centre of Prague, on Wednesday between 5.30 and 7 pm – the historian of our institute, Pavel Šturm, will be there to show you around. This is not on the official programme, but ask one of the student helpers, they will point you in the right direction.

You will have read that this is the second time that the Institute of Phonetics in Prague is hosting ICPhS. It was Chlumský's successor, Bohuslav Hála, who was the chair of the sixth congress in 1967. It seems fitting that I show you several comparisons of the two congresses, which are based on a comparison which was conducted by Pavel Šturm for the 2015 ICPhS.

There were nearly 600 registered participants and 237 presentations in 1967, and 1,144 participants and 869 presentations this year. An interesting detail concerns authorship: 87%

of papers were written by a single author for the 1967 congress, while the same applies only to 18% of papers in 2023. The comparison of participants according to country is as follows:

6th ICPhS 1967 (active part.)		20th ICPhS 2023	
Czechoslovakia	37	United States	198
Soviet Union	35	Germany	167
United States	31	United Kingdom	105
Germany (East)	25	France	72
Germany (West)	21	Japan	66
Netherlands	10	Canada	64
Poland	10	Hong Kong	40
Romania	9	Czech Republic	38
United Kingdom	9	Netherlands	38
France	6	China	33
Italy	6	Australia	30
Yugoslavia	6	Switzerland	25
		Korea	24
		Taiwan	22
		Sweden	20
		Poland	20
		Spain	19
		Italy	17
		Austria	15
15 other countries	33	36 other countries	131

I hope you will agree with me that the comparison is a strong indication that phonetic sciences are a thriving interdisciplinary area, and I would like to think that professors Chlumský and Hála would be proud of what the entire field of phonetic sciences, and our institute have achieved.

Let us add a few statistics which are specific to this congress. The acceptance rate was slightly below 85%, and there are 266 oral presentations and 603 posters on the programme.

Of course, organizing a congress is an undertaking which has many people in the background, and it would not be possible to thank all of them. However, I would like to thank the most important ones: my colleagues from the Scientific Programme Committee (Tomáš Bořil, Katka Chládková, Jonáš Podlipský and Jan Volín), the representative of Guarant International, the professional congress organizer, Romana Kluková, and for the army of volunteers, Míša Svatošová. Naturally, this congress would not take place without the International Phonetic Association, and I would particularly like to thank its President, Michael Ashby, for support. And most of all, thank you, all of you – authors for submitting your contributions and for coming to Prague, to the reviewers and scientific area chairs who helped us with selecting the contributions!

I hope you will enjoy your stay in Prague and the Czech Republic, I hope you will be able to explore some of the beautiful things this city and country have to offer. And most importantly, I wish all of us a fruitful congress – five days of exciting presentations, stimulating discussions, chatting with old friends and meeting new ones – simply a very successful ICPhS!



Radek Skarnitzl
Chair of the 20th ICPhS 2023



Welcome from Michael Ashby, President of the International Phonetic Association 2019–2023

As President of the International Phonetic Association I have the privilege and the pleasure of welcoming you all to this historic twentieth International Congress of Phonetic Sciences. We have a full and exciting programme ahead of us, much to see, much to learn, friendships to renew, and no doubt new ones to make. But before we look ahead, it is appropriate that we should spend a few moments remembering those of our colleagues who have, sadly, died in the last 4 years or so.

They include:
Luc van Buuren, Anne Cutler, Grzegorz Dogil, Anthony Fox, Erik Fudge, Wiktor Jassem, Hermann Künzel, John Laver, Philip Lieberman, Robert Mannell, Silvia Moosmüller, Terrance Nearey, Ohala, John Włodzimierz Sobkowiak, Elmar Ternes, Hans Tillmann, Jack Windsor Lewis, John Woodhead.

This list is certainly incomplete, and I apologise for omissions. But it is already a soberingly long list, no doubt in part on account of the Covid pandemic. And as you look though it you will notice several particularly prominent names: Anne Cutler (1945–2022), who was formerly a member of both the IPA Council and the Permanent Council for ICPhS; and two former Presidents of the IPA, John Laver (1938–2020), and John Ohala (1941–2020). There is a story to tell about each one on this list, and I am not going to hurry through all the names, but rather invite you to spend a few moments in silence, remembering them all.

The International Congress is only one of the IPA's many activities. Our central concern is of course the protection and maintenance and development of the International Phonetic Alphabet, by virtue of which the IPA serves effectively as an international standards organisation. But we support phonetic research in numerous other ways. I am proud of the fact that no fewer than 69 student researchers from around the world are here at the Congress with substantial financial support from the IPA. At the same time we are supporting important language documentation work. And, of course, we publish a major journal, The Journal of the International Phonetic Association (JIPA), which began in 1886 and continues to grow in significance and impact.

Why am I boasting about the IPA's achievements and activities? Well, I want to draw your attention to two forthcoming meetings. In the lunch break tomorrow (Tuesday) there will be a meeting devoted to JIPA, and on Wednesday at the same time we have the four-yearly General Meeting of the IPA, at which we will hear reports from the Association's committees and officers, and welcome the newly-elected Council and Executive. Both of these meetings are open to all, whether members or not, and I encourage you to attend.

So, again a warm welcome from the IPA, and I return you to the capable hands of our hosts here in Prague.



Michael Ashby
President of the International Phonetic Association

Scientific area Chairs for ICPhS 2023 congress

Scientific Area	Area Chairs
1. Speech Perception	Joseph Casillas
1. Special Session - Interplay or intermezzo? Structures and processes in prosody and music	Jianjing Kuang, Oliver Niebuhr
2. Speech Acoustics	Adrian Simpson
2. Special Session - Bridging linguistic and clinical perspectives through computational models of speech production	Ben Parrell, Doris Mücke, Antje Mefferd, Sarah Harper
3. Speech Production and Speech Physiology	Peter Birkholz
3. Special Session - Novel approaches to studying vocal development	Kasia Hitczenko , Marisa Casillas, Alejandrina Cristia, Meg Cychosz , Amanda Seidl
4. Speech Prosody	Plínio Barbosa
4. Special Session - The pedagogy of the Laryngeal Articulator Model	John Esling, Lise Crevier-Buchman, Míša Hejná, Scott Moisik
5. Phonation and Voice Quality	Heriberto Avelino
5. Special Session - Qualitative research in phonetics	Joe Pearce , Nate Lindgren Haj Bakir
6. Tone	Liquan Liu
6. Special Session - Changing research culture - Toward big team speech science	Timo Roettger, Joseph Casillas, Stefano Coretta
7. Syllable	Pavel Šturm
7. Special Session - The sounds of ageing	Míša Hejná, Anna Bothe Jespersen
8. Laboratory Phonology	Eva Reinisch
8. Special Session - The history of phonetic sciences	Michael Ashby, Rüdiger Hoffmann, Jürgen Trouvain
9. Phonology-Phonetics Interface	Elizabeth Zsiga
9. Special Session - Acoustic manifestations of speech register	Melanie Weirich, Stefanie Jannedy
10. Phonetics of First Language Acquisition	Nikola Paillereau

Scientific area Chairs for ICPhS 2023 congress

Scientific Area	Area Chairs
10. Special Session - Prosody Visualization Challenge	Katie Jepson, Debbie Loakes, Olga Maxwell
11. Phonetics of Second and Foreign Language Acquisition	Šárka Šimáčková
12. Bilingual/Multilingual Phonetics	Charles Chang
13. Phonetic Universals and Typology	Jeff Mielke
14. Phonetics of Sound Change	Marton Soskuthy
15. Speech Evolution	Bart de Boer
16. Speech Technology	Barbara Schuppler
17. Speech Corpora and Big Data	Adrian Leemann
18. History of Phonetics	Ruediger Hoffmann
19. Phonetics of Lesser Documented and Endangered Languages	Rachid Ridouane
20. Field Methods in Phonetics	Gareth Walker
21. Phonetics of Conversation	Martin Kohlberger
22. Sociophonetic Variation	Melanie Weirich
23. Forensic Phonetics and Speaker Characteristics	Gea de Jong-Lendle
24. Clinical Phonetics and Speech Disorders	Mathias Scharinger
25. Phonetic Neurolinguistics	Mirjam Broersma
26. Phonetic Psycholinguistics	Sona Patel
27. Phonetics of Affective Speech	Lise Crevier-Buchman
28. Multimodal Phonetics	Pilar Prieto
29. Phonetics Pedagogy	Kristýna Červinková Poesová

1. Speech Perception

Lasting stress 'deafness' after auditory training: French listeners revisited (ID: 19) - Peperkamp Sharon, Brazeal Jason 1
Exposure-independent comprehension of Greek-accented speech: evidence from New Zealand listeners (ID: 46) - Cooper Stephanie, Cooper Sarah 6
Effects of aging and age-related hearing loss on emotion identification in prosodic and semantic channels (ID: 51) - Wang Yingyang, Xu Min, Shao Jing, Wang Lan, Yan Nan 11
Perception, recognition, and encoding of Cantonese sound change variants (ID: 62) - Soo Rachel, Babel Molly 17
Phonetic correlates of gender in prepubertal voices (ID: 75) - Funk Riccarda, Simpson Adrian P., Weirich Melanie 22
Unmasking the truf: Impact of community masks on the perception of voiceless fricatives in English (ID: 77) - Canzi Massimiliano, Rathcke Tamara 27
Syllable rate drives rate normalization, but is not the only factor (ID: 81) - Severijnen Giulio, Bosker Hans Rutger, McQueen James 32
Pause particles influencing recollection in lectures (ID: 85) - Elmers Mikey 37
Differential perception of vowel and stop contrasts in Spanish-English bilingual preschoolers (ID: 86) - Montanari Simona, Steffman Jeremy, Mayr Robert, Flores Patricia, Feeney Carissa 42
Generalized perceptual adaptation to second-language (L2) speech: variability, similarity, and intelligibility (ID: 87) - Bradlow Ann R., Bassard Adrianna M., Paller Ken A. 47
Linguistic release from masking in simulated electric hearing (ID: 88) - Lung Huizi, Liao Shangdi, Chen Fei 52
Are alveolar trills perceived as "strong" rhotics? (ID: 89) - Reinisch Eva, Mitterer Holger 57
Vowel perception under prominence: Dual roles for F0 and duration (ID: 93) - Steffman Jeremy, Zhang Wei 62
American listeners' recognition of sentences unaffected by racial and ethnic primes (ID: 98) - McLaughlin Drew, Van Engen Kristin 67
No evidence for convergence to sub-phonemic F2 shifts in shadowing (ID: 100) - Ulusahin Orhun, Bosker Hans Rutger, McQueen James M., Meyer Antje S. 72
Perception of coarticulatory vowel nasalisation in German listeners and its relationship to production (ID: 101) - Kunay Esther, Hoole Philip, Harrington Jonathan 77
Neutralisation and the perception of close-mid and open-mid vowels: The gradient between phonological categories (ID: 116) - Im Megumi, Veloso João, Costa Maria Conceição, Jesus Luis 82
Cross-domain pitch recognition abilities of Mandarin-speaking prelingually deaf children with cochlear implants (ID: 118) - Jin Qianwen, Yu Jue, Sun Han, Zhang Kexin, Du Xinran 87
Rapid speech adaptation and its persistence over time by non-standard and non-native listeners (ID: 120) - Du Fengting 92
Integration of multiple cues in native and non-native speech perception (ID: 132) - Ren Xiaomu, Cohen Clara 97
The effects of accent familiarity on narrative recall in noise (ID: 135) - Krug Andreas, Khattab Ghada, White Laurence 102
Feature-specific perceptual sensitivity to within-category variations (ID: 136) - Nielsen Kuniko, Scarborough Rebecca 107
Incremental cue training: a study of lexical tone learning by non-tonal listeners (ID: 140) - Li Yanyu, White Laurence, Khattab Ghada 112
Testing the perceptual basis of laryngeal metathesis and rarity of preaspirated stops (ID: 146) - Gilbert Madeline 117
Sibilant perception by merged speakers: the case of Taiwan Mandarin (ID: 155) - Lee-Kim Sang-Im, Tung Hsiang-Yu 122
Early exposure to an Indigenous heritage language and the transfer of perceptual abilities into a foreign language (ID: 156) - Mulík Stanislav, Amengual Mark, Carrasco-Ortiz Haydee 127
Word-level prominence and "stress deafness" in Maltese-English bilinguals (ID: 173) - Lialiou Maria, Bruggeman Anna, Vella Alexandra, Grech Sarah, Schumacher Petra B., Grice Martine 132
Perceptual adaptation to altered cue informativeness: Distributional, auditory, and lexical factors (ID: 179) - Gao Jiayin, Kirby James 137
Production and perception of Mandarin /i/-nasal rhymes (ID: 182) - Liu Suyuan, Babel Molly 142
The role of boundary tone in chunking of spontaneous speech (ID: 190) - Weidinger Lucas, Wan Xinyuan, Koivusalo Liisa, Dobrego Aleksandra 147
Talker variability in cross-dialect lexical processing (ID: 208) - Ross Jory, Clopper Cynthia 152
Sound categorization after speaking with a bite block (ID: 227) - Zhang Xinyu, Schoonen Rob, Janse Esther 157
Perception of the Japanese word-medial singleton/geminate contrast by Kelantan Malay speakers (ID: 249) - Hamzah Hilmi, Tsukada Kimiko, Hajek John 162
The effect of shortening onset consonants on speech segmentation by Taiwanese Southern Min listeners (ID: 268) - Ou Shu-chen, Guo Zhe-chen 167
Individual differences in non-native phonetic perception: potential links to native perception, attention, & memory (ID: 269) - Honda Claire, Clayards Meghan, Baum Shari 172
Clear speech facilitates word segmentation: evidence from eye-tracking (ID: 283) - Guo Zhe-chen, Smiljanic Rajka 177
Cue shifting in the perception of Shanghainese sandhi patterns (ID: 286) - Wei Zihao 182
Top-down effects of lexical pitch accent on phonetic categorization in Japanese (ID: 299) - Ariga Terumichi, Matsubara Risa 186
The role of allophones in phoneme perception models: Do devoiced vowels trigger vowel epenthesis? (ID: 309) - Kishiyama Takeshi, Huang Chuyu, Furukawa Kei, Hirose Yuki 191

Physiological vs. Subjective sleepiness: what can human hearing estimate better? (ID: 326) - Martin Vincent P., Ferron Aymeric, Rouas Jean-Luc, Shochi Takaaki, Dupuy Lucile, Philip Pierre 196
Sleepiness estimation from voice: a French perceptual replication study (ID: 327) - Martin Vincent P., Rouas Jean-Luc, Ferron Aymeric, Philip Pierre 201
Effects of different levels of amplitude variation on perceived speaker dominance (ID: 340) - Pearsell Sara, Pape Daniel 206
Perceptual Training Enhances the Use of Vowel Quality Cues to Lexical Stress: The Benefits of Intonational Variability (ID: 348) - Tremblay Annie, Kim Hyoju, Kim Sahyang, Cho Taehong 211
Effect of speech enhancement at phonetic level perception of speech between native English and Mandarin listeners (ID: 360) - Zhang Yunqi C., Watson Catherine I., Hui C.T. Justine, Hioka Yusuke 216
The perception of visual phonetic information: eye-tracking while searching for segmental versus prosodic cues (ID: 363) - Deng Xizi, Jastrzebski Erin, McClay Elise, Yeung H.Henny, Wang Yue 221
The role of duration in the perception of checked versus unchecked tones in Taiwanese Southern Min (ID: 365) - Zhang Wei, Lu Yu-An 226
Perception of Japanese pitch accent by typically developing children and children with autism spectrum disorders (ID: 366) - Shinohara Yasuaki, Uchida Mariko, Matsui Tomoko 231
A perceptual experiment testing the Italian /ʎ/ - /j/ contrast (ID: 368) - Anastaseni Anna, Romano Antonio 236
Intelligibility of spontaneous casual speech in background noise across the lifespan (ID: 373) - Tuomainen Outi, Taschenberger Linda, Soengadie Amanda, Hazan Valerie 241
Clear speech improves word recognition, but may increase listening effort (ID: 389) - Meemann Kirsten, Smiljanic Rajka 246
Noise does not affect weighing of speaker information in spoken-word recognition (ID: 392) - Reese Helen, Reinisch Eva 251
The online processing of anticipatory tonal information in Standard Chinese by native and non-native listeners (ID: 393) - Hasanah Hana Nurul, Chen Yiya, Yang Qing 256
Predicting intelligibility from pronunciation distance metrics (ID: 404) - Bent Tessa, Holt Rachael 262
Nonnative accent adaptation in the initial moments and over a month (ID: 425) - Xie Xin, Kurumada Chigusa 267
An amodal study of phonological abstraction in early infancy (ID: 428) - Altuntas Eylem, Best Catherine T., Kalashnikova Marina, Götz Antonia, Burnham Denis K. 272
Perception of metaphony: a comparison between two dialects of the Lausberg Area (southern Italy) (ID: 458) - Greca Pia 277
Perceptual cues to ejective stops across languages (ID: 464) - Percival Maida 282
Korean listeners' weighting and integration of cues to the three-way korean stop contrast (ID: 470) - Kim Hyoju, Tremblay Annie, Cho Taehong 287
Developmental differences in intelligibility of unfamiliar accents (ID: 478) - Holt Rachael, Bent Tessa, Lind-Combs Holly, Kunath Audrey 292
Perceiving speech produced with face masks in competing talker environments (ID: 493) - Chiu Faith, Bartoseviciute Laura, Lee Albert, Yao Yujia 297
Production and Perception of Long and Short Vowel Contrast in Mongolian (ID: 502) - Min Ao, Baiyila Dahu, Li Aijun 302
Does talker variability help adults learn novel words? (ID: 523) - Abu El Adas Sandy, Yen Ivy, Levi Susannah 307
Sound change in the lab: Exposure-induced vocalic chain shifts (ID: 526) - Podlipský Václav Jonáš, Olejarczuk Paul 312
Investigating native Chinese speakers' identification of English consonants in noise (ID: 538) - Sheng Kai, Gong Jian, Sun Yuhong, Zhang Weizhong, Zhou Weijing, Wang Feng 317
Initial curvature of pitch contour affects fine-grained lexical tone perception of Mandarin but not Cantonese (ID: 545) - YE Yanyuan, RONG Yicheng, PENG Gang 322
Perception and imitation of period doubling (ID: 560) - Huang Yaqian 327
Perceptual similarity space of Serbian lexical pitch accents: evidence from Serbian and French listeners (ID: 571) - Nikolić Dušan, Winters Stephen 332
Does learning English vowels aid the perception of English stress for learners with small vowel inventory background? (ID: 580) - Allehyani Manal, Smith Rachel, Cohen Clara 337
Conditions on adaptation to an unfamiliar lexical tone system: the role of quantity and quality of exposure (ID: 601) - Zhao Liang, Sloggett Shayne, Chodroff Eleanor 342
Prose or poetry? A perceptive phonetic study (ID: 604) - Colonna Valentina, Romano Antonio 347
Clear speech sounds fast: The impact of speaking mode variation on perceived tempo (ID: 605) - Plug Leendert, Zheng Yue, Smith Rachel 352
Effects of FFP2 face masks on consonant identification (ID: 621) - Lengeris Angelos, Zatrazemi Joanna 357
Perception of standard Croatian pitch accents by speakers from pitch accent and from stress accent regiolects (ID: 627) - Pletikos Olof Elenmari, Vlašić Duić Jelena, Bićanić Jordan 361
Speech rate effects in French voicing production and perception (ID: 639) - Kang Yoonjung, Haggarty Nicholas 366
Mismatched coarticulatory information hinders lexical access of coronal stops in Malayalam (ID: 657) - G P Seema, Krishnaswamy Meghavarshini, Mishra Ramesh, Dutta Indranil 371
Perception of L2 English affricate onsets in native Mandarin listeners (ID: 661) - Wang Yizhou, Bundgaard-Nielsen Rikke 376
The role of immediately prior exposure and talker accent on sociolinguistic variant identification (ID: 663) - Li Aini, Tamminga Meredith 381
Free classification predicts discrimination of Finnish vowels by naïve Japanese listeners (ID: 667) - Daidone Danielle, Lidster Ryan, Kruger

Franziska 386
Effects of slowed rate of speech on eye gaze to the mouth in young adults: Clinical implications (ID: 671) - Shaw Virginia, Bernar Claire, Josafatow Nik, Bonnell Tyler, Li Fangfang 391
The representation of onset sibilant variants in Taiwan Mandarin: an ERP study (ID: 678) - Hsu Yun-Han, Lee Chia-Lin, Fon Janice 396
Perception of Malayalam three-way stop contrast among American English speakers (ID: 682) - Krishnaswamy Meghavarshini, Warner Natasha 401
Perception of Japanese moraic nasal in intervocalic position by Japanese native speakers and Korean learners of Japanese (ID: 701) - Han Heesun, Namba Koji 406
Testing the sound-driven hypothesis: an online rating experiment on the phonaesthetics of constructed languages (ID: 715) - Mooshammer Christine, Xia Qiang 411
Final lengthening in line end perception of re-synthesized recitations of German poems (ID: 723) - Schauffler Nadja, Koch Julia, Ketschik Nora, Bernhart Toni, Dieterle Felix, Eschenbach Gunilla, Kinder Anna, Richter Sandra, Viehhauser Gabriel, Vu Thang, Kuhn Jonas 416
The role of L2 fluency in the perception of non-native clusters by Turkish speakers (ID: 745) - Canalis Stefano 421
Language-specific stress discrimination by European Portuguese-learning infants – An ERP study (ID: 755) - Lu Shuang, Severino Cátia, Vigário Marina, Frota Sónia 426
Segmental Accuracy and Intelligibility: Native and nonnative listeners' perspectives (ID: 760) - Na In Young 431
Effect of context F0 on quantity in Finnish (ID: 770) - O'Dell Michael, Nieminen Tommi, Vakkilainen Joonas 436
Perception of Japanese pitch accent by Dutch listeners: from acoustic to phonological processing (ID: 781) - Hu Shuangshuang, Kager René .. 441
Perception of coarticulated tones in Taiwan Mandarin and Taiwan Southern Min (ID: 783) - Huang Po-Hsuan, Chiu Chenhao 446
Perception of voicing distinction in syllable-initial stops by multilingual speakers (ID: 788) - Cal Zuzanna, Wrembel Magdalena 451
The role of rhythmic stress in the perception of words (ID: 790) - Łukaszewicz Anna, Łukaszewicz Beata, Mołczanow Janina 456
Weighted phonetic distances in the evaluation of phonotactics (ID: 851) - Orzechowska Paula, Dziubalska-Kołaczyk Katarzyna 461
Animated film character profiles: the roles of voice and lexical content (ID: 862) - Crochiquia Alice, Eriksson Anders, Madureira Sandra, A. Barbosa Plinio 466
Energetic and informational masking effects on Spanish vowel discrimination (ID: 876) - Gibson Mark, Schlechtweg Marcel, Ayala Alcalde Judit, DiCiaccio Andrea, Wang Xianhui, Xu Li 471
Perceptual assimilation of Spanish, Thai, and Korean stop consonants by native English speakers (ID: 890) - Nagle Charlie, Bruun Shelby, Baese-Berk Melissa 476
Trochaic foot benefit in speech comprehension (ID: 898) - Scharinger Mathias, Domahs Ulrike 481
Effects of vowel acoustics on the intelligibility and comprehensibility of China English (ID: 905) - WANG QIAN 486
Generalization of training to an untrained phonetic environment: Initial /s-z/ to final /s-z/ (ID: 908) - Poulsen Birgitte, Rasmussen Sidsel, Bohn Ocke-Schwen 491
Investigating the the role of production in perceptual learning (ID: 913) - Huttner Lena-Marie, Nguyen Noel, Pickering Martin 496
Hearing beg and bag: Variation in the perception of pre-velar /æ/-raising (ID: 929) - Sullivan Lisa 501
"Kriech nicht da rein" – a new corpus of naturalistic misperception of German misheard sung speech (ID: 932) - Nieder Jessica, Tang Kevin 506
Voice discrimination across speaking styles in Persian (ID: 952) - Pellegrino Elisa, Asadi Homa, Dellwo Volker 511
The effect of speaker on speech intelligibility (ID: 978) - Wright Richard, Tucker Benjamin, Kelley Matthew C. 516
Illusions of ungrammaticality in the perception of non-native accented speech (ID: 979) - Wesołek Sarah, Gulgowski Piotr, Błaszczak Joanna, Żygis Marzena 521
Differences between L1 and L2 processing of meaningful sentences revealed by combining EEG and eye tracking (ID: 992) - Mulder Kimberley, Boves Lou, Brand Sophie 526
Perception of oral and nasal alveolar flap sounds in American English (ID: 994) - Garai Luca 531
Perception of pop-out voice in various conditions of babble noise (ID: 995) - Kitahara Mafuyu, Kawahara Hideki, Amano Shigeaki 536
L2 speech segmentation for word recognition: the role of lexical stress and syllable structure (ID: 1026) - Garrido-Pozu Juan 540
Tracking speaker-specific speech rate: habitual vs local influences on English stop voicing perception (ID: 1036) - Ting Connie, Kang Yoonjung 545
Evaluating the impact of disfluencies on the perception of speaker competence using neural speech synthesis (ID: 1040) - Kirkland Ambika, Włodarczak Marcin, Gustafson Joakim, Székely Éva 550
Listening with style: the effect of bidialectal style-shifting on PRICE vowel perception in Britain's Black Country (ID: 1051) - Packer-Stucki Thomas 555
Reorganization of the auditory-perceptual space across the human vocal range (ID: 1070) - Friedrichs Daniel, Dellwo Volker 560
Can L2 speakers perceive sociopragmatic nuances at the sound level? Exploring Chinese Listeners' Perceptions of (ING) (ID: 1075) - Lin Ke ... 565
When is enough, enough? VOT judgments vary by voicing intensity, aspiration intensity, voice quality, and rate of change (ID: 1080) - Kingston John, Rysling Amanda 570
Audiovisual cues in developmental speech perception: data from a voiceless fricative place discrimination (ID: 1082) - Bosch Laura, Birulés Joan, Pons Ferran 575

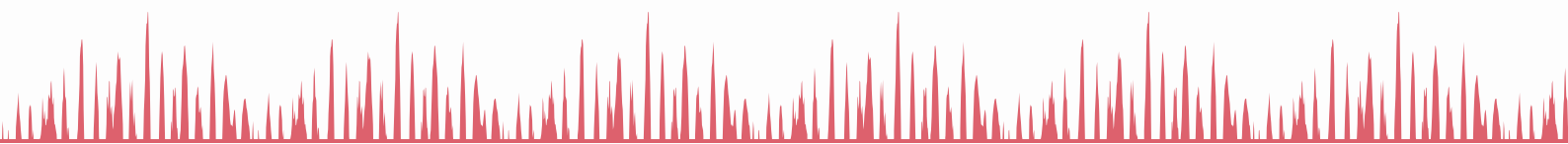


1. Special Session - Interplay or intermezzo? Structures and processes in prosody and music

Text-to-tune alignment in Japanese “Happy Birthday to You” (ID: 35) - Kubozono Haruo, Mizoguchi Ai 580
Spectral cues in pitch perception in English and Cantonese speech and music (ID: 332) - Chan May Pik Yu, Kuang Jianjing 585
The Decline of Tone and Melody Correspondence in Modern Thai Pop Songs?: A Preliminary Analysis (ID: 588) - Ketkaew Chawadon, Suwannarat Junyawan 590
VIP-Radar: a model for the phonetic study of poetry reading (ID: 709) - Colonna Valentina, Romano Antonio 594
The influence of musical abilities on the processing of L2 focus prosody: An eye-tracking study (ID: 722) - Jansen Nelleke, Loerts Hanneke, Harding Eleanor E., Başkent Deniz, Lowie Wander 599

2. Speech Acoustics

Durational adjustment in Cantonese focus marking of native and non-native speakers (ID: 9) - Yang Yike 604
On the sound source locations of ‘glottal fricative’ [h] (ID: 50) - Yoshinaga Tsukasa, Maekawa Kikuo, Iida Akiyoshi 609
Automatic classification of vowels in semi-spontaneous Pakistani Punjabi (ID: 82) - Jabeen Farhat, Wackerbarth Moritz, Wagner Petra 614
Non-uniform cue-trading: Differential effects of surprisal on pause usage and pause duration in German (ID: 90) - Yuen Ivan, Ibrahim Omnia, Andreeva Bistra, Möbius Bernd 619
Variation in perception and production of /s/-/ʃ/ in English (ID: 96) - Lo Roger, Liu Suyuan, Vaughn Charlotte, McAuliffe Michael, Babel Molly 624
Unreleased Plosive Consonants in Thai: an Acoustic Study (ID: 180) - CANO CÓRDOBA Paula Alejandra, TRAN Thi Thuy Hien, VALLÉE Nathalie, SAVARIAUX Christophe, GERBER Silvain, YAMLAMAI Nicha, ROSSIGNOL Morgane 629
Voice onset time (VOT) in Kurdish EFL learners (ID: 229) - Naser Lureen 634
Is focalization the keystone connecting vowel perception and production? (ID: 234) - Polka Linda, Masapollo Matthew, Menard Lucie, Schwartz Jean-Luc, Rapin Lucile 639
Fundamental frequency normalization and statistical power: an assessment of 15 normalizing techniques (ID: 280) - Genette Jérémy, Verhoeven Jo, Gillis Steven 644
Vowel production changes under noise with consideration of low-order formant masking (ID: 292) - Uezu Yasufumi, Akagi Masato, Unoki Masashi 649
Speaker-specific anticipatory labial coarticulation in French (ID: 320) - GUITARD-IVENT Fanny, WOHMANN-BRUZZO Louise, AUDIBERT Nicolas, FOUGERON Cécile 654
Viable signal periodicities in speech rhythm (ID: 346) - Campbell Jessica, Byrd Dani, Goldstein Louis 659
Amplitude envelopes as a means to study length contrasts: evidence from L1 and L2 Italian (ID: 381) - Einfeldt Marieke, Braun Bettina 664
A perceptual and acoustic study of melody in whispered Czech words (ID: 402) - Hanžlová Adléta, Bořil Tomáš 669
Prominence effects on Dutch low vowels (ID: 407) - Bruggeman Anna 674
Acoustic absement in detail: Quantifying acoustic differences in time-series speech data (ID: 415) - Kelley Matthew 679
The impact of the intensity ratio between vowels and voiceless plosives on the intelligibility of sung text (ID: 433) - Meister Einar, Vurma Allan, Dede Tuuri, Kala Veeda, Meister Lya, Raju Marju, Ross Jaan 684
Acoustic properties of bilabial trills in Mangbetu (ID: 446) - Olson Kenneth 689
The effect of stress and speech rate on vowel quality in spontaneous Central Mexican Spanish (ID: 450) - Hernandez Laura A., Perry Scott James, Tucker Benjamin V. 694
Measuring and modelling the duration of intervocalic alveolar taps in Peninsular Spanish (ID: 451) - Perry Scott James, Kelley Matthew C., Tucker Benjamin V. 699
Improved formant frequency measurements in Praat (ID: 463) - Weenink David 704
Constructing a short reading passage with phonemic and phonetic coverage for Thai: Traimit passage (ID: 474) - Onsuwan Chutamanee, Bairerkdee Kanwara, Chonmahatrakul Nawasri, Duangmal Juthatip, Amornlaksananon Atita, Denmuenwong Pimchanok 708
Infant Directed Speech expands F1/F2 space for /i a u/: But what happens to central vowels /ɜ ʊ/ in Australian English? (ID: 485) - Götz Antonia, Altuntas Eylem, Swinfield Nicole, Burnham Denis K., Best Catherine T. 713
Contribution of wavelet-based features to identification of overlapping Azerbaijani vowels using machine learning (ID: 517) - Ghaffarvand-Mokari Payam 718
Rime changes and mergers in Beijing retroflex suffixation: an acoustic study (ID: 518) - Wang Yutong, Sabev Mitko 723
Low density Assamese alveolar stops show greater coarticulatory resistance compared to labials and velars (ID: 522) - Thakuria Abhilasha, Gogoi Pamir, Dutta Indranil 728
The effects of voiced stops on adjacent vowel duration in Japanese (ID: 551) - Ishibashi Shoji, Koya Shonosuke 733
Acoustic properties of the four-way laryngeal contrast in Bengali infant directed speech (ID: 561) - Narkar Jahnavi 738
Spectral implications of codec compression on voiceless fricatives (ID: 587) - V. Christensen Krestina 743
Temporal vs. area-sum formulae of vowel nasality in simulated and nasometric corpora (ID: 600) - Dow Michael 748
Evolution of Voices in French Audiovisual Media Across Genders and Age in a Diachronic Perspective (ID: 631) - Rilliard Albert, Doukhan David, Uro Rémi, Devauchelle Simon 753



Differences in adult speech: Does boy-directed and girl-directed speech exist (ID: 641) - Keeping Tori, Robertson Cassandra, Li Fangfang 758
Neutralization and secondary acoustic cues of voicing contrast: A Tohoku and Tokyo Japanese production experiment (ID: 650) - Huang Chuyu, Matsui Sanae, Watabe Naoya, Noguchi Hiroto, Hashimoto Ayako, Mizoguchi Ai, Kitahara Mafuyu 763
Park or bark? VoT Production of English Bilabial Stops in Arabic Learners of English (ID: 699) - Al-Kendi Azza, Al-Aghbari Khalsa 768
The phonetic realization of intervocalic obstruents in Suzhou Wu Chinese (ID: 708) - Ge Chunyu, Mok Peggy 773
A production study of Korean consonants (ID: 726) - Perkins Jeremy, Lee Dahm, Lee Seunghun 778
Using machine learning to model the three-way laryngeal contrast in Korean (ID: 727) - Perkins Jeremy, Yan Yu, Lee Dahm, Lee Seunghun 783
The effect of prose and poetry text layout on the acoustics of reading aloud (ID: 793) - Nenadić Filip, Tucker Benjamin 788
What makes a geminate? An acoustic study of consonants in Garo (ID: 822) - A'gitok Cheman Baira, Athanasopoulou Angeliki 793
A diachronic study of vowel harmony in French broadcast speech since 1940 (ID: 823) - Cecelewski Juliusz, Gendrot Cédric, Boula de Mareüil Philippe, Adda-Decker Martine 798
Vocal effort in human interactions with voice-AI (ID: 834) - Bradshaw Leah, Perepelytsia Valeriia, Dellwo Volker 803
A comparison of acoustic convergence to an unfamiliar language and an unfamiliar accent in shadowed speech (ID: 889) - Hutchinson Amy, Zhou Alexis, Seo Yuhyeon, Dmitrieva Olga 808
Variability in Swedish voiceless fricative contrasts (ID: 906) - Wikse Barrow Carla, Włodarczak Marcin, Heldner Mattias, Strömbergsson Sofia 813
Acoustic phonetic classification of primary infant protophones (ID: 919) - Buder Eugene H., Oller D. Kimbrough 818
Dentofacial Disharmony patients' sibilants differ from controls' more in source than filter properties (ID: 947) - Oakley Madeleine, Tran Auvi, Paye Ciana, Trudan Emma, Turvey Timothy, Blakey George, Zajac David, Mielke Jeff, Jacox Laura 823
Acoustic Analysis of L2 English Phonemes for Language Identification (ID: 982) - Williams Samantha, Hughes Vincent, Foulkes Paul 828
Methodological trends in acoustic phonetic analysis (ID: 984) - Grabowski Emily 833
Impacts of remote and traditional recording methods on quality of formant analysis for vowel reduction (ID: 990) - Conklin Jenna 838
Acoustic correlates of the Javanese light vs. heavy distinction: a large-scale corpus study (ID: 1004) - Xu S.C. Angela, Wilson Colin 843
What French speakers' nasal vowels tell us about anticipatory nasal coarticulation (ID: 1049) - Rodriquez Francesco, Pouplier Marianne, Alderton Roy, Lo Justin J.H., Evans Bronwen G., Reinisch Eva, Carignan Christopher 848
Spirantization of word-final plosives in Standard Dutch (ID: 1074) - Pfiffner Alexandra, Martinez-Garcia Julian 853
Spontaneous imitation of English sibilants by native and non-native speakers (ID: 1106) - Hauser Ivy, Graham Emily, Zhang Xinwen 858

2. Special Session - Bridging linguistic and clinical perspectives through computational models of speech production

Measuring Gradient Effects of Alcohol on Speech with Neural Networks' Posterior Probability of Phonological Features (ID: 469) - Wayland Ratree, Tang Kevin, Wang Fenqi, Vellozzi Sophia, Sengupta Rahul 863
Different sensitivity to temporal auditory feedback perturbation in adults who stutter: syllable structure effects (ID: 861) - Oschkinat Miriam, Hoole Philip 868
Auditory acuity and response to altered auditory feedback in children with and without residual speech sound disorder (ID: 927) - Eads Amanda, Kearney Elaine, Koda Asuka, Hitchcock Elaine, Shiller Douglas, Guenther Frank, McAllister Tara 873
Using computational models to characterize the role of motor noise in speech: The case of amyotrophic lateral sclerosis (ID: 988) - Parrell Benjamin, Mefferd Antje, Harper Sarah, Roessig Simon, Mücke Doris 878

3. Speech Production and Speech Physiology

A review of palatographic measurement devices developed at the TU Dresden from 2011 to 2022 (ID: 27) - Birkholz Peter, Stone Simon, Wagner Christoph, Kürbis Steffen, Wilbrandt Alexander, Bosshammer Mats 883
Preaspiration in Italian voiceless geminate and singleton stops (ID: 74) - Dian Angelo, Hajek John, Fletcher Janet 888
Responses to auditory perturbation involve a mix of opposing and following responses in steady vowels and glissandos (ID: 79) - Ning Li-Hsin . 893
An articulatory study of Estonian palatalization (ID: 106) - Malmi Anton, Lippus Pärtel 898
Lateral and central tongue bracing coordination in typical speakers and cochlear implant users (ID: 119) - Liker Marko 903
Coarticulatory resistance and aggression in pitch-accented vowels under the effect of vowel-to-vowel coarticulation (ID: 121) - Deme Andrea, Bartók Márton, Csapó Tamás Gábor, Gráczi Tekla Etelka, Juhász Kornélia, Markó Alexandra 908
Towards a material point method vocal tract simulation (ID: 134) - Moisik Scott 913
Co-registration of simultaneous high-speed ultrasound and electromagnetic articulography for speech production research (ID: 145) - Kirkham Sam, Strycharczuk Patrycja, Gorman Emily, Nagamine Takayuki, Wrench Alan 918
5-minute formant adaptation task in Dutch children (ID: 209) - Rebernik Teja, Tienkamp Thomas, Polsterer Katharina, Buurke Raoul, Hukker Vera, Medvedeva Masha, van der Ploeg Mara, Schepers Iris, Sekeres Hedwig, de Vries Wietse, Abur Defne, Jonkers Roel, Noiray Aude, Wieling Martijn 923

What does successful L2 vowel acquisition depend on? A conceptual replication (ID: 225) - Witteman Jurriaan, Schiller Niels, McQueen James, Karaseva Ekaterina 928
Talking chorally alters speech rhythm and induces fluency in people who stutter- but are these things connected? (ID: 235) - Meekings Sophie, Eijk Lotte, Scott Sophie, Maruthy Santosh 932
Syllable position effects with French and Spanish /l/ (ID: 237) - Colantoni Laura, Kochetov Alexei, Steele Jeffrey 937
Optimal control theory of speech production using probabilistic articulatory-acoustic models (ID: 261) - Elie Benjamin, Šimko Juraj, Turk Alice .. 942
Being looked at in the face attenuates articulatory movements in both sighted and blind speakers (ID: 279) - Saito Haruka, Trudeau-Fisette Paméla, Vidou Camille, Ménard Lucie 947
Ejectives in Georgian. A real-time MRI analysis of vertical larynx movement (ID: 301) - Sulaberidze Nato, Brandt Erika, Hoole Phil, Krämer Martin, Reichenbach Jürgen R., Simpson Adrian P. 952
The retroflex lateral [ɭ] as an allophone for /l/ in Standard Austrian German (ID: 316) - Luttenberger Jan, Reinisch Eva 957
Assessing kinematic relations with high speech rate resolution data (ID: 322) - Kuberski Stephan R., Gafos Adamantios I. 962
Visualising phonetics at the neuromuscular level reveals the discrete pulsatile gated nature of speech (ID: 334) - Wrench Alan 967
The role of listener feedback in prosodic cue production: an interactive task (ID: 377) - Chatterjee Lahari, Schneider Kathleen, Wartenbuger Isabell, Hanne Sandra, Hofmann Andrea, Tuomainen Outi 972
Aerodynamic parameters of ejectives and pulmonic stops (ID: 391) - Solé Maria-Josep 977
Effects of autonomic arousal and cognitive load on sensorimotor adaptation of voice (ID: 396) - Tomassi Nicole, Turashvili Dea, Williams Alyssa, Stepp Cara 982
Stop voicing and devoicing as articulatory tasks: A cross-linguistic RT-MRI study (ID: 416) - Zhang Yubin, Goldstein Louis 987
Articulatory characteristics of the Japanese /r/: A real-time MRI study (ID: 443) - Maekawa Kikuo 992
An articulatory study of the word-initial /stɹ/ cluster in British English (ID: 483) - Nakamura Mitsuhiro 997
Markerless tongue pose estimation in cross-linguistic spoken and mimed speech using ultrasound (ID: 507) - Lukose Annslin Maria, Gully Amelia, Bailey George 1002
Nasal coda merger in Hangzhou Mandarin: a dynamic ultrasound study (ID: 559) - Huang Yin, Ma Liang, Carignan Christopher 1007
Articulation of the nasal vowels in Shanghai Chinese (ID: 574) - Chen Changhe, Havenhill Jonathan 1012
Labialization in Amazigh: acoustic and articulatory marking over time (ID: 592) - Buech Philipp, Hermes Anne, Ridouane Rachid 1017
Synchrony and stability of articulatory landmarks in English and Mandarin CV sequences (ID: 611) - Kramer Benjamin M., Stern Michael C., Wang Yichen, Liu Yuyang, Shaw Jason A. 1022
The coarticulatory behavior of Standard Mandarin apical vowels (ID: 628) - Foley Sean 1027
Quantifying tongue-tip to upper incisor distance in an Ultrasound Tongue Imaging study of Swedish /i/ (ID: 634) - Westerberg Fabienne, Lawson Eleanor, Stuart-Smith Jane 1032
The role of feedback in lip-tube perturbation of Taiwan Mandarin rounded vowels (ID: 656) - Huang Jian-zhi, Chiu Chenhao 1037
Effect of individual vocal tract geometry on the tongue shaping for American English /ɹ/ (ID: 690) - Lu Yijing, Hsu Haley, Goldstein Louis, Toutios Asterios 1042
Individual- and group-level associations between articulatory and acoustic variability for English consonants (ID: 697) - Harper Sarah 1047
Lip postures of high vowels in Taiwan Mandarin (ID: 703) - Chiu Chenhao, Huang Po-Hsuan 1052
Phase relations between the tongue body and the jaw across rate modifications in younger and older speakers (ID: 707) - Thies Tabea, Mücke Doris, Ding Angela, Mefferd Antje 1057
Anticipatory labialization in C[u] clusters: a cross-linguistic study (ID: 711) - Volenec Veno 1062
Articulation of Cantonese and English sibilants among bilingual speakers in Hong Kong (ID: 742) - Havenhill Jonathan, Liu Ming, Li Tak Wang 1067
The relationship between vowel change and nasal loss in the Chengdu dialect of Mandarin Chinese: Evidence from rt-MRI (ID: 778) - Liao Sishi, Hoole Phil, Harrington Jonathan 1072
Dynamic articulatory and acoustic features of Hungarian sibilants as a function of phonological voicing (ID: 786) - Gráczi Tekla Etelka, Juhász Kornélia, Csapó Tamás Gábor, Deme Andrea, Markó Alexandra 1077
Speech breathing during physical activity is opportunistic (ID: 802) - Weston Heather 1082
Articulatory and acoustic analysis of coarticulated /u/ and /y/ produced by female and male speakers of German (ID: 803) - Bučar Shigemori Lia Saki, Franzke Rosa, Hoole Philip 1087
The l~n merger in Southwestern Mandarin: an articulatory study (ID: 813) - Huang Jing, Shibata Kye, Hsieh Feng-fan, Chang Yueh-chin, Tiede Mark 1092
Comparing apples to oranges - asynchrony in jaw & lip articulation of syllables (ID: 842) - Svensson Lundmark Malin, Erickson Donna 1097
Individual differences in speech production: What is "phonetic substance"? (ID: 867) - Johnson Keith 1102
Modeling the temporal evolution of the vocal tract shape with deep learning techniques (ID: 887) - Laprie Yves, Ribeiro Vinicius, Isaieva Karina, Vuissoz Pierre-André, Leclere Justine 1107
Modification of EEG brain activity induced by respiratory muscle regulation during phonation (ID: 912) - Hashemi Iraj, Demolin Didier, Cheron Guy, Cebola Alvarez Anita, Petieau Mathieu 1112
Similar and different tongue surface contours: intra-speaker controls in ultrasound analysis (ID: 921) - Scobbie James 1117

How speech rate, syllabic complexity and diversity affect the emergence of speech rhythm in speeded syllable repetition (ID: 923) - Lancia Leonardo, Li Jinyu, Fougeron Cécile 1122
Respiratory and supralaryngeal effects on speech breathing noise across loudness conditions and speaking tasks (ID: 924) - Fuchs Susanne, Koenig Laura L., Werner Raphael 1127
Speakers compensate for temporal perturbations in auditory feedback regardless of syllable structure (ID: 931) - Karlin Robin, Naber Chris, Parrell Benjamin 1132
CV coordination: The case of enchaînement and liaison in French (ID: 961) - Oh Sejin, Fougeron Cecile, Buech Philipp, Hermes Anne 1137
The nature of acoustic goals referenced during hyperarticulated speech (ID: 969) - Diantoro Carissa, Redford Melissa 1142
Monolingual and plurilingual strategies in the articulation of French R : a case study (ID: 971) - Lara Andres Felipe, Hermes Anne, Oh Sejin, Pillot-Loiseau Claire 1147
Non-pulmonic initiation in human beatboxing : a real-time mri study (ID: 998) - Dehais-Underdown Alexis, Vignes Paul, Crevier-Buchman Lise, Demolin Didier, Vuissoz Pierre-André, Isaieva Karyna, Fauvel Marc, Laprie Yves, Felblinger Jacques 1152
Speed rate effects on human beatboxing: an aerodynamic study (ID: 1001) - Dehais-Underdown Alexis, Vignes Paul, Crevier-Buchman Lise, Demolin Didier 1157
Piaroa voiceless stops as partial undergoers of nasal harmony (ID: 1005) - Rosés Labrada Jorge, Faytak Matthew, Schnoor Tyler, Lapierre Myriam, Michael Lev 1162
The lombard effect on the vowel space of northern vietnamese (ID: 1013) - Le Giang, Tang Yan 1167
Triggering and maintenance of distinct tongue postures (ID: 1019) - Purnomo Annabelle, Liu Yadong, Gick Bryan, Ebbutt Nicole 1172
Speech sound stability over time: Evidence from Norwegian vowels in spontaneous speech production (ID: 1030) - Li Peng, Flege James E., Martin Clara D., Kartushina Natalia 1177
Individual differences in vowel compactness persist under intoxication across first and second languages (ID: 1035) - Chang Charles, Tang Kevin, Nevins Andrew 1182
Article]{Domain adaptation towards speaker-independent ultrasound tongue imaging-based articulatory-to-acoustic Mapping (ID: 1095) - You Kang, Xu Kele, Ming Feng, Wang Jilong, Zhu Boqing, Tam'as G'abor Csap'o, Feng Dawei 1187
Larynx raising in English word-final ejective stops: A real-time MRI study (ID: 1111) - Price Lavinia, Pouplier Marianne, Hoole Philip 1192

3. Special Session - Novel approaches to studying vocal development

The Predictability of naturalistic evaluation of all-day recordings for speech and language development (ID: 691) - Ha Seunghee, Yoon Tae-Jin, So Jungmin 1197
Canonical babbling ratio extracted from day-long audio recordings: A preliminary report from India (ID: 698) - Meera Shoba, Swaminathan Divya, V Sri Ranjani, Srikar Malavi, Raju Reny 1201
Reinforcing infant vocalisations through self-initiated feedback: A tool for basic research and clinical interventions (ID: 750) - Keren-Portnoy Tamar, Daffern Helena, DePaolis Rory, Cox Christopher, Brown Kenneth, Oxley Florence, Kanaan Mona 1206
The development of canonical proportion continues past toddlerhood (ID: 774) - Hitczenko Kasia, Bergelson Elika, Casillas Marisa, Colleran Heidi, Cychosz Margaret, Grosjean Pauline, Hamrick Lisa R., Kelleher Bridgette L., Scaff Camila, Seidl Amanda, Walker Sarah, Cristia Alejandrina 1210
The acquisition of vowels differing in phonemic vowel length: The case of Hungarian (ID: 989) - Zajdó Krisztina 1215

4. Speech Prosody

Prosody of pre-focal background depends on following focus (ID: 17) - Roessig Simon 1220
Poem declamation in two varieties of Portuguese: the link between prosody and pleasantness (ID: 26) - Barbosa Plinio 1225
The influence of syllable structure on lexical tones in Croatian: no effect? (ID: 30) - Tabain Marija, Kapović Mate, Gordon Matthew, Gregory Adele, Beare Richard 1230
Perception of lexical pitch accent and rising shape in South Kyungsang Korean (ID: 33) - Joo Hyunjung, D'Imperio Mariapaola 1235
Testing the locus of speech-act meaning in english intonation (ID: 37) - Sostarics Thomas, Cole Jennifer 1240
Correlation of prosodic boundary cues in German (ID: 41) - Zerbian Sabine, Schubö Fabian 1245
Uptalk and the Frequency Code: how gender affects iconic associations of pitch (ID: 73) - Calhoun Sasha, Warren Paul, Mills Joy, Agnew Jemima 1250
Continuation rises in pre-1974 Cypriot Greek (ID: 95) - Przedlacka Joanna, Baltazani Mary, Armostis Spyros, Unal Özlem, Coleman John 1255
To what extent do French listeners perceive differences in accent location within words? An EEG investigation (ID: 107) - Rassili Outhmane, Michelas Amandine, Dufour Sophie 1260
Realizations of the tritonal pitch accent in Paraguayan Guarani (ID: 109) - Jun Sun-Ah, Siah Jian-Leat, Cabrera Marisabel, Johnson Hunter 1265
Perceived naturalness of accents in noun phrases by mono- and bilingual listeners of Russian (ID: 111) - Zerbian Sabine, Barabashova Kristina 1270
Evaluating prosodic aspects of oral reading proficiency in schoolchildren: effects of gender, genre and grade (ID: 113) - Wehrle Simon,

Sappok Christopher 1275
The influence of empathy and autistic-like traits in prominence perception (ID: 175) - Orrico Riccardo, Gryllia Stella, Kim Jiseung, Arvaniti Amalia 1280
So near yet so far: fine phonetic differences in tonal alignment and temporal structure in close contact varieties (ID: 197) - Pöhnlein Jasmin, Kleber Felicitas 1285
Prosody-syntax interface and Hungarian noun phrases (ID: 201) - Langer Corinna, Kügler Frank 1290
Sometimes low really is just the opposite of high: perception of low F0 targets by tone and non-tone language speakers (ID: 211) - Barnes Jonathan, Lee Chris K. C., Brugos Alejna, Shattuck-Hufnagel Stefanie, Veilleux Nanette 1295
Contour shape and prosodic focus marking: the case of Hong Kong Cantonese (ID: 213) - Lee Chris K. C., Barnes Jonathan 1300
Interactions of lexical stress, vowel length and pharyngealisation in Palestinian Arabic (ID: 218) - Kelly Niamh 1305
Prosodic convergence across varieties of Italian (ID: 228) - Savino Michelina, Torreira Francisco, Grice Martine 1310
Rhythmic Similarity hypothesis for cross-cultural vocal emotion recognition (ID: 253) - Liang Yachan, Choi Jiyoun, Broersma Mirjam, Goudbeek Martijn, Konopka Agnieszka 1315
Redundancy and individual variability in the prosodic marking of information status in German (ID: 278) - Lorenzen Janne, Roessig Simon, Baumann Stefan 1320
Perceptual Assimilation Model for Suprasegmentals predicts Cantonese tone sequence recall in Punjabi minorities (ID: 289) - YU Doris, LAI Veronica Ka Wai, CHOI William 1325
The Role of Prosody and Beat Gesture in Enhancing Memory for Discourse Information in Mandarin (ID: 294) - Yan Mengzhu, Calhoun Sasha 1330
Accent realization in spontaneous speech: Greek H* and L+H* (ID: 295) - Gryllia Stella, Orrico Riccardo, Kim Jiseung, Hu Na, Arvaniti Amalia 1335
Prosody and word order in marking focus within Djambarrpuyŋu noun phrases (ID: 310) - Jepson Kathleen 1340
The perception of parallel encoded emotional and linguistic prosody in Mandarin Chinese (ID: 311) - Zhang Qianyutong, Li Shanpeng 1345
Boundary strength and predictability effects on durational cues at Tone Sandhi Group boundaries in Taiwan Southern Min (ID: 312) - Wang Sheng-Fu 1350
Reject?! On the prosody of non-acceptance (ID: 313) - Repp Sophie, Seeliger Heiko 1355
Contrast or context, that is the question (ID: 378) - cangemi francesco, Grice Martine, Jeon Hae-Sung, Setter Jane 1360
Implicit imitation of intonation contours in word shadowing (ID: 394) - Song Yu Jin, Clopper Cynthia G. 1365
Enhancement of intonational contrasts in American English (ID: 400) - Cole Jennifer, Steffman Jeremy 1370
Temporal determinants of phrasing in coordinated structures in Greek (ID: 411) - Nicolaidis Katerina, Baltazani Mary 1375
F0-based Pairwise Variability Index: A Prosodic Metric for Holistic Language Processing (ID: 412) - Chen Alvin Cheng-Hsien 1381
The interaction of discourse markers with prosodic boundaries in ESP (ID: 432) - Ballier Nicolas, Henderson Alice, Méli Adrien, Arnold Taylor 1386
The effect of F0 and post-focal compression on the perception of narrow focus in Hebrew (ID: 479) - Amir Noam, Anavi Ori, Deouell Ella, Mixdorff Hansjoerg, Sliber-Varod Vered 1391
Gender-related variation of nasality and sound change of denasalization driven by prosodic boundaries in Seoul Korean (ID: 484) - Lee Jungah, Cho Taehong, Kim Sahyang 1395
Word-level and syllable-level predictability effects on syllable duration in Taiwan Mandarin (ID: 496) - Wang Sheng-Fu, Tseng Shu-Chuan 1400
Boundary-driven account for downstep in Japanese (ID: 497) - Furukawa Kei, Nakamura Satoshi 1405
The Interaction of Intonation Boundary and Lexical Tones under Two Focus Conditions in the Wuzhi Dialect of Chinese (ID: 498) - Zhu Yuzhu, Li Zhiqiang, Li Aijun 1410
An acoustic study of V-to-V coordination within and across words in French (ID: 511) - Yildiz Alice, D'Alessandro Daria, Oh Sejin, Hermes Anne, Fougeron Cécile 1415
Variations of focus prominence in three tone languages (ID: 512) - Li Katrina Kechun, Nolan Francis, Post Brechtje 1420
Mandarin question intonation patterns in the production of Hungarian learners of Chinese (ID: 520) - Juhász Kornélia, Bartos Huba 1425
The absence of pauses in spoken narratives and memory recall (ID: 535) - Volín Jan, Šturm Pavel 1430
The effect of prosodic prominence on diphthong reduction in Taiwan Mandarin spontaneous speech–using /aɪ/ as an example (ID: 543) - Chen Chieh-Ching, Fon Janice 1435
Prosodic boundary strength and phrase-initial syllable duration (ID: 549) - Kentner Gerrit, Franz Isabelle, Knoop Christine, Menninghaus Winfried 1440
Prosody of corrective focus in Japanese complex DPs (ID: 556) - Chan Le Xuan, Furusawa Rina, Tsujita Rin, Lee Seunghun 1444
Regional variation, articulation rate, and pausing patterns in three varieties of Spanish (ID: 615) - Garcia-Amaya Lorenzo, Kendro Kelly, Henriksen Nicholas 1449
What does BERT learn about prosody? (ID: 622) - Kakouros Sofoklis, O'Mahony Johannah 1454
The phonetics-phonology-syntax interface: a computational implementation (ID: 629) - Bögel Tina, Hill Romi, Hofenbitzer Justin, Zhao Tianyi . 1459
Macro- and Micro-Prosodic Changes in the Duration of Brazilian Portuguese Vowels (ID: 632) - de Moraes João Antônio, Rilliard Albert 1464
American English pitch accent dynamics: A minimal model (ID: 646) - Iskarous Khalil, Cole Jennifer, Steffman Jeremy 1469

Differential effects of prosodic boundary on glottalization of word-initial vowels in Korean (ID: 651) - Hwang Sungwok, Kim Sahyang, Cho Taehong 1474
Speech rhythm metrics: a typological survey (ID: 664) - Gordon Matthew, Katsika Argyro, Chien Sherry, Jang Jiyoung, Lai Ryan Ka Yau 1479
How does focus-induced prominence influence realization of edge tones and segmental anchoring in Seoul Korean? (ID: 677) - Hatcher Richard, Joo Hyunjung, Kim Sahyang, Cho Taehong 1484
Prosodic prominence of Swahili TU in polar interrogatives (ID: 679) - Lee Seunghun J., Kamano Shigeto, Abe Yuko, Miyazaki Kumiko 1489
Some asymmetrical pre- versus post-focal effects on articulatory realization of prominence distribution in Korean (ID: 681) - Im Suyeon, Kim Sahyang, Cho Taehong 1494
Examining variability in the production and perception of prosodic attitudes (ID: 695) - Kim Jeesun, Raue Kate, Davis Chris 1499
Alignment of prosodic and syntactic junctures and vowel-initial glottalization in syntactic disambiguation in English (ID: 714) - Shin Jae-Eun, Kim Sahyang, Cho Taehong 1504
Variability in the stress group pattern of Copenhagen Danish with a focus on the Copenhagen multiethnolect (ID: 733) - Pharao Nicolai 1509
Prosodic alignment in different conversational feedback functions (ID: 734) - Figueroa Carol, Benus Stefan, Skantze Gabriel 1514
Extraction of emotional prosody from telephone calls to hospital emergency departments (ID: 744) - WEI Dongjun, EMBARKI Mohamed, BARAKAT Oussama, DESMETTRE Thibaut, MARX Tania, VAUCHEREY Coralie, ROBERT Stephan 1519
The domain of boundary-related lengthening in Tokyo Japanese, a mora-timed language (ID: 753) - Chien Sherry, Tsai Karen, Katsika Argyro 1524
Functions of Amfo'an (Meto) prosody: a preliminary acoustic study (ID: 754) - Culhane Kirsten 1529
The perception of surprise questions in Estonian (ID: 787) - Lippus Pärtel, Asu Eva Liina, Leppik Katrin, Sahkai Heete 1533
Oral reading proficiency and prosody – a perceptual pilot study on especially fluent German students (grade 3 to 7) (ID: 795) - Sappok Christopher 1538
Contrast and givenness in biased declarative questions (ID: 808) - Repp Sophie, Seeliger Heiko 1543
The effect of (non-)native language and task complexity on speech entrainment (ID: 811) - Mareková Lucia, Kruyt Joanna, Beňuš Štefan 1548
Perceptual Evaluation of Natural and Synthesized Prosody in Chinese Learners' English Production (ID: 819) - Chen Ying, Li Yan, Xiao Wentao, Liu Li 1553
Macro- and Micro-rhythm in L2 English: Exploration and Refinement of the Measures (ID: 830) - Ortega-Llebaria Marta, SilvaJr. Leonidas, Nagao Jun 1558
The edge tones inventory in Standard Russian prosody: are phrase accents required? (ID: 832) - Duryagin Pavel 1563
Evidence for preplanned and adaptive F0 control (ID: 839) - Kim Seung-Eun 1568
Investigations of F0 control: pitch targets vs. pitch register (ID: 852) - Kim Seung-Eun, Tilsen Sam 1573
Intonational preferences for lexical contrast and verum focus (ID: 865) - Röhr Christine, Grice Martine, Baumann Stefan 1578
Prosody in contact: Polar questions in Cypriot varieties of Greek and Turkish (ID: 866) - Payne Elinor, Armostis Spyros, Lombardo Eftychia, Simon Rainer, Baltazani Mary 1583
Prosodic predictors of temporal structure in conversational turn-taking (ID: 869) - Franich Kathryn 1588
Language redundancy effects on f0: A preliminary controlled study (ID: 877) - Zhang Cong, Lai Catherine, Napoleão de Souza Ricardo, Turk Alice, Bögel Tina 1593
Predicting pitch accent patterns in Osaka Japanese using a machine translation method with a transformer (ID: 878) - Noguchi Hiroto 1598
Do words sing their own tunes? Word-specific pitch realizations in Mandarin and English (ID: 895) - Chuang Yu-Ying, Baayen R. Harald, Bell Melanie J. 1603
Prosodic cues of distinguishing neutral and non-neutral yes/no questions in Hungarian: the acoustics of surprise (ID: 909) - Mády Katalin, Reichel Uwe D., Kohári Anna, Molnár Cecília Sarolta, Szalontai Ádám 1608
Explicit rules or implicit imitation: a comparative study of two approaches to teaching English prosody (ID: 925) - Zhang Xiaodan, Romero Joaquin 1613
Boundary perception as the interplay of acoustic cues, syntactic constituency and prominence (ID: 940) - Kuang Jianjing, Chan May Pik Yu, Rhee Nari 1618
Between tone and stress: acoustic evidence for lexical prominence in Siriano (ID: 943) - Ni Tianyi, Köhnlein Björn, Morley Rebecca 1623
The Phonetics of Downtrends in Chokri (ID: 948) - Gogoi Tulika, Tetseo Sekholu, Gope Amalesh 1628
French Rhythm in Read Speech by French Natives and Chinese L2 Learners (ID: 953) - XI Lei, Ridouane Rachid 1633
The Phonetics of Prosodic Marking of Focus in Sylheti (ID: 957) - Gogoi Tulika, Gope Amalesh 1638
The phonetics and phonology of corrective focus in xiangshan wu chinese: the interaction with disyllabic tone sandhi (ID: 960) - Shi Yibing, Nolan Francis, Post Brechtje 1643
Methods for PoLaR Exploration with Machine Learning: Grammatical Analysis of Intonation without Grammatical Labels (ID: 1007) - Veilleux Nanette, Ahn Byron, Brugos Alejna, Jeong Sunwoo, Shattuck-Hufnagel Stefanie 1648
The kinematic properties of prosodic boundaries in conversational turn-taking (ID: 1008) - Purse Ruaridh, Krivokapić Jelena 1653
The prosody of high and low affective engagement in Polish and German parliamentary speeches (ID: 1033) - Karpiński Maciej, Klessa Katarzyna, Sawicka-Stępińska Brygida, Kasperek Hanna 1658
Prosodic realization of syntactic phrase and clause boundaries in Tokyo Japanese (ID: 1038) - Ishihara Shinichiro 1663

Perceptual learning of novel intonation patterns by Cantonese and Wu speakers (ID: 1056) - Lam Man Yan Priscilla, Xu Wenwei, Chen Yiya .. 1668
Marking prominence: Towards cue-based annotation of prosodic prominence (ID: 1064) - Brugos Alejna, Breen Mara, Shattuck-Hufnagel Stefanie, Veilleux Nanette, Barnes Jonathan 1673
Focus prosody in Fijian: In-situ focus marking (ID: 1065) - Mut Tsz Ching Summer, Simard Candide, Tamata Apolonia, Lee Albert 1678
Detecting the Accentual Phrase boundaries in Seoul Korean using tonal and segmental cues (ID: 1079) - Lee Seung Suk 1681
Articulatory manifestation of focus-induced prominence in Seoul Korean (ID: 1086) - Jang Jiyoung, Katsika Argyro 1686
Temporal scope of articulatory slowdown under informational focus: Data from English and Mandarin (ID: 1089) - Liu Yuyang, Wang Yichen, Stern Michael C., Kramer Benjamin M., Shaw Jason A. 1691
A hierarchy of prominence: The production and perception of focus in American English (ID: 1090) - Katsika Argyro, Jang Jiyoung, Krivokapic Jelena, Goldstein Louis, Saltzman Elliot 1696

4. Special Session - The pedagogy of the Laryngeal Articulator Model

The pedagogy of the Laryngeal Articulator Model (ID: 185) - Esling John H., Benner Allison, Calamai Silvia, Celata Chiara, Crevier Buchman Lise, Hejná Míša, Moisik Scott 1701

5. Phonation and Voice Quality

Voice quality in Norwegian: effects of dialect, speaker age, sex, and speaking mode (ID: 42) - van Dommelen Wim 1706
Mandarin prosodic focus by speakers with autism spectrum disorders (ID: 56) - Pan Ho-hsien, Lyu Shaoren, Lin Hai-Ti, Gau Susan Shur-Fen 1711
Creapy: A Python-based tool for the detection of creak in conversational speech (ID: 59) - Paierl Michael, Röck Thomas, Wepner Saskia, Kelterer Anneliese, Schuppler Barbara 1716
The influence of common colds and its interaction with lexical tones on voice quality of Mandarin speakers (ID: 133) - Mou Yifan, Zhu Lei 1721
Exploring the role of formant frequencies in the classification of phonation type (ID: 186) - Ludusan Bogdan, Heldner Mattias, Włodarczak Marcin 1726
Vocal fold vibratory patterns in bilingual speakers of Low and High German (ID: 212) - Peters Jörg, Frank Marina, Rohloff Marina 1731
Effects of fundamental frequency and harmonics-to-noise ratio on the perception of Danish laryngealized phonation (ID: 219) - Peña Jailyn 1736
Duration as a cue for phonological voicing contrast in whispered Czech (ID: 243) - Svatošová Michaela, Bořil Tomáš 1741
Speech Emotion Recognition from Glottal Flow Signals using 1D Convolutional Neural Networks (ID: 251) - Ben-Dom Itay, Watson Catherine, McCann Clare 1746
Phonetics of glottalized phonations in Yateé Zapotec (ID: 270) - Chai Yuan, Fernández Adrian, Mendez Briseida 1751
Labels for voices (ID: 347) - Kreiman Jody 1756
What is a 'substantial' voice? An LTAS study on two Cantonese accents (ID: 385) - Wong Eugene, Fung Roxana 1761
Vowel-initial words and glottalization: a corpus study of Chichicastenango K'iche' (ID: 395) - Wood Elizabeth 1766
Laughter patterns in multi-speaker conversation data: comparison between spontaneous laughter and intentional laughter (ID: 449) - Wang Kexin, Ishi Carlos, Hayashi Ryoko 1771
Cross cultural perception of valence and arousal (ID: 477) - Erickson Donna, Rilliard Albert, Li Yongwei, Menezes Caroline, Kawahara Shigeto, Sadanobu Toshiyuki, Hayashi Ryoko, Shochi Takaaki, de Moraes João, Obert Kerrie 1776
Prosodic word as the domain of breathiness: Evidence from disyllabic words in Kunshan Wu (ID: 480) - Xu Wenwei, Mok Peggy 1781
A Perception Study on Voice Quality and Stance in Mandarin Chinese (ID: 492) - Li Qiuyuan, Mok Peggy 1786
Convergence of creaky voice use in Australian English (ID: 505) - White Hannah, Gibson Andy, Penney Joshua, Szakay Anita, Cox Felicity ... 1791
The role of creaky voice in Cantonese tonal production (ID: 519) - Ho Chu Yan 1796
Perceptual effects of aliasing distortion in glottal flow modelling (ID: 537) - Wang Zihan, Gobl Christer 1801
Contributions of acoustic measures to the classification of laryngeal voice quality in continuous English speech (ID: 772) - Xu Chenzi, Foulkes Paul, Harrison Philip, Hughes Vincent, Wormald Jess 1806
On the role of glottal stop: from boundary marker to correlate of focus. An experimental study on Italian and French (ID: 821) - De Paolis Bianca Maria, Abbà Bianca, Romano Antonio 1811
Creaky voice in three Scots varieties: Using f0-based identification to consider social and linguistic factors (ID: 824) - Pearce Joe 1816
What is glottal whistle? – Exploring extremly high fundamental frequencies in human vocal production (ID: 841) - Grawunder Sven, Wingerath Angela 1821
Cepstral peak prominence in normophonic adults: the effect of gender and speech task segmental composition (ID: 857) - O Regan Saoirse, Yanushevskaya Irena 1826
Variation and dialectal differences in the segmental scope of phonologically contrastive laryngealisation (ID: 881) - Siem Andrea 1831
Interaction between prosodic position and obstruent voicing: a case of English lettered word in Wenzhou Wu Chinese (ID: 896) - Zhang Weijun, Mok Peggy Pik Ki 1836
Creaky voice in L2 English and L1 Dutch (ID: 964) - Sebregts Koen, Vriesendorp Hielke, Quené Hugo, White Yosiane 1841

Principles of voice quality theory: A rtMRI study on labiodental production (ID: 1071) - Lee Daniel, Moisik Scott, Vijayaragavan Vimalan 1846

5. Special Session - Qualitative research in phonetics

I can be both? (Pre-)aspiration and (pre-)glottalisation do not have to be mutually exclusive (ID: 215) - Hejná Míša (Michaela) 1851
The devil is in the detail: An interactional-phonetic study of G-word interjections and some methodological implications (ID: 263) - Cantarutti Marina 1856
Game of phones: a socio-phonetic analysis of stylised media performance of yorkshire english (ID: 737) - Jackson Lucy 1861
Teaching language rhythm by music to older English learners of Italian (ID: 930) - Zambianchi Laura, Jeon Hae-Sung 1866

6. Tone

The influence of tone experience and native tone and intonation categories on nonnative tone learning (ID: 63) - Chin Jessica L. L., Antoniou Mark 1871
Tone as a factor influencing the dynamics of diphthong realizations in Standard Mandarin (ID: 70) - Li Chenyu, Al-Tamimi Jalal, Wu Yaru 1876
The role of prior knowledge in second-language learners' overnight consolidation of Cantonese tones (ID: 157) - Qin Zhen, Jin Rui, Wu Ruofan 1881
Voicing and register in Mnong Râlâm (ID: 177) - Brunelle Marc, Đinh Lư Giang, Tạ Thành Tấn 1886
Realization of low tone sequences in disyllabic words in a large Mandarin speech corpus (ID: 224) - Wu Yaru, Chen Yiya, Lamel Lori 1891
L2 phonological contrasts are not necessarily lost: Evidence from French listeners detecting Mandarin tones (ID: 272) - Dellwo Volker, Schwab Sandra, Shi Rushen 1896
Tone is not predominant; Tone is not primordial (ID: 303) - Maddieson Ian 1901
Cue weighting in Mandarin tone perception: A comparison between native speakers and learners of Mandarin (ID: 351) - Jiang Song 1906
Discrimination of Cantonese lexical tones by listeners with and without learning experience of tone languages (ID: 358) - Wong Yee-ping 1911
Grammatical tone in Kanise Khumi: exclusive/inclusive pronominal agreement (ID: 421) - Ikeda Elissa, Gehrmann Ryan 1916
Contact-induced tonogenesis in Hong Kong English (ID: 440) - Yeung Ping Hei 1921
Phonetic realizations of neutral tones in wuhan chinese disyllabic words (ID: 488) - ZHOU Duanchi 1926
Tone change in Mandarin: Comparing acoustic analyses from the early 20th century to today (ID: 500) - Yang Cathryn 1930
Intermingling tone systems: the relationship of Nanning Mandarin to Nanning Cantonese and Standard Mandarin (ID: 516) - Hou Xinzi, Zhao Liang, Chodroff Eleanor 1935
Disyllabic tones in Mandarin preschool children and child-directed speech (ID: 524) - Wang Zhiwei, Li Aijun, Gao Jun 1940
An acoustic study of tone sandhi in Lishui Wu (ID: 542) - LAN WUPENG, CHEN XIAOCONG, ZHANG CAICAI 1945
Syntactic and prosodic effects on reconstruction of tone sandhi domains: Evidence from Xiamen Dialect (ID: 643) - Hu Yiying, Feng Hui, Li Aijun 1950
Non-native tone perception – When music outweighs language experience (ID: 694) - Götz Antonia, Liu Liquan 1955
Dialectal differences in the realisation of Tone 1 in Taiwan Mandarin (ID: 720) - Fon Janice, Chuang Yu-Ying 1960
Disyllabic tone sandhi patterns in Shaoxing Wu Chinese (ID: 785) - Wen Xinyi, Chen Yiya, Cheng Lisa 1965
Production of Cantonese tones by Mandarin-speaking immigrants: Acoustic and perceptual measurements (ID: 800) - Yang Yike, Han Dong, Wong Sze Man, Chan Chak Sum, Leung Chun Yeung, Chen Xiaocong 1970
Lexical tones are timed to articulatory gestures (ID: 840) - Burroni Francesco 1975
A preliminary exploration of tones in paite (ID: 972) - Hangshing Rosie Nempalching, Mahanta Shakuntala 1980
Phonetic implementation of phonologically different high tone spans in Luganda (ID: 1083) - Hughes Cerys, Lee Seung Suk, Farinella Alessa, Yu Kristine 1985

7. Syllable

Duration and Sonority in Blackfoot Syllable Structure (ID: 137) - Derrick Donald 1990
Voiceless stops in syllable coda in Tuscan Italian (ID: 388) - Marotta Giovanna, Cossu Piero, Avano Giuditta 1995
Intrusive vocoids and syllable structure in Georgian (ID: 613) - Crouch Caroline, Chitoran Ioana, Goldstein Louis, Katsika Argyro 2000
Temporal organization of word-initial consonant sequences in Najdi Arabic (ID: 616) - Alkhonini Omar, Kwon Harim 2005
A simple acoustic measure of onset complexity (ID: 838) - Durvasula Karthik 2010
Syllabic reduction in Italian connected speech: towards the integration of linguistic and computational approaches (ID: 959) - Schettino Loredana, Vitale Vincenzo Norman, Cutugno Francesco 2015
Emergent syllable types in Brazilian Portuguese (ID: 1050) - Cristófaro Silva Thaïs, Cantoni Maria 2020

7. Special Session - The sounds of ageing

Tracking creak from early to late adulthood: A panel study from the North East of England (ID: 167) - Grama James, Eiswirth Mirjam,

Buchstaller Isabelle 2025
Longitudinal investigation of the aging voice (ID: 206) - Tucker Benjamin, Hedges Stephanie, Berardi Mark, Hunter Eric 2030
Variability in space, compensation in time: Effects of aging on tongue body kinematics in prosodic focus marking (ID: 380) - Mücke Doris, Thies Tabea, Buech Phlipp, Furlani Noemi, Schreen Janine, Hermes Anne 2035
Change across the lifespan in GOAT: Evidence from a panel study of Tyneside English (ID: 530) - Bauernfeind Lea, Ahrens Carina, Grama James 2040
Age-related vowel variation in French (ID: 705) - Hermes Anne, Audibert Nicolas, Bourbon Angélina 2045
Phonation and Aging in White Hmong (ID: 796) - Khan Sameer, Esposito Christina, Schäfer Kevin 2050
Phonetic and pragmatic features of the French discourse marker 'et puis': a longitudinal study in 10 female speakers (ID: 1068) - Gerstenberg Annette, Fuchs Susanne, Koenig Laura 2055

8. Laboratory Phonology

SPRAAKLAB: a mobile laboratory for collecting speech production data (ID: 28) - Wieling Martijn, Rebernik Teja, Jacobi Jidde 2060
/R/ lenition in Quebec French : Evidence from the distribution of 9 allophones in large corpora (ID: 39) - Lancien Mélanie, Hutin Mathilde, Adda-Decker Martine, Vasilescu Ioana, Stuart-Smith Jane 2065
Production planning Hypothesis: evidence from palatalization in American English (ID: 105) - Kul Malgorzata, Kaźmierski Kamil 2070
Laxing harmony in Laurentian French: Coarticulation or phonological process? (ID: 149) - Burness Phillip, Tahtadjian Talia, MacLeod Beth, Ahn Suzy 2075
Auditory and written perception of Japanese sequential voicing (ID: 153) - Masuda Hinako, Hirayama Manami, Vance Timothy J. 2080
The window of opportunity: Anticipatory nasal coarticulation in three languages (ID: 159) - Pouplier Marianne, Rodriquez Francesco, Alderton Roy, Lo Justin J. H., Reinisch Eva, Evans Bronwen G., Carignan Christopher 2085
Modelling the parameters of fixed stress in Slavic with dynamic computational networks (ID: 248) - Hudson Toby 2090
Dynamics of the tongue contour in the production of guttural consonants in Levantine Arabic (ID: 273) - Al-Tamimi Jalal, Palo Pertti 2095
The effect of phonological factors on variable word onset vowel deletion in najdi arabic (ID: 291) - Alsabhan Rana, Setter Jane 2100
Language specificity vs speaker variability of anticipatory labial coarticulation in German and English (ID: 314) - Lo Justin J. H., Carignan Christopher, Pouplier Marianne, Alderton Roy, Rodriquez Francesco, Evans Bronwen G., Reinisch Eva 2105
Investigating overlapping gestures in acoustic signals: the case of fricatives and sonorants in clusters (ID: 325) - Genette Jérémy 2110
Variable rhoticity in Glasgow English: Sociolinguistic factors and abstractions vs. exemplars (ID: 336) - Kaźmierski Kamil 2115
An Electropalatographic study of the Singleton-Geminate contrast in Arabic (ID: 424) - Issa Amel 2120
Intergestural CV timing of homophonous words with different morphological structures:A case study of liquid/l/ in Korean (ID: 437) - Lee Jiyoung, Kim Sahyang, Cho Taehong 2125
Perception of Tashlhiyt consonant quantity contrasts by native vs. nonnative listeners from three languages (ID: 465) - Hallé Pierre, Ridouane Rachid, Buech Philipp, Gao Jiayin, Chang Yueh-Chin, Hsieh Feng-Fan 2130
Intra- and intersegmental durational compensation of Korean plosives (ID: 467) - Choi Yeongeun 2135
Articulatory Correlates of Stop Phonation in L1 Korean and L2 English (ID: 539) - Yun Gwanhi, Sung Jae-Hyun 2140
/s/-lenition and resyllabification in Southern Cone Spanishes (ID: 562) - Marchini Gilly, Ramsammy Michael 2144
Interacting sound symbolic patterns: The case of vowel sound symbolism in Korean given names (ID: 619) - Sullivan Lisa, Kang Yoonjung 2149
Preboundary lengthening and its kinematic characteristics in Mandarin Chinese in interaction with focus and lexical tone (ID: 672) - Li Hongmei, Kim Sahyang, Cho Taehong 2154
Obstruent voicing and laryngeal feature in Arabic (ID: 700) - Dallak Abdulrahman, Khattab Ghada, Al-Tamimi Jalal 2159
Pre-aspiration in Southern Swedish and its relationship to prosody (ID: 710) - Tronnier Mechtild 2164
How measures of gestural overlap relate to dynamics: evidence from German & English word-initial stop-lateral clusters (ID: 721) - Du Shihao, Kuberski Stephan, Gafos Adamantios 2169
Glottalization of voiceless stops in Multicultural London English (ID: 730) - Chong Adam, Garellek Marc 2174
Formant characteristics of unstressed unreduced vowels in American English: only explained by duration? (ID: 735) - Sugahara Mariko 2179
Disappearing supralaryngeal articulatory distinction of the three-way laryngeal contrast of Korean velar stops (ID: 752) - Song Jiyeon, Kim Sahyang, Cho Taehong 2184
O Lo/r/d, open thou ou/r/ lips: rhoticity in choral singing from Glasgow and Cambridge (ID: 810) - Marshall Edward 2189
Stress conditioned phonological process: a case study of italian palatalization (ID: 814) - Shao Bowei, Buech Philipp, Hermes Anne, Giavazzi Maria 2194
Compensatory breathiness in Canary Islands Spanish – an acoustic study of vowels in /s/ weakening contexts (ID: 836) - Broś Karolina 2199
Consonant-vowel coarticulation patterns in Swedish and Mandarin (ID: 892) - Gao Man, Svensson Lundmark Malin 2204
A simulated analysis of repeated measures in VOT: we need more tokens! (ID: 900) - Parrish Kyle 2209
The timing of secondary dorsal articulations across syllable positions in irish (ID: 950) - Padgett Jaye, Bennett Ryan, McGuire Grant, Ní Chiosáin Máire, Bellik Jenny 2214
Coarticulation and contrast in a vowel harmony system: coarticulatory propensity in Khalkha Mongolian V-C-V sequences (ID: 1043) - Mitra

Auromita, Krishnaswamy Meghavarshini, Dutta Indranil 2219

8. Special Session - The history of phonetic sciences

Early activities in the field of x-ray imaging in experimental phonetics and speech training (ID: 44) - Hoffmann Ruediger 2224
Bent vs. straight models of the human vocal tract based on measurements by Chiba and Kajiyama (ID: 288) - Arai Takayuki 2229
The first X-ray sound film (1935): its origins, impact and restoration (ID: 473) - Ashby Michael, Angelika Braun 2234
Restoration of Portable Kymograph (ID: 904) - Amelot Angelique, Ghio Alain, Demolin Didier 2239

9. Phonology-Phonetics Interface

Acoustic correlates and contrast maintenance in Burmese voiceless sonorants (ID: 104) - Repetti-Ludlow Chiara 2244
Analyzing variability in closure voicing and co-intrinsic F0 in Central Standard Swedish (ID: 356) - Kirby James, Tan Maryann 2249
Morpho-phonological effects on the phonetic characteristics of tense consonants in Korean compounds (ID: 359) - Kwon Chloe Dokyung, Tilsen Sam, Whitman John 2254
Positional asymmetry in length and voicing cues in the alveolar plosives of Ikema Miyako Ryukyuan (ID: 386) - Shinohara Shigeko, Amelot Angelique, Hussain Qandeel 2259
Fortis stops in Polish - evidence from acoustic measures of voice quality (ID: 434) - Schwartz Geoffrey, Asiaee Maral, Wojtkowiak Ewelina 2264
PlanArt: A modular platform for computational modeling of articulatory planning (ID: 439) - Turk Alice, Elie Benjamin, Šimko Juraj 2269
Acoustic properties of palatalized consonants in Japanese (ID: 548) - Tanaka Yu 2274
Aspiration length and intensity as potential acoustic cues to [voice] in final position: evidence from Shughni (ID: 582) - Makarov Yury 2279
Geminate types and their acoustic effects on adjacent vowels in Hungarian (ID: 777) - Neuberger Tilda 2284
Revisiting CV timing with a new technique to identify inter-gestural proportional timing (ID: 853) - Durvasula Karthik, Wang Yichen 2289
Measuring variation in Central Pame vowels (ID: 1101) - Herce Borja, Torres Catalina 2294

9. Special Session - Acoustic manifestations of speech register

Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production (ID: 45) - Aoki Nicholas, Zellou Georgia 2299
Acoustic measures of non-native addressee register for mid to high proficient English learners of German (ID: 126) - Sell Bianca, Terada Megumi, Belz Malte, Mooshammer Christine 2304
A study on caregivers speech in retirement homes (ID: 296) - Rouas Jean-Luc, Wu Yaru, Shochi Takaaki 2309
Acoustic characteristics of dog-directed speech: The role of individual empathy level and personality traits (ID: 584) - Nuscheler Jasmin Devi, Elsässer Nathalie, Braun Angelika, Zahner-Ritter Katharina 2314
Assessing register variation in local speech rate (ID: 608) - Duran Daniel, Weirich Melanie, Jannedy Stefanie 2319

10. Phonetics of First Language Acquisition

How do children mark gender phonetically? An analysis of voice quality (ID: 207) - Wong Eugene, Munson Benjamin 2324
Mandarin tone production can be learned under perceptual guidance — A machine learning simulation (ID: 282) - Meng Hanyi, Chen Chien-Tzu, Chen Yue, Liu Zirui, Xu Yi 2329
Vowel duration before a geminate in Italian children and adults: is lexicality important? (ID: 333) - Colombo Lucia, Arciuli Joanne 2334
Infants' learning of novel segments is modulated by prosody (ID: 343) - Chládková Kateřina, Podlipský Václav Jonáš, Nudga Natalia, Paillereau Nikola, Kynčlová Kateřina, Šimáčková Šárka 2339
Vowel length in infant-directed speech: the realisation of short-long contrasts in Czech IDS (ID: 350) - Svoboda Michaela, Chládková Kateřina, Kocjančič Antolík Tanja, Paillereau Nikola, Slížková Petra 2344
Influence of bilingualism or caregiver input? Variation in VOT in simultaneous bilingual preschoolers in Singapore (ID: 361) - Sim Jasper 2349
The effects of vision on speech production in children (ID: 410) - Menard Lucie, Turgeon Christine 2354
The Origins of Phonology and Lexicon in Infancy: Phonological Abstraction Before Perceptual Attunement? (ID: 674) - Burnham Denis, Best Catherine, Götz Antonia, Kalashnikova Marina, Johnson Elizabeth Johnson, Altuntas Eylem, Cutler Anne 2359
Input distribution of Korean affricates in the context of a diachronic sound change: comparison between IDS and ADS (ID: 831) - Nam Minji, Choi Youngon, Hwang Hyun Kyung, Mazuka Reiko 2364
A longitudinal study of pauses, interpausal units and clauses in infant-directed speech (ID: 911) - Kohari Anna, Mády Katalin 2369
Evidence for preserved phonetic sensitivity to lexical tones in Chinese international adoptees (ID: 1077) - Chang Andrea, Perrachione Tyler, Lim Sung-Joo 2374

10. SpecialSession - Prosody Visualization Challenge

Using spectral coherence between temporal envelope and mouth movements to reveal rhythmic regularities in speech (ID: 1118) - He Lei 2379
Disentangling intonation in tone languages (ID: 1119) - Li Katrina Kechun, Nolan Francis, Post Brechtje 2381

11. Phonetics of Second and Foreign Language Acquisition

Relation between musical aptitude and L2 stress perception in French- and Korean-speaking listeners (ID: 25) - Martinez Garcia Maria Teresa, Schwab Sandra 2383
Perception of word stress amongst French learners of English: tone & suffix (ID: 43) - Kamiyama Takeki, Amand Maelle 2388
Production of Mandarin Disyllabic Tones by Vietnamese Speakers (ID: 48) - Tu Jung-yueh 2393
Perception of Japanese consonant length by advanced learners from Mandarin- and Mongolian-speaking backgrounds (ID: 52) - Tsukada Kimiko, NA Yurong 2398
The contribution of temporal cues to perceived nativeness in the native speech of English migrants to Austria (ID: 53) - Mennen Ineke, Reubold Ulrich, Mayr Robert 2403
Effects of L1 phonotactic constraints on learning novel words (ID: 65) - Katayama Tamami 2408
Predicting nonnative consonant identification: English consonants as perceived by native speakers of Kalaallisut (ID: 68) - Bohn Ocke-Schwen, Korneliussen Kulunnguaq 2413
Oral recast and task-based interaction on the prosody of L2 Chinese learners (ID: 69) - Chen Chun-Mei 2418
Filler particles in English and Spanish L1 and L2 speech (ID: 110) - Muhlack Beeke 2423
L2 acquisition of Polish sibilant fricatives by L1 Czech speakers (ID: 143) - Howson Phil, Madathodiyil Irfana 2428
The impact of cognitive load on speech production in German-English bilinguals (ID: 150) - Göbel Jessica, McCarthy Kathleen 2432
The acoustic realization of L2 Spanish phonetic categories and allophonic alternations by English-speaking immigrants (ID: 154) - Amengual Mark 2437
Exploring the problem in native danish listeners' perception of the english hot and hut vowels (ID: 194) - Villumsen Jonas, Bohn Ocke-Schwen, Horslund Camilla Søballe 2442
Dynamic tongue movements in L1 Japanese and L2 English liquids (ID: 198) - Nagamine Takayuki 2447
Is the effect of L2 vocabulary size on lexical encoding modulated by L1-L2 similarity? (ID: 199) - Llompart Miquel, Rocca Brian, Darcy Isabelle 2452
Revising parameters for predicting L2 speech fluency and proficiency (ID: 221) - Kallio Heini, Kuronen Mikko 2457
L2-Mandarin regional accent variability facilitates Mandarin-naive English listeners' learning of Mandarin tone-words (ID: 244) - LI Yanping, Tyler Michael, Burnham Denis, Best Catherine 2462
How Japanese L2 listeners perceive English silent-centre front vowels (ID: 245) - Tokuma Shinichi 2467
Nuclear contours of Spanish echo questions produced by proficient Chinese learners of Spanish: A dynamic analysis (ID: 265) - Li Peng, Xi Xiaotong 2472
Exploring the Phonetic Context Effects on the Production of Portuguese /r/ by L1-Mandarin Learners (ID: 275) - Cao Yidan, Zhou Chao, Hristovsky Gueorgui 2477
Perception of vowel and consonant quantity contrasts by Cantonese, English, French, and Japanese speakers (ID: 290) - Lee Albert, Shinohara Yasuaki, Chiu Faith, Mut Tsz Ching 2482
Assessing intelligibility of L2 speech by native and non-native listeners: a longitudinal study (ID: 302) - Pesantez Pesantez Alejandra Carolina, Lima Jr Ronaldo 2487
Effects of dialect exposure on the processing of Argentine regional phones by English-speaking L2 learners of Spanish (ID: 305) - Schmidt Lauren 2492
L1 Effects on Naive Perception and Production of [\textbaru] by Mandarin and Italian Speakers (ID: 306) - Chen Xuanda, Clayards Meghan, Goad Heather, Kim Donghyun 2497
Neural predisposition for L2 stress learning in French and German listeners (ID: 330) - Schwab Sandra, Jost Lea, Mouthon Michael, Etter Faustine, Kamber Julie, Rogenmoser Lars, Annoni Jean-Marie 2502
Word-final devoicing in Brazilian Portuguese and L2 English (ID: 338) - Mendes Junior Wellington, Cristófaro Silva Thaïs 2507
Characterization of Mandarin accent of French across three different speaking styles (ID: 341) - FU Wenxun, Adda-Decker Martine, Kühnert Barbara 2512
The link between perception and production of English vowel contrasts in non-native speakers (ID: 353) - Chang Chih-Chao, Chien Chun-Ting, Lu Yu-An 2517
A preliminary study of Mandarin neutral tone production by Japanese and Korean L2 learners (ID: 354) - Shu Tong, Mok Peggy Pik Ki 2522
The classification of native and non-native productions of Italian consonant length contrast (ID: 367) - Feng Qiang, Busà M. Grazia 2527
Acquisition of the low vowel in L2 Spanish by Chinese speakers (ID: 372) - Font Rotchés Dolors, Cao Yongfa, Rius-Escudé Agnès 2532
Phonetic transcription of atypical speech: second or a foreign language (ID: 390) - Carović Ines 2537
Investigating predictors of foreign accentedness in L3 acquisition (ID: 403) - Wrembel Magdalena, Kaźmierski Kamil, Dziubalska-Kołaczyk Katarzyna, Weckwerth Jarosław, Cal Zuzanna 2542
Perception of English lexical stress in different intonations by Mandarin listeners (ID: 413) - Yang Tzu-Hsuan, Tremblay Annie 2547

Orthography and the mental lexicon: the effects of English silent letters on French learners (ID: 435) - Mouquet Marine, Mairano Paolo 2552
Influences of multi-language experience on Mandarin consonant perception (ID: 460) - Chen Shuwen, Wong Janice W. S. 2557
Distributional learning and overnight consolidation of a difficult non-native tone contrast (ID: 499) - CHUI Yin To, QIN Zhen 2562
Static and dynamic features of English monophthongal vowels production by Hijazi Arabic L2 learners (ID: 501) - Almurashi Wael, Al-Tamimi Jalal, Khattab Ghada 2567
The role of intonation type and stress position in native Mandarin speakers' perception of English lexical stress (ID: 510) - Zhou Weijing, Nolan Francis, Jing jiawei, Gong Jian 2572
The effect of learning context on Mandarin listeners‘ perception of English vowels (ID: 528) - Lu Jiayi, Ota Mitsuhiko 2577
Exploring second language attitudinal prosody with a machine learning approach (ID: 566) - Xu Xiaolin, Grenon Izabelle 2581
Nativeness and intelligibility of Japanese accented English consonants by French listeners (ID: 572) - PEREZ-RAMON RUBEN, KONDO MARIKO, DETEY SYLVAIN, FONTAN LIONEL, AMAND MAELLE, KAMIYAMA TAKEKI 2586
French /ø-u/ contrast in Japanese learners with/without gesture feedback in monosyllables: a descriptive acoustic study (ID: 586) - Pillot-Loiseau Claire, Kamiyama Takeki, Kitamura Ayako 2591
Perception of Spanish Declarative Questions and Statements by L2 Spanish Speakers (ID: 606) - Bedialauneta Txurruka Izaro 2596
Schwa in French obstruent-liquid clusters by native speakers and Mandarin learners (ID: 609) - Yang Minmin, Ridouane Rachid 2601
Revisiting intelligibility: The role of listeners' L1 and proficiency in perceiving Chinese EFL learners' speech (ID: 630) - Cao Yating, Chen Hua . 2606
The perception of Japanese lexical contrasts by L2 learners: The role of the lexicon (ID: 638) - Fonseca Marco 2611
Factors determining perceptual and acoustic similarity between native and non-native vowels (ID: 644) - Balas Anna, Wrembel Magdalena, Kaźmierski Kamil, Weckwerth Jarosław 2617
Phonetic Realization of Narrow Focus by Chinese EFL Learners (ID: 652) - Zhou Xiaoyu, Sun Binbin, Feng Hui 2622
Tongue contours for the Japanese moraic nasal by speakers of Standard Chinese (ID: 683) - Mizoguchi Ai, Morimoto Maho, Li Weiyu, Arai Takayuki 2627
Suprasegmental cues to perception of English lexical stress by Mandarin speakers (ID: 706) - He Mosi, Zhang Ting, Li Bin 2632
Regional Dialect Influence on L2 English Production: BATH – STRUT – LOT triangle for L1 Turkish speakers (ID: 716) - Aksu Bahar 2637
Geminate rearticulation is psychoacoustically robust: Imitation of Polish geminates by Spanish-Catalan bilinguals (ID: 729) - Rojczyk Arkadiusz, Rallo Fabra Lucrecia 2641
The effect of L2 experience on Mandarin speakers’ production of Australian English short and long vowels (ID: 747) - Liu Shuting, Maxwell Olga, Fletcher Janet 2646
Does visuospatial working memory predict L2 perceptual learning from phonetic training with hand gestures? (ID: 749) - Xi Xiaotong, Li Peng, Prieto Pilar 2651
Effects of language training on mandarin lexical tone perception by Japanese speakers (ID: 763) - Qi Wu 2656
Perception and Production of English Tense and Lax Vowels by Chinese EFL Learners (ID: 764) - Wang Lijuan, Guo Jiwei, Feng Hui 2661
Intelligibility and related features of English sounds produced by young Japanese learners (ID: 768) - Joto Akiyo, Nishio Yuri 2666
L3 Norwegian /ʉː/ and /uː/ in L1 Polish/L2 English learners: different patterns of cross-linguistic interactions (ID: 775) - Weckwerth Jarosław, Wrembel Magdalena, Kaźmierski Kamil, Balas Anna 2671
f0 enhancement in Japanese voicing contrast by Japanese native speakers and its implication to L2 perception/production (ID: 792) - Iwamoto Keiji 2676
Phonetic realization of multiple stress levels by speakers of a non-stress language: A case of Japanese-accented English (ID: 815) - Tajima Keiichi, Kitahara Mafuyu, Yoneyama Kiyoko 2681
Accommodation to passive exposure in the L2 (ID: 817) - Weber Julika, Geissler Christopher 2686
Learning to distinguish the three-way laryngeal contrast of Korean plosives by native English speakers (ID: 818) - Perrachione Tyler, Gabrieli John, Finn Amy 2691
Articulatory timing of the Japanese singleton and geminate /t/ produced by speakers of Standard Chinese (ID: 820) - Morimoto Maho, Mizoguchi Ai, Li Weiyu, Arai Takayuki 2696
Comparison of L2 Korean pronunciation error patterns from five L1 backgrounds by using automatic phonetic transcription (ID: 879) - Yeo Eun Jung, Ryu Hyungshin, Lee Jooyoung, Kim Sunhee, Chung Minhwa 2701
Learning phonotactically complex L3 words: Are bilinguals more successful? (ID: 884) - Hanulikova Adriana 2706
Perceptual representation and production of American English simplex vowels by adolescent Palestinian-Arabic learners (ID: 891) - van Heuven Vincent J., Farran Bashar M. 2711
Different realizations of German underlying vs. derived diphthongs between two generations of Cantonese L1 speakers (ID: 914) - Vuong Michelle, Einfeldt Marieke, Braun Bettina 2716
Effects of stress and prominence on Spanish stops and lenition in L2 speech of proficient Mandarin learners of Spanish (ID: 918) - Xi Xiaotong, Li Peng 2721
Automatic feedback on pronunciation and Anophone – a tool for L2 Czech annotation (ID: 922) - Holaj Richard, Pořízka Petr 2726
Exploring the relationship between individual differences in cross-language perceived similarity and L2 vowel perception (ID: 941) - Cebrian Juli, Gorba Celia 2731
Second language learners' acquisition of Mandarin fricative contrasts: an ultrasound study (ID: 965) - Oakley Madeleine 2736

Identification of non-native English speakers' L1s via patterns of prosodic feature deviance from native speaker norms (ID: 1010) - Talley Jim . 2741
Is rhoticity on the tip of your tongue? Investigating tongue shapes for English /r/ in French learners with ultrasound. (ID: 1011) - Léger Alice, King Hannah, Ferragne Emmanuel 2746
L1 phonotactic constraints mediate phonological awareness of non-native gemination (ID: 1032) - Mairano Paolo, Santiago Fabian 2751
Influence of primary and secondary accent on the duration of French vowels in production of Czech and French speakers (ID: 1046) - Suková Vychopňová Kateřina 2756
Individual differences in VOT realizations of English plosives by Palestinian pupils (ID: 1057) - Tamim Nour, Hamann Silke, Boersma Paul 2761
The effect of age of entry and contact on backward transfer of Glaswegian English (L2) on Hindi and Indian English (L1) (ID: 1063) - Shaktawat Divyanshi 2766
French prosodic stress produced by Twi speakers compared to French speakers: an acoustic investigation (ID: 1105) - Adu Manyah Kofi, Owusua Amponsah Rejoice 2771

12. Bilingual/Multilingual Phonetics

Native-Language Phonotactic Processing in Bilinguals (ID: 12) - Freeman Max, Marian Viorica 2776
Mid vowel and nasal vowel production in young adult French speakers on Mayotte Island in the Indian Ocean (ID: 72) - Mori Miki 2781
The relationship between L2 acquisition and L1 attrition in the phonetic domain (ID: 83) - Kornder Lisa, Reubold Johann Ulrich Balthasar, Mayr Robert, Mennen Ineke 2785
Prosodic cue weighting in the perception of lexical stress by French-English bilinguals: Preliminary behavioural data (ID: 122) - Gilbert Annie, Honda Claire T., Baum Shari R. 2790
Acoustic analyses of Mandarin tones produced by bilingual elementary students in Canada (ID: 123) - Lin Youran, Pollock Karen, Tucker Benjamin, Li Fangfang 2795
Hesitating with and without Language Heritage - Prosodic Aspects of Filler Particles in the RUEG Corpus (ID: 169) - Böttcher Marlene, Zellers Margaret 2800
Switching between phonological biases is not free: evidence from a bilingual reconstruction task (ID: 183) - Bramlett Adam A., Wiener Seth 2805
Assessing the effects of auditory feedback perturbation on perception of similar vowels in a different language (ID: 192) - Rees Madeleine, Davis Matt, Bradshaw Abbie, Post Brechtje 2810
L2-induced phonetic drift in L1 Polish vowel productions (ID: 205) - Wojtkowiak Ewelina 2815
Effects of cognitive load on vowel production in bilingual speakers of High and Low German (ID: 216) - Frank Marina, Rohloff Marina, Peters Jörg 2820
Vowel reduction in spontaneous Bulgarian Judeo-Spanish (ID: 220) - Grünke Jonas, Sabev Mitko, Andreeva Bistra, Gabriel Christoph 2825
Crosslinguistic influence and language-specific strategies in bilingual clear speech (ID: 362) - Jung Ye-Jee, Dmitrieva Olga 2830
Cross-linguistic influence in the mid vowels of Hausa/Ekhwa speakers (ID: 398) - Bawa Emmanuel, Burdin Rachel, Cassiani Taylor 2835
Foreign-accented phonetic detail in L1 word processing: L1 Spanish L2 English listeners and English-accented Spanish (ID: 420) - Forst Hadley, Simonet Miquel 2840
L2 proficiency predicts L1 accentedness and comprehensibility (ID: 430) - Yazawa Kakeru, Pérez Ramón Rubén, Kondo Mariko 2845
Demystifying heritage accent: Heritage accent perceived by heritage speakers (ID: 481) - Kim Joo Kyeong, Kim Ji Young, Carter John, Choi Jenny 2850
The production of Mandarin /r/ by early and late Japanese-Mandarin bilinguals: articulatory and acoustic findings (ID: 482) - Zhu Zhiqiang, Mok Peggy 2855
Perception-production mismatch in L1 drift (ID: 503) - Sypiańska Jolanta, Cal Zuzanna 2860
Does gemination resist linguistic attrition? A study on Italian migrant speech in Melbourne Australia (ID: 565) - De Iacovo Valentina, Mairano Paolo, Hajek John 2865
Do Chinese-English Bilinguals Speak English Words with Lexical Tone in Mind? (ID: 653) - Yang Qing, Chen Yiya 2870
Acquiring new contrasts in L2 vs. L3: Chinese dialectal speakers' acquisition of Mandarin and English phonetic contrasts (ID: 670) - Ju Tianqi, Gu Beiwei, Huang Fumo, Liu Xinyi, Yue Cen, Yao Yao 2875
Crosslinguistic influence on perceptual cue-weighting in Korean-English bilinguals (ID: 676) - Seo Yuhyeon, Dmitrieva Olga 2880
Individual variability in L1 category compactness on L2 production compactness and accuracy (ID: 702) - Zhai Weiyi, Clayards Meghan, Goad Heather 2885
A preliminary study on Cantonese vowel development by young heritage children (ID: 858) - Lan Chen, Mok Peggy 2890
The perception of lexical stress in Indian Englishes (ID: 871) - Payne Elinor, Maxwell Olga, Fuchs Robert, Wang Yizhou 2895
Phonetic integration in Heritage Korean language mixing (ID: 933) - Butler Nicoline 2900
The role of psychotypology in the acquisition of rhotic consonants by adolescent emergent multilinguals (ID: 966) - Krzysik Iga 2905
Revisiting the Vowels of Hong Kong English – the Post-Handover Generation (ID: 997) - Chan Hoi Ling, Fuchs Robert 2910
Lexically-guided phonetic recalibration transfers across languages in French-English bilinguals (ID: 1034) - Caudrelier Tiphaine, Martin Clara D., Samuel Arthur G., Beausoleil Marie-Michèle, Tiede Mark K., Ménard Lucie 2916
Examining the effects of stress on vowel production in heritage vs. monolingual Spanish school-aged child speakers (ID: 1104) - Repiso Puigdelliura Gemma, Pape Daniel 2921

13. Phonetic Universals and Typology

Diphthong acoustics and articulatory uniformity (ID: 203) - Hu Fang 2926
Vowels and Diphthongs in Chengdu Mandarin (ID: 204) - Hu Fang, Li Nan 2931
Patterns of the syllable-final consonants in Chinese dialects (ID: 371) - Lee Wai-Sum, Zee Eric 2936
Spatial sound symbolism: a cross-linguistic study (ID: 596) - Wikström Alexandra, Vainio Lari, Vainio Martti 2941
Reconsidering favored phonotactic patterns for ejective and implosive consonants (ID: 637) - Easterday Shelece 2946
Effect of vowel context on stop place identification in Yoruba (ID: 645) - Burns Roslyn, Shaw Jason 2951
Bora /ɨ/ is produced with a low tongue dorsum and dental contact: an ultrasound study (ID: 897) - Mielke Jeff, Oakley Madeleine, Vásquez-Aguilar Alonso, Zariquiey Roberto, Parker Steve 2956

14. Phonetics of Sound Change

Ghost segments in the Flemish Tussentaal (ID: 11) - Bozhinoski Mishko, Boersma Paul 2961
Interaction of quantity, foot structure, and stress in the 2nd/3rd syllable sonorants of Soikkola Ingrian trisyllables (ID: 31) - Kuznetsova Natalia, Brodskaya Irina, Markus Elena 2966
Predicting sound change: acoustic evidence from a longitudinal study of /s/-retraction in English /str/ (ID: 91) - Stevens Mary 2971
Individuals' perceptual retuning predicts articulatory accommodation: implications for sound change (ID: 114) - Beddor Patrice, Coetzee Andries, Calloway Ian, Tobin Stephen, Purse Ruaridh 2976
Pre-/l/ vowel change in Australian English pool-pull (ID: 124) - Szalay Tuende, Benders Titia, Cox Felicity, Proctor Michael 2981
Diphthongisation of /iː/ in West Frisian (ID: 164) - Kingma Martijn, Boersma Paul, Van de Velde Hans, Versloot Arjen 2986
Change over time in [ɛ] and [æ] in Hawai'i Creole Creole (ID: 170) - Grama James 2991
From pasta to pHizza: variation and change in VOT in Swiss German (ID: 171) - Steiner Carina, Jeszenszky Péter, Leemann Adrian 2996
Stop voicing in Drehu: Effects of place of articulation, speaker sex, and language attitudes (ID: 172) - Torres Catalina, Steffman Jeremy 3001
Simultaneous vowel changes in British Columbian English (ID: 181) - Cardoso Amanda, Liu Suyuan, Pritchard Robert, Babel Molly 3006
Sociolinguistic factors predicting sound change in Swiss German (ID: 191) - Leemann Adrian, Steiner Carina, Siebenhaar Beat, Jeszenszky Péter 3011
Phonotactically driven cue weighting in a sound change in progress: acoustic evidence from West Central Bavarian (ID: 231) - Thon Katharina, Kleber Felicitas 3016
Possible back prelateral mergers in Oklahoma (ID: 355) - Freeman Valerie, Landers Molly 3021
Investigating the Irish English PEN-PIN merger: patterns and origins (ID: 383) - Tallon Kate 3026
Age- and gender-based variation in the perception of voicing contrast in Tokyo Japanese (ID: 509) - Kang Yoonjung, Hirayama Manami 3031
Increasing aspiration of Swiss German plosives: a sound change in progress? (ID: 515) - Zebe Franka, Watter Camille, Schmid Stephan 3036
Voice quality is not an obligatory stage in tonogenesis: A case study of Eastern Khmu (ID: 532) - Maspong Sireemas 3041
The effects and interaction of age and region on the acoustic realization of the Korean stop contrast (ID: 544) - Holliday Jeffrey, Kong Eun Jong, Lee Hyunjung 3046
Investigating sound change through computational agent-based modelling: An R package (ID: 583) - Cronenberg Johanna, Gubian Michele, Harrington Jonathan, Reichel Vanessa 3051
Cross-dialectal comparison of /st/ clusters in Spanish: implications for sound change (ID: 597) - Galvano Amber, Henriksen Nicholas 3056
Acoustic analysis of vowel nasalization in Taizhou Chinese (ID: 662) - Han Qianwen, Lee Wai-Sum 3061
An acoustic study of the realisation of KIT in the conversational speech of young English speakers in Australia (ID: 684) - Docherty Gerard, Gonzalez Simón, Foulkes Paul 3066
The loss of intervocalic short /h/ in Estonian (ID: 825) - Teras Pire 3071
Phonetic variation of word-final affricates in Romanian connected speech (ID: 873) - Niculescu Oana, Avram Andrei A. 3076

15. Speech Evolution

Phonetic approaches to analyses of great ape quasi-vowels (ID: 22) - Ekström Axel G., Moran Steven, Sundberg Johan, Lameira Adriano R. . 3081

16. Speech Technology

Prosody under control: Controlling prosody in text-to-speech synthesis by adjustments in latent reference space (ID: 60) - Šimko Juraj, Törö Tuukka, Vainio Martti, Suni Antti 3086
Who's speaking? Predicting speaker profession from speech (ID: 223) - Wu Yaru, Chen Lihu, Elie Benjamin, Suchanek Fabian, Vasilescu Ioana, Lamel Lori 3091
ASRLux: Automatic speech recognition for the low-resource language Luxembourgish (ID: 266) - Gilles Peter, Hillah Léopold,

Hosseini-Kivanani Nina 3096
Using Tacotron 2 to develop synthetic child speech in South African English (ID: 293) - Terblanche Camryn, Tucker Benjamin V, Schnoor Tyler, Harty Michal 3101
Using word-level features for prosodic prominence detection in conversational speech (ID: 298) - Linke Julian, Kubin Gernot, Schuppler Barbara 3106
Discerning dimensions of quality for state of the art synthetic speech (ID: 438) - Seebauer Fritz, Kuhlmann Michael, Haeb-Umbach Reinhold, Wagner Petra 3111
A biological investigation of performative speech through synthesis (ID: 466) - Lau Emily, Post Brechtje, Knill Kate 3116
Automatic sound event detection and classification of great ape calls using neural networks (ID: 508) - Jiang Zifan, Soldati Adrian, Schamberg Isaac, Lameira Adriano, Moran Steven 3121
Prak: An automatic phonetic alignment tool for Czech (ID: 525) - Hanžl Václav, Hanžlová Adléta 3126
Speaker interpolation based data augmentation for automatic speech recognition (ID: 585) - Kerle Lisa Kristina, Pucher Michael, Schuppler Barbara 3131
Automatic speech recognition: system variability within a sociolinguistically homogenous group of speakers (ID: 593) - Harrington Lauren, Hughes Vincent 3136
SLAM 3: An Updated Stylization Model for Speech Melody (ID: 617) - Strickland Emmett, Evrard Marc, Lacheret-Dujour Anne 3141
Neural Speech Synthesis with Controllable Creaky Voice Style (ID: 717) - Lameris Harm, Włodarczak Marcin, Gustafson Joakim, Székely Éva 3146
Recovering implicit pitch contours from formants in whispered speech (ID: 736) - Pérez Zarazaga Pablo, Malisz Zofia 3151
Nasality detection from acoustic data with a convolutional neural network and comparison with aerodynamic data (ID: 934) - Elmerich Amélie, Kim Lila, Gendrot Cédric, Amelot Angélique, Crevier-Buchman Lise, Maeda Shinji 3156
Analysis-by-synthesis: phonetic-phonological variation in deep neural network-based text-to-speech synthesis (ID: 939) - Tånnader Christina, House David, Edlund Jens 3161
Natural choice: Comparing place classification between natural and Tacotron fricatives (ID: 949) - Pandey Ayushi, Le Maguer Sebastien, Edlund Jens, Harte Naomi 3166

17. Speech Corpora and Big Data

Knowledge-driven vs. data-driven methods for filtering acoustic measures in phonetics corpora (ID: 103) - Lancien Mélanie, Adda-Decker Martine, Stuart-Smith Jane 3171
The everyday speech environments of preschoolers with and without cochlear implants (ID: 276) - Cychosz Margaret, Newman Rochelle, Edwards Jan, Munson Benjamin, Romeo Rachel, Kosie Jessica 3176
Turning podcasts into a training corpus for conversational text-to-speech synthesis (ID: 323) - Sahkai Heete, Piits Liisi, Ermus Liis, Hein Indrek, Mihkla Meelis, Kiissel Indrek, Suluste Kristjan, Männisalu Egert, Altrov Rene, Pajupuu Hille, Pajupuu Jaan 3181
Using Mahalanobis distance to filter erratic vowel features in less documented languages : exploration on Quebec French (ID: 337) - Lancien Mélanie, Stuart-Smith Jane, Adda-Decker Martine 3186
Stop devoicing and place of articulation: a cross-linguistic investigation using large scale corpora (ID: 429) - Popescu Anisia, Hutin Mathilde, Vasilescu Ioana, Adda-Decker Martine, Lamel Lori 3191
Prosodic annotation of large speech corpora (ID: 471) - Martin Philippe 3196
How variable are English sibilants? (ID: 569) - Sonderegger Morgan, Stuart-Smith Jane, Mielke Jeff, SPADE Consortium The 3201
PhonD2: a database on phonotactic structures of German dialects (ID: 602) - Lameli Alfred, Werth Alexander, Bunkov Valeria, Link Samantha 3206
Articulation rate in consonants and vowels: results and methodological challenges from a cross-linguistic corpus study (ID: 618) - Lo Roger Yu-Hsiang, Sóskuthy Márton 3211
PEASYV: a procedure to obtain phonetic data from subtitled videos (ID: 837) - Méli Adrien, Ballier Nicolas 3216
IDEAR: A speech database of identity-marked, clear, and read speech (ID: 928) - Perepelytsia Valeriia, Bradshaw Leah, Dellwo Volker 3221

18. History of Phonetics

Phonlab project: Coimbra – Harvard. Rethinking 20th-century scientific centres and peripheries (ID: 40) - Lopes Quintino, Ferreira Carla, Pereira Gilberto, Ferreira Carolina, Silva Diogo, Quaresma Helena 3226
On the history of the phonetic institute of the Universite de Paris from 1911 to 1950 (ID: 256) - FOUGERON Cécile, AMELOT Angélique, RIALLAND Annie 3231
The 6th ICPhS in Prague and the organized life of phonetics (ID: 773) - Šturm Pavel 3236
On the evolution of ICPhS between 1932 and 2019 (ID: 856) - Trouvain Jürgen 3241

19. Phonetics of Lesser Documented and Endangered Languages

Phonetic description of filled pauses as discourse markers in Totoli (ID: 10) - Kaland Constantijn, Bardají Maria 3246
Aerodynamics and articulation of word-final ejectives in Eastern Armenian (ID: 15) - Toparlak Tabita, Dolatian Hossep 3251
Final lengthening across various sound classes in a language documentation corpus of Lower Sorbian (ID: 64) - Paschen Ludger 3257
Compensatory lengthening and polysyllabic shortening in the Gheg dialect of Albanian (ID: 129) - Riverin-Coutlée Josiane, Kapia Enkeleida, Cunha Conceição, Harrington Jonathan 3262
The /t/ release in Jutland Danish: Decomposing the spectrum with functional PCA (ID: 139) - Puggaard-Rode Rasmus 3267
An MRI examination of lingual fricatives in Upper Sorbian (ID: 142) - Howson Phil, Birkholz Peter 3272
The utility of turning-point analysis methods for Welsh and Irish intonation (ID: 239) - Pulman-Slater Jack, McCabe Connor 3276
Variable realization of the Arapaho glottal stop, despite its being distinctive and frequent (ID: 241) - Whalen D. H., DiCanio Christian, Chen Wei-Rong 3281
The realization of the voiceless nasal consonants in Upper Tanana (Dene) (ID: 247) - Lovick Olga, Reyes Herrera Kristy, Stewart Jesse 3286
Cotopaxi Media Lengua Vowels: A Preliminary Analysis (ID: 250) - Stewart Jesse, Onosson Sky 3291
The phonetic basis for modern Mongolian vowel harmony: An analysis by means of real-time MRI (ID: 318) - Saitô Yoshio, Maekawa Kikuo, - Yurong 3296
The Acoustics of Vowel Harmony in Sakha (ID: 331) - Chan May Pik Yu, Kuang Jianjing 3301
Phonetic variability in the realization of glottalized stops in Uspanteko (Mayan) (ID: 339) - Bennett Ryan, Henderson Robert, Harvey Meg 3306
Phonetics-prosody interplay is implicated in typologically rare and complex morphophonology in Tohono O'odham. (ID: 352) - Kim Daejin, Cruz Robert 3311
Focus prosody in Fijian: A pilot study (ID: 374) - Lee Albert, Simard Candide, Tamata Apolonia, Sun Jiaying, Mut Tsz Ching 3316
An acoustic study of the fricative vowel in two endangered Ryukyuan languages (ID: 387) - Shinohara Shigeko, Shao Bowei 3321
Vowel space and acoustic characteristics of stressed syllables in the two dialects of Kibushi: A pilote study (ID: 401) - Kassime Ahamada, Mori Miki, Hirsch Fabrice 3326
Word position and stress combine to affect voice onset time in Hawaiian (ID: 406) - Davidson Lisa, Parker Jones 'Ōiwi 3330
Acoustic correlates of metrical feet in two Yupik languages (ID: 419) - Alden McKinley, Arnhold Anja 3335
Vowel Raising In Cotopaxi Quichuan Languages (ID: 472) - Onosson Sky, Stewart Jesse 3340
A study on tones in Jiuhe Bai produced by Naxi speakers (ID: 529) - Wan Meihao, Wang Feng 3345
Intrinsic fundamental frequency differences in two tonal Austronesian languages (ID: 577) - Arnold Laura, Gao Jiayin, Kirby James 3350
Mapping phonetic variation to phonological units: Feature sharing in Comox-Sliammon (Salish) vowels (ID: 654) - Mellesmoen Gloria, Cardoso Amanda 3355
Evaluating forced alignment for under-resourced languages: A test on Squliq Atayal data (ID: 655) - Wang Chi-Wei, Chen Bo-Wei, Huang Po-Hsuan, Lai Ching-Hung, Chiu Chenhao 3360
Morphologyand phonetic realization: lexical vs. morphological prenasalization in Sà'án Sàvĭ ñà ñuù Xnúvíkó (Mixtec) (ID: 675) - Belmar Viernes Guillem, Salazar Jeremías 3365
Development of speech syntheses for lower sorbian and upper sorbian using marytts (ID: 687) - Schmiedel Astrid, Steiner Ingmar 3370
Resources and tools for pre-processing speech data in a lesser-known variety of English (ID: 725) - Vella Alexandra, Grech Sarah, Padovani Ian, Micallef Maria-Christina 3374
Acoustic correlates of allophonic plosive voicing in Romeyka Greek (ID: 743) - Yazar Merve, Bağrıaçık Metin, Canalis Stefano, Dadan Mehmet Can, Eryılmaz Metehan 3379
Word class and frequency effects in Hawaiian stressed vowel clusters (ID: 798) - Kettig Thomas 3384
Acoustic characterization of Najdi Arabic Vowel System (ID: 868) - Algethami Ghazi 3389
Acoustic and aerodynamic features of Hadza clicks (ID: 916) - Demolin Didier, Ghio Alain, Harvey Andrew 3392
Gemination in dogri and punjabi: evidence of deviation among 'like' languages (ID: 942) - Badyal Pranav 3397
Variation and palatalisation in the production of the plural prefix in Meru: A study of three dialects (ID: 970) - Cunha Conceição, Kanana Fridah, Harrington Jonathan 3402
The quantitative phonetic analysis of languages with small accessible populations (ID: 974) - Sóskuthy Márton, Chow Una, Mellesmoen Gloria, Sadlier-Brown Emily 3407
Evaluating formant estimations and discrete cosine transform to differentiate between pharyngeal fricatives in Mehweb (ID: 1027) - Arkhipov Alexandre, Daniel Michael, Shiryaev Alexander, Shepel Ekaterina 3412
Rate-sensitive differences between modal and non-modal vowels in San Martín Peras Mixtec (ID: 1029) - Eischens Ben 3417
The acoustic and articulatory properties of the prenasalised coronal trill in two Oceanic languages (ID: 1031) - Rangelov Tihomir 3422
Vowel duration in Enenlhet (ID: 1037) - Wheeler Paige-Erin 3427
A nasalance-based study of the /h/ vs. /ñ/ opposition in Zuberoan Basque (ID: 1047) - Egurtzegi Ander, García-Covelo Andrea, Urrestarazu-Porta Iñigo 3432
Pre-Pausal Glottalization in Chicontepec Nahuatl (ID: 1055) - Aguilar Andres 3437
Frication and formant frequencies in the Mundabli high vowels (ID: 1061) - Faytak Matthew, Shao Bowei, Taffre Douanla Angèle, Tschonghongei Nelson 3442
Pre-aspirated sonorants in Shehret, a Modern South Arabian language (ID: 1103) - Watson Janet, al-Kathiri Amer, Tome Lourido Gisela 3447

20. Field Methods in Phonetics

Evaluating cluster analysis on f0 contours: an information theoretic approach on three languages (ID: 108) - Kaland Constantijn, Ellison T. Mark 3452
Data from online production experiments and challenges for collecting good-quality recordings for prosodic analyses (ID: 152) - Bevivino Dorotea, Hemforth Barbara, Turco Giuseppina 3457

21. Phonetics of Conversation

Getting from A to B: complexities of turn change and retention in conversation (ID: 14) - Gilmartin Emer, Wlodarczak Marcin 3462
Characteristics and distribution of silent pauses in conversations between autistic and non-autistic dyads (ID: 24) - Wehrle Simon, Vogeley Kai, Grice Martine 3467
Points of maximum grammatical control – The prosody of a turn-holding practice (ID: 66) - Kelterer Anneliese, Wepner Saskia, Linke Julian, Schuppler Barbara 3472
Timing of laughter in conversation: Better late than never? (ID: 112) - Rathcke Tamara, Kapogianni Eleni, Page Lucy 3477
Voice quality dynamics of turn-taking events in Swedish and German (ID: 160) - Wlodarczak Marcin, Heldner Mattias, Bruggeman Anna, Wagner Petra 3482
People with Parkinson's disease exhibit phonetic flexibility (ID: 210) - Delvaux Véronique, Goeseels Eva, Huet Kathy, Piccaluga Myriam, Roland Virginie, Verhagen Clemence, Harmegnies Bernard 3487
Temporal aspects of turn-taking in Zoom conversations (ID: 448) - Xia Qiang 3492
Investigating the relationship between prosodic entrainment and interaction style (ID: 514) - Kruyt Joanna, Huttner Lena-Marie, O'Mahony Johannah 3497
Exploring global and local articulation rate entrainment in typical and atypical speakers (ID: 568) - Eijk Lotte, Stankova Stefany, Meekings Sophie 3502
TSCR: A comprehensive coding system for task-oriented vocal interaction analysis (ID: 624) - Copping Brandon, Holzmeyer Elizabeth, Kumar Santosh, Ones Deniz S., Buder Eugene H. 3507
Do humans converge phonetically when talking to a robot? (ID: 712) - Offrede Tom, Mishra Chinmaya, Mooshammer Christine, Skantze Gabriel, Fuchs Susanne 3512
What makes a good pause? Investigating the turn-holding effects of fillers (ID: 828) - Jiang Bing'er, Ekstedt Erik, Skantze Gabriel 3517

22. Sociophonetic Variation

Changes in parents' phonetic parameters during the first year of a child's life. Comparative study of German and Swedish (ID: 54) - Weirich Melanie, Simpson Adrian P. 3522
Variation in FACE and FLEECE trajectories in Australian English adolescents according to community language diversity (ID: 71) - Penney Joshua, Cox Felicity, Gibson Andy 3527
Implicit vs. explicit perception of French optional liaison as a marker of formality (ID: 84) - Dailey Megan, Peperkamp Sharon 3532
Pronunciation of gender-neutral German with respect to speaker attitude (ID: 147) - Belz Malte, Müller Miriam, Mooshammer Christine 3537
Being North Korean in Koreatown: The role of identity & experience in accommodation in the London North Korean community (ID: 148) - Evans Bronwen G., Shin Rakgyun, Lo Justin 3542
What you see is (not) what you hear? The role of social expectations on speech intelligibility in a British context (ID: 151) - Guo Ruohan, Cheng Fengjie, Evans Bronwen G. 3547
Changes in fundamental frequency and f1 across chronological age (ID: 163) - Parkman Seren 3552
Reconstructing the perception of gender identity, sexual orientation, and gender expression in American English (ID: 178) - Lang Benjamin 3557
Attitudes in the German-Polish context based on perceptual evidence (ID: 184) - Zygis Marzena, Blaszczak Joanna, Wesolek Sarah, Saint-Pettersen Markus Saint-Pettersen, Cwiek Aleksandra, Gulgowski Piotr 3562
Indexing femininity in vowel acoustics: A comparison between speakers in eastern and western parts of Germany (ID: 193) - Weirich Melanie, Zahner-Ritter Katharina 3567
The sociolinguistics of /r/-vocalization in spoken Swiss Standard German (ID: 195) - Vonlanthen Selma, Steiner Carina, Jeszenszky Péter, Leemann Adrian 3572
Acoustic and articulatory characteristics of rhoticity in the North-West of England (ID: 217) - Nance Claire, Dewhurst Maya, Fairclough Lois, Forster Pamela, Kirkham Sam, Nagamine Takayuki, Turton Danielle, Wang Di 3577
Evaluating the role of self-description in demarcating accents (ID: 238) - Cole Amanda, Strycharczuk Patrycja 3582
Vowel raising across Syria and Jordan in the DiVaL Corpus (ID: 240) - Hellmuth Sam, Almbark Rana, Lucas Chris, Brown Georgina 3587
Developmental changes of fundamental frequency and temporal characteristics in Estonian adolescent speech (ID: 257) - Meister Einar, Meister Lya 3592
Effects of dialect priming on phonetic convergence (ID: 281) - Clopper Cynthia, Dossey Ellen, Gonzalez Roberto 3597

Quantifying nasalance in Korean aegyo (ID: 297) - Crosby Drew, Dalola Amanda ... 3602
Apparent time change in the articulation of onset rhotics in Southern British English (ID: 315) - Strycharczuk Patrycja, Lloyd Susan, Scobbie James ... 3607
SOUNDS HEIMISH: A sociophonetic study of London's orthodox Jewish community (ID: 375) - Rubner Keren, McCarthy Kathleen ... 3612
What's that phthong? Automated classification of dialectal mono- and standard diphthongs (ID: 382) - Oppermann Simon, Siebenhaar Beat ... 3617
Towards a model of the listener in perception of queer speech (ID: 399) - Bedin Cooper ... 3622
The effects of gender and social class on the deaffrication of /ʤ/ in Jordanian Arabic (ID: 417) - Salman Amenah, Dalola Amanda ... 3627
The urge to unmerge: a case of structural change across the lifespan (ID: 442) - Barnard Marc, Hellyer Robert James, Holmes-Elliott Sophie . 3632
Voice types and voice quality in Japanese anime (ID: 476) - Ishi Carlos T., Utsugi Akira, Ota Ichiro ... 3637
Stop voicing in two settler communities on the Canadian Prairies (ID: 490) - Pfiffner Alexandra, Rosen Nicole ... 3642
Front vowel patterning in the Interlake region of Manitoba, Canada (ID: 491) - Rosen Nicole, Sullivan Lisa ... 3646
How gender affects immigrants' accent variation--An acoustic study on Hunan immigrants in Beijing (ID: 495) - ZOU Ziyan, YOU Shuxiang ... 3651
Context-induced phonetic shift in Silesian Czech speakers (ID: 567) - Šimáčková Šárka, Podlipský Václav Jonáš ... 3656
Connected speech strategies at word boundaries: taps and glottal stops in RP (ID: 570) - Sánchez-Leira Ana ... 3661
Acoustic merger between /e/ and /æ/ in Singapore English: insights into stylistic variation and sub-varietal difference (ID: 612) - Lan Canaan, Maxwell Olga, Diskin-Holdaway Chloé ... 3666
Front vowel lowering in New Zealand English: investigating a regional sound change (ID: 640) - Ross Brooke, Ballard Elaine, Waston Catherine ... 3671
Effects of Sociolinguistic Factors on Second Dialect Acquisition of North Korean Refugees in Seoul (ID: 680) - Lee Jungah, Idemaru Kaori, Vaughn Charlotte ... 3676
Segmental and suprasegmental modification in Rakugo-style Japanese (ID: 685) - Kamano Shigeto, Sakamoto Chikau, Tsujita Rin, Lee Seunghun J. ... 3681
Variation in the Scottish BIT vowel: Comparing two corpora (ID: 740) - Hall-Lew Lauren, Elliott Zuzana, Göbel Jessica, Cowie Claire, Markl Nina ... 3686
The Voiceless Velar Stop /k/ in Rijal Alma Arabic Revisited (ID: 741) - Yaqoub Leila, Hellmuth Sam, Bailey George ... 3691
A preliminary investigation of vocalic realizations of Hispanic micro-workers from Latin America (ID: 794) - Vasilescu Ioana, Wu Yaru, Lamel Lori, Torres Cierpe Juana, Tubaro Paola ... 3696
Redefining the variable context(s) for English (t/d)-deletion (ID: 807) - Walker James ... 3701
Aspiration of fortis plosives in multiethnolectal Zurich German (ID: 826) - Morand Marie-Anne, Watter Camille, Schmid Stephan ... 3705
Acoustic Correlates of Self-Ascribed Masculinity/Femininity in an American English sample (ID: 844) - Zhou Chen, Bishop Jason ... 3710
The sociophonetic variation of /j/ and /ɻ/ in malappuram malayalam (ID: 855) - Thesni K Naseera, Punnoose Reenu ... 3715
The impact of social information on VOT shadowing by nonbinary speakers (ID: 917) - Rechsteiner Jack, Sneller Betsy ... 3720
Modeling social meanings of phonetic variation amid variable co-occurrence: A machine-learning approach (ID: 945) - Villarreal Dan, Grama James ... 3725
Biases and speech-to-text efficacy for British English varieties (ID: 956) - Hung Kyra, Cardoso Amanda, Sharma Devyani, Levon Erez ... 3730
Quantifying Prosodic Variation in AAE as a Function of Gender and Interlocutor Using the PoLaR Framework (ID: 963) - Holliday Nicole ... 3735
Regional and individual variation in acoustic targets of /aɪ/ and /aʊ/ in American English (ID: 975) - Shport Irina, Bissell Marie, Berkson Kelly, Carmichael Katie ... 3740
The emergence of rhotic vowels in Quebec French: a change from below? (ID: 1025) - Lipari Massimo ... 3745
Linguistic stability and malleability in long-term dialect contact: The cot/caught contrast among mobile speakers (ID: 1039) - Nycz Jennifer ... 3750
Acquiring allophony: GOOSE and SCHOOL vowels in the speech of Australian children (ID: 1072) - Gibson Andy, Cox Felicity, Penney Joshua ... 3755

23. Forensic Phonetics and Speaker Characteristics

Bias and consistency of individual lingual articulatory behavior and its relationship to the first and second formants (ID: 131) - Lins Machado Carolina, He Lei ... 3760
Investigating the effects of talker age and listener age on trait perception from adult voices (ID: 144) - McGettigan Carolyn, Lavan Nadine ... 3765
What can voice line-ups tell us about voice similarity? (ID: 165) - Passetti Renata Regina, Madureira Sandra, Almeida Barbosa Plínio ... 3770
"YouTube space" – A preliminary investigation of vowel spaces and charisma perception on YouTube (ID: 168) - Berger Stephanie, Zellers Margaret, Niebuhr Oliver ... 3775
Just one word: an analysis of just as a speaker discriminant using various acoustic measures (ID: 176) - Gibb-Reid Ben ... 3780
Sensitivity of x-vectors and automatic speaker recognition scores to vocal variation (ID: 187) - Wormald Jessica, Foulkes Paul, Harrison Philip, Hughes Vincent, Kelly Finnian, van der Vloed David, Welch Poppy, Xu Chenzi ... 3785
Automatic assessment of voice similarity within and across speaker groups with different accents (ID: 196) - Gerlach Linda, McDougall Kirsty, Kelly Finnian, Alexander Anil ... 3790
Voice distinctiveness: an investigation of the role of speakers' position in a population with respect to f0 (ID: 349) - McDougall Kirsty, Paver Alice, Nolan Francis ... 3795

Evaluating discrete forensic voice evidence: a preliminary investigation based on filled pause occurrence (ID: 408) - Carne Michael 3800
Speaker discriminatory power of voice quality acoustics under forensic conditions (ID: 456) - Chan Ricky K. W. 3805
Between-speaker syllable intensity variability in Persian (ID: 457) - Asadi Homa, Alinezhad Batool 3809
Speaker idiosyncratic intensity and mouth opening-closing variations: the case of English (ID: 506) - Zhang Yu, He Lei, Dellwo Volker 3814
The Acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech (ID: 686) - Cao Grace Wenling, Mok Peggy 3819
ASR-based development of challenging speaker discrimination tests (ID: 758) - Fröhlich Andrea, Dellwo Volker, French Peter, Ramon Meike . 3824
The effect of targeted voice manipulations on long-term acoustic characteristics (ID: 766) - Houzar Alžběta, Nechanský Tomáš, Skarnitzl Radek 3829
Analysis of spectral features' change for speaker comparison (ID: 780) - Grászi Tekla Etelka, Krepsz Valéria, Huszár Anna, Száraz Bettina, Deme Andrea, Juhász Kornélia, Markó Alexandra 3834
Reviewing the performance of formants for forensic voice comparison: a meta-analysis of forensic speech science research (ID: 885) - Fairclough Lois, Brown Georgina, Kirchhübel Christin 3839
Some measures of phonetic similarity for use in legal trade mark disputes (ID: 894) - Ferrari Disner Sandra, van Heuven Vincent J. 3844
Voice comparison analysis of forensic recordings using the VoiceSauce program (ID: 973) - Jessen Michael, Konrat Canan, Horn Jonathan ... 3849
Exploration of simple methods of likelihood ratio calculation in forensic voice comparison (ID: 983) - Jessen Michael, Braun Almut, Menges Sonja 3854
Are there individual hesitation patterns? (ID: 1112) - Braun Angelika, Elsässer Nathalie 3859

24. Clinical Phonetics and Speech Disorders

Acoustic characteristics of stop consonants in Mandarin-speaking children with cochlear implants (ID: 61) - Yang Jing, Xu Li 3864
Artificial grammar learning in individuals with Down syndrome: the role of prosodic cues (ID: 117) - Stojanovik Vesna, Zimmerer Vitor, Setter Jane 3869
Rhythm deficits in developmental dyslexia: evidence from Italian (ID: 127) - Rossi Marina, Smit Eline, Rathcke Tamara 3874
A pilot fieldwork ultrasound study of tongue shape variability in children with and without speech sound disorder (ID: 200) - Smith Amy, Dokovova Maria, Lawson Eleanor, Kuschmann Anja, Cleland Joanne 3879
A Speech Therapy Animation and imaging Resource (STAR) (ID: 236) - Lawson Eleanor, Cleland Joanne, Stuart-Smith Jane, Aitken Brian, Beck Janet 3884
Probing production-perception relationships in children with myotonic dystrophy (ID: 242) - Menard Lucie, Bellavance-Courtemanche Marie, Trudeau-Fisette Paméla, Vidou Camille, Uribe Cristina 3888
Effect of auditory feedback on high lexical tones in deaf adults with cochlear implants (ID: 254) - Wan Tsung-Lun Alan 3893
Effect of hearing loss on prosodic intervention in Parkinson's speech (ID: 259) - Menezes Caroline, Pakulski Lori 3898
Exploring the Use of Phonological Features for Parkinson's Disease Detection (ID: 300) - Hosseini-Kivanani Nina, V´asquez-Correa Juan Camilo, Schommer Christoph, Nöth Elmar 3902
Towards disfluency features for speech technology based automatic dementia classification (ID: 459) - Thomas Megan, Hollands Samuel, Blackburn Daniel, Christensen Heidi 3907
Characterizing rhythm in dysarthric speech using the temporal envelope (ID: 578) - San Segundo Fernández Eugenia, Delgado Jonathan, He Lei 3912
Accommodation to vocal pitch in children with autism (ID: 581) - Hu Anqi, Qi Zhenghan, Franich Kathryn 3917
Speech metrics that differentiate variants of Primary Progressive Aphasia depend on language typology (ID: 598) - Elvira-García Wendy, Marrero-Aguiar Victoria, Borràs-Comes Joan, Santos-Santos Miguel Ángel 3922
A bimodal approach to study the effects of rhythmic priming (ID: 620) - Contreras Roa Leonardo, Mairano Paolo, Moreau Caroline, Basirat Anahita 3927
Phonetic-phonological factors in stuttering in German (ID: 756) - Jorschick Annett, Rachko Dinah, Cholin Joana 3932
Strategies of vocal adaptation to background noise of dysphonic and control schoolteachers (ID: 779) - Pettirossi Amelia, Audibert Nicolas, Crevier-Buchman Lise 3937
Phonological and sociophonetic information in dysarthric speech: a first investigation on Italian (ID: 875) - Gili Fivela Barbara, d'Apolito Sonia, Pagliaro Anna Chiara 3942
Synchronization type matters: Articulatory timing in different rhythmic conditions in persons who stutter (ID: 926) - Franke Mona, Benker Nicole, Falk Simone, Hoole Philip 3947
Nonword repetition in children with cochlear implants using different speech and language rehabilitation approaches (ID: 937) - Van Bogaert Lucie, Machart Laura, Lœvenbruck Helene, Vilain Anne 3952
Unsupervised clustering reveals several subtypes in speakers with ataxia (ID: 976) - Slis Anneke, Karlin Robin, Parrell Ben 3957
Vowel space area changes in children with dysarthria (ID: 1002) - Choi Jiyoung, Hwang KyungHae, Levy Erika 3962
Prenasalization in Greek-learning children with cochlear implants and typical hearing (ID: 1024) - Koenig Laura L., Okalidou Areti, Kyriafinis George 3967

Speech breathing in children with a cleft palate (ID: 1042) - Charuau Delphine, Vaxelaire Béatrice, Sock Rudolph 3972
Acoustic change over time in speech of one bilingual individual with Parkinson's disease (ID: 1073) - Verkhodanova Vass, Coler Matt, Jonkers Roel, de Bot Kees, Lowie Wander 3977
Atypical coarticulation in autism: evidence from sibilant-vowel interaction in Cantonese (ID: 1076) - Yu Alan, McAllister Robert, To Carol 3982

25. Phonetic Neurolinguistics

Aging effects on sentence-length velocity distribution characteristics of the tongue and jaw (ID: 563) - Raines Claudia, Thies Tabea, Ding Angela, Mefferd Antje 3987
Effects of speech modalities in acquiring neural markers of voice recognition: An ERP experiment using voice lineups (ID: 797) - Plante-Hebert Julien, Boucher Victor, Jemel Boutheina 3992
Interactions of speech and speaker processing in early evoked components (ID: 835) - Rinke Paula, Lavan Nadine, Scharinger Mathias 3997
Impaired vowel perception in congenital amusia: Evidence from event-related potentials (ID: 1020) - Pfeifer Jasmin, Hamann Silke 4002
Not all phonological neighbors affect production equivalently: predictions from a neural dynamic model (ID: 1023) - Stern Michael, Shaw Jason 4007

26. Phonetic Psycholinguistics

Words or rules: Phonological mutations in Wubuy (ID: 32) - Bundgaard-Nielsen Rikke, Baker Brett, Wang Yizhou 4012
Self-interruptions in breast cancer patients who complain of anomia (ID: 128) - RICHARD Amélie, HIRSCH Fabrice, JACQUIN-COURTOIS Sophie, REILLY Karen T. 4017
The influence of face masks on native and non-native memory recall of spoken sentences (ID: 138) - Truong Thanh Lan, Weber Andrea 4022
The use of syntactically redundant phonetic cues in speech perception (ID: 166) - Cohen Clara 4028
Self-monitoring of speech errors: effects of phonetic contrast (ID: 226) - Quené Hugo, Nooteboom Sieb G. 4033
Labels speak volumes: Outgroup labels and accent strength perception (ID: 307) - Calkins Kasandra, Puhacheuskaya Veranika, Kim Jiseung, Järvikivi Juhani 4038
Effects of stress and segment position within a word on speech errors in Croatian (ID: 384) - Vidović Zorić Ana 4043
Can segmental detail influence prosodic analysis? The case of contractions in English (ID: 444) - Mitterer Holger, Kim Sahyang, Cho Taehong 4048
Logarithmic durations for classifying and predicting Japanese voiced and voiceless plosives (ID: 504) - Yamakawa Kimiko, Amano Shigeaki, Kondo Mariko 4053
Logarithmic durations for classifying and predicting Japanese short and long vowels (ID: 527) - Amano Shigeaki, Hirata Yukari, Yamakawa Kimiko 4057
Phonological network properties of nonwords influence their learnability (ID: 555) - Turnbull Rory 4061
Mechanisms of generalization for phonetic learning of accented speech (ID: 590) - Melguy Yevgeniy, Johnson Keith 4066
Phonetic imitation of stops and vowels: Individual stability and perceptual underpinnings (ID: 610) - Zhang Hanna, Shertz Jessamyn 4071
The processing of prosodic prominence in German (ID: 849) - Zeyer Barbara, Penke Martina 4076
Temporal organization in lexical self-repair in Colombian Spanish unscripted speech (ID: 1067) - Vera Diettes Kelly Johanna, Plug Leendert .. 4081

27. Phonetics of Affective Speech

Towards Automating the Assessment of Teachers' Teaching Emotions in Preschool Classrooms (ID: 16) - Huo Mingyue, Tang Yan 4086
Production and perception of vocal emotions: A comparison of Mandarin Chinese and German emotional prosody (ID: 260) - Wei Huan, Scharinger Mathias, Domahs Ulrike 4091
An acoustic analysis of Berlin database of emotional speech based on bio-informational dimensions (ID: 489) - Cai Zhengyang, Xu Yi 4096
Modality Effects on Emotion Perception in English by Chinese L2 English Users: An Eye-tracking Study (ID: 541) - Liu Xiazhen, Zhou Weijing, Nolan Francis 4101
Acoustic analysis of Mandarin speech in a mixed emotion of pleasure and disgust (ID: 812) - GONG Mengjie, Gu Wentao 4106
Effects of adult attachment style on emotion: Evidence from physiological and acoustic measures (ID: 816) - Hu Han, Gu Wentao 4110

28. Multimodal Phonetics

Multimodal signalling: the interplay of oral and visual feedback in conversation (ID: 49) - Spaniol Malin, Janz Alicia, Wehrle Simon, Vogeley Kai, Grice Martine 4115
Kijk: a praat plugin to visualise articulatory trajectories (ID: 130) - Lins Machado Carolina, He Lei 4120
A multimodal account of listener feedback in face-to-face interactions (ID: 202) - Rossi Martina, Schröer Marin, Ludusan Bogdan, Zellers Margaret 4125
Iconic gestures in focus - synchronization of prosody and gesture in prominence (ID: 232) - Kügler Frank, Gregori Alina 4130

A kinematic analysis of visual prosody: Head movements in habitual and loud speech (ID: 319) - Pagel Lena, Sóskuthy Márton, Roessig Simon, Mücke Doris 4135
Oro-facial expressions and acoustic cues in German questions (ID: 441) - Zygis Marzena, Sardhaei Nasim, Krifka Manfred 4140
Computational models for articulatory learning of English diphthongs: One dynamic target vs. two static targets (ID: 453) - Xu Anqi, Gerazov Branislav, van Niekerk Daniel, Krug Paul, Prom-on Santitham, Birkholz Peter, Xu Yi 4145
The relation between pitch accent types, head movements and perceived prosodic prominence in L2 French (ID: 454) - Baills Florence, Baumann Stefan, Rohrer Patrick 4150
Developing a corpus of audio-visual attitudinal expressions in Hindi (ID: 462) - Mixdorff Hansjörg, Nayan Navneet, Rilliard Albert, Rao Preeti, Ghosh Debashis 4155
The development of audiovisual speech perception in Mandarin-speaking children (ID: 521) - WENG YI, PENG Gang 4160
Gestural alignment and accommodation in speaker-listener head gestures (ID: 623) - Garvin Karee, Franich Kathryn 4165
The prosody-gesture coordination in Chinese EFL learners' academic oral presentation (ID: 633) - Cao Yating, Chen Hua 4170
Comparing manual vs. semi-automated methods for the coding of co-speech gestures (ID: 635) - Dych Walter, Garvin Karee, Franich Kathryn 4175
Discrepant Combinations of Acoustic and Visual Speech Signals and Its Relation to Acoustic Distance in Vowels (ID: 718) - Matsubara Risa, Ariga Terumichi, Kishiyama Takeshi, Huang Chuyu 4180
Gestures-prosody (a)synchrony: Temporal interaction between head and eyebrow peaks in statements and yes-no questions (ID: 776) - Cruz Marisa, Frota Sonia 4185
Can segmental or syllabic durations be predicted by the presence of co-speech gestures? (ID: 801) - Nabrotzky Jule, Ambrazaitis Gilbert, Zellers Margaret, House David 4190
Is gesture-speech physics at work in rhythmic pointing? Evidence from Polish counting-out rhymes (ID: 955) - Kadavá Šárka, Ćwiek Aleksandra, Stoltmann Katarzyna, Fuchs Susanne, Pouw Wim 4195
Gesture and prosodic prominence in Ambonese Indonesian (ID: 967) - Farinella Alessa, Kaland Constantijn, Kaufman Daniel 4200
Towards identifying articulatory gestures with pixel difference and audio segmentation (ID: 1096) - Orlando Giulia, Palo Pertti 4205
Analysing speech data with SATKIT (ID: 1097) - Palo Pertti, Moisik Scott R., Faytak Matthew 4210

29. Phonetics Pedagogy

A survey of phonetics education in Italian universities (ID: 58) - Spreafico Lorenzo 4215
Re-evaluating the assessment of phonetic skills: what we learned during the pandemic (ID: 97) - Setter Jane, Knight Rachael-Anne, Whitworth Nicole 4219
Distant rhythms: computing fluency (ID: 174) - Lin Xuewei, Gibbon Dafydd 4224
Explaining voice characteristics to novice voice practitioners - How successful is it? (ID: 188) - Wiechmann Jana, Rautenberg Frederik, Wagner Petra, Häb-Umbach Reinhold 4229
An inclusive course on the history of phonetics (ID: 267) - Garellek Marc, Nielson Shai, Rhodes Tamara L., Clem Emily 4234
Intelligibility of Maritime English terms pronounced by Japanese deck cadets (ID: 426) - Uchida Yoko, Sugimoto Junko 4239
Objective versus perceived proficiency in pronunciation assessment of phonetic detail in advanced L2 English learners (ID: 575) - Romero Joaquín, Quesada Vázquez Leticia 4244
The Integration of Acoustic-Cue Visualization in HVPT of English Vowel Perception by Chinese ESL Learners (ID: 666) - Chen Hsueh Chu, Han Qianwen, Tian Jingxuan 4249

LASTING STRESS 'DEAFNESS' AFTER AUDITORY TRAINING: FRENCH LISTENERS REVISITED

Sharon Peperkamp and Jason Brazeal

Laboratoire de Sciences Cognitives et Psycholinguistique (CNRS, ENS-PSL, EHESS)
sharon.peperkamp@ens.psl.eu, mail@jasonbrazeal.com

ABSTRACT

Using a sequence recall task, we trained native speakers of French–a language without contrastive stress–to perceive a stress contrast. Contrary to a previous training study with French listeners, but in line with the persistent stress 'deafness' effect observed in advanced French learners of Spanish–a language with contrastive stress–we found no effect of training. We discuss these results in terms of differences in processing mode tapped by different perception tasks.

Keywords: speech perception, stress 'deafness', auditory training, French.

1. INTRODUCTION

A large body of research has shown that the perception of even the most difficult non-native sound contrasts can be improved by means of auditory training. Much of this successful training uses the high-variability phonetic training paradigm [1], focusing on a variety of segmental [1-7] or tonal [8-11] contrasts. Here, we examine another dimension, i.e. word stress.

Native speakers of French, a language without contrastive stress, have difficulty perceiving stress contrasts [12,13]. Previous research found that eight 30-minute training sessions improve stress perception in French listeners as assessed in an odd-one out task, but only in the absence of a large amount of phonetic variability [15]. (The results were the same for implicit training with a shape-word matching task and explicit training with several different tasks.) This moderate result may seem surprising, given that stress is signalled by substantial acoustic cues, i.e. duration, pitch, and intensity. Yet, research with L2 speakers has shown that the so-called stress 'deafness' effect is extremely robust: French speakers who are advanced learners of Spanish, a language with contrastive stress, have as much difficulty with perceiving stress as French monolinguals [14]. Thus, from this perspective, what is surprising is not so much that the training was only successful with little phonetic variability but that it was successful at all.

It is of course possible that focused training with explicit feedback is more effective than years of L2 learning. In the present study we examine this issue, and train French listeners to perceive stress by means of the same task as used with the advanced bilinguals in [14], i.e. sequence recall [13]. In this task, participants must recall sequences of two auditorily presented non-words that differ either in the position of stress (test condition) or in a phoneme (control condition), with the various tokens of the non-words being phonetically different. The combination of a memory load and phonetic variability ensures that this task taps a phonological processing level. The control condition concerns a native segmental contrast and provides an individual baseline with regard to phonological short-term memory.

Here, we use the sequence recall task in a pretest-posttest design, with six intervening training sessions.

2. METHOD

2.1. Materials

We used eight minimal pairs of CVCV non-words, one involving a phoneme contrast, i.e. /fíku/-/fítu/ (for the pre- and posttest), and seven involving a stress contrast (one for a pre- and posttest, and six for the training sessions), for instance /páku/-/pakú/. Each pair was instantiated by eight tokens, four per non-word item, recorded either by a single speaker or by two speakers (see Table 1); in the latter case, each speaker contributed two tokens per non-word item. Overall, there were three different speakers.

A resynthesis algorithm in the audio editor Adobe Audition was used to introduce global pitch variation into all stimuli except those used in three of the training sessions (see Table 1). For items that were recorded by a single speaker, the percentages of pitch modification were 94, 98, 102, and 106. For items that were recorded by two speakers, the percentages were 97 and 103, each applied to one token of each speaker.

The mean duration of the stimuli with the phoneme contrast was 454ms (/fiku/: 450; /fitu/: 459). The mean durations of the different stimuli sets with the stress contrast are shown in Table 2. The same table shows that in four sets, stressed vowels are longer, louder, and have a higher F0 peak than unstressed vowels. In the remaining three sets, one or two of these measures differ between stressed and unstressed vowels.

Table 1. Materials.

Session	Contrast(s)	ISI	Pitch change	Speaker(s)
Pre- and posttest	/fíku/ - /fítu/ /númi/ - /numí/	80ms	yes	French male
Training				
Session 1	/mípa/ - /mipá/	240ms	no	Dutch female
Session 2	/páku/ - /pakú/	240ms	yes	Dutch female
Session 3	/mítu/ - /mitú/	160ms	no	English male
Session 4	/támi/ - /tamí/	160ms	yes	English male
Session 5	/núta/ - /nutá/	80ms	no	Dutch female + English male
Session 6	/kánu/ - /kanú/	80ms	yes	Dutch female + English male

Table 2. Mean durations of stress-initial and stress-final stimuli, and mean differences between stressed and unstressed vowels in terms of duration, F0, and energy.

Session	Total duration		Difference between stressed and unstressed vowels		
	Stress-initial	Stress-final	Duration (ms)	Max F0 (Hz)	Energy (dB)
Pre- and posttest	452	406	38.9 **	72.8 ***	5.1 ***
Training					
Session 1	504	473	17.6 **	51.7 ***	4.3 ***
Session 2	445	427	10.5	54.8 ***	5.1 ***
Session 3	451	523	48.3 **	25.9 ***	4.6 ***
Session 4	455	508	60.6 ***	9.9 **	5.7 ***
Session 5					
Speaker 1	503	435	-5.4	65.9 ***	5.4 ***
Speaker 2	511	505	14.3 ~	13.4 ***	4.8 ***
Mean	*507*	*505*	*4.5*	*39.7*	*5.1* ***
Session 6					
Speaker 1	416	394	10.2	61.8 **	5.7 ***
Speaker 2	411	426	39.8 **	15.4 *	4.9 ***
Mean	*414*	*426*	*25.0* *	*38.6*	*5.3* ***

2.2. Procedure

The pre- and posttest were identical. They consisted of two parts, for the phoneme and the stress contrast, respectively. In each part, participants were first asked to press the number keys 1and 2, upon which they heard all the tokens of the first and second item, respectively. Subsequently, they could continue listening to the various tokens of the two items by pressing the associated keys; pressing each one of these keys resulted in the playing of one token of the corresponding item. Next, it was verified that they had learned the distinction between the two items as well as the correct association between the items and the number keys. That is, they heard a token of one of the items and had to press the associated key, 1 or 2. A message on the screen informed participants whether their response was correct or incorrect. After having given seven correct responses in a row, participants turned to the test phase, during which they were presented with 5 blocks of 16 sequences constituted by repetitions of the two items. The sequence length increased at each block, varying from two in the first block (e.g., /pakú páku/) to six in the last block (e.g. /pakú pakú páku páku pakú páku/). Participants had to reproduce each sequence by typing the associated keys in the correct order. The order of the 16 sequences in each block was randomized, and each item was instantiated randomly by one of the four recorded tokens. Each trial consisted of a sequence with an ISI of 80ms, followed by the word 'OK'. Participants could not begin typing their response until they had heard this word, and the short ISI prevented them from mentally translating the non-words into the associated numbers while they listened to the sequence. A 1500ms pause separated each response from the next trial.

The pre- and posttest lasted around 30 minutes. They were separated by between 9 and 17 days (mean controls: 12.0; mean trainees: 11.5; $t<1$). During this period, the trainees but not the controls went through six training sessions, in which they were trained on novel pairs of non-words differing only in the position of stress. The training started the day after the pretest and ended between one and six days before

the posttest (mean: 1.7). Trainees could not complete more than one session per day, and no more than four days elapsed between sessions, except for one person who had a seven-day lapse between two training sessions. Each session lasted around 30 minutes.

At the beginning of the first training session the objective of the study was explained. That is, participants were told that in some languages words can differ solely by their stress pattern; that since this type of contrast does not exist in their language, French listeners find it hard to perceive; and that the goal of the study was to train their ears to perceive it better. Each of the following training sessions began with a reminder that the two items they would be trained on differ only in the position of stress and that stressed syllables are longer and pronounced with a higher and louder voice than unstressed ones.

The procedure for the training sessions differed from that of the pre- and posttest in two aspects. First, there was no phoneme contrast. Second, participants received feedback during the test blocks: If the answer was correct, positive feedback was displayed on the screen. An error message was displayed otherwise, together with a message asking the participant to re-listen to the sequence and re-enter a response; after 2500ms, the same sequence as before was played. Once more, if the answer was correct, positive feedback was provided. If the second answer was still incorrect, the computer screen displayed the error message as well as the correct transcription and an invitation to listen to the sequence once more; after 2500ms, the sequence was played one last time with the correct answer still displayed on the screen.

Differences across the training sessions with regard to the ISI, the presence vs. absence of global pitch variation, and the number of speakers (Table 1) were loosely meant to gradually increase the level of difficulty.

2.3. Participants

Twenty native speakers of French were tested in Paris. Half of them, two men and eight women aged between 20 and 30 (mean: 24), participated in the pre- and posttest only. The other half, three men and seven women aged between 20 and 31 (mean: 23), also participated in the training sessions. None had ever lived abroad, but all had studied one or more foreign languages—including languages with non-predictable stress—at school.[1]

3. RESULTS AND DISCUSSION

3.1. Pre- and posttest

Responses in the test phase that were a 100% correct transcription of the input sequence were coded as correct; all other responses were coded as incorrect. Table 3 shows the mean error rates.

Table 3. Mean error rates in the pre- and posttest. Standard errors are shown in parentheses.

		Pretest	Posttest
Trainees	Phoneme	27.8 (4.2)	22.1 (5.6)
	Stress	55.9 (7.1)	48.6 (7.4)
Controls	Phoneme	25.6 (3.5)	21.5 (3.3)
	Stress	53.1 (5.9)	50.5 (6.0)

Using the *lme4* package [16] in the R environment [17], we analysed these data in a logistic regression model with contrast-coded fixed factors Group (trained vs. control), Test (pre- vs. posttest), Contrast (phoneme vs. stress), and all the interactions, and a random intercept for Participant (with one or more added slopes the model would not converge). Statistical significance was assessed by means of model comparison using a likelihood ratio test. The results showed effects of Contrast (β=0.66, SE=0.03, z=23.2, χ^2=575, p<.0001) and Test (β=-0.13, SE=0.03, $|z|$=4.4, χ^2=19.8, p<.0001), but no effect of Group (β=-0.001, SE=0.15, $|z|$<1) and no interactions (Group × Test: β=0.04, SE=0.03, z=1.31, χ^2=1.72, p>.1; Group × Contrast: β=0.004, SE=0.03, z<1; Group × Test × Contrast: β=-0.02, SE=0.03, $|z|$<1).

Thus, for all participants performance was better on the phoneme than on the stress contrast, and better in the post- than in the pretest. Even though the interpretation relies on a null result (i.e. the absence of a triple interaction), this pattern of results strongly suggests that training did not have an effect. Rather, compared to the pretest, both trainees and controls showed a small, contrast-independent, increase in performance in the posttest, presumably due to having been familiarized with the task.

3.2. Training sessions

Trials with a 100% correct transcription of the input sequence on either the first or the second try were coded as correct, all others as incorrect. Table 4 shows the mean error rates. (When only the first try is taken into account, the error rates are between 11.5 and 19.0 percentage points higher (mean: 15.8).)

Recall that variations in the training sessions were loosely meant to gradually increase the level of difficulty. It is easy to see that this was not the case.

Table 4. Mean error rates in the six training sessions. Standard errors are shown in parentheses.

1	2	3	4	5	6
15.12 (11.3)	39.12 (15.4)	7.50 (8.3)	12.12 (10.3)	33.00 (14.9)	36.40 (15.2)

As to the factors that may influence task difficulty, it is impossible to draw conclusions because of collinearity and the fact that each session contained unique items. Yet, comparing sessions 1, 3, and 5 to sessions 2, 4, and 6, respectively, we note that the added global pitch variation in the latter always yielded higher error rates, as expected given [13]. Manipulation of ISI, however, did not seem to have a linear effect, contrary to what we expected. In particular, while compared to the ISI of 80ms (sessions 5-6), lower error rates were obtained with the ISI of 160ms (sessions 3-4), presumably because it leaves participants the time to mentally translate the non-words online into the associated numbers, there was no additional drop in error rate for the ISI of 240ms (sessions 1-2). Rather, the 240ms ISI yielded higher error rates than the 160ms ISI.

We speculate that the 240ms ISI may be hard because the associated numbers have to be kept in short-term memory for a longer time (see [18,19] for a similar effect in AX discrimination). According to this interpretation, the difference in performance with a 240ms compared to a 160ms ISI should increase as the sequence length—and hence the total trial duration—increases. And indeed, from shortest to longest sequence length the differences in mean error rates are 4.4, 11.7, 16.6, 28.1, and 26.3 percentage points, showing a robust linear increase ($R^2 = 0.89$, $p < 0.04$).

4. GENERAL DISCUSSION

Using the sequence recall task of [13], we failed to improve stress perception in French listeners. This finding meshes well with the persisting stress 'deafness' effect seen in advanced French L2 learners [14], but it contrasts with a previous study reporting successful training with an odd-one out task [15].

What might account for the difference between the two training studies? The amounts of training were comparable, i.e. six 30-minute sessions for a total of 480 trials consisting of 1920 tokens in our study vs. eight 30-minute sessions for a total of 1728 single-token trials in [15]. Furthermore, in studies on segmental and tonal contrasts that provide analyses of the training sessions, the largest improvement is often reported for the earlier sessions [2,3,6,9,20]. Thus, it seems unlikely that making the training last longer would allow us to observe an effect.

Rather, we argue that our trainees failed to improve their stress perception for other reasons. It is well-known that differences in both stimulus complexity and task demands influence non-native sound perception [18,19,21]. For instance, the automatic selective perception model [23] distinguishes a phonetic and a phonological processing mode. The latter is called upon when rapid processing is required, and performance in this mode is worse than in the former. As to stimuli complexity, our test stimuli had a large amount of phonetic variability (four single-speaker tokens per item, with additional global pitch variation), while successful training as assessed in an odd-one out task was obtained only when there was little variability (each item recorded once by two female speakers) [15]. As to task demands, the sequence recall task requires participants to encode the auditory input in their short-term memory buffer in order to recall the sequence. With phonetically varied stimuli this encoding cannot use a low-level acoustic or phonetic representation but must be phonological in nature. As French listeners do not represent stress phonologically, they fail to improve in the sequence recall task. By contrast, the odd-one out task of [15] is less constraining, especially in the version with very limited phonetic variability.

We also note that a short ISI prevents participants from rehearsing the stimuli subvocally. Findings that successful training can generalize from perception to production [22-24] suggests that standard training tasks may trigger subvocal rehearsal. A long ISI of 500ms was used in the odd-one out task of [15], leaving ample time for this strategy; as there were only three items in each trial, this benefit of a long ISI may not have been counteracted by the drawback of the increased memory load that we saw in the sequence recall task.

Overall, then, successful training appears to rely on a low-level response strategy and/or activation of the perception-production loop in pre- and posttest [25,26], both of which are largely ruled out in the sequence recall task. Yet even successful training can fail to withstand a stricter test: in [15], the exact same training did not improve stress perception when the amount of phonetic variability in pre- and posttest was increased. We note that possibly, improved perception of segmental and tonal contrasts after training in the high-variability phonetic training paradigm [1] may similarly be observed only in the relatively unconstraining identification and discrimination tasks that are typically used in this paradigm. Thus, future research could investigate the extent to which the benefits of this widely used training paradigm resist a cognitively more demanding perception task. This is especially relevant when considering the question of the extent to which training aids real life L2 processing, where the cognitive load is particularly high.

To conclude, we have provided evidence that French listeners' inability to accurately represent and process stress at a phonological level resists not only L2 learning but also explicit auditory training.

7. REFERENCES

[1] Logan, J, Lively, S, Pisoni, D. 1991. Training Japanese listeners to identify English /r/and/l/: A first report. *J. Acoust. Soc. Am.* 89, 874–886.

[2] Flege, J. 1995. Two procedures for training a novel second language phonetic contrast. *Appl. Psycholinguist.* 16, 425–442.

[3] Pruitt, J., Jenkins, J., Strange, W. 2006. Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *J. Acoust. Soc. Am*. 119, 1684-1696.

[4] Shinohara, Y., Iverson, P. 2018. High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *J. Phonetics*, 66, 242–251.

[5] Kingston, J. 2003. Learning foreign vowels. *Lang. Speech* 46, 295-349.

[6] Nishi, K., Kewley-Port, D. 2007. Training Japanese listeners to perceive American English Vowels. Influence of training sets. *J. Speech Lang. Hear. R*. 50, 1496-1509.

[7] Iverson, P., Evans, B. 2009. Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *J. Acoust. Soc. Am.* 126, 866-877.

[8] Sadakata, M., McQueen, J. 2014. Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Front. Psychol.* 5, 1318.

[9] Wang, Y., Spence, M., Jongman, A., Sereno, J. 1999. Training American listeners to perceive Mandarin tones. *J. Acoust. Soc. Am*. 106, 3649–3658.

[10] Wang, Y., Jongman, A., Sereno, J. 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *J. Acoust. Soc. Am*. 113, 1033–1043.

[11] Kaan, E., Barkley, C., Bao, M., Wayland, R. 2008. Thai lexical tone perception in native speakers of Thai, English and Mandarin Chinese: An event-related potentials training study. *BMC Neurosci.* 9, 53.

[12] Dupoux, E., Pallier, C., Sebastián-Gallés, N., Mehler, J. (1997) A destressing 'deafness' in French? *J. Mem. Lang.* 36, 406-421.

[13] Dupoux, E., Peperkamp, S., Sebastián-Gallés, N. (2001) A robust method to study stress 'deafness'. *J. Acoust. Soc. Am*., 110, 1606-1618.

[14] Dupoux, E., Sebastián-Gallés, N, Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': the case of French learners of Spanish. *Cognition* 106, 682-706.

[15] Schwab, S., Dellwo, V. 2021. Explicit versus non-explicit prosodic training in the learning of Spanish L2 stress contrasts by French listeners. *J. Second Lang. Stud*. DOI:10.1075/jsls.21017.sch

[16] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 671, 1-48.

[17] R Core Team 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[18] Werker, J., Tees, R. 1984. Phonemic and phonetic factors in adult cross-language speech perception. *J. Acoust. Soc. Am*. 75, 1866–1878.

[19] Werker, J., Logan, J. 1985. Cross-language evidence for three factors in speech perception. *Percept. Psychophysics* 37, 35–44.

[20] Lively, S., Pisoni, D., Yamada, R., Tohkura, Y., Yamada, T. 1994. Training Japanese listeners to identify English R and L. III. Long-term retention of new phonetic categories. *J. Acoust. Soc. Am*. 89, 874–886.

[21] Strange, W. 2011. Automatic selective perception ASP of first and second language speech. A working model. *J. Phonetics* 39, 456-466.

[22] Rochet, B. 1995. Perception and production of second-language speech sounds by adults. In: W. Strange (ed) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press, 379-410.

[23] Bradlow, A., Pisoni, D., Akahane-Yamada, R., Tohkura, Y. 1997. Training Japanese listeners to identify English R and L: IV. Some effects of perceptual learning on speech production. *J. Acoust. Soc. Am*. 101, 2299-2310.

[24] Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., Molholt, G. 2005. The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Appl. Psycholinguist*, 262, 227-247.

[25] Baddeley, A., Gathercole, S., Papagno, C. 1998. The phonological loop as a language learning device. *Psych. Review* 105, 158–173.

[26] Jacquemot, C., Scott, S. 2006. What is the relationship between phonological short-term memory and speech processing? *Trends Cogn. Sci*. 10, 480–486.

[1] One additional participant was withdrawn from the experiment, because for the stress contrast in the pretest they had not passed the success criterion of the verification phase after 200 trials. The data from a second additional participant were excluded, because in the posttest they made twice as many complete reversals (for instance, 1211 instead of 2122) than correct responses on the stress contrast, suggesting a potential confusion in the association between the items and the response keys.

Exposure-independent comprehension of Greek-accented speech: evidence from New Zealand listeners

Stephanie Cooper[1] & Sarah Cooper[2]

[1]Phonetics Laboratory, University of Cambridge, [2]School of Arts, Culture and Language, Bangor University
sc2346@cam.ac.uk, s.cooper@bangor.ac.uk

ABSTRACT

Accents provide extensive variation in speech, which can cause difficulties for comprehension. However, after a period of exposure to a foreign accent, listeners are usually able to understand it better. This study examined whether New Zealanders could adapt to consider Greek-accented raised /ɪ/ in English as part of their native centralised /ɪ/ phonetic category. Participants listened to a story in either Greek-accented or New Zealand-accented speech, and then underwent a cross-modal priming and lexical decision task as well as a rating task to examine perception at different levels of speech processing. Participants correctly identified words in Greek-accented English whether they had previous exposure to the accent or not. This effect extended throughout the speech processing system, from automatic lexical activation to deliberate categorisation. We discuss reasons for these results, including perceptual flexibility for variant forms of vowels and a high level of familiarity New Zealanders have with the Australian accent.

Keywords: Speech perception, perceptual learning, accent comprehension, dual mapping

1. INTRODUCTION

Foreign accents are known to be a challenge for speech comprehension. Previous research has shown that listeners are capable of adapting to accented forms of words by learning from prior exposure to the accent [7], [16]. They allegedly achieve this by broadening their mental representations of sounds in their native language, also known as phonetic categories, to accommodate more variation in sounds. They also adjust the internal structure of these categories so that their exemplar sound is closer to the accented form [18].

As groups of people speak with relatively similar accents, generalising comprehension from one speaker to another should be beneficial. Xie & Myers [17] found that when listening to Mandarin-accented English speech, which usually devoices word-final /d/, listeners' generalisation of /d/ comprehension from one speaker to another was constrained by how acoustically similar the two speakers were in their /d/ production. Bottom-up information from the speech signal (i.e. similar vowel duration, closure duration and burst duration in word-final /d/) therefore drives generalisation more than top-down knowledge that two speakers are from the same accent group (e.g. explicit knowledge that both speakers have Mandarin accents).

"Dual mapping" of multiple pronunciations to one lexical representation may assist in specific instances of generalisation. For instance, Samuel & Larraza [12] found that highly proficient Basque-Spanish bilingual listeners accepted incorrect pronunciations of Basque sounds by L1 Spanish speakers even when no exposure phase was given. Similar results were found in Catalan-Spanish bilinguals [13]. This suggests that dual mapping develops when listeners are accustomed to hearing accented speech, and could aid rapid comprehension of the accent when listeners consider its sounds as allophones of native categories.

Exposing New Zealand-accented participants to Greek-accented speech is a particularly interesting way of examining potential accent adaptation. Firstly, these accent groups have not been examined in previous accent adaptation studies. Secondly, Greek speakers often produce /ɪ/ as [i] in English, due to the lack of /ɪ/ in Greek [11]. Finally, the New Zealand-accented centralised /ɪ/ is further from /iː/ in the vowel space than other English accents [2], so there should be even further ground for New Zealand listeners to cover when broadening their native /ɪ/ phonetic categories to include [i]-like productions.

In the present study, we exposed one experimental group of participants to Greek-accented English and one control group to New Zealand English, to see whether this experience would cause the experimental group to shift their phonetic category boundary for /ɪ/ to include /i/-like productions. We then used a cross-modal priming and lexical decision task and a categorisation task to measure this potential accent adaptation. Priming examines subconscious, automatic, online speech processing, while categorisation examines conscious, deliberate, offline speech processing [18]. Performance on these tasks may reveal if a boundary shift occurs for the experimental group compared to

the control group, and then how far through the speech processing system it extends.

2. METHOD

2.1. Participants

45 participants took part in the accent adaptation study online. They were aged 18-29, currently living in New Zealand, and spoke New Zealand English. Most participants were raised in and currently living in Tāmaki Makaurau (Auckland), but some were from other regions such as Te Whanganui-a-Tara (Wellington) and Waitaha (Canterbury).

2.2. Materials

2.2.1. Priming task

180 pairs of English prime words, target words and non-words were chosen for the online cross-modal priming and lexical decision task. Word frequencies per 100 million words ranged from 200-7000 [6]. All words and non-words were one syllable, with CV, CVC, CVCC or CCVC structures. All words and non-words were 3-5 orthographic characters long.

There were 9 types of trial in total, with the most relevant critical and filler trials described in more detail in Table 1. These trials were designed to examine whether hearing a Greek-accented word containing /ɪ/ primed that same word, or an /iː/ minimal pair of the word, and how this compared with normal same word or minimal pair priming. There were 15 of each relevant trial.

Trial type	Description	Example
Critical-same	/ɪ/ primes, same word targets	*kiss*-KISS
Critical-MP	/ɪ/ primes, minimal pair targets	*ship*-SHEEP
Filler-same	/ɔː, ɜː, ɒ, uː/ primes, same word targets	*boss*-BOSS
Filler-MP	/ɔː, ɜː, ɒ, uː/ primes, minimal pair targets	*lawn*-LEARN

Table 1: Relevant trials in cross-modal priming and lexical decision task. Trials are depicted as *auditory prime*-VISUAL TARGET.

Other types of trials were designed to indicate how participants responded to targets when they were different to the primes. 90 trials also used word primes and nonword targets, so participants would properly attend to the lexical decision task.

It was predicted that the experimental group would respond more rapidly to critical-same trials where the target word matched the prime, as they would be more likely to correctly identify the prime.

2.2.2. Categorisation task

25 English words containing /ɪ/ that have /iː/ minimal pairs were used as auditory stimuli in the offline categorisation task. The words used in the offline task were also used in the online processing task in the critical-same or critical-MP trials.

2.3. Speakers

Speech was recorded from a Greek-accented speaker and a New Zealand-accented speaker.

The Greek-accented speaker speaks Greek as a first language. She is highly fluent in English. She recorded the story “Comma Gets a Cure” [8] for the experimental group exposure phase. She also recorded the 180 words that were used in the priming task, 25 of which were also used in the categorisation task.

The New Zealand-accented speaker is a monolingual New Zealand English speaker. She recorded the story “Comma Gets a Cure” for the control group exposure phase. She also recorded 10 words for practice trials of the priming task.

2.4. Procedure

The experiment was designed using the Gorilla Experiment Builder [1]. Participants completed the experiment virtually using their own laptop. They were instructed to sit in a quiet place and use headphones for the experiment.

Participants were provided with details of the study, including information on how to withdraw. After consenting to take part, they were randomly assigned to either the experimental group or the control group. The experimental group heard the story “Comma Gets a Cure” [8] in Greek-accented English, while the control group heard the story in New Zealand English. They then progressed to the remainder of the experiment.

2.4.1. Priming task

The first task employed cross-modal priming and lexical decision. In each trial, participants heard a Greek-accented prime word then made a lexical decision about a target on the screen, which may have been the same word, a different word or a

nonword. They indicated their response using the keyboard. Response times were used to explore whether the target word had been activated when the auditory stimulus was presented. This task reflects online processing by probing instantaneous activation of lexical representations.

Participants were given instructions to the task, followed by 10 practice trials with auditory stimuli (using /ɛ/ and /a:/ vowels) produced by the New Zealand-accented speaker.

The main task was presented with breaks every 45 trials for a total of 180 trials. As soon as a button press was recorded, the next trial began. No feedback was given. Trials were presented in a randomised order, but primes and targets always appeared in the same pairs.

2.4.2. Categorisation task

After the priming task, the categorisation task began. All words in this task were also produced by the Greek-accented speaker. Words contained /ɪ/ sounds and had /i:/ minimal pairs, such as *ship* and *sheep*. Participants heard a word, then marked on a sliding 7-point scale between the two words where what they heard fell. A rating of 1 meant the word was heard strongly as its /ɪ/ version, such as *ship*, while a rating of 7 meant the word was heard strongly as its /i:/ version, such as *sheep*.

There were 25 trials in this task. This task probed offline processing by asking participants to make a conscious decision about word category.

2.5. Analysis

For the priming task, the effects of accent exposure and trial type on log response times were analysed using a mixed effects model. The group exposed to Greek-accented speech were the experimental group, and the group exposed to New Zealand-accented speech were the control group. For the categorisation task, effects of accent exposure on ratings of where words sat on the /ɪ/-/i:/ spectrum were analysed using a Mann-Whitney U test.

3. RESULTS

3.1. Priming task

3.1.1. Mean response time analysis

7759 out of 8100 trials were included for analysis, with those excluded being incorrect lexical decisions or outliers. A linear mixed-effects model with maximum likelihood was then used to assess response time averages. Exposure and Trial Type were included as fixed effects, while Participant and Item were included as random effects. Figure 2 shows the two groups' responses to the relevant trial types used in this model.

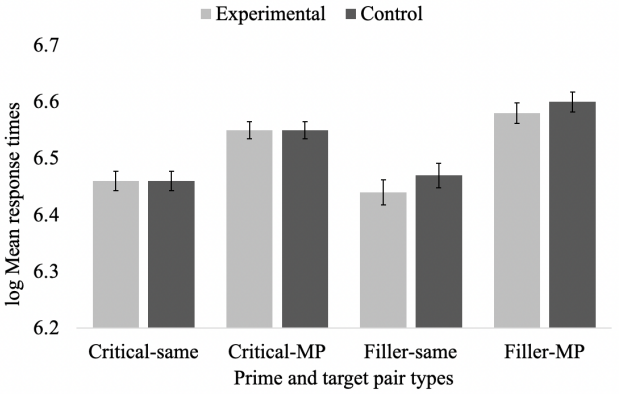


Figure 2: Log mean response times to different trial types by experimental group and control group participants in priming task.

ANOVAs between potential models found that a model with Exposure removed was not significantly different from the original model (X^2 = 0.07, p = 0.7859), indicating that participants performed at similar levels in this task irrespective of the accent they had been exposed to at the start. Further analysis was then completed using this model, with Category as a fixed effect and Participant and Item as random effects. This model's explanatory power (R^2 = 0.41) is substantial.

To investigate the potential effects of priming in the different types of trials, a post-hoc Tukey's test for mixed effects models was performed. This revealed that critical-same (e.g. *kiss*-KISS*)* and filler-same trials (e.g. *boss*-BOSS) mostly showed significantly different mean response times to six other trial types. The one exception was critical-MP trials (e.g. *ship*-SHEEP), which did not show significant differences in mean response times with both critical-same (p = 0.054) and filler-same (p = 0.08) trials. There was no significant difference in mean response times between critical-same and filler-same trials (p = 1.00).

These results suggest that priming was occurring, as trials in which primes and targets were the same word were responded to faster than trials with contrasting primes and targets. Additionally, accented /ɪ/ words successfully primed their written counterpart in the same way as words with no effect of accent, although there did appear to also be a small priming effect for /i:/ minimal pairs in the critical-MP trial.

3.1.1. Learning across task

To examine whether the groups showed different patterns of response times throughout the course of the experiment, mean response times at trial numbers 1-180 were calculated for each group.

All means showed a general downwards trend as trials progressed, which is likely to reflect a growing familiarity with the experiment. Mean response times for critical-same and critical-MP trials do not appear to show different directionality to the trend across all trials. This suggests that little learning occurred in these types of trials that was different to learning across the experiment. Control participants were therefore unlikely to have learned the Greek accent by hearing the speech in the task.

3.4. Offline processing task

A Mann-Whitney U test was conducted to measure the potential difference between groups in their responses in this task. This non-parametric test was chosen instead of an independent samples t-test as the responses are ordinal rather than continuous, and were not normally distributed.

There was no significant difference in median responses between the group exposed to Greek-accented speech (Mdn = 2) and the group exposed to New Zealand-accented speech (Mdn = 2) ($W = 155604$, $p = 0.63$), This indicates that both groups tended towards perceiving what they heard as /ɪ/ versions of the words, as intended by the speaker.

4. DISCUSSION

The results of this study indicate that participants correctly identified potentially ambiguous words in a foreign accent even without prior knowledge of this accent. This was indicated in the priming task, where accented /ɪ/ words successfully primed their written form, and in the categorisation task, where participants mostly rated the word they heard as being on the /ɪ/ side of the minimal pair scale. This accurate comprehension therefore occurred at both subconscious online and conscious offline levels of processing, which indicates that knowledge of the intended sounds could be extended throughout the speech system. Crucially, these results were almost identical for participants who had exposure to Greek-accented speech and for those who did not.

These results are mostly inconsistent with previous findings. Other studies [7], [18] found that experimental groups who had been exposed to accented speech responded rapidly to an equivalent of critical-same trials, but control groups did not. Xie et al. [18] also examined offline processing of accented words, and again found differences between experimental and control groups. Similar results were found in studies examining learning of accented vowel sounds [4], [10], [16].

Xie et al. [18] did also find that participants were unable to eliminate minimal pairs as competitors to accented words. They suggested that multiple aspects of accent that do not appear in lab-generated speech, among other factors, may have contributed to this. This may apply to the present study, in addition to participants having little prior knowledge of the accent (3 minutes for the experimental group and 0 for the control group). Minimal pairs may not have been eliminated as strongly as they would be if the acoustic signal only contained one altered feature from native speech or if participants had more exposure to the accent. It is worth noting that the difference in response times between critical-same and critical-MP trials was close to reaching significance, so the activation of minimal pairs when hearing an /iː/-like /ɪ/ word appeared to be low.

It is unlikely that the control group's ability to correctly identify accented /ɪ/ words was due to learning the accent in the online processing task, as response times in critical trials were consistent across the whole experiment. Various other factors may have contributed to this ease of processing.

The Greek-accented speaker is highly fluent in English, so her accent may have been relatively easy to understand compared to someone who does not speak English as fluently. Listeners may also generally be more willing to accept variation from their native categories in vowels than in consonants, as has been suggested by previous studies [14], [15]. New Zealand being highly multicultural [9], and therefore perhaps highly multilingual, may also contribute to this perceptual flexibility.

Another possibility is that New Zealanders may be highly familiar with the Australian accent, which is also known for its raised /ɪ/ [3]. Due to regular exposure to the Australian accent in their daily lives, New Zealand listeners may have accommodated Australian-accented vowels as allophones of their native categories, in line with the "dual mapping" theory proposed by Samuel & Larraza [12]. Greek-accented /ɪ/ may then follow these routes to be correctly categorised with ease. Further research is required to investigate whether these types of relationships do exist not only in bilingual environments, but in environments where multiple accents occur at high frequencies.

This study found that participants could understand accented speech at all levels of processing without necessarily having prior exposure to that accent. Further research may reveal factors that facilitate quicker and more successful accent comprehension.

5. ACKNOWLEDGEMENTS

This study was made possible by Bangor University and The Philological Society, whose funding enabled this MSc research to take place. The first author is currently funded for her PhD by Harding Trust, who she would also like to express gratitude to. Thank you also to Brechtje Post for her comments on the initial draft of this article.

6. REFERENCES

[1] Anwyl-Irvine, A. L., Massonié J., Flitton, A., Kirkham, N. Z., Evershed, J. K. (2019). Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods*.

[2] Bauer, L., Warren, P., Bardsley, D., Kennedy, M., & Major, G. (2007). New Zealand English. *Journal of the International Phonetic Association*, *37*(1), 97-102.

[3] Cox, F., & Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association*, *37*(3), 341-350.

[4] Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations?. *Cognition*, *108*(3), 710-718.

[5] Cox, F., & Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association*, *37*(3), 341-350.

[6] Davies, M. (2004). *British National Corpus*. Oxford University Press. Retrieved from https://www.english-corpora.org/bnc/

[7] Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*, 148.

[8] Honorof, D., McCullough, J., & Somerville, B. (2000). Comma gets a cure. Available at: https://www.dialectsarchive.com/comma-gets-a-cure

[9] International Organisation for Migration (2020). *World Migration Report 2020*. https://worldmigrationreport.iom.int/wmr-2020-interactive//

[10] Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543-562.

[11] Papachristou, V. (2011). Explicit vs. implicit pronunciation teaching to Greek children: The case of the acquisition of English vowels. *Selected papers on theoretical and applied linguistics, 19*, 371-381.

[12] Samuel, A. G., & Larraza, S. (2015). Does listening to non-native speech impair speech perception?. *Journal of Memory and Language*, *81*, 51-71.

[13] Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language*, *52*(2), 240-255.

[14] Shaw, J. A., Best, C. T., Mulak, K., Docherty, G., Evans, B., Foulkes, P., & Hay, J. 2014. Effects of short-term exposure to unfamiliar regional accents: Australians' categorization of London and York English consonants. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 72–75. University of Canterbury, New Zealand.

[15] Shaw, J., Best, C. T., Docherty, G., Evans, B. G., Foulkes, P., Hay, J., & Mulak, K. E. (2018). Resilience of English vowel perception across regional accent variation. *Laboratory Phonology*, *9*(1), 1-36.

[16] Weber, A., Di Betta, A. M., & McQueen, J. M. (2014). Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners. *Journal of Phonetics*, *46*, 34-51.

[17] Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, *97*, 30-46.

[18] Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 206-217.

Effects of Aging and Age-Related Hearing Loss on Emotion Identification in Prosodic and Semantic Channels

Yingyang Wang[1], Min Xu[2], Jing Shao[3*], Lan Wang[1], Nan Yan[1]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]Institute of Corpus Studies and Applications, Shanghai International Studies University
[3]Department of English Language and Literature, Hong Kong Baptist University
yingyangwx@gmail.com, xumin@shisu.edu.cn, jingshao@hkbu.edu.hk, lan.wang@siat.ac.cn, nan.yan@siat.ac.cn

ABSTRACT

Accurate perception of emotional status conveyed in speech signal is crucial for communication. Aging is known to have negative effect on speech perception. Age-related hearing loss (ARHL) may exert additional effects on speech processing. In this study, we investigated how these two factors influence emotion identification in prosodic and semantic channels by examining the performance of three groups of participants: younger adults (YA) and older adults (OA) with normal hearing and with hearing loss. The results revealed that OA with hearing loss demonstrated degraded performance than their normal hearing counterparts and YA in both prosodic and semantic channels. Moreover, only OA with hearing loss showed the influence of channel in that their recognition accuracy in the prosodic channel was lower. These findings suggested that ARHL significantly contributes to the inferior emotion perception in OA, and it may especially manifest itself in the conditions where fine-tuning of auditory perception is required.

Keywords: emotional perception, aging, age-related hearing loss, communication channels

1. INTRODUCTION

The processing of emotions plays a crucial role in social interactions [1, 2]. In speech communication, emotions are coded by two main channels: semantic content, which conveys the linguistic meaning behind spoken words; and prosody, which conveys paralinguistic information through acoustic cues such as pitch, intensity, and rhythm. To correctly interpret speakers' communicative intentions, listeners must decode the emotions conveyed in these channels.

Age-related decline of cognitive and social understanding, together with changes of specific neuropsychological in the brain make it more difficult for older adults (OA) to recognize basic emotions across a variety of modalities than younger adults (YA) [3, 4]. These difficulties even exist for the healthy OA with normal hearing. With respect to emotional prosody, a previous study found that subjects' performance began to decline at the age of 45, and the highest error rates occurred over the age of 65 by examining the recognition of emotions across lifespan [5]. Later studies further confirmed an age-related decrease in emotion perception accuracy [6–10]. With respect to emotional semantics, the accuracy of identifying emotions in lexical stimuli was also influenced by aging [11, 12]. Moreover, cross-channel studies revealed aging effects on channel dominance. In the perception of spoken sentences containing different prosody and semantics emotion combination matrixes, YA relied more on prosodic information , whereas OA showed a bias toward semantics [13, 14]. When asked to focus on one channel, OA had difficulty in using the prosodic cues and scored higher in the semantic channel [14]. As these studies required integrating multichannel information simultaneously, it is still unclear whether the semantic biases found in OA are intrinsic or simply result from selective attention during multichannel emotional processing. Examining OA's emotional perception in each channel (prosody vs. semantics) separately is necessary.

Aside from aging, another factor that exacerbates the difficulty of speech perception is age-related hearing loss (ARHL). OA with ARHL have a higher threshold to recognize frequency and temporal information changes [15–17], resulting in deficits in processing fine-tuned auditory cues. However, studies about emotional perception and ARHL are still scarce. Existing studies only focused on emotional prosody, suggesting that OA with hearing loss had a specific mediation effect on the auditory emotion recognition [18]. They also showed reduction in emotional valence, arousal ratings [19] and vocal emotion perception [20]. However, among the studies, none of them has investigated the perception of emotional semantics, which is one of the most important channels for expressing emotions.

As stated above, the effect of aging on emotional perception has been demonstrated in previous studies. However, it is still unclear whether ARHL has an additional impact and, if so, whether and how it manifests itself in different communication channels. In the present study, we aim to fill this research gap. First, we tested three groups of participants, i.e., YA

and OA with and without hearing loss, in the emotion identification tasks to investigate the effects of aging and ARHL on emotion processing. Second, we examined how prosodic and semantic channels affect identification accuracy and reaction time (RT). It is expected that aging and ARHL would adversely affect emotional perception, which means that OA, especially those with ARHL, perform worse than YA. Regarding communication channels, it is likely that OA, particularly for those with hearing loss, are worse at recognizing emotional prosody than semantics, due to their declined ability in processing acoustic cues.

2. METHOD

2.1. Participants

Three groups of native Mandarin speakers were recruited as participants. They all scored 26 or higher in the Montreal Cognitive Assessment-Basic test (MoCA-B, Chinese Version March 16, 2019), indicating normal cognitive abilities. All participants' hearing conditions were measured by audiometric thresholds at 6 octave frequencies (from 250 Hz to 8000 Hz) using an audiometer (GSI 18). The group of YA with normal hearing (YNH) contained 20 college and post-graduate students (10 females and 10 males) with an age range of 20–25 years old (Mean = 22.70, SD = 1.51). Their binaural hearing thresholds were below 20 dB HL at all octave frequencies. The group of OA with normal hearing (ONH) contained 18 participants (10 females and 8 males), ranging between 62 and 70 years of age (Mean = 66.39, SD = 2.53). Their binaural audiometric thresholds were lower than 20 dB HL below 4000 Hz and no more than 30 dB HL at 8000 Hz. The last group was OA with moderate or moderate-severe hearing loss (OHL), with 17 participants (9 females and 10 males) ranging in age from 62 to 77 years (Mean = 67.05, SD = 3.52). All of them have binaural hearing thresholds of more than 35 dB HL at 1000 Hz or higher, and most have more than 20 dB HL at 500 Hz or below. Figure 1 shows the audiogram of three groups.

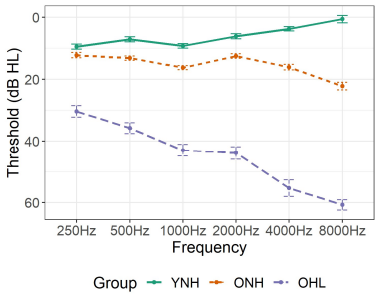


Figure 1: The mean audiometric thresholds (±1 standard error) for binaural hearing.

Each participant signed a written consent form which was approved by the Human Subjects Ethics Subcommittee at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.

2.2. Stimuli

Two channels were used to convey emotions, namely prosody and semantics. All stimuli were selected from Age-Related Differences in Affective Norms for Chinese Words (AANC) which was the first large-scale Chinese four characters word system on age-related affective norms [21]. Using 9-point scales, the AANC assessed the semantic valence of the pleasantness evoked (positive words: valence > 6; negative words: valence ≤ 4; neutral word: 4 ≤ valence < 6) and the level of familiarity with words. In the present study, we selected stimuli with familiarity levels greater than 7, ensuring that both younger and older participants were familiar with the words.

The prosodic stimuli were 16 semantically neutral concrete words (4 < valence < 6, such as "玩具手枪 /uan^{35} tɕy^{51} ʂəu^{214} tɕʰiɑŋ^{55}/." (toy pistol)) which were enunciated in happy, sad and neutral tone of voice. The semantic stimuli were semantically happy (valence > 6.5, such as "笑口常开. /ɕiɑu^{51} kʰəu^{214} tʂʰɑŋ^{35} kʰai^{55}/" (often laugh)), sad (valence < 3.5, such as "万箭穿心. /uan^{51} tɕiɛn^{51} tʂʰuan^{55} ɕin^{55}/" (in extreme grief)) and neutral (4<valence<6, such as "春去秋来. /tʂʰuən^{55} tɕʰy^{51} tɕʰiəu^{55} lai^{35}/" (time rolls on)) words. The semantic stimuli were 16 words in each emotional state and all were uttered in a neutral tone. Overall, there were a total of 96 stimuli (16 words × 3 emotions × 2 channels). In this experiment, "happy" and "sad" stimuli were targets, while "neutral" stimuli were treated as fillers.

Each stimulus was recorded three times in a sound booth by three female Mandarin speakers sampling at 44100 Hz with a 16 bits resolution. We firstly selected the most emotional and clearly evocative one from 3 repetitions of each speaker. We then evaluated the emotional intensity of the selected tokens by conducting an emotion category identification test (happy or sad), as well as the emotional intensity rating on a 7-point scale (1 = not intense, 7 = very intense), following the procedures used in previous studies [22, 23]. Participants were 8 younger native Chinese speakers (4 males and 4 females) who did not take part in the following perception experiment and are naïve to the experimental design. Only stimuli that were of higher than 88% of identification accuracy and a mean rating of greater than 4.5 were adopted. The mean accuracy of identifying emotional categories and rating emotional intensity for the target stimuli were shown in Table 3.

Channel	Emotional status	Accuracy		Intensity	
		Mean	SD	Mean	SD
Prosody	Happy	99%	0.03	5.46	0.34
	Sad	99%	0.04	5.77	0.15
Semantics	Happy	100%	0.00	5.32	0.38
	Sad	98%	0.04	5.42	0.47

Table 1: The accuracy of identifying emotional categories and rating emotional intensity for the target stimuli selected in the experiment

2.3. Procedure

We implemented the experiment in a standard laboratory soundproof room through E-prime 3.0. Auditory stimuli were presented over Sennheiser HD280 PRO headphones binaurally. In order to preserve the details of the stimuli and to ensure that every listener perceived the same volume, a 1k Hz tone of the same root-mean-square level as all stimuli was generated, and it was used to calibrated the system volume at approximately 60 dB SPL (measured by sound pressure meter (Rion NL-21)).

The tasks of two channels were presented in separate blocks. In prosodic task, 48 stimuli were repeated twice and presented randomly to the participants, resulting in 96 trails. Participants were asked to identify the emotion according to prosody of speaker's tone by pressing one of the three emotion coded keys ("v" for happy, "b" for sad and "n" for neutral) on the keyboard as quickly as possible. In semantics task, trails present in the same way and participants were asked to response according to semantic meaning of the stimuli. The order of these two tasks was counterbalanced across participants. Practices were conducted before each task to help participants familiarize with the experiment.

2.4. Data analysis

Data analyses only included trials involving happy and sad emotions, with fillers excluded. Identification accuracy and RT were analysed. To avoid the influence of ceiling effect, the identification accuracy was converted to rationalized arcsine transform units (RAU) [24]. For RTs, incorrect responses and responses over ±3 SD from the mean of each subject in each channel were excluded. And then the original data were log-transformed. Transformed accuracy and RT data were entered as dependent variables in a linear mixed effects model, in which group (YNH, ONH and OHL), channel (prosody and semantics), as well as their interactions were treated fixed effects, while subject and item were treated as random effects. Pairwise comparisons with significant interactions were conducted using Tukey's post hoc tests. The data analyses were carried out with R, using the lmeTest package [25] and the emmeans package [26].

3. RESULTS

Figure 2 illustrated the mean identification accuracy in two different communication channels. Results of linear mixed effects model suggested a significant main effect of group (χ^2 (2) = 31.938, $p < 0.001$), an approximate significant main effect of channel (χ^2 (1) = 3.068, $p = 0.080$). The interaction between group and channel was also significant (χ^2 (2) = 15.733, $p < 0.001$). Post-hoc tests indicate that OA with hearing loss identified emotional stimuli less accurately than YA and OA with normal hearing in both prosody (OHL vs YNH: $p < 0.001$; OHL vs ONH: $p < 0.001$) and semantic (OHL vs YNH: $p < 0.001$; OHL vs ONH: $p = 0.003$) channels, whereas there was no significant difference between YA and OA with normal hearing (prosodic channel: $p = 0.873$; semantic channel: $p = 0.628$). Additionally, in the group of OA with hearing loss, the identification accuracy of the prosodic channel was significantly lower than that of the semantic channel ($p < 0.001$).

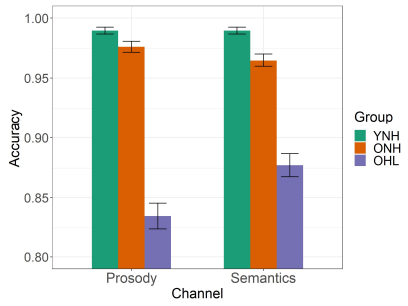


Figure 2: Mean identification accuracy (±1 standard error) across two channels of three groups.

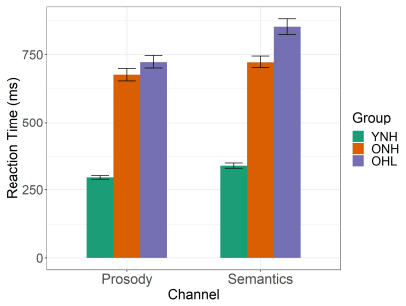


Figure 3: Mean reaction time (±1 standard error) across two channels of three groups.

The RT of the three groups with respect to prosodic and semantics channels was illustrated in Figure 3. Linear mixed-effects analyses revealed a significant main effect of group (χ^2 (2) = 29.766, $p < 0.001$), an

approximate significant main effect of channel (χ^2 (1) = 3.158, p = 0.076), but no significant interaction between two fixed effects (χ^2 (2) = 2.610, p = 0.271). In general, older adults, both those with and without hearing loss, employed longer RT than younger adults (ONH vs YNH: β = 0.250, SE = 0.060, t = 3.930, p < 0.001; OHL vs YNH: β = 0.310, SE = 0.063, t = 4.717, p < 0.001), suggesting greater cognitive efforts in performing the task.

4. DISCUSSION

The present study explored how aging and ARHL influence emotion perception in two channels. The findings showed that OA with hearing loss scored lower than the other two groups. Moreover, their accuracy in prosodic condition was lower than semantic condition, demonstrating the influence of channels. In addition, OA both with and without hearing loss responded slower than YA.

For OA with hearing loss, the most important finding was that their accuracy in recognizing emotional stimuli was lower than both YA and OA with normal hearing, suggesting that hearing ability played a crucial role in emotional speech perception and prosody comprehension. It has been found that there was degeneration of neural coding in the central auditory nervous system in OA with hearing loss [27, 28], which may influence their emotion speech processing. Additionally, previous study also found that when processing emotion information, hearing loss led to more engagement of the prefrontal cortex and less involvement of the amygdala, which may also have resulted in deficient emotion recognition in this population [29].

Another important finding was that OA with hearing loss showed channel effect in that their recognition accuracy in the prosodic channel was lower. A possible explanation was that the comprehension of speech involves a combination of sensory information from external sound and our prior knowledge in the brain. On the one hand, emotional prosody was modulated by detailed changes in acoustic cues. Mandarin speakers encoded happy prosody by employing higher F0, stronger intensity, shorter duration and higher first formant than sad prosody [30, 31]. OA with hearing loss had difficulties in decoding cues in speech signal [19, 20], which would result in deficiency in processing the fine-tuning information contained in the prosodic channel. On the other hand, human brain automatically utilizes prior knowledge to facilitate speech comprehension, especially in adverse hearing environments [32], and OA might rely more heavily on these knowledge than YA [33]. In our study, the stimuli used in the semantic condition were idioms which contain rich semantic information to assist the listeners process the emotion status. OA with hearing loss might be able to use their prior to identify the emotion. In contrast, such information was absent in prosodic condition where the stimuli were semantically neutral. Consequently, they were able to perform better in the semantic channel.

We also found that OA with normal hearing achieved comparable identification accuracy compared with YA, which is inconsistent from previous studies. This might be caused by the different task complexity. Previous studies usually included three or more emotions as targets [5, 7–12], while our study only investigated two (happy vs. sad). The lower task demand may make the group difference (YA vs. OA with normal hearing) less obvious in terms of accuracy. However, it is worth noting that both OA with and without hearing loss took longer RTs than YA. RT is a reliable measurement of listening effort [34]. The longer RT suggests that OA might require more cognitive resources to process speech and connect it to specific emotions compared with YA.

Future studies should increase complexity to reveal the possible group difference between OA with normal hearing and with hearing loss. This might include increasing the numbers of target emotions and investigating the emotion processing across multiple channels. Moreover, the association between cognitive ability and emotion speech perception is also a worthwhile direction for future research.

5. CONCLUSION

In the current study, we found that ARHL significantly affected emotion perception. OA with hearing loss performed worse than their normally hearing peers and YA. Furthermore, they showed channel influence in that they recognized emotion better in semantic channel. These findings broadened our understanding of emotion recognition in OA with hearing loss. It also provided empirical data to better understand the underlying mechanism of the deficient emotional processing in OA with ARHL.

6. ACKNOWLEDGEMENTS

This work was jointly supported by grants from National Key R&D Program of China (2020 YFC2004100), National Natural Science Foundation of China (NSFC 11904381, NSFC 81371900, NSFC 62271477, NSFC U1736202), Shenzhen KQTD Project (KQTD 20200820113106007), Shenzhen Science and Technology Program, Grant Award (JCYJ 20210324115810030) and the Start-up Grant from Hong Kong Baptist University (162646).

7. REFERENCES

[1] Davidson, R. J., Scherer, K. R. & Goldsmith, H. H. 2003. Handbook of Affective Sciences, *Adolescence*, 38, 390–391.

[2] Fischer, A. H.,& Manstead, A. S. R. 2008. Social functions of emotion. *Handb. Emot.* 3, 456–468.

[3] Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. 2008. A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neurosci. Biobehav. Rev.* 32, 863–881.

[4] Sullivan, S., & Ruffman, T. 2004. Social understanding: How does it fare with advancing years? *Br. J. Psychol.* 95, 1–18.

[5] Brosgole, L., & Weisman, J. 1995. Mood recognition across the ages. *Int. J. Neurosci.* 82, 169–189.

[6] Mitchell, R. L. C. 2007. Age-related decline in the ability to decode emotional prosody: Primary or secondary phenomenon? *Cogn. Emot.* 21, 1435–1454.

[7] Sen, A., Isaacowitz, D. & Schirmer, A. 2018. Age differences in vocal emotion perception: on the role of speaker age and listener sex. *Cogn. Emot.* 32, 1189–1204.

[8] Martzoukou, M., Nasios, G., Kosmidis, M. H. & Papadopoulou, D. 2022. Aging and the Perception of Affective and Linguistic Prosody. *J. Psycholinguist. Res.*.

[9] Dupuis, K., & Pichora-Fuller, M. K. 2015. Aging affects identification of vocal emotions in semantically neutral sentences. *J. Speech, Lang. Hear. Res.* 58, 1061–1076.

[10] Mitchell, R. L. C., Kingston, R. A. & Barbosa Bouças, S. L. 2011. The Specificity of Age-Related Decline in Interpretation of Emotion Cues From Prosody. *Psychol. Aging*. 26, 406–414.

[11] Grunwald, I. S. *et al.* 1999. The effects of age and gender on the perception of lexical emotion. *Appl. Neuropsychol.* 6, 226–238.

[12] Isaacowitz, D. M. *et al.* 2007. Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychol. Aging*. 22, 147–159.

[13] Ben-David, B. M., Gal-Rosenblum, S., van Lieshout, P. H. H. M. & Shakuf, V. 2019. Age-related differences in the perception of emotion in spoken language: The relative roles of prosody and semantics. *J. Speech, Lang. Hear. Res.* 62, 1188–1202.

[14] Kim, D., Sim, H. S. & Lee, Y. 2022. Age-Related Differences in the Perception of Emotion in Emotional Speech: The Effects of Semantics and Prosody. *Audiol. Speech Res.* 18, 48–59.

[15] Patterson, R. D., & Weber, L. 1982. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *J. Acoust. Soc. Am.* 76, 1788-1803.

[16] Strouse, A., Ashmead, D. H., Ohde, R. N. & Grantham, D. W. 1998. Temporal processing in the aging auditory system. *J. Acoust. Soc. Am.* 104, 2385–2399.

[17] Xu, M., Shao, J. & Wang, L. 2021. Effects of aging and age-related hearing loss on talker discrimination. *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*. 2917–2921.

[18] Lena, L., Kreifelts, B. & Wildgruber, D. 2012. Age-related decrease in recognition of emotional facial and prosodic expressions. *Emotion*. 12, 529–539.

[19] Picou, E. M. 2016. How hearing loss and age affect emotional responses to nonspeech sounds. *J. Speech, Lang. Hear. Res.* 59, 1233–1246.

[20] Christensen, J. A., Sis, J.,Kulkarni, A. M., Chatterjee, M., Prostheses, A. & National, B. T. 2020. Effects of age and hearing loss on the recognition of emotions in speech. *Ear and hearing.* 40, 1069–1083.

[21] Liu, P., Lu, Q., Zhang, Z., Tang, J. & Han, B. 2021. Age-Related Differences in Affective Norms for Chinese Words (AANC). *Front. Psychol.* 12. 1–17.

[22] Lin, Y., Ding, H. & Zhangb, Y. 2020. Prosody Dominates Over Semantics in Emotion Word Processing: Evidence From Cross-Channel and Cross-Modal Stroop Effects. 63, 896–912.

[23] Lin, Y., & Ding, H. 2020. Effects of communication channels and actor ' s gender on emotion identification by native Mandarin speakers. *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*. 3151–3155.

[24] Studebaker, G. A., 1985. A "rationalized" transform arcsine. *J. Speech, Lang. Hear. Res.* 28, 455–462.

[25] Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. 2017. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26.

[26] Lenth, R.,Singmann, H.,Love, J.,Buerkner, P. & Herve, M. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R Packag. version*. 1.

[27] Frisina, R. D. 2001. Possible neurochemical and neuroanatomical bases of age-related hearing loss - Presbycusis. *Semin. Hear.* 22, 213–225.

[28] Tremblay, K. L., Piskosz, M. & Souza, P. 2003. Effects of age and age-related hearing loss on the neural representation of speech cues. *Clin. Neurophysiol.* 114, 1332–1343.

[29] Husain, F. T., Carpenter-Thompson, J. R. & Schmidt, S. A. 2014. The effect of mild-to-moderate hearing loss on auditory and emotion processing networks. *Front. Syst. Neurosci.* 8, 1–13.

[30] Zhang, S., Ching, P. C. & Kong, F. 2006. Acoustic analysis of emotional speech in Mandarin Chinese, in *International symposium on chinese spoken language processing*. 57–66.

[31] Yuan, J., Shen, L. & Chen, F. 2002. The acoustic realization of anger, fear, joy and sadness in Chinese. *7th Int. Conf. Spok. Lang. Process. ICSLP*. 2025–2028.

[32] Davis, M. H., & Johnsrude, I. S. 2007. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hear. Res.* 229, 132–147.

[33] Lesicko, A. M. H., & Llano, D. A. 2017. Impact of peripheral hearing loss on top-down auditory processing. *Hear. Res.* 343, 4–13.

[34] Pals, C., Sarampalis, A., van Rijn, H. & Başkent, D. 2015. Validation of a simple response-time measure of listening effort. *J. Acoust. Soc. Am.* 138, EL187–EL192.

PERCEPTION, RECOGNITION, AND ENCODING OF CANTONESE SOUND CHANGE VARIANTS

Rachel Soo and Molly Babel

Department of Linguistics, The University of British Columbia, Canada
rachel.soo@ubc.ca, molly.babel@ubc.ca

ABSTRACT

In Cantonese, a sound change sees /n/ produced as [l] syllable-initially, creating [n]- and [l]-initial pronunciation variants for underlying /n/-initial words. We examine the perception, recognition, and encoding of these sound change variants through four online experiments utilizing Bayesian frameworks for equivalence interpretations. An immediate repetition priming paradigm with [l] targets (Exp. 1) demonstrates recognition equivalence between [n] and [l] forms, in spite of clear phonetic sensitivity to [n] and [l] in AX discrimination (Exp. 2a) and categorization tasks (Exp. 2b). A long distance repetition priming task (Exp. 3) suggests listeners encode both pronunciation variants by dually mapping them to a single lexical representation. Taken together, these findings contribute to our understanding of the [n]-[l] sound change and uniquely situates the study of phonetic variation, which has traditionally been studied through the lens of within-/cross-dialect pronunciation variants, in the context of diachronic sound change variants.

Keywords: Psycholinguistics, Speech perception, Social encoding, Cantonese, Sound change

1. INTRODUCTION

Listeners encounter phonetic variation in spoken language. Regular contextual exposure to pronunciation variants may provide listeners with the perceptual flexibility to recognize and encode multiple pronunciation variants [1, 2]. For example, [2] posit that Basque listeners are regularly exposed to Spanish-accented Basque, which allows for the dual-mapping of a "correct" laminar [t͡s̻] and "incorrect" palato-alveolar [t̠ʃ] pronunciation variant to a single Basque lexical item. Encoding strength may also vary as a function of variant typicality, or social weight, as canonical and socially prestigious dialect forms may be preferentially encoded [3, 4, 5, 6].

Outside of regional/foreign accent variation, pronunciation variants are also apparent in diachronic sound changes and this dual-mapping can also be realized for changes-in-progress. Much like within-/cross-dialect variation [1, 2, 3], particular pronunciation variants in a sound change may be specific to a demographic group [7]. Pronunciation variants may also be socially meaningful just as particular regional accents may be socially stigmatized or recognized as the standard [8]. This may be the case with a Cantonese sound change, where /n/ categorically becomes [l] syllable-initially [9, 10]. Words such as *nou5* 腦 "brain", historically pronounced with an initial /n/ (hereafter, historical variant) are pronounced with [l] (hereafter, innovative variant), making it homophonous with /l/-initial words such as *lou5* 老 "old". [n] pronunciations are considered "standard", associated with prestige and used in formal contexts [11], while [l] pronunciations are socially stigmatized. This sound change was initially observed in the first half of the 20th century [12], with early sociolinguistic work indicating that younger speakers were leading the sound change [11, 13, 14]. More recent phonetic work reveals that the sound change may be complete at the community-level [10, 15], with some hypercorrection to [n] pronunciations. We build on this work by examining the consequences of this sound change on the perception, recognition and encoding of [n] and [l].

2. GENERAL PROCEDURE

All experiments were delivered online in Gorilla [16]. Participants provided consent, completed a short listening task to establish that they were wearing adequate headphones [17], and listened to a short Cantonese story to familiarize them to the talker's voice. Then, they proceeded to the main task in each experiment. 39 Cantonese talkers completed Exp. 1; 16 were counterbalanced into Exp. 2a, and 23 into Exp. 2b. A separate group of 35 Cantonese talkers completed Exp. 3. All participants were recruited from student participant pools at Canadian and Hong Kong

institutions, ranged in age from 18-35, acquired Cantonese before age 5, and spent > 10 years in a Cantonese-speaking family. This linguistic profile conspires to produce lifelong exposure to Cantonese, and given that the sound change has been around since the mid-20th century, regular exposure to [n] and [l] pronunciations as well. Participants completed the Bilingual Language Profile (BLP) [18], which quantifies relative dominance in English and Cantonese on a scale of +/- 218, respectively. Participants had a BLP dominance score below 50.

3. EXPERIMENT 1

Exp. 1 was an immediate repetition priming task. On each trial, participants were presented with two auditory items, a prime and a target, separated by a 500 ms inter-stimulus interval (ISI) and asked to make a lexical decision to the target. If the pronunciation affects word recognition, we predict that both [n]- and [l]-initial primes should facilitate the recognition of [l]-initial targets equivalently.

3.1. Materials

40 [l]-initial target words were used. Each [l]-initial target word (e.g., *lou5* 老 "old") was preceded by one of four different primes, representing different pair types. In identity pairs, the prime and target were identical words (though physically-distinct). In historical pairs, the prime was the historical and less frequent [n]-initial variant of the target (*nou5* 腦 "brain"). Critical pairs were matched with two control types. In rime controls, the prime and target shared rimes and differed in their initial consonants (*pou5* 抱 "embrace"). Unlike the historical trials, the prime did not share a consonant that is in a sound change with [l]. In unrelated controls, the prime and the target did not share any features (*caa4* 茶 "tea").

80 non-[l]-initial filler target words that were likewise preceded by different primes, representing different pair types, were used. This produced a 1:2 ratio of critical to filler items in the experiment, and a total of 120 real word targets. To balance the number of real word and nonword targets, 120 nonwords targets were created from gaps in Cantonese. Each of these prime-target pairs were counterbalanced across 4 lists. Each list contained a total of 240 trials, in which there were 120 real word targets (40 critical pairs and 80 filler pairs), and 120 nonword targets.

3.2. Analysis and results

Data wrangling for all experiments used the {tidyverse} package [19] in R [20]. For all

Table 1: Sample stimuli for Exp. 1 and 3.

Exp. 1	Pair	Prime	Target
Critical	Identity	*lou5* 老 "old"	*lou5* 老 "old"
	Historical	*nou5* 腦 "brain"	
Control	Rime	*pou5* 抱 "embrace"	
	Unrelated	*caa4* 茶 "tea"	
Exp. 3	**Pair**	**Block 1**	**Block 2**
Critical	Identity	*lou5* 老 "old"	*lou5* 老 "old"
	Historical	*naam4* 男 "male"	*laam4* 藍 "blue"
Control	Unmatched	–	*laai1* 拉 "pull"

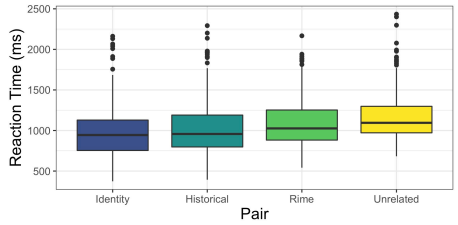


Figure 1: Reaction time (ms) to targets in critical and control pairs in Exp. 1.

experiments, reaction times < 200 ms or > 3000 ms were removed (1% of the responses), as well as reaction times < or > 2.5 standard deviations from each subject's mean (3% of the responses). Reaction times for correct responses to critical and control pairs are shown in Fig. 1.

The reaction times to correct responses were analyzed using a Bayesian mixed-effects model fitted in Stan using the {brms} package [21] in R [20]. The model included Pair (Historical, Identity, Rime, Unrelated) as a dummy-coded fixed effect. Historical pairs were used as the reference level to allow for the comparison of Historical pairs ([n]-initial primes) with Identity pairs ([l]-initial primes) and Rime pairs (primes with shared rimes). The model also included by-item random intercepts and by-subject random slopes for the effect of Pair. A lognormal response distribution was assumed for the reaction times. Regularizing, weakly informative priors were selected for all parameters. Hamiltonian Monte-Carlo sampling with four chains (each with 2000 iterations, 1000 warm-up) were used to draw samples from the posterior distribution. We report the mean of the posterior distribution, the 95% credible interval (CrI), the proportion of the CrI that falls within a region of practical equivalence (ROPE) of ±0.03, and the probability of direction (PD). The mean of the posterior distribution is a point estimate; the average value of the distribution. In general, if the 95% CrI does not encompasses 0, this is taken as evidence for a meaningful effect, while a 95% CrI that encompasses 0 is taken as evidence for a negligible effect. The ROPE specifies a region around 0 that is practically equivalent to 0 (in this case, ± 0.03) and thus, provides a measure of the probability that the credible effect values are non-negligible. A meaningful effect size

should have a smaller proportion of 95% CrI within the ROPE, while negligible effects should have a larger proportion of the 95% CrI within the ROPE. Finally, the PD provides a measure of how probable the specific direction (positive or negative) of the effect is, with strong evidence for a non-null effect represented by a PD greater than 95% [22].

The model revealed weak evidence for a difference and strong evidence for equivalence between historical and identity pairs (Identity pairs: β =-0.03, CrI = [-0.07, 0.00], % ROPE = 38.60%, PD = 96.66%). While there is slightly greater priming for identity pairs, as indicated by the negative β value, the upper limit of the 95% CrI is 0, and 38.60% of the posterior distribution falls within the ROPE. In contrast, there is strong evidence for a difference and weak evidence for equivalence between historical and rime pairs, and between historical and unrelated pairs. For both comparisons, the 95% CrI does not encompass 0, 0% of the posterior distribution falls within the ROPE, and the PD is greater than 95% (Rime pairs: β =0.07, CrI = [0.03, 0.11], % ROPE = 0%, PD = 99.98%; Unrelated pairs β =0.14, CrI = [0.10, 0.18], % ROPE = 0%, PD = 100%).

3.3. Interim discussion

Exp. 1 showed that [n]-initial primes (in historical pairs) and [l]-initial primes (in identity pairs) equivalently facilitated the recognition of [l]-initial targets. This suggests that the sound change has rendered [n]- and [l]-initial items as recognition equivalents. There was also strong evidence for a difference between the historical and rime pairs, suggesting that the priming observed for [n]-initial primes in historical pairs to [l]-initial targets is unlikely to be a product of shared rimes between the prime and target. These data may be accounted for by (at least) two explanations. First, listeners may simply be poor at distinguishing [n] and [l] at a phonetic level. Alternatively, listeners may be able to perceive the difference between [n] and [l] but map the two pronunciation variants to a single lexical representation (e.g., *lou5* and *nou5* both map to *lou5* 老 "old") [2]. Exps. 2a and 2b were carried out to test the phonetic-level explanation.

4. EXPERIMENT 2A AND 2B

In Exp. 2a, AX discrimination task participants were presented with two items separated by a 500 ms ISI, and asked to make a same-different judgement. In Exp. 2b, categorization task participants were presented with an individual token and asked to categorize the initial consonant as "N" or "L".

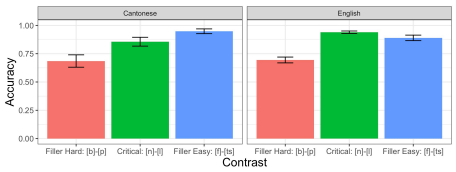


Figure 2: Discrimination accuracy for each nonword pair contrast in Exp. 2a for Cantonese and English listeners.

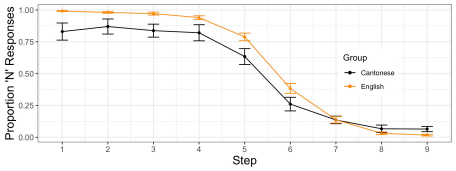


Figure 3: Proportion of 'N' responses in Exp. 2b for Cantonese and English listeners.

4.1. Materials

AX discrimination task items were three nonword pairs each representing different contrasts: *noe6-loe6* (critical), *foe6-tsoe6* (easy), and *boe6-poe6* (hard). The natural *noe6* and *loe6* tokens were used as end points from which nine-step continua were synthesized in TANDEM-STRAIGHT [23]. The individual tokens from this continuum were the items for the categorization task.

4.2. Analysis and results

Due to space, we do not detail the Bayesian analyses and instead provide figures visualizing the discrimination accuracy for pairs in Exp. 2a (see Fig. 2) and the proportion of 'N' responses in Exp. 2b (see Fig. 3). Responses from English control listeners are provided for comparison.

In the AX discrimination task, Cantonese talkers demonstrate high discrimination sensitivity to [n] and [l] in the critical *noe6-loe6* pair; this is on par with an acoustically clear contrast in *foe6-tsoe6*. Likewise, the sigmoidal response function in the categorization task demonstrates clear categorical perception of [n] and [l]. Together, these results demonstrate that listeners can distinguish [n] from [l] at a phonetic level. We turn now to testing the second, dual-mapping explanation in Exp. 3.

5. EXPERIMENT 3

This was a long distance repetition priming task. Items are presented across two blocks. On each trial, participants made a lexical decision to an individual auditory token. Block 1 contained the primes from

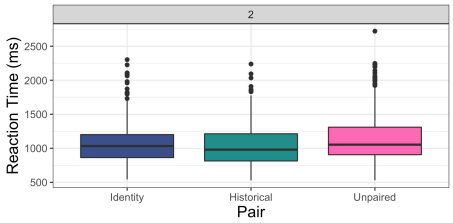


Figure 4: Reaction time (ms) to targets in critical and control pairs in block 2 of Exp. 3.

the pairs in Exp. 1 (e.g., *nou5* 腦 "brain") and block 2 had the corresponding targets (*lou5* 老 "old"). If [n]- and [l]-initial pronunciations are separately encoded variants, priming is not expected, as minimal pairs are known to not show priming across long distances. However, if both pronunciations are fully encoded as acceptable variants of a single lexical item, long distance priming is predicted.

5.1. Materials

Exp. 3 used a subset of stimuli from Exp. 1 (see Table 1). Block 1 contained the primes from 10 critical identity pairs (*lou5* 老 "old") and the primes from 10 critical historical pairs (*naam4* 男 "male"). Block 2 had the 10 corresponding targets of the critical identity pairs (*lou5* 老 "old") and the 10 corresponding targets of the critical historical pairs (*laam4* 藍 "blue"). 20 additional [l]-initial words with no prime in the preceding block were included as unmatched controls (*laai1* 拉 "pull"). Each block contained 160 real word fillers and 240 nonwords.

5.2. Analysis and results

Reaction time outlier removal was the same as in Exp. 1 (3% of the responses removed). These data are shown in Fig. 4. The reaction times to correct responses in block 2 were analyzed using a Bayesian mixed-effects model fitted in Stan using the {brms} package [21] in R [20]. The model included Pair (Historical, Identity, Unmatched), with Historical pairs as the reference level. The model also included by-item random intercepts and by-subject random slopes for the effect of Pair. A lognormal response distribution was assumed for the reaction times. Regularizing, weakly informative priors were selected for all parameters. Hamiltonian Monte-Carlo sampling with four chains (each with 2000 iterations, 1000 warm-up) were used to draw samples from the posterior distribution.

We observed weak evidence for a difference and strong evidence for equivalence between identity and historical pairs. While there was slightly more priming by [n]-initial primes in Historical pairs, indicated by the positive β estimate, the 95% CrI encompasses 0, and 23.27% of the posterior distribution falls within the ROPE (Identity pairs: β =0.06, CrI = [-0.02, 0.14], % ROPE = 23.27%, PD = 91.87%). However, there was strong evidence for a difference and weak evidence for equivalence between unmatched and historical pairs. The 95% CrI does not include 0, only 3.03% of the posterior distribution falls in the ROPE, and the PD is greater than 95% (Unmatched pairs: β =0.09, CrI = [0.01, 0.17], % ROPE = 3.03%, PD = 98.9%).

These results align with and extend those of Exp. 1; historical [n]-initial primes (in historical pairs) and innovative [l]-initial primes (in identity pairs) equivalently facilitate the recognition of [l]-initial targets not just at short time lags (Exp. 1), but also across a more extended interval of time (Exp. 3).

6. DISCUSSION AND CONCLUSION

In Exp. 1, historical [n]-initial primes and innovative [l]-initial primes equivalently facilitated the recognition of [l]-initial targets. Exps. 2a and 2b ruled out the possibility that these results were due to poor discriminability of [n] and [l] at a phonetic level. Listeners showed high discrimination sensitivity to [n] and [l] (on par with an acoustically clear contrast in Cantonese) in Exp. 2a and had crisp sigmoidal categorization functions, suggesting [n] and [l] are distinct perceptual categories. Having ruled out this phonetic explanation, Exp. 3 tested a lexical explanation, wherein both [n]- and [l]-initial pronunciations were dually mapped to a single lexical representation. We observed the same recognition equivalence between [n]- and [l]-initial primes even across a more extended interval of time in a long distance repetition priming task. As minimal pairs do not show priming in this paradigm, these results suggest that [n] and [l] are dually-mapped as acceptable pronunciation variants for a single lexical representation.

As mentioned, Cantonese talkers have metalinguistic awareness about these socially meaningful phonetic forms. As such, this dual-mapping may be a consequence of regular exposure to the pronunciation variants in the community [2]. Additionally, while the sound change has been described as complete [10], these findings also point toward enregistered, context-conditioned variation between [n] and [l] [24]. In this respect, a dual-mapping structure is an ideal representation, as it affords the ability to process both [n] and [l] pronunciation variants in the appropriate social contexts [10].

7. ACKNOWLEDGEMENTS

We thank Fion Fung and members of the Speech in Context Laboratory for technical assistance on this project. We also thank Ashley Farris-Trimble, Yao Yao, and Phil Monahan for assistance with participant recruitment. This work was funded by a SSHRC Insight Grant awarded to Molly Babel.

8. REFERENCES

[1] M. Sumner and A. G. Samuel, "The effect of experience on the perception and representation of dialect variants," *Journal of Memory and Language*, vol. 60, no. 4, pp. 487–501, 2009.

[2] A. G. Samuel and S. Larraza, "Does listening to non-native speech impair speech perception?" *Journal of Memory and Language*, vol. 81, pp. 51–71, 2015.

[3] M. Sumner and A. G. Samuel, "Perception and representation of regular variation: The case of final/t," *Journal of Memory and Language*, vol. 52, no. 3, pp. 322–338, 2005.

[4] M. Sumner and R. Kataoka, "Effects of phonetically-cued talker variation on semantic encoding," *The Journal of the Acoustical Society of America*, vol. 134, no. 6, pp. EL485–EL491, 2013.

[5] Z. Jones and C. G. Clopper, "Subphonemic variation and lexical processing: Social and stylistic factors," *Phonetica*, vol. 76, no. 2-3, pp. 163–178, 2019.

[6] M. Sumner, S. K. Kim, E. King, and K. B. McGowan, "The socially weighted encoding of spoken words: A dual-route approach to speech perception," *Frontiers in Psychology*, vol. 4, p. 1015, 2014.

[7] M. Rae, P. Warren *et al.*, "Goldilocks and the three beers: sound merger and word recognition in NZE," *New Zealand English Journal*, vol. 16, p. 33, 2002.

[8] M. Sumner, "A phonetic explanation of pronunciation variant effects," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. EL26–EL32, 2013.

[9] S.-l. Wong, "A Chinese syllabary pronounced according to the dialect of Cantonese," *Canton: Lingnam University*, 1941.

[10] L. S. Cheng, M. Babel, and Y. Yao, "Production and perception across three Hong Kong Cantonese consonant mergers: Community-and individual-level perspectives," *Laboratory Phonology*, vol. 13, no. 1, 2022.

[11] P. G. Pan, "Prestige forms and phonological variation in Hong Kong Cantonese speech." Master's Thesis, The University of Hong Kong, 1981.

[12] J. D. Ball, *Cantonese Made Easy: A book of simple sentences in the Cantonese dialect, with free and literal translations, and directions for the rendering of English grammatical forms in Chinese*. Kelly & Walsh, Limited, 1907.

[13] H. S.-W. Yeung, "Some aspects of phonological variations in the Cantonese spoken in Hong Kong," Master's Thesis, The University of Hong Kong, 1980.

[14] D. S. Bourgerie, "A quantitative study of sociolinguistic variation in Cantonese," Doctoral Dissertation, The Ohio State University, 1990.

[15] L. Cheng, "Mergers-in-Progress in Hong Kong and Vancouver Cantonese-English Bilinguals," Bachelor's Thesis, University of British Columbia, 2017.

[16] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior research methods*, vol. 52, no. 1, pp. 388–407, 2020.

[17] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[18] D. Birdsong, "Dominance and Age in Bilingualism," *Applied Linguistics*, vol. 35, no. 4, pp. 374–392, 2014, publisher: Oxford University Press. [Online]. Available: https://journals.scholarsportal.info/details/01426001/v35i0004/374_daaib.xml

[19] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[21] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of statistical software*, vol. 80, no. 1, pp. 1–28, 2017.

[22] B. Nicenboim and S. Vasishth, "Statistical methods for linguistic research: Foundational ideas—part ii," *Language and Linguistics Compass*, vol. 10, no. 11, pp. 591–613, 2016.

[23] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3933–3936.

[24] A. Agha, "Voice, footing, enregisterment," *Journal of Linguistic Anthropology*, vol. 15, no. 1, pp. 38–59, 2005.

PHONETIC CORRELATES OF GENDER IN PREPUBERTAL VOICES

Riccarda Funk[1], Melanie Weirich[2], Adrian P. Simpson[3]

[123]Friedrich Schiller University Jena, Germany
[1]riccarda.funk@uni-jena.de, [2]melanie.weirich@uni-jena.de, [3]adrian.simpson@uni-jena.de

ABSTRACT

This study investigates acoustic and perceptual correlates of gender in prepubertal voices. 60 German-speaking pupils (29 girls, 33 boys, aged 6-8 years) were recorded in first and again in second grade. A listening experiment was conducted to find out how female or male the voices are perceived to be. Stimuli of the five most female- and male-sounding children were played in a second experiment to two groups of listeners who judged each stimulus against six perceptual attribute pairs. Linear-mixed-models and Spearman's ρ correlations were run to evaluate the relationship between perception and acoustics.

There are significant acoustic differences between both genders and the most female- and male-sounding children. Significant correlations between perception ratings and acoustic parameters are reported. Furthermore, linear-mixed-models verify significant relationships between perception and acoustics within and across gender, indicating that certain phonetic parameters of the voice and speech are used by children to construct gender.

Keywords: Prepubertal voice, gender perception, acoustics, correlates of gender

1. INTRODUCTION

Many phonetic correlates of gender in the adult voice are due to anatomical differences in the larynx and the vocal tract. In prepubertal voices, these differences are marginal [1, 2]. Nevertheless, many studies have found acoustic differences between prepubertal boys and girls as young as two and a half years of age [3]. However, these differences are rarely systematic. Indeed, in some cases the results contradict each other:

- **f0:** girls > boys [4, 5], girls < boys [6], girls = boys [7]
- **F1:** girls > boys [8, 9, 10, 11]
- **HNR:** girls > boys [12, 13, 14, 15]
- **Tempo:** girls > boys [7, 15], girls < boys [7, 16]
- **/s/:** girls > boys in center of gravity (CoG) and girls < boys in skewness [17, 18, 19]

Despite such inconsistencies, other studies have shown that listeners are able to determine gender in children's voices at above-chance levels [14, 15, 20, 21, 22]. In most cases, the listeners' judgements coincide with biological sex. More interesting, however, are gender-identification rates of individual children. For some children identification rates are at chance level, whereas others receive a very clear gender assignment [14, 15]. The latter seem to exhibit a set of robust acoustic cues encoding gender that listeners overwhelmingly agree upon. When listeners rate stimuli of such children against a number of perceptual attribute pairs, e.g. fast–slow, high–low, strong correlations between perception ratings and acoustic measurements can be found [14].

The present study concentrates on identifying the acoustic and perceptual characteristics of children's voices which lead to a relatively unambiguous gender attribution by adult listeners. How do these voices differ from those with acoustic patterns that lead to more ambivalent, variable gender responses? Which acoustic parameters can be used to predict the perception of children's voices? Can we find relationships between perceptual attribute ratings and the corresponding acoustic parameters?

The study is part of a longitudinal project analysing recordings of ca. 60 German primary school children from the first to fourth grade (6-10 years). In this article, we focus on the results of the first and second graders.

2. METHOD

2.1 Listeners

Acoustic voice recordings were made of 62 German-speaking first-graders (29 girls, 33 boys) aged 6–7 years old ($\bar{x}$ = *6.16*). The recordings took place in September 2020 at two primary schools in neighbouring East German villages near Gotha. The same children were recorded again in October 2021. In this case, 60 children participated (28 girls, 32 boys), aged 7–8 years old ($\bar{x}$ = *7.18*). Two of the children could no longer participate due to relocation. In this study, we only analysed the data of the 60 children common to both recordings.

2.2 Recordings

All recordings were made using a USB microphone with integrated audio interface (Røde NT-USB) located in a portable sound wall (Marantz Pro Sound Shield) using *audacity* [23]. The microphone was

positioned 20-30 cm from the speaker's mouth. The children were recorded individually in a quiet room at the school. Both spontaneous and read (copied) speech was elicited. Children were asked to repeat versions of ten simple prose sentences pre-recorded by an adult female speaker (first author). We recognise that children copying pre-recorded utterances may accommodate to aspects of the adult speaker's voice. However, given the level of (il)literacy in first graders, this was the only way to elicit tightly controlled sentence-length utterances. In this study, we focus on the recordings of the first two sentences:

1. *Im Sommer blühen die Blumen.*
   (Flowers blossom in summer.)
2. *Kannst du die Rose riechen*?
   (Can you smell the rose?)

2.3 Acoustic analysis

The stimuli of the children were segmented and annotated using *Web-MAUS* [24]. The *praat* [25] text grids were manually checked and corrected. All acoustic analyses were also performed in *praat* using different scripts.

Mean fundamental frequency was measured in both sentences and semitone range was calculated with the formula

(1) $st = \frac{12}{ln(2)} ln\left(\frac{fmax}{fmin}\right)$.

Mean formant values were estimated in the vowels /ɔ/, /y:/ and /u:/ of sentence 1 and /a/, /o:/ and /i:/ of sentence 2 (time step 0.01; window length 0.025 s; maximum formant 5500 Hz). A two-dimensional acoustic vowel space (AVS) was created using the F1 and F2 of these vowels. The area of this polygon was calculated and for legibility and ease of comparison will be stated in kHz². The same vowels were used for determining voice quality (HNR). Tempo was calculated in canonical syllables per second (without pauses). center of gravity (CoG) and skewness were calculated from the sibilant spectra of /z/ in "Sommer" (sentence 1) and "Rose" (sentence 2) and /s/ in "kannst (sentence 2).

2.4 Listening experiment 1 – gender identification

The aim of the first listening experiment was to find out how female or male the voices of the children are perceived to be, i.e. which children produce robust phonetic correlates of gender. For this, listeners heard the recordings of sentences 1 and 2. The stimuli were normalised to 70 dB.

The listening experiment was conducted online using *Percy* [26] – one experiment for the voices of the first graders and another one for the second-graders. Stimuli were presented in random order. Each stimulus could be listened to twice. Listeners decided on a seven-point scale how boy-like or girl-like a stimulus was. The duration of the experiment for each subject was shortened by dividing the stimuli into three approximately equally sized groups.

167 listeners participated in the experiment with first-graders (119 female, 46 male, 2 diverse; age: 16–71 years, $\bar{x} = 33.2$, $\sigma = 13.7$). 117 different subjects listened to the voices of the second-graders (94 female, 21 male, 2 diverse; age: 16-70 years, $\bar{x} = 33.2$, $\sigma = 13.4$). None of the listeners reported any hearing problems and the mother tongue of all was German.

2.5 Listening experiment 2 – perceptual attribute rating

The second listening experiment explored possible relationships between perceptual attribute ratings and the corresponding acoustic parameters. Stimuli of the five girls and five boys with the highest and lowest perception ratings (see 3.1) were used. Listeners judged perceptual attribute pairs for sentences 1 (recording from first grade) and 2 (recording from second grade) of each child on a seven-point-scale: fast–slow, high–low, clear–mumbled, hoarse–clear, loud–quiet, monotonous–melodious. The stimuli were normalised to 70 dB using *praat* [25] and the experiment was conducted using *SoSci survey* [27].

To avoid a possible influence of gender stereotypes on the perception ratings, the listeners were split into two approximately equally sized groups. As in [14], both groups heard the same stimuli, but the first group was told they were listening to boys, the second that they were listening to girls. 102 listeners took part in this listening experiment (80 female, 21 male, 1 diverse; 16–77; $\bar{x} = 34.8$, $\sigma = 12.8$). None of the participants reported any hearing problems and the mother tongue of all was German.

3. RESULTS

3.1 Gender identification

In listening experiment 1, the gender identification score of the individual children is between 1.55 and 6.56 on the seven-point scale (mean of both sentences and grade levels). As expected, boys and girls differ significantly in a Wilcoxon rank-sum test (girls: $\bar{x} = 5.09$, $\sigma = 1.01$; boys: $\bar{x} = 3.04$, $\sigma = 1.18$; $W = 85$, $p < .01$*). When we take a score less than 2 or greater than 6 as a sign of systematic gender assignment, 15 children are producing phonetic patterns that receive unambiguous gender attribution by the listeners. The remaining 45 children are producing acoustic patterns that lead to more ambivalent gender responses. Therefore it is not surprising that

	all (N = 60)							selected (N = 10)						
	girls		boys					females		males				
	x̄	σ	x̄	σ	p	W	D	x̄	σ	x̄	σ	p	W	d
F0 (Hz)	250	21	249	24	1.0	6845	.08	256	22	232	16	.01*	78	1.0
Semitones	6.82	2.58	6.66	1.91	.99	7170	.01	7.10	1.81	6.42	1.67	1.0	162	.33
AVS (kHz²)	1.44	1.10	1.30	0.94	1.0	6853	.08	1.64	0.99	1.67	1.19	.99	201	.01
HNR (dB)	19.10	3.01	20.10	3.02	.04*	8620	.35	18.74	2.75	18.45	3.77	1.0	196	.03
Tempo (syll/s)	3.92	0.58	4.15	0.57	.01*	8914	.42	4.04	0.48	4.08	0.65	1.0	196	.03
CoG (Hz)	4198	1987	3599	2013	.10	5909	.30	4209	2381	3699	2036	1.0	172	.23
Skewness	0.59	1.49	1.14	2.14	.08	8408	.30	0.64	1.35	1.33	2.06	1.0	245	.38

Table 1: Arithmetic means and standard deviations of acoustic parameters (girls and boys / selected female- and male-sounding children). Wilcoxon rank-sum tests (p-values Bonferroni-Holm-corrected) and Cohen's d on the right.

the intraclass correlation coefficient (ICC) including the ratings of all children shows a poor inter-rater reliability (*ICC (1, 1) = .50, p < .01**), while the inter-rater reliability of the other 15 children is quite good (*ICC (1, 1) = .82, p < .01**).

For listening experiment 2 and the acoustic analysis, we choose the five most female- and male-sounding children (mean of both grades). In this case, perception coincides with biological gender.

3.2 Acoustic analysis

Table 1 shows the acoustic differences between girls and boys (mean of both sentences and both grade levels). When all 60 children are included, significant differences in the Wilcoxon rank-sum tests can be found in HNR and speech tempo with higher speech rates and HNR values for boys. Cohen's d was also calculated to evaluate the effect sizes. Here we can assume medium effects of gender for HNR, speech tempo, CoG and skewness. In line with earlier studies [17, 18, 19], girls have higher frequencies of CoG and lower values for skewness than boys. However, the standard deviations of both parameters are quite high, which suggests a high interindividual variability. By contrast to other studies [12, 13, 14, 15], HNR values for the boys are higher than for the girls indicating a higher modality in the boys' voices. The girls have a slightly larger AVS than the boys; the fundamental frequency of both genders is almost the same.

Interestingly, if we only compare the most female- and male sounding children (selected from listening experiment 1, see 3.1), only the fundamental frequency differs significantly between both groups with a large effect – the female children speak with higher fundamental frequency than the male children. The semitone range of the females is also higher than it is for the males (medium effect). AVS, HNR and speech tempo are almost the same. As in the whole group, the girls show higher values for CoG and lower values for skewness than the male children, with a small effect for CoG and a medium effect for skewness.

3.3 Acoustic analysis and perception

To find out which acoustic parameters predict the perception of the children's voices, linear-mixed-models were run using the *lme4 package* [28] in *R* [29] for all children and separately for girls and boys. Perception was the dependent variable acoustic parameters of sentence 1 and 2 were tested as fixed factors. We entered a random intercept for each child. P-values were obtained using Likelihood ratio tests comparing the model with the acoustic parameters with the model without them. Interactions between each acoustic parameter and grade level or sentence have been calculated.

When all children are tested, the model is best when skewness is added as fixed factor (χ^2 *(1) = 4.26, p = .02**). There is no interaction with grade level (χ^2 *(2) = 2.69, p = .26*) or sentence (χ^2 *(2) = 1.51, p = .47*). For the boys, we get the best result when we only add f0 as fixed factor (χ^2 *(1) = 5.36, p = .02**). There is no interaction with grade level (χ^2 *(2) = 4.10, p = .13*) or sentence (χ^2 *(2) = 2.87, p = .24*). For the girls, the best result can be achieved when AVS (χ^2 *(1) = 4.07, p = .04**), HNR (χ^2 *(1) = 6.92, p = .01**) and tempo (χ^2 *(1) = 4.35, p = .04**) are included as fixed factors. Interactions can be found between grade level and AVS (χ^2 *(2) = 7.74, p = .02**) and grade level and tempo (χ^2 *(2) = 6.82, p = .02**), whereby the influence of AVS on gender perception is greater in the first grade and tempo in the second grade. There is no interaction between grade level and HNR (χ^2 *(2) = 5.90, p = .05*). For sentence, interactions can be found with AVS (χ^2 *(2) = 8.82, p = .01**) and tempo, (χ^2 *(2) = 9.58, p = .01**). The influence of AVS on gender perception is greater in sentence 1, whereas tempo has a greater effect in sentence 2. There is no interaction between sentence and HNR (χ^2 *(2) = 3.93, p = .14*).

3.4 Perceptual attribute ratings

The perceptual attributes high–low (pitch), monotonous–melodious (range), fast–slow (tempo), hoarse–clear (hoarseness), loud–quiet (volume) and clear–mumbled (precision) were assessed on a seven-point-scale. The first term has the value 1 (e.g.

"high"), the second the value 7 (e.g. "low"). If we compare these perception ratings (mean of both stimuli) of the most female- and male-sounding children, significant differences can be found for pitch, range and precision, which means that the female children sound higher, more melodic and seem to have a more precise articulation than the males (see table 2). In these parameters, large effects can be found. Furthermore, there is a large effect of hoarseness with males being perceived as hoarser than females.

	female		male				
	$\bar{x}$	σ	$\bar{x}$	σ	p	t	d
Pitch	2.98	0.58	5.10	0.30	.001*	9.53	4.3
Range	4.94	0.29	3.56	0.37	.001*	-4.40	2.0
Tempo	4.06	0.56	3.99	0.86	.87	-0.17	.08
Hoarsen.	5.11	0.54	3.49	0.63	.27	-5.03	2.2
Volume	3.76	0.33	4.03	0.49	.46	1.25	.56
Precis.	2.85	0.40	4.47	0.64	.001*	4.57	2.0

Table 2: Arithmetic means and standard deviations of attribute ratings. T-tests (p-values Bonferroni-Holm-corrected) and Cohen's d on the right.

It makes no significant difference whether listeners thought they were listening to boys or girls ($p > .40$ in all cases).

Spearman's ρ correlations between attribute ratings and corresponding acoustic parameters (mean of both stimuli) were run. Perceived precision was correlated with acoustical vowel space and tempo. There are strong significant correlations between perceived pitch and f0 ($\rho = -.84$, $p < .01$*) and between perceived tempo and measured tempo ($\rho = -.83$, $p < .01$*): With increasing f0 and tempo, the children are perceived as higher pitched and faster speaking (see figure 1). No significant correlations between semitone range and perceived range ($\rho = .27$, $p = .45$) or between HNR and perceived hoarseness ($\rho = .53$, $p = .12$) can be found. For precision, there is neither a significant correlation with AVS ($\rho = -.27$, $p = .44$) nor with tempo ($\rho = .25$, $p = .48$).

4. DISCUSSION

Studies have shown that the gender perception of boys and girls differs significantly in a listening experiment. Similar to [14, 15], 15 of the 60 children seem to produce unequivocal acoustic patterns and achieve very high or low perception values. When all 60 children are compared to each other, significant differences between girls and boys can be found in HNR and tempo, with girls having higher levels of noise and lower tempo. This is contrary to the results of [14]. It is possible that the girls of the present study were shyer and as a result used quieter, breathier voice qualities than the boys of the same age. In [14], the sentences were read by the children. Therefore it is probable that reading competence rather than speech tempo was reflected there.

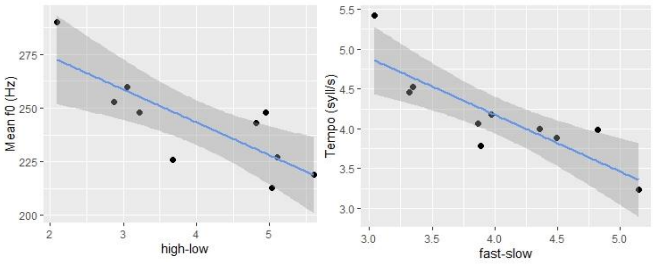


Figure 1: Acoustic parameters plotted as a function of perceptual attribute ratings. Left: "high–low" and mean f0; Right: "fast–slow" and tempo.

When we only compare the five most female- and male-sounding children, only f0 differs significantly. This would suggest that f0 plays a strong role in gender perception. Linear-mixed-models confirm this for boys, whereas skewness is also decisive for perception across gender. For girls, AVS, HNR and tempo could be crucial. Interactions between AVS, tempo and grade level or sentence show that in different stimuli, different parameters could be used for gender construction and perception in the female group.

A listening experiment with perceptual attribute ratings shows significantly higher ratings of pitch, pitch range and precision for the female-sounding children. In addition, there are strong correlations between perception and corresponding acoustic parameters, especially for pitch and speech tempo. In other words, listeners are able to recognize acoustic parameters associated with gender perception.

From these results we can conclude that there is a range of acoustic parameters that influence gender perception. Furthermore, it seems that a particular high or low gender rating does not rely on a single acoustic parameter. In further listening experiments using sophisticated voice morphing techniques [30] we will be able to examine in more detail the role played by individual acoustic parameters. In addition, due to the longitudinal nature of the project from which this study is a part, we will be able to analyse how the vocal expression of gender in individual children develops over the first years in primary education before reaching puberty.

5. ACKNOWLEDGMENTS

The project is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grant SI 743/9-1, http://www.dfg.de/). We would like to thank Christoph Draxler (MLU Mu-

nich) for his help in constructing and creating the first listening experiment. Most of all we are very grateful to the children and their parents agreeing to take part in this project.

6. REFERENCES

[1] Kahane, J. C. 1978. A morphological study of the human prepubertal and pubertal larynx. Journal of Anatomy 151, 11–20.

[2] Fitch, W. Tecumseh & J. Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America 106(3), 1511–1522.

[3] McCormack, P. F. & T. Knighton. 1996. Gender differences in the speech patterns of two and a half year old children. In Proc. Sixth Australian International Conference on Speech Science and Technology. Adelaide, 337–342.

[4] Hasek, C. S., S. Singh, & T. Murry. 1980. Acoustic attributes of children's voices. Journal of the Acoustical Society of America 68, 1262–1265.

[5] Ferrand, C.T. & R. L. Bloom. 1996. Gender differences in children's intonational patterns. Journal of Voice 10(3), 284–29.

[6] Glaze, L. E., D. M. Bless, P. Milenkovic, & R. D. Susser. 1988. Acoustic characteristics of children's voice. Journal of Voice 2(4), 312–319.

[7] Whiteside, S. P. & C. Hodgson. 2000. Speech patterns of children and adults elicited via a picture-naming task: An acoustic study. Speech Communication 32, 267–285.

[8] Bennett, S.. 1981. Vowel formant frequency characteristics of preadolescent males and females. Journal of the Acoustical Society of America 69(1), 231–239.

[9] Busby, P. A. & G. L. Plant. 1995. Formant frequency values of vowels produced by preadolescent boys and girls. Journal of the Acoustical Society of America 97, 2603–2607.

[10] Pettinato, M., O.Tuomainen, S. Granlund, & V. Hazan. 2016. Vowel space area in later childhood and adolescence. Effects of age, sex and ease of communication. Journal of Phonetics 54, 1–14.

[11] Whiteside, S. P. 2001. Sex-specific fundamental and formant frequency patterns in a cross-sectional study. Journal of the Acoustical Society of America 110, 464–478.

[12] Ferrand, C. T. 2000. Harmonics-to-noise ratios in normally speaking prepubescent girls and boys. Journal of Voice 14(1), 17–21.

[13] Kallvik, E., E. Lindström, S. Holmqvist, J. Lindman, & S. Simberg. 2015. Prevalence of hoarseness in school-aged children. Journal of Voice 29(2), 260–e1.

[14] Simpson, A. P., R. Funk, & F. Palmer. 2017. Perceptual and acoustic correlates of gender in the prepubertal voice. In Interspeech 2017. Stockholm, 914–918.

[15] Funk, R., S.Voigt-Zimmermann, & A. P. Simpson. 2018. Junge oder Mädchen? Zur Geschlechtsidentifikation pr¨apubertärer Stimmen. In Proc. Phonetik und Phonologie 14. Vienna.

[16] Walker, J. F., L. M. D. Archibald, S. R. Cherniak, & V. G. Fish. 1992. Articulation rate in 3-and 5-year-old children. Journal of Speech, Language and Hearing Research 35(1), 4–13.

[17] Li, F., D. Rendall, P. L. Vasey, M. Kinsman, A. Ward-Sutherland, & G. Diano. 2016. The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. Journal of Phonetics 57, 59–70.

[18] Shadle, C. H., L. L. Koenig, & J.L. Preston. 2011. Acoustic characterization of /s/ spectra of adolescents: moving beyond moments. In Proceedings of Meetings on Acoustics 12. Seattle.

[19] Flipsen, P., L. Shriberg, G. Weismer, H. Karlsson, & J. McSweeny. 1999. Acoustic characteristics of /s/ in adolescents. Journal of Speech, Language, and Hearing Research 42, 663–677.

[20] Günzburger, D., A. Bresser, & M. ter Keurs. 1987. Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects. Language and Speech 30, 47–58.

[21] Karlsson, I. 1987. Sex differentiation cues in the voices of young children of different language background. Journal of the Acoustical Society of America 81, 68–69.

[22] Kaya, H., A. A. Salahb, A. Karpovc, O. Frolovae, A. Grigoreve, & E. Laykso. 2017. Emotion, age, and gender classification in children's speech by humans and machines. Computer Speech and Language 46, 268–283.

[23] Audacity Team. 2021. Audacity(R): Free Audio Editor and Recorder. Version 3.0.0 retrieved March 17th 2021 from https://audacityteam.org/.

[24] Kisler, T., Reichel U. D., & F. Schiel. 2017. Multilingual processing of speech via web services. Computer Speech & Language 45, 326–347.

[25] Boersma, Paul & David Weenink. 2018. Praat, doing phonetics by computer. http://www.praat.org, version 6.0.40.

[26] Draxler, Christoph. 2014. Online experiments with the Percy software framework – experiences and some early results. Reykjavik, Iceland, 235–240.

[27] Leiner, D. J. (2019). SoSci Survey. Version 3.1.06. Available at https://www.soscisurvey.de

[28] Bates D, Maechler M, Bolker B & Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4, Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[29] R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

[30] Kawahara, Hideki, M. Morise, Toru Takahashi, R. Nisimura, T. Irino, & Hideki Banno. 2008. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 3933–3936.

UNMASKING THE TRUF: IMPACT OF COMMUNITY MASKS ON THE PERCEPTION OF VOICELESS FRICATIVES IN ENGLISH

Massimiliano Canzi & Tamara Rathcke

University of Konstanz
{massimiliano.canzi, tamara.rathcke}@uni-konstanz.de

ABSTRACT

The current study aims at quantifying the effects of wearing a face mask on speech perception, by investigating performance of native English listeners in a phoneme monitoring task with monosyllabic words containing voiceless fricatives. Previous experimental work on the topic has mainly focussed on the effects of acoustic filtering caused by the use of face coverings with mixed results and weak effects of mask wearing on speech perception. In this experiment, we explore the interplay of acoustic filtering with other potentially relevant factors such as the presence of visual cues, lexical frequency and listener-specific background. We provide evidence that suggests the impact of face coverings (esp. FFP-2 face mask) on speech perception is not directly moderated by the acoustic properties of masked speech. Rather, it is inked to an interplay of audio-visual integration, the absence of visual cues for (some) target fricatives, and the listener-specific sociolinguistic background.

Keywords: face mask, fricative perception, th-fronting, lexical frequency, COG, intensity

1. INTRODUCTION

Following the initial outbreak of the SARS-CoV-2 pandemic in March 2020, wearing face masks had become obligatory in public places around the world. There exist many anecdotal reports of speech perception being negatively affected by the presence of a face mask cf. [1]. In theory, such negative impact of a face mask might arise for several reasons. Speech perception is multimodal and it relies on a number of cues [2], including visual cues which are lacking in masked speech. The visual modality is particularly important for the perception of labial and dental sounds whose articulation has more readily visible articulatory movements (e.g. [3]). Such visual cues are arguably less crucial for the perception of alveolar or post-alveolar sounds [4]. Moreover, a face mask can be considered a filter to the acoustic properties of speech, due to the air stream passing through the fabric [5] [4].

There has been a long-standing interest in the impact of culturally motivated face coverings on speech acoustics and perception in forensic contexts [4] [6] [7]. For example, the study by [7] analysed recordings of 9 speakers (F = 5) with seven types of face coverings, including surgical masks, motorcycle helmets, ski and cycling masks. The speakers produced words of the CVC structure, with either onsets or codas containing /f/, /θ/ or /ʃ/. Presence of a face covering was shown to affect the centre of gravity (COG) in all fricatives, especially (labio)dentals. Spectral skewness and curtosis were also affected in all fricatives while only dentals additionally showed a large modification of the spectral peak.

Other studies focused on the perceptual effects of face coverings [6] [8]. For example, the impact of of niqāb, balaclava and a surgical mask was examined with regards to the perception of voicing, manner and place of articulation of stops [6]. Some subtle misperceptions were observed in place of articulation of fricatives (especially /f/ vs. /θ/) and nasals (esp. /n/ and /ŋ/), though overall, a surprisingly small number of misperceptions was attested. In contrast to the study of the isolated sound perception [6], the experiment in [8] examined listeners' transcriptions of whole words and sentences produced by speakers wearing different types of face masks (N95, surgical, cloth). The impact on intelligibility was found to be rather negligible and also comparable across all face masks (3-5%).

A combined effect of the presence of face masks or other face coverings and adverse listening conditions (like an impoverished signal-to-noise ratio, SNR) has also been examined [7] [9]. Overall, the use of surgical or FFP-2 masks shows little negative impact on perception in quiet listening conditions. In lower SNR-levels, an audio-visual presentation can give a 10% boost in intelligibility as compared to the same signal presented as audio-only [7]. Speech perception can be similarly affected in older and younger adults, though older listeners tend to perform slightly worse on the intelligibility tasks in noise and find listening to masked speech more strenuous and effortful [9]. Surprisingly though, a face mask with a transparent window to expose

speaker mouth does not seem to generally promote intelligibility [9].

Overall, existing studies demonstrate that despite subjective beliefs that masked speech is more difficult to understand [1], the documented effects on speech perception seem to be rather subtle (if at all present). However, few studies examined how speech intelligibility may be affected by an interplay of the acoustic filtering effects and the absence of visual cues arising due to face masks, lexical frequency and listener-specific background. The present study was based on the following three hypotheses:

- Acoustic degradation (due to filtering [4]-[5]) along with the absence of visual cues may be jointly contributing to the lower level of intelligibility in masked speech.

- Lexical frequency and lexical competition may moderate the intelligibility effect, with high-frequency words and no minimal pairs causing less difficulties than low-frequency words and the presence of minimal pairs.

- Listener-specific background (specifically the presence of sound mergers in their variety, e.g. due to th-fronting [10]) may further mediate the perception of masked speech.

2. METHODS

2.1. Participants

Seventy native speakers of English (F = 44; Age: Mo = 20-29, R = 18-65) volunteered to participate. The participants were hailing from nine different English-speaking countries, including Australia, Canada, England, India, Ireland, New Zealand, Scotland, Singapore and the United States. The participants from England and Scotland were coded as th-fronters due to the exposure to /θ/-/f/ merger in these varieties [10]. The participants were recruited via social media and could decide to enter a prize draw in recognition of their time and efforts. They were screened for visual and hearing impairments and reported normal hearing and normal or corrected-to-normal vision.

2.2. Stimuli

The target fricatives of the present study included /f/, /θ/, /s/ and /ʃ/. They were chosen because of the presence (/f/, /θ/) or the absence (/s/, /ʃ/) of visual cues during their production. Moreover, the discriminability of /θ/ and /f/ is reduced in some varieties of English, due to a fricative merger [10].

A total of 100 monosyllabic English word stimuli were devised for the purposes of the experiment. Five minimal quadruplets contained target fricatives in onset positions (e.g. *fie/thigh/shy/sigh*). Ten minimal pairs were created with targets in coda positions (e.g. *deaf/death* and *mess/mesh*). Forty monosyllabic words contained the target fricatives without a lexical competitor (e.g. *fact,shame*). In addition, 20 monosyllabic distractors without a target fricative were created. Lexical frequency measures were obtained for all stimuli from the SUBTLEX-UK database [11]. A total of 100 experimental items were recorded audio-visually in two experimental conditions (mask, no mask) by one white, male, native speaker of Southern British English in his thirties. For the masked speech condition, the speaker was wearing a FFP-2 mask that was commonly used in Germany at the time of the recording. The mask prevents listeners from observing any lip movements of the speaker.

2.3. Procedure

The study utilized an online platform called Gorilla [12]. Participants were instructed to perform the experiment in a quiet room using only wired or built-in mouse and keyboard, headphones, or speakers. All stimuli were presented audio-visually and twice, once in the no mask condition and once in the mask condition. Each participant was presented with two distinct blocks of stimuli, one without a mask and one with a mask. The order of presentation of the two blocks was counterbalanced across participants. Within each block, the order of all stimuli was fully randomized. Participants were tasked with monitoring the stimuli for the occurrence of one of the four fricatives and were prompted to respond as quickly as possible by clicking on the corresponding button (F, TH, S, and SH) shown on the experimental screen. If none of the four fricatives were detected, participants were instructed to click on the X button located in the center of the screen. Following the phoneme monitoring task, participants completed a questionnaire that included general demographic questions and queries about their experience wearing face masks in public.

2.4. Data treatment

Data collected in the phoneme monitoring task were analysed with R (v4.2.1) [13] using RStudio [14]. Two main outcome variables were selected: (1) accuracy (correct, incorrect) which refers to whether or not the participant correctly selected the right phoneme in each trial and (2) reaction times (measured in ms and logarithmically transformed to assume a log-normal distribution [15]). In compari-

son to the binary variable of accuracy, reaction times might provide a more nuanced look at the data, by providing an indication of changes in cognitive load and processing ease across conditions [16]. For the statistical analysis, only reaction times of correct answers were selected.

A stepwise backwards model fitting procedure was deployed, by using model comparisons (through functions such as `anova` and `drop1`) [13] [17] [18] to obtain best-fit models. Post-hoc comparisons of the relevant factor levels were carried out using the `emmeans` (v1.8.0) package [19].

For accuracy, a mixed logistic regression model was devised using the lme4 (v1.1.3) package [17]. Participants and stimuli were added as random intercepts. Slopes were allowed over the random intercept of participant where appropriate. Reaction time data were analysed using linear mixed-effects regression, again with package lme4 [17]. Varying intercepts were allowed for participant, item and browser. Slopes were allowed over participant where appropriate.

3. RESULTS

3.1. Acoustics of masked speech

We tested for acoustic differences in the spoken stimuli across all four fricatives as well as across conditions to see whether acoustic properties such as COG and intensity should be added to our main, perceptual analysis. We ran two linear models and we found a significant interaction of experimental condition and fricative for both COG and intensity. Post-hoc comparisons [19] show that mask wearing affects COG the most in /f/ [t(76) = -8.91, $p < .0001$] followed by /θ/ [t(76) = -5.18, $p < .0001$]. However, COG is not affected by mask wearing for both /ʃ/ [t(76) = -0.79, n.s.] and /s/ [t(76) = -1.77 , n.s.]. Conversely, /s/ appears to be the one fricative where intensity is the most affected by mask wearing [t(76) = -9.68, $p < .0001$], followed by /f/ [t(76) = -7.03, $p < .0001$], /ʃ/ [t(76) = -5.16, $p < .0001$] and /θ/ [t(76) = -3.117, $p < .001$]. All the p-values are corrected for multiple comparisons within the `emmeans` package.

3.2. Accuracy

To test the effects of mask wearing on phoneme monitoring accuracy, we devised a generalised linear logistic regression model. The best-fitting model included the interaction of experimental condition (mask, no mask) and target (F, TH, S, SH) [χ^2 = 53.51, $p < .0001$]. COG [χ^2 = 8.72, $p < .005$] and lexical frequency [χ^2 = 8.00, $p < .005$] entered as

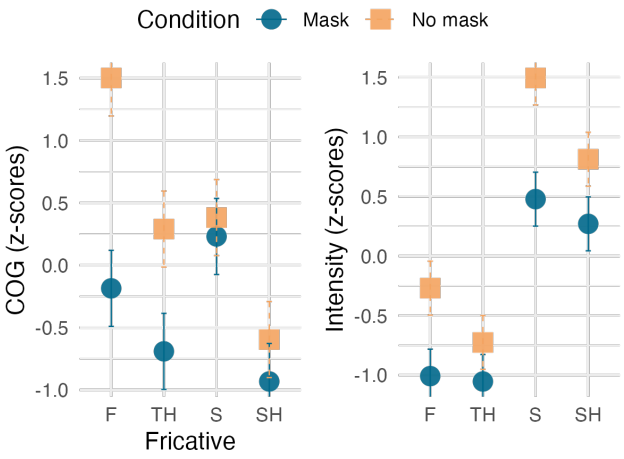


Figure 1: Model estimates of COG and intensity by experimental condition across all fricatives

main effects. Interactions of these factors with the experimental condition were not significant. Pairwise post-hoc comparisons for the interaction of interest reveal that the effect of condition is only significant for non-sibilants that normally have visual cues in the no-mask condition. The difference is particularly large for /θ/ (z = -10.01, $p < .0001$) compared to /f/ (z = -6.838, $p < .0001$). However, /f/ has the lowest accuracy rate across all conditions. The main effect of COG shows a negative relationship between accuracy and COG. The main effect of lexical frequency suggests that higher frequency goes hand-in-hand with higher accuracy scores.

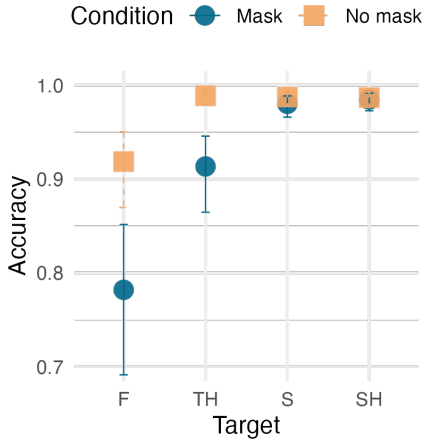


Figure 2: Model estimates of accuracy by experimental condition across all target fricatives

We tested the effect of th-fronting on non-sibilant monitoring accuracy, by devising a supplementary model with only /θ/ and /f/ as targets and with participants grouped into varieties that present th-fronting (England, Scotland) and those that do not (all other countries). The best-fitting model included

the non-significant interaction of experimental condition and merger [χ^2 = 0.57, n.s.], as well as the significant interaction of experimental condition and target [χ^2 = 22.01, p < .0001]. COG [χ^2 = 14.60, p < .001] and log-transformed lexical frequency [χ^2 = 5.7, p < .0001] also entered as main effects. Participant and item were allowed as varying intercepts. We found no effect of th-fronting on accuracy specifically. We then decided to not distinguish participants by the merger status in their variety for the main analysis.

3.3. Reaction Times

To determine whether reaction times were also affected by masked speech during phoneme monitoring, we fit a linear mixed regression model to the log-transformed reaction times of correct responses. The best-fitting reaction-time model included significant main effects of experimental condition [χ^2 = 9.14, p < .005], target fricative [χ^2 = 44.19, p < .0001], device (mobile, computer) [χ^2 = 17.52, p < .0001] and log frequency [χ^2 = 7.91, p < .005]. However, no significant interaction between condition and target was discovered, suggesting that differences in reaction times happen uniformly across all target fricatives. We ran post-hoc analyses using the `emmeans` package to determine marginal means and the direction of effects. In particular, reactions times appear to be higher for the mask (β = 6.45, SE = .04) than the no mask (β = 6.37, SE = .04) condition. Overall, sibilants appear to be identified significantly faster than non sibilants [z = -6.35, p = .0001].

We also tested the effects of th-fronting on reaction times of correct responses in a separate model that only included /θ/ and /f/ as target fricatives. The best-fitting model, among significant main effects of experimental condition, COG, log-transformed lexical frequency and target fricative, included a main effect of th-fronting [χ^2 = 3.85, p = .001]. However, no significant interaction of th-fronting and experimental condition was discovered. This finding suggests that participants whose English varieties present instances of th-fronting are generally characterised by slower reaction times. However, they are not impacted by masked speech any differently than participants who speak English varieties where th-fronting is not as prevalent.

4. DISCUSSION

The present study was designed to determine the effects of FFP-2 masks on the acoustics and the perception of voiceless fricatives in English. The acoustics of /f/ (COG) and /s/ (Intensity) are affected more

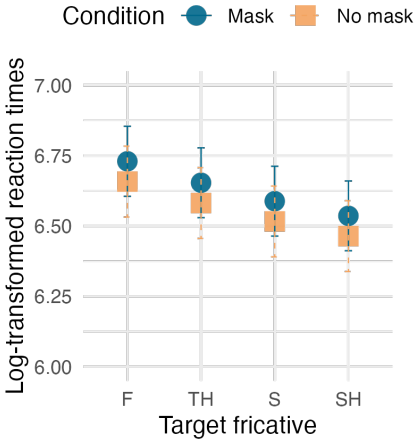


Figure 3: Model estimates of log-transformed reaction times by experimental condition across all target fricatives

than the acoustics of /θ/ and /ʃ/. However, perceptually, the fricatives most affected by a face mask are /θ/ and /f/ (not at all /s/, despite having the largest change in intensity overall). These findings suggest that the source of the effect is not (primarily) in the acoustic modification of masked speech (cf. [6], [8]). Rather, it is a combination of the lack of visual cues in (labio)dentals and a relatively low level of acoustic energy in non-sibilants. [20].

The experiment found that lexical frequency significantly predicted overall higher rates of phoneme monitoring accuracy and lower reaction times for correct responses. However, this is equally true for both experimental conditions. The presence of lexical competition did not significantly affect either the accuracy or reaction times. It is possible that the results were affected by non-uniform distribution of (competitor) stimuli across the lexical frequency range.

Among all fricatives, /θ/ is most affected by a face mask, while /f/ is also not recognised at ceiling in the unmasked audio-visual condition. The lower accuracy rates are not, however, a consequence of th-fronting in varieties in which /f/ maps onto both /f/ and /θ/. Rather, the result might be to do with the ambiguously encoded (labio)dental contrast in English [21].

This work contributes to existing knowledge on masked-speech perception by integrating previous work on the effects of acoustic filtering with factors such as the absence of visual cues, lexical frequency, and listener-specific background. Future research should investigate the effects of face masks on a wider range of phonemes and cross-cultural aspects of audio-visual integration that may mediate speech perception.

5. ACKNOWLEDGEMENTS

We thank George Walkden for his contribution to the recording of all experimental stimuli. We are also grateful to the students of the *Speech Perception* class (Aikaterini Tsaroucha, Hanna Kim and Katharina Hölzl) for their help in collecting experimental data.

6. REFERENCES

[1] Goldin, A., Weinstein, B., Shiman, N., others, 2020. How do medical masks degrade speech perception. *Hearing review* 27(5), 8–9.

[2] Partan, S., Marler, P. 1999. Communication goes multimodal. *Science* 283(5406), 1272–1273.

[3] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(5588), 746–748.

[4] Fecher, N., Watt, D. 2011. Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. *ICPhs*, 663–666.

[5] Coniam, D. 2005. The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-sars study in hong kong. *Language Assessment Quarterly: An International Journal* 2(4), 235–261.

[6] Llamas, C., Harrison, P., Donnelly, D., Watt, D. 2009. Effects of different types of face coverings on speech acoustics and intelligibility.

[7] Fecher, N. 2014. *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants*. PhD thesis University of York.

[8] Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., others, 2020. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *The Journal of the Acoustical Society of America* 148(6), 3562–3568.

[9] Brown, V. A., Van Engen, K. J., Peelle, J. E. 2021. Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications* 6(1), 1–12.

[10] Levon, E., Fox, S. 2014. Social salience and the sociolinguistic monitor: A case study of ing and th-fronting in britain. *Journal of English Linguistics* 42(3), 185–217.

[11] Van Heuven, W. J., Mandera, P., Keuleers, E., Brysbaert, M. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology* 67(6), 1176–1190.

[12] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods* 52(1), 388–407.

[13] Team, R. C., others, 2013. R: A language and environment for statistical computing.

[14] Racine, J. S. 2012. Rstudio: a platform-independent ide for r and sweave.

[15] Baayen, R. H., Milin, P. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3(2), 12–28.

[16] Eggemeier, F. T. 1988. Properties of workload assessment techniques. In: *Advances in psychology* volume 52. Elsevier, 41–62.

[17] Bates, D. M. 2010. lme4: Mixed-effects modeling with r.

[18] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software* 82, 1–26.

[19] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R package version* 1(1), 3.

[20] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America* 108(3), 1252–1263.

[21] Babel, M., McGuire, G. 2013. Listener expectations and gender bias in nonsibilant fricative perception. *Phonetica* 70(1-2), 117–151.

SYLLABLE RATE DRIVES RATE NORMALIZATION, BUT IS NOT THE ONLY FACTOR

Giulio G.A. Severijnen [1], Hans Rutger Bosker [1,2], James M. McQueen [1,2]

[1]Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
giulio.severijnen@donders.ru.nl, hansrutger.bosker@donders.ru.nl, james.mcqueen@donders.ru.nl

ABSTRACT

Speech is perceived relative to the speech rate in the context. It is unclear, however, what information listeners use to compute speech rate. The present study examines whether listeners use the number of syllables per unit time (i.e., syllable rate) as a measure of speech rate, as indexed by subsequent vowel perception. We ran two rate-normalization experiments in which participants heard duration-matched word lists that contained either monosyllabic vs. bisyllabic words (Experiment 1), or monosyllabic vs. trisyllabic pseudowords (Experiment 2). The participants' task was to categorize an /ɑ-aː/ continuum that followed the word lists. The monosyllabic condition was perceived as slower (i.e., fewer /aː/ responses) than the bisyllabic and trisyllabic condition. However, no difference was observed between bisyllabic and trisyllabic contexts. Therefore, while syllable rate is used in perceiving speech rate, other factors, such as fast speech processes, mean F0, and intensity, must also influence rate normalization.

Keywords: Rate normalization, speech rate, syllable rate.

1. INTRODUCTION

Talkers vary in their speech rate, leading to differences in the duration of sentences and the speech sounds in those sentences (e.g., words, syllables, vowels etc.). This poses a problem for speech sounds that are contrasted through temporal information, such as vowels or plosives. Listeners deal with this variability by normalizing vowel perception for the surrounding speech rate [1], [2]. This is a contrastive effect, in which sounds in a fast context are perceived as relatively slow (long), and sounds in a slow context as relatively fast (short). While this effect has been well established, it remains unclear what information listeners use to compute a talker's rate. The present study examined whether the syllable rate (number of syllables per unit time) alone can drive rate normalization.

Previous studies that examined rate normalization have mostly manipulated speech rate by recording a sentence at a normal rate and linearly compressing or expanding the sentence [2], [3], or instructing speakers to produce a natural fast or slow sentence [4], [5]. While these manipulations succeeded in inducing rate normalization, there are many possible cues that could have driven the perception of fast vs. slow tempo, including the number of produced segments [6], [7], the presence of fast speech processes such as reduction [8], but also prosodic cues such as intensity and mean fundamental frequency [9], [10].

Koreman [6] examined whether the number of produced vs. intended syllables/segments affected perceived rate in an explicit rate judgement task. Results showed that sentences with more produced segments (e.g., "He probably said") were perceived as faster than duration-matched sentences with deletions, and thus fewer produced segments per unit time (e.g., "He pro'bly said"). Moreover, sentences with deletions ("He pro'bly said") sounded faster than sentences with the same number of segments but without deletions ("He always said"), and thus a lower number of intended segments. This suggests that listeners track the number of produced and intended segments to compute speech rate. Converging results for syllables, but not for phones, were found by Plug et al. [7]. However, both studies involved explicit rate judgments, not rate normalization.

Reinisch [8] extended these findings to implicit rate normalization. She examined how sentences with or without fast speech processes (deletions and reductions) affected subsequent /a-aː/ perception in German. In contrast to [6] and [7], results showed the exact opposite pattern: duration-matched sentences with fast speech processes (and thus fewer produced segments) resulted in more /aː/-responses than those without, implying that these sentences were perceived as faster. She concluded that the presence of fast speech processes is an indication to listeners of a faster speech rate, in turn influencing subsequent vowel perception.

Despite the diverging results in these previous studies, they all seem to agree that listeners use the number of units (either segments or syllables) per unit time to compute speech rate. However, sentences contain various speech units that differ in their

relative salience. For instance, vowels usually have a greater intensity than consonants, and stressed syllables usually have a greater intensity than unstressed syllables. Considering this difference in salience, do listeners then use all available units alike to compute speech rate, or just the most salient ones?

The present study examined this for syllables, asking: Do listeners use the number of syllables per unit time (i.e., syllable rate) to compute speech rate when strictly controlling for the number of stressed syllables? Alternatively, listeners could use only the stressed syllables (i.e., stress rate) to compute speech rate. We ran two implicit rate normalization experiments using Dutch /ɑ-aː/ targets. In Experiment 1, we compared target perception after duration-matched monosyllabic vs. bisyllabic word lists (i.e., same stress rate, different syllable rate). Experiment 2 compared duration-matched monosyllabic vs. trisyllabic pseudoword lists. If listeners use the syllable rate as the measure for speech rate, monosyllabic word lists should be perceived as two times slower than duration-matched bisyllabic lists and as three times slower than trisyllabic lists. Moreover, if listeners weigh all syllables equally, we might expect a linear relation between the number of syllables in the word lists and the rate normalization effect size.

2. METHODS

2.1. Participants

We recruited 40 native speakers of Dutch from the Radboud University participant pool (Experiment 1: 16 female, 4 male, age range: 18 – 34, M_{age} = 22, SD_{age} = 4.44; Experiment 2: 12 female, 8 male, age range: 21 – 32, M_{age} = 26, SD_{age} = 3.21). All participants gave informed consent and received a monetary reward or course credits for their participation. No participant reported having any speaking or hearing problems.

2.2. Materials

We created word lists in which the first four words served as context to a fifth word (target) that could contain either a short /ɑ/ or a long /aː/. The complete stimulus list has been made available at: https://osf.io/36anr/?view_only=ad1cb3f67e2d4e749620119f7c88c788. All materials were recorded from a female native talker of Dutch.

2.2.1. Context words

For Experiment 1, we selected ten minimally different Dutch word pairs in which one word was monosyllabic, the other bisyllabic with final stress (e.g., *klom,* /klɔm/, 'climbed', vs. *kolom,* /ko.ˈlɔm/, 'column'). Critically, the bisyllabic words differed from the monosyllabic words only in the presence of an unstressed vowel between the first two consonants (1:2 syllable ratio, same stress rate). Next, we set both words of each pair to the same mean duration using PSOLA in Praat [11] (see Figure 1). This was done within word pairs (i.e., *klom* and *kolom* had the same duration, but could differ in duration from another pair).

We then created 12 word lists. Each word list contained either four monosyllabic or four bisyllabic words, with 50 ms of silence between the words, resulting in six monosyllabic lists and six bisyllabic lists. We created six different combinations of the context words, and for every monosyllabic combination of context words, there was a duration-matched bisyllabic version (e.g., *klom, fruit, trein, fluit* vs. *kolom, vooruit, terrein, voluit*).

For Experiment 2, we created ten pairs of pseudowords, following Dutch phonotactics (e.g., *tralk,* /trɑlk/ vs. *tarallok*, /ta.ˈrɑ.lɔk/). In each pair, one pseudoword was monosyllabic, the other trisyllabic with a stressed second syllable (1:3 syllable ratio, same stress rate). Similar to Experiment 1, both pseudowords within a pair were set to the mean duration for the corresponding pair and we again created 12 word lists (6 with monosyllabic items, 6 with trisyllabic items) by concatenating the pseudowords.

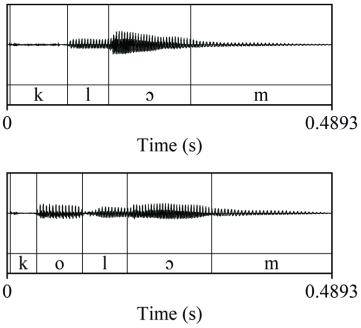


Figure 1: Oscillogram of the context word pair *klom* (top panel) vs. *kolom* (bottom panel)

2.2.2. Target words

We selected six Dutch monosyllabic word pairs that differed only in their vowel, containing either a short /ɑ/ or a long /aː/ (e.g., *stad,* /stɑt/, 'city' or *staat,* /staːt/, 'state').

Next, we required a duration continuum for each word pair, ranging from step 1 (long /aː/) to step 7 (short /ɑ/). Since in Dutch, the /ɑ-aː/ vowel contrast is cued by both temporal and spectral cues, we wanted to control for spectral information by selecting a (close to) ambiguous (i.e., midway between /ɑ-aː/) value for the first and second formant (F1; F2). We

manipulated the F1 and F2 of the vowel based on Burg's LPC method in Praat [11], and set the F1 to 777 Hz and the F2 to 1456 Hz. These values were shown to be perceptually ambiguous in a pretest. For the duration continuum, we manipulated vowel duration using PSOLA in Praat (203 ms (step 1) - 73 ms (step 7); step size = 21.7 ms; the same durations applied to all word pairs).

We then concatenated three middle steps (steps 3, 4, and 5) and the two unambiguous steps (1 and 7) with the word lists to create the stimuli for the experiments (i.e., 'target lists'). Each target word was linked to one word list combination, and all steps of that target word were concatenated with the monosyllabic and bisyllabic versions of that combination (see OSF). The same procedure was also applied in Experiment 2.

2.3. Procedure

The experiment was built and hosted on the Gorilla Experiment Builder (www.gorilla.sc). Participants first performed a headphone screening [12], ensuring that the majority of participants were likely to be wearing headphones.

The main task was a 2AFC experiment in which participants heard the target lists with the target words in list-final position and were instructed to press a button to indicate which word they had heard (e.g., *stad* vs. *staat*).

On each trial, participants were visually presented with a fixation cross in the middle of the screen for 500 ms. Afterwards, we auditorily presented the target lists and at target word offset, two response options (i.e., the two members of the target word pair) appeared on the screen in the middle left and right. Participants were instructed to respond as quickly and accurately as possible from target word offset with button presses ([Z] or [M] for the left or right option, respectively). Response positions were counterbalanced across participants. The next trial started 1 s after their response or after a timed out trial (3 s). Steps 3-5 were repeated 6 times while steps 1 and 7 were repeated twice, serving only as perceptual anchor points. All steps were presented once with a monosyllabic, and once with a bisyllabic word list resulting in 264 trials.

Experiment 2 was almost identical but differed in two ways. First, the 2AFC task was preceded by a familiarization task in which participants were visually presented with the orthography of the pseudowords and auditorily presented with the pseudowords. Participants could repeat the audio up to five times. Second, in the 2AFC task, we visually presented the orthography of the target lists (containing the pseudowords, but not the target word) during the auditory presentation of the target lists. These two changes were added to increase the possibility that the pseudowords were perceived correctly.

3. RESULTS

We excluded trials without a response (Exp1: 0.8%; Exp2: 0.6%). Participants indeed demonstrated ceiling/floor performance on the 'anchor point' steps 1 and 7 (proportion long /a:/ responses: 0.96 and 0.02 (step 1 and 7, Experiment 1), and 0.97 and 0.03 (step 1 and 7, Experiment 2)). Further analysis was restricted to the critical middle steps 3-5. Categorization curves for Experiment 1 and 2 are presented in Figure 2. We ran two separate Generalized Linear Models (GLMM), one for each experiment, with a logistic linking function using the lmerTest package [13] in R [14]. The models tested the binomial categorization responses (long vowel coded as 1; short vowel coded as 0).

The final model for Experiment 1, as obtained through forward modelling, contained the following fixed factors: Rate (categorical predictor with two levels, deviance coded with monosyllabic coded as -0.5 and bisyllabic coded as 0.5), Continuum step (continuous predictor, scaled to z-scores) and their interaction. Next to random intercepts for Participants and Items, the model further included by-Participant random slopes for Rate and by-Item random slopes for Continuum step. Models with a more complex structure failed to converge.

The model revealed a significant effect of Continuum step (β = -1.37, SE = 0.06, z = -22.67, p < .001), confirming that the shorter vowel durations resulted in fewer long /a:/ responses. Second, the model revealed a significant effect of Rate (β = 0.32, SE = 0.11, z = 2.88, p < .01), illustrating that word lists containing bisyllabic context words resulted in a higher proportion of long /a:/ responses compared to word lists with monosyllabic context words. No significant interaction was found.

The final model for Experiment 2 was identical to that for Experiment 1, except that, next to random intercepts for Participants and Items, the model included by-Participant random slopes for Rate and Continuum step.

The model revealed a significant effect of Continuum step (β = -1.59, SE = 0.15, z = -10.64, p < .001), confirming the same effect of vowel duration on /ɑ/-/a:/ responses as in Experiment 1. Second, the model revealed a significant effect of Rate (β = 0.35, SE = 0.11, z = 3.16, p < .01), again illustrating that word lists with trisyllabic context words lead to more long /a:/ responses. However, the size of this effect was numerically not greater than (instead, seemed

comparable to) the monosyllabic vs. bisyllabic contrast in Experiment 1. Indeed, an omnibus model of the combined data showed no interaction between Rate and Experiment ($\beta = 0.01$, $SE = 0.14$, $z = 0.08$, $p = .93$), demonstrating a lack of evidence for a stronger rate effect in Experiment 2 compared to Experiment 1.

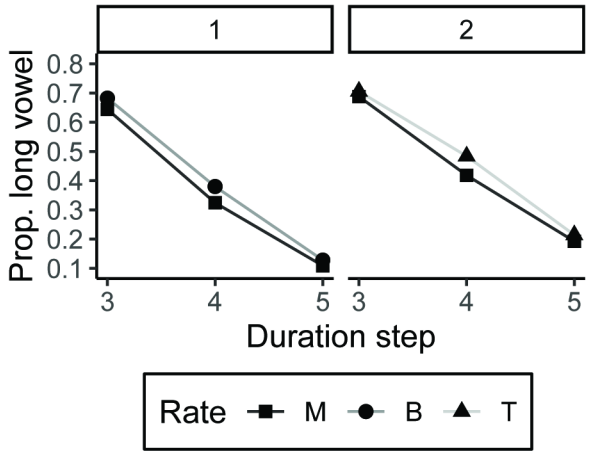


Figure 2: Mean proportion of long /a:/ responses, plotted separately for Experiment 1 (left panel) and Experiment 2 (right panel). Target words were preceded by words lists containing monosyllabic (pseudo)words (M), bisyllabic words (B), or trisyllabic pseudowords (T).

4. DISCUSSION

The present study examined whether the syllable rate is used as a measure of speech rate in an implicit rate normalization task. We found that word lists containing only bisyllabic words led to more /a:/ responses and were thus perceived as faster than duration-matched lists with only monosyllabic words. This illustrates that, despite both words lists having the same stress rate, the number of syllables (i.e., syllable rate) affected vowel perception.

Further, we found that lists containing trisyllabic pseudowords also led to more /a:/ responses compared to lists with monosyllabic pseudowords. However, unexpectedly, the rate effects in Experiment 1 (1:2 syllable ratio) and Experiment 2 (1:3 syllable ratio) were comparable. That is, despite having three times as many syllables in the trisyllabic condition compared to the monosyllabic condition, we did not find a larger rate effect in Experiment 2. This suggests that there is no linear relation between the number of syllables per unit time and the size of the rate effect. These results are in line with Bosker & Ghitza [15], who found that there is an upper limit to the speech rate that can drive rate normalization.

The results of both experiments are consistent with previous explicit rate judgement experiments [6], [7], that found that the number of articulated syllables leads to the perception of a faster tempo. Note that in [6] and [7], the critical comparison was between sentences with or without deletions. Our findings show that, in the absence of a mismatch between the intended and produced number of syllables, listeners also use syllable rate as a measure for speech rate.

An interesting avenue for future research would be to combine the present study with the findings in Koreman [6], Plug et al. [7], and Reinisch [8]. Specifically, if one would compare the bisyllabic lists with new bisyllabic lists in which the vowel of each first syllable is replaced by a reduced vowel (e.g., *kəlom, fəruit, tərein, fəluit*), how would that affect vowel perception? Following [8], the reduced list would be perceived as faster because of the presence of reductions. On the other hand, [6] and [7] would predict the opposite, because the fully articulated bisyllabic words would contain more segments per unit time. Testing such a reduced word list could potentially shed light on the diverging results between [6], [7], and [8].

Moving on to the comparable rate effects in both experiments, the question arises why there is no linear relation between syllable rate and the rate effect. We offer three possible explanations. First, following the conclusions in Bosker & Ghitza [15], syllable rate can only drive speech rate perception to a certain degree. While the present study did not explicitly test where this upper limit lies, it does seem that, with the present stimuli, the 1:2 syllable ratio is close to ceiling. We would predict that if other cues, such as fast speech processes [8] and prosodic cues [7] would be present in the stimuli, we would observe a larger rate effect in Experiment 2. Second, it could be that unstressed syllables carry less weight than stressed syllables in computing speech rate. While unstressed syllables are still important (as is evident from the effect of syllable rate in Experiment 1), listeners might rely more heavily on the stress rate. Third, there is a possibility that the results in the present study could partially be driven by the use of pseudowords in Experiment 2. However, to our knowledge, there is no evidence indicating that non-words would show smaller effects, as rate normalization effects have even been found with non-speech stimuli [1].

In conclusion, the present study provided evidence from an implicit rate normalization task that listeners use syllable rate to compute speech rate. However, syllable rate can only partly account for speech rate normalization, and other cues, such as fast speech processes and prosodic cues, must also play a role in distinguishing different speech rates.

5. ACKNOWLEDGMENTS

This research was funded by the DCC internal round grant, awarded to J.M. and H.R.B. Funding was also received from an ERC Starting Grant (HearingHands, 101040276) from the European Union, awarded to H.R.B. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We would further like to thank Annelies van Wijngaarden whose voice was recorded for the speech materials used in this study.

6. REFERENCES

[1] H. R. Bosker, "Accounting for rate-dependent category boundary shifts in speech perception," *Atten Percept Psychophys*, vol. 79, no. 1, pp. 333–343, Jan. 2017, doi: 10.3758/s13414-016-1206-4.

[2] L. C. Dilley and M. A. Pitt, "Altering Context Speech Rate Can Cause Words to Appear or Disappear," *Psychol Sci*, vol. 21, no. 11, pp. 1664–1670, Nov. 2010, doi: 10.1177/0956797610384743.

[3] M. Maslowski, A. S. Meyer, and H. R. Bosker, "Eye-tracking the time course of distal and global speech rate effects.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 46, no. 10, pp. 1148–1163, Oct. 2020, doi: 10.1037/xhp0000838.

[4] G. R. Kidd, "Articulatory-Rate Context Effects in Phoneme Identification," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 4, pp. 736–748, 1989.

[5] R. S. Newman and J. R. Sawusch, "Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another," *Journal of Phonetics*, vol. 37, no. 1, pp. 46–65, Jan. 2009, doi: 10.1016/j.wocn.2008.09.001.

[6] J. Koreman, "Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 582–596, Jan. 2006, doi: 10.1121/1.2133436.

[7] L. Plug, R. Lennon, and R. Smith, "Measured and perceived speech tempo: Comparing canonical and surface articulation rates," *Journal of Phonetics*, vol. 95, pp. 1–15, 2022, doi: 10.1016/j.wocn.2022.101193.

[8] E. Reinisch, "Natural fast speech is perceived as faster than linearly time-compressed speech," *Atten Percept Psychophys*, vol. 78, no. 4, pp. 1203–1217, May 2016, doi: 10.3758/s13414-016-1067-x.

[9] S. Feldstein and R. N. Bond, "Perception of Speech Rate as a Function of Vocal Intensity and Frequency," *Lang Speech*, vol. 24, no. 4, pp. 387–394, Oct. 1981, doi: 10.1177/002383098102400408.

[10] A. C. M. Rietveld and C. Gussenhoven, "Perceived speech rate and intonation," *Journal of Phonetics*, vol. 15, no. 3, pp. 273–285, Jul. 1987, doi: 10.1016/S0095-4470(19)30571-6.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Feb. 26, 2019. [Online]. Available: www.praat.org

[12] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Atten Percept Psychophys*, vol. 79, no. 7, pp. 2064–2072, Oct. 2017, doi: 10.3758/s13414-017-1361-2.

[13] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "**lmerTest** Package: Tests in Linear Mixed Effects Models," *J. Stat. Soft.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[14] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[15] H. R. Bosker and O. Ghitza, "Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation," *Language, Cognition and Neuroscience*, vol. 33, no. 8, pp. 955–967, Sep. 2018, doi: 10.1080/23273798.2018.1439179.

PAUSE PARTICLES INFLUENCING RECOLLECTION IN LECTURES

Mikey Elmers

Language Science and Technology, Saarland University, Saarbrücken (Germany)
elmers@lst.uni-saarland.de

ABSTRACT

This study investigated the influence of pause-internal phonetic particles (PINTs) on recall for native and non-native listeners of English. Participants were 45 monolingual English and 45 L1 German listeners who heard segments from university lectures, in English, and answered content-based questions. Three versions of lecture stimuli were created: an unmanipulated original version, a "silence" version, and a "no PINTs" version where all PINTs were removed including silences. In the original and "silence" versions, half of the key information was preceded by PINTs material. The results indicated that material preceded by PINTs was less likely to be recalled. Additionally, the participant's first language was not significant for understanding the speaker. However, English listeners tended to score higher during the "no PINTs" condition, while German listeners tended to score higher during the original condition. This study was unable to replicate the recall benefit of PINTs found in single sentence laboratory setting experiments.

Keywords: pause particles, breath noises, memory, lectures, second language education

1. INTRODUCTION

Pause-internal phonetic particles (PINTs) encompass a variety of phenomena such as acoustic-phonetic silence, breath noises (i.e., exhalations and inhalations), filler particles (FPs) like "uh" or "um", and tongue clicks. These phonetic particles can exhibit an influence on recollection. For example, [1] found that the recollection of story plot points was improved when including FPs. In word recognition studies, [2] found that disfluencies improved the recollection of the following word, while [3] found that silent pauses improved the recollection of the following word. Importantly, [3] claims that a feature of disfluencies is that they provide additional time. [4] found that native and non-native listeners exhibited shorter response times for complex phrases that were preceded by FPs or silence compared to a no pause condition. Overall, these studies show that PINTs can affect recollection in laboratory settings. However, these studies do not utilize material from a real-world setting and focus on smaller contexts (i.e., words or sentences). This study expected to find a PINTs benefit for recollection in university lecture segments, similar to the previously mentioned smaller contexts.

Similar to [5], this study does not advocate for 'lab speech' or 'natural speech', rather the goal is to improve awareness around the types of data and methods used. This study explored the influence of PINTs on memory, using real-world data, rather than in a laboratory setting and with material larger than a single sentence. Another main goal was to evaluate the effect of PINTs on both native speakers (NSs) and non-native speakers (NNSs). I opted to examine English monolingual listeners and L1 German listeners due to the English language stimuli used in the study.

2. METHOD

Lectures were collected from Open Yale Courses [6] which contains free and open access courses from Yale University. English-language lectures were chosen based on the speaker's PINTs profile. After selecting a specific speaker, annotations were made for a subset of their lectures. The chosen speaker displayed a relatively high number of PINTs during his lectures, with upwards of 40% of his total time incorporating PINTs material. Fig. 1 shows an example for this speaker.

2.1. Participants

This study used a web-based experiment created with Labvanced [7] to present the audio stimuli to participants, and to collect their answers and questionnaire information. Participants were recruited using the crowd-sourcing platform Prolific [8] and consisted of 45 monolingual English participants (mean age 38 years; age range 21–62 years) and 45 L1 German participants (mean age 35 years; age range 21–72 years) who were paid for their participation. One monolingual English participant reported hearing impairment and was not included in the results.

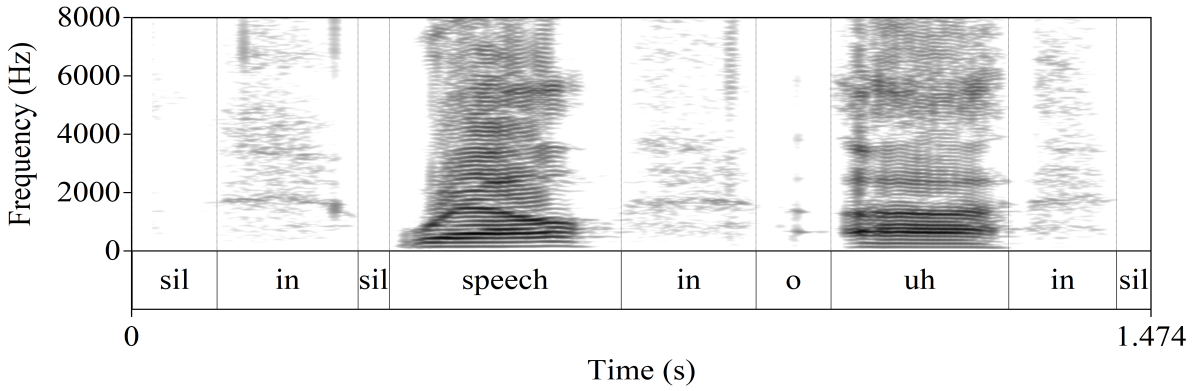


Figure 1: Example section from speaker. Annotations of PINTs: silence (sil), inhalation noise (in), exhalation noise (ex), filler particles uh and um, tongue click (cl), and other (o).

2.2. Stimuli

Stimuli consisted of four three-minute sections extracted from full length lectures. Each audio segment was followed by two multiple-choice content-based questions, with one question preceded by PINTs material and the other not. The study was balanced so that the key material was equally preceded, or *not* preceded, by PINTs. However, neither question was preceded by PINTs material in the "no PINTs" condition, since all PINTs material was removed. An example question was: "According to Paul Fussell, what is the essential trope or rhetorical figure of World War One poetry?" The possible answers were: a) hyperbole, b) metaphor, c) oxymoron, d) irony. The participants did not need to know what these concepts meant, or any other encyclopedic or background knowledge. Instead, they needed to answer based on the content as presented by the lecturer.

The different conditions were created using a Praat [9] script that would remove or replace the PINTs material. In the "silence" condition, non-silence PINTs were replaced with a silence taken from the audio and matched to the duration of the cut material. Therefore, the "silence" condition maintained the same duration as the original audio. The original and "silence" conditions provided the same amount of processing time, while the "no PINTs" condition provided less processing time (see Fig. 2). The "no PINTs" condition did not include any acoustic pause whatsoever. Participants only heard one of the three conditions, i.e., one third of participants heard four original clips, one third of participants heard four "silence" clips, and one third of participants heard four "no PINTs" clips. Each of the conditions included the same textual material, however, the order of the four audio clips was randomized to prevent ordering effects.

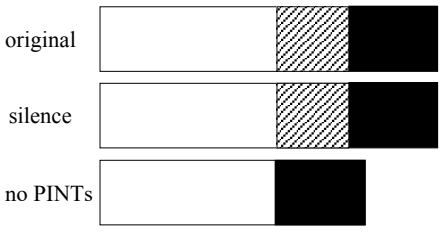


Figure 2: Schematic of the duration for the three conditions showing speech (white), PINTs (grey), and speech material that contained the key information (black).

2.3. Procedure

Participants were informed that they would hear four audio clips, each approximately three minutes, and answer content-based questions immediately following each clip. Participants were instructed to use headphones and test their audio volume before starting. They were told to not take notes. They were also told that the recordings were from a non-ideal microphone and included some background noise. This was in order to draw their attention away from some of the minor artefacts that occurred from the audio manipulation in the "silence" and "no PINTs" conditions. Participants were told that they would

receive a score at the end of the test as an additional incentive to perform well. While listening to the audio, participants saw "Listen closely!" on their screen. They heard each audio clip only once.

After completing the listening section, participants answered a questionnaire that included: age, hearing impairment, L1, self-assessed English skills (for the German listeners), a test score if possible (for the German listeners), highest completed education (high school, university, or other), level of interest in the audio contents (1: very uninterested to 5: very interested), how easy the speaker was to follow and understand (1: very difficult to 5: very easy), and how prepared they found the speaker (1: very unprepared to 5: very prepared). Total completion time was between 15-20 minutes.

3. RESULTS

Participants were scored based on how many questions they answered correctly with a maximum score of 8 (1 point per correct answer). The monolingual English participants scored higher than the L1 German participants in all conditions except in the original condition, however, the monolingual English speakers usually had a higher variance (see Table 1). Monolingual English participants scored highest during the "no PINTs" condition while the L1 German participants scored highest during the original audio condition.

condition	L1	mean	median	sd	n
noPINTs	*EN*	6.26	7	1.83	15
silence	*EN*	6.07	6	1.44	13
original	*DE*	6.00	6	1.07	15
original	*EN*	5.88	6	1.92	16
noPINTs	*DE*	5.87	6	1.41	15
silence	*DE*	5.67	5	1.76	15

Table 1: Mean score, median score, standard deviation score, and count information for the different conditions and L1s.

This project's data and scripts can be found at https://github.com/MikeyElmers/paper_icphs23. The data was pre-processed using the dplyr [10] (Version 1.1.1), stringr [11] (Version 1.5.0), and tidyr [12] (Version 1.3.0) packages. Homogeneity of variance was evaluated using Levene's test from the car [13] (Version 3.1.2) package. Statistical models were analyzed with linear regression and binomial generalized linear mixed models (binomial GLMMs) with the lme4 [14] (Version 1.1.31) and lmerTest [15] (Version 3.1.3) packages in R [16] (Version 4.0.4). Models were compared with the Akaike information criterion (AIC) [17] to calculate unexplained variance. The best fit model was selected as the model with the lowest AIC.

3.1. Preceding PINTs

This study investigated the effect of PINTs material immediately before key information on participant score. These models did not include the "no PINTs" audio condition since all PINTs material was removed. Scores are out of 1 rather than 8 since the evaluation is done on a by-question basis rather than a subject's collective score. Table 2 shows that when key information was preceded by PINTs, the result was an overall lower score. The data showed violations for normality, as indicated by the Shapiro-Wilk test, and homogeneity of variances, as indicated by Levene's test. Therefore, the non-parametric Wilcoxon rank sum test was used. Results indicated a significant difference between the preceding PINTs conditions ($W = 32096$, $p < 0.001$). Participants performed significantly better when critical information was *not* preceded by PINTs information.

preceding PINTs	mean	sd
no	0.81	0.39
yes	0.66	0.47

Table 2: Mean and standard deviation score based on whether the answer was immediately preceded by PINTs material.

Binomial GLMMs were used to evaluate which variables influenced score. The model with the best fit was: $glmer(score \sim precede + (1 \mid id), family = binomial)$. This model predicts score based on the answer being preceded by PINTs information as a fixed effect, and subject with intercept as a random effect. This model performed better than models that incorporated L1, condition, or the questionnaire variables. The analysis revealed a main effect for preceding PINTs ($Estimate = -0.88$, $SE = 0.23$, $z = -3.87$, $p < 0.001$). This main effect indicates that the presence of PINTs material before the answer lowered participants' score.

3.2. Ease

Participant's reported how easy it was to follow and understand the speaker (1: very difficult to 5: very easy). Overall, the mean ease was 2.82. The condition that removed all the pause material was considered the easiest to follow when averaging over

all participants. However, an ANOVA showed no significant differences between conditions ($F(2, 86) = 0.88$, $p > 0.05$). This finding was interesting since substituting or deleting PINTs material created minor artefacts within the audio. The original, unmanipulated version was found to be the most difficult to follow, possibly because this speaker uses a high frequency of PINTs ($\sim$ 40% of his total speaking time). When comparing means of ease by L1, I found that the monolingual English group (M = 3.05 , SD = 1.29) and the L1 German group (M = 2.60, SD = 1.16) were not significantly different ($t(87) = 1.71$, $p > 0.05$, $d = 0.36$). These results indicate that the NNSs found the English-language lecturer as easy to understand as the monolingual NSs of English.

Linear regression models were tested with L1, condition, and the different questionnaire variables. The model with the best fit (lowest AIC) predicted total score with ease as the only fixed effect. Table 3 shows that with an ease value of 1, participants' total score was 4.84 (out of a total of 8) and that the higher the ease value, the higher the total score. Significant effects for all levels of ease were found, except for a value of 3. Importantly, an ease value of 5 improved participants' total score more than an ease value of 4 which improved more than an ease value of 2. The ease value of 3 did not follow this trend.

	Est	SE	t	p-value
$(Intercept)$	4.84	0.42	11.60	$< 0.001 ***$
$ease2$	1.15	0.51	2.28	$< 0.05 *$
$ease3$	0.95	0.54	1.77	0.08
$ease4$	1.54	0.55	2.81	$< 0.01 **$
$ease5$	1.95	0.64	3.09	$< 0.01 **$

Table 3: Summary information of linear model with ease as predictor.

3.3. Correlation

Table 4 contains the Pearson correlations between total score and the questionnaire variables. No correlation was found between age or interest and total score. However, I found a weak correlation between ease and total score ($t(87) = 2.98$, $p < 0.01$, $r = 0.30$), and between preparation and total score ($t(87 = 2.77$, $p < 0.01$, $r = 0.28$).

	age	ease	interest	prep
total score	0.09	0.30	0.10	0.28

Table 4: Pearson correlations between total score and questionnaire responses for all participants.

4. DISCUSSION

This experiment evaluated whether PINTs improved the recall of English-language lecture material for native and non-native listeners. I found that PINTs immediately preceding key information negatively impacted score. While I found that L1 did not influence the ease rating of the speaker, non-native students may encounter significant problems when listening to lectures, such as word recognition or with creating meaning [18]. These issues are related to linguistic aptitude and awareness of the lecture material. This study found that monolingual English listeners tended to scored better in the "no PINTs" situation, while L1 German listeners tended to scored better with the original audio. This may be due to monolingual English listeners being above the threshold of needing the time-buying aspect of PINTs. Conversely, the L1 German listeners still benefited from the time-buying aspect of PINTs. [19] found that a base skill level might be required before the benefits of additional processing time from pauses can be seen, and importantly, that beyond a certain skill level pauses may no longer aid and, instead, be an irritant to the listener. [20] found that pauses may increase comprehension, but only for advanced students. [20] also found that increasing the duration of pauses has a ceiling and once above that ceiling, comprehension will decrease. Therefore, it is important to consider the impact of PINTs in environments where the recall of key information is crucial, such as educational settings, and for both native and non-native listeners.

This study investigated the influence of PINTs on recollection in an ecologically valid scenario with longer material lengths. I found PINTs to be detrimental to the recollection of upcoming content in a lecture scenario, which is contradictory to what previous studies have found. These results might differ from other studies due to the longer material lengths, or because there are many variables that are difficult to control in a real-world scenario. Additionally, this experiment treated all PINTs equally. In real-world scenarios, speakers have distinct PINTs profiles and often many PINTs will co-occur making it difficult, if not impossible, to evaluate individual PINTs separately. Additionally, while non-native participants were asked for their test score, I was not able to get an accurate picture of the influence of PINTs for different skill levels. In this study most of the L1 German listeners were advanced in English. Future work should continue to evaluate longer material lengths, with a variety of language backgrounds and skill levels.

5. REFERENCES

[1] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of memory and language*, vol. 65, no. 2, pp. 161–175, 2011.

[2] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[3] L. J. MacGregor, M. Corley, and D. I. Donaldson, "Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners," *Neuropsychologia*, vol. 48, no. 14, pp. 3982–3992, 2010.

[4] M. Watanabe, Y. Den, K. Hirose, and N. Minematsu, "The effects of filled pauses on native and non-native listeners' speech processing," in *Disfluency in Spontaneous Speech*, 2005.

[5] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[6] "Open yale courses," https://oyc.yale.edu/, accessed: June 6th, 2022.

[7] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified JavaScript framework for online studies," in *International Conference on Computational Social Science (Cologne)*, 2017.

[8] "Prolific," Oxford, UK, 2014, accessed: 12.06.2022. [Online]. Available: https://www.prolific.co

[9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org

[10] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, *dplyr: A Grammar of Data Manipulation*, 2023, r package version 1.1.1. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[11] H. Wickham, *stringr: Simple, Consistent Wrappers for Common String Operations*, 2022, r package version 1.5.0. [Online]. Available: https://CRAN.R-project.org/package=stringr

[12] H. Wickham, D. Vaughan, and M. Girlich, *tidyr: Tidy Messy Data*, 2023, r package version 1.3.0. [Online]. Available: https://CRAN.R-project.org/package=tidyr

[13] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[15] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[17] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *International Symposium on Information Theory*, 1973, pp. 267–281.

[18] J. Kilbon, "L2 student perceptions regarding their comprehension of academic lectures–a longitudinal study," Ph.D. dissertation, University of Leicester, 2022.

[19] E. K. Blau, "The effect of syntax, speed, and pauses on listening comprehension," *TESOL quarterly*, vol. 24, no. 4, pp. 746–753, 1990.

[20] G. Jacobs, W. Chuawanlee, B. K. Itoga, D. Sakumoto, S. Saka, and K. A. Meehan, "The effect of pausing on listening comprehension," in *The Eighth Second Language Research Forum*. Honolulu, HI: ERIC, 1988, pp. 1–19.

DIFFERENTIAL PERCEPTION OF VOWEL AND STOP CONTRASTS IN SPANISH-ENGLISH BILINGUAL PRESCHOOLERS

Simona Montanari[a], Jeremy Steffman[b], Robert Mayr[c], Patricia Flores[a] & Carissa Feeney[a]

[a]California State University, Los Angeles, [b]University of Edinburgh, [c]Cardiff Metropolitan University

ABSTRACT

Despite extensive work on perceptual development in bilingual infants, we know little about speech perception in bilingual preschoolers. This study investigates English vowel and stop voicing perception in 20 Spanish-English bilingual preschoolers and 20 English monolingual peers. Perception was assessed through a forced-choice minimal-pair identification task in which children heard synthesized audio stimuli that varied systematically along an /i-ɪ/ vowel continuum and a /b-p/ Voice Onset Time (VOT) continuum and were asked to match them with one of two pictures for each contrast. The results of Bayesian mixed-effects logistic regression analyses indicated no credible difference in vowel perception between monolinguals and bilinguals. In contrast, the bilinguals' category boundary for English /b-p/ was impacted by their experience with Spanish, with more short-lag VOT tokens being perceived as voiceless in line with Spanish VOT. We interpret the observed asymmetry as stemming from differences in L1-L2 mappings across the contrasts.

Keywords: speech perception; Spanish-English bilinguals; preschoolers; voice onset time; vowels

1. INTRODUCTION

Despite an abundance of research on speech perception in simultaneous bilingual infants (see e.g. [1]) and adult second/foreign language learners (e.g. [2]), few studies have examined how young bilinguals perceive speech sounds in the preschool years, when important preliteracy skills whose emergence is dependent on speech perception abilities are being developed [3]. McCarthy et al. [4], who assessed the perception of the English voicing contrast in bilabial and velar stops in Sylheti-English bilingual children, found that English perception patterns at preschool entry (i.e., at age 4;4) were different from monolinguals' and appeared affected by children's existing Sylheti phonemic categories. However, a year later, the bilingual children's productions were no longer significantly different from those of their monolingual peers, suggesting more refined phonemic categories. The authors speculated that phonemic categorization may be initially affected by language dominance in young bilinguals. However, phonemic categories in the other language can be acquired and refined with language experience. The limited research on speech perception in sequential bilingual children confirms that even in the case of early exposure to a second language (L2) (e.g., in infancy or early childhood), native-like perception may take years to develop. Darcy and Krüger [5], who tested the perception of four different German vowel contrasts in 10-year-old Turkish/German bilingual children who began to be exposed to German between age 2 and 4, found that the children had difficulties with the discrimination of those contrasts that were perceptually similar for Turkish monolingual speakers. These findings were interpreted as evidence of perceptual assimilation of German phonemes to similar Turkish phonemes even in the case of early bilingualism. Similarly, Tsukada et al. [6], who compared the perception of 4 English vowel contrasts in Korean-English bilingual school-age children who had spent up to 6 years in the United States, found that the children discriminated English vowels more accurately than Korean-English bilingual adults but less accurately than monolingual English-speaking children. Overall, these studies suggest that even early and intensive exposure to two languages may not be enough to prevent phonemic representations from one language influencing speech perception in the other [7]. However, it is unclear whether some L2 sound contrasts are more difficult to perceive than others for bilingual preschoolers

This study expands the limited literature on bilingual speech perception at preschool age by examining L2 speech perception on two types of contrast, a vowel contrast and a stop voicing contrast, in Spanish-English bilingual children. We ask: 1) To what extent does vowel or VOT continuum step predict categorization responses? 2) To what extent do Spanish-English bilingual children and age-matched English monolingual children differ in their categorization responses on the English /i-ɪ/ vowel continuum? 3) To what extent do they differ in their categorization responses on the English /b-p/ VOT continuum?

2. METHOD

2.1. Participants

The participants are 20 Spanish-English bilingual (4 males and 16 females, mean age 4;9) and 20 English monolingual (12 males and 8 females, mean age 4;6) children growing up in Los Angeles, California, matched in age ($t(38)=1.78$, $p=.084$). The bilingual children are raised in homes where they have regularly and consistently heard both languages from early in life, although they are exposed to English more overall and through a larger variety of speakers/sources as compared to Spanish, which they mainly hear from family members. On that basis, they were classed as English-dominant. The monolinguals hear mainly English from their input providers, but due to the bilingual nature of the community in which they live, they also have limited exposure to Spanish. Nevertheless, based on the information provided they fit the description of "functional monolinguals" with no active use or knowledge of Spanish [8]. The children were recruited and tested by trained Child Development majors as part of their coursework for a language development course at a public, 4-year urban university.

2.2. Stimuli

Two phonetic continua were created, manipulating vowel formants and vowel duration, and VOT alongside other cues to voicing. These continua ranged between "sheep" and "ship" for the /i-ɪ/ contrast, which is known to pose difficulties for Spanish monolinguals [9,10], and "penny" and "Benny" for the /p-b/ contrast. To obtain initial values, natural tokens of the target words were recorded by a female native speaker of American English.

Manipulation of the vowel and voicing contrast continua was carried out using a Praat scripts. For the vowels, we employed the Burg Method of LPC decomposition and resynthesis ([11], www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt). The formant values for each endpoint of the continuum were based on the naturally produced formants for /i/ and /ɪ/. The resynthesis process estimated source and filter models for one endpoint (the word with /i/). The filter model's F1, F2, and F3 were then interpolated linearly (in Bark space, [12]) to the values of the other continuum endpoint, resulting in 10 intermediate filter steps. Phase-locked higher frequencies from the starting base file (/i/) that were lost in the process of LPC resynthesis were restored to all steps, improving the naturalness of the continuum. The result was a 10-step continuum ranging from /i/ to /ɪ/. We next manipulated vowel duration to co-vary with F1, F2 and F3 based on the speaker's natural productions, whereby /i/ was longer than /ɪ/ (in agreement with large-scale studies measuring vowel duration, [13]). The duration manipulation thus interpolated between endpoint duration values in 10 linearly equidistant steps in the way that duration and vowel formants would normally co-vary (e.g., the /i/ endpoint had the longest duration, the /ɪ/ endpoint had the shortest duration). For the stop voicing contrast, VOT, F0 and the duration of the following vowel were manipulated over a total of 10 steps [14]. VOT ranged from 0 (short-lag VOT) to 90 ms (long-lag VOT). The duration of the following vowel was shortened incrementally as VOT was lengthened. F0 also varied along the continuum, showing a localized increase following longer VOT (+30 Hz for the longest continuum step relative to the base file), which interpolated to a slight dip (-5 Hz relative to the base file). The duration of the F0 perturbation was set to be 75 ms, which is a typical range for voicing-induced F0 changes in vowels [15].

In order to reduce the number of trials in an attempt to minimize fatigue effects, we selected 8 steps by excluding step 2 and step 9 for both continua (keeping the 6 most central steps in the continuum and the endpoints).

2.3. Procedures

We created a child-friendly forced-choice minimal-pair picture identification task in each language in which children heard an auditory stimulus that varied systematically along the vowel or VOT continuum and were asked to match it with one of two pictures representing a minimal pair ("sheep" and "ship" for /i-ɪ/ and "penny" and "Benny" for /p-b/). In the experiment, they were exposed to the 8 tokens from each contrast in random order alongside other materials not reported here, for a total of 24 trials. Children were familiarized with the stimuli one week before the experiment and had to show they knew the words in order to participate.

2.4. Analyses

We analyzed the data with Bayesian mixed-effects logistic regression analyses of categorization data (carried out using *brms* [16]). We analyzed the stop and vowel data separately. In each model we predicted binomial responses (/p/ and /i/ mapped to 1 in the stop and vowel model, respectively). This was predicted as a function of continuum step (scaled), gender (female mapped to -0.5, male to 0.5), a background variable that some literature has found to be predictive of bilingual development (see [17] for a review), and language background (bilingual mapped

to -0.5, monolingual to 0.5). A quadratic term for continuum step was also included to allow the model to capture potentially larger effects in the middle of the continuum (as compared to either end, as with the linear continuum step term). All interactions between fixed effects were included as well. The random effects consisted of a random intercept for speaker and by-speaker slope for continuum step. We used weakly informative normal priors for both intercept and fixed effects: normal (0,1.5) in log-odds space. In reporting effects, we give the 95% credible interval (CrI), which indexes the range in which 95% of the estimated posterior falls. When this interval *excludes* the value of zero it indicates a robust effect: that is, a non-zero effect estimate with a consistent directionality. We also report the probability of direction "pd" [18], which is more intuitively interpreted as strength of evidence for an effect. This metric gives the percentage of a distribution with a given positive/negative sign. It ranges from 50 (distribution centered on 0, no effect) to 100 (strong evidence for an effect). Pd values in excess of 97.5% correspond to 95% CrI excluding 0, which we take as a reliable effect [19]. Pd values approaching this threshold constitute weaker evidence for an effect. Full model summaries are given in Table 1.

We excluded participants who did not show sensitivity to the acoustic continua we used. This was done by running individual-level regression analyses predicting each participants' categorization as a function of continuum step. We excluded participants who showed a zero estimate (no effect of step, i.e. flat categorization across all steps) or a negative estimate (the opposite of the predicted effect given the coding of variables, i.e. more /b/ responses for longer VOT, or more /ɪ/ responses with /i/-like formants and duration). Either of these indicates potential inattention to the task, or lack of perception of the contrast at all. This was done on a by-contrast basis, leaving 17 bilinguals and 18 monolinguals for the /b/~/p/ contrast and 19 bilinguals and 18 monolinguals for the /ɪ/~/i/ contrast.

3. RESULTS

3.1. Stop results

The model of stop voicing perception found an expected effect of (linear) *VOT continuum step* (β = 2.57, 95%CrI = [1.82,3.43], pd = 100), showing that listeners increase /p/ responses with increasing VOT. There was not a main effect of *gender* (pd = 63), but crucially, there was a main effect of *language background* (β = -1.12, 95%CrI = [-2.27,-0.00], pd = 98), showing that monolinguals show decreased /p/ responses overall (Fig. 1A). We can consider this in

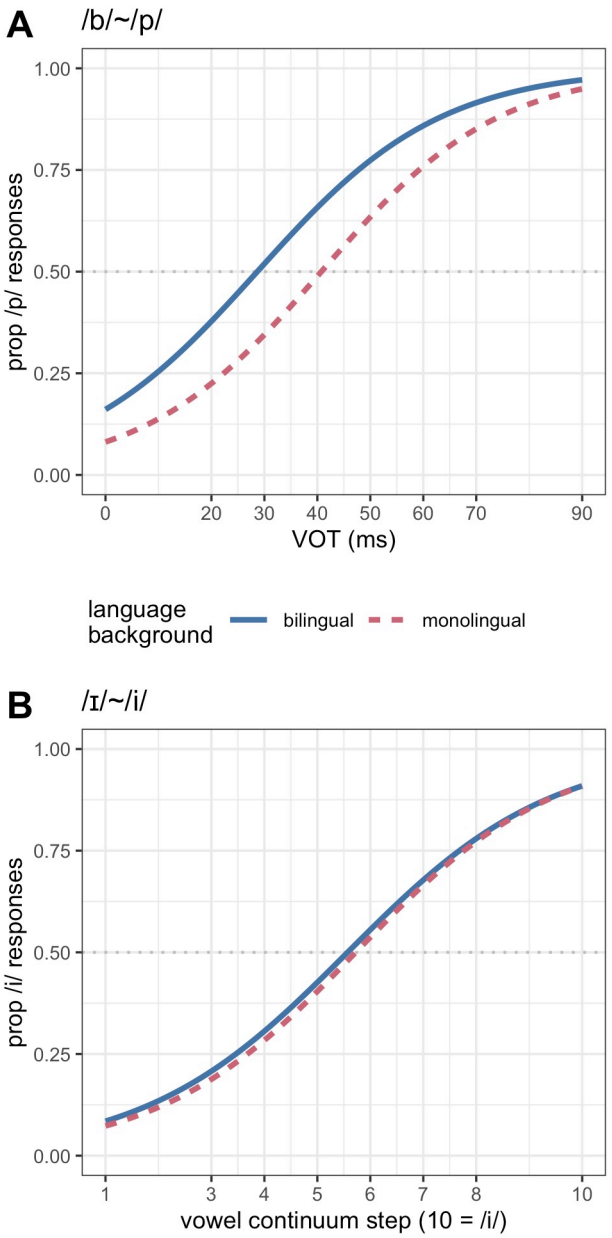


Figure 1: Categorization for the /b/~/p/ continuum (A) and /ɪ/~/i/ continuum (B), along the continuum (x axis). Lines are logistic fits to the data. Line type and coloration shows language group. The dotted horizontal line placed at 0.50 on the y axis represents the category boundary.

terms of the category boundary for each group, operationalized as the point on the continuum at which listeners give 50% /p/ responses. The category boundary is shifted numerically lower for bilingual speakers (about 30 ms) as compared to monolinguals (about 40 ms). Put differently, bilinguals require overall shorter VOT to perceive /p/, and perceive short-lag VOT as /p/ more often. Neither the interaction between *gender* and *continuum step*, nor *language background* and *continuum step* was credible pd = 63 and pd = 88, respectively).

3.2. Vowel results

The model of vowel perception found an expected effect of the *formant continuum* (β = 2.24, 95%CrI = [1.78, 3.21], pd = 100), showing that listeners' increase /i/ responses with changes in formants and vowel duration. There was not a main effect of *gender* (pd = 84). Unlike with the stop voicing model, there was no effect of *language background* (pd = 55), as shown in Figure 1B. The interaction between *gender* and *continuum step*, and *language background* and *continuum step* were not credible (pd = 94, and pd = 68 respectively). The weaker evidence for the former interaction was inspected via the *estimate slopes* function in [20], showing a larger effect of step for male speaker versus female speaker (male β = 2.86, 95%CrI = [1.93, 4.06]; female β = 1.93, 95%CrI = [1.26, 2.82]).

stop model	β	error	95%CrI
Intercept	0.79	0.32	[0.19,1.44]
continuum step	2.57	0.42	[1.82,3.47]
language backgr.	-1.12	0.57	[-2.27,-0.00]
gender	0.15	0.57	[-0.98,1.28]
step:lang backgr.	0.67	0.57	[-0.46,1.79]
step:gen	0.24	0.67	[-1.05,1.60]
gen:lang backgr.	1.54	0.92	[-0.29,3.35]
gen:step:lang backgr.	0.04	1.06	[-2.07,2.09]
step.quad	-0.02	0.27	[-0.54,0.52]
step.quad:lang backgr.	0.25	0.49	[-0.70,2.78]
step.quad:gen	1.70	0.54	[0.70,2.78]
vowel model	β	error	95%CrI
Intercept	-0.24	0.40	[-1.04,0.53]
continuum step	2.24	0.36	[1.78,3.21]
language backgr.	-0.12	0.76	[-1.62,1.37]
gender	0.75	0.76	[-0.75,2.26]
step:lang backgr.	-0.28	0.61	[-1.49,-0.91]
step:gen	0.94	0.61	[-0.23,2.17]
gen:lang backgr.	0.31	1.23	[-2.09,2.74]
gen:step:lang backgr.	0.02	1.09	[-2.18,2.10]
step.quad	0.23	0.26	[-0.28,0.76]
step.quad:lang backgr.	-0.47	0.47	[-1.42,0.44]
step.quad:gen	-0.06	0.51	[-1.05,0.95]

Table 1: Model summaries for the stop model (top) and vowel model (bottom). Estimates and errors are given with 95% CrI. "step.quad" indicates the quadratic term for continuum step.

4. DISCUSSION AND CONCLUSION

This study sought to extend our understanding of speech perception abilities in bilingual preschoolers. To this end, we investigated the perception of the /i-ɪ/ vowel contrast and the /b-p/ stop voicing contrast in Spanish-English bilinguals aged 4-5 years, and compared it to that of their English monolingual peers. The results revealed an asymmetry across the two contrasts, with monolinguals and bilinguals differing in their perception of the stop voicing contrast, but not the vowel contrast. The bilinguals categorized shorter VOT tokens as voiceless than the monolinguals and hence the two sets of preschoolers differ in their category boundary between voiced and voiceless bilabial stops. The preschoolers' gender, in turn, did not affect categorization responses on either contrast. How can we explain these findings?

To begin with, the direction of the category boundary difference on the VOT continuum suggests an influence of the bilinguals' first language (L1) as Spanish contrasts prevoiced and short-lag VOTs [21]. These results are in line McCarthy et al.'s findings [4]. However, in their study, the children were L1-dominant when interactions surfaced in L2 English VOT patterns. In contrast, in the present study, the bilingual preschoolers were dominant in L2 English, and hence, crucially, what this study adds is that cross-linguistic interactions in speech perception are not limited to bilinguals' weaker language.

At the same time, the results also suggest that L1-L2 interactions do not occur across the board since the bilinguals' perception of the English-specific /i-ɪ/ vowel contrast matched that of their monolingual peers, with identical location of the category boundary. What may have caused interactions on the VOT continuum, but not the vowel one, is that the former involves a difference in phonetic implementation for the *same* phonological contrast. On the other hand, the vowel contrast involves acquisition of the novel L2 category /ɪ/, which requires perceptual attunement to its temporal and spectral properties. As such, it is a cross-linguistically more dissimilar sound than the two stops, which according to the *Speech Learning Model* (SLM, [22]) and its recent revision, the SLM-r ([23]) makes it comparatively easier to acquire – in particular since early bilinguals are more likely to establish a novel L2 category than late bilinguals according to the model.

Taken together, this paper extends the limited literature on speech perception in bilingual preschoolers by documenting that subtle differences from age-matched monolinguals may occur on cross-linguistically similar contrasts and may even manifest in the dominant language. At the same time, despite these differences, the bilinguals managed to differentiate both contrasts successfully, and hence the data reported here do not raise any concerns in terms of the development of the children's preliteracy skills. Future work is needed that investigates which perceptual dimensions are particularly prone to L1-L2 interactions in this population, and to what extent increased use of the L2 may shift L1 perceptual boundaries.

5. REFERENCES

[1] Fennell, C.T., Tsui, A.S-M., & Huddon, T.M. 2016. Speech perception in simultaneous bilingual infants. In: Nicoladis, E., & Montanari, S. (eds), *Bilingualism across the Lifespan: Factors Moderating Language Proficiency*. American Psychological Association, 43-62.

[2] Ingvalson, E.M., Ettlinger, M., & Wong, P.C.M. 2014. Bilingual speech perception and learning: A review of recent trends. *Int. J. Biling.* 18, 35–47.

[3] Nittrouer, S., Burton, L. 2005. The role of early language experience in the development of speech perception and phonological processing abilities: Evidence from 5-year-olds with histories of otitis media with effusion and low socioeconomic status. *J. Commun. Disord.* 38, 29–63.

[4] McCarthy, K., Mahon, M., Rosen, S., & Evans, B.G. 2014. Speech perception and production by sequential bilingual children: A longitudinal study of voice onset time acquisition. *Child. Dev.* 85, 1965-1980.

[5] Darcy, I. & Krüger, F. 2012. Vowel perception and production in Turkish children acquiring L2 German. *J. Phon.* 40*,* 568-581.

[6] Tsukada, K., Birdsong, D., Bialystok, E. et al. 2005. A developmental study of English vowel production and perception by native Korean adults and children. *J. Phon.* 33, 263–290.

[7] Simonet, M. 2010. Dark and clear laterals in Catalan and Spanish: Interaction of phonetic categories in early bilinguals. *J. Phon. 38*, 663–678.

[8] Best, C.T. & Tyler, M.D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, O.-S. & Munro, M.J. (eds), *Language Experience in Second Language Speech Learning. In Honor of James Emil Flege*. John Benjamins, 13–34.

[9] Morrison, G.S. 2008. L1-Spanish speakers' acquisition of the English /i/-/ɪ/ contrast: Duration-based perception is not the initial developmental stage. *Lang Speech* 51, 285-315.

[10] Morrison, G.S. 2009. L1-Spanish speakers' acquisition of the English /i/—/ɪ/ contrast II: Perception of vowel inherent spectral change. *Lang Speech* 52, 437-62.

[11] Winn, M.B. (2019). GUI-based wizard for creating realistic vowel formant continua (v. 38). http://www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt

[12] Traunmüller, H. 1990. Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.* 88, 97-100.

[13] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099-3111.

[14] Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *J. Acoust. Soc. Am.*147, 852-866.

[15] Hombert, J. M. (1975). *Towards a theory of tonogenesis: An empirical, physiologically and perceptually based account of the development of tonal contrasts in languages*. University of California, Berkeley.

[16] Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5), 1-54.

[17] Iglesias, A., & Rojas, R. 2022. Bilingual language development of English learners: Estimating the growth of two languages. In: Goldstein, B. (ed.), *Bilingual Language Development and Disorders in Spanish-English Speakers.* Brookes, 9-44.

[18] Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, 4(40), 1541.

[19] Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in psychology*, *10*, 2767.

[20] Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). *Estimation of Model-Based Predictions, Contrasts and Means.* CRAN.

[21] Cho, T., & Ladefoged, P. 1999. Variation and universals in VOT: Evidence from 18 languages. *J. Phon.* 27, 207–229.

[22] Flege, J.E. 1995. Second language speech learning: Theory, findings, and problems. In: Strange W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. York Press, 233–272.

[23] Flege, J.E., & Bohn, O.S. 2021. Flege, J.E. & Bohn, O.S. 2021. The revised Speech Learning Model (SLM-r). In: Wayland, R. (ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge University Press, 3–83.

GENERALIZED PERCEPTUAL ADAPTATION TO SECOND-LANGUAGE (L2) SPEECH: VARIABILITY, SIMILARITY AND INTELLIGIBILITY

Ann R. Bradlow[1], Adrianna M. Bassard[2], and Ken A. Paller[2]

[1]Department of Linguistics, [2]Department of Psychology, Northwestern University, Evanston, IL, USA
abradlow@northwestern.edu, AdriannaBassard2022@u.northwestern.edu, kap@northwestern.edu

ABSTRACT

Recent work suggests that generalization of perceptual adaptation to L2 speech depends on similarity between training and test talkers as well as on high-variability training. We explored this proposal for both talker- and accent-independent adaptation with a round-robin design involving multi-talker testing following single-talker (ST) and multiple-talker (MT) training. Test sentences were produced by four L2 English talkers from different L1 backgrounds. Training involved four single-talker (ST) and four multiple-talker (MT) conditions with all talkers serving as both test and training talkers. On average, greater intelligibility improvement resulted from MT than ST training; however, some training-test talker combinations also showed generalized adaptation following ST training. Critically, variation in training talker baseline intelligibility related to variation in adaptation and generalization. Together with prior work, these data suggest that generalized adaptation to L2 speech is sensitive to a combination of high-variability training, training-test talker similarity, and baseline intelligibility of the training talker(s).

Keywords: perceptual adaptation, speech intelligibility, L2 speech.

1. INTRODUCTION

Both human and computer listeners often struggle to recognize speech by second-language (L2) talkers [1-4]. However, prior work has shown substantial exposure-induced improvement in L2 speech recognition for human listeners (for a review see, [5]) indicating remarkably flexible human speech perception in response to extensive variation in speech production. A key issue for this research is to identify the exposure conditions that lead to the most extensive generalization. Optimal adaptation would generalize from a small set of training stimuli to improved speech recognition accuracy for speech by novel (i.e., untrained) L2 talkers from both trained and novel accents. In addition to the benefits for enhancing speech communication across a language barrier, this research enterprise provides a window into the processes and representations that support perceptual constancy in the face of extensive speech variation.

One approach emphasizes exposure to variability during training as a means of directing listener attention toward systematic and away from random variation. Research in this approach has successfully demonstrated cross-talker [6-8] and cross-accent [9,10] generalization after exposure to multiple talkers of a single accent (e.g., Chinese-accented English) or multiple talkers from different accent groups (e.g., Chinese-accented, Spanish-accented, Korean-accented, and Hindi-accented English), respectively. In contrast, learning following exposure to a single talker or accent was limited to the trained talker or accent, respectively [e.g., 6].

Other work has emphasized similarity between training and test stimuli rather than exposure to variability [11-13]. For example, [13] demonstrated cross-talker generalization following both multiple- and single-talker training with a design that involved 20 unique training-test talker combinations with all talkers appearing in both training and test phases. This design effectively disentangled training format (single-talker versus multiple-talker) from the specific training and test talkers, and thereby revealed variation in generalization across different training-test talker pairs with some single-talker conditions leading to as much cross-talker/within-accent generalization as multiple-talker training. This finding is consistent with a similarity-based account according to which generalization of adaptation beyond training stimuli depends on sufficient overlap between training and test talkers in their patterns of phonetic realization of phonological contrasts [11-13]. In this view, any observed benefit of high variability over low variability training is not due to exposure to variability per se; instead, the benefit of training set variability is the increased likelihood of exposure to training stimuli that are sufficiently similar to the test stimuli to facilitate cross-talker generalization.

The present study explored the variability-based and similarity-based approaches for both cross-talker and cross-accent generalization with a design that like [13] and [10] but unlike [9] disentangled training format (single-talker, ST, versus multiple-talker, MT) from the specific talkers presented in the training and test phases. Moreover, like [9] and [13] but unlike

[10] the present study investigated both talker-general and accent-general adaptation with sentences rather than words as the training and test stimuli. This design feature is motivated by the fact that phrase-level and other supra-segmental and supra-lexical aspects of phonetic structure are salient, language-general features of naturally produced L2 speech [14, 15, 16] and therefore potentially important cues for perceptual adaptation particularly across accents.

2. METHOD

The overall design for this study involved a training phase followed by a test phase with an 11-12-hour delay between training and test. This delay was in anticipation of a subsequent study (not reported here) involving sleep consolidation. Eight different training conditions (between-participants) were followed by an identical multiple-talker, multiple-accent sentence-in-noise recognition test. A round-robin arrangement of training and test talkers allowed us to compare perceptual adaptation to L2 English across various training-test talker combinations (n=16) following both single-talker (ST) and multiple-talker (MT) training. Importantly, the four talkers involved in the round robin all came from different L1 backgrounds, allowing examination of both talker-specific and talker-general/accent-general adaptation to L2 speech.

2.1. Participants

A total of 195 first-language (L1) American English listeners participated in this study. All participants were between 18-35 years and self-reported as having no deficits in speech, language, or hearing, and as having normal or corrected-to-normal vision.

2.2. Materials and procedure

A total of 300 sentence recordings were downloaded from an open-access corpus of L2 speech that includes recordings from over 100 L2 talkers from over 20 L1 backgrounds [17]. For the present study, we compiled a set of simple sentences (e.g., “A towel is near the sink.”) from four L2 talkers from four different L1 backgrounds (75 sentences per talker), two males (L1 Brazilian Portuguese and L1 Spanish), and two females (L1 Farsi and L1 Turkish). These talkers were selected based on informal, subjective judgements by the authors as clearly L2-accented with moderate-to-good comprehensibility.

In both the training and test phases, participants listened over headphones or earbuds to sentence recordings that had been digitally mixed with speech-shaped noise at a fixed signal-to-noise ratio of 0 dB (i.e., the speech and noise were presented with equal loudness). The sentences were presented one at a time with no possibility of repetition. Participants typed what they heard using the computer keyboard before advancing to the next sentence. No feedback was provided in either the training or test phase.

At test, participants were presented with 15 sentences by each of the four talkers (total = 60 sentences). The talker-sentence pairings and trial order were held constant across all training conditions. Participants were randomly assigned to one of eight training conditions, four single-talker (ST) and four multiple-talker (MT) (n=20-22 per condition). All training conditions included the same 60 sentences (different from the test sentences). An additional untrained control group (n=27) took the multiple-talker test without any prior training phase.

In the ST training conditions, all 60 sentences were produced by one of the four talkers. Thus, at the test phase, listeners in ST training conditions encountered one trained talker and three novel talkers. In the MT training conditions, three of the four talkers produced 20 sentences each, while the fourth talker was excluded from the training set. Thus, at the test phase, listeners in MT training conditions encountered three trained talkers and one novel talker.

All sentence transcriptions were scored using an open-source automated scoring tool, Autoscore [18], which counts a sentence as correctly (score=1) or incorrectly (score=0) recognized if and only if the transcription exactly matches the talker’s script. Obvious spelling errors or homophones of intended words counted as correct. Sentence-level rather than word-level intelligibility was the dependent variable (DV) in all analyses because words in sentences are not equally independent of each other (due to sentence-level structure). In keeping with prior work, we chose intelligibility (word recognition accuracy) as the DV rather than comprehensibility or accentedness (see [19] for discussion on this point).

For analysis, proportional scores were log-odd transformed: $\log((p)/(1-p))$ where p is the intelligibility score ranging from 0 to 1. Separate ST and MT analyses were conducted because of the different balance of trained versus novel talkers encountered in the test following ST versus MT training (1 trained and 3 novel for ST versus 3 trained and 1 novel for MT). Within each analysis (ST or MT), t-tests with Bonferroni correction for multiple comparisons (n=16) compared performance at test for each training talker-test talker pairing relative to that test talker’s baseline intelligibility (control condition). Additionally, relative entropy (Kullback-Leibler divergence) of the log-odd transformed test score means was calculated for each training condition (n=8) with respect to the control condition. This information theoretic metric is a non-zero

number that quantifies "informativity" (or suprisal) of the distribution of scores across the four test talkers relative to the untrained control condition. Relative entropy of 0 would indicate no divergence from control to test (i.e., no adaptation); while, higher and lower relative entropy values indicate more or less divergence from baseline, respectively (i.e., more or less improvement in speech recognition accuracy relative to untrained controls). The equation for relative entropy is shown in (1), where P(x) is average intelligibility score for a given test talker, X is the number of test talkers within each training condition (4), and Q(x) is the given test talker's baseline intelligibility score.

(1) $$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

3. RESULTS

	Test talker			
	BRP	FAR	SPA	TUR
Control	0.784 (0.022)	0.722 (0.025)	0.712 (0.021)	0.545 (0.025)
Single-talker training (ST)				
BRP	**0.843 (0.023)**	0.725 (0.032)	0.727 (0.03)	0.572 (0.032)
FAR	0.886 (0.016)	**0.778 (0.023)**	0.767 (0.019)	0.599 (0.02)
SPA	0.852 (0.024)	0.723 (0.035)	**0.725 (0.035)**	0.581 (0.034)
TUR	0.794 (0.026)	0.676 (0.039)	0.708 (0.029)	**0.566 (0.031)**
Multiple-talker training (MT)				
noBRP	**0.854 (0.024)**	0.76 (0.025)	0.727 (0.029)	0.626 (0.025)
noFAR	0.827 (0.021)	**0.749 (0.035)**	0.714 (0.033)	0.609 (0.034)
noSPA	0.817 (0.028)	0.739 (0.033)	**0.727 (0.032)**	0.566 (0.031)
noTUR	0.878 (0.014)	0.794 (0.016)	0.778 (0.021)	**0.617 (0.024)**

Table 1: Average proportion of sentences correctly recognized for each test talker in untrained control, single-talker (ST), and multiple-talker (MT) training conditions. For ST, talker-specific ("old") trials are in bold. For MT, novel talker trials are in bold. Std. error in parentheses.

Table 1 shows average sentence recognition accuracy across all participants in the untrained control condition, the four ST conditions, and the four MT training conditions. The speech recognition accuracy scores are broken down by test talker (columns). For the ST training format, trials with matched training and test talkers are shown in bold. L1 codes for the ST training and test talkers are BRP = Brazilian Portuguese, FAR = Farsi, SPA = Spanish, TUR = Turkish. For MT training, condition codes indicate the excluded talker (all three of the other talkers were included): noBRP = Brazilian Portuguese excluded, noFAR = Farsi excluded, noSPA = Spanish excluded, noTUR = Turkish excluded. Test trials with the novel talker are shown in bold.

Figure 1 shows sentence recognition accuracy for all ST (top) and MT (bottom) training conditions by test talker. In each plot, the line indicates baseline intelligibility for the test talkers (untrained control condition). Boxes match bold entries in Table 1, i.e., indicate talker-specific ('old' talker) for ST and talker-generalization ('new' talker) for MT.

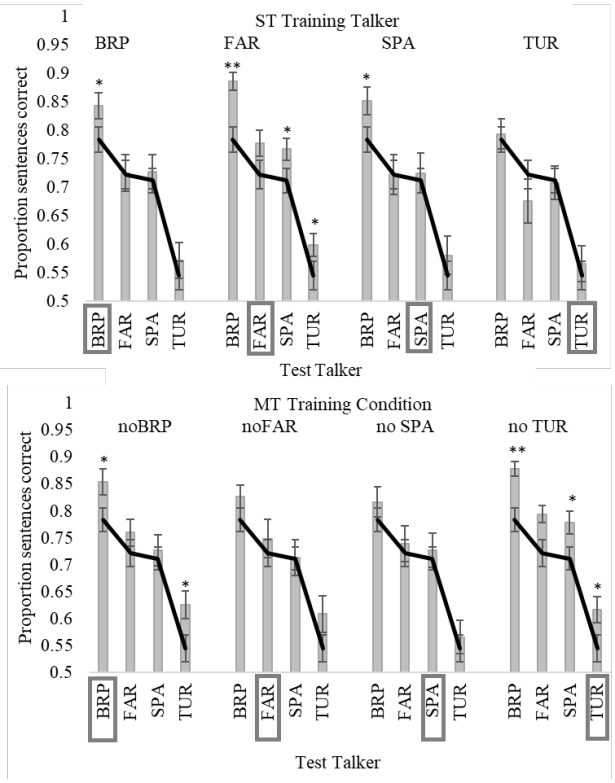


Figure 1: Sentence recognition accuracy for all ST (top) and MT (bottom) training conditions. Error bars are SE. Lines indicate baseline intelligibility for each test talker. Boxes indicate 'old' talker trials for ST and 'new' talker trials for MT. For pair-wise comparisons, * = p<.05, ** = p<.003 (Bonferroni correction for 16 comparisons).

Within the ST training format we see variation across the 16 combinations of training and test talkers. Significant training-to-test improvement (at the Bonferroni corrected level of p<.003) emerged for one test talker, BRP, following ST training with one talker, FAR. All other differences from baseline intelligibility failed to reach significance at this level although several others were significant at the less conservative level of p<.05 (FAR training for SPA and TUR test talkers, SPA training for BRP test talker, and BRP training for BRP test talker). Similarly, within the MT training format, we see

substantial variation across the 16 combinations of training condition and test talker. Pair-wise comparisons showed significant training-to-test improvement (at the Bonferroni corrected level of $p<.003$) for the BRP test talker following the MT training condition that excluded the TUR talker. All other differences from baseline intelligibility failed to reach significance at this level although several others were significant at the less conservative level of $p<.05$ (noTUR training for SPA and TUR test talkers; noBRP training for BRP test talker).

Figure 2 shows relative entropy for each training condition grouped by training format (ST vs. MT). Each data point represents overall divergence of test from baseline scores for the four test talkers in the indicated training condition. Lines connect ST and MT training conditions that include (ST) or exclude (MT) the talker indicated, e.g., the FAR-noFAR line connects the ST condition with FAR as the training talker and the MT condition that excluded this talker, noFAR. The overall pattern reveals slightly higher average relative entropy following MT than ST training. However, the individual training conditions vary in the direction of change from ST to MT. In particular, relative entropy decreases for the MT condition that excludes the "best" ST trainer (FAR to noFAR) and increases for the MT condition that excludes the "worst" ST trainer (TUR to noTUR).

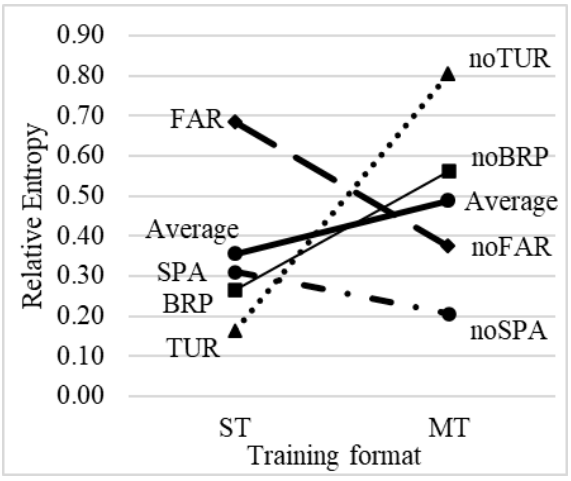


Figure 2: Relative entropy (Kullback-Leibler divergence) for each training condition grouped by training format (ST vs. MT). See text for more explanation.

4. DISCUSSION

The present study revealed exposure-induced cross-talker, cross-accent improvement in L2 English recognition by L1 English listeners following both ST and MT training. The extent of adaptation varied across training and test talker pairings; however, in both formats we found evidence of significant cross-accent as well as cross-talker generalization.

The relative entropy (Kullback-Leibler divergence) analysis showed that average relative entropy for MT training was higher than for ST training. This reflects greater divergence from baseline of test scores across the four test talkers following MT training than following ST training. Moreover, Figure 2 shows a cross-over effect for the least (TUR) and most (FAR) effective ST training talkers. Specifically, the upward sloping TUR-noTUR line indicates that exclusion of the least effective ST training talker (TUR) from the MT format (noTUR) benefitted adaptation, while the downward sloping FAR-noFAR line indicates that exclusion of the most effective ST training talker (FAR) detracted from adaptation.

The patterns emerging in this dataset suggest a possible role for baseline intelligibility in determining the extent of generalized perceptual adaptation to L2 speech. Listeners showed most consistent improvement across all training conditions for the test talker with the highest baseline intelligibility, BRP. Conversely, the talker with the lowest baseline intelligibility, TUR, was the least effective training talker. Variation in baseline intelligibility cannot account for all aspects of the present dataset, e.g., SPA and FAR have similar baseline intelligibility but they do not show qualitatively or quantitively equivalent performance as either training or test talkers. Nevertheless, association of the training and/or test talker's baseline intelligibility with extent of generalized adaptation to L2 speech noted above (i.e., for BRP and TUR) is consistent with other research on lexically-guided perceptual learning for speech which has demonstrated a crucial role for lexical knowledge in exposure-induced recalibration of phonetic category boundaries [20, 21 and many others]. If word recognition accuracy is too low (as for a low-intelligibility L2 talker) then lexical knowledge may not be available to guide the mapping of phonetic variation to linguistically meaningful categories, and consequently adaptation to L2 speech may be constrained for both talker-specific and talker/accent-general adaptation. In these cases, more exposure or other source of lexical access (e.g., through written medium) may be needed for effective adaptation. Conversely, if word recognition accuracy is high then lexically-guide recalibration of phonetic categories can presumably proceed readily.

In conclusion, together with prior work [5-13], the present study suggests that variability, similarity, and intelligibility of both training and test talkers are all relevant for determining the balance of specificity and generality of perceptual adaptation to L2 speech. Future research should directly test the separate and combined effect of each of these factors on generalized perceptual adaptation to L2 speech.

7. REFERENCES

[1] Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97.

[2] Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency on intelligibility of Chinese-accented English. *Language and Speech*, 47(2), 139–154.

[3] McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, & Psychophysics*, 80(6), 1559–1570.

[4] Harwell, D. (2018). The Accent Gap. *Washington Post*, 1–11.

[5] Bent, T. & Baese-Berk, M. (2021). Perceptual learning of accented speech. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni, eds. The Handbook of Speech Perception, 2nd ed. John Wiley & Sons, Inc., pp. 428–464.

[6] Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729.

[7] Tzeng, CY, Alexander, JD, Sidaras, SK, and Nygaard, LC (2016). The role of training structure in perceptual learning of accented speech. *J. Exp. Psychol. Hum. Percept. Perform*. 42, 1793–1805.10.1037/xhp0000260

[8] SK Sidaras, JED Alexander, LC Nygaard (2009) Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America* 125(5), 3306-3316.

[9] Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America,* 133(3), EL174-EL180.

[10] Alexander, J., and Nygaard, LC (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, 145(6), 3382.

[11] X Xie, EB Myers (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language* 97, 30-46

[12] X Xie, FS Earle, EB Myers (2018) Sleep facilitates generalization of accent adaptation to a new talker. *Language, Cognition and Neuroscience* 33 (2), 196-210, https://doi.org/10.1121/1.5110302

[13] Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, 150(11), e22–e56.

[14] Baker, R. E., Baese-Berk, M., Bonnasse-Gahot, L., Kim, M., Van Engen, K. J., & Bradlow, A. R. (2011). Word durations in non-native English. *Journal of phonetics*, 39(1), 1-17.

[15] Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, 138(3), EL223-EL228.

[16] Bradlow, A. R. (2022). Information encoding and transmission profiles of first-language (L1) and second-language (L2) speech. *Bilingualism: Language and Cognition*, 25(1), 148-162.

[17] Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar

[18] Borrie, S.A., Barrett, T.S., & Yoho, S.E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *Journal of Acoustical Society of America*, 145, 392-399.

[19] Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68(1), 115-146.

[20] Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47, 204–238.

[21] Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, 13, 262–268.

LINGUISTIC RELEASE FROM MASKING IN SIMULATED ELECTRIC HEARING

Huizi Lung [1], Shangdi Liao [2], Fei Chen [2]

[1] Unit of Human Communication, Development, and Information Sciences, Faculty of Education, The University of Hong Kong, Hong Kong, China
[2] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

fchen@sustech.edu.cn

ABSTRACT

Many works revealed linguistic release from masking (LRM) in speech perception from normal-hearing listeners. The present work investigated the LRM in understanding vocoded Mandarin speech simulating cochlear implants (CIs) listening. Mandarin sentences were corrupted by 2- and 6-talker babble maskers spoken in languages with varying degrees of linguistic similarity to Mandarin, namely Mandarin, Cantonese, and English. Target and masker sentences were mixed and processed by an 8-channel noise vocoder to simulate CI listening, and presented to normal-hearing listeners to recognize. Mandarin recognition scores were significantly higher in English babble masker than in Cantonese and Mandarin babble maskers, and significantly higher in Cantonese than Mandarin talker babble masker, demonstrating LRM effect in simulated CI listening. Listeners benefited from a target-masker linguistic mismatch even though spectral resolution was low, and when competing talkers also spoke Mandarin, more talkers might be less disruptive to Mandarin speech recognition in CI listening.

Keywords: Mandarin speech perception, cochlear implants, linguistic release of masking (LRM).

1. INTRODUCTION

Speech perception in noise is a challenging task for people with normal hearing, and even more so for people with hearing loss. Normal-hearing (NH) listeners experience a release from masking (MR) when the masker is fluctuating, compared to when the masker is a steady state noise (SSN) [e.g., 1-2]. The target speech can be "glimpsed" during temporal dips in a fluctuating masker when the overall intensity of the masker is low or during spectral dips when target and masker differ in their frequency composition [2].

Linguistic dissimilarity between the target and masker speech has been shown to improve speech understanding. Such improvement is known as linguistic release from masking (LRM). The target-masker linguistic similarity hypothesis states that the more dissimilar the target and the masker language, the easier it is for the listener to segregate the target from the masker, especially at challenging signal-to-noise ratios (SNRs) [3]. LRM has been found in English [e.g., 4] and Mandarin speech perception [e.g., 5] among NH listeners. Both adults and children can benefit from a linguistic mismatch between target and masker [e.g., 6].

LRM studies also showed that masker meaningfulness plays a role in speech masking. Van Engen and Bradlow found that at SNR of 0 dB or lower, English sentence recognition by monolingual English speakers was worse in English 2-talker babble masker than in Mandarin 2-talker babble masker [7]. Bilingual listeners who can understand both the target and masker languages also showed LRM. Calandruccio and Zhou found that English-Greek bilinguals whose dominant language was English showed improved English speech recognition in Greek maskers compared to in English maskers [8]. In [5], Mandarin-English bilinguals showed better Mandarin sentence recognition when English rather than Mandarin maskers were used. These results suggested that a more familiar masker language interfered more with target speech perception.

For listeners with profound-to-severe hearing impairment, cochlear implants (CIs) provide an efficient way for them to restore their hearing [9]. CIs extract useful acoustic information (i.e., primarily multi-channel temporal envelope waveforms, or discarding spectral details) from the original speech input, and electrically stimulate the auditory nerves to deliver the acoustic information and subsequently evoke sound perception in brain.

Vocoder processing has been long used to simulate CI listening in NH subjects [10]. Real CI users may vary significantly due to factors such as etiology of hearing loss and implant device differences. Vocoder processing is useful in measuring the effect of a processing parameter

change in CI as the amount of spectral and temporal cues, such as the number of frequency channels, can be manipulated independent of patient-related factors.

In CI users, speech perception is worse in modulated noise than SSN or not significantly different in the two kinds of noise, showing negative MR or no MR in a fluctuating masker [e.g., 11-12]. Liu et al. found that children who used CIs also showed less MR than NH adults and children [13]. Fu and Nagoki proposed that the reduced or absent MR benefit may be due to reduced spectral resolution and spectral smearing in CI listening [11], as they found that NH subjects listening to vocoded speech performed similarly to CI users when spectral resolution was low (i.e., 4 channels). CI signal processing is limited by the number of electrodes implanted and the number of spectral channels. Ihlefeld et al. also found that NH listeners demonstrated decreased MR when presented with vocoded stimuli [14].

While there is ample evidence that hearing-impaired listeners show reduced MR benefit than NH listeners, few studies investigated LRM in hearing-impaired listeners. Viswanathan et al. reported LRM benefit in simulated CI listening [15]. NH listeners listened to vocoded English sentences in competing English or Dutch masker sentences. English speech recognition performance was better when the masker was Dutch.

The purpose of this study was to investigate whether Mandarin-speaking CI users may experience LRM by using a vocoder processing to simulate CI listening. The target language in this study was Mandarin (which is a tonal language), and three masker languages, Mandarin, Cantonese, and English, which vary in linguistic similarity to Mandarin, were used.

2. METHODS

2.1. Subjects and materials

Sixteen native Mandarin Chinese listeners (10 males and 6 females, age range 21 to 27 years) were recruited from the Southern University of Science and Technology, and paid to participate in the study. Participants were students at the university and were bilingual (Mandarin and English), and had hearing within normal range. The experimental procedure involving human subjects was approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

The speech material comprised sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database [16]. The MHINT corpus includes 24 lists, each with 10 sentences and 10 keywords per sentence. All sentences were spoken by a male native Mandarin Chinese speaker with a fundamental frequency (F0) of 75–180 Hz and recorded at a sampling rate of 16 kHz. The target MHINT sentences were corrupted by competing talker babble masker (2-talker or 6-talker babble, in Mandarin, Cantonese, or English) at SNR levels of 0 or –5 dB.

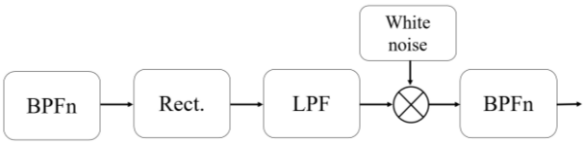


Figure 1: Schematic of Implementing the Noise Vocoder at the nth Channel. Rect.: Waveform rectification.

Six native-Mandarin speakers (3 male and 3 female) and 6 native-Cantonese speakers (3 male and 3 female) recorded sentences in Mandarin and Cantonese respectively for the Mandarin and Cantonese maskers. The speakers were each instructed to read 8 sentences extracted from the newspaper in a natural style. Recording was done in a sound-proof booth with a digitization rate of 16 kHz. All sentences were equated for root-mean-square (RMS) level. All 8 sentences spoken by each speaker were combined to form a one-talker masker signal, resulting in 6 different masker signals for each language (Mandarin and Cantonese). English masker speech was taken from the TIMIT database [17]. Eight sentences were taken from each of 6 (3 male and 3 female) speakers. Sentences from each speaker were combined to form 6 one-talker masker signals. Each of the 18 (=6 speakers × 3 languages) one-talker masker signal ranged from 15 to 25 seconds in duration.

2.2. Signal processing

Two male voices from each language were used to produce the 2-talker babble signals to match the speaker gender of the target materials (MHINT sentences spoken by a male speaker). A random segment of masker speech with duration equal to that of the target MHINT sentence was selected from each of the 2 male voice masker signals. Then, the two segments were equalized for RMS level and summed up. The 2-talker signal was then mixed with the target MHINT sentence at SNR levels of 0 and –5 dB. The corrupted MHINT sentence was then adjusted to have the same RMS value as the original sentence. For the 6-talker babble maskers, segments randomly selected from each of 6 masker sentences (in the same language) were used to corrupt the

MHINT sentence at SNR levels of 0 and –5 dB. The processing was the same as that used in the 2-talker babble (i.e., RMS equalization, summation, and mixture with the MHINT sentence).

Vocoder processing was used to simulate CI listening [e.g., 10]. MHINT sentences, after being corrupted with the maskers, were processed by a noise vocoder. Figure 1 shows the schematic of implementing the noise vocoder in one channel. Speech signals were first processed through a pre-emphasis high-pass filter with 1200 Hz cut off frequency. Then, signals were band-pass filtered (BPF) into 8 frequency channels between 80 and 6000 Hz. The corner frequencies for the 8 channels were 80, 221, 425, 724, 1158, 1790, 2710, 4050, and 6000 Hz. The temporal envelope was extracted by half-wave rectification followed by low-pass filtering (LPF) with 250 Hz cut off frequency. The extracted envelope from each frequency band was then used to modulate a white noise signal, and BPF again into 8 frequency bands. The BPF signals at each frequency channel were then summed up to generate the noise-vocoded speech stimulus with its amplitude readjusted to match the RMS value of the original signal.

2.3. Procedure

Participants took part in a training session before the speech recognition experiment. The training session and the experiment were conducted in a soundproof room. Speech materials were played to listeners through circumaural headphones at a comfortable listening level indicated by the listeners. Participants were instructed that they would listen to Mandarin sentences that had been processed. Each sentence could be played three times at maximum.

In the training session, listeners listened to vocoded MHINT sentences while reading the transcript for the sentences to become familiarized with the processed speech materials. In the experiment, participants listened to vocoded MHINT sentences without reading. They were instructed to repeat the words they heard, and to guess if they were not sure. The experimenter scored participants' response by the number of words correctly identified in the sentence. The intelligibility score for each condition was computed as the percentage of words in the target MHINT sentences correctly identified in each list of MHINT sentences.

In the experiment, all subjects participated in a total of 12 conditions [=3 masker languages × 2 types of talker babbles × 2 SNR levels]. The babble maskers in different languages with varying linguistic contents are denoted as "Eng 2", "Eng 6", "Can 2", "Can 6", "Man 2" and "Man 6". "Eng",

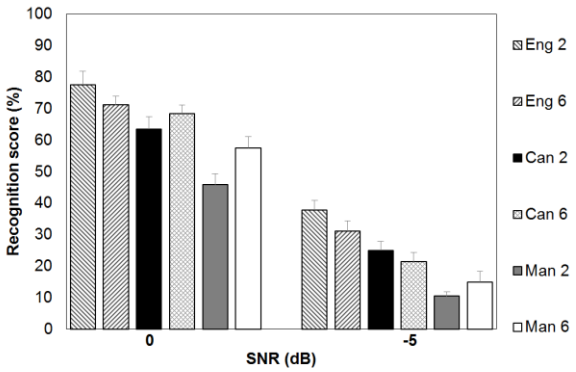


Figure 2: Mean Sentence Recognition Scores for All Conditions. Error bars denote 1 standard error of the mean.

"Can" and "Man" refer to English, Cantonese, and Mandarin respectively, and "2" and "6" indicated the number of talkers in each masker babble. A different list of MHINT sentences was used in each condition and none of the target sentences was repeated across conditions. The order of the test conditions was randomized across subjects to minimize potential learning effects. Subjects were given a 5-minute break every 30 minutes during testing. The whole experiment took about 2 to 3 hours to complete for each subject.

3. RESULTS

Figure 2 shows the mean Mandarin sentence recognition scores for all conditions. Statistical significance was determined using percent correct score as the dependent variable, and masker language (English, Cantonese, or Mandarin), type of talker babble (2-talker or 6-talker), and SNR level (0 dB or –5 dB) as within-subject factors.

Three-way repeated measures analysis of variance (ANOVA) indicated a significant effect of masker language ($F_{2,30}$ = 70.890, p < 0.001, η^2_P = 0.825), a non-significant effect of type of talker babble ($F_{1,15}$ = 0.498, p = 0.491, η^2_P = 0.032), and a significant effect of SNR level ($F_{1,15}$ = 984.877, p < 0.001, η^2_P = 0.985). Analysis also revealed a significant interaction between masker language and type of talker babble ($F_{2,30}$ = 7.745, p = 0.002, η^2_P = 0.341). The interaction between masker language and SNR level ($F_{2,30}$ = 0.742, p = 0.485, η^2_P = 0.047) and between type of talker babble and SNR level ($F_{1,15}$ = 3.164, p = 0.096, η^2_P = 0.174) were not significant. The interaction among masker language, type of talker babble, and SNR level ($F_{2,30}$ = 0.660, p = 0.524, η^2_P = 0.042) was not significant.

Post-hoc analysis with Bonferroni correction showed that with respect to the effect of masker language, Mandarin sentence recognition score was

significantly better (higher) in English maskers than in Cantonese maskers ($p < 0.001$) or Mandarin maskers ($p < 0.001$), and significantly better in Cantonese maskers than in Mandarin maskers ($p < 0.001$). Comparisons across conditions showed that this was true when the type of talker babble and the SNR level were the same, except in the condition pairs "Eng 6" vs. "Can 6" at 0 dB SNR, and "Can 6" vs. "Man 6" at –5 dB SNR. With respect to the effect of SNR level, recognition score was significantly better at 0 dB SNR than at –5 dB SNR ($p < 0.001$). Comparisons across conditions indicated that this was true in all condition pairs when the masker language and type of talker babble were the same.

4. DISCUSSION AND CONCLUSION

The present study investigated whether CI users experience LRM by using vocoder processing to simulate CI listening in NH listeners. The results indicate that simulated CI listening benefited from a linguistic mismatch between target and masker languages and showed LRM in Mandarin speech recognition.

Mandarin sentence recognition scores were significantly better in conditions with English talker babble masker compared to Cantonese or Mandarin babble masker, and significantly better in Cantonese compared to Mandarin babble masker, demonstrating LRM effect in simulated CI listening. The most effective masker was Mandarin, and English was the least effective masker. These findings are similar to results seen in NH listeners, which indicated that when the masker language was different from the target language, listeners showed MR benefit, compared to when the masker and target speech were spoken in the same language [e.g., 4-5, 18]. Furthermore, findings were consistent with results from Viswanathan et al. [15] which also demonstrated LRM benefit in simulated CI listening.

Despite a lack of access to F0 cue and reduced spectral resolution in the 8-channel noise-vocoded stimuli, listeners were able to benefit from a linguistic mismatch between target and masker speech. This suggests that linguistic properties which are not represented through F0 or spectral cues may contribute to the LRM benefit in a target-masker mismatch condition.

Mandarin speech recognition may be more vulnerable to interference from competing talkers also speaking Mandarin in CI listening because they may not be able to resolve concurrent tonal information in competing streams of speech. Luo and colleagues investigated concurrent vowel and tone recognition in NH listeners, simulated CI listeners and real CI listeners in a series of experiments [e.g., 19-20]. NH listeners achieved nearly perfect vowel and tone identification across conditions. But when the same group of subjects listened to 8-channel or 4-channel noise-vocoded speech simulating CI listening, tone recognition performance in single syllable or concurrent syllables became significantly poorer [19]. Real CI users also performed the recognition tasks, and their tone recognition performance in the single-syllable condition was significantly poorer than 4- and 8-channel simulation results [20], indicating that real CI users were even less able to make use of F0 and pitch information for tone recognition. CI users' performance of concurrent tone recognition was also poorer than performance of NH subjects and simulation performance.

In [5] and [21], NH listeners showed better Mandarin speech perception with fewer number of talkers in competing speech. Their findings were also consistent with findings from studies of English speech-on-speech masking in NH English speakers [e.g., 7]. A masker with fewer talkers may provide more opportunities for glimpsing the target speech in NH listeners. However, Chen et al. found that adult Mandarin-speaking CI users performed worse in a single-talker masker than in a 2-talker or 4-talker masker [21]. Results from the present study with regards to the effect of number of competing talkers were somewhat consistent with results from [21], but in contrast to studies of English speech masking [22]. In the present study, when the masker language was Mandarin, subjects performed worse with fewer number of competing talkers in the Mandarin babble masker. This effect was the same at both SNR levels used.

In conclusion, consistent with early findings from NH listeners, the present study simulating CI speech perception showed the effect of linguistic release from masking. The recognition of vocoder-processed Mandarin sentences in speech masker was notably influenced by the linguistic content of masker, with an increased influence from English, Cantonese to Mandarin. As CI speech perception primarily relies on limited amount of temporal envelope information, the findings in this work provide evidence on the perceptual impact of temporal envelope in linguistic release of masking.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61971212). Part of this study was the basis for the Master's dissertation of the first author (H.Z.L.).

6. REFERENCES

[1] Festen, J. M., Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America,* 88, 1725-1736.

[2] Peters, R. W., Moore, B. C. J., Baer, T. 1998. Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America,* 103, 577-587.

[3] Brouwer, S., Van Engen, K. J., Calandruccio, L., Bradlow, A. R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America,* 131, 1449-1464.

[4] Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., Bradlow, A. R. 2013. Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology,* 22, 157-164.

[5] Chen, F., Li, J., Wong, L. L. N., Yan, Y. 2013. Effect of linguistic masker on the intelligibility of Mandarin sentences. *Proc. of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2098-2102.

[6] Calandruccio, L., Leibold, L. J., Buss, E. 2016. Linguistic masking release in school-age children and adults. *American Journal of Audiology,* 25, 34-40.

[7] Van Engen, K. J., Bradlow, A. R. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America,* 121, 519-526.

[8] Calandruccio, L., Zhou, H. 2014. Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance. *Journal of Speech, Language, and Hearing Research,* 57, 1089-1097.

[9] Chen, F., Ni, W., Li, W., Li, H. 2019. Cochlear implantation and rehabilitation. In: Li, H., Chai, R. (eds), *Hearing Loss: Mechanisms, Prevention and Cure.* Springer, Singapore, 129-144.

[10] Chen, F., Zheng, D. C., Tsao, Y. 2017. Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language. *The Journal of the Acoustical Society of America*, 142, 1157–1166.

[11] Fu, Q. J., Nogaki, G. 2005. Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology,* 6, 19-27.

[12] Zirn, S., Polterauer, D., Keller, S., Hemmert, W. 2016. The effect of fluctuating maskers on speech understanding of high-performing cochlear implant users. *International Journal of Audiology,* 55, 295-304.

[13] Liu, J. S., Liu, Y. W., Yu, Y. F., Galvin, J. J., Fu, Q. J., Tao, D. D. 2021. Segregation of competing speech in adults and children with normal hearing and in children with cochlear implants. *The Journal of the Acoustical Society of America,* 150, 339-352.

[14] Ihlefeld, A., Deeks, J. M., Axon, P. R., Carlyon, R. P. 2010. Simulations of cochlear-implant speech perception in modulated and unmodulated noise. *The Journal of the Acoustical Society of America,* 128, 870-880.

[15] Viswanathan, N., Kokkinakis, K., Williams, B. T. 2018. Listeners experience linguistic masking release in noise-vocoded speech-in-speech recognition. *Journal of Speech, Language, and Hearing Research,* 61, 428-435.

[16] Wong, L. L. N., Soli, S. D., Liu, S., Han, N., Huang, M. W. 2007. Development of the Mandarin Hearing in Noise Test (MHINT). *Ear and Hearing,* 28, 70S–74S.

[17] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., Zue, V. 1993. TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium, 1993.*

[18] Calandruccio, L., Dhar, S., Bradlow, A. R. 2010. Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America,* 128, 860-869.

[19] Luo, X., Fu, Q. J. 2009. Concurrent-vowel and tone recognition in acoustic and simulated electric hearing. *The Journal of the Acoustical Society of America,* 125, 3223–3233.

[20] Luo, X., Fu, Q. J., Wu, H. P., Hsu, C. J. 2009. Concurrent vowel and tone recognition by Mandarin-speaking cochlear implant users. *Hearing Research, 256*, 75–84.

[21] Chen, B., Shi, Y., Zhang, L., Sun, Z., Li, Y., Gopen, Q., Fu, Q. J. 2020. Masking effects in the perception of multiple simultaneous talkers in normal-hearing and cochlear implant listeners. *Trends in Hearing,* 24, 1-12.

[22] Cullington, H. E., Zeng, F. G. 2008. Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *The Journal of the Acoustical Society of America,* 123, 450-461.

ARE ALVEOLAR TRILLS PERCEIVED AS "STRONG" RHOTICS?

Eva Reinisch[1], Holger Mitterer[2,3]

[1]Austrian Academy of Sciences, [2]University of Malta, [3]Hanyang University Seoul, South Korea
Eva.Reinisch@oeaw.ac.at, holger.mitterer@um.edu.mt

ABSTRACT

Rhotics are a highly variable sound class. In German, realizations as different as alveolar trills [r] and uvular fricatives [ʁ] co-occur as allophones of a single rhotic phoneme. Previous evidence indicated that listeners perceive [r] as the strongest/most salient version of /r/. The present experiments tested whether this leads to a form of speaker adaptation, in which weak forms of /r/ are deemed less acceptable, if the speaker also produces the strong from (i.e., [r]). In a 2AFC task, listeners categorized a continuum from [ʁ] to [h] while also hearing the same speaker use [r] or not. In two experiments, we found that participants in the [r]-exposure group accepted fewer stimuli from the [ʁ]-[h] continuum as /r/ than those in the [ʁ]-only exposure group. A third experiment tested whether this effect is speaker specific by exposing participants to [r] from a different speaker, but those results were inconclusive.

Keywords: speech perception, rhotics, German, perceptual saliency, speaker specificity

1. INTRODUCTION

Rhotics form a highly variable sound class [1]. They can be produced at different places and manners of articulation with variants as different as alveolar trills and uvular fricatives. Cross-linguistically, the most frequent variant is the alveolar trill, despite its articulatory complexity [1]. Trills are also considered the perceptually "strongest" or "most salient" variants, which is due to their "clearly modulated signal, distinct from other speech segments" ([2], see also [3],[4]). In the present study, we asked whether different rhotic allophones are interpreted relative to each other such that the presence of a strong [r] makes listeners less willing to accept other, "weaker" rhotic allophones as an instance of /r/. Note that such a finding may partly explain the popularity of [r] in the world's languages despite its articulatory complexity.

Specifically, we assessed the perception of alveolar trills compared to the articulatorily and acoustically maximally different uvular fricatives. German was used as the example language, since in German both, alveolar trills and uvular fricatives (among others) co-occur as allophones of a single rhotic phoneme [5]. While in syllable onset [ʁ] or [ʀ] are considered the standard, in the south of Germany, many speakers use [r]. In general, speakers are relatively consistent in their use of rhotic allophones in prevocalic position. With regard to the status of [r] as the perceptually strongest rhotic allophone, one study on German already found some support using a selective adaptation paradigm [6].

Selective adaptation is the effect that if a given stimulus is heard repeatedly, a following ambiguous stimulus is less likely to be perceived as an instance of the previously heard adaptor. In the case of rhotics, if listeners are exposed to /r/-initial words produced with [ʁ], then ambiguous stimuli from a test continuum between [ʁ] and [h] should be less likely perceived as the rhotic. In [6], this full match condition (among others) was compared to a condition in which the adapter stimuli were produced with [r], an allophone with little acoustic overlap to [ʁ] in the test continuum. A stronger effect of selective adaptation was *expected* for the full match condition but *found* for the acoustically dissimilar [r]-adaptors. This greater effectiveness of the alveolar trill as adaptor was interpreted to result from this allophone's greater perceptual saliency compared to other German rhotic allophones.

The present study set out to further assess the role of trilled rhotic allophones relative to other variants in speech perception. Using a different experimental setup, we asked whether native German listeners who have heard /r/-initial words produced with [r], would be less willing to accept [ʁ] as an instance of a rhotic than when no alveolar trills have been heard.

In three experiments, native German listeners performed blocks of forced-choice minimal word pair identification tasks categorizing [ʁ]-[h] and [r]-[l] continua. Different continua for the different rhotic allophones were necessary to minimize the acoustic distance between the rhotics and the other option. In Experiment 1 the rhotic allophones were produced on different words but spoken by the same speaker. Experiment 2 then went on to test categorization of the different allophones on the same word in two different minimal pairs, again spoken by the same speaker. Experiment 3 used different word pairs and the [r]-[l] continuum was spoken by a different speaker. Responses to the [ʁ]-[h] continuum were compared between a group of listeners who categorized both continua and one who only ever heard the [ʁ]-[h] continuum. We hypothesized that if

[r] is perceived as a strong/salient variant of the rhotic, then hearing a speaker produce this allophone should reduce listeners' willingness to categorize sounds along the [ʁ]-[h] continuum as the rhotic.

2. EXPERIMENTS 1 AND 2

2.1. Methods

2.1.1. Participants

Thirty and 34 participants for Experiments 1 and 2 respectively were recruited via the online platform prolific.co [7]. Participants were between 20 and 42 years of age and declared themselves as native speakers of German. They performed the experiments via the internet and were paid for their services.

2.1.2. Materials

A female native speaker of German recorded the minimal word pairs *Reis-heiß* (rice-hot), *Reiher-Leiher* (heron-lyre), *Rose-Hose* (rose-pants) and *Rose-Lose* (rose-lottery ticket). The rhotics were produced as [ʁ] in the minimal pairs contrasting with /h/ and as [r] when contrasting with /l/, matching the places of articulation. For [r], a token with a clear amplitude modulation was chosen; for [ʁ] one with a steady-state fricative noise. The words with /h/ and /l/ were selected to match the /r/-initial words in duration and f0 to facilitate the creation of continua, for which we used audio morphing with STRAIGHT [8]. A set of pre-tests with participants who did not take part in the main experiments was run to find 6 continuum steps (from originally 21) that spanned the range from good tokens of /r/ to the other phoneme. The continua *Reis-heiß* and *Reiher-Leier* were used in Experiment 1, *Rose-Hose* and *Rose-Lose* in Experiment 2.

2.1.3. Procedure

Except for the use of different minimal word pairs, the procedure was the same in Experiments 1 and 2. Participants performed a 2-alternative forced choice task (2AFC). On each trial they were presented with an audio stimulus (i.e., one token from the continua) and saw two pictures on the computer screen that represented the words from the respective minimal pair. They had to respond by button press (left vs. right arrow key) whether the left or right picture better matched the audio stimulus.

In each experiment, participants were randomly assigned (by the JavaScript program based on [9]) to the "same" condition or the "mixed" condition. Participants in the same condition only ever heard stimuli from the [ʁ]-[h] continuum. Participants in the mixed condition received alternating blocks from the continuum in which the rhotic was produced as [ʁ] or [r]. Each participant completed 30 blocks, where one block consisted of all 6 stimuli from a given continuum. Table 1 illustrates the design.

Block	Same condition	Mixed condition
1	uvular fricative	alveolar trill
2	uvular fricative	uvular fricative
3	uvular fricative	alveolar trill
...		
30	uvular fricative	uvular fricative

Table 1: Design of Experiments 1 and 2.

2.2. Analyses and Results

The two experiments were analysed separately but in identical manner. Analyses were conducted in R [10] using generalized linear mixed-effects models with a logistic linking function using the lme4 [11] and lmerTest [12] packages. In all models, the dependent variable was whether participants heard the word as /r/-initial or not. Fixed factors were Continuum Step, Block (both variables rescaled to range from -0.5 to 0.5), and Condition ("mixed" coded as 0.5, "same" as -0.5). Condition was allowed to interact with the other two factors. The models used the maximal random-effects structure that converged.

Two types of analyses were conducted. First, we compared responses to only the even numbered blocks, that is blocks, in which participants in both conditions responded to the [ʁ]-[h] continuum. Secondly, we compared the first 15 blocks of the same condition to the even numbered blocks in the mixed condition, again only comparing responses to the [ʁ]-[h] continuum. Here conditions matched on experience with the [ʁ]-[h] continuum.

Tables 2 and 3 report the statistics for the analyses of even numbered blocks in Experiments 1 and 2 respectively, and Fig. 1 and 2 illustrate the results. Analyses for both experiments show a significant effect of Condition, suggesting that fewer /r/-responses were given by participants in the mixed conditions. That is, if listeners are presented with rhotics produced as [r], then [ʁ] was less likely to be accepted as a rhotic than when only [ʁ] was heard. This was the case when different /r/-initial words were used in Experiment 1 as well as when the same word was heard in both conditions in Experiment 2.

The second analyses that took into account the amount of experience with [ʁ] showed similar results. The effect of Condition in Experiment 1 just reached significance ($b_{(\text{Condition})}$=-0.82, z=-1.98, p=0.048) as did the effect of Condition in Experiment 2 ($b_{(\text{Condition})}$=-1.15, z=-2.25, p=0.024). All data and analyses are made available on OSF [13].

Factor	b	z	p
Intercept	-0.20	-1.03	0.303
Step	3.66	13.72	< 0.001
Condition	-0.99	-2.52	0.012
Block	0.15	0.92	0.359
Step:Condition	0.61	1.25	0.260
Condition:Block	-0.42	-1.32	0.186

Table 2: Statistics for Experiment 1; comparison of even numbered blocks between conditions.

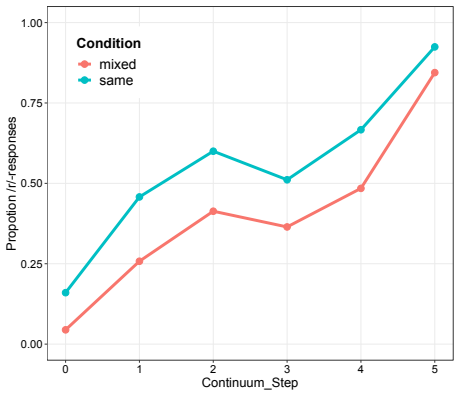


Figure 1: Results of Experiment 1: /r/-responses along the [ʁ]-[h] continuum depending on whether the speaker was also heard producing [r] on a different /r/-initial word.

Factor	b	z	p
Intercept	-0.93	-3.96	< 0.001
Step	2.95	7.44	< 0.001
Condition	-1.06	-2.28	0.023
Block	0.16	1.03	0.306
Step:Condition	0.42	0.54	0.587
Condition:Block	1.07	3.36	< 0.001

Table 3: Statistics for Experiment 2; comparison of even numbered blocks between conditions.

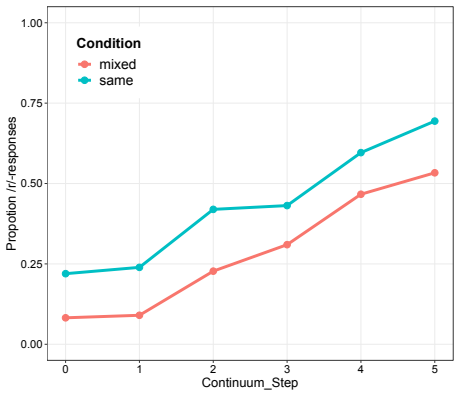


Figure 2: Results of Experiment 2: /r/-responses along the [ʁ]-[h] continuum depending on whether the speaker was also heard producing [r] on the same /r/-initial word.

2.3. Discussion

Experiments 1 and 2 confirmed our hypothesis that German listeners would be less willing to accept [ʁ] as an instance of the rhotic if they heard the speaker produce [r] as well. This effect is clearly visible in the descriptive data with a shift of roughly 15% fewer /r/-responses if an alveolar trill was heard. However, the effect is not strong statistically, just passing the p.<05 criterion. This is due to large interindividual differences in how often listeners perceive the rhotic on the [ʁ]-[h] continuum. However, our manipulation has to be between participants. Nevertheless, the effects were found in both types of analyses and replicated when the two rhotic allophones were heard on different words as well as on the same words.

3. EXPERIMENT 3

Experiment 3 addressed the question to what extent the decrease in /r/-responses to [ʁ] when also [r] was heard might be driven by listeners' expectation that a given speaker produces only one rhotic allophone. It therefore set out to replicate the mixed condition of Experiment 1, but with a male speaker producing the [r]-[l] continuum. The [ʁ]-[h] continuum was the same as in Experiment 1. We hypothesized that the effect of fewer /r/-responses in this new mixed rhotics/ different speaker condition might disappear relative to the mixed rhotics/ same speaker condition.

3.1. Methods

3.1.1. Participants

A new set of 40 participants was recruited via the online platform prolific.co [9]. Additionally, the data from 15 participants from Experiment 1 was re-used - those who had been assigned to the mixed condition.

2.1.2. Materials

The same [ʁ]-[h] continuum was used as in Experiment 1. In addition, a male native speaker of German recorded the minimal pair *Reiher-Leier* with the rhotic produced as [r]. According to the same criteria as reported above, tokens were selected and morphed into a 21-step continuum. Again, six steps were selected in a pre-test.

2.1.3. Design and Procedure

The "same speaker" condition here was identical to the mixed condition in Experiment 1. The female speaker was heard producing both, the [r]-[l] and [ʁ]-[h] continua, again in 30 alternating blocks, where one block refers to all 6 steps of a given continuum.

The "mixed speaker" condition was identical in design and procedure, with the only difference that the [r]-[l] continuum was produced by the male speaker. The randomization procedure assigned participants to the same- or mixed-speaker conditions with a probability of 1/3 vs. 2/3, which led to 12 versus 28 of the new participants assigned to the same versus mixed condition, respectively (15 were added to the same-speaker condition from Experiment 1).

3.2. Analyses and Results

We analyzed responses to the female speaker's [ʁ]-[h] continuum, which was identical across conditions. Condition now refers to whether the [r]-[l] continuum was spoken by the female or male speaker. The statistical model was identical to that in the first two experiments, again using the maximally converging random-effects structure (see OSF [13]). Table 4 reports the statistics and Fig. 3 illustrates the results.

Factor	b	z	p
Intercept	-0.43	-2.19	0.082
Step	3.99	15.96	< 0.001
Condition	-0.51	-1.32	0.187
Block	-0.31	-1.25	0.211
Step:Condition	-0.17	-0.35	0.725
Condition:Block	0.64	1.31	0.192

Table 4: Statistics for Experiment 3. Comparison of responses to the [ʁ]-[h] continuum.

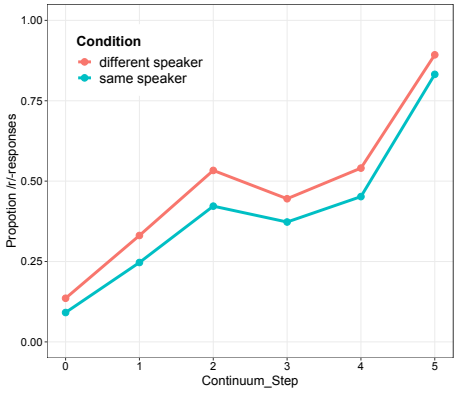


Figure 3: Results of Experiment 3: /r/-responses along the [ʁ]-[h] continuum spoken by the female speaker depending on whether the [r]-[l] continuum was produced by the same or a different speaker.

3.3. Discussion

Experiment 3 showed a numerical difference in how often [ʁ] was accepted as an instance of the rhotic between the conditions in which [r] was heard from the same versus a different speaker. This difference was, however, not statistically significant, due to the large interindividual differences in how often /r/ was perceived on the [ʁ]-[h] continuum. Therefore, we interpret the results of Experiment 3 as inconclusive. We cannot say whether the tendency to reject [ʁ] as an instance of the rhotic depends on hearing them from the same speaker.

4. GENERAL DISCUSSION

The present study built on the common assumption that trilled rhotics are perceptually "strong" or "salient" variants of rhotic phonemes [2,4,5]. It asked whether this strength of trills would affect the perception of other rhotic allophones, specifically the uvular fricative [ʁ]. We used German as the test language since a variety of rhotic allophones is found in native speakers, including [r] and [ʁ]. Moreover, as discussed in the introduction, a recent study [6] has shown that [r] triggers stronger selective adaptation than other rhotic allophones. The present study adds to this finding by showing that listeners are less willing to accept [ʁ] as a rhotic if also [r] is heard. This may be one of the reasons that [r] is able to remain the standard rhotic in so many languages despite its articulatory complexity.

That is, the data support hypothesis that the alveolar trill can be interpreted as a perceptually strong/ salient from of the rhotic. However, this finding might be somewhat surprising in the light of the language that we used to show the effect. While the saliency of the alveolar trill has been used as an argument to explain the prevalence of this allophone across languages despite its articulatory complexity [1], German does not entirely fit the pattern. Specifically, German has undergone a recent historical shift from [r] to [ʁ] as the standard rhotic allophone (for discussions see [4] and [5]).

We speculate that the fact that [ʁ] is considered the German standard, in addition to the occurrence of a large variety of rhotic allophones in German local varieties and dialects, might have contributed to the large interindividual differences in the present study. Note that a previous study on spoken-word recognition in German [14] showed that native speakers of German were faster to recognize /r/-initial words if they were produced with [ʁ] rather than [r]. Further research will therefore have to assess to what extent acoustic properties interact with language or speaker-specific characteristics of allophones during speech perception.

5. ACKNOWLEDGMENTS

This work was supported by a grant to (a.o.) HM by the Ministry of Education of the Republic of Korea

and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884).

6. REFERENCES

[1] Ladefoged, P., Maddieson, I. (1996). *The Sounds of the World's Languages.* Oxford: Blackwell.
[2] Solé, M.-J. (2002). Aerodynamic characteristics of trills and phonological patterning. *J. Phon.* 30, 655–688.
[3] Sebregts, K. (2015). *The sociophonetics and phonology of Dutch r.* Poland: LOT Publications.
[4] Schiller, N. O. 1999. The phonetic variation of German /r/. In: Butt, M., Fuhrhop, N. (eds), *Variation und Stabilität in der Wortstruktur: Untersuchungen zu Entwicklung, Erwerb und Varietäten des Deutschen und anderer Sprachen*. Hildesheim, 261-287.
[5] Wiese, R. 2003. The unity and variation of German /r/. *Zeitschrift für Dialektologie Linguistik.* 70, 25–43.
[6] Mitterer, H., Reinisch, E. 2023. Selective adaptation of German /r/: A role for perceptual saliency. *Attention, Percept., Psychophys.*, 85, 222-233.
[7] Prolific, https://www.prolific.co/.
[8] Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. *Speech Comm.* 27, 187–207
[9] de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47, 1–12.
[10] R Core Team. 2021. R: A language and environment for statistical computing. (version 4.1.2) R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.R
[11] Bates, D. M., & Sarkar, D. 2007. lme4: Linear mixed-effects models using S4 classes (Version 0.999375–27) [Computer software]. http://www.r-project.org/: R Foundation for Statistical Computing.
[12] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *J. Statist. Softw.* 82:13.
[13] OSF, supplementary materials. https://osf.io/2jpba/.
[14] Llompart, M., Eger, N. A., Reinisch, E. 2021. Free allophonic variation in native and second language spoken word recognition: The case of the German Rhotic. *Frontiers in Psychology* 12, 5289.

Vowel perception under prominence: Dual roles for F0 and duration

Jeremy Steffman[1] Wei Zhang[2]

[1]The University of Edinburgh [2]McGill University
[1]jeremy.steffman@ed.ac.uk [2]wei.zhang16@mail.mcgill.ca

ABSTRACT

Prosodic prominence modulates vowel production and acoustics. In the present study we test if the same effects play out in perception. We test perception of four American English vowel contrasts, varying formants and duration along a continuum. We test how F0-based prominence modulates vowel categorization, and explore the role of vowel duration as an intrinsic and prominence-lending cue. We find that F0-based prominence modulates vowel perception, dependent on vowel height. High vowels undergo perceptual recalibration in line with hyperarticulation, non-high vowels instead reflect sonority expansion. Lengthening of the vowel is interpreted as an intrinsic vowel quality cue, not as prominence. However, both F0-based prominence and duration interact with formants: more prominent F0 and longer duration enhance categorization along the formant continuum, showing a stronger influence of formant cues under prominence. Results thus show a dual influence for F0 and duration. Both independently impact categorization, and additionally mediate formant cue use.

Keywords: speech perception, vowels, prominence, sonority expansion, hyperarticulation.

1. INTRODUCTION

From a phonological view, a phenomenon in spoken language can be either segmental (e.g., vowels and consonants) or suprasegmental (e.g., presence or absence of a pitch accent). However, as [1] pointed out, it's hard to find a domain of speech dealing with only "suprasegmentals" (F0, duration and intensity). These intricacies have only more recently begun to be explored in terms of speech perception and speech processing [2,3]. Segments differ intrinsically in features such as F0 and intensity [4-6]. In addition, prosodic prominence modifies "segmental features" including VOT and formant structure in vowels [e.g., 7,8]. In other words, many phonetic dimensions play dual roles (both suprasegmental and segmental). For vowels in particular, F0, duration and formants have each been shown to vary systematically both as function of contrastive segmental features, and as a function prosodic prominence. In this study we ask: How does the listener deal with this multiple-mapping in speech perception?

Vowel perception under prominence is an interesting test case because there are two existing frameworks predicting acoustic variation under prosodic prominence in vowel production: *hyper-articulation* and *sonority expansion*. The hyper-articulation model [9] predicts that the vowel's distinctive features are hyper-articulated when the vowel is produced as prominent (e.g., pitch-accented). The sonority expansion model [10] predicts that for a prominent vowel, the vocal tract is more open (i.e., the jaw is lower) so that more energy can radiate from it, resulting in expanded sonority. The speech production literature suggests that the hyper-articulation hypothesis better predicts acoustic variation in high vowels and the sonority expansion hypothesis better predicts the patterns in non-high vowels, in American English vowels at least [7, 11, cf. 12]. This begs the question if perceptual analogs of these effects occur, and if they too vary based on vowel height, as in the production literature.

Previous work has shown that listeners consider prosodic features in segmental perception [e.g., 13, 14]. For example [13] showed that listeners expected a longer VOT to identify a sound as voiceless English /p/ after an intonational phrase boundary than a word boundary, reflecting phrase-initial lengthening (or, strengthening) of VOT in speech production. [14] recently explored how prosodic prominence influenced vowel categorization, finding that listeners used phrasal prosodic prominence to recalibrate perception of the /ɛ/-/æ/ contrast. A prominent vowel was perceived as having undergone sonority expansion with acoustically lower and backer (in the vowel space) formants perceived as /ɛ/ when prominent. Extending [14,15], we test how pitch accent type/shape influences vowel perception, and how vowel duration (along a continuum) impacts vowel perception for four vowel contrasts varying in height and front-ness: /i-ɪ/, /u-ʊ/, /ɛ-æ/, /ʌ-ɑ/. The vowels in each pair differ in their intrinsic duration [5] and F0 [6] making duration and F0 (potential) intrinsic cues for their identification. F0 is predicted to cue prominence, based on the finding that F0 is

more critical for prominence perception than duration [16], and based on the nature of our F0 manipulation.

Predictions for F0 are that prominence-lending F0 patterns should lead listeners to expect a prominent vowel variant: that is, hyperarticulation in high vowels and sonority expansion in non-high vowels. Empirically, this predicts that prominent F0 leads to *fewer* higher vowel responses in high vowel pairs /i-ɪ/, /u-ʊ/ (acoustically higher-vowel F1/F2 required to perceive /i/ and /u/ under an expectation of hyper-articulation). Conversely, prominent F0 is predicted to lead to *more* higher vowel responses in non-high vowel pairs /ɛ-æ/, /ʌ-ɑ/ (aligning with sonority expansion, whereby vowels become lower and backer in the vowel space). F0 as an intrinsic cue predicts the opposite effect for high vowels (more /i/ and /u/ responses with high F0). Duration also cues prominence [17], in which case longer duration may generate the same effects as prominent F0. Duration as an *intrinsic* cue predicts the opposite (more /i/, /u/, /æ/ and /ɑ/ responses with longer duration, as each of these is longer than their counterpart in the pairs tested [5]). We also examine how duration and F0 influence vowel perception in a carrier phrase versus isolation. This allows us to test the potential up-weighting or down-weighting of any of these cues as a function of the presence/absence of phrasal context.

2. METHODS

Stimuli were created by re-synthesizing the speech from a female speaker of American English, recorded in a sound-attenuated booth, using a Shure SM27 Large-diaphragm Condenser Microphone and pop filter. The audio was recorded at 44.1 kHz and 32 bit depth. The speaker was instructed to produce the phrase "I will say *x* now" (where *x* is one of eight target words) with 4 levels of emphasis (no-emphasis, emphasized, more emphasized, very emphasized) on *x*. The emphasis instructions were intended to elicit different levels of F0-based prominence which were used as the basis for resynthesis, described below.

All target words used a /k_d/ frame. The following words were used: "keyed", "kid", "cooed", "could", "ked" (a type of insect), "cad", "cud", "cod". The starting point was the speaker's production of the third level of emphasis for each word. We resynthesized F1 and F2 [18] to create a 7-step formant continuum between each pair of vowels: "kid-keyed", "ked-cad", "could-cooed", "cud-cod". Formant resynthesis was followed by pitch accent (F0) resynthesis. The goal of F0 resynthesis was to create two F0 conditions, one with a relatively prominent F0 contour over the target vowel, and one with a relatively non-prominent contour. To keep the F0 resynthesis relatively controlled while also being

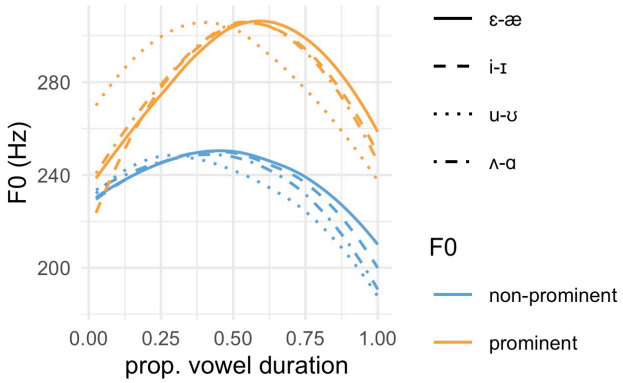


Figure 1: The F0 values for the F0 manipulation for each contrast (indicated by line type).

naturalistic, we resynthesized F0 to match naturally produced contours for each vowel contrast. For a given contrast, a representative less-prominent "no-emphasis" (roughly H*) contour was selected, and a representative more-prominent "more emphasized" (roughly L+H*) contour was selected. These were each overlaid on the continuum using the PSOLA method in Praat [19], with both conditions thus created by resynthesis. The F0 contours for each contrast and prominence condition are shown in Figure 1. Within a contrast the same distinction is apparent: the prominent condition shows a later F0 peak, higher scaling of the peak, and overall higher F0. We are not testing the effect of a specific F0 contour or shape (i.e. if we had controlled the contours across contrasts) but rather testing the effect of these relative differences which are *of the same nature* across contrasts. Finally, we created a duration continuum with five equidistant steps [20], whose endpoints varied by contrast and were set according to the speaker's productions:110-270ms for i/-/ɪ/, 80-250ms for /ɛ-æ/, 150-320ms for /u/-/ʊ/ and 60-250 for /ʌ/-/ɑ/. Altogether, there were 280 stimuli (4 contrasts * 7 formant steps * 2 pitch accents * 5 duration steps). Target words were spliced into the same sentence frame "I'll say __ now". In one version of the experiment they were presented in this carrier phrase, and in another were presented in isolation (a between-subjects manipulation).

Participants were recruited online from Prolific, 30 for each carrier phrase condition for a total of 60. All were self-reported native speakers of American English with normal hearing. Participation was remote, with participants seated in a quiet space using headphones. **The Experiment** was a two-alternative forced choice task. Stimuli were presented auditorily with orthographic representations onscreen and categorized by key press. Each stimulus was presented once for a total of 280 trials (fully randomized). All trials were analyzed.

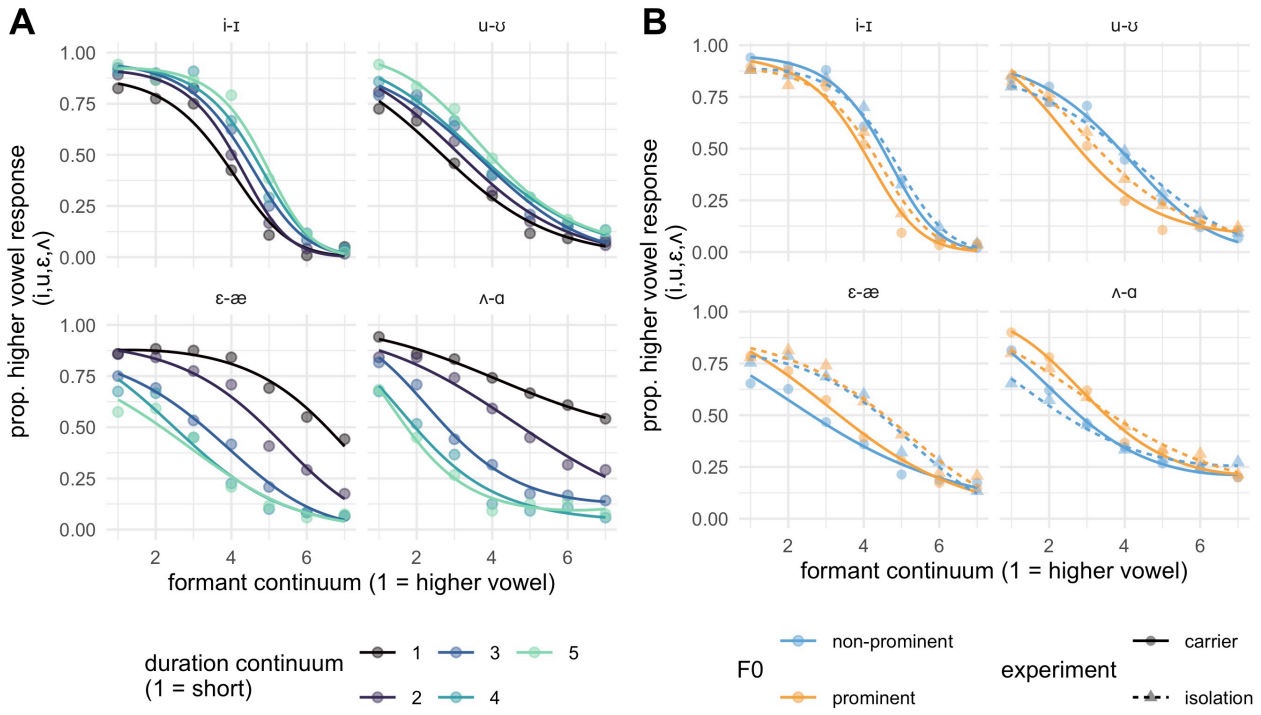


Figure 2: Panel A shows categorization responses as a function of formants (x axis) and duration (line coloration), with the proportion of higher vowel responses for each contrast plotted (/i,u,ɛ,ʌ/). Panel B shows categorization pooled by duration and split by the carrier phrase (line type) and F0 (coloration) manipulations.

Statistical analysis of the data was carried out with Bayesian mixed effects logistic regression, implemented using *brms* [21]. We fit a separate model for each of the four contrasts we tested. The dependent variable was coded with the higher vowel in the pair mapped to 1 (/i/, /ɛ/, /u/, /ʌ/ mapped to 1; /ɪ/, /æ/, /ʊ/, /ɑ/ mapped to 0). The fixed effects in each model were formant continuum step, duration continuum step (both scaled and centered), prominence condition (prominent mapped to 0.5, and non-prominent mapped to -0.5), and experiment (isolation mapped to 0.5 and carrier phrase mapped to -0.5). All interactions were included as well. Random effects were specified as by-participant random intercepts, with all fixed effects and interactions as random slopes, save for experiment (as it was not a within-participant manipulation). We report the median for an estimate's posterior and the probability of direction (pd), computed using the R package *bayestestR* [22] which gives the percentage of a posterior distribution with a given directionality. A distribution precisely centered on 0 (no effect) would have a pd of 50, while one with a consistently estimated positive or negative effect will have a pd approaching 100. pd > 97.5 corresponds to the 95% credible intervals (CrI) excluding zero, which we consider a credible effect. All of the data, code for running the statistical models, full model summaries, and an expanded version of Fig. 2 are available on the OSF at: https://osf.io/cfsru/.

3. RESULTS

First we consider the effects of formants and duration for each vowel contrast. We find an expected effect of the formant continuum for each contrast (/i/-/ɪ/: β = -3.70, pd = 100; /u/-/ʊ/: β = -3.14, pd = 100; /ɛ/-/æ/: β = -2.15, pd = 100; /ʌ/-/ɑ/: β = -2.05), shown as a decrease in categorization responses left to right along the x-axis in Fig 2A.

Duration credibly impacted categorization for each contrast as well. For both high vowel contrasts, longer duration led to an increase in /i/ and /u/ responses, respectively (/i/-/ɪ/: β = 0.60; /u/-/ʊ/: β = 0.40, pds = 100), while longer duration in the non-high contrasts increased /æ/ and /ɑ/ responses (/ɛ/-/æ/: β = -1.48, /ʌ/-/ɑ/: β = -1.83, pds = 100). Duration effects clearly support the intrinsic duration predictions [5]: longer duration favors perception of an intrinsically longer vowel, especially for non-high vowels [23].

F0 additionally had a credible effect for each contrast, which showed opposing directionality as a function of vowel height. In high vowel contrasts higher F0 decreased higher vowel responses (/i/-/ɪ/: β = -0.75, /u/-/ʊ/: β = -0.65, pds = 100). The effect showed an opposite directionality credible effect in non-high vowel contrasts whereby higher F0 *increased* higher vowel responses (/ɛ/-/æ/: β = 0.28, pd = 99; /ʌ/-/ɑ/: β = 0.67, pd = 100). Both of these

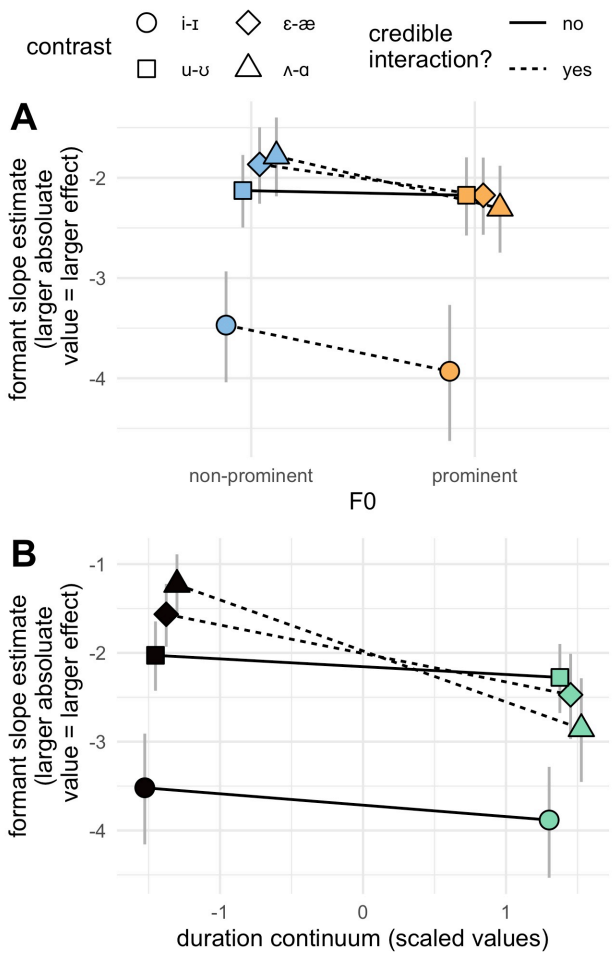


Figure 3: Estimates and 95%CrI from the models for the effect of formant step (y axis), as a function of F0 condition (Panel A) and duration continuum (Panel B).

effects are visible as differences across line coloration in Figure 2B. The carrier phrase manipulation exerted a credible main effect for only /ɛ/-/æ/ ($\beta = 0.97$, pd = 100), and otherwise did not interact with any predictors or show a main effect for other contrasts. The F0 effect confirms the prominence predictions: prominent F0 for high vowels decreases higher vowel responses, consistent with an expectation of acoustic hyperarticulation. Prominent F0 for non-high vowels increases higher vowel responses, consistent with an expectation of acoustic sonority expansion.

We next considered the potential *enhancement role* of both F0 and duration in vowel perception in examining the interaction of each with formant cues. Figure 3A shows the estimate for the effect of formants across F0 conditions, extracted using the *estimate_trends* function [24]. Note that a larger absolute value (more negative) effect corresponds to an up-weighting of formant cues. All contrasts but /u/-/ʊ/ showed a credible interaction between F0 (prominence) and formants (pds > 99). As shown in Figure 3A, this interaction reflected a stronger/larger effect of formants in the prominent F0 condition: prominence-based enhancement of formant cue use. Comparable duration-based enhancement was found for non-high vowel contrasts, shown in Figure 3B, both of which showed a credible interaction between vowel duration and formant continuum. In line with the F0 effect, this suggests longer (more prominent) durations lead listeners to upweight formant cues in perception, for non-high vowels in particular.

4. DISCUSSION AND CONCLUSION

This study finds clear support for duration as an intrinsic cue to vowel quality, while, conversely, the pattern of F0 results is consistent with F0 cuing prominence. Critically, the prominence effect varied by vowel height, comporting with the patterns of prominence strengthening in the speech production literature [9,10], while adding new data on how pitch accent type mediates prominence perception in this regard, thus building on [14,15]. We also found that F0-based prominence enhanced formant cue usage for three of the four contrasts tested. The same was observed for the non-high vowel contrasts as a function of duration, in which longer durations led a larger effect of formant cues. We suspect this effect is restricted to non-high vowels due to the higher weight of duration cues for these contrasts (see βs for duration in text and Figure 2A), which may lead to more salient duration-based enhancement.

In summary, results show a dual role for F0-based prominence in vowel perception: one that mediates which vowel category is perceived (as a function of prominence strengthening), and also enhances the perception of formant cues. Duration also serves a dual role in the sense that it constitutes an intrinsic cue to vowel quality, and additionally leads to upweighting of formants (for non-high vowels only). The present study thus advances our understanding of prominence effects on vowel perception across the vowel space in American English and addresses the question of multiple prominence functions in speech perception. More generally, this result, in the vein of other recent work, underlines the importance of considering prosodic features in segmental/lexical processing [cf. 25,26]. In future research, we plan to test the possibility of enhancing effects for other prominence cues and segmental cues, with the general prediction that prominence should enhance the use of segmental cues, e.g. VOT. Future work will also benefit from testing how these effects extend to L2 learners of English, and how they are impacted by speaker-specific properties and generalize across speakers. Finally, we are interested to test how they relate to distributional learning by varying the co-occurrence distributions of duration and F0 in the stimuli (e.g., longer durations tend to occur with a prominent F0).

5. REFERENCES

[1] Lehiste, I. (1970). Suprasegmentals. Cambridge, MA: MIT Press.

[2] Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141-201.

[3] McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In *The handbook of phonetic sciences* (pp. 489-520). Blackwell.

[4] Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, *31*(4), 428-435.

[5] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

[6] Chen, W. R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, *87*, 101063.

[7] Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878.

[8] Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, *59*(2–3), 134–149.

[9] De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

[10] Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty & D. R. Ladd (Eds.), *Gesture, Segment, Prosody (Papers in Laboratory Phonology)*. (pp. 68–86). Cambridge: Cambridge University Press.

[11] Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: Evidence from formant structure. *INTERSPEECH*, 2535–2538.

[12] Garellek, M., & White, J. (2015). Phonetics of Tongan stress. *Journal of the International Phonetic Association*, *45*(1), 13-34.

[13] Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19-EL25.

[14] Steffman, J. (2021). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, *36*(5), 586-611.

[15] Steffman, J. A. (2020). *Prosodic prominence in vowel perception and spoken language processing*. University of California, Los Angeles.

[16] Jasmin, K., Tierney, A., Obasih, C., & Holt, L. (2022). Short-term perceptual reweighting in suprasegmental categorization. *Psychonomic Bulletin & Review*, 1-10.

[17] Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126–152.

[18] Winn, M. (2016). Vowel formant continua from modified natural speech (Praat script). http://www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt.

[19] Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (version 6.1.09). http://www.praat.org

[20] Winn, M. (2014). Make duration continuum (Praat script).http://www.mattwinn.com/praat/Make_Duration_Continuum.txt

[21] Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.

[22] Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. Journal of Open Source Software, 4(40), 1541.

[23] Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013-3022.

[24] Makowski D, Ben-Shachar M, Patil I, & Lüdecke D (2020). Estimation of Model-Based Predictions, Contrasts and Means. *CRAN*.

[25] McQueen, J. M., & Dilley, L. C. (2020). Prosody and spoken-word recognition. In *The Oxford handbook of language prosody* (pp. 509-521). Oxford University Press.

[26] Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, *108*, 104034.

AMERICAN LISTENERS' RECOGNITION OF SENTENCES UNAFFECTED BY RACIAL AND ETHNIC PRIMES

Drew J. McLaughlin[1] & Kristin J. Van Engen[2]

[1]Basque Center on Cognition, Brain and Language; [2]Washington University in St. Louis
[1]d.mclaughlin@bcbl.eu; [2]kvanengen@wustl.edu

ABSTRACT

A listener's ability to accurately understand speech can be affected by racial and ethnic information about the speaker. For example, the presentation of an East Asian versus a White face has been shown to lead to better understanding of Mandarin-accented English and poorer understanding of American- or Canadian-accented English. This social priming effect has yet to be examined with images of Latinx or Black faces for Standard American English (SAE) in American listeners. We used a matched-guise paradigm to present listeners with Black, East Asian, Latinx, and White primes (images of faces), as well as a control prime (a blurred silhouette), during transcription of SAE-accented sentences.

Results indicated no effect of the priming manipulation. In particular, the lack of an East Asian prime effect on recognition accuracy differs from prior work. We discuss the possibility that these effects may be dependent on characteristics of specific social contexts.

Keywords: speech perception; social priming; race and ethnicity

1. INTRODUCTION

Spoken language carries paralinguistic information such as cues to the speaker's race, gender, and social class [1]. For example, American listeners can identify a non-standard American English dialect from a snippet of speech as short as "hello" [2]. It is important to note that the "standard" dialect of a region is typically determined by which group(s) of speakers hold the most power [3]. As such, in the United States, the Standard American English (SAE) dialect primarily reflects the dialect(s) of White (majority race) speakers. There are also multiple other dialects spoken by communities of minority racial and ethnic groups in the United States. The best documented of these is African American Language (AAL), which differs from SAE grammatically [4] and phonologically [5]. Other dialects appear to be borne from second language (L2; "foreign") accents, such as Asian-influenced English and Spanish-influenced English. Most notable for the present study, these dialects differ from AAL because they are less homogeneous, representing influences from many different L2 accents and languages.

Capturing this difference among AAL, Spanish-influenced, and Asian-influenced dialects, a recent racial position model of Americans [6] indicates that Asian and Latinx Americans are typically stereotyped as more *culturally foreign* (a term encompassing multiple attitudes and stereotypes about immigration, as well as having a foreign accent) than White and Black Americans. Thus, while some Americans hold negative stereotypes toward Black Americans, they do not appear to be associated with foreign accents in the same way that Asian and Latinx Americans are.

1.1. Social Priming

Prior work indicates that a listener's ability to understand speech may be affected by racial and ethnic information about the speaker (here forward referred to as *social priming*). One study, for example, found reduced recognition accuracy for L1 Canadian-accented English speech (presented in pink noise) when still images of Chinese-Canadian talkers were presented (as compared to presenting the audio with a fixation cross only). This same reduction in accuracy did not occur when images of White-Canadian talkers were presented [7]. However, complementing this work, [8] examined American listeners' recognition accuracy for Mandarin Chinese-accented speech (presented in babble). Listeners were assigned to images of an East Asian, White, or control (i.e., a silhouette) face, and results indicated that the East Asian prime *facilitated* speech recognition compared to a White prime (differences with the control prime were non-significant). Most notably, when examined in conjunction with the results of [7], it appears that minority race/ethnicity primes do not always negatively affect speech perception. Instead, such cues may prime listeners to expect specific dialect/accent qualities, which can either facilitate or inhibit perception.

This interpretation aligns with an exemplar model of speech recognition [9, 10, 11], in which phonetically-detailed episodic traces ("exemplars") are stored in the lexicon and linked to social information. Across a listener's life, patterns are extrapolated across to create larger social categories that are linked to phonetic (and higher-order

linguistic) patterns. These socio-phonetic connections thus provide a framework by which top-down information can influence speech perception.

However, while the results of [7] and [8] complement each other, recent work has not consistently revealed "negative" social priming effects for standard, L1-accented speech. In L1 German listeners of various ages, the effects of East Asian and White visual primes on recognition of standard L1-, non-standard (regional) L1-, and L2-(Korean-) accented speech were mixed [12]. Korean-accented German presented with an East Asian prime was more accurately perceived by teens and older adults (not younger adults), but there were no effects of priming for L1-accented speech.

1.2. Study Aims and Hypotheses

In the present study, we examined potential effects of racial and ethnic primes on perception of SAE speech for White L1 American listeners. We thus attempt to conceptually replicate the work of [7] in American listeners and extend it by examining additional racial and ethnic primes (Black and Latinx, in addition to East Asian and White). Our modified experiment includes a full matched-guise design between visual primes and SAE voices.

Based on the results of [7], we predicted that listener recognition accuracy for the East Asian priming condition would be poorer than the White priming condition. For Latinx priming condition, we predicted a similar outcome, given that Latinx and Asian Americans are similarly stereotyped as having foreign accents [6] and Spanish-influenced accents are prominent in the United States.

For the Black priming condition, however, we did not have a predicted outcome. Based on an exemplar model of speech perception, one could expect L1 American listeners to associate Black faces with AAL. This top-down effect on speech processing could result in poorer speech recognition accuracy given the present study's use of SAE stimuli. However, it is also possible that social priming effects are primarily driven by associations between racial/ethnic groups and L2 accents. In this case, one may expect no negative priming effect for the Black condition, because Black Americans are not associated with cultural foreignness in the same way as Asian and Latinx Americans [6].

2. METHODS

Data collection occurred online during Fall 2020. We acknowledge that the generalizability of our findings may be limited by the extenuating circumstances of the COVID-19 pandemic. Approval for the present study was acquired from the Washington University in St. Louis Institutional Review Board.

2.1. Participants

White young adult participants ($N = 126$) with normal hearing were recruited to participate online from the Washington University in St. Louis Psychology Participants Pool using SONA Systems. From this sample, 25 participants were excluded from analyses for one or more of the following reasons: Reporting that English was not a language learned from birth, reporting that they did not use headphones during the task, reporting that their data should be excluded (e.g., "I was too distracted"), and data collection error. This resulted in a final sample of 101 participants (age mean = 19.11; age $SD = 1.03$). All participants received course credit as compensation. Seventy-two participants reported that they were female and 29 reported that they were male.

2.2. Stimuli

Auditory stimuli included 60 unique sentence-length items collected from six separate corpora stored in the Northwestern SpeechBox [14] database. All recordings were from L1 speakers of Standard American English. One additional male was recorded in-lab to bring the numbers to an even 30 male voices and 30 female voices. The sentences were all semantically normal and contained 3-7 (mean = 4.62) common keywords and 4-10 (mean = 7.08) total syllables each. All words were included in the keywords count with the exception of the determiners "a" and "the".

All auditory stimuli were levelled to 65 dB sound pressure level using the UCLA Phonetics Lab's intensity scaling Praat script [15]. Speech-shaped noise was created using the average long-term spectrum of the entire batch of target files, and then combined with each file at a -8 dB signal-to-noise ratio (SNR). During the experiment, speech-shaped noise began two seconds before sentence onset and continued two seconds after sentence offset.

In order confirm that the speakers' accents were perceived to be SAE, we conducted an online pilot. Fourteen young adult subjects (none overlapping with the present study) rated the stimuli on a scale from 0 to 10, where 0 indicated "not at all standard" and 10 indicated "extremely standard". Stimuli were presented in speech-shaped noise at -8 dB SNR as described above. Overall, the mean rating of the stimuli was 6.56 ($SD = 1.44$). We explore whether the variation in these accent ratings affects social priming in the analyses of the main dataset.

Two additional recordings were created for attention-check trials. These recordings were of the

same female voice saying either "Please type a single G" or "Please type a single Q." These catch trials were presented without background noise in combination with a female control prime.

For the visual stimuli, images of faces were selected from the Chicago Face Database [16]. Six female and six male faces were selected for each race and ethnicity (White, Black, East Asian, and Latinx; 48 faces total) based on norming data. The faces were all rated highly for prototypicality of race/ethnicity and approximately matched for age, attractiveness, and happiness. Using these images, two control images were also created. Webmorph [17] was used to combine all of the female faces and (separately) all of the male faces into two morphed images. Next, Adobe Photoshop [18] was used to blur the images. First the face was selected (i.e., from hairline to chin, and from ear to ear) with a lasso tool and blurred with a 120 pixel-radius Gaussian blur, and then the entire image was blurred with a 40 pixel-radius Gaussian blur. This resulted in a silhouette image that was ambiguous in terms of race and ethnicity.

2.3. Procedure

The experiment was hosted on Gorilla [19], an online experiment-building platform.

The task began with a consent document and instructions. Participants were asked to complete the experiment in one sitting with their full attention. For the speech transcription task, they were told to look at the picture on the screen while listening, and type the full sentence when prompted. Instructions indicated that headphones were required for the task. Before beginning, an example file was presented for participants to adjust their volume as needed. The file could be replayed multiple times. The voice in the example did not occur in the experiment.

The transcription task contained 62 trials (60 test items and two attention checks). The attention-check stimuli occurred at fixed points one-quarter and three-quarters through the task. Halfway through the task, a self-timed break was offered to participants. The sentences-in-noise and image primes were randomly paired across subjects, with the exception that female faces were always assigned to female voices and male faces were always assigned to male voices. With the exception of the male and female control images, none of the face images repeated during the task. Audio stimuli also did not appear more than once. The order of presentation of each audio file and prime was also randomized across subjects. Thus, if any files were inadvertently harder than others, this challenge was equally represented for all priming conditions across subjects. Within each trial, the image was on screen for the full length of the audio file. Thus, the priming image appeared when the speech-shaped noise began, allowing two seconds of exposure before sentence onset.

After the transcription task, subjects completed two additional unrelated pilot tasks. Demographic and language information was collected at the end of the experiment. Additionally, after a reminder that their responses would not affect their compensation, subjects were prompted to report if they used headphones (as instructed) and if their data should be excluded for any reason (e.g., "I was not paying enough attention").

2.4. Analyses

Transcription accuracy was determined using the R package Autoscore [20]. Common misspellings predetermined by the Autoscore package were scored as correct, and differences in tense (*kick* versus *kicked*) and plurality (*cat* versus *cats*) were allowed. Generalized linear mixed-effects regression (GLMER) was used to model the data. The predicted variable, transcription accuracy, was treated as a grouped binomial of correctly identified keywords versus incorrectly identified (or missed) keywords. For example, if for the target sentence *"the **gray mouse ate** the **cheese**"* a subject transcribed *"the **gray** house had **cheese**"*, this would be scored as two keywords correct and two keywords incorrect/missing. Although there are multiple keywords per sentence/trial, the GLMER model is nonetheless able to predict the counts of the correct versus incorrect/missing groups of data using a binomial regression.

Random intercepts by subject and item (i.e., audio file) were specified, and random slopes of prime were included by subject.

3. RESULTS

The fixed effect of prime was dummy-coded with the Control condition as the reference level. The log-likelihood comparison of GLMER models with and without the effect of prime indicated that it did not significantly improve model fit ($\chi^2 = 3.06$, $df = 4$, $p = .55$). As shown in Figure 1, accuracy did not differ across priming conditions.

We also examined the effect of sex (i.e., male versus female speaker/prime), which significantly improved model fit ($\chi^2 = 5.89$, $df = 1$, $p = .02$). Participants had better transcription accuracy for the female voices/primes than the male voices/primes ($\beta = -1.30$, $p = .01$). The effects of prime and sex did not significantly interact ($\chi^2 = 0.71$, $df = 4$, $p = .95$).

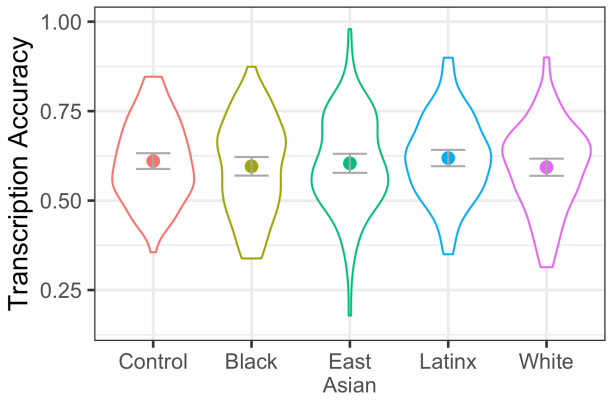


Figure 1: The non-significant effect of prime on transcription accuracy is visualized with violin distributions, mean points, and 95% confidence intervals.

Lastly, we explored whether individual differences among the speakers' voices may influence social priming. Mean standardness ratings of each stimulus (where 0 indicated "not at all standard" and 10 indicated "extremely standard") from the pilot data were added to the GLMER model with the effect of prime. Model fit was significantly improved by the effect of ratings ($\chi^2 = 37.61$, $df = 1$, $p < .001$) but not the interaction between ratings and prime ($\chi^2 = 3.37$, $df = 4$, $p = .50$). The direction of the effect of ratings indicated that higher (more standard) rated stimuli were also more intelligible ($\beta = 1.01$, $p < .001$).

4. DISCUSSION

Speakers with non-standard accents can face prejudice and stigmatization that impacts their everyday lives, including their ability to secure housing and employment [21, 22]. For second language- (L2-) accented speakers these negative effects of listener attitudes are also present, such that reduced comprehensibility of their speech affects how listeners perceive their intelligence, among other traits [23].

Prior work indicates that another challenge to successful communication may be perceived "incongruencies" between a speaker's race or ethnicity and their accent. However, these negative social priming effects have not been observed consistently across listener groups [12]. The present study focused on the effects of visually-presented racial and ethnic primes on the perception of SAE speech. Our stimuli included Black, East Asian, Latinx, and White primes, as well as a control image.

To our surprise, no significant effects of the priming images emerged. Presentation of different racial and ethnic primes did not affect listeners' recognition accuracy for SAE-accented sentences presented in noise. One methodological explanation may be the effect of the noise on listeners' ability to identify each speaker's accent. Indeed, our exploration of accent ratings indicated that more intelligible stimuli were also rated as sounding more standard, suggesting that listeners may be less able to identify a speaker's accent under adverse listening conditions. Similar results were found in [24], which showed that adding noise to the speech signal affected judgments of L1 speakers, such that listeners rated them as sounding more L2-accented. In other words, it is possible that by degrading the speech signal with noise, the present study reduced the possibility of observing a negative social priming effect. However, the initial study to observe this effect [7] presented sentences in pink noise at an exceptionally difficult signal-to-noise ratio (speech was only approximately 20% intelligible). It is possible that the findings of [7] are specific to the area in which the research was conducted (Vancouver, B.C., Canada). A key future direction for work examining social priming effects will be the influence of listeners' attitudes and implicit expectations of foreign and non-standard accents. The present work examined listeners who were White, young adult, Americans. Although the participants were all recruited from a university in the Midwestern United States, it is likely that their individual linguistic experiences vary. By examining individual differences in social networks as well as exposure to dialect and accent variation, future work may be able to determine the role of previous experience on social priming.

To the best of our knowledge, the present study is the first to examine Black and Latinx primes for American listeners and SAE accent. We hypothesized that, based on an exemplar model, the White listeners in our sample may associate Black faces with AAL and Latinx faces with Spanish-influenced English. If this were the case, a "negative" social priming effect could occur, given the use of SAE stimuli. Our results did not reveal this to be the case. It remains unclear whether social priming effects for standard L1 accent, as documented in [7], occur for additional minority race and/or ethnicity primes for American listeners.

5. SUMMARY

A listener's ability to understand speech can be affected by racial and ethnic information about the speaker. We examined this social priming effect using images of Black, East Asian, Latinx, and White faces (as well as a control silhouette) and SAE accent in White American listeners. Results indicated no effect of the priming manipulation. We suggest that social priming effects may only occur in specific populations of listeners and recommend examination of individual listener differences in future work.

6. ACKNOWLEDGMENTS

This research was supported by Doctoral Dissertation Research Improvement Grant 2116319 from the National Science Foundation.

7. REFERENCES

[1] Kraus, M. W., Torrez, B., Park, J. W., & Ghayebi, F. (2019). Evidence for the reproduction of social class in brief speech. *Proceedings of the National Academy of Sciences, 116*(46), 22998-23003.

[2] Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of language and social psychology, 18*(1), 10-30.

[3] Lippi-Green, R. (2012). English with an accent: Language, ideology, and discrimination in the United States. *Routledge.*

[4] Cukor-Avila, P., & Bailey, G. (1995, June). Grammaticalization in AAVE. *In Annual Meeting of the Berkeley Linguistics Society* (Vol. 21, No. 1, pp. 401-413).

[5] Edwards, W. F. (2008). African American Vernacular English: phonology. *Varieties of English, 2,* 181-191.

[6] Zou, L. X., & Cheryan, S. (2017). Two axes of subordination: A new model of racial position. *Journal of personality and social psychology, 112*(5), 696.

[7] Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America, 137*(5), 2823-2833.

[8] McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502-521.

[9] Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *Working papers in linguistics-Ohio State University Department of Linguistics*, 101-113.

[10] Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review, 105*(2), 251.

[11] Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency. *Frequency and the emergence of linguistic structure, 45*(137), 10-1075.

[12] Hanulíková, A. (2021). Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PloS one, 16*(10), e0259230.

[13] McLaughlin, D. J., & Van Engen, K. (under review). Social priming: Exploring the effects of speaker race and ethnicity on perception of nonnative accents.

[14] Bradlow, A. R. (n.d.) *SpeechBox.* Retrieved from https://speechbox.linguistics.northwestern.edu

[15] Vicenik, C. (2022, October 1). *Noise and Speech Manipulation*. Praat Script Resources. http://phonetics.linguistics.ucla.edu/facilities/acoustic/praat.html

[16] Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47*(4), 1122-1135.

[17] Lisa DeBruine. (2017, December 4). *Webmorph* (Version v0.0.0.9001). Zenodo. doi: 10.5281/zenodo.1073696

[18] Adobe Systems. (2002). *Adobe Photoshop 7.0*. Adobe Press.

[19] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods, 52*(1), 388-407.

[20] Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America, 145*(1), 392-399.

[21] Carlson, H. K., & McHenry, M. A. (2006). Effect of accent and dialect on employability. *Journal of employment counseling, 43*(2), 70-83.

[22] Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of language and social psychology, 18*(1), 10-30.

[23] Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and affective response toward accented English. *Language & Communication, 22*(2), 171-185.

[24] Gittleman, S., & Van Engen, K. J. (2018). Effects of noise and talker intelligibility on judgments of accentedness. *The Journal of the Acoustical Society of America, 143*(5), 3138-3145.

NO EVIDENCE FOR CONVERGENCE TO SUB-PHONEMIC F2 SHIFTS IN SHADOWING

Orhun Uluşahin[1], Hans Rutger Bosker[1,2], James M. McQueen[1,2], Antje S. Meyer[1,2]

[1] Max Planck Institute for Psycholinguistics, [2] Donders Institute for Brain, Cognition and Behaviour
orhun.ulusahin@mpi.nl, hansrutger.bosker@donders.ru.nl, james.mcqueen@donders.ru.nl, antje.meyer@mpi.nl

ABSTRACT

Over the course of a conversation, interlocutors sound more and more like each other in a process called convergence. However, the automaticity and grain-size of convergence are not well established. This study therefore examined whether female native Dutch speakers converge to large yet sub-phonemic shifts in the F2 of the vowel /e/. Participants first performed a short reading task to establish baseline F2s for the vowel /e/, then shadowed 120 target words (alongside 360 fillers) which contained one instance of a manipulated vowel /e/ where the F2 had been shifted down to that of the vowel /ø/. Consistent exposure to large (sub-phonemic) downward shifts in F2 did not result in convergence. The results raise issues for theories which view convergence as a product of automatic integration between perception and production.

Keywords: *Phonetic convergence, sub-phonemic variation, alignment, F2, speech shadowing.*

1. INTRODUCTION

During conversation, we constantly switch roles between speaker and listener, shifting between cognitive tasks. Despite this dynamism of natural conversation, language perception and production have historically been studied in isolation, and the resulting theoretical frameworks have largely focused on only one of these primary modes of language processing. Consequently, recent research has called for more comprehensive unified theories of language perception and production which can better explain how perception and production interact [1], [2].

One possible avenue for investigating the interaction between perception and production is afforded by the phenomenon of phonetic convergence, which broadly refers to the tendency of interlocutors to sound more like each other over time. It has been observed both in laboratory settings and in natural conversation [3], and has been operationalized as increasing similarity between interlocutors' speech rates [4], vowel formants [5], and VOTs [6] among other acoustic features. However, the automaticity of the mechanisms that underlie convergence and the sensitivity of these mechanisms to sub-phonemic variation is unclear. One account asserts that alignment between interlocutors (i.e., the gradual establishment of common ground in linguistic representations on the syntactic, lexical, and phonological level) is automatic, and that phonetic convergence may be a natural product of this alignment process [7], [8]. Conversely, an alternative account presents evidence for the influence of various social factors on convergence [3], [9]–[12]. However, much of the previous empirical work on convergence presents evidence based on a combination of acoustic measures for multiple (often covarying) acoustic properties and/or perceptual similarity measures [3]. Furthermore, sub-phonemic convergence to individual formants has only been documented for a restricted set of vowels using strictly acoustic measures [13]. Thus, the questions of whether interlocutors converge to phonemes or to acoustic features, and whether this occurs automatically, remain unanswered.

To investigate the automaticity of convergence to sub-phonemic variation, we created a unidimensional and low-level acoustic metric. Specifically, we measured convergence as a single formant shift (i.e., F2) on a single vowel (i.e., /e/) in trisyllabic words repeated in a shadowing task. We manipulated target words in the shadowing task such that the critical vowel /e/ had a substantially lower F2 than usual, around that of the Dutch vowel /ø/. Other formants and acoustic features, most importantly the F3 which also distinguishes the two vowels, were left untouched during the manipulations to conceal the aim of the experiment and prevent perceptual category shifts (i.e., manipulated vowels would still be perceived as /e/). The unidimensional measure therefore not only provided a strictly acoustic convergence metric, but also afforded extensive experimental control.

We tested whether female native speakers of Northern Standard Dutch converged to a sub-phonemic, unidimensional downward shift in the F2 of the vowel /e/ produced by a female native speaker. Our methodology employed a pre-test/post-test paradigm consisting of 3 tasks: (1) A reading task for obtaining participants' baseline F2 values for the vowels /e/ and /ø/, (2) a shadowing task in which participants repeated trisyllabic Dutch words containing the F2-downshifted critical vowel /e/

(henceforth referred to as /?/), and finally (3) a phonological 2-alternative forced-choice (2AFC) task where participants identified trisyllabic stimuli containing /e/, /ø/, or /?/ (i.e., the manipulated vowel) as containing /e/ or /ø/, to verify that our manipulations were indeed sub-phonemic. In line with theories which argue for automatic convergence, we hypothesized that participants' F2s for the vowel /e/ would decrease (i.e., converge in the direction of /?/) as a function of exposure to /?/ during the shadowing task, and that participants would still categorize manipulated (i.e., F2 downshifted) stimuli as /e/ in the 2AFC task.

2. METHOD

2.1. Participants

40 adult female native speakers of Northern Standard Dutch participated in the experiment. All participants reported normal hearing and normal or corrected-to-normal vision. We recruited only female participants to minimize the acoustic distance between the typical formants of our female speaker and the participants. All participants gave informed consent (Project code: ECSW-2019-019) and received monetary compensation.

2.2. Materials

We selected the vowels /e/ and /ø/ as reference points based on existing reports [14] of their mean formant values[1]. The low F1 difference between the two vowels and the large F2 difference enabled participants' convergence to be measured as a function of observed changes in their F2 relative to their baseline levels. Crucially, the F3 difference, which is essential for distinguishing the two vowels, allowed our manipulations to remain sub-phonemic.

Using CELEX [15], we selected 40 trisyllabic Dutch words as experimental items for the shadowing task. These words all had exactly one instance of /e/ (always in a stressed syllable), and no instances of /ø/ (e.g., *orchidee* [ɔrxiˈde] "orchid"). Moreover, replacing /e/ with /ø/ would render all experimental words non-words. Additionally, a set of 120 trisyllabic filler words was generated. The vowels /e/ or /ø/ did not occur in any of these filler words. Thus, a total of 160 words were selected for the shadowing task[1].

Another set of 20 trisyllabic words was selected from CELEX to be used in the reading task[1]. 10 of these words contained only one instance of the vowel /e/ in a stressed syllable and no vowel /ø/. The other 10 contained only one instance of the vowel /ø/ in a stressed syllable and no vowel /e/.

We recorded a female native speaker of Northern Standard Dutch. For the experimental words in the shadowing task (i.e., words containing /e/), we also recorded "/ø/ nonword versions" in which the vowel /e/ was replaced with the vowel /ø/ (e.g., [ɔrxiˈde] → [ɔrxiˈdø]). In this case, speech was elicited using prompts following Dutch orthography (e.g., *orchidee* → *orchideu*). We recorded these versions to measure a reference F2 value for our speaker's typical /ø/ for each word, thus assigning word-by-word "targets" for our F2 manipulations. The mean F2 and F3 measures (in the middle 50% of the vowels) in our speaker's /e/ and /ø/ recordings were comparable to the reference values in [14] [1].

The 40 experimental words were then manipulated. For each word, the F2 of the vowel /e/ was shifted to match the F2 of its "/ø/ version" based on our recordings (e.g., *orchidee*'s new F2 value was set to our speaker's F2 on *orchideu*). The formant manipulations were carried out using the Burg LPC method implemented in Praat [16]. The F1 and the F3 were left untouched. For each word, a lower-F2 version of the original vowel /e/, labelled /?/, was generated and recombined with the original signal. Measurements confirmed that our manipulations had the desired effect, resulting in an average F2 difference of 490 Hz between /e/ and /?/[1].

2.3. Procedure

After providing consent, participants entered a sound-attenuating room with a monitor, keyboard and microphone. The experiment was run using Presentation® (Version 22.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

The experiment began with the reading task. Participants saw trisyllabic Dutch words on screen and were asked to read them out loud. Each of the 20 words selected for the reading task was presented twice for a total of 40 trials. The words used in the reading task did not appear elsewhere in the experiment. On each trial, a random word was selected and presented on screen. After exactly 1 second, the caption *OPNAME* ("recording") appeared on screen, and participants had 3 seconds to read the word out loud.

After the reading task, participants performed three shadowing blocks. Each block contained a randomized combination of all 40 experimental words and all 120 filler words. In each trial, participants saw a fixation cross while listening to the auditory stimulus for the duration of the whole word (i.e., about 1 second). At word offset, the fixation cross was replaced with the caption *OPNAME*, indicating the 3-second-long response period for each trial. We instructed participants to "repeat [the words]

as quickly and as accurately as possible" without explicitly demanding that they imitate any unusual features they might hear in some of the trials.

Finally, participants performed a 2AFC task in which they categorized stimuli as either containing the vowel /e/ or /ø/. This task used the 40 experimental words from our speaker that were initially recorded for the shadowing task. However, it also incorporated the alternative recordings in which the speaker had replaced the vowel /e/ with the vowel /ø/ (e.g., nonword *orchideu*), as well as F2-downshifted versions of each word. Thus, any item could be heard in any of the three forms. Four evenly distributed pseudo-randomized trial sets each contained a total of 40 trials, and each trial contained only one utterance of one critical vowel (e.g., a trial could feature the word *orchidee*, *orchideu* or *orchid?* where "?" denotes the manipulated vowel). Of the 40 words in each set, 8 contained the vowel /e/, 16 contained the vowel /ø/, and another 16 contained /?/. We used a higher proportion of /ø/ and manipulated words to gather more data that would reveal whether our manipulations triggered a perceptual category shift (i.e., /?/ perceived as /ø/).

During the task, participants saw two options in written form on screen (i.e., the /e/ spelling as in *orchidee* and the /ø/ spelling as in *orchideu*) upon word onset. They were instructed to indicate which written form represented the word they were hearing using keyboard controls.

2.4. Pre-processing

We used the WebMaus webservice [17] with the "nld-NL" parameters for the automatic word-level and phoneme-level parsing of the tokens recorded during the experiment. Following a data-driven approach for formant analysis (e.g., [18]), participants' audio recordings were analyzed with two sets of parameters: 1) 4 maximum formants at a 5 KHz range, 2) 5 maximum formants at a 5.5 KHz range. For all participants, the reliability of either set of parameters was evaluated as a function of F2 variability across experimental trials for a given participant (the greater the variability, the less accurate the tracking) as well as manual inspection of the estimated formant tracks in the spectrograms of 10 items per participant. All subsequent processing of a given participant's acoustic data was carried out using the best performing set of parameters. All formant values were extracted from the middle 50% of vowels.

3. RESULTS

6400 audio files ((120 experimental shadowing trials + 40 reading trials) * 40 participants) were subjected to the pre-processing steps identified above. We excluded the data from one participant who consistently produced F2s that were higher than 2400 Hz. Thus, formant values from 6240 files were used.

The participants' baseline F2s for the vowels /e/ (i.e., 2271 Hz) and /ø/ (i.e., 1842 Hz) were comparable to reports of typical Northern Standard Dutch [14] and our speaker's recordings.

We used a linear mixed-effects model [19] with the "lmerTest" package [20] in R [21] to test whether exposure to manipulated stimuli significantly predicted F2 convergence. With F2 (Hz) as the dependent variable, we included fixed effects of Task (Reading and Shadowing, the latter mapped onto the intercept) and Trial Number (z-scored) as predictors, and their interaction. Finally, we entered Word and Participant Number as random factors.

We found no effect of Task (β = -16.302, SE = 19.976, t = -0.816, p = 0.36) or Trial Number (β = -0.291, SE = 0.319, t = -0.915, p = 0.42) on F2 frequency. We also did not observe any significant interaction between these two factors (β = 0.3096, SE = 0.3183, t = 0.973, p = 0.33). The lack of an effect of Trial Number suggests that exposure to manipulated stimuli did not significantly predict any gradual change in F2 in either task. The lack of an effect of Task suggests that participants' F2s were not different overall (i.e., compared to the "baseline" reading measurement).

To analyze the output of the 2AFC task, we coded responses where participants categorized /?/ as /ø/ as "critical" given that they might indicate a perceptual category shift. Critical responses constituted only 1.56% of responses to all trials in which /?/ was used (N = 640, 16 trials with /?/ * 40 participants). Thus, the results from the 2AFC task suggested that, despite our large F2 shifts, we succeeded in avoiding perceptual category shifts.

4. DISCUSSION

The present study investigated whether consistent exposure to a manipulated (i.e., sub-phonemically downshifted) vowel formant (i.e., F2) influenced participants' speech production and resulted in phonetic convergence in a shadowing task (i.e., lower F2s for the vowel /e/ in participants' speech). We had hypothesized that participants would gradually shift their F2s for the vowel /e/ towards the *targets* represented by our manipulated stimuli, as a function of increased exposure. This hypothesis was in line with a view of convergence as an automatic process which would be sensitive to (consistent) sub-phonemic variation irrespective of participants' awareness of the F2 manipulation. Although a few participants (n = 4) followed a pattern of consistent

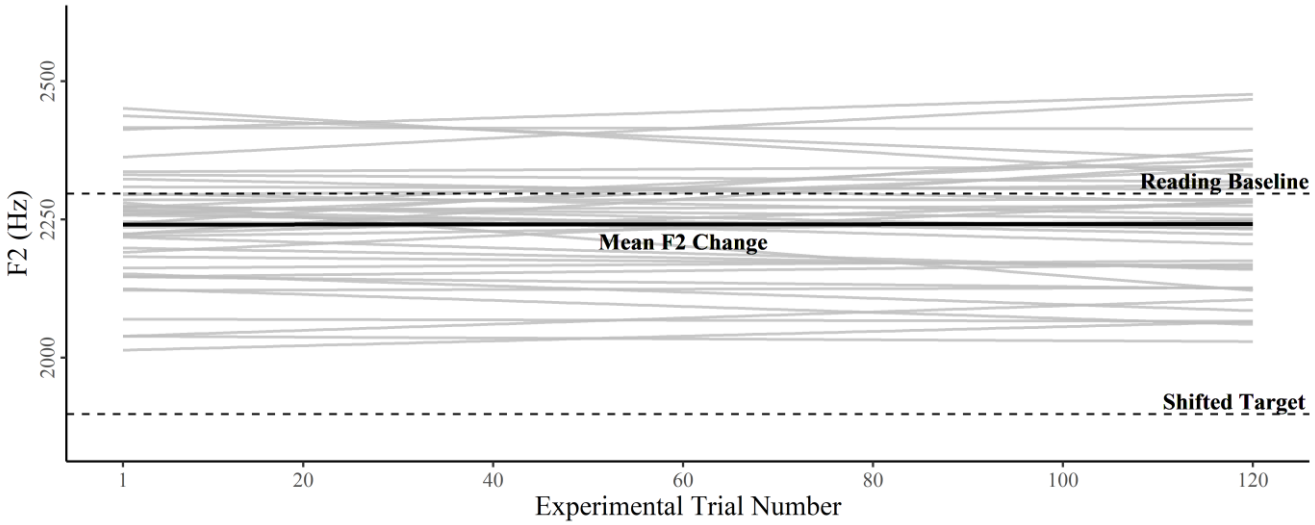


Figure 1: Each solid grey line follows the mean F2 trend of a participant. The solid black line follows the mean across participants. The dashed lines indicate, respectively, participants' baseline F2 values for /e/ and the speaker's average F2 in experimental items containing /?/ ("Shifted Target"). Trial numbers on the x-axis represent the sequence of the experimental trials (i.e., excluding fillers) across all 3 blocks.

downward shifts in F2 across all trials, most participants' F2s remained consistent across trials with some participants even displaying a marginal increase in F2 across all trials (see Figure 1). Although numerically there was a small overall downward shift in F2 from the reading task to the shadowing task across all participants, this difference was not significant.

The results of the phoneme categorization task suggest that we successfully avoided perceptual category shifts. However, our design did not include a task which could directly measure whether any perceptual adjustments occurred in participants' category boundary between /e/ and /ø/. We had theoretical and empirical precedent for assuming that exposure to our manipulated stimuli would indeed result in adjustments in the perceptual stream. Our F2 manipulations were substantially larger than the just noticeable difference for F2 (i.e., our manipulations resulted in a minimum of 20% reduction in F2, with the just noticeable difference reported at 1.5% [22]), and there is evidence that perceptual adjustments to consistent novel stimuli are not only common but also quickly formed and tracked [23]. Future research may introduce intermittent or post-test perceptual measures into a similar design to address the question whether perceptual adjustments to F2 manipulations do occur but fail to translate into production, or whether a sub-phonemic F2 shift alone fails to facilitate perceptual adjustments in the first place.

Although the absence of perceptual category shifts due to the sub-phonemic nature of the acoustic manipulations was an explicit goal of our design, the phonological non-salience of the manipulations (i.e., the manipulations did not introduce comprehension challenges) might also have been a factor behind the lack of evidence for convergence. Nevertheless, it is worth noting that convergence has been reported for other acoustic features of incoming speech that are sub-phonemic such as speech rate [4], and F0 in non-tone languages [24]. Thus, the influence of the non-salience of the manipulation must be evaluated in combination with other factors.

Finally, recent research [25] has highlighted that task engagement can play an important role in facilitating phonetic convergence. Our experimental paradigm, given its non-dialogic setting and the high number of repetitive trials, may have reduced our participants' level of engagement. Given the present study's success in concealing large sub-phonemic F2 shifts (i.e., /?/ was almost always perceived as /e/), future work may increase the experimental word ratio, while retaining sufficient experimental control to be able to apply precise sub-phonemic acoustic manipulations and maintain perceived naturalness.

Given these limitations, the lack of evidence for convergence in F2 despite consistent exposure to large F2 shifts in a consistent target (i.e., the vowel /e/) suggests that convergence in F2 might be dependent on covarying acoustic factors and resulting category shifts. Alternatively, our manipulations may have resulted in perceptual adjustments in our participants without producing observable differences in production. Nevertheless, the lack of evidence for convergence to large sub-phonemic shifts presents a potentially interesting issue for comprehensive theories of language perception and production that claim that there is automatic integration between these two primary modes of language cognition [1], [2], [7].

7. REFERENCES

[1] M. J. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behav. Brain Sci.*, vol. 36, no. 4, pp. 329–347, Aug. 2013, doi: 10.1017/S0140525X12001495.

[2] A. S. Meyer and J. M. McQueen, "Key Issues and Future Directions: Toward a Comprehensive Cognitive Architecture for Language Use," in *Human Language: From Genes and Brains to Behavior*, P. Hagoort, Ed. The MIT Press, 2019. doi: 10.7551/mitpress/10841.001.0001.

[3] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Atten. Percept. Psychophys.*, vol. 79, no. 2, pp. 637–659, Feb. 2017, doi: 10.3758/s13414-016-1226-0.

[4] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evol. Hum. Behav.*, vol. 34, no. 6, pp. 419–426, Nov. 2013, doi: 10.1016/j.evolhumbehav.2013.08.001.

[5] A. Schweitzer and N. Lewandowski, "Social factors in convergence of F1 and F2 in spontaneous speech," in *Proceedings of the 10th International Seminar on Speech Production, ISSP 2014*, 2014, pp. 391–394. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039164612&partnerID=40&md5=f2bfd3147d28baabd4e03c1b7891ecea

[6] K. Nielsen, "Specificity and abstractness of VOT imitation," *J. Phon.*, vol. 39, no. 2, pp. 132–142, Apr. 2011, doi: 10.1016/j.wocn.2010.12.007.

[7] S. Garrod and M. J. Pickering, "Why is conversation so easy?," *Trends Cogn. Sci.*, vol. 8, no. 1, pp. 8–11, Jan. 2004, doi: 10.1016/j.tics.2003.10.016.

[8] M. J. Pickering and S. Garrod, "Alignment as the Basis for Successful Communication," *Res. Lang. Comput.*, vol. 4, no. 2, pp. 203–228, Oct. 2006, doi: 10.1007/s11168-006-9004-0.

[9] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, "Phonetic convergence in college roommates," *J. Phon.*, vol. 40, no. 1, pp. 190–197, Jan. 2012, doi: 10.1016/j.wocn.2011.10.001.

[10] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, 2012, doi: 10.1016/j.wocn.2011.09.001.

[11] M. Babel, "Dialect divergence and convergence in New Zealand English," *Lang. Soc.*, vol. 39, no. 4, pp. 437–456, Sep. 2010, doi: 10.1017/S0047404510000400.

[12] A. Schweitzer and N. Lewandowski, *Social Factors in Convergence of F1 and F2 in Spontaneous Speech*. 2014. doi: 10.13140/2.1.3709.5689.

[13] S. Tilsen, "Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production," *J. Phon.*, vol. 37, no. 3, pp. 276–296, Jul. 2009, doi: 10.1016/j.wocn.2009.03.004.

[14] P. Adank, R. van Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1729–1738, Sep. 2004, doi: 10.1121/1.1779271.

[15] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX Lexical Database (CD-ROM)," 1996, Accessed: Jun. 18, 2022. [Online]. Available: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2339741

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Nov. 15, 2022. [Online]. Available: http://www.praat.org/

[17] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Comput. Speech Lang.*, vol. 45, pp. 326–347, Sep. 2017, doi: 10.1016/j.csl.2017.01.005.

[18] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1379–1393, Sep. 2009, doi: 10.1121/1.3180321.

[19] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *J. Mem. Lang.*, vol. 59, no. 4, pp. 390–412, Nov. 2008, doi: 10.1016/j.jml.2007.12.005.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[22] D. Kewley-Port and C. S. Watson, "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.*, vol. 95, no. 1, pp. 485–496, Jan. 1994, doi: 10.1121/1.410024.

[23] X. Zhang and L. L. Holt, "Simultaneous tracking of coevolving distributional regularities in speech.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 44, no. 11, p. 1760, 20181001, doi: 10.1037/xhp0000569.

[24] M. Simonet, "Intonational convergence in language contact: Utterance-final F0 contours in Catalan-Spanish early bilinguals," *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 157–184, 2011, doi: 10.1017/S0025100311000120.

[25] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Commun.*, vol. 138, pp. 50–66, Mar. 2022, doi: 10.1016/j.specom.2022.02.002.

[1] See supplementary material (formant settings used for each participant, mean reference and speaker formant values, a list of stimuli, and a log of F2 manipulations) at https://osf.io/wb2mn/

PERCEPTION OF COARTICULATORY VOWEL NASALISATION IN GERMAN LISTENERS AND ITS RELATIONSHIP TO PRODUCTION

Esther Kunay, Philip Hoole, Jonathan Harrington

Institute for Phonetics and Speech Processing (IPS), LMU Munich
es.kunay | hoole | jmh@phonetik.uni-muenchen.de

ABSTRACT

When making linguistic decisions, language users typically use fine phonetic details from the input they perceive. In this paper we present data from an adaptive staircase discrimination experiment in which German listeners were exposed to different patterns of nasalisation in VN sequences in a synthetical CVNCV target word. Results indicated that listeners had less difficulties in stimuli discrimination when differences affected the post-vocalic nasal consonant rather than the vowel alone, a finding which suggests that listeners are especially sensitive to those phonetic details they are experienced with. Production data from German speakers suggest that systematic differences in nasal timing occur in German depending on the nature of the postvocalic NC sequence. We further investigated whether listeners' individual perception results were related to their specific coarticulatory vowel nasalisation patterns in production.

Keywords: coarticulation, vowel nasalisation, adaptive staircase experiment, perception-production link.

1. INTRODUCTION

The emergence of distinctive nasal vowels is assumed to have its seeds in VN sequences in which coarticulatory vowel nasalisation is increased and the nasal consonant is lost over time [1, 2, 3, 4]. Specifically, vowels have been attested to become coarticulatorily nasalised more extensively if the nasal stop is followed by a voiceless obstruent, an issue that can be explained by physiological and aerodynamic principles [1, 5].

From a perceptual point of view, fine acoustic variations in vowel nasality may be used by listeners when making linguistic decisions [6, 7, 8, 9], albeit language groups differ with respect to perceptual sensitivity depending on their linguistic experience with these cues. For example, discrimination perception tests with American English and Botswana Ikalanga listeners showed that when the extent of overall nasality varied across the vowel and nasal stop, listeners could easily differentiate between two stimuli. However, discrimination became more challenging when the nasal gesture had a similar overall extent but was differently distributed across the segments [10]. This effect of perceived equivalence was stronger for the American English listeners, i.e. for the language group with more experience in coarticulatory vowel nasalisation. Further research has focused on the relationship between production patterns and perceptual sensitivity in the individual. With respect to contextual vowel nasality, findings suggest that such a relationship may be present for some language users, but not for others [9, 11, 12]. Data indicate that there is reason to postulate a greater flexibility in perception than in production, such that speakers who are inconsistent in producing a target property may as listeners nonetheless be sensitive to it (cf. [11]). This is even expected because speech perception is adaptive due to the need of constant adjustment to the linguistic environment.

In the current study we present data obtained from native speakers of Standard German, a language that does not exhibit contrastive nasal vowels. We tested whether a) German language users showed differences in perceptual sensitivity in two different tasks when the nasal gesture was variably shifted across VN segments and b) whether there was evidence for a relationship between the individual usage of coarticulatory vowel nasalisation in perception and production. Participants' production data came from a large-scale real-time magnetic resonance imaging (rt-MRI) study in which data were collected from the mid-sagittal plane to investigate the interacting gestures of the lips, tongue and soft palate during fluent speech [13, 14, 15]. In addition, synchronic acoustic data were recorded and analysed with respect to segmental boundaries. Some of the production data served as input in the current study for investigating the relationship between the usage of coarticulatory nasal cues in production and perception in the individual.

2. METHOD

2.1. Participants

An adaptive staircase perception experiment [16, 17] was run with 20 German participants (age: 19–35 years) who had previously participated as speakers in the rt-MRI study.

2.2. Stimuli

Stimuli patterns of AABA, ABAA, BBAB and BABB sequences were created, with the target stimulus in second or third position. As stimuli, the word [baːntə] was generated by articulatory synthesis and edited with regard to the velum lowering gesture; this was carried out with the software tool VocalTractLab [18], version 2.2 API for Windows (10 November 2017). The discrimination experiment comprised two conditions of nasalisation patterns: one in which a constant velum lowering interval was shifted along the vowel and nasal consonant, and one which contained a consistent nasal consonant but a temporally modified portion of vowel nasalisation. For both conditions, stimulus A was constructed in the same manner, representing a near-natural utterance of the target word with a nasal consonant of 80 ms and no vowel nasalisation present. Condition one ('constant condition') implied a constant interval of a total of 80 ms velum lowering. At the start, the difference ('delta') between the A and B stimuli corresponded to this maximum of 80 ms. In the course of the experiment, delta was decreased by shifting the nasalised interval as a whole in the B stimuli (details in sec. 2.3). Fig. 1 exemplifies the shifting procedure for the constant condition. Condition two ('extended condition') involved B stimuli with a constant nasal consonant of 80 ms but varying temporal nasalisation on the vowel, which ranged from zero to full nasalisation of 350 ms. Fig. 2 exemplifies three trial modifications of the nasalised time span during the vowel. At the beginning of the experiment, and as with the constant condition, delta corresponded to the maximum nasalisation difference between A and B (350 ms). Further, the stimuli were constructed by implementing the post-vocalic consonantal portion (/n/+/t/) with a tongue-tip gesture of a total of 160 ms. For stimulus A, the velum lowering duration was set to 80 ms, such that the nasal stop resulted from the synchronous onset of the tongue tip and the velar gesture. For all stimuli, the post-nasal voiceless stop was generated by placing a glottal gesture with the VocalTractLab specification "slightly breathy" over the remaining part of the tongue-tip gesture.

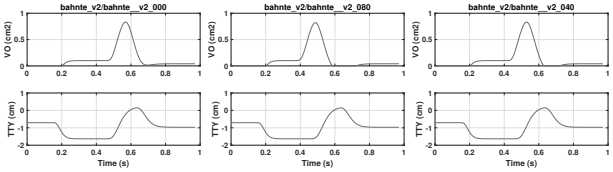


Figure 1: Shifting pattern for the constant condition. Left: basic configuration in stimulus A: tongue tip (TTY) and velum opening (VO) start at the same time point. Middle: VO interval shifted into the vowel by 80 ms (stimulus B): delta of stimuli A and B is at its maximum. Right: VO interval shifted into the vowel by 40 ms (stimulus B); delta of stimuli A and B is half of its maximum.

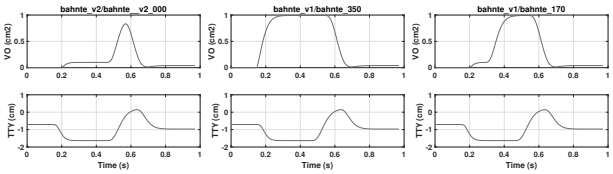


Figure 2: Shifting pattern for the extended condition. Left: basic configuration in stimulus A: tongue tip (TTY) and velum opening (VO) start at the same time point. Middle: overall VO of 350 ms during the vowel (stimulus B): delta of stimuli A and B is at its maximum. Right: VO of 170 ms during the vowel (stimulus B); delta of stimuli A and B is about half of its maximum.

2.3. Procedure

Listeners were informed that the differing stimulus was either in second or third position and they were told to press the appropriate key and to guess if necessary. Listeners were also informed that there were two conditions which would be tested separately. All participants started with the constant condition. No time limit was given and information about the outcome appeared immediately on the screen after each response was given. The overall advantage of the adaptive staircase method is an ongoing adjustment to listeners' discrimination thresholds: depending on each response outcome, delta is either increased or decreased. In this experiment, delta was divided by 2 after two consecutive correct responses and immediately increased by 50% if one answer was wrong. Thus, the gradual delta modification was expected to approach listeners' individual discrimination thresholds after several trials and was assumed to stay closely above and below this value. Since only delta was adapted, the trial patterns

randomly alternated with respect to the direction of the nasalisation approach. As the listeners were expected to finally oscillate around their threshold boundaries, the experiment ended after 12 turning points but with a fixed upper limit of 50 trials (fig. 3). The specific by-participant thresholds were determined from a fixed number of reversals that corresponded to the six consecutive turning points with the lowest threshold values (in Fig. 3: the last six points). Thus, listeners' individual perceptual thresholds referred to a period during which the listeners were relatively constant in their decisions.

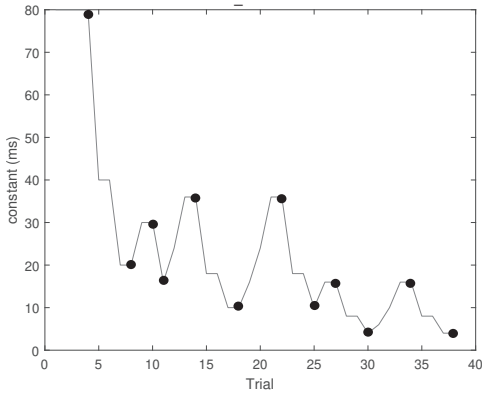


Figure 3: Exemplary staircase plot for one participant (constant condition). Delta is indicated by the y-axis, the trial number by the x-axis. The plot shows 12 turning points (black): one turn after the first two correct responses plus 10 turns visible in the curve plus one final turn when the twelfth turning point is reached.

3. RESULTS

3.1. Perception

Fig. 4 shows the means for the constant condition as a function of the means of the extended condition. Evidently, most participants showed overall finer discrimination abilities in the constant condition compared to the extended type. In addition, a more consistent pattern is evident for the constant condition, where 14 of the 20 data points are located below the value of 20 ms in contrast to a broader distribution for the extended condition. Moreover, responses for the two conditions did not necessarily correspond for individual listeners. For instance, while participants S14, S09 and S10 showed relatively low perception thresholds in both conditions, participant S04 had one of the lowest values in the constant condition but the highest in the extended version. Conversely, S05 exhibited a relatively high threshold in the constant

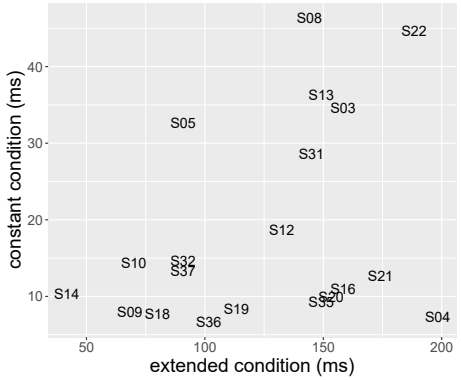


Figure 4: Means of the lowest six consecutive turning points for 20 listeners. X-axis: extended condition. Y-axis: constant condition.

condition but had less discrimination difficulties in the extended condition compared to most other listeners.

3.2. Perception and production

The perception results of the individual participants were combined with the production data obtained from the rt-MRI measurements. To elaborate a ranking in terms of by-speaker temporal vowel nasalisation in production, a total of 26 CVNV and CVNCV words in carrier phrases (CVNV n=303; CVNCV n=215) with tense vowels in pre-nasal position were analysed. The time span of the velum lowering gesture was determined by means of kinematic analyses of the velum signal obtained from the MR images. Vowel nasalisation (vow.nas) was defined as the time span between the point of maximum velocity during velum opening and the acoustic vowel offset. The overall velum lowering gesture (OVL) including the nasal stop referred to the interval between the points of maximum velocity during the velum opening and closure gesture[1]. For each participant, the amount of vowel nasalisation was divided by OVL for each target word. Thus, higher ratios corresponded to a larger extent of vowel nasalisation relative to OVL. Fig. 5 shows the by-speaker ratios ranked by the median based on the production target words. A Pearson correlation test was applied to the data of the extended condition (fig. 6 right). Results revealed no correlation between the production and perception values (r=0.230, p=0.33). For the constant condition (fig. 6 left), a Spearman correlation test[2] was run, which also did not reveal any significant correlation effect (r=-0.158, p=0.51). Data sets were combined by calculating the by-speaker ratio means, which were then defined as a function of the individual

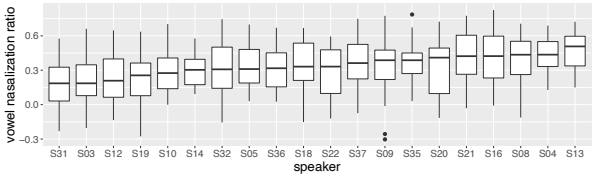


Figure 5: Ratio of vowel nasality to OVL (vow.nas/OVL) of target words with CVNV and CVNCV sequences for 20 speakers.

delta means from the perception test. Figure 6 shows the combined data sets for both experimental settings separately (left: constant condition, right: extended condition).

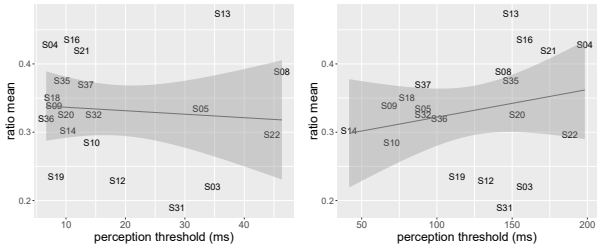


Figure 6: Means of the nasality ratios (vow.nas/OVL) and the perception thresholds of the lowest six consecutive turning points for 20 participants. Results are shown for the constant (left) and extended (right) condition.

4. DISCUSSION AND CONCLUSION

This study investigated the perceptual sensitivity of German listeners to vowel nasalisation as well as the relationship between speakers' nasality patterns in production and their discrimination skills as listeners in perception. Results revealed that listeners were overall more precise in identifying the target words when the constant interval of the nasalised portion was shifted as a whole across the vowel and nasal segment. More difficulties arose when the overall temporal extent of nasality solely varied within the vowel. Evidently, German listeners were highly insensitive to alterations of the temporal extent of vowel nasalisation and probably to the presence of coarticulatory vowel nasalisation at all. Consequently, the low thresholds in the constant condition probably do not indicate listeners' perception of vowel nasality but rather their sensitivity to other acoustic modifications coming along with the overall shifting. For example, listeners may have attended to the silent interval between the nasal offset and the offset of the /t/, which increased the more the 80 ms interval was shifted into the vowel. This increased interval may have been interpreted as a longer time span of alveolar stop closure for /t/, inducing the effect of a rather 'strong' /t/ compared to an oral stop with a short closure period similar to a weak /t/ or even /d/. As German language users are familiar with alveolar stop closure differences between /d/ vs. /t/, this might have helped them in identifying the target word more precisely. Alternatively, listeners may have relied on the stepwise absence (or in reverse trials on the increasing presence) of the nasal consonant. As with the extent of alveolar stop closure, German listeners can be assumed to have some experience with nasal weakening in certain contexts, as indicated by data discussed in [14, 15]. As shown in [15], velum lowering in CVNCV contexts was not just shifted more into the vowel compared to CVNV contexts but also reduced in its temporal extent, resulting in a shortened nasal consonant. If articulatory modifications of the nasal consonant systematically occur in production, German listeners might be sensitive to such nasal weakening patterns in perception as well, at least more than to alterations of vowel nasalisation. Considering participants' relationship between vowel nasalisation in production and perception, speakers indeed showed quite a range of anticipatory vowel nasalisation (Fig. 5). However, no correlation to their perception performances could be attested. Thus, the finding that across listeners perception skills might be related to some general experience with nasal weakening patterns in production cannot be applied to the individual language user in our data. Possibly for German, the variation patterns within a specific speaker's own production are too inconspicuous to systematically affect their perceptual sensitivity.

Overall, our experiment provides further insights into the question of which fine perceptual details listeners are able to detect when confronted with different nasality patterns. Results provide further evidence that listeners are especially sensitive to those acoustic variations they are more experienced with.

5. ACKNOWLEDGEMENTS

This research was funded by DFG (German Research Foundation) grant HA 3512/15-1 'Nasal coarticulation and sound change: A real-time MRI study' (J. Harrington & J. Frahm) and supported also by European Research Council Advanced Grant 742289 'Human interaction and the evolution of spoken accent' (J. Harrington).

6. REFERENCES

[1] P. S. Beddor, "A coarticulatory path to sound change," *Language*, vol. 85, no. 4, pp. 785–821, 2009.

[2] R. A. Krakow, P. S. Beddor, L. M. Goldstein, and C. A. Fowler, "Coarticulatory influences on the perceived height of nasal vowels," *The Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 1146–1158, 1988.

[3] J. J. Ohala and C. Jones, "The phonetics of sound change," in *Historical linguistics: Problems and perspectives*, C. Jones, Ed., 1993, pp. 237–278, longman, London.

[4] G. Zellou and M. Tamminga, "Nasal coarticulation changes over time in Philadelphia English," *Journal of Phonetics*, vol. 47, pp. 18–35, 2014.

[5] M. Ohala and J. J. Ohala, "Nasal epenthesis in Hindi," *Phonetica*, vol. 48, no. 2-4, pp. 207–220, 1991.

[6] P. S. Beddor and R. A. Krakow, "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2868–2887, 1999.

[7] P. S. Beddor, K. B. McGowan, J. E. Boland, A. W. Coetzee, and A. Brasher, "The time course of perception of coarticulation," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2350–2366, 2013.

[8] C. A. Fowler and J. M. Brown, "Perceptual parsing of acoustic consequences of velum lowering from information for vowels," *Perception & Psychophysics*, vol. 62, no. 1, pp. 21–32, 2000.

[9] G. Zellou, "Individual differences in the production of nasal coarticulation and perceptual compensation," *Journal of Phonetics*, vol. 61, pp. 13–29, 2017.

[10] P. S. Beddor, "Nasals and nasalization: The relation between segmental and coarticulatory timing," in *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany, 2007, pp. 249–254.

[11] ——, "The relation between language users' perception and production repertoires." in *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, U.K., 2015.

[12] M. Grosvald and D. Corina, "The production and perception of sub-phonemic vowel contrasts and the role of the listener in sound change," in *The Initiation of Sound Change: Perception, Production, and Social Factors*, M.-J. Solé and D. Recasens, Eds. Benjamins Amsterdam, the Netherlands, 2012, vol. 323, pp. 77–100.

[13] C. Carignan, P. Hoole, E. Kunay, A. Joseph, D. Voit, J. Frahm, and J. Harrington, "The phonetic basis of phonological vowel nasality: Evidence from real-time MRI velum movement in German," in *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia, 2019, pp. 413–417.

[14] C. Carignan, S. Coretta, J. Frahm, J. Harrington, P. Hoole, A. Joseph, E. Kunay, and D. Voit, "Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German," *Language*, vol. 97, no. 2, pp. 333–364, 2021.

[15] E. Kunay, "Vowel nasalization in German: a real-time MRI study," Ph.D. dissertation, Ludwig Maximilians University of Munich, Faculty of Speech and Literature Sciences, 2021.

[16] E. Gerrits and M. E. Schouten, "Categorical perception depends on the discrimination task," *Perception & Psychophysics*, vol. 66, no. 3, pp. 363–376, 2004.

[17] C. Kaernbach, "Simple adaptive testing with the weighted up-down method," *Perception & psychophysics*, vol. 49, no. 3, pp. 227–229, 1991.

[18] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013, download: https://www.vocaltractlab.de/index.php?page=vocaltractlab-download.

[1] The production data indicated a clear difference with respect to the vowel nasalisation ratios relative to OVL for the two coda contexts (higher ratios in the CVNCV context). However, irrespective of the coda environment broad inspection of the relation between the production and perception data did not reveal any systematic patterns: for each perception condition, each coda condition and each vowel category (/a:, e:, i:, o:, u:, ø:, y:/), no context gave reason to consider a more detailed analysis in addition to the overall data presented in fig. 6. Although we were able to test only 20 of the former participants in the MRI production study, there was no indication that results may considerably change with a higher number of participants.

[2] The Spearman method was applied for the constant condition because data deviated from normal distribution.

NEUTRALISATION AND THE PERCEPTION OF CLOSE-MID AND OPEN-MID VOWELS: THE GRADIENT BETWEEN PHONOLOGICAL CATEGORIES

Megumi Im[1]; João Veloso[1]; Maria Conceição Costa[2]; Luis M. T. Jesus[3]

[1]University of Porto, Porto, Portugal; [2]Department of Mathematics (DMat) and Centre of Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal [3]School of Health Sciences (ESSUA), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Aveiro, Portugal
megumiim1004@gmail.com; jveloso@letras.up.pt; lopescosta@ua.pt; lmtj@ua.pt

ABSTRACT

A prototype theoretical framework was the basis of a study on the impact of neutralisation on the perception of Portuguese close-mid and open-mid vowels and the gradient between phonological categories.

Fifteen Portuguese listeners from Lisbon participated in identification and goodness rating tasks of 20 random repetitions of 31 stimuli from a /i/-/e/-/ɛ/-/a/ continuum. Boundaries between categories were explored using logistic regression curves and analysis of variance.

Results from the identification task revealed four vowel categories marked by three statistically distinct boundaries. The prototype of a category, i.e., the stimulus with the highest goodness rating score, was approximately in the centroid and the goodness score decreased as the stimulus moved away from this centroid. The boundary between /e/ and /ɛ/, the underlying opposition that neutralises at surface level, was less steep than between /ɛ/ and /a/, the opposition that never neutralises, a possible influence of the phonological system on speech perception.

Keywords: Speech perception, Phonological system, Mid front vowels, The Prototype Theory.

1. INTRODUCTION

The way we decode speech signals into phonemic categories is still unclear, but some studies [4], [11], [15] show that prototypes play an important role in structuring the categories of speech sounds.

Prototypical phonemes stored in long-term memory may serve as cognitive reference points against which real items are judged, that is, listeners compare the *input* acoustic signals with the phonemes idealised in their linguistic knowledge [15]. When an *input* is similar to the prototype (the best representative of a given phonemic category), it is recognised by the listener as a member of that category [4], [15]. This model implies a hypothetical mechanism in which phonemic categories are represented in terms of prototypes as sole representatives of all members of the same category, and that it is the distance to these prototypes, that defines phonemic categories [11]. Thus, the members of a phonemic category do not have the same representative relevance.

Phonemes have a distinctive function in a given language and are abstract entities, stored as part of the speakers' knowledge of the language, and activated in the production and perception of speech sounds [23]. Listeners are less accurate in the perception of partial contrasts which neutralise in certain contexts [10] and are less sensitive to the difference between allophones than to the contrasting sounds of their own language [2].

The difference in the perception of the neutralizable opposition of the mid vowels and the constant opposition of other vowels has been previously observed in Brazilian Portuguese (BP) studies [21], [22]. As in European Portuguese (EP), the phonological system of BP has oppositions of mid vowels in the tonic syllable. Regarding the process of neutralising the mid vowels, BP presents different realisations in relation to those of EP within the unstressed, non-final context: [e] for front vowels and [o] for back vowels [1], [25].

Silva and Neves [21], [22] carried out two experiments to study the potential differences in the way in which the contrasts between mid-back vowels [o] and [ɔ] [21] and mid front vowels [e] and [ɛ] [22] are represented in the perceptual system of BP speakers. Results showed that the boundary between /e/ and /ɛ/ was less well defined and required a more elaborate processing, thus justifying the complex phonological descriptions [8], [9], [19] of these vowels.

This study was based on two theoretical assumptions: The perception of speech sounds is influenced by the relationships they have within their language system [2], [10], [23]; the categories of speech sounds, especially vowels, have an internal structure [13], [14]. Therefore, this study's main goal was to explore whether the neutralisation of vocalic

phonemes had an influence on the perception of EP, more specifically, ascertaining the gradient of the boundary between the categories of close-mid and open-mid front vowels as defined by Portuguese listeners, comparing it with the gradient of boundaries between other front vowel categories.

2. HYPOTHESIS

Considering the complexity observed in the Portuguese vowel system [3], [17], [18], [24], [26], [25], the proposals of the studies mentioned above, and the results obtained for BP [21], [22], we formulated the following hypotheses: H1. There are four phonemic categories with a gradient structure (although there may be dispersion intra and inter-listeners, as previously observed in speech production studies); H2. The gradient of the boundary between the mid vowels /e/ and /ɛ/ is lower than other boundaries.

To test the first hypothesis (H1) the following questions were posed: Q1. Can four phonemic categories be observed? Q2. Do these categories have a gradient internal structure?

To test the second hypothesis (H2), we analysed four front vowel categories and their boundaries, the presence, in these categories, of an internal structure with prototypes, and the gradient of the boundaries. Postulating an internal structure of the phonemic category and, consequently, distinct boundaries between categories, the gradient of these boundaries was studied using goodness rating tasks.

3. METHOD

The stimuli used in the perception experiments were synthesised with a morphing technique [12] that is based on natural speech samples. Having observed [5] a convergence in the production of mid front vowels by young speakers from Lisbon and considering the role perception units play in speech production [6], [7], [20], stimuli were based on a male speaker from Lisbon, maintaining the characteristics of natural speech during the synthesis process.

3.1. Selection of the speaker

We selected a speaker with the same characteristics as those recruited for the study of Escudero et al. [5], that is, young males with permanent residence in Lisbon, in order to generate stimuli from natural speech. Five speakers (aged between 21 and 35) were recruited who met the criteria and participated in the recording, having previously signed an informed consent.

The speech samples were recorded individually in Lisbon and Porto, and participants were asked to speak spontaneously for 2 minutes on a common topic. The recordings were classified by 8 experts, on whether the recording had the typical characteristics of the vowels of the Lisbon dialect and whether the quality of the voice on the recording was pleasant. We selected the speaker with the highest average score and the lowest standard deviation for the two criteria.

3.2. Recording of the target-vowels

New recordings were made in Lisbon with the selected speaker (21 years of age; university student of the 3rd year of an Undergraduate Degree). We used an *AKG Perception 120 USB* condenser microphone connected to a laptop computer. The recordings were made with *Praat version 6.0.37*, at a sampling frequency of 48000 Hz, with 16 bits per sample, and the data stored in mono .wav format (Windows PCM) with no compression. The speaker produced the target vowels [i, e, ɛ, a] in isolation three times. We chose to record samples with a drop in tone at the end, a procedure similar to that used in the study by Silva and Neves [22], thus seeking to generate the stimuli in the most natural fashion possible. We used a duration of 400 ms, similar to that of the study by Masapollo et al. [16].

3.3. Creation of stimuli

We selected one repetition of each of the four vowels, with the following criteria: A drop in the f_o of 20-25 Hz from beginning to end; values of the frequency of F_1 and F_2, as close as possible to those of Escudero et al. [5].

These four samples were used as anchors to generate 31 stimuli with the *morphing* procedure in the *TANDEM-STRAIGHT monolithic Package 014* (function *Morphing Menu Last Modified by GUIDE v2.5 19-Jul-2016 01:42:59*), while dividing each trajectory into 11 stages: The stimulus stimulus001, stimulus011, stimulus021 and stimulus031 correspond to the anchors of the vowels /i/, /e/, /ɛ/ and /a/, respectively. As the value of F_1 increased, the values of F_2 and F_3 decreased, thus reflecting the characteristics of the stimulus filter.

3.4. Perception pre-test

The pre-test was used to verify the experimental procedures and if there was any issue with the stimuli generated. We recruited a homogeneous group, all of them from the North of Portugal, namely, Braga and Porto. All pre-test experiments were carried out in a quiet room at the University of Minho.

Six (females = 3) EP speakers, from the North of Portugal, university students, aged between 20 and 24

(average = 22.3 years of age), participated voluntarily in the pre-test.

3.5. Perception test

Fifteen (females = 9) EP speakers from Lisbon, with academic qualifications above an undergraduate degree or attending an undergraduate degree, aged between 19 and 34 (average = 23.7 years of age), voluntarily participated in the main test.

The participants listened to each stimulus 20 times, randomly ordered (620 stimuli in total), with an optional pause after every 155 stimuli. There was no possibility of repetition. The experiment included two tasks: 1) to choose one of the words presented on the screen containing the tonic vowel they heard, with an option X in case of no match (identification task); 2) to classify the quality of the sound heard on a scale of 1 to 5 according to the vowel chosen in the task, 1 being a bad example of this vowel and 5 a good example (goodness rating task).

The main test was carried out in the city of Lisbon, and the stimuli presented via *Praat version 6.0.37*, installed in a *TOSHIBA dynabook PT65DGP-RJA* laptop computer, using *Sennheiser* HD 380 Pro headphones connected to its internal sound card. During the presentation of the stimuli, the two-syllable words with the target-vowel present in the tonic syllable, <pico>, <medo>, <teto>, <pato> and <X> were displayed on the computer screen. In addition to the words mentioned above, a scale of 1 to 5 was also displayed at the bottom of the screen.

The answers were given using the computer mouse, by selecting a word and a number for each sound heard. Data from the identification task (task 1) and from the goodness rating task (task 2) were collected with *Praat version 6.0.37*.

3.6. Data processing

The data were statistically analysed with *R studio software Version 1.1.453* and *Excel 2016*, which were also used in the creation of graphs. Since it was intended to statistically analyse the boundaries between categories in the identification task (task 1), and the internal structures of each category and the gradient of each boundary in the goodness rating task (task 2), the number of stimuli were considered as independent variables, each being a distinct identity defined by the values of the acoustic parameters, and the responses of the participants in each task as dependent variables.

4. RESULTS

This section presents the analysis and discussion of the results.

4.1. Identification task

The first research hypothesis (H1) has to do with the existence of speech perception boundaries between the phonemic categories /i/-/e/, /e/-/ɛ/ and /ɛ/-/a/. In order to determine the boundaries between two categories, data from the identification task was initially explored using logistic regression curves that allowed us to estimate the boundary between two categories, which corresponds to 0.5 probability of response. Three boundaries, shown in Fig. 1, between categories were estimated for each participant.

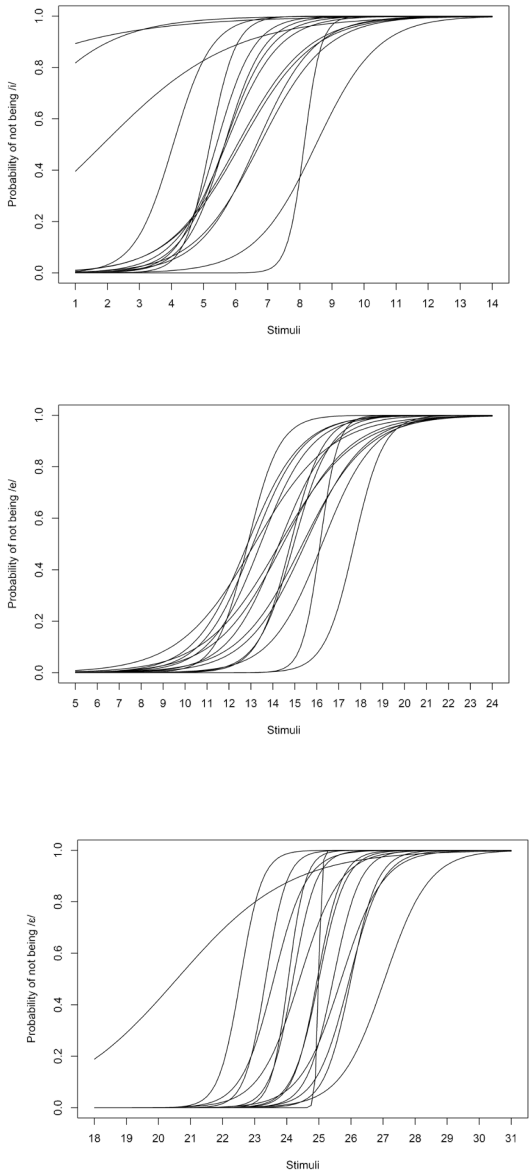


Figure 1: The logistic regression curves of all participants for the /i/-/e/ (top), /e/-/ɛ/ (middle) and /ɛ/-/a/ (bottom) transitions.

Estimates (median / mean ± standard deviation) for the three boundaries were obtained from the logistic regression curves adjusted for all participants:

/i/-/e/ (boundary 1) – 5.65 / 5.28 ± 2.40; /e/-/ɛ/ (boundary 2) – 14.46 / 14.61 ± 1.42; /ɛ/-/a/ (boundary 3) – 24.95 / 24.52 ± 1.59.

Analysis of Variance (ANOVA) was carried out to assess whether there was a significant distance between the samples of the three boundaries for each participant. The null hypothesis of equality of medians between groups was rejected (F=403.7, df=2, *p*-value $<2 \times 10^{-16}$) and multiple comparisons were performed using the Tukey's HSD test. None of the 95% confidence intervals (CI) for the mean of the differences between groups contained zero: Comparison between boundaries 1 and 2 – CI (-10.97; -7.68); boundaries 1 and 3 – CI (-20.89; -17.60); boundaries 2 and 3 – CI (-11.56; -8.27).

4.2. Goodness rating task

The sum of all points (1 to 5) from the goodness rating task of each stimulus were calculated for each participant and represented in a bubble chart. The median value was used to represent the group of all participants (see Fig. 2).

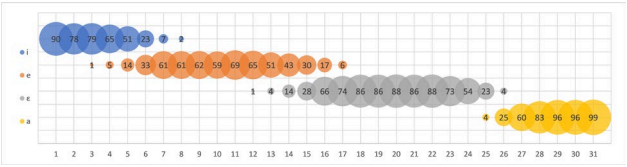


Figure 2: Median goodness rating of all participants, for each of the 31 stimuli (horizontal axis). Scores are represented by the sizes of the circles. The number inside each circle is the median score for each stimulus, considering all participants, which may range from 1 (worst) to 100 (best).

Fig. 2 shows that as the stimulus moves away from the centroid, the goodness score decreases. Most listeners used four categories with three boundaries, exhibiting an internal structure that included their prototype.

Considering the three boundaries revealed by the identification results and the goodness score, a central question of our study was finally addressed, that is, if the boundary between /e/ and /ɛ/ was less steep than other boundaries in the perception of EP vowels.

Median/mean comparison tests for paired samples (Wilcoxon tests) revealed that there were no significant differences between the median values (*p-value* = 0.5760) at boundary 1, but there were significant differences for both boundary 2 (*p-value* = 0.0499) and boundary 3 (*p-value* = 0.0170).

The gradient of the boundaries was also analysed empirically, by calculating the ratio between the median goodness scores at the boundaries between categories, while considering boundary 1 = stimulus 6, boundary 2 = stimulus 15 and boundary 3 = stimulus 25. Our assumption was: The closer to 1 the ratio between goodness scores on a boundary is, the more confusing the categorisation decision will be. When comparing the three boundaries, it was observed that boundary 2 was the one with the ratio closest to 1 (28 / 30 = 0.93), which suggests a greater degree of confusion. Second, there was boundary 1, with a ratio of 23 / 33 = 0.70, and finally, boundary 3 with a ratio of 4 / 23 = 0.17, which suggests a lower degree of confusion than in the previous two.

The gradient of boundary 3 could be considered the highest, because it corresponded to the lowest value of the ratio, and it was also the only boundary for which significant differences were found between the medians of the goodness scores. According to the statistical analysis, there was a considerable degree of confusion in the first two boundaries, since there were no significant differences between the mean/median values of their goodness scores. The empirical analysis of the ratio revealed boundary 2 as the least steep of the three.

5. CONCLUSIONS

The identification task revealed that the participants used four distinct vowel categories, which are in accordance with the phonological theory and with the what has been observed in Portuguese speech production studies. From the answers obtained in the goodness rating task, we were able to observe the internal structure of each category, which has its prototype in the centroid.

The boundaries between /i/ and /e/, and /e/ and /ɛ/ were less steep than the boundary between /ɛ/ and /a/, the latter being non-neutralizable. The gradient of the /e/-/ɛ/ boundary could be due to the complex phonological relations that these sounds establish in the vowel system of EP. There is also a possibility of dialectal specificity but this question remains open for future work.

Stimuli beyond /i/ and /a/ would have covered the entire categories of these two vowels. Variance comparison tests could be carried out, which would allow us to analyse the amplitude and variability of the categories. The same study could be extended to the BP variant and to other EP dialects, providing new evidence for the cognitive and/or phonological theoretical model.

6. ACKNOWLEDGEMENTS

This work was supported by National Funds through the FCT - Foundation for Science and Technology, in the context of the projects UIDB/00022/2020 (Centro de Linguística da Universidade do Porto – CLUP),

UIDB/00127/2020 (Institute of Electronics and Informatics Engineering of Aveiro – IEETA), UIDB/04106/2020 and UIDP/04106/2020 (Center for R&D in Mathematics and Applications – CIDMA).

7. REFERENCES

[1] L. Bisol, *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre: Edipucls, 2001.

[2] A. Boomershine, K. C. Hall, E. Hume, and K. Johnson, "The impact of allophony versus contrast on speech perception," in *Phonological contrast: Perception and acquisition*, P. Avery, E. Dresher, and K. Rice, Eds. New York: Mouton de Gruyter, 2008, pp. 146–172.

[3] J. B. Carvalho, "Contrastive hierarchies, privative features, and portuguese vowels," *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, vol. 6, no. 1, pp. 51–66, 2011.

[4] O. Eerola, J. Savela, J.-P. Laaksonen, and O. Aaltonen, "The effect of duration on vowel categorization and perceptual prototypes in a quantity language," *J Phon*, vol. 40, no. 2, pp. 315–328, Mar. 2012.

[5] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *J Acoust Soc Am*, vol. 126, no. 3, pp. 1379–1393, Sep. 2009.

[6] B. G. Evans and P. Iverson, "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences," *J Acoust Soc Am*, vol. 115, no. 1, pp. 352–361, Jan. 2004.

[7] B. G. Evans and P. Iverson, "Plasticity in vowel perception and production: A study of accent change in young adults," *J Acoust Soc Am*, vol. 121, no. 6, pp. 3814–3826, 2007.

[8] J. Harris, *English Sound Structure*. Cambridge: Blackwell, 1994.

[9] J. Harris and G. Lindsey, "Vowel patterns in mind and sound," in *Phonological knowledge: Conceptual and empirical issues*, N. B. Roberts, P. Carr, and G. Docherty, Eds. Oxford University Press, 2000.

[10] E. Hume and K. Johnson, "The impact of partial phonological contrast on speech perception," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003, pp. 2385–2388.

[11] P. Iverson and P. K. Kuhl, "Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism?," *Percept Psychophys*, vol. 62, no. 4, pp. 874–886, Jan. 2000.

[12] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, "Development of exploratory research tools based on TANDEM-STRAIGHT," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2009, pp. 111–120.

[13] P. K. Kuhl, "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," *Percept Psychophys*, vol. 50, no. 2, pp. 93–107, Mar. 1991.

[14] P. K. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science (1979)*, vol. 255, no. 5044, pp. 606–608, Jan. 1992.

[15] S. E. Lively and D. B. Pisoni, "On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception.," *J Exp Psychol Hum Percept Perform*, vol. 23, no. 6, pp. 1665–79, Dec. 1997.

[16] M. Masapollo, L. Polka, M. Molnar, and L. Ménard, "Directional asymmetries reveal a universal bias in adult vowel perception," *J Acoust Soc Am*, vol. 141, no. 4, pp. 2857–2869, Apr. 2017.

[17] M. H. M. Mateus, "Aspectos Fonológicos e Prosódicos da Gramática do Português," in *Gramática da Língua Portuguesa*, A. M. Brito, I. Duarte, and I. H. Faria, Eds. 2003, pp. 987–1053.

[18] M. H. M. Mateus and E. Andrade, *The Phonology of Portuguese*. Oxford: Oxford University Press, 2000.

[19] A. Nevins, "Vowel lenition and fortition in Brazilian Portuguese," *Let Hoje*, vol. 47, pp. 228–233, 2012.

[20] R. S. Newman, "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," *J Acoust Soc Am*, vol. 113, no. 5, pp. 2850–2860, May 2003.

[21] D. M. R. Silva and R. R. Neves, "Um estudo experimental sobre a percepção do contraste entre as vogais médias posteriores do português brasileiro," *DELTA*, vol. 25, no. 2, pp. 319–345, 2009.

[22] D. M. R. Silva and R. R. Neves, "Perception of height and categorization of Brazilian Portuguese front vowels," *DELTA*, vol. 32, no. 2, pp. 355–373, Aug. 2016.

[23] N. S. Troubetzkoy, *Grundzuge der Phonologie*. 1939.

[24] J. Veloso, "O sistema vocálico e a redução e neutralização das vogais átonas em português europeu contemporâneo," in *Manual de Linguística Portuguesa*, A. Martins and E. Carrilho, Eds. Berlin: De Gruyter, 2016, pp. 636–662.

[25] W. L. M. Wetzels, "Mid vowel neutralization in Brazilian Portuguese," *Cadernos de Estudos Lingüísticos*, vol. 23, pp. 19–55, 1992.

[26] W. L. M. Wetzels, "The representation of vowel height and vowel height neutralization in Brazilian Portuguese (Southern Dialects)," in *Tones and Features: Phonetic and Phonological Perspectives*, J. A. Goldsmith, E. Hume, and W. L. M. Wetzels, Eds. 2011, pp. 331–360.

CROSS-DOMAIN PITCH RECOGNITION ABILITIES OF MANDARIN-SPEAKING PRELINGUALLY DEAF CHILDREN WITH COCHLEAR IMPLANTS*

Qianwen Jin, Jue Yu, Han Sun, Kexin Zhang, Xinran Du

School of Foreign Languages, Tongji University, Shanghai, China
wonderlandjin@163.com, erinyu@126.com, sunhansh1030@163.com, 17602170322@163.com, dxr3232869@163.com

ABSTRACT

The study focused on how Mandarin-speaking children with cochlear implants (CIs) perceived domain-specific pitch information (linguistic pitch vs. musical pitch), under different pitch contexts (context with both segment and pitch information vs. context with mere pitch information). The results showed that when provided with linguistic information involving both segment and pitch, the cochlear-implanted (CI) children were likely to achieve desirable accuracy in identifying different Mandarin lexical tones, though consuming more reaction time (RT) compared with normal-hearing (NH) peers. However, if there was only pitch information available, CI children underperformed their NH counterparts. Furthermore, though they had domain-general barriers in perceiving the subtle pitch variations, comparatively CI children showed a more profound deficiency in pitch recognition under music-domain than linguistic-domain. Finally, RT measurement was found to be negatively correlated with the chronological age, while positively correlated with the age at implantation.

Keywords: Cochlear-implanted children, pitch perception, language and music.

1. INTRODUCTION

Normal-hearing people can recognize more than 90% of the tone through extracting the fundamental frequency from the complex sound signals [7]. But most hearing-impaired children failed even with the help of implanted cochlea [15]. Although cochlear implantation can help hearing-impaired children restore part of their auditory sensation, linguistic disorders still frequently emerge when it comes to daily communication [3, 13, 14]. It was reported that CI children have a lower capacity for recognizing tones and intonations [8, 11, 16], as well as decoding pragmatic and affective meanings [4, 6, 12, 17], in which pitch is a vital element. Therefore, in order to improve the intelligibility and naturalness of language production performance by CI children, it is urgent to well understand their perception performance in various pitch events.

Previous studies have mostly focused on the description of the inefficiency of CI children in pitch perception but barely provide a deep analysis into the internal and external factors that induce the pitch perceptive weaknesses. Besides, though correlations between language and music laid down a theoretical basis for studying pitch perception from a cross-domain perspective [1, 9, 10], there is still a handful of studies investigating the cross-domain pitch perception by Mandarin-speaking children. Therefore, the main purpose of the current study was to obtain a comprehensive understanding about CI children's pitch perception defects from a cross-domain perspective. Only by figuring out what induces the weaknesses of their pitch perception in essence, can clinicians and therapists design and put forward effective intervention in point. The study focused on cross-domain pitch recognition performance (linguistic pitch vs. musical pitch) of the CI children under different phonetic contexts (context with both segment and pitch information vs. context with mere pitch information). By comparing their linguistic and music pitch perception performances, it can be understood whether the CI children's deficit in pitch perception is domain-specific or domain-general; by comparing pure pitch and pitch with intact information, it can be examined how the Mandarin-speaking CI children perceive the Mandarin tones, in terms of bottom-up processes (cortex acoustic analysis) or top-town (stored knowledge gained from linguistic experience). In this way, the following three research questions were expected to be addressed:

1) Are there any differences between the CI children's and the NH children's ability in cross-domain pitch perception in terms of, pitch context and pitch movement?

2) Are there any differences between the CI and the NH group in pitch perception confusion pattern?

3) How the cochlear usage and musical experience impact the CI Children's pitch perception performance?

*This paper is supported by CDPF Hearing and Speech Program (No. 2022CDPFHS-08).

2 METHOD

2.1 Participants

Totally 11 monolingual Mandarin-speaking CI children (7 boys and 4 girls), aged from 6;9 to 11;10 (M ± SD = 9.54 ± 2.88 years), and 20 NH children (10 boys and 10 girls) aged from 5;2 to 10;3 (M ± SD = 7.23 ± 1.96 years), were recruited to take part in the experiment. They were matched in hearing age, language background and education to avoid possible further deviation in the experiment results.

2.2 Audio Recording

Recordings were collected in a sound-proof booth where the noise was 15.6 dB. The recorder is a native Mandarin-speaking female qualified with a Mandarin Proficiency Test certificate of 2nd Grade, Level A. She was seated in front of a laptop, with a Sennheiser head-worn microphone situated approximately 2cm from the mouth. For music stimuli, the piano note C4 (Middle C) was played and recorded with a fundamental frequency of 261.63 Hz. All audios were saved in the wav. format, 16 bits, 44100 Hz.

2.3 Materials and Design

2.3.1 Linguistic Pitch Event

The subtypes of linguistic pitch events include the Mandarin tone and the Mandarin pure tone. Three Mandarin tones were chosen as the experimental materials for perception, namely, T1 (high level), T2 (middle rising) and T4 (high falling). Tone 3 was not included in that CI children are reported to strategically rely more on "creaky voice" instead of pitch to realize the low-dipping property of Tone 3 [5]. Mandarin pure tones were the sinusoidal carrier waves of the lexical tones, being regarded as partially lexical. Specifically, they are comprised of a sine wave and the fundamental frequency shape extracted from recorded Mandarin tones through the LPC manipulation [2].

2.3.2 Non-Linguistic Pitch Event

The subtypes of non-linguistic pitch events include the musical (piano) note and the musical pure note. We generated the musical stimuli in the form of a pair of notes with different pitch directions in accordance with lexical tones' pitch trend. The F0 of the head note was set to the value of the minimum F0 in the first 250 ms of lexical tone, and that of the tail note was modified to the maximum F0 in the latter 250 ms. As with Mandarin pure tones, musical pure notes were the sinusoidal analogues of the musical notes.

2.4 Procedures

All children were brought to participate in a 3 AFC picture-choice task in a visual-world paradigm. Before the formal experiment, they were invited for a session for five trials under instruction. Stimuli were played on a computer screen in a volume reported as audibly comfortable to the participants. At the end of the audio, they were asked to choose one of the three pictures, which represented respectively T1, T2 and T4, shown on the computer screen. Their response accuracy and reaction time were recorded.

3 RESULTS AND DISCUSSION

3.1 Linguistic Pitch Event Perception

In Mandarin lexical tone perception, the groups showed no significant differences both in perception accuracy and reaction time (ACC: $\beta = 0.01$, $p = 0.89$; RT: $\beta = -2.054.98$, $p = 0.19$). It seemed that the CI Children could make the comparable performance as those typically developing children when identifying Mandarin lexical tones, though they had longer reaction time. With regard to the Mandarin pure tone perception, the "group" factor unfolded a significant main effect both on the perception accuracy and the reaction time (ACC: $\beta = 0.27$, $p = 0.01$; RT: $\beta = -2382.33$, $p < 0.05$). CI children were more likely to have obstacles than the NH children when the segmental information was missed. In other words, if the only available attribute was the speech pitch contour, hardly can the CI children match them with the familiar linguistic pitch contour in their mother tongue. In addition, the interaction effect between group and tone type was found significant ($\beta = 0.208$, $p < 0.05$). For the NH children, the identification for a flat tone was far easier than that for a falling tone. Conversely, the CI children performed undesirably in perceiving flat tone but much better in falling tone, though generally worse than the NH children. Since the CI children relied much on larger F0 range to perceive the pitch changes, the detection of flat tone came to be more troublesome.

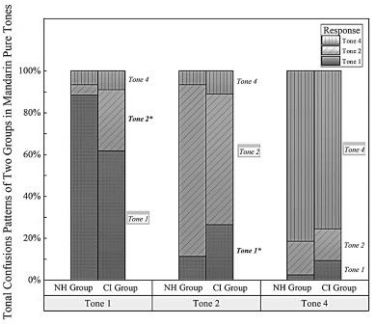


Figure 1: Confusion Patterns for Responses in Mandarin Pure Tones of CI Group and NH Group

Tonal identification errors were also examined to investigate the tonal confusion of the two groups. It showed there was no significant difference between the CI and the NH group in perceiving Mandarin lexical tones ($\chi 2 = 24.557, p = 0.80$); while significant difference in perceiving Mandarin pure tone ($\chi 2 = 54.422, p < 0.001$). Kenward-roger method in linear mix model was further applied to investigate the details in the perception of Mandarin pure tone (see Figure 1).

Generally speaking, all the children can correctly perceive most of the tones, for both groups showed ACC above the chance level 33.3%, which indicated that the cochlear device did facilitate those hearing-impaired children to perceive the linguistic pitch variations. However, when there was merely natural pitch contour modulated in the sine wave without any segmental information, the CI children performed much worse than the NH peers. Among all the Mandarin pure tones, the most confusing error pair was T1-T2. The CI children frequently recognized flat pure tone as rising pure tone or vice versa by mistake.

3.2 Non-Linguistic Pitch Event Perception

Musical notes consisted of two sequential equal-long piano notes which retained the identical pitch direction and pitch range of their corresponding linguistic pitch. Different from the perception result of Mandarin lexical tones, significant group effect was found in musical notes perception in terms of ACC, though not RT (ACC: $\beta = 0.33, p<0.01$; RT: $\beta = -311.64, p = 0.58$). It showed that the accuracy of CI children was significantly far from satisfactory than that of NH children. The musical correlates maintained the same pitch direction and pitch range of the linguistic tones, but the changes in musical pitch were discrete instead of continuous, which demanded more precise resolution in frequencies. The intrinsic hearing loss and imperfect "artificial ear" of the CI children caused the incapability in resolving frequency precisely, thus constraining them to perceive the musical note pitch correctly.

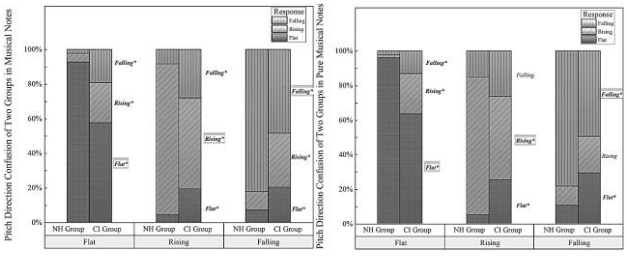


Figure 2: Confusion Patterns for Responses in Linguistic Pitch Event Perception of CI Group and NH Group; Pitch Direction Confusion of Two Groups in Musical Notes (the left figure); Pitch Direction Confusion of Two Groups in Pure Musical Notes (the right figure)

Musical pure notes excluded the musical timbre and thus took a stricter form of pitch that required elementary ability in acoustic processing. The Results were quite similar to that of musical notes, in which group effect was found significant (ACC: $\beta = 0.28$, $p<0.01$; RT: $\beta = -161.88, p = 0.83$). It indicated that different from linguistic timbre, which exerted an enormous impact on the pitch perception of Mandarin tones, for the CI children, the timbre of musical instrument had little effect on their pitch perception. In other words, with or without instrument information, CI children achieved the same accuracy performance.

The Chi-square test showed that the CI children displayed different confusion patterns from their NH counterparts in identifying musical pitch (musical notes: $\chi 2 = 126.59, p < 0.001$; musical pure notes: $\chi 2 = 112.56, p < 0.001$). In musical notes perception, two groups performed significantly differently in the accuracy of the nine target-response pairs. In musical pure notes, the CI children were also vulnerably trapped into the "pitch change trick", in which they mistakenly perceived those pitch with subtle variances as pitch with dynamic movement, either going up or falling down. But with regard to the rising pitch, the CI children would be less likely to take the rising pitch as a falling one or vice versa. The above results confirmed that CI children were more sensitive to large pitch range not only in linguistic domain, but also in music domain.

3.3 Asymmetrical Patterns Between Different Pitch Event Perception

Figure 3 presented the different ACC and RT performance between the CI and the NH group in the cross-domain pitch perception tasks. Pitch event 1 to 4 in the figure represented respectively Mandarin lexical tone, Mandarin pure tone, musical note and musical pure note.

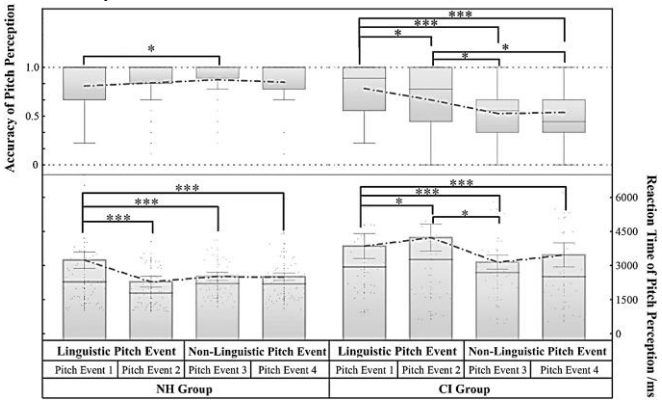


Figure 3: Patterns in Linguistic and Non-Linguistic Pitch Event Perception of Two Groups

Compared to the NH children, the CI children performed diversely in different pitch events. First, the CI children perceived Mandarin lexical tone significantly better than the other three pitch events. Secondly, their perception accuracy of linguistic pure tone was lower than lexical tone but higher than musical pitch perception, which revealed that though the lack of segmental information would somewhat hinder CI children from recognizing the pitch information, they still retained the ability to take advantage of pitch contour to achieve tone recognition. Moreover, they demanded more RT time to organize and coordinate the cognitive resources to process the natural pitch contour in daily life. Finally, it seemed that the perception consistency only resided in non-linguistic pitch events, for the instrumental timbre helped little to perceive pitch information under the music domain. Comparing with linguistic pitch perception, non-linguistic pitch perception of the CI group was extremely poorer. This further reflected the biased development of pitch recognition ability between the music and the linguistic domain.

3.4 Factors Influencing CI Children's Pitch Perception

The former analysis revealed the internal deficiency of CI children in sensing subtle pitch variation in both linguistic and non-linguistic domains and their weaknesses in detecting discrete pitch in music domain. It was attributed not only to their intrinsic bottom-up defect, but also to their insufficient language experience and thus impaired the top-down modulation of speech contour. In this section, other external factors, including musical training, chronological age and cochlear device usage were investigated to fully explain the pitch perception ability of the CI children.

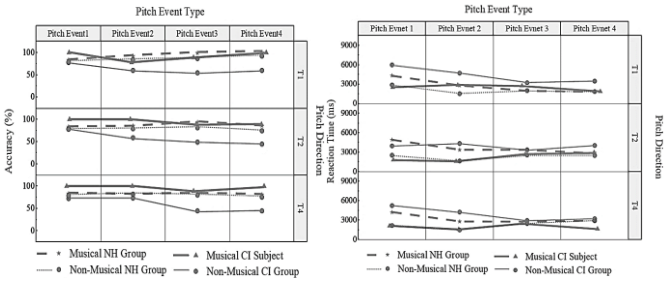


Figure 4: Pitch Perception of Musically Trained and Untrained Children

In our data, half of the NH group (10 children) had received formal musical training. The LMM results showed that the musical training had little influence on normally-developing children in pitch perception but only a difference in the reaction time of Mandarin lexical tone perception (ACC: $\beta = 0.05$, $p = 0.62$; RT: $\beta = 1961.71$, $p<0.05$). As for the CI group, there were 10 non-musically trained subjects. Only one 9-year-old CI boy received 2 years of piano training. Conversely, one-sample T-test indicated that the trained child performed extremely differently in both pitch perception accuracy and reaction time, compared with those non-musical training CI peers (ACC: $t = 16.02$, $p < 0.001$; RT: $t = -12.17$, $p < 0.001$). Furthermore, the musical CI subject showed comparable perception performance to the 20 NH participants, no matter in language or music. The results may not substantiate but indicate the possible positive impact of musical training on the CI children's ability in pitch perception in both music and language.

Spearman coefficients were calculated between the chronological age, age at implantation, length of CI device and pitch perception results. It was found that the chronological age and the age at implantation were closely related to the CI children's reaction time ($\rho = -0.43$, $p < 0.001$; $\rho = 0.29$, $p < 0.001$). The elder the CI children were, the shorter reaction time they demanded in recognizing the pitch. Those who received cochlear implantation earlier could reach faster speed in their way of working out pitch in musical pure notes and pitch with the falling trend.

4. CONCLUSION

In conclusion, CI children could achieve desirable accuracy in identifying different Mandarin tonal pitches, though consuming more reaction time. On the one hand, results in different pitch contexts proved that CI children's top-down regulation was rather weak when perceiving the linguistic pitch contour. They relied more on the segmental information in words. The deficiency in extracting stored pitch knowledge was possibly attributed to the inadequate linguistic experience, causing them failing to internalize the linguistic pitch contour. Therefore, the study further confirmed the significance of linguistic input and experience in the development of pitch awareness and pitch memory. On the other hand, the cross-domain comparison indicated that the CI children may apply domain-specific mechanisms when perceiving the music-domain and linguistic-domain pitch information. The linguistic mechanism took charge of processing continuous coarse pitch contour, while music mechanism dealt with discrete and precise pitch and thus, had higher requirements for the CI children's pitch perception competence. Several important factors were responsible for the CI children's deficiency in pitch perception performance. The case study of the CI boy proved that musical training could potentially strengthen his ability both in language and music. In addition, earlier implantation of cochlear devices may benefit the CI children when processing the pitch information.

5. ACKNOWLEDGEMENT

We would like to express our special gratitude to Yihao Huang, Ziyun Zhang, Jie Liu, Hanlin Zou for helping with data collection and analysis.

6. REFERENCES

[1] Besson, M., Schön, D. 2001. Comparison between language and music. Annals of the New York Academy of Sciences, 930, 232–258.

[2] Boersma, P., Winnik, D. 2019. Praat: doing phonetics by computer (Version 6.0.53), Computer program.

[3] Deroche, M. L., Kulkarni, A. M., Christensen, J. A., Limb, C. J., Chatterjee, M. 2016. Deficits in the Sensitivity to Pitch Sweeps by School-Aged Children Wearing Cochlear Implants. Frontiers in neuroscience, 10, 73.

[4] Hopyan-Misakyan, T. M., Gordon, K. A., Dennis, M., Papsin, B. C. 2009. Recognition of affective speech prosody and facial affect in deaf children with unilateral right cochlear implants. Child neuropsychology: a journal on normal and abnormal development in childhood and adolescence, 15(2), 136–146.

[5] Jue, Y., Meiqi, Z. 2019. Lexical-tone production in prelingually deafened Mandarin-speaking children with cochlear implants. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, 1059–1063.

[6] Kalathottukaren, R. T., Purdy, S. C., Ballard, E. 2017. Prosody Perception and Production in Children with Hearing Loss and Age- and Gender-Matched Controls. Journal of the American Academy of Audiology, 28(4), 283–294.

[7] Liang, Z.A. 1963. Auditory perceptual cues in Mandarin tones. Acta Physiol. Sin., 26, 85-91

[8] Mao, Y., Xu, L. 2017. Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants. International journal of audiology, 56(sup2), 23–30.

[9] Osterhout, L. and Holcomb, P.J. (1992) Event-Related Brain Potentials Elicited by Syntactic Anomaly. Journal of Memory and Language, 31, 785-806

[10] Patel A. D. 2005. The relationship of music to the melody of speech and to syntactic processing disorders in aphasia. Annals of the New York Academy of Sciences, 1060, 59–70.

[11] Peng, S. C., Tomblin, J. B., Cheung, H., Lin, Y. S., Wang, L. S. 2004. Perception and production of mandarin tones in prelingually deaf children with cochlear implants. Ear & Hearing, 25(3), 251-264.

[12] Van, D. J., Schiller, N. O., Levelt, C. C., VAN Heuven, V. J., Beers, M., Briaire, J. J., & Frijns, J. (2019). Prosody perception and production by children with cochlear implants. Journal of child language, 46(1), 111–141.

[13] Volkova, A., Trehub, S. E., Schellenberg, E. G., Papsin, B. C., Gordon, K. A. 2013. Children with bilateral cochlear implants identify emotion in speech and music. Cochlear implants international, 14(2), 80–91.

[14] Vongpaisal, T., Trehub, S. E., Schellenberg, E. G. 2006. Song recognition by children and adolescents with cochlear implants. Journal of speech, language, and hearing research: JSLHR, 49(5), 1091–1103.

[15] Wang, J. 2009. Auditory plasticity induced by cochlear implants in early life. Journal of audiology and speech pathology, 17(1), 7.

[16] Wang, S., Liu, B., Zhang, H., Dong, R., Mannell, R., Newall, P., Chen, X., Qi, B., Zhang, L., Han, D. 2013. Mandarin lexical tone recognition in sensorineural hearing-impaired listeners and cochlear implant users. Acta Oto-Laryngologica, 133(1), 47–54.

[17] Wysocka, M., Mackiewicz, L., Logopedyczne, P. T. 2016. Intonation perception in hearing-impaired children. Logopedia, 45, 53-68.

RAPID SPEECH ADAPTATION AND ITS PERSISTENCE OVER TIME BY NON-STANDARD AND NON-NATIVE LISTENERS

Fengting Du

Newcastle University
f.du4@newcastle.ac.uk

ABSTRACT

Speech produced by L2 talkers is more intelligible to listeners with the same L1 background as the talker than to listeners with a different one. This study extended this line of research to a non-standard native variety and two non-native tonal varieties with similar syllable structures. It examined how these are perceived and adapted to by native and L2 listeners speaking the same L1 or a different L1 from the talker. A post-test after 24 hours detected how the adaptation to accented speech persists. The results found that non-standard and non-native English listeners were better at perceiving and adapting to speech from talkers sharing the same but not a similar or different native background. Such identification accuracy increased after exposure and could continue to improve over a period of sleep. The adaptation effect was not found to extend to new talkers. Additionally, the tested syllabic characteristics were found to have inconsistent effects on speech perception or adaptation.

Keywords: speech perception, speech adaptation, regional- and foreign-accented English

1. INTRODUCTION

Many studies have investigated the influence of both linguistic and non-linguistic traces of a foreign accent on speech perception. The speech identification accuracy or the intelligibility of accented speech to native speakers relates to phonetic and phonological features, such as segmental properties and prosodic patterns [11, 12, 14, 15, 16], and demographic characteristics like race and social status [13, 17]. Recently, more literature has investigated whether an intelligibility difference exists in cross-linguistic communication [2, 3, 5, 7, 8]. These studies found that when an L2 talker talks to an L2 listener from the same native language background, it will give the listener an intelligibility benefit.

Listeners familiar with foreign accents or language teaching experience were found to view L2 speech more favourably. They are likely to exert less effort in accented speech processing and be able to understand more speech that is less intelligible to others without accent familiarity. Additionally, it has been found that exposure to systematic variability of foreign accents facilitates the generalisation of accent features, which can develop into the adaptation to foreign-accented speech [2, 8]. Studies have shown that listeners can adapt to both talker-specific and talker-varied accented speech, as well as accent-independent adaptation. Moreover, studies also found that greater talker variabilities in the exposure can lead to better adaptation.

Although the foreign-accented speech is usually labelled as hard to understand, it has been found that listeners can adapt to accented speech with 2-4 sentences, about one minute of exposure to the target accent [3]. Studies also found that listeners' perceptual ability can be 'improved' if they put in more effort, further suggesting that the intelligibility of accent speech depends not only on the talker but also on the listener [9, 10]. Such inconsistency also shows that speech perception is a complex progress involving a range of interlocuters' language-related background factors, such as native language, L2 proficiency level, language experience, language attitudes and so on.

Many studies compared the intelligibility of native speech with that of L2 speech produced by the talker having the same native language background as the listener. However, such studies cannot answer where the speech intelligibility benefit comes from. Is it from the same articulatory features of segments or the same phonological change transferred into L2 production? To answer these questions, this study examines the scenarios where the interlocutors share a similar L1 background (e.g., having similar syllable structures in terms of the consonant clusters in the onset and coda positions) and where the talker is a native speaker having a regional accent. Additionally, the persistence of rapid speech adaptation is understudied but interesting to look at as it can shed some light on the mechanisms of maintenance with short-term memory.

Therefore, the present study aims to investigate how regional and foreign English varieties are perceived and adapted to by native listeners, L2 listeners with the same L1 background as the talker and L2 listeners with a different L1 background from the talker. It also touches on how the adaptation will change over time and investigates the role of language proficiency in perceiving and adapting to

regional and foreign-accented speech. Burmese and Mandarin accented learners were enrolled, as both their L1s are monosyllabic and only permit consonant clusters in the syllable onset but not the coda [1, 4]. The native variety studied is Geordie, a dialect spoken by people living in Newcastle, England.

2. MAIN EXPERIMENT

2.1. Participants

Six female participants were recorded. They were talkers with the three tested accent backgrounds under study (two from each).

Twenty-one listeners (five Geordie, five Mandarin advanced and five Mandarin intermediate and six Burmese intermediate English learners) aged 19-25 were enrolled and remunerated £10 for participation. The English proficiency of L2 learners was categorised as intermediate if their latest score of *The International English Language Testing System* (IELTS), a standard test to measure L2 learners' English proficiency level, was between 6-7 out of 9 and advanced if it was above 7.

2.2. Materials

2.2.1. Lexical transcription

Twelve lists of 20 words were selected from the reading tests of IELTS. In a lexical list, ten of them had a consonant cluster at the onset position, while the rest ten with a cluster at the coda. Most stimulus words were monosyllabic or disyllabic, with only a few trisyllabic.

2.2.2. Accent exposure

Six picture pairs with 12 differences between each pair were taken from *London UCL Clear Speech in Interaction* (LUCID Corpus) [6] and used as exposure materials. Participants saw a picture and heard someone (with one of the accents under study) describing how the picture differs from another in a spot-the-difference task. In the picture presented, 20 parts were numbered randomly, 12 target parts and eight distractors. To ensure consistency in the linguistic content and exposure time, the researcher wrote the scripts for the recordings. Each recording lasted about 1.5 to 2 minutes long.

2.3. Procedures

This online experiment was built on *LabVanced* and had seven phases in two days. The participant was required to find a quiet place with their computer connected to the Internet and wear earphones if necessary. The first-day experiment had six tests: four main tests involving lexical transcription (pre-test: P0 and three post-tests: P1, P2, P3) and two exposure tests (before P1 and P2, respectively).

Each main task had three 20-word lists recorded in each accent under study. In each word list, half words had a consonant cluster in the onset, and the other half had a cluster in the coda. Stimulus words in P1 and P2 were all recorded by the same talker with each accent as in P0. Listeners were exposed to a new talker of each accent in P3 to see if listeners could generalise the idiosyncrasies of the studied accents to achieve talker-independent perception. One point was awarded if the participant could correctly hear and transcribe a word.

Exposure tasks aimed to help the participant focus on speech decoding. The accuracy from such tasks was only used as criteria to remove the outlier data but not in data modelling.

Another post-test was conducted 24 hours away from the first exposure. The participant transcribed a combined word list of 40 tokens produced by the talker from exposure and the novel one (coded as P4 and P5, respectively). It was scored in the same way as the other post-tests.

2.4. Analysis

The data from one Burmese participant was removed as an outlier. The variables expected to affect the accuracy of perceiving accented speech were the L1 relationship between the interlocutors (same, similar or different), the position of the consonant cluster (onset or coda), test phases (P0, P1, P2, P3, P4 and P5), the talker (same as or different from the exposure recording) and the language proficiency of the listener (intermediate-I, advanced-A). To be specific about the relationship between two L1s, a listener listening to a talker from the very native language background was coded as the same. An L2 listener listening to an L2 talker was coded to have a similar language background. Other scenarios were categorised as having a different background between interlocutors (Geordie listeners listening to Burmese or Mandarin talkers and vice versa).

The accuracy was analysed under separate conditions: pre-test, adaptation to and generalisation of accented speech and the role of language proficiency. Although each condition would not involve all variables, they were all modelled using mixed-effect models. The inclusion of variables was based on theoretical relevance and the significance of the added effects achieved by the likelihood ratio test. The basic model had two random effects: the participants and lexical stimuli. All modelling started from a null model and added one relevant fixed effect each time. The variable stayed in the model only

when it contributed significantly to model fit, the same as the interactions among variables.

3. RESULTS

3.1. Performance across the whole experiment

Fig. 1 illustrates the performance of each listener group across the tests. Generally, Burmese-accented speech achieved the lowest perceptive accuracy. The perceptive accuracy of Geordie and Burmese accents kept increasing from P0 to P2, suggesting a potential adaptative effect. Although the accuracy of Mandarin-accented speech seems to decrease, it may be due to fatigue, as it was the last test in each testing round. The Geordie listeners had the highest accuracy in perceiving three accents and nearly perfect in perceiving their own accented speech. The perceptive performance of the Mandarin-A listeners was lower than that of the Geordie listeners but higher than the Burmese and Mandarin-I listeners. The perception of a different talker's speech was lower than the same.

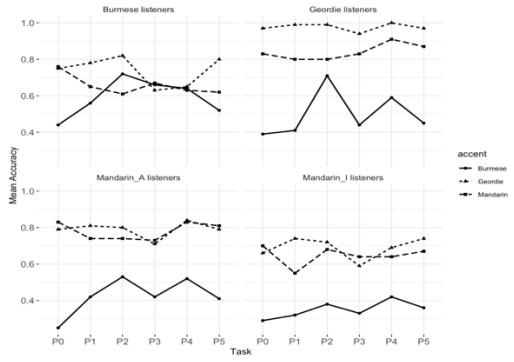


Figure 1: Performance of listeners in each scenario.

3.2. Pre-test

The relevant fixed effects to model pre-test performance were the L1 relationship and the consonant cluster position. The addition of the language background of interlocutors significantly contributed to data likelihood ($\chi^2(5) = 19.5$, $p < .0001$), but the addition of the consonant cluster position did not ($\chi^2(6) = 0.18$, $p > .1$).

Table 1 summarises the model's output. It shows that compared with the listener with a different native language background from the talker, the listener with the same L1 performed better in perceiving speech (running a Plogis function generated an 80.7% chance of transcribing correctly). By comparison, the listener with a different L1 had a lower performance (66.1%), slightly higher than the one with a similar L1 (65.3%).

Table 1: Output from the pre-test model.

Fixed effect	Estimate	SE	z-value
Intercept	0.67	0.31	2.15
Same L1	0.77	0.22	3.47
Similar L1	-0.03	0.26	-0.13

3.3. The adaptation to and generalisation of accented speech

In modelling the adaptation to accented speech, the potential fixed effects were added in the order of the L1 relationship, the test phase and the consonant cluster position. The test phases included were only P0, P1, P2 and P4, as the recordings were from the same talker. The final model included the L1 relationship ($\chi^2(5) = 110.27$, $p < .001$), the test phase ($\chi^2(8) = 13.70$, $p < .01$) and the consonant cluster position ($\chi^2(9) = 5.02$, $p < .05$), without detected interactions. Table 2 summarises the model output.

Table 2: Output from the adaptation model.

Fixed effect	Estimate	SE	z-value
Intercept	0.15	0.34	0.44
Same L1	1.06	0.14	7.61
Similar L1	-0.16	0.16	-0.97
Onset cluster	0.64	0.28	2.28
P1	0.61	0.37	1.66
P2	0.94	0.38	2.46
P4	1.28	0.36	3.51

The outputs demonstrate that compared with the interlocutors having different L1 backgrounds, the intelligibility of accented speech produced by the talker with the same L1 background was higher. However, a similar native language background did not benefit the listener's perception, if not lowered it. It was contrary to the hypothesis that L2 learners may share similar learning strategies in language learning, making L2 learners familiar with other L2-accented speech features. The results also suggest that it was easier for the listeners to transcribe words that had an onset consonant cluster.

The effect of the test phase significantly influenced the listener's perceptive performance in that the speech perception accuracy increased over the experimental progress. A higher probability of correctly perceiving accented speech in the first two post-tests demonstrates the instant benefits of short exposure to phonological processing. In the post-tests starting 24 hours after the start of the experiment, the accentual processing ability of the listeners kept improving. It indicated that the listeners could extract and generalise the accentual characteristics and retain

their perceptual adaptation to improve the efficacy of decoding accented speech in the short future.

To model if the listener could generalise talker-variability from an individual talker to match a new talker's speech characteristics in phonological processing, the L1 relationship, the test phase, the talker and the cluster position were involved. The test phases involved were P2, P3, P4 and P5, as the exposure factor was controlled. The final model included the L1 relationship ($\chi^2(5) = 116.34$, $p < .0001$) and the talker ($\chi^2(6) = 18.97$, $p < .0001$).

Table 3: Output from the adaptation model.

Fixed effect	Estimate	SE	z-value
Intercept	1.10	0.22	5.04
Same L1	0.91	0.13	7.21
Similar L1	-0.13	0.14	-0.90
Different talker	-0.45	0.10	-4.39

The significant talker factor indicates that brief exposure to accented speech from a single talker may not be sufficient for listeners to develop a talker-independent accentual adaptation.

3.4. The role of language proficiency of the listener

In modelling the role of language proficiency, only the data from two groups of Mandarin listeners were analysed. The potential fixed effects were the same as those modelled in the last section analysis except for the inclusion of the proficiency effect. The final model of the adaptation effect contained the fixed effects of the L1 relationship ($\chi^2(5) = 83.94$, $p < .001$) and the proficiency $\chi^2(6) = 5.64$, $p < .05$).

Table 4: Output from the L2 proficiency model.

Fixed effect	Estimate	SE	z-value
Intercept	2.15	0.30	7.18
Same L1	-0.31	0.31	-0.99
Similar L1	-2.60	0.31	-8.39
Proficiency-I	-0.79	0.29	-2.76

As shown in Table 4, the intermediate Mandarin listeners performed worse than their advanced peers, which is expected given the known advantage of higher proficiency in L2 learners. However, the effect of the L1 relationship in this model differs from the previous results. When Mandarin talkers and listeners shared the same L1, it appeared to have a negative effect on their perceptual accuracy, as with other interlocutors who shared similar L1s. The factor of cluster position was found to be insignificant, suggesting that it does not consistently influence accentual perception. The test phase was also insignificant, indicating that taking a one-day break did not result in significant perceptual improvement.

In Table 5, the generalisation effect model is shown. It is similar to the previous results but includes the proficiency level variable. The results indicate that advanced listeners had better perceptual accuracy ($\chi^2(7) = 6.01$ and $p < .05$). Mandarin listeners were unable to transfer accentual adaptation to a novel talker after minimal accentual exposure as well.

Table 5: Output from the adaptation model.

Fixed effect	Estimate	SE	z-value
Intercept	1.44	0.30	4.77
Same L1	-0.11	0.34	-0.33
Similar L1	-2.15	0.34	-6.40
Different talker	-0.55	0.14	-4.00
Proficiency-A	0.73	0.25	2.88

4. DISCUSSION AND CONCLUSION

The results of the current study demonstrate that the listener can adapt to accented speech quickly. It also sheds further light on how the relationship between the native languages of the interlocutors influences phonological processing. One limitation is that it only tested the role of listeners' proficiency and did not strictly control the L2 talkers' proficiency, which may cause the generally low intelligibility of Burmese-accented speech in the study. Although the research included native talkers whose English proficiency levels were naturally higher than L2 learners', the proficiency of L2 learners still should have been controlled to avoid the perceptive variation caused by different L2 intelligibility levels.

The significant talker effect on perceptual accuracy implies that listeners may not be able to transfer accentual adaptation to a novel talker through minimal accentual exposure. It is crucial to note that the talker effect may also arise from the surprise effect of a change in voice. Therefore, further investigation is needed in order to determine what will occur once the surprise effect wears off.

The inconsistency regarding the cluster position factor between analysing Geordie, Burmese, and Mandarin advanced groups and only two Mandarin groups implies that the impact of cluster position on accented speech perception could vary depending on individual factors, such as L1 background and proficiency level. Thus, more work should be undertaken to investigate the perception of and adaptation to L2 accented speech in more contexts and the roles of more aspects of previous linguistic knowledge in driving accentual adaptation, such as phonetic property and prosodic patterns.

5. ACKNOWLEDGEMENTS

I thank my supervisors, Dr. Rory Turnbull, Dr. Niamh Kelly and Prof. Martha Young-Scholten, for their guidance and support. I am also grateful to all informants and participants involved in this research and members of the Phonetics & Phonology Research Group at Newcastle University who provided constructive feedback on an earlier draft.

6. REFERENCES

[1] Watkins, J. W. 2001. Burmese. *Journal of the International Phonetic Association* 31(2), 291–295.

[2] Bent, T., A. R. Bradlow. 2003. The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America* 114(3), 1600-1610.

[3] Clarke, C. M., Garrett, M. F. 2004. Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America* 116(6), 3647-3658.

[4] Duanmu, S. 2007. *The Phonology of Standard Chinese.* Oxford University Press.

[5] Bradlow, A. R., Bent, T. 2008. Perceptual adaptation to non-native speech. *Cognition* 106(2), 707-729.

[6] Baker, R., V. Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3), 761-770.

[7] Grynpas, J., Baker, R., Hazan, V. 2011. Clear speech strategies and speech perception in adverse listening conditions. *International Phonetic Association* XVII, 17-21.

[8] Baese-Berk, M. M., Bradlow, A. R., Wright, B. A. 2013. Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America* 133(3), EL174-EL180.

[9] Pichora-Fuller, M. K., Kramer, S. E., Eckart, M. A., Edwards, B., Hornsby, B., Humes, L. E.,...Wingfield, A. 2016. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing* 37, 5S–27S.

[10] Peng, Z. E., Wang, L. M. 2019. Listening effort by native and non-native listeners due to noise, reverberation, and talker foreign accent during English speech perception. *Journal of Speech, Language, and Hearing Research* 62(4), 1068-1081.

[11] Steffman, J., Jun, S. A. 2019. Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception. *The Journal of the Acoustical Society of America* 146(3), EL251-EL257.

[12] Alves, U. K., Brisolara, L. B. 2020. Listening to accented speech in Brazilian Portuguese: On the role of fricative voicing and vowel duration in the identification of/s/–/z/minimal pairs produced by speakers of L1 Spanish. *Journal of Portuguese Linguistics* 6, 1-23.

[13] Kutlu, E. 2020. Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different English varieties. *Journal of Multilingual and Multicultural Development*, 1-15.

[14] Baese-Berk, M. M., Drake, S., Foster, K., Lee, D. Y., Staggs, C., Wright, J. M. 2021. Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions. *Frontiers in Psychology* 12, 661415.

[15] Severijnen, G. G., Bosker, H. R., Piai, V., McQueen, J. M. 2021. Listeners track talker-specific prosody to deal with talker-variability. *Brain Research* 1769, 147605.

[16] Steffman, J., Katsuda, H. 2021. Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese. *Language and Speech* 64(4), 839-858.

[17] Kutlu, E., Tiv, M., Wulff, S., Titone, D. 2022. The impact of race on speech perception and accentedness judgements in racially diverse and non-diverse groups. *Applied Linguistics* 43(5), 867-890.

INTEGRATION OF MULTIPLE CUES IN NATIVE AND NON-NATIVE SPEECH PERCEPTION

Xiaomu Ren and Clara Cohen
Glasgow University Laboratory of Phonetics, University of Glasgow
x.ren.1@research.gla.ac.uk

ABSTRACT

In this visual-world eye-tracking study, native and non-native listeners' integration of prosody and semantics in speech perception was examined. Participants' eye movements were recorded as they listened to English sentences with variations in prosodic accent, target object information, and semantic match. The results revealed a complex three-way interaction between L1, prosody, and verb semantics. Mandarin listeners did not always prioritize semantics over prosody; instead, they relied more on prosody for processing new information. They also exhibited less flexibility in their integration strategy compared to English listeners, employing a similar approach for both repeated and new information. In contrast, English listeners utilized prosodic cues to enhance semantic processing with old information sentences, but they did not consistently give equal weight to different speech cues. Specifically, they used fewer prosodic cues and prioritized semantics for sentences with new information, where verb semantics were beneficial for new nouns.

Keywords: Pitch accent; Verb semantics; Information status; Eye movements; Speech perception

1. INTRODUCTION

Understanding spoken language demands fluency in two skills: perception and comprehension. Low-level perception processes segmental and suprasegmental cues [1], encoding discrete segments, like consonants and vowels [2], along with prosodic information about prominence (stress and pitch accent) [3, 4]. Higher level comprehension processes extract lexical, semantic, and syntactic information, deriving meaning from the incoming speech stream. Listeners must attend to both sources of information, incorporating them in real time.

Native listeners perform this task effortlessly, drawing on a wide range of information to maximize the efficiency of speech perception and minimize adverse interference [5]. They can draw on low-level prosodic cues to make sense of information structure [6,7], disambiguate structural ambiguities [8,9], and predict upcoming referents [7,10,11]; and they can draw on semantic cues from verb to predict the upcoming direct object on hearing a transitive verb but not an intransitive verb [12]. Furthermore, they can combine these levels of information interactively [13], observing that semantic processing of verbs is strengthened when those verbs are highlighted by a prosodic pitch accent.

Non-native listeners can also use high-level and low-level information in speech perception. At high levels, non-native listeners use contextual semantic information to predict the upcoming lexical form [14, 15], and use the morphosyntactic information like article-noun gender agreement to predict the upcoming target nouns [16, 17]. At lower levels, too, non-native listeners draw on prosodic characteristics to judge the naturalness of the pronunciation [18], and build an understanding of information structure [7]. However, non-native speech perception has been claimed to struggle in integrating information across multiple domains [19], relying on semantic information over syntax or prosody [20, 21].

To examine how native and non-native listeners differ in their attention to high-level and low-level information, we ask whether (i) non-native listeners are indeed less attentive to low-level prosodic information, relative to high-level semantic information, and (ii) whether they are less able to combine high-level and low-level information interactively.

To answer this question, we designed a visual-world eye-tracking experiment in which auditory stimuli were presented with manipulations across three types of speech cues: pitch accent, information status, and verb semantics. Pitch accent and information status test listeners' attention to low-level prosodic information, as the expected pitch accent will differ depending on whether a target word is old or new information. Verb semantics tests listeners' attention to high-level semantic information.

If non-native listeners are less attentive to prosody than semantic information, then we expect that non-native listeners will show reduced effects of pitch accent in their ability to identify a target noun, relative to native listeners. We also expect the effects

of prosody in non-native listeners to be smaller than the effects of semantic information. Finally, if non-native listeners are also less able to combine the three types of speech cues interactively, then we expect that the effect of semantic information will be smaller, regardless of pitch accent, whereas native listeners should show increased semantic processing when target nouns are accented, especially when that accent correctly signals new information status in the target noun.

2. EXPERIMENTAL METHOD

2.1. Participants

Participants were 32 English native speakers ("English listeners") and 40 Mandarin native speakers learning English as L2 ("Mandarin listeners"). The Mandarin listeners used English daily with at least intermediate proficiency (IELTS above 5.5). All participants were university students.

2.2. Materials

Participants were presented with a visual display containing four pictures: a target noun (e.g., *cabbage*), a competitor noun sharing minimally the onset and vowel of the first syllable (e.g., *captain*), and two distractor images (e.g., *badger* and *crocus*). Auditory stimuli took the form of two sentences, with participants instructed to click on the noun mentioned in the second sentence.

All sentences were in English, and were designed to manipulate three binary variables: Information Status (Repeated/New), Verb Appropriateness (Appropriate/Inappropriate), and Pitch Accent (Target/Neutral). In New sentences, the competitor was named in the first sentence and the target named in the second sentence (e.g., *Here is a captain. Susan is going to shred the cabbage after school*). In Repeated sentences, the target was named in both sentences (e.g., *Here is a cabbage. Susan is going to shred the cabbage after school*).

In Appropriate sentences, the verb in the second sentence could only apply to the target as a direct object (e.g., *shred the cabbage*). In Inappropriate sentences, the verb could not plausibly apply to the target (e.g., *drink the cabbage*).

In Target accented sentences, a contrastive pitch accent was placed on the target noun (e.g., *Susan is going to shred the CABBAGE after school*), while in neutral accented sentences, the prosody was a more neutral topic-content melody, with any pitch accent placed on the final adverbial.

These three binary variables were crossed to create 8 conditions for each item. Forty different sentence-pair items were recorded, and rotated in the 8 conditions across 8 experimental lists in a Latin square design. Each list also contained a further 40 filler sentence pairs. All filler sentences were constructed so as to balance each combination of information status, verb appropriateness, and pitch accent across all trials within each experimental list. Stimuli were presented in pseudo-random order, in a different order for each participant. All sentences were recorded by a phonetically-trained native speaker of English in a soundproofed booth.

2.3. Procedure and Analysis

Data were collected with an Eyelink 1000+ eye-tracker running Experiment Builder software. Gaze data was sampled at 1000Hz. Each experiment began with calibration, two practice trials, and then the system was recalibrated before the main experiment began. Participants completed 8 blocks of 10 trials, with a break and recalibration between each block.

Looks to target and competitor were binned and converted to proportions over 50ms windows, over an interest period stretching from 200 ms before the target onset to 1800ms after target onset. Target advantage (TA) was calculated by subtracting the proportion of looks to competitor from proportion of looks to target. This TA metric formed the dependent variable.

TA was analysed with generalised additive mixed models (GAMMs), using the `mgcv` package (version 1.8.4 [22]) in R (version 4.2.1; [23]). Repeated and new information sentences were analysed separately. For both repeated and new sentences, a simple GAMMs was built, containing parametric effects and difference smooths for each of the three binary main variables: Accent (neutral/target); Verb Appropriateness (appropriate/inappropriate); and L1 (Mandarin/English). All factors were treatment-coded, with default levels set as Neutral Accent, Appropriate Verb, and Mandarin L1. Interaction terms were added progressively as two-way and then three-way interactions, individually considered both as parametric terms, difference smooths, or both. Each subsequent model was tested against the simpler model with a log-likelihood ratio test, as implemented in the package `itsadug` (version 2.4.1 [24]).

3. RESULTS

Figure 1 shows the gaze traces for both repeated and new target nouns, while Table 1 summarises the final models. For repeated information conditions, the optimal model contained all two-way interactions in

the parametric terms, and a three-way interaction in the difference smooths. For new information, the model ended up needing three-way interactions in both parametric terms and smooths.

For repeated information sentences, Mandarin listeners had a reduced TA for Inappropriate verbs ($\beta = -0.068, p < .001$) and Appropriate verbs showed very little effect of Accent ($\beta = 0.001, p = .91$). However, the Accent by Appropriateness interaction reveals that Accent did affect Inappropriate verbs, which suffered more under Target Accent than under Neutral Accent ($\beta = -0.029, p < .01$). English listeners showed higher TA than Mandarin listeners with Appropriate verbs under Neutral Accent ($\beta = 0.214, p < .001$), and unlike Mandarin listeners, English listeners benefited from Target Accent, with both Appropriate and Inappropriate verbs ($\beta = 0.047, p < .001$). English listeners also showed a smaller deficit for Inappropriate verbs than Mandarin listeners ($\beta = 0.042, p < .001$), but the lack of three-way interaction means that, like Mandarin listeners, English listeners also showed a stronger Appropriateness effect under Target Accent than Neutral Accent. The significantly different curvature of the gaze traces indicates that this Appropriateness effect was concentrated in the first 25 time windows for English listeners (Fig 1, top right black lines).

In other words, with repeated nouns, Mandarin listeners were more sensitive to verb Appropriateness than accent, and accenting the target increased the effect of Appropriateness. English listeners shared the heightened effect of Appropriateness under Target Accent, but also showed a stronger sensitivity to Accent than Mandarin listeners.

With new Information nouns, Mandarin listeners demonstrate a response pattern that was clearly slower than that of English listeners. English listeners reached peak TA at about Time Window 25, while Mandarin listeners did not peak until Time Window 35. For Mandarin listeners, there was little net effect of Appropriateness for Neutral Accent ($\beta = -0.004, p = .7$), but the significantly different curvatures of the gaze traces reveal that TA for Appropriate verbs rose faster and peaked higher than for Inappropriate verbs (Fig 1, bottom left grey lines). With Target Accent, Inappropriate Verbs had significantly lower net TA than Appropriate verbs ($\beta = -0.061, p < .001$). English listeners showed a much larger effect of Appropriateness for Neutral Accent than Mandarin listeners ($\beta = -0.061, p < .001$), but did not share the benefit of Target Accent on Appropriate verbs with Mandarin speakers ($\beta = -0.08, p < .001$). They also showed a smaller effect of Appropriateness under Target Accent than Neutral Accent ($\beta = 0.071, p < .001$), although again the significantly different curvatures of the traces suggest that this reflects a delayed peak in TA for Inappropriate verbs, allowing net TA to 'catch up' to Appropriate verbs (Fig 1, bottom right black lines).

In other words, English listeners showed a larger effect of Appropriateness than Mandarin speakers, but it was slightly reduced under Target Accent. The effect of Appropriateness emerged more slowly in Mandarin speakers, but by contrast to English speakers it was more pronounced under Target Accent than Neutral Accent.

4. DISCUSSION AND CONCLUSION

The study confirms some, but not all of our predictions. First, when processing repeated information sentences, Mandarin listeners were more sensitive to semantics (Appropriateness) than prosody (Accent), while English listeners attended to both. Further, when English listeners listened to repeated information sentences, adding target accent increases the attention to semantics, which was also predicted. This finding replicates a pattern observed in a previous study [13], where placing a contrastive pitch accent on a target noun deepened semantic processing.

However, contrary to prediction, this deepening of semantic processing disappeared for English listeners with new information sentences, while it was present for Mandarin listeners in both contexts. Further, although we predicted that Mandarin listeners would rely more inflexibly on semantics than English listeners, instead Mandarin listeners showed more sensitivity to prosody than semantics with new information, while it was English listeners for whom semantics had the largest effect.

The results suggest that, counter to our predictions, Mandarin listeners are, in fact, capable of integrating speech cues at different levels, and they do not always prioritize semantics over prosody. Nevertheless, we can confirm that Mandarin listeners integrate speech cues *differently* from English listeners, and *less flexibly*. Mandarin speakers showed a similar integration strategy between semantics and prosody across both repeated and new information. The addition of target accent always increased semantic processing, perhaps because the accent rendered the target noun louder and more salient. The accent also aided perception generally in new information sentences, where that prosodic structure was appropriate to the information status of the sentence.

English listeners, by contrast, allowed prosodic accent to deepen semantic processing only with old

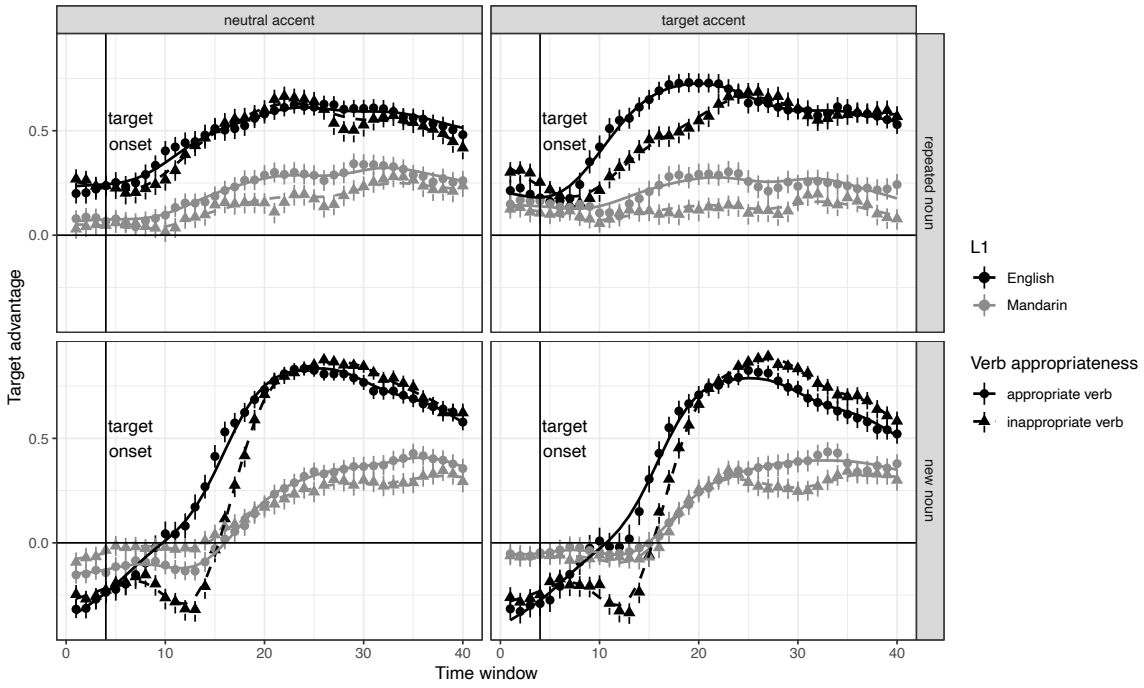


Figure 1: Gaze traces with repeated nouns, showing target advantage (TA) across English (black line) and Mandarin (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. Neutral accent is on the left, while Target Accent is on the right. Target accent is the unexpected prosody for repeated information (top row), and the expected prosody for new target nouns (bottom row).

Table 1: GAMM models for sentences with repeated (left) and new (right) information. Levels for Accent (abbreviated *acc*) are Target (neutral, abbreviated *neu*) and target (abbreviated *tar*). Levels for Appropriateness (*App*) are Appropriate (reference, *a*) and Inappropriate (*i*). Levels for L1 are Mandarin (reference, *ma*) and English (*eng*). Difference smooths are calculated on an 8-level factor representing the interaction between Acc, App, and L1.

Repeated information

Parametric	Est.	SE	*t*	*p*
Intercept	0.240	0.030	7.85	$< .001$
Acc=tar	0.001	0.009	0.11	.91
App=i	-0.068	0.009	-7.71	$< .001$
L1=eng	0.214	0.043	4.95	$< .001$
Acc=tar:App=i	-0.029	0.011	-2.74	$< .01$
Acc=tar:L1=eng	0.047	0.011	4.36	$< .001$
App=i:L1=eng	0.042	0.011	3.88	$< .001$

Difference smooths	edf	*F*	*p*
Window	6.65	11.25	$< .001$
Win:Acc=tar,App=a,L1=eng	6.82	8.03	$< .001$
Win:Acc=neu,App=i,L1=eng	5.56	2.73	$< .05$
Win:Acc=tar,App=i,L1=eng	6.55	5.26	$< .001$
Win:Acc=neu,App=a,L1=man	1.02	0.82	.36
Win:Acc=tar,App=a,L1=man	1.01	6.34	$< .05$
Win:Acc=neu,App=i,L1=man	3.08	3.61	$< .05$
Win:Acc=tar,App=i,L1=man	3.22	6.97	$< .001$
Window, by subj	434.00	8.38	$< .001$

New information

Parametric	Est.	SE	*t*	*p*
Intercept	0.154	0.029	5.31	$< .001$
Acc=tar	0.036	0.010	3.72	$< .001$
App=i	-0.004	0.10	-0.39	.70
L1=eng	0.260	0.043	6.02	$< .001$
Acc=tar:App=i	-0.041	0.014	-2.93	$< .005$
Acc=tar:L1=eng	-0.080	0.014	-5.59	$< .001$
App=i:L1=eng	-0.061	0.014	-4.28	$< .001$
Acc=tar:App=i:L1=eng	0.071	0.021	3.40	$< .001$

Difference smooths	edf	*F*	*p*
Window	8.51	63.01	$< .001$
Win:Acc=tar,App=a,L1=eng	1.00	1.20	.27
Win:Acc=neu,App=i,L1=eng	8.63	32.96	$< .001$
Win:Acc=tar,App=i,L1=eng	8.63	31.26	$< .001$
Win:Acc=neu,App=a,L1=man	6.39	18.92	$< .001$
Win:Acc=tar,App=a,L1=man	6.31	19.56	$< .001$
Win:Acc=neu,App=i,L1=man	6.67	21.94	$< .001$
Win:Acc=tar,App=i,L1=man	6.49	20.17	$< .001$
Window, by subj	445.41	8.12	$< .001$

information sentences, where the accent was inappropriate for the information structure. Perhaps this was because the prosodic pattern, by virtue of its unexpectedness, served to direct their attention to the target noun. In new information sentences, the prosodic pattern was expected, and so almost invisi-

ble to English listeners.

5. REFERENCES

[1] A. Cutler and A. Jesse, "Word stress in speech perception," in *The Handbook of Speech Perception*. John Wiley & Sons, Ltd, 2021, pp. 239–265.

[2] X. Tong, C. McBride, C.-Y. Lee, J. Zhang, L. Shuai, U. Maurer, and K. K. H. Chung, "Segmental and suprasegmental features in speech perception in Cantonese-speaking second graders: An ERP study: Speech perception in Cantonese children: An ERP study," *Psychophysiology*, vol. 51, no. 11, pp. 1158–1168, 2014.

[3] D. Dahan, "Prosody and language comprehension," *WIREs Cognitive Science*, vol. 6, no. 5, pp. 441–452, 2015.

[4] K. A. Wenrich, L. S. Davidson, and R. M. Uchanski, "Segmental and Suprasegmental Perception in Children Using Hearing Aids," *Journal of the American Academy of Audiology*, vol. 28, no. 10, pp. 901–912, 2017.

[5] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press, 2012.

[6] D. Dahan, M. K. Tanenhaus, and C. G. Chambers, "Accent and reference resolution in spoken-language comprehension." *Journal of Memory and Language*, vol. 47, pp. 292–314, 2002, place: Netherlands Publisher: Elsevier Science.

[7] M. Perdomo and E. Kaan, "Prosodic cues in second-language speech processing: A visual world eye-tracking study," *Second Language Research*, vol. 37, no. 2, pp. 349–375, 2021.

[8] A. Weber, M. Grice, and M. W. Crocker, "The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements." *Cognition*, vol. 99, pp. B63–B72, 2006, place: Netherlands Publisher: Elsevier Science.

[9] C. Nakamura, J. A. Harris, and S.-A. Jun, "Integrating prosody in anticipatory language processing: how listeners adapt to unconventional prosodic cues," *Language, Cognition and Neuroscience*, vol. 37, no. 5, pp. 624–647, 2022.

[10] A. Weber, B. Braun, and M. W. Crocker, "Finding Referents in Time: Eye-tracking Evidence for the Role of Contrastive Accents." *Language and Speech*, vol. 49, pp. 367–392, 2006, place: United Kingdom Publisher: Kingston Press.

[11] K. Ito and S. R. Speer, "Anticipatory effects of intonation: Eye movements during instructed visual search," *Journal of Memory and Language*, no. 58(2), pp. 541–573, 2008.

[12] M. Arai and F. Keller, "The use of verb-specific information for prediction in sentence processing," *Language and Cognitive Processes*, vol. 28, no. 4, pp. 525–560, 2013.

[13] L. Wang, M. Bastiaansen, Y. Yang, and P. Hagoort, "The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect," *Neuropsychologia*, vol. 49, no. 5, pp. 813–820, 2011.

[14] A. Foucart, E. Ruiz-Tada, and A. Costa, "Anticipation processes in L2 speech comprehension: Evidence from ERPs and lexical recognition task," *Bilingualism: Language and Cognition*, vol. 19, no. 1, pp. 213–219, 2016.

[15] H. Hopp, "Semantics and morphosyntax in predictive L2 sentence processing," *International Review of Applied Linguistics in Language Teaching*, vol. 53, no. 3, pp. 277–306, 2015.

[16] ——, "Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability." *Second Language Research*, vol. 29, no. 1, pp. 33–56, 2013.

[17] S. De Deyne, D. J. Navarro, and G. Storms, "Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations," *Behavior Research Methods*, vol. 45, no. 2, pp. 480–498, 2013.

[18] C. Tsurutani, K. Tsukada, and S. Ishihara, "Comparison of Native and Non-native Perception of L2 Japanese Speech Varying in Prosodic Characteristics," Queensland, Australia, 2010, pp. 122–125.

[19] A. Sorace, "Pinning down the concept of âinterfaceâ in bilinguals," *Linguistic Approaches to Bilingualism*, vol. 1, pp. 1–33, 2011.

[20] H. Clahsen and C. Felser, "Grammatical processing in language learners," *Applied Psycholinguistics*, vol. 27, no. 1, pp. 3–42, 2006.

[21] T. Gruter, E. Lau, and W. Ling, "How classifiers facilitate predictive processing in L1 and L2 Chinese: the role of semantic and grammatical cues," *Language, Cognition and Neuroscience*, vol. 35, no. 2, pp. 221–234, Feb. 2020, publisher: Routledge.

[22] S. N. Wood, "Thin-plate regression splines," *Journal of the Royal Statistical Society (B)*, vol. 65, no. 1, pp. 95–114, 2003.

[23] R. C. Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2022, r package version 2.4.1.

THE EFFECTS OF ACCENT FAMILIARITY ON NARRATIVE RECALL IN NOISE

Andreas Krug, Ghada Khattab and Laurence White

School of Education, Communication and Language Sciences, Newcastle University, UK
a.krug2@newcastle.ac.uk, ghada.khattab@newcastle.ac.uk, laurence.white@newcastle.ac.uk

ABSTRACT

Listeners frequently hear spoken utterances in different accents. It has been suggested that L2 accents are processed and represented in less lexical and semantic detail by L1 listeners. This has been related to listeners' lower expectations of L2 linguistic proficiency, but the impact of other factors, such as accent familiarity, has not been fully explored. Here we test whether the L2 recall disadvantage extends to unfamiliar L1 accents, to which reduced proficiency expectations should not apply. Tyneside (UK) and New Zealand listeners performed a change detection task, during which they heard story pairs in Tyneside and New Zealand English. Stories were masked with multi-speaker babble noise to increase task difficulty. Accent did not mediate New Zealand listeners' recall, but there was a familiarity benefit for Tyneside listeners, whose recall was more accurate for their own accent. Thus, similar to L2 accent studies, L1 accent recall can be mediated by familiarity.

Keywords: speech perception, recall, familiarity, regional accents, noise masking

1. INTRODUCTION

This paper investigates how narrative recall is affected by accent familiarity when stories are presented in noise. Past research has shown that short-term recall of stories is affected by the L1/L2 status of the narrator [1]. Specifically, word changes between an orally narrated story and a corresponding transcript were detected more reliably if the narrator was an L1 rather than L2 speaker of English, though only if participants were asked to listen for comprehension rather than memory [1].

These findings have been interpreted within the framework of "less-detailed" or "good-enough" representations [1, 2, 3]: thus, the processing and subsequent representation of linguistic stimuli is only as detailed as required for the task at hand. This account predicts that factors such as task instructions and linguistic structure will influence the processing and representation of narrations. For example, stored representations during the task will be more detailed if listeners are instructed to listen for memory rather than comprehension. In terms of linguistic structure, focus encourages more detailed representations. Thus, semantically related changes are detected more frequently when presented in focus versus out of focus [4]. Only in linguistic focus are representations detailed enough to detect the changes. Lev-Ari and Keysar's [1] results suggest that speaker identity also influences processing. When stories were narrated by L1 speakers of English, listener performance in a change detection paradigm was better than when the narrator was an L2 speaker, suggesting more detailed lexical and semantic representations of L1 narrations [1]. Importantly, this L1/L2 accent effect was not found if listeners were instructed to focus on potential word changes rather than comprehension beforehand: they presumably engage more detailed processing under such task conditions, regardless of speaker accent [1].

It remains unknown whether the L2 speech recall disadvantage extends to different L1 regional accents. Lev-Ari and Keysar [1] suggest that the poorer recall of L2 speech is due to listeners' expectation that L2 speakers will be less proficient in their language use than L1 speakers. As a result, listeners process and represent L2 speech in less detail to make the processing mechanism more efficient [5, 6]. Less proficient language use would not be expected from speakers of different regional accents of the same L1. However, listeners also hold varying expectations about regional accents in terms of prestige and pleasantness [7, 8]. Moreover, listeners tend to be better at processing familiar rather than unfamiliar accents – as indexed, for example, by lexical access – especially under adverse listening conditions [9, 10, 11]. This disparity could extend to narrative recall if less successful lexical processing results in poorer lexical and semantic representations of speech. Given these considerations, poorer recall of unfamiliar L1 accents is plausible.

A partial replication of Lev-Ari and Keysar's [1] study found no differences in recall due to variation in Spanish L1 accents [12]. However, design differences may have influenced outcomes: for example, listeners had to recall the changed words in Lev-Ari and Keysar [1] while Frances et al. [12] only analysed change detection accuracy.

The current study specifically investigates the effect of accent familiarity on recall. Adverse listening conditions are employed to increase task difficulty and thus potentially elicit effects of accent familiarity that might not otherwise emerge [9]. The study uses a change detection and recall paradigm similar to that of Lev-Ari and Keysar [1] to address the research question: How does familiarity with different L1 accents affect recall of narratives spoken in these accents under adverse listening conditions?

2. METHOD

2.1. Participants

Two groups of participants were recruited via universities and social media platforms: native Tynesiders (UK) and native New Zealanders. A pre-screening questionnaire was used to confirm that the participants had been born and raised in Tyneside or New Zealand [13]. These locations were chosen as providing participant groups unlikely to be familiar with each other's accents. Participants received course credit or vouchers as a token of appreciation. Data from 42 Tynesiders (mean age: 28.6 years) and 37 New Zealanders (mean age: 24.5 years) were analysed. None of the participants reported any diagnosed speech, language or hearing difficulties. Accent familiarity was tested by means of an accent matching task as well as a demographic and language background questionnaire (Section 2.3), which confirmed that the two groups were relatively unfamiliar with each other's accents.

2.2. Stimuli

The stimulus set for the recall task consisted of story pairs differing only in a single target word [see Sturt et al., 4, for a similar design]. Each story consisted of three sentences. For experimental pairs, the target occurred in the second sentence, although never sentence-initially. As the example below shows, the changed target is semantically related to the original one in the 'related' version of the story (*rubies → diamonds*). In the 'unrelated' version, the two targets are unrelated (*rubies → drugs*). Semantically related targets could be subsumed under one hyperonym (e.g. *gems*). For filler pairs, the target occurred in the first or the third sentence.

Example story: *Sam and Kate made an interesting discovery yesterday. They found a small chest full of rubies/diamonds/drugs in their attic. They had no idea where it came from or who put it there.*

Stories were recorded at Newcastle University (UK) and Canterbury University (New Zealand) in similar studio conditions. For each accent, there were two middle-aged female speakers, judged by phonetically trained assessors to be representative of their accents. The Tyneside English (TE) speakers had lived in or close to Newcastle upon Tyne (UK) for most of their lives. The New Zealand English (NZE) speakers were of European descent, born and raised in the Christchurch region and had lived in New Zealand most of their lives. European NZE accents are relatively homogeneous [14], which is why those two speakers were deemed representative of not just Christchurch accents, but of European NZE in general.

For each accent, there were 12 experimental and 8 filler story pairs. All stimuli were masked with amplitude-stable multi-speaker babble noise [15], using an adapted Praat script originally developed by McCloy [16]. The amplitude of the stories was first normalised to 65dB at the sentence level. The stories were then overlaid with the babble noise at a signal-to-noise ratio of 0dB. The noise signal was faded in and out for 500 ms before and after the speech signal.

2.3. Procedure

The recall task was part of a larger study, conducted online via LabVanced [17]. The study included: 1. a headphone check; 2. a lexical decision task (reported elsewhere); 3. the recall task; 4. the accent matching task; 5. the demographic and language background questionnaire.

The recall task included 40 story pairs (24 experimental), preceded by four practice pairs. For each speaker, there were 10 trials (6 experimental), 5 with changes and 5 without, as shown in Table 1. The story pairs were blocked by accent, with speaker nested under accent. The order of speakers was varied between participants via a Latin-square design.

Experimental Trials			**Filler Trials**	
un-changed 2	related change 2	unrelated change 2	un-changed 3	changed 1
Total = 6			Total = 4	
Grand total = 10				

Table 1: Trials per speaker per participant during the recall task

On each trial, participants first saw centrally located fixation crosses for 500 ms. They then heard the first story, followed by another 500 ms of fixation crosses. Then they heard the second story, which: a. was identical to the first one; b. included a related word change (*rubies → diamonds*); c. included an unrelated word change (*rubies → drugs*). Participants had to respond, via key presses, whether the two

stories were identical or different. If they responded "identical", they moved on to the next trial. If they responded "different", they saw the second story on a new screen, including two text boxes for participants to input the changed word from the second story and the original word from the first story.

2.4. Data Analysis

The data were analysed within the 'tidyverse' package in R [18, 19]. The data analysis was conducted in two steps. First, participants' key press accuracy was analysed, i.e., if they correctly responded that the two stories were identical or different. Second, the accuracy of the correction was analysed. This measure was only considered for trials which did include a change and for which the participants had pressed the correct key initially.

Logistic mixed-effects models were used for the statistical analysis of both datasets, which consisted of 1896 and 955 trials for the key press and correction accuracy, respectively. The following predictors were included in the regression models:

- **Accent**: TE *vs* NZE.
- **Story pair**: unchanged *vs* related change *vs* unrelated change. For correction accuracy, unchanged story pairs were irrelevant and this level of story pair was dropped.
- **Participant origin:** Tyneside *vs* New Zealand.
- All interactions (two-way and three-way) between accent, story pair and participant origin were included.

The models further included two control predictors:

- **Block** distinguished the first and second five trials in a specific speaker's voice to check for adaptation effects.
- **Articulation rate**, in syllables per second (z-transformed).

The random effects structure of the models included random intercepts for participant and item, plus random slopes for accent by participant.

For both key press and correction accuracy, the analysis procedure started with the full model, provided below in 'lme4' syntax [20]:

accuracy ~ accent x story pair x participant origin + block + articulation rate + (1 + accent|participant) + (1|item)

Significant effects were identified by means of log-likelihood model comparisons, using the 'afex' package in R [21]. Predictors that did not reach significance were dropped. This resulted in the following models, for which Bonferroni-corrected pairwise comparisons were conducted with the 'emmeans' package [22]:

accuracy_key_press ~ accent x story pair x participant origin + (1 + accent|participant) + (1|item)

accuracy_correction ~ story pair + (1| participant) + (1|item)

3. RESULTS AND DISCUSSION

3.1. Key Press Accuracy

Overall, 1388 trials (73.2%) were correct and 508 (26.8%) were incorrect. There was a significant effect of story pair, $\chi^2(2) = 107.0$, $p < .001$, along with interactions between accent and participant origin, $\chi^2(2) = 9.5$, $p = .002$, and between accent, story pair and participant origin, $\chi^2(2) = 6.6$, $p = .038$. Regarding story pair, pairwise comparisons showed significant differences between unchanged and related story pairs, $\beta = 5.8$, $SE = 1.2$, $p < .001$, as well as unchanged and unrelated story pairs, $\beta = 5.5$, $SE = 1.1$, $p < .001$, but not between related and unrelated story pairs, $\beta = 1.0$, $SE = 0.1$, $p > .999$.

Regarding the interaction between accent and participant origin, participants' key press accuracy was higher for their own accent. Pairwise comparisons supported this for the difference between TE and NZE speakers in participants from Tyneside, $\beta = 2.1$, $SE = 0.8$, $p = .039$. The accuracy difference between accents was not significant for New Zealand participants, $\beta = 0.7$, $SE = 0.3$, $p = .418$.

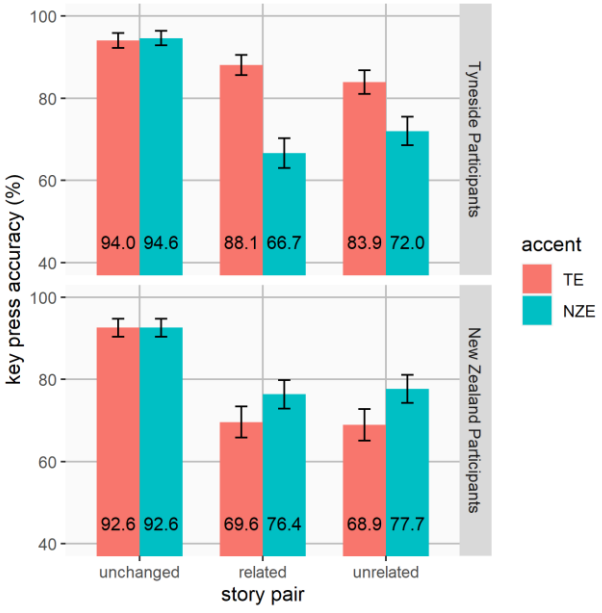


Figure 1: Key press accuracy by participant origin, story pair and accent (error bars: ± one SE)

Figure 1 illustrates the three-way interaction between accent, story pair and participant origin. The

main effect of story pair is obvious in that accuracy was highest for unchanged pairs. Pairwise comparisons show that the two-way interaction between accent and participant origin is further modulated by story pair: thus, whilst accuracy did not vary by speaker accent for NZE participants for any story pairs, Tyneside participants were more accurate for TE speakers than NZE speakers, both in related story pairs, $\beta = 4.4$, $SE = 1.9$, $p < .001$, and unrelated story pairs, $\beta = 2.4$, $SE = 1.0$, $p = .038$. Moreover, the largest difference in key press accuracy was found for related story pairs, where Tynesiders' accuracy was 21.4 percentage points lower for the unfamiliar accent, as compared to a difference of 11.9 percentage points for unrelated story pairs.

Taken together, all participants did best for unchanged story pairs. Where story pairs included a change, Tyneside participants performed better with their own accent than for the NZE accent, especially when the changed words were semantically related. For New Zealand participants, there was a similar, albeit not significant, trend in the data. These findings suggest a familiarity benefit for change detection under adverse listening conditions. The increased task difficulty induced by presenting stories in noise might have been one reason why an effect of L1 accent was found here, but not in previous research [12].

3.2. Correction Accuracy

Out of the 955 change trials with a correct key press response, 797 trials (83.5%) were correct and 158 (16.5%) were incorrect as regards target recall (both original and changed word recalled correctly). Story pair was the only robust predictor of correction accuracy, $\chi^2(2) = 6.54$, $p = .011$. The difference between related and unrelated story pairs is shown in Figure 2. Participants were better at correcting semantically related rather than unrelated changes, regardless of the accent they heard and where they were from. There was an accuracy difference of 6.6 percentage points between the story pairs.

Thus, semantic proximity is helpful for the correction of the changes, with no difference in target recall due to familiar versus unfamiliar L1 accents, here presented in noise. The only evidence for a familiarity-based difference comes from the change detection results, where there was a familiarity benefit for the Tyneside participants. Further research is needed to clarify the mechanism behind this benefit. While a less-detailed processing account could explain the change detection results, it is not useful for the data regarding recall of changed words: for the latter, semantically related changes were corrected more accurately for both accents. Likewise, if the pattern of results is underpinned by a lexical processing imbalance between familiar and unfamiliar L1 accents, the differential effects of semantic proximity remain to be clarified.

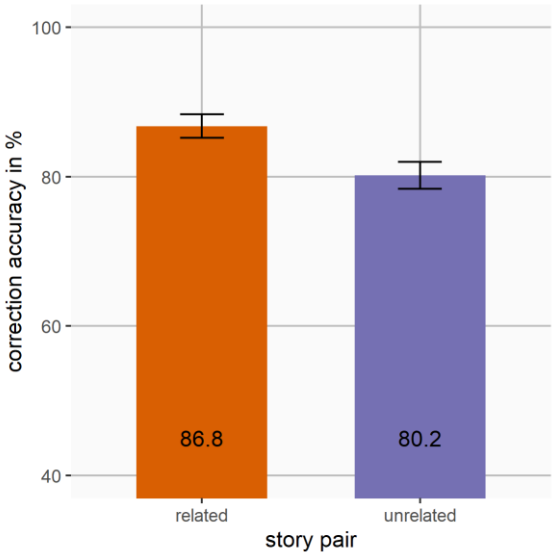


Figure 2: Target word correction accuracy by story pair (error bars: ± one SE)

Finally, auditory inspection by trained phoneticians determined that intelligibility was more greatly affected by noise masking for the NZE speakers than the TE speakers. This auditory impression was supported by long-term average spectra of the four speakers' voices, which showed that the energy of frequency components above 4 kHz was consistently lower for both NZE speakers. This might have artificially lowered the recall of NZE speech and, thus, the interaction between accent and participant origin could have been stronger if different speakers had been used. Importantly, this would not change the primary interpretation of the results because all participants were exposed to both NZE speakers, and the effects described here would likely remain significant.

4. CONCLUSION

This study used a change detection paradigm to investigate the effects of L1 accent familiarity on narrative recall in noise. The results showed a familiarity benefit for change detection for Tyneside participants, but not for New Zealand participants. This familiarity benefit was greater for Tynesiders' detection of semantically related story changes than for unrelated changes. While the exact mechanism behind these findings remains to be determined, it is clear that accent familiarity effects on narrative processing are not only mediated by the native language of the speaker: L1 accents, as well as L2 accents, can affect memory for spoken narratives.

5. ACKNOWLEDGEMENTS

The study presented here is part of the first author's doctoral research, which is funded by the Economics and Social Research Council (ESRC) in the United Kingdom (grant number: 2235617). We would like to thank Donald Derrick and Jen Hay at the University of Canterbury for their help with sourcing and recording the NZE speakers. We would also like to thank the speakers that lend their voices to this research and all participants that took part in it.

6. REFERENCES

[1] Lev-Ari, S., Keysar. B. 2012. Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Processes*, 49, 7, 523–538.

[2] Ferreira, F., Bailey, K. G. B., Ferraro, V. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 1, 11–15.

[3] Ferreira, F., Patson, N. D. 2007. The 'good-enough' approach to language comprehension. *Language and Linguistics Compass*, 1, 1–2, 71–83.

[4] Sturt, P., Sanford, A. J., Stewart, A., Dawydiak, E. 2004. Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, 11, 5, 882–888.

[5] Goslin, J., Duffy, H., Floccia, C. 2012. An ERP investigation of regional and foreign accent processing. *Brain and Language*, 122, 2, 92–102.

[6] Hanulíková, A., van Alphen, P. M., van Goch, M., Weber, A. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience,* 24, 4, 878–887.

[7] Montgomery, C. 2018. The perceptual dialectology of England. In: Braber, N., Jansen, S. (eds.), *Sociolinguistics in England.* Palgrave, 127–164.

[8] Sharma, D., Levon, E., Yang, Y. 2022. 50 years of British accent bias: Stability and lifespan change in attitudes to accents. *English World-Wide,* 43, 2, 135–166.

[9] Adank, P., Evans, B. G., Stuart-Smith, J., Scott, S. K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 2, 520–529.

[10] Impe, L., Geerarts, D., Speelman, D. 2008. Mutual intelligibility of standard and regional Dutch language varieties. *International Journal of Humanities and Arts Computing*, 2, 1–2, 101–117.

[11] Sumner, M., Samuel, A. G. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60, 4, 487–501.

[12] Frances, C., Costa, A., Baus, C. 2018. On the effects of regional accents on memory and credibility. *Acta Psychologica*, 186, 63–70.

[13] Newcastle University IT Service. 2022. Newcastle University Form Builder. https://services.ncl.ac.uk/itservice/collaboration-services/form-builder/.

[14] Burridge, K., Kortmann, B. 2004. Introduction: Varieties of English in the Pacific and Australasia. In: Schneider, E., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (eds.), *A handbook of varieties of English. Volume 1: Phonology*. De Gruyter, 567–579.

[15] Simion, D. 2022. Sound Bible: Party crowd. https://soundbible.com/2163-Party-Crowd.html.

[16] McCloy, D. 2013. Praat script: Noise masking of stimuli. https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat.

[17] Finger, H., Goeke, C., Diekamp, D., Standvoß, K., König, P. 2017. LabVanced: A unified JavaScript framework for online studies. In: *2017 International Conference on Computational Social Science IC^{2}S^2, Cologne, Germany, 10–13 July, 2016.*

[18] R Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/.

[19] Wickham, H. et al. 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, 4, 43, 1–6.

[20] Bates, D. et al. 2022. lme4: Linear mixed-effects models using 'Eigen' and S4. https://cran.r-project.org/web/packages/lme4/index.html.

[21] Singman, H. et al. 2022. afex: Analysis of factorial experiments. https://cran.r-project.org/web/packages/afex/index.html.

[22] Lenth, R. V. et al. 2022. emmeans: Estimated marginal means, aka least-squares means. https://cran.r-project.org/web/packages/emmeans/index.html.

Feature-specific perceptual sensitivity to within-category variations

Kuniko Nielsen[1], Rebecca Scarborough[2]

[1]Oakland University (USA), [2]University of Colorado Boulder (USA)
nielsen@oakland.edu, rebecca.scarborough@colorado.edu

ABSTRACT

This study investigates perceptual sensitivity to types of within-category variation that are produced systematically by speakers. The systematic subcategorical increase of certain features is sometimes hypothesized to be targeted at perceptual benefit. We ask, then, whether there is an asymmetry with respect to the perceptual salience of the increase vs. the decrease of phonetic features. In particular, we examine the within-category perception of voice onset time (VOT) and coarticulatory vowel nasality in English. Two perceptual experiments, an AX discrimination task and an AXB perceptual similarity task, with words with manipulated VOT and nasality were conducted. Results revealed a perceptual asymmetry of within-category variations for VOT: participants were more accurate in discriminating stimulus pairs containing increased (vs. decreased) VOT, and were less likely to identify increased VOT stimuli as similar to a natural VOT token. On the other hand, no such asymmetries were observed for coarticulatory vowel nasality in either experiment.

Keywords: perceptual asymmetry, within-category sensitivity, VOT, nasal coarticulation

1. INTRODUCTION

Speakers produce systematic within-category phonetic variation motivated by lexical and contextual factors that may relate to communicative success. For example, potentially confusable words such as those with many phonological neighbors are produced with more peripheral vowel qualities and greater nasal coarticulation (in appropriate contexts) than words with fewer neighbors [1, 2, 3]. Similarly, voiceless stop-initial words with minimal pair neighbors (e.g., pat, with minimal pair bat) are produced with longer VOT than words with no minimal pair (e.g., pal, with no minimal pair bal) [4, 5]. These adjustments resemble adaptations in pronunciation that are made in communicative contexts that explicitly demand clarity. For example, speakers produce hyperarticulated vowels and more released final stops in “clear” speech [6].

In the case of clear speech, it is generally assumed that the adjustments in pronunciation are made with the purpose of improving clarity, i.e., to result in perceptual benefit. And it has been suggested that the adjustments made in cases of lexical confusability or competition may have the same purpose: hyperarticulated vowels or increased VOT in hard words or contexts could serve to make those words more perceptible. Such effects can only be useful, though, if listeners are sensitive to the variation being produced.

In generating these effects, speakers are manipulating phonetic properties at a subcategorical level of detail. For example, a stop with a 60ms VOT and one with a 100ms VOT are both equally voiceless under the classical assumption that listeners perceive categorically. In categorical perception, while listeners are highly sensitive to differences between categories, listeners are insensitive to differences within a category (e.g., the difference between 60ms and 100ms VOT) [7, 8]. However, growing evidence indicates that, in certain tasks, listeners can hear and respond to within-category detail [e.g., 9, 10, 11]. Furthermore, McMurray et al. [10] showed that fine-grained phonetic details within phonemic categories are preserved in memory and have effects on lexical access. Crucially, such sensitivity could allow listeners to make use of the subcategorical variation produced by speakers in clear speech or confusable words.

Note, though, that it is an increase of vowel peripherality or VOT or coarticulatory nasality that is produced by speakers in harder contexts. Thus, it is the increase of these features, rather than their decrease, that could lead to perceptual benefit. It is possible that listeners might be sensitive only to within-category variation of the kind that contributes to perceptual benefit, but not to kinds of variation that do not. We ask, then, 1) to what extent listeners are sensitive to within-category variation, and 2) whether they are more sensitive to differences that may be useful than those that are not.

Both VOT and vowel nasality play a role in perception at a subcategorical level: the presence of nasality can contribute to lexical perception by predicting an upcoming nasal consonant [11], and the presence of greater VOT leads to perceptual rating of a better instance of a voiceless stop [12]. However, these features also differ in that VOT is an explicitly contrastive feature for English, serving as the primary

cue to stop voicing, while vowel nasality is generally non-contrastive for English (despite the fact that it can help to cue the contrastive nasality of an adjacent nasal consonant).

Nielsen and Scarborough [13] showed that the increased presence of coarticulatory vowel nasality and VOT were more accurately discriminated in an AX (same–different) discrimination task than decreased nasality and VOT. However, that analysis included only the correct response rate for the "different" pairs, which fails to take into account response bias, rendering the results less interpretable. The current study aims to examine within-category perceptual sensitivity in VOT and vowel nasality by extending [13] in two respects: 1) by applying signal-detection analysis [14] for the AX discrimination data focused on differences, and 2) by conducting an additional experiment with an AXB task that examines perceptual similarity.

We predict that listeners should be sensitive to subcategorical variation for both of these features and that they may be more sensitive to variation where these features are increased rather than decreased. Thus, we investigate within-category perceptual sensitivity at two positions within a phonological category: where there is more of a feature (cf., hyperarticulation) and where there is less (cf., hypoarticulation). Further, since phonological contrast influences category structure, we are interested in whether these two features, which differ in phonological status, may show different patterns of perceptual sensitivity.

2. EXPERIMENT 1

2.1. Participants

Twenty-four native speakers of American English participated in a two-alternative forced-choice (AX) discrimination task. All were undergraduates and received course credit for their participation.

2.2. Stimulus selection and construction

Stimuli included 18 monosyllabic words with nasal codas and 19 monosyllabic words with onset /p/. Each set of words was recorded by a phonetically trained male native speaker of American English. Three versions of each word (i.e., more nasality/VOT, less nasality/VOT, and natural) were created. Degree of vowel nasality was increased or decreased by additively combining the waveform of the vowel in each nasal test item (e.g., ban) with the waveform of a more nasal vowel (e.g., from man) or a less nasal/oral vowel (e.g., from bad) respectively in varying ratios by formula using Praat. This process yielded tokens with intermediate spectral characteristics. Test tokens were selected to have approximately equal change in nasality, measured as ±2.5dB A1-P0. (See [15] for nasality measurement details.) VOT was increased or decreased by 40ms by splicing in aspiration from hyper-aspirated tokens or by deleting aspiration from original tokens.

2.3. Procedures

In each trial of the AX task, stimuli were presented in pairs comprising either different stimuli (e.g., moreVOT - naturalVOT) or the same stimulus presented twice. Participants indicated "different" or "same" as quickly as possible. There were 260 trials in each session, equally distributed between same and different.

2.4. Analysis

Signal detection sensitivity, measured as d-prime (d') [14], was calculated for each condition/type for each participant. Statistical analysis of the data was based on mixed-effects modeling using the lmer function in the lme4 package for R [16], with Condition (more vs. less) and Type (nasal vs. VOT) as fixed factors, and random intercepts by Subject.

2.5. Results

Overall average d-primes (across more and less conditions) for VOT and nasal were 0.94 and 0.95, respectively, indicating that the degree of manipulation for the two types of stimuli in terms of their discriminability were comparable.

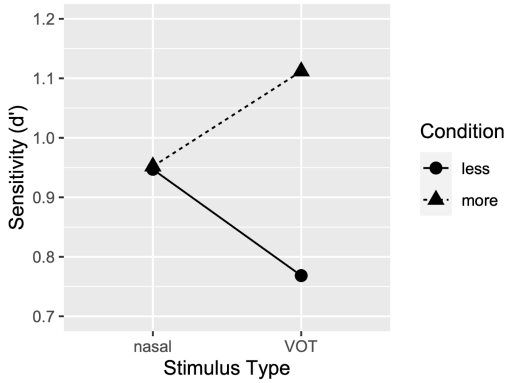


Figure 1: Summary of Experiment 1 (AX discrimination). d-prime is presented by stimulus type (Nasality vs. VOT) and condition (more vs. less).

Figure 1 summarizes the results of Experiment 1 and plots the d' separately for the two types of stimuli and two conditions. There is a clear asymmetry between the two conditions for VOT, with a d' of 1.11 for the more_VOT condition and a d' of 0.77 for the less_VOT condition. However, there is no such difference for nasality, with d' of 0.95 for both the

more_nasal and less_nasal conditions. These patterns of d' were statistically assessed using linear mixed effect regressions. While there is no significant main effect observed for either Type (t=-1.56, p>0.1) or Condition (t<1, p>0.1), the interaction between Type and Condition was significant (t=2.096, p<0.05), confirming that the asymmetry between the two conditions is dependent on the stimulus type; in particular, only VOT shows a within-category asymmetry.

3. EXPERIMENT 2

3.1. Participants

Twenty-six native speakers of American English (different from those in Experiment 1) participated in an AXB perceptual similarity task. All were university undergraduates and received course credit for their participation.

3.2. Stimuli and procedures

The same stimuli from Experiment 1 were used. In each trial of the AXB task, participants heard three repetitions of the same word: an unchanged/natural token (X) and tokens with increased (more) or decreased (less) nasality or VOT (A and B). Participants indicated which of the two flanking items (A or B) sounded more similar to the middle item (X) as quickly as possible. The interstimulus interval (ISI) within triples was either 50ms or 500ms to explore effects of task difficulty.

3.3. Analysis

Responses (coded as *more* or *less*) were analyzed using Generalized Linear Mixed Effects regression (glmer function in the lme4 package for R [16]), with Type (VOT or nasality), Order ("more" as A vs. B), and ISI (50ms vs. 500ms) as fixed factors, and by-participant and by-item random intercepts.

3.4. Results

Figure 2 summarizes the results of Experiment 2 and plots the rate of "more" responses in the AXB task separately for the two types of stimuli. Overall, the "less" item and unchanged/natural item were judged as more similar in 52.6% of responses (vs. 47.4% for "more"). A significant intercept in an intercept-only model (est.=0.66, z=–3.12) confirms that the differences between "more" and unchanged/natural tokens were more perceptible than the differences between "less" and unchanged/natural tokens.

With respect to stimulus type, it can be seen that nasality responses are at 51% (cf., chance at 50%) with respect to 'more' (vs. 'less') responses. VOT, on the other hand, appears to show a 'less' bias, with only 42% 'more' responses (vs. 50% chance). A generalized linear mixed effect regression indicates, however, that the Type difference (nasality vs. VOT) was not significant (z= -1.05, p>0.1).

The effect of Order was significant, where A was chosen as more similar to X more often than B (z=6.84, p<0.001), but there was no main effect of ISI (z= -0.31, p>0.1) nor any significant interactions among the fixed effects (p>0.1).

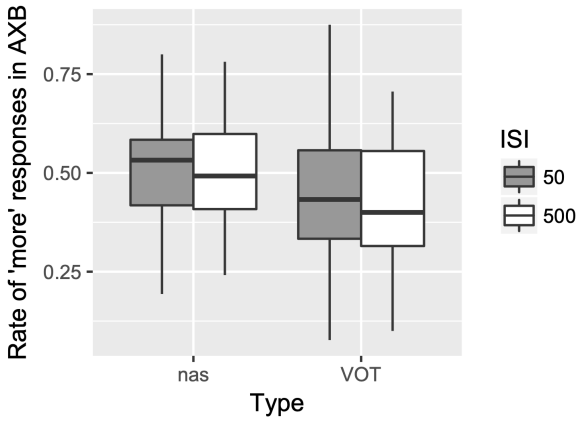


Figure 2: Summary of Experiment 2 (AXB similarity). Rate of 'more' responses (where the token with increased feature was selected as more similar to the natural token), presented by stimulus type (Nasality vs. VOT).

4. DISCUSSION

We were interested in exploring the potential perceptual consequences of subcategorical variation of the type produced systematically by listeners in various "hard" contexts. Speakers tend to hyperarticulate in such contexts, and it is generally suggested that this hyperarticulation has perceptual benefit for listeners. However, this perceptual benefit could not occur unless listeners have the ability to perceive the relevant subcategorical variation. To test this, we investigated listeners' sensitivity to variation in two phonologically relevant features (VOT and nasality) and explored a possible perceptual asymmetry between increased vs. decreased degrees of these features using an AX discrimination test (Experiment 1) and an AXB similarity test (Experiment 2).

Experiment 1 showed that listeners were more accurate in discriminating stimulus pairs containing tokens with increased VOT than pairs containing a decreased VOT token, while no difference in discrimination was observed for increased vs. decreased nasality pairs. Contrary to Nielsen and Scarborough [13], which showed that nasal tokens were discriminated more accurately than VOT tokens in the AX task, the overall discriminability for the two feature types were comparable in our data.

Experiment 2 showed that participants judged decreased-feature stimuli as more similar to the natural stimuli in both stimulus types. This suggests that an increase in these features (VOT and vowel nasality) is perceptually more salient than a decrease.

Both experiments showed listener sensitivity to within-category variation in that listeners could discriminate tokens with subcategorical differences and/or make systematic comparisons of tokens differing subcategorically. These findings contribute to a body of evidence that, despite what categorical perception would predict, listeners do have access to low-level phonetic distinctions within categories.

Both experiments also illustrated a bias in sensitivity, with greater sensitivity to variation in parts of a category that exhibit "more" rather than "less" of a feature, at least for some features. Since in at least the cases examined here, and probably more generally as well, it is an increase in a feature that is realized in contexts where clarity is required, this greater sensitivity to "more" suggests a link between within-category sensitivity and communicative benefit.

However, the asymmetrical perceptual sensitivity between "more" and "less" is shown only for VOT and not for nasality in Experiment 1 (with a non-significant difference in Experiment 2 that trends in the same direction). We consider several factors that might contribute to this feature-conditioned difference in result. In other words, why do listeners show evidence of more structured patterns of sensitivity to VOT than to vowel nasality?

An important difference between VOT and nasality lies in the phonological status of the two features. VOT is an explicitly contrastive feature in English, capturing the contrast between voiced and voiceless stops. In fact, VOT is the primary cue conveying this contrast. Vowel nasality, on the other hand, is generally not contrastive for English; rather it is the coarticulatory consequence of an adjacent nasal consonant. To the extent that vowel nasality carries information in English, then, it is information about the adjacent nasal (or oral) consonant. And there are cues other than vowel nasality that signal the nasal/oral contrast on the following consonant.

One might imagine that the contrastive status of VOT would cause the feature to be more categorical and therefore would lead to reduced perceptual sensitivity. In fact, although listeners were clearly able to attend to within-category variation in VOT, there are parts of the VOT category that listeners were not as sensitive to (namely, the reduced VOT parts of the category). With respect to nasality, listeners' sensitivity is distributed more broadly across the category.

We note an additional (probably related) difference between VOT and nasality. VOT is systematically structured in its variation, which is relatively predictable by speaker [17]. Degree of nasality, on the other hand, is quite variable across speakers, and even within speakers, as measured acoustically [18]. Not only are some speakers simply more nasal than others at their baseline, but the acoustic complexity of nasality leaves it vulnerable to influence from other properties of the signal (e.g., vowel quality or voice quality). This structured variation may contribute directly to a more structured perceptual sensitivity for VOT, which can be matched to our manipulation of increased and decreased VOT. The general variability in nasality, on the other hand, which is reflected in our manipulated tokens, may make it harder for listeners to assess what is more nasality vs. less.

The fact that this fine-grained featural difference in sensitivity is more evident in Experiment 1 (AX discrimination) than Experiment 2 (AXB similarity judgment) is unsurprising, given the much greater difficulty of the AXB task. While AX discrimination requires a single comparison ("same" or "different" between the members of a pair), AXB tasks require the listener to compare both A and B to X, and then to compare those two comparisons gradiently. Previous research has shown that such differences in task difficulty lead to perceptual difference limens that are double or more when calculated from AXB tasks, relative to AX tasks [18], indicating greater perceptual sensitivity in the AX task. Thus, the AX task allows listeners to better respond to low-level within-category acoustic differences.

5. CONCLUSION

Taken together, our results demonstrate that listeners are sensitive to within-category differences in both VOT and vowel nasality. Further, both experiments reveal an asymmetry in this subcategorical perceptual sensitivity: stimuli with increased VOT were perceived more accurately than the stimuli with decreased VOT. For nasality, however, the "more" vs. "less" asymmetry in sensitivity is weaker or absent. In other words, neither feature shows the flat within-category sensitivity that Categorical Perception would suggest; rather, listeners are sensitive to some within-category differences. They are especially sensitive to those differences where a feature is increased, rather than decreased, mirroring the systematic variation produced by speakers in hyperarticulation contexts. In other words, listeners have greatest perceptual sensitivity to those phonetic details that may be more communicatively useful.

6. REFERENCES

[1] R. Wright, “Factors of lexical competition in vowel articulation,” in *Papers in Laboratory Phonology VI*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, Cambridge), pp. 26–50, 2004.

[2] B. Munson and N. P. Solomon, “The effect of phonological neighborhood density on vowel articulation,” *Journal of Speech, Language and Hearing Research* 47(5), 1048–1058, 2004.

[3] R. Scarborough, “Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation,” *Journal of Phonetics* 41(6), 491–508, 2013.

[4] M. Baese-Berk and M. Goldrick, “Mechanisms of interaction in speech production,” *Language and Cognitive Processes* 24(4), 527–554, 2009.

[5] N. P. Fox, M. Reilly., and S. E. Blumstein, *“*Phonological neighborhood competition affects spoken word production irrespective of sentential context,” *JML* 83, 97-117, 2015.

[6] M. Picheny, N. Durlach, and L. Braida, *“*Speaking clearly for the hard-of-hearing II: Acoustic characteristics of clear and conversational speech,” *JSLHR* 29, 434–446, 1986.

[7] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of experimental psychology*, 54(5), 1957.

[8] S. Harnad, "Psychophysical and cognitive aspects of categorical perception: A critical overview," In *Categorical perception: The groundwork of cognition,* 1-52, 1987.

[9] P. K. Kuhl, "Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not." *Perception & psychophysics* 50.2, 93-107, 1991.

[10] B. McMurray, M. K. Tanenhaus, and R. N. Aslin, "Gradient effects of within-category phonetic variation on lexical access," *Cognition*, 86(2), B33-B42, 2002.

[11] P. S. Beddor, K.B. McGowan, J. E. Boland, A. W. Coetzee, and A. Brasher, “The time course of perception of coarticulation,” *JASA* 133(4), 2350-2366, 2013.

[12] J. S. Allen, & J. L. Miller, “Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate,” *Perception & Psychophysics,* 63(5), 798-810, 2001.

[13] K. Y. Nielsen & R. Scarborough, “Perceptual asymmetry between greater and lesser vowel nasality and VOT,” Proc. of ICPhS, 2015.

[14] N. A. MacMillan and C. D. Creelman, “*Detection theory: A users guide,*” Cambridge, England: Cambridge University Press, 1991.

[15] M. Y. Chen, “Acoustic correlates of English and French nasalized vowels,” JASA, 102, 2360–2370, 1997.

[16] D. Bates, M. Mächler, B. Bolker B, and S. Walker S, “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software,* 67(1), 1–48, 2015.

[17] E. Chodroff and C. Wilson, “Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English,” *Journal of Phonetics,* 61, 30-47, 2017.

[18] W. Styler, “On the acoustical features of vowel nasality in English and French,” *JASA*, 142(4), 2469-2482, 2017.

[19] W. A. Rosenblith and K. N. Stevens, “On the DL for frequency”, *JASA,* 25, 980-985, 1953.

Incremental cue training: a study of lexical tone learning by non-tonal listeners

Yanyu Li, Laurence White & Ghada Khattab

Newcastle University
y.li218@newcastle.ac.uk, laurence.white@newcastle.ac.uk, ghada.khattab@newcastle.ac.uk

ABSTRACT

Speech perception requires sensitivity to acoustic dimensions relevant to meaningful distinctions. Training with exaggerated cues has been used to direct second language (L2) learners' attention to dimensions relevant for segmental contrasts; here we applied similar incremental cue training to lexical tone contrasts. Two groups of native English listeners were trained, either on a fixed tonal contrast (*high-rising* vs *high-falling*) or with incremental stimuli, where pitch movements relevant to the rising/falling contrast were exaggerated at training outset and reduced, stepwise, to the fixed contrast level by training completion. An ABX task tested discrimination, with feedback during training. Final tonal discrimination accuracy was similar between groups, but prior learning trajectories contrasted: the Incremental group had higher initial performance, i.e., with the exaggerated contrasts, whilst the Fixed group improved gradually over training. Finally, despite contrasting training, the groups had comparable pre-test to post-test improvement on a separate untrained tonal contrast (*high-level* vs *mid-rising*).

Keywords: lexical tone, perception, incremental cue training, second language acquisition

1. INTRODUCTION

Speech perception is shaped by early exposure to first languages (L1s). Attention-to-dimension models, such as the Native Language Magnet Model (NLM) [1, 2], describe the perceptual system as a distorted space where speech sound discrimination is based on perceived acoustic distances. Critical dimensions for sound discrimination are "stretched", leading to increased perceptual sensitivity, while the less relevant dimensions are "compressed", resulting in reduced sensitivity [3–8]. Perceptual distance along critical dimensions is additionally warped for efficient categorisation: distance within category prototypes is compressed, while distance at category boundaries is stretched.

Such adaptation to L1 exposure leads to efficient L1 perception, but potentially interferes with non-native perception (what we call here "L2 perceptual bias"). Thus, non-native listeners behave differently from L1 listeners in phonetic discrimination. For instance, Japanese L1 listeners tend to confuse English /l/ and /r/ [9, 10]. Iverson et al. [3] found that in perceiving the contrast, Japanese listeners relied more on F2, which was a less important dimension for English L1 listeners. Similarly, Spanish L1 listeners' perception of English /i/-/ɪ/ is influenced more by vowel duration than spectral features [11, 12]. L2 perceptual bias is also seen in perception of suprasegmental features, such as lexical tone. For instance, in lexical tone perception, English L1 listeners rely more on pitch height, while Thai and Mandarin L1 listeners are differentially sensitive to pitch contour [13–15].

Despite L2 perceptual bias, the adult listener's perceptual space is not invariable. Francis et al. [16] trained Mandarin and English L1 listeners on Cantonese tone perception. Multi-dimensional scaling (MDS) showed that L1 listeners of both languages increased sensitivity to previously less-attended dimensions. Mandarin listeners improved discrimination on two Cantonese *rising* tones 23 and 25 (as represented in Chao digits [17]), with common pitch trajectory, but differing pitch height. English listeners improved discrimination between Cantonese 23 and 33, which contrast in pitch trajectory but are similar in pitch height. Likewise, Chandrasekaran et al. [13] trained English L1 listeners on Mandarin tones which contrasted mainly on pitch contour. MDS indicated that listeners improved discrimination based on pitch trajectory, but showed no training benefit for pitch height-based discrimination.

As phonetic learning seems to be strongly influenced by perceptual attention [13, 16], training techniques based on attention-to-dimension theories – including adaptive cue training and inhibition training – have been tested for segmental learning [11, 19, 20]. There is evidence of listener discrimination improvement with these techniques, whilst their relative benefits are still debated [11, 19]. Adaptive or incremental cue training, for example, aims to redirect listeners' attention to previously less attended dimensions to improve phonetic perception. Training expands critical acoustic dimensions to boost phonetic discrimination at the beginning of training. This enhancement gradually reduces to a natural range by the end of training.

Whilst there is evidence that the tonal perceptual space can be modified by adult experience [13, 16], the efficacy of incremental cue training for L2 lexical tone learning has not been tested. Therefore, the present study examines whether the initial use of exaggerated pitch cues to lexical tone contrasts improves inexperienced English listeners' tonal discrimination, compared to a training condition using non-exaggerated cues.

2. METHOD

2.1. Participants

We tested 90 adult native English-speaking listeners (F=74, M=16). Only the 86 participants without tonal language experience were included in the analyses. These 86 participants were randomly assigned to the Incremental (N=40) and Fixed groups (N=46).

2.2. Stimuli

The stimuli were 36 stylised tonal syllables, combining 4 pseudo-tones and 9 syllables. The four synthesised tones shared the same onset pitch height, to encourage listeners' attention to pitch contour. The four tones were *high-rising* 45, *high-falling* 41, *level* 44, and *falling-rising* 425 (Figure 1).

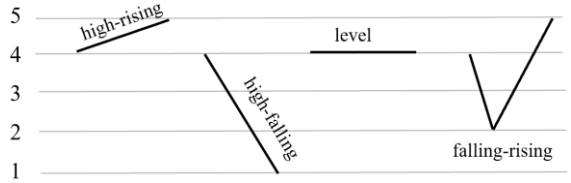


Figure 1: The pseudo-tone system used as the basis for the experimental stimuli.

These tones were presented in nine consonant-vowel (CV) syllables consisting of three voiceless aspirated plosives /p/, /t/, /k/ and three open long vowels /a/, /i/, /u/. All syllable combinations obeyed Mandarin and English phonotactic constraints.

2.2.1. Recording

Two female Mandarin L1 speakers recorded the nine syllables with four Mandarin tones (*high-level* 55, *mid-rising* 35, *falling-rising* 214, *high-falling* 51) in a sound-attenuated booth. Each syllable was embedded in a Mandarin sentence, e.g., (English translation) "I read '/paa/ 35' three times". All sentences were presented three times on a screen. Speakers were asked to read sentences aloud in a normal, even pace.

2.2.2. Pitch manipulation

Extraction. For each speaker, one example of each token was selected from the three recorded repetitions. The target tonal syllables were then extracted from the carrier utterances using Praat [21].
Manipulation. The recorded syllables were overlaid with the four pseudo-tones via Pitch-Synchronous Overlap and Add (PSOLA) [22, 23] in Praat. The PSOLA manipulation is intended to change F0 trajectories without affecting other acoustic features.

The endpoints of the pseudo-tones, as represented in Chao digits, were converted into Hertz values according to each speaker's fundamental frequency (F0) range, based on the formula "T = [(lg X - lg L) / lg H - lg L] * 5" [24], where H and L are each speaker's highest and lowest F0 points and X is the F0 point to be converted. The converted F0 values for the onset and offset of each tone are shown Table 1, also the turn point for the falling-rising tone.

Pseudo-tone (Chao digits)	Speaker 1 (Hz)	Speaker 2 (Hz)
High-rising **45**	297-375	294-335
High-falling **41**	297-147	294-201
Level **44**	297-297	294-294
Falling-rising **425**	297-240-375	294-260-335

Table 1: F0 values (Hz) of the four pseudo-tones.

The extracted Mandarin syllables were manipulated to the pseudo-tone pitch values (Table 1). Mandarin tone *mid-rising* 35 was converted to pseudo-tones *high-rising* 45 and *falling-rising* 425, Mandarin *high-falling* 51 to pseudo-tone *high-falling* 41, and Mandarin *high-level* 55 to pseudo-tone *level* 44. The natural *falling-rising* tone 214 was not used in the conversion, due to phrase-medial realisation as 21, as well as creakiness and other phonetic confounds.
Test stimuli. The contrasting tonal pair *level* 44 and *falling-rising* 425 (Table 1) were used in pre- and post-training tests.
Training stimuli. The contrasting tonal pair *high-rising* 45 and *high-falling* 41 were used in training blocks. For the incremental group, the F0 endpoints of the tones were manipulated to enhance the difference between *rising* and *falling*. F0 offset values for *high-rising* 45 were raised by two 30Hz steps. Likewise, F0 offsets for *high-falling* 41 was lowered by two 30Hz steps. Table 2 shows the F0 values for each level of enhancement for each tone.

2.3. Procedure

The experiment consisted of 5 blocks, taking about 30 minutes per participant. Blocks 1 and 5 were the pre-test and post-test, using the pseudo-tone contrast *level* 44 *vs falling-rising* 425. Blocks 2, 3, and 4 were training blocks, using the contrast *high-rising* 45 *vs high-falling* 41. Each testing or training block contained 36 ABX trials: 9 syllables x 2 tonal orders (AB/BA) x 2 test stimuli for each order (ABA/ABB).

In each ABX trial, participants heard two segmentally-identical syllables (spoken by the same speaker), one for each pseudo-tone, and then the same syllable (spoken by the other speaker), with one of the contrasting pseudo-tones (i.e., A or B). All trials were randomised within each block.

Pseudo-tone (Chao digits) by block	Speaker 1 (Hz)	Speaker 2 (Hz)
High-rising **45**		
Block 2	297-435	294-395
Block 3	297-405	294-365
Block 4 (baseline)	297-375	294-335
High-falling **41**		
Block 2	297-87	294-141
Block 3	297-117	294-171
Block 4 (baseline)	297-147	294-201

Table 2: Enhanced and baseline F0 values (Hz) for pseudo-tones 45 and 41 for the Incremental group. Block 2 has the most exaggerated contrasts; Block 4 is identical to the Fixed group baseline.

In Blocks 1 and 5, participants had up to 5 seconds to respond in each trial, to indicate whether the third sound heard was the same as the first or the second. No feedback was provided regarding the response.

In Blocks 2 to 4 (training), the task was the same, but the timeout was 10 seconds and visual feedback was given ("correct"/ "incorrect"). For the Fixed group, the stimuli were the unaltered (baseline) syllables for all blocks. For the Incremental group, the most exaggerated versions of the pseudo-tones were used in Block 2, intermediate versions in Block 3, and the baseline contrasts in Block 4 (Table 2).

3. RESULTS

Separate accuracy analyses were conducted for test and training blocks. In test blocks, as shown in Figure 2, there was an overall accuracy improvement from pre-test (Block 1) to post-test (Block 5) for both Incremental (67% to 76%) and Fixed groups (70% to 76%.). In training blocks, the Fixed group's accuracy improved from 69% to 75% from Block 2 to Block 4, whilst the Incremental group's accuracy dropped from 81% (Block 2) to 76% (Block 4). Thus, whilst the Incremental group performed numerically better at the start of training, with exaggerated stimuli, the Fixed group performed comparably at the end of training, when the two groups were exposed to the same (baseline) stimuli.

We analysed these accuracy patterns via logistic mixed-effect regression models, with Group (Incremental vs Fixed) and Block (coded as a numeric variable) as fixed effects. There was a random intercept for participants and a random slope for block-by-participant. A random intercept for syllable type was excluded because it did not improve model fit. Significance of predictors was determined by likelihood ratio tests.

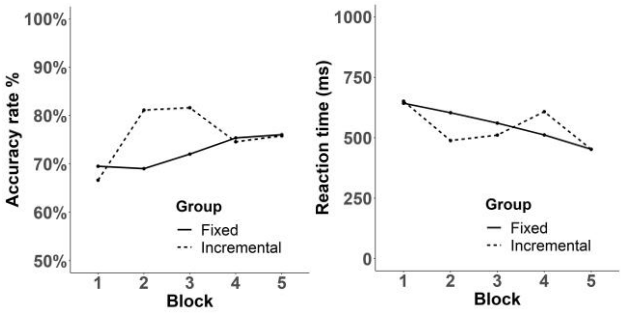


Figure 2: Tone discrimination accuracy and reaction time across the five blocks (Blocks 1 & 5: test-only; Blocks 2, 3 & 4: training) by group (Incremental vs Fixed)

3.1. Discrimination accuracy analyses

3.1.1. Test block accuracy

There was an effect of Block in the test-only blocks, $\chi^2(1) = 31.96$, $p < .001$: thus, there was an overall improvement in accuracy from pre-test to post-test (Figure 2 left: Block 1 *vs* Block 5) in distinguishing the pseudo-tone contrasts *level* and *falling-rising*. There was no effect of Group, $\chi^2(1) = 0.53$, $p = .47$, and no Group x Block interaction, $\chi^2(1) = 0.37$, $p = .54$.

3.1.2. Training block accuracy

In training blocks, Group had an effect, $\chi^2(1) = 20.44$, $p < .001$, indicating that the Incremental group performed better than the Fixed group overall. There was also an effect of Block, $\chi^2(2) = 8.69$, $p = .003$, and a Block x Group interaction, $\chi^2(2) = 20.10$, $p < .001$. Post-hoc pairwise analyses were conducted between every two blocks within each group and between groups for each block. Comparing Blocks 2 and 4, the Incremental group dropped accuracy, $\chi^2(1) = 10.07$, $p = .002$, whereas the Fixed group improved accuracy, $\chi^2(1) = 13.31$, $p < .001$. Looking at adjacent blocks, the Incremental group decreased accuracy between Blocks 3 and 4, $\chi^2(1) = 19.09$, $p < .001$, whilst the Fixed group improved accuracy between Blocks 2 and 3, $\chi^2(1) = 4.76$, $p = .03$. Comparing within blocks, the Incremental group were more accurate than the Fixed group in Block 2, $\chi^2(1) = 9.42$, p = .002, and Block 3, $\chi^2(1) = 5.80$, p = .02. Despite their contrasting learning trajectories, the two groups had comparable Block 4 discrimination accuracy, $\chi^2(1) = 0.10$, p = .75 (Figure 2 left: Blocks 2 – 4).

3.2. Reaction time analyses

Reaction times (RTs) on correct-response trials were analysed via linear mixed-effect regression models.

The fixed effects and random effects structure was as for the accuracy data. We eliminated RT outliers, via lower and upper thresholds for each participant based on individual means and standard deviations (SDs). Thus, RTs below or above 2.5SD from the mean were removed from further analyses.

3.2.1. Test block reaction times

RTs for both groups reduced from pre-test to post-test (Figure 2 right: Block 1 *vs* Block 5), $\chi^2(1) = 46.57$, $p < .001$. There was no effect of Group, $\chi^2(1) = 0.15$, $p = .70$, nor an interaction between Group and Block, $\chi^2(1) = 0.01$, $p = .92$.

3.2.2. Training reaction times

For training block RTs, there was no effect of Block, $\chi^2(2) = 0.54$, $p = .50$, or Group, $\chi^2(1) = 0.08$, $p = .77$, nor an interaction, $\chi^2(2) = 1.33$, $p = .25$.

4. DISCUSSION AND CONCLUSION

Overall, listeners' tonal discrimination accuracy improved with experience. Firstly, performance was better and reaction times were quicker on the untrained tonal contrast (*level* 44 vs *falling-rising* 425) in the final block (after training) than in the initial block. This improvement did not differ between Incremental and Fixed training groups.

For the trained tonal contrast (*high-rising* 45 *vs high-falling* 41), learning trajectories for the two training groups were in opposite directions. The Fixed group started with lower accuracy which increased over training. The Incremental group started with higher accuracy which decreased over training. The two groups had comparable accuracy at the end of training, i.e., when both were exposed to the same baseline tonal contrast (Block 4).

For the trained tonal contrast (Blocks 2-4), RTs did not show statistically-robust (overall or pairwise) variation. There was a non-significant numerical tendency for faster RTs in more accurate blocks, which at least runs counter to any possible speed/accuracy trade-off.

The overall training improvement, on both trained and untrained stimuli, is consistent with previous findings on lexical tone learning [13, 16, 25]. However, the higher accuracy for the Incremental group at training outset was not seen in other studies which also used enhanced stimuli at training outset [11, 19]. The initial performance benefit in the Incremental group appears to be due to the stimulus enhancement, where pitch trajectory differences between the pseudo-tonal contrasts were maximised and required less effort to distinguish. That this difference was observed here, but not in other similar studies, may relate to training design contingencies.

First, the study on training Spanish listeners on English /i/-/ɪ/ [11] used adaptive stimuli and found them to be more effective than fixed stimuli. One difference with our findings is in the training implementation. In Kondaurova and Francis [11], participants had to reach an accuracy threshold in consecutive sessions before stimulus enhancement was reduced. Potentially this adaptive manipulation is more sensitive to individuals' learning trajectories than our training regime, where the enhancement reduced automatically in successive training blocks.

Second, in our study, lack of Block 2 to 3 accuracy fall-off for the Incremental group suggests that we used relatively salient levels of cue enhancement in both blocks, clearly above the just-noticeable-difference for tonal contours [26]. Starting with a less exaggerated contrast tailored to provide a reliable, but minimal discrimination boost – along with more, but smaller steps of decreasing enhancement – may help to maintain higher performance back to baseline.

Third, studies of L2 segmental learning using enhanced stimuli [11] have typically implemented training for 4 to 10 days [11, 19]; where found, the boost to learning manifested as improvement over multiple sessions [11]. Insofar as accuracy is high in early training blocks, our findings support an overall beneficial effect of enhanced stimuli, which may be sustained with extended training, although we note that Iverson et al. [19] found no ultimate benefit for initially-enhanced stimuli over multiple training sessions. The nature of the learned L2 contrast, the degree of enhancement and the length/frequency of training are all likely to influence relative outcomes.

Finally, there were no robust response latency differences between the Incremental and Fixed groups (notwithstanding the numerically contrasting patterns across training blocks, see Figure 2). These null results were comparable to [11], which used an adaptive training method over a longer period. This reinforces our accuracy data conclusions regarding the similar final impact, over one session, of Incremental and Fixed stimuli. Further work will examine how stimulus enhancement affects reaction time over extended training.

In summary, F0 exaggeration of training stimuli initially boosted discrimination of lexical tones. The improvement in performance for this Incremental group, relative to the Fixed group (who were presented with non-exaggerated stimuli throughout), was, however, not sustained when stimulus exaggeration was eliminated in the final training block. Discrimination of untrained tonal contrasts improved equally for the Fixed and Incremental training groups.

5. ACKNOWLEDGEMENTS

This study is supported by the Chinese Scholarship Council (CSC).

6. REFERENCES

[1] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, 'Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)', *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 363, no. 1493, pp. 979–1000, Mar. 2008, doi: 10.1098/rstb.2007.2154.

[2] P. K. Kuhl and P. Iverson, 'Linguistic experience and the "perceptual magnet effect"', *Speech Percept. Linguist. Exp. Issues Cross-Lang. Res.*, pp. 121–154, 1995.

[3] P. Iverson *et al.*, 'A perceptual interference account of acquisition difficulties for non-native phonemes', *Cognition*, vol. 87, no. 1, pp. B47–B57, Feb. 2003, doi: 10.1016/S0010-0277(02)00198-1.

[4] A. L. Francis, N. Kaganovich, and C. Driscoll-Huber, 'Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English', *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1234–1251, Aug. 2008, doi: 10.1121/1.2945161.

[5] P. W. Jusczyk, 'From general to language-specific capacities: the WRAPSA Model of how speech perception develops', *J. Phon.*, vol. 21, no. 1, pp. 3–28, Jan. 1993, doi: 10.1016/S0095-4470(19)31319-1.

[6] R. M. Nosofsky, 'Attention, Similarity, and the Identification-Categorization Relationship', p. 19, 1986.

[7] R. L. Goldstone, 'Feature distribution and biased estimation of visual displays.', *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 19, no. 3, p. 564, 1993.

[8] R. L. Goldstone, 'Influences of categorization on perceptual discrimination.', *J. Exp. Psychol. Gen.*, vol. 123, no. 2, p. 178, 1994.

[9] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, 'Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/', *J. Phon.*, vol. 32, no. 2, pp. 233–250, Apr. 2004, doi: 10.1016/S0095-4470(03)00036-6.

[10] H. Goto, 'Auditory perception by normal Japanese adults of the sounds "l" and "r."', *Neuropsychologia*, vol. 9, pp. 317–323, 1971, doi: 10.1016/0028-3932(71)90027-3.

[11] M. V. Kondaurova and A. L. Francis, 'The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods', *J. Phon.*, vol. 38, no. 4, pp. 569–587, Oct. 2010, doi: 10.1016/j.wocn.2010.08.003.

[12] P. Escudero and P. Boersma, 'Bridging the gap between L2 speech perception research and phonological theory', *Stud. Second Lang. Acquis.*, vol. 26, no. 4, pp. 551–585, 2004.

[13] B. Chandrasekaran, P. D. Sampath, and P. C. M. Wong, 'Individual variability in cue-weighting and lexical tone learning', *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 456–465, Jul. 2010, doi: 10.1121/1.3445785.

[14] J. Gandour, 'Tone perception in Far Eastern languages', *J. Phon.*, vol. 11, no. 2, pp. 149–175, Apr. 1983, doi: 10.1016/S0095-4470(19)30813-7.

[15] J. Gandour, 'TONE DISSIMILARITY JUDGMENTS BY CHINESE LISTENERS/声调异同辨别的测试', *J. Chin. Linguist.*, pp. 235–261, 1984.

[16] A. L. Francis, V. Ciocca, L. Ma, and K. Fenn, 'Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers', *J. Phon.*, vol. 36, no. 2, pp. 268–294, 2008.

[17] Y.-R. Chao, 'A system of tone letters', *Maître Phon.*, 1930.

[18] P. C. Wong and T. K. Perrachione, 'Learning pitch patterns in lexical identification by native English-speaking adults', *Appl. Psycholinguist.*, vol. 28, no. 4, pp. 565–585, 2007.

[19] P. Iverson, V. Hazan, and K. Bannister, 'Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults', *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3267–3278, 2005.

[20] D. G. Jamieson and D. E. Morosan, 'Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones', *Percept. Psychophys.*, vol. 40, no. 4, pp. 205–215, Jul. 1986, doi: 10.3758/BF03211500.

[21] P. Boersma and D. Weenink, 'Praat: doing Phonetics by Computer', 2005. https://www.fon.hum.uva.nl/praat/ (accessed Dec. 16, 2021).

[22] P. Boersma, 'Praat: doing phonetics by computer', *Httpwww Praat Org*, 2006.

[23] E. Moulines and J. Laroche, 'Non-parametric techniques for pitch-scale and time-scale modification of speech', *Speech Commun.*, vol. 16, no. 2, pp. 175–205, Feb. 1995, doi: 10.1016/0167-6393(94)00054-E.

[24] Y. Wang, A. Jongman, and J. A. Sereno, 'Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training', *J. Acoust. Soc. Am.*, vol. 113, no. 2, pp. 1033–1043, Feb. 2003, doi: 10.1121/1.1531176.

[25] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, 'Training American listeners to perceive Mandarin tones', *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3649–3658, Dec. 1999, doi: 10.1121/1.428217.

[26] C. Liu, 'Just noticeable difference of tone pitch contour change for English- and Chinese-native listeners', *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3011–3020, Oct. 2013, doi: 10.1121/1.4820887.

TESTING THE PERCEPTUAL BASIS OF LARYNGEAL METATHESIS AND RARITY OF PREASPIRATED STOPS

Madeline Gilbert

Laboratoire de Phonétique et Phonologie (UMR 7018, CNRS/Sorbonne Nouvelle, Paris)
madeline-besse.gilbert@sorbonne-nouvelle.fr

ABSTRACT

Pre-stop coda [h] is often argued to be perceptually weaker than post-stop [h]. This perceptual property of [h] is often used to explain why laryngeal metathesis moves [h] from before a stop to after a stop (HC → CH), and why preaspirated stops are rarer than (post)aspirated stops. This study tests Arabic, English, and French-speaking listeners' perception of [h] in h-stop (HC) and stop-h (CH) sequences using an ABX perception task. The status and phonotactics of /h/ and aspirated stops differ in these languages. Results provide no evidence that [h] is easier to perceive after a stop than before a stop. For all groups, [h] was either equally perceptible in HC and CH sequences (Arabic, English), or more perceptible in HC sequences (French). Perception is largely driven by native language experience.

Keywords: cross-linguistic perception, metathesis, laryngeals, preaspiration

1. INTRODUCTION

The perceptibility of [h] is often argued to differ based on phonotactic position. According to Bladon [1], preconsonantal [h] is difficult to perceive because it has weak acoustic cues (low-intensity, dispersed energy), and because it has formants in the same regions as the preceding vowel, allowing the vowel to 'mask' it (see also [2]). In addition to weak acoustic cues, Silverman [2] and Kingston [3] argue that [h] is not salient before a stop because it occurs at the vowel offset, where air pressure is low. They argue that [h] is easier to perceive *following* a stop because stop releases are highly salient (high-intensity, high-volume, noisy), and because the auditory system is more sensitive to the onset of a stimulus (at a stop release, following silence) than to its offset (at the end of a vowel). The asymmetry in perceptibility between pre- and post-stop position is used to explain both laryngeal metathesis and the cross-linguistic rarity of preaspirated stops.

Metathesis is relatively uncommon, but laryngeals are frequent participants [4]. From one perspective, metathesis occurs to optimize the perception of one or both sounds [5, 6, 7].[1] Steriade [6] argues that regular, systematic metathesis improves perceptibility. This condition seems to hold for laryngeal metathesis, which typically operates in the direction [hC] → [Ch]. Most accounts assume that this directionality is because [h] is easier to perceive in the resulting [Ch] sequences than in the original [hC] sequences [8, 9, 4]. Cho [9] points out that the domain of laryngeal metathesis is often limited in ways that support this proposal. Metathesis in Cherokee and Korean is limited to contexts in which [h] would dock on a voiceless stop. In contrast, when [h] would dock on a sonorant, [h] fails to metathesize in Cherokee and deletes in Korean [9]. He argues that metathesis might be limited to occurring around voiceless stops because it is only these contexts where metathesizing improves perceptibility; docking [h] on a sonorant would not.

The supposed perceptual difficulty of preconsonantal [h] has also been used to explain the cross-linguistic rarity of preaspirated stops relative to postaspirated stops [2, 10]. Preaspirated stops are allophonically conditioned in ways that might optimize perception. For example, Icelandic aspirated stops are preaspirated word-medially and word-finally, but postaspirated word-initially [11, 12], where there would be no vowel preceding HC to provide the transitional cues necessary to perceive [h] [13]. Furthermore, the aspiration portion of preaspirated stops tends either to be lost, or to be strengthened by adding an oral constriction ([hp, ht, hk] → [fp, ɕt, xk]; [2, 10]), which might also suggest that preconsonantal glottal [h] is difficult to perceive on its own.

Thus, accounts of laryngeal metathesis and preaspiration rely on the hypothesis that [h] is less perceptible before a stop than on its release. Clayton [10] is one of the few studies that tests this hypothesis. In an AX task, he found that word-medial preaspiration ([VhCV]) was *not* universally more difficult than word-initial postaspiration ([#Ch]). However, the duration of

[h] was not controlled, and high accuracy on HC sequences could be due to longer [h] in these words.

I test the perceptual optimization account of laryngeal metathesis around voiceless stops, whereby [h] is supposedly easier to perceive in the CH sequences resulting from metathesis than in its original position (HC). An alternative hypothesis is that [h] perception is driven by native language experience. Listeners have been found to perceive sounds that exist in their native language better than sounds that do not (e.g., [14, 15, 16]), and Mielke [17] finds similar effects for [h]. To test language-specific perception, the current study was run on Arabic, English, and French-speaking listeners, whose native languages differ in the presence and status of HC and CH sequences.

2. METHOD

2.1. Materials and procedure

The stimuli were trisyllabic nonce words (penultimate stress) with medial consonants /p, t, k/ flanked by /a/. Word sets were built off three base items: *famapa*, *lamaka*, *lanata*. Each word set contained versions with no [h] (NoH; [VCV]), pre-posed [h] (HC; [VhCV]), and post-posed [h] (CH; [VChV]). These words were organized into HC/C and CH/C comparisons, which test listeners' accuracy in perceiving [h] in pre-stop and post-stop position, respectively (Table 1). Listeners heard all possible presentation orders (AAB, ABA, BAA, BAB), with an ISI of 500ms between words. These stimuli are part of a larger study with other medial consonants and an additional comparison type, which are not reported here.

Table 1: Sample trials for medial consonant /t/

Type	Order	Example
HC/C	HC-None-HC	lanahta-lanata-lanahta
CH/C	CH-None-CH	lanatha-lanata-lanatha

Stimuli were recorded by a male native speaker of Turkish (a trained linguist) in a sound booth, using a Shure KSM44 microphone with a pop filter. Turkish was used because it allows clusters with [h] and gives no group a native-listener advantage. Two productions of each word were used, so that X would not be the same physical token as A or B. Test nonce words were manipulated so that words in the same set (*lanata, lanahta, lanatha*) matched for duration of the preceding vowel, duration of [h], pitch contour, and intensity.

The ABX task was run on PCIbex. Listeners heard three words, and pushed a key indicating whether the final word (X) was the same as the first (A) or second (B). The full experiment consisted of 132 test trials and 72 control trials. Again, this paper reports only a subset of the results.

2.2. Participants

Listeners were 20 native speakers each of Arabic (Levantine), English (U.S.), and French (France), recruited on Prolific. Most English and French listeners resided in their respective countries, while most Arabic listeners resided in European countries. All listeners reported a single native language, except for two Arabic speakers who reported native multilingualism with English and Hebrew. Participants had intermediate proficiency or higher in other languages (English speakers: French, Japanese, Spanish; French speakers: English, German, and Spanish; Arabic speakers: English, Italian, French, German, Hebrew, Czech, Hungarian). I do not believe this knowledge substantially affected the results (Section 4). Four participants were excluded due to low accuracy on controls (1 Arabic, 2 English, 1 French).

2.3. Hypotheses

Under the perceptual optimization account, listeners should perceive [h] more accurately in CH than in HC sequences. In contrast, if perception is driven by native language experience, listeners should perceive [h] more accurately in contexts where it occurs in their native language (Table 2). Arabic listeners should accurately distinguish [h] in both HC and CH sequences, since Arabic has phonemic /h/ and is phonotactically permissive.[2] French lacks /h/ and aspirated stops; these listeners should have low accuracy on both HC and CH sequences. English has aspirated stops and phonemic /h/, but /h/ is not permitted as a coda. They should perceive [h] in CH better than in HC, and [h] in CH more accurately than French listeners.

Table 2: Properties of listeners' native languages

Language	HC	CH
Arabic	Yes - /h/	No
English	No	Yes - Asp. Stops
French	No	No

3. RESULTS

Linear mixed-effects models (*lme4* [18], in R) predicted Accuracy (Incorrect, Correct) from the

independent variables Language (Arabic, English, French), Condition (HC/C, CH/C), and their interaction. Models also had random intercepts of Participant and Word Set. There was a significant interaction between Language and Condition, so I report results from *emmeans* pairwise comparisons (p-values Tukey adjusted) [19].

The results suggest that [h] is easier to hear before a stop than after it. Figure 1 shows each group's accuracy by condition. Arabic and English listeners had higher accuracy in HC/C than in CH/C, but the difference between conditions was not statistically significant (Table 3). French listeners had significantly higher accuracy in HC/C than in CH/C. Figure 2 compares groups within each condition. In the CH/C condition, the groups all differ significantly in accuracy, in the following order: Arabic > English > French (Table 4a). In the HC/C condition, Arabic listeners' accuracy is higher than French listeners', but Arabic-English and English-French groups do not differ significantly (Table 4b).

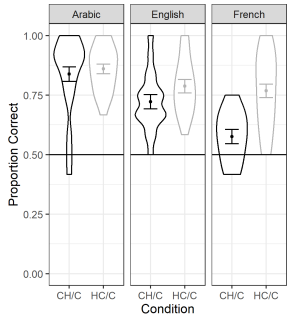


Figure 1: Group x Condition

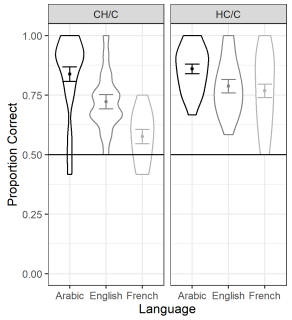


Figure 2: Condition x Group

Table 3: *emmeans* Comparisons within Group

		Est	SE	z.ratio	p
Ar.	HC/C-CH/C	0.17	0.26	0.66	0.51
Eng.	HC/C-CH/C	0.36	0.23	1.58	0.11
Fr.	HC/C-CH/C	0.95	0.21	4.56	0.00***

Table 4: *emmeans* Comparisons within Condition

		Est	SE	z.ratio	p
a. CH/C	Ar. - Eng.	0.70	0.25	2.76	0.02*
	Ar. - Fr.	1.41	0.24	5.78	0.00***
	Eng. - Fr.	0.71	0.22	3.16	0.00***
b. HC/C	Ar. - Eng.	0.52	0.27	1.92	0.13
	Ar. - Fr.	0.63	0.26	2.40	0.04*
	Eng. - Fr.	0.12	0.25	0.47	0.88

4. DISCUSSION

The results provide no evidence that [h] is easier to hear in CH than in HC sequences, and thus no evidence that HC → CH metathesis improves the perceptibility of [h] around voiceless stops. No listener group perceives [h] better in CH than in HC sequences. Instead, perception is driven by native language background. Arabic listeners have high accuracy in both HC/C and CH/C conditions, as expected since Arabic has /h/ and allows it as a coda. Although Arabic does not have CH sequences, /h/ occurs after other consonants, which may explain high accuracy in CH/C [17]. English listeners appear to benefit from experience with aspirated stops: their accuracy on CH/C is higher than French listeners'. However, this experience only puts accuracy in CH/C on par with HC/C. Given that English lacks coda [h], we might expect higher accuracy on CH/C. The fact that English has both /h/ and aspirated stops may have tempered the boost for CH/C. Beyond deciding if [h] was present, these listeners must decide if it is a realization of /h/ or aspiration on a stop. The CH sequences in my stimuli were also onsets to unstressed syllables, a position where English stops are not aspirated.

French listeners' accuracy was unexpectedly high, given that French lacks both HC and CH sequences. The Perceptual Assimilation Model [20] provides possible explanations. French listeners may have mapped (glottal) [h] to French /ʁ/, which occurs as a coda and in onset clusters. While /ʁ/ is typically uvular, there is extensive variation [21]. If [h] were mapped to /ʁ/, discrimination is expected to be good. Another possibility is that [h] was uncategorized, which could lead to good discrimination because listeners may attend to strictly acoustic properties [22].

Mielke [17] also tested the general and language-specific perception of [h]. Listeners heard nonce words recorded by a Turkish speaker and clicked where they heard <h> (if at all) on a word printed on the screen. When they clicked, <h> appeared in that location. He reports sensitivity to [h], not accuracy. His results differ from mine in ways that may be driven by task effects. In his experiment, Arabic listeners were more sensitive to [h] in HC sequences than English and French listeners; my Arabic listeners had higher accuracy than English and French listeners HC sequences, but Arabic and English listeners did not differ from each other. In Mielke's task, English-speaking listeners may not have clicked on orthographic <h> in HC stimuli,

since coda [h] does not exist in English. Another difference is that Mielke's English and French listeners had similar sensitivity in CH sequences; numerically, French listeners were actually *more* sensitive to [h] in CH sequences than English listeners. My study finds the opposite. Mielke suggests that English listeners may have categorized both the intervocalic stops and CH sequences in his stimuli as aspirated stops. In this case, they might not click on <h> since aspiration on stops is not represented orthographically in English. My ABX task removes the interference of orthography that may have affected English listeners' sensitivity to [h] in Mielke's study. A final difference is that Mielke's English and French listeners were more sensitive to [h] in CH than in HC sequences. In my study, English listeners' accuracy is similar in both conditions, and French listeners have lower accuracy CH than on HC sequences. This may be due to phonetic differences in the stimuli.

Recall that Clayton's [10] AX task also found that listeners' ability to discriminate [h] depended on their native language. Pre-stop [h] was not universally difficult. His Gaelic listeners, whose native language has pre- and postaspirated stops, had the highest accuracy on all sequences and perceived CH and HC sequences equally well. His Polish listeners, whose native language does not have coda [h] or aspirated stops, had higher accuracy on HC than on CH sequences, like my French listeners. Similar to my suggestion for French, Clayton proposes that they may have mapped coda [h] onto a native phonemic category. His English listeners were the only ones who had more difficulty with medial HC sequences than with initial CH sequences, which he argues is because only the CH sequences are native-like.

Why, then, are HC sequences rare as preaspirated stops, and 'resolved' by metathesis in the direction HC $\rightarrow$ CH? For preaspirated stops, Clayton argues that they are perceptible, but are difficult to innovate. For both preaspirated stops and metathesis, articulatory pressures, or nuanced perceptual factors not captured in either Clayton's or my study (e.g., listening conditions affecting [h] perception more in some positions, details of coarticulation), may also be at play.

Listeners' knowledge of other languages could have influenced the results, although participants mostly knew languages where the distribution of [h] is *less* permissive than in their native language (e.g., no French listeners knew Arabic or Hebrew, which have coda [h]). However, most Arabic and French-speaking listeners knew English, which could have provided them with experience with aspirated stops. This experience does not seem to have been overly helpful. Arabic listeners actually outperform English listeners on CH sequences, and French listeners have lower accuracy than English listeners. If knowledge of English helps French listeners, the effect is weak.

HC sequences of various types are accompanied by other cues in natural speech, including differences in the duration of the preceding vowel and stop closure (e.g., [23, 2, 24, 10]). This study isolated only the perceptibility of [h] in different positions. The other cues may enhance perceptual contrastiveness of the stops or [h], and future studies on preaspiration and laryngeal metathesis would benefit from studying the interaction between cues.

Another interesting avenue of research would be to analyze the differences in [h] perception in languages where the status of preaspiration differs. In some languages, preaspirated stops are usually considered to be a phonological category (e.g., Icelandic, see [25] for overview). In others, preaspiration is optional and variable (e.g., [23, 26, 27]). In these latter cases, [h] may not need to be as perceptible since it functions as a secondary cue or sociolinguistic variant.

5. CONCLUSION

This study tested the claim that [h] is more difficult to perceive in h-stop sequences than in stop-h sequences, an assumption often used to explain laryngeal metathesis and the rarity of preaspirated stops. A cross-linguistic ABX task provides no evidence for this assumption. Instead, [h] is either equally difficult to perceive in HC and CH sequences (Arabic and English speakers), or harder to perceive in CH sequences (French speakers). Instead of a general difficulty with preconsonantal [h], listeners' ability to perceive [h] depends on the distribution of [h] and aspirated stops in their native language.

6. ACKNOWLEDGMENTS

Thanks to Maria Gouskova, Lisa Davidson, Gillian Gallagher, Juliet Stanton, and Joseph Casillas for helpful feedback. Thanks also to Deniz Özyıldız for recording Turkish stimuli, and to the Department of Language and Literature at NTNU for use of their recording facilities. This work has also benefited from a government grant managed by the ANR under the 'Investissements d'Avenir' program (ANR-10-LABX-0083) (LABEX-EFL).

7. REFERENCES

[1] A. Bladon, "Phonetics for hearers," in *Language for hearers*, G. McGregor, Ed. Great Britain: Pergamon Press, 1986, pp. 1–24.

[2] D. Silverman, "On the rarity of pre-aspirated stops," *Journal of Linguistics*, vol. 39, no. 3, pp. 575–598, 2003.

[3] J. Kingston, "Articulatory binding," in *Papers in laboratory phonology 1: Between the grammar and physics of speech*, J. Kingston and M. E. Beckman, Eds. Cambridge: Cambridge University Press, 1990.

[4] T.-J. Yoon, "Laryngeal metathesis as perceptual optimization," *Studies in Phonetics, Phonology and Morphology*, vol. 18, no. 3, pp. 413–435, 2012.

[5] M. Grammont, *Traité phonétique*. Paris: Librarie Delagrave, 1933.

[6] D. Steriade, "Directional asymmetries in place assimilation: A perceptual account," in *Perception in Phonology*, E. Hume and K. Johnson, Eds. San Diego: Academic Press, 2001, pp. 291–250.

[7] E. Hume, "The indeterminacy/attestation model of metathesis," *Language*, vol. 80, no. 2, pp. 203–237, 2004.

[8] E. Flemming, "Laryngeal metathesis and vowel deletion in Cherokee," in *Cherokee papers from UCLA*, ser. UCLA Occasional Papers in Linguistics. Los Angeles: Department of Linguistics, UCLA, 1996, vol. 16, pp. 23–44.

[9] H. Cho, "Laryngeal feature mobility in Cherokee and Korean," *Language and Linguistics*, vol. 56, pp. 283–306, 2012.

[10] I. Clayton, "On the natural history of preaspirated stops," PhD Dissertation, University of North Carolina Chapel Hill, Chapel Hill, 2010.

[11] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Cambridge, MA: Blackwell Publishers, 1996.

[12] C.-K. Suh, "Aspects of phonetics and phonology of Icelandic preaspiration," *The Phonology-Morphology Circle of Korea*, vol. 6, pp. 63–83, 2001.

[13] D. Steriade, "Phonetics in phonology: The case of laryngeal neutralization," 1997.

[14] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behavior and Development*, vol. 7, no. 1, pp. 49–63, 1984.

[15] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler, "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 6, pp. 1568–1578, 1999.

[16] L. Davidson and J. A. Shaw, "Sources of illusion in consonant cluster perception," *Journal of Phonetics*, vol. 40, no. 2, pp. 234–248, 2012.

[17] J. Mielke, "The interplay of speech perception and phonology: Experimental evidence from Turkish," *Phonetica*, vol. 60, pp. 208–229, 2003.

[18] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[19] R. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means," 2020. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[20] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language experience in second language teaching and learning: In honor of James Emil Flege*, ser. Language Learning & Language Teaching, O.-S. Bohn and M. J. Munro, Eds. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2007, vol. 17, pp. 13–34.

[21] E. R. Webb, "Minimalism and French /R/: Phonological representations in phonetically based phonology," *Journal of French Language Studies*, vol. 19, no. 1, pp. 87–115, 2009.

[22] C. T. Best, G. W. McRoberts, and N. M. Sithole, "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 14, no. 3, pp. 345–360, 1988.

[23] P. Helgason, "Preaspiration in the Nordic Languages," PhD Dissertation, Stockholm University, Stockholm, Sweden, 2002.

[24] P. O'Neill, "S-aspiration and occlusives in Andalusian Spanish: Phonetics or phonology?" *Oxford University Working Papers in Linguistics, Philology & Phonetics*, vol. 12, pp. 73–85, 2009.

[25] L. Heimisdóttir, "The phonology of aspiration in Icelandic: A gesture-based approach," PhD Dissertation, Cornell University, Ithaca, NY, 2015.

[26] M. Stevens and U. Reubold, "Pre-aspiration, quantity, and sound change," *Laboratory Phonology*, vol. 5, no. 4, pp. 455–488, 2014.

[27] M. Hejná, "Pre-aspiration in Welsh English: A case study of Aberystwyth," PhD Dissertation, University of Manchester, Manchester, 2015.

[28] J. Blevins and A. Garrett, "The evolution of metathesis," in *Phonetically based phonology*, B. Hayes, D. Steriade, and R. Kirchner, Eds. New York: Cambridge University Press, 2004, pp. 117–156.

[1] Alternatively, metathesis may be phonetically natural, but not functionally motivated [28].

[2] Arabic does not have /p/. I included words with /p/ for Arabic-speaking listeners because (a) all of my Arabic-speaking listeners had exposure to English, which does have /p/; (b) their accuracy on /p/ was parallel to English and French listeners'. All listener groups had lower accuracy on /p/ than on /t, k/, but Arabic listeners' accuracy was higher than English and French listeners on all three consonants. In a model including Consonant (/p, t, k/), no interaction terms were significant.

SIBILANT PERCEPTION BY MERGED SPEAKERS: THE CASE OF TAIWAN MANDARIN

Sang-Im Lee-Kim & Hsiang-Yu Tung

HIPCS Hanyang University; National Yang Ming Chiao Tung University
sangimleekim@hanyang.ac.kr; loveh89301@gmail.com

ABSTRACT

In many sound changes, the merging of categories in perception often precedes the merger in production, while the reverse, namely merger in production but distinction in perception, has been argued to be rare [1]. The present study examined the perception of the alveolar-retroflex sibilants in Taiwan Mandarin (TM), focusing on the relationship between perception and production by merged speakers. The results of an AXB discrimination task showed that merged speakers were able to discriminate the sibilants above chance level but did so less accurately than non-merged distinct speakers. In an identification task with social guises, merged speakers were again shown to be less sensitive to phonetic cues and more sensitive to social cues. The results indicate that the frequent encounters with distinct forms in the speech community may have helped merged speakers keep the categories separate, to some extent, in their memory.

Keywords: sibilant merger, speech perception, Taiwan Mandarin, social guises

1. INTRODUCTION

In many sound changes, categories are often merged in perception before being merged in production [1]. This tendency translates to the frequent attestation of near-mergers—speakers maintain small but consistent distinctions between categories in production but fail to perceive the small differences in perception [2, 3]. The opposite pattern, namely merger in production but distinction in perception, has also been observed [4, 5, 6, 7, 8, 9]. For example, although speakers of the PIN-PEN merger were outperformed by their distinct counterparts, their identification accuracy was consistently higher than chance level [9]. In the case of the Cantonese tone merger, the perception of merged speakers was shown to remain largely intact in an AX discrimination, despite their slower responses than distinct speakers [10].

One possible reason for the apparent perception-production mismatch in phonemic mergers is that merged speakers are exposed to distinct speech in a speech community in which large variation exists among the speakers [9, 11]. Frequent encounters with the exemplars of those distinct tokens may keep the categories from being completely merged in their mental representation. It has also been proposed that the contrasts maintained in a non-merging context may help facilitate the perception of the distinct categories [8, 9]. For example, /ɪ/ and /ɛ/ are merged before nasals in the PIN-PEN merger, but the vowel contrast is robust in non-pre-nasal contexts. Merged speakers may, therefore, extend their sensitivity to the acoustic cues available in non-merging contexts to the merging context.

The present study examined the perception of the alveolar-retroflex sibilants in Taiwan Mandarin (TM), focusing on the relationship between perception and production by merged speakers. The sibilants in TM appear only in the syllable initial position and are known to be variably implemented, ranging from full merger to clear contrasts (e.g., /s~ʂ/) [12]. This pattern is unconditional as the merger is not limited to particular phonological contexts, and is clearly above the level of social awareness [13]. Two perception tasks (identification and discrimination), were conducted with two goals in mind. First, we aimed to rigorously examine the merged speakers' perception of the contrasts. It was pointed out that a commutation test may be too conservative as a measure to assess the merged speakers' actual perceptual ability [8, 9]. Often used in sociolinguistic research, commutation tests utilize 100% accuracy as the cut-off point; failure in this task is interpreted as the speaker being merged in perception. This measure is likely to underestimate the merged speakers' actual perceptual sensitivity [8], and a more sophisticated method was thus warranted.

Second, we investigated the combined effects of social expectations and one's production characteristics on the perception of the TM sibilants. The impact of socio-indexical information prevails in speech perception [14, 15, 16], and speakers of mergers-in-progress are also sensitive to the social background associated with particular phonetic variants [6]. The TM sibilant merger is socially structured in that it is often attributed to contact with a substratum language, Taiwanese Southern Min (TSM), which lacks the retroflex category in its inventory [17, 18]. We examined whether the model talker's social background, namely expected TSM

fluency, would have a systematic influence on sibilant perception. The identification experiment was, therefore, designed to include socially motivated guises such that a model talker was described as being from either a city where TSM is not widely spoken or a city where TSM is more common.

2. EXP 1: SIBILANT IDENTIFICATION WITH SOCIAL GUISE

2.1. Participants

Seventy-three Taiwanese college students (40 females, 33 males, aged 20-29) participated in both production and perception experiments. A wordlist reading task was performed for all the participants to establish their merger status. The wordlist consisted of four disyllabic words with word-initial sibilants (alveolars /s ts^h/ vs. retroflexes /ʂ tʂ^h/) followed by the vowels /a/; many filler items were also included. The randomized wordlist presented in Chinese characters was repeated five times, and the speech signals were recorded at a sampling rate of 44.1 kHz.

The frication noise was labeled manually in Praat [19] and was submitted to a multitaper spectral analysis in Matlab [20] to obtain spectral mean (M1) and peak values. The mean values of the retroflexes were then subtracted from those of the alveolars for each individual speaker to obtain spectral peak and spectral mean distances, respectively. As shown in Figure 1, the two values were highly correlated (r = .950, p < .0001). Individuals' merger status was further determined through a two-sample t-test: a non-significant difference in the distribution of either spectral means or peaks was taken to indicate that the speaker was merged. Two groups were identified with respect to the merger status: 28 MERGED (11F, 18M) and 44 DISTINCT (29F, 15M) (Table 1). A linear regression model fitted to the spectral distance data in R [21] revealed a significant effect of sex but not TSM-fluency; males were more likely to merge the sibilants than females (p = .0055), but TSM-fluency (p = .8486) and its interaction with sex were not significant predictors of the merger (p = .7234). This is consistent with previous findings [12, 22] that merger is not necessarily triggered by TSM fluency—the pattern has become widespread throughout the speech community in Taiwan.

	F	M	total
MERGED	11	18	29
DISTINCT	29	15	44
total	40	33	73

Table 1. Numbers of participants by merger status and sex.

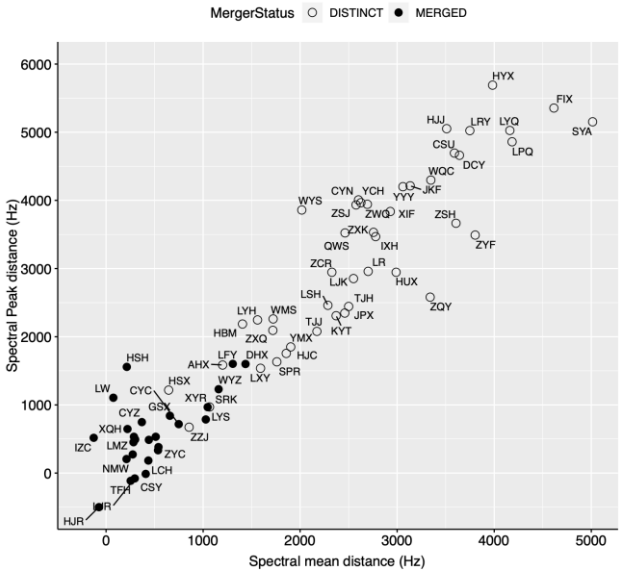


Figure 1. The distribution of spectral peak distance against spectral mean distance by each participant. MERGED speakers are marked with filled circles, and DISTINCT speakers with empty circles.

2.2. Stimuli

For the perception study, a TM male speaker who conveyed clear alveolar-retroflex distinction produced twelve monosyllabic words containing initial sibilants /s ts ts^h/ and /ʂ tʂ tʂ^h/ followed by /a/ and /u/ carried by Tone 1 (X^{55}). These original tokens were digitally manipulated in Praat. The frication noise /s/ and /ʂ/, for example, was first equated for duration and combined with different amplitude proportions to create 8-step continua. To increase acoustic ambiguity, digital manipulation was conducted twice (Figure 2). The vowels were spliced from the original alveolar tokens. Figure 3 presents the spectra of the /sa-ʂa/ and /su-ʂu/ continua. As shown in the figures, the spectral peaks at higher frequencies (8-9 kHz) gradually decreased in amplitude while those at lower frequencies (1-3 kHz) were amplified as the signals changed from most /s/-like to most /ʂ/-like tokens. Note that the acoustic difference of the sibilants from one end to the other was larger for the sibilants before /u/ than before /a/. This was not intended, nor expected, but would likely lead to perceptual consequences, which will be discussed later.

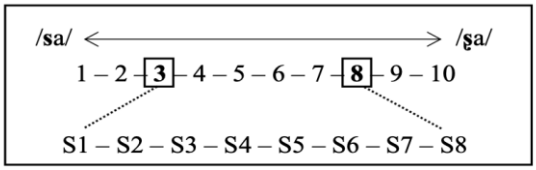


Figure 2. Acoustic manipulation for sibilant continuum

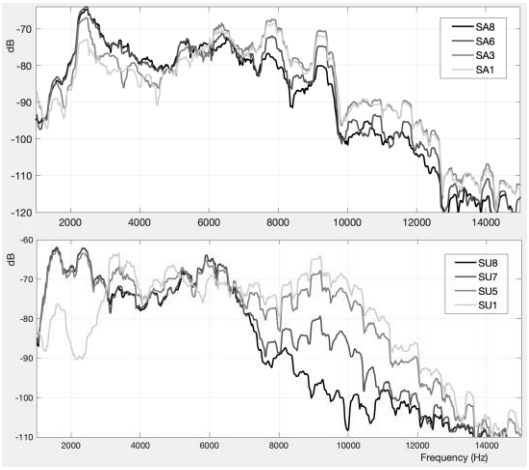


Figure 3. /s/-/ʂ/ spectral continua in the /a/ (top) and /u/ vowel context (bottom).

2.3. Procedure

The perception experiment was conducted in a sound-attenuated booth at National Yang Ming Chiao Tung University in Taiwan. Participants were randomly assigned to one of the social conditions: in the TAIPEI condition, they were told that the talker was originally from Taipei in Northern Taiwan (more likely to speak "standard" distinct variety TM), and in the TAINAN condition, they were told the talker was from Tainan in Southern Taiwan (more likely to speak TSM). To provide further social information, participants watched a video clip of a singer performing a popular song in their respective languages prior to the experiment, and landmarks of the respective cities were presented on the computer screen during the experiment (Figure 4).



Figure 4. Photos of representative buildings in each city: Anping Castle in Tainan (left) and the Taipei 101 in Taipei (right).

In a 2-AFC identification task, participants categorized the target stimuli as alveolar or retroflex by selecting the corresponding letter in the *Zhuyin* phonetic alphabet. There were three blocks in total, each of which consisted of 90 trials containing 48 targets (3 sibilant sets * 8 steps * 2 vowels) and 42 filler items. Fillers included various non-sibilant sounds (e.g., vowels [i] or [u]). The experiment was administered in E-prime 3 [22] and took about 20 minutes to complete.

2.4. Results

The results of the identification task are summarized in Figure 5. The data were analyzed using mixed-effects logistic regression in R. The model revealed three prominent findings. First, in the baseline condition (MERGERSTATUS = MERGED, VOWEL = /a/), merged speakers perceived more retroflexes in the TAIPEI than in the TAINAN condition ($p < .01$), reflecting their implicit bias that a speaker from Tainan would not produce "proper" retroflexes. This was also the case for the distinct speakers, as indicated by the lack of a significant interaction between MERGERSTATUS and SOCIALCONDITION. It is interesting that the speakers' biases were reflected in their perception, even though TSM fluency has lost predictive power on the sibilant merger in production (see §2.1). Second, the merged speakers were less sensitive to the fine-grained acoustic details of the stimuli than the distinct speakers (STEP*MERGERSTATUS: $p < .0001$). As signified by the flatter slopes, merged speakers were reluctant to make either choice even for the stimuli from the two ends, potentially leading to a greater reliance on social cues. Third, the effect of social cues was greatly attenuated in the /u/ vowel context, particularly for the merged speakers (SOCCOND*VOWEL: $p < .0001$). This finding seems to reflect the relative saliency of the acoustic cues in frication noise conditioned by the following vowels. Recall that the acoustic differences in frication noise were larger for the /u/ than for the /a/ vowel context (Figure 3). Cue-trading between social and phonetic cues is evident in this case. Taken together, the results showed that the merged speakers were less sensitive to phonetic cues but more sensitive to social cues.

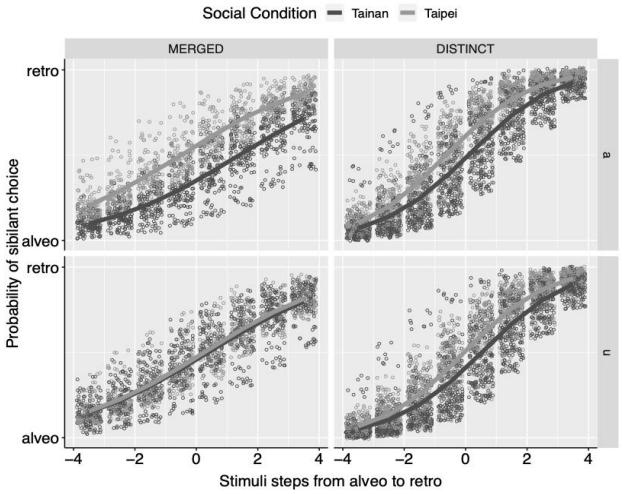


Figure 5. Predicted sibilant choices by the social guise broken down by the speakers' merger status and vowel context.

3. EXP 2: SIBILANT DISCRIMINATION

3.1. Participants

Thirty-five speakers were invited to participate in the discrimination task. All but two participants had participated in the identification task about six months before. Sixteen participants were from the MERGED group (4F, 12M) and nineteen were from the DISTINCT group (16F, 3M).

3.2. Stimuli and procedure

Among the tokens used for the identification task (Figure 2), S1 and S2 tokens were chosen as exemplars of alveolars and S7 and S8 tokens were as exemplars of retroflexes for the discrimination task. The two tokens from one end were similar but not identical acoustically, which was thought to make the experimental task sufficiently difficult. For example, the sequence S1(alveo)-S7(retro)-S8(retro) was created for an AXB trial, in which case the correct answer would be B. 96 tokens (4 pairs * 4 orders * 3 manners of sibilants * 2 vowels) were repeated three times across three blocks. Participants were given 3 seconds to respond in E-prime. The experiments were conducted individually in a sound-attenuated booth.

3.3. Results

The results of the discrimination task are summarized in Figure 6. The accuracy data were analyzed using a mixed-effects logistic regression, and the log response time data were analyzed using a mixed-effects regression in R.

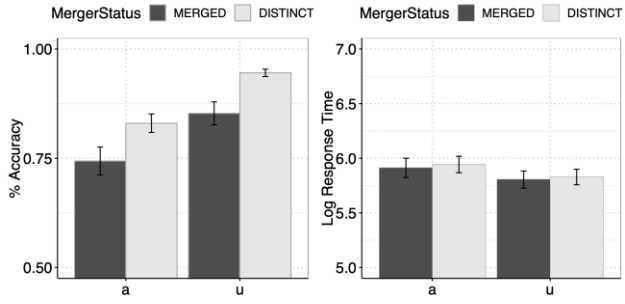


Figure 6. Mean discrimination accuracy (left) and log response time (right) by vowel and the speakers' merger status.

Two major findings were drawn from the model. First, the merged speakers' discrimination of the sibilants (M = 80%) was significantly less accurate than that of the distinct speakers (M = 88%) ($p < .01$), but they performed well above chance level, in line with previous works [4, 5, 6, 7, 8, 9]. Second, both the discrimination accuracy and response time data converged on the vowel effect: the sibilants were discriminated more accurately ($p < .0001$) and faster ($p < .0001$) in the /u/ than in the /a/ context, regardless of the speaker's merger status. Consistent with the results of the identification task, this can be attributed to the larger acoustic differences in the noise properties before /u/ than before /a/.

4. DISCUSSION

This study investigated the perception of the TM alveolar-retroflex sibilants by merged speakers. The results showed that merged speakers were able to discriminate the sibilants above chance level, but did so less accurately than distinct speakers. The TM sibilant merger is unconditional, and merged speakers' moderate perceptual ability was thus likely to arise from their exposure to distinct forms carried by other speakers in the speech community. In addition, the merged speakers' sensitivity to the distinct categories is likely to be further enhanced by the high social awareness of this variant. The sibilant merger is subject to overt comments among TM speakers [13], which may have facilitated the maintenance of the separate exemplars for the two categories. The results of the identification task accompanied by the social guises lend further support for this analysis. Merged as well as distinct speakers in this task were sensitive to the social cues, in which the implicit biases systematically modulated the perception of the sibilants. This result demonstrates the pattern—merger in production and distinction in perception in phonemic mergers—may be more widely attested than previously assumed and should, therefore, be reconsidered as a genuine linguistic pattern of sound change.

The vowel context effects are also worth discussing. In the identification task, merged speakers, in particular, showed a greater weighting of phonetic cues in the /u/ vowel context accompanied by reduced reliance on social cues; in the discrimination study, both groups performed better in the /u/ than in the /a/ vowel context. The acoustic differences in the frication noise were greater in the rounded vowel context (Figure 3), which seems to have reduced the reliance on non-acoustic cues in speech perception. Lip rounding of the following vowel may enhance retroflexion of the retroflex category [24], which could explain the salient acoustic cues in the /u/ context and hence the reduced reliance on other cues.

5. ACKNOWLEDGMENTS

We would like to thank Yun-Chieh Chou, Kuei-Hong Lin, Wei-Cheng Weng for help with data collection and processing. Special thanks also to Yu-an Lu, Li-Hsin Ning, and Meredith Tamminga for their helpful

comments. This work was supported by MOST110-2410-H-A49-058-MY3, Taiwan and start-up research funds from Hanyang University, Korea.

6. REFERENCES

[1] Labov, W. 2011. *Principles of Linguistic Change. Vol. 3: Cognitive and Cultural Factors*. John Wiley & Sons.
[2] Labov, W., Yaeger, M., Steiner, R. 1972. A quantitative study of sound change in progress. In *U.S. Regional Survey*. Philadelphia.
[3] Yu, A. C. L. 2007. Understanding near mergers: the case of morphological tone in Cantonese. *Phonol* 24, 187–214.
[4] Hay, J., Drager, K., Thomas, B. 2013. Using nonsense words to investigate vowel merger. *Eng. Lang. Ling.* 172, 241–269.
[5] Thomas, B., Hay, J. 2005. A pleasant malady: The ellen/allan merger in New Zealand English. *Te Reo* 48.
[6] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *J. Phon.* 344, 458–484.
[7] Baranowski, M. 2013. On the role of social factors in the loss of phonemic distinctions. *Eng. Lang. Ling.* 172, 271–295.
[8] Wade, L. 2017. The role of duration in the perception of vowel merger. *Lab. Phon. 8*(1), 1–34.
[9] Austen, M. 2020. Production and perception of the Pin-Pen merger. *J. Ling. Geog.* 82, 115–126.
[10] Mok, P. P. K., Zuo, D., Wong, P. W. Y. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Lang. Var. Chan.* 25, 341–370.
[11] Maguire, W., Clark, L., Watson, K. 2013. Introduction: what are mergers and can they be reversed? *Eng. Lang. Ling.* 17(2), 229–239.
[12] Lee-Kim, S., Chou, Y. I. 2022. Unmerging the sibilant merger among speakers of Taiwan Mandarin. *Lab. Phon.* 13.
[13] Chung, K. S. 2006. Hypercorrection in Taiwan Mandarin. *J. As. Pac. Comm.* 16(2), 197–214.
[14] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *J. Lang. Soc. Psy.* 181, 62–85.
[15] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Ling.* 48(4), 865–892.
[16] Drager, K. 2010. Sociophonetic variation in speech perception. *Lang. Ling. Comp.* 47, 473–480.
[17] Ing, R. 1984. Issues on the pronunciations of Mandarin. *Wor. Chn. Lang.* 35, 6–16.
[18] Kubler, C. C. 1985. The influence of Southern Min on the Mandarin of Taiwan. *Anth. Ling.* 272, 156–176.
[19] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.0. 37. 2018.
[20] Blacklock, O. S., Shadle, C. H. 2003. Spectral moments and alternative methods of characterizing fricatives. *J. Acoust. Soc. Am.* 113(4), 2199.
[21] R Core Team. 2022. *R: A language and environment for statistical computing*.
[22] Chuang, Y.-Y., Sun, C.-C., Fon, J., Baayen, R. H. 2019. Geographical variation of the merging between dental and retroflex sibilants in Taiwan Mandarin. *Proc. 19th ICPhS.* Melbourne.
[23] Schneider, W., Eschman, A., Zuccolotto, A. 2002. *E-Prime: User's guide*. Psychology Software Incorporated.
[24] Rau, D., Li, J. 1994. Phonological variation of (ts), (tsh), and (s) in Mandarin Chinese. *Proc. 23rd NWAV*, Stanford University.

Early exposure to an Indigenous heritage language and the transfer of perceptual abilities into a foreign language

Stanislav Mulík[1], Mark Amengual[2], Haydée Carrasco-Ortíz[3]

[1]Universidad Nacional Autónoma de México, Mexico;
[2]University of California, Santa Cruz, USA; [3]Universidad Autónoma de Querétaro, Mexico
stanmulik@gmail.com

ABSTRACT

This study investigates a potential facilitatory effect in the perception of the French vowel contrasts /a-ɔ/ and /ɔ-o/ by Spanish-dominant heritage speakers of Hñäñho (HHSs), an understudied and vulnerable Indigenous language in Mexico, as a result of transferring their perceptual abilities of comparable contrasts from their heritage language (Hñäñho) to a language unknown to them (French). In an AXB task, 12 HHSs and 12 monolingual speakers of Mexican Spanish (MSMs) were presented with minimal pairs containing these vowels in Hñäñho and French. Group comparisons showed similar discrimination patterns for these groups in both Hñäñho and French. However, response time data indicate that only the HHSs exhibited significant correlations between their individual perceptual abilities in Hñäñho and French. These results suggest a possible transfer of perceptual abilities from a native, heritage language to an unknown foreign language, even if the heritage language is not actively used by its heritage speakers.

Keywords: Hñäñho, Indigenous heritage language, non-native phonetic learning, early exposure.

1. INTRODUCTION

Heritage speakers (HSs) are early bilinguals who have been exposed to a minority (i.e., heritage) language and the majority language early in life. These HSs may have grown up speaking both languages since birth (i.e., simultaneous bilinguals), may have been brought up in a monolingual setting in early childhood and become bilingual after starting school in the majority language (i.e., sequential bilinguals), or may have had exposure to their heritage language (HL) during their childhood to acquire native-like comprehension of the language but have no active command in it (i.e., passive bilinguals). The latter scenario is especially common in contexts of language shift, like those of speakers of endangered Indigenous languages, in which HL use is usually limited to their nuclear family.

There has been an increased interest in the phonetic abilities of HSs in recent years [9, 26], especially since there is the general assumption that HSs have a benefit in perceiving and producing sounds in their HL as a result of early exposure to the language [4, 10, 19]. However, fewer studies have investigated if this early exposure is advantageous for the perception and/or production of sounds in another language [3, 25]. In other words, does early exposure to an HL facilitate non-native phonetic learning?

The aim of the present study is to examine the potential transfer of perceptual abilities from a native, HL to an unknown foreign language. More specifically, we focus on the perception patterns of Spanish-dominant, Mexican Indigenous HSs in their HL (Hñäñho) and in a foreign language unknown to them (French). We compare their perception in both languages to those of a control group of Spanish-speaking Mexican monolinguals of similar socioeconomic status and age, who have no knowledge of French nor Hñäñho and may therefore only transfer their perceptual abilities from Spanish.

1.1. Heritage speakers of Hñäñho

Santiago Mexquititlán Otomi (Hñäñho) is an Oto-Manguean language variety spoken in Central Mexico. Hñäñho is the first language (L1) or second language (L2) almost exclusively of the Otomi Indigenous people from the rural community of Santiago Mexquititlán in the state of Querétaro.

Even though the term HS has not traditionally been used in Mexico to refer to Indigenous language speakers, the linguistic reality of these speakers very much aligns with the definition of a HS [23]. For instance, Hñäñho is the minority language used in the household, whereas Spanish is the majority language mostly learned either early or later in life as an L2 [20, 16, 24]. The different degrees of Hñäñho-Spanish bilingualism observed in urban areas span from limited knowledge of Hñäñho, changes in language dominance and an increase in Spanish language use, and an eventual language shift from Hñäñho to Spanish in as few as three generations [8]. The relevance of the Ñäñho culture and a connection to the language have been reported to be experienced even by those who have little or no competence in the Hñäñho language.

1.2. Spanish, Hñäñho, and French vowels

The vowel systems of Spanish, Hñäñho, and French are considerably different. Spanish has a simple five-vowel symmetrical system [18, 27]. Along the height dimension, Spanish has two high vowels (/i/ and /u/), two mid vowels (/e/ and /o/), and one low vowel (/a/); and, along the frontness/backness dimension, there are two front vowels (/i/ and /e/), one central vowel (/a/), and two back vowels (/u/ and /o/). In contrast, Hñäñho has a nine-vowel symmetrical system [17, 21, 22] with an additional contrast in height, distinguishing higher-mid vowels /e/ and /o/ from lower-mid vowels /ɛ/ and /ɔ/, and a high-mid central vowel /ə/ also contrasting with the high central vowel /ɨ/. It is also worth noting that the Hñäñho vowel system has one nasal vowel /ɑ̃/. Finally, the vowel system of Standard French consists of up to 13 vowels (/i, y, u, e, ø, ə, o, ɛ, ɛː, œ, ɔ, a, ɑ/) and four nasal vowels (/ɛ̃, œ̃, ɔ̃, ɑ̃/) [13].

Of relevance to the present study, the Hñäñho and French vowel systems both contain mid-vowel contrasts, even though they may not be acoustically identical. Most importantly, however, the acoustic spaces of the Hñäñho and French pairs /e-ɛ/ and /o-ɔ/ are different from the Spanish mid vowels /e/ and /o/. Therefore, it remains to be seen if the perceptual abilities regarding the mid-vowel contrasts in one language (HL) can transfer to an unknown foreign language, for which the combination of Spanish as the dominant language, Hñäñho as L1/HL, and French as a foreign language is an idoneous setting.

2. METHOD

2.1. Participants

A total of 24 adults participated in the study. Twelve were Hñäñho heritage speakers (HHSs) with Mexican Spanish as their dominant language and 12 were Mexican Spanish monolinguals (MSMs). Both groups were proficient speakers of Spanish, but they differed in their mother tongue: HHSs' native language was Hñäñho and/or Spanish, while MSMs' only native language was Spanish. HHSs learned Spanish simultaneously since birth or sequentially during early childhood as an L2; the average age of Spanish acquisition was 5. The two groups were matched in age (mean age 29.9 for HHSs and 29.4 for MSMs) and low socio-economic status.

2.2. Materials

The experimental stimuli consisted of two Hñäñho minimal pairs /do/ *do* 'rock' - /dɔ/ *da̲* 'eye'; /pa/ *pa* 'day' - /pɔ/ *pa̲* 'snake', and one French minimal triplet /ot/ *hôte* 'host' - /ɔt/ *hotte* 'hood' - /at/ *hâte* 'haste'. The stimulus recordings in Hñäñho and French were obtained from three female native speakers of each language in a sound-attenuated booth. The stimuli were recorded using a head-mounted microphone (Shure SM10A) and a solid-state digital recorder (Marantz PMD660), digitized (44 kHz, 16-bit quantization), and computer-edited for subsequent acoustic analysis. All tokens were automatically extracted in *Praat* [7], adjusted to 200ms duration, and normalized in terms of intensity. One acoustic realization per speaker was selected for each Hñäñho and French stimulus and used in the experiment. Table 1 displays the mean vowel formant values for all experimental stimuli.

		F1 (Hz)	**F2** (Hz)	**F3** (Hz)
Hñäñho	/a/	934	1530	2224
	/ɔ/	808	1315	2615
	/o/	626	1078	3021
French	/a/	822	1522	2812
	/ɔ/	589	1172	2602
	/o/	455	819	2598

Table 1: Mean formant values for the vowels used as experimental stimuli.

2.3. Experimental procedure

On the day of the experiment, HHSs filled in the BLP questionnaire [6] and all participants completed the experimental tasks in an individual computerized session. Response times and accurate answers from all tasks were recorded on a Cedrus Response Pad RB740. The AXB task was carried out in a sound-attenuated room, where the participants were comfortably seated in front of a computer screen. Auditory stimuli were delivered by means of earphone inserts and the instructions, as well as other visual aids, were presented at the center of the screen in white font on black background.

There were 24 randomly ordered trials in each AXB categorical discrimination task (Hñäñho and French) consisting of a 500ms fixation cross, followed by the presentation of three auditory Hñäñho/French words (A, X, and B), each pronounced by a different speaker in random order, with a duration of 200ms and 1000ms stimulus onset asynchrony. At the beginning of each task, participants were instructed to pay close attention to the three words and decide whether the vowel in the word they heard second (X) was the same as the vowel in the word they heard first (A) or in the word they heard third (B), and to press the button designated for each scenario. There were four prior AXB training trials on an easy vowel contrast /o-i/ to familiarize the participants with the task.

2.4. Data analyses

Correct answers and corresponding reaction times (RTs) were obtained from all participants for both vowel contrasts in each language. Data from 1 HHS participant were removed, because they failed to follow the instructions. All remaining data were analyzed for outliers. Trials with RTs above 2 standard deviations (SD) from the mean were removed for each participant in each experimental condition (N= 46, 3.3% of the data). Additionally, data from 1 HHS and 1 MSM participant were removed, since their average RTs were above 2SD from the group mean in all 4 experimental conditions. The final dataset consisted of data from 10 HHSs and 11 MSMs.

ANOVAs and Pearson's correlations were used for the statistical analyses of the data, which consisted of accuracy scores and mean RT values per participant and per condition. The factors included in the ANOVAs were participant *Group* (HHSs and MSMs), *Language* (Hñäñho and French), and vowel *Contrast* (/a-ɔ/ and /ɔ-o/).

3. RESULTS

3.1. Group differences

Figures 1 and 2 show the group averages for participants' RTs and accuracy scores in the AXB task, respectively. At the group level, HHSs and MSMs showed similar perceptual patterns for both contrasts in Hñäñho and French, both in terms of RTs and accuracy.

For RTs (Fig. 1), the ANOVA showed no main effect of *Group*, *Language*, or *Contrast* (all *F*s<1). However, there was a significant *Language* Contrast* interaction ($F(1,19)=15.130$, $p<0.001$). Bonferroni-corrected pairwise comparisons showed that participants took longer to identify the /a-ɔ/ contrast than the /ɔ-o/ contrast in Hñäñho, whereas the inverse was true for French (both *p*s<0.05). There were no other significant interactions.

For accuracy scores (Fig. 2), the ANOVA yielded a main effect of *Group* ($F(1,19)=5.791$, $p<0.05$), which did not interact with any other factor. There was a main effect of *Contrast* ($F(1,19)=4.485$, $p<0.05$), but no main effect of *Language* ($F<1$).

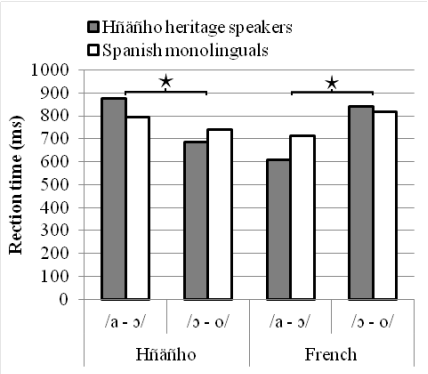


Figure 1: Reaction times in the AXB task, divided by language, vowel contrast, and participant group.

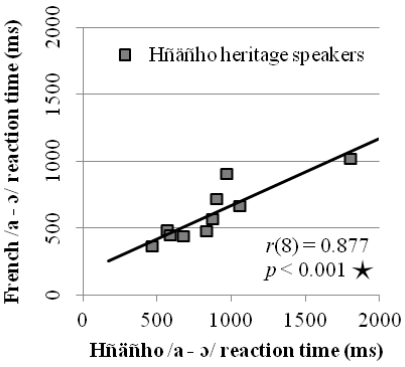


Figure 3: Pearson's correlation between RTs for the /a-ɔ/ contrast in Hñäñho and French for HHSs.

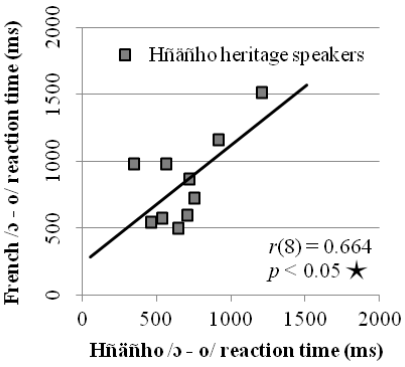


Figure 5: Pearson's correlation between RTs for the /ɔ-o/ contrast in Hñäñho and French for HHSs.

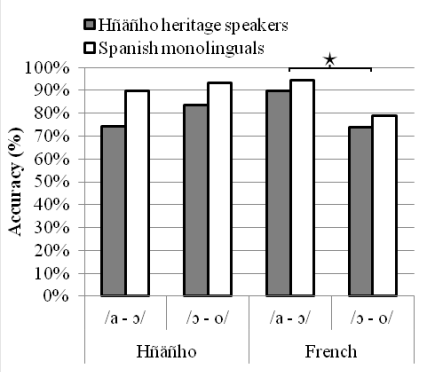


Figure 2: Accuracy scores in the AXB task, divided by language, vowel contrast, and participant group.

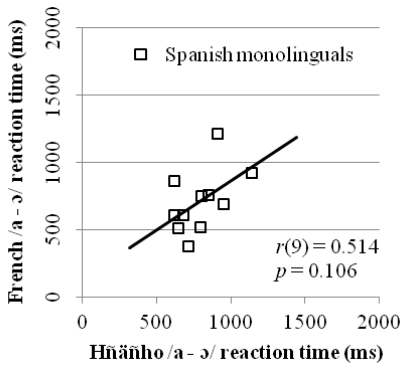


Figure 4: Pearson's correlation between RTs for the /a-ɔ/ contrast in Hñäñho and French for MSMs.

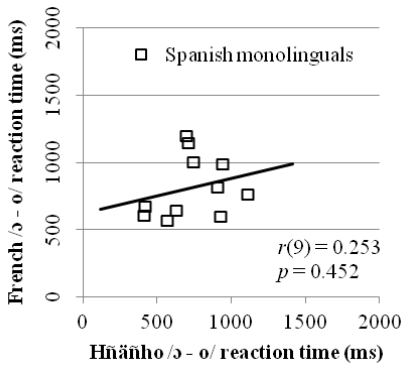


Figure 6: Pearson's correlation between RTs for the /ɔ-o/ contrast in Hñäñho and French for MSMs.

However, there was a significant *Language* Contrast* interaction ($F(1,19)=23.515$, $p<0.001$). Bonferroni-corrected pairwise comparisons showed that participants were more accurate in identifying the /a-ɔ/ contrast than the /ɔ-o/ contrast in French ($p<0.001$), whereas there was an inverse, non-significant trend in Hñäñho ($p=0.072$), with participants being more accurate in identifying the /ɔ-o/ contrast than the /a-ɔ/ contrast. There were no other significant interactions.

In summary, despite MSMs' higher accuracy, the lack of interactions between the participant group and the vowel contrast suggests that there are no group differences between HHSs and MSMs in their patterns of perceptual sensitivity to vowel contrasts /a-ɔ/ and /ɔ-o/ in Hñäñho and French.

3.2. Individual differences

At the individual level, there were no significant correlations between participants' accuracy scores in Hñäñho and French, neither in the HHS group nor in the MSM group, and neither for the /a-ɔ/ nor for the /ɔ-o/ contrast (all *p*s>0.2 except for HHSs' non-significant trend for the /ɔ-o/ contrast, $p=0.063$). However, there were significant correlations between RTs in Hñäñho and French, both for the /a-ɔ/ and for the /ɔ-o/ contrast, but only in the HHS groups and not in the MSM group. These results are shown next in more detail.

Figures 3 and 4 display the relationship between the /a-ɔ/ contrast in Hñäñho and French for the HHS and MSM group, respectively. For this vowel contrast, Pearson's correlation was significant only for the HHS group (Fig. 3) and not for the MSM group (Fig. 4). Similarly, Figures 5 and 6 show the relationship between the /ɔ-o/ contrast in Hñäñho and French for the HHS and MSM group, respectively. For this vowel contrast, too, Pearson's correlation was significant only for the HHS group (Fig. 5) and not for the MSM group (Fig. 6).

4. DISCUSSION AND CONCLUSIONS

The aim of this study was to examine the possible transfer of HHSs' perceptual abilities from their HL (Hñäñho) to a foreign language that was unknown to them (French). HHSs' performance on an AXB categorical discrimination task, using Hñäñho and French vowel contrasts that do not exist in Spanish, was compared to that of a matched group of monolingual Mexican Spanish speakers with no previous exposure to Hñäñho nor French.

The comparison of the speaker groups' RTs and accuracy data revealed similar discrimination patterns for these groups in Hñäñho and French. However, an analysis of the RT data yielded significant correlations between the individual perceptual abilities in Hñäñho and French, but this was only found for the HHSs, and not for the MSMs. These results are interpreted as evidence of a potential link between their perceptual abilities in the HL and in the foreign language. In other words, they may indicate transfer of perceptual abilities from a native, heritage language to a foreign language, even in cases in which the HL is not actively used by the HSs.

Theoretical models of L2 phonetic acquisition, such as the Speech Learning Model (SLM; [14]), the Perceptual Assimilation Model of Second Language Learning (PAM-L2; [5]), and the Second Language Perception Model (L2LP; [12]), have postulated that L2 perception, especially at the initial stages of L2 acquisition, is influenced by L1 perception [5; 15; 11]. These models propose that success in the acquisition of L2 sounds depends on the establishment of new phonetic categories for the L2 segments based on the perceived similarity between the L2 sound and an existing L1 category.

It has been claimed that bilinguals acquire a third language (L3) easier than monolinguals acquire an L2 [1, 25] and, more specifically, bilinguals may acquire certain phonetic features in an L3 easier than monolinguals acquire them in an L2 [2, 3], but phonetic similarities between an unknown foreign language to the native language may also facilitate the perception of universally 'difficult' contrasts [3]. The results of this study suggest that this might also be the case for passive bilinguals (e.g., HHSs) whose HL (Hñäñho) contains similar vowel contrasts as the foreign language (French), but crucially not comparable to the vowel system of their dominant language (Spanish), thus extending the previously reported effects of L1 to heritage bilinguals.

Our results, of course, must be taken with caution because they are based on a relatively small number of participants. Future studies should collect data from a larger participant pool in order to be able to support stronger and generalizable claims about potential bilingual advantages in phonetic learning, and more specifically, the ability of perceiving sounds in an unknown language by HSs. Importantly, these efforts will contribute to a necessary diversification of our field to include Indigenous languages and Indigenous speakers, who have been severely understudied.

5. ACKNOWLEDGMENTS

The authors would like to thank Gloria Avecilla Ramírez for her contribution to the study design, as well as to two anonymous reviewers for their insightful comments.

5. REFERENCES

[1] Abu-Rabia, S., Sanitsky, E. 2010. Advantages of bilinguals over monolinguals in learning a third language. *Bilingual Research Journal, 33*, 173–199.

[2] Amengual, M. 2021. The acoustic realization of language-specific phonological categories despite dynamic cross-linguistic influence in bilingual and trilingual speech. *The Journal of the Acoustical Society of America, 149*(2), 1271-1284.

[3] Antoniou, M., Liang, E., Ettlinger, M., Wong, P. 2015. The bilingual advantage in phonetic learning. *Bilingualism: Language and Cognition*, *18*(4), 683-695.

[4] Au, T.K., Oh, J.S., Knightly, L.M., Jun, S-A., Romo, L. 2008. Salvaging a childhood language. *Journal of Memory and Language, 58*, 998-1011.

[5] Best, C., Tyler, M. D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Munro, M.J., Bohn, O.-S. (eds.), *Language experience in second language speech learning: In honor of James Emil Flege*, 13-34. Amsterdam: John Benjamins.

[6] Birdsong, D., Gertken, L. M., Amengual, M. 2012. Bilingual language profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin*.

[7] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer. http://fon.hum.uva.nl/praat

[8] Canuto Castillo, F. 2015. Otomíes en la ciudad de México. La pérdida de un idioma en tres generaciones. *Lengua y migración 7*(1), 53–81.

[9] Chang, C. B. 2021. Phonetics and phonology of heritage languages. In: Montrul S., Polinsky, M. (eds.), *The Cambridge Handbook of Heritage Languages and Linguistics*, 581-612. Cambridge, UK: Cambridge University Press.

[10] Chang, C.B., Haynes, E., Rhodes, R., Yao, Y. 2008. A tale of two fricatives: consonant contrast in heritage speakers of Mandarin. *University of Pennsylvania Working Papers in Linguistics, 15*, 37-43.

[11] Elvin, J.,Escudero, P. 2019. Cross-linguistic influence in second language speech: implications for learning and teaching. In: Gutierrez-Mangado, M., Martínez-Adrián, M., Gallardo-del-Puerto, F. (eds) *Cross-linguistic influence: from empirical evidence to classroom practice*, pp. 1-20. Springer, Cham.

[12] Escudero, P. 2005. *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. Netherlands Graduate School of Linguistics.

[13] Fagyal, Z., Kibbee, D., Jenkins, F. 2006. *French: A linguistic introduction*. Cambridge University Press.

[14] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In Strange, W. (ed.), *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277. York Press: Timonium, MD.

[15] Flege, J. E., Bohn, O. S. 2021. The revised speech learning model (SLM-r). In: Wayland, R (ed.), *Second language speech learning: Theoretical and empirical progress*, 3-83. Cambridge University Press.

[16] Guerrero Galván, A. 2009. Otho 'bui. Migrantes otomíes en la ciudad de México. *Lengua y Migración 1*(2), 39–56.

[17] Hekking Sloof, E.F.R., Andrés de Jesús, S., de Santiago Quintanar, P., Guerrero Galván, A., Núñez López, R.A. 2010. *HE'MI MPOMUHÑÄ AR HÑÄÑHO AR HÑÄMFO NDÂMAXEI. DICCIONARIO BILINGÜE OTOMÍ-ESPAÑOL DEL ESTADO DE QUERÉTARO.* México: INALI.

[18] Hualde, J. I. 2005. *The sounds of Spanish*. Cambridge University Press.

[19] Knightly, L.M., Jun, S-A., Oh, J.S., Au, T.K. 2003. Production benefits of childhood overhearing. *J. Acoust. Soc. Am., 114*, 465-474.

[20] Martínez Casas, R. 2000. Nuevos espacios para las lenguas y culturas indígenas: los otomíes en Guadalajara. Nueva Antropología. *Revista de Ciencias Sociales, 57*, 43–55.

[21] Mulík, S., Amengual, M., Avecilla-Ramírez, G., & Carrasco-Ortíz, H. 2019. An acoustic description of the vowel system of Santiago Mexquititlán Otomi (Hñäñho). In: Calhoun S., Escudero P., Tabain M., Warren P. (eds.), *Proc 19th ICPhS* Melbourne, Australia 2019, 1377-1381.

[22] Mulík, S., Amengual, M., Avecilla-Ramírez, G, Carrasco-Ortíz, H. 2021. The vowel system of Santiago Mexquititlán Otomi (Hñäñho). *Journal of the International Phonetic Association.* 1-21.

[23] Mulík, S., Amengual, M., Maldonado, R., Carrasco-Ortíz, H. 2021. Hablantes de herencia: ¿una noción aplicable para los indígenas de México? *Estudios de Lingüística Aplicada, 73*, 7-37.

[24] Mulík, S., Corona-Dzul, B., Amengual, M., Carrasco-Ortíz, H. 2021. Perfil psicolingüístico de los bilingües otomí (hñäñho)-español, migrantes de Santiago Mexquititlán a Santiago de Querétaro, México. *Cuadernos de Lingüística de El Colegio de México*, 8, e154, 1-50.

[25] Polinsky, M. 2015. When L1 becomes an L3: Do heritage speakers make better L3 learners? *Bilingualism: Language and Cognition*, *18*(2), 163-178.

[26] Polinsky, M., Scontras, G. 2020. Understanding heritage languages. *Bilingualism: Language and Cognition*, *23*(1), 4-20.

[27] Quilis, A., Esgueva, M. 1983. Realización de los fonemas vocálicos españoles en posición fonética normal. In: Esgueva M., Cantero, M. (eds.). *Estudios de fonética,* vol. 1., 159-252. Madrid: Centro Superior de Investigaciones Científicas.

WORD-LEVEL PROMINENCE AND "STRESS DEAFNESS" IN MALTESE-ENGLISH BILINGUALS

Maria Lialiou[1], Anna Bruggeman[2], Alexandra Vella[3], Sarah Grech[3], Petra Schumacher[1], Martine Grice[1]

[1]University of Cologne, Germany; [2]University of Bielefeld, Germany; [3]University of Malta, Malta
{mlialiou, petra.schumacher, martine.grice}@uni-koeln.de; anna.bruggeman@uni-bielefeld.de;
{alexandra.vella, sarah.grech}@um.edu.mt

ABSTRACT

This study investigates "stress deafness" in bilingual speakers of Maltese and Maltese English. Although both reportedly have lexical stress, the acoustic cues to prominence appear to be relatively weak. Further, word-initial pitch peaks make pitch an unreliable cue to lexical stress, which can be elsewhere in the word. In a sequence recall task, we show that speakers dominant in Maltese exhibit a classic "stress deafness" effect, similar to speakers of French. Speakers who identify as balanced or Maltese English dominant have more diverse results and do not show such a strong tendency towards "stress deafness". These speakers may rely on their exposure to other varieties of English to identify (and recall) word prominences. This study suggests that the nature of stress in Maltese might need to be revisited.

Keywords: stress, prominence, perception, Maltese, Maltese English, sequence recall

1. INTRODUCTION

The difficulty of processing prominence contrasts at a more abstract level under high cognitive load has been the subject of research yielding the somewhat controversial "stress deafness" (SD) phenomenon [e.g. 1-9]. There is consensus that the perceptual (in)sensitivity to prominence contrasts differs as a function of the presence or absence of lexical stress, or some form of word-level prominence, in the native language. Specifically, speakers of languages that lack lexical stress (under other accounts, languages that lack contrastive or have unpredictable lexical stress [4;5]), are expected to perform worse on recalling sequences of stimuli with prosodic prominence contrasts compared to speakers of languages with stress. For example, native speakers of French, Indonesian, and Persian form a cluster of populations exhibiting the SD effect in contrast to Dutch and Japanese listeners who exhibit better performance [6].

This study is concerned with "stress deafness" in Maltese-English bilinguals differing in language dominance. Maltese and English co-exist in the repertoire of most Maltese, in a context shaped by historical events and demographic shifts which have resulted in a rich and complex linguistic landscape [10]. Thus, Maltese/English language contact as well as the prosodic characteristics of the two languages make Maltese and Maltese English (hereafter, MaltE) interesting test cases for the distinction between "stress deafness" and stress sensitivity in speakers/listeners of different languages.

Maltese and MaltE are both reported to have weight-sensitive lexical stress [11;12]. In fact, although MaltE prioritises weight over other factors when assigning stress, there is also a strong tendency to avoid stress falling earlier than on the penultimate syllable, e.g. in compounds, making stress placement more regular than in mainstream varieties.

To date, there has been no full-scale empirical study examining the phonetic cues to lexical stress in Maltese. The few studies on Maltese stress report an inconsistent relationship between stress and vowel duration [13;14]. They also report that neither Maltese nor MaltE reduce the quality of unstressed vowels [12;15;16]. Thus, in both Maltese and MaltE the cues to stress appear to be relatively weak. This does not seem to be compensated for by pitch. Although Maltese and MaltE are intonational languages that have regular pitch accents [17], they commonly exhibit additional initial peaks on unstressed syllables [18-21], making the location of a pitch peak an unreliable cue to stress too. It appears that words in both Maltese and MaltE can have more than one type of prominence: lexical stress prominence and pitch prominence on the initial syllable. The existence of these different types of prominence raises questions about the perceptual status of word prominence for Maltese-English bilingual speakers as well as about their ability to process it at a more abstract level.

The main question we put forward in this study is whether Maltese listeners are sensitive to varying locations of word prominence. Secondly, given the bilingual situation in Malta, we ask whether language dominance (Maltese vs. English) affects listeners' sensitivity and consequently their processing of word prominence at a more abstract level. In the local Maltese context, most bilingual Maltese-English

speakers operate in both languages on a daily basis. Whilst the notion of *balanced* bilingual persists, it is more likely that many of these bilinguals will be more dominant in one of their two languages than in the other [22;23]. The construct of language dominance, although increasingly invoked in the literature, is not easy to define or operationalise [23]. It nevertheless continues to feature as important in research involving bilinguals since there is clear evidence that it impacts on participants' performance in different tasks [24]. A number of studies on Maltese bilinguals [e.g. 25] acknowledge that bilinguals' choices in language use are as important to determining dominance as is proficiency, perceived or actual [23]. To this end, we hypothesize that given the weak and competing cues to word prominence, Maltese listeners will exhibit a SD effect. Yet, we expect the degree of SD to covary with the nature of the individual's bilingual profile, with more English dominant bilinguals possibly being more sensitive to stress-related prominence, in the absence of competing prosodic enhancement, than more Maltese dominant bilinguals. This is related to the likelihood for speakers who make more use of MaltE to have greater exposure to other varieties of English, in which the acoustic cues to lexical stress are stronger.

2. METHODOLOGY

This study is concerned with understanding the perceptual status of word prominence in Maltese using the Sequence Recall Task (SRT) developed by [1]. In this task participants are required to learn to associate two words representing a phonetic difference (segmental or prosodic) with a specific key on the keyboard. Once the association is established, participants are presented with sequences of varying lengths (in past studies ranging from 2 to 6) made up of contrasting words and asked to reproduce the order of the words in the sequence by pressing the relevant keys. To allow for a direct comparison with SD (e.g. Indonesian, Persian and French) and non-SD (e.g. Dutch, Japanese) reference group listeners, we replicate the design of [6] using the same experimental setup including the exact acoustic stimuli and experimental details such as inter-stimulus interval and sequence orders. The only change we made was implementing it to run in OpenSesame [26].

2.1. Experiment

2.1.1. Stimuli

Two sets of minimal pairs consisting of pseudowords are used as stimuli in the task: the segmental pair /ˈmuku/~/ˈmunu/ which tests the ability to discern a segmental contrast, and the prosodic pair /ˈnumi/~/nuˈmi/ which tests a prominence contrast. The minimal pairs are produced by 2 Dutch and 2 Persian speakers (1 male, 1 female in each language group). The realisation of the prosodic minimal pair varies in terms of phonetic cues (duration/pitch/vowel quality for the Dutch stimuli, and pitch only for the Persian stimuli, cf. [6]). Each speaker produced three instantiations per pseudoword and all tokens were time-compressed to 450ms for comparable duration.

2.1.2. Experimental procedure

The recall tasks (segmental/prosodic) were presented in a counterbalanced order across participants. Each task consisted of 4 parts: word presentation, word identification, practice SRT (2 words), main SRT (3 to 5 words) (for additional information see [6]). At the end of each sequence in the main SRT, the word "OK" was played to flush acoustic memory. The response key for /ˈmuku/ and /ˈnumi/" was "1", for /ˈmunu/ and /nuˈmi/ it was "2". In the main SRT 5 different patterns were presented per sequence length (e.g. for length 3: 122, 121, 122, 211, 212), yielding a total of 15 test items (3 lengths x 5 items), repeated twice, once for each stimulus language. The resulting 30 test items were presented in random order, and speaker (male/female) and acoustic token combinations were determined randomly. In total this yielded 60 items per experiment (30 for each SRT, segmental/prosodic). The entire experiment lasted ca. 35 minutes.

2.1.3. Participants and exclusion criteria

64 Maltese native speakers (age range: 18-28; 29-40; 41-60; > 61, with the majority, 47 out of 64, falling into the 18-28 range) were recruited for this study. Participants were divided into two groups according to their language use choices [23], specifically, as to their choice of language when speaking to close friends and family. The two resulting groups were a Maltese dominant (N=32) and a balanced/MaltE dominant group (N=32). 9 participants were excluded from the analysis, 4 due to complex linguistic background (e.g. exposure to multiple other languages at home), and 5 due to technical issues. We additionally decided on a cut-off minimum performance in the segmental SRT in order to exclude participants who apparently failed to concentrate on the task: we excluded 4 participants who scored correctly on fewer than 10 (i.e. 33%) of the sequences in the segmental task. Following this criterion for the joint analysis with data from [6] (see section 3.2), we also excluded 1 French participant reported on in [6] as scoring < 10. After the above exclusions, we are

able to report on results from a total of 25 Maltese dominant and 26 MaltE dominant participants.

2.2. Analysis and statistics

The response variable is binary i.e. is a correct response given for an item. We ran binary logistic regression models with *glmer()* [27] in R [28] with main effects for: TASK (segmental/prosodic), LENGTH (3/4/5), SPEAKER (Dutch-f/Dutch-m/Persian-f/Persian-m), and listener GROUP (Maltese/MaltE dominant, M and ME below). Scripts and data are available at https://osf.io/62nc3/ on the OSF platform.

We ran one model on the full dataset from both SRTs, which included interactions for TASK with SPEAKER and TASK with GROUP, and an interaction between SPEAKER and GROUP. The random effects structure included an intercept for participant and a random slope for by-task variation in performance across participants. Following the observation that there were no GROUP differences on the segmental SRT, we continued the analysis with a model that exclusively ran on the prosodic SRT (dropping the main effect of TASK and its possible interactions), allowing the inclusion of by-listener random slopes for SPEAKER. Estimates and SEs were generated with the package *emmeans* [29], with Tukey correction for multiple comparisons. *p*-values were generated through LRTs between full and null models for the predictors tested. Predicted probabilities and 95% CIs were also from *emmeans*.

3. RESULTS

3.1 Maltese versus MaltE dominant speakers

Mean model predictions (back transformed from the logit scale), for the segmental contrast, were the following: sequences of length3 had a probability of a correct answer of 0.87 (ME) / 0.83 (M), dropping to 0.60 (ME) / 0.52 (M) for sequence length5. For the prosodic contrast, the mean probabilities of a correct response for sequences of length3 and length5 were 0.69 (ME) / 0.60 (M) and 0.33 (ME) / 0.25 (M), respectively. The model showed a main effect of LENGTH ($\chi^2(2) = 224.97$, $p<0.0001$, and all pairwise significant at $p<0.0001$: length3-logit = 1.17, SE = 0.13; length4-logit = 0.49, SE = 0.12; length5-logit = −0.34, SE = 0.12). There was also a main effect of TASK with higher scores on the segmental task (prosodic-logit = −0.12, SE = 0.11; segmental-logit = 1.01, SE = 0.12, $p<0.0001$), but no interaction between GROUP and SPEAKER, nor main effects.

For subsequent analysis we ran models on the prosodic SRT separately. Predicted mean correct scores and 95% CIs on the prosodic task are shown in Figure 1, with separate estimates for the different stimulus speakers. Performance on the prosodic task again showed an effect of sequence LENGTH ($\chi^2(2) = 93.21$, $p<0.0001$, all pairwise significant at $p<0.01$: length3-logit = 0.49, SE = 0.20; length4-logit = −0.02, SE = 0.20; length5-logit = −0.94; SE = 0.20). There were no group differences but there was a SPEAKER effect: the Dutch female speaker's stimuli yielded more incorrect responses than the Persian male speaker's ($p<0.05$: Dutch-f-logit = −0.45, SE = 0.21; Persian-m-logit = 0.12, SE = 0.20).

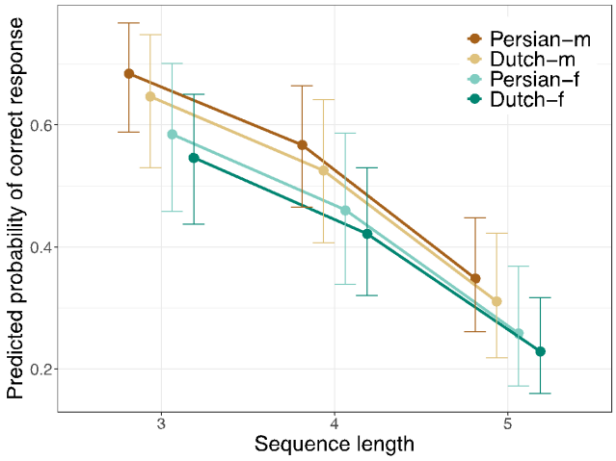


Figure 1: Predicted probability and 95% CIs for a correct response on the prosodic SRT, pooled across listener groups and separated by stimulus speaker (Persian/Dutch, m/f).

3.2. Comparison with results from [6]

For the purpose of a direct comparison with the SD and non-SD groups in [6], we ran a model on the joint prosodic data from the present experiment, and the original raw data from [6]. Model specification is identical to the prosodic model from 2.2, with the exception of the change from stimulus SPEAKER to stimulus LANGUAGE (speaker is not available for the data from [6]). Predicted probability of a correct response is shown in Figure 2.

The scores of the Maltese dominant group are similar to those of the SD groups from [6] (Indonesian, French, Persian), while the MaltE dominant group performs better. Model comparisons on the GROUP level confirm that: i) the known "non-stress deaf" Dutch and Japanese groups perform significantly better (all comparisons are significant at $p<0.05$) than the known "stress deaf" groups Indonesian, French, and Persian (confirming the analysis from [6] with a different statistical implementation); ii) the Maltese dominant group performs significantly worse than the "non-stress deaf" groups, but not different from the known "stress deaf" groups; iii) the MaltE group does not differ significantly from any other group, including the Maltese dominant one.

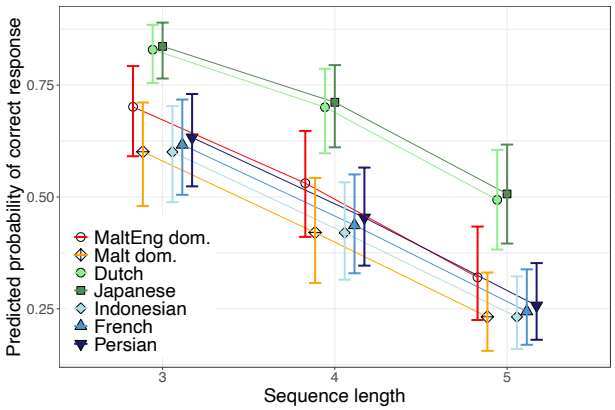


Figure 2: Predicted probability and 95% CIs for a correct response on the prosodic SRT for the combined Maltese data and [6], per participant group, averaged across stimulus language.

4. DISCUSSION

4.1. A "stress deafness" effect

The above results provide clear evidence that native Maltese speakers who are dominant in Maltese pattern with the known SD groups. In contrast, native Maltese speakers who are balanced users of Maltese and MaltE or dominant in the latter, exhibit more varied behaviour. Whereas these speakers behave no differently from Maltese dominant speakers (section 3.1), in a larger comparison with known SD and non-SD groups (section 3.2), they additionally perform neither better nor worse than any other group. The result is that they straddle the SD vs. non-SD dichotomy that exists between the other groups.

In contrast to most previous studies, we did not aggregate response scores across participants prior to statistical analysis, yielding more insight into within-group variation in performance. The large overlap in CIs between the MaltE and SD groups suggests that the MaltE group is tendentially more SD-like, although this sample did not perform significantly worse than non-SD groups. Future work is planned to seek the individual-level factors that yield better performance and model Maltese/MaltE dominance as a continuous factor to investigate whether there is a degree of English dominance that serves as a threshold for "non-stress deaf" performance.

4.2. Acoustic properties of stimuli

In section 3.1. we discussed a SPEAKER effect: the stimuli spoken by the Dutch female speaker yielded lower performance than those from the Persian male speaker (the speaker yielding the highest correct scores). In an earlier replication of [6], [9] reported a similar effect for Tashlhiyt Berber and Moroccan Arabic participants. The stimulus effect was explained with reference to the phonetic details of the stimuli spoken by the Dutch-f speaker, which diverged considerably from the productions of the other three speakers. Firstly, no durational difference is found between the initial syllables of the two pseudowords as a function of stress (i.e. stress in these tokens is not actually cued by durational differences). Secondly, the fundamental frequency (f0) patterns for all tokens of this speaker's renderings of /nuˈmi/ terminate high as opposed to low (as in the tokens produced by the other speakers). These diverging realisations could potentially result in conflicting prominence cues, affecting participants' performance by hindering a straightforward identification of one of the two syllables as the most prominent one. Although this is a limitation of the study, which we were aware of from the outset [cf. 9], we retained the stimuli from the original study [6] for the purposes of direct comparison across studies [6; 9].

5. CONCLUSION

With a sequence recall task replicating the design of [6] we showed that Maltese speakers, if they are dominant in Maltese, exhibit a classic "stress deafness" effect. Maltese listeners who are balanced bilinguals or dominant in English, in contrast, do not exhibit the effect to the same extent, as their performance falls between that of known "stress deaf" and "non-stress deaf" groups.

Maltese dominant speakers do not appear to process word-level prominence in the same way as speakers of languages with Germanic-type stress or Japanese pitch accent. We argue that this can be explained with reference to the prosody of Maltese and MaltE, in which lexical stress and prosodic enhancement are not straightforwardly linked. First, weak cues to stress in their native language might prevent them from being fully attuned to the prosodic contrast tested here. Second, they can only make limited use of pitch prominence, which is not exclusively reserved to mark the syntagmatic contrast of lexical stress, since it can often occur on an initial (unstressed) syllable.

The situation might be somewhat different for speakers identifying as Maltese-English dominant or *balanced*. Their performance in recalling stress patterns might be enhanced through greater exposure to mainstream varieties of English, which often lack pitch prominence on unstressed syllables and instead overwhelmingly place pitch prominence on, or in the vicinity of, lexical stress. This could lead them to more easily relate the location of prosodic prominence to that of lexical stress, allowing them to recall the prosodic contrast as an unequivocal categorical contrast.

6. ACKNOWLEDGMENTS

First, we would like to thank Nadia Pelageina for the great help with recruiting participants and running the experiment in Malta. We would also like to thank Hamed Rahmani and colleagues for making publicly available the stimuli and raw participant results. Lastly, we thank the Collaborative Research Center 1252 Prominence in Language, project A01, funded by the German Research Council, for supporting this work.

7. REFERENCES

[1] E. Dupoux, S. Peperkamp, and N. Sebastián-Gallés, "A robust method to study stress "deafness"," *J. Acoust. Soc. Am.*, vol. 110, pp. 1606–1618, 2001.

[2] E. Dupoux, C. Pallier, N. Sebastián-Gallés, and J. Mehler, "A destressing "deafness" in French?" *Journal of Memory and Language*, vol. 36, pp. 406–421, 1997.

[3] E. Dupoux, and S. Peperkamp, "Fossil markers of language development: phonological 'deafnesses' in adult speech processing," in *Phonetics, phonology and cognition*, J. Durand, and B. Laks, Eds. Oxford University Press, 2002, pp. 168–190.

[4] S. Peperkamp, and E. Dupoux, "A typological study of stress 'deafness'," in *Laboratory Phonology*, C. Gussenhoven, and N. Warner, Eds. vol. 7, De Gruyter, 2002, pp. 203–240.

[5] S. Peperkamp, I. Vendelin, and E. Dupoux,"Perception of predictable stress: A cross-linguistic investigation," *Journal of Phonetics*, vol. 38, pp. 422–430, 2010.

[6] H. Rahmani, T. Rietveld, and C. Gussenhoven, "Stress 'deafness' reveals lexical marking of stress and tone in the adult grammar," *PLoS ONE*, vol. 10, pp. , 2015.

[7] S. Correia, J. Butler, M. Vigário, and S. Frota, "A stress "deafness" effect in European Portuguese," *Language and Speech*, vol. 58, pp. 48–67, 2015.

[8] A. Kimball, H. Keupdjio, K. Franich, and C. Kouankem, "Expanding field studies using online speech perception experiments," in *Proc. 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019, pp. 315-319.

[9] A. Bruggeman, "Lexical and postlexical prominence in Tashlhiyt Berber and Moroccan Arabic," Universität zu Köln, Doctoral Dissertation, 2020.

[10] A. Vella, "Language and language varieties in Malta," *International Journal of Bilingual Education and Bilingualism*, vol. 16, pp. 532–552, 2013.

[11] A. Vella, "Maltese intonation and focus structure," in *Maltese linguistics: A snapshot in memory of Joseph A. Cremona (1922-2003) [Il-lingwa tagħna]*, vol. 1, R. Fabri, Ed. 2009, pp. 63-92.

[12] A. Vella, "Prosodic structure and intonation in Maltese and its influence on Maltese English," University of Edinburgh, Doctoral dissertation, 1995.

[13] M. Azzopardi, "The phonetics of Maltese: Some Areas Relevant to the Deaf," University of Edinburgh, Doctoral Dissertation, 1981.

[14] E. Hume, J. Venditti, A. Vella, and S. Gett, "Vowel duration and Maltese 'għ'," in *Introducing Maltese Linguistics: Selected Papers from the 1st International Conference on Maltese Linguistics*, B. Comrie, Ed. John Benjamins, 2009, pp. 15-46.

[15] S. Grech, "Variation in English: Perceptions and patterns in Maltese English," University of Malta, Doctoral dissertation, 2015.

[16] S. Grech, and A. Vella, "Rhythm in Maltese English," *The languages of Malta*, vol. 18, pp. 203-223, 2018.

[17] S.-A. Jun, "Prosodic typology: by prominence type, word prosody, and macro-rhythm," in *Prosodic Typology II: The phonology of intonation and phrasing*, 2nd ed., S.-A. Jun, Ed. Oxford University Press, 2014, pp. 520-539

[18] A. Vella, "The phonetics and phonology of wh-question intonation in Maltese," in *Proc. 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 1285-1288.

[19] A. Vella, "Alignment of the 'early' HL sequence in Maltese falling tune wh-questions," in *Proc. 17th International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 2062-2065.

[20] M. Grice, A. Vella, and A. Bruggeman, "Stress, pitch accent, and beyond: Intonation in Maltese questions," *Journal of Phonetics*, vol. 76, 100913, 2019.

[21] M. Grice, J. S. German, and P. Warren, "Intonation systems across varieties of English", in Oxford Handbook of Language Prosody, C. Gussenhoven, and A. Chen, Eds. Oxford University Press, pp. 285-302.

[22] S. Romaine, *Bilingualism*, Oxford: Blackwell Publishers, 1995.

[23] J. Treffers-Daller, "What defines language dominance in bilinguals," *Annual Review of Linguistics*, vol. 5, pp. 375-393, 2019

[24] E. Argyri, and A. Sorace, "Crosslinguistic influence and language dominance in older bilingual children," *Bilingualism: Language and cognition*, vol. 10, pp. 79-99, 2007

[25] S. Caruana, and L. Mori, "Rethinking Maltese English as a continuum of sociolinguistic continua through evaluations of written and oral prompts," *English World-Wide: A journal of varieties of English*, vol. 42, pp. 245-272, 2021.

[26] S. Mathôt, D. Schreij, and J. Theeuwes, "OpenSesame: An open-source, graphical experiment builder for the social sciences," *Behavior Research Methods*, vol. 44, pp. 314-324, 2021.

[27] D. Bates, M. Maechler, B. Bolker, and S. Walker. *Lme4: Linear mixed- effects models using Eigen and S4*, R package. (2015). [online]. Available: https://cran.r-project.org/web/packages/lme4/lme4.pdf

[28] R Core Team, *R: A language and environment for statistical computing*. (2021). [online]. Available: https://www.R-project.org/

[29] R. Lenth, V. Russel, P. Buerkner, M. Herve, J. Love, F. Miguez, H. Riebl, and H. Singmann. Emmeans: *Estimated Marginal Means, aka Least-Squares Means*, R package. (2022). [online]. Available: https://cran.r-project.org/web/packages/emmeans/index.html

PERCEPTUAL ADAPTATION TO ALTERED CUE INFORMATIVENESS: DISTRIBUTIONAL, AUDITORY, AND LEXICAL FACTORS

Jiayin Gao[1,2,a)], James Kirby[2,b)]

[1]LPP (UMR7018, CNRS/Sorbonne Nouvelle),
[2]Institute for Phonetics and Speech Processing (IPS), LMU Munich
a) jiayin.gao@cnrs.fr, b) jkirby@phonetik.uni-muenchen.de

ABSTRACT

This study examines French-speaking listeners' perceptual adaptation to altered informativeness of acoustic cues signalling the voicing contrast. We attempt to condition a shift of perceptual attention from a primary cue to a secondary cue. Listeners were exposed to stimuli in which the distributional characteristics of either the primary or secondary cue were altered: either informativeness of F0 was increased, or that of VOT was decreased. When lexical feedback was given, increased informativeness of F0 led to listeners' increased use of high F0 in identifying voiceless plosives, whereas decreased informativeness of VOT led to increased use of low F0 in identifying voiced plosives. We suggest this asymmetry is due to general auditory mechanisms in the perception of F0 and VOT. Our findings highlight the importance of non-distributional characteristics such as auditory mechanisms and lexical knowledge in cue (re-)weighting, shedding light on how a secondary cue is phonologised in sound change.

Keywords: perceptual learning; perceptual adaptation; cue weighting; voicing; French

1. INTRODUCTION

Phonetic categories are signalled by multiple cues. For example, VOT (voice onset time) and F0 (fundamental frequency) (among other acoustic dimensions) signal the voicing contrast of plosives in many languages, with larger/positive VOTs associated with a higher F0 at the vowel onset (or onset F0) and smaller/negative VOTs with a lower onset F0 (e.g., [1]). But the relative weight of these cues differs. When the informativeness of one acoustic dimension is greater, listeners give a greater perceptual weight to this dimension [2, 3].

Following [4, 5, 6, 7], we define "informativeness" in distributional and probabilistic terms. When the distributions of two phonetic categories are more separated along one acoustic dimension, this dimension has greater informativeness than when their distributions are more overlapped (all else being equal). In languages like French or English, the VOT distributions of the voiced and voiceless categories are more separated than the F0 distributions. The VOT dimension is thus more informative, and listeners typically give a greater perceptual weight to VOT than onset F0 in categorising voicing [8, 9, 10].

Cue weighting is also dynamic. When the informativeness of a cue is increased, listeners rapidly shift their attention to this cue through exposure [4, 6, 11], a process known as "perceptual adaptation". But how is listeners' perception affected by a *decrease* in the informativeness of the primary cue? This question is important because in real speech, the primary cue cannot always be reliably recovered, which, in the long term, could lead to a cue re-weighting process. The current study tackles this question by comparing speech perception in two experimental conditions: (1) in which the informativeness of a secondary cue is increased, and (2) in which the informativeness of the primary cue is decreased. The rationale for the second condition is that although manipulation of the primary cue does not modify the informativeness of secondary cues *per se*, it impacts their *relative* informativeness. Our goal is to understand the conditions under which listeners may come to upweight a secondary cue.

We explore this issue in the context of the French voicing contrast. In Experiment 1, we examine the effect of increased informativeness of F0 (a secondary cue) on categorisation of voiced and voiceless plosives. In Experiment 2, we examine the effect of decreased informativeness of VOT (a primary cue) on the same categorisation task.

We also examine the influence of lexical feedback. Although some studies (e.g., [6, 12]) have shown that changes to distributional informativeness alone can lead to perceptual adaptation, others have found that perceptual learning is more robust when lexical information is provided [13, 14, 15, 16]. We therefore probe for a potential effect of lexical feedback by displaying or not the appropriate image.

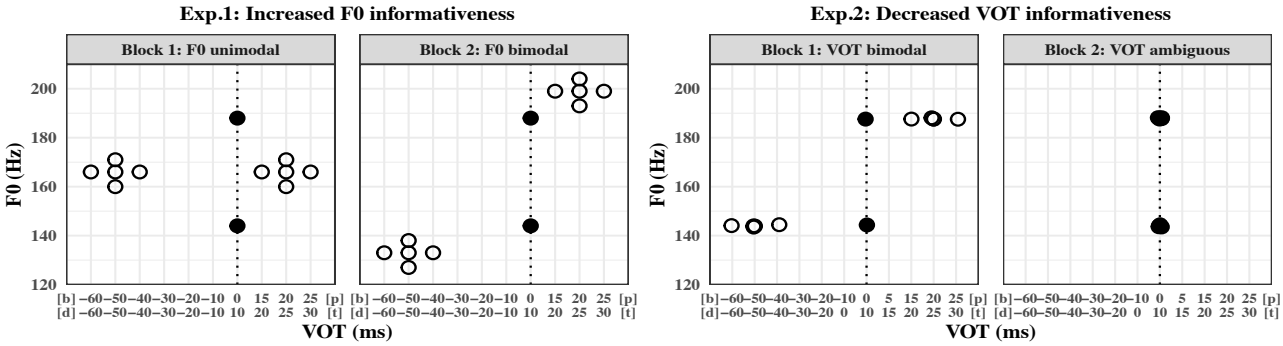


Figure 1: Stimuli of Experiments 1 (left) and 2 (right): each dot for a VOT-F0 association. Black dots are the test stimuli located at the ambiguous VOT step (dotted line). White dots are the exposure stimuli. For Experiment 2, jittered dots at -50, 20/25 ms (Block 1) and 0 ms (Block 2) represent higher numbers of stimulus trials.

2. EXPERIMENTS

2.1. Participants

120 native French-speaking listeners, recruited and paid through Prolific, participated remotely in two two-alternative-forced-choice (2AFC) experiments via PsyToolkit [17, 18]. 60 participated in Experiment 1 and the other 60 in Experiment 2. They were instructed to use earbuds or headphones and to perform the experiment in a quiet room. We discarded the data from one listener due to a high error rate (10 out of 30) in the training session.

2.2. Stimulus creation and protocol

A stimulus pool was created based on two minimal pairs, *pain* 'bread' vs. *bain* 'bath' and *thym* 'thyme' vs. *daim* 'deer', produced by a female native speaker of French. All four words contain the same rime [ɛ̃]. Continua along three parameters were created using re-synthesis. First, voice lead was manipulated from −60 to 0 ms with a step of 10 ms by reducing the original voice lead in *bain* or *daim* using cosine-function squared smoothing [19]. Second, voice lag was manipulated from 5 to 30 ms with a step of 5 ms by modifying the duration of the release within the original *pain* or *thym* using TD-PSOLA [20]. Third, onset F0 was manipulated from 127 to 204 Hz with a step of 5.5 Hz, again using TD-PSOLA.

For each experiment, a subset of stimuli were selected to form the distributional configurations of VOT (involving voice lead and voice lag) and onset F0 (Fig. 1), as detailed below. Participants performed the 2AFC task using their own personal computer. On each trial, the listener was presented with an auditory stimulus drawn either from the *pain–bain* pair or the *thym–daim* pair, and invited to choose the word they heard by clicking on one of the two pictures presented on the screen. As detailed below, the number of trials used to create an effect of exposure was decided following [6].

2.3. Experiment 1: Increased F0 informativeness

2.3.1. Procedure

Experiment 1 examined listeners' perceptual use of F0 when the informativeness of F0 was increased. In two blocks, listeners were exposed to a series of randomly ordered stimuli that fell in a given distributional configuration of VOT and F0, schematised in Fig. 1 (left), where each dot represents an association between VOT and onset F0 for a given stimulus. Block 1 was followed by Block 2.

In Block 1 (F0 unimodal), the distributions of voiced and voiceless categories were well separated along the VOT dimension but unimodal along the F0 dimension. That is, only VOT, but not F0, was informative. In Block 2 (F0 bimodal), the distributions of the two categories were equally well separated along the VOT dimension as in Block 1, but also separated along the F0 dimension. The F0 distance between the two categories was 65 Hz, above the post-obstruent F0 difference canonically observed in French. That is, the informativeness of VOT was unchanged, but that of F0 was increased.

The exposure stimuli, represented by the empty dots, fell within the distributional configurations described above. The two blocks also contained stimuli that were ambiguous for voicing identification, determined by the results of a baseline experiment conducted with a different listener group. They were the test stimuli, represented by the black dots. The task of the listener was the same for both exposure and test stimuli, but only their responses to the test stimuli were measured and compared across blocks.

In order to examine the effect of lexical feedback

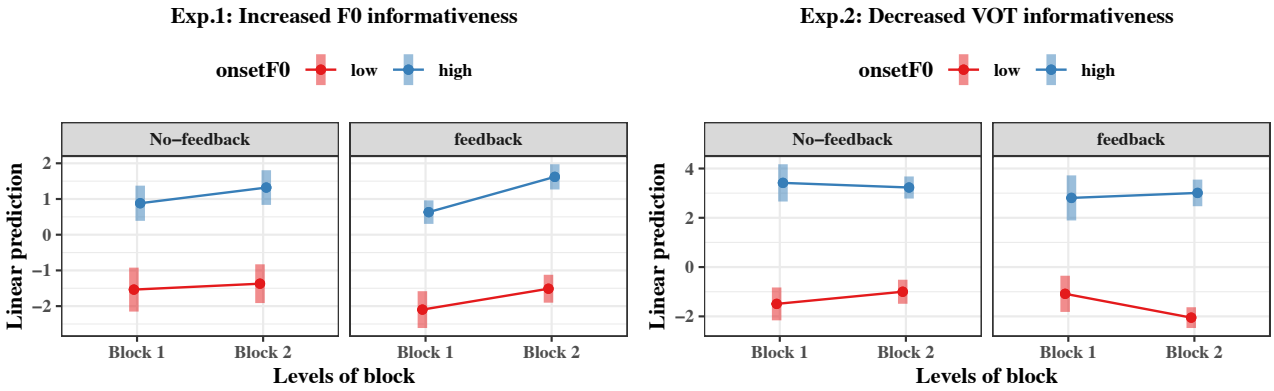


Figure 2: Interaction plots: voiceless response by F0, block, and feedback in the two experiments. Bars: 95% CIs.

associated with the F0 informativeness on listeners' performance, half of the listeners received feedback in Block 2 and the other half received no feedback. Feedback was provided by displaying the appropriate picture in the middle of the screen for 1.5 sec once the listener responded. Regardless of VOT, the pictures representing *bain* and *daim* were given as feedback to the auditory stimuli with a low onset F0 (from 127 to 144 Hz), and the pictures representing *pain* and *thym* were given as feedback to the auditory stimuli with a high onset F0 (from 188 to 204 Hz).

The no-feedback version contained 416 trials in total (12 exposure and test stimuli × 2 word pairs × 8 repetitions × 2 blocks + 32 attention checks). The feedback version contained 312 trials in total (12 exposure and test stimuli × 2 word pairs × 6 repetitions × 2 blocks + 24 attention checks).

2.3.2. Prediction

If an increase in the informativeness of onset F0 in distributional terms *alone* prompts listeners to upweight their use of onset F0 in voicing categorisation, we predict an increase in the use of F0 to the test stimuli in Block 2 compared to Block 1 (i.e., a greater use of high onset F0 in identifying voiceless plosives and/or a greater use of low onset F0 in identifying voiced plosives). If lexical feedback plays a role in addition to the distributional characteristics, we predict that such an increase will occur only or especially when lexical feedback is given.

2.3.3. Analysis and results

A generalised linear mixed model (GLMM), using the `lmerTest` package [21] in R [22], was fitted to listeners' response data to the test stimuli (i.e., the black dots in Fig. 1). The Helmert-coded predictors *Block* (1 vs. 2), *onsetF0* (high vs. low), *feedback* (no-feedback vs. feedback) and their three-way interaction were included. Random intercepts were included for participants, and by-participant random slopes for *Block*, *onsetF0*, and *feedback*. Interaction plots and pairwise comparisons were made using the `emmeans` package [23].

As shown in Fig. 2 (left), in the absence of lexical feedback, listeners' voiceless response rate increased in Block 2 as compared to Block 1 for high onset F0 only ($est. = 0.44, z = 2.45, p < .05$) but response rate did not differ between the two blocks for low onset F0 ($est. = 0.17, n.s.$). When lexical feedback was given, listeners' voiceless response rate increased in Block 2 compared to Block 1 for both F0 levels, but the increase was of a greater magnitude when onset F0 was high ($est. = 0.99, z = 4.87, p < .0001$) than when it was low ($est. = 0.59, z = 2.54, p < .05$). In sum, an increase in the informativeness of F0 led listeners to upweight the use of *high* onset F0 in identifying voiceless plosives. Lexical feedback strengthened this effect.

2.4. Experiment 2: Decreased VOT informativeness

2.4.1. Procedure

Experiment 2 examined listeners' perceptual use of F0 as the informativeness of VOT decreases, with two different distributional configurations, schematised in Fig. 1 (right). The test stimuli are the same as for Experiment 1. In Block 1 (VOT bimodal), the exposure stimuli are represented by the empty dots and the test stimuli by the black dots. Voiced and voiceless categories were well separated along both the VOT and F0 dimensions. Furthermore, VOT covaried with F0, with positive VOTs associated with a high onset F0 and negative VOTs with a low onset F0. That is, both VOT and F0 were informative. In Block 2 (VOT ambiguous), the test stimuli and the

exposure stimuli (black dots) were the same. Stimuli were well separated along the F0 dimension but shared the same VOT step, which was ambiguous for voicing identification (i.e., VOT was distributionally uninformative).

As in Experiment 1, half the listeners received lexical feedback (display of the appropriate picture in Block 2), while the other half did not. Regardless of VOT, the pictures representing *bain* and *daim* were displayed as feedback accompanying auditory stimuli with a low onset F0 (144 Hz), and the pictures representing *pain* and *thym* were given as feedback to the auditory stimuli with a high onset F0 (188 Hz).

The no-feedback version contained 406 trials in total (12 exposure and test stimulus trials (Block 1) $\times$ 2 word pairs $\times$ 8 repetitions + 1 test stimulus (Block 2) $\times$ 2 word pairs $\times$ 90 repetitions + 34 attention checks). The feedback version contained 310 trials in total (12 exposure and test stimulus trials (Block 1) $\times$ 2 word pairs $\times$ 6 repetitions + 1 test stimulus (Block 2) $\times$ 2 word pairs $\times$ 70 repetitions + 26 attention checks).

2.4.2. Prediction

If the decrease in the informativeness of VOT in distributional terms *alone* upweights listeners' use of F0 in voicing categorisation, we predict an increase in the use of F0 to the test stimuli in Block 2 compared to Block 1. Again, if lexical feedback plays a role, we predict that such an increase will occur only or especially when lexical feedback is given.

2.4.3. Analysis and results

A GLMM with the same structure as for Experiment 1 was fitted to the response data to the test stimuli. As shown in Fig. 2 (right), when lexical feedback was *not* given, listeners' voiceless response rate was slightly increased in Block 2 as compared to Block 1 for low onset F0 ($est. = 0.49, z = 1.88$, $p = .061$), but did not differ between the two blocks for high onset F0 ($est. = -0.19, n.s.$). When lexical feedback was given, listeners' voiceless response rate was decreased in Block 2 as compared to Block 1 for low onset F0 ($est. = -0.97, z = -3.12$, $p < .005$), but did not differ for high onset F0 ($est. = 0.20, n.s.$).

In sum, while listeners did not respond to a decrease in the informativeness of VOT alone, the presence of lexical feedback led to an upweighted use of F0. Contrary to the increased use of high onset F0 in Experiment 1, here listeners increased their use of *low* onset F0 in identifying voiced plosives.

3. DISCUSSION

Our results first show that increased informativeness of F0 leads to a greater use of *high* onset F0 in identifying voiceless plosives, whereas a decreased informativeness of VOT leads to a greater use of *low* onset F0 in identifying voiced plosives, the latter being achieved only when lexical feedback is given. We hypothesise that general auditory plays a role in this asymmetry. In Experiment 1, while the raising and lowering of F0 is symmetrical in acoustic terms, listeners may be more sensitive to F0 raising at the auditory/attentional level [24]. In Experiment 2, the removal of prevoicing from voiced plosives may trigger a perceptual shift to low onset F0 as a replacement of the low-frequency energy at the vowel onset [25], giving rise to an increased perceptual weight of low onset F0. In contrast, the removal of voice lag from voiceless plosives does not affect the low-frequency energy at the vowel onset, hence F0 is not used as a rescue option. A similar asymmetry may then be expected in other languages where the voicing contrast is primarily signalled by the presence vs. absence of closure voicing such as Italian.

Our second finding, that listeners' increased use of high or low onset F0 only occurs or is strengthened when lexical feedback is given, is in line with other studies showing a crucial effect of lexical feedback in perceptual learning [13, 14, 15, 16]. Furthermore, it suggests that changes in distributional informativeness are processed by the listeners, but that for at least some kinds of contrasts or regions of phonetic space, distributional changes alone may not be sufficient to trigger a perceptual adaptation.

In sum, our study shows how a particular combination of distributional, auditory, and lexical factors contributes to the dynamic processing of VOT and F0 in a true voicing language. They suggest that F0 raising after voiceless plosives is perceptually more salient than F0 lowering after voiced plosives [26], mirroring the production pattern [27]. However, the exposure to devoiced tokens (when combined with listeners' implicit lexical knowledge) seems most favourable for increasing the perceptual weight of F0 lowering. Such laboratory findings may have implications for cue re-weighting in sound change (e.g., [28, 29, 30, 31]), pointing to perceptual mechanisms involved in the phonologisation of co-intrinsic F0 (aka tonogenesis) or a secondary cue in general.

4. ACKNOWLEDGEMENTS

This work was funded by ERC StG 758605 to J. Kirby. Thanks to M. Clayards, J. Schertz, J. Riverin-Coutlée and M. Gubian for helpful comments.

5. REFERENCES

[1] House, A. S., Fairbanks, G. 1953. The influence of consonant environment upon the secondary acoustic characteristics of vowels. *JASA* 25(1), 105–113.

[2] Holt, L. L., Lotto, A. J. 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. *JASA* 119(5), 3059–3071.

[3] Schertz, J., Clare, E. J. 2020. Phonetic cue weighting in perception and production. *WIREs Cognitive Science* 11(2), e1521.

[4] Clayards, M., Tanenhaus, M. K., Aslin, R. N., Jacobs, R. A. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108(3), 804–809.

[5] McMurray, B., Aslin, R. N., Toscano, J. C. 2009. Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science* 12(3), 369–378.

[6] Idemaru, K., Holt, L. L. 2011. Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance* 37(6), 1939–1956.

[7] Kirby, J. 2013. The role of probabilistic enhancement in phonologization. In: Yu, A. C. L. (ed), *Origins of Sound Change: Approaches to phonologization*. Oxford University Press, 228–246.

[8] Serniclaes, W. 1987. *Étude expérimentale de la perception du trait de voisement des occlusives du français*. PhD Thesis. Université Libre de Bruxelles.

[9] Abramson, A. S., Lisker, L. 1985. Relative power of cues: F0 shift versus voice timing. *Phonetic linguistics: Essays in honor of Peter Ladefoged*, 25–33.

[10] Summerfield, Q., Haggard, M. 1974. Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. *Journal of Phonetics* 2(4), 279–295.

[11] Francis, A. L., Nusbaum, H. C. 2002. Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance* 28(2), 349–366.

[12] Kim, D., Clayards, M., Kong, E. J. 2020. Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics* 81, 100984.

[13] Norris, D., McQueen, J. M., Cutler, A. 2003. Perceptual learning in speech. *Cognitive Psychology* 47(2), 204–238.

[14] Harmon, Z., Idemaru, K., Kapatsinski, V. 2019. Learning mechanisms in cue reweighting. *Cognition* 189, 76–88.

[15] Kraljic, T., Samuel, A. G. 2006. Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review* 13(2), 262–268.

[16] Goudbeek, M., Cutler, A., Smits, R. 2008. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication* 50(2), 109–125.

[17] Stoet, G. 2010. PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods* 42(4), 1096–1104.

[18] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44(1), 24–31.

[19] Mitterer, H. Generating fade-in and fade-outs (personal web page).

[20] Valbret, H., Moulines, E., Tubach, J.-P. 1992. Voice transformation using PSOLA technique. *Speech Communication* 11(2-3), 175–187.

[21] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13).

[22] R Core Team, 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

[23] Lenth, R. 2022. emmeans: Estimated marginal means, aka Least-squares means.

[24] Hsu, C. H., Evans, J. P., Lee, C. Y. 2015. Brain responses to spoken F0 changes: Is H special? *Journal of Phonetics* 51, 82–92.

[25] Kingston, J., Diehl, R. L., Kirk, C. J., Castleman, W. A. 2008. On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics* 36(1), 28–54.

[26] Gao, J., Yun, J., Arai, T. 2019. VOT-F0 coarticulation in Japanese: Production-biased or misparsing. *Procs. of the 19th ICPhS*, paper 619.1–5.

[27] Kirby, J., Ladd, D. R. 2016. Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages. *JASA* 140(4), 2400–2411.

[28] Beddor, P. S. 2009. A coarticulatory path to sound change. *Language* 85(4), 785–821.

[29] Coetzee, A. W., Speeter, P., Shedden, K., Styler, W., Wissing, D. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics* 66, 185–216.

[30] Kuang, J., Cui, A. 2018. Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics* 71, 194–214.

[31] Harrington, J., Kleber, F., Reubold, U., Schiel, F., Stevens, M. 2019. The phonetic basis of the origin and spread of sound change. In: Katz, W. F., Assmann, P. F. (eds), *The Routledge Handbook of Phonetics*. New York: Routledge.

PRODUCTION AND PERCEPTION OF MANDARIN /I/-NASAL RHYMES

Suyuan Liu and Molly Babel

University of British Columbia
suyuan.liu@ubc.ca, molly.babel@ubc.ca

ABSTRACT

This study compares how speakers of Beijing Mandarin (BjM) and Shanghai Mandarin (ShM) differ in their production and perception of nasal codas as a means of evaluating the (mis)alignment of perception and production. Speaker-listeners completed a word-level production task, a word recognition task, and a goodness rating task. Results suggest that BjM users contrast /in/-/iŋ/ words in perception and production by consistently mapping [in] and [iɲ] to the former and [iŋ] and [iəŋ] to the latter. ShM users show no /in/-/iŋ/ word contrast in production while consistently mapping the [iəŋ] variant to /iŋ/-words and rating it as the best /iŋ/-exemplar in perception. This indicates a mismatch in perceptual and production, providing insights into cross-dialect variation.

Keywords: nasals, production, perception, Mandarin

1. INTRODUCTION

Mandarin includes diverse varieties that emerge through the promotion of standard Mandarin to speakers of Chinese languages. Different Mandarin varieties have distinct production variants that may be unfamiliar to speakers of other varieties. However, the term "Mandarin" obscures multiple language concepts [1]. Two terms are used in this paper to differentiate the concepts. *Standard Mandarin* is used to refer to the idealized Mandarin promoted in China. *[City Name] Mandarin* (e.g. Shanghai Mandarin) is used in reference to the Mandarin variety spoken in or around the city region, which is *not* the same as the local Chinese language (e.g. Shanghainese).

Standard Mandarin only allows two codas /n/ and /ŋ/. The degree of distinction between /n/ and /ŋ/ is conditioned by three main factors: (1) the preceding vowel, (2) the Mandarin variety, and (3) whether the investigation is in speech perception or production [2, 3, 4, 5]. Among all Mandarin varieties, Taiwan Mandarin (TwM), Shanghai Mandarin (ShM) and Beijing Mandarin (BjM) are the most well-studied when it comes to Mandarin nasal codas. In all three varieties, the degree of distinction in production and perception is vowel-dependent. The high front vowel rhyme /iN/ – N is used to represent an unspecified nasal – has the lowest degree of distinction in TwM and ShM production [3, 5, 6, 7, 8, 9, 10] and the lowest perceptual classification accuracy [3, 4, 5, 11]. In production, both groups use a (fronted) velar tongue gesture when producing the nasal in /iN/ [3, 6]. BjM, however, makes such a distinction by realizing /in/ as [in] and /iŋ/ as [iəŋ], a foreign variant to ShM and TwM [12, 13, 14]. But interestingly, even BjM listeners are poor at perceiving /iN/ rhymes [3, 15, 16], which suggests a misalignment between perception and production. However, the perception studies did not take into account the vowel differences that are present in BjM production. In addition, the majority of previous perception studies focused on phonetic distinction [3, 16] or instructed listeners to map phonemes to a Pinyin representation (e.g. select "n" or "ng" for the nasal you heard) [5, 15]. It is unclear whether the poor perception performance generalizes to a lack of sensitivity at the lexical level. To address the perception and production misalignment in [3], the current work reassesses the nasal coda distinction in perception and production of /iN/ rhymes when considering the vowel differences and focuses perceptual sensitivity to the word level for BjM and ShM.

2. GENERAL PROCEDURE

Participants were recruited through Chinese social media from Beijing and Shanghai. 29 participants' data were analyzed: 14 in the ShM group (9 male, 5 female, M = 31.9 years, range = 22-56) and 15 in the BjM group (7 male, 8 female, M = 26.7 years, range = 20-32). All participants completed language background and residential history questionnaires and a headphone [17] and microphone check. They then took part in a production experiment and a set of perception experiments: a two-way alternative forced choice task and a goodness rating task with Chinese instructions. Including both tasks examines category boundaries and corresponding prototypes, which could differ by dialectal groups [18]. All

experiments were hosted online using Gorilla [19].

3. PRODUCTION STUDY

Participants were instructed to record themselves producing the Chinese character shown on the screen. Four practice trials were completed prior to the actual task. Trials were separated by a 450ms fixation pause. The screen automatically advanced to the next trial if no recording was made in 3s.

3.1. Material

Production stimuli consisted of 9 minimal pairs of frequency-matched monosyllabic words in Chinese characters. Each character appeared twice per block for three blocks (totaling 108 trials). The 3 pairs of target stimuli are shown on the left in Table 1. The other 6 pairs were codaless fillers items (e.g. [p^{h}wo] 'splash' 泼- [pwo] 'spread' 播).

Production Stimuli	Perception Stimuli			
	in	iɲ	iŋ	iəŋ
/jin/-/jiŋ/ 'sound' - 'eagle' 音- 鹰	[jin]	[jiɲ]	[jiŋ]	[jiəŋ]
/bin/-/biŋ/ 'guest' - 'ice' 宾- 冰	[bin]	[biɲ]	[biŋ]	[biəŋ]
/tɕin/-/tɕiŋ/ 'gold' - 'shock' 金- 惊	[tɕin]	[tɕiɲ]	[tɕiŋ]	[tɕiəŋ]

Table 1: Target production and perception stimuli

3.2. Production coding

All tokens were manually coded by a Spanish-English bilingual and a Northern Mandarin-English bilingual. Both coders are phonetically trained. Coders listened to all randomized tokens and labeled each token as (1) [in], (2) [iŋ], (3) [iəŋ], (4) [iɲ], (5) no nasal coda, or (6) other with text for specification. Two coders agreed on 622 out of 1149 tokens (unweighted Cohen's kappa = 0.581, $p < 0.001$) and jointly relabeled all tokens disagreed on.

3.3. Production results

Impressionistically, the distribution of /iN/-words pronunciations show similarity between BjM and ShM groups. To assess this and facilitate model interpretation, separate models were run for the /in/ and /iŋ/ words, allowing for cross-dialect comparisons within each lexical set. Two Bayesian mixed-effect multinomial logistic regression models were fit separately for /in/ and /iŋ/ words in Stan

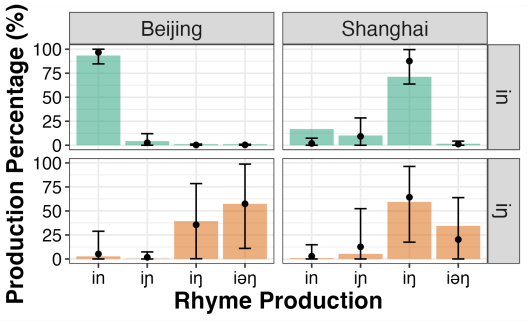


Figure 1: Participants' proportion of produced categories when presented with a character for /in/-word or /iŋ/-word (rows) for the two dialects (columns). The error bars show the 95% HDI of posterior predictions for each group.

using the {brms} package [20] in R [21]. Both models included an outcome variable of dummy-coded Production Category ([in], [iŋ], [iɲ], and [iəŋ]; ref level: [iŋ]) and a predictor variable of dummy-coded Mandarin varieties (ShM vs. BjM; ref level: BjM). Both models also included by-speaker and by-word random intercepts. Samples were drawn from the posterior distribution using a four-chain Hamiltonian Monte-Carlo sampling (2000 iterations, 100 warm-ups). Weakly informative priors were used for all parameters. The mean of the posterior distribution, the 95% credible interval (CrI), and the probability of direction (PD) are reported. A CrI that does not encompass 0 suggests a meaningful effect and a PD greater than 95% suggests a probable direction of effect [22]. The bars in Figure 1 show the average percentage of producing one of the four variants when seeing a /in/-word or /iŋ/-word for each Mandarin variety in the empirical data. The error bars show the range of posterior predictions for each grouping.

The models give evidence that BjM speakers distinguish /iN/ by consistently producing [in] tokens for /in/-words (β = 7.46, CrI = [3.57, 11.81], PD = 99.95%) while producing a mix of [iŋ] and [iəŋ] (β = 0.61, CrI = [-1.85, 3.11], PD = 94.77%) for /iŋ/-words. For ShM speakers, the model gives evidence that they consistently produce more [iŋ] for /in/-words than [in] (β = -5.95, CrI = [-9.42, 0.08], PD = 97.32%), and produce more [iŋ] tokens than [iəŋ] tokens (β = -1.6, CrI = [-4.57, 0.11], PD = 97.08%) for /iŋ/-words.

4. PERCEPTION: LEXICAL CATEGORIZATION

In this two-way alternative forced choice task, participants heard a stimulus and saw two Chinese

characters for a minimal pair on the screen. Left and right arrow keys were used to indicate which character they thought they heard. Participants started with four practice trials with feedback. No feedback was given after the practice trials. The screen advanced to the next trial if no answer was submitted in 2.5s. A 1-sec fixation reset pause was added between trials.

4.1. Materials

Perception stimuli consisted of 4 production variants of the same 3 minimal pairs listed in Table 1 and 2 productions of the canonical endpoints of 6 filler minimal pairs. All tokens were produced by a 24-year-old female Northern Mandarin-speaking trained linguist who grew up in Shandong, China. Stimuli were digitally recorded at a sampling frequency of 44.1 kHz. All production tokens were then coded by two Spanish-English bilingual trained linguists as [n], [ŋ], [ɲ], or [ɴ]. Only tokens that both coders agreed to match the target production were selected as stimuli. Two tokens were used for each variant. Each auditory stimulus was repeated twice per block for two randomized blocks (totaling 192 trials, 8 trials for each lexical target). The intensity of all stimuli was scaled to 75 dB and .wav files were converted to the Ogg vorbis container format.

4.2. Categorization task results

Within dialect group quantification was made to evaluate the perceptual equivalence of the pronunciation variants. Two Bayesian mixed-effect logistic regression models were fit separately for ShM and BjM groups using the {brms} package [20] in R [21] with the dummy-coded Probability of Velar Responses (/iŋ/ = 1, /in/ = 0) as the outcome variable. The dummy-coded auditory Stimuli Type ([in], [iŋ], [iɲ], [iəŋ]; ref level: [iŋ]) was the predictor variable. A by-subject random intercept with Stimuli Type as a by-subject slope and a by-word random intercept were included. Samples were drawn from the posterior distribution using a four-chain Hamiltonian Monte-Carlo sampling (2000 iterations, 500 warm-ups). Weakly informative priors were used for all parameters. The bars in Figure 2 show the average percentage of responses where a velar vs. alveolar character is selected in response to the auditory stimuli. The error bars show the mean and range of posterior predictions of velar responses for each stimuli type.

The models give evidence that BjM listeners distinguish /in/-/iŋ/ by consistently choosing /in/-words for [in] (β = -4.58, CrI = [-8.66, -4.76], PD = 100%) and [iɲ] (β = -0.82, CrI = [-3.59, -1.80], PD = 100%), and /iŋ/-words for [iŋ] (β = 1.82, CrI = [-0.03, 3.64], PD = 97.37%) and [iəŋ] (β = 5.08, CrI = [1.99, 5.51], PD = 100%). ShM listeners only distinguish /in/-/iŋ/ by consistently selecting /iŋ/-words for [iəŋ] (β = 2.47, CrI = [1.82, 4.05], PD = 100%) and /in/-word for the rest. The posterior distributions in Figure 2 also show that BjM listeners are more likely to choose /in/-words for [in] than [iɲ], and more likely to choose /iŋ/-words for [iəŋ] than [iŋ].

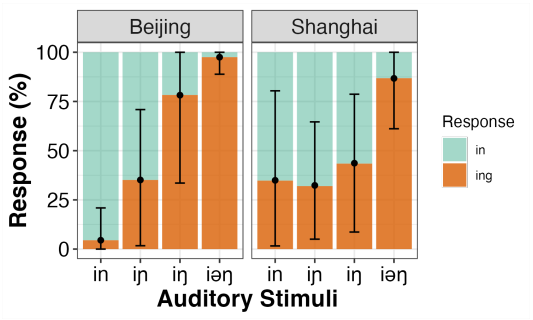


Figure 2: Word categorization results for ShM and BjM by stimuli type (x-axis). The y-axis plots the percentage of /iŋ/ word responses. Error bars present the mean and 95% HDI of posterior predictions of velar responses.

5. PERCEPTION: GOODNESS RATING

Only the four variants of target /iN/ minimal pairs were included in this task to abbreviate the experiment; no filler trials were included. In each trial, participants heard the stimulus while seeing a Chinese character and a continuum ranging from 1 to 7 in whole numbers. Listeners were asked to click on the rating that indicates how well the auditory stimulus matched the character, with 1 being the least and 7 being the most accurate. Participants started with four practice trials. The screen automatically advanced to the next trial in 5s. A 1-sec fixation reset pause was added between trials to avoid habitual clicking.

5.1. Materials

Same critical items were used as in Section 4.1. Each variant had two production tokens repeated twice with each appropriate Chinese character in each block for three randomized blocks (totaling 144 trials, 6 trials for each lexical target).

5.2. Goodness rating results

Similar to the categorization task, within-group comparisons were made. Four Bayesian mixed effect linear regression models were fitted separately for each dialect group and for when the displayed word ends with /n/ and /ŋ/ using the {brms} package [20] in R [21]. The outcome variable was the Rating and the predictor variable was the dummy-coded Auditory Stimuli Type ([in], [iŋ], [iɲ], [iəŋ]; reference level: [iŋ]). A by-subject random intercept with Stimuli Type as by-subject slope and a by-word random intercept were included. Samples were drawn from the posterior distribution using a four-chain Hamiltonian Monte-Carlo sampling (2000 iterations, 500 warm-ups). Weakly informative priors were used for all parameters. Figure 3 shows the group ratings for the four variant types with error bars indicating the 95% HDI range of posterior predictions.

The models provide evidence that [iəŋ] is rated as the best /iŋ/ exemplar by both BjM (β = 6.76, CrI = [0.45, 1.60], PD = 99.95%) and ShM group (β = 5.66, CrI = [0.02, 3.04], PD = 97.58%). BjM listeners rated [in] as the best /in/ exemplar (β = 6.61, CrI = [2.34, 3.82], PD = 100%), [iɲ] as a better match for /in/ than /iŋ/ (β = 5.22, CrI = [1.12, 2.22], PD = 100%) and [iŋ] as a better match for /iŋ/ than /in/ (β = 5.74, CrI = [4.60, 6.85], PD = 100%). The models show little evidence for difference across ShM listeners' rating for [in], [iɲ], and [iŋ] (PD < 75%) and the posterior predictions in Figure 3 show that ShM listeners consistently give higher /in/-word ratings to all these three variants.

6. DISCUSSION AND CONCLUSION

In production, it is clear that BjM speakers contrast /iN/ and produce variants with no overlap. ShM speakers, at a group level, show a low yet existing degree of distinction of [iŋ]-[iəŋ] for /iN/ rhymes. This pattern could not have been observed when the investigation of ShM production was limited to a single nasal segment level, since the nasals in both variants are /ŋ/. The role of vowel for ShM suggests that some contrast at the rhyme level remains. This could indicate a shift in contrastive acoustic cues from nasal coda to vowel, but this requires further investigation.

Perceptually, both BjM and ShM listeners consider [iəŋ] as the best exemplar for /iŋ/. ShM listeners exhibit lower-to-no degree of distinction in lexical mappings for variants that only differ on the nasal segments – that is, at the group level there is not strong evidence that a particular nasal coda maps

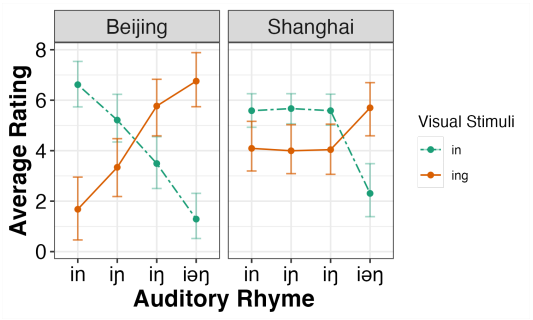


Figure 3: Goodness ratings for ShM and BjM with the y-axis showing the rating from 1 (worst) to 7 (best) for the stimuli types. The error bars indicate the 95% HDI of posterior predictions.

to a particular lexical item. However, BjM listeners still exhibit contrastive lexical mappings of variants that have no vowel quality difference ([in] vs. [iŋ]), which contradicts previous results [3, 16]. This suggests that the type of perceptual task (AXB task vs. lexical categorization task) and response (Pinyin label vs. Chinese character) elicit different patterns and conclusions. Lexical information can facilitate categorization [23], and the use of a single voice could have improved listener performance [24].

When comparing each group's perception to their own production, it is interesting to see that ShM users' exemplar for /iŋ/ does not match across perception and production. One possible explanation is ShM users' frequent exposure to BjM and Standard Mandarin via media and education. Due to the promotion of Standard Mandarin, speakers of all Mandarin varieties were taught and tested on the contrast in coda nasal through the Pinyin representation even when they do not adopt such distinction in production. Therefore, when presented with the auditory stimuli [iəŋ], which is "foreign" to ShM listeners, they could still map it to the canonical BjM productions of /iŋ/ words. This suggests a possibility of social factors and language exposure influencing listeners' perceptual exemplars even when such exemplars are not present in their production space.

To sum up, the current project revisits the distribution of Mandarin coda nasals in ShM and BjM users' perception and production. It provides insights into the necessity of considering the degree of distinction above sound segment level in both perception and production. This project focusing on group-level does not account for the large degree of individual variation. Follow-up studies investigating the individual perception-production (mis)alignment and potential generational differences are in progress.

7. ACKNOWLEDGMENTS

Thank you to Roger Lo for statistical guidance. Kathleen Currie Hall provided comments that improved the clarity of this work.

8. REFERENCES

[1] J. Weng, "What is mandarin? the social project of language standardization in early republican china," *The Journal of Asian Studies*, vol. 77, no. 3, pp. 611–633, 2018.

[2] L. Wang, J. Cui, and Y. Chen, "Wuxi speakers' production and perception of coda nasals in mandarin." in *INTERSPEECH*, 2018, pp. 2559–2562.

[3] M. Faytak, S. Liu, and M. Sundara, "Nasal coda neutralization in shanghai mandarin: Articulatory and perceptual evidence," *Laboratory Phonology*, vol. 11, no. 1, 2020.

[4] C. C. Lin, "Nasal endings of taiwan mandarin: Production, perception, and linguistic change," in *35th International Conference on Sino-Tibetan Languages and Linguistics, Arizona State University*, 2002.

[5] M. Luo, M. Sloos, and J. van de Weijer, "Perception and production of mandarin nasal codas by shanghainese speakers," *Chinese Accents and Accented Chinese CAAC2014*, p. 16, 2015.

[6] C. Chiu, Y. Lu, Y. Weng, S. Jin, W. Weng, and T. Yang, "Uncovering syllable-final nasal merging in taiwan mandarin: Ultrasonographic investigation of tongue postures and degrees of nasalization," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 398–402.

[7] J. Fon, J.-m. Hung, Y.-H. Huang, and H.-j. Hsu, "Dialectal variations on syllable-final nasal mergers in taiwan mandarin," *Language and Linguistics*, vol. 12, no. 2, pp. 273–311, 2011.

[8] H.-j. Hsu and J. K.-P. Tse, "Syllable-final nasal mergers in taiwan mandarin-leveled but puzzling," *Concentric: Studies in Linguistics*, vol. 33, no. 1, pp. 1–18, 2007.

[9] Y.-H. Lai, "Production of mandarin chinese nasal coda by l1 and l2 speakers of mandarin chinese," 华语文教学研究*[Huayu Wenjiaoxue Yanjiu]*, vol. 5, no. 1, pp. 155–180, 2008.

[10] Y. Guan, "Nasal coda realisation in speech production of shanghai mandarin," 2019.

[11] Y.-h. Lai, "Perceptual processing of mandarin nasals by l1 and l2 mandarin speakers," *Journal of psycholinguistic research*, vol. 41, no. 4, pp. 237–252, 2012.

[12] X. Mou, "Nasal codas in standard chinese: a study in the framework of the distinctive feature theory," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[13] Q. Ran, "汉语鼻音韵尾的实验研究[An experimental study of Mandarin nasal codas]," 南开语言学刊*[Nankai Yuyanxue Kan]*, no. 2, pp. 37–44, 2005.

[14] F. Shi, "北京话的元音格局[The Vowel Distribution of Beijing Hua]," 南开语言学刊*[Nankai Yuyanxue Kan]*, no. 1, pp. 30–36, 2002.

[15] Y. Chen and S. Guion-Anderson, "Perceptual confusabiltiy of word-final nasals in southern min and mandarin: Implications for coda nasal mergers in chinese." in *ICPhS*, 2011, pp. 464–467.

[16] W. Wu, "普通话鼻音韵尾的听辨实验研究[a perception discrimination study of mandarin nasal codas]," Master's thesis, 天津师范大学[Tianjin Normal University], 2019.

[17] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[18] X. Xie, R. M. Theodore, and E. B. Myers, "More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 43, no. 1, p. 206, 2017.

[19] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior research methods*, vol. 52, no. 1, pp. 388–407, 2020.

[20] P.-C. Bürkner, "Bayesian item response modeling in R with brms and Stan," *Journal of Statistical Software*, vol. 100, no. 5, pp. 1–54, 2021.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[22] B. Nicenboim and S. Vasishth, "Statistical methods for linguistic research: Foundational ideas—part ii," *Language and Linguistics Compass*, vol. 10, no. 11, pp. 591–613, 2016.

[23] M. H. Davis and I. S. Johnsrude, "Hearing speech sounds: top-down influences on the interface between audition and speech perception," *Hearing research*, vol. 229, no. 1-2, pp. 132–147, 2007.

[24] C. E. Stilp and R. M. Theodore, "Talker normalization is mediated by structured indexical information," *Attention, Perception, & Psychophysics*, vol. 82, pp. 2237–2243, 2020.

THE ROLE OF BOUNDARY TONE IN CHUNKING SPONTANEOUS SPEECH

Lucas Weidinger[1], Xinyuan Wan[1], Liisa Koivusalo[1], Aleksandra Dobrego[1]

[1]University of Helsinki, Finland
lucas.weidinger@helsinki.fi xinyuan.wan@helsinki.fi liisa.koivusalo@helsinki.fi aleksandra.dobrego@helsinki.fi

ABSTRACT

This study aimed to explore how listeners segment spontaneous speech of different nature and what role boundary tone plays in the process. Participants (n = 41) were asked to listen to extracts of scientific and conversational speech while seeing corresponding transcripts on a screen and to segment the speech into chunks by placing markers between orthographic words. No explanation on what chunks are or should be was given to participants. The speech samples were annotated for silent pauses and boundary tones. Agreement on chunks between participants was high, with scientific samples showing stronger agreement than conversational ones (Fleiss' kappa, 0.56 vs 0.396, $p<.05$). Boundary tone appears to be a significant cue in both types of speech. The level of tone did not matter. The study suggests that spontaneous speech processing is affected by boundary tone independently of the nature of speech, and listeners use boundary tones as cues for chunking.

Key words: chunking, spontaneous speech

1. INTRODUCTION

The speech input we encounter every day is mostly spontaneous. It is filled with disfluencies, false starts, hesitations, and repetitions, but they nevertheless do not affect people's ability to effectively understand it. One of the prerequisites for understanding is chunking - the fast and largely automatic process by which the listener determines where one meaningful unit ends and the next begins [1]. What these chunks are, whether listeners agree on them, and what cues mark a chunk, are debatable issues.

Here, speech is studied from the perspective of Linear Unit Grammar (LUG) [2], which assumes that language is linear, unfolds in real time and is segmented into chunks by a listener or a reader. These chunks then can be treated as cohesive building blocks that carry meaning and give the speech structure. Even without a thorough explanation of what chunks are, people tend to intuitively chunk speech in the same way [3][4][5]. This was shown on English [3], Finnish, Russian and Swedish materials [6]. The participants listened to the speech extracts while simultaneously seeing the text on screen and were asked to mark boundaries between chunks (ChunkitApp, [7]). The researchers assessed comprehension, agreement, and segmentation strategies of different groups of speakers. It was found that participants agreed on chunks in their places and size and that their segmentation strategies did not differ. Interestingly, speakers of all languages studied (English/Finnish/Russian/Swedish) and all language levels (native/non-native, [5]) used prosodic cues rather than syntactic when choosing where to place a boundary in speech.

This paper sets out to explore this crucial role of prosody in chunking by looking at boundary tone (a rise or fall in pitch). The initial idea was to see whether boundary tone is important in chunking different kinds of speech. extracts of conversational and scientific speech were chosen, and participants' agreement and segmentation choices were investigated using ChunkitApp. ChunkitApp is a tablet application that displays transcripts and plays corresponding extracts simultaneously. The agreement served as a method validation: if, as in previous studies, the agreement was high, then this method did capture intuitive segmentation that could be analysed. If the agreement was low, some other method would need to be found to better capture participants' choices in segmentation.

It was expected that the boundary tone would be a significant indicator for scientific speech, but not for conversational speech. This hypothesis stems from the expectation that the prosody of the more formal, scientific speech is more premeditated than the prosody of the purely conversational speech. When the structure and prosody of the uttered sentence are planned slightly ahead of time, chunk boundary markers, like changes in pitch for example, would be placed more deliberately by the speaker. It was expected to be the case for what is here described as *scientific speech*: recordings of lectures, in which a lecturer knows what they are going to say ahead of time.

However, it was found that there was no difference between two types of speech, spontaneous and scientific, and the boundary tone was a reliable cue in chunking both. This suggests that the role of pitch fluctuations in processing spontaneous speech is independent of the setting and the purpose of the talk: Whether it is a predominantly monologic lecture or a conversation, dialogistic in nature.

2. METHODS

2.1. Experimental setup

The experiment was conducted using ChunkitApp [7]. Participants listen to the extract and marks chunk boundaries by pressing on a tilde symbol (~) between orthographic words in the transcript on screen. When pressed, the tilde symbol turns into a vertical line (|). This is then considered a boundary. After each extract, participants must answer a comprehension question. Participants were free to place boundaries wherever they felt they belonged, and all boundary placements were considered valid for analysis, given high enough agreement between participants on the location of the boundary. They received no instructions as to how many and where to place boundaries, with the exception that they should place at least one boundary in each extract.

The stimuli were extracts of spontaneous speech and their transcripts. They were taken from the AE minicorpus [8] in case of the conversational speech, while the scientific speech samples were taken from open-source lecture recordings. A full list of the sources for the "scientific speech" samples can be found in the Appendix. 20 stimuli of each type, 40 in total, were included in the experiment. Additionally, one test stimulus and two training stimuli were used to familiarise the participants with the task. The durations of the stimuli ranged from 10 to 15 seconds, and they consist of between one and three sentences. All extracts were in English.

The transcripts were created manually to get them to reflect the speech samples as accurately as possible. An automatic speech-to-text system was found to be unsuitable for this task, as machine generated transcripts contained too many errors. They omitted some of the information specifically needed for this experiment, like repetitions and non-starts. It was ultimately faster and easier to transcribe by hand than to fix the machine generated transcripts.

ChunkitApp was also used to collect informed consent and background information of participants. Participants were asked about their education, language background, proficiency in English, and whether they had diagnosed reading or hearing disorders. Questions regarding native languages, proficiency in English, and reading and hearing disorders were obligatory to be able to control the participant group. Personal information, such as participants' names or voices, was not collected, and the participants could not be identified from their answers in any phase. Participants could take part in the study using their own device and a pair of headphones. They were allowed to perform the experiment at their own discretion and withdraw at any time.

2.2. Participants

Participants were recruited through the University of Helsinki Faculty of Arts mailing list as well as social media. The total number of recruited participants was 89, however, only results from 41 participants were used in the final analysis (see 3.1 Outliers). All of them were naïve to the purpose of experiment. Prior to the main task, participants were asked to self-report their English skills in terms of the six CEFR levels of English proficiency for adults (A1-C2). The description of these were given to the participants as a guideline to help them self-report their own English skills. All participants included in the analysis reported at least an upper intermediate level of English skill (CEFR B2). Five participants were English native speakers, 36 were L2- English speakers. The distribution of self-reported English skills is shown below. Participants were asked to self-report their English skills on a six-point Likert-scale. Figure 1 shows that 34 participants indicated their overall English skills as either advanced (5) or proficient (6).

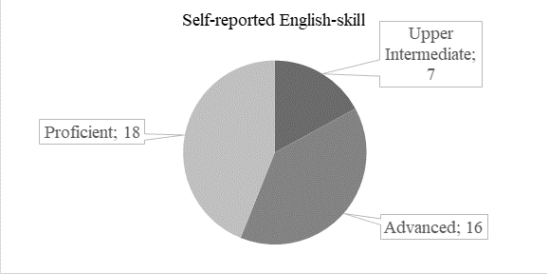


Figure 1: Participants' self-reported English skills, given in CEFR levels

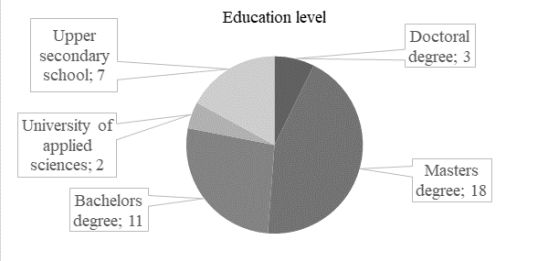


Figure 2: Participants' highest completed level of education

Participants were also asked to report the highest level of education they completed. Most of the participants (34 people) had some description of a tertiary degree. The exact distribution of completed degrees by participants can be seen in Figure 2.

2.3. Annotations

All speech samples were annotated using three tiers in Praat [9]: (1) orthographic, (2) break index, and (3) tonal. The orthographic tier was an interval tier, the others point tiers. On the orthographic tier, all utterances were marked using the transcripts. These utterances were generally separated by pauses; this made annotation of the extracts easier. Silent pauses 50-250 ms (SP) and above 250 ms (LSP) in duration were marked on the break index tier. On the tonal tier, either a high (H%) or a low (L%) tone was marked, based on the F0 pitch just before a silent pause. The full scheme for the annotations can be seen in Table 1.

The annotations were done by two annotators independently before being reviewed by all four authors.

Tier	Annotations		Count
Break index	silent pause	SP	65
	long silent pause	LSP	114
Tonal	high	H%	65
	low	L%	227

Table 1: Annotation scheme

3. RESULTS

3.1. Outliers

Participants were excluded from the analysis based on their background and the minimum number of boundaries they had put in total.

89 participants were recruited. 31 participants were excluded due to their education background in linguistics, applied linguistics and some other programs with a focus on language studies. Recruiting linguists for the task was avoided since they often have an elaborate knowledge of language science, and they could stand out from the target sample. Additionally, 4 participants were excluded because they were diagnosed with dyslexia and hearing disorders. 13 participants had put less than 40 boundaries per 40 extract, and were excluded as well. Since the task was to put at least one boundary in each extract, in their case the task was considered incomplete. Therefore, in total, 41 users were included in the final analysis.

3.2. Analysis of agreement

All following analysis was done using programming language R version 4.0.5 and RStudio [10][11]. The resulting boundary data from ChunkitApp was a data frame with all possible boundaries as rows and participants as columns. A boundary was marked with 1 and no boundary with 0. Agreement was used as a measure to ensure the observed behaviour was not just random but was motivated for all participants with certain factors. Agreement was calculated similarly to [3]. Fleiss' kappa [12] was chosen as a measure to assesses the reliability of agreement between three or more raters. The measure ranges from 0 to 1, with 0 indicating no agreement and 1 indicating absolute agreement. Several studies have shown that Fleiss' kappa values are comparable across populations and experimental conditions [3][12], and results from previous studies show that it is conceptually possible to use Fleiss' kappa to investigate participants' agreement rate [3][4][5]. It was found that Fleiss' kappa for the boundary agreement on conversational speech is 0.396 ($p < .05$), and Fleiss kappa for the agreement on scientific speech is 0.56 ($p < .05$). The result suggests that participants agree on boundaries in scientific speech more, since probably conversational speech has more variation in syntactic and prosodic structure.

As there was some agreement for both types of speech, it was valid to proceed with investigating the boundary tone.

3.3. Analysis of boundary tone

This analysis was done to understand how the boundary tone influences the segmentation choices in different types of speech. The hypothesis was that boundary tone will be the predominant cue for segmentation in scientific speech, whereas for conversational speech the boundary tone will be not significant.

The dependent variable was boundary frequency, i.e., how many people marked a boundary in a particular place. The independent variable was the boundary tone (high / low / none).

Boundary tone was fitted to a binomial logistic regression with mixed effects (fixed + random intercept [extract]). The full model included one fixed effect: boundary tone as a factor with three levels (high, low and none), with no tone as reference level; and extract as random intercept. The model was fitted using the lme4 package [13], separately for each kind of speech. The summaries of the model coefficients can be seen in Table 2 (conversational) and Table 3 (scientific).

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-6.443	1.001	-6.438	1.21e-10 ***
tonalH	6.063	1.064	5.701	1.19e-08 ***
tonalL	5.125	1.022	5.017	5.26e-07 ***

Table 2: Logistic regression coefficients for conversational speech.

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-4.9819	0.5017	-9.930	< 2e-16 ***
tonalH	4.9213	0.6108	8.057	7.79e-16 ***
tonalL	4.8641	0.5468	8.896	< 2e-16 ***

Table 3: Logistic regression coefficients for scientific speech.

The regressions revealed a significant influence of boundary tone on placing a boundary, whether it is high, low, or none, for both scientific and conversational speech. TukeyHSD comparison of means was used to assess the differences in influence between these levels. The results are in Table 4 (conversational) and Table 5 (scientific).

	Estimate	Std. Error	z value	Pr (>|z|)
H - NONE == 0	6.0632	1.0635	5.701	< 1e-04 ***
L - NONE == 0	5.1253	1.0217	5.017	< 1e-04 ***
L - H == 0	-0.9378	0.4145	-2.262	0.0549

Table 4: TukeyHSD for conversational speech.

The results of TukeyHSD show that in both types of speech, boundary frequency is significantly affected by the presence [high and low] and absence [none] of the boundary tone. However, there are no differences between the low and high tones, i.e., it does not matter how prominent the boundary tone is to place a boundary.

	Estimate	Std. Error	z value	Pr (>|z|)
H - NONE == 0	4.92127	0.61078	8.057	< 1e-06 ***
L - NONE == 0	4.86411	0.54676	5.896	< 1e-06 ***
L - H == 0	-0.05716	0.41054	-0.139	0.989

Table 5: TukeyHSD for scientific speech.

4. DISCUSSION

Current study aimed to investigate the role of boundary type in chunking conversational and scientific speech. Extracts of spoken speech were played back to participants who were asked to mark segments in corresponding transcripts shown on a screen using ChunkitApp. Boundary tone (rise and fall in pitch) was found to be an important cue for pacing a boundary marker in both types of speech. This suggests that spontaneous speech processing is affected by the boundary tone independently of the nature of speech. Besides, it did not matter how strong the boundary tone is - if it is present at all, a listener would most probably mark it as a boundary, i.e., the beginning or the end of a chunk.

This study, of course, has its limitations. First, the annotation convention is largely based on the ToBI annotation system which marks the boundary tone if the word carrier is at the end of an intonational phrase or is followed by a detectable pause (i.e., BI No.4 in ToBI). Therefore, it is not certain that it is the boundary tones that participants used as the predominant cues for chunking since boundary tone and pause are always shown up together. A different annotation convention could be proposed to explore the roles of pauses versus boundary tones in chunking.

Secondly, theories in intonational phonology and discourse analysis often consider L-tone (i.e., L%, low static or falling pitch contours) as the default boundary tone. In turn, H-tone (i.e., H%, high static or rising pitch contours) is usually marked for indicating a question or the continuation of speech with pauses in between during turn taking in conversation. Due to this function of H%, it was initially hypothesised that participants were more likely to mark a boundary when it is associated with a H%, rather than L%. However, the difference between them was not found to be significant. An interesting continuation of this study would be to have the same task conducted the by musicians or linguists who regularly have much more experience with pitch fluctuations.

Finally, apart from acoustic cues, other linguistic factors such as words with indications may also be used as important cues in chunking. A potential future study could be to explore whether and how lexical indicators are used as boundary markers.

5. ACKNOWLEDGEMENTS

This study was conducted as a part of an Experimental Laboratory course at the University of Helsinki, Finland. We thank the Department of Digital Humanities and the course leader Juraj Šimko for the environment that inspired this study.

6. REFERENCES

[1] Christiansen, M. H. & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *The Behavioral and Brain Sciences, 39*, E62.
[2] Sinclair, J. M., & Mauranen, A. (2006). Linear unit grammar: Integrating speech and writing (Vol. 25). John Benjamins Publishing.
[3] Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2022). Perceptual chunking of spontaneous speech: Validating a new method with non-native listeners. Research Methods in Applied Linguistics, 1(2), 100012.
[4] Anurova, I., Vetchinnikova, S., Dobrego, A., Williams, N., Mikusova, N., Suni, A., ... & Palva, S. (2022). Event-related responses reflect chunk boundaries in natural speech. *NeuroImage, 255*, 119203.
[5] Dobrego, A., Konina, A., & Mauranen, A. (2022). Continuous speech segmentation by L1 and L2 speakers of English: the role of syntactic and prosodic cues. *Language Awareness*, 1-21.
[6] Konina, A., Dobrego, A., Mauranen, A. (forthc.) Natural speech segmentation strategies: cross-linguistic evidence.
[7] Vetchinnikova, S., Mauranen, A., & Mikušová, N. (2017). ChunkitApp: Investigating the relevant units of online speech processing. In INTERSPEECH 2017: 18th Annual Conference of the International Speech Communication Association. pp. 811– 812. Interspeech 2017, Stockholm, Sweden.
[8] Cavalcante, F. A. & Ramos, A. C. (2016). The American English spontaneous speech minicorpus: Architecture and comparability. *CHIMERA. Romance Corpora and Linguistic Studies 3.2*, 99–124
[9] Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.55, retrieved 25 October 2021 from http://www.praat.org/
[10] R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[11] RStudio Team. (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.
[12] Fleiss, D. J., DiGiovanni, C. W., Sangeorzan, B. J., Hansen Jr, S. T., Kuo, R., Tejwani, N., & Price, R. (2003). Diagnostic tests for gastrocnemius tightness. *JBJS, 85*(4), 760.
[13] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. (2014). Fitting linear mixed-effect models using lme4, *Journal of Statistical Software*

7. APPENDIX

Link to the document containing links to all sources for scientific, practice and training stimuli:

https://drive.google.com/file/d/19dzSonA9aujUj1FMyKUhHg_KowzZDhns/view?usp=share_link

TALKER VARIABILITY IN CROSS-DIALECT LEXICAL PROCESSING

Jory Ross, Cynthia G. Clopper

Ohio State University
ross.1589@osu.edu, clopper.1@osu.edu

ABSTRACT

Listeners must adapt to immense variability in the speech signal. This study examined the effects on lexical activation of two sources of variability: dialect variation and individual talker variability. A cross-modal lexical decision task revealed robust evidence of lexical activation of competing /æ ɛ/ minimal pairs for the Northern dialect of American English, but not of competing /ɑɪ ɑ/ minimal pairs for the New England dialect of American English. This dialect difference in lexical activation likely reflects greater phonetic ambiguity for the Northern vowel pair and greater phonological confusion for the New England vowel pair. Unexpectedly, performance did not differ as a function of either the number of different talkers in each experimental block or whether the talkers within each block were from the same or different dialects. The cross-modal lexical decision task, which requires a response to a visual target, may not be sensitive to these talker variability effects.

Keywords: dialect variation, talker variability, lexical competition, cross-modal priming

1. INTRODUCTION

Talker variability, both within and across dialects, affects lexical processing. Across dialects, listeners process familiar dialects faster and more accurately than unfamiliar dialects in a range of tasks [1, 2, 3]. Listeners also process prestigious dialects faster and more accurately than non-prestigious dialects [4, 5, 6]. This processing benefit for prestigious dialects may reflect dialect familiarity through media exposure and/or the social status associated with prestigious forms. For example, Sumner et al. [7] have proposed that prestigious forms are more robustly encoded in memory, granting them a processing advantage over non-prestigious forms.

Within dialects, listeners process speech produced by familiar talkers more accurately than speech produced by unfamiliar talkers [8, 9]. In addition, trial-to-trial talker variability increases demands on attention [10] and working memory [11] during speech processing tasks. Thus, as the number of talkers within a speech processing task increases, so do response times [12]. Notably, effects of talker variability emerge even if the acoustic differences between talkers are minimal [13] or the dimensions of variability are irrelevant to the task [12].

Processing demands due to talker variability are further magnified if the talkers have different dialects. For example, Clopper [1] observed longer response times in mixed-dialect blocks than same-dialect blocks in a speeded lexical classification task, suggesting that dialect variation imposes processing demands over and above talker variability effects.

The goal of the current study was to investigate the joint influence of dialect variation and talker variability on lexical activation. To this end, we examined priming in a cross-modal lexical decision task with auditory primes produced in two regional dialects, New England and Northern American English, in mixed-dialect and same-dialect conditions. Both dialects were relatively unfamiliar to the listeners to allow processing costs of both prime dialect and talker variability to emerge.

In a previous cross-modal lexical decision task, Clopper and Walker [14] presented Midland American English listeners with Northern primes that contained perceptually confusable /æ ɛ/. Matching auditory primes facilitated access to the visual target words, but competing minimal pair primes, such as auditory prime /blæst/ preceding visual target *blessed*, inhibited lexical access. Similarly, in an auditory form priming lexical decision task, Sumner and Samuel [6] presented General American English listeners with New York City English primes that contained perceptually confusable non-rhotic forms (e.g., /ɚ/ realized as [ə]). Matching primes facilitated access to General American targets, but facilitation was reduced for New York City non-rhotic primes. Sumner and Samuel [6] did not consider competing minimal pairs in their design.

The current study included Northern /æ ɛ/ auditory primes and visual targets, as in Clopper and Walker's [14] study, as well as New England /ɑɪ ɑ/ auditory primes and visual targets, parallel to Sumner and Samuel's [6] study. Like New York City English, New England American English is non-rhotic, leading to perceptually confusable /ɑɪ ɑ/, as in minimal pairs such as *card* and *cod*. We expected to replicate the previous cross-dialect lexical processing findings [6, 14] showing facilitation for matching primes and inhibition for competing primes. We also expected to observe slower response times overall in the mixed-dialect condition relative to the same-

dialect condition, as in previous work [1]. Given that slower response times tend to be more variable than faster response times [15], we expected to observe weaker overall facilitation and inhibition in the slower, more variable mixed-dialect condition than in the faster, less variable same-dialect condition.

2. METHOD

2.1. Participants

Participants were recruited through Prolific, an online research recruitment platform. Data from 98 participants (female = 42, male = 53, non-binary = 1, unreported = 2) were included in the analysis. All participants were native speakers of American English, born in the United States, reported normal or corrected-to-normal vision, and no history of speech, language, or hearing disorders. Participants ranged in age from 18-68 years old (M = 30.7 years). Participants' residential histories varied, but none had lived in either the Northern or New England dialect regions, so that both prime dialects were equally unfamiliar. Although participants had limited direct experience with the stimulus dialects, these dialects are perceptually distinctive to listeners who have not lived in these regions [16, 17]. All participants passed at least one of two attention checks during the experiment. They were asked to wear headphones while completing the task, and 91 of the included participants reported that they did so.

2.2. Stimulus materials

The stimulus materials comprised auditory prime words and visual target words and nonwords in English. The primes and targets contained one of four target stressed vowels /æ ɛ ɑɪ ɑ/. These vowels were selected because /æ/ is shifted in the Northern dialect to be confusable with /ɛ/ and /ɑɪ/ is non-rhotic in the New England dialect and confusable with /ɑ/. The real words had a mean familiarity rating of at least 5.5 out of 7 in the Hoosier Mental Lexicon [18].

The auditory prime words were produced by two Northern talkers (one female, one non-binary) and two New England talkers (one female, one male). The Northern talkers were recorded in a sound-attenuated booth at Ohio State University in Columbus, OH, and the New England talkers were recorded in a sound-attenuated booth at the University of Massachusetts in Amherst, MA.

The cross-modal lexical decision task included three critical trial types, as shown in Table 1. On Matching trials, the auditory prime and visual target were the same word, containing either /æ/ or /ɑɪ/. On Competing trials, the prime and target were minimal pairs. The prime contained /æ/ or /ɑɪ/ and the target was its minimal pair with /ɛ/ or /ɑ/, respectively. Thus, all Matching and Competing critical trials contained primes with potentially confusable /æ/ or /ɑɪ/. On Unrelated trials, the prime and target were phonologically and semantically unrelated.

Trial Type	Vowel Contrast	Prime	Target
Matching	/æ ɛ/	mass	mass
	/ɑɪ ɑ/	sharp	sharp
Competing	/æ ɛ/	blast	blessed
	/ɑɪ ɑ/	card	cod
Unrelated	/æ ɛ/	spar	fed
	/ɑɪ ɑ/	hatch	scoff

Table 1: Examples of critical trial primes and targets in the cross-modal lexical decision task.

The experiment included eight Matching trials, eight Competing trials, and 16 Unrelated trials for each vowel contrast, for a total of 64 critical trials. In addition, 128 filler trials were presented, including 32 unrelated trials with a word target, 32 competing trials with a nonword target, and 64 unrelated trials with a nonword target. No primes or targets were repeated within-listener, although the same word could appear as both a prime and a target either within (Matching) or across (Competing, Unrelated) trials for the same listener. Critical primes and targets were counterbalanced for trial type across three experimental lists. Each listener was presented with a single list. Within each list, all four vowels in all trial types were counterbalanced across talkers.

2.3. Procedure

Participants completed the cross-modal priming task on their own personal computers. On each trial, they were presented with an auditory prime, and after a 50 ms interstimulus interval, a visual orthographic target. Participants indicated whether the visual target was a real word or nonword in English by pressing 'f' or 'j' on their keyboard, respectively.

Participants were randomly assigned to one of three blocking conditions. In the Blocked condition (N = 22), listeners were presented with one block containing 96 trials with all primes produced by the two Northern talkers and a separate block of 96 trials with all primes produced by the two New England talkers. In the Mixed condition (N = 24), listeners were presented with primes produced by all four talkers throughout both blocks of the experiment. To ensure that any observed block effects were not due to the number of talkers presented within each block, in the Control condition (N = 52), listeners were presented with two blocks, each with one of the two talkers from each of the two dialects. Two versions of

the Control condition were presented to fully counterbalance talker pairings within blocks. Block order was counterbalanced across participants in the Blocked and Control conditions. Trial order was randomized within each block. Participants were permitted to take a break between the two blocks.

2.4. Analysis

Overall mean accuracy in the cross-modal lexical decision task was 93%, so the analysis focused on reaction times to correct trials. Prior to analysis, trials with reaction times shorter than 250 ms or longer than 2500 ms were excluded. Trials were then excluded if they were more than 2.5 standard deviations above or below the mean reaction time for each participant, target word, or prime token. Prime tokens with mean accuracy below 75% were also excluded (total excluded N = 255 trials, 1.5%).

The comparisons of interest were between the Matching and Unrelated critical trials, for which we expected to observe facilitation, and between the Competing and Unrelated critical trials, for which we expected to observe inhibition. Linear mixed-effects regression models were built to explore these trial type comparisons. The facilitation model included the Matching and Unrelated critical trials with /æ ɑɪ/ targets. The inhibition model included the Competing and Unrelated critical trials with /ɛ ɑ/ targets.

In both models, log-transformed response times were predicted by trial type (Matching/Competing, Unrelated), prime dialect (Northern, New England), condition (Blocked, Mixed, Control), vowel contrast (/æ ɛ/, /ɑɪ ɑ/), and all interactions. The maximal data-driven random effects by participants, target words, and prime tokens were used [19]. Statistical significance was determined using the Satterthwaite approximation of degrees of freedom for *F*- and *t*-statistics via the *lmerTest* package in R [20].

3. RESULTS

3.1. Facilitation

The facilitation analysis revealed significant main effects of trial type ($F(1, 151) = 39.9$, $p < .001$), prime dialect ($F(1, 86) = 6.8$, $p = .011$), and vowel contrast ($F(1, 47) = 8.3$, $p = .006$), as well as a significant trial type x prime dialect interaction ($F(1, 150) = 15.5$, $p < .001$). No effects or interactions involving blocking condition were significant. As expected, responses were faster to Matching trials (M = 668 ms) than Unrelated trials (M = 712 ms) overall, consistent with facilitation. Responses were also faster following Northern primes (M = 679 ms) than New England primes (M = 702 ms) and for /æ/ targets (M = 666 ms) than /ɑɪ/ targets (M = 716 ms). The significant interaction is shown in Fig. 1. Although the main effect of trial type was significant, post-hoc estimated marginal means comparisons confirmed significant facilitation on Matching trials following Northern primes only ($t(158) = -7.0$, $p < .001$).

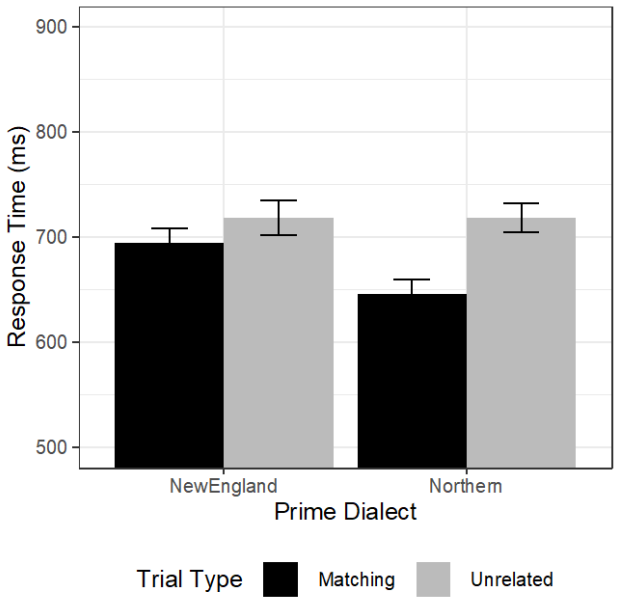


Figure 1: Mean response times for Matching and Unrelated trials following New England and Northern primes. Error bars are standard error of subject means.

3.2. Inhibition

The inhibition analysis revealed a significant main effect of prime dialect ($F(1, 93) = 4.3$, $p = .041$) and a significant trial type x prime dialect interaction ($F(1, 190) = 11.8$, $p < .001$). No effects or interactions involving blocking condition were significant. Unlike in the facilitation analysis, responses were faster following New England primes (M = 689 ms) than Northern primes (M = 709 ms). The significant interaction is shown in Fig. 2. As in the facilitation analysis, post-hoc estimated marginal means comparisons confirmed significant inhibition on Competing trials following Northern primes only ($t(209) = 3.0$, $p = .003$).

4. DISCUSSION

The analysis uncovered the expected effects of facilitation and inhibition following Northern primes, but not following New England primes. No effects or interactions involving blocking condition were observed in either analysis. The results therefore replicated the facilitation and inhibition observed for Northern /æ ɛ/ by Clopper and Walker [14] but failed to extend these findings to New England /ɑɪ ɑ/.

The lack of facilitation for the New England non-rhotic primes in our study contrasts with Sumner and

Samuel's [6] finding of facilitation for non-rhotic New York City primes. This difference may reflect the phonological confusability of our materials, which included non-rhotic forms with minimal pair competitors, such as target *card* with competitor *cod*. Sumner and Samuel's [6] materials involved non-rhotic forms without minimal pair competitors, such as target *baker* with no real word competitor */beikə/. Our New England non-rhotic primes likely activated phonological competitors with /ɑ/, reducing the benefits of the matching prime for the visual target, which contained orthographic <r>.

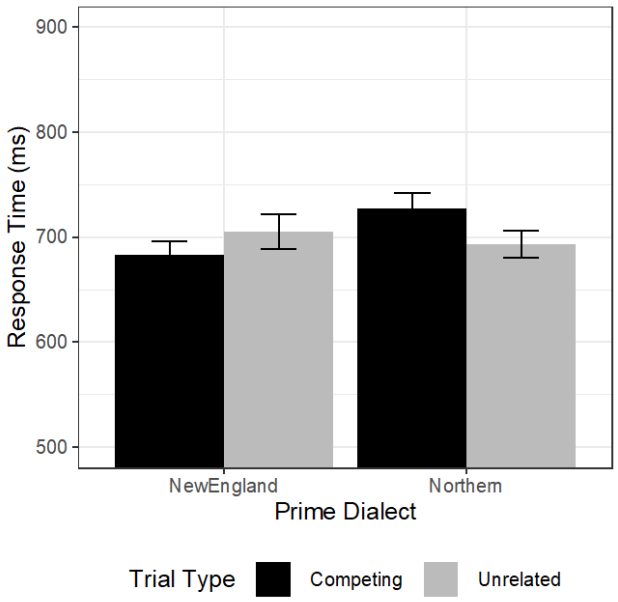


Figure 2: Mean response times for Competing and Unrelated trials following New England and Northern primes. Error bars are standard error of subject means.

Although Sumner and Samuel [6] observed facilitation for non-rhotic New York City primes and General American targets, the magnitude of this facilitation was less than for General American primes and targets. Moreover, no facilitation was observed in their study for non-rhotic New York City targets, regardless of prime. Their results suggest relatively weak lexical activation for the non-rhotic forms for General American listeners for whom the New York City dialect is unfamiliar. Our results are broadly consistent with this weak overall activation for unfamiliar non-rhotic forms.

The facilitation analysis revealed faster responses overall following Northern primes than New England primes and for /æ/ targets than for /ɑɪ/ targets. The inhibition analysis further revealed faster responses overall following New England primes than Northern primes. These main effects of prime dialect and vowel contrast provide insight into the dialect differences in facilitation and inhibition. In both cases, the non-rhotic New England primes had less impact overall on response times than the Northern /æ/ primes, further suggesting less robust lexical activation for the New England primes than the Northern primes.

One explanation for the asymmetric effect of dialect on facilitation and inhibition could be the negatively stereotyped status of non-rhotic forms in the New England dialect in comparison to the non-stereotyped status of Northern /æ/ [21, 22]. Sumner and Kataoka [23] found semantic priming for prestigious British English non-rhoticity, but no semantic priming for negatively stereotyped New York City non-rhoticity for American English listeners, suggesting weaker lexical activation for the non-prestigious variant.

However, Clopper [1] found that negatively stereotyped dialects may be more robustly encoded than non-stereotyped dialects, predicting processing advantages for New England non-rhotic forms relative to Northern /æ/ variants. Moreover, Clark et al. [24] observed facilitation for both non-stereotyped Midland American English forms and negatively stereotyped Southern American English forms in the same cross-modal priming task as in the current study. Thus, stereotypes and prestige may not be the critical factors underlying differences in lexical processing among unfamiliar dialects.

An alternative explanation for the differences in lexical activation between the New England and Northern primes could be the level of representation that each contrast involves. The New England non-rhotic forms may map more directly onto a competing phonological form than the shifted Northern /æ/, such that New England /ɑɪ/ is more likely to be perceived as /ɑ/ than Northern /æ/ is to be perceived as /ɛ/. Thus, the phonetically ambiguous Northern /æ/ primes might produce greater effects because words with both /æ/ and /ɛ/ are strongly activated, whereas the New England non-rhotic primes might primarily activate phonological competitors with /ɑ/, reducing both facilitation for matching primes and inhibition for competing minimal pair primes.

Unexpectedly, the results uncovered no effects of blocking condition on response times. We expected slower performance in the Mixed condition with four talkers per block than in the Blocked and Control conditions with only two talkers per block [12]. Previous studies showing talker variability effects have used speeded lexical classification tasks [1, 12], whereas the current study used a cross-modal lexical decision task. Talker variability will have the greatest effect on processing of the auditory prime, and that effect may be attenuated or have subsided by the time listeners process the visual target. Consequently, the use of a task requiring responses to visual targets in the current study may have reduced our ability to observe effects of talker variability on performance.

5. ACKNOWLEDGMENTS

This project was partially supported by the National Science Foundation (BCS-1843454). We would like to thank Meghan Armstrong-Abrami and Joe Pater for assistance with collecting the New England speech stimuli.

6. REFERENCES

[1] Clopper, C. G. 2017. Dialect interference in lexical processing: Effects of familiarity and social stereotypes. *Phonetica* 74, 25–59.

[2] Floccia, C., Goslin, J., Girard, F., Konopczynski, G. 2006. Does a regional accent perturb speech processing? *J. Exp. Psych. Human Percep. Perf.* 32, 1276–1293.

[3] Labov, W., Ash, S. 1997. Understanding Birmingham. In: Bernstein, C., Nunnally, T., Sabino, R. (eds.), *Language Variety in the South Revisited*. University of Alabama Press, 508–573.

[4] Clopper, C. G., Bradlow, A. R. 2008. Perception of dialect variation in noise: Intelligibility and classification. *Lang. Speech* 51, 175–198.

[5] Dufour, S., Nguyen, N., Frauenfelder, U. H. 2007. The perception of phonemic contrasts in a non-native dialect. *J. Acoust. Soc. Am.* 121, EL131–EL136.

[6] Sumner, M., Samuel, A. G. 2009. The effect of experience on the perception and representation of dialect variants. *J. Mem. Lang.* 60, 487–501.

[7] Sumner, M., Kim, S. K., King, E., McGowan, K. B. 2014. The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Front. Psych.* 4(1015), 1–13.

[8] Nygaard, L. C., Pisoni, D. B. 1998. Talker-specific learning in speech perception. *Percept. Psychophys.* 60, 355–376.

[9] Nygaard, L. C., Sommers, M. S., Pisoni, D. B. 1994. Speech perception as a talker-contingent process. *Psychol. Sci.* 5, 42–46.

[10] Wong, P. C. M., Nusbaum, H. C., Small, S. L. 2004. The neural basis of talker normalization. *J. Cogn. Neurosci.* 16, 1173–1184.

[11] Nusbaum, H. C., Morin, T. 1992. Paying attention to differences among talkers. In: Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, Y. (eds.), *Speech Perception, Production and Linguistic Structure*. IOS Press, 113–134.

[12] Mullennix, J. W., Pisoni, D. B. 1990. Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys.* 47, 379–390.

[13] Magnuson, J. S., Nusbaum, H. C. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psych. Human Percep. Perf.* 33, 391–409.

[14] Clopper, C. G., Walker, A. 2017. Effects of lexical competition and dialect exposure on phonological priming. *Lang. Speech* 60, 85–109.

[15] Wagenmakers, E.-J., Brown, S. 2007. On the linear relation between mean and standard deviation of a response time distribution. *Psych. Rev.* 114, 830–841.

[16] Clopper, C. G., Pisoni, D. B. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *J. Phon.* 32, 111–140.

[17] Clopper, C. G., Pisoni, D. B. 2004. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Lang. Var. Change* 16, 31–48.

[18] Nusbaum, H. C., Pisoni, D. B., Davis, C. K. 1984. Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In: *Research on Speech Perception Progress Report No. 10.* Speech Research Laboratory, Indiana University, 357–376.

[19] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255–278.

[20] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Software* 82(13), 1–26.

[21] Campbell-Kibler, K. 2012. Contestation and enregisterment in Ohio's imagined dialects. *J. Eng. Linguist.* 40, 281–305.

[22] Niedzielski N., Preston D. R. 2000. *Folk Linguistics*. Mouton de Gruyter.

[23] Sumner, M., Kataoka, R. 2013. Effects of phonetically-cued talker variation on semantic encoding. *J. Acoust. Soc. Am.* 134, EL485–EL491.

[24] Clark, H., Bissell, M., Clopper, C. G., Walker, A. 2022. Effects of lexical competition, dialect familiarity, and dialect exposure on lexical processing. Poster presented at Laboratory Phonology 18, virtual, June 23-25.

SOUND CATEGORIZATION AFTER SPEAKING WITH A BITE BLOCK

Xinyu Zhang, Rob Schoonen, Esther Janse

Radboud University
xinyu.zhang@ru.nl, rob.schoonen@ru.nl; esther.janse@ru.nl

ABSTRACT

Earlier research has investigated how speech sound categorization is affected by displacement of the articulators during listening. We investigated perceptual behavior *following* a manipulation to restrict tongue raising, and tested whether the experience of having spoken with a bite block affects subsequent categorization of an /ɪ/-to-/ɛ/ continuum. Furthermore, we investigated whether being able to hear one's own manipulated production affects this potential perceptual recalibration (by including a condition with noise-masking during production). Surprisingly, participants in the bite block group gave fewer /ɛ/ responses than the no bite block group at pretest, but response patterns were similar for the two groups at posttest. This suggests that anticipating speaking with a bite block affected categorization behavior more than the actual speech experience. As such, the results do not provide evidence that sound representations change due to articulator displacement.

Keywords: speech perception, perceptual recalibration, speech production

1. INTRODUCTION

The much-debated motor theory of speech perception [1] proposed that part of perceiving speech sounds is perceiving the articulatory movement (but cf. e.g., [2, 3] for critical notes). Later empirical studies, using e.g., 'virtual-lesion' methods like rTMS [4]), showed that disruption of premotor cortex during perception impairs speech perception. But can changes in speech production experience induce changes in speech perception?

Several studies have perturbed the articulators *while* participants perform a perception task in order to investigate the interaction between speech production and perception. Yeung and Scott [5] had participants breathe either through the mouth or nose while giving nasality judgments to auditory stimuli. Participants were more likely to judge a sound as 'nasal' when breathing through the nose. Similarly, Trudeau-Fissette et al. [6] applied facial stretch in the direction of the mouth shape as if producing the vowel /e/, and observed that participants perceived more stimuli as /e/ when there was skin stretch, as compared to no skin stretch.

Another study investigated the subsequent perceptual effect *following* a manipulation of the articulators [7]. Sato and colleagues investigated perceptual categorization of a noise-masked /pa/ vs. /ta/ before and after articulator training, with half of their participants getting lip training (repeatedly protruding lips as if making a kiss-like gesture) and the other half tongue training (repeatedly pressing tongue to palate). Following training, the 'lip group' was biased towards perceiving /p/, whereas the 'tongue group' was biased towards /t/. Whereas Sato and colleagues engaged speakers' articulators, their training did not involve speech production.

Other studies manipulated speakers' auditory feedback during a speech task [8, 9]. Using altered auditory feedback (AAF), Shiller and colleagues [8] upshifted speakers' auditory feedback of centre of gravity of /s/, inducing a compensatory downshifted /s/. Subsequent sound categorization aligned with the altered auditory input during speaking: those who had had the AAF gave more /s/ responses than at pretest, which was not observed for a control group (but cf. [9] for different results).

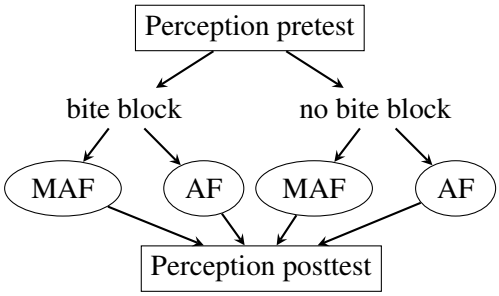
The current study used a bite block to alter speech production, impairing tongue height distinctions. More specifically, we tested for a potential recalibration in perceptual categorization of a vowel height contrast. Importantly, as our focus was on *subsequent* rather than *simultaneous* effects of articulator displacement, the bite block was removed during perceptual testing. Additionally, in line with earlier research on what drives perceptual change [8, 9], we isolated the effect of hearing one's altered speech from the motor experience associated with the altered production, by noise-masking the auditory feedback during speech production for half of our participants. Thus, we address the research question of whether having full access to auditory feedback of one's own altered speech matters. Lastly, if the bite block mainly affects response bias rather than perceptual representations per se [7], the bite-block effect might be strongest for the most ambiguous continuum steps.

2. METHODS

2.1. Experimental Design

Sound categorization was compared between pretest and posttest and was always tested without bite block obstruction. A production task was administered between pretest and posttest, in which participants read aloud a set of pseudowords from a computer screen (see §2.3.1). Participants were randomly assigned to a bite-block or no-bite-block (control) group. Within each of the two groups, participants would either have access to normal auditory feedback of their own productions, or would hear speech-shaped noise to mask their production. Figure 1 illustrates the experimental design[1].

Figure 1: Experimental design. Ellipses represent the production task; rectangles represent perception tasks; MAF: masked auditory feedback; AF: regular auditory feedback.



2.2. Participants

Sixty participants aged between 18 and 30 years were recruited through the Radboud Research Participation System. They were randomly assigned to one of the four experimental Biteblock-Auditory Feedback conditions (see Figure 1). All were native speakers of Dutch, with normal or corrected-to-normal vision, and they all reported having normal hearing and speech abilities (i.e., none reported stuttering or any other speech problems), and none reported dyslexia. They were all compensated for their participation.

2.3. Materials

2.3.1. Production stimuli

The target vowels for the perception and production tasks were /ɪ/ and /ɛ/. The production stimuli were bisyllabic pseudowords conforming to Dutch phonotactics. The syllable structure for the pseudowords (targets and fillers alike) was C1VC2-/ɪf/, such that the target vowel is always embedded in a stressed CVC structure (e.g., /ˈtɪsɪf/, /ˈtɛsɪf/). The addition of the standard second syllable made the pseudowords less 'wordlike'. Care was also taken that no C1VC2 combination (either in targets or in fillers) formed real Dutch morphemes to further minimize lexical frequency effects. For target stimuli, in order to allow easy segmentation of the target vowels (for the production analysis not reported here), C1 and C2 were voiceless plosives and fricatives. Filler stimuli followed the same syllable structure, and contained more diverse vowels in the stressed V position, as well as more diverse consonants in the C1 and C2 positions with other manners of articulation. The places of articulation of C1 were evenly distributed in the target items: /p/, /t/, and /k/ each appeared twice. Each of the stop consonants was paired with each of the vowels (3∗2 = 6 target stimuli). In addition to the six target stimuli, the production set contained 12 filler stimuli, all repeated three times [(6 targets + 12 fillers)∗3 repetitions = 54 items for production].

2.3.2. Masking noise

Speech-shaped noise was used as masking noise for the masking-noise-as-auditory-feedback condition. Speech-shaped noise was created by filtering random noise with a filter made by averaging the long-term average spectrum of 10 male and 10 female adult speakers reading a phonetically balanced text. The resulting 25-minutes-long noise file, with a sampling frequency of 44100 Hz, was played during the entire production block for those in the masked auditory feedback (MAF) groups.

2.3.3. Perception Stimuli

The perception pretest and posttest blocks contained three continua of seven steps of the target contrast /ɪ/ to /ɛ/, as well as nine filler continua of seven steps [(3 targetContrasts + 9 fillerContrasts) ∗ 7 stepsEach = 84 items in total]. The end points of each (target or filler) continuum were a monosyllabic real-word minimal pair in Dutch (e.g., [pɪp]-[pɛp]). Each step of each continuum was presented once in each perception block. Three types of filler continua were included: different vowel formant continua (3 pairs, e.g., [biːr] - [byːr]), consonant place of articulation continua (3 pairs, e.g., [tau] - [kau]), and voice onset time continua (3 pairs, e.g., [toːs] - [doːs]).

2.4. Procedures

The experiment was written and run in Open Sesame [10]. All procedures were carried out at the Centre for Language Studies Lab of Radboud University in one experimental session that took less than 20 minutes in total. Upon arrival in the lab, participants were informed that they were assigned to the bite-block group or to the control (no bite-block) group. Before the start of the perception pretest, participants who were assigned to the bite block group received instructions on how to insert the bite block. Instructions included verbal description (keep tongue flat under the bite block, where to place the incisors), and being shown a photo of the experimenter with the bite block in the mouth. Importantly, participants did not put the bite block in their mouths, nor tried speaking with it, before the perception pretest.

The bite block used in the experiment was an unopened plastic bottle of a probiotic drink (75 mm tall, smallest circumference 24 mm, largest circumference 38 mm), with the original 65 ml of product still inside, see Figure 2. Each participant assigned to the bite-block group used a new bottle which they could take with them afterwards.

In each perception block, participants were presented with tokens from different contrast continua and asked to decide which of the words of the real-word minimal pair they heard. At the start of each trial, two options (the two words of the pair) appeared on the screen 500 ms before the audio was presented. After the initial 500ms, the two options remained on the screen while an auditory token (one step from the corresponding auditory continuum) was played through a pair of Sennheiser HME 110 headphones. Participants were then asked to indicate which word they heard, by pressing either the left or right key on a button box, corresponding to the left or right option on the screen, respectively[2]. The order of the stimuli was randomized per participant and per block.

The production task was self-paced by key press. The pseudoword prompts were displayed on a BenQ PD2700U computer screen (font size 32, resolution 1920 * 1080). Audio recordings were made using a Sennheiser ME 64 microphone and stored per token. Participants went through a practice phase of five items for the perception task before completing the perception pretest. The order of the production stimuli was randomized per participant and per block. Masking noise during the production block was played through a pair of Avantree ANC 031 active noise canceling headphones.

Figure 2: Bite block used in the present study.



2.5. Statistical Analysis

For each trial, the randomized stimuli and the response by the participant were automatically recorded by the experimental interface (§2.4). All analyses were carried out in R (version: 4.0.4) [11]. A generalized logistic regression mixed-effects model was run on the data using the lme4 package [12], with the perceptual response as the dependent variable. Binary contrasts were coded using treatment coding. Predictor variables include continuum step (7 steps, centered on step 4), auditory feedback condition (levels AF, MAF; with AF mapped on the intercept), bite block condition (levels BB, NB; with NB mapped on the intercept), as well as time of measurement (levels pretest, posttest; with pretest mapped on the intercept). We included random intercepts for participant and minimal pair. The analysis started with the four-way interaction model including all predictor variables[3]: $responseE \sim cstep * prepost * AF * BB + (1|subject) + (1|pair)$. Insignificant effects were taken out in a step-wise manner (taking out higher-order interactions before lower-order ones, followed by removal of simple effects) to arrive at the most parsimonious model. Model comparisons were applied using the anova() function in R after each removal step to verify that exclusion of an effect did not lead to significantly worse model fit.

3. RESULTS

Figure 3 shows categorization behavior at pretest and posttest for bite block and no bite block groups. The most parsimonious model of the likelihood of getting an /ɛ/ response was: $responseE \sim cstep * AF * BB + prepost * BB + AF * BB + prepost * AF + (1|subject) + (1|pair)$. Model estimates are provided in Table 1.

As expected, participants were more likely to give an /ɛ/ response for the higher steps on the continuum (β = 1.970 (0.158), $p < 0.001$), and the no bite block group (on the intercept of the logistic regression model) did not change their categorization behavior from pretest to posttest (β = -0.368 (0.258), *n.s.*). At pretest, participants in the bite block group were less likely to give an /ɛ/ response than the no

Figure 3: Percentage of /ɛ/ response at pretest and posttest for bite block group (BB) and no bite block group (NB) at each continuum step (collapsing over auditory feedback conditions).

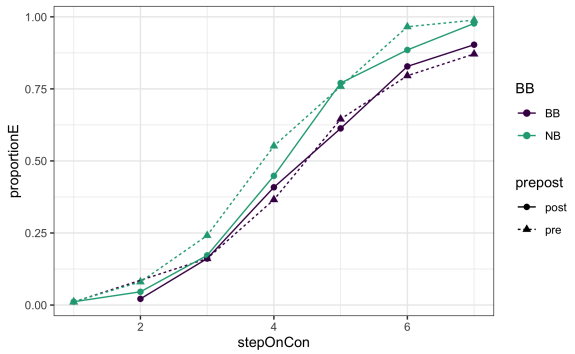
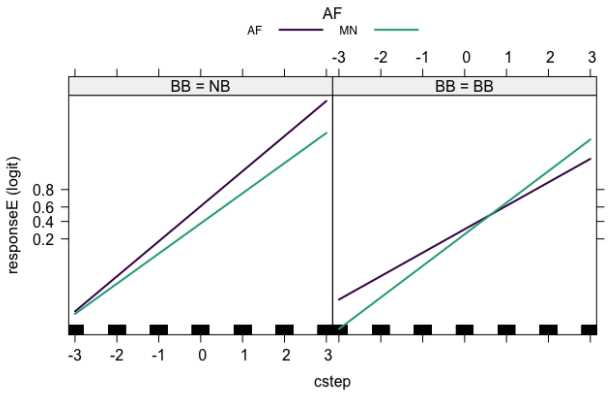


Table 1: Coefficients for sound categorization model; β coefficients represent logits.

	β	SE	z value	p value
(Intercept)	0.632	0.608	1.035	0.300
cstep	1.969	0.158	11.834	<0.001
MAF	-0.820	0.604	-1.357	0.175
BB	-1.603	0.595	-2.675	0.007
posttest	-0.368	0.258	-1.420	0.153
cstep:BB	-0.653	0.179	-3.637	<0.001
cstep:MAF	-0.276	0.196	-1.407	0.159
MAF:BB	0.682	0.818	0.833	0.405
BB:posttest	0.625	0.280	2.229	0.026
cstep:MAF:BB	0.733	0.255	2.870	0.004

bite block group (β = -1.603 (0.595), $p < 0.05$). The significant interaction between posttest and bite block condition ($\beta = 0.625$ (0.280), $p < 0.05$) shows that the difference between the groups is smaller at posttest. This was verified by running the same model with posttest (rather than pretest) mapped on the intercept, showing no bite block effect at posttest. For the no bite block group, masking noise did not affect categorization behavior at the center of the continuum (β = -0.820 (0.604), *n.s.*). This null effect of masking noise also holds for the other continuum steps (β = -0.276 (0.196), *n.s.*). For the bite block group, there was an effect of masking noise on categorization behaviour, but only at the more extreme continuum steps (as evident from the triple interaction in Table 1, and as visualized in Figure 4, $\beta = 0.6733$ (0.255), $p < 0.01$). However, as these effects held at pretest and posttest alike, they cannot be attributed to (noise-masking effects on) the actual bite-block production experience, but rather seem to reflect pre-existing group differences between the two bite-block groups (cf. design overview in Figure 1).

Figure 4: Model effects plot showing the three-way interaction (MN: masking noise; AF: regular auditory feedback).



4. DISCUSSION

We set up a sound categorization study with a pretest-posttest design to investigate whether tongue height restriction due to speaking with a bite block affects subsequent categorization of a vowel height contrast continuum. Our second question was whether being able to hear one's own altered speech would affect this potential shift. Noise masking during production was used to address this second point.

Regarding the first research question, the bite block and no bite block groups were expected to differ at posttest, if at all, rather than at pretest. Had our results aligned with studies that tested perception during articulation manipulation [5, 6], those in the bite block group would give more /ɛ/ responses at posttest, as the bite block fixes the tongue to a more /ɛ/-like configuration. Recall that instruction about the bite block occurred before pretest, without speakers actually trying it themselves. It is unclear whether these pretest results would replicate. If so, we can only speculate that those anticipating speaking with tongue restriction mapped their /ɪ/ onto a more open position, hence the 'regular' /ɛ/ might seem more /ɪ/-like to them. However, the observation of no bite block group difference at posttest does not provide reliable evidence that sound representations change due to articulator displacement.

Now that our results have not shown a reliable perceptual shift driven by articulator displacement, our second research question does not apply anymore. Follow-up research is planned to investigate bite block and noise masking effects on speech acoustics, and to test whether individual amount of acoustic change in production predicts perceptual change.

5. REFERENCES

[1] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[2] G. Hickok, "The role of mirror neurons in speech and language processing," *Brain and Language*, vol. 112, no. 1, p. 1, 2010.

[3] A. J. Lotto, G. S. Hickok, and L. L. Holt, "Reflections on mirror neurons and speech perception," *Trends in Cognitive Sciences*, vol. 13, no. 3, pp. 110–114, 2009.

[4] I. G. Meister, S. M. Wilson, C. Deblieck, A. D. Wu, and M. Iacoboni, "The essential role of premotor cortex in speech perception," *Current Biology*, vol. 17, no. 19, pp. 1692–1696, 2007.

[5] H. H. Yeung and M. Scott, "Postural control of the vocal tract affects auditory speech perception." *Journal of Experimental Psychology: General*, vol. 150, no. 5, p. 983, 2021.

[6] P. Trudeau-Fisette, T. Ito, and L. Ménard, "Auditory and somatosensory interaction in speech perception in children and adults," *Frontiers in Human Neuroscience*, vol. 13, p. 344, 2019.

[7] M. Sato, K. Grabski, A. M. Glenberg, A. Brisebois, A. Basirat, L. Menard, and L. Cattaneo, "Articulatory bias in speech categorization: Evidence from use-induced motor plasticity," *cortex*, vol. 47, no. 8, pp. 1001–1003, 2011.

[8] D. M. Shiller, M. Sato, V. L. Gracco, and S. R. Baum, "Perceptual recalibration of speech sounds following speech motor learning," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1103–1113, 2009.

[9] D. R. Lametti, A. Rochet-Capellan, E. Neufeld, D. M. Shiller, and D. J. Ostry, "Plasticity in the human speech motor system drives changes in speech perception," *Journal of Neuroscience*, vol. 34, no. 31, pp. 10 339–10 346, 2014.

[10] S. Mathôt, D. Schreij, and J. Theeuwes, "Opensesame: An open-source, graphical experiment builder for the social sciences," *Behavior Research Methods*, vol. 44, no. 2, pp. 314–324, 2012.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[1] The flow chart presented here was part of a larger experimental design, which included more production blocks and a production analysis.

[2] The option containing '[ɪ]' did not always appear on the left. The '[ɪ]' option was shown on one side of the screen for two of the minimal pairs and on the other side for the other minimal pair.

[3] responseE: the likelihood of getting an /ɛ/ response; cstep: continuum step, centered on step 4 (continua always go from /ɪ/ to /ɛ/ from step 1 to step 7); AF: auditory feedback condition; BB: bite block condition; pair: minimal pair continuum.

PERCEPTION OF THE JAPANESE WORD-MEDIAL SINGLETON/GEMINATE CONTRAST BY KELANTAN MALAY SPEAKERS

Mohd Hilmi Hamzah[1], Kimiko Tsukada[2,3], John Hajek[3]

[1]Universiti Utara Malaysia, [2]Macquarie University, [3]The University of Melbourne
hilmihamzah@uum.edu.my, kimiko.tsukada@gmail.com, j.hajek@unimelb.edu.au

ABSTRACT

Both Japanese and Kelantan Malay (KM) use consonant length contrastively, though they differ in terms of word position, i.e., Japanese permits the contrast in word-medial position, while KM restricts such a contrast to word-initial position only. It is thus of theoretical interest to examine the extent to which speakers of these languages process familiar consonant length contrasts (i.e., short/singleton vs long/geminate) in unfamiliar word position.

To this end, we examined the perception of Japanese consonant length by KM participants who had no experience with Japanese. The KM and control Japanese participants responded to 200 trials via an AXB discrimination task. The overall mean discrimination accuracy was 84% and 99% for the KM and Japanese groups, respectively. While there was a clear between-group difference in discrimination accuracy, compared to participants from other first language backgrounds, it appears that the KM group benefitted from their experience with L1 consonant length.

Keywords: Consonant length, short/singleton, long/geminate, Japanese, Kelantan Malay

1. INTRODUCTION

In this study, perception of Japanese consonant length contrasts by native and non-native listeners was compared to examine the extent to which Japanese sounds are processed accurately by speakers of languages with length contrasts differing in word position: word-medial and word-initial positions.

Like Japanese, consonant length is contrastive at the level of words across all consonant groups in Kelantan Malay (KM), as empirically established [1-4]. It was reported that closure duration is the most robust acoustic feature of the singleton/geminate consonant contrast in KM, with non-durational parameters such as root mean square (RMS) amplitude and fundamental frequency (F0) playing important secondary roles in enhancing the word-initial length contrast in this language.

On the one hand, Japanese allows contrastive consonant length in word-medial position, where it is considered to be more perceptually salient [e.g., 5]. On the other hand, consonant length in KM only occurs word-initially, which is more marked and generally avoided across languages [e.g., 6]. It is well known that the word-initial contrast is more perceptually indiscernible, particularly for utterance-initial geminates beginning with voiceless stops as there is ostensibly insufficient acoustic information available for listeners in this particular utterance position [e.g., 7].

However, results from a series of perception experiments [8, 9] show that KM native speakers can reliably differentiate natural stimuli of isolated tokens with word-initial singletons and geminates, including those with voiceless stops in utterance-initial position. The results confirm the primary role of closure duration in geminate perception for KM, at least utterance-medially, while RMS amplitude and F0 show limited perceptual functions in both utterance-initial and medial positions. Given the expectation that word-medial geminates are relatively easier to be distinguished from their corresponding singletons, we hypothesise that KM speakers can potentially perceive the word-medial length contrast in Japanese. That is, their L1 experience with consonant length in KM may transfer optimally to cross-language processing. The roles of other prosodic features, such as syllable structure and intonation with regard to consonant length contrast, appear to be underresearched in KM and cannot be ascertained at this stage.

Empirical work on word-initial geminates has been conducted in a related variety of Pattani Malay [e.g., 10], though research on cross-language speech processing has never been reported between this variety and other foreign languages (FLs), including Japanese. Findings of the current study will have potential contributions for FL pronunciation pedagogy and multilingualism in Malaysia where Japanese is taught as a FL for undergraduate students in most public universities in this country. In this regard, American English learners of Japanese demonstrated a clear advantage over non-learners (93% vs 78%) in their perception of Japanese singleton/geminate contrasts [11]. It would be of theoretical as well as pedagogical importance to determine if KM speakers benefit from a combination of L1 consonant length and FL Japanese learning.

2. METHOD

2.1. Stimuli preparation

2.1.1. Speakers

The experimental stimuli and procedures were identical to those used in previous research ([11, 12]). Six (3 males, 3 females) native speakers of Japanese participated in the recording sessions, which lasted between 45 and 60 minutes. The speakers' age ranged from late twenties to early forties. According to self-report, which was confirmed by the second author, all speakers spoke standard Japanese, having been born or having spent most of their life in the Kanto region surrounding the Greater Tokyo Area. The speakers were recorded in the recording studio at the National Institute of Japanese Language and Linguistics, Tokyo.

2.1.2. Speech materials

	Singleton		Geminate	
	word	Gloss	word	gloss
Alvolar	he**t**a	unskilled	he**tt**a	decreased
	ka**t**o	transient	ka**tt**o	cut
	ma**t**e	wait	ma**tt**e	waiting
	o**t**o	sound	o**tt**o	husband
	sa**t**e	well, then	sa**tt**e	leaving
	wa**t**a	cotton	wa**tt**a	broke
Velar	a**k**e	open	a**kk**e	appalled
	ha**k**a	grave	ha**kk**a	mint
	i**k**a	below	i**kk**a	lesson one
	ka**k**o	past	ka**kk**o	parenthesis
	sa**k**a	slope	sa**kk**a	author
	shi**k**e	rough sea	shi**kk**e	humidity

Table 1: Twelve pairs of Japanese words used with target sounds underlined and bolded.

Table 1 shows 12 Japanese word pairs used in this study. The /(C)V**C(C)**V/ tokens contained singleton (n = 96) or geminate (n = 96) consonants intervocalically (underlined and bolded). Only tokens with stops were considered in this study. As voiced geminates are limited in Japanese [13-15], only voiceless stops (/t, k/) were used. On average, the closure durations were 96 ms and 262 ms for singletons and geminates, respectively. The geminate-to-singleton ratios were 2.7 for alveolars (/t/-/t:/) and 2.8 for velars (/k/-/k:/), respectively. These durational values are in good agreement with what has been reported in previous research [e.g., 16] (see, however, [15] for alveolars).

2.2. Participants

Two groups of young adults participated in an AXB discrimination task. The first group consisted of 12 (6 males, 6 females) native speakers of KM (*mean age* = 22.5 years, *sd* = 0.8) who were undergraduate students at Universiti Utara Malaysia in Kedah, Malaysia. They were born and raised in Kelantan, Malaysia and are fluent in Kelantan Malay. They can also speak the standard variety of Malay. None of these participants had experience learning Japanese.

The second and a control group consisted of 10 (2 males, 8 females) native speakers of Japanese (*mean age* = 21.0 years, *sd* = 0.8) who were students at University of Oregon in Eugene, OR, USA. All Japanese speakers were born and spent the majority of their life in Japan. Their mean length of residence in the US was 0.4 years (*sd* = 0.22) at the time of participation. None of the Japanese speakers participated in the recording sessions. According to self-report, all participants had normal hearing.

All participants were tested individually in a session lasting approximately 20 to 25 minutes in a quiet room at their university. The experimental session was self-paced. The participants heard the stimuli at a self-selected, comfortable amplitude level over the high-quality speakers on a notebook computer.

2.3. Procedure

The participants completed a two-alternative forced-choice AXB discrimination task, in which they were asked to listen to trials arranged in a triad (A-X-B). The presentation of the stimuli and the collection of perception data were controlled by the PRAAT program [17]. In the AXB task, the first (A) and third (B) tokens always came from different length categories, and the participants had to decide whether the second token (X) belonged to the same category as A (e.g., 'saka$_2$'-'saka$_1$'-'sakka$_3$') or B (e.g., 'oto$_3$'-'otto$_1$'-'otto$_2$'; where the subscripts indicate different speakers).

The participants listened to a total of 200 unique trials. The first eight trials were for practice and were not analyzed. The three tokens in all trials were spoken by three different speakers. Thus, X was never acoustically identical to either A or B. This was to ensure that the participants focused on relevant phonetic characteristics that group two tokens as members of the same length category without being distracted by audible but phonetically irrelevant within-category variation (e.g., in voice quality). This was considered a reasonable measure of participants' perceptual capabilities in real world situations [18]. All possible AB combinations (i.e., AAB, ABB, BAA, and BBA, 48 trials each) were tested.

The participants were given two ('A', 'B') response choices on the computer screen. They were asked to select the option 'A' if they thought that the first two tokens in the AXB sequence were the same

and to select the option 'B' if they thought that the last two tokens were the same. No feedback was provided during the experimental sessions. The participants could take a break after every 50 trials if they wished. The participants were required to respond to each trial, and they were told to guess if uncertain. A trial could be replayed as many times as the participants wished in order to reduce their anxiety, but responses could not be changed once given. The interstimulus interval in all trials was 0.5 s.

3. RESULTS

We used R version 3.6.0 for statistical analyses and data visualization reported below [19]. The packages used include ez [20] and tidyverse [21].

3.1. Overall results

Figure 1 shows the distributions of percentages of correct discrimination by the two groups of participants. The overall mean discrimination accuracy was 84% and 99% for the KM and Japanese groups, respectively. The Japanese group was at near ceiling with little individual variation. A comparison via the Welch two-sample *t*-test showed that the difference between the KM and Japanese groups was significant [$t(11.5) = -5.3$, $p < .001$].

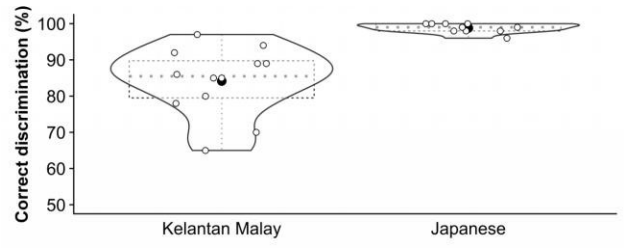


Figure 1: Accuracy (%) of length discrimination by two groups of participants. The horizontal line and the black circle in each box indicate the median and mean, respectively.

3.2. Comparison of the direction of category change (Geminate (G) > Singleton (S) or Singleton > Geminate)

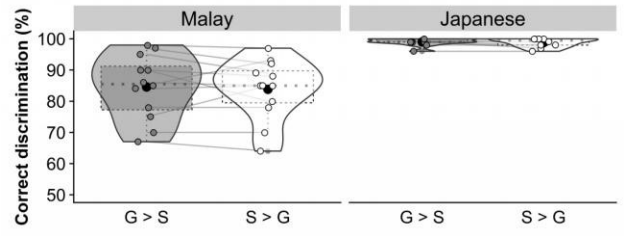


Figure 2: Accuracy (%) of length discrimination for trials differing in the direction of category change. The light lines connect individual participants' scores.

Figure 2 shows the distributions of percentages of correct discrimination for trials differing in the direction (from G to S, from S to G) of length category change. The question of interest was if the participants' discrimination accuracy differed between trials that started with a geminate (and ended with a singleton (i.e., G-S-S, G-G-S)) and trials that started with a singleton (and ended with a geminate (i.e., S-G-G, S-S-G)).

Two-way analysis of variance (ANOVA) with group (KM, Japanese) and direction of category change (G > S, S > G) reached significance only for the main effect of group [$F(1, 20) = 23.2$, $p < .001$, $\eta_G^2 = .52$]. As clearly seen in Figure 2, the Japanese group was significantly more accurate than the KM group whether the length category changed from geminate to singleton or from singleton to geminate within a trial. Neither group was biased with respect to the direction of category change.

3.3. Comparison of the length category (Geminate vs Singleton) of the target token (X in AXB)

Figure 3 shows the distributions of percentages of correct discrimination for trials differing in the length category (geminate, singleton) of the target token. The question of interest was if the participants' discrimination accuracy differed between trials in which X in AXB was a geminate and trials in which X in AXB was a singleton.

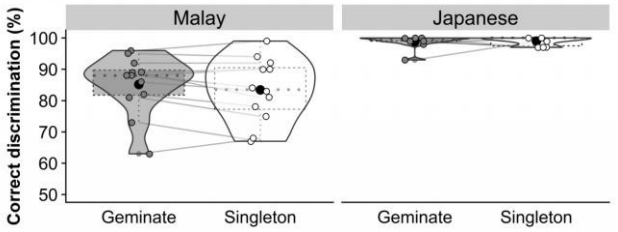


Figure 3: Accuracy (%) of length discrimination for trials differing in the length category of the target token.

Two-way ANOVA with group (KM, Japanese) and length (geminate, singleton) reached significance only for the main effect of group [$F(1, 20) = 22.8$, $p < .001$, $\eta_G^2 = .52$]. As clearly seen in Figure 3, the Japanese group was significantly more accurate than the KM group whether X in the AXB sequence was singleton or geminate. Neither group was biased with respect to the length category of the target token.

3.4. Comparison of the place of articulation (alveolar vs velar) of the target token (X in AXB)

Figure 4 shows the distributions of percentages of correct discrimination for trials differing in the place of articulation (alveolar, velar) of the target token. The question of interest was if the participants' discrimination accuracy differed between trials in which X in AXB was alveolar and trials in which X in AXB was velar.

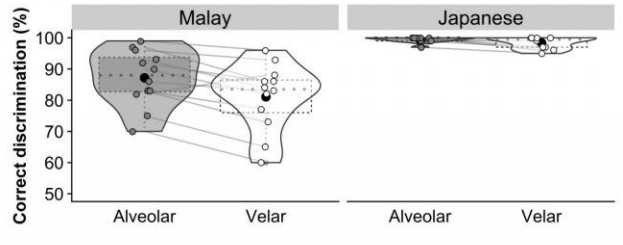


Figure 4: Accuracy (%) of length discrimination for trials differing in the place of articulation of the target token.

Two-way ANOVA with group (KM, Japanese) and place (alveolar, velar) reached significance for the main effects of group [$F(1, 20) = 23.3$, $p < .001$, $\eta_G^2 = .52$] and place [$F(1, 20) = 22.3$, $p < .001$, $\eta_G^2 = .07$]. The two-way interaction was also significant [$F(1, 20) = 10.0$, $p < .01$, $\eta_G^2 = .03$]. As seen in Figure 4, while the Japanese group was not affected by the place factor (99% for alveolar vs 98% for velar), the KM group was more accurate when the target token was alveolar (87%) than when it was velar (81%).

3.5. Comparison of length discrimination at alveolar (/t/-/t:/) and velar (/k/-/k:/) places of articulation

Figure 5 shows the distributions of percentages of correct discrimination for trials differing in the place of articulation (alveolar, velar) and the length category (geminate, singleton) of the target token. The question of interest was if the participants' discrimination accuracy differed for trials in which X in AXB varied in both place and length at the same time (i.e., alveolar geminate, velar geminate, alveolar singleton, velar singleton).

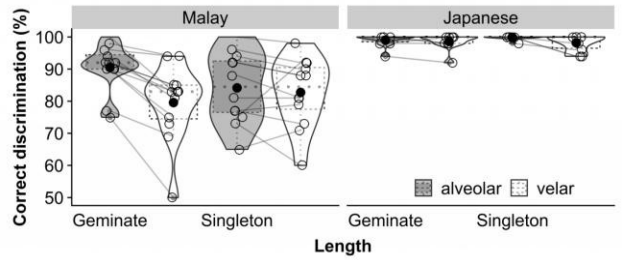


Figure 5: Accuracy (%) of length discrimination for trials differing in the place of articulation and the length category of the target token.

Three-way ANOVA with group (KM, Japanese), length (geminate, singleton) and place (alveolar, velar) reached significance for the main effects of group [$F(1, 20) = 22.9$, $p < .001$, $\eta_G^2 = .49$] and place [$F(1, 20) = 25.4$, $p < .001$, $\eta_G^2 = .06$]. The interactions involving the place factor were all significant [group x place: $F(1, 20) = 11.5$, $p < .01$, $\eta_G^2 = .03$, length x place: $F(1, 20) = 7.3$, $p < .05$, $\eta_G^2 = .02$, group x length x place: $F(1, 20) = 9.6$, $p < .01$, $\eta_G^2 = .03$]. As seen in Figure 5, the influence of place, which was virtually limited to the KM group, was clearer when the target token was geminate (91% for alveolar vs 80% for velar) than when it was singleton (84% for alveolar vs 83% for velar).

4. DISCUSSION

This study examined how KM speakers may perceive Japanese consonant length contrasts known to be difficult for non-native speakers. Consonant length is contrastive in their L1 Malay, but only word-initially. Thus, we were interested in determining if the KM speakers are able to successfully transfer their L1 experience and perceive familiar singleton/geminate contrasts in unfamiliar word-medial position.

While the KM group was less accurate than the Japanese group in discriminating Japanese consonant length, it is notable that their overall score (84%) was intermediate between the previously noted two groups of American English speakers with (93%) and without (78%) Japanese language experience [11]. It thus appears that the KM group benefitted, to some extent, from their experience with L1 consonant length despite lack of Japanese experience.

5. CONCLUSIONS

In the current study, the speakers of KM did not match the Japanese speakers in discriminating Japanese singleton/geminate contrasts word-medially. This may be because they lack specific knowledge of the phonetic characteristics of Japanese singletons and geminates. Only the non-native group was affected by the place factor and discriminated Japanese length contrasts more accurately when alveolar rather than velar occurred in the target position (Figure 4). Our results suggest that experience with Malay singletons and geminates may be helpful but may not automatically transfer to native-like processing of Japanese singletons and geminates.

In the future, we intend to include speakers of other varieties of Malay (e.g., Kedah Malay) who do not use consonant length contrastively in any word positions. Further, given that there are many Malaysian learners of Japanese, it would be pedagogically valuable to include Malay-speaking *learners* of Japanese from different backgrounds and examine if there is additional benefit of Japanese learning within different settings.

6. ACKNOWLEDGEMENTS

This research was supported by Universiti Utara Malaysia (UUM) through University Grant (SO Code: 21403) and the Sumitomo Foundation Fiscal 2020 Grant for Japan-related Research Projects (Grant Number: 208012). We thank reviewers for their time and input for this work.

7. REFERENCES

[1] Hamzah, M. H., Fletcher, J., Hajek, J. 2011. Durational correlates of word-initial voiceless geminate stops: The case of Kelantan Malay. *Proc. 17th ICPhS* Hong Kong, 815–818.

[2] Hamzah, M. H., Fletcher, J., Hajek, J. 2015. Word-initial voiceless stop geminates in Kelantan Malay: Acoustic evidence from amplitude/F0 ratios. *Proc. 18th ICPhS* Glasgow.

[3] Hamzah, M. H., Fletcher, J., Hajek, J. 2016. Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay. *J. Phon.* 58, 135–151.

[4] Hamzah, M. H., Fletcher, J., Hajek, J. 2020. Non-durational acoustic correlates of word-initial consonant gemination in Kelantan Malay: The potential roles of amplitude and f0. *JIPA* 50(1), 23–60.

[5] Tsukada, K, Cox, F., Hajek, J., Hirata, Y. 2018. Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Language Research* 34(2), 179–200.

[6] Frej, M. S., Carignan, C., Best, C. T. 2017. Acoustics and articulation of medial versus final coronal stop gemination contrasts in Moroccan Arabic. *Proc. 18th Interspeech* Stockholm, 210–214.

[7] Ridouane, R., Hallé, P. A. 2017. Word-initial geminates: From production to perception. In: Kubozono, H. (ed.), *The Phonetics and Phonology of Geminate Consonants*. Oxford University Press, 34–65.

[8] Hamzah, M. H., Hajek, J., Fletcher, J. 2016. The role of closure duration in the perception of word-initial geminates in Kelantan Malay. *Proc.* 16th *SST* Sydney, 85–88.

[9] Hamzah, M. H., Fletcher, J., Hajek, J. 2019. The secondary roles of amplitude and F0 in the perception of word-initial geminates in Kelantan Malay. *Proc. 19th ICPhS* Melbourne, 2753–2757.

[10] Abramson, A. S. 2003. Acoustic cues to word-initial stop length in Pattani Malay. *Proc. 15th ICPhS* Barcelona, 387–390.

[11] Tsukada, K., Idemaru, K., Hajek, J. 2018. The effects of foreign language learning on the perception of Japanese consonant length contrasts. *Proc.* 17th *SST* Sydney, 37–40.

[12] Tsukada, K., Hajek, J. 2022. Adaptation to L3 phonology? Perception of the Japanese consonant length contrast by learners of Italian. *Proc.* 18th *SST* Canberra, 171–175.

[13] Hussein, Q., Shinohara, S. 2019. Partial devoicing of voiced geminate stops in Tokyo Japanese. *J. Acoust. Soc. Am*. 145, 149–163.

[14] Kawahara, S. 2015. The phonetics of sokuon, or geminate obstruents. In: Kubozono, H. (ed.), *Handbook of Japanese Phonetics and Phonology*. Walter de Gruyter, 43–78.

[15] Sano, S. 2019. The distribution of singleton/geminate consonants in spoken Japanese and its relation to preceding/following vowels. *Proc. 19th ICPhS* Melbourne, 1833–1837.

[16] Hayes-Harb, R. 2005. Optimal L2 speech perception: Native speakers of English and Japanese consonant length contrasts. *Journal of Language and Linguistics* 4, 1–29.

[17] Boersma, P., Weenink, D. Last viewed June 13, 2016. *Praat: Doing Phonetics by Computer* [version 6.0.19], retrieved from http://www.praat.org

[18] Strange, W., Shafer, V. L. 2008. Speech perception in second language learners: The re-education of selective perception. In: Hansen Edwards, J. G., Zampini, M. L. (eds), *Phonology and Second Language Acquisition.* John Benjamins, 153–191.

[19] R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[20] Lawrence, M. A. 2016. *ez: Easy Analysis and Visualization of Factorial Experiments.* R package version 4.4-0. https://CRAN.Rproject.org/package=ez

[21] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R. et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43), 1686.

THE EFFECT OF SHORTENING ONSET CONSONANTS ON SPEECH SEGMENTATION BY TAIWANESE SOUTHERN MIN LISTENERS

Shu-chen Ou[1], Zhe-chen Guo[2]

[1]National Sun Yat-sen University, [2]The University of Texas at Austin
sherryou@mail.nsysu.edu.tw, zcadamguo@utexas.edu

ABSTRACT

This study examined whether and how shortening voiced syllable-onset consonants impacts listeners' speech segmentation. Cross-linguistically, listeners tend to use longer onset consonants to locate word beginnings, possibly because of the greater phonetic richness and hence auditory-perceptual salience of these consonants. An alternative account, however, holds that it is the increased consonantality that guides the use of longer onsets as word beginning cues. Since shortening voiced stops and nasals enhances their consonantality, we tested the competing explanations in an artificial language learning experiment where shorter voiced onsets provided a potential segmentation cue. Taiwanese Southern Min listeners segmented the words of an artificial language through listening to continuous repetitions of the words. Shortening the voiced onsets in word-initial syllables did not significantly improve segmentation. The results thus did not support the consonantality-based explanation but was in line with the phonetic richness account. Possible mechanisms underlying the findings are discussed.

Keywords: speech segmentation, prosody, consonant shortening, artificial language learning.

1. INTRODUCTION

Speech is continuous without reliable pauses between words [1]. Yet, listeners can still parse it into discrete words as if they were beads on a string. One well-established finding from research into this ability is that listeners across languages use longer vowels to locate word-final positions [2]–[4]. More recently, studies have begun to explore whether and how consonant duration is also used in segmentation. In this paper, we aim to add further insight into this issue by testing whether shortened voiced syllable-onset consonants are useful for speech segmentation.

Like longer vowels, longer syllable-onset consonants can guide speech segmentation, but in a different way. Specifically, they are interpreted as signalling word beginnings rather than word-finality. This was demonstrated by White et al. [5] in a study with English listeners using the artificial language (AL) learning paradigm [2]. The listeners learned to segment trisyllabic nonsense words (e.g., /pabiku/) by listening to speech streams in which the words were concatenated continuously. The results showed that their segmentation improved when the word-initial onset consonants were lengthened. Similar findings were replicated in Italian, Hungarian, and Taiwanese Southern Min (TSM) [6], [7], even though in languages like Italian, consonant duration contrast is phonemically relevant. Using longer onsets to locate word-initial positions seems to be a cross-linguistic tendency superseding language-specific functions or patterns of consonant duration.

Kim et al. [8] observed a similar tendency in an AL learning study with Dutch and Korean listeners. They found that when the AL words began with a voiceless stop onset, both groups segmented speech better when the onset was longer with a longer voice onset time (VOT) than when it was shorter with a shorter VOT. This ran contrary to the prediction that due to the different phonetic realizations of domain-initial strengthening in the two languages [9], [10], Dutch listeners would exploit stops with shorter VOTs better than those with longer VOTs to locate word beginnings, while Korean listeners would show the opposite pattern. Kim et al. attribute their findings to the richer phonetic information in stops with longer VOTs, which confer perceptual salience on the stops to make them so useful for segmentation as to override language-specific strengthening patterns.

However, while phonetic richness can explain the cross-linguistic parallel in [8], there is an alternative possibility. Domain-initially, voiceless onsets such as voiceless stops can be produced with longer VOTs to become more consonant-like and enhance the syntagmatic contrast with the following vowel [11]. Yet, in the same environment, voiced onsets become more consonant-like through reduced sonority such that nasals, for example, are shorter with decreased nasal airflow [11]–[13]. Since [8] tested only voiceless stops, it is possible that the cross-linguistic preference for stops with longer VOTs is not driven by phonetic richness, but by a universal tendency to interpret heightened consonantality as signalling word beginnings. Voiced nasals and stops can provide a test of this alternative account as their consonantality increases when they are shortened.

Here, we examined the effect of shortening voiced onset consonants on speech segmentation by TSM

listeners. Ou and Guo [7] have shown that TSM listeners exploit lengthening of voiced onsets to locate word-initial positions. Currently, it has not been explicitly tested whether and how they would use shortening of these consonants. To assess the consonantality-based account, we conducted an AL learning experiment with TSM listeners which was modelled after [7] but with the cue voiced onset lengthening replaced by voiced onset shortening.

2. EXPERIMENT

2.1. Design and hypotheses

As in previous studies (e.g., [2]), our experiment began with a learning phase in which listeners were exposed to continuous repetitions of the (nonsense) words of an AL. Next, they recognized the words in a forced-choice test, with higher accuracy suggesting more successful segmentation during the learning.

If there is a tendency to interpret increased consonantality as cueing word beginnings, TSM listeners' segmentation should improve in an "initial shortening (IS)" condition where voiced onsets in the initial syllables of the AL words were shortened, compared with a "no shortening" (NS) condition where segment duration did not vary. Also, for the consonantality-based explanation to be supported, it had to be shown that improved segmentation was not due to shortening in any boundary-adjacent syllables. That is, the listeners were not expected to benefit from a "final shortening (FS)" condition where onsets of word-final syllables were shortened. Yet, given the evidence for longer onsets as cues to word-initiality [5]–[7] and the phonetic richness hypothesis [8], it may be that only onset lengthening is used to identify word beginnings. This predicted that TSM listeners would not benefit from onset shortening at all.

2.2. Stimuli

Following [7] and [14]–[16], we created an AL consisting of six trisyllabic CVCVCV words: /banume/, /bimɔna/, /genigɔ/, mimabu/, /nebɔgi/, and /nɔgamu/. These words were constructed from four voiced consonants (/b, g, m, n/) and five vowels (/a, i, e, u, ɔ/), all occurring in TSM. Each CV syllable was produced in a sound-treated room by a native male TSM speaker with phonetic training, who was instructed to read the syllable aloud in monotone in a carrier sentence (/gua kɔŋ __ tsit e dʑi/ "I said the word __"). His speech was recorded with Audacity and a Zoom H4n Handy Recorder at a 44.1 kHz sampling rate and digitized as WAV files.

The syllables were extracted from the carrier sentence and prosodically neutralized using Praat [17]: the F0 contours were flattened, the amplitudes of the syllables were equalized, and the durations of all consonants and vowels were set to a base duration of 150 ms. The resulting syllables were concatenated to form the AL words for the NS condition. In the IS and FS conditions, the onsets of the first syllables and third syllables were shortened by a factor of 1.5 (i.e., to 100 ms, or 0.67 of the base duration), respectively. The shortening factor was based on the magnitudes of lengthening used in [6], [7]. Fig. 1 shows the waveforms of an AL word under the three conditions.

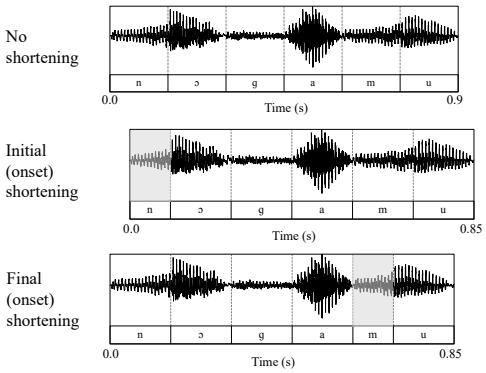


Figure 1: Waveforms of /nɔgamu/ in the three conditions

For each condition, repetitions of the AL words were concatenated without pauses to produce six long speech streams, which were presented as stimuli in the learning phase. Each stream contained 20 repetitions of each word in a pseudo-random order such that the same word never followed itself. The total duration of the six streams was 10-11 minutes. Five-second fade-in and fade-out effects with logarithmic ramps were applied to each stream to prevent listeners from identifying the very first and last syllables and using them as word boundary cues.

The test phase was a two-alternative forced-choice task which presented two stimuli, separated by a 500-ms silence, in each trial: a word of the AL and a "partword." The partword was a trisyllabic CVCVCV sequence formed by combining substrings of two AL words (e.g., /mabuge/, from /mimabu/ and /genigɔ/). There were six partwords. The average between-syllables transitional probability, computed based on the learning-phase speech streams following [2], was around 0.6 for the partwords and one for the AL words. The test contained 36 trials, resulting from all word-partword pairings. The orders of the word and partword were balanced and no stimuli in the test phase had the onset shortening cue. The experiment was compiled using E-prime 2.0 [18].

2.3. Procedure

Seated individually in front of a desktop computer and randomly allocated to either condition, participants first learned the AL by listening to the six

speech streams passively via headphones. They were encouraged to pay as much attention to the streams as possible and made aware that they would be tested on how much they knew the words of the AL. No cues to the words (e.g., their lengths) were provided. After listening to the speech streams, participants completed the forced-choice test, in which they heard two stimuli in each trial and selected, in five seconds, the one that they thought was a word of the AL. The experiment took about 30 minutes.

2.4. Participants

Ninety native speakers of TSM were recruited (45 males, 45 females; $N = 30$ for each condition). Their mean age in years was 21.8 (range: 20–35). All had also acquired Taiwan Mandarin as a native language and learned English as an academic subject. None reported speech or hearing impairments.

3. RESULTS

Participants' test response accuracy was analyzed. Timeout trials (0.25% of all trials) were discarded. The remaining responses ($N = 3{,}232$) were coded as correct (1) if the AL word was selected and incorrect (0) if not. Fig. 2 shows the response accuracy of each participant and the mean of each condition.

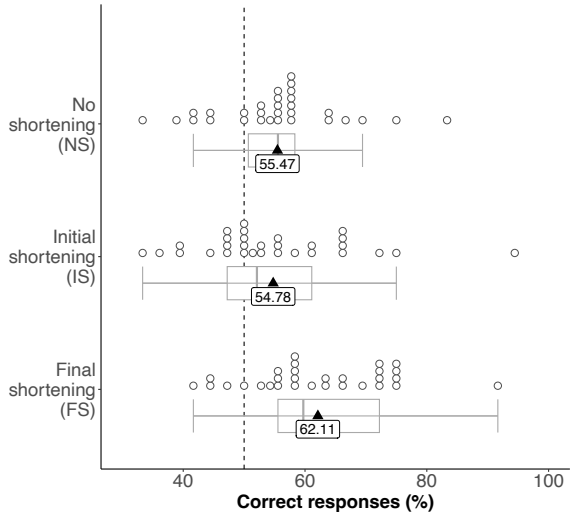


Figure 2: Response accuracy of individual participants (circles) and mean accuracy of each condition (triangles). The gray bar in each box marks the median and the dashed vertical line represents the chance accuracy (50%).

A Bayesian hierarchical logistic regression model was fitted to the responses using the *brms* [19] package of R [20]. The population-level (fixed) effect of main interest was Condition, which contrasted the IS and FS conditions with NS (baseline). As in [21], the model also contained the order number of each trial in the test (Trial) and log-transformed response time (LogRT) to capture any effects of practice or fatigue and response speed-accuracy relationship, respectively. These two continuous predictors were scaled and entered as population-level covariates. The group-level (random) effects included intercepts varying by participant, AL word, and partword and slopes for Condition varying by AL word. Weakly informative priors (e.g., Normal(0, 10)) were assigned to all population- and group-level effects. Following [22], we used a weakly informative LKJ(2) prior [23] on the intercept-slope correlation in the group-level effects. Posterior distributions of model parameters were approximated using four 2,000-iteration Monte Carlo Markov Chains with 1,000 warm-ups. $\hat{R}$ values [24] reached one for all parameters, confirming chain convergence.

Fig. 3 shows the marginal posterior distributions of estimates for the population-level effects. We determined statistical significances based on the 95% highest density interval (HDI) criterion [25]. That is, the effect of a predictor was significant if the 95% HDI of the relevant posterior distribution excluded zero. Thus, LogRT had a significant effect suggesting that slower responses tended to be incorrect, 95% HDI = [–0.261, –0.094], possibly due to listeners' response certainty or confidence [7], [21]. Trial had no significant effect, 95% HDI = [–0.082, 0.069].

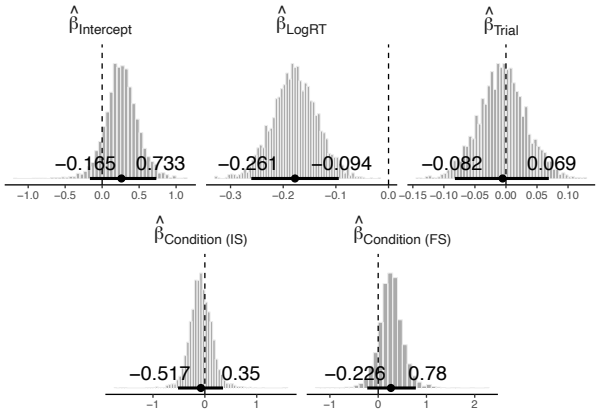


Figure 3: Marginal posterior distributions of population-level effect estimates. The bar under each distribution represents the 95% HDI with the circle marking the mean. The dashed vertical line indicates zero.

Of main interest was the Condition predictor contrasting the NS condition with the other two conditions with shortening cues. The model revealed that neither IS, 95% HDI = [–0.517, 0.350], nor FS, 95% HDI = [–0.226, –0.780], differed significantly from NS. There was thus no evidence that shortening voiced onset consonants, either word-initially or word-finally, affected speech segmentation.

As no shortening conditions differed significantly from NS, we further explored the listeners'

performance by comparing it against the chance level. To do so, we computed the posterior distributions of correct response percentages for the conditions as shown in Fig. 4. No condition was significantly better than chance (as none had a 95% HDI that excluded 50%). However, the posterior probability that the accuracy was better than chance was high, especially for the FS condition (NS: 90%; IS: 77%; FS: 96%). Thus, it was likely that the listeners might still be able to segment out some of the AL words simply by listening to them repeated continuously.

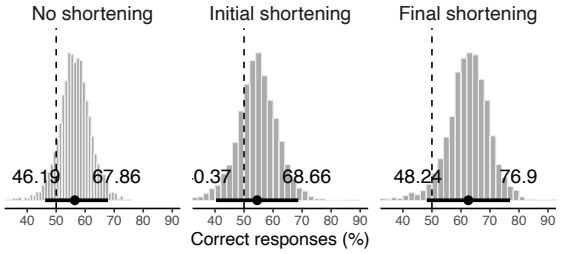


Figure 4: Posterior distributions of probability of correct responses (%) for the three conditions. The bar represents the 95% HDI with the circle marking the mean. The dashed line indicates the chance level (i.e., 50%).

4. DISCUSSION AND CONCLUSION

This study examined whether shorter onset consonants are exploited in speech segmentation. One explanation for the cross-linguistic use of longer onsets in segmentation is that listeners interpret increased onset consonantality as signalling word beginnings. We tested it in an AL learning experiment in which the voiced stops or nasals in syllable onset were shortened to enhance their consonantality. The results showed that onset shortening, either in the initial or final syllables of the AL words, did not significantly improve segmentation. Thus, at least for TSM listeners, there is no evidence for the above explanation. Our findings indirectly align with the phonetic richness hypothesis [8], which holds that the greater auditory-perceptual salience of longer onsets underlies their cross-linguistic use in segmentation.

A question then arises as to the specific mechanisms leading listeners to use onset lengthening as reported in previous studies [5]–[7] but not onset shortening. We assume that the listeners' behavior may be explained from a functional-informational perspective. According to the Effort Code account of intonational variation [26], listeners associate what speakers produce with high pitch or greater pitch range (or more generally, with greater articulatory effort) with greater informational load. When the onsets of word-initial syllables in an AL learning task are made longer and perceptually more salient, they are interpreted as the speaker's effort to signal important information, which directs listeners' attention to word beginnings. In contrast, though word-initial onsets were consistently shortened in our IS condition to provide a potentially useful segmentation cue, listeners would perceive them as informationally trivial and ignore them.

The patterns of findings here and in previous AL learning studies may also be accounted for based on the timing of the perceptual center (p-center) of a syllable, or the perceptual moment of its occurrence [27]. White et al. [5] suggest that lengthening the onset of the initial syllable of an AL word delays its p-center relative to the p-center of the last syllable in the previous word, and such a delay is interpreted as a marker of prosodic boundary. On the other hand, shortening a word-initial onset brings its p-center closer to that of the final syllable in the previous word, producing no boundary percept. Moreover, since the shortening reduces the temporal distance between the p-centers of the syllables across a word boundary, it could lead listeners to incorrectly analyze the two syllables as belonging to the same word.

In addition to the lack of an effect of onset shortening, it was found that although no conditions showed segmentation better than chance, the posterior probability of above-chance performance was relatively higher in the FS condition. We speculate that shortening the onsets of word-final syllables might cause the following vowels to be perceived as longer and hence word-final, reflecting the well-documented use of longer vowels to locate word-finality [2], [3]. However, further confirmatory evidence would be needed to verify this speculation.

To conclude, the current study did not find evidence that shortened voiced onsets with increased consonantality are used to identify word beginnings. Word-initial onset consonants, whether they are voiced or voiceless, facilitate segmentation when they are lengthened. Yet, it remains to be tested in future work if onset shortening can become useful when combined with other cues. For instance, while a nasal onset in the domain-initial position is produced with a shorter duration to enhance consonantality, there is a concomitant reduction in coarticulatory nasalization in the following vowel [12], [28]. It is possible that shorter onsets are exploited when accompanied by a coarticulatorily appropriate reduction of their influence on the vowel. Additionally, our AL stimuli were shortened artificially and their F0 and amplitude were neutralized. It is then important to examine onset shortening effects on segmentation of meaningful, naturalistic speech in which other segmental and prosodic phenomena co-vary with onset duration. Further research is also needed to test if the lack of onset shortening effects is specific to TSM or can generalize to other languages.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the grant support to the first author from National Science and Technology Council, Taiwan (MOST109-2410-H-110-069-MY3).

6. REFERENCES

[1] Cole, R. A., Jakimik, J., Cooper, W. E. 1980. Segmenting speech into words. *J Acoust Soc Am* 67, 1323–1332.

[2] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: The role of distributional cues. *J Mem Lang* 35, 606–621.

[3] Tyler, M. D., Cutler, A. 2009. Cross-language differences in cue use for speech segmentation. *J Acoust Soc Am* 126, 367–376.

[4] Ordin, M., Polyanskaya, L., Laka, I., Nespor, M. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Mem Cognit* 45, 863–876, 2017.

[5] White L., Mattys, S. L., Stefansdottir, L., Jones, V., 2015. Beating the bounds: Localized timing cues to word segmentation. *J Acoust Soc Am* 138, 1214–1220.

[6] White, L., Benavides-Varela, S., Mády, K. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J Phon* 81, 100982.

[7] Ou, S.-C., Guo, Z.-C. 2021. The differential effects of vowel and onset consonant lengthening on speech segmentation: Evidence from Taiwanese Southern Min. *J Acoust Soc Am* 149, 1866–1877.

[8] Kim, S., Cho, T., McQueen, J. M. 2012. Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *J Phon* 40, 443–452.

[9] Cho T., McQueen, J. M. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *J Phon* 33, 121–157.

[10] Cho T., Keating, P. A. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *J Phon* 29, 155–190.

[11] Cho T., Keating, P. 2009. Effects of initial position versus prominence in English. *J Phon* 37, 466–485.

[12] Cho, T., Kim, D., Kim, S. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *J Phon* 64, 71–89.

[13] Hsu, C.-S. K., Jun, S.-A. 1998. Prosodic strengthening in Taiwanese: Syntagmatic or paradigmatic? *UCLA Working Papers in Phonetics* 96, 69–89.

[14] Kim, S., Cho, T., McQueen, J. M. 2012. Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *J Phon* 40, 443–452.

[15] Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., Christiansen, M. H. 2015. Factors influencing sensitivity to lexical tone in an artificial language: Implications for second language learning. *Stud Second Lang Acquis* 37, 335–357.

[16] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: The role of distributional cues. *J Mem Lang* 35, 606–621.

[17] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/.

[18] Psychology Software Tools. 2012. E-Prime 2.0. Pittsburgh, PA.

[19] Bürkner, P. C. 2017. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw* 80, 1–28.

[20] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[21] Ou, S.-C., Guo, Z.-C. 2021. The language-specific use of fundamental frequency rise in segmentation of an artificial language: Evidence from listeners of Taiwanese Southern Min. *Lang Speech* 64, 437–466.

[22] Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., Kong, E. J. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *J Phon* 71, 147–161.

[23] Lewandowski, D., Kurowicka, D., Joe, H. 2009. Generating random correlation matrices based on vines and extended onion method. *J Multivar Anal* 100, 1989–2001.

[24] Gelman, A., Hill, J. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

[25] Kruschke, J. K., Aguinis, H., Joo, H. 2012. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organ Res Methods* 15, 722–752.

[26] Gussenhoven, C. Intonation and interpretation: phonetics and phonology. *Proc. Speech Prosody 2002*, 47–57.

[27] Morton, J., Marcus, S., Frankish, C. 1976. Perceptual centers (P-centers). *Psychol Rev* 83, 405–408.

[28] Jang, J., Kim, S., Cho, T. 2018. Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *J Acoust Soc Am* 144, EL33–EL39.

INDIVIDUAL DIFFERENCES IN NON-NATIVE PHONETIC PERCEPTION: POTENTIAL LINKS TO NATIVE PERCEPTION, ATTENTION, & MEMORY

Claire Honda[1,2], Meghan Clayards[1,2], Shari Baum[1,2]

[1]McGill University, Montreal, Canada; [2]Centre for Research on Brain, Language and Music, Montreal, Canada.
claire.honda@mail.mcgill.ca, meghan.clayards@mcgill.ca, shari.baum@mcgill.ca

ABSTRACT

Adults differ considerably in their ability to distinguish non-native sounds. The predictors of these individual differences remain poorly understood. We investigated whether differences in the discrimination of unfamiliar non-native phonemes were predicted by differences in performance on two native phonetic perception tasks (2AFC and VAS) and on two non-linguistic cognitive tasks. While preregistered analyses showed no significant predictors of non-native discrimination, exploratory analyses revealed a potential role of consistent native perception and cognitive factors in promoting successful non-native perception. These findings contribute to our understanding of individual differences in non-native phonetic perception, and have implications for encouraging successful acquisition of new languages in adulthood.

Keywords: phonetic perception, gradiency, response consistency, sustained attention, working memory

1. INTRODUCTION

Adults show great variability in their ability to successfully discriminate non-native phonemic contrasts (e.g., [1]). Some of this variability can be attributed to factors including native language background [2], musical ability [3], and auditory acuity [4]. Nonetheless, these factors explain only a portion of the observed variability and may depend on the specific non-native sounds being discriminated, so the predictors of successful non-native perception remain underspecified.

Adults also show individual differences in their perception of native speech sounds. A common way of measuring native phonetic perception is using 2-alternative forced choice (2AFC) tasks, in which listeners hear continua of sounds and classify each sound into one of two categories. A more recent alternative is visual analog scaling (VAS) tasks, in which listeners indicate where each sound falls along a continuous line between two categories. When a participant's responses on these tasks are plotted against the changes in stimulus properties along the continuum, the resulting identification slope can be steep (reflecting a sudden change in perception from one sound to the other) or shallow (reflecting gradual changes in the percept). Listeners differ in the steepness of their identification slopes and also in how consistently they respond to stimuli (e.g., [5]).

It is thought that steep slopes on 2AFC tasks reflect an ability to make fine-tuned and accurate judgments about which category a sound belongs to, which is important for making sense of the variable speech signal [6]. Conversely, shallow slopes on VAS tasks appear to reflect gradient perception, i.e. an ability to distinguish fine-tuned differences between sounds [5]. Gradient perception may grant listeners greater perceptual flexibility by enabling them to activate and manage competing phonemic representations, which facilitates reinterpretation of ambiguous acoustic input as new information becomes available [7]. Such flexibility in perception could be useful when learning a new language given that non-native sounds often sound like ambiguous instances of native sounds [8] (though see [9] for a study where native perception on a 7-point VAS task did not predict non-native perception).

There do appear to be links between native and non-native perception under some circumstances. In participants who underwent perceptual training to identify non-native vowels, better native vowel discrimination predicted better non-native vowel perception [10]. In bilinguals who were divided into poor vs. good perceivers of a non-native contrast, the poor perceivers were less sensitive when processing both native and non-native vowel contrasts [11]. Furthermore, L1 phonetic deficits have been linked to difficulty in L2 phonetic mastery [12]; and successful learning of an L2 contrast recruits the same neural regions involved in processing L1 contrasts [13].

Despite some evidence in support of a relationship between native and non-native phonetic perception, other studies have failed to find such a relationship. Native phonological processing has been shown not to predict learning of a non-native contrast [14], and perception of a continuum of native sounds also did not predict identification of non-native sounds during the same session [9]. As such, the predictors of individual differences in non-native perception and the links between native and non-native perception remain to be clarified.

We aimed to elucidate the potential predictors of successful non-native perception—

including categorization slopes on 2AFC and VAS tasks of native perception—and in so doing, to also clarify the nature of any relationship that might exist between native and non-native perception.

We hypothesized that the ability to distinguish fine-tuned differences in native speech sounds would relate to the ability to accurately discriminate non-native speech sounds. If this were the case, we would find that better non-native perception relates to steeper 2AFC slopes (reflecting an ability to accurately categorize sounds) and shallower VAS slopes (perhaps reflecting gradient perception that is fine-tuned and flexible).

2. METHODS

The methods, exclusion criteria, and analyses for this study were preregistered based on a pilot study (https://osf.io/ez5qh/?view_only=8e4a1498e04f4ee0946752ee93b9ce71).

2.1. Participants

139 monolingual English speakers (97 females) were recruited through Prolific.co. They were aged 18-35 (mean: 25), right-handed, born and living in the United States or Canada, and had no language, cognitive, or hearing impairments. They completed the study at home on Gorilla.sc, using their own headphones.

2.2. Native phonetic perception tasks

Two native phonetic perception tasks were used. On both tasks, participants listened to two minimal pairs (*bet–bat* and *dear–tear*, publicly available at https://osf.io/369my/ [15]). The minimal pairs were manipulated so that each one varied systematically in two relevant acoustic cues (formant frequency and vowel duration for *bet–bat*, voice onset time and onset F0 for *dear–tear*). Each cue varied in 5 steps, and each version of the first cue was paired with each version of the second cue, resulting in 25 stimuli per pair. Stimuli whose cue values were both at the extremes (i.e., step 1 or step 5) consequently sounded clear and unambiguous, while stimuli with more intermediate cue values sounded more ambiguous. The same stimuli were used in both tasks.

On each trial of the 2AFC task, participants chose between two options on the screen via mouse click to indicate what they heard (e.g., *bet* or *bat*). On each trial of the VAS task, the screen displayed a slider with the two members of the minimal pair on opposite ends, and participants indicated where along the continuous slider they perceived the stimulus to be (from 0 to 100). Each stimulus from each minimal pair was presented 5 times per task. Participants completed the VAS task first to avoid biasing responses based on the categorical demands of the 2AFC task.

2.3. Non-native phonetic perception task

Participants discriminated German vowels and consonants (/øː/–/œ/, /yː/–/ʏ/, /ʃ/–/ç/) that are perceptually challenging sounds for native English speakers [16]. To construct the stimuli, three native German speakers were each recorded producing 14 minimal pairs containing these phonemes. Stimuli were presented in a 3-interval oddity task. On each trial, participants heard three words—one from each speaker—and clicked "1", "2", or "3" to indicate which word sounded different or "None" if all three words sounded the same. There were 12 trials per minimal pair: half were switch trials, where one word was the other member of the minimal pair (e.g., "Kirche", "Kirche", "Kirsche"), and the other half were catch trials, where all 3 words were the same.

2.4. Cognitive tasks

To determine whether any observed relationships between performance on the speech tasks might be driven by individual differences in general cognitive capacity, participants completed two non-linguistic cognitive tasks. A version of the Continuous Performance Task (CPT) was used as a measure of sustained attention [17], and a backwards digit span task was used as a measure of working memory [18]. On the AX-CPT task, participants saw a string of letters and had to press one key whenever they saw the letter X after the letter A or press a different key for any other letter combination. There were 200 trials: 140 AX trials (A followed by X), 20 AY trials (A followed by a consonant other than X), 20 BX trials (B followed by X), and 20 BY trials (B followed by a consonant other than X). On the backwards digit span task, participants listened to a recorded series of numbers and had to type them out in the reverse order. The number of digits to be recalled increased every three trials, starting with two digits and increasing to a maximum of ten. When a participant incorrectly answered all three trials of a difficulty level, the task was terminated.

2.5. Preparatory data analysis

Participants were excluded if they reported phonetic training, German exposure, or very good listening or speaking ability in a language other than English. They were also excluded for low-effort task performance based on preregistered criteria related to reaction times and accuracy. This left 91 participants with complete data across all tasks.

2.5.1. Native phonetic perception tasks

Two mixed-effects logistic regression models were fit to the 2AFC data. In both models, the outcome was participants' responses (first model: responses to *bet–bat* stimuli; second model: responses to *dear–tear*). Fixed effects were the first and second acoustic cues, and correlated random effects were by-participant random intercepts and by-participant random slopes for each acoustic cue. The random slopes coefficients for both cues were extracted as the variables of interest, because they quantify how each participant's use of a given cue differs from the group average [15]. These random slopes coefficients—4 per participant—were reduced to two dimensions using principal component analysis (PCA) in R [19] using the lme4 package [20] and the prcomp() function.

The rotated logistic function in [5], which models response slope (gradiency), was fit to participants' VAS responses in MATLAB [21]. This generated two slopes and two consistency measures per participant (*bet–bat*, and *dear–tear*). For each trial, the difference between the participant's actual VAS response and the response predicted by the rotated logistic was calculated. The standard deviation of these residuals was averaged per minimal pair to provide an estimate of each participant's response consistency. The four final VAS variables—2 gradiency measures and 2 consistency measures—were reduced to two PCA components in R as for the 2AFC task.

2.5.2. Non-native phonetic perception task

From participants' responses on the oddity task, the non-parametric sensitivity index *A* was calculated, which takes into account hits (correct selection of the odd item on a switch trial) and false alarms (incorrect selection of an item as being odd on a catch trial).

2.5.3. Cognitive tasks

Bin scores were calculated from participants' AX-CPT responses [22]. These scores account for both reaction time (RT) and accuracy. Lower bin scores indicate better sustained attention as reflected by higher accuracy and/or smaller RT differences between non-switch (AX) and switch (non-AX) trials.

The highest number of digits successfully recalled on the backwards digit span task was taken as a measure of working memory.

2.6. Primary data analysis

A multiple regression model was fit with oddity *A* scores as the outcome. Predictor variables were the first two principal components derived from the PCA of the 2AFC coefficients, the first two components derived from the PCA of the VAS measures, and the two control predictors (AX-CPT bin scores and digit span levels). We predicted that the 2AFC and VAS measures would relate to oddity scores after accounting for the control predictors.

3. RESULTS

There were individual differences in performance on all tasks, as anticipated. Examples of four participants' responses on the VAS task are shown in Figure 1, illustrating differences both in slopes and in the consistency of responses.

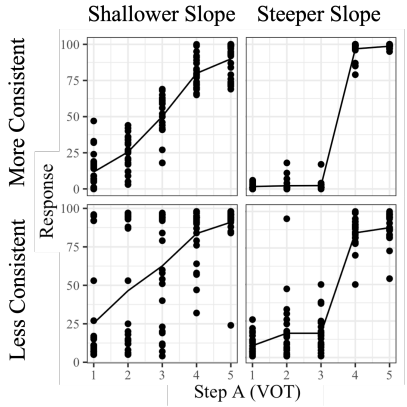


Figure 1: Four different participants' response curves for dear-tear cue A on the VAS task, illustrating individual variability.

The preregistered multiple regression model revealed that, contrary to our hypothesis, none of the predictors was significantly related to non-native perception scores ($p > 0.5$ for all).

Cook's distance was calculated to check for participants who might be particularly influencing the model's fit. This revealed one participant with higher influence than the others ($d = 1.6$), who upon closer examination had responded to the VAS task in a more categorical way than most (primarily using endpoints instead of the whole range). The multiple regression model was refit without this influential participant. The resulting exploratory model is summarized in Table 1. The 2AFC and VAS slope measures were again not significant predictors. Instead, the first principal component derived from the VAS variables was a significant predictor of non-native perception. This component primarily reflected consistency (see Table 2). Figure 2 shows non-native perception plotted against VAS consistency (averaged across both contrasts), revealing how more consistent VAS responders tended to have better non-native discrimination. The attention and memory control measures were also marginally significant ($p = 0.052$ and $p = 0.056$, respectively). We contextualize and elaborate upon these findings below.

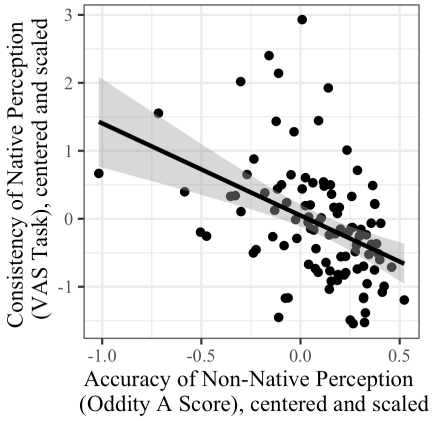


Figure 2: Scatterplot of non-native discrimination and native VAS consistency (higher values indicate less consistency and better non-native perception).

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
(Intercept)	0.078	0.025	3.129	0.002
2AFC principal component (PC) 1	-0.008	0.021	-0.353	0.725
2AFC PC2	<0.001	0.023	0.002	0.999
VAS PC1	0.099	0.029	3.391	**0.001**
VAS PC2	-0.033	0.028	-1.194	0.236
AX-CPT bin score	-0.055	0.028	-1.974	**0.052**
Backwards digit span	-0.056	0.029	-1.936	**0.056**
Multiple R^2 = 0.217, Adjusted R^2 = 0.161, Residual SE = 0.231 (df = 84), n = 91				

Table 1: Regression table for the multiple regression model predicting oddity A scores. Model equation: Oddity A score ~ 2AFC PC1 + 2AFC PC2 + VAS PC1 + VAS PC2 + AX-CPT bin score + Backwards digit span.

	PC1	PC2
2AFC variables		
bet-bat acoustic cue A	-0.360	0.537
bet-bat acoustic cue B	-0.170	0.766
dear-tear acoustic cue A	-0.647	-0.266
dear-tear acoustic cue B	-0.651	-0.232
VAS variables		
bet-bat slope	0.395	-0.217
dear-tear slope	0.259	0.813
bet-bat consistency	-0.632	-0.262
dear-tear consistency	-0.614	0.472

Table 2: Correlations between the original 2AFC (top) and VAS (bottom) variables and the first two principal components extracted from them. PC1 and PC2 explained 52% and 29% of the variance in 2AFC responses; PC1 and PC2 explained 40% and 29% of the variance in VAS responses respectively.

4. DISCUSSION

Our objective was to investigate possible predictors of successful non-native perception. Contrary to our hypothesis, we did not find a relationship between fine-tuned or gradient native perception (as measured by 2AFC and VAS tasks) and accurate non-native perception. However, we found preliminary evidence that consistent native perception and non-linguistic cognitive factors (attention and memory) may play a role in discriminating unfamiliar language sounds.

The capacity to consistently perceive native sounds could conceivably help with reliably and accurately perceiving non-native sounds as well. The importance of consistent perception is hinted at by work showing that the ability to consistently assimilate a given non-native phoneme to a native category has been linked to having greater non-native perceptual proficiency [23] and a larger non-native vocabulary [24]. Future work should investigate the potential functions and implications of consistency during native and non-native phonetic perception.

Our attention and memory measures were marginally related to non-native discrimination, which is also supported by some previous work. For instance, non-native phonemes that are attended to are learned better than ones not attended to [25], and attention on the CPT has been related to non-native listening test scores [26]. It has also been shown that people with better working memory learn non-native phonemes faster and use more optimal learning strategies [27]. Attention and memory likely contribute to promoting robust processing and representation of non-native sounds.

The lack of strong relationships observed between our variables could be due in part to differences between the tasks used to measure native and non-native perception (e.g., the number of stimuli presented per trial, unmanipulated vs. ambiguous tokens). It could also be that native and non-native perception are simply not strongly linked – at least not initial non-native perception and the native constructs we measured ([9]). Perhaps native and non-native perception involve different mechanisms or strategies to some extent, as suggested by research showing that native speakers integrate phonetic information more gradiently while non-native speakers show a more categorical pattern [28].

If future work replicates our finding that consistent native perception, attention, and memory relate to the success of non-native perception, then this could lead to the development of brief pre-screenings to identify language learners who would benefit from more support during learning. Overall, our findings contribute to the understanding of individual differences in native and non-native phonetic perception, and we encourage further investigation of the factors (such as consistency) that may play a role in the successful perception of non-native phonemes.

5. REFERENCES

[1] K. Hattori and P. Iverson, "English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy," *J Acoust Soc Am*, vol. 125, no. 1, p. 12, 2009.

[2] J. E. Flege, O.-S. Bohn, and S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *J. Phon.*, vol. 25, no. 4, pp. 437–470, Oct. 1997.

[3] L. R. Slevc and A. Miyake, "Individual Differences in Second-Language Proficiency: Does Musical Ability Matter?," *Psychol. Sci.*, vol. 17, no. 8, pp. 675–681, Aug. 2006.

[4] M. Kachlicka, K. Saito, and A. Tierney, "Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound," *Brain Lang.*, vol. 192, pp. 15–24, May 2019, doi: 10.1016/j.bandl.2019.02.004.

[5] E. C. Kapnoula, M. B. Winn, E. J. Kong, J. Edwards, and B. McMurray, "Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 43, no. 9, pp. 1594–1611, Sep. 2017, doi: 10.1037/xhp0000410.

[6] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries.," *J. Exp. Psychol.*, vol. 54, no. 5, pp. 358–368, 1957.

[7] E. C. Kapnoula, J. Edwards, and B. McMurray, "Gradient Activation of Speech Categories Facilitates Listeners' Recovery From Lexical Garden Paths, But Not Perception of Speech-in-Noise," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 47, no. 4, pp. 578–595, 2021.

[8] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Learning & Language Teaching*, vol. 17, O.-S. Bohn and M. J. Munro, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 13–34.

[9] P. Fuhrmeister, "Neural Correlates of Native-language Speech Perception and Non-native Speech Sound Learning," Doctoral thesis, University of Connecticut, 2020.

[10] A. Lengeris and V. Hazan, "The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek," *J. Acoust. Soc. Am.*, vol. 128, no. 6, Dec. 2010.

[11] B. Díaz, C. Baus, C. Escera, A. Costa, and N. Sebastián-Gallés, "Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language," *Proc. Natl. Acad. Sci.*, vol. 105, no. 42, pp. 16083–16088, Oct. 2008.

[12] R. Sparks and L. Ganschow, "The Impact of Native Language Learning Problems on Foreign Language Learning: Case Study Illustrations of the Linguistic Coding Deficit Hypothesis," *Mod. Lang. J.*, vol. 77, no. 1, pp. 58–74, 1993.

[13] N. Golestani and R. J. Zatorre, "Learning new sounds of speech: reallocation of neural substrates," *NeuroImage*, vol. 21, no. 2, pp. 494–506, Feb. 2004.

[14] F. S. Earle and D. T. Arthur, "Native phonological processing abilities predict post-consolidation nonnative contrast learning in adults," *J. Acoust. Soc. Am.*, vol. 142, no. 6, pp. EL525–EL531, Dec. 2017.

[15] M. Clayards, "Differences in cue weights for speech perception are correlated for individuals within and across contrasts," *J. Acoust. Soc. Am.*, vol. 144, no. 3, pp. EL172–EL177, Sep. 2018.

[16] R. Mayr and P. Escudero, "Explaining individual variation in L2 perception: Rounded vowels in English learners of German," *Biling. Lang. Cogn.*, vol. 13, no. 3, pp. 279–297, Jul. 2010.

[17] C. K. Conners, J. N. Epstein, A. Angold, and J. Klaric, "Continuous Performance Test Performance in a Normative Epidemiological Sample," p. 8, 2003.

[18] D. Weschler, "Wechsler Adult Intelligence Scale, (WAIS-IV)," *San Antionio TX Psychol. Corp.*, 2008.

[19] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2020.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *J. Stat. Softw.*, vol. 67, no. 1, 2015.

[21] The MathWorks Inc., "MATLAB." Natick, Massachusetts, 2015a.

[22] M. M. Hughes, J. A. Linck, A. R. Bowles, J. T. Koeth, and M. F. Bunting, "Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity," *Behav. Res. Methods*, vol. 46, no. 3, pp. 702–721, Sep. 2014.

[23] Y. Kang and J. Schertz, "The influence of perceived L2 sound categories in on-line adaptation and implications for loanword phonology," *Nat. Lang. Linguist. Theory*, vol. 39, no. 2, May 2021.

[24] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, "Vocabulary Size is Associated with Second-Language Vowel Perception Performance In Adult Learners," *Stud. Second Lang. Acquis.*, vol. 33, no. 3, pp. 433–461, Sep. 2011.

[25] E. Pederson and S. Guion-Anderson, "Orienting attention during phonetic training facilitates learning," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. EL54–EL59, Feb. 2010, doi: 10.1121/1.3292286.

[26] S.-P. Lee, "Computer-Detected Attention Affects Foreign Language Listening but Not Reading Performance," *Percept. Mot. Skills*, vol. 123, no. 1, pp. 33–45, Aug. 2016.

[27] J. R. McHaney, R. Tessmer, C. L. Roark, and B. Chandrasekaran, "Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry," *Brain Lang.*, vol. 222, Nov. 2021.

[28] F. Desmeules-Trudel, "Spoken word recognition in native and second language Canadian French: Phonetic detail and representation of vowel nasalization," Doctoral thesis, University of Ottawa, 2018.

CLEAR SPEECH FACILITATES WORD SEGMENTATION: EVIDENCE FROM EYE-TRACKING

Zhe-chen Guo, Rajka Smiljanic

The University of Texas at Austin
zcadamguo@utexas.edu, rajka@austin.utexas.edu

ABSTRACT

Listener-oriented hyperarticulated clear speech improves word recognition in noise and auditory memory. Using visual-world eye-tracking, this study examined whether clear speech benefits spoken word segmentation and how its effect develops over time. English-speaking listeners heard clear and conversational sentences in which the target word (e.g., *ham*) was temporarily ambiguous with a competitor (e.g., *hamster*) across a word boundary (e.g., *she saw the ham starting*…). Analysis of the listeners' eye fixations indicated that the likelihood of fixating the picture of the target compared to that of the competitor was higher for clear speech than conversational speech. The difference emerged even before segmental information for resolving the ambiguity was available in clear speech. The findings showed that speaking clearly facilitated listeners' discovery of word boundaries. Improved speech segmentation and reduced lexical competition may in part underlie the clear speech processing benefits.

Keywords: Clear speech, speech segmentation, hyperarticulation, eye-tracking

1. INTRODUCTION

Words in fluent speech are not consistently separated by pauses [1]. To successfully recognize words and comprehend the spoken message, the listener needs to find the word boundaries or to achieve speech segmentation. This study aims to provide further insight into such an ability by examining how intelligibility-enhancing hyperarticulated speaking styles affect the listener's speech segmentation.

Past research has revealed a wide variety of signal-dependent and -independent cues used during segmentation (see [2] for a review). The phrase *night rates* can be distinguished from *nitrates* based on allophonic differences [3]. Stressed syllables [4], [5] and vowel lengthening [6], [7] are used to locate word-initial and -final positions, respectively. Signal-independent lexical cues, such as knowing that *anything* is a word, allows listeners to detect the fragment *corri* more rapidly in *anythingcorri* than in *imoshingcorri* [8]. Speech does not even have to be meaningful to be segmented. Exposed to nonsense syllable sequences repeated continuously, listeners can still extract the sequences by tracking syllable co-occurrence frequencies [6], [9].

While these findings improve the understanding of what cues are useful for identifying word boundaries, speech segmentation is rarely explored in connection with phonetic variation in real-word communication (cf. [10]). In daily interactions, speakers dynamically adjust their output along a hypo-hyperspeech continuum, reflecting a balance between speaker- and listener-oriented forces (H&H theory: [11]). Under challenging listening conditions (e.g., noisy environment or non-native listener), they produce intelligibility-enhancing hyperarticulated "clear speech" [12]. Relative to casual or conversational speech, clear speech shows various acoustic-phonetic enhancements (e.g., longer segment duration, vowel space expansion) and benefits perceptual processes including word recognition and auditory memory (for reviews, see [13], [14]).

Recent work has begun to explore these perceptual benefits in finer detail by examining how clear speech impacts listeners' segmentation. In [15]'s artificial language learning study, English listeners heard speech streams containing continuous repetitions of nonsense words and were tasked with segmenting the words based on the statistical patterns of the words' component syllables. Results showed that in quiet, their segmentation was more accurate with the speech streams produced clearly than with those produced conversationally. These findings suggest that the well-established clear speech benefits for word recognition may partly be aided by improved segmentation. Yet, unlike the participants in [15], listeners in real-world communication rarely parse nonsense speech. It remains unclear the extent to which speaking clearly improves segmentation of meaningful words as the speech signal unfolds.

We addressed this issue in a visual-world eye-tracking experiment [16]. Unlike traditional behavioral tasks (e.g., word-spotting [17]), eye-tracking provides rich temporal information of how the evolving speech is integrated to locate word boundaries. In our experiment, listeners heard "lexical garden-path" sentences [18], [19] such as *She saw the ham starting to get crispy and brown*, in which the string *ham st-* overlaps with the unintended *hamster*. Models of spoken word recognition (e.g., [20], [21]) assume that both *ham* and *hamster* are

activated and compete for lexical selection until the signal mismatches *hamster* at the "disambiguation point (DP)" (e.g., onset of /ɑ/ in *starting* in the above sentence). The question of interest concerns whether clear speech provides more word boundary robust cues to allow listeners to consider *ham* over *hamster* and at what point in the evolving speech signal this benefit occurs. To explore this, we tracked English listeners' gaze at related images as they listened to the sentences produced clearly and conversationally.

2. EYE-TRACKING EXPERIMENT

2.1. Hypothesis

Based on [15]'s findings, it was hypothesized that relative to conversational speech, clear speech similarly aids segmentation of meaningful words such that listeners can locate the word boundary more effectively even before the DP is reached. If this was the case, we expected that within the analysis time window (see Section 2.5), the likelihood of fixating the image of the target (e.g., *ham*) as compared with that of fixating the competitor image (e.g., *hamster*) would be significantly higher for clear speech than for conversational speech.

2.2. Stimuli

Critical items were twenty-six pairs of picturable English nouns, each containing a monosyllabic target word (e.g., *ham*) and a disyllabic, initially stressed competitor in which the target was embedded as onset (e.g., *hamster*). The log frequencies per million words of the targets (mean: 2.76) and competitors (mean: 2.06) were matched as closely as possible using the CLEARPOND database [22] to minimize frequency effects [23]. For each pair, a target-bearing sentence was constructed such that the target plus the onset of the next word overlapped partially with the competitor (e.g., *She saw the* ***ham*** *starting to get crispy and brown*). Two additional sentences were created for each pair: one in which the intended word was the competitor (e.g., *She saw the* ***hamster*** *running quickly on the wheel*) and the other in which it was a mono- or di-syllabic distractor word semantically and phonologically unrelated to the target and competitor (e.g., *He wanted to find the* ***turkey*** *under the tree*). The sentences began in a way that would not bias towards either the target, competitor, or distractor. Seventy-eight unique sentences (26 target-competitor pairs × three sentence types) were constructed.

A female native speaker of American English produced the sentences first conversationally and then clearly. For the conversational style, she was instructed to speak "as if to a friend or someone familiar with her voice." For the clear style, she was instructed to speak "as if to someone hard-of-hearing or a non-native speaker." The instructions have been shown to be successful in eliciting the style distinctions [14], [24], as confirmed by a preliminary acoustic analysis showing that, for example, segments in clear speech were longer. The 156 sentence stimuli (78 sentences × two styles) were recorded using a Shure SM10A head-mounted microphone at a 44.1k Hz sampling rate as WAV files.

The stimuli were each paired with a display containing four drawings in the quadrants to form 156 trials. The drawings depicted the target, competitor, a monosyllabic distractor, and a disyllabic distractor, as shown in Fig. 1. In the critical trials, the drawing of the target was mentioned in the sentence while in the filler trials, the mentioned drawing was the competitor or one of the distractors. Drawings of each type appeared in each quadrant roughly equally often. They were taken mostly from normed drawing databases (e.g., [25]) and edited to ensure consistent visual features when necessary.

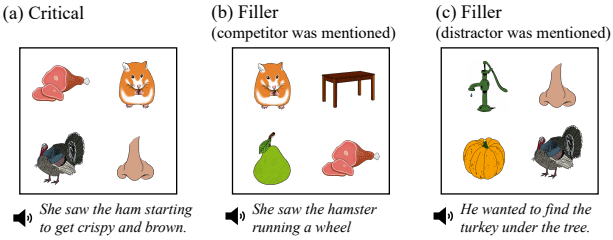


Figure 1: Sample visual arrays of the three trial types

The trials were distributed over two lists of 78 trials such that participants were presented with clear and conversational speech but never with the same sentence in both styles. Each list was divided into three blocks, with block order and trial order within a block randomized. Trials of the three types in Fig. 1 and of the two speech styles occurred (approximately) an equal number of times.

2.3. Procedure

Participants were first familiarized with the drawings and their labels. They were then seated about 60 cm from a computer screen with their heads on a chin rest and assigned to either list. Each trial began with a fixation cross that served as a drift check. When fixated, the cross was displaced by four drawings and three seconds later, a sentence was played in quiet via headphones. Participants clicked the drawing mentioned in the sentence as quickly and accurately as possible. They completed five trials with feedback as practice before the experiment proper. Their right eyes were tracked at 500 Hz with an EyeLink Portable Duo (SR research) eye-tracker, calibrated with a

standard nine-point grid before the practice and each block and when recalibration was needed. The experiment was run using Experiment Builder.

2.4. Participants

Forty-eight native American English speakers participated in the study (29 females; age range: 16-29, mean age: 19.1). All passed a pure-tone screening binaurally at 25 dB HL for 1000, 2000, and 4000 Hz. None reported hearing, speech, or vision disorders. Nine participants were excluded for calibration failure, not fixating any image in almost all trials, or having response accuracy 2.5 standard deviations below the mean for at least one speaking style.

2.5. Data analysis

We tested the hypothesis by comparing proportions of fixations to target versus competitor images in the critical trials within a selected analysis time window. This window began 200 ms after target word onset as it takes about 200 ms to launch a saccade [26]. The end point of the window was sentence-specific but the same for the clear and conversational versions of each sentence—it was the DP of the clear stimulus. This was done as the durations from target onset to the DP in conversational stimuli (mean: 259 ms) could be too short to provide meaningful data when the initial 200 ms was excluded. Thus, since conversational speech was faster, the analysis window included some post-DP segments for the conversational counterpart of each clear stimulus. Such a window provided a stringent and realistic test of the hypothesis: clear speech prior to the DP may contain more robust word boundary cues than conversational speech with post-DP information for ruling out the competitor.

The analysis window of each sentence was divided equally into 10 time bins. Within each bin, proportions of fixations were calculated separately for the target and competitor images and transformed into empirical logits [27]. With this we derived from each trial two empirical-logit curves representing fixation likelihoods over normalized time—one for the target and the other for the competitor.

3. RESULTS

The valid participants (N = 39) clicked with >92% mean accuracy across trial types and styles. Fixations were coded by the eye-tracker's default algorithm as directed to one (or none) of the images on the display and those in the critical trials with correct responses were analyzed as in Section 2.5. Fig. 2 shows the mean empirical logits over the 10 time bins within the analysis window by image type and speaking style.

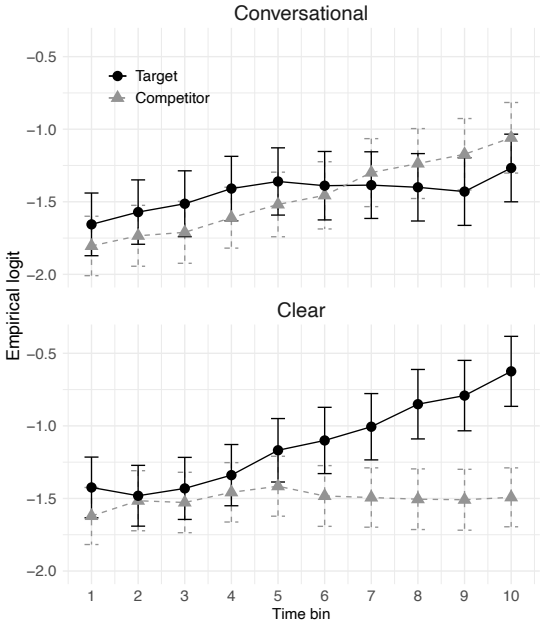


Figure 2: Mean empirical logit at each time bin within the analysis window for fixation data in correct critical trials by speaking style and image type.

Inspecting Fig. 2 revealed that the empirical-logit curve representing the likelihood of target fixations diverged from the curve for competitor fixations to a larger extent for clear speech than for conversational speech. To test the statistical significance of the pattern, we analyzed the curves using generalized additive mixed modeling (GAMM: [28]) implemented in the *mgcv* [29] package of R [30]. Following [31], we modelled the time-varying effects of image type and style and their interaction with smooth functions for three binary predictors: IsTarget (coded as 1 for the curves of the target image and 0 for those of the competitor), IsClear (coded as 1 for the curves of clear speech and 0 for those of conversational speech), and IsTargetClear (coded as 1 for the curves of the target in the clear condition; otherwise, 0). Of interest was the smooth of IsTargetClear, for which positive estimates indicated that target fixation proportions compared with competitor fixation proportions were relatively higher for clear speech than for conversational speech. By-participant and by-sentence factor smooths for the three predictors were included as random effects. Autocorrelated residuals were corrected using an AR(1) error model with $\rho = 0.823$.

All smooth terms were significant ($p < .001$). As these terms were not immediately interpretable, the hypothesis was tested by visualizing the smooth of IsTargetClear over time normalized between 0 and 1. As Fig. 3 shows, its effect was positive and significant from 0.576 to the end. That is, starting around halfway of the analysis window, the difference of the target from the competitor in fixation proportions began to be significantly larger for clear speech than

conversational speech. This result was consistent with the hypothesized clear speech segmentation benefit.

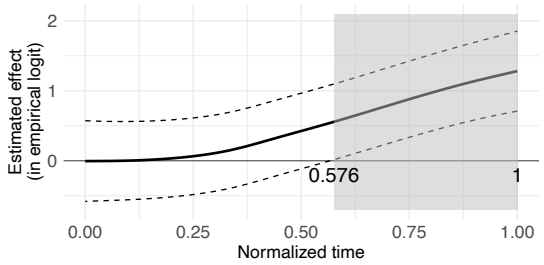


Figure 3: Estimated effect of IsTargetClear over normalized time. The dashed lines mark the 95% confidence interval. The time interval during which the effect was significant is shaded.

4. DISCUSSION

Using visual-world eye-tracking, we examined whether clear speech facilitates word segmentation relative to conversational speech. English-speaking listeners heard clear and conversational sentences with temporary ambiguity between the target and a competitor across a word boundary. Analyzing their eye fixation patterns revealed that the proportions of target fixations as compared with those of competitor fixations were significantly higher for clear speech than conversational speech at 0.576 into the analysis time window. This suggested that clear speech led listeners to locate the word boundary more effectively and consider the intended target word more. Thus, speaking clearly improves not only segmentation of nonsense speech streams as found in [15] but also segmentation of meaningful words during real-time speech processing. Improved word segmentation may in part underlie the clear speech benefits for linguistic and cognitive processes, including word recognition and auditory memory [13], [14].

The findings cannot be attributed to poor intelligibility of conversational speech. This possibility is discounted by a further analysis comparing fixations to the two related images (i.e., target and competitor) versus the unrelated distractor images across the two styles. The results showed that throughout the analysis window, the target and competitor on the one hand were distinguished from the distractors on the other equally well for both styles. This means that the listeners could accurately perceive the critical segments (e.g., *ham st-*) and reject the distractors in conversational speech. The conversational stimuli were just more ambiguous between the target and competitor than the clear ones, resulting in the patterns in Fig. 2.

The current evidence for the clear speech segmentation benefit is noteworthy considering the time window selected for the analysis. As mentioned, the window ended at the DP for the clear sentence stimuli but included some segments after the DP for their conversational counterparts. Word recognition models (e.g., [20], [21]) would posit that during the time window, conversational speech should disambiguate the target from the competitor more than clear speech as post-DP segments mismatching the competitor are available. The results, however, suggested the opposite. As our acoustic analysis found segmental and prosodic enhancements (e.g., longer duration, greater F0 range, etc.) in clear speech, boundary cues like stress [5] and vowel lengthening [19] might also be exaggerated. These enhanced cues then jointly promote activation of the target while reducing competition from unintended competitors—an effect not captured by the lexical competition dynamics assumed in the word recognition models. A pressing goal would be to examine the extent to which each boundary cue contributes to segmentation when enhanced through clear speech modifications.

Our findings further suggest that the relative importance of signal-dependent and signal-independent segmentation cues may be altered under hyperarticulation. It has been proposed that at least in optimal listening conditions, lexical and semantic cues generally outweigh signal-dependent acoustic-phonetic and prosodic cues [10], [32], [33]. Yet, this view is based on experimentation with a single cue from each cue type without consideration of hyperarticulated clear speech. The advantage of clear speech over conversational speech with segmental information for ruling out the competitor implies that listeners may put more weight on the signal-dependent cues enhanced through hyperarticulation, potentially outranking lexico-semantic information. Future work is needed to examine how the clear speech benefit changes in the presence of, for example, semantic context to the target word.

Another further question is to examine the extent to which clear speech facilitates word segmentation in adverse conditions, such as environmental noise. Since noise increases lexical competition from competitor words [34], it is of interest to examine whether the clear speech segmentation benefit will be reduced or delayed in noise as a result. Another issue concerns whether clear speech also aids segmentation for other listener populations such as non-native or hard-of-hearing listeners. The findings would add to the understanding of the mechanism underlying the clear speech intelligibility benefit in realistic communicative contexts.

5. ACKNOWLEDGMENT

The authors would like to thank Madeline Smith, Eliana Spradling, and Madison Rider for their assistance with data collection.

6. REFERENCES

[1] Cole, R. A., Jakimik, J., Cooper, W. E. 1980. Segmenting speech into words. *J. Acoust. Soc. Am.* 67, 1323–1332.

[2] Cutler, A. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press.

[3] Gow, D. W., Gordon, P. C. 1995. Lexical and prelexical influences on word segmentation: Evidence from priming. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 344–359.

[4] Cutler, A., Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. *Comput. Speech. Lang.* 2, 133–142.

[5] Tyler, M. D., Cutler, A. 2009. Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am.* 126, 367–376.

[6] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: the role of distributional cues. *J. Mem. Lang.* 35, 606–621.

[7] White, L., Benavides-Varela, S., Mády, K. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J. Phon.* 81, 100982.

[8] White, L., Melhorn, J. F., Mattys, S. L. 2010. Segmentation by lexical subtraction in Hungarian speakers of second-language English. *Q. J. of Exp. Psycho.* 63, 544–554.

[9] Erickson, L. C., Thiessen, E. D. 2015. Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review* 37, 66–108.

[10] White, L., Mattys, S. L., Wiget, L. 2012. Segmentation cues in conversational speech: robust semantics and fragile phonotactics. *Front. Psychol.* 3, 1–9.

[11] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W. J., Marchal, A. (eds), *Speech Production and Speech Modelling*. Springer Netherlands, 403–439.

[12] Uchanski, R. M. 2005. Clear speech. In: Pisoni, D., Remez, R. (eds), *The Handbook of Speech Perception*. Blackwell, 207–235.

[13] Smiljanić, R., 2021. Clear speech perception. In: Nygaard, L. C., Pardo, J., Pisoni, D., Remez, R. (eds), *The Handbook of Speech Perception*. Wiley, 177–205.

[14] Pichora-Fuller, M. K., Goy, H., van Lieshout, P. 2010. Effect on speech intelligibility of changes in speech production influenced by instructions and communication environments. *Semin. Hear.* 31, 77–94.

[15] Guo, Z.-C., Smiljanić, R. 2021. Speaking clearly improves speech segmentation by statistical learning under optimal listening conditions. *Laboratory Phonology* 12, 14.

[16] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634.

[17] McQueen, J. 1996. Word spotting. *Lang. Cogn. Process.* 11, 695–699.

[18] Davis, M. H., Marslen-Wilson, W. D., Gaskell, M. G. 2002. Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 218–244.

[19] Salverda, A. P., Dahan, D., McQueen, J. M. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90, 51–89.

[20] Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102.

[21] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86.

[22] Marian, V., Bartolotti, J., Chabal, S., Shook, A. 2012. CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS One* 7, e43230.

[23] Dahan, D., Magnuson, J. S., Tanenhaus, M. K. 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cogn. Psychol.* 42, 317–367.

[24] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: talker and listener factors in speaking style changes. *Lang. Linguist. Compass* 3, 236–264.

[25] Duñabeitia, J. A., et al. 2018. MultiPic: a standardized set of 750 drawings with norms for six European languages. *Q. J. of Exp. Psychol.* 71, 808–816.

[26] Fischer, B. 1992. Saccadic reaction time: implications for reading, dyslexia, and visual cognition. In: Rayner, K. (ed), *Eye Movements and Visual Cognition*. Springer, 31–45.

[27] Barr, D. J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457–474.

[28] Lin, X., Zhang, D. 1999. Inference in generalized additive mixed models by using smoothing splines. *J. R Stat. Soc. Series B Stat. Methodol.* 61, 381–400.

[29] Wood, S. N. 2021. mgcv: mixed GAM computation vehicle with automatic smoothness estimation.

[30] R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing.

[31] van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., Wood, S. N. 2019. Analyzing the time course of pupillometric data. *Trends Hear.* 23, 1–22.

[32] Mattys, S. L., White, L., Melhorn, J. F. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol. Gen.* 134, 477–500.

[33] Mattys, S. L., Bortfeld, H. 2017. Speech segmentation. In: Gaskell. M. G., Mirković, J. (eds), *Speech Perception and Spoken Word Recognition*. Routledge, 55–75.

[34] Ben-David, B. M., Chambers, C. G., Daneman, M., Pichora-Fuller, M. K., Reingold, E. M., Schneider, B. A. 2011. Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. *J. Speech Lang. Hear. Res.* 54, 243–262.

CUE SHIFTING IN THE PERCEPTION OF SHANGHAINESE TONE SANDHI PATTERNS

Zihao Wei

Chinese University of Hong Kong
zihaowei@cuhk.edu.hk

ABSTRACT

The role of cue reweighting in reshaping tone sandhi patterns is rarely discussed. The consensus in Shanghainese disyllabic prosodic words is that F0 lexical tonal differences are neutralized in the second syllable and that VOT and closure duration are the primary cues in maintaining the voicing/register contrast. To investigate perception strategies in different age groups, an AXB forced choice experiment is used. The findings of the current study are generally consistent with the hypothesis that change occurs when listeners reanalyze coarticulatory effects used in speech production (e.g., Ohala, 1993). The F0 split in sandhi patterns is most likely due to younger speakers no longer associating stable VOT differences with the onset stop voicing contrast in the second syllable and instead begin to associate the difference between high- and low-register words with the voicing-condition F0 difference. This cue shift exemplifies how perception leads to the process of reanalysis.

Keywords: Tonogenesis, Cue weighting, Sound change, Sandhi pattern, Shanghainese

1. INTRODUCTION

It is well-known that obstruents affect the F0 of the following vowel, and the vowels following a voiceless onset tend to have relatively higher F0 in many languages (e.g., Kirby & Ladd, 2015). Furthermore, in tonal languages where F0 differences indicate lexical contrasts, the relation between voicing and F0 is often manifested as co-occurrence restrictions between tone and the onset consonant of tone-bearing syllables.

In Shanghainese, a disyllabic prosodic word gets its surface sandhi tone by delinking the tone of the second syllable and spreading the tonemes of the first syllable over the two syllables. For example, the prosodic word /tsɔ^{3a} vɛ^{3b}/ 'fried rice' that could be autosegmentally represented as /ts^{h}ɔ^{MH} vɛ^{LH}/ has a surface form [ts^{h}ɔ^{M} vɛ^{H}] (e.g., Yip, 1995). In Shanghainese, as traditionally described, the functional cost of neutralizing tone in second syllables is mitigated by the preservation of vocal fold vibration. Some previous studies suggested that vocal fold vibration is no longer the primary cue (e.g., Chen & Wang, 2012), while others show that the two categories have multidimensional acoustic and articulatory correlates, including a significant difference in Closure of Duration of onset stop (CD) (Shen, Wooter & Wang, 1987) and in breathiness concentrated mainly at the onset of the following vowel (e.g., Ren, 1992).

The lexical contrasts can still be maintained when the primary cue changes. Such a change is reminiscent of tonogenesis, a type of cue shift that has been studied extensively (e.g., Hyman, 1973). The result of our previous production study (Wei, 2022) shows that the tone sandhi pattern in Shanghainese is changing: in the second syllable of a sandhi domain, F0 is overtaking CD and VOT as the primary cue in younger generations. In this work, we will examine the role of these cues in perception and whether there are generational differences in the relative importance of these different cues. Research on the mapping between production and perception cues is very inspiring when looking at an ongoing diachronic change. Only when we know how cues shift in both production and perception can we further draw a picture of the first trigger in a particular case of sound change where the cue shifting in both production and perception on the timeline is well established.

Cue weighting is a process of conceptualizing, quantifying, and ranking the cues associated with a contrast, which encompasses both perception and production (Schertz, 2019). The weighting of a cue for a phoneme may show variations in both production and perception depending on age, and the weights of cues in production and perception do not always coincide. Research on tonogenesis in Korean finds that speakers of different age groups prefer different strategies in speech perception (e.g., Kang & Guion, 2008; Kang, 2009, 2014; Silva, 2006; Wright, 2007). Meanwhile, studies on the production and perception of /u/- and /ʊ/-fronting of Standard Southern British English (e.g., Harrington, 2012; Harrington, Kleber, & Reubold, 2008; Kleber, Harrington, & Reubold, 2012) revealed that the relationship between production and perception is unstable during diachronic change. As a result, the comparison of cue weighting in perception and production bears on the theoretical issue of the processes underlying sound change.

2. EXPERIMENT

To test the hypothesis that both F0 and VOT are prominent cues in the perception of domain-medial onset stops, two separate sets of AXB forced choice experiments are used to investigate the participants' perception strategies in different age groups of Shanghainese speakers.

2.1. Participants

All the experiments in this study were conducted online using the Gorilla online application (https://app.gorilla.sc/admin/home) with 76 native Shanghainese speakers. There were 36 young participants (age range 20-40, mean 27.83), 24 middle-aged participants (age range 40-60, mean 49.63), and 16 elder participants (age > 60, mean 64.23). All participants spoke Standard Mandarin with different degrees of Shanghainese interference.

2.2. Stimuli

Three acoustic properties of the second syllable were manipulated along acoustic continua: the F0 slope, the VOT and the closure duration of the onset. The synthesized stimuli were produced based on the natural speech of a female Shanghainese speaker in her 20s. We chose two minimal pairs: 小巴 /ɕiɔ^{hL} pa^{H}/ "minibus" vs. 小排 /ɕiɔ^{hL} ba ^{H}/ "rib", 大巴 /da^{lL} pa^{H}/ "bus" vs. 大排 /da^{lL} ba^{H}/ "chop". The duration of the target vowel was normalized to 160 ms, and the onset duration of the domain-medial /p/ (i.e., Closure duration + VOT) was set to 125 ms, while that of the domain-medial /b/ was 35 ms. Next, the F0 at the offset of the vowel of /pa/ was set to 260 Hz, and that of the offset of the vowel of /ba/ was set to 220 Hz as in a mid-high flat tone. The normalization is based on the natural range of the speaker. Finally, five fundamental frequency curves with different slopes were generated on each of the second syllables by decreasing the F0 of their initial points in 10Hz steps. The specific stimuli information of F0 after the manipulation is presented in Figure 1.

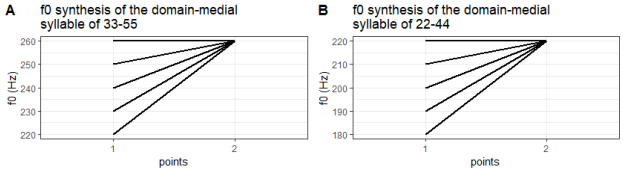


Figure 1: Steps of F0 onset of synthesized stimuli

Table 1 illustrates the VOT, closure duration, and F0 manipulations for one token as a cube in a 3-dimensional acoustic space (2 sandhi patterns × 4 F0 onset × 3 closure duration × 4 VOT = 60 stimuli). The 3a-1a/3b sandhi tone is shown on the top left side, and the 3b-1a/3b sandhi tone is shown on the bottom left side. Altogether, the stimuli varied in VOT, closure duration, and F0 slopes.

	ɕiɔ^{33} Pa^{55}	da^{22} Pa^{44}
	2nd syllable	2nd syllable
F0 onset (Hz)	260, 250, 240, 230, 220	220, 210, 200, 190, 180
Closure Duration (ms)	115, 70, 25	115, 70, 25
VOT (ms)	Voiceless: 10 Voiced: -25, -70, -115	Voiceless: 10 Voiced: -25, -70, -115

Table 1: Demonstration of the manipulation of VOT (four levels), Closure Duration of the onset (three levels), and F0 onset (five levels) of the second syllable.

2.3. Procedures

Each participant was then asked to listen to each stimulus five times in the experiment. As a result, each speaker had to identify 300 tokens (60*5). Participants were asked to complete a training session consisting of 20 disyllabic prosodic words to understand the requirements of the experiment. The training instructions asked participants to listen to a stimulus and to select one of the two corresponding disyllabic prosodic words that appeared on the screen by pressing A or D within ten seconds after hearing the recording. The formal experiment consisted of four blocks containing 75 tokens, with a mandatory one-minute break after each section. All responses given in less than 160 ms were removed as we considered them too fast to be valid. For the remaining data points, we remove data points smaller than Q1 - 1.5 * (Q3 - Q1) (Q1: lower tertile, Q3: higher tertile), and data points larger than Q1 + 1.5 * (Q3 - Q1). The final response times obtained for all data points are between 160 ms and 3000 ms. In the perception study, we use the LDA to weight different cues in perception (Schertz, 2019).

3. RESULTS

Looking at Figure 2 below, we can see that the slope of the F0 curve affected the perception of the voicing contrast in second syllables in all age groups, irrespective of VOT and closure duration. Broadly speaking, the lower the F0 was (i.e., the larger the slope of the F0 curve was), the more likely the stop onset was perceived as voiced, and the higher it was,

the more likely the stop onset was perceived as voiceless. In general, CD played a secondary role in the perception of second syllable, but this secondary role was much stronger in older than younger participants. VOT, on the other hand, had little effect on the perception in either of the three age groups.

In general, young speakers highly relied on the slope of F0 in their perception, and their perception was categorical: changes in CD had a relatively small effect on the perception of the voicing contrast. In the perception of elder and middle-aged participants, F0 was less categorical and CD played a greater role (middle-aged speakers fall between older and younger ones). Finally, VOT had little effect in the perception of all three age groups.

In summary, we found that the perceptual strategies of the three age groups changed over time:

- The elder group relied heavily on both CD and the slope of the F0 curve, but did not seem to rely on VOT in perception, despite using in in production.
- The middle-aged group also used F0 slope as a primary cue but relied significantly less on CD.
- The young group tended to use the slope of the F0 curve as the primary cue and used CD as a much more peripheral cue than other groups.

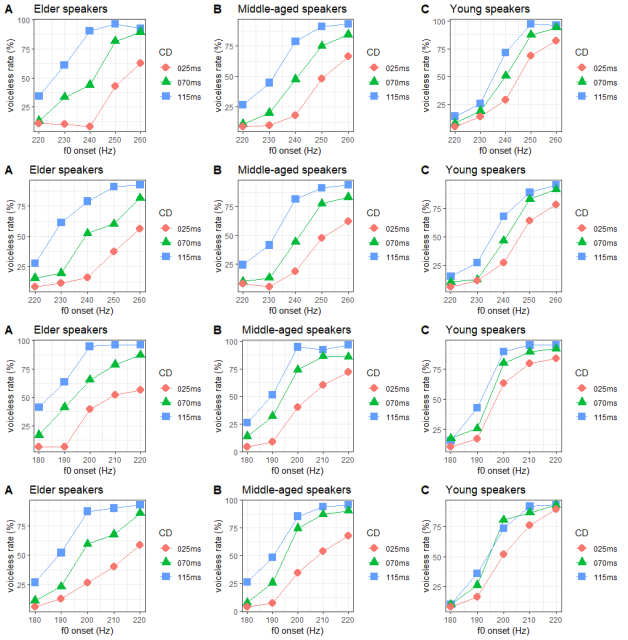


Figure 2: Rate of perceived as voiceless in 3a-1a/3b sandhi tone, VOT > 0 (1st row); in 3a-1a/3b sandhi tone, VOT < 0 (2nd row); in 3b-1a/3b sandhi tone, VOT > 0 (3rd row), and in 3b-1a/3b sandhi tone, VOT < 0 (4th row)

	Age group	X-squared	*p*-value	error rate
	Elder	666.86	$< 2.2e^{-16}$	23.37%
/pɔTɛ3a-1a/3b/	Middle	1057	$< 2.2e^{-16}$	22.32%
	Young	2016.3	$< 2.2e^{-16}$	18.41%
/bɔTɛ3b-1a/3b/	Elder	603.88	$< 2.2e^{-16}$	24.22%
	Middle	1166.1	$< 2.2e^{-16}$	20.45%
	Young	1982.1	$< 2.2e^{-16}$	17.75%

Table 2: LDA Results of the perception experiments

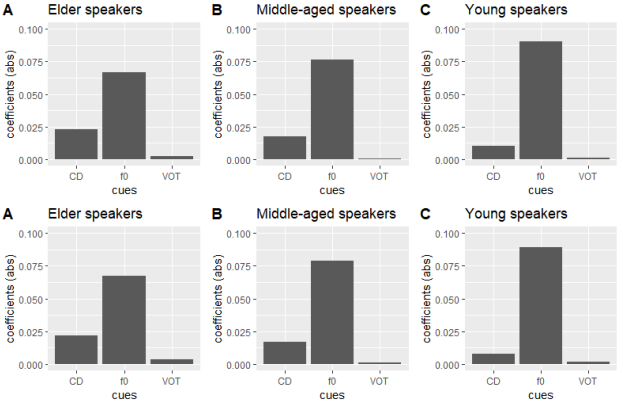


Figure 3: LDA results of perception, in sandhi tone 3a-1a/3b (top) and 3b-1a/3b (botton).

An examination of LDA results, in Figures 6 and 7 confirmed the observations made from the raw results. VOT, an outstanding production cue among the elder speakers, was the least valued perceptual cue for all three age groups. For elder speakers, F0 was already the most important perceptual cue, followed by closure duration. This pattern was equally valid for all three age groups, the main difference between them being the relative importance of F0 slope, that gradually increased for the middle-aged and young groups. In contrast, the importance of the closure duration diminished with age. Based on the findings of the perception experiment, we can say that the perception strategies used by the three generations are similar, but vary quantitatively.

4. DISCUSSION AND CONCLUSION

In perception, F0 is always the primary cue for all three age groups, and CD is gradually losing its importance as a secondary cue. VOT, as opposed to F0 and CD, has little effect in the perception of the three age groups. However, the limited role of VOT could be due to our speaker being a 24-year-old young woman. Although in her actual pronunciation, the VOT of syllables with phonologically voiced onset stops in the second syllable is negative, we cannot exclude that Shanghainese listeners are aware that young speakers rely less on VOT and more on F0 and adjust their perception to the young voice used in the experiment. Therefore, to further investigate the changes in the tone sandhi pattern of disyllabic prosodic words in Shanghainese, we would need to test stimuli produced by middle-aged and elder speakers as well.

5. ACKNOWLEDGMENT

This work is based on a section of the author's memoir in uOttawa, supervised by Professor Marc Brunelle.

6. REFERENCES

[1] Chen, Z., & Wang, Y., Roles of F0 and closure in the voiced-voiceless distinction for initial stops in Wu dialects—take Shanghainese for example. *In Tonal Aspects of Languages-Third International Symposium*. 2012.

[2] Harrington, J., The coarticulatory basis of diachronic high back vowel fronting. In M.-. J. Solé & D. Recasens (Eds.), The initiation of sound change. Perception, production, and social factors (pp. 103–122). Amsterdam: John Benjamins Publishing Company. 2012.

[3] Harrington, J., Kleber, F., & Reubold, U., Compensation for coarticulation, /u/-fronting, and sound change in Standard Southern British: An acoustic and perceptual study. Journal of the Acoustical Society of America, 123, 2825–2835. 2008.

[4] Hyman, L., (Ed.) Consonant types and tone: Southern California occasional papers in linguistics. *Los Angeles: The Linguistic Program, University of Southern California*. 1973.

[5] Kang, K. H., & Guion, S. G., Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. The Journal of the Acoustical Society of America, 124(6), 3909–3917. 2008.

[6] Kang, K. H., Clear speech production and perception of Korean stops and the sound change in Korean stops (Doctoral dissertation). Eugene: University of Oregon. 2009.

[7] Kang, Y., Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. Journal of Phonetics, 45, 76–90. 2014

[8] Kirby, J. P., Ladd, D. R., Stop voicing and F0 perturbations: Evidence from French and Italian[C]//ICPhS. 2015.

[9] Kleber, F., Harrington, J., & Reubold, U., The relationship between the perception and production of coarticulation during a sound change in progress. Language and Speech, 55, 383–405. 2012.

[10] Ohala, J. J., Coarticulation and phonology [J]. Language and speech, 1993, 36(2-3): 155-170.

[11] Ren, N., *Phonation types and consonant distinctions: Shanghai Chinese*. Ph.D. Dissertation. The University of Connecticut. 1992.

[12] Schertz, J., Clare, E. J., Phonetic cue weighting in perception and production[J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2020, 11(2): e1521.

[13] Shen, Z., Wooters, C., & Wang, W. S. Y., Closure duration in the classification of stops: A statistical analysis. *OSU Working Papers*, 35, 197–209. 1987.

[14] Silva, D. J., Variation in Voice Onset Time for Korean stops: A case for recent sound change. Korean Linguistics, 13(1), 1–16. 2006.

[15] Wei, Z. H., *Cue reweighting in Shanghainese sandhi patterns*. ExLing 2022: Proceedings 13th International Conference of Experimental Linguistics. 2022.

[16] Wright, J. D., Laryngeal contrast in Seoul Korean (Doctoral dissertation). Philadelphia: University of Pennsylvania. 2007.

[17] Yip, M., Tone in east Asian languages[J]. The handbook of phonological theory, 1995, 476: 494.

TOP-DOWN EFFECTS OF LEXICAL PITCH ACCENT ON PHONETIC CATEGORIZATION IN JAPANESE

Terumichi Ariga, Risa Matsubara

Graduate School of Arts and Sciences, The University of Tokyo
ariga@phiz.c.u-tokyo.ac.jp, matsubara-risa761@g.ecc.u-tokyo.ac.jp

ABSTRACT

Although recent studies have shown that spoken word recognition is achieved with reference to not only segmental but also lexical prosodic information, the role of lexical prosody in speech perception is not well established. The present study investigated whether Japanese pitch accent has effects on phonetic categorization of voiceless–voiced continua (e.g., *tansu* LHH – *dansu* LHH, in which only *tansu* LHH "wardrobe" is a real word, and *tansu* HLL – *dansu* HLL, in which only *dansu* HLL "dance" is a real word). The results demonstrated that listeners were inclined to identify the voicing of the initial consonant so that the entire input resulted in a real word with consideration of pitch accent information. This suggests that Japanese pitch accent has a top-down effect on phonetic categorization, providing further evidence that lexical prosody plays a role in spoken word recognition in Japanese.

Keywords: speech perception, lexical prosody, phonetic categorization, top-down effect, Japanese

1. INTRODUCTION

Psycholinguistic studies have proposed various models of spoken word recognition by human listeners [8, 16, 18, 21, 30, 33]. Spoken word recognition is achieved under various levels of linguistic information. For example, the TRACE model of spoken word recognition [18] hypothesizes that there are three levels of processing: acoustic-feature, phoneme, and lexical levels. Each level of processing interacts with each other and gives both positive and negative feedback on other levels of processing.

This allows the redundancy of information in speech perception, making the mechanism of spoken word recognition robust against adverse situations in speech comprehension. For example, even when a phoneme is missed in auditory sentence comprehension, listeners can restore the missing input of a phoneme by referring to the context [31, 32]. When an acoustically equivocal phoneme is presented within a lexical environment, listeners interpret the phoneme so that the entire word becomes a real word [12, 19]. These effects are offered by top-down effects that a higher-level processing has on a lower-level processing. The likelihood of existing real words computed in the lexical-level processing has a bias on the processing of perception of segments.

Recently, ample studies have offered evidence that spoken word recognition is achieved with reference to not only segmental but also suprasegmental information, such as lexical prosody [9]. Words spoken with incorrect lexical prosody impose difficulty in recognizing the words appropriately [7, 20, 27]. When listeners listen to a spoken word, they activate words whose lexical prosody is identical to the perceived prosody but not words whose lexical prosody contradicts the perceived prosody [2, 11, 23, 28]. In languages in which prosody is lexically specified, lexical prosody is a potential cue for identifying spoken words.

The role of lexical prosody is not restricted to constraining lexical access as a bottom-up processing. In English, lexical prosody has a top-down influence on lower-level segmental processing. Connine and others [6] investigated perceptual phonetic categorization of consonants [15] in voice onset time (VOT) continua of a stress-contrastive pair *diGRESS – tiGRESS* and *DIgress – TIgress* (*diGRESS* and *TIgress* are real words; capital letters indicate stressed syllables). They reported that listeners tended to identify the voicing of initial consonants as voiceless (/t/) if the stress pattern of continua was strong-weak (i.e., *DIgress – TIgress*) than weak-strong (i.e., *diGRESS – tiGRESS*). This was because the strong-weak continua would become a real word if the initial consonant was /t/ (*TIgress*) whereas the weak-strong continua would become a real word if the initial consonant was /d/ (*diGRESS*), suggesting that the acoustically equivocal consonant was perceptually categorized so that the entire input becomes a real word rather than a nonword.

This replicated the finding that the categorization of phonemes is subject to be influenced by lexical status [12, 19]. Nevertheless, since the lexical status was determined by the stress pattern of the items, these results could be accounted for by the top-down effect of lexical prosody on phoneme-level processing.

However, it is still unclear whether these effects of

lexical prosody obtained in the English study can be generalized to other languages. In the present study, we investigated whether Japanese lexical prosody has top-down effects on the categorization of an equivocal input of phonemes.

Unlike stress languages such as English, Japanese lexical prosody is pitch accent, realized by high-or-low tonal patterns specified for each lexical item [13, 17, 29]. In Japanese, Cutler and Otake [10] reported that listeners can infer the entire input of a word based on pitch accent information in a gating paradigm. This suggests that lexical prosodic information is beneficial for compensating for the upcoming input of segments. However, it remains unknown whether Japanese pitch accent influences the perceptual categorization of an ambiguous input of segments.

In addition, Japanese is suitable for assessing the top-down effects of lexical prosody on the segment-level processing for the following reasons. First, compared to English, Japanese has relatively many lexical items with prosodic contrasts [25]. Thus, it is easy to find experimental materials. Contrary to the previous study [6] which used only a single item (a *TIgress–diGRESS* pair), we used multiple items in Japanese in the present experiment.

Second, Japanese pitch accent is unique in that the produced form of lexical prosody is not constant. In Japanese, the surface pitch pattern of a lexical item is variable due to phonological rules such as compound accent rule and deaccentuation [13]. There is also a dialectical variety of pitch accent [14]. Despite this variability of pitch accent, listeners of Japanese can recognize words presented with different lexical prosody from their lexical knowledge with minimal difficulty [3, 24]. Considering this, there is a possibility that pitch accent in Japanese may not be a reliable cue for categorizing ambiguous inputs.

Therefore, the present study addressed whether Japanese lexical pitch accent plays a role in phonetic categorization. We conducted an identification task of VOT continua, following Connine and others' experiment [6].

2. EXPERIMENT

2.1. Materials

Six pairs of experimental materials were prepared. All pairs were tri-moraic real words whose onset was a stop, segmentally contrastive by the voicing of the initial consonant, and prosodically contrastive by lexical pitch accent (either unaccented LHH or initially accented HLL). The full pairs are shown in Table 1. Pitch accent types of voiceless materials were counterbalanced. By referring to the Japanese lexical database [1], all materials were confirmed that they were not pronounced by any other pitch patterns in standard Tokyo Japanese, and highly familiar words among Japanese lexical items.

voiceless biased	voiced biased
koten LHH "classic"	*goten* HLL "palace"
kazai HLL "valuables"	*gazai* LHH "art supplies"
penki LHH "paint"	*benki* HLL "toilet"
panchi HLL "punch"	*banchi* LHH "address"
tansu LHH "wardrobe"	*dansu* HLL "dance"
teashi HLL "limb"	*deashi* LHH "start"

Table 1: Full sets of the experimental materials.

In these items, the voicing of the initial consonant was dependent on the type of pitch accent. Therefore, lexical status should be determined only by specifying the lexical prosodic pattern. For example, unaccented and voiceless-initial *tansu* LHH "wardrobe" was a real word, whereas *dansu* LHH was a nonword. Similarly, accented and voiced-initial *dansu* HLL "dance" was a real word, whereas *tansu* HLL was a nonword. Thus, if there are any effects of lexical prosody in categorizing the initial consonants, listeners should be inclined to report the voicing of the initial consonants so that the entire word becomes a real word. Items which would be recognized as real words if the initial consonant was voiceless were categorized as a voiceless-biased condition, while items which would be recognized as real words if the initial consonant was voiced were as a voiced-biased condition. Note that this bias was offered by the pitch accent of the items.

Auditory materials were made by the following procedure. Firstly, we recorded natural tokens of both voiced and voiceless items in both unaccented and accented pitch patterns (i.e., we made four tokens for each item: *tansu* LHH, *dansu* LHH, *tansu* HLL, and *dansu* HLL). The tokens were produced by a female native Tokyo Japanese speaker in a sound attenuated room, digitized at a sampling rate of 44.1 kHz with 16-bit resolution using a Samson Q2U microphone. The intensity of all the sound files were normalized using Praat 6.2.12 [5]. Then, using a Praat script [34], we synthesized six-level continua of the words in which the voicing of the initial consonant gradually ranged from a voiced consonant to a voiceless consonant. Following the previous study [6], these continua were made by manipulating VOT by 12 ms increments from 10 to 70 ms (i.e., 10, 22, 34, 46, 58, and 70 ms). We must note that this range of the VOT steps was based on English voicing contrast and the contrast of voicing in Japanese takes slightly different range [26]. However, the continua of a shorter VOT should be categorized as words with voiced consonants, whereas the continua of a longer VOT should be as words with voiceless consonants.

2.2. Procedure

The experiment was conducted online via PCIbex [35]. Participants were instructed to listen to the auditory materials through a headphone at a comfortable volume.

The task was an identification task of the voicing of the initial consonant in a two-alternative forced choice (2AFC) paradigm. At the beginning of each trial, a fixation point (+) was displayed at the center of the display. After 1000 ms, participants listened to either one of the six-level continua. Participants were asked to judge whether the initial consonant of the auditory material was voiced or voiceless. They answered by choosing either one of the buttons with labels of the voiced or voiceless initial mora in *hiragana* (Japanese syllabary characters) presented on the display. While there was no timeout for the response, participants were forced to select one of the choices. After the response, the next trial started automatically with a 1000 ms interval.

We emphasized the following notes in the instruction. First, the aim of the experiment was to investigate the mechanism of human speech perception and not to test the participants' listening ability: participants should answer intuitively without too much serious consideration. Second, therefore, there were no absolute correct answers, and the ratio of voiced and voiceless responses need not to be 1:1.

The task consisted of 72 main trials (2 bias conditions × 6 VOT lengths × 6 items), after twelve practice trials (where a *koban* HLL – *goban* LHH item with six-level VOTs was used). The items were presented in a random order. Main trials were divided into two sessions and participants could take breaks between the sessions.

After the identification task, we conducted a pronunciation task to confirm participants' knowledge on the pitch accent of the words used in the experiment. This confirmation was necessary because lexical knowledge on prosody could slightly vary from listener to listener even if they were standard Tokyo Japanese listeners.

Twelve words used as the auditory materials were presented on the display one by one. Participants pronounced the words naturally. The recording files were submitted to the experimenter after the experiment and the experimenter checked the listeners' pronunciation. The experiment lasted for 10–20 minutes in total.

2.3. Participants

Twenty-five native speakers of Tokyo Japanese participated in the experiment. Their mean age was 24.40 (SD = 6.73). They received monetary compensation for their participation.

2.4. Results

Data of the trials were excluded if the trials used auditory materials which each participant failed to pronounce with a standard Tokyo Japanese pitch pattern in the pronunciation task. This totaled 13.3 % of all the trials. Responses were coded as 1 if they answered that the initial consonant of the auditory material was voiceless and as 0 if they answered that the initial consonant was voiced.

The rates for voiceless response for each condition are shown in Figure 1. Changes in response rates as a function of VOT were regressed with logistic curves individually for each bias condition.

Response rates were analyzed by generalized linear mixed-effects models (GLMM), using lme4 package [4] in R 4.2.1 [22]. Participants' response (either 0 or 1) was set as the dependent variable. For the independent variables, the bias conditions, the VOT conditions, and their interaction were set as fixed effects whereas intercepts of both participants and items were set as random effects. The bias conditions were numerically coded as –0.5 or 0.5 for the voiced-bias condition and the voiceless-biased condition, respectively. The VOT conditions were also numerically coded as 10, 22, 34, 46, 58, or 70 for each six-level length (70 is the most likely to be categorized as voiceless).

The GLMM analysis suggested that there were significant main effects of bias conditions (β = 1.01, SE = 0.50, z = 2.00, p = .045) and VOT (β = 0.17, SE = 0.01, z = 18.03, p < .001), whereas their interaction was not significant (β = 0.02, SE = 0.02, z = 0.95, p = .342). As predicted, the rate for the voiceless response became higher as the VOT became longer. However, listeners tended to report more that the initial consonant was voiceless in the voiceless-

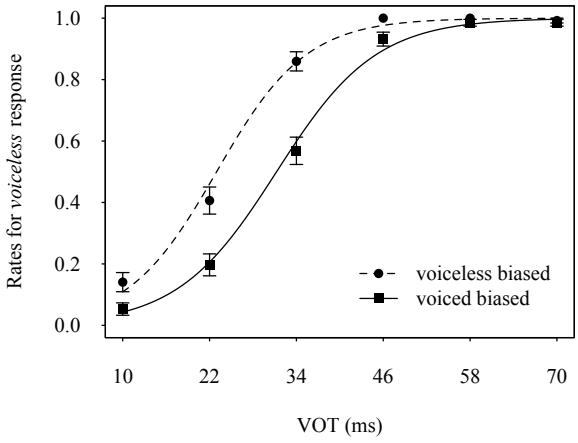


Figure 1: Rates for voiceless response for each bias condition and for each VOT. Error bars indicate the SE. The dashed line is a logistic curve for the voiceless-biased condition and the solid line is for the voiced-biased condition.

biased condition than in the voiced-biased condition.

Then, to investigate the difference in response rates by the bias conditions in detail, we made six post-hoc pairwise comparisons for each VOT condition with the Bonferroni correction (adjusted α = 0.008). When VOTs were 46, 58 and 70 ms, response rates did not differ significantly between the voiced-bias condition and the voiceless-biased condition (46 ms: p = .948; 58 ms: p = .907; 70 ms: p = .976). However, when VOTs were 10, 22, and 34 ms, listeners tended to identify the initial consonant as voiceless in the voiceless bias condition than in the voiced-bias condition (10 ms: p = .002; 22 ms: p < .001; 34 ms: p < .001). In these conditions, a bias occurred in phonetic categorization so that the entire input of auditory materials became a real word rather than a nonword.

3. CONTROL EXPERIMENT

The bias in phonetic categorization obtained in the experiment can be interpreted not only by the effects of lexical prosodic pattern but also by some unexpected VOT properties of the materials. To exclude this confounding interpretation, we conducted a supplemental control experiment.

We extracted the first mora of the original tri-moraic materials (e.g., *da–ta* continua with either H or L tone). These mono-moraic materials (72 trials in total) were presented in the same procedure as an identification task in the main experiment.

Nine native speakers of Tokyo Japanese who did not participate in the main experiment participated in the experiment. Their mean age was 26.40 (SD = 2.07).

The rates for voiceless response for each condition are shown in Figure 2. The GLMM analysis suggested that there was a significant main effect of VOT (β = 0.19, SE = 0.02, z = 11.34, p < .001), whereas there were neither significant main effect of the bias conditions (β = –0.36, SE = 0.84, z = –0.43, p = .194) nor their interaction (β = 0.01, SE = 0.03, z = 0.18, p = .857). The trend of categorical perception of VOT continua did not differ between the bias conditions when the auditory materials were the extractions of the initial mora. This suggested that the inclination of phonetic categorization in the main experiment occurred not due to the bias in VOT lengths of the materials themselves, but due to the bias that offered by the lexical prosodic pattern of the materials.

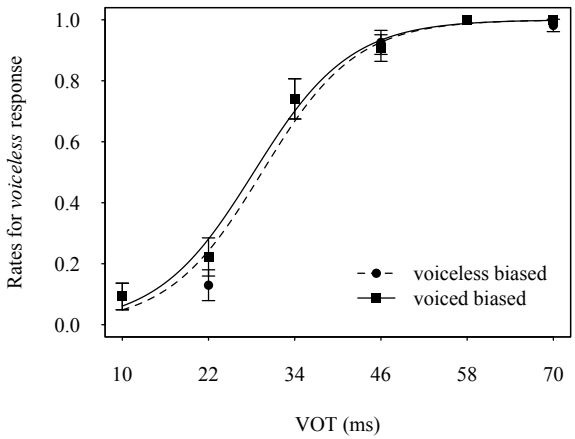


Figure 2: Rates for voiceless response for each bias condition and for each VOT (control experiment). Error bars indicate the SE.

4. DISCUSSION

The present study investigated whether lexical prosody plays a role in phonetic categorization in Japanese speech perception. Results demonstrated that listeners' categorization of the initial consonants in VOT continua was affected by the lexical-status bias offered by the pitch accent information. In the *tansu–dansu* pair, for example, when the pitch pattern was unaccented LHH, listeners tended to perceive /t/ more than /d/ since the continua became a real word if the initial consonant was /t/ (*tansu* LHH "wardrobe"). Similarly, when the pitch pattern was initially accented HLL, listeners tended to perceive /d/ more than /t/ since the continua became a real word if the initial consonant was /d/ (*dansu* HLL "dance").

This replicated the previous study in English [6], suggesting that the previous findings in English can be extended to the case of Japanese. Contrary to the hypothesis that Japanese pitch accent may not be a reliable cue due to the variability of lexical prosody [24], pitch accent may play a role in computing lexical status of perceived spoken words.

These behavioral results can be accounted for within the paradigm of the parallel processing models of spoken word recognition [18], assuming that there is a level for processing lexical prosody in the architecture. In Japanese, the processing of lexical prosody fundamentally plays a role in constraining lexical activation [2, 23]. In addition, this processing also affects the phonetic categorization that should be processed in the lower phoneme-level processing.

Therefore, pitch accent has both bottom-up and top-down effects in Japanese spoken word recognition. The mechanism of spoken word recognition should be robust against variety of acoustic signals by utilizing both segmental and lexical prosodic information. The findings in the present study provide further evidence that pitch accent plays a role in spoken word recognition in Japanese.

5. REFERENCES

[1] Amano, S., Kondo, T. 1999. *NTT Deetabeesu Shiriizu: Nihongo no Goitokusei* [*NTT Database Series: Lexical Properties of Japanese*]. Sanseidoo.

[2] Ariga, T. 2022. Pitch accent constrains lexical activation in Japanese spoken word recognition: A semantic priming study. *Language and Information Sciences* 20, 1–17.

[3] Ariga, T. 2022. Bummyakuka de teiji sareta ayamatta akusento no saikaishaku no katei [Process of reanalysis of mispronounced prosody in sentential context]. *IEICE Technical Report* TL2022-20, 11–16.

[4] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[5] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. https://www.praat.org/

[6] Connine, C. M., Clifton, C. E., Cutler, A. 1987. Effects of lexical stress on phonetic categorization. *Phonetica* 44(3), 133–146.

[7] Cutler, A., Clifton, C. E. 1984. The use of prosodic information in word recognition. In: Bouma, H., Bouwhuis, D. G. (eds), *Attention and Performance X: Control of Language Processes*. Lawrence Erlbaum Associates, 183–196.

[8] Cutler, A. 1995. Spoken word recognition and production. In: Miller, J. L., Eimas, P. D. (eds), *Speech, Language, and Communication*. Academic Press, 97–136.

[9] Cutler, A., Dahan, D., van Donselaar, W. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40(2), 141–201.

[10] Cutler, A., Otake, T. 1999. Pitch accent in spoken-word recognition in Japanese. *The Journal of the Acoustical Society of America* 105(3), 1877–1888.

[11] van Donselaar, W., Koster, M., Cutler, A. 2005. Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology* 58A(2), 251–273.

[12] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1) 110–125.

[13] Kawahara, S. 2015. The phonology of Japanese accent. In: H. Kubozono (ed), *The Handbook of Japanese Phonetics and Phonology*. De Gruyter Mouton, 445–492.

[14] Kubozono, H. 2021. *Ippan Gengogaku kara Mita Nihongo no Purosodii* [*Japanese Prosody from General Linguistics Perspectives*]. Kuroshio Shuppan.

[15] Liberman, A. M., Harris, K. S., Hoffman, H. S., Griffith, N. C. 1957. The discrimination of speech sounds within and across phonemic boundaries. *Journal of Experimental Psychology* 54(5), 358–368.

[16] Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102.

[17] McCawley, J. D. 1968. *The Phonological Component of a Grammar of Japanese*. Mouton.

[18] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18(1), 1–86.

[19] McQueen, J. M. 1991. The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance* 17(2), 433–443.

[20] Minematsu, N., Hirose, K. 1995. Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese. *Journal of the Acoustical Society of Japan (E)* 16(5), 311–320.

[21] Norris, D. 1994. Shortlist: A connectionist model of Continuous Speech recognition. *Cognition* 52(3), 189–234.

[22] R Core Team 2022. R: A language and environment for statistical computing. https://www.R-project.org/

[23] Sekiguchi, T., Nakajima, Y. 1999. The use of lexical prosody for lexical access of the Japanese language. *Journal of Psycholinguistic Research* 28(4), 439–454.

[24] Shibata, T. 1961. Nihongo no akusento [Japanese accent]. *Gengo Seikatsu* 117, 14–20.

[25] Shibata, T., Shibata, R. 1990. Akusento wa dooongo wo donoteido bembetsu shiuru ka: Nihongo, eigo, chuugokugo no baai [How significant is word accent in differentiating homonyms in Japanese, English, and Chinese?]. *Keiryoo Kokugogaku* 17(7), 317–327.

[26] Shimizu, K. 2018. Heisa shiin no yuusei sei musei sei no onkyoo teki tokutyoo ni kansuru koosatsu [A Study on Phonetic Characteristics of Voicing Contrasts of Stop Consonants]. *Journal of the Phonetic Society of Japan* 22(2), 69–80.

[27] Slowiaczek, L. M. 1990. Effects of lexical stress in auditory word recognition. *Language and Speech* 33(1), 47–68.

[28] Soto-Faraco, S., Sebastián-Gallés, N., Cutler, A. 2001. Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language* 45, 412–432.

[29] Vance, T. J. 1987. *An Introduction to Japanese Phonology*. State University of New York Press.

[30] Vitevitch, M. S., Siew, C. S. Q., Castro, N. 2018. Spoken word recognition. In: Rueschemeyer, S., Gaskell, M. G. (eds), *The Oxford Handbook of Psycholinguistics* (2nd edition). Oxford University Press, 31–47.

[31] Warren, R. M. 1970. Perceptual restoration of missing speech sounds. *Science* 167, 392–393.

[32] Warren, R. M., Warren, R. P. 1970. Auditory illusions and confusions. *Scientific American* 223, 30–36.

[33] Weber, A., Scharenborg, O. 2012. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science 3*, 387–401.

[34] Winn, M. 2020. Praat script to manipulate VOT in natural speech. https://github.com/ListenLab/VOT

[35] Zehr, J., Schwarz, F. 2018. PennController for Internet Based Experiments (IBEX). https://www.pcibex.net/

THE ROLE OF ALLOPHONES IN PHONEME PERCEPTION MODELS: DO DEVOICED VOWELS TRIGGER VOWEL EPENTHESIS?

Takeshi Kishiyama[1], Chuyu Huang[2], Kei Furukawa[3], Yuki Hirose[1]

[1]Graduate School of Arts and Sciences, The University of Tokyo, Japan
[2]Faculty of Foreign Studies, Nagoya Gakuin University, Japan
[3]Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan
kishiyama.t@gmail.com, huang@ngu.ac.jp, furukawa.kei.fi4@is.naist.jp, hirose@boz.c.u-tokyo.ac.jp

ABSTRACT

This study investigated how phoneme perception models should incorporate allophones, leveraging dialectical differences among Japanese. It has been reported that listeners of a language that disallows consonant clusters insert epenthetic vowels, or illusory vowels, to repair illegal consonant clusters, thus, for example, perceiving VCCV as VCVCV. In addition to the roles of phonotactic constraints and acoustic cues, recent studies have indicated that allophones, such as devoiced vowels in the Tokyo dialect, also facilitate perceptual epenthesis. We compared the discrimination accuracies of VCCV and VCVCV perception of Tokyo-dialect speakers to those of Kansai-dialect speakers, who are reported to devoice vowels less frequently. Both Tokyo and Kansai speakers perceived illusory vowels to the same degree, indicating that illusory vowels were perceived even by speakers without devoicing. Furthermore, the results suggest that discriminative models other than probabilistic models assuming an auditory realization distribution can be psychologically valid.

Keywords: speech perception, phonotactics, context effects, perceptual epenthesis, dialect.

1. INTRODUCTION

When hearing speech sounds that violate the phonotactics of the native language, the listener perceives the phonemes according to the native language's rules. For example, native Japanese speakers perceive [ebzo] as /ebuzo/ according to their phonotactics, perceptually inserting the *illusory vowel* [1, 2]. Vowel devoicing has been reported to affect the illusion [3]. Our study re-examined these effects through experiments on speakers of different Japanese dialects with different degrees of devoicing of high vowels following certain consonants, and discussed how we should incorporate them into perceptual models.

First, phonotactics provide insights into the vowel illusion [4, 5, 6], as shown in an experiment with native French and Japanese speakers [1]. Unlike French, Japanese prohibits consonant clusters such as /bz/, and in this experiment the speakers of each language tried to discriminate between [ebuzo] and [ebzo]. While French speakers successfully distinguished the two sequences, Japanese speakers perceived [ebzo] as /ebuzo/, which decreased their discrimination accuracy.

Second, experiments with native speakers of Brazilian Portuguese and Japanese have revealed the influence of formant cues on illusory vowels [2]. Both languages do not allow /bz/, and the default epenthetic vowels were /i/ and /u/, respectively. The study deleted the vowels between /b/–/z/ in [ebizo] and [ebuzo], leaving the vowels' acoustic cues in /b/. The remnant of formant cues affected the participants of both languages, increasing the insertion rate of the non-default /i/ for Japanese.

One-step models can explain these results, where phonotactics and acoustic cues simultaneously affect the illusion [2]. These one-step models can be represented by hidden Markov models (HMMs), which can also represent perceptual assimilation models [7] and models in exemplar theory [8]. The statistical inference [9, 10] reproduced the illusory vowels [11] by calculating

$$\hat{c} = \arg\max_{c} P(c|c_{t-1})P(S_{\text{new}}|c). \quad (1)$$

Eq. (1) calculates two likelihoods: $P(c|c_{t-1})$ is the likelihood of the phoneme array c_{t-1} to c, and $P(S_{\text{new}}|c)$ is that of the input speech S_{new} for a phoneme c. Consequently, the phoneme c that maximizes the likelihoods is activated as $\hat{c}$. Traditionally, all elements c above were phonemes [10]. However, recent studies have suggested that allophones should be treated as discrete units in perceptual models.

First, devoiced/deleted vowels in Japanese are reported to increase the perceptual epenthesis [3].

The Tokyo dialect in Japanese deletes the high vowel /u/ in /esupo/, and it would be [espo] when the vowel follows fricatives or affricates [12, 13]. In other cases, the vowel in /epuso/ becomes devoiced as [epu̥so]. The study reported that the deletion/devoicing encouraged assimilation and increased the perceptual epenthesis. Second, transition probabilities between allophones are reported to cause the vowel illusion [14]. Japanese [ɕ], a realization of /s/ precedes /i/, but not other vowels. The study investigated whether [ɕ] induces the illusion of /i/ [14], and found that /i/ was perceived more frequently after [ɕ], than after [g]. This result suggested that allophones and their transition probabilities affect the vowel illusion.

The question is whether phoneme perception models should consider allophones as categorical, rather than as continuous. Without the assumption of categorical allophones, the one-step models can explain the results above. The difference in the auditory cues within the deleted and devoiced vowels yield different $P(S_{\text{new}}|c)$. Furthermore, the stimuli [ɕ] is supposed to have acoustic cues of /i/ [15], which can also change $P(S_{\text{new}}|c)$. To explain the results, therefore, these models do not need to have categorical allophones.

The multiple interpretations for the experiments are allowed by confounding the allophones and the acoustic cues. First, in [epu̥so] and [espo], not only the allophones (devoiced/deleted) but the preceding speech sounds ([p]/[s]) are also different. Second, when comparing [egpo] and [eɕpo], the transition probability from [g] or [ɕ] to the next vowel is confounded with the acoustic cues of [g] or [ɕ].

We leveraged the dialectal differences between the Kansai and Tokyo dialects in Japanese to deal with this issue. We compared the discrimination accuracies of VCCV and VCVCV perception of Tokyo-dialect speakers to those of Kansai-dialect speakers, who are reported to devoice/delete vowels less frequently [16, 17], while keeping the test items constant. According to the assumption that the deletion and devoicing caused the illusion, Kansai dialect speakers, who are less exposed to these allophones, should rarely yield illusions and more accurate responses in discriminative AXB tasks.

2. PERCEPTION EXPERIMENT

2.1. Materials and methods

We used an outsourcing service to recruit participants older than 18 years from the Tokyo or Kansai region. Those who had never lived in other areas since birth had priority during the recruitment process. We had 62 participants, including 33 Tokyo speakers (mean age 38, $SD = 10.9$) and 29 Kinki speakers (mean age 37, $SD = 11.9$), and excluded the data from four participants whose answers were under 75% accuracy in the categorization task (§3). The research was conducted in accordance with ethical requirements approved by the institution, obtaining informed consent from participants.

In the AXB task, they were asked to distinguish between VC_1VC_2V and VC_1C_2V, from a list of target stimuli, with 32 items based on a combination of the four conditions below. (1) For environments where deletion and devoicing occur, we established four pairs for C_1C_2: s–p, k–t, p–s, and ts–k. (2) We also had stimuli VC_1VC_2V with /u/ between C_1 and C_2 that were discriminated from VC_1C_2V. (3) We created two patterns, depending on whether a trial's stimulus in the X position was either VC_1VC_2V or VC_1C_2V. (4) We prepared voiced counterparts of C_1 and C_2, such as b–z, where neither deletion nor devoicing occurs. The positions of the correct answers were counterbalanced, A or B.

We recorded stimuli from three male speakers, and adjusted the loudness to 75 dB. The duration before C_2 was about 50 ms (about 400 ms for the all three moras) so that the "geminate" consonant, a distinctive phoneme *sokuon* in Japanese, would not be perceived. We removed the voicing parts between consonants in /egudo/ and /ebuzo/ based on formant and power, and extended the silent interval. In this case, the cycle of each preceding waveform was duplicated and extended. In contrast, the /u/ in /epuso/ was cut so that it would be of the same duration as /epuso/.

After three trials of practice tasks, 32 target and 24 filler items were randomized, and presented in an experimental environment on web browsers. A "+" was displayed in the center of the screen for 1000 ms before the audio stimuli, and the three audio items (AXB) were presented at intervals of 200 ms. After the presentation, they were to respond as quickly as possible. We didn't provide feedbacks on correctness to avoid the learning effect, and the program only provided answers in the practice task. The participants could start the next trial by pressing the space key.

2.2. Results

The y-axis in Figure 1 shows the average discrimination accuracy of VC_1VC_2V and VC_1C_2V in the AXB discrimination task. The left and right panes indicate whether the stimulus was in devoicing environment (e.g., k–t if C_1C_2 was devoicing condition, g–d if not). The x-axis shows

the Tokyo–Kansai ratio, the ratio of residential history between the Tokyo and Kansai region. For example, if a participant has lived in Tokyo for 25 years and in Kansai for 15 years, the ratio is $(25-15)/40$, or 0.25. We showed four items vertically, to see the tendencies for each item.

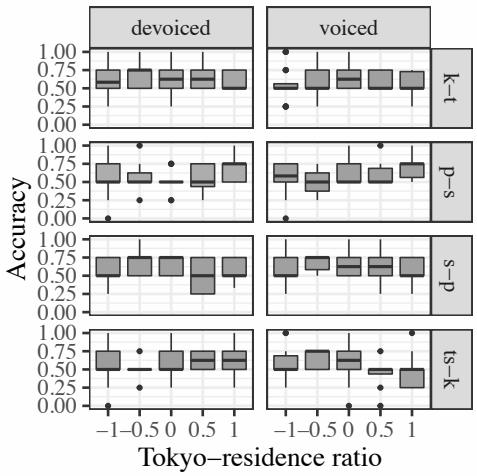


Figure 1: A boxplot that shows the relationship between residential history and response accuracy.

We checked the distribution of the reaction time data, and excluded 0.61% of data with a reaction time of 10 seconds or longer as outliers. We calculated the mean accuracy, averaging four patterns based on the X in AXB (VC_1VC_2V or VC_1C_2V) and the correct answer position (A or B), and excluded missing values.

We created a Bayesian linear mixed-effects model [18, 19, 20] with the Tokyo–Kansai ratio as the independent variable, and the AXB task discrimination accuracy as the dependent variable. Items and participants were modeled as random effects. The analysis revealed an overall mean accuracy of 0.59, without consistent effects for environmental differences. The estimates of dialectal differences were positive (0.019), and not in the direction of decreasing precision.

To evaluate the null hypothesis, we calculated a Bayes factor (BF) for the Tokyo–Kansai ratio. The Bayes factor, which is used to compare the models in the framework of Bayesian statistics, was 0.150 and provided substantial evidence for the null hypothesis. In this study, we calculated the BF as $Posterior\ Odds/Prior\ Odds$, where each $Odds$ is calculated as $P(b \notin [0,\infty]|Data)/P(b \in [0,\infty]|Data)$ and $P(b \notin [0,\infty])/P(b \in [0,\infty])$, respectively. Given that b, the estimates of the Tokyo–Kansai ratio, was supposed to be negative, the null region [21], considered to have no effect, was set to a positive value [0, ∞].

The BF of the Tokyo–Kansai ratio was 0.15, which indicated that the null hypothesis of a positive effect was $1/0.15(=6.67)$ times more plausible than the alternative hypothesis. This result contradicts the interpretation that the allophone increases the rate of illusions, and decreases its accuracy. All data and script are available at the corresponding author's repository https://github.com/kishiyamat/icphs2023 .

2.3. Discussion

If these allophones trigger the perceptual epenthesis, Tokyo dialect speakers should have lower accuracy in the discrimination experiment, than those without devoicing. However, the results of the AXB task showed that the discrimination accuracy of both Kansai and Tokyo dialect speakers, regardless of their language experience, was around 0.59, which is close to a chance level. In addition, the BF suggested that the language experience and devoicing did not affect the rate of illusions in the discrimination task.

Furthermore, we examined whether dialect differences in language experience affect the offline goodness rating of deletion/devoicing. Note that we performed this categorization task after the online process, to avoid any effect on the validation of the online process.

3. CATEGORIZATION TASK

3.1. Materials and methods

The participants in the perceptual task also completed the categorization task. All sounds presented in the perceptual task were categorized and rated. There were eight different patterns: four that were devoiced in the Tokyo dialect (esupo, ekuto, epuso, etsuko) and four that were not devoiced (ezubo, egudo, ebuzo, edzugo). Each of them had a devoiced counterpart, and the choices were shown in Japanese letters on the screen. Because three speakers recorded 16 patterns, the participants listened to 48 stimuli, selected one of the above eight options above that was closest to them, and rated its acceptability on a 7-point scale (0–6). In the rating, zero was considered "not appropriate at all," and six was considered "completely appropriate."

3.2. Results

The x-axis in Figure 2 shows the Tokyo–Kansai ratio, and the y-axis shows the goodness of fit rating.

The left and right panes indicate the environment of speech (devoicing or not), and the color difference shows whether the actual stimulus was devoiced or not. The overall ratings were above the mean of 4. The devoiced stimuli were rated higher in the devoiced environment (left pane), while they were rated lower in the no-devoicing environment (right pane). The residential history did not contribute to these differences.

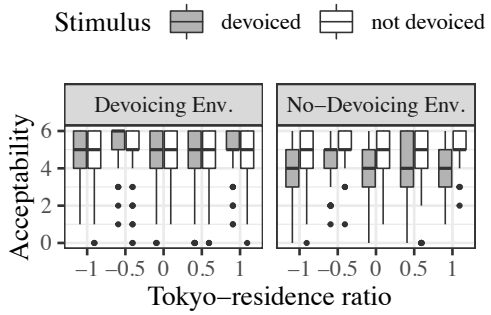


Figure 2: A boxplot that shows the relationship between the residential history and the goodness of fit rating.

To test these points, we created a Bayesian linear mixed-effects model, with the goodness of fit rating as the dependent variable. The independent variables were environment (devoicing/no-devoicing), stimulus type (devoiced/not devoiced), the Tokyo–Kansai ratio, and their interactions. Participants and items were random effects, and we employed a normal distribution ($M = 0$, $SD = 10$), as a prior distribution for each parameter.

The mean rating was 4.87, indicating that the evaluation was generally natural, given that the rating range was 0–6. The estimates of the devoicing environment and the devoiced stimulus were -0.37 and -0.79, respectively. Furthermore, the interaction, that is, the devoiced stimulus in the devoiced environment, increased the rating by 1.19. The *BF* for the Tokyo–Kansai ratio was 0.012 when the null region was set to $[-\infty, 0]$, indicating very strong evidence for the null hypothesis.

3.3. Discussion

The results showed that the overall responses were above 4, which means that the quality of the stimuli in this study was not lower than those in previous studies [3]. If the participants can judge the devoiced speech in the Tokyo dialect offline without epenthesis, the rate of devoiced stimuli in a no-devoicing environment would be less than three. Considering the results of the categorization task, both the Tokyo and Kansai dialect speakers showed the increased rating when the speech was devoiced in a devoicing environment, while it decreased when the devoiced stimuli were in a no-devoicing environment. However, the interaction's b was 1.19, and not high enough to cross the threshold between unnatural and natural.

4. GENERAL DISCUSSION

Previous research proposed that devoicing could cause the perceptual epenthesis. In addition, probabilistic models included the distribution of auditory realizations. This study proposes two different interpretations, based on the results of perceptual experiments and categorization tasks.

First, the results of the discrimination accuracy and categorization tasks in the previous studies might not be due to the effect of devoicing, but because of acoustic differences. If the perceptual epenthesis were due to the devoicing or deletion, the discrimination accuracy of the Tokyo dialect speakers would be lower than that of the Kansai dialect speakers, who do not devoice or delete the vowels. However, no such effect was observed. Instead, Kansai dialect speakers perceived the illusory vowels almost the same as the Tokyo dialect speakers, in our study.

Second, regarding how acoustic cues affect speech perception, previous studies [9, 10, 11] have assumed Eq. (1) during inference. On the right-hand side, $P(S_{\text{new}}|c)$ is thought to represent the distribution of auditory realizations for phonemes or the perceptual likelihood of S_{new} given c, and it does not need to be a generative model. This makes the results consistent with the previous study, because $P($ [ɯ̥] $|$ /u/ $)$ or $P($ [s] $|$ /su/ $)$ is not zero, given that the Kinki dialect speakers hear them in Tokyo dialect, which is prevalent in Japan. Given the Bayes' Theorem, the equation can be rewritten as $\hat{c} = \arg\max_c P(c|c_{t-1})P(c|S_{\text{new}})/P(c)$, and it integrates discriminative models instead of generative models. This allows us to employ other discriminative models, such as neural networks and multinomial logistic regressions. Given that previous studies have already supported the validity of HMMs [11, 22], we will test whether we can incorporate the above discriminative model into HMMs to explain behavioral data, in future studies.

5. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 20H01254 and 21K20027.

6. REFERENCES

[1] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler, "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, pp. 1568–1578, 1999.

[2] E. Dupoux, E. Parlato, S. Frota, Y. Hirose, and S. Peperkamp, "Where do illusory vowels come from?" *Journal of Memory and Language*, vol. 64, no. 3, pp. 199–210, 2011.

[3] A. Kilpatrick, S. Kawahara, R. Bundgaard-Nielsen, B. Baker, and J. Fletcher, "Japanese vowel devoicing modulates perceptual epenthesis," 2018.

[4] P. Hallé, J. Segui, A. Dominguez, and F. Cuetos, "Special is especial but stuto is not astuto: Perception of prothetic /e/ in speech and print by speakers of Spanish," *Psicolinguıstica en Espanol. Homenaje a Juan Segui*, pp. 31–47, 2014.

[5] P. J. Monahan, E. Takahashi, C. Nakao, and W. J. Idsardi, "Not all epenthetic contexts are equal: Differential effects in Japanese illusory vowel perception," *Japanese/Korean Linguistics*, vol. 17, pp. 391–405, 2009.

[6] W. Mattingley, E. Hume, and K. C. Hall, "The influence of preceding consonant on perceptual epenthesis in Japanese," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015, p. 888.

[7] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794, 2001.

[8] F. Lacerda, "The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 2. Stockholm University Stockholm, 1995, pp. 140–147.

[9] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference." *Psychological review*, vol. 116, no. 4, pp. 752–782, 2009.

[10] C. Wilson and L. Davidson, "Bayesian analysis of non-native cluster production," in *Proceedings of NELS*, vol. 40, 2013.

[11] T. Kishiyama, "The influence of parallel processing on illusory vowels," *Proc. Interspeech 2021*, pp. 1708–1712, 2021.

[12] M. Fujimoto and S. Kiritani, "Tookyoo hoogen-to Oosaka hoogen-niokeru boin-no museika-no hikaki," *Journal of the Phonetic Society of Japan*, vol. 7, no. 1, pp. 58–69, 2003.

[13] J. A. Shaw and S. Kawahara, "The lingual articulation of devoiced /u/ in Tokyo Japanese," *Journal of Phonetics*, vol. 66, pp. 100–119, 2018.

[14] A. Kilpatrick, S. Kawahara, R. Bundgaard-Nielsen, B. Baker, and J. Fletcher, "Japanese perceptual epenthesis is modulated by transitional probability," *Language and Speech*, pp. 1–21, 2020.

[15] H. Kubozono, "Japanese phonetics and phonology," *Iwanami Shoten*, 1999.

[16] H. Byun, "High vowel devoicing in the Keihanshin area of Japan," *Journal of the Phonetic Society of Japan*, vol. 15, no. 2, pp. 23–37, 2011.

[17] ——, "High vowel devoicing in Japanese : As an indicator of standardization of dialect," Ph.D. dissertation, The University of Tokyo, Nov 2012.

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[19] B. Goodrich, J. Gabry, I. Ali, and S. Brilleman, "rstanarm: Bayesian applied regression modeling via Stan." 2022, r package version 2.21.3. [Online]. Available: https://mc-stan.org/rstanarm/

[20] D. Lüdecke, M. S. Ben-Shachar, I. Patil, B. M. Wiernik, E. Bacher, and D. Makowski, "easystats: Framework for easy statistical modeling, visualization, and reporting," *CRAN*, 2022, r package. [Online]. Available: https://easystats.github.io/easystats/

[21] J. K. Kruschke, "What to believe: Bayesian methods for data analysis," *Trends in cognitive sciences*, vol. 14, no. 7, pp. 293–300, 2010.

[22] T. Kishiyama, , C. Huang, and Y. Hirose, "One-step models in pitch perception: Experimental evidence from Japanese," *Proc. Interspeech 2022*, pp. 1841–1845, 2022.

PHYSIOLOGICAL VS. SUBJECTIVE SLEEPINESS: WHAT CAN HUMAN HEARING ESTIMATE BETTER?

Vincent P. Martin[1,2], Aymeric Ferron[1], Jean-Luc Rouas[1],Takaaki Shochi[1,3], Lucile Dupuy[4], Pierre Philip[2]

[1]LaBRI, Univ. Bordeaux, CNRS, Bordeaux INP, UMR 5800, F-33400 Talence, France
[2]SANPSY, Univ. Bordeaux, CNRS UMR 6033, F-33076 Bordeaux, France
[3]CLLE, Université de Toulouse, CNRS UMR 5263, F-31058, Toulouse, France
[4]Bordeaux Population Health, Univ. Bordeaux, INSERM U1219, F-33076, Bordeaux, France
Correspondence should adressed to: vincent.martin@labri.fr and rouas@labri.fr

ABSTRACT

This article investigates the perception of vocal manifestations of excessive sleepiness. Although previous efforts have demonstrated that naive listeners are able to estimate behavioral sleepiness, we aim to assess in this study this ability on both subjective (medical questionnaires) and physiological (polysomnography) measurements of sleepiness. We asked 71 naive French-speaking listeners to annotate a subset of the Multiple Sleep Latency Test corpus, each listener participating in two annotation sessions, using one among three annotation tools, with two different annotation paradigms (with or without reference). Based on these data, we then evaluated their ability to correctly annotate subjective or physiological sleepiness depending on the annotation tool or the paradigm of the test they undertook. We also measured the interaction between their performance and their characteristics, as well as the correlation between the listeners' performances and speakers' characteristics.

Keywords: Sleepiness, Voice, Perceptual study, Experimental study, Paralinguistics

1. INTRODUCTION

Voice analysis is a promising tool for measuring sleepiness-related symptoms frequently and under normal patient living conditions [1]. Sleepiness is a peculiar symptom of interest in neuropsychiatric diseases that has a very high prevalence in the general population (up to 40% of the general population [2]) and induces negative consequences in both personal and public health, increasing the risk of disability and mortality [3].

Hence, automatic detection of sleepiness using speech samples has been the focus of two international challenges, proposed in parallel to the 2011 and 2019 Interspeech conferences. During the last challenge on the SLEEP corpus, the proposed systems did not lead to significant improvements in performance, the winner achieving a correlation coefficient between estimated and ground-truth values of sleepiness levels of $\rho = .387$ [4]. More recent propositions using the latest machine learning technologies have not reached much better performances [5, 6, 7].

To investigate the feasibility of the task, the team of Huckvale *et al.* conducted a perceptual study based on the SLEEP corpus [8]. Using 90 samples of the corpus, annotated by 26 listeners, they obtained a correlation of $r = .72$ between annotations processed using Wisdom of the Crowd and ground-truth values. Thus, they concluded that human hearing can estimate sleepiness through voice samples and that the bottleneck in the automatic estimation performances relies on the corpus content. However, this work suffers from three major limitations: (1) In a study using the exact same data, we failed to replicate these conclusions using the annotations made by 30 French-speaking listeners [9]; (2) In the SLEEP corpus, the samples are very short (between 3 and 5 seconds), whereas the minimum duration to estimate sleepiness with computational methods seems to be around 20 seconds [10]; (3) The ground-truth sleepiness label in the SLEEP corpus has never been validated in Sleep Medicine and is more a behavioral than a subjective sleepiness measurement [11], since it relies on the annotation by investigators of the behavioral manifestation of sleepiness, including voice (label contamination) [10].

Thus, to investigate the ability of human hearing to estimate sleepiness annotated with medicine-validated tools, we use in this study the Multiple Sleep Latency Test corpus (MSLTc) [10, 12], which

has been recorded during a MSLT, a gold standard polysomnographic test in sleep medicine [11], and contains physiological (polysomnography) and subjective (medical questionnaire) sleepiness labels. A specificity of MSLTc compared to other corpora is the recorded population: sleep clinic patients affected by various forms of hypersomnia, for which physiological and subjective sleepiness is uncorrelated [13].

This paper presents the results of the Endymion study, which takes advantage of this particularity to investigate which of subjective or physiological sleepiness is more easily detected in voice by naive listeners. The paper is organized as follows. We first introduce the methodology of our perceptual experiment in Section 2. Then, we present and discuss our results in Section 3. Finally, we draw conclusions in Section 4.

2. METHOD

2.1. Perceptual experiment

We recruited 71 native French-speaking annotators who had no experience in sleepiness assessment and no hearing problems. Since they can change the perception of speech, the sex [14], the age [15] and the musical sensibility [16] (any hobby or job related to music in this study) of the listeners were collected, as reported in Table 1.

N	Values	71
Age (years)	20-25	30
	25-30	26
	>30	15
Sex	F	22
	M	49
Musical Sensibility	Sensible	30
	Not. Sensible	41

Table 1: Listeners' characteristics.

The samples used in our experiment are extracted from the Multiple Sleep Latency Test corpus (MSLTc) [10, 12]. This corpus contains the recordings of 93 patients admitted to the sleep medicine department of the Bordeaux University Hospital for the diagnosis and/or treatment of rare hypersomnia diseases. They undertook a Multiple Sleep Latency Test (MSLT), consisting of asking them to take 20-minute naps every two hours, from 9a.m. to 5p.m. Before each nap, the patients were recorded reading texts of approximately 250 words extracted from *Le Petit Prince* (Saint-Exupéry). In order to reduce the variability of annotations due to the large number of speakers in the corpus, we selected 20 patients (10M/10F) from the MSLTc so that their variations of physiological (sleep latency) and subjective (Karolinska Sleepiness Scale and Cartoon Faces Scale) sleepiness across the naps were maximal. Furthermore, since the MSLTc recordings are long (mean duration: 77s), we kept only the first 30 seconds of each audio file to avoid fatiguing the listeners. The samples are then normalized in amplitude to -3dB. This constitutes the Endymion subcorpus.

The perceptual test consisted of asking the listeners to annotate one of the following three sleepiness measurements on these samples: (1) the duration after which patients fall asleep after the reading task, assessed by polysomnography. This value, named *sleep latency*, is a physiological measurement of short-term sleep propensity [11] and ranges from 0 to 20 minutes; (2) the score of the patients on the Karolinska Sleepiness Scale (KSS) [17], a nine-level scale that asks them to rate their level of sleepiness "during the last 10 minutes", filled just after each voice recording; (3) their score on the Cartoon Faces Scale (CFS) [18], also filled just after each voice recording, ranges from 0 to 4. The listeners annotated the KSS and CFS using the corresponding questionnaire, while sleep latency is annotated using a 100-level slider whose end anchors are 'Perfectly awake' on the left, and 'Extremely sleepy' on the right. The three annotation tools are represented in Figure 1.

During their contribution to the experiment, listeners were asked to participate in two annotation paradigms. On the one hand, a *Random* paradigm, during which they blindly annotated 10 samples drawn pseudorandomly among the Endymion subcorpus. On the other hand, a *Baseline* paradigm, for which the listeners had access to a reference sample of the patients when they are awake (recorded before an iteration of the MSLTc during which they stayed awake). The annotators then annotated the remaining four samples of the same patient, corresponding to the other four iterations of the MSLT. In this paradigm, each listener was asked to annotate the samples of two patients from the study subcorpus. In both paradigms, the samples were annotated one after the other (no browsing back). The order of the two paradigms and the used annotation tool was randomized.

2.2. Data Analysis

First, we scaled the annotations per annotator to remove their specific annotation behavior (z score). Then, to facilitate the comparison between the

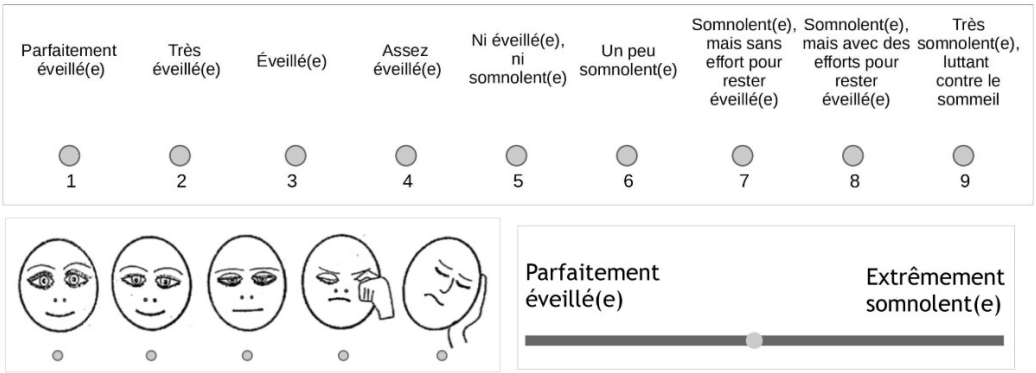


Figure 1: Annotation tool used in this perceptual experiment. (Top) Karolinska Sleepiness Scale. (Bottom left) Cartoon Faces Scale. (Bottom Right) Slider.

scales, we normalized the scores of each scale between 0 and 1. The same transformation is applied to the ground-truth values. To be compared with other scales, the annotated score on the slider and the corresponding ground-truth sleep latency are inverted, so that 0 corresponds to sleep latencies of 20 min (the patient did not fall asleep) and 1 corresponds to sleep latencies of 0 min (the patients felt asleep immediately). We thus obtain for each task a "sleepiness score" between 0 and 1 that can be compared between paradigms and annotation tools.

Since the number of points per annotator is too small to use correlation metrics [19], we measured annotation performance using the Mean Absolute Error (MAE), which is the arithmetic mean of the absolute error between annotation and ground truth across the samples. As a consequence, a lower MAE means fewer errors and thus better annotation performances.

3. RESULTS AND DISCUSSION

3.1. Paradigm and task

To identify the combination of paradigm and task favoring annotation performances, the MAE between the annotation of the listeners and the expected ground-truth values for each combination of task and paradigms are reported in Table 2.

The predominant effect affecting performances is the task: the best MAEs (.31-.28) are obtained annotating subjective sleepiness (KSS or CFS – with few differences between them) while the performances of annotating physiological sleepiness using the slider are distinctly worse (.47-.39). One hypothesis to explain this observation is that subjective sleepiness affects predominantly the acoustic quality of voice [20], while physiological sleepiness interfer with reading abilities [21, 22].

Paradigm	KSS	CFS	Sleep Latency
Random	**.30** (n=220)	.31 (n=250)	.40 (n=240)
Baseline	.29 (n=192)	**.28** (n=192)	.45 (n=184)
Both	**.29** (n=412)	.30 (n=442)	.46 (n=424)

Table 2: MAE between normalized annotation and ground truth for each paradigm and each annotation task.

Since the listeners only had access to short extracts (30s), we hypothesize that their annotation relies mainly on acoustic and prosodic information, hence detecting subjective sleepiness.

The second predominant effect is a small but notable difference between the *Random* and the *Baseline* paradigms, the latter leading to a better MAE (-1% of absolute MAE for the KSS, -3% for the CFS, -2% for the slider). We hypothesize that these differences are due to the reference audio file. Indeed, when picking a random audio sample, it is almost impossible to distinguish the sleepiness state of the speaker from all the other traits expressing through voice. By proposing such a reference, the listener may estimate accurately the sleepiness state of the patients independently from their traits, that are also expressing in the reference audio file in which they are not sleepy. Moreover, annotating the same speaker four times in a row in four different states could reinforce this distinction between the expression of the speaker's state and traits through voice.

3.2. Influence of listeners characteristics

In order to identify features of the listeners that could improve or interfere with their annotation capabilities, we tested (Mann-Whitney's U) on all tasks and paradigms whether the MAE per

listener before normalization was different between the following different categories: "first annotation session" vs. "second annotation session"; "M" vs. "F"; "Musical Sensitivity" vs. "No musical sensitivity". Moreover, regarding differences through age, we computed an univariate ANOVA across the three categories.

None of the factors taken into account led to significant differences in annotation performance in terms of MAE ($p > .05$): if some listeners' characteristics influence the way they annotate sleepiness from voice samples, they have not been captured by our study and would eventually require further investigation.

3.3. Influence of patients' characteristics

Factor	Paradigm	Task	MAE	
			ρ	p
Fatigue	Random	CFS	.51	.02
		All	.48	.04
Anxiety	Random	Slider	.56	.01
		All	.49	.03
Education level	Random	Slider	-.47	.04
	Baseline		-.57	.009
	All		-.57	.009

Table 3: Significant correlations between MAE per patient and patients' characteristics.

To measure the characteristics of the speakers that could influence the annotation of their level of sleepiness, we calculated the correlation (Spearman ρ) between MAE per speaker (before normalization) and the following factors: age, BMI, neck circumference, educational level (highest level of study after the French Certificate of general education), fatigue [23], Alertness [24], and Anxiety and Depression [25]. We also tested sex differences (Mann-Whitney's U). The significant correlations ($p < .05$) are reported in Table 3.

The speakers' fatigue level interferes with the annotation performance on the *Random* paradigm: when speakers report high levels of fatigue, listeners annotating their samples make more errors (i.e. MAEs are higher). This difference was not observed in the baseline paradigm. Similarly, the speaker's anxiety level interferes with the performance of listeners on the *Random* paradigm. Finally, the speakers' education level anticorrelates with the MAE using the slider: when the education level of the speaker increases, the errors made by the listeners when estimating their sleep latency using the slider decrease.

Fatigue and anxiety interfere with the correct annotation of sleepiness based on voice samples on the *Random* paradigm but not on the *Baseline* paradigm: we assume that the *Baseline* paradigm effectively allows listeners to annotate the speaker's sleepiness states independently of their trait features – such as fatigue or anxiety. An exception is the educational level of the speakers, which interferes with the annotation of physiological sleepiness on all the paradigms (*Random*, *Baseline*, and All). We hypothesize that in that task, physiological sleepiness has the same impact over voice as a lower educational level: the sleepy patients make more reading errors [21] and have different reading behaviors regarding the location and duration of the reading pauses [22].

4. CONCLUSION

This perceptual experiment is the first to investigate the ability of human hearing to differentiate between subjective and physiological sleepiness. The collected annotations show a better estimation of subjective sleepiness than physiological sleepiness by naive listeners. This observation leads us to hypothesize that the acoustic correlates of a subjective level of sleepiness are more numerous than those of its physiological counterpart, which is corroborated by the interference of education level with the estimation of physiological sleepiness only. Assessing physiological sleepiness using voice and speech may thus require additional information, i.e. features related to cognitive planning such as reading proficiency, length of pauses or sustained vowels. Moreover, the better performances obtained using the *Baseline* paradigm than using the *Random* one suggest that human earing estimate better variations of sleepiness states than its absolute level, opening new research paradigms in the automatic estimation of sleepiness.

Finally, no measured listeners' characteristics interacted with performances, while errors were correlated with speakers' fatigue, anxiety, and educational level: these results should incite to consider these parameters when automatically estimating sleepiness using voice features. However, such results may also be linked to the reading task on which the patients have been recorded: they still have to be confirmed using spontaneous speech recordings.

5. ACKNOWLEDGEMENTS

We are deeply indebted to all the participants who spent time participating in the Endymion study.

6. REFERENCES

[1] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," *Digital Biomarkers*, pp. 78–88, Apr. 2021.

[2] T. B. Young, "Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[3] A. J. Scott, T. L. Webb, M. Martyn-St James, G. Rowse, and S. Weich, "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Medicine Reviews*, vol. 60, p. 101556, 2021.

[4] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*, 2019, pp. 2413–2417.

[5] S. Amiriparian, P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, and B. W. Schuller, "A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech," arXiv 2005.08722, 2020, _eprint: 2005.08722.

[6] J. V. Egas-Lopez and G. Gosztolya, "Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness," in *ICASSP 2021*, Toronto, ON, Canada, 2021, pp. 7288–7292.

[7] E. L. Campbell, L. Docio-Fernandez, C. Garcia-mateo, A. Wittenborn, J. Krajewski, and N. Cummins, "Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism." in *Workshop on Speech, Music and Mind*, 2022.

[8] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of Sleepiness Ratings from Voice by Man and Machine," in *Interspeech 2020*, 2020.

[9] Anonymous, ""Prediction of Sleepiness Ratings from Voice by Man and Machine": the Endymion replication study," in *sumbitted to ICPhS 2023*, 2023.

[10] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, P. Philip, and J. Krajewski, "How to Design a Relevant Corpus for Sleepiness Detection Through Voice?" *Frontiers in Digital Health*, vol. 3, p. 686068, Sep. 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2021.686068/full

[11] V. P. Martin, R. Lopez, Y. Dauvilliers, J.-L. Rouas, P. Philip, and J.-A. Micoulaud-Franchi, "Sleepiness in adults: An umbrella review of a complex construct," *Sleep Medicine Reviews*, p. 101718, Nov. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1087079222001319

[12] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, and P. Philip, "The Objective and Subjective Sleepiness Voice Corpora," in *LREC 2020*, Marseille, France, 2020, p. 6525-6533. [Online]. Available: https://aclanthology.org/2020.lrec-1.803

[13] R. B. Sangal, "Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy," *Clinical Neurophysiology*, vol. 110, no. 12, pp. 2131–2135, 1999.

[14] M. Sato, "The neurobiology of sex differences during language processing in healthy adults: A systematic review and a meta-analysis," *Neuropsychologia*, vol. 140, p. 107404, Mar. 2020.

[15] H. Goy, M. Kathleen Pichora-Fuller, and P. van Lieshout, "Effects of age on speech and voice quality ratings," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1648–1659, Apr. 2016.

[16] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, 2013.

[17] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual." *Int J Neurosci*, vol. 52, pp. 29–37, 1990.

[18] C. C. Maldonado, A. J. Bentley, and D. Mitchell, "A Pictorial Sleepiness Scale Based on Cartoon Faces," *Sleep*, vol. 27, no. 3, pp. 541–548, 2004.

[19] M. Rhemtulla, P. Brosseau-Liard, and V. Savalei, "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions." *Psychological Methods*, vol. 17, no. 3, pp. 354–373, Sep. 2012.

[20] V. P. Martin, J.-L. Rouas, P. Thivel, and J. Krajewski, "Sleepiness detection on read speech using simple features," in *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, 2019.

[21] V. P. Martin, G. Chapouthier, M. Rieant, J.-L. Rouas, and P. Philip, "Using reading mistakes as features for sleepiness detection in speech," in *Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 985–989.

[22] V. P. Martin, B. Arnaud, J.-L. Rouas, and P. Philip, "Does sleepiness influence reading pauses in hypersomniac patients?" in *Speech Prosody 2022*. ISCA, 2022, pp. 62–66.

[23] L. B. Krupp, N. G. LaRocca, J. Muir-Nash, and A. D. Steinberg, "The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus," *Archives of Neurology*, vol. 46, no. 10, pp. 1121–1123, 1989.

[24] A. Shahid, S. Chung, L. Maresky, A. Danish, A. Bingeliene, J. Shen, and C. Shapiro, "The Toronto Hospital Alertness Test scale: relationship to daytime sleepiness, fatigue, and symptoms of depression and anxiety," *Nature and Science of Sleep*, p. 41, 2016.

[25] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatrica Scandinavica*, vol. 67, no. 6, pp. 361–370, 1983.

"PREDICTION OF SLEEPINESS RATINGS FROM VOICE BY MAN AND MACHINE": THE ENDYMION REPLICATION PERCEPTUAL STUDY

Vincent P. Martin[1,2], Jean-Luc Rouas[1], Aymeric Ferron[1], Pierre Philip[2]

[1]LaBRI, Univ. Bordeaux, CNRS, Bordeaux INP, UMR 5800, F-33400 Talence, France
[2] SANPSY, Univ. Bordeaux, CNRS UMR 6033, F-33076 Bordeaux, France
Correspondence should be addressed to: vincent.martin@labri.fr and rouas@labri.fr

ABSTRACT

Following the release of the SLEEP corpus during the Interspeech 2019 paralinguistic continuous sleepiness estimation challenge, a paper presented at Interspeech 2020 by Huckvale et al. examined the reasons for the poor performance of the models proposed for this task. They conducted a perceptual experiment on a subset of this corpus that seems to indicate that human hearing is, however, able to estimate sleepiness in this corpus.

In this study, we present the results of the Endymion replication study, in which the same samples were rated by thirty French-speaking naive listeners. We then discuss the causes of the differences between the two studies and examine the effect of listener and sample characteristics on annotation performances.

Keywords: Sleepiness, Voice, Perceptual study, Experimental study replication, Paralinguistics

1. INTRODUCTION

Sleepiness is a public health problem that increases the risk of disability and mortality [1]. The significant imbalance between the amount of sleep specialists and the prevalence of sleepiness (up to 40% of the general population [2]) and the need for physicians to better follow up their patients between consultations has led them to adopt Ecological Momentary Assessment with which they have access to patients' symptoms very regularly in their usual living conditions, paving the way to personalized treatments and real-time relapse prevention [3]. A promising tool to do so is speech processing. Indeed, voice is associated with the physiological state of the speaker [4] and it is possible to implement voice measurements in passive situations without requiring the patient to perform a specific task (e.g., interacting with a connected device).

In this way, sleepiness detection in speech has been at the heart of two international challenges proposed in parallel with the 2011 and 2019 Interspeech conferences. During the 2019 challenge, the SLEEP corpus was introduced [5] with the task of estimating sleepiness (correlation between predictions and ground truth values). It contains 16,492 random samples from 915 German-speaking subjects, whose sleepiness levels are annotated with the truncated average of three Karolinska Sleepiness Scale [6, KSS]: one is filled in by the subjects themselves, while the other two are annotated by assistants using video and audio [7]. Contrary to the expectations of the challenge organizers, the proposed systems did not show much improvement from the baseline (ρ = .387 for the best system [8] versus .343 for the baseline). Even more recent work on this corpus using the latest deep learning techniques did not perform better [9, 10].

To investigate the causes of this glass ceiling, Huckvale et al. [11] conducted a perceptual study to test the suitability of the corpus for the proposed regression task. Based on the annotations by 26 British English listeners of 90 samples extracted from the SLEEP corpus, and using *Wisdom of the Crowd*, they achieved performances far beyond those ever achieved by the systems proposed for sleepiness estimation tasks (r = .72). Thus, we claim that the study by Huckvale et al. supports that human hearing can estimate sleepiness from speech samples of the SLEEP corpus.

In this article, we propose to reproduce the perceptual study conducted in [11] (denoted as "original study") with naive French-speaking listeners to confirm or infirm this hypothesis. This paper is organized as follows. In Section 2, we introduce the methodology of our replication study on the SLEEP corpus. We present our results in Section 3 and discuss them in Section 4. Finally, we draw conclusions in Section 5.

2. METHOD

Thirty French-speaking listeners without hearing impairment were recruited by word-of-mouth.

Since we hypothesize that they can improve the perception of sleepiness in speech, we collected their understanding of German and their musical sensitivity [12]. All the characteristics of the listeners available in the original and replication studies are presented in Table 1. Using a KSS, they annotated the same 100 samples (10 for training) of the SLEEP corpus as in the original study. The order of the samples is the same in both studies, and the samples are shown and annotated one after the other (no browsing back).

The annotation tools used in each study are shown in Figure 1. While the version used in the original study combines at the same time a Lickert-like scale (gradual textual description) and a Visual Analog Scale (continuous line with two anchors), the scale used in our replication study is the standard Lickert scale as it has been presented to speakers.

Characteristic	Huckvale et al. $n = 26$	Endymion $n = 30$
Age	18-60	20-60
Sex	-	M: 17 F: 13
Impairments in hearing	None	None
Native language	English	French
German language level	German $\neq$ first language	"Not at all" ($n = 19$) "At least a little" ($n = 11$)
Specific Musical Sensibility	-	No ($n = 16$) Yes ($n = 14$)
Compensation	£5 ($n = 20$) attendance credits ($n = 6$)	None

Table 1: Listeners' characteristics.

3. RESULTS

The results of the same analysis as in the original study are reported in Table 2.

Z-normalized raw scores. First, the annotations are z-scaled per listener to eliminate their individual characteristics. The resulting distributions in both studies are shown in Figure 2 (left). The raw z-scaled annotations in the Endymion replication study resulted in a slightly better Person and Kendall correlations than in the original study, but these achievements are still insufficient to accept the hypothesis that human hearing is able to estimate sleepiness from speech samples extracted from the SLEEP corpus.

Wisdom of the crowd. In a second step, a Wisdom-of-the-Crowd (WoC) procedure is applied: for each sample, all z-scaled annotations are averaged, resulting in an average predicted score. The resulting distributions are shown in Figure 2 (right). Compared to the original study, applying WoC to listener annotations in the Endymion replication study brings a smaller gain in the correlation between estimated values and ground truth.

Metric	Huckvale et al.	Endymion
	Z-scaled annotations	
Correlation Kendall's coefficient	$r = .249$ $\tau = .117$	$r = .318$ $\tau = .23$
	WoC z-scaled annotations	
Correlation Friedman test UAR F1 SL/NSL	$r = .72$ 1|2,3,4,5,6,7|8,9 93.6% .87/.96	$r = .41$ 1|2,3,4|5,6,7|8,9 69.6% .51/.81
	Complementary results (no normalization)	
ICC2-10	.668	**.975**
Std/listener *mean (std)*	1.83 (.38)	2.34 (.21)

Table 2: Comparison metrics between the original study and our replication study. *WoC*: Wisdom of the Crowd, *SL*: Sleepy, *NSL*: Not Sleepy.

In order to determine underlying groups in the annotations, a Friedman test and the corresponding post-hoc analysis are calculated with the Python package `Pingouin` v.0.4.0 [13]. In the original study, the annotations are grouped into three sleepiness levels based on a Friedman test: a 'sleepy' group (KSS > 7), a 'normal' group ($2 \leq$ KSS ≤ 8) and an 'aroused' group (KSS $= 1$). The same analysis applied to our replication study suggests a division into four groups ($W = .263, ddof = 8, p = .007$, pairwise Mann-Whitney): a 'sleepy' and an 'aroused' groups (resp. KSS $>$ 7 and KSS $= 1$), and two 'slightly aroused' and 'slightly sleepy' subgroups (KSS $\in \{2,3,4\}$ and KSS $\in \{5,6,7\}$). Sleepiness subgroups for each study can be observed in Figure 2 (right).

To calculate binary classification performance, the ground-truth KSS is binarized into two classes: Sleepy (KSS > 7) and Not Sleepy (KSS ≤ 7). A threshold of .26 on the z-scaled WoC annotation gives a binary classification UAR (Unweighted Average Recall) of 93.6% in Huckvale et al., while in our replication study the best cutoff value of .31 yielded a UAR of only 69.6%, which is lower than the classification performances typically achieved on the task. To complement these results for further discussion, we also calculated the F1 value for each class: in both studies, the F1 values are better in the NSL class than in the SL class.

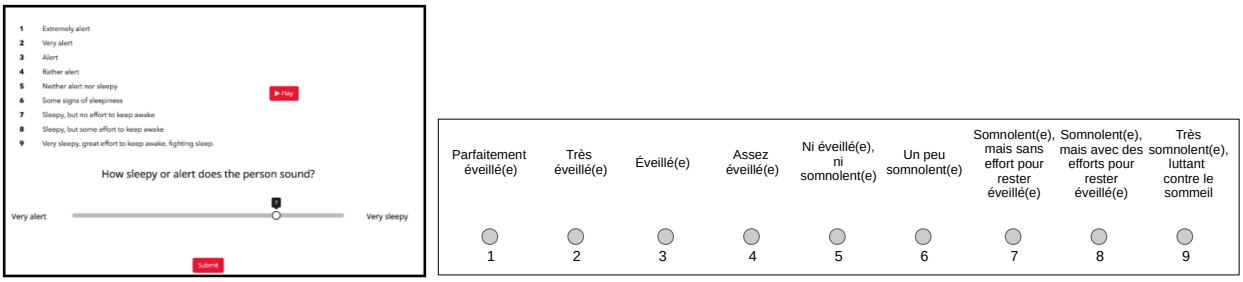


Figure 1: KSS annotation tool proposed in the original study (left), and our replication study (right).

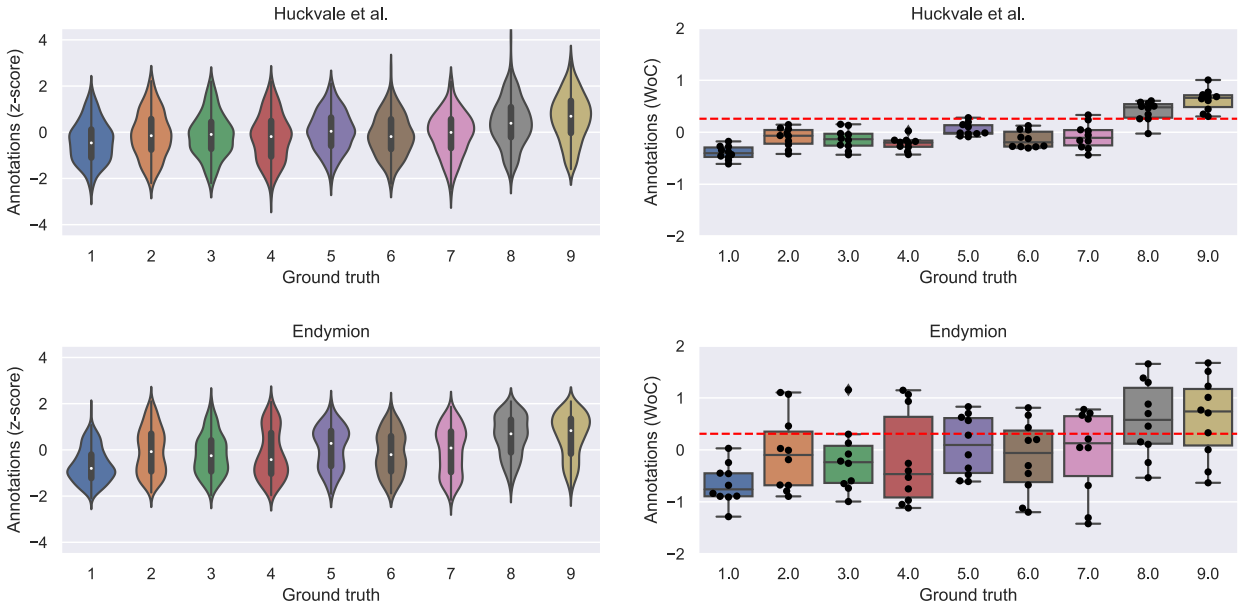


Figure 2: (Left) Violin plots of z-scaled annotations by ground-truth value, in the study of Huckvale et al. (top) and in the Endymion study (bottom). (Right) Box plot of the WoC z-scaled annotations depending on the ground-truth KSS. Each dot represents a sample, and the red dashed line represents the cut-off value giving the best UAR.

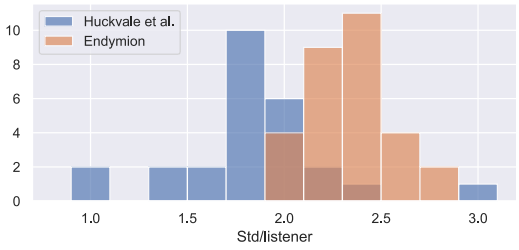


Figure 3: Distribution of standard deviation of annotations per listener (before z-normalization). The observed difference is statistically significant (t-test, $p < .0001$).

4. DISCUSSION

Differences between the studies To find the underlying cause of the differences between these analyzes, we calculate the intraclass correlation (ICC2-10) on the raw annotations before z-normalization for each study, which is an indicator of overall agreement between the annotators [14]. It shows that there is a lower agreement between annotators in the original study (ICC = .668) than in our replication study (ICC = .975). This could be the source of the small performance gain bought by WoC in the Endymion study: averaging already converging opinions over a sample yields much less information than averaging dissenting opinions. To account for the variety of levels used by listeners in each study, we also calculate the standard deviation per listener of their annotations before z-normalization (see Figure 3).

The listeners of the original study use significantly fewer different levels than those of the Endymion study. However, in the present study, listeners use a greater variety of levels, creating greater finesse in the annotations, albeit with less contrast (4 subgroups in the Endymion study vs. 3 subgroups in Huckvale et al.). These

different behaviors may find their source in the presentation of the annotation scales between the two studies: while in the present study the textual description is directly above the selected value, some listeners may have inadvertently used the annotation scale in Huckvale et al. as a simple visual analog scale, without referring to the text description at the top of the screen, creating their own rating scale [15].

What is the influence of the listeners' characteristics on their annotation performance? In the Endymion study, two listener characteristics of particular interest for the task were collected: (1) their musical sensitivity, since music practice or a hobby related to music might improve the perception of speech [12, 16]; (2) their understanding of the German language, as annotators who understand the language might have access to additional linguistic information. Therefore, for each annotator, we calculate the Mean Absolute Error (MAE) between the annotations (after z-scaling) and the corresponding ground truth values. Then, we min-max normalize them and compute Mann-Whitney tests with the aim of distinguishing the MAE between each group.

We find no significant difference for either of the two previous factors (resp. $p = .190$ and $p = .228$ for musical sensitivity and German language comprehension). Thus, if some listener characteristics influence the way they estimate sleepiness from speech samples, they are not captured by our study.

Which samples are the hardest to annotate? For a given sample, we hypothesize that two effects could explain the differences between annotations and ground truth. First, listener fatigue, who may not annotate the last samples as carefully as the first ones, for whom concentration may be easier (influence of the order of the samples). Second, the speaker's level of sleepiness, since human hearing might be able to detect some levels of sleepiness more easily than others (influence of the KSS). For each study and each sample, we computed the MAE between the ground truth values and the listeners' z-score annotation (MAE per sample). To make the ground-truth and z-scale annotation values comparable, we min-max normalize them so that their minimum value is 0 and their maximum value is 1. Then we calculated the correlation (Spearman's ρ) between the MAE per sample and their order and between the MAE per sample and the ground-truth KSS. The results are presented in Table 3.

The MAE does not correlate with the sample index in either study, which excludes the hypothesis of listener fatigue. However, the per-sample MAE correlates weakly with the KSS ground truth of the original study and more strongly with the Endymion study: the higher the KSS, the larger the errors between annotations and ground truth. We have put forward two hypotheses about this result. First, the human auditory system may be more sensitive to vocal expressions of alertness than to sleepiness, which explains the increase in MAE with KSS. Coming back to the classification results from Section 3, the F1 values are better in the NSL class than in the SL class, which supports this hypothesis. Second, we cannot exclude the hypothesis that some speakers could have completed a KSS indicating a high level of sleepiness at the time of their evaluation, but were then stimulated by the various recording tasks. Therefore, sleepy subjects may make (involuntary) efforts to compensate for their sleepiness in order to complete the task, creating a difference between their self-reported level of sleepiness, assessed before the task, and the expression of their sleepiness in their voice.

Factor	Huckvale et al.	Endymion
Order	$\rho = .09, p = .39$	$\rho = -.13, p = .24$
KSS	$\rho = .20, p = .05$	$\rho = .40, p < 10^{-3}$

Table 3: Correlation between the MAE per sample and their order and ground truth KSS.

5. CONCLUSION

To conclude, our replication study did not provide results as conclusive as the previous study conducted by Huckvale et al. [11]. Therefore, we cast doubt on the assumption that the human ear is capable of correctly assessing sleepiness from speech samples extracted from the SLEEP corpus. Regarding the factors influencing the annotation of the samples, we did not identify any influence of the annotators' characteristics. On the other hand, we found a link between the level of sleepiness of the speakers and the quality of sleepiness annotation in these two studies, with listeners having more difficulty in estimating very sleepy speakers.

6. ACKNOWLEDGEMENTS

We are deeply indebted to all the participants who spent time participating in the Endymion study. We gratefully thank Pr. Jarek Krajewski for having given us access to the SLEEP corpus, and Pr. Mark Huckvale for having given us the necessary material for the replication of his study.

7. REFERENCES

[1] A. J. Scott, T. L. Webb, M. Martyn-St James, G. Rowse, and S. Weich, "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Medicine Reviews*, vol. 60, p. 101556, 2021.

[2] T. B. Young, "Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[3] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological Momentary Assessment," *Annual Review of Clinical Psychology*, vol. 4, no. 1, pp. 1–32, 2008. [Online]. Available: https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[4] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," *Digital Biomarkers*, pp. 78–88, Apr. 2021.

[5] B. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychocz, R. Vollman, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*, 2019.

[6] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual." *Int J Neurosci*, vol. 52, pp. 29–37, 1990.

[7] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech - Between Performance and Interpretation," in *Speech Prosody 2014*, 2014, pp. 864–868.

[8] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*, 2019, pp. 2413–2417.

[9] J. V. Egas-López, R. Busa-Fekete, and G. Gosztolya, "On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection," in *Speech and Computer*, ser. Lecture Notes in Computer Science, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. Cham: Springer International Publishing, 2022, pp. 178–187.

[10] E. L. Campbell, L. Docio-Fernandez, C. Garcia-mateo, A. Wittenborn, J. Krajewski, and N. Cummins, "Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism." in *Workshop on Speech, Music and Mind*, 2022.

[11] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of Sleepiness Ratings from Voice by Man and Machine," in *Interspeech 2020*, 2020.

[12] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, 2013.

[13] R. Vallat, "Pingouin: statistics in Python," *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, 2018.

[14] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016.

[15] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, no. 1, pp. 1–15, Mar. 2000.

[16] W. F. Thompson, E. G. Schellenberg, and G. Husain, "Decoding speech prosody: Do music lessons help?" *Emotion*, vol. 4, no. 1, pp. 46–64, 2004.

THE EFFECTS OF DIFFERENT LEVELS OF AMPLITUDE VARIATION ON PERCEIVED SPEAKER DOMINANCE

Sara Pearsell[1], Daniel Pape[1]

[1] McMaster University, Hamilton, Canada
pearsels@mcmaster.ca , paped@mcmaster.ca

ABSTRACT

The purpose of this study is to examine how suprasegmental amplitude manipulation on different linguistic levels (sentence, word, syllable) would affect perceptual dominance ratings. Six (Canadian) English speakers produced sentences/paragraphs with neutral valence. Amplitudes were then manipulated on three levels: (i) *sentence/paragraph* (i.e. amplitude enhancement of entire phrases or paragraphs), (ii) *word* (one specific word in focus, realized with enhanced amplitude), and (iii) *syllable* (i.e. increasing/decreasing amplitude differences between all stressed versus unstressed syllables). 71 participants of an online perception experiment then rated the stimuli on 4 continuous sliding scales corresponding to concepts of social and physical dominance. Results showed that amplitude increases on *sentence* level significantly increased dominance ratings. In contrast, on the *syllable* level significantly decreased dominance ratings were observed for increased amplitude levels. *Word* level manipulations and the decreased syllable condition did not have an effect on ratings.

Keywords: amplitude variation, perceived personality traits, speech perception, variability in speech

1. INTRODUCTION

In more recent research ([1], [2], [3] [4]), there has been an increase in examining the acoustic properties that influence the perception of speakers' personality traits. Research has identified influences of pitch ([5], [6], [1], [2]), pause-filled-gaps ([7], [8]) and other parameters on the perceived personality and/or vocal traits of a speaker, such as dominance or charisma. Despite this interest, to our knowledge, there remains limited research examining the influence of varying levels of suprasegmental amplitude variation on personality trait perception, particularly within the realm of dominance. Here we are focused on amplitude differences, with amplitude and intensity being classified as acoustic parameters while loudness is classified as the corresponding perceptual parameter. We are interested how different linguistic domains such as *sentence/phrase*, *word* or *syllable* level would influence perception of dominance.

1.1. Previous Research

Previous research has found that differences in (1) amplitude ([9], [10]), (2) loudness ([11]), and (3) intensity ([12]) impact listeners' perceptions of dominance for a given speaker. Results from Tusing & Dillard [13] found increased overall mean vocal amplitude is positively associated with increases in dominance ratings (here defined as a more general concept of dominance) in the production of spontaneous speech. In that study, speech productions were short audio messages which were examined for amplitude differences over several phrases. Due to the short nature of audio length, the experimental design essentially examined amplitude variation perception at sentence level.

Other research has found that more confident individuals speak with greater amplitude ([14]). In this study, participants verbally answered a series of questions, and then scored how confident they were in their own answers. The recorded participant audio was converted to amplitude measurements. The amplitude measurements of these verbal responses were compared to the confidence scores participants gave themselves. Since these verbal answers were short, (one to two words), it could be argued that this study was looking at influences of word stress, thus examining the effect of word stress on *social* dominance (as opposed to overall dominance) perception through confidence scores.

Lastly, high signal intensity is associated with perceptions of dominance ([12], [10]). In the study by Aronovitch [12], participants listened to short audio recordings (around 10 seconds) and were asked to judge several personality traits. One of these scales looked at dominance (ranging from dominant to submissive). The presented recorded audio stimuli were several different speakers spontaneously describing a presented image. Results demonstrated that higher signal intensity increases perceptions of dominance. Like the study by Tusing & Dillard [13], intensity levels were examined for the whole phrase, and again had a more general definition of dominance (overall dominance).

Although these studies often do not differentiate between social or physical dominance and generally emphasize amplitude variations at sentence level, results clearly show there is a relationship between ratings of perceived dominance and acoustic amplitude variation.

For the current study, we opted for two different definitions of dominance. We modelled our definition as proposed by Mueller and Mazur [15, p. 570]: "a dominant person tells other people what to do, is respected, influential, and often a leader, while submissive or subordinate people are not influential or assertive and are usually directed by others,". This definition was later used in a study by Puts et al. [5] which provided the framework for the current study. More specifically, *physical* dominance is a measure of physical behaviours or traits, such as physical strength, which increase an individual's status within a given hierarchy, while *social* dominance measures the social component of behaviours or traits resulting in a speaker being viewed as influential and respected as a leader. Essentially, for *physical* dominance, speakers are judged whether they are likely to win in a physical fight. For *social* dominance, it is the likelihood that the speaker is a respected leader.

1.2. Aims & Hypotheses

The present study investigates how suprasegmental amplitude variation on different linguistic/phonetic levels affects listeners *social* and *physical* dominance perception of these same speakers, and additionally which of these dominance types are the most salient in high (positive) versus low (negative) listener ratings. Three different linguistic/phonetic levels were chosen to capture numerous possible amplitude variations seen in acoustic phonetics and linguistics in general: sentence-level (i.e., amplitude manipulation of the entire phrase/paragraph), word-level (i.e., one specific word in focus), and syllable-level (i.e., amplitude differences contrasting stressed compared to unstressed syllables).

We have the following hypotheses: First, we assume that increases in amplitude over the whole sentence will increase ratings of dominance for both *social* and *physical* dominance. Second, increasing amplitude differences at word level (focus) should increase ratings for *social* dominance (see [14]), but not necessarily for *physical* dominance. Lastly, for syllable stress, we expect to see an influence on ratings for *both* types of dominance (similar to our expectations for *sentence* level), however, due to the lack of previous research, it is unclear if listener ratings will be affected more positively or more negatively on the *syllable* level.

2. METHODS AND MATERIALS

2.1. Stimuli

As word material we chose 2 paragraphs (consisting of multiple simple sentences) and 4 isolated sentences. Both the sentences and paragraphs were constructed to have a neutral valence (i.e. the connotations of the stimuli had neither positive nor negative valence to prevent any influence of word material on listeners interpretations). Each paragraph was approximately 12 seconds long.

Six native Canadian English speakers (3 female, 3 male) produced the stimuli. 4 were professional voice actors and 2 were Linguistics graduate students (McMaster University). Due to COVID restrictions, the professional voice actors used their own high-quality microphones and adequate recording environments. During the recording sessions, the speakers were monitored and directed via Zoom by the authors of the current study. The 2 graduate students were recorded on a high-quality microphone (Rode NT1A) via Focusrite Scarlet audio interface inside the soundproof booth of the Phonetics Lab at McMaster University. For all recordings, microphone distance was around 10 cm (microphone being horizontally off-centre from the lips) by 45 degrees. Recordings for each stimulus was repeated 3 times, with the best production selected as the perceptual stimulus for the listeners. Prosodic differences were as tightly controlled as possible across speakers and conditions. After recording, the audio samples were screened ([16]) and carefully checked for achieved accuracy and consistency of produced vocal parameters, the absence of undesired voice quality changes (e.g., sentence-final glottalization) and differences in articulation (e.g., different speech rates and hyper/hypoarticulation). Additionally, a high pass filter (80 Hz for male speakers and 150 Hz for female speakers; 24dB/octave linear-phase) was applied to remove and attenuate any additional low frequency noise which may have been a part of the original recordings. Before amplitude manipulations, the loudness for all stimuli per condition ((i.e., applied to each sentence stimulus or each paragraph) was normalized to 65 dB using the Praat (Boersma & Weenink [17]) intensity normalization procedure.

2.2. Amplitude Manipulations

Each sentence/paragraph's amplitude was then manipulated on three levels: (1) sentence stress, (2) word stress (*focus*) and (3) syllable stress (i.e. stressed versus unstressed syllable amplitudes). Each of the three manipulation levels was performed on the baseline audio file (i.e. the speaker-produced audio

file for each speaker and each paragraph/sentence) in Adobe Audition ([18]), resulting in an *enhanced condition* (i.e., the *enhanced* condition had increased amplitude values at sentence, word or syllable level). Only for the syllable stress condition we also included a *reduced* condition.

Sentence stress (phrasal stress) was manipulated for the entire sentence or paragraph. The overall sentence intensity was increased by 6 dB. This means that by increasing the intensity on sentence/paragraph level and comparing these manipulations to the originally produced utterance (i.e., baseline) we can examine how the dominance perception of that speaker in the higher intensity condition compares to the perception of the same speaker producing the same sentence/paragraph, but with normal (i.e. -6 dB) intensity as baseline.

Word stress was manipulated by increasing the amplitude of one particular word of each sentence, so that this one word is now in stressed or *focus* position for that sentence, but only with regard to intensity differences (thus not pitch, duration, or vowel quality differences) compared to the rest of the sentence. The selected focus word which was increased in amplitude was selected before the manipulation and was identical for all speakers.

Lastly, *syllable stress* intensity, or more precisely the intensity difference between stressed and unstressed syllables, was manipulated in two ways: an *enhanced* condition (as in the previous two descriptions) and additionally a *reduced* condition, in order to investigate the effect of varying stressed/unstressed syllable intensity differences on listener dominance ratings. For the *enhanced* condition, the intensity of all syllables in stressed positions within a sentence/paragraph was increased by 6 dB, while all unstressed syllables remained unaltered. For the *reduced* condition, both the intensity of stressed syllables was reduced, *and* the intensity of unstressed syllables was increased. We aimed for this split approach (decrease of stressed syllable intensity plus increase of unstressed syllable intensity) to generate the most natural-sounding production for this condition. The *reduced* condition is meant to give the effect of a speaker with diminished distinctions of stressed/unstressed syllable intensity. Stimuli were not repeated, and the final acoustic stimuli count was 180.

2.3. Experimental Setup

The current study explored an academic setting with a focus on dominance ratings for professors by students. The statements used were modelled from the Puts et al. [5] study which asked listeners to judge if a speaker was more likely to win a physical fight (classified as *physical* dominance) and whether the listener was “extremely dominant” or “extremely submissive” (*social* dominance). In addition to these 2 statements, we created and added two additional scales (one social and one physical). Participants were presented acoustic stimuli through headphones containing either one paragraph or one sentence. Simultaneously, participants were presented with a screen with 2 statements (one for physical and one for social dominance), paired with a sliding scale for each statement. On a following screen, 2 different statements were presented (again, one social and one physical) along with the repetition of the exact same audio stimulus from the previous screen. Participants were asked to listen to the audio stimuli and rate the statements on a continuous sliding scale. In sum, listener ratings were based on a duplet of two statements, relating to whether the speaker was likely to win in a physical fight or whether a speaker was a respected leader. The responses gave a rating on a scale from 0% (strongly disagree) to 100% (strongly agree).

71 participants took part in the perception study. It was conducted online (due to pandemic restrictions) via Prolific, participants reported normal hearing and cognition and were compensated for their time (around 30 minutes). Ethics clearance was obtained from the McMaster Ethics Board (MREB).

3. RESULTS

Since our data was not normally distributed, we conducted multiple Wilcoxon signed-rank pairwise significance tests ([19], [20]) to determine if each manipulated amplitude condition is significantly different from the corresponding unaltered baselines, for both types of dominance. We corrected for multiple comparisons by adjusting p-level thresholds (Bonferroni correction).

We found significant rating differences for the *enhanced* conditions (compared to baseline) at both *sentence* level (physical dominance: $p<.025$ **, social dominance: $p<.025$ **) and at *syllable* level (physical: $p=.008$ *, social: $p<.0025$ **). *Word* level and the *reduced* intensity *syllable* condition were however not significantly different from their baselines. Figure 1 illustrates these findings with a mean plot.

In the figure, it can be seen that for the significantly different *sentence* level condition, participants, as expected, showed higher dominance ratings for the *enhanced* condition compared to baseline. However, the opposite pattern was found for the *syllable* level: Here the *enhanced* condition was rated less dominant (or to turn it around: the baseline condition was rated more dominant).

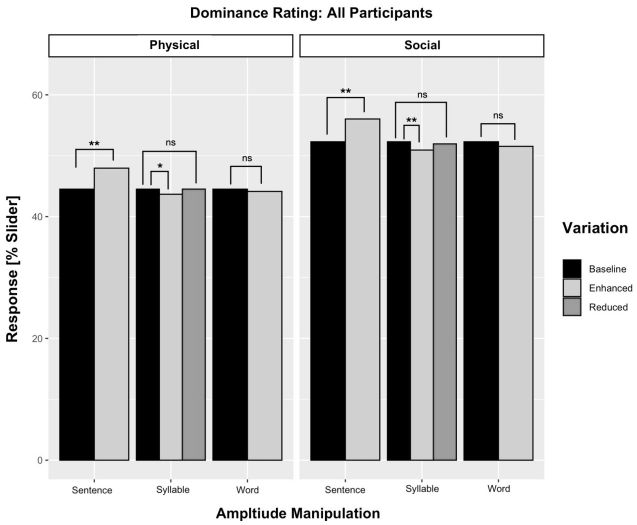


Figure 1: Mean dominance ratings over all speakers and all listeners, split by *physical* versus *social* dominance. The y-axis displays slider position percentages with 0% corresponding to the left extreme value and 100% as the right extreme value. The x-axis shows the different amplitude manipulations levels (*sentence*, *syllable*, *word*)

We found opposite directions for the significantly different dominance ratings of our listeners: Enhancing amplitude differences at *sentence* level significantly increased dominance ratings, whereas enhancing amplitude differences at *syllable* level significantly *decreased* dominance ratings. Manipulations at *word* level or the reduced condition at *syllable* level did not (significantly) affect listeners' dominance ratings.

4. DISCUSSION

Not all amplitude increases lead to higher dominance ratings, the effect strongly depends on the actual linguistic level manipulated. Also, the two examined dominance domains (*physical* versus *social*) show similar patterns across all three linguistic domains.

For the high-level *sentence* condition, our data confirm the results of previous research: increasing the overall amplitude increases the perceptual ratings of dominance [12], [13]. This is true for both dominance domains, and shows that overall increases in sentence level amplitudes positively correlate to increases in perceived *social* and *physical* dominance.

For the mid-level *word* condition, our results contradict results from [14] which found increases in amplitude for *word* stress result in higher scores within the *social* dominance environment. Our data did not confirm these differences for the social dominance domain, and we furthermore also did not find differences for *physical* dominance.

With respect to the two low-level *syllable* conditions, we are not aware of previous literature examining differences on this level. Our results show, for both the *social* and *physical* dominance domain, that enhancing amplitude differences between all stressed and unstressed *syllables* leads to significant dominance rating differences, but opposite to our findings for the *sentence* level: the *enhanced* condition significantly lowers dominance ratings compared to the baseline, whereas the *reduced* condition had no effect on dominance. It is of interest to note that the effect size for the *syllable* level condition is much smaller than *sentence* level condition (see figure 1), suggesting intensity changes on *sentence/phrase* level are more salient and robust compared to *syllabic* level changes.

5. CONCLUSION

We manipulated amplitudes on three different linguistic levels (sentence, word, syllable) and asked listeners to rate these stimuli using dominance scales with continuous slider scales. We found that our *enhanced* amplitude conditions significantly influenced listener ratings for *sentence* and *syllable* level (but not *word* level), and there was no difference comparing *physical* and *social* dominance domain ratings. However, the direction of dominance ratings was different when comparing *sentence* to *syllable* level. One explanation for the unexpected lower dominance ratings for the *enhanced syllable* condition could be due to increased attention: Unlike sentence or focus conditions, stressed syllable manipulations/variations across an entire utterance are not common. This amplitude manipulation may have marked it more auditorily salient to listeners causing decreased dominances ratings for this level.

Finally, with respect to the more salient and pronounced effect found on *sentence* level, it is worth noting that in real-life conditions listeners will not commonly judge identical speech/word material to compare matched increased versus normal amplitude conditions for full sentences and paragraphs. Thus, possible aims to increase a speaker's dominance by simply increasing overall spoken amplitude levels for a full sentence would not succeed, since the matched amplitude baseline condition would be missing here. Although the results are interesting and appear to demonstrate increased amplitude for whole utterances increases dominance ratings, without the baseline to compare the increased environment, the results remain limited in terms of real world application.

6. ACKNOWLEDGEMENTS

This work was partly funded by the Canadian NSERC grant RGPIN-2018-06518 to the second author.

7. REFERENCES

[1] Quené, H., Boomsma, G., & van Erning, R. 2016. Attractiveness of male speakers: effects of voice pitch and of speech tempo. In *Proceedings of Speech Prosody*, 8, 1086-1089.

[2] Berger, S., Niebuhr, O., & Peters, B. 2017. Winning over an audience–A perception-based analysis of prosodic features of charismatic speech. In *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, 1454-1457.

[3] Yang, Z., Huynh, J., Tabata, R., Cestero, N., Aharoni, T., & Hirschberg, J. 2020. What makes a speaker charismatic? producing and perceiving charismatic speech. In *Proc. 10th International Conference on Speech Prosody*, 2020, 685-689.

[4] Rafikova, A., Valueva, E., & Panfilova, A. 2022. Voice and Psychological Characteristics: A Contemporary Review. *Psychology. Journal of Higher School of Economics*, 19, 195-215.

[5] Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, 28(5), 340-344.

[6] Rosenberg, A., & Hirschberg, J. 2009. Charisma perception from text and speech. *Speech Communication*, 51(7), 640-655.

[7] Möbius, B. (2003). "Gestalt Psychology meets phonetics–An early experimental study of intrinsic F0 and intensity," in *Proceedings of the 15th International Congress.of Phonetic Sciences*. Barcelona, IPA and UAB, 3, 2677–2680.

[8] Niebuhr, O., & Fischer, K. 2019. Do not Hesitate!- Unless You Do it Shortly or Nasally: How the Phonetics of Filled Pauses Determine Their Subjective Frequency and Perceived Speaker Performance. In *INTERSPEECH*, 544-548.

[9] Harrigan, J. A., Gramata, J. F., Lucic, K. S., & Margolis, C. 1989. It's how you say it: Physicians' vocal behavior. *Social Science & Medicine*, 28(1), 87-92.

[10] Scherer, K. R., & London, H., & Wolf, J. J. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. Journal of Research in Personality, 7, 31-4.

[11] Buller, D. B., & Burgoon, J. K. (1986). The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. *Human Communication Research*, 13, 126-144.

[12] Aronovitch, C. D. 1976. The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology,* 99, 207-220.

[13] Tusing, K. J., & Dillard, J. P. 2000. The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, 26(1), 148-171.

[14] Kimble, C. E., & Seidel, S. D. 1991. Vocal signs of confidence. *Journal of Nonverbal Behavior*, 15(2), 99-105.

[15] Mueller, U., & Mazur, A. 1997. Facial dominance in Homo sapiens as honest signaling of male quality. Behavioral Ecology, 8, 569–579.

[16] Hairer, M. 2021 [software]. Amadeus Pro (version 2.8.8). London, United Kingdom; HairerSoft.

[17] Boersma, P, & Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.23, retrieved 8 October 2022 from http://www.praat.org/

[18] Adobe Inc. 2021. *Adobe Audition*. Retrieved from https://adobe.com/products/audition

[19] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[20] RStudio Team. 2009-2018. *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Perceptual Training Enhances the Use of Vowel Quality Cues to Lexical Stress: The Benefits of Intonational Variability

Annie Tremblay,[1] Hyoju Kim, [2] Sahyang Kim, [3] & Taehong Cho[4]

1. The University of Texas at El Paso, 2. University of Kansas, 3. Hongik University, 4. Hanyang University
actremblay@utep.edu, kimhj@ku.edu, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study investigates whether high-variability phonetic training, also known as multi-talker phonetic training, enhances Seoul Korean listeners' weightings of acoustic cues to English lexical stress and does so more than single-talker perceptual training. Seoul Korean listeners at an intermediate proficiency in English completed a cue-weighting stress perception task (pre-test), eight perceptual training sessions (over eight consecutive days) in which they heard noun-verb stress minimal pairs produced by one or four talkers and identified the word they heard, and the same cue-weighting stress perception task (post-test). In both training conditions, the stimuli varied in their intonational realizations. The results showed that both training types similarly enhanced Korean listeners' use of vowel quality cues to English lexical stress, and training increased listeners' use of pitch cues in the absence of vowel quality cues. The comparable effects of training type are attributed to the intonational variability in the training stimuli.

Keywords: Stress perception, cue weighting, perceptual training, Korean, English

1. INTRODUCTION

High-Variability Phonetic Training (HVPT), also known as multi-talker phonetic training, has been shown to be highly successful for improving listeners' discrimination and identification of difficult second-language (L2) sound contrasts, more so than single-talker phonetic training (STT) [1, 2]. HVPT has been shown to enhance listeners' perception of L2 phonetic categories [3], L2 syllable structure [4], and L2 lexical tones [5]. Phonetically variable speech has been deemed beneficial to L2 speech learning because it enables listeners to weight multiple dimensions of linguistic contrasts relative to the phonetic context in which they are heard, thus aiding listeners' development of robust L2 perceptual representations. What is less clear from this research, however, is whether the benefits of HVPT extend to the weighting of acoustic cues to lexical stress for listeners whose first language (L1) does not have lexical stress. The present study seeks to answer this question with Seoul Korean L2 learners of English.

English has lexical stress. For example, in English, words such as *DEsert* and *deSSERT* (with capitalized letters representing the stressed syllables) have different stress patterns. The stress contrast in noun-verb pairs is reflected in the alternation of full-reduced vowels. Since full and reduced vowels are realized differently in the spatial dimension, the primary acoustic correlate of lexical stress is often considered to be vowel quality [6]. Importantly, the full-reduced forms are also distinguished in the temporal dimension, such that a difference in vowel quality is accompanied by an intrinsic difference in duration, with full vowels being longer than reduced ones. Since a difference in duration does not necessarily induce a difference in vowel quality (unlike the reverse), duration is considered a secondary phonetic correlate of lexical stress [7]. Pitch is also known to be an important correlate of lexical stress [8], but pitch is realized differently as a function of phrase-level pitch accent types (L*, H*, L+H*, L*+H) [9,10], such that there is no one-to-one relationship between pitch and lexical stress, making the pitch change a less consistent correlate. In contrast, Seoul Korean does not have lexical stress. Prominence is realized intonationally by phrasal edge tones, with the Accentual Phrase (AP) having the underlying LHLH or HHLH tonal pattern (where L = low and H = high). The first tone of the AP varies as a function of the phrase-initial segment (H for fortis and aspirated segments, and L for all other segments) [11]. Consequently, pitch is an intonationally driven correlate of this segmental (also lexical) contrast.

Previous research has shown that Korean L2 learners of English (L1 dialect(s) unspecified) have more difficulty recalling sequences of English nonwords differing in stress, where the stress contrast is realized primarily with suprasegmental cues (e.g., ['mipa] vs. [mi'pa]), than sequences of English nonwords differing in a segment (e.g., ['kupi] vs. ['kuti]) [11]. However, research findings differ as to whether Korean L2 learners of English can use vowel quality cues to lexical stress in spoken word recognition. Lin et al. [12] report that Korean L2 learners of English are not more accurate at rejecting nonwords whose incorrect stress placement is signaled by vowel quality cues (e.g., **HOrizon* ['hɑɪaɪzən]) compared to incorrectly stressed nonwords without vowel quality changes

(e.g., **Enough* ['ɪnʌf]). By contrast, Connell et al. [13] found that Korean L2 learners of English show higher target-over-competitor fixation proportions when the first syllable of the target and competitor differ segmentally and suprasegmentally (e.g., *PArrot* vs. *paRADE*) than when they do not (e.g., *PArrot* vs. *PArish*), an effect not found for target and competitor words whose first syllable differ only supra-segmentally (e.g., *SURface* vs. *surPRISE* or *SURplus*). To explain this, the authors proposed that Korean L2 learners of English may have assimilated full and reduced English vowels to different Korean vowels and use vowel quality differences to distinguish target from competitor words. In other words, Korean listeners may transfer the use of spectral cues from the perception of vowels in Korean to the perception of lexical stress in English.

One important remaining issue, however, is whether HVPT can enhance the weighting of acoustic cues to English lexical stress in listeners whose L1 does not have lexical stress, and result in more target-like cue weighting than STT. Given Seoul Korean listeners' sensitivity to pitch as a prosodic cue that signals a segmental contrast (lenis vs. fortis and aspirated segments in phrase-initial position), we might expect Korean listeners to rely on pitch cues to English lexical stress. What remains to be seen is whether HVPT can help them rely less on pitch and more on vowel quality when perceiving English lexical stress. The present study will elucidate whether this is the case, comparing the efficiency of HVPT and STT for enhancing cue-weighting. Since this is the first study that seeks to answer this question, it is unclear whether perceptual training should target a specific cue distribution or whether a distribution of cues that mimics spoken English (as established from corpus studies) would be sufficient to enhance learning. We opted for the latter as a starting point into this investigation. The cue-weighting task that served as pre- and post-test in this study is the one that Tremblay et al. [14] used to test Dutch listeners' perception of English lexical stress.

2. METHOD

2.1. Participants

The participants were 54 Seoul Korean L2 learners of English (mean age: 24, 32 females) (for a comparison with native English listeners, see [14]). The Korean listeners were tested at a Korean university in Seoul. Among them, 27 were randomly assigned to HVPT and 27 to STT. All participants completed a detailed language background questionnaire and the Lexical Test for Advanced Learners of English (LexTALE) [15] to measure lexical proficiency in English (mean: 68.1, SD: 8.4). The two L2 subgroups did not differ significantly in their age, age of first exposure to English, years of English education, English proficiency self-ratings, English accent self-ratings, or LexTALE score. No participant reported a history of speech, language, or hearing impairments.

2.2. Materials

2.2.1. Training stimuli

The words that served as training stimuli were noun-verb minimal pairs that differed in lexical stress. The lexical items were 28 English noun-verb pairs for which at least one of the two vowels was reduced when unstressed (e.g., *REcall* vs. *reCALL*), and 8 English noun-verb minimal pairs without vowel reduction (e.g., *PERmit* vs. *perMIT*). This distribution of words with and without vowel reduction cues (78% vs. 22%) was based on Cutler and Carter [16]'s corpus study. The token frequency of the words as a noun and that of a verb was controlled based on the Corpus of Contemporary American English [17].

The auditory stimuli for HVPT were recorded by four native speakers of American English (two male and two female). The auditory stimuli for STT were recorded by a female speaker who also recorded items for HVPT. The 36 word pairs were elicited and recorded with three different intonations (H*L–, L*H–, and flat intonational contour). The target words with H*L– were elicited in the declarative sentence *Mary said ____ before*. The words with L*H– were elicited in the interrogative sentence *Mary said ____ before?* The words with a flat intonational contour were elicited in the carrier phrase *MARY said ____ before*, where *MARY* had a contrastive pitch accent and where the target word was deaccented. The intensity of the words was normalized to 70 dB.

2.2.2. Pre-/post-test stimuli

The word pair that was used for the cue-weighting stress perception task that served as pre- and post-test was *DEsert-deSSERT.* It was recorded with a H* pitch accent by a female native speaker of American English who did not record the training stimuli; thus, any learning from the training is evidence for generalization to a new talker. The word pair was elicited in the carrier sentence *Click on* ____. One token of *DEsert* served as the base token and was manipulated to have seven steps for each acoustic dimension (vowel quality, pitch, and duration). The values corresponding to step 1 (*DEsert*) and step 7 (*deSSERT*) in each dimension were based on the naturally produced tokens. The voiced portion of the syllables had its formant structure (F1, F2, F3, and corresponding bandwidths), duration, and pitch

manipulated, and intensity between the two stress patterns neutralized. Two of the dimensions (formant structure and pitch, formant structure and duration, pitch and duration) were orthogonally manipulated in 7 steps, holding the other two dimensions at Step 4 (for details, see [14]). This manipulation yielded 147 auditory stimuli (3 matrices of 7 × 7 stimuli).

2.3. Procedures

The complete experimental procedure lasted ten days. Participants completed the pre-test (i.e., the cue-weighting speech perception task) in the lab on the first day of participation. In each trial of the pre-test, participants heard an auditory stimulus over headphones and were asked to press the left arrow on the keyboard if they thought they heard *DEsert* and the right arrow if they thought they heard *deSSERT*. Each trial ended with the participant's response followed by a 1,000 ms inter-trial interval. The main session—including a total of 441 trials (147 stimuli × 3 repetitions)—was divided into three blocks, and test items were randomized across participants. The task lasted approximately 15 minutes.

The training was conducted remotely. On eight consecutive days following the pre-test, participants completed eight 20-minute training sessions, with no more than one training session per day. On each trial of the training, participants heard an auditory stimulus and were asked to decide whether the stimulus was a noun or a verb (they were told nouns would be stressed word-initially verbs word-finally), and they received explicit feedback on the accuracy of their responses. In each training session, participants heard 36 noun-verb pairs repeated 4 times (total: 288 stimuli). For the HVPT group, in each of Sessions 1-4, participants heard 2 talkers, with the stimuli being repeated twice per talker, and in each of Sessions 5-8, participants heard all 4 talkers. In the STT training, where participants heard only 1 talker, the stimuli were repeated four times. In each training session, 41.7% of the words had a H*L– intonation, 16.6% had an L*H– intonation, and 41.7% had a flat intonational contour, mimicking the distribution of word-level intonations reported in Im, Cole, and Baumann [18]. The distribution of intonational patterns was counterbalanced across talkers in the HVPT training and across the two stress patterns in both types of training.

On the day following the last training session, participants completed the post-test, which was identical to the pre-test.

2.4. Data analysis

Mixed-effects logistic regression models were conducted on Seoul Korean listeners' proportion of *DEsert* (coded as 1) and *deSSERT* (coded as 0). Separate models were built for each of the three stimulus matrices (i.e., vowel quality by pitch, vowel quality by duration, and pitch by duration). For each model, the fixed effects included two manipulated dimensions (each centered), their interactions, test (pre- vs. post-), and training type (HVPT vs. STT). Korean listeners' response on the pre-test in the HVPT condition served as baseline. Random intercepts were participant and item. The largest model was backward fit using log-likelihood ratios. Only the models with the best fit are presented.

3. RESULTS

Seoul Korean listeners' proportion of *DEsert* selection for each of the three stimulus matrices, test, and training type is shown in Fig. 1., together with English listeners' proportion of *DEsert* selection (for reference).

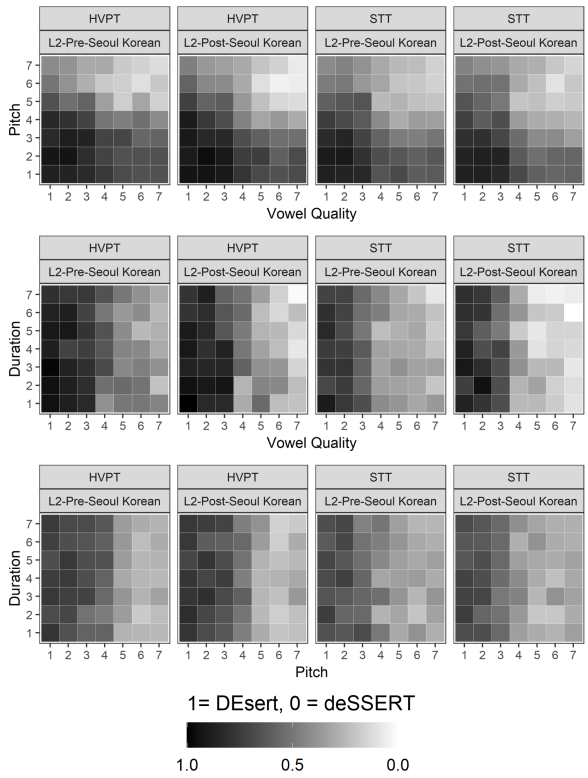


Figure 1: Seoul Korean listeners' proportion of *DEsert* vs. *deSSERT* selection as a function of test and training type when the stimuli varied by vowel quality and pitch (top), vowel quality and duration (middle), and pitch and duration (bottom). The two left panels show the results of HVPT, and the two right panels show the results of STT.

When the stimuli varied by vowel quality and pitch (top panels), the model with the best fit (Table 1) had the following structure: responses ~ (vowel.quality + pitch) * test + (vowel.quality + pitch) * training.type + (1|participant) + (1|item). The model revealed

significant effects of vowel quality, pitch, and test. Importantly, the model yielded a two-way interaction between vowel quality and test, with Korean listeners in the HVPT group showing a stronger effect of vowel quality (decrease in *DEsert* selection as step increased) in the post-test than in the pre-test, and an interaction between pitch and training type, with the effect of pitch (decrease in *DEsert* selection as step increased) being stronger in the pre-test results of the STT group than in those of the HVPT group. The lack of three-way interaction between cue, test, and training type suggests that HVPT is not superior to STT for altering listeners' cue weighting.

Table 1. Mixed-effects Logistic Regression with Best Fit for Vowel Quality (VQ) x Pitch Stimuli

	Est.	SE	*z*	*Pr*(>|*z*|)
(Intercept)	0.25	0.09	2.77	.006
VQ	−0.39	0.03	−13.02	< .001
Pitch	−0.54	0.03	−18.30	< .001
Test (Post-test)	−0.11	0.04	−3.10	.002
Training (STT)	−0.05	0.10	<|1|	>.1
VQ × Test (Post-test)	−0.06	0.02	−2.86	.004
Pitch × Test (Post-Test)	−0.03	0.02	−1.34	>.1
VQ × Training (STT)	0.01	0.02	<|1|	>.1
Pitch × Training (STT)	0.12	0.02	6.17	< .001

When the stimuli varied by vowel quality and duration (middle panels), the model with the best fit had the following structure: response ~ (vowel.quality + duration) * test + (vowel.quality + duration) * training.type + (1|participant) + (1|item). As can be seen in Table 2, the model revealed significant effects of vowel quality, duration, and test. Again, the model yielded a significant two-way interaction between vowel quality and test, with Korean listeners in the HVPT condition showing a greater effect of vowel quality in the post-test than in the pre-test. No other interaction was significant, suggesting again that the two training types did not differ in their ability to alter listeners' cue weightings.

When the stimuli varied by pitch and duration (bottom panels), the model with the best fit had the following structure: response ~ (pitch + duration) * test * training.type + (1|participant) + (1|item). As presented in Table 3, the model revealed significant effects of pitch, duration, and test. Crucially, the model also yielded a significant two-way interaction between pitch and test, indicating that Korean listeners in the HVPT condition relied more on pitch from pre-test to post-test, and a significant pitch-by-training-type interaction, suggesting that Korean listeners' use of pitch in the pre-test was stronger for the HVPT group than for the STT group. Again, the lack of three-way interaction between cue, test, and training type indicates that the two training types did not differ in their altering of listeners' cue weighting.

Table 2. Mixed-effects Logistic Regression with Best Fit for Vowel Quality (VQ) x Duration Stimuli

	Est.	SE	*z*	Pr(>|*z*|)
(Intercept)	0.47	0.14	3.29	.001
VQ	−0.42	0.02	−18.09	< .001
Duration	−0.05	0.02	−2.33	.020
Test (Post-test)	−0.33	0.04	−9.32	< .001
Training (STT)	−0.19	0.20	<|1|	>.1
VQ × Test (Post-test)	−0.08	0.02	−4.45	< .001
Duration × Test (Post-test)	−0.02	0.02	−1.26	>.1
VQ × Training (STT)	0.00	0.02	<|1|	>.1
Duration × Training (STT)	0.00	0.02	<|1|	>.1

Table 3. Mixed-effects Logistic Regression with Best Fit for Pitch x Duration Stimuli

	Est.	SE	*z*	Pr(>|*z*|)
(Intercept)	−0.02	0.09	−0.18	>.1
Pitch	−0.44	0.02	−19.95	< .001
Duration	0.04	0.02	2.00	.045
Test (Post-test)	−0.14	0.03	−4.09	< .001
Training (STT)	−0.03	0.12	<|1|	>.1
Pitch × Test (Post-test)	−0.06	0.02	−3.25	.001
Duration × Test (Post-test)	−0.02	0.02	−1.26	>.1
Pitch × Training (STT)	0.13	0.02	7.43	< .001
Duration × Training (STT)	−0.02	0.02	−1.31	>.1

4. DISCUSSION AND CONCLUSION

This study investigated whether HVPT can help Seoul Korean listeners rely less on pitch and more on vowel quality when perceiving English lexical stress, and whether the benefits of HVPT are superior to those of STT. The results showed comparable effects of training types on Korean listeners' use of vowel quality cues to English lexical stress. The similar training type effects are attributed to the considerable intonational variability that was introduced in the stimuli for both training types, resulting in greater ability for listeners in the STT group to extract vowel quality cues to English lexical stress. The increased use of pitch with training in the absence of vowel quality cues indicates that the greater occurrence of H* than of L* in the training may lead listeners to rely more on this cue after the training (especially since this was the cue in the pre- and post-test stimuli), suggesting that future training should make pitch cues to lexical stress unpredictable. This is the first study to show beneficial effects of perceptual training on listeners' weighting of acoustic cues to lexical stress.

5. REFERENCES

[1] Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *J. Acoust. Soc. Am., 89*, 874-886.

[2] Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am., 94*, 1242-1255.

[3] Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *J. Acoust. Soc. Am., 118*, 3267-3278.

[4] Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *J. Phon., 52*, 105-120.

[5] Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *J. Acoust. Soc. Am., 106,* 3649-3658.

[6] Gay, T. (1978). Physiological and acoustic correlates of perceived stress. *Lg. and Speech, 21,* 347-353.

[7] Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7-33). Cambridge: Cambridge University Press.

[8] Lehiste, I. (1970). *Suprasegmentals.* Cambridge: MIT Press.

[9] Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in English and Japanese. *Phonology Yearbook, 3,* 255-310.

[10] Ladd, D. R. (2012). *Intonational Phonology.* Cambridge: Cambridge University Press.

[11] Jun, S.-A. (1998). The Accentual Phrase in the Korean prosodic hierarchy. *Phonology, 15,* 189-226.

[12] Lin, C. Y., Wang, M. I. N., Idsardi, W. J., & Xu, Y. I. (2014). Stress processing in Mandarin and Korean second language learners of English. Bilingualism: *Lg. and Cog., 17,* 316-346.

[13] Connell, K., Hüls, S., Martínez-García, M. T., Qin, Z., Shin, S., Yan, H., & Tremblay, A. (2018). English learners' use of segmental and suprasegmental cues to stress in lexical access: An eye-tracking study. *Lg. Learning, 68,* 635-668.

[14] Tremblay, A., Broersma, M., Zeng, Y., Kim, H., Lee, J., & Shin, S. (2021). Dutch listeners' perception of English lexical stress: A cue-weighting approach. *J. Acoust. Soc. Am., 149,* 3703-3714.

[15] Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behav. Res. Methods, 44,* 325-343.

[16] Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Lg., 2,* 133-142.

[17] Davies, M. (2008-) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus.byu.edu/coca/.

[18] Im, S., Cole, J., & Baumann, S. (2018). The probabilistic relationship between pitch accents and information status in public speech. *Proceedings of 9th International Conference on Speech Prosody* (pp. 508-511). Poznán, Poland.

EFFECT OF SPEECH ENHANCEMENT AT PHONETIC LEVEL PERCEPTION OF ENGLISH SPEECH BETWEEN NATIVE ENGLISH AND MANDARIN LISTENERS

Yunqi C. Zhang[1], Catherine I. Watson[2], C.T Justine Hui[1], Yusuke Hioka[1]

[1]Acoustics Research Centre, Department of Mechanical and Mechatronics Engineering, University of Auckland, New Zealand
[2]Department of Electrical, Computer and Software Engineering, University of Auckland, New Zealand
yzhb694@aucklanduni.ac.nz; c.watson@auckland.ac.nz; justine.hui@auckland.ac.nz; yusuke.hioka@ieee.org

ABSTRACT

This study investigates the effect of commonly used speech enhancement algorithms in improving the intelligibility of noisy New Zealand English speech on native New Zealand English and native Mandarin listeners at the phonetic level. A phonetic error analysis was carried out to analyse the errors made by the listeners from a subjective listening test to find systematic errors. Results show that existing speech enhancement algorithms performed similarly by causing more onset deletion and coda deletion or insertion than when listening to clean speech. Overall, the algorithms did not improve the listeners' speech intelligibility on the phonetic level. Although native New Zealand English listeners always made fewer errors than native Mandarin listeners, they all made similar onset and coda errors. The nucleus errors made by the native Mandarin listeners were mainly caused by unfamiliarity with the accent.

Keywords: phonetic analysis, Mandarin, non-native, speech intelligibility, speech enhancement

1. INTRODUCTION

Speech intelligibility indicates how well speech can be comprehended by listeners. Non-native (L2) listeners have been known to experience more severe speech intelligibility degradation than native (L1) listeners under noise, such as listening or talking in a noisy environment [1, 2, 3]. To improve the speech intelligibility of listeners when listening to noisy speech, assistive listening devices implementing speech enhancement (SE) algorithms could be used [4]. While numerous numbers of SE algorithms have been developed and their performance has been thoroughly evaluated [5], to date their effect on L2 listeners is not well understood.

Listeners' speech perception was found to be influenced by their L1 language due to various factors such as the contrasts in phonetic familiarity, lexical structure, and acoustic structure between the L1 and L2 languages [3, 6]. As reported by the Ministry of Education of the People's Republic of China in 2019 [7], there were 703,500 Chinese students studying overseas and mainly entering English-speaking countries [8]. Hence, this study focuses on the perception of English speech by native Mandarin listeners.

Our previous study [9] investigated the effect of different SE algorithms on two groups of listeners: native English listeners residing in New Zealand as the control group and native Mandarin listeners residing in Mainland China who have never been immersed in an English-speaking country as the experimental group. It was found that all SE algorithms tested showed little improvement or even degradation in speech intelligibility compared to the original noisy speech for both groups, and the perception gap between the native English and native Mandarin listeners was not narrowed. So far, most studies on enhanced speech evaluated speech intelligibility by the correctness of each word, but none has investigated the response error pattern at the phonetic level. To ensure the corpus is in the L1 language of the control group, the previous study used a corpus in New Zealand English (NZE), despite most students in Mainland China leant General American (GenAm) or British English at school [10]. Therefore, NZE is considered less familiar to the Mandarin participants and the effect of using NZE corpus in such listening tests remains unknown. NZE shares the same vowel system as other non-rhotic standard varieties of English such as Received Pronunciation, which consists of eleven monophthongs plus a neutral schwa [11], but differs from other non-rhotic varieties for the /ɪ, e, iː, æ/

vowels [12]. Despite NZE being less commonly used, 35% of the international students in New Zealand were imported from Mainland China [8], showing that NZE is used and can be understood by a part of the local Mandarin speakers in noisy environments.

The current study analyses the errors made by the two groups of listeners outlined in [9] for each phonetic component. The analysis explores the possible common or unique errors made by each group and gives a glimpse of the perception of native Mandarin listeners on NZE.

2. METHODOLOGY

Phonetic level analysis was applied to the results of two subjective listening tests conducted in [9], which were 1) the *baseline test* that evaluated the clean speech from a NZE corpus by native Mandarin listeners and 2) the *speech enhancement test* that investigated the intelligibility of noisy speech processed by selected SE algorithms on native English and native Mandarin listeners. The response keywords were divided into phonetic components for marking and the systematic errors made were analysed. The study was approved by the University of Auckland Human Participants Ethics Committee (UAHPEC24202).

2.1. Subjective listening tests

The subjective listening tests were conducted online where the participants were asked to use their own headphones and transcribe the stimuli. The speech intelligibility was evaluated by counting the number of words correctly answered, which was normalised by the total number of words. Details of the test design can be found in [9].

2.1.1. Baseline Test

Nine native Mandarin listeners based in Mainland China (C_{clean} group) were recruited. All participants reported they had learnt or had been learning American (n = 7) or British (n = 2) English. Each participant transcribed 48 clean (without any noise added) English sentences.

2.1.2. Speech Enhancement (SE) Test

Twenty native NZE listeners residing in New Zealand (NZE_{SE} group) and 19 native Mandarin listeners residing in Mainland China (C_{SE} group, referred as CC group in [9]) were recruited. All participants in the C_{SE} group reported have learnt either British (n = 6) and/or American English (n = 13). Each participant transcribed 108 noisy speech sentences, which included six types of speech: the original noisy speech (before SE was applied) and the speech enhanced by five widely used SE algorithms, namely Wiener filter (WF) [13], subspace (SS) [14], non-negative matrix factorisation (NMF) [15], Conv-TasNet (Conv) [16], and complex U-Net (Unet) [17].

2.1.3. Stimuli

To generate the stimuli used in the test, the Bamford-Kowal-Bench (BKB) [18] sentences from the Speech Perception Assessment New Zealand (SPANZ) corpus [19] were used. The corpus is designed for participants speaking NZE, i.e. the sentences were recorded in NZE accent and revised to accommodate commonly used expressions in New Zealand. According to the description in the original BKB sentences, it was assumed that the occurrence of the vowels and phonemes in the keywords was balanced. Each sentence contains 3 - 4 keywords to be marked; any words other than the keywords were not marked.

2.2. Phonetic Analysis

The correctness of participants' responses from the subjective listening tests was marked manually by the first author according to the suggestions in the BKB corpus [18]. All keywords that received incorrect answers were collected for phonetic analysis. This is because even though the response words differed for each participant, the phoneme they tend to make mistakes on may be the same.

A phoneme that occurs at different locations can have various pronunciations according to the surrounding phoneme environments, which is known as co-articulation [20]. Therefore, recording the location of the mistaken phonemes can be crucial for analysing the errors specifically. This was realised by dividing every syllable into its phonetic components: onset, nucleus, and coda. For multi-syllable words, the phonetic components were numbered to the syllable they were in.

Overall, each word was divided into their syllables and further divided by phonetic components. Absence of a component was recorded as *NA*. An example is the two syllable word, "happy" [hæ-piː], being separated into onset 1 /h/, nucleus 1 /æ/, coda 1 *NA*, onset 2 /p/, nucleus 2 /iː/, and coda 2 *NA*. Responses with inconsistent number of syllables compared to that of the keywords were recorded with the difference

in the number of syllables, where a positive number indicates more syllables were answered than that in the keyword, and vice versa. The phonetic analysis was applied to the baseline and SE test results separately. The analysis for the SE test also recorded different SE algorithms and groups of participants separately.

This study only investigated the phonemes in the first two syllables in each word as all of the baseline test keywords (83.3% mono-, 16.7% double-syllable) and 98.4% of the SE test keywords were either mono- (69.4%) or double-syllable (28.8%). The responses of three-syllable keywords were excluded as the errors made by the participants were too few to provide any systematic error, i.e. all combinations of keyword and error occurred less than twice. Errors of the same phonetic component in different syllables were not combined as the second syllable had a different phoneme error pattern compared to the first syllable.

3. RESULTS AND DISCUSSION

According to the phonetic analysis, common phonetic errors can be classified into deletion, missing the whole phonetic component; insertion, making up non-existing phonetic component; and substitution, substituting certain phonemes or the whole phonetic component with other phonemes. The errors were analysed by each phonetic component separately and listed in tables. Since the main purpose of this study is to investigate the performance of SE algorithms on native Mandarin listeners, the tables in this paper will focus on the error occurrence of the C_{SE} group. Table 1 shows the three most common nucleus errors for the C_{SE} group for each SE algorithm from the first syllable, where the corresponding error occurrence of the NZE_{SE} group is also reported. Tables 2, 3, and 4 show the most common onset, nucleus, and coda errors of all three groups, respectively, where for the C_{SE} group only the errors occurred over 15 times are presented due to space constraint. The onset errors of the C_{clean} group are excluded due to the absence of systematic onset errors, i.e. all combinations of keyword and error appeared only once.

Errors of the phonetic components were first analysed for each SE algorithm, but little difference was found between algorithms. For example, as shown in Table 1, most of the common nucleus errors reappear in every algorithm. The unique [/ɜː/-/ou/] confusion in noisy speech also occurred in most SE algorithms for the C_{SE} group (twice for SS and Unet, three times for WF and NMF). The onset and coda errors caused by each SE algorithm were mostly deletion errors, which are not shown in this paper. The only exception was the /ʃ/-/h/ confusion in speech enhanced by NMF for the C_{SE} group (n = 5) in onset. As a result, analysis of aggregated results from all SE algorithms was conducted.

For common onset errors, as reported earlier, no systematic errors (n = 1) were found from the C_{clean} group. However, for the C_{SE} group, the occurrence of the most frequent systematic onset errors increased (n = 39) similar to that of the nucleus (n = 46) and coda errors (n = 32). This increase is expected as the consonants at the beginning of an English word are often voiceless [21]. Therefore, the onset, especially the onset of the first syllable, can be masked easily by noise and/or removed by the SE algorithm. The most frequent

Table 1: The three most common occurred nucleus errors for C_{SE} group in the first syllable for each SE algorithm.

SE Algorithm	Nucleus error (C_{SE}count, NZE_{SE}count)		
Noisy	eɪ → aɪ (7, 3)	e → iː (4, 1)	ɜː → ou (4, 0)
WF	eɪ → aɪ (7, 7)	ə → *NA* (5, 5)	ɒ→ ou (5, 1)
SS	eɪ → aɪ (9, 9)	e → eɪ (8, 0)	ɒ→ ʌ (5, 0)
NMF	eɪ → aɪ (9, 10)	ə → *NA* (8, 5)	ɒ→ oː (7, 1)
Conv	eɪ → aɪ (11, 9)	ə → *NA* (6, 4)	e → eɪ (6, 0)
Unet	ɒ→ ou (7, 0)	ə → *NA* (6, 2)	ɒ→ oː (5, 0)

Table 2: Most frequent onset errors (C_{clean} group is excluded as no systematic onset errors).

C_{SE} group	Count	NZE_{SE} group	Count
Onset 1 (first syllable)			
ð → *NA*	38	ð → *NA*	39
ð → h	17	p → *NA*	15
k → h	16	f → *NA*	13
p → *NA*	15	h → *NA*	13
f → *NA*	15	p → p l	11
Onset 2 (second syllable)			
d → *NA*	17	l → *NA*	6
s → *NA*	17	r → *NA*	6
l → *NA*	10	b → *NA*	5
d → t	9	d → *NA*	5
r → *NA*	9	s → *NA*	5

Table 3: Most frequent nucleus errors.

C_{clean} group	Count	C_{SE} group	Count	NZE_{SE} group	Count
Nucleus 1 (first syllable)					
e → ɪ	9	eɪ → aɪ	46	eɪ → aɪ	41
ʌ→ æ	7	ə → *NA*	31	ə → *NA*	25
e → iː	6	e → eɪ	26	æ→ aː	14
æ→ e	3	ɒ→ ou	20	æ→ e	11
ɪə → iː	3	**ɒ→ oː**	19	eɪ → ə	11
ou → oː	2	**e → ɪ**	17	au → e	11
æ→ ʌ	2	ɒ→ ʌ	17	ou → ʌ	7
eɪ → aɪ	2	**æ→ e**	16	eɪ → æ	7
ʌ→ aː	2	**e → iː**	16	uː → iː	7
ɪ → iː	2	**ɜː → uː**	16	iː → ə	7
ɪ → ʌ	2	**eɪ → iː**	15	æ→ eɪ	6
Nucleus 2 (second syllable)					
iː → ə	2	ə → *NA*	65	ə → *NA*	32
		ɪ → *NA*	21	ɪ → *NA*	8
		iː → ɪ	21	iː → ɪ	7

Table 4: Most frequent coda errors.

C_{clean} group	Count	C_{SE} group	Count	NZE_{SE} group	Count
Coda 1 (first syllable)					
l → t	3	*NA* → n	32	*NA* → n	12
NA → d	3	*NA* → d	20	n t → t	11
t → tʃ	3	n → *NA*	15	k → t	8
d → t	2	*NA* → l	15	*NA* → d	8
Coda 2 (second syllable)					
		ŋ → *NA*	27	ŋ → *NA*	10
		NA → ŋ	18	n → *NA*	8
		n → *NA*	16	*NA* → ŋ	8

type of onset error in the C_{SE} group was deletion, which was the same for the NZE_{SE} group, indicating that the deletion of phonemes is not due to their language nativity but a consequence of the noise and SE algorithms.

Many of the frequent nucleus errors made by the Mandarin listeners (C_{clean} and C_{SE} groups) can be explained by the unfamiliarity with the NZE accent, which are shown in bold in Table 3, including the well-known centralised /ɪ/, the raised /e/ [11], and the more recently studied merging [/ɜː/-/uː/] [22]. Some vowel confusions may be caused by the listeners being more familiar with the rhotic GenAm accent, which is learnt by most of the Mandarin participants, than NZE accent. For example, the [/ʌ/-/æ/] confusion can be explained by the fact that /æ/ in GenAm is usually used as /ʌ/ in NZE; the [/ɒ/-/oː/] confusion may be due to the fact that NZE distinguishes /ɒ/ and /oː/ while /ɒ/ and /ɑ/ merged in GenAm [23, 21]. The most common nucleus errors in the second syllable were identical between the C_{SE} and NZE_{SE} groups, indicating that under noisy conditions, even the native listeners have difficulty distinguishing the vowel lengths between /ɪ/ and /iː/. Moreover, the schwa in both syllables was ignored frequently, 96 times by the C_{SE} group, 57 times by the NZE_{SE} group in total, as it is an unstressed vowel with a very short duration and usually appears in the unstressed syllable in the corpus, e.g. "along" to "long". Hence, it can be easily missed or masked by noise. All three groups made two common vowel errors, the [/eɪ/-/aɪ/] confusion (e.g. "hay" to "hi") and the [/æ/-/e/] confusion (e.g. "bag" to "bed"), where the latter is regarded as a characteristic of NZE accent. Both confusions imply that under noisy conditions, even native listeners encounter difficulties in distinguishing certain vowel pairs.

As seen in Table 4, the common coda errors from the C_{clean} group were mainly substitution errors, which can be explained by the language structure of Mandarin in which no obstruent consonant in the word-final position is allowed [24]. Hence, the C_{clean} group may be less sensitive to the word-final sound even when listening to clean speech. For the SE test, most coda errors by both groups can be concluded as deletion or insertion errors. This can be explained by the same reason for the missing onset, as the consonants in English at the end of a word are also mostly voiceless [21]. The NZE_{SE} group showed much fewer coda errors than the C_{SE} group, which may be because the NZE_{SE} group was able to notice small but significant vowel duration differences to distinguish voicing minimal pairs compared to the C_{SE} group [25]. The coda errors in the second syllable for both C_{SE} and NZE_{SE} groups were the same, related to the deletion or insertion of /n/ and /ŋ/. One possible explanation of this finding is that the SE algorithms tend to manipulate the noisy signal in a way that the nasal consonants sound similar to noise. The high occurrence of /ŋ/ may be due to the fact that many double-syllable keywords in the corpus are gerund, e.g. "playing".

4. CONCLUSION

This study analysed the phonetic errors from three groups of results from listening tests that evaluated the intelligibility of speech: native Mandarin listeners' responses to clean NZE speech (C_{clean} group), native Mandarin (C_{SE} group) and English listeners' (NZE_{SE} group) responses to enhanced NZE speech. The errors between different algorithms showed little difference, indicating that the existing SE algorithms did not improve the speech intelligibility of either native English or native Mandarin listeners at phonetic level. The result showed that the NZE_{SE} group had much fewer errors than the C_{SE} group for all phonetic components. Compared to the C_{clean} group, the C_{SE} group tended to have much more deletion errors for onset, and deletion and insertion errors for coda. Such error patterns in the onset and coda were also similar for the NZE_{SE} group.

The main limitation of this study is that the corpus is not designed for phonetic analysis. Since meaningful sentences were used in the test, participants with higher language proficiency would have been able to use semantic cues to guess the correct keywords. This top-down effect should have been considerably reduced as most of the keywords in a sentence had multiple possible substitutions. The meaningful sentences also lead to the test containing unbalance phones, resulting in a less comprehensive analysis for each phoneme. Also, part of the manual marking involved guessing the response, hence, the result may not fully reflect the participants' perception. Further experiments with a corpus that is designed for phonetic analysis would be recommended.

ACKNOWLEDGEMENT

We thank the participants for their participation, Dr. Suzanne Purdy for providing the SPANZ corpus, and Dr. Elaine Ballard for her insightful advice.

5. REFERENCES

[1] M. Cooke, M. L. G. Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception." *The Journal of the Acoustical Society of America 123, no. 1*, pp. 414–427, 2008.

[2] S. L. L. M. C. Mattys, C. K. W. Li, and S. L. Y. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers." *Speech Communication, Non-native Speech Perception in Adverse Conditions, 52, no. 11*, pp. 887–99, 2010.

[3] M. L. G. Lecumberri, M. Cooke, and A. Cutle, "Non-native speech perception in adverse conditions: A review." *Speech Communication, 52, no. 11*, pp. 864–86, 2010.

[4] P. Loizou, *Speech Enhancement: Theory and Practice*, 2007.

[5] D. S., R. Deshmukh, and P. P., "A review of speech signal enhancement techniques." *International Journal of Computer Applications 139*, pp. 23–26, 2016.

[6] L. Polka, "Characterizing the influence of native language experience on adult speech perception." *Perception & Psychophysics 52, no. 1*, pp. 37–52, 1992.

[7] Statistics on chinese learners studying overseas in 2019 - ministry of education of the people's republic of china [www document]. [Online]. Available: http://en.moe.gov.cn/news/press_releases/202012/t20201224_507474.html

[8] OECD, "What is the profile of internationally mobile students?" 2021.

[9] Y. C. Zhang, C. J. Hui, Y. Hioka, and C. I. Watson, "Performance of speech enhancement algorithms for native Mandarin listeners on English perception," in *Proc. 15th ICA International congress on acoustics*, South Korea, 2022, pp. 111–116.

[10] W. Wenfeng and G. Xuesong, "English language education in china: A review of selected research." *Journal of Multilingual and Multicultural Development 29, no. 5*, pp. 380–99, 2008.

[11] C. I. Watson, M. Maclagan, and J. Harrington, "Acoustic evidence for vowel change in new zealand english." *Language Variation and Change 12, no. 1*, pp. 51–68, 2000.

[12] M. Maclagan and J. Hay, "Getting fed up with our feet: Contrast maintenance and the new zealand english 'short' front vowel shift." *Language Variation and Change 19, no. 1*, pp. 1–25, 2007.

[13] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, 1996, pp. 629–632 vol. 2.

[14] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[15] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," *ICLR*, p. 20, 2019.

[18] J. Bench, Ã. Kowal, and J. Bamford, "The BKB (bamford-kowal-bench) sentence lists for partially-hearing children," *British Journal of Audiology*, vol. 13, no. 3, pp. 108–112, 1979.

[19] J.-H. Kim and S. Purdy, "Speech perception assessments new zealand (SPANZ)," *New Zealand Audiological Society Bulletin*, vol. 24, pp. 9–16, 2014.

[20] R. Daniloff and R. Hammarberg, "On defining coarticulation." *Journal of Phonetics 1*, pp. 239–248, 1973.

[21] H. Rogers, *The Sounds of Language: An Introduction to Phonetics.* Taylor & Francis Group, London, United Kingdom, 2000.

[22] M. Maclagan, C. I. Watson, R. Harlow, J. King, and P. Keegan, "Investigating the sound change in the new zealand english nurse vowel /ɜː/," *Australian Journal of Linguistics 37, no. 4*, pp. 465–85, 2017.

[23] M. Hay, J.and Maclagan, E. Gordon, J. Beal, P. Honeybone, and A. McMahon, *New Zealand English.* Edinburgh University Press, Edinburgh, United Kingdom, 2008.

[24] C.-C. Cheng and C.-C. Cheng, *A Synchronic Phonology of Mandarin Chinese*. De Gruyter, Inc., 1973.

[25] J. E. Flege, M. J. Munro, and L. Skelton, "Production of the word-final english /t/-/d/ contrast by native speakers of english, mandarin, and spanish." *The Journal of the Acoustical Society of America 92, no. 1*, pp. 128–43, 1992.

The perception of visual phonetic information: Eye-tracking while searching for segmental versus prosodic cues

Xizi Deng, Erin Jastrzebski, Elise McClay, H. Henny Yeung, Yue Wang

Simon Fraser University

ABSTRACT

This eye-tracking study investigated visual scanning of a face when searching for different types of phonetic information. Twenty-four native English perceivers heard two audio sentences and then saw a silent talking face. The audio sentences sometimes differed in segmental and/or prosodic information, and perceivers had to decide which sentence the silent face articulated. Several types of prosodic differences were tested (contrastive word focus, phrasal bracketing, and intonation). Results consistently showed more overall mouth looking when identifying segments than prosody, and there were distinct patterns of shifts in looking between the mouth and non-mouth areas for processing prosody versus segments, although this was not consistent across different types of prosodic differences. Overall, these results show that visual perception of phonetic cues is affected by the type of phonetic cue itself: More mouth looking for segments than prosody. Moreover, we also observed variable looking patterns for different types of visual prosody, suggesting avenues for future research.

Keywords: audio-visual processing, speech, perception, phonetic, prosody, eye-tracking

1. INTRODUCTION

Useful speech information is encoded in many sensory modalities besides audition, like visual cues from an interlocutor's face, which can enhance auditory speech comprehension [1]–[15]. One leading reason why visual cues enhance auditory speech comprehensibility is that critical speech cues are embedded in particular areas in the face. For example, a number of studies have suggested that adults pay attention to the mouth during speech tasks when identifying segments of consonant-vowel (CV) syllables [16], or when identifying which face is the source of auditory speech [17]. For example, a classic study, Lansing & McConkie (1999), found that participants' gaze duration at the lower part of the face was longer when making decisions about segments than intonation. These results suggest that the mouth area of a talking face provides useful cues for decoding segmental information. On the other hand, prosodic information also carries critical cues in speech communication by accentuating pragmatic or syntactic information, and facial movement of the whole head or around the eyes may be helpful for linguistic prosody [7]–[14]. Correspondingly, Swerts & Krahmer (2008) found that the upper part of a face has stronger cue value in the detection of word-level prosodic prominence in a three-word sentence. Similarly, the production of visual cues in the upper part of the face was found to be more related to distinguishing echoic questions from statements [10], as these features are linked with variations in F0 [18]. However, many others have found instead that the lower part of the face [10], including chin-opening displacement [19], lip area and jaw-opening [20], can be more facilitative of processing contrastive word focus than other facial gestures, even though head movement and eye-brow displacement exerted a smaller but still significant contribution to the correct perception of this word focus [19]. Overall, since prosody may convey information in a wide range of linguistic domains (from local word to global sentence levels), research has not been conclusive as to how a face is scanned for different types of prosodic information and how looking patterns differ for prosody and segments.

The current study expands upon the previous studies in several novel ways. First, we conducted a more comprehensive investigation on the looking patterns when processing segments versus three types of prosodic information differing in the linguistic domains that the information is conveyed: contrastive word focus (more local), sentence intonation (more global), and intermediately: phrasal bracketing, where constituents of a sentences were prosodically grouped together in different ways. Second, we examined not only how long participants looked at one particular facial area but also investigated the overall scanning dynamics on the visual face by conducting a growth curve analysis on looking patterns over time (or how eye gaze patterns changed as the distribution of critical information

changed). Third, we focused on how perceivers adjusted their scanning strategies as task difficulty increased across conditions by varying the concordance of prosodic and segmental information.

In the current study, native English speakers scanned a talking face while looking for segmental and/or different prosodic information. We hypothesized more overall fixations to the mouth when identifying segmental differences and comparatively fewer fixations to the mouth when identifying prosodic differences. Second, we predicted that word-focus would show more and/or a faster shift of attention between the mouth and non-mouth areas when identifying prosodic information compared to when identifying segmental information, since visual cues from the both the lower part and upper part of the face contributed to the perception of prosody in these sentence types [9], with the upper face area being a bit less salient than the lower area [19]. However, we also predicted no such gaze shifting differences for the intonation type: Participants might fixate their eye gaze at just the mouth when searching for segmental information in this sentence type, and just the eyes when searching for intonational information [9]. Thus, we predicted less of a need to shift between the eyes and mouth in this condition [10]. Finally, we predict that the looking patterns for phrasal bracketing to be intermediate to those for word focus and intonation.

2. METHOD

2.1. Participants

Twenty-four native Canadian English speakers (5 males; 19 females), aged from 19 to 24 (mean = 21.2, SD = 2.3), were recruited. All participants had normal or corrected-to-normal vision, and none had hearing impairments or language-related pathologies.

2.2. Stimuli and Procedure

Each trial consisted of two auditory sentences and one silent video, which showed a brief clip of a face articulating an auditory sentence. A female, monolingual speaker of Canadian English recorded all auditory and visual stimuli, and these trials were arranged into *Segment*, *Prosody*, and *Both* conditions in word focus, phrasal bracketing and intonation sentences (see Figure 1 and Table 1). In experimental trials, auditory sentences differed from each other, and the silent video only matched one of them: In the *Prosody* condition, the two auditory sentences differed such that words were stressed using contrastive focus (word focus), or words were prosodically grouped together in distinct ways (phrasal bracketing), or one sentence was a question while the other was a statement (intonation), but both sentences had the same words; in the *Segment* condition, sentences differed in which words were used (having visually distinct consonants and vowels), but both sentences had the same patterns of prosody; in the *Both* condition, sentences difference in both prosody and segments. All stimuli were created in quadruplet groups, hereafter referred to as "topics." An example of stimuli from three topics are presented in Table 1.

During the experiment, each participant was calibrated in an Eyelink 1000 eye-tracker in binocular mode (SR Research) using a standard nine-point procedure on a 1280 * 1024 pixel screen. Afterwards, participants were informed that they would be presented with two auditory sentences followed by a single silent video that matched one (or both) of the audio-only stimuli in each trial. As shown in Figure 1, each trial began with presentation of both auditory sentences (with a 300 ms ISI). This was followed by a 500 ms pause, and then a fixation point turned blue before a video of a silent talking face was played. During video playback participants could select a keyboard button to indicate which of the audio recordings (or both) matched the video they saw (left arrow button indicating the first audio recording, right arrow button indicating the second and down arrow button indicating both), in which case, the video continued to play to the end, and then advanced automatically to the next trial. If no responses were made during the video, they were instructed to make a choice when the video ended, which then triggered the beginning of the next trial. Participants could thus respond anytime from the beginning to the end of stimulus presentation.

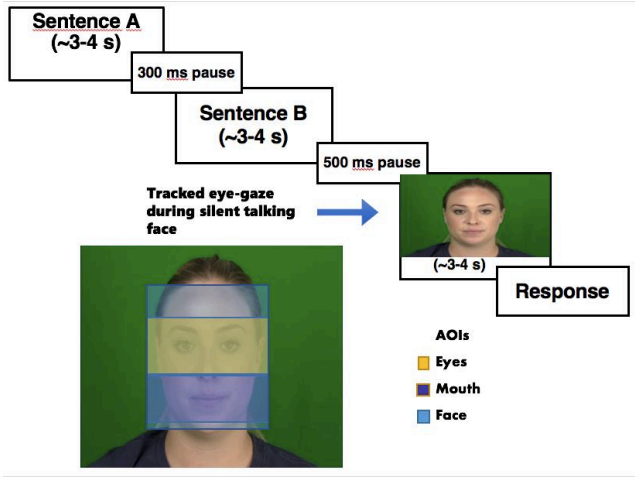


Figure 1: The structure of a trial. Participants first heard a sentence followed by another. Then, they saw a silent

talking face and began scanning the face for useful speech cues in order to make a decision about which sentence matched the video. The trial ended when the response was recorded, or when the video ended (whichever was later).

Word focus	Condition
No, **JESS** found her **PLAIN** dress for **KYLA'S** wedding.	Reference
No, Jess **FOUND** her plain **DRESS** for Kyla's **WEDDING**.	Prosody
No, **ANNE** found her **BLUE** dress for **JIMMY'S** wedding.	Segment
No, Anne **FOUND** her blue **DRESS** for Jimmy's **WEDDING**.	Both
Phrasal bracketing	
Vanessa and **[Alexander or Katrina]** will pick up the popcorn.	Reference
[Vanessa and Alexander] or Katrina will pick up the popcorn.	Prosody
Christina and **[Zachariah or Maria]** will pick up the popcorn.	Segment
[Christina and Zachariah] or Maria will pick up the popcorn.	Both
Intonation	
Sally will make a call **[before the salon closes?]**	Reference
Sally will make a call **[before the salon closes.]**	Prosody
Sally will make a call **[after she finishes lunch?]**	Segment
Sally will make a call **[after she finishes lunch.]**	Both

Table 1: Stimuli from an example topic in the three sentence types.

3. RESULTS

We first calculated the mean of raw proportion fixations to the mouth relative to the whole face (Prop) and then performed the empirical logit transformation (Elog) for the raw values through the equation below, in order to avoid the boundness issue of proportion measures, which violates the assumptions of statistical parametric models [21]:

$$Elog = Log(Prop + e/(1 - Prop + e))$$

The critical analysis was done in the time window from the onset of the video to the point when participants made their responses. This was done for only correct trials and averaged independently for all the three conditions in each sentence type.

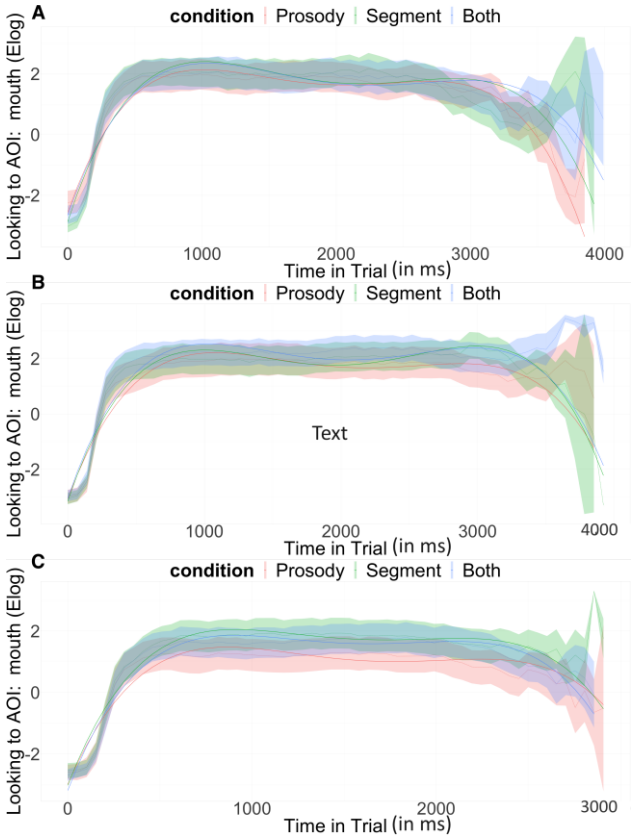


Figure 2: Elog values of fixations to the mouth (y -axis, averaged across participants and items), along the time in a trial (x-axis). Plot A, Plot B and Plot C reflect raw Elog values and model fits (solid lines) during the time window of interest in word-focus, phrasal bracketing and intonation type respectively. Note that model were calculated across 57 time bins.

Figure 2 exhibits the average Elog values of proportion fixations to the mouth during *Segment*, *Prosody* and *Both* conditions in each of the three different sentence types (word focus, phrasal bracketing, intonation). For statistical analysis, three growth curve analyses of the Elog values of proportion fixations to the mouth were performed within the above-mentioned time window for the three sentence types using R [22]. We used fourth-order orthogonal polynomial models, where Elog values were the dependent variable, and condition (*Segment*, *Prosody*, *Both*) was a fixed factor, with participant and topic (i.e. item) specified as random effects. The maximal model was performed first and if it did not converge, the random effects structure was simplified following Mirman (2014) (see https://dmirman.github.io/GCA/GCA2019.html). Models were fit using maximum-likelihood estimation [23] and estimates for the parameters used in significance tests for model fit components in the pair-wise comparisons were calculated from Satterthwaite's method.

Results for the word focus model showed that there was a significant effect of condition on the intercept (X^2 = 22.14, p < .001), linear (X^2 = 10.87, p = .004), quadratic (X^2 = 8.37, p = .02), cubic (X^2 = 48.09, p < .001) and quartic terms (X^2 = 10.07, p = .007). As shown in Figure 2a, there were greater overall fixations to the mouth in *Segment* condition compared to *Prosody* condition, which is also confirmed by the main effect of this difference on the intercept (estimate = .35, p < .001). More importantly, the significant effect of condition on the cubic term (estimate = 2.16, p < .001) revealed that there was a faster and more transient shift of looking to the mouth, then away from the mouth and back to the mouth again in *Prosody* condition than that in *Segment* condition, because the cubic (and quartic) term reflects the steepness of the curve around the inflection points at the extremities and a greater parameter estimate indicates a shallower curve [23]. When comparing the fixation curves in the *Both* condition, there was again a more acute change of looking in the *Prosody* condition than in the *Both* condition (quartic: estimate = 1.05, p = .004). However, the *Segment* condition showed the most abrupt shift of looking between the mouth and non-mouth area, relative to *Both* condition (quartic: estimate = .85, p = .01).

Results for the phrasal bracketing model showed that there was a significant effect of condition on the intercept (X^2 = 122.94, p < .001), linear (X^2 = 17.45, p < .001), quadratic (X^2 = 14.88, p < .001), and quartic terms (X^2 = 14.21, p < .001), indicating overall differences among these conditions. When comparing the differences between *Segment* and *Prosody* condition, again, there was a significant effect of condition on the cubic (estimate = -1.03, p = .03) and quartic terms (estimate = -1.27, p < .001), indicating that there was a more abrupt change of looking in *Prosody* condition than in *Segment* condition, between the mouth and non-mouth area. Likewise, when comparing the fixation curves in the *Both* condition versus the other two, there was a faster shift of looking in the *Both* condition than in the *Prosody* condition (quartic: estimate = -.85, p = .009) and a marginally faster shift of looking in *Segment* condition than *Both* condition (cubic: estimate = .87, p = .05).

Results for the intonation model showed that there was a significant effect of condition on the intercept (X^2 = 471.33, p < .001), linear (X^2 = 38.72, p < .001), quadratic (X^2 = 36.57, p < .001), and quartic terms (X^2 = 6.67, p = .04). When comparing the differences between *Segment* and *Prosody* condition, there was only a significant main effect of this difference on the intercept (estimate = .50, p < .001) and on the quadratic term (estimate = -1.58, p < .001), indicating more overall fixations to the mouth in *Segment* condition compared to *Prosody* condition, and an overall less steep fixation curve around the central inflection point, which reflects the participants initially attended to and eventually moved away from the mouth area more slowly in *Prosody* condition than in *Segment* condition. However, there were no significant cubic (estimate = -.32, p = .33) or quartic terms (estimate = -.41, p = .12) in the intonation type, reflecting a more parallel looking pattern between the *Segment* and *Prosody* conditions. When comparing the *Both* condition to the other two, there was also relatively more constant and stable looking, though less to the mouth, in *Prosody* condition than in *Both* condition (quartic: estimate = -.66, p = .01). However, there were no significant differences between the shapes of the curves around the inflection points in *Both* and *Segment* condition (quartic: estimate = -.26, p = .30).

4. DISCUSSION

The results showed a greater emphasis on looking to the mouth for decoding segmental information, relative to prosodic information, supporting our first hypothesis. However, this emphasis was also manifested in different ways across the three prosodic types. For word focus, there was faster shift between the mouth and non-mouth area in *Prosody* condition than in *Segment* condition, in line with our second hypothesis. For the phrasal bracketing type, differences across conditions may instead have indicated the necessity of searching the mouth for useful visual cues of critical information at different positions of the sentences as different individuals processed the syntactic ambiguity cued by the prosodic patterns, resulting in less “urgent” shift to the mouth (a curve with less steep curvatures). In the intonation sentence type, participants instead fixated facial areas where visual correlates were embedded for prosodic and segmental information respectively, without the need to shift their eye gaze back and forth, again corroborating the second hypothesis. Finally, the presence of both segmental and prosodic cues in the *Both* condition suggested that perceivers showed patterns of looking in between the *Prosody* and *Segment* conditions.

6. REFERENCES

[1] J. Navarra, H. H. Yeung, J. F. Werker, and S.

Soto-Faraco, “24 Multisensory and sensorimotor interactions in speech perception,” *The Handbook of Multisensory Processing.*, pp. 435, 2012.

[2] K. W. Grant and P.-F. Seitz, “The use of visible speech cues for improving auditory detection of spoken sentences,” *J. Acoust. Soc. Am.*, vol. 108, no. 3, p. 1197, 2000.

[3] M. E. Król, “Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study,” *PLoS One*, vol. 13, no. 3, pp. 1–15, 2018.

[4] W. H. Sumby and I. Pollack, “Visual Contribution to Speech Intelligibility in Noise,” *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[5] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, “Eye movement of perceivers during audiovisual speech perception,” *Percept. Psychophys.*, vol. 60, no. 6, pp. 926–940, 1998.

[6] M. Dohen and H. Lœvenbruck, “Interaction of Audition and Vision for the Perception of Prosodic Contrastive Focus,” *Lang. Speech*, vol. 52, no. 2–3, pp. 177–206, 2009.

[7] M. Swerts and E. Krahmer, “Facial expression and prosodic prominence: Effects of modality and facial area,” *J. Phon.*, vol. 36, no. 2, pp. 219–238, 2008.

[8] C. R. Lansing and G. W. McConkie, “Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences,” *Percept. Psychophys.*, vol. 65, no. 4, pp. 536–552, 2003.

[9] C. R. Lansing and G. W. McConkie, “Attention to Facial Regions in Segmental and Prosodic Visual Speech Perception Tasks,” *J. Speech, Lang. Hear. Res.*, vol. 42, pp. 526–539, 1999.

[10] E. Cvejic, J. Kim, C. Davis, and G. Gibert, “Prosody for the eyes: Quantifying visual prosody using guided principal component analysis,” *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010*, no. September, pp. 1433–1436, 2010.

[11] J. M. Foxton, L. D. Riviere, and P. Barone, “Cross-modal facilitation in speech prosody,” *Cognition*, vol. 115, no. 1, pp. 71–78, 2010.

[12] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, “Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception,” *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[13] S. Garg, G. Hamarneh, A. Jongman, J. A. Sereno, and Y. Wang, “Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories,” *Speech Commun.*, vol. 113, no. April, pp. 47–62, 2019.

[14] R. E. Ronquest, S. V. Levi, and D. B. Pisoni, “Language identification from visual-only speech signals,” *Attention, Perception and Psychophysics*, vol. 72, no. 6, pp. 1601–1613, 2010.

[15] P. Prieto, C. Puglesi, J. Borràs-Comes, E. Arroyo, and J. Blat, “Exploring the contribution of prosody and gesture to the perception of focus using an animated agent,” *J. Phon.*, vol. 49, pp. 41–54, 2015.

[16] L. G. Lusk and A. D. Mitchel, “Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation,” *Front. Psychol.*, vol. 7, no. February, pp. 1–11, 2016.

[17] D. J. Lewkowicz, M. Schmuckler, and V. Agrawal, “The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony,” *Cognition*, vol. 214, no. September, pp. 104743, 2021.

[18] C. Cavé *et al.*, “About the Relationship between Eyebrow Movements and F0 Variations,” *Proceeding Fourth Int. Conf. Spok. Lang. Process.*, vol. 4, pp. 2175–2178, 1996.

[19] R. Scarborough, P. Keating, S. L. Mattys, T. Cho, and A. Alwan, “Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English,” *Lang. Speech*, vol. 52, no. 2–3, pp. 135–175, 2009.

[20] M. Dohen, H. Lœvenbruck, M. Cathiard, and J. Schwartz, “Can we see Focus ? A Visual Perception Study of Contrastive Focus in French,” *Speech Prosody 2004, International Conference*, 2004.

[21] J. W. Dink and B. Ferguson, “Eyetracking R: An R Library for Eye-Tracking Data Analysis.,” 2015. *URL http://www.eyetrackingr.com*

[22] D. Mirman, J. A. Dixon, and J. S. Magnuson, “Statistical and computational models of the visual world paradigm: Growth curves and individual differences,” *J. Mem. Lang.*, vol. 59, no. 4, pp. 475–494, 2008.

[23] S. Kalénine, D. Mirman, E. L. Middleton, and L. J. Buxbaum, “Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 38, no. 5, pp. 1274–1295, 2012.

THE ROLE OF DURATION IN THE PERCEPTION OF CHECKED VERSUS UNCHECKED TONES IN TAIWANESE SOUTHREN MIN

Wei Zhang[1], Yu-An Lu[2]

[1]Departmant of Linguistics, McGill University, Canada.
[2]Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Taiwan
wei.zhang16@mail.mcgill.ca; yuanlu@nycu.edu.tw

ABSTRACT

Taiwanese Southern Min has checked verses unchecked tonal contrast the perception of which has received limited investigation. This study examined the role of duration in perceiving this contrast in TSM, for both the mid and high registers. Identification experiments were carried out on stimuli sampled from 'All-cue' Continuum where all cues were proportionally manipulated and from 'Duration' Continuum where only the duration cue was manipulated. For the mid register, similar categorization patterns were found for Duration and All-cue Continua, indicating that duration is indeed used reliably to categorize checked vs. unchecked tones. Additionally, we observed a suppression effect of unchecked tone identification on tokens with glottalization, an accompanying cue to checked tones. For the high register, however, no categorical response was found on the Duration Continuum, suggesting that the reliance on duration cue is subtle and that glottalization cues guided the perception of checked vs. unchecked tones for the high register.

Keywords: Taiwanese Southern Min, checked tone perception, duration cue, glottalization cue

1. INTRODUCTION

Taiwanese Southern Min (TSM), also known as Taiwanese Hokkien, is the second popular language spoken in Taiwan [1]. TSM has a rich tonal system including seven lexical tones, two of which are checked tones occurring in CV[p, t, k, ʔ] syllables where the final stops are unreleased [2]. Unchecked tones, on the other hand, occur in CV(N) syllables. An example of the tonal system [3] is shown in Fig. 1. As can be seen, the checked tones (T5, T3, the numbers in the tone names indicate the pitch height [4], same below), are shorter than the unchecked tones (T55, T13, T51, T31, T33).

In recent years, there has been growing attention to the sandhi schemes of TSM [3], [5], [6], and the acoustic features of checked and unchecked tones [7]–[9]. However, to the best of our knowledge, investigations on how listeners perceive the checked vs. unchecked tone pairs in TSM has been limited. There are two pairs of checked vs. unchecked tones for mid and high registers, respectively: T3 vs. T33 and T5 vs. T55. Phonologically, the distinction between checked and unchecked tones are the final unreleased stop. However, as shown in Fig. 1, acoustically, the duration difference is also very salient: unchecked tones are about twice as long as checked tones. Furthermore, previous studies using electroglottographic and acoustic data have shown that, while final stop closures may be absent, energy dipping and irregular glottal vibration can still be observed on the preceding vowels [8]. [3] also reported cases of coda deletion, in which over 80% of /ʔ/ codas were deleted, followed by /t/, /k/ and then by /p/, suggesting a possible increasing role of duration in checked tone perception. Similar to how voiceless stop codas in English which are often cued by a shorter preceding vowel and the presence of glottalization [10], the checked tones in TSM, which end with voiceless stops, are also shorter and often accompanied by glottalization. This study is aimed to investigate to what extent listeners make use of duration and spectral cues informed by the final stop in perceiving the TSM checked vs. unchecked contrast for both registers.

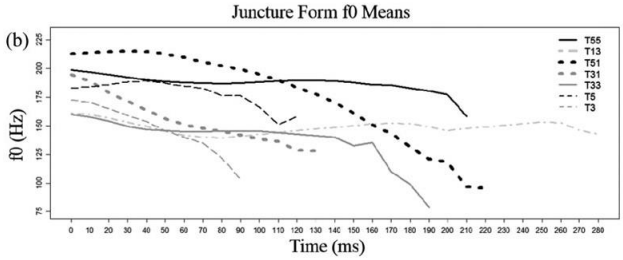


Figure 1: Tonal system of TSM (From Pan (2017: 38), the speaker was a female speaker over 40 years old. The numbers in the tone names indicate pitch height.)

There are relatively fewer investigations of the perception (than production) of checked tones or syllables (see [11] for a review). From the few investigations of checked tone perception in other languages/Chinese dialects, *F0*, *duration* and *voice quality* (such as *glottalization*) are the most important cues reported. Relative importance of the three cues varies across languages. In Burmese, voice quality and duration might be the distinguishing cues for the checked tone from the rest of the tones [12]. For both

Southern and Northern Vietnamese, [13] found that final-glottalization was a more important cue than F0 to the checked tone perception. For White Hmong, [14] found that listeners rely more on F0 and duration than glottalization for identifying checked tones. Additionally, for Xiapu Min, it was found that glottalization is less important than duration and F0 [11], and duration seems to be the most important cue [15] in the checked tone perception.

In terms of the checked vs. unchecked tonal contrast in TSM, as shown in Fig. 1, both the checked and unchecked tones have a falling contour acoustically. Additionally, [16] showed that when the durations are normalized, checked tones and unchecked tones exhibit very similar F0 contours, suggesting that F0 may not be an important cue to the checked vs. unchecked tonal contrast. With a special interest in duration, this study investigates the importance of the duration cue in perceiving checked vs. unchecked tones in TSM. We selected the CV[ʔ] structure for the checked tones based on the aforementioned finding that [ʔ] was more frequently deleted than [p, t, k] as the final coda of TSM checked tones. This may suggest that duration plays a relatively greater role in checked tones with a [ʔ] coda than in those with the other three codas.

2. METHOD

2.1. Material

The segment was controlled for both checked and unchecked tones as well as for both registers. Specifically, the syllable /la/ was used for T33 (‘扒’, meaning ‘stir’) and T55 (‘拉’, meaning ‘pull’) while /laʔ/ was used for T3 (‘垃’, meaning ‘garbage’) and T5 (‘蠟’, meaning ‘wax’). A female speaker of TSM recorded each of the four tones/syllables naturally, in a sound-attenuated booth. From these recordings, one sample of each tone was selected based on less perturbation and a smoother amplitude envelope. These selected samples were then used to generate checked-unchecked tonal continua.

To evaluate the importance of duration in checked tone perception, we created two types of tonal continua. Type 1 is created by manipulating only the duration of the syllable between the checked and unchecked tones, named Duration Continuum hereafter. The Duration Continuum was created in two ways using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat [17]): (1) by increasing the duration based on the checked tone sample, and (2) by decreasing the duration based on the unchecked tone sample. The main difference between the two methods was whether vowel glottalization naturally contained in the checked stimuli was preserved or not. Type 2 is created by using the checked and unchecked tones as endpoints, hence every contributing cue should be proportionally manipulated. This is called All-cue Continuum hereafter. Tandem-Straight [18], a speech analysis, modification and resynthesis framework, was used to create the All-cue Continuum. The spectrograms of the endpoint samples are shown in Fig. 2. To assess the extent to which TSM listeners rely on *duration* in categorizing checked and unchecked tones, we compared these listeners’ categorization of the same tonal pair in the two types of continua, one only differs in duration while the other in all relevant acoustic cues.

Altogether there were six continua used in this study: Duration Continuum based on T3, Duration Continuum based on T33, All-cue Continuum of register 3, and the same three continua for register 5.

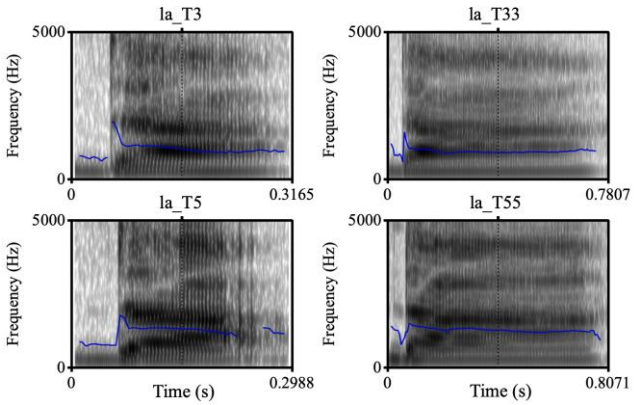


Figure 2: Spectrograms of base samples (top left: T3; top right: T33; bottom left: T5; bottom right: T55).

There were seven steps in each continuum. The duration of the checked and unchecked tone endpoints are 160 ms and 340 ms (~30 ms difference each step), respectively. The duration values were selected according to the authors’ previous production results of TSM tones and adjustment was made for balancing the categorical boundary to the middle steps.

2.2. Procedure and Participants

We recruited 17 TSM speakers (10M, 7F, $M = 23$ yrs, sd = 1.6) at Experimental Phonology Lab, National Yang Ming Chiao Tung University for the experiment. Two-alternative forced choice paradigm was used. Since there were twice as many stimuli on the Duration Continua as on the All-cue Continua, we doubled the number of stimuli on the All-cue Continua. Hence there were 8 continua in each block and the block repeated four times. Participants were allowed to take a break after each block. Altogether there were 224 (8[continua]*7[steps]*4[blocks]) trials in the experiment. A short instruction with audio

examples and 6 trials were provided before the experiment for practice.

Before the main task, the participants were asked to complete a questionnaire on their TSM, such as age of acquisition, frequency of use, listening level and speaking level (assessed on a 1-7 scale).

2.3. Statistical analysis

We excluded participants who self-reported their TSM listening level as 3 or lower (4 participants). For the rest of the data, statistical analysis was carried out with Bayesian mixed effects logistic regression, implemented using brms [19] in R [20]. The dependent variable was the response of the tone type (checked T3 and T5 were mapped to 0 and unchecked T33 and T55 were mapped to 1). Three fixed predictors were used: *step* index of the continuum (scaled and centered), manipulation *method* (treatment coded: All-cue Continuum was taken as the reference level), and *register* (register 3 as the reference level). The interactions between any two of them were included. By-participant random effects were specified as random intercept with the three fixed effects as random slopes.

3. RESULTS

The categorization results for each continuum are visualized in Fig. 3. Panel A and B of Fig. 3 show the patterns of responses for register 3 and register 5, respectively. For register 3, the results of the Duration Continuum show that increasing the duration between the checked and unchecked tones (moving from left to right on the x-axis) resulted in a gradual shift towards the unchecked tone, similar to the pattern observed on the All-cue Continuum, in which all cues were proportionally manipulated. In terms of register 5, however, no categorical response pattern was observed in the two Duration Continua. This suggests that, for the T3-T33 contrast but not for the T5-T55 contrast, duration plays a relatively larger role in distinguishing between checked and unchecked tones.

Results of the mixed effect logistic Bayesian model are presented in Table 1. *Rhat* values of all effects are 1, indicating the model does not show convergence issues. The probability of direction (*pd*) is computed using the R package *bayestestR* [21]. A *pd* value larger than 95% indicates a credible effect.

In Table 1, larger step index credibly increases the unchecked tone responses (β = 4.25, *pd* = 100%), consistent with the expectation that unchecked tones are longer, when other predictors hold constant. Compared to the All-cue Continua, the Duration Continua based on checked tones credibly decrease the unchecked tone response (β = -0.74, *pd* = 95%), whereas the Duration Continua based on unchecked tones do not show credible effect (β = 0.53, *pd* = 91%). On average, register 5 has credibly less unchecked tone responses (β = -0.68, *pd* = 95%).

The interaction between step and method, and the interaction between method and register are all credible. Comparison results of '*step:method*', using the *emmeans* package [22], revealed that step contributes a credible larger effect to the unchecked tone response in All-cue Continua than in Duration Continua based on checked tones (β = 1.3; *hpd*: [0.4, 2]), and a credible smaller effect to the unchecked tone response in All-cue Continua than in Duration Continua based on unchecked tones (β = -2.6; *hpd*: [-3.4, -1.8]).[1]

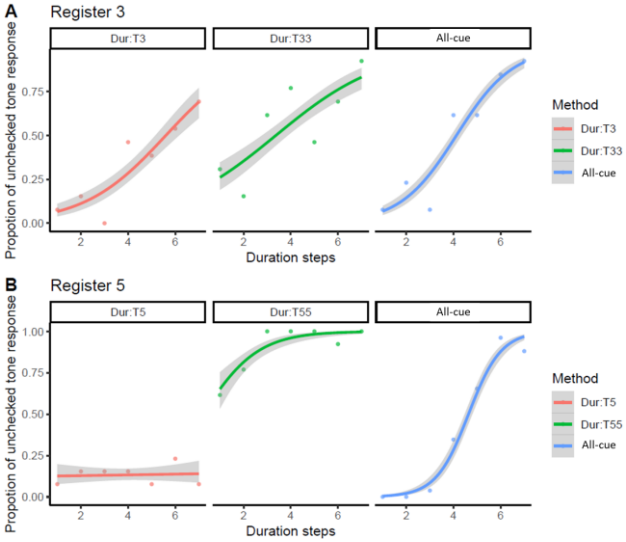


Figure 3: Tonal categorication results, grouped by register (top: Register 3; bottom: Register 5) and colored by manipulation method (left: Duration Continua based on checked tones; middle: Duration Continua based on unchecked tones; right: All-cue Continua). The abbreviation, for example, 'Dur:T3' indicates the Duration continuum based on T3 sample.

Comparison results of '*method:register*' showed firstly that there is no credible difference between Duration Continuum based on T3 and All-cue Continuum of register 3, or between Duration Continuum based on T33 and All-cue Continuum of register 3, indicating that duration manipulation elicits comparable response as manipulating all cues in perception for the mid register. However, the effect of Duration Continuum based on T3 was credibly smaller than that of Duration Continuum based on T33 (β = -1.3; *hpd*: [-2.6, 0]), suggesting that the glottalization, which exists in the T3 base but not in the T33 base, plays a role in perception and reduces the unchecked tone response. Secondly, for register 5, there is a credible difference between any two of the three continua. Specifically, All-cue Continuum of register 5 shows a credible larger effect on the unchecked tone responses than the Duration Continuum based on T5 (β = 1.7; *hpd*: [0.8, 2.5]), and

a credible smaller effect than the Duration Continuum based on T55 (β = -4.6; *hpd*: [-5.6, -3.7]). The effects indicate that, in contrast to the mid register, duration manipulation does not elicit any comparable response as manipulating all cues for the high register. Thirdly, when comparing across registers, the Duration Continuum based on T3 has a credible larger effect than that based on T5 (β = 1.7; *hpd*: [0.7, 2.6]). Specifically, the Duration Continuum based on T33 has a credible smaller effect than that based on T55 on the response variable (β = -3.4; *hpd*: [-4.5, -2.4]). These results suggest again that glottalization plays a predominant role in checked tone perception for the high register.

	β	l-95% CI	u-95% CI	pd
Intercept	-0.13	-1.04	0.78	62 %
step	4.25	2.2	5.89	**100%**
methodChecked	-0.74	-1.52	0.18	**95%**
methodUnchecked	0.53	-0.28	1.29	91%
register	-0.68	-1.54	0.18	**95%**
step:methodChecked	-3.33	-3.98	-2.68	**100%**
step:methodUnchecked	-2.38	-3.08	-1.67	**100%**
step:register	0.17	-0.41	0.76	72%
methodChecked:register	-0.98	-1.53	-0.43	**100%**
methodUnchecked:register	4.1	3.42	4.82	**100%**

Table 1: Results of the mixed-effect logistic Bayesian model. Probabilities of direction (*pd*s) of credible effects are bolded.

4. DISCUSSION

This study investigated the perception of the checked vs. unchecked tonal contrast in TSM. As the spectral cue, glottalization, is disappearing with the deletion of syllable coda in checked tones, this study examined whether duration shows an increased cue weight. Through identification experiments on All-cue and Duration Continua for both registers, we found that for the mid tones, duration is indeed used reliably to identify checked vs. unchecked tones based on the linear function of their responses as the duration step increased. The tokens without glottalization were categorized similarly as the all-cue tokens. We did see a suppression effect of unchecked tone identification on tokens with glottalization cues. The reliance on duration cue, however, was subtle for the high tones. It appears that glottalization cues guided the perception of checkedness for the high tones.

We provide some possible sources of why duration plays different roles in perceiving the checked tone for the two registers. First, literature on TSM revealed that coda [ʔ] was deleted more often in the high checked tone than in the mid checked tone [23-24, cited after 25], which could result in different weightings of glottalization and duration for the two registers in general. Second, [25] found that in TSM when the deletion of [ʔ] happened, the preceding vowel was lengthened for the high tones, causing a possible disconnection between the high checked tone and shorter duration. For the mid tones, however, the duration remained short after the deletion of the coda, strengthening the connection between the mid checked tone and shorter duration. That is, the duration cue may not be reliable in categorizing high checked vs. unchecked tones but remains reliable for the mid register. Third, when creating the stimuli, we used separate base samples for the two registers, hence the degree of glottalization was not strictly controlled across the registers. As shown in Fig. 2, there seems to be more glottalization for 'la_T5' than for 'la_T3', which might bring about larger glottalization effect in the Duration Continuum based on T5 than on T3. However, this difference could not be the only reason accounting for the asymmetrical results between Duration continua based on T33 and T55, since both continua did not contain glottalization. What might be a reason instead is the psycho-acoustic effect: a sound with higher pitch is perceived longer than with lower pitch when their physical durations are equal [26-27]. Hence, steps in Duration Continuum based on T55 can be perceived longer (than steps in the continuum based on T33), which may increase the unchecked tone responses. As shown in Fig. 3, as duration decreases from step 4 to 1, there are increasing checked tone responses. If even shorter durations were tested, the categorical pattern might emerge, and this direction awaits further investigation.

The results of this study may have implications for the tone merger in TSM. The two checked tones are merging and the final codas are undergoing loss [1, 3]. The confirmed role of duration in distinguishing between the checked and unchecked tones, along with the ongoing trend of stop coda deletion [3], suggests a possible outcome of tone merging in TSM: a long vs. short tonal contrast. The deletion of stop codas suggests a decreasing phonological awareness of checkedness, while the confirmed role of duration in this study indicates awareness of quantity. It is possible that the contrast between checked and unchecked tones could be phonologized into a long vs. short contrast in TSM.

One limitation of this study is the relatively small sample size and the variability in the degree of glottalization across the different registers. Additionally, further investigation is needed to explore the perception of checkedness in the high register under shorter duration ranges, and the individual-level perception-production link on checkedness.

5. ACKNOWLEDGEMENTS

We would like to thank Yin-Ching Chang for her help on collecting data. This work was supported by Ministry of Science and Technology of Taiwan (MOST110-2628-H-A49-001-MY2) to Yu-An Lu.

6. REFERENCES

[1] U. Ang, "The distribution and regionalization of varieties in Taiwan," *Language and Linguistics*, vol. 14, no. 2, p. 315, 2013.

[2] J. Norman, *Chinese*. Cambridge University Press, 1988.

[3] H. Pan, "Glottalization of Taiwan Min checked tones," *Journal of the International Phonetic Association*, vol. 47, no. 1, pp. 37–63, Apr. 2017, doi: 10.1017/S0025100316000281.

[4] Y.-R. Chao, "A system of tone letters," *Le maître phonétique*, 1930.

[5] C.-H. Kuo, "Perception and acoustic correlates of the Taiwanese tone sandhi group," PhD Thesis, UCLA, 2013.

[6] Y.-F. Chien and A. Jongman, "Tonal Neutralization of Taiwanese Checked and Smooth Syllables: An Acoustic Study," *Lang Speech*, vol. 62, no. 3, pp. 452–474, Sep. 2019, doi: 10.1177/0023830918785663.

[7] H. Pan, M.-H. Chen, and S.-R. Lyu, "Electroglottograph and acoustic cues for phonation contrasts in Taiwan Min falling tones," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] H. Pan, H. Huang, and S. Lyu, "Coda Stop and Taiwan Min Checked Tone Sound Changes.," in *INTERSPEECH*, 2016, pp. 1011–1015.

[9] H. Pan, "Voice quality of falling tones in Taiwan Min," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[10] J. Penney, F. Cox, and A. Szakay, "Weighting of Coda Voicing Cues: Glottalisation and Vowel Duration.," in *INTERSPEECH*, 2018, pp. 1422–1426.

[11] Y. Chai, "Phonetics and Phonology of Checked Phonation, Syllables, and Tones," Ph.D., University of California, San Diego, United States -- California, 2022. Accessed: Sep. 30, 2022. [Online]. Available: https://www.proquest.com/docview/2714115642/abstract/26EACB3B47A6499FPQ/1

[12] J. F. Gruber, *An articulatory, acoustic, and auditory study of Burmese tone*. Georgetown University, 2011.

[13] M. Brunelle, "Tone perception in northern and southern Vietnamese," *Journal of Phonetics*, vol. 37, no. 1, pp. 79–96, 2009.

[14] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1078–1089, 2013.

[15] Y. Chai, "Perception of checked tones in Xiapu Min," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. A309–A310, Oct. 2021, doi: 10.1121/10.0008390.

[16] M.-H. Lin, "Pitch and duration reflexes in Taiwanese Southern Min tones across generations.," M.A., National Yang Ming Chiao Tung University, Hsinchu, 2022.

[17] P. Boersma, "Praat: doing phonetics by computer [Computer program]. Version 6.1.16," *http://www.praat. org/*, 2020.

[18] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[19] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of statistical software*, vol. 80, no. 1, pp. 1–28, 2017.

[20] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria, 2013.

[21] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework," *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019.

[22] H. Singmann, J. Love, P. Buerkner, and M. Herv, "Package 'emmeans,'" *CRAN R Project. Available online: https://cran. r-project. org/web/packages/emmeans/emmeans. pdf (accessed on 10 April 2020)*, 2019.

[23] U. Ang, "The motivation and direction of sound change: On the competition of Minnan dialects Chang-chou and Chuan-chou, and the emergence of General Taiwanese," *Unpublished PhD dissertation, National Tsing Hua University, Taiwan*, 2003.

[24] S. Chen, "New Sound Variation in Taiwan Southern Min: Vowel Systems and the Lower Register Entering Tone in Taipei, Changhua, and Tainan," *Language and Linguistics*, vol. 11, no. 2, pp. 425–468, 2010.

[25] H. Pan and S. Lyu, "Taiwan Min Nan (Taiwanese) Checked Tones Sound Change," in *Interspeech 2021*, Aug. 2021, pp. 2641–2645. doi: 10.21437/Interspeech.2021-672.

[26] C. Gussenhoven and W. Zhou, "Revisiting pitch slope and height effects on perceived duration," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1365–1369.

[27] Y.-A. Lu and S.-I. Lee-Kim, "The effect of linguistic experience on perceived vowel duration: Evidence from Taiwan Mandarin speakers," *Journal of Phonetics*, vol. 86, p. 101049, May 2021, doi: 10.1016/j.wocn.2021.101049.

[1] (HPD interval probability: 0.95. An HPD interval excluding 0 is assumed to indicate a credible effect.)

PERCEPTION OF JAPANESE PITCH ACCENT BY TYPICALLY DEVELOPING CHILDREN AND CHILDREN WITH AUTISM SPECTRUM DISORDERS

Yasuaki Shinohara[1-3], Mariko Uchida[4,5], Tomoko Matsui[4]

[1]Waseda University, [2]CUNY Graduate Center, [3]University of Delaware, [4]Chuo University, [5] Tokyo Gakugei University
y.shinohara@waseda.jp

ABSTRACT

This study examined how typically developing (TD) children and children with autism spectrum disorders (ASD) identified Japanese pitch-accent minimal-pair words in three stimulus conditions: natural recordings (NatRec), sine-wave speech (SWS), and noise-vocoded sine-wave speech (NzVocSWS). It was hypothesized that TD children would show a significant decrease in identification accuracy from the NatRec condition to degraded speech (SWS and NzVocSWS) conditions, and a significant increase from the SWS to NzVocSWS condition, consistent with the results in a previous study testing Japanese-speaking adults. On the contrary, children with ASD were expected to show a different identification pattern because of their difficulty in phonological processing and detecting subtle differences in duration and intensity cues. The results demonstrated that both groups showed a similar pattern, although the identification accuracy of children with ASD was significantly lower than that of TD children across the three stimulus conditions.

Keywords: speech perception, pitch accent, autism spectrum disorders, noise-vocoded sine-wave speech

1. INTRODUCTION

Speech sounds carry multiple acoustic cues signaling phonological contrasts [1–4], and degraded speech is often used to manipulate audible acoustic cues. Sine-wave speech (SWS) is a type of synthetic speech, consisting of multiple time-varying sinusoids that replicate the first three or four formants of natural speech with its amplitude pattern [5]. People understand words and sentences from SWS [5–7], but since there is no fundamental frequency (F0) information in SWS, Japanese speakers cannot identify pitch-accent minimal-pair words correctly [8]. Instead of F0, they rely on the first formant frequency (F1) to identify pitch accents, resulting in poor identification [9]. However, when SWS is noise-vocoded, individual formants become less audible, and Japanese speakers use other acoustic cues (e.g., duration and intensity), leading to better identification [9]. In the present study, pitch-accent perception of children with autism spectrum disorders (ASD) was examined with the hypothesis that their pitch-accent perception would be different than typically developing (TD) children.

ASD is often described as a neurodevelopmental disorder affecting social interaction and speech communication [10]. People with ASD often have enhanced pitch perception, but their outperformance is observed on a low-level perception (e.g., auditory discrimination), and not on a higher-level (e.g., categorical) perception [11–13]. This means that people with ASD are auditorily sensitive, but experience difficulty in phonological processing. This claim has been supported by both behavioral and neurophysiological experiments [12].

The Japanese language has lexical contrasts based on a high (H) vs. low (L) bitonal pitch accent. For example, [nizi] (HL) with a relatively high-pitched first mora [ni] and a relatively low-pitched second mora [zi] means *two o'clock*, whereas [nizi] (LH) with a low-pitched first mora and a high-pitched second mora means *rainbow*. The primary acoustic cue for pitch-accent contrast is F0, and Japanese speakers use the location of the F0 peak to identify accented morae [1]. Although there are many debates about whether duration and intensity are the secondary acoustic cues for pitch-accent perception [2–4], a previous study showed that Japanese speakers tend to rely on the duration and intensity of vowels to identify pitch-accent minimal-pair words when F0 is not available [9].

This study examined the identification accuracy of Japanese pitch-accent minimal-pair words in three testing conditions: natural recordings (NatRec), SWS, and noise-vocoded sine-wave speech (NzVocSWS). In NatRec, where all relevant acoustic cues, such as F0, duration, and intensity, are available, it was hypothesized that the identification accuracy would be at ceiling for TD children. However, in SWS, where F0 is not available, TD children may rely on F1 for pitch-accent perception, as Japanese-speaking adults do [9]. Consequently, TD children would not

be able to identify pitch accents correctly. In NzVocSWS, where the effect of quasi-periodicity (i.e., F1) is reduced, TD children were expected to increase their identification accuracy compared to the SWS condition. Since each formant is less audible due to the noise-vocoding effect, they may rely on other acoustic cues (e.g., duration and intensity), resulting in higher identification accuracy [9]. Conversely, it was hypothesized that children with ASD would not increase their identification accuracy from SWS to NzVocSWS due to their difficulty in phonological processing and detecting subtle differences in both duration and intensity cues [12], [14–19]. For children with ASD, identification accuracy was expected to be at a chance level in both the SWS and NzVocSWS conditions, and not as high as that of TD children in the NatRec condition. Based on these hypotheses, we predicted a significant interaction between the group and condition factors.

2. METHOD

2.1. Participants

Table 1 displays the participants' information. Sixteen children participated in the experiment. They were all Tokyo Japanese speakers who had lived most of their lives in Tokyo [20]. Eight of the 16 participants were TD children, and the remaining eight were diagnosed by pediatricians as having ASD. The two groups (TD and ASD) were matched in chronological age, but significantly differed in the scores of the Parent-Interview ASD Rating Scales – Text Revision (PARS-TR) [21]. The PARS-TR is an interview-based rating scale widely used in Japan to assess ASD symptoms. The symptoms were assessed during the preschool period and at the time of the study.

	TD (n = 8) Mean (*SD*)	ASD (n = 8) Mean (*SD*)	Group differences
Sex	6 girls, 2 boys	2 girls, 6 boys	
Age (years)	9.64 (1.74)	9.48 (1.74)	n.s. ($p > .05$)
PARS-TR (present)	4.00 (6.76)	20.88 (9.72)	$p < .01$
PARS-TR (preschool)	7.00 (10.62)	30.88 (10.87)	$p < .01$
PVT-R score	63.38 (15.58)	46.50 (23.81)	n.s. ($p > .05$)
PVT-R (SS)	13.63 (3.93)	8.88 (6.22)	$p = .089$

Note: Independent sample t-test and Wilcoxon rank sum test were used for the group difference analyses.

Table 1: Participants' information of typically developing (TD) children and children with autism spectrum disorders (ASD).

There was no significant difference in the Peabody Picture Vocabulary Test of receptive vocabulary (Japanese version; PVT-R score) between the two groups, but the difference in the PVT-R evaluation score (SS) was marginally significant. The PVT-R score is often used to assess verbal intelligence quotient (IQ) [22], and the PVT-R (SS) is often used to assess language delay [23].

2.2. Stimuli

Three types of stimuli were used in our experiment: NatRec, SWS, and NzVocSWS. Figure 1 displays waveforms and narrowband spectrograms of the word [ame] with the HL and LH pitch accents in each stimulus type.

2.1.1. Natural recordings (NatRec)

Two Japanese pitch-accent minimal pairs ([ame] with HL meaning *rain* vs. LH meaning *candy* and [niꜜzi] with HL *two o'clock* vs. LH *rainbow*) were recorded by six Tokyo Japanese speakers (three women and three men), using a Rode NT2-A microphone with 44,100 16-bit samples per second. The intensity was normalized across all tokens based on the root-mean-square method. Tokens from two speakers (two women) were used only for practice sessions, and the tokens from the remaining four speakers were used for the identification test. Sixteen tokens (4 minimal-pair words × 4 speakers) were used in the identification test of the NatRec condition.

2.1.2. Sine-wave speech (SWS)

The duration of the NatRec stimuli was averaged between minimal-pair words produced by each speaker. All tokens were then transformed to SWS with a Praat script written by Darwin [24]. The first three formants (F1, F2, and F3) and their amplitude were tracked every 10 ms with the respective gender setting. Other parameters were set as the default (e.g., amplitude tracks were low-pass filtered at 50 Hz). The three individual sinusoids were combined to construct the SWS version of the 16 tokens (4 minimal-pair words × 4 speakers).

2.1.3. Noise-vocoded sine-wave speech (NzVocSWS)

Finally, the SWS stimuli went through a peak-picking noise vocoder using a Praat script written by Winn [25]. The frequency range was set from 70 Hz to 10 kHz. Spectral analysis and synthesis were conducted using 33 filters, with the width of the synthesis filter being 6 dB per mm. As each formant was less audible, it was expected that TD Japanese speakers would not rely on F1 but would instead use duration and

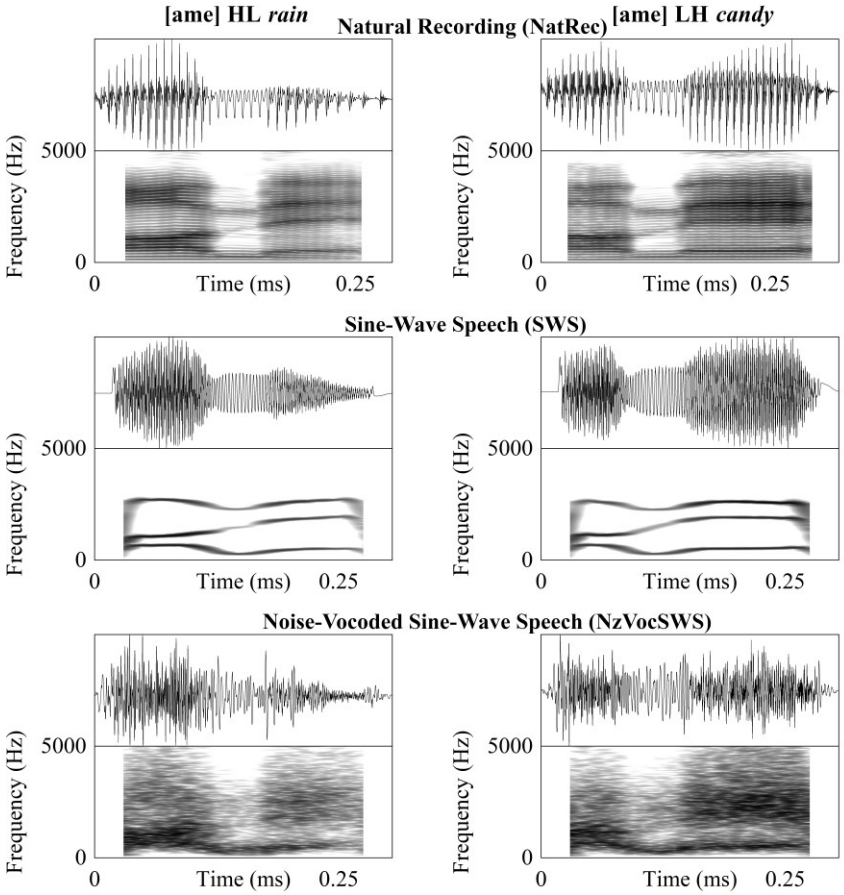


Figure 1: Waveforms and narrowband spectrograms of the word [ame] with the HL (high-low) and LH (low-high) pitch accents in three stimulus types: natural recordings (NatRec), sine-wave speech (SWS), and noise-vocoded sine-wave speech (NzVocSWS).

intensity for pitch-accent perception in the NzVocSWS condition.

It should be noted that when a noise vocoder with square-shaped synthesis filters was used, Japanese speakers did not show a significant difference in their pitch-accent identification accuracy between the SWS and NzVocSWS conditions [8].

2.3. Procedure

Participants took a Japanese pitch-accent identification test under three conditions (NatRec, SWS, NzVocSWS). They heard a word and saw a minimal pair displayed on the screen with both *kanji* (a Japanese writing system with Chinese characters) and *hiragana* (Japanese moraic orthography) (e.g., 雨(あめ), 飴(あめ)), then clicked on the word they thought they had heard. Half of the participants took the test in the NatRec, SWS, and NzVocSWS order, and the other half took it in the NatRec, NzVocSWS, and SWS order. Before each condition, participants had a practice session with two tokens. They received no feedback during the test and heard eight tokens for each minimal pair of [ame] and [niʑi]. Each participant completed 48 trials (8 tokens × 2 minimal pairs × 3 conditions).

3. RESULTS

Figure 2 displays the identification accuracy of pitch-accent contrast under three stimulus conditions (NatRec, SWS, NzVocSWS) by TD children and children with ASD. A logistic mixed effects model based on correct and incorrect binomial responses was used for the statistical analysis. The fixed factors were group (TD, ASD), condition (NatRec, SWS, NzVocSWS), and their interaction. Orthogonal contrasts were set for these categorical variables. By-participant random intercepts were also included in the model.

The logistic mixed effects model demonstrated a significant effect of group (TD vs. ASD), $\beta = 0.32$, $SE = 0.15$, $z = 2.11$, $p = .035$, suggesting that the TD children had higher identification accuracy across all three conditions than the children with ASD. Significant effects were also found for condition. The identification accuracy was significantly higher in NatRec than in degraded speech (i.e., SWS and NzVocSWS) conditions, $\beta = 0.70$, $SE = 0.08$, $z = 8.82$, $p < .001$, and accuracy in the NzVocSWS condition was significantly higher than that in SWS, $\beta = -0.28$, $SE = 0.09$, $z = -3.05$, $p < .01$. However, the interaction

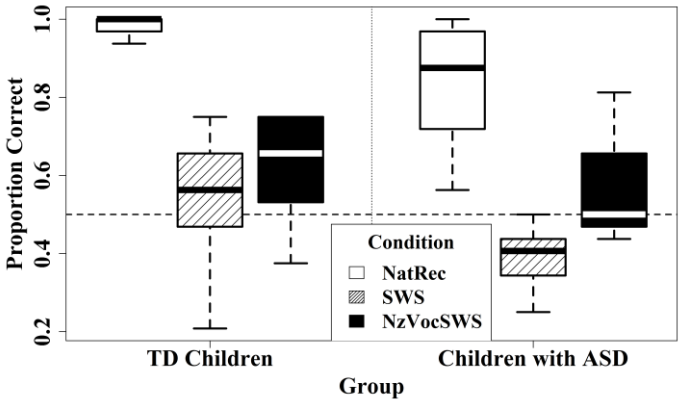


Figure 2: Identification accuracy of two Japanese pitch-accent minimal pairs ([ame] HL *rain* vs. LH *candy*, [niꜜzi] HL *two o'clock* vs. LH *rainbow*) under three stimulus conditions (NatRec, SWS, NzVocSWS) by typically developing (TD) children and children with autism spectrum disorders (ASD). The horizontal dashed line represents the chance level of correct proportion.

between group and condition was not significant in any contrast ($p > .05$).

4. DISCUSSION

This study examined identification accuracy of Japanese pitch-accent minimal-pair words by TD children and children with ASD under three stimulus conditions (NatRec, SWS, NzVocSWS). For TD children, it was hypothesized that pitch-accent identification accuracy would decrease from the NatRec to degraded speech (SWS and NzVocSWS) conditions, but significantly increase from the SWS to NzVocSWS condition. As Japanese-speaking adults did in a previous study [9], TD children may rely on F1 for pitch accent perception in SWS where F0 information is not available, resulting in poor identification accuracy. However, because the effect of misleading voice pitch contour (i.e., F1) may be reduced due to less audible formants in NzVocSWS, we predicted that their pitch accent identification accuracy would increase, with more reliance on other acoustic cues (e.g., duration and intensity). In contrast to TD children, children with ASD were not expected to significantly increase their identification accuracy even after the SWS was noise-vocoded, due to their difficulty in phonological processing and detecting subtle differences in duration and intensity cues [14–19]. Because of these difficulties, identification accuracy in the NatRec condition may also not be as high for children with ASD, predicting a significant interaction between the group and condition factors.

The results demonstrated that, although identification accuracy was lower for the ASD group than the TD group across all three conditions, no significant interaction was found—both groups showed similar pitch-accent perception patterns. Their identification accuracy in the degraded speech (SWS and NzVocSWS) conditions was lower than that in the NatRec condition, and accuracy in the NzVocSWS condition was higher than in the SWS condition. These results do not fully support our hypotheses, but are consistent with the results of tests with Japanese-speaking adults. Since this study did not examine how each acoustic cue is used by each group, further research is necessary.

There are multiple limitations that need to be addressed in a future study. First, language ability was not controlled between the TD and ASD groups. There were no significant differences in the PVT-R (SS) or PVT-R score between the TD and ASD groups, which suggests that the ASD group had neither significant language delay nor significantly lower verbal IQ compared to the TD group. Since autism spectrum disorder is an umbrella term including a wide range of disorders, more-detailed language-ability backgrounds should be analyzed. Second, since these are preliminary results from only eight participants in each group, there may have been a lack of statistical power to show a significant interaction between group and condition—it is necessary to increase the number of participants. Finally, this study only examined pitch-accent identification, not auditory discrimination. It is also important to test auditory sensitivity differences between the two groups. Identifying difficulties in pitch-accent perception in further studies would help clinical diagnoses and education practices for these populations.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers 19H01753 and 23K00490. We thank our participants and research assistants (Ms. Juri Fukuda, Mr. Takumi Fujimura, and Mr. Satsuki Kurokawa) for their support.

6. REFERENCES

[1] Kitahara, M. 2001. *Category Structure and Function of Pitch Accent in Tokyo Japanese*. Indiana University.

[2] Sugiyama, Y. 2022. Identification of minimal pairs of Japanese pitch accent in noise-vocoded speech. *Front. Psychol.* 31. doi: 10.3389/fpsyg.2022.887761

[3] Beckman, M. E., Pierrehumbert, J. B. 1986. Intonational structure in Japanese and English. *Phonol. Yearb.* 3, 255–309. doi: 10.1017/S095267570000066X.

[4] Cutler, A., Otake, T. 1999. Pitch accent in spoken-word recognition in Japanese. *J. Acoust. Soc. Am.* 105, 1877–1888. doi: 10.1121/1.426724

[5] Remez, R. E., Rubin, P. E., Pisoni, D. B., Carrell, T. D. 1981. Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191.

[6] Hillenbrand, J. M., Clark, M. J., Baer, C. A. 2011. Perception of sinewave vowels. *J. Acoust. Soc. Am.* 129, 3991–4000. doi: 10.1121/1.3573980

[7] Remez, R. E., Dubowski, K., Broder Hytowitz, R., Davids, M., Grossman, Y. S., Moskalenko, M., Pardo, J. S., Hasbun, S. M. 2011. Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 968–977. doi: 10.1037/a0020734.

[8] Shinohara, Y. 2022. Perception of noise-vocoded sine-wave speech of Japanese pitch-accent words. *JASA Express Lett.* 2, 085204. doi: 10.1121/10.0013423.

[9] Shinohara, Y. 2022. Japanese pitch-accent perception of noise-vocoded sine-wave speech. *J. Acoust. Soc. Am.* 152, A175. doi: 10.1121/10.0015940.

[10] Haesen, B., Boets, B., Wagemans, J. 2011. A review of behavioural and electrophysiological studies on auditory processing and speech perception in autism spectrum disorders. *Res. Autism Spectr. Disord.* 5, 701–714. doi: 10.1016/j.rasd.2010.11.006.

[11] Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., Bonnel, A.-M. 2003. Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *J. Cogn. Neurosci.* 15, 226–235. doi: 10.1162/089892903321208169.

[12] Wang, X., Wang, S., Fan, Y., Huang, D., Zhang, Y. 2017. Speech-specific categorical perception deficit in autism: An event-related potential study of lexical tone processing in Mandarin-speaking children. *Sci. Rep.* 7, 43254. doi: 10.1038/srep43254.

[13] Hudry, K. et al. 2010. Preschoolers with autism show greater impairment in receptive compared with expressive language abilities. *Int. J. Lang. Commun. Disord.* 45, 681–690. doi: 10.3109/13682820903461493.

[14] Bruneau, N., Bonnet-Brilhault, F., Gomot, M., Adrien, J.-L., Barthélémy, C. 2003. Cortical auditory processing and communication in children with autism: electrophysiological/behavioral relations. *Int. J. Psychophysiol.* 51, 17–25. doi: 10.1016/S0167-8760(03)00149-1.

[15] Bruneau, N., Roux, S., Adrien, J. L., Barthélémy, C. 1999. Auditory associative cortex dysfunction in children with autism: evidence from late auditory evoked potentials (N1 wave–T complex). *Clin. Neurophysiol.* 110, 1927–1934. doi: 10.1016/S1388-2457(99)00149-2.

[16] Isaksson, S., Salomäki, S., Tuominen, J., Arstila, V., Falter-Wagner, C. M., Noreika, V. 2018. Is there a generalized timing impairment in Autism Spectrum Disorders across time scales and paradigms? *J. Psychiatr. Res*., 99, 111–121. doi: 10.1016/j.jpsychires.2018.01.017.

[17] Kargas, N., López, B., Reddy, V., Morris, P. 2015. The relationship between auditory processing and restricted, repetitive behaviors in adults with autism spectrum disorders, *J. Autism Dev. Disord*., 45, 658–668. doi: 10.1007/s10803-014-2219-2.

[18] Vlaskamp, C. et al. 2017. Auditory processing in autism spectrum disorder: Mismatch negativity deficits, *Autism Res*., 10, 1857–1865, doi: 10.1002/aur.1821.

[19] Lepistö, T., Kujala, T., Vanhala, R., Alku, P., Huotilainen, M., Näätänen, R. 2005. The discrimination of and orienting to speech and non-speech sounds in children with autism. *Brain Res*. 1066, 147–157. doi: 10.1016/j.brainres.2005.10.052.

[20] Akinaga, K., Kindaichi, H. 2014. *Shin Meikai Nihongo Akusento Jiten*, 2nd edition. Sanseido.

[21] Pervasive Developmental Disorders Autism Society Japan Rating Scale (PARS) Committee. 2013. *Pervasive Developmental Disorders Autism Society Japan Rating Scale: Text Revision*. Spectrum Publishing Co.

[22] Ueno, K., Nagoshi, N., Konuki, S. 2008. *PVT-R kaiga goi hattatsu kensa. [Picture Vocabulary Test]*. Nihon Bunka Kagakusha.

[23] Kjelgaard, M. M., Tager-Flusberg, H. 2001. An investigation of language impairment in autism: Implications for genetic subgroups. *Lang. Cogn. Process.* 16, 287–308. doi: 10.1080/01690960042000058.

[24] Darwin, C. 2005. SWS [Praat Script]. http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS.

[25] Winn, M. 2021. Vocoder [Praat Script]. http://www.mattwinn.com/praat/vocode_all_selected_v45.txt.

A perceptual experiment testing the Italian /ʎ/ - /j/ contrast

Anna Anastaseni[1,2] & Antonio Romano[1]

[1]LFSAG, Laboratorio di Fonetica Sperimentale "Arturo Genre", Univ. degli Studi di Torino
[2]GIPSA-lab, CNRS/Université Grenoble Alpes

ABSTRACT

Due to the articulatory proximity, Italian /ʎ/ and /j/ are confused by some speakers. The neutralisation of the /ʎ/-/j/ contrast lets prevail [j]-like sounds. Furthermore /lj/ and /llj/ are sometimes realised with a [ʎ]-like sounds in Northern regions.

Two classification tasks were proposed to a sample of 113 native Italian listeners to assess the incidence of the duration factor in distinguishing intervocalic segments [ʎ], [j] and [lj], manipulating the duration of the sounds within the same contexts.

It emerged that participants were always able to tell apart [lj]-like stimuli, while they showed more variation in the definition of /ʎ/ and /j/ categories. The duration plays a significant role in the identification of /ʎ/ and /j/: with less than 140-160 ms, the perceived stimulus originated from a [ʎ]-sound is generally classified as /j/, whereas stimuli with increasing durations of an original [j]-sound need an upper threshold (~200 ms) to be mostly attributed to /ʎ/.

Keywords: Auditory phonetics, identification task, palatal lateral, palatal approximant, duration.

1. INTRODUCTION

The phonological inventory of Italian distinguishes between the lateral palatal /ʎ(ʎ)/ and the approximant palatal /j/; however, this distinction has low-functional load. In neutral pronunciation the palatal /ʎ/ is always geminate in postvocalic position [1]. The alternation as short and long is governed by position: is not phonemic. Its duration exceeds 100 ms in standard Italian and in several regional varieties [2], but it does not reach the length of the corresponding non-palatal lexical geminate /ll/ [l:] and, apparently, it does not induce the preceding vowel to shorten [3]. On the other hand [j] can be considered as inherently short even though in various regions may be lengthened [1].

The lateral one has a complex articulatory mechanism ([1], [4], [5]). It appears late in the phonological inventory of Italian-speaking children [6] and it is often simplified ([7], [8]). The phoneme /ʎ/ is rare in majority of the world's languages; it is present in only 20 of the 451 languages considered in UPSID (*UCLA Phonological Segment Inventory Database*). Moreover, it occurs in only 16 languages whose inventory also includes the palatal approximant /j/ (by contrast, in 378 languages in the database).

The [ʎ]-sound and [j]-sound can be defined as approximants: they have a periodic structure, but a reduced intensity of about 10 dB compared to vowels [9]. The neutralisation of the opposition between /ʎ/ and /j/ is widespread and is favoured by acoustic similarity, as shown by the evolution of phonological systems in other Romance languages ([8], [10], [11]). In Italy, the delateralisation of /ʎ/ (/ʎ/ → [j]) is a widespread phenomenon in the Centre-South (with considerable micro-areal variation, [12]–[14]), with the exception of Tuscany and Sardinia [1]. The neutralisation of the /ʎ/ - /j/ contrast allows [j]-like sounds to prevail. On the other hand, in the same regions a lengthening of /j/ in intervocalic position is observed ([1], [14]). Furthermore, the phonological status of /ʎ/, mainly in Northern regions, has to be assessed taking into account the extension of [ʎ]-like sounds to the realisation of /lj/ and /llj/ [1], [15], [16].

Moreover, in pronunciation, among speakers of younger generations especially in the North, regional features are weak and often phonetic features coming from different regional Italian varieties are co-present [17]. As the three corpora ParlaTO, KIP and CLIPS show, in the variety of Italian spoken in Turin, [ʎj] and [j:] are also attested, not only the expected variants [ʎ] and [lj] [18].

Within the community we have been observing in the last two years by means of acoustical analysis, the loss of the lateral feature seems not to give rise to a real neutralisation between /ʎ/ and /j/ since their common realisations are respectively [j:] and [j] and a length contrast is often preserved.

A perceptual experiment has been set up in order to verify the hypothesis that a duration feature is sufficient to avoid a full merger. It is important to check the ability of Italian speakers to classify different phonemes. In this pilot study, we have prepared two identification tasks.

Firstly, stimuli with /ʎ/, /j/, /lj/ and /llj/ were presented to the listeners. The results showed that participants had more difficulty distinguishing the palatal lateral and palatal approximant. For this reason, we decided to focus our attention on /ʎ/ and /j/, with an exploratory reanalysis of the first experiment.

2. METHOD

In both experiments, an identification task was proposed to a group of Italian speakers.

Participants had to listen to a list of pseudo-words and had to associate each pseudo-word heard to an orthographic transcription within a closed list of possible labels. The orthographic labels were spelled with ‹gl›, which in Italian is usually associated with /ʎ/, ‹i›, associated with /j/, ‹li›, associated with /lj/, and ‹lli› associated with /llj/ in the first experiment. In the second experiment responses were associated to pseudo-words with ‹gl› and ‹i›.

The tasks were disseminated online via the Folerpa platform [19], an online tool developed at the Instituto da Lingua Galega (ILGA) of the University of Santiago de Compostela (Galicia, Spain).

In the first task 32 pseudo-words were presented to each participant, in the second task 40; each item was tested twice (total stimuli: 64 for the first test; 80 for the second). To avoid list effects the stimuli were presented at each replay in random order. In the middle of the task there was a pause.

The response times were calculated in milliseconds, between the end of the repetition of each stimulus and the listener's response.

2.1. Participants

The first task was proposed to 67 university students. Age ranged from 18 years to 29 years (mean: 21.20; SD: 2.52), all the participants were Italian native speakers. 62% of the subjects indicated Piedmont as their region of origin.

On the other hand, the second experiment was carried out with 46 students of the University of Turin aged between 21 and 31 (mean 23.93; SD: 2.07). 57% of the participants indicated Piedmont as their region of origin.

Participants were asked to indicate all the languages in which they have active and passive competence. The mother tongue and the L2, in fact, can modify the performance of identity task. Comparative studies conducted on Catalan and Spanish show that Catalan speakers have better performance in distinguishing [ʎ:] and [j] than Castilian speakers [20], [21].

All participants had no hearing problems, nor had they ever suspected hearing loss.

Finally, it was controlled that no subject had given more than 85% of identical answers and that there weren't any technical problems.

Each student was asked to complete the Vinegrad Plus Questionnaire [23], according to the modified version of Montesano and Valenti [23]. This test requests answers to a variety of questions designed to identify common traits of dyslexia. 33% of the participants were in the risk range, while 28% said they suspected to be dyslexic. These data are particularly important because phonological deficits are linked with dyslexia (see 2.3.1). In fact, many studies have shown the presence of phonological difficulties even in dyslexic adults, especially in phonological tasks that require access to phonology from visual inputs ([24]-[26]).

This fact is relevant since the identification task in the proposed modalities has an important visual component. Participants, in fact, had to make an effort of grapho-phonematic conversion to associate the heard stimuli with the proposed orthographic labels ([27]–[30]).

2.2. Stimuli

The stimuli consisted of four groups of pseudo-words with [ʎ:], [j] [lj] and [l:j] (Table 1).

All stimuli had the same phonotactic structure. They were disyllabic and paroxytone pseudo-words, with the target sound in intervocalic position. The vowels preceding it were stressed.

Pseudo-words were chosen to avoid the influence of lexical variables of existing words on the identification test.

	Task 1, 2		*Task 1*	
	/ʎ/ ‹gl›	/j/ ‹i›	/lj/ ‹li›	/llj/ ‹lli›
[ˈa]_[a]	/ˈtʃaʎa/ ‹ciaglia›	/ˈtʃaja/ ‹ciaia›	/ˈtʃalja/ ‹cialia›	/ˈtʃallja/ ‹ciallia›
[ˈa]_[o]	/ˈnaʎo/ ‹naglio›	/ˈnajo/ ‹naio›	/ˈnaljo/ ‹nalio›	/ˈnalljo/ ‹nallio›
[ˈɔ]_[o]	/ˈrɔʎo/ ‹roglio›	/ˈrɔjo/ ‹roio›	/ˈrɔljo/ ‹rolio›	/ˈrɔlljo/ ‹rollio›
[ˈɔ]_[a]	/ˈtɔʎa/ ‹toglia›	/ˈtɔja/ ‹toia›	/ˈtɔlja/ ‹tolia›	/ˈtɔllja/ ‹tollia›

Table 1: Selected stimuli for the identification tasks (task 1 and 2).

Once the pseudo-words had been selected for the test, a professional speaker was recorded in a silent room, he repeated the stimuli in isolation.

Afterwards, the durations of the vowels preceding the target phonics [ʎ:], [l:j], on the one hand, and [j], [lj], on the other hand, were equalised to 145 ms (± 5ms) and 160 (± 5 ms) using GoldWave audio editing software.

The mean of the duration of [ʎ:] in the four pseudo-words (*ciaglia, naglio, roglio, toglia*) was 158 ms. All stimuli were standardised to this value (± 2 ms), and then, keeping the previous vowel identical, the duration of the target sound was varied by reducing it by 40 ms each time and creating four new stimuli with [ʎ] durations of 220, 180, 140 and 100 ms.

During the first task, all stimuli were used, whereas for the second task, only stimuli with [ʎ:] and [j] were used.

Due to the impact of the duration factor for the classification of [ʎ] (§2.3.1), on the second task we focused on the rule of duration on [ʎ]-[j] distinction. The same manipulation was done with the stimuli presenting [j]-sound, in the original audio of 180 ms duration, manipulated to reproduce the same *continuum* as [ʎ] (260, 220, 180, 140, 100 ms).

2.3. Data Analysis

2.3.1. *Accuracy of participants' responses in the dyslexia risk range*

As mentioned, the Vinegrad Plus Questionnaire was administered to the participants of the two experiments in order to check whether there was a percentage of people in the sample who were in the dyslexia risk group.

It emerged that 33% were actually in this group. Given the possible correlation between dyslexia and the outcome of the tasks, the data on the accuracy of the answers were cross-referenced with the presence or absence of dyslexia risk.

To verify this hypothesis, the Chi-square test was applied and the χ^2 value and Cramer's V were calculated. This analysis did not yield significant results ($\chi^2 = 0.626$; V = 0.013). This fact suggests that difficulties in the classification of /ʎ/ and /j/ are not characteristic only of those with specific difficulties, but more widespread in the population.

Because of dyslexia is not affecting the results of the tasks we proposed, we did not delete the data of any participant for analysis.

2.3.2. *Experiment 1*

The first task showed that the participants were able o correctly associate /lj/ and /llj/ to the corresponding orthographic label ‹li› and ‹lli› (see Table 2).

	‹gl›	‹i›	‹li›	‹lli›
/ʎ/*	69%	22%	6%	4%
/j/	23%	70%	3%	4%
/lj/	2%	1%	91%	7%
/llj/	3%	1%	5%	92%

Table 2: Summary of identification task's results. Experiment 1. Classification of /ʎ/, /j/, /lj/, /llj/.

* Here are the reported results of stimuli with a [ʎ] duration of 180 ms, the average value among those tested. By entering the duration values of [ʎ] =180 ms here, [ʎ] and [j] with the same duration are compared.

The data were processed statistically with SPSS for Analytics (by IBM), v. 28. As we were working with categorical data, crosstabs were used to check the relationship between two variables (type of stimulus and listener categorisation). We conducted Chi-square test and Cramer V. The significance value was set at 0.05 (α=0.5).

As a result, the degree of association of /lj/ and /llj/ with the respective graphic labels ‹li› and ‹lli› is significantly higher (χ^2= 606.696; V=0.871) than for /ʎ/ and /j/ (χ^2= 191.701; V= 0.490).

In addition, we assessed the participants' sensitivity to the duration of [ʎ] by evaluating whether it affects classification. As discussed, in fact, duration is one of the variables of the realisation of these phonemes. For this reason, each pseudo-word was proposed with the target sound [ʎ] having durations of 260, 220, 180, 140 and 100 ms.

It was found that the duration factor played a significant role in the correct recognition of the sound (Pearson's r=0.947**, p=0.01). As the duration of [ʎ] decreased, the number of target responses ‹gl› dropped from 87% at 260 ms to just 9% at 100 ms (Figure 1).

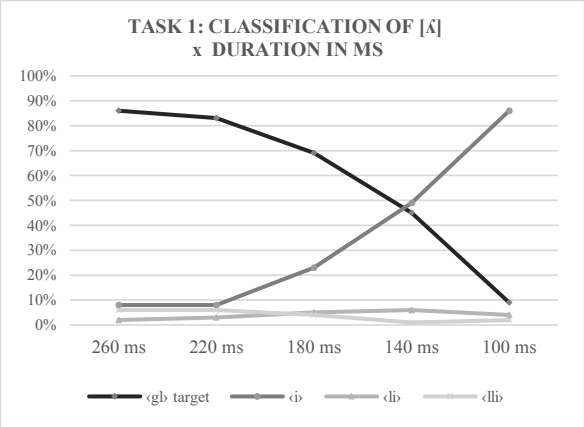


Figure 1*:* Classification of [ʎ] as a function of sound duration (in ms). Experiment 1.

Given these initial results, it was decided to investigate this further by proposing a task in which not only the length of [ʎ] but also that of [j] was manipulated (Experiment 2).

2.3.3. *Experiment 2*

Our results of the first experiment demonstrate that the duration of [ʎ] is decisive in the subjects' choices in the identification task. The data collected in the first task was then confirmed by the second experiment and are presented below.

In this experiment, listeners had to associate manipulated stimuli with /ʎ/ and /j/ (100-260 ms) and orthographic labels with ‹gl› and ‹i›.

The results show that in more than 90% of the responses the sound [ʎ]≥220 ms it is solidly associated with the target transcription ‹gl› and the stimuli with [j] at 100 ms is associated with ‹i›.

There is a directly proportional relationship between the duration of [ʎ] and the target association ‹gl› (Pearson's r = 0.935**; p = 0.02): the longer the duration of the sound, the better its recognition.

Whereas, surprisingly, an inversely proportional relationship appears in the recognition of [j] (Pearson's r = –0.988; p = 0.002; Figure 22). The [ʎ]-sound below the 180 ms threshold is no longer identified with certainty. On the other hand, [j] is almost unambiguously associated with ‹i› when it has durations of 100 ms and 140 ms.

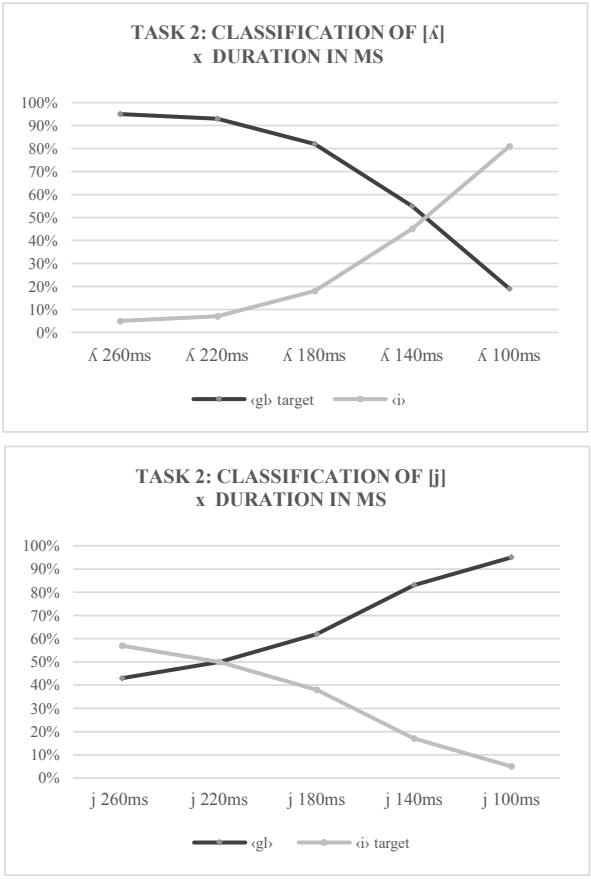


Figure 2: Experiment 2. Classification of [ʎ] and [j] as a function of duration (in ms).

Significantly, however, the curves describing the classification of [ʎ] and [j] as ‹gl› and ‹i› according to duration do not show an entirely symmetrical profile.

The most significant finding concerns pseudo-words with [j] of longer duration, which are alternately associated with ‹gl› and ‹i›.

It appears that different parameters come into conflict: the quality of the sound and its duration. It follows that [j] of duration superior than 180 ms is clearly classifiable neither as /ʎ/ nor as /j/.

This finding suggests that the duration feature is regarded as a significant index by the listeners.

Indeed, in Italian, duration characterises and distinguishes the two phonemes.

3. CONCLUSIONS AND PERSPECTIVES

Two perceptual identification tasks were submitted to a total of 113 Italian-speaking university students.

The results of these tasks showed that participants were always able to tell apart [lj]-like stimuli, while they showed more variation in the way they classified the stimuli in the /ʎ/ and /j/ categories. In this case, it is confirmed that the duration plays a significant role in the identification of /ʎ/: with less than 140-160 ms. The perceived stimulus originated from a [ʎ]-sound is generally classified as /j/, whereas stimuli with increasing durations of an original [j]-sound need an upper threshold (around 200 ms) to be mostly attributed to /ʎ/.

The [ʎ]-sound is associated with longer duration than [j]. What has been observed has important consequences for spelling. Indeed, it must be taken into account that if at a perceptual level the distinction between phonemes is not clear, in writing the indirect route cannot be used or in any case, it has an increased chance of not being reliable.

Moreover, it is necessary to complete further research in order to reach more solid conclusions. It would therefore be necessary to verify the discrimination and identification capacity of the two phonemes in different age groups, as well as to extend the research geographically.

This is the reason why we decided to collect data from a more controlled sample. Therefore, it was decided to propose a reading task, a pseudo-words dictation task and an auditory discrimination task to 60 Piedmontese speakers, divided into age groups (15-20, 20-25, 25-30). All data were collected in person. Data analysis is still in progress.

4. REFERENCES

[1] Canepari, L. 1999. *Manuale di pronuncia italiana.* Zanichelli.

[2] Celata, C. 2006. *Analisi del processo di retroflessione consonantica in area romanza, con dati sperimentali dal còrso e dal siciliano*. Pisa: Scuola Normale Superiore (unpublished PhD dissertation).

[3] Endo, R., Bertinetto, P.M. 1999. Caratteristiche prosodiche delle così dette 'rafforzate' italiane. *Atti delle IX Giornate di Studio del Gruppo di Fonetica Sperimentale.* 243–255.

[4] Bladon, R.W., Carbonaro, E. 1978. Lateral consonants in Italian. *Journal of Italian Linguistics.* 3 (1), 43–54.

[5] Recasens, D. 1984. Vowel-to-vowel coarticulation in Catalan VCV sequences: an articulatory and acoustical study. *Journal of Phonetics*. 12, 61–73.
[6] Zanobini, M., Viterbori, P., Saraceno, F. 2012. Phonology and Language Development in Italian Children: An Analysis of Production and Accuracy. *Journal of Speech, Language, and Hearing Research*. 55, 16–31. https://doi.org/10.1044/1092-4388(2011/10-0228).
[7] Galatà, V., Meneguzzi, G., Conter, L., Zmarich, C. 2012. Primi dati sull'acquisizione fonetico-fonologica dell'italiano L2 in prescolari rumeni. In: Paoloni, A.; Falcone, M. (eds), *La voce nelle applicazioni*. Bulzoni, 35–50.
[8] Zampaulo, A., Haug, D.T.T. 2015. The evolution of the (alveolo) palatal lateral consonant in Spanish and Portuguese. In: Haug, D.T.T. (ed.), *Historical Linguistics: Selected papers from the 21st International Conference on Historical Linguistics* (Oslo 2013), Benjamins, 69–86.
[9] Pettorino, M., Giannini, A. 1991. Indagine acustica sulle approssimanti dell'italiano. *Atti del XIX Convegno Nazionale dell'Associazione Italiana di Acustica*. 10–12.
[10] Porras, J.E., 2013. Spanish Yeísmo: A Cognitive Linguistic Approach to Phonological Change. *Spanish Yeísmo: A Cognitive Linguistic Approach to Phonological Change*. 335–352.
[11] Rost Bagudanch, A. 2014. Una panorámica del yeísmo: ¿un proceso acabado o en construcción? *Revista Internacional de Lingüística Iberoamericana*. XII, 1(23), 141–163.
[12] Romano, A., Badin, P. 2009. An MRI Study on the Articulatory Properties of Italian Consonants. *Estudios de Fonética Experimental*. XVIII, 327–344.
[13] Colonna, V., Romano, A. 2018. La variazione diatopica nel micro-spazio dialettale leccese: il dialetto salentino delle frazioni di Vernole. In: Caramuscio G., Romano A. (eds), *Una d'arme, di lingua, d'altare, di memorie, di sangue, di cor* - Omaggio a Luciano Graziuso, Grifo, 105–123.
[14] Marotta, G., Nocchi, N. 2001. La liquida laterale nel livornese. *RID. Rivista italiana di dialettologia*. 25, 285–326.
[15] Romano, A., Miletto A.M. 2017. *Argomenti scelti di glottologia e linguistica.* Omega.
[16] Bertinetto, P.M., Loporcaro, M. 2005. The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association.* 35, 131-151.
[17] Cerruti, M. 2013. Varietà dell'italiano, In: Iannaccaro, G. (ed.), *La Linguistica Italiana All'alba Del Terzo Millennio (1997-2010)*. Bulzoni, 91–127.
[18] Fontanot, A. 2019, *Aspetti di italiano "composito" nella pronuncia di giovani torinesi: un'indagine su corpora*, Università degli studi di Torino (unpublished).
[19] Fernández Rei, E., Cajiao A. A., Peláez C. O., Garabal J. A. C. 2021. *FOLERPA: Ferramenta On-Line para ExpeRimentación PerceptivA* , Santiago de Compostela: Instituto da Lingua Galega. Available online: https://ilg. usc. gal/folerpa (accessed on 20 October 2021).
[20] Rost Bagudanch, A. 2016. La percepción de/ʎ/ y /j/en catalán y en español. Implicaciones en la explicación del yeísmo. *Estudios de Fonética Experimental*. XXV, 40–80.
[21] Rost Bagudanch, A. 2020. Bilingualism and sound change: perception in the /ʎ/-/j/ merger process in Majorcan Spanish. *Zeitschrift für romanische Philologie*. 136, 106–133.
[22] Vinegard, M., 1994. A revised adult dyslexia check list. *Educare-London-National Bureau For Handicapped Students*, 21–21.
[23] Montesano, L., Valenti, A., 2018. SLD: promoting continuity between School and University education. Results of a screening program in Calabria. *Form@ re-Open Journal per la formazione in rete.* 18, 192–206.
[24] Svensson, I., Jacobson, C., 2006. How persistent are phonological difficulties? A longitudinal study of reading retarded children. *Dyslexia*. 12, 3–20. https://doi.org/10.1002/dys.296.
[25] Danelli, L., Berlingeri, M., Bottini, G., Borghese, N.A., Lucchese, M., Sberna, M., Price, C.J., Paulesu, E. 2017. How many deficits in the same dyslexic brains? A behavioural and fMRI assessment of comorbidity in adult dyslexics. *Cortex*. 97, 125–142. https://doi.org/10.1016/j.cortex.2017.08.038.
[26] McCrory, E.J., Mechelli, A., Frith, U., Price, C.J. 2005. More than words: a common neural basis for reading and naming deficits in developmental dyslexia? *Brain*. 128, 261–267. https://doi.org/10.1093/brain/awh340.
[27] Yasuhara, M., Kuklinski, T.T. 1978. Category boundary effect for grapheme perception. *Perception & Psychophysics*. 23, 97–104. https://doi.org/10.3758/BF03208288.
[28] Gibson, E.J., Pick, A., Osser, H., Hammond, M., 1962. The Role of Grapheme-Phoneme Correspondence in the Perception of Words. *The American Journal of Psychology*. 75, 554–570. https://doi.org/10.2307/1420279
[29] Bassetti, B., Cerni, T., Masterson, J. 2022. The efficacy of grapheme-phoneme correspondence instruction in reducing the effect of orthographic forms on second language phonology. *Applied Psycholinguistics.* 43, 683–705. https://doi.org/10.1017/S014271642200008X.
[30] Stoehr, A., Martin, C.D. 2022. The impact of orthographic forms on speech production and perception: An artificial vowel-learning study. *Journal of Phonetics*. 94, 1–17 https://doi.org/10.1016/j.wocn.2022.101180.

Intelligibility of spontaneous casual speech in background noise across the lifespan

Outi Tuomainen[1,2], Amanda Soengadie[2], Linda Taschenberger[1], Valerie Hazan[1]

Speech, Hearing and Phonetic Science, University College London (UCL)[1], Linguistics, University of Potsdam[2]
tuomainen@uni-potsdam.de

ABSTRACT

This study investigated speech-in-noise intelligibility of child, young, middle aged and older adult talkers (N=48; 24 Female). The aim was to investigate whether age and sex differences in intelligibility remain when masking noise is individually tailored to talker characteristics. A total of 83 native English listeners took part in an online listening study. The listeners heard short sentences extracted from spontaneous diapix conversations; these were mixed with speech-shaped noise that was tailored to each talker's long-term average spectrum. The number of correct keywords was calculated. There were no significant differences between talker groups based on sex or age apart from the middle-aged male talkers who were less intelligible than other talker groups. These results suggest that some of the previous findings with regards to intelligibility differences between different talker groups (female vs male; younger vs older adults) may be driven by the masking potential of background noise.

Keywords: Speech intelligibility, speech perception, speech-in-noise, lifespan

1. INTRODUCTION

How well or poorly a particular talker is understood by a listener is determined by multiple parameters. For example, it has been shown that individual talker characteristics (e.g., talker sex and age), listener characteristics (e.g., listener's hearing acuity, degree of shared language experience with the talker), speaking style (casual vs clear speech), and the listening environment (quiet vs noisy backgrounds and background noise type and level) can all contribute to perceived speech intelligibility (for a review see [1]). Therefore, one could argue that speech intelligibility is highly constrained by the individual parameters that were manipulated in each perceptual assessment of intelligibility.

Assessment of speech intelligibility in background noise, in particular, has gained much attention in past research [2, 3, 4, 5, 6]. Overall, multiple studies have shown that with regards to effects of age and sex, for casual everyday speech heard in noisy backgrounds, female talkers are more intelligible than male talkers [7], and that younger adults are more intelligible than older adults (60+ years) [e.g., 6]. For children, it has been shown that children acquire intelligible speech gradually in early childhood. For example, the intelligibility of casual speech heard in quiet, is at 96% for speech produced by 3 year olds, and 99% for speech produced by 8 year olds [8]. For older children and casual speech presented in background noise, studies have shown that speech intelligibility for 11-12-year-old child talkers does not differ from adults [3, 5].

The interesting question then is that what explains these differences in intelligibility between different talker groups. There are anatomical and physiological differences between female and male talkers and younger and older talkers that result in differences in their acoustic-phonetic profiles: for example, in fundamental frequency (f0) and temporal characteristics, as well as in the amount of mid-frequency (1-3 kHz) energy in their speech [e.g., 2, 9]. However, despite these well-documented age- and sex-related differences in speech intelligibility and acoustics, studies have often failed to identify any systematic acoustic-phonetic correlates of perceived intelligibility. For example, some studies have highlighted speaking rate as a factor, with slower speech facilitating intelligibility in different multitalker and reverberant environments [2] whereas other studies have reported that f0 characteristics [7] or the amount of energy within the 1-3 kHz frequency range [4,6] are better predictors of intelligibility than speaking rate. Furthermore, these three acoustic cues are also shown to be important for speech perception in noisy environments as there is evidence that listeners use these cues to help segregate target and competing talkers in various multitalker contexts [e.g., 10].

However, one potential confound for the unclear and conflicting results with regards to predictors of perceived speech intelligibility in background noise is the plethora of different background noise types and levels used between these different studies. Because some background noise types might be less effective in masking talkers from different age groups, it is hard to draw parallels between the results from these different studies. For example, multi-talker babble that is based on a mixture of adult

female and male voices may be less effective in masking child voices as the target talkers. Furthermore, signal-to-noise ratios that are either very hard [11] or very easy [3] can lead to floor and ceiling effects in listener performance which can make it difficult to statistically assess the true extent of listeners' abilities. Finally, to our knowledge, no other study has previously investigated speech intelligibility in a more lifespan design, that is, mapping differences between children and younger, middle aged and older adults in the same study.

Therefore, in the current study, we extend some of the previous findings by investigating the intelligibility of spontaneous casual speech in background noise across the lifespan (here: 10-76 years of age) at moderate difficulty levels. Here, we focus on how deriving the background noise from the individual long-term average spectrum (LTAS) of the target talker influences the perceived intelligibility of their speech. We predict that if age- and sex-related differences in perceived speech intelligibility are mainly related to the degree to which the background noise is masking the acoustic characteristics of the talker, these group differences related to age and sex will disappear, or be significantly reduced, if the background noise provides a similar degree of masking for each talker. However, if perceived intelligibility is primarily driven by individual talker characteristics such as speaking rate, mid-frequency energy and f0, we expect to replicate previous findings with regards to these sex- and age-related differences.

2. METHODS

2.1. Listeners

Participants were recruited via Prolific (www.prolific.co) [12]. Only participants with a Prolific score (i.e., participant quality rating) of 70/100 or higher were invited to the study. A total of 83 native monolingual English listeners (recruitment criteria for country of birth: United Kingdom) aged 20-42 years completed the study (M=29 years, SD=6.6). All participants self-declared no history of language/hearing impairments, and they passed the initial online automated screening for headphone use and the auditory attentional screening during experimental testing (see below for details). Ethical approval was obtained from the UCL Research Ethics Committee.

2.2. Speech materials

The speech materials used were taken from the LifeLUCID corpus [13], which consists of interactive speech from native Southern British English talkers across the lifespan (8-80 years) communicating with an unfamiliar partner of the same sex and similar age. The pair of talkers had to complete a spot-the-difference picture-based task (diapix; [14]) within 10 minutes or less. One person in the pair was told to take the lead in the task. We extracted speech samples from child (CH), young (YA), middle-aged (MA), and older talkers (OA) communicating in favourable listening conditions (i.e. no noise in background), thereby eliciting a conversational speaking style. The samples were short sentences or part-sentences containing between 4 and 10 words (e.g., "there's a dog sitting on a blue seat" or "the big green bush") containing 3 keywords. Most samples were simple descriptions of the diapix pictures containing high frequency lexical items similar to BKB (Bamford-Kowal-Bench) sentences [15]. Criteria for the selection of the samples were that they be equally spaced throughout the task, did not contain any disfluencies and were not produced in response to a request for repetition by the interlocutor. Each speech sample was extracted from the original audio recording using Praat (v 6.0.21). In total, 288 speech samples were extracted from 48 lead talkers (12 CH, 10-17 years old, M = 13.85 years; 12 YA, 20-26 years old, M = 22.32 years; 12 MA, 35-49 years old, M = 42.95 years; 12 OA, 59-76 years old, M = 66.50 years; 6 female per group). Six speech samples were taken from each talker, resulting in 72 samples per age group. For the male CH group, three pre-puberty and three post-puberty talkers were included. This classification was based on the talker's median f0 with pre-pubescent males ranging from 94-99 semitones (relative to 1 Hz) and post-pubescent from 80-85 semitones.

A Praat script was used to produce tailored speech-shaped noise for each talker. This was done by creating noise with a spectrum matching their individual long-term average spectrum, which was calculated from their complete diapix speech recording (duration of noises 20 sec). These tailored noises were then mixed with the speech samples and root-mean-square normalised to 70 dB in Matlab (2016b). Initial piloting showed a fixed signal-to-noise ratio (SNR) of -4 dB to achieve an average intelligibility score (% keywords correct) of around 50-70% (avoiding floor/ceiling effects in intelligibility).

2.3. Procedure

The perception test was run online, using the Gorilla platform to run and host the experiment [www.gorilla.sc]. Data was collected between 28 May and 1st June 2020. Samples were divided into six lists with each list comprised of one sample from each of the 48 talkers. Additionally, each list included eight cross-reference trials and 4 attentional trials. The latter consisted of one sample presented in quiet from each age group (2 female, 2 male). The cross-reference trials, which were identical for all listeners, were included to check consistency in intelligibility scores across listener groups. These were additional masked samples from the existing 48 talkers (1 female, 1 male from each age group). This resulted in each participant hearing one sample from each of 48 talkers (disregarding attentional and cross-reference trials); six participants therefore made up a complete data set for each talker. Trials were randomised with attentional trials occurring every 12 trials.

After hearing each sample online once, participants were asked to type out the target sentence as accurately as possible. They were then presented with the keywords on a following screen and had to select those that they had correctly identified. Participant self-scorings were manually checked and corrected for any errors before statistical analyses.

2.4. Data analysis

Data are presented as raw scores of correctly recognised keywords where a score of 18 represents 100% correct (3 keywords x 6 talkers). Statistical analysis were conducted in R (v 1.4.1106) using an linear mixed-effects models (*lmer* function) for main effects and interactions of Talker sex (F,M) and Talker age (CH, YA, MA, OA) with listener as random effect. Post hoc comparisons were conducted using the *emmeans* package in R (Tukey p-value adjustment).

3. RESULTS

3.1 Attentional and cross-reference trials

All participants were near or at ceiling in the attentional trials (overall performance 96%). Furthermore, the performance for the cross-reference trials did not differ between the six different lists ($p=.185$), indicating that there were no significant differences for identical samples across listener groups who were hearing different sample lists.

3.2 Sentences-in-noise results

Raw scores (out of 18) for keywords correctly identified for female and male and CH, YA, MA and OA groups separately are displayed in Fig. 1. The best fitting model included significant interaction between Talker sex and Talker age ($\chi2$ (3)=64.40, $p<.001$) and, therefore, the main effects of Talker sex and Talker age were retained in the final model.

For Talker sex, female talkers (M=10.7, SD=3.05) were more intelligible than male talkers (M=9.3, SD=3.44; $p<.001$).

For Talker age, we found that OA and CH talkers achieved highest overall intelligibility and YA and MA talkers achieved the lowest overall intelligibility (see Fig. 1). MA talkers were the least intelligible group and they were significantly less intelligible than CH and OA talkers (both comparisons; $p<.001$). YA talkers were also less intelligible than OA talkers ($p<.001$; see Fig. 1). MA and YA talkers did not differ from each other ($p=0.393$).

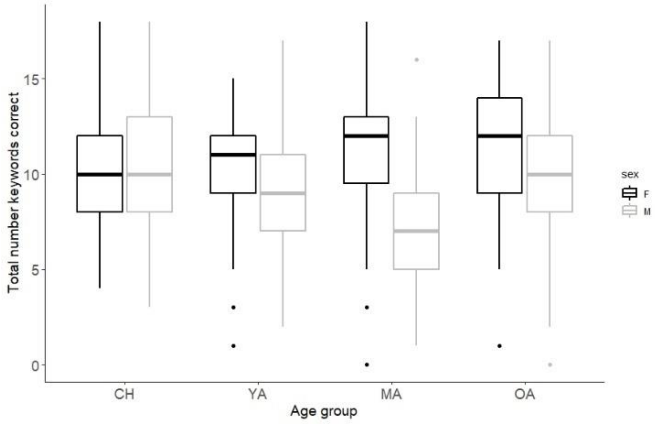


Figure 1. Number of keywords correct (out of 18) for the child (CH), young adult (YA), middle aged adult (MA) and older adult (OA) talkers separated by Talker sex (Female, Male).

However, Talker sex * Talker age interaction indicated that these sex- and age-related differences were mainly driven by the MA male talker group. First, post-hoc comparisons between the four different age groups separately for the female and male talkers revealed that none of the age groups differed in female talkers (all comparisons, $p\geq.110$) whereas in the male talkers, the MA group was significantly less intelligible than the CH, YA and OA groups (all comparisons $p<.001$). Second, post-hoc comparisons between female and male talkers within each of the four age groups revealed that female talkers were significantly more intelligible than male talkers only in the MA group ($p<.001$); all other comparisons $p\geq.134$).

4. CONCLUSIONS

Our data supports the hypothesis that age- and sex-related differences in perceived speech intelligibility are, at least partially, related to the properties of the background masking noise. When the masking noise was individually tailored to the individual talker characteristics with respect to its long-term average spectrum, we generally found very little differences in perceived intelligibility in different-aged talker groups.

Contradicting many previous studies, we found here that child and older adult talkers were (at least marginally) easiest to understand and middle aged adults were most difficult to understand in individually tailored background noise. On the other hand, aligning with many previous findings, we reported here too that female talkers were more intelligible than male talkers. However, in the current study these group differences with regards to age and sex were primarily driven by the low score achieved for the middle aged male talkers. Although male talkers achieved a lower score than female talkers in all four groups, the difference generally remained relatively small at one to two fewer keywords correctly identified (out of 18 total), whereas the intelligibility score for the middle aged male talkers was approximately four to five keywords correct lower than that for their female peers.

The reasons for this difference in this particular age group are unclear. Perceptually, it appeared that, despite tailoring the background noise masker to the acoustics of individual talkers, some target talkers might still be masked better by the noise than others. For example, it is possible that for some of the talkers the masking noise better covered lower frequency regions (that are rich in phonetic detail for word identity) leaving fewer audible phonemes/words to aid recognition [16]. Furthermore, as stated earlier, speech intelligibility is very likely a result of multiple parameters involving the individual characteristics of both the listener and the speaker as well as the content of the message itself. Therefore, we plan to further investigate these individual differences and look at which talkers are more/less intelligible and what, if any, acoustic parameters might help explain these results and predict speech intelligibility in our dataset. We will also conduct qualitative analyses of the content of the speech samples used in the study to further rule out the effect of sentence predictability on speech intelligibility in our sample.

Lastly, it is worth noting that although the talkers in this study all spoke the Southern British English variety, we have no information about the regional accents of the listeners. However, listener characteristics are an unlikely explanation for our results as the different listener groups performed consistently and did not differ on the cross-reference trials. Furthermore, despite the fact that it has been well established that shared language and accent background between the speaker and the listener (e.g., in terms of phonology) increase perceived intelligibility [see e.g., 17], it is nevertheless unlikely that a discrepancy between the accents influence intelligibility of one particular sub-group of talkers, such as middle aged male talkers.

In summary, this study highlights the importance of carefully considering masking characteristics in studies of speech in noise that investigate correlates of speech intelligibility.

5. REFERENCES

[1] Smiljanić, R., & Bradlow, A. R. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass*, *3*(1), 236-264.

[2] Cox, R., Alexander, G., and Gilmore, C. 1987. Intelligibility of average talkers in typical listening environments *The Journal of the Acoustical Society of America*. 81, 1598 – 1608.

[3] Hazan, V., & Markham, D. 2004. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

[4] Hazan, V., Tuomainen, O., Tu, L., Kim, J., Davis, C., Brungart, D., Sheffield, B. 2018. How do aging and age-related hearing loss affect the ability to communicate effectively in challenging communicative conditions?, *Hearing Research*, 369, 33-41.

[5] Smiljanic, R., & Gilbert, RC. 2017. Intelligibility of Noise-Adapted and Clear Speech in Child, Young Adult, and Older Adult Talkers. *Journal of Speech, Language, and Hearing Research*, 60(11), 3069–3080. https://doi.org/10.1044/2017_JSLHR-S-16-0165.

[6] Tuomainen, O., Hazan, V., Davis, C., & Kim, J. 2019. Intelligibility of conversational and clear speech in young and older talkers as perceived by young and older listeners. *The Journal of the Acoustical Society of America*, *146*(1), EL28.

[7] Bradlow, AR., Torretta, GM., Pisoni, DB. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

[8] Flipsen P. 2006. Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics*, 20(4), 303–312.

[9] Tuomainen, O., Taschenberger, L., Rosen, S., & Hazan, V. 2022. Speech modifications in interactive speech: effects of age, sex and noise type. *Philosophical transactions of the Royal Society of*

London. Series B, Biological sciences, 377(1841), 20200398. https://doi.org/10.1098/rstb.2020.0398

[10] Cushing, IR., & Dellwo, V. 2010. The role of speech rhythm in attending to one of two simultaneous speakers. In *proceedings of speech prosody*, Chicago.

[11] Smiljanic R. 2013. Can older adults enhance the intelligibility of their speech?. *The Journal of the Acoustical Society of America*, 133(2), EL129–EL135. https://doi.org/10.1016/j.heares.2018.06.009.

[12] www.prolific.co. London, UK (accessed March 2020).

[13] Tuomainen, O, Taschenberger, L., Hazan, V. 2022. LifeLUCID Corpus: Recordings of Speakers Aged 8 to 85 Years Engaged in Interactive Task in the Presence of Energetic and Informational Masking, 2017-2020. Data Collection. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854350.

[14] Baker, R., Hazan, V. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43, 761–770

[15] Bench, J., Kowal, A. and Bamford, J. 1979. The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, 13, 108-112. http://dx.doi.org/10.3109/03005367909078884

[16] Cooke M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

[17] Hansen Edwards, JG., Zampini, ML., & Cunningham, C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, 37(4), 538-557

Clear Speech Improves Word Recognition, but may Increase Listening Effort

Kirsten Meemann & Rajka Smiljanić

The University of Texas at Austin
kirsten.meemann@utexas.edu, rajka@austin.utexas.edu

ABSTRACT

Listener-oriented clear speech (CS) improves word recognition in noise and auditory memory. Using a dual-task paradigm, the current study examined whether this processing benefit is related to the effect of intelligibility-enhancing CS on listening effort in the presence of energetic and informational maskers. Native English listeners repeated sentences produced in a conversational and clear speaking style and mixed with either two-talker babble or speech-shaped noise. They simultaneously performed a visual task on the computer screen and their response times were measured.

The results showed that hearing clear speech sentences increased response times on the visual task in the presence of an energetic masker. The response times on the visual task in informational masker were not different between the two speaking styles. This suggests that listeners direct their attentional resources to the more salient hyperarticulated clear speech and that a decrease in listening effort may not underlie the CS intelligibility benefit.

Keywords: listening effort, clear speech, speech perception, dual task

1. INTRODUCTION

When an acoustic signal is degraded or inaccessible, accurate sound and word recognition becomes considerably more challenging resulting in a decrease in intelligibility. Even when word recognition is high, processing speech under challenging listening conditions can increase cognitive load and listening effort [1], [2]. The task demands, such as presence of noise or unfamiliar accent, can increase listening effort, or "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task, with listening effort applying more specifically when tasks involve listening" (Framework for Effortful Listening (FUEL), [1, p. 5S]; Ease of Language Understanding (ELU) model, [3]). According to these models, listeners have limited cognitive capacity and they draw on the available resources accordingly when performing several tasks simultaneously [4]. The more effort listeners expend to understand degraded speech, the fewer resources they have left to complete simultaneous tasks or remember spoken information [1], [5]. Effortful listening can lead to decreased motivation and task disengagement [6] and it can have long-term health consequences such as increased stress, fatigue, anxiety and social withdrawal [7].

One way in which speakers can increase intelligibility is through a listener-oriented hyperarticulated clear speech (CS) [8]. In addition to enhancing word recognition, CS aids various linguistic and cognitive processes, such as speech segmentation and auditory memory [9]. Using pupillometry, Borghini & Hazan [10] recently showed that the presence of the enhanced acoustic-phonetic cues present in CS improved intelligibility and significantly reduced listening effort. However, the relationship between intelligibility and effort is non-monotonic [11]. Furthermore, different measures of listening effort may assess different aspects of listening effort (e.g., [12]). The FUEL indicates that behavioral measures may be most sensitive to listening demands, while physiological measures may be more sensitive to both listening demands and motivation [1]. The current study examined how variation in intelligibility through speaking style modifications and different masker types affects listening effort using a dual-task paradigm. Previous work showed that maskers that vary in the degree of informational (IM) and energetic masking (EM) impact the intelligibility benefit listeners obtain from CS [13]. Here, we focus on investigating whether listener-oriented speaking style modifications affect listening effort differently in EM and IM ([14] for a review).

Native English listeners performed a word recognition task in speech-shaped noise (SSN) and 2-talker (2T) babble and a visual response task first separately and then simultaneously. Self-report measures have additionally been used to quantify subjective listening effort (e.g., [15]). We hypothesized that 1) processing easier-to-understand CS will reduce listening effort, 2) listening effort will be greater in 2T babble, where the speech signal is masked by both EM and IM, than in SSN, which imposes only EM, and 3) the CS benefit on effort will be smaller in 2T babble, where word recognition accuracy is expected to be higher than in SSN [13]. These hypotheses are consistent with the FUEL and ELU models in that more intelligible CS will reduce

cognitive load leaving additional processing resources for the secondary visual task. The aim was to provide a more nuanced understanding of the CS intelligibility benefit and its association with listening effort.

2. METHODS

2.1. Participants

Fifty-one native English listeners (average age: 21.0 years, SD: 4 years, range: 18-34 years) participated in the study. All participants were recruited at the University of Texas at Austin. They provided written consent to use their responses anonymously. All listeners passed hearing screening bilaterally at 25 dB HL at 500, 1000, 2000, and 4000 Hz. They received class credit or a small monetary compensation for participation.

2.2. Materials

Primary Task

One 21-year-old female native speaker of American English read 60 semantically meaningful sentences (e.g., *Mice like to eat cheese*; [16]) first in conversational (CO) and then in clear (CS) speech. For CO speech, the talker was told to speak in a casual manner as if she was talking to a friend or family member. For CS, the talker was instructed to speak as if she was communicating with someone who has a low proficiency in English and who has difficulty following them conversationally (following [17]). All sentences were normalized for root-mean-square amplitude and mixed with competing speech and with SSN at -5dB signal-to-noise ratio (SNR). The competing speech masker consisted of 2-talker (2T) babble (2 female talkers) [18]. The speech-shaped noise masker had been generated by shaping white noise to match the long-term average spectrum of 6-talker (6T) babble (3 male, 3 female talkers; [18]). 12 additional sentences produced in CO and CS were used as practice stimuli for the word recognition task. They were mixed with 2T babble and SSN at -5dB SNR.

Secondary Task

Visual stimuli for the secondary non-linguistic visual task consisted of two square boxes (approximately 5 cm across), one on the left and one on the right side of the screen. At each trial, a left- or a right-pointing arrow appeared in one of the two boxes.

Subjective Effort

We adapted four questions from the original NASA-TLX [21] related to exerted effort, mental demand, perceived performance (error rate), and frustration to assess subjective listening effort [12].

2.3. Procedure

The experiment began with the primary word recognition task alone followed by the secondary visual stimulus response task alone presented in SuperLab 5 (Cedrus). The word recognition single task consisted of 12 sentences, alternating between speaking style and masker combinations, CO+2T, CO+SSN, CS+2T, CS+SSN, three times. The order of presentation in the single task was fixed. Sentences were presented at set intervals of 4000 ms. For each trial, the noise started 100 ms before the target sentence and stopped 100 ms after the target ended. Participants were instructed to listen to the sentences, one at a time, and to repeat each target sentence verbally. Oral responses were recorded and scored off-line afterwards.

For the secondary visual task, a left- or right-facing arrow appeared in the left- or the right-hand side box (at 500-2000 ms intervals). In the visual single-task condition, 40 trials were presented. Participants were instructed to press ‘z’ if a left-pointing arrow appeared in either of the boxes on the screen, and to press ‘m’ if a right-pointing arrow appeared on the screen. Accuracy and response times were recorded for each trial.

In the dual task following the two single tasks, participants heard and verbally repeated target sentences while simultaneously completing the visual task described above. They were told to complete both tasks to the best of their ability, but to prioritize the word recognition task [19]. In the dual task, participants were presented with four 15-sentence blocks of speaking style (CO and CS) and masker (2T and SSN) combinations, one combination per block. With three keywords per sentence, there were 45 target words for each of the four speaking style and masker combinations. All listeners heard sentences produced in both speaking styles and presented with both types of masking noise. The speaking style and noise masker combination was counterbalanced across blocks such that each listener heard a different style and masker combination order. The order in which the blocks of sentences were presented was pseudorandomized. Listeners never heard the same sentence twice. In the primary word recognition task, sentences were presented at fixed intervals of 4000 ms, and in the concurrent secondary visual task, arrows remained on the screen until participants responded or after 2500 ms had elapsed. The

presentation of visual and auditory stimuli in the dual-task condition did not systematically coincide. Response times to correct responses in the secondary visual task were collected to quantify listening effort.

After each of the four dual-task blocks, subjective ratings of listening effort were collected. The order of the five questions after each block was consistent. For each question, participants were told to click on a blank square on an unnumbered 21-point scale that corresponded to their experience during the most recent dual-task block. The full experimental session took about 30 minutes to complete.

3. ANALYSES

To assess effort, response times (RTs) for answers to the visual task were analyzed for speaking style and masker combinations. Only RTs for correct responses (96.57% of stimuli) were included in the analysis [15]. RTs slower than 2500 ms and faster than 200 ms were excluded from the data analysis.

Data were analyzed with linear mixed effects regression using the *lme4* package in R. Model comparisons were done via likelihood ratio tests, where models differed only in fixed effects. All models included by-participant and by-item (word) random intercepts. Significant results were further analyzed with estimated marginal means (*emmeans* package in R [20]).

For intelligibility, each sentence used in the primary task contained four target words which were scored as correct (1) or incorrect (0). Data were analyzed with logistic mixed-effects regression using the *glmer* function in R [21]. All models had word recognition score as a dependent variable with a binomial link function, and random intercepts for listeners and words. Likelihood ratio tests were conducted to determine the models with the best fit. Further post-hoc analyses were done via estimated marginal means tests.

Subjective listening effort data were analyzed via mixed effects modeling (*lmer*). Self-report scores for the effect of speaking style and noise masker were analyzed on only the effort question. All models included by-subject random intercepts.

4. RESULTS

Response Times

Descriptive statistics for response time data in the dual task are shown in Figure 1. Results revealed that RTs to visual stimuli were significantly slower in the dual task condition (mean = 652.8 ms; SD = 296.2 ms;) than when performing the visual task alone (mean = 566.1 ms; SD = 186.2 ms) (z = 13.39, p < .001).

Post-hoc analyses of a two-way interaction between speaking style and masker ($\chi^2(1) = 9.14$, p < .01) showed that RTs were significantly shorter for sentences produced in CO than in CS when presented in SSN (z = -4.71, p < .001). There was no difference in RTs between speaking styles when presented in 2T babble (z = -0.41, p = .68). Lastly, listeners responded significantly more slowly to CO sentences presented in 2T babble than in SSN (z = 3.65, p < .001), but not CS sentences (z = -0.57, p = .57).

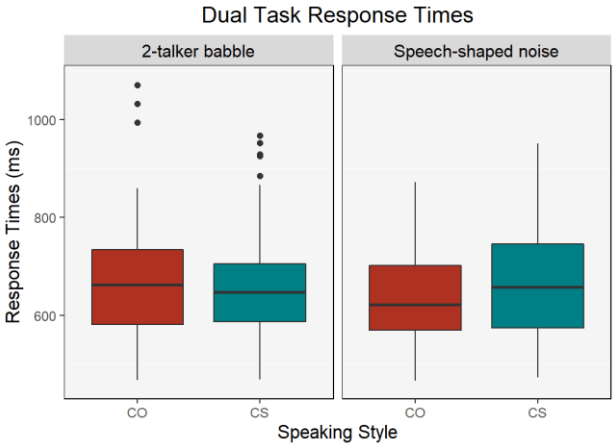


Figure 1. Listeners' response times in Conversational speech (CO) and Clear speech (CS) in 2-talker (2T) babble (left panel) in speech-shaped noise (right panel).

Speech Intelligibility

A two-way interaction between speaking style and noise masker ($\chi^2(1) = 31.50$, p < .001) revealed that intelligibility was significantly better for CS than CO sentences in both masker conditions (SSN: z = 8.37, p < .001; 2T babble: z = 2.62, p < .001). The odds of successful word recognition were 90% higher for CS than CO speech in SSN (OR = 1.90, 95% CI, 1.64-2.21, and 22% higher in 2T babble (OR = 1.22, 95% CI, 1.05-1.42). Additionally, post-hoc comparisons showed that listeners correctly identified significantly more words when target sentences were presented in SSN than in 2T babble, but only when the sentences were produced in CS (z = 4.45, p < .001). Specifically, listeners were 41% more likely to recognize CS-produced words in SSN than in 2T babble (OR = 1.41, 95% CI, 1.21-1.64).

Subjective Effort

To evaluate subjective effort, we examined participants' ratings for how effortful they found the task. A model with main effects of speaking style and masker yielded the best fit ($\chi^2(1) = 7.70$, p < .01). Results showed that listeners perceived CO sentences as more difficult than CS sentences ($t(79) = 2.79$, p < .01) and sentences presented in 2T babble as more

effortful than sentences presented in SSN ($t(79) = 3.67$, $p < .001$).

5. DISCUSSION

In contrast to our hypothesis, the results of the dual-task experiment showed that CS did not reduce listening effort despite increasing intelligibility. Listening effort, as measured by the RTs in response to the visual task, did not differ between CS and CO speaking styles in 2T babble, and it even increased in CS compared to CO when speech was masked by SSN. Additionally, listening effort was increased in 2T babble compared to SSN when sentences were produced in CO, but there was no difference between masker types when speech was produced in CS.

The results suggest that the well-established CS intelligibility benefit does not arise from a decrease in listening effort. This finding is in contrast with Borghini and Hazan [10] who showed smaller pupil dilation for CS than CO, suggesting a decrease in listening effort. Behavioral measures (e.g., response times) are hypothesized to be most sensitive to listening demands, while physiological measures, such as pupillometry, are sensitive to both listening demands and motivation (cf., [1]). This is also in line with other recent studies that found a lack of correlation between different measures of listening effort [22], [23].

Dual-task paradigms assume limited cognitive capacity, such that the difficulty of the primary task can determine how many resources are left for simultaneously performing a secondary task: The more difficult the primary task, the more cognitive resources are directed to completing that task and the fewer resources are left to perform the secondary task, leading to a performance decrease in the secondary task [24]. Consequently, this suggests that RTs in a dual task can reflect listeners' direction of attention, regardless of instructions to prioritize the primary task.

Our results suggest that it is not only the difficult primary task that may draw listeners' attention. Listeners seem to direct their attentional resources to the more acoustically salient and easier-to-process clear speech compared to conversational speech (cf., [25]). This finding is in line with Brown and Strand [15] who reported increased use of cognitive resources for processing speech with visual cues than without, despite better word recognition in the audiovisual modality. CS may thus be more intelligible not because of lower processing cost but because of the allocation of cognitive resources. Under this view, listeners are automatically drawn to the exaggerated acoustic-phonetic cues of the listener-oriented clear speech, and to the presence of visual cues, resulting in enhanced intelligibility via the engagement of additional resources.

Listeners' subjective effort ratings seem to stand in contrast to the dual-task findings. Listeners perceived CS and SSN as less effortful than CO and 2T babble. However, the combined findings show that objectively measured listening effort was larger in the speaking style and masker combination that listeners subjectively perceived as easier. This further supports the notion that different measures tap into different kinds of effort, such as the effort listeners exert during a task versus the emotional response to their performance and exerted effort.

Regarding masker type, word recognition accuracy was better in the purely energetic masker (SSN) than in the predominantly informational masker (2T babble) only for clearly produced speech. Conversely, RTs did not differ between the two masker types in CS, but they were longer in 2T babble than in SSN for CO speech. This indicates that listeners spent more effort to understand conversationally produced speech masked by 2T babble compared to SSN to achieve similar recognition accuracy. In terms of subjective effort, listeners rated 2T babble as more effortful compared to sentences presented in SSN. Taken together, the objective and subjective measures of effort show that informational masking (IM) from 2T babble is more effortful than energetic masking from SSN. This is consistent with previous studies that reported an association between IM and increased objectively measured listening effort as indicated by pupillometry [26] and increased subjective listening effort [27]. The finding that objectively measured effort and subjective effort aligned here suggests that suppressing irrelevant lexical intrusion from IM is more effortful. It seems that listeners correctly assessed that they were working harder in order to attend to the target speech. This further suggests that the subjective measures of effort may not be equally sensitive to the challenges presented by speaking style variation vs. acoustic degradation through signal masking. The link between the subjective ratings and various types of listening challenges warrants further exploration.

In summary, this study provides evidence that the well-established CS processing benefits may not be arising from a decrease in listening effort. Instead, listeners seem to direct their cognitive resources to more attention-grabbing, salient speech. The results also showed that speech degraded predominantly by informational masking increased listening effort. This research has the potential of improving the quality of treatments for hearing-impaired individuals by relying on insights from both intelligibility and listening effort.

REFERENCES

[1] M. K. Pichora-Fuller *et al.*, "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)," *Ear Hear.*, vol. 37, pp. 5S-27S, 2016.

[2] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response.," *Ear Hear.*, vol. 32, no. 4, pp. 498–510, 2011.

[3] J. Rönnberg *et al.*, "The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances," *Front. Syst. Neurosci.*, vol. 7, no. July, pp. 1–17, 2013.

[4] D. Kahneman, *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall, 1973.

[5] P. M. A. Rabbitt, "Channel-capacity, intelligibility and immediate memory," *Q. J. Exp. Psychol.*, vol. 20, no. 3, pp. 241–248, 1968.

[6] M. Richter, "The moderating effect of success importance on the relationship between listening demand and listening effort," *Ear Hear.*, vol. 37, pp. 111S-117S, 2016.

[7] B. W. Y. Hornsby, G. Naylor, and F. H. Bess, "A taxonomy of fatigue concepts and their relation to hearing loss," *Ear Hear.*, vol. 37, pp. 136S-144S, 2016.

[8] R. Smiljanić, "Clear Speech Perception: Linguistic and Cognitive Benefits," in *The Handbook of Speech Perception*, 2nd ed., D. B. Pisoni, R. E. Remez, L. C. Nygaard, and J. S. Pardo, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc, 2021, pp. 177–205.

[9] S. Keerstock and R. Smiljanić, "Clear speech improves listeners' recall," *J. Acoust. Soc. Am.*, vol. 146, no. 6, pp. 4604–4610, 2019.

[10] G. Borghini and V. Hazan, "Effects of acoustic and semantic cues on listening effort during native and non-native speech perception," *J. Acoust. Soc. Am.*, vol. 147, no. 6, pp. 3783–3794, 2020.

[11] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started," *Trends Hear.*, vol. 22, pp. 1–32, 2018.

[12] J. F. Strand, V. A. Brown, M. B. Merchant, H. E. Brown, and J. Smith, "Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures," *J. Speech Lang. Hear. Res.*, vol. 61, no. 6, pp. 1463–1486, 2018.

[13] K. Meemann and R. Smiljanić, "Intelligibility of Noise-Adapted and Clear Speech in Energetic and Informational Maskers for Native and Nonnative Listeners," *J. Speech, Lang. Hear. Res.*, vol. 65, no. 4, pp. 1263–1281, 2022.

[14] J.-P. Gagné, J. Besser, and U. Lemke, "Behavioral assessment of listening effort using a dual-task paradigm: A review," *Trends Hear.*, vol. 21, pp. 1–25, 2017.

[15] V. A. Brown and J. F. Strand, "About Face: Seeing the Talker Improves Spoken Word Recognition but Increases Listening Effort," *J. Cogn.*, vol. 2, no. 1, p. 44, 2019.

[16] M. Fallon, S. E. Trehub, and B. A. Schneider, "Children's use of semantic cues in degraded listening environments," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2242–2249, 2002.

[17] A. R. Bradlow and J. A. Alexander, "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners.," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2339–2349, 2007.

[18] K. J. Van Engen and A. R. Bradlow, "Sentence recognition in native- and foreign-language multi-talker background noise," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 519–526, 2007.

[19] V. A. Brown and J. F. Strand, "Noise increases listening effort in normal-hearing young adults, regardless of working memory capacity," *Lang. Cogn. Neurosci.*, vol. 34, no. 5, pp. 628–640, 2019.

[20] R. V. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." R package version 1.3.1, 2018.

[21] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[22] C. Visentin, C. Valzolgher, M. Pellegatti, P. Potente, F. Pavani, and N. Prodi, "A comparison of simultaneously-obtained measures of listening effort: pupil dilation, verbal response time and self-rating," *Int. J. Audiol.*, vol. 61, no. 7, pp. 561–573, 2021.

[23] A. L. Francis, T. Bent, J. Schumaker, J. Love, and N. Silbert, "Listener characteristics differentially affect self-reported and physiological measures of effort associated with two challenging listening conditions," *Attention, Perception, Psychophys.*, vol. 83, no. 4, pp. 1818–1841, 2021.

[24] S. Fraser, J.-P. Gagné, M. Alepins, and P. Dubois, "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues.," *J. speech, Lang. Hear. Res.*, vol. 53, no. 1, pp. 18–33, 2010.

[25] V. Boswijk, H. Loerts, and N. H. Hilton, "Salience is in the eye of the beholder: Increased pupil size reflects acoustically salient variables," *Ampersand*, vol. 7, p. 100061, 2020.

[26] S. Villard, T. Perrachione, S.-J. Lim, A. Alam, and G. Kidd, "Listening effort elicited by energetic versus informational masking," *Proc. Meet. Acoust.*, vol. 45, pp. 1–12, 2022.

[27] J. Rennies, V. Best, E. Roverud, and G. Kidd, "Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort," *Trends Hear.*, vol. 23, pp. 1–21, 2019.

NOISE DOES NOT AFFECT WEIGHING OF SPEAKER INFORMATION IN SPOKEN-WORD RECOGNITION

Helen Reese & Eva Reinisch

Acoustics Research Institute, Austrian Academy of Sciences
hreese@kfs.oeaw.ac.at, ereinisch@kfs.oeaw.ac.at

ABSTRACT

Information about the speaker, such as gender, is known to affect phoneme categorization. However, studies examining this effect have mostly been conducted in quiet laboratory conditions. In addition to more closely resembling realistic listening, background noise has been shown to impact listeners reliance on phonetic detail relative to lexical information. Noise is therefore a useful manipulation for testing how different parts of the speech signal are processed. We tested the effect of noise on the weighing of speaker gender. Speaker gender is encoded in the acoustic signal but provides information about the speaker rather than what is being said. Its processing thus may pattern with 'lower-level' phonetic or 'higher-level' indexical information. The results of a minimal pair categorization experiment indicate that listeners' use of speaker gender during phoneme categorization is not affected by speech-shaped noise masking the speech signal. This suggests that speaker information does not pattern with phonetic detail.

Keywords: speech perception, speaker information, adverse conditions, noise

1. INTRODUCTION

Listeners are able to recognize words with remarkable accuracy, despite the immense variability in speech across speakers and listening conditions. Some factors underlying this variability have been shown to systematically affect speech perception. For instance, the identity of the speaker, such as their (inferred) age or gender, has been shown to affect how listeners perceive individual sounds [4, 6, 10]. This was first demonstrated in an experiment testing the effect of speaker gender on the perception of ambiguous stimuli along an /s/-/ʃ/ continuum [17]. Listeners were more likely to categorize stimuli along the continuum as an /s/ when it was paired with a male voice and as /ʃ/ when paired with a female voice [17]. Similar effects have since been replicated with other contrasts, such as /ʊ/-/ʌ/ [6] and /s/-/θ/ [10], and via other manipulations aimed at isolating the effect of gender from other general acoustic effects. This was achieved, for example, by presenting the same audio stimulus with images of male and female faces [10, 18].

Importantly, it has been suggested that these gender effects arise from unconscious expectations about how men and women differ in the production of certain phonemes. For example, with regards to the /s/-/ʃ/ contrast, listeners appear to expect fricatives produced by a man to have lower spectral centers of gravity than when the same phoneme is produced by a woman, corresponding to the typical characteristics of female versus male speakers. Therefore, listeners report hearing /s/, which has a higher spectral center of gravity than /ʃ/, at lower values along an /s/-/ʃ/ continuum if the speaker is perceived as male [17].

This speaker gender effect has traditionally been studied in 'ideal' laboratory conditions, that is, without any background noise or distracting tasks. More recently, however, interest has increased in how listeners fare in so-called 'adverse' conditions that more closely resemble everyday listening situations [8]. Those are often characterized by noise disrupting the speech signal or listeners whose attention is dived among multiple tasks. Not only are these conditions more realistic, but they have been shown to also allow for examining how listeners weigh different parts of the speech signal during perception, that is, to what extent they rely on acoustic properties versus, for instance, lexical information to recognize words.

In general, listening under increased cognitive load in the form of a concurrent visual search task has been shown to increase listeners' reliance on lexical information while decreasing attention to acoustic cues. That is, in experiments testing the lexical bias (i.e., "Ganong") effect [5], participants showed a larger effect, that is, they were more likely to perceive an ambiguous sound as forming part of a real word as opposed to a non-word when completing a simultaneous visual search task with the phoneme categorization task [9]. In contrast, when listening in noise, listeners up-weighed (i.e., relied more on) phonetic as opposed to lexical information [7]. In an experiment testing word segmentation, participants were more willing to rely on acoustic cues to boundaries even if it led to semantically unacceptable combinations when listening in noise than when presented with the same stimuli in clear listening conditions [7].

By examining the effect of speaker gender on phoneme categorization with the additional manipulation of adverse listening conditions, the present study aimed at better understanding the processing of speaker information during speech perception. The role of speaker gender in speech perception is difficult to categorize in this respect. It deals with higher-level information about the speaker, but it is conveyed in the acoustic signal. Specifically, we asked whether the effect of speaker gender would pattern with the processing of higher-level (e.g., lexical) information or lower-level acoustic information.

One previous study [15] already examined the role of additional cognitive load on the speaker gender effect. It was found that performing a visual search task while listening to the audio stimuli did not affect the magnitude of speaker effect in either direction. That is, listeners' reliance on speaker gender when categorizing an /s/-/ʃ/ continuum neither increased nor decreased as compared to the control condition without or a lower added cognitive load. A possible explanation for this is that speaker information is processed as acoustic information as opposed to higher-level indexical or lexical information.

The current study set out to further investigate this issue by examining the effect of background noise on the speaker gender effect. As the addition of noise to the speech signal has been shown to increase reliance on phonetic information, we hypothesized that when listening in noise participants will up-weigh speaker information together with acoustic detail, thereby increasing the size of the speaker effect. We thus conducted an experiment in which the same audio stimuli as used in [15] were overlayed with speech-shaped noise. Participants completed a phoneme categorization task under two conditions: with and without noise.

2. METHODOLOGY

2.1. Participants

Thirty native speakers of Austrian German were recruited via the online platform prolific.co [13] and compensated according to the site's payment scheme. They were aged between 18 and 40 and reported no hearing difficulties. The experiment was completed remotely. All participants reported using headphones.

2.2. Design and Materials

The experiment design followed that of Strand and Johnson's seminal study on the speaker gender effect in the perception of /s/-/ʃ/ continua [17]. Audio recordings were taken from [15] in which native speakers of Austrian German were recorded producing word initial /s/-/ʃ/ minimal word pairs (note that in Austria, as opposed to Germany, German word-initial /s/ is produced voiceless).

Recordings of four speakers (two female and two male) were selected for manipulation. Using a script in PRAAT [2], the word initial fricatives (/s/ and /ʃ/) were extracted from the recordings and interpolated to create a 15-step /s/-/ʃ/ continuum. This continuum was then spliced onto the words replacing the original word-initial fricatives. Multiple continua were created, and one was selected that the experimenters judged to sound the most natural when combined with all voices. Minimal pairs of all speakers with the same fricative continuum were then subjected to piloting. Two speakers (one female, one male) and two minimal pairs (*sein-schein* "to be" – "appearance", *senken-schenken* "to sink" – "to bestow") were selected for use in the final experiments. They showed the most typical s-shaped categorization curves suggesting that the continuum is perceived as intended, hence allowing to test for additional effects. Of the 15 continuum steps, the two endpoints and seven intermediary steps were selected for use in the experiment, for a total of nine stimuli per speaker and minimal pair. For more information on the creation and piloting of the continua as well as the spectral properties of the fricatives, see [15].

For the present study those stimuli were further manipulated. Speech shaped noise was created using a modified PRAAT script based on [19] that averaged the long-term average spectra (LTAS) of the speech stimuli and filtered a segment of white noise according to this LTAS. For the noise condition, the minimal pair continua were then overlayed with this speech shaped noise at a signal-to-noise ratio (SNR) of -3. This was the strongest distortion that still elicited typical categorization curves in a series of pilot experiments testing different SNRs. Fig. 1 illustrates the effect of the distortion on the speech signal by showing spectrograms of the same stimulus with and without the added noise.

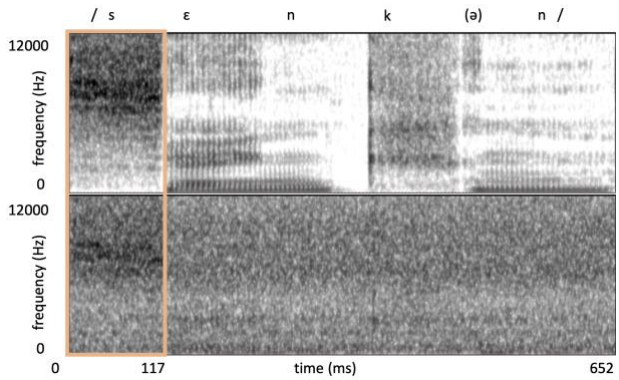


Figure 1: Spectrograms of the word *senken* (to sink) produced by the female speaker without (top) and with speech-shaped noise masking the signal (bottom).

Note that since the fricatives were also included in the noise mask, higher SNRs made it impossible for listeners to discern the acoustic properties of the fricatives.

2.3. Procedure

The experiment was created with PSYCHOPY3 [12] and hosted online via pavlovia.org [11]. It consisted of a two-alternative forced choice task in which participants were asked to identify the words along the /s/-/ʃ/ continuum. Per trial, participants were presented with a single audio stimulus, which they could hear only once per trial. They selected by button-press which of two words on the computer screen better matched the audio. In one condition, the audio files were presented overlayed with noise and in the other without noise. Context conditions were blocked, and condition order was counterbalanced across participants. Trial order was randomized for each participant. Each block consisted of 192 trials: for each minimal pair, the two end points were presented three times and the seven intermediary steps six times, thus 2 speakers × 2 minimal pairs × (2 end points × 3 repetitions + 7 intermediary steps × 6 repetitions) = 192. The experiment took approximately 15 minutes to complete.

3. RESULTS

Fig. 2 illustrates the categorization data as proportion /ʃ/-responses over the continuum.

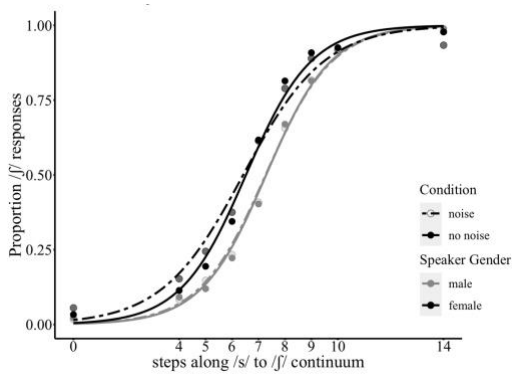


Figure 2: Proportion /ʃ/ responses over continuum steps for the two speakers (female in black, male in gray), by condition (solid lines -> noise; dashed lines -> no noise). Dots show the raw, lines the fitted data.

For statistical analyses the most extreme continuum steps were excluded due to near ceiling identification. Statistical analyses were conducted by means of a generalized linear mixed-effects model with a logistic linking function using the lme4 [1] package in R [14]. The dependent variable was response (/ʃ/ coded as 1, /s/ as 0). Fixed effects were Continuum Step (centered on 0), Condition (noise coded as 0.5, no noise as -0.5), Speaker Gender (female coded as 0.5, male as -0.5) and all interactions. The random-effects structure included an intercept for participants with random slopes for Continuum Step and Condition. This was the maximal random-effects structure that converged. Results are reported in Table 1.

Factor	b	z	p
Intercept	0.05	0.40	0.689
Step	0.95	18.52	<0.001
Condition	0.03	0.54	0.588
Gender	0.73	8.36	<0.001
Step:Condition	-0.07	-2.16	0.031
Step:Gender	-0.03	-0.91	0.362
Condition:Gender	-0.01	-0.05	0.959
Step:Condition:Gender	-0.10	-1.60	0.110

Table 1: Results of the statistical analysis. The model structure was: (phoneme classification ~ continuum step * condition * speaker gender + (1 + continuum step + speaker gender | participant)).

Significant effects were found for Continuum Step (more /ʃ/-responses the more /ʃ/-like the fricative) and for Speaker Gender, thus replicating the speaker effect that more /ʃ/-responses were given for the stimuli produced by the female speaker. The interaction Continuum Step x Condition was also significant, indicating that the effect of continuum step is smaller, that is, the categorization function is less steep in the noise than in the clear condition. No significant interaction was found for Step x Gender, indicating that the categorization function did not differ in steepness for the stimuli produced by the male and female speaker. Importantly, as for our main manipulation of listening condition, neither the interaction Condition x Gender nor the three-way interaction Step x Condition x Gender were significant, which signals that the noise condition did not affect the use of speaker gender in the categorization task.

4. DISCUSSION

The present study set out to further explore the effect of speaker gender in phoneme categorization, since the mechanism behind it is not clear. Information about the gender of the speaker is (mostly) contained in the acoustic signal (e.g., F0, voice quality etc.). At the same time, speaker gender provides higher-level indexical information about the speaker. In speech perception it could hence either pattern with the processing of lower-level acoustic or higher-level (e.g., indexical; also, lexical) information. While overall, listeners have been shown to use all types of information available (or accessible) to them for

speech perception, differences in the weighing of different types of information have been demonstrated especially in adverse listening conditions [7, 8].

The present study made use of adverse listening conditions to assess the nature of the speaker gender effect. We asked whether the addition of speech-shaped noise masking the acoustic signal would increase listeners' reliance on speaker gender in spoken-word recognition. This is because previous studies have shown an increase in listeners' reliance on acoustic relative to higher-level lexical information when listening in background noise [7]. We hence tested the hypothesis that the effect of speaker gender in phoneme categorization patterns with low-level acoustic listening in noise.

To examine this, we conducted an experiment in which participants categorized /s/-/ʃ/ minimal pair words. The words contained an initial fricative taken from a single /s/-/ʃ/ continuum and then combined with word ends (i.e., the rest of the word following the fricative) produced by one male and one female speaker. In one condition, listeners heard the audio stimuli in clear listening conditions and in the other the stimuli were overlayed with speech-shaped noise. In both conditions we replicated the speaker effect as it was first established in [17], that is, that more /ʃ/ responses were given for the female than the male speaker. However, the magnitude of this effect did not differ significantly between the noise and the clear listening conditions. The hypothesis that the speaker gender effect would pattern with the up-weighing of acoustic relative to higher-level information in noise could hence not be supported.

Interestingly, a previous study on the impact of another type of adverse listening condition on the speaker effect could not find support for its processing to pattern with higher-level lexical information either [13]. In this study, the addition of cognitive load in the form of a simultaneous visual search task was likewise shown not to affect the magnitude of the speaker effect [13]. In other words, the addition of a visual search task did not lead listeners to up-weigh speaker gender as a cue in the categorization task, while the same manipulation has been shown to cause listeners to rely more on lexical information by increasing the size of the lexical bias effect in experiments testing word segmentation and phonetic categorization under cognitive load [7, 9].

Evidently, the speaker effect does not pattern with effects found in previous studies implementing cognitive load and noise to test the differential weighing of acoustic versus higher-level information. Additional findings might speak to this apparent null result regarding the modulation of the speaker effect. First, it must be noted that noise "covers" the speech signal, hence acoustic cues are masked and not as easily available for listeners to use as in clear listening conditions. The supposed up-weighing of acoustic cues during listening in noise must therefore be seen relative to the use of other types of information [7]. In the present study, similarly to [15] in their study on cognitive load, the speaker effect was not directly compared to other (e.g., lexical) effects. Secondly, other studies that specifically targeted low-level acoustic processing effects such as speechrate-dependent phonetic categorization also failed to find modulating effects of cognitive load or noisy environments [3, 16].

Finally, while not affecting the magnitude of the speaker effect, the noise manipulation in the present study did show an effect on the perception of the continuum (as indicated by the Condition by Continuum Step interaction), namely resulting in a shallower categorization curve in the noise condition than in the clear listening condition. This suggests that listeners were less able to access acoustic information in the fricatives (also a kind of noise) in the categorization task when listening in noise. Similar effects on the categorization curve were found in the experiment on the effect of cognitive load on the speaker gender effect [13].

Where does this leave us with regard to the processing of speaker information? The role of speaker gender information in spoken-word recognition cannot be clearly categorized as patterning with acoustic or lexical information. One explanation for this could be that speaker gender information is processed independently of or in a different fashion than low-level acoustic information or higher-level (lexical) information. However, the exact mechanism remains to be described, possibly by using different experimental paradigms, such as by studying the time-course of when speaker (gender) information is employed during word recognition.

5. REFERENCES

[1] Bates, D. M., & Sarkar, D. 2007. lme4: Linear mixed-effects models using S4 classes (Version 0.999375–27) [Computer program]. http://www.r-project.org/: R Foundation for Statistical Computing.

[2] Boersma, P. & Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.50.

[3] Bosker, H.-R., Reinisch, E. & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. Journal of Memory and Language, 94, 166-176.

[4] Drager, K. 2011. Speaker age and vowel perception. *Language and Speech* 54, 99-121.

[5] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of experimental*

psychology: Human perception and performance, 6, 110-125.
[6] Johnson, K., Strand, E. A., & D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics* 27, 359-384.
[7] Mattys, S. L., Brooks, J., & Cooke, M. 2009. Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive psychology* 59, 203-243.
[8] Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27, 953-978.
[9] Matty, S. L., & Wiget, L. 2011. "Effects of cognitive load on speech recognition," J. Mem. Lang. 65, 145–160.
[10] Munson, B. 2011. The influence of actual and imputed talker gender on fricative perception, revisited (L). *J. Acoust. Soc. Am*. 130, 2631-2634.
[11] Pavlovia, https://www.pavlovia.org/
[12] Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. 2019. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*.
[13] Prolific, https://www.prolific.co/.
[14] R Core Team. 2021. R: A language and environment for statistical computing. (version 4.1.2) R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.R
[15] Reese, H., & Reinisch, E. 2022. Cognitive load does not increase reliance on speaker information in phonetic categorization. *JASA Express Letters* 2.
[16] Reinisch, E., & Bosker, H. R. (2022). Encoding speech rate in challenging listening conditions: white noise and reverberation. Attention, Perception, & Psychophysics, 84, 2303-2318.
[17] Strand, E. & Johnson, K. 1996. Gradient and Visual Speaker Normalization in the Perception of Fricatives. *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference*, 14-26.
[18] Strand, E. 1999. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology* 18, 86-100.
[19] Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication* 52, 911-918.

THE ONLINE PROCESSING OF ANTICIPATORY TONAL INFORMATION IN STANDARD CHINESE BY NATIVE AND NON-NATIVE LISTENERS

Hana Nurul Hasanah [1,3], Qing Yang [1,2], Yiya Chen [1,2]
[1] Leiden University Centre for Linguistics, [2] Leiden Institute for Brain and Cognition, [3] Chinese Studies Universitas Indonesia
h.n.hasanah@hum.leidenuniv.nl, q.yang@hum.leidenuniv.nl, yiya.chen@hum.leidenuniv.nl

ABSTRACT

This study investigated whether and how anticipatory tonal coarticulatory cues are utilized during online speech processing. Native speakers of Standard Chinese (SC) and Indonesian learners of SC participated in an eye-tracking experiment with the Visual World Paradigm (VWP). They were asked to identify two-syllable nonce words that were printed on a screen while listening to auditory stimuli that had varying levels of appropriate anticipatory tonal cues. The eye movement data revealed that (a) both groups showed some sensitivity to anticipatory tonal information when selecting a target item; (b) native listeners of SC demonstrated a predictive effect of tonal processing when the anticipatory tonal cues in the first syllable were strong, but this effect was not present for Indonesian-speaking SC learners.

Keywords: online speech processing, anticipatory tonal information, eye-tracking

1. INTRODUCTION

Listeners are known to use cues of anticipatory coarticulation in an incoming speech stream (e.g., [1, 2]) to speed up spoken word recognition (e.g., [3, 4]). Evidence, however, has been drawn mainly from segmental coarticulation in West-Germanic languages like English and Dutch. There is much less research on the possible utilization of anticipatory tonal cues in languages with lexical tone, such as Sinitic varieties. We know that in Standard Chinese (SC), tones show anticipatory coproduction in connected speech (e.g., [5] for bisyllabic context and [6] for trisyllabic context). Studies of real-time spoken word recognition in SC have revealed the important role of tonal (in addition to segmental) information in listeners' recognition of a target word before the auditory stimulus is entirely unmasked [7–10], and even for ambiguous acoustic signals with gradient tonal manipulation (e.g., for native SC listeners: [11]; for both native and non-native (English) listeners:[12]).

To date, however, only [13] have investigated the use of anticipatory tonal cues in SC spoken word recognition. Using the VWP, they examined listeners' looks to target-competitor words in Tone3 (T3) which differed in segments. They found that native SC listeners could utilize the T3 sandhi information to anticipate an upcoming T3 word. Note that tonal variation, including the anticipation cues, can be broadly classified as due to phonological alternation, known as tone sandhi, and to phonetic alternation, known as coarticulation (even though their boundary can be difficult to draw) (see [14] and references therein). Tone sandhi typically induces salient pitch changes, e.g., the T3 sandhi in SC [15]; its magnitude of tonal anticipatory effect is larger than coarticulation-induced anticipatory effects. It remains unclear whether subtle anticipatory cues due to tonal coarticulation are sufficient for listeners to speed up lexical tone access and to what extent coarticulation effects differ from tone sandhi effects. It is also unclear how general listeners' ability is to use anticipatory tonal information for speech processing and to what extent L1 speaker and L2 learners of a tonal language may differ in their utilization of anticipatory pitch cues.

To address these questions, the current study was designed to understand the role of different types of anticipatory tonal information in the online speech processing of both native and non-native listeners of SC. We chose Indonesian learners of SC which helps to generalize our findings to non-tonal L2 learners of SC whose first language is not a commonly studied Germanic language. We implemented the Visual World Eye-tracking Paradigm with two printed nonce words (NWs) (see [16] for a review of the paradigm). Listeners saw the nonce words on a screen while listening to auditory stimuli. We examined their eye fixations towards the printed stimuli.

The printed stimuli differed in the tone of the second syllable; the auditory stimuli either had coarticulatorily appropriate or inappropriate (i.e., matched vs. mismatched) anticipatory pitch cues. Comparing listeners' responses between the two types of cues is not only a classic approach to examine the benefit of

anticipatory cues in perception (e.g., [1]) but also recommended for the investigation of anticipatory tonal cues in spoken word recognition [17]. We included two tonal contexts that differed in the magnitude of tonal anticipatory cues: strong magnitude due to tone sandhi (Tone3; T3), to replicate the findings in [13], and weak magnitude due to coarticulation (Tone1; T1). Our goal was to compare whether degrees of anticipatory coarticulation affect how the coarticulatory cues are used.

Assuming listeners of SC use anticipatory tonal cues in online speech processing, we expected that when an auditory stimulus contains a mismatched tonal cue, there would be a delay in listeners' fixation on the target and consequently, temporarily decreased looks to the target. Furthermore, a strong anticipatory cue (T3 context) would amplify the cue effect and speed up listeners' recognition of the spoken nonce word, relative to the weaker coarticulatory cue (T1 context). Last, comparisons in the gaze patterns of native vs. non-native listeners would shed light on the general ability of listeners to utilize anticipatory tonal information for online speech processing and the role of L1 in modulating the use of such cues.

2. METHOD

2.1. Participants

Thirty-two native SC speakers (mean age: 26.3, SD: 3.6, 26 females) and thirty-one Indonesian L2 learners of SC (mean age: 20.9, SD: 0.7, 27 females) participated in this study.[1] Two SC speakers also reported speaking another Northern Mandarin variety. All L2 learners had learned SC for 2.5 to 3.5 years (2.5 years = 14 learners) and achieved ±2500 words (Intermediate level).[2] All participants reported normal to corrected vision and no speech and hearing disorders history.

2.2. Design

Table 1 illustrates the 4 x 2 within-subject design.

Factor	Level
Trial condition	1. critical 2. baseline
Appropriateness of coarticulatory cue	1. matched cue 2. mismatched cue
Magnitude of the anticipation (Tone type)	1. strong (T3) 2. weak (T1)
L1 background	1. tonal language 2. non-tonal language

Table 1: Experimental factors

Apart from the baseline and critical trials (66 each), fillers were added (164 trials) to balance the occurrence of items in all levels of trial condition (Appendix 1). These 296 items were presented in four blocks. We also included one practice block of 10 trials. Trial randomization was implemented within each block, and the position of target-competitor items was randomized and counterbalanced. Each item only appeared once in each block in either matched or mismatched cues.

2.3. Visual stimuli

The target NWs carried T3 or T1 followed by T3 (T3T3; T1T3) and had either a sonorant or a stop (/b/ or /d/) onset in the 1st syllable. As visual stimuli, they were typed in black KaiTi font and appeared with either a phonological competitor or a distractor. All items were controlled for visual complexity, measured in their number of strokes [$F(2,96)=0.731$, $p=0.484$], log morphemic frequency according to the SUBTLEX-CH corpus [18] [$F(2,96)=1.951$, $p=0.148$], and phonological neighborhood density based on the Chinese Lexical Database [19] [$F(2,96)=0.856$, $p=0.428$]. Furthermore, the goodness of the NWs was judged by fourteen native speakers of SC on a 4-point scale (1 meaning "the word is not a real word in SC" and 4 meaning "the word is a real word in SC"). We only included pairs where each member had an average value of less than 2.

2.4. Auditory stimuli

Bisyllabic NWs were recorded at the Phonetics Lab of the Leiden University Centre for Linguistics by one male native SC speaker, who was born and raised in Beijing. The NWs were produced in a carrier sentence, "*wo3 nian4* ___", and recorded in a randomized order three times. Two tokens of each word were selected to create spliced stimuli. The selection was based on accuracy and clarity in the patterns of anticipatory coarticulation and sandhi alternation (see Appendix 2). In total, 33 NW triplets were included in this experiment.

To manipulate the tonal coarticulatory information, the T3 in the second syllable was spliced onto the first syllable of another word with either an identical tonal context (i.e., matched cue) or a different tonal context (i.e., mismatched cue). Furthermore, the duration of the first syllable was normalized to 330ms, which was the mean duration of the first syllable of target-competitor NW pairs. The manipulation and normalization were done in Praat using customized scripts [20], [21]. The filler items were also manipulated in this way.

2.5. Procedures

Our participants were tested in a dim room and seated on a height-adjusted chair. A chin-rest was placed in front of an Eyelink Portable DUO (version 6.10.01) eye-tracker, which was situated central-horizontally from a 23-inch monitor. The task was created and presented using SR Research Experiment Builder (version 2.2.245) with auditory stimuli played using a soundcard (M-Track 2x2M) over a Beyer DT-770 dynamic headphone. The visual display was left-right symmetric on-screen against a white background.

Prior to the eye-tracking task, there was a training session to familiarize participants with the intended coarticulatory tonal variations and ensure they knew the pronunciation of all NWs stimuli. They listened to a list of real bisyllabic words, then asked to read aloud the NW stimuli, which was recorded. Only participants who produced all items as intended could begin the eye-tracking task. Native and non-native participants differed significantly in this session, with L1 speakers spending ±25 minutes and L2 learners 50-120 minutes to complete the task.

We included a 9-point calibration and validation procedure before the eye-tracking recording. Each trial began with a central fixation cross (500ms) with an invisible boundary for a drift check. Participants' gaze should be within the boundary to allow visual stimuli to be presented for 1000ms. Otherwise, recalibration and validation were enforced. After an auditory stimulus was played, participants were instructed to select the corresponding printed NW using a mouse click. For the entire session, written instructions were given in the native language of the respective groups, and participants were allowed to rest during the block transition. Nobody took a break of more than 5 minutes.

2.6. Data analysis

Data containing fixation samples and proportions were extracted in 20ms bin length using the Time Binning Report in SR research Data Viewer (version 4.1.1). We only included trials identified correctly by participants and fixation samples that fell on the target or competitor area region. The starting time-point of analysis was 200ms post stimulus onset to account for the oculomotor delay for programming a saccade [22]. While the ending time-point was where, approximately, the proportion of looks to target had reached the maximum, i.e., 1000ms (native listeners) and 1200ms (non-native listeners). For this report, statistical modeling for native and non-native data was performed separately to reduce model complexity. Patterns of fixation proportions across conditions over time for each group were modeled using generalized additive mixed modeling (GAMM) (*mgcv* package version 1.8-41) [23], and model plots were done in the *itsadug* package (version 2.4.1) [24] in R (version 4.2.0) [25]. The ineffectiveness of binomial distribution to account for autocorrelation led us to the continuous response distribution and we logit transformed the fixations proportions ([26, 27]).

Different smoothing parameters were implemented for the effects of our interest, which allowed significance tests from the conditional effects to be reported by pairs of parametric and smooth terms. The final model for each listener group included time as a covariate, trial condition (Trial), appropriateness of coarticulatory cue (Cue), the magnitude of anticipation (Tone) and their interactions as fixed effects, and by-participant factor smooth as a random effect. We also incorporated an AR(1) error model to reduce autocorrelation in the residuals. All models were fitted using a scaled-t family to alleviate the non-normal distribution of the residuals. *P*-values were adjusted using the Holm-Bonferroni method to control the family-wise error rate [28]. Model summary outputs of fixed effects are presented in Appendix 3.

3. RESULTS

Proportions and trajectories of target fixations varied over time, with the native listeners demonstrating a higher average proportion of fixations (POF) than the non-native listeners. The parametric and smooth terms in each group's model summary showed a significant effect of Trial ($p<0.01$). Specifically, the estimated average POF in critical trials was lower than in baseline trials and the difference smooth term contrasting the shape of fixation between the two conditions was far above 0. Last but not least, we also found a significant interaction of Trial and Tone ($p<0.05$), but only for native listeners.

As we zoomed into the critical trials, we found a common pattern shared by native and non-native listeners, i.e., higher target POF means in the matched-cue condition than in the mismatched-cue condition (visually depicted as a concave shape in **Fig. 1**). However, it was only in the T3 context that the difference was significant. Moreover, for the native listeners, the difference smooth term also reported a significant effect of Cue ($p<0.05$). Fixation trajectories between matched- and mismatched-T3 started to differ significantly at around 200ms before the offset of the 1st syllable (see **Fig. 1*a*)**. These patterns were not observed in the T1 context (see **Fig. 1*b*)**, which confirmed the significant effect of Tone ($p<0.05$) reported in the model summary. This difference between T3 and T1 context suggests that degrees of

anticipatory coarticulation affected how the coarticulatory cues were used.

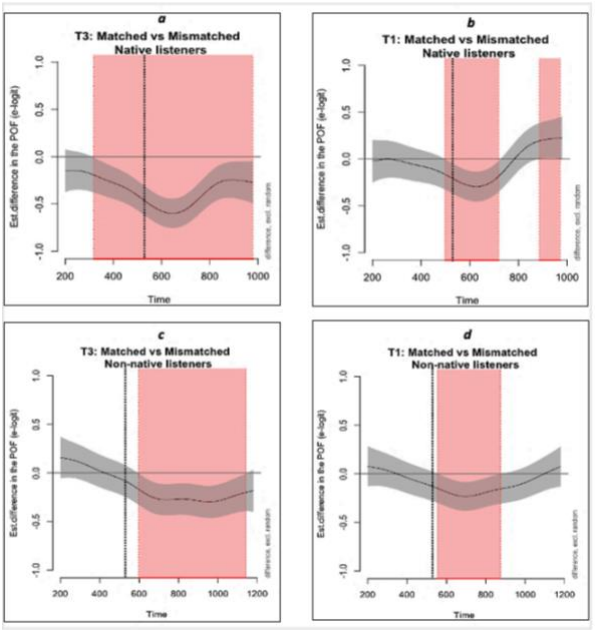


Fig. 1: Estimated difference smooth between matched- and mismatched-cue conditions in critical trials. Top panels: native listeners' T3 (***a***) and T1 (***b***). Bottom panels: non-native listeners' T3 (***c***) and T1 (***d***). The vertical shade indicates that a horizontal confidence band significantly differs from zero, that is, the interval of an effect. The vertical dotted line represents the first syllable offset (530ms).

As for the non-native listeners, the main effect of Cue was absent; looks between the matched vs. mismatched cue conditions became significantly different only after the offset of the 1st syllable, even when the anticipatory cues were salient (see the start of the vertical shade in **Fig. 1*c*** and **1*d***). The absence of the effect of Cue, and also Tone, indicates the difference between native and non-native listeners in their utilization of anticipatory tonal cues in online speech processing.

4. CONCLUSION AND DISCUSSION

Our results found significant differences between the baseline and critical trials, suggesting that the presence of phonological competitors in the critical trials significantly inhibited both native and non-native listeners from fixating on the target. Moreover, for the critical trials, both groups were slower in fixating on the targets and showed fewer looks when the auditory stimuli contained inappropriate anticipatory tonal cues. This pattern supports the prediction that listeners are sensitive to mismatched tonal coarticulatory cues, as evidenced by a delayed onset of target fixation and temporarily decreased looks to the target in a cue-mismatched condition.

Furthermore, the magnitude of anticipatory cues modulated listeners' utilization of anticipatory tonal information in online speech processing. In the T1 coarticulation context, neither group of listeners was able to use the cue to anticipate the upcoming tone in a timely manner (**Fig. 1*b*** and **1*d***). In the T3 sandhi context, native listeners were able to utilize the anticipatory cue at approximately 200ms before the end of the first syllable (**Fig. 1*a***), suggesting a predictive effect of the tonal cue. Our findings thus confirm the perceptual advantage of anticipatory tone sandhi cues for native listeners reported in [13].

Regarding non-native listeners, though they demonstrated a longer interval of difference effect in the T3 condition than the T1 condition (**Fig. 1*c*** and **1*d***), they could not use the T3 sandhi anticipatory cue predictively, like the native listeners (**Fig. 1*a*** and **1*c***). Another difference between native and non-native listeners is that non-native listeners had fewer target fixations for all conditions. This may be due to their less efficient (therefore slower) decoding for lexical or sentence processing [29, 30], and in this study, less efficiency in utilizing the tonal cues. The longer training session for the non-native group also suggests it required more cognitive effort to process the novel (nonce) words.

To sum up, both groups utilized anticipatory tonal cues during online speech processing as they fixated on a target more quickly in the cue-matching condition than in the cue-mismatching condition. However, non-native listeners did not show an immediate use of T3 sandhi cue to anticipate an upcoming tone, as native listeners did. The observed differences between native and non-native listeners in their ability to utilize anticipatory tonal cues confirm that the tonality of L1 prosody significantly impacts their processing efficiency of tonal information. In the future, it would be important to directly compare the native and non-native patterns within the same models, with a larger corpus. What is also needed is further research to investigate the effect of carry-over coarticulatory information on speech processing and the impact of different L1 tonal and non-tonal backgrounds on the utilization of such cues for speech processing.

5. ACKNOWLEDGMENTS

HNH is currently supported by the Indonesia Endowment Funds for Education (201903223013970). QY is currently supported by the China Scholarship Council (201707110012). YC is currently supported by the NWO Vici grant (vi.c.181.040). We thank Dr. Marjolein Fokkema for sharing her knowledge in GAMM analysis.

6. REFERENCES

[1] Martin, J. G., Bunnell, H. T. 1981. Perception of anticipatory coarticulation effects. *J. Acoust. Soc. Am*. 69(2), 559–67.

[2] Fowler, C. A., Brown, J. M. 2000. Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics* 62(1), 21–32.

[3] Dahan, D., Tanenhaus, M. K. 2004. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *J Exp Psychol Learn Mem Cogn* 30(2), 498–513.

[4] Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., Brasher, A. 2013. The time course of perception of coarticulation. *J. Acoust. Soc. Am*. 133(4), 2350–66.

[5] Xu, Y. 1997. Contextual tonal variations in Mandarin. *J. Phon.* 25, 61–83.

[6] Shen, X. S. 1990. Tonal coarticulation in Mandarin. *J. Phon.* 18(2), 281–295.

[7] Malins, J. G., Joanisse, M. F. 2010. The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *J. Mem. Lang.* 62, 407–420.

[8] Zou, T. 2017. *Production and perception of tones by Dutch learners of Mandarin* [Doctoral dissertation]. Leiden University.

[9] Yang, Q., Chen, Y. 2022. Phonological competition in Mandarin spoken word recognition. *Language, Cognition and Neuroscience* 37, 820–843.

[10] Zou, T., Liu, Y., Zhong, H. 2022. The roles of consonant, rime, and tone in Mandarin spoken word recognition. *Front. Psychol.* 12, 740444.

[11] Shen, J., Deutsch, D., Rayner, K. 2013. On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *J. Acoust. Soc. Am*.133(5), 3016–29.

[12] Qin, Z., Tremblay, A., Zhang, J. 2019. Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: An eye-tracking study. *J. Phon.* 73, 144–157.

[13] Sun, Y., Green, J., Shih, C. 2021. Anticipatory tonal variations can facilitate spoken-word recognition: A Visual World Eye-tracking study. *Book of Abstracts TAI 2021*, 210–211.

[14] Chen, Y. 2012. Message-related variation. In: Cohn, A., Fourgeron, C., Huffman, M. (eds.), *The Oxford handbook of laboratory phonology*. Oxford University Press, 103–114.

[15] Yuan J.H., Chen Y. (2014), 3rd tone sandhi in Standard Chinese: A corpus approach. *Journal of Chinese Linguistics* 42(1): 218–237.

[16] Magnuson, J. S. 2019. Fixations in the visual world paradigm: where, when, why?. *J Cult Cogn Sci* 3, 113–139.

[17] Qin. Z, Zhang, J. 2022. The use of tonal coarticulation cues in Cantonese spoken word recognition, *JASA Express Letters* 2, 035202.

[18] Cai, Q., Brysbaert, M. 2010. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS ONE* 5(6), e10729.

[19] Sun, C.C., Hendrix, P., Ma, J., Baayen, R. H. 2018. Chinese lexical database (CLD). *Behav Res* 50, 2606–2629.

[20] Pacilly, J. 2022a. CrossSplice.praat. [Praat script]

[21] Pacilly, J. 2022b. ModDuration.praat. [Praat script]

[22] Hallett, P. E. 1986. Eye movements. In: Boff, K. R., Kaufman, L., Thomas, J. P. (eds), *Handbook of perception and human performance* Vol. 1. Wiley Interscience, 10–11.

[23] Wood, S. N. 2020. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. https://cran.r-project.org/web/ packages= mgcv

[24] van Rij, J., Wieling, M., Baayen, R. H., van Rijn, H. 2022. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. https://CRAN.R-project.org/package=itsadug

[25] R Development Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/

[26] Porretta, V., Kyröläinen, A.-J., van Rij, J., Järvikivi, J. 2018. Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In: Czarnowski, I., Howlett, R., Jain, L. (eds), *Intelligent Decision Technologies 2017*. Smart Innovation, Systems and Technologies, vol 73. Springer, Cham.

[27] Ito, A., Knoeferle, P. 2022. Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behav Res,* 10.3758/s13428-022-01969-3. Advance online publication.

[28] Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.

[29] Duyck, W., Vanderelst, D., Desmet, T., Hartsuiker, R. J. 2008. The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review* 15, 850–855.

[30] Hopp, H. 2017. Individual differences in L2 parsing and lexical representations. *Biling: Lang. Cogn.* 20(4), 689–690.

[1] We excluded data from 3 native SC speakers born and grew up in the Southern part of China and 2 learners spoken Hakka or Hokkien in their households.

[2] The difference in the length of learning did not influence learners' eye-fixations patterns.

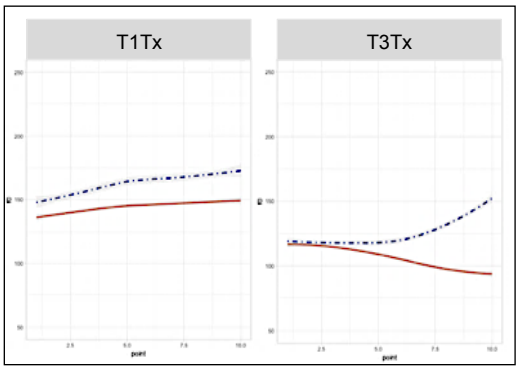
7. APPENDICES

Appendix 1. Stimuli presentation across trial types and example. Number in brackets indicates the original lexical tone of the second syllable before splicing.

Trial type	Visual Stimuli	Auditory Stimuli	
	Tar Com	Matched	Mismatched
Critical [1]	宾瓦 - 宾哇	*bin1$_{(3)}$wa3*	*bin1$_{(1)}$wa3*
Baseline [2]	宾瓦 - 垃副	*bin1$_{(3)}$wa3*	*bin1$_{(1)}$wa3*
Filler 1	宾哇 - 宾瓦	*bin1$_{(1)}$wa1*	*bin1$_{(3)}$wa1*
Filler 2	垃副 - 宾哇	*la1$_{(4)}$fu4*	*la1$_{(2)}$fu4*
Filler 3	当条 - 垃副	*dang1$_{(2)}$tiao2*	*dang1$_{(4)}$tiao2*

Note: [1] The target-competitor contrast is on the 2nd syllable (differ in tone). [2] The contrast lies on both syllables: 1st syllables differ in segments; 2nd syllables differ in segments and tone. The inclusion of baseline trials was aimed at creating a comparable design to that in a four-image VWP.

Appendix 2. Averaged f0 contours of T1 and T3 in target-competitor NW pairs produced by one speaker (n=34 for T1; n=32 for T3).



Tonal contrasts for each panel.
Left: T1T3 (dot-dashed); T1T1 (solid).
Right: T3T3 (dot-dashed); T3T2 (solid).

Appendix 3. Parametric and smooth terms of GAMM analysis of fixations proportion to targets.

		Native listeners				Non-native listeners			
		Est.	SE	*t*	*p*<	Est.	SE	*t*	*p*<
Parametric coefficients	(Intercept)	1.004	0.055	18.307	0.01	0.432	0.054	8.066	0.01
	Cue	-0.242	0.044	-5.495	0.05	-0.109	0.041	-2.694	0.05
	Tone	-0.199	0.044	-4.535	0.05	0.048	0.040	1.183	n.s.
	ToneCue	0.203	0.062	3.263	0.05	0.045	0.057	0.783	n.s.
	Trial	0.398	0.044	9.073	0.01	0.289	0.040	7.151	0.01
	TrialCue	0.185	0.062	2.974	0.05	0.117	0.057	2.044	n.s.
	TrialTone	0.191	0.062	3.062	0.05	0.020	0.057	0.355	n.s.
	TrialToneCue	-0.081	0.088	- 0.917	n.s.	-0.156	0.079	-1.963	n.s.
		edf	Ref.df	F	*p*<	edf	Ref.df	F	*p*<
Approximate significance of smooth terms	s(Time)	6.137	7.072	19.79	0.01	5.494	6.400	5.395	0.01
	s(Time):Cue	6.036	7.299	7.299	0.05	5.921	7.062	1.988	n.s.
	s(Time):Tone	2.269	2.913	2.871	0.05	2.676	3.430	0.912	n.s.
	s(Time):ToneCue	1.000	1.000	1.875	n.s.	3.019	3.880	1.176	n.s.
	s(Time):Trial	7.022	8.108	45.07	0.01	6.855	7.862	4.160	0.01
	s(Time):TrialCue	1.044	1.085	1.532	n.s.	5.541	6.729	2.646	n.s.
	s(Time):TrialTone	1.000	1.001	9.468	0.05	2.671	3.436	0.300	n.s.
	s(Time):TrialToneCue	1.014	1.026	0.416	n.s.	1.001	1.001	1.150	n.s.

We set "critical" as the reference level of Trial, "matched" as the reference level of Cue, and "Tone 3" as the reference level of "Tone". The first four rows of parametric and smooth terms refer to critical trials, while the last four rows refer to differences in effects (of Time, Cue, and Tone) between critical and baseline trials.

PREDICTING INTELLIGIBILITY FROM PRONUNCIATION DISTANCE METRICS

Tessa Bent and Rachael F. Holt

Indiana University; Ohio State University
tbent@iu.edu; holt.339@osu.edu

ABSTRACT

Unfamiliar native and non-native accents can cause word recognition challenges, particularly in noisy environments, but few studies have incorporated quantitative pronunciation distance metrics to explain intelligibility differences across accents. Here, intelligibility was measured for 18 talkers -- two from each of three native, one bilingual, and five non-native accents -- in three listening conditions (quiet and two noise conditions). Two variations of the Levenshtein pronunciation distance metric, which quantifies phonemic differences from a reference accent, were assessed for their ability to predict intelligibility. An unweighted Levenshtein distance metric was the best intelligibility predictor; talker accent further predicted performance. Accuracy did not fall along a native - non-native divide. Thus, phonemic differences from the listener's home accent primarily determine intelligibility, but other accent-specific pronunciation features, including suprasegmental characteristics, must be quantified to fully explain intelligibility across talkers and listening conditions. These results have implications for pedagogical practices and speech perception theories.

Keywords: Speech perception, intelligibility, non-native accents, regional accents

1. INTRODUCTION

Theories of speech perception must account for how listeners can recognize words amidst the variability present in speech signals [1, 2]. One factor that introduces substantial variability across talkers is a talker's accent, stemming either from regional differences or from influences of the first language when communicating in a second language. When these accent variants are unfamiliar, they can cause word recognition decrements [3, 4]. The challenge for understanding unfamiliar accents can be particularly acute in cases in which communication is occurring in the presence of background noise [5-8]. Although poorer accuracy has been observed for both unfamiliar native and non-native varieties, some research suggests that the pronunciation patterns found in unfamiliar non-native accents are more challenging to overcome than those in unfamiliar native varieties [9, 10]. However, most studies that include both native and non-native variants have been limited by including a small number of unfamiliar accents (e.g., [9, 10]). Therefore, it is difficult to determine the generalizability of the findings. [11] included a much wider range of accents but was limited by including only one talker per accent. Thus, there was a confound between the talker and accent variables. Studies that include both a wider number of accent varieties and multiple talkers representing each variety can provide greater insight into the factors that cause word recognition difficulties across native and non-native accents.

One open question in these studies is what specific talker or accent characteristics are leading to word recognition decrements. There is not a consensus regarding methods for quantifying how the speech differs from the listener's home accent. Researchers have taken different approaches to this problem. One approach is to focus on the impact of very specific production features (e.g., VOT differences [12]). Another approach is to synthetically modify the speech so that specific dimensions are more or less native-like [13, 14]. Work using this approach suggests that while keeping phonemic properties constant, changing the rhythmic properties of non-native speech to be more native-like increases intelligibility [13] but changing intonation patterns to be more native-like does not [14]. These approaches have provided insight into which dimensions of speech impact intelligibility for non-native talkers, but they do not allow for understanding how naturally produced interactions among multiple phonemic differences or between segmental and suprasegmental features may impact intelligibility. Finally, some researchers have incorporated general descriptions of the accents included (e.g., vowel differences across regional variants [8] or descriptions of general phonemic and suprasegmental differences across varieties [9]) but fell short of incorporating quantitative pronunciation metrics into their statistical modelling.

A recent approach to understanding how phonemic variability across accents impacts speech perception has been to incorporate edit-distance metrics into investigations of non-native accent strength ratings [15-17] and intelligibility [11, 18]. All these metrics measure pronunciation distance for

a speaker relative to a reference set (e.g., distance of a non-native talker to a specific native variant) using phonemic transcription. Several variations have been evaluated in relation to non-native accent strength ratings including Levenshtein Distances, Native Discriminative Learning (NDL) and Pointwise Mutual Information (PMI) metrics [15-17]. These metrics have shown substantial promise in accounting for variance in accent strength ratings across talkers. An advantage of the NDL and PMI metrics is that they incorporate the frequency of pronunciations within a corpus such that there are greater penalties for less-frequent pronunciation patterns, for which listeners are likely to assign higher accent strength ratings. However, these metrics require large amounts of data to calculate (e.g., hundreds of talkers), limiting their utility for most researchers.

Few investigations have incorporated these pronunciation distance metrics into studies of intelligibility. One study used the Levenshtein distance metric to characterise the talkers' utterances but did not incorporate the distance scores into the statistical models [18]. [11] incorporated the Levenshtein distance metric to model intelligibility of seven different accents. Levenshtein scores were a significant predictor of intelligibility in both quiet and noise-added conditions. Furthermore, a model including both Levenshtein distances and talker accent was a better fit than the one using only the Levenshtein distances. However, they only used one talker per accent, thus confounding talker- and accent-specific effects. Here, we build on [11] by including two more accents than used previously and, more importantly, including two talkers per accent. Incorporating more than one talker per accent will begin to address whether variability in intelligibility can be traced to pronunciation features that are characteristics of a specific accent or whether effects that have been described as "accent effects" are talker-level effects. Finally, we evaluate two versions of the Levenshtein distance metric to determine whether the weighted version used in [11] and [18] explains more variability in intelligibility than an unweighted version used in [19].

2. METHOD

2.1. Participants

There were 370 American English monolingual listeners between the ages of 18 – 35 years (average = 26). All listeners had self-reported typical speech, language, and hearing abilities. Midland American English was familiar to all participants due to its similarity to Standard American English. Participants who reported daily exposure to one of the other accents in their condition were excluded.

2.2. Stimuli

Sixty sentences from the Hearing in Noise Test for Children [20] were produced by 18 talkers representing 9 different accents: three native (Midland American English, Southern Standard British English, Scottish English), five non-native (French-, Spanish-, German-, Japanese-, and Mandarin-accented English), and one bilingual (Hindi-Indian English) variety. Two speakers (1 female and 1 male) represented each variety.

2.3. Procedure

Listeners were recruited through Prolific (https://www.prolific.co/) and tested online, using Pavlovia [21]. Prior to the intelligibility experiment, participants completed a consent form, a background questionnaire, and a headphone screening [22]. Listeners were presented with sentences from three talkers of the same gender including a Midland American English talker and talkers with two less familiar accents, each contributing 20 sentences. Listeners were randomly assigned to one of three listening conditions: quiet, +4 dB signal-to-noise ratio (SNR), or 0 dB SNR. The noise was an 8-talker babble with talkers matched in gender to the target speech. A randomly selected section of the babble file that was 1 second longer than the sentence was selected as the masker for each item. For each accent / listening condition combination, 15–18 participants were tested.

Listeners were presented with 9 practice trials, including three from each talker in the experimental trials. These were presented in the same listening condition as the experimental trials. Experimental trials were blocked by talker and randomized within a block. Participants were instructed to listen to a sentence and type in what they heard. They were not provided with any feedback.

2.4. Analysis

2.4.1 Levenshtein distances

All sentences were phonemically transcribed by two trained research assistants. These transcriptions were compared, and any discrepancies were resolved by discussion with a third transcriber. The transcriptions for the non-Midland speakers were compared to four talkers representing the familiar Midland American English referent for the calculation of two different versions of the Levenshtein distance metric: weighted and unweighted. In the unweighted version [19], the

target stimulus is compared to the reference and any phonemic difference between the two (i.e., addition, deletion, substitution) is given a one-point penalty. These penalties are then summed and divided by the total possible number of phonemes. This total is determined by the maximum logical alignment of the two strings (Table 1). The weighted Levenshtein distance [18] assigns different penalties depending on the type of difference between the two strings which are then summed at the word level:

- Vowel substituted by a vowel = 0.5
- Consonant substituted by a consonant = 0.75
- Phoneme insertion = 1.0
- Change to word length = 1/log10(max(length(Word1), length(Word2)))
- Other (e.g., deletion, vowel to consonant substitution, consonant to vowel, etc.) = 0.4

These weightings were developed based on theoretical assumptions from prior literature, but this metric has not been compared to other Levenshtein variants to determine whether these weightings add further explanatory power.

Levenshtein scoring	
Target word	*stopped*
Midland American accent	Ø s t ɑ p t
	| | | | | |
Spanish-accented English	ɛ s t ʌ p Ø
Unweighted Levenshtein scores	1 0 0 1 0 1
Final score	3/6 or .5
Weighted Levenshtein scores	1 0 0 .5 0 .4
Final score	1.9

Table 1: Unweighted and weighted Levenshtein scoring examples

2.4.2 Intelligibility scoring

Intelligibility scores were determined at the sentence level using Fuzzy String Matching, specifically the token sort ratio from [23] using the online implementation (https://tokensortratio.netlify.app). These scores range from 0 to 100. Sentences that match the target exactly are given a score of 100 and those without any matching characters are given a 0. For example, for the target sentence "Mother read the instructions", a response of "motherly subscriptions" received a score of 36, "follow the instructions" 63, "read the instructions" 86 and "mother read the instruction" a 98. These scores tend to be higher than percent words correct scores with a strict scoring criterion in which the examples above would have received scores of 9, 50, 75, and 75, because responses are given credit for partial matches. However, [23] showed that scores obtained via Fuzzy String Matching are highly correlated with traditional hand calculated measures of percent words correct.

3. RESULTS

3.1 Weighted vs. unweighted Levenshtein scores

We first evaluated whether the weighted or unweighted Levenshtein distance scores were a better predictor of intelligibility. Intelligibility score for each sentence was entered into a linear mixed effects model with fixed effects of Levenshtein score and SNR and their interaction with random intercepts for participant and item as well as by-participant and by-item varying slopes for talker accent. Two models were built, one with the weighted and one with the unweighted Levenshtein scores. From Type III Wald chi-square tests, both models showed significant effects of the Levenshtein scores, SNR, and their interaction (all p-values < 0.001). Therefore, we computed model comparisons to determine which Levenshtein score resulted in a better fit for the intelligibility scores. The model comparisons showed that the AIC was lower for the model with the unweighted (179969) than the weighted (179972) Levenshtein scores; thus, the unweighted Levenshtein scores were a better fit for predicting intelligibility. In further modelling, we employed the unweighted scores.

3.2 Levenshtein scores and talker accent

We next investigated whether talker accent contributed to intelligibility scores (Figure 1). In this model, we included fixed effects of unweighted Levenshtein scores, SNR, and talker accent, as well as their interactions. The model also included random intercepts for participant and item as well as by-participant and by-item varying slopes for talker accent. The results are shown in Table 2.

	F-value	*p*-value
Levenshtein	35.0	< 0.001
SNR	153.9	< 0.001
Accent	23.2	< 0.001
Levenshtein x SNR	2.1	n.s.
Levenshtein x Accent	11.3	< 0.001
SNR x Accent	29.7	< 0.001
Levenshtein x SNR x Accent	2.9	< 0.001

Table 2: Output of Type III Analysis of Variance Table with Satterthwaite's method for full model of intelligibility

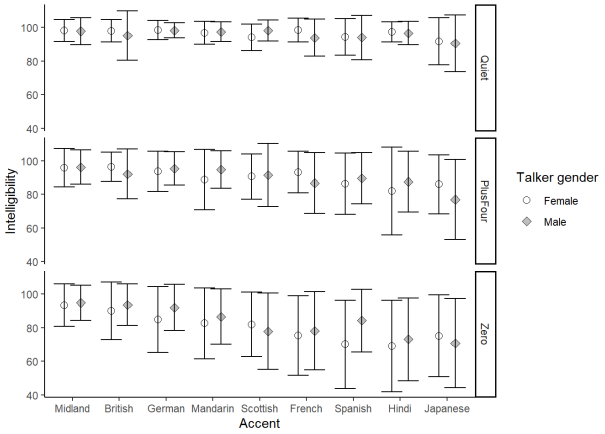


Figure 1: Intelligibility scores by accent and talker across the three listening conditions.

All three main effects were significant. The main effect of Levenshtein scores arose because items with higher Levenshtein scores were less intelligible than those with lower Levenshtein scores. That is, productions that diverged more from the local accent were more difficult for listeners to understand than those closer to the local dialect. The main effect of SNR resulted from highest accuracy in quiet and lowest accuracy in the 0 dB SNR with intermediate performance in the +4 dB SNR condition. The main effect of accent was due to the differences in intelligibility across accents with highest accuracy for the Midland American English and the Southern Standard British English accents and lowest accuracy for the Hindi and Japanese accents. The two-way interaction between Levenshtein distance and SNR was not significant suggesting that the effect of pronunciation distance from the local accent was similar across listening conditions. The other two-way and the three-way interaction were significant. The SNR by Accent interaction arose because some accents were highly intelligible even in the most difficult SNRs (e.g., Midland and British) whereas other accents showed much larger intelligibility declines, particularly in the most difficult SNR. The Levenshtein by Accent interaction and the three-way interaction among Levenshtein, Accent, and SNR suggest that effects of phonemic distance from the local accent were not consistent across accents and further, that these impacts were influenced by listening environment.

4. DISCUSSION

This study investigated intelligibility in quiet and two noise-added conditions across nine accents with two talkers representing each accent. The explanatory value of two variants of the Levenshtein distance metric, which measures phonemic differences from a reference accent, were evaluated. Results showed that higher Levenshtein distance scores were associated with lower intelligibility. Further, the unweighted Levenshtein metric was a better fit for the data than the weighted one. This result further supports the inclusion of these pronunciation distance metrics into studies of intelligibility to capture how differences in segmental productions across both unfamiliar native and non-native accents can impact intelligibility. They also suggest that relatively simple metrics (i.e., unweighted scores that are not adjusted for frequency of pronunciations within a large corpus) may be sufficient to capture the impact of phonemic differences on intelligibility. That said, the continued evaluation of other distance metrics (e.g., PMI, NDL) would still be valuable to determine the extent to which they can capture variance in intelligibility across talkers and accents, beyond their utility with explaining accent strength ratings [15-17]. This study extends the use of these pronunciation distance metrics to both native and non-native varieties whereas most previous studies (except [11]) used only non-native varieties. These results also cast doubt on any strong claims about non-native accents causing more disruption to word recognition than unfamiliar native ones.

These distance measures do not capture any pronunciation differences beyond the phoneme level, such as prosodic differences. Fully explaining why a specific talker or accent causes word recognition challenges, particularly for stimuli longer than a word, will very likely require measures of prosodic characteristics, such as rhythm, intonation, and speaking rate. Indeed, results in the full model suggest that intelligibility is impacted by interactions among phonemic distance, accent features, and listening conditions. Establishing standardized measures of prosodic distance from the home accent will be an essential next step in explaining the source of the accent effects. Similarly, a finer-grained analysis is necessary to interpret the impacts of listening condition (i.e., SNR). As has been seen for regional American English accents [8] and non-native accents [7], intelligibility differences across varieties tend to be much larger in adverse listening conditions. Additional analyses of the current data set will allow for pinpointing which pronunciation characteristics underlie the challenges that arise only in noisy environments. Finally, future studies should also include other social or sociolinguistic factors, such as listener familiarity with accents or language attitudes, to determine their impact on communication success.

5. ACKNOWLEDGEMENTS

We would like to thank our research assistants who supported this project: Lindsey Altum, Payton Bastie, Jessica Bell, Megan Hancock, Malachi Henry, Holly Lind-Combs, Yi Liu, Alondra Rodriguez, Ali Stallons, and Amy Warrington. We are grateful for the contributions of Lilian Golzarri Arroyo, who provided statistical consulting support, and for our funding from the National Science Foundation (Award Numbers: 1941691 and 1941662).

6. REFERENCES

[1] Pierrehumbert, J. B. 2016. Phonological representation: Beyond abstract versus episodic. *Annu. Rev. Linguist*. 2, 33–52.

[2] Kleinschmidt, D. F., Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev*. 122, 148–203.

[3] van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T. 2002. Quantifying the intelligibility of speech in noise for non-native listeners. *J. Acoust. Soc. Am*. 111, 1906–1916.

[4] McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A. 2018. Van Engen, K. J. Coping with adversity: Individual differences in the perception of noisy and accented speech. *Atten. Percept. Psycho.* 80, 1559–1570.

[5] Rogers, C. L., Dalby, J., Nishi, K. 2004. Effects of noise and proficiency on intelligibility of Chinese-accented English. *Lang. Speech* 47, 139–154.

[6] Munro, M. J. 1998. The effects of noise on the intelligibility of foreign-accented speech. *Stud. Second Lang. Acq.* 20, 139–154.

[7] Wilson E. O., Spaulding, T. J. 2010. Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *J. Speech Lang. Hear. R*. 53, 1543–1554.

[8] Clopper, C. G., Bradlow, A. R. 2008. Perception of dialect variation in noise: Intelligibility and classification. *Lang. Speech* 51, 175–98.

[9] Bent, T., Baese-Berk, M., Borrie, S. A., McKee, M. 2016. Individual differences in the perception of regional, nonnative, and disordered speech varieties. *J. Acoust. Soc. Am*. 140, 3775–3786.

[10] Adank, P., Evans, B. G., Stuart-Smith, J., Scott, S. K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol. Human* 35, 520–529.

[11] Bent, T., Holt, R. F., Van Engen, K. J., Jamsek, I. A., Arzbecker, L. J., Liang, L., Brown, E. 2021. How pronunciation distance impacts word recognition in children and adults. *J. Acoust. Soc. Am*. 150, 4103–4117.

[12] Sumner, M. 2011. The role of variation in the perception of accented speech. *Cognition* 119131–136.

[13] Tajima, K., Port, R., Dalby, J. 1997. Effects of temporal correction on intelligibility of foreign-accented English. *J. Phonetics* 25, 1–24.

[14] Sereno, J., Lammers, L., Jongman, A. 2016. The relative contribution of segments and intonation to the perception of foreign-accented speech. *Appl. Psycholinguist*. 37, 303–322.

[15] Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., Nerbonne, J. 2014. Measuring foreign accent strength in English. *Lang. Dynamics Change* 4253–269.

[16] Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., Baayen, R. H. 2014. A cognitively grounded measure of pronunciation distance. *PloS One* 9, e75734.

[17] Bartelds, M., Richter, C., Liberman, M., Wieling, M. 2020. A new acoustic-based pronunciation distance measure. *Fr. Art. Int.* 3, 39.

[18] Levy, H., Konieczny, L., Hanulíková, A. 2019. Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience. *J. Child Lang.* 46, 368–392.

[19] Gooskens, C., Heeringa, W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Lang. Var. Change* 16, 189–207.

[20] Nilsson, M., Soli, S. D., Gelnett, D. J.1996. *Development of the Hearing in Noise Test for Children (HINT-C)*. House Ear Institute.

[21] Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. K. 2019. PsychoPy2: Experiments in behavior made easy. *Behav. Res.* 51, 195–203.

[22] Woods, K. J. P., Siegel, M. H., Traer, J., McDermott, J. H. 2017. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psycho.* 79, 2064–2072.

[23] Bosker, H. R. 2021. Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies *Behav. Res*. 53, 1945–1953.

NONNATIVE ACCENT ADAPTATION IN THE INITIAL MOMENTS AND OVER A MONTH

Xin Xie[1] & Chigusa Kurumada[2]

[1]Department of Language Science, University of California, Irvine;
[2]Department of Brain and Cognitive Sciences, University of Rochester
xxie14@uci.edu, ckuruma2@ur.rochrster.edu

ABSTRACT

Perceptual difficulties associated with unfamiliar talkers or accents are known to dissipate—sometimes within minutes of exposure. How do the adaptive changes in perception evolve beyond these initial moments of encounter? The current study examined incremental changes in native listeners' recognition of L2-accented speech (Mandarin-accented English) over two timescales: within the first few minutes (Exp.1) and across five sessions spanning a month (Exp.2). We developed a new repeated exposure-test paradigm to track the trajectory of recognition improvements for a category that was initially confusable with another (i.e., word-final /d/ as in *feed* sounding like /t/ as in *feet*). L1 listeners' adaptation was detected as early as after ~30 instances of exposure and continued to evolve over the subsequent sessions, suggesting a common process guiding the rapid adaptation to, and the longer-term accommodation of, an initially unfamiliar accent.

Keywords: Speech perception, nonnative (L2) accents, short- and long-term adaptation, repeated and incremental testing, computational modelling

1. INTRODUCTION

How listeners navigate the substantial amount of cross-talker variability is a central question in speech perception. The "same" phoneme or word is produced with distinct acoustic-phonetic properties across talkers with different characteristics (e.g., height, gender, accent). This variability is known to make the recognition of an unfamiliar talker or accent difficult, [1], [2]. Such difficulties can, however, dissipate as listeners adapt to the current input [3]–[6]. For example, native listeners of English become faster and more accurate in responding to Spanish- and Mandarin-accented English after initial exposure to these previously unfamiliar accents within as few as two to four sentence-length utterances [4], [7].

Alongside the short-term adaptation, real-world speech recognition and accent adaptation also evolve over episodes of social interaction across talkers and contexts [8], [9]. Long-term experiences with an accent gained over months and years have been reported to facilitate the comprehension of, and adaptation to, a novel talker from a similar accent background e.g., [10], [11]. However, environmental exposure to a previously unfamiliar accent alone does not always lead to a significant change of perception [12]. Currently, little is known about the nature and the amount of exposure needed to support the L1-/L2-accent accommodation. Questions also remain open about whether rapid adaptation seen in the initial moments of encounter and the longer-term accent accommodation draw on a single or multiple distinct mechanism(s) [13].

This is the knowledge gap the current study set out to address. Specifically, we asked two questions: (Q1) ***How rapid is the adaptation?*** In the initial moments of encountering an *a priori* unfamiliar accent, how do the recognition speed and accuracy change? (Q2) ***How does the adaptation continue?*** In response to repeated exposure to the same (*a priori*) unfamiliar talker or accent, do the effects of exposure accumulate over time? Or alternatively, does adaptation begin anew each time?

Our current experiments built upon Xie et al. (2017) [15], who used a typical exposure-test design to examine L1-English listeners' adaptation to a Mandarin-accented word-final /d/-/t/ contrast in English. During exposure, Xie et al. (2017) used a lexical decision task where a target group of native listeners heard words produced by a Mandarin-accented talker. Importantly, some words included /d/ in a final position (e.g., *lemonade*), which is initially confusable with /t/ (i.e., *lemonade* sounding like *lemonate)*. In contrast, control group heard speech from the same Mandarin-accented talker but no /d/-final words. At test, both groups provided phoneme categorization responses to /d/-/t/ minimal pairs (e.g., *feed-feet*) produced by the same Mandarin-accented talker. Changes in the categorization responses between the two groups—elicited by the critical exposure to /d/-final words in the L2 accent (target group) or lack thereof (control group)—were taken as evidence for exposure-elicited adaptation to the L2 accented talker.

Extending this, we developed a new repeated testing paradigm in which listeners undergo three test blocks within a session (Fig. 1, top). This allows for finer-gained observations of adaptative changes in the

recognition of unfamiliar speech inputs (Q1, Exp(eriment) 1). To examine long-term development of adaptive speech perception (Q2, Exp. 2), this repeated exposure-test design was administered five times spanning four weeks (Fig. 1, bottom). The test items were repeated across multiple blocks/days to increase the power of tracking trajectories of adaptation both within and across sessions.

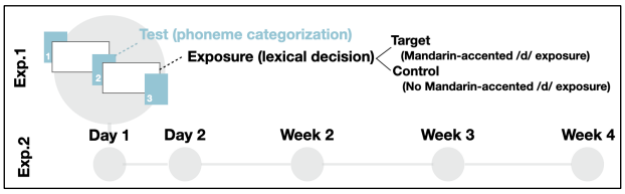


Figure 1 Exp. 1 consisted of three test blocks (10 items each) interspersed with two exposure blocks (90 items each) within a single session. Exp. 2 repeated the procedure from Exp. 1 five times: Day 1, one day later (Day 2), one week later (Week 2), two weeks later (Week 3), and three weeks later (Week 4).

2. EXPERIMENTS

Using the new repeated testing paradigm described above, we examined native English listeners' adaptation to L2- (Mandarin-)accented English.

2.1. Experiment 1 (One day)

Following the design of Xie et al. (2017), we compared recognition of L2-accented speech during test from listeners who were exposed to the critical /d/-final words vs. those who did not (the "target" vs. the "control" groups, respectively). Everything else was identical across the exposure conditions.

2.1.1. Participants

55 undergraduate students, all native speakers of American English, were recruited from the UC Irvine community and completed the experiment online via an online testing platform FindingFive. They were randomly assigned to the control exposure condition (n = 30) and target exposure condition (n = 25).

2.1.2. Stimuli

Both exposure and test stimuli were produced by a male native-Mandarin speaker of medium intelligibility in English. (The intelligibility was assessed by a separate norming study.) **Exposure stimuli** for the target group consisted of 90 English words (30 critical and 60 filler items) and 90 phonologically-legal nonwords. The critical items were all multi-syllabic words ending in /d/ (e.g., *lemonade, overload*). The exposure list for the control group was identical except that the 30 /d/-final critical items were replaced by 30 filler items. Filler words and nonwords did not contain any /d/ or /t/ sounds; no stop sounds other than /d/ appeared in word-final position. The exposure items were evenly distributed across the two exposure blocks.

Test stimuli consisted of five /d/- or /t/-final minimal pairs (e.g., *feed-feet*). The smaller number of test stimuli (8% of those used in Xie et al., 2017) was motivated by the current repeated testing paradigm. Because listeners go through multiple test blocks, presenting a large number of test items would run the risk of providing additional information about the target contrast and thereby neutralizing between-group differences (i.e., listeners in the control group may be able to adapt their recognition of the target contrast through the exposure to the test items alone). In addition, there is a growing awareness that the statistical information of the test tokens can attenuate (or potentially override) the effects of the exposure [16], [17]. Ideally, therefore, both exposure and test tokens should be sampled from the same natural cue distributions that characterize the target contrast.

Given these constraints, we applied a new method for sampling test tokens according to their acoustic-phonetic features. Specifically, we chose five minimal pairs based on: (a) Bayesian ideal observer simulations [18] and (b) human listener responses [15]. In (a), we simulated distributional learning of the underlying L2-accented categories along three cue dimensions critical for word-final /d/-/t/ contrast (Fig. 2). By contrasting the expected recognition accuracy after the L2- accented exposure (simulating the target group) vs. L1-accented exposure (simulating control participants who do not have exposure to L2-accented /d/), we computed the relative advantage expected from the L2-accented exposure (shown by the colorscale in Fig. 2). We then identified test items that were predicted to receive distinct categorization responses if native-English listeners indeed update their internal representations of /d/ vs. /t/ categories to match those in the L2-accented exposure input. In (b), we further narrowed our choices down to those test items that are *a priori* ambiguous to native listeners to avoid ceiling effects.

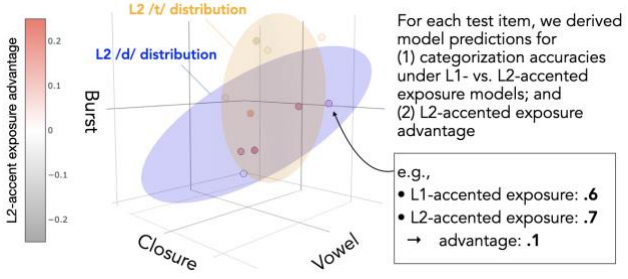


Figure 2: Model predictions derived for the test items. The color of the dots indicates the degree of advantage expected from L2-(vs. L1-) accent exposure for individual test items.

2.1.3. Procedure

During exposure, participants completed a lexical decision (i.e., word vs. non-word) task. During test, participants provided 2AFC phoneme categorization responses to the minimal pair words (e.g., *feed-feet*). All participants completed an exit questionnaire about their language background and familiarity with L2 accents.

2.1.4 Results

Linear mixed effect logistic regression models were fit to the test data using the *lme4* package in R [19]. (Exposure) condition, test block, and category (/d/ vs. /t/) were examined as fixed effects. All models were specified with the maximal random effects structure justified by the design. Condition and category were sum contrast coded; test block was Helmert coded to compare (i) test2 vs. test1 and (ii) test3 to the mean of test1 and test2.

As predicted, /d/ words were recognized less accurately than /t/ words overall (β = -1.29, SE = .35, z = -3.66, p = .0002). A significant interaction between test block and category suggested that the improvement in test3 over the previous two tests was larger for /d/ and /t/ (β = .21, SE = .09, z = 2.26, p = .02). There was a marginally significant three-way interaction between condition, test block, and category (β = .18, SE = .09, z = 1.89, p = .06): Simple effects analyses revealed that only the target group significantly improved their recognition of /d/ (β = .52, SE =.15, z = 3.36, p = .0008).

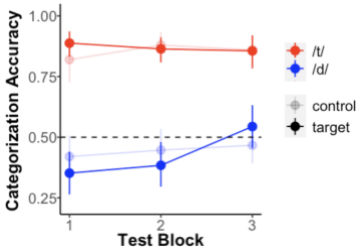


Figure 3: Results from Exp. 1 (1 day). Y-axis shows proportion of correctly categorized tokens in each exposure condition, over the three test blocks. Error bars represent 95% confidence intervals.

2.2. Experiment 2 (Five days spanning a month)

Exp. 2 repeated the same exposure-test session in Exp. 1 five times with one important change. In this experiment, the control condition used the stimuli identical to the target condition, except that they were produced by a native-English speaker (i.e., L1-accented exposure). This follows the design adopted by many influential studies on accent adaptation (e.g., [4]-[6], [20]), and it allowed us to keep constant the identities of words and the number of /d/ words heard across the two exposure conditions. Notably, neither of the target groups in Exp. 1 and 2 received exposure to Mandarin-accented /d/ (or /t/) sounds and therefore lacked the critical exposure needed for adaptation.

2.2.1. Participants

70 native speakers of American English, aged 18-45, were recruited via Prolific and completed the experiment online via FindingFive. Participants were randomly assigned to the L1-accented exposure condition (control, n = 38) and L2-accented exposure condition (target, n = 32).

2.2.2. Stimuli and procedure

The stimuli materials in both conditions were identical to those used in the target condition in Exp.1. Five exposure-test sessions, each identical to that of Exp. 1, were administered across five time points over the course of a month (Fig. 2).

2.3.3. Results

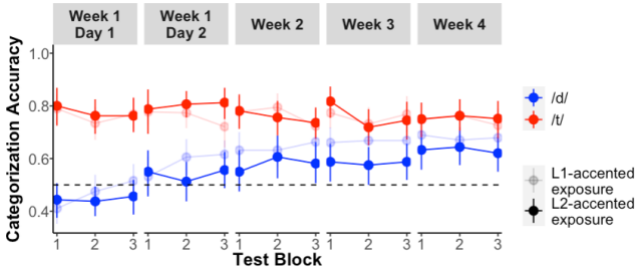


Figure 4: Results from Exp. 2 (multi-session spanning four weeks), collapsed across exposure conditions. Y-axis shows proportion of correctly categorized tokens within and across experimental sessions. Error bars represent 95% confidence intervals.

Analysis procedures were identical to those in Exp. 1, except that test block was recoded as a continuous variable (i.e., a total of 15 test blocks over five test sessions). This enabled us to examine whether test performance continued to change cumulatively across all test blocks. Both groups' overall performance improved significantly across time (β = .04, SE = .006, z = 6.63, p < .0001). Similar to Exp. 1, a test block-by-category interaction was found (β = .06, SE = .006, z = 9.54, p < .0001). This interaction was driven by an increase in recognition accuracy for /d/ (β = .10, SE = .008, z = 11.92, p < .0001) and a smaller decrease for /t/ (β = -.02, SE = .009, z = -1.95, p = .052). Unlike in Exp. 1, however, no three-way interaction was found: Both the target and the control groups in Exp. 2 improved their recognition of the L2-accented categories to a similar degree.

3. DISCUSSION

The current study combined two lines of work on native listeners' processing of L2-accented speech: (1) rapid adaptation in the initial moments of encountering an (*a priori*) unfamiliar talker/accent, and (2) longer-term accommodation of an accent through repeated exposure. Replicating previous findings, Exp. 1 demonstrated that listeners did indeed improve their perception within the first few minutes of exposure. The new repeated testing paradigm revealed detectable exposure effects *after* the second exposure block. This supports the idea that "rapid" adaptation still requires some amount of cumulative input from which listeners can extrapolate distributional structures of acoustic-phonetic cues characterizing the unfamiliar accent.

The results of Exp. 2 presented an intriguing puzzle. As expected, repeated exposure to a Mandarin-accented talker helped native listeners improve their recognition of the /d/ category of that talker without a corresponding decrease in the accuracy for the /t/ category, resulting in an increase in overall accuracy (from .61 to .69). We demonstrate, for the first time, that adaptation to naturally produced L2-accented categories continues to evolve over a month (see also [22], [23] for shorter-term adaptation to simpler stimuli).

Unexpectedly, control participants who received L1-accented exposure also achieved a comparable level of adaptation. By-item analysis confirmed that both groups responded to each of the five minimal pairs in a similar manner throughout the five sessions. In short, while the target group consistently benefited from exposure to L2-accented /d/ sounds over both timescales, the behavioural patterns of the control groups differed between Exp. 1 and Exp. 2.

What caused this difference? One might attribute it to differences in participant profiles, namely the campus population in Exp. 1 and the Prolific participants in Exp. 2. However, the fact that the participants had comparable recognition accuracy for /d/ and /t/ in the very first test block makes this an unlikely reason for the observed difference. Rather, it could be due to the number of /d/ tokens, which differed between the experiments. Recall that in Exp. 2, the control group received as many /d/ tokens as the target group did albeit from an L1-accented talker. This exposure, combined with the small number of minimal pair test items repeated across test blocks, may have drawn their attention to the /d/-/t/ contrast and led the control group to derive some strategic response patterns to improve their performance (e.g., I have responded /t/ to a similar item already, so I will respond /d/ to the current item).

Another, non–mutually exclusive, possibility is that *the test items alone* may have been sufficient to support the adaptation. i.e., Listeners in both the control and target conditions may have adapted to the accent-specific features through the exposure to the /d/-/t/ minimal pairs without labeling information. If this is indeed the case, it means that the L2-accented /d/ words heard during exposure did not provide any additional benefit to the target group. To address this, our future follow-up tests will include a condition in which listeners are tested on the same schedule as in the current study (= three test blocks / session * five sessions over a month) without any L2-accented exposure in between. If L1-listeners improve their recognition of the /d/-/t/ contrast in this condition, that would suggest that unsupervised adaptive changes in perception can happen faster and be more efficiently than has previously been believed.

More generally, the current data highlight the complexity of identifying the mechanism(s) for adaptive changes of recognition. Is there a single mechanism that explains adaptation across different timescales? Did participants in the control vs. target conditions use the same or different mechanism(s) to improve their recognition? By simply examining behavioral data aggregated over trials and subjects, as we did for the current study, we cannot know **how**—by what mechanism—the observed adaptive changes occurred.

To address this question, we have recently proposed a combinational framework (called ASP for "Adaptive Speech Perception") [24]. ASP instantiates how listeners' responses may adapt to exposure based on: (a) changes in auditory perception (e.g., low-level signal transformation and normalization), (b) changes in linguistic representations (e.g., distributional learning of phoneme categories), and (c) changes in decision making (e.g., updating of decision biases). Pursued in independent lines of work, these three mechanisms have so far rarely been contrasted or tested against each other using the same data.

ASP allows researchers to model adaptive changes in responses under any (combination) of the three mechanisms. This, in turn, allows us to select exposure and test items for human perception experiments in a targeted manner to more effectively contrast the predictions of the different mechanistic models. The new incremental testing paradigm developed in the current study will facilitate this approach: The improved ability to sample human responses to different types and amounts of exposure will increase the statistical power for reliable model comparisons. The paradigm, combined with model-based hypothesis-testing, thus holds promise to provide new insights into adaptive speech perception.

5. REFERENCES

[1] P. Adank and E. Janse, 'Comprehension of a novel accent by young and older listeners.', *Psychol Aging*, vol. 25, no. 3, pp. 736–740, Sep. 2010, doi: 10.1037/a0020054.

[2] V. Porretta, B. V. Tucker, and J. Järvikivi, 'The influence of gradient foreign accentedness and listener experience on word recognition', *J Phon*, vol. 58, pp. 1–21, Sep. 2016, doi: 10.1016/j.wocn.2016.05.006.

[3] M. M. Baese-Berk, D. J. McLaughlin, and K. B. McGowan, 'Perception of non-native speech', *Lang Linguist Compass*, vol. 14, no. 7, pp. 1–20, 2020, doi: 10.1111/lnc3.12375.

[4] X. Xie *et al.*, 'Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker', *J Acoust Soc Am*, vol. 143, no. 4, pp. 2013–2031, Apr. 2018, doi: 10.1121/1.5027410.

[5] A. R. Bradlow and T. Bent, 'Perceptual adaptation to non-native speech.', *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[6] C. Y. Tzeng, J. E. D. Alexander, S. K. Sidaras, and L. C. Nygaard, 'The role of training structure in perceptual learning of accented speech', *J Exp Psychol Hum Percept Perform*, 2016, [Online]. doi.org/10.1037/xhp0000260%5Cnhttp://

[7] C. M. Clarke and M. F. Garrett, 'Rapid adaptation to foreign-accented English.', *J Acoust Soc Am*, vol. 116, no. 6, pp. 3647–3658, 2004, doi: 10.1121/1.1815131.

[8] K. B. McGowan, 'Social expectation Improves speech perception in noise', *Lang Speech*, vol. 58, no. 4, pp. 502–521, Feb. 2015, doi: 10.1177/0023830914565191.

[9] A. Hanulíková, 'Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups', *PLoS One*, vol. 16, no. 10 October, Oct. 2021, doi: 10.1371/journal.pone.0259230.

[10] M. J. Witteman, A. Weber, and J. M. McQueen, 'Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation', *Atten Percept Psychophys*, vol. 75, no. 3, pp. 537–556, 2013, doi: 10.3758/s13414-012-0404-y.

[11] A. Weber, A. M. di Betta, and J. M. McQueen, 'Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners', *J Phon*, vol. 46, no. 1, pp. 34–51, 2014, doi: 10.1016/j.wocn.2014.05.002.

[12] B., Evans, and Iverson, P. 'Plasticity in vowel perception and production: A study of accent change in young adults.' *J. Accoustic Soc. Am. 121* (6), pp. 3814-3826, 2007, https://doi.org/10.1121/1.2722209

[13] Y. Zheng and A. G. Samuel, 'The relationship between phonemic category boundary changes and perceptual adjustments to natural accents', *J Exp Psychol Learn Mem Cogn*, vol. 46, no. 7, pp. 1270–1292, Jul. 2020, doi: 10.1037/xlm0000788.

[14] T. Bent and M. Baese-Berk, 'Perceptual learning of accented speech', *The Handbook of Speech Perception*, pp. 428–464, 2021, doi: 10.1002/9781119184096.ch16.

[15] X. Xie and E. B. Myers, 'Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers', *J Mem Lang*, vol. 97, pp. 30–46, 2017, doi: https://doi.org/10.1016/j.jml.2017.07.005.

[16] L. Liu, and T. F. Jaeger, 'Inferring causes during speech perception', *Cognition*, vol. 174, pp. 55–70. 2018. doi: 10.1016/j.cognition.2018.01.003

[17] Y. Zheng, and A. G. Samuel, 'Flexibility and stability of speech sounds: The time course of lexically-driven recalibration', *J Phon,* vol. 97, 2023. doi.org/10.1016/j.wocn.2023.101222

[18] M. Tan, X. Xie, and T. F. Jaeger, 'Using rational models to interpret the results of experiments on accent adaptation', *Front Psychol*, vol. 12, Nov. 2021, doi: 10.3389/fpsyg.2021.676271.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting linear mixed-effects models using lme4', *J Stat Softw*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[20] S. K. Sidaras, J. E. D. Alexander, and L. C. Nygaard, 'Perceptual learning of systematic variation in Spanish-accented speech.' *J Acoust Soc Am* **125**, 3306, 2009.

[21] X. Xie, L. Liu, and T. F. Jaeger, 'Cross-talker generalization in the perception of non-native speech: A large-scale replication.' *Journal of Experimental Psychology: General*, 150, e22–e56, 2021. https://doi.org/10.1037/xge0001039

[22] D. Saltzman and E. Myers, 'Listeners are initially flexible in updating phonetic beliefs over time', *Psychon Bull Rev*, 2021, doi: 10.3758/s13423-021-01885-1.

[23] M. J. Witteman, N. P. Bardhan, A. Weber, and J. M. McQueen, 'Automaticity and stability of adaptation to a foreign-accented speaker', *Lang Speech*, vol. 58, no. 2, pp. 168–189, Jun. 2015, doi: 10.1177/0023830914528102.

[24] X. Xie, T. F. Jaeger, and C. Kurumada, 'What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review'. *Cortex.* To appear.

An Amodal Study of Phonological Abstraction in Early Infancy

Eylem Altuntas[1], Catherine T. Best[1], Marina Kalashnikova[2], Antonia Goetz[1], Denis Burnham[1]
[1]MARCS Institute for Brain, Behaviour, and Development, Western Sydney University, Sydney, Australia
[2]Basque Center on Cognition, Brain, and Language
eylem.altuntas@westernsydney.edu.au

ABSTRACT

Adults learn non-native phonological contrasts faster if they experience the stimulus language in their first postnatal 6 months than if they never experienced it [1]. The resulting implication that young infants engage in phonological abstraction was examined here. We trained infants to associate words from two artificial mini-languages distinguished by consonant place (labial vs coronal), presented to two subgroups either as audio-only speech (A) or video-only speech (V) followed by different animal images for each of the two mini-languages. In Test, novel words of each mini-language were presented in the opposite modality to Training (A→V or V→A), followed by either the Congruent (trained) or the Incongruent (opposite mini-language) animal. The A→V mode infants attended longer to Congruent than Incongruent Test trials, but the V→A infants did not. Thus, infants show (i) phonological abstraction of the labial vs coronal distinction, (ii) in a cross-modal task, (iii) but only for A→V transfer.

Keywords: infant speech perception, phonological abstraction, language acquisition, artificial language training, perceptual learning

1. INTRODUCTION

While very young infants as early as 1 month discriminate consonant or vowel contrasts from just about any language, later in their first year they begin to pay less attention to foreign, non-native contrasts and attune to the contrasts in their ambient (native) language [1]. This process of perceptual attunement to native speech contrasts occurs around 6 months of age for vowels and around 9-10 months for consonants, and it is a well-established phenomenon [2-4]. As perceptual attunement involves the infant establishing the phonemes of the language spoken around them, speech perception *before* 6 months is generally assumed to involve phonetic-level or even acoustic-level [5] rather than phonemic-level speech perception. An implicit corollary of this assumption is that before 6 months, prior to perceptual attunement, there is no phonological abstraction, i.e., that perceptual attunement is a necessary precursor of phonological abstraction.

This view has recently been challenged by studies which show that adults retain phonologically abstract knowledge that could only have been acquired before 6 months, that is, before perceptual attunement begins. Choi and colleagues [6, 7] found that 30-year-old Dutch speakers who were adopted at 3 to 5 months of age from Korea (no subsequent contact with Korean) learned to discriminate the Korean 3-way fortis, lenis, and aspirated alveolar stop contrast, [t*]-[t]-[th], faster than their Dutch counterparts who had no previous exposure to Korean. In addition, with no further perceptual or production training, the adoptees were significantly better than the Dutch controls at (i) generalising discrimination of the 3-way contrast from the trained alveolar place of articulation to the bilabial and the velar places, and (ii) producing the contrast at all three places.

Those findings indicate that the adoptees had not only retained knowledge of speech heard early in life, but they also stored that knowledge at an abstract phonological level. As that abstract representation must have been laid down before 6 months [9, 10], these results imply that perceptual attunement (which occurs from around 6 months) is not a prerequisite for phonological abstraction. Indeed, it may even be possible that phonological abstraction paves the way for perceptual attunement. However, this evidence for abstract phonological representation prior to 6 months and storage thereafter into adulthood is indirect as it comes from 30-year-old adults. The present study investigates the development of abstract phonological knowledge directly in young infants.

Two forms of abstraction were studied here:

1. Infants first had to learn two mini-languages of word sets that differed only in consonant place of articulation: labial (lips) vs coronal (tongue tip). Infants were exposed to pseudo-words (non-words) paired with animal images that differed between the languages. Tests for this place of articulation learning presented novel words of each mini-language paired with either the same (Congruent) or the opposite (Incongruent) animal as during exposure. *This manipulation tests whether infants can form contrasting mini-language categories based solely on place of articulation, irrespective of the actual consonants manifested at these places.*

2. Infants received the word-animal pairings in the exposure phase in either audio-only (A) or video-only (V: silent talking face) modality. Test trial words

were presented in the opposite modality from the exposure phase (A→V or V→A). *This manipulation tests whether infants can learn the articulatory contrast between word sets (lips vs tongue tip consonants) and generalise it to the opposite modality, and whether this depends on direction of modality change.*

With respect to 1, research suggests that even in their earliest months of life, infants are capable of tracking the distributional statistics of vowels and consonants in their environment, allowing them to begin learning the phonetic categories used in their native language. For instance, studies such as Wanrooij, Boersma, and van Zuijen [8] have shown that infants as young as 2-3 months old can rapidly learn phonetic distinctions, as evidenced by event-related potential (ERP) studies. Additionally, Zacharaki and Sebastian-Galles [9] found evidence that infants can begin to learn and differentiate between the sounds of their native language even prior to the onset of perceptual narrowing. Furthermore, infants can discriminate consonant place of articulation contrasts by as young as young as 1-3 months of age [10]. In addition, infants as young as 2 months old can abstract consonant place contrasts across vowel context changes, as demonstrated by Bertoncini et al. [11]. Hillenbrand's [12] research further suggests that infants can perform this abstraction across contexts and talkers by the age of 6 months.

Regarding 2, it is now well-established that acoustic and dynamic optical information (from the lips, face and head of the talker) yield a collaborative effect on infants' speech perception: audio and video information together facilitate infants' recognition of their mother [13]; infants match an audio phoneme to a silent video of the face articulating it [14]; they also integrate information from the two modalities to perceive a hybrid phoneme [15]. Thus, it appears that infants detect associations between what they hear and what they see, suggesting that infants' speech perception is intermodal.

It is also possible that beyond simple transfer from one modality to another, or intermodal, speech perception may in fact be amodal, *rather than auditory or visual*. That is, speech may be perceived phonologically, which is more abstract than the specific sensory modalities. In this regard, by 4 months of age (prior to perceptual attunement) infants can recognise the particular mouth movements that correspond to specific speech sounds [16] even for phonemes that do not occur in their native language [17]. This is posited to be modality-neutral or amodal in nature, i.e., not being simply auditory, visual, tactile, and/or gestural but above them all [18].

Further evidence for the amodal nature of speech perception comes from another cross-modal study investigating Spanish-learning and English-learning infants [19]. In that study, 6- and 11-month-old infants were familiarised to silent videos (video-only) of the English /b/-/v/ contrast, which is non-existent in Spanish, then habituated to either audio-only /b/ or /v/, followed by a test on the same video-only /b/-/v/ contrast as in the familiarisation period. Six-month-olds in both language groups showed a preference for the video-only consonant they had been habituated to in the audio-only mode. But only the English-learning infants continued to show this video-only preference at 11 months, implying that the Spanish-learning infants showed a developmental decline in cross-modal discrimination of this non-native consonant contrast. This cross-modal matching of audio-only to subsequent video-only speech information is a true instance of infants' amodal speech perception.

The evidence for amodal perception in the above-mentioned study [19] comes from auditory followed by visual (A→V) presentations. There are also studies using a V→A design For instance, Teinonen et al. [20] presented silent videos of Finnish vowels to 4-month-old infants and found that the infants were better able to discriminate between audio-only vowels when they had previously seen the corresponding silent videos. Similarly, Kuhl et al. [21] presented silent videos of Mandarin Chinese syllables to 6-month-old infants and found they could better discriminate between audio-only syllables after viewing the videos. Nonetheless, other studies have not found that silent videos enhance infants' later perception of audio-only speech. Weikum et al. [22] presented silent videos of English or French vowels to 4-month-old infants and found no evidence that the videos improved their perception of audio-only vowels. These inconsistencies suggest that the relationship between visual and auditory information in speech perception is complex and may depend on various factors that require further exploration.

Thus, this study assessed POA and cross-modality (auditory, visual) abstraction, including both A→V and V→A conditions. This was achieved using a task created for a larger project on phonological abstraction (see ICPhS 2023 paper 674). In our POA/amodal task, infants were familiarised with two artificial languages containing non-words that differed solely on corresponding audio and video information – labial POA includes not only audible place information but also visible 'lip' articulatory gestures, and coronal POA includes both audible place information and visible 'tongue tip' gestures. The words were presented in either audio-only (A) or video-only (V) form, followed by tests in the opposite modality (A→V or V→A).

2. EXPERIMENT

Due to the COVID-19 pandemic, we initiated data collection online using the Lookit platform [23] and later continued it in our infant speech perception laboratory. In both contexts, we examined infants' ability to differentiate between mini-languages that differed on a single phonological feature of their consonants. In one mini-language all consonants were articulated using the lips (i.e., the *labial* feature for place of articulation) whereas in the other mini-language all consonants were articulated using the tongue tip (i.e., the *coronal* place feature).

2.1. Method

2.1.1. Participants

Eighty-three monolingual English-learning infants participated: 53 infants completed the A→V modality (exposed to audio-only words then tested on video-only) and 30 infants completed the V→A mode (exposed to video-only and tested on audio-only).

Another 20 participants were excluded from the A→V group due to technical difficulties (n = 5), fussiness (n = 4), experimenter error (n = 4), language background (n = 6), and background noise (n = 1), leading to a success rate of 72%. In the V→A group, another 21 participants were excluded due to technical complications (n = 5), fussiness (n = 3), experimenter error (n = 4), language background (n = 5), background noise (n = 3), and caregiver interference (n = 1), resulting in a success rate of 58%, which was lower than for A→V.

A→V infants had a mean age of 5.49 months (SD = 1.42 months, range = 3.91-8.94 months), and V→A ones had a mean age of 6.31 months (SD = 1.64 months, range = 4.08-8.98 months). 47 infants were North American English-learning monolinguals in the USA and 36 were Australian-English learning monolinguals in Sydney, Australia. In terms of ethnicity, 91.5% of the infants were Caucasian.

All infants were born full-term, without any known risks of cognitive or language delay, and with normal hearing and vision. Demographic information was either collected at registration or on the test day. Families received either a $5 gift card (for online participation) or $20 travel reimbursement (for lab participation), and a certificate of completion.

2.1.2. Stimuli

The two mini-languages were composed of non-words that only differed in place of articulation of the syllable-initial consonants: lips (labial place: /b, v, w/) versus tongue tip (coronal place: /d, z, l/). All words had three consonant-vowel (CV) syllables, with different consonants and vowels in each of the three syllables (e.g., *bi-va-wo* for lips, or *dæ-zu-la* for tongue tip). 240 non-words were generated for each of the two mini-languages.

A female native speaker of Australian English was video- and audio-recorded producing all 480 unique non-words for the two mini-languages in infant-directed speech (IDS). Video recordings consisted of a close-up of the speaker's face from the neck up. Audio-only (A) and video-only (V) stimulus words were separately extracted from the recordings.

2.1.3. Procedure

We used an adaptation of an associative learning task [24], which includes an Exposure Phase and a Test Phase. During the Exposure Phase, infants were exposed to two sets of word-image pairings, with modality of the words depending on Audio-only *or* Video-only modality subgroup: Set A (lips) words were followed by a cartoon image of a jellyfish and Set B (tongue tip) words by a cartoon crab image. Infants completed a set number of 36 Exposure trials divided into 3 Blocks always in the same order: 12 Set A pairings, 12 Set B pairings, then 12 mixed Set A and Set B pairings (6 of each presented in pseudo-random order). Test blocks followed immediately after Exposure with no break. The Test Phase was cross-modal for both subgroups (opposite modality to Exposure: A→V or V→A), and included two Blocks of 12 trials, a total of 24 Test trials. 25% of the Test trials were Incongruent trials [12.5% A-words, 12.5% B-words] such that the word-animal pairings were mismatched with respect to Exposure pairings (Lips-Crab; Tongue Tip-Jellyfish); and 75% of trials were Congruent trials [37.5% A-words, 37.5% B-words], such that the word-animal pairings were the same as in Exposure (Lips-Jellyfish; Tongue Tip-Crab). Test trials were presented in pseudo-random order, such that 3-trial sequences of Incongruent trials were separated by at least two 3-trial sequences of Congruent trials (e.g., Incongruent, Congruent, Congruent, Incongruent).

Different subsets of the nonwords in each mini-language were used for each phase (Training and Test) – the experiment program randomly selected words from the relevant mini-language without replacement. Thus, infants never heard or saw the same word twice, and the words were randomised across task phases and infants.

Infants sat on the parent's lap facing a computer monitor and watched a 6-minute series of the non-words, which they either heard *or* saw in Exposure and Test according to their cross-modal subgroup (A→V or V→A), paired with the designated cartoon animal for each mini-language. Parents wore

headphones and listened to music during the task so as not to inadvertently influence the baby.

2.2. Data Analysis

Infants' looking times were recorded during the experimental session and were coded offline using ELAN. Looking time was averaged over triplets (3-trial sequences) due to individual trials being very short (4.443 seconds). We transformed the raw looking times into proportions of looking time over each 3-trial sequence because there were minor variations in trial length due to internet transmission in the online test context. Statistical analyses were performed on the Test Phase data with linear mixed models in *R* [25], using the *lmer* function from the package *lme4* [26]. The *p* values for the fixed-effects factors were computed using the Kenward-Roger approximation to the degrees of freedom, as recommended by [27] and the *anova* function from the *car* package [28] was used to calculate *F*. The dependent variable was proportion of looking time per triplet. Only the Incongruent triplet and the immediately preceding Congruent triplet in each block were included in the analysis, a total of 2 Congruent-Incongruent sequences. The fixed effects were Condition (Congruent, Incongruent), Blocks (Block 1, Block 2), and Mode (A→V, V→A). Participants was a random effect.

2.3. Results

Results from the linear mixed model analysis are displayed in Figure 1. The main effect of Condition was significant, $F(1, 243) = 4.28$, $p = .03$); infants looked longer in Congruent than Incongruent triplets. There was also a significant interaction between Condition and Mode: $F(1, 243) = 4.46$, $p = .03$. Therefore, we next tested whether Condition was significant in each Mode separately. In the A→V Mode, the main effect of Condition indicates that infants looked significantly longer for Congruent than Incongruent triplets ($F(1, 314) = 27.44$, $p < 0.001$), However, in V→A Mode, Condition was not significant ($F(1, 176) = 0.31$, $p = .5$).

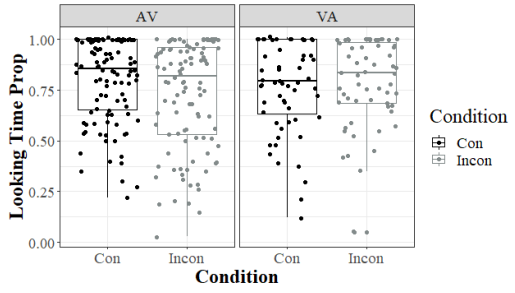


Figure 1: Looking time proportions across Test conditions in A→V and V→A cross-modal subgroups.

3. DISCUSSION AND CONCLUSIONS

This study investigated amodal phonological abstraction in infancy. Specifically, we asked whether 1. infants can learn two artificial languages based on consonants on a consonant place of articulation difference that is both auditorily and visually distinct, and whether 2. they can then transfer this learning across modalities, i.e., from audio-only to video-only (A→V mode) and video-only to audio-only (V→A).

In the A→V, but not the V→A mode, infants showed phonological abstraction of the difference between two artificial languages that differed in using labial vs tongue-tip consonants. Infants generalised this abstract category learning across the A→V speech modality change, but not across the V→A speech modality change. This shows that learning and cross-modal generalisation occur in A→V, but not in V→A. This modality difference could suggest that (i) there is no visual learning and thus no cross-modal transfer or that (ii) there is visual speech learning, which is not amodally represented thus obviating cross-modal transfer to auditory speech. The latter *cannot be correct*, because there *must be* amodal representation of visual speech given the A→V results. Thus, it is possible that in the V→A mode there is no visual learning in the Exposure phase for some more peripheral reason. The most likely candidate is that when the initial trials, in Exposure, consist of visual-only presentations (in V→A), these are not as interesting or attention-getting as are initial auditory-only trials (in A→V) such that there is little to no learning of the word-animal associations. This could be tested by detailed examination of the Exposure phase in the A→V and V→A groups, but this is beyond the scope of this presentation.

These preliminary analyses suggest that young infants show two types of phonological abstraction in the same context (i) phonological abstraction of the labial vs coronal consonant distinction, (ii) additional phonological abstraction in the form of amodal speech perception, but only in the A→V mode. Thus, not only is there phonological abstraction of phonetic /articulatory features of speech, but there is also abstract amodal representation of this phonologically abstract information prior to completion of phonological attunement in the first year of life.

4. REFERENCES

[1] J. F. Werker, J. H. Gilbert, K. Humphrey, and R. C. Tees, "Developmental aspects of cross-language speech perception," *Child development,* vol. 52, no. 1, pp. 349-355, 1981.

[2] J. F. Werker and R. C. Tees, "Influences on infant speech processing: Toward a new synthesis," *Annual review of psychology,* vol. 50, no. 1, pp. 509-535, 1999, doi: 10.1146/annurev.psych.50.1.509.

[3] P. K. Kuhl, "A new view of language acquisition," *Proceedings of the National Academy of Sciences,* vol. 97, no. 22, pp. 11850-11857, 2000, doi: 10.1073/pnas.97.22.11850.

[4] L. Polka and J. F. Werker, "Developmental changes in perception of nonnative vowel contrasts," *Journal of Experimental Psychology: Human perception and performance,* vol. 20, no. 2, p. 421, 1994, doi: 10.1037//0096-1523.20.2.421.

[5] D. K. Burnham, M. D. Tyler, and S. Horlyck, *Periods of speech perception development and their vestiges in adulthood*. 2002.

[6] J. Choi, M. Broersma, and A. Cutler, "Early phonology revealed by international adoptees' birth language retention," *Proceedings of the National Academy of Sciences,* vol. 114, no. 28, pp. 7307-7312, 2017.

[7] J. Choi, A. Cutler, and M. Broersma, "Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory," *Royal Society Open Science,* vol. 4, no. 1, p. 160660, 2017.

[8] K. Wanrooij, P. Boersma, and T. L. Van Zuijen, "Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study," *Frontiers in psychology,* vol. 5, p. 77, 2014.

[9] K. Zacharaki and N. Sebastian-Galles, "Before perceptual narrowing: The emergence of the native sounds of language," *Infancy,* vol. 27, no. 5, pp. 900-915, 2022.

[10] P. D. Eimas, "Auditory and linguistic processing of cues for place of articulation by infants," *Perception & Psychophysics,* vol. 16, no. 3, pp. 513-521, 1974.

[11] J. Bertoncini, R. Bijeljac-Babic, P. W. Jusczyk, L. J. Kennedy, and J. Mehler, "An investigation of young infants' perceptual representations of speech sounds," *Journal of experimental psychology: General,* vol. 117, no. 1, p. 21, 1988.

[12] J. Hillenbrand, "Speech perception by infants: Categorization based on nasal consonant place of articulation," *The Journal of the Acoustical Society of America,* vol. 75, no. 5, pp. 1613-1622, 1984.

[13] D. Burnham, "Visual recognition of mother by young infants: facilitation by speech," *Perception,* vol. 22, no. 10, pp. 1133-1153, 1993, doi: 10.1068/p221133.

[14] B. Dodd and D. Burnham, "Processing speechread information," *The Volta Review,* 1988.

[15] D. Burnham and B. Dodd, "Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages," in *Speechreading by humans and machines*: Springer, 1996, pp. 103-114.

[16] P. K. Kuhl and A. N. Meltzoff, "The bimodal perception of speech in infancy," *Science,* vol. 218, no. 4577, pp. 1138-1141, 1982, doi: 10.1126/science.7146899.

[17] G. E. Walton and T. Bower, "Amodal representation of speech in infants," *Infant Behavior and Development,* vol. 16, no. 2, pp. 233-243, 1993.

[18] P. K. Kuhl and A. N. Meltzoff, "Speech as an intermodal object of perception," in *Perceptual development in infacy: The Minnesota symposia on child psychology*, 1988, vol. 20, pp. 235-266.

[19] F. Pons, D. J. Lewkowicz, S. Soto-Faraco, and N. Sebastián-Gallés, "Narrowing of intersensory speech perception in infancy," *Proceedings of the National Academy of Sciences,* vol. 106, no. 26, pp. 10598-10602, 2009, doi: 10.1073/pnas.0904134106.

[20] T. Teinonen, V. Fellman, R. Näätänen, P. Alku, and M. Huotilainen, "Statistical language learning in neonates revealed by event-related brain potentials," *BMC neuroscience,* vol. 10, no. 1, pp. 1-8, 2009.

[21] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science,* vol. 255, no. 5044, pp. 606-608, 1992.

[22] W. M. Weikum, A. Vouloumanos, J. Navarra, S. Soto-Faraco, N. Sebastián-Gallés, and J. F. Werker, "Visual language discrimination in infancy," *Science,* vol. 316, no. 5828, pp. 1159-1159, 2007.

[23] K. Scott and L. Schulz, "Lookit (part 1): A new online platform for developmental research," *Open Mind,* vol. 1, no. 1, pp. 4-14, 2017.

[24] C. Kabdebon and G. Dehaene-Lambertz, "Symbolic labeling in 5-month-old human infants," *Proceedings of the National Academy of Sciences,* vol. 116, no. 12, pp. 5805-5810, 2019.

[25] R. C. Team, "R: The R Project for Statistical Computing [Internet]. 2021 [cited 2022 Aug 22]," *Available from: Available from:* https://www. *r-project. org*.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823,* 2014.

[27] U. Halekoh and S. Højsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," *Journal of Statistical Software,* vol. 59, pp. 1-32, 2014.

[28] J. Fox and S. Weisberg, *An R companion to applied regression*. Sage publications, 2018.

PERCEPTION OF METAPHONY: A COMPARISON BETWEEN TWO DIALECTS OF THE LAUSBERG AREA (SOUTHERN ITALY)

Pia Greca

Institute of Phonetics and Speech Processing (IPS), LMU Munich
greca@phonetik.uni-muenchen.de

ABSTRACT

In several sound changes, differences in production do not necessarily match those in perception. This study explores this relationship in two varieties (MZ: Mittelzone; MM: Mormanno) spoken in the Lausberg area of Southern Italy that differ in the degree to which inflectional phonetic cues are transferred from a suffix to a phonetically mid stem vowel. In MZ, the transfer can be complete (e.g. [mɛs, mis]; cf. Standard Italian [meze, mezi], month/months), whereas in MM it is more marginal and with suffix preservation. Listeners of both varieties identified in a forced-choice test inflectional information from stems in MM- and MZ-produced stimuli with all suffix information removed. The results showed a perceptual advantage in morphological categorisation when listeners perceived stems of their own variety, and suggest that, compatibly with speech production, MM listeners use a phonetic mode of perception in categorising morphological information, whereas in MZ the perceptual strategy is more phonological.

Keywords: Metaphony, perception, dialect variation, sound change, phonologisation

1. INTRODUCTION

Over the last few years, several studies have investigated the relationship between production and perception in the progression of a sound change [1, 2, 3]. However, this relationship is often complex and cross-linguistically not quite predictable. As several studies [1, 3, 4, 5, 6] have shown, sound change does not necessarily progress at the same rate in production and perception. Whether or not perception leads production can depend both on the type of sound change and whether the sound change has reached (or is near to) completion in a group of speakers instead of another [5, 6, 7]. This study explores the perception-production link in some dialects belonging to the so-called Lausberg area in Southern Italy [8], in which different forms of metaphony (an especially pervasive phenomenon

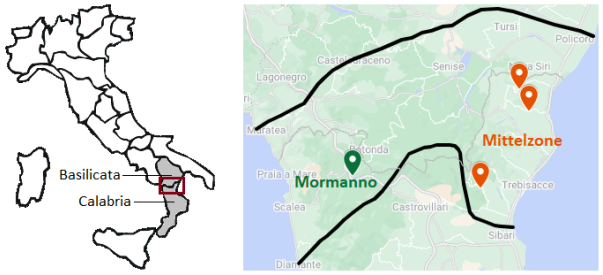


Figure 1: Geographical localisation of the Lausberg area and the dialects in this study (map data ©2022 Google).

No metaphony	Metaphony		Inflectional meaning
	MM	MZ	
[mɛse]	[mesi]	[mis(ə)]	sing./pl.
[bɔna]	[bonu]	[bun(ə)]	fem./masc.

Table 1: Examples of metaphony in MM and MZ

in Italo-Romance varieties [9, 10], which derives originally from V-to-V coarticulation) coexist with different degrees of suffix vowel quality erosion.

In metaphony, the cues to the suffix vowel are manifested in the stem to different degrees. In the Lausberg area, the preferred targets for metaphony are the mid stem vowels /e, o/, while its triggers are the high suffix vowels /i, u/ [8, 11]. These suffixes may indicate a variety of morpho-syntactic categories, such as gender and number in nouns and adjectives. Thus, whereas in Standard Italian the phonetic distinction between e.g. 'month'/'months' [meze, mezi], or between the masculine and the feminine singular form of 'good' [bwɔna, bwɔno] is carried by the final inflectional suffix (and not by the stem), in regions affected by metaphony high-vowel suffixed forms are marked by different degrees of stem vowel raising. Thus, from the examples in Table 1 we can observe that the suffix vowels /i, u/ have caused stem vowel raising in the plural and in the masculine forms in Mormanno (MM). On the east coast of the Lausberg area in the so-called Mittelzone (MZ) region, the degree of metaphony is, however, far more advanced (see Table 1) [8, 11, 12]. The evidence that metaphony has been phonologised [13, 14] to a greater extent in MZ than in MM is

shown by the degree of suffix reduction: whereas in MM the suffix is typically present, in MZ it is usually absent (or at least strongly reduced; see Table 1) [12, 15]. The issue to be considered in the present study is the extent to which perception matches production as far as these different degrees of metaphonic progression are concerned. An initial hypothesis to be tested is that Lausberg listeners can identify morphological inflection in the stem better than chance. This is likely to be so, given that in these regions information about the inflectional suffix is manifested in varying degrees of strength in the stem. Another hypothesis is that MZ listeners should be more sensitive to morphological information in the stem than MM listeners. This follows from studies of sound changes in progress in other languages showing greater listener sensitivity to phonologised than non-phonologised forms [16] and also because MZ listeners are likely to be more reliant on phonetic information in the stem as a cue to morphological inflectional information, given that the suffix is often absent in their variety. Both hypotheses apply within each variety, i.e. when listeners from MZ (MM) are listening to MZ (MM) stimuli. The predictions about cross-variety responses (MZ listening to MM and vice-versa) are, however, less clear. Firstly, given that the direction of sound change in MM is towards that of MZ (in that both show information about morphological inflection in the stem but to different degrees) and taking into account that some MM listeners are likely to have come into contact with MZ speakers, then MM listeners might be able to parse inflectional information from the stem as accurately as MZ listeners when both groups listen to MZ speakers. On the other hand, MZ listeners might not be attuned to the fine, coarticulation-based phonetic differences that signal morphological information in MM stems. This hypothesis is supported by the observation from lexical phonology that these gradient phonetic differences are no longer parsed by speakers at a final stage of sound change, in which categorical boundaries between phonemes have already gained morpho-lexical relevance [14, 17]. In addition, given that MM is a mountainous village whereas MZ is an entire region, MZ listeners are perhaps less likely to have been exposed to the MM variety than the other way round. For these reasons, MZ listeners might be less accurate than MM listeners when listening to the MM variety. These predictions were tested in a lexical decision task in which listeners had to judge morphological information given only the stem without any suffix information in both varieties.

2. METHOD

2.1. Participants

All participants were recruited either from personal contacts or on social media. The specific requirement was that participants spoke either the MM or an MZ variety. Before starting the experiment, participants were asked to fill out a questionnaire asking for some biographic metadata including which specific local dialect they speak (to be sure that they were proficient in one of the target dialects). A total of 73 Lausberg listeners took part in the experiment. However, responses from listeners that either came from one of the very few villages in MZ whose dialect has diphthongising metaphony (based on [8, 11]) and/or completed less than 50% of the experiment (n=39) were excluded from analysis. Therefore, the final analysis was based on 34 participants (mean age 36.7, age range 19–65, 17 females) who completed the experiment either in its entirety (n=31) or responded to more than 50% of the stimuli (n=3). 17 participants were analysed from MM (mean age 33.3, age range 19–54, 12 females) and 17 from MZ (mean age 40.1, age range 21–65, 5 females). Upon completion of the task, participants were rewarded with an Amazon voucher.

2.2. Materials and design

The audio recordings used for the experiments were selected from the production data described in [12] and included 25 speakers (14 for MZ and 11 for MM stimuli). Three of these speakers also took part in the perception experiment. The target words were elicited from these speakers by means of a picture-naming task in which words had to be first produced in isolation and then in an embedded sentence ("I say ... two times"). The selected word pairs (Table 2) were formed from the same lexical stem followed by different inflectional suffixes: either /i, u/ (which cause phonetic raising in the stem vowel) or /e, a/ suffixes (which cause no raising). The auditory stimuli were formed from 7 word pairs for /e/ and 7 for /o/ stems. Each word type (n=14) was repeated 4 times for each region to give 4 (repetitions) $\times$ 14 (word types) $\times$ 2 (stems) $\times$ 2 (regions: MM, MZ) = 224 stimuli. For each audio file used, the suffix vowel was manually excised on the basis of an audiovisual inspection of the sound wave and the spectrogram by using the software Praat (version 6.1.51) [18], so that the suffix vowel quality was no longer audible. In a minority of cases, part of the duration of the consonant preceding the suffix



Figure 2: Screenshot from the experiment interface: the response was logged by clicking on one of the two options.

vowel was also removed if vowel-to-consonant coarticulation still enabled the suffix vowel to be perceived.

2.3. Procedure

The experiment was run on the web by means of the online platform Percy [19]. Participants participated remotely by using either a computer or a mobile device. They were also encouraged to complete the task in a quiet place and preferably using headphones. The experiment was structured so that participants were exposed to stimuli produced by speakers from both regions and in equal measure, without being previously advised that the stimuli were produced in any particular dialect. Having completed the initial questionnaire, the participant could start the task, which consisted first in listening to the audio stimulus and then providing a forced choice response between two morpho-lexical options (singular or plural; masculine or feminine). Participants were not explicitly informed that the last vowel in the audio recording had been intentionally deleted. The order of the two written options was randomised (right and left-hand side of the screen). In order to listen to the stimulus, the participant had to either click on the headphone symbol that appeared on the experiment web page or to press the space bar.

Each audio stimulus was presented only once, and the experiment was programmed so that each participant could listen to each stimulus only twice. Six extra stimuli were added at the beginning of the experiment and were used as a training phase: responses to these stimuli were not analysed. The software was able to register all those cases in which the participants answered without listening to the stimulus: such cases were excluded from further analysis. The audio stimuli used in both the training phase (first six stimuli) and the 'real' experimental phase were presented to participants in random order. The task took approximately 15 to 20 minutes. The total number of responses collected

Stem	Suffixes	Word pairs		Meaning
/e/	/a, u/	bella	bellu	beautiful (f./m.)
		pettsa	pettsu	cloth, piece
		vekkja	vekkju	old (f./m.)
	/e, i/	dente	denti	tooth, teeth
		mese	mesi	month, months
		pede	pedi	foot, feet
		verme	vermi	worm, worms
/o/	/a, u/	bona	bonu	good (f./m.)
		korna	kornu	horns, horn
		kotta	kottu	cooked (f./m.)
		morta	mortu	dead (f./m.)
		ossa	ossu	bones, bone
		ova	ovu	eggs, egg
		tsoppa	tsoppu	lame (f./m.)

Table 2: Word types used in the experiment (phonological transcription) and their meaning.

and used for the analysis (n=7334) included 3670 responses to stimuli taken from words containing /e/ stems and 3664 observations taken from stem-/o/ words.

2.4. Statistical analysis

The statistical analysis tested, by means of a logistic mixed model, whether there were significant differences in the accuracy of participants' answers between regions. The model was computed with the `lme4` package [20] (version 1.1.31), while post-hoc comparison tests and associated Bonferroni corrections were run with the `emmeans` package [21] (version 1.8.2) in R [22]. The dependent variable was binary (two levels: correct or incorrect answer), while the model was fitted with the fixed factors *Suffix vowel* (four levels: /a, e, i, u/), the region of the speaker who produced each stimulus (henceforth *Stimulus region*), the *Listener's region* (two levels, MM and MZ), and all possible interactions. The random factors included the lexical *Stem* (e.g. for word pairs like /bella, bellu/ the lexical stem was /bell/, see Table 2) with slope in *Stimulus region*, the *Speaker* who produced each stimulus with slope in *Suffix vowel*, and the random intercept *Listener*, i.e. each participant in the perception experiment.

3. RESULTS

Fig. 3 shows the proportion with which morphological information was correctly identified in stem-only stimuli. We can observe that identification accuracy was above chance (50%) level except in two cases: MM listeners perceiving MM stimuli spliced from /i/ suffixes, and MZ listeners perceiving MM stimuli spliced from /u/ suffixes. Given that all main effects (p<.05, with the exception of the *Listener's region*) and all interactions (p<.001) were significant, the results

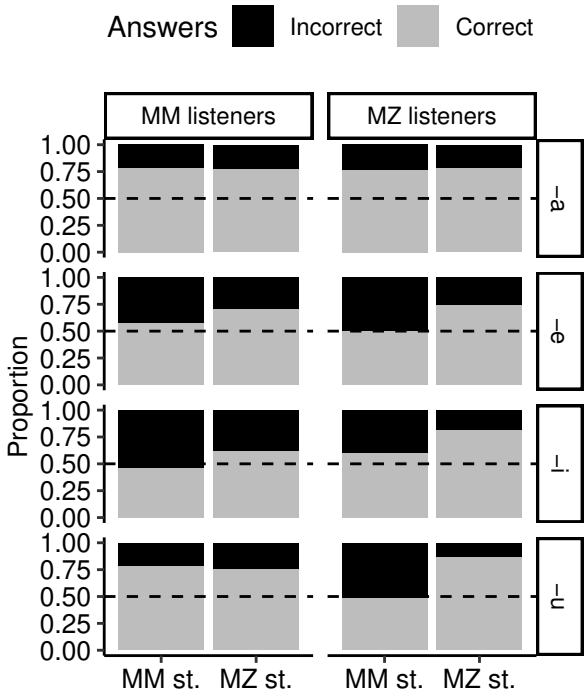


Figure 3: Proportion of correct (grey) and incorrect (black) responses by listener's region, stimulus region, and suffix. The dashed horizontal line shows chance level.

reported below are for Bonferroni-corrected post-hoc tests, which were all significant at $p<.001$ unless otherwise stated. MZ listeners were significantly more accurate in perceiving MZ- than MM-stimuli excised from /u/ suffixes (Fig. 3, row 3: col. 3 vs col. 4) and were also more accurate than MM listeners in perceiving MZ stimuli for both suffix /i/ (row 3: col. 2 vs 4) and /u/ ($p<.01$; row 4: col. 2 vs 4). On the other hand, MM listeners perceiving MM stimuli were more accurate than MZ listeners for suffix /u/ (Fig. 3, row 4: col. 1 vs 3). For MM stimuli only, there were firstly more accurate responses to /a/-suffixed stems perceived by MM listeners than to /i/-suffixed ones (Fig. 3, col. 1: row 1 vs row 3); secondly, there were more /a/-suffixed stems accurately perceived by MZ listeners than /u/-suffixed ones (col. 3: row 1 vs row 4); finally, there were more /u/-suffixed stems accurately perceived by MM listeners than /i/-suffixed ones (col. 4: row 1 vs row 4).

4. DISCUSSION

The study tested listeners' ability to identify inflectional morphological information in the stem for two varieties of the Lausberg area. Recent production [12] and earlier auditory studies [8, 11, 15] suggest that these varieties differ in the extent to which cues have been transferred from the suffix to the stem, being greater for the Mittelzone (MZ) than for Mormanno (MM). A central hypothesis was that MZ listeners should be more sensitive to morphological information in the stem than MM listeners: if cue information from the suffix has been transferred to the stem to a greater extent in MZ, then MZ listeners should be more sensitive to this type of information. This was found to be so when MZ listeners categorised MZ-produced stems spliced from /i, u/ suffixes, but MZ listeners had no such perceptual advantage in categorising MM-produced stems in the same contexts. Perhaps this was because the cues to the suffix in MM-produced stems were too weak to be perceived. However, MM listeners were also more accurate than MZ listeners in categorising MM-produced stems spliced from /u/ suffixes. These region-specific results suggest that MM listeners might be more sensitive to fine phonetic information in the stem as a cue to the suffix, whereas MZ listeners perceive the morphological information in the stem more categorically. However, this interpretation is not entirely compatible with the following finding: when MM listeners categorised MM-produced stems that had been spliced from /i/ suffixes, then their performance was at chance level. MM listeners may have performed poorly in this task because the acoustic phonetic differences relative to the suffix /e/ (vs suffix /i/) competitor is marginal, but acoustically far greater in judging /u/ suffixes (for which the competitor was always suffix /a/, see Table 2). Overall, the results are consistent with different, phonetic and phonological modes of perception for MM and MZ listeners respectively. When perceiving the own variety, perceptual differences between MM and MZ match those from production [12] showing a greater trade-off between coarticulatory source (the suffix) and effect (the stem) for MZ than for MM. However, results relating to cross-variety responses confirm that the progressively greater perceptual trade-off between coarticulatory source and effect as a sound change progresses [7, 23, 16] can be inhibited when the change is completed [5, 6]. Finally, these results lend support to some theories from lexical phonology [17, 24] that sound changes can progress from being under predominantly phonetic control (as in MM) towards a phonological stage (as in MZ) in which the original phonetic factors at the sound change's origin have a negligible influence on categorical judgements.

5. ACKNOWLEDGMENTS

This study was funded by the European Union (ERC, *SoundAct*, project N. 101053194). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Many thanks to Jonathan Harrington for his help with the analysis, and to two anonymous reviewers for their helpful comments.

6. REFERENCES

[1] J. Harrington, F. Kleber, and U. Reubold, "Compensation for coarticulation,/u/-fronting, and sound change in standard southern British: An acoustic and perceptual study," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2825–2835, 2008.

[2] T. Kendall and V. Fridland, "Variation in perception and production of mid front vowels in the US Southern Vowel Shift," *Journal of Phonetics*, vol. 40, no. 2, pp. 289–306, 2012.

[3] F. Kleber, J. Harrington, and U. Reubold, "The relationship between the perception and production of coarticulation during a sound change in progress," *Language and speech*, vol. 55, no. 3, pp. 383–405, 2012.

[4] J. Harrington, F. Kleber, and U. Reubold, "The production and perception of coarticulation in two types of sound change in progress," in *Speech planning and dynamics*, S. Fuchs, M. Weirich, D. Pape, and P. Perrier, Eds. Frankfurt: Peter Lang, 2012, pp. 39–62.

[5] A.-F. Pinget, "The actuation of sound change," Ph.D. dissertation, University of Utrecht, 2015.

[6] A. W. Coetzee, P. S. Beddor, K. Shedden, W. Styler, and D. Wissing, "Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis," *Journal of Phonetics*, vol. 66, pp. 185–216, 2018.

[7] J. Kuang and A. Cui, "Relative cue weighting in production and perception of an ongoing sound change in Southern Yi," *Journal of Phonetics*, vol. 71, pp. 194–214, 2018.

[8] H. Lausberg, *Die Mundarten Südlukaniens*. Halle an der Saale: Niemeyer, 1939.

[9] G. Rohlfs, *Grammatica storica della lingua italiana e dei suoi dialetti. Fonetica*. Torino: Einaudi, 1966, vol. 1.

[10] M. Maiden and L. M. Savoia, "Metaphony," in *The dialects of Italy*, P. M. Maiden M., Ed. London/New York: Routledge, 1997, pp. 15–25.

[11] K.-H. Rensch, *Beiträge zur Kenntnis nordkalabrischer Mundarten.* MÃ¼nster: Aschendorffsche Verlagsbuchhandlung, 1964.

[12] P. Greca, M. Gubian, and J. Harrington, "The relationship between the coarticulatory source and effect in sound change: evidence from Italo-Romance metaphony in the Lausberg area," Submitted.

[13] L. M. Hyman, "Enlarging the scope of phonologization," in *Origins of sound change: Approaches to phonologization*, A. C. Yu, Ed. Oxford: Oxford University Press, 2013, pp. 3–28.

[14] P. Kiparsky, "Phonologization," in *The Oxford handbook of historical phonology*, J. Honeybone and J. Salmons, Eds. Oxford: Oxford University Press, 2015, vol. 1, pp. 563–582.

[15] J. Trumper, "Calabria and southern Basilicata," in *The dialects of Italy*, M. Maiden, M.; Parry, Ed. London/New York: Routledge, 1997, pp. 355–364.

[16] P. S. Beddor, K. B. McGowan, J. E. Boland, A. W. Coetzee, and A. Brasher, "The time course of perception of coarticulation," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2350–2366, 2013.

[17] R. Bermúdez-Otero, "Amphichronic explanation and the life cycle of phonological processes," in *The Oxford handbook of historical phonology*, P. Honeybone and J. Salmons, Eds. Oxford: Oxford University Press, 2015, pp. 374–399.

[18] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.

[19] C. Draxler, "Percy – an HTML5 framework for media rich web experiments on mobile devices," in *Proceedings of the 12th annual conference of the International Speech Communication Association (Interspeech 2011)*. Florence, Italy, 2011, pp. 3339–3340.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[21] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022, r package version 1.8.2. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[23] P. S. Beddor, "A coarticulatory path to sound change," *Language*, vol. 85, pp. 785–821, 2009.

[24] M. Ramsammy, "The life cycle of phonological processes: Accounting for dialectal microtypologies," *Language and Linguistics Compass*, vol. 9, no. 1, pp. 33–54, 2015.

PERCEPTUAL CUES TO EJECTIVE STOPS ACROSS LANGUAGES

Maida Percival

University of Toronto
maida.percival@mail.utoronto.ca

ABSTRACT

The goal of this paper is to identify perceptual cues to an ejective - non-ejective stop contrast from Hul'q'umi'num' for listeners from four languages with ejectives: Hul'q'umi'num', Q'anjob'al, Dene, and Eastern Oromo, chosen for their typologically different stop laryngeal contrasts. Results indicated that listeners from each language were similar in perception: all used as primary cue to the perception of ejectives the period of silence following the stop burst and had similar usage of properties of the stop burst and coarticulation in a following vowel as secondary cues. Cross-linguistic differences in perception could mainly be explained based on inventory differences in terms of which other stop types occur across the languages. This is the first study to systematically test perceptual cues to ejectives and has implications for ejective typology.

Keywords: speech perception, ejectives, typology

1. INTRODUCTION

It has been proposed that there are "strong" (or tense or fortis) and "weak" (or slack or lenis) ejective stops in languages, which are distinguished from one another through groupings of acoustic, articulatory, and perceptual characteristics [1, 2]. For perception, the intuitions are that some languages' ejectives sound "weak", less "poppy", or more similar to non-ejective stops than others [3, 4], but these intuitions have never been tested and in fact no prior studies have systematically examined which acoustic cues are used to perceive ejectives. The contribution of this paper is to start to fill this gap by presenting the results of an experiment whose goals are to identify perceptual cues to an ejective - non-ejective contrast from Hul'q'umi'num' and to identify any cross-linguistic differences in the use of these cues between listeners from four languages with ejectives with typologically different stop laryngeal contrasts (Table 1). Two of these languages, Hul'q'umi'num' and Q'anjob'al, also impressionistically differ in the realization of their ejectives: they are thought to be on average strong in Hul'q'umi'num' but weak in Q'anjob'al (c.f. closely related languages [1, 5]).

It is expected that listeners will use as cues acoustic dimensions that differ between ejective and non-ejective stops. If languages have different inventories of stop laryngeal contrasts, then they might also be expected to differ in their use of cues to the two-way Hul'q'umi'num' contrast.

Language (Family)	Contrast
Hul'q'umi'num' (Coast Salish)	T' - T^h
Q'anjob'al (Mayan)	T' - T
Dene[1] (Dene/Athabaskan)	T' - T^h - T
Eastern Oromo (Cushitic)	T' - T^h - D

Table 1: Summary of languages' laryngeal contrasts. T' = ejective, T = voiceless unaspirated, T^h = voiceless aspirated, D = voiced

2. METHODOLOGY

2.1. Participants

100 participants took part in the experiment, with about equal numbers from each language. Table 2 gives the participant details by language. Because some of the languages had few speakers, second language and less fluent first language (L1) speakers were sometimes included as participants in the study. However, all participants grew up hearing their language in their homes and/or community even if they only spoke it a little themselves.

Language	# (# L1)	Gender	Mean age
Hul'q'umi'num'	26 (7)	18F, 8M	56 (24-87)
Q'anjob'al	25 (25)	13F, 12M	33 (19-61)
Dene	24 (19)	17F, 8M	51 (19-76)
Oromo	25 (23)	12F, 13M	42 (17-65)

Table 2: Summary of participants by language

2.2. Materials

A recording of a real-word ejective - non-ejective minimal pair (t^hahw [tax^w] "later, right now, then" and t'ahw [t'ax^w] "go downhill to water") by a female speaker of Hul'q'umi'num' was manipulated in Praat [6] to create stimuli for a forced-choice identification task. The two baseline tokens

were modified in pieces (burst, post-burst release, following vowel), as described in the following paragraphs. Modification values were based on production [7]. The pieces were cross-spliced to make all possible combinations, a total of 80 stimuli.

Burst: The two bursts (baseline non-ejective and baseline ejective) were manipulated to have two intensities (low, where the mean intensity of the burst was set to 40 dB, and high, where it was set to 50 dB) for a total of four bursts. Burst intensity was manipulated because high intensity bursts are a characteristic of strong ejectives, and so listeners might be expected to use them in perception.

Post-burst release: The baseline non-ejective stop had an aspirated release while the baseline ejective stop had a release of silence following the burst. These two release types (aspiration, silence) were given three durations (0 ms, 50 ms, 120 ms) for a total of five releases (0 ms, 50 ms aspiration, 50 ms silence, 120 ms aspiration, 120 ms silence). 0 ms was included to see how listeners respond when there is no post-burst release: the lack of one is a characteristic of weak ejectives, as well of voiceless unaspirated stops (found in Q'anjob'al and Dene) and voiced stops (found in Oromo).

Following vowel: The two vowels (baseline non-ejective, baseline ejective) were each manipulated to have two F0 patterns (raised, where F0 is raised by 25 Hz for the first 30 ms, and lowered, where F0 is lowered by 25 Hz for the first 30 ms) for a total of four following vowels. After the first 30 ms, the vowels' F0 was set to the mean F0 contour for the speaker who recorded the baseline tokens. The voiceless coda consonant was kept together with its baseline vowel type during cross-splicing (i.e. all tokens with baseline ejective vowels included [x^w] from the baseline ejective [t'ax^w]). Depressed F0 at the vowel onset is a characteristic of weak ejectives and so it was thought that listeners might use vowel F0 to distinguish ejective and non-ejective stops, especially if their language has weak ejectives.

2.3. Procedures

Participants listened to the stimuli in a quiet location: for Hul'q'umi'num', this was at a language school in Duncan, BC, Canada, for Q'anjob'al, this was at a home in Santa Eulalia, Huehuetenango, Guatemala, and for Dene (Déline, NT, Canada) and Oromo (Canada, mostly Toronto), this was typically at the participants' homes. Most participants wore headphones; except some did not have access to any.

The stimuli were presented in random order on a webpage using jsPsych javascript framework. For each trial, participants heard a stimulus play automatically and saw two boxes, one for each word, with the word written in Hul'q'umi'num' orthography. Participants would then select the box with the word that they heard. Most participants completed the task by themselves, but those with accessibility issues indicated their responses by pointing at or verbalizing the word they wanted to select and a researcher would select it for them.

The instructions participants received were to listen and indicate whether the word they heard was tahw [t^hax^w] starting with a plain <t> or t'ahw [t'ax^w] starting with an ejective <t'>. Participants were shown the two words in advance and were played an example recording of them (by a different female speaker). The language groups other than Hul'q'umi'num' also had illustrations included with the text of the words and were given an explanation that there is a language called Hul'q'umi'num', which has ejective sounds reportedly like those of their language and that one of the goals of the study was to see whether they can hear these sounds like they do the equivalent sounds in their own language.

2.4. Statistical analysis

Generalized linear mixed regression models were performed in R [8] using the lme4 package [9] to test the extent of listeners' reliance on the manipulated acoustic dimensions in the stimuli for perception of ejectives. The template for the models is given in 1.

(1)

$$response \sim language * dimension + (1 \mid participant)$$

For each model, the response variable was the participants' response of ejective or non-ejective. The predictor variables were language (4 levels: Hul'q'umi'num', Q'anjob'al, Dene, Oromo) and the manipulated dimension for the given model (i.e., one of release type, release duration, burst type, burst intensity, vowel type, or vowel F0). The dimensions, all simple coded, are summarized in Table 3. The

Dimension	Levels	
Release duration	0 ms	*50 ms* vs. 120 ms
Release type		*silence* vs. aspiration
Burst type	*ejective* vs. non-ejective	
Burst intensity	*high* vs. low	
Vowel type	*ejective* vs. non-ejective	
Vowel F0	*lowered* vs. raised	

Table 3: Summary of dimensions

interaction between language and dimension was also included to test if the use of a dimension as a cue varied across languages. Separate models

were done for responses to stimuli with zero (post-burst) release duration (1597 responses) and those to stimuli with non-zero (post-burst) release durations (6392 responses). For language, the reference level was Hul'q'umi'num', since that was the language the stimuli were from, and for all models, the significance level was p <0.05.

3. RESULTS

3.1. Post-burst releases

Figure 1 shows the results for release duration and release type together in one graph. % ejective response is on the y-axis, release duration is on the x-axis, release type is represented by the different line colours, and the responses for each language are grouped by panel. Each dot represents one participant's mean. Because stimuli of zero post-burst release duration have no release type, they are omitted from the models for release dimensions in this section, and separate models are performed on them in the following sections for the burst and following vowel to see if these dimensions play a bigger role in the absence of release type cues.

In general, the responses for each language group in Figure 1 are very similar to one another: listeners perceive as more ejective stimuli with a period of silence while they perceive as more non-ejective stimuli with aspiration. As for stimuli with neither aspiration nor silence, listeners perceived them about equally as ejective or non-ejective.

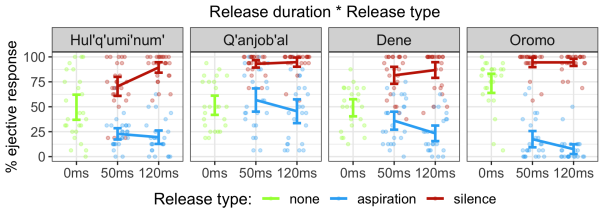


Figure 1: % ejective response to releases by language

Two exceptions to this are reflected across all the models by significant effects of language. First, Q'anjob'al listeners responded with significantly more ejective responses overall than Hul'q'umi'num' listeners for stimuli with non-zero post-burst release durations, as they did not consistently classify stimuli with aspiration (in blue) as non-ejective. Second, Oromo listeners gave significantly more ejective responses for stimuli with zero post-burst release duration (in green).

Figure 2, where dimensions are on the x-axis and languages are represented by line colour, reflects the results of the models for release type and release duration. For release type, stimuli with silence received significantly more ejective responses than those with aspiration (β = -3.468, z = -40.501, p <0.001) in all languages. However, Oromo listeners (in yellow) relied on release type to a greater extent than Hul'q'umi'num' (in red), as indicated by a significant interaction between language and release type (β = -2.433, z = -9.639, p <0.001).

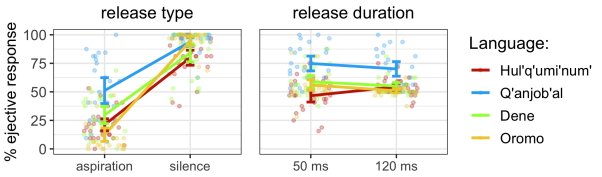


Figure 2: % ejective response to release type and release duration by language

For release duration, no significant difference was found in listeners' responses to stimuli with releases of 50 ms and 120 ms, but there were significant interactions for release duration for Hul'q'umi'num' and each of the other languages. These interactions are a reflection of the patterns seen in Figure 1: for Hul'q'umi'num' there is a greater increase in % ejective response for releases of silence (in red) between 50 ms and 120 ms than for the other languages, while for the other languages the % ejective response to releases of silence is very high (even at 50 ms), but there is a greater decrease in % ejective response for releases of aspiration (in blue) between 50 ms and 120 ms.

3.2. Bursts

Figure 3 (top) shows the results for burst type. In stimuli with post-burst releases, there was a significant effect of burst type (β = -0.416, z = -7.829, p <0.001) and no significant differences across language. For each language, the stimuli with baseline ejective stop bursts were perceived as more ejective than stimuli with baseline plain stop bursts. In stimuli with zero post-burst release duration, the significant effect of burst type remains and in fact listeners use burst type more (β = -1.222, z = -7.829, p <0.001), especially Q'anjob'al listeners (in blue) (β = -0.931, z = -2.71, p = 0.007).

Burst intensity results are shown in Figure 3 (bottom). Burst intensity was significant in stimuli with both non-zero (β = -0.268, z = -9.701, p <0.001) and zero post-burst release durations (β = -0.333, z = -2.85, p = 0.004). Stimuli with high burst intensity were perceived as more ejective than those with low burst intensity. However, there were differences across languages in the extent

that this was so in stimuli with non-zero post-burst release durations. Significant interactions for Hul'q'umi'num' compared to Dene (β = 0.471, z = 3.25, p = 0.001) and Oromo (β = 0.469, z = 3.292, p <0.001) suggested that while listeners from Hul'q'umi'num' and Q'anjob'al do use burst intensity, Dene and Oromo listeners do not differ in % ejective response based on high vs. low intensity bursts. These interactions were not significant in stimuli with zero post-burst release duration; Dene listeners used burst intensity similarly to Hul'q'umi'num' and Q'anjob'al listeners, but with Oromo listeners' mean % ejective response (in yellow) being lower with high burst intensity, it's not clear, despite the model, that they use it at all.

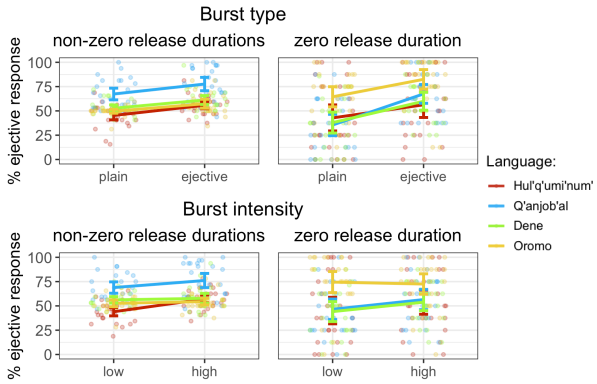


Figure 3: % ejective response to burst type (top) and intensity (bottom) by language in stimuli with non-zero (left) and zero (right) post-burst release durations

3.3. Following vowels

The results for vowel type (Figure 4) resemble those for burst type but with slightly smaller effects. Stimuli with baseline ejective vowels received significantly more ejective responses than those with baseline plain vowels (β = -0.206, z = -3.898, p <0.001). In stimuli with zero post-burst release duration, vowel type's significant effect was greater (β = -0.984, z = -8.005, p <0.001), especially for Q'anjob'al (β = -0.875, z = -2.532, p = 0.011).

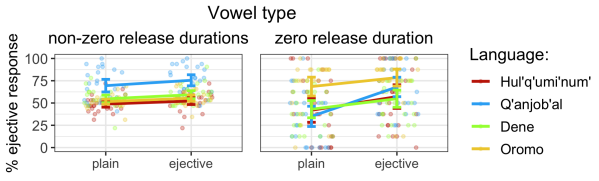


Figure 4: % ejective response to vowel type by language in stimuli with non-zero (left) and zero (right) post-burst release durations

F0 at the vowel onsets was not significant in stimuli with any release duration, nor were any interactions of vowel F0 and language significant. This suggests that vowel F0, at least as manipulated for this study was not used in perception of ejectives.

4. DISCUSSION AND CONCLUSIONS

Listeners from each language were remarkably similar in perception: they all used as primary cue to ejectives the period of silence following the stop burst. Burst type (likely correlated with spectral properties of the burst and/or change in amplitude over time) and vowel type (likely correlated with creaky (ejective) and breathy (aspirated) voice quality from stop coarticulation) were secondary cues in each of the languages, that listeners relied on even more in stimuli with no period of silence. Burst intensity was also a secondary cue but perhaps as it encompassed only a single acoustic dimension, it was used less consistently.

While both strong (long release, loud burst) and weak (creaky vowel) ejective characteristics consistently cued ejectives, small differences across language groups were present, which can be attributed to their inventories of laryngeal stop types. Q'anjob'al listeners more often hearing aspiration as ejective is likely due to the lack of aspirated stops in their language, and their greater burst type and vowel type cue usage in stimuli without aspiration or silence may reflect greater exposure to short stops, given that Q'anjob'al only has voiceless unaspirated pulmonic stops. Dene and Oromo listeners relying less consistently on burst intensity may be due to intensity being too subtle to consistently differentiate three stop laryngeal types. Oromo listeners classifying stimuli with no post-burst release as mainly ejective may relate to their lack of voiceless unaspirated pulmonic stops (their only pulmonic stops without post-burst releases are voiced). Their greater use of release type may also be evidence of the importance of aspiration to voiceless non-ejective stops for them. An alternate explanation for Oromo's greater use of release type, and Q'anjob'al's of vowel and burst type is demographics: Hul'q'umi'num' and Dene's L1 listeners being fewer and older than Q'anjob'al and Oromo's may have led to less consistent cue usage.

Given that the present study was based on non-native speech for three language groups, further experiments of language-specific perception and production are underway for each language to add to this study's main finding: that there seems to be little evidence for cross-linguistic differences in the perception of ejectives, outside of how they contrast within a language's stop laryngeal type inventory.

5. REFERENCES

[1] J. Kingston, "The Phonetics and Phonology of the Timing of Oral and Glottal Events," Ph.D. dissertation, University of California Berkeley, 1985.

[2] M. Lindau, "Phonetic differences in glottalic consonants," *Journal of Phonetics*, vol. 12, pp. 147 – 155, 1984.

[3] S. Bird, "Dakelh ejectives: Evidence for new ways of classifying sounds," 76th Meeting of the Linguistic Society of America.San Francisco, CA, 2002.

[4] P. D. Fallon, *The Synchronic and Diachronic Phonology of Ejectives*. New York, NY: Routledge, 2002.

[5] S. Russell, "Some acoustic characteristics of pulmonic and glottalic stops in Mam," Master's thesis, Simon Fraser University, 1997.

[6] P. Boersma and D. Weenink, "Praat," http://www.praat.org, University of Amsterdam, 1992−2017. [Online]. Available: ttp://www.fon.hum.uva.nl/praat/

[7] M. Percival, "Contextual variation in the acoustics of Hul'q'umi'num' ejective stops," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, M. T. . P. W. Sasha Calhoun, Paola Escudero, Ed., 2019, pp. 3270–3274.

[8] R. C. Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: http://www.R-project.org/

[9] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[1] This language is also called Sahtúot'ine kedé by its speakers in Déline but I shall refer to it as Dene in this paper, because this term is shorter and because it is a term now used for both the language family and the particular language.

KOREAN LISTENERS' WEIGHTING AND INTEGRATION OF CUES TO THE THREE-WAY KOREAN STOP CONTRAST

Hyoju Kim[1], Annie Tremblay[2], Taehong Cho[3]

[1]University of Kansas, USA, [2]University of Texas at El Paso, USA, [3]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS), Hanyang University, Korea
kimhj@ku.edu, actremblay@utep.edu, tcho@hanyang.ac.kr

ABSTRACT

The current study investigates whether listeners' cue weighting predicts their cue integration as the speech signal unfolds over time. It does so by testing the time course of acoustic cue integration in the processing of Korean stop contrasts by native Korean listeners. Listeners' weighting of voice onset time (VOT) and fundamental frequency (F0) to perceive Korean stop contrasts was measured with a cue-weighting speech perception task, and the time course of VOT and F0 integration was examined with a visual-world eye-tracking task. The results revealed that the timing of VOT integration is predicted by listeners' reliance on F0, with delayed integration of VOT in target-competitor pairs where F0 is a primary cue to process the stop contrast. These results suggest that listeners adapt their integration of acoustic cues in spoken word recognition based on the importance they assign to individual cues.

Keywords: Korean laryngeal stop contrast, speech perception, cue weighting, cue integration.

1. INTRODUCTION

Phonological contrasts in natural speech sounds are distinguished by multiple acoustic dimensions. The acoustic cues involved in a phonological contrast, however, do not have equal importance in speech processing. While some acoustic cues contribute more reliably to distinguishing phonological contrasts (primary cues), other cues are more variable (secondary cues) [1, 2].

An important issue in the processing of cues to phonological contrasts is that acoustic information from different cues often arrives asynchronously as the speech signal unfolds over time. One influential account for listeners' integration of asynchronous acoustic cues is the continuous integration approach [3, 4]. The approach assumes that each cue provides partial evidence for lexical access as soon as it arrives. This partial information results in the greater activation of lexical candidates that are consistent with it, and lexical activation is updated as the remaining cues become available over time. However, it is not clear how such cue integration that develops over the time course of speech processing is related to cue weighting and the primacy of available cues in perceiving speech segments.

The present study investigates the relationship between listeners' cue-weighting and cue integration: Are acoustic cues integrated continuously when the weight of the cue available later is greater than that of the cue available earlier? As a test case, we examined how individual native Seoul Korean listeners weight and integrate voice onset time (VOT) and onset fundamental frequency (F0) when hearing Korean words that begin with a laryngeal stop.

Korean has a three-way stop contrast, with VOT and onset F0 of the following vowel playing a role in distinguishing the contrasts. In word-initial position, the fortis stop has a short VOT and high F0, the lenis stop has an intermediate VOT and low F0, and the aspirated stop has a long VOT and high F0 [5]. In Seoul Korean, however, the VOTs of lenis and aspirated stops have merged over time, resulting in speakers being more likely to depend on the onset F0 difference of the following vowel when distinguishing lenis from aspirated stops [6, 7, 8, 9].

In the perception of the Korean stop contrasts, Seoul Korean listeners use VOT as a primary cue for perceiving the aspirated vs. fortis contrast but F0 as a primary cue for perceiving the aspirated vs. lenis contrast, and both VOT and F0 as primary cues for perceiving the lenis vs. fortis contrast [10]. These acoustic properties of the Korean stop contrasts suggest that it is important to study the time course of cue integration because they raise the question of whether the timing of cue integration is contingent on the primary cue being heard earlier (VOT) or later (F0) in the speech signal.

The present study used a cue-weighting speech perception task (Experiment I) to quantify listeners' relative weighting of cues to the stop contrast and a visual-world eye-tracking experiment (Experiment II) to test the time course of cue integration. Under the continuous integration account, two hypotheses can be made regarding the relationship between cue integration and cue weighting. One is that the time course of cue integration is associated with cue weighting (the *associated view*). Under this view, listeners would not wait for F0 to integrate VOT if the earlier cue (VOT) is a primary cue (aspirated vs.

fortis), but they would wait for F0 to integrate VOT if the later cue (F0) is a primary cue (aspirated vs. lenis; fortis vs. lenis). The alternative hypothesis is that cue weighting is not associated with the time course of cue integration (the *independent view*). This view predicts that listeners would not wait for F0 to integrate VOT regardless of whether the earlier cue (VOT) is a primary or secondary cue to the contrast.

2. METHODS

2.1. Participants

The participants were 62 native Seoul Korean listeners (38 female, mean age: 24.1, SD: 3.2). They were tested in Korea and received monetary compensation for their participation. All participants had normal or corrected-to-normal vision and no speech and hearing disorder history.

2.2. Experiment I

2.2.1. Materials

The auditory stimuli for Experiment 1 were based on a Korean triplet *ppul* [p*ul] 'horn', *bul* [pul] 'fire', and *phul* [p^{h}ul] 'grass' [10]. The triplet was recorded by a female Seoul Korean speaker. The VOT of the word-initial stop and the onset F0 were each manipulated along a seven-step continuum. The minimum and maximum values of the VOT (0-110 ms) and F0 (180-290 Hz) continua were selected based on the acoustic analysis of the recorded tokens. The manipulation resulted in 49 auditory stimuli (7 steps of VOT × 7 steps of F0). All stimuli were normalized to a mean amplitude of 70 dB.

2.2.2. Procedure

The complete stimulus matrix was heard three times, once for each contrast pair (i.e., fortis vs. lenis choice, lenis vs. aspirated choice, and fortis vs. aspirated choice). On each trial, participants were asked to select the word that best matched what they heard by pressing either the left or the right arrow on the keyboard. The word labels were given with corresponding buttons on the computer monitor, and the position of the word labels was counterbalanced across trials. A total of 441 trials (49 stimuli × 3 pairs of choices × 3 repetitions) were randomly presented in three blocks, with each stimulus being heard once per block.

2.2.3. Data analysis

Mixed-effects logistic regression analyses were conducted on the participants' responses in each trial. The first model focused on listeners' responses to the aspirated vs. lenis contrast (aspirated response = 1). The fixed effects included VOT, F0, and their interaction. The initial model included random intercepts for item, participant, and repetition, and random slopes of VOT and F0 for each participant. The largest model was backward fit, and the best model was selected using the log-likelihood ratio comparisons. The second and third models focused on the lenis vs. fortis contrast (lenis response = 1) and the aspirated vs. fortis contrast (aspirated response = 1), respectively. When two cues showed significant effects, the fixed-effect coefficients were compared to determine which cue had a stronger effect than the other cue [11].

2.3. Experiment II

2.3.1. Materials

Fifteen disyllabic Korean words (five disyllabic noun triplets) that begin with a bilabial stop were selected as the critical stimuli. The words within a triplet share the same syllable structure and phonemes in the first syllable except for the word-initial stop, such that there is a temporary lexical ambiguity contingent on the initial consonant (e.g., *ppaltae* 'straw' - *balgul* 'digging' - *paljji* 'bracelet'). The critical words differed further at the onset of the second syllable. The words were controlled for (log-transformed) token frequency and number of letters.

The words in the display were presented orthographically in pairs. Along with the critical stimuli, there were 15 filler stimuli (five disyllabic triplets) with a three-way affricate contrast in word-initial position. Like the critical stimuli, the words within a filler triplet shared the same syllable structure and phonemes in the first syllable except for the word-initial affricate.

Critical trials consisted of six target-competitor types (3 target types × 2 competitor types): the fortis target and lenis competitor; the fortis target and aspirated competitor; the lenis target and fortis competitor; the lenis target and aspirated competitor; the aspirated target and fortis competitor; and the aspirated target and lenis competitor. On each trial, a target-competitor pair was grouped with one filler pair in the visual display, making a quadruplet (e.g., *ppaltae* 'straw' - *paljji* 'bracelet', *jjokji* 'note' - *chokgam* 'feel').

The auditory stimuli were recorded by a female native Seoul Korean speaker. The speaker produced each word three times in the carrier sentence *jigeum __(l)eul nuleusejo* 'Now click on __'. One recording of the carrier phrase was selected for the experiment. The target words were elicited from their original

carrier sentence. Unlike Experiment 1, the VOT of the word-initial stop and onset F0 of the following vowel were manipulated to have specific VOT and onset F0 values (fortis stop: 0 ms VOT and 250 Hz F0; lenis stop: 55 ms VOT and 180 Hz F0; aspirated stop: 110 ms VOT and 290 Hz F0); these values fell within the natural distribution of values for the corresponding stops [10], and they ensured that VOT and F0 would arrive as asynchronously as possible.

Since manipulating VOT yields different lengths of the first syllable within a triplet and results in longer syllables, thus taking slightly more time to reach the point of disambiguation, the duration of the following vowel was manipulated such that the syllable would have the same duration across tokens within a triplet. All stimuli were normalized to a mean amplitude of 70 dB.

2.3.2. Procedure

Eye movements were recorded with a head-mounted SR EyeLink II eye-tracker, sampling at 250 Hz. The experiment began with calibrating the eye tracker, which was followed by four practice trials and the main session. In each trial, participants saw a display with four printed Korean words arrayed in an invisible 2 × 2 grid for 2,000 ms. The arrangements of the printed words in the grid were randomized on each trial. The preview was followed by a fixation cross in the center of the screen for 500 ms. As soon as the fixation cross disappeared, the four printed words appeared on display in the same location, and the auditory stimulus was given over headphones. Participants were asked to click on the target word corresponding to the auditory stimulus as quickly and accurately as possible.

There were 180 trials, including 90 critical trials (6 target-competitor types × 5 triplets × 3 repetitions) and 90 filler trials (6 target-competitor types × 5 triplets × 3 repetitions). These trials were presented in three blocks, each containing 60 trials, with randomized orders across participants.

2.3.3. Data analysis

For the eye movement data, the average proportion of fixations directed to each word was calculated in 4 millisecond time windows for each target-competitor type. Statistical analyses were conducted on the difference between the empirical log-transformed proportions of target and competitor fixations (i.e., target-over-competitor advantage). The time course of each target-competitor type was aligned with the VOT disambiguation point, which is defined as the point where listeners could begin using VOT to distinguish the target from the competitor. We used generalized additive mixed modeling (GAMM) to analyze the data due to the nonlinearity of the eye-fixation trajectories. Three binary smoothing models were built using the function *bam* of the R package *mgcv* [12]. Each model examined the time point from the VOT disambiguation onset where the target-over-competitor advantage becomes significantly above 0 (i.e., significant target-over-competitor advantage). The GAMM specification of each model included the difference between the non-linear patterns of a target-competitor pair, the non-linear difference over time in the general time pattern for each participant (random effects), and the by-listener linear random slope for the two target-competitor pairs.

3. RESULTS

3.1. Experiment I

Fig. 1 shows listeners' responses and fixed-effect coefficients extracted for each listener on each contrast type.

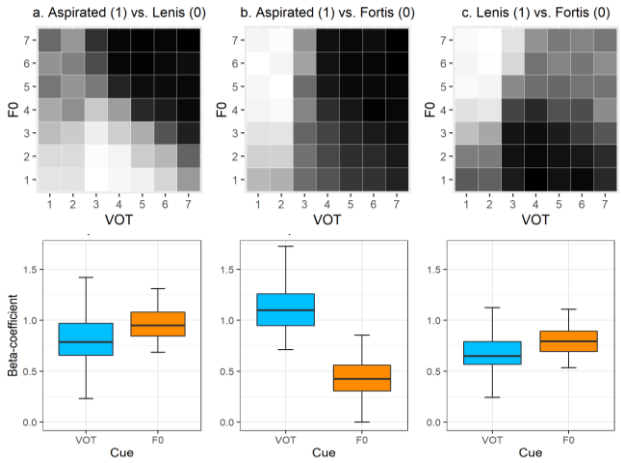


Figure 1: Listeners' responses (upper) and fixed-effect coefficients (lower) for the aspirated vs. lenis (left), the aspirated vs. fortis (middle), and the lenis vs. fortis (right) contrasts. The darker the cell, the more stop responses coded as 1.

For the aspirated vs. lenis contrast, the model found a significant simple effect of VOT ($\beta = 0.67$, $p < 0.001$) and F0 ($\beta = 0.94$, $p < 0.001$), with listeners' proportion of aspirated stop responses increasing as step number increased, and a significant two-way interaction between VOT and F0 ($\beta = 0.11$, $p < 0.001$), with the effect of VOT in Korean listeners' aspirated stop responses being greater at higher levels of F0 and with the effect of F0 being greater at higher levels of VOT. A comparison of cue importance revealed that the effect of F0 was stronger than that of VOT [$t(12535) = 2.92$, $p < 0.01$]. For the aspirated vs. fortis contrast, there was a significant simple

effect of VOT (β = 1.27, p < 0.001), with listeners' proportion of aspirated stop responses increasing as VOT step number increased, but no simple effect of F0. There was also a significant two-way interaction (β = 0.19, p < 0.001), with the effect of VOT in listeners' aspirated stop responses being greater at higher levels of F0 and the effect of F0 being greater at higher levels of VOT. For the lenis vs. fortis contrast, there was a significant simple effect of VOT (β = 0.56, p < 0.001) and F0 (β = −0.75, p < 0.001), with listeners' proportion of lenis stop responses increasing as VOT step number increased and F0 step number decreased. A comparison of cue importance showed that the effect of F0 was stronger than that of VOT [t(12497) = −17.87, p < 0.001].

These results suggest that Korean listeners rely more on onset F0 than on VOT to distinguish aspirated stops from lenis stops and lenis stops from fortis stops, and they rely more on VOT than on F0 to distinguish aspirated stops from fortis stops.

3.2. Experiment II

Fig. 2 shows listeners' target-over-competitor advantage in the target-competitor conditions.

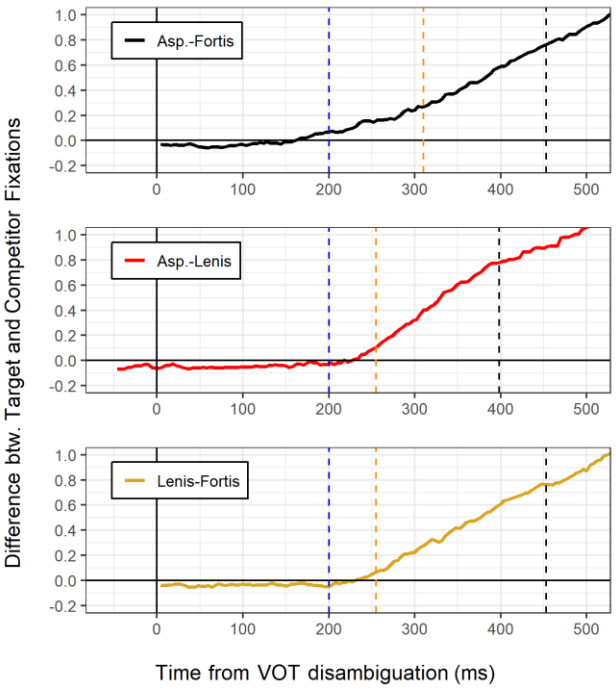


Figure 2: Listeners' eye movement in the condition with an aspirated target and a fortis competitor (top), an aspirated target and a lenis competitor (middle), and a lenis target and a fortis competitor (bottom). The vertical solid line represents the VOT disambiguation onset in the stimuli. The blue, yellow, and black dotted lines demarcate corrected VOT disambiguation onset, F0 onset, and second syllable onset (assuming a 200-ms oculomotor delay).

For the condition with an aspirated target and a fortis competitor (top panel), the GAMM analysis revealed that listeners start integrating VOT at 209 ms from the VOT disambiguation onset, suggesting that listeners showed a target-over-competitor advantage as soon as VOT became available in the speech signal. For the condition with an aspirated target and a lenis competitor (middle panel), the model showed that listeners started looking at the aspirated target at 248 ms from the VOT disambiguation onset, indicating a somewhat delayed lexical activation compared to the condition where the competitor began with a fortis stop. Similarly, for the condition with a lenis target and a fortis competitor (bottom panel), listeners started integrating VOT at 246 ms from the VOT disambiguation onset, suggesting that listeners' VOT integration was somewhat delayed.

These results indicate that listeners' real-time processing of acoustic cues for lexical access is modulated by their cue weighting, with delayed integration of VOT (early cue) in the target-competitor pairs where onset F0 (later cue) is a primary cue to process the stop contrast.

4. DISCUSSION

This study used a cue-weighting task and a visual-world eye-tracking paradigm to investigate how Seoul Korean listeners weight and integrate VOT and onset F0 to process Korean stop contrasts. The results support the *associated view* of cue integration, in that listeners' real-time processing of the cue available earlier in the speech signal was affected by the weight of the cue available later in the speech signal. These results suggest that the time course of cue integration is associated with the importance that listeners assign to different acoustic dimensions. From a theoretical perspective, the results elaborate more on the nature of the cue integration mechanism in spoken word recognition [3, 4] by providing the first empirical evidence for the effect of the relative perceptual weight of phonetic information on the time course of cue integration. Further research is needed to find converging evidence of the associated view of cue integration and to better understand the mechanism underlying the time course of cue integration, as it remains unclear whether cue integration is a consistent property of individual listeners or varies depending on the nature of acoustic cues involved in the phonological contrast.

Lastly, it is worth mentioning that a cue-weighting task showed that listeners do use VOT as a cue to the aspirated vs. lenis contrast, which is in line with the account discussed in [7] regarding the ongoing sound change in Seoul Korean lenis-aspirated distinction.

5. ACKNOWLEDGEMENTS

This work was supported by the KU Linguistics Graduate Student Research Scholarship.

6. REFERENCES

[1] Francis, A. L., Baldwin, K., Nusbaum, H. C. 2000. Effects of training on attention to acoustic cues. *Perception & Psychophysics* 62, 1668-1680.

[2] Holt, L. L., Lotto, A. J. 2006. Cue weighting in auditory categorization: implication for first and second language acquisition. *J. Acoust. Soc. Am*. 119, 3059-3071.

[3] McMurray, B., Clayards, M. A., Tanenhaus, M. K., Aslin, R. N. 2008. Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic bulletin & review* 15, 1064-1071.

[4] Toscano, J. C., McMurray, B. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 436-464.

[5] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384-422.

[6] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287-308.

[7] Choi, J., Kim, S., Cho, T. 2020. An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. PloS ONE, 15 e0240682.

[8] Kang Y. 2014. Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: a corpus study. *J. Phon.* 45, 77-90.

[9] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., Yoon, T. J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: implications for mechanisms of tonogenesis. *J. Phon.* 66, 120-144.

[10] Lee, H., Politzer-Ahles, S., Jongman, A. 2013. Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *J. Phon.* 41, 117-132.

[11] Tremblay, A., Broersma, M., Zeng, Y., Kim, H., Lee, J., Shin, S. 2021. Dutch listeners' perception of English lexical stress: A cue-weighting approach. *J. Acoust. Soc. Am*. 149, 3703-3714.

[12] Wood, S. 2006. Generalized additive models: an introduction with R. CRC Press.

DEVELOPMENTAL DIFFERENCES IN INTELLIGIBILITY OF UNFAMILIAR ACCENTS

Rachael Frush Holt[1], Tessa Bent[2], Holly Lind-Combs[1], and Audrey Kunath[1]

[1] The Ohio State University, Columbus, OH, USA, [2] Indiana University, Bloomington, IN, USA
holt.339@osu.edu

ABSTRACT

Early school-aged children have more difficulty understanding speech produced with unfamiliar accents than adults. However, little is known about the development of these abilities. Furthermore, most investigations use one or two unfamiliar talkers, limiting generalizability of the findings. We examined younger (5-7 years; n=304) and older children's (11-13 years; n=147) perception of English sentences spoken by two unfamiliar native, one bilingual, five non-native, and one familiar native (Midland American English) accented speakers in quiet and noise-added conditions. In quiet, both groups were near ceiling for familiar and unfamiliar native accents but showed performance decrements for some non-native and bilingual accents. In noise, younger children's word recognition accuracy declined across all accents whereas the older children maintained high accuracy for the unfamiliar native accents and one non-native accent. Thus, children's word recognition with unfamiliar accents improves during middle childhood but may continue to mature into adolescence, particularly in noise.

Keywords: Speech perception, intelligibility, children, non-native accents, unfamiliar accents

1. INTRODUCTION

Both adults and children are sensitive to the social-indexical variability present in speech due to speaker accent [1-3]. This sensitivity emerges in infancy [4-6] and continues to develop through childhood [2,7-10] and, likely, into adolescence [11-14]. Some research suggests that young children and even toddlers can understand unfamiliar accented speech [15-16], consistent with findings that certain fundamental perceptual skills that likely contribute to accent perception are in place early in development, including the observation that young children show lexically guided retuning of phoneme boundaries [17] and toddlers can learn phoneme mappings of artificial accents [18]. However, other research suggests that accent perception follows a protracted developmental trajectory certainly through the early elementary years [3,7,19,20] and well into adolescence [11-14]. Data supporting a protracted developmental trajectory are consistent with findings that some general auditory and speech perception skills are developing into adolescence, including auditory perceptual learning [21], understanding speech in the presence of competing noise [22], and phoneme categorization [23]. Differences in tasks and task demands, the number of accents studied, whether the accents are regional or non-native, and how the accents differ from the listener's home accent likely contribute to the disparate developmental findings.

Several recent investigations on the perception of regional American English varieties suggest that sociolinguistic abilities continue to develop through middle childhood and possibly into adolescence. Using a free classification paradigm with four regional American English dialects, monolingual American English listeners ranging in age from 4-86 years completed two experiments [12]. In the first, 8- to 86-year-olds classified sentence recordings from five females and five males of each regional dialect from the TIMIT Corpus [24] based on their perceived region of origin. The youngest children were less accurate than children 12 years of age and older, as well as the adults, at placing talkers from the same regional dialect into the same group. It was not until children were 16 years of age that their accuracy was adult-like. A second experiment reduced the number of sentences to make it more child-friendly and included children between 4-11 years. The 4- to 5-year-olds produced fewer talker groups than the older children, and children between 4-7 years performed less accurately than children between 8-11 years. These results suggest that classification of these regional dialects by monolingual American English speakers incrementally improves through childhood and adolescence, with more substantial growth in middle (7-8 years) and late childhood (11-12 years), with adult-like performance appearing by 16 years.

Two related studies [13-14] employed different tasks to examine dialect perception of monolingual English speakers ranging from 4-70+ years. The stimuli were TIMIT sentences spoken by 12 females with three from each of the same regional American English dialects as [12]. In the first study [13],

listeners completed two tasks: dialect identification using ad-hoc labels and dialect discrimination. The discrimination task was easier for the younger children than the identification task, which would be predicted based on well-known foundations of auditory perceptual development [25]. Specifically, 4- to 5-year-olds could discriminate New England talkers from Midland and Northern talkers but could not identify them with ad-hoc labels until 10-11 years of age. Reliable discrimination was achieved across all four dialects by 12-13 years of age, whereas ad-hoc labelling of all four dialects was adult-like by 14-15 years of age. These results suggest that the cognitive capacity to discriminate some of these regional dialects is present in preschool, but it takes more than 10 years of exposure to linguistic variation in their native language to reliably discriminate and identify these four regional dialects. The second study [14] examined the development of language attitudes and talker regional background using rating scales of locality, status, and solidarity. A fundamental understanding of dialect perception based on locality judgements was observed in children as young as 6 years for some of the dialects; adult-like performance was achieved by 8 years of age, which is earlier than was observed on the discrimination and ad-hoc labelling tasks in [13].

Together, these investigations suggest that (1) task demands influence when sociolinguistic abilities with regional accents appear to fully mature, and (2) while early underpinnings of sensitivity to accent variability appear before children enter school, other dimensions of accent perception likely take more than 10 years to fully mature. Although these studies have revealed important insights into the development of accent perception skills with a focus on regional accents, they have not tested whether listeners understand the linguistic content of speech. While there is evidence that young children struggle to understand talkers with unfamiliar nonnative and regional accents [2,7-9], few have examined the development of this ecologically important skill. One of the only studies to include adolescents [11] employed a single non-native accent of English (Japanese). Word recognition accuracy was evaluated in quiet and speech-shaped noise for children 5 to 15 years and young adults. Accuracy significantly improved with age for the native accent in noise and for the non-native accent in quiet and noise; accuracy for the native accent in quiet was at ceiling for all age groups. The age at which children achieved adult-like performance varied with accent and noise condition: 8-9 years and 11-12 years for the native talker in quiet and noise, respectively, and 14-15 years for non-native talker in quiet; the oldest children (15 years) did not achieve adult-like performance for the non-native talker in noise. The inclusion of a single non-native accent limits the generalization of these findings. Because of its importance to functioning in everyday life in a global world, the current study examines intelligibility of eight unfamiliar native and non-native accents in school-age children through adolescents to examine differences in accent intelligibility development more fully.

2. METHOD

2.1. Participants

Two groups of children participated. The younger group consisted of 304 children between 5 to 7 years (mean age = 6.5 years, SD = 0.8 years; 163 female, 141 male, 0 nonbinary). The older group consisted of 147 children between 11 and 13 years (mean age = 12.4 years, SD = 0.9 years; 74 female, 72 male, 1 nonbinary). All children's caregivers reported typical hearing, speech, and language development. All children were monolingual American English-speaking and had minimal to no exposure to the unfamiliar accent varieties to which they were assigned. Testing took place in the Midland dialect region of the United States, so all children had extensive exposure to Midland American English.

2.2. Materials

The stimuli were 60 English sentences from the Hearing In Noise Test for Children (HINT-C) [26]. HINT-C sentences are simple, declarative sentences containing words familiar to children. The sentences were recorded from nine adult females representing each the following English accents: Midland American, Southern Standard British, Scottish, Bilingual Hindi, German, Mandarin, French, Japanese, and Spanish. Root Mean Square amplitude was equalized across the sentences to equate sound pressure level. To create the noise-added condition, sentences were mixed with 8-talker babble [from 27] at a signal-to-noise ratio (SNR) of +4 dB. Each sentence was centered on a random selection of babble that was 1-second longer than the sentence.

2.3. Procedure

Children were recruited from the Center of Science and Industry (COSI) in Columbus, OH, United States and tested in a quiet lab within the museum. Caregivers completed a demographic questionnaire about their child's language background. The study was approved by the local IRB.

Sentences were presented at a comfortable listening level under Audiotechnica headphones (model 8TH-770COM). Children were randomly assigned to either quiet or +4 dB SNR, and to one of four accent conditions all of which contained Midland: 1) German and Scottish, 2) Mandarin and Bilingual Hindi, 3) British and Japanese, or 4) French and Spanish. Prior to testing, children completed nine practice trials representing three sentences from each accent of their assigned condition. Twenty test sentences were contributed by each talker per condition (60 total). Sentences were randomly blocked by accent and randomized within each block. They were asked to repeat each sentence as best they could and guess when necessary. An experimenter recorded each response in real time and scored responses offline. Words were scored as incorrect if they included added or deleted phonemes. Only a(n)/the, has/had, have/had, are/were, and is/was were exceptions to this rule, consistent with the HINT-C scoring method.

Note that some of the data from the younger group have been reported previously: British and Japanese [28], and Mandarin, Bilingual Hindi, German, and Scottish [29]. All data were collected on a Dell Optiplex 790 Desktop Computer using Pavlovia (http://pavlovia.org/) or E-Prime 2.0 [30]. Data not reported elsewhere include all conditions from the older group and the French and Spanish condition from the younger group. Data were combined across studies to more fully understand developmental differences in accent intelligibility across many accent varieties.

3. RESULTS

Mean accuracy and standard error are presented across all 9 accents as a function of age group in Figure 1. Word recognition accuracy was analyzed using a generalized linear mixed effects model with a logit link function (see Table 1). Random effects included age group (older vs. younger, which was the reference) and accent (9 levels, dummy coded with Midland as the reference). Listening condition (noise vs. quiet, which was the reference) was included as a fixed effect. Participant and item (word) were included as random intercepts.

As expected, older children had better word recognition accuracy than younger children (χ^2 = 58.44, p < .001) and word recognition accuracy was better in quiet than in noise (χ^2 = 159.32, p < .001). The random effect of accent was also significant (χ^2 = 798.46, p < .001). Accuracy was better for Midland than all other accents (ps < .001) except German (p = .418), a non-native accent that, like English, is a Western Germanic language. Four accents interacted with age, reflecting that relative to older children, younger children had poorer word recognition with one unfamiliar, native accent (British), two non-native varieties (Mandarin and Spanish), and Bilingual Hindi. The seven two-way interactions between accent and listening condition suggest that children overall have more difficulty with the majority of the accents tested in noise relative to quiet. Two three-way interactions were significant (age, listening condition, and Mandarin; age, listening condition, and Spanish) indicating that

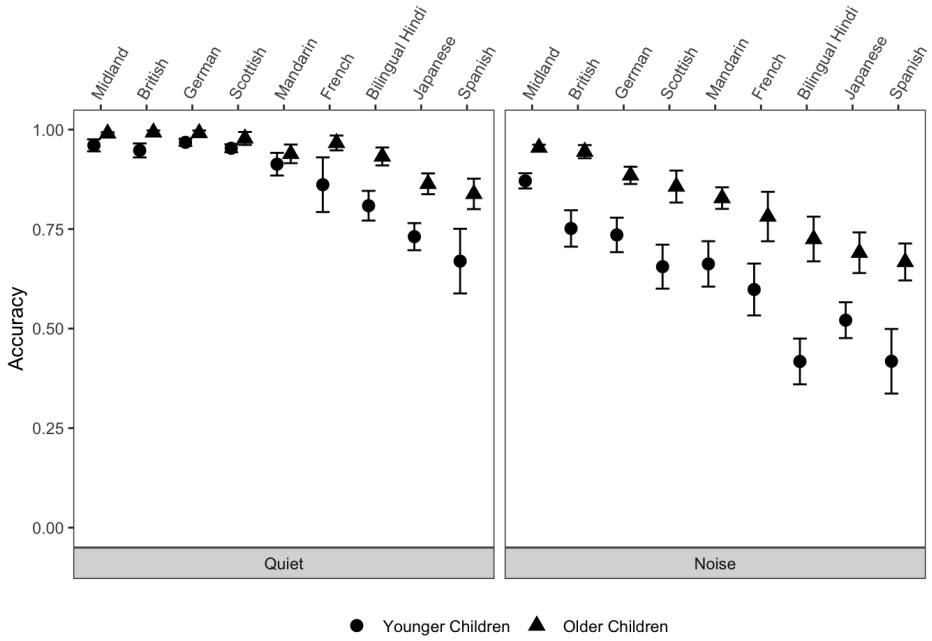


Figure 1: Mean word recognition accuracy (proportion correct) +/- 1 standard error across all 9 accents as a function of age group. Circles reflect younger children's performance and triangles that of the older children. Accuracy in quiet is shown in the left-hand panel and accuracy in +4 dB SNR is shown in the right-hand panel.

noise influences younger and older children's word recognition differently for two of the non-native varieties relative to Midland.

	Estimate	SE	*z*-value	*p*-value
Intercept (Midland)	4.25	.12	36.30	< .001
Age (Older)	1.25	.16	7.64	< .001
Condition (Noise)	-1.64	.13	-12.62	< .001
Acc (British)	-.47	.13	-3.47	< .001
Acc (Scottish)	-.70	.16	-4.77	< .001
Acc (Mand)	-.96	.12	-8.21	< .001
Acc (French)	-.53	.16	-3.40	< .001
Acc (BL Hin)	-2.06	.12	-16.88	< .001
Acc (Japanese)	-2.75	.14	-20.28	< .001
Acc (Spanish)	-2.43	.16	-15.28	< .001
Age x Acc (British)	.85	.33	2.58	.010
Age x Acc (Mand)	-.83	.19	-4.29	< .001
Age x Acc (BL Hin)	.40	.19	2.05	.040
Age x Acc (Spanish)	-.53	.21	-2.52	.012
Acc (British) x Cond	-.77	.13	-6.14	< .001
Acc (German) x Cond	-1.08	.14	-7.48	< .001
Acc (Scottish) x Cond	-1.11	.14	-8.19	< .001
Acc (Mand) x Cond	-.60	.11	-5.39	< .001
Acc (French) x Cond	-1.23	.16	-7.58	< .001
Acc (BL Hin) x Cond	-1.06	.11	-9.78	< .001
Acc (Spanish) x Cond	-.43	.15	-2.83	.005
Age x Mand x Cond	.73	.23	3.15	.002
Age x Spanish x Cond	.73	.25	2.85	.004

Table 1: Significant results of the mixed effects model, which included fixed effects for listening condition, and random effects for age group and accent, as well as all two- and three-way interactions. *Note:* Acc = Accent, Mand = Mandarin, BL = Bilingual, Cond = Condition, Hin = Hindi.

4. DISCUSSION

This study is the first to include both a relatively large number of unfamiliar native and non-native accents in a spoken word recognition task and a large group of adolescent participants to examine developmental differences in accent intelligibility cross- sectionally. Intelligibility of accents is important to study developmentally because understanding speech of people with different language and dialect histories is a real-life, everyday skill needed to communicate in a global society. Including many unfamiliar accent varieties, both native and non-native, is valuable for generalizability, as well as for allowing a more nuanced understanding of accent perception development.

The results suggest that in quiet both groups maintain relatively high spoken word recognition accuracy with a variety of unfamiliar accents but show slightly poorer word recognition with a subset of the non-native talkers. The younger children also showed some decrement in word recognition with the bilingual talker. These results contrast with previous findings for adults showing that adult listeners tend to show high word recognition accuracy with unfamiliar native and non-native accents in quiet and only show decrements in noise-added conditions [1,19]. For both groups of children, the addition of noise resulted in substantial decrements in word recognition accuracy relative to performance in quiet for most of the unfamiliar accents. There was some evidence that word recognition for Spanish- and Mandarin-accented English talkers are influenced differently by noise for the two age groups of children. The underlying causes for why listening condition and age interact with these two accents but not the others should be explored in future research. In particular, the contribution of factors such as the acoustic-phonetic characteristics of the accents, children's prior familiarity with the accents, and sociolinguistic variables should be investigated.

These results suggest that word recognition for talkers with unfamiliar accents improves during early and middle childhood, particularly in noise, which is a more common listening environment than quiet, and that improvement from early childhood to later childhood does not occur to the same degree across accents. Moreover, while non-native accents tend to be more difficult to understand than native ones, there is no evidence of a native/non-native divide. For example, the German-accented talker was one of the easier speakers to understand for both age groups.

Future research should examine adults' word recognition of these materials to determine if, when, and under what conditions adolescents attain adult-like intelligibility scores with these accents. Still, these data support previous studies [11-14] that accent perception has a protracted period of development. Further, it builds on prior work by investigating both unfamiliar native and non-native varieties in the same paradigm. Follow-up work should examine the acoustic-phonetic characteristics of the accents that contribute to the intelligibility differences between the younger and older listener groups to provide insight into the developmental changes underpinning the improvements observed with development.

5. ACKNOWLEDGEMENTS

Supported by the National Science Foundation (Award numbers: 1941691 [Bent], 1941662 [Holt]). We are grateful to contributions of Xiwei Chen to

the statistical analyses, and to Emma Canino, Jada Hudgins, Katherine Miller, Mone Skratt Henry, Melissa Martin, and Polina Surikova to data collection. We would also like to acknowledge the Center of Science and Industry and the children and families who participated in this work.

6. REFERENCES

[1] Adank, P., Evans, B. G., Stuart-Smith, J., Scott, S. K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J Exp Psychol. Human Percept. Perform.* 35, 520-529.

[2] Bent, T., Holt, R. F. 2018. Shhh… I need quiet! Children's understanding of American, British, and Japanese-accented English Speakers. *Lang. Speech.* 61, 657-673.

[3] Nathan, L., Wells, B., Donlan, C. 1998. Children's comprehension of unfamiliar regional accents: A preliminary investigation. *J. Child Lang.* 25, 343-365.

[4] Butler, J., Floccia, C., Goslin, J., Panneton, R. 2011. Infants' discrimination of familiar and unfamiliar accents in speech. *Infancy.* 16, 392-417.

[5] Kinzler, K. D., Dupoux, E. Spelke, E. S. 2007. The native language of social cognition. *Proc. Nat. Acad. Sci. USA.* 104, 12577-12580.

[6] van Heugten, M., Johnson, E. K. 2014. Learning to contend with accents in infancy: Benefits of brief speaker exposure. *J. Exp. Psychol.: Gen.* 143, 340-350.

[7] Bent, T. 2014. Children's perception of foreign-accented words. *J. Child Lang.* 41, 1334-1355.

[8] Bent, T. Atagi, E. 2017. Perception of nonnative-accented sentences by 5- to 8-year-olds and adults: The role of phonological processing. *Lang. Speech.* 60, 110-122.

[9] Holt, R. F., Bent, T. 2017. Children's use of semantic context in perception of foreign-accented speech. *J. Speech Lang. Hear. Res.* 60, 223-230.

[10] Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., Quann, C. A. 2009. Development of phonological constancy: Toddlers' perception of native- and Jamaican-accented words. *Psychol. Sci.* 20, 539-542.

[11] Bent, T. 2018. Development of unfamiliar accent comprehension continues through adolescence. *J. Child Lang.* 45, 1400-1411.

[12] Jones, Z., Yan, Q. Y., Wagner, L., Clopper, C. G. 2017. The development of dialect classification across the lifespan. *J. Phon.* 60, 20-37.

[13] McCullough, E. A., Clopper, C. G., Wagner, L. 2019. Regional dialect perception across the lifespan: Identification and discrimination. *Lang. Speech.* 62, 115-136.

[14] McCullough, E. A., Clopper, C. G., Wagner, L. 2019. The development of regional dialect locality judgments and language attitudes across the life span. *Child Dev.* 90. 1080-1096.

[15] Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., Irwin, J. R. 2013. Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Dev.* 84, 2064-2078.

[16] van Heugten, M., Johnson, E. K. 2016. Toddlers' word recognition in an unfamiliar regional accent: the role of local sentence context and prior accent exposure. *Lang. Speech.* 59, 353-363.

[17] McQueen, J. M., Tyler, M. D., Cutler, A. 2012. Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds. *Lang. Learn. Dev.* 8, 317-339.

[18] White, K. S., Aslin, R. N. 2011. Adaptation to novel accents by toddlers. *Dev. Sci.* 14, 372-384.

[19] Bent, T., Atagi, E. 2015. Children's perception of nonnative-accented sentences in noise and quiet. *J. Acoust. Soc. Amer.* 138, 3985-3993.

[20] O'Connor, C., Gibbon, F. E. 2011. Familiarity of speaker accent on Irish children's performance on a sentence comprehension task. *J. Clin. Speech Lang. Stud.* 18, 1-17.

[21] Huyck, J. J., Wright, B. A. 2011. Late maturation of auditory perceptual learning. *Dev. Sci.* 14, 614-621.

[22] Johnson, C. E. 2000. Children's phoneme identification in reverberation and noise. *J. Sp. Lang. Hear. Res.* 43, 144-157.

[23] Hazan, V., Barrett, S. 2000. The development of phonemic categorization in children aged 6-12. *J. Phon.* 28, 377-396.

[24] Fisher, W. M., Doddington, G. R., Goudie-Marshal, K. M. 1986. The DARPA speech recognition database: Specifications and status. *Proceedings of the DARPA speech recognition workshop,* 93-99.

[25] Aslin, R. N., Smith, L. B. 1988. Perceptual development. *Ann. Rev. Psych.* 39, 435-473.

[26] Nilsson, M., Soli, S. D., Gelnett, D. J. 1996. *Development of the Hearing in Noise Test for Children (HINT-C).* Los Angeles, CA: House Ear Institute.

[27] Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., Chandrasekaran, B. 2014. Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Lang. Acquist.* 23, 89-111.

[28] Bent, T., & Holt, R. F. (2018). Shhh… I need quiet! Children's understanding of American, British, and Japanese-accented English speakers. *Lang. Sp.* 61, 657-673.

[29] Bent, T., Holt, R. F., Van Engen, K. J., Jamsek, I. A., Arzbecker, L. J., Liang, L., Brown, E. 2021. How pronunciation distance impacts word recognition in children and adults. *J. Acoust. Soc. Amer.* 150, 4103-4117.

[30] Psychology Software Tools. 2007. *E-Prime* Version 2.0.

PERCEIVING SPEECH PRODUCED WITH FACE MASKS IN COMPETING TALKER ENVIRONMENTS

Faith Chiu[1], Laura Bartoševičiūtė[2], Albert Lee[3], Yujia Yao[1]

[1]Glasgow University Laboratory of Phonetics, University of Glasgow, [2]University of Essex, [3]The Education University of Hong Kong
faith.chiu@glasgow.ac.uk, albertlee@eduhk.hk

ABSTRACT

This paper examines the perception of speech produced with face masks in everyday multi-talker environments. Three groups of participants listened to English target sentences produced with or without a face mask in the presence of English or Lithuanian competing speech. Participants were monolingual English listeners, and second language English listeners with either Lithuanian or Mandarin Chinese as first language (L1). Lithuanian listeners also completed the experiment with Lithuanian target sentences. Participants were generally more accurate perceiving sentences produced without a face mask, and when listening in L1. Competing speech in a language matching the target lowered perception accuracy. Exceptionally, only when Lithuanian participants (with both English and Lithuanian knowledge) listened for Lithuanian targets was there no added challenge from matching language of target and competing speech. We conclude that acoustic distortions from face masks present an across-the-board difficulty while linguistic knowledge can reduce distraction from competing speech.

Keywords: face masks, competing talker, speech perception

1. INTRODUCTION

The COVID-19 pandemic has reshaped speech communication. Face-to-face communication often includes one or both parties sporting a face mask. The listener's comprehension effort now involves adapting to mask-imposed distortions to the acoustic speech signal [1, 2], not to mention other challenges such as deprivation of visual cues. It has been shown that individuals struggle with understanding speech produced with a face mask when presented in noise [3], even for native speakers [4]. Furthermore, most everyday speech communication also takes place in less than ideal listening environments such as a noisy background with a competing talker. It is also not uncommon these days to converse in one's second or additional language. Previous research has found that non-native listeners perform significantly worse than native listeners in perceiving speech presented with intelligible competing background speech [5, 6]. This study aims to understand the difficulty imposed by speech produced with face masks in an everyday multi-talker environment. Target sentences produced with and without a face mask were presented to listeners in the presence of a competing talker. The competing speech either matched or differed in language from target sentences. Participants' linguistic background determined the intelligibility of the competing talker.

2. METHOD

2.1. Stimuli

The speech material consisted of target sentences in English and in Lithuanian, and competing speech in English and in Lithuanian.

2.1.1. Target sentences

English target sentences were based on the British English version of the International Matrix sentence test for speech audiometry in noise [7]. This standardised test consists of a 50-word base matrix (10 names, 10 verbs, 10 numerals, 10 adjectives, and 10 nouns) from which grammatically correct but contextually unpredictable five-word sentences can be built, using a random combination of one word of each word category. All matrix sentences have fixed syntactic structure ('Alan bought two big beds'; name + verb + number + adjective + noun). Up to 100,000 different sentences can be generated.

Name	Verb	Number	Adjective	Noun
Alan	bought	two	big	beds
Barry	gives	three	cheap	chairs
Hannah	got	four	dark	desks
Kathy	has	five	green	mugs
Lucy	kept	six	large	rings
Nina	likes	eight	old	ships
Peter	sees	nine	pink	shoes
Rachel	sold	ten	red	spoons
Steven	wants	twelve	small	tins
Thomas	wins	some	thin	toys

Table 1: Matrix for English target sentences.

Lithuanian target sentences (only presented to Lithuanian listeners) follow the same format and are constructed as original stimuli for this experiment.

Name	Verb	Number	Adjective	Noun
Ieva	laiko	kelis	pilkus	raktus
Miglė	rodo	visus	juodus	krepšius
Agnė	turi	devynis	žalius	peilius
Dalia	mato	du	sunkius	ratus
Paulius	rado	tris	svarbius	batus
Marius	perka	penkis	mažus	rūbus
Nojus	gavo	šešis	naujus	šaukštus
Benas	neša	keturis	senus	žaislus
Juozas	davė	aštuonis	pigius	stalus
Rūta	piešia	septynis	baltus	daiktus

Table 2: Matrix for Lithuanian target sentences.

Audio stimuli from these target sentences were generated by recording a native female speaker of each language in sound attenuated booths. Individual words were produced in sentence frames specified above, with and without a cotton fabric face mask. The recordings were segmented and recombined on Praat [8] to generate the presented stimuli.

2.1.2. Competing speech

Competing speech was semantically meaningful sentences in either English or Lithuanian. These sentences were presented continuously at a difficult −10dB Signal-to-Noise ratio, meaning that the competing speech was 10dB louder than the presented target sentences. Male voices were chosen so participants can utilise speaker sex differences as a segregation cue. Low-level acoustic cues such as fundamental frequency differences can help with separating and tracking sentences in competition [9].

English competing speech contained sentences from the IEEE lists [10] produced by a native British English male speaker [11]. Lithuanian competing speech consisted of text being read aloud by a native Lithuanian male speaker from the LIEPA corpus [12]. Both these speakers did not wear a face mask.

2.2. Participants

2.2.1. Experiment 1

Twenty-four native Lithuanian listeners (thirteen female and eleven male) completed Experiment 1. All participants were between 18 and 37 years old and reported no history of hearing or language impairment. All participants had second language knowledge of English, and had completed either high school or tertiary education.

2.2.2. Experiment 2

Participants were twenty-two monolingual British English speakers (sixteen female and six male, age range: 18–34) and twenty-two second language speakers of English with Mandarin Chinese as first language (nineteen female and three male, age range: 20–31). All participants reported no history of hearing or language impairment and had completed either high school or tertiary education.

2.3. Procedure

2.3.1. Experiment 1

The experiment was conducted online using Psytoolkit [13, 14]. Participants were instructed to complete the task in a quiet room with headphones, and at a computer with keyboard input. After initial surveys on demographic information and English experience and proficiency (abbreviating LEAP-Q [15]), the experimental session started with a headphone check [16]. Then, participants were told that they would hear target sentences by a female talker in the presence of a male competing talker. The target sentences were cued by sex of the speaker. Participants were instructed to listen only to the female voice and ignore the male voice. They were then asked to type what they heard after each sentence, and were requested to report individual words if they had not heard the whole sentence.

In the main task, participants heard a total of 160 trials: from 2 Target language conditions (ENGLISH/LITHUANIAN) × 2 Mask conditions (YES/NO) × 2 Competing Speech language conditions (ENGLISH/LITHUANIAN) with 20 sentences each. These 160 trials were presented over four blocks of 40 sentences each. Presentation order of blocks were counterbalanced across participants and trials were randomised within each block. The task was self-paced and participants heard each sentence only once; there was no repetition of sentences in the same configuration. No feedback was provided. The total duration of the experimental session, without breaks, was around forty-five minutes. Participants were allowed to take breaks between blocks.

2.3.2. Experiment 2

The procedure in Experiment 2 was identical to Experiment 1, except that participants heard only English target sentences produced with and without mask, in the presence of either competing English or Lithuanian speech. There were thus a total of 80 main trials in Experiment 2.

2.4. Data scoring

Responses were scored based on the number of words accurately reported in each sentence. There

were five keywords per sentence; a full mark on a sentence was scored as 100%. Each error took away 20% (4 accurate keywords = 80%). Keywords were considered accurate regardless of the position they appeared in the participant's response. If the participant missed some of the keywords in the sentence, the remaining accurately reported words were still scored (e.g. 'Lucy small desks'). Variations in spelling (e.g. 'Kathy' as 'Cathy') were considered an accurate response. Phonologically similar words that change the meaning of the word were considered incorrect (e.g. 'wings' for 'rings').

3. RESULTS

3.1. Experiment 1

Lithuanian listeners' perception performance is given in Figure 1.

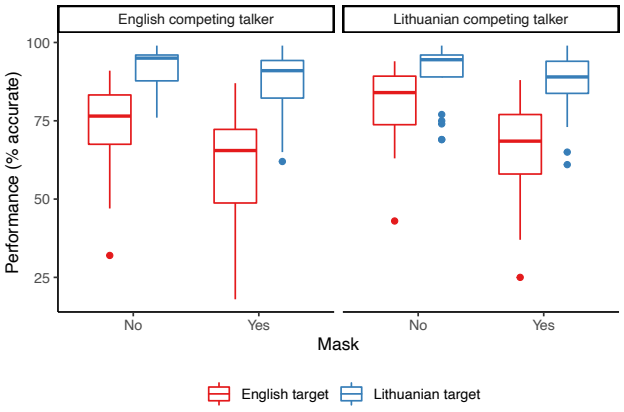


Figure 1: Performance (% accurate) of Lithuanian listeners across all conditions.

A 2 × 2 × 2 analysis of variance (ANOVA) was conducted on the percentage of accurately reported keywords as a function of language of Target (English versus Lithuanian), Mask (with or without a face mask), and language of Competing Speech (English versus Lithuanian). The results revealed two significant two way-interactions: Target × Mask ($F(1, 23) = 26.001$, $p < .001$, $\eta p^2 = .531$) and Target × Competing Speech ($F(1, 23) = 25.123$, $p < .001$, $\eta p^2 = .522$). All three main effects were significant, for Target ($F(1, 23) = 59.409$, $p < .001$, $\eta p^2 = .721$), Mask ($F(1, 23) = 59.536$, $p < .001$, $\eta p^2 = .721$) and Competing Speech ($F(1, 23) = 18.771$, $p < .001$, $\eta p^2 = .449$). The three-way interaction Target × Mask × Competing Speech was not significant, nor was the two-way interaction Mask × Competing Speech (both $p > .050$).

Individual 2 × 2 ANOVAs were performed for each Target language in order to follow up these observations, comparing the percentage of accurately reported keywords as a function of Mask (with or without a mask) and language of Competing Speech (English versus Lithuanian). For English target sentences, the interaction term was not significant ($p > .050$). There was a significant main effect of Mask ($F(1, 23) = 54.448$, $p < .001$, $\eta p^2 = .703$). More keywords were accurately reported on sentences produced without a face mask, in both English and Lithuanian competing speech. There was also a main effect of Competing Speech ($F(1, 23) = 31.304$, $p < .001$, $\eta p^2 = .576$). Lithuanians listeners were less accurate when the competing speech was in a language which matches the English target sentences; this was true both when the targets were produced with and without a mask.

When listening to Lithuanian target sentences, there was only a main effect of Mask on the percentage of accurately identified keywords ($F(1, 23) = 15.544$, $p < .001$, $\eta p^2 = .403$). Lithuanian target sentences produced with a face mask were more poorly perceived, and this was true regardless of whether it was presented in both English and Lithuanian competing speech. Here, unlike the case with English target sentences, there was no main effect of Competing Speech. The interaction term was also not significant. (Both $p > .050$.) Thus, only when Lithuanian listened in their first language was there no added challenge from matching language of target sentences and competing speech.

In addition, Lithuanian listeners reported more accurate keywords when listening to target sentences in their first language (Lithuanian) compared to English. Running planned comparisons, this was found in all four conditions: when target sentences were produced with a mask and presented in English competing speech ($t(23) = 8.730$, $p < .001$) and in Lithuanian competing speech ($t(23) = 6.890$, $p < .001$), and when target sentences were produced without a face mask and delivered with an English ($t(23) = 6.834$, $p < .002$) and Lithuanian competing talker ($t(23) = 4.314$, $p < .001$).

3.2. Experiment 2

Figure 2 shows the percentage of accurately reported keywords by English and Chinese listeners.

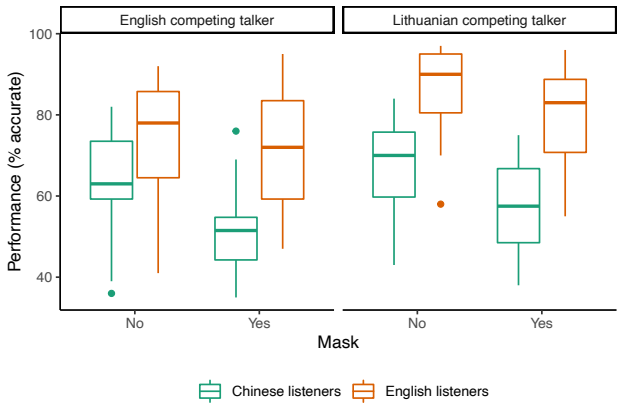


Figure 2: Performance of English listeners and Chinese listeners across all conditions.

A mixed 2 × 2 × 2 ANOVA was conducted on the percentage of accurately reported keywords as a function of Mask (with or without a face mask), language of Competing Speech (English versus Lithuanian), and Group (English versus Chinese listeners). The results indicated a significant three-way interaction of Mask × Competing Speech × Group ($F(1, 42) = 6.497$, $p = .015$, $\eta p^2 = .134$).

Planned comparisons showed English listeners outperforming Chinese listeners in all conditions: when target sentences were produced with a face mask and presented in English competing speech ($t(42) = 5.009$, $p < .001$) and in Lithuanian competing speech ($t(42) = 6.162$, $p < .001$), and when target sentences were produced without a face mask and delivered with an English ($t(42) = 2.767$, $p = .008$) and Lithuanian competing talker ($t(42) = 5.776$, $p < .001$).

Individual 2 × 2 ANOVAs were performed for each listener group. For English listeners, there was a main effect of Mask ($F(1, 21) = 5.439$, $p = .030$, $\eta p^2 = .206$) as well as a main effect of Competing Speech ($F(1, 21) = 78.729$, p. < .002, $\eta p^2 = .789$). The interaction term was not significant ($p > .050$). Chinese listeners were similar; there was a main effect of Mask. They too reported more accurate keywords when the target sentences were produced without a mask. This held, both when the target was presented in English and in Lithuanian competing speech ($F(1, 21) = 21.960$, $p < .001$, $\eta p^2 = .511$). There was also a main effect of Competing Speech in Chinese listeners ($F(1, 21) = 19.869$, $p < .001$, $\eta p^2 = .486$). Like English listeners, there was more of a detrimental effect on perception when the competing talker matched the language of the target sentences produced with and without a mask.

4. DISCUSSION

4.1. Across-the-board difficulty in perceiving speech produced with face masks

Overall, a main effect of Mask was found across all participant groups: in monolingual English listeners, and in second language English listeners with either Lithuanian or Chinese as first language. Target sentences produced with a mask were less accurately perceived. With Lithuanian listeners, the Mask effect was observed for target sentences both in their first (Lithuanian) and second (English) languages.

This finding echoes other reports of decreased perception performance when listening to speech produced with a face mask and presented in noise [3, 4]. Notably, all listener groups in all conditions experienced a drop in performance in perceiving speech produced with masks. This across-the-board effect could be due to attenuation of the acoustic signal from mask-wearing in the form of dampening. In particular, high frequency information is lost [17]. Further data analysis and future iterations of this study could consider specifically the identification of keywords involving fricatives and plosives. These segments tend to be commonly misperceived when produced with a face mask even in quiet [18]. The study could also benefit from additional auditory and acoustic analyses of the target sentence stimuli in terms of speaking style. Clear speech strategies have been documented with speakers wearing face masks and this instead increases signal intelligibility [19].

4.2. Target and competing talker language similarity

We also found that perception accuracy was higher when listening in one's first language. In Experiment 1, Lithuanian listeners reported more accurate keywords to Lithuanian rather than English target sentences. In Experiment 2, the native English group performed better than the second language group which had Chinese as first language. These results corroborate the findings of previous work [5, 6] showing that speech perception with a competing talker is more difficult in one's non-native language.

Additionally, we found that a competing talker in a language which matches the target sentences had more of a detrimental effect on perception accuracy compared to one that was mismatched. This was found in all listener groups (Lithuanian, English and Chinese) when listening to English target sentences. The results with Lithuanian and Chinese listeners on English target sentences replicate existing findings documenting a benefit of linguistic mismatch between target and competing speech for non-native speakers [20, 21, 22]. This is in line with Brouwer and colleagues' linguistic similarity hypothesis [20].

However, there was no effect of Competing Speech language in our study when Lithuanians listened to target sentences in their native language. We believe this is due to the group's knowledge of both the target and competing talker languages. Our study contradicts [20] in their native bilingual group. When Dutch speakers of second language English listened in their native language, they behaved instead like monolinguals presented with a competing talker in a foreign unintelligible language [20] (i.e. less interference listening to English-on-Dutch as opposed to Dutch-on-Dutch). To validate our finding, we could attempt replicating our results with another speaker for Lithuanian target sentences or to include different competitor talkers. This will help ensure that the observed effects are not just due energetic masking differences across our two competing talkers: that the English competing talker was not merely more effective at swamping (rendering more time-frequency regions inaudible from more overlaps of the target signal), and that the Lithuanian competing talker has not just allowed more glimpses of the target signal (regions in which the target is least affected) [23].

Acknowledgements: This research was supported by a British Academy grant awarded to FC and AL (COV19\2008370).

5. REFERENCES

[1] Corey, M. R., Jones, U., Singer, C. A. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am.* 148, 2371–2375.

[2] Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C. S., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., Kenyon, K., Caverlé, M. W., Schultz, B. G., Vogel, A. P. 2020. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *J. Acoust. Soc. Am*. 148, 3562–3568.

[3] Toscano, J.C., Toscano, C.M. 2021. Effects of face masks on speech recognition in multi-talker babble noise. *PLOS ONE* 16, e0246842.

[4] Yi, H., Pingsterhaus, A., Song, W. 2021. Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Front. Psychol.* 12, 682677.

[5] Bradlow, A. R., Alexander, J. A. 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J Acoust. Soc. Am.* 121, 2239–2249.

[6] Cooke, M., Garcia Lecumberri, M.L., Barker, J. 2008. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.* 123, 414–427.

[7] Hall, S. J. 2006. *The development of a new English Sentence in Noise Test and an English Number Recognition Test*. MSc thesis. University of Southampton.

[8] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03.

[9] Brungart, D.S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am.* 109, 1101–1109.

[10] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*. 17, 225–246.

[11] Millman R.E., Mattys S.L. 2017. Auditory Verbal Working Memory as a Predictor of Speech Perception in Modulated Maskers in Listeners With Normal Hearing. *J. Speech Lang. Hear. Res.* 60, 1236–1245.

[12] Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V. 2018. Lithuanian Speech Corpus Liepa for Development of Human-Computer Interfaces Working in Voice Recognition and Synthesis Mode. *Informatica* 29, 487–498.

[13] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behav. Res. Methods* 42, 1096–1104.

[14] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teach Psychol.* 44, 24–31.

[15] Marian, V., Blumenfeld, H.K., Kaushanskaya, M. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967.

[16] Woods, K.J.P., Siegel, M.H., Traer, J., McDermott, J.H. 2017. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* 79, 2064–2072.

[17] Corey, R.M., Jones, U., Singer, A.C. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am.* 148, 2371–2375.

[18] Llamas, C., Harrison, P., Donnelly, D., Watt, D. 2008. Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistics* 2, 80–104.

[19] Cohn, M., Pycha, A., Zellou, G. 2021. Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition* 210, 104570.

[20] Brouwer, S., Van Engen, K.J., Calandruccio, L., Bradlow, A.R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *J. Acoust. Soc. Am.* 131, 1449–64.

[21] Calandruccio, L., Zhou, H. 2014. Increase in speech recognition due to linguistic mismatch between target and masker speech: monolingual and simultaneous bilingual performance. *J. Speech Lang. Hear. Res.* 57, 1089–1097.

[22] Van Engen, K.J. 2010. Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Commun.* 52, 943–953.

[23] Cooke, M. 2006. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* 119, 1562–73.

PRODUCTION AND PERCEPTION OF LONG AND SHORT VOWEL CONTRAST IN MONGOLIAN

Ao Min[1], Dahu Baiyila[2], Aijun Li[3]

[12]Inner Mongolia University, China
[3]University of Chinese Academy of Social Sciences (UCASS) and IL, CASS, China
aomin@imu.edu.cn dabhvrbayar@163.com liaj@cass.org.cn

ABSTRACT

Long and short vowels are phonologically contrasted in many languages, but their related acoustic features are not always identical. In this study, we aim to explore the encoding and decoding mechanism of long-short vowel contrast in Mongolian. To this end, production and perception experiments were carried out for long-short vowel contrast which were embedded in meaningful and meaningless words.

This study found that, firstly, vowel duration is the most important acoustic production parameter, which has a relatively stable duration ratio pattern between long and short vowels. Secondly, the perception on a physical continuum of duration for meaningful and meaningless words indicates that there is a clear categorical boundary, while long vowels have a broader perceptual space than that of short vowels. Finally, there is a strong complementary relationship between word meaning and vowel duration in the process of long-short vowel encoding and decoding.

Keywords: Mongolian, long-short vowel contrast, acoustic feature, categorical perception

1. INTRODUCTION

1.1. Long-short vowel contrast

Long-short vowel contrast is ubiquitous in the vowel systems of world ' s languages, such as Uyghur, English, Korean, Czech, Finnish, Swedish, Arabic, Danish, Estonian and Ewe. Studies show that the difference between long and short vowels lies in duration, F0, formant, accented or not, adjacent consonants, and prosodic boundary types, among which duration is the most stable distinguishing feature [1]. Although duration is positively correlated with change of F0 for both long and short vowels in Swedish [2], there is a relatively stable duration ratio of 2.4: 1 between them in Japanese [3].

On the other hand, duration of long and short vowels change under individual conditions. The adjacent consonant and syllable structure will influence the acoustic duration pattern between long and short vowel [4]. However, few studies have been done on the variation patterns of vowel duration between long-short vowel contrast in continued speech.

1.2. Long-short vowel contrast in Mongolian

There are stable long-short vowel contrast in Mongolian, one of the Mongolian languages in Altaic language family [5,6]. However, the pairs of long and short vowels and specific vowels are differ in Mongolian dialects. There are 8 pairs of long-short vowel contrast in standard Mongolian [7], while there are 11 pairs in Khalkha dialect [8]. For example, the duration contrast of vowel /ɐ/ distinguishes the meanings between the words "ab" (/ɐw/, means " Take ") and "abu" (/ɐ:w/, means " Father "). Figure 1 is the spectrogram, where the vowels of both words have basically the same formant pattern but remarkably different durations.

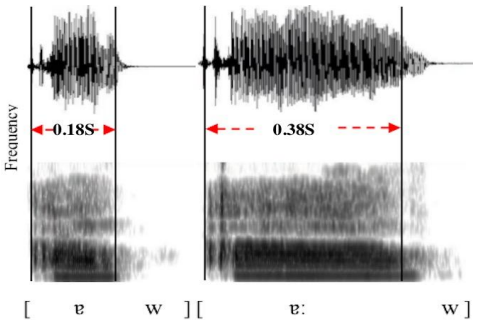


Figure 1: The spectrogram of the words /ɐw/ (left side, means " Take") and /ɐ:w/ (right side, means"Father") .

Long-short vowel contrast in Mongolian is widely studied in both segmental and suprasegmental aspects. For example, it is found that long and short vowels in Mongolian are different not only in duration, but also in tongue position [9]. According to the length of duration, vowels can be divided into ultra-long vowels, long vowels, short vowels and ultra-short vowels [10]. The prolongation of non-syllabic vowel has prosodic function in Mongolian [11]. However, there are many problems to be further discussed about the duration of long-short vowel contrast in Mongolian.

To this end, this study investigates the duration patterns and the encoding and decoding mechanism

of long-short vowel contrast in Mongolian under the regulation of word meaning.

2. METHOD

2.1. Corpus design

The speech data of this study includes seven short vowels in Mongolian, namely /ɐ/, /ə/, /i/, /ɔ/, /ʊ/, /o/ and /u/, and their corresponding long counterparts /ɐː/, /əː/, /iː/, /ɔː/, /ʊː/, /oː/ and /uː/. Two kinds of materials were designed based on monosyllabic words in fixed context, including 50 pairs of meaningful words (common words) and 116 pairs of meaningless words, with long-short contrast vowels in the same context. The meaningful words include 15 pairs of VC and 35 pairs of CVC, such as el_e(/əl/, each) and egeli(/əːl/, affectionate), and the meaningless words include 84 pairs of CVC and 32 pairs of VC obtained by the combination of /m/, /s/, /ʧʰ/ and /p/ with seven pairs of contrasted vowels respectively, such as /sɐs/ ~ /sɐːs/.

2.2. Production and perception experiments

Six speakers (gender balanced) participated in the production experiment. Another twenty listeners participated in the perception experiment. All the speakers, aged 20 to 30, were students with the Class I Type A certificate of standard Mongolian test. They didn't live in other dialect areas before going to senior high school. Speakers read each word which randomly presented on the screen three times at a normal speed. Finally 5,976 speech sounds were collected.

The perception experiment was conducted on duration continuum manipulated from 30 meaningful and 28 meaningless words with long vowel in CVC context by linear test method [12]. Altogether 10 stimulus sounds were synthesized for each vowel contrast. finally 920 stimuli were made in Praat and 11,600 perceptual results were obtained.

2.3. Tools for data collection and processing

Three kinds of experimental tools were used in the data processing: xSpeech tools [13] for recording, annotation and perception experiments; ProsodyPro [14] for extracting vowel duration. Formulas (1) for calculating the prolongation rate of vowel duration and formulas (2) for calculating the multiple of duration between long and short vowels are shown below [15].

(1) $$VDxv = \frac{VD_{LV} - VD_{SV}}{VD_{LV-SV}} \times 100\%$$

(2) $$K_V = \frac{VD_{LV}}{VDsv}$$

In Formula (1), VD_{XV} is the prolongation rate of the average vowel duration for a certain long-short vowel contrast (xv), VD_{LV} and VD_{SV} are the average duration of long-short vowel contrast (xv) respectively, and $VD_{LV\text{-}SV}$ represents the difference between the average duration of all long and short vowels. In Formula (2), K_V is the multiple of duration between long and short contrasted target vowels.

3. RESULTS

3.1. Acoustic features of vowel production

The result of acoustic analysis are shown in Table 1, and the values in brackets are meaningful monosyllabic words, while the values before the brackets are meaningless monosyllabic words. Each value is an average from three repeated readings by six speakers.The IF, PR and PD represent for influence factors, prolongation rate and prolongation direction,respectively.

No	IF	Type	Mean duration	PR	PD
1	Lip shape	L	Unround: 0.227 (0.286) Round: 0.231 (0.261)	1.5% (10%)	+(+)
		S	Unround: 0.123 (0.116) Round: 0.123 (0.104)	0.2% (11%)	+(+)
2	Tongue high and low dimensions	L	High: 0.232 (0.657) Medium: 0.227 (0.283) Low: 0.226 (0.305)	1.5% (6.5%)	+(+)
		S	High: 0.123 (0.657) Medium: 0.123 (0.122) Low: 0.123 (0.137)	0.2% (11%)	+(+)
3	Tongue front and low dimensions	L	Front: 0.228 Middle : 0.227 Back: 0.230	0.80%	+(+)
		S	Front: 0.123 Middle: 0.2123 Back: 0.123	0.20%	+(+)
4	Adjacent consonants	L	Unvoiced: 0.268 (0.238) Voiced: 0.227 (0.265)	18%	+(-)
		S	Unvoiced: 0.151 (0.078) Voiced: 0.097 (0.0965)	16%	-(+)
5	Syllable features	L	VC: 0.301 (0.301) CVC: 0.252 (0.251)	15% (6.5%)	-(-)
		S	VC: 0.146 (0.147) CVC: 0.087 (0.087)	16% (11%)	-(-)

Table 1: The influence degree and statistical significance of each factor on vowel durations.

The results indicate that adjacent consonants, syllable length and word meaning have a statistically significant influence on vowel duration in the processing of acoustic production although the duration is affected by multiple factors in varying degrees. First, the influence of vowel inherent features on vowel duration is not statistically significant (P>0.05, df=426, F=0.004). Although there are differences in the actual vowel duration in the classification of vowel inherent features, the dispersion is small, and the prolongation rate is between 0.2% and 1.5%, there is no significant difference. Second, syllable features are significant to vowel duration. The syllable length is inversely proportional to the vowel duration, and the duration shortening of short vowels (40%) is larger than that of long vowels (16%). Similar studies showed that [16], the vowel duration gradually decreases with the gradual increase of phonemes in syllables. Third, the influence of adjacent consonants on vowel duration is complicated. Although, vowel duration is shortened by unvoiced consonants it is prolonged by voiced consonants, which proved to be a cross-language phonetic feature [17].

Duration is the distinctive feature in the acoustic production of long-short vowel contrast in Mongolian. Figure 2 shows the duration distribution of 7 long-short vowel contrast according to their categories, where black represents short vowels and gray represents long vowels.

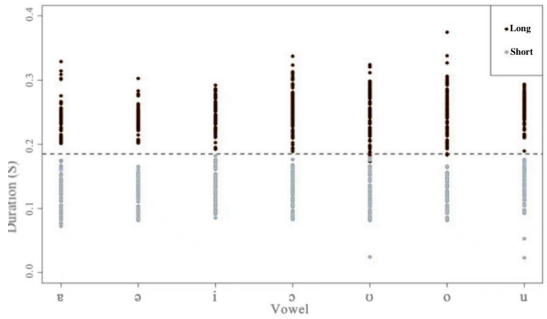
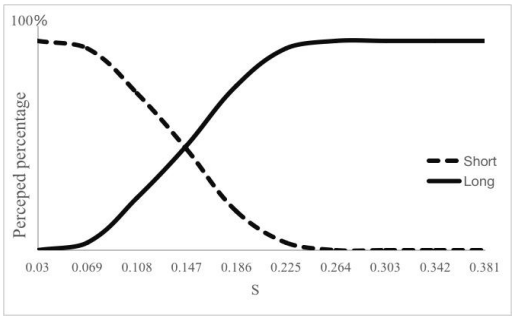


Figure 2: Duration contributions of 7 long and short vowel contrasts.

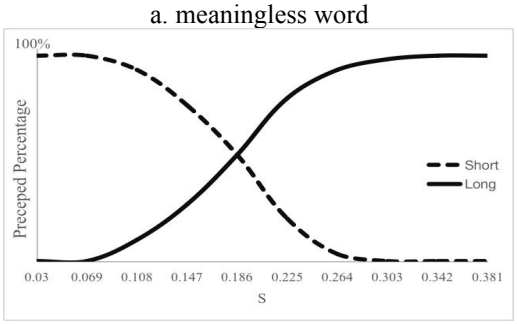
The physical duration of long and short vowels as a whole shows a remarkable boundary at 0.18s as indicated by black dotted line in the Figure. 95% of vowel duration is distributed around this boundary. Short vowels are shorter than 0.18s and long vowels are longer than 0.18s. As far as all the data are concerned, the standard deviation of duration among seven short vowels is 0.038s, and that among seven long vowels is 0.033s, and the difference between short and long vowels is 0.069s, which is much larger than the internal standard deviation of short vowels or long vowels. That is, on the duration dimension, the distribution between short and long vowels is scattered, while the internal distribution is concentrated. The duration difference between long and short vowels is statistically significant (P<0.01, df=2987, F=0.02).

3.2. Perceptual features

The identification results for continuum perception experiment are plotted in Figure 3. The total number of stimulus sounds in the identification task for meaningless words (Figure 3a) is 5,600, while that for meaningful words (Figure 3b) is 6,000.



a. meaningless word



b. meaningful words

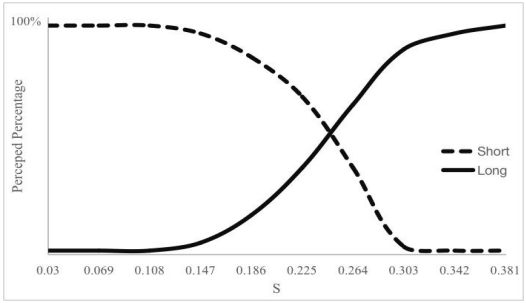
Figure 3: Perception results of long and short vowels.

Based on the binomial distribution of the identification scores and the sigmoid shape of the response function, the results show that:

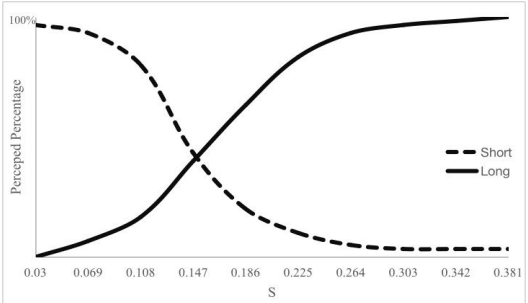
First of all, the perception boundary of long-short vowel contrast in Mongolian is 5/10 for meaningful words (see Figure 3b). When the duration of long vowels is shortened by half, it has a very obvious influence on vowel category. Most listeners can not identify them as a long vowel when the duration is shorter than half. Thus, duration is the primary phonetic feature for the classification of long-short vowel contrast in Mongolian at phonological level. Listeners basically identify long or short vowels by vowel duration.

Second, the perception space of long vowels is broader than that of short vowels. The experimental results show that the auditory space of long vowels covers all the duration segments after stimulus 2/10, while the space of short vowels ranges from stimulus 1/10 to 7/10. For meaningless words, the space of short vowels ranges from 1/10 to 6/10, and that of long vowels spans from 2/10 to 10/10.

The production and perception experiments of long-short vowel contrast prove that duration is the most important acoustic feature between long vowels and short vowels, but it is not the only one. The experimental results presented in Table 1 and Figure 2 fully illustrate that word meaning is a critical factor. Based on this finding, all experimental words, including 30 meaningful words and 28 meaningless words, are divided into long vowel words with meaning, whose short counterparts had no meaning, and short vowel words with meaning, whose long counterparts had no meaning, so as to re-examine the actual influence of word meaning on vowel type and duration pattern. For example, the /mus/ (ice) is meaningful, but the /mu:s/ is meaningless; the /mɐm/ is meaningless, but the /mɐ:m/ (Lama) is meaningful. The identification results in Figure 4 illustrates there is a complementary relationship between word meaning and vowel duration in decoding or perceiving vowel category. When experimental words are meaningless, listeners perceive long vowels or short vowels depending on vowel duration. However, when experimental words are meaningful, listeners first identify the vowel category based on the phonological and phonetic knowledge of the mother tongue. This is proved by the fact that meaningful short vowel words move the perceptual boundary in Figure 4a (1/10-6.5/10) more rightward than that of the meaningless words in Fig. 4b (1/10- 4/10); meaningful long vowels move the boundary more leftward in Figure 4b (4/10-10/10) than that of the meaningless words in Figure 4a (6.5/10-10/10) .



a. short (meaningful)-long (meaningless)



b. short (meaningless) - long (meaningful)

Figure 4: *P*erceptual results of vowel types in meaningful meaningless pairs.

The influence of word meaning on vowel duration is polarized. On the one hand, word meaning will shorten the acoustic duration of vowels. When the speech data are classified by syllable features, the influence of vowel duration in meaningless words is greater than that in meaningful words. As you can see in Table 1, the influence rate of long vowels is 15% (6.5% for meaningful words) and that of short vowels is 16% (11% for meaningful words). On the other hand, word meaning expands the perceptional space of short vowels and moves the perceptual boundary rightward as mentioned above which was shown in Figure 3. The identification curve of long and short vowels of meaningful words is less sharpen than that of meaningless words. Therefore, word meaning is an important factor besides duration in the identification of vowel category. For meaningful words, listeners identify the vowel comprehensively by duration and word meaning; or, they identify only by duration for meaningless words.

According to the duration data of this study, duration pattern of long-short vowel contrast in Mongolian is 2:1. that is, long vowels are twice as long as short vowels. Acoustically, although vowels differ in the proportion of duration, a comprehensive survey shows that long vowels are twice as long as short vowels. In term of perception boundary, the dividing point of duration between long-short vowel contrast is at 1/2, but it moves slightly when word meaning is involved. When vowel duration is less than or greater than 1/2 of the long vowel, the vowel type changes in phonology.

4. CONCLUSION

This study analyzed the acoustic production and perception category of Mongolian long-short vowel contrast. It was preliminarily found that duration is the most significant distinguishing feature between long vowels and short vowels in acoustic production. In addition, it also found the complementary mechanism between word meaning and vowel duration in the process of perception and identification of long and short vowels. When experimental words are meaningless, listeners perceive long vowels or short vowels depending on vowel duration. However, when experimental words are meaningful, listeners first identify the vowel category based on the phonological and phonetic knowledge of the mother tongue.

5. ACKNOWLEDGMENTS

This work is supported by national social science fund "The Construction of Mongolian Speech Database and the In-depth Study of Intonation

Related Issues" (22XYY003), Key projects of National Social Science Fund "Quantitative Syntactic Research Based on the Mongolian Annotated Corpus' (19AYY018) and, Project of Cultural Experts and "Four Batches" of Talents awarded to Aijun LI .

6. REFERENCES

[1] An acoustic study of the Japanese short and long vowel distinction. Master's thesis, Lawrence: University of Kansas, 2008.

[2] Lyberg. B, "Some observations on segment duration and its interaction with the fundamental frequency contour in Swedish," J. Acoust. Soc. Am, 77, S54, 1985.

[3] Kozasa.T, An acoustic and perceptual investigation of long vowels in Japanese and Phonpeian. PhD dissertation, Honolulu: University of Hawai'I at Manoa, 2005.

[4] Myers.S, "Vowel duration and neutralization of vowel length contrasts in Kinyarwanda," Journal of Phonetics, vol. 33: 2005, pp. 427-446.

[5] Svantesson. J. O. , Tsendina. A. , & Karlsson. A, Phonology of Mongolian, Oxford: Oxford university press, 2013, pp. 183-185.

[6] Huhe, Qoijungjab, Acoustic Analysis on the Mongolian Phonemes, Hohhot: Inner Mongolia University Press, 1999, pp. 131–153.

[7] Chinggeltei. Mongolian Grammer, Hohhot: Inner Mongolia People's Publishing House. 1991, pp. 11-21.

[8] Batdvrj. Acoustic Analysis of Khalha Dialect in Mongolian. Ph.D dissertation, Hohhot: Inner Mongolia university, 2014.

[9] Huhe, "Long and Short Vowels in Mongolian: A Multivariate Analysis of Variance Approach," Minority Languages of China.vol. 5, 2021, pp. 101-107.

[10]Yurong, Analysis of Vowel Duration based on the Mongolian Spoken Corpus, Journal of Inner Mongolia University, vol, 6, 2012, pp. 1-9.

[11]Aomin, "The Prosodic value of Mongolian non-Syllabic vowel," Applied Linguistics, vol. 1, 2018, pp. 108-115.

[12]F.Shi, Auditory Patterns: A Preliminary Study on the Characteristics of Chinese Speech Perception, Beijing: Commercial Press, 2019.

[13]Ziyu Xiong, "xSegmenter: Segment automatic segmentation and annotation tools," Chinese journal of phonetics, vol. 1, 2009, pp. 27-34.

[14] Xu Y, ProsodyPro - A tool for large-scale systematic prosody analysis, Tools and Resources for the Analysis of Speech Prosody, 2013.

[15]LI Bin. REN Rui-chao, "Mathematic Analysis of English Monophthongs' Duration and its Modeling," Mathematics in Practice and Theory, vol. 46, 2016, pp. 158-166.

[16]Xueguang, Aomin, "Statistical analysis of Mongolian standard phonetic vowel duration," Mongolian Language, vol. 7, 2019, pp. 47-51.

[17]Mcallister.R. , Flege.J.E. , & Piske.T, "The Acquisition of Swedish Long vs. Short Vowel Contrasts by Native Speakers of English, Spanish and Estonian," Applied Psycholinguistics, vol. 22, 2001, pp. 283-299. https://scripts.sil.org/cms/scripts/page.php?item_id=IPAhome.

DOES TALKER VARIABILITY HELP ADULTS LEARN NOVEL WORDS?

Sandy Abu El Adas[1], Ivy Yen[2], Susannah V. Levi[2]

[1]Basque Center on Cognition, Brain and Language, [2]The Department of Communicative Sciences and Disorders, New York University
[1]s.abueladas@bcbl.eu

ABSTRACT

Exposure to multiple talkers has been shown to be beneficial for adults learning speech sounds in a second language and for children learning novel words in a lab-based word learning task. The current study examines the extent to which variability is beneficial for adults in a word learning task and how individual differences in reading and language skills affect performance. Seventy-six participants were randomly assigned to either the multiple-talker (four talkers) or single-talker condition and learned novel object-nonword pairings. To assess word learning, listeners completed three naming tasks throughout the study (after 4 exposures, after 8 exposures, and after a brief delay) in which they had to generate the name of the object. Overall, listeners with higher language skills showed better word learning than those with poorer language skills. No effects of reading ability or talker variability were found. These results suggest that variability may only be beneficial in some tasks.

Keywords: speech processing, talker processing, phonetic variability, individual differences.

1. INTRODUCTION

In the domain of spoken language processing, different areas of research have examined the benefit of variability on learning. The most widely studied area is learning non-native sound contrasts. Numerous studies have shown that when adults learn a novel non-native sound contrast, exposure to multiple talkers improves both perception and production [1–5]. For example, Japanese speakers trained on the English /r/-/l/ contrast in the multiple talker condition performed better than those in the single talker condition, demonstrating the importance of variability on the development of phonemic categories [5]. The explanation for this benefit is that talker variability allows listeners to determine which variability is tied to the phonemic contrast and which is not [6], and that this type of perceptual warping helps reduce the focus on irrelevant acoustic information [5].

Not all studies, however, found a benefit of multiple talkers for learning non-native sound contrasts (e.g., [7]). The lack of consistency across studies may be explained by differences in the stimuli used, whether listeners are tested on new items (generalization), amount of exposure during training, and age of the participants (infants, children, adults).

While many studies have examined the benefits of variability on L2 sound learning, fewer studies have examined its effects on word learning. Richtsmeier et al. [8] investigated the effect of talker variability (single talker vs. 10 talkers) on word learning in pre-school children and found that the multiple talker condition resulted in more accurate and faster naming. A similar study of children with and without Developmental Language Disorder (DLD) also found a benefit of talker variability [9]. Together, these two studies suggest that the exposure to multiple talkers facilitates word learning in children, presumably by helping the learner create abstract phonological representations of the novel words.

In addition to talker variability, phonological factors such as phonological complexity and phonotactic probability have been shown to impact word learning [10–11]. For example, studies show that adults perform better on a word learning task when words are shorter and less phonologically complex [10]. Phonotactic probability also plays a role on word learning. However, there are inconsistencies across children and adults. While children show a high phonotactic probability advantage [11], adults show a high phonotactic probability disadvantage [12].

Individual factors such as language and reading ability also impact word learning. Studies show that individuals with poor reading or language skills perform differently [13–15]. However, studies have examined these skills separately. It is important to examine both reading and language skills simultaneously (within the same individuals) because there is an overlap in the underlying skill differences in phonological processing that then impact word learning.

While most research on the effects of talker variability in word learning focused on children that have not yet reached phonological maturity, little is known on how adults cope with talker variability when learning new words and whether this varies based on individual differences in the domains of reading and language ability. This is particularly

important in the context of studies showing inconsistent results of how people process and use talker variability and how age plays a role on whether variability facilitates learning [16].

In the current study, we address this gap by examining how talker variability and individual differences in reading and language skills affect word learning in adults. In addition, we investigate how phonological information (phonotactic probability) affects word learning given the literature showing a relationship between phonological processing skills and word learning.

2. METHODS

2.1. Participants

Seventy-six participants ages 18-40 (M=30.8, SD=6.47) completed the study. All listeners were native speakers of U.S. English with no history of speech, language, or hearing disorders. All participants passed a web-based headphone check [17]. Participants were recruited through Prolific (www.prolific.co) and paid $20.

2.2. Stimuli

Sixteen CVC nonwords were selected from Storkel and Hoover's database [18]. Half had high phonotactic probability (M=0.0096, SD=0.0031, range=0.0075-0.0172) and half had low phonotactic probability (M=0.0032, SD=0.0001, range=0.0019-0.0048). Phonotactic probability was determined by the biphone frequency corresponding to the adult corpus. Each nonword was paired with an object from the Novel Object and Unusual Name (NOUN) Database [19].

Four female native speakers of U.S. English ages 19-24 from the New York City area produced the nonwords. Recordings took place in a sound-attenuated booth with a SHURE 10A head-mounted microphone and a Fostex FR-2LE recorder using a 44.1 kHZ sampling rate. All items were amplitude normalized for the word learning task.

2.2. Procedure

Participants were randomly assigned to one of two conditions: (a) high talker variability (n=29) or (b) low talker variability training (n=47). Those in the high variability condition heard the nonwords produced by all four talkers, and those in the low variability training were randomly assigned to one of the four talkers. Participants also completed additional tasks described below. The experiment took approximately one hour to complete and was hosted on Gorilla experiment builder [20].

2.2.1. Word learning task

During the training phase, participants heard the nonword-object pair, repeated the nonword aloud, and then answered a yes/no question about the semantic features of the object (e.g., "Is the object black and white?") to check engagement (Figure 1).

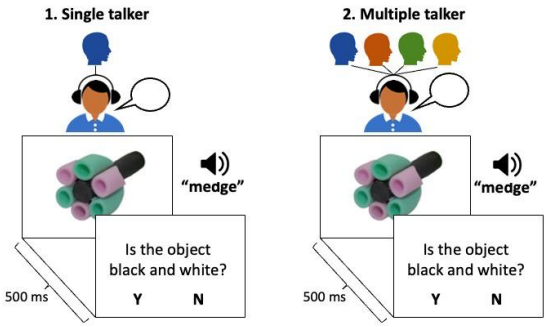


Figure 1: Procedure for **(1)** single-talker and **(2)** multiple-talker conditions in the word learning task.

During *Training 1* participants heard the 16 nonwords repeated 4 times in random order (16*4=64 trials), followed by a naming task (*Naming 1*), where they were presented with the object and asked to say the name out loud. They then completed a second identical training block (*Training 2*) and naming task (*Naming 2*). Following a brief delay (approximately 10 minutes) in which they completed filler tasks, participants then completed a final naming task (*Naming 3*).

2.2.2. Individual difference measures

Language ability was assessed using a modified version of the Recalling Sentences subtest of the CELF-4 [21]. Participants heard pre-recorded sentences and were instructed to repeat them. All scoring procedures from the CELF-4 were followed. Reading ability was assessed with the Rapid Online Assessment of Reading (ROAR), a quick and automatic online assessment previously tested with both adults and children [22-23]. In this task, participants complete a lexical decision task with orthographic real words and nonwords.

One limitation is that the score distribution on the ROAR was positively skewed with more participants with higher scores (see Table 1). This is expected given that our participants were typically developing adults.

Assessment	Mean (SD)	Range
Recalling Sentences	84.17 (10.91)	44-96
ROAR	78.76 (4.26)	58-84

Table 1: Mean, standard deviation (SD), and range for the language and reading measures.

3. RESULTS

A logistic mixed-effect model was fit to the naming accuracy data (coded as 0,1) using the lme4 package [24] in RStudio [25-26]. The fixed effects included naming (1, 2, 3), condition (multiple talkers, single talker), language ability, reading ability, phonotactic probability (high, low), a 3-way interaction between naming, condition, and phonotactic probability along with their 2-way interactions, and also 2-way interactions for condition by reading ability and condition by language ability. Raw scores (scaled and centered) were used for our individual measures (Recalling Sentences, ROAR) to allow for more variability in the data. The model also included random intercepts for participant and nonword (item). All categorical fixed effects (phonotactic probability, task, condition) were sum-coded.

The analysis revealed a significant effect of naming and the emmeans analysis [27] showed that participants performed significantly better on Naming 2 and 3 compared to Naming 1, and that there was a significant decline in performance between Naming 2 and Naming 3 (Table 2 and Figure 2).

Contrast	*Estimate*	*SE*	*z-ratio*	*p**
1 vs. 2	-2.17	0.119	-18.17	**<.001**
1 vs. 3	-1.93	0.118	-16.29	**<.001**
2 vs. 3	0.24	0.099	2.43	**0.040**

Table 2: emmeans analysis of naming. *p* values are adjusted using Holm's method and indicated as *p**

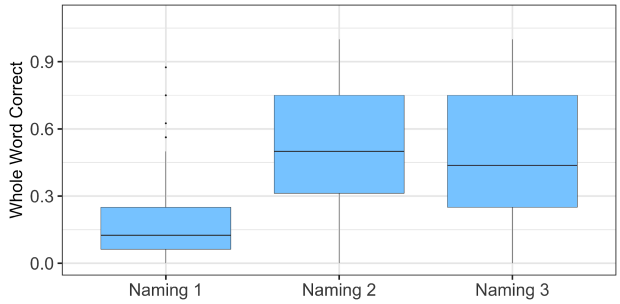


Figure 2: Naming by proportion correct.

In addition, there was a significant effect of language ability (β=0.55, *SE*=0.16, *p*<0.001), with better naming for participants with higher language scores (see Figure 3).

In addition, the analysis revealed a significant interaction between condition and phonotactic probability (β=-0.10, *SE*=0.04, *p*=0.0239). No pairwise comparisons from the emmeans analysis for condition within each of the levels of phonotactic probability reached significance, but the interaction suggests that the difference for condition was greater in the high than the low phonotactic probability items (see Figure 4).

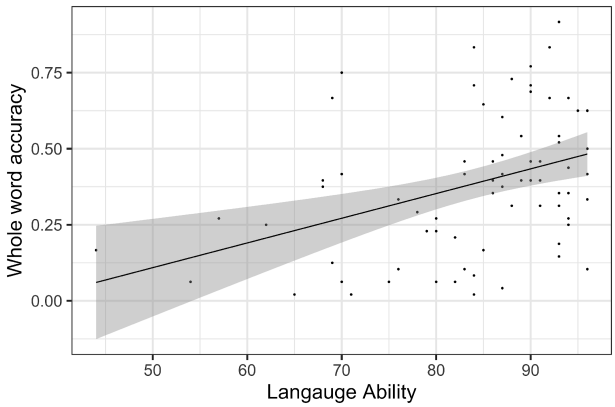


Figure 3: Relationship between language ability (raw scores on Recalling Sentences) and whole nonword accuracy.

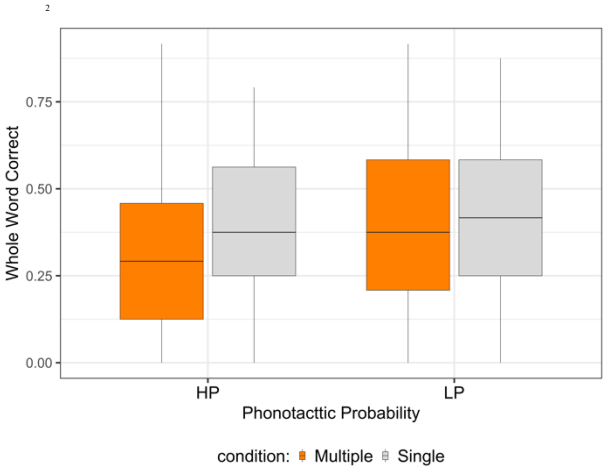


Figure 4: Proportion correct by condition (multiple=orange, single=grey) and phonotactic probability (HP=high, LP=low).

Unexpectedly, there was no effect of talker variability (β=0.05, *SE*=0.16, *p*=0.75) or reading ability (β=0.24, *SE*=0.17, *p*=0.15). None of the other effects or interactions reached significance.

4. DISCUSSION

In the current study, we examined the extent to which talker variability modulates performance on a word learning task with adults, and whether individual differences in reading or language ability influence performance. Consistent with previous studies, our results revealed an effect of exposure on word learning, where participants showed gains in learning over time and performed better after two exposure blocks compared to after only one. However, these gains were not fully maintained after a brief delay. This decrease might be explained by the fact that the participants were in the early stages of learning,

where integration of the newly learned words have not been fully integrated in the lexicon.

Consistent with previous studies on word learning (e.g., [14]), we found that participants with higher language skills performed better on the naming tasks. Of note is that this study with children examined learning of low frequency words rather than nonwords. Here, we show that language ability plays an important role on word learning even after carefully controlling the stimuli.

Contrary to our hypothesis and previous studies (e.g., [13]), we found no effect of reading ability on word learning. There are several possible explanations for the difference between our study and previous studies. First, the type of task we used to assess reading ability may have been less sensitive to reveal differences in reading ability among adults who – as far as we know – had typical reading development. Here, we used the ROAR, which required participants to read the items and decide if the items are real words, not a decoding task. Second, it is possible that the lack of reading ability effect on word learning is related to statistical power and the overall variance of the ROAR scores. In fact, most of our participants scored high on the ROAR, which may have reduced the effects of reading on word learning.

Also surprisingly, talker variability did not facilitate learning, and participants in the single- and multiple-talker conditions performed similarly on the word learning task. These results go against the word learning studies with children that found a benefit for talker variability. One possible reason for this discrepancy may relate to the number of talkers used during training. The two studies with children used ten talkers in the multiple talker condition, while the current study used only four. Apart from the quantity of talkers, the variability of the talkers themselves may also play a role. While both our study and the previous studies used only female speakers, it may be possible that talker variability gains depend on how much acoustic differences exist between talkers. Indeed, in the current study, all female speakers were from the same age range and had a similar dialect, reducing the amount of acoustic variability between speakers. More research is needed to understand the effects of quantity and quality of talker variability on word learning.

Another possible reason for the lack of benefit of talker variability may relate to the amount of exposure during training. It is possible that our listeners did not have enough exposure to the talkers to show a talker variability benefit. This idea is in line with the effect of naming block on word learning where participants did not maintain their learning after a short delay. This is also consistent with a previous study showing that at the early stages of learning, variability can slow processing and reduce the differences between multiple and single talker conditions, but after sufficient exposure, variability is beneficial [28].

Finally, it is possible we did not find any talker variability effects due to age. The two previous studies that found a talker effect on word learning [8-9] tested children, whereas our study tested adults. Studies show age influence on talker processing where listeners become more sensitive to talkers and talker differences in their native language compared to talkers in an unfamiliar language [e.g., 29].

Most of the research on talker variability in word learning has focused on children that have not yet reached an adult-like phonological maturity. Here, we add to a body of literature on talker variability in word learning by testing adults on a set of phonologically legal nonwords. Because our stimuli contained nonwords that are all phototactically legal, it is possible that we did not find differences between single- and multiple-talker conditions because our participants have matured phonological capabilities that washed-out the differences between the groups. The interaction between condition and phonotactic probability suggests that individuals show talker variability benefits depending on how phonologically familiar nonwords are. While post-hoc testing did not find any significant differences between the two groups, there was a greater difference between the multiple and single talker conditions on nonword with high phonotactic probability. This may suggest that when the newly learned words are more word-like, talker variability negatively impacts learning, especially in the early stages of learning. These results are consistent with previous studies on word learning in adults showing better performances on items from low phonotactic probability compared to high phonotactic probability [12].

5. CONCLUSION

Our study found no benefit of talker variability in adults in a word learning paradigm, although this might have been driven by differences in design and stimuli. In addition, language ability, but not reading ability, influenced learning. Participants with higher language ability performed better on the naming tasks (regardless of training condition). We also found that phonological familiarity impacts whether talker variability is helpful during a word learning task, with better gains in words that are more phonologically dissimilar to other words in the language.

6. REFERENCES

[1] Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura Y.I. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

[2] Kartushina, K., & Martin, C.D. (2019). Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training. *Language Learning*, *69*(1), 71–105.

[3] Lambacher, S. G., Martens, W. L., Kakehi, K. , Marasinghe, C. A. , & Molholt, G.. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics,* 26(*2*), 227–247.

[4] Lively, S.E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English/r/and/l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America,* 96(4), 2076–2087.

[5] Logan, J. S., Lively, S. E. & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America,* 89(*2*), 874–886.

[6] Holt, L. L. & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America,* 119(5), 3059–3071.

[7] Brosseau-Lapré, F., Rvachew, S., Clayards, M. & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, 34(*3*), 419–441.

[8] Richtsmeier, P. T., Gerken, L., Goffman, L. & Hogan, T.P. (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, 111(3), pp. 372–377.

[9] Plante, E., Bahl, M. , Vance, R. & Gerken, L. (2011). Beyond phonotactic frequency: Presentation frequency effects word productions in specific language impairment. *Journal of Communication Disorders,* 44(*1*), 91–102.

[10] Papagno, C. & Vallar, G. (1992). Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. *The Quarterly Journal of Experimental Psychology Section A*, 44(*1*), 47–67.

[11] Storkel, H. L. & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics*, 14(*6*), 407–425.

[12] Storkel, H. L., Armbrüster, J. & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech Language and Hearing Research*.

[13] Alt, M., Hogan, T., Green, S., Gray, S., Cabbage, K. & Cowan, N. (2017). Word learning deficits in children with dyslexia. *Journal of Speech Language Hearing and Research*, 60(*4*),1012–1028.

[14] Windfuhr, K. L. & Snowling, M. J. (2001). The Relationship between Paired Associate Learning and Phonological Skills in Normally Developing Readers. *Journal of Experimental Child Psychology*, 80(2), 160–173.

[15] McGregor,K., Gordon, K., Eden, N., Arbisi-Kelm, T. & Oleson, J. (2017). Encoding deficits impede word learning and memory in adults with developmental language disorders. *Journal of Speech Language and Hearing Research*, 60(*10*), 2891–2905.

[16] Quam, C., & Creel, S. C. (2021). Impacts of acoustic-phonetic variability on perceptual development for spoken language: A review. *Wiley Interdisciplincary Reviews: Cognitive Sciences*, 12(*5*), e1558.

[17] Woods, K. J., Siegel, M. H., Traer, J. & McDermott J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention Perception & Psychophysics.*, 79(*7*), 2064–2072.

[18] Storkel, H. L. & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English. *Behavioural Research. Methods*, 42(*2*), 497–506.

[19] Horst, J. S. & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavioural Research Methods*, 48(*4*), 1393–1409.

[20] Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N. & Evershed J. (2018). Gorillas in our midst: Gorilla. sc. *Behavioural Research Methods*.

[21] Semel, E., Wiig, E. H. &. Secord, W. A. (2003). Clinical evaluation of language fundamentals (CELF-4). *San Antonio TX*: Psychological Corporation.

[22] Yeatman J. D., et al. (2021). Rapid online assessment of reading ability. *Scientific Reports*, 11(*1*),1–11.

[23] Amrita, B., Annaliese, B., Tau, N., Pablo, R. & Saloni, K. (2020). The role of intrinsic reward in adolescent word learning.

[24] Bates, D., Maechler, K., Bolker, B. & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014*.

[25] Team, R. C. (2017). *R: A language and environment for statistical computing.*

[26] RStudio Team. (2016). *RStudio Team. RStudio: Integrated Development for R.*

[27] Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means (Version R Package Version 1.3.4).

[28] Lee D. & Baese-Berk, M. (2021). Non-native English listeners' adaptation to native English speakers. *JASA Express Letters.*, 1(*10*), 105201.

[29] Levi, S. V., & Schwartz, R. G. (2013). The Development of Language-Specific and Language-Independent Talker Processing. *Journal of Speech, Language, and Hearing Research*, *56*(3), 913–925.

SOUND CHANGE IN THE LAB: EXPOSURE-INDUCED VOCALIC CHAIN SHIFTS

Václav Jonáš Podlipský[1] and Paul Olejarczuk[2]

[1]Palacký University Olomouc, Czechia, [2]California State University, Chico, USA
vaclav.j.podlipsky@upol.cz, polejarczuk@csuchico.edu

ABSTRACT

Previous research on perceptual adaptation in vowels has attested generalization of retuning across vowel-height counterparts. Our study tested for a lab-induced chain shift. In a lexical decision task, we exposed U.S. English participants to words with either raised or lowered tokens of /i/, along with unmanipulated back vowels but no other front vowels. In subsequent categorization of the members of a [i]~[ɪ]~[ɛ] continuum, we found a difference between the raised- and lowered-/i/ participants in the location of the /i/-/ɪ/ boundary (replicating perceptual adaptation in vowels) as well as the /ɪ/-/ɛ/ boundary. Thus, the exposure-induced shift of /i/-/ɪ/ triggered a similar shift for the unexposed /ɪ/-/ɛ/ contrast, pointing to adaptation at the level of the vocalic system rather than the level of perceptual cues or low-level auditory compensation.

Keywords: perceptual recalibration, generalization of adaptation, chain shifts, English vowels

1. INTRODUCTION

Languages tend to converge on vocalic and consonantal systems with optimal (articulatory and/or perceptual) distance between categories. Theoretical accounts of this phenomenon include the Dispersion Theory [1, 2], Dispersion-Focalization Theory [3], Self-Organizing Systems Theory [4], or evolutionary phonological accounts [5, 6]. One diachronic process purportedly motivated by the pressure for distance optimization is the *chain shift*, where an imbalance in perceptual/articulatory space created by an initial shift of one sound category is rectified by subsequent, cascading adjustments of other sounds throughout the system. Amongst the well-known examples of historical chain shifts, where the contrasts between categories may have remained while the members of these contrasts changed, are the Great Vowel Shift affecting Middle English, the Cockney Vowel Shift, Grimm's Law, and the Romance intervocalic lenition. Many diachronic chain shifts are also now underway, including Californian /ɪ/, /ɛ/, /æ/ lowering [7, 8].

Explanations like Dispersion Theory, which appeal to the notion of perceptual space, in essence constitute hypotheses about individuals (since perception is an aspect of human cognition). It therefore follows that, if perceptual space prefers optimal distance between elements, we should be able to observe this preference in a controlled laboratory setting. Our study tests this prediction using the paradigm of immediate perceptual adaptation to non-canonical speech-sound realizations. Listeners routinely face the challenge of adapting to high phonetic variability in natural speech input. Past research has demonstrated that listeners retune their speech sound categories when exposed to shifted sound tokens in the lab. In a seminal paper, Norris et al. [9] showed that a sound halfway between [f] and [s] was perceived as /f/ or as /s/ depending on lexical context in Dutch, and that this category adaptation persisted in subsequent perception. Apart from lexical cues, adaptation has been shown to also be guided by visual cues [10], or even simply by the distribution of phonetic properties in the input [11, 12]. Most previous research focused on the recalibration of Cs (see e.g. the review in [13]) and studies on Vs have, until recently, been scarce [11, 12, 14, 15, 16].

Interestingly, perceptual adaptation can generalize to other categories within the system. For instance, Kraljic & Samuel [17] revealed that exposure to shifted VOT in English [d]-[t] led to shifted categorization of /d/-/t/ and also /b/-/p/. Chládková et al. [11] found that exposure to shifted Greek [i]-[e] resulted in adjusted categorization of /i/-/e/ and also /u/-/o/ (Greek has a symmetrical 5-vowel system). This indicates that category recalibration is distinct from the lower-level auditory compensation attested e.g. by Ladefoged & Broadbent's study [18] showing that the same [bVt] word was more likely perceived as /bɛt/ than /bɪt/ when it followed a synthetic carrier phrase with a lowered F1 making the F1 of [bVt] seem relatively higher, i.e. more appropriate for /ɛ/ than /ɪ/. This was replicated with natural speech [19], but also reversed speech [20, 21] and spectrally rotated speech analogs [22] as precursors.

Aiming at category recalibration rather than auditory compensation, this study (using a lexical decision task) exposed U.S. English listeners to words with the full high-to-low range of back Vs, but no front Vs except for /i/, whose tokens were lowered for one group of listeners and raised for another. Then, a

V categorization task tested for a potential difference between the groups in the location of the boundary between /i/ and /ɪ/ as well as the boundary between /ɪ/ and /ɛ/ (both latter absent from exposure). We asked whether a retuning of /i/~/ɪ/ would trigger a chain-shift reaction, namely adapted perception of /ɪ/~/ɛ/.

2. METHODOLOGY

2.1. Stimuli

All stimuli were based on sound-booth recordings of a male U.S. English speaker from Portland, OR. The initial phase, a lexical decision (LD) task, consisted of 208 exposure trials. These included 3× repetitions of 30 target items (English /i/-words, /i/ between obstruents, no other front Vs, none forming minimal pairs with /ɪ/-words, e.g. *cheese*, *fever*, *seafood*), 1×30 English word fillers (no front Vs, e.g. *assault*, *carbon*, *snuggle*), and 2×44 phonotactically plausible non-word fillers (no front Vs, dissimilar to words with front Vs, e.g. [ˈkupəɹə], [dəˈtɑɪ], [pʌlˈpun]).

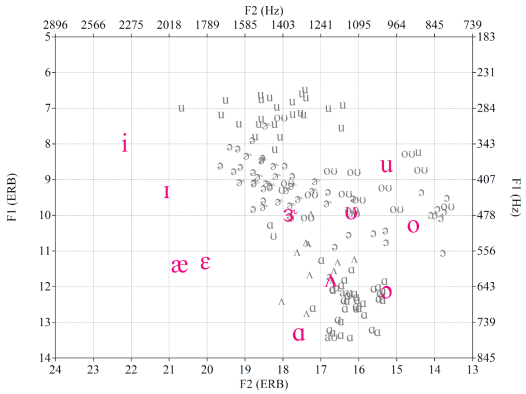


Figure 1: An F1 by F2 plot of all the Vs in the exposure words apart from /i/. U.S. English male means from [23] given for reference (in larger font).

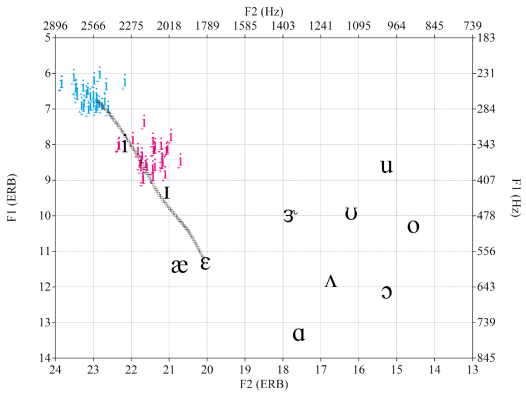


Figure 2: An F1 by F2 plot of all the /i/s in the exposure words, either lowered [i̝] (magenta) or raised [i̞] (cyan) artificially. The line of '+' symbols represents the 85-member [i]~[ɪ]~[ɛ] continuum used in the categorization task. U.S. English male means from [23] given for reference (in larger font).

Fig. 1 displays the F1 and F2 values of all the Vs contained in the exposure stimuli except for /i/, covering the full high-to-low range. Two resynthesized copies of each /i/-word were prepared, one with lowered [i̝] and the other with raised [i̞]. The resynthesis involved shifting the F1, F2, and F3 in Praat [24], using the method described in [11]. Table 1 gives the mean formant values of the original [i]s (not used in exposure) and the magnitude of each shift. The lowering was such that the resulting [i̝]s were between our speaker's prototypical /i/ and /ɪ/ but still more peripheral than /ɪ/ (see Fig. 2), and the raising ensured that the more extreme [i̞]s were without artifacts and still sounded natural. The shifted [i] tokens are shown in Fig. 2.

values in ERB	F1	F2	F3
Original [i] means	7.06	22.42	24.08
Lowering shift	+1.15	-1.11	-0.91
Raising shift	-0.575	+0.557	+0.455

Table 1: Mean F1, F2 and F3 frequencies of the [i] tokens before manipulation and the shift size for each direction and formant.

The stimuli for the post-exposure categorization task were 85 unique, isolated vowels forming a finely sampled [i]~[ɪ]~[ɛ] continuum, produced by resynthesizing (the same method as /i/ shifting) a single token of our speaker's naturally-produced [ɛ] (265 ms in duration). The qualities ranged between the [ɛ] endpoint with F1 = 11.22, F2 = 20.08 and F3 = 23.65 ERB, and the [i] extreme with F1 = 6.79, F2 = 22.9, and F3 = 25.46 ERB, with near-perfect interpolation for the steps in between. The test continuum is shown in Fig. 2.

2.2. Procedure

The whole experiment was implemented as a Praat Demo window script [24] and run on a computer with a headset in a quiet environment. Participants first completed the LD task, i.e. the exposure phase. On each of the 208 trials, they heard one of the LD stimuli (in random order) and decided whether it was an existing English word by clicking on either a button labeled 'word' or one labeled 'not a word' (button position random). They were offered a short break after every 52 trials.

After the exposure task and a short silent break, they completed a 3-alternative forced-choice categorization test. On each of the 85 trials, they heard a stimulus (random order) and identified the vowel by clicking on one of 3 buttons marked '[i] *beef*, *see*', '[ɪ] *kiss*, *kid*', and '[ɛ] *fresh*, *egg*' (button order random). They took a short break halfway through.

2.3. Participants

The participants were 65 native U.S. English speakers (51 female, 14 male), aged 19 to 44 years (mean 23), all college students who participated in exchange for course credit. They were randomly assigned to either the lowered-/i/ (n = 33) or the raised-/i/ exposure condition (n = 32). About ⅔ in each condition were Californians, most others came from Nebraska, 5 from other (mid)western states, and 1 from Texas. None reported any hearing impairments.

3. RESULTS

All (anonymous) data and the analysis script in R [25-28] as well as the experiment Praat script and all stimuli are available at https://osf.io/czu5g/. We first inspected responses on the exposure task. Overall, the lexical decisions were correct 91% of the time. Words with lowered [i̞] were identified as English words 93% of the time, words with raised [i̝] 95% of the time, indicating that the manipulated Vs were mostly indeed perceived as /i/. No participants were thus excluded due to frequent misidentification of the manipulated words as non-words.

Inspection of the V categorization data revealed that the numbers of responses were not balanced across categories, with most listeners giving fewer /ɪ/ than /i/ or /ɛ/ responses. Four listeners with 5 or fewer /ɪ/ responses (out of 85 per listener) were excluded from subsequent analyses, resulting in the numbers of participants in the lowered and raised conditions of 29 and 32, respectively. Fig. 3 is a tile plot showing all responses for the included participants. Fig. 4 shows the proportion of each response category for each stimulus, pooling data across participants. The figures suggest a somewhat higher proportion of /i/ responses in the lowered than in the raised condition, and in turn a higher proportion of /ɛ/ in the raised than in the lowered condition.

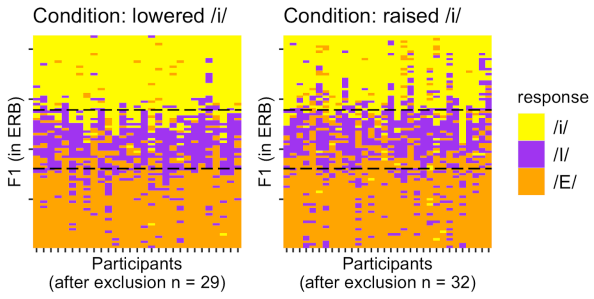


Figure 3: Tile plots displaying all categorization responses to the 85 stimuli (shown by their F1) in the two conditions. The dashed lines highlight the points of equal distance between category means, used for centering the F1 as predictor (see text). '/I/' = /ɪ/, '/E/' = /ɛ/.

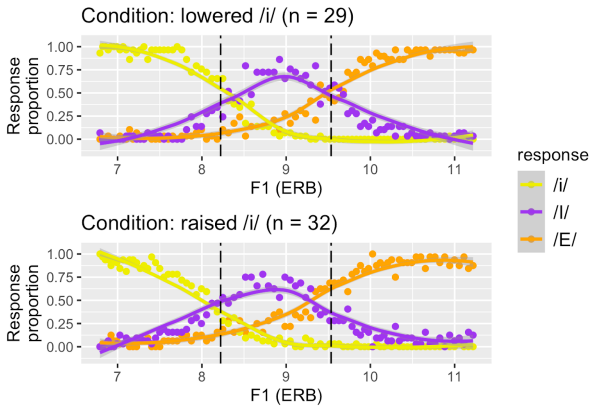


Figure 4: Scatterplots showing the proportion of each response category for each continuum stimulus (ranked by F1) across participants, with loess smoothing, split by condition. The dashed lines: same as Fig. 3.

The data were fitted to two mixed-effects logistic regression models, separately for the /i/-/ɪ/ contrast (excluding /ɛ/ responses from the data) and the /ɪ/-/ɛ/ contrast (excluding /i/ responses). Each model had Vowel Spectrum (centered) and Condition (sum-coded: lowered /i/ 1, raised /i/ -1) as the fixed effects and Participant as the random effect, with by-participant varying intercepts and slopes for Vowel Spectrum. Like in [11], F1 (in ERB) was used as a proxy of stimulus vowel spectrum (F2 and F3 steps being almost perfectly correlated with F1 steps across the continuum). Since the numbers of responses across categories were unbalanced (see above), the spectrum predictor was centered, differently for each contrast, to the points of equal distance between category means along the F1: first we computed mean F1 (in ERB) for responses elicited for each category and then the means of adjacent categories' means, shown as the dashed lines in Fig. 3 and 4, which served as the respective predictor centers.

Tables 2 and 3 give the models' coefficient estimates and Fig. 5 plots the fitted values for each model. The predicted probability of an /i/ response against /ɪ/ was 0.56 for the raised and 0.81 for the lowered condition (logit difference: +0.492, SE = 0.23, z = 2.14, p = 0.033), whereas the predicted probability of an /ɛ/ response against /ɪ/ was 0.74 for the raised and 0.67 for the lowered condition (logit difference: -0.206, SE = 0.093, z = -2.22, p = 0.026). The top panels of Fig. 5 also show the predicted category boundaries, i.e. points along the F1 continuum where $P(response)$ = 0.5 for each Condition. Expressed as category boundary shifts, the effect of Condition was 0.18 ERB for the the /i/-/ɪ/ contrast and 0.26 ERB for the /ɪ/-/ɛ/ contrast. To visualize individual variability, the bottom panels of Fig. 5 plot by-participant logistic functions derived from the random effect coefficients of each model.

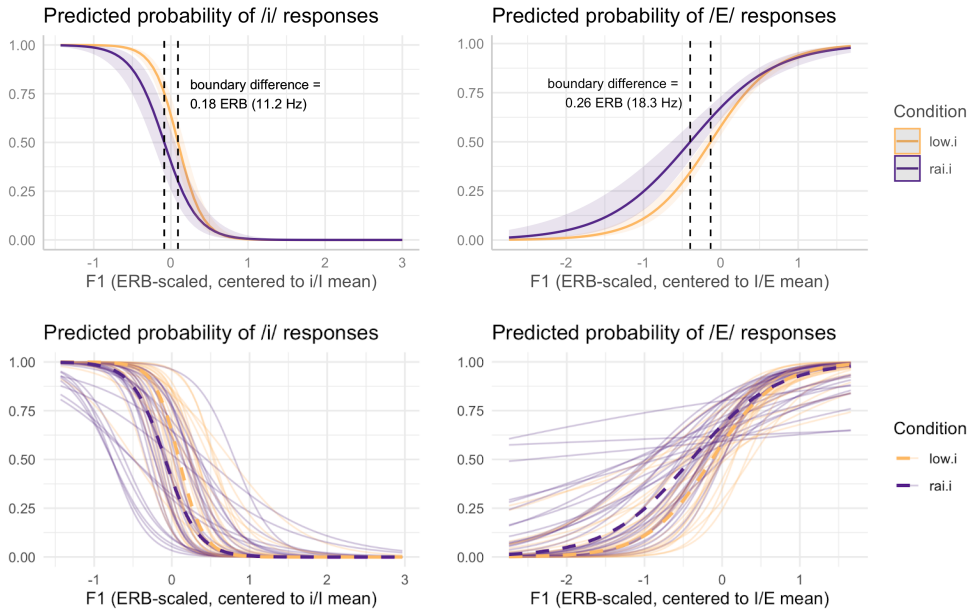


Figure 5: Values fitted by the logistic regression models for the /i/-/ɪ/ (left) and the /ɪ/-/ɛ/ contrast (right) split by Condition. Top panels show the overall logistic functions (repeated as thick dashed curves in the bottom panels) with dashed, vertical lines highlighting the corresponding category boundaries where *P(response)* = 0.5. Bottom panels show by-participant functions derived from the random effect coefficients.

	Estimate	SE	z	Pr(>|z|)
(Intercept)	0.088	0.235	0.375	0.707
F1_iɪ	-5.505	0.415	-13.260	**< 0.001**
Condition1	0.492	0.230	2.138	**0.033**
F1_iɪ : Condition1	-0.793	0.351	-2.255	**0.024**

Table 2: Coefficients estimated by the model of the probability of /i/ against /ɪ/ responses.

	Estimate	SE	z	Pr(>|z|)
(Intercept)	0.525	0.094	5.597	**< 0.001**
F1_ɪɛ	2.130	0.171	12.431	**< 0.001**
Condition1	-0.206	0.093	-2.221	**0.026**
F1_ɪɛ : Condition1	0.285	0.166	1.720	0.085

Table 3: Coefficients estimated by the model of the probability of /ɛ/ against /ɪ/ responses.

4. DISCUSSION

We exposed U.S. English listeners to words with either raised or lowered tokens of /i/, along with unmanipulated back vowels but no other front vowels. In the categorization test that followed, we found a difference between the raised- and lowered-/i/ listeners in the probability of the /i/ vs. /ɪ/ response and in the location of the /i/-/ɪ/ boundary in the expected direction, thus replicating the effect of perceptual recalibration for vowels documented previously [11, 12, 16].

At the same time, categorization differed in the expected direction for the /ɪ/-/ɛ/ contrast as well, although no tokens of these vowels had been presented in the exposure phase (unlike in [14, 15]), showing a generalization of the perceptual adjustment from /i/-/ɪ/ to an adjacent contrast. Previous studies found generalization of perceptual recalibration for vowels across vowel-height counterparts in Greek [11] and, using the shorter-term selective adaptation paradigm, also across vowel-frontness counterparts in German [29]. Our findings are indicative of chain-shift category retuning which has not been attested conclusively so far (cf. [14]).

Importantly, our results cannot be ascribed to speaker-specific formant cue calibration or auditory compensation as a general sensory mechanism of adjusting to the acoustic shifts of the stimuli (see [30] for these accounts) since all participants heard the same back Vs spanning the full high-to-low range, ensuring the two exposure conditions differed only in the realization of /i/ and not in the overall F1 span of the exposure stimuli.

Therefore, we interpret our findings as evidence of adaptation at the level of the vocalic system, induced in a controlled laboratory setting and potentially driven by a preference to maintain optimal distance between categories, whereby shifting one category triggered the shift of a neighboring category (though see [5, 31] for alternative explanations of the mechanisms underlying chain shifts). Assuming a link between perceptual processing and phonological categories, accumulated perceptual recalibration may thus serve as a mechanism of longitudinal chain shifts occurring in individual language users (e.g. [32]) and, by extension, in entire language communities.

4. ACKNOWLEDGEMENTS

This research was supported the Czech Science Foundation grant number 21-09797S.

5. REFERENCES

[1] Lindblom, B. (1986). Phonetic universals in vowel systems. In: Ohala, J., Jaeger, J. (eds.) *Experimental Phonology*. Academic Press, 13–44.

[2] Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus*. PhD diss., UCLA.

[3] Schwartz, J. L., Boë, L. J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *J. Phonetics*, 25(3), 255–286.

[4] de Boer, Bart. 2001. *The origins of vowel systems*. OUP.

[5] Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology*, 25(2), 217–270.

[6] Vaux, B., & Samuels, B. (2015). Explaining vowel systems: dispersion theory vs natural selection. *Linguist. Rev.*, 32(3), 573–599.

[7] Kennedy, R., & Grama, J. (2012). Chain shifting and centralization in California vowels: An acoustic analysis. *Am. Speech*, 87(1), 39–56.

[8] D'Onofrio, A., Pratt, T., & Van Hofwegen, J. (2019). Compression in the California vowel shift: Tracking generational sound change in California's Central Valley. *Lang. Var. Change*, 31(2), 193–217.

[9] Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychol.*, 47(2), 204–238.

[10] Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *J. Phonetics*, 45, 91–105.

[11] Chládková, K., Podlipský, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *J. Exp. Psychol. Human*, 43(2), 414–427.

[12] Colby, S., Clayards, M., & Baum, S. (2018). The role of lexical status and individual differences for perceptual learning in younger and older adults. *J. Speech Lang. Hear. R.*, 61(8), 1855–1874.

[13] Nelson, S., & Durvasula, K. (2021). Lexically-guided perceptual learning does generalize to new phonetic contexts. *J. Phonetics*, 84, 101019.

[14] McQueen, J., & Mitterer, H. (2005). Lexically-driven perceptual adjustments of vowel categories. *Proc. of ISCA Workshop on Plasticity in Speech Perception* London, 233–236.

[15] Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Sci.*, 32(3), 543–562.

[16] Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *J. Phonetics*, 81, 100984.

[17] Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychon. B. Rev.*, 13(2), 262–268.

[18] Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.*, 29(1), 98–104.

[19] Ladefoged, P. (1989). A note on "Information conveyed by vowels." *J. Acoust. Soc. Am.*, 85(5), 2223–2224.

[20] Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral–envelope distortion. *J. Acoust. Soc. Am.*, 90(6), 2942–2955.

[21] Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *J. Acoust. Soc. Am.*, 96(3), 1263–1282.

[22] Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Atten. Percept. Psycho.*, 73(4), 1195–1215.

[23] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97(5), 3099–3111.

[24] Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.20, retrieved 24/09/2022 from http://www.praat.org/.

[25] R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL https://www.R-project.org/.

[26] Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67, 1–48.

[27] Singmann, H., et al. (2022). Afex: Analysis of Factorial Experiments. R package version 1.1-1. https://CRAN.R-project.org/package=afex.

[28] Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

[29] Llompart, M., & Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in German vowels. *J. Phonetics*, 67, 34–48.

[30] Chodroff, E., & Wilson, C. (2020). Acoustic–phonetic and auditory mechanisms of adaptation in the perception of sibilant fricatives. *Atten. Percept. Psycho.*, 82(4), 2027–2048.

[31] Ettlinger, M. (2007). An exemplar-based model of chain shifts. *Proc. 16th ICPhS* Saarbrücken, 685–688.

[32] Harrington, J., Palethorpe, S., & Watson, C. (2000). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *J. Int. Phon. Assoc.*, 30(1-2), 63–78.

INVESTIGATING NATIVE CHINESE SPEAKERS' IDENTIFICATION OF ENGLISH CONSONANTS IN NOISE

Kai Sheng[1,*], Jian Gong[1,✉], Yuhong Sun[1], Weizhong Zhang[1], Weijing Zhou[2,†], Feng Wang[3]

1. Phonetics Lab, Jiangsu University of Science and Technology, China
2. School of Foreign Languages, Yangzhou University, China
3. Deep Blue Honour School, Jiangsu University of Science and Technology, China

[*]ricky_shk@163.com, [✉]j.gong@just.edu.cn, [†]zwjzj513@163.com

ABSTRACT

The present study investigated the perception of English consonants by native Chinese speakers in quiet and three noise conditions. The results demonstrated that the perception accuracy for Chinese speakers in quiet was significantly higher than that in noise. Evidences of "language-independency" in acoustic/auditory processing was found by analysing the individual consonant identification in the four test conditions. A weak correlation was found between Chinese speakers' performance in quiet and their deterioration in noise, suggesting that Chinese speakers' English consonant phonetic category learning can be fragile, especially in adverse listening conditions.

Keywords: L2 consonant perception, perception in noise, language independent

1. INTRODUCTION

Background noise can greatly affect listeners' speech perception. Although bilinguals and L2 learners with high proficiency could achieve the same performance as native speakers in quiet condition, their performances still decrease significantly in noisy conditions [1, 2]. The noise produces masking effects on the target speech signal, which causes the reduction of the acoustic information that can be obtained from the speech [3]. The masking effect can be divided into energetic masking and informational masking. Previous studies have reported that native speakers and non-native speakers showed higher English consonant identification scores in informational masking noise than in pure energetic masking noise with the same signal to noise ratio, however, the strongest masking effect were induced by the combination of informational and energetic masking [4, 5, 6].

The study of consonant perception in noise by listeners from European countries has been carried out for a long time, and English is usually the target language. For example, [5] investigated the effect of masker type on Spanish listeners' perceptions of English consonants; [7] examined Dutch listeners' identification of English consonants, and [8] studied the L1 influence on Norwegian listeners' English intervocalic consonants perception. More recently, [6, 9] have compared the English consonant perception in noise between native English speakers and Chinese EFL learners, to explore the native advantage and the transmitted information changes in different noise conditions. However, in [6, 9], the Chinese EFL learners were students studying in England, who all had considerable second language experience. It is suggested that people with more target language experience perform better in speech perception [10]. Therefore, in order to have a complete picture of non-native perception of English consonants in noise, it is worth studying those listeners with less experience to English, e.g., native Chinese speakers living in China.

In a large study of English consonant perception in noise, the full set of English consonant identification scores from speakers of eight European languages were compared. A significant degree of similarity on several factors such as the effect of noise type, consonant and phonetic feature was found across all language groups, indicating an "language-independent" processing in acoustic and auditory considerations for the speakers from different European languages [4]. Therefore, the purpose of the current study was to extend the English consonant identification in noise experiment in [4] to a non-European language. More specifically, native Chinese speakers' identification of English consonant in quiet and three noise conditions (Speech shaped noise, Competing Speaker noise and eight-talker babble noise) would be compared to examine the evidence of "language-independency". Possible L1 influences would also be investigated.

2. METHOD

2.1. Listeners

A group of 36 Chinese listeners, including 9 males and 27 females, participated in the current study. These listeners were students from a Chinese university, ranging in age from 19 to 27 years (M=22.7). No listener had reported hearing or language problems, and all the listeners had passed a hearing test with pure-tone thresholds $\leq$ 20 dB HL at octave intervals between 1000 and 8000Hz [11]. Most listeners were from Northern Mandarin dialect spoken region. All of them had a certification in level II grade A or above in the National Proficiency Test of Putonghua (Mandarin). These listeners were studying various majors in university, and all of them had passed the College English Test Band 6 (CET-6). Listeners were paid for their participation.

2.2. Stimuli

The English consonant stimuli used in the current study were nonsense vowel-consonant-vowel (VCV) tokens derived from the Interspeech 2008 Consonant Challenge corpus [11]. The vowel contexts for each VCV tokens in this corpus were the 9 combinations of the 3 vowels /æ, i, u/ in initial and final positions. The test sets in the corpus were produced by 4 male and 4 female speakers, containing 24 British English consonants (/p, b, t, d, k, ɡ, tʃ, ʤ, f, v, θ, ð, s, z, ʃ, ʒ, h, m, n, ŋ, l, r, j, w/[12]). In each test condition, 16 VCV tokens were used for each of the 24 consonants, making 384 VCV tokens altogether. The vowel contexts were balanced for each consonant. Another 10 VCVs were used as practice items at the beginning of the test in quiet condition.

2.3. Noise Maskers

Listeners identified VCVs in quiet and three different additive noise backgrounds, i.e. speech-shaped noise (SSN), competing speaker (CS) and 8-talker babble (8BB). The three noise backgrounds were selected on the basis that they permit the roles of informational and energetic masking to be examined [4, 11]. More specifically, speech-shaped noise is a pure energetic masker with a fixed spectrum and no significant temporal modulations; a competing speaker contains significant modulations in both frequency and time and produces both energetic and informational masking since audible components of the masker can compete with those of the target; the 8-talker babble can be seen as located in the middle of a SSN to CS continuum.

2.4. Procedure

Listeners completed the perception test individually at a sound-treated laboratory. A customized MATLAB program was used to present speech stimuli and collect the responses. Participants were required to identify the speech stimuli that presented through headphones, by clicking the corresponding button on a 4×6 on-screen button grid. Real English words with capital letters to indicate the corresponding consonant were shown on the buttons. All listeners completed the test in quiet first, and then the three tests with different noise maskers in a random order. The signal to noise ratio (SNR) for CS and SSN conditions were -6dB, while the SNR for 8BB was set at -2dB [11]. These SNR values were chosen to avoid floor and ceiling effects for listeners [5].

3. RESULTS

3.1. The overall results of consonant identification

Figure 1 shows the overall consonant identification rates by native Chinese listeners in four test conditions. It is clear listeners performed best in quiet, with a mean accuracy of 79.9%. As for the noise conditions, listeners' performance in SSN (57.5%) was worse than in the other two maskers, while 8BB (60.4%) seemed to produce similar masking effect to CS (61.9%). One-way ANOVA confirmed a significant main effect of test condition ($p < 0.001$). Post-hoc pairwise comparison showed that there were significant differences between quiet and the three noise conditions respectively ($p < 0.001$). A significant difference was found between SSN and the other two noise conditions (CS, 8BB: $p < 0.05$). No significant difference was found between CS and 8BB ($p = 0.233$).

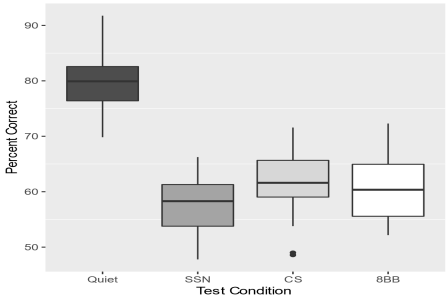


Figure 1: Identification rates in 4 test conditions.

3.2. Individual consonants

3.2.1. *Identification rate of individual consonants*

Figure 2 demonstrates listeners' 24 consonants' mean accuracies in four test conditions. It can be seen that, the rank of the size of masking effect for different types of noise varied for different consonants. Different from the overall result shown in Figure 1, listeners didn't always perform worst in SSN for individual consonants. For example, six consonants (/p, t, ʤ, θ, s, r/) demonstrated clear higher identification accuracy in SSN than in CS, while ten consonants (/b, k, tʃ, f, h, m, n, l, j, w/) showed clear higher accuracy in CS than in SSN. Meanwhile, consonants with the best and worst performance in different noise conditions also varied. For example, consonant /ɡ/ has the highest accuracy of 99.17% in quiet. /t/ has the highest accuracy of 88.39% in SSN. /l/ has the highest accuracy of 92.47% in CS. /ʃ/ has the highest accuracy of 85.03% in 8BB. In all test conditions, the accuracy of nasal /ŋ/ is the lowest, and the accuracy of /ð/ is the second lowest. A repeated-measures of ANOVA with two within-subjects factor (test condition and consonant) confirmed the main effect of test condition $[F(3,105) = 526.20, p < 0.001, \eta_p^2 = 0.938]$ and consonant $[F(23,805) = 119.35, p < 0.001, \eta_p^2 = 0.773]$ and a significant interaction between test condition and consonant $[F(69,2415) = 30.00, p < 0.001, \eta_p^2 = 0.462]$. Further simple effect analysis of test condition suggested that the accuracy of 24 consonants except for /ʃ/ has a significant difference among four test conditions. However, no significant difference was found for /ʃ/ in all test conditions $[F(3,35) = 2.11, p = 0.116 > 0.05, \eta_p^2 = 0.153]$.

3.2.2. *Identification degradation in noise conditions*

It can be observed from Figure 2 that, compared with the performance in quiet, /p/ has the biggest performance degradation in noise (CS in particular). Other thirteen consonants, /b, d, k, tʃ, f, θ, h, m, n, l, r, w/, also showed large identification accuracy drop in at least one noise condition. It is interesting to see that most of these sounds were among the best identified in quiet condition. For consonants that with relative low identification accuracies in quiet, such as /ð, ʒ, ŋ/, their performance degradations in noise were also relatively small. For consonants like /ʤ, s, z, ʃ/, although they were not within the best identified sounds in quiet, however, they were quite able to resist the effect of noise.

3.3. Relation between identification in quiet and noise

Statistical analyses were carried out to explore the relation between listeners' performance in quiet and their performance in noise. A clear correlation can be seen that the better their performance in quiet, the better their mean performance across three noise conditions ($r = 0.775, p < 0.001$). In fact, native Chinese listeners' performance in quiet was significantly correlated with the identification rates in each of the three masking conditions respectively ($p < 0.001$).

Correlation analyses were also carried out between listeners' performance in quiet and their performance degradation in three noise conditions. Although a significant weak correlation was found between listeners' performance in quiet and their mean performance degradation across the 3 noise conditions ($r = 0.349, p < 0.05$), detailed analyses revealed that a significant medium correlation between listeners' performance in quiet and performance degradation only existed in SSN ($r = 0.539, p < 0.01$) but not in CS ($r = 0.339, p > 0.05$) and 8BB ($r = 0.321, p > 0.05$).

In addition, the correlations of participants' performance between each noise condition were also explored. Findings showed that participants' performance between SSN and CS ($r = 0.790, p < 0.001$), between SSN and 8BB ($r = 0.757, p < 0.001$) and between CS and 8BB ($r = 0.795, p < 0.001$) were all significantly correlated, suggesting some possible universal strategies were applied by listeners in all 3 noise conditions.

4. DISCUSSION

The present study examined English consonant identification in quiet and in three noise conditions (SSN, CS and 8BB) by native Chinese listeners. The highest mean consonant identification score was shown in quiet and the lowest score in SSN, with CS and 8BB in between. This result is consistent with previous study on English consonant perception by speakers of eight European languages [4]. Cooke *et. al.* [4] reported that listeners' performance for SSN was always worse than the two modulated maskers (Speech Modulated Noise and Competing Speaker), which was independent of a listener's first language. Therefore, the results of the current study support the claim in [4] that the rank of masking effect for different noise type is language-independent. However, detailed analysis showed that listeners didn't always perform worst in SSN for individual consonants, indicating that different types of noise have different masking effect on individual sounds.

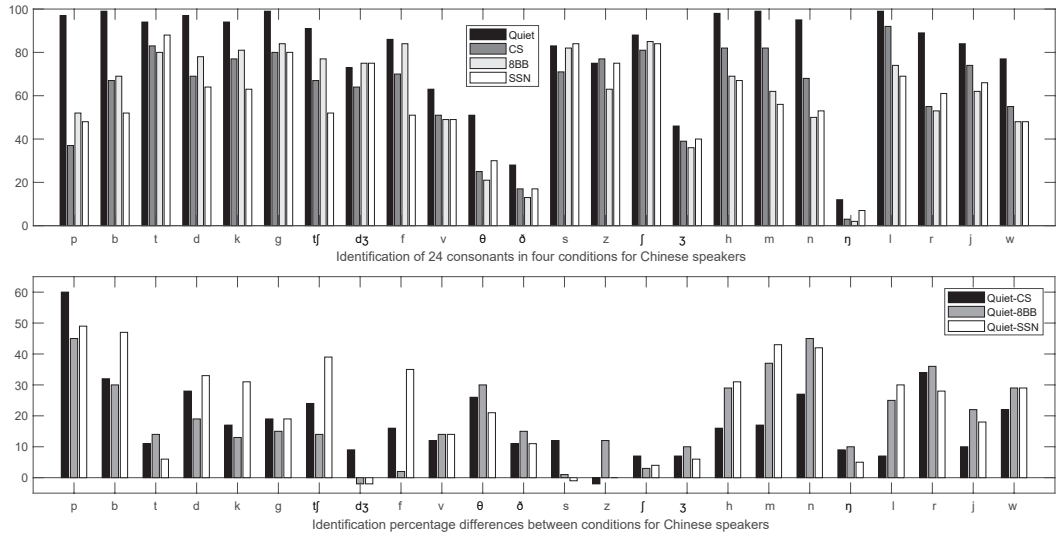


Figure 2: The identification rate of 24 consonants in four test conditions (upper). The identification percentage differences between conditions for Chinese speakers(lower).

From the results of identification in quiet condition, some clear L1 influences can be observed. Those English consonants with a similar phonetic counterpart in Mandarin Chinese led to a relative high identification accuracy (e.g., /p, b, t, d, k, g, tʃ, h, m, n/), whereas those English consonants without good phonetic counterparts in Mandarin Chinese resulted in a relative low accuracy (e.g, /θ, ð, ʒ/). This is consistent with the results reported in many previous studies of Chinese perception of English consonants [6, 13]. It was found that almost all consonants had a reduction on identification scores to some extent in three noise conditions. However, the identification accuracies of consonant /ʃ/ were almost the same in all test conditions, and consonant /s/ had a similar result to /ʃ/ except for in CS. The high identification rates for sibilant fricatives /s/ and /ʃ/ in SSN agree with previous study's finding [14]. Sibilants have much more intense noise components than other fricatives, and their high frequency energy allows them to escape some of the masking effect of SSN [4].

Cooke *et. al.* [4] compared the responses of English consonant perception in noise from speakers of eight European languages. Consonant /t/ was among the best identified sounds and /θ/, /ð/ were often found to be the worst identified for most of participant groups, indicating some possible language-independent process for certain sounds. Similar results were also seen in the current study, that /t/ had the highest mean identification accuracy and /θ, ð/ were among the most difficult sounds for Chinese listeners across three noise conditions, providing more evidence that this kind of language-independency may also exist in non-European language. Another noteworthy result is the extremely low identification accuracies for /ŋ/ in quiet and all three noise conditions. The main reason for this may due to the fact that some speakers in the corpse produced the /ŋ/ sound as a combination of a nasal plus a stop, and even the native English speakers confused /ŋ/ and /g/ in noise conditions [7]. However, both native English speakers and non-native Dutch speakers could achieve near perfect performance in quiet condition. Therefore, the extremely poor performance for Chinese speakers may also due to some language-specific reasons such as phontactic restriction for /ŋ/ in Chinese [15]. What's more, compared to the Dutch speakers whose English proficiency was relatively high, the lack of English experience for the Chinese speakers may also affect their perception judgment.

In the present study, it is noted that there was a significant correlation between performance in quiet and performance in noise, and a significant medium correlation between performance in quiet and degradation in SSN. These findings are consistent with previous study [5] about English consonant identification by native Spanish listeners, suggesting that Chinese listeners' English consonant phonetic category learning can be fragile, especially in some adverse listening conditions. A related and interesting result was shown in a training study [16] that for Chinese learners, the more their

performance of English consonant identification improved in SSN, the more their performance improved in quiet, suggesting that some robust perceptual cue in adverse conditions might not be easily learned in normal situation.

5. ACKNOWLEDGMENTS

This study was supported by grants from the Scientific and Technological Innovation Team of Jiangsu University of Science and Technology (2020), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX22_3729).

6. REFERENCES

[1] L. H. Mayo, M. Florentine, and S. Buus, "Age of second-language acquisition and perception of speech in noise," *Journal of speech, language, and hearing research*, vol. 40, no. 3, pp. 686–693, 1997.

[2] C. L. Rogers, J. J. Lister, D. M. Febo, J. M. Besing, and H. B. Abrams, "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," *Applied Psycholinguistics*, vol. 27, no. 3, pp. 465–485, 2006.

[3] H. Zhang and Z. Wang, "Usage of speech perception in noise in the selection and evaluation of hearing aids," *Journal of Clinical Otorhinolaryngology*, vol. 7, no. 2, pp. 69–72, 1993.

[4] M. Cooke, M. L. G. Lecumberri, O. Scharenborg, and W. A. Van Dommelen, "Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight european languages," *Speech Communication*, vol. 52, no. 11-12, pp. 954–967, 2010.

[5] M. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2445–2454, 2006.

[6] J. Gong, W. Zhou, and S. Zhang, "Effect of noise condition on the perception of l2 english consonants," *Language Education*, vol. 4, no. 2, pp. 44–52, 2016.

[7] M. Broersma and O. Scharenborg, "Native and non-native listeners' perception of english consonants in different types of noise," *Speech Communication*, vol. 52, no. 11-12, pp. 980–995, 2010.

[8] W. A. Van Dommelen and V. Hazan, "Perception of english consonants in noise by native and norwegian listeners," *Speech Communication*, vol. 52, no. 11-12, pp. 968–979, 2010.

[9] J. Gong, W. Zhou, and X. Ji, "Transmitted information analysis on distinctive features for chinese listeners' perception of english consonants in noise," *Journal of Jiangsu University of Science and Technology (Social Science Edition)*, vol. 15, no. 3, pp. 31–36, 2015.

[10] O.-S. Bohn and J. E. Flege, "Interlingual identification and the role of foreign language experience in l2 vowel perception," *Applied psycholinguistics*, vol. 11, no. 3, pp. 303–328, 1990.

[11] M. Cooke and O. Scharenborg, "The interspeech 2008 consonant challenge," 2008.

[12] P. Roach, "British english: received pronunciation," *Journal of the International Phonetic Association*, vol. 34, no. 2, pp. 239–245, 2004.

[13] J. Gong and W. Zhou, "Effect of experience on chinese assimilation and identification of english consonants." in *ICPhS*, 2015.

[14] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[15] Y. Lin, *The Sounds of Chinese*. Cambridge University Press, 2007.

[16] J. Gong, Y. Yu, W. Bellamy, F. Wang, and X. Ji, "Effect of perceptual training with noise on chinese learners' english consonant reception thresholds," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1087–1091.

INITIAL CURVATURE OF PITCH CONTOUR AFFECTS FINE-GRAINED LEXICAL TONE PERCEPTION OF MANDARIN BUT NOT CANTONESE

Yanyuan Ye, Yicheng Rong, Gang Peng

The Hong Kong Polytechnic University, Hong Kong
yanyuan.ye@connect.polyu.hk, rosy.rong@polyu.edu.hk, gang.peng@polyu.edu.hk

ABSTRACT

While the initial curvature (*Wāntóu* in Mandarin), a short curving portion at the onset of a pitch contour in tone languages, has been found inconsequential for distinguishing lexical tones, little is known about the role of this portion in fine-grained tone perception. To address this issue, the current study examined categorisation in 20 Mandarin and 19 Cantonese speakers using two tone continua (/i/ varying from Tone 1 to Tone 2): one continuum was synthesised based on the level tone with initial curvature; the other was based on the one without it. Participants were required to identify lexical tones. Results of Mandarin speakers revealed the initial curvature shifted the labelling boundary towards the rising tone, suggesting the initial curvature may affect fine-grained tone perception. However, this effect was absent in the Cantonese group, indicating the effect of initial curvature might be subject to the complexity of the tone system and different phonetic composition.

Keywords: initial curvature; fine-grained perception; lexical tones; Mandarin and Cantonese.

1. INTRODUCTION

Fundamental frequency (F0) has been widely believed to be the primary cue for tone perception, in which the pitch contour and the register mainly contribute to forming distinctive tones [1], [2]. Although citation tones seem to be stable and well-defined by their pitch contours compared to their sandhi forms, there would still be transitions at both ends of the contour. The transition at the beginning portion of the pitch contour has been described as a prolonged rising [3] and been referred as onset-curving [4] and pre-onset [5]. Given the use of the term 'onset' which usually refers to consonants is ambiguous, this portion will be named initial curvature in the current study. The shape of the initial curvature can vary depending on the surrounding phonetic environment.

Previous studies have been trying to define the domain of tone, only several of which focus on the effect of the initial curvature on tone perception. Howie [3] described nine types of F0's shape depending on different onsets for Mandarin lexical tones and concluded that the domain of tone was the syllabic vowel and the following voiced segment and that the initial curvature of the pitch contour was merely the anticipatory adjustment of the voice.

Lin divided the contour into three parts: the onset-curving section, the tone section, and the end-falling section [4]. He emphasised the essential function of the tone section in distinguishing four tones in Mandarin, which has been supported by his later empirical study [5]. Specifically, he examined the role of length of tone section in tone identification by cutting the sounds into small segments lasting for 80/100ms. The result showed that the longer the segment consisted of the tone section, the higher the identification accuracy would be, while the segments that mainly contained the onset-curving section or the end-falling section would yield non-targeted responses [5]. He thus considered these two sections insignificant for tone classification. Yang [6] conducted another set of perceptual experiments and figured out the perceptual centre of tone was located at the middle parts of the vowel, regardless of any type of onset. Since then, the initial portion of the pitch contour has been considered insignificant for lexical tone perception (see [7] for a review). However, those studies mainly focused on general tone distinguishment and only adopted simple word labelling tasks. Moreover, the materials that previous research used were fragmented or unnaturally synthesised, which might not reflect the authentic pattern of pitch categorisation. Thus, it remains unknown about the role of the initial curvature in fine-grained lexical tone perception.

Although non-tonal languages, such as English and German, do not use lexical tones to differentiate word meanings, changes in pitch may alter the expression of meaning at the sentence level. The transitional F0 movement towards the targeted accented syllable in these languages, i.e., the tonal onglide, has been considered to gain phonological status because it can be used as an acoustic cue to identify pragmatic meanings by German listeners [8], [9]. As the tonal onglide and initial curvature are both transitional adjustments observed at the beginning of pitch contour and thus share articulatory and phonetic similarities, the question that arises is whether the initial curvature would be the perceptual cue like

tonal onglide rather than being merely the phonetic variance.

The current study investigates the effect of initial curvature in fine-grained tone perception in Mandarin and Cantonese by conducting categorical perception experiments, a classical paradigm examining the fine-grained perceptual pattern of categorisation from a micro perspective [10], [11]. We focused on the perception of Tone 1 (high-level tone) and Tone 2 (high-rising tone), as these two tones are acoustically comparable in terms of the overall pitch contour in the two languages we examined [12].

2. METHODOLOGY

2.1. Participants

Twenty Mandarin native speakers (10 males, mean = 21.95 ± 2.98 yrs.) and 19 Cantonese speakers (11 males, mean = 21.88 ± 2.28 yrs.) were recruited. They were all right-handed. None of them majored in music, linguistics, or psychology, nor reported hearing or mental injuries.

2.2. Materials

The Mandarin Tone 1 stimulus *yi1* "clothes" and Tone 2 *yi2* "aunt", and the Cantonese Tone 1 stimulus *ji1* "doctor" and Tone 2 *ji2* "to lean on" were recorded by native female speakers, respectively. As shown in Figure 1d, the Cantonese *ji1* has a small rising at the beginning portion of its pitch contour, while this initial curvature is not observed in the Mandarin *yi1* (Figure 1a), although they share the same overall pitch contour. The four original stimuli were normalised to 300ms and 70dB before further manipulation.

The initial curvature of Cantonese *ji1* was straightened in PRAAT software [13] to form the *ji1* as a pure level tone, without any F0 changing at the beginning portion (see Figure 1e); the Mandarin *yi1* was manipulated with an additional initial rising like the original Cantonese *ji1*, giving rising to *yi1* with initial curvature (Figure 1b).

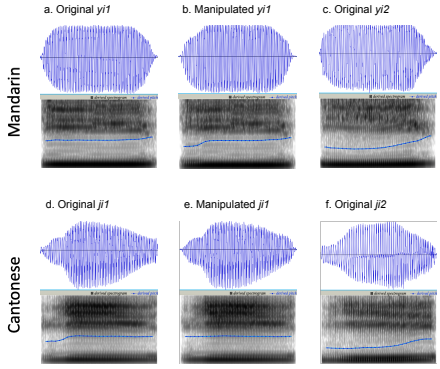


Figure 1: Mandarin tone tokens (the original *yi1* without initial curvature, the manipulated *yi1* with initial curvature, and *yi2*); Cantonese tone tokens (the original *ji1* with initial curvature, the manipulated *ji1* without curvature, and *ji2*)

The four Tone 1 stimuli, together with Cantonese *ji2* and Mandarin *yi2*, were synthesised into four seven-step continua varying from Tone 1 to Tone 2 by applying the pitch-synchronous overlap and add method [14]. As shown in Figure 2, the four continua can be divided into two sets: the Mandarin *yi* with and without initial curvature; the Cantonese *ji* with and without initial curvature. Each continuum has seven stimuli gradually varying from high-level to rising contour with an equal step size (stimulus 1 stands for the typical high-level tone, and stimulus 7 is the rising tone).

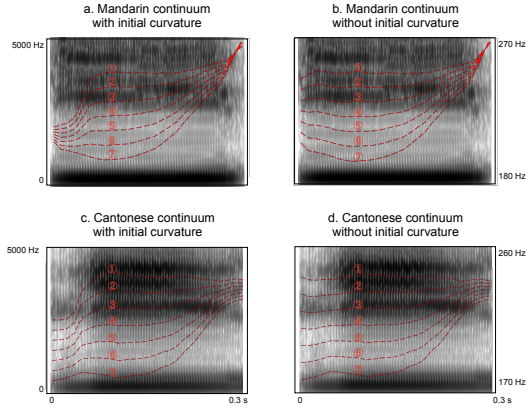


Figure 2: Two Mandarin and two Cantonese continua (the red curves indicate pitch contours along each continuum)

2.3. Procedures

The experiments were carried out remotely via Praat and monitored online via the VOOV platform, which enabled instructing participants as well as ensuring their full concentration during the whole process.

In each language group, there were two identification tasks (two continua: with and without initial curvature) and one rating task. The identification task, a two-alternative-forced choice task, is part of the categorical perception paradigm [10]. Participants were asked to identify the sounds they heard by clicking the mouse. Specifically, the Mandarin speakers need to identify whether they heard was the word "clothes" or "aunt", and the Cantonese participants needed to respond between the word "doctor" and "to lean on". In identification tasks, there were 2 continua * 7 stimuli * 9 repetitions = 126 trials. The order of each task, as well as the presentation of each stimulus, were counterbalanced.

After completing the identification tasks, participants from each language group would rate the naturalness of stimuli 1 and 7 of continua of their native language. Two nonspeech stimuli were taken into the rating task as fillers. On the basis of the templates with Tone 1 and Tone 2, nonspeech

counterparts were synthesized using 64 iteration steps [15]. The rating score ranged from 5 (most natural) to 1 (least natural). The expected rating result was that the naturalness of stimuli with and without initial curvature did not differ in the scores.

2.4. Scoring and data analysis

For each identification task, two key parameters were calculated – namely, the boundary position and the identification slope. First, for each stimulus in one continuum, an identification score, defined as the percentage of responses with which participants identified that stimulus as being Tone 1, was calculated. Next, boundary position, located where the identification score reaches 50%, and identification slope were assessed using a logistic regression equation described as [16]. To determine the influence of initial curvature on boundary position and identification slope, paired samples t-tests were conducted in Mandarin and Cantonese groups, respectively. For the rating task, the naturalness of stimuli with and without initial curvature would be compared in each group using the Wilcoxon signed-ranks test.

3. RESULTS

Figure 3 displays the identification curves for the Mandarin group. The result of paired-sample t-test showed the boundary position of the continuum *yi* with initial curvature (M = 3.70; SD = .36) was significantly larger than that without it (M = 3.46; SD = .31) in the Mandarin group; [$t(19) = -3.11$, $p = .006$, $d = .70$]. In terms of identification slope, there is no significant difference between the continuum *yi* with (M = -1.84; SD = .13) and without initial curvature (M = -1.83; SD = .06); [$t(19) = 0.14$, $p = .891$, $d = .03$].

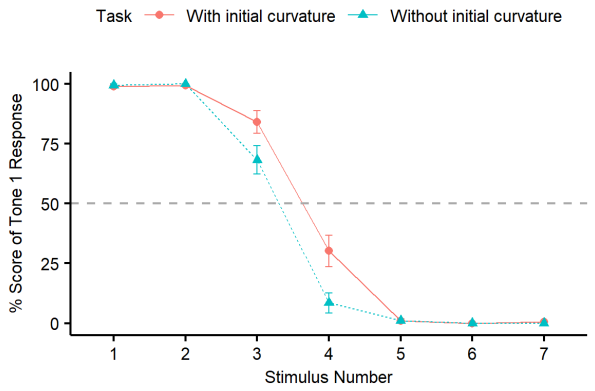


Figure 32: The identification curves of Tone 1 responses for identification tasks with and without initial curvature in Mandarin speakers. Error bars: ± 1 standard error.

As for the Cantonese group, the identification curves are shown in Figure 4. Statistical analyses revealed no significant differences in terms of either boundary position (*ji* with curvature: M = 3.95, SD = .49 vs *ji* without curvature: M = 3.86, SD = .41; $t(18) = 1.35$, $p = .194$, $d = .22$) or identification slope (*ji* with curvature: M = -1.86, SD = .14 vs *ji* without curvature: M = -1.88, SD = .11; $t(18) = .71$, $p = .488$, $d = .08$).

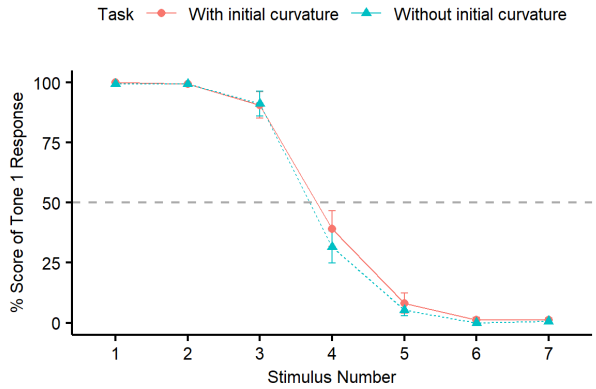


Figure 43: The identification curves of Tone 1 responses for identification tasks with and without initial curvature in Cantonese speakers. Error bars: ± 1 standard error.

Wilcoxon signed-ranks tests were conducted to determine the effect of initial curvature on perceptual naturalness. The results have shown no significant difference between different types of stimuli in either the Mandarin group (*yi* with curvature: M = 4.58, SD = .58 vs. *yi* without curvature: M = 4.61, SD = .43; Z = 102, $p = .472$) or Cantonese group (*ji* with curvature: M = 4.66, SD = .62 vs. *ji* without curvature: M = 4.68, SD = .55; Z = 9.5, $p = .833$), indicating that the naturalness of stimuli would not be affected by the existence or absence of the initial curvature statistically. Moreover, the relatively high mean rating score of each group (almost approaching the highest score of five) has ensured the quality of stimuli in our study.

4. DISCUSSIONS

4.1 The initial curvature contributes to fine-grained lexical tone perception

When Mandarin speakers identified words from the continuum whose base of Tone 1 has additional initial curvature, the identification boundary position would shift towards the rising tone compared to the case when the base had no initial rising. In other words, for Mandarin speakers, if there was initial curvature in the base, there would be more stimuli labelled as the level tone; if the base had no initial rising, the stimuli would be identified more as the rising tone. The initial curvature in Mandarin tones would yield more level-tone responses. However, this effect was not observed in the Cantonese group. There was no significant difference between the two continua in either the identification boundary position or the slope.

Regarding perceptual naturalness, neither the Mandarin nor Cantonese group showed significantly different rating scores between the continua with and without initial rising. The existence or absence of initial curvature would not influence the naturalness of the synthesised stimuli.

Previous studies have argued from the acoustic perspective that the tone domain bears on the syllabic vowel and possible voiced segment following it and that the F0 of the initial voiced consonant or non-syllabic vowel was merely anticipatory movements [3]. Some studies conducted perceptual experiments and concluded that the initial curvature should be excluded from the domain of tone [5], [6]. The results of the present study have suggested that the initial curvature of the pitch contour would affect fine-grained lexical tone perception in some languages. This seemly inconsistency with previous research might result from different designs of experiments and materials: the tasks in previous research were to label the tone class directly. Their primary purpose was to determine which portion contributes to tone distinguishment. The general contour was, without doubt, believed to be foremost in lexical tone perception. However, the issue was far from fully addressed. Having determined the primary perceptual cue did not guarantee the complete irrelevance of other cues. This is what our research cares about. In the present study, we focused on the effect of the short initial portion. Participants were asked to identify between two tones, and the stimuli were from a continuum varying between two tone categories. The stimuli used in the current research were synthesised from the base that contains the entire citation tone and of high naturalness and quality, while those from previous studies were cut and fragmental. To summarise, while the initial curvature has been found inconsequential for distinguishing lexical tones, this portion would affect fine-grained lexical tone perception.

4.2 The effect of initial curvature is subject to the tone system

The effect of initial curvature upon fine-grained lexical tone perception was observed in Mandarin but absent in Cantonese. The phenomenon that needs to be illustrated first is the shifting effect of the initial curvature on the perception boundary of the syllabic vowel *yi* in Mandarin. As introduced, the initial curvature of manipulated *yi* was created by adding a short rising contour at the beginning portion of the vowel [i]. Therefore, we can narrow this question and discuss why changing a small initial portion of F0 would change the boundary position. The shifted boundary position observed among Mandarin speakers, as shown in Figure 3, reflects that the participants were led to yield more level-tone responses, indicating listeners' perception of the pitch direction became less sensitive, and their judgement criteria of the level pitch became looser. One of the possible reasons is that the initial prolonged rising of the F0 contour has reduced the perceptual sensitivity for the rise in the overall pitch pattern. This could also be explained from the perspective of the relationship between acoustics and perception. The presence of the initial curvature gentled the overall slope of the tonal continuum, which in turn affects the perception.

However, this effect observed among Mandarin speakers seems weak in Cantonese speakers. One possible explanation is that the segment bearing the initial F0 curvature of Cantonese is less salient. Different from Mandarin, the phonetic composition of Cantonese stimuli starts with a more fricative onset [j] [17] (see Figure 1 d-e). Another possible reason for the absence of the effect of initial curvature in Cantonese listeners is that the Cantonese tone system is more complex compared to the Mandarin one [12]. Specifically, in Mandarin, Tone 1 and Tone 2 are the only level tone and rising tone, respectively. As more "psychoacoustic space" has been provided, Mandarin listeners would show leniency towards the acoustic changes in the level-tone and rising-tone categories. On the other hand, for Cantonese listeners, the category of Tone 2 is compressed due to the existence of Tone 5, the other rising tone in the Cantonese tone system. Likewise, the other two level tones in Cantonese (Tone 3 and Tone 6) will press the category of Tone 1 into a smaller psychoacoustic space, making this Tone 1 category less flexible and changeable relative to that of Mandarin. In short, the different complexity of the tone system might alter the likelihood of adjusting the perceptual categories of tones: the more complex and denser the system, the less it is likely to be affected.

5. CONCLUSIONS

The current study investigated the influence of initial curvature of pitch contour on the fine-grained lexical tone perception among speakers of Mandarin and Cantonese. Our findings have revealed that the presence of the initial curvature would yield more level tone responses in Mandarin listeners by reducing their sensitivity to pitch direction. This effect was not observed in Cantonese, indicating that the effect of initial curvature might be subject to the complexity of the tone system and different phonetic composition.

6. ACKNOWLEDGEMENTS

This research was partly supported by a fellowship award from the Research Grants Council of the Hong Kong SAR, China (Project No. PolyU/RFS2122-5H01).

7. REFERENCES

[1] Y. R. Chao, 'Mandarin primer', in *Mandarin Primer*, Harvard University Press, 2013.
[2] Y. Xu, 'Contextual tonal variations in Mandarin', *J. Phon.*, vol. 25, no. 1, pp. 61–83, Jan. 1997, doi: 10.1006/jpho.1996.0034.
[3] J. M. Howie, 'On the domain of tone in Mandarin', *Phonetica*, vol. 30, no. 3, pp. 129–148, 1974.
[4] M.-C. Lin, 'The pitch indicator and the pitch characteristics of tones in standard Chinese', *China ACTA Acust.*, vol. 1, pp. 8–15, 1965.
[5] M.-C. Lin, 'A perceptual study on the domain of tones in Beijing Mandarin', *China ACTA Acust.*, vol. 20, no. 6, pp. 437–445, 1995.
[6] S. Yang, 'A preliminary study on the perceptual center of tones in Standard Chineseb', *Acta Psychol. Sin.*, vol. 3, pp. 247–253, 1992.
[7] Y. Song, 'Hanyu Shengdiao Shiyan Yanjiu Huiwang [A review of Mandarin tone studies]', *Yuwen Yanjiu*, no. 1, pp. 41–45, 2006.
[8] D. Crystal, 'Prosodic Systems and Intonation in English. Cambridge Studies in Linguistics 1.', 1969.
[9] S. Ritter and M. Grice, 'The Role of Tonal Onglides in German Nuclear Pitch Accents', *Lang. Speech*, vol. 58, no. 1, pp. 114–128, Mar. 2015, doi: 10.1177/0023830914565688.
[10] B. H. Repp, 'Categorical Perception: Issues, Methods, Findings', in *Speech and Language*, vol. 10, Elsevier, 1984, pp. 243–335. doi: 10.1016/B978-0-12-608610-2.50012-1.
[11] McMurray B., 'The Myth of Categorical Perception'. PsyArXiv, Feb. 21, 2022. doi: 10.31234/osf.io/dq7ej.
[12] G. Peng, 'Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese', *J. Chin. Linguist.*, vol. 34, no. 1, p. 134, 2006.
[13] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (Version 6.1. 09)[Computer program]. Retrieved May 1, 2019*. 2019.
[14] E. Moulines and J. Laroche, 'Non-parametric techniques for pitch-scale and time-scale modification of speech', *Speech Commun.*, vol. 16, no. 2, pp. 175–205, Feb. 1995, doi: 10.1016/0167-6393(94)00054-E.
[15] J. Swaminathan, A. Krishnan, J. T. Gandour, and Y. Xu, 'Applications of static and dynamic iterated rippled noise to evaluate pitch encoding in the human auditory brainstem', *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 281–287, 2007.
[16] Y. Xu, J. T. Gandour, and A. L. Francis, 'Effects of language experience and stimulus complexity on the categorical perception of pitch direction', J. Acoust. Soc. Am., vol. 120, no. 2, pp. 1063-1074, Aug. 2006, doi: 10.1121/1.2213572.
[17] X. Li, J. Huang, Q. Shi, Y. Mai, and D. Chen, 'Guangzhou Fangyan Yanjiu [Research on Guangzhou Cantonese]', 1995.

PERCEPTION AND IMITATION OF PERIOD DOUBLING

Yaqian Huang

University of California San Diego; Acoustics Research Institute, Austrian Academy of Sciences
yqhuang1016@gmail.com

ABSTRACT

Period doubling consists of voicing periods alternating in amplitude and/or frequency and is often perceived as rough sounding and with an indeterminate pitch. Studies have found that pitch during period doubling is perceived as lower as the degree of amplitude/frequency modulation increased [1, 2]. However, it is unclear how listeners perceive period doubling when identifying linguistic tones. In an artificial language learning study, we test tonal identification and imitation using resynthesized period-doubled tones in native Mandarin and English speakers. We show that period doubling drives a low-tone percept, especially as the modulation, particularly that of frequency, increases. However, higher f0 (300Hz) led to more low-tone responses in amplitude-modulated tokens than lower f0 (200Hz). Period doubling is also imitated with lowered f0. Both Mandarin and English listeners behaved similarly, suggesting that period doubling is not perceived language-specifically. Period doubling is predicted to signal low tones, even when the f0 is high.

Keywords: f0, period doubling, pitch, tone, voice

1. INTRODUCTION

Typical modal voice possesses a single f0 as the primary correlate of pitch. However, when voicing is irregular such that multiple periods can be identified, determining the f0, and consequently the pitch, becomes problematic. This is the case with period doubling, a type of voice that carries at least two simultaneous periodicities, with an alternation in amplitude and/or frequency. This leads to an indeterminate pitch with low and rough quality [3, 4, 5]. It is considered to be a special case of "multiply pulsing", a subtype of creaky voice [3], commonly observed in ~25% of normal speakers' utterances [6].

How is pitch perceived during this voice given the presence of multiple frequencies and periods? Past pitch-matching studies in period doubling found that the perceived pitch became lower as the degree of amplitude and frequency modulation between the two alternating periods increased. The perceptual outcome also differed across f0s and modulation types: a lower f0 facilitated identification of a lower pitch, and the matched pitch dropped more quickly in frequency- than amplitude-modulated tokens [1, 2]. However, it is still unclear how period doubling is used to identify linguistic tones. For example, Mandarin tones are often realized with creaky voice, itself manifested as vocal fry and period doubling [4, 7]. But Huang [7] found that resynthesized tones with period doubling created by the "double pulsing" parameter in the Klatt synthesizer had a negative impact on tonal identification in Mandarin, even with the frequently-creaky dipping Tone 3. Thus, period doubling could hinder rather than facilitate tone identification. It might be perceived as roughness or as competing pitches that disrupt tone perception.

In this study, using an artificial language learning and shadowing paradigm with implicit categories of 'high' and 'low' tones, we test both English and Mandarin listeners' ability to perceive and imitate tonal stimuli manipulated with period doubling. We show that regardless of language background, higher modulation degrees and frequency modulation bias listeners to hear a low tone more frequently; and period doubling is often imitated with lowered f0 and creaky voice quality. Also contrary to past findings, a higher stimulus f0 leads to a low-tone percept in more amplitude-modulated tokens than a lower f0. We further discuss its implications for tonal perception.

2. METHODS

2.1. Stimuli

The stimuli consisted of resynthesized tokens of a vowel [a]. One pulse of the vowel was extracted from a region of stable formants. All stimuli contain a basis formed by duplicating and/or manipulating this extracted pulse according to the experimental condition (illustrated in Fig. 1). To reproduce the characteristics of period doubling, three experimental conditions were created based on the empirical ratios calculated from electroglottography in a scripted Mandarin corpus [7]: amplitude modulation, frequency modulation, and combined amplitude and frequency modulations of every other cycle. The resulting stimuli have alternating pulses of "long-short-long" periods and/or "high-low-high" amplitudes (as described in [8, 9]). All stimuli were 300 ms long and scaled to 70 dB. The following steps detail the process of stimuli creation, achieved using a custom PRAAT script [10].

- Non-modulated: duplicate and concatenate the extracted [a] pulse. (Fig. 1a)
- Amplitude modulation (*am*): the amplitude of the first pulse (a1) is retained, and that of the second pulse (a2) is reduced based on amplitude ratio a1/a2. (Fig. 1b)
- Frequency modulation (*fm*): the duration of the first pulse (d1) lengthens and that of the second pulse (d2) shortens based on frequency ratio d1/d2, while maintaining a fixed presumptive f0 given by 2/(d1+d2). (Fig. 1c)
- Combined amplitude and frequency modulation: the first pulse lengthens, and the second pulse both shrinks and lowers in amplitude. (Fig. 1d)

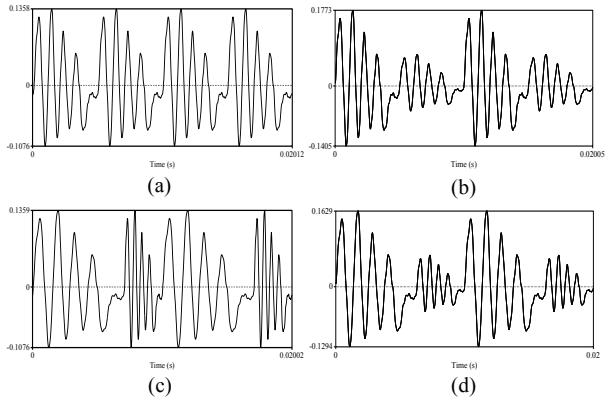


Figure 1: Resynthesized period-doubled pulses of 200 Hz: original unmodified token (a), amplitude-modulated at ratio 2.4 (b), frequency-modulated at 2.4 (c), and amplitude- (2.4) plus frequency-modulated (2.4) (d).

The range of ratios covered at least two standard deviations from the median of either distribution (*am*: 1.43+0.73*2=2.89; *fm*: 1.38+0.56*2=2.5). Table 1 shows the specifics adopted to create the stimuli. Non-modulated stimuli have a ratio of 1, meaning identical amplitudes and/or frequencies, expected to create a modal-sounding tone. Extreme values (4 for *am*; 3 for *fm*) were included to anchor the other end of the distributions to ensure period doubling. We expect to see that the perception changes along with the degrees of modulation from modal (a single pitch) to period doubling (multiple pitches). When the ratio increases, the modulation extent increases, and an increasingly stronger percept of the lower pitch (half of the original one) is expected, because the two alternating pulses become more and more distinct to facilitate the lower pitch percept cued by the lower f0, given by 1/(d1+d2).

	Ratio	Step	Total
Amplitude ratio: a1/a2	Range: (1, 3); Extreme: 4	0.2	11
Frequency ratio: d1/d2	Range: (1, 2.6); Extreme: 3	0.1	17

Table 1: Amplitude and frequency ratios used for modulation based on empirical data.

The pitch of the training stimuli was manipulated using the overlap-add method in PSOLA through the *Manipulation* function in PRAAT. Depending on the experimental condition, 40 training tokens were generated from a gaussian distribution around 200 Hz and 100 Hz (=200/2), or 300 Hz and 150 Hz (=300/2). The standard deviations were chosen considering the non-linearity of pitch perception after converting hertz to semitones to better simulate a comparable distance between different pitches by the human ear.

2.2. Participants and procedure

Thirty native Mandarin speakers (18F, mean age = 20.43), and thirty-one native English speakers (22F, mean age = 20.26) were recruited from the undergraduate population at an institution. No hearing or language disorders were reported. Recordings from twenty-five Mandarin and twenty-nine English speakers were analyzed for the shadowing experiment, due to equipment failure or participants' error during the experiment.

Both perception and shadowing experiments were implemented in PsychoPy [11] in a sound-attenuated booth. Stimuli were played from Focusrite Scarlett 8i6 pre-amplifier connected to a lab computer and participants used headphones throughout. To begin with, participants were told they were learning a novel language which has two tonal categories, represented iconically as ↑ and ↓. The participants went through a familiarization phase where they heard 40 modal tones of a single pitch between half of the f0 and the f0 of the experimental stimuli. They pressed an up or down arrow key upon hearing the sound to get familiarized with the corresponding category ↑ or ↓. Then they were tested in a training phase to categorize the aforementioned 40 tokens, with feedback provided. They were required to pass the training phase with a cumulative accuracy at the minimum of 75% after categorizing at least two rounds of the training tokens (80 tokens). Then they proceeded to the two repetition test blocks to categorize 380 [= (11 *am_degree* x 17 *fm_degree* + 3 extreme) x 2] resynthesized tokens of period doubling into the same categories. The two categories were always given by symbols without explicit explanation. Order of stimulus presentation was fully randomized across listeners and phases. The entire experiment was strictly timed, and participants were warned if they responded too slow (after 1.3 seconds upon hearing the audio). Two misses on the same token were counted as an invalid response. Between the two test blocks, participants were prompted to take a break. The entire perception experiment lasted approximately 20 minutes.

Immediately after the perception experiment, participants were asked to participate in the shadowing experiment. They were recorded using a desk-mounted microphone, with the headphones on. They were asked to reproduce the period-doubled tones (same test stimuli as those in the perception experiment) heard three times by imitating the pitch and voice of the stimuli. They were also allowed to play the audio as many times as needed and were encouraged to practice as many times as needed before they were ready to produce the sounds. Prior to the experiment, they were instructed to produce tone sweeps [12] to assess their vocal range to compare to the pitch imitation of period doubling.

2.3. Analysis

The mean f0s and voice quality correlates of the shadowing productions were extracted using VoiceSauce [13]. Each individual's vocal range was assessed by extracting the max and min f0s in their tone sweeps and coded as covariates included in the statistical models for predicting mean imitated f0s. Voice quality correlates include H1–H2, harmonics-to-noise ratio < 500 Hz (HNR), subharmonic-to-harmonic ratio (SHR), and strength of excitation (SoE). In general, a lower H1–H2 indicates a higher degree of glottal constriction; a lower HNR indicates a noisier quality; a higher SHR indicates stronger subharmonics; and a lower SoE indicates lower energy and possibly more constriction. All these measures in the explained direction would signal a creakier quality [14, 15, 16].

For both perception and production data, logistic and linear mixed-effects models were used to predict binomial categorization [↑ ('up') or ↓ ('down')], imitated f0 and voice quality correlates, given *manipulation type* (no, amplitude, frequency, combined modulation), *f0 condition* (200, 300 Hz), their interactions, *language* (Mandarin, English) and random intercepts of *subject* and *repetition* of the stimuli (first, second) if applicable. The baselines used for within-factor comparison were unmodulated, 200 Hz, and 'up' responses. For voice quality models, imitated f0 was included as a covariate. Further, we included the prior *categorization* as a predictor to probe whether the productions are correlated with the perceptual results.

Data were trimmed to remove outliers determined by log-transformed f0 or reaction time, or acoustic measures that are larger than 2.5 standard deviations from the mean (2.5% of the original data). Following Yuan and Liberman [17], for better interpretation, the imitated f0 was converted to semitones ($= 12 * log_2\,(f0/f0_{base})$), using a speaker-dependent f0 base: 5th percentile of all f0 values per speaker.

3. RESULTS

3.1. Perception during period doubling

No effect of language was found. Participants' categorization varied as a function of modulation type [$\chi^2(3)=1786.61$, $p<.001$], f0 [$\chi^2(1)=6.44$, $p<.05$], and their interactions [$\chi^2(3)=55.24$, $p<.001$]. Frequency modulation had a stronger effect ($\beta=5.00$, $p<.001$) than amplitude modulation ($\beta=2.21$, $p<.05$). The combined frequency and amplitude modulation had the strongest effect ($\beta=5.32$, $p<.001$), showing an additive effect. Pairwise comparisons confirmed this order of effect strength: combined modulation > frequency modulation > amplitude modulation > no modulation (Fig. 2). A linear trend can be observed in both amplitude and frequency modulation: higher modulation degrees lead to a larger proportion of low-tone responses (Fig. 3). Further, both Figs. 2 and 3 show the significant interaction between modulation type and f0 was driven by the increased proportion of 'down' responses for tokens with higher f0 (300 Hz), specifically for amplitude-modulated tokens.

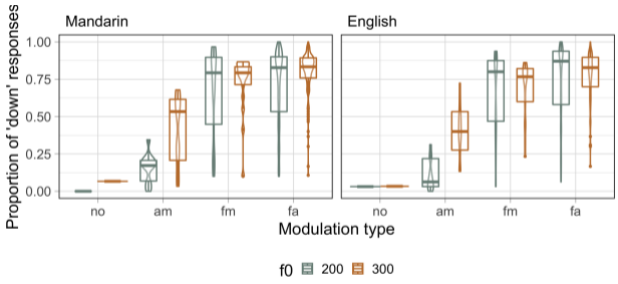


Figure 2: Proportion of 'down' responses as a function of modulation types in Mandarin and English listeners. 'no': unmodulated; 'am': amplitude modulation; 'fm': frequency modulation; 'fa': combined frequency and amplitude modulation.

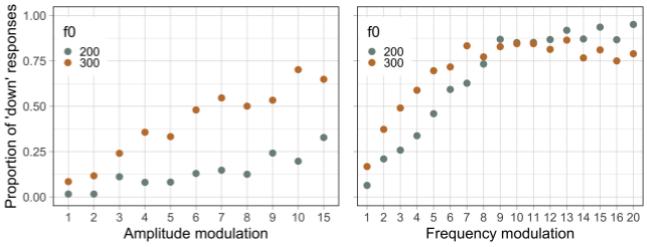


Figure 3: Proportion of 'down' responses vary by amplitude (left) and frequency (right) modulation degrees in purely amplitude- or frequency-modulated tokens.

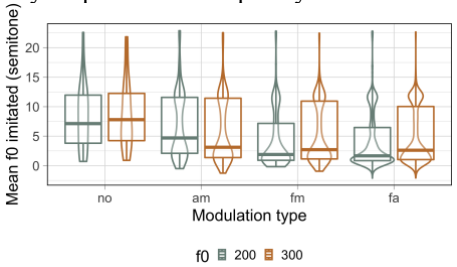


Figure 4: Imitated mean f0 (semitone) as a function of modulation types.

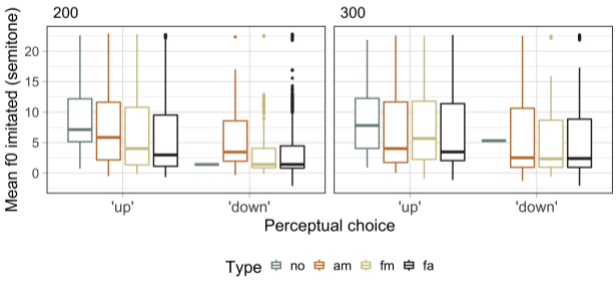


Figure 5: Imitated mean f0 (semitone) in different f0 conditions as a function of perceptual responses. There were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few.

3.2. Shadowing during period doubling

Modulation type and the interaction between modulation type and f0 condition were significant in predicting mean f0 (p<.001). In general, amplitude and frequency modulations were associated with lower imitated f0, and combined modulation had an additive effect. With this main lowering effect of modulation, imitation of amplitude-modulated tokens of 300 Hz tended to be even lower than 200 Hz, similar to the perceptual results (Fig. 4).

3.3. Perception-production link

The imitated f0 averaged across three repeated tokens produced by participants can be predicted using their own *categorization* response in perception. The main effect of *categorization* [F(1)=571.83, p<.001], and its interactions with modulation type [F(3)=5.62, p<.001] and f0 [F(1)=113.44, p<.001] were all significant. Specifically, when participants chose 'down' responses, their imitated f0 was also lower (β=-3.44, p<.05), showing a positive correlation between pitch perception and imitation. A lowering trend according to modulation types was largely observed within the 'up' responses for both stimulus f0s, except that amplitude-modulated 300 Hz tokens with 'up' responses had lower imitated f0 (Fig. 5). In terms of voice quality, when using perceptual response as a predictor, their imitations tended to have lower H1–H2, HNR, and SoE, but higher SHR, especially with frequency and combined modulation. This indicates that the imitated tones were produced constricted, noisier, and quieter, but with more subharmonics, as in the actual tokens heard.

4. DISCUSSION

This study adopted an artificial language learning paradigm to investigate how Mandarin and English listeners perceive and process (pseudo-)linguistic tones realized by period doubling. In both cases, we observe a general trend that the perceived pitch and imitated f0 are lower when the stimuli are modulated to have correlates of period doubling, and when those modulations increase in degree. This suggests that listeners tend to identify a lower pitch during period doubling, which was especially found for frequency modulations or when the modulation degree (of both frequency and amplitude) reaches a certain threshold. This threshold differs depending on the modulation type: amplitude modulation only has around 70% of the tokens identified as low tones even when its modulation degree reaches the extreme (the stronger cycle is 4 times louder than the weaker one).

Thus voicing with amplitude modulation can still signal a 'high' tone, even when the modulation is strong, and especially when the original f0 is *lower*. Both perception and shadowing results support this, though it seems counterintuitive as lower f0 is typically associated with low tones. But because pitch and tone perception is relative, it is possible that if the original unmodulated tokens with an already low f0 (200 Hz) are categorized as the 'high' tone baseline, the effect on the f0 induced by amplitude modulation would not be salient enough to signal pitch lowering for a 'low' tone category.

Yet frequency modulation biases listeners to hear nearly 100% of the tokens as low tones. Listeners are probably more sensitive to changes in period than amplitude of glottal pulses when detecting periodicity or extracting pitch of speech signals. This may be related to findings that listeners tend to be influenced by changes in the frequency rather than time domains. For example, temporal noise measures like jitter and shimmer are not perceptually relevant independently of spectral HNR [18, 14].

Based on the findings that period doubling leads to a low tone bias – regardless of the original f0 – we would predict that the presence of period doubling could be used to signal low tones in languages, even when the f0 of the original tone is high. It will also interfere with high-tone perception, at least with moderate to high modulation.

Speakers also imitated typical characteristics found for period-doubled voice, as well as other creaky voice subtypes. Thus, they not only imitated irregular voicing to match the rough quality, but they also could be reproducing period doubling, or using it to realize roughness in the stimuli. However, we cannot be entirely sure of the precise type of creaky voice produced – whether listeners were able to target the specific subtype of creaky voice. A study designed to test imitation of different creaky voice subtypes would be useful and generalizable.

Lastly, the similar behaviors between Mandarin and English speakers suggest that pitch perception during period doubling may not be language-specific and is likely not influenced by inherent knowledge of lexical tone.

5. REFERENCES

[1] Sun, X., Xu, Y. 2002. Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice*, *16*(4), 443-459.

[2] Bergan, C. C., Titze, I. R., 2001. Perception of pitch and roughness in vocal signals with subharmonics. *Journal of Voice*, *15*(2), 165-175.

[3] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th ICPhS* Glasgow, 0821-1.

[4] Yu, K. M. 2010. Laryngealization and features for Chinese tonal recognition. In *Eleventh Annual Conference of the International Speech Communication Association*. Makuhari, 1529-1532.

[5] Schreibweiss-Merin, D., Terrio, L.M., 1986. Acoustic analysis of diplophonia: A case study. *Perceptual and motor skills*, *63*(2), 755-765.

[6] Klatt, D., Klatt, L. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am*. 87, 820-857.

[7] Huang, Y. 2022. Articulatory properties of period-doubled voice in Mandarin. *Proc. Speech Prosody 2022*, 545-549.

[8] Titze, I. R., 1994. Fluctuations and perturbations in vocal output. *Principles of voice production*, 209-306.

[9] Gerratt, B. R., Kreiman, J. 2001. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, *29*(4), 365-381.

[10] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.24, retrieved 5 September 2021 from http://www.praat.org/

[11] Peirce, J. W. 2007. PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162* (1-2), 8-13.

[12] Keating, P., Kuo, G. 2012. Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050-1060.

[13] Shue, Y. L., Keating, P., Vicenik, C., Yu, K. 2010. VoiceSauce: A program for voice analysis. *Energy*, *1*(H2), H1-A1.

[14] Garellek, M. 2019. The phonetics of voice. In: W. F. Katz & P. F. Assmann (eds), *The Routledge Handbook of Phonetics*. Routledge, 75-106.

[15] Garellek, M., Chai, Y., Huang, Y., Van Doren, M. 2021. Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association*, 1-28.

[16] Kim, S., Matachana, C., Nyman, A., Yu, K. M. 2020. Creak in the phonetic space of low tones in Beijing Mandarin, Cantonese, and White Hmong. In *10th International Conference on Speech Prosody 2020*. Tokyo, 523-527.

[17] Yuan, J., Liberman, M. 2014. F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, *65*, 67-74.

[18] Kreiman, J., Gerratt, B. R. 2005. Perception of aperiodicity in pathological voice. The *Journal of the Acoustical Society of America, 117*, 2201–2211.

PERCEPTUAL SIMILARITY SPACE OF SERBIAN LEXICAL PITCH ACCENTS: EVIDENCE FROM SERBIAN AND FRENCH LISTENERS

Dušan Nikolić, Stephen Winters

University of Calgary
dusan.nikolic1@ucalgary.ca, swinters@ucalgary.ca

ABSTRACT

Serbian lexical pitch accents have been characterized by two related perceptual dimensions: length and pitch [1]. In line with these two dimensions, the present study explored perceptual similarity space of Serbian lexical pitch accents as perceived by Serbian (native) speakers, the speakers of a pitch-accent language, and French (non-native) speakers, the speakers of a language with no word-level prosody. Participants completed an online free classification task which required them to position the given lexical pitch accents close to each other based on their similarity. The multi-dimensional scaling and results revealed length as a clear dimension of perceptual salience of lexical pitch accents, while pitch movement distinctions were often confused. The findings suggest that the perceptual space created by listener groups is largely shaped by listeners' language experience, and that a lack of word-level prosodic categories in one's language does not necessarily induce perceptual "insensitivity" to lexical pitch accents.

Keywords: lexical pitch accents, perceptual space, Serbian, French, free classification task

1. INTRODUCTION

Using a sequence recall task (SRT), Dupoux and colleagues [2, 3] found that French speakers often have difficulty identifying stress contrasts because they fail to encode stress in their phonological representations. This "stress deafness" effect--which we are recasting as "stress insensitivity"--was also investigated by Rahmani and colleagues [4] in a similar study that involved speakers of Japanese, French, Persian, Dutch, and Indonesian. The authors discovered that French and Indonesian listeners exhibit similar insensitivity patterns when perceiving and processing stress contrasts, while no such effect was found for listeners from the rest of the language groups. This combination of results suggests that speakers of *any* language without lexical stress, tone markings or, in fact, any word-level prosodic category fail to perceive the distinctions between stress contrasts. Nikolić and Winters [5] also used the SRT to explore whether listeners might be similarly insensitive to another word-prosodic category – lexical pitch accent. Nikolić and Winters found that both English and Serbian listeners could identify Serbian lexical pitch accent categories at a better than chance rate. This finding is consistent with the observation that speakers of a language which has word-prosodic contrasts can identify other word-prosodic contrasts better than speakers of languages which have no word-prosodic distinction [4].

These studies used the sequence recall task (SRT) to explore the topic of perceptual insensitivity at the word-prosodic level [3, 4, 5]. One of the aims of using this methodological approach is to preclude the listeners in the task from superficially processing and encoding the stimuli at an acoustic level only, which makes it easier for the listeners to detect fine-grained prosodic differences between tokens of different prosodic categories. Instead, the SRT encourages listeners to develop and rely on phonological representations of word-prosodic categories in order to complete the experimental task. With the SRT, listeners are first required to associate arbitrary category labels with example tokens from one of the two word-prosodic categories under study. Listeners then hear a sequence of tokens from different word-prosodic categories and, after a delay, are required to input the sequence of labels associated with those tokens, in the same order. Only entirely correct sequences are then counted as indicative of "non-insensitivity" to the contrast.

Instead of arbitrarily associating labels with categories in this study, we used a free classification task to investigate how listeners might classify tokens of word-prosodic categories based on their similarity, without imposing any labels upon them. Imai and Gardner [6] first developed the free classification task and showed that there were a number of advantages to using it, the main one being that the task did not need to employ experimenter-imposed categories. Interest in the task resurged in linguistic research when Clopper and Pisoni [7] used it to investigate the perceptual similarity of regional dialects in the United States. Listeners in [7] carried out a free classification task in which they were instructed to group regional dialects based on their similarity into as many groups as they heard among the stimuli. Multidimensional scaling analyses of the data showed that gender, geography, and linguistic markedness were the three

determining dimensions of similarity among the dialects. The listeners' performance on the task therefore reflected relevant linguistic and sociolinguistic factors in the perception of regional dialect variation.

Daidone and colleagues [14] used a free classification task to investigate the perceptual space of German vowels and Finnish phonemic length contrasts by L2 learners of these languages. The results showed that the task could be effectively used to describe discriminability of both segmental contrasts (vowels) and suprasegmental contrasts (length), and that the task could be a replicable and reliable method for examining the perception of non-native speech categories.

Clopper [8] described the free classification task by emphasizing that the method could be applied to "examine the perceptual similarity of a range of speech and nonspeech stimulus materials, including linguistic structures such as segments, tones, or intonation contours, and indexical properties such as voices, speaking styles, emotions, dialects, foreign accents, or languages" (p. 575). Apart from [14], no study has yet used this task for examining any of the linguistic purposes described in [8]. With the present study, we addressed this large research gap.

2. THE PRESENT STUDY

In the present study, our broad goal was to use the free classification task to examine the perception of *lexical pitch accents* by listeners of languages both with and without word-prosodic contrasts in their native phonological inventories. To this end, we specifically explored whether native Serbian and French speakers could distinguish between Serbian lexical pitch accents. Serbian is a language that has four lexical pitch accents which combine short and long vowels (S and L) and rising and falling pitch contours (R and F). The lexical pitch accents in Serbian are thus descriptively categorized as short-falling (SF), short-rising (SR), long-falling (LF), and long-rising (LR). Unlike Serbian, French has no word-level prosodic contrasts, and perceptual prominence in French is expressed at the phrasal level only [9]. Based on the previous stress insensitivity findings, we expected Serbian listeners to display greater distances between the lexical pitch accent categories in the perceptual spaces that emerged from the free classification task than the French listeners would. However, since we were applying a new methodology to the study of this task, we also expected the analysis to yield subtler, finer-grained differences in perceptual categorization (for both groups of listeners) that the Sequence Recall Task could not capture.

2.1. Study Design

40 native speakers of Serbian and 60 native speakers of French participated in the study. Both groups of participants completed an online free classification task, in which they freely grouped recordings based on their similarity. The free classification task was adapted from an online rating toolbox built by Donhauser and Klein [10]. The participants listened to 2 groups of 2 nonwords, "rejav" [rejav] and "kavan" [kavan], both of which were patterned on Serbian phonotactic rules. Each word was produced with 4 Serbian lexical pitch accents by one male and one female speaker. 16 nonwords were produced in total (2x4x2). In the free classification task interface, the listeners could hover the computer cursor over distinct icons, spaced around a circle, each of which played a unique recording of a nonword production (Fig. 1). The listeners' task was to place the icons representing each of these words next to each other in the circle, based on similarity. The listeners were instructed to place more similar tokens closer to each other, and more dissimilar tokens further away from each other (Fig. 2).

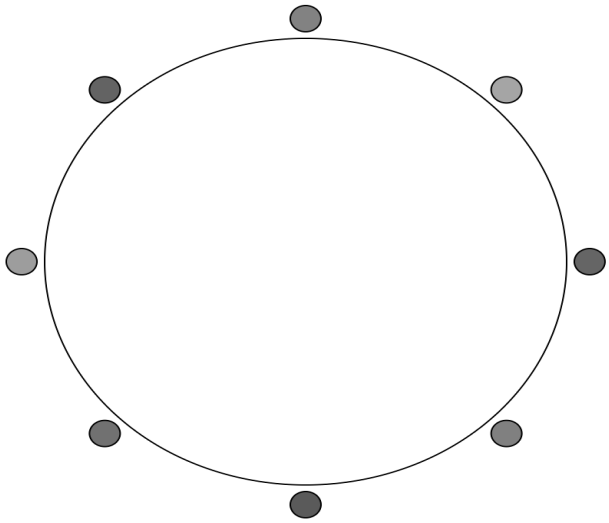


Figure 1: The set-up of the free classification task. Each small circle is a recording. Hovering the cursor over a circle plays the recording.

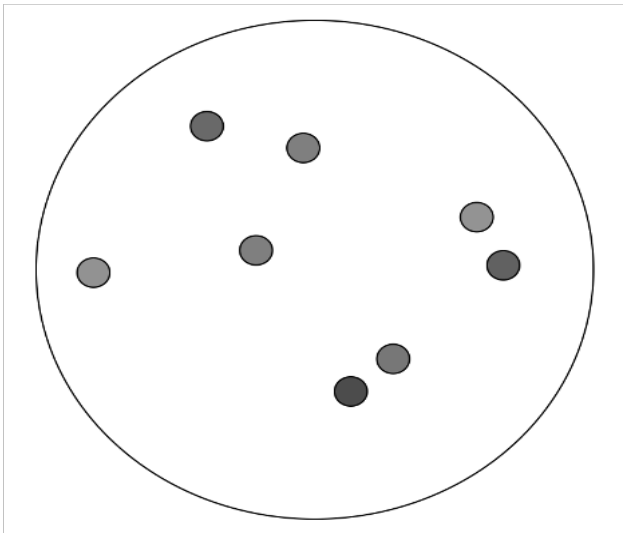


Figure 2: The display of the free classification task after the classification has been completed.

There was no time limit on the task and the participants were instructed *not* to group the words simply based on their segmental structure (i.e., all "kavan" and all "rejav" words) or based on gender (i.e., all male- and all female-produced words next to each other). Since there were sixteen tokens in total, we separated them into distinct groups for two iterations of the task, which we will refer to free classification task I and free classification task II.

2.2. Data Analysis

The perceptual similarity structure of the nonwords was extracted from the free classification data following the guidelines in [10]. Using a python script, we converted the raw coordinates into similarity matrices reflecting the normalized Euclidean distances. The resulting matrices were submitted to multi-dimensional scaling (MDS) to "produce the best fit for the entire distance matrix in a specified number of orthogonal dimensions" ([8], p. 578). All the analyses were performed in RStudio following the procedure outlined in [11].

3. RESULTS

We created an average similarity matrix per listener group (Serbian and French) out of all the received similarity matrices. For Serbian listeners, the most optimal number of dimensions for the free classification task I was 3, while the most appropriate number of dimensions for the free classification task II was 2.

The MDS analysis for the Serbian listeners' performance on the free classification tasks yielded the results displayed in figures 3 and 4. The stress values for the two free classification tasks were 0.177 and 0.24, respectively. The "M" stands for a male-produced item, while the "F" stands for a female-produced item. Thus, the "MRejav_LR" is a male-produced "rejav" nonword in a long-rising lexical pitch accent.

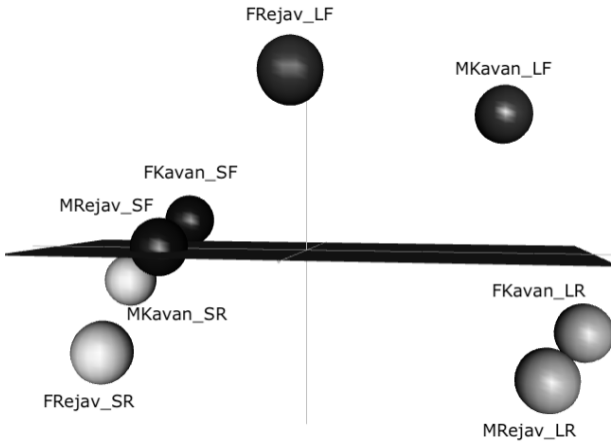


Figure 3: The 3-dimensional multidimensional scaling of Serbian listeners' classification of lexical pitch accents on free classification task I.

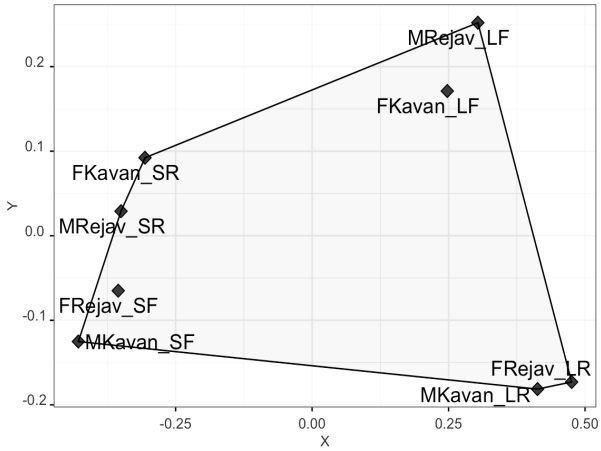


Figure 4: The 2-dimensional multidimensional scaling of Serbian listeners' classification of lexical pitch accents on free classification task II.

An identical MDS analysis was performed on the French listeners' classifications of the lexical pitch accents. Both analyses yielded 3-d models with stress values of 0.18 and 0.06. Figures 5 and 6 below display the perceptual spaces received from these analyses.

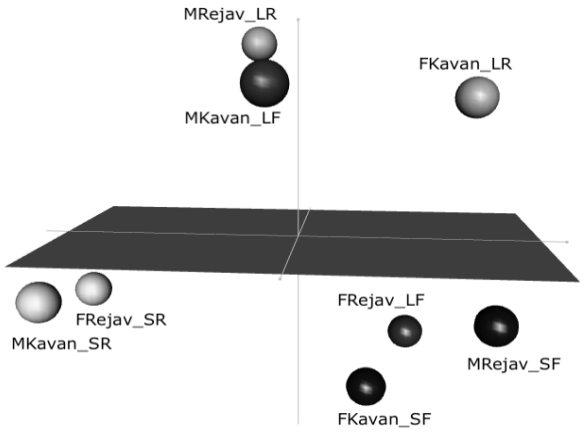


Figure 5: The 3-dimensional multidimensional scaling of French listeners' classification of lexical pitch accents on free classification task I.

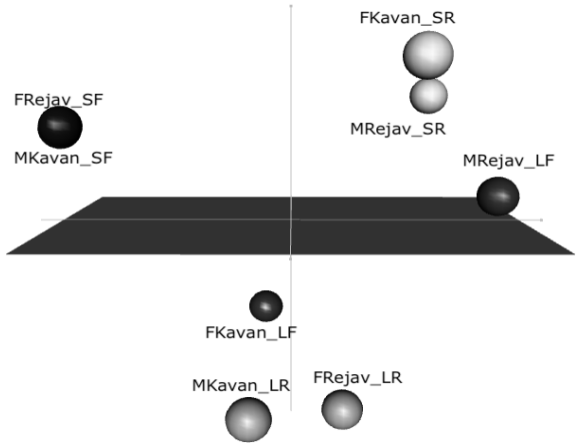


Figure 6: The 3-dimensional multidimensional scaling of French listeners' classification of lexical pitch accents on free classification task II.

4. DISCUSSION

The French listeners clearly distinguish the SF, SR and LR groups, but the classification of the LF tokens is inconsistent. The Serbian listeners distinguish between the LR and the LF tokens, while the SR and SF stimuli form "one group". The Serbian listeners maintain better coherence between tokens of the four existing pitch accent categories than the French listeners do (with the LF category causing the big perceptual grouping problems for the French listeners).

Given all the perceptual support that the task provides, the listeners would likely process the stimuli at an acoustic level. However, the perceptual categories they form are not strictly based on the obvious acoustic distinctions in the stimuli, and they also differ from one language group to the other. There are two main findings of the study. First, language experience and knowledge must be playing some role in shaping those perceptual categories. The classifications formed by both groups consisted of different lexical pitch accent categories. Considering that the stimuli that both groups heard were identical, this distinction is likely reflected in the prosodic system of the listeners' first language. The French listeners could hear some nuanced phonetic differences in the pitch movement as the pitch movement is characteristic of the French post-lexical phrases (e.g., AP, iP, IP). The Serbian listeners could clearly hear the differences in length, and the pitch distinctions in items with long lexical pitch accents. Pitch distinctions were less obvious in items with short vowels, which indicates that the Serbian listeners tend to conflate short-falling and short-rising lexical pitch accents.

Second, methodologically, the free classification task can reveal perceptual sensitivities to word-level prosodic distinctions in listeners of languages with no native word-level prosodic contrasts. That is, unlike what the previous research suggests [4], a lack of word-level prosodic categories does not strictly impose perceptual "insensitivity" to a prosodic category in a task such as the free classification task. Non-native listeners could be processing lexical pitch accent differently from native listeners, i.e., non-native listeners' processing of a word-prosodic category that does not exist in their first language could be phonetic, while native listeners' processing could be phonological (cf. [12, 13]). The future research should attempt to address the processing mechanisms of native and non-native listeners when listening to word-level prosodic categories.

5. CONCLUSION

The study explored the perceptual similarity space of lexical pitch accents by using an adapted and novel version of the free classification task showing more fine-grained phonetic and phonological subtleties of the perceptual sensitivity towards native and non-native prosodic categories. The study has revealed that taking a more nuanced methodological approach to the investigation of prosodic insensitivity may require a revision of our understanding of when and why the phenomenon occurs.

6. REFERENCES

[1] Sredojević, D. 2017. How much do phonetic realisations of Serbian accents actually differ from each other in various dialects?. *Годишњак Филозофског факултета у Новом Саду*, 42(1), 323-337.

[2] Dupoux, E., Pallier, C., Kakehi, K., Mehler, J. 2001. New evidence for prelexical phonological processing in word recognition. *Language and cognitive processes*, 16(5-6), 491-505.

[3] Dupoux, E., Sebastián-Gallés, N., Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': The

case of French learners of Spanish. *Cognition*, 106(2), 682-706.

[4] Rahmani, H., Rietveld, T., Gussenhoven, C. 2015. Stress "deafness" reveals absence of lexical marking of stress or tone in the adult grammar. *PloS one*, 10(12), e0143968.

[5] Nikolić, D., Winters, S. 2022. Are Serbian and English listeners insensitive to lexical pitch accents in Serbian?. *Phonetica*, 79(4), 397-423.

[6] Imai, S., Garner, W. R. 1965. Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, 69(6), 596.

[7] Clopper, C. G., Pisoni, D. B. 2007. Free classification of regional dialects of American English. *Journal of phonetics*, 35(3), 421-438.

[8] Clopper, C. G. 2008. Auditory free classification: Methods and analysis. *Behavior Research Methods*, 40(2), 575-581.

[9] Post, B. 2000. *Tonal and phrasal structures in French intonation* (Vol. 34). The Hague: Thesus.

[10] Donhauser, P. W., Klein, D. 2022. Audio-Tokens: A toolbox for rating, sorting and comparing audio samples in the browser. *Behavior Research Methods*, 1-8.

[11] Levshina, N. 2015. *How to do linguistics with R. Data Exploration and Statistical Analysis*, Amsterdam-Philadelphia.

[12] Chen, J., Best, C. T., & Antoniou, M. 2020. Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, 83, 101013.

[13] Post, B., & Alter, K. (2015). Neural correlates of categorical linguistic and gradient paralinguistic intonation. *Proc. 18th ICPhS* Glasgow, 1-5.

[14] Daidone, D., Lidster, R., & Kruger, F. (2023). Free classification as a method for investigating the perception of nonnative sounds. *Studies in Second Language Acquisition,* 1-27.

Does Learning English Vowels Aid the Perception of English Stress for Learners with Small Vowel Inventory Background?

Manal Allehyani, Rachel Smith, & Clara Cohen
2503139a@student.gla.ac.uk
Glasgow University Laboratory of Phonetics, University of Glasgow

ABSTRACT

This study is part of a larger longitudinal project investigating the relationship between acquiring English vowel quality and acquiring English stress by Arabic learners of English. In this paper, two experiments were conducted to test 1) learners' ability in perceiving vowel contrasts of different difficulty levels, and 2) the relationship between the ability identify vowels and the ability to perceive stress placement in disyllabic nonsense words. An analysis using Signal Detection Theory revealed that all dialect groups were less sensitive to the quality of vowels in the difficult pairs---i.e., pairs distinguished by contrasts absent in the native Arabic dialects---relative to medium and simple pairs. Further, the participants' d-prime scores in vowel identification and stress placement were positively correlated. These findings suggest that acquisition of English stress patterns may benefit from learners' ability to perceive vowel contrasts.

Keywords: stress perception, vowel perception, signal detection theory, language acquisition, Arabic

1. INTRODUCTION

It is well-established in the field of second language acquisition (SLA) that the first language (L1) influences the way second language (L2) learners perceive the target language [1-2]. Arabic and English differ in their vowel inventories, in the phonetic cues used to signal stress, and their phonological stress patterns. Each of these differences makes perceiving and producing English stress a challenging task for Arabs.

Languages' vowel inventories vary drastically in size. Modern Standard Arabic has a small inventory of 3 short vowels /a/, /ɪ/, /u/ and their long counterparts /a:/, /ɪ:/, and /u:/. Some regional variants contain other vowels, such as /e:/ and /o:/, or schwa in Moroccan and Syrian dialects [3]. English, by contrast, has a large inventory of twelve monophthongs and eight diphthongs. This makes it challenging for beginner Arab learners to discriminate English contrastive vowels: learners tend to assimilate several contrastive L2 vowels to the same L1 [4]. Most Arabic dialects also lack central vowels, leading learners to assimilate English central vowels to high or low Arabic categories. This not only hinders their perception and production of vowel contrasts but also creates difficulties in perceiving and producing English stress, which exploits vowel quality as a segmental cue to stress.

The two languages' differences in vowel inventory have implications for the phonetic parameters they use to maintain perceptual contrast between vowels. In terms of vowels' intrinsic F0 (IF0), English speakers maximise the perceptual difference between the F0 of high and low vowels to free more perceptual space for half-open vowels. This is reported to enhance vowel contrasts [5] which is imperative since the crowded vowel space causes some vowels' ellipses to overlap [6]. However, the scenario is different in Arabic. The very small vowel inventory and the lack of many half-open vowels encourage the perceptual difference of IF0 between high and low vowels to be kept to the minimum since the vowel space is not as crowded as it is in English. However, Arabic vowels differ mainly in the dimension of height, with little variation in IF0. Arabic speakers rely more on lengths to maximize vowel contrasts [7,3]. Among the Arabic dialects there are variation in the duration and ratio between long/short vowels. Long and short vowels in Sudanese dialect were reported as longer than their long and short counterparts in Jordanian and Libyan dialects. The Libyan dialect have the shortest vowels durations and the smallest short to long ratios whereas the Jordanian dialect lies in between [8]. In addition, Arabic vowels undergo a process of velarization after velarized consonants. In such vowels, the tongue is lowered and retracted, raising F1 and lowering F2, and resulting in higher intensity values compared to the plain vowels.

Intensity and f0, along with duration, are the major phonetic cues to stress [9]. This brings us to theorize that if IF0 values of the vowels are necessary to enhance vowel contrasts in English, English speakers should use F0 and vowel quality to signal stress because those cues also serve to enhance vowel contrasts. However, since Arabic does not link stress and vowel quality in this way, then Arabic speakers may rely more on their native cues that do not signal vowel contrasts in English—namely, intensity and length. Indeed, studies have shown that the perceptual weighting of stress cues is language-specific [10] in ways that are broadly consistent with this hypothesis.

In English, vowel quality is weighted as the strongest cue to stress (i.e., where a stressed syllable triggers vowel reduction in an adjacent unstressed syllable [11-13]) followed by pitch [14,15], which is in turn more important than duration and intensity [16-19]. In Arabic, by contrast, vowel quality does not play a role in perceiving or producing stress [20-25]. Instead, Arabic speakers rely more on using higher amplitudes and longer vowel durations when producing contrastive stress in Arabic. When learning English, Arabs produced higher F0 than native English speakers which indicates that they are aware of the need to use this cue to highlight stress in producing English stressed syllables.

Regarding stress placement, English is considered as a non-predictable stress language which does not rely on syllable weight when assigning stress. On the other hand, Arabic is classified as a highly predictable-stress language whose stress assignment rules are sensitive to syllable weight. A superheavy ultimate (i.e., CVVC or CVCC) receives stress, otherwise the penultimate receives stress only if heavy (i.e., CVC or CVV) [26]. For this reason, English stress is partially encoded in the lexicon whereas Arabic stress is assigned post-lexically. Moreover, since Arabic stress is predictable, Arabs struggle in Arabic stress perception tasks, producing lower accuracy and higher reaction times than English and Chinese speakers [27]. This deafness of Arabic native speakers may be a by-product of predictability.

We argue that there is a close connection between the factors that make English vowels hard to discriminate for Arabic learners, and the factors that complicate English stress perception for them. The main hypothesis in this paper is thus that learning English vowels will help native speakers of Arabic to better perceive English stress. As a preliminary test of this hypothesis, two perceptual experiments were administered: a vowel perception experiment and a stress placement perception experiment. If improving English vowel perception helps listeners with English stress perception, then learners who are better at perceiving English vowel contrasts should also show better stress perception than learners who struggle to hear those contrasts. It is important to note that work is in progress to gauge native Arabic speakers' sensitivity to each of the stress cues (i.e., intensity, vowel length, and pitch), and how their performance in each of these experiments is correlated to their performance in the perception of English vowels. After six months of exposure to English, the same participants will reconduct the same experiments to see if causation can be established between their vowel perception and their perception in the other three experiments of 1) stress placement perception, 2) perception of pitch experiment, 3) perception of vowel length, and 4) perception of intensity in the two phases of the project. Here, we produce the preliminary results of this project, demonstrating a strong connection between vowel perception and stress placement perception in exactly the direction that would be expected to aid listeners in acquiring English vowel contrasts.

2. METHODS

2.1. Vowel perception experiment

2.1.1. Participants

40 Jordanians, 40 Libyans and 39 Sudanese (mean ages of 33.7, 33.4, and 30.5, respectively) participated in the two experiments in this paper and all other experiments of the larger project. They all live in England.

2.1.2. Materials and procedure

The participants were tested on their ability to discriminate 18 contrastive vowel pairs in an AXB task. Table 1 shows pairs varied in difficulty level (i.e., high, medium, low, plus four easy pairs used as distractors). Difficulty levels were determined based on 1) frequency of reporting problematic English pairs for speakers from different Arabic backgrounds in the TESOL literature [28] and 2) results of perceptual assimilation tasks and vowel confusion matrices in phonetic studies [29, 30].

Pair (A-B)	Target (A)	Difficulty
BATH-TRAP	BATH	High
DRESS-KIT	DRESS	High
LOT-STRUT	LOT	High
LOT-FOOT	LOT	High
STRUT-TRAP	STRUT	High
STRUT-BATH	STRUT	High
LOT-CAUGHT	LOT	Medium
NURSE-FLEECE	NURSE	Medium
CAUGHT-NURSE	CAUGHT	Medium
DRESS-STRUT	DRESS	Medium
LOT-BATH	LOT	Medium
STRUT-FOOT	STRUT	Medium
CAUGHT-GOOSE	CAUGHT	Low
NURSE-STRUT	NURSE	Low
NURSE-BATH	NURSE	Low
GOOSE-FOOT	GOOSE	Low
KIT-FLEECE	KIT	Low
CAUGHT-BATH	CAUGHT	Low

Table 1: The contrastive pairs, their targets and difficulty level.

The A and B stimuli were of the form "bVd" and the X options were of the form "hVd". Each pair was manipulated as AAB, ABB, BAA, and BBA. For example, the pair DRESS-KIT was manipulated as DRESS-DRESS-KIT, DRESS-KIT-KIT, KIT-DRESS-DRESS, and KIT-KIT-DRESS. The X

options were equally either A (e.g., DRESS) or B (i.e., KIT). Each manipulation was repeated twice yielding 8 repetitions per contrastive pair (i.e., 21 contrastive pairs * 8 repetitions = 168 responses/participant). Using Gorilla, the participants were asked to press a key to indicate whether X resembled the first or second word. The experiment lasted for 45 minutes.

2.1.3. Analysis

A signal detection analysis was conducted. The target was always defined as the L2 phoneme that, according to the literature, the participants should find more difficult to learn. This difficulty may be because the phoneme is perceptually new, or because, if it exists in L1, it must be licensed by a specific phonological context. Note that the phoneme difficulty level is relative, based on the other phonemic member within the contrastive pair. For example, the vowel of BATH was the target for the BATH-TRAP pair, but not for the STRUT-BATH pair. In this sense, (A) of each pair was specified as the target (Table 1). AAB and BAA sets were specified as the target groups (i.e., X=A in AAB and BAA) whereas ABB and BBA as the nontarget group (i.e., X=B). Hits, misses, false alarms (FA), and correct rejections (CR) were thus coded as in Table 2.

Stimuli	**Response**	
	X=A	X=B
Target (X =A) in AAB and BAA stimuli	Hit	Miss
Non-target (X=B) in BBA and ABB) stimuli	FA	CR

Table 2: Vowel experiment coding for signal detection analysis.

It was hypothesized that all learner groups would perform worst for high difficulty pairs, represented by lower d-primes suggesting that the pairs are still perceived as the same L1 category, Medium and low difficulty pairs should produce higher d-prime scores. We also hypothesised that the Sudanese group would perform the worst since they have the largest length difference between long and short vowels. The Libyans were hypothesized to perform better than the Sudanese for having the smallest ratio of long and short vowel durations.

2.2. Stress placement experiment

2.2.1. Materials and procedure

The stimuli in this task are disyllabic nonsense words of the form CVC.CVC and CVC.CV:C (e.g., [tɪn.zɪn] and [nɪm.dɪ:f], respectively). The main vowel was in the second syllable of each word form; the 11 monophthongs shown in Table 1 each formed the main vowel in nine different stimuli, with a high, mid and low first-syllable vowel respectively. For each stimulus, stress either fell on the first or second syllables. This resulted in 198 stimuli/participant (i.e., 11 vowels * 9 stimuli/vowel * 2 stress locations). The participants listened to each stimulus once and were asked to press a key to indicate whether stress fell on the first or second syllable. The whole task lasted for a maximum of thirty minutes.

2.2.2. Analysis

A signal detection analysis was again conducted. Based on Arabic's predictable rules, stimuli of the form CVC.CVC attract stress to their first syllable and CVC.CV:C forms attract stress to their second syllable. If the participants were answering based on these Arabic rules, this would indicate that they were insensitive to the phonetic cues to stress in English. Thus, the target for the performance matrices was defined as a response that showed a correct ability to ignore the phonological rules of Arabic stress. This means in CVC.CVC forms, the target is the second stressed syllable and in CVC.CV:C forms, the target is the first stressed syllable. The two targets were combined to generate one d-value per participant (Table 3).

Stimuli	**Response**	
	Target	Non-target
Targets (**CVC**.CV:C and CVC.**CVC**)	Hit	Miss
Non-targets (CVC.**CV:C** and **CVC**.CVC)	FA	CR

Table 3: The targets in CVC.CVC and CVC.CV:C forms combined. Bold font indicates stress.

We hypothesised that the participants who could successfully discriminate vowels in the vowel perception experiment would perform well in this experiment too, showing a positive correlation between acquiring English vowels and acquiring English-like cues to stress. This will be evaluated by testing the strength and direction of the correlation between the d-primes of the two tasks for each participant.

3. RESULTS

3.1. Vowel perception results

Participants' d-prime scores were calculated for each difficulty level per pair, producing three d-prime scores per participant. These scores were then analysed with linear mixed effects regression modelling, using the `lme4` package [31] in `R` [32] P-values for coefficient estimates were calculated based on Satterthwaite denominator degrees of freedom, as provided by the `lmerTest` package [33]. D-prime

was modelled as a function of Difficulty Level, Dialect, and their interaction as fixed effects, with a random intercept included for Participant.

The model summary is provided in Table 4, with a partial effects plot illustrating the patterns in Figure 1. There was a significant interaction between Difficulty Level and Dialect. For Jordanian speakers, d-prime was significantly higher with low and medium difficulty pairs than for high difficulty pairs (low $\beta = 0.93$, $p < .001$; med $\beta = 0.76$, $p < .001$). Libyan speakers did not differ significantly from Jordanian speakers in this respect, and for medium difficulty pairs, Sudanese speakers also did not differ from Jordanians. For low-difficulty pairs, however, Sudanese speakers showed less improvement than Jordanians ($\beta = -0.64$, $p < .001$).

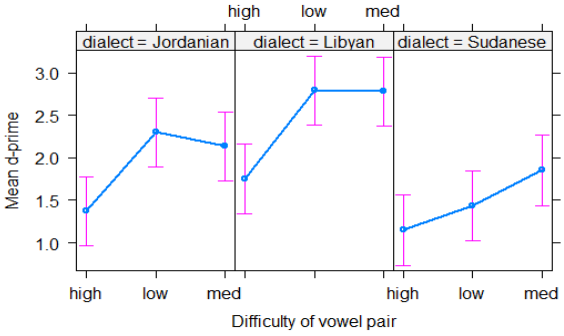


Figure 1: Partial effects plot showing model effects of Difficulty (low, med, high) and Dialect (Jordanian, Libyan, Sudanese) on vowel perception D-prime scores.

Effect	β	SE(β)	t	
Intercept	1.37	0.21	6.64	***
Dif=low	0.93	0.12	7.48	***
Dif=med	0.76	0.12	6.16	***
Dia=Lib	0.38	0.29	1.30	
Dia=Sud	-0.23	0.29	-0.77	
Dif=low × Dia=Lib	0.11	0.17	0.66	
Dif=med × Dia = Lib	0.27	0.17	1.52	
Dif=low × Dia=Sud	-0.64	0.18	-3.65	***
Dif=med × Dia = Sud	-0.06	0.18	-0.33	

Table 4: Summary of mixed effects model predicting vowel perception d-prime against Difficulty (low, med, high) and Dialect (Jordanian, Libyan, Sudanese). All factors were treatment-coded. For Dialect, Jordanian was the reference level; for Difficulty, High was the reference level. Three asterisks indicate $p < .001$.

3.2. Stress placement perception results

To examine the relationship between vowel and stress-placement perception, D-prime scores were calculated on the stress placement task, and re-calculated for the vowel perception task, this time collapsing across all difficulty levels. This produced one d-prime score per listener for each task. As illustrated in Figure 2, the d-prime scores showed a significant positive correlation across tasks ($\rho = .363$, $t(116)=4.20$, $p < .001$). The R^2 of 0.13 indicates that about 13% of the variance in stress placement d-prime was explained by variance in vowel perception d-prime.

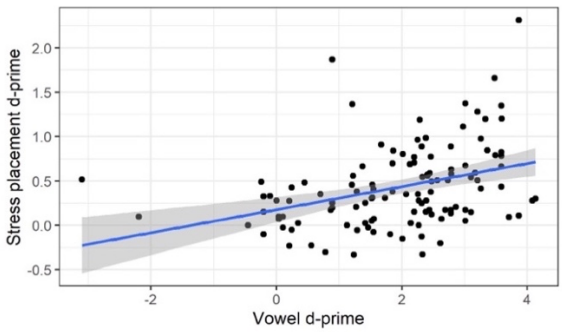


Figure 2: Correlation between vowel perception D-prime scores and stress placement D-prime scores.

4. DISCUSSION AND CONCLUSION

This paper argued for the possible role that learning English vowel contrasts might play, for Arab learners of English, in perceiving English stress location in contrastive disyllabic words. We showed that Arabic learners do struggle to discriminate vowel pairs that prior literature has identified as challenging. Speakers of Jordanian, Libyan and Sudanese Arabic pattern similarly to each other, except that Sudanese learners struggle slightly more with easy contrasts. Crucially, across the two AXB tasks we observed the predicted significant positive correlation between perceiving English vowels and perceiving English stress placement.

We think this relationship is best explained in terms of the two languages' differences in vowel inventories. More specifically, we speculate that the secondary cues that a language uses to differentiate the vowels in its inventory, will be the same as those used to signal stress. The condensed vowel space of English necessitates maximising the differences of IF0 values between high and low vowels to create more space for half-open vowels. In Arabic, by contrast, as a small vowel inventory language with only two half-open vowels (i.e., /e:/ and /o:/), no maximization of the difference in IF0s is required since the vowel space is not crowded. In addition, Arabic vowels are differentiated by length, and pharyngealization processes contribute to intensity differences. We therefore hypothesize that when applying stress, the two languages use different cues that distinguish their vowels. English uses vowel quality and pitch whereas Arabic uses duration and intensity. Work is in progress to test this relation by assessing Arabic learners' ability to discriminate subtle variation in the three stress cues of pitch, intensity and duration, and exploring the relationship between these abilities and their vowel perception performance.

5. REFERENCES

[1] Best, C. (1995) 'A direct realist view of cross-language speech perception'. In W. Strange (ed.) *Speech Perception and Linguistic Experience: Issues in cross-language research.* York Press.

[2] Flege, J. (1995) 'Second language speech learning: theory, findings, and problems.' In W. Strange (ed.) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press.

[3] Al-Tamimi, J (2007) Static and Dynamic cues in Vowel Production: a cross dialectal study in Jordanian and Moroccan Arabic

[4] Iverson, P. and Evans, G. (2009) 'Learning English vowels with different first-language vowel systems II: auditory training for native Spanish and German speakers.' *Journal of Acoustical Society of America*, 126 (2), pp 866-877.

[5] Van Hoof, J. and Verhoeven, J. (2011) 'Intrinsic vowel F0, the size of vowel inventories and second language acquisition.' *Journal of Phonetics*, 39, pp 168-177.

[6] Lee, W. (2011) 'Dispersion and variability of vowels of different vowel inventory sizes.' *Journal of Acoustical Society of America*. *Proceedings of Meetings on Acoustics,* 14, pp 1-9.

[7] Newman, D. and J. Verhoeven (2002) Frequency analysis of Arabic vowels in Connected Speech. *Antwerp Papers in Linguistics*. 100: 77-86.

[8] Ahmed, A. (2008) Production and Perception of Libyan Arabic Vowels. *PhD Thesis*. University of Newcastle.

[9] Tatham, M and Morton, K. (1988) The Perception and Production of Speech. A textbook.

[10] Peperkamp, S and Dupoux, E. (2002) 'A Typological study of stress deafness'. *Lab Phonol* (7), pp 203–240.

[11] Fry, D. (1964) 'The dependence of stress judgments on vowel formant structure', In: S. Karger, Basel (eds.). *Proceedings of the 5th International Congress of Phonetic Science*. University of Munster, pp 306–311.

[12] Zhang, Y. and Francis, A. (2010) 'The weighing of vowel quality in native and non-native listeners' perception of English lexical stress'. *Journal of Phonetics*, 38, pp 260-271.

[13] Frost, D. (2011) 'Stress and cues to relative prominence in English and French: a perceptual study.' *Journal of International Phonetic Association*, 41, 1, pp 67-84.

[14] Chrabaszcz, A., Winn, M., Lin, C. & Idsardi, W. (2014) 'Acoustic cues to the perception of word stress by English, Mandarin, and Russian speakers.' *Journal of Speech, Language, and Hearing*. Vol. 57, Issue 4, pp 1468-1479.

[15]Cooper, N., Cutler, A.,& Wales, R. (2002). 'Constrain ts of lexical stress on lexical access in English: Evidence from native and non-native listeners.' *Language and Speech*, 45 (3), pp 207–228.

[16] Fry, D. (1955) 'Duration and intensity as physical correlates of linguistic stress'. *Journal of Acoustical Society of America*, 27, 765.

[17] Fry, D. (1958) 'Experiments in the perception of stress'. *Language and Speech*, 1(2), pp126–152.

[18] Lieberman, P. (1960) 'Some acoustic correlates of word stress in American English'. *The Journal of Acoustical Society of America.* 32 (4), pp 451- 454.

[19] Morton, J. and Jassem, W. (1965) 'Acoustic correlates of stress'. *Language and Speech*, (8), pp 148-158.

[20] Zuraiq, W. (2005) The Production of Lexical Stress by Native Speakers of Arabic and English and by Arab Learners of English. *A PhD Thesis*. University of Kansas.

[21] Bouchhioua, N. (2008) 'The role of vowel quality in cuing stress and accent in Tunisian Arabic, native English, and L2 English.' *In Speech Prosody*. *Fourth Conference on Speech Prosody, pp* 539-542.

[22] Bouchhioua, N. (2016a) 'Typological variation in the phonetic realization of lexical and phrasal stress: Southern British English vs. Tunisian Arabic'.

[23] Bouchhioua, N. (2016b) 'Stress and accent in Tunisian Arabic'. *Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics*.

[24] Zuraiq, W. and Sereno, J. (2007) 'English lexical stress cues in native English and non-native Arabic speakers.' *Ratio 1.*

[25] Almbark, R., Bouchhioua, N., & Hellmuth, S. (2014) 'Acquiring the phonetics and phonology of English word stress: comparing learners from different L1 background.' *Proceedings of the International Symposium of the Acquisition of Second Language Speech*. Concordia Working Papers in Applied Linguistics.

[26] Watson, JCE (2011) 'Word stress in Arabic'. In: Ostendorp, M., Ewen, C., Hume, E. and Rice, K., (eds.) The Blackwell companion to phonology.

[27] Lin, Ch. (2018) The Perception and Production of Arabic lexical Stress by Learners of Arabic: A Usage-Based Account. *PhD dissertation*. University of Michigan.

[28] Othman, A. and Abubaker, S. (2021) 'Investigating pronunciation difficulties encountered by Arab students in learning English sounds.' *Global Libyan Journal* 50.

[29] Almbark, R. (2014) 'Perception and production of SSBE vowels by foreign language learners: towards a foreign language model'. *Proceedings of the Int'l Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics* (5).

[30] Alshangiti, W. and Evans, B. (2018) 'The perception and production of British English vowels and consonants by Arabic learners of English'. *Journal of Phonetics* (68), pp 15-31.

[31] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4.

[32] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[33] Kuznetsova A, Brockhoff PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. doi:10.18637/jss.v082.i13.

Conditions on Adaptation to an Unfamiliar Lexical Tone System: The Role of Quantity and Quality of Exposure

Liang Zhao[1], Shayne Sloggett[1], and Eleanor Chodroff[1,2]

[1]University of York, Dept. of Language and Linguistic Science,
[2]University of Zurich, Dept. of Computational Linguistics
liang.zhao@york.ac.uk, shayne.sloggett@york.ac.uk, eleanor.chodroff@uzh.ch

ABSTRACT

Variation in lexical tone systems across dialects presents a perceptual challenge to speech adaptation. For instance, several Mandarin regional dialects have four phonological tones, but the phonetic realization of those tones differs considerably. Previous studies have demonstrated that listeners readily accommodate dialectal variation in lexical tone systems through incidental exposure. The present study examined whether adaptation was facilitated through the presence of explicit minimal-pair tone contrasts or increased exposure to the dialect. We found that rapid adaptation to the novel tone system was persistent even when minimal-pair sentences were removed from the stimuli and about one minute of incidental exposure was available with no repetition. Minimal-pair contrast was not necessary for adaptation. Increased exposure through repeated trials reliably enhanced sensitivity to the novel tone system.

Keywords: speech perception, lexical tone, perceptual adaptation, tone perception, Mandarin dialects

1. INTRODUCTION

Perceptual adaptation to unfamiliar speech requires adequate exposure to the target speech especially when the initial input is received later in adulthood. What makes the exposure "*adequate*" has become a central question to research on speech adaptation. Previous findings suggest that the adaptation outcome may be modulated by both quality and quantity of the spoken stimuli.

Quality refers to the type, structure and source of information in the experimental exposure. For example, lexical information has been considered beneficial for adaptation to a novel sound contrast [1, 2]. Hayes-Harb [2] tested English speakers' discrimination of /g/- and /k/-like novel sounds after auditory training with either minimal pairs of the sounds or with members of a [g]-[k] continuum without lexical meaning in a bimodal distribution that favored tokens towards the endpoints of the continuum. The results showed that adaptation to the novel contrast occurred with statistical learning alone, but discrimination was significantly enhanced when a lexical contrast was present. That said, listeners are able to rapidly discriminate between previously unheard linguistic contrasts: listeners are able to generalize heard patterns to new segments [3] and words unheard in the training [4]. Nevertheless, it is likely that the adaptation process could be facilitated if an explicit lexical contrast, such as a minimal pair, was heard in the exposure.

Listeners also frequently rely on explicit training to gather ample information for adaptation. If more information is available in the test stimuli, such as in sentences and passages, adaptation may happen without explicit training. Clarke & Garrett [5] reported listeners' rapid adaptation to Spanish- and Chinese-accented speech with one minute of incidental exposure to the sentence stimuli. In addition, exposure with about 16 sentences was found sufficient to initiate adaptation to a foreign-accented talker [6]. Crucially, however, listeners significantly improved over the course of the experiment. Even though adaptation can be reasonably successful in a short period, increased exposure may help listeners to generalize heard patterns to novel sounds or speakers.

In the present experiment, we investigated aspects of quality and quantity of incidental exposure in adaptation to a novel lexical tone system. Recently, Zhao, Sloggett, & Chodroff [7] identified that native Standard Mandarin listeners adapted to a novel lexical tone system from the Chengdu Mandarin dialect with less than two minutes of incidental exposure. Chengdu Mandarin has the same underlying four-tone system as Standard Mandarin, but disparate phonetic tone realizations (Figure 1) [8, 9, 10]. In the experiment, 24 sentence frames were manipulated to have either a semantically plausible (low-surprisal) or an implausible (high-surprisal) meaning by altering the tone category of one target word in the sentence-medial or sentence-final position. Native speakers of Standard Mandarin were asked to judge whether the spoken sentence was plausible or not with a "yes" or a "no" response in both their native dialect, Standard Mandarin, and the unfamiliar dialect, Chengdu Mandarin. Accurate judgment on the sentence plausibility and/or different response times between the high- and low-surprisal conditions indicated awareness of the surprisal tone manipulation and adaptation to the unfamiliar tone system. The results showed that listeners consistently slowed down for the high-surprisal sentences in Chengdu Mandarin starting from the beginning of the experiment, which strongly indicated rapid adaptation to the unfamiliar phonetic tones without explicit exposure to the dialect. This finding also leads to the hypothesis that phonetic tone information is always processed — even if such information may have little influence in lexical decision [11]. This has broader implications for models of speech perception involving lexical tone [11, 14, 15].

A critical aspect in the design of the experiment was that participants heard both the low- and high-surprisal versions of the sentence in each dialect, which provided minimal-pair sentences that contrasted in semantic plausibility and lexical tone category [7]. The presentation of minimal-pair sentences may facilitate rapid adaptation to a novel tone system. To test the potential conditions for adapting to a novel tone system with incidental exposure, the present study investigated 1) whether adaptation can still be achieved when minimal pairs are removed, and 2) if increasing the amount of incidental exposure would facilitate adaptation. We removed the minimal-pair

contrast in the dialect-specific stimuli and introduced three repetitions of all trials in the new experiment. Minimal-pair presentation may be necessary for adaptation, in which case, we would expect that listeners have comparable response times for Chengdu Mandarin between high- and low-surprisal sentences, at least in the initial trials. However, a difference in response time may still emerge as incidental exposure increases with repetition.

The present study first investigated the effect of repetition on adaptation without minimal-pair presentation via the main experiment. To single out the effect of minimal-pair presentation, the response-time data in the first repetition block from the present experiment was extracted and compared to that from the previous experiment [7]. The absence of minimal-pair sentences was expected to impede adaptation. However, the effect of non-minimal-pair presentation was expected to be overcome by increasing the amount of ambient exposure through repetition.

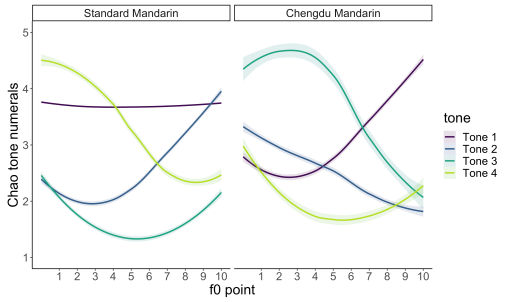


Figure 1: Smoothed lexical tone contours of Standard Mandarin and Chengdu Mandarin converted to Chao Tone numerals [10]. Ribbons reflect ±1 standard error of the mean.

2. METHODS

The current experiment replicated the design of the previous experiment of Standard Mandarin and Chengdu Mandarin [7], but with three repetition blocks and without the presence of minimal pair sentences in each dialect.

2.1. Participants

Thirteen native speakers of Standard Mandarin who reported little or no knowledge of Chengdu Mandarin participated in the experiment. No participant reported any hearing or reading impairments.

2.2. Materials

The experiment used the same 24 sentence frames as in Zhao et al. [7]. For each sentence frame, the lexical tone category of one target word was manipulated to have a semantically plausible (high-surprisal) or an implausible (low-surprisal) meaning (see Table 1). To avoid having both low- and high-surprisal versions of the sentence presented in the same dialect, each surprisal version was assigned to a different dialect. In addition, each critical-word tone category and position (medial or final, balanced evenly) were presented approximately the same number of times in each dialect.

2.3. Procedure

The experiment was built using Gorilla Experiment Builder [12]. Participants were asked to complete the semantic plausibility judgment task on their personal device with internet access in a quiet room, and with headphones if possible. The participants were made aware that they would be listening to sentences spoken in either the familiar Standard Mandarin or an unfamiliar Mandarin dialect. They were then presented with a test audio so they could adjust the volume of the sound output to a comfortable level.

The practice phase consisted of two trials that introduced the test phase procedure. To start the trial, the participant pressed a "play audio" button on the screen. After listening, participants responded to the question *"Does this sentence make sense"* by clicking on either the "yes" or "no" response button on the screen. Feedback was then provided by displaying the orthographic form of the sentence on the screen. The two trials contained one high-surprisal and one low-surprisal Standard Mandarin sentence. These sentences were not repeated in the main experiment.

In the test phase, the participants received one of the two lists described in the materials. The task was identical to the practice trials except that no feedback was given. There were 24 trials in each block, and a total of three blocks for each dialect (24 trials × 3 repetitions × 2 dialects). The participants heard Chengdu Mandarin blocks first, then the Standard Mandarin blocks to avoid task-based learning effects in the adaptation process due to prior familiarity with the sentence frames in the native dialect. The trials in each repetition block were randomized.

low-surprisal sentence	a) 有 一只 鹰 在 天上 飞 *You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei1* *There is an eagle in the sky flying* "There is an eagle flying in the sky"
high-surprisal sentence	b)* 有 一只 鹰 在 天上 肥* *You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei2** *There is an eagle in the sky gaining weight** "There is an eagle gaining weight in the sky"

Table 1: An example sentence item across surprisal conditions.

2.4. Data analysis

The effects of *surprisal* (high surprisal vs. low surprisal), *dialect* (Standard Mandarin vs. Chengdu Mandarin) and *repetition* were assessed on accuracy and response time. Responses that matched the expected plausibility judgment were considered *correct*: "yes" responses to low-surprisal plausible sentences and "no" responses to high-surprisal implausible sentences. Response time was calculated as the interval between the end of the audio file and the click registering a judgment.

Accuracy was modeled with a Bayesian logistic mixed-effects regression, and response time with a Bayesian log-normal mixed-effects regression, both with weakly informative priors [13]. Each model included fixed effects of surprisal, dialect, two repetition contrasts, and the full set of interactions. The random effect structure for participants included an intercept and slopes for surprisal, dialect, repetition contrasts and all the interactions, and for sentence frame, an intercept and a random slope for surprisal. The priors for the accuracy model were $N(0, 20)$ for the intercept, main effects and interactions, and $N(0, 0.05)$ for random effects. For the response-time model, the priors were $N(7, 1)$ for the intercept, $N(0, 1)$ for main effects and interactions, and $N(0, 0.01)$ for random effects. The model was run for 2000 iterations with a burn-in period of 1000 iterations. Surprisal and dialect were sum-coded (*surprisal:* high-surprisal = 1, low-surprisal = −1; *dialect:*

Chengdu Mandarin = 1, Standard Mandarin = −1). Reverse Helmert coding was used for the three-level repetition factor, comparing one level to the mean of the previous level(s) (*repetition contrast 1*: block 2 = 1/2, block 1= -1/2, block 3 = 0; *repetition contrast 2*: block 3 = 2/3, block 2 = − 1/3, block 1 = −1/3). An estimate was deemed credible in its direction of influence on the dependent variable if the 95% credible interval excluded 0 (i.e., no effect).

For the comparison between the present and previous experiments, response time was modeled with fixed effects of surprisal, dialect, trial, presentation (*presentation*: with-minimal-pair design = 1, no-minimal-pair design = −1), and the interaction of surprisal, dialect and presentation. The random effects for participants included an intercept and slopes for surprisal and dialect, and for frame an intercept and a slope for surprisal. The same priors were used as the above response-time model.

3. RESULTS

3.1. Results from the current experiment

3.1.2. Accuracy

The overall accuracy across dialect and surprisal conditions (Figure 2) closely resembled the previous findings: 1) high near-ceiling accuracy in both surprisal conditions for the familiar dialect and 2) considerably lower accuracy in the high-surprisal condition for the unfamiliar dialect, Chengdu Mandarin. A gradual improvement can be seen over the three blocks in the Chengdu high-surprisal conditions.

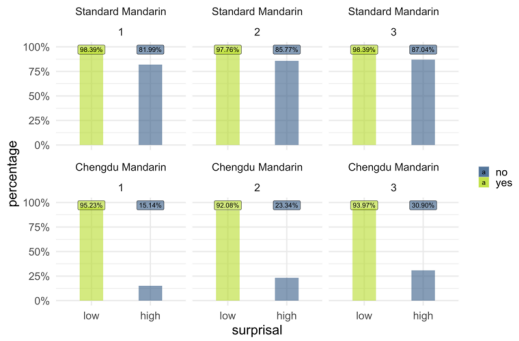


Figure 2: Percentage of *correct* responses across dialect, surprisal and repetition ("1, 2, 3" refer to the repetition blocks).

The model revealed credible main effects of surprisal, dialect, and the interaction between dialect and surprisal as in the previous study. Specifically, accuracy was higher in the low-surprisal condition than in the high-surprisal condition (*surprisal:* β = −7.64, 95% CI = [−14.89, −3.10]); accuracy was higher for sentences spoken in Standard Mandarin than in Chengdu Mandarin (*dialect:* β = −7.04, 95% CI = [−14.24, −2.51]); the credible interaction between surprisal and dialect indicated higher accuracy for Standard Mandarin low-surprisal sentences (*dialect x surprisal:* β = 5.13, 95% CI = [0.58, 12.32]). For the effect of repetition, accuracy was reliably different from the first to the second block of repetition (*repetition contrast 1*: β = −7.78, 95% CI = [−21.09, −0.19]), but not from the first two blocks to the third block (*repetition contrast 2*: β = 4.79, 95% CI = [−6.06, 19.63]). The second repetition block reliably interacted with surprisal (*surprisal x repetition contrast 1*: β = 8.29, 95% CI = [0.73, 21.59]) and dialect (*dialect x repetition contrast 1*: β = 7.70, 95% CI = [0.11, 20.95]), while the third repetition block showed no credible interaction with the other factors (*surprisal x repetition contrast 2*: β = −4.23, 95% CI = [−18.96, 6.62]; *dialect x repetition contrast 2*: β = −4.39, 95% CI = [−19.14, 6.43]; *surprisal x dialect x repetition contrast 2*: β = 4.63, 95% CI = [−6.13, 19.48]). This suggested that accuracy improved after the second repetition of the trials for the high-surprisal sentences compared to the low-surprisal, and for Chengdu sentences compared to Standard Mandarin, but these did not reliably improve in the third block.

3.1.2. Response time

The response-time model identified credible effects of all tested factors and their interactions, except for the interaction between surprisal and the second repetition contrast. To be exact, the credible effects of surprisal (*surprisal:* β = 0.15, 95% CI = [0.12, 0.18]), dialect (*dialect:* β = 0.11, 95% CI = [0.08, 0.13]) and the interaction between surprisal and dialect (*surprisal x dialect:* β = −0.07, 95% CI = [−0.09, −0.05]) replicated the patterns found in the previous study: listeners were reliably slower in the high-surprisal condition in both dialects, with greater difference between the surprisal conditions in Standard Mandarin than in Chengdu Mandarin (Figure 3). Nevertheless, a difference was still observed between high- and low-surprisal conditions in Chengdu Mandarin.

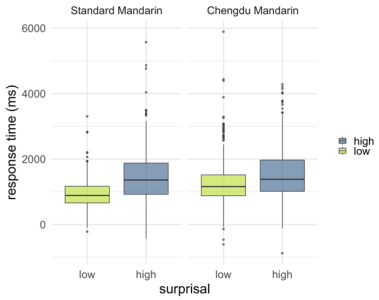


Figure 3: Response times across dialect and surprisal conditions.

For the effect of repetition (Figure 4), all responses generally accelerated block by block (*repetition contrast 1*: β = −0.16, 95% CI = [−0.21, −0.11]; *repetition contrast 2*: β = −0.19, 95% CI = [−0.24, −0.14]). Moreover, slower responses were found for Chengdu sentences after each repetition, relative to Standard Mandarin sentences (*dialect x repetition contrast 1*: β = 0.11, 95% CI = [0.06, 0.17]; *dialect x repetition contrast 2*: β = 0.10, 95% CI = [0.05, 0.14]). However, response times were credibly faster for high-surprisal sentences from the first to the second block (*surprisal x repetition contrast 1*: β = −0.06, 95% CI = [−0.11, −0.003]), possibly driven by the faster responses to Standard Mandarin sentences, but there was no difference towards the third block (*surprisal x repetition contrast 2*: β = −0.02, 95% CI = [−0.07, 0.02]). In fact, response times were reliably modulated by the three-way interactions between surprisal, dialect and both repetition contrasts (*surprisal x dialect x repetition contrast 1*: β = 0.07, 95% CI = [0.02, 0.13]; *surprisal x dialect x repetition contrast 2*: β = 0.06, 95% CI = [0.01, 0.11]), suggesting block-wise slowdown for Chengdu high-surprisal sentences, but block-wise speed-up for Standard Mandarin high-surprisal sentences.

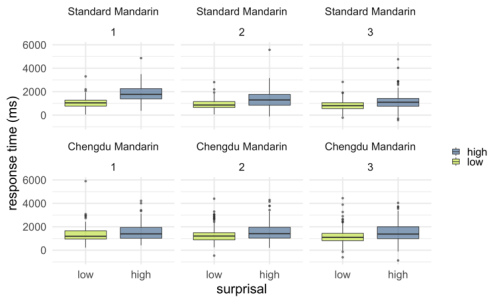


Figure 4: Response times across dialect, surprisal and repetition conditions ("1, 2, 3" refer to the repetition blocks).

3.2. The effect of minimal pairs

To examine the effect of minimal-pair presentation on response time, we compared the first block of data in the present study to the data in the previous study, which only differed in the presence of minimal pairs (Figure 5). Accuracy was not examined as surprisal did not influence responses to Chengdu sentences in either study. No credible effect of presentation was detected between the two designs (*presentation*: $\beta = 0.05$, 95% CI = [−0.03, 0.12]), indicating that the effect of minimal-pair presentation was not as salient as expected. Response times were consistently slower for high-surprisal sentences in both experiments (*surprisal*: $\beta = 0.20$, 95% CI = [0.15, 0.25]), reflecting listeners' awareness of surprisal manipulation even without minimal pairs or repetition. The high-surprisal slowdown did not reliably interact with presentation (*surprisal x presentation*: $\beta = 0.01$, 95% CI = [−0.03, 0.06]; *surprisal x dialect x presentation*: $\beta = -0.0006$, 95% CI = [−0.04, 0.03]), indicating that the removal of minimal pairs did not credibly affect listeners' sensitivity to the surprisal manipulation. Nevertheless, the estimated mean difference between high and low surprisal differed numerically between the two sets of data in the expected direction: about 170 ms without minimal pairs, and 270 ms with minimal pairs. Minimal-pair presentation may have numerically facilitated adaptation to the novel tone system, resulting in greater distinction between the surprisal manipulations; removal of the minimal pairs reduced, but did not obviate the effect of surprisal. There was also a credible interaction between presentation and dialect (*dialect x presentation*: $\beta = 0.06$, 95% CI = [0.003, 0.11]), indicating slower responses to Chengdu sentences when minimal pairs were present.

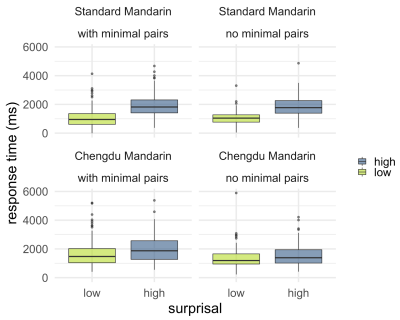


Figure 5: Response times across dialect, surprisal and presentation conditions in the previous (with-minimal-pair) and the new (no-minimal-pair) experiments

4. DISCUSSION

The previous experiment on Chengdu Mandarin revealed rapid adaptation to the novel tone system with incidental exposure containing minimal-pair sentences [7]. The current experiment found that rapid adaptation to the novel tone system was persistent even when minimal-pair sentences were removed from the stimuli and only minimal incidental exposure was available. Enhancement in quality (minimal pairs) and quantity (repetition) of the exposure both facilitated adaptation in terms of accuracy and response time, but they were not necessary factors.

The present study showed similar results for the effects of dialect and surprisal as in the previous study, which indicated successful adaptation and shared perceptual mechanisms for novel tone processing with [7] and without minimal pairs. Increased incidental exposure reliably boosted adaptation to the unfamiliar tone system, evidenced by improved accuracy and increased difference in response times between the low- and high-surprisal conditions starting from the second block of repetition.

Specifically for the effect of minimal-pair presentation, it was surprising to find that listeners were sensitive to the surprisal manipulations even when the minimal pairs were removed from exposure. In fact, adaptation occurred under rather adverse conditions, where incidental exposure was limited to one repetition and with no minimal-pair sentences in the same dialect. Nevertheless, inclusion of minimal pairs in the stimuli can assist discrimination between the low- and high-surprisal meanings, and may potentially direct more attention to the tone contrast and ease the process of adaptation or learning of the new tone system.

As minimal incidental exposure was sufficient for adapting to the unfamiliar tone system with or without minimal pairs in the exposure, it is unlikely that listeners relied on increased exposure to initiate adaptation. Repetition was more likely a consolidating factor as the mappings between the phonological categories and the novel phonetic tone realizations were reinforced over repetitions.

5. CONCLUSION

The present study investigated the potential factors impacting adaptation to an unfamiliar tone system by testing the effects of minimal-pair presentation of the stimuli and increased incidental exposure. We found that both factors were adequate to induce adaptation, though neither was necessary. The current study also has implications of more readily processed tones for models of speech perception. Further analysis should delve into the tone-specific adaptation which was not reported due to limited space. As the current experiment examined adaptation with incidental exposure, further research could be done to assess the effect of explicit training for adaptation to unfamiliar lexical tone systems.

6. ACKNOWLEDGMENTS

The final author was supported by SNSF grant 208460. The authors would also like to thank Sam Hellmuth and Paul Foulkes for helpful feedback.

7. REFERENCES

[1] Norris, D., McQueen, J. M., & Cutler, A. "Perceptual learning in speech," *Cognitive Psychology*, *47*(2), 204–238. 2003.

[2] Hayes-Harb, R. "Lexical and statistical evidence in the acquisition of second language phonemes," *Second Language Research*, 23(1), 65-94. 2007.

[3] Maye, J., Aslin, R. N., & Tanenhaus, M. K. "The weckud wetch of the wast: Lexical adaptation to a novel accent," *Cognitive Science*, *32*(3), 543–562. 2008.

[4] McQueen, J. M., Cutler, A., & Norris, D. "Phonological abstraction in the mental lexicon," *Cognitive science*, 30(6), 1113-1126. 2006.

[5] Clarke, C. M., & Garrett, M. F. "Rapid adaptation to foreign-accented English," *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. 2004.

[6] Bradlow, A. R., & Bent, T. "Perceptual adaptation to non-native speech," *Cognition*, *106*(2), 707–729. 2008.

[7] Zhao, L., Sloggett, S., & Chodroff, E. "Top-down and Bottom-up Processing of Familiar and Unfamiliar Mandarin Dialect Tone Systems," in *Proc. Speech Prosody*, May.2022, pp. 842-846.

[8] Hou, J.Y. (侯精一). *The Modern Outline of Chinese Dialects* (现代汉语方言概论). China: Shanghai Education Press (上海教育出版社), 2002.

[9] Li, Rong. (李荣). "Chengdu Dialect Dictionary (成都方言词典)," *The Modern Dictionaries of Chinese Dialects* (现代汉语方言大词典). China: Jiangsu Education Press (江苏教育出版社), 2002.

[10] Zhao, L., & Chodroff, E. "The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects," in *Proc. 13th Language Resources and Evaluation Conference*, May.2022, pp. 1985-1990.

[11] Gao, X., Yan, T. T., Tang, D. L., Huang, T., Shu, H., Nan, Y., & Zhang, Y. X. "What makes lexical tone special: a reverse accessing model for tonal speech perception," *Frontiers in psychology*, 2830 (10). 2019.

[12] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods*, *52*(1), 388–407. 2020.

[13] Bürkner, P.-C. (n.d.). *Advanced Bayesian Multilevel Modeling with the R Package brms*. Aug.2022. Available: https://github.com/paul-buerkner/brms.

[14] Strauss, T. J., Harris, H. D., & Magnuson, J. S. "jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition," *Behavior Research Methods*, 39(1), 19-30. 2007.

[15] L. Shuai and J. G. Malins, "Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese," *Behavior Research Methods*, 49 (1), 230– 241. 2017.

PROSE OR POETRY? A PERCEPTIVE PHONETIC STUDY

Valentina Colonna[1], Antonio Romano[2]

[1]University of Granada (Spain), Department of General Linguistics and Literary Theory
[2]University of Turin (Italy), Department of Foreign Languages, Literatures and Modern Cultures
valentinacolonna@ugr.es, antonio.romano@unito.it

ABSTRACT

In their spoken form, poetry and prose distinguish in many and acknowledged aspects, but are they discernible at a perceptual level by their prosody? This topic received very little attention. This paper aims to present the results of a perceptive Phonetics study based on the recognition, by a group of students, of prose and poetry among manipulated delexicalised stimuli. Taking into account a selection of variables, the perceptive approach, combined with statistical analysis, enabled us to evaluate also the influence of the textual construction, the speaker's vocal type, and the audio manipulation. Among the main outcomes, we observed poetry recognition in its "metrical" reading and through characterizing traits, such as longer pauses and a slower speech rate, connected to a specific use of melody. This task has proved to be a useful tool for deepening the acoustic dimension of the literary text, putting the focus on the listener's perception.

Keywords: Perceptive Phonetics; Experimental Task; Poetry; Prose; Voices of Italian Poets

1. INTRODUCTION

Poetry and Prose diverge in many aspects that have effects also on the prosodic dimension: first of all, silence and layout, length, density, rhythm, and melody. However, although the rhythmic issue of both has been addressed from a theoretical point of view, Linguistics works concerning the recognition of poetry and prose on a perceptual level are still little considered[1].

Among the most representative, we can mention Fónagy's [1] phonological approach, Bolinger's [2] studies, Bröggelwirth's [3] work in the German context, Wagner's [4] approach, which focused on metrical feet, and MacArthur and Miller's [5], focusing on variables in spontaneous speech and poetry readings.

While studies showed poetry and prose readings' distinctions in the prosodic production, on a perceptual level, in the German context [3] [6], the issue has not yet been addressed for Italian, of which this paper represents an initial proposal of study.

This work takes into consideration previous linguistic studies aimed at other forms of speech. The main reference works include 't Hart et al. [7] and Romano [8] [9]: in this study, a particular reference will be to the AMPER project methodology [10]. Further literature references were Moulines & Verhelst [11], Van Bezooijen & Gooskens [12], and Knoll *et al.* [13].

2. THE STUDY

This study has been conducted as part of the VIP-*Voices of Italian Poets* project [14] and consists of the analysis of a perceptive Phonetics task submitted to the attention of a group of native Italian students and composed of manipulated audio data, which we go to introduce.

2.1. Data

The acoustic data forming part of the experimental task consist of recordings by an actress-announcer (woman, an actor (man), and a poetess (woman)). The three speakers read two poems by two different authors of the Italian 20th century: one by Giorgio Caproni (*Alba*) and one by Giuseppe Ungaretti (*Sirene*).

The first composition is in fixed metrics (sonnet), but with a language and style closer to our time; the second poem is in free metrics (mainly with hendecasyllables and septenaries), with more conservative language and style. Although with overall short lengths, the poems are very different in form and content. The selection criteria also depend, besides the existence of the original recording by the poet (see Colonna [15]), on the hypothesis of stylistic reading changes depending on the theme-form combination (rhythmic-metric) (see Bologna [16]) and the poem's rhythmical, syntactical, and rhetorical issue.

The choice of different types of speakers was made considering the need to test different prosodic styles and approaches to the poetry reading, to assess their influence on perception.

The speakers were asked to read firstly "metrically" (see [15]), following the layout of the poem as closely as possible, respecting the verse as the unit of measurement, and, then, to read "syntactically", following the punctuation and syntaxis, the same text transposed into prose.

The recordings were made in the "Arturo Genre" Laboratory of Experimental Phonetics of the University of Turin. Each file has been sampled at 16 kHz, and agreed to let their voice be used for this study under the GDPR.

2.2. Methodology

Poetry and prose presented clear characteristic traits in all three speakers: lower Fluency and Speech Rate, internal melodic variety, and greater use of pauses, broader overall, in the first case; higher Fluency and Speech Rate, higher internal melodic variety and less use of pauses, shorter overall, in the second case.

Overall, we collected 6 recordings, including prose/poetry prosodic readings of every poem (2) among the 3 speakers. For each recording, a selected portion of a maximum of 20 s has been detected and annotated at a vowel level through PRAAT program. To compare the degree of perception between different types of alterations, we manipulated the data. More in detail, we used a script to extract some data (duration, f_0, and intensity), then manually corrected and manipulated in 4 ways: energy normalisation; flattening of f_0 average and energy normalisation; flattening of durations and energy normalisation; further flattening of durations (shorter) and energy normalisation. Later, these data were entered into MATLAB software which, through the AMPER routines [17], produced 4 synthetic audios per recording (delexicalised stimuli), as well as one synthetic audio without further manipulation.

Fig. 1 and Fig. 2 show two examples of manipulation (fixed intensity and fixed f_0) in announcer-actress' prose reading.

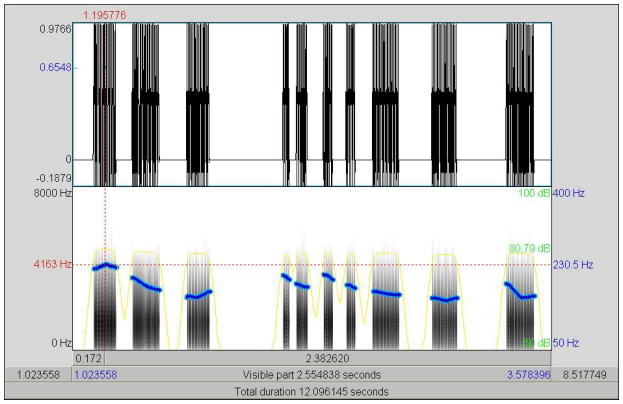


Figure 1: PRAAT window of synthetic voice stimulus (fixed intensity).

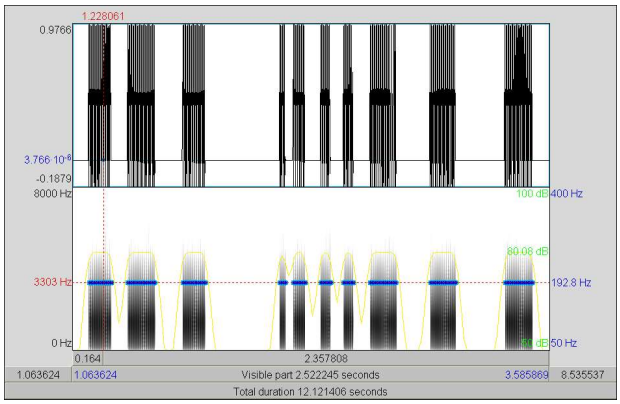


Figure 2: Praat window of synthetic voice stimulus (fixed f_0).

Before the creation and submission of the final task (by PRAAT), we create and test a training test (identification test) aimed at the recognition of prose or poetry (in random order). The task we will present in this paper is a discrimination test composed of 32 random input pairs of stimuli, each proposing one prose and one poetry, with equal handling characteristics and the same speaker. The listeners were invited to choose only one of the two stimuli, after a complete listening of both, by pressing a button on the left (if poetry was detected in the first stimulus) or right (if poetry was detected in the second stimulus), to distinguish poetry between the two stimuli.

More in detail, the task included the following manipulated data: 5 audios of 1 poetess-Caproni-poetry (1 natural synthetic+4manipulations), 5 audios of 1 poetess-Caproni-prose (1 natural synthetic + 4manipulations), 5 audios of 1 announcer-Caproni-poetry (1 natural synthetic + 4manipulations), 5 audios of 1 announcer-Caproni-prose (1 natural synthetic + 4manipulations), 3 audio of 1 announcer-Ungaretti-poetry (1 natural synthetic + 2 manipulations), 3 audio of 1 announcer-Ungaretti-prose (1 natural synthetic + 2 manipulations), 3 audio of 1 actor-Ungaretti-poetry (1 natural synthetic + 2 manipulations), 3 audio of 1 actor-Ungaretti-prose (1 natural synthetic + 2 manipulations).

The task submission was carried out thanks to the involvement of 40 students of the Dep. of Foreign Languages and Literatures and Modern Cultures of the Univ. of Turin (General Linguistics Course, Master, Prof. A. Romano). They have been informed and consented to the study to be conducted, under a general University agreement and according to the GDPR.

After the training task (with double listening to each stimulus) each student participated in the discrimination task.

2.3. Results

The study results showed a major convergence in the recognition of poetry from prose[2]. In particular, with the discrimination task participants reached 72.6% of correct answers (23.55/32 correct answers). Limited are the cases of less than half of correct answers and all correct or wrong answers. On the other hand, the margin of chance (16/32) is almost absent.

The results highlighted a different reception according the reading styles of the three speakers: in particular, the reception of the poetess's reading appeared to be quite distinct from the actor's reading, which was the most frequently recognised, followed by that of the announcer-actress. Furthermore, the stimulus pair with the highest number of correct responses was that of the actor-reading of Ungaretti, with manipulation of duration (average). The percentages of recognised stimuli were 80% for the actor's voice, 76% for the announcer's voice, and 67% for the poetess' reading (Fig. 3).

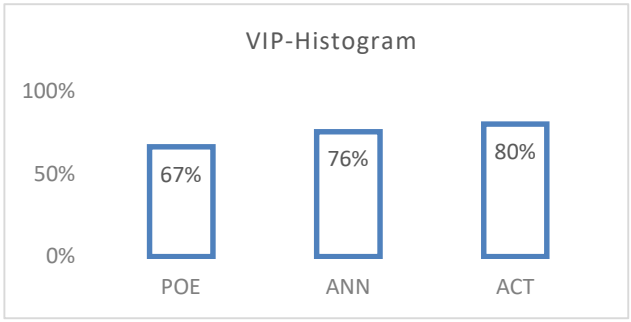


Figure 3: VIP-Histogram. Percentage of correct answers per vocal type.

Considering instead the percentages of correct responses concerning manipulation (see Fig. 4), we noted that the highest percentage of correct responses is found with the synthetic stimuli with flattening of f_0, followed by the unaltered stimuli and the stimuli with flattening of syllable mean durations (1st type).

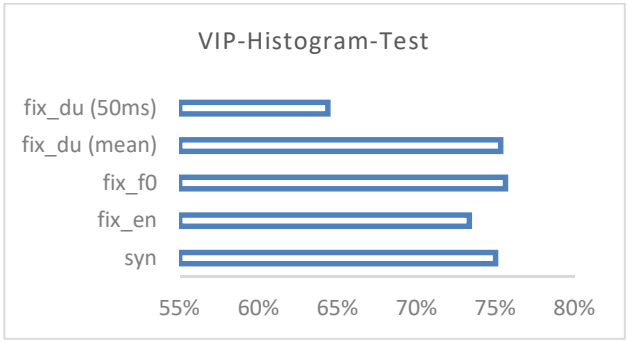


Figure 4: VIP-Histogram-test. Percentage of correct answers per type of stimulus.

Slightly lower are the results for the stimuli with energy flattening and even lower are those in which we adopted a second duration flattening (50 ms). In conclusion, speech rate, mean f_0, and intonation patterns appear to be central elements in recognition, even when syllable duration is normalised.

Two statistical ANOVA tests enabled verification of the significance of this analysis. In the first, the *type of speaker* factor (actor, announcer, poet) was fixed; in the second, the factor of the *type of manipulation* was fixed together with the *type of poem* read.

The first test showed that the type of voice had a statistically significant effect: the number of correct answers has been higher depending on the type of voice (higher with actor-voice, medium with announcer-actress, low with the poet): $F(2, 29) = 7.794$, $p<0.002$. We report in Fig. 4 the related Distribution Plot (ANOVA).

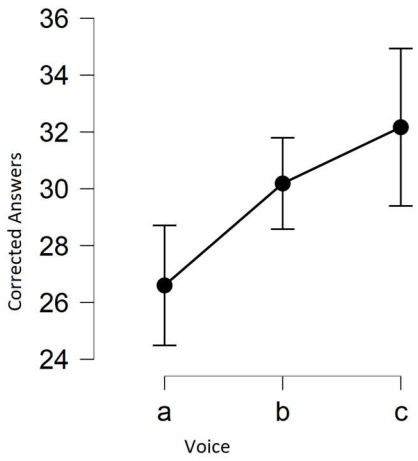


Figure 5: Distribution Plot (ANOVA) (a-poetess, b-announcer/actress, c-actor).

In Fig. 6 we report also the related Q-Q-plot, which shows the standardised residuals fit on the diagonal, suggesting that the hypotheses, normality, and linearity were not violated.

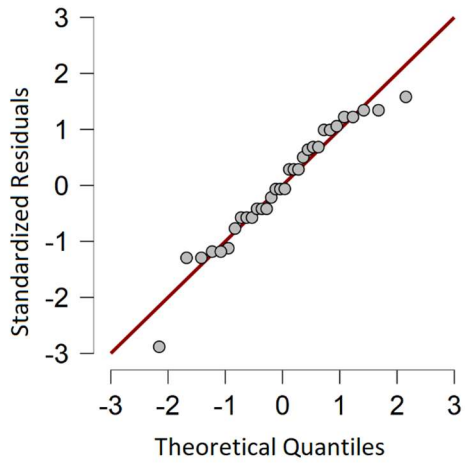


Figure 6: Q-Q-plot related to the first ANOVA test.

The second ANOVA test, with the fixed factors of manipulation (5 types set out above) and type of poem (Ungaretti/Caproni's compositions) shows that the type of manipulation test is statistically non-significant: F (4, 22) = 0.301, p<0.874. In contrast, the type of poem turns out to be a significant factor, F (1, 22) = 7.961, p<0.010. Finally, a non-significant interaction is given by the interaction of the two factors instead, F (4, 22) = 0.711, p<0.593. We report in Fig. 5 the related Distribution Plot (ANOVA).

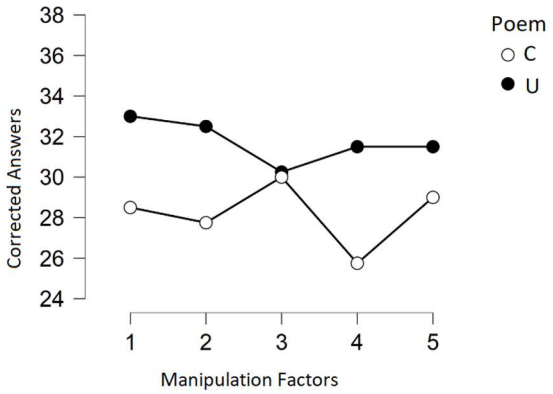


Figure 7: Second Distribution Plot (ANOVA) (5 Manipulation Factors).

This result confirms our hypothesis of an influence of the textual-formal and content difference (e.g. the substantial metrical difference between a text in hendecasyllables, such as Caproni's poem and one with various meters, as well as the clear linguistic difference) on different performances and receptions in the reader, conditioning the prosody (perceptible also in the delexicalised stimuli). At the same time, different voices (and ways to manipulate them) represent a further element of condition and influence in reception. In this way, the theme of the text and its form, together with the individual vocality, can be considered elements affecting the different modes of reading and their perception.

Furthermore, it can be deduced that excessive alteration of duration (too fast) prevents discrimination, as vocalic duration is a determining element and is considered to be distinctive. It also turns out that speech rate, mean f_0, and intonation patterns are central elements in recognition, even though the syllable duration is normalised.

3. CONCLUSION

This perceptual investigation showed that recognition of poetry from prose, through a metrical reading of the composition, is possible, even in the presence of delexicalised synthetic stimuli and possible further alterations of their acoustic parameters. This type of experimental task has proved to be a useful tool to deepen the acoustic dimension of the literary text.

It has emerged that the percentage of exact recognition of the poetry reading (with "metric" style) is greater where f_0 and duration are smoothed out, as well as in the unaltered synthetic stimuli.

The results highlighted that the perception of poetry prosody passes through recognisable and specific traits: first and foremost a specific speech rate, that appears central in poetic reading, which cannot be separated from a specific melodic use of the voice, melodically more homogenous, and a larger dimension of pauses.

Furthermore, the selected audience, familiar with literary and linguistic studies, easily associated poetry with a primarily actorial style. This result led to the hypothesis that a non-expert audience predominantly depicts its imagery of the poetic voice with this vocal modality, better known because it is usually proposed as a support in teaching. It must be connected to a historical tradition of actors' poetry reading characterised by a recitative-declamatory approach and a divergent tendency in poets' reading to increase the speech rate, in general terms.

The little association of the poetic voice with the voice of a poet(ess) can be understood as a manifestation of a lesser familiarity in listening to this kind of vocality and could be ascribable to the Italian didactic tradition mentioned above, on the whole, little linked to contemporary poets and their readings in the last twenty years. We, therefore, believe that a consolidated tradition of readings by authors and actors of the second half of the Twentieth century could be considered more settled in the ear of listeners. It would be related to stylistic features, as an emphatic and declamatory style, and slower speech rates (see [15]). In the common imagination, it would better correspond to the "painting" of the poetic voice, which would present a specific cadence in which the verse retains its distinguishable measure even when it is free and files are manipulated.

The perceptual dimension is to be understood as a significant part of what we consider a possible literary vocal imagery. A broadening of the research to a richer variety of texts and authors, as well as a larger number of vocal samples per category and speakers categories, together with a more differentiated listening audience and further differentiations of stimuli's combination, would represent a further step to describe the vocal literary imagery. As much as we believe that it can only be truly complete with the support of interdisciplinary studies, including cognitive sciences, we believe that research such as this can be considered an initial contribution to exploring the discrimination between poetry and prose (and their respective more typical reading styles).

4. REFERENCES

[1] Fónagy, I. 1987. Le lettere vive in poesia. In: *Le lettere vive*, 289-342. Bari: Dedalo.

[2] Bolinger, D. 1989. *Intonation and its uses.* London: Edward Arnold.

[3] Bröggelwirth, J. 2007. *Ein rhythmisch-prosodisches Modell lyrischen Sprechstils*. Ph.D. Dissertation, Universität Bonn. Then in: Bröggelwirth, J. 2010. *Ein rhythmisch-prosodisches Modell lyrischen Sprechstils.* Munich: GRIN.

[4] Wagner, P. 2012. Meter Specific Timing and Prominence in German Poetry and Prose. In: Niebuhr, O. (ed.) *Prosodies – Context, Function, Communication*, 219-236. Berlin/New York: De Gruyter.

[5] MacArthur, M., Miller, L. M. 2018. After Scansion: Visualizing, Deforming and Listening to Poetic Prosody. *Stanford ARCADE. Literature, the Humanities, & the World*, Dec. 13. https://arcade.stanford.edu/content/after-scansion-visualizing-deforming-and-listening-poetic-prosody.

[6] Hussein, H., Meyer-Sickendiek, B., Baumann, T. 2020. Free Verse and Beyond: How to Classify Post-modern Spoken Poetry. *Proc. 10th International Conference on Speech Prosody*. Tokyo (Japan), May 2020, 690-694.

[7] 't Hart, J., Collier, R., Cohen, A. 1990. *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press.

[8] Romano, A. 1997. Persistence of prosodic features between dialectal and standard Italian utterances in six sub-varieties of a region of Southern Italy (Salento): first assessments of the results of a recognition test and an instrumental analysis. *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 175-178, doi: 10.21437/Eurospeech.1997-71

[9] Romano, A. 1999. Rising-Falling contours in Speech: a Metaphore of Tension-Resolution Schemes in European Musical Traditions? Evidence from Regional Varieties of Italian. *Atti del 8th workshop CSNLP – "Language, Vision & Music" (Galway, Irlanda, 8-12 Agosto 1999)*, 232-238.

[10] AMPER *Project-Atlas Multimédia Prosodique de l'Espace Roman* http://www.lfsag.unito.it/amper/amper.html

[11] Moulines, E., Werner, V. 1995. Time-domain and frequency-domain techniques for prosodic modification of speech. In: Bastiaan Kleijn, W., Paliwal, K. K. (eds.) *Speech coding and synthesis*, 519-555. Amsterdam: Elsevier.

[12] Van Bezooijen, R., Gooskens, Ch. 1999. Identification of language varieties: The contribution of different linguistic levels. *Journal of language and social psychology*, 18(1), 31-48.

[13] Knoll, M. A., Uther, M., Costall, A. 2009. "Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners". *Speech Communication*, 51(3), 210-216.

[14] Colonna, V. 2021. *Voices of Italian Poets. Analisi fonetica e storia della lettura della poesia italiana dagli anni Sessanta a oggi*. PhD Dissertation, University of Genoa-University of Turin.

[15] Colonna, V. 2022. *Voices of Italian Poets. Storia e analisi fonetica della lettura della poesia italiana del Novecento*. Alessandria: Dell'Orso.

[16] Bologna, C. 1990. Les interprètes. *Le grand atlas des littératures a été réalisé par Encyclopoedia Universalis*, 80-83. Paris: Encyclopaedia Universalis.

[17] AMPER Routines: http://www.lfsag.unito.it/amper/amper.html.

[1] This study takes up and reworks part of the first author's PhD Dissertation.

[2] The training test gave the same result: participants reached 68% of correct answers related to the identification of poetry and 66% related to prose.

CLEAR SPEECH SOUNDS FAST: THE IMPACT OF SPEAKING MODE VARIATION ON PERCEIVED TEMPO

Leendert Plug[1], Yue Zheng[1], Rachel Smith[2]

[1] University of Leeds, [2]University of Glasgow
l.plug@leeds.ac.uk

ABSTRACT

Research has identified multiple acoustic cues for perceived speech tempo variation. Together, these warrant the proposal that speech which conveys more complex spectral information is heard as taking more time than speech conveying less complex information; if timing is controlled, the former sounds faster than the latter. We therefore hypothesize that hyper-articulated speech sounds faster than normal speech when articulation rates are controlled. We report on two listening experiments which address this hypothesis using clear and normal sentence productions. Experiment 1 assesses listeners' ability to separate tempo and speaking mode judgements; Experiment 2 uses the same stimuli to probe the direction of any effect of speaking mode variation on perceived tempo. The results confirm that compressed clear sentence productions sound faster than normal productions with the same duration, although the precise mechanism underlying this effect remains to be established.

Keywords: speech perception, tempo, clear speech, English

1. INTRODUCTION

Research on speech tempo perception has identified multiple non-temporal acoustic cues for perceived tempo variation: increases in *f*0 span and level, intensity level and vowel space size have all been shown to raise perceived tempo [1–4]. Generalizing across these findings, Weirich and Simpson [4] propose that 'there is a link between the perception of spectral information and the perception of time', such that '[a]n event that conveys more information in terms of acoustic parameters (a greater range in *f*0, a greater vowel space traversed) is perceived as taking more time than an event conveying less information'.

The acoustic parameters listed above are all implicated in variation along the 'H&H continuum' [5, 6], including shifts between normal and clear speaking modes [7]. According to [5, 6], speakers adjust their articulatory effort and precision on a continuum between loosely and firmly controlled ('hypo-' and 'hyper-') articulation on the basis of the estimated need for speech clarity given situational constraints. Many studies have investigated speakers' adjustments when asked to speak *clearly*—as if communicating in noise, or with a hearing-impaired or second-language listener [8–14]. These studies have shown that recurrent correlates of clear speech include increased *f*0 and intensity levels and *f*0 span, and greater dispersion of vowels in the F1–F2 space [7–9, 14]. Clear speech is also associated with a decrease in coarticulation, resulting in more easily delimitable, 'canonical' articulations [10, 12, 13, 15].

These findings warrant the hypothesis that when articulation rate is controlled, the non-temporal features of clear speech will make it sound fast relative to normal speech, as clear speech conveys more spectral information in the same time window [4]. A complicating factor is that clear speech is generally articulated slowly [7–9, 14], and listeners may draw on their knowledge of associations among non-temporal and temporal parameters in making tempo judgements [16, 17]. This would predict that when listeners recognize speech as clear, they are biased towards hearing it as relatively slow. Either way, we can hypothesize that speech mode variation between normal and clear has an impact on speech tempo perception when articulation rate is controlled.

We report on two experiments that address these related hypotheses. Both were pairwise comparison tasks with pairs of normal and clear sentence productions, controlled for articulation rate, serving as stimuli. In Experiment 1, participants judged whether the productions in each pair differed in tempo, speaking mode, both, or neither. This allowed us to assess the extent to which listeners can separate these two perceptual parameters and test the general hypothesis that speaking mode variation has an impact on tempo perception. In Experiment 2, participants heard the same pairs and judged how the productions differed in tempo; here their attention was not drawn to, nor were they asked to make judgements about, speaking mode. This allowed us to test the more specific hypothesis that clear speech sounds faster than normal speech when articulation rate is controlled.

2. EXPERIMENT 1

2.1. Participants

82 native British English speakers aged 18–35 years were recruited through *Prolific* (www.prolific.co). All reported normal hearing and passed a screening task in which they identified sentence production pairs with identical members. All were paid.

2.2. Stimuli

We used the LUCID Corpus [14], available via *SpeechBox* [18], to create our stimuli. The corpus includes a set of 144 sentences which were read by 40 Southern Standard British English speakers in both normal and clear speaking modes. For the normal mode, speakers were instructed to speak 'casually, as if talking to a friend'; for the clear mode, 'clearly as if talking to someone who is hearing impaired' [14]. We selected the ten sentences in Table 1 as produced by six speakers—four female and two male—who did not make large loudness and voice quality adjustments in speaking clearly, but whose 'clear' productions still had distinct articulatory and prosodic characteristics.

Table 1: Stimulus sentences with summary statistics for articulation rate (mean and SD, N=6) split by speaking mode.

Sentence	Articulation rate (sylls/sec)	
	normal	clear
The bear belongs to the children.	4.52 (0.69)	3.13 (0.67)
I've lost my box of pins.	5.36 (1.33)	3.47 (0.65)
The dog barked at the sheep.	4.41 (0.65)	3.32 (0.61)
The old lady ate the peach.	5.12 (0.73)	3.50 (0.26)
The music blared from the shack.	4.26 (0.35)	3.07 (0.26)
The pear belongs to the teacher.	5.12 (0.41)	3.30 (0.60)
The seat came with the car.	4.03 (0.20)	2.83 (0.54)
She's going to sue the firm.	4.52 (0.39)	3.09 (0.26)
The suit was full of holes.	4.49 (0.82)	3.08 (0.26)
Jonathan gave his wife a bush.	5.22 (0.79)	3.28 (0.59)

Acoustic analysis following [14] confirmed that across the 120 sentence productions, all clear ones are longer and more slowly articulated than their corresponding normal production (Table 1); most also have a higher Long-Term Average Spectrum and greater *f*0 dispersion. Segmentations produced by the BAS tools *G2P* and *WebMAUS* [19] and additional auditory analysis highlighted systematic differences in phone-level articulation between the normal and clear sentence productions consistent with descriptions of SSBE full forms and connected speech phonetics. For example, in the normal productions [t, k, g] are mostly unreleased when followed by another consonant: *dog barked* [dɒgˀbɑːkˀt]. In the clear sentence productions, these plosives are predominantly released.

For each speaker's production of each sentence in Table 1, we created four sentence pairs, as outlined in Table 2. Pair members were separated by a 1-second within-pair silence. Compressions were done through PSOLA in Praat [20]. The target of the compression of a clear production was always the duration of the matching normal production. In SPEED pairs, members differ in (utterance-level) articulation rate but not speaking mode. In PRECISION pairs, they differ in speaking mode but not articulation rate. In BOTH pairs, they differ in speaking mode and articulation rate. In NEITHER pairs, members are identical. To manage the number of trials per participant, we created four lists (N=120), counterbalancing for sentence and within-pair order.

Table 2: Stimulus pair types; '~' indicates that pairs were included in both possible orders.

Type	Pair members
SPEED	clear ~ $\text{clear}_{\text{compressed}}$
PRECISION	normal ~ $\text{clear}_{\text{compressed}}$
BOTH	normal ~ clear
NEITHER	normal – normal clear – clear

2.3. Task and procedure

We used *Gorilla* to run the experiment online [21]. Participants were assigned randomly to one of the four lists. In each trial, participants were informed on-screen that in the sentence production pair they were going to hear, the productions might sound different in the speed of the articulation, the precision of the articulation, both speed and precision, or neither (i.e. the same). They were asked to decide which of these four descriptions seemed most appropriate to them given the following audio. The production pair then played twice with a 2-second between-pair silence, before the four response options 'neither', 'speed', 'precision' and 'both' appeared on screen. The next trial began once the participant had submitted a response, with a 0.5-second between-trial silence. Trial order was randomised by participant.

2.4. Predictions

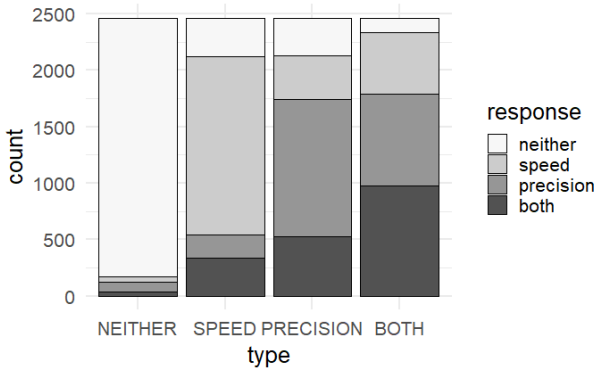
The hypothesis that speech mode variation has an impact on speech tempo perception when articulation rate is controlled yields several specific predictions. We expected participants to be less accurate at identifying PRECISION pairs ('precision') than at identifying SPEED pairs ('speed'). More specifically, we expected PRECISION pairs to attract more 'speed' and 'both' responses than SPEED pairs attract 'precision' and 'both' responses—indicating that

speaking mode differences trigger tempo percepts more than the other way around. We also expected PRECISION pairs to attract more 'speed' and 'both' responses than NEITHER pairs—indicating that participants hear precision differences as tempo differences, rather than simply producing occasional false alarms when no rate difference is present.

2.5. Results

Figure 1 shows the response proportions by stimulus type. Listeners are very good at identifying NEITHER pairs (93% 'neither'). For SPEED pairs, 'speed' is also the majority response (64% 'speed'). For PRECISION pairs, 'precision' is the majority response (50% 'precision'), but this majority is significantly smaller than that for SPEED pairs (χ^2= 108.64, df=1, p<0.001).

Figure 1: Cumulative bar chart of the Experiment 1 response proportions by stimulus type.



As predicted, PRECISION pairs attract significantly more 'speed' and 'both' responses than SPEED pairs attract 'precision' and 'both' responses (37% vs. 22%; χ^2=131.2, df=1, p<0.001). This is partly because the number of 'speed' responses for PRECISION pairs is almost twice that of 'precision' responses to SPEED (16%~8%). PRECISION pairs also attract significantly more 'speed' and 'both' responses than NEITHER pairs (37% vs. 4%; χ^2=843.60, df=1, p<0.001). Participants are least accurate at identifying BOTH pairs (40%): a majority of responses suggests participants were hearing difference in one parameter only (33% 'precision', 22% 'speed').

Having shown that speaking mode variation is heard as tempo variation more often than vice versa, we tested the direction of the effect in Experiment 2: is clear speech heard as faster, or slower, than normal speech with the same articulation rate?

3. EXPERIMENT 2

3.1. Participants

26 native British English speakers aged 18–35 years were recruited primarily from student cohorts. None had participated in Experiment 1. All reported normal hearing and passed the same screening task as in Experiment 1. Most were paid.

3.2. Stimuli, task and procedure

Experiment 2 was identical to Experiment 1 except for the on-screen instructions and response options. In each trial (N=120), participants were informed that the pair of productions might sound the same or different in tempo. They were asked to identify the faster pair member ('first', 'second', 'neither').

3.3. Predictions

Our key prediction concerned PRECISION pairs, in which articulation rate was identical but speaking mode differed. We expected participants to perceive a tempo difference in at least the same percentage of PRECISION trials as observed in Experiment 1 (37%), and we expected that when a tempo difference was heard, the (compressed) clear production would be heard as faster. Beyond this, we expected participants to veridically report articulation rate differences where these were present. That is, in SPEED pairs we expected the member with the higher articulation rate to be heard as faster, and in NEITHER pairs we expected the members to be heard as the same. In BOTH pairs, where uncompressed clear and normal productions were presented, we expected participants to accurately identify the clear pair member as slower.

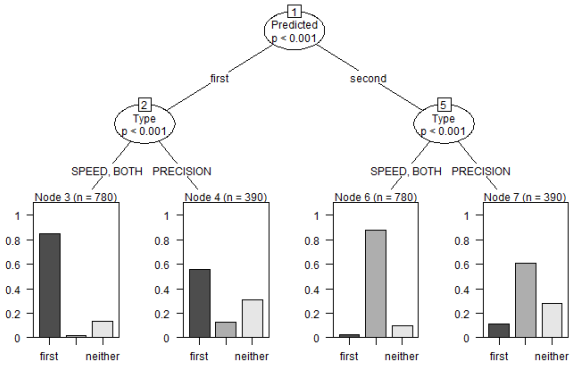
3.4. Results

Consistent with our prediction for PRECISION pairs, 71% of responses reflect a tempo difference; this is almost twice as many as in Experiment 1 (37%). The responses are split near-evenly as to which pair member ('first' or 'second') was heard as faster. While in these pairs, both productions have the same duration, 'neither' is the minority response (29%). For SPEED and BOTH pairs, 88% and 90% of responses identify a tempo difference, again split evenly between 'first' and 'second'. As expected, listeners accurately identified NEITHER pairs (96% 'neither').

To examine the responses to PRECISION, SPEED and BOTH pairs further we created a variable to reflect our predictions as to which pair member should be heard as faster: the clear production for PRECISION pairs and the higher-rate production for SPEED and BOTH pairs. We fitted a mixed ordinal regression model (random intercept for participant) [22] and a conditional inference regression tree model [23] to the responses with this variable (*Predicted*) as well as the pair type (*Type*: 'PRECISION', 'SPEED' or 'BOTH') as predictors. Both models revealed the same effects; we visualize them using the tree model in Figure 2.

The tree algorithm iteratively implements binary data splits according to the strongest predictor for the relevant data subset. The first split is at the top of the tree (node 1). In Figure 2, this split is for *Predicted*, such that 'predicted first' maps to a clear majority of 'first' responses and 'predicted second' to a clear majority of 'second' responses across PRECISION, SPEED and BOTH pairs. Within each of the two subsets of responses, *Type* gives rise to a further split (nodes 2, 5) which shows that responses to PRECISION pairs are significantly different from those to SPEED and BOTH pairs. The bar plots suggest that this is partly due to the greater proportion of 'neither' responses. Moreover, while for SPEED and BOTH pairs (nodes 3, 6) nearly all 'first' and 'second' responses are in the predicted direction, for PRECISION pairs (nodes 4, 7) the majority is smaller: about 10% of responses are in the opposite direction. Still, the modelling confirms that responses to PRECISION pairs show a significant listener preference for hearing the compressed clear sentence productions as relatively fast.

Figure 2: Regression tree for the Experiment 2 responses to PRECISION, SPEED and BOTH pairs; predictors explained in the text.



4. DISCUSSION

We have reported on two experiments that address the hypothesis that clear, hyper-articulated speech sounds faster than normal speech when articulation rates are controlled. The Experiment 1 results show that listeners are able to separate the parameters of 'speed' and 'precision' reasonably well when encouraged to do so. However, the fact that responses to PRECISION and BOTH pairs were more variable (and less accurate) than responses to SPEED pairs is consistent with the notion that speaking mode variation triggers tempo variation percepts. The Experiment 2 results show that when listeners assess tempo only, speaking mode clearly influences judgements: as predicted, clear speech sounds faster than normal speech when articulation rates are matched.

Our results show no support for the notion that listeners are biased by their knowledge of typical production patterns towards hearing clear speech as relatively slow [16, 17]. Participants in Experiment 2 showed remarkably robust sensitivity to the duration differences between PRECISION and BOTH pairs: when clear and normal productions differed naturally in duration, clear ones were heard as slower; when the duration difference was absent, clear productions were mostly heard as faster. This is not to dispute the possible perceptual relevance of listeners' knowledge of typical production patterns: listeners are acute at drawing on *any* relevant dimension of variation in making tempo judgements and may weight cue parameters depending on the task at hand [24]. The large difference between Experiments 1 and 2 in the proportions of responses to PRECISION pairs which reflect a perceived tempo difference demonstrates this: participants responded differently to the same stimuli as the details of their tasks were different.

Of course we cannot tell exactly what acoustic parameter(s) participants were weighting highly in Experiment 2. The clear productions were spectrally more complex in the sense of [4], but it remains an open question as to whether this is the best way to characterize the crucial perceptual dimension(s). Note that the increased spectral complexity in clear speech enhances intelligibility [7, 25]. Studies have shown that listeners judge speech as fast in conditions of high cognitive load [26] and low intelligibility [27]. In these contexts, 'complexity' is of a different type, posing a perceptual challenge. While a unified account for all of these effects on perceived tempo is clearly desirable, we cannot propose one at present.

Interestingly, one Experiment 2 participant observed after debriefing that in some of the PRECISION pairs, some specific segment transitions sounded fast. We assume this relates to the low degree of coarticulation in the compressed clear productions; we also cannot rule out some impact of our linear compression method on perceived segment-to-segment timing. Another participant noted that some sentences sounded 'urgent', which made them sound fast. It would appear, then, that both detailed utterance-internal timing patterns and more holistic interpretations of speaker intentions may feed into listeners' tempo judgements. These observations should inform further experimental work.

5. ACKNOWLEDGEMENTS

This research was made possible by a Leverhulme Trust Research Grant (RPG-2017-060).

REFERENCES

[1] Kohler, K. J. 1986. Parameters of speech rate perception in German words and sentences:

Duration, f0 movement, and f0 level. *Language and Speech* 29, 115–139.
[2] Rietveld, A. C. M., Gussenhoven, C. 1987. Perceived speech rate and intonation. *Journal of Phonetics* 15, 273–285.
[3] Feldstein, S., Bond, R. N. 1981. Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech* 24, 387–394.
[4] Weirich, M., Simpson, A. P. 2014. Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics* 43, 1–10.
[5] Lindblom, B., 1990, Explaining phonetic variation: A sketch of the H & H theory, in *Speech production and speech modelling*, ed. W. J. Hardcastle and A. Marchal, Amsterdam: Kluwer Academic, 403–439.
[6] Lindblom, B. 1996. Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America* 99, 1683–1692.
[7] Smiljanić, R., Bradlow, A. R. 1999. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass* 3, 236–264.
[8] Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., Brungart, D. 2018. Clear speech adaptations in spontaneous speech produced by young and older adults. *Journal of the Acoustical Society of America* 144, 1331–1346.
[9] Krause, J. C., Braida, L. D. 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America* 115, 362–378.
[10] Searl, J., Evitts, P. M. 2013. Tongue-palate contact pressure, oral air pressure, and acoustics of clear speech. *Journal of Speech, Language, and Hearing Research* 56, 826–839.
[11] Bradlow, A. R., Torretta, G. M., Pisoni, D. B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20, 255–272.
[12] Smiljanić, R., Bradlow, A. R. 2008. Temporal organization of English clear and conversational speech. *Journal of the Acoustical Society of America* 124, 3171–3182.
[13] Picheny, M. A., Durlach, N. I., Braida, L. D. 1986. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research* 29, 434–446.
[14] Hazan, V., Baker, R. 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America* 130, 2139–2152.
[15] Matthies, M., Perrier, P., Perkell, J. S., Zandipour, M. 2001. Variation in anticipatory coarticulation with changes in clarity and rate. *Journal of Speech, Language, and Hearing Research* 44, 340-353.
[16] Koreman, J. 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* 119, 582–596.
[17] Reinisch, E. 2016. Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception and Psychophysics* 9, 9–23.
[18] Bradlow, A. R. n.d. SpeechBox. Retrieved from https://speechbox.linguistics.northwestern.edu.
[19] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
[20] Boersma, P., Weenink, D. 2017. Praat: Doing phonetics by computer. Version 6.0.25. Retrieved from www.praat.org.
[21] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52, 388–407.
[22] Bürkner, P.-C., Vuorre, M. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science* 2, 77–101.
[23] Tagliamonte, S.A., Baayen, R.H. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24**,** 135–178.
[24] Plug, L., Lennon, R., Smith, R. 2022. Measured and perceived speech tempo: Comparing canonical and surface articulation rates. *Journal of Phonetics* 95, 101–193.
[25] Smiljanic, R., Gilbert, R. C. 2017. Intelligibility of noise-adapted and clear speech in child, young adult, and older adult talkers. *Journal of Speech, Language, and Hearing Research* 60, 3069–3080.
[26] Bosker, H. R., Reinisch, E., Sjerps, M. J. 2017. Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language* 94, 166–176.
[27] Bosker, H. R., Reinisch, E. 2017. Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology* 8.

EFFECTS OF FFP2 FACE MASKS ON CONSONANT IDENTIFICATION

Angelos Lengeris & Ioanna Zatrazemi

National and Kapodistrian University of Athens
alengeri@phil.uoa.gr, joanna.zatrazemi@gmail.com

ABSTRACT

The COVID-19 pandemic has led most countries to recommend or mandate the use of face masks in public places. Although masks have been shown to reduce virus transmission, they also affect communication by attenuating speech sounds and concealing speakers' facial expressions and lip movements. Previous research examining adverse effects of face masks on speech perception has mainly focused on global measures of intelligibility. This study examined the effects of wearing face masks on consonant identification. Greek listeners identified 20 Greek consonants (embedded in VCV tokens) uttered with and without FFP2 face masks. The results showed that identification was significantly lower for consonants uttered with face masks than for consonants uttered without face masks, with female listeners outperforming male listeners in both tasks. Across genders and tasks, voicing was the most salient perceptual feature, place of articulation was the least salient feature and manner of articulation was in between.

Keywords: Identification, consonants, FFP2 face masks.

1. INTRODUCTION

Successful communication requires perceiving speech in various listening environments. Under ideal conditions and for adults with normal hearing, speech perception in one's native language generally approaches ceiling performance (although there are cases where two phonemes may be confused with one another, e.g., /f/ and /θ/ for English listeners, see [1]; /v/ and /ð/ for Greek listeners, see [2]). When communication takes place in adverse listening conditions, however, listeners often have difficulty in understanding speech, as shown by numerous studies comparing different types of maskers, signal-to-noise ratios (SNRs), and language backgrounds, among others [3, 4, 5].

The COVID-19 pandemic has added yet another difficulty in communication as most countries have recommended or mandated the use of face masks in public places to reduce virus transmission [6]. Wearing a face mask alters the acoustic features of speech and conceals speakers' facial expressions and lip movements resulting in decreased speech intelligibility, depending on mask type (cloth, surgical, respirator) [7, 8, 9, 10]. [7] for example, evaluated different mask types showing that most masks had little acoustic effect below 1 kHz, but they attenuated higher frequencies by different amounts depending on the fabric, shape and fit of the mask.

Previous research examining adverse effects of wearing face masks on speech perception has mainly focused on global measures of intelligibility at word or sentence level. This study examined the effects of face masks on consonant identification thus removing any word frequency effects. Greek listeners identified 20 Greek consonants embedded in aCa tokens uttered with and without FFP2 face masks. FFP2 masks, widely used in Europe, filter at least 94% of all aerosols, including COVID-19, offering similar protection to America's N95 and China's KN95 masks [11]. The study complements work in [2], examining the identification of 20 Greek consonants in quiet and in two types of background noise, a competing talker at an SNR of -6 dB and an 8-speaker babble at an SNR of -2 dB. The results in [2] showed that Greek listeners achieved excellent results in quiet across consonants except for fricatives θ/ and /ð/. Consonant identification was significantly lower in the two noise conditions, with the 8-speaker babble having a larger deteriorating effect in listeners' identification compared to the competing speaker (despite the difference in SNR). Across listening conditions, voicing was more resistant to noise than manner of articulation, which was in turn more resistant to noise than place of articulation (for similar results, see [12, 13]).

The goals of the current study were therefore to (a) examine the effects of wearing face masks on consonant identification (b) pinpoint the most difficult consonants for Greek listeners, especially in the masked condition and (c) compare the salience of voicing, place, and manner of articulation. Both male and female listeners were tested because of gender-related anatomical and functional differences of the brain and the auditory system [14] that could be affecting performance. For example, females generally outperform males in verbal fluency, perceptual speech, and fine motor skills while males are better in spatial and working memory [15, 16, 17].

2. METHOD

2.1. Participants

Participants were 14 native speakers of Greek (7 m, 7 f) with a mean age of 28 years (age range 19 to 45 years). They all reported normal hearing and no language impairment.

2.2. Stimuli

Stimuli consisted of aCa tokens (stressed on the second syllable) containing the stops /p/, /b/, /t/, /d/, /k/, /g/, the fricatives /f/, /v/, /θ/, /ð/, /s/, /z/, /x/, /ɣ/, the nasals /m/ and /n/, the liquids /l/ and /ɾ/, and the affricates /ts/ and /dz/. Stimuli were recorded by four native speakers of Greek (2 m, 2 f) with and without wearing FFP2 masks (MASKED vs. UNMASKED condition, respectively) for a total of 160 stimuli (20 consonants × 4 speakers × 2 conditions).

2.3. Procedure

Testing was administered in TP software [18]. Following each stimulus presentation, participants indicated which consonant they heard by clicking on a computer screen one of 20 consonant options written in Greek orthography.

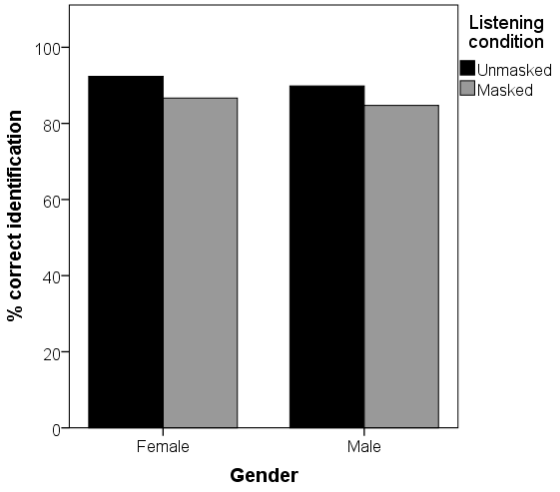


Figure 1: Mean percent correct identification across Greek consonants by female and male listeners in UNMASKED and MASKED condition.

3. RESULTS

Figure 1 shows percent correct identification scores for female and male listeners in each condition. Both groups achieved higher scores in the UNMASKED condition compared to the MASKED condition. For female listeners, mean identification dropped from 92.4% to 86.7% correct, respectively. For male

	Response																			
Stim.	p	b	t	d	k	g	ts	dz	f	v	θ	ð	s	z	x	ɣ	m	n	l	ɾ
p	**97**	2	1																	
b		**97**										3								
t	3	3	**81**							3		3			3	3				
d				**73**		9				3		15								
k			3		**97**															
g		3				**85**										12				
ts							**90**	7			3									
dz							3	**94**								3				
f									**60**		34		6							
v										**85**	3	12								
θ									3	3	**94**									
ð				3						6	6	**84**								
s													**100**							
z														**100**						
x															**100**					
ɣ						6										**94**				
m																	**100**			
n																		**100**		
l																			**100**	
ɾ																2			3	**95**

Table 1: Greek consonant confusion patterns in UNMASKED condition. Percentages have been pooled over participants and genders. Responses less than 3% are not shown.

	Response																			
Stim.	p	b	t	d	k	g	ts	dz	f	v	θ	ð	s	z	x	ɣ	m	n	l	ɾ
p	**97**		3																	
b	3	**94**										3								
t	13		**78**		9															
d				**71**								29								
k			3		**94**					3										
g						**70**						2				28				
ts							**94**	6												
dz							9	**91**												
f				3					**44**		53									
v										**81**		19								
θ				13					19		**69**									
ð				3						6	16	**63**				9				
s													**100**							
z														**100**						
x															**100**					
ɣ																**97**				3
m																	**94**	6		
n			3															**94**	3	
l																			**100**	
ɾ																9				**91**

Table 2: Greek consonant confusion patterns in MASKED condition. Percentages have been pooled over participants and genders. Responses less than 3% are not shown.

listeners, mean identification dropped from 89.8% to 84.7% correct, respectively. A mixed ANOVA on identification scores with Gender as between-subjects factor and Condition as within-subjects factor showed a significant main effect of Condition, F (1, 12) = 121.6, $p < .001$, confirming that the use of FFP2 face masks lowered consonant identification. Female listeners outperformed male listeners across conditions, but the differences between the two groups did not reach significance nor did the group × condition interaction (the small sample size of the study may have masked the effect).

Confusion matrices in the two conditions are shown in Tables 1 and 2 (pooled over genders). Identification scores in UNMASKED condition (Table 1) ranged from 60% to 100% correct. The most problematic consonant was /f/ (confused with /θ/ 34% of time), followed by /d/, /t/, /ð/, /g/, and /v/ (all below 90% correct). Identification scores in MASKED condition (Table 2) ranged from 44% to 100% correct. The most problematic consonant was, again, /f/ (confused with /θ/ 53% of time), followed by /θ/ and /ð/ (both below 70% correct).

Table 3 shows percent correct identification of Greek consonants as a function of the three dimensions along which consonants are characterized, namely voicing, place of articulation, and manner of articulation. These were coded as

	Dimensions		
Condition	Voicing	Place-of-articulation	Manner-of articulation
UNMASKED	97.7	92.5	95.8
MASKED	97.5	90.9	95.3

Table 3: Mean percent correct identification for voicing, place-of-articulation, and manner-of-articulation in two conditions pooled over genders.

follows: Voicing had two values, voiced and voiceless. Place of articulation had four values, labial, dental, alveolar, and velar. Manner of articulation had five values, plosive, fricative, affricate, liquid, and nasal. Across genders and conditions, voicing was the most salient perceptual feature, place of articulation was the least salient feature and manner of articulation was in between.

4. DISCUSSION

This study examined the effects of wearing face masks on consonant identification. Greek listeners identified 20 Greek consonants uttered with and without FFP2 face masks. The stimuli were aCa tokens so that listeners could not make use of lexical information when identifying consonants. Results showed that wearing face masks significantly lowered identification. Across conditions, fricatives were found to be the most problematic consonants to

identify. This is is consistent with previous research in Greek and other languages showing that fricatives are generally difficult to perceive [1, 2]. Across genders and tasks, voicing was the most salient perceptual feature, place of articulation was the least salient feature and manner of articulation was in between [12, 13].

Female listeners outperformed male listeners in both conditions, but differences did not reach significance. Given the anatomical and functional differences between males and females [14] and the superiority of females in speech perception tasks reported in some studies (e.g., [17]), it would not be surprising if differences reached significance using a larger sample size (but note that there are studies in the literature reporting no differences in performance between males and females, e.g., [19]).

A 5% mask-induced decrease in mean identification accuracy is not expected to create much difficulty to normal-hearing individuals in otherwise ideal listening conditions. The decrease is larger, however, in the presence of background noise [10], when noise is combined with poor room acoustics (e.g., in school classrooms [20]) and when listeners are D/deaf or hard of hearing [21]. In those situations, clear speech production and use of transparent face masks that do not block visual information can significantly improve communication [10].

6. ACKNOWLEDGMENTS

This work has been supported by a travel grant from the National and Kapodistrian University of Athens.

7. REFERENCES

[1] Jongman, A. 1989. Duration of frication noise required for identification of English fricatives. *J. Acoust. Soc. Am*. 85, 1718–1725.

[2] Lengeris, A., Nicolaidis, K. 2014. Greek consonant confusions by native listeners in quiet and noise. *Proc. 11th International Conference on Greek Linguistics*, Rhodes, Greece, 866–873.

[3] Benkí, J. R. 2003. Analysis of English nonsense syllable recognition in noise. *Phonetica* 60, 129–157.

[4] Garcia Lecumberri, M. L, Cooke, M. 2006. Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am*. 119, 2445–2454.

[5] Wang. X., Xu, L. 2021. Speech perception in noise: Masking and unmasking. *Journal of Otology* 16(2), 109–119.

[6] Liang, M., Gao, L., Cheng, C., Zhou, Q., Uy, J. P., Heiner, K., Sun, C. 2020. Efficacy of face mask in preventing respiratory virus transmission: a systematic review and meta-analysis. *Travel Medicine and Infectious Disease* 36, 101751.

[7] Corey, R. M., Jones, U., Singer, A. C. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am*. 148, 2371–2375.

[8] Rahne, T., Fröhlich, L., Plontke, S., Wagner, L. 2021. Influence of surgical and N95 face masks on speech perception and listening effort in noise. *PLoS One 16:e0253874*.

[9] Bernstein, J. G., Venezia, J. H., Grant, K. W. 2020. Auditory and auditory- visual frequency-band importance functions for consonant recognition. *J. Acoust. Soc. Am*. 147, 3712–3727.

[10] Yi, H., Pingsterhaus, A., Song, W. 2021. Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Frontiers in Psychology* 12:682677.

[11] Li, K. K., Joussen, A. M, Kwan, J. K, Steel, D. 2020. FFP3, FFP2, N95, surgical masks and respirators: what should we be wearing for ophthalmic surgery in the COVID-19 pandemic? *Graefe's Archive for Clinical and Experimental Ophthalmology* 258(8), 1587–1589.

[12] Wang, M. D., Bilger, R. C. 1973. Consonant confusions in noise: a study of perceptual features. *J. Acoust. Soc. Am*. 54, 1248–1266.

[13] Phatak, S., Allen, J. B. 2007. Consonant and vowel confusions in speech-weighted noise. *J. Acoust. Soc. Am*. 121, 2312–2326.

[14] Ruigrok, A. N., Salimi-Khorshidi, G., Lai, M. C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., Suckling, J. 2014. A meta-analysis of sex differences in human brain structure. *Neuroscience and Biohehavioural Reviews* 39, 34–50.

[15] de Lima Xavier, L., Hanekamp, S., Simonyan, K. 2019. Sexual dimorphism within brain regions controlling speech production. *Frontiers in Neuroscience* 13(795), 1–7.

[16] Zaidi, Z. F. 2010. Gender Differences in Human Brain: A Review. *The Open Anatomy Journal* 2, 37–55.

[17] Watson, C. S., Qiu, W. W., Chamberlain, M. M., Li, X. 1996. Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *J. Acoust. Soc. Am*. 100, 1153–1162.

[18] Rauber, A., Rato, A., Kluge, D., Santos, G. 2011. TP-S [Application software]. http://www.worken.com.br/tp_regfree.php.

[19] Baxter L., Saykin A., Flashman L., Johnson S., Guerin S., Babcock D., et al. 2003. Sex differences in semantic language processing: a functional MRI study. *Brain and Language* 84, 264–272.

[20] Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., Kirk, K. I. 2020. Effect of masks on speech intelligibility in auralized classrooms. *J. Acoust. Soc. Am*. 148, 2878–2884.

[21] Saunders, G. H., Jackson, I. R., Visram, A. S. 2021. Impacts of face coverings on communication: An indirect impact of COVID-19. *International Journal of Audiology* 60, 495–506.

PERCEPTION OF STANDARD CROATIAN PITCH ACCENTS BY SPEAKERS FROM PITCH ACCENT AND FROM STRESS ACCENT REGIOLECTS

Elenmari Pletikos Olof, Jelena Vlašić Duić and Jordan Bićanić

University of Zagreb, Croatia
epletikos@ffzg.hr, jvduic@ffzg.hr, jbicanic@ffzg.hr

ABSTRACT

Previous acoustic studies have found that Croatian speakers use either standard four pitch-accent system (PA) or a stress accent (SA). The focus of the present study is on perception of four pitch accents: we tested how fast and how accurately the Croatian listeners from pitch and from stress system can perceive the words with correct and incorrect pitch accents, in total 80 tokens were tested. The results show that there is a significant difference in perception between listeners from the PA and listeners from the SA. PA listeners (N=67) recognise frequent words with an average reaction time of 1417 ms and with an average accuracy of 87.89 %, whereas SA listeners (N=90) react more slowly, with an average reaction time of 1625 ms and with an accuracy of 69.97 %. This suggests that pitch accents make word recognition difficult for the speakers who are from the stress accent regiolect.

Keywords: pitch-accent, stress-accent, perception, Croatian, word recognition

1. INTRODUCTION

Croatian is a South Slavic Language with the pitch-accent system containing four pitch accents and short or long post-accentual syllables [1, 2, 3]. Names for accents partly reflect their acoustics and they are called short falling (SF) e. g. [kíʃa] kȉša ‘rain’); long falling (LF), e.g. [sú:ntse] sûnce ‘sun’); short rising (SR) e.g. [rósá] ròsa ‘dew’); long rising (LR), ([trá:ʊá] tráva ‘grass’). Since IPA offers different opportunities for tone marking, we will use the possibility of marking the high tone of "rising" accents on both the accented and post-accented vowels, since there is no rise, but a high tone is spreading. Post-accentual length (L) is also marked in dictionaries and grammars, e.g. [mjése:ts] mjȅsēc ‘moon’. All Croatian dictionaries, both monolingual and bilingual, mark the four tones and long post-accentual vowels, although they are not commonly marked in standard orthography. Croatian pitch accents, when compared to some other Indo-European languages with pitch accents, are similar to those in Slovenian and Lithuanian, but differ from those in Swedish [4]. In contrast to tone languages, all pitch languages share a small phonological distinction in the distinctive role of pitch accent, with minimal pairs occurring infrequently in the same linguistic context, and with spoken dialects varying from pitch to stress accent [5]. While Croatian Standard is described with four pitch accents, fieldwork involving acoustically and/or perceptually analysed recordings of many Croatian speakers with the task of reading standard-written sentences revealed that less than half of Croatian speakers use all four accents [6, 7], and the use of post-accentual length is also rare [5, 8]. The four accents are mostly used in the Štokavian area in eastern Croatia (mainly Slavonia) and in Štokavian and Čakavian southern Croatia (Dalmatia), while in the Kajkavian area (north and central Croatia, including the capital Zagreb) and in the Čakavian northwest (Istria and Kvarner), most inhabitants use the stress accent system. Some Kajkavian and Čakavian dialects have a rising monosyllabic pitch accent called the "acute" [9], but this accent is used only in dialects and not in the standard variety spoken in towns. To refer to different standard varieties, we will use the term "regiolect", a colloquial variety of a standard language spoken in a particular region or town and influenced by the native dialect. To preserve the standard four pitch accents and post-accentual length, the teaching of four pitch accents is an inevitable part of study programs in Croatian language, phonetics, and acting [10]. It is also taught in elementary and high school [11, 12].

2. PREVIOUS WORK

Previous studies on the perception of Croatian pitch accents, as well as Serbo-Croatian which includes Štokavian pitch accents spoken in Serbian and Bosnian [13, 14, 15, 16], have investigated specific acoustic correlates (such as duration or tonal contour) and their effect on the perception of certain accent types, which accents and post-accentual lengths can be identified correctly, and which accents are preferred. Research on duration perception in monosyllabic words [17] suggests that the boundary between long falling (LF) and short rising (SF) accents occurs at 147 ms, and in another study [18], at 146 ms. In two-syllabic words [19], the boundary

is at 118 ms, with durations of 108 ms perceived as belonging to the correct SF category, and durations of 128 ms perceived as too long. There has also been some research on lexical tone perception [20, 21, 22, 23]. Study [21] demonstrates that the F0 peak, which occurs in the last 24% of the duration of the stressed vowel, serves as a cue for the perception of rising accents. Meanwhile, study [22] shows that the later position of the F0 peak and the larger rising range of the initial and final F0 in the stressed syllable determine the perception of a long rising accent. However, study [23] strongly criticizes these results and argues that the most important acoustic information for the perception of rising accents is not being considered – namely, the F0 on the post-stressed syllable. As a result, study [23] shows that a speaker who can identify their own accents with 100% accuracy cannot differentiate between a short rising and a short falling accent if the post-stressed syllable, including the voiced sound that precedes the vowel, is "cut off". The authors conclude that the first part of the contour is less informative than the second, which is supported by studies [24, 25] that show that tonal changes are not perceived if they occur at the beginning of the vowel, but rather at the transition from 2/3 to 3/3 of the vowel duration.

Other research [26] tested over 1600 speakers in 20 cities in Yugoslavia on the perception of minimal pairs. The results show that the tone difference on long vowels is maintained in some cities for some minimal pairs. For example, the difference between Lûka – lúka is perceived in Dubrovnik, and the difference between râdio – rádio is perceived in Osijek and Split but not in Dubrovnik. However, the difference between SR and SF on the first syllable of a word is poorly recognized in all Croatian cities.

Another study [27] tested accent recognition as metalinguistic ability and the results show that the accuracy of recognition of stressed syllables is 75%, and the accuracy for different pitch accents is 86.03% for LF, 65.52% for LR, 80.67% for SF and 61.25% for SR. Recognition of the same categories with pseudowords is 1 to 5% less than with existing words. Other results [18] show very similar results: stressed syllables are perceived with accuracy 77.50%, and pitch accent type in common words 72.58%, and in pseudowords 59.47%. There is also a big difference regarding the type of accent: LR 63.06%, SR 57.67%, LF 76.41%, SF 63.78%. The difference between [10] and [26] is in the recognition of LR and SF, while both papers give evidence that LF is recognised most accurately, and SR less accurately.

3. RESEARCH QUESTIONS

n this research, we attempt to answer the following research questions (RQ):

RQ1: Which accents in Croatian contribute to faster word processing, and which ones slow it down?

RQ2: Which word accents in Croatian are perceived with greater accuracy, and which with less?

RQ3: What is the difference in the processing of Croatian standard pitch accents between speakers/listeners from different Croatian regiolects, those with and without PA?

4. METHODOLOGY

4.1. Participants

200 native Croatian speakers (age range 18-54; mean 24.07) participated in a pitch-accent perception experiment. 82.6% of the participants were female, 16.4% were male, and 1% did not answer. Most participants had either completed their studies or were currently pursuing higher education, with 35% studying phonetics, 16% psychology, 14% linguistics, 11% Croatian, and 24% in other fields. Based on the data obtained from a survey about the participants' place of longest residence, place of elementary and high school education, and the origin of parents or guardians, the participants were classified into three groups for further analysis, based on the accent system that exists in their regiolects: the pitch accent system PA (N=67), accounting for 33.5%; the stress accent system SA (90), accounting for 45%; and the mixed system M (N=43), accounting for 21.5%. The mixed system includes individuals whose regiolect of at least one parent is different from the regiolect in which the speaker grew up and was educated, or who lived in two different regiolects with pitch and stress accent systems during primary and high school (years 6 to 18). To ensure that the data about the speaker's place of residence reliably points to the regiolect and accent system that listeners use in speech, one part of the participants (N=58) was tested by the authors for pitch accent production by reading sentences with target words containing the standard four pitch accents. The results of classification into three groups (PA, SA, M) based on data on residence and data obtained from production tests matched 84.48%. In cases where the data did not match (15.52%), the speakers were mostly classified by production assessment into PA or SA, and by data from the questionnaire, they were classified into a mixed system. Since the results based on the questionnaire coincide to a significant extent with the production testing, we can consider them reliable for further analysis and interpretation.

4.2. Stimuli

For stimuli, we used 20 accentually indisputable, common, disyllabic Croatian words with an initial accent, which were balanced in the accented vowel /a e i o u/. These words were selected during pre-testing where 74 participants were presented with a word displayed on the screen. The participants had to make a lexical decision as to whether the word was known or not. Frequent words that were identified as known with a recognition rate of over 99% and an average reaction time of 694 ms were selected. Rare words that were recognized as known by an average of 21% and had an average reaction time of 1408 ms were discarded.

Selected words with LF are 'škola' (school), 'sunce' (sun), 'meso' (meat), 'tajna' (secret), 'biljka' (plant); with LR 'glava' (head), 'vino' (wine), 'ruka' (hand), 'torba' (bag), 'leđa' (back); with SF 'kuća' (house), 'kiša' (rain), 'nebo' (sky), 'vatra' (fire), 'soba' (room); and with SR 'žena' (woman), 'voda' (water), 'magla' (fog), 'život' (life), 'bubreg' (kidney). All words were pronounced by phonetician Jordan Bićanić and recorded in an acoustic studio with all four pitch accents, one correct (C) and three false (F) accents. The total number of produced tokens was 80. All stimuli were verified by accentologist Blaženka Martinović. Figure 1 illustrates the tone movement and duration in four accents produced on the same word. The figure of tonal movement were generated using Praat [28].

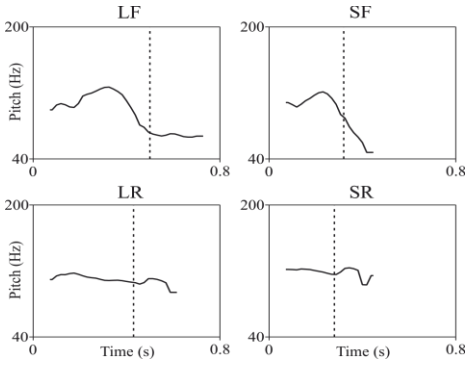


Figure 1: Pitch movement in the word 'glava' (head) produced with four pitch accents: LR is correct, other three are wrong; dotted line marks the syllable boundary

4.3. Procedure

The participants in this experiment took part through the online platform SoSci Survey. The questionnaire collected data on the participants' age, gender, place of origin, etc. It included practice tasks and experimental tasks. The acoustic stimulus was presented with a visual stimulus (a word) on the screen. The presentation of the stimuli was randomized, and each stimulus was presented once. Participants were explicitly asked to judge the correctness of the pitch accents of standard Croatian by forced choice using the keyboard (for the answer "correct," key "A," and for the answer "incorrect," key "L"). The responses about correctness, as well as the reaction time, were recorded. The reaction time was measured from the beginning of the token, and the duration of words varied between 626 and 689 ms.

5. RESULTS

5.1. Perception of correctness

When analysing the proportion of correct responses for all words pronounced with a particular accent (of which 1/4 have correct and 3/4 have false pitch accents), the data indicate that the highest recognition accuracy is for long accents (LF and LR), and significantly lower for short accents (SF and SR) (Figure 2). The overall analysis shows that PA listeners recognise frequent words with an average accuracy of 87.89%, whereas SA listeners recognise them with significantly lower accuracy of 69.97% ($Z=23.72$, $p<0.01$).

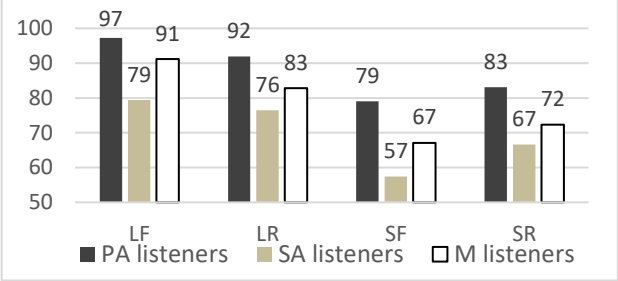


Figure 2: Results for all tokens: proportion of correct recognition (%) by listeners from PA, SA and M systems

The results, divided into two groups of words pronounced with correct and false accents (Figure 3), show a significant difference in recognition accuracy between different accent types. In both systems, for correctly pronounced words, the short falling accent is the most accurately recognised with 99.4% accuracy in the PA and 96% in the SA. The long falling accent follows in second place with 98.8% accuracy in the PA and 92% in the SA. The rising accents are the most difficult to recognise, with the long rising accent having the lowest score of 83% in the PA and 52% in the SA. The results for the short rising accent are 89% in the PA and 63% in the SA. Regarding the words pronounced with false pitch accents, the lowest recognition score is for the short falling accent with 72% by PA listeners and 45% by

SA listeners. This finding indicates that SF was not recognised as false on words where standard Croatian has three other accents.

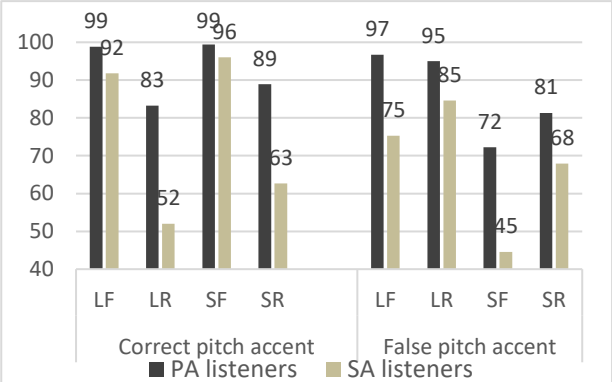


Figure 3: Proportion of correct recognition of words with correct and words with the false pitch accent: difference between PA and SA listeners

5.2. Reaction time

Analysis of the reaction time for all words spoken with a certain accent (with correct and incorrect pitch accents) shows that PA listeners recognize frequent words with an average reaction time of 1417 ms (sd 805) and SA listeners react more slowly, with an average reaction time of 1625 ms (sd 929) (t-test $p<0.01$). Meanwhile, results for M listeners are similar to those of PA listeners.

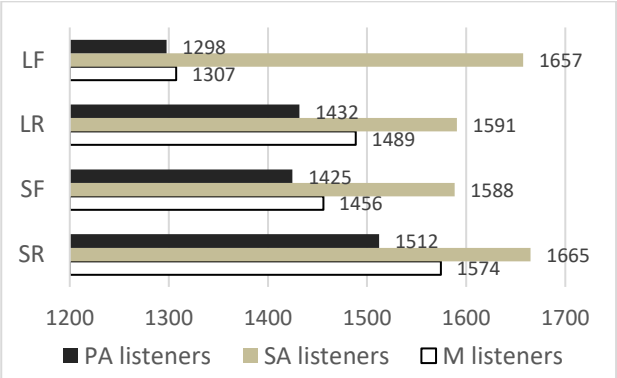


Figure 4: Results for all tokens: reaction time (ms) difference between listeners from PA, SA and M

When analysed by specific accent (see Figure 4), it can be concluded that SA listeners require a similar amount of time for all four accents, while PA listeners recognise LF accent very quickly, in only 1298ms. Recognition of LR and SF accents requires a similar amount of time, and SR accent recognition is slightly slower. When we distinguish between correctly and falsely pronounced accents, the reaction time data (Figure 5) give us interesting insight into processing: LR accent, as well as SR, take a long time to be processed on words where they are correct.

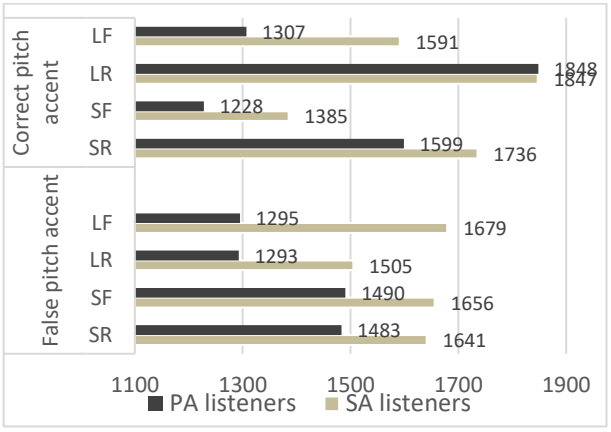


Figure 5: Reaction time for words with correct and words with the false pitch accent: difference between PA and SA listeners

6. DISSCUSSION

The fact that SF was not recognized as false, not only by SA listeners (which was expected) but also by PA listeners, suggests that the short falling accent (which is acoustically similar to stress accent [5]) strongly permeates standard Croatian and is also acceptable to PA speakers, as shown in previous research [29]. This is not unexpected because previous studies have shown that stress accent also exists in the speech of competent Croatian speakers (e.g., linguists, professors, actors) and if such SA occurs in the prescribed syllable in the word (as is the case in our research), listeners will still perceive it as standard [11]. The reaction time for short falling accent might suggest that SF sounds acceptable to all listeners, and it takes a long time to recognize it as an intruder.

7. CONCLUSION

Based on a significant sample size, our study has demonstrated that Croatian speakers who were raised and educated in areas with a stress accent system regiolect process and perceive the four "standard Croatian" pitch accents with considerably less accuracy and slower reaction times. These findings, along with previous research [2, 10, 11, 30, 31], suggest that the pitch accent system as the only variety in the standard language might be challenging and impractical. Instead, it may be more appropriate to accept the stress accent system as a part of the standard Croatian language. However, such a change would require a revision of the current accent norm.

8. ACKNOWLEDGEMENTS

We would like to express our gratitude to accentologist Blaženka Martinović for verifying the

accents of the stimuli, and fellow phoneticians Marko Liker and Ana Vidović Zorić for testing the online questionnaire and providing valuable feedback to improve the study. Special thanks to mathematician Barbara Brzić for conducting the statistical analysis of the data. Finally, we would like to thank all the participants who willingly took part in the experiment without any financial incentive.

9. REFERENCES

[1] Landau, E., Lončarić, M., Horga, D., Škarić, I. 1995. Croatian. *Journal of the International Phonetic Association,* 25(2), 83-86.

[2] Martinović, B., Pletikos Olof, E., Vlašić Duić, J. 2021. *Naglasak na naglasku*. Pula: Sveučilište Jurja Dobrile u Puli.

[3] Škarić, I. 2007. Fonetika hrvatskoga književnoga jezika. In: Babić, S., Brozović, D., Škarić, I., Težak, S., (Eds), *Glasovi i oblici hrvatskoga književnoga jezika.* Nakladni zavod Globus, 17–157.

[4] Pletikos, E. 2005. Percepcija sličnosti prozodije riječi stranih jezika s hrvatskim naglascima. *Govor* XXII/2, 89–126.

[5] Pletikos, E. 2008. Akustički opis hrvatske prozodije riječi. PhD thesis University of Zagreb.

[6] Pletikos Olof, E., Bradfield, J. 2019. Croatian pitch-accents: fact and fiction. *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia 2019. Calhoun, S., Escudero, P., Tabain, M., Warren, P. (Eds.), Canberra, Australia: Australasian Speech Science and Technology Association Inc., 855-858.

[7] Pletikos, E., Vlašić, J. 2007. Acoustic description of Croatian accents. *Studia Phonetica Posnaniensia* 8, 27–49.

[8] Rajle, L., Pletikos Olof, E., Martinović, B. 2020. Zanaglasne dužine u osječkom govoru. *Jezikoslovlje,* 21(1.), 53-80.

[9] Vlašić Duić, J., Pletikos Olof, E. 2014. Akustičke posebnosti akuta u čakavskom, kajkavskom i staroštokavskom govoru. *Sarajevski filološki susreti II: Zbornik radova,* 21–44.

[10] Pletikos Olof, E., Vlašić Duić, J, Martinović, B. 2016. Metode mjerenja naglasnih kompetencija. *Metodologija i primjena lingvističkih istraživanja. Zbornik radova s međunarodnoga znanstvenog skupa Hrvatskoga društva za primijenjenu lingvistiku.* Eds. Udier, S. L.; Cergol Kovačević, K. Srednja Europa. Zagreb, 293–307.

[11] Colnago, M., & Pletikos Olof, E. 2022. Uspješnost učenja mjesta naglaska u hrvatskome standardnome jeziku učenika osmih razreda osnovne škole u Zagrebu. *Govor,* 39(1), 39-72.

[12] Gusak Bilić, I., Alerić, M., Vlašić Duić, J. 2018. Teaching word stress patterns in primary school. *Croatian Journal of Education* 20 (Sp.Ed.2), 21–44.

[13] Ivić, P., Lehiste, I. 2002. *O srpskohrvatskim akcentima.* Izdavačka knjižarnica Zorana Stojanovića.

[14] Smiljanić, R. 2004. *Lexical, pragmatic and postitional effects on prosody in two dialects of Croatian and Serbian: an acoustic study.* Routledge.

[15] Sredojević, D. 2017. *Fonetsko-fonološki opis akcenata u standardnom srpskom jeziku – od specifičnog ka opštem.* Filozofski fakultet.

[16] Sredojević, D., Subotić, L. 2011. The neoŠtokavian accent shift and the phonological significance of suprasegmental features in different Štokavian dialects: Acoustic and phonetic analysis. *Baltistica Supplement* VII, 227–242.

[17] Mildner, V., Lisker, L. 1987. Acoustic vs. lexical judgements in the perception of falling accents in Serbo-Croatian: a preliminary study. *Proceedings of the XIth International Congress of Phonetic Science,* Tallin, 1, 324-321.

[18] Mildner, V. 1994. Perceptual acquisition of the long-short distinction in the falling accents of standard Croatian. *Language and Speech*, 37, 2, 163-170.

[19] Bakran, J. 1988. Tolerance of segmental duration variation of continuous speech. Speech '88. *Proceedings of Seventh FASE Symposium*, 22. - 26. August, 1988, Edinburgh, Book 2, 509-513.

[20] Purcell, E. T. 1971. The Acoustic Differentiation of Serbo-Croatian Accents in Statements. *Phonetica*, 24, 1-8.

[21] Purcell, E. T. 1976. Pitch peak location and the perception of Serbo-Croatian word tone. *Journal of Phonetics*, 4, 265-270.

[22] Purcell, E. T. 1981. Two parameters in the perception of Serbo-Croatian tone. *Journal of Phonetics*, 9, 189-196.

[23] Lehiste, I., Ivić, P. 1986. *Word and sentence prosody in Serbocroatian*. MIT Press.

[24] Rossi, M. 1971. Le seuil de glissando ou seuil de perception des variationa tonales pour les sons de la parole. *Phonetica,* 23, pp. 33.

[25] Rossi, M. 1978. La perception des glissandos descendats dans les contours prosodiques. *Phonetica,* 35, 11-40.

[26] Magner, T. F., Matejka, L. 1971. *Word accent in modern Serbo-Croatian*. University Park and London: The Pennsylvania State University Press

[27] Vrban Zrinski, K., Varošanec-Škarić, G. 2004. Slušno prepoznavanje hrvatskih naglasaka. *Govor*, XXI, 2, 93-110.

[28] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer.* [Computer program]. Version 6.3.03, retrieved 1 December 2022 from http://www.praat.org/

[29] Varošanec-Škarić, G., Škavić, Đ. 2001. Neutralizacija kratkouzlaznoga i kratkosilaznog naglaska u suvremenom hrvatskom prihvaćenom izgovoru. *Govor,* XVIII, 2, 87-104.

[30] Kapović, M. 2018. The unattainable standard – Zagreb dialect meets Standard Croatian accentuation. *Int. J. Slavic Studies* 7(1), 337–362.

[31] Magner, T. 1966. *Zagreb Kajkavian Dialect*. University Park: Pennsylvania State University.

SPEECH RATE EFFECTS IN FRENCH STOP VOICING PRODUCTION AND PERCEPTION*

Yoonjung Kang, Nicholas Haggarty

University of Toronto Scarborough, University of Toronto
yoonjung.kang@utoronnto.ca, n.haggarty@mail.utoronto.ca

ABSTRACT

Studies on speech rate variation have reported shortening of prevoicing and aspiration of stops in fast speech. Perception studies have also found that the boundary between short and long-lag VOT stops shifts to shorter VOTs when the speech rate increases in the carrier phrase. However, no previous study has explored how speech rate variation affects the perception of prevoiced stops. Our study aims to fill this gap by investigating the effects of speech rate variation on French stops, which contrast prevoiced and short-lag stops. We collected production and perception data from the same participants using the same speech materials. Thirty-one self-identified native speakers of French participated in an online experiment. The production results confirmed previous observations of shortening of prevoicing and aspiration in fast speech. The perception results showed an expected shift in perception for positive VOTs based on speech rate but not for negative VOTs.

Keywords: French, voicing, production, perception, speech rate

1. INTRODUCTION

Previous production studies have found that in fast speech compared to slow speech, the duration of prevoicing (negative VOTs) and aspiration (positive VOTs) shortens [1, 2]. Parallel to the variation in production, perception studies have found that the boundary between short and long-lag VOT stops shifts to shorter VOTs when the speech rate increases in preceding phrases and identical stimuli are more likely to be heard as the long-lag category (i.e., voiceless or aspirated) when embedded in fast speech [3-6]. To our knowledge, no previous study has examined if a similar perceptual adjustment is found for the perception of contrast between prevoiced and short-lag stops and, in particular if a rate-modulated boundary shift occurs in the negative VOT continuum. Given that languages can contrast multiple stop categories along the negative VOT dimension [7, 8], we should expect that listeners can develop sensitivity to not only the presence or absence of prevoicing but to the degree of prevoicing [9]. Thus, they should also be able to perform the type of rate normalization found in the positive VOT range. Our study fills this gap in the literature by investigating the effects of speech rate variation in French stops, which contrast prevoiced and short-lag stops [10, 11].

2. METHODS

2.1. Participants and experiment overview

Participants were recruited through prolific.co. The eligibility requirements were to reside in Canada or France and to consider French as their first and primary language. A total of 43 participants completed the study. Out of those, 11 were excluded for not meeting the eligibility requirement (for listing English as either their first language or their dominant language in the detailed background questionnaire). Out of the 32 participants who were retained for analysis, nine were residing in Canada, and the other 23 were residing in France. The experiment itself was built and hosted on gorilla.sc [12]. The full experiment included informed consent, a background questionnaire, a perception task, and a production task and took 28.2 minutes on average. The perception and the production tasks took 12.1 and 9.6 minutes on average, respectively.

2.2. Perception experiment

The perception task was an identification task, in which participants heard target words (*coût* vs. *gout*) embedded in a carrier phrase, “Mon mot favori est___.” and selected the word they heard. A male Canadian French speaker produced the speech materials. Two baseline stimuli were created by splicing together a token of carrier phrase (carrier), a voiceless stop closure from a voiceless stop production (clos), prevoicing of a voiced stop production (prev), an aspiration of a voiceless stop production (asp), and a post-stop vowel (v) of either a voiced or voiceless stop production. All acoustic analysis and manipulation were done in Praat [13]. (See Figure 1). The spliced baselines were subsequently manipulated to create two 18-step VOT continua of 10 ms intervals that range from -100 ms

to 70 ms. To create positive VOT tokens, the duration of prevoicing (prev) was reduced to 0 ms while the duration of aspiration (asp) was adjusted to match the intended VOT value. The voiceless closure that precedes the stop release was also manipulated to be equal to the mean of stop closure of all stop productions (voiced or voiceless) for the speaker so that the closure duration itself does not provide an independent cue for voicing. For the negative VOT continua, the aspiration (asp) portion was reduced to 0.3 ms to give a perception of stop release but not of aspiration, and the duration of prevoicing (prev) was adjusted to match the intended negative VOT value. The voiceless closure (clos) duration was also adjusted so that the total stop closure (clos + prev) was equal to the overall average closure duration.

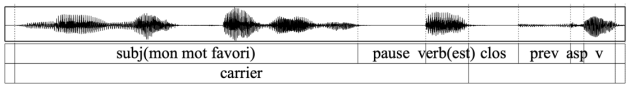


Figure 1: Segmentation of spliced baseline stimulus

The subparts of the carrier phrase (subj, pause, and verb) were also manipulated to the average value of all productions for the speaker to remove any potential durational cue to target voicing. This modified carrier was further manipulated to create fast and slow versions. The slow version was 1.5 times longer (108% of the overall average duration of the speaker's production) than the fast version (72%). These ratios were chosen to ensure that the two versions sound noticeably fast and slow without sounding unnatural. The speech rate of the carrier phrase, the VOT of the target word, and the base vowel following the target stop were orthogonally varied to create a total of 72 stimuli (18 VOT steps * 2 vowel bases (voiced vs. voiceless) * 2 speech rates). The stimuli were randomized and repeated twice for a total of 144 trials.

2.3. Production experiment

For the production task, participants heard one of two prompts from the model talker, which were the same as the fast and slow versions of the carrier phrase used in the perception task (without the target word). After hearing the prompt, participants were instructed to produce the full sentence including the target word displayed on the screen while trying to imitate the speech rate of the model talker's prompt as closely as possible. The recording was self-paced, and for each trial, participants began and ended the recording by clicking a button. In order to minimize data loss caused by participants stopping the recording too early and cutting off the end of the utterance, which is common in online production experiments [14], participants were asked to produce the sentence twice per prompt, and only the first production of each trial was analyzed.

Each target word was presented 20 times, ten times with a slow carrier prompt and ten times with a fast prompt. Each word was presented in its own block, within which the fast and slow trials were randomized. The order of the two blocks was randomized for each participant, with half of the participants starting with the voiced block and the other half starting with the voiceless block. Each participant was expected to produce a total of 40 tokens (2 target words * 2 speech rates * 10 repetitions). After excluding mispronunciations and other recording anomalies and omissions, 573 tokens for *coût* and 581 tokens for *gout* were retained for analysis. The production data were converted from .weba to mono .wav format for acoustic analysis. Each production was segmented for subj, pause, verb, clos, prev, asp, and v(owel) (see Figure 1) and measured for duration.

2.4. Statistical analysis

For the perception analysis, we built two separate mixed-effects logistic regression models using the *lme4* package [15] in R [16], one for the positive VOT values (>0 ms) and one for the negative VOT values (<0 ms). The response variable was the participants' RESPONSE (voiced = 0, voiceless = 1), and the fixed-effect predictors included VOT (ms) and BASE.VOWEL (voiced, voiceless), SPEECH.RATE (fast, slow), and DIALECT (Canadian, European), and full interaction terms. All fixed-effect predictors were centred by z-score transformation to promote model convergence. The random effects included a by-SUBJECT random intercept and by-SUBJECT slope adjustments to the two phonetic predictors, VOT and BASE.VOWEL, to account for individual variation in phonetic cue use. The full model was trimmed down using stepwise regression, as implemented in the *buildmer* [17] function in R [16].

We predict that VOT will exhibit a positive, statistically significant coefficient in the positive VOT model. However, the effect of prevoicing duration, as opposed to its mere presence or absence, on voicing perception is less clear. We also predict that BASE.VOWEL will have a significant effect, with voiced and voiceless vowels promoting voiced and voiceless percepts, respectively, due to the numerous secondary voicing cues present in post-stop vowels. We also included DIALECT and its interaction with other predictors to assess the differences in cue use between dialects. Finally, the primary factor of interest in our study is SPEECH.RATE. If both prevoicing and aspiration shorten in fast speech and

participants calibrate their perception of VOT duration cues depending on the ambient speech rate, we predict that listeners will be more willing to accept an ambiguous positive VOT token as voiceless (i.e., longer aspiration) and an ambiguous negative VOT token as voiced (i.e., long prevoicing) in fast speech.

For production analysis, we built mixed-effects linear regression models to compare the duration of each segment in slow vs. fast speech. The model included DURATION (ms) as the response variable and SPEECH.RATE (fast = -0.5, slow = 0.5), VOICING of the target stop (voiced = -0.5, voiceless = 0.5), and their interaction as fixed-effect predictors. A by-SUBJECT random intercept and by-SUBJECT slope adjustments to VOICING and SPEECH.RATE were also included. We predict that other things being equal, segments will be longer in slow than fast speech.

In the next section, we will first discuss the production results and establish the rate-conditioned change of prevoicing and aspiration in our pool of participants. We will then discuss the perception results and interpret them relative to the production variation.

3. RESULTS

3.1. Production results

The boxplots in Figure 2 summarize the distribution of the duration of each subpart by speech rate and stop voicing. For every segment, the duration is longer in the slow condition than in the fast condition, and the difference is statistically significant (subj: $b = 487.0$, $p < 0.001$; paus: $b = 95.5$, $p < 0.001$; verb: $b = 37.9$, $p < 0.001$; clos: $b = 63.0$, $p < 0.001$; prev: $b = 12.6$, $p = 0.021$; asp: $b = 16.2$, $p < 0.001$; vowel: $b = 10.0$, $p < 0.001$).

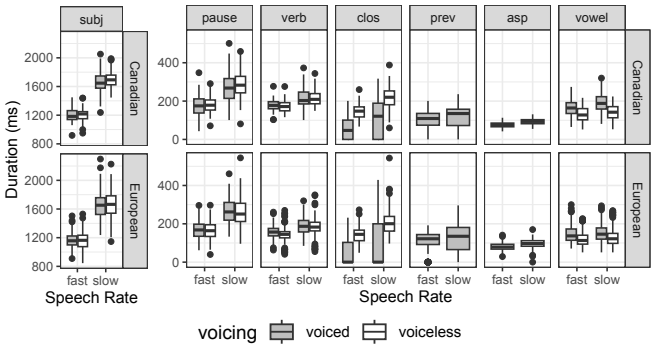


Figure 2: Boxplots of duration by segment, voicing, speech rate, and dialect

The mean duration of each segment by speech rate and the ratio of slow to fast duration are summarized in Table 1. For many of the segments, the slow-to-fast ratio is close to the model talker's ratio of 1.5, indicating that the task successfully elicited the intended speech rate in the participants' production. Although the ratios become smaller later in the sentence, and the ratios for prevoicing (prev), aspiration (asp) and the following vowel (vowel) are much smaller than 1.5., the rate difference was maintained until the end of the sentence.

	fast	slow	ratio
subj	1170.7 (106.8)	1663.1 (168.9)	1.42
pause	170.6 (45.6)	266.8 (71.8)	1.56
verb	157.3 (34.3)	195.4 (46.3)	1.24
clos	95.1 (73.4)	157.8 (104.3)	1.66
prev	104.9 (53.0)	118.5 (75.1)	1.13
asp	79.6 (19.3)	95.7 (21.2)	1.20
vowel	139.2 (45.9)	148.9 (50.1)	1.07

Table 1: Mean durations (standard deviation) in ms by speech rate and slow-to-fast duration ratios

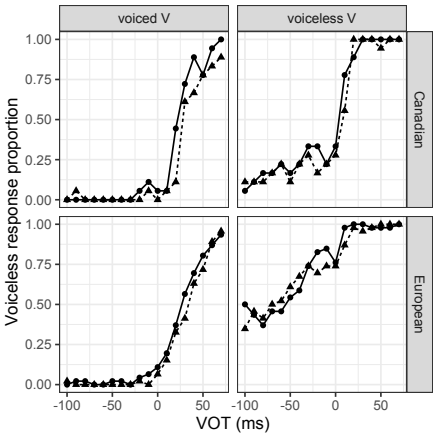
For subj, pause, and verb, the voicing of the target stop does not affect their duration significantly (subj: $b = 6.9$, $p = 0.337$; paus: $b = -2.9$, $p = 0.528$; verb: $b = -6.6$, $p = 0.117$). For the post-stop vowel, the duration is shorter for the voiceless than the voiced stop condition (vowel: $b = -26.1$, $p < 0.001$). The (voiceless) closure duration excluding prevoicing (clos: $b = 99.2$, $p < 0.001$) is longer for the voiceless than the voiced stops, as expected. For the preceding vowel (verb: $b = 8.9$, $p = 0.007$), there's a significant interaction of stop voicing and rate, such that the effect of rate is stronger for the voiceless stop than for the voiced stop.

3.2. Perception results

Figure 3 (a) displays the proportion of voiceless responses by VOT, BASE.VOWEL, DIALECT, and SPEECH.RATE. The results align with the expectation as higher VOT values and a voiceless base vowel induced more voiceless responses. These phonetic predictors also interacted with each other, with VOT having a more gradient effect when the positive VOT values were combined with the voiced base vowel or the negative VOT values were combined with the voiceless base vowel which created incongruence in the acoustic cues. The two DIALECT groups had similar responses, although European speakers heard a higher proportion of voiceless stops than Canadian speakers, particularly when negative VOTs were combined with the voiceless base vowel. We can also see the effect of SPEECH.RATE, the main predictor of interest, is visible in the positive VOT ranges. The circles and solid lines, representing voiceless stop responses in the fast rate condition, were generally higher than the triangles and dotted lines, representing voiceless stop responses in the slow rate condition, except when voiceless responses were

close to 100% regardless of speech rate. The rate effect for the negative VOT range, on the other hand, was inconsistent.

(a)



(b)

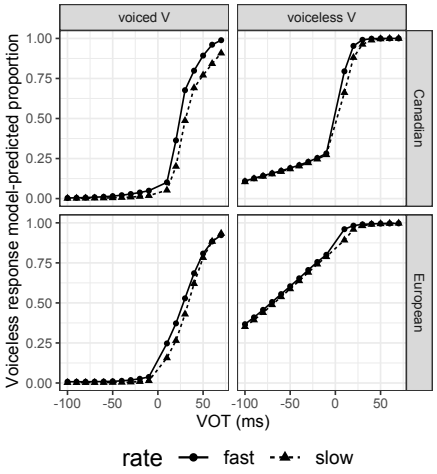


Figure 3: The average proportion of voiceless responses by VOT, BASE.VOWEL, DIALECT, and SPEECH.RATE: (a) Observed proportion; (b) Model predicted proportion

We now turn to the results of our statistical tests. First of all, in both the positive and the negative VOT models, by-PARTICIPANT random slopes for VOT and BASE.VOWEL were both retained as well as the random intercept, which means that participants varied in their phonetic cue use. By including these random slopes, our models test the phonetic predictors' effects at the group level while controlling for individual variation. Figure 3 (b) plots model-predicted voiceless response proportions.

We discuss the positive VOT model first. We found significant effects of the three linguistic predictors all in the expected direction (VOT: b = 3.16, $p < 0.001$; BASE.VOWEL: b = 2.78, $p < 0.001$; SPEECH.RATE: b = -0.43, $p < 0.001$). The main effect of DIALECT was not significant (b = - 0.24, p =0.641), but DIALECT interacted with VOT (b = - 0.71, p =0.018) and SPEECH.RATE (b = 0.31, p = 0.019) significantly. The three-way interaction of DIALECT x VOT x SPEECH.RATE was also significant (b = 0.26, p = 0.043). This means that the rate effect was stronger for Canadian speakers, and the dialect difference in speech rate effect was more pronounced in higher than lower VOT values.

For the negative VOT model, we also found a significant main effect of VOT (b = 1.08, p = 0.003), which means that not only the presence or absence of prevoicing but the duration of prevoicing can affect voicing perception. The main effect of BASE.VOWEL was also significant (b = 2.29, $p < 0.001$) in the expected direction. The interaction of VOT and BASE.VOWEL was significant (b = 0.40, p = 0.043), which means that the VOT duration effect was stronger when the base vowel comes from a voiced stop token. We found a significant effect of DIALECT (b = 1.08, p = 0.014) and a significant interaction of DIALECT and BASE.VOWEL (b = 0.81, p = 0.017). This means that the European speakers heard more voiceless stops than the Canadian speakers and this dialect difference is more pronounced when the base vowel is voiceless. Finally, we found a significant main effect of SPEECH.RATE (b = -0.34, p = 0.033) and a marginal interaction of SPEECH.RATE and BASE.VOWEL (b = 0.28, p = 0.082). This means that more voiceless stops were heard in fast speech than in slow speech, the opposite of the predicted pattern if the speech rate adjustment mirrors the shift in prevoicing duration in production, and this effect holds only when the base vowel is voiced and not when it is voiceless.

4. DISCUSSION

Our study examined if and how listeners adjust their perception of contrast between prevoiced and short-lag stops depending on the ambient speech rate. The results show that similar to the previous findings from the contrasts between short-lag and long-lag voiceless stops, listeners adjusted their expectation of VOT length and heard more voiceless stops (longer VOT category) in fast than in slow rate conditions. For the negative VOT values, fast speech induced more voiceless responses in the voiced base vowel condition. We interpret this marginal and unexpected effect as a lack of predicted effect. Despite the fact that listeners attend to the duration of prevoicing in perception, they were not more willing to accept a prevoiced stop as voiced in fast than in slow speech condition. We note that while both aspiration and prevoicing shortened in fast speech production, the effect was less consistent across speakers for prevoicing than for aspiration. This difference in production may explain the observed difference in perception between prevoicing and aspiration.

5. REFERENCES

[1] Kessinger, R., Blumstein, S. 1997. Effects of speaking rate on voice onset time in Thai, French and English. *Journal of Phonetics* 25, 143-168.

[2] Allen, J. S., Miller, J. L. 1999. Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America* 106 (4), 2031-2039.

[3] Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception* 7(5), 1074.

[4] Kang, Y., Kung, K. Li, L. J., Ting, C., Yeung, J. 2018. Speech rate variation in English stop production and perception. *Toronto Working Papers in Linguistics 40*.

[5] Kidd, G. R. 1989. Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance* 15(4), 736.

[6] Miller, J. L., Green, K. P., Reeves, A. 1986. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43 (1-3), 106-115.

[7] Cho, T., Whalen, D. H., Docherty, G. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *Journal of Phonetics* 72, 52-65.

[8] Atta, F., van de Weijer, J., Zhu, L. 2022. Saraiki. *Journal of the International Phonetic Association 52 (3),* 541-561.

[9] Hazan, V. L., Boulakia, G. 1993. Perception and production of a voicing contrast by French-English bilinguals. *Language and Speech* 36(1), 17-38.

[10] Abdelli-Beruh, N. B. 2004. The stop voicing contrast in French sentences: Contextual sensitivity of vowel duration, closure duration, voice onset time, stop release and closure voicing. *Phonetica* 61(4), 201-219

[11] Keating, P. 1984. Phonetic and phonological representation of stop consonant voicing. *Language* 60(2), 286-319.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.05, retrieved 5 January 2022 from http://www.praat.org/

[13] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[14] Sullivan, L. 2022. Participant error in online production data collection in Gorilla. Challenges for change: a crowd-sourced brainstorming session. A talk given at *LabPhon 18 Satellite Workshop.*

[15] Bates, D., Mächler M., Bolker B., Walker S., Christensen, B. R. H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, Pavel N. 2022. *Lme4:* Linear mixed-effects models using Eigen and S4. R package version 1.1-20.

[16] R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna. R Foundation for Statistical Computing.

[17] Voeten, C. C. 2022. *Buildmer:* Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 2.4.

* This study was funded by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209) and the University of Toronto Excellence Award.

MISMATCHED COARTICULATORY INFORMATION HINDERS LEXICAL ACCESS OF CORONAL STOPS IN MALAYALAM

Seema GP, Meghavarshini Krishnaswamy, Ramesh Mishra, Indranil Dutta

Technische Universität Dresden, University of Arizona, University of Hyderabad, Jadavpur University

gp.seema@gmail.com, mkrishnaswamy@arizona.edu, rkmishra@uohyd.ac.in, indranildutta.lnl@jadavpuruniversity.in

ABSTRACT

Coarticulatory information affects lexical access, more so among dense inventories. We explore the Malayalam dental, alveolar and retroflex stop contrast. In this eye-tracking study, VC:V words were used where 'C' was one of the three stops. Consonants were cross-spliced with either the same stop (match condition) or one of the other two from an identical vocalic environment (mismatch condition) to generate real word audio stimuli. Participants viewed two words, followed by an audio. Pupillary fixations were analysed for 3 conditions- target matched with audio; audio from the 'mismatch condition' with the the distractor and audio either sharing a stop or not.

Participants looked at the target and distractor at the same rate in the second condition. Coarticulatory information from two different consonants in the audio impacts lexical access. Presence of matched coarticulatory information results in higher looks at the target, even if the distractor is from the same phonological cohort.

Keywords: coarticulation, lexical access, dense coronal inventory, eye-tracking, Malayalam

1. COARTICULATORY INFORMATION AND LEXICAL ACCESS

One thread of research highlights the significance of segmental coarticulation as an aid to speech perception ([1, 2]). Coarticulatory information also effects lexical access and activation [3]. Research on coarticulatory resistance has also shown that segments exhibit variable degrees of coarticulatory resistance ([4, 5, 6]). While coarticulatory information is used successfully by listeners to both perceive and access lexical items, it has not been shown how coarticulatory resistance interferes in this process, especially when the segmental contrasts are relatively dense or comparably sparse. In this eye-tracking study, we show how coarticulatory information as a result of dynamic gestural overlap interacts with lexical access. This experiment is aimed to tease out the fine-grained interaction of contrastive and coarticulatory information in aiding lexical access.

Match-mismatch studies where auditory information is presented to subjects alongside visual stimuli within a modified visual world paradigm show that listeners pay keen attention to coarticulatory information for successful lexical access and activation. There are a number of studies that explore the subtle relationship between fine-grained phonetic information and lexical activation. For instance, listeners are found to use prosodic cues such as vowel duration to anticipate word length and constrain competing lexical choice [7]. In an eye-tracking paradigm, subjects look preferentially at those bisyllabic items that are consistent with auditory cues exhibiting vowel durations stemming from bisyllabic items. Similarly, when cross-spliced words where misleading coarticulatory cues for final consonants are tested against consistent patterns, subjects tend to show effects of these inconsistencies vis-a-vis lexical activation and competition (within the visual world paradigm) [8]. More recently, Beddor, McGowan, et al. show that listeners' moment-by-moment fixations on visual displays are mitigated by coarticulatory acoustic effects in real-time [3]. While these studies point to how coarticulatory information is successfully utilized for lexical access; the interaction of articulatory-motor and auditory-perceptual information, and the attendant effect on lexical access has not been studied. There is hardly any information about how varying inventory sizes could impact both coarticulatory influence and lexical access in complex ways.

In this study, we report on an eye-tracking study where cross-spliced matched and mismatched auditory stimuli are presented to the subjects in a modified visual world paradigm.

2. CORONAL STOPS IN MALAYALAM

Malayalam exhibits a number of coronal segments that makes it interesting to test the predictions

of articulatory-motor and auditory-perceptual constraints in languages. Malayalam has a three-way contrast in place of articulation among plosives showing differences in tongue tip and tongue blade involvement (and sublaminal regions in the case of the retroflex), namely dental /t̪ː/, alveolar /tː/ and retroflex /ʈː/ [9]. This three-way contrast is only available intervocalically, and the plosives are always geminate in this context. Thus, Malayalam poses interesting questions on both these counts due to the fairly large number of contrasting segments in the dental/alveolar region. We are especially interested in determining how dynamic information (that is, the relationship between coarticulatory cues in the preceding and succeeding vowels, and the place cues in the articulation of the consonant itself) about the consonant place of articulation in the adjoining vowels can be successfully utilized by listeners to access the lexicon and what (if any) effect phonological cohorts may have in mitigating lexical access and activation. To that end, we present subjects with auditory stimuli under matched coarticulatory information, and mismatched coarticulation in a visual world experiment. Our results indicate that listeners gaze at printed words with coarticulatory information regardless of the presence or absence of a phonological cohort word. We argue that in dense consonantal systems, in addition to the activation of the cohort, coarticulatory information (ostensibly, the dynamics of variable articulatory overlap) is encoded in the lexicon.

3. MATERIALS AND METHODS

3.1. Materials

72 words were chosen for this experiment containing a $V_1C\text{ː}V_2$ sequences, where V_1 belonged to the /ə, o, i, u, e/, V_2 was from the set /ə, aː, i, o, u, e/, and Cː is one of the three coronal geminate stops. The words chosen for this study are a subset of the ones used in [10], with four words excluded because the vowel neighbourhood only had one of the three stops.

First, three instances of each word were recorded in isolation by a female native speaker of Malayalam, using a fixed cardioid condenser microphone (Shure Beta 53), and a sampling rate of 22.5 kHz. Next, these recordings were used to create stimuli using using Praat [11] for two conditions: match and mismatch. For each condition, the stop closure and burst of the coronal stop were removed and replaced with the stop closure and burst from the audio of a different word. The stop closure was identified as the earliest point with low or no energy in the signal. The end of the burst was marked as the last point before the appearance of formant bands in the signal while ensuring that the burst was included completely in the extraction. This process separates the acoustic information of the stops' place of articulation observed in the stop (closure duration, burst amplitude) from that observed in the neighbouring vowels (formant transitions, F3 changes (retroflexes)).

For the match condition, closure and burst from one of the other instances of the same word were used for the cross-splicing (coarticulatory information before closure matched information of the consonant). For the mismatch condition, two stimuli were generated where possible, using a word with the same vocalic environment (i.e. identical V_1 and V_2), but a different coronal stop. This yielded 72 tokens for the match condition, and 105 tokens for the mismatch condition, with every word represented at least once in each condition. Figure-1 is an example of the original recording, the match condition, and the mismatch condition for the token [et̪ːi]. The stimuli in the second frame was generated by replacing the closure and burst from one of the other two recordings of [et̪ːi], while the stimuli in the third frame used the closure and burst from the token [eʈːi] (Figure 1). Following the cross-splicing, four native speakers of Malayalam listened to the stimuli in the mismatch condition, and provided qualitative feedback on the stimuli, to ensure that the words were still identifiable in the altered tokens.

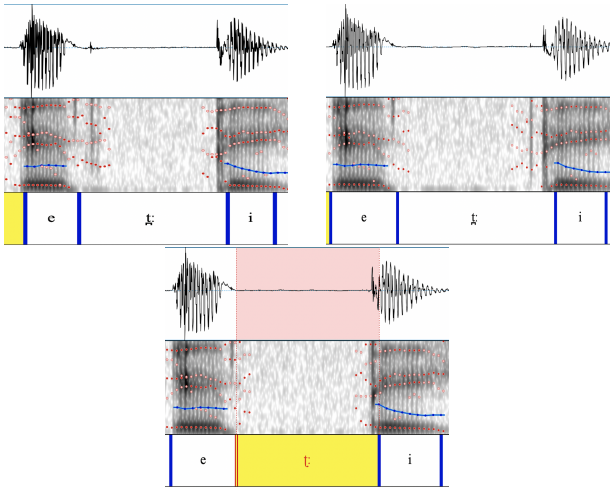


Figure 1: original recording, matched token and mismatched token for the word "et̪ːi"

3.2. Methods

We chose eye-tracking because it allowed us to study the effects of keeping or removing given acoustic cues in a stimuli in a task involving lexical

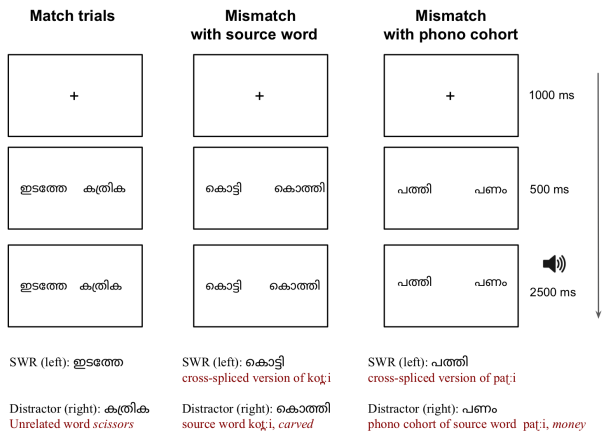


Figure 2: Sequence of events on a sample trial for the three experimental conditions. Note: SWR denotes spoken word referent

processing. 12 participants (4 female, mean age = 25 years, SD = 2.4) took part in the study. All participants were students at University of Hyderabad and had acquired Malayalam as their L1. All participants reported normal or corrected-to-normal vision. Eye movement data was recorded using a desktop mounted Eyelink 1000 eyetracker (SR research, Ontario) with a sampling rate of 1000 Hz under binocular viewing. Participants rested their head on a chin rest for stable viewing. A 9-point calibration was used for each participant. Stimuli was presented on a 19 inch LCD monitor with a refresh rate of 60 Hz placed at a distance of 60 cm from the participant.

Each trial began with a fixation cross for 1000 ms. A preview of the visual world display consisting of two printed Malayalam words was presented for 500 ms. Next, the Malayalam spoken word was presented and the visual world display was shown for 2500 ms longer. All the stimuli were presented in black against a grey background. There were four types of trials:

1. *Match trials* where the spoken word contained matching coarticulatory information to one of the printed words. The other word in the display was an unrelated distractor. These trials were included to serve as a baseline to demonstrate that our paradigm can capture the biased looks towards spoken word referents.
2. *Mismatch trials with source word, or "Mismatch (source)"* where the spoken word contained mismatched coarticulatory information with one of the printed words.The other word in the display was the source word that matched the coarticulatory information in the spoken word.
3. *Mismatch trials without source word "Mismatch (phono cohort)"* where the spoken word included mismatched coarticulatory information. The display contained the printed version of the spoken word and another word that was phonologically similar to the spoken word (but didn't match the coarticulatory information).
4. *Filler trials* where both the words in the display were unrelated to the spoken word. The filler trials were included to break anticipation and strategy.

Each participant was administered 10 match trials, 10 mismatch trials with source word, 6 mismatch trials without source word and 10 filler trials.

4. RESULTS

Fixations to the two printed words were extracted using SR research Dataviewer (SR research, Ontario). A square box of 5° by 5° around each word was considered an area of interest (AOI) for the eye movement analyses. The fixations falling within the AOIs from the onset of the spoken word till end of trial were considered for analyses. For each participant, the proportion of fixation to each word was calculated as the number of fixations to that word divided by the total number of fixations in the 1500 ms duration on each trial.

Statistical analyses on fixation data was performed using "lme4" package in R ([12]). Fixation data was logit transformed and analysed using mixed effects logistic regression models. *Glmer* function was used with family specified as "binomial" and "link logit". Participants and items were added as random effects in all analyses.

The fixation data on match trials was analysed to first investigate if spoken words elicited looks towards the referent word. Object type ("spoken word referent", "distractor") was added as a fixed effect with "Spoken word referent" as the baseline. Subject-wise and item-wise random slopes for Object type were also included in the model. Looks to the spoken word referent were significantly greater than looks to the unrelated distractor (t = - 3.6, p = 0.002) indicating that our paradigm was effective in capturing language-mediated eye movements.

Similar analyses was conducted on filler trials with object type ("distractor1", "distractor2") as a fixed effect. As expected, there was no difference in looks towards the two words (t = -0.33, p = 0.751).

The two mismatch conditions were then analysed by including condition (source word, phonological cohort of the source word) and object type (spoken word referent, distractor) as fixed effects with

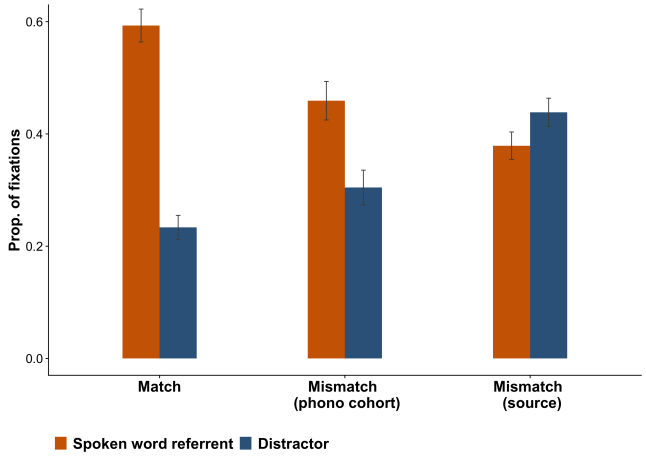


Figure 3: Proportion of fixations for the “Match" and the two “Mismatch" conditions. In the “Match" condition, the distractor was an unrelated word. In the “Mismatch (source)" condition, the distractor was the source word whereas in the “Mismatch (phono cohort)" condition, the distractor was the phonological cohort of the source word

“source word" and “spoken word referent" as baseline, respectively. Subject- and item-wise random slopes for both the fixed effects were also added to the model. There was no effect of object type (t = -0.18, p = 0.321) with descriptively greater looks to the spoken word referent compared to the distractor across both conditions. Crucial to our hypothesis, there was a significant interaction between condition and object type (t = -2.75, p = 0.015) indicating that participants looked at the spoken word referent more often only when the distractor word was a phonological cohort of the source word. Interestingly, when the distractor word was the source word, participants looked at the source word and the spoken word referent equally often because of the source-word related coarticulatory information present in the spoken word.

5. DISCUSSION

In this experiment, we investigated the influence of coarticulatory information on lexical access in a sample of Malayalam-English bilinguals in India using the visual world paradigm with printed words. We showed that, upon hearing a spoken word cross-spliced with coarticulatory information from a source word with a different coronal stop, participants gazed at this source word as often as they looked at the printed word referred to by the spoken word. We verified that this pattern of results was driven by the coarticulatory information in the spoken word (and not by the phonological similarity between the spoken word and the source word) by including a condition where the phonological cohort of the source word was presented on the screen along with the spoken word referent. In this condition, more looks were seen towards the spoken word referent, confirming that phonological similarity was not the driving our main results. We also had a baseline condition where the spoken-word referent was presented along with an unrelated word. More looks towards the referent word were seen as is typically observed in visual world studies.

These findings, in addition to general findings of Marslen-Wilson and Welsh [13], help us argue for encoding of the dynamic nature of articulatory overlap and the attendant acoustic variation in the lexicon. This, we feel, is more so the case in consonantal systems that are dense in terms of the potential for articulatory overlap. The Malayalam three-way coronal contrast serves as a testing ground for the nature of this encoding. In recent acoustic investigation of the formant dynamics of Malayalam, it has been found that in VC:V contexts, Malayalam alveolars offer the greatest coarticulatory resistance compared to the retroflex and dentals [10].

Previous research shows that despite the greater relative articulatory complexity of the retroflex (where both tongue tip and tongue dorsum are involved), alveolars resist coarticulation from the adjoining vowels, more so than the retroflexes and dentals, in that order [10, 14]. This seemingly unusual result has been argued to be due to the low neighborhood density of the alveolars in Malayalam. Alveolars appear only in the medial position, and without a voiced counterpart, compared to the ubiquitous retroflexes and (to a lesser degree) dentals.

The absence of the burst information in the mismatched stimuli also points to the strength of the acoustic cues relevant for place of articulation discrimination when compared to dynamics of the formants, as found in the offset of the preceding vowel and onset of the following vowel.

Our results, when read along with the acoustic study of Dutta, Redmon, et al.[10], underscore the need to understand encoding of coarticulatory information at a deeper level for successful lexical access – when listeners need to access a denser cohort. They also dovetail with the results of studies that find that coarticulatory information aids speech perception [1, 2], in addition to the need to encode lexicons with fine-grained phonetic detail.

6. REFERENCES

[1] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[2] C. A. Fowler, "An event approach to the study of speech perception from a direct–realist perspective," *Journal of phonetics*, vol. 14, no. 1, pp. 3–28, 1986.

[3] P. S. Beddor, K. B. McGowan, J. E. Boland, A. W. Coetzee, and A. Brasher, "The time course of perception of coarticulation," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2350–2366, 2013.

[4] R. A. W. Bladon and A. Al-Bamerni, "Coarticulation resistance in English/l," *Journal of Phonetics*, vol. 4, no. 2, pp. 137–150, 1976.

[5] C. A. Fowler and L. Brancazio, "Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation," *Language and Speech*, vol. 43, no. 1, pp. 1–41, 2000.

[6] D. Recasens and A. Espinosa, "An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan," *Journal of the Acoustical Society of America*, vol. 125, pp. 2288–2298, 2009.

[7] A. P. Salverda, D. Dahan, and J. M. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," *Cognition*, vol. 90, no. 1, pp. 51–89, 2003.

[8] D. Dahan, J. S. Magnuson, M. K. Tanenhaus, and E. M. Hogan, "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition," *Language and Cognitive Processes*, vol. 16, no. 5-6, pp. 507–534, 2001.

[9] S. N. Dart and P. Nihalani, "The articulation of Malayalam coronal stops and nasals," *Journal of the International Phonetic Association*, vol. 29, pp. 129–142, 1999.

[10] I. Dutta, C. Redmon, M. Krishnaswamy, S. Chandran, and N. Raj, "Articulatory complexity and lexical contrast density in models of coronal coarticulation in Malayalam," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019.

[11] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[13] W. D. Marslen-Wilson and A. Welsh, "Processing interactions and lexical access during word recognition in continuous speech," *Cognitive Psychology*, vol. 10, no. 1, pp. 29–63, 1978. [Online]. Available: https://www.sciencedirect.com/science/article/pii/001002857890018X

[14] M. Krishnaswamy, I. Dutta, and M. Bhaumik, "Alveolar stops exhibit greater coarticulatory resistance than retroflexes and dentals in malayalam," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2582–2582, 2020.

PERCEPTION OF L2 ENGLISH AFFRICATE ONSETS IN NATIVE MANDARIN LISTENERS

Yizhou Wang[1] & Rikke L. Bundgaard-Nielsen[2]

[1]The University of Melbourne, [2]MARCS Institute (Western Sydney University)
yizwang3@unimelb.edu.au, rikkelou@gmail.com

ABSTRACT

This paper reports on a study investigating native Mandarin listeners' perceptual discrimination and identification of two English phonetic affricate onset contrasts, /tʃ/-/tɹ/ and /dʒ/-/dɹ/, in two vowel contexts, /i/ and /u/. Results suggest that vowel quality influences discrimination accuracy such that /u/ creates a difficult condition for Mandarin listeners, especially for the long voice of onset (VOT) contrast, /tʃ/-/tɹ/. Mouse-tracking data revealed rich information about online processing during the identification procedure, and, consistent with the results from the discrimination task, multiple metrics suggested a robust effect of vowel context. Interestingly, and in contrast to the discrimination results, VOT was not found to be an influencing factor in cursor movements in the identification, indicating that VOT may have a more nuanced role as compared to vowel context.

Keywords: affricate, perception, Mandarin, mouse-tracking.

1. INTRODUCTION

Perceiving unfamiliar nonnative and second language (L2) consonants can be challenging for listeners, especially when certain consonants form phonemic contrasts in the L2 but not in the listener's native language (L1), e.g., the English /l/-/ɹ/ contrast is notoriously difficult for native Japanese listeners to accurately perceive [1]. Many phenomena of this kind stem from differences between L1 and L2 inventories [2]. Another important line of research additionally shows that nonnative category confusions can be influenced by phonological and phonotactic contexts, e.g., L1 English and L1 French listeners are unable to accurately perceive */tl/-/kl/, and */dl/-/gl/ contrasts because both */tl/ and */dl/ are unattested sequences according to English or French phonology [3], [4]. Close analysis of such patterns indicates that */dl/-/gl/ is slightly easier than */tl/-/kl/, suggesting a potential effect of voice onset time (VOT). Consistent with this observation, a recent study [5] showed that L1 Japanese listeners' perception of English /s/-/ʃ/ contrast is conditioned by the nucleus vowel context: Japanese listeners can perform good discrimination in a familiar phonotactic context (*/su/-/ʃu/), but their performance significantly worsens in an unfamiliar phonotactic context (where the nonnative sequence is unattested in the L1, */si/-/ʃi/). These studies indicate the perception of L2 consonant onsets can be influenced by the vowel context, the phonotactic quality of the sequence (unattested or attested), and potentially, the voicing specification of the plosives of the contrast (long- or short-lag).

In the present study, we investigate L1 Mandarin listeners' perception of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ diphone contrasts, which have sometimes been reported to be challenging for Mandarin listeners [6], [7]. It remains unknown, however, how the processing of these two contrasts is influenced by their phonological contexts. Importantly, while English /tɹ/ and /dɹ/ are stop-rhotic sequences phonologically, their realisations are phonetically more similar to affricate-rhotic sequences [8]. Mandarin has long- and short-lag affricates /tʃ, dʒ/ and a rhotic category /ɹ/. For complex onsets, Mandarin phonology permits affricate-/w/ sequences, but affricate-rhotic sequences are not attested [9].

As a result, English /tɹ, dɹ/ may be perceived as *unfamiliar* phonetic affricate categories by Mandarin listeners, while English /tʃ, dʒ/ may be perceived as *familiar* affricate categories. The aim of the present study examines Mandarin listeners' perceptual sensitivity to the cues associated with the /ɹ/ segment in unfamiliar phones, which includes the labial (rounding) and lingual (narrowing) gestures. In case Mandarin listeners do not accurately perceive the rhotic segment in the onset, they would perceive /tɹ, dɹ/ as similar to English /tʃ, dʒ/, which are phonetically similar and phonotactically attested according to Mandarin phonology. Alternatively, Mandarin listeners may exploit some (but not all) the gestural cues in perception, e.g., attending to the labial gesture and thus perceiving English /tɹ/ as similar to the Mandarin sequence /tʃw/. This is supported by evidence in Mandarin loanword adaption patterns, e.g., the English name *Trump* is adapted as '川普' /tʃwan-pu/. If that is the case, English /tʃ/-/tɹ/ will be perceptually mapped to Mandarin /tʃ/-/tʃw/ in perception, and thus even unfamiliar categories may still be discriminated accurately given their acoustic and gestural cues are perceptually salient.

However, if Mandarin listeners heavily rely on the labial gesture (i.e., substituting /ɹ/ with /w/), then the perception of /tʃ/-/tɹ/ will still be difficult when the following segment also has the [+labial] feature, e.g., a rounded vowel /u/, due to anticipatory coarticulation, because then the two affricate categories would be difficult to differentiate based on whether the labial gesture is present. In contrast, an unrounded vowel such as /i/ will create a condition where the labial gesture should be maximally salient (since now only /tɹi/ will be produced with the labial gesture), and good perception would be expected. Our general prediction is thus that Mandarin listeners' discrimination and identification of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ will be influenced by the vowel context: /u/ is predicted to induce poor discrimination and identification but /i/ is predicted to induce more accurate perception.

In the following, we present an AXB discrimination task and a mouse-tracking identification task to investigate Mandarin listeners' perception of the /tʃ/-/tɹ/ and /dʒ/-/dɹ/. The discrimination task tests the pairwise discriminability of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ by analysing the discrimination accuracy, which reflects the outcome of perception [10]. The mouse-tracking identification task is deployed to investigate the online processing patterns, i.e., how phonetic-phonological information is integrated during the decision-making process, because vision, cognition, and hand motion are tightly coupled, and goal-approaching movement is a valid index of cognitive conflicts [11]–[14].

2. METHODS

2.1. Participants

Twenty right-handed, native Mandarin listeners (18 females; two males) participated in the study. None reported any hearing or speech disorders. All were international students at an Australian university (M_{age} = 24.3), and all spoke English as an L2. In addition to Standard Mandarin, eleven of them also spoke a regional Mandarin dialect, and two spoke Cantonese. None spoke a third language fluently. Their average length of residence in Australia was 2.2 years, and on average, they had learned English in foreign language classroom settings for 13.3 years. Their average age at the onset of acquisition was 6.6 years old, and their mean age of arrival was 22.1 years old. All participants completed a vocabulary size test (VST) [15], and their mean VST score was 8075. Based on these measures, the participants are advanced L2 English speakers.

2.2. Stimuli

The stimuli were eight English CVCV pseudowords, /tʃu-ti, dʒu-ti, tɹu-ti, dɹu-ti, tʃi-ti, dʒi-ti, tɹi-ti, dɹi-ti/, produced by a male native speaker of Australian English, who was phonetically trained. These stimuli were used to create six contrasts, including four critical contrasts /tʃi/-/tɹi, /tʃu/-/tɹu/, /dʒi/-/dɹi/, /dʒu/-/dɹu/, plus two filler contrasts, /dɹu/-/tɹu/, /dʒu/-/dʒi/. The target syllables differ systematically in terms of vowel context (/u/ vs /i/), phonological structure (true affricate vs stop-rhotic sequence), and VOT (short- vs long-lag, or voiced vs voiceless). The second syllable, /ti/, was added to generate a controlled phonological context across all stimuli. The speaker produced each pseudoword three times in a clear speech style in order to maximise the acoustic differences. Stress was on the first syllable of each stimulus word.

2.3. Procedures

The participants completed the two experiments on two subsequent days: First the discrimination task, and then the mouse-tracking task. The tasks were served online and data was collected using *PsyToolkit* [16], [17] and *PsychoPy* [18]. In the discrimination task, participants were presented with 144 trials (six contrasts, four triplets, and six repetitions) testing discrimination of On each trial, the listener was served a sequence of three stimuli (A, X, B) with an interstimulus interval (ISI) at 1.0 s, and the middle stimulus (X) was either phonologically identical to the first or the last stimulus. The long ISI was used to encourage phonological processing [19]. The listener had 3.0 s to decide whether the first two or the last two stimuli were more similar by pressing the "F" key (X = A) or the "J" key (X = B) on their keyboard.

In the mouse-tracking identification task, the participant's computer screen was normalised to a 2 units by 2 units canvas. Following the conventional mouse-tracking paradigm, a "start" box was located in the centre of the screen bottom [0, 0], while the two response labels were printed at the top-left [-1, 2] and top-right [1, 2] corner of the screen. On each trial, the listener clicked the "start" box to play the auditory stimulus, and had to move their mouse and click on the appropriate category label "CH" vs "TR" or "J" vs "DR", representing /tʃ, tɹ, dʒ, dɹ/, respectively. After a trial, the "start" box was printed again, and the listener needed to click on the box to enter the next trial. This procedure ensured that the mouse cursor started from approximately the same location each time. The stimuli were presented in a randomised order, while the directions (left vs right) of correct responses were counterbalanced. The task had 288 trials (four consonants, two vowels, three tokens per combination, two directions, and six repetitions.

The mouse trajectories were recorded during the response procedure. The sample rate of mouse-trajectory recording was 60 frames per second (FPS) for all participants, i.e., two adjacent mouse locations represent cursor displacement within 17 ms. For data analysis, all rightward trajectories were mirrored as leftward trajectories.

When the decision is easy, we expect that the mouse tracking trajectory resembles a straight line connecting the "start" button and the correct response, while cognitive conflicts can lead to more or less curved trajectories. Following previous research [11]–[13], we analysed multiple aspects of mouse movement, including response latency and curvature complexity. In particular, response latency was measured in terms of the identification RT and motor pauses (defined as the idle time after movement initiation, in seconds). Trajectory curvature complexity was measured by total trajectory distance and the maximal deviation from the ideal straight line (both in normalised units).

3. RESULTS

3.1. AXB discrimination

Participants achieved very good performance in discriminating the two filler contrasts (/dɹu/-/tɹu/, 95%, /dʒu/-/dʒi/, 97%), and the filler trials were systematically removed in the statistical analysis. For AXB accuracy (Table 1), we built a generalised linear mixed-effects model (GLMM, binomial link) to evaluate the effect of the onset, vowel, and their interaction whilst controlling participants as a random factor. Tukey-adjusted *post hoc* tests revealed that the contrast /tʃu/-/tɹu/ had significantly lower accuracy as compared to /tʃi/-/tɹi/ (p = .013), /dʒi/-/dɹi/ (p = .017), and /dʒu/-/dɹu/ (p = .027), but these three contrasts did not differ significantly. We further compared mono-dialectal Mandarin speakers with participants who spoke an additional dialect, but no difference was observed in all contrasts (p's > .05, t-tests). The correlation between accuracy data and the speakers' vocabulary sizes was checked, but no significant coefficients were observed (p's > .05, Pearson's r), suggesting that the cohort of participants tested here was homogenous.

3.1. Mouse-tracking identification

The accuracy data of the identification task are summarised in Table 2. In general, the participants achieved very high accuracy (93-97%) in the /i/ context, but the /u/ context led to much poorer performance (57-85%). For statistical analysis, we built a GLMM (binomial link) to model the effect of vowel condition (/u/ vs /i/), phonological structure (affricate vs sequence), and VOT (short vs long). A Wald Chi-squared test revealed a main effect of vowel condition [χ^2 = 479, p < .0001], structure [χ^2 = 149, p < .0001], and VOT [χ^2 = 16, p < .0001]. In addition, there was a significant vowel-structure interaction effect [χ^2 = 9.1, p = .0026], vowel-VOT interaction effect [χ^2 = 13.0, p = .0003], and a vowel-structure-VOT interaction [χ^2 = 7.3, p = .0070]. In addition, a short VOT led to higher identification accuracy in the /u/ context, and the true affricates (/tʃu, dʒu/) had lower accuracy than the corresponding sequence categories (/tɹu, dɹu/).

CV-CV	Condition	Acc. (SD)
/tʃu/-/tɹu/	Long VOT	88 (13)
/dʒu/-/dɹu/	Short VOT	95 (6)
/tʃi/-/tɹi/	Control	98 (4)
/dʒi/-/dɹi/	Control	97 (9)

Table 1: AXB discrimination accuracy.

CV	Acc. (SD)	CV	Acc. (SD)
/tʃu/	57 (31)	/tʃi/	97 (8)
/dʒu/	66 (30)	/dʒi/	93 (20)
/tɹu/	78 (26)	/tɹi/	96 (6)
/dɹu/	85 (18)	/dɹi/	97 (8)

Table 2: Identification accuracy.

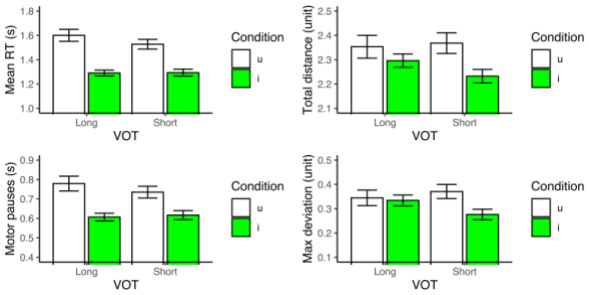


Figure 1: Mouse-tracking metrics.

For the mouse trajectory analysis, we focused on the correct responses where /tʃ/ and /dʒ/ were the target responses, while /tɹ/ and /dɹ/ were the corresponding distractors, as the accuracy data indicated that our listeners had more difficulty in identifying the true affricate categories as compared to the sequence categories. First, we analysed two latency metrics in the responses, including the mean RT and motor pauses, defined as the mouse idle time after the motion initiation, see Figure 1. We applied logarithmic transformations and built two linear mixed-effects models (LMMs) to evaluate the effect of vowel conditions as well as VOT.

For RT, a Wald Chi-squared test revealed a significant effect of the vowel [χ^2 = 69.4, p < .0001], but no significant effect of VOT [χ^2 = 1.9, p = .1686], or vowel-VOT interaction [χ^2 = 1.5, p = .2223], indicating that vowel condition but not VOT

condition affected the response RT. In the /u/ context, listeners spent a significantly longer time making identification choices. For motor pauses, we again found a main effect of the vowel [χ^2 = 32.8, p < .0001], but no significant effect of VOT [χ^2 = 0.6, p = .4533], or vowel-VOT interaction [χ^2 = 1.7, p = .1935], indicating that the listeners had significantly longer pauses in the /u/ context than the /i/ context.

Next, we analysed two curvature metrics of mouse trajectories, including total trajectory distance (length), and maximal deviation from the ideal straight line, see Figure 1. These metrics were similarly analysed using LMMs. For total distance, we found a significant effect of the vowel [χ^2 = 18.7, p < .0001], but no significant effect of VOT [χ^2 = 0.8, p = .3637], or vowel-VOT interaction [χ^2 = 0.9, p = .3474]. These results indicated that vowel condition but not VOT affected the trajectory lengths, and participants had significantly longer mouse trajectories in the /u/ context than in the /i/ context. Similarly, maximal deviation showed a significant effect of the vowel [χ^2 = 15.0, p = .0001], but no significant effect of VOT [χ^2 = 1.6, p = .2025], or vowel-VOT interaction [χ^2 = 1.5, p = .2184], suggesting a similar pattern.

4. GENERAL DISCUSSION

Both the AXB task and the identification task confirmed our prediction that /u/ can create a more challenging condition as compared to /i/ in Mandarin listeners' perception of English (phonetic) affricate onsets, but the two tasks also revealed slightly different patterns. In the AXB task, accuracy data indicated that /u/ created a difficult scenario only in the long VOT condition. However, the identification task revealed that the /u/ context caused perceptual confusion in both short and long VOT conditions, although the accuracy was still higher in the short VOT context as compared to the long VOT context. It is worth noting that Mandarin and English have similar VOT-based contrasts for perceptually distinguishing short- and long-lag obstruents [20]. In general, our finding thus indicates that VOT has a more nuanced interfering effect than the vowel context, and such effect is more apparent when the task complexity increases, because the identification task but not the AXB task requires extra knowledge to draw the correspondence between orthographic and phonological representations. Nonetheless, these findings echo previous research findings that the difficulty level of nonnative consonant perception is influenced by the phonological and phonotactic context [5], and VOT may also play a role in the perceptual easiness of unfamiliar onset categories [3], [4]. One potential explanation is that VOT differences can affect the temporal structure and the phasing relations between the articulatory gestures, and thus the perceptual salience of other gestures (e.g., the lingual and labial gestures for producing the rhotic sound) can be reduced due to an increased salience of aspiration (wide laryngeal). Or perhaps strong aspiration leads to partial devoicing of the following sonorants, and therefore the gesture cues are weakened and become more difficult to attend to, especially for L2 listeners. As for the differences in the results from our two experiments, it is possible that auditory learning precedes the learning of the correspondences between phonology and orthography, or perhaps the explicit metalinguistic knowledge creates another layer of phonological representations beyond perception itself [21], so that the discrimination and identification tasks tap into different subsystems of L2 phonology. This is also shown in the listeners' potential decision bias against the true affricate categories in the /u/ context, see Table 2.

Finally, the four online-processing metrics in mouse-tracking showed a consistent vowel effect, while the VOT effect was non-significant. The findings therefore confirmed our prediction that Mandarin listeners rely on the labial cue, but not the lingual cue, in perceiving English /tɹ, dɹ/ categories. More broadly, this finding indicates that L2 segment perception is sensitive to its immediate phonological context, e.g., the /ɹ/ segment in /tɹ, dɹ/ is likely to be substituted as a /w/ in the /i/ context, while it can be regarded as perceptually 'deleted' in the /u/ context.

To further investigate the nuanced effect of VOT, future research could recruit a group of L2 listeners who are relatively inexperienced with the target language to determine whether and the extent to which English /dʒu/-/dɹu/ can cause perceptual confusion at the beginning of L2 learning. Nonetheless, the mouse-tracking technique provides a wide range of additional metrics for understanding the cognitive processes during decision-making [11]–[14]. We argue that mouse-tracking can complement keystroke paradigms (e.g., AXB/AX tasks or identification by key-pressing) by offering additional perspectives into the online processing of L2 speech input, e.g., the change of mind *en route* as indicated by the curved mouse trajectories. A future study should analyse the types of mouse trajectories and their distribution in different experimental conditions.

5. ACKNOWLEDGEMENT

We would like to thank the twenty participants and Alexander Kilpatrick for his help in generating the experiment stimuli.

5. REFERENCES

[1] A. Sheldon and W. Strange, "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception," *Appl. Psycholinguist.*, vol. 3, no. 3, pp. 243–261, 1982.

[2] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[3] C. T. Best and P. A. Hallé, "Perception of initial obstruent voicing is influenced by gestural organization," *J. Phon.*, vol. 38, no. 1, pp. 109–126, 2010.

[4] P. A. Hallé and C. T. Best, "Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 2899–2914, 2007.

[5] A. Kilpatrick, R. L. Bundgaard-Nielsen, and B. J. Baker, "Japanese co-occurrence restrictions influence second language perception," *Appl. Psycholinguist.*, vol. 40, no. 2, pp. 585–611, 2019.

[6] Y. Lan, "Perception of English fricatives and affricates by advanced Chinese learners of English," in *Proceedings of INTERSPEECH 2020*, H. Meng, B. Xu, and T. Zheng, Eds. Shanghai, China: International Speech Communication Association, 2020, pp. 4467–4470.

[7] Y. Lan, "Vowel effects on L2 perception of English consonants by advanced learners of English," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, M. Le Nguyen, M. C. Luong, and S. Song, Eds. Hanoi, Vietnam: University of Science, Vietnam National University, 2020, pp. 149–157.

[8] L. Magloughlin, */tɹ/ and /dɹ/ in North American English: phonologization of a coarticulatory effect*. Unpublished doctoral thesis: University of Ottawa, 2018.

[9] S. Duanmu, *The phonology of Standard Chinese*. Oxford, UK: Oxford University Press, 2007.

[10] W. Strange and V. L. Shafer, "Speech perception in second language learners," in *Phonology and second language acquisition*, J. G. Hansen-Edwards and M. L. Zampini, Eds. Amsterdam: Benjamins, 2008, pp. 153–192.

[11] M. J. Spivey, M. Grosjean, and G. Knoblich, "Continuous attraction toward phonological competitors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 29, pp. 10393–10398, 2005.

[12] P. E. Stillman, X. Shen, and M. J. Ferguson, "How mouse-tracking can advance social cognitive theory," *Trends Cogn. Sci.*, vol. 22, no. 6, pp. 531–543, 2018.

[13] D. U. Wulff *et al.*, "Movement tracking of cognitive processes: A tutorial using mousetrap," *PsyArXiv*, 2021.

[14] Y. Wang, R. L. Bundgaard-Nielsen, B. J. Baker, and O. Maxwell, "Native phonotactic interference in L2 vowel processing: Mouse-tracking reveals cognitive conflicts during identification," in *Proceedings of INTERSPEECH 2022*, H. Ko and J. H. L. Hansen, Eds. International Speech Communication Association, 2022, pp. 5223–5227.

[15] P. Nation and D. Beglar, "A vocabulary size test," *Lang. Teach.*, vol. 31, no. 7, pp. 9–13, 2007.

[16] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behav. Res. Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[17] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teach. Psychol.*, vol. 44, no. 1, pp. 24–31, 2017.

[18] J. W. Peirce, "PsychoPy: Psychophysics software in Python," *J. Neurosci. Methods*, vol. 162, no. 1–2, pp. 8–13, 2007.

[19] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, pp. 35–44, 1985.

[20] J. Gong, M. Cooke, and M. L. Garcia-Lecumberri, "Towards a quantitative model of mandarin Chinese perception of English consonants," in *Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech New Sounds 2010*, Poznań, Poland, 2010.

[21] A. Cutler, "Representation of second language phonology," *Appl. Psycholinguist.*, vol. 36, no. 1, pp. 115–128, 2015.

[22] P. Cisek and J. F. Kalaska, "Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action," *Neuron*, vol. 45, no. 3, pp. 801–814, 2005.

THE ROLE OF IMMEDIATELY PRIOR EXPOSURE AND TALKER ACCENT ON SOCIOLINGUISTIC VARIANT IDENTIFICATION

Aini Li and Meredith Tamminga

Department of Linguistics, University of Pennsylvania
liaini@sas.upenn.edu; tamminga@ling.upenn.edu

ABSTRACT

In everyday speech perception, recognition of variable pronunciations is crucial, as different pronunciation variants may signal social information. Research in speech perception has investigated how perception can be robust to gradient phonetic variability, but less is known about how listeners integrate contextual cues in identifying discrete sociolinguistic variants. We present three experiments investigating the role of immediately prior exposure and its interaction with talker accent in perception of the English sociolinguistic alternation between *-ing* (/ɪŋ/) and *-in'* (/ɪn/). We hypothesize that having recently encoded one variant will increase the probability of perceiving the same variant subsequently during perception, and that this priming effect will be modulated by talker accent via social expectations. We find that the *-ing/-in'* variants can be primed in perception. However, this variant priming effect does not significantly interact with talker accent, suggesting that perception of variable pronunciations may not integrate local statistics and social unexpectedness simultaneously.

Keywords: speech perception, sociolinguistic variants, priming, talker accent

1. INTRODUCTION

Human speech is highly variable such that words can have different pronunciations. For instance, the word WALKING can be pronounced as either *walking* or *walkin'*. Crucially, these variable pronunciations often carry rich social information about the talker (i.e., age, gender, dialect, accent) [1, 2, 3]. A key challenge for listeners in everyday language use thus is to recognize different pronunciation variants reliably. Previous research in speech perception has examined why and how human speech perception, and by extension, sociolinguistic perception can be robust regardless of talker and speech variability [4, 5]. Recently, it has been suggested that speech perception can be viewed and modelled as a dynamic process of inference under uncertainty [6, 7, 8, 4]. That is, as speech signals unfold over time, listeners constantly draw linguistic inferences, update expectations, and generalize knowledge across words, sentences and talkers given their pre-existing knowledge of probabilistic distributions of acoustic, linguistic and social cues [5]. For instance, during processes such as phoneme categorization, when given an ambiguous sound halfway in the /s/-/ʃ/ continuum, listeners would be more likely to perceive the ambiguous sound as either /s/ or /ʃ/ based on what cue distributions they were exposed to and how they upweighted or downweighted certain cues. This approach offers a useful window to understand how listeners overcome gradient variability through cue integration during abstraction and generalization. However, we still have less idea of how listeners integrate different contextual cues in identifying equally well-formed discrete sociolinguistic variants, where the variants are distinct phonemes but are not lexically contrastive in the relevant context.

To probe this question, this study investigates how listeners recognize the well-studied sociolinguistic alternants *-ing* (/ɪŋ/) and *-in'* (/ɪn/) in English. We focus on the role of two different types of cues: immediately prior exposure and its interaction with talker accent. More specifically, we first test whether identification of discrete sociolinguistic variants can be primed by the immediately prior variant that the listener perceived (**Experiment 1**). In addition, we ask whether this priming effect, if it exists, can be modulated by talker accent via social expectation. Previously, it has been shown that linguistic expectations modulate priming effects in various experimental priming paradigms such as lexical repetition priming [9] and syntactic priming [10]. Of particular interest to our current investigation is the generalization that unexpectedness enhances priming; for example, unexpected verb constructions elicit stronger syntactic priming [10]. This raises the question of whether social unexpectedness might also enhance priming. We first ask whether a Southern accent in

American English influences how listeners identify *-ing*/*-in'* variants (**Experiment 2**). A Southern accent was chosen for this experiment because the variant *-in'* has been found to be strongly associated with the American South [11, 12, 13]. Despite intra-regional differences, Southern accents are among the most recognizable accents in the U.S. [14], which makes it easier for listeners to detect the difference between Southern and non-Southern accents. We then ask whether variant priming is modulated by talker accent (**Experiment 3**). Taken together, this study aims to use three experiments to develop a better understanding of how listeners integrate contextual cues in sociolinguistic variant identification.

2. THE CURRENT STUDY

2.1. Experiment 1: Priming variant choices in the lab

Exp. 1 tested the hypothesis that hearing one variant of *-ing*/*-in'* would make listeners more likely to perceive that same variant when given an ambiguous target for categorization. We compared whether a target word such as *sleepING*, with an ambiguous final nasal, is more likely to be categorized as *sleepin'* after *walkin'* than after *walking* (note that here we use <-ING> notation to indicate the ambiguous suffix form).

The experiment utilized a paradigm that combined two established tasks: lexical decision and forced-choice categorization. The lexical decision task was used to prime listeners' perception of different variants, while the forced-choice categorization task was used to probe listeners' perception of ambiguous targets. The ambiguity in the critical targets was created by performing source extraction on the *-in'* suffix through inverse filtering, resulting in the information filtered by the vocal tract, such as place of articulation, being masked while the information produced by the vocal folds, such as the intonational contour, remains unchanged. A norming study was used to estimate that the baseline *-ing* perception rate for these ambiguous items was 70%. The lexical decision task was adopted for prime trials to make it possible to filter out incorrect prime responses, which suggest the prime was not perceived as intended and therefore might not have been effective.

We adopted a within-subjects design with two critical conditions: *-in'*-primed and *-ing*-primed. To rule out the possibility that participants were converging to the model talker's overall *-ing*/*-in'* rate rather than being influenced by the intended prime, we controlled for the overall rates of the two variants throughout the experiment: whenever a participant responded to a target, they had been exposed to exactly 50% clear *-in* and 50% clear *-ing* primes from the model talker up to that point, regardless of which prime condition they were in.

Exp. 1 was implemented online in PCIbex [15]. 102 participants (self-reported monolingual American English speakers) were recruited from Prolific and received $10 for their participation (age range: 17-50y). Participants were informed that they would hear real English words and non-words (possible but non-existing), and that they needed to respond to these items in two different ways. On trials where they saw the prompt question, "Is this a word?", they needed to decide whether the word that they heard was a real English word or not by pressing a key (J for real words and F for non-words). On trials where they saw the prompt question, "Which word did you hear?", they needed to identify the word exactly the way they heard it by clicking on one of two options presented orthographically. To improve the acceptability of informal pronunciations in an experimental setting, following [16], participants were told that some of the words they would hear might be pronounced in a casual way, but that casually pronounced words were still real words of spoken American English. Participants were also given audio examples, then given a practice session with feedback (including words, non-words, and *-ing*/*-in'* variants) before the experiment.

38 critical ambiguous targets paired with 38 clear primes were created and included. Primes and targets were matched in whole word frequency using the SUBTLEX_{US} Log10CD measure [17]. Primes and targets also differed in the consonants immediately preceding the suffix *-ing*/*-in'*. 200 filler trials of various types including distractor sequences (e.g., sequences where targets after *-ing* or *-in'* did not have *-ing*/*-in'*. The word-to-nonword ratio for the lexical decision trials was 50%. All the experimental items were recorded by an adult white male native speaker of American English from New Jersey.

Analyses were conducted using R version 4.0.5 [18]. Mixed effects logistic regression was run using the *lme4* package version 1.1-27.1 [19], and plots were created using the *sjPlot* package version 2.8.9 [20]. Thirty-one participants were removed due to their accuracy in the lexical decision task being lower than 80%. Data from the remaining 71 participants were analyzed. Critical sequences where participants judged *-in'*-containing words as "nonword" were also excluded

from the final analysis. A mixed-effects logistic regression was conducted to predict listeners' *-ing*/*-in'* responses when identifying ambiguous targets (0 for *-in'*, 1 for *-ing*). Fixed effects were PRIME CONDITION (*-in'*-primed vs. *-ing*-primed), TARGET FREQUENCY, TRIAL NUMBER and the two-way interaction: PRIME CONDITION * TARGET FREQUENCY. PRIME CONDITION by SPEAKER and TARGET FREQUENCY by SPEAKER were included as random slopes correlated with a by-speaker random intercept. TARGET was also included as a random intercept. All the categorical predictors were sum-coded (-1, 1) and continuous predictors were scaled and centered using z-scores. Model output revealed a significant main effect of PRIME CONDITION ($\beta = 0.77$, $p < 0.001$), as Figure 1 illustrates.

Participants were significantly more likely to categorize an ambiguous target as containing *-ing* when they had just heard an *-ing* variant on the previous trial. No other predictors were statistically significant. On top of this predicted effect of prime, in both conditions, participants were more likely to perceive ambiguous targets as containing *-in'*, i.e., compared to the average *-ing* response baseline of 70% in the norming study (the red dashed line). It seems that when inferring which of two variants they heard, in the *-ing*-primed condition, listeners were matching their *-ing* responses to the model talker's overall *-ing* rates (the blue solid line), perhaps due to a global expectation of *-ing* from the experimental context. This could suggest that the priming effect is mostly driven by the *-in'*-primed condition, consistent with an account where unexpected forms elicit stronger priming.

2.2. Experiment 2: Talker accent modulates variant choices

Having established that phonological variant choices in perception can be primed, Exp. 2 tested whether listeners also integrated social information, specifically talker accent attributes, during discrete phonological variant categorization. In particular, we compared whether listeners would be more likely to perceive an ambiguous ING-suffixed word as *-in'*-containing, as opposed to *-ing*-containing, when the talker had a noticeable Southern accent instead of a less regionally-marked American English accent, which we refer to for convenience as a "general" accent.

Exp. 2 was implemented online through PCIbex. A total number of 102 participants (self-reported monolingual American English speakers) were recruited from both Prolific (N=80) and the undergraduate subject pool (N=32) in return for either monetary compensation (2 dollars) or course credit. The participant make-up from both sources was held consistent for each condition. Participants were randomly assigned to one of two critical conditions: a Southern accent condition and a general accent condition. In the Southern accent condition, all the items participants heard were Southern-accented. In the general accent condition, all stimuli were presented in a general accent instead. Participants were *not* given explicit information about what accent they would hear. The experiment used a simple forced-choice categorization task. For each stimulus item, participants were asked to identify whether they heard the word as containing *-ing* or *-in'*. The whole experiment lasted around 10 minutes.

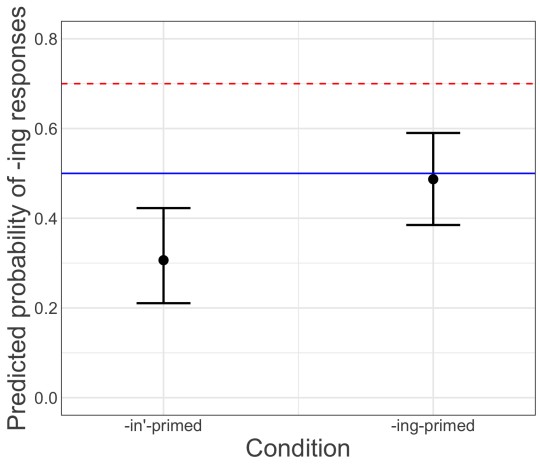


Figure 1: Predicted probability of perceived *-ing* responses in **categorization**, when all other predictors held at their average values, in probability terms. The red dashed line indicates the baseline *-ing* perception rates. The blue solid line refers to the overall *-ing* rates listeners were exposed to.

Target items (N = 38) from Experiment 1 were used as critical stimuli for this experiment. To keep the two accent conditions maximally similar, an adult white female bidialectal speaker from Louisiana who speaks both Southern-accented American English and a less-marked general American English was asked to produce critical items in both accents (henceforth: Southern guises vs. general guises). Ambiguous stimuli were created using the same procedure as in Experiment 1, with all ambiguous items derived from originally *-in'*-suffixed forms. In the end, 38 Southern-accented and 38 general-accented ambiguous targets were created and served as the final critical stimuli.

A mixed-effects logistic model was fit to predict

listeners' perceived *-ing* responses to ambiguous targets with TALKER ACCENT (General accent vs. Southern accent, sum-coded) as the independent variable. PARTICIPANT and TARGET were included as random intercepts. There is a main effect of TALKER ACCENT: Overall, listeners were significantly more likely to identify ambiguous targets as containing the *-ing* variant in the general accent condition, as opposed to the *-in'* variant in the Southern accent condition ($\beta = 1.30$, $p < 0.001$).

2.3. Experiment 3: Do variant priming and talker accent interact?

Exps. 1 and 2 showed independent effects of prior exposure and talker accent on variant perception. In Exp. 3 we ask whether these influences interact. If the priming in Exp. 1 partly reflected the unexpectedness of *-in'*, then the priming effect might be weaker in the Southern accent condition, as the association between *-in'* and Southern speech could make *-in'* primes less surprising.

For Exp. 3, 284 self-reported monolingual American English speakers were recruited from Prolific and the subject pool. As in Exp. 1, the two prime conditions (*-in'*-primed vs. *-ing*-primed) were manipulated within participants. Different from Exp. 1, talker accent was included as a between-participants factor. In the general-accent condition, participants were presented with general-accented stimuli (i.e. general-accented fillers, clear primes, and ambiguous targets), whereas in the Southern-accent condition, participants heard Southern-accented items. The same items from Exp. 1 were used as items for this experiment (both critical and filler items). Stimuli were produced in either a general accent or a Southern accent by the same bidialectal speaker from Exp. 2.

After excluding participants whose overall accuracy in lexical decision trials was lower than 80%, results from 155 participants were analyzed (general accent N = 83, Southern accent N = 72). A mixed-effects logistic regression was fit to predicted listeners' perceived *-ing* responses with the following fixed effects: TALKER ACCENT and PRIME CONDITION (in a two-way interaction), PRIME CONDITION and TARGET FREQUENCY (in a two-way interaction), and TRIAL NUMBER. PRIME CONDITION by SPEAKER and TARGET FREQUENCY by SPEAKER were included as random slopes and TARGET was included as a random intercept. Categorical predictors were sum-coded and continuous predictors were scaled and centered. We found a significant main effect of PRIME CONDITION, indicating that listeners on average were less likely to perceive *-ing* in ambiguous targets in the *-in'*-primed condition ($\beta = -0.60$, $p < 0.001$). There was also a significant effect of TALKER ACCENT, which indicates that listeners were significantly more likely to perceive *-ing* when the talker was general-accented ($\beta = 1.45$, $p < 0.001$). The interaction between PRIME CONDITION and TALKER ACCENT was not significant, which fails to support our hypothesis that the priming effect would be significantly weaker when the talker accent was Southern ($\beta = -0.24$, $p = 0.16$). Predicted effects of PRIME CONDITION and TALKER ACCENT are shown in Figure 2.

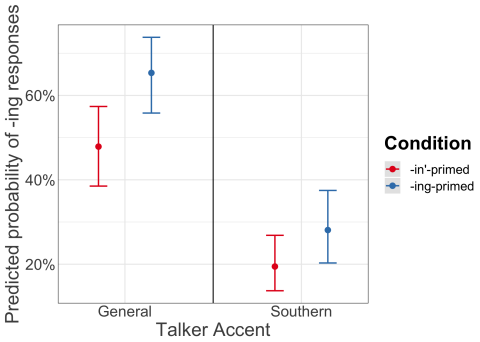


Figure 2: Predicted probability of perceived *-ing* responses across two talker accent conditions

3. DISCUSSION

We have demonstrated that first, the variant participants were recently exposed to can influence which variant they perceive subsequently. This is evidence that in speech perception, discrete sociolinguistic variants can be primed. The difference between the two conditions cannot be attributed to convergence towards the talker's overall *-ing*/*-in'* rate because the conditions do not differ in that rate. In addition, when listeners need to categorize ambiguous stimuli, talker accent influences their decisions as well. Even without being explicitly told that they would be hearing from a general or Southern-accented talker, listeners were able to infer this information from the acoustic cues in the stimuli, then use this information to associate variants with different talker accents based on participants' existing sociolinguistic knowledge (i.e., the relationship between regional accents and *-ing*/*-in'* variation). Finally, our results fail to provide evidence for a possible interaction between variant priming and talker accent, raising new questions about whether social unexpectedness modulates variant priming in the same way as linguistic unexpectedness does to other types of priming.

4. REFERENCES

[1] N. Niedzielski, "The effect of social information on the perception of sociolinguistic variables," *Journal of Language and Social Psychology*, vol. 18, no. 1, pp. 62–85, 1999.

[2] E. A. Strand, "Gender stereotype effects in speech processing," Ph.D. dissertation, Ohio State University, 2000.

[3] J. Hay and K. Drager, "Stuffed toys and speech perception," *Linguistics*, vol. 48, no. 4, pp. 865–892, 2010.

[4] D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," *Psychological Review*, vol. 122, no. 2, pp. 148–203, 2015.

[5] D. F. Kleinschmidt, K. Weatherholtz, and T. Florian Jaeger, "Sociolinguistic perception as inference under uncertainty," *Topics in cognitive science*, vol. 10, no. 4, pp. 818–834, 2018.

[6] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[7] J. S. Allen, J. L. Miller, and D. DeSteno, "Individual talker differences in voice-onset-time," *Journal of the Acoustical Society of America*, vol. 113, no. 1, pp. 544–552, 2003.

[8] R. S. Newman, S. A. Clouse, and J. L. Burnham, "The perceptual consequences of within-talker variability in fricative production," *J Acoust Soc Am*, vol. 109, no. 3, pp. 1181–1196, 2001.

[9] D. L. Scarborough, C. Cortese, and H. S. Scarborough, "Frequency and repetition effects in lexical memory," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 1, pp. 1–17, 1977.

[10] S. Bernolet and R. J. Hartsuiker, "Does verb bias modulate syntactic priming?" *Cognition*, vol. 114, no. 3, pp. 455–61, Mar 2010.

[11] A. Houston, "Continuity and change in English morphology: the variable (ING)," Ph.D. dissertation, University of Pennsylvania, 1985.

[12] W. Labov, *Principles of Linguistic Change. Vol. 2: Social Factors*. Oxford: Blackwell, 2001.

[13] K. Campbell-Kibler, "Accent, (ING), and the social logic of listener perceptions," *American Speech*, vol. 82, no. 1, pp. 32–64, 2007.

[14] R. Lippi-Green, "The standard language myth," in *English with an accent: language, ideology, and discrimination in the United States*. Routledge, 2012, ch. 4.

[15] J. Zehr and F. Schwarz, "PennController for Internet Based Experiments (IBEX)," Retrieved from https://doi.org/10.17605/OSF.IO/MD832, 2018.

[16] Y. White, "The processing and mental representation of ing variation," Ph.D. dissertation, University of Pennsylvania, 2021.

[17] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavioral Research Methods*, vol. 41, no. 4, pp. 977–90, Nov 2009.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] D. Lüdecke, P. D. Waggoner, and D. Makowski, "Insight: A unified interface to access information from model objects in r," *Journal of Open Source Software*, vol. 4, no. 38, p. 1412, 2019.

FREE CLASSIFICATION PREDICTS DISCRIMINATION OF FINNISH VOWELS BY NAÏVE JAPANESE LISTENERS

Danielle Daidone[1], Ryan Lidster[1], and Franziska Kruger[2]

[1]University of North Carolina Wilmington and [2]Indiana University
daidoned@uncw.edu, lidsterr@uncw.edu, fkruger@indiana.edu

ABSTRACT

This study explores whether listeners' groupings of vowels by similarity in a free classification task can be used to predict the discriminability of those vowels. Recent research has shown that free classification results correlate strongly with discrimination, which could potentially enable researchers to predict discrimination across a variety of phenomena not easily studied with other existing tools. However, the method's generalizability has yet to be tested. In particular, we are interested in the utility of free classification for listeners with a small first language vowel system.

Naïve Japanese listeners, whose native system contains five vowels, performed a free classification task and an oddity discrimination task with the eight vowels of Finnish. Results showed that free classification was a strong predictor of discrimination scores. These findings suggest that free classification is a useful method for predicting the discriminability of stimuli, even when listeners have a small number of native categories.

Keywords: free classification, discrimination, non-native perception, oddity, laboratory phonology

1. INTRODUCTION

Free classification (FC), also known as free sort, is a task in which listeners group stimuli simply according to how similar they perceive them to be. Participants are presented with all of the stimuli at once on the screen, and they listen to them as many times as they like and in any order that they choose. There are no researcher-imposed labels or categories; rather, the participants form groups of any size, arranging together those stimuli that they believe to sound similar. This method has been used for dialect similarity and accentedness research [1, 3, 5] and has recently been extended into investigating the similarity of non-native segments [6]. There is emerging evidence that a free classification task can be used to predict the discriminability of non-native contrasts as well [7]. In that study, we found that FC was a strong predictor of discrimination scores in experiments examining the perception of German vowels and Finnish phonemic length by American English listeners.

FC has many potential advantages as a task type: completing the task does not require knowledge of orthography or any linguistic terminology, and since all stimuli are available for comparison at once, FC typically takes a short time. Since FC can in principle be used for any type of stimuli, it could potentially enable researchers to predict confusability and differences across a variety of phenomena not easily studied with other existing tools. However, FC's generalizability has yet to be tested, and the particular phonological system of listeners' first languages (L1) may make free classification more or less predictive due to the way results are scored. When analyzing FC results, each pair of stimuli are coded dichotomously: 1 if the pair is grouped together by that participant, and 0 if not. To obtain more continuous ratings of perceived similarity, researchers average the results of multiple trials for stimuli of the same category (e.g., presenting participants with the same vowel in 2-3 different voices or consonant contexts) and/or by averaging across many participants [see 4]. For example, if Participant X grouped both male and female /e/ with female /ɛ/, but put the male /ɛ/ token in a different group, the /e-ɛ/ contrast overall would have a grouping rate of 50% for that individual (two possible pairings made, and two not made). Similarly, if Participant X grouped all of the /e/ and /ɛ/ tokens together, but Participant Y put all /e/ and /ɛ/ tokens in separate groups, the combined grouping rate for /e-ɛ/ would be 50%.

Given this type of analysis, one concern for generalizability is the potential for participants to form large groups of stimuli that they consider to be vaguely "similar" but are nonetheless discriminable. In such a scenario, FC grouping rates would fail to predict differences for discriminability within the large groups. This is especially a concern when listeners have few L1 categories; thus, many non-native stimuli could represent within-category differences for them, making it more likely for them to group many stimuli together regardless of their phonetic differences.

Therefore, to test whether FC is still a useful predictor of discrimination accuracy in the case where the listeners' L1 has few categories, we tested Japanese listeners on their perception of Finnish short

vowels. Japanese has only five vowels /a e i o ɯ/ which can be long or short, raising concern that Japanese listeners may not make sufficiently detailed groupings of Finnish vowels to predict differences in their perception performance across contrasts.

2. METHOD

2.1. Participants

Forty-one L1 Japanese listeners participated. Results from 14 participants were later excluded for various reasons: 10 had timeouts on more than 5% of oddity trials, one had lived abroad in Austria, one did not follow instructions on the FC task, one had had speech therapy, and one had phonetics training. All 27 participants (age range 18-31, M=20) whose results were included passed a bilateral hearing screening and did not report any prior hearing or speech problems. All had studied English and had one semester of study in a third language as part of university requirements (9 Chinese, 6 German, 6 French, 3 Russian, 1 Korean), in which all participants reported having beginner-level proficiency. Because none had prior exposure to Finnish, we have termed the participants "naïve" listeners with respect to Finnish.

2.2. Stimulus material and vowel analysis

The 8 short vowels of Finnish /i e y ø u o æ ɑ/ were embedded in alveolar (/tVhVt/) and velar contexts (/kVhVt/). In each of the stimuli, the same vowel was repeated, e.g. /tihit/; /kuhuk/. Three female native speakers of Finnish from Helsinki recorded the stimuli. The recordings were then analyzed in Praat [2]. Vowel boundaries were marked at a zero-crossing at the beginning and end of each vowel and the values for f0, F1, F2 and F3 were extracted for each speaker's tokens at 25%, 50%, and 75% of the total vowel duration. This was done for both vowels within a word. Following Flynn [9], formants were normalized with Gerstman's formula. The average normalized F1 and F2 values at the vowel midpoint for each context separately are shown in Figure 1. A square represents values from Speaker 1, a diamond for Speaker 2, and a triangle for Speaker 3. Solid shapes represent the alveolar context; border-only shapes represent the velar context.

2.3. Procedure

Participants completed a hearing screening, free classification (FC) task, oddity task, background questionnaire, and a perceptual assimilation task (not discussed here). All participants wore headphones and completed the tasks in this order.

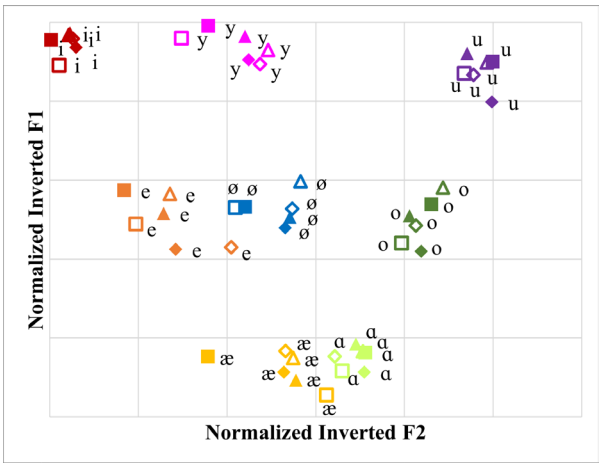


Figure 1: Average formant values of the Finnish stimuli.

FC was administered via PowerPoint. On each slide, a 16x16 grid was shown on the left and sound files in randomly numbered squares were presented on the right. Oral and written instructions in Japanese asked participants to make groups of any size according to what sounds they thought were similar to each other. Participants were instructed to click on the sound files to listen to them and drag them onto the grid to form groups. They could listen to each sound file and rearrange groups as many times as they liked. Groups had to consist of at least two sound files, and they were instructed to ignore differences in speaker. Participants completed two slides, one for each consonant context. Each slide presented 24 sound files (8 vowels x 1 context x 3 speakers), and the order of slides was counterbalanced across participants. Participants took approximately 10-20 minutes to complete the task.

The oddity task was administered via a web browser using jsPsych [8]. During each trial, participants saw three different colored robots, one representing each Finnish speaker, and an X on the screen. Each robot "said" a word, and participants clicked on the one that said something different, or clicked X to indicate that all three robots said the same thing. There were first 8 practice trials using words with a /o-e/ contrast (e.g. /tehet-tohot/) to confirm participants understood the task. If they made more than one mistake on the practice, they repeated it until they could pass and move on to the experimental trials.

The experiment tested discrimination performance on 10 different vowel contrasts: /u-y/, /u-ø/, /y-ø/, /o-ø/, /e-ø/, /e-i/, /e-æ/, /æ-ɑ/, /ɑ-o/, and /ɑ-i/, chosen because they were thought to present a range of difficulties based on pilot data from American English listeners. Each target contrast appeared once in the 6 possible sequences of "different" trials (ABB, BAA, ABA, BAB, AAB, BBA) and twice in each of the 2 possible "same" trials (AAA, BBB), per

consonant context, giving a total of 200 trials ([6 different trials + 4 same trials] * 10 contrasts * 2 consonant contexts). Trials had an ISI of 400ms, ITI of 1000ms, and a time limit of 3500ms to respond. The trials were blocked by consonant context with a short break in between, presented in random order within blocks. The task took approximately 25 minutes to complete.

3. ANALYSIS AND RESULTS

3.1. Free Classification

The researchers coded which stimuli were grouped together by each participant on each slide. An R script was then used to tabulate all possible pairs of stimuli, coded as 1 if participants put them in the same group and 0 if not. For example, if participants grouped the /ø æ ɑ/ tokens of one speaker together, then the pairs /ø-æ/, /ø-ɑ/, and /æ-ɑ/ were all given a 1 for that participant. The total number of times each pairing was made was divided by the maximum number possible, resulting in the percent of the time those stimuli were rated as similar by participants. These formed a matrix of similarity ratings between each pair of stimuli. Overall similarity ratings are displayed in Table 1. Cell shading reflects the degree of similarity.

Table 1: Free classification similarity ratings.

	ɑ	æ	e	i	o	ø	u	y
ɑ		0.70	0.04	0.01	0.13	0.35	0.07	0.01
æ			0.08	0.01	0.07	0.35	0.02	0.00
e				0.07	0.01	0.07	0.01	0.01
i					0.00	0.00	0.01	0.05
o						0.14	0.76	0.01
ø							0.10	0.19
u								0.02

Similarity ratings were then inverted (1-similarity) to create dissimilarity matrices which were analyzed using SPSS 28's ALSCAL function for multidimensional scaling (MDS) with a convergence criterion of 0.001 [4]. MDS is an analytical procedure used to analyze the similarity of multiple stimuli simultaneously by finding a positioning for all stimuli in one common space that best recreates the "distances" (dissimilarities) observed in the data. After examining solutions ranging from 1 to 5 dimensions, the 3-dimensional solution was found to be the best fit for the data for both consonant contexts separately based on the low stress values (/t/: 0.111, /k/: 0.099), high R^2 (0.937 for both), and location below the elbow of the stress plots [4]. Only minute differences between consonant contexts were found, so for brevity, Figures 2a and 2b display the averaged locations across both consonant contexts in 3D perceptual space as estimated by MDS. Both the grouping rates as well as the averaged distances between the rescaled stimuli locations in 3D MDS space were compared to oddity results to evaluate their predictive power.

The MDS solution visually shows how acoustics are warped in perceptual space, suggesting that /u/ and /o/ were perceived as very similar to each other, while /e/ and /ø/ were not. The overlap between /u/ and /o/ is a surprising result that differs from the pilot data with American English listeners.

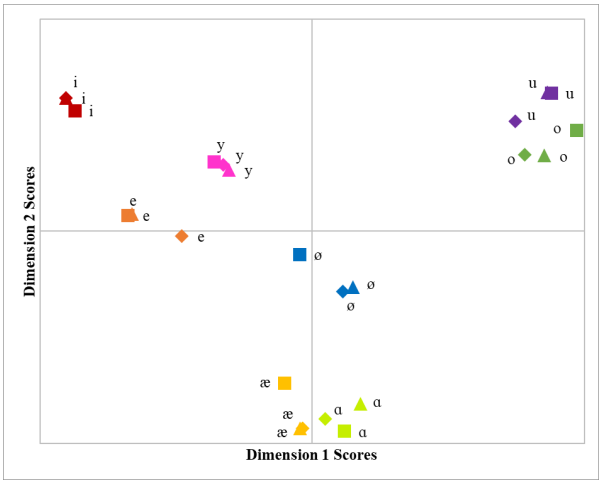


Figure 2a: MDS solution for Dim1 x Dim2.

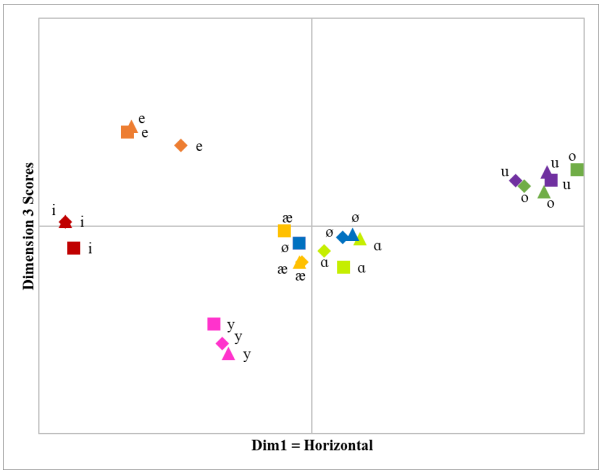


Figure 2b: MDS solution for Dim1 x Dim3. (The same data as Figure 2a if viewed "from above")

To help interpret these results, we compared dimension scores to our measurements of the vowels' acoustic properties, and also coded vowels for two phonological properties: rounding (1 for rounded, 0 for unrounded), and specifically front-rounding (1 for front rounded vowels, 0 for all else), since they represent a category of vowel not present in Japanese at all. The results, shown in Table 2, indicate that Japanese listeners seemed to predominantly use vowel backness (F2) and height (F1) to distinguish stimuli. Dimension 3 scores are less easily

interpretable, but it appears that Japanese listeners are separating front-rounded vowels from others. Duration and pitch were not related to how Finnish short vowels were grouped on the FC task.

Table 2: Correlations with MDS solutions by context.

	F1	F2	F3	Dur	f0	Δf0	Rnded	Front-Rnded
Dim1	0.25	**-0.95***	-0.33	0.09	0.25	-0.16	0.56*	-0.11
Dim2	**-0.79***	0.08	0.33	0.24	0.22	-0.05	0.30	-0.16
Dim3	-0.44*	-0.03	-0.28	0.10	-0.01	0.05	0.56*	**0.70***

* $p < 0.05$ after False Discovery Rate corrections.

3.2. Oddity

Oddity performance on each contrast was evaluated for accuracy. Accuracy rates by contrast are displayed in Figure 3. Mauchly's test for Sphericity was significant, so we ran a one-way ANOVA with sphericity corrections, showing that accuracy rates significantly differed by contrast ($F_{(5.4,140)} = 41.669$, $p < 0.001$, $\eta^2 = 0.499$). Post-hoc pairwise comparisons with Bonferroni corrections revealed 25 significant comparisons, including that /æ-ɑ/ was more difficult than every other contrast. The next most difficult contrasts all involved /ø/, showing that the Japanese listeners found it confusable with several other Finnish vowels, whereas /u-y/ was relatively easy for Japanese listeners.

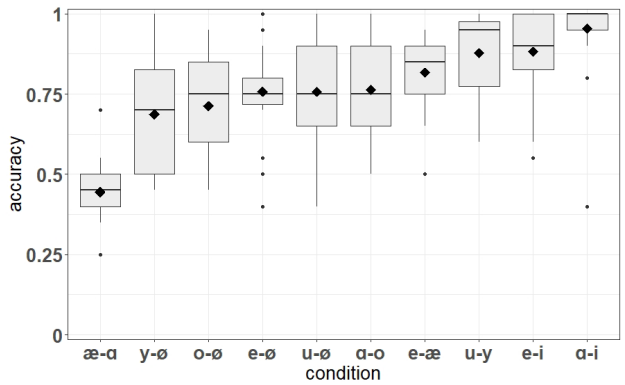


Figure 2b: Accuracy on oddity task by contrast.

3.3. Task Comparison

Linear regressions were run to determine how well the FC results predicted the accuracy of discrimination of contrasts in oddity. For the first regression analysis, the independent variable was FC similarity rates by contrast and the dependent variable was oddity accuracy scores by contrast. This regression equation was significant, $F_{(1, 18)} = 90.54$, $p < .001$, $r = 0.913$, $R^2 = .834$ (CI[.72, .93]), showing that FC similarity rates strongly predicted performance on oddity. In the second regression, distances between stimuli in the 3D MDS solution were entered as the independent variable, also yielding a significant result, $F_{(1, 18)} = 28.93$, $p < .001$, $r = 0.785$, $R^2 = .616$ (CI[.05, .91]).

4. DISCUSSION AND CONCLUSION

The performance of Japanese listeners on a variety of Finnish vowel contrasts ranged widely by contrast in ways that may not have been easily predicted simply by examining the phonemic inventories alone. For example, /u-y/ was surprisingly easy, and /y/, /u/, /o/, and /e/ all appear to be similarly confusable with /ø/, while /æ-ɑ/ was quite difficult to discriminate for Japanese listeners.

We tested whether an FC task could be used to predict this pattern of performance, and the results show that FC grouping rates were a strong predictor of variable discrimination. For example, /æ-ɑ/ was found to largely overlap according to FC, while /u/ and /y/ were perceived as much more distinct from each other. This shows that, despite Japanese having relatively few native language vowel categories, listeners nevertheless demonstrated sensitivity to phonetic detail in their grouping behavior. The surprisingly tight correlation between FC grouping rates and oddity accuracy ($r = 0.913$) suggests that the task is a valuable tool for investigating non-native perception.

Furthermore, by examining the FC results using Multidimensional Scaling (MDS), we obtained an informative visualization of perceptual warping. FC also provided a concise means of predicting performance on contrasts beyond those tested in this current experiment, opening questions for future research. For example, the unexpected overlap between /u/ and /o/ suggests that Japanese listeners may have considerable difficulty discriminating this contrast. If true, determining why this is different from Japanese listeners' attested ease with English /u-o/ [10], for example, could help advance our understanding of Japanese perception in general.

The task features of FC—in particular the lack of researcher-imposed labels and the ability to present all stimuli together—make it a promising tool for phonologists investigating the perception of a wide variety of phenomena. More research is needed to determine whether this predictive power generalizes further to suprasegmentals and other features. By continuing to explore the relationship between task performance across languages, we believe phonologists can arrive at a more explanatory theory of L2 perception.

5. ACKNOWLEDGEMENTS

We thank Aaron Albin, Rebecca Muth, and the IU L2 Psycholinguistics Lab for their contributions.

6. REFERENCES

[1] Atagi, E., Bent, T. 2013. Auditory free classification of nonnative speech. *Journal of Phonetics 41*, 509-519.

[2] Boersma, P., Weenink, D. 2015. Praat: doing phonetics by computer. Version 5.4.08 http://www.praat.org/.

[3] Bradlow, A. R., Clopper, C., Smiljanic, R., Walter, M. A. 2010. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication 52*, 930–942.

[4] Clopper, C. G. 2008. Auditory free classification: Methods and analysis. *Behavior Research Methods 40*, 575–581.

[5] Clopper, C. G., Pisoni, D. B. 2007. Free classification of regional dialects of American English. *Journal of Phonetics 35*, 421–438.

[6] Daidone, D., Kruger, F., Lidster, R. 2015. Perceptual assimilation and free classification of German vowels by American English listeners. *Proc. 18th ICPhS.* University of Glasgow.

[7] Daidone, D., Lidster, R., Kruger, F. 2023. Free classification as a method for predicting perceptual discriminability of non-native contrasts. *Studies in Second Language Acquisition*, 1-27.

[8] de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods 47*, 1–12.

[9] Flynn, N. 2011. Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series 211*, 1–28.

[10] Nishi, K., Kewley-Port, D. 2007. Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research 50,* 1496–1509.

EFFECTS OF SLOWED RATE OF SPEECH ON EYE GAZE TO THE MOUTH IN YOUNG ADULTS: CLINICAL IMPLICATIONS

Virginia L. Shaw, Claire Bernar, Nik Josafatow, Tyler Bonnell, Fangfang Li

University of Lethbridge, Lethbridge, Alberta, Canada
vl.shaw@uleth.ca

ABSTRACT

The slowed rate of speech used in clinical speech therapy protocols for fluency and motor speech disorders is known to confer improvements in motor-speech coordination and speech intelligibility. We investigated whether a slowed rate of speech might also support improved speech production *indirectly*, by changing the location and duration of eye gaze behaviors of listeners in favor of the mouth, where visual cues to articulation are centered. An eye tracking protocol was designed to measure listeners' total duration of fixations to the mouth in single sentences in both regular and slowed rate conditions, and comparison blocks (regular and loud volume, sentence repetition with no change). Preliminary analysis from 25 typical young adult listeners suggests greater attention to the mouth area occurs in a slowed rate condition. The results add to supportive evidence for the clinical use of a slowed rate of speech (SR) as a therapy strategy.

Keywords: speech rate, perception, attention, gaze

1. INTRODUCTION

Voluntary production of a slowed rate of speech (SR) is well-documented as an effective speech modification strategy for persons with motor speech disorders [1], as well as fluency disorders [2]. Improvements in the smooth, forward flow of speech and/or speech intelligibility with slowed speech rate are primarily attributed to coordination benefits [3, 4] and beneficial sensory feedback [5] during slowed movements.

As an addition to the known motor coordination benefits, the current study aims to explore *whether use of a slowed rate of speech (SR) may provide important secondary benefits for speech perception and motor-speech production, by* ***spontaneously drawing visual attention*** *to the sound placement cues and movements of the oral articulators at the mouth.*

The visual information provided by movement and placement of the articulators during speech production is known to affect *speech perception* [6-9]. Attention to visual speech cues can confirm an ambiguous, incomplete or poorly perceived auditory speech signal, thereby improving message intelligibility and comprehension [6]. Fewer cognitive processing resources may be needed to perceive and understand speech with supportive visual speech cues [10]. Visual cues are known to support the speech perception abilities of hearing impaired speakers [11], and can improve typical listeners' abilities to perceive speech in background noise [6, 7], when learning speech sounds in a second language [8], and when hearing accented [9] or less intelligible speech [12].

The positive effect of adding visual speech cues for accurate perception of the auditory signal is reinforced in recent *eye tracking* studies. Yi, et al. [7] found that listeners achieve higher perceptual accuracy when they spontaneously move their eye gaze closer to the center of the mouth in difficult listening environments (i.e., low signal-noise ratio, multiple speakers), presumably to take additional advantage of mouth movements for intelligibility and speaker identity cues. Banks, et al. [13] noted that longer eye gaze fixations to the mouth improved speech perception accuracy when listening to degraded (i.e., noise vocoded) speech signals.

The visual cues from articulatory movements may also support learning of *speech sound production*. Through Infant Directed Speech (IDS), young children learning speech sounds are exposed to exaggerated articulatory movements from caregivers [14]. In addition, adults have been found to produce "hyperarticulations" of speech sounds in response to less intelligible speech or speech sound production errors [15] by preschool children, providing "both enhanced input to children and an error-corrective signal" (p. 1836). As hyperarticulation creates an audible change in the speech signal, it is likely that a change in the visible movement component of the oral articulators would also occur.

Finally, in clinical speech-language pathology practice, speech therapists routinely use cued visual attention to the mouth to enhance their clinical models and provide effective feedback, particularly in speech sound treatment protocols. Cueing methods may include: verbal prompts to "look at my mouth," use of a mirror for clients to see their own articulatory movements [16], and use of specific hand shapes or touch cues at the mouth area on the clinician's (or client's) face to draw client's

observation and imitation of multisensory cues during speech sound production attempts [17].

Given the benefits for speech perception and speech production learning provided by the visual cues just described, we proposed that the mechanisms for *natural visual attention* may provide a rationale for why a clinically slowed rate of speech would draw more visual attention to the mouth than a regular speech rate. Visual attention can be described as "stimulus-driven" [18], wherein the perceptual characteristics of stimuli (across modalities) are subconsciously evaluated and visual attention is drawn to more salient stimuli [19]. One of the high saliency conditions, that of "novelty" or "surprise", is created when sensory input differs from prior experience or expectations (i.e., stronger, atypical, new, or unpredictable information) [20]. Evidence shows that "surprise is a strong attractor of human attention" (p. 1295), as measured via eye gaze [20].

With respect to speech rate, adult listeners have been found to have expectations for prosodic features based on their experience with both the particular language and speaker [21], providing a baseline or foundation for "typical" prosody characteristics such as speech rate, volume, and pitch contours. Therefore, use of a clinically slowed rate of speech, which is significantly less than the lower boundaries of typical speech rates [22], would be expected to create a "surprise" effect. Moreover, unexpected changes in movement features and timing have been found to draw visual attention [23, 24]. As the mouth is the *source* of the slowed speech signal (both extended movement and auditory components), the "novelty" or surprise effect would likely focus on the mouth. Once visual attention is focused at the mouth, all of the possible benefits of visual speech cues for speech perception and production would be available.

With the foundation of a significantly slowed rate of speech as a novel or surprising stimulus condition, we hypothesized that typical young adult participants would increase their eye gaze to the mouth area significantly more during the Slowed Rate (SR) condition than in the Regular Rate (RR) condition. A Volume condition was added to allow comparison of the effects on visual attention of a less novel change in stimuli. Our study accepted the key assumption that measurement of the location and duration of visual fixation(s) of the eyes will generally relate to the focus of a person's visual attention [25].

2. METHODS

2.1. Study design

A screen-based eye tracker was used to record participants' eye gaze behaviours while passively viewing a video-recorded model speaking single sentences aloud. Three (3) blocks were presented to all participants (within-subject, repeated measures design). Each block contained 9 novel sentences (control condition) and 9 repeated sentences (manipulated condition). Sentence stimuli order was randomized within blocks 2 & 3. The order of blocks 2 & 3 was alternated.

- Block 1 - **Repetition (Rep)**: sentences repeated unchanged (Rep A, Rep B), Rep A always presented before Rep B, 200 syllables per minute (i.e., spm)
- Block 2/3 – **Volume (Vol)**: Regular Volume, Loud Volume (+6 dB) [26], 200 spm
- Block 3/2 - **Rate**: Regular Rate (200 spm), Slowed Rate (120 spm)

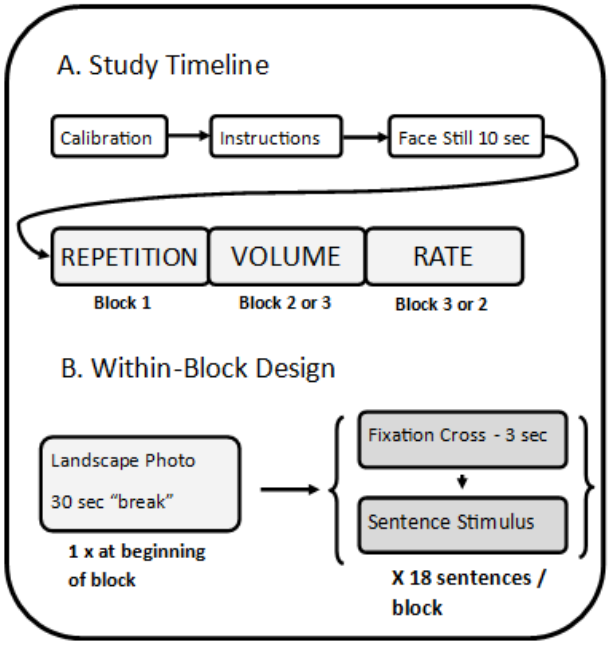


Figure 1: Depiction of study timeline and example block design.

2.2 Participants

A convenience sample of 32 young adults was recruited through a program for psychology student involvement in research at the University of Lethbridge. Twenty-five participants (15 females, 10 males, aged 19-31 years, mean = 22.2 years, sd ± 3.0) met pre-selected inclusion criteria: English as a first language, normal or corrected-to-normal vision, and no history of speech, language, learning or hearing difficulties (self-reported). Pre-study ethics approval was obtained from the Research Ethics Board of the University of Alberta. Participants

provided voluntary written consent prior to testing and completed a demographic questionnaire.

2.3. Stimuli design & recording

Sentence stimuli were designed to be linguistically consistent (sentence structure, length, gender-neutral subject) and were produced with natural, but neutral intonation.

The slowed speech rate (SR) was achieved through differentiation in the prolongation of vowel sounds [4], to maintain the relative syllable durations consistent with spoken English pronunciation and stress patterns. Continuing consonant sounds were not increased in duration. All speech rates were standardized to include 0.5 seconds of pre-speech inspiration time for each sentence.

Sentence stimuli were video recorded on a Logitech web cam (1080 x 720p x 60fps), with online monitoring of audio recording levels and a consistent mouth-to-microphone distance of 10" (Shure Beta 87A cardioid microphone). The speaker was a 20-year-old female undergraduate student with clear articulation.

Using Tobii Pro Lab software (v.1.181) [27], 3 dynamic Areas of Interest (AOIs) were placed on each video stimulus (Eyes, Mouth, Face).



Figure 2. Designated Areas of Interest (AOIs): Eyes, Mouth, Face. Marked facial areas were not visible to participants.

2.4 Eye tracking recordings and analyses

Eye tracking recordings were performed with a Tobii Pro Fusion screen-based eye tracker and Tobii Pro Lab presentation software (v.1.181) [27]. Participants were seated at a standard distance of 65cm. Participants viewed one of four randomly selected timelines. Preset cut-off values for validation of calibration accuracy and precision (<1.0 deg) and percentage of eye gaze samples (>70%) were met for all participants included in data analysis. Using the Tobii Pro Lab preset "I-VT Fixation" filter, measurements were taken of the Total Duration of Fixations (summed) for the mouth Area of Interest (AOI) over the Time of Interest (TOI) (i.e., "Talk Time" or speech portion) for each stimulus. The amount of time participants spent viewing the mouth was compared across blocks and conditions.

To account for differences in sentence stimuli durations, the Total Duration of Fixation values were normalized to a proportion (Total Duration of Fixations to the Mouth [in ms] / Duration of "Talk Time" per stimuli [in ms] = Proportion of Total Duration of Fixations to Mouth per stimuli).

3. RESULTS

Data analyses were performed in R, v. 4.1.2 [28]. The dependent variable of interest (Normalized Total Duration of Fixation of Eye Gaze to the Mouth) was analyzed using a Bayesian hierarchical linear mixed-effects model [29], with Main effects = Block and Condition, and Random factors = Participant (intercept and slope) and Sentence (slope). A normal distribution was used to model the dependent variable and set weakly informative priors for the parameters (normal), (0,1).

(1) Total_Duration_Fixation_Mouth (Normalized) ~ Block * Condition + (1 + Block + Condition | Participant) + (1 | Sentence)

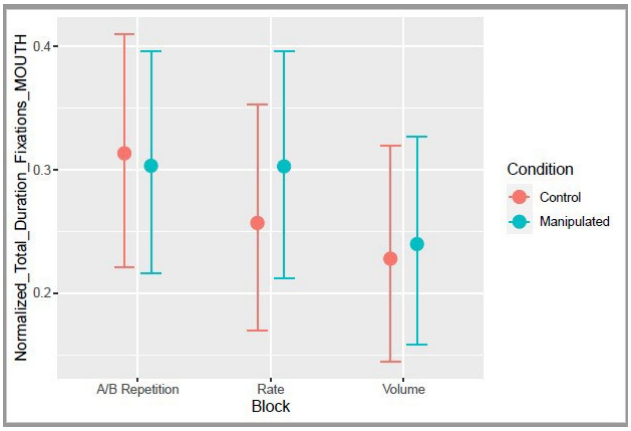


Figure 3: Conditional Effects plot: Amount of predicted change in the (duration normalized) Total Duration of Fixations to the Mouth Area. All Blocks and Conditions (Scale: 0.0 to 1.0, Control – L, Manipulated - R)

The model converged successfully with Rhat values of <1.01 (four chains, 2000 iterations). The significant interaction between Block (Rate) and Condition (Manipulated) of 0.06 (**Table 1**) indicates that when an "average" participant hears a slowed rate of speech, they look 6% more to the mouth area

than when they hear a regular rate of speech (Control condition).

Population Parameter Estimates: Change in Eye Gaze Duration to Mouth				
	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept (Block Rep) (Rep A)	0.32	0.05	0.22	0.41
BlockRate (Regular Rate)	-0.06	0.03	-0.12	0.01
BlockVol (Regular Volume)	-0.09	0.03	-0.14	-0.03
BlockRep: ConditionMan (Rep B)	-0.01	0.02	-0.05	0.03
BlockRate: ConditionMan (Slow Rate)	**0.06**	0.03	**0.00**	**0.11**
BlockVol: ConditionMan (Loud Volume)	0.02	0.03	-0.03	0.08

Table 1: Estimated parameters of the model

The non-significant Block x Condition interaction estimates in the Loud Volume condition (2% increase) and Repetition B condition (1% decrease) indicate that not all selected experimental manipulations of the speech signal result in a meaningful change in eye gaze to the mouth.

The estimated Bayesian Conditional R2 was 0.471, explaining 47% of the variability in the full model. The estimated Marginal R2 value, (excluding the Random Effects of Participant and Sentence) was 0.017. The large difference between the R2 values highlights the importance of the Random Effects in the prediction of the model parameters. With the addition of Participant as a Random Factor, the reference Intercept estimate of 0.32 (Population Level) is estimated to vary substantially (i.e., Intercept estimate for Group Level Effect of Participant: sd Estimate = 0.23, Est Error = 0.04). In contrast, only a minimal effect of including Sentence as a Random Factor was noted (Group Level Effect of Sentence: sd Estimate = 0.02, Est Error = 0.01).

4. DISCUSSION

In confirmation of our hypothesis, preliminary analysis of model results showed an increase in the duration of eye gaze to the mouth in the Slowed Rate (SR) condition compared to the Regular Rate (RR) condition for the average participant. We interpreted this as evidence of increased visual attention to the mouth due to the novel prosody and movement characteristics of a clinically slowed rate of speech (~120spm).

The smaller magnitude of the visual attention response to the mouth in the Loud Volume (LV) condition was thought to have occurred because a significantly slowed speech rate was likely more novel or surprising to listeners than a perceptual "doubling" of volume (+6dB) [26], especially when the volume changes were not paired with sentence content indicating urgency or strong emotion. In addition, the increased volume of the auditory signal was not associated with any significant changes in the movement characteristics of the speaker's mouth.

No increase in focus to the mouth was noted in the second presentation of sentences in the Repetition block. This was expected, as the unchanged sentences would likely have maintained a neutral or decreased salience value for participants.

To interpret the magnitude of the changes in duration of eye gaze fixations to the mouth (6%) with Slowed Rate we considered the context of typical eye gaze behaviours, as the amount of spontaneous gaze to the eyes is generally greater than to the mouth [30]. The preference for eye-directed gaze may be related to factors such as physiological visual biases [31] or the significance of eye contact in social interactions [32]. According to Thompson, et al. [30], there is also a strong, difficult to inhibit, natural tendency to fixate first on the eyes. This "first fixation" bias towards the eyes may be easier to sustain, rather than change to another measurement area or AOI (such as the mouth). As such, even a relatively small increase in eye gaze to the mouth, as shown in the model results, may arguably be interpreted as relevant.

Finally, with respect to potential benefits of a slowed rate of speech for modelling of speech production by clinicians, the study results are complex to interpret. Although the increase in attention to the mouth is positive, the individual variability indicated by the difference between the Conditional and Marginal R2 values suggests it may be difficult to predict how a particular individual might change their eye gaze in response to a slowed rate (SR) of speech. Future analysis of the idiosyncratic eye gaze responses of the participants in our study to SR may give insight into the breadth of personal responses to a slowed rate of speech, with individualization of clinical models to follow.

5. CONCLUSION

Study results indicated a small increase in the duration of eye gaze fixations to the mouth for typical young adult participants, when passively observing a video model of a speaker using a significantly slowed rate of speech. Our preliminary findings provide initial evidence for potential benefits of a slowed rate of speech in enhancing speech perception and/or speech production through spontaneous eye gaze to the mouth.

6. REFERENCES

[1] K. M. Yorkston, M. Hakel, D. R. Beukelman, S. Fager. 2007. Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review. *J. Med. Speech-Lang. Pathology* 15, 1-26.

[2] A. K. Bothe, J. H. Davidow, R. E. Bramlett, R. J. Ingham. 2006. Stuttering treatment research 1970-2005: I. Systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. *Amer. J. of Speech-Lang. Pathology* 15, 321-341.

[3] A. S. Mefferd, M. S. Dietrich. 2020. Tongue- and jaw-specific articulatory changes and their acoustic consequences in talkers with dysarthria due to Amyotrophic Lateral Sclerosis: Effects of loud, clear, and slow speech. *JSLHR (Online)* 63, 2625-2636.

[4] E. Boberg, D. Kully. 1994. Long-term results of an intensive treatment program for adults and adolescents who stutter. *Journal of Speech and Hearing Research* 37, 1050-1059.

[5] A. S. Mefferd. 2019. Effects of speaking rate, loudness, and clarity modifications on kinematic endpoint variability. *Clinical Linguistics & Phonetics* 33, 570-585.

[6] Y. Yuan, Lleo, Y., Daniel, R., White, A., Oh, Y. 2021. The impact of temporally coherent visual cues on speech perception in complex auditory environments. *Front. in Neuroscience* 15.

[7] A. Yi, W. Wong, M. Eizenman. 2013. Gaze patterns and audiovisual speech enhancement. *JSLHR* 56, 471-480.

[8] A. Chauvin, N. A. Phillips. 2021. Bilinguals show proportionally greater benefit from visual speech cues and sentence context in their second compared to their first language. *Ear and Hearing*.

[9] Y. Zheng, A. G. Samuel. 2019. How much do visual cues help listeners in perceiving accented speech? *Appl. Psycholing.* 40, 93-109.

[10] R. Holt, L. Bruggeman, K. Demuth. 2020. Visual speech cues speed processing and reduce effort for children listening in quiet and noise. *Appl. Psycholing.* 41, 933-961.

[11] J. B. Frtusova, N. A. Phillips. 2016. The auditory-visual speech benefit on working memory in older adults with hearing impairment. *Front. Psychol.* 7, 1-14. Art no. 490.

[12] C. K. Keintz, K. Bunton, J. D. Hoit. 2007. Influence of visual information on the intelligibility of dysarthric speech. *Amer. J. of Speech-Lang. Pathology* 16, 222-234.

[13] B. Banks, E. Gowen, K. J. Munro, P. Adank. 2021. Eye gaze and perceptual adaptation to audiovisual degraded speech. *JSLHR* 64, 3432-3445.

[14] A. Cristià. 2010. Phonetic enhancement of sibilants in infant-directed speech. *J. Acoust. Soc. Amer.* 128, 424-434.

[15] H. M. Julien, B. Munson. 2012. Modifying speech to children based on their perceived phonetic accuracy. *JSLHR* 55, 1836-1849.

[16] W. Secord. 1981. *Eliciting Sounds: Techniques for Clinicians*. The Psychological Corporation.

[17] A. K. Namasivayam, A. Huynh, F. Granata, V. Law, P. van Lieshout. 2021. PROMPT intervention for children with severe speech motor delay: a randomized control trial. *Ped. Res.* 89, 613-621.

[18] O. Räsänen, S. Kakouros, M. Soderstrom. 2018. Is infant-directed speech interesting because it is surprising? – Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition* 178, 193-206.

[19] C. Ranganath, R. Gregor. 2003. Neural mechanisms for detecting and remembering novel events. *Nature Rev. Neuroscience* 4, 193-202.

[20] L. Itti, P. Baldi. 2009. Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295-1306.

[21] S. Kakouros, N. Salminen, O. Räsänen. 2018. Making predictable unpredictable with style - Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. *Neuropsychologia* 109, 181-199.

[22] T. Kendall. 2013. *Speech Rate, Pause and Sociolinguistic Variation : Studies in Corpus Sociophonetics*. Palgrave Macmillan.

[23] C. J. Howard, A. O. Holcombe. 2010. Unexpected changes in direction of motion attract attention. *Attention, Perception and Psychophysics* 72, 2087-2095.

[24] A. von Mühlenen, M. I. Rempel, J. T. Enns. 2005. Unique temporal change is the key to attentional capture. *Psychol. Sci.* 16, 979-986.

[25] B. T. Carter, S. G. Luke. 2020. Best practices in eye tracking research. *Int. J. Psychophysiol.* 155, 49-62.

[26] E. Sengpiel. Decibel levels and perceived volume change. http://www.sengpielaudio.com/calculator-levelchange.htm (accessed March 28, 2022).

[27] *Tobii Pro Lab*. (2014). Tobii Pro AB, Danderyd, Sweden.

[28] *R: A Language and Environment for Statistical Computing*. (2022). R Foundation for Statistical Computing, Vienna, Austria. [Online]. https://www.R-project.org/

[29] P. Bürkner. 2018. Advanced Bayesian multilevel modeling with the R package brms. 10, 395-411.

[30] S. J. Thompson, T. Foulsham, S. R. Leekam, C. R. G. Jones. 2019. Attention to the face is characterised by a difficult to inhibit first fixation to the eyes. *Acta Psychologica* 193, 229-238.

[31] K. Guo, C. Smith, K. Powell, K. Nicholls. 2012. Consistent left gaze bias in processing different facial cues. *Psychol. Res.* 76, 263-269.

[32] T. Foulsham, J. T. Cheng, J. L. Tracy, J. Henrich, A. Kingstone. 2010. Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition* 117, 319-331.

THE REPRESENTATION OF ONSET SIBILANT VARIANTS IN TAIWAN MANDARIN: AN ERP STUDY

Yun-Han Hsu, Janice Fon & Chia-Lin Lee

Graduate Institute of Linguistics, National Taiwan University, Taiwan
r09142006@ntu.edu.tw; jfon@ntu.edu.tw; chialinlee@ntu.edu.tw

ABSTRACT

Previous behavioural studies on Mandarin sibilant variants showed visual words are processed slower following their auditory forms of infrequent variants than frequent or canonical forms. This study further examined the processing timeline of variant frequency using the ERP technique. Forty-two native speakers viewed 120 sibilant-initial words following their auditorily presented canonical or variant pronunciations. Results showed larger N400 responses in the infrequent variants than the frequent ones during the auditory prime, suggesting greater effort in lexical access for the former. Responses to the subsequent visual words showed enhanced P2 for both variants relative to the canonical forms, but no such difference was found for N400 responses, indicating elevated attention following variants but similar semantic facilitation from the audio stimuli for all forms after a short delay. Our findings thus suggested the variant frequency effect emerges during lexical access prompted by variant pronunciation, and is quickly annihilated after visual presentation.

Keywords: speech perception, variant frequency, ERPs, retroflex sibilants, dental sibilants

1. INTRODUCTION

In Taiwan, aside from the official language, Taiwan Mandarin, other languages such as Min, Hakka, and Austronesian languages are used as well. Among them, Min is the second largest language and over 70% of the population in Taiwan are various degrees of Mandarin-Min bilinguals [1]. Due to the frequent contact between the two languages, the phonological system of Mandarin is inevitably affected by Min [2, 3]. In Mandarin, there are three retroflex sibilants (/tʂ/, /tʂʰ/, /ʂ/) and three dental sibilants (/ts/, /tsʰ/, /s/). On the other hand, Min only has the dental series but not the retroflex ones. As a result of language contact, many Mandarin speakers show various degrees of merging between retroflex and dental sibilants [2, 3], resulting in deretroflexion of retroflexes [4]. This is a rather common variation, and could be found in both male and female speakers in casual speech [5]. However, males generally made smaller sibilant contrasts than females, but inconsistent cross-regional comparisons were observed, in which males from the southern regions made larger sibilant contrasts than ones from the northern regions [6]. Interestingly, there is also a reverse merging process reported, i.e., merging dentals with retroflexes, termed hypercorrection [7]. This likely occurs because speakers associate a negative connotation with dental realisations and thus substitute dentals with retroflexes even when the former is in fact the canonical form. Hypercorrection is mainly found in formal or read speech [7].

Previous behavioural studies showed inconsistent findings in the processing of pronunciation variants. Some studies claimed a canonical form advantage, which predicts that canonical forms are always processed more efficiently than variant forms. Studies on the medial /t/ deletion in English (e.g., *center* pronounced as *cenner*) showed that canonical forms (e.g., *center*) have advantages over variant ones (e.g., *cenner*) in processing even though /t/-deleted variants occur fairly frequently in that language [8]. However, other studies found that frequency plays an important role. High frequency pronunciations could facilitate spoken word processing even when they are not canonical. Sumner and Samuel [9] examined the r-less variant in the New York City (NYC) dialect and found that NYC residents process the frequent r-less variants faster than the canonical r-full forms. Results of Chuang and Fon [10] on Mandarin sibilants also sided more with Sumner and Samuel [9], and showed that visual words are processed slower following infrequent audio variants than frequent variants or canonical forms, while frequent variants are processed as quickly as the canonical forms.

Although frequency is found to be the major factor in variant processing, the intrinsic drawback of a behavioural study automatically prevents us from pinpointing when exactly such an effect occurs on the processing timeline, as reaction time is generally a reflection of a rather late stage of processing. It is thus unclear whether lexical

activation is successful immediately after one hears the audio variant or some time is required before the correct lexical entry is accessed. This study thus examined when the variant frequency effect emerges during the early stages of processing using the event-related potential (ERP) techniques. By adopting a cross-model paradigm adapted from Chuang and Fon [10], participants first heard a word with one of the three sibilant variants, followed by a visual presentation of the word or an unrelated word to examine how variant frequency affects the early stage of processing. In order to make further comparison with Chuang and Fon [10] within a similar time course of the pronunciation variant processing, the interstimulus interval between the auditory and visual stimuli in this study was equivalent to the design in Chuang and Fon [10].

To examine how variant frequency affects the early stage of processing, we measured the three ERP components, the N400 time-locked to both auditory and visual stimuli, the visual P2, and the late positivity component (LPC) time-locked to the visual targets. The N400, a broad component that goes negatively and peaks around 400 ms post-stimuli, is usually associated with lexical access and semantic facilitation [11]. It also reflects the ease of retrieving phonological information at the lexical level. When more effort is required in the retrieval, larger N400 effects are elicited [12-14]. The visual P2 component, a positively going waveform that peaks at around 200 ms post-stimuli over anterior electrode sites, is known to be modulated by attention and is enhanced when the targets are relatively infrequent [15]. Finally, the LPC component reflects the explicit detection of semantic anomalies [16] and is thought to involve in conscious perception and recollection of an item [17]. For example, in a semantic congruency detection task, LPC was found to be larger for the unfamiliar dialectal pronunciations than for familiar ones [18]. We predicted that these ERP components (P2, N400, and LPC) would be modulated by the two different rules, the frequent deretroflexion rule [4] and the less frequent hypercorrection one [7].

2. METHOD

2.1. Participants

Forty-two native speakers of Taiwan Mandarin (21M, age=20-30) were recruited. None of them reported any history of hearing or neurological disorders.

2.2. Materials

Stimuli were 60 retroflex-initial (e.g., *she^{4}bei^{4}* 'equipment') and 60 dental-initial bisyllabic words (e.g., *zi^{1}xun^{4}* 'information'). Each word included three conditions: canonical, variant and unrelated pronunciations. Canonical forms were pronounced according to the dictionary pronunciation, while variant forms were pronounced by merging sibilants through deretroflexion or hypercorrection. Retroflexes were pronounced as dentals, while dentals were pronounced as retroflexes. Unrelated words were words that were not related to the sibilant-initial words in any way. An analogous set of 60 stop-initial pseudo variants were also included as a control baseline. This control baseline was designed to ensure that participants did not distinguish the variant pronunciations from the canonical ones based on merely similar phonological features they shared. In fact, only one phonological feature differed between canonical and variant pronunciations for the sibilant groups, [±retroflex]. Pseudo variants were thus created from these bisyllabic words by changing unaspirated stops to aspirated ones (e.g., *dian^{4}nao^{3}* 'computer' becomes '*tian^{4}nao^{3}*'), which was also one phonological feature apart. We expected that the maximum effect would be observed in the pseudo stop variants, compared to the sibilant groups. Table 1 is an example of the stimulus design. All priming stimuli were recorded by the third author, a female phonetician, and were later synthetically adjusted to one second using Praat [19].

	Auditory Prime			Visual Target
	Canonical	Variant	Unrelated	
R	*she^{4}bei^{4}* 'equipment'	*se^{4}bei^{4}*	*jie^{2}gou^{4}* 'structure'	*she^{4}bei^{4}* 'equipment'
D	*zi^{1}xun^{4}* 'information'	*zhi^{1}xun^{4}*	*guo^{2}nei^{4}* 'domestic'	*zi^{1}xun^{4}* 'information'
S	*dian^{4}nao^{3}* 'computer'	*tian^{4}nao^{3}*	*fa^{1}xian^{4}* 'discover'	*dian^{4}nao^{3}* 'computer'

Table 1: The stimulus design (R: retroflex; D: dental; S: stop).

2.3. Procedure

Three stimulus lists were created for the cross-modal experiment so that the same visual target occurred on a list only once. Word frequency was matched across lists. Participants were divided into three groups and were randomly assigned to one list. The experiment was presented via E-prime 3.0. Figure 1 illustrates the experimental procedure. To ensure that participants attend to the stimuli, they

were asked to memorise both the auditory and visual stimuli presented in each block. A memory task was conducted right after the end of each block. Before the audio-visual experiment, a picture-naming task was conducted to record participants' production of the sibilants because the two sets of retroflex and dental sibilants are in free variation with each other and speakers often show various degrees of merging. Therefore, the picture-naming task was to check the speakers' degrees of merging and see whether there is an interaction between their production and perception.

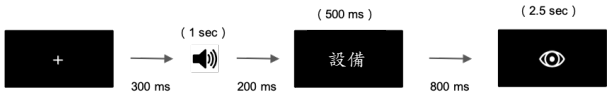


Figure 1: The flowchart of the experimental procedure.

2.4. EEG recordings

The electroencephalography (EEG) measurement was recorded using a 32-channel electrode cap via a *SYNAMPS2* amplifier with a 1000 Hz sampling rate. All electrodes were referenced to the left mastoids online and re-referenced to the average of the left and right mastoids offline. Four additional electrodes placed above/below the left eye and at the outer canthus of each eye in a bipolar montage measured the electrooculogram such as blinks and eye movements. Impedances were kept below 5 kΩ for all electrodes. The continuous EEG was segmented into epochs. Trials contaminated by artifacts were rejected before averaging. Artifact-free ERPs were averaged by stimulus type after subtraction of the 200 ms pre-stimulus baseline. Prior to measurement, ERPs were digitally filtered with a bandpass filter of 0.1-30 Hz.

3. RESULTS

For data time-locked to the onset of the second syllable of the auditory stimuli, a repeated measures ANOVAs with the within-subjects factor VARIANT (canonical/variant) was conducted for each sound on the mean amplitudes of the N400 responses measured between 300 and 600 ms in the central-posterior regions. The N400 responses on the retroflex showed no significant canonical-variant difference. For the dental and stop conditions, the N400 effects showed canonical-variant differences [dental: $F(1, 40) = 4.76$, $p < .05$; stop: $F(1, 40) = 13.72$, $p < .001$]. Results in Figure 2 showed that the N400 effect was larger in the variant condition than the canonical one in both dental and stop.

In addition, differences in mean voltage between the canonical and the variant condition were significantly different among the three sounds [$F(2, 80) = 6.00$, $p < .01$]. Post hoc analyses showed that the canonical-variant differences among three sounds were significantly different ($p < .001$). Figure 3 showed that the canonical-variant differences in stops was larger, followed by the ones in dentals and the differences were the least in retroflexes.

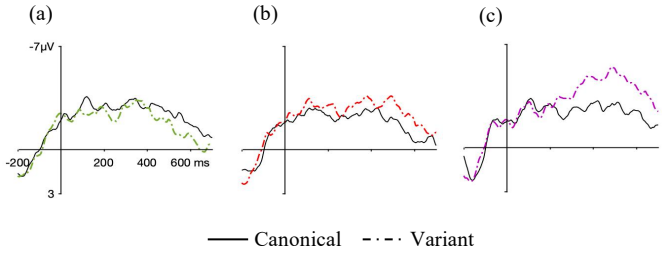


Figure 2: Grand average waveforms for the three conditions in the (a) retroflexes, (b) dentals and (c) stops, respectively, at electrode CZ when time-locked to the onset of the second syllable of the auditory stimuli.

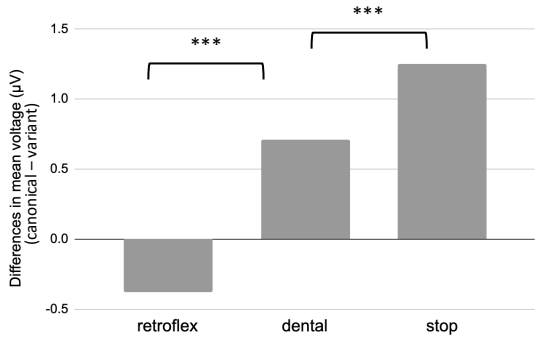


Figure 3: Canonical-variant differences in mean voltage between canonical and variant conditions when time-locked to the onset of the second syllable of the auditory stimuli (*** $p < .001$).

For data time-locked to the onset of the visual target, separate repeated measures ANOVAs with the within-subjects factor CONDITION (canonical/variant/unrelated) were conducted for each of the three components—the P2, N400, and LPC, for each sound. The P2 was computed within the 150-250 ms window in the frontal areas. The mean amplitudes of N400 were computed between 300 and 600 ms in the central-posterior regions. Lastly, The LPC was measured between 600 and 1000 ms in the central-posterior areas. Figure 4 respectively presents the waveforms time-locked to the onset of the visual stimuli in the (a) retroflexes, (b) dentals and (c) stops at electrodes FCZ, CZ, CPZ when time-locked to the onset of the visual stimuli.

For the P2 component, the main effect of CONDITION was significant among the three sounds [retroflex: $F(2, 82) = 6.20$, $p < .01$; dental: $F(2, 82) = 4.21$, $p < .05$; stop: $F(2, 82) = 4.25$, $p < .05$], with enhanced P2 effect in the variant condition across all three sounds (retroflex: $p < .001$; dental: $p < .01$; stop: $p < .001$), as shown in Figures 4(a), (b) and (c) at around 200 ms at electrode FZ.

Next, among all three sounds, the main factor CONDITION was significant in the N400 responses [retroflex: $F(2, 82) = 19.02$, $p < .001$; dental: $F(2, 82) = 15.21$, $p < .001$; stop: $F(2, 82) = 33.24$, $p < .001$]. Post hoc comparisons reported that the N400 effects were larger in the unrelated condition than the related condition across all sound types (*ps* < .001). However, the N400 canonical-variant difference was only significant in the stop group ($p < .001$), as shown in Figure 4(c).

The LPC responses showed that the conditional differences were only found in the stops [$F(2, 82) = 18.80$, $p < .001$]. Post hoc analyses showed that the LPC responses elicited larger responses in the variant condition than the other two conditions ($p < .001$), as shown in Figure 4(c).

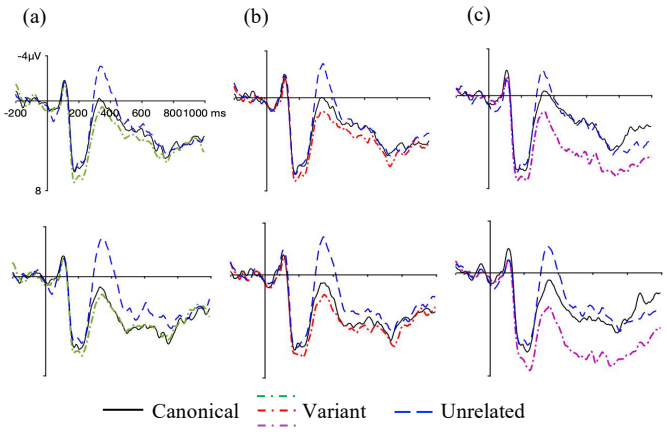


Figure 4: Grand average waveforms for the three conditions in the (a) retroflexes, (b) dentals and (c) stops, respectively, at electrodes FZ (upper row) and CZ (bottom row) when time-locked to the onset of the visual target.

Furthermore, based on the sibilant production, participants were divided into two groups, the unmerged (N=24, 10M) and the merged group (N=18, 11M). The unmerged group clearly distinguished retroflex from dental sounds while the merged group tended to replace the retroflex with dental sounds. No significant group differences were found in any of the factors.

4. DISCUSSION

This study examined the variant frequency effect of Taiwan Mandarin sibilant variants during the early processing stage. Results from the auditory stimuli showed that compared to the canonical conditions, the N400 responses were largest for the pseudo variants during auditory priming, followed by the infrequent variants. In contrast, no significant N400 difference was observed between the frequent variants and the canonical conditions. This suggests that, upon hearing the words, infrequent and pseudo variants required greater effort in lexical access than frequent ones, demonstrating a variant frequency effect in the early stage of processing.

When data was time-locked to the subsequent visual stimuli, responses showed enhanced P2 for both frequent and infrequent variants relative to the canonical forms, suggesting elevated attention. It indicated that regardless of variant frequency, variants were more "marked" compared to canonical pronunciations when a visual form was present, indicating a possible source for the canonical form advantage. Interestingly, no N400 difference was found between canonical and variant conditions, indicating similar semantic facilitation from the audio stimuli for both types of variants and their canonical forms after a short delay. In contrast to the data patterns from the existing frequent and infrequent variants, responses to visual words following the pseudo-variants showed enhanced P2 and LPC, reflecting attention and conscious recollection of the prime-target relationship. Our findings thus suggest that the variant frequency effect observed in the behavioural study by Chuang and Fon [10] emerges during the early stage of lexical access prompted by variant pronunciation. Moreover, this variant frequency effect is quickly annihilated after visual presentation, suggesting that the lexical access is successful in the later stage of processing. In addition, the null group effect suggested that there is a discrepancy between production and perception. Listeners can make a distinction between canonical and variant pronunciations even for the merged group. Put all the results together, our findings provided a possible reconciliation for the inconsistent patterns seen in previous behavioural literature.

5. REFERENCES

[1] S.-f. Huang, *Language, Society, and Group Awareness: Studies in Taiwan Sociolinguistics*. Taipei: Crane Publishing, 1994.

[2] C. C. Kubler, *The development of Mandarin in Taiwan: A case study of language contact*. Cornell University, 1981.

[3] C. C. Kubler, "The influence of Southern Min on the Mandarin of Taiwan,"

Anthropological Linguistics, vol. 27, no. 2, pp. 156-176, 1985.

[4] J. Fon, "The Phonetic Realizations of the Mandarin Phoneme Inventory: The Canonical and the Variants," in *Speech Perception, Production and Acquisition*: Springer, 2020, pp. 11-36.

[5] C. C. Lin, "A sociolinguistic study of the use of the retroflex sounds in Mandarin in College students in Taiwan," *Bulletin, College of Arts and Letters, National Central University,* pp. 1-15, 1983.

[6] Y. Chuang, "An acoustic study on voiceless retroflex and dental sibilants in Taiwan Mandarin spontaneous speech," *Unpublished master's thesis. National Taiwan University, Taipei, Taiwan,* 2009.

[7] K. S. Chung, "Hypercorrection in Taiwan Mandarin," *Journal of Asian Pacific Communication,* vol. 16, no. 2, pp. 197-214, 2006.

[8] M. A. Pitt, "The strength and time course of lexical activation of pronunciation variants," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 35, no. 3, p. 896, 2009.

[9] M. Sumner and A. G. Samuel, "The effect of experience on the perception and representation of dialect variants," *Journal of memory and language,* vol. 60, no. 4, pp. 487-501, 2009.

[10] Y.-Y. Chuang and J. Fon, "Cross-dialectal Perception of Voiceless Dental and Retroflex Sibilant Variants in Taiwan Mandarin," in *ICPhS*, 2011, pp. 496-499.

[11] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP)," *Annual review of psychology,* vol. 62, p. 621, 2011.

[12] N. Dumay, A. Benraiss, B. Barriol, C. Colin, M. Radeau, and M. Besson, "Behavioral and electrophysiological study of phonological priming between bisyllabic spoken words," *Journal of Cognitive Neuroscience,* vol. 13, no. 1, pp. 121-143, 2001.

[13] Y. Liu, H. Shu, and J. Wei, "Spoken word recognition in context: Evidence from Chinese ERP analyses," *Brain and language,* vol. 96, no. 1, pp. 37-48, 2006.

[14] A. S. Desroches, R. L. Newman, and M. F. Joanisse, "Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity," *Journal of Cognitive Neuroscience,* vol. 21, no. 10, pp. 1893-1906, 2009.

[15] S. J. Luck and S. A. Hillyard, "Electrophysiological correlates of feature analysis during visual search," *Psychophysiology,* vol. 31, no. 3, pp. 291-308, 1994.

[16] A. J. Sanford, H. Leuthold, J. Bohan, and A. J. Sanford, "Anomalies at the borderline of awareness: An ERP study," *Journal of Cognitive Neuroscience,* vol. 23, no. 3, pp. 514-523, 2011.

[17] M. Misra and P. J. Holcomb, "Event–related potential indices of masked repetition priming," *Psychophysiology,* vol. 40, no. 1, pp. 115-130, 2003.

[18] J. C. Bühler, F. Waßmann, D. Buser, F. Zumberi, and U. Maurer, "Neural processes associated with vocabulary and vowel-length differences in a dialect: An ERP study in pre-literate children," *Brain topography,* vol. 30, no. 5, pp. 610-628, 2017.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. http://www.praat.org/ (accessed July 23, 2021)." 2021.

PERCEPTION OF MALAYALAM THREE-WAY STOP CONTRAST AMONG AMERICAN ENGLISH SPEAKERS

Meghavarshini Krishnaswamy, Natasha Warner

University of Arizona
mkrishnaswamy@arizona.edu, nwarner@arizona.edu

ABSTRACT

Acoustic proximity of novel segments to existing categories affects perception. Competition also erodes relationships between a category and familiar segments; affecting identification. This study examines the pattern of perception of three novel Malayalam coronal stops (dental, alveolar, retroflex) by American English listeners. Participants responded to three-way forced-choice questions on consonant identity in VCV sequences. Accuracy and proportion of responses were analysed before and after exposure to the contrast.

Alveolars were more likely to be identified correctly post-exposure, but not retroflexes, and the trend for dentals was significantly different from alveolars. Accuracy trends were not significant post-exposure. Retroflexes were significantly less confusible, and participants responded 'retroflex' fewer times for audio without them after exposure.

Participants do not start with a biased towards any stop, they could tell the difference between retroflexes and the other two stops- 'coronal stop' exist as one category, its perception has directionality and an observable trend.

Keywords: Novel contrast, Malayalam, speech perception, three-way contrast

1. INTRODUCTION

This study examines the perception of the Malayalam three-way geminate coronal stop contrast (dental /t̪ː/, alveolar /tː/ and retroflex /ʈː/) among adult speakers of American English with no prior exposure.

1.1. Coronal three-way stop contrast in Malayalam

Malayalam is a Dravidian language with a large phonemic inventory, mainly spoken in the Indian State of Kerala [1]. It exhibits a three-way contrast intervocalically among geminate stops involving the tongue-tip, blade and dorsum: dental /t̪ː/, alveolar /tː/ and retroflex /ʈː/ [2]. In this paper, "coronal contrast" refers to this three-way contrast. It occurs only with certain neighboring vowels, and there are few minimal triplets (example: /pət̪ːi/ 'snake's hood', /pət̪ːi/ 'what was said', /pəʈːi/ 'dog').

Articulatory differences among coronal stops are well-researched in this and other Indic languages. Dart and Nihalani show that dental stops are articulated as 'interdentals' or 'dentialveolars'; alveolars make a constriction in the region behind the upper teeth and about 5mm away from it, and retroflexes make a constriction well away from the alveolar ridge, with both tongue tip and underblade contact [2]. Dentals also show longer constriction duration than alveolars. Thus the dental-alveolar stop distinction is one of constriction placement and length. Retroflexes stand out in terms of their place of articulation and tongue region contact, and have been reported to show tongue curling (constriction involving the sublaminal region) in other Dravidian languages [3]. At the acoustic level there are differences in Voice Onset Time (VOT) [4, 5], F3 lowering near retroflex stops (in Tamil, Hindi and Punjabi), and F2 raising (Tamil) [3]. Further, preceding vowels also show colouring and lengthening [3]. Dutta & Redmon et al. also show differing VCV coarticulatory patterns [6]. Thus, there are potential acoustic cues to the three-way distinction.

1.2. Coronals in English

Among British and American varieties of English, word-medially, /t/ is realised as a flap ('water'), an oral stop ('attention, stem'), or a glottal stop, depending on stress and environment. Dental stops only occur before an interdental fricative ('width', 'breadth'), and are rather rare [7]. Thus English speakers frequently hear alveolar oral stops and rarely hear a context-sensitive dental allophone.

Based on acoustic correlates and point of articulatory contact, the closest candidate for retroflection in English would be the rhotic and the rhotic schwa. While rhotics in American English can be bunched or retroflex [8], both have third formant transitions that are similar to Indic retroflection [3]. Coronal stops may coarticulate

with a following rhotic (‘attraction’, ‘truck’, ’trash’), creating allophonically retroflex stops.

The three coronal stops would thus constitute a novel contrast for native English listeners that uses familiar acoustic cues. This experiment investigates whether the three-way distinction can be perceived by naive English listeners at all, and whether they have a bias toward any of the categories.

1.3. Novel Contrast Perception and the PAM Model

The Perceptual Assimilation Model (PAM) [9] discusses the perception and categorization of novel sounds outside of the native sound inventory of a language, based on proximity to native sound categories. A novel sound may be perceived as a ‘good’ or ‘bad example’ of a native category, it may fall outside of any native sound category, or be perceived as outside the speech stream.

PAM proposes four possibilities for how a distinction between two sounds of another language might relate to native sound categories:

1. ‘Single category (SC)’: novel sounds match one native category equally well. For example, Japanese listeners perceive both English /l/ and /r/ as poor examples of the Japanese category /r/ [10].
2. ‘Different category goodness (CG)’: one novel sound is a better example of a native category than the other. Example, Japanese perception of the English /w/-/r/ contrast shows better recognition of /r/ than of /w/ [11].
3. ‘Two categories (TC)’: novel sounds are binned in different categories, thus improving identification of both. For example, voicing contrast in Zulu lateral fricatives is perceived as two different categories by English speakers [10].
4. ‘Undefined space (UC/UU)’: one or both segments may fall outside any perceptual category. For example, perception of Zulu clicks by English listeners [12].

Among the first three, the TC shows the highest level of discrimination, while SC shows the lowest. CG falls somewhere in between. [10].

Training on Indic two-way coronal contrasts has shown SC or CG type of perception. Experiments on Hindi dental-retroflex stop contrast comparing infants and adults from an English-speaking background showed that adults performed only at chance, both with and without training [13]. Comparison between performance of American English monolinguals, strong Bengali-English and Spanish-English bilinguals showed that bilinguals exhibited better levels of discrimination (that is, more participants moved to the CG and TC type of perception) for alveolar-retroflex contrast among laterals, rhotics and nasal stops after long training sessions (for non-native contrasts) [14].

Previous work on phonemic perception has been limited to two-way contrasts (e.g. dental vs. retroflex) and did not include Malayalam alveolar stops. The current study analyzes a three-way contrast within a small articulatory space.

1.4. Research Questions

The first research objective is to test whether the three Malayalam stops will be perceived differently at all. If the contrast is perceived, the second question is whether any of the three stops is more detectable than another. Third, we can examine whether the listeners exhibit bias toward any of the categories.

2. MATERIALS AND METHODS

2.1. Materials

This perception experiment used 65 nonsense minimal triplets, in the pattern $V_1C\text{ː}V_2$ (vowel-coronal geminate-vowel). For example, the stimuli for the vowel context ‘iCːo’ are [iṯːo, itːo, iʈːo]. The vowel set used in this experiment is [ə, i, o, u, e, æ, ɔ]. In this paper, ‘ə’ symbol is used for the central mid vowel, which can be both stressed and unstressed in Malayalam. The last two are not phonemic in Malayalam, but were included to ensure that the study had a sufficient number of trials. The stimuli were recorded by a phonologist whose L1 is Malayalam to ensure that both the coronal contrast and the [æ-a, ɔ-o] distinction could be produced, while keeping the prosody of all words as similar as possible. The list of words included thirteen real words. Since the listeners have no knowledge of Malayalam this is not expected to affect results.

2.2. Participants

37 Native speakers of American English were recruited through Mechanical Turk (MT) (9 female, 15 male, others unknown), with a 90% or above rating on previous studies, and previous participation in at least 10 MT tasks. Of these 13 reported speaking only English, 16 reported studying Spanish or other languages in school, and 7 chose not to answer. One participant reported Malayalam and English as their L1 and their data was excluded.

2.3. Procedure

The experiment was run using Finding Five [15]. It began with a description of tongue shapes for the three-way contrast (see Appendix) and the symbols used for them, and also included instructions for keypress responses. The description was included because the native English listeners would otherwise have no information on which response symbol meant what. After the introduction, three trials tested their retention of the three symbols, with questions on tongue gestures associated with each symbol (ex."choose the symbol for the sound made when the tongue cuts off the flow of air behind the teeth"). The next three trials were practice trials identical to the test trials discussed below.

The task was three-way alternative forced choice (AFC), in which the participant listened to a stimulus, and chose one of three options displayed on the screen. Listeners first responded to 60 pre-exposure test items, after having received only the written description of the categories described above. A filler stimulus containing an unrelated consonant was included after every 6 stimuli to make the task less repetitive. (This pre-exposure phase was included to test another hypothesis, but will not be discussed further here, as listeners were not able to identify the categories well based on abstract descriptions and the number of pre-exposure items gave low power.) Listeners pressed a key on the keyboard to indicate their response.

Listeners were then exposed to the three-way contrast by hearing 30 items of the same type as the test stimuli, with the correct response symbol shown on the screen while they heard it, and no response required. The first 15 items were presented as triplets in the order dental, alveolar, retroflex. The second 15 items consisted of the same minimal triplets with order within each triplet being random.

After this brief exposure phase, the listeners were presented with an additional 132 post-exposure test trials in the same format as the pre-exposure trials. All vowel neighbourhoods were included for this phase, resulting in 65×3 trials. The post-exposure test phase was followed by a language background questionnaire.

3. RESULTS

Table-1 shows the post-exposure results (36 participants, 44 minimal triplets) as a confusion matrix (accuracy in the diagonal, in bold). Chance is 33%. All statistical analyses on this data was conducted using the *lme4* package in R [16]. Maximal models was attempted; i.e. if a factor was within subjects, random slopes were included for that factor by subjects, and if a factor was within-minimal-triplets, random slopes were included for that factor by minimal triplets. If the maximal model failed to converge, random slopes were removed until convergence was achieved.

Target→ *Response↓*	**Dental**	**Alveolar**	**Retroflex**	**Response Given**
Dental	**612 (38.64%)**	550 (34.72%)	508 (32.07%)	1670 (**35.41%**)
Alveolar	561 (35.42%)	**635 (40.09%)**	525 (33.14%)	1721 (**36.22%**)
Retroflex	411 (25.95%)	399 (25.19%)	**551 (34.79%)**	1361 (**28.64%**)
Total	1584	1584	1584	4752

Table 1: Confusion Matrix for post-exposure trials

3.1. Accuracy

A GLME was fitted to examine the effect of Place on accuracy (Equation (1)). Accuracy for alveolars was not significantly different from accuracy for dentals and retroflexes ($p > 0.1$ for both dentals and retroflexes). If listeners thought almost all the stimuli sounded like alveolars (bias toward alveolar), they might respond with, say, 80% alveolar, regardless of what the stimulus was. This would be reflected as high accuracy for alveolar stimuli and low accuracy for the other two categories, but would only reflect bias, not ability to perceive a difference. The lack of a significant effect of Place on accuracy suggests that any bias is small.

$$\begin{aligned} Accuracy \sim\ & Place\ Of\ Articulation\ + \\ & (1 + Place\ Of\ Articulation \mid Participant)\ + \\ & (1 + Place\ Of\ Articulation \mid Minimal\ Triplet) \end{aligned} \quad (1)$$

However this does not show if listeners can perceive the difference among the three coronal categories. For this, the next subsection investigates detectability of the three-way distinction by examining proportion of a particular response instead of accuracy.

3.2. Detectability of distinctions among coronal stops

In order to assess whether listeners can detect the difference among the coronals, the analysis assessed the number of times a listener responded 'dental' when the stimulus really had a dental stop versus when it had an alveolar or a retroflex stop. This was done correspondingly for the 'alveolar' and 'retroflex' responses. These results appear in Table-1 (distributed row-wise). For example, participants responded 'alveolar' 635 times when presented with alveolar stops, 561 times when presented with dental stops, and 525 times when presented with retroflex stops.

For the dependent variable 'alveolar response' (yes/no), a GLME was fit with place of articulation

of the stimulus as the fixed effect. Random intercepts were included for each triplet and place of articulation, and random slopes were included for each triplet. The same was done with 'dental response' and 'retroflex response' as dependent variables. In each case the category that represented the dependent variable was also used as the reference level of the factor 'place of articulation'. The model used is represented in (2):

$$Response \sim Place\,Of\,Articulation + (1 + Place\,Of\,Articulation \mid Participant) + (1 \mid Triplet) \quad (2)$$

Listeners were significantly more likely to respond 'alveolar' to alveolar stimuli than to retroflex stimuli ($p < 0.02$). However, they were not significantly more likely to respond 'alveolar' to alveolar stimuli than to dental stimuli ($p > 0.05$). This suggests that the listeners were able to hear some difference between alveolar and retroflex stimuli, but not between alveolar and dental stimuli. With 'dental' response as the dependent variable, alveolar and dental stimuli did not differ significantly ($p > 0.1$), but retroflex stimuli were significantly less likely to be identified as dental than dentals stimuli ($p < 0.05$). With 'retroflex' response as the dependent variable, participants were significantly more likely to identify retroflex stimuli as retroflex than to identify either other category as retroflex. Retroflex stimuli was also misidentified significantly fewer times as alveolars stops ($p < 0.005$) and dentals stops ($p < 0.02$). This analysis indicate that listeners could detect some difference between retroflexes and the other two categories, but could not hear a difference between alveolars and dentals.

4. DISCUSSION

This experiment was designed to examine whether a novel three-way contrast could be perceived by speakers with access to allophonic acoustic cues associated with alveolar stops, interdental stops, and retroflection in their language; and whether a short exposure to the contrast was enough to allow them to detect differences among the three categories.

The analysis with response accuracy as the dependent variable verifies that listeners do not have a strong bias either toward or away from any one of the three categories. They do not, for example, respond 'alveolar' predominantly because of that sound being most similar to their native category.

The analysis of detectability shows that even after a brief exposure to this novel three-way distinction, listeners have some significant ability to distinguish retroflexes from the other two categories. Listeners respond in the same way regardless of whether the stimulus is alveolar vs. dental, showing that they do not detect any difference between these two categories. In terms of detectibility of the three-way distinction, American English listeners show some ability to distinguish retroflexes from the other two categories, collapsing this three-way distinction to a two-way one. This could be due to the availability of robust acoustic cues that speakers are already sensitive to in the retroflection of rhotics (since low F3 is a strong cue for American English /r/). Familiarity with retroflection among sonorant rhotics may be helpful in the identification of retroflection among stops. Alternatively, assimilation of English /t/ to a following rhotic (as in 'attraction, truck, trash') might be providing enough experience with a rhotic stop to allow listeners to detect that category, even though that assimilation is allophonic rather than phonemic in American English.

The three-way contrast appears to show the 'different category goodness' (CG) pattern of assimilation, with all three stops mapping to one category, but retroflexes being less confusable. One possibility here is that retroflex stops are perceived to be closer to the exemplar 'coronal' category, even with a brief exposure. The CG prediction and results are further supported by the low accuracy results for each category, and lower detectablility for alveolars and dentals than for retroflexes. This result is consistent with previous research on Indic two-way coronal contrasts (dentals vs retroflexes, alveolars vs retroflexes).

There are interesting acoustic and lexical effects that could be playing a role here. The low lexical frequency of allophonic interdentals in English (e.g. 'width, breadth' and very few other words) could make the alveolar/dental distinction more difficult to learn. Acoustically, the English /tr/ sequence includes strong perceptual cues to retroflection, including increased VOT/aspiration along with low F3. This may contribute to the listeners' ability to distinguish retroflex from other categories. Future research on novel phoneme perception, especially when expanded to multiple choices for a given small perception space as in this three-way distinction, can provide us with greater insights into the interaction between acoustic cues, allophonic patterns, and amount of past exposure to a sound type.

5. REFERENCES

[1] C. Chandramouli, "Census of India 2011," *Language, States and Union Territories. New Delhi: Government of India*, pp. 13–14, 2018.

[2] S. N. Dart and P. Nihalani, "The articulation of Malayalam coronal stops and nasals," *Journal of the International Phonetic Association*, vol. 29, no. 2, pp. 129–142, 1999.
[3] S. Bakst, "Rhotic and Retroflexes in Indic and Dravidian," Master's thesis, Trinity College, University of Cambridge, June 2012.
[4] E. Chodroff, A. Golden, and C. Wilson, "Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. EL109–EL115, 2019.
[5] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.
[6] I. Dutta, C. Redmon, M. Krishnaswamy, S. Chandran, and N. Raj, "Articulatory complexity and lexical contrast density in models of coronal coarticulation in malayalam," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 1992–1996.
[7] D. Eddington, "Flaps and other variants of /t/ in American English: Allophonic distribution without constraints, rules, or abstractions," *Cognitive Linguistics*, vol. 18, no. 1, pp. 23–46, 2007. [Online]. Available: https://doi.org/10.1515/COG.2007.002
[8] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4466–4481, 2008.
[9] C. T. Best, "Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development," in *Developmental neurocognition: Speech and face processing in the first year of life*. Springer, 1993, pp. 289–304.
[10] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794, 2001. [Online]. Available: https://doi.org/10.1121/1.1332378
[11] C. T. Best and W. Strange, "Effects of phonological and phonetic factors on cross-language perception of approximants," *Journal of phonetics*, vol. 20, no. 3, pp. 305–330, 1992.
[12] C. T. Best, G. W. McRoberts, and N. Sithole, "The phonological basis of perceptual loss for non-native contrasts: Maintenance of discrimination among Zulu clicks by english-speaking adults and infants," *Journal of Experimental Psychology: Human perception and performance*, vol. 14, no. 3, pp. 345–360, 1988.
[13] J. F. Werker, J. H. Gilbert, K. Humphrey, and R. C. Tees, "Developmental aspects of cross-language speech perception," *Child development*, pp. 349–355, 1981.
[14] D. V. Gogoi, *Acquisition of novel perceptual categories in a third language: the role of metalinguistic awareness and feature generalization*. University of Florida, 2009.
[15] Team FindingFive, "Findingfive: A web platform for creating, running, and managing your studies in one place. findingfive corporation (nonprofit), NJ, USA," 2019.
[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

6. APPENDIX

The text used to introduce the three-way contrast to participants is as follows:

"Malayalam language has three types of stops produced with the tongue tip and tongue body. They are: Dental (tongue tip stops air-flow at the back of the teeth) , Alveolar (tongue front stops air-flow at the alveolar ridge,bumpy part behind the teeth), Retroflex (the tongue is curled such that the tongue tip stops air-flow behind the alveolar ridge)"

The following table lists words used as stimuli in the post-exposure condition. Real words are highlighted in grey.

Exposure Condition	Triplet ID	Triplet Template	Real Word?	Repeated in both exposure conditions?
post-exposure	1	ɔCɔ	N	N
post-exposure	2	ɔCo	N	N
post-exposure	3	ɔCe	N	N
post-exposure	4	ɔCu	N	N
post-exposure	5	ɔCi	N	N
post-exposure	6	ɔCɔ	N	N
post-exposure	7	ɔCæ	N	N
post-exposure	8	əCɔ	N	N
post-exposure	9	oCɔ	N	N
post-exposure	10	eCɔ	N	N
post-exposure	11	uCɔ	N	N
post-exposure	12	iCɔ	N	N
post-exposure	13	æCæ	N	N
post-exposure	14	æCə	N	N
post-exposure	15	æCo	N	N
post-exposure	16	æCe	N	N
post-exposure	17	æCu	N	N
post-exposure	18	æCi	N	N
post-exposure	19	æCɔ	N	N
post-exposure	20	æCa	N	N
post-exposure	21	əCæ	N	N
post-exposure	22	oCæ	N	N
post-exposure	23	eCæ	N	N
post-exposure	24	uCæ	N	N
post-exposure	25	iCæ	N	N
post-exposure	26	əCa	N	N
post-exposure	27	oCa	N	N
post-exposure	28	eCa	Y	N
post-exposure	29	uCa	N	N
post-exposure	30	iCa	N	N
post-exposure	31	ɔCa	N	N
post-exposure	32	aCa	N	N
post-exposure	33	aCə	N	N
post-exposure	34	aCe	N	N
post-exposure	35	aCu	Y	N
post-exposure	36	aCi	N	N
post-exposure	40	əCa	N	N
post-exposure	41	oCa	N	N
post-exposure	42	eCa	Y	N
post-exposure	43	uCa	N	N
post-exposure	44	iCa	N	N

Perception of Japanese moraic nasal in intervocalic position by Japanese native speakers and Korean learners of Japanese

Heesun Han[1], Koji Namba[2]

[1]Osaka Shoin Women's University, [2]Osaka University
Han.heesun@osaka-shoin.ac.jp, cozy.numbat.ciee@osaka-u.ac.jp

ABSTRACT

This study examines the difference in the perception of the Japanese moraic nasal (/ɴ/) by followed by a vowel between Japanese native speakers (JN) and Korean learners of Japanese (KA: advanced learners, KB: beginners). The participants were instructed to identify stimuli by choosing one of two choices (/goseɴeɴ/, /gose:eɴ/).

The results showed that regarding the stimuli in the test word /goseɴeɴ/, the JN judgment rate of /ɴ/ is higher than that of the KB. This means that KB depends on "the degree of closure" more than JN in judgments of /ɴ/. KA's judgment was similar to that of JN. On the other hand, regarding the stimuli on the test word /gose:eɴ/, KA showed the lowest percentage of judgment rate of /ɴ/ among the three groups. They correctly perceived the long vowel part as a vowel ([e]). It indicates that KA did not reach the level of acceptance by JN in terms of the acceptance of vowels as free allophones of /ɴ/ in the intervocalic position.

Keywords: Japanese moraic nasal, perception, degree of closure, variations, learning level

1. INTRODUCTION

Japanese is a language with mora units, generally structured /V/ and /CV/. In addition, there are also the so-called special morae, such as Japanese moraic nasal (hereafter /ɴ/), the geminate consonant, and long vowels. Previous studies have focused on the lengths of special morae. Although /ɴ/ is characterized by being generated as various sounds in contrast to long vowels and geminate consonants, few studies have focused on /ɴ/ sound variations.

In this study, perceptual experiments were conducted with a focus on the variation of /ɴ/. This study includes Korean learners of Japanese, to compare /ɴ/ recognition between native and Korean learners. This study will also investigate the mastery of /ɴ/ by the level of learning.

2. LITERATURE

In general, /ɴ/ is generated as a nasal consonant assimilated to the place of articulation when an obstruent follows. In Japanese phonemes, /ɴ/ has the most conditional allophones. Furthermore, while /ɴ/ is known to be followed by vowels and semi-vowels (/w, j/), the resulting sounds have been considered as nasal vowels (/ã, ĩ, ũ, ẽ, õ/). However, previous studies disagree in regard to the /ɴ/ sounds followed by vowels and semi-vowels in the intervocalic position.

There is a variety of observations about the intervocalic /ɴ/ sounds, which are nasal vowels (e.g. [ẽ]), vowels (e.g. [e]), nasalized approximants (e.g. [ũɻ]), nasalized fricatives (e.g. [ỹʒ]), voiced velar nasal([ŋ]), and voiced uvular nasal ([ɴ]) ([1], [2], [3], [4], [5], [6], [7], [8], [9], [10]). In the case of the intervocalic /ɴ/ sounds, many variations can be called a "free variant." The cause for such variations has been attributed to individual differences ([9], [11]), speech styles ([5], [7]), and speech rates ([5], [7]). Meanwhile, according to [9], there are no clear regional differences in the articulations of /ɴ/ sounds.

When generated as vowels or nasal vowels, as with words like "/seɴi/ fiber" and "/se:i/ sincerity", they can be confusing even among Japanese native speakers, so it is necessary to judge the words depending on the context ([4], [8]). However, as described above, the intervocalic /ɴ/ is generated not solely as nasal vowels (or vowels) but includes variations such as completely closed nasal stop [ŋ] (or [ɴ]) in the oral cavity, and the intermediates between these two. In order to avoid confusion, it is effective to generate /ɴ/ as a nasal stop, however, this gives the impression that the sounds are being emphasized ([12]). In other words, "the degree of closure (manner of articulation)" is important for the recognition of /ɴ/, while it may make it feel unnatural.

In the Korean language, on the other hand, it is necessary to distinguish clearly the place of articulation (/m, n, ŋ/) on the nasal codas. Furthermore, these nasal codas need to be

completely closed in the oral cavity. Thus, regarding the perception of /ɴ/ sounds followed by a vowel, there is the possibility that differences between Japanese native speakers and Korean learners of Japanese may be due to attentiveness toward the degree of closure.

Therefore, this study incorporates native Korean speakers alongside Japanese speakers, as Korean provides a prime example of a language with contrastive phonemes on the nasal coda.

3. METHOD

In Japanese phonology, /ɴ/ and long vowels have been thought of as contrasts. However, there are cases when variations in /ɴ/ obscure phonetic differences between phonemes despite the apparent phonological differences between the two.

Accordingly, this study employs meaningful words that contain "/ɴ/ + vowel (/goseɴeɴ/: five thousand yen)" and add "vowel + vowel (/gose:eɴ/: encouragement)" consisting of five morae. Seven Japanese native speakers (S1, S2, S3, S4, S5, S6, S7) were asked to produce various sounds at the third mora of /goseɴeɴ/ in four speech rates. The four speech rates consist of "A (slow)," "B (normal)," "C (relatively fast)," and "D (fastest)," in order of the rates. Each utterance was measured in length and respectively classified as intended speech rate. Speakers were allowed to practice reading the sentences before recording. While the test words were embedded in carrier sentences when they were recorded, they were omitted from the sounds tested to prevent interference in judgments due to the influence of context ([4], [8]).

The listeners consisted of groups of 30 native speakers of Japanese (JN), 60 native speakers of Korean; 30 beginner learners of Japanese (KB), and 30 advanced learners (KA). KB, in their 20s, studied Japanese for less than one year and majored in Japanese at a Korean university, and 30 KA, also in their 20s, have passed the JLPT Level 1 and studied at a Japanese university for more than one year.

The total number of stimuli is 56 sounds (two test words × seven speakers × four speech rates). The stimuli were randomly ordered, and listeners were asked to listen to each stimulus only once. After being given the instructions, listeners had 10 minutes to practice and become familiar with the process. The perception experiment took about 50 minutes, including breaks. A pause of three seconds was placed between each stimulus. Listeners were asked to choose one of two choices (/goseɴeɴ/ five thousand yen, /gose:eɴ/ encouragement) to identify whether the stimuli are /ɴ/ or the long vowel. Two types of test forms (type1: /goseɴeɴ/ five thousand yen or /gose:eɴ/ encouragement, type2: /gose:eɴ/ encouragement or /goseɴeɴ/ five thousand yen) were prepared to offset any influence that the choice order may have. The choices were written in Japanese. The experiments were conducted in a silent room, with participants left alone for the test duration.

4. ANALYSIS OF STIMULI

We asked three researchers specializing in phonetics or phonology (one Japanese Native speaker, one Korean Native speaker and one English Native speaker) to listen to the stimuli and describe them in IPA. As a result, the stimuli are categorized into three groups as shown in Table 1 ((1) [ẽ], [ĩ], (2) [ũɰ̃], and (3) [ɴ]). Only six sounds were described as nasal vowels.

As mentioned in Section 2, some studies ([5], [7]) claimed that the sounds of /ɴ/ vary according to the speech rate. However, there was no obvious consistent correlation found between these sound variations (nasal vowels or [ũɰ̃] or [ɴ]) and the speech rate (the length of each stimulus) for the twenty-eight stimuli of /goseɴeɴ/ in this study (r2=-0.082). This result disagrees with the previous studies ([5], [7]). This study shows that the distributions of the stimuli vary across the speaker. Eventually, various sounds with different degrees of closure were collected in this study.

Furthermore, we observed the spectra of the third mora in the test word (/goseɴeɴ/). Regarding nasal vowels ([ẽ], [ĩ]), almost all frequencies keep their energy evenly like in stimulus B from S6 (figure 1). The same tendency was observed in the third mora /e/ of another test word /gose:eɴ/. On the other hand, stimulus B from S4 (figure 2) which was categorized [ũɰ̃], shows the sudden decays of certain frequency regions (e.g. around 2,000Hz). These phenomena are called anti-formants. Anti-formants are frequency regions in which the amplitudes of the source signal are attenuated because nasal cavities absorb energy from the sound waves when the oral cavity is closed. This is known as a characteristic of nasal consonants ([13]). Although stimulus D from S2 was described as [ɴ] in figure 3, it displays a spectrum that resembles that of the [ũɰ̃] in figure 2. Thus, the results show that while anti-formant possibly contributes to the /ɴ/ recognition, it may not be the only cue for perception.

	[ẽ] or [ĩ]	[ɯ̃]	[ɴ]
S1		A, C, D	B
S2	A, B	C, D	
S3		A, D	B, C
S4			A, B, C, D
S5		A, B	C, D
S6	A, B, D	C	
S7	B	A	C, D

Table1: IPA description of the third mora of /gosenɴ/

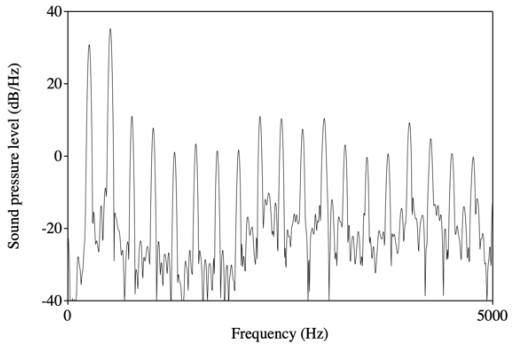


Figure 1: [ẽ] the spectra of the third mora in /gosenɴ/

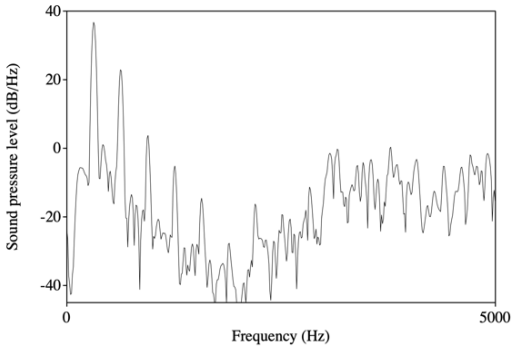


Figure 2: [ɯ̃] the spectra of the third mora in /gosenɴ/

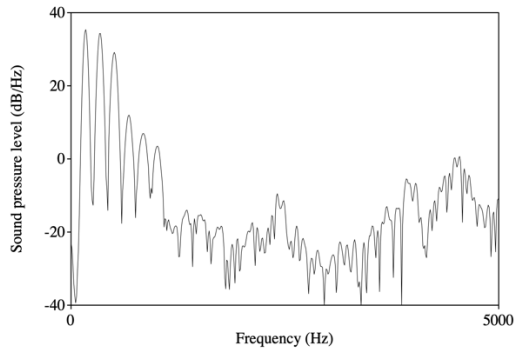


Figure 3: [ɴ] the spectra of the third mora in /gosenɴ/

5. RESULTS

5.1 Number of correct answers for test words

Regarding the test word /gosenɴ/ (Table 2), JN correctly judged /gosenɴ/ in 639 instances out of a total of 840 (seven speakers x four times x thirty listeners), while KB only correctly judged in 496 instances. KA correctly judged /gosenɴ/ in 609 instances out of a total of 840 KA showed a tendency similar to JN. Significant differences were found among the three groups (one-way ANOVA (p=0.05), F (2, 2517) =11.206, ***p<0.001***). Multiple comparisons among the three groups were conducted based on the Benjamini-Hochberg method (q=0.017). The correct judgment of JN and KA are significantly higher than that of KB (JN vs. KB: ***p<0.001***, KA vs. KB: ***p<0.001***). There was no significant difference between JN and KA (JN vs. KA: p=0.109).

On the other hand, regarding the test word /gose:ɴ/ (Table 3), significant differences were found among the three groups (one-way ANOVA (p=0.05), F (2, 2517) =12.300, ***p<0.001***). JN correctly judged /gose:ɴ/ in 555 instances out of a total of 840 (seven speakers x four times x thirty listeners). KB and KA correctly judged /gose:ɴ/ in, respectively, 641 and 623 instances out of a total of 840. KA and KB showed similar tendencies, with no significant differences between them (p=0.325). The judgment of JN was significantly lower than that of KB and KA (JN vs. KB: ***p<0.001***, JN vs. KA: ***p<0.001***).

Summarizing these results, the recognition of the two test words as /ɴ/ to the stimuli was higher for JA (924 out of 1680), KA (826 out of 1680), and KB (695 out of 1680), in that order. However, in terms of the seven speaker's intentions for the test words (the original meaning of each word), KA had the highest percentage of correct responses among the three groups.

	/ɴ/	long vowel
JN	**639**	201
KB	**496**	344
KA	**609**	231

Table 2: Recognition of /gosenɴ/ (Max 840)

	/ɴ/	long vowel
JN	285	**555**
KB	199	**641**
KA	217	**623**

Table 3: Recognition of /gose:ɴ/ (Max 840)

5.2 Correlation between degree of closure for /ɴ/ and its judgment

In this section, we investigated the difference in the judgment rate for /ɴ/ in the third mora of the test word /goseɴeɴ/ depending on the degree of closure ((1) [ẽ], [ĩ], (2) [ũ̯], and (3) [ɴ]).

As shown in figure 4, the judgment rate for /ɴ/ approached 100% as the degree of closure for /ɴ/ increased in all groups. However, even for /ɴ/ sounds with insufficient closure ([ẽ], [ĩ], [ũ̯]), JN had a significantly higher /ɴ/ judgment rate than KB ($p<0.001$). Regarding [ũ̯], KB had a significantly lower recognition as /ɴ/ than JN and KA (Table 4).

Thus, it indicates that the sensitivity of closure in recognition of /ɴ/ varies depending on the native language and the learning level of the target language.

6. DISCUSSION AND CONCLUSION

In the perception of stimuli of the two test words, JN judged /ɴ/ more frequently than the learners, and judged vowels less frequently than the learners. Particularly, in comparison with KB, JN had higher rates of judgment as /ɴ/ even for sounds with insufficient closure ([ẽ], [ĩ], [ũ̯]). This means that KB depends on "the degree of closure" more than JN in judgments of /ɴ/. On the other hand, the KA judgment rate had results similar to the JN. However, regarding the stimuli on the test word /goseːeɴ/, KA showed the lowest percentage of judgment rate of /ɴ/ among the three groups. In other words, they correctly judged the stimuli ([goseːeɴ]) as intended /goseːeɴ/, and they correctly perceived the long vowel part as a vowel ([e]). This indicates that KA did not reach the level of acceptance by JN in terms of the acceptance of vowels as free allophones of /ɴ/ in the intervocalic position.

Complete closure makes it easier to recognize stimuli as /ɴ/ for both JN and the learners, while a sound closer to a vowel with no restriction is perceived as /ɴ/ more naturally with the intervocalic position to JN. This result supports the view of [14].

In contrast, the Korean language has three nasal codas (/m, n, ŋ/), which have independent places of articulation, and places of articulation must be strictly maintained to distinguish them. The same opinion has been reported that English speakers do not perceive /ɴ/ like the English nasal coda, according to [15].

In conclusion, we found it more difficult for Korean speakers to recognize /ɴ/, which varies widely depending on "the degree of closure." Furthermore, it indicates that the sensitivity of closure in recognition of /ɴ/ varies depending on the native language and the learning level of the target language.

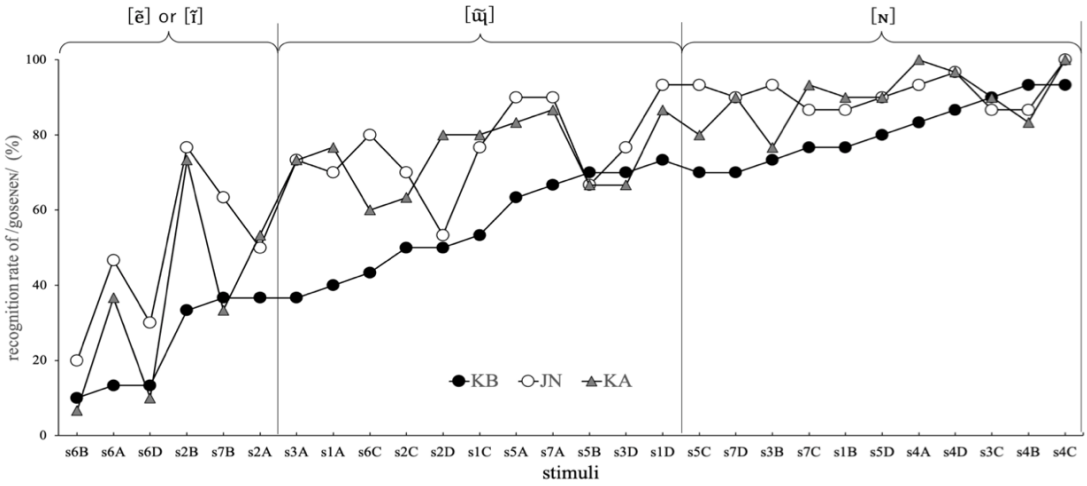


Figure 4: Recognition rates of /ɴ/ for stimuli /goseɴeɴ/ in each classification

	JN	KB	KA	one-way ANOVA (p=0.05)	Benjamini-Hochberg method (q=0.017)		
					JN vs. KB	JN vs. KA	KB vs. KA
[ẽ], [ĩ] (Max 180)	86 (48%)	43 (24%)	64 (36%)	F (1, 537) = 11.600, $p<0.001$	***≤.001***	0.057	0.042
[ũ̯] (Max 330)	252 (76%)	185 (56%)	247 (75%)	F (1, 987) = 20.530, $p<0.001$	***≤.001***	0.903	***≤.001***
[ɴ] (Max 330)	301(91%)	268 (81%)	296 (90%)	F (1, 987) = 9.423, $p<0.001$	***0.001***	0.932	***≤.001***

Table 4: Comparison between three listener groups in each classification for /ɴ/ sounds

7. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP22K00529 and JP19K13179.

8. REFERNCES

[1] Hattori, S. 1951. *Onseigaku [Phonetics]*, Tokyo, Japan: Iwanami Shoten.

[2] Onuma, Y., K. Otsubo and O. Mizutani. 1979. *Nihongo onseigaku [Japanese Phonetics]*, Kuroshio Shuppan,

[3] Muraki M. and N. Nakaoka, Hatsuon to sokuon: Eigo, chuugokugo washa no baai [The moraic nasal and the geminate consonant: Cases of English native speakers and Chinese native speakers]. 1989. In: Sugito, M. Ed., *Kouza Nihongo to nihongo Kyouiku dai3 kan [Series in Japanese and Teaching Japanese as a Foreign Language: Number 3]*. Meiji Shoin, 139–177.

[4] Kurosaki, N. 2002. A Study of a Nasal Sound Succeeded by a Vowel. *Kanagawa University studies in language* 25, 11-22.

[5] Toki, S. 2006. Gendai no onseigaku/oninron [Phonetics and phonology of modern Japanese]. In: Nitta et. al. (eds), *Nihongo yousetsu [Essentials in Japanese Linguistics]*. Hitsuji Shobo, 129-168.

[6] Vance, T. J. 2008. *The sound of Japanese*. Cambridge University Press.

[7] Saito, S. 2013. *Nihongo onseigaku. (kaichouban) [Japanese Phonetics (Revised Edition)]*. Sanseido.

[8] Uwano, Y. 2014. Fun'iki > funiki no henka kara onitenkan ni tsuite kangaeru [A Thought on the Metathesis from Change of the Sounds between 'Fun'iki' and 'Fuinki'], In: *Hokkaido Hougen Kenkyuukai 40 Shuunen Kinenronbunshuu [Journal for 40th Anniversary of Association of Dialectology in Hokkaido Prefecture]*. Hokkaido Hougen Kenkyuukai, 8-19.

[9] Hashi, M., A. Kodama, T. Miura, S. Daimon, Y. Takakura and R. Hayashi. 2014. Articulatory Variability in Word-Final Japanese Moraic-Nasals: An X-ray Microbeam Study. *Journal of the Phonetic Society of Japan, Vol. 18. Number 2*, 95-105.

[10] Matsui, M. 2018. Nihongo tokusyuhaku onso no youso to kouzou ni tsuite [On the elements and the structures of special moraic phonemes in Japanese], *in Theoretical and Applied Linguistics at Kobe Shoin, Vol. 21*, pp.105-150.

[11] Maekawa, K. 2019. A real-time MRI study of Japanese moraic nasal in utterance-final position. *Proc. 19th ICPhS* Melbourne, 1987-1991.

[12] Sudo, J. 2007. The Japanese Interjection Un: From Its Meanings and Functions to an Analysis of Its Phonetic Features. *Journal of the phonetic Society of Japan. Vol. 11, Number 3*, 94-106.

[13] Fujimura, O. 1962, Analysis of nasal consonants, *J. Acoust. Soc. Am. Vol. 34*, 1865-1875.

[14] Kawakami, S. 1987. Nihongo no iwayuru bionsetu shin no jittai [Fact aoubt the so-caled Japanese Nnasal syllabic consonants]. *The bulletin of the Phonetic Association of Japan.* 185, 18-21.

[15] Nozawa, T. and S. Cheon. 2012. The Identification of Nasals in a Coda Position by Native Speakers of American English, Korean and Japanese. *Journal of the Phonetic Society of Japan.*16:2, 5-14.

TESTING THE SOUND-DRIVEN HYPOTHESIS: AN ONLINE RATING EXPERIMENT ON THE PHONAESTHETICS OF CONSTRUCTED LANGUAGES

Christine Mooshammer, Qiang Xia

Humboldt-Universität zu Berlin, Germany
christine.mooshammer@hu-berlin.de

ABSTRACT

The sound-driven hypothesis [1, 2] assumes that aesthetic judgements of languages are based on how they sound in terms of their intrinsic sound properties. For languages that are known and recognized, social factors and cultural norms play a more important role than their sound structure. In this study, the sound-driven hypothesis is tested by an online rating experiment in which 14 constructed languages, mostly from the Fantasy and Science Fiction genre, are rated regarding their pleasantness on a seven-point Likert scale. The ratings show significant correlations with several phonetic and phonological characteristics, e.g. syllable structure and sonority related measures. In order to test the dependency on the linguistic background we compared listeners from two syllable-timed tone languages, Cantonese and Mandarin Chinese, with listeners from two stress-timed intonational languages, English and German. The results indicate a stronger agreement on negative judgements whereas pleasantness depends on the L1 of the listeners.

Keywords: language attitude, phonaesthetics, typology, sonority, constructed languages

1. INTRODUCTION

Why do French and Italian sound beautiful and pleasing to most people whereas German and Arabic sound harsh and unpleasant [3, 4]? This difference in aesthetic attitude towards languages has long been discussed. Giles, Bourhis, and Davies [1] proposed the 'inherent value hypothesis' and the 'imposed norm hypothesis' to account for the causes of these language attitudes. The former hypothesis suggests that some languages have phonetic and phonological properties that are intrinsically more pleasant than others, while the latter presumes that aesthetic judgements are the result of a complex of social, cultural and political associations with the speakers of a given language. Van Bezooijen [2] suggested 'sound-driven hypothesis', 'context-driven hypothesis' and 'norm-driven hypothesis' as more precise terms. Ample evidence shows that listener's attitudes towards a language are, to some extend, biased by its imposed social prestige and cultural norms [1, 3, 5]. At least for the better-known languages that are easy to recognize, it is not straightforward to test whether the aesthetic judgement of a language is truly sound-driven. Therefore, the current study uses so-called constructed languages, i.e. languages that are invented and, in our case, used in Fantasy and Science Fiction genres. Since they are mostly unknown and unintelligible for the general public, the influence of imposed norms and subconscious associations to its speakers can be minimized in terms of language attitudes. Constructed languages (conlangs) are therefore an ideal test field to examine the sound-driven hypothesis.

Several studies have shown that phonetic and phonological characteristics contribute to aesthetic judgements of languages. Based on a sample of 16 European languages Reiterer et al. [4] found that languages with a higher sonority index [6] and a larger vocalic share are judged more positively than languages with a low sonority index and more consonants. Furthermore, a wide pitch range and frequent stressed syllables contribute to positive language perceptions [7, 8].

Certain sound-symbolic features also affect the perception of languages. For example, Crystal [9] investigated the phonological characteristics of English words that are regarded as sounding beautiful by poets and readers in newspaper polls. He found that front vowels and sonorants occur more frequently in pleasant-sounding words relative to words used in everyday conversations (see also [7] for an overview). According to Stockwell [10], guttural consonants convey a certain harshness. The production of pharyngeal and glottal sounds has been associated with negative emotions due to the similarity and the anatomical closeness with choking and coughing [11, 12]. These

sound-symbolic and phonological properties are an important design feature of conlangs [13, 14]. For example, Tolkien's Elvish languages Sindarin and Quenya [15], intended to sound beautiful, are indeed composed of a higher vocalic proportion, a higher percentage of open syllables, more sonorants, front vowels and consonants [11, 16, 17] compared to conlangs designed for evil or aggressive fictional peoples, such as Dothraki [14], Klingon [18] and Orkish [15].

The aims of this study are to test the 'sound-driven hypothesis' with languages that are unknown and unintelligible for most people, thereby minimizing the potential effect of imposed norms. Therefore, fourteen conlangs from different genres were rated in an online rating study. The second aim is to investigate whether the aesthetic judgements are affected by phonetic and phonological characteristics of the conlangs used in this study. Third, in order to test whether this hypothesis is valid for typologically divers languages, we compare listeners from two syllable-timed tone languages, Cantonese and Mandarin Chinese [19], with listeners from two stress-timed intonation languages, English and German.

2. METHODS

2.1. Materials

In this experiment, 14 constructed languages were tested. Adûnaic, (Neo-)Khuzdul, (Neo-)Orkish, Quenya and Sindarin from the Middle-earth universe designed by J.J.R. Tolkien and featured in his work *The Lord of the Rings* [15, 20] as well as by David Salo for the film adaptation; Atlantean and Klingon by Marc Okrand [21, 18, 22]; Gaulic Vulcan [23]; Dothraki by David Peterson [14]; Na'vi by Paul Frommer [24]; Fjerdan by David Peterson and Christian Thalmann [25]; Kesh by Ursula Le Guin [26]; and the unpublished conlangs Horn and ʕUiʕuid by Dominique Bobeck. According to the designers' intention, Quenya and Sindarin are supposed to sound pleasant and elegant, while Dothraki, Klingon, Horn and Orkish should sound harsh, unpleasant and evil [14, 22, 15]. For each of the 14 conlangs, three sentences were recorded that were at least 10 syllables long and did not consist of any well-known buzzwords, such as *Khaleesi* in Dothraki. They were recorded by a female and a male speaker in a neutral voice without emotional involvement or sound effect at a 44100 Hz sampling rate and then downsampled to 22050 Hz for better presentation online.

2.2. Procedure

Online experiments with instruction in the four natural languages were conducted on the browser-based Percy platform [27]. Prior to the experiment participants filled out a questionnaire with information, for example about their age, gender, native language. Then participants were asked to rate the stimuli on three seven-point Likert scales: *pleasantness* (pleasant vs. unpleasant), *goodness* (good vs. evil), and *peacefulness* (peaceful vs. aggressive), solely based on their personal impression. Only the pleasantness scale will be considered here. Each stimulus could be played twice. The order of experiment stimuli and the order of three rating scales were randomized for each participant. They were offered an optional break after the 28th stimulus. The fictionality of the constructed languages was highlighted in the title of experiment 'Assess fantasy languages', so that participants were able to see it throughout the entire experiment. In total, 56 stimuli (14 conlangs x 2 sentences x 2 speakers) were rated. After finishing rating, participants had the option to listen to one additional stimulus of each conlang and guess which conlang was being played. The results are not considered here. The entire experiment took about 20 minutes. The link was distributed via email lists and social media.

2.3. Participants

After excluding non-native speakers and participants who showed a standard deviation of less than 0.5 for their ratings, the data from 22 Cantonese (YUE), 63 Chinese Mandarin (CMN), 40 English (ENG) and 91 German (GER) speakers are considered here.

2.4. Phonetic and phonological parameters

For calculating the sonority of each constructed language, we used the sonority index developed by [6]. Each speech sound group was assigned a sonority value, according to its manner of articulation, voicing and vowel height. For example, low vowels as [a, æ, ɑ] have the maximal value of 100, followed by mid back vowels [o, ɤ] with a value of 80, while voiced and voiceless stops have the minimal value of 2. The sonority index was calculated as the mean of sonority values of the phonological transcriptions of each stimulus.

Syllable boundaries were determined by the sonority profile based on the values in [6], as occurring prior to a local sonority minimum with a following rise [17]. Based on that, other indices such

as syllable rate and percentage of open syllables were derived.

With regard to phonetic properties, voiced and voiceless intervals of each recorded stimulus were manually annotated in Praat [28]. The fundamental frequency contours were calculated by using the R package wrassp [29, 30]. The f0 ranges were set to 80–500 Hz and 50–400 Hz for the female and the male speaker respectively.

3. RESULTS

3.1. Ratings

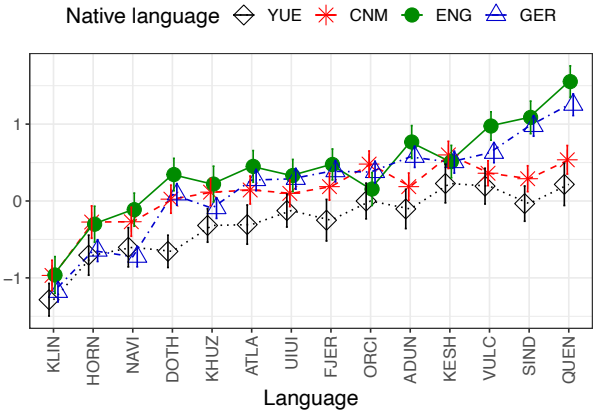


Figure 1: Mean ratings of conlangs with standard errors per listener group.

Fig. 1 shows the means and the standard errors of the pleasantness ratings for the four listener groups. As can be seen, Klingon is rated as the most unpleasant language, regardless of the L1 of the participants. The second most unpleasant conlang is Horn, except for German speaking participants who disliked Na'vi more than Horn. The most pleasant conlang for English and German speaking participants are the Elvish languages Quenya and Sindarin. Cantonese and Mandarin speaking participants chose Kesh and Quenya as most pleasant. As can be seen in Fig. 1, Cantonese and Mandarin speakers generally showed smaller positive mean ratings compared to English and German. To test whether the native language of the raters and the model speaker of the stimuli (m/f) affect the overall pleasantness ratings, a linear mixed effects model was calculated with speaker and conlang as random factors.

Fig. 2 shows the effects from the linear mixed effects model. Post hoc tests for listener group indicate that the listeners from the two tone languages Cantonese and Mandarin Chinese rate the conlangs significantly less pleasant than listeners from the stress-timed languages English and German. There are no significant differences within the language groups. Additionally, there is a significant interaction between the factors speaker and listener group: Mandarin Chinese listeners rate stimuli from the female speaker significantly more pleasant than the male speaker.

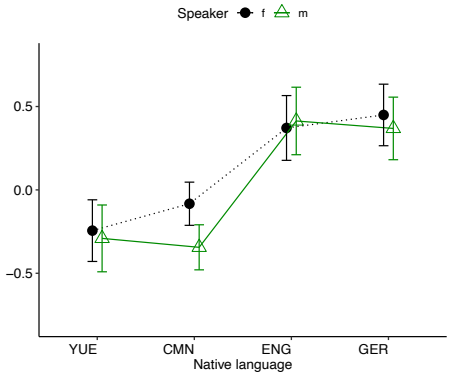


Figure 2: Effects of speaker and native language on ratings.

3.2. Relationship between ratings and phonological and phonetic characteristics

	YUE	CMN	ENG	GER
SonIndex	0.25	0.21	0.43	**0.45**
SonIndexCons	0.00	0.01	0.39	0.28
PctVowels	**0.44**	0.41	**0.48**	**0.61**
PctObstr	-0.37	-0.39	**-0.73**	**-0.67**
PctObstrOfCons	-0.23	-0.26	**-0.68**	**-0.54**
PctVoiced	**0.44**	0.42	**0.67**	**0.68**
PctOpenSyll	0.25	0.23	0.12	0.22
PctOpenNasSyll	**0.47**	**0.50**	**0.44**	**0.55**
PctComplOns	-0.34	-0.39	**-0.52**	**-0.51**
PctBackVOfV	-0.03	-0.07	-0.23	-0.09
PctGuttVelar	-0.40	-0.36	**-0.76**	**-0.65**
SyllRateN	0.18	0.23	0.23	0.22
F0N	-0.34	-0.37	-0.39	**-0.55**
IQR_F0N	-0.03	-0.01	-0.14	-0.15

Table 1: Correlation coefficients between mean ratings and phonological and phonetic characteristics of the stimuli. Coefficients with $p<0.05$ are grey shaded, coefficients with $p<0.001$ are additionally printed in bold.

Table 1 shows the Pearson correlation coefficients between the mean ratings and several sound-related characteristics. The ratings of all listener groups are positively affected by the percentage of vowels (PctVowels), the percentage of acoustically

measured voicing (PctVoiced) and the percentage of syllables that are open or closed by a nasal (PctOpenNasSyll). Negative correlations for all groups were found for percentage of obstruents (PctObstr), percentage of syllables with complex onsets (PctComplOns), percentage of velar and guttural sounds (PctGuttVel) and normalized F0 (F0N). Therefore, stimuli were rated as more pleasant if they contained more vowels and sonorants, longer stretches of voicing, empty or single onsets, empty or nasal codas, labial and alveolar consonants, and were spoken with a lower f0. The overall sonority (SonIndex), the sonority of the consonants (SonINdexCons) and the percentage of obstruents relative to the number of consonants in the stimulus (PctObstrOfCons) correlated only with the ratings for English and German speaking participants, not for Cantonese and Mandarin Chinese participants.

4. DISCUSSION

The first aim of this study was to test whether aesthetic judgments can be triggered based only on how the language sounds, without being recognized. The results of this rating study confirm the sound-driven hypothesis. Klingon, with the remarkable presence of uvular and guttural sounds among other things [18], was rated as the least pleasant conlang; whereas the Elvish language Quenya, having a high portion of vowels, open syllables and sonorants etc., was given a high pleasantness rating. It is noteworthy that the majority of participants was unfamiliar with the presented conlangs and did not recognize them. The linkage between aesthetic judgments and the intended impressions of Klingon speakers being aggressive and of Elves being pure and beautiful can hardly be established by the participants. Hence, biases by cultural norms and social traits can be excluded.

The second aim was to pinpoint which phonetic and phonological characteristics create the impression of pleasantness or unpleasantness. In general, stimuli with a high percentage of obstruents, complex syllable structures and velar or guttural consonants were rated less pleasant than stimuli with a large portion of voicing and vowels. This is in agreement with findings for natural languages. For example, Reiterer et al. [4] found a positive correlation between several phonological characteristics and ratings of 14 European languages. In their study, Italian was rated as beautiful which correlated with the percentage of vowels and a larger vocalic share whereas German had the lower ratings on beauty and had also low values of sonority and vocalic share.

The third aim was to investigate whether listeners from typologically divers languages have similar aesthetic preferences. Therefore, we compared the ratings from the two language groups, Germanic and Sinitic. The former are represented here by English and German that are stress-timed languages with an intonational prosodic system. Cantonese and Mandarin Chinese are syllables-timed languages with lexical tones and more restricted syllable structures [19]. Even though listeners from all four languages showed similar preferences, Cantonese and Mandarin Chinese listeners were more reluctant with positive judgements. As can be seen in Fig. 1, English and German listeners gave larger positive ratings than Cantonese and Mandarin Chinese listeners. One possible reason for this difference could be that none of the presented conlangs was a tonal language. Furthermore, the model speakers who produced the stimuli for this study were native speakers of German. Since the inventors of conlangs rarely describe the prosodic features of their languages, our speakers might have imposed their native German stress-timed system. This might also lead to less positive ratings from the Cantonese and Mandarin Chinese listeners. Concerning the sound-inherent phonological features, listeners from tone languages seem to be less sensitive to sonority-related measures (see Table 1) which might also account for the lower pleasantness values.

In conclusion, based on several conlangs our study confirms the sound-driven hypothesis for participants from four different languages, two tonal languages and two stress-timed languages. In a next step, we are planning to include the results from a larger variety of typologically diverse languages, such as Arabic, Russian and Italian, to further test the universality of the sound-driven hypothesis, and to examine whether phonaesthetic judgements are based on similar sets of phonological features.

5. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Christoph Draxler for his ongoing and patient support on the Percy Platform. Thanks to F. Wegmann and K. Meinhardt for recording the stimuli. We are also grateful to D. Bobeck, H. Hornecker, O. Olina, M. Walch for fruitful discussions and to the students of the class "Warum klingt Orkisch böse?" for assistance with material preparation and audio annotation.

6. REFERENCES

[1] H. Giles, R. Bourhis, and A. Davies, "Prestige speech styles: The imposed norm and inherent value hypotheses," in *Language and Society*. Mouton Publishers, 1979, pp. 589–596.

[2] R. van Bezooijen, "Aesthetic evaluation of dutch," in *Handbook of Perceptual Dialectology*, D. Long and D. R. Preston, Eds. John Benjamins Amsterdam, 2002, vol. 2, pp. 13–30.

[3] H. Giles and N. Niedzielski, "German sounds awful, but Italian is beautiful," in *Language Myths*, L. Bauer and P. Trudgill, Eds. Pinguin, 1998, pp. 85–93.

[4] S. M. Reiterer, V. V. Kogan, A. Seither-Preisler, and G. Pesek, "Foreign language learning motivation: Phonetic chill or Latin lover effect? Does sound structure or social stereotyping drive FLL?" *Psychology of Learning and Motivation*, vol. 72, pp. 165–205, 2020.

[5] A. Schüppert and G. Charlotte, "Investigating the role of language attitudes for perception abilities using reaction time," *Dialectologia*, vol. Special Issue II, pp. 119–140, 2011.

[6] J. G. Fought, R. L. Munroe, C. R. Fought, and E. M. Good, "Sonority and climate in a world sample of languages: Findings and prospects," *Cross-Cultural Research*, vol. 38, no. 1, pp. 27–51, 2004.

[7] N. H. Hilton, C. Gooskens, A. Schüppert, and C. Tang, "Is Swedish more beautiful than Danish? matched guise investigations with unknown languages," *Nordic Journal of Linguistics*, vol. 45, no. 1, pp. 30–48, 2021.

[8] R. van Bezooijen, "The relative importance of pronunciation, prosody, and voice quality for the attribution of social status and personality characteristics," in *Language Attitudes in the Dutch Language Area*, R. van Hout and U. Knops, Eds. Berlin: De Gruyter Mouton, 1988, pp. 85–104.

[9] D. Crystal, "Phonaesthetically speaking," *English Today*, vol. 11, no. 2, pp. 8–12, 1995.

[10] P. Stockwell, "Invented language in literature," in *Encyclopedia of Language & Linguistics*, 2nd ed., K. Brown, Ed. Elsevier, 2006, vol. 6, pp. 3–11.

[11] J. Podhorodecka, "Is lámatyáve a linguistic heresy? Iconicity in JRR Tolkien's invented languages," in *Insistent images*, O. Fischer, C. Ljungberg, and E. Tabakowska, Eds. John Benjamins Publishing, 2007, pp. 103–130.

[12] I. Fónagy, "Why iconicity?" in *Form Miming Meaning. Iconicity in Language and Literature*, M. Nänny and O. Fischer, Eds. John Benjamins Publishing Company, 1999, pp. 3–36.

[13] B. Beinhoff, "Why are alien languages inherently human?" *Foundation*, vol. 44, no. 122, pp. 5–19, 2015.

[14] D. J. Peterson, *The Art of Language Invention: From Horse-Lords to Dark Elves to Sand Worms, the Words Behind World-Building*. Penguin, 2015.

[15] J. R. R. Tolkien, *The Lord of the Rings: One volume*. Harper Collins, 2021.

[16] H. Elsen, "Lautsymbolik in phantastischen Sprachen," *Wirkendes Wort*, no. 1, pp. 103–119, 2019.

[17] N. Johannesson, "Quenya, the Black Speech and the Sonority scale," in *Proceedings of the First International Conference on J.R.R. Tolkien's Invented Languages*, A. Stenström, Ed., pp. 14–21.

[18] M. Okrand, *The Klingon Dictionary*, 2nd ed. Pocket Books, 1992.

[19] P. P. K. Mok, "On the syllable-timing of Cantonese and Beijing Mandarin," *Chinese Journal of Phonetics*, vol. 2, pp. 148–152, 2009.

[20] J. R. R. Tolkien, *Sauron Defeated. The End of the Third Age: The History of the Lord of the Rings, part four*, ser. The History of Middle-earth, C. Tolkien, Ed. Harper Collins, 2002.

[21] K. Shadlag. (2006) Atlantean Language Institute. [Online]. Available: https://web.archive.org/web/20140323053118/http://www.freewebs.com/keran_shadlag/

[22] M. Okrand, *The Klingon Way: A Warrior's Guide*. Pocket Books, 1996.

[23] M. R. Gardner and The Vulcan Language Institute. (1980–2004) The Vulcan language. [Online]. Available: http://surak.nu/vulcanlanguage.pdf

[24] S. Andrews. (2010) Paul Frommer sounds off on Avatar language. [Online]. Available: https://news.usc.edu/32485/Paul-Frommer-Sounds-Off-on-em-Avatar-em-Language/

[25] C. Thalmann and D. J. Peterson. (2022) Aldjen Helt: Creating the Fjerdan language for Netflix' Shadow and Bone. [Online]. Available: https://sites.google.com/unito.it/i-conlangs22/book-of-abstracts

[26] U. K. Le Guin, *Always coming home*. London: Gollancz, 2016.

[27] C. Draxler, "PercyConfigurator – Perception Experiments as a Service," in *Interspeech*, 2017, pp. 823–824.

[28] P. Boersma and D. Weenink. (2019) Praat. [Online]. Available: http://www.praat.org/

[29] L. Bombien, R. Winkelmann, and M. Scheffers, *Wrassp: An R Wrapper to the ASSP Library*, 2020, R package version 0.1.9.

[30] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

FINAL LENGTHENING IN LINE END PERCEPTION OF RE-SYNTHESIZED RECITATIONS OF GERMAN POEMS

Nadja Schauffler[1], Julia Koch[1], Nora Ketschik[1], Toni Bernhart[2], Felix Dieterle[3], Gunilla Eschenbach[3], Anna Kinder[3], Sandra Richter[3], Gabriel Viehhauser[2], Thang Vu[1], Jonas Kuhn[1]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Institute of Literary Studies/Digital Humanities, University of Stuttgart, Germany
[3]German Literature Archive (DLA), Marbach, Germany
nadja.schauffler@ims.uni-stuttgart.de

ABSTRACT

This paper presents a perception study investigating the perceptual relevance of lengthening at the end of lines with enjambment in poem recitations. An enjambment occurs when the end of a line disrupts a syntactic unit. In recitation, a speaker has different possibilities to realize this conflict between syntactic coherence on the one hand, and versification on the other. In a previous study, lengthening of the verse-final segment was found in lines both with and without enjambment. In the present study, we investigated whether this lengthening effect is perceptually meaningful in the prediction of line breaks. We re-synthesized the original recitation and manipulated the verse-final segments' duration. While predicting line breaks seemed to be an overall difficult task for listeners, we found that verse-final lengthening increases the success rate and was thus perceived by listeners as a prosodic cue to versification, even when it occurred within syntactic constituents.

Keywords: perception, speech synthesis, enjambment, poetry, digital humanities

1. INTRODUCTION

Lyric poetry is characterized by the use of lines and stanzas as primary structural units, creating an additional layer of structure and aesthetics beyond the syntax of clauses and sentences. This can lead to tensions between the linguistic form (syntax) and the poetic form (versification). Enjambments - when the syntactic unit continues past the end of a line - are an example of this tension. This mismatch between verse and syntax offers the speaker the opportunity to emphasize either the syntactic unity or the verse structure.

It has been suggested that there is a third way for speakers to serve both layers by simultaneously using prosodic cues signalling continuation as well as prosodic cues signalling discontinuation [1, 2]. In a corpus study investigating one professional speaker and his recitations of poems by the German poet Friedrich Hölderlin, Schauffler et al. [3] looked at prosodic features used by the speaker to mark the end of the line. They compared lines with enjambment to lines without enjambment and analyzed cues typically found at prosodic boundaries, namely lengthening of the phrase-final segment(s) (cf. [4, 5, 6]), the insertion of a silent pause (e.g., [7]) and F_0 reset (e.g., [8, 9]). They found that the speaker indeed reliably marks the end of a line by means of final lengthening, even in cases of enjambment, while on the other hand, silent pauses and F_0-register effects were used significantly less often in lines with enjambment.

In the study at hand, we are investigating whether verse-final lengthening is perceived by listeners as a cue marking the end of the line. It has been shown before that phrase-final lengthening is a cue in syntactically ambiguous structures, such as in the perception of bracketed lists [10]. In an online perception experiment, we have now tested whether verse-final lengthening also has a structuring effect when it occurs within syntactic units, such as in enjambments. To this end, we used the original recordings from [3] and re-synthesized them by cloning the speakers' F_0, energy and duration for each segment [11]. In a second step, we manipulated the duration of the verse-final segments (nucleus and coda) in three steps in order to investigate whether the segments' duration affects the identification of line ends.

This experiment also serves as a baseline study to explore human-in-the-loop re-synthesis as a methodological tool in hypothesis testing, as suggested in the »textklang« project.[1] Our synthesis approach allows to generate test items for perception experiments in which particular prosodic parameters can be manipulated while keeping the prosodic realisation of the original. This can be especially

advantageous for questions in the realms of digital humanities where natural data is often too limited to systematically evaluate hypotheses.

2. STIMULI

The stimuli were taken from the same corpus of Hölderlin poems as in the study by [3]. Only lines with free verse and no rhymes were considered.[2] In order to avoid traces of other prosodic phrasing cues, we picked two-line pairs with enjambment where neither a silent pause nor F_0 reset was used by the speaker (cf. [3]). Furthermore, we excluded lines in which the speaker made an audible caesura in close proximity to the line break (within two words before or after the line end). After re-synthesis (see below) the stimuli were auditorily evaluated. We excluded cases where the synthesis model produced an audible pause at the end, and one case where the re-synthesized verse-final word was apparently unknown (and thus lead to an unnatural re-synthesis). This resulted in 16 items. We divided them into two groups, 11 items with "strong" enjambments, i.e. the line-break disrupts a syntactic constituent which cannot be further decomposed, and 5 items with "weak" enjambments, where the line break takes place between syntactic phrases. See 1[3] for an example of a strong enjambment and 2[4] for an example of a weak enjambment.

(1) Den Stromgeist fern, und schaudern regt [im Nabel [der Erde]$_{NP}$]$_{PP}$ der Geist sich wieder.

(2) Doch [die ewige Sonne]$_{NP}$ goß
[ihr verjüngendes Licht]$_{NP}$ über das alternde

3. RE-SYNTHESIS AND MANIPULATION

3.1. Synthesis approach

We used the open source code and pretrained multilingual models provided by the IMS Toucan Speech Synthesis Toolkit release v2.2 [12] to create our stimuli. This toolkit implements a text-to-speech (TTS) pipeline consisting of multiple components: First, the input text is transcribed to a phonetic transcription according to IPA. Then a FastSpeech 2 based TTS model [13] generates spectrograms from the input phonemes and finally a HiFi-GAN vocoder [14] converts the spectrograms to waveform. A main advantage of FastSpeech 2 is its inherently high level of controllability due to dedicated submodules for predicting duration, F_0 and energy. The implementation in IMS Toucan comes with a lightweight reconstruction-based aligner that is trained jointly with the TTS module to extract phone durations from the reference audio [12, 15]. Thereby, alignment is directly integrated into the TTS pipeline instead of relying on an external tool for forced-alignment as seen in the original FastSpeech 2 paper [13]. To exactly re-synthesize a reference recording, we follow the approach on prosody cloning described in [15]: Duration, F_0 and energy values are extracted from the original recording. To obtain phone-wise controllability instead of dealing with individual spectrogram frames, pitch and energy values are averaged over the spectrogram frames that belong to a single phoneme according to the aligner as first proposed in [16]. These extracted prosodic values are then used to overwrite model predictions given by the duration, pitch and energy predictors respectively. Koch et al. [11] further explore the idea of overwriting prosodic values in a human-in-the-loop setup. They first clone the prosody of a reference audio and in a second step, a human expert can adjust prosodic values of individual phones using the same overwriting technique in order to make fine-grained manipulations to the prosody of the original recording. We followed this approach in our experiment.

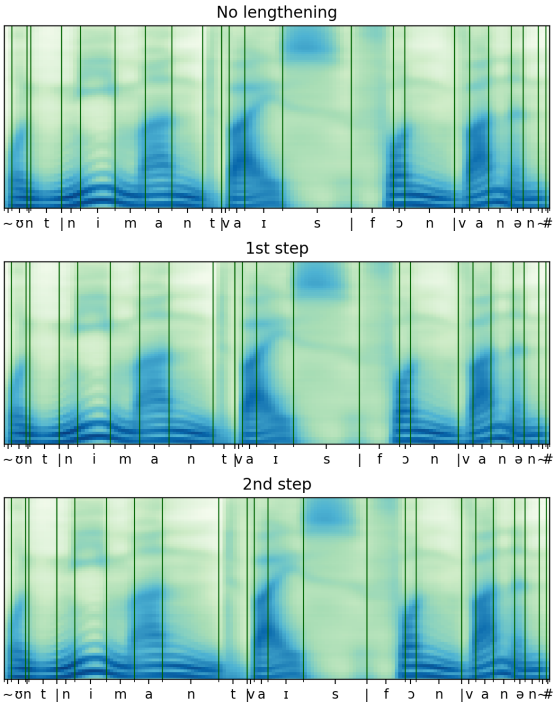


Figure 1: Spectrograms of the three different manipulations of the word *niemand* in the sentence "[...] und niemand weiß von wannen [...]"[5]. All phones in the last syllable rhyme of this word (a, n and t) are affected by the manipulations. The green bars mark phone boundaries. | denotes word boundaries.

3.2. Manipulation

We manipulated the final syllable rhyme in each line of the re-synthesized first lines since this is where final lengthening has been shown to be particularly pronounced (e.g. [4, 6, 17]). Each segment in this rhyme was manipulated in three steps. Each step is based on the original data:

no lengthening The original duration of the line-final segments was replaced by the mean duration of the respective segment in the respective syllable position (syllable-final for segments in the coda, syllable-vowel for non-final vowels) as taken from the original corpus.

1st step The segments were lengthened to the mean duration of the respective segments at the end of a line, i.e. to the mean z-score of these segments compared to non-final segments.

2nd step The segments were further lengthened corresponding to the upper quartile of the z-score of the respective final segments' duration in the original data.

Segment duration of the original data was extracted by means of the Festival [18] version of the University of Stuttgart [19]. Since our TTS models segment duration in terms of number of spectrogram frames we convert the extracted durations to spectrogram frames with the python library librosa [20] time_to_frames function according to equation 1 with $frames_{phone}$ denoting the number of frames assigned to a given phone and dur_{phone} the duration of that phone in seconds as extracted with Festival, sr being the sampling rate and hop_length meaning the number of samples between successive frames. With our settings of $sr = 16kHz$ and a $hop\_length = 256$ one spectrogram frame depicts 0.016 seconds, and hence, phone durations are modelled accurately to 16 milliseconds in our approach. To improve the audio quality of the synthetic stimuli, we perform super resolution to $48kHz$ at the stage of vocoding as shown in [21, 12].

$$(1)\quad frames_{phone} = floor(dur_{phone} * sr / hop\_length)$$

Figure 1 shows a comparison of spectrograms visualizing the different steps of manipulations. In this example the line break occurs immediately after the word *niemand* (no one). Hence, we manipulate all phones within the final syllable rhyme of this word, i.e. the phones a, n and t. While the difference for the phone a is rather small we observe a clearly visible difference in duration for n and t.

4. PARTICIPANTS AND PROCEDURE

For our online perception study, participants were recruited via several mailing lists and directed to www.soscisurvey.de where the experiment was made available. The call for participation was addressed to German L1 speakers. In a questionnaire following the experiment 40 participants indicated that they have a background in literary studies, and 32 participants indicated a linguistic background. The items with their three manipulations were distributed over three lists in a between-subject Latin Square design so that each participant only listened to one manipulation per item and to 16 items in total. The items were presented in randomized order in an online questionnaire which was implemented using SoSci Survey [22]. 112 participants took part in the experiment from beginning to end. They were randomly assigned to one of the three lists. The participants were instructed to listen to the audio. They were given the two verses in written form in one line without the line break. If more context was needed in order to make sense of the two lines parts of the preceding or following line (or both) were given in brackets. A sequence of four words was underlined and participants were instructed to choose the one they thought was the last in the first verse. For each item they could indicate how certain they are with their choice using a slider on a 0-10 scale. They could listen to the audio files as often as necessary. After the experiment, participants were asked about how difficult the task was, and about their background (see above). They were also given the possibility to comment on the experiment.

5. ANALYSIS

Four participants indicated in the comment section that they did not listen to the audio file but judged the line break according to their own reading. These participants were excluded from the analysis so that the answers of 110 participants were statistically analysed, namely 585 cases for the baseline condition without lengthening, 585 cases for the 1st step lengthening and 590 cases for the 2nd step lengthening.

The statistical analysis was performed in R 4.2.2 [23], using the function glmer from package lme4 [24]. In order to investigate the relationship between the identification of the line break and the amount of verse-final lengthening, we performed a generalized linear mixed effects analysis using the logit link function with whether the last word in the verse was correctly identified or not as the dependent

variable. As fixed factors we included *lengthening* (no lengthening, 1st step and 2nd step), and the *type of enjambment* (strong or weak). Since it may be possible that cues correlating with final lengthening may still be present in the signal, we also included the last segments' *duration in the original* as phoneme-specific z-score as fixed factor. As random factors we included intercepts for *participants* and *items*. All factors were tested for their significance by means of likelihood ratio tests by comparing the model including the effect in question to the model without it.

6. RESULTS

Most participants thought the task was rather difficult or depending on the item sometimes rather difficult and sometimes rather easy (mean 3.6 on the 1-5 point scale where 1 is very easy and 5 is very difficult).

Figure 2 gives an overview of the effect of the amount of lengthening (x-axis) on the probability of correctly recognizing the last word of the line (y-axis). We can see that the text itself hardly gives any indication as to where the line break occurs since when there was no lengthening (intercept: β=-0.57, SE=0.79, p=.47) participants correctly identified the last word in only 40% of the time. The ability to identify the last word increases significantly with the first step length manipulation to about 52% (β=0.63, SE=0.15, p<.0001) and increases further with the second step length manipulation to about 59% (β=1.25, SE=0.16, p<.0001). There was no effect for type of enjambment. Also the original duration of the line-final segments did not affect the probability to correctly identify the last word.

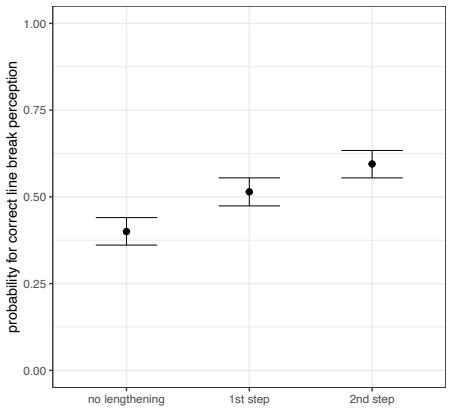


Figure 2: The probability for the indication of the correct line end by condition.

7. DISCUSSION AND CONCLUSION

Given the complexity of the poetic language, the lack of meter, and the limited context the task of finding the correct line break was rather difficult as reflected in the degree of difficulty reported by the participants on the one hand and the overall low success rate on the other hand. In order to make the task more approachable, only four options were given per line. These options may, of course, also have an effect on the probability for choosing the correct word. The other words were not manipulated in any way other than being re-synthesized so that we cannot exclude the presence of phrase-final cues in these words. Since the non-final words were always the same, differences between the length conditions can, however, be attributed to the manipulation of the line-final segments' duration. This manipulation of length had a significant effect: it facilitated the identification of the line end. The more pronounced the last segments' lengthening was the more likely listeners perceived the lengthening as a prosodic cue marking the end of the line. This result suggests that final lengthening is available to listeners as a structure-giving cue in this genre-specific context, even if it is in conflict with the syntactic coherence continuing over the line break. The finding that enjambments without a silent pause can still be prosodically conveyed by using more subtle cues such as final lengthening supports the idea of a "rhythmical performance" as suggested by [2, 1].

Surprisingly, the type of enjambment (strong or weak) did not have an effect on the line end perception even though we would have expected that weak enjambments are more likely to be identified given the weaker syntactic cohesion. With only 5 items in the weak-enjambment group, the data was, however, not balanced in this respect. Further research is needed to investigate the effects of different types of enjambments on both the production of line-final prosodic cues and their effect in perception.

With respect to re-synthesis we can say that it presents a valuable tool in experimental research by making it possible to manipulate fine-phonetic detail while cloning everything else from the original. A few participants reported that the audio sounded to be of not the best quality, or "tinny", and only one participant mentioned that some words sounded manipulated. In our next step we will work on improving the synthesis model with respect to speech quality and accuracy of prosody cloning.

8. ACKNOWLEDGEMENTS

This research is supported by funding from BMBF for the »textklang« project.

9. REFERENCES

[1] R. Tsur and C. Gafni, "Enjambment – irony, wit, emotion. A case study suggesting wider principles," *Studia Metrica et Poetica*, vol. 5, pp. 7–28, 01 2019.

[2] R. Tsur, *Poetic Rhythm: Structure and Performance – An Empirical Study in Cognitive Poetics*. Brighton and Portland: Sussex Academic Press, 2012.

[3] N. Schauffler, F. Schubö, T. Bernhart, G. Eschenbach, J. Koch, S. Richter, G. Viehhauser, T. Vu, L. Wesemann, and J. Kuhn, "Prosodic realisation of enjambment in recitations of German poetry," in *Proc. Speech Prosody*, Lissabon (Portugal), 2022.

[4] D. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1220, 1976.

[5] A. Turk and S. Shattuk-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, pp. 445–472, 2007.

[6] K. J. Kohler, "Prosodic boundary signals in German," *Phonetica*, vol. 40, pp. 445–472, 1983.

[7] M. Nespor and I. Vogel, *Prosodic Phonology*. Dordrecht, Holland: Foris, 1986.

[8] C. G. Berg, Rob van den and T. Rietveld, "Downstep in Dutch: Implications for a model," in *Papers in laboratory phonology II: Gesture, segment, prosody*, G. J. Docherty and D. R. Ladd, Eds. Cambridge: Cambridge University Press, 1992, pp. 335–367.

[9] H. Truckenbrodt, "The syntax phonology interface," in *The Cambridge handbook of phonology*, P. de Lacy, Ed. Cambridge: Cambridge University Press, 2007, pp. 435–456.

[10] C. Petrone, H. Truckenbrodt, C. Wellmann, J. Holzgrefe-Lang, I. Wartenburger, and B. Höhle, "Prosodic boundary cues in german: Evidence from the production and perception of bracketed lists," *Journal of Phonetics*, vol. 61, pp. 71–92, 2017.

[11] J. Koch, F. Lux, N. Schauffler, T. Bernhart, F. Dieterle, J. Kuhn, S. Richter, G. Viehhauser, and N. T. Vu, "PoeticTTS - Controllable Poetry Reading for Literary Studies," in *Proc. Interspeech 2022*, 2022, pp. 1223–1227.

[12] F. Lux, J. Koch, and N. T. Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2022.

[13] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *ICLR*, 2020.

[14] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, vol. 33, 2020.

[15] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *Proc. IEEE SLT*, 2022.

[16] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP*. IEEE, 2021, pp. 6588–6592.

[17] A. Schweitzer, *Production and Perception of Prosodic Events – Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart, 2010.

[18] A. W. Black, "The Festival speech synthesis system," www.cstr.ed.ac.uk/projects/festival.html, November 1997.

[19] Festival, "Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. IMS German Festival home page," www.ims.uni-stuttgart.de/phonetik/synthesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2010.

[20] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, L. Nickel, P. Friesch, M. Vollrath, and T. Kim, "librosa/librosa: 0.9.2," Jun. 2022.

[21] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li *et al.*, "DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021," *Proc. Blizzard Challenge Workshop*, vol. 2021, 2021.

[22] D. J. Leiner, "Sosci survey (version 3.4.04)," 2021. [Online]. Available: https://www.soscisurvey.de

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear mixed-effects models using S4 classes*, 2013, r package version 0.999999-2. [Online]. Available: http://CRAN.R-project.org/package=lme4

[1] http://hdl.handle.net/11022/1007-0000-0007-F6C5-5

[2] In one item, we manually exchanged the rhyming word with a phonetically similar non-rhyming word.

[3] *The stream spirit far away, and shuddering in the // Navel of the earth the spirit stirs again* (own translation).

[4] *But the eternal sun poured // its rejuvenating light over the aging* (own translation).

[5] *and no one know from where [...]*)

THE ROLE OF L2 FLUENCY IN THE PERCEPTION OF NON-NATIVE CLUSTERS BY TURKISH SPEAKERS

Stefano Canalis

Boğaziçi University
stefano.canalis@boun.edu.tr

ABSTRACT

The interplay between perception and the phonological constraints of a language plays a role in loanword adaptation; for example, native phonotactics may induce speakers to perceive 'illusory' epenthetic vowels in clusters that would be ill-formed in their language. Perceptual accuracy is also modulated by the degree of L2 verbal fluency; more fluent L2 learners tend to perceive L2 sounds more accurately than less fluent learners.

Word-initial consonant clusters are not attested in the native lexicon of Turkish. However, Turkish has borrowed many words containing such clusters in the donor languages. A learned pronunciation may preserve them, but a short epenthetic vowel between the two consonants is generally observed in a colloquial pronunciation. Two perceptual experiments showed that Turkish native speakers who are more fluent in an L2 are much more successful in discriminating word-initial CC tokens from VCV tokens than less fluent speakers are.

Keywords: Perceptual illusion, Turkish phonology, loanword adaptation, L2 fluency, vowel epenthesis.

1. INTRODUCTION

Many factors may influence the phonetic and phonological adaptation of loanwords (see [1] for an overview), but there is now strong and reliable evidence that one of them is the way segmental and phonotactic constraints of the native language shape the perception of foreign words (see e.g. [2, 3]). For instance, when speakers of Japanese – a language that does not allow clusters of obstruents – are presented with auditory stimuli such as the pseudoword [ebzo], they are often found to perceive an 'illusory' vowel between the two consonants [4]. This parallels Japanese speakers' production of loanwords, which display an epenthetic vowel in similar clusters (e.g. English *basket* > Japanese [basɯketto]). Similar 'illusory' vowels have been reported for other languages as well [3], suggesting that loanword adaptation often takes place in perception.

At the same time, since perception is supposedly affected by a speaker's native phonological system, it is plausible to expect that speakers will be able to faithfully perceive loanwords from a given L2 if they are proficient in that language, as they master its phonology as well as their L1's phonology.

Indeed, several studies suggest that perceptual accuracy of L2 sounds is modulated by the degree of L2 proficiency; for example, learners with an earlier age or acquisition and/or longer length of exposure tend to perceive L2 sounds more accurately than late learners or learners with a shorter exposure [5, 6]. L2 proficiency not only affects perception accuracy, but also the mode of adaptation: less fluent Korean listeners are more attentive to non-contrastive phonetic information, such as coda release, than more fluent listeners when perceiving English sounds [7], and in environments with degrees of bilingualism, highly bilingual speakers are more likely to refer to phonological representations over phonetic representations of the input language in loanword adaptation [8].

Turkish provides an excellent testing ground to understand both the role of perception in loanword adaptation and the impact of different levels of L2 proficiency on the adaptation process. As for the first aspect, word-initial consonant clusters are unattested in the native lexicon of Turkish. However, Turkish has extensively borrowed lexical items from other languages, including languages which have more complex syllable types. Turkish words such as *tren* (from French *train*) or *tramvay* (from English *tramway*) generally show at least two variants. A careful or learned pronunciation may preserve the cluster. However, in a more common colloquial pronunciation an epenthetic high vowel between the two consonants of the cluster is typically observed. It is often shorter than lexical vowels, but it may acquire the full duration of normal vowels, for example under contrastive emphasis [9] (although see [10] for an analysis in terms of vowel intrusion rather than epenthesis). In conformity with Turkish vowel harmony, the frontness/backness and roundness of this vowel appear to match that of the following vowel (however, the epenthetic vowel is always back after velar plosives [9]).

As for the impact of different levels of L2 proficiency on the adaptation process, knowledge of foreign languages is unevenly distributed across the

population, providing ample room to compare the performance of proficient vs. less skilled L2 speakers.

2. EXPERIMENTS AND RESULTS

We carried out two perceptual experiments, one ABX test and one identification test, to explore the relationship between L2 proficiency and perception of non-native clusters by Turkish speakers.

In both experiments the subjects heard pairs of pseudowords, the only difference between them being the presence or absence of a vowel between the first two consonants ($C_1V_1C_2V_2C_3V_3$ vs. $C_1C_2V_2C_3V_3$). Our goal was to investigate two related hypotheses: first, given the role played by L1 phonology on the faithful perception of clusters, and since the native Turkish lexicon does not allow initial CC clusters, Turkish speakers are expected to discriminate between #CCV and #CVC with difficulty. Second, subjects who are fluent in at least one L2 that allows #CC clusters are expected to perceive them more accurately than subjects who are monolingual or have a limited knowledge of such L2s.

2.1. Experiment 1

In this experiment we used an ABX paradigm: subjects first heard two stimuli, one with and one without a word-initial consonant cluster, and then had to decide whether the third stimulus they heard was identical to the first or to the second.

2.1.1. Participants

Thirty-eight native Turkish speakers participated in this experiment. One group (n=23, mean age 23.2 years, SD = 2.6 years) consisted of speakers proficient in at least one language allowing word-initial consonant clusters (English, French, German, and so on); they had started acquiring it between the age of 6 and 11, and have studied or been exposed to it continuously for at least the last 10 years. The second group (n=15, mean age 43.1 years, SD = 12.3 years) consisted of speakers with limited or no knowledge of any language allowing word-initial consonant clusters (the age difference reflects subject availability; subjects in the first group are mostly university students, while subjects in the second group reflect less widespread L2 acquisition among the older generations. We do not think that age is in itself a major factor in their different performance).

2.1.2. Materials

Ten triplets of the form (CiCeCV, CCeCV, CaCeCV) were created. All stimuli were pseudowords in Turkish and, except for the initial CC clusters, consisted of phonologically licit Turkish syllables. CCeCV pseudowords provided the cluster; The first vowel in CiCeCV is the most likely Turkish speakers would perceive in case of perceptual illusion (as said above, epenthetic vowels are usually high and conform to Turkish vowel harmony, thus a following [e] would imply an epenthetic front unrounded vowel); the correct identification of CaCeCV is used as a baseline.

When asked about identity, it may be possible that subjects base their decision more on acoustic similarity than on phonological representations; that is, they might judge the AB stimuli CiCeCV and CCeCV to be different not because they perceive a cluster in the latter, but because they conclude that it is a [CiCeCV] token with a very short [i], and hence not identical to the other stimulus. To reduce the role of phonetic similarity and induce subjects to rely on a more abstract representation, we used two different talkers; one of them produced the A and B stimuli, while the other produced the X stimuli.

Therefore, the stimuli were recorded twice, once by a female native Turkish speaker and once by a male native Turkish speaker. The recordings were made in a quiet room using a cardioid condenser microphone powered by a preamplifier connected to a PC laptop, where the signal was digitized at a 44.1 kHz sampling rate using the software *Audacity* [11].

Since both speakers were native Turkish speakers, they would be expected to insert an epenthetic vowel when asked to produce the CCeCV stimuli. However, both have training in phonetics and phonology, were informed about the goal of the experiments and the kind of stimuli needed, are fluent L2 speakers of languages (English and French) allowing word-initial clusters, and were explicitly instructed to produce clusters. Their production was immediately checked, and they were asked to repeat a CCVCV stimulus if the initial recording showed vowel-like characteristics between the first two consonants. This procedure was repeated until a satisfactory cluster was obtained. ABX trials were then created. Each experimental trial consisted of the stimuli A, B, and X, where A and B were from the female speaker, while X (repeating either A or B) was from the male speaker. Half of the AB pairs differed in the presence vs. absence of a cluster (CCeCV vs. CiCeCV). Each pair could appear in the two different possible orders, and either the first or the second item could

be repeated, resulting in four (2x2) trial triplets for each CCeCV / CiCeCV pair.

Additionally, ABX pairs CiCeCV / CaCeCV were created to provide a baseline. In principle, inability to distinguish CCeCV from CiCeCV could be due to some other confounding variable rather than perceptual illusion. By measuring their ability to perceptually discriminate CiCeCV from CaCeCV, which should be a very simple task for native Turkish speakers, we have a point of reference against which the CiCeCV vs. CaCeCV discrimination rate can be compared. For each CiCeCV / CaCeCV pair we created four ABX trials as above, for a total of 80 trials (10x2x2x2).

2.1.3. Procedure

After describing the task to the subjects, the stimuli were presented with a laptop and a low-noise headphone to each participant through an ABX task scripted in *jsPsych* [12]. The participants were asked to listen to the ABX trials, determine whether the third sound was a repetition of the first or the second, and click on the corresponding box (*Birinci* 'The first' or *İkinci* 'The second') on the screen. All the instructions were in Turkish. The inter-stimulus interval was 500 ms and the inter-trial interval was 1500 ms. The experiment started with a brief practice session to familiarize the subjects with the task. This session had 10 trials with a set of pseudowords similar to the experiment proper's. All trials were randomized for each participant. The experiment took about 10-15 minutes per person.

2.1.4. Results and discussions

The ratio of 'incorrect' responses as a function of L2 proficiency is shown in Figure 1. Turkish native speakers with limited or no exposure to languages allowing word-initial CC clusters tend to confuse CiCeCV with CCeCV much more often than speakers who are fluent speakers of at least one such L2. This result is consistent with the hypothesis that proficiency in an L2 phonology may influence the way the sounds and structures of a previously unheard word are categorized.

Turkish native speakers with limited or no exposure to other languages also confused [a] and [i] more often than the other group (presumably because, being on average older and less educated that the latter, they were relatively less at ease with the use of a laptop and the kind of task they were asked to perform). However, this kind of mistake is substantially less pronounced than the CiC / CC confusability, suggesting that their difficulty in correctly distinguishing clusters from CiC sequences is not merely the effect of unfamiliarity with the task.

Since the accurate perception of the CiC / CC stimuli – the dependent variable – is binary, we modelled the data with a mixed-effects generalized linear model (*glmer* function in the *lme4* R package [13]), with subjects and items as random variables. We used each speaker's number of misidentifications between [a] and [i] as a proxy of their general ability to perform the task, as opposed to the language-specific ability to correctly perceive consonantal clusters. The results are shown in Table 1. L2 fluency highly increases the likelihood of a correct CiC / CC discrimination, while apparently the general ability to perform the test has a minor negative effect on it it. The influence of both factors is statistically significant, while their interaction is not.

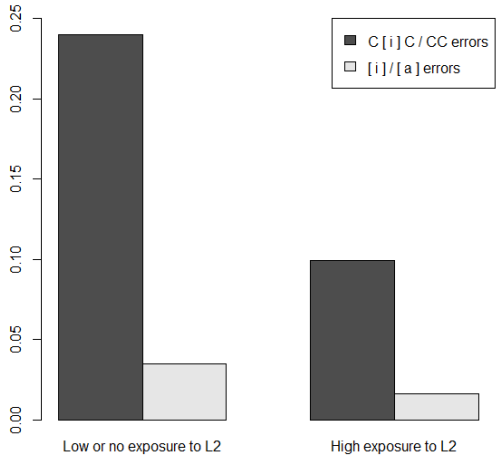


Figure 1: Proportions of misidentifications between C[i]C and CC in ABX test depending on speakers' level of exposure to L2s. Proportions of misidentifications between C[i]C and C[a]C added for comparison.

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.4219	0.1917	7.42	***
FluentL2	1.1603	0.2241	5.18	***
GeneralAbility	-0.123	0.0646	-2.01	*
FluentL2: GeneralAbility	-0.185	-1.51	-1.31	0.131

Table 1: Mixed-effects generalized linear model of the interaction between 'correct' identification of the CiC / CC and L2 fluency.

2.2. Experiment 2

In this experiment we used an identification paradigm: the subjects heard a stimulus, and then had to decide whether it contained the vowel [i] or not. Compared to the ABX test, the judgement required in this type of test entails a higher degree of metalinguistic awareness from the subject. At the same time, it avoids the problem of what participants really consider to be a repetition of the 'same' word; therefore, we decided to use both experimental paradigms.

In this case as well, we expected that subjects with higher L2 proficiency would perform better in correctly identifying the absence of an [i] in pseudowords starting with a consonantal cluster.

2.2.1. Participants

Twenty-eight native Turkish speakers participated in the experiment. Their demographics is similar to that of the first experiment. One group (n=14, mean age 22.4 years, SD = 2.1 years) consisted of proficient L2 speakers, while the other (n=14, mean age 38.1 years, SD = 7.3 years) consisted of less skilled L2 speakers or monolingual speakers.

2.2.2. Materials

In this experiment the same stimuli used in the first experiment were used (either CCeCV or CiCeCV or CaCeCV). As in the first experiment, CaCeCV stimuli were included to provide a baseline.

2.2.3. Procedure

After describing the task to the subjects, the stimuli were presented with a laptop and a low-noise headphone to each participant through an identification task scripted in *jsPsych* [12]. The participants were asked to listen to each stimulus and say whether it included an [i] clicking on the corresponding box (*Hayır* 'No' or *Evet* 'Yes') on the screen. All the instructions were in Turkish. Each pseudoword was heard four times (uttered twice by the female voice and twice by the male voice). The inter-stimulus interval was 750 ms. As with the first experiment, this experiment too started with a brief practice session to familiarize the subjects with the task. All stimuli were randomized for each participant. The experiment took about 8-10 minutes.

2.2.4. Results and discussion

Picture 2 shows that also the identification test reveals a pattern comparable to the one provided by the ABX test: Turkish speakers who have had little or no exposure to languages that admit word-initial consonant clusters perceive a vowel in such clusters significantly more often than proficient L2 speakers.

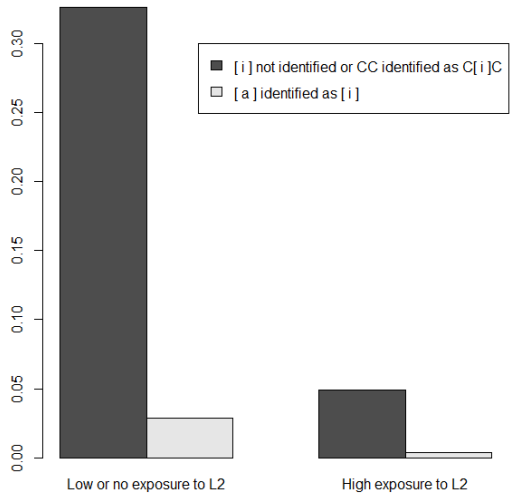


Figure 1: Proportions of misidentified [i]'s in identification test.

We modelled these data as well with a mixed-effects generalized linear model in which subjects and items are random variables and the correct/incorrect identification of [i] in a token is the dependent variable. The analysis showed a highly significant positive correlation (estimate 2.858, std. error 0.1704) between being a proficient bilingual speaker and the ability to accurately identify the presence or absence of [i] in CiCeCV / CCeVCV stimuli.

3. CONCLUSIONS

The results of both experiments indicate that proficiency in an L2 phonology may interact with and complement the role of native phonology in phonological perception, facilitating the correct perception of phonotactic configurations unattested in a speaker's L1 if these configurations are part of L2.

4. ACKNOWLEDGMENTS

I would like to thank Pavel Logačev and Elif Sarmış for their comments and suggestions on the design of the experiments, and Can Aydemir and Göksu Gamlı for their invaluable support in the collection of the data. This research was funded by Boğaziçi University's Bilimsel Araştırma Projeleri Koordinatörlüğü [Pr. No: 16021, 'Syllabic phonology'].

5. REFERENCES

[1] Kang Y. 2011. Loanword phonology. In: Oostendorp M., Ewen C, Hume E., Rice K. (eds), *The Blackwell Companion to Phonology*. Blackwell, 2258–2282.

[2] Peperkamp, S., Dupoux, E. 2003. Reinterpreting loanword adaptations: The role of perception. *Proc. 15th ICPhS* Barcelona, 367–370.

[3] Dupoux E., Parlato E., Frota S., Hirose Y., Peperkamp S. 2011. Where do illusory vowels come from? *Journal of Memory and Language* 64, 199–210.

[4] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Mehler, J. 1999. Epenthetic vowels in Japanese: a perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25: 1568–1578.

[5] Flege, J. E., Bohn, O.-S., Jang, S. 1997. Effects of experience on nonnative subjects' production and perception of English vowels. *Journal of Phonetics* 25, 169–186.

[6] Flege, J. E., MacKay, I. R. A., Meador, D. 1999. Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America* 106: 2978–2987.

[7] Kwon, H. 2017. Language experience, speech perception and loanword adaptation: variable adaptation of English word-final plosives into Korean. *Journal of Phonetics* 60, 1–19.

[8] Heffernan, K. 2007. The role of phonemic contrast in the formation of Sino-Japanese. *Journal of East Asian Linguistics* 16, 61–86.

[9] Clements, G. N., Sezer, E. 1982. Vowel and consonant disharmony in Turkish. In: van der Hulst, H., Smith, N. (eds), *The Structure of Phonological Representations, Part II*. Foris, 213–255.

[10] Bellik, J. 2018. An acoustic study of vowel intrusion in Turkish onset clusters. *Laboratory Phonology* 9(1), 16.

[11] Audacity Team. 1999–2021. Audacity® software [Computer program].

[12] de Leeuw, J. R. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12.

[13] Bates D., Mächler M., Bolker B., Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48

Language-specific stress discrimination by European Portuguese-learning infants – An ERP study

Shuang Lu[1], Cátia Severino[2], Marina Vigário[2], Sónia Frota[2]

[1]School of Foreign Languages, Renmin University of China, Beijing, China
[2]Center of Linguistics, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal
shuanglu2013@outlook.com, catiaseverino@campus.ul.pt, mvigario@edu.ulisboa.pt, sfrota@edu.ulisboa.pt

ABSTRACT

European Portuguese (EP) is a language with variable stress, a mixed prosodic profile and conflicting frequency distributions of trochaic and iambic stress patterns. Besides, duration and vowel quality, instead of pitch, have been claimed as the primary cues for stress perception in EP. Previous ERP and behavioral studies have revealed diverging results regarding EP adult speakers' stress discrimination in the absence of vowel quality cues: they were able to perceive the stress contrasts at the pre-attentive stage, whereas exhibited a stress "deafness" effect similar to that found in speakers of languages with fixed stress at the attentive stage. Nonetheless, both measures on EP adults have demonstrated a processing advantage for the iambic stress pattern. Using a passive oddball paradigm, the present study showed that EP-learning infants at 5-7 months exhibit a mismatch response for iambic stress only, indicating a processing advantage for the iambic stress pattern.

Keywords: infant stress perception, ERP, mismatch response, iambic advantage, European Portuguese.

1. INTRODUCTION

Development of stress perception plays an important role in infants' language acquisition. It may facilitate infants' abilities to segment and categorize words [1], [2], [3], [4], [5], or even be an early marker of later language development [6], [7]. Infants' early ability in stress discrimination has been reported for Italian newborns [8], English infants at 2 months [9], German infants at 4-6 months [10], [11], [12], Spanish infants at 6 months [13], and French infants at 9-10 months [14], [15]. Importantly, this ability has been shown to develop in language-specific ways [16], [17].

Infants' ability in stress discrimination may be manifested by preference for one of the stress patterns [17]. For example, German infants have been shown to develop a processing advantage for trochaic stress at 4-5 months [12]. English infants have been found to develop a trochaic preference between 6-9 months [18]. Hebrew infants at 9 months revealed a preference for iambic stress over trochaic stress [19]. However, native Spanish and Catalan infants did not show a preference for either stress pattern at 6 and 9 months [20]. These language-specific preference patterns have been explained by a rhythmic account [11], [21], which proposed that infants of stress-timed languages tend to develop a trochaic preference, while learners of syllable-timed languages are expected to develop no preference [17]. Another explanation that could account for the language-specific stress preference pattern is the frequency of stress patterns. Infants are expected to develop a preference for the dominant stress pattern of the language. Lastly, infants may be more sensitive to certain cues (i.e., pitch, intensity, or duration) that signal stress in their native language [19].

EP has a mixed prosodic profile that includes both stress-timed and syllable-timed rhythm [22]. Besides, it is difficult to tell which is the dominant stress pattern based on the frequency distribution of stress patterns in EP [17]. Finally, duration and vowel quality, instead of pitch, have been claimed as the primary cues for EP stress perception, which makes it different from English [23], Spanish, Catalan [20] or Hebrew [19]. Previous behavioral and ERP studies have revealed diverging results regarding EP adult speakers' stress discrimination in the absence of vowel quality cues: they could perceive the stress contrasts at the pre-attentive stage, whereas exhibited a stress "deafness" effect at the attentive stage [24], [25]. Nonetheless, both measures on EP adults have shown a processing advantage for iambic stress. A recent eye-tracking study found that, in the absence of vowel quality cues, EP-learning infants at 5-6 months look longer at the iambic pattern [17]. To date, no electrophysiological study has investigated the developing stress perception abilities in EP-learning infants. Using a passive oddball paradigm, the present study recorded ERPs from 5-7 month-old infants to examine: 1) whether EP-learning infants would show stress discrimination at 5-7 months of age; and 2) whether they would reveal a language-specific stress preference.

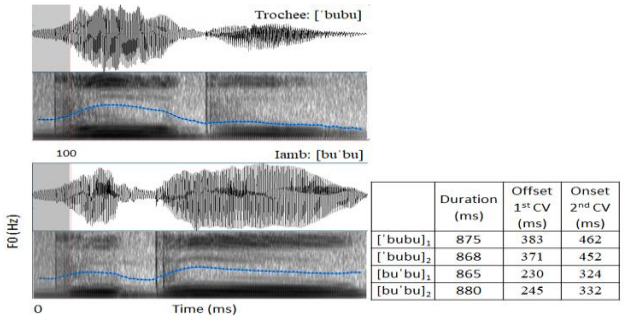
2. METHODS

2.1. Participants

Twenty-three infants were recruited from the wider Lisbon area (12 females). All infants were raised in monolingual EP homes, and their age range was between 5 months 7 days to 7 months 10 days (M = 6 months 11 days, SD = 14 days). All participants were typically developing infants according to an EP adapted version of the CSBS-DP Checklist [26]. Fifteen additional infants were tested but excluded from data analysis due to fussiness and technical problems. All caregivers completed an informed consent prior to data collection.

2.2. Stimuli

The stimuli in the present study were the same as in a previous adult study [24]. A female native speaker of EP naturally produced the disyllable [bubu] with either a trochaic or an iambic stress pattern. Each stress pattern was produced twice, resulting in four tokens in total. The productions were recorded at a sampling rate of 22050Hz. The mean duration is 872 milliseconds for the trochaic tokens and 873 milliseconds for the iambic tokens. The timings of the offset of the first CV and the onset of the second CV for each token are presented in Fig. 1. Following [12], we replaced the first 100 milliseconds of [ˈbubu]$_1$,[ˈbubu]$_2$, and [buˈbu]$_2$ by the first 100 milliseconds of [buˈbu]$_1$ to control the onset acoustic differences. After the manipulation, physical differences between tokens started at 100 milliseconds, and no pitch discontinuity was observed in any of the tokens. All stimuli were nonsense words in EP, and were judged as perceptually natural by three native EP speakers.



	Duration (ms)	Offset 1st CV (ms)	Onset 2nd CV (ms)
[ˈbubu]$_1$	875	383	462
[ˈbubu]$_2$	868	371	452
[buˈbu]$_1$	865	230	324
[buˈbu]$_2$	880	245	332

Figure 1: Spectrograms of the trochaic and iambic stress patterns, with the duration and timings of the offset of the first CV and the onset of the second CV for each token.

2.3. Procedure

Two blocks were created in a passive oddball paradigm: (1) trochaic block – the frequently occurring iambic tokens were occasionally replaced by the deviant trochaic tokens; (2) iambic block – the iambic tokens served as deviants, while the trochaic tokens functioned as standards. Within each block, each standard token was presented 250 times and each deviant token occurred 50 times, resulting in 600 trials in total (250 x 2 tokens + 50 x 2 tokens). The stimuli were delivered in a pseudo-random order, with at least two standards preceding each deviant. A hundred clean standards (50 x 2 tokens) that were not immediately preceded or followed by a deviant were selected from each block to compare with the same stress pattern presented as deviants in the other block. In order to prevent participants' automatic anticipation of stimulus onset, we randomly varied the offset-to-onset inter-stimulus interval between 800, 825, and 850 milliseconds. Each block was split equally into two sub-blocks with each one lasting for about 8 minutes. The order of the four sub-blocks was counterbalanced across participants. E-Prime 2.0 software [27] was used for stimulus presentation.

The experiment was conducted in a sound attenuated booth. Infants sat on their parents' laps while the stimuli were presented via loudspeaker at a constant and comfortable hearing level. During the EEG recordings, infants were entertained by watching a silent cartoon.

2.4. EEG recording and averaging

Continuous EEG was recorded from 32 Ag/AgCI electrodes according to the International 10-20 electrode placement standard. The electrodes were mounted on an elastic cap (Quik-cap, Compumedics, NeuroScan, Victoria, Australia). The vertical eye movements were recorded from electrodes placed above and below the left eye. The EEG was referenced online to the left mastoid and was amplified using the SynAmps RT 128-channel Amplifier (Compumedics NeuroScan, Victoria, Australia), with a sampling rate of 1000Hz. Impedances were kept below 10 kΩ.

The EEG data were processed using MATLAB toolboxes (MathWorks): EEGLAB and ERPLAB [28]. The EEG signals were re-referenced to average reference and were band-pass filtered from 1 to 30 Hz. Eye artifact was removed through independent component analysis (ICA, EEGLAB). The raw EEG data were segmented into epochs of 1000 milliseconds, with a 200 ms pre-stimulus baseline and 800 ms after the onset of the stimulus. Trials exceeding ±150μV in any channel on the entire epoch were rejected. On average, 58 trials for each stimulus type were included in data analysis. Finally, the ERPs were averaged for each stimulus type,

electrode and participant. The grand-averaged difference waves were generated for the trochaic and iambic stress patterns respectively by subtracting the average responses to the clean standards from the average responses to the corresponding deviants.

2.5. Data analysis

Based on visual inspection, mean amplitudes within five consecutive time windows of 100 milliseconds were computed from 200 to 700 milliseconds after stimulus onset. The mean amplitudes were analysed in four regions: left-frontal (LF) contained the electrodes F7, F3, FT7, and FC3; right-frontal (RF) included the electrodes F4, F8, FC4, and FT8; left posterior (LP) comprised the electrodes TP7, CP3, P7, and P3; right-posterior (RP) consisted of the electrodes CP4, TP8, P4, and P8.

The mean amplitudes for each stress pattern and time window were submitted to 2 x 2 x 2 repeated measures ANOVAs with Discrimination (deviant vs. standard), Hemisphere (left vs. right), and Anteriority (anterior vs. posterior) as within-subject factors. The Greenhouse-Geisser correction was applied to all the p-values and the F-values, and the Bonferroni correction was applied for multiple comparisons.

3. RESULTS

Grand averages at electrodes Fz and Pz are displayed in Fig. 2a for trochaic stress and in Fig. 2b for iambic stress. For trochaic stress, a mismatch negativity (MMN) was observed between 500 to 700 milliseconds. However, for iambic stress, a MMR-like component was found between 500 to 700 milliseconds, with a prominent frontal distribution.

3.1. Trochaic stress

Table 1 summarizes the main effects and interactions in the five time windows of 100 milliseconds for trochaic and iambic stress. The main effect of Discrimination was not significant in any of the time window (200-300ms: [F (1, 22) = .74, p = .40]; 300-400ms: [F (1, 22) = .44, p = .51]; 400-500ms: [F (1, 22) = .64, p = .43]; 500-600ms: [F (1, 22) = .38, p = .55]; 600-700ms: [F (1, 22) = .51, p = .48]). In the time windows of 200-300ms [F (1, 22) = 10.95, p = .003, η^2 = .33], 300-400ms [F (1, 22) = 24.4, p < .001, η^2 = .53], and 600-700ms [F (1, 22) = 28.9, p < .001, η^2 = .57], the main effect of Anteriority was significant, with the mean amplitudes being positive in the frontal area, while negative in the posterior area. In the time window of 400-500ms, effects of Hemisphere [F (1, 22) = 10.89, p = .003, η^2 = .33] and Anteriority [F (1, 22) = 22.2, p < .001, η^2 = .50] were found. The mean amplitude was more negative in the left hemisphere than in the right hemisphere. For the time window of 500-600ms, the main effects of Hemisphere [F (1, 22) = 7.33, p = .013, η^2 = .25], Anteriority [F (1, 22) = 20.78, p < .001, η^2 = .49], and the interaction of Hemisphere x Anteriority [F (1, 22) = 6.04, p = .022, η^2 = .22] reached significance. No other significant main effect or interaction was found. In sum, even though visual inspection seemed to suggest that the trochaic deviant stimulus elicited a MMN-like response in the 500-600ms time window, statistical analyses did not reveal any significant discrimination effect.

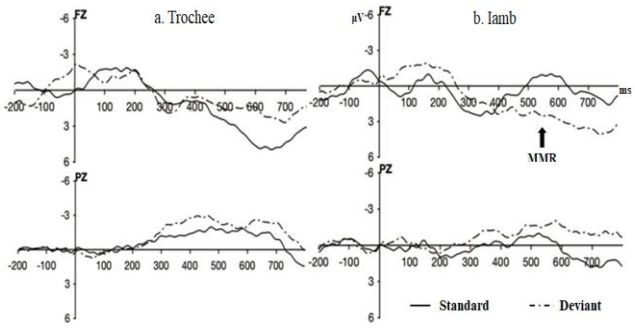


Figure 2: Grand averages for a) the trochaic and b) the iambic stress at electrodes Fz and Pz.

a. Trochee	200-300	300-400	400-500	500-600	600-700
Ant.	**	***	***	***	***
Hem.			**	*	
Ant. x Hem.				*	
a. Iamb					
Ant.	*	***	**		*
Dis. x Ant.				**	*

Table 1: Main effects and interactions in the five time windows for a) trochaic stress and b) iambic stress. *** $p \leq .001$, ** $p \leq .01$, * $p < .05$.

3.2. Iambic stress

In the time window of 200-300ms, the main effect of Discrimination almost reached significance [F (1, 22) = 4.21, p = .052, η^2 = .16], with the iambic deviant stimulus being more negative than the iambic standard stimulus. In addition, the main effect of Anteriority was significant [F (1, 22) = 4.50, p = .045, η^2 = .17]. The main effect of Anteriority was also significant in the time windows of 300-400ms [F (1, 22) = 18.13, p < .001, η^2 = .45], 400-500ms [F (1, 22) = 8.84, p = .007, η^2 = .29], and 600-700ms [F (1, 22) = 6.61, p = .017, η^2 = .23]. There was a significant interaction of Discrimination x

Anteriority in both the time windows of 500-600ms [F (1, 22) = 10.95, p = .003, η^2 = .33] and 600-700ms [F (1, 22) = 7.01, p = .015, η^2 = .24]. Post hoc analyses only yielded a significant discrimination effect in the posterior region in the 500-600ms time window [t (22) == 2.84, p = .01], but not in the 600-700ms time window [t (22) == 2.04, p = .054]. Besides, neither the 500-600ms [t (22) == -1.62, p = .12] or the 600-700ms [t (22) == -1.70, p = .10] time window revealed a significant discrimination effect in the frontal area.

Further 2 x 3 repeated measures ANOVAs were performed on the midline electrodes (Fz, Cz, and Pz), with Discrimination (deviant vs. standard) and Anteriority (anterior, central, and posterior) as within-subject factors. The results showed that in both the time windows of 500-600ms [F (2, 44) = 4.21, p = .025, η^2 = .16] and 600-700ms [F (2, 44) = 7.44, p = .004, η^2 = .25] the interaction of Discrimination x Anteriority was significant. Post hoc analyses revealed that there was a significant discrimination effect at Fz in both time windows (500-600ms: [t (22) = -2.42, p = .024]; 600-700ms [t (22) = -2.17, p = .041].

Taken together, a discrimination related positive response was only elicited by iambic stress in 5-to-7 month-old EP-learning infants.

3.3. Difference waves

Fig. 3 shows the Grand-average difference wave (deviant minus standard) for trochaic and iambic stress at electrode Fz. Five 2 x 2 x 2 repeated measures ANOVAs with Stress (Trochee vs. Iamb), Hemisphere (left vs. right), and Anteriority (anterior vs. posterior) as within-subject factors were performed on the difference waves, in order to directly compare the differences between trochaic and iambic stress in the five time windows. The results showed a significant interaction of Stress x Anteriority in the 500-600ms [F (1, 22) = 6.71, p = .017, η^2 = .23] and 600-700ms [F (1, 22) = 9.83, p = .005, η^2 = .31] time windows. However, post hoc analyses revealed no significant difference between the trochaic and iambic stress in either the time windows.

Further 2 x 3 repeated measures ANOVAs were performed on the midline electrodes (Fz, Cz, and Pz), for the difference waves, with Stress (trochee vs. iamb) and Anteriority (anterior, central, and posterior) as within-subject factors. A significant interaction of Stress x Anteriority was only found in the 600-700ms [F (2, 44) = 7.46, p = .003, η^2 = .25] time window. The trochaic difference wave and iambic difference wave revealed diverging response at Fz.

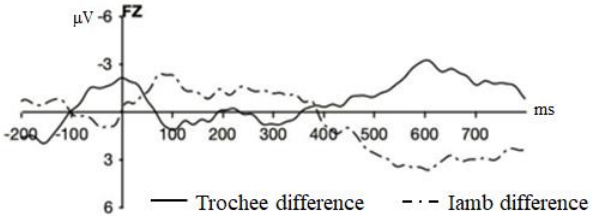


Figure 3: Grand-average difference wave for the trochaic and iambic stress at electrode Fz.

4. DISCUSSION

In the present study, we used the electrophysiological method to investigate stress discrimination in 5-to-7 month-old EP-learning infants. No significant discrimination effect was found for the trochaic stress pattern. However, the iambic stress pattern yielded a discrimination response (MMR) with a positive polarity at 500-600ms after stimulus onset. In addition, a significant discrimination effect was elicited in the posterior region at the 500-600ms time window, with the deviant iambic stimulus being more negative than the standard iambic stimulus. Previous infant studies have also reported this MMR component [10], [12], [29]. Some studies proposed that MMR reflects a genuine discrimination response due to the immaturity of the infant brain [10]. Others claimed that the MMR at fronto-central sites may be considered as a P3a, which reflects automatic novelty detection [30]. No definite explanation has been provided to account for this difference in polarity of the mismatch response in infants. In the present study, the MMR, as well as the negative response in the posterior region, seemed to suggest a processing advantage for the iambic stress in infants. This result is consistent with both the ERP study on EP adult speakers and the eye-tracking study on EP-learning infants, which also revealed a processing advantage for the iambic stress [17], [24]. Thus, infants seem to develop their stress perception ability through an asymmetrical perception mechanism triggered by iambic stress. Future studies need to address when the pre-attentive discrimination of stress contrasts found in adult speakers develops, beyond the iambic preference.

5. ACKNOWLEDGEMENTS

This research was supported by the Portuguese Foundation for Science and Technology (Grants PTDC/LLTLIN/29338/2017, UIDB/00214/2020).

6. REFERENCES

[1] Jusczyk, P. W., Houston, D. M., and Newsome, M. 1999. The beginnings of word segmentation in English-learning infants. *Cogn. Psychol.* 39, 159–207.

[2] Christophe, A., Guasti, M. T., Nespor, M., and Van Ooyen, B. 2003. Prosodic structure and syntactic acquisition: the case of the head-complement parameter. *Dev. Sci.* 6, 211–220.

[3] Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., and Alcantara, C. 2006. Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *J. Mem. Lang.* 54, 283–299.

[4] Frota, S., Butler, J., Correia, S., Severino, C., and Vigário, M. 2012. "Pitch first, stress next? Prosodic effects on word learning in a intonation language" in *Proceedings of the 36th Annual Boston University Conference on Language Development*, Boston, MA, 190–201.

[5] Polka, L., and Sundara, M. 2012. Word segmentation in monolingual infants acquiring Canadian English and Canadian French: native language, cross-dialect, and cross-language comparisons. *Infancy* 17, 198–232.

[6] Weber, C., Hahne, A., Friedrich, M., and Friederici, A. D. 2005. Reduced stress pattern discrimination in 5-month-olds as a marker of risk for later language impairment: neurophysiologial evidence. *Cogn. Brain Res.* 25, 180–187.

[7] Friedrich, M., Herold, B., and Friederici, A. D. 2009. ERP correlates of processing native and non-native language word stress in infants with different language outcomes. *Cortex* 45, 662–676.

[8] Sansavini, A., Bertoncini, J., and Giovanelli, G. 1997. Newborns discriminate the rhythm of multisyllabic stressed words. *Dev. Psychol.* 33, 3–11.

[9] Jusczyk, P. W., and Thompson, E. 1978. Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Percept. Psychophys.* 23, 105–109.

[10] Friederici, A. D., Friedrich, M., and Christophe, A. 2007. Brain responses in 4-month-old infants are already language specific. *Curr. Biol.* 17, 1208–1211.

[11] Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., and Nazzi, T. 2009. Language specific prosodic preferences during the first year of life: evidence from German and French infants. *Infant Behav. Dev.* 32, 262–274.

[12] Weber, C., Hahne, A., Friedrich, M., and Friederici, A. D. 2004. Discrimination of word stress in early infant perception: electrophysiological evidence. *Cogn. Brain Res.* 18, 149–161.

[13] Skoruppa, K., Pons, F., Bosch, L., Christophe, A., Cabrol, D., and Peperkamp, S. 2013. The development of word stress processing in French and Spanish infants. *Lang. Learn. Dev.* 9, 88–104.

[14] Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., et al. 2009. Language-specific stress perception by 9-month-old French and Spanish infants. *Dev. Sci.* 12, 914–919.

[15] Bijeljac-Babic, R., Serres, J., Höhle, B., and Nazzi, T. 2012. Effect of bilingualism on lexical stress pattern discrimination in French-learning infants. *PLoS One* 7:e30843.

[16] Bhatara, A., Boll-Avetisyan, N., Höhle, B., and Nazzi, T. 2018. Early sensitivity and acquisition of prosodic patterns at the lexical level. In: Prieto, P., Esteves-Gibert, N. (eds), *The development of prosody in first language acquisition 23*. Amsterdam: John Benjamins, 37–57.

[17] Frota S, Butler J, Uysal E, Severino C and Vigário M 2020. European Portuguese-Learning Infants Look Longer at Iambic Stress: New Data on Language Specificity in Early Stress Perception. *Front. Psychol. 11:1890.*

[18] Jusczyk, P. W., Cutler, A., and Redanz, N. J. 1993. Infants' preference for the predominant stress patterns of English worlds. *Child Dev.* 64, 675–687.

[19] Segal, O., and Kishon-Rabin, L. 2012. Evidence for language-specific influence on the preference of stress patterns in infants learning an Iambic language (Hebrew). *J. Speech Lang. Hear. Res.* 55, 1329–1341.

[20] Pons, F., and Bosch, L. 2007. "The perception of lexical stress patterns by Spanish and Catalan infants" in *Segmental and prosodic issues in Romance phonology*. eds. P. Prieto, J. Mascar and M. J. Sol (Amsterdam: John Benjamins), 199–218.

[21] Butler, J., and Frota, S. 2018. Emerging word segmentation abilities in European Portuguese-learning infants: new evidence for the rhythmic unit and the edge factor. *J. Child Lang.* 45, 1294–1308.

[22] Frota, S., and Vigário, M. 2001. On the correlates of rhythmic distinctions: the European/Brazilian Portuguese case. *Probus* 13, 247–273.

[23] Chrabaszcz, A., Winn, M., Lin, C. Y., and Idsardi, W. J. 2014. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *J. Speech Lang. Hear. Res.* 57, 1468–1479.

[24] Lu, S., Vigário, M., Correia, S., Jerónimo, R., and Frota, S. 2018. Revisiting stress "Deafness" in European Portuguese – a behavioral and ERP study. *Front. Psychol.* 9:2486.

[25] Correia, S., Butler, J., Vigário, M., and Frota, S. 2015. A stress "deafness" effect in European Portuguese. *Lang. Speech* 58, 48–67.

[26] Wetherby, A., and Prizant, B. 2002. *Communication and symbolic behavior scales – diagnostic profile, first normed edition*. Baltimore: Brookes Publishing.

[27] Schneider, W., Eschman, A., and Zuccolotto, A. 2012. E-Prime User's Guide. Pittsburgh, PA: Psychology Software Tools, Inc.

[28] Lopez-Calderon, J., Luck, S. J. 2010. *ERPLAB Toolbox*. http://erpinfo.org/erplab

[29] Dehaene-Lambertz, G. 2000. Gerebral specialization for speech and non-speech stimuli in infants, *J. Cogn. Neurosci.* 12, 449-460.

[30] Cote, A. K. 2002. Probing awareness during sleep with the auditory oddball paradigm, *Int. J. Psychophysiol*. 46, 227-241.

SEGMENTAL ACCURACY AND L2 INTELLIGIBILITY: NATIVE AND NONNATIVE LISTENERS' PERSPECTIVES

XXX

XXX
XXX

ABSTRACT

This study investigated the contribution of listeners' L1 and stimulus properties to the perception of L2 speech. Thirty-nine listeners from five L1 backgrounds (native English, Korean, Mandarin, Indonesian, Turkish) completed a minimal-pairs forced choice task recorded by twenty Korean EFL talkers; the percentage of accurate word identification (i.e., intelligibility) was measured across listeners. Stimulus properties were measured by error coding focusing on ten target phonemes. Moderate to strong intergroup correlations suggested similarities in listeners' perceptions regardless of L1 backgrounds. No evidence for interlanguage speech intelligibility benefit (ISIB) was found for the nonnative listeners; in fact, native listeners always outperformed the nonnative listeners who shared (Korean) and did not share L1 with the talkers. Consonant errors (but not vowel errors) were linked to lower intelligibility. The effect of stimulus properties was further discussed in relation to the role of L1 Korean-specific interlanguage features in listeners' perception of L2 speech.

Keywords: intelligibility, pronunciation, segmental errors, listener factors, nonnative

1. INTRODUCTION

The importance of intelligibility over nativeness has become central to L2 pronunciation with the recognition that a speaker can be highly understandable despite a strong foreign accent [1]. To Munro and Derwing [2], intelligibility indicates listeners' actual understanding, whereas comprehensibility refers to listeners' perceived ease of understanding. Intelligibility has most commonly been measured using listeners' orthographic transcription, yet this study adopted a forced choice task, where listeners identify sounds to directly measure intelligibility in a way that precludes listeners from using information outside the speech signal, such as world and lexical knowledge [3].

An active line of intelligibility research has revealed that variability in intelligibility is linked to speakers' production accuracy in segmentals (e.g., [4]) and prosodic features such as stress, speech rate, and rhythm (e.g., [5]). In addition to the speaker-related factors or *stimulus properties*, *listener-related factors* also play an important role in intelligibility, such as listeners' L1 [6]; expectations [7]; and familiarity with the speakers' accents [8]. Previous intelligibility research has generally focused on the perception of native listeners. Given the increasing L2 use among nonnative speakers globally, more research needs to consider nonnative listeners and different L1 backgrounds.

Some studies examined nonnative listeners' perceptions in terms of an interlanguage speech intelligibility benefit (ISIB), which refers to the benefit nonnative listeners receive when comprehending L2 speech better than native listeners. Several studies support this (e.g., [6, 9]). An ISIB was found for Mandarin listeners who were more accurate than native listeners at discriminating between a voiced or voiceless consonant (e.g., cub/cup) produced by Mandarin talkers on a forced-choice task [6]. Other studies found no support for an ISIB (e.g., [10, 11]). In [10], no evidence was found for Polish listeners listening to Polish-accented English speech on a transcription task. The Spanish listeners showing overall poorer performance indicated an interlanguage speech intelligibility *detriment* for nonnative listeners who differ in the L1 background from the nonnative talkers. Interestingly, no relationship was found between segmental or word stress errors and intelligibility, although segmental errors did affect the different perceptual measures studied.

Any investigation of listeners' understanding of L2 speech should consider both speaker-related (stimulus properties) and listener-related factors [12], but few studies have systematically examined stimulus properties in terms of segmental errors in relation to how they affect the intelligibility of L2 speech perceived by both native and nonnative listeners. Therefore, the current study aimed to address the following questions:

1. To what extent do native and nonnative English listeners share a response in their perception of L2 intelligibility?
2. What is the effect of the listener's L1 on the intelligibility of L2 speech?

3. What is the relationship between segmental errors and Korean learners' English intelligibility as perceived by native and nonnative English listeners?
   a. Which segmental sounds are most difficult for Korean learners to produce intelligibly as perceived by native and nonnative English listeners?

2. METHODOLOGY

2.1. Speech materials

The speech samples were the same as those used by [13] with native listeners. The speech samples were recorded by twenty native Korean speakers of L2 English in Seoul, South Korea; all speakers were in their final year at a high school at the time of the study (16 females, 4 males; Age = 15-17). Each speaker recorded their reading aloud ten target English words from a list of ten minimal pairs items (presented in Table 1). These minimal pairs were carefully chosen referring to a chapter on problematic phonemes specifically targeting Korean learners of English [14]. Recordings were embedded into each item of a survey on an online experiment platform *Gorilla* (https://gorilla.sc); items were randomized using the Randomise Trial function in Gorilla.

No.	Target		Foil	
1	bit	/ɪ/	beat	/i/
2	fan	/f/	pan	/p/
3	lagging	/g/	lacking	/k/
4	rhyme	/r/	lime	/l/
5	ripping	/p/	ribbing	/b/
6	sand	/æ/	send	/ɛ/
7	sheet	/ʃ/	seat	/s/
8	song	/ɔ/	sung	/ʌ/
9	taste	/ey/	test	/ɛ/
10	vote	/v/	boat	/b/

Table 1: Minimal-pairs word list.

2.2. Listeners

Thirty-one nonnative English-speaking listeners were recruited at a large university in the Midwestern US from various academic disciplines. The participants were assigned to four listening groups based on their L1 background: Mandarin Chinese (n = 9), Indonesian (n = 7), Turkish (n = 9), and Korean (n = 6). All listener groups completed a language background questionnaire upon finishing the task. It was noted that both the Korean and the native English listener groups reported very high familiarity with Korean-accented English (self-rated; mean rating 4.67 and 4 out of 5, respectively).

2.3. Data collection procedure

Listeners completed a 200-item (ten recordings each from twenty Korean talkers) minimal-pairs forced-choice task administered using Gorilla. Listeners listened to each speech sample recorded by the speakers on each minimal-pairs forced-choice trial. The corresponding target-foil pair was then simultaneously displayed on the computer monitor, and listeners chose which word they heard. Listeners were allowed as long as they needed to respond. They pressed the play button when they were ready to hear the next speech sample. Listeners were given two short breaks throughout the task, which was completed in 25 minutes or less. Responses were scored automatically.

2.4. Data analysis

2.4.1. Listener task

The overall intelligibility scores were calculated by the mean percentage of words identified correctly by the listeners (henceforth, percentage correct) for each of the twenty Korean EFL speakers. For each of the ten target words, the mean percentage of items correctly identified was also computed across all twenty speakers (out of 620 items possible for each target word). The ten target words were then divided into those containing consonant targets (n = 6) and vowel targets (n = 4), and speakers' mean intelligibility scores for each category were also calculated. This led to a total of four types of intelligibility scores that were considered during analyses: overall intelligibility (i.e., mean percentage correct for each speaker), item intelligibility (i.e., mean percentage correct for each of the ten target items), consonant and vowel intelligibility (i.e., mean percentage correct for items containing consonant/ vowel targets).

2.4.2. Error coding

Each target phoneme in each sample (n = 200) was coded as either "on target" or "not on target", the latter being perceptually salient as indicating phoneme substitutions (e.g., /p/ for /f/ in 'faint') from General American (GA) English pronunciation. Note that phonetic errors (e.g., differences in aspiration, and vowel quality variations) that did not deviate substantially from GA English were disregarded and coded as "on target".

3. RESULTS

The results of the Cronbach alpha analyses confirmed the acceptable inter-rater reliability for the thirty-one

nonnative listeners' scores (α = 0.765). All statistical analyses were carried out using software R and r effect sizes and eta-squared effect sizes were interpreted following [15] and [16], respectively.

3.1. Intergroup correlations

The intelligibility scores were averaged across all listeners separately for each of the nonnative listener groups, the distribution of which is presented in Figure 1. Table 2 summarizes the intergroup correlations on overall scores expressed as Pearson r values for all possible pairings of the five listener groups. The correlations were moderate to strong, ranging from .44 to .89.

Regarding consonant intelligibility, all the correlations between the groups were moderate to strong (r = .58 – .85). Vowel intelligibility revealed moderate to strong correlations for all (r = .40 – .70) but the Korean listener group, who showed nonsignificant and very weak correlations. In general, there were moderate to strong intergroup correlations on intelligibility, yet groups' correlations with the Korean listener group were somewhat weaker (and greatly weaker when it comes to vowel intelligibility) than with other groups.

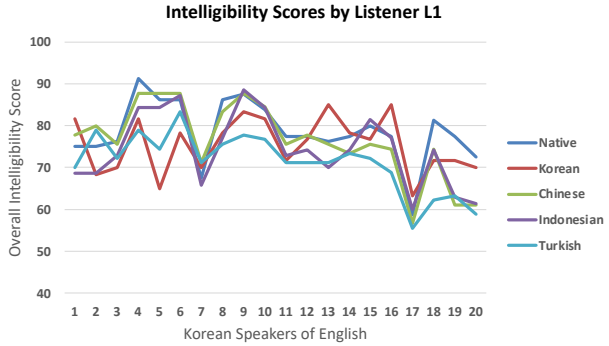


Figure 1: Mean intelligibility scores for the twenty Korean speakers of English by listener L1.

	EN	KR	CN	IN	TK
EN	1.00	–	–	–	–
KR	0.49 (.028)	1.00	–	–	–
CN	0.83 (<.001)	0.46 (.043)	1.00	–	–
IN	0.86 (<.001)	0.49 (.029)	0.88 (<.001)	1.00	–
TK	0.70 (.001)	0.44 (.050)	0.89 (<.001)	0.76 (.001)	1.00

Table 2: Intergroup correlations on the intelligibility scores.

3.2. Effects of listeners' L1 on intelligibility

To see the effect of listeners' L1 on intelligibility (see Figure 2), a one-way repeated measure ANOVA was performed using ez Package in R with listener L1 as a within-group factor (5 levels) on Korean speakers' intelligibility scores, and subsequently on consonant and vowel intelligibility. The results from the ANOVA analyses showed that there is extremely strong evidence of a difference between L1 listener groups in their intelligibility scores (F (3, 48) = 7.482, p < .001, η^2_G = .09) (Greenhouse-Geiser corrected). Post hoc (Bonferroni adjusted) analyses revealed that the Turkish listener group performed significantly lower than the native English and Chinese listener groups: native English (M_{diff}= 7.23, 95% CI [3.23, 11.22]), Chinese (M_{diff}= 5.11, 95% CI [2.13, 8.09]). The Indonesian listener group also performed significantly poorly on the task than the native listener group (M_{diff} = 4.06, 95% CI [0.91, 7.22]) (p < .001).

The effect of listener L1 was also statistical for consonant intelligibility (F (4, 76) = 5.40, p < .001, η^2_G = .048) and vowel intelligibility (F (3, 49) = 4.47, p =.003, η^2_G =.11). Regarding consonant intelligibility, post hoc analyses found that the native listener group performed significantly better than all the other groups: Chinese (M_{diff} = 4.7, 95% CI [0.3, 9.1]), Indonesian (M_{diff} = 5.2, 95% CI [1.0, 9.4]), Korean (M_{diff} = 5.6, CI [1.2, 10.0]), and Turkish (M_{diff} = 6.6, CI [2.2, 10.9]) (p <.05). No significant differences were found for any other L1 pairs (p-value approximates 1). Concerning vowel intelligibility, the Turkish listener group performed significantly poorer than the Chinese (M_{diff} = 10.1, p <.001, 95% CI [3.8, 16.4]) and the native English groups (M_{diff} = 8.6, p = .017, 95% CI [1.1, 16.0]).

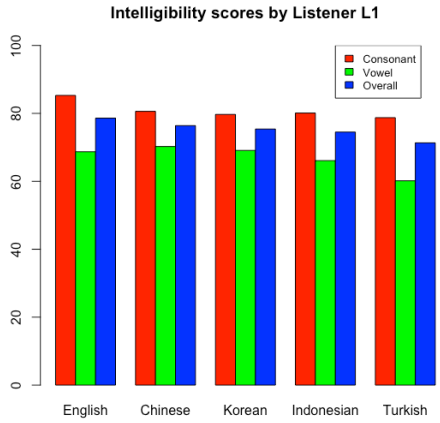


Figure 2: Intelligibility scores by five listener groups.

3.3. Segmental errors and intelligibility

When the overall intelligibility scores were compared with the number of target segment errors made by each speaker (M = 2.6, SD = 1.27, Range = 0-4; out of a maximum of 10 errors), moderate to strong negative associations were obtained for the native English (r = -.75), Chinese (r = -.56), and Indonesian (r = -.68) groups (p < .001). Slightly weaker but nonsignificant associations were found for the Korean (r = -.32, p = .164) and the Turkish group (r = -.28, p =.221). Consonant errors revealed moderate to strong negative correlations with overall intelligibility for all the listener groups ranging from -.79 to -.56 (p <.0025), whereas none of the correlations regarding vowel errors were significant (p > .05).

Table 3 shows item responses for native versus nonnative listeners combined on the listener task, along with the percentage error (identified in error coding) for different targets. Overall, native listeners and nonnative listeners showed similar responses to the items on the listener task. For both listeners, the three most intelligible target items were *sheet*, *lagging*, and *taste*, in which target segments /ʃ/, /g/, and /ey/ showed 0% error. The item *ripping* (/p/) also showed high intelligibility with 0% error. The three least intelligible items were *bit*, *sand,* and *vote*, where target segments /ɪ/, /æ/, and /v/ were of rather higher percentage error: 70%, 60%, and 25%, respectively. This is closely followed by the item *rhyme* (target consonant /r/ contrasting with /l/), which also showed lower item intelligibility with a 35% error rate.

	Listener-based intelligibility (%)		Production errors (%)
Items	Native	Nonnative	
b**i**t	36.88	53.23	70
fan	87.5	76.77	0
la**gg**ing	98.13	93.06	0
rhyme	68.13	70.48	35
ri**pp**ing	94.38	90.16	0
s**a**nd	50.63	54.84	60
sheet	98.75	90.16	0
s**o**ng	89.38	80.16	70
t**a**ste	97.5	93.55	0
v**o**te	64.38	48.39	25

Table 3: Item intelligibility by native vs. nonnative listeners; production errors on the right.

4. DISCUSSION

The questions posed at the beginning of this paper are now answered.

Listeners, regardless of their native status and L1, share a response to L2 speech in terms of intelligibility. On the other hand, the mean intelligibility scores for individual Korean EFL speakers varied greatly. It appears that the perception of L2 speech is more likely affected by the properties of the speech itself than the linguistic background of the listeners. Listener group comparisons showed no ISIB. In fact, native English listeners overall outperformed all the other listener groups. Rather, the overall poorer performance of the Turkish and the Indonesian listeners suggests an interlanguage speech intelligibility *detriment* for nonnative listeners who differ in the L1 background from the nonnative speakers.

Speakers with more segmental errors overall were found to be perceived as less intelligible by listeners; yet speakers' consonant accuracy, but not vowel accuracy, was negacorrelated with the speakers' overall intelligibility. Concerning the item intelligibility for individual target sounds, the finding of the present study showed that nonnative listeners found *sheet* (90.16%), *lagging* (93.06%), and *taste* (93.55%) the most intelligible, whereas *vote* (48.39%), *bit* (53.23%), and *sand* (54.84%) the least intelligible; this agrees with the native listener data [13]. Language-specific factors of the Korean language, such as having no distinction between lax and tense vowels (as in the case of bit vs. beat, and send vs. sand), and no equivalent sounds of labiodentals in the L1 (e.g., vote) could account for Korean speakers' overall difficulty with the three items.

5. CONCLUSION

Intelligibility has predominantly been evaluated from the perspective of native English listeners responding to L2 speech, yet there is no valid reason to suggest native listeners' perceptions as basing the pedagogical solutions for teaching pronunciation. The results of this study found no essential differences between native and nonnative listeners in their perception of L2 speech. Concerning listener effects, no evidence for interlanguage speech intelligibility benefit was found for nonnative listeners. In this study, the nonnative listeners never outperformed the native English listeners in the forced-choice word identification task. In contrast, the properties of the L2 speech derived from the speakers' L1 phonological system appeared to be related to intelligibility. These findings should be refined by future research considering other listener and speaker factors related to intelligibility not examined in this study.

6. REFERENCES

[1] J. Levis, "Revisiting the Intelligibility and Nativeness Principles," *JSLP*, vol. 6, no. 3, pp. 310–328, Nov. 2020.

[2] M. J. Munro and T. M. Derwing, "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, Mar. 1995.

[3] R. Thomson, "Measurement of accentedness, intelligibility, and comprehensibility," in *Assessment in Second Language Pronunciation,* O. Kang and A. Ginther, Eds., 1st ed. Routledge, 2017, pp. 11– 29.

[4] C. L. Rogers and J. Dalby, "Forced-Choice analysis of segmental production by Chinese-accented English speakers," *J Speech Lang Hear Res*, vol. 48, no. 2, pp. 306–322, Apr. 2005.

[5] O. Kang, R. I. Thomson, and M. Moran, "Which features of accent affect understanding? exploring the intelligibility threshold of diverse accent varieties," *Applied Linguistics*, vol. 41, no. 4, pp. 453–480, Aug. 2020.

[6] R. Hayes-Harb, B. L. Smith, T. Bent, and A. R. Bradlow, "The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts," *Journal of Phonetics*, vol. 36, no. 4, pp. 664–679, Oct. 2008.

[7] O. Kang and D. L. Rubin, "Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation," *Journal of Language and Social Psychology*, vol. 28, no. 4, pp. 441–456, Dec. 2009.

[8] B. Chung and H. K. M. Bong, "Intelligibility of Korean-accented English: Effects of listener's familiarity," *English Teaching*, vol. 76, no. 1, pp. 33–56.

[9] X. Xie and C. A. Fowler, "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English," *Journal of Phonetics*, vol. 41, no. 5, pp. 369–378, Sep. 2013.

[10] I. A. Jułkowska and J. Cebrian, "Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech," *JSLP*, vol. 1, no. 2, pp. 211–237, Oct. 2015.

[11] M. J. Munro, T. M. Derwing, and S. L. Morton, "The mutual intelligibility of L2 speech," *Stud. Sec. Lang. Acq.*, vol. 28, no. 01, Mar. 2006.

[12] B. Zielinski, "The Segmental/Suprasegmental Debate," in *The Handbook of English Pronunciation*, 1st ed., M. Reed and J. M. Levis, Eds. Wiley, 2015, pp. 397–412.

[13] I. Y. Na, "The impact of segmental accuracy on intelligibility," [Master's Scholarly Paper, University of Hawai'i at Mānoa]. ScholarSpace.

[14] P. Avery, S. Ehrlich, *Teaching American English pronunciation*. Oxford, England: Oxford University Press, 1992.

[15] L. Plonsky and F. L. Oswald, "How Big Is 'Big'? Interpreting effect sizes in L2 Research: Effect sizes in L2 Research," *Language Learning*, vol. 64, no. 4, pp. 878– 912, Dec. 2014.

[16] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Routledge, 2013.

EFFECT OF CONTEXT *F*0 ON QUANTITY IN FINNISH

Michael O'Dell[1], Tommi Nieminen[1,2], Joonas Vakkilainen[2]

[1]University of Helsinki, [2]Tampere University
michael.odell@helsinki.fi, tommi.k.nieminen@tuni.fi, joonas.vakkilainen@tuni.fi

ABSTRACT

There is evidence that perceived quantity in Finnish is affected by *F*0 manipulation, at least within stressed vowels in accented position. We investigated whether unstressed vowels would be affected by *F*0 manipulation outside the vowel. Recordings of two Finnish sentences were manipulated for total duration and *F*0 range of context, keeping unstressed target vowel duration and *F*0 constant. Native subjects then judged whether the target vowel was long or short. Results show that a long context sentence makes a constant target vowel sound short and vice versa, but more importantly, that changes in *F*0 range can also shift the quantity boundary of an unstressed vowel, even though the *F*0 contour of the vowel itself does not change. This suggests that *F*0 can interact with quantity because manipulations which change perceived tempo will also affect quantity perception, which is judged relative to tempo rather than being based on absolute durations.

Keywords: Finnish, quantity, *F*0, tempo

1. INTRODUCTION

Finnish is a full fledged quantity language, with a long vs. short distinction for both consonants and vowels with very few syntagmatic restrictions. The opposition between long and short is possible for vowels in all syllables, stressed or unstressed, and likewise for consonants at all word internal syllable boundaries [1], [2].

Perception tests have shown that varying the *F*0 contour ("dynamic" vs. flat, high vs. low, increasing slope, shifting location of a fall, etc.) within a vowel can shift the duration boundary between long and short for many languages that have a quantity opposition (cf. e.g. [3] for a partial survey). The same has been found for Finnish [2], [4]–[6]. These perception tests have studied accented vowels (in isolated words or in carrier sentences) in word initial stessed position, a position which often exhibits an *F*0 peak in Finnish. The interaction between quantity and *F*0 in Finnish has been taken to show an affinity with tone languages, since in both cases a change in *F*0 can affect the perceived lexical identity of words [4]–[7].

Another possible interpretation is that manipulating *F*0 contour affects perceived local tempo (speaking rate). Naturally this should also affect perceived quantity, since quantity distinctions cannot be based on absolute durations, but must be judged relative to speaking rate. This type of explanation has been offered for perception data from Danish [8], Japanese [9], and Finnish [2], as well as for Finnish production data [10]. If we assume that *F*0 movements associated with sentence accent progress at a fairly constant rate for constant tempo, then more of that movement will happen during a longer vowel than during a shorter vowel. An *F*0 contour familiar to the listener will then also provide a cue for tempo—*F*0 movement progressing faster than expected will cue a faster tempo and vice versa. A local tempo model has also been suggested in connection with perception data from Estonian, Finnish, Norwegian and Italian [11]–[13], but for temporal rather than *F*0 manipulation of the context.

Any factor which affects perceived tempo in the environment of a possible quantity distinction should automatically also affect perception of that quantity. It might be expected that such an effect would be strongest for manipulations (including *F*0) during the relevant quantity domain itself, but manipulations in the immediate vicinity of the quantity being perceived might also have a sizeable effect. For instance, several studies have shown that changing context rate by means of manipulating durations can have a large impact on whether whole syllables are perceived (so-called distal speech rate effect [14]–[16]).

In the present experiment, we wanted to see whether manipulating *F*0 in the *environment* of a vowel, rather than during the vowel itself, might also shift the quantity boundary in the perception of that vowel. In addition, we wanted to investigate Finnish vowel quantity perception in unstressed, unaccented positions, where there is typically a lack of *F*0 movement.

2. MATERIAL AND METHODS

Two native speakers of Finnish (TN & JV, the second and third authors) were recorded producing several sets of minimal pair sentences at three different tempos. The sentence pairs differed only in the phonological length (short vs. long) of the final vowel of one word. Speakers produced the sentences two times each for three raters, first at a normal rate, then at a faster than normal rate, and finally at a slower than normal rate.

As expected, durations of carrier sentences and vowels both increase from fast to normal to slow tempo. On average it would appear that long vowels are affected by tempo changes more than short vowels, but long vowels are always longer in duration than the corresponding short vowels at all tempos, even though the ratio of long to short is reduced somewhat at faster tempos.

2.1. Stimuli

Two minimal pairs were selected for further use. In the first pair (speaker JV) the word (and final vowel) in question occurred at the end of the sentence:

(I) *Se on tosi ärsyttävä./ärsyttävää.*

In the second pair (speaker TN) the word and final vowel occurred in the middle of the sentence:

(II) *Se on tosi helppo/helppoo tuottaa ittekkin.*

Two tokens from each sentence pair were selected to calculate target durations for synthesizing stimuli for the perception experiment: the fast tempo sentence with the shortest long vowel and the slow tempo sentence with the longest short vowel.

A five step stimulus series was constructed for each sentence using Praat to manipulate durations in the following manner. The duration of the target vowel was kept constant at the geometric mean of the fast long vowel and the slow short vowel: 92 ms for stimulus series I (*Se on tosi ärsyttävä/ärsyttävää*), and 67 ms for stimulus series II (*Se on tosi helppo/helppoo tuottaa ittekkin*). The duration of the context (the rest of the sentence excluding the target vowel) was varied systematically so that the range between the durations of the two original contexts was equally divided into six equal steps on a logarithmic scale (i.e. the ratio of durations for adjacent steps in the series was constant). Stimuli 1–5 in the stimulus series were synthesized using the resulting central five durations. In other words, the context durations of the original sentences (not used in any stimuli) could be thought of as steps 0 and 6 in the series of equal steps. The resulting durations of target vowels and contexts are shown in Table 1 for both series,

Table 1: Stimulus structure

	Duration (ms)				
	series I:		series II:		
	preceding context	vowel	preceding context	vowel	following context
1:	772	92	568	67	958
2:	874	92	624	67	1033
3:	989	92	685	67	1114
4:	1119	92	751	67	1202
5:	1267	92	824	67	1296
	$F0$ range (semitones)				
A:	14.78		11.12		
B:	7.10		5.04		

and also graphically in Figure 1 for series I.

As expected, in spite of having the same duration, the target vowel in stimulus 1 with the shortest (fastest) carrier sentence unambiguously sounds long, whereas in stimulus 5 with the longest (slowest) carrier sentence it sounds short.

Because our main interest was to see whether $F0$ fluctuation affects quantity perception (presumably mediated by a difference in perceived tempo), we also manipulated the $F0$ contours of stimuli. For each stimulus series two versions A and B were synthesized using two different $F0$ contours. Both versions kept $F0$ of the target vowel constant (original contour). Average $F0$ during the target vowel also served as a baseline around which the $F0$ contour for the rest of the sentence was expanded or reduced. For version A, to create an expanded $F0$ range, deviation of $F0$ (in semitones) from baseline outside the target vowel was multiplied by 1.5. For version B, to create a reduced $F0$ range, deviation of $F0$ (in semitones) from baseline outside the target vowel was divided by 1.5. Thus the $F0$ ranges for the A and B versions differed from each other (in semitones) by a factor of 2.25. The resultant $F0$ ranges are shown in Table 1. Figure 2 shows the resulting $F0$ contours for series I.

One original normal tempo sentence with a long test vowel was selected for the spectral basis of the manipulated stimuli in each series. Thus no stimulus was identical to an original token, all stimuli were manipulated for both duration and $F0$. All in all a total of 20 stimuli were thus produced: 5 context duration steps (1, 2, 3, 4, 5) $\times 2$ $F0$ versions (A, B) $\times 2$ series (I, II).

2.2. Perception test

19 subjects took part in the perception test, 5 male, 13 female, 1 other, all university students aged 20–29 years (median age 22) at the time of the test.

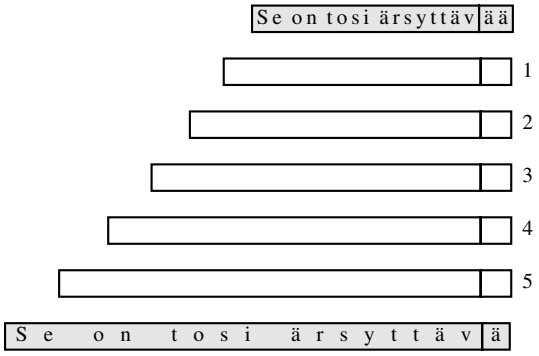


Figure 1: Durations for stimulus series I, steps 1–5: *Se on tosi ärsyttävä/ärsyttävää* (speaker JV, original sentences in gray)

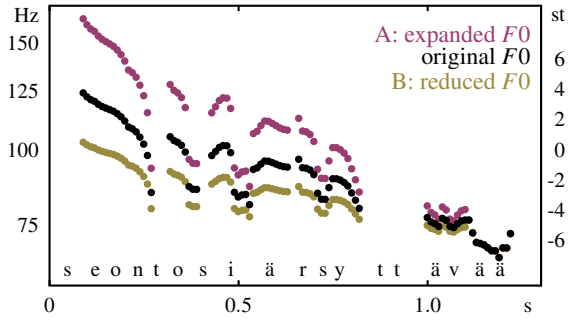


Figure 2: $F0$ for simulus series I, versions A & B: *Se on tosi ärsyttävä/arsyttävää* (speaker JV)

Five repetitions of each stimuli were presented to subjects in random order, with the restriction that a stimulus from the same series (sentence I or II) was never presented more than three times in a row. Subjects were required to underline on an answer sheet which of the two possible words they heard (i.e. *ärsyttävä* vs. *ärsyttävää* or *helppo* vs. *helppoo*). They were instructed to make a decision even if they were unsure which word they heard.

2.3. Statistical model

A Bayesian hierarchical logistic regression model [17] was used to analyze the responses of the perception test. In this model probability of a long vowel response p is modeled as a logistic function of the context duration (or tempo) step x.

$$\text{(1)} \quad \text{logit}(p) = \beta \cdot (x - \mu)$$

Both intercept μ (50 % boundary between long and short vowel response) and slope β are allowed to vary by main effects subject (SUBJ), stimulus series (sentence, SENT) and $F0$ version (F0), as well as all interaction effects. Slope parameters β are restricted to be positive using a logarithmic scale.

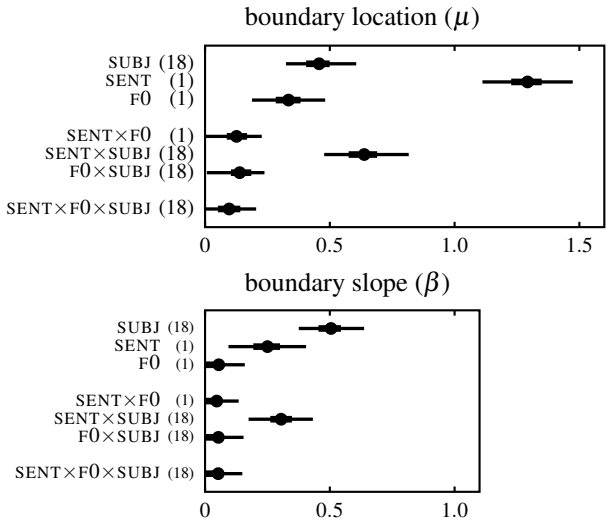


Figure 3: Posterior distributions (median, 50 % and 95 % HDI) for size of boundary location (μ) and slope (β) effects (standard deviations of effects, degrees of freedom in parentheses)

3. RESULTS

Estimated (posterior) standard deviations (SD) of the effects in the model are presented visually in Figure 3. These standard deviations give an idea of the importance of each effect [17]. In this figure the posterior distributions have been stylized: the thin line shows the 95 % HDI (Highest Density Interval), the thick line shows the 50 % HDI, and the dot shows the median of the posterior distribution.

Not surprisingly, the SD for the SUBJ main effects are clearly non-zero, indicating that subjects differed among themselves, both in (50 %) boundary location μ, and in slope β. Also the two sentences clearly produced different μ and β, and there was a clear SENT×SUBJ interaction. The difference in slope for the two sentences (sentence II had a slightly shallower slope) is probably due to the fact that the stimulus steps in series II were smaller because the original endpoints used to construct the stimuli were not as far apart.

The main interest for this experiment is whether expanding and reducing the $F0$ contour of the carrier had any affect on perceived quantity of the target vowel. As can be seen in Figure 3, the F0 main effect is clearly non-zero for boundary location μ. The F0 main effect for slope β is very small or zero, as are all interactions involving F0, for both μ and β. This indicates that changing context $F0$ contour (from range A to range B) shifted the identification curve towards shorter (faster) or longer (slower) contexts, but did not greatly affect its slope (or boundary width). Figure 3 does not indicate the

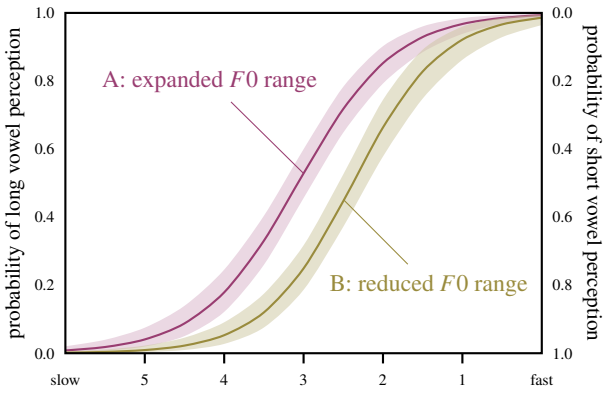


Figure 4: Posterior average identification curves (median and 95 % CI) for stimulus series I

direction of this shift, but looking at the posterior distributions of the individual parameters reveals that A (expanded $F0$ range) shifts the boundary towards slower context and B (reduced range) towards faster context. This difference is illustrated in Figure 4, which shows the posterior average (with all SUBJ effects removed) identification curves for stimulus series I.

The shift observed is exactly what is expected if greater $F0$ range has the effect, *ceteris paribus*, of increasing perceived tempo, because in that case it should be easier to hear a long vowel in the context of expanded $F0$ range (A).

4. DISCUSSION AND CONCLUSION

In a series of experiments on Estonian, Finnish, Norwegian and Italian [11]–[13] looking at the influence on quantity perception of durational variation outside the domain of quantity, Krull *et al.* concluded that "The results were compatible with a model of speech perception where an 'inner clock' handles variations in the speaking rate ..." [13]. The present results are also compatible with a local speaking rate model such as Traunmüller's modulation model [18].

Aulanko found a greater $F0$ range for Finnish long vowels compared with short vowels in the first, stressed syllable of two syllable words embedded in carrier sentences, due to the fact that although the beginning of the falling $F0$ contour was similar in both long and short vowels, it extended further downward in the long vowels. He speculated that "This might be a feature that is used unconsciously by Finns in perceiving linguistic quantity degrees: a vowel with a relatively short duration but with a great F_0 movement might tend to be perceived as long." [19, p. 48]. Vihanta reported similar results in a larger investigation utilizing pairs of semantically plausible sentences differing in quantity patterns. He concluded that it is probably the case that a change in $F0$, determined by sentence prosodic factors, simply progresses further during a long phoneme [10]. Suomi's production data for Finnish words of various lengths and quantity patterns embedded in carrier sentences in several accentual conditions [20], [21] showed an almost invariant $F0$ contour for each accentual condition, independent of target word length and quantity structure. All these results suggest that $F0$ contour, which is fairly constant as long as tempo and intonational structure remain unchanged, provides a robust indication of local tempo in Finnish.

It has been known for some time that changes in $F0$ (range, level, dynamics) can affect perceived quantity. Previous studies have usually manipulated $F0$ *within* the segment perceived as phonologically long or short. It now appears that changes in $F0$ *outside* that segment can affect perceived quantity as well. This means that rather than postulating various "lexical tones" associated with various quantities it is simpler, at least in the case of Finnish, to explain the effects of $F0$ on perceived quantity as mediated by tempo. In any case quantity differences must be perceived in relation to tempo, and perceived tempo (and the passage of time) is known to be affected by many factors including $F0$ movements.

The inevitable interaction between $F0$ contour and durational or rhythmic aspects of speech does, of course, mean that there can be no clear cut distinction between pure tone languages and pure quantity languages. Indeed, it is misleading to think of any language in terms of pure, mutually exclusive categories. In the course of history it is possible that a language such as Finnish might gradually come to rely more and more on local $F0$ movements to distinguish words previously distinguished mainly by durational or rhythmical means, especially if quantity distinctions came to be restricted to enviroments, such as accents, with stereotypical $F0$ movements.

However, in present day Finnish, it may be that $F0$ contours provide a very reliable cue to local tempo (thereby also influencing quantity) precisely because Finnish does not have lexical tone. Rather, $F0$ "is tied more closely to a different level of timing (eg. sentence intonation), and may thus provide a reference for perception of quantity." [2]. Put in a different way, Finnish quantity is not tonal, rather tonality (intonation) provides the listener with a strong cue for synchronizing to the time scale of the speaker at the level of quantity.

5. REFERENCES

[1] J. Lehtonen, *Aspects of Quantity in Standard Finnish* (Studia Philologica Jyväskyläensia VI). University of Jyväskylä, 1970.

[2] M. O'Dell, *Intrinsic Timing and Quantity in Finnish* (Acta Universitatis Tamperensis 979). Tampere University Press, 2003.

[3] R. Cumming, "The effect of dynamic fundamental frequency on the perception of duration," *Journal of Phonetics*, vol. 39, pp. 375–387, 2011.

[4] M. Vainio, D. Aalto, J. Järvikivi, and A. Suni, "Quantity and tone in Finnish lexically stressed syllables," in *Proceedings of the 2nd International Symposium on Tonal Aspects of Languages*, A. Belotel-Grenié and M. Grenié, Eds., 2006, pp. 121–124.

[5] J. Järvikivi, D. Aalto, R. Aulanko, and M. Vainio, "Perception of vowel length: Tonality cues categorization even in a quantity language," in *Proceedings of the 16th International Congress of Phonetic Sciences*, J. Trouvain and W. J. Barry, Eds., Universität des Saarlandes, Saarbrücken, Germany, 2007, pp. 693–696.

[6] J. Järvikivi, M. Vainio, and D. Aalto, "Real-time correlates of phonological quantity reveal unity of tonal and non-tonal languages," *PLoS ONE*, vol. 5, no. 9, e12603, 2010.

[7] J. Šimko, D. Aalto, P. Lippus, M. Włodarczak, and M. Vainio, "Pitch, perceived duration and auditory biases: Comparison among languages," in *Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK*, The Scottish Consortium for ICPhS 2015, Ed., University of Glasgow, 2015, paper 575, ISBN: 978-0-85261-941-4.

[8] N. Thorsen, "F_0 timing in Danish word peception," *Phonetica*, vol. 41, pp. 17–30, 1984.

[9] Y. Nagano-Madsen, *Mora and Prosodic Coordination: A Phonetic Study of Japanese, Eskimo and Yoruba*. Lund University Press, 1992.

[10] V. V. Vihanta, "F_0:n osuudesta suomen kvantiteetti-oppositiossa," in *Fonetiikan päivät — Espoo 1988 / Papers from the 15th Meeting of Finnish Phoneticians*, M. Karjalainen and U. K. Laine, Eds., ser. Teknillinen Korkeakoulu, Sähkötekniikan osasto, Akustiikka ja äänenkäsittely / Helsinki University of Technology, Faculty of Electrical Engineering, Acoustics Laboratory, (résumé: Sur le role de la F_0 dans l'opposition de quantité en finnois), Helsinki University of Technology, 1988, pp. 13–37.

[11] D. Krull, H. Traunmüller, and W. A. van Dommelen, "The effect of local speaking rate on perceived quantity: A comparison between three languages," in *Proceedings of the 15th International Congress of Phonetic Sciences*, M. J. Solé, D. Recasens, and J. Romero, Eds., Universitat Autònoma de Barcelona, Spain, 2003, pp. 1739–1742.

[12] H. Traunmüller and D. Krull, "The effect of local speaking rate on the perception of quantity in Estonian," *Phonetica*, vol. 60, pp. 187–207, 2003.

[13] D. Krull, H. Traunmüller, and P. M. Bertinetto, "Local speaking rate and perceived quantity: An experiment with Italian listeners," *Working Papers (Lund University, Dept. of Linguistics)*, vol. 52, pp. 81–84, 2006. [Online]. Available: https://www2.ling.su.se/staff/hartmut/quantity_italian.pdf.

[14] L. C. Dilley and M. A. Pitt, "Altering context speech rate can cause words to appear or disappear," *Psychological Science*, vol. 21, no. 11, pp. 1664–1670, 2010. [Online]. Available: http://journals.sagepub.com/doi/pdf/10.1177/0956797610384743.

[15] L. C. Dilley, T. H. Morrill, and E. Banzina, "New tests of the distal speech rate effect: Examining cross-linguistic generalization," *Frontiers in Psychology*, vol. 4, no. 1002, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2013.01002/full#B18a.

[16] M. O'Dell and T. Nieminen, "Distal rate effect for Finnish epenthetic vowels," in *Proceedings 9th International Conference on Speech Prosody 2018*, 2018, pp. 646–650. DOI: 10.21437/SpeechProsody.2018-131.

[17] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[18] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, no. 1–3, pp. 170–183, 1994.

[19] R. Aulanko, "Microprosodic features in speech: Experiments on Finnish," in *XIII Fonetiikan päivät Turku 1985 / XIII Meeting of Finnish Phoneticians — Turku 1985*, O. Aaltonen and T. Hulkko, Eds., ser. Publications of the Department of Finnish and General Linguistics of the University of Turku, 1985, pp. 33–54.

[20] K. Suomi, "Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language," *Journal of Phonetics*, vol. 33, pp. 291–309, 2005.

[21] K. Suomi, "On the tonal and temporal domains of accent in Finnish," *Journal of Phonetics*, vol. 35, no. 1, pp. 40–55, 2007.

PERCEPTION OF JAPANESE PITCH ACCENT BY DUTCH LISTENERS: FROM ACOUSTIC TO PHONOLOGICAL PROCESSING

Shuangshuang Hu[1], René Kager[2]

[1] Shanghai Normal University, China [2] Utrecht University, The Netherlands
[1] shuangshuanghu@shnu.edu.cn, [2] r.w.j.kager@uu.nl

ABSTRACT

Both Dutch and Japanese employ positional marking for lexical contrasts, i.e., word stress position in Dutch and pitch accent position in Japanese. The present study investigated whether naive Dutch listeners could transfer a native positional sensitivity to perceive the position of Japanese lexical pitch accents. An AX discrimination task and a modified sequence recall task were applied to examine Dutch listeners' perception of Japanese pitch accents at the acoustic level and the phonological level, respectively. It was found that naive Dutch listeners showed as good performance as Japanese natives when perceiving Japanese pitch accent contrasts at the acoustic level. However, they showed difficulty when perceiving Japanese pitch accent contrasts at the phonological level while the natives remained good performance. The findings suggest that the positional sensitivity in the native language may facilitate the perception of non-native positional contrasts acoustically, but it may not be reliable enough at the phonological level.

Keywords: perception of Japanese pitch accent, phonological processing, positional marking.

1. INTRODUCTION

Languages differ in the use of prosodic cues to contrast lexical meanings. For instance, Dutch uses word stress to differentiate words. One syllable of any multi-syllabic word is typically marked with higher stress than the other syllable(s) [1, 2]. The meanings of words with identical segments can then be signaled by the position of stress, e.g., *VOORnaam* (stress on the first syllable) means "first name" while *voorNAAM* (stress on the second syllable) means "distinguished".

Tokyo Japanese (henceforth Japanese) uses pitch accent for lexical contrast. The meaning of a word in Japanese is determined by the presence or absence of an abrupt pitch fall (termed "pitch accent") and, if present, by the position of pitch accent in any multi-moraic word [3, 4]. For instance, the disyllabic (bimoraic) word "hashi" /haʃi/ signals three lexical meanings: "chopsticks" (initial-accented), "bridge" (final-accented) and "edge" (unaccented).

Prima facie, the two languages employ different word prosodic cues in their lexicons. Nonetheless, they do share commonalities: Japanese and Dutch use the abstract feature of positional marking: position of pitch accent in Japanese and position of stress in Dutch.

Previous research has underscored that Dutch listeners are sensitive to perceive stress position and can exploit positional marking for word recognition in their native languages [5, 6]. It is of interest to ask whether Dutch listeners can transfer a native positional sensitivity to perceive a non-native positional cue, Japanese pitch accent.

Studies on the perception of non-native positional cues have mainly focused on word stress. It is found that stress language listeners can perceive the positional contrasts of a non-native word stress at the acoustic level, i.e., via acoustic approaches such as AX and AXB discrimination tasks. However, they have difficulty to process the non-native stress when the task taps into phonological representation of stress, namely in a sequence recall task which incorporated phonetic variability and memory load [7, 8]. These findings suggest that the ability to perceive positional cues at the acoustic level does not guarantee a success at the phonological level.

Thus, the present study aims to examine whether stress language listeners, Dutch listeners, can generalize their sensitivity to native positional marking to positional contrasts of Japanese pitch accent (including the positional marked vs. unmarked contrast - accented vs. unaccented), not only acoustically but also phonologically. Via a discrimination task and a sequence recall task, respectively, the study attempts to investigate to what extent the native word prosody influences perception of non-native positional marking at the acoustic level and the phonological level in terms of perceptual models.

2. METHOD

2.1. Subject

40 Dutch listeners (mean age: 22 years old, SD=3.8, 16 males) and 36 Tokyo Japanese listeners (mean

age: 24 years old, SD=5.2, 15 males) participated in the two tasks in the study. All the listeners reported normal hearing without language impairment. None of the Dutch participants have been exposed to any tonal or pitch accent languages. All the participants were non-musicians.

2.2. Stimuli

AX discrimination task and sequence recall task were applied to investigate perception of Japanese pitch accent contrasts at the acoustic level and the phonological level, respectively.

2.2.1. AX discrimination task

Given acoustic properties of consonants and vowels [9], /no/ and /jo/ were used for constructing disyllabic nonwords /nono/ and /jojo/. Both the two words do not exist in Dutch or Japanese[i]. According to [10], disyllabic words in Japanese carry three accentual patterns with corresponding surface tonal patterns (apostrophe ’ symbolizes the accent) as shown in Table 1.

accentual contrast (positional contrast)	surface tonal pattern
initial-accented vs. final-accented	H’L vs. LH’
initial accented vs. unaccented	H’L vs. LH
final-accented vs. unaccented	LH’ vs. LH

Table 1: Stimuli.

All tokens were embedded in a carrier sentence with declarative intonation:

kinoo _____ (target nonword) ga kita.

“Yesterday_____(target nonword) nominative marker came”.

The nominative marker “ga” serves to differentiate the unaccented and the final-accented word [11]. All sentences were produced by a female Japanese phonetician in Praat in a sound-proof room. The targeted tokens were extracted from the carrier sentences and were normalized as 96 ms, 70 dB.

The overall design of AX discrimination task is: contrast (3) x the identity of X (2: XA, XA) x token set (2: /nono/, /jojo/) x repeat times (2) with fillers (12 AA pairs), containing 36 trials in total. The inter-stimulus interval (ISI) in each trial was 400 ms. Note that the same trials did not show up consecutively.

2.2.2. Sequence recall task

The sequence recall task integrated a memory load and phonetic variability [12]. It aims to eliminate listeners’ reliance on fine-grained acoustic details and taps into phonological representation.

In this modified task, four contrasts were constructed: one segmental contrast, i.e., nonword /nogo/ vs. /nopo/ as the baseline and three accentual contrasts as shown in Table 1, carried by nonword /nono/. All the accentual tokens were produced six times each by three female and three male Japanese phoneticians and were extracted and manipulated in the same fashion as those in AX discrimination task. Segmental tokens were produced in isolation with a flat tone by the same six speakers and normalized as same as the accentual ones. For each speaker, three items of the best quality were selected.

Two-word, three-word and four-word sequences were used, with ISI set to 80 ms as in [12]. In each sequence, the nonwords were produced by different speakers. The order of the voices was counterbalanced over the sequences.

There were eight trials of each sequence length, as follows:

Two-word length: AA, BB, AB, BA, AA, BB, AB, BA

Three-word length: AAA, AAB, ABA, ABB, BBB, BBA, BAB, BBB

Four-word length: AABA, ABAA, ABBA, BAAB, BABB, BBAB, ABAB, BABA

The overall design is: contrast (4) x trials (24), yielding 96 trials in total.

2.3. Procedure

The two tasks were programmed and conducted in ZEP [13] on an experiment laptop.

2.3.1. AX discrimination task

All the participants were instructed to listen to two tokens from a foreign language. They were required to judge as soon as possible whether the two tokens were the same or not by pressing the corresponding buttons “Same” or “Different” on the button box. The trial proceeded only after the participant made the response.

The task consisted of a practice phase (four trials) and a test phase. Feedback was only provided in the practice phase. All the trials were counterbalanced in the test phase.

2.3.2. Sequence recall task

All the participants were instructed that they were going to learn four pairs of new words in a foreign language. They learned the segmental contrast first, and then the three accentual contrasts which were counterbalanced across participants.

Each contrast contained 5 phases. The participants were instructed that they were going to learn new words A and B, associating with buttons

A and B, respectively. In Phase 1, participants first listened to 6 tokens of word A followed by 6 tokens of word B. After learning words A and B, they proceeded to Phase 2 where they could press button A or B to listen to the words repeatedly as many times as they wished to make sure they had memorized the two words. In Phase 3, they did a practice in which they heard a word and judged whether the word was word A or B by pressing the corresponding button. There were 20 trials in the practice, in which they were required to reach a criterion of 75% accuracy. If they failed in the practice, they would go back to Phase 2 to learn the two words again until they could reach the passing criterion for Phase 3. They then proceeded to Phase 4 where they listened to words A and B in two-word and three-word sequences. They were required to recall the sequences by pressing the corresponding buttons. For instance, if they heard A-B, they should press button A firstly and B secondly in an A-B order. After finishing Phase 4, they proceeded to Phase 5, the test phase, containing 3 blocks (8 trials per block). Blocks 1, 2 and 3 were words A and B in two-word, three-word, and four-word sequences, respectively. A feedback message "OK" would appear on the screen, simultaneously presented with a sound message "okay" once they responded to each trial to eliminate the possible use of echoic memory by the participants [14].

3. RESULTS

3.1. AX discrimination task

A correct response made by a participant was marked as "1" while an incorrect response was marked as "0". Any missing response was regarded as an incorrect response. To compare the performance between the two groups, a generalized Linear Mixed Model (GLMM) was conducted in SPSS 26. Contrast (3 levels) and Language (2 levels) were taken as fixed factors and intercepts for participant and item were taken as random effects into the model. Neither Contrast (F (2, 1818)= 2.523, $p=0.051$) nor Language (F (1, 1818) = 0.036, $p=0.849$) were found significant. No interaction between Contrast and Language was found as well (F (2, 1818) = 0.382, $p=0.683$).

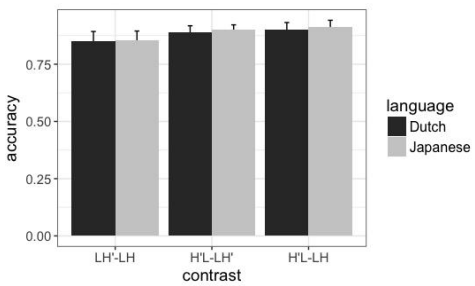


Figure 1: Dutch and Japanese listeners' performance in AX discrimination task.

As shown in Figure 1, Dutch listeners perceived all the three accentual contrasts as well as Japanese natives, both groups reaching around 85% accuracy for LH' vs. LH and around 90% accuracy for H'L vs. LH' and H'L vs. LH.

3.2. Sequence recall task

A response is correct if the entire sequence is recalled correctly. Correct responses were marked as "1" while incorrect responses were marked as "0". Any missing response was regarded as an incorrect response. A GLMM was computed in SPSS 26 to analyze whether the two groups differed in perception of the accentual contrasts at the phonological level. Contrast (4 levels: one segmental contrast and three accentual contrasts), Sequence Length (3 levels: 2-, 3- and 4-word length) and Language (2 levels) were taken as fixed factors, and intercepts for participant and item were taken as random effects into the model. The F-tests showed that Contrast (F (3, 7278) = 74.846, $p < 0.001$) and Language (F (2, 7278) = 20.050, $p < 0.001$) had main effect. The interaction between the two factors was found significant as well (F (3, 7278) = 13.998, $p<0.001$). Sequence Length (F (2, 7278) = 1.919, $p = 0.304$) was not significant, with no interaction with Language (F (2, 7278) = 1.027, $p = 0.358$), or Contrast (F (6, 7278) = 0.725, $p = 0.630$), indicating that the differences among language groups and contrasts were comparable in each sequence length.

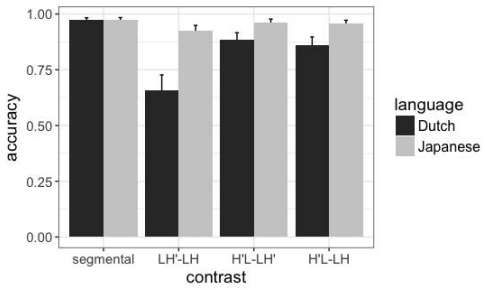


Figure 2: Dutch and Japanese listeners' performance in sequence recall task.

Figure 2 displays the performance of each language group in perceiving segmental contrast (control condition) and accentual contrasts (accentual condition) with sequences of different lengths collapsed. In perceiving segmental contrast, the two groups achieved nearly ceiling effect, both with accuracy around 97%. However, they showed significant differences in perceiving accentual contrasts. Japanese natives outperformed Dutch

listeners in perceiving all the accentual contrasts (F (3, 7278) = 25.721, $ps<0.001$).

Dutch listeners performed the best when perceiving segmental contrast, compared with their perception of accentual patterns (F (3, 3834) = 27.739, $ps<0.001$). They achieved around 85% accuracy when perceiving H'L vs. LH' and H'L vs. LH, but their performance of perceiving LH' vs. LH dropping to 65.7% accuracy, significantly worse than that of perceiving the other two contrasts (F (3, 3834) = 27.739, $ps<0.001$). In contrast, Japanese natives remained good performance, with an accuracy around 95% when perceiving all the contrasts. Their perception of accentual contrasts was as good as that of segmental contrast (F (3, 3450) = 1.193, $ps>0.05$).

4. DISCUSSION

In the acoustic task (AX discrimination task), Dutch listeners were found as sensitive as Japanese natives to positional contrasts of Japanese pitch accent. This can be explained by the influence of native phonology [15], which predicts if the native grammar has the phonological feature that differentiates a particular non-native contrast, listeners are able to perceive the contrast. Dutch uses position of word stress, which may benefit them to transfer the native positional sensitivity to perceive Japanese pitch accented contrast initial-accented vs. final-accented (H'L vs. LH'). For accented vs. unaccented contrasts (H'L vs. LH, LH' vs. LH), although Dutch lexicon does not employ positional marked vs. unmarked, Dutch listeners may still distinguish the accented word from the unaccented one because one in the pair is positional marked. Another account is that Dutch listeners may resort to perceive the surface tonal patterns of the accentual contrasts acoustically. According to PAM [16], Dutch listeners may map H'L, LH' and LH onto the native nuclear pitch accents H*L, L*H and L*(H%) [17] in intonation categories, respectively, which led to good perception. However, such perceptual association between the non-native Japanese accentual patterns and Dutch intonation categories is in nature speculative.

When the task highlights the phonological representation of Japanese pitch accent, Dutch listeners' perceptual sensitivity seemed be mediated. They were not as advantageous as Japanese natives who remained good performance. The findings suggest that Dutch listeners' ability of perceiving Japanese pitch accents at the acoustic level does not entail the success at the phonological level. It is notable that Dutch listeners showed significant difficulty, with accuracy around 65%, when processing final-accented vs. unaccented (LH' vs. LH) at the phonological level. If according to the model in terms of the influence of native phonology [15], Dutch listeners should be able to employ the native positional marking to differentiate the marked final-accented from the unaccented word, which contradicted the finding.

Alternatively, the trouble Dutch listeners encountered in processing final-accented vs. unaccented (LH' vs. LH) phonologically could be due to the failure to perceive the surface tonal contrasts realized by the accentual patterns. Phonetically speaking, the H tone in the final accented position in a word is slightly higher than the H tone in the final position when the word is unaccented [11]. Dutch listeners were sensitive to perceive LH' vs. LH at the acoustic level, which could be in that they could rely on fine-grained acoustic information in AX task. Previous studies have reported that they were good at discriminating subtle pitch differences when perceiving non-native lexical tones acoustically [18, 19]. However, when it taps into phonological representation of lexical pitch accent, Dutch listeners may no longer rely on pitch differences. In Dutch lexicon, pitch is one of the acoustic correlates of stress, that is, the stressed syllable in a word is realized by a higher pitch together with longer duration and larger amplitude [20]. Pitch is not the exclusively acoustic cue to word stress in Dutch whereas tonal patterns in Japanese lexicon are realized and determined by the positional cue, the accent. The lack of use of lexical pitch in Dutch might lead to the failure of perceiving Japanese pitch accent contrasts by Dutch listeners at the phonological level.

To sum up, the current findings showed that Dutch listeners were able to perceive the positional contrasts of Japanese pitch accent acoustically. However, they had difficulty when processing the non-native positional cues at the phonological level. This seemed to indicate that when perceiving the positional contrasts of Japanese pitch accent at the phonological level, the abstract positional marking in Dutch may not be reliable enough.

5. REFERENCES

[1] Kager, R. 1989. *A Metrical Theory of Stress and Destressing in English and Dutch*. Dordrecht: Foris.

[2] Trommelen, M., Zonneveld, W. 1999. Word stress in English and Dutch. In: Hulst, H. van der (ed), *Word Prosodic Systems of the Languages of Europe*. Berlin: Mouton de Gruyter, 478-515.

[3] Poser, W. J. 1984. *The Phonetics and Phonology of Tone and Intonation in Japanese*. Ph.D. dissertation. Massachusetts Institute of Technology.

[4] Kubozono, H. 2008. Japanese accent. In: Miyagawa, S., Saito, M. (eds.), *The Oxford Handbook of Japanese Linguistics*. Oxford: Oxford University Press, 165-191.
[5] Cutler, A., van Donselaar, W. 2001. Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech* 44, 171–195.
[6] Jongenburger, W. 1996. *The Role of Lexical Stress during Spoken-word Processing*. Ph.D. dissertation, University of Leiden. The Hague: Holland Academic Graphics.
[7] Peperkamp, S., Dupoux, E. 2002. A typological study of stress "deafness". *Laboratory Phonology* 7, 203-240.
[8] Peperkamp, S., Vendelin, I., Dupoux, E. 2010. Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics* 38, 422-430.
[9] Lehiste, I. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.
[10] Haraguchi, S. 1999. Accent. In: Tujimura, T. (ed) *The Handbook of Japanese Lingusitics.* Oxford: Blackwell, 1-30.
[11] Vance, T. J. 1995. Final accent vs. no accent: Utterance-final neutralization in Tokyo Japanese. *Journal of Phonetics* 23(4), 487–499.
[12] Dupoux, E., Peperkamp, S., Sebastian-Galle, N. 2001. A robust method to study stress "deafness". *Journal of the Acoustical Society of America* 110 (3), 1606-1618.
[13] Veenker, T. J. G. 2012. The Zep Experiment Control Application. Utrecht Institute of Linguistics OTS, Utrecht University.
[14] Morton, J., Crowder, R., Prussin, H. 1971. Experiments with the stimulus suffix effect. *Journal of Experimental Psychology* 91, 169-190.
[15] Brown, C. 1998. The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research* 14, 136 ‑ 193.
[16] So, C. K., Best, C. T. 2010. Cross-language perception of nonnative tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech* 53, 273-293.
[17] Gussenhoven, C. 2005. Transcription of Dutch Intonation. In: S.-A., Jun (ed.), *Prosodic Typology and Transcription: A Unified Approach*. Oxford: Oxford University Press, 118-145.
[18] Leather, J. 1990. Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In: Leather, J., James, A. (eds.), *New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second Language Speech*. Amsterdam: University of Amsterdam, 72-89.
[19] Liu, L., Chen, A., Kager, R. 2017. Perception of tones in Mandarin and Dutch listeners. *Language and Linguistic*s 18(4), 622-646.
[20] Sluijter, A. M. C. 1995. *Phonetic Correlates of Stress and Accent*. Holland Institute of Generative Linguistics Dissertations 15. The Hague: Holland Academic Graphics.

[1] Note that the disyllabic word /jojo/ (the approximant [j] and a short [o]) is a non-word in both the languages in that the Japanese word "jojo" in the animation "Jojo's Bizarre Adventure" is /dʒo.dʒo/, not /jojo/. "Jojo" (a toy) in Dutch is /ˈjəʊ.jəʊ/.

PERCEPTION OF COARTICULATED TONES IN TAIWAN MANDARIN AND TAIWAN SOUTHERN MIN

Huang, Po-Hsuan[1]
Chiu, Chenhao[2]

Graduate Institute of Linguistics, National Taiwan University[1, 2]
Neurobiology and Cognitive Science Center, National Taiwan University[2]

r09142003@ntu.edu.tw[1]
chenhaochiu@ntu.edu.tw[2]

ABSTRACT

Lexical tones in tone languages have been found to coarticulate with ambient tones, and speakers are observed to normalize for such coarticulation. Beijing Mandarin is one of the examples. In languages with larger tone inventories, however, the recoverability of the target tones is lower, and faithful comprehension of coarticulated tones based solely on normalization may be less likely. In this study, perception of coarticulated tones in Taiwan Mandarin and Taiwan Southern Min is investigated. It was revealed that compared with Taiwan Mandarin, Taiwan Southern Min tones induced much less normalizing effects. Crucially, Taiwan Southern Min was found to have stricter tone boundaries than Taiwan Mandarin, presumably allowing for effective tone perception under tonal coariculation despite the larger tone inventory and a lower degree of normalization. This study sheds light on the perception strategies of coarticulated tones in languages with tone inventories of different sizes and the nature of tonal coarticulation.

Keywords: tone perception, normalization, tonal coarticulation, Taiwan Mandarin, Taiwan Southern Min.

1. INTRODUCTION

Lexical tones in tone languages are ideally distinguished with their tone values and shapes. In real-world communication, however, the production and perception of tones is highly variable, subject to factors including prosody [1], syllable duration [2], and tonal coarticulation [1, 3, 4, 5, 6, 7]. In this study, we focus on the last factor, i.e., tonal coarticulation, where a target tone's pitch value and/or pitch contour varies owing to preceding and/or following tones, to the extent that it might be acoustically similar to another lexical tone. Specifically, we focus on the perception of coarticulated tones in Taiwan Mandarin (TM) and Taiwan Southern Min (TSM).

Lexical tones are known to coarticulate with ambient tones [1, 3, 4, 5, 6, 7, 8], and such coarticulation has been found to be perceptible to the speakers [1, 4, 6, 9], and could influence tone perception [4, 6, 9]. For example, in Beijing Mandarin (BM), [4] found that changing ambient tones could affect the identification accuracies of the target tones. The possible confusion of tone perception caused by such variation has been found to be dealt with by means of normalization in BM [4, 9], where the target tones were perceived as higher when preceded by lower tones, and vice versa. This mechanism therefore serves to aid the perception of coarticulated tones in BM.

However, while such normalization is possible in BM and presumably TM, which have only four lexical tones, normalizing effects as strong would be less probable in languages with larger tone inventories, such as TSM, with the lower recoverability [10] led by the larger number of potential target tones.

Logically, there could be three types of languages pertaining the perception of coarticulated tones. One would have no tonal coarticulation, and therefore does not require any further mechanism. The other two, with their tones coarticulated, would have to develop strategies to deal with such coarticulation. One possibility is to have looser tone boundaries, and for the listeners to accept coarticulated tones as other tones, and then retrieve the speakers' intended tones through normalization. The other is to have stricter boundaries, and reject coarticulated tones as other lexical tones. These languages are referred to as Type III, I and II languages hereafter, as illustrated in Fig. 1. Under this account, BM, and likely TM, would be Type I languages. A complete reliance on normalization to resolve the possible confusion of coarticulated tones is however less likely for TSM, since normalization in a language with a large

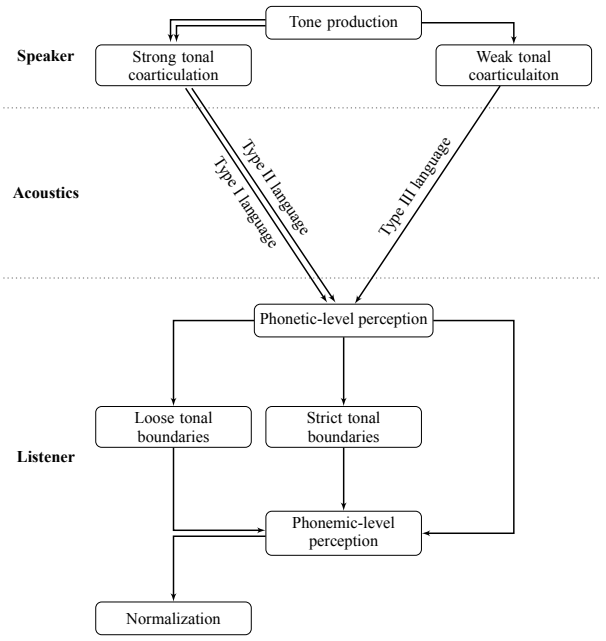


Figure 1: Three kinds of languages with reagrd to tone perception under tonal coarticulation.

tone inventory would be more difficult as mentioned previously. On the other hand, it has also been found that while TSM has more lexical tones than TM, it has identical magnitude of tonal coarticulation as TM [11]. On this account, one may predict that TSM is more likely a Type II language, which has to maintain a stricter tone boundary in order for the coarticulated tones to not be perceived as other tones.

Since previous studies on the perception of coarticulated tones in TM and TSM are sparse and inconsistent, and no direct comparison between the two languages could be found, this study aims to provide a systematic comparison of the two languages. Crucially, this study seeks to find out:

- whether linguistic differences exist between the degrees of normalization for coarticulated tones in TM and TSM.
- if TSM has weaker normalizing effect, how else TSM speakers could cope with the higher probability of perceptual confusion.

2. PERCEPTION OF COARTICULATED TONES IN PREVIOUS STUDIES

2.1. Perception of coarticulated tones in BM

In general, while no studies of TM could be found, past studies generally found coarticulated tones to be perceptible by the listeners in BM. In [4], swapping of ambient tones was found to affect the identification of target tones. Target tones were perceived with higher accuracies in compatible positions than in conflicting ones. Coarticulated tones were therefore hypothesized to be perceived through normalization. A more direct result supporting such stipulation was borne out in [9], where the low-tone-falling-tone boundaries were found to be higher when the ambiguous target tones were preceded by lower F0 offsets, and vice versa.

2.2. Perception of coarticulated tones in TSM

Tonal coarticulation was also found to be perceivable in TSM, and certain normalizing effects seemed to be present. In [1], subjects could predict truncated target tones with its ambient tones above chance rates, and in [6], target tones were found to be perceived as lower when preceded by higher F0 offsets, and vice versa. The audios were however not processed, and therefore the acoustic cues other than F0 might have given hints to the subjects in these two studies.

2.3. Sectional summary

While normalizing effects seemed to be found in both BM and TSM, no previous research of TM was conducted. Crucially, a quantifiable and consistent design would be required to make a direct comparison between the two languages.

3. METHODS

Two experiments were conducted in this study. The first was a replication of [9], where the magnitude of normalization of tonal coarticulation was measured. The second was a word-non-word test, aiming to examine the subjects' strictness on tone boundaries.

3.1. Experiment 1

Experiment 1 measures the normalization effect for tonal coarticulation in TM and TSM. Following the design of [9], a continuum from low-level tone to falling tone preceded by tones of different pitch levels of offsets were synthesized into disyllabic stimuli, and subjects were asked to judge whether they heard a low-level tone or a falling tone. If normalization was at work, different pitch levels of preceding offsets should render different thresholds of low-level tone/falling tone judgement.

3.1.1. Participants

This study recruited 43 Taiwanese college students as participants (25 females; 20–27 y.o., mean=21.93). 15 of them were native speakers of TM. These subjects were not speakers of TSM. 28 of them were native speakers of TSM (and

also TM). Among these 28 participants, 11 were advanced TSM speakers.

Both Experiments 1 and 2 had a Mandarin version and a Southern Min version. The monolingual group participated in the Mandarin versions. The bilingual group participated in both the Southern Min and Mandarin versions.

3.1.2. Stimuli and procedure

3 Pairs of TM disyllabic words and 4 pairs of TSM words were chosen for each of the two versions. These pairs were minimal pairs that differed only in the tones of the second syllables. Within each pair, the first syllables were the same and served as the primes, while the second syllables were the targets, and were a low-level tone and a falling tone, respectively. The pairs were 55+21/55+51, 35+21/35+51, 51+21/51+51 in TM version and 33+21/33+51, 55+21/55+51, 51+21/51+51, 21+21/21+51 in TSM version.

The audios used for syntheses were produced by two native speakers. The F0 values of the primes were then synthesized into five-level tone scales with the equation of [12], with the f_{min} being 100 Hz:

$$scale = \frac{1}{2}(39.86 \times \log(\frac{f_i}{f_{min}})) + 1$$

The F0 values of the targets were divided into 10 F0 levels on Bark's scale from 0.9 to 1.9. The durations of all syllables were 0.3s, and the mean intensities were scaled to be the same. These are illustrated in Fig. 2.

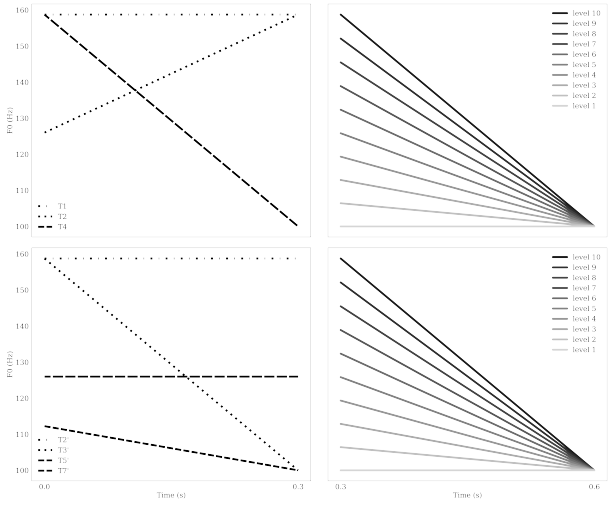


Figure 2: Illustration of synthesized stimuli.

Each stimuli were randomly shuffled, and repeated five times. In each trial, subjects saw an ear picture and were asked to judge whether they heard the low tone or the falling tone.

3.1.3. Analyses

The responses of the subjects were calculated as percentages of falling tone responses, and then fitted through GAMMs. The looseness of the fitted splines was taken as indicator of the magnitudes of perceptual compensation for tonal coarticulation.

3.2. Experiment 2

Experiment 2 was conducted to see whether TM speakers and TSM speakers had different acceptance ranges of falling and low-level tones.

3.2.1. Stimuli

In each version, 10 disyllabic words were used. All their first syllables were the high-level (55) tone. Half of them had as their second syllables the falling (51) tone, and another half the low-level (21) tone. The words were checked to have no minimal pair counterparts; that is, the words in the 55+21 group would be non-words if the second syllables were changed into a falling tone, and vice versa. The target tones were a continuum from low-level to falling tones. The targets and the preceding high-level tones were synthesized as in Experiment 1. For each level, there were 10 repetitions, randomly presented. In each trial, subjects saw the intended words, and were asked to decide whether the tokens were good production of the intended words or not.

3.2.2. Analyses

The subjects' responses were calculated as percentages and taken as acceptance rates of each of the 10 levels. The first derivatives of the acceptance rates were then fitted through GAMMs and examined with simple t-tests.

The maximum of the first derivatives was taken as indicators of the rate of acceptance climb, and the level on which this maximum point falls was determined as the threshold of tone acceptance. Both larger maximum "velocity" (first derivatives) of acceptance climb and a threshold closer to the two ends of the continuum (i.e., a lower threshold level for the low tone, and a higher threshold level for the falling tone) would indicate a stricter tone boundary, and vice versa.

4. RESULTS AND DISCUSSIONS

4.1. Normalizing effects for coarticulated tone in TM and TSM

GAMM results of Experiment 1 showed that the distances between the fitted splines of the monolingual group's TM falling tone responses (the maximum distance being 2.78 level) were significantly wider than those in TSM (the maximum distance being 1.53 level), as in Fig. 3, suggesting TM had stronger normalizing effects than TSM.

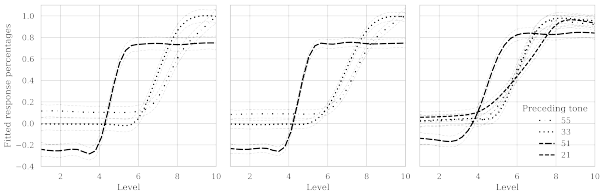


Figure 3: GAMM fitted falling tone response percentages in Experiment 2 (left: Mandarin (monolingual); middle: Mandarin (bilingual); right: Southern Min).

This is in accordance with our prediction. Since TSM has 7 lexical tones, while TM has only 4, retrieving the intended tone by the speaker would be much more difficult in the former. While in TM, a simple one-to-one mapping could be made during normalization, in TSM, multiple possibilities could exist. The lower level of retrievability may lead TSM listeners to rely less on normalization.

4.2. Tone boundaries in TM and TSM

Experiment 2 showed that the falling tone acceptance rates in TSM had substantially larger maximum of first derivatives of the acceptance rates (p=.019*), and marginally significantly higher levels of thresholds (p=.059), as in Fig. 4. This

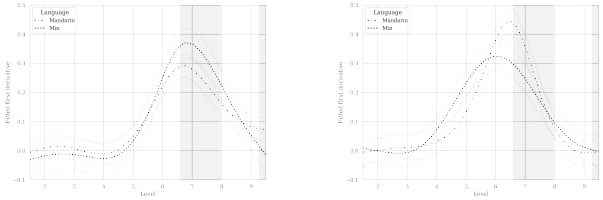


Figure 4: GAMM fitted first derivatives of acceptance rates in Experiment 3 (left: low tone; right: falling tone). The gray areas indicate intervals of significant differences.

pattern was more significant on the advanced bilingual group (p=.005***/.091**, respectively). This pattern, however, was not seen for the low tone, where the maximums of first derivatives of the acceptance rates were higher in TM than in TSM (p=.018*), and the levels of the thresholds were not significantly different (p=.974).

The results suggested that the boundary was stricter in TSM for the falling tone, while for the low tone, this pattern was absent. Nevertheless, such absence of a stricter boundary for the low tone in TSM was likely due to the tone sandhi rules, where the TSM low tone would become the falling tone. Though the stimuli was isolated disyllabic words, and the subjects were reminded that the words were produced in isolation and therefore should not undergo sandhi changes, several subjects reported feeling both the low tone versions and falling tone versions to be acceptable for them. This might explain why the boundaries for the low tone were not significantly stricter for TSM subjects as seen in the falling tone. The results thus indicated stricter tone boundaries in TSM than in TM.

4.3. Perception strategies of coarticulated tones in TM and TSM and the nature of tonal coarticulation

The results in the two experiments indicated that while TM, like BM in past studies, behaved like a Type I language that could have accepted coarticulated tones as other lexical tones, and then retrieved the targets through normalization, TSM did not have as strong normalization, and could have perceived coarticulated tones as the intended tones owing to the stricter tone boundaries.

Our results thus suggest that tonal coarticulation may be a universal constraint that cannot be voluntarily contolled. One could imagine effective perception to be difficult under tonal coarticulation in a language with a large tone inventory, with the higher possibility of confusion and the lower reliability of normalization. As seen, TSM listeners had to keep stricter tone boundaries to attain faithful comprehension. The reason why TSM does not choose to resolve such confusion with less tonal coarticulation may be that it simply is not voluntary and may well exist in all tone languages. This, however, is beyond the scope of this study and will require further typological research.

5. CONCLUSIONS

This study is the first to investigate the perception of coarticulated tones in TM, and to make quantifiable comparison between languages with different sizes of tone inventories. We hope to shed light on the interaction of production and perception, and to provide further clues pertaining the nature of tonal coarticulation.

6. REFERENCES

[1] S.-H. Peng, "Production and perception of Taiwanese tones in different tonal and prosodic contexts," *Journal of Phonetics*, vol. 25, no. 3, pp. 371–400, 1997.

[2] Y. Xu and M. Wang, "Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin," *Journal of Phonetics*, vol. 37, no. 4, pp. 502–520, 2009.

[3] X. S. Shen, "Tonal coarticulation in mandarin," *Journal of Phonetics*, vol. 18, no. 2, pp. 281–295, 1990.

[4] Y. Xu, "Production and perception of coarticulated tones," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2240–2253, 1994.

[5] ——, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[6] H. Wang, "The prosodic effects on Taiwan Min tones," *Language and Linguistics*, vol. 3, pp. 839–852, 2002.

[7] Y.-C. Chang and F.-F. Hsieh, "Tonal coarticulation in Malaysian Hokkien: A typological anomaly?" *The Linguistic Review*, vol. 29, pp. 37–73, 2012.

[8] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley: University of California Press, 1968.

[9] H. Zhang, H. Ding, and W.-S. Lee, "The influence of preceding speech and nonspeech contexts on Mandarin tone identification," *Journal of Phonetics*, vol. 93, p. 101154, 2022.

[10] E. Flemming, *The Grammar of Coarticulation*. Paris: L'Harmattan, 2011.

[11] B.-X. Huang and C. Chiu, "Tonal coarticulation in Taiwan Mandarin and Taiwan Southern min," unpublished.

[12] J. Fon and W.-Y. Chiang, "What does Chao have to say about tones?—A case study of Taiwan Mandarin," *Journal of Chinese Linguistics*, vol. 27, no. 1, pp. 13–37, 1999.

PERCEPTION OF VOICING DISTINCTION IN SYLLABLE-INITIAL STOPS BY MULTILINGUAL SPEAKERS

Zuzanna Cal and Magdalena Wrembel

Adam Mickiewicz University, Poznań, Poland
zuzanna.cal@amu.edu.pl, magdala@amu.edu.pl

ABSTRACT

The objective of the current study is to explore how multilingual learners categorise fortis and lenis stops in the three languages they know. In particular, we aim to investigate whether patterns of categorisation of VOT continua are specific to a language and place of articulation (PoA), as well as whether trilinguals experience boundary shifts in any of the languages and PoA. The study was conducted on L1 Polish-L2 English-L3 Norwegian speakers (n=22, aged 20), who had just started studying Norwegian as part of intensive formal training. The experiment included preparation of three continua (/b-p/, /d-t/, /g-k/) in Polish, English and Norwegian and was run in PsychoPy. The participants listened to the words from the continua and decided whether they hear a voiced or voiceless sound at the beginning of each word. The results point to rather unique patterns of VOT categorisation across languages and PoA with possible L1-L2 and L1-L3 interference.

Keywords: speech perception, L3 phonology, multilingualism, VOT, third language acquisition

1. INTRODUCTION

Research on multilingual perception of speech has gained increasing interest in the recent years. Extensive investigations [1, 2, 3, 4, 5] have pointed to a potential multilingual advantage according to which prior linguistic experience has facilitative effects on the perception of novel L3 contrasts. What is more, recent research also pointed out to rather specific trends. For example, [6] found combined effect of L1, markedness and L2/L3 proficiency in multilingual speech perception of rhotics by children. Furthermore, in the study on perception of sibilant pairs, [7] showed that multilinguals assimilate some L3 sounds to both L1 and L2 categories, with a preference for the latter and that beginner L3 learners are likely to perceive subtle acoustic differences in novel phonological contrasts.

As far as VOT perception is concerned, a considerable amount of literature has been published over the last few decades [8, 9, 10, 11, 12, 13, 14]. In these studies, which assume, however, the perspective of second language acquisition, VOT has been shown to be susceptible to the influence of a number of factors, including the age of onset, L1/L2 status or individual differences. Recent research attested different patterns of perception while comparing monolinguals and bilinguals. For example, [13] found that monolingual speakers showed a consistent pattern while in bilinguals, the two languages were perceived in significantly different ways.

Although VOT perception in L2 acquisition has been widely researched, to the best of our knowledge, only a few studies have attempted to explore VOT perception from the multilingual perspective. [15] investigated the effect of language learning on the perception of VOT in L2/L3 and L1. The study was designed to tap into the influence of a newly acquired language on the previously known ones. The task involved identification of a /bi-pi/ continuum by two groups (bi- and trilingual) of language learners that differed in learning settings and experience. The results pointed to both regressive and progressive cross-linguistic influence. Another study that tackled VOT perception in multilingual learners [16] aimed to investigate the perception and production of word-initial stops by L1 Mandarin Chinese L2 English speakers learning Japanese or Russian as L3. The participants performed an identification task of /p, t, k, b, d, g/ in word-initial position of monosyllable or disyllabic words. The results pointed to the role of phonetic similarity in different stop categories between L1, L2, and L3 that might have contributed to learners' confusion in perception. What is more, higher accuracy scores were found in the perception of voiced stops than in voiceless ones in L3 Japanese/Russian. As voiced stops in the participants L3s differ from those in their L1, they might have formed a new category, and therefore, be more easily perceived by the learners, which is in line with Flege's Speech Learning Model [17].

All in all, research on the subject, though quite prominent, has been mostly restricted to limited comparisons of languages, places of articulation and types of speakers. This paper attempts to expand the existing body of literature by examining perception of word-initial stops by trilingual speakers of L1 Polish L2 English and L3 Norwegian. These three languages seem to be complementary with respect to the features of their stop consonants. Firstly, Polish is a

true voicing language that has prevoicing in the voiced series of stops and short-lag VOT in voiceless stops [8]. English, on the other hand, is an aspirating language in which /b, d, g/ are partially voiced and /p, t, k/ are aspirated [18]. When it comes to Norwegian, the voiced stops are prevoiced, similarly to Polish ones, but its voiceless stops are aspirated and resemble those in English [19, 20].

2. METHOD

2.1. Aims and predictions

The aim of this study is to investigate how L1 Polish - L2 English - L3 Norwegian speakers categorise fortis and lenis stops in the three languages. The following research questions were asked in the study: (1) What are the patterns of VOT categorisation in multilinguals? Are they language- and PoA-specific? (2) What are the perceptual boundary locations for the perception of voiced and voiceless stops in all three languages? Do they point to potential sources of CLI?

There are two predictions regarding the outcomes of the study. In relation to the first RQ, it is hypothesised that multilingual advantage might trigger more language- and PoA- specific patterns of VOT categorisation. The second RQ considers perceptual boundary locations. We foresee two possible scenarios depending on the language status and/or linguistic similarities. First of all, due to existing similarities in the process of learning foreign languages [21], as well as intrinsic phonetic characteristics, i.e., aspiration being present in the voiceless series of stops in L2 English and L3 Norwegian, we may hypothesise that the participants experience boundary shifts towards the English values. In other words, the boundaries in L3 Norwegian might be found to be later in the continuum and yield similar values to those of L2 English, particularly for the fortis series. On the other hand, at the initial stages of L3 acquisition, and due to the existing similarities between Polish and Norwegian voiced stops, the participants might also rely heavily on their L1. In this sense, we hypothesise that the perceptual boundaries in L3 Norwegian might be shifted towards those of L1 Polish and appear earlier in the continua.

2.2. Participants

The participants included 19 L1 Polish speakers, who had been learning English as their L2 and had recently started learning L3 Norwegian extensively by attending Norwegian Studies programme at two Polish universities. There were 14 females and 5 males, whose mean age was 20 years old. Prior to the experiment the participants had had limited, almost none, contact with Norwegian. Their mean age of onset in English and Norwegian was 6.3 and 19.6 years respectively.

They were asked to fill in the Language History Questionnaire (LHQ) [22] in order to obtain information about their language learning history and use, including the age of onset, declared exposure to languages, length and type of instruction etc. The participants also completed proficiency placement tests in both of their foreign languages – English and Norwegian. Based on the grading criteria they were classified as intermediate (approximately B2 level) learners of L2 English and beginner (A1) learners of L3 Norwegian. Knowledge of any other foreign languages has been reported as minute and the participants claimed not to have used any of them for an extensive period of time.

2.3. Procedure and stimuli

The study design involved the preparation of VOT continua with the use of a Praat script [23] based on one-syllable minimal pairs with word initial stop sounds /p,b,t,d,k,g/. The token words included *pas-bas*, *tam-dam*, *kas-gaz* for Polish, *pen-ben*, *tan-Dan*, *cast-gas* for English and *par-bar*, *ta-da*, *kar-gard* for Norwegian and were recorded in carrier phrases by native speakers of the respective languages. Separate continua were created for each place of articulation and for each language. Each token from the continua differed from the next one by 10 ms. Table 1 shows specific values and the number of steps in each of the continua. The ranges were based on the values obtained from native speakers' recordings in all three languages.

	Polish	English	Norwegian
b-p	-90 - 30 ms (13 steps)	0 - 70 ms (8 steps)	-140 - 80 ms (23 steps)
d-t	-130 - 20 ms (16 steps)	0 - 90 ms (10 steps)	-130 - 90 ms (23 steps)
g-k	-80 - 60 ms (15 steps)	0 - 70 ms (8 steps)	-140 - 90 ms (24 steps)

Table 1: Ranges and number of steps of VOT continua (in ms) for the three languages.

The experiment was programmed in PsychoPy [24] and conducted in three separate language blocks on different days. In each session, the participants were presented with audio stimuli and a visual representation of two sounds respective to the continuum from which the sound was played. The task was to identify which of the presented sounds appeared at the beginning of each word by clicking an appropriate key on a computer keyboard. Every token was presented 3 times. The experiment was

conducted in a quiet room and the participants were instructed to wear headphones. Each language block lasted around 7-10 minutes.

3. RESULTS

The obtained data included accuracy rates and reaction times (RT). First of all, the Pearson correlation was run to compare accuracy scores (%) for each step of the continuum with RT (ms). The statistical analysis was run in SPSS [25]. Moderate and significant negative correlation coefficients were obtained for all three continua in L1 Polish (/b-p/: r=-0.44, p=.035; /d-t/: r=-0.54, p=.008, /g-k/: r=-0.55, p=.005), see Table 2. Strong and significant negative correlation was found in the /b-p/ continuum in L2 English (r=-0.81, p=.016) and the /d-t/ continuum in L3 Norwegian (r=-0.72, p=.002). Table 2 presents correlation coefficients for all continua. All r-values turned out to be negative, except for the /g-k/ continuum in English, which yielded quite weak and positive correlation. All in all, the values show that the lower the accuracy scores, the higher RT, which means that the participants took longer to react when they were less confident about the answer.

Language	Continuum	*r*	*p*
Polish	b-p	-0.44	.035
	d-t	-0.54	.008
	g-k	-0.55	.005
English	b-p	-0.81	.016
	d-t	-0.38	.285
	g-k	0.23	.578
Norwegian	b-p	-0.54	.056
	d-t	-0.72	.002
	g-k	-0.47	.080

Table 2: The Pearson correlation coefficients for each continuum and language.

Figures 1-3 show accuracy rates for the identification of voiced plosive sounds in all three continua in the three languages. What can be observed in these charts is that there are some discrepancies across L1/L2/L3 languages, especially visible in the /b-p/ continuum, but also slightly in the /d-t/. However, in the /g-k/ continuum, the participants were more consistent in their answers across the languages.

To obtain the perceptual boundary location, the data was converted using a logistic regression and the boundary location was calculated with the use of the following formula: *-LN(b0)/LN(b1)*, where *b0* corresponded to the constant and *b1* to the obtained slope of the function. Table 3 provides mean perceptual boundaries and standard deviations for each continuum and language.

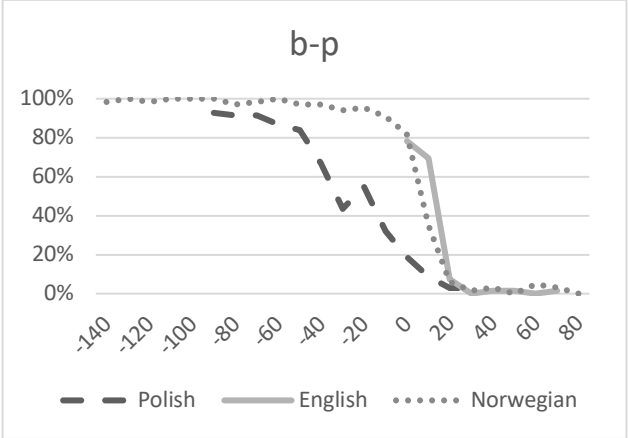


Figure 1: Accuracy rates for the identification of /b/ in L1 Polish, L2 English and L3 Norwegian.

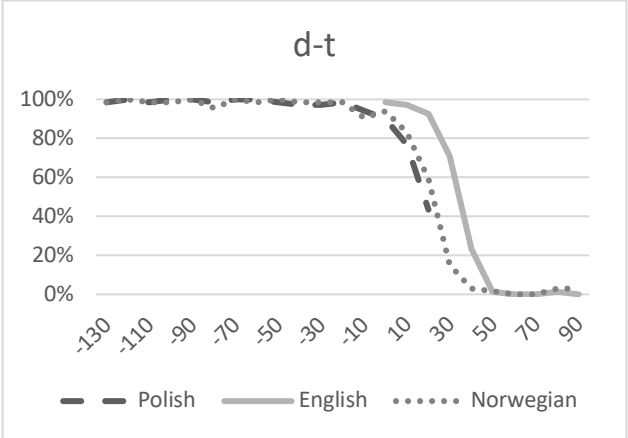


Figure 2: Accuracy rates for the identification of /d/ in L1 Polish, L2 English and L3 Norwegian.

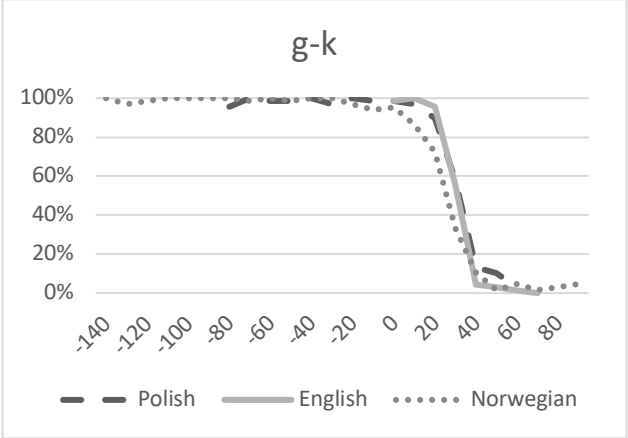


Figure 3: Accuracy rates for the identification of /g/ in L1 Polish, L2 English and L3 Norwegian.

The mean perceptual boundary location for the /b-p/ continuum was the earliest in Polish (-25.42 ms), followed by Norwegian (-2.13 ms) and English (14.65 ms). In /d-t/ continuum, the mean perceptual boundary turned out to be the lowest in Norwegian (10.11 ms), then Polish (14.05 ms) and, finally, English (35.97 ms). As far as /g-k/ continuum is concerned, the earliest perceptual boundary was

found in Norwegian (18.77 ms), followed by Polish (29.87 ms) and English (30.82 ms), which yielded similar values.

Language	Continuum	Mean	Std. Deviation
Polish	b-p	-25.42	18.89
	d-t	14.05	15.82
	g-k	29.87	14.00
English	b-p	14.65	8.05
	d-t	35.97	6.52
	g-k	30.82	3.93
Norwegian	b-p	-2.13	8.20
	d-t	10.11	9.01
	g-k	18.77	11.54

Table 3: Mean perceptual boundary (in ms) in each continuum and language.

A Linear Mixed Model was run to investigate the differences in perceptual boundary locations with language (Polish, English, Norwegian), place of articulation (PoA; labial, coronal, velar), as well as interaction between language and PoA as fixed factors, and participant as a random effect. The results showed significant main effects of Language (F=43.878, p<.001), PoA (F=108.036, p<.001) and their interaction (F=18.822, p<.001).

Pair-wise comparisons between Languages and PoA pointed to the following results. In the /b-p/ continuum, statistically significant differences were found between all three languages, that is between L1 Polish and L2 English (t=-9.279, p<.001), L1 Polish and L3 Norwegian (t=-6.653, p<.001), as well as between L2 English and L3 Norwegian (t=3.800, p<.001). With regards to the /d-t/ continuum, statistically significant differences were shown between L1 Polish and L2 English (t=-2.954, p=.004), as well as L2 English and L3 Norwegian (t=7.613, p<.001). As far as the /g-k/ continuum is concerned, statistically significant differences were found between L1 Polish and L3 Norwegian (t=3.080, p=.003), but also between L2 English and L3 Norwegian (t=3.546, p=.001).

4. DISCUSSION AND CONCLUSION

The aim of this paper was to investigate VOT perception by multilingual learners. We intended to address research questions as to whether patterns of categorisation of VOT continua are specific to a language and PoA, as well as whether trilinguals experience boundary shifts in any of their languages and PoA.

The first prediction set out in this study was that multilingual advantage would trigger more language- and PoA-specific patterns of VOT categorisation. This prediction was, to a great extent, confirmed by our data. Statistical analysis pointed to significant differences in most of the cases. That is, perceptual boundaries in /b-p/ continuum turned out to be statistically different in all three languages. In /d-t/ continuum, the differences appeared between L1 and L2, as well as L2 and L3. Finally, in /g-k/ continuum significant differences were observed between L1 and L3, as well as L2 and L3. These obtained differences suggest that multilingual learners generally maintain language- and PoA- specific patterns of VOT categorisation. This finding might be an indication of a multilingual advantage, according to which, multilingual learners are more prone to perceive subtle linguistic contrasts [1, 2, 3, 4, 5].

The second prediction dealt with perceptual boundary locations. Two possible scenarios were hypothesised depending on the language status (L1/L2/L3) and inherent phonetic similarities. Firstly, it was predicted that the boundaries of L2 English and L3 Norwegian would be closer due to the existing similarities between voiceless stops, but also due to the process of foreign language learning [21]. On the other hand, it was predicted that the participants might also rely heavily on their L1 while learning an additional language [26], which, in turn, would contribute to L1 and L3 boundaries being closer together. We found confirmation of the second scenario only in /d-t/ continuum, where no statistically significant difference was found between L1 Polish and L3 Norwegian. This suggests possible interdependence between the two languages. What was also found is no statistically significant difference between L1 Polish and L2 English in /g-k/ continuum, which suggests possible, unexpected, interactions between the L1 and L2. No other traces of CLI were attested in the data. Again, pointing to the possible role of multilingual advantage and confirming the findings of [13].

Recapitulating, this paper explored categorisation of VOT continuum by multilingual learners of L1 Polish, L2 English, L3 Norwegian. It was found that the learners maintain patterns of VOT categorisation specific to a language and place of articulation. Possible, though limited interactions were attested between L1-L2 and L1-L3, in one of the three places of articulation, which suggests plausible, though rather weak L1 interference in the perception of VOT in both the L2 and L3. Ongoing comparisons with a longitudinal data collection design and a control group performance will allow for a more in-depth analyses and further conclusions.

5. ACKNOWLEDGEMENT

This research is supported by a project grant of the Polish National Science Centre (NCN) OPUS-19-HS (UMO-2020/37/B/HS2/00617) CLIMAD "Cross-linguistic influence in multilingualism across domains: Phonology and syntax".

6. REFERENCES

[1] Enomoto, K. 1994. L2 perceptual acquisition: The effect of multilingual linguistic experience on the perception of a "less novel" contrast. Edinburgh *Working Papers in Applied Linguistics* 5, 15–29.

[2] Tremblay, M.-C., Sabourin, L. 2012. Comparing behavioral discrimination and learning abilities in monolinguals, bilinguals and multilinguals. *J. Acoust. Soc. Am*. 132, 3465–3474.

[3] Antoniou, M., Liang, E., Ettlinger, M., Wong, P. 2015. The bilingual advantage in phonetic learning. *Bilingualism: Language and Cognition* 18, 683–695.

[4] Kopečková, R. 2015. Differences in the perception of English vowel sounds by child L2 and L3 learners. In: U. Gut, R. Fuchs, E.-M. Wunder (eds.), *Universal or diverse paths to English phonology*. Mouton de Gruyter, 71–90.

[5] Onishi, H. 2016. The effect of L2 experience on L3 perception. *Int. J. Multilingual*. 13, 459–475.

[6] Balas, A., Kopečková, R., Wrembel, M., & Warren, P. (2019). Perception of rhotics by Multilingual Children. *Proc. 19th ICPhS* Melbourne, 3725–3729.

[7] Wrembel, M., Marecka, M., & Kopečková, R. 2019. Extending perceptual assimilation model to L3 phonological acquisition. *Int. J. Multilingual.* 16, 513–533.

[8] Keating, P., Mikoś, M., Ganong III, W. 1981. A cross-language study of range of VOT in the perception of stop consonant voicing, *J. Acoust. Soc. Am*. 70, 1260–1271.

[9] Flege, J. E. 1987. The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *J. Phonetics* 15, 47–65.

[10] Rojczyk, A. 2010. *Temporal and spectral parameters in perception of the voicing contrast in English and Polish*. Wydawnictwo Uniwersytetu Śląskiego.

[11] Stölten, K. 2013. "The Effects of Age of Onset on VOT in L2 Acquisition and L1 Attrition: A Study of the Speech Production and Perception of Advanced Spanish-Swedish Bilinguals," Ph.D. dissertation, Stockholm University.

[12] Kong, E. 2019. Individual differences in categorical perception: L1 English learners' L2 perception of Korean stops. *Phonetics and Speech Sciences* 11, 63–70.

[13] García-Sierra, A., Schifano, E., Duncan, G. M., Fish, M. S. 2021. An analysis of the perception of stop consonants in bilinguals and monolinguals in different phonetic contexts: A range-based language cueing approach. *Attention, perception & psychophysics* 83, 1878–1896.

[14] Pan, L., Ke, H., Styles, S. J. 2022. Early linguistic experience shapes bilingual adults' hearing for phonemes in both languages. *Scientific Reports* 12.

[15] Liu, Z., Gorba, C., Cebrian, J. 2019. Effects of learning an additional language on VOT perception. *Proc. 19th ICPhS Melbourne*, 260–254.

[16] Liu, J., & Lin, J. 2021. A cross-linguistic study of L3 phonological acquisition of stop contrasts. *SAGE Open* 11.

[17] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In: Strange, W. (ed), *Speech perception and linguistic experience: Issues in cross-language research*. York Press, 233–277.

[18] Lisker, L., Abramson A. 1964. A cross-language study of voicing in initial stops, *Word* 20, 384–422.

[19] Ringen, C., Dommelen, W.A. 2013. Quantity and laryngeal contrasts in Norwegian. *J. Phonetics* 41, 479–490.

[20] Czarnecki, P. 2016. *The Phonology of Quantity in Icelandic and Norwegian*. Wydział Neofilologii UAM w Poznaniu.

[21] Bardel, C., & Falk, Y. 2007. The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research* 23, 459–484.

[22] Li, P., Zhang, F., Yu, A., Zhao, X. 2020. Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition* 23, 938–944.

[23] Winn, M. B. 2020. Praat script.

[24] Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, Sogo, H., Kastman, E., Lindeløv, J.K. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51, 195–203.

[25] IBM Corp. Released 2021. IBM SPSS Statistics for Macintosh, Version 28.0. Armonk, NY: IBM Corp.

[26] Hermas, A. 2010. Language acquisition as computational resetting: verb movement in L3 initial state. *Int. J. Multilingual* 7, 343–362.

THE ROLE OF RHYTHMIC STRESS IN THE PERCEPTION OF WORDS

Anna Łukaszewicz[1], Beata Łukaszewicz[1], Janina Mołczanow[1]

[1]University of Warsaw
a.b.lukaszewicz@uw.edu.pl, b.lukaszewicz@uw.edu.pl, jmolczanow@uw.edu.pl

ABSTRACT

Polish has a grammatical stress system with two degrees of stress. Apart from the main stress falling regularly on the penult, it has iterative rhythmic stresses applying from the left word edge. Previous studies show that rhythmic stresses are optional and acoustically weak. The functionality of rhythmic stress is unclear. In this paper, we test the hypothesis that rhythmic stress facilitates word segmentation: despite being phonologically redundant, it may have a differential effect on making lexical decisions. We report a perceptual experiment, in which we investigated whether rhythmically congruent and incongruent lexical items vs. segmentally matched nonce words would trigger different response times (RTs) in a word/non-word decision task. The results point to a statistically significant interaction between wordhood and rhythm. Faster RTs were obtained for rhythmically congruent lexical items compared to rhythmically incongruent ones. For nonwords, the pattern was reversed. Rhythmic structure is thus associated with word-likeness.

Keywords: perception, Polish, rhythm, RTs, stress

1. INTRODUCTION

This paper investigates the relationship between phonologically redundant rhythmic (secondary) stress and lexical processing, on the basis of behavioural data (response times; RTs) obtained in a perception experiment involving a word/non-word decision task. While it is clear that prosody plays an important role in speech recognition, facilitating utterance parsing and word comprehension (e.g. see [1] and the literature cited therein), it is not obvious whether the optional alternation of rhythmic beats and metrically weak syllables, as the one occurring in iterative secondary stress systems, plays a similar role. We focus here on rhythmic stress in Polish (the so-called Warsaw and Cracow standard varieties), which is a well-known example of an iterative stress system in the phonological literature (e.g. [2], [3], [4], [5], [6]). Rhythmic stress in Polish is fully predictable and hence phonologically redundant. It is also known to be optional, which prompts questions concerning its functionality.

Little is known about the perception of rhythmic stress in Polish. To our knowledge, the only perception experiment conducted thus far was the one reported in [7]; it revealed no sensitivity of native speakers of Polish to the presence of rhythmic stress. On the one hand, that negative result could be associated with rhythmic stress being acoustically weak. (The lack of acoustic robustness of subsidiary degrees of stress has often been raised as an issue cross-linguistically, putting into question the physical reality of the phenomenon in general [8].) On the other hand, the method used in that experiment was based on the participants' transcription of stress, and thus it relied heavily on introspection. Here we report an experiment based on a standard behavioural (RT) paradigm, which circumvents those methodological issues. The hypothesis that we put forward is that, though acoustically weak and phonologically redundant, rhythmic stress affects word segmentation and lexical processing, facilitating word recognition.

The Polish rhythmic stress pattern is illustrated below. It is a well-documented phenomenon. Basic generalizations are found, among others, in an early kymographic study by Dłuska, published in 1932 ([2]), traditional descriptions of the Polish prosody (such as [9]), phonological analyses couched in various theoretical frameworks (the metrical grid in [3]; optimality theory in [4]), as well as in recent studies of acoustic correlates of rhythmic stress ([5], [6], [10], [11]). As can be seen in the examples adduced below, Polish has primary stress on the penult and an iteration of secondary (rhythmic) beats on every odd-numbered syllable. Important from the point of view of the current study is that the iterative pattern, which begins at the left edge of the word, is [ˌσσ] (stressed – unstressed), and not (σˌσ) (unstressed – stressed). As clarified below, this observation is the basis for the rhythmically congruent and rhythmically incongruent stimuli in the experiment we report here.

- *poˈmidor* ('tomato' N., nom. sg.)
- *ˌpomiˈdora* ('tomato' N., gen. sg.)
- *ˌpomidoˈrowy* ('tomato' Adj., nom. sg.)
- *ˌpomiˌdoroˈwego* ('tomato' Adj., gen. sg.)

In what follows, we describe the details of a rhythmic stress perception experiment involving RTs, which we conducted for the purpose of the present

study. We test whether native speakers of Polish are sensitive to a change in the temporal structure across the first two syllables in a word, resulting in prominence reversal, [ˌσσ] (rhythmically congruent stimuli) > [σˌσ] (rhythmically incongruent stimuli), and whether this has a differential effect on the word recognition process.

2. METHOD

Overall, the design of our RT experiment was based on a classic stress perception study [1], the main difference being that the semantically congruent vs. incongruent condition in [1] is replaced by the word vs. nonword condition in our experiment. Analogously to [1], we aimed to determine whether the metrical and lexical aspects of spoken language are processed independently or in interaction. Those aspects were thus manipulated orthogonally, so as to create four conditions. Words and nonwords appeared in the rhythmically congruent vs. rhythmically incongruent versions, forming segmentally matched quadruplets, described in more detail in 2.1 below.

2.1. Stimuli

Overall, the stimuli consisted of 128 tokens (32 lexical items and 32 nonce words, both in the rhythmically congruent and rhythmically incongruent versions. The stimuli can thus be described as having the following structure:

- W+R+, a rhythmically congruent word
- W+R– , a rhythmically incongruent word
- W-R+, a rhythmically congruent nonword
- W–R–, a rhythmically incongruent nonword

2.1.1. Lexical vs. nonce stimuli

All target stimuli consisted of five syllables: the first three syllables constituted the (lexical or nonce) 'stem'; the remaining two included an inflectional ending. To ensure maximal segmental and prosodic comparability between lexical and nonce items, in both the last two syllables contained the same inflectional ending *-ami* (instr. pl.). The ending contains the penultimate syllable of the word, which bears primary stress.

The lexical items and the nonce words were segmentally matched: the nonce word differed from its lexical counterpart only in the consonant occurring at the beginning of the third syllable; cf. *czekoladkami* [t͡ʂɛkɔlatkami] 'chocolates' (instr. pl.) – *czekoszadkami* [t͡ʂɛkɔʂatkami] (nonce word). Another such pair (out of the total of 32 paired words and nonwords) is illustrated in Table 1. It is well-known that phonotactic probability influences on-line processes and has an impact on how quickly words are recognised on the basis of fluent speech (e.g. [12], [13], [14]). The lexical items were thus frequency matched on the basis of the balanced subsections of the National Corpus of Polish (Narodowy Korpus Języka Polskiego). The corresponding nonce words were created in such a way as to avoid an overlap between their stems and existing lexical stems. (Because Polish is a highly inflectional language, for all items lemmatic searches were conducted, i.e. all possible grammatical forms were taken into account.)

Word	Nonce
samochod+ami 'car' (instr.pl) /samɔxɔdami/	samoziod+ami /samɔʑɔdami/

Table 1: An example of a lexical item - nonce word pair in the stimuli set.

2.1.2. Rhythmically congruent/incongruent stimuli

All word and nonword stimuli were prepared in two versions: rhythmically congruent and rhythmically incongruent. The rhythmically congruent version was the one agreeing with the secondary stress pattern of Polish, i.e. with the initial syllable having enhanced duration relative to the second syllable. For the purpose of obtaining such stimuli, target words and nonwords were recorded by a phonetically trained native speaker of the standard variety of Polish (the so-called Warsaw Polish). This variety of Polish is known to exhibit the required pattern (e.g. [5], [6]). The target stimuli were incorporated in a frame, which was the same for all items. (*Powiedziała __ po raz pierwszy,* 'She said __ for the first time'.) The recordings were conducted in a sound-treated booth.

The rhythmically incongruent counterparts were obtained by changing the relative duration of the first and second syllables. Care was taken to change the lengths within the ecological bounds. To this end, the original recordings were time-compressed by the 0.7 factor to obtain the initial syllable and time-stretched by the 1.4 factor to obtain the second syllable, using the 'overlap-add' algorithm in *Praat* (v. 6.2.09; [15]). The relevant syllables were then extracted from the continuous acoustic signal using a high-resolution waveform editor (Sound Forge). This allowed to align the boundaries of vowels with zero crossings, so that each excised segment contained a complete number of cycles (e.g. [5], [6], [16], [17]). (This was necessary to offset a potential auditory effect of the resulting speech signal being artificially distorted.) The manipulated syllables were then spliced to replace the two initial syllables in the original sound

files, using a *Praat* script. (We did not use a single manipulated version as the lexical-nonce pairs differed in the third syllable's onset, which has an effect on the second syllable's vowel formants.) All items were placed within the same frame – a single token of the frame was selected, such that the formant structure of the final vowel of *Powiedziała* was least affected by the following sound. Fig. 1 shows an example of the original version, a rhythmically congruent pattern, juxtaposed against the manipulated version, a rhythmically incongruent pattern of the word *balonikami* /balɔɲiˈkami/ 'balloon' (Dimin. instr. pl.) The first syllable (*ba*) is longer, and the second syllable (*lo*) is shorter in the rhythmically congruent item (top panel); the pattern is reversed in the rhythmically incongruent item (bottom panel).

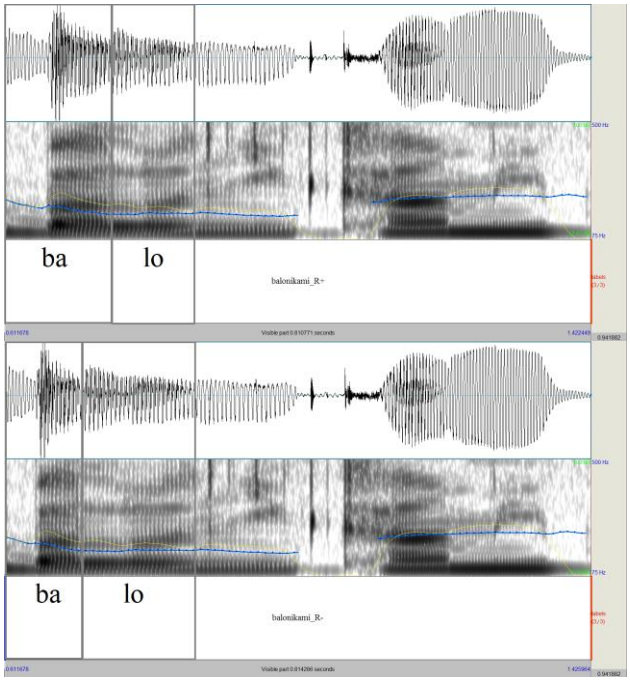


Figure 1: The rhythmically congruent (original) version (top panel) and the rhythmically incongruent version (bottom panel) of the same word.

2.2. Participants

24 (16F) participants took part in the pilot study that we report here. All of them were native speakers of Polish, living in Warsaw, with no reported history of speech or hearing impairment. The mean age was 25.6 (age range 19-41). Most speakers were right-handed (20 out of 24). As pointed out e.g. in [5] and [6], apart from being optional and weakly expressed in acoustic terms, rhythmic stress may also be subject to regional variation. It was thus necessary to take the dialectal factor into account and to base the experiment on the dialects known to exhibit the rhythmic pattern (the so-called Warsaw or Cracow Polish). As required for research involving human subjects, the protocol of the experiment was approved by the Committee for the Ethics of Research Involving Human Participants at University of Warsaw (131/2022).

2.3. Technical apparatus

The experiment was programmed and presented using *E-Prime 3.0*. RTs were measured with the Chronos device. The stimuli were recorded in a sound-treated booth at the sampling frequency of 44.1 kHz, 32-bit depth, using an AT897 microphone. They were manipulated using *Praat* (v. 6.2.09; [15]) and Sound Forge (a high-resolution waveform editor). The participants were wearing closed headphones (ATH-M50x) throughout the experiment. The statistics were conducted in *SPSS* (v. 28).

2.4. Procedure

The participants' task was to press, as quickly as possible, the extreme right or extreme left button on the Chronos panel, depending on whether the stimulus they heard was a word or a nonword. For half of the participants, the extreme right and left buttons corresponded to the choice "word" and "nonword", respectively. For the other half, the buttons were reversed. The experiment was preceded by a short training session, during which the participants could practice pressing the buttons in response to a set of stimuli designed analogously to those appearing later in the experiment. As in [1], each participant heard only one item from a given word – nonword – rhythmic – nonrhythmic quadruplet during the experiment. In this way, we avoided a potential bias created by the participant's being repeatedly exposed to the same initial sequences. The stimuli were balanced across the sessions and were randomized within each session, using the "random order" function in *E-Prime*; all stimuli were equally represented in the experiment.

3. RESULTS

RTs were measured from the onset of the auditory stimulus. The target item began at 0.61 s (which was the fixed length of the initial part of the frame). Only accurate responses were submitted to the statistical analysis. The accuracy was 95.7 %. The statistics were obtained through a fitted generalized mixed effects model with the Gamma probability distribution and the log link function. The model included 'participant' as the random effect. The fixed effects were: wordhood (word/nonword), rhythmicity (congruent/incongruent), the interaction term

rhythmicity*wordhood, and the trial number. The dependent variable was RT. We also included the 'word length' as a covariate. Most factors were significant (we report the F-statistics for brevity); wordhood: $F(1, 698) = 15.8$, $p < 0.001$; rhythmicity*wordhood: $F(1, 698) = 26.5$, $p < 0.001$; trial: $F(31, 698) = 3.9$, $p < 0.001$; word length $F(1, 698) = 4.2$, $p = 0.041$). The only effect that turned out insignificant was 'rhythm'. The estimated RT means (in [ms]) depending on rhythmic congruence and wordhood are graphed in Fig. 2. The interaction between the two factors is clear: rhythmicity has the opposite effect for words and nonwords.

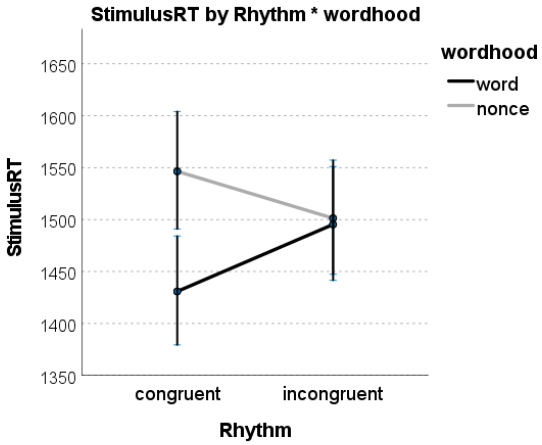


Figure 2: Estimated marginal means for RTs depending on rhythm and wordhood.

4. DISCUSSION

Our results corroborate the hypothesis that rhythmic stress in Polish, though phonologically redundant and acoustically weak, plays an essential role in lexical access and word recognition. The rhythmic structure has a crucial effect on the word-likeness. Rhythmically congruent words are recognised faster than rhythmically incongruent words. The word-likeness judgements of nonwords also depend on the rhythmic factor: rhythmically incongruent nonwords trigger faster decisions. In sum, wordhood and rhythmicity seem to go hand in hand: the recognition of a speech chunk as a word is facilitated if it is well-formed from the point of view of the rhythmic pattern.

This result is important because it is based on rhythmic structure which is entirely redundant from a phonological point of view. Interestingly, while wordhood and the interaction between rhythmicity and wordhood turned out to be significant factors in our analysis, rhythmicity appeared non-significant as the main effect. However, it is by no means a negligible factor in language processing. Our results concerning the role of rhythmic structure corroborate its influence on on-line processing and word recognition, parallel to the one exhibited by sublexical segmental structures (phonotactic probability; e.g. [13]).

Finally, it is important to add that the rhythmic stress in Polish has sometimes been reported to be difficult to detect in the acoustic speech signal ([18]; see also [6] and the literature cited therein), which likely depended on the dialect choice and methodological questions (as discussed thoroughly in [6]). Nevertheless, its being phonologically optional and acoustically weak poses an intriguing question about whether it is perceptible and what its functionality can be. In some previous work (e.g. [3]), the presence of rhythmic stress was associated with slow careful speech. A recent experimental study (11) has demonstrated no such tendency in hyperarticulated speech produced at slower rates (Lombard speech). This would suggest that the presence of rhythmic stress is more important from the point of view of perception than production. The only previous perception study that we know of (i.e. [7]) pointed to the lack of awareness of the presence of rhythmic stress in native speakers of Polish, based on introspection and explicit judgement. Our behavioural experiment, based on a standard RT paradigm, points to the association between rhythmic stress and word-likeness. Native speakers of Polish may not be aware of the rhythmic stress pattern, but its presence has an effect on word segmentation and recognition.

5. CONTRIBUTIONS

Initial conceptualization: AŁ, BŁ, JM; experiment design and preparation of stimuli: AŁ, BŁ; E-prime programming: AŁ; participant recruitment: JM; conducting the experiment: AŁ, BŁ, JM; statistical analyses: BŁ; writing the paper: AŁ, BŁ, JM; revisions: AŁ, BŁ.

6. ACKNOWLEDGEMENTS

We would like to thank two anonymous reviewers for their questions and comments. Research was funded by the National Science Centre (Poland), grant 2017/27/B/HS2/00780.

7. COMPETING INTERESTS

The authors declare no conflict of interest.

8. REFERENCES

[1] Magne, C., Astésano, C., Aramaki, M., Ystad, S., Kronland-Martinet, R., Besson, M. 2007. Influence of

Syllabic Lengthening on Semantic Processing in Spoken French: Behavioral and Electrophysiological Evidence. *Cerebral Cortex* 17, 2659-2668.
[2] Dłuska, M. 1932. *Rytm spółgłoskowy polskich grup akcentowych.* Polska Akademia Umiejętności.
[3] Rubach, J., Booij, G. E. 1985. A grid theory of stress in Polish. *Lingua* 66, 281–319.
[4] Kraska-Szlenk, I. 2003. *The phonology of stress in Polish*. Lincom.
[5] Łukaszewicz, B. 2015. Polish stress revisited: Phonetic evidence of an iterative system. Paper presented at 23rd Manchester Phonology Meeting.
[6] Łukaszewicz, B. 2018. Phonetic evidence for an iterative stress system: The issue of consonantal rhythm. *Phonology* 35, 115–150.
[7] Steffen-Batogowa, M. 2000. *Struktura akcentowa języka polskiego*. Wydawnictwo Naukowe PWN.
[8] Tabain, M., Fletcher, J., Butcher, A. 2014. Lexical stress in Pitjantjatjara. *Journal of Phonetics* 42, 52–66.
[9] Dłuska, M. 1974. *Prozodia języka polskiego.* Państwowe Wydawnictwo Naukowe.
[10] Łukaszewicz, B., Zajbt, E., Krawczyk, U. 2018. The rhythm of heptasyllabic words: Evidence for metrical bidirectionality. *Speech Prosody* 9, 676‑679.
[11] Łukaszewicz, B., Zajbt, E., Mołczanow, J. 2020. Polish iterative stress and its phonetic parameters in quiet vs. noisy environments. *Lingua* 240, 102835.
[12] Pitt, M. A., Samuel, A. G. 1995. Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology* 29, 149–188.
[13] Vitevitch, M. S., Luce, P. A. 1999. Probabilistic phonotactics and spoken word recognition. *Journal of Memory and Language* 40, 374–408.
[14] Vitevitch, M. S., Luce, P. A. 2003. The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistic Phonetics* 17, 487–499.
[15] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer. http://www.praat.org/. (v. 6.2.09)
[16] Łukaszewicz, B., Rozborski, B. 2008. Korelaty akustyczne akcentu wyrazowego w języku polskim dorosłych i dzieci. *Prace Filologiczne* 54, 265–283.
[17] Łukaszewicz, B., Rozborski, B. 2008. Acoustic correlates of word stress in child and adult Polish and the acquisition of phonological rhythm. Paper presented at Generative Linguistics in Poland (GLiP-6), Warsaw.
[18] Dogil, G. 1999. The phonetic manifestation of word stress in Lithuanian, Polish, German and Spanish. In: van der Hulst, H. (ed.), *Word prosodic systems of the languages of Europe.* Mouton de Gruyter, 273–311.

Weighted phonetic distances in the evaluation of phonotactics

Paula Orzechowska, Katarzyna Dziubalska-Kołaczyk

Adam Mickiewicz University in Poznań
paulao@amu.edu.pl, dkasia@amu.edu.pl

ABSTRACT

Net Auditory Distance is a phonotactic principle that computes the preferability of consonant clusters based on a universally preferred distribution of phonetic distances between adjacent segments. The distances pertain to place, manner and laryngeal features, all of which are expected to contribute to cluster structure. In this paper, we make an attempt at extending the principle by determining the most optimal combination of weighted distances required to predict the frequency of word-onset consonant groups in German. Statistical modelling using gradient boosted decision trees has revealed that, among others, the manner (sonority) distance of the consonant-vowel transition is the most important predictor of cluster frequency. The findings make it possible to assign weights to distances used in the computation of phonotactic preferability, ultimately leading to a more refined classification of German word-onsets.

Keywords: consonant clusters, phonetic features, weights, frequency, German

1. INTRODUCTION

One of the main themes in phonological research is the relationship between structure and usage. Efforts have been invested in the study of combinations of consonants in the languages of the world, revealing that their distribution patterns can be related to markedness and frequency [1-3]. This relationship has been also investigated for such clusters in Standard German, with markedness defined in terms of phonetic / phonological properties [4-6]. For instance, [6] asked which phonetic distances characterizing neighbouring segments best predict the frequency of word-initial clusters. The distances were derived from the Net Auditory Distance (NAD) principle (Sect. 2) that measures the size of contrast between adjacent consonants (C) and vowels (V) in terms of the place of articulation, manner of articulation and laryngeal features. The results of the study suggest that type frequency increases with an increase in the manner of articulation distance between two consonants (see [7] for similar results).

In this paper, we expand on the previous study by analysing the same dataset using different statistical methods. Although linear regression has been used to model the relationship between structure and usage, it is not always best-fitted to the nature of frequency data. Outliers lower the efficiency of linear models; [6] point to the lack of normality in residuals caused numerous low-frequency clusters and high-frequency /ʃt/. Thus, this analysis compares the predictions of linear regression and XGBoost, a technique used to model non-normal distributions.

In the following sections, we introduce the phonotactic model (2), propose a method of weighing distances based on XGBoost values (3), and discuss implications for a new weighted version of NAD (4).

2. PHONOTACTIC MARKEDNESS

2.1. Net Auditory Distance

NAD [8] is a measure of auditory distances between pairs of segments in a cluster, proposed in the framework of Natural Phonology [9,10]. Phonotactic preferability is computed based on well-formedness conditions incorporated from higher linguistic levels.

NAD is grounded in perceptual contrast, following the psychological principle of figure and ground [11]. It assumes that segments forming a universally preferred cluster should be sufficiently different from each other in order to be clearly perceived. Perceptual clarity is best achieved by the combination of a quieter C and a louder V [12,13] for two reasons. First, auditory cues in CV transition are more robust and richer in place cues than in VC [14,15]. Additionally, CV interface facilitates more precise articulation compared to VC [16].

The relevance of phonetic contrast is phonologically captured by the principle of clarity of perception [9], suggesting that larger phonetic distances between adjacent segments are expected to facilitate perception. This is reflected in NAD, where markedness conditions are based on a well-defined arrangement of net distances between pairs of

consonants and vowels neighbouring on them in word-initial, medial and final (C)CCs.

2.2. Computation of individual distances

Calculations are computed over 3 types of distances: the manner (MOA) and place (POA) of articulation as well as the sonorant/obstruent (SO) contrast, along the values in Table 1 and well-formedness conditions. The condition for a preferred initial CC requires that the distance between two consonants be larger than or equal to the distance between a consonant and the following vowel.
NAD(C1C2) ≥ NAD(C2V), where:
NAD(C1C2) = |(MOA1-MOA2)| + |(POA1-POA2)| + |SO|, and NAD(C2V) = |(MOA1-MOA2)| + |SO|.

In NAD, MOA distances correspond to sonority distances [17], which rise by 1 from a vowel towards the least sonorous plosives. POA distances, in turn, relate to the segments' anatomical location in the vocal tract [18]. The SO distance captures the sonorant/obstruent distinction. Distance=0 specifies S+S and O+O sequences, while distance=1 holds for sequences that belong to different classes.

S	A	F	N	L		G	V	
5.0	4.5	4.0	3.0	2.5	2	1.0	0	
p b			m				1.0	bilabial
	pf	f v					1.5	lab-dent
t d	ʦ	s z	n	l			2.0	alveolar
		ʃ ʒ					2.5	post-alv.
		ç j				j	3.0	palatal
k g		x	ŋ				3.3	velar
		ʁ			ʀ		3.6	uvular
							4.0	(radical)
		h					5.0	(glottal)

Table 1: NAD distances for German (capitals refer to Stop, Fricative, Affricate, Nasal, Liquid, Glide).

For example, the NAD computation for /bj/ involves: C1C2=|5-1|+|1-3|+|1|=7, and C2V=|1-0|+0=1. The cluster meets the well-formedness condition (7≥1), and is thus preferred. The total distance value for a cluster, referred to as 'NAD product', is an index expressing the degree of cluster preferability. NAD product is calculated by means of subtraction: NAD(C1C2)-NAD(C1V), and amounts to 6 for /bj/. The larger the NAD product value, the better the cluster. (For a detailed exposition of other conditions and more detailed NAD categories see [8].)

The numerical values assigned to each consonant reflect the fact that various languages may require phonological distinctions that are more or less fine-grained. Thus, in line with the sonority scale in [19], the most recent version of NAD contains 12 MOA classes for consonants and 5 classes of vowels based on height and peripherality (see [20] for an overview of the revised NAD version).

3. ANALYSIS

3.1. Data

The data analysed in this study were drawn from [4], who extracted a cluster list from phonological descriptions, coursebooks, dictionaries [21-23]. Some clusters are found in rare words, proper nouns and loans, e.g. /bj skv/ (*Björn*, *Squash*). For reasons of space, the proposed analysis is based only on CC types, as presented in Table 2.

Clusters (n=46)
bj bl bʀ dʀ fj fl fʀ gl gm gn gʀ kl km kn ks kʀ kv pfl pfʀ pl pn ps pʀ sf sk sl sm sn sp sʀ st sts sv ʃk ʃl ʃm ʃn ʃp ʃʀ ʃt ʃv tj tʀ tv tsv vʀ

Table 2: Initial CC clusters in German.

Table 3 presents type frequency of each cluster, i.e. the cumulative frequency of all words starting with this cluster. The data were drawn from the corpus of newspaper texts *Leipziger Wortschatz-Portal* [24], which contains 1.65 million word types. In rows with multiple clusters, the frequency value refers to an individual CC.

Cl	Freq	Cl	Freq	Cl	Freq
ʃt	1261	bl	248	ps	27
pʀ	754	gl	247	sv, sm	14
ʃp	648	tsv	235	sp	13
gʀ	619	pl	188	ks, bj, gn	10
fʀ	599	kv	144	sts	9
kʀ	569	ʃʀ	121	sn, tj	8
tʀ	538	ʃn	114	fj, sʀ	6
bʀ	458	ʃm	99	tv	5
kl	429	kn	98	vʀ	4
dʀ	324	sk	80	sf, gm	2
ʃv	312	pfl	53	pn, ʃk, km, pfʀ	1
fl	300	st	45		
ʃl	292	sl	28		

Table 3: Type frequency of initial CC clusters.

3.2. Procedure

We ran a series of analyses using linear regression and XGBoost models, and the R environment [25]. The goal was to identify those independent variables

that best predict type frequency (i.e. variables generating models with the lowest mean squared error). We compared linear regression with models using XGBoost [26], i.e. gradient-boosted decision trees. Both techniques make it possible to find relations between variables but linear regression is limited to linear dependencies, while XGBoost is also sensitive to non-linear relations. Informally speaking, the latter method divides data into homogenous groups depending on the values of the dependent variable. The contribution of a variable to a model is expressed with a numerical value, indicating the variable's importance. The comparison of models was based on the mean squared error obtained through 5-fold cross-validation process (C-V MSE). A complete list of tested models M (numbered 1-20) is given in Table 4. The MSE values are rounded to the nearest tens.

M	Independent variables	MSE
1	MOA_C2V (0.46)+MOA_C1C2 (0.28)+ POA_C1C2 (0.27)+SO_C1C2 (0)+SO_C2V (0)	260
2	MOA_C2V (0.35)+POA_C1C2 (0.34) +MOA_C1C2 (0.19)+SO_C1C2 (0.12)	261
3	MOA_C2V (0.45)+POA_C1C2 (0.29)+MOA_C1C2 (0.26)+SO_C2V (0)	261
4	MOA_C2V (0.43)+MOA_C1C2 (0.29) +POA_C1C2 (0.28)	257
5	MOA_C1C2 (0.56)+POA_C1C2 (0.44)	247
6	MOA_C2V (0.57)+MOA_C1C2 (0.43)	252
7	MOA_C2V (0.51)+POA_C1C2 (0.49)	258
8	**MOA_C1C2 (1)**	**239**
9	POA_C1C2 (1)	252
10	**MOA_C2V (1)**	**242**
11	MOA_C1C2 (0.54)+POA_C1C2 (0.44) +SO_C1C2 (0.03)	248
12	**MOA_C1C2 (0.82)+SO_C1C2 (0.18)**	**240**
13	POA_C1C2 (1)+SO_C1C2 (0)	259
14	SO_C1C2 (1)	259
15	**MOA_C2V (0.77)+SO_C2V (0.23)**	**234**
16	SO_C2V (1)	259
17	NAD_C2V (0.56)+NAD_C1C2 (0.44)	278
18	NAD_C1C2 (1)	291
19	**NAD_C2V (1)**	**242**
20	NAD_product (1)	290

Table 4: Independent variables constituting tested models. The estimated importance of each variable in XGBoost is given in brackets. The best models (with the lowest C-V MSE values) are marked in bold type.

Type frequency was coded as a dependent variable. The tested models involved different combinations of independent predictors derived from the well-formedness condition in Sect. 2.2, i.e. distances for pairs of segments (MOA_C1C2, POA_C1C2, SO_C1C2, MOA_C2V, SO_C2V), summated distances for pairs of segments (NAD_C1C2, NAD_C2V) and NAD product.

3.3. Results

In this section, we present XGBoost analyses for CC word onsets based on 5 best XGBoost models, and compare them with linear regression models.

The best model (15, C-V MSE=234.31) features two variables; MOA and SO distances for the C2V transition. The model constitutes a better fit to the frequency data than linear regression (C-V MSE=273.22). The primacy of the non-linear model is visualized below.

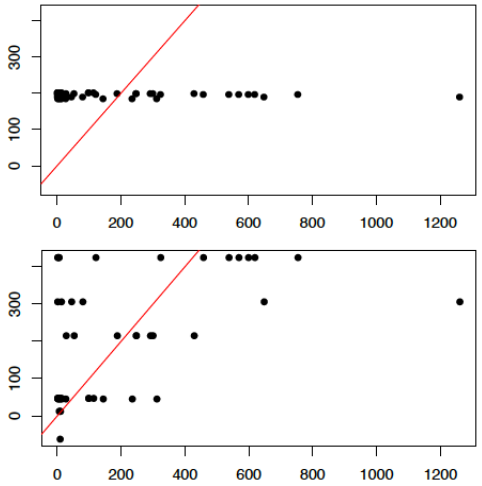


Figure 1: Scatterplots presenting observed vs. expected values for linear regression (top) and XGBoost (bottom) using MOA_C2V and SO_C2V variables (model 15). In both plots, the diagonal line represents the line of perfect fit.

Figure 1 juxtaposes the distribution of the predicted (y axis) and observed (x axis) values in linear regression and XGBoost. As can be seen, the predicted values of XGBoost are much closer to the observed values. All expected values of linear regression lie within a very narrow range (ca. 180-210), which is not in line with the observed values (most of which range from 1 to ca. 600). The XGBoost model is much more flexible, displaying a greater variety of expected values (mostly in the range from 1 to ca. 400). The strength of XGBoost lies in the lack of assumption on linearity, and its ability to capture non-linear dependencies. Model 1 5 captures the contribution of MOA_C2V and

SO_C2V, whereby the importance of the manner distance (77%) outweighs the SO distance (23%).

Other models with low C-V MSE values are listed under (8, 10, 12, 19). The summary of the prediction error obtained through cross-validation for the models is given in Table 5. In each case, the decision tree method provides better results. For clarify of presentation, we also include model 15.

M	Linear regression	XGBoost
8	251.69	238.94
10	264.93	241.84
12	260.11	239.66
15	273.22	234.31
19	266.14	241.84

Table 5: Comparison of the CV-MSE measures for 5 best linear regression and XGBoost models.

4. TOWARDS WEIGHTED AVERAGE NAD

The findings of the present study offer a starting point for introducing weights to the NAD principle. The analyses have revealed not only a novel statistical method that can be successfully used in determining such weights but also a selection of distances that could be possibly eliminated from future NAD calculations due to being of little use in predicting type frequency.

The model that best predicts the frequency of CCs includes two predictors: MOA_C2V and SO_C2V, among which the manner distance plays a prime role (ca. 80%). Other best models in Table 5 partially overlap with model 15: (8, 12) are also based on MOA_C1C2, and the remaining models (10, 19) include MOA_C1C2 and/or SO_C1C2.

It must be emphasized that models including a wider range of variables tend to display a greater C-V MSE compared to simpler models (e.g. 2, 3). This holds true particularly for models that contain POA_C1C2 and/or SO_C1C2, SO_C2V. In fact, the importance of the latter two variables tends to be low or none in models with three or more independent variables (1-3) or models using POA (9, 13). This suggests that information provided by SO duplicates information arising from POA, but can supplement the information from MOA.

Finally, cumulative distances were shown to be poor predictors of cluster frequency. NAD_C1C2 and NAD product are characterized by the highest C-V MSE values.

5. CONCLUSIONS

The findings of the study support previous theoretical and empirical work on phonotactics. First of all, they suggest that manner (or sonority) distances [27-29] constitute a relevant phonotactic primitive and motivate the core structure of word-initial clusters in German. A similar observation was made in an independent study by [7]. Using the same data and a different set of predictors, [7] demonstrated that manner of articulation distances and voicing constitute the backbone structure of German clusters. That is, average frequency values of clusters were shown to be higher for clusters displaying larger manner distances and contrast in voicing. These observations go in line with [30] arguing that an adequate description of phonotactic possibilities of a language requires operating on weighted phonological features.

Moreover, the results, along with the previous contributions, testify to the relevance of fine-grained categories in the study of phonotactics. Neither NAD product nor cumulative NAD values were shown to account for the frequency data. This observation might also explain why the sonority slope of clusters, usually expressed in binary terms as well-formed/ill-formed, preferred/dispreferred, turns out to be an insufficient criterion in accounting for complex sequences of segments. In turn, subtle sub-segmental cues and gradient classification of structures are expected to be more informative, as argued in [30].

Previous work [7] using the same dataset revealed the importance of CC distances. Here, we have showed that MOA and SO distances for the consonant-vowel transition constitute the best correlates of type frequency. The CV transition, representing an alternation of a quieter C and louder V, is the most salient sequence cross-linguistically [13, 14]. This observation is captured by the distance matrix in [20]: in CCV, the sequence of consonants should display larger contrast compared to CV.

Finally, let us note that type frequency should be viewed as reflecting the phonotactic potential of a language. The relevance of the manner of articulation might also suggest a direction through which the phonotactic inventory of German has developed historically (see also [31] for similar results), and a way in which it might expand. The question on how every-day usage shapes cluster inventories (token frequencies) will be addressed in another study.

6. REFERENCES

[1]Jusczyk, P.W; Luce, P.A., Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *J. of Memory and Language* 33(5), 630-645.

[2] Stites, J., Demuth, K., Kirk, C. 2004. Markedness vs frequency effects in coda acquisition. In: Brugos, A., Micciulla, L., Smith, C.E. (eds.), *Proceedings of the 28th Annual Boston University Conference on Language Development,* 565-576.

[3] Van de Vijver, R., Baer-Henney, D. 2012. Sonority intuitions are provided by the lexicon. In: Parker, S. (ed.), *The Sonority Controversy*. Berlin: de Gruyter, 195-215.

[4] Orzechowska, P., Wiese, R. 2015. Preferences and variation in word-initial phonotactics: A multi-dimensional evaluation of German and Polish. *Folia Linguistica* 49, 439-486.

[5] Ulbrich, C.; Alday, P., Knaus, J., Orzechowska, P., Wiese, R. 2016. The role of phonotactic principles in language processing. *Language, Cognition and Neuroscience* 31(5), 662-682.

[6] Orzechowska, P., Dziubalska-Kołaczyk., K. 2022. Gradient phonotactics and frequency: A study of German initial clusters. *Italian J. of Linguistics* 34(1), 103–138.

[7] Wiese R., Orzechowska, P. (in press). Structure and usage do not explain each other: An analysis of German word-initial clusters, *Linguistics*.

[8] Dziubalska-Kołaczyk, K. 2019. On the structure, survival and change of consonant clusters. *Folia Linguisitca Historica* 40(1), 107-127.

[9] Donegan, P. J. & Stampe, D. 1979. The study of natural phonology. In: Dinnsen, D.A. (ed.), *Current Approaches to Phonological Theory*. Bloomington: Indiana University Press, 126-173.

[10] Dressler, W.U. 1985. *Morphonology: The dynamics of derivation*. Ann Arbor: Karoma Publishers.

[11] Dressler, W. U. 2009. Natural phonology as part of natural linguistics. *Poznań Studies in Contemporary Linguistics* 45(1), 32-42.

[12] Greenberg, J. H. 1978. Some generalizations concerning initial and final consonant clusters. In: Greenberg, J. H. (ed.), *Universals of Human Language*. Stanford: Stanford University Press, 243-279.

[13] Maddieson, I. 2013. Syllable structure. In: Dryer, M., Haspelmath, M. (eds.), *The World Atlas of Language Structure Online*. Munich: Max Plank Digital Library.

[14] Maddieson, I. 1999. In search of universals. In: Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., Bailey, A. C. (eds.), *Proceedings of the 14th ICPhS,* 2521-2528.

[15] Ohala, J.J. 1990. The phonetics and phonology of aspects of assimilation. In: Kingston, J. & Beckman, M. (eds.), *Papers in Laboratory Phonology I*. Cambridge: CUP, 258-275.

[16] Ohala, J.J. & Kawasaki, H. 1984. Prosodic phonology and phonetics. *Phonology Yearbook* 11, 113-129.

[17] Goldsmith, J.A. 1990. *Autosegmental and Metrical Phonology*. Blackwell.

[18] Ladefoged, P. 2006. *A Course in Phonetics* (5th ed.). Boston: Heinle & Heinle.

[19] Parker, S. 2008. Sound level protrusions as physical correlates of sonority. *J. of Phonetics* 36, 55-90.

[20] Dziubalska-Kołaczyk, K., Pietrala, D. 2020. The NAD Phonotactic Calculator – an online tool to calculate cluster preferability across languages. In: Sauer, H., Chruszczewski, P.P. (eds.). *Mostly Medieval. In Memory of Jacek Fisiak.* San Diego: Academic Publishing, 445-458.

[21] Meinhold, G., Stock, E. 1980. *Phonologie der Deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut Leipzig.

[22] Duden Aussprachewörterbuch 1990 / 2009 (6th ed.). *Duden: Wörterbuch der deutschen Standardaussprache*. Mannheim / Wien / Zürich: Dudenverlag.

[23] Wiese, R. 2000. *The Phonology of German* (2nd ed). Oxford: Clarendon Press.

[24] *Leipziger Wortschatz-Portal*. Access online: <wortschatz.informatik.uni-leipzig.de/de>.

[25] R Core Team, R: *A language and environment for statistical computing*, R foundation for statistical computing, Vienna, Austria, 2020.

[26] Chen, T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785-794.

[27] Parker, S. 2012. Sonority distance versus sonority dispersion—A typological survey. In: Parker, S. (ed.), *The Sonority Controversy*. Berlin: Walter de Gruyter, 101–166.

[28] Selkirk, E.O. 1984. On the major class features and syllable theory. In: Aronoff, M., Oehrle, R.T. (eds.), *Language sound structure*. Cambridge, MA: The MIT Press, 107–136.

[29] Steriade, D. 1982. *Greek prosodies and the nature of syllabification*. PhD diss., Cambridge, MA: The MIT Press.

[30]Orzechowska, P. 2019. *Complexity in Polish phonotactics: On features, weights, rankings and preferences*. Springer.

[31] Baumann, A., Wissing, D. 2018. Stabilizing determinants in the transmission of phonotactic systems: Diachrony and acquisition of coda clusters in Dutch and Afrikaans. *Stellenbosch Papers in Linguistics Plus* 55, 77-107.

ANIMATED FILM CHARACTER PROFILES: THE ROLES OF VOICE AND LEXICAL CONTENT

Alice Crochiquia [1,2], Anders Eriksson[1], Sandra Madureira[2], Plinio A. Barbosa[3]

[1]Stockholm University, Sweden; [2]PUC-SP, Brazil; [3]UNICAMP, Brazil
alice.crochiquia@ling.su.se; anders.eriksson@ling.su.se;
sandra.madureira.liaac@gmail.com; pabarbosa.unicampbr@gmail.com

ABSTRACT

Our study examines the attribution of physical and psychosocial features to the characters of a dubbed animated film by comparing the results of a listening test experiment with the results of a reading test experiment. The aim of this study was to investigate the potential influence of lexical content on the perceptual evaluation of speaker characteristics. In the listening test, respondents were asked to listen to audio samples of dubbed dialogues in the film produced by voice actors and rate the characters on fourteen bipolar continuous scales. The task in the reading test was the same, except for the stimuli, which were transcriptions of the audio files. Results from both tests are congruent, as the character profiles based on the perception scores from each test are nearly identical for all characters. Differences are mainly due to the degree judgement scores are amplified by the characters' voice quality profiles.

Keywords: dubbing, perception, animated film, voice acting, character features.

1. BACKGROUND

A professional film dubbing is a collaborative effort by translators, dialogue writers, dubbing directors, voice actors and sound engineers, and the result is a type of production in which features of spoken and written text are present [1]. An important step in the process of dubbing is the selection of actors to lend their voices to the characters in a new language, as audiences expect "character synchrony" [2], that is, vocal stereotyping and the association of specific vocal features to certain types of characters.

Compared to dubbing for live-action films, dubbing for animated films, especially those aimed at children, presents far fewer constraints regarding another type of synchronization, phonetic or lip synchrony, which is the matching of articulatory movements shown on screen and sounds produced by the voice actors. That is due both to the less precise articulatory movements of the characters in comparison to real actors and child audiences being less demanding [3]. Thus, in those cases, priority can be given to translations that are semantically similar to the wording in the original text but not in articulatory detail, and in that way, keep the content of character dialogues in the target language close to the one in the source text.

This paper is about the influence of lexical content on the perception of characters from an animated film, where we compare the results from a reading test experiment to results from a previous study, which included a listening test experiment.

The previous study [4] covered the character synchrony in the Brazilian Portuguese dubbing of an animated film, in a combined analysis of the impressions of listeners regarding the attitudes, physical and social characteristics of the characters, and the acoustic parameters of each of the voice actors who performed as the characters. The results of that experiment showed that actors' voice acoustic profiles were mainly differentiated in relation to *f0* mean and range, and harmonic to noise ratio (HNR). Based on the perception scores of physical, psychological, and social features of the Listening Test, five distinct character profiles were generated. From those two sets of data, correlations between acoustic, physical, social, and attitudinal characteristics were estimated. Among the correlations with the highest values were *f0* maximum and Size, *f0* maximum and Age, *jitter* and Honest and Kind.

A major concern in studies of the perception of physical, psychological, and social features based on speech stimuli is the influence of semantic and syntactic content of the recorded material to be judged [5].

There are several masking techniques [6] to isolate voice from linguistic information, among them reverse speech, listening to unknown languages, meaningless content recordings, randomizing splicing, content filtering, and even acoustic-driven phoneme removal transformation techniques [7] but all of them have drawbacks. Listeners perceive them as odd and perceptual judgments are affected. Neutralizing techniques, which use the same text voiced by different speakers, are also often applied to avoid lexical influence but as [5] mentions despite the same semantic and syntactic content perceptual judgments are not essentially neutral because they can be influenced by judges' social, behavioral, and individual characteristics.

Given the semantic, syntactic, and vocal interacting factors influencing listeners' judgments of speech stimuli in perceptual tests, investigating how lexical

content affects listeners' judgments is an important goal and the focus of the present work.

2. METHODS

2.1. Material

The material for both experiments was taken from the Brazilian dubbing of the Disney animated film Zootopia. The stimuli in the Listening Test were audio samples of dialogues produced by voice actors performing as the four main characters in the film (Assistant Mayor Bellwether, Chief Bogo, Judy Hopps, and Nick Wilde). For the experiment, five audio samples were selected for the corpus as Assistant Mayor Bellwether, who is revealed as the film's villain, has two drastically different characterizations throughout the narrative, and was thus considered as two distinct characters. The stimuli in the Reading Test were the transcriptions of those audio samples.

2.2. Listening Experiment

The experiment was conducted entirely online, on SurveyMonkey. The first page of the link given to the respondents presented a brief description of the experiment and a questionnaire to collect data on the respondents' background, including whether or not they had seen the film. A group of 77 Brazilian Portuguese native speakers, aged between 20 and 50, took part in the experiment.

The respondents, 46 women and 31 men, were asked to listen to the audios and rate the characters on 14 bipolar continuous scales regarding their size, age, temper, attitudes, character, social and vocal features. The respondents had to do that by using a sliding tool to indicate the point of the scale they felt best described the character, which was translated into a score from 0 to 100. Table 1 shows the scales and the pairs of descriptors used in the experiment in Brazilian Portuguese and their translation to English.

Based on the mean values of the perception scores for the five characters, a profile was established for each of them. Table 2 shows the profiles of the friendly Bellwether (B1), Bellwether's villainous persona (B2), Chief Bogo (CB), Judy Hopps (JH), and Nick Wilde (NW).

For most of the features considered for the profiles, we selected those with scores higher than 65.5 and lower than 35.5 as significant and added the qualifiers based on the scores within those ranges. For Age and Size, however, all scores were considered, as the values around the mid-points of those scales do not represent a neutral reading of the character on those scales, but rather an impression of the speaker as mid-sized or middle-aged. More details regarding the characters' profiles can be found in [4].

Scale	Descriptors
Tamanho	*Pequeno - Grande*
(Size)	(Small - Big)
Idade	*Jovem - Velho*
(Age)	(Young - Old)
Temperamento	*Agressivo - Dócil*
(Temper)	(Aggressive - Docile)
Atitudes	*Dominador - Submisso*
(Attitudes)	(Dominant - Submissive)
	Preguiçoso - Ativo
	(Lazy - Energetic)
	Durão - Gentil
	(Tough - Gentle)
	Charmoso - Desinteressante
	(Charming - Uninteresting)
	Bondoso - Maldoso
	(Kind - Mean)
	Cauteloso - Audacioso
	(Cautious - Bold)
	Alegre - Triste
	(Happy - Sad)
Caráter	*Disonesto - Honesto*
(Character)	(Dishonest - Honest)
Características sociais	*Sofisticado - Simples*
(Social features)	(Sophisticated - Simple)
Voz	*Animada - Monótona*
(Voice)	(Lively - Monotonous)
	Agradável - Desagradável
	(Pleasant - Unpleasant)

Table 1: Scales and descriptors in the Listening test.

Characters	Features
B1	Small, Young
	Happy, Kind, Very Gentle,
	Charming
	Lively, Pleasant
B2	Small, Young
	Rather Tough, Rather Aggressive,
	Not Charming
	Rather Lively, Not Pleasant
CB	Big, Not Young
	Very Tough, Very Aggressive,
	Not Charming
	Not Lively, Not Pleasant
JH	Small, Young
	Honest, Happy,
	Charming
	Rather Lively, Pleasant
NW	Mid-sized
	Not Young, Charming Not
	Lively, Pleasant

Table 2: Character profiles.

2.3. Reading Experiment

The design for this test was nearly identical to the previous one. Among the few differences were the

exclusion of the question about hearing impairments in the questionnaire and the change for the last two scales from *Voz* to *Maneira de expressão* (Manner of expression) as the test did not include listening to the voices. The descriptors all remained the same.

The test was presented, also on SurveyMonkey, to a group of 73 Brazilian Portuguese native speakers, 43 women and 30 men, aged between 19 and 58. The stimuli in the two tests were not referred to by the names of the characters in the film, nor did any of those names occur in the audio and text samples.

2.4. Interrater reliability

For a perception experiment based on ratings by listeners to be reliable, there must be a high degree of agreement between judges. As a measure of interrater agreement, we have used Cronbach's Alpha. With this method, values can range from 0 to 1, indicating complete disagreement and perfect agreement, respectively. Values above 0.7 are considered acceptable, and scores higher than 0.9 are considered excellent.

3. RESULTS

Table 3 shows the Cronbach's Alpha values between male and female judges, and between those who had seen the film (**S**) and those who had not (**nS**) in the Reading Test.

			Male	Female
	z-norm	**M/F**	**S/nS**	**S/nS**
B1	0.987	0.990	0.980	0.988
B2	0.934	0.984	0.973	0.986
CB	0.951	0.965	0.916	0.968
JH	0.955	0.967	0.873	0.911
NW	0.940	0.931	0.890	0.927
All	0.961	0.976	0.953	0.965

Table 3: Interrater reliability for the Reading Test.

Table 4 presents the mean score values. Low values are closer to the descriptor on the left end of the scale, and higher values to the right end of it. To compile more detailed impressions of the characters, we've chosen to highlight scores higher than 80 and lower than 20 in this experiment as strong, as opposed to the 65.5 and 35.5 values in the first experiment. These cutoff points are not definitive and may still be modified in planned future analyses and experiments.

Table 5 shows the comparison between the mean values of the perception scores from listeners and those from readers who had not seen the film, who did not have any information on the characters beyond the reading of the transcriptions during the test. The correlations show that even in the complete absence of speech information and without previous knowledge of the characters in the film, the reading profiles are similar to listening profiles.

4. DISCUSSION

Results of the Listening Test and the Reading Test were found to be coherent, that is, the same characteristics were attributed to the characters in the two tests. However, some of the characteristics were evaluated with stronger scores in one of them.

Strong scores in both tests were attributed to Bellwether 1 on the following characteristics: Docile, Energetic, Happy, Lively; to Bellwether 2, Energetic; and to Chief Bogo, Dominant and Tough.

The similarities between the results for the two tests extended beyond the very high or very low scores. Judy and Nick received overall weaker scores in both tests. The strongest score attributed to Judy was Energetic (83 in the Listening Test and 74 in the Reading Test) and to Nick was Dominant (34 in the Listening Test and 21 in the Reading Test) and Bold (67). Nick did not have scores below 31 or above 60 on any of the other scales. In the Listening Test, his strongest scores were for Charming and for a Pleasant voice.

Coherence in relation to both the Listening Test and the Reading Test can be found in the characteristic attributions to Bellwether 1 as Docile, Gentle and Honest and to Bellwether 2 as Aggressive, Tough, and Dishonest; to Chief Bogo as Big, Old, and Tough; to Judy as Energetic and Honest. Nick had a few close scores in Listening and Reading, however, most of them were around the mid-mark of the scales, unlike the strong scores of both Bellwethers and Chief Bogo.

Comparison inter-characters shows that Chief Bogo got the strongest scores for Old, Aggressive and Tough; Bellwether 1 for Young, Kind, Docile, Pleasant, Happy, Lively, and Charming; Bellwether 2 for Bold.

For many of the acoustic parameters considered in the previous study, a similar pattern to the perception scores emerged, especially those related to *f0*. The values for Bellwether 1, Bellwether 2 and Chief Bogo were placed at the ends of the spectra, and Judy and Nick fell somewhere in the middle.

The slightly stronger scores on the Listening Test than in the Reading Test convey the impact of pitch and voice quality settings on characteristic attributions. Chief Bogo's low-pitched voice was paired to Bigger, Older, more Dominant, Tougher judgments; Bellwether 1's, Bellwether 2's and Judy's high-pitched voices were paired with Younger, and more Lively judgments. Nick's Whispery and Harsh voice quality with slightly higher scores for Charming and Age (Older).

	B1		*B2*		*CB*		*JH*		*NW*	
Scales	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**
Small - Big	30	**19**	24	29	**87**	63	34	35	53	48
Young - Old	23	27	23	36	65	59	28	36	60	34
Aggressive - Docile	**87**	**83**	31	28	**14**	25	51	49	57	32
Dominant - Submissive	56	50	**12**	22	**9**	**16**	31	32	34	21
Lazy - Energetic	**88**	**83**	84	83	78	68	**83**	74	44	60
Tough - Gentle	**87**	77	31	24	**9**	**17**	42	45	55	31
Charming - Uninteresting	22	67	37	41	49	61	32	39	34	46
Kind - Mean	**18**	**22**	63	55	28	60	38	38	48	57
Cautious - Bold	50	36	74	72	72	64	68	52	46	67
Happy - Sad	**11**	**16**	35	37	55	48	29	39	49	41
Dishonest - Honest	73	71	39	55	41	49	73	70	46	43
Sophisticated - Simple	55	53	36	45	50	55	41	44	45	59
Lively - Monotonous	**7**	**15**	26	31	52	48	22	40	42	39
Pleasant - Unpleasant	26	**18**	44	49	49	65	29	40	33	57
Correlation	0.875		0.937		0.841		0.928		-0.198	

Table 4: Mean values for perception scores for the characters' features in the two experiments.

	B1		*B2*		*CB*		*JH*		*NW*	
Scales	**L**	**R(nS)**	**L**	**R(nS)**	**L**	**R(nS)**	**L**	**R(nS)**	**L**	**R(nS)**
Small - Big	30	23	24	27	87	56	34	38	53	47
Young - Old	23	33	23	36	65	54	28	38	60	31
Aggressive - Docile	87	81	31	32	14	26	51	45	57	32
Dominant - Submissive	56	46	12	22	9	17	31	32	34	20
Lazy - Energetic	88	81	84	84	78	66	83	73	44	64
Tough - Gentle	87	72	31	24	9	20	42	47	55	29
Charming - Uninteresting	22	35	37	40	49	64	32	38	34	45
Kind - Mean	18	23	63	52	28	60	38	38	48	59
Cautious - Bold	50	38	74	71	72	62	68	50	46	67
Happy - Sad	11	18	35	36	55	46	29	41	49	38
Dishonest - Honest	73	72	39	58	41	46	73	70	46	44
Sophisticated - Simple	55	53	36	46	50	55	41	40	45	55
Lively - Monotonous	7	17	26	30	52	44	22	42	42	35
Pleasant - Unpleasant	26	21	44	45	49	67	29	40	33	58
Correlation	0.966		0.921		0.767		0.893		-0.215	

Table 5: Perception scores of listeners and readers who had not seen the film.

5. RESULTS

Overall, results indicate that the lexical content in the dialogues was enough to lead readers to an accurate assessment of the characters especially in the case of Bellwether and her two personas, and we must point out that the majority of the dialogue lines did not contain qualifiers that would immediately make the features of the characters explicit to the listeners and readers.

However, the results in this study have not been enough to lay out in detail how the lexical content may have been a deciding factor in the perception of these characters. We hope to gain more insight into the influence of the lexical content with additional analyses of the dialogue lines and with similar listening and reading experiments with Zootopia in different languages.

With the comparison of the results of both experiments, evidence of the influence of parameters like pitch means and phonation voice quality settings can be inferred from the higher intensity of characteristics attributed to size, age, and some attitudinal features. The voice performances did match the characters' features displayed in the text and in most cases served to amplify them.

Even in Nick's case, where none of his perception ratings met the new threshold for strong scores, the similar patterns in the results for the perception experiment and the acoustics may still provide some insight into the use of voice as characterization as he is portrayed in the film as a hard-to-read character.

Some of these correlations highlighted as significant in the previous study are already well-documented in the literature on vocal expressivity, like high values for parameters related to *f*0 and small size, and the opposite for low values, as well as low values for Harmonic-to-Noise Ratio as a vocal aging index. However, correlations like low H1-H2 values and higher scores for Dominance, Aggressiveness are not as efficiently explained by findings in previous studies and require further inspection.

6. ACKNOWLEDGEMENTS

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brasil (CAPES) - Finance Code 001, grant #88881.689915/2022-01. Participation at ICPhS 2023 by the first author has been supported by a grant from *The Swedish Foundation for International Cooperation in Research and Higher Education* (STINT, BR2027-7253). The third author acknowledges a grant from PIPEq #21672, PUC-SP. The fourth author acknowledges a grant from CNPq #302194/2019-3.

7. REFERENCES

[1] C. Bosseaux, "Investigating dubbing: Learning from the past, looking to the future," in *The Routledge Handbook of Audiovisual Translation*, L. PérezGonzález, Ed. Routledge, 2018, pp. 48–63.

[2] C. Whitman-Linsen, *Through the dubbing glass: the synchronization of American motion pictures into German, French, and Spanish*. Peter Lang, 1992.

[3] F. Chaume, "Synchronization in dubbing: a translational approach," in *Topics in Audiovisual Translation*, P. Orero, Ed. John Benjamins Publishing Company, 2004, pp. 35–52.

[4] A. Crochiquia, A. Eriksson, P. Barbosa, S. Madureira, "A perceptual and acoustic study of dubbed voices in an animated film", Proc. Speech Prosody 2022, pp. 565–569, 2022.

[5] J. Pittam, *Voice in Social Interaction: An Interdisciplinary Approach*. SAGE, 1994.

[6] K. R. Scherer, S. Feldstein, R. N. Bond, and R. Rosenthal, "Vocal cues to deception: a comparative channel approach," *Journal of Psycholinguistic Research*, vol. 14, pp. 409–425, 1985.

[7] C. Noufi, J. Berger, M. Frank, K. Parker, and D. L. Bowling, "Acoustically-driven phoneme removal that preserves vocal affect cues," arXiv. Preprint, 2022.

Energetic and informational masking effects on Spanish vowel recognition

Mark Gibson[1], Marcel Schlechtweg[2], Judit Ayala[1], Andrea DiCiaccio[3], Xianhui Wang[3], Li Xu[3],
[1]Universidad de Navarra, [2]Carl von Ossietzky Universität Oldenburg, [3]Ohio University,
Corresponding author: mgibson@unav.es

ABSTRACT

We report results of an ongoing study examining the effects of different masking conditions (background babble, or the so-called cocktail party effect, and varying levels of noise saturation in relation to the signal, i.e. signal-to-noise ratio, henceforth SNR) on vowel recognition for two groups: native English speakers taking a test in Spanish, L1-EN, and native Spanish speakers, L1-SP, taking the same Spanish test. Results for the current study are commensurate with our earlier work showing notable confusion when discerning the rounded back vowels [o] and [u] for both masking conditions and language groups (though the L1-EN group showed less release from masking), which we surmise ensues from masking of the first formant (F1) by the third formant (F3), and a lack of visual information referencing jaw angle and/or lip aperture. The biological sex of the listener made no difference on recognition, though both groups showed a bias for the female voice targets (stimuli). Effects of language-specific biases seem likely, but the nature of the interaction of these effects is still unknown.

Keywords: Speech perception in noise, vowel recognition, language effects.

1. INTRODUCTION AND BACKGROUND

Formant frequencies and spectral shape have both been shown to be crucial cues for vowel recognition [1]. It is known that noise affects the access listeners have to these acoustic cues, but it is not altogether clear how different types of noise mask different acoustic dimensions necessary for vowel recognition (see [2] for a good explanation on how noise affects information flow in speech). On one hand, steady-state noise physically interferes with the intended signal blocking access to the acoustic cues transmitted by the sender. This has become known as *energetic masking* (see [3]), which falls out when peripheral neural activity intended to parse the transmitter's signal is overwhelmed by a noise source. A simple example of energetic masking would be where high-volume white noise cuts off access to an intended object of perception. On the other hand, *informational masking* occurs when linguistic, or other higher order dimensions of the noise source create uncertainty, or entropy, with regard to the information flow, which may lead to perceptual errors. A common example of informational masking would be where linguistic/acoustic-phonetic information, say from speakers in the background (i.e. cocktail party effect) conditions the correct interpretation of a stimulus [4].

Previous studies have shown that informational masking is reduced when similarity of the noise source and target stimulus is reduced (when they are more spatially distinct [5, 6]), or when the background speech (i.e. noise) is produced by a speaker who is the opposite biological sex of the speaker producing the stimulus [7, 8]. Additionally, effects of masking have been shown to be higher for intelligible noise (say, where the background babble is clear, and in the perceiver's native language) as compared to non-intelligible noise (say where speech has been modified, or a foreign language is used for background babble). Disentangling the effects of each type of masker is problematic due to difficulties in segregating the informational from energetic mechanisms of speech-on-speech masking.

A dearth of evidence shows that a listener's capacity to discriminate phonological contrasts is greatly reduced by exogenous factors such as the level of noise in relation to the signal (SNR) and the number of speakers in multi-speaker background babble [9-11]. Information provided by complementary signals or modalities, such as a visual input, has also been reported to have a role in parsing speech in noise [12].

Fewer studies have focused on the effects that language (both the native language, henceforth L1, of the perceiver and the language of the targets/maskers), and other endogenous factors related to the perceiver, have on speech processing in noise. However, the results from this line of research are often contradictory, and to a large extent mutually exclusive. Those studies that have addressed language-specific effects for listening in noise have mainly focused on the effects of second language, henceforth L2, deficits, and the advantages exposure in a L2 has on recognition accuracy [13], recognition accuracy increasing as a function of proficiency in the L2. At the same time, previous studies also show that release from masking is higher when background babble is unintelligible (see [5-8]). Thus, it is an intuitively straightforward observation that if unintelligibility of the background masker enhances release from masking, then exposure to that language should not increase recognition accuracy. Would not

exposure to the phonetic/acoustic dimensions of the language of exposure go part and parcel with higher morpho-syntactic/lexical awareness, which is supposedly the source of informational masking?

Thus, it is our working assumption that language-specific interactions on speech recognition in noise may not be so simplistic. We assert that L1 biases may aid in vowel recognition given the right circumstances (based on inventory size across the languages involved in the tests), and the ability to suppress informational maskers (multi-speaker background babble) may be enhanced when the masker is in a second language [5, 14, 15].

With regard to how L1 biases may enhance L2 vowel recognition in noise, a speaker's native phonological inventory (size) may affect the subsequent weighting of acoustically salient cues when perceiving speech, conditioning the attention of the listener to hear specific contrasts. Acoustic cue-weighting as a function of its reliability in speech recognition has a long tradition in the perception literature [16], which may be pertinent to speech recognition in noise. The more reliable a cue is in signaling a phonological contrast, the more perceptual weight that cue receives, and hence directs a speaker's attentional focus to that cue (see [16] for a good review). Accordingly, a speaker from a language with a large vowel inventory (with high inter-categorical overlap and high intra-categorical variation) like English may have an advantage over a speaker with a relatively small vowel inventory (with low inter-categorical overlap and low intra-categorical variation), such as Spanish, in vowel recognition in noise because their attention is attuned to finer acoustic cues.

In addition to any advantage L1 vowel inventory would provide the L1-EN group, previous studies addressing the language of background babble itself in speech recognition in noise have reported that participants show a higher capacity to block out background babble from a foreign language than from their L1 [17]. For the current study, this would provide a further advantage for the L1-EN group, since not only would their native cue weights help them focus on the fine phonetic differences of the vowel stimuli, but the language of the background babble makes them more impervious to informational masking.

At the same time, previous studies have shown that recognition in noise increases as a function of exposure to a language [see 13]. The logical conclusion in this setting then is that L1 speakers will generally outperform even advanced L2 speakers given the groups are matched for age, and the L1 speakers reside in a place where their L1 is the dominant language. Hence, in this case, our L1-SP speakers will have an advantage over the L1- EN group, even if the L1-EN speakers have heightened attention to different acoustic cues, due to their increased exposure to Spanish. In this paradigm, we expect competition between attentional focus and exposure.

In the following sections we present results for our on-going study in which we examined the effects of exogenous (energetic, SSN/SNR) and endogenous (informational) masking on vowel recognition in noise.

2. EXPERIMENT AND SPEECH MATERIALS

2.1. Hypotheses

We formulate the following hypotheses:

- H1: If L1 inventory size modulates recognition, we expect to see effects even for speakers with little to no knowledge of Spanish, and differences between the Spanish and English participants, recognition by L1-EN being generally better than recognition by L1-SP speakers. This is buttressed by the idea that non-native speakers can more efficiently suppress non-L1 multi-speaker background babble [6].
- H2: Past research has shown that exposure to a language enhances recognition in noise. Hence, if exposure to language enhances vowel recognition in noise, we expect a trend toward better recognition as a function of proficiency in Spanish (as proficiency is a function of exposure), and differences between L1-SP and L1-EN whereby L1-SP exhibit the highest percentage of correct responses.

2.2. Experiment design and masking conditions

The perception experiment was designed using MATLAB. Two masking conditions were programmed: background babble and the signal-to-noise ratio (SNR henceforth), which is a ratio of signal power to noise power, which is expressed in decibels (dB). Background babble was generated randomly according to differing numbers of speakers: 1, 2, 4, 6, 8, 10, 12, and 16. In addition, a speech-shaped noise (SSN) was included. The SNR was set to three [0, -6, -12] dB levels. A total of 1080 stimuli (i.e., 5 vowels × 4 repetitions × 2 voices x 9 masking types × 3 SNRs) were presented randomly to each participant.

2.3. Target stimuli

Target stimuli were recorded in a sound-proof recording booth at the Speech Laboratory of the

Universidad de Navarra. One female and one male read isolated syllables [da, de, di, do, du]. These specific syllables were used because they exist in both English and Spanish, and do not confound possible effects of VOT, as voiceless stops would.

2.4. Testing procedure

The tests were administered at the Speech Laboratories at the Universidad de Navarra and Ohio University, in sound-proof spaces. The participants heard the stimuli using Audio-Technica ATH-R70X studio headphones. Volume was set to a comfortable listening level, which could be changed following an initial trial session programmed into the MATLAB-based test. Participants were instructed that they would hear an isolated syllable [da, de, di, do, du] in different noise conditions and that their task was to listen and identify the syllable they heard. The response options appeared in text boxes on the screen and the participants were instructed to select the correct syllable by left-clicking over it. Reaction times were not registered, though this option may be interesting in future rounds of testing

2.5. Participants

Ten L1-EN (5 female/5 male) and ten L1-SP (5 female/5 males) speakers (18-35 years) were randomly selected for results (we have collected for 35 L1-EN and 35 L1-SP but we only report results here for a subset). For the L1-EN group, three had an initial level of Spanish, four had an intermediate level, and three had an advanced level. Personal data related to speech and/or auditory problems, attention deficit disorder and general cognitive capacities were collected for the participants. Additional lifestyle information was solicited to isolate possible error effects due to deficient sleep, alcohol abuse, or fatigue.

3. RESULTS

Results were by and large commensurate with our earlier pilot in that both groups showed notable confusion in recognizing the rounded back vowels [o] and [u]. It is noteworthy that 0 dB SNR produced nearly perfect vowel recognition but any errors at 0 dB SNR were almost all due to [o,u] confusions. Figures 1 and 2 illustrate this confusion in matrix form, for L1-EN and L1-SP respectively, where the y-axis displays the stimuli presented to the subject and the x-axis shows responses. Numbers in the individual cells are given in percentages. Figures 3 and 4 plot recognition accuracy as a function of SNR and number of background speakers, also for L1-EN and L1-SP respectively.

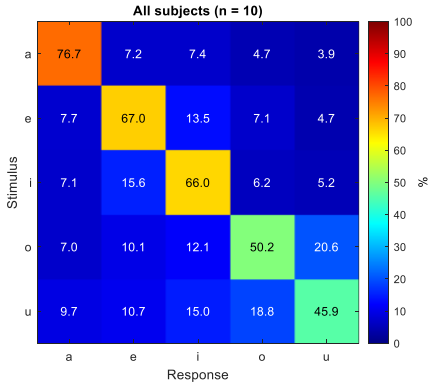


Figure 1. Confusion matrix for L1-EN. Data were pooled across all 10 L1-EN participants and across all masking conditions and SNRs.

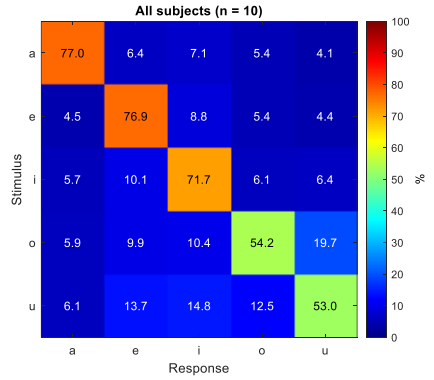


Figure 2. Confusion matrix for L1-SP. Data were pooled across all 10 L1-SP participants and across all masking conditions and SNRs.

As can be seen in the matrices, the L1-EN group showed lower ($p < 0.01$, $F = 21.67$) accuracy for [o] and [u] (50.2% accuracy for [o] and 45.9% for [u]) than the L1-SP group (54.2% accuracy for [o] and 53% for [u]).

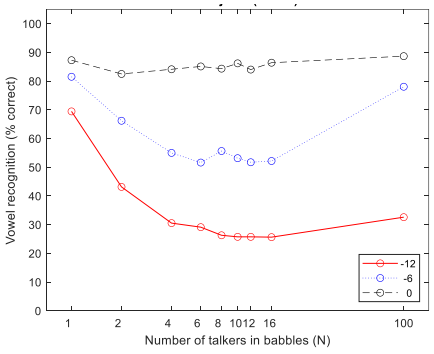


Figure 3. Vowel recognition as a function of SNR and number of background speakers for L1-EN. The results for the SSN condition were plotted arbitrarily at 100 on the abscissa.

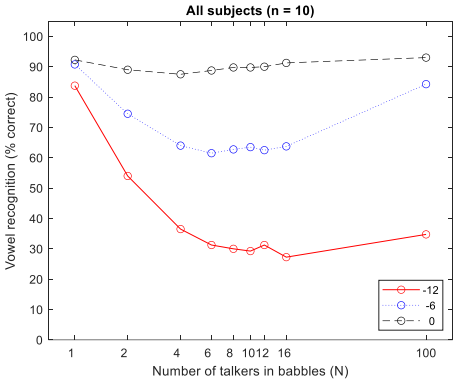


Figure 4. Vowel recognition as a function of SNR and number of background speakers for L1-SP. The results for the SSN condition were plotted arbitrarily at 100 on the abscissa.

Both groups showed relatively high release from masking at 0 dB (accuracy at between 87-92% for all numbers of background speakers), though at higher noise saturations, recognition accuracy drops off notably at around 4 speakers.

As far as responses reported as a function of the biological sex of the voice target, results were commensurate across groups. The L1-SP showed lower accuracy in recognizing vowels produced by the male talker (57.3% accuracy for male targets, 75.76% accuracy for the female voice targets). For the L1-EN group, accuracy in recognizing vowels produced by the male and female talkers was 54.9% and 67.4%, respectively. Overall accuracy was better for the L1-SP group, though significance was only reached for female voice targets. For the male voice targets, effects of group were non-significant ($p>0.05$, $F = 2.13$). There were no differences in recognition between female and male participants.

As regard accuracy scores across different levels of language proficiency, results are similar with our earlier findings in that accuracy did not improve as a function of language level past a certain point. Ceiling was reached by a L1-EN participant with a B2 level of Spanish (intermediate), at a nearly identical accuracy as for our ceiling L1-SP participant. Thus, we do not find support for the argument that L2 exposure conditions recognition, at least for the relatively small vowel inventory of Spanish.

4. CONCLUSIONS AND DISCUSSION

The results of the recognition experiments suggest notable confusion in discerning the back vowels [o] and [u], which increases with noise saturation and the number of speakers, but not as a function of L1 or language exposure. These results lead to a couple of interesting insights. First, the confusion of the back vowels in noise is curious in that [o] is the masculine marker in Spanish, meaning a vast majority of all masculine nouns and adjectives end in [o], and most native speakers do not confuse [o] and [u] (an anecdotal observation). However, it cannot go unsaid that, across the Romance languages, alternations of [o] and [u] are quite common, even for dialects of Spanish [18], between Portuguese and Spanish [19] and Romanian [20], in addition to other non-Romance languages such as German [21].

As for the specific reason listeners may confuse [o] and [u], we surmise two scenarios. First, the confusion of [o] with [u] (and vice versa) may be due to lexical effects. The stimuli [da], [de] and [di] are all words in Spanish, while [do] and [du] are not. There is a possibility that this imbalance in the stimuli is skewing responses. However, if that were the case, we should expect to see differences across language and proficiency groups since this would only influence native and advanced speakers of Spanish, which we did not find evidence for in our results. Returning to the idea that L1 vowel inventory size may suppose an advantage for the L1-EN group, our results here are inconsistent with the notion that inventory size supposes any advantage in recognizing non-native vowel contrasts. However, results for the L1-SP group are not so much better than for the L1-EN group that we can rule out competing forces that play out while recognizing phonological contrasts in noise

Nevertheless, another possibility to account for the [o]-[u] confusion could be that tongue height (which is what the listeners are finding difficult to discern), as expressed acoustically as F1, is obfuscated by the rounding of the lips (expressed by F3). In normal speech conditions, a speaker not only has access to the auditory signal, but also has access to visual information that is used for information gain (i.e., reducing entropy or uncertainty). This means that in noisy conditions, the extra visual input may aid the listener in discerning the back vowels. This scenario finds support in a set of machine-learning models that were created [22] where it was shown that the addition of a visual variable (maximum lip aperture, obtained through motion capture images of the lips) does aid in classification of the vowels in the face of a noise source.

5. ACKNOWLEDGMENTS

This work was financed by a grant [ref. PID2019-105929GA-I00] by the Ministry of Science and Innovation (Spain).

7. REFERENCES

[1] Parikh, G., Loizou, P. 2005. The influence of noise on vowel and consonant cues. *Journal of the Acoustical Society of America*, 118, 6, 3874–3888.

[2] Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 623–656,

[3] Pollack, I. 1975. Auditory informational masking. *Journal of the Acoustical Society of America,* 57, S5.

[4] Summers, R., Roberts, B. 2020. Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *The Journal of the Acoustical Society of America* 147(2), 1113-1125.

[5] Freyman, R. L., Balakrishnan, U., Helfer, K.S. 2001. Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America*, 109, 2112–2122.

[6] Arbogast, T.L., Mason, C.R.., Kidd, G. 2002. The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America*, 112, 2086–2098.

[7] Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott, K.R. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110, 2527–2538.

[8] Kidd, G. Jr., Mason, C.R., Swaminathan., J., Roverud, E., Clayton, K.., Best, V. 2016. Determining the energetic and informational components on speech-on-speech masking. *Journal of the Acoustical Society of America*, 140, 132–144.

[9] Liu, C., Kewley-Port, D. 2004. Formant recognition in noise for isolated vowels. *Journal of the Acoustical Society of America*, 116, 5, 3119–3129.

[10] Wang, X., Xu, L. 2020. Mandarin tone perception in multiple-talker babbles and speech-shaped noise. *Journal of the Acoustical Society of America*, 147, 4, EL307-EL313.

[11] Wang, X., & Xu, L. 2021. [Speech perception in noise: Masking and unmasking]. Journal of Otology, 16, 2, 109-119, DOI: 10.1016/j.joto.2020.12.001.

[12] Yuan, Y., Lleo, Y., Daniel, R., White, A., Oh, Y. 2021. The impact of temporarily coherent visual cues on speech perception in complex auditory environments. *Frontiers in Neuroscience*, 15, 1–7

[13] Li, M., Wang, W., Tao, S., Dong, Q., Guan, J., Liu, C. 2016. Mandarin Chinese vowel-plus-tone identification in noise: Effects of language experience. *Hearing Research*, 331, 109–118.

[14] Calandruccio, L., Dhar, S., Bradlow, A.R. 2010. Speech-on-speech masking with variable access to the linguistic content of the masker speech. *Journal of the Acoustical Society of America*, 128, 860–869.

[15] Brouwer, S., Van Engen, K. J., Calandruccio, L. Bradlow, A. R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *Journal of the Acoustical Society of America*, 131, 1449–1464.

[16] Toscano, J.C., McMurray, B. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34, 3, 434–464.

[17] Van Engen, K.J., Bradlow, A.R. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America*, 121, 1, 519–526.

[18] Hualde, J. I. 2014. Los sonidos del español. Cambridge, U.K: Cambridge University Press.

[19] Mateus, M. H. & de Andrade, E. 2000. *The Phonology of Portuguese.* Oxford, U.K: Oxford University Press.

[20] Renwick, M. E. L. 2012. Vowels of Romanian: Historical, Phonological and Phonetic Studies. Phd dissertation, Cornell University.

[21] Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R. & Fleischer, M. 2020. Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. *Scientific Data* 7, 255.

[22] Gibson, M., González, M. & Schlechtweg, M. (accepted). A machine learning model to assess the integration of visual and auditory cues during speech perception in noise. *Proceedings of Forum Acusticum* 23, 10 Conference of the European Acoustics Associaion.

PERCEPTUAL ASSIMILATION OF SPANISH, THAI, AND KOREAN STOP CONSONANTS BY NATIVE ENGLISH SPEAKERS

Charlie Nagle[1], Shelby Bruun[1], Melissa Baese-Berk[2]

The University of Texas at Austin[1], University of Oregon[2]
cnagle@austin.utexas.edu, sbruun@utexas.edu, mbaesebe@uoregon.edu

ABSTRACT

Models of second language (L2) sound learning argue that learners' initial perceptual assimilation patterns are important because they shape perceptual learning pathways and outcomes. To contribute to this line of literature, we examined native English listeners' perceptual assimilation of Spanish, Thai, and Korean stop consonants, asking participants to map stops from all three languages onto English stop categories and rate their goodness-of-fit using a 5-point scale.

Results showed that several categories were unequivocally assimilated to a single English category (e.g., Korean Aspirated and Lenis stops to English voiceless stops). In contrast, the Spanish Voiceless category, Thai Plain category, and Korean Fortis category were mapped onto both English voiced and voiceless stops. Yet, this group trend masks substantial individual variation in the assimilation of these ambiguous categories. Some participants showed split categorization, aligning with the group, whereas others showed a strong preference for one of the English categories.

Keywords: second language, speech perception, perceptual assimilation, stop consonants, individual differences

1. INTRODUCTION

During the first year of life, infants can easily perceive differences between sounds, including sounds that do not occur in their native language (L1) [1]. However, as L1 categories coalesce, the ability to distinguish between non-native (L2) sound contrasts rapidly diminishes. L1 speech learning "warps" perceptual space, such that the L1 acts as a perceptual filter for the L2 [2]. The individual no longer perceives acoustic reality but rather their perception is pushed and pulled by L1 categories.

Models of L2 speech sound learning such as the Speech Learning Model [3, 4] and the Second Language Perceptual Assimilation Model [5] formalize how the L1 and L2 should interact during learning and perception. For example, the Speech Learning Model argues that the timing of L2 learning relative to L1 learning is critical. For individuals who learn the L2 later in life, by the time L2 learning begins the L1 system is already firmly in place, which means that it will exert a strong influence on the L2. Likewise, according to the L2 Perceptual Assimilation Model [5], initial perceptual assimilation patterns determine the ease or difficulty of the learning task. For instance, if both members of an L2 contrast are assimilated to a single L1 category and judged to be relatively equally good exemplars of that category, then learning to perceive the L2 contrast could be very challenging. Despite their differences, the Speech Learning Model and the L2 Perceptual Assimilation Model share the assumption that crosslinguistic assimilation patterns matter, insofar as they may dictate both the starting point and the rate of L2 perceptual learning.

Additionally, there is likely to be substantial individual variation in these patterns, even among listeners who come from the same L1 background. The L1 sets the boundaries on what is possible in terms of potential L2 assimilation patterns, but within those general, language-specific guardrails, individual listeners appear to exhibit a range of meaningful subpatterns [6]. Such findings mirror the extensive individual variability documented in other areas of perceptual performance and learning, including individual differences in phonetic cue weights [7, 8].

Stop consonants are one of the most studied L2 segments. Surprisingly, however, there have not been many perceptual assimilation studies targeting L2 stops. This may be because stop consonant systems have a smaller set of categories than vowel systems, where multicategory mappings (i.e., many-to-one or many-to-many) are common (e.g., [6]). In other words, most languages have two to three stop consonant categories at a given place of articulation, which places an inherent limit on the types of crosslinguistic patterns that may emerge. Yet, even if patterns are relatively fixed for L2 stops, it is important to understand how strong those patterns are, that is, how well the L2 sound fits into the L1 category, which could have an impact on learning new L2 stop categories.

The few studies that have been carried out on the perceptual assimilation of word-initial stops show that listeners tend to respond based on the phonetic cues present in their L1. For instance, Mandarin [9], Japanese [10], and Thai listeners [11] assimilate

Korean stops to the L1 category that is the best match in terms of aspiration, or the amount of voice onset time (VOT) present in the stop, presumably because VOT is the primary cue to stop consonant contrasts in each language. Thus, these listeners tend to show a single category assimilation for Korean Lenis and Aspirated stops, which are similar in terms of VOT but differ with respect to fundamental frequency (F0), while making a distinction between those categories and Korean fortis stops, which are produced with shorter VOT. Oliveira and Rato [12] reported similar findings for Cantonese listeners' perception of European Portuguese stops. It bears mentioning that these trends may mask substantial variability in individual response patterns [10].

Further, current research has focused on the perceptual assimilation of stops from one additional language by listeners who are unfamiliar with that language. Investigating how individuals assimilate sounds from several languages would shed light on the relative difficulty listeners might experience when learning those languages. It would also be interesting to examine individual patterns of variation to document the range of assimilation patterns and the strength of those patterns within languages and within listeners.

In this study we therefore examined English listeners' perceptual assimilation of word-initial stops in Spanish, Thai, and Korean. Briefly stated, English and Spanish have two phonological stop consonant categories, whereas Thai and Korean have three. In all languages, VOT is a relevant phonetic cue to stop consonant identity, and it serves as the primary cue in English, Spanish, and Thai [13, 14]. In Korean, however, both VOT and F0 are essential cues to the three-way contrast, which cannot be fully differentiated by either cue independent of the other [8, 15]. Importantly, although English speakers are sensitive to F0 differences, they primarily rely on VOT, which is a more robust cue to stop consonant voicing in English [16].

2. METHOD

We created the experiment using Gorilla Experiment Builder (www.gorilla.sc) and integrated it with Prolific (www.prolific.co) for participant recruitment. Access was restricted to desktop devices located in the United States.

2.1. Participants

Twenty-eight monolingual speakers of English (13 females, 15 males) completed the study. Mean participant age was 35.39 *(SD* = 11.82). 89% (*n* = 25) of participants identified as White, 4% (*n* = 1) as Black, and 7% (*n* = 2) as Mixed or Other. All participants were born and raised in the United States with one participant reporting a period of residency in the United Kingdom. All reported normal hearing.

2.2. Tasks

Participants completed a background questionnaire targeting known languages and language use. None reported proficiency in the target languages.

Participants then completed a forced-choice perceptual assimilation task including 320 items, described below. The instructions made it clear that all stimuli would begin with either a "p" sound or a "b" sound. Participants were asked to indicate the sound they heard by selecting one of two choices: "Sounds like English 'p'" or "Sounds like English 'b'". Goodness ratings were elicited following each trial on a 5-point scale where only the endpoints were labelled (1 = "bad" and 5 = "good"). The task was completed in three blocks (2 blocks of 107 trials, 1 block of 106 trials) to allow for two two-minute rest periods between blocks. Tokens from each language were intermixed across the task and presentation order was randomized across participants.

2.2.1 Perceptual Assimilation Stimuli

We used a subset of stimuli from a larger stimulus set containing Thai, Korean, and Spanish words and nonsense words. All nonsense words obeyed the phonotactic constraints of their respective language. To validate the stimuli, we asked native speakers of each language to evaluate how good of an example each stimulus was of the intended word on a 4-point scale (1 = "poor", 4 = "excellent"). All stimuli used in this experiment were rated 3 or higher.

In this study, there were eight stop consonant categories: two for Spanish and three each for Thai and Korean. Within each language, the stimuli were produced by four talkers, two males and two females, and the initial stops were combined with five following vowels for two repetitions (4 talkers × 5 vowels × 2 repetitions = 40 stimuli per category). Thus, there were 320 total stimuli.

2.3.1. Acoustic characteristics of stimuli

Stimuli were coded for VOT and F0 using Praat version 6.1.53 and values were extracted using a script. Positive VOT was measured from the release of the stop consonant burst to the onset of voicing of the following vowel. Negative VOT was measured from the onset of vocal fold vibration, evident as low-frequency periodicity in the waveform, to the burst of the stop consonant. Ten percent of the data (50/320

tokens) was cross-coded to ensure intercoder reliability. Following Schertz et al. [8], F0 was measured 5 ms after the onset of the following vowel. Table 1 summarizes the acoustic characteristics of the stops in each language.

	VOT	F0
	M (*SD*)	*M* (*SD*)
Spanish		
Voiced	–104.08 (33.31)	155.45 (16.99)
		139.05 (11.59)
Voiceless	22.68 (9.06)	182.35 (31.27)
		160.75 (28.78)
Thai		
Voiced	–77.08 (26.82)	244.25 (15.55)
		134.95 (5.52)
Plain	11.38 (4.29)	267.85 (14.54)
		155.90 (23.02)
Aspirated	91.26 (19.13)	271.50 (18.65)
		176.61 (30.02)
Korean		
Fortis	13.05 (5.23)	288.15 (31.78)
		150.50 (13.03)
Lenis	53.08 (15.67)	245.05 (14.00)
		124.95 (8.91)
Aspirated	70.72 (17.99)	326.10 (29.51)
		163.15 (11.22)

Table 1: Acoustic characteristics of the stimuli. F0 is shown separately for female (top) and male (bottom) talkers.

3. RESULTS

For each language, we examined the proportion of English voiced and voiceless stop responses per category and the corresponding category goodness rating. We computed proportions and mean goodness ratings for all participants to shed light on group trends. We also examined individual categorization patterns.

As shown in Figure 1, in Korean, the Aspirated and Lenis categories were mapped almost exclusively onto the English voiceless category and rated as good exemplars of that category. In contrast, the Fortis category was variably mapped onto both voiced and voiceless stops and received category goodness ratings that were highly variable. In Thai, the Aspirated category was mapped nearly exclusively onto English voiceless stops, whereas the Voiced category was mapped nearly exclusively onto English voiced stops, and both stops were deemed to be good exemplars of their English counterpart. Perception of the Plain category was more variable, though participants showed a slight preference for English voiced stops. Finally, for Spanish stops, the Voiced category was mapped onto English voiced stops, but assimilation of the Voiceless category was variable. In this case, contrary to the pattern evident in Thai, Spanish Voiceless stops were generally mapped onto English voiceless stops.

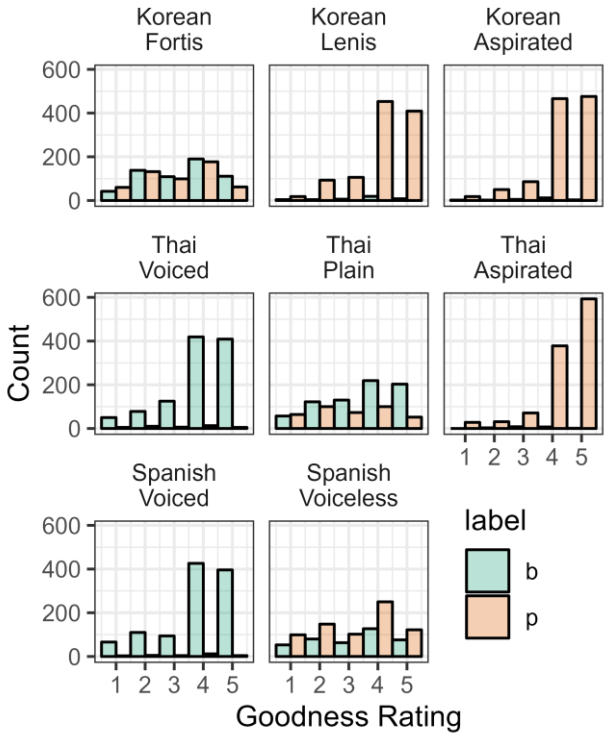


Figure 1: Assimilation patterns and goodness ratings for Korean, Thai, and Spanish stops.

We were particularly interested in examining assimilation patterns and goodness ratings for the categories that showed the most variability, namely Korean Fortis, Thai Plain, and Spanish Voiceless stops. To do so, we computed difference scores for each participant by subtracting the percentage of /b/ responses from the percentage of /p/ responses. Thus, in Figure 2 positive values index a preference for /p/, negative values a preference for /b/, and a value of zero no preference, or equal proportions of /b/ and /p/ responses. The plot reveals a wide range of patterns with respect to the direction of assimilation in each language. For instance, while nearly all individuals assimilated the Spanish Voiceless category to English /p/, for some participants this assimilation was relatively weak, near the zero mark, whereas for others it was quite strong at 75%. The opposite pattern is evident for Thai Plain stops, which most participants mapped onto English /b/ to varying degrees. This divergent pattern could be due to the slight difference in VOT across the languages, where Thai Plain stops were produced with shorter VOT than Spanish Voiceless stops (10 ms on average). Korean Fortis stops showed a relatively split pattern, with some participants mapping them onto English /b/ and others onto English /p/, and the difference scores

suggest a narrower range than what was observed for the other two languages. This narrower range seems to index greater ambiguity in terms of how the Korean Fortis category maps onto English stop categories, which could be due to a combination of VOT intermediate to the Spanish Voiceless and Thai Plain categories and higher F0, which would be associated with voicelessness in English.

Figure 2 also provides a visual summary of how distinctly each participant categorized the three ambiguous categories. Individuals on the extreme left and extreme right of the plot assimilated all three categories to the same English category (/b/ on the left and /p/ on the right), and for the rightmost participant on the plot, proportions were nearly identical across languages. Individuals shown in the middle portion of the plot tended to map Thai Plain stops onto /b/ and Spanish Voiceless stops onto /p/, showing the greatest uncertainty (i.e., values closest to zero) for Korean Fortis stops. Yet, even among these individuals the proportion of stops mapped onto one category varied considerably.

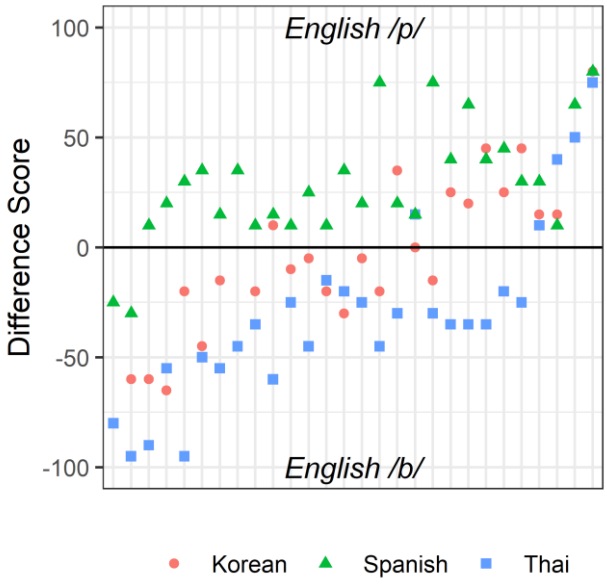


Figure 2: Individual assimilation patterns for Korean Fortis, Spanish Voiceless, and Thai Plain stops.

4. DISCUSSION

In this study, we examined perceptual assimilation of voicing contrasts from three languages by native English listeners. Each language showed slightly different assimilation patterns. Below, we review these assimilation patterns and the individual variability within each language and category type.

We found evidence for a single category assimilation for Korean Aspirated and Lenis stops, both of which were perceived as good exemplars of English voiceless stops. This finding mirrors the results of previous studies involving participants whose L1 predominantly marks stop consonant contrasts using VOT [9, 10, 11]. In other words, because those two Korean categories are both produced with a considerable amount of VOT, they are mapped onto the L1 category that is the closest phonetic match, which for English listeners is the voiceless category. In contrast, Korean Fortis stops were variably mapped onto English voiced and voiceless stops. The individual data revealed that most listeners mapped Korean Fortis stops onto English voiced stops, but there were several who mapped them onto voiceless stops. This means that for some listeners, all three Korean stops were mapped onto a single English category, albeit with varying degrees of fit, but the Lenis and Aspirated categories were always rated as better exemplars of English voicelessness than the Fortis category was.

For Thai, Voiced stops were mapped onto English voiced stops and Aspirated stops onto English voiceless stops, which fits with the distribution of VOT values in both languages. The Thai Plain category, however, was more ambiguous, insofar as it was mapped onto both English categories. Yet, most listeners perceived Thai Plain stops as a better exemplar of English voiced stops, at least in terms of response proportions. Thus, there seemed to be less ambiguity in the classification of Thai Plain stops than there was in the classification of Korean Fortis stops. In fact, some individuals classified Thai Plain stops as instances of English voiced stops nearly categorically. This also fits with the phonetic reality of English voiced stops, which can be realized either as fully (pre)voiced stops or as short-lag stops.

Finally, Spanish Voiced stops were mapped onto English voiced stops and rated as good exemplars of the English category, but the Spanish Voiceless category was more ambiguous. Most participants assimilated Spanish Voiceless stops to English voiceless stops, but the fit was variable.

Overall, these results provide baseline data for English listeners' perceptual assimilation of stops in three unfamiliar languages. They also highlight the importance of going beyond group trends to look at individual assimilation patterns [6]. The English listeners recruited for this study were mostly monolingual, reporting little to no L2 learning experience with the languages included in this study (or any other), and therefore represented a relatively linguistically homogenous participant sample. Yet, even these individuals showed varying perceptual assimilation patterns that could lead to different learning outcomes. Thus, based on these results, it is important to categorize individuals based on individual performance rather than group data.

5. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. 2117664, 2117665, and 2309561.

6. REFERENCES

[1] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., Iverson, P. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9(2), F13–F21.

[2] Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., Nelson, T. 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B* 363, 979–1000.

[3] Flege, J. E. 1995. Second language speech learning: Theory, findings, problems. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Langauge Research*. York Press, 233–277.

[4] Flege, J. E., Bohn, O.-S. 2021. The revised Speech Learning Model. In: Wayland, R. (ed), *Second Langauge Speech Learning: Theoretical and Empirical Progress*. Cambridge University Press, 3–83.

[5] Best, C. T., Tyler, M. D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Munro, M. J., Bohn, O.-S. (eds), *Second Langauge Speech Learning: The Role of Language Experience in Speech Perception and Production*. John Benjamins, 13–24.

[6] Mayr, R., Escudero, P. 2010. Explaining individual variation in L2 perception: Rounded vowels in English learners of German. *Bilingualism: Language and Cognition* 13(3), 279–297.

[7] Kim, D., Clayards, M., Goad, H. 2018. A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics* 67, 1–20.

[8] Schertz, J., Cho, T., Lotto, A., Warner, N. 2015. Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics* 52, 183–204.

[9] Holliday, J. J. 2014. The perceptual assimilation of Korean obstruents by native Mandarin Listeners. *The Journal of the Acoustical Society of America* 135, 1585.

[10] Holliday, J. J. 2019. The perception and production of word-initial Korean obstruents by native speakers of Japanese. *Language and Speech* 62(3), 494–508.

[11] Wayland, R. 2000. Perceptual assimilation and categorical discrimination of Korean stop consonants by native Thai speakers. *The Journal of the Acoustical Society of America* 108, 2652.

[12] Oliveira, D., Rato, A. 2018. Assimilação percetiva das oclusivas orais do português europeu L2 por falantes nativos de cantonês. *Diacrítica* 32(2), 133–156.

[13] Lisker, L., Abramson, A. 1964. A cross-language study of voicing in stops: Acoustical measurements. *Word* 20, 384–422.

[14] Raphael, L. 2021. Acoustic cues to the perception of segmental phonemes. In: Pardo, J. S., Nygaard, L. C., Remez, R. E., Pisoni, D. B. (eds), *The Handbook of Speech Perception*. Wiley Blackwell, 603–631.

[15] Cho, T., Jun, S.-A., Ladefoged, P. 2002. Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics* 30, 193–228.

[16] Shultz, A. A., Francis, A. L., Llanos, F. 2012. Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America* 132, EL95–101.

Trochaic foot benefit in speech comprehension

Mathias Scharinger[1,3,4]
Ulrike Domahs[2,3,4]

[1]Research Group »Phonetics«, Institute of German Linguistics, Philipps-Universität Marburg, Germany
[2]Research Group »Neurolinguistics«, Institute of German Linguistics, Philipps-Universität Marburg, Germany
[3]Research Center »Deutscher Sprachatlas«, Philipps-Universität Marburg, Germany
[4]Center for Mind, Brain & Behaviour »CMBB«, Universities of Gießen & Marburg, Germany

emails: mathias.scharinger@uni-marburg.de; ulrike.domahs@uni-marburg.de

ABSTRACT

There is evidence that speech comprehension benefits from prosodic regularities, possibly through segmentation of continuous acoustic signals into foot structure. We probed this assumption in a modified Sternberg memory task, using 8-digit-numbers with either trochaic or iambic grouping obtained by prosodic manipulations. A visual 4-digit probe number was used to test recall performance. The probe was either congruent or incongruent with the foot structure of the 8-digit-numbers. Recall performance was measured by accuracy and reaction times. Accuracy was significantly better for probes congruent with the foot structure, evidencing that feet are indeed used during the encoding of speech. Reaction times for correct recall were faster for probes with congruent foot structure and in the trochaic condition. Together, these findings provide evidence for the role of feet during speech comprehension in a memory task during speech comprehension, and for a trochaic bias in German.

Keywords: speech comprehension, prosody, foot structure, trochee

1. INTRODUCTION

There is mounting evidence that speech comprehension involves a segmentation mechanism describable by oscillator models [1-3]. The main assumption is that the continuous speech signal is segmented into temporally discrete units that are optimal for working-memory processing [1, 4]. The sizes of these units roughly correspond to the syllable (150-300 ms) and to short phrases (500-2000 ms). Correspondingly, the neural oscillations underlying the segmentation process during speech perception are most prominently in the theta- (3-7 Hz) and in the delta-band (0.5-2 Hz) [5, 6]. In this tradition, Ghitza [1] has focused on delta-oscillations, suggesting their particular role for prosodic segmentation. In his study, he used a modified Sternberg memory task wherein 10-digit numbers, grouped either regularly (2-2-2-2-2 pattern) or irregularly (3-3-2-2 pattern), had to be memorized. Participants were auditorily presented with these number strings, followed by a 2- or 3-digit probe. Probes could be either inside a chunk (i.e., **589** from the string 334 **589** 33 22, grouping congruency), be split between two successive chunks (i.e., **933** from the string 334 58**9** **33** 22, grouping incongruency) or not at all be contained in the string (i.e., **555**). The results of Ghitza's study demonstrated an effect of grouping congruency: Error rates significantly increased if probes were incongruent to the grouping pattern of the 10-digit strings. Furthermore, error rates also increased for grouping patterns whose cycle durations deviated from the delta-oscillation (>3 Hz).

While these results have provided a mechanistic explanation for the importance of delta-oscillations in speech segmentation, the precise functional relevance of the particular segmentation (or grouping) has rather been left unexplored. As delta-oscillations correspond to temporal windows which may comprise of up to two syllables, a possible functional role of delta could be the tracking of binary foot structure in speech comprehension. This assumption will be tested in the below-described experiment, attempting to answer the research question of whether delta-oscillations provide a foot-based segmentation of continuous speech, capitalizing on prosodic regularities and perhaps the trochaic bias in German [7, 8].

2. CURRENT STUDY

The rationale of this study is to explore the influence of prosodic foot structure on speech comprehension. Following Ghitza [1], we assume that the encoding of continuous speech information benefits from prosodic regularities, and in particular, foot structure. We hypothesize that strong-weak (trochaic) and weak-strong (iambic) prosodic patterns drive delta oscillations. Assumedly, segmenting the speech signal into delta-sized units (i.e., between 500-2000

ms) provides discrete information for optimized working-memory processing. We thus also use a modified Sternberg memory task, with the following important differences compared to Ghitza's study [1]: (1) 8-digit number strings, regular grouping (2-2-2-2); (2) prosodic manipulations to either conform to a trochaic (strong-weak) or an iambic (weak-strong) pattern; (3) cross-modal design, with auditorily presented 8-digit number strings and visually presented 4-digit probe strings. Using a cross-modal design, we hope to avoid that congruency effects stem merely from acoustic matching between the large string and the probe.

2.1. Methods

2.1.1. Stimuli

Stimulus construction was based on naturally spoken single numbers of German. A female native speaker of German pronounced the numbers from 0 to 9 in several renditions, which were recorded on a Windows PC with 44.1 kHz sampling rate and 16-bit amplitude resolution. The best token per number (based on comparability in intensity and pitch) was selected for further stimulus construction. The two-syllable number "7" was realized as monosyllable ([zi:m]). Subsequently, all natural tokens were lengthened to 250 ms using the overlap-add algorithm in PRAAT [9]. A silence of 50 ms was added to all renditions, resulting in 300 ms recordings of each number.

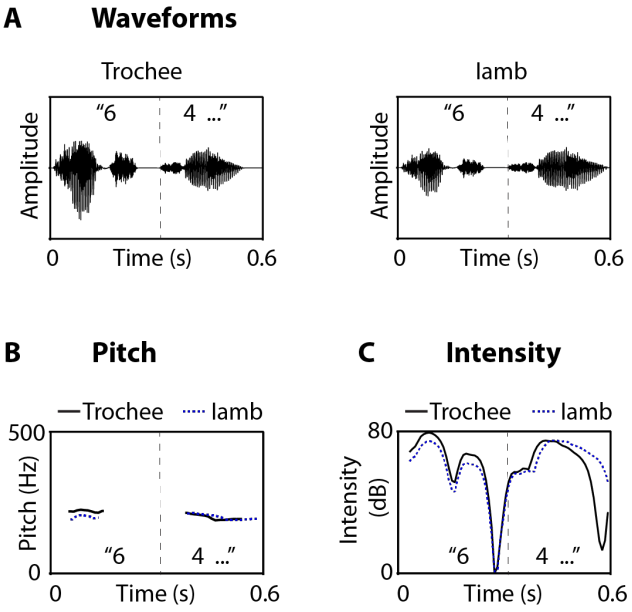


Figure 1: Illustration of stimulus construction. Naturally-produced single digits were modified to either elicit a strong-weak (trochee) or weak-strong (iamb) grouping (A). In odd-numbered positions in the 8-digit strings, pitch and intensity were enhanced (pitch: 20 Hz, intensity: 5 dB), yielding the strong-weak (trochaic) pattern. In even-numbered positions in the 8-digit strings, duration was enhanced (from 250 to 300 ms), yielding the weak-strong (iambic) pattern (B, C).

The 8-digit strings were then constructed by concatenating a random selection of the digits between 0 and 9 (without replacement). Strong-weak (trochaic) and weak-strong (iambic) prosodic patterns were achieved based on the insights of [8] (see Figure 1). In total, 100 8-digit strings with a trochaic and 100 8-digit strings with an iambic pattern were constructed. For half of the trochaic and iambic strings, a 4-digit number was selected as probe, being a proper subset (e.g., **4183** from 64183297). 50% of these probes were congruent with the foot structure of the 8-digit string, and 50% were not congruent (as in the above-example, corresponding to the split-chunk-condition of Ghitza [1]). For the other half of the trochaic and iambic strings, the 4-digit probe was not a proper subset (e.g., **4138** from 64183297). Note that the first digit pair of these probes was always contained in the larger string, congruent or incongruent with the foot structure, while the second digit pair provided the mismatching information. Thereby, a fully-crossed 2x2x2 design (foot: trochee/iamb; congruency: congruent/incongruent; match: yes/no) could be achieved.

2.1.2. Design

The 2x2x2 design (see Figure 2) determined the Sternberg memory task in which an auditory 8-digit string was followed by a visual probe that could be either contained in the larger string (match: yes) or not (match: no). In half of these cases, the probe was congruent or incongruent with the foot structure (grouping) of the larger string.

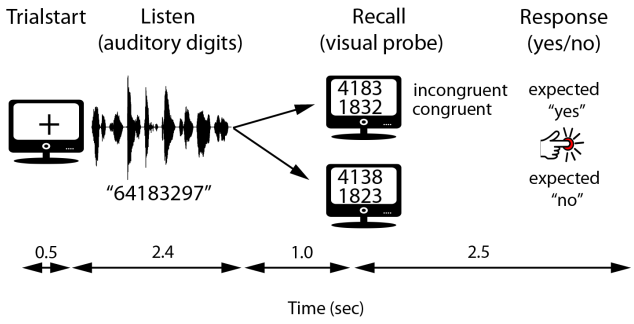


Figure 2: Trial structure. A trial started with a fixation cross displayed for 0.5 sec. The auditory 8-digit string was presented subsequently. After 1-sec pause, the probe was displayed on the screen. Participants had then to respond "yes" if the probe was contained in the string and "no" if it was not.

2.4. Participants

Participants (N=20) were native speakers of German (mean age: 27 yrs, SD: 9 yrs., 14 females, 5 males 1 diverse). They were right-handed and had normal hearing. Participants provided written informed consent before the experiment started. The study was approved by the Ethics Committee of the German Society for Linguistics (DGfS) and in accordance to the declarations of Helsinki. Participants received monetary compensation for their participation.

2.5. Procedure

The memory task was conducted in a shielded chamber because participants' EEG was recorded in addition to their behavioural responses (data presented elsewhere). Participants were seated approximately 2 m in front of a computer screen (27-inch). A trial started with a fixation cross displayed in the middle of the screen for 0.5 secs. Subsequently, participants heard the 8-digit strings over loudspeakers (2.4 sec). After a pause of 1 sec, the probes were displayed in the middle of the screen and participants had 2.5 secs to give their responses on a computer keyboard, using the keys "f" and "j". Response-to-key attribution to indicate a probe match- or mis-match was counter-balanced across participants. Each participant received a different randomization of the 200 trials. Trials were presented with the open-source software OpenSesame [10]. The experiment lasted for about 30 minutes.

2.7. Statistics

Reaction times (starting from the onset of probe presentation) were used as dependent variable in a Linear Mixed Effects (LME) model, with *subject* (participant ID) and *item* (Trial) as random effects (intercept only). The fixed effects comprised *match* (yes/no), *foot* (trochee/iamb) and *congruency* (congruent/incongruent). Accuracy was analysed as logistic dependent variable (0: incorrect, 1: correct detection) in a Generalized Mixed Effects model with the same effect structure as in the reaction time model. Additionally, participants' response sensitivity was modelled with *d'* in a Linear Mixed Effects model with *subject* (participant ID) as random effect (intercept only) and the fixed effects *foot* and *congruency*. Statistical analyses were calculated in jamovi [11].

3. RESULTS

The LME model on reaction times showed main effects of *match, foot* and *congruency*. Reaction times were faster for mismatching compared to matching probes (t=-3.08, p<0.01), for trochaic compared to iambic digit strings (t=-3.98, p<0.001) and for congruent compared to incongruent probes (t=-9.2, p<0.001). The interaction of the effects *match* and *congruency* was based on larger differences between congruent and incongruent probes when the probe was contained in the larger string than when it was not.

The GME model on accuracy revealed main effects for *match* and *congruency*. Probes not contained in the larger digit strings and probes congruent with the foot structure of these larger strings were responded to more accurately (z=19.8, p<0.001; t=6.75, p<0.001).

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Foot	15.87853	1	3916	<.001
Match	9.48576	1	3957	0.002
Congruency	84.57592	1	3950	<.001
Foot ✻ Match	0.00602	1	3951	0.938
Foot ✻ Congruency	0.08743	1	3952	0.767
Match ✻ Congruency	4.94331	1	3942	0.026
Foot ✻ Match ✻ Congruency	0.14500	1	3943	0.703

Note. Satterthwaite method for degrees of freedom

Table 1: Results of the LME model on reaction times.

The interaction of the effects *match* and *congruency* was based on larger accuracy differences between incongruent and congruent probes when the probes were contained in the larger strings than when they were not. The three-way interaction of *match*, *congruency* and *foot* resulted from a stronger congruency effect in matching probes when the large digit strings had trochaic compared to iambic foot structure. Accuracy dropped almost to chance level when probes were contained in the trochaic 8-digit strings but were incongruent with the foot structure.

Fixed Effect Omnibus tests

	X^2	df	p
Congruency	45.556	1.00	<.001
Foot	1.675	1.00	0.196
Match	393.626	1.00	<.001
Congruency ✻ Foot	0.113	1.00	0.737
Congruency ✻ Match	22.590	1.00	<.001
Foot ✻ Match	4.41e-4	1.00	0.983
Congruency ✻ Foot ✻ Match	11.226	1.00	<.001

Table 2: Results of the GME model on accuracy.

The LME model on *d'* showed a main effect of congruency ($F(1,57)=38.13$, $p<0.001$), based on higher *d'* for congruent than for incongruent probes ($t=6.17$, $p<0.001$, see Figure 3).

4. DISCUSSION

The results of the memory task suggest that the prosodically induced grouping of numbers is beneficial for memory encoding, particularly if the grouping adheres to a trochaic (strong-weak) pattern. Furthermore, the congruency effect, replicating the "split-chunk" finding of Ghitza [1], provides evidence that foot structure plays an important role during encoding and/or retrieval from memory.

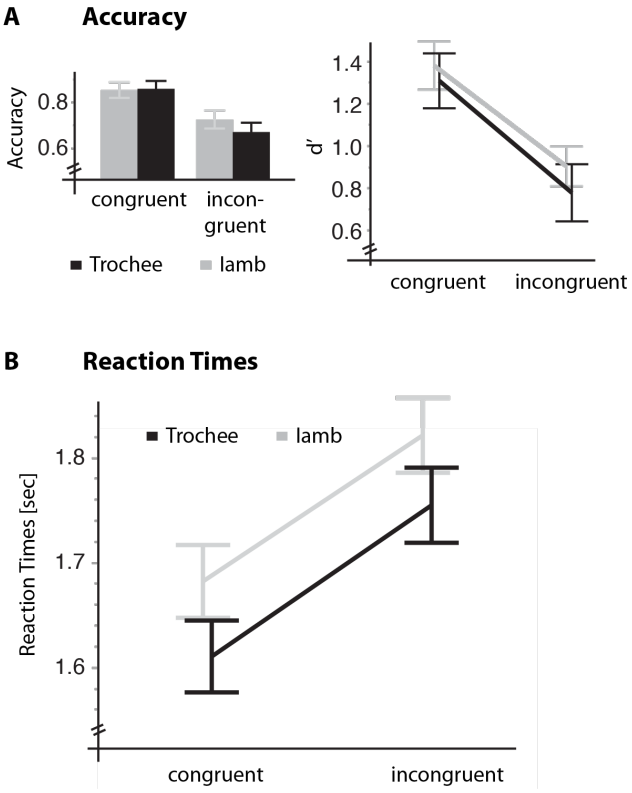


Figure 3: Summary of results. A. (left): Accuracy (proportion correct responses), pooled across matching/mismatching probes. (right) The *d'* analysis corroborates the accuracy pattern, with higher *d'* for congruent than incongruent probes. B. Reaction times differed between trochaic and iambic digit-string patterns and between congruent and incongruent probes, with probes congruent to trochaic digit-strings being responded to fastest. Error-bars/whiskers indicate standard errors of the means.

Considering the experimental memory task as a proxy to speech comprehension, the aforementioned research question can be answered as follows.

In addition to Ghitza's findings [1], the present study's results suggest that the encoding of continuous speech information benefits from prosodic regularities in terms of foot structure. Segmenting the speech signal into delta-sized units (i.e., between 500-2000 ms) seems to provide chunks for optimized working-memory processing, supporting [12] who showed that delta-sized units are more related to meaningful utterances compared to theta-sized units that relate to syllable units. In order to test whether the type of foot structure plays a role for the memory task, we compared the performance for trochees and iambs that varied with regard to the position of the prosodic prominence as well as the acoustic properties establishing prominence, following the iambic/trochaic law [13]. As can be seen in Figure 3B, reaction times were significantly shorter for trochaic patterns, irrespective of the congruency of the 4-digit probe with the 8-digit sequence. According to the oscillation models, we would have expected that both disyllabic structures drive delta oscillations and facilitate the chunking of all congruous sequences and their memorability. The significant advantage of trochees over iambs, however, speaks in favour of strong-weak segmentation strategies and sensitivity to specific prosodic cues (pitch and intensity) as has previously been observed in first language acquisition [14] and adult processing [15] (but see [7] for iambic and trochaic grouping preference in German monolinguals and French-German bilinguals). The latter study suggests that future research has to include languages with iambic preferences or languages in which word stress patterns are less important than in German or English, as is the case in French, for instance. The inclusion of such languages enables us to test whether working memory abilities can be improved by chunks of the shape of any language-specific prosodic pattern or – for languages without any word stress pattern – does not play a role in the recall of sequences at all. Moreover, future research has to also speak to the (alleged) universality of the Iambic/Trochaic law [8, 15-17] as well as to (possibly) universal constraints on working memory [18, 19].

5. CONCLUSION

Taken together, the congruency effect for trochees shows that delta-sized units are only advantageous for segmentation and processing if the processing units exhibit initial prominence.

6. REFERENCES

[1] O. Ghitza, 2017. Acoustic-driven delta rhythms as prosodic markers, *Lang Cogn Neurosci,* 32, 545-561

[2] L. Henke and L. Meyer, 2021. Endogenous Oscillations Time-Constrain Linguistic Segmentation: Cycling the Garden Path, *Cereb. Cortex,* 31, 4289-4299

[3] B. Zoefel and R. VanRullen, 2015. EEG oscillations entrain their phase to high-level features of speech sound, *Neuroimage,* 124, 16-23

[4] A. Baddeley, 1992. Working memory, *Science,* 255, 556-559

[5] N. Ding and J. Z. Simon, 2014. Cortical entrainment to continuous speech: functional roles and interpretations, *Frontiers in Human Neuroscience,* 8,

[6] A.-L. Giraud and D. Poeppel, 2012. Cortical oscillations and speech processing: emerging computational principles and operations, *Nat. Neurosci.,* 15, 511-517

[7] N. Boll-Avetisyan, A. Bhatara, A. Unger, T. Nazzi, and B. Höhle, 2020. Rhythmic grouping biases in simultaneous bilinguals, *Bilingualism: Language and Cognition,* 23, 1070-1081

[8] M. J. Crowhurst and A. T. Olivares, 2014. Beyond the Iambic-Trochaic Law: The joint influence of duration and intensity on the perception of rhythmic speech, *Phonology,* 31, 51-94

[9] *PRAAT: Doing phonetics by computer (version 6.1.24)*. (2020). Institut for Phonetic Sciences, Amsterdam.

[10] S. Mathôt, D. Schreij, and J. Theeuwes, 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences, *Behavior Research Methods,* 44, 314-324

[11] D. J. Navarro and D. R. Foxcroft. 2019. *Learning Statistics with Jamovi: A Tutorial for Psychology Students and Other Beginners*. .

[12] Victor J. Boucher, Annie C. Gilbert, Boutheina Jemel; The Role of Low-frequency Neural Oscillations in Speech Processing: Revisiting Delta Entrainment. *J Cogn Neurosci* 2019; 31 (8): 1205–1215.

[13] B. Hayes, 1985. Iambic and trochaic rhythm in stress rules, *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society,* 429-446

[14] K. Johnson and P. W. Jusczyk, 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics, *J. Mem. Lang.,* 44, 548-567

[15] D. M. de la Mora, M. Nespor, and J. M. Toro, 2013. Do humans and nonhuman animals share the grouping principles of the Iambic - Trochaic Law?, *Atten. Percept. Psychophys.,* 75, 92-100

[16] M. J. Crowhurst, 2020. The iambic/trochaic law: Nature or nurture?, *Language and Linguistics Compass,* 14, e12360

[17] A. Revithiadou, 2004. The iambic/trochaic law revisited: Lengthening and shortening in trochaic systems, *Leiden Papers in Linguistics,* 1, 37-62

[18] K. Schulze and S. Koelsch, 2012. Working memory for speech and music, *Annals of the New York Academy of Sciences,* 1252, 229-236

[19] J. Segawa, M. Masapollo, M. Tong, D. J. Smith, and F. H. Guenther, 2019. Chunking of phonological units in speech sequencing, *Brain Lang.,* 195, 104636

Acknowledgments:

We wish to express our thanks to Sharon Stürmer and Michelle Morgenstern for their invaluable help with data collection.

Effects of Vowel Acoustics on the Intelligibility and Comprehensibility of China English

Qian Wang
The Chinese University of Hong Kong (Shenzhen)
wangqian@cuhk.edu.cn

ABSTRACT

The present study examines the impact of vowel acoustic properties on the perceived intelligibility and comprehensibility of China English (CE), an emergent, fast-growing, yet underdescribed variety of English spoken primarily by L1 Mandarin speakers. The acoustic analysis reveals that certain vowel distinctions present in Inner-Circle English dialects are less prominent in CE, while vowel length contrasts remain preserved within the CE vowel system. Additionally, there is a tendency to monophthongize the diphthongs in CE. Multiple regression analysis was conducted to examine the effects of vowel quality contrast, vowel length contrast, degree of vowel centralization, and degree of monophthongization on listeners' vowel recognition performance and comprehensibility rating. The results suggest that the length contrast between /ɑː/ and /ʌ/ and the degree of monophthongization of /eə/ and /oʊ/ significantly impact the perceived intelligibility and comprehensibility of CE.

Keywords: vowel acoustics, intelligibility, perception, comprehensibility, China English

1. INTRODUCTION

In today's globalized world, where English is increasingly used as a lingua franca, many new English varieties are emerging in different parts of the world. China English (CE), as one of the fastest-growing varieties of English with the largest number of speakers, has gained a great deal of attention in recent years [7, 9, 12]. Although the intelligibility of native Englishes has been studied extensively over the past few decades, remarkably little research has been conducted on the intelligibility of CE. Nevertheless, research has been conducted on the intelligibility of Chinese learners of English, revealing that CE exhibits distinct phonological features that can influence listeners' perceptual judgments [1, 8, 14, 17].

One of the most contentious topics in English as a lingua franca (ELF) research is the extent to which a new English variety is intelligible to international listeners, as it plays a critical role in determining the legitimacy of the variety. Levis [13] outlined the distinction between broad and narrow views of intelligibility. The former defines intelligibility as the accuracy of listeners' orthographic transcriptions of L2 speech, while the latter refers to listeners' impressionistic perception of how easily they understand L2 speech. Munro and Derwing [15] refer to the broad sense of intelligibility as comprehensibility, and they pointed out that these two constructs did not measure the same thing because it is common that listeners can transcribe the speech perfectly and yet perceive it to be difficult/effortful to understand. In this study, both narrow intelligibility and broad intelligibility (i.e., comprehensibility) were measured via a forced-choice word identification task and a scalar rating task, respectively.

Previous studies examining the factors that make non-native speech unintelligible suggest that pronunciation has a dominant role in the loss of intelligibility of non-native speech. Derwing and Rossiter [5] claim that prosody severely impairs intelligibility. On the other hand, other studies have emphasized the importance of segmental features, such as vowel and consonant sounds, in causing unintelligibility [16]. Jenkins [10] argued that pronunciation issues account for the biggest source of loss of comprehensibility and intelligibility, and these most commonly occur at the segmental level. In Jenkins' lingua franca core, vowel length contrast (especially before voiced/unvoiced consonants) and vowel quality of /ɜː/ were core pronunciation features that hindered intelligibility [10].

Previous studies have reported that certain vowel characteristics in CE have the potential to impact both the intelligibility and comprehensibility of CE. Hung [9], in his seminal work on the phonology of CE, found that CE has a simplified system of 6-7 vowels with a general lack of long/short or tense/lax contrast and diphthong simplification. Previous research has shown that variability in intelligibility can be related to differences in vowel space area [2,3], i.e., speakers who have larger vowel spaces tend to be more intelligible.

Given the above-mentioned, the study aims to investigate the impact of CE speakers' vowel production on the intelligibility and comprehensibility of CE. In particular, this study analyzes vowel quality contrast, vowel length

contrast, degree of vowel centralization, and degree of monophthongization of CE diphthongs and their effect on CE's intelligibility and comprehensibility.

2. METHOD

The study had two interrelated phases: production (identifying CE's acoustic features) and perception (examining listeners' recognition and evaluation of the speaker's comprehensibility) to establish the link between CE's vowel realizations and listener perception.

2.1. Participants

CE is defined as the English variety spoken by well-educated and highly proficient Chinese speakers of English. According to this working definition, six female CE speakers, aged 19-24 years (mean age of 21.5), were recruited from a high-caliber university where English is the primary language of instruction. They had been learning English for 14-16 years (mean of 14.67). Female speakers were chosen to maintain homogeneity and avoid confounding acoustic patterns caused by gender.

A total of 30 adult listeners with normal speech and hearing participated in the perception experiment. They were included into five groups based on their national backgrounds and Kachru's three-circle model of World Englishes [11]: six from Inner Circle (US), six from Outer Circle (Germany, Pakistan, India, Nigeria, South Africa), six from Expanding Circle (Vietnam, South Korea, Thailand, Poland, Myanmar), six Mandarin-speaking listeners from mainland China, and six from Indonesia.

2.2. Speech task

A word list with 19 permissible vowel phonemes in CE (five long vowels, six short vowels, and eight diphthongs) in the /hVd/ frame was used to test the intelligibility of CE vowels. Uniform word frames were used to eliminate other factors that could affect word recognition. Speakers read the words with the same vowel as the test words and then read the test words in the sentence frame "I say ___ again." Each word was read twice at a normal speed.

2.3. Listening task

Vowel intelligibility was gauged through a forced-choice word identification task. In this task, participants were asked to listen and identify each word they heard by selecting the correct word from a list of 19 words containing the target vowels and beginning with "h" and ending with "d." Before listening to the test items, each participant listened to three practice items to familiarize themselves with the task and computer interface. Each stimulus was played a maximum of two times, and all the stimuli were presented in random order for each of the listeners. Intelligibility was measured by the ratio of correctly identified words to the total number of words in the study, which is also called the "percent of correct identification" (PCI). After completing the word identification task for each speaker, listeners were asked to report their level of difficulty understanding the speaker's speech on a scalar of 1 (extremely hard to understand) to 9 (extremely easy to understand), which is how comprehensibility was measured in this study.

2.4. Acoustic measurements

The study measured several acoustic properties of vowels, including formant frequency (F1 and F2), duration contrast, quality contrast (as indicated by the Euclidean distance between four vowel pairs), and vowel space area. For diphthongs, the degree of formant movement was measured by calculating the rate of change of formant frequency values (ROC). F1 and F2 values were measured at the temporal midpoint of the monophthongs using Praat and were converted into a Bark scale [20] after obtaining each speaker's average F1 and F2 values.

(1) $Z = 13 \arctan(0.00076\,F) + 3.5 \arctan(F/7500)^2$

In this formula, F is the frequency in Hertz, and Z refers to the frequency in Bark. The vowel chart of CE plotted using the Bark values of F1 and F2 better reflects the relative perceptual distance between any two vowels. The comparison between vowels in CE and other varieties of English can be easily made in the vowel chart by using this scale.

To measure the quality contrasts between tense/lax vowels, the average Euclidean Distance (ED) was calculated between two vowels within each tense/lax vowel pair, based on their F1 and F2 values plotted in a two-dimensional vowel space. For example, to determine the ED between /i/ and /ɪ/, the two vowels are plotted as points in the vowel chart, with /i/ represented by (F1i, F2i) and /ɪ/ represented by (F1ɪ, F2ɪ). The ED can be calculated using the Pythagorean theorem, expressed as:

(2) $ED(i, ɪ) = \sqrt{(F1_i - F1_ɪ)^2 + (F2_i - F2_ɪ)^2}$

After plotting all the monophthongs in a two-dimensional vowel chart, the CE vowel space area (area of the vowel polygon based on four corner vowels /i:, u:, ɑ:, æ/) was calculated for each CE

speaker and used to indicate vowel space area and dispersion [3, 6].

To determine vowel duration, the onset and offset of each vowel were manually determined from the waveform using Praat [2, 18], with reference to a spectrogram. To compare vowel durations, mean long-to-short vowel duration ratios were calculated instead of using absolute values due to expected differences in articulation rate.

To measure diphthongs, the rate of change (ROC) of formant values (F1 and F2) was calculated using the formula "y (D-C)/x (B-A) (Hz/sec)", as shown in Figure 1 (adapted from [19]). In this figure, the vertical axis represents the formant frequency of a vowel, and the horizontal axis refers to the duration of the vowel. Point A refers to the point in time where t=20% of the entire vowel duration; Point B refers to the point where t=80% of the entire vowel duration; Point C refers to the formant value at Point B; and Point D refers to the formant value at Point A. Therefore, the ROC can be calculated by "y (D-C)/x (B-A) (Hz/sec)."

2.4. Statistical analysis

In the perception study, the acoustic features of CE vowels discussed above, including vowel quality contrast (i.e., ED), vowel length contrast (i.e., DR), degree of vowel centralization (i.e., vowel space area), and degree of monophthongization (i.e., ROC), were submitted to statistical analysis to explore how these acoustic features of vowels affect listeners' perception of CE. Stepwise multiple regression analysis was performed to further explore the degree to which the acoustic properties of vowels predicted the intelligibility and comprehensibility of CE. Stepwise regression enables the building of good models when many predictors are available, particularly if correlated. The alpha level for significance for all the tests was set as $p < 0.05$.

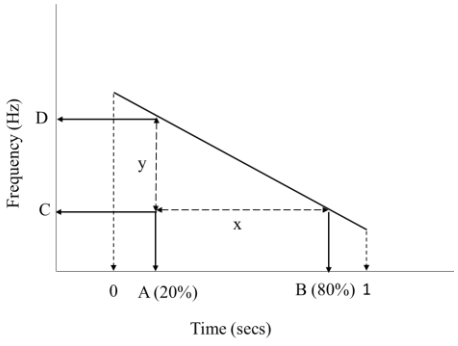


Figure 1: Schematic representation of how the ROC (rate of change) was obtained

3. RESULT

3.1. Production phase

The acoustic analyses of CE vowels show that there are some distinctive features of CE. Specifically, some vowel distinctions that exist in Inner-Circle English varieties are less evident in CE (as illustrated in Figure 2), while vowel length contrasts are still maintained. Furthermore, the study revealed a tendency to monophthongize the diphthongs in CE.

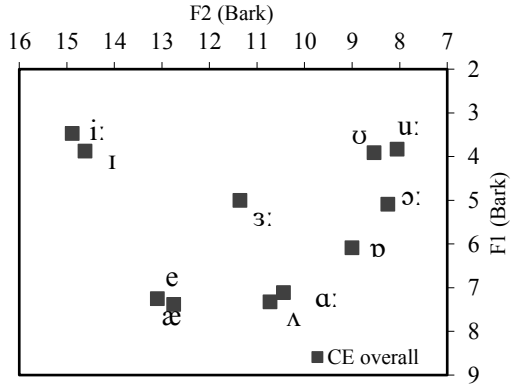


Figure 2: Formant plot for average CE monophthongs

The acoustic analyses of CE vowels revealed that vowel length contrasts were maintained in CE, with all long vowels being longer than their short counterparts (as shown in Figure 3). Paired T-tests confirmed that the durational differences between long and short values were statistically significant, except for the pair /e/-/æ/. Specifically, /iː/ was significantly longer than /ɪ/, $t(5) = 3.10$, $p < .05$; /ɑː/ was significantly longer than /ʌ/, $t(5) = 2.72$, $p < .05$; /ɔː/ was significantly longer than /ɒ/, $t(5) = 3.36$, $p < .05$; and /uː/ was significantly longer than /ʊ/, $t(5) = 4.71$, $p < .05$. However, there was no significant difference in vowel length between /e/ and /æ/ (Mean = 0.17, SD = 0.04), $t(5) = 1.85$, $p > .05$). On average, long vowels were around 1.34 times the length of short vowels.

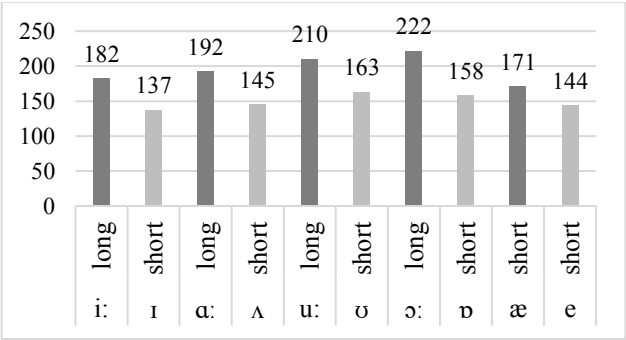


Figure 3: Duration (ms) of long and short vowels in CE

The data presented in Figure 4 indicates that the six CE speakers had varying degrees of vowel expansion, as reflected by the size of their vowel space area. The vowel space area for the CE speakers in this study ranged from 15.18 to 40.64, with S4 having the largest area (40.64) and S5 having the smallest (15.18). However, there were no significant differences in vowel space area between S1 (29.85), S2 (20.18), and S6 (24.21).

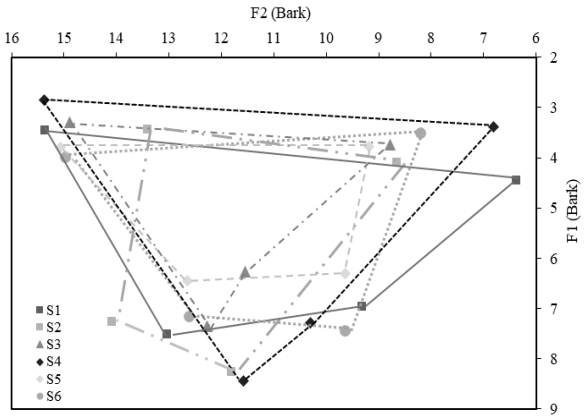


Figure 4: Vowel polygons of six CE speakers

3.2. Perception phase

The results of the statistical analyses indicate that two acoustic variables have a significant impact on the intelligibility scores. These variables are the duration ratio between /ɑː/ and /ʌ/ (Beta = .37, $p < .05$) and the degree of monophthongization of /oʊ/ as in "goat" (Beta = .16, $p < .05$), as presented in Table 1. Specifically, the study found that the intelligibility scores were positively correlated with a higher duration ratio between /ɑː/ and /ʌ/ and a larger ROC of /oʊ/. This suggests that the length contrast between /ɑː/ and /ʌ/ and the degree of monophthongization of /oʊ/ are important factors in speech intelligibility.

Predictor variable	Beta	Partial correlation	t	Sig.
Duration ratio of /ɑː/ and /ʌ/	.37	.38	5.40	.000
ROC (F2) of /əʊ/	.16	.17	2.31	.022
Note. Final model R^2 = .17, F (2, 179) = 18.59, $p <$ .01, Adjusted R^2= .16				

Table 1: Result of multiple regression analysis using vowels variables as predictors of intelligibility

The results of the regression analysis indicate that only two variables had a significant impact on comprehensibility: the duration ratio between /ɑː/ and /ʌ/ (Beta = .29, $p < .05$) and the ROC of /eə/ (Beta = -.19, $p < .05$). These findings suggest that the length contrast between /ɑː/ and /ʌ/ and the degree of monophthongization of /eə/ are significant factors affecting comprehensibility ratings.

Predictor variable	Beta	Partial correlation	t	Sig.
Duration ratio of /ɑː/ and /ʌ/	.29	.29	3.96	.000
ROC (F1) of /eə/	-.19	-.18	-2.50	.013
Note. Final model R^2 =.09, F (2, 179) = 8.97, p <. 01, Adjusted R^2= .08				

Table 2: Result of multiple regression analysis using vowel variables as predictors of comprehensibility

4. DISCUSSION AND CONCLUSION

The study has identified certain acoustic qualities of vowels that have a significant impact on listeners' vowel recognition performance and comprehensibility rating. These qualities include the length contrast between /ɑː/ and /ʌ/ and the degree of monophthongization of /eə/ as in "hair" and /oʊ/ as in "goat." These findings suggest that vowel length contrast and the degree of monophthongization of specific diphthongs are important factors in enhancing the intelligibility and comprehensibility of non-native speakers of English.

The study found no statistically significant impact of vowel space area and quality contrast on the intelligibility or comprehensibility ratings of CE speakers. This does not align with Bent et al. [2] and Bradlow et al. [3], which showed higher overall intelligibility for speakers with greater vowel areas. This discrepancy may be due to the high proficiency level and homogeneity of the participants. However, a closer examination of individual speakers revealed a potential positive correlation between vowel space area and word recognition ease, as S4 with the largest space area received higher comprehensibility ratings than S5 with the smallest space area.

It should be noted that these preliminary findings are limited to the word identification task and further research is needed to explore the influence of acoustic variables on the intelligibility and comprehensibility of CE speakers in natural conversation.

5. REFERENCES

[1] Bent, T., Bradlow, A. R. 2003. The interlanguage speech intelligibility benefit. *J. Acoust. Soc. Am. 114,* 1600–1610.

[2] Bent, T., Bradlow, A. R., Smith, B. L. 2008. Production and Perception of Temporal Patterns in Native and Non-Native Speech. *Phonetica 65*(3), 131–147.

[3] Bradlow, A. R., Torretta, G. M., Pisoni, D. B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication 20*, 255–272.

[4] Brato, T. 2018. Basic Vowel Formant Analysis for NORM (Version: 3.0). Retrieved from https://www.uni-regensburg.de/language-literature-culture/english-linguistics/staff/brato/praat-scripts/index.html.

[5] Derwing, T. M., Rossiter, M. J. 2003. The effects of pronunciation instruction on the accuracy, fluency and complexity of L2 accented speech. *Applied Language Learning 13,* 1-18.

[6] Fox, R. A., Jacewicz, E. 2008. Analysis of total vowel space areas in three regional dialects of American English. *Proceedings of Acoustics 08* (pp. 495–500). Paris, France.

[7] Hansen Edwards, J. G. 2017. China English: Attitudes, legitimacy and the native speaker construct: Is China English becoming accepted as a legitimate variety of English? *English Today 1*(2), 1–8.

[8] Hardman, J. B. 2010. *The intelligibility of Chinese-accented English to international and American students at a US university* (Unpublished doctoral thesis). The Ohio State University.

[9] Hung, T. N. 2002. *Phonological features of Chinese varieties of English.* Paper presented at the 7th English in Southeast Asia Conference, Hong Kong Baptist University, December 6–8.

[10] Jenkins, J. 2000. *The phonology of English as an international language.* Oxford: Oxford University Press.

[11] Kachru, B. B. 1985. Standards, codification and sociolinguistic realism: the English language in the outer circle. In: R. Quirk and H.G. Widdowson (Eds), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge: Cambridge University Press.

[12] Kirkpatrick, A. 2007. *World Englishes: implications for international communication and English language teaching.* Cambridge, England: Cambridge University Press.

[13] Levis, J. 2006. Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–270). New York: Palgrave Macmillan.

[14] Munro, M.J., Derwing, T.M. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning 45*(1), 73–97.

[15] Munro, M. J., Derwing, T. M. 1999. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Article reprinted in *Language Learning 49*(1), 285-310.

[16] Munro, M. J., Derwing, T. M. 2006. The functional load principle in ESL pronunciation instruction: An exploratory study. *System 34*(4), 520–531

[17] Rogers, C. L. 1997. *Intelligibility of Chinese-accented English* (Unpublished doctoral dissertation). Indiana University, Bloomington.

[18] Smith, B. L., Hillenbrand, J., Ingrisano, D. 1986. A comparison of temporal measures of speech using spectrograms and digital oscillograms. *Journal of Speech & Hearing Research 29*(2), 270–274.

[19] Tsukada, K. 2008. An acoustic comparison of English monophthongs and diphthongs produced by Australian and Thai speakers. *English World-Wide 29*(2), 194–211.

[20] Zwicker, E., Terhardt, E. 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am. 29*(5), 1523–1525.

GENERALIZATION OF TRAINING TO AN UNTRAINED PHONETIC ENVIRONMENT: INITIAL /S-Z/ TO FINAL /S-Z/

Birgitte Poulsen, Ocke-Schwen Bohn, and Sidsel Rasmussen

Department of English
Aarhus University, Denmark

ABSTRACT

This study examined whether native speakers of Danish, which has no voiced fricatives, can be trained to perceive the initial English /s-z/ contrast, and whether training on the initial contrast affects the perception of English /s-z/ in final position. 25 native Danish speakers participated as either controls (n=9, no training) or trainees (n=16). The trainees conducted 10 training sessions on /zV/ and /sV/ tokens over three weeks. All participants were tested for identification accuracy on the initial /zV/, /sV/ tokens, which were trained, as well as on final untrained /Vs/, /Vz/ tokens before and after the 3-week training period. The trainees, but not the controls, were significantly more accurate in perception of initial /zV, sV/ after the training period. Interestingly, post-training accuracy for untrained /Vs/-/Vz/ also improved for the trainees, and more so for stimuli with voicing cues than for stimuli with vowel duration cues only.

Keywords: Perceptual training of nonnative contrasts, Generalization of training, English /s-z/.

1. INTRODUCTION

Many studies has shown that adults' perception of non-native speech sounds can be improved through perceptual training [1], [2]. In addition, several studies have shown that the efficacy of training extends beyond trained tokens in that learning may generalize to new words spoken by familiar talkers or to novel tokens produced by an unfamiliar talker [3]. However, very few studies have examined whether training for a contrast in one syllable position (e.g., initial) affects perception of this contrast in a different position (e.g., final), but see [4]. Pursuing this question would help clarify whether speech sound learning happens on the level of specific acoustic dimensions, on the level of position-sensitive allophones, or on a more abstract phonemic level.

In initial position, the voicing contrast between English /s/ and /z/ is implemented by the absence (for /s/) or presence (for /z/) of glottal pulsing during the periods of frication. In final position, the nominal "voicing" contrast is produced with presence or absence of glottal pulsing during frication and/or different duration ratios of vowel to consonant: low (short vowel and long fricative) for the voiceless member, and high (long vowel and short fricative) for the voiced member of the contrast. Native English listeners rely mainly on these relational cues for the identification of voicing contrasts of fricatives in syllable-final position [5], but nonnative listeners without native contrasts in this position have been reported to rely more on other cues [6].

The present study examines how perceptual training affects native Danish listeners' perception of the English /s-z/ contrast in the trained initial and the untrained final position. Danish has no voiced fricatives, and previous studies have shown that native Danish speakers assimilate English initial [s] and [z] tokens to the same native category, /s/ [7], [8]. In terms of the Perceptual Assimilation Model (PAM, [9]), this is a Category Goodness assimilation type, in which the goodness-of-fit of English [s] to Danish /s/ is much better than that of English [z].

Previous studies have suggested that experience with a contrast in one phonetic environment (e.g., syllable-initial) can be exploited to perceive the same contrast in a different position [6], [10]. Broersma [6], [10] showed this for Dutch listeners' perception of obstruent voicing contrasts in final position, noting that Dutch has obstruent voicing contrasts in initial and medial, but not final position. Relatedly, Trapp and Bohn [4] reported that training Danish adolescents on the syllable-final English /s-z/ contrast led to improved identification accuracy for this contrast not only in final position, but also in the untrained syllable-initial position. These studies thus address an important issue raised in both the Speech Learning Model [11] and its revised version [12], namely, the level at which speech sounds are perceived and learned: as position-sensitive allophones or at the more abstract level of the phoneme.

The present study examined:

1) To what extent adult L1 Danish speakers' accuracy in identifying the English fricative voicing contrast /s-z/ would improve through internet-based training.
2) Whether any training effect would be allophone specific or phoneme-general, i.e., if trainees rely on trained cues in a different environment.

2. METHOD

2.1. Participants

25 native speakers of Danish (19 f, 6 m, mean age = 23.4 years, range 20-30) participated. None of the participants had spent any extended period in an English-language environment. Self-reported English proficiency (speaking and understanding) averaged 4.6 (range 3-5) on a scale from 1 (very low) to 5 (very high). Participants were randomly assigned to either the experimental (training) group (n = 16) or the control group (n = 9). Participants received an honorarium equivalent to 50 euros (training group) or 25 euros (controls). None of the participants reported hearing impairments.

2.2. Stimuli

The pre- and post- initial-fricative tests consisted of 60 CV tokens produced by two native English speakers: 30 each by a female and a male speaker. Each of the speakers' tokens consisted of 15 /sV/ and 15 /zV/ syllables, with 5 different tokens of V = /ɑ, i, u/ (3 vowels x 5 different tokens x 2 fricative contrasts x 2 speakers = 60). These tokens were selected from the Shannon et al. (1999) corpus [13]. The final-fricative tokens for the pre- and post-training sessions were recorded by two native English speakers (1f, 1m) and validated by three native English speakers. Each of the talkers produced five tokens each of /Vs/ and Vz/ syllables with V = /ɑ, i, u/ for a total of 60 tokens. The stimuli for the training sessions were the initial fricative tokens, presented in two randomizations, for a total of 120 trials.

2.3. Procedure

The current study consisted of three phases all conducted using the web-tool PERCY [14]: 1) a pre-training session (including first training), 2) nine evenly spaced sessions of internet-based training at the participant's home resulting in ten training sessions per participant, and 3) a post-test, three weeks after pre-test. The control group received no training. The pre- and post-tests were identical for the training and control groups, testing the perception of not just the trained initial /s-z/ contrast but also the untrained final /s-z/ contrast to examine whether any training effect would be allophone-specific or phoneme-general. The pre- and post-training procedures differed on two points: Only the pre-training session contained an explanation and instruction on the articulation of English /s/ and /z/. Here the participants were asked to touch the front of their neck to feel the absence/presence of vocal fold vibrations during production of /s/ and /z/. In addition, only the pre-training session included a familiarization task (identification of initial /fV/ and /vV/) to acquaint participants with the format of the identification tasks. The English initial /f-v/ contrast is unproblematic for native speakers of Danish, who map these fricatives consistently and with high goodness ratings to their native (initial) /f/ and /ʋ/, respectively [7], [8]

The pre- and post-tests consisted of two separate tasks: Identification of final /s/ and /z/, and identification of initial /s/ and /z/. The ten training sessions with feedback trained only initial /s-z/.

At both pre- and post-training sessions participants also took part in a delayed-repetition production task, the results of which are not included in this paper.

Immediately following the identification task on initial /s/-/z/ at our lab, participants assigned to the trainee group completed their first training session. During the following three weeks, the training group completed nine more identical training sessions online at home. Participants were reminded of each training session, and the results obtained via PERCY allowed the experimenters to make sure that participants followed the training schedule. Each session was self-paced and took ca. 10-12 minutes to complete depending on accuracy.

During tests and training, the participants were instructed to listen to the tokens through high quality headphones provided by the experimenters and were asked to identify whether the initial (or, in the case of final fricative, the final) consonant was a /s/ or a /z/. The participants indicated whether they heard a syllable with /s/ or /z/ by clicking on one of two buttons on a computer screen labeled <s> and <z>. During training, immediate feedback was provided after each response: After a correct response, the selected button would turn green, and the next token would be played. After an incorrect response, the button would turn red, the token would be replayed, and, following a short interval, the next token would be played. At the end of each training session, trainees received information on their accuracy rates.

3. RESULTS

The results for the trained initial /s-z/ contrast are presented in Fig. 1, and the results for the untrained final /s-z/ contrast, in Fig. 2.

We first compared the identification accuracy of the trainee and the control group at pre-training. For the initial /s-z/ contrast, which would later be trained, the mean accuracy was 75.8% (SD=11.6) for the trainee group and 70.7% (SD=16.1) for the control group. A t-test revealed that the difference between the groups was nonsignificant ($t(23) = 0.924$, $p > .3$).

Likewise, the difference in identification accuracy for the final /s-z/ contrast (trainees: 65.3%, SD=14.3; controls: 67.4%, SD=13.9) was also nonsignificant ($t(23) = 0.361$, $p > 0.7$). These results suggest that any difference between the trainee and the control groups at post-training can be attributed to the training regime.

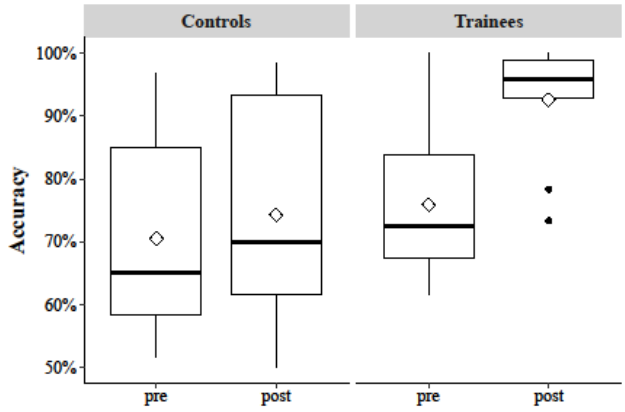


Figure 1: Accuracy rates of initial /s-z/ for control group and training group at pre- and post-training. Diamonds indicate means, bold bars indicate medians.

With respect to the trained initial /s-z/ contrast, the trainees' mean accuracy increased significantly from 75.8% (SD=11.6) at pre-training to 92.5% (SD= 9.3) at post-test, $t(15) = 5.870$, $p < .001$). The control group's accuracy did not change significantly during the interval between pre- and post-training (70.7%, SD=16.1 and 74.2%, SD=17.8), $t(8) = 1.631$, $p > .07$.

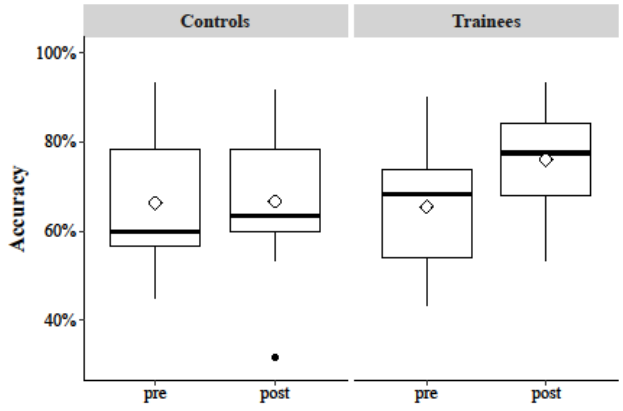


Figure 2: Accuracy rates for final /s-z/ for control group and training group at pre- and post-training. Diamonds indicate means, bold bars indicate medians.

With respect to the untrained final /s-z/ contrast, the group that was trained on the initial fricative contrast improved significantly from an accuracy rate of 65.3% (SD=14.3) to 75.9% (SD=12.0) from pre- to post-training ($t(15) = 3.915$, $p < .01$), whereas the control group's accuracy between the interval of pre- and post-training (67.4.7%, SD=13.9 and 70.7%, SD=13.1) did not change significantly, $t(8) = 0.193$, $p > .1$).

An additional and unexpected finding was that the trainees' overall identification accuracy for final /s/ and /z/ at post-training differed significantly for the two talkers who provided the stimuli, as shown in Fig. 3. At pre-test, the numerical accuracy difference for the tokens from talker LH (mean: 60.6%, SD= 13.3) and talker ZB (mean: 70.2%, SD=17.8) did not differ significantly, $t(29) = 1.727$), $p > .09$). However, after training ZB's tokens were identified significantly more accurately (mean: 82.9%; SD=13.9) than the tokens provided by LH (mean: 69.2% SD=13.5), indicating that something about ZB's tokens signaled the contrast between /s/ and /z/ more clearly than did LH's tokens, $t(29)=2.839$, $p < .01$.

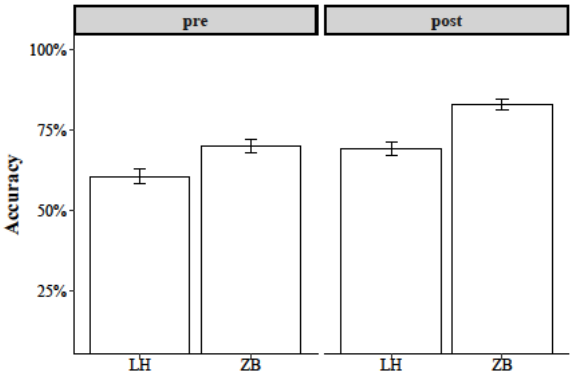


Figure 3: Identification accuracies at pre- and at post-training for syllable-final /s-z/ tokens provided by talkers LH and ZB. Error bars represent +/-1SE.

Acoustic analyses of the final /z/ tokens revealed that the duration ratio of vowel to consonant (V:C) differed significantly for the two talkers: The mean ratio for talker LH was 2.4 (SD=0.34), and the ratio for talker ZB, 2.0 (SD=0.27), $t(28) = 3.6732$, $p < .001$. This was also true for the V:C duration ratios for final /s/ tokens, for which LH's ratio (mean: 0.852, SD= 0.29) was significantly larger than ZB's (Mean: 0.513, SD= 0.08), $t(28) = 4.176$, $p < .001$.

Raphael [5] showed that native English listeners use the V:C ratio as a cue to decide whether the final consonant is voiced or voiceless. This means that LH's V:C ratio would signal the voicing of the final fricative more clearly than ZB's ratio. However, Fig. 3 shows that the nonnative listeners in the present study apparently benefitted less from this cue (duration ratio of V:C) than native speakers would. Further acoustic analyses revealed a systematic

difference between the two talkers' use of voicing in the fricative portions of their syllable final /z/ tokens: Talker LH partly extended voicing from the preceding vowel into the final fricative in only 5 of the 15 tokens, whereas talker ZB consistently extended the voicing from the preceding vowel into all his final /z/ tokens. Fig. 3 shows that the identification accuracy for the tokens for the talker who did not consistently employ partial voicing in the final fricative (LH) was lower at both pre- and post-training than for the talker who did consistently voice (ZB). These findings suggest that the participants in the present study benefited more from the presence of multiple cues (ZB's final voicing as well as a large V:C ratio) than LH's single cue (an even larger V:C ratio).

4. DISCUSSION

The present study examined the effect of internet-based perception training on native Danish speakers' perception of the initial English fricative contrast /s/-/z/. As reported in previous training studies, e.g., [1], [4], the results of the present study showed that perceptual training significantly increased perceptual accuracy for initial English /s/ and /z/. At pre-test, both the trainee and the control group were informed about the articulatory difference in the implementation of the contrast (presence vs absence of vocal fold vibration). The identification accuracy of the two groups did not differ significantly at pre-test. Trainees, who completed ten sessions of training over a period of three weeks, were significantly more accurate in identifying initial fricatives at post-training than at pre-training. The controls, who had not received any training between testing, did not significantly improve their identification accuracy of initial fricatives /s/ and /z/.

We also examined whether training of the fricative contrast in one position, syllable-initial, would transfer to an increase in identification accuracy of the contrast in a different position, syllable-final. Interestingly, the results indicate a moderate but significant training effect for the untrained final fricative voicing contrast. A similar albeit reverse effect is reported in [4] where training of final /s-z/ led to an improvement in identification accuracy of initial /s-z/ by adolescent L1 Danish listeners.

Regarding the perception of /s/ and /z/ in final position, we found a numerical difference between the accuracy for the two talkers' tokens at pre-test, and a significant difference at post-test. As expected, see [15], both native talkers' vowel-to-consonant ratio was larger for final /z/ than for final /s/ tokens. However, only one of the talkers, ZB, consistently extended voicing for the preceding vowel into the final fricative portion for /z/, whereas the other talker, LH, did so only for 5 of her 15 tokens. The tokens from the talker who consistently used voicing were more accurately identified by the trainees especially at post-test. This suggests that training on the syllable-initial voicing contrast (where the fricative portions differ with respect to absence vs presence of vocal fold vibration) could have sensitized the trainees to voicing duration during frication. At pre-training, participants were acquainted with the presence of vocal fold vibration for fricative voicing, and in both tests and training tasks, the same two response labels, <s> and <z>, represented the voicing contrast in syllable-initial as well as in syllable-final position. Thus, since the participants were trained to apply vocal fold vibration to the perception of voicing, vocal fold vibration may have been emphasized as a voicing cue at the expense of V:C duration ratio cues [16]. In the present case, this would indicate that speech sound learning happened on the level of specific acoustic dimensions.

Future studies should explore more systematically whether acoustic cues which trainees successfully learn to use in one position can be transferred to a different, untrained position. Addressing this question could contribute to a clarification of what it is that learners attend to and learn in non-native speech learning: Phonemes, position-sensitive allophones, or the specific acoustic cues which are used to signal speech sound differences in different positions.

5. REFERENCES

[1] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *J. Acoust. Soc. Am.*, vol. 89, no. 2, pp. 874–886, Feb. 1991.

[2] K. Saito, K. Hanzawa, K. Petrova, M. Kachlicka, Y. Suzukida, and A. Tierney, "Incidental and Multimodal High Variability Phonetic Training: Potential, Limits, and Future Directions," *Lang. Learn.*, vol. 72, no. 4, pp. 1049–1091, Dec. 2022.

[3] S. E. Lively, J. S. Logan, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1242–1255, Sep. 1993.

[4] L. Trapp and O.-S. Bohn, "Training Danish listeners to identify English word-final /s/ and /z/: Generalization of training and its effect on production accuracy.," in *Proceedings of New Sounds*, 2002, pp. 343–350.

[5] L. J. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *J. Acoust. Soc. Am.*, vol. 51, no. 4B, pp. 1296–1303, 1972.

[6] M. Broersma, “Perception of final fricative voicing: Native and nonnative listeners’ use of vowel duration,” *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1636–1644, 2010.
[7] C. S. Horslund and O.-S. Bohn, “Assimilation patterns predict L2 identification accuracy of English initial consonants,” *J. Second Lang. Pronunciation*, vol. 8, no. 2, pp. 218–247, 2022.
[8] O.-S. Bohn and A. A. Ellegaard, “Perceptual assimilation and graded discrimination as predictors of identification accuracy for learners differing in L2 experience: The case of Danish listeners’ perception of Danish fricatives,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 2070–2074.
[9] C. T. Best and M. Tyler, “Nonnative and second-language speech perception,” *Lang. Exp. Second Lang. Speech Learn. Honour James Emil Flege*, pp. 13–34, 2007.
[10] M. Broersma, “Perception of familiar contrasts in unfamiliar positions,” *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3890–3901, 2005.
[11] J. E. Flege, “Second language speech learning: theory, findings, and problems.,” in *Speech perception and linguistic experience: theoretical and methodological issues*, W. Strange, Ed. Timonium, MD: New York Press, 1995, pp. 229–273.
[12] J. E. Flege and O.-S. Bohn, “The revised Speech Learning Model (SLM-r),” in *Second Language Speech Learning.*, R. Wayland, Ed. Cambridge, UK: Cambridge University Press., 2021, pp. 3–83.
[13] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, “Consonant recordings for speech testing,” *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. L71–L74, 1999.
[14] C. Draxler, “Online experiments with the Percy software framework - experiences and some early results,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, May 2014, pp. 235–240. Accessed: Dec. 07, 2022. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/564_Paper.pdf
[15] A. S. House and G. Fairbanks, “The influence of consonant environment upon the secondary acoustical characteristics of vowels,” *J. Acoust. Soc. Am.*, vol. 25, no. 1, pp. 105–113, 1953.
[16] B. Bassetti, P. Escudero, and R. Hayes-Harb, “Second language phonology at the interface between acoustic and orthographic input,” *Appl. Psycholinguist.*, vol. 36, no. 1, Art. no. 1, 2015.

THE ROLE OF PRODUCTION IN PERCEPTUAL LEARNING

Lena-Marie Huttner[1], Noël Nguyen[1], Martin J. Pickering[2]

[1]Aix Marseille University, CNRS, LPL, Aix-en-Provence, France, [2] The University of Edinburgh
lena-marie.huttner@univ-amu.fr, noel.nguyen-trong@univ-amu.fr, martin.pickering@ed.ec.uk

ABSTRACT

Speech perception and production are often said to be intrinsically linked, though the nature of the link is contested. In this pilot study, we investigated whether the production of a VOT contrast would influence changes in perception. We designed an online experiment with 2x2 conditions, in which participants were asked to categorize nine stimuli on a voice onset time continuum between /tin/ and /din/ before and after playing a categorization game with a bot. During the game participants were trained on a sound to word correspondence with a bias either towards /din/ or /tin/. In one condition participants alternated categorizing and producing the stimulus, while in another participants only categorized the stimuli. Significant differences in categorization between conditions occurred only during the interactive phase of the experiment. This could mean that while perception and production may be related, production does not appear to mitigate a long-term adjustment in perception.

Keywords: Speech perception, perceptual learning, speech production.

1. INTRODUCTION

In any interaction people have to adjust their perception of the speech signal to their interlocutor's variable production. In laboratory settings this plasticity in speech sound perception can be observed in the phenomenon of perceptual learning [1, 2]: After exposure to an ambiguous stimulus, people will tend to adjust their perceptual boundaries to the input. Theories of speech perception offer several explanations for this observation. However, the nature of the underlying mechanisms that allow for this plasticity are still debated. Yet, it serves a fundamental communicative purpose without which communication would be impossible. In a conversation, interlocutors will come to share their representations on every linguistic level [3]. Perceptual learning may be a manifestation of this alignment on the level of speech sounds: A listener will hear their interlocutor's production, supposedly a manifest of their own representation of a sound, and adjust their perception to it. Similarly, social interaction does not only influence a speaker's perception of speech sounds but also their production. Experiments on phonetic convergence have found that interlocutors will become more similar in their acoustic phonetic realization over time [4, 5]. It is assumed that this phenomenon is caused or accompanied by a change in perception: Speakers are assumed to adjust their production to the perceived input [6]. [3] further stipulate that in order for this process of alignment to function, perception and production must be intrinsically linked. This idea is also iterated in sensorimotor theories of speech perception and production [7]. Yet, it is important to note that though coordinated, speech perception and production are still regarded to be separate processes. A change in one domain may not suffice for adjustment in the other [8, 9]. There is also conflicting experimental evidence on the co-occurrence of perceptual adaptation and convergence in speech production: On the one hand [8] found that native English speakers' perception of a contrast adapted to the speaker when it was presented as an idiolect rather than a dialect, however, production of the contrast was unaffected by this training. On the other hand[10] found that training native Japanese learners of English in the perception of a novel phonemic contrast lead to the production of that contrast. Contrary to [8] and [9], [10] argue that a close link between speech perception and production is a requirement for category formation. These studies ([8, 10] are just two examples of experiments examining the mechnanisms that may govern perceptual adaptation. A common denominator of these studies is the inclusion of social factors that inevitably influence the participant's behavior. Language is a social practice and linguistic behaviors cannot always be isolated from the social environment in which they occur. Yet, when trying to understand the perception production link, it can be useful to break down social interaction to a minimum and exclude any extralinguistic information. It

appears to be debated how and if perception and production influence one another. We here ask what role does the production of a contrast throughout a conversation play in the adjustment of perception? An influence of speech production on speech perception has been observed in [11]. Their ([11]) findings are in line with the assumptions made by [7], which stipulate that a speech sound representation is composed of both perceptual and motor information. In this pilot study we aim to shed a light on the link between perception and production by examining role of speech production on perceptual learning.

2. METHODS

To investigate whether an adaptation in perception was influenced by the production of a contrast, we designed an online experiment in Labvanced [12] with two by two conditions. The experiment was run on participants' smartphones. [13] compared the quality of speech data recorded on smartphones and a head mounted condenser microphone and found no significant influence of recording device on fitness the data for phonetic research. In this study we combined the experimental paradigms of perceptual learning and phonetic convergence studies; using a pre-post paradigm to measure changes in categorization as well as changes in speech production: We measured participants categorization of the stimuli as well as their production before and after playing a categorization game with a bot. We here present and discuss the results of the perceptual side of the experiment.

2.0.1. Stimuli

We measure phonetic adaptation and convergence on voice onset time (VOT). Perceptual adaptation and phonetic convergence on VOT has been shown by [14] and [15] respectively. We created a 9-step VOT continuum ranging form 23 to 55 ms of the words din and tin spoken by a female native speaker of English (dialect region: Northern England). The continuum was created in Praat [16] using a script [17]. Prior to the experiment, we conducted a stimulus test in which 30 participants (native English speakers from the UK) were asked to categorize the stimulus continuum.

2.1. Experimental setup

We recruited 88 native English speakers from the UK between the ages of 18 and 40 on Prolific. Participants were first presented with a consent form followed by a basic demographic questionnaire. Prior to the experimental tasks, we conducted a headphone screening test[18]. Participants were also repeatedly told they should be in a quiet room to participate in the experiment. Participants whose audio data included noticeable background noise that indicated otherwise were excluded from analysis and replaced. Participants were instructed to turn off notifications for the duration of the experiment and to place their phones on a surface in front of them. Participants were told that they would be interacting with a partner whose voice they would be hearing throughout the experiment. To make this claim more believable participants exchanged a small greeting with the experiment script before beginning the first task. The experiment began with the pre-test which consisted of a production and categorization task. In the production task participants were first asked to produce 10 instances of each tin and din; the words were shown in randomized order along with an image of a red microphone to indicate that recording was taking place. In the categorization task, participants first heard one of the nine stimuli before the words tin and din appeared on their screens. Participants were asked to tap on the word they had perceived. The perception task ran for 45 trials in which 5 iterations of each acoustic stimulus were presented. Participants were then told they would be playing an interactive game with another participant. During this interactive phase of the experiment, participants again heard one of the acoustic stimuli and were asked to tap on the word they just heard. They were then shown the word their supposed partner had read on their screen, i.e. they were given feedback on the intended categorization of the stimulus they heard. Here the experiment script was biased towards one of the endpoints. In two of the conditions, VOT-steps 1 through 6 were identified as /din/ (d-bias conditions) whereas in the other two, steps 4 through 9 were identified as /tin/ (t-bias condition). Participants in the interactive conditions were then asked to produce a word (either tin or din) which they read on their screen. In the control condition, participants did not speak and went on to the next categorization task. The interactive game consisted of 90 trials in which each acoustic stimulus was presented 10 times. Participants then proceeded to the post-test in which the two pre-test tasks were repeated in randomized order.

2.2. Hypotheses

In light of the findings by [10, 1, 2] we expect perceptual shifts to occur in all four conditions in

the direction of the bias of the script. For the two conditions with a bias towards /tin/ we expect participants to categorize more stimuli as /tin/ and in the conditions with a bias towards /din/ we expect participants to categorize more stimuli as /din/ in the post-test in comparison to the pre-test. We further expect there to be a difference in categorization between interactive and control condition.

3. RESULTS

We compare the categorization curves between and within conditions. The curves are depicted below 1.

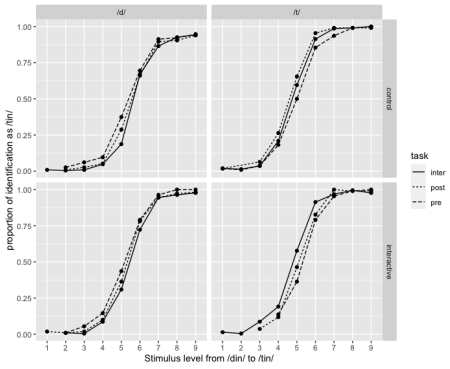


Figure 1: Proportion of identification of the stimuli as /tin/ by condition.

To compare between as well as within group differences, we ran a generalized mixed effects model using the lme4 package in R [19]. With task (pre-test, post-test, interaction) and condition as fixed effects and subject as a random effect (formula: response ∼ condition * task + (1+ stimulus|subject), family = binomial). In this model, the effect of task (interaction) is statistically significant and negative (95% CI [-1.08, -0.44], p < .001) as well as the effect of task (post) (95% CI [-0.90, -0.17], p = 0.004). We conducted a post-hoc Tukey test to examine the differences within and between conditions. Table 1 shows the significant differences between conditions.

task	contrast	estimate	SE	Z-score	p-value
interaction	control t - control d	1.6288	0.338	4.825	0.0001
interaction	interactive t - control d	1.6471	0.334	4.928	0.0001
interaction	control t - interactive d	1.3729	0.316	4.341	0.0009
interaction	interactive t - interactive d	1.3912	0.316	4.405	0.0006
post	control t - control d	1.7238	0.359	4.797	0.0001
post	control t - interactive d	1.3676	0.339	4.033	0.0032

Table 1: Between group differences

There are the following statistically significant within group differences in categorization:

condition	contrast	estimate	SE	Z-score	p-value
control d	pre - interaction	0.7458	0.160	4.653	0.0002
control t	pre - post	-0.8035	0.188	-4.278	0.0011
interactive t	pre - interaction	-0.9798	0.165	-5.939	<.0001

Table 2: Within group differences

We expect any differences in categorization to be more pronounced for the three midpoint stimuli than the endpoint stimuli. We therefore ran the same model as above, but only on the three midpoint stimuli (points 4, 5, and 6 in 1). In this model, with the intercept corresponding to control d-bias and pre-test, The effect of task (interaction) is statistically significant and negative (95% CI [-1.12, -0.38], p < .001) as well as the effect of task (post) (95% CI [-0.89, -0.04], p = 0.033) We again conducted a post-hoc Tukey test to compare contrasts. The statistically significant between group results are reported below in table 3:

task	contrast	estimate	SE	Z-score	p-value
interaction	control t - control d	2.1335	0.403	5.298	<.0001
interaction	interactive t - control d	1.9063	0.401	4.757	0.0001
interaction	control t - interactive d	1.6017	0.408	3.930	0.0048
interaction	interactive t - interactive d	1.3745	0.400	3.434	0.0294
post	control t - control d	2.2412	0.429	5.220	<.0001
post	control t - interactive d	1.6837	0.433	3.892	0.0056

Table 3: Between group differences in the categorization of the three midpoint stimuli

In this analysis the within group comparisons also yielded statistically significant results:

condition	contrast	estimate	SE	Z-score	p-value
control t	pre - post	-0.8496	0.213	-3.994	0.0038
interactive d	pre - interaction	0.6116	0.179	3.423	0.0305
interactive t	pre - interaction	-0.9744	0.185	-5.268	<.0001

Table 4: Within group differences in the categorization of the three midpoint stimuli

4. DISCUSSION

The results show that the categorization behavior between bias conditions differed significantly in the predicted direction: in the post test the /d/ and /t/ bias conditions differed significantly both in the overall analysis and the analysis of the three midpoint stimuli (see tables 1 and 3). There are no statistically significant differences between the categorizations of interactive and control conditions (of the same bias). Within condition there are significant differences between the categorization during the pre-test and the interactive game, a significant difference between pre- and post-test categorization could only be found for the control-t condition. The within condition comparisons of

the overall analysis and the analysis of the three midpoint stimuli show slightly different results: Overall, there are significant differences between pre-test and interactive task categorization in one control and one interactive condition, whereas analysis of the midpoint stimuli shows a significant difference between pre-test and interaction only for the two interactive conditions. A significant difference between pre- and post test categorization can be found in the control-t condition. As shown in figure 1, the control t-bias condition diverts from a pattern the other three exhibit: in the control-t condition, more stimuli are identified as /tin/ in the post-test than in the interactive task. Whereas in the other three conditions the categorization of the post-test lied in between pre-test and interactive task. It is unclear why the control-t condition differs in this regard. It could be that there is an issue of power and that a repetition of the experiment with more participants would result in clearer picture. Further, we only presented five iterations of each acoustic stimulus in the pre- and post-test perception tasks, keeping in mind that in an online experiment, the participants' attention may wane. Follow-up experiments should include attention checks and perhaps more varied stimuli. While interaction does not appear to elicit a robust effect on categorization, the bias condition lead to a significant contrast between groups. Perceptual learning experiments do not always yield a robust effect. For example [20] discuss that some acoustic properties may be easier to learn than others, noting that learning of VOT contrasts by native speakers may be more pronounced than other contrasts. According to [8], adjustments to perceptual representations occur as a function of their cause. In our experiment, participants were asked to categorize non-ambiguous stimuli during the interactive task. Perhaps this impeded participants from adapting their perceptual categories. By simply guessing the categroization of the midpoint stimuli, the participant could still ensure that the understanding was good enough. As the within group contrasts show, categorization of the stimuli was more likely to divert from the pre-test pattern in the interactive task. This could imply that participants adjust their categorization when interacting, but the interaction has no long lasting effect on the perception. This is in contrast to arguments made by [1], who state that changes elicited by perceptual learning are long lasting, but would be similar to what [8] state: Adjustment of perceptual categories is costly, if there is another way to resolve the tension between the acoustic signal and the category, participants will readily do so. In [10] participants further have an external motivation - learning a foreign language - to adapt their perception of a contrast. In our experiment there was no such motivation. Perhaps long lasting changes to perceptual categories are less likely to be elicited if participants do not have a social reason that would yield such changes beneficial. It should further be noted that [17] recommend VOT ranges for stimulus construction which we took into account. However, we noticed that the talent producing the unaltered recordings of our stimuli regularly produced VOTs of more than 100ms for /t/-initial words. Further, initial inspection of the production data showed that the participants themselves regularly produced VOTs of more than 100ms. It is possible that our stimuli were therefore perceived as so extreme that the participant attributed an origin to the stimulus that would inhibit perceptual learning, which would be similar to the findings by [8]. We did however not ask the participants any debrief questions, these interpretations of the participants' possibly socially motivated behavior are therefore speculative. It is also possible that the perceptual system treats the participant's own production as corollary discharge. [21] have shown that self vocalizations are attenuated by the auditory system. Such a suppression of the signal could mean that experimental condition of "interaction" had little influence on the perceptual changes. [9] further argue that perception and production are independent processes that exhibit coordination, but aren't required for each other's function. The results presented here are in line with that perspective.

5. CONCLUSION

Overall our results suggest that adjustments in perception are independent of the speaker's own production. How and whether the participants' production changed throughout the experiment and how that relates to the changes in perception will have to be determined in further analysis. Rather than adjusting their perception long term, participants may have opted to employ another strategy to resolve the tension between the acoustic stimulus and the intended category. Future experiments may further explore the social factors that influence the participant's behavior.While this experiment still raises many questions, it offers a starting point to explore the link between speech sound perception and production.

6. ACKNOWLEDGEMENTS

This work has been conducted in the framework of the Conversational Brains (COBRA) Marie Sklodowska-Curie Innovative Training Network and has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 859588.

7. REFERENCES

[1] R. L. Goldstone, "Perceptual learning," *Annual review of psychology*, vol. 49, no. 1, pp. 585–612, 1998.

[2] T. Kraljic and A. G. Samuel, "Perceptual learning for speech: Is there a return to normal?" *Cognitive psychology*, vol. 51, no. 2, pp. 141–178, 2005.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[4] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 790, 1975.

[5] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[6] N. Nguyen and V. Delvaux, "Role of imitation in the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 46–54, 2015.

[7] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The perception-for-action-control theory (pact): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.

[8] T. Kraljic, S. E. Brennan, and A. G. Samuel, "Accommodating variation: Dialects, idiolects, and speech processing," *Cognition*, vol. 107, no. 1, pp. 54–81, 2008.

[9] J. S. Pardo and R. E. Remez, "On the relation between speech perception and speech production," *The Handbook of Speech Perception*, pp. 632–655, 2021.

[10] R. Akahane-Yamada, Y. Tohkura, A. R. Bradlow, and D. B. Pisoni, "Does training in speech perception modify speech production?" in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE, 1996, pp. 606–609.

[11] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1245–1248, 2009.

[12] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified javascript framework for online studies," in *International conference on computational social science (cologne)*, 2017.

[13] E. U. Grillo, J. N. Brosious, S. L. Sorrell, and S. Anand, "Influence of smartphones and software on acoustic voice measures," *International journal of telerehabilitation*, vol. 8, no. 2, p. 9, 2016.

[14] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychonomic bulletin & review*, vol. 13, no. 2, pp. 262–268, 2006.

[15] K. Nielsen, "Specificity and abstractness of vot imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, 2011.

[16] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[17] M. B. Winn, "Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 852–866, 2020.

[18] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/rl/by japanese adults learning english," *Perception & psychophysics*, vol. 36, no. 2, pp. 131–145, 1984.

[21] J. D. Greenlee, A. W. Jackson, F. Chen, C. R. Larson, H. Oya, H. Kawasaki, H. Chen, and M. A. Howard III, "Human auditory cortical activation during self-vocalization," *PloS one*, vol. 6, no. 3, p. e14744, 2011.

HEARING *BEG* AND *BAG*: VARIATION IN THE PERCEPTION OF PRE-VELAR /Æ/-RAISING

Lisa Sullivan

University of Manitoba
lisa.sullivan@umanitoba.ca

ABSTRACT

Pre-velar /æ/-raising is a process in which some speakers of North American English raise /æ/ before /g/. Documentation of its production is limited, but even less is known about its perception. Perception studies using identification tasks [1, 2] found that perception does not vary based on listeners' production of the process. However, there is some evidence that individuals from regions without pre-velar /æ/-raising have greater metalinguistic awareness of it than those from regions with it [3], suggesting that there may be differences in perception. An AX discrimination task was conducted and revealed that individuals who raise less discriminated raised and unraised /æg/ more than those who raised more, suggesting that there is a relationship between perception and production of this process. Furthermore, differing results from previous studies suggest that this relationship is related to some aspect of perception that is captured by discrimination, but not identification, tasks.

Keywords: sociophonetics, sociophonetic perception, speech perception, North American English, pre-velar /æ/-raising

1. INTRODUCTION

Pre-velar /æ/-raising (hereafter, æg-raising) is a process in which some speakers of North American English raise (and front) the vowel /æ/ before the voiced velar stop (/g/). For speakers who æg-raise, the /æ/ in words like *bag* /bæg/ is pronounced higher than the /æ/ in words like *back* /bæk/, such that it may sound closer to the vowel in *beck* /bɛk/ or *bake* /bek/. In contrast, those who don't æg-raise would pronounce the /æ/ in *bag* very similarly to the vowel in *back*. Acoustic evidence shows that æg-raising occurs across Canada (e.g. [4, 5, 6, 7, 3]), as well as in the Pacific Northwest (e.g. [8, 9]), Upper Midwest (e.g. [10, 11]) and California [12, 13] in the US, but not in Nevada [14] or Colorado [3]. Outside these regions, evidence is limited to self-report data [15] which suggests that the prevalence of æg-raising decreases as you move away from the US-Canada border.

While documentation of the production of æg-raising is limited, even less is known about its perception. Freeman [1] and Sullivan [2] found no relationship between individual participants' production and perception of æg-raising in vowel identification tasks. However, evidence of greater metalinguistic awareness of æg-raising in non-æg-raising regions compared to æg-raising regions suggests that there may be a link that is not captured by these studies.

Variation in speech perception may be due to factors such as mobility, exposure and geography, as well as both the speaker's and listener's native dialects (e.g. [16, 17, 18]). The current study focuses on the link between production and perception. There is some evidence that there are links between perception and the listener's dialect. Listeners have been shown to classify sounds based on the phonemic system of their own dialect [19, 17, 20]. Furthermore, the perceptual saliency of certain dialect features is linked to listeners' native dialects [16, 19]. These dialect features are more salient to non-native dialect listeners than native ones. While these studies don't directly compare individual production and perception, differences between dialect regions suggests a possible link between production and perception.

Several studies [21, 22, 23] have shown that such a link may exist, even within dialects. Sumner and Samuel [23] found differences in the perception of /ɪ/-dropping between New York City speakers who /ɪ/-drop and those who don't. Fridland and Kendall [22] found that speakers' boundary between /e/ and /ɛ/ was correlated with their production, even within dialect. Evans and Iverson [21] found that Northern English speakers with more Northern English accents perceived vowels more closely to Northern productions than those who had more Southern productions. These studies suggest that differences in production within a dialect region may explain differences in perception within that region.

The current study uses a modified AX discrimination task to test if there is a relationship

between an individual's production of æg-raising and their ability to discriminate between raised and unraised /æg/. If individual production has an effect, participants who æg-raise less should differentiate raised and unraised /æg/ more than those who æg-raise more.

2. METHOD

2.1. Participants

49 participants, 25 from Ontario (13 F, 12 M) and 24 from Colorado (15 F, 9 M), completed the experiment. All participants were native speakers of English between the ages of 18 and 35 at the time of the study and who had spent at least 5 years of their childhood in their respective province/state. Ontario and Colorado were chosen as regions for this study as they were expected to produce a range of degrees of raising from no raising (Colorado) to large degrees of raising (Ontario) [24, 25].

2.2. Production Stimuli

The word list consisted of 60 monosyllabic English words including 22 words containing /æ/, /ɛ/ and /e/ followed by /g/ or /k/. See Sullivan [3] for a complete word list.

2.3. Perception Stimuli

Two 5-step (vowel quality) by 5-step (vowel duration) /æ/ to /ɛ/ continua were resynthesized in Praat [26] from the /ɛ/ natural tokens of a male native North American English speaker from the Midwest. The first continuum ranged from /bægsən/ to /bɛgsən/, and the second from /bæksən/ to /bɛksən/. These tokens served as X tokens for the AX discrimination task. The speaker's natural productions of /bægsən/ and /bæksən/ were used as A tokens. These tokens were run through the same script as the X tokens, except that the vowel quality and duration wasn't modified. Praat's [26] change gender feature was then used to make these sound female.

2.4. Procedure

Participants completed the tasks on their own computers using built-in or external microphones. They were asked to wear headphones for the perception task. They first completed the word list reading task in Gorilla [27]. Words were displayed on the screen individually, and participants said each word once. They completed the word list 3 times. The words were randomized each time.

Next, they complete the perception task in jsPsych [28]. For each trial, they heard the unraised A token of /bægsən/ or /bæksən/ followed by an X token from the corresponding continuum. They were asked to indicate if the second token was a different word from the first one, the same word as the first, but spoken with a different accent, or the same word as the first, spoken with the same accent. The 'same word, different accent' option was included so that participants could indicate when they were able to discriminate the two sounds, but they were not so different that they considered them to be different vowels (and thus different words) altogether. Participants heard each X token once (50 trials), and the order of items was randomized by participant.

3. ANALYSIS

3.1. Production

Production tokens were force aligned using the Montreal Forced Aligner [29] and manually corrected in Praat [26]. F1 and F2 measures were taken from the midpoint of each target vowel. Individual degree of raising for each speaker was calculated as a proportion of the F1 of /æk/ to the F1 of /ek/ using the formula in (1). Larger positive scores indicate more raising, as they indicate /æg/ is closer to /ek/. Scores close to 0 indicate no raising, and negative scores indicate that /æg/ is lower than /æk/.

$$(1) \qquad \frac{M_{\text{æk}} - M_{\text{æg}}}{M_{\text{æk}} - M_{\text{ek}}}$$

3.2. Perception

Since the primary research question was about whether or not participants were able to distinguish /æg/ from /æk/, participant responses to the discrimination class were collapsed into two categories "Same", which includes the response "Same word with the same accent" and "Different" which includes both "Different word" and "Same word with a different accent" for the purposes of the statistical analysis. However, the 3 way response variable is considered descriptively, as well.

A mixed effects logistic regression model was constructed with discrimination as the response variable using the *glmer*() function from the lme4 [30] package in R [31]. The lmerTest package [32] was used to obtain significance values. Predictor variables, interactions and random effects were

included as listed in (2).

(2) glmer(Discrimination ~(Vowel Quality + Duration) * Phonological Context * (Production + Metalinguistic Awareness) + (1 + Phonological Context + Vowel Quality + Duration | Participant), data = Perception Data, family = "binomial", control = glmerControl(optimizer = "bobyqa", optCtrl=list(maxfun=2e6)))

The continuous variables (duration, vowel quality, metalinguistic awareness and production) were z-score normalized. Phonological context coded with /k/ as -0.5 and /g/ as 0.5. For post-hoc analysis and descriptive statistics, participants divided into low and high æg-raising groups based on the mean degree of æg-raising.

4. RESULTS

4.1. Production

Individual degree of raising scores are shown in Figure 1. While Ontarians raise more than Coloradans, there is a large degree of variation both within and across groups, ranging from no raising, to raising /æg/ over 75% of the way from /æk/ to /ek/.

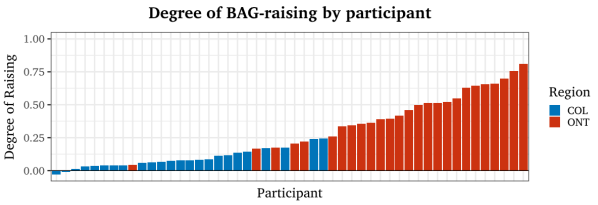


Figure 1: Individual degree of BAG-raising

4.2. Perception: 2-way response variable

Overall, the results are consistent with what would be found if participants were doing the task as expected. X tokens are discriminated from A tokens less when the vowel quality is more /æ/-like (Est. = -1.7, *SE* = -.11, z = -13.48, p < 0.001) and duration is longer (Est. = -1.4, *SE* = -.10, z = -13.59, p < .001).

There are overall differences in how the vowel is distinguished before /g/ and /k/ in line with the fact that the distinction in production only happens before /g/. Participants are more likely to distinguish more /ɛ/-like tokens from unraised /æ/ when the vowels occur before /k/, where the process doesn't happen, than before /g/, where it does. This is born out in the interaction between vowel quality and phonological context in the regression model (Est. = 1.58, *SE* = .15, z = 10.30, p < .001).

This effect is modulated by individual production. As shown in Figure 2, participants who æg-raise less show similar levels of discrimination of more /ɛ/-like tokens before /g/ and /k/. On the other hand, those who æg-raise more show a much lower level of discrimination before /g/, than /k/, suggesting they are less sensitive to distinctions between /ɛg/ and /æg/. This is born out in a three way interaction between vowel quality, phonological context and production in the regression model (Est. = 0.59, *SE* = .15, z = 3.83, p < .001). Consistent with what Figure 2 shows, post-hoc analyses shows a larger effect size of the interaction between vowel quality and phonological context for the high æg-raising group (Est. = 2.09) than the low one (Est. = 1.20).

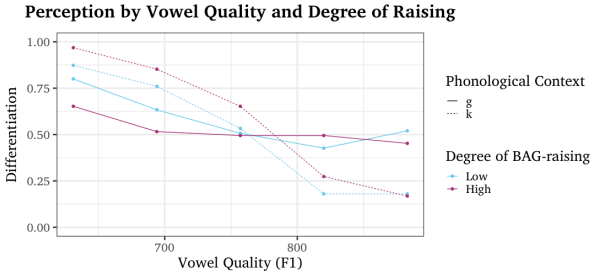


Figure 2: Discrimination Results (2 Levels)

4.3. Perception: 3-way response variable

For the statistical analysis, the three way response variable was collapsed into two levels; however, in 29.5% of the "different" responses participants indicated that the vowel was the same as in the natural unraised /æ/ token, but spoken in a different accent, rather than being a different vowel altogether. Figure 3 shows the proportion of participant responses divided by phonological context, low or high degree of BAG-raising in perception and vowel quality of the vowel in the continuum token.

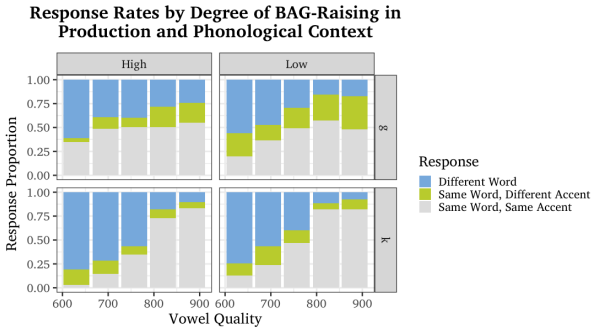


Figure 3: Discrimination Results (3 Levels)

The different accent responses in the /k/ context are similar for both groups, and occur at a lower rate

than in the /g/ context, where they seem to occur more with more /æ/-like tokens, perhaps because speakers are less willing to recategorize these tokens than the more /ɛ/-like ones. Furthermore, low æg-raisers were more likely to have different accent responses in the /g/ context, especially for the more /ɛ/-like tokens, suggesting that these speakers differentiate more in this context, even if they don't recategorize.

In summary, the results show that (1) participants are more likely to differentiate a vowel from /æ/ the less /æ/-like it is, (2) they are more likely to discriminate less /æ/-like vowels before /k/ than /g/, (3) high æg-raisers discriminate less before /g/ than low æg-raisers.

5. DISCUSSION

The goal of this study was to determine if there is a relationship between an individual's production of æg-raising and their ability to discriminate between raised and unraised /æg/. The results indicate that participants are able to differentiate variants of /æg/ and that this is modulated by individual production of æg-raising. Furthermore, the use of a discrimination task with a three way response variable may have contributed to detecting this difference.

The results of this study are in line with the metalinguistic evidence [3] that individuals from non-æg-raising regions are more aware of æg-raising than those from æg-raising regions, as both findings show greater sensitivity to æg-raising for those who do not æg-raise themselves or aren't exposed to it in the community. Furthermore, the results are consistent with previous findings on other phenomena that show a link between production (or dialect) and perception, particularly those that show links between individuals' production and perception [21, 22, 23] and those that show non-native dialect speakers have greater perceptual saliency of certain dialect features than native dialect speakers [16, 19].

On the other hand, the results run contrary to previous æg-raising studies which found no link between production and perception [1, 2]. There are a number of methodological considerations which could explain the difference between this study and the previous ones, including a higher number of participants, the possibility that these participants had a larger range of degrees of raising, and differences in the synthesis and choice of stimuli; however, the most likely explanation seems to be in the task itself.

Where as Freeman [1] and Sullivan [2] used identification tasks which required the participant to categorize the vowel they were perceiving as an English vowel, the current task used a discrimination task and allowed participants to note within-category differences. It could be that the nature of the tasks (discrimination vs identification), explains why the results are different. Perhaps, discrimination better taps into the aspect of perception that results in differences in metalinguistic awareness.

However, it is possible that the third option in the task, which allowed the participant to say that the second word was the same as the first, but just had a different accent, allowed for this difference to emerge. This seems to be supported by the 3-way categorization results where much of the difference between the low and high æg-raising groups in the /g/ context arises from the different accent responses, and not the different word responses. It may, therefore, be that the link between perception and production doesn't have to do with recategorization of the vowel, but rather with hearing it as a variant of the same vowel, a distinction the discrimination tasks previously employed weren't able to capture.

6. CONCLUSION

The current study finds that there appears to be a link between the production and perception of æg-raising. It also suggests that perception studies of dialectal variation should incorporate options for identifying variants of the same vowel, and not just perception of different vowels. Future work should continue to examine how æg-raising is perceived, including examining its perception in context, and with real words. Furthermore, an identification task which allows for identification of variants of the same vowel could be developed to tease apart possible reasons for the difference in results between this study and previous ones [1, 2].

7. REFERENCES

[1] V. Freeman, "Prevelar merger in production vs. perception," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 457–461.

[2] L. Sullivan, "The production and perception of prevelar/æ/-raising by Canadian and American English speakers," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, p. 2809, 2020. [Online]. Available: https://doi.org/10.1121/

1.5147827
[3] ——, "Pre-velar /æ/-raising in Ontario and Colorado English: Production, perception and metalinguistic awareness," Ph.D. dissertation, Univesrity of Toronto, Toronto, ON, 2022.
[4] C. Boberg, "Regional phonetic differentiation in standard Canadian English," *Journal of English Linguistics*, vol. 36, no. 2, pp. 129–154, 2008.
[5] J. H. Esling and H. J. Warkentyne, "Retracting of/æ/in Vancouver English," in *Focus on Canada*, S. Clarke, Ed. John Benjamins, 1993, vol. 11, pp. 229–246.
[6] S. Onosson, "Prevelar vowel raising and merger in Manitoba English," *Journal of English Linguistics*, vol. 50, no. 3, pp. 247–280, 2022.
[7] N. Rosen and C. Skriver, "Vowel patterning of mormons in southern Alberta, Canada," *Language & Communication*, vol. 42, pp. 104–115, 2015.
[8] V. Freeman, "Vague eggs and tags: Prevelar merger in Seattle," *Language Variation and Change*, vol. 33, no. 1, pp. 57–80, 2021.
[9] J. M. Riebold, "The social distribution of a regional change:/æg, ɛg, eg/in Washington State," Ph.D. dissertation, University of Washington, 2015.
[10] E. Koffi, "The acoustic vowel space of central Minnesota English: Focus on female vowels," *Linguistic Portfolios*, vol. 2, no. 1, pp. 2–16, 2013.
[11] C. Zeller, "The investigation of a sound change in progress: /ae/to/e/in midwestern American English," *Journal of English linguistics*, vol. 25, no. 2, pp. 142–155, 1997.
[12] A. Cardoso, L. Hall-Lew, Y. Kementchedjhieva, and R. Purse, "Between California and the Pacific Northwest: The front lax vowels in San Francisco English," *American Speech: A Quarterly of Linguistic Usage*, vol. 2016, pp. 33–54, 2016.
[13] L. Esposito and E. Lake, "Complicating prevelar raising in the West," *American Speech*, pp. 1–45, 07 2021.
[14] K. Gunter, I. Clayton, and V. Fridland, "Pre-velar raising and categorization in Nevada," in *New Ways of Analyzing Variation 47*, New York City, N.Y., 2018.
[15] J. A. Stanley, "Regional patterns in prevelar raising," *American Speech*, vol. 97, no. 3, pp. 374–411, 2022.
[16] R. Allbritten, "Sounding Southern: Phonetic features and dialect perceptions," Ph.D. dissertation, Georgetown University, Washington, DC, 2011.
[17] B. O. Flanigan and F. P. Norris, "Cross-dialectal comprehension as evidence for boundary mapping: Perceptions of the speech of southeastern Ohio," *Language Variation and Change*, vol. 12, no. 2, pp. 175–201, 2000.
[18] J. Hay and K. Drager, "Stuffed toys and speech perception," *Linguistics*, vol. 48, no. 4, pp. 865–892, 2010.
[19] C. G. Clopper, J. B. Pierrehumbert, and T. N. Tamati, "Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise," *Laboratory Phonology*, vol. 1, no. 1, pp. 65–92, 2010.
[20] E. Jacewicz and R. A. Fox, "The effects of cross-generational and cross-dialectal variation on vowel identification and classification," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1413–1433, 2012.
[21] B. G. Evans and P. Iverson, "Plasticity in vowel perception and production: A study of accent change in young adults," *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3814–3826, 2007.
[22] V. Fridland and T. Kendall, "Exploring the relationship between production and perception in the mid front vowels of US English," *Lingua*, vol. 122, no. 7, pp. 779–793, 2012.
[23] M. Sumner and A. G. Samuel, "The effect of experience on the perception and representation of dialect variants," *Journal of Memory and Language*, vol. 60, no. 4, pp. 487–501, 2009.
[24] C. Boberg, *The English language in Canada: Status, history and comparative analysis*. Cambridge University Press, 2010.
[25] C. Holland and T. Brandenburg, "Beyond the front range: The Coloradan vowel space," *Publication of the American Dialect Society*, vol. 102, no. 1, pp. 9–30, 2017.
[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.1.38," 2021. [Online]. Available: http://www.praat.org/
[27] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior research methods*, vol. 52, no. 1, pp. 388–407, 2020.
[28] J. R. de Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, 2015.
[29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.
[30] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://R-project.org/
[32] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

Acknowledgements: This work draws on research supported by the Social Sciences and Humanities Research Council of Canada.

"KRIECH NICHT DA REIN!" - A NEW CORPUS OF NATURALISTIC MISPERCEPTION OF GERMAN MISHEARD SUNG SPEECH

Jessica Nieder[1] & Kevin Tang[2*]

[1]Department of Linguistics & [2]Department of English Language and Linguistics

Heinrich-Heine University Düsseldorf, Germany; [2]Dept. of Linguistics, University of Florida, USA

nieder@phil.hhu.de, kevin.tang@hhu.de

ABSTRACT

Naturally-occurring misperception [1] can help establish the ecological validity of laboratory findings of speech perception and generate new hypotheses. In this study, we report on a corpus of misheard German sung speech which contains instances of misperception reported by individuals. We validated the corpus by examining segmental confusions, and word mis-segmentation. Approximately 1,000 segment confusions were found. Our naturalistic segment confusions were significantly correlated with acoustic distances (r = 0.559) and with speech-in-noise-induced confusions in an experimental study (vowel: r = 0.364; consonant: r = 0.210). Our mis-segmentation patterns only partially confirmed the rhythmic segmentation hypothesis [2] and findings from previous studies. While boundaries inserted before strong syllables created content words following the preferred rhythmic properties of German, we find an unexpected amount of boundary deletion before strong syllables, resulting in nonce percepts which might reflect the expectation of listeners with neologisms in lyrics [3].

Keywords: German lyrics, mondegreen, slips of the ear, phonetic similarity, rhythmic segmentation.

1. INTRODUCTION

In conversations, listeners are faced with the task of identifying speech segments and segmenting the speech stream they receive into meaningful words. When communication operates smoothly, one can only assume that the listener's understanding is identical to the utterance of the speaker, but it is only when communication breaks down that we are able to disentangle what was said from what was received [4]. These naturally-occurring instances of speech misperception or *slips of the ear* have the potential of revealing our speech perceptual processes and have been the focus of corpus-based or experimental studies over the past years, e.g., [1, 5, 2, 6, 7, 8, 9]. Using a corpus of naturalistic misperception of spoken English [1] demonstrated how naturalistic misperception can help establish the ecological validity of experimental studies (e.g., [10, 11, 12]). However, compared to spoken speech, the misperception of sung speech is much less well researched. Listening to sung speech is a very different activity as compared to listening to a conversation due to words and melodies being processed interactively [13]. When listening to sung speech, the nature of the context is more complex; it involves not only the listener's immediate surrounding and the activity that the listener is doing while listening, but also the general expectation of the artist and the genre of the song containing the sung speech [14]. Moreover, sung speech is produced differently from spoken speech in terms of duration, vowel formants [15], intonation [16] and pronunciation. It is almost always masked with music, while spoken speech may be masked with different noise types.

In this study, we present a new corpus of naturalistic misperceptions of misheard song lyrics by focusing on misperceptions of sung speech in German by German native speakers. We demonstrate how such a naturalistic corpus can shed light on patterns in segmental confusions, and word mis-segmentation. We examine the patterns in the light of existing findings from experimental and corpus-based studies. Our data is openly available at https://osf.io/xajvf/ and https://osf.io/acqp6/.

1.1. Speech segmentation in German

For English, a stress-based language, it has been established that listeners make use of their knowledge of the opposition of strong, i.e., stressed, vs. weak, i.e., unstressed syllables for inserting or

*We gratefully acknowledge Gerrit Kentner for sharing his data [18], the data contributions of the students in the linguistic seminar, the research assistance of Lukas Ostrowski, and the feedback of Ulrike Domahs and our anonymous reviewers.

deleting word boundaries in speech segmentation [5, 2, 11]. As a result, the rhythmic segmentation hypothesis was proposed that predicts a deletion of boundaries before weak and an insertion of word boundaries before strong syllables for English mis-segmentation data [2]. In German, the preferred rhythmic pattern is trochaic. Polysyllabic words show at least one stressed syllable, while stress in monosyllabic words shows an opposition between grammatical vs. lexical words: monosyllables are stressed in lexical words but can be unstressed in grammatical words such as clitics [17]. From this it follows that the same pattern of insertions before strong and deletions before weak syllables as compared to the English data is expected [17, 18]. Due to the tendency of unstressed monosyllables being grammatical words in German, a boundary insertion before weak syllables is expected to lead to the creation of a grammatical word [18] in case of mis-segmentations.

2. CORPUS

2.1. Compilation and annotation

For this study, a data set of 176 German misheard song lyrics was collected in a linguistics seminar at Heinrich-Heine University Düsseldorf and from three books containing a collection of German misheard sung speech [19, 20, 21][1]. To examine segmental confusions, we first obtained a phonetic transcription of the data using the CELEX pronunciation dictionary [22]. The transcriptions of the intended lyrics and the misheard lyrics were then segmentally aligned using an automatic alignment method (Pointwise-Mutual-Information-based Levenshtein distance, see [23]) that has been used to successfully align naturalistic misperceptions in English [1].

To examine word mis-segmentation, syllable stress was annotated using CELEX. A syllable was marked as strong if it had primary stress within the word it appears in, otherwise a syllable was marked as weak. Secondary stress was not marked because CELEX does not encode secondary stress and its existence is not well supported by acoustic evidence [24]. Subsequently, the data was coded manually for possible boundary misperception. Following [2, 18] we identified 1) **boundary insertions**, e.g., an additional word boundary in the perceived lyrics without a corresponding word boundary in the intended lyrics and 2) **boundary deletions**, e.g., a missing word boundary in the perceived lyrics that is available in the intended lyrics:

1. Griech - isch - er | Wein (original)
   Kriech | nicht | da | rein (perceived)
   'Greek wine/Do not crawl in there.'
2. Die | Crew | hat | da | noch | Fra - gen (original)
   Mir - ko | hat | da | noch | Fra - gen (perceived)
   'The crew/Mirko still has questions.'

69 data points were left for the mis-segmentation analysis.

3. STUDY I: SEGMENTAL CONFUSIONS

In the first study we focus on evaluating the segmental confusions found in misheard sung speech by comparing them with i) acoustic measurements and ii) speech-in-noise-induced segmental confusions.

3.1. Data for Validation

To validate the role of phonetic similarity in segmental confusions, we extracted the vowel frequency measurements of 16 vowels /a, ɪ, eː, ɛ, ɛː, ɪ, i̇ː, ɔ, oː, ʊ, uː, ʏ, yː, œ, øː, ə/ of 69 male and 58 female speakers from [25]. The Hertz frequencies were converted to Bark scale [26] to better match with human perception. The Euclidean distance was computed for each pair of vowels using the Bark frequencies of the first two formants. The distance was computed for men and women separately before averaging to obtain a single set of distances.

To demonstrate the ecological validity of laboratory research, we extracted the confusion data from two speech-in-noise experiments (one vowel, one consonant) from [27]. Ten listeners with normal hearing were tested in a close-set forced choice phoneme-recognition task. Ten vowels /a, aː, ɛ, e, ɪ, i, ɔ, o, ʊ, u/ were presented in a CVC frame. 14 consonants /p, t, k, b, d, ɡ, s, f, v, n, m, ʃ, ʦ, l/ were presented in a VCV frame. Five Signal-to-Noise ratios (SNRs) were examined (0, -5, -10, -15, and -20 dB). We focused on the confusion response rates for consonants and vowels at -15 dB SNR because they are least influenced by ceiling effects and have the largest variation across phonemes.

3.2. Results

995 segment confusions were found with 212 insertions, 252 deletions and 531 substitutions. Focusing on the vowel-vowel and consonant-consonant substitutions, two confusion matrices were tabulated and smoothed by adding 0.01 to all cells. The smoothed count matrices were transformed into confusion proportions. The proportion matrices were then converted into similarity matrices using the formula $S_{x,y} = (p_{x,y} +$

$p_{y,x})/2$, where $p_{x,y}$ is the proportion of times that the segment x was perceived as the segment y. The similarity matrices were then converted into distance matrices using $-ln(S_{x,y})$, a well-established metric for estimating the perceptual distance between two segments from similarity [28]. Similarly, the experimental confusion matrices were converted into distance matrices but using a more appropriate similarity metric $S_{x,y} = (p_{x,y} + p_{y,x})/(p_{x,x} + p_{y,y})$.

Pearson correlation was used to compare the global similarity of matrices. The statistical significance was evaluated using the Mantel test [29] with 10,000 permutations (upper-tailed) because sound distances are not completely independent. Our naturalistic segment distances were correlated with acoustic vowel distances (r = 0.559, p = 0.0001) and with the two sets of experimental-induced segment distances (vowel: r = 0.364, p = 0.0064; consonant: r = 0.210, p = 0.0245).

4. STUDY II: WORD MIS-SEGMENTATION

In the second study we focus on native word mis-segmentation and build on what has been shown by [18] for German in a cross-linguistic setting. In doing so, we test the predictions of the rhythmic segmentation hypothesis [2].

4.1. Data for Validation

To validate our findings we compare our data to four previous studies on mis-segmentation: [2] present natural and laboratory-induced mis-segmentations of continuous speech on English data and confirm the validity of the rythmic segmentation hypothesis for English native misperception (henceforth, ENG). Using experimental evidence, [7] investigate mis-segmentation in Dutch and confirm the predictions of the rhythm segmentation hypothesis for their data (henceforth, DUT). Focusing on mis-segmentation of English song lyrics by German native speakers (henceforth, ENG♪-GER♪), [18] shows that the rhythm segmentation hypothesis holds for non-native song perception. [30] presents experimental data for non-native (English-German) and native misperception (German-German). We took the German-German data from the supplementary materials of [30] and coded it for boundary type and syllable strength. This left us with 19 observations (henceforth, GER).

4.2. Results

Overall, we find more boundary deletions than insertions (42 vs. 27), a finding that is also confirmed by [18] for German in a cross-linguistic setting (ENG♪-GER♪). Table 1 displays the results of this study in comparison to the results of the aforementioned previous studies [2, 18, 7, 30].

Focusing on boundary insertions, more insertions before strong syllables are reported for ENG and DUT [2, 7]. For ENG♪-GER♪, the same amount of boundary insertions before strong as compared to weak syllables is found [18]. Interestingly, for German, this study and GER [30] shows the opposite pattern: we find more insertions before weak than before strong syllables. Focusing on boundary deletions, for ENG and ENG♪-GER♪ more deletions before weak than before strong syllables are reported [2, 18]. For DUT, roughly the same amount of boundary deletions before weak as compared to strong syllables is found [7]. For German, this study and GER show the opposite pattern, i.e., more boundary deletions before strong than before weak syllables [30] are observed.

Inspecting the boundary deletions in our data reveals that in 20 of 39 cases the deletion before a strong syllable created a nonce word, e.g., "*Flaggenhof*" for "*Flaggen hoch*" (nonce word vs. 'raise flags'). In addition, all deletions before weak syllables in our data created nonce words (3 of 3 cases). Compared to the deletion data, only 2 of the 27 cases of insertions created nonce words.

Omitting the nonces, and thus taking into account existing percepts only, leaves us with a total of 25 insertions (4 before strong syllables, 21 before weak syllables) and a total of 19 deletions (19 before strong). However, the general pattern of more insertions before weak and more deletions before strong syllables, contrary to the predictions of the rhythmic segmentation hypothesis, remains the same: $\chi^2 = 27.258$, df = 1, $p < .001$.

Another prediction of the rhythmic segmentation hypothesis states that an insertion before a weak syllable leads this weak syllable to likely be a grammatical word [18]. We observe the following pattern in our data of existing percepts that is displayed in Table 2.

In case of boundary insertions before strong syllables, our data supports the rhythmic segmentation hypothesis: If a boundary is inserted before a strong syllable, this strong syllable became a lexical word, e.g., "*Geduld ist ungesund*" (original) vs. "*Der Tod ist ungesund*" (perceived) 'Patience/The death is unhealthy'. However, if a word boundary is inserted before a weak syllable, the data does not support the rhythmic segmentation hypothesis: in only 3 cases of a total of 21 cases the weak syllable became a grammatical word,

Table 1: All boundary insertions and deletions before strong vs. weak syllables in our data as compared to data from four other studies. Chi-square test of independence for the relationship between boundary type and syllable strength: GER♪ (this study): $\chi^2 = 39.368$, df = 1, $p < .001$; GER: $\chi^2 = 0.30536$, df = 1, $p = 0.5805$; ENG: $\chi^2 = 22.484$, df = 1, $p < .001$; DUT: $\chi^2 = 16.208$, df = 1, $p < .001$; ENG♪-GER♪: $\chi^2 = 8.2073$, df = 1, $p < .005$.

	Insertion					Deletion				
	GER♪	GER	ENG	DUT	ENG♪-GER♪	**GER♪**	GER	ENG	DUT	ENG♪-GER♪
Strong	**4**	4	90	101	22	**39**	1	68	72	19
Weak	**23**	14	45	36	22	**3**	0	107	73	63

Table 2: Boundary insertions before weak vs. strong syllables leading to grammatical vs. lexical words within the group of existing percepts only.

	Grammatical	Lexical
Strong	0	4
Weak	3	18

e.g., "*lasset uns gemeinsam*" (original) - "*lasst uns gemein sein*" (perceived) 'Let's [...] together/Let's be mean'. In the remaining 18 cases, the weak syllable became a lexical word, e.g., "*Dies Kind soll unverletzet sein*" (original) - "*Dies Kind soll unser letztes sein*" (perceived) 'This child should be unharmed/our last'. While in these cases the weak syllable became a lexical word, contrary to the prediction, the syllable before the word boundary involves the creation of grammatical words in 7 of 11 cases, as in the following example:

3. wir | fah-ren | auf | Feu - er - rä - dern | Rich - tung | Zu - kunft (original)
   wir | fah - ren | auf | eu - ern | Rä - dern | Rich - tung | Zu - kunft (perceived)
   'we are riding on flaming bikes/your bikes into the future'

Here, the former weak syllable of the nonce word *Feuerrädern* in the original lyrics becomes a strong syllable in the onset of the new lexical word and the resulting percept is a correct grammatical phrase consisting of a grammatical word and a lexical word: *euern Rädern*.

5. DISCUSSION

Study I suggested that vowel confusions from misheard sung speech is strongly influenced by phonetics (r = 0.559). This is a surprising finding, since the confusion matrix was extracted regardless of any phonological environments and words. However, vowel and consonant confusions were only weakly correlated with experimental confusions (rs = 0.210-0.364). This could be due to how the experimental conditions differ greatly from the naturalistic one, such as the lack of music, and that segments were spoken (not sung) and were presented in VCV or CVC frames in isolation. Despite all the potential top-down influences and the influences of music on sung speech, listeners still heavily rely on phonetic/acoustic similarity during the processing of sung speech.

Study II demonstrated that the predictions of the rhythmic segmentation hypothesis cannot explain the pattern in our data. This raises the question as to why German native misperception behaves so differently from the data in previous studies. Against the predictions of the rhythmic segmentation hypothesis, we find more boundary deletions before strong than before weak syllables, often leading to a creation of new word forms or nonces. While the nature of these nonces remains a topic for further studies, a tentative explanation for their occurrence could lie in the importance of affective signals in perception, as has been shown by [14]. The nonces appear to be more humorous than the original lyrics, leading the listener to perceiving a more amusing nonce word instead of the original lyrics. Another reason for the amount of nonces lies in the nature of song lyrics. The production of song lyrics often use innovations such as neologisms for stylistic reasons or to express certain emotions [3]. The familiarity of the listeners in our study with nonces in song lyrics might have lead them to create neologisms themselves when misperceiving the original lyrics. These results open the path for further research on the listener's expectations in speech perception.

Taken together, this study reported on a new corpus of naturally-occurring misperception of sung speech. Despite its relatively small size and all the inherent differences with listening to spoken speech and sung speech, we demonstrated the role of both bottom-up (phonetics) and top-down (lexical expectation) factors in the processing of sung speech. This reinforces the idea of how examining our everyday perceptual errors has the potential to establish the ecological validity of laboratory findings of speech perception and generate new hypotheses [1, 5, 2, 6, 7].

6. REFERENCES

[1] K. Tang, "Naturalistic speech misperception," Ph.D. dissertation, University College London, 2015.

[2] A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *Journal of Memory and Language*, vol. 31, no. 2, pp. 218–236, 1992.

[3] V. Werner, "Love is all around: a corpus-based study of pop lyrics," *Corpora*, vol. 7, no. 1, pp. 19–50, 2012.

[4] J. Laver, "The production of speech," in *New Horizons in Linguistics*, J. Lyons, Ed. Harmondsworth: Penguin, 1970, pp. 53–75.

[5] K. Tang and A. Nevins, "Measuring segmental and lexical trends in a corpus of naturalistic speech," in *Proceedings of the 43rd Meeting of the North East Linguistic Society*, H.-L. Huang, E. Poole, and A. Rysling, Eds., vol. 2. GLSA (Graduate Linguistics Student Association), 2014, pp. 153–166.

[6] Z. Bond, *Slips of the Ear: Errors in the Perception of Casual Conversation*. Leiden, The Netherlands: Brill, 1999.

[7] J. Vroomen, M. Van Zon, and B. de Gelder, "Cues to speech segmentation: Evidence from juncture misperceptions and word spotting," *Memory & Cognition*, vol. 24, no. 6, pp. 744–755, 1996.

[8] M. A. Tóth, "A microscopic analysis of consistent word misperceptions," Ph.D. dissertation, Universidad del Paiś Vasco, 2017.

[9] R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri, "A corpus of noise-induced word misperceptions for english," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL458–EL463, 2016.

[10] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[11] M. S. Vitevitch, "Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear," *Language and Speech*, vol. 45, no. 4, pp. 407–434, 2002.

[12] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.

[13] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, p. 321, 2013.

[14] C. Beck, B. Kardatzki, and T. Ethofer, "Mondegreens and Soramimi as a Method to Induce Misperceptions of Speech Content â Influence of Familiarity, Wittiness, and Language Competence," *PLoS ONE*, vol. 9, no. 1, 2014.

[15] J. Sundberg, "Formant structure and articulation of spoken and sung vowels," *Folia Phoniatrica et Logopaedica*, vol. 22, no. 1, pp. 28–48, 1970.

[16] R. J. Zatorre and S. R. Baum, "Musical melody and speech intonation: Singing a different tune," *PLoS Biology*, vol. 10, no. 7, p. e1001372, 2012.

[17] P. Eisenberg, "Syllabische Struktur und Wortakzent. Prinzipien der Prosodik deutscher Wörter," *Zeitschrift für Sprachwissenschaft*, vol. 10, no. 1, pp. 37–64, 1991.

[18] G. Kentner, "Rhythmic segmentation in auditory illusions - evidence from cross-linguistic mondegreens," in *Proceedings of 18th ICPhS*. Glasgow: ICPhS, 2015.

[19] A. Hacke and M. Sowa, *Der weisse Neger Wumbaba: kleines Handbuch des Verhörens*, ser. Die Wumbaba-Trilogie. Kunstmann, 2004.

[20] ——, *Der weisse Neger Wumbaba kehrt zurück: zweites Handbuch des Verhörens*, ser. Die Wumbaba-Trilogie. Kunstmann, 2007.

[21] ——, *Wumbabas Vermächtnis: Drittes Handbuch des Verhörens*. Kunstmann, 2012.

[22] H. R. Baayen, R. Piepenbrock, and L. Gulikers, *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania, 1995.

[23] M. Wieling, J. Prokić, and J. Nerbonne, "Evaluating the pairwise string alignment of pronunciations," in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. Association for Computational Linguistics, 2009, pp. 26–34.

[24] F. Kleber and N. Klipphahn, "An acoustic investigation of secondary stress in German," Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel, Tech. Rep. 37, 2006.

[25] W. F. Sendlmeier and J. Seebode, "Formantkarten des deutschen Vokalsystems," TU Berlin, Institut für Sprache und Kommunikation, Tech. Rep., 2006.

[26] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 97–100, 1990.

[27] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2635–2648, 2009.

[28] R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.

[29] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[30] B. Voss, *Slips of the Ear: Investigations Into the Speech Perception Behaviour of German Speakers of English*, ser. Tübinger Beiträge zur Linguistik. G. Narr, 1984.

[1] We do not approve of the racist word in the titles.

VOICE DISCRIMINATION ACROSS SPEAKING STYLES IN PERSIAN

Elisa Pellegrino[1], Homa Asadi[2], Volker Dellwo[1]

[1]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
[2]University of Isfahan, Isfahan, Iran

{elisa.pellegrino; volker.dellwo}@uzh.ch; h.asadi@fgn.ui.ac.ir

ABSTRACT

Human listeners have a remarkable ability to recognize speakers by their voice, but within-speaker voice variability through different speaking styles, for example, can reduce recognition performance. In this study, we investigated voice discrimination across speaking styles in Persian. One hundred and forty-three naïve Persian listeners were asked to decide whether pairs of style-matched utterances in child-directed, spontaneous, read and clear speech originated from the same or different speakers. Listeners' performance across speaking styles was examined using the bias-free sensitivity measure A', and the bias measure b''D. Results showed that listeners performed accurately across all speaking styles but more poorly in child-directed speech. They had a bias toward responding 'different' regardless of the speaking style, thus suggesting a more general difficulty telling people together than apart.

Keywords: speaker discrimination, speaking styles, vocal variability.

I. INTRODUCTION

Human voices are individual and convey numerous cues about talkers' identity, including age, gender, health conditions, or emotional states [1]. Beyond being an "auditory face" [2], human voices also act as a behavioural biometric [3]. However, they are susceptible to strong variability for the effect of spontaneous or intentional modulations [4]. Human voice changes, for example, across the life span [5], as an effect of fatigue [6] or physical and mental health [7], [8]. Still, we also adapt the way we speak according to the age of our interlocutors [[9], [10], their linguistic origin [11], as well as to the background conditions [12]. For a review of volitional and unvolitional changes see [13]. How human listeners can distinguish speakers despite the enormous variability that individual voices reveal is far from being fully understood. In this study, we investigated the effect of the variability introduced to voices by speaking styles on unfamiliar speaker discrimination. What are the factors affecting voice processing? It has been amply documented that human listeners are fairly accurate at distinguishing unfamiliar speakers, especially if exposed to long, phonetically and prosodically rich utterances [14]–[17]. Nevertheless, studies on speaker perception and forensic speaker identification have shown that linguistic and extra-linguistic sources of within-speaker variability challenge the perception of speaker identities. More specifically, it has been observed that listeners' accuracy to recognize or discriminate between unfamiliar voices is reduced by mismatch in the language spoken in the stimulus pairs [18][19], type of vocalizations (vowels vs laughter, [20]; sung vs spoken word [21]) and in the linguistic relationship between lexical items within-one stimulus pair [22]. Detrimental to speaker perception are also changes in degree of speakers' vocal expressiveness [23] and in voice quality obtained by disguising voice source or resonance features (e.g. creaky voice, falsetto, hyponasality, whispering [24] [25] [26].

Another factor affecting vocal identity processing, that has recently received more scholarly attention, is speaking style. Numerous studies have indeed focussed on the effect of style-matched and style mismatched conditions on the speaker discrimination abilities in humans and machines. Style matched trials typically featured read-read speech only [27] or both read-read and spontaneous-spontaneous speech [28], [29], [30]. Style mismatched trails, instead, combined read speech with spontaneous speech [27]–[30] or with pet-directed speech [31]. Overall findings from this line of research have revealed that both humans and machines performed better when the speaking style was coherent within one stimulus pair compared to when it changed. Moreover, between the style-matched conditions (read-read, spontaneous-spontaneous), the read–read condition scored the highest accuracy. Taken together, results suggest that less controlled speaking styles (e.g. conversational or free speech) and the exaggerated prosody of pet-directed speech introduce more variability in the speakers' acoustic spaces [29] that make it difficult for listeners to establish stable identity percepts. Lower accuracy in style mismatched as compared to matched automatic voice recognition, yet with novel

insights on the role of high variable speaking styles on speaker perception is also documented in [32]. In [31], speaking styles under comparisons were clear speech and conversational speech on the one hand, adult- and infant-directed speech on the other. Between the matched conditions, mean within-register recognition led to ceiling effects for adult-directed, clear and conversation speech. Slightly lower accuracy was documented for infant-directed speech. More interestingly, in the style mismatched conditions, it was observed that when the systems were trained with more acoustically variable speaking styles (spontaneous or infant-directed speech) and tested with less variable registers (clear speech or adult-directed speech), the system performed better as compared to the opposite conditions (training with clear or adult-directed speech and testing respectively with spontaneous and infant-directed speech) [32]. This suggests that — at least for machines — learning vocal identity from acoustically more variable speaking styles may also lead to speaker recognition benefits.

To shed further light on the effect of speaking styles on speaker perception, we tested the speaker discrimination performance of Persian listeners on stimuli produced in most of the speaking styles employed in previous research: read, clear, spontaneous and child-directed speech, this latter being comparable to pet directed speech in many relevant features [33]. In line with findings showing that listeners' performance degrades with the more inherently acoustically variable speaking styles ([28]–[31], we hypothesize that listeners are more accurate in clear and read speech than in spontaneous or infant-directed speech. If this hypothesis holds, we also expect listeners to be more likely to incorrectly assign intraspeaker variability to different individuals (telling people together) in spontaneous and child-directed speech, that should measurably result in a bias towards responding 'different'. Based on the novel observations that speaker-specific acoustic variability can instead benefit voice processing by machines [32] we cannot fully exclude that listeners may perform better in spontaneous and child-directed speech and exhibit a bias toward responding the same (e.g. incorrectly assign interspeaker variability to same individuals) in clear and adult directed speech.

2. MATERIALS AND METHOD

2.1. Speakers

52 male Persian speakers, aged between 25 and 45, y.o. (Mean=29.66 , SD=4.8) were recruited for the collection of the database of within-speaker speaking style variability. Participants were BA, MA or PhD students at the University of Isfahan. All participants were monolingual with no language other than Persian spoken in their homes. They were all from the Isfahan province where they had lived their entire lives. None of the speakers reported any history of speech or hearing disorders and they were all naïve as to the purposes of the experiment. Participants from the Department of Linguistics were rewarded a grade, while the others received monetary compensation.

2.2 Speech Materials

To study the effect of speaking style on speaker discriminability, we collected a speech corpus of child-directed, read, clear and spontaneous speech. For the read speech part, each speaker was instructed to read aloud a list of 20 sentences, at their natural pace and intonation, with a pause between each sentence. They were also told to repeat any disfluent sentences before moving on. For the clear speech part, the same set of sentences was used but participants were asked to imagine that they were talking to hard of hearing persons. Speakers had several attempts for this task, and we eventually selected the repetition of sentences which sounded more clearly enunciated. To elicit between 2-4 minutes of spontaneous speech, a set of topics was developed (e.g. study field, recent vacations, and plans for the future). Speakers were instructed to select one or two topics which they would be comfortable speaking about. To elicit child-directed speech, participants were shown a picture of a baby and asked to tell a story to him. The speaking styles together yielded between 7 to 9 minutes of speech per speaker for a total corpus of 416 minutes of speech.

Due to COVID restrictions, speech data collection was carried out remotely, in a quiet room at participants' home, via vocal messages through the WhatsApp application. To control for the effect of differences in mobile equipment in the quality of recording, the corpus was collected exclusively using one specific smartphone brand and model. Recording instructions, sentences and pictures were sent to the speakers prior to the recording session. To supervise the data collection procedure, the recordings were taken during video call on ZOOM in the presence of the experimenter (2nd author). Recording sessions for each speaker took place on the same day. To permit acoustic analysis in Praat, participant's audio tracks were converted from their original format supported by the device (.m4a) to .wav format using the free online app Online Audio Converter but the original sampling rate and bit depth (44 kHz; 32 bit) were not modified. Eighty audio clips of about 3 seconds were extracted from each speech samples of the dataset. Each excerpt was selected such that the influence of

semantic cues to speaking styles on listeners' responses was minimized.

2.2 Listeners

143 native Persian listeners (male=75, female=68), ranged in age from 18 to 36 y.o. participated in the speaker discrimination test. The listeners were all students at the University of Isfahan. None was a trained phonetician. All listeners declared not to have hearing, vision, and/or dyslexia problems.

2.3. Procedure

Listeners were tested individually in a quiet room at their home. Their task was to decide whether the two stimuli in a trial were spoken by the same speaker or by different speakers (voice discrimination task). Each stimulus in a stimulus pair was scaled to an average intensity of 70 dB SPL. The stimuli in each stimulus pair were separated by a 1s silent interstimulus interval. Each listener was presented with 40 stimulus pairs (10 per speaking styles) over high-quality earphones, equally divided by same and different speakers trials. The stimulus order and item pairs were randomised for each participant but the stimulus set was identical to all listeners. The pairing of voices in different-speaker trials was random. Experiment was designed through learning management system (LMS) of the University of Isfahan, which provided an accessible online platform for conducting behavioural studies. Listeners were shown the "same" and "different" options on the screen and asked to click on the corresponding button after listening to each trial. Listeners were familiarised with the experiment interface and the stimuli through a demonstration session which presented four random stimuli containing voices and lexical items not present in the stimulus set. Testing was self-paced; participants generally took approximately 5-8 minutes to complete the experiment. Participants used their in-built microphones and headphones when doing the experiment. The study was conducted within the guidelines of the Ethics Committee of the University of Isfahan. Participants gave their informed consent to participate in the study.

2.4. Data Analysis and Statistics

To analyse the effect of speaking styles on listeners' ability to discriminate between same and different speakers, we calculated the bias-free sensitivity measure A' from signal detection theory [34]. A' values range from 0.0 to 1.0, with 0.5 signifying chance level sensitivity and 1.0 indicating highest sensitivity. A' sensitivity measure (dependent variable was calculated per listener and speaking style. To examine the effect of speaking style on listeners bias towards responding the same or different, we also calculated the measure b''D per listeners and speaking styles. B''D scores range between 1 and -1, with negative values indicating a bias toward responding 'different speakers', and positive values 'same speaker'. To test the significance of the effect of speaking style on A' and b''D, we ran Linear Mixed Effect Model with Speaking Styles as a fixed factor, A' and b''D as dependent variables, listeners as random intercept. Statistical analyses were performed with R Core Team 2022 [35].

3. RESULTS

As shown in Fig. 1, speaker discrimination was performed accurately in all speaking styles, albeit to a different extent. The lowest sensitivity was scored in stimuli produced in child-directed speech followed, in increasing order of sensitivity, by read, spontaneous and clear speech. The results of statistical analysis confirmed in part these observations: the effect of speaking style on A'[χ^2 (3)=22.162, p <0.001] was significant. Post-hoc analysis with Tukey corrections was conducted that showed that only the comparisons including stimuli in child-directed speech were significant (Fig. 2). The negative sign of the estimate indicated a poorer performance in child-directed speech as compared to clear, read and spontaneous speech. Concerning the bias, Fig. 3 shows that the scores of b''D were mostly negative for all the examined speaking styles, thus pointing to a more general bias towards responding 'different'. This pattern was confirmed by the statistical analysis that revealed no significant changes in b''D between read, clear, spontaneous and child-directed speech [χ^2(3)=6.7, p = 0.082].

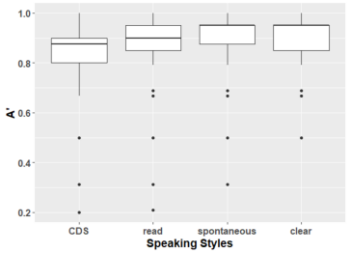


Figure 1: A' values by Speaking Style. The styles in the plot are ordered by increasing listeners' sensitivity

```
$contrasts
 contrast           estimate     SE  df t.ratio p.value
 CDS - clear       -0.05490 0.0127 435  -4.333  0.0001
 CDS - read        -0.03683 0.0127 435  -2.906  0.0200
 CDS - spontaneous -0.04872 0.0127 435  -3.845  0.0008
 clear - read       0.01808 0.0127 435   1.427  0.4833
 clear - spontaneous 0.00618 0.0127 435  0.488  0.9618
 read - spontaneous -0.01190 0.0127 435 -0.939  0.7839

Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 4 estimates
```

Figure 2: Results of post-hoc test with Tukey correction for multiple comparisons across speaking styles.

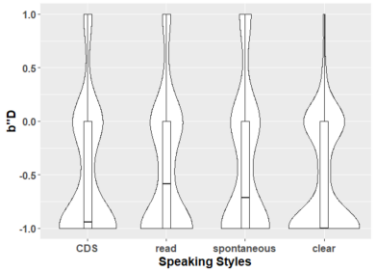


Figure 3: b''D values by Speaking Style.

5. DISCUSSION

In this study, we investigated the effect of speaking styles on Persian listeners' ability to discriminate between speakers. We quantified listeners' performance in terms of the bias-free sensitivity measure A' and the bias measure b''D. The overall results of this study confirm that speaking style variability affects the speaker discrimination performance. The direction of such effect, i.e. the lower sensitivity in child-directed speech, supports previous research indicating that inherently more variable speaking styles pose challenges to speaker discrimination [28]–[31]. Surprisingly though, spontaneous speech did not worsen listeners' performance compared to read speech as expected and reported in previous research [28]–[30]. No significant differences in A' were, indeed, obtained between clear, spontaneous and read speech. What causes this discrepancy is unclear and will be object of future acoustic investigations. Here we can only speculate that with the controlled elicitation method used for this study we may have not obtained a clear acoustic divide between clear, spontaneous and read speech, which results in measurably comparable difficulty when trying to discriminate between unfamiliar speakers. Alternatively, individual speakers may have produced the four speaking styles differently, with some of them marking the inter-style differences to a higher degree as compared to others (cf. [36] for individual differences in spontaneous and read speech). Within-speaker acoustic variability across speaking styles draws paths for future acoustic investigations.

Another unpredicted finding of this study was the listeners' bias towards responding different, irrespective of speaking styles. This type of bias was expected only for the more inherently variable speaking styles of the corpus but not for clear and read speech. Such bias indicates that participants were more likely to incorrectly assign intraspeaker variability to different individuals than they were to incorrectly assign inter-speaker variability to the same individual. We propose here two alternative non-exclusive explanations for this finding. First, one may impute the observed bias to the cautious behaviour of our listeners that - in line with previous research with not degraded stimuli nor embedded in background noise [26] seem to prefer responding different when unsure about the origin of intra-stimulus voices (for different results with degraded stimuli see [26]). This result also echoes recent research showing that listeners have more difficulties telling people together than apart (see a.o. [12]). The alternative interpretation is that the stimuli used in the perception contain comparable within-speaker variability that induced listeners to mistake within for between-speaker variability. Because the acoustic signal is the input to human perceptual processes, future step of this research will examine the acoustic properties in the stimuli used in the perception test to understand the perceptual strategies used by Persian listeners when discriminating between unfamiliar speakers in different speaking styles.

6. ACKNOWLEDGEMENTS

This study was supported by a Swiss Research Partnership Grant with South Asia and Iran funded by the Zürcher Hochschule für Angewandte Wissenschaften (ZHAW).

7. REFERENCES

[1] V. Dellwo, M. Huckvale, and M. Ashby, "How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification," in *Speaker Classification I*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–20.

[2] P. Belin, S. Fecteau, and C. Bédard, "Thinking the voice: neural correlates of voice perception," *Trends Cogn Sci*, vol. 8, no. 3, pp. 129–135, Mar. 2004.

[3] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *J Acoust Soc Am*, vol. 144, no. 1, pp. 375–386, Jul. 2018, doi: 10.1121/1.5045323.

[4] Y. Lee and J. Kreiman, "Acoustic voice variation in spontaneous speech," *J Acoust Soc Am*, vol. 151, no. 5, pp. 3462–3472, May 2022, doi: 10.1121/10.0011471.

[5] B. V. Tucker, C. Ford, and S. Hedges, "Speech aging: Production and perception," *WIREs Cognitive Science*, vol. 12, no. 5, Sep. 2021, doi: 10.1002/wcs.1557.

[6] N. R. Williams, "Occupational groups at risk of voice disorders: a review of the literature," *Occup Med (Chic Ill)*, vol. 53, no. 7, pp. 456–460, Oct. 2003, doi: 10.1093/occmed/kqg113.

[7] H. Steurer, E. Schalling, E. Franzén, and F. Albrecht, "Characterization of Mild and Moderate Dysarthria in Parkinson's Disease: Behavioral Measures and Neural Correlates," *Front Aging Neurosci*, vol. 14, May 2022.

[8] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed

people: a cross-situation study," *BMC Psychiatry*, vol. 19, no. 1, p. 300, Dec. 2019.
[9] S. Kemper and S. Kemper, "Elderspeak : Speech accommodations to older adults Elderspeak : Speech Accommodations to Older Adults," vol. 5585, 2007.
[10] C. Kitamura, C. Thanavishuth, D. Burnham, and S. Luksaneeyanawin, "Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language," *Infant Behav Dev*, vol. 24, no. 4, pp. 372–392, 2001, doi: 10.1016/S0163-6383(02)00086-3.
[11] E. Lattey, "Interlinguistic Variation and Similarity in Foreigner Talk," in *The Dynamic Interlanguage*, Boston, MA: Springer US, 1989, pp. 87–100.
[12] M. Garnier, N. Henrich, and D. Dubois, "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect," *Journal of Speech, Language, and Hearing Research*, v 53, no. 3, pp. 588–608, 2010
[13] N. Lavan and R. Holloway, "Flexible voices : identity perception from variable vocal signals," *Psychonomic Bulletin & Review*, pp. 1–27, 2013.
[14] S. Cook and J. Wilding, "Earwitness Testimony: Never Mind the Variety, Hear the Length," *Appl Cogn Psychol*, vol. 11, no. 2, pp. 95–111, Apr. 1997.
[15] R. Roebuck and J. Wilding, "Effects of vowel variety and sample lenght on identification of a speaker in a line-up," *Appl Cogn Psychol*, vol. 7, no. 6, pp. 475–481, Nov. 1993, doi: 10.1002/acp.2350070603.
[16] G. E. Legge, C. Grosmann, and C. M. Pieper, "Learning unfamiliar voices.," *J Exp Psychol Learn Mem Cogn*, vol. 10, no. 2, pp. 298–303, Apr. 1984, doi: 10.1037/0278-7393.10.2.298.
[17] P. D. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," *J Acoust Soc Am*, vol. 40, no. 6, pp. 1441–1449, Dec. 1966, doi: 10.1121/1.1910246.
[18] M. Wester, "Talker discrimination across languages," *Speech Commun*, vol. 54, no. 6, pp. 781–790, 2012, doi: 10.1016/j.specom.2012.01.006.
[19] J. M. Zarate, X. Tian, K. J. P. Woods, and D. Poeppel, "Multiple levels of linguistic and paralinguistic features contribute to voice recognition.," *Sci Rep*, vol. 5, p. 11475, 2015, doi: 10.1038/srep11475.
[20] N. Lavan, B. Short, A. Wilding, and C. McGettigan, "Impoverished encoding of speaker identity in spontaneous laughter," *Evolution and Human Behavior*, vol. 39, no. 1, pp. 139–145, Jan. 2018, doi: 10.1016/j.evolhumbehav.2017.11.002.
[21] Z. F. Peynircioğlu, B. E. Rabinovitz, and J. Repice, "Matching Speaking to Singing Voices and the Influence of Content," *Journal of Voice*, vol. 31, no. 2, pp. 256.e13-256.e17, Mar. 2017, doi: 10.1016/j.jvoice.2016.06.004.
[22] C. R. Narayan, L. Mak, and E. Bialystok, "Words Get in the Way: Linguistic Effects on Talker Discrimination," *Cogn Sci*, vol. 41, no. 5, pp. 1361–1376, Jul. 2017, doi: 10.1111/cogs.12396.
[23] N. Lavan, L. F. Burston, P. Ladwa, S. E. Merriman, S. Knight, and C. McGettigan, "Breaking voice identity perception: Expressive voices are more confusable for listeners," *Quarterly Journal of Experimental Psychology*, vol. 72, no. 9, pp. 2240–2248, 2019.
[24] A. R. Reich and J. E. Duke, "Effects of selected vocal disguises upon speaker identification by listening," *J Acoust Soc Am*, vol. 66, no. 4, pp. 1023–1028, 1979.
[25] A. Hirson and M. Duckworth, "Glottal fry and voice disguise: a case study in forensic phonetics," *J Biomed Eng*, vol. 15, no. 3, pp. 193–200, May 1993.
[26] A. Bartle and V. Dellwo, "Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech," *International Journal of Speech, Language and the Law*, vol. 22, no. 2, pp. 229–248, 2015.
[27] H. M. J. Smith, T. S. Baguley, J. Robson, A. K. Dunn, and P. C. Stacey, "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance," *Appl Cogn Psychol*, vol. 33, no. 2, pp. 272–287, Mar. 2019, doi: 10.1002/acp.3478.
[28] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for 'easy' versus 'hard' voices in style-matched and -mismatched speech," *J Acoust Soc Am*, vol. 151, no. 2, pp. 1393–1403, Feb. 2022, doi: 10.1121/10.0009585.
[29] A. Afshan, J. Kreiman, and Alwan Abeer, "Speaker discrimination in humans and machines: Effects of speaking style variability," in *INTERSPEECH 2020*, Shangai, China, 2020.
[30] S. v. Stevenage, R. Tomlin, G. J. Neil, and A. E. Symons, "May I Speak Freely? The Difficulty in Vocal Identity Processing Across Free and Scripted Speech," *J Nonverbal Behav*, vol. 45, no. 1, pp. 149–163, 2021
[31] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *J Acoust Soc Am*, vol. 144, no. 1, pp. 375–386, Jul. 2018, doi: 10.1121/1.5045323.
[32] T. Kathiresan *et al.*, "Mothers Reveal More of Their Vocal Identity When Talking to Infants," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4088888.
[33] L. Lansade, M. Trösch, C. Parias, A. Blanchard, E. Gorosurreta, and L. Calandreau, "Horses are sensitive to baby talk: pet-directed speech facilitates communication with humans in a pointing task and during grooming," *Anim Cogn*, vol. 24, no. 5, pp. 999–1006, Sep. 2021, doi: 10.1007/s10071-021-01487-3.
[34] J. B. Grier, "Nonparametric indexes for sensitivity and bias: Computing formulas.," *Psychol Bull*, vol. 75, no. 6, pp. 424–429, 1971, doi: 10.1037/h0031246.
[35] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J Stat Softw*, vol. 67, no. 1, 2015.
[36] G. P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Commun*, vol. 22, no. 1, pp. 43–65, Jul. 1997, doi: 10.1016/S0167-6393(97)00012-5.

THE EFFECT OF SPEAKER ON SPEECH INTELLIGIBILITY

Richard Wright[1], Benjamin V. Tucker[2,3], Matthew C. Kelley[1]

[1]University of Washington, [2]Northern Arizona University, [3]University of Alberta
rawright@uw.edu, benjamin.tucker@nau.edu, mattck@uw.edu

ABSTRACT

Most perception and intelligibility studies assume that noise masks spoken-sentence stimuli equivalently regardless of the speaker. That is, any sentence is equally intelligible across speakers when controlled for style and rate. However, previous studies have shown significant intelligibility variation between speakers. In this study, we predict speaker-specific intelligibility differences in noise for 720 IEEE/Harvard sentences produced by 20 Pacific Northwest speakers (10 women and 10 men). In this corpus, style and rate are controlled in the productions to avoid clear speech. Each sentence recording ($n = 14,400$) was mixed with corpus-shaped noise at three levels (+2, 0, -2 dB) resulting in 43,200 total stimuli. Sentence intelligibility was calculated from orthographic transcriptions provided by 1,868 English-speaking listeners to 120 randomly selected stimuli presented online. We find that different speakers have different intrinsic intelligibility even when reading the same set of sentences. Moreover, some speakers' speech was more affected by noise than others.

Keywords: speech intelligibility, speaker intelligibility, IEEE Corpus, speech in noise.

1. INTRODUCTION

The use of natural speech, as opposed to synthetic speech or pure tones, is well motivated in perceptual testing and hearing-related testing [1]. One might even say that *a word is worth a thousand pure tones*. However, the greater ecological validity of natural speech brings greater variation to the test or experiment. Previous studies have found individual differences in speaker intelligibility related to varying tasks [2, 3]. Even when controlling for task, individual speaker differences in intelligibility are strong enough to overwhelm other factors such as dialect differences [4]. Despite these previous findings, most perception and intelligibility studies implicitly assume that noise masks spoken-sentence stimuli equivalently regardless of the speaker. That is, any one sentence is equally intelligible across speakers when controlled for style and rate. This is particularly problematic if a small number of speakers are used in the stimuli and especially when different listeners are presented with different speaker-stimuli. While there has been significant work on individual speaker differences in clear speech [5, 6], with a few exceptions [7] less work has been done on individual differences in normal speaking styles with controlled materials.

In this study, we examine whether sentences are equivalently intelligible across speakers using the same speaking style. Our hypothesis is that there will be a significant effect of individual speakers on speech intelligibility in noise. Furthermore, we predict that there will be speaker-by-noise interactions resulting in non-linear intelligibility differences at different noise levels. To test this, we used an online perceptual-transcription procedure presenting stimuli from the Pacific Northwest subset of the UWNU IEEE Corpus [8] in three levels of noise. We chose this corpus because it has a large number of sentence stimuli read by a relatively large number of speakers ($n = 20$) and because the recordings were carefully controlled for regional accent and curated to select for equivalent, non-hyperarticulated, speech styles. In examining the results we used individual speaker, English as a first language, and noise as independent variables.

2. METHODS

To investigate the intelligibility across speakers, we conducted a massive online perception study where we asked listeners to orthographically transcribe what they heard.

2.1. Stimuli

The stimuli for this experiment came from recordings of 720 IEEE ('Harvard') sentences [8]. All 720 IEEE sentences were read by 20 native English Speakers (10 women, 10 men) originating from the Pacific Northwest (Washington, Oregon, & Idaho). The recording of the IEEE sentences was controlled for speaking style and rate to avoid the production of hyperarticulated or clear speech. This resulted in a total of 14,400 items. Each recorded

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
NoiseCond	7745201.56	3872600.78	2.00	178440.16	78403.12	0.0001
English FL	86248.14	86248.14	1.00	1819.56	1746.15	0.0001
Speaker ID	622507.55	32763.56	19.00	179752.55	663.32	0.0001
NoiseCond×English FL	40903.00	20451.50	2.00	178468.19	414.05	0.0001
English FL×Speaker ID	12636.05	665.06	19.00	179778.33	13.46	0.0001
NoiseCond×Speaker ID	266782.01	7020.58	38.00	178257.03	142.14	0.0001

Table 1: ANOVA table of the final linear mixed-effects model. NoiseCond = Noise Condition, English FL = English is first language.

item was subsequently masked using steady corpus-shaped noise at three different signal-to-noise ratios: +2 dB (N1), 0 dB (N2), -2 dB (N3). After the masking manipulation was applied, the total number of stimuli was 43,200.

2.2. Listeners

In the present analysis, 1,868 English-speaking adults were recruited as listeners from the University of Alberta and the University of Washington. Participants received course credit for their participation in the experiment. Participants' age ranged from 16-70 years-old with a mean age of 20.93 (standard deviation: 4.11). A total of 1,143 listeners reported English as their first language and 752 reported another first language. In addition, 1,242 participants reported their gender as female, 589 as male, and 37 as other.

2.3. Procedure

Data was collected using a custom online experiment, presented using a browser, over a two-year period (Nov. 2020 – Nov. 2022). At the beginning of each experimental session, participants were asked to wear headphones and then completed a demographic questionnaire. Each participant was presented with 120 randomly selected test sentences from the full list of stimuli and was asked to type each sentence to the best of their ability. Sentence lists were controlled so that no sentence was repeated in any given experimental session.

There are many potential measures of intelligibility that can be used to investigate speech intelligibility [9, 10, 11]; we chose Levenshtein Distance (LD) to provide a general measure of accuracy. This is a string-edit metric for calculating the number of edits needed to change one string into another. For each response, we calculated the Levenshtein Distance from the participants' typed response to the correct response across the entire sentence [12, 11]. All responses that were blank or contained some variant of "I don't know" were removed from the data leaving just over 180,000 total responses.

2.4. Statistical Analysis

A linear mixed effects analysis was performed using the lme4 package [13] in R [14]. With listener and item as random intercepts, we used the calculated Levenshtein distance as the dependent variable with predictors of `Noise Condition` (NoiseCond: *N1*, *N2*, *N3*) `Speaker` (20 speakers), and whether English was the listener's first language, `English is First Language` (English FL: *yes* or *no*). We also included all possible two 2-way interactions of `English as a First Language`, `Noise Condition`, and `Speaker`. While `Speaker` as a predictor has 20 levels and is a bit difficult to interpret, we chose to include it as a predictor in the models as opposed to a random effect to allow for an investigation of speaker differences in terms of listener accuracy.

3. RESULTS

The statistical model is summarized in the ANOVA table in Table 1 of the linear-mixed effects model. An ANOVA table was selected to summarize the model effects without printing all the interactions in a coefficient table of 20 levels of `Speaker`. The model indicates that all interactions were significant, as were the main effects.

The results indicate that listeners' accuracy decreases as the SNR becomes worse (0dB: $\hat{\beta} = 9.28$, SE=0.18, $t = 51.41$; -2dB: $\hat{\beta} = 13.93$, SE=0.18, $t = 76.09$). When `English is First Language` listeners are more accurate ($\hat{\beta} = -7.63$, SE=0.26, $t = -29.12$) and the effect of `Speaker` is also significant. However, there are also two two-way interactions in the model. The results of the interaction between `English is First Language` and `Noise Condition` are illustrated in

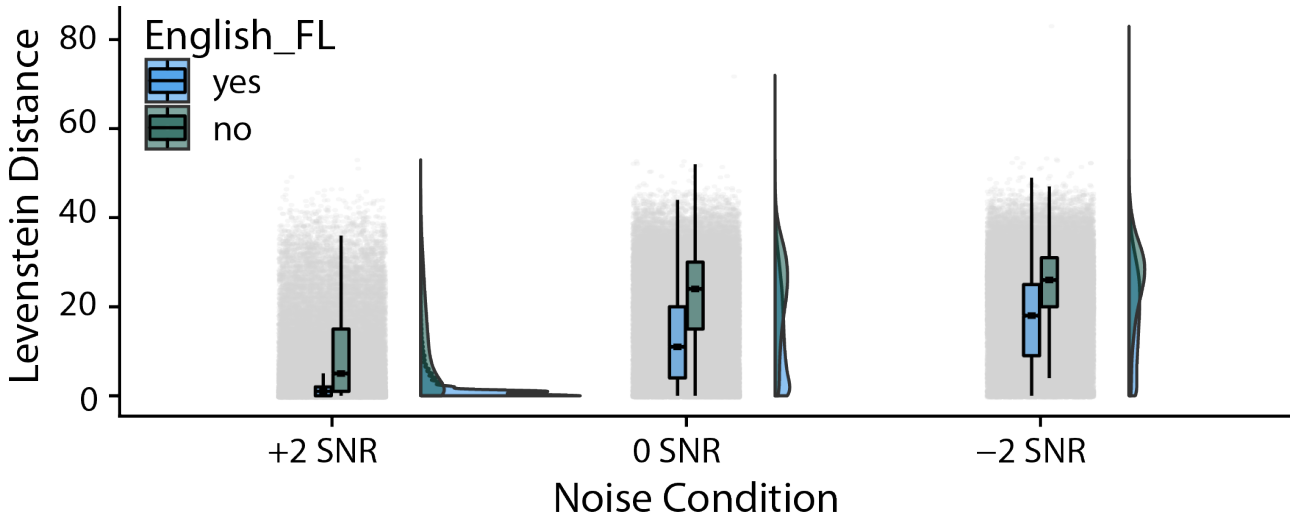


Figure 1: Raincloud plot of interaction between `Noise` and `English is First Language`.

Figure 1. We see that listeners with English as their first language are more accurate than listeners with other first languages and that the difference between these two groups increases as the SNR is degraded. The interaction between `English is First Language` and `Speaker` is also significant.

Figure 2 illustrates the main effects of `Speaker` and the three different noise conditions. It is clear that the individual speakers differ in terms of general intelligibility with some speakers being significantly more intelligible than other speakers. For example, PNM055 and PNF143 are the two most intelligible speakers in all noise conditions while PNM085 and PNF142 are the two least intelligible. The statistical model does not show a speaker gender effect; we simply find that some speakers are more intelligible than others.

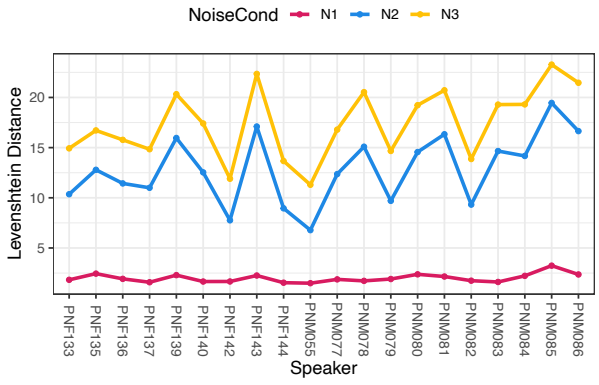


Figure 2: Levenshtein Distance split by talker and noise condition. The 10 speakers on the left are female and the 10 speakers on the right are male.

There is also a significant interaction between `English is First Language` and `Speaker`. Since there are 20 speakers and thus 20 levels in the model it is difficult to visualize the interaction between `Noise Condition` and `Speaker`. Figure 3 illustrates the least intelligible and the most intelligible speakers and their Levenshtein distance for individual items. Items have been sorted in this figure based on their average accuracy. It is interesting to observe the massive impact of increased noise in the signal for the least intelligible speaker while that impact is greatly reduced for the most intelligible speaker.

4. DISCUSSION

In the present study, we used a massive data approach (43,200 total stimuli, 1,868 listeners, and over 180,000 total data points after cleaning) to probe the effect of an individual speaker on sentence intelligibility. The advantage of a large data set like this is that we have increased power which allows us to truly investigate individual speaker effects. We replicated previous studies' findings that there is no reliable effect for speaker-gender on intelligibility [15, 16]. We saw a significant effect of English L1 on intelligibility scores, as expected with lower scores for L2 English listeners [17, 18] and interactions with both noise and speaker (see Figure 1. The main hypothesis, that we would see a significant effect of speaker on sentence intelligibility was confirmed (see Figures 2 & 3).

As is illustrated in Figure 3, showing the most and least intelligible speakers, there are large differences in LD scores and an interaction of speaker with noise. The interaction of speaker with

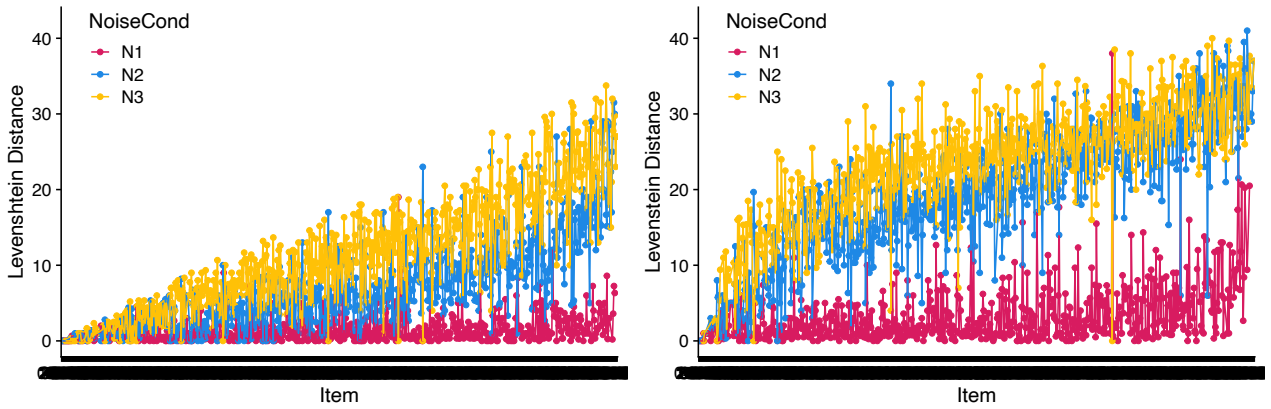


Figure 3: Levenshtein distance for each item (sorted by average Levenshtein distance) split across the three noise conditions for the most intelligible speaker (top, PNM085) and the least intelligible speaker (bottom, PNM055).

noise is important because it indicates a non-linear degradation of speech intelligibility with increased masking noise. That is, a low-intelligibility speaker is masked more effectively than a high-intelligibility speaker even when the linguistic content of the speech is controlled. We also observe with the least intelligible speaker that the difference between N2 and N3 in the noise condition is diminished while it seems to be maintained for the most intelligible speaker. There also appear to be speaker and noise interactions (see Figure 3) with individual sentences, which remains to be investigated in future work with this dataset.

Previous research calculating different measures of intelligibility, including phoneme and feature distance measures [9, 10], makes us confident that the LD distances are representative of listeners' sentence perception. Future work will include analysis of the acoustic characteristics of the individual talkers using vowel space, intensity, fundamental frequency, and dynamic measures [2, 19, 4]. It will also include an investigation of the effect of sentence material on talker intelligibility; as can be seen in Figure 3, there is a sharper rise for the least intelligible speaker in LD scores across sentences (items), hinting that there may be talker-by-sentence interactions.

One implication of this study is that when you recruit a speaker for an intelligibility experiment, or when you select a speaker from a corpus, you don't know where they land on the spectrum of speaker intelligibility and how stimuli based on their speech will interact with masking noise. It is also possible that when sampling a small number of speakers, for example only 2-4, effects of individual speaker intelligibility may inadvertently influence a generalization about gender or dialect. There is a speaker-by-second language interaction that suggests the need for more research on the impact of speaker and noise on second language speakers. It is also possible that a speaker's intelligibility may interact with L2 learners' proficiency or age of acquisition.

5. ACKNOWLEDGMENTS

This work was supported by NIH NIDCD R01 DC006014.

6. REFERENCES

[1] T. Beechey, "Ecological Validity, External Validity, and Mundane Realism in Hearing Science," *Ear and Hearing*, vol. 43, no. 5, pp. 1395–1401, Oct. 2022.

[2] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, no. 3–4, pp. 255–272, Dec. 1996.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing I," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 96–103, Mar. 1985.

[4] D. R. McCloy, R. A. Wright, and P. E. Souza, "Talker Versus Dialect Effects on Speech Intelligibility: A Symmetrical Study," *Language and Speech*, vol. 58, no. 3, pp. 371–386, Sep. 2015.

[5] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *The Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 272–284, Jul. 2002.

[6] S. H. Ferguson and D. Kewley-Port, "Talker Differences in Clear and Conversational Speech: Acoustic Characteristics of Vowels," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 5, pp. 1241–1255, Oct. 2007.

[7] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, Jan. 2004.
[8] L. Panfili, J. Haywood, D. R. McCloy, P. Souza, and R. A. Wright, "The UW/NU Corpus 2.0," Tech. Rep., 2017. [Online]. Available: https://depts.washington.edu/phonlab/projects/uwnu.php
[9] E. Felker, M. Ernestus, and M. Broersma, "Evaluating Dictation Task Measures for the Study of Speech Perception," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Melbourne: Australasian Speech Science and Technology Association Inc., 2019, pp. 2690–2694.
[10] R. G. Podlubny, T. M. Nearey, G. Kondrak, and B. V. Tucker, "Assessing the importance of several acoustic properties to the perception of spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. 2255–2268, Apr. 2018.
[11] H. R. Bosker, "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies," *Behavior Research Methods*, Mar. 2021.
[12] E. Sohoglu and M. H. Davis, "Perceptual learning of degraded speech by minimizing prediction error," *Proceedings of the National Academy of Sciences*, vol. 113, no. 12, pp. E1747–E1756, Mar. 2016.
[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *Journal of Statistical Software*, vol. 67, no. 1, 2015.
[14] R. C. Team, "R: A language and environment for statistical computin," Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org/
[15] D. McCloy, L. Panfili, C. John, M. Winn, and R. Wright, "Gender, the individual, and intelligibility," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1722–1722, Sep. 2018.
[16] R. A. Wright, B. V. Tucker, M. C. Kelley, and M. Oganyan, "Effects of noise, native language, age, and speaker gender on intelligibility in a large corpus of read speech," *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. A264–A264, Apr. 2022.
[17] M. L. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2445–2454, Apr. 2006.
[18] S. L. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953–978, Sep. 2012.
[19] R. Smiljanić and A. R. Bradlow, "Production and perception of clear speech in Croatian and English," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1677–1688, Sep. 2005.

Illusions of Ungrammaticality in the Perception of Non-Native Accented Speech

Sarah Wesołek[1,2], Piotr Gulgowski[3], Joanna Błaszczak[3], Marzena Żygis[1,2]

[1]Leibniz-Centre General Linguistics, [2]Humboldt University Berlin, [3]University of Wrocław

wesolek@leibniz-zas.de, piotr.gulgowski@uwr.edu.pl, zygis@leibniz-zas.de

ABSTRACT

Studies have shown a reduced sensitivity to grammatical violations in non-native accented speech. The same expectation of ungrammaticality that reduces sensitivity to errors might also prime listeners to perceive illusory errors in accented utterances. In our study, we hypothesized that listeners exposed to non-native accented speech perceive more illusory grammatical errors. This effect should increase when sentences contain phonological violations.

To test our predictions, we conducted two mirror experiments for German and Polish, contrasting the perception of: (a) well-formed, (b) phonologically anomalous, and (c) ungrammatical sentences in native and non-native speech.

Our results show that in both languages grammatical sentences were more often judged as ungrammatical in non-native than native accent, a phenomenon called 'grammatical tinnitus'. However, the effect did not increase when phonologically anomalous sentences were presented in the Polish experiment, even if the overall number of incorrect answers increased. For German, the 'tinnitus' was absent in this condition.

Keywords: non-native accent, perception, grammatical illusion, German, Polish

1. INTRODUCTION

The current study assesses whether native listeners perceive illusions of ungrammaticality when listening to foreign-accented speech. The investigated phenomenon can be illustrated with the real-life example of a German politician, Raed Saleh, who was born in Palestine and moved to Germany in early childhood. As a candidate for the office of the mayor of Berlin in 2014, he participated in a TV debate. After the debate, Saleh's German skills were judged as inadequate because of many grammatical mistakes. However, a journalistic analysis [1] demonstrated that Saleh's use of German grammar was in fact better (one mistake every 140 words) than the use of grammar of a native German TV host in the debate (one mistake every 34 words). It is possible that Saleh's migrant background and non-standard pronunciation (e.g., using the palato-alveolar [ʃ] instead of the palatal [ç]) triggered this illusion of ungrammaticality.

Correspondingly, empirical investigations revealed that foreign accents modulate speech perception. Accented speech is more challenging to understand than native productions [2]. It has been experimentally linked to a general processing slow-down [3, 4] and increased processing cost [5]. Generally, non-native productions deviate from native utterances on a number of linguistic properties. The degree of accentedness has been shown to strongly depend on segmental errors [6, 7], and distorted suprasegmental features, such as intonation, timing, speech rate, and loudness [6, 8, 9]. Additionally, L2 speakers often use grammatical structures deviating from native language usage. The presence of grammatical errors could increase the degree of perceived accentedness, but results from the few available studies are mixed. Munro & Derwing [7] found a correlation between the number of grammatical errors spontaneously produced by accented or native speakers and the accent scores. However, when Asano & Weber [10] manipulated the grammaticality of German sentences produced by L1 and L2 speakers, participants rated speech containing grammatical errors as more foreign only when it was produced by L1 but not by L2 speakers. Furthermore, evidence from EEG research suggests that the grammaticality of utterances is evaluated differently for native and foreign-accented speech. In an ERP study by Hanulíková et al. [11], grammatical errors in sentences spoken with a foreign accent did not lead to a P600 effect, typically evoked by the same errors in sentences spoken by native speakers. Because the P600 component is commonly associated with structural reanalysis and repair, the authors interpreted their findings as an effect of listeners expecting incorrect utterances from foreign-accented speakers and consequently reducing attempts to repair such mistakes. The same expectation of ungrammaticality that results in reduced sensitivity to errors might also prime listeners to perceive illusory errors in well-formed utterances, resulting in a so-called 'grammatical tinnitus'.

Research has also shown that a foreign accent can trigger specific biases in listeners reflected, for example, in decreased ratings of traits like status, solidarity, or dynamism for accented speakers [12]. Crucially, such biases and expectations can also exert a top-down influence on perceptual processes, as reported for the phoneme categorization [13] and lexical recognition [14] tasks.

The discussion of the Raed Saleh's candidacy, as well as the research findings listed above motivated our first hypothesis that the number of incorrect responses in a grammaticality judgment task for grammatical sentences (both well-formed and phonologically anomalous) will be higher in the foreign-accented speech condition than in the native-accented speech condition. We also expect that the foreign accents will lead to a greater increase in the number of incorrect responses in sentences with phonological anomalies in comparison to well-formed sentences.

2. EXPERIMENTS

2.1. Experimental design

To test our hypotheses, we conducted two mirror experiments taking place in Germany and Poland. Participants listened to native and foreign-accented sentences further divided into well-formed, phonologically anomalous, and ungrammatical utterances. Sentences from the foreign accent condition were recorded by L2 speakers and were marked mostly by a subtle accent characterized by suprasegmental variation. Both experiments combined the ERP technique of recording brain potentials with the behavioral task of grammaticality judgments. As part of the task, participants had to decide whether the auditorily presented sentence was grammatically correct. Only the behavioral results of the study are reported below, and we limit our discussion to the well-formed and phonologically anomalous conditions essential for answering the hypotheses concerning the 'grammatical tinnitus'. The ungrammatical condition served to provide clear cases of ungrammaticality for the judgment task.

2.1.1. Materials

For each experiment, we selected 180 unique experimental items of comparable length following the same syntactic structure. For every item, three sentence versions were created: well-formed, phonological substitution, and grammatical error. This resulted in 540 experimental sentences. The critical middle region started with a preposition followed by a possessive determiner and a two or three syllabic noun. The three versions of each item differed only in the critical region, see Examples (1) and (2) below. For German and Polish audio examples please visit the project's OSF repository (https://osf.io/4w2ct/?view_only=d005b1e78dba4d1a8c7b07eebff7e56f). In the phonologically anomalous condition, three native vowels were replaced in an equal number of sentences by sounds easier to pronounce by L2 learners. Note that Polish learners of L2 German typically struggle with the German lax-tense contrast, e.g., [15]. For German data, we decided to use opposite vowel substitutions. Additionally, we have chosen substitutions between [ʏ] and [ɪ], since [ʏ] does not belong to the Polish phoneme system and [ɪ] is not part of the German vowel inventory. In the German experiment, the substitutions were: [e]→[ɛ], [o]→[ɔ], and [ʏ]→[ɪ]. In the Polish experiment, the substitutions were [ɛ]→[e], [ɔ]→[o], and [ɪ]→[y]. The phonological substitutions always appeared in the stressed, penultimate syllable. Grammatical errors consisted of L2-typical gender mismatches between the possessive determiner and the noun, see e.g., [16].

(1) Example: *German sentence triple*

Lena befragt die Lehrerin ...
Lena consults the teacher

(i) well-formed
...zu ihrem Fehler in der Klausur.
about her[masc] mistake[masc] in the exam

(ii) phonological substitution
...zu ihrem F[ɛ]ler in der Klausur.
about her[masc] mistake[masc] in the exam

(iii) grammatical error
... zu ihrer Fehler in der Klausur.
about her[fem] mistake[masc] in the exam

(2) Example: *Polish sentence triple*

Nina robi porządek…
Nina makes order

(i) well-formed
...w swojej kwaterze przed przyjazdem gości.
in her[fem] flat[fem] before arrival guests

(ii) phonological substitution
...w swojej kwat[e:]rze przed przyjazdem gości.
in her[fem] flat[fem] before arrival guests

(iii) grammatical error
...w swoim kwaterze przed przyjazdem gości.
in her[masc] flat[fem] before arrival guests

Digital auditory recordings of the materials were created. For each experiment, half of the sentences were recorded by native speakers and the other half by L2 speakers with a non-native accent. The same four phonologically trained male speakers recorded sentences for both the Polish and German experiment: two native speakers of Polish who were also L2 speakers of German and two native speakers of German who were also L2 speakers of Polish. The experimenters provided instructions for the speakers how to produce segmental violations in the critical

word, avoiding segmental changes in other parts of the sentence. Additionally, the recordings were inspected by two phonetically trained linguists for categorical changes and their accent strength. If a file contained any other categorical violations than the intended one, it was re-recorded. Thus, materials belonged to six conditions: 3 related to Sentence Type (well-formed, phonological substitution, grammatical error) and 2 related to Sentence Accent (native, foreign). The different versions of experimental items were distributed across three lists such that each participant listened to only one version of a given item. 180 experimental sentences (30 per condition) appeared on one list. Additionally, 72 fillers (36 native-accented, 36 foreign-accented) were recorded by the same speakers and added to each list. This resulted in a total number of 252 sentences per list.

2.1.2. Procedure

The experiments were conducted in EEG laboratories in Germany and Poland. Participants were seated in front of a computer screen. The experimental session started with written instructions appearing on the screen. Participants were asked to listen to the recordings of sentences while looking at the screen and to respond to questions following the sentences. Once the recording stopped, the question 'Is the sentence grammatically correct?' was displayed in the respective language. The participants' task was to respond by pressing one key for 'Yes' or another for 'No' as fast and accurately as possible. Their responses were recorded. A comprehension question appeared after the grammaticality question in one third of trials. A short practice session preceded the experiment proper.

2.1.3. Participants

33 native speakers of German (17 women, 16 men) aged between 21 and 35 years (M=26, SD=3.8) participated in the German experiment. They were mostly students of Phillips-Universität Marburg, except for 8 participants who were non-students. None of the participants reported having a linguistic background or knowledge of Polish.

In the Polish experiment, 30 native speakers of Polish (16 women, 14 men) took part. They were mostly students of the University of Wrocław, aged between 21 and 31 years (M=23, SD=2.4). None of the students had a linguistic background nor reported knowledge of German.

2.2. Statistics

All statistical analyses were conducted with R (version: 4.2.0, [17]). A binomial logistic regression model (package 'Lme4', [18]) was fitted with JUDGMENT ACCURACY [correct, incorrect] as the dependent variable. Judgment accuracy values were derived from 'Yes' or 'No' answers to the question 'Is the sentence grammatically correct?'. For grammatical (well-formed and phonologically anomalous) sentences, 'No' constituted an incorrect answer. For ungrammatical sentences, it would constitute a correct answer. We fitted separate models for the German and Polish experiment. In both models, the fixed factors were ERROR TYPE [well-formed, phonological substitution, grammatical error], ACCENT TYPE [native, foreign], and their interactions. We also included PARTICIPANT and SENTENCE TRIPLE as random intercepts and ERROR TYPE, ACCENT TYPE and their interaction as slopes. Because of convergence issues, we removed the interactions between ERROR TYPE and ACCENT TYPE from all random structures. For multiple comparisons we used the 'emmeans' function [19].

For the statistical analysis of German data, 5917 data points were submitted. For Polish, 5369 data points were analysed.

3. RESULTS

3.1. German experiment

The analysis of the German experiment revealed a significant interaction for ERROR TYPE and ACCENT TYPE between well-formed sentences and phonologically anomalous sentences (z=-3.49, p<.001).

The results indicate that sentences with no grammatical or phonological error (well-formed), spoken with a foreign accent were more likely to be (incorrectly) judged as ungrammatical than well-formed sentences spoken with a native accent (z=-3.3, p<.05), see Figure 1. A comparison of sentences with a phonological substitution revealed no significant difference between the native and foreign accent condition.

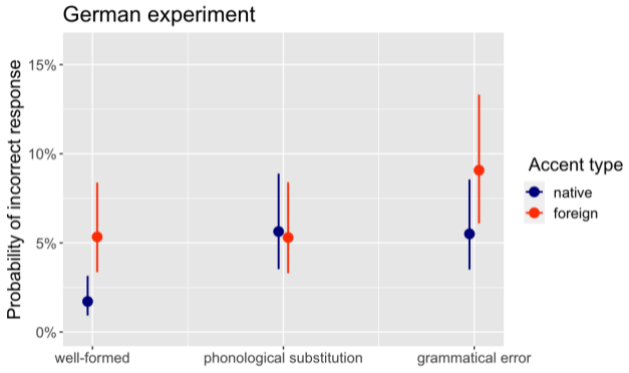


Figure 1: Probability of incorrect response for Error Type and Accent Type, German experiment

3.2. Polish experiment

The data analysis of the Polish experiment revealed a simple effect of ACCENT TYPE in the well-formed condition, with foreign-accented sentences being more likely to be judged as ungrammatical than native-accented sentences (z=3.25, p<.01), see Figure 2. Likewise, sentences containing phonological anomalies were more likely to be judged as ungrammatical when being foreign-accented compared to native-accented sentences (z=2.33, p<.05).

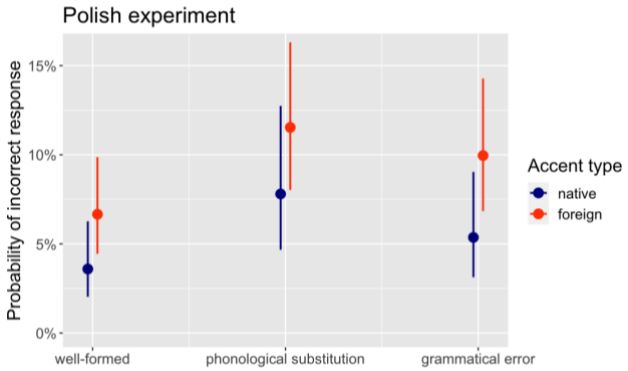


Figure 2: Probability of incorrect response for Error Type and Accent Type, Polish experiment

However, the statistical modelling indicated no significant interaction between ERROR TYPE and ACCENT TYPE. As demonstrated in Figure 2, the overall probability of incorrect responses increased for both the native and the foreign accent when sentences contained phonological anomalies.

4. CONCLUSIONS

The present study investigated whether listeners exposed to foreign-accented speech experience a form of perceptual illusion where non-existent grammatical errors are heard in grammatically correct speech. We predicted that the number of incorrect grammaticality judgments in the well-formed and phonologically anomalous condition would be higher for foreign-accented, compared to native-accented speech. Additionally, we predicted the number of incorrect responses to increase for sentences with vowel substitutions in the critical phrase.

The data from both languages suggest that listeners were more likely to judge well-formed utterances with no segmental substitutions as ungrammatical when they were uttered by foreign-accented speakers compared to fully native-sounding speakers. This effect can be interpreted as the 'grammatical tinnitus', a form of (un)grammaticality illusion. Various accounts of the mechanism underlying this phenomenon are possible. First, the results are consistent with the hypothesis that exposure to non-native accents induces a perception of non-existent grammatical errors in well-formed sentences. In other words, the locus of the effect would be at the online perception and comprehension stage. Second, the sentence as a whole can be judged as ungrammatical without the participant actually hearing any 'phantom' errors during speech perception.

Results also revealed unexpected differences between the experiments. In Polish, the overall probability of judging utterances as ungrammatical increased both in well-formed sentences and sentences containing phonological substitutions, as hypothesized. Surprisingly, the German data revealed no difference between the native and foreign-accented sentences in the phonologically anomalous condition, suggesting that Polish listeners were affected more by phonological anomalies than German listeners. We propose three explanations for the differences between the languages. Firstly, Germans might have a greater experience with segmental variation because of their exposure to various dialects, sociolects, and foreign accent variants. Due to their familiarity with segmental variation, they might be less sensible to phonological substitutions, leading to less confusion about the grammatical correctness of an utterance. Secondly, there might be differences between the degree of familiarity with the vice-versa accent. Previous ERP studies [20, 21] indicated that more familiarity with an accent leads to more native-like processing of language errors. Due to the awareness of the segmental patterns of a particular accent, listeners might be less prone to judge segmental violations as grammatical errors, resulting in no extra 'grammatical tinnitus'. However, a questionnaire filled by the participants showed little difference in self-reported–exposure to the non-native accents between both groups. Thirdly, the perceived strength of segmental violations might have diverged between the German and Polish experiments. While for two of three segmental violations the change was from tense to lax in the German experiment, it was the other way around for the Polish experiment.

In summary, the present investigation demonstrates that a foreign accent affects the way listeners perceive the grammaticality of utterances. The 'grammatical tinnitus' can result in foreign speakers being perceived as less (linguistically) competent, as illustrated by the example of Raed Saleh.

5. ACKNOWLEDGMENT

This work was supported by Deutsch-Polnische Wissenschaftsstiftung within the project 'Grammatical tinnitus and its role in the perception of foreign language accent. A comparison of German and Polish' (grant number: 2020-07). We are grateful to Ulrike Domahs for giving Sarah Wesolek the opportunity to run the experiment with German speakers at Marburg University. Finally, we also thank the Leibniz-Society.

5. REFERENCES

[1] Heiser, Sebastian. 2014. Berlins Bürgermeisterkandidat Saleh: Ein dubioses Hörproblem. *TAZ*: Gesellschaft/Medien. [Online]. https://taz.de/Berlins-Buergermeisterkandidat-Saleh/!5034466/ [Accessed: January 6, 2023].

[2] Bent, T., Bradlow, A. 2003. The Interlanguage Speech Intelligibility Benefit. *The Journal of the Acoustical Society of America*, 114, 1600–1610.

[3] Clarke, C. M., Garrett, M. F. 2004. Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116 (6), 3647–3658.

[4] Floccia, C., Butler, J., Girard, F., Goslin, J. 2009. Categorization of regional and foreign accent in 5- to 7-year-old British children. *International Journal of Behavioral Development*, 33 (4), 366–375.

[5] Van Engen, K. J., Peelle, J. E. 2014. Listening effort and accented speech. *Frontiers in Human Neuroscience*, vol. 8, 1–4.

[6] Magen, H. S. 1998. The perception of foreign-accented speech. *Journal of Phonetics*, 26 (4), 381–400.

[7] Munro, M. J., Derwing, T. M. 1995. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 45 (1), 73–97.

[8] Munro, M. J. 1995. Nonsegmental Factors in Foreign Accent: Ratings of Filtered Speech. *Studies in Second Language Acquisition*, 17 (1), 17–34.

[9] de Mareüil, P. B., Vieru-Dimulescu, B. 2006. The Contribution of Prosody to the Perception of Foreign Accent. *Phonetica*, 63 (4), 247–267.

[10] Asano, Y., Weber, A. 2016. Listener sensitivity to foreign-accented speech with grammatical errors. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, 1775–1780.

[11] Hanulíková, A., van Alphen, P. M., van Goch, M. M., Weber, A. 2012. When one person's mistake is another's standard usage: the effect of foreign accent on syntactic processing. *J Cogn Neurosci*, 24 (4), 878–887.

[12] Fuertes, J., Gottdiener, W., Martin, H., Gilbert, T., Giles, H. 2012. A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, 42, 120–133.

[13] Jannedy, S., Weirich, M. 2014. Sound change in an urban setting: Category instability of the palatal fricative in Berlin. *Laboratory Phonology*, 5, 91–122.

[14] Ingvalson, E. M., Lansford, K. L., Federova, V., Fernandez, G. 2017. Listeners' attitudes toward accented talkers uniquely predicts accented speech perception. *The Journal of the Acoustical Society of America,* 141 (3), EL234–EL238.

[15] Nimz, K., Khattab, G. 2020. On the role of orthography in L2 vowel production: The case of Polish learners of German. *Second Language Research*, 36 (4), 623–652.

[16] Grüter, T., Lew-Williams, C., Fernald, A. 2012. Grammatical gender in L2: A production or a real-time processing problem? Second Language Research, 28 (2), 191–215.

[17] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 4.2.0. https://www.R-project.org/.

[18] Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software*, 67 (1).

[19] Lenth, R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.3, https://CRAN.R-project.org/package=emmeans.

[20] Grey, S., van Hell, J. G. 2017. Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, 42, 93–108.

[21] Holt, R., Kung, C., Demuth, K. 2018. Listener characteristics modulate the semantic processing of native vs. foreign-accented speech. *PLoS ONE*, 13 (12), 1–24.

DIFFERENCES BETWEEN L1 AND L2 PROCESSING OF MEANINGFUL SENTENCES REVEALED BY COMBINING EEG AND EYE TRACKING

Kimberley Mulder, Sophie Brand, Lou Boves

Amsterdam Center for Language and Communication, University of Amsterdam; Radboud Teachers Academy; Center for Language Studies, Radboud University Nijmegen
k.mulder2@uva.nl; Sophie.Brand@ru.nl; Lou.Boves@ru.nl

ABSTRACT

We investigate how native French listeners and advanced Dutch learners of French use visual and auditory information while processing meaningful spoken sentences containing a reduced noun in mid-sentence position. Using a version of the visual world paradigm combined with EEG recordings, we investigate whether these reduced nouns affect the processing –as evidenced by the EEG signals– of a semantically related noun at the end of the sentence. We also investigate whether a newly developed feature that summarizes the eye movements in some time window affects the processing of those words. It appears that the reduction status of the first noun affects the processing of the related word and that this effect is different for the two listener groups. We also find that adding the new eye movement feature explains a highly significant amount of the variance in the EEG signals.

Keywords: visual world paradigm; EEG; eye tracking; reduced pronunciation; L1-L2

1. INTRODUCTION

Understanding reduced words in everyday speech can pose difficulties for second language learners, even if they are highly proficient in that language [1, 2], but little is known about the cognitive processes that make understanding reduced words more difficult for learners than for native listeners. To investigate this, we designed a visual world setting aimed to uncover those processes and their time course during listening comprehension, by combining EEG recordings and eye tracking. Native French listeners and advanced Dutch learners of French listened to meaningful sentences that contained a reduced or full form noun towards the middle of the sentence and a noun that was semantically related to that target noun towards the end of the sentence. Listeners who struggle to recognize a reduced target noun might need to use other sources of information to eventually correctly understand the message. This information could be semantically related information that occurs at the end of the sentence and the visual information about the target provided by the pictures. Through the simultaneous recordings of EEG and eye tracking we hope to see how native and non-native listeners deal with reduction over time and how they use visual information and semantically related words to solve any processing problems caused by reduction. We expect that effects of reduction last for a longer period of time for learners than for natives. In addition, the visual information (the pictures that appear on the screen one second before the start of the sentences and remain there until sentence end) that could help interpreting the reduced words might be used differently by natives and learners. If the reduced forms are not recognized and not matched with a visual target, this could affect they way in which semantically related word and the sentence as a whole are processed.

In order to explore long-distance effects, we need to go beyond standard analysis based on EEG time-locks on the target noun in the sentence. In [3] EEG signals were time-locked to fixations during the presentation of the audio signals and the presentation of the pictures, in addition to the start of the target noun, where the effect of reduction (perhaps better: the interaction between reduction and L1/L2) was assumed to be largest. In this study we focus on saccade data, and we investigate whether eye movements in the preview interval and in intervals aligned with the target and the related word affect ERP responses.

2. METHOD

We analyze the eye tracks and EEG signals obtained from 27 L2 and 28 native French listeners, who processed 115 stimulus sentences, the first three of which were practice trials. All sentences had the same structure: subject + verb form (+ que 'that') + target noun (+ verb form) + prepositional phrase / noun (e.g. *Elle a pris la tenaille pour enlever le clou*, "She took the pliers to remove the nail.") The

subject and verb form at the start of the sentences were chosen to minimize the predictability of the target word that could be pronounced in full, or with a heavily reduced ə in the first syllable. From 1000 ms before the start of the spoken sentences up to their offset, the listeners saw four pictures on a screen, one of which corresponded to the target word. After the end of a sentence the four pictures were replaced by a photo of a scene, and the listeners had to decide whether that scene corresponded to the message conveyed by the sentence.

The 54 target nouns were all concrete nouns that can easily be depicted and comprise two or three syllables, the first of which contains a ə. The 58 filler nouns in 'target position' were one-, two- or three-syllable words that can easily be depicted. The semantically related nouns at sentence end were not subject to phonetic criteria. All stimuli were produced by a female speaker born and raised in the Northern part of France. Due to the phonetic and visual restrictions imposed on the target nouns it was not possible to impose additional restrictions, such as frequency of occurrence. The `lemFreqFilm` counts of the target nouns in the Lexique database [4] vary between 0.1 and 739.12, μ = 50.14, σ = 127.45, and 0.34 -1061,92, μ = 113.56 and σ = 239.24, for the related nouns.

We simultaneously recorded EEG (using an Acticap with 64 active electrodes and BrainVision© hardware/software) and eye movements (using an SR Research Ltd. Eyelink 1000 eye tracker). To ensure temporal synchronization of the EEG and eye signals within Presentation (www.neurobs.com), an eye-tracking extension was developed in-house.

The distance to the screen was such that the four pictures could easily be identified without the need for fixations. Actually, we found that fewer than half of the trials contained a fixation of at least 100 ms [5]) on any picture, both during the preview period and during a time interval of 1000 ms aligned with the onset of the target noun. Therefore, we developed a novel feature that summarizes five measures that can be derived from the eye movement: (1) the number of pictures fixated, (2) the total time the eyes are in a Region of Interest of any picture, (3) the number of saccades, (4) the total distance travelled by the eyes, (5) the area of the screen covered by the eyes, defined as the area of the ellipse with radii equal to the standard deviation of the X- and Y-coordinates. It appeared that we could compress the five measures using Principle Component Analysis. The first component alone explains more than 95% of the variance in the data.

The EEG epochs were subjected to automatic Independent Component Analysis (ICA) [6] to remove artifacts. Epochs that contained too many artifacts were discarded.

To remove the impact of exogenous auditory excitation and eye movements we subtracted the Multivariate Temporal Response Function (mTRF) [7] from the EEG signals. For that purpose we constructed an artificial excitation function that is the sum of the Hilbert envelope of the audio signal and the 1-D eye movement function representing the distance travelled by the eye gaze.

We limit the analysis to the 35 sensors that are assumed to be least affected by eye movements. Furthermore, we average the signals from individual sensors in nine subsets that cover the scalp in left-center-right in one and front-center-back in the other direction.

The impact of factorial and numerical predictors on the EEG signals for the factors of interest, here the reduction status of the target word and the listener group (L1 or L2), or perhaps better: on the interaction between the factors `reduction` and `language`, changes during the course of a stimulus [8]. However, we are interested in the time course of possible effects of control factors such as the frequency of the target and the related words, as well as of the predictors derived from the eye track signals. For that reason, we apply Linear Mixed Effect Models, using the `lme4` [9] package in `R` version 4.1.2 [10] to overlapping 100 ms windows shifting along the time axis, cf. [11, 12]. Here, we use 20 such windows; the first one starts at the onset of the target or related word; the last one ends at 1000 ms after word onset. Using partially overlapping time windows may give rise to the repeated analysis problem. To counteract that problem we only consider an effect as truly significant if its t-value is significant in at least three consecutive time windows.

To investigate the impact of predictors derived from the eye tracks we start with a basic model:

```
model.0 = lmer(amplitude~Alday+trial+
block+marker*language*logFreqFilm+time+
(1+logFreqFilm|ppn)+(1|word),data=window,
```

in which `amplitude` represents the 10 EEG samples in the window; `Alday` is the average amplitude of the EEG signals in a 240 ms interval preceding the onset of the word that replaces the conventional baseline correction [13]; `trial` is the number of the trial in either `block` #1 or #2 in a session; `marker` denotes the reduction status of a stimulus (full or reduced); `language` denotes the listener group; `logfreqFilm` is the log of the frequency count of

the target word (when analyzing the data with time-lock on the onset of the target word) or of the related word (when time-lock is on the related word) in *Lexique* and `time` corresponds to the sequence of the samples in a window. `ppn` is the code of the 27+28 participants, and `word` the lexical identity of the target word. Factor levels `reducedFull` and `languageDutch` are on the intercept.

We then add additional predictors derived from the eye movement data, and use window-based AIC values to determine these improve the model fit, in which case we consider them as relevant.

3. RESULTS

We analyzed the features derived from the eye tracks to see whether there are significant differences between the two listener groups and/or between the three windows (preview, target, or related word) using *t*-tests for independent samples. In all three time intervals the distance traveled by the eyes of the L2 listeners is much larger than for the L1 listeners, but these movements appear to be constrained in a much smaller area of the display. Also, the L1 listeners fixate more often on a picture during the target window, and in the two other windows their fixations are longer. Thus, it appears that the two listener groups use the visual information in quite different manners.

Fig. 1 summarizes one detail of the analysis of the EEG signals related to the target word (red symbols) and the related word (blue symbols) in 20 time windows shown along the horizontal axis, viz. the *t*-values of the three-way interaction `reduction:language:frequency`. The nine panels show the results for the nine scalp areas, from left to right and front (top) to back (bottom). Note that the horizontal axis is labeled 'time from time-lock': the absolute time differs between target and related word. The vertical axes represent the *t*-values of the interaction `reduction:language:frequency`. The brown horizontal dashed lines indicate the value $t = 1.96$, and the yellow dotted line the value $t = -1.96$. The red '+' symbols show the windows in which the interaction is considered as significant in a brain area for the target words; the blue '*' symbols show the same data for the semantically related words.

Note that the factor `frequency` is tied to the word under analysis. Therefore, the predictors for the target and semantically related words are different. The results show a significantly more positive P200 component for the words with higher frequency in the native listeners, both for the reduced target

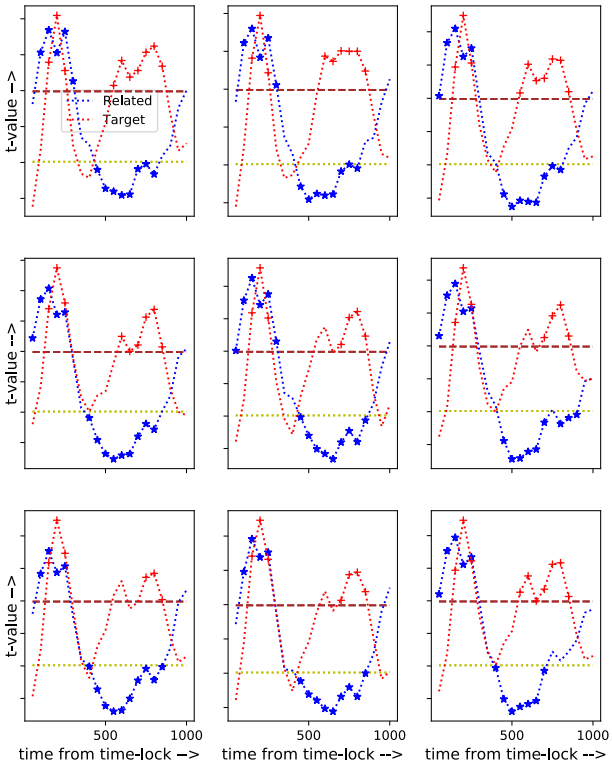


Figure 1: Three-way interaction `reduction:language:frequency` for the target and for the related word.

nouns, and -surprisingly- also for the semantically related words. This happens in all brain areas. For the target nouns there is also a significantly more positive late and somewhat broad P600 effect. The related words at the end of the sentences show a very different pattern. Here, in a broad time window starting 400 ms after word onset and lasting until 700 ms after onset, high frequency words following a reduced target show a significantly negative-going ERP.

Next, we built `LME4` models for the target and the related nouns in which we added the eye track features as additional predictors. For the target nouns we built models with the eye track feature obtained in the 1000 ms preview period and aligned with the onset of the noun; we also added both predictors simultaneously. The latter model yielded by far the smallest AIC values in all 20 window positions. For the related noun we also added the eye track feature of the window aligned with that word, plus all combinations of the three features. Here too, the model that included all three eye track features had by far the smallest AIC in all 20 window positions.

Given the substantial contributions of the eye track features to the fit of the `lme4` models one would expect that adding these features should affect

the significance of the factors that we are interested in, i.e., `reduction`, `language` and `frequency`. However, that appeared not to be the case. It appears that eye movement predictors that we add to the models do account for substantial amounts of variation in the EEG signals, but it is almost exclusively variance that has little or nothing to do with the linguistic factors that we are interested in.

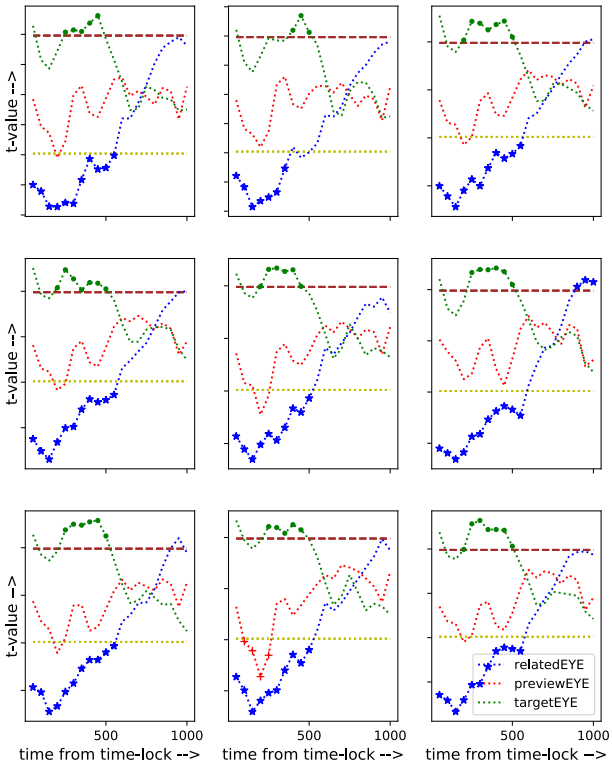


Figure 2: The t-values of the eye track features for the semantically related words.

While the eye track features do not affect the significance pattern of the main factors, it is still interesting to investigate the time course of their own significance, especially in the processing of the related word. The results are shown in Fig. 2, for the model that contained all three features. The first observation is that the patterns are very similar in all nine areas, and that significant t-values are restricted to a time interval that ends at about 550 ms after the onset of the words. Perhaps not surprisingly, in the presence of the eye track features aligned with the two nouns, the preview feature is almost never significant. The eye track features aligned with target and related word are antagonistic: a large value of the eye track feature during the target word yields less negative going ERPs in the related word, while a high value of the feature in the semantically related word is associated with more negative-going ERPs. For the target words (not shown) high eye activity is associated with more negative-going ERPs, again up to 550 ms after word onset. Here, however, there are significant contributions of the eye activity in the preview interval, in the form of less negative-going ERPs in the time interval between 600 and 800 ms after word onset.

4. DISCUSSION AND CONCLUSION

In this study we ask two questions: does a reduced word form in the first part of a sentence affect the recognition of a semantically related word at the end of that sentence differently between native listeners and advanced learners of the language? And does information in eye movements recorded in a visual world paradigm reveal how the visual information is used?

The significant interaction between `reduction`, `language` and `frequency` displayed in Fig. 1 confirms that reduction in the first part of a sentence does affect processing differently in native listeners and advanced learners, and that the difference is larger for high-frequency words. The effect is limited to the first 350 ms after the start of the words, which suggests that it is mainly related to phonetic decoding. We find the same effect of three-way interaction in the time interval up to 350 ms after the start of semantically related words at the end of the sentence. Thus, differences between natives and learners in phonetic decoding of reduced words in the first part of a sentence appear to affect their phonetic decoding of fully-produced related words at sentence end. ERP components in a broad time interval starting around 500 ms after the onset of the target en the related words are usually associated with semantic integration. In that time interval the effects of the three-way interaction differ between the targets and the related words. In the targets the interaction enhances a P600 effect, while that effect is diminished in the semantically related words.

The eye movements of the natives differ from those of the learners in all three time intervals. Therefore, it is safe to conclude that the natives use the visual information differently that the learners. The results shown in Fig. 2 show that eye movements make significant contributions to predicting EEG signals in the time interval up to 600 ms after the onset of the semantically related words. This suggests that visual information affects phonetic decoding and lexical access, but that its role in semantic integration is limited.

5. REFERENCES

[1] S. W. L. Wong, P. P. K. Mok, K. K.-H. Chung, V. W. H. Leung, D. V. M. Bishop, and B. W.-Y. Chow, "Perception of native english reduced forms in chinese learners: Its role in listening comprehension and its phonological correlates," *TESOL QUARTERLY*, vol. 51, pp. 7 – 31, 2015.

[2] D. Nouveau, "Limites perceptives de l'e caduc chez des apprenants nÃ©erlandophones," *Revue Canadienne de Linguistique Appliquée*, vol. 15, pp. 60 – 78, 2012.

[3] K. Mulder, S. Brand, M. Ernestus, and lou Boves, "Processing reduced speech in the l1 and l2: A combined eye-tracking and erp study," *Language, Cognition and Neuroscience*, Submitted, The data used in this study were collected as part of the European Research Council Consolidator grant 284108 awarded to Mirjam Ernestus.

[4] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2 : A new French lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516 – 524, 2004.

[5] B. R. Manor and E. Gordon, "Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks," *Journal of Neuroscience Methods*, vol. 128, pp. 85–93, 2003.

[6] S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, pp. 145–151.

[7] M. Crosse, G. Di Liberto, A. Bednar, and E. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016, article 604.

[8] L. ten Bosch, L. Boves, and K. Mulder, "Analyzing reaction time and error sequences in lexical decision experiments," in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 2280 – 2284.

[9] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[11] K. Winsler, K. J. Midgley, J. Grainger, and P. J. Holcomb, "An electrophysiological megastudy of spoken word recognition," *Language, Cognition and Neuroscience*, pp. 1 – 22, 2018.

[12] A. Nijveld, L. ten Bosch, and M. Ernestus, "ERP signal analysis with temporal resolution using a time window bank," in *Proceedings of Interspeech*, Graz, 2019, pp. 1208 – 1212.

[13] P. M. Alday, "How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits," *Psychophysiology*, vol. 56, no. 12, p. e13451, 2019.

PERCEPTION OF ORAL AND NASAL ALVEOLAR FLAP SOUNDS IN AMERICAN ENGLISH

Luca Garai

Department of English Linguistics, Eötvös Loránd University
garailuc@gmail.com

ABSTRACT

The perception of /t d/ alveolar flaps is well-documented in phonetic literature, unlike that of nasal flaps. The study examines how segment duration, preceding vowel quality, and vowel nasality influence the categorical perception of /t/ and /n/ uttered in a V́_ɚ flapping environment, in American English words and nonwords. A *Praat* script was used to manipulate pre-recorded words via acoustic mixing and temporal manipulation, yielding a 5-level nasality scale of preceding vowels and five durations of medial /t/ and /n/ each. A binary forced-choice perception test was conducted with six monolingual North American male participants, and a logistic regression model was used to analyze the effect of acoustic features on the perception of the medial consonant. The results show that the type of underlying consonant (t/n) and preceding vowel quality (ɪ/ɛ/æ/ʌ/ɑ) had a significant effect on consonant perception, whereas consonant duration and vowel nasality did not influence participants' decisions significantly.

Keywords: speech perception, alveolar flapping, nasalization, American English

1. INTRODUCTION

Flapping is a lenition process that results in stop consonants becoming shorter and more sonorous. In American English, it primarily targets intersonorant alveolars /t d/ [1]; for instance, the medial consonant in *shouting* /ʃaʊtɪŋ/ is shortened and voiced: [ʃaʊrɪŋ]. The most typical flapping environment is V́_V, between a stressed and an unstressed vowel [2, 3]. The main distinguishing feature between flapped and unflapped realizations of consonants is segment duration. Flapped /t d/ are usually between 10 and 40 ms long, while their unflapped realizations are articulated longer than 100 ms [4]. When a plosive is flapped, a short sonorous segment (similar to liquids) is produced instead of a longer closure+hold phase.

Many studies have investigated the perception of alveolar flaps in English [5, 6, 7, 8, 9]. Despite the detectable acoustic differences between flapped minimal pairs containing /t/ and /d/ (such as the duration of the preceding vowel due to pre-fortis clipping [10]), listeners cannot identify the underlying consonant based on these differences alone [8, 9]. Lexical bias plays a role in the perception of flap-containing minimal pairs, wherein the more frequent word form is detected more often when the minimal pairs are real lexical items [5, 7].

In recent decades, the theoretical literature documented the spread of flapping to intersonorant alveolar nasal /n/ and nasal-plosive cluster /nt/, resulting in homophony between word pairs such as winner–winter [wɪr̃ɚ] [3, 11, 12]. One previous acoustic observation regarding nasal flaps is that flaps produced from underlying /n/ and /nt/ are significantly longer in duration than oral /t/ flaps [13]. Figure 1 exhibits this comparison in sound samples recorded for the current experiment. For /n/ and /t/ appearing in a V́_ɚ environment after front and central vowels /ɛ ɪ ʌ/, the difference in consonant duration follows the previous findings of Garai [13].

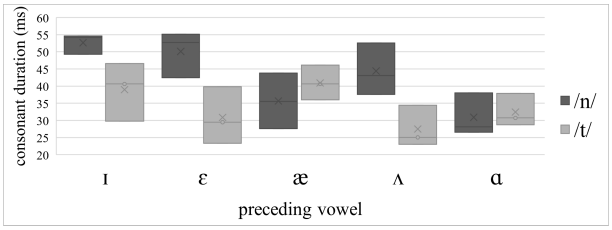


Figure 1: Duration of nasal /n/ and oral /t/ articulated in a V́_ɚ flapping environment, after stressed vowels of various vowel qualities.

During the articulation of nasals such as /n/ or /m/, the velum is lowered and air flows through the nasal cavity as well as the pharynx, in a process called velopharyngeal coupling [14]. The involvement of the nasal tract influences the acoustics of speech sounds due to the interference of resonances in the oral and nasal cavities, which

heightens the intensity in some frequency ranges and dampens it in others. Fujimura [14] observed several nasal formants (or poles) and antiformants resulting from this interference, the first of which (P0) is the most prominent around 250–300 Hz.

Nasality can spread to neighboring segments, primarily onto vowels [15]. The movement of the velum is not synchronous with the opening and closing of the lips, producing a coarticulatory effect where nasal airflow is present in the production of the adjacent vowel(s), thus changing the formant structure of these vowels [16]. Vowel nasalization is spontaneous in most languages and it can be measured by looking at the difference between the amplitude of the first oral formant (A1*) and the amplitude of the first nasal pole (P0) [17, 18]. The value of A1*–P0 is lower in nasalized vowels compared to oral ones [18]. In nonhigh vowels, P0 is identical to either the first (H1) or the second (H2) harmonic, whichever is higher in amplitude. In high vowels /i u/, the first formant is low enough that it possibly interferes with the range of P0 and in these vowels, the second nasal pole (P1) is used as a reference point for comparison [18]. Figure 2 shows the spectral difference between an oral and a nasalized realization of the same quality of vowel; the main difference lies in the relative intensity of the lower frequency range compared to the intensity of the first formant, and the first few harmonics have a higher intensity than A1 in nasalized vowels.

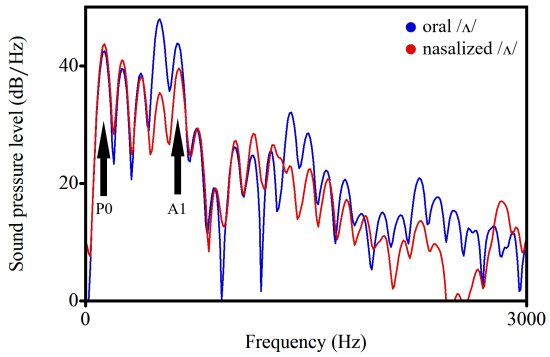


Figure 2: Spectral slice of oral and nasalized /ʌ/, with arrows pointing to the amplitude of the first nasal pole (P0) and the first oral formant (A1).

The difference between A1*–P0 values in oral and nasalized vowels is not constant throughout the duration of the vowel. Due to the coarticulatory effect of neighboring consonants, the beginning and/or the end of the vowel (depending on which side has a neighboring nasal consonant) is considerably more nasalized than the rest of the segment. This change can be observed in Figure 3 in vowels of various qualities. The A1*–P0 value of post-oral vowels changes based on whether the following consonant has a nasalizing effect on them. Vowels before nasal /n/ tend to get lower in their A1*–P0 values throughout the duration of the segment, whereas the same cannot be said for vowels before oral /t/.

This study focuses on the perception of oral and nasal alveolar flap sounds by North American listeners and aims to find preliminary cues as to whether there are distinct acoustic features (or clusters of features) by which listeners categorize the underlying consonant heard in a given sample. Based on the phonological and phonetic literature discussed above, the following two research questions are put forward:

1. How does the duration of the target consonant influence the perception of underlying /t/ and /n/ uttered in a flapping environment?
2. How does the quality and the degree of nasalization of the preceding vowel influence perception of underlying /t/ and /n/ uttered in a flapping environment?

2. METHODS

Five pairs of bisyllabic English (non)word minimal pairs were selected as the stimuli for the experiment, as seen in table 1. Each pair of words contained a stressed first vowel of a different quality, /t/ or /n/ as the medial consonant, and the initial consonant was /ʃ/ in all cases for ease of segmentability. The stimuli were read aloud three times by a 31-year-old male speaker from North Carolina and were recorded with a Blue Yeti USB microphone in a quiet environment. Recordings were segmented and annotated using *Praat* [19], and the duration of the medial consonant and the A1*–P0 value of the preceding vowel were extracted using VoiceSauce [20]. For the perception experiment, the recording with a stressed vowel exhibiting the highest A1*–P0 value in /t/-samples was chosen (out of three per word), while for samples with medial /n/, the one with the lowest A1*–P0 preceding vowel was selected.

Stimuli were manipulated using a *Praat* [19] script. The oral vowel preceding /t/ and the nasalized vowel preceding /n/ in every minimal pair was isolated from the word, matched in duration and pitch, then mixed in various ratios (100:0; 75:25; 50:50; 25:75; 0:100) creating a scale of nasality, and finally spliced back into the containing word. Furthermore, the duration of the medial consonant

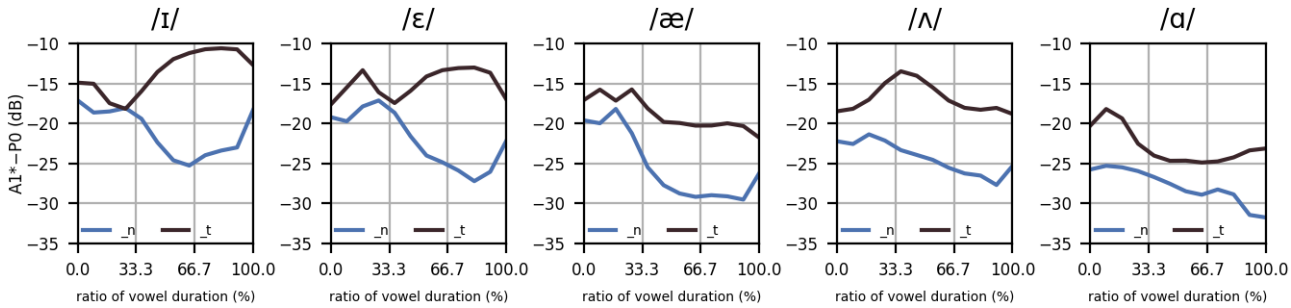


Figure 3: A1*–P0 values of vowels with various qualities articulated before /n/ and /t/ consonants throughout the duration of the vowel.

V́_	/t/	/n/
ɪ	SHITTER	SHINNER
ɛ	SHETTER	SHENNER
æ	SHATTER	SHANNER
ʌ	SHUTTER	SHUNNER
ɑ	SHOTTER	SHONNER

Table 1: Sample words recorded for the perception experiment.

was changed so that each level of vowel nasality was followed by consonants of 20, 30, 40, 50, and 60 ms, resulting in $5 * 5 * 2 = 50$ items per vowel quality, 250 target items overall.

A series of 250 bisyllabic filler items were also recorded with a word-initial /t/ or /z/, containing a diphthong as the stressed vowel, and /g k/ or /m p/ as the medial consonant, forming minimal pairs (e.g. *toager–toaker, zoomer–zooper*). Filler items were not manipulated acoustically.

Participants were recruited locally in the Raleigh, NC area with the aim of forming a homogeneous group of speakers to avoid regional and societal discrepancies. Six North American male speakers participated in the experiment with an average age of 31 years (min. 27, max. 32). All participants were native to North Carolina and were NC residents at the time of the experiment. They were all monolingual English speakers and none of them reported any conditions affecting their hearing or speech production.

A binary forced-choice test was conducted using *PsychoPy* [21] in a quiet room in each participant's home. Stimuli were presented in isolation and participants were tasked with deciding which medial consonant they heard out of the two shown on the left and right side of the screen (N and T in the case of target items), by pressing the left or right arrow key on the keyboard. The experiment contained three instances of the 250 target items each, presented in random order, with an equal number of filler items. Overall, participants judged the medial consonant in 1500 items, which were split up into nine equal parts with optional short breaks in between. Answers were recorded within *PsychoPy* [21] and exported to an Excel sheet for each participant by the software.

The analysis of the data was conducted in *RStudio* [22], with the *lme4* [23] package, using a logistic regression model. The participant's decision (N/T) for each target stimulus was used as the dependent variable, while the explanatory variables were the following: underlying consonant type and the resulting word-final nasality (n/t); duration of the consonant after manipulation (20/30/40/50/60 ms); quality of the preceding vowel (ɪ/ɛ/æ/ʌ/ɑ); nasalization of the preceding vowel after manipulation (0%/25%/50%/75%/100%). The participant was also included in the model to account for individual differences and the interaction of explanatory variables was also analyzed. McFadden's R^2 [24] was obtained to test the fit of the logistic model, and Tukey's post-hoc tests were conducted in cases where variables had a significant effect.

3. RESULTS AND DISCUSSION

Based on the results of the logistic regression model, the type of the underlying consonant (n/t) had a significant effect on the decisions of participants ($z = 2.606$; $p = 0.009$). Meanwhile, there wasn't a substantial difference in perception based on the duration of the medial consonant. The ratio of perception-based categorical decisions in relation to consonant features is shown in Figure 4 (a). Regardless of consonant duration, stimuli with an underlying /t/ were perceived as /t/ in a much higher proportion of instances (72.58% of all stimuli with underlying /t/) compared to those with

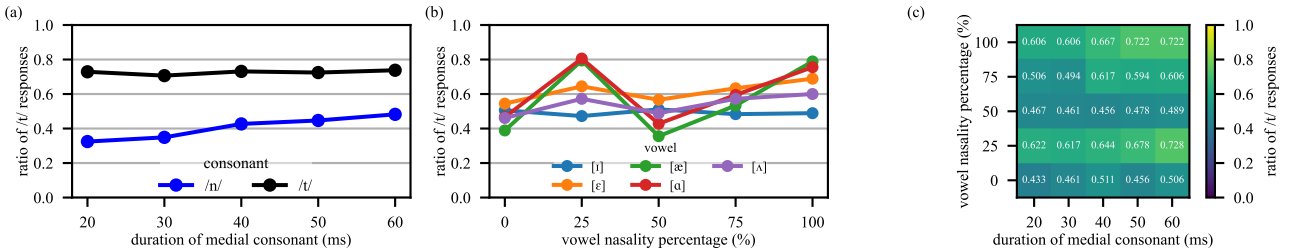


Figure 4: (a) Ratio of /t/ answers for each underlying consonant, in relation to the duration of the consonant. (b) Ratio of /t/ answers for each preceding vowel quality, in relation to the nasality of the preceding vowel. (c) Interaction of the two manipulated values (consonant duration and preceding vowel nasality).

underlying /n/ (40.58%). Whether this difference is due to the acoustic characteristics of the consonants themselves or the nasalization of the word-final [ə˞] segment supplied by the preceding /n/ cannot be determined based on the results of the current experiment. We see a positive correlation between the duration of the realized consonant and the ratio of /t/ answers in the case of stimuli with underlying /t/; this does not fall in line with previous acoustic observations detailed above in Section 1, which stipulated that /n/ uttered in a flapping environment is longer in duration than flapped /t/.

Based on the regression model, the quality of the preceding vowel had a significant effect on the perception of the medial consonant ($z = -6.535$; $p < 0.001$). The post-hoc test showed that there was a significant difference in the perception of consonants after the following pairs of vowels: /ɪ/ and /æ/ ($p = 0.005$), /ɪ/ and /ɑ/ ($p < 0.001$), /ɪ/ and /ɛ/ ($p < 0.001$), /ʌ/ and /ɛ/ ($p = 0.009$). These differences are shown in Figure 4 (b). The perception of consonant nasality before mid-high front /ɪ/ tended to differ from most other preceding vowel environments in that participants were more likely to perceive these consonants as /t/, which may be attributed to the fact that the first formant of high front vowels usually falls in the same frequency range as P0, as noted by Styler [18]. The logistic model did not show a significant effect of vowel nasality on the perception of the following consonant, therefore the degree of vowel nasalization did not influence participants' judgments in categorizing the medial consonant.

A significant interaction was observed between underlying consonant type and preceding vowel quality ($z = 6.828$; $p < 0.001$), between consonant duration and vowel quality ($z = 2.858$; $p = 0.004$), between preceding vowel nasality and vowel quality ($z = 2.526$; $p = 0.01$), and between consonant duration and preceding vowel nasality ($z = 1.976$; $p = 0.04$; this interaction is shown in Figure 4 (c)). These results reveal that while certain features do not have a significant effect on the perception of oral v. nasal alveolar consonants in a flapping environment, their influence can be detected in conjunction with other acoustic features. Furthermore, vowel quality seems to have a substantial effect on the perception of these consonants, both on its own and interacting with other features. McFadden's R^2 for the logistic regression model was $\rho^2 = 0.462$, indicating an excellent model fit, as described by McFadden [24].

4. CONCLUSION

The features with the most substantial effect on the categorical perception of /t/ and /n/ in a flapping environment were the type of underlying consonant and the quality of the preceding vowel. Underlying /t/ realizations were more likely to be perceived as /t/ than underlying /n/ realizations, regardless of consonant duration. In terms of preceding vowel quality, lexical bias might have been a potential influencing factor (in minimal pairs with real lexical items like *shitter, shatter, shutter–shunner*), since it has been shown that listeners tend to favor word forms with a higher lexical frequency [5, 7]. The formant structure of the preceding vowels might have also played a role in whether listeners identified the medial consonant as nasal.

Due to the small number of participants, the results detailed above should be viewed as preliminary. To avoid the influence of lexical bias on the results in the future, the stimulus set should be limited to nonwords or controlled for lexical frequency. Other future steps of perception research on nasal flapping are the inclusion of underlying /nt/ as the medial consonant, and diphthongs as the preceding vowel. The current results posit that there might be previously undocumented acoustic features distinguishing oral and nasal alveolar flaps.

5. REFERENCES

[1] J. S. Kenyon, *American Pronunciation. 12th ed.* Ann Arbor: George Wahr Publishing Company, 1994.

[2] D. Patterson and C. M. Connine, "Variant frequency in flap production," *Phonetica*, vol. 58, no. 4, pp. 254–275, 2001.

[3] D. Kahn, "Syllable-Based Generalizations in English Phonology," Ph.D. dissertation, Massachusetts Institute of Technology, 1976.

[4] V. W. Zue and M. Laferrière, "Acoustic study of medial /t, d/ in American English," *The Journal of the Acoustical Society of America*, vol. 66, no. 4, pp. 1039–1050, 1979.

[5] C. M. Connine, "It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition," *Psychonomic Bulletin & Review*, vol. 6, no. 11, pp. 1084–1089, 2004.

[6] N. Warner, "Cues to perception of reduced flaps," *The Journal of the Acoustical Society of America*, vol. 125, no. 3317, 2009.

[7] W. Herd, A. Jongman, and J. Sereno, "An acoustic and perceptual analysis of /t/ and /d/ flaps in American English," *Journal of Phonetics*, vol. 38, no. 4, pp. 504–516, 2010.

[8] A. Braver, "Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping," *Lingua*, vol. 152, pp. 24–44, 2014.

[9] G. Yun, "A mismatch in completeness between acoustic and perceptual neutralization in English flapping," *Korean Journal of English Language and Linguistics*, vol. 22, pp. 1133–1158, 2022.

[10] T. Scheer, "Voice-induced vowel lengthening," *Papers in Historical Phonology*, vol. 2, pp. 116–151, 2017.

[11] M. Picard, "English flapping and the feature [vibrant]," *English Language and Linguistics*, vol. 1, no. 2, pp. 285–294, 1997.

[12] B. Vaux, "Flapping in English," Linguistic Society of America, Chicago, IL, 7 January 2000, 2000.

[13] L. Garai, "Influencing factors of nasal flapping in English," in *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2021*. Budapest: Nyelvtudományi Kutatóközpont, 2021.

[14] O. Fujimura, "Analysis of nasal consonants," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1865–1875, 1962.

[15] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of vocal-tract characteristics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 541–558, 1971.

[16] G. Zellou, R. Scarborough, and K. Nielsen, "Phonetic imitation of coarticulatory vowel nasalization," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3560–3575, 2016.

[17] M. Garellek, A. Ritchart, and J. Kuang, "Breathy voice during nasality: A cross-linguistic study," *Journal of Phonetics*, vol. 59, pp. 110–121, 2016.

[18] W. Styler, "On the acoustical features of vowel nasality in English and French," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.2.21)," http://www.praat.org/, 2022, accessed: 10-01-2022.

[20] Y.-L. Shue, P. A. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 1846–1849.

[21] J. W. Peirce, J. R. Gray, S. Simpson, M. R. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. Lindelø v, "PsychoPy2: experiments in behavior made easy," *Behavior Research Method*, 2019.

[22] RStudio Team, "RStudio: Integrated development environment for R," http://www.rstudio.com/, Boston, MA, 2022.

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[24] D. McFadden, "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, P. Zarembka, Ed. Academic Press, 1973, pp. 105–142.

PERCEPTION OF POP-OUT VOICE IN VARIOUS CONDITIONS OF BABBLE NOISE

Mafuyu Kitahara[1], Hideki Kawahara[2], and Shigeaki Amano[3]

[1]Sophia University, [2]Wakayama University, [3]Aichi Shukutoku University
mafuyu@sophia.ac.jp, kawahara@wakayama-u.ac.jp, psy@asu.aasa.ac.jp

ABSTRACT

A perception experiment was conducted to investigate the characteristics of pop-out voice which is easy to listen to even in a noisy background. Stimuli consisted of 30 target voices (either in a high, mid, or low pop-out rank) overlapped by a forward or backward babble noise which was made by mixing normal or time-reversed short sentences in 4 to 100 voices. The authors expected that these babble noise operations would obscure temporal information of phoneme in the time-reversed babble noise, and linguistic information decreases as the number of voices of babble noise increases. Twelve listeners rated a pop-out score of the stimuli using a 5-point scale. Results indicate that the scores are similar in both forward and backward babble noises across the different number of voices in the babble noise, suggesting that the characteristics of pop-out voice are robust regardless of temporal and linguistic properties of babble noise.

Keywords: pop-out voice, babble noise, backward play-back, linguistic information

1. INTRODUCTION

There are some voices which stand out and reach the listener even in very noisy background. Amano and his colleagues have been focusing on the acoustic and perceptual properties of such voices and name those as “pop-out voice” [1-3]. It is different from so-called clear speech which concentrates on the intelligibility of linguistic information in utterances [4]. Even if the intelligibility is low, the voice itself can stand out. Pop-out voice is a concept that focuses on properties of voice, irrespective to the content of the linguistic message conveyed by that voice.

The degree of pop-out depends on various factors on the speaker’s side, such as Fo, intensity, formant frequencies, spectral envelope, speech rate and phonological structure [2]. Another important set of affecting factors is the properties of background noise. The source of the noise in natural setting may vary widely, including environmental noise, artificial noise, and speech by other people as in a cocktail party. The so-called cocktail party effect can be simulated by using the babble noise which is created by overlapping multiple voices.

The focus of the present paper is to test various conditions of babble noise in order to investigate how robust pop-out voice is. The number of voices overlapped in the generation of babble noise is an easily manipulatable property and has frequently been discussed in the literature [5-6]. For example, one past study investigated the effect of the number of voices in babble noise in detecting isolated words. Two to ten voices were overlapped for babble noise and the performance of listeners was worse in the 2-voice condition but improved when the number of voices increased [7]. In another study, the number of voices was by far wider, from 1 to 512 to find that the performance was worst when 8 voices were overlapped [8].

These studies focused on the intelligibility of target items, though. The dependent variable was the rate of correct identification of the word, syllable, or segment in most cases. In contrast, pop-out is supposed to be a property of target voice which must be independent from the intelligibility issues. To clarify the characteristics of pop-out voice, it is worth testing whether the number of voices in babble noise affect the pop-out voice in the way similar to the results obtained for intelligibility-oriented experiments.

Another issue here is the interference of language and/or linguistic content in babble noise. When the number of voices is only a few, word fragments in the noise are clearly identifiable and supposedly interfere the detection of target items. However, word fragments may not interfere if the listener does not know the language or, at least, the linguistic content is unattainable for the listener. Only a few studies have investigated the linguistically crossed speech-in-noise problem [9,10].

When the source of the babble noise came from the same language as the target item (English-English), the performance in the two-voice babble was worse than that in the different language condition (English-Mandarin) [10]. This “linguistic interference” is of particular interest for intelligibility-oriented studies by its nature. However, if the effect of linguistic interference is minimal or irrelevant for the degree of pop-out voice, this will in

turn establish that pop-out voice is independent from intelligibility.

In the present study, a method to generate unintelligible babble noise was adopted where time-reversed voices were overlapped. Due to unnatural time course of the signal, any word fragments in such backward babble noise were difficult to detect even when a very few voices were overlapped. The merit of using backward babble noise is that overall acoustic properties can be kept the same as forward babble noise created from the same set of voices: long-term spectra, mean and dispersion of Fo, mix of male and female voice, among others, are the same across the two types of noise.

2. METHOD

2.1. Participants

Twelve native Japanese speakers (eight males and four females) participated in the perception experiment. Their dialectal background varies across western and eastern regions in Japan. Their average age was 53.9 years (SD = 7.3 years). They had no significant history of hearing problems.

2.2. Pre-experiment

A pre-experiment was conducted to obtain a pop-out score of speech items for the main experiment. Speech items /gaizin saɴ wa kampeki ʃɯgi deaɾɯ/ (The foreigner is a perfectionist) spoken by 100 native Japanese speakers were randomly selected from the items used in [1]. The speech items were filtered with K-weighting [11], and then their root mean square (RMS) was calculated. Intensities of the items were adjusted to the same level using the RMS. After the adjustment, the items were mixed at -4dB signal-to-noise ratio with a babble noise that was generated by overlapping various sentences spoken by 10 native Japanese speakers (see [1] for precise procedure). Using a five-point scale (1: not pop-out - 5: pop-out), the 12 participants in Section 2.1 rated degree of pop-out of the 100 speech items in the babble noise. They were instructed to use the full range from 1 to 5 whenever possible, and to make intuitive decisions without too much trouble. A practice session consisting of 18 trials was given beforehand for the participants to familiarize the notion of pop-out voice [cf. 1]. The pop-out score for each item was obtained as an average of their ratings.

2.3. Stimuli

2.3.1. Target voices

Using the obtained pop-out score, target voices of 30 native Japanese speakers (15 males and 15 females) were selected from the 100 speech items in the pre-experiment to form three ranks: low (L: 1.0-2.0), middle (M: 2.5-3.5), and high (H: 4.5-5.0). Each rank contained 10 target voices spoken by five each of males and females.

2.3.2. Babble noise

The source of babble noise was the speech database having 30 Japanese sentences spoken by 14 each of male and female native Japanese speakers [12]. All sentences in the database were filtered with K-weighting [11] and their level was adjusted to the same level. A 100-second babble noise was generated by overlapping either 4, 8, 20, 40, or 100 sentences that were randomly selected from the database but had the same number of male and female speakers in each sentence group. Intensity of all the babble noises was adjusted to the same level. In addition to simple overlapping of multiple voices, reversed babble noises were prepared in this study where the source sentences were inverted in a time domain before the overlap. These backward babble noises and the usual forward babble noises were both used to test the effect of whatever linguistic information left in the babble noise on the target voice. Thus, the direction of the babble noise is one of the independent variables in our experiment.

2.3.3. Stimulus preparation

Thirty target voices (10 each for H, M, and L pop-out score ranks) were mixed at -7dB signal-to-noise ratio with the babble noises generated in the way described in the previous section. The mixing timing was such that the babble noise started 500 ms before the beginning of the target voice and ended 500 ms after the end of the target voice.

The babble noise was obtained by randomly selecting a segment from the entire 100-second babble noise so that each target voice was mixed with a different part of babble noise. Under these mixing conditions, two stimulus sets were prepared: one by using the forward babble noise and the other by using the backward babble noise.

2.4. Procedure

Stimuli stored in a personal computer were diotically presented to each participant in a random order in a quiet room at a constant intensity through

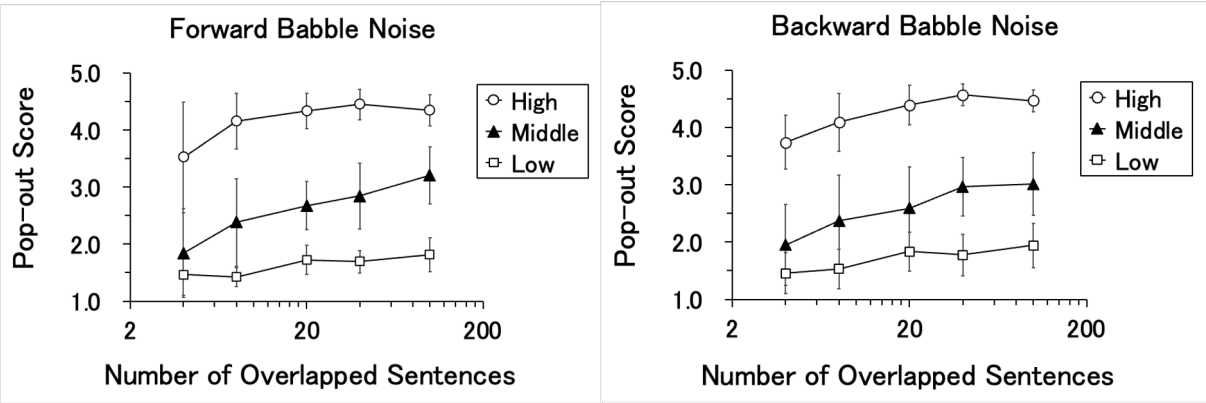


Figure 1: Average pop-out score as a function of the number of sentences (log scale) in the babble noise and pop-out rank of the target voice. Left: forward babble noise. Right: backward babble noise.

circumaural headphones (SONY, MDR-7506) with an audio interface (Roland, Rubix24). The participants rated the degree of pop-out using a five-point scale (1: not pop-out - 5: pop-out). They rated the set of forward babble noise (150 trials) and then the set of backward babble noise (150 trials).

3. RESULTS

Figure 1 depicts average ratings by the participants as a function of the number of sentences. The left and the right panel show the forward and backward babble noise condition, respectively. Three lines in each panel correspond to the rank of the pop-out score of target voices, H, M, and L. Statistical analyses using linear mixed effects models were applied with rating score as the dependent variable, and the number of sentences in the babble noise, the pop-out score rank, and the direction of the babble noise as the independent variables. Intercepts of participants and target voices were random factors.

Table 1: LME results for the number of sentences in the babble (n_babble), the direction of the babble (dir_babble: backward as the baseline), and the rank of pop-out score (rank_L/M: H as the baseline)

	Estimate	*t value*	*Pr(>|t|)*
intercept	3.962	27.161	<.001
n_babble	.006	5.573	<.001
dir_babble	-.014	-.191	.848
rank_L	-2.349	-14.481	<.001
rank_M	-1.695	-10.450	<.001
n_babble:dir_babble	-.003	-.214	.830
n_babble:rank_L	-.001	-.876	.381
n_babble:rank_M	.003	2.125	.033
dir_babble:rankL	-.064	-.590	.555
dir_babble:rankM	-3.672	-.341	.733

The statistical model indicates that the main effects of the number of sentences and the pop-out score rank were significant ($p < .001$), but the main effects of the direction was not significant (n.s.). The interaction between the number of sentences and M-ranked voices was significant ($p < .05$). These results suggest that when the number of sentences in the babble noise increase, the pop-out score increases for the M-ranked target voices. However, for the H- and L-ranked voices, the effect of the number of voices was not confirmed. Moreover, the direction of the babble noise did not change the overall picture.

4. DISCUSSION AND CONCLUSION

The results of the experiment have several implications on the nature of the pop-out voice. First, the linguistic information left in the babble noise had negligible effect on the recognition of the target voice. The fact that forward and backward babble noise conditions did not significantly make a difference in the performance of listeners imply that the backward noise whose linguistic information was destroyed by inverting the source also had a similar effectiveness to the forward babble noise in disturbing the target voice. This, in turn, suggests that pop-out voice is independent from the notion of intelligibility. It is a property of voice irrespective to the linguistic content.

Second, the fewer the number of sentences in the babble noise, the more disturbing effect the babble noise had on the target voice. As shown in Figure 1, the average pop-out score becomes larger as the number of sentences increases, which was statistically evident for the M-ranked target voices.

As for the H-ranked target voices, there seems to be a ceiling effect across babble noise conditions,

while the L-ranked target voices were identified as close to “not pop-out” across the board.

There are remaining issues to be addressed for the future study of pop-out voice. First, the effect of other noises than babble noise needs to be explored. A body of research about “speech-in-noise” has already explored a variety of different noises, such as vocoded speech, speech-shaped noise, and environmental noise, but it is an intelligibility-oriented study which did not explore the characteristics of “voice-in-noise”.

Second, the linguistic information of the target voice itself needs further investigation. There is an on-going study of pop-out voice perception by listeners who have no experience in the language used for the target voice [13], which suggests that pop-out voice is independent from language. That is, English monolingual listeners can detect pop-out voice spoken in Japanese. Then, what characteristics of the target voice did the listeners pick up even though they did not know the language? Paralinguistic and prosodic information in the voice may be relevant for this issue, which remains largely unexplored.

Third, the procedure of presenting forward babble noise first, and backward babble noise second might have a confounding effect on our conclusions in which the direction of the babble noise do not matter. Counter-balancing the material is necessary in the future experiments.

Application of pop-out voice is another possible area of future study. Voices that pop out in noisy environment is in itself useful for public announcement. Disaster and other emergency situations need such voice that addresses people who need to evacuate immediately.

5. ACKNOWLEDGMENT

This research was supported by JSPS Kakenhi 20H00291.

6. REFERENCES

[1] Amano, S., Kawahara, H., Banno, H., Maki, K., Yamakawa, H. 2022. Evaluation experiment of pop-out voices in babble noise. *Proc. of the Acoustical Society of Japan, Spring 2021 meeting*.

[2] Amano, S., Kawahara, H., Banno, H., Maki, K., Yamakawa, H. 2021. Acoustic features of pop-out voice in babble noise. *Acoustic Science and Technology*. 43, 105-112 (in Japanese).

[3] Kitahara, M., Tajima, K., Yoneyama, K., Kitamura, T., Kawahara, H., Amano, S. 2022. Detection of pop-out voice at various SN ratios. *Proc. of the Acoustical Society of Japan, Fall 2022 meeting* (in Japanese).

[4] Uchanski, R. M. 2005. Clear speech. In: D. B. Pisoni and R. E. Remez (eds), *The Handbook of Speech Perception*. Blackwell, 207-235.

[5] Miller, G. A. 1947. The masking of speech. *Psychol. Bull*. 44, 105–129.

[6] Bronkhorst, A. W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions, *Acust. Acta Acust*. 86, 117–128.

[7] Freyman, R. L., Balakrishnan, U., Helfer, K. S. 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition, *J. Acoust. Soc. Am*. 115, 2246–2256. doi:10.1121/1.1689343

[8] Simpson, S. A., Cooke, M. 2005. Consonant identification in N-talker babble is a nonmonotonic function of N, *J. Acoust. Soc. Am*. 118, 2775–2778. doi:10.1121/1.2062650

[9] Van Engen K. J. and Bradlow A. R. 2007. Sentence recognition in native- and foreign-language multi-talker background noise, *J Acoust. Soc Am*. 121, 519–526. doi: 10.1121/1.2400666.

[10] Van Engen, K. J. 2012. Speech-in-speech recognition: a training study. *Lang. Cogn. Processes* 27, 1089–1107. doi: 10.1080/01690965.2012.654644

[11] *Recommendation ITU-R BS*.1770-4. 2015.

[12] Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S., Shikano, K. 2000. Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components. *IEICE Trans. Inf. Syste*. (Jpn Ed.), J83-DII, 2077-2086 (in Japanese).

[13] Amano, S., Yamakawa, K., Kawahara, H. 2022. Effects of native language on detection of pop-out voice. *Proc. of the Acoustical Society of Japan, Fall 2022 meeting* (in Japanese).

L2 SPEECH SEGMENTATION FOR WORD RECOGNITION: THE ROLE OF LEXICAL STRESS AND SYLLABLE STRUCTURE

Juan J. Garrido-Pozú

Furman University
juan.garrido@furman.edu

ABSTRACT

Current understanding of speech perception suggests that the routines that listeners use to segment the speech signal for word recognition are language specific. English speakers exploit stress patterns for lexical segmentation whereas Spanish speakers use syllable structure. This study investigated L2 lexical segmentation and specifically examined whether L2 learners with a syllable-based L1 (Spanish) exhibit stress-based segmentation in L2 English. 81 English–Spanish learners and 72 Spanish–English learners completed a word-spotting task that manipulated lexical stress and syllable structure. Participants aurally identified English monosyllabic words embedded in disyllabic pseudowords.

Results showed that Spanish-English learners transferred L1 segmentation routines to L2 English. Spanish-English learners exhibited syllable-based segmentation of English regardless of differences in lexical stress, which was more prominent as L2 proficiency increased. English-Spanish learners showed stress-based segmentation regardless of differences in syllable structure. This study suggests that listeners develop only the segmentation routines that correspond to the phonological composition of their L1.

Keywords: L2 speech segmentation, L2 speech perception, lexical stress, syllable.

1. INTRODUCTION

For successful communication, listeners need to segment the speech stream to identify where words begin and end. Word boundaries do not emerge implicitly in the speech signal, but they are rather imposed by the listener using explicit procedures [2]. Each language has specific features that can be exploited for speech segmentation. Early studies found that statistical correlates of the sound patterns of words can be useful for speech segmentation due to the co-ocurrence of syllables within a word being more frequent than those across words [3, 12, 13]. In addition, phonetic cues specific to pre and post boundary phonemes can be exploited for segmentation [4, 17, 20]. In English, for example, voiceless stops are aspirated at word onsets but unaspirated at word offsets. Such cues can be useful for word boundary identification. Listeners can also use implicit knowledge of the structural restrictions of syllables in a particular language to identify the sequences that are permissible as onsets and offsets. Using syllable patterns is notably more efficient for segmentation of languages that have more consistent syllable structure, like Spanish, but not languages that have more variable syllable patterns, like English. In English, however, suprasegmental information provides reliable cues for segmentation [10].

2. STRESS-BASED SEGMENTATION

Prosodic features such as stress and rhythm can serve as cues for the identification of word boundaries in a language like English. The Metrical Segmentation Strategy (MSS) [5, 10] claims that placing word boundaries before stressed syllables is an efficient segmentation strategy that works for at least 90% of content words in English because most content words begin with a metrically stressed syllable. This strategy works for content words that begin with a stressed syllable, but not for content words that begin with an unstressed syllable or for function words. Consequently, the MSS assumes two separate strategies: content words are detected through stressed initial syllables whereas function words are detected through unstressed initial syllables [6].

Statistical and distributional analyses of the structure of the vocabulary of English supports the efficacy of treating strong syllables as cues for word onsets. Cutler & Carter [6] reported that in a computer-readable English dictionary with over 33,000 word entries, 73% of the words had a strong initial syllable and only 27% had a weak initial syllable. Cutler & Carter [6] analyzed a corpus of natural speech samples containing over 190,000 words from British English conversations. Over 90% of these words had a strong initial syllable. Clearly, words with strong initial syllables are more frequent than words with weak initial syllables in average speech contexts, which supports the reliability of strong syllables as cues for word onsets [5].

More evidence in favor of the MSS is found in Cutler & Norris [10]. English monolinguals completed a word-spotting task in which they listened to real monosyllabic words (e.g., *mint*) embedded in

disyllabic nonwords that had either two strong syllables (e.g., *mintayf*) or one strong and one weak syllable (e.g., *mintef*). Results revealed that listeners detected real words faster when the second syllable of the nonword was weak than when it was strong. Listeners had difficulty detecting *mint* in *mintayf* because they placed a boundary before the second strong syllable (*-tayf*) and treated it as the beginning of a new word. In *mintef*, on the other hand, there was no segmentation within the word because the second syllable is weak, facilitating recognition of the target *mint*.

3. L2 STRESS-BASED SEGMENTATION

The type of segmentation strategies that listeners use for a L2 may be constrained by the nature of their L1. Studies have reported that bilinguals with a syllable-based L1 like Italian or Spanish can develop L2 segmentation routines that are not syllable-based. However, bilinguals with a stress-based L1 like English struggle to develop syllable-based routines for a L2. In Cutler et al. [9], French-English and English-French early bilinguals completed two syllable monitoring tasks (in English and French) and one word-spotting task in English. Only the French-dominant bilinguals exhibited syllable-based segmentation in the French syllable-monitoring task, and no syllabic effects were observed in the English syllable-monitoring task. In the word-spotting task, only the English-dominant group showed stress-based segmentation. Bilinguals exhibited only the segmentation routine that was motivated by the phonology of their dominant language. French-dominant participants behaved like French monolinguals, and English-dominant participants behaved like English monolinguals; however, the lack of syllabic effects for French-dominant bilinguals with English words shows that these bilinguals may have employed two different segmentation routines, a syllabic one for their dominant language (French) and a non-syllabic one for their non-dominant language (English).

Studies with L2 learners of English with different L1s have reported that L2 learners exhibit evidence of stress-based segmentation. In Katayama [15], the results of a syllable monitoring task revealed that English monolinguals and Japanese learners of English identified target syllables faster in disyllabic words with initial stress than with final stress, which is evidence of stress-based segmentation. Similarly, in Cutler & Shanley [11], Arabian and Chinese learners of English exhibited stress-based segmentation in a word-spotting task in English where they detected monosyllabic words more accurately when they were embedded into disyllabic pseudowords with initial stress than with final stress. These studies showed that L2 learners who speak a L1 that is not stress-based still exhibited L2-specific stress-based segmentation, suggesting that L2 learners can develop segmentation strategies that are not necessarily motivated by L1 phonology. However, it is still unclear whether L2 learners of English with a syllable-based L1 like Spanish can develop L2 stress-based segmentation. Notably, how segmentation routines change as L2 proficiency increases also needs further investigation.

4. THE PRESENT STUDY

The present study investigated whether L2 learners with a syllable-based L1 (Spanish) exhibit evidence of L2 stress-based segmentation. The specific goals of the present study are to test whether L2 learners exhibit stress-based segmentation of English, to test whether L2 learners exhibit syllable-based segmentation of English, and to assess whether L2 segmentation changes as L2 proficiency increases. The study is driven by the following research questions:

1. Are L2 learners sensitive to lexical stress during segmentation of English?
2. Are L2 learners sensitive to syllable structure during segmentation of English?
3. Does L2 proficiency modulate learners' sensitivity to lexical stress and/or syllable structure during segmentation of English?

5. METHODOLOGY

5.1. Participants

The participants included 81 English L1 learners of Spanish and 72 Spanish L1 learners of English. The English L1 subjects were born and raised in the United States and had no intensive exposure to L2 Spanish before puberty. The Spanish L1 subjects were born and raised in Peru and had no intensive exposure to L2 English before puberty. The mean age of English L1 subjects was 22.2 (*SD* = 5.06) and of Spanish L1 subjects was 31.4 (*SD* = 6.76). English L1 subjects started L2 instruction at age 11.9 (*SD* = 6.33), studied L2 Spanish for 8.51 years (*SD* = 3.82), and were exposed to L2 Spanish 18% of the time per week (*SD* = 15.3). Spanish L1 subjects started L2 instruction at age 15.2 (*SD* = 5.12), studied L2 English for 4.67 years (*SD* = 4.03), and were exposed to L2 English about 31.6% of the time per week (*SD* = 20.7). All participants completed the LexTALE [18] and the LexTALE-ESP [14] which provided a measure of participants' proficiency in English and Spanish respectively.

5.2. Stimuli and procedure

Participants completed four tasks in the following order: a background questionnaire, a word-spotting task, the LexTALE test, and the LexTALE-ESP test. Participants completed all tasks in one session which lasted about 40 minutes.

5.2.1. Word-spotting task

The word-spotting task (WST) followed the design in Cutler et al. [9]. Participants listened to isolated English pseudowords (e.g., *mintef*, *mintayf*) and indicated whether the pseudowords contained a real English word embedded at the beginning (e.g., *mint* in *mintef*, *jump* in *jumpev*). In each trial, participants first saw the question "*Does it begin with a real word?"* at the top center of the screen and the words *YES* and *NO* on each side of the screen. 500 ms later, participants heard a pseudoword (e.g., *mintef*). They pressed the left arrow key to answer *YES* and the right arrow key to answer *NO*. If participants answered *YES*, they were asked to type the real English word that they had identified. The task recorded accuracy and response times.

Participants listened to a total of 104 disyllabic pseudowords: 64 experimental items, and 40 filler/distractors. The 64 experimental items began with a monosyllabic real English word. Half of the experimental items contained a real word with a CVCC structure (e.g., *mint*, *risk*) while the other half had a CVC structure (e.g., *thin*, *kiss*). Real words were made into pseudowords by adding segments to the end. CVCC words were attached to a final VC sequence with either a strong or weak syllable (e.g., *mint*: *mintayf*, *mintef*; *risk*: *riskeeb*, *riskeb*). CVC words were attached to a final CVC sequence with either a strong or weak syllable (e.g., *thin*: *thintayf*, *thintef*; *kiss*: *kisskeeb*, *kiskeb*). CVCC pseudowords were matched with a CVC pseudoword with which they shared the same ending. Fillers and distractors consisted of disyllabic pseudowords without any embedded real words. The pseudowords were recorded by a female native speaker of US English from New Jersey. The speaker was instructed to read the items at a normal rate and did not have knowledge of the target words. Target words were not recorded as list initial or list final to avoid intonational differences.

5.3. Data analysis

The data were analyzed using mixed effects models. The WST measured response times (RTs) and accuracy. RTs were analyzed using linear mixed effects models while correct responses were analyzed using logistic mixed effects models. The independent variables were coded as *stress type* (initial, final), *word type* (CVC, CVCC), and *proficiency* (treated as a continuous variable). The models had random intercepts by participant and by item and included all interactions. Main effects and interactions were assessed by partitioning the variance hierarchically via nested model comparisons. The statistical analyses were carried out using R [22]. The analyses used the packages *lme4* [1], *lmerTest* [16], and *emmeans* [19].

6. RESULTS

For English L1 subjects, there was a main effect of *stress type* on the proportion of correct responses ($\chi2(1) = 3.87$, $p = .04$). English L1 subjects identified more accurately words embedded in pseudowords with initial stress than final stress (see Fig. 1). The log odds of identifying the target word correctly when the pseudoword had final stress were 1.07 (β: 1.07, $SE = 0.22$, $p < .001$), but when the pseudoword had initial stress, the log odds increased by 0.51 (β: 0.51, $SE = 0.26$, $p = .04$). The models also revealed a main effect of *stress type* on RTs ($\chi2(1) = 4.49$, $p = .03$). Overall, English L1 subjects identified real words embedded in pseudowords with initial stress about 169ms faster than those in pseudowords with final stress ($SE = \pm.08$) (see Fig. 2). There was not an effect of *word type* on the proportion of correct responses or RTs of English L1 subjects.

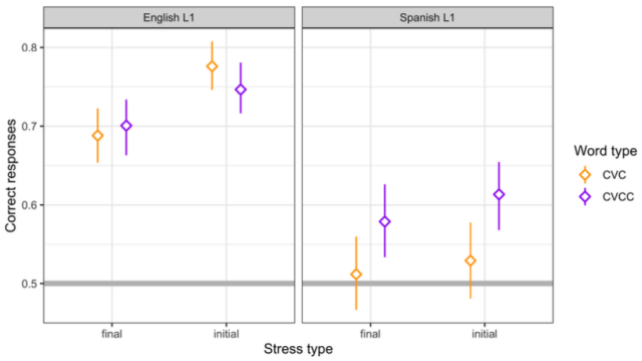


Figure 1: Proportion of correct responses of English L1 subjects (left panel) and Spanish L1 subjects (right panel) as a function of stress type and word type.

For Spanish L1 subjects, there was a main effect of *proficiency* ($\chi2(1) = 21.03$, $p < .001$) and *word type* ($\chi2(1) = 4.06$, $p = .04$) on the proportion of correct responses. There was also a two-way interaction between *word type* and *proficiency* ($\chi2(4) = 4.06$, $p < .04$). Spanish L1 subjects identified more accurately real words with a CVCC structure than with a CVC structure (β: 0.45, $SE = 0.22$, $p = .04$) (see Fig. 1 and 3). Overall, the log odds of identifying the target word correctly increased by 0.8 as English L2 proficiency increased ($SE = \pm.16$). Syllable structure played a role

at higher levels of L2 proficiency but not much at lower levels (see Fig. 3). The models yielded no effect of *stress type* on the proportion of correct responses of Spanish L1 subjects and no main effects or interactions on RTs. On average, the Spanish L1 group exhibited slower responses than the English L1 group overall (see Fig. 2), a difference of approximately 830 milliseconds (SE = ±.115) and lower rates of correct responses (M = .56, SD = .49) than the English L1 group (M = .73, SD = .44) (β: -0.93, SE = 0.22, p < .001) (see Fig. 1).

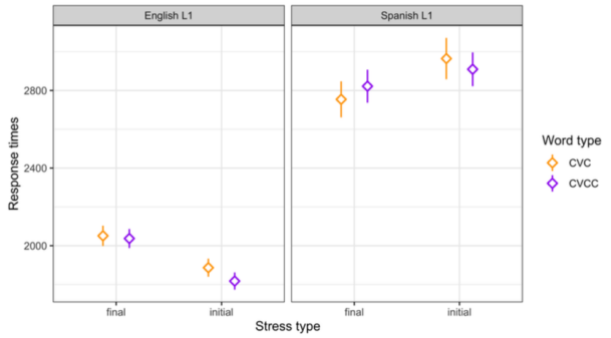


Figure 2: Response times in milliseconds of English L1 subjects (left panel) and Spanish L1 subjects (right panel) as a function of stress and word type.

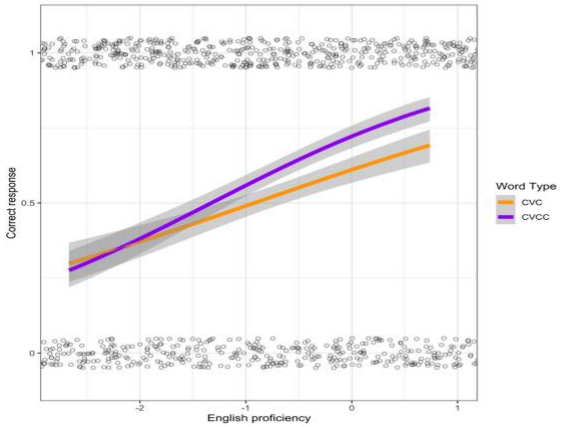


Figure 3: Proportion of correct responses of Spanish L1 subjects as a function of L2 English proficiency and word type.

7. DISCUSSION

This study investigated whether L2 learners develop separate segmentation routines that are specific to the phonology of their L2. Specifically, the study tested whether L2 learners with a syllable-based L1 exhibit L2 stress-based segmentation in a word-spotting task that manipulated lexical stress and syllable structure. The first research question asked whether L2 learners use lexical stress for segmentation of English. Lexical stress affected the responses of English L1 learners of Spanish but not of Spanish L1 learners of English. English L1 subjects identified real words embedded in pseudowords with initial stress faster and more accurately than in pseudowords with final stress, which is considered evidence of stress-based segmentation and coincides with previous studies reporting similar results in English monolinguals [5, 6, 21] and English-dominant bilinguals [8, 9]. The results support the Metrical Segmentation Strategy [10] which suggests that strong syllables can serve as reliable cues for locating possible word onsets. However, the results of this study differ from other studies showing evidence of stress-based segmentation in L2 learners with Japanese L1 [15], Mandarin Chinese L1, and Arabic L1 [11]. Such difference may be due to typological differences between Spanish (syllable-based) and the other languages in question.

The second research question asked whether L2 learners use syllable structure for segmentation of English. Only Spanish L1 subjects showed evidence of syllable-based segmentation of English. The results suggest that learners transfer L1 segmentation routines to a L2, which supports previous studies claiming that L2 segmentation is constrained by L1 segmentation [7, 8, 9]. However, the direction of the syllabic effects was not given as expected. Spanish L1 subjects identified CVCC words more accurately than CVC words, which may indicate that they placed the syllable boundary after the medial consonant cluster (e.g., *thint.ayf /thint.ef; mint.ayf/mint.ef*), facilitating recognition of CVCC words. This syllabification pattern differs from the pattern expected for Spanish and could be evidence of L2-specific syllabification.

The last question focused on the role of L2 proficiency. English L1 subjects did not exhibit syllabic effects even at higher levels of L2 Spanish proficiency. Spanish L1 subjects exhibited stronger syllabic effects at higher levels of English proficiency. Considering the complexity of a word-spotting task, the lack of syllabic effects in lower proficiency learners of English may correspond to difficulties in lexical recognition rather than the absence of a syllable-based segmentation routine.

8. CONCLUSION

This study showed that whereas English L1 learners of L2 Spanish exhibited stress-based segmentation of English, Spanish L1 learners of L2 English displayed syllable-based segmentation of English. The study brings support to the Metrical Segmentation Strategy [10] and suggests that segmentation routines are language specific and L2 segmentation is constrained by L1 segmentation even at higher levels of L2 proficiency.

9. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

[2] Carroll, S. E. 2004. Segmentation: learning how to hear words in the L2 speech stream. *Transactions of the Philological Society*, *102*(2), 227–254.

[3] Chomsky, N. 1955. *The logical structure of linguistic theory*. Cambridge, MA: Cambridge University Press.

[4] Christie, W. M. 1974. Some cues for syllable juncture perception in English. *J. Acoust. Soc. Am.*, *55*(4), 3.

[5] Cutler, A., & Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*(2), 218–236.

[6] Cutler, A., & Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3), 133–142.

[7] Cutler, A., Mehler, J., Norris, D., & Segui, J. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*(4), 385–400.

[8] Cutler, A., Mehler, J., Norris, D., & Segui, J. 1989. Limits on bilingualism. *Nature*, *340*(6230), 229–230.

[9] Cutler, A., Mehler, J., Norris, D., & Segui, J. 1992. The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, *24*(3), 381–410.

[10] Cutler, A., & Norris, D. 1988. The Role of Strong Syllables in Segmentation for Lexical Access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121.

[11] Cutler, A., & Shanley, J. 2010. Validation of a training method for L2 continuous speech segmentation. *Interspeech*, 1844–1847.

[12] Harris, Z. S. 1955. From phoneme to morpheme. *Language*, *31*, 190–222.

[13] Hayes, J., & Clark, H. 1970. Experiments in the segmentation of an artificial speech analog. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 221–234). New York: Wiley.

[14] Izura, C., Cuetos, F., & Brysbaert, M. 2014. Lextale-ESP: A test to rapidly and efficiently assess the Spanish vocabulary size. *PSICOLOGICA*, *35*(1), 49–66.

[15] Katayama, T. 2015. Effect of Phonotactic Constraints on Second Language Speech Processing. *I-Perception*, *6* (6).

[16] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

[17] Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. *Phonetica*, *5*(s1), 5–54.

[18] Lemhöfer, K., & Broersma, M. 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343.

[19] Lenth, R. V. 2022. *Emmeans: Estimated marginal means, aka least-squares means*.

[20] Nakatani, L., & Dukes, K. 1977. Locus of segmental cues for word juncture. *The Journal of the Acoustical Society of America*, *62*(3), 714–719.

[21] Norris, D., & Cutler, A. 1988. The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, *43* (6), 541–550.

[22] R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tracking speaker-specific speech rate: Habitual vs. local influences on English stop voicing*

Connie Ting[1,2] & Yoonjung Kang[3]

[1]McGill University, [2]Centre for Research on Brain, Language and Music, [3]University of Toronto Scarborough
connie.ting@mail.mcgill.ca, yoonjung.kang@utoronto.ca

ABSTRACT

Studies have shown listeners can track individual speakers' speech rate and adjust their perception of duration contrasts. However, studies show mixed results regarding whether listeners adjust their perception similarly for vowels vs. consonants. One study showed that after hearing a dialogue between a fast and slow speaker, listeners adjusted their perception of German vowel length contrasts (/a/ vs. /a:/) in a speaker-specific way. A replication study showed no such rate effect for the English voicing contrast. In our study, English listeners heard a dialogue between a fast and slow speaker, with a greater rate difference compared to previous studies, followed by two identification tasks. The first included isolated stop-initial syllables (e.g. pig~big) manipulated along a VOT continuum. The second included the same syllables embedded in a fast or slow carrier sentence. Our results showed listeners adjusted their perception of word-initial stops in both the rate-manipulated carrier sentences and dialogue condition.

Keywords: speech rate, VOT, speech perception, individual variation

1. INTRODUCTION

Research has shown that speakers can vary substantially in speech rate not only across speakers but also within individuals [1, 2]. Speech rate variation is particularly pertinent to durational properties, since faster speech compresses utterance duration while slower speech expands duration. Considering the amount of variation that occurs, listeners must be able to normalize for speech rate to correctly interpret the speech signal from their interlocuter. That is, listeners need to perceive durational acoustic cues relative to the surrounding speech rate and/or to the speech rate that the listener associates with a particular speaker.

The effects of speech rate on listeners' perception have been studied at least as far back as the 1960's [3]. In general, the reported effect is that a slower contextual speech rate leads listeners to perceive an ambiguous unit of speech as relatively short, whereas a faster contextual speech rate leads listeners to perceive the same stimulus as relatively long [4, 5, 6, 7]. Previous work has also shown that listeners can track durational properties in a speaker-specific fashion [8].

Importantly, however, speech rate effects may differ across speech contexts. Many studies have shown that speech perception is influenced by relatively short adjacent contexts, such as an adjacent phoneme [7, 9] or a carrier sentence containing the target sound/word [10]. More recent studies have shown that listeners' perception can be influenced by the global or habitual speech rate of a speaker [11, 12]. That is, listeners can track the average speech rate of a speaker over a longer period of time. For example, Reinisch [12] examined speech rate effects in the context of a dialogue between two speakers. After hearing a 2-minute dialogue between two female native speakers of German, varying in rate (fast vs. slow) and order (first vs. second speaker), listeners completed a phonetic categorization task in which they categorized words of minimal pair continua differing in the /a/-/a:/ duration contrast. The stimuli were presented as words in isolation, intermixed across speakers. Their results showed that listeners' perception of the vowel contrast shifted depending on the speech rate of the speaker. Stimuli produced by the fast speaker in the preceding dialogue elicited more /a:/ responses compared to the slow speaker.

These results suggest that listeners are able to track speaker-specific speech rate information in a dialogue context and make use of this information in later perception. However, it is possible that the results are due to the presence of target vowels in the dialogue which could not be avoided, rather than perceptual adjustments due to speech rate. A similar study [13] aimed to replicate Reinisch's findings with English listeners' perception of a consonantal contrast, namely voice onset time (VOT). They presented listeners with a dialogue between two speakers, matching the rate manipulation of Reinisch's study [12], but crucially did not include any instances of stressed syllable-initial voiceless stops in the dialogue. Results from their categorization task showed no effect of speech rate.

Since these studies examined two different types of contrasts (vowels vs. consonants) and also differed in the presence or absence of the target structure in

the dialogue, there are multiple reasons why the two studies diverged in their findings. Given the lack of speech rate effect on a consonantal contrast, it remains unclear whether a speaker-specific speech rate effect can be found for a vowel length contrast using a dialogue which omits all instances of the target stimuli. In other words, it is possible that no speech rate effect emerges when listeners are not exposed to target stimuli in the dialogue, regardless of the kind of durational contrast being examined.

On the other hand, it is possible that the different results are due to the particular type of contrast being examined. One possibility is that the degree of rate difference contributes to the presence of rate normalization and that a greater rate difference is required to elicit rate normalization of consonantal contrasts than of vowel contrasts. That is, the lack of the speech rate effect found for the consonantal contrast may be attributed to a rate difference between speakers that was not salient enough.

Another difference between [12] and [13] is the former used multiple minimal pairs as target words while the latter used a single monosyllable word pair, making the task far less natural and more prone to decay of dialogue rate effects through the experiment.

The current study replicates [13] with two crucial modifications – making the rate difference larger and using multiple word pairs. In addition, we included another identification task following the main task, whereby the same word stimuli used in the main task is presented embedded in a rate-manipulated carrier sentence, where previous studies [14] have shown robust rate effects and we should expect the same. Significant rate effects in the sentential condition could assure us that the target stimuli manipulation is natural and the online participants are performing the tasks as intended. The results of our main tasks can thus be evaluated in a firmer footing.

2. METHODS

2.1. Participants

80 listeners were recruited through the Amazon Mechanical Turk (Mturk) platform and were paid for their participation. All listeners were self-identified native speakers of American English and reported normal speech and hearing.

2.2. Stimuli

Two male speakers (M1, M2) individually recorded a 460-word dialogue between two people, which was scripted such that no stressed syllable-initial voiceless stops were included. Each speaker recorded both roles of the dialogue (A and B) and were instructed to read the dialogue at a comfortable rate.

The dialogue recordings were segmented at phrase boundaries and labelled according to the speaker role (A or B). Phrase durations were measured to determine the natural speech rate for each speaker. Phrase durations were then manipulated to create two speech rate conditions (fast and slow), by compressing or expanding the speech to be 20% shorter or longer than the average of the two speakers' natural speech rate (exaggerated from rate conditions used in previous studies, which ranged from 10% shorter to 15% longer [12, 13, 14]). Manipulated phrases were spliced back together leaving 250 ms of silence between the utterances. After duration manipulation, the resulting dialogue was ~2 minutes. Four versions of the dialogue were used in this study such that each speaker was heard in each role (A and B) and at each speech rate (fast and slow).

Each speaker also recorded 5 repetitions of 8 words (4 pairs: time-dime, pig-big, toe-doe, pan-ban) in a carrier sentence: 'Now I will say ___'. A carrier sentence with the target word in final position was chosen following previous studies which have shown that VOT is perceived by listeners relative to preceding speech rate context [13, 14]. Measurements were made for the duration of the carrier sentence, as well as the closure, aspiration, vowel, and coda (when applicable) of the target word to determine the average duration of each speaker's natural production of the carrier sentence and target words. To make the sentence stimuli as comparable as possible, one repetition of the carrier sentence was extracted from each speaker's production such that the sentences sounded most similar between speakers in terms of overall pitch contour and included minimal pause between words. Base tokens of the chosen carrier sentences were created by manipulating the duration of the sentence as well as the closure duration between the carrier sentence and target word to be the mean duration across speakers.

For the target word stimuli, a base token for each word pair was created by splicing together the aspiration of a voiceless token and the vowel+coda from its voiced counterpart, all originally produced in the same carrier sentence context. The durations of the vowel and the coda were manipulated to be equal to the word-pair means across speakers. The VOT duration of base tokens was manipulated to create VOT continua ranging from 0 to 70 ms in 8 steps.

For sentential condition stimuli, the carrier sentence was manipulated to create two rate conditions (fast and slow) similar to the dialogue manipulation (20% faster or slower) and the target words were spliced back onto one of the rate-manipulated carrier sentences. In total, the manipulations were used to create 4 dialogue conditions, 64 words in isolation (8 VOT steps * 4

word pairs * 2 speakers) and 128 sentences (8 VOT steps * 4 word pairs * 2 speakers * 2 sentence rates).

Each speaker also recorded 5 repetitions of the 8 words in isolation. The third repetition for each word was used for control stimuli (16 total: 8 words * 2 speakers). These tokens were not manipulated and were used to represent naturally produced tokens.

A pretest was run to determine the range of the VOT continuum that would be sufficient in obtaining a balance of voiced vs. voiceless responses. 30 participants who did not take part in the main experiment participated online through Mturk and were paid for their participation. The final VOT continuum used for the main experiment ranged from 10 to 70 ms in 7 steps, as stimuli beyond these boundaries consistently elicited expected responses.

2.3. Procedure

The experiment was built and hosted on the Gorilla platform (gorilla.sc) [15]. Before beginning the experiment, listeners completed an audio test to ensure that the sound was playing properly and at a comfortable volume.

Each participant heard one of four versions of the dialogue. Each version was heard by 20 participants. After hearing the dialogue, participants completed two categorization tasks. The first presented words in isolation. On each trial, participants saw a fixation cross for 500 ms, with 100 ms of pause before and after the fixation cross. They then heard an individual word and were asked to click the corresponding word on the screen. The stimuli for the two speakers were presented intermixed. For each trial, two words were displayed on the screen: the button on the left side of the screen always corresponded with the word with a voiceless stop (e.g., time) while the button on the right side of the screen always corresponded with the voiced counterpart (e.g., dime). After each response by button click, the next trial would begin. Each stimulus was repeated 2 times, resulting in a total of 112 trials for each listener (7 VOT steps * 4 word pairs * 2 speakers * 2 repetitions).

After hearing the words in isolation, listeners were presented with a second categorization task with the same procedure, but with the target word embedded in a carrier sentence. Listeners saw a fixation cross, then heard the sentence 'Now I will say ___', and were asked to indicate the last word of the sentence by clicking the corresponding word on the screen. The buttons were arranged in the same way as the previous task. Each stimulus was repeated once, resulting in a total of 112 sentential trials (7 VOT steps * 4 word pairs * 2 speakers * 2 sentence rates).

Note that speech rate was a between-subject factor in the dialogue condition; the same speaker was heard as fast by some participants but as slow by others. In the sentential condition, all participants heard the same stimuli, and rate was a within-subject factor.

Finally, listeners were presented with a short block of control stimuli with the same procedure as the previous tasks. The control block consisted of 16 tokens (8 words * 2 speakers) randomized. These were unmanipulated natural tokens, intended to elicit the correct response by listeners and used as a control such that listeners who did not correctly identify all of the natural tokens were excluded from the analysis (n=4). The full experiment took approximately 20 minutes to complete.

3. RESULTS

All statistical analyses were conducted in R [16]. Starting with the dialogue condition, in which listeners heard the dialogue then categorized words in isolation, Figure 1 shows the proportion of voiceless responses over the VOT continuum for the fast versus slow speaker in the dialogue aggregated over all four versions of the dialogue. Figure 1 shows a shift in category boundary, particularly in the ambiguous region of the VOT continuum (~30 ms), such that the same VOT duration was perceived more often as voiceless for speakers' fast speech than slow speech. This is represented by the solid line, representing fast speech, appearing above the dotted line, representing slow speech.

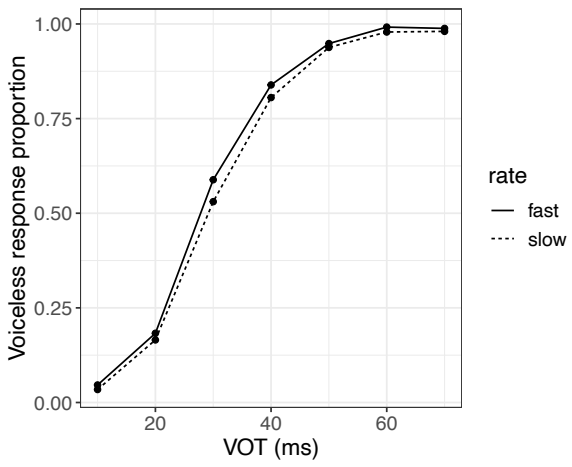


Figure 1: Proportion of voiceless responses over the VOT continuum for the fast (solid line) versus slow (dotted line) speaker in the dialogue.

A logistic mixed-effects model was fit using the *glmer* function of the *lme4* package [17]. The model included response (voiceless coded as 1, voiced as 0) as a dependent variable and VOT (ms, centred), SPEAKER (M1 = -0.5, M2 = 0.5), SPEECH RATE (fast = -0.5, slow = 0.5), and interactions between SPEECH

RATE and VOT and between SPEECH RATE and SPEAKER, as fixed factors. By-PARTICIPANT and by-WORD PAIR random intercepts were included, in addition to by-PARTICIPANT random slope adjustments to VOT. The results of the model showed a significant effect of VOT (b_{VOT} = 0.20, z = 27.41, p <0.001), with more /p/ responses as VOT duration increased, as expected. Results also showed a significant effect of SPEAKER ($b_{Speaker}$ = -0.80, z = -9.94, p = <0.001), indicating that speaker M2 had significantly less voiceless responses than speaker M1. The results also showed a significant effect of SPEECH RATE ($b_{Speech\ Rate}$ = -0.40, z = -4.02, p = <0.001), indicating that when compared across dialogue conditions, a speaker's fast speech condition elicited more /p/ responses compared to their slower condition. There were no significant interactions ($b_{Speech\ Rate * VOT}$ = -0.01, z = -1.65, p = 0.099; $b_{Speech\ Rate * Speaker}$ = 0.06, z = 0.15, p = 0.885).

We turn next to the sentential condition, in which listeners heard target words embedded in a carrier sentence which varied in rate (fast or slow). Figure 2 shows the proportion of voiceless responses over the VOT continuum for the fast versus slow sentential contexts. The results of this task as plotted in Figure 2 show a consistent difference in VOT categorization due to speech rate, such that a faster sentential context elicited more voiceless responses. This is shown by the solid line, representing fast speech, appearing above the dotted line, representing slow speech.

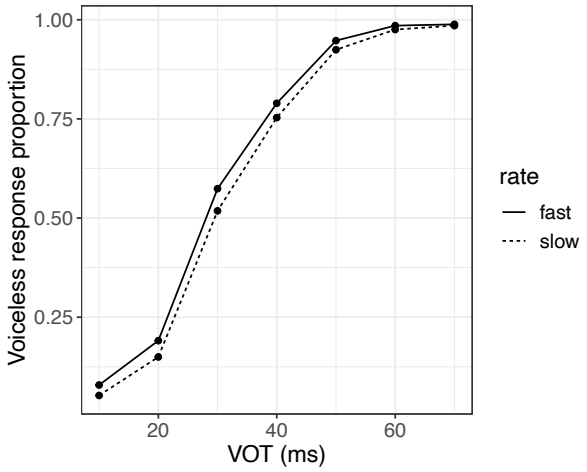


Figure 2: Proportion of voiceless responses over the VOT continuum for fast (solid line) versus slow (dotted line) sentential contexts.

The model specification for the sentential task was identical to the dialogue model. The results showed a significant effect of VOT (b_{VOT} = 0.17, z = 27.60, p <0.001), with more /p/ responses as VOT duration increased, as expected. Results also showed a significant effect of SPEAKER ($b_{Speaker}$ = -0.25, z = -3.35, p = <0.001), indicating that speaker M2 had significantly less voiceless responses than speaker M1. The results also showed a significant effect of SPEECH RATE ($b_{Speech\ Rate}$ = -0.33, z = -3.72, p <0.001), indicating that the fast speech rate sentences elicited more voiceless responses compared to the slow speech rate sentences. There were no significant interactions ($b_{Speech\ Rate * VOT}$ = 0.0003, z = 0.05, p = 0.962; $b_{Speech\ Rate * Speaker}$ = -0.24, z = -1.61, p = 0.108).

4. DISCUSSION

The current study tested whether listeners keep track of individual speakers' speech rate and whether rate normalization differs across speech rate contexts. In the first part of the experiment, listeners heard a dialogue between two male speakers and then categorized individual words. The dialogue was created such that it provided listeners with each speaker's speech rate in direct contrast (one fast, one slow) and did not include any of the target stimuli (stressed word-initial stops). Results from the categorization task following the dialogue showed a significant effect of speech rate, suggesting that the speaker-specific speech rate affected listeners' perception of the English voicing contrast.

In the second part of the experiment, listeners categorized target words which were embedded in a carrier sentence that was either fast or slow. The results of this task also revealed a significant effect of speech rate, showing that faster speech in the sentential context elicited more voiceless responses. The significant effect of speech rate found in both contexts suggests that rate normalization in listeners' perception of consonantal contrasts can occur in both local (sentential) and habitual (dialogue) contexts.

Note that the rate manipulation in this study was greater than that used in previous studies (see [12] for vowels, [13] for consonants) allowing us to test whether a more exaggerated rate difference would lead to a clear speech rate effect. Given that a smaller speech rate difference has led to rate effects for vowel contrasts but not consonant contrasts [12, 13], the significant rate effect in the current study for English VOT using a more exaggerated rate difference could indicate that speech rate normalization does not behave in exactly the same way across different durational contrasts. Additionally, the significant rate effects found for vowel contrasts might be attributable to the presence of target structure in [12], rather than speech rate alone. More work is required to test the presence and degree of rate normalization across speech rate contexts and different durational contrasts in the absence of target structure tokens in the rate-defining speech context.

5. REFERENCES

[1] Miller, J. L., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41, 215-225.

[2] Jacewicz, E., Fox, R. A., Wei, L. 2010. Between-speaker and within-speaker variation in speech temp of American English. *J. Acoust. Soc. Am.* 128, 839-850.

[3] Pickett, J., Decker, L. R. 1960. Time factors in perception of a double consonant. *Language and Speech* 3, 11-17.

[4] Miller, J. L. 1987. Rate-dependent processing in speech perception. In Ellis, A. W. (ed.), *Progress in the psychology of language*. London: Erlbaum, 119-157.

[5] Miller, J. L., Dexter, E. R. 1988. Effects of speaking rate and lexical status on phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 369-378.

[6] Kidd, G. R. 1989. Articulatory-rate context effects in phoneme identification. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 736-748.

[7] Sawusch, J. R., Newman, R. S. 2000. Perceptual normalization for speaking rate: II. Effects of signal discontinuities. *Perception & Psychophysics* 62, 285-300.

[8] Allen, J. S., Miller, J. L., DeSteno, D. 2003. Individual talker differences in voice-onset-time. *J. Acoust. Soc. Am.* 113, 544-552.

[9] Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1074-1095.

[10] Dilley, L. C., Pitt, M. A. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21, 1664-1670.

[11] Baese-Berk, M. M., Heffner, C. C., Dilley, C. L., Pitt, M. A., Morrill, T. H., McAuley, J. D. 2014. Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science* 25, 1546-1553.

[12] Reinisch, E. 2016. Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics* 37, 1397-1415.

[13] Ting, C., Kang, Y. 2019. The effect of habitual speech rate on speaker-specific processing in English stop voicing perception. *Proc. 19th ICPhS* Melbourne, 3230-3234.

[14] Toscano, J. C., McMurray, B. 2015. The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience* 30(5), 529-543.

[15] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[16] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[17] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[18] Fox, J., Weisberg, S. 2019. An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA.

[19] Newman, R. S., Sawusch, J. R. 2009. Perceptual normalization for speaking rate: III. Effects of the rate of one voice on perception of another. *J. Phon.* 37, 46-65.

*This study was funded by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209).

EVALUATING THE IMPACT OF DISFLUENCIES ON THE PERCEPTION OF SPEAKER COMPETENCE USING NEURAL SPEECH SYNTHESIS

Ambika Kirkland[1], Marcin Włodarczak[2], Joakim Gustafson[1], Éva Székely[1]

[1]KTH Royal Institute of Technology, [2]Stockholm University
kirkland@kth.se, wlodarczak@ling.su.se, jkgu@kth.se, szekely@kth.se

ABSTRACT

Perceptions of speaker competence can be influenced by various factors, including the presence of disfluencies such as false starts and different types of repetitions. In this study, we generated utterances with false starts and each of three different types of repetitions using a text-to-speech system with implicit prosody control trained on spontaneous speech. A web-based listening task was conducted to evaluate the impact of these different types of disfluencies on perceptions of speaker competence. False starts were found to have the greatest negative impact, consistent with previous research showing a relationship between processing fluency and perceived competence. The results of this study have implications for public speaking training, as speakers can work on minimizing false starts and other disfluencies to improve perceived competence. Additionally, understanding the impacts of different types of disfluencies can help speakers choose strategies for minimizing their use and improving their overall fluency and effectiveness in communication.

Keywords: Speech perception, speech synthesis, public speaking, spontaneous speech, disfluencies

1. INTRODUCTION

Disfluencies, such as filled pauses, repetitions and speech repairs, are ubiquitous in spontaneous speech and play an important role in maintaining the smooth flow of conversation when delays occur in speech planning [1]. Disfluencies can inform listeners about upcoming delays and provide an insight into speech planning [2]. Nonetheless, there is evidence that disfluencies negatively impact a range of judgments about speakers, including their perceived competence [3, 4]. This is consistent with more general evidence that disrupting the fluent processing of information is one factor that can make the source of this information seem less competent [5]. However, the term “disfluency” encompasses a wide range of speech behaviors and not all of these affect the listener in the same way.

False starts, also called speech repairs, occur when a speaker begins a word or phrase and then stops and starts over to correct it. These follow the general pattern of *reparandum* (the original words), *editing term* (e.g. "I mean" or "uh"), and *alteration* (the correction) [6]. False starts have been found to negatively impact word monitoring latency when they occur mid-sentence, likely due to the disruption they cause in the flow of the speech [7]. In contrast, word repetitions seem to be less disruptive to comprehension. Operantly conditioning speakers to reduce their production of silent pauses increases the incidence of function word repetition [8], which may indicate that repeating function words plays a similar role to that of silent pauses in buying time while planning the next part of the utterance. On the other hand, fluency failures on content words do not seem to serve the same function and may occur when the speaker begins saying a word before the entire plan is ready [8].

Despite the link between speech disfluencies and perceived competence, and the evidence that repetitions and false starts reflect different cognitive processes and affect processing fluency to different degrees, some questions remain. Do specific types of repetitions have different effects on perceived competence? Do differing impacts on processing fluency actually translate to differing impacts on competence perception?

One challenge in trying to answer these questions is that disfluencies are peculiar to and characteristic of spontaneous speech [9]. This means that as an object of investigation, they are not very well suited to the common strategy of using lab-recorded speech to create stimuli for perception experiments. This is a more general dilemma in studying speech perception. Differences between read and spontaneous speech are apparent in turn-taking behaviors [10], stress position and boundaries between tone units [9], articulation rate [11], vowel reduction [11], and f0 range [12], and are even reflected in the neurophysiological correlates of speech processing [13]. Using spontaneous speech

would be ideal but this requires relinquishing control over the content of utterances. Many spontaneous speech corpora exist but it is highly unlikely that one could find specific utterances in these corpora which meet a given set of criteria in terms of structure and content and of which there are several instances from the same speaker varying only in terms of which disfluencies are present.

One potential alternative is to use neural text-to-speech (TTS) trained on spontaneous speech to create stimuli for speech perception experiments. Neural TTS has come to rival human speech in quality and naturalness in recent years [14], making spontaneous TTS an increasingly viable alternative to natural speech. Many stimuli can be produced without the time and expense of recording speech in a lab, and can be easily tailored to the specific demands of the experiment. Some spontaneous neural TTS systems also grant a measure of control over various prosodic features as well as the location and frequency of disfluencies such as filled pauses [15, 16]. Such systems have been leveraged recently to investigate the role of filled pauses and their interaction with prosodic features in the perception of uncertainty [17, 18].

In the present study, we used spontaneous TTS to investigate how false starts and different kinds of repetitions (function word repetitions, verb phrase repetitions, or the repetition of key content words) impact the perception of speaker competence. We hypothesized that false starts would have the most negative impact on perceived competence given their disruption of listening comprehension, while function word repetitions would have the smallest impact, with the other types of repetitions falling somewhere in between.

2. METHOD

2.1. Stimulus creation

2.1.1. Data and system

The training data for the TTS system were derived from a public-domain technology podcast which involves two male speakers of American English discussing technology news and reviewing products. Audio from the speaker with the most airtime was used in the training corpus. The audio was segmented into breath groups using a speaker-dependent breath detection method [19]. The Google Cloud Speech API [20] was used for the initial transcription. Filled pauses such as *uh* and *um* were identified using IBM Watson Speech to Text, combined with the output of the Gentle forced aligner [21]. These tokens were then inserted into the Google API transcription. Then the transcriptions were manually corrected, as especially the pronunciations of technology acronyms were not good enough. Finally, mean f0 and mean speech rate were automatically computed for each breath group in the training corpus using the Wavelet Prosody Analyzer toolkit [22].

The text-to-speech system was based on the sequence-to-sequence Tacotron 2 TTS engine, modified with a style-unit-level prosody control method similar to that described in [23] and [18] to enable implicit control of f0 and speech rate at inference. As opposed to setting specific values for speech rate and fundamental frequency, the system can be prompted to produce an f0 or speech rate based on a target percentile of the sampling distribution of these values in the training data. The neural vocoder HiFi-GAN [24] was used to decode the speech signal.

2.1.2. Stimuli

The sentences synthesized for the experiment all consisted of definitions of scientific terms with the general form: "The [description of the phenomenon] is called the **[eponym]**.", e.g., "The proportionality factor that relates temperature to kinetic energy in a gas is called the **Boltzmann constant**."

These sentences were created with a few factors in mind. First, we wanted to ensure that the repetition or false start would occur at a point in each sentence that was as similar as possible in terms of syntactic context, and that the actual repeated word or phrase would not vary. This allowed for some diversity in the content of the stimuli while constraining the immediate context of the disfluency. Secondly, we did not want participants to base their judgment on whether the sentences seemed accurate or plausible, so we chose factual statements about real-world phenomena.

Finally, we wanted to evoke a public speaking context where speaker competence would be of utility to the listener without actually testing comprehension by asking participants to place themselves in the shoes of someone trying to learn new information. Hence, we chose technical terms from domains such as physics, astronomy and medicine which would not be common knowledge to most people. These definitions were adapted from Wikipedia entries [25] to fit the stimulus format. We created four variations of 15 different sentences (60 items in total) by adding a repetition or false start: a function word repetition (*fw-rep*), verb

Condition	Example
fw-rep	"The proportionality factor that relates temperature to kinetic energy in a gas is called **the um the** Boltzmann constant."
vp-rep	"The proportionality factor that relates temperature to kinetic energy in a gas **is called um is called** the Boltzmann constant."
n-rep	"The proportionality factor that relates temperature to kinetic energy in a gas is called the **Boltzmann um Boltzmann** constant."
false-start	"The proportionality factor that relates temperature to kinetic energy in a gas **is known um is called** the Boltzmann constant."

Table 1: Examples of the four utterance types synthesized for the perception experiment (fw-rep = function word repetition, vp-rep = verb phrase repetition, n-rep = name repetition)

phrase repetition (*vp-rep*), name repetition (*n-rep*) and false start (*false-start*). A filled pause (*um*) was placed between the repeated word or phrase, or in the case of false starts, as an editing term between the reparandum and alteration. Examples of each utterance type are shown in Table 1. We initially considered including a fluent version of each sentence but ultimately used only non-fluent stimuli due to concerns that participants would attend only to the mere presence or absence of disfluencies. When synthesizing the stimuli we used prosody modification to ensure that all utterances had a falling intonation contour.

2.2. Perception experiment

A web-based subjective listening task was carried out to assess the impact of different kinds of repetitions on perceptions of a speaker's competence. We used a MUSHRA-like [26] design where participants viewed all four versions of each utterance side-by-side (in randomized order) and rated them each on a 5-point scale ranging from 1 ("not at all competent") to 5 ("extremely competent"). Participants were able to play each stimulus as many times as needed. To insure that participants were attending to the stimuli we implemented attention checks. Two extra sets of sentences were synthesized, with one audio clip in each set instructing participants to to give the

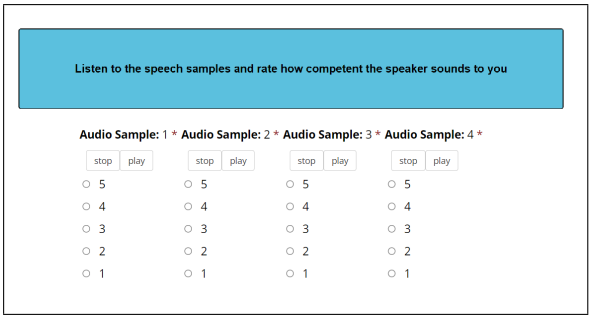


Figure 1: Interface of the perception experiment

stimulus a rating of "1". Responses to these stimuli were not included in the results.

Participants were native speakers of English recruited using the crowdsourcing platform Prolific. They were pre-screened to check that they met the inclusion criteria and affirmed that they were wearing headphones for the duration of the experiment and had no hearing impairments. We recruited new participants to replace any who failed one or both attention checks, encountered technical issues, or were unable to complete the experiment for other reasons, until we had collected data from a total of 40 participants. Of these, 43.6% were female and 56.4% were male.

3. RESULTS

A Friedman test with repeated measures on participants was carried out to determine whether responses differed between conditions. This test is a non-parametric equivalent to repeated-measures analysis of variance, suitable for non-normal ordinal-scale data. The test was significant, χ^2=14.03, $p < 0.005$, indicating a difference in participants' ranking of stimuli on the 5-point competence scale across conditions. Figure 2 shows the distribution of responses for each condition. Post-hoc Conover tests on mean ranks with a Holm-Bonferroni correction for multiple comparisons showed that utterances with false starts were rated as sounding significantly less competent than both utterances with function word repetitions (t=2.63, $p < 0.05$) and utterances with verb phrase repetitions

Condition	Mean	SD
fw-rep	3.11	0.77
vp-rep	3.17	0.80
n-rep	3.03	0.70
false-start	2.80	0.83

Table 2: Mean and standard deviation of competence ratings by condition

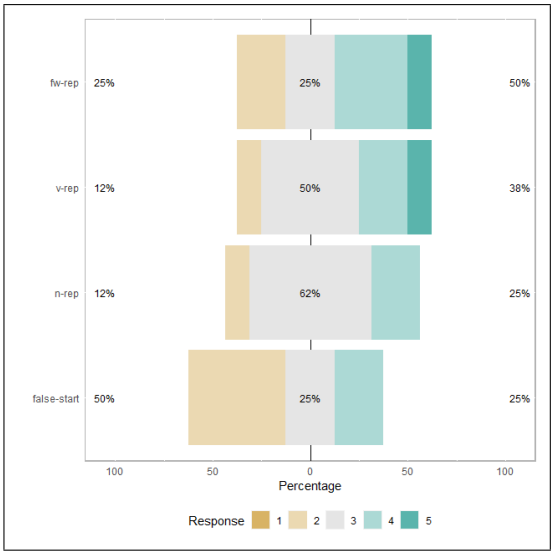


Figure 2: Distribution of responses by condition

(t=3.55, $p < 0.005$). No other pairwise comparisons were significant.

4. DISCUSSION

The results of the perception experiment partially confirmed our predictions. Utterances containing false starts were rated as sounding less competent than utterances with function word or verb phrase repetitions. However, there was no significant difference between false starts and name repetitions, or between any of the other repetition types. The relative predictability of where a disfluency might occur in an utterance may have attenuated some of their impact. The stimulus order was randomized in each trial, meaning participants would not know which type of repetition they were about to hear the first time they listened to a given audio file, but the amount of variation was nonetheless constrained by the stimulus design. If the effect of disfluencies is at least partly a function of how much they disrupt processing it makes sense that predictability would blunt this effect to some extent. Nonetheless, the disfluency that appears most disruptive to processing was associated with lower perceived competence. These findings are hence consistent with, and build upon, previous work showing a link between fluency and perceived competence [3, 4, 5].

It is also important to note that in addition to potentially indicating a speaker's competence, repetitions play a whole host of other roles in conversation. Bazzanella [27] provides a list of dozens of such roles, which include cognitive and meta-cognitive functions such as planning an utterance or marking or facilitating comprehension; interactional functions like marking surprise, agreement or disagreement; turn-taking behaviors such as holding or yielding a turn; and stylistic functions such as emphasis. The context, the speaker, and characteristics of the listener may all shape how disfluencies are interpreted and in turn how they influence perceptions of competence.

Spontaneous neural text-to-speech (TTS) has the potential to be a valuable tool for investigating various factors that impact speech perception. One advantage of this method is its ability to generate a wide range of stimuli, allowing researchers to systematically vary certain features while others are constrained. For example, sentences with the desired semantic content or syntactic structure could be synthesized with different speaker characteristics (such as age or gender, both of which have been shown to influence speaker competence), voice quality (e.g., creak/vocal fry), or modifications to pitch, intensity or speech rate. Since the use of spontaneous TTS as a tool for studying speech perception is relatively new, the methodology will need continued development as the capabilities of TTS evolve. More studies are needed to better understand the types of research questions this method is suited to investigate and to chart out pitfalls and best practices.

5. CONCLUSIONS

This study investigated the effects of false starts and different types of repetitions on a speaker's perceived competence, building upon previous research utilizing speech synthesis as a tool for studying speech perception. The results demonstrated that false starts had a negative impact on perceived competence, consistent with the idea that they may disrupt processing fluency for the listener. There were no significant differences observed between the various types of repetitions. These findings have practical implications for public speaker training, suggesting that minimizing false starts may improve perceived competence. While it has been generally assumed that all disfluencies negatively impact perceived competence, our results indicate that certain types of disfluencies may be more detrimental than others. The study demonstrates an example of how spontaneous neural TTS can be used to generate a range of stimuli that are systematically varied to investigate the various factors that influence speech perception.

6. ACKNOWLEDGEMENTS

This research is supported by the Riksbankens Jubileumsfond project CAPTivating – Comparative

Analysis of Public Speaking with Text-to-Speech (P20-0298).

7. REFERENCES

[1] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, no. 3, pp. 201–242, 1998.

[2] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[3] G. R. Miller and M. A. Hewgill, "The effect of variations in nonfluency on audience ratings of source credibility," *Quarterly Journal of Speech*, vol. 50, no. 1, pp. 36–44, 1964.

[4] J. K. Barge, D. W. Schlueter, and A. Pritchard, "The effects of nonverbal communication and gender on impression formation in opening statements," *Southern Communication Journal*, vol. 54, no. 4, pp. 330–349, 1989.

[5] D. M. Oppenheimer, "Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 20, no. 2, pp. 139–156, 2006.

[6] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California at Berkeley, Berkeley, CA, 1994.

[7] J. E. F. Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of memory and language*, vol. 34, no. 6, pp. 709–738, 1995.

[8] P. Howell and S. Sackin, "Function word repetitions emerge when speakers are operantly conditioned to reduce frequency of silent pauses," *Journal of psycholinguistic research*, vol. 30, no. 5, pp. 457–474, 2001.

[9] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," *Speech communication*, vol. 10, no. 2, pp. 163–169, 1991.

[10] C. Aruffo, "Reading scripted dialogue: Pretending to take turns," *Discourse Processes*, vol. 57, no. 3, pp. 242–258, 2020.

[11] G. P. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, no. 1, pp. 43–65, 1997.

[12] P. Wagner and A. Windmann, "Re-enacted and spontaneous conversational prosodyâhow different?" *Proceedings of Speech Prosody 2016*, pp. 518–522, 2016.

[13] M. Drolet, R. I. Schubotz, and J. Fischer, "Authenticity affects the recognition of emotions in speech: behavioral and fmri evidence," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 12, no. 1, pp. 140–150, 2012.

[14] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 346–352.

[15] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[16] S. Wang, J. Gustafson, and É. Székely, "Evaluating sampling-based filler insertion with spontaneous tts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1960–1969.

[17] E. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *INTERSPEECH*, 2017, pp. 804–808.

[18] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[19] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6925–6929.

[20] G. LLC, "Google cloud speech api video model," https://cloud.google.com/speech-to-text.

[21] R. Ochshorn and M. Hawkins, "Gentle forced aligner[computer program]," 2017.

[22] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[23] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] Wikimedia Foundation, "Wikipedia, the free encyclopedia," http://en.wikipedia.org/, 2023, [Online; accessed 06-January-2023].

[26] International Telecommunication Union, Radiocommunication Sector, "Method for the subjective assessment of intermediate quality levels of coding systems," ITU Recommendation ITU-R BS.1534-3, 2015.

[27] C. Bazzanella, "Redundancy, repetition, and intensity in discourse," *Language sciences*, vol. 33, no. 2, pp. 243–254, 2011.

LISTENING WITH STYLE: The effect of bidialectal style-shifting on PRICE vowel perception in Britain's Black Country

Thomas Packer-Stucki

Queen Mary University of London
t.j.packer-stucki@qmul.ac.uk

ABSTRACT

Speech perception research has found that manipulating listener attitudes towards a dialect area can lead to convergence or divergence in production style, as well as a simultaneous convergence or divergence in perception. The present study investigates whether these two effects are connected, such that the changes in production style are the cause of changes in perception. The potential link is interrogated directly through an experiment which pushes participants' stylistic production towards the extreme ends of their repertoires, followed by a perceptual test in which they match vowel tokens to a vowel line-up resynthesized along the same standard-dialect continuum. Analysis of the data shows that style does affect perception: as participants are asked to switch from a more standard to a more dialectal style, their perceptual categories shift towards the standard end of the vowel continuum, i.e., when they speak in dialect, vowels are more likely to be rated as standard-sounding, and vice-versa.

Keywords: Vowel perception, experimental sociolinguistics, bidialectalism, style-shifting

1. INTRODUCTION

1.1. Priming and speech perception

Research in experimental sociolinguistics has shown that a listener's perception of an incoming speech signal can be influenced by external factors, for example, acoustically identical sounds can be perceived differently when listeners are primed with differing biographical information about the speaker [12]. Many of these studies make use of vowel rating tasks, wherein participants hear a target vowel, and must identify the most similar sounding vowel from a synthesised line-up. This methodology was originated by Niedzielski [16], whose study had Detroit participants rate instances of the MOUTH vowel along a unraised-raised continuum, some being told they were hearing a Detroit speaker and others that they were hearing a Canadian. As this vowel is raised in both dialect areas, but stereotyped only in Canada, participants only reported hearing the raising where they expected it, in the Canada condition. While the vowel stimuli were identical across both conditions, the participants' stereotypes of the perceived speaker's dialect biased their perception in the expected direction.

1.2. Attitudinal priming

More recently, Walker et al. [21] uncovered an effect whereby the *attitudes* of New Zealand listeners towards Australia seem to bias their perception of the KIT vowel along a NZE-AuE continuum of possible realisations. Their experiment involved a listening task in which participants heard a series of target KIT vowels embedded within sentences and were asked after each to identify the most similar-sounding vowel from a resynthesized line-up, ranging from an AuE to NZE realisation. The listening task was completed twice, before and after treatment in one of three conditions, which involved the reading of positive or negative facts about Australia, or a control with random facts. Analysis of results revealed a shift in perception to more AuE-like KIT in response to the *negative* facts, while the *positive* facts had the opposite effect. Specifically, Walker et al. interpreted that participants had a defensive reaction to conditioning with positive facts about Australia; Kiwi pride leads to more Kiwi-sounding vowel perceptions. However, it is unclear by which mechanism such attitude-based shifts occur.

1.3. Listening with style

While the motivation for a link between attitude and perception has so far been difficult to identify [21], the link between attitudes and style is well theorised and attested. Research contributing to Communication Accommodation Theory (CAT) has shown that the propensity to converge towards or diverge away from a particular style or variety can be motivated by attitudes towards the speakers of that code [6], [11]. The present paper hypothesises that in the case of Walker et al. it is an attitude-motivated change in style which is responsible for the change in perception. We directly test the possibility of style-motivated shifts in perception by means of an experiment which manipulates production style rather than attitude and observes changes in vowel productions and perceptions. We did this by

following a similar experimental paradigm, but instead of manipulating attitude between each of the listening task iterations, we expressly manipulate bidialectal participants' production style by means of a reading task, pushing them from the most standard to the most dialectal end of their stylistic repertoire. What follows is a presentation of the perceptual side of this larger study involving both production and perception elements.

2. METHODOLOGY

2.1. Participants

While it is attested that all speakers exhibit *style-shifting*, we need to be certain that our participants will shift decisively between two distinct styles, if we are to argue that this shift is responsible for any perceptual effects observed. We should also be certain that such a decisive shift is naturalistic; something which our participants would usually do. To this end, we identified *bidialectal* speakers [20] as the ideal candidates for our experiment – according to Hazen [14] their speech can occupy the, "extreme ends of the style-switching continuum."

In the United Kingdom, one area which retains a unique, productive dialectal vernacular is the *Black Country*, whose dialect forms an important part of local identity, and is in common use today by the majority of the region's 1.2 million inhabitants [3], [17]. Recruitment of participants took place from the pool of presenters at local community radio station *Black Country Radio*. These speakers were chosen as they regularly access both standard and dialect codes. A total of eight participants were recruited for participation – their ages ranged from 27 to 69 (avg. 54) with five men and three women. Participants were not compensated for their time.

2.2. Procedure

The experiment took place in a recording studio at the radio station and lasted an average of 30 minutes. In the reading tasks, participants were asked to read text in either a standard or dialectal style. In the listening tasks, participants saw a series of target sentences with a PRICE-containing word highlighted (see Figure 1); after each they had to select the most similar sounding vowel from a continuum (see Figure 2). The reading and listening tasks were combined to create the following experimental procedure:

Reading Task 1: news bulletins (standard)
Listening Task 1: 4 sentences
Reading Task 2: news bulletins (standard)
Listening Task 2: 3 sentences
Reading Task 3: Poem 1 (dialect)
Listening Task 3: 4 sentences
Reading Task 4: Poem 2 (dialect)
Listening Task 4: 3 sentences

In our analysis, the main effect will therefore be the location of a particular vowel trial within the experiment, following either a standard or dialect reading task. While the reading tasks always appeared in the same order, the order of the 14 listening task sentences was randomised for each participant to mitigate any potential order effect of the stimuli.

2.3. Perception: Listening task

Following Niedzielski [16] and Walker et al. [21], a vowel perception task was developed for use in the Black Country. The PRICE vowel was selected, as differences between BCE and SSBE are notable, with [ɒɪ] given as the typical Black Country realisation and [ʌɪ] appearing in the south [7].

2.3.1. Stimuli

14 sentences containing PRICE were selected from the Bamford-Kowal-Bench list of sentences controlled for lexical frequency and average age of acquisition [4], and were recorded in a broadcast-standard recording booth by a 25-year-old female speaker from Stourbridge, in the south-west of the Black Country. When the sentences were presented to participants, they were shown orthographically on a screen with the target vowel-containing word underlined and played twice through headphones, as shown in Figure 1. Participants then heard the SSBE-BCE PRICE vowel continuum (see section 2.3.2.) and were asked to select the most similar-sounding vowel to the target. Additionally, each stimuli sentence was presented only *once* in the experiment to decrease the likelihood of participants guessing the aim – average vowel ratings between reading task styles will therefore be compared.

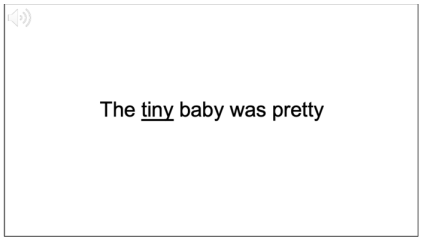


Figure 1: An example stimulus sentence as seen by participants, where the target vowel for matching with the continuum is underlined

2.3.2. Vowel continuum

In order to create the 6-step vowel continuum for PRICE, a bidialectal speaker from the Black Country with phonetics training 'performed' both the SSBE and BCE endpoints for the vowel. For the resynthesis itself, Praat software [5] was used along with a script [15] which takes two natural vowel endpoints, estimates their formants using linear prediction, and then uses inverse filtering to synthesise 6 vowels equidistant from each other in terms of formant values from each endpoint. In the resulting continuum, '1' represented the most SSBE-type vowel, and '6' the most BCE-type vowel. Following presentation of the target vowel, participants saw six boxes on a screen which represented the six vowel choices available (in order). They could then use a mouse to hover over each of the boxes, which would then cause that particular vowel to be played to them. They were asked to select the vowel which sounded most similar to the target, as shown in Figure 2.

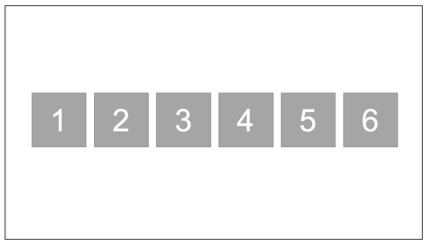


Figure 2: The vowel continuum as seen by participants, follow the presentation of each target vowel

2.4. Production: Reading task

The production element of the study was designed to manipulate the participants' style and aimed to push the participants to the most extreme ends of their stylistic repertoires. The instructions were explicit in doing so: participants were asked to speak firstly in a formal style i.e., for a 'national audience', and then secondly in a dialectal style i.e., 'broad Black Country'. This was done to ensure the biggest shift in style possible, given that in previous research, perceptual shifts tend to be slight even where they are significant [9].

2.4.1. SSBE style script (standard)

Given that the participants are radio presenters, the task used to elicit the participants' most standard style was the reading of two 'fake' news bulletins, adapted from BBC News stories selected for their financial or academic focus. Participants were asked to read these stories with a national audience in mind.

2.4.2. BCE style script (regional dialect)

To elicit the most dialectal style possible, participants were tasked with reading two poems written in Black Country dialect, chosen for their setting in the region and their frequent use of dialect grammar and vocabulary.

Both the standard and dialect tasks, having two parts each, were split in two and interspersed within the listening task sentences to increase the likelihood that participants remained under the influence of each style as they completed the perceptual element. To mitigate potential biases, participants were told that the reading and listening tasks were separate experiments mixed together to prevent the procedure from becoming tiresome.

3. RESULTS

A total of 112 vowel rating tokens were analysed across the preceding production condition (dialect or standard). As shown in Figure 3, on average, participants selected vowels from the opposite end of the continuum compared to the preceding production condition. In essence, when participants produced more dialect, they perceived the same vowels as sounding more standard.

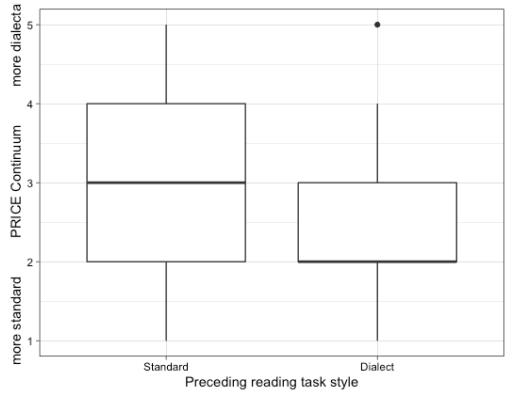


Figure 3: Average PRICE vowel ratings split between preceding reading task style (standard and dialect)

Following previous vowel perception research [13], [12], a gender effect also seems to be present – while the shift in rating between conditions is similar for men and women, the men's ratings seem to occupy a lower part of the scale than those of the women (Figure 4), such that the male participants rated the vowels as more standard-sounding overall.

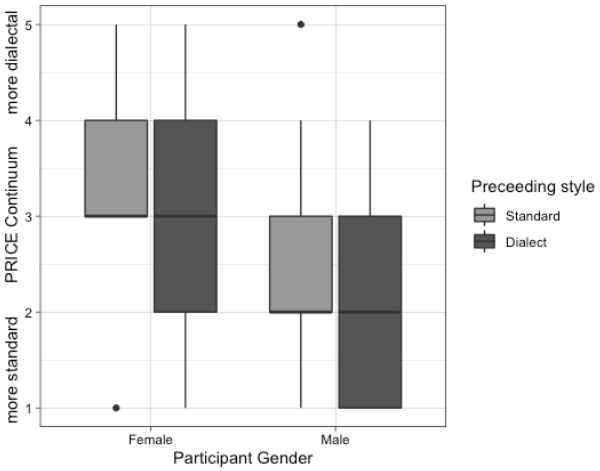


Figure 4: Average PRICE vowel ratings across the two conditions by participant gender

To check the significance of the effect observed, and to control for the potential gender effects, a linear mixed effects regression model was built in R [13] using the 'lme4' package [2], [19], with the vowel ratings as the dependent variable, and the condition and gender as fixed effects, with random intercepts and slopes fitted across the condition for participants.[i] The output model, shown in Table 1, indicates that the *reading task style* is a significant predictor of *vowel rating* response. When participants shift from standard to dialectal production, the coefficient for vowel rating shows a negative difference of -0.37 (Cohen's d = -0.42[ii]), meaning they rate the stimuli vowels as sounding more standard than before.

Fixed effects	Estimate	Std. error	*t*	*p*-value
(Intercept)	3.46	0.27	12.62	<0.0001***
Preceding reading task: dialect	-0.37	0.17	-2.18	0.0317*
Gender: male	-1.02	0.32	-3.19	0.0129*

Number of observations: 112; Groups: Participant (8, SD=0.01)

Table 1: lmer analysis of experimental results

In sum:

- Participants report hearing more dialectal realisations of PRICE when they are asked to produce more standard language and vice versa.
- Male participants report hearing more standard realisations overall, while female participants hear more dialectal realisations, though the direction and extent of the shift is the same.

4. DISCUSSION

The present paper makes a contribution to an emerging field of research at the intersection of language attitudes, speech production and speech perception. We provide evidence that style can be a significant mediator of speech perception, and by extension the interpretation that style may be the source of seemingly attitude-based effects. Additionally, we have presented evidence for such a production-perception relationship alluded to by Walker et al. [21], that as participants are encouraged to shift their production style in one direction, their vowel perceptions shift in the *other*.

While Walker et al. found that manipulation of attitudes produced a change in participants' vowel perceptions, the motivation for such an effect is somewhat unclear. When interpreting their results, the authors refer to previously published production data from the same paradigm [10] and identify that the attitude treatment had a stylistic effect on the participants' own KIT productions. Namely, for the participants who were sports fans, the *positive* facts about Australia led them to produce more *Kiwi* vowels. These findings were interpreted using an *exemplar theory* framework, in which linguistic variables are represented as individual mental entries which must reach a certain level of activation to outcompete others – recent activation can therefore introduce bias as these exemplars have a head start against others [18]. In this way, it could be argued that greater 'activation' of more AuE-type vowel exemplars in production means those same exemplars are more easily activated in perception. However, given that at least *some* of the participants style-shifted to more AuE productions in response to the positive facts, there would remain to be identified a mechanism by which shifts in production can occur in one direction, while shifts in perception occur in the other.

While our data support the intuition that style should be considered a candidate for mediator between attitude and perception, they do not lend themselves to an exemplar-based explanation. While an exemplar model can elegantly explain effects introduced by priming with biographical information such as age or dialect area, it has difficulty in accounting for the findings presented here, which suggest that perception can be biased *away* from rather than *towards* the more recently accessed vowel realisations.

6. CONCLUSIONS

Making use of speakers with especially board stylistic repertoires, the present study contributes evidence that production style seems to influence vowel perception. In this way style, being sensitive to attitudes, may be the 'missing link' where affect seems to influence speech perception. However, it still remains to identify the mechanism by which style can mediate perception.

7. REFERENCES

[1] T. Baguley, 'Standardized or simple effect size: What should be reported?', *British journal of psychology (London, England : 1953)*, vol. 100, pp. 603–17, Dec. 2008.

[2] D. Bates, M. Maechler, B. Bolker, and S. Walker, 'lme4: Linear Mixed-Effects Models using "Eigen" and S4'. 2022.

[3] BBC Archive, 'Black Country dialect', *BBC Local*, 24-Sep-2014. [Online]. Available: https://www.bbc.co.uk/blackcountry/features/2002/12/accents/black_country_dialect.shtml. [Accessed: 11-Aug-2022].

[4] J. Bench, A. Kowal, and J. Bamford, 'The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children', *Br J Audiol*, vol. 13, no. 3, pp. 108–112, Aug. 1979.

[5] P. Boersma and D. Weenink, 'Praat: doing Phonetics by Computer'. 2022.

[6] B. Chakrani, 'Arabic interdialectal encounters: Investigating the influence of attitudes on language accommodation', *Language & Communication*, vol. 41, pp. 17–27, Mar. 2015.

[7] U. Clark and E. Asprey, *West Midlands English: Birmingham and the Black Country*, Annotated edition. Edinburgh: Edinburgh University Press, 2013.

[8] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Routledge, 1988.

[9] K. Drager, 'Speaker Age and Vowel Perception', *Lang Speech*, vol. 54, no. 1, pp. 99–121, Mar. 2011.

[10] K. Drager, J. Hay, and A. Walker, 'Pronounced Rivalries: Attitudes And Speech Production', *Te Reo*, vol. 53, pp. 27–53, Jan. 2010.

[11] P. Garrett, 'Communication accommodation theory', in *Attitudes to Language*, 1st ed., Cambridge University Press, 2010, pp. 105–120.

[12] J. Hay, A. Nolan, and K. Drager, 'From fush to feesh: Exemplar priming in speech perception', vol. 23, no. 3, pp. 351–379, Oct. 2006.

[13] J. Hay, P. Warren, and K. Drager, 'Factors influencing speech perception in the context of a merger-in-progress', *J. Phonetics*, 2006.

[14] K. Hazen, 'An Introductory Investigation Into Bidialectalism', *University of Pennsylvania Working Papers in Linguistics*, vol. 7, no. 3, Jan. 2001.

[15] D. Lawrence, 'A Brief Introduction to Vowel Resynthesis for Speech Perception Research'. University of Edinburgh, 2015.

[16] N. Niedzielski, 'The Effect of Social Information on the Perception of Sociolinguistic Variables', *Journal of Language and Social Psychology*, vol. 18, no. 1, pp. 62–85, Mar. 1999.

[17] Office for National Statistics, 'Population and household estimates, England and Wales', *Census 2021*, 28-Jun-2022. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/latest. [Accessed: 11-Aug-2022].

[18] J. B. Pierrehumbert, 'Exemplar dynamics: Word frequency, lenition and contrast', in *Typological Studies in Language*, vol. 45, J. L. Bybee and P. J. Hopper, Eds. Amsterdam: John Benjamins Publishing Company, 2001, p. 137.

[19] R Core Team, 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[20] J. Smith and M. Durham, 'Bidialectalism or Dialect Death? Explaining Generational Change in the Shetland Islands, Scotland', *American Speech*, vol. 87, no. 1, pp. 57–88, Mar. 2012.

[21] A. Walker, J. Hay, K. Drager, and K. Sanchez, 'Divergence in speech perception', *Linguistics*, vol. 56, no. 1, pp. 257–278, Jan. 2018.

[i] Call formula: lmer(VowelRating ~ Condition + Gender + (1+Condition|Participant), PRICE_data, REML=FALSE)

[ii] It is important to acknowledge the limitations [1] of relying on effect size statistics such as Cohen's [8] *d,* and for this reason interpretation of the effect is also based on the coefficients.

REORGANIZATION OF THE AUDITORY-PERCEPTUAL SPACE ACROSS THE HUMAN VOCAL RANGE

Daniel Friedrichs, Volker Dellwo

Department of Computational Linguistics, University of Zurich
daniel.friedrichs@uzh.ch, volker.dellwo@uzh.ch

ABSTRACT

We investigated the auditory-perceptual space across a substantial portion of the human vocal range (220-1046 Hz) using multidimensional scaling analysis of cochlea-scaled spectra from 250-ms vowel segments. Three native German speakers produced vowels /i y e ø ɛ a o u/ (N=240) spanning a wide range of their respective voice frequency ranges, which were subsequently identified in a closed-set task by five phonetically trained listeners (>80% correct response). Our findings revealed systematic spectral alterations associated with vowel height and frontness as fundamental frequency increased and a distinct clustering around /i a u/ above 523 Hz. These results emphasize the importance of the overall spectral shape in vowel perception and demonstrate listeners' utilization of acoustic anchors at higher pitches. Moreover, this study sheds light on the quantal nature of these vowels and their potential impact on language evolution, offering a possible explanation for their widespread presence in the world's languages.

Keywords: auditory-perceptual space, fundamental frequency, pitch, vowels, quantal theory

1. INTRODUCTION

For much of the last century, it has been widely believed that vowels produced at high fundamental frequencies (f_o) are challenging to comprehend, supported by numerous studies in the field of singing research (e.g., [1,2]). In 2013, Sundberg [3] provided a comprehensive overview, summarizing the results from various studies in the field of Western classical singing, and concluded that vowel recognition declines at an f_o of around 523 Hz (i.e., the musical note C5). Above this frequency, only the open vowels /a/ and /ɑ/ (i.e., the vowels with the highest F_1) remain identifiable, while identification rates for all other vowels drop towards chance level at higher f_o. This outcome is expected due to the sparse distribution of harmonics at high f_o, leading to an undersampling of the vocal tract transfer function and, consequently, a less precise specification of the formant frequency distribution, a crucial acoustic representation in the vowel identification process.

Contrary to these findings, a few early studies outside Western classical singing showed accurate vowel identification at high f_o but did not consider secondary cues' impact [4,5]. Recent research has addressed these secondary cues at high f_o, such as vowel duration, formant transitions, and coarticulation (see [6,7,8] for more information on secondary cues). For example, Friedrichs et al. [9] found that the phonological function of the steady-state vowels /i y e ø ɛ a u o/ could be maintained at f_o up to 880 Hz in a two-alternative forced-choice task. By increasing the number of response options and talker variability, another study [10] demonstrated that the point vowels /i a u/ could even remain identifiable up to 1046 Hz. These findings align with those reported by Zhang et al. [11], who discovered that isolated vowels /i a u/ produced by a female Chinese Yue Opera singer could be identified with high accuracy up to around 932 Hz when presented to phonetically trained listeners in a free-choice identification task.

Considering the conflicting observations in singing research, it is possible that vowel recognition at high f_o is compromised by articulatory and acoustic adjustments made by classical Western Opera singers. Joliveau et al. [12] demonstrated, by measuring broadband acoustic excitation at the mouth of soprano singers, that they shift their first resonance frequency (f_{R1}) towards f_o to increase vocal power in their performances. Such modifications were typically observed when f_o was high and approached f_{R1}. Altering the vocal tract's first resonance frequency may, therefore, significantly impact the perceived vowel category.

To date, little is known about the mechanisms that allowed vowel identification at very high f_o in the aforementioned studies outside Western classical singing. These mechanisms likely involve specific spectral properties and auditory-perceptual strategies employed by listeners [13]. In this study, we investigate the spectral properties of vowels within a range of 220 to 1046 Hz to explore the underlying auditory-perceptual mechanisms that drive the vowel identification process throughout the human vocal range. By examining these auditory-perceptual mechanisms and spectral properties, we aim to provide a better understanding of how vowel identification can be maintained at high f_o and the factors that contribute to the discrepancies observed

between Western classical singing and other singing styles. This knowledge can provide insights for researchers in singing, phonetics, and linguistics and contribute to the improvement of acoustic models for vowel recognition at high f_0.

2. METHOS

2.1. Participants

Five phonetically trained listeners (3 females, 2 males; mean age = 37.1 years, standard deviation = 8.4 years) participated in the listening experiment. All participants had audiometric thresholds of less than 20 dB hearing level at octave frequencies between 125 and 8000 Hz.

2.2 Stimuli

Three female native German speakers (mean age = 32.2, standard deviation: 2.5) with professional vocal training in singing or acting were selected from a corpus of 70 speakers [14] based on their extended vocal range and similar f_0 in producing citation-form words and reading a story (mean 215.4 Hz, standard deviation 12.4 Hz). Recordings were made using a cardioid condenser microphone (Sennheiser MKH 40 P48) at a sampling frequency of 44.1 kHz and a constant distance of 30 cm from the speakers. The speakers produced the eight vowels /i y e ø ɛ a o u/ in isolation at ten target f_0 ranging from 220 to 1046 Hz (i.e., 220, 330, 440, 523, 587, 698, 784, 880, 988, and 1046 Hz). The target f_0 were played to the speakers as reference tones via studio monitor loudspeakers before each recording. f_0 was measured with an autocorrelation method [15,16] in Praat [17] and manually checked. For each vowel, multiple recordings were made, and those with the closest distance to the target f_0 (maximum deviation of 5%) were selected, resulting in a total of 240 vowels (10 f_0 x 8 vowels x 3 speakers). 250-ms segments were extracted from the vowel centers, normalized to an arbitrary intensity, and onset, and offsets were faded using raised cosines.

2.3. Procedure

The phonetically trained listeners participated in a multiple-choice transcription task in which they were presented with the stimuli vowels via closed headphones (Beyerdynamic DT 770 Pro, 250 Ω) and asked to assign each vowel to one of the eight vowel categories displayed on a computer screen. Listeners could also assign the presented vowel to a category between two vowels (e.g., /i¬–e/ instead of /i/ or /e/). The output level could be chosen individually for comfort, and listeners were allowed to replay each presented vowel multiple times. Only vowels with more than 80% correct responses were selected for subsequent analyses.

2.4. Cochlea-scaled spectra

Cochlea excitation patterns were simulated using a 200-channel linear gammatone filter bank, with bandwidths and center frequencies calculated using the ERB formulae by Glasberg and Moore [18]. The RMS level of the output wave was calculated and converted to dB for each filter channel. Frequency weighting was applied based on measurements made by Puria et al. [19] to account for the transmission properties of the middle ear.

2.5. Data analysis

Classical multidimensional scaling (MDS) analysis [20, 21] was performed on the cochlea-scaled spectra to model changes in the estimated auditory-perceptual space throughout the f_0 range. MDS has been previously used to assess perceptual proximity or dissimilarity in vowels [22, 23]. At each target f_0, all eight vowels were assigned to specific coordinates in a two-dimensional space derived from the calculated distances of the cochlea-scaled spectra. The distances shown in the MDS space are linearly related to the spectral distance.

Considering a vowel clustering phenomenon at high f_0 reported in previous studies [10,11], we aimed to investigate potential clustering of vowels at different f_0 by calculating the Euclidean distances between vowel pairs that appeared to cluster together in the MDS space (cluster 1: /i e y/, cluster 2: /ø ɛ a/, cluster 3: /u o/). For each speaker and f_0, we calculated the Euclidean distance between normalized cochlea-scaled spectra of all vowel pairs within each cluster using the formula:

$$(1) \qquad d(p,q) = \sqrt{\Sigma(q_i - p_i)^2}$$

where $d(p,q)$ is the Euclidean distance between two spectra, and p_i and q_i are the amplitude values at frequency bin i in spectra p and q, respectively. The summation is taken over all frequency bins i from the first to the last frequency bin in the spectra.

To assess the relationship between the f_0 and the sum of Euclidean distances, we performed a piecewise mixed-effects linear regression with the Euclidean distances as the dependent variable and f_0 as a fixed effect. Based on Sundberg (2013), which suggests that listeners' vowel identification performance decreases at around 523 Hz, the model allowed for different slopes in the relationship between f_0 and the Euclidean distances, starting at 523 Hz. The random

effects in this model included the speaker and listener. This analysis helped us determine if the changes in the Euclidean distances were significantly different across the f_o range, providing insights into the clustering phenomenon and the underlying auditory-perceptual mechanisms involved throughout the human vocal range.

3. RESULTS

Fig. 1 displays the estimated auditory-perceptual space, averaged across the three speakers at all recorded frequencies between 220 and 1046 Hz. At the lowest f_o (220 Hz), the MDS space derived from cochlea-scaled spectra resembles a typical F_1–F_2 space. As the f_o increases, the vowel height dimension (vertical axis) partially collapses while the frontness dimension (horizontal axis) expands, altering the ratio of these dimensions from 1:1.44 (220 Hz) to 1:0.47 (1046 Hz). Despite this reorganization, the point vowels /i a u/ remain at the corners of the vowel space.

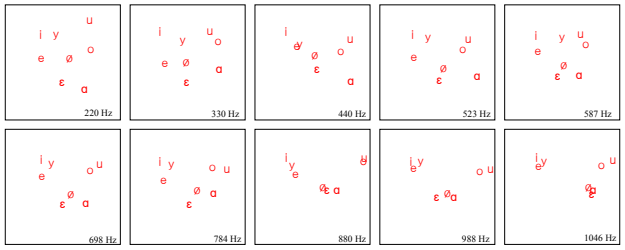


Figure 1: MDS plots showing the estimated auditory-perceptual space derived from cochlea-scaled spectra of the vowels used in this study throughout the f_o range from 220–1046 Hz (averaged across speakers). From around 523 Hz, the vowels start to cluster around the point vowels /i a u/.

Starting around 523 Hz, shifts towards the categories /i a u/ are observed: /y e/ move towards /i/, /o/ towards /u/, and /ø ɛ/ towards /a/. Fig. 2 illustrates the average decrease in Euclidean distance measurements for vowels in these three clusters as f_o increases. Using the False Discovery Rate (FDR) controlling procedure [23], analyses of differences reveal no significant change in Euclidean distance measurements between the lowest f_o (220 Hz) and 330 Hz (adjusted $p = 0.43$) as well as 440 Hz (adjusted $p = 0.21$), but significant differences occur between 220 Hz and 523 Hz (adjusted $p < .05$), as well as between 220 Hz and all f_o up to 784 Hz (all adjusted $p < .05$) and up 1046 Hz (all adjusted $p < .001$).

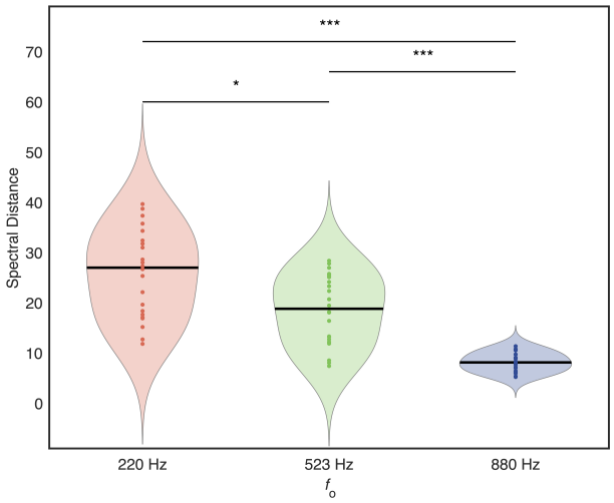


Figure 2: Violin plots showing the distribution of Euclidean distance measurements within the three clusters around the point vowels /i a u/ for a low, medium, and high f_o. The Euclidean distance between cochlea-scaled spectra decreases significantly towards the highest f_o.

Fig. 3 provides examples of cochlea-scaled spectra of /i y e/ forming a vowel cluster at high f_o. Dissimilarity is evident at the lowest f_o (220 Hz), but spectral proximity becomes more apparent at higher f_o (523 and 880 Hz). At 880 Hz, the Euclidean distance between the spectral vectors approaches the lower end of the scale and is only observable in high-frequency bands above 7 or 8 kHz.

Euclidean distances between the cochlea-scaled spectra of vowel pairs reveal a clear pattern across the investigated f_o. A piecewise linear mixed-effects regression with Euclidean distance as the dependent variable, f_o as a fixed effect (split at 523 Hz), and speaker and listener as random effects showed no significant relationship between f_o and Euclidean distance below 523 Hz ($\beta 1 = -0.005$, $p = 0.65$). At these lower f_o, the perceptual distance between vowels remains relatively constant. However, for f_o above 523 Hz, a significant negative relationship emerges between f_o and Euclidean distance ($\beta 2 = -0.045$, $p < .001$), suggesting that vowel clustering becomes more pronounced at higher f_o, with vowels becoming increasingly perceptually similar.

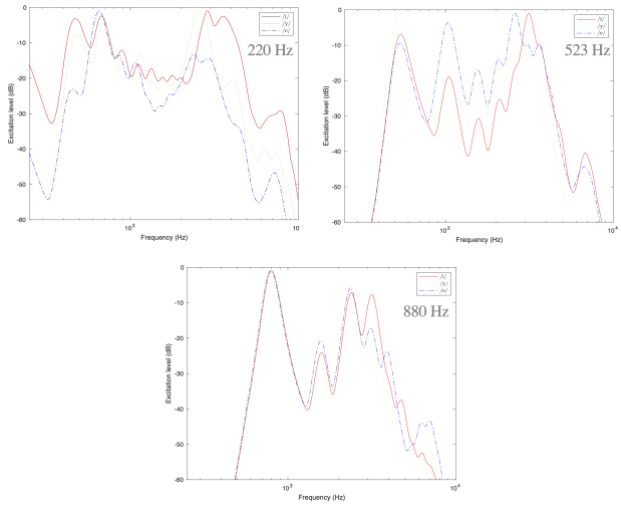


Figure 3: Cochlea-scaled spectra of the vowels /i y e/ produced by one speaker at fundamental frequencies of about 220, 523, and 880 Hz (top to bottom).

At the lowest f_o of 220 Hz, the median Euclidean distance was about 28 units, indicating a relatively large perceptual distance between vowels. As f_o increased up to 523 Hz, the median Euclidean distance remained relatively stable. From 523 Hz onwards, the median Euclidean distance gradually decreased, reaching a minimum of about 8 units at 880 Hz.

The two higher f_o, 988 Hz and 1046 Hz exhibited nearly identical median Euclidean distances as observed at 880 Hz ($p > .05$), indicating that the vowel clustering phenomenon persists at these higher f_o. The piecewise linear mixed-effects regression also confirmed the lack of a significant difference in Euclidean distance between these three f_o ($p > .05$). Overall, our results support the existence of a clustering phenomenon in vowel perception at higher f_o.

4. DISCUSSION

The findings presented in this study suggest that the auditory-perceptual space undergoes a systematic reorganization throughout the human vocal range and a clustering phenomenon around the vowels /i a u/ at very high f_o. The patterns found in the MDS analysis likely allow vowels to remain distinguishable even when the vocal tract transfer function is undersampled to a high degree. Furthermore, this might explain why studies on vowel perception outside the field of Western Opera singing reported vowel intelligibility at high f_o.

The findings presented in this study suggest that the auditory-perceptual space undergoes a systematic reorganization throughout the human vocal range and a clustering phenomenon around the vowels /i a u/ at very high f_o. The patterns found in the MDS analysis likely allow vowels to remain distinguishable even when the vocal tract transfer function is undersampled to a high degree. This might explain why studies on vowel perception outside the field of Western Opera singing reported vowel intelligibility at high f_o.

Our results indicate a tripartite organization of the auditory-perceptual space. At lower f_o (220–440 Hz), the arrangement of the vowels resembles that of a vowel quadrilateral or a typical F_1–F_2 space. This suggests that well-established mechanisms, such as processing formant frequency distributions, remain unaffected in this range. An auditory-perceptual reorganization begins at frequencies around 523 Hz, previously identified as the absolute frequency at which vowel intelligibility drops for all except /a/-like vowels [3]. This reorganization process evolves until around 880 Hz. From this f_o on (i.e., at 880, 988, and 1046 Hz), the auditory-perceptual clustering around the vowels /i a u/ is present, with three clusters formed by the vowels /i y e/, /ø ɛ a/, and /o u/. While accurate vowel category perception may be maintained between 523–880 Hz due to distinctive spectral patterns still observable at these frequencies, it seems unlikely that vowel perception can be as accurate at the three highest f_o due to spectral proximity within the clusters. However, it is probable that listeners can distinguish vowels between the three clusters up to the highest f_o humans are capable of producing. This finding may also have implications for the quantal theory of speech [25] and speech evolution theories, as it could shed further light on why these vowels can be found in almost all the world's languages.

Several questions remain, however. The present study relied on five phonetically trained listeners who participated in the identification test and achieved an identification score of >80% for the selected vowels. Their phonetic expertise and the ability to replay a stimulus may have contributed to their performance. These listeners also made similar judgments on >30,000 vowels as part of creating a larger vowel and voice corpus [14,26,27]. Further studies would be necessary to determine the factors that helped them in the identification. Results from previous studies utilizing vowels from the same corpus and other sources are more consistent with the auditory-perceptual analyses presented here, as they also suggested a special role of the point vowels at high f_o [10,11,28]. These results propose that whole spectrum features may play an important role in vowel perception generally and that they provide a more comprehensive account than theories based solely on formant frequency patterns.

ACKNOWLEDGEMENTS

This study was supported by the NCCR Evolving Language, funded through the Swiss National Science Foundation Agreement #51NF40_180888. We are also grateful for the partial funding received from the SNSF Grants No. P2ZHP1_168375, P400PG_180693, and 100016_143943/1. We would like to express our gratitude to Dieter Maurer for his impactful work on vowel acoustics that inspired this research. Special thanks go to Stuart Rosen for his help with the auditory excitation patterns and to Paul Iverson for his contributions to the MDS analysis.

5. REFERENCES

[1] H. von Helmholtz, Die Lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik. Vieweg+Teubner Verlag, 1913.

[2] H. Hollien, A. P. Mendes-Schwartz, and K. Nielsen, "Perceptual confusions of high-pitched sung vowels," J. Voice, vol. 14, pp. 287-298, 2000.

[3] J. Sundberg, "Perception of Singing," in The Psychology of Music, Elsevier Inc., 2013, doi:10.1016/B978-0-12-381460-9.00003-1.

[4] D. Maurer, P. Mok, D. Friedrichs, and V. Dellwo, "Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer," in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2132-2133, 2014.

[5] D. Maurer and T. Landis, "Intelligibility and spectral differences in high-pitched vowels," Folia Phoniatr. Logop., vol. 48, pp. 1-10, 1996.

[6] W. Strange, R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman, "Consonant environment specifies vowel identity," J. Acoust. Soc. Am., vol. 60, pp. 213-224, 1976.

[7] D. Friedrichs, D. Maurer, H. Suter, and V. Dellwo, "Vowel identification at high fundamental frequencies in minimal pairs," in Proc. 18th Int. Congr. Phonetic Sci., paper no. 0438, pp. 1-5, 2015.

[8] D. Friedrichs, F. Nolan, and S. Rosen, "Do listeners rely on dynamic spectral properties in the recognition of high-pitched vowels?" in The Journal of the Acoustical Society of America, vol. 146, no. 4, p. 3051, 2019.

[9] D. Friedrichs, D. Maurer, and V. Dellwo, "The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz," J. Acoust. Soc. Am., vol. 138, no. 1, pp. EL36-EL42, 2015.

[10] D. Friedrichs, D. Maurer, S. Rosen, and V. Dellwo, "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks," J. Acoust. Soc. Am., vol. 142, no. 2, pp. 1025-1033, 2017.

[11] Y. Zhang, F. Nolan, and D. Friedrichs, "Perceptual clustering of high-pitched vowels in Chinese Yue Opera," Speech Commun., vol. 137, pp. 60-69, 2022.

[12] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," Nature, vol. 427, no. 6970, p. 116, 2004.

[13] D. Friedrichs, S. Rosen, P. Iverson, D. Maurer, and V. Dellwo, "Mapping vowel categories at high fundamental frequencies using multidimensional scaling of cochlea-scaled spectra," J. Acoust. Soc. Am., vol. 140, no. 4, p. 3219, 2016.

[14] D. Maurer, C. d'Heureuse, H., Suter, V., Dellwo, D., Friedrichs, T., Kathiresan, "The Zurich Corpus of Vowel and Voice Quality, Version 1.0," 2018.

[15] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. Inst. Phonetic Sci., vol. 17, pp. 97-110, 1993.

[16] D. Friedrichs, D. Maurer, H. Suter, and V. Dellwo, "Methodological issues in the acoustic analysis of steady state vowels," in Trends in Phonetics and Phonology: Studies from German Speaking Europe, A. Leemann, M. J. Kolly, S. Schmid, and V. Dellwo, Eds. 2016.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program] (version 6.3.09)," 2023. [Online]. Available: http://www.praat.org/ (Last viewed March 22, 2023).

[18] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hear. Res., vol. 47, pp. 103-138, 1990.

[19] S. Puria, W. T. Peake, and J. J. Rosowski, "Sound-pressure measurements in the cochlear vestibule of human cadaver ears," J. Acoust. Soc. Am., vol. 101, pp. 2754-2770, 1997.

[20] R. N. Shepard, "The analysis of proximities: multidimensional scaling with an unknown distance function. I," Psychometrika, vol. 27, pp. 125-140, 1962.

[21] R. N. Shepard, "The analysis of proximities: multidimensional scaling with an unknown distance function. II," Psychometrika, vol. 27, pp. 219-246, 1962.

[22] P. Iverson and P. Kuhl, "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," J. Acoust. Soc. Am., vol. 97, no. 1, pp. 553-562, 1995.

[23] D. Kewley-Port and B. S. Atal, "Perceptual differences between vowels in a limited phonetic space," J. Acoust. Soc. Am., vol. 85, pp. 1726-1740, 1989.

[24] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. Roy. Stat. Soc. Ser. B (Methodological), vol. 57, no. 1, pp. 289-300, 1995.

[25] K. N. Stevens, "On the quantal nature of speech," J. Phonetics, vol. 17, pp. 3-45, 1989.

[26] D. Maurer, H. Suter, D. Friedrichs, and V. Dellwo, "Acoustic characteristics of voice in music and straight theatre: Topics, conceptions, questions," in Trends in Phonetics and Phonology: Studies from German Speaking Europe, A. Leemann, M. J. Kolly, S. Schmid, and V. Dellwo, Eds., 2016.

[27] Maurer, D., Acoustics of the Vowel—Preliminaries. Bern, Switzerland: Peter Lang AG, International Academic Publishers, 2016.

[28] D. Friedrichs, "Beyond formants: vowel perception at high fundamental frequencies," Doctoral dissertation, University of Zurich, Zurich, Switzerland, 2017.

Can L2 Speakers Perceive Sociopragmatic Nuances at the Sound Level? Exploring Chinese Listeners' Perceptions of (ING)

Ke Lin

Georgetown University
kl863@georgetown.edu

ABSTRACT

The variation between the alveolar and velar forms (ING) has been widely used in conveying social meanings for native speakers in many English varieties. L2 speakers' perception of (ING), however, remains unknown. The presented study employs a matched guise task exploring Chinese listeners' mental representations of the (ING) variable, and a social network questionnaire, assessing how their associations interact with the levels of integration in the host society. Results from multivariate analysis show that (1) listeners with higher social network strengths perceived the speakers to be less educated, regardless of the (ING) variant; (2) the velar variant was perceived as more educated; (3) speakers' gender, and regional background had greater impacts on listeners' ratings than the (ING) guise.

Keywords: sociophonetics variation, perception, L2 processing

1. INTRODUCTION

We conduct a perception study on L2 speakers' perceptibility and judgments of a well-documented, stable phonological English variable (ING). For native speakers, the most common realizations of (ING) is the standard velar form, *-ing*, and the non-standard, alveolar form, *-in.* A vast amount of research has documented that in addition to the standard and non-standard differences, the variations have been used consistently in conveying social meanings in many English varieties [1,2,4,5]. Cross-dialectically, the standard velar variant is associated with prestige and formality while the alveolar variant with lack of sophistication and articulateness. While variations in productions and perceptions of (ING) among native speakers received much attention, no consistent patterns or clear explanations have been given to the variation of (ING) for immigrants in Anglophone countries. To bridge the gap, we investigate the perception of (ING) by native speakers of Chinese living in the U.S. We seek to uncover the following meanings that L2 speakers ascribe to the variants of (ING):

i. How salient is (ING) in speech to Chinese speakers in the U.S.?
ii. What are the meanings of (ING) for L2 speakers in the U.S. and to what extent do associations between (ING) variants and the social meanings resemble those of the native speakers, as demonstrated in Campbell-Kibler [5] and Kiesling [4]?
iii. Is hearers' degree of integration into the L1 society related to their evaluation of (ING)?

We envision results of this study may inform e.g., language and communication training. By understanding how L2s perceive and interpret the sociolinguistic cues associated with phonological variants, educators can improve lessons to support learners in developing sociolinguistic competence and navigating social meanings. By emphasizing the influence of social networks on perception, we also highlight the importance of intercultural communication skills for successful social integration.

2. BACKGROUND

2.1 Meanings of (ING) for L1 speakers

Early studies found that both in the U.S. [1] and the U.K. [3], men used the informal alveolar variant *-in* more than women. The alveolar variant was used more in Southern U.S. states [2], and the velar variant was used more often in formal speech. Newer studies uncovered more nuanced social meanings that (ING) contribute to the speakers' identity and stance. Kiesling [4] observed that boys in fraternity groups used the informal variant *-in* to index *solidarity, roughness, masculinity*, and *-ing* when exhibiting structural *power*. Thus, the social meanings of the variants evoke a wide range of impressions depending on the context. At the perception level, impressions of the (ING) variants

are even more complex. Campbell-Kibler [5] performed a matched guise study in which she matched variants with both Southern and Californian speakers' speech. The alveolar *-in* variant strengthened perceptions of the Southern accent, *informal*, *masculine, country, laidback,* and *lack of education,* the velar *-ing* evoked perceptions of *formal, urban, gay,* and *metrosexual*.

2.2 (ING) production by L2 speakers

Considering L2 English speakers, Schleef [6] found that Polish teenagers who interacted mainly with Scottish or mixed Polish-Scottish teens preferred to use -in more often than Polish teens who interacted only with other Polish teens. Drummond [7] observed that adult Polish women produced the alveolar form more frequently than men, which is the opposite pattern observed in native speakers. They speculated that women worked jobs where it was easier to adopt the local vernacular variant. However, this explanation does not apply to the teenagers in Schleef's study, as both boys and girls had similar exposure to native speech. Therefore, other factors must contribute to the difference in usage between gender and local variants. While the velar form is associated with sophistication and formality for L1s, L2 speakers (of varied proficiency and contact level with L1s) may not share the same understanding. As such, L2 speakers' backgrounds and the types of interactions with L1 speakers may contribute to a different encoding process of the (ING) forms. Expanding on how social clusters shape linguistic practice (e.g., as demonstrated in Milroy [8,9]), we include L2 speakers' levels of interaction with native speakers in the analysis.

3. METHOD

We conducted a matched guise study combined with a perception questionnaire to investigate L2 speakers' interpretations of social meanings associated with variants of (ING). Our study largely followed Campbell-Kibler's [5] procedure for consistency with previous work, but we used different strategies for speech elicitation and generating audio. In addition, we included a survey on language use and a social network to assess participants' integration in academic or professional life, living arrangement, and social and recreational life based on Ager & Strang [10]. While the matched guise task captured participants' covert associations with (ING) variants, the language use and social network survey explored the social conditions influencing L2s' perceptions of the variants.

3.1 Participants

Study participants were recruited from various parts of the U.S. as individuals who identified any dialect of Chinese as their first L1 language, and who have lived in the U.S. for at least one year. The minimum length of stay requirement ensured that they had first-hand exposure to local native speakers in their communities, and enough time to develop their local networks. 89 participants at ages ranging from 18 to 55 took part in this study virtually from home.

3.2 Audio stimuli

Audio recording comes from two native speakers of American English from California and two from the South (Georgia, Mississippi), collected over Zoom. First, each speaker was asked to tell a short story based on a series of comic strips [11] using as many present progressive verb forms as possible. The speakers then re-recorded segments with (ING) tokens in both alveolar and velar forms, matching the original intonation and speed. Using the velar and alveolar (ING) variants of the segments, we used the *cut* and *concatenate* functions in Praat [12] to replace the (ING) words with the reproduced alveolar and velar counterparts, yielding two pairs of manipulated recordings for each original story. We selected two 20- to 30-second-long segments from each speaker in each guise, yielding 8 audio clips, each containing 4 to 8 tokens of (ING) words. Every pair of recordings matches in every aspect except for the production of (ING) words. To reduce artifacts due to the manipulation, we manually adjusted the volume, pitch, and speed of each inserted segment to match the surroundings.

Qualities	*arrogant, irritable, hardworking, effective communication skills, perfunctory, confident, good management skills, logical/organized, intelligent, well-educated, has a high salary, humorous, lazy, not fit*
Personas	*student, boss, stoner, jock, urban dweller*
Stereotypes	*redneck, stoner, blue-collar*

Table 1: Social impressions elicited from Chinese listeners selected into the survey instrument.

3.3 Pilot interview

We first played the eight manipulated clips to five Chinese listeners and elicited opinions on the audio quality and speakers. We then incorporated the most frequently reported impressions into the experiment materials (see Table 1 for the selected impressions).

3.4 Experiment materials

3.4.1 Matched guise task survey

Next, we exposed each participant to 8 different parts of the 8 audio clips, each with a fixed *speaker guise:* 4 speakers with the *-ing* form, and 4 using *-in*. Participants evaluated each speaker based on (1) the levels of their 7 perceived personal *traits* on a continuous 6.0 Likert scale (*formal, educated, warm, extroverted, intelligent, engaging, talks very fast*), (2) the presence of social attributes from Table 1 (*qualities*, *personas*, and *stereotypes)*, as yes/no/skip answers, and (3) gender (*male, female*), ethnicity (*White, Black, Asian, Hispanic or Latino, unsure*), region (*New England, South, East Coast, West Coast, Other, unsure)*. The presented audio parts were selected by splitting the 8 clips in halves, re-assembling in two survey versions A, B ([14,15]), and assigning to participants at random such that each would hear all original 4 speakers with both -ing and -in in their survey.

3.4.2 Language and social network questionnaire

In the questionnaire, we collected participants' anonymized *id*, *age, age of arrival in the U.S.,* and *gender*. We characterized their engagement with the host community by surveying six social networks: *workplace, collaborators, interest groups, friends, previous housing, and current housing*. Participants assessed each network based on if it consisted of *mostly Americans (+1), mostly Chinese (+0), balanced Chinese and Americans (+0.5),* or *others (+0)*. We assigned scores in the brackets and aggregated them across all 6 networks to obtain the overall *social network strength* score, between 0 (lowest) and 6 (highest strength).

4. DATA ANALYSIS

Surveys were delivered via Qualtrics [15]. After eliminating incomplete and invalid responses (e.g., submitted in under 5 minutes), we obtained 41 responses for Survey A [13], and 42 responses for Survey B [14]. We applied multivariate analysis using Rbrul [16] to infer how various aspects of speakers and survey participants interacted with participants' social perceptions of the speakers. We included seven potential predictors: *speaker guise*, *speaker gender*, *speaker region, participant id*, *participant age of arrival to the U.S., participant gender,* and *participant social network strength*. The outcome variables were set to the ratings of the seven participant's impressions of the speakers: *formal, educated, warm, extroverted, intelligent, talks very fast,* and *engaging*. We used linear regression for modeling.

5. RESULTS

Between the seven impressions of social traits, only *formal* showed notable differences in the distribution of mean scores across all factor groups. The difference in the mean scores for *formal* was the greatest when interacting with the speakers' gender (see Table 2). The high range in the mean scores suggests that, overall, participants were least reluctant to express their opinions on the formality for speakers, while for the other social traits, they expressed mostly neutral opinions. Interestingly, while the scores for *educated* interacted with speaker *gender*, speaker *region*, (ING) *guise*, and self-reported *social network strength*, *intelligence* – a trait very close in meaning – did not show the same effects. Although the interaction of *educated* and *network strength* is statistically significant (see Table 2), the coefficient is only –.11. The network strength makes an impact on the directionality of the perception, but the effect is small. For both *educated* and *formal* (see Figure 1)*,* the 2 female speakers were rated higher than male speakers, regardless of (ING). The Southern speaker received higher average ratings for both *educated* and *formal* than the west coast speaker, revealing a reversed pattern from impressions previously documented for native English speakers [5].

Qualitatively, when performing the pilot study, we found the saliency of the distinction between velar and alveolar (ING) to vary by the participant's native dialect in Chinese, and the language proficiency in L2. Those who expressed having been occupied by the goal to comprehend the content did not notice the difference. Speakers that did not distinguish between alveolar and velar forms in their dialects found it more difficult to do so in English. Quantitatively in the experiment, the distinction between the (ING) variants did shape the pattern in the perception of *formal* and *educated*.

For both traits, the velar form received slightly higher mean scores than the alveolar form. While this pattern seems to match previous results [4,5], it should be noted that both forms received a quite neutral score (ranging from 2.21 to 2.94 on a 6.0 Likert scale), which suggests that (ING) is not highly indicative of the social meanings for the average Chinese listener. Finally, regardless of (ING) guises, participants with higher *social network strengths* were likely to give lower *educated* ratings to all speakers: perhaps, the more integrated individuals felt more confident in the host society and did not need to inflate their scores.

FORMAL ~ Speaker gender ($p < 0.0001$)		
Factors	Coefficient	Mean Score
female	0.747	3.08
male	–0.747	1.58
FORMAL ~ Speaker region ($p < 0.0001$)		
South	0.295	2.63
West Coast	–0.295	2.04
FORMAL ~ Speaker guise ($p = 0.04$)		
velar	0.125	2.46
alveolar	–0.125	2.21

EDUCATED ~ Speaker gender ($p < 0.0001$)		
Factors	Coefficient	Mean Score
female	0.506	3.36
male	–0.506	2.35
EDUCATED ~ Speaker region ($p < 0.0001$)		
South	0.286	3.14
West Coast	–0.286	2.57
EDUCATED ~ Social network score ($p = 0.002$)		
+1	–0.11	
EDUCATED ~ Speaker guise ($p = 0.02$)		
velar	0.0843	2.94
alveolar	–0.0843	2.55

Table 2. Variables significantly predicting the perception of *formal* (top) and *educated* (bottom).

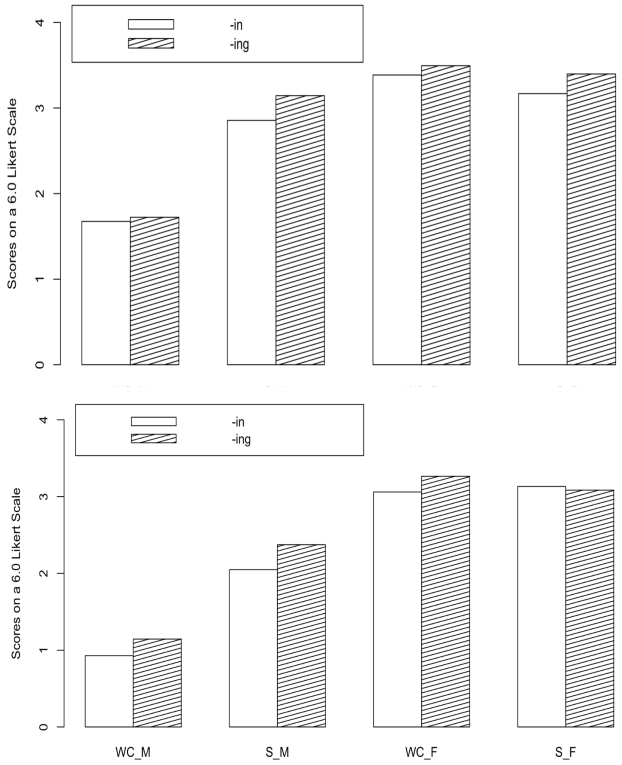


Figure 1. Average ratings for *educated* (top) and *formal* (bottom) across guises (*-in, -ing*) and speakers by gender (M, F) and region (West Coast: WC, Southern: S).

6. CONCLUSION

This work is the first attempt at investigating L2 listeners' social perceptions of (ING), a stable and socially meaningful English variable for L1 speakers. The study encompassed both a theoretical and a practical aim. For theoretical considerations, it aimed to provide a perspective on people's ability in sociopragmatic acquisition in L2, elucidating whether adults can acquire social meanings and nuanced linguistic variation at the phonological level in their second language. By co-investigating both L2 speakers' sociophonetic perceptions and their levels of interaction with the native speakers of the host community, we may eventually uncover the social conditions that facilitate the acquisition of sociopragmatic competence in their L2. At the practical level, when we establish that speech features at the phonological level can affect the way L2 speakers' perceive the social traits of the speakers, we can facilitate communication and enhance understanding in intercultural scenarios.

7. REFERENCES

[1] Labov, W. (1966). The Social Stratification of English in New York City. Washington, DC: Center for Applied Linguistics.

[2] Labov, W. (2001). Principles of linguistic change Volume 2: Social factors. *LANGUAGE IN SOCIETY-OXFORD-*, *29*.

[3] Trudgill, P., & Trudgill, S. (1974). *The social differentiation of English in Norwich* (Vol. 13). CUP Archive.

[4] Kiesling, S. F. (1998). Men's identities and sociolinguistic variation: The case of fraternity men. *Journal of Sociolinguistics*, *2*(1), 69-99.

[5] Campbell-Kibler, K. (2007). Accent,(ing), and the social logic of listener perceptions. American Speech, 82(1), 32-64. https://doi.Org/10.1215/00031283-2007-002

[6] Schleef, E., Meyerhoff, M., & Clark, L. (2011). Teenagers' acquisition of variation. English World-wide, 32(2), 206-236. https://doi.Org/10.1075/eww.32.2.04sch

[7] Drummond, r. (2012). Aspects of identity in a second language: ING variation in the speech of Polish Migrants living in Manchester, UK. Language variation and change, 24(1),107-133.

[8] Milroy, L., & Margrain, S. (1980). Vernacular language loyalty and social network. *Language in Society*, 43-70.

[9] Milroy, L. (2000). Social network analysis and language change: introduction. European journal Of English studies , 4 (3), 217-223

[10] Ager, A., & Strang, A. (2008). Understanding integration: A conceptual framework. Journal of Refugee Studies, 21(2), 166-191.

[11] Plauen, E. O., & Song, Rui. (2019). *Fu yu zi quan ji: Cai se shuang yu ban = Father and son*. Beijing: Bei jing jiao yu chu ban she

[12] Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.13)

[13] Lin, K. (2021, March). *Survey A*. Qualtrics. https://georgetown.az1.qualtrics.com/jfe/form/SV_exuqaUqfPk52lfM

[14] Lin, K. (2021, March). *Survey B*. Qualtrics. https://georgetown.az1.qualtrics.com/jfe/form/SV_PYfkRyRbwZXijc

[15] Qualtrics LLC. (2005). Qualtrics XM. https://www.qualtrics.com

[16] Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass*, *3*(1), 359-383.

WHEN IS ENOUGH, ENOUGH? VOT JUDGMENTS VARY BY VOICING INTENSITY, ASPIRATION INTENSITY, VOICE QUALITY, AND RATE OF CHANGE

John Kingston and Amanda Rysling

University of Massachusetts, Amherst and University of California, Santa Cruz
jkingston@linguist.umass.edu, rysling@ucsc.edu

ABSTRACT

Listeners categorized a short-long lag VOT /ba-pa/ continuum in eight conditions, which fully crossed (a) manipulating voicing intensity vs. voice quality, (b) whether time to ramp up from either 0 to peak voicing intensity or from lax to modal voice quality was short (10 ms) vs long (70 ms), (c) aspiration intensity (21 vs 33 dB). Listeners responded "pa" much more often when voicing reached peak intensity slowly, and measurably more often when voice quality became modal slowly. Apparently, voicing is not perceived until its intensity or quality crosses a threshold. Listeners' reliance on aspiration intensity was conditional: they only responded "pa" more with more intense aspiration if peak voicing intensity or modal voice were reached quickly.

Keywords: perception, VOT, aspiration and voicing intensity, voice quality

1. INTRODUCTION

Voice-onset time (VOT) is the most studied acoustic correlate of and perceptual cue to voicing contrasts between stops [1, 2]. In a typical study of its perception, stimuli are made by incrementally varying when a harmonically rich, periodic sound source, AKA "voicing," is turned on relative to when a high-pass filtered,[1] aperiodic sound source, AKA "noise," is abruptly turned on. The abrupt onset of the aperiodic source simulates the stop burst, and its continuation simulates any ensuing aspiration. What is typically reported is when, relative to the onset of that aperiodic source, listeners cross over from responding "voiced" to "voiceless," expressed as a VOT value. This category boundary has been found to vary as a function of other acoustic correlates, e.g., F1 onset frequency [3] and F0 [4, 5]. Even so, what the correlates of voicing are and how and why they serve as perceptual cues remain open questions.

Here, we take up these questions by manipulating voicing's intensity and quality independently of the timing of its onset. Repp [6] showed that listeners required a shorter lag in voice onset to respond "ta" to a stimulus when the intensity of the vowel was weaker relative to the intensity of the preceding aspiration. Using an aerodynamic model of stop consonant production, Ohala [7] showed that subglottal air pressure (Ps) is predicted to be lower during vowels following voiceless aspirated stops than those following voiced stops because glottal resistance is lower. Lower Ps would produce the less intense voice source observed following voiceless aspirated than voiced stops in English [8, 9, 10]. Lower Ps and weaker vowel intensity are expected after long- than short-lag stops, too, for the same reason: the glottal aperture is larger and resistance lower when the delay in voice onset is longer.

This expectation was confirmed: voice quality was laxer and breathier and intensity was lower for at least 25 ms following voiceless aspirated than voiceless unaspirated stops, presumably because the glottis did not close completely during early glottal cycles after voicing began, cf. [1, 11]. These observations led us to ask: do listeners use the laxer onset of voicing from coarticulation with an aspirated stop itself as a cue, or do they instead use the concomitant lesser intensity of voicing? Do a shorter VOT and weaker aspiration intensity suffice to elicit a "pa" response to a short-long lag VOT /ba-pa/ continuum when the vowel's initial voice quality was laxer, or do they only suffice if the vowel's initial voicing intensity were itself weaker?

2. METHOD

2.1. Stimuli

The Klatt synthesizer [12] was used to create eight short-to-long lag /ba-pa/ continua, which fully crossed three two-level factors (2x2x2): (a) dimension (voicing intensity, 0 to 72 dB, vs. voice quality, 0.9 to 0.5 open quotient and 18 to 0 dB spectral tilt), (b) ramp duration (from initial to final value in 10 vs. 70 ms), (c) aspiration intensity (21 vs. 33 db).[2] These manipulations produced steep or

gradual ramps in voicing intensity or voice quality, which occurred with either 21 dB or 33 dB of aspiration intensity. In all continua, VOT was 7, 18, 29, or 40 ms. The voiced portion of the vowel was constant in duration, cf. [13, 14, 15, 16]. Stimuli were presented synchronously to both headphone earpieces or 0.45 ms earlier to one earpiece; this tested the ultimately unsubstantiated hypothesis that intensity differences woould be more detectable at the midline [17], and so we do not return to it here.

The waveforms in Figure 1 show differences between the steep and gradual intensity ramps. Figure 2 shows that voicing reached peak intensity nearly as quickly for the gradual as the steep voice quality ramp. Figure 3 shows by standard measures that voice quality was laxer when the ramp was longer: the first harmonic, H1, was more intense relative to the second, H2, and the harmonic nearest F1, A1. The top of Table 1 shows that H1-H2 and H1-A1 were both greater when the voice quality ramp was gradual. The bottom shows H1-H2 didn't differ between 10- and 70-ms intensity ramps, and that H1-A1 did not differ nearly as much as it did between long and short voice quality ramps. All harmonics were at least 15 dB less intense when the intensity ramp was longer. Figure 4 shows aspiration intensities for the shortest and longest VOTs.

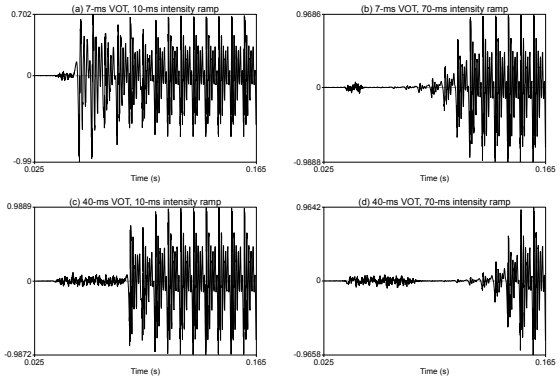


Figure 1: (a) 7-ms VOT, 10-ms intensity ramp, (b) 7-ms VOT, 70-ms intensity ramp, (c) 40-ms VOT, 10-ms intensity ramp, (b) 40-ms VOT, 70-ms intensity ramp. (33-dB aspiration intensity.)

2.2. Participants

34 listeners reported exposure only to English before age 6; 36 listeners reported early exposure to other languages (18 in all). 17 native and 16 non-native listeners responded to voicing intensity; 17 native and 20 non-native listeners to voice quality ramps.

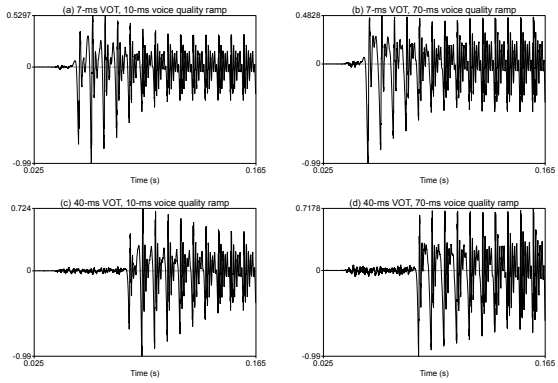


Figure 2: (a) 7-ms VOT, 10-ms voice quality ramp, (b) 7-ms VOT, 70-ms voice quality ramp, (c) 40-ms VOT, 10-ms voice quality ramp, (b) 40-ms VOT, 70-ms voice quality ramp. (33-dB aspiration intensity.)

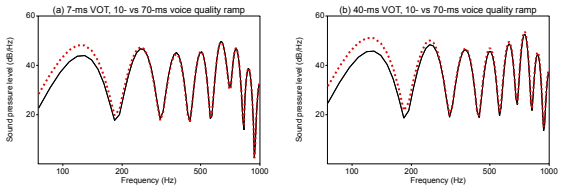


Figure 3: Spectra of 70 ms after voice onset; log 75-1000 Hz: (a) 7-ms VOT, (b) 40-ms VOT; (black solid) 10-ms voice quality ramp, (red dashed) 70-ms ramp.

VOT (ms)	7		40	
Ramp (ms)	10	70	10	70
H1-H2	-2.7	0.8	-2.5	1.2
H1-A1	-5.7	-1.2	-6.7	-2.3
H1-H2	-2.7	-2.6	-2.5	-2.5
H1-A1	-6.3	-4.2	-13.4	-5.2

Table 1: H1-H2 and H1-A1 in dB, after 7- and 40-ms VOTs, for 10- and 70-ms ramps: (top) voice quality, (bottom) intensity.

2.3. Procedure

A training block and then 9 test blocks were presented twice. Participants were trained with correct-answer feedback to respond "ba" to the stimuli with the shortest VOT and "pa" to those with the longest. Each of the 2 VOTs was presented once during each training block with the 2 aspiration intensities, the 2 ramp durations, and the 3 apparent locations, for 48 training trials. The 2 intermediate VOTs in all their combinations with the other

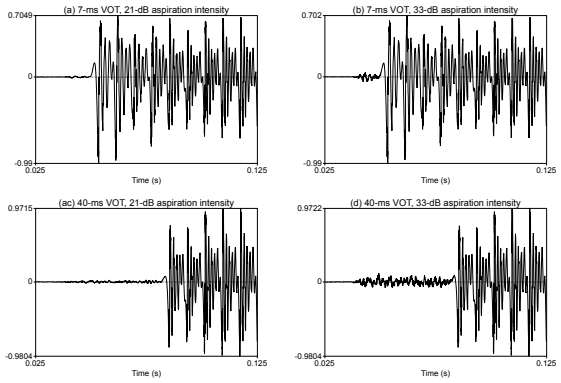


Figure 4: (a) 7-ms VOT, 21-dB aspiration intensity, (b) 7-ms VOT, 33-dB aspiration intensity, (c) 40-ms VOT, 21-dB aspiration intensity, (b) 40-ms VOT, 33-dB aspiration intensity. (10-ms intensity ramp.)

stimulus manipulations were added during each ensuing test block. Each stimulus was presented 18 times, for 864 total test trials. The order of stimulus presentation was randomized differently within each training and testing block for each participant.

Each trial begin with a display of a cross for 500 ms. The response options, "BA" and "PA," then appeared on opposite sides of the screen; which one appeared on the left vs right was counter-balanced between participants. The stimulus began when the response options appeared. Participants could respond while the stimulus was playing; afterward, they had up to 1500 ms more to respond. If they did not respond before that additional time had elapsed, they were encouraged to respond sooner. After the participant responded or the trial timed out, correct answer feedback was displayed for 500 ms on training trials. The interval before the next trial began was 750 ms in both training and test trials.

3. RESULTS

Figure 5 shows that, as expected, listeners responded "pa" more often as VOT lengthened. They also responded "pa" more when ramps were longer, but more so for voicing intensity than voice quality. Listeners only responded "pa" more often when aspiration was more intense if ramp duration was also short. When ramp duration was long, "pa" responses increased for less intense aspiration as well, enough to shrink aspiration intensity's effect when voicing intensity was manipulated. This shrinkage of aspiration intensity's effect across the two voicing intensity ramps occurred even though "pa" responses also increased *in general* for more intense aspiration. The difference in "pa" proportions between the stronger and weaker aspirations shrank for the long voice quality ramp.

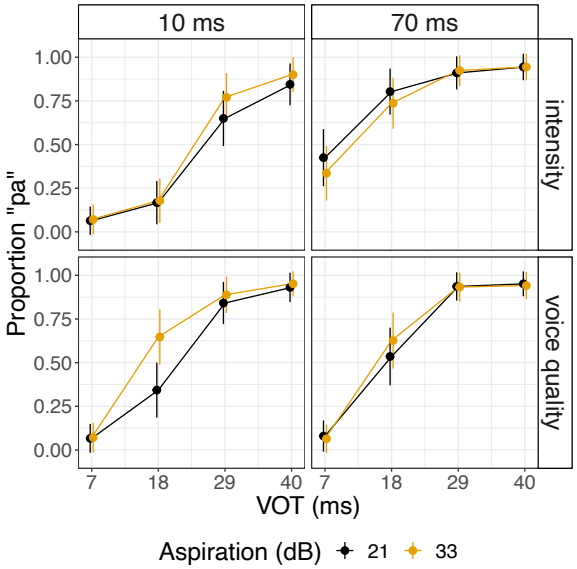


Figure 5: Mean proportions of "pa" responses (95% confidence intervals) by ramp duration, ramp dimension, VOT, and aspiration intensity

A Bayesian mixed effects logistic regression model was fitted to "pa" responses using the *brms* package [18]. Population-level effects included ramp dimension (voicing intensity = 0.5, voice quality = -0.5), ramp duration (10 ms = -0.5, 70 ms = 0.5), VOT (centered and scaled such that its standard deviation = 0.5) [19], and aspiration intensity (22 dB = -0.5, 33 dB = 0.5), and all interactions between them. A group-level effect of subject was included on the intercept and slopes of the within-subjects population-level effects (ramp duration, VOT, voicing intensity) and their interactions. Listener language exposure was not included; other analyses showed that the estimate's 0.95 credible interval (CI) included 0. Table 2 shows estimates of the included population-level effects.

The log Bayes factor (332.55) showed this model fit the data better than one without interactions between population-level effects. Chains did not diverge; autocorrelation within chains was modest; the model converged with `R-hat` values of 1.00 and ample effective sample sizes for all parameters; posterior predictive checks showed that replications closely matched the observed data.

The 95% CI for the intercept's positive estimate does not include 0, which shows a bias to respond "pa." No 95% CIs for the estimates of the individual population-level effects included 0; their signs show that listeners responded "pa" less often when the

	Est	Err	Lwr	Upr
Intercept	1.13	0.11	0.91	1.35
Dim(ension)	-0.71	0.22	-1.14	-0.27
Dur(ation)	1.47	0.12	1.24	1.71
VOT	5.39	0.27	4.86	5.93
Asp(iration)	0.47	0.09	0.29	0.66
Dim:Dur	1.84	0.23	1.38	2.30
Dim:VOT	-1.97	0.54	-3.03	-0.92
Dim:Asp	-0.60	0.18	-0.96	-0.24
VOT:Asp	0.47	0.16	0.15	0.79
Dur:Asp	-0.73	0.11	-0.96	-0.51
Dim:Dur:VOT	-2.02	0.35	-2.71	-1.34
Dim:Dur:Asp	0.35	0.22	-0.08	0.79
Dim:VOT:Asp	0.32	0.31	-0.29	0.93
Dur:VOT:Asp	-0.19	0.25	-0.69	0.30
Dim:Dur:VOT:Asp	0.24	0.46	-0.65	1.16

Table 2: Estimates of population-level effects, errors, and lower and upper bounds of 95% CIs.

manipulated dimension was intensity rather than voice quality (top vs. bottom panels in Figure 5), but more often when ramp duration was longer (right vs. left), VOT was longer (*x*-axis), and aspiration was more intense (colors). These effects are moderated by interactions with 95% CIs that do not include 0. When the manipulated dimension was voicing intensity rather than voice quality, the proportions of "pa" responses increased more for a longer ramp duration but less for a longer VOT or more intense aspiration (right vs. left in the top vs. bottom). When aspiration was more intense, the proportions of "pa" responses also increased more for longer VOTs but less when ramp duration was longer (left vs. right). The negative estimate for the three-way interaction of manipulated dimension, ramp duration, and VOT shows that "pa" proportions increased less for longer VOTs and the longer ramp duration when voicing intensity was manipulated rather than voice quality (top vs. bottom right).

4. DISCUSSION

The long, gradual ramp in voicing intensity dramatically increased the proportion of "pa" responses compared to the short, steep one; a finding that qualitatively and quantitatively resembles the effect of a slow, 60-ms long rise vs abrupt rise to peak voicing intensity in Darwin and Pearson's Experiment 1 [20]. Their /ba:pa/ category boundary was much earlier, 12.7 ms, when the rise was slow, and later, 30.6 ms, when it was abrupt, perhaps because voicing intensity was too weak relative to aspiration's intensity for listeners to use it when it rose slowly. Our category boundaries differed similarly between the long and short intensity ramp, 2.0 vs 25.5 ms. Both findings suggest that voice onset is not detected until its intensity reaches a criterial level. The proportion of "pa" responses also increased when the voice quality ramp was long and gradual rather than short and steep, but that increase is nearly entirely attributable to an aspiration intensity of 21 dB eliciting more "pa" responses when VOT was 18 ms; otherwise, "pa" responses increased equally and modestly for both aspiration intensities at VOTs of 29 and 40 ms when the voice quality ramp was longer. The effects of the duration of the voicing intensity ramp and aspiration intensity were more uniform across VOTs: "pa" responses increased for both aspiration intensities when the voicing intensity ramp was gradual.

But the potential three-way interaction was not supported: "pa" responses were no longer more numerous with more intense aspiration when the ramp duration was longer for either voicing intensity or voice quality. This collapse of the difference in voiceless responses between weaker and stronger aspiration levels with both longer ramps is surprising given Repp's original finding that listeners responded "ta" more often for a given voicing lag when aspiration was more intense [6]. The present results are also surprising given Darwin and Seton's finding that listeners responded "pa" more often when aspiration level was increased by 10 dB, and more intense aspiration increased "pa" responses more when the level of voicing was lower [21]. The surprise is most acute for the long intensity ramp because the voicing level remained low for so long, and voiceless endpoint responses might have been expected in the presence of only subtle voicing. But Darwin and Pearson did not find that category boundaries shifted to earlier VOT values when aspiration intensity was increased by 15 dB in stimuli in which voicing reached its peak intensity slowly [20]. There is thus some precedent for the finding that aspiration intensity is not determinative of voicelessness judgments when voicing intensity begins gradually. We can still conclude that those judgments depend on how late the periodic source was perceived to be too weak or too lax.

[1] To simulate acoustic coupling to the trachea while the glottis is open, the filter's lower cutoff is high enough that the aperiodic source does not excite the first formant.
[2] A ∼10 ms-long burst began with aspiration onset. Its intensity at successive 5-ms intervals was 0-45-35-0 dB.

5. REFERENCES

[1] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[2] A. S. Abramson and D. H. Whalen, "Voice onset time (VOT) at 50: Theoretical and practical issue in measuring voicing distinctions," *Journal of Phonetics*, vol. 63, pp. 75–86, 2017.

[3] K. R. Kluender, "Effects of first formant onset properties on vot judgments can be explained by auditory processes not specific to humans," *Journal of the Acoustical Society of America*, vol. 90, pp. 83–96, 1991.

[4] D. H. Whalen, A. Abramson, L. Lisker, and M. Mody, "Gradient effects of fundamental frequency on stop consonant voicing judgments," *Phonetica*, vol. 47, pp. 36–49, 1990.

[5] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody, "F0 gives voicing information even with unambiguous voice onset times," *Journal of Acoustical Society of America*, vol. 47, pp. 36–49, 1993.

[6] B. H. Repp, "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," *Language and Speech*, vol. 22, pp. 173–189, 1979.

[7] J. J. Ohala, "A mathematical model of speech aerodynamics," *Annual Report of the Institute of Phonetics, University of Copenhagen*, vol. 8, pp. 11–28, 1974.

[8] A. S. House and G. Fairbanks, "The influence of consonant environment upon the secondary acoustical characteristics of vowels," *Journal of Acoustical Society of America*, vol. 25, pp. 105–113, 1953.

[9] P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," *Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 454–460, 1963.

[10] I. Lehiste and G. E. Peterson, "Vowel amplitude and phonemic stress in American English," *Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 428–435, 1959.

[11] E. Fischer-Jørgensen and B. Hutters, "Aspirated stop consonants before low vowels, a problem of delimitation, – its causes and consequences," *Annual Report of the Institute of Phonetics, University of Copenhagen*, vol. 15, pp. 77–102, 1981.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Retrieved: 21 November 2022, 2022.

[13] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *Journal of the Acoustical Society of America*, vol. 32, pp. 693–703, 1960.

[14] A. J. Lotto, K. R. Kluender, and K. P. Green, "Spectral discontinuities and the vowel length effect," *Perception & Psychophysics*, vol. 58, no. 7, pp. 1005–1014, 1996.

[15] J. C. Toscano and B. McMurray, "Cue-integration and context effects in speech: Evidence against speaking-rate normalization," *Attention, Perception, & Psychophysics*, vol. 74, pp. 1284–1301, 2012.

[16] ——, "The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments," *Language, Cognition and Neuroscience*, vol. 30, no. 5, pp. 529–543, 2015.

[17] J. K. Maier, P. Hehrmann, N. S. Harper, G. M. Klump, D. Pressnitzer, and D. McAlpine, "Adaptive coding is constrained to midline locations in a spatial listening task," *Journal of Neurophysiology*, vol. 108, no. 7, pp. 1856–1868, 2012.

[18] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[19] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[20] C. J. Darwin and M. Pearson, "What tells when voicing has started?" *Speech Communication*, vol. 1, no. 1, pp. 29–44, 1982.

[21] C. J. Darwin and J. Seton, "Perceptual cues to the onset of voiced aspiraion in aspirated initial stops," *Journal of the Acoustical Society of America*, vol. 74, no. 4, pp. 1126–1135, 1983.

AUDIOVISUAL CUES IN DEVELOPMENTAL SPEECH PERCEPTION: DATA FROM A VOICELESS FRICATIVE PLACE DISCRIMINATION

[1]Laura Bosch, [2]Joan Birulés, [1]Ferran Pons

[1]Universitat de Barcelona, [2]Université Grenoble-Alpes
laurabosch@ub.edu, joanbirules@gmail.com, ferran.pons@ub.edu

ABSTRACT

Infants' perceptual narrowing to native speech sounds has been mainly described from auditory paradigms. Previous results from fricative place of articulation discrimination in infancy offer a non-convergent view on the factors modulating discrimination of noise contrasts. The present research focuses on infants' capacity to discriminate a native voiceless fricative place of articulation contrast, [s] - [f]. The contribution of audiovisual (AV) cues in reaching successful discrimination was explored. Three groups of 6-month-old infants were tested with versions of the familiarization-preference procedure involving two auditory-only conditions (adult-directed and infant-directed speech) and an AV presentation. Discrimination was absent and not favored by the AV format. Positive evidence was only obtained in an older group of 12-month-olds from the auditory-only version. The developmental pattern emerging from these data deviates from the expected maintenance of early language-general skills and suggests a late emergence of this fricative contrast with limited contribution of AV cues.

Keywords:voiceless fricatives, place discrimination, audiovisual cues, infants, perceptual narrowing.

1. INTRODUCTION

The field covering developmental speech perception research in infancy and its connection to later language learning skills has established that this process starts well before infants begin to utter their first words. Infants' attention towards the language spoken in their environment, increases their learning of the specific sound properties of their native language [1]. The speech signal is rich in information and it can be encoded based on a number of different dimensions. Infants' representation of these dimensions emerges from their initial perceptual biases, the acoustic-perceptual saliency of the information, the regularities present in the input language and the implicit learning mechanisms that are at work [2]. Phonetic learning in the first year of life has been characterized as a transition from language-general to language-specific speech perception resulting from experience with the native input. Attunement to the native sound properties leads to perceptual narrowing processes, described as a focus on language-specific segmental properties that sets the foundation of a contrastive, phonological system [3-4]. However, a review of the studies addressing discrimination of consonant contrasts in infancy reveals inconsistencies relative to the so-called canonical developmental pattern [5], especially affecting some nasal, liquid and fricative segments. The discrimination trajectory deviates from the universal model that posits an initial broad sensitivity for many different segmental contrasts, followed by an enhanced discrimination capacity for the native ones and a loss of sensitivity for the non-native ones. Alternative trajectories have been described, often revealing a pattern of a slower, protracted development, with gradual gains resulting from experience after an initial stage of poor or inconsistent discrimination [6]. The present research has placed the focus on infants' capacity to discriminate a native voiceless fricative place of articulation contrast, [s] - [f], to inform about its developmental trajectory and, additionally, to assess the possible contribution of audiovisual cues in the discrimination process of this consonant contrast. The role of visual speech, that is, seeing the articulation and paying attention to distinctive visual speech cues might enhance discrimination, as shown in research with young infants, tested with both native [7] and non-native consonant contrasts [8].

2. BACKGROUND RESEARCH

The study of fricative place of articulation discrimination in infancy has offered inconsistent data. Research has covered different contrasts, relative to different places of articulation, some of them more confusable than others. The location of the broadband aperiodic energy (higher vs. lower frequencies) can make discrimination relatively easy (e.g. [f] - [s]), or more confusable when the acoustic difference is reduced (e.g. [f] - [θ]). The former pair seems to be discriminated by 6-month-old infants [9], while discrimination of the latter contrast was not obtained till 12-14 months of age, still limited to

certain phonetic contexts [10]. However, a different study found positive evidence in 6-month-olds [11]. The alveolar - post-alveolar distinction, involving spectral and temporal differences was also found difficult to discriminate in 6- to 14 month-olds, contrary to data from an earlier study [12]. More recently, research exploring the effect of exposure on the discrimination of the alveopalatal - retroflex fricatives revealed restricted learning in 4- to 6-month-olds [13].

Overall, these studies offer a rather heterogeneous pattern of results, likely to be affected by acoustic as well as contextual factors, with some fricative pairs easier or less prone to confusion than others. Even though methodological approaches are slightly different, all have in common the use of auditory-only paradigms to test infants' discrimination or learning. Unfortunately, audiovisual (AV) paradigms have scarcely been used in this research domain, while they can offer relevant information to better understand and clarify the inconsistent pattern of results just described. Some place of articulation contrasts (such as the target contrast in this research, the labiodental - alveolar, [f] - [s] distinction) are produced with clearly visible articulatory cues, which can favour discrimination. It has already been shown, by studies exploring the distribution of attention to talking faces, that by around 6 months of age infants' attention to a talker's mouth gradually increases and the detection of the perceptual equivalence between auditory and visual properties of speech is likely to be connected to speech processing and phonetic learning [14]. Although in a different study, data suggesting a more restrictive attention shift to the mouth in 6 to 9 month-old infants has been reported limited to contexts in which auditory and visual information were in conflict [15], the possible contribution of AV cues in phonetic discrimination studies on challenging contrasts does need to be further assessed.

3. AIMS OF THE STUDY

Preliminary research in our Lab failed to find evidence of discrimination for a voiceless fricative place of articulation contrast in 6-month-olds, tested on the [f] - [s] native contrast, in spite of a clear difference in acoustic salience (larger distance in the center of broadband noise in the spectrum). Distributional properties of these segments in the native input favoring [s] might preclude an early differentiation or lead to an asymmetrical pattern of discrimination. Common temporal features of these fricatives might also constrain differentiation, while visual cues, if available, might have a facilitation effect. The present research aims at exploring the discrimination of this contrast using different conditions, controlling for the auditory characteristics of the material, and also including an AV version. To better understand the perceptual trajectory for this contrast, a group of one-year-olds was also been tested.

4. METHODOLOGICAL APPROACH

The experimental approach adopted in this research used versions of the familiarization-preference procedure [16-17], except for the AV modality experiment that required an habituation phase. Test trials were restricted to four (two Same and two Switch). Differential (longer) attention time to Switch trials in the test phase was interpreted as positive evidence of discrimination. Stimuli were CVC monosyllabic non-words, with the target fricative as syllable onset, followed by the vowel [i] and [t] as a coda. Variability, a common feature of the material in the these experiments, was implemented by using recordings from several speakers and a multi-token presentation.

5. DISCRIMINATION IN SIX-MONTH-OLD INFANTS

5.1. Experiments 1a and 1b: Auditory-only modality

5.1.1. Participants

A sample of N=18 (7 boys, M age in days = 186.1, range = 108-223) full-term 6-month-old infants with no history of hearing problems according to parents' report participated in Experiment 1a. An independent sample of N=18 (10 boys, M age in days 192.4, range 148-203) full-term 6-month-olds with similar characteristics participated in Experiment 1b. All participants came from Spanish-speaking or Catalan-speaking homes. The fricative contrast under study is present in both languages and has a similar distribution. The target stimuli were non-words in both these languages.

5.1.2. Material

For Experiment 1a, eighteen [sit] and [fit] tokens were selected from the recordings of six female speakers that were requested to produce the stimuli in a *motherese* style. In each category, twelve tokens were used for the familiarization phase and six for the test. Variability among tokens in each category was high, but significant differences were only found for the expected acoustic dimensions that characterize the target fricatives in this study (see Table 1).

For Experiment 1b, a selection of 16 tokens for each category (from those used in Experiment 1a, same six speakers) was used and they were classified in terms of speech register based on their proximity to Adult- or Infant-directed speech typical properties (ADS vs. IDS). Tokens to be used in the IDS condition differed from those in the ADS condition only in overall duration, vowel duration and F2 frequency at vowel transition.

	[sit]	[fit]	*p*-value
Duration (s)	1.06	1.11	n.s.
Noise duration (s)	0.27	0.29	n.s.
Vowel F0 (Hz)	256.7	280.9	n.s.
CoG (Hz)	6006.7	3634.5	< 0.001
F2 vowel trans. (Hz)	2609.9	2539.1	0.03

Table 1: Mean values of acoustic measures obtained for tokens in each category of fricatives used in Experiment 1a.

5.1.3. Procedure

In both a and b Experiments infants were tested with a version of the familiarization-preference procedure that consisted in a 2-minute infant-contingent familiarization phase with tokens of the same stimulus type (either [sit] or [fit]) presented on alternating trials from two lateral loudspeakers and followed by four test trials, quasi-randomly presented. Experiment 1b also involved two different conditions (ADS / IDS material) for each familiarization condition. Participants were randomly assigned to each testing condition.

5.1.4. Results

Experiment 1a: A one-way ANOVA ensured that both familiarization conditions ([sit] or [fit]) did not differ in number of familiarization trials needed to accumulate the 2 minutes' attention in this phase [$F(1, 16) = 1.95$, $p = .182$, petasq = .11]. Each participant's mean looking time (LT) to Same and Switch test trials was computed. A mixed ANOVA with Trial (Same, Switch) as a within-subjects factor and Familiarization Stimulus ([fit],[sit]) as a between-subjects factor. No significant results were obtained (all $p >$. 1; see Figure 2, left).

Experiment 1b: A mixed ANOVA with Trial (Same, Switch) as a within-subjects factor, and Stimulus type (ADS, IDS) and Familiarization Stimulus ([fit],[sit]) as between-subjects factors was run. No significant effects were found (all $p >$. 1).

5.2. Experiment 2: audiovisual modality

5.2.1. Participants

N=18 healthy, full-term and normally hearing infants (9 boys; M age=191.55, range=180-205). They came from the same language environments as participants in Experiment 1.

5.2.2. Material

Video stimuli were recorded for this experiment. Four different female speakers were recorded uttering several [fit] and [sit] tokens (see Figure 1) in a natural voice and using different intonation patterns as in IDS. It was first confirmed that adults were able to discriminate the fricative segments from a silent presentation.

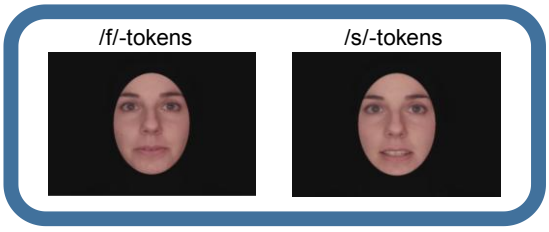


Figure 1: Two examples of frames from the video recordings with speakers producing [f] and [s] tokens for Experiment 2.

For each category ([fit] and [sit]) two blocks of video tokens were built to be used in the habituation phase trials. Each block was made of six different video tokens from two different speakers and lasted approximately 15 seconds. Another 12 different tokens from the same speakers were selected for the test phase, 6 for each category ([fit] and [sit]).The faces of the speakers were centered and faded in and out of a black background in each presentation.

5.2.3 Procedure

Infants were tested with a habituation-preference paradigm [18]. The high perceptual saliency of the audiovisual talking faces precluded the use of a 2-minutes' familiarization phase as in the previous experiments. The paradigm included a habituation phase using two different blocks of either [sit] or [fit] AV recordings, presented in a random order, until participants reached the habituation criterion (a decrease of 40% of attention time from the baseline). The test phase included 4 test trials, two Same and two Switch trials in a counterbalanced order. Testing took place in a soundproof room. The experimenter monitored infants' behavior from an adjacent control room, using Habit 1.0 software [19].

5.2.4. Results

A one-way ANOVA first ensured similar habituation times in both conditions (M habituation trials = 15.11; [fit] 15.33, [sit] 14.88; F(1, 16) = 0.04, p = .846, petasq < .01]. LT Scores were analyzed with a mixed ANOVA with Trial (Same, Switch) as a within-subjects factor and Habituation Stimulus ([fit], [sit]) as a between-subjects factor. The ANOVA yielded no significant results (see Figure 2), except for a trend in Trial [F(1, 16) = 3.95, p = .064, petasq = .20].

6. DISCRIMINATION IN ONE -YEAR-OLD INFANTS

6.1. Experiment 3: auditory-only modality

5.2.1. Participants

N=18 healthy, full-term 12-month-old infants (11 boys, M age = 370.94, range = 345-393), with no history of hearing problems. The language background was similar to the one described in the previous experiments.

5.2.2. Material and Procedure

Stimuli, procedure and apparatus were identical to Experiment 1a.

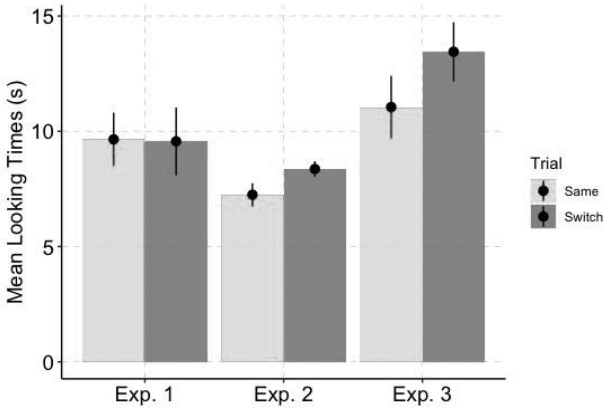


Figure 2: Mean attention time to Same and Switch trials for 6-month-old and 12-month-old infants tested in the auditory-only condition (Experiments 1 and 3), and for 6-month-old infants in the AV condition (Experiment 2).

5.2.4. Results

A one-way ANOVA on number of trials in each familiarization condition revealed that those familiarized with [fit] needed more trials to reach criterion [F(1, 16) = 4.88, p = .042, petasq = .23]. A mixed ANOVA with LT Scores and Trial (Same, Switch) as a within-subjects factor, and Familiarization Stimulus ([fit], [sit]) as a between-subjects factor revealed a significant main effect of Trial [F(1, 16) = 6.49, p = .022, petasq = .29], with longer attention to the Switch. Although the interaction of Trial with Stimulus Familiarization did not reach significance [F(1, 16) = 2.63, p = .124, petasq = .14], due to differences in the Familiarization phase infants' attention in the two groups were separately analyzed. Paired t-tests revealed that only infants familiarized with [fit] detected the change to [sit] [t(8) = 4.05, p = .004, d = 1.35], but not in the other direction (t <1).

7. CONCLUSIONS

The pattern of results in these experiments suggests a non-canonical developmental trajectory for a place of articulation fricative contrast present in the native language of the participants. Only by age 12 months first evidence of discrimination could be gathered, together with an asymmetrical pattern favouring one particular direction in change detection. Crucially, neither AV cues, nor exaggerated (IDS) presentation in the auditory-only modality were sufficient to increase the likelihood of contrast detection. This is especially relevant considering the variability in our testing material, based on different tokens produced by multiple speakers and designed to test phonetic categorization. Variability resulting from different voices and different stimuli used in our reserach paradigm clearly constrains discrimination by age 6 months, a result the diverges from data in [9], although the aim of that study and the paradigm used were different. The lack of a facilitation effect in the AV experiment, offering visible cues to support differentiation, reveals some limitation in the use of the redundant AV cues by 6 months, at least in this task. Further research should explore this issue with eye-tracking technology so as to identify how attention to the visual information is deployed during the task and which changes can be observed when older groups of infant participants are tested. Additional experiments using a different vowel to form the syllabic stimuli are also needed so as to confirm or modulate the present results.

8. ACKNOWLEDGEMENTS

The authors wish to thank the participating infants and their families. Research funded by Spanish MICINN projects (PGC 2018-097487-B-100, PID 2021-128159NB-100) to LB and FP. The authors also thank M. Ramon-Casas, J. Sánchez-Galán, M. Konjevod, N. Sanz-Gauntlett, and W. Elvira for their support with experiment preparation, phonetic analyses, data collection and coding.

9. REFERENCES

[1] Panneton, R., Newman, R. 2012. Development of speech perception. In: Werner, L.A. (ed.) *Human auditory development*. Springer, 197-222.

[2] Werker, J. F., Curtin, S. 2005. PRIMIR: A developmental framework of infant speech processing. *Language learning and development*, *1*(2), 197-234.

[3] Kuhl, P. K. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831-843.

[4] Werker, J. F., Yeung, H. H., Yoshida, K. A. 2012. How do infants become experts at native-speech perception?. *Current Directions in Psychological Science*, *21*(4), 221-226.

[5] Narayan, C. R. 2019. An acoustic perspective on 45 years of infant speech perception, Part 1: Consonants. *Language and Linguistics Compass*, *13*(10), e12352.

[6] Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental science*, *13*(3), 407-420.

[7] Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850-855.

[8] Danielson, D. K., Bruderer, A. G., Kandhadai, P., Vatikiotis-Bateson, E., & Werker, J. F. (2017). The organization and reorganization of audiovisual speech perception in the first year of life. *Cognitive Development*, *42*, 37-48.

[9] Beach, E. F., Kitamura, C. 2011. Modified spectral tilt affects older, but not younger, infants' native-language fricative discrimination. *Journal of Speech, Language and Hearing Research*, 54, 658-667.

[10] Eilers, R. E., Wilson, W. R., Moore, J. M. 1977. Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Research*, *20*(4), 766-780.

[11] Tyler, M. D., Best, C. T., Goldstein, L. M., Antoniou, M. 2014. Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non-native fricative place contrasts. *Developmental psychobiology*, *56*(2), 210-227.

[12] Nittrouer, S. 2001. Challenging the notion of innate phonetic boundaries. *The Journal of the Acoustical Society of America*, *110*(3), 1598-1605.

[13] Cristià, A., McGuire, G. L., Seidl, A., Francis, A. L. 2011. Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of phonetics*, *39*(3), 388-402.

[14] Lewkowicz, D. J., Hansen-Tift, A. M. 2012. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences, 109*(5), 1431- 1436.

[15] Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611-624.

[16] Jusczyk, P. W., Aslin, R. N. 1995. Infants′ detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1-23.

[17] Bosch, L., Sebastián-Gallés, N. 2003. Language experience and the perception of a voicing contrast in fricatives: Infant and adult data. In *Proceedings of the 15th international conference of phonetic sciences* (pp. 1987-1990). UAB/Casual Prods.

[18] Fennell, C. T. 2012. Habituation procedures. In: E. Hoff (ed.) *Research methods in child language: A practical guide*, Blackwell, 1-16.

[19] Cohen, L. B., Atkinson, D. J., & Chaput, H. H. (2000). Habit 2000: A new program for testing infant perception and cognition.

TEXT-TO-TUNE ALIGNMENT IN JAPANESE "HAPPY BIRTHDAY TO YOU"

Haruo KUBOZONO[1] and Ai MIZOGUCHI[1,2]

NINJAL[1], Maebashi Institute of Technology[2]
kubozono@ninjal.ac.jp, aimizoguchi@maebashi-it.ac.jp

ABSTRACT

This paper examines how text is aligned with musical tunes in Japanese "Happy Birthday to You". Although this song is sung with the same melody as its English version, words are segmented in different ways from English. Based on a production experiment involving twenty native speakers of Tokyo Japanese, this paper demonstrates that (i) four-mora or longer words are segmented based on the syllable rather than the mora, while shorter words are parsed by the mora, (ii) unlike lexical stress in English, lexical pitch accent in Japanese plays no role in the alignment, resulting in tonal neutralizations, (iii) vowel length is also neutralized in most cases, and (iv) vowel devoicing exerts a certain effect.

This paper also considers implications of these findings for Japanese syllable structure, with main focus on the questions of what vowel sequences constitute diphthongs and how (much) superheavy syllables are avoided in the language.

Keywords: text-to-tune alignment, Japanese, syllable and mora, lexical pitch accent, word segmentation

1. INTRODUCTION

How text is aligned with musical tunes has attracted serious attention in linguistics since the pioneering work by Liberman [1], illuminating not only the relationship between text and tune, but also the roles of the syllable and lexical stress as well as the strategies involved in word segmentation. Specifically, lexical stress in English and German plays an important role in the alignment, where words with non-initial stress behave differently from initially-stressed words.

The effect of lexical stress manifests itself in vocative intonation or 'calling contour', too, where the stressed syllable is aligned with the nuclear pitch accent H* before a downstepped !H ([2]):

(1) Calling contour in English

H* !H H* !H vs. L H*!H
Johnny! Jona than! Re becca!

(2) Calling contour in German

H* !H H* !H vs. L H*!H
Lisa Ursula! Hans-Jürgen

Text-to-tune relationship has been studied in Japanese, too, centering around the roles of the mora and the syllable. Linguistic analyses of Japanese songs have revealed the dominant role of the mora over the syllable (e.g. [3]), which is in line with the well-known fact that the mora is the basic prosodic unit in traditional Japanese poetry such as *haiku* and *tanka*. More recently, [4] has presented statistical evidence that syllabic alignment is used along with moraic alignment, especially in translated (vs. native) songs.

More interesting facts have been revealed by the series of work on the baseball chant sung by baseball fans to cheer on their favorite players ([5, 6, 7]). This chant consists of three musical notes followed by a pause as shown in (3), to which long names as well as short ones are to be aligned.

(3) | ♩ ♩ ♩ 𝄽 |
| | |
i chi roo

Linguistic analyses of this chant have clarified that the syllable as well as the mora plays a pivotal role ([5]) and that distinctions in lexical pitch accent and vowel length are both neutralized in some restricted contexts ([7]). A phonological test based on the role of the syllable has further revealed that Japanese has only three tautosyllabic vowel sequences, or diphthongs: i.e. /ai/, /oi/, and /ui/ (as well as /ei/ if it is recognized as a diphthong rather than a monophthong /e:/). It has also shown that what appear to be trimoraic syllables such as /ain/, /oin/, and /uin/ are actually processed as a sequence of two syllables, e.g. /a.in/, /o.in/, and /u.in/ ([8, 9]).

2. PRODUCTION EXPERIMENT

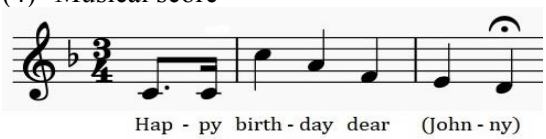
While the text-to-tune relationship in the baseball chant has revealed many things about Japanese phonology, it remains unanswered how general the alleged relationship is in the language, how words are segmented in other chants in Japanese, and how the language is different from English and other languages. To answer these questions, this current study has conducted a production experiment with 'Happy Birthday to You', a song that is sung with the same melody across languages. It is expected that this

experiment will reveal how Japanese speakers segment words and how their strategies are different from those used by English speakers.

2.1. Method

A total of 20 native speakers of Tokyo Japanese (10 males and 10 females) participated in the experiment. They are in two age groups, one aged between 19 and 46 years old, and the other in their seventies. The elderly group participated in a face-to-face experiment, while the younger group participated in an online experiment using Zoom. They were presented with a list of 109 names and were asked to sing the following portion of the Happy Birthday song twice with each name in the slot of *John-ny*. The list includes real Japanese names, English names, and new names coined from them.

(4) Musical score



To see the effects of various linguistic factors, the test words were selected in such a way that they range in phonological lengths (one mora to seven moras), syllable structures (different types of syllable weight), accent structures (presence/absence of a lexical accent as well as its position), morphological structure (monomorphemic vs. polymorphemic), and the nature of the final vowel (voiced vs. devoiced).

The subjects' responses in the online experiment were recorded using the recording function of the Zoom as well as a digital recorder placed near the speaker on the experimenter's side. Those in the face-to-face experiment were recorded in a silent room using a digital recorder. The data thus obtained consist of 2,180 tokens (109 stimuli x 20 subjects). They were analyzed auditorily by the experimenters as to how the test words are aligned with the tune in (4), that is, how they are segmented into two parts.

2.2. Results

The analysis has revealed the following pattern as the basic text-setting pattern in the Japanese song. Namely, the second musical note in the two-note slot in (4) is basically aligned with the last mora or syllable of the word, with the remaining syllables/moras all linked with the first note. This is illustrated in Table 1 with words ending in a light, i.e., monomoraic, syllable. Dots and hyphens indicate syllable boundaries and musical note boundaries, respectively.

Input		Segmentation pattern	Gloss
length	Word		
1μ	ri	ri-i	Li
2μ	ma.ma	ma-ma	mama
3μ	pa.me.ra	pa.me-ra	Pamela
4μ	ya.ma.ga.ta	ya.ma.ga-ta	Yamagata
5μ	fe.ru.nan.de	fe.ru.nan-de	Fernande
6μ	ma.ku.do.na.ru.do	ma.ku.do.na.ru-do	McDonald's

Table 1: Basic segmentation pattern.

Since words with a final monomoraic syllable like those in Table 1 are ambiguous between the two strategies, i.e. syllable-based vs. mora-based segmentation, we focused on stimuli that end in a heavy, i.e. bimoraic, syllable. A typical example is given in (5), which actually exhibits two segmentation patterns, a syllable-based pattern as in (5a) and a mora-based one as in (5b).

(5) te.he.ran 'Tehran' → (a) te.he-ran, (b) te.he.ra-n

This analysis has shown that the choice between the syllabic and moraic segmentation patterns in (5) is determined primarily by the length of the word. This can be confirmed by Fig. 1, where the proportion of the mora/syllable-based segmentation patterns is shown for each word length, for all speakers as well as for each age group.

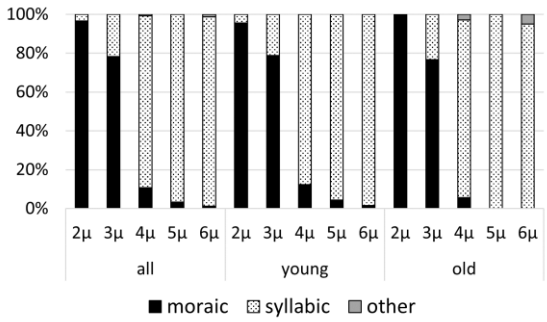


Figure 1: Segmentation strategy as a function of word length by the mora.

A crucial boundary exists between three-mora and four-mora stimuli: words consisting of two or three moras are aligned with the tune basically by the mora, with the last mora of the name associated with the second musical note on its own, whereas four-mora or longer words favor the syllable-based segmentation as shown in (5a). This result is in line with the result of the baseball chant reported in the literature ([5]). Moreover, the two age groups (young and old) show no noticeable difference, suggesting that word length plays a pivotal role independently of the speaker's age.

As for shorter stimuli, the most interesting case is that of light-heavy bisyllables, which favor segmentation at a syllable-internal mora boundary:

e.g. *a.ran* → *ara-n* 'Alan', *ta.roo* → *taro-o* 'Taro', *ji.mii* → *jimi-i* 'Jimmy', *ma.sai* → *masa-i* 'Masai'. This segmentation is strikingly different from the one used by native speakers of English and German, who respect syllable boundaries faithfully: e.g. *a-lan* 'Alan', *ma-sai* 'Masai'. The latter segmentation pattern is marginally observed in our Japanese data, too, but the mora-based segmentation overwhelms the syllable-based one in light-heavy bisyllables (72.5% vs. 27.5%).

The role of the mora can also be confirmed by the fact that bimoraic monosyllables are segmented at the mora boundary, with the coda consonant (moraic nasal) aligned with the second musical note on its own: e.g. *jon* → *jo-n* 'John'. This contrasts with the pattern usually observed in English and German, where coda consonants are aligned with the second musical note together with the preceding vowel: e.g. *jon* → *jo-on* 'John'.

While words consisting of one to three moras follow a moraic segmentation pattern, most four-mora or longer words are subject to the syllable-based segmentation shown in (5a): e.g. *wa.shin.ton* → *wa.shin-ton* 'Washington', *ku.ro.ma.tii* → *ku.ro.ma-tii* 'Cromartie', *an.do.rei* → *an.do-rei* 'Andrei'.

A careful analysis of the data further reveals that the syllabic vs. moraic proportion varies slightly depending on the type of heavy syllable. In four-mora stimuli, for example, the syllabic segmentation in (5a) accounts for 93% of stimuli ending in a long vowel but 85% of those ending in a moraic nasal.

3. DISCUSSION

3.1. Pitch accent's role

While our analysis has demonstrated the roles of the mora and the syllable in the text-to-tune alignment, it shows no evidence for the role of the lexical pitch accent. Our stimuli actually include many pairs of words like those in (6), which contrast either in the presence (') or absence (^{0}) of the pitch accent or in its position. However, none of these pairs showed a difference in the segmentation pattern: all test words simply followed the regular segmentation pattern in Table 1 and the musical melody shown in (4) above, hence showing no difference in tonal melody.

(6) a. ma'.ma → ma-ma 'mom'
ya.no^0 → ya-no 'Yano'
b. be.ru.gi'i → be.ru-gii 'Belgium'
be.ru.rin^0 → be.ru-rin 'Berlin'
c. a'.ki.ko → a.ki-ko 'Akiko'
a.ki.o^0 → a.ki-o 'Akio'
d. te'.he.ran → te.he-ran 'Tehran'
te.he'n.ra → te.hen-ra 'Tehenra (new name)'

It is worth adding here that not just the contrasts in lexical pitch accent but also those in vowel length are often neutralized. Thus, the pairs or triplets given in (7) are temporally neutralized as short vowels are often lengthened in the song.

(7) a. ma.no → maa-noo 'Mano'
maa.no → maa.noo 'Maano'
maa.noo → maa-noo 'Maanoo'
b. o.ta.ni → oo.ta-ni 'Otani'
oo.ta.ni → oo.ta-ni 'Ohtani'

3.2. Morphology's role

Our analysis has also shown that morphological structure plays no role. The test stimuli include both morphologically complex words as in (8a) and simplex words as in (8b). However, the text-to-tune alignment is not affected by the presence or absence of a word-internal morpheme boundary (#) or its position.

(8) a. ya.ma'#ga.ta → ya.ma.ga-ta 'Yamagata'
oo#ta.ni → oo.ta-ni 'Ohtani'
b. pi.no'.ki.o → pi.no.ki-o 'Pinocchio'
i.ta.ri.a^0 → i.ta.ri-a 'Italy'

3.3. Devoiced vowels

On the other hand, devoiced vowels in word-final position exerts a certain effect on the segmentation pattern. This can be seen from the examples in (9): words usually pronounced with a final devoiced vowel favor the seemingly irregular segmentation pattern in (i), where the final two syllables are aligned with the second musical note, over the pattern in (ii), where only the final syllable (with a full vowel) is aligned with the note. This irregular behavior of devoiced vowels is consistent with the finding from modern songs ([4]), as well as the well-known fact that moras with a devoiced vowel behave like a heavy syllable together with their preceding mora in the assignment of lexical pitch accent (e.g. [10]).

(9) a. fe.ru.na'n.de.su̥ 'Fernandes'
→ (i) fe.ru.nan-de.su̥, (ii) fe.ru.nan.de-su
b. ban.gu.ra.de'.shu̥ 'Bangladesh'
→ (i) ban.gu.ra-de.shu̥, (ii) ban.gu.ra.de-shu
c. ma.ru.ti'.ne.su̥ 'Martinez'
→ (i) ma.ru.ti-ne.su̥, (ii) ma.ru.ti.ne-su
d. moo.ri'.sha.su̥ 'Mauritius'
→ (i) moo.ri-sha.su̥, (ii) moo.ri.sha-su

The effect of devoiced vowels can be confirmed with the data in Fig. 2, which shows the proportion of the two segmentation patterns in (9)—(i) and (ii)— as a function of the nature of the word-final vowels in

five-mora or six-mora words ending in a sequence of two light syllables like those in (9).

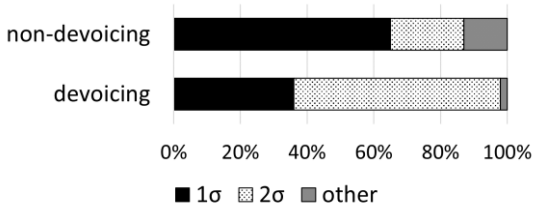


Figure 2: Segmentation patterns as a function of the nature of the word-final vowels.

4. IMPLICATIONS FOR SYLLABLE STRUCTURE

Having understood the basic principles underlying the text-to-tune alignment in the Happy Birthday song, let us finally consider the implications of these principles for the syllable structure of Japanese.

4.1. Diphthongs

One issue that is still debated in the literature concerns the inventory of diphthongs in the language ([8]). Since words longer than three moras predominantly favor the syllabic segmentation in (5a), we can use this strategy to test which vowel sequence actually belongs to the same syllable and which one does not in our alignment data. This analysis has revealed that the four types of vowel sequences in (10a) tend to be aligned with the second musical note, those in (10b) usually split into two musical notes, and /au/ in (10c) fluctuates between the two patterns.

(10) a. /ai/ o.chi-ai 'Ochiai'
/oi/ ma.ki-roi 'McIlroy'
/ei/ a.do-rei 'Adlay'
/ui/ kin-sui 'Kinsui'
b. /ae/ oo.ma-e 'Oomae'
/ao/ a.sa.ga-o 'Asagao, morning glory'
/oe/ ka.wa.go-e 'Kawagoe'
/eo/ mai.ne-o 'Mineo'
/ia/ i.ta.ri-a 'Italy'
/ua/ ma.na.gu-a 'Managua'
/ea/ e.ku.re-a 'éclair'
/oa/ e.ka.no-a 'Ecanoa'
/io/ pi.no.ki-o 'Pinocchio'
/ie/ yo.ri.i-e 'Yoriie'
/uo/ a.ka.u-o 'Akauo, red fish'
/ue/ ta.me.su-e 'Tamesue'
c. /au/ rin-dau ~ rin.da-u 'Lindau'

(10b) includes six vowel sequences that are permitted in native morphemes: /ae/, /ao/, /oe/, /ie/, /uo/, and /ue/. It is interesting to find that not all vowel sequences that are permitted in native morphemes form diphthongs. Overall, the results in (10) accord with the claim by [11] and [9], who demonstrated with dialectal accent data that only vowel sequences ending in /i/ form diphthongs in Japanese. One thing that is not reported in these previous studies is the behavior of /au/ in (10c): in the assignment of lexical pitch accent, this vowel sequence behaves as a sequence of two syllables just like those in (10b).

4.2. Superheavy syllables

Our data also provide some evidence pertaining to the debate about superheavy, i.e. trimoraic, syllables. Japanese permits what appear to be trimoraic syllables in loanwords: e.g. *wain* 'wine', *ainshutain* 'Einstein', *saaroin* 'sirloin', *rinkaan* 'Lincoln'. Accentual analyses in the literature have shown that these trimoraic sequences do not form one syllable, but actually a sequence of two syllables ([12, 9]).

Our data from the text-to-tune alignment support this previous finding to a large extent. They contain many words with a final moraic nasal preceded by a long vowel or diphthongal sequence as exemplified in (11). Of these, diphthong-like vowel sequences plus the moraic nasal in (11a) generally exhibit a bisyllabic behavior in our data, with the second element of the vowel sequences and the moraic nasal aligned with the second musical note. Here, it is interesting indeed that the vowel sequences in (10a)—/ai/, /oi/, /ei/, and /ui/—split into two syllables when they are followed by a coda consonant, although they otherwise behave as one syllable as shown in (10a). This reinforces the argument that superheavy syllables are disfavored, if not entirely prohibited, in Japanese (e.g. [8]).

(11) a. ko.ka-in, *ko-kain 'cocaine'
ba.ren.ta-in, ?ba.ren-tain 'Valentine'
kya.ro.ra-in, *kya.ro-rain 'Caroline'
ma.ke-in, *ma-kein 'McCain'
hi.ro-in, *hi-roin 'heroine'
saa.ro-in, *saa-roin 'sirloin'
b. rin-kaan > rin.ka-an 'Lincoln'
ha.ro-wiin > ha.ro.wi-in 'Halloween'
ha.ri-keen ~ ha.ri.ke-en 'hurricane'
to.ron-boon > to.ron.bo-on 'trombone'

On the other hand, monophthongal vowel sequences plus a moraic nasal in (11b) favor the monosyllabic pattern over the bisyllabic one (66.3% vs. 31.3%), suggesting that they resist being split to two syllables. This agrees with the result obtained from the analysis of the baseball chant ([9]), although it is different from what has been reported on the basis of a phonological test on the assignment of lexical accent ([8, 9]).

5. ACKNOWLEDGEMENTS

We would like to thank the subjects of our experiment. The work reported in this paper is supported by JSPS KAKENHI grants (19H00530 and 20H05617) to the first author.

6. REFERENCES

[1] Liberman, M. 1979. *The Intonational System of English.* Garland Publishing, Inc.

[2] Ladd, D. R. 1996. *Intonational Phonology*. Cambridge University Press.

[3] Kubozono, H. 1999. Kayō ni okeru mōra to onsetsu [the mora and the syllable in song]. In: Speech and Grammar Study Group (ed.), *Bunpō to Onsei II [Speech and Grammar 2]*. Kurosio, 241–260.

[4] Starr, R. L., Shih. S. S. 2017. The syllable as a prosodic unit in Japanese lexical strata: Evidence from text-setting. *Glossa*. 2(1), 93, 1–34.

[5] Tanaka, S. 1999. Nihongo no onsetsu to yonhaku no tenpurēto: Senryū to puroyakyū seien ni okeru ziamari no bunseki [Syllables and four-mora templates in Japanese: Analysis of 'ji-amari' in senryu poems and baseball chants]. In: Spoken Language Working Group (ed.), *Bunpō to Onsei II [Speech and Grammar 2]*. Kurosio, 261–290.

[6] Ito, J., Kubozono, H., Mester, A., Tanaka, S. 2019. Kattobase: The linguistic structure of Japanese baseball chants. *Proceedings of Annual Meeting of Phonology (AMP) 2018* (online). Washington, DC.

[7] Kubozono, H. 2020. Neutralizations in vowel length and word accent in Japanese. *Kobe Papers in Linguistics*. 12, 69–79.

[8] Kubozono, H. 2018. Prosodic evidence for syllable structure in Japanese. In: Guillemot, C., Yoshida, T., Lee, S. J. (eds.) *Proceedings of the 13th Workshop on Altaic Formal Linguistics. (MIT Working Papers in Linguistics 88)*. Tokyo, 35–50.

[9] Kubozono, H. 2022. *Word and Sentence Prosody: The Endangered Dialect of Koshikijima Japanese.* De Gruyter Mouton.

[10] Kubozono, H. 2002. Prosodic structure of loanwords in Japanese: Syllable structure, accent and morphology. *Onsei Kenkyu*, 79–97.

[11] Kibe, N. 2000. *Seinanbu Kyushu Nikei Akusento no Kenkyū [Studies in the Two-Pattern Accent Systems in South-west Kyushu]*. Bensei Shuppan.

[12] Kubozono, H. 1999. Mora and syllable. In: Natsuko Tsujimura (ed.), *The Handbook of Japanese Linguistics*, Blackwell, 31–61.

SPECTRAL CUES IN PITCH PERCEPTION IN ENGLISH AND CANTONESE SPEECH AND MUSIC

May Pik Yu Chan, Jianjing Kuang

University of Pennsylvania
pikyu@sas.upenn.edu, kuangj@sas.upenn.edu

ABSTRACT

Listeners integrate multidimensional cues (e.g., spectral shape) during pitch processing in both speech and music. While prior works have shown that this effect applies to tonal and non-tonal speakers, it remains unclear whether listeners' pitch processing strategies differ on different linguistic levels (e.g. tone vs intonation). Furthermore, work has yet to show whether differences in pitch processing lie between natural (both speech and music) vs artificial sounds, or speech vs non-speech sounds. A pitch classification experiment investigated these questions. Listeners were either English or Cantonese speakers listening to stimuli modeled after Cantonese prosody, or English speakers listening to comparable violin stimuli. Results show that all participants integrate spectral cues in pitch perception. However, the effect of musicality on shift but not categoricity differed between Cantonese and English speech conditions. Overall results suggest shared pitch processing for speech and music.

Keywords: Pitch perception, Spectral Slope, Speech and Music, Musicality, Prosodic effects

1. INTRODUCTION

Listeners rely on pitch in both linguistic (e.g. tone, lexical stress, phrasal prosody and intonation) and musical contexts (e.g. melody, harmony). Although fundamental frequency, F0, is understood as an important acoustic property of pitch, various works have shown that pitch perception is driven by multidimensional cues, such as timbral cues in music [1, 2, 3] or spectral slope cues in speech [4, 5, 6], intensity [7].

One core question of the present work is at what level of processing cue integration happens. For example, while work comparing speech and non-speech has shown that integrate spectral slope cues in both sawtooth and speech sounds, listeners are more sensitive to relative differences between spectral slope cues in speech but not non-speech [6]. In other studies focusing on music only, listeners were also found to integrate timbral cues, though works have tended to focus on fully instrumental comparisons [8, 9] or artificially synthesized stimuli [1, 2, 3], none of which have been directly comparable to speech.

Apart from speech vs non-speech, another question within speech is whether there is a processing difference between linguistic vs non-linguistic levels of processing. Prior work on integration comparing speakers' language backgrounds have not found language effects, regardless of whether speakers speak a tonal or non-tonal language [4], or a tonal language with multiple pitch height contrasts [10]. Furthermore, [11] has found that Cantonese participants integrate gender cues to the same extent with respect to lexical tones, though Cantonese musicians and non-musicians process pitch differently when listening to non-linguistic vocalization. We therefore speculate that the degree to which pitch is phonologically defined or specified may affect the extent of cue integration in pitch perception. In other words, when a "speech" mode encompasses both linguistic and non-linguistic speech, it is unclear whether the processing of such "language mode" differs from non-linguistic speech.

On a broader view, differences in processing may also potentially lie in natural vs artificial sounds. While [6] has shown that listeners process speech sounds differently from sawtooth sounds, it remains unclear whether this effect is due to sawtooth sounds not evoking a "speech mode", or because sawtooth sounds are not naturally occurring at all. In this study, we seek to compare speech sounds with violin-sounding stimuli. This is because violin sounds also contrast in "voice quality" with *sul ponticello* and *sul tasto* (playing closer to vs away from the bridge) resulting in differences in spectral slope. These sounds are comparable to "tenser" and "breathier" sounding voicing in speech respectively. Since bowed instruments like the violin are fixed-formant instruments with a clear F1 and F2 structure, it is directly comparable to speech with a stable vowel. The use of violin stimuli would

therefore allow us to compare speech vs non-speech, using stimuli that are still natural sounding.

We seek to answer the following questions: First, whether linguistic structure affects listeners' degree of cue integration in pitch perception. While prior work has shown that listeners' experience with tonal languages does not have effects on cue integration [6, 4, 10], it remains unclear whether the prosodic structure of the stimuli plays a role. Specifically, since previous work included stimuli modeled off English prosody, it is possible that, for Mandarin and Cantonese listeners, the stimuli based on foreign prosody do not trigger listeners' "linguistic mode" thereby dampening potential language effects. Therefore, in this study, we test whether the effect is transferable to stimuli of a syllable-timed prosodic structure that could be mapped to tonal categories that may affect listeners' pitch perception strategies. Second, whether musical stimuli showing within-instrument differences are as integrative as speech. And third, as various works have shown that listeners' musical competence affects pitch processing in both linguistic [12, 13, 14, 15] and non-linguistic tasks [16, 17] as well as integration strategies [4], we also test whether the effects of musicality remain across both languages and modalities. A pitch classification experiment was performed to test these questions.

2. METHODS

A total of 183 participants participated in a pitch classification experiment. 88 were native English speakers, 47 of which participated in the speech condition, 41 of which participated in the music condition. 95 participants were native Cantonese speakers and participated in the speech condition.

2.1. Pitch classification experiment

To test whether listeners of different backgrounds integrate spectral cues in pitch perception differently in speech and music modalities, a pitch classification experiment following the methods in [6] was done. Stimuli involving two pitch contours were resynthesized from a native Cantonese female talker's production of the pseudoword "ma-ma-ma", or from a violin's recording of the note sequence G3 (196 Hz), D4 (294 Hz), G3 (196 Hz). Unlike the stimuli in [6], the production of the pseudoword was in Cantonese T4-T1-T4 (low falling - high level - low falling, as in 麻媽麻) tones, which matched closely to the pitch interval of the violin note sequence. By extension, the tempo of our current stimuli includes three syllables being comparably long in duration, contrasting stimuli from [6] where lexical stress in the second syllable is significantly longer than the first and third unstressed syllables. Both recordings were made in the same recording environment and were around 700 ms.

Resynthesis was done in Praat, where the recordings were first normalized to 700 ms. The pitch tier from the Violin recording was extracted. It was then hand-corrected to smooth out transitions between notes, and to create a little pitch declination on the two G3 notes, referencing the low falling tone regions of the Cantonese recordings. An eleven-step continuum, each 0.35 semitones apart, was then created based on the manipulated pitch tier. Step 6 corresponded to the original pitch, and steps 1-5 were lower in F0, while steps 7-11 were higher in F0. Each stimulus consisted of two pitch peaks (in other words, 6 syllables, or 6 violin notes). The first pitch peak has a consistent F0 range, and only included the pitch peak at Step 6; the second pitch peak included pitch peaks from the 11-step continuum.

Spectral slopes of the sound sources were manipulated by inverse filtering of an LPC (burg) object in Praat, after resampling both sound files to 20k Hz. The original sound source was used as the "tenser" voice quality stimuli source, and a "breathier" version of the sound source was created by using the Filter (de-emphasis) command, which decreases the spectral slope by 6 dB/octave. The pitch tiers were then replaced by the manipulated pitch tiers described above, and the stimuli were filtered using the original LPC objects of the speech and violin recordings. The manipulated pitch peaks were then concatenated with the baseline (step 6) pitch peak in all "tenser" or "breathier" spectral slope permutations (TB, BT, TT, BB). A 500 ms gap was created between the two pitch peaks to avoid listeners comparing the last note of the first peak and the first note of the second peak, instead of listening to the overall contour. 88 stimuli were created in total (2 sources (speech/violin) x 4 spectral slope permutations x 1 baseline F0 x 11 F0 steps.)

Participants completed the experiment using Qualtrics. Cantonese-speaking participants only completed the speech condition, and English participants completed either the speech or the violin condition. A between-subject design was used to avoid training effects and to allow for sufficient repetitions of trials. Listeners listened to one concatenated pitch peak pair per trial and were asked to judge whether the second pitch peak was higher in pitch than the first. For English-

speaking participants, each stimulus was repeated 5 times; and for Cantonese-speaking participants, each stimulus was repeated 3 times. All trials were randomized.

To quantify participants' cue integration extents, a categoricity and a shift score were calculated per participant. The categoricity score was calculated by the mean 'the second peak is higher' response rates in steps 7-11 minus that of steps 1-5. Participants with more categorical responses will have a higher score, and those with less categorical responses will have a lower score. A shift score was calculated by summing the distance between the TB/BT condition and the baseline conditions. In other words, we take the absolute value of average responses from the TB minus BB, TB minus TT, TT minus BT, and BB minus BT. We then divide the summed value by two, as there are two baseline conditions. This allows us to scale the shift score from 0 to 1, such that it is comparable to the categoricity score. A high score would suggest that participants integrate spectral cues more, resulting in more shift from the baseline. A low score would suggest that the participant attends to F0 more, thus shifting less from the baseline.

2.2. Musicality test

We use a continuous musicality score as a proxy for listeners' musical competence, which can capture both the musical proficiency achieved from musical training and from innate musical disposition. Listeners completed the Montreal Battery of Musical Abilities (MBEMA) [18] after the pitch classification task, also administered on Qualtrics in one setting. The task includes a melody, rhythm and memory task. The melody and rhythm task asks listeners to distinguish whether two melodic phrases, that may differ slightly in the melody or the rhythm, were the same or different, The memory task asks participants to identify whether a melodic phrase was novel or repeated from the former two tasks. Each task includes 20 questions, resulting in 60 questions total. The percentage of participants' overall correct responses is used as a measure of participants' musical competence.

3. RESULTS

The proportion of "the second peak is higher" responses faceted by experiment condition is shown in figure 1. Across all three conditions, listeners integrate spectral cues in the same direction. When listeners hear tenser-breathier stimuli for example, listeners are biased towards thinking that the second pitch peak is higher in pitch, even in cases when the F0 is not necessarily higher. The opposite direction holds for the breathier-tenser condition. Among the baseline conditions, the tenser-tenser condition biases participants towards a higher second pitch peak compared to the breathier-breathier condition. Across the three experiments, participants had a mean musicality score of 50.66 (SD = 7.86), and their mean melody, rhythm, and memory scores were 17.03 (SD = 2.42), 16.57 (SD = 3.30) and 17.06 (SD = 3.28) respectively.

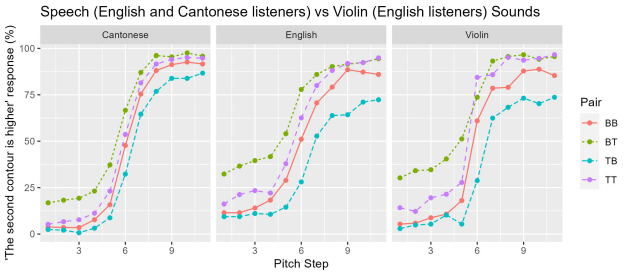


Figure 1: Percent the second contour is higher response rates

We fitted a mixed effect logistic regression model to evaluate the effect of spectral condition (4-factor levels), experiment condition (3-factor levels, reference: English), and pitch step (11 numeric steps) on participants' responses. The reference levels of the spectral condition were rotated to achieve pairwise level comparison. By participant random intercepts were included. Results showed a significant main effect of F0 (β = -0.609, SE = 0.007, $z = -93.024$, $p < 0.0001$), and that responses from the Cantonese condition differed significantly from the English condition (β = 0.229, SE = 0.074, $z = 3.114$, $p = 0.002$), but the Violin condition did not differ significantly from the English condition (β = 0.003, SE = 0.086, $z = 0.039$, $p = 0.969$). The main effect of spectral slope differed significantly from each other across all pairs, reaching p<0.0001 even after applying Bonferroni correction for multiple testing,

To understand the effect of musicality on listener responses, figure 2 shows participant responses after they were grouped into musicality quartiles within their respective experiment. Participants with the highest musicality scores are in the group 4 facet, who tends to have the most categorical responses with less differences between spectral conditions compared to other groups. By contrast, listeners in group 1, who have the lowest musicality scores in their experiment condition, show the

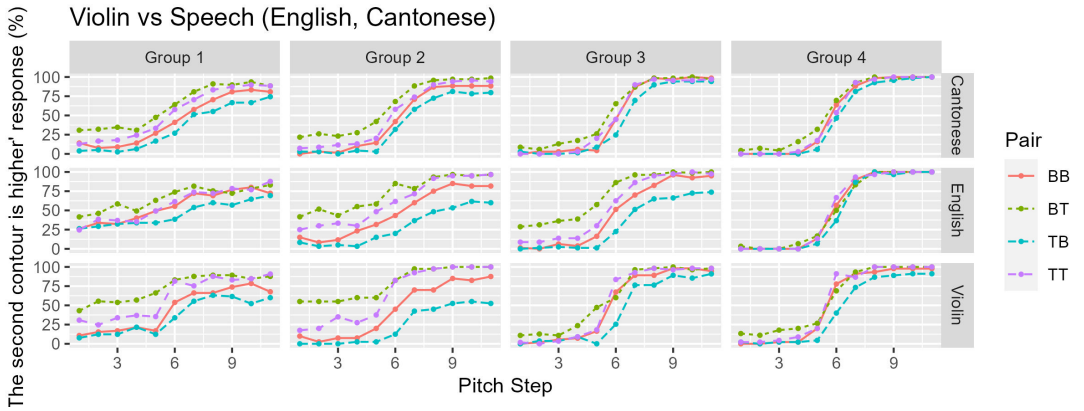


Figure 2: Responses faceted by musicality quantiles within each experiment

least categorical responses with more shift between spectral slope conditions. The difference between groups appears to be least robust in the Cantonese condition, where participants even in group one, appear to have relatively categorical responses.

We model participants' categoricity scores and shift scores by their musicality scores. We ran two linear regressions with the same model specifications, predicting either categoricity or shift. Main effects of musicality (numeric) and experiment condition (3-factor levels, reference: English) were included, as well as their interaction term.

Results for the categoricity model observed a significant main effect of musicality (β = 1.32, SD = 0.27, t = 4.84, p < 0.0001), suggesting that listeners with higher musicality have more categorical responses. No other terms were significant. Results for the shift model found no significant main effect of musicality, unlike results in [6, 4]. However, there is a significant interaction effect of musicality and the Cantonese experiment condition (β = -0.81, SD = 0.41, t = -1.98, p = 0.049). The results suggest that there is a stronger negative correlation between shift and musicality for Cantonese, which was significantly differing from English responses. Figure 3 illustrates the effect of musicality on categoricity and shift.

4. DISCUSSIONS

In this study, we aim to understand (1) whether linguistic level affects listener's pitch perception strategies, (2) whether speech vs non-speech mode results in different pitch processing strategies, and (3) whether the effect of musicality is consistent across conditions.

We test our first question by using stimuli modeled after Cantonese tones and comparing

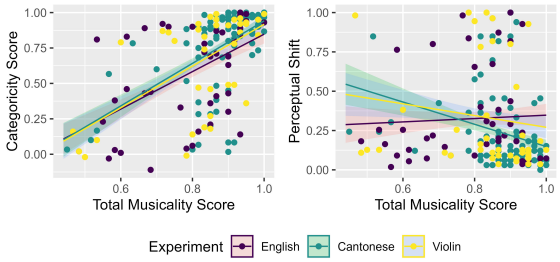


Figure 3: The relationship between musicality and categoricity and shift scores

responses from English and Cantonese speakers. Our results show that listeners of both languages attend to spectral cues in similar directions during pitch perception. Listeners are biased towards thinking tenser sounding stimuli is higher in pitch and vice versa. However, our results show that when the stimuli are modeled after Cantonese prosody, responses from Cantonese speakers do differ from that of English speakers. Moreover, the effect of musicality on shift is more apparent for Cantonese speakers than English speakers, though categoricity patterns similarly. These results suggest that linguistic levels may play a minor role in affecting pitch perception, though cue integration is overall consistent in speech processing by speakers of different languages. We also show that listeners of the violin condition exhibit similar responses compared to the speech responses. This contrasts with results consisting of sawtooth sounds in [6], suggesting that listeners integrate spectral cues in pitch perception similarly in natural sounding stimuli, and not only in speech. Overall results suggest that cue integration in pitch perception spans multiple general and linguistic domains.

5. REFERENCES

[1] E. J. Allen and A. J. Oxenham, "Symmetric interactions and interference between pitch and timbre," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1371–1379, 2014.

[2] P. G. Singh and I. J. Hirsh, "Influence of spectral locus and f 0 changes on the pitch and timbre of complex tones," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2650–2661, 1992.

[3] C. M. Warrier and R. J. Zatorre, "Influence of tonal context and timbral variation on perception of pitch," *Perception & Psychophysics*, vol. 64, no. 2, pp. 198–207, 2002.

[4] A. Cui and J. Kuang, "The effects of musicality and language background on cue integration in pitch perception," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4086–4096, 2019.

[5] J. Kuang and M. Liberman, "The effect of spectral slope on pitch perception," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] ——, "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," *Frontiers in psychology*, vol. 9, p. 2147, 2018.

[7] J. G. Neuhoff and M. K. McBeath, "The doppler illusion: The influence of dynamic intensity change on perceived pitch." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 4, p. 970, 1996.

[8] C. L. Krumhansl and P. Iverson, "Perceptual interactions between musical pitch and timbre." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 3, p. 739, 1992.

[9] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg, "The dependency of timbre on fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2946–2957, 2003.

[10] M. P. Y. Chan and J. Kuang, "Spectral cue integration in pitch perception by Cantonese listeners," 2021, 1st International Conference on Tone and Intonation.

[11] W. Lai and J. Kuang, "The effect of speaker gender on Cantonese tone perception," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 4119–4132, 2020.

[12] J. A. Alexander, P. C. Wong, and A. R. Bradlow, "Lexical tone perception in musicians and non-musicians," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[13] D. Behne, Y. Wang, M. H. Ro, A.-K. Hoff, H. A. Knutsen, and M. Schmidt, "Effects of musical experience on linguistic pitch perception training," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3168–3168, 2006.

[14] T. L. Gottfried and D. Riester, "Relation of pitch glide perception and Mandarin tone identification," *Journal of the Acoustical Society of America*, vol. 108, no. 5, p. 2604, 2000.

[15] C.-Y. Lee and T.-H. Hung, "Identification of Mandarin tones by English-speaking musicians and nonmusicians," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3235–3248, 2008.

[16] A. Houtsma, N. Durlach, and D. Horowitz, "Comparative learning of pitch and loudness identification," *The Journal of the Acoustical Society of America*, vol. 81, no. 1, pp. 129–132, 1987.

[17] R. Wayland, E. Herrera, and E. Kaan, "Effects of musical experience and training on pitch contour perception," *Journal of Phonetics*, vol. 38, no. 4, pp. 654–662, 2010.

[18] I. Peretz, N. Gosselin, Y. Nan, E. Caron-Caplette, S. E. Trehub, and R. Béland, "A novel tool for evaluating children's musical abilities across age and culture," *Frontiers in systems neuroscience*, vol. 7, p. 30, 2013.

The Decline of Tone and Melody Correspondence in Modern Thai Pop Songs?: A Preliminary Analysis

Chawadon Ketkaew, Junyawan Suwannarat

Chiang Mai University, Thailand
chawadon.k@cmu.ac.th, junyawan.suwannarat@cmu.ac.th

ABSTRACT

The purpose of this research is to investigate the correspondence between tone and melody in modern Thai pop songs, as there is a belief that there is no appropriate matching between lexical tones and musical notes in contemporary Thai pop songs. To examine this, the researchers analyzed a corpus of ten modern Thai songs released in 2022 with the highest views on YouTube. The results revealed a statistically significant correspondence between tonal transitions and musical note transitions in these songs, consistent with the findings of previous studies [1,2]. The results also indicated that Thai pop music composition does not necessarily strive to maximize similar transitions, but tends to avoid contrary transitions. This finding refutes the idea that today's music neglects the correspondence between tone and melody, which can negatively impact the intelligibility and aesthetics of the lyrics in a song.

Keywords: Tone-melody correspondence, Thai music, Thai lexical tones

1. INTRODUCTION

One of the challenges of singing in tone languages is how listeners can comprehend the lyrics when the pitch of the words is being used to convey meaning, and the melody of the music may also involve changes in pitch. According to several notable Thai songwriters, they said that if the transitions between lexical tones and musical notes do not match, it can make it more difficult for listeners to understand the lyrics, as the tone and melody may not align with the meaning of the words being sung. This can lead to confusion or misunderstandings and make it harder for listeners to engage with the song. Additionally, the mismatch between tones and melody can affect the aesthetics of native listeners. The combination of mismatched between lexical tone and musical note transitions can create a disjointed or dissonant sound that may be less appealing to listeners, which can negatively impact the enjoyment of the song and potentially make it less successful or popular.

However, there has been significant discussion recently surrounding the matching of lexical tones and musical notes in modern Thai pop songs, in contrast to pop songs of the past. Therefore, this preliminary analysis aims to determine whether the degree of correspondence between lexical tones and musical notes has been reduced in modern Thai pop songs.

2. BACKGROUND

2.1 Overview of Thai pop music

Thai pop music refers to popular music in Thailand. Its early history can be traced back to the 1960s, when Western music and culture began to influence the country. During this time, Thai pop was heavily influenced by Western pop and rock, and artists adopted Western instruments and styles [3]. In the 1980s and 1990s, Thai pop underwent significant transformation and became more diverse and experimental, absorbing elements from various genres, including dance-pop, R&B, hip-hop, and electronic music. Recently, Thai pop has continued to evolve and diversify, incorporating elements from various global music styles and trends. Modern Thai pop encompasses a wide range of genres and styles and remains an essential part of Thai culture and society. However, the influence of Western music, especially in terms of melody, has resulted in some difficulties with the intelligibility of lyrics in modern Thai pop songs.

2.2 Tone-melody correspondence in Thai songs

The Thai language serves as a good case study for the correspondence between tone and melody due to its five tones, as shown in Table 1.

Tone	Tone Value	Examples
MID	[33]	khā: “to be stuck”
LOW	[21]	khà: “galangal”
FALLING	[42]	khâ: “value”
HIGH	[44]	khá: “to trade”
RISING	[35]	khǎ: “legs”

Table1: Thai lexical tones

The proper use of tones is essential in Thai communication, as the meaning of a word can be changed by the tone used. It is intriguing to examine how these tones are reconciled with the music's melody when set to song.

Previous research [4,5] has revealed that there is a high level of correspondence between lexical tones and musical notes in Thai for traditional music, classical music, and recitals, approximately 90%. The correspondence is reduced to 60-70% in contemporary songs [5]. More recent studies [6] have also examined the degree of parallelism in Thai pop songs and found it to be around 80%. These studies have shown that the musical genre can significantly impact the level of correspondence, and that mismatches between tones and melody in Thai pop music often involve the Falling tone. Additionally, more recent studies [1,2] using a larger corpus have found correspondence between lexical tone and musical note transitions, and have suggested that falling and rising tones in Thai behave as if they were a high tone for the purposes of tone-music alignment.

3. METHOD

In this study, the term "modern Thai pop songs" was defined based on [7] and [8] as mainstream popular music designed solely for commercial purposes and characterized by a large scale of consumption. Thus, this preliminary research investigated the correspondence between tonal transitions and musical note transitions in the 10 most-viewed modern Thai pop songs on YouTube in 2022 (the data was collected in December 2022), requiring the songs to be sung in standard Thai. Professional musicians transcribed the melodies of each song using musical notation, and another professional musician double-checked the transcriptions. The lyrics were aligned vertically with their corresponding musical notes, and transitions across musical phrases were excluded, as illustrated in Figure 1.

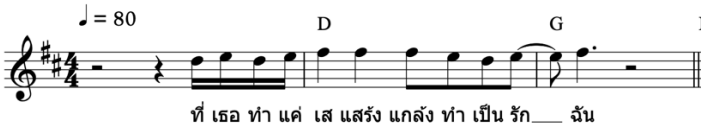


Figure1: Example of a musical phrase consist of musical notes and lyrics

Furthermore, to avoid any potential noise, syllables described as "surface toneless" [9] were excluded. Tonal transitions in this study were defined according to the research by Ketkaew and Pittayaporn [1]. Specifically, tonal transitions were grouped into three categories based on their directions, as summarized in Table 2.

Ascending tonal transition	Descending tonal transition	Level tonal transition
MID→HIGH	MID→LOW	MID→MID
MID→RISING	FALLING→LOW	LOW→LOW
MID→FALLING	FALLING→MID	FALLING→HIGH
LOW→MID	FALLING→FALLING	FALLING→RISING
LOW→FALLING	HIGH→MID	HIGH→FALLING
LOW→HIGH	HIGH→LOW	HIGH→HIGH
LOW→RISING	RISING→LOW	HIGH→RISING
	RISING→MID	RISING→FALLING
		RISING→RISING
		RISING→HIGH

Table 2: Tonal transition categories (following Ketkaew and Pittayaporn [1])

The direction of tonal transitions was then paired with musical note transitions, and the mapping between tonal transitions and musical note transitions was coded in terms of similar, contrary, and oblique motions, according to [10,11], as shown in Figure 2.

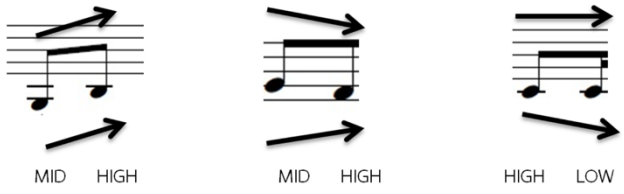


Figure 2: Examples of tonal transition and musical note transition in different motions

A tonal transition that aligned with a musical transition in terms of pitch change direction was coded as a similar motion. It was coded as contrary if the tonal transition and musical note transition moved in opposite directions. A tonal transition and a musical note transition that did not agree in direction but did

not move in opposite directions were coded as oblique. The similar, contrary, and oblique motions were then analyzed using the Chi-squared test and the Holm-Bonferroni method with R version 4.3.0 for statistical analysis.

4. Result

The result revealed a correspondence between lexical note transitions and musical note transitions in 10 modern Thai pop songs, as shown in Table 3.

Tonal Transition	Melodic Transition		
	Up	Down	Same
Up	**340 (19.15%) (similar)**	100 5.63% (contrary)	112 (6.31%) (oblique)
Down	116 (6.5%) (contrary)	**413 (23.27%) (similar)**	119 (6.70%) (oblique)
Same	189 (10.65%) (oblique)	212 (11.94%) (oblique)	**174 (9.8%) (similar)**

Sum of diagonal cell 52.23%

Table 3: Correspondence between tonal transitions and musical note transitions

As presented in Table 3, when the tonal transitions are upward, they are paired with upward melodic transitions (19.15%) more frequently than with downward and level melodic transitions. Similarly, when the tonal transitions are downward, they tend to correspond with the same direction of melodic transitions (23.27%). In the case of level tonal transitions, the pairing with the same direction of melodic transitions occurs less frequently than with upward and downward melodic transitions. For further analysis, all pairs of tonal and melodic transitions were grouped into similar, contrary, and oblique motions, as depicted in Figure 3

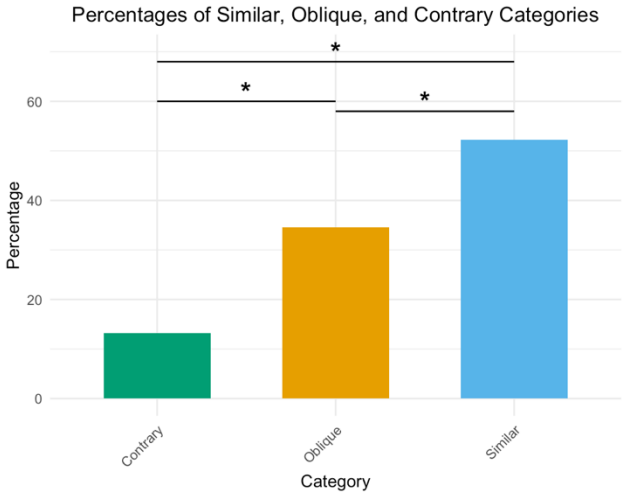


Figure3: Occurrence of similar, contrary and oblique motions in modern Thai pop songs

The analysis of the data demonstrates a significant relationship between tonal transitions and melodic transitions (Chi-squared = 405.83, df = 2, $p < 0.0001$). Further investigation using pairwise comparisons with the Holm-Bonferroni method confirms significant differences between all pairs of motion categories ($p < 0.0001$ for all comparisons). As shown in Figure 3, the tonal transitions and musical note transitions in this corpus consist of 1,775 pairs. Of these, 927 pairs (52.23%) exhibit similar motions. This occurrence is more frequent than that of contrary and oblique motions at a statistically significant level ($p<0.001$). There are 216 pairs (12.17%) of mapping between tonal and musical transitions that are contrary. Interestingly, the number of oblique motions (35.6%) occurs more often than contrary motion at a statistically significant level ($p<0.001$). These results suggest that Thai pop music allows for transitions that are not in contrary motions.

Together, these results provide insights into the underlying structure of Thai pop music and its patterns of tonal and musical transitions. Moreover, these findings contradict the idea that modern Thai music disregards the relationship between tone and melody, which can negatively impact the intelligibility and aesthetics of the lyrics in a song.

5. CONCLUSION AND DISCUSSION

The results of this study showed that there was a correspondence between lexical tones and musical notes in 10 modern Thai pop songs, consistent with previous research [1,2], which found correspondences of 55.3% and 57.6% respectively. This finding refutes the notion that contemporary music disregards the correspondence between tone

and melody, which can negatively impact the intelligibility and aesthetic of lyrics in a song.

In our observation, it is worth considering the potential influence of genre on the correspondence between lexical tones and musical notes in modern Thai pop songs. For example, four songs ทักครับ /tʰákkʰráp/, คำๆเดียว /kʰāmkʰāmdīaw/, เสแสร้ง /sě:sê:ŋ/, and Kiminoto that are a combination of pop music and rap or hip-hop showed correspondences of less than 50%. Additionally, other factors such as prominent beats, note duration, and the position of lyrics within the musical phrase may also be relevant for future analysis.

Furthermore, the prevalence of oblique motions in the analyzed songs indicates that Thai pop music allows for transitions that are not in contrary motions, providing a more nuanced understanding of the musical structure in this genre. This insight can be useful for composers, musicologists, and researchers interested in exploring the intricacies of Thai pop music and its relationship with linguistic features.

In conclusion, this study highlights the importance of examining the correspondence between lexical tones and musical notes in modern Thai pop songs, as it sheds light on the intricate relationship between language and music. Future research can expand on this analysis by investigating a larger dataset, including a diverse range of Thai music genres, and exploring other linguistic and musical factors that may influence the correspondence between lexical tones and musical notes.

6. REFERENCES

[1] Ketkaew, C., & Pittayaporn, P. 2014. Mapping between lexical tones and musical notes in Thai pop songs. In: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 28), Phuket, 160–169.
[2] Ketkaew, C., & Pittayaporn, P. 2015. Do note values affect parallelism between lexical tones and musical notes in Thai pop songs? In: Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow.
[3] Cheng, G., & Charupatanapongse, T. [Online]. The trends and tunes of Thai pop. Kontinentalist. Available: https://kontinentalist.com/stories/how-did-thai-pop-songs-become-so-popular. [Accessed: Apr. 20, 2023].
[4] List, G. 1961. Speech melody and song melody in central Thailand. Ethnomusicology, vol. 5, no. 1, 16–32.
[5] Saurman, M. E. 1999. The agreement of Thai speech tones and melodic pitches. Notes on Anthropology, vol. 3, no. 3, 15–24.
[6] Ho, W. S. V. 2006. The tone-melody interface of popular songs written in tone languages. In: Baroni, M., Addessi, A. R., Caterina, R., & Costa, M. (eds), Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC 2006), 1414–1422.
[7] Wuttipong, N. 2011. The Thai popular music industry: industrial form and musical design. Doctoral dissertation, University of Nottingham, Nottingham, United Kingdom.
[8] Burnett, R. 1996. The Global Jukebox: The International Music Industry. Psychology Press, vol. 18.
[9] Luksaneeyanawin, S. 1983. Intonation in Thai. Thesis, University of Edinburgh, Edinburgh.
[10] Kirby, J. & Ladd, D. R. 2016. Tone-melody correspondence in Vietnamese popular song. 5th International Symposium on Tonal Aspects of Languages (TAL 2016).
[11] Ladd, D. R. & Kirby, J. 2020. Tone–melody matching in tone-language singing. In: The Oxford Handbook of Language Prosody, 675–688.

7. APPENDIX: List of 10 modern Thai pop songs

	Songs	*YouTube view*
1.	พิพิธพันธ์ /pʰíʔpʰíttàpʰān/	169m
2.	คำๆเดียว /kʰāmkʰāmdīaw/	160m
3.	ขวานบิ่น /kʰă:nbìn/	154m
4.	นะหน้าทอง /náʔnâ:tʰɔ̄:ŋ/	153m
5.	ทักครับ /tʰákkʰráp/	116m
6.	เสแสร้ง /sě:sê:ŋ/	98m
7.	วาดไว้ /wâ:wáj/	91m
8.	ลืมแทบไม่ไหว /lɯ̄:mtʰê:pmâjwăj/	97m
9.	โต๊ะริม /tóʔrīm/	71m
10.	Kiminoto	67m

VIP-RADAR: A MODEL FOR THE PHONETIC STUDY OF POETRY READING

Valentina Colonna[1], Antonio Romano[2]

[1]University of Granada (Spain), Department of General Linguistics and Literary Theory,
[2]University of Turin (Italy), Department of Foreign Languages, Literatures and Modern Cultures
valentinacolonna@ugr.es, antonio.romano@unito.it

ABSTRACT

The prosody of a poetry reading is comparable to a musical performance. However, experimental studies on this acoustic data finalized to deepen the music of the page through its vocal production still received little attention. The *Voices of Italian Poets* (VIP) project developed a model of Phonetic study for the poetry reading, created for Italian and yet applied to other idioms, based on the connection with music. The core of this methodology is the VIP-Radar, a graphical representation of the poem's recordings. Among its 20 indexes, 5 depend on musical praxis classification, inspired also by Baroque Rhetoric, in relationship to speech analysis. This is the case of S*ynonymia & Palilogia* (intonation level), *Appoggiato* and *Articolato* (rhythmical level), and *Trattenuto* and *Accelerando* (timing dimension). VIP's interdisciplinary approach aims to be an innovative tool for poetry reading studies.

Keywords: Acoustic Phonetics; Music; Prosody; Poetry; Voices of Italian Poets

1. INTRODUCTION

Poetry reading, as well as a music concert, represents the last step of the composition and the living moment of the page. Performance sheds written art to light enabling the common listening with the audience and the last concretization of its musical form. Through the performance, the immortal form of art is reborn every time in a unique, unrepeatable, and new life, as a single event. In poetry, the page, as a sheet in music, represents the core of a performance: studying poetry reading and, in particular, its prosody, allows, as in music, to know more about the most musical literary form and its author.

Since the 1970s, prosodic research on poetry reading has grown in four areas, focusing on their languages. More in detail, we find the English-speaking school of Byers [1] [2] and Barney [3] [4] [5]; Bernstein [6]; Grobe [7]; MacArthur [8], and MacArthur *et al.* [9]. The German tradition is by Wagner [10], Meyer-Sickendiek *et al.* [11], Schauffler *et al.* [12]. The French approaches are by Puff [13], Rumsey [14], Martin [15], and the Italian school goes from Bertinetto [16] and Schirru [17], to Colonna [18] [19] and Romano [20].

For Italian, the *Voices of Italian Poets* (VIP) project [18], which also includes the first Italian vocal archive for Phonetic research, developed a model of study for the poetry reading, created for Italian and yet applied to other idioms, based on the connection with music[1]. The core of this methodology is the VIP-Radar, a graphical representation of the poem's recordings we go to introduce in this paper[2].

2. THE VIP-RADAR

The VIP-Radar aims to track the main features of poetry reading and represents an instrument for its qualitative analysis. Its creation required a specific terminology, identified and coined, inspired on the whole by music and its language, and functional in particular to the introduction of specific indexes detected for the description of the poetry prosody.

This model made it possible to draw a prosodic overview of the Italian poetry reading style of the last century, enabling also a comparison and quantitative analysis related to 18 authors of the 20th century, grouped into two sections: First (1tv) and Second (2tv) Radio-Television (8 in the 1tv and 10 in the 2tv).

To create the Radar graph, previously we annotated the selected audio materials (.wav, mono, 44100 Hz, 16 bit), following a protocol we go to explain (§2.1.), and then we extracted data, through a script and manually.

2.1. Annotation

The annotation took place using the PRAAT program and is structured on the following 4 levels, which highlight the relationship between text and prosodic realization: VS (*Verso*: "Verse"), EN (*Enunciato*: "Poetic utterance"), CP (*Curva prosodica*: "Prosodic curve"), PR (*Parola ritmica*: "Rhythmic word"). The identification of ENs and PRs occurs on a perceptual level. EN is an independent linguistic act, presenting a terminal boundary and a unitary prosodic signification, which may consist of a single or several internal sections. CP is an inter-pausal section, whereas PRs are tonal-accentual units, with rhythmical cadence falls,

including one or more words, phonetically produced in only one unit with strong stress.

Part of the annotation system also concerns the question of pauses: an evaluation of pauses according to their length was carried out in a general indicative measurement. Specifically, we identified short pauses <pb>, medium pauses <pm>, long pauses <pl>, very long pauses <pll>, and reset pauses <P> (coinciding with the end of the verse). We report in Fig. 1 a sample of audio, including the 4 annotation levels.

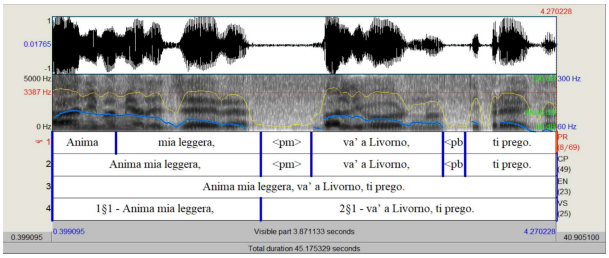


Figure 1: PRAAT window of annotation example.

2.2. The indexes

We present the two main categories included in the VIP-R: parameters related to the prosodic structure and organization; phonetic and stylistic indexes. The organizational indexes are the following:

- *Verse-curve* VS(CP), prosodic curve (CP) coinciding with the verse on the page (VS);
- *Hemi-verse curve* CP(VS), prosodic curve (CP) including a portion of the verse (VS);
- *Inter-verse curves* CP(VS)CP, prosodic curve (CP), placed between two verses (VS) and including a portion of both;
- *Bi-/Poly-verse curve* VS(CP)VS, prosodic curve (CP) entirely including two or more verses (VS).

The stylistic indexes are the following:

- *Pitchspan*: melodic extension (in semitones);
- *Rel_meanpitch*: average relative frequency f_0 (in Hertz);
- *Rel_meanI*: average relative loudness (in dB);
- *Voice Setting Changes*: tonal or register jumps on the total CPs;
- *Speech Rate*: ratio of total phonetic syllables and CPs duration;
- *Accelerando*: perception of acceleration in the Speech Rate, detected at the perceptual level;
- *Trattenuto*: perceived slowing down Speech Rate;
- *Plenus*: ratio of total CP duration and P (pauses);
- *Focus*: focused intonations;
- /Da//: declarative intonation (assertive and poetic declarative, with a terminal boundary on a final medium-low level);
- *Interrupt*: fragmentary pronunciation in CPs and/or in the use of pauses;
- *Appoggiato*: from musical language, stressing mode, marking the CPs by PRs;
- *Articolato*: from musical language, stressing mode marking the ENs by CPs, using pauses;
- *Synonymia & Palilogia Intonation*: rhetorical elements in the CPs intonation through figures of repetition on the same tone – *palilogia* ("palillogy") – or on different tones – *synonymia* ("synonymy"), with the possibility of *variation.* Expression from the Baroque musical lexicon;
- *Enjambment*: prosodic realization of the enjambment by a pause;
- *Plan*: planning of prosodic poetic speech, given by the ratio of EN and VS.

Figg. 2 and 3 are examples of VIP-Radar: we will focus on musical indexes composing it.

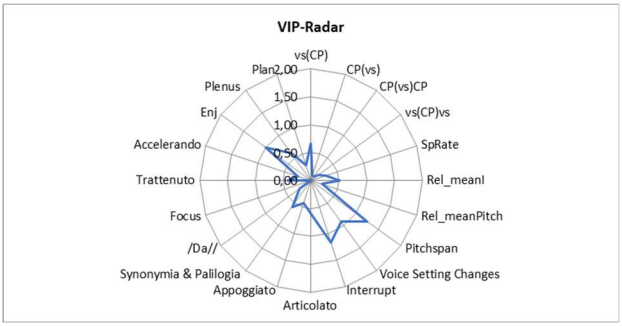


Figure 2: VIP-Radar example (reading of *Sono una creatura* by Giuseppe Ungaretti).

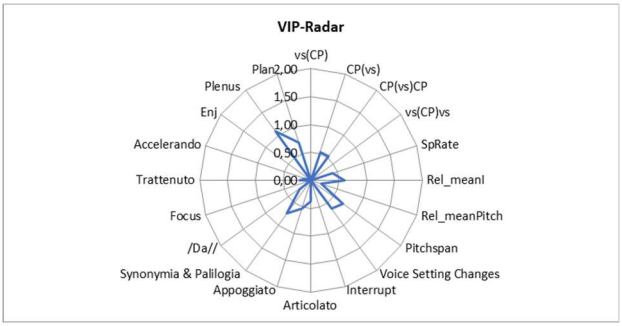


Figure 3: VIP-Radar example (reading of *Alba* by Giorgio Caproni).

2.2. Musical Indexes

Musical indexes are borrowed from musical agogics and praxis and are the following: S*ynonymia & Palilogia* (intonation level), *Appoggiato* and *Articolato* (rhythmical level), and *Trattenuto* and *Accelerando* (timing dimension).

More in detail, starting from the rhythmical level, *Articolato* is connected to the musical concept of *articulation* "articulation" (or more markedly *caesura*, more incisive on a semantic level), inherited

from the Baroque musical language, in which it indicates the divisions into *phrasings*, according to the choices of breath, functional to harmony and melody, and created by performance. It enables the enhancement of accents, melodic trends, and structure, necessary for the transmission of a message in a rhetorical work. In the 18th century, the reference was the literary punctuation (see Couperin [21], to justify the use of commas in the notation of his scores) and flows into a system defined by Mattheson [22] with the idea of *phrasing*. The production of musical articulations is linked to a specific technical and praxis modality and intention. The phrasing scansion allows underlining parts and internal significations, through a breathing division. Inspired by this, in this linguistic work, *Articolato* refers to the utterance scansion in prosodic curves: pauses turn out to be a determining element and accentuate on a macro-structural level major units with minor units, which contribute to making the reading more "breathy". Therefore, a reading with more scansion in CP is more articulated.

Appoggiato, on the other hand, is a musical term that indicates not only a specific accentuation mode, through lengthening the duration and greater intention of weight and intensity of a note, but also a connection with the harmonic structure of the part in which it is. In our study, starting from musical reference, *Appoggiato* scans the microstructure of CPs in PRs, which, like on the score, are marked (with prominence and with tonal evidence). A reading in which the CPs present a greater number of PRs will be considered more *appoggiata*. *Articolato* is calculated as the ratio CP/EN and *Appoggiato* as the ratio PR/CP[3].

The stylistic-rhetoric apparatus is not secondary in intonation. The index of rhetorical figures of intonation, specifically of repetition, is indicated as *Synonymia & Palilogia (Intonation)*, and takes up and extends Colonna [23], which draws on musicological studies of rhetoric (Civra [24] and López Cano [25]). In detail, this parameter is intended to show the frequency of repeated melodic movements on the same tone (*Palilogia*) or different tones (*Synonymia*), sometimes also with *variatio*. The consistent presence of intonation repetitions in poetic speech is a non-negligible element to take into account for this prosodic typology, based on a rhetorical composition as poetry[4]. Figg. 4 and 5 show an example from the same reading, the first in comparison with the second.

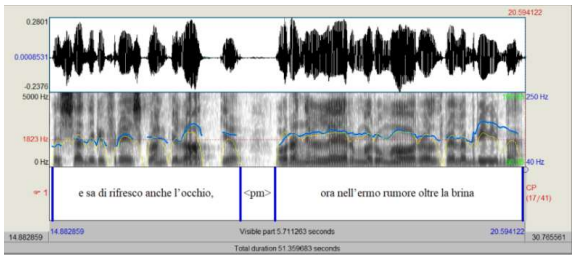


Figure 4: PRAAT window of *synonymia*.

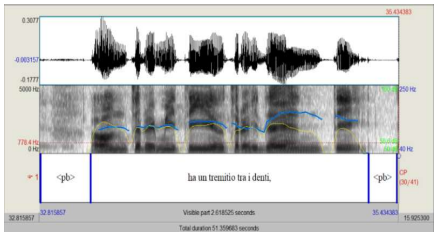


Figure 5: PRAAT window of *palilogia*.

Finally, *Trattenuto* and *Accelerando*, detected on a perceptual and subjective level, indicate the perception of a slower or faster tempo than the general, inspired by the musical term, often used in correspondence of movements or short passages with changes in velocity/agogics intentions.

2.3. Some analysis of *musical indexes*

The VIP corpus analysis, including 32 recordings by 18 authors (quite entirely of free metrics texts), showed that some changing behaviours, together with common aspects, emerged in the passage from the first to the second radio-tv [18].

In particular, a growing *Appoggiato* and decreasing *Articolato* have been detected in the transition over time. The VIP-*Appoggiato*-Boxplots in Fig. 6 show an interquartile *range* higher in the 2tv (together with quite higher mean and median), with symmetric distribution and minimum dispersion in the first quartile, high on the high values of the third quartile (with very high *outlier*). Major variability is detected in 2tv, whereas in the 1tv the distribution is asymmetric, the range is reduced and the variation is limited to a low level.

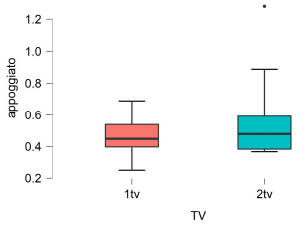


Figure 5: VIP-*Appoggiato*-Boxplots.

Considering the *Articolato* index, the central distribution appears higher in the 1tv, within an interquartile range extended over higher values than

in the 2tv and slightly wider. Both distributions appear asymmetrical, even though greater dispersion in the 2tv, in the first and third quartiles.

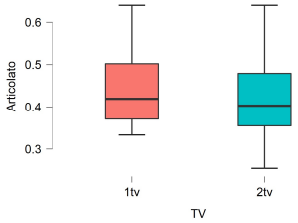


Figure 6: VIP-*Articolato*-Boxplots.

This comparison shows how the way of rhythmically marking the reading has changed over time, moving from a system that predominantly makes use of *Articolato* to one in which an *Appoggiato* style prevails. This data led us to support the hypothesis of a "prosodic accentual verse" (PAV), based on the theory by Fortini [26] of the Italian free verse as an accentual verse. PAV would be marked by PRs, more or less combined with the articulation level. In this way, the melodic-intensive-durative markings, associated with the reduction of silences, would scan the accentual/rhythmic prosodic system, creating dynamism in the unity of reading. Further considerations about text structure and meter are developed in [18].

This path is in line with what happened in music in the transition to an accentual notation, with a gradual loss of a quantitative system, during the transition from the Baroque rhetorical system to the Classical and, later, Romantic system. From a system in which the technique of "articulation" between musical phrases was functional to the marking of the accent itself, we move to a qualitative accent system, with intensive, dynamic, rather than "breathing" markings. The path in the structure of poetic speech would seem to be similar.

Finally, related to *Synonymia & Palilogia*, we found a consistent use of the repetition strategy: in particular, as Fig. 7 shows, 50% of the readings in 1tv made high use of *Synonymia & Palilogia*, as shown by the distribution concentrated on the highest band of the graph. Half of the readings in 2tv, on the other hand, are spread out over a wider range, oriented in a medium-high band: there is a greater heterogeneity in the frequency of repetition figures, more varied, as is also confirmed by the strong dispersion over low values (in the first quartile).

Furthermore, a wider analysis of the VIP-Radars showed that one of the most typical intonations, /Da// (assertive utterance, also coinciding with poetic declarative), represents one of the most repeated.

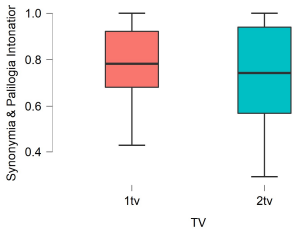


Figure 7: VIP-*Synonymia&Palilogia*-Boxplots.

This parameter makes it possible to note how important the repetitive element of intonation is, confirming the presence of a prosodic rhetorical apparatus of speech that overlaps with that of the written text. Instead of music, poetry intonation rhetoric is detectable only after performance and can be very detached from the page figures. Focused on repetition figures, this index suggests an attempt to return to original (melodic/rhythmic) formulas of poetry speech, connected to memory and orality.

Trattenuto and *Accelerando* gave as a result the presence of isolated high levels in readings, rather than slight records, revealing a reduced perception of speech rate changes or much-connoted styles.

The general VIP-Musical-Radar (Fig. 9), including only the musical indexes, confirms the many global convergences and variety of these. In particular, we find the globally high level of *Palilogia & Synonymia* and a global major level of *Appoggiato*, with isolated highest peaks.

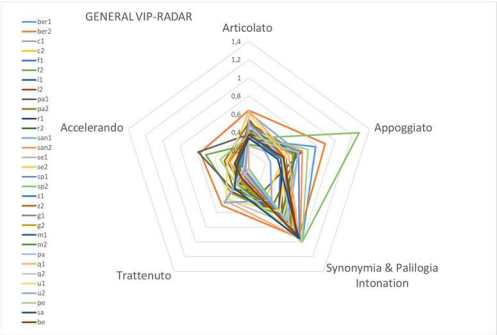


Figure 9: General-VIP-Musical-Radar.

3. CONCLUSION

The VIP-Radar enabled a qualitative image of a poetry reading, based on the strong connection between music and poetry prosody. We found common features, e.g. on a rhythmical level, showing a qualitative change in performance "praxis", as well as common points on melodic and agogics levels. Together with rhetoric, they resulted to be an essential part of a phonetic study of poetry and, in synergy with the objective data analysis, need to be considered for more complete research on the poetry reading, and artistic performance strictly connected to the musical art since its origins.

4. REFERENCES

[1] Byers, P.P. 1977. *The Contribution of Intonation to the Rhythm and Melody of Non-Metrical English Poetry*. Ph.D. Dissertation, University of Wisconsin–Milwaukee.

[2] Byers, P.P. 1979. A formula for poetic intonation. *Poetics*, 8 (4), 367-380.

[3] Barney, Th. H. 1998. *Style in performance: The prosody of poetic recitation*. Ph.D. Dissertation, University of Lancaster.

[4] Barney, Th. H. 1999. Readers as text processors and performers: a new formula for poetic intonation, *Discourse Processes*, 28 (2), 155-167.

[5] Barney, Th. H. (1995). Phonetics and the Empirical Study of Poetry. In: Kreuz Roger, J., MacNealy, M. S. (eds.), *Empirical Approaches to Literature and Aesthetics*. Norwood: Ablex.

[6] Bernstein, Ch. (ed.) 1998. *Close Listening: Poetry and the Performed Word*. New York: Oxford University Press. Then in: *My Way: Speeches and Poems*. Chicago: University of Chicago Press, 1999.

[7] Grobe, Chr. 2016. On Book: The Performance of Reading. *New Literary History*, 47(4), 567-589.

[8] MacArthur. 2016. Monotony, the Churches of Poetry Reading, and Sound Studies. *Monotony, the Churches of Poetry Reading, and Sound Studies PMLA*, 131(1), 38-63.

[9] MacArthur, M., Zellou G., Miller, L.M. 2018. Beyond Poet Voice: Sampling the (Non-) Performance Styles of 100 American Poets. *Journal of Cultural Analytics*. DOI: 10.31235/osf.io/5vazx.

[10] Wagner, P. 2012. Meter Specific Timing and Prominence in German Poetry and Prose. In: Niebuhr, O. (ed.). *Understanding Prosodies. The Role of Context, Function and Communication.* Berlin/New York: De Gruyter.

[11] Meyer-Sickendiek, B., Hussein, H., Baumann, T. 2017. Rhythmicalizer: Data Analysis for the Identification of Rhythmic Patterns in Readout Poetry. In: Eibl, M., Gaedke, M. (eds.). *Informatik 2017. Lecture notes in Informatics (LNI), Gesellschaft für Informatik*, 2189-2200.

[12] Schauffler, N., Schubö, F., Bernhart, T., Eschenbach, G., Koch, J., Richter, S., Viehhauser, G., Vu, Th., Wesemann, L., Kuhn, J. 2022. Prosodic realisation of enjambment in recitations of German poetry, *Proc. Speech Prosody 2022*, 530-534, DOI: 10.21437/SpeechProsody.2022-108.

[13] Puff, J. Fr. (ed.) 2015. *Dire la poésie?* Nantes: Cécile Defaut.

[14] Rumsey, L. 2015. Da-DA-da-da-da: Intonation and Poetic Form. *Thinking Verse*, 5, 15-49.

[15] Martin, Ph. 2019. Prosodic structure and poetic diction. In : Colonna, V., Romano A. (eds.) *RiCognizioni*, 6(11), 65-74.

[16] Bertinetto, P. M. 1973. *Ritmo e modelli ritmici: analisi computazionale delle funzioni periodiche nella versificazione dantesca.* Torino: Rosenberg & Sellier.

[17] Schirru, G. 2002. Correlati acustici del ritmo linguistico. In: Buffoni, F. (ed.) *Ritmologia.* Milano: Marcos Y Marcos, 267-306.

[18] Colonna, V. 2022. *Voices of Italian Poets. Storia e analisi fonetica della lettura della poesia italiana del Novecento*. Alessandria: Dell'Orso.

[19] Colonna, V. 2021. *Voices of Italian Poets*. *Analisi fonetica e storia della lettura della poesia italiana dagli anni Sessanta a oggi*. PhD Dissertation, University of Genoa-University of Turin.

[20] Romano, A. 2019. Voci e letture poetiche da Wordsworth a Heaney". In : Colonna, V., Romano, A. (eds.). «La musica della poesia», *RiCognizioni*, 6 (11), 97-110.

[21] Couperin, F. 1722. *Pièces de clavecin (III)*. Paris: Durand.

[22] Mattheson, J. 1737. *Kern melodischer Wissenschaft*. Hamburg: Herold.

[23] Colonna, V. 2017. *Prosodie del «Congedo». Analisi fonetica comparativa di dodici letture.* Tesi magistrale, Università degli Studi di Torino, A.A. 2016-2017. Unpublished.

[24] Civra, F. 1991. *Musica poetica. Introduzione alla retorica musicale*. Torino: UTET Università.

[25] López Cano, R. 2000. *Música y retórica en el Barroco.* Barcelona: Amalgama.

[26] Fortini, F. 1957. Metrica e libertà. *Ragionamenti*, III, 10-12, 267-74. Now in: Fortini, F. 2003. *Saggi ed epigrammi* (ed.). Milano: Mondadori, 783-808.

[1] This methodology is currently used in the VSP (*Voices of Spanish Poets*) project, funded by the European Union, within the framework of a Marie Skłodowska-Curie Action (MSCA Postdoctoral Fellowship) at the University of Granada, lead by Valentina Colonna.

[2] Further levels of analysis, which we will not consider in this study, compose the methodology (e.g. VIP-EnjambmentLine; VIP-CP-P-Histogram, ecc.)

[3] The index of *Articolato* and *Appoggiato* are reported on the radar on a 0-5 scale. Values <0.4 are considered low, ≥0.4, <0.6 medium, >0.6 high; >1 very high.

[4] These intonations are identified perceptually, together with the visualization of f_0 curves on the spectrogram.

THE INFLUENCE OF MUSICAL ABILITIES ON THE PROCESSING OF L2 FOCUS PROSODY: AN EYE-TRACKING STUDY

Nelleke Jansen[1,2], Hanneke Loerts[1], Eleanor E. Harding[1,2], Deniz Başkent[1,2], Wander Lowie[1]

[1]University of Groningen; [2]University Medical Center Groningen
n.g.jansen@rug.nl

ABSTRACT

The perception of speech prosody in a second language (L2) remains challenging for proficient L2 users. Previous eye-tracking evidence indicates that Dutch listeners show difficulty in the processing of pitch accents signalling contrastive focus in English, whereas native English listeners use this cue in perception to anticipate upcoming information. We investigated whether musical abilities influenced the processing of contrastive focus accents by 40 Dutch adult L2 English users. In a visual-world eye-tracking paradigm, participants listened to sentences with the focus particle *only* while viewing images of the objects and characters mentioned. We measured participants' anticipatory fixations on the image showing the alternative of the contrast. Participants also completed a music perception test. Initial analyses indicate that individuals with higher music perception scores show more anticipatory fixations in L2 listening. This suggests that having stronger perceptual resources underlying both music and speech perception may benefit the processing of L2 focus prosody.

Keywords: L2, prosody, focus, musical abilities, visual world paradigm

1. INTRODUCTION

1.1. Contrastive focus accents in Dutch and English

Focus is an aspect of information structure which indicates new, contrastive, or otherwise important information in an utterance. In many languages, including Dutch and English, focus is realized by means of a pitch accent on the focal word [1, 2]. Users of English as a second language (L2) need to process this prosodic cue during sentence comprehension, which is challenging even to advanced L2 users [3, 4]. In this study, we investigated the interpretation of L2 contrastive narrow focus accents, which can play a role in the semantic interpretation of a sentence. In sentences that contain the focus particle *only*, *only* associates with a word or word group that carries focus [5]. Because focus is signalled by a pitch accent, this type of sentence needs an accent to resolve ambiguity. For example, the sentence *I only gave a spoon to the girl* has several possible readings indicated by different patterns of accentuation (focus indicated by square brackets, accent indicated by capitalisation):

(1) *I only gave [a SPOON]$_F$ to the girl*
'I gave a spoon to the girl and I didn't give anything else to the girl'

(2) *I only gave a spoon [to the GIRL]$_F$*
'I gave a spoon to the girl and I didn't give a spoon to anyone else'

While both Dutch and English use pitch accents to signal focus, and focus accents have similar phonetic realisations in Dutch and English [6], these languages differ somewhat in the way *only* is used in sentences. In Dutch, *alleen* ('only') preferably directly precedes the focal word (group) [7], which means that the word order of a sentence with *alleen* usually changes with a different focus. In English, *only* preferably precedes the verb, rather than the focal word [8, 9]. This difference is illustrated by comparing Dutch example (3) to (2).

(3) *Ik heb alleen [aan het MEISJE]$_F$ een lepel gegeven*
I have only to the GIRL a spoon given

In Dutch sentences with *alleen*, word order is thus an important focus cue besides prosody, whereas in English sentences with *only*, prosody has a higher relative importance as a focus cue.

The online processing of focus requires the integration of prosodic cues into syntactic, semantic, and discourse domains [10]. The integration of prosodic information happens fast in L1 English processing of sentences with *only* [4, 11] and L1 Dutch processing of sentences with *alleen* [12], where L1 listeners are able to use focus prosody to anticipate upcoming information [4,

12]. However, the integration of information from different linguistic (e.g., syntax) and extra-linguistic domains (e.g., discourse) is notoriously difficult for L2 users, forming an obstacle in the L2 attainment of proficient users [13]. Furthermore, the differences between languages regarding focus cues may hinder the processing of focus in a second language. For L1 Dutch users of English as an L2, difficulties could especially arise when the focus particle *only* does not directly precede the word(group) in focus, which means that listeners have to rely on accentuation alone to get to the correct interpretation. A previous eye-tracking study by Ge et al. [4] has shown that proficient L1 Dutch L2 English listeners do not show the anticipatory eye-movements demonstrated by L1 English listeners. They found that when listening to a sentence like *The dinosaur is only CARRYING the bucket, not throwing the bucket*, L1 English listeners showed anticipatory fixations on picture of the focus alternative (the dinosaur throwing the bucket) upon hearing the word *not*. However, Dutch L2 English users shifted their gaze later, upon hearing the verb *throwing*. These findings suggest that despite their high proficiency and language similarities, L1 Dutch advanced L2 English users have difficulty with the integration of prosodic cues to update their interpretation of sentences with contrastive focus in online processing.

1.2. Musical abilities and prosody perception

Music training and music perception abilities have been related to a more refined perception of prosody in foreign languages [14]. It has been proposed that such a connection may be due to neural plasticity in overlapping neural networks for music and speech [15, 16]. However, previous research has mainly considered the perception of unfamiliar foreign languages, rather than perception by proficient L2 users [14]. Following previous research supporting a link between musical abilities and speech prosody perception, we expected musical abilities to benefit the integration of prosody in determining the location of the focus when processing L2 speech. Findings from a previous ERP study [17] on the processing of L2 focus prosody suggested that refined music perception abilities were related to a reduced processing effort and better online adjustment of expectations for focus accentuation.

In the current study, we used eye-tracking to test if Dutch adults' musical abilities influenced their online focus interpretations in English sentences with *only*. We hypothesised that listeners with higher music perception scores would show more anticipatory eye-movements than listeners with lower music perception scores, reflecting more target-like processing of focus prosody in the L2.

2. METHOD

2.1. Participants

So far, we have included 40 L1 Dutch adults (ages 19-37) who were advanced L2 English users. Participants reported normal hearing and (corrected-to-)normal eyesight. We used a C-test [18] as a general measure of English proficiency. Participants had an average score of 85.90% (SD = 8.42), which indicates high proficiency levels (comparable to CEFR levels B2 to C2).

2.2. Materials and design

2.2.1. Musical abilities

We used the Short Profile of Music Perception Skills (Short-PROMS) test battery [19] to measure music perception abilities. The Short-PROMS consists of the subtests Melody, Rhythm, Rhythm-to-Melody, Tuning, Accent (i.e. intensity), Timbre, Tempo, and Pitch. In each trial, the participant listens to two musical excerpts and indicates whether they are the same or different. The PROMS gives a score for each subtest, which we used in the current analysis. Because contrastive focus is signalled by a pitch accent, we expected Pitch scores to be most strongly related to the processing of focus in speech.

2.2.2. Visual world paradigm

We used the visual world paradigm to investigate sentence processing. Participants listened to spoken sentences while looking at pictures on a screen. Auditory stimuli consisted of English sentences with the focus particle *only* in two conditions: direct object focus (DO condition) and indirect object focus (IO condition). An example of a stimulus with DO focus is given in (4) and Fig. 1.

(4) *I only gave a SPOON to the girl. [pause] I didn't give a FORK to the girl.*

Before the target sentences were presented, the speaker introduced two possible indirect objects (a girl and a boy) and two possible direct objects (spoons and forks), giving potential focus alternatives. In the first clause of the target stimulus, the focus accent indicated the location of focus, which disambiguated the meaning of the sentence. The second clause gave a lexical disambiguation, which could be anticipated based on the focus

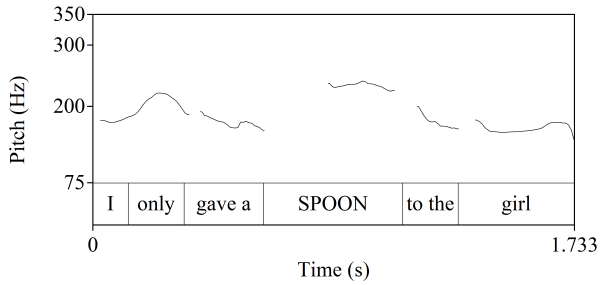


Figure 1: Pitch contour of the first clause in a DO focus trial.

accent. While listening to the introductory stories and target sentences, participants viewed pictures [20, 21] on a screen in a 2x2 grid showing the potential direct objects and indirect objects in four combinations (Fig. 2).

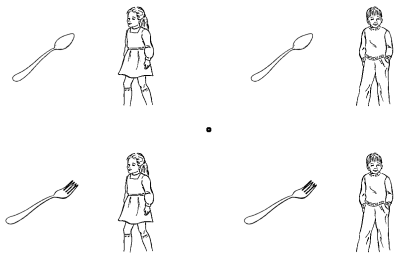


Figure 2: Example of a visual display.

The experiment consisted of 48 experimental trials and 48 fillers. The fillers contained different sentence constructions. Stimuli were recorded by a female L1 speaker of American English.

2.3. Procedure

Participants were tested in the lab. They first completed the C-test using pen and paper. Next, they took part in the eye-tracking task. Trials were presented in eight blocks. Participants listened to the stimuli using headphones. They were instructed to look at the pictures while listening to the speaker, without any active task. For 25% of the trials, participants had to answer a question about what happened in the story, to ensure they kept paying attention. Fixations were recorded with the SR Research EyeLink Portable Duo eye tracker. Finally, participants completed the Short-PROMS test.

2.4. Analysis

We measured proportions of fixation time (%FT) on four interest areas: target 1, the target given in the first clause; target 2, which is the focus alternative (given in the second clause); the competitor, which would be the focus alternative in the other focus condition; and a distractor. Taking example (4), participants were expected to look at the girl with the spoon (target 1) upon hearing the first clause. We assessed to what extent participants anticipated the alternative of the focus by fixating on the girl with the fork (target 2) before hearing the disambiguating words *FORK to the girl*. Higher %FT on target 2 compared to the competitor indicated correct interpretations of the focus accent. For visualisation, stimuli were divided into small time windows (see Fig. 3), each starting 200 ms after the onset of the relevant word, taking into account the timing of saccades in response to language [22]. To answer our research question, we measured %FT in a larger time window where anticipation may occur, i.e. *[I didn't give a]*. Length in ms differed between items.

We analysed the %FT on target 2 and the competitor with a zero-inflated beta regression (ZIBR) model, using the package glmmTMB [23] in R [24]. Beta regression is suitable for the analysis of proportion data which follows a beta distribution, and a zero-inflated model was required because %FT was often zero for both target 2 and competitor, because participants kept looking at target 1. We included a random-effects structure of `Item` nested within `Participant`. The influence of proficiency and musical abilities on anticipatory fixations was assessed by modelling interactions between the dummy variable `Competitor` (0 = target 2, 1 = competitor) and `Proficiency`, and between `Competitor` and the Short-PROMS subtest scores. We included the subtests that significantly improved the model in model comparisons, namely `Pitch` and `Tempo`. We expected higher proficiency and musical abilities would lead to higher %FT (and fewer %FTs of zero) on the target, and lower %FT (and more zeros) on the competitor.

3. RESULTS

Averaged %FT on the interest areas are given in Fig. 3. Fixations on target 2 start to deviate from the competitor before participants have heard the direct object in the second clause. This suggests that to some extent, participants anticipate the focus alternative based on the pitch accent in the first clause, while %FT on target 1 remains high.

Model outcomes are given in Table 1. The beta model, which models the %FT, shows a negative main effect of `Competitor`. This indicates %FT is higher for target 2 than for the competitor.

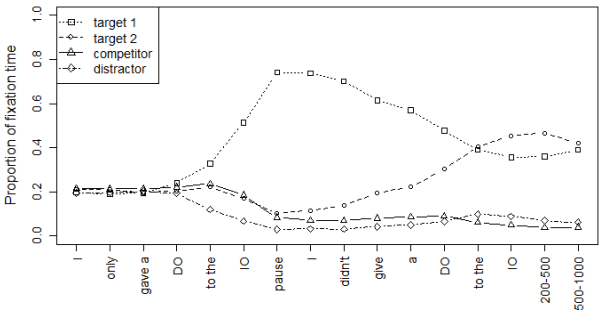


Figure 3: Averaged %FT (y-axis) on the four interest areas per time window (x-axis).

The zero-inflation model estimates the probability to get extra %FTs of zero. The main effect of `Competitor` indicates this probability is higher for the competitor than for target 2. The interaction `Competitor*Proficiency` indicates proficiency is related to fewer zeros on target 2 (negative slope on the probability of zeros), and more zeros on the competitor (positive slope). Similarly, the interaction `Competitor*Pitch` shows musical pitch perception is related to fewer zeros on target 2 and more on the competitor, in line with our expectations. The interaction `Competitor*Tempo` shows an effect of musical tempo perception in the opposite direction. However, while the effect of `Tempo` differs between target and competitor, the positive slope for target 2 is not significant, and `Tempo` does not affect the probability of zeros on the competitor (flat slope).

Parameter	Beta model Estimate (SE)	Zero-infl. model Estimate (SE)
Intercept	0.320 (0.087)*	1.189 (0.158)*
Competitor	-0.194 (0.094)*	0.838 (0.089)*
Proficiency	0.016 (0.010)	-0.027 (0.019)
Pitch	0.077 (0.055)	-0.351 (0.111)*
Tempo	-0.056 (0.076)	0.228 (0.136)
Competitor:Proficiency	-0.016 (0.010)	0.042 (0.010)*
Competitor:Pitch	-0.104 (0.065)	0.486 (0.062)*
Competitor:Tempo	0.030 (0.086)	-0.232 (0.082)*

* significant at α-level of 0.05

Table 1: ZIBR model (coefficients are log-odds).

4. DISCUSSION

In this eye-tracking study, we investigated the influence of musical abilities on the processing of contrastive focus prosody in sentences with *only* by L1 Dutch proficient L2 English users. We examined to what extent participants showed anticipatory fixations on the alternative of the contrast (i.e., the second target), versus the competitor (reflecting a different contrast). Anticipatory fixations indicated the participant had come to the correct focus interpretation by processing the contrastive pitch accent in the first clause of the stimulus.

Results showed that participants (to some degree) anticipated the correct focus alternative, as fixations proportions were higher on the second target than on the competitor. However, participants also continued looking at the target of the first clause, in line with previous findings [4]. Continued fixations on the first target may partly reflect participants' attention towards *what happened*, rather than the alternative situation (*what did not happen*).

Based on previous research indicating that individuals with stronger music perception abilities showed a more fine-grained prosody perception [14], we hypothesised that these individuals would show more anticipation based on the contrastive pitch accent. Although results showed no effect on the fixation time proportions, musical pitch perception ability increased the probability to fixate on the target (i.e., lower probability of no fixation on the target) and decreased the probability to fixate on the competitor. This increased anticipation indicates that musical abilities facilitate the integration of prosodic information in language processing, which may be challenging for L2 users [3, 13]. Our study thus extends previous findings on the perceptual advantage related to musical abilities in foreign language perception to an advantage in prosody-to-meaning mapping in proficient L2 perception.

A beneficial influence of musical abilities on the processing of L2 prosody may be explained on the basis of transfer between music and speech in overlapping neural networks [15, 16]. Having strong perceptual resources underlying both music and speech perception may thus also benefit the processing of L2 focus prosody. Such strong perceptual resources could be the result of genetic predispositions, music training, or both [25]. Because we currently only used music perception scores as our measure of musical abilities, further analyses are planned to elucidate the potential role of participants' music training.

English proficiency also had a positive effect on anticipation. This indicates participants with higher (written) proficiency show more efficient L2 prosodic processing, perhaps because they can use processing resources that would otherwise be dedicated to lexical or syntactic processing.

Finally, this study will be extended to include a control group of L1 English participants. This will allow us to compare performance of the L2 listeners to L1 listeners, and investigate whether musical abilities also play a role in L1 processing.

5. ACKNOWLEDGEMENTS

We thank Victor Bernal from the Data Science team of the Center for Information Technology at the University of Groningen for advice on the statistical analysis.

6. REFERENCES

[1] C. Gussenhoven, *On the Grammar and Semantics of Sentence Accents*. Foris Publications, 1984.

[2] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 1996.

[3] E. Akker and A. Cutler, "Prosodic cues to semantic structure in native and nonnative listening," *Biling.: Lang. Cogn.*, vol. 6, no. 2, pp. 81–96, 2003.

[4] H. Ge, I. Mulders, X. Kang, A. Chen, and V. Yip, "Processing focus in native and non-native speakers of English: An eye-tracking study in the visual world paradigm," *Appl. Psycholinguist.*, vol. 42, no. 4, pp. 1057–1088, 2021.

[5] M. Rooth, "A theory of focus interpretation," *Nat. Lang. Semant.*, vol. 1, no. 1, pp. 75–116, 1992.

[6] A. M. Sluijter and V. J. Van Heuven, "Acoustic correlates of linguistic stress and accent in Dutch and American english," in *Proc. 4th International Conference on Spoken Language Processing*. IEEE, 1996, pp. 630–633.

[7] A. Foolen, R. van Gerrevink, L. Hogeweg, and P. Prawiro-Atmodjo, "The placement of focus particles in dutch," *Linguistics in the Netherlands*, vol. 26, no. 1, pp. 51–63, 2009.

[8] G. Bouma, P. Hendriks, and J. Hoeksema, "Focus particles inside prepositional phrases: A comparison of Dutch, English, and German," *J. Comp. Ger. Linguist.*, vol. 10, no. 1, pp. 1–24, 2007.

[9] M. M. Vann, "Is it only about only? A study on focus particle placement acceptability in L2 English learners and English monolinguals," *Annali di Ca' Foscari. Serie occidentale*, vol. 56, 2022.

[10] K. Lambrecht, *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, 1994.

[11] S. Gennari, L. Meroni, and S. Crain, "Rapid relief of stress in dealing with ambiguity," in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, J. C. Trueswell, M. K. Tanenhaus, and M. K. Tanenhaus, Eds. MIT Press, 2005, pp. 245–259.

[12] I. Mulders and K. Szendrői, "Early association of prosodic focus with alleen 'only': Evidence from eye movements in the visual-world paradigm," *Front. Psychol.*, vol. 7, p. 150, 2016.

[13] A. Sorace, "Pinning down the concept of "interface" in bilingualism," *Linguist. Approaches Biling.*, vol. 1, no. 1, pp. 1–33, 2011.

[14] N. Jansen, E. Harding, H. Loerts, D. Başkent, and W. Lowie, "The relation between musical ability and sentence-level intonation perception: A meta-analysis comparing L1 and non-native listening," in *Proc. Speech Prosody*, 2022, pp. 713–717.

[15] A. D. Patel, "Why would musical training benefit the neural encoding of speech? The OPERA hypothesis," *Front. Psychol.*, vol. 2, p. 142, 2011.

[16] A. Tierney and N. Kraus, "Auditory-motor entrainment and phonological skills: precise auditory timing hypothesis (PATH)," *Front. Hum. Neurosci.*, vol. 8, p. 949, 2014.

[17] N. Jansen, R. Ganga, M. Struiksma, and A. Chen, "Musicality influences the processing of accentuation in sentences with *only* by L2 English learners," in *Hanyang International Symposium on Phonetics and Cognitive Sciences of Language*, 2019, pp. 53–54.

[18] M. Daller, A. Müller, and Y. Wang-Taylor, "The C-test as predictor of the academic success of international students," *Int. J. Biling. Educ. Biling.*, vol. 24, no. 10, pp. 1502–1511, 2021.

[19] M. Zentner and H. Strauss, "Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS," *Ann. N. Y. Acad. Sci.*, vol. 1400, no. 1, pp. 33–45, 2017.

[20] E. Bates, S. D'Amico, T. Jacobsen, A. Székely, E. Andonova, A. Devescovi, D. Herron, C. Ching Lu, T. Pechmann, C. Pléh *et al.*, "Timed picture naming in seven languages," *Psychon. Bull. Rev.*, vol. 10, no. 2, pp. 344–380, 2003.

[21] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity." *J. Exp. Psychol.: Learn. Mem. Cogn.*, vol. 6, no. 2, p. 174, 1980.

[22] G. T. Altmann and Y. Kamide, "Now you see it, now you don't: Mediating the mapping between language and the visual world," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, J. Henderson and F. Ferreira, Eds. Psychology Press, 2004, pp. 347–386.

[23] M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *The R Journal*, vol. 9, no. 2, pp. 378–400, 2017.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[25] E. G. Schellenberg, "Music training and speech perception: A gene–environment interaction," *Ann. N. Y. Acad. Sci.*, vol. 1337, no. 1, pp. 170–177, 2015.

DURATIONAL ADJUSTMENT IN CANTONESE FOCUS MARKING OF NATIVE AND NON-NATIVE SPEAKERS

Yike Yang

Hong Kong Shue Yan University
yyang@hksyu.edu

ABSTRACT

The speech production of non-native speakers usually differs from that of native speakers, but very few investigations on the prosody of non-native speech have been conducted. Although duration has been regarded as an important acoustic cue for focus marking in Cantonese, precisely how native Cantonese speakers manipulate duration in different types of syllable structures remains to be explored. This study aims to fill these research gaps by examining how native and non-native speakers of Cantonese mark prosodic focus via durational adjustment. The native speakers demonstrated an asymmetrical effect of focal lengthening in different types of syllable structures in Cantonese, and the word-final nasals were expanded substantially. The non-native speakers revealed different patterns, which was attributed to their failure to fully acquire Cantonese focus marking. Discussions about the amount of focal lengthening in Cantonese and the acquisition and learnability of second language (L2) prosody are provided.

Keywords: second language, speech production, speech prosody, focus, duration

1. INTRODUCTION

1.1. Durational adjustment in focus marking

Focus is a linguistic device that introduces new information and highlights the portion of a sentence that contains this information [1]. Cross-linguistically, there are various linguistic means of focus marking [2], and there are even different ways to manipulate acoustic cues to differentiate the new information from the old information within languages that make use of prosody to mark focus [2]. More specifically, while cues related to fundamental frequency (F0), such as mean F0 and maximum F0, have been found to correlate with focus in Mandarin [3], duration is an important acoustic cue that marks focus in Cantonese [4]. However, although the durational adjustment of focus marking has been examined previously in Mandarin [5] and Yoloxochitl Mixtec [6], a thorough investigation of the durational adjustment of focus marking in Cantonese is still lacking.

The effect of focus-induced lengthening is not always symmetrical. In stressed languages such as English and Swedish, the stressed syllables often exhibit greater lengthening within the focal words than do their unstressed counterparts [7], [8]. Moreover, whether the focal lengthening is symmetrical for the segments within a syllable is a controversial issue [5], [8].

1.2. Non-native speech

Non-native speakers usually differ from native speakers in various aspects (e.g., word choice and cohesion), and there has been reports of a perceived foreign accent in non-native speakers' speech production [9]. Although some domains of linguistic knowledge have been proven to be acquirable at the end state of second language (L2) acquisition (such as syntax [10] and lexicon [11]), it has been argued that the attainment of native-like pronunciation is unlikely for late L2 learners [12]. The foreign accent in L2 ultimate attainment has been well documented, but consensus regarding which segmental and suprasegmental features may contribute to a non-native accent has not been reached. Some attempts have been made to explore the effects of segments on accent, but very few investigations have concerned the suprasegmental level [13].

1.3. This study

As discussed above, although duration has been found to be a salient acoustic cue in Cantonese focus marking, precisely how native speakers of Cantonese manipulate duration in different types of syllable structures remains to be explored. While many studies have focused on the production of L2 segments [14], there are few investigations of non-native speakers' marking of focus prosody [15]. To fill these gaps, this study examines the manipulation of duration in Cantonese focus marking by native and non-native speakers.

Two research questions will be addressed in this paper:

1) How do native speakers of Cantonese mark focus via durational adjustment?

2) Can non-native speakers of Cantonese acquire the durational adjustment patterns in Cantonese focus marking?

2. METHODS

2.1. Speakers

Forty-three participants were recruited to participate in a production experiment at a speech laboratory of a local university. Twenty-one of the participants were native speakers of Hong Kong Cantonese (10 females and 11 males, aged 20.78 ± 2.56), who were born and raised in Hong Kong; the other 22 were native speakers of Mandarin (19 females and 3 males, aged 30.14 ± 4.30), who were born in Northern China and had been exposed to Cantonese since their arrival in Hong Kong. The Mandarin speakers had arrived in Hong Kong after puberty (average age: 22.73 ± 4.21) and the average length of residence was 7.41 ± 3.11 years. The participants completed a language background questionnaire [16] prior to the recording session, according to which the non-native speakers were all fluent speakers of Cantonese. None of the participants had any history of speaking, hearing or language difficulties.

2.2. Materials and procedures

Two types of syllable structures were considered, namely consonant-vowel combinations (CV) and consonant-vowel-consonant combinations (CVC). Two monosyllabic verbs with the high-level Tone 1 were chosen for each syllable type, and the verbs were then embedded in meaningful Cantonese sentences, as shown in Sentence 1 below. The target syllables are presented in Table 1. As Cantonese is a subject-verb-object (SVO) language, the verb normally occurs in the sentence-medial position.

(1) *go2 go3 si1gei1 zaa1 cyun1baa1*
that CL driver drive residential_bus
'That driver drives the residential route.'

Table 1: Target syllables.

Initial consonant	Vowel	Final consonant	Meaning
/ ts/	/a/	N/A	to drive
/f/	/ei/	N/A	to fly
/l/	/ɪ/	/ŋ/	to carry
/ts/	/ɔ/	/ŋ/	to assemble

Different types of focus were elicited via precursor questions that were asked by the experimenter and the participants were instructed to answer the questions as naturally as possible using the target sentences. The question and answer pairs were presented randomly on a computer screen in E-Prime 2.0 [17] and the dialogues were recorded at the sampling rate of 44,100 Hz in Audacity [18] on another computer. As the broad focus and the verb focus were of particular relevance to the current study, only the results for these two focus types are reported in this paper.

2.3. Data analysis

The author, who is a trained phonetician, performed the segmentation of the target verbs manually at both the segmental and syllabic levels using Praat [19]. Following the segmentation, the duration of each segment and syllable was extracted automatically via a Praat script. The duration was measured in milliseconds (ms).

The duration values at the segmental and syllabic levels were analysed via linear mixed-effects modelling using the 'lme4' package [20] in R [21], [22]; the *focus type* was the fixed effect, and the *speaker* and *repetition* were included as the random effects. The broad focus type was the baseline condition. The native and non-native data were first analysed separately and were then compared. The figures were plotted using the 'ggplot2' package in R [23].

3. RESULTS

3.1. Durational adjustment of native speakers

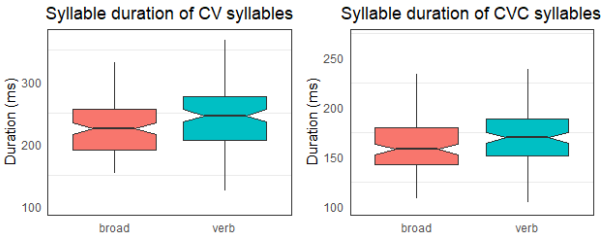


Figure 1: Syllable duration of native speakers.

We first examined how native speakers of Cantonese made durational adjustments for different syllable types at the syllabic level. Linear mixed-effects models suggested that the native speakers lengthened the syllable duration for both syllable types, but the lengthening effect was more robust for CV syllables (28.656 ± 4.908 ms; $\chi^2(1) = 31.918$, $p < .001$) than it was for CVC syllables (14.662 ± 3.115 ms; $\chi^2(1) = 21.229$, $p < .001$). The mean syllable duration of the syllables produced by native speakers is plotted in Figure 1.

Separate models for the segment duration of CV and CVC syllables were then fitted. For the CV syllables, there were main effects of focus ($\chi^2(1) = 7.733$, $p = .005$) and segment ($\chi^2(1) = 569.570$, $p < .001$) as well as a marginal interaction between focus and segment ($\chi^2(1) = 3.746$, $p = .053$) on the duration values. Post-hoc tests revealed a minor

lengthening of the consonant duration (8.94 ± 3.94, p = .107) and a significant lengthening of the vowel duration (19.72 ± 3.94, p < .001) under focus. For the CVC syllables, there were main effects of focus ($\chi^2(1)$ = 6.526, p = .011) and segment ($\chi^2(2)$ = 235.520, p < .001) on duration values, but the interaction between focus and segment did not reach significance ($\chi^2(1)$ = 2.027, p = .363). Post-hoc tests showed no differences for the first consonant (p = .964) and the vowel (p = .614) in the two focus types, but the second consonant (syllable-final nasal sound) was slightly lengthened under focus (11.87 ± 4.23 ms, p = .057). The mean segment duration of the syllables produced by native speakers is plotted in Figure 2.

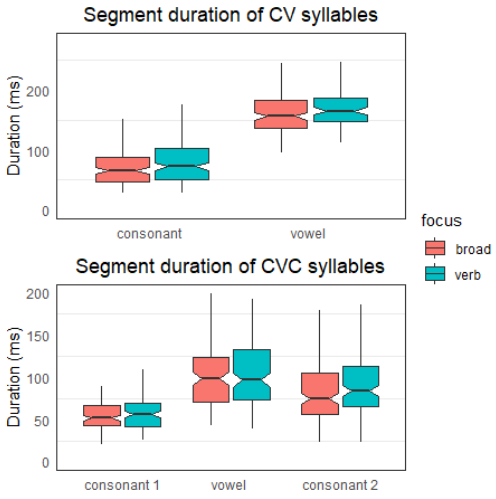


Figure 2: Segment duration of native speakers.

3.2. Durational adjustment of non-native speakers

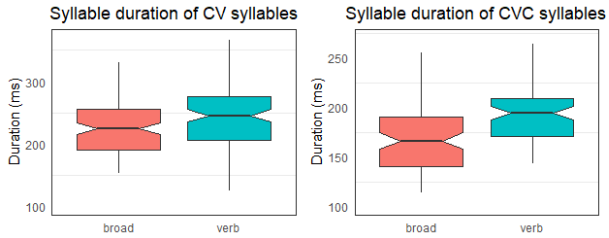


Figure 3: Syllable duration of non-native speakers.

This section reports the non-native speakers' data. At the syllabic level, there was a main effect of focus on syllable duration for the CV syllables ($\chi^2(1)$ = 44.038, p < .001) and for the CVC syllables ($\chi^2(1)$ = 41.129, p < .001), lengthening the on-focused syllable by 56.39 ± 7.92 ms and 26.56 ± 3.88 ms, respectively. The mean syllable duration of the syllables produced by non-native speakers is plotted in Figure 3.

Separate models were then fitted for the segment duration of CV and CVC syllables. For the CV syllables, there were main effects of focus ($\chi^2(1)$ = 13.287, p < .001) and segment ($\chi^2(1)$ = 385.17, p < .001) as well as an interaction between focus and segment ($\chi^2(1)$ = 9.658, p = .002) on the duration values. Post-hoc tests revealed a lengthening of the consonant duration (15.7 ± 5.69, p = .032) and a lengthening of the vowel duration (40.7 ± 5.69, p < .001) under focus. For the CVC syllables, there were main effects of focus ($\chi^2(1)$ = 14.023, p < .001) and segment ($\chi^2(2)$ = 303.35, p < .001) as well as an interaction between focus and segment ($\chi^2(1)$ = 9.843, p = .007). Post-hoc tests showed no differences for the first consonant (p = .942) between the two focus types, but the vowel (23.66 ± 4.44 ms, p < .001) and the second consonant (12.12 ± 4.44 ms, p = .071) were lengthened under focus. The mean segment duration of the syllables produced by non-native speakers is plotted in Figure 4.

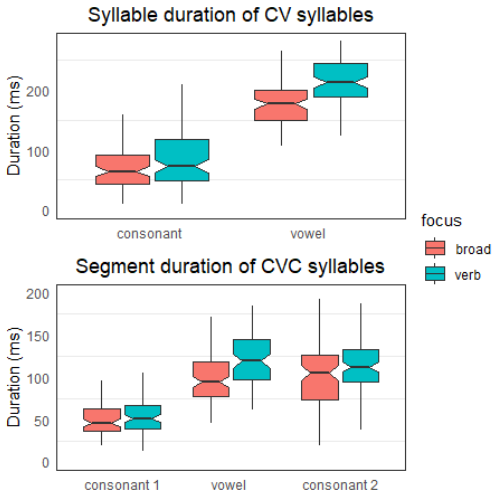


Figure 4: Segment duration of non-native speakers.

3.3. Comparing the two speaker groups

The duration patterns of the CV and CVC syllables produced by the native and non-native speakers are compared in this section.

Table 2: Duration patterns of the CV syllables in milliseconds.

	Focus type	C	V	CV
Native speakers	Broad	71.42	160.59	232.01
	Verb	80.36	180.30	260.66
	Increase	12.52%	12.28%	12.35%
Non-native speakers	Broad	70.46	175.61	246.06
	Verb	86.12	216.33	302.45
	Increase	22.23%	23.19%	22.91%

Table 2 presents the duration patterns for the CV syllables produced by the two speaker groups. The native and non-native speakers lengthened the duration of the initial consonant and the vowel equally, but the non-native speakers lengthened the

duration to a much greater extent than did the native speakers (around 12.5% for native speakers and 23% for non-native speakers).

Table 3: Duration patterns of the CVC syllables in milliseconds.

	Focus type	C1	V	C2	CVC
Native speakers	Broad	56.22	103.95	83.72	243.89
	Verb	59.67	110.62	95.59	265.88
	Increase	6.15%	6.41%	14.17%	9.02%
Non-native speakers	Broad	51.55	99.29	104.35	255.19
	Verb	55.62	122.94	116.47	295.03
	Increase	7.89%	23.83%	11.61%	15.61%

Table 3 shows the duration patterns for the CVC syllables produced by native and non-native speakers. The native speakers lengthened the first consonant and vowel slightly (6.15% and 6.41%), and the major lengthening was on the syllable-final second consonant (14.17%); this suggested that although the Cantonese speakers also marked focus via duration for CVC syllables, the lengthening was asymmetrical. Specifically, the initial consonant and the vowel were lengthened to a lesser extent and the final consonant was lengthened to a greater extent. However, the non-native speakers showed an extremely different pattern from the native speakers, as the duration of the vowel was expanded extensively (23.83%), and the initial and final consonants were expanded to a much lesser extent (7.89% and 11.61%, respectively).

4. DISCUSSION AND CONCLUSION

This study examined the manipulation of duration in Cantonese focus marking by native and non-native speakers using a production experiment. The first research question concerned how native speakers of Cantonese marked focus via durational adjustment. Two types of syllable structures were included in this study, namely CV and CVC syllables. Although there was on-focus lengthening in both syllable types, the durational adjustment for the CV and the CVC syllables differed. For the CV syllables, the consonant and the vowel were equally lengthened, and the contribution of the consonant and vowel to the lengthening of the syllable was similar. For the CVC syllables, the initial consonant and vowel underwent similar degrees of lengthening, and the final consonant experienced the greatest expansion. A question that then arose was why the segments of the CV syllables were expanded equally regardless of whether the vowel was a monophthong or diphthong, while the segments of the CVC syllables varied dramatically in the degree of focal lengthening. One may argue that the obvious lengthening of the final nasal sounds was an additive effect of word-final lengthening and focal lengthening, but if this was the case, a similar pattern should also have been observed for the CV syllables. Moreover, an earlier study of English suggested that the lengthening effects did not add up [24]. Based on the current data, whether the patterns in the CV syllables or in the CVC syllables were the fundamental ones in Cantonese focus-induced durational adjustment is far from conclusive. Thus, it is necessary to use all the possible syllable structures available in Cantonese to obtain a full picture of Cantonese focal lengthening in future studies. In addition to syllable structure, it is interesting to examine whether there are interactions of prosodic focus and lexical tones in Cantonese focal lengthening, as has been shown in other tonal languages (e.g., [25]).

Note that the amount of focal lengthening in Cantonese was around 10% (12.35% in CV syllables and 9.02% in CVC syllables), which was far less than indicated in the results reported previously in other languages [8]. A possible explanation for this might be that the target words in this study were all sentence-medial verbs. As the lengthening effect may be sensitive to the position of the focal components [5], further studies should include more positions to test whether this was caused by the present test design or by the nature of Cantonese focus lengthening.

With regard to the second research question, we showed that the non-native speakers of Cantonese did not have similar patterns to those of the native speakers. In general, the non-native speakers expanded the syllable duration to a greater extent than did the native speakers. For the CVC syllables, whereas the native speakers mainly lengthened the final consonant and only lengthened the initial consonant and the vowel slightly, the non-native speakers mainly lengthened the vowel, thus suggesting that the non-native speakers' Cantonese was non-native-like in terms of durational adjustment. Despite the non-native speakers all being fluent speakers of Cantonese, they had failed to fully acquire focus prosody in Cantonese. This indicates that suprasegmental features may be problematic for advanced L2 speakers and require more time and effort to acquire. It also remains to be explored whether the non-native-like lengthening patterns would be salient in revealing foreign accent.

In summary, this study examined the durational adjustment in Cantonese focus marking by native and non-native speakers. The native speakers demonstrated an asymmetrical effect of focal lengthening in different types of syllable structures. There were differences between the native and non-native speakers, which were attributed to the failure of the non-native speakers to acquire focus marking in Cantonese.

5. ACKNOWLEDGEMENTS

The author would like to thank all the informants for their participation and the ICPhS committee and reviewers for their constructive suggestions. This work was partially supported by an ASA International Student Grant from the Acoustical Society of America.

6. REFERENCES

[1] D. Büring, "Focus and intonation," in *The Routledge Companion to the Philosophy of Language*, G. Russell and D. G. Fara, Eds. London: Routledge, 2012, pp. 103–115.

[2] Y.-C. Lee, "Prosodic focus within and across languages," University of Pennsylvania, 2015.

[3] Y. Yang and S. Chen, "Revisiting focus production in Mandarin Chinese: Some preliminary findings," in *Proc. Speech Prosody 2020*, 2020, pp. 260–264.

[4] W. L. Wu, "Cantonese prosody: Sentence-final particles and prosodic focus," Ph.D dissertation, Univ. College London, London, 2013.

[5] Y. Chen, "Durational adjustment under corrective focus in Standard Chinese," *J. Phon.*, vol. 34, no. 2, pp. 176–201, 2006.

[6] C. DiCanio, J. Benn, and R. Castillo Garcia, "The phonetics of information structure in Yoloxóchitl Mixtec," *J. Phon.*, vol. 68, pp. 50–68, 2018.

[7] A. E. Turk and L. White, "Structural influences on accentual lengthening in English," *J. Phon.*, vol. 27, no. 2, pp. 171–206, Apr. 1999.

[8] M. Heldner and E. Strangert, "Temporal effects of focus in Swedish," *J. Phon.*, vol. 29, no. 3, pp. 329–361, 2001.

[9] J. E. Flege, G. H. Yeni-Komshian, and S. Liu, "Age Constraints on Second-Language Acquisition," *J. Mem. Lang.*, vol. 41, pp. 78–104, 1999.

[10] A. Sorace and F. Filiaci, "Anaphora resolution in near-native speakers of Italian," *Second Lang. Res.*, vol. 22, no. 3, pp. 339–368, Jul. 2006.

[11] K. Saito, "The Role of Age of Acquisition in Late Second Language Oral Proficiency Attainment," *Stud. Second Lang. Acquis.*, vol. 37, no. 4, pp. 713–743, 2015.

[12] D. Singleton, "The Critical Period Hypothesis: A coat of many colours," *IRAL - Int. Rev. Appl. Linguist. Lang. Teach.*, vol. 43, no. 4, pp. 269–285, 2005.

[13] K. Idemaru, P. Wei, and L. Gubbins, "Acoustic Sources of Accent in Second Language Japanese Speech," *Lang. Speech*, vol. 62, no. 2, pp. 333–357, Jun. 2019.

[14] J. E. Flege, I. R. A. MacKay, and D. Meador, "Native Italian speakers' perception and production of English vowels," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2973–2987, Nov. 1999.

[15] Y. Yang, "First Language Attrition and Second Language Attainment of Mandarin-speaking Immigrants in Hong Kong: Evidence from Prosodic Focus," Ph.D dissertation, The Hong Kong Polytechnic Univ., Hong Kong, 2022.

[16] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism," *COERLL, University of Texas at Austin*, 2012. [Online]. Available: https://sites.la.utexas.edu/bilingual/.

[17] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime User's Guide*. Pittsburgh: Psychological Software Tools Inc, 2012.

[18] Audacity Team, "Audacity(R): Free Audio Editor and Recorder." 2019.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2015.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[21] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2018.

[22] RStudio Team, "RStudio: Integrated Development for R." RStudio, Inc., Boston, MA, 2016.

[23] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer, 2016.

[24] W. Cooper, S. Eady, and P. Mueller, "Acoustical aspects of contrastive stress in question-answer contexts," *J. Acoust. Soc. Am.*, vol. 77, no. 6, pp. 2142–2155, 1985.

[25] Y. Yang, S. Chen, and K. Li, "Effects of focus on duration and intensity in Chongming Chinese," in *Proc. ICPhS 2019*, 2019, pp. 3578–3582.

On the sound source locations of 'glottal fricative' [h]

Tsukasa Yoshinaga[1], Kikuo Maekawa[2], Akiyoshi Iida[1]

[1]Toyohashi University of Technology, [2]National Institute for Japanese Language and Linguistics
yoshinaga@me.tut.ac.jp, kikuo@ninjal.ac.jp, iida@me.tut.ac.jp

ABSTRACT

To explore the sound source locations of the consonant usually represented as [h], we extracted the vocal tract and glottal geometries from magnetic resonance imaging (MRI) scans of a Japanese male in three contexts [ha], [he], and [ho], and physically measured the sound produced by the vocal tract replicas set on the rigid vocal fold models with different degrees of adduction. The results showed that the sound amplitudes of vocal tracts with open glottis were in the similar range to those with the adducted 3 mm glottal opening in the contexts of [ha] and [he]. In contrast, the sound amplitudes were significantly decreased by the abducted glottis in the context of [ho]. These results indicate that the vocal tract constriction mainly produces the sound sources of [h] in the context of [ha] and [he], whereas the glottal constriction mainly produces the sound sources of [h] in [ho].

Keywords: fricatives, aeroacoustics, turbulence, glottis, sound source.

1. INTRODUCTION

Fricative consonants are known to be realized by producing turbulent jet flows in a vocal tract. The jet flow is produced at a constriction in the vocal tract, and the position of the constriction is commonly called the place of articulation. While most of the fricative consonants are realized by forming the constriction in the oral tract, the glottal fricative [h] is known to show no clear place of articulation in the tract. Therefore, the glottal fricative is often described as a voiceless counterpart of the following vowels [7].

After the major revision of the IPA chart in 1989, there was much discussion on the manner of articulation of the glottal consonants. Some researchers argued that the turbulent airflow passing through the oral cavities produces the sound of [h] and those should be removed from the consonant chart [5-6], while others argued that there is obvious glottal constriction in some languages and [h] should be listed as the glottal fricative [8]. The recent review of phonetics [2] commented that this discussion is far from settled, and the IPA chart is expected to be re-examined in a few years' time.

From a point of view of acoustics or fluid dynamics, the acoustic characteristics of [h] have been theoretically analysed by Stevens [11-12]. Based on some assumptions on the vocal tract and glottal configurations (*i.e.*, cross-sectional areas), Stevens predicted that the spectral characteristics of [h] keep the same formant characteristics of following vowels, while the amplitudes of lower frequency ranges (around F1 to F4) decreases due to the difference of sound source characteristics between the turbulent airflow and glottal voice source. Although he additionally showed the effects of the vocal tract constriction on the acoustic characteristics, it is still unclear whether the main sound source of [h] is located near the glottis or in the oral cavity because no measurement data was provided regarding the vocal tract geometry of speakers pronouncing [h]. One attempt for measuring glottal openings using glottal electromyography (EMG) in [h] also lacks information of the vocal tract constriction [16], and the location of the sound source was not discussed.

Therefore, in this study, we explore the sound source locations of [h] by utilizing a recently developed MRI database. By extracting three-dimensional vocal tract geometry and conducting mechanical experiments, we clarify the effects of the glottal and oral vocal tract constrictions on the acoustic characteristics of [h] in various contexts.

2. MATERIALS AND METHODS

2.1. Vocal tract acquisition

The three-dimensional vocal tract geometry of [h] was obtained as a side product of the construction of the real-time MRI articulatory movement database [9]; the three-dimensional static MRI scans of selected phonemes were recorded for some participants.[1] The subject of this study is an adult male who speaks Tokyo Japanese, and he sustained [h] for 15 s in three phonemic contexts: /ha/, /he/, and /ho/. Since /hi/ and /hu/ in Japanese are known to be realized as [çi] and [ɸɯ] [13-14], we excluded these contexts in this study. The image dimensions were set to 256 × 256 × 28 and the voxel spacing became 1× 1 × 3 mm^3.

The vocal tract geometry above vocal folds was extracted based on luminosity values using the software itk-SNAP [17]. The mid-sagittal planes of

the MRI and extracted vocal tracts of [h] in three vowel contexts are shown in Fig. 1. Since no close contact of upper and lower incisors was observed in [h] of this subject, the reconstruction procedure of incisors was omitted. In the vocal tract of [h] in /ha/, the tongue retracted and the airflow channel at the pharynx was constricted. In the context of /he/, the size of the frontal oral cavity was reduced, and a constriction appeared at the hard palate. This vocal tract shape is relatively close to that of [ç], though the cross-sectional area of [h] (~30 mm^2) was much larger than that of [ç] (~9 mm^2) [15]. The vocal tract shape of [h] in /ho/ was similar to that of /ha/ in that the frontal cavity has no constriction, but the area of the pharyngeal constriction in /ho/ was larger than that of /ha/ and lips were rounded. Moreover, small openings of the velopharyngeal valve were observed in /ha/ and /ho/, but the flow channel above the velum was cut to assume the velopharyngeal closure during [h].

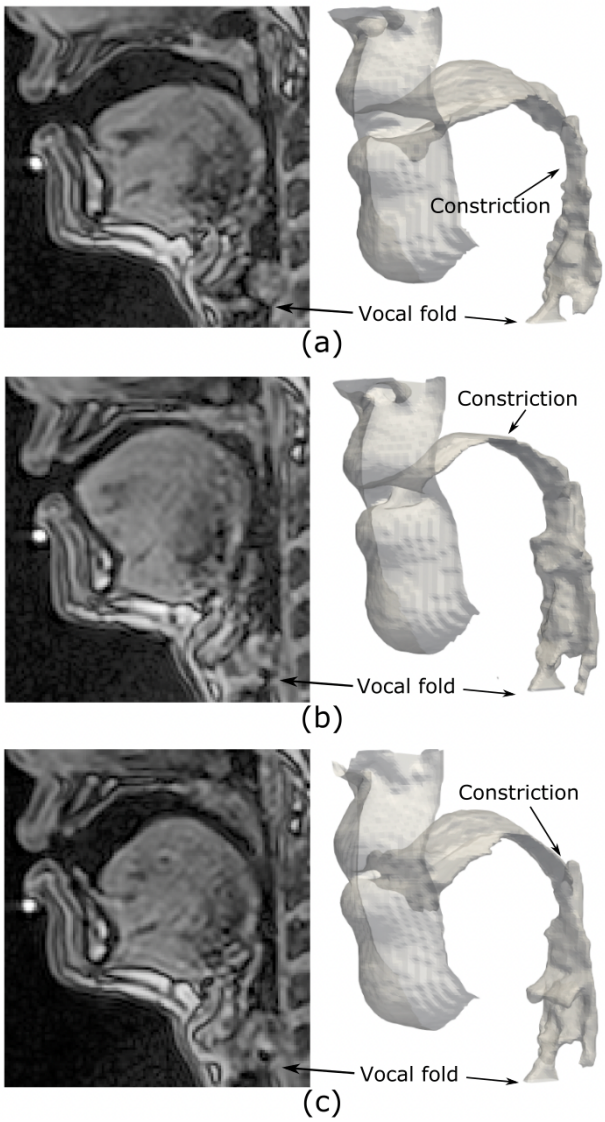


Figure 1: MRI scans of the mid-sagittal plane and vocal tracts of [h] in three phonemic contexts: /ha/ (a), /he/ (b), and /ho/ (c).

Table 1: Constriction areas of the vocal tracts and glottal channels between the epiglottis and arytenoid cartridge in the three vowel contexts.

Context	Constriction area (mm^2)	
	Vocal tract	Arytenoid cartridge
/ha/	28.1	20.2
/he/	28.4	27.4
/ho/	40.6	24.2

The characteristic cross-sectional areas of each vocal tract are summarized in Table 1. While each vocal tract had a constriction in the vocal tract, another constriction appeared at the glottal channel between the epiglottis and arytenoid cartridge. This constriction near the glottis was narrower than the vocal tract constrictions for all vowel contexts.

2.2. Vocal fold model

Since the glottal constriction (degree of adduction) cannot be observed with the current resolution of MRI, we constructed a rigid vocal fold model with three slit widths and set at the position of vocal folds shown in Fig. 1. The vocal folds were assumed to have a triangular opening (Fig. 2) and the distance between arytenoid cartridges was changed from L_g = 1 to 3 mm, while the vocal fold length and thickness were fixed to L_m = 15 mm and L_t = 3 mm, respectively. These dimensions are typical values of adult males [3]. With the change of L_g from 1 to 3 mm, the constriction area of the glottis varies from 7.5 to 22.5 mm^2.

In addition to these adducted glottises, we constructed geometry of an open glottis of normal breathing from the MRI scan of the same subject.[2] The images were taken with the resolution of 0.33 × 0.33 × 1 mm^3, and the glottal geometry was extracted as shown in Fig. 3. The vocal tract with open vocal folds was constructed by replacing the glottal cavity just below the pharynx of [h] for three vowel contexts. The minimum cross-sectional area of the open glottis was 71.9 mm^2.

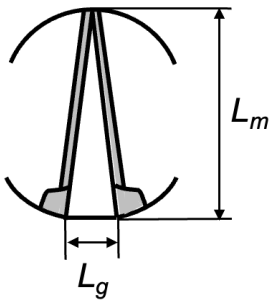


Figure 2: Vocal fold model from the top view. The glottal opening L_g was changed from 1 to 3 mm.

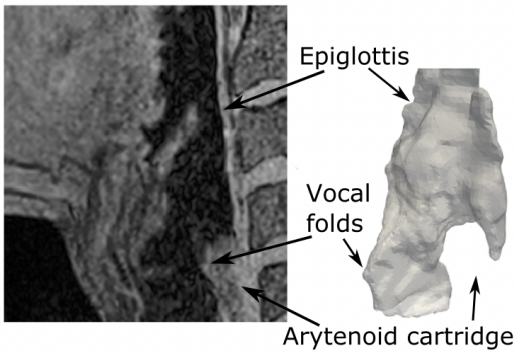


Figure 3: MRI scan of the mid-sagittal plane and open glottal geometry.

2.3. Mechanical experiment

To measure the sound produced by the vocal tracts and glottises specified in the previous sections, we constructed vocal tract replicas and the vocal fold model with acrylic resin using rapid prototyping (Foam3, Foamlabs, USA). The replica was set on the pressure chamber (57 × 32 × 42 mm^3) that mimics a volume of the subglottal tract, and airflow was inserted from a compressor (SOL-2039, Misumi, Japan) through a flow meter (PFM750-01, SMC, Japan) and air tubes with an inner diameter of 16 mm. The flow rate was changed from 83 to 500 cm^3/s (typical range of speech production [12]) by adjusting a flow valve.

The replica was set in an anechoic chamber (Fig. 4) and the sound produced by the replica was measured by a microphone (46AF, G.R.A.S., Denmark) at 300 mm from the lips. The sound was recorded with a sampling frequency of 50 kHz, and the spectrum was calculated by discrete Fourier transform with 60 averages of 1000 point-frames and Hann-window. The overall sound pressure level (OASPL) was calculated by integrating sound pressures in the frequency range from 0.5 to 16 kHz for the comparison of sound amplitudes for each glottal condition.

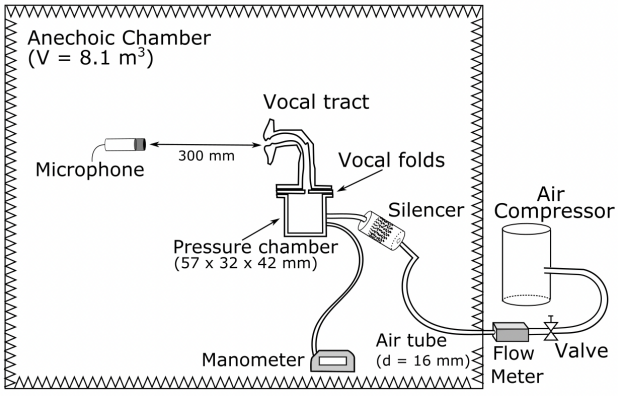


Figure 4: The experimental apparatus.

3. RESULTS

The spectra of sound measured for three vocal tracts with adducted (L_g = 3 mm) and abducted glottises are shown in Fig. 5. The background noise (BGN) indicates the sound amplitudes without the airflow from the equipment. For all three vocal tracts, clear peaks appeared in the lower frequency range (< 2 kHz) compared to the typical fricative consonants [4]. The main peaks were located at 1300 and 2600 Hz for /ha/, 600 and 1950 Hz for /he/, and 700 and 1000 Hz for /ho/. While the peaks of [h] in /he/ and /ho/ fell within the range of typical formant frequencies of the following vowels (i.e., [ɛ] and [ɔ]) [10], the peaks of [h] in /ha/ had much higher frequencies than the typical values of the vowel /a/. The typical F1 of /a/ is in the range from 800 to 1000 Hz, and a fragment of F1 was observed at the shoulder of the first peak (at 900 Hz), indicating the turbulent airflow sound sources are dominant in the higher frequency range [12]. It should be noted that the first peaks of [h] in /he/ and /ho/ only appeared with the glottal adduction, suggesting the vocal folds have an acoustic role in forming the first peak.

The amplitudes of the sounds produced by the vocal tracts with adducted and abducted vocal folds were in the similar range for [h] in /ha/ and /he/. In contrast, the sound amplitudes of [h] in /ho/ were increased 10 to 15 dB by narrowing the glottis with L_g = 3 mm. This is because the vocal tract constrictions in /ha/ and /he/ were more prominent than that in /ho/, and the sound sources were mainly produced by the vocal tract constrictions. Meanwhile, the sound sources of [h] in /ho/ were mainly produced at the glottis, so that the amplitudes of sounds were decreased by the opening of the vocal folds.

The OASPLs of the sounds produced by each vocal tract with different flow rates are plotted in Fig. 6. The vocal tracts of each vowel context showed different tendencies depending on the degree of adduction. For the vocal tract of [h] in /ha/, the OASPLs increased in a similar way for all glottal conditions with the increment of the flow rate. This is because the main sound source was located at the vocal tract constriction. In the context of /he/, the OASPLs were decreased about 10 dB with the increase of L_g from 1 to 3 mm, and those also decreased about 5 dB by opening the glottis. For the vocal tract of [h] in /ho/, the OASPLs were almost the same from L_g = 1 to 3 mm, whereas the OASPLs decreased 15 dB by opening the glottis. Those results indicate that the sound sources were produced at both vocal tract and glottis in /he/, and the amplitudes were largely changed by the glottal width, while the sound was also produced with the vocal tract constriction as shown in Fig. 5 (b). In the context of /ho/, the main

sound source was located at the glottis, and the OASPLs significantly increased with the adducted glottises.

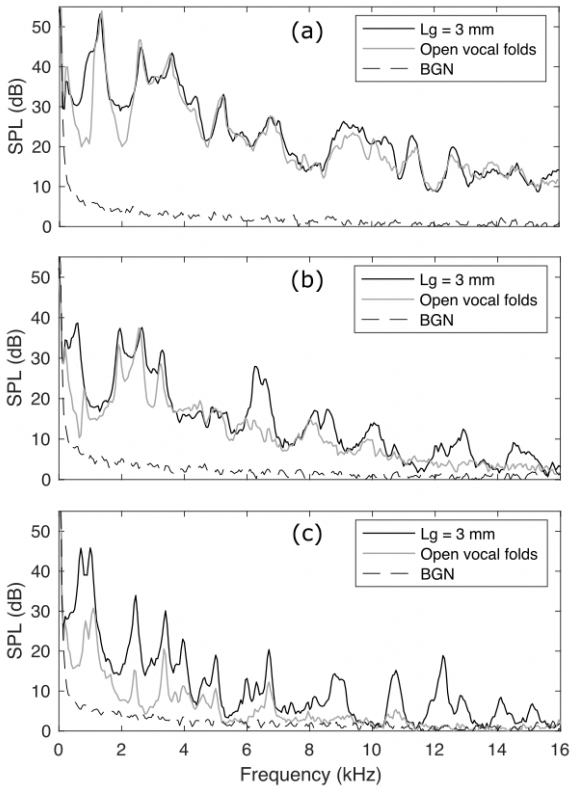


Figure 5: Spectra of sounds measured for the vocal tracts with adducted (L_g = 3 mm) and abducted glottis. The vowel contexts of /ha/, /he/, /ho/ are plotted in (a), (b), and (c), respectively.

4. DISCUSSION AND CONCLUSIONS

The current study showed that the role of glottis is largely depending on the degree of constriction in the vocal tract. When the constriction area was less than 30 mm^2 as in /ha/ and /he/, the large amplitudes of sounds were produced only with the oral constriction, and this suggests the unnecessity of the glottal constriction in [h]. In contrast, when the maximum oral constriction area was about 40 mm^2 as in /ho/, only small noise was produced, and adducted vocal folds significantly increased the sound amplitudes.

Although the previous theoretical analysis suggested the limited effects of the vocal tract constriction around 30 mm^2 of /hi/ and /hu/ [11-12], we found that the vocal tract constriction within that range of the size can produce the main sound source for [h], especially in the context of /ha/. This result suggests that [h] in /ha/ could be replaced by other phonemes like velar [x] or pharyngeal [ħ]. Indeed, the vocal tract geometry in Fig. 1 (a) looks close to that of [x] observed with an X-ray scan in a previous study [1].

In conclusion, this study demonstrated the possibility of three states of sound source locations in [h]. First, there are cases the sound source is mainly produced at the vocal tract and few effects of the glottal constriction are observed as in /ha/ of this study. Second, there are cases the sound sources are produced at both the oral tract and the glottis as in /he/, where the sound amplitudes are increased by the glottal constriction. Third, there are cases like /ho/ where the glottal constriction produces the main sound source with the result that the sound is difficult to be heard without the glottal adduction.

To further validate these phenomena for the next step, it is necessary to increase the number of subjects to be examined with the MRI and the mechanical experiment.

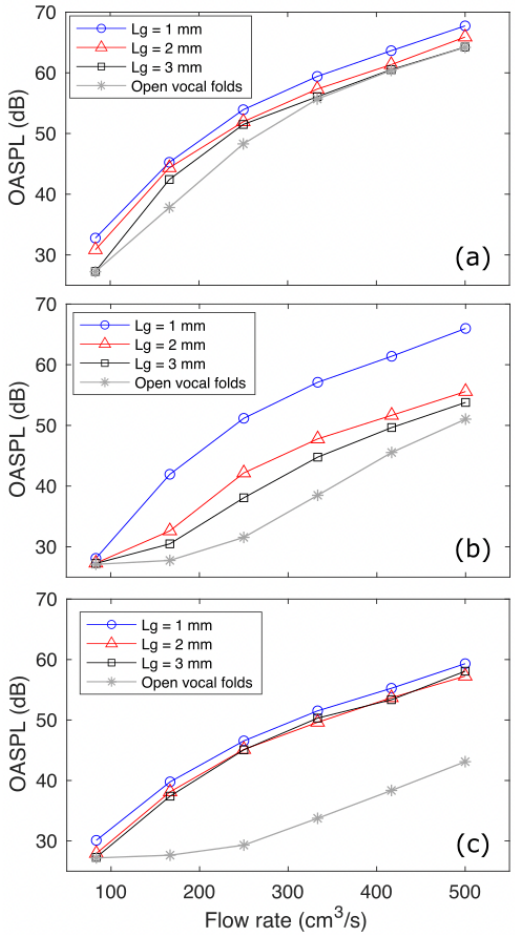


Figure 6: The overall sound pressure levels (OASPLs) of the sound produced by the vocal tracts with different degrees of glottal adduction. The vowel contexts of /ha/, /he/, /ho/ are plotted in (a), (b), and (c), respectively.

5. ACKNOWLEDGEMENTS

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number: JP19K21641 and JP20H01265.

6. REFERENCES

[1] Badin, P. 1991. Fricative consonants: acoustic and X-ray measurements. *J. Phonetics*, 19, 397-408.

[2] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *J. Phonetics*, 94, 101155.

[3] Hirano, M., Kurita, S. Nakashima, T. 1981. The structure of the vocal folds. *Vocal Fold Physiology*. (University of Tokyo Press, Tokyo), 33-41.

[4] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, 108(3), 1252-1263.

[5] Kloster-Jensen, M. 1991. Letters to the Editor. *J. Int. Phonetic Assoc.*, *21*(1), 42-42.

[6] Ladefoged, P. 1990. Some proposals concerning glottal consonants. *J. Int. Phonetic Assoc.*, 20(2), 24-25.

[7] Ladefoged, P., Maddieson, I. 1998. *The sounds of the world's languages.* (Blackwell, London).

[8] Laufer, A. 1991. The 'glottal fricatives.' *J. Int. Phonetic Assoc.*, *21*(2), 91-93.

[9] Maekawa, K. 2021. Production of the utterance-final moraic nasal in Japanese: A real-time MRI study. *J. Int. Phonetic Assoc.* 53(1), 189-212.

[10] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24(2), 175-184.

[11] Stevens, K. N. 1971. Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am.*, *50*(4), 1180-1192.

[12] Stevens, K. N. 2000. *Acoustic phonetics.* (MIT press, Cambridge).

[13] Vance T.J. 2008. *The Sounds of Japanese.* (Cambridge Univ. Press, Cambridge).

[14] Yoshinaga, T., Maekawa, K., Iida, A. 2021. Aeroacoustic differences between the Japanese fricatives [ɕ] and [ç]. *J. Acoust. Soc. Am.*, 149(4), 2426-2436.

[15] Yoshinaga, T., Maekawa, K. Iida, A., 2022. Variability in Production of Non-Sibilant Fricative [ç] in /hi/. *Proc. Interspeech 2022*, 620-624.

[16] Yoshioka, H. 1981. Laryngeal adjustments in the production of the fricative consonants and devoiced vowels in Japanese. *Phonetica*, 38(4), 236-251.

[17] Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., Gerig, G. 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3), 1116-1128.

[1] The URL of the Realtime MRI Articulatory Movement Database: https://rtmridb.ninjal.ac.jp (Noted that the three-dimensional static scans have not been included in the database yet).

[2] This MRI scan was conducted for the purpose of obtaining the upper and lower incisor angles, and thus the glottis was abducted for normal breathing.

AUTOMATIC CLASSIFICATION OF VOWELS IN SEMI-SPONTANEOUS PAKISTANI PUNJABI

Farhat Jabeen, Moritz Wackerbarth, Petra Wagner

Bielefeld University, Germany
firstname.lastname@uni-bielefeld.de

ABSTRACT

This is the first machine learning based analysis of oral vowels produced in semi-spontaneous Pakistani Punjabi. Our investigation is based on lower frequency characteristics (f0, F1, F2, F3) and vowel duration. We use random forest classification to investigate the prediction of Punjabi vowels based on these acoustic features. For allophonic variation, we use a confusion matrix to identify the false positive and false negative vowel predictions. Our data indicate that frequency characteristics do not reliably classify different vowels of the same class (e.g., front or high). This results in varying rates of successful predictions in the classification of vowels. The analysis of feature importance shows that F1 and F2 play the most important role in the classification of Punjabi vowels, followed by duration. The influence of F3 and f0 is negligible in this regard. The misclassification of vowels in the same group highlights the phonemic vs. allophonic variation in Pakistani Punjabi.

Keywords: vowels, formants, semi-spontaneous speech, Punjabi, automatic classification

1. BACKGROUND

Punjabi is an Indo-Aryan language spoken mainly in India and Pakistan. Indian Punjabi is written in Gurumukhi script, whereas Pakistani Punjabi uses the Shahmukhi script. Notwithstanding, speakers of Pakistani and Indian Punjabi appear to understand each other fairly well. The literature review in the following subsections specifies if an analysis is based on an Indian or a Pakistani dialect of Punjabi.

1.1. Vowel inventory

Being an Abugida language, every consonant in Punjabi is associated with an inherent vowel whose quality changes depending on the associated diacritic [1]. The ten oral vowels in Punjabi are distinguished on the basis of their quality and quantity. [2] reported an inventory of three short /ɪ ʊ ə/ and seven long vowels /i e ɛ u o ɔ a/. [3] described that the tense/lax distinction in Punjabi is based on vowel quantity. However, [4] stated that vowel quality plays a more important role in Punjabi compared to vowel duration. They made a distinction between central lax vowels /ɪ ʊ ə/ and peripheral tense vowels /i e ɛ u o ɔ a/. Examples of words listed by [5] show that this difference is reflected in their distribution as peripheral vowels occur at all positions in a word, while central vowels are placed word finally in close syllables only.

For their analysis of oral vowels produced by speakers of Majhi dialect in India, [6] used word lists consisting of monosyllabic words. Their proposed inventory of vowels differed from the ones reported in existing literature. For example, [6] included the open vowel /æ/ in their inventory instead of /ɛ/ reported by [7] in their analyses of Punjabi spoken in Delhi. Moreover, [6] characterized schwa as a near-open vowel compared to the schwa reported as a central vowel by [3]. The source of these differences in vowel quality remains unclear as [6] did not provide much information about their participants.

1.2. Allophonic variation

The issue of allophonic distribution contributes to the disagreement regarding the vowel inventory of Punjabi. [2] reported that the vowel pairs /ɪ e/ and /ʊ o/ are allophones. But she offered no explanation regarding their distribution. It is interesting to note that the allophonic alteration of these vowels is mentioned for Pakistani Punjabi only and has not been reported in the analyses of Indian Punjabi.

Most of the existing analyses of Punjabi vowels are based on Indian dialects. [5] is the only acoustic analysis of Punjabi spoken in Pakistan (Lyallpuri dialect). Their data showed a closely shared acoustic space by the following set of vowels: /i ɪ e/, /ʊ o/, and /ə a/. Despite this, [5] do not refer to the allophonic variation reported by [2]. Moreover, as [5]'s analysis is based on one speaker's data, further investigation is warranted.

The acoustic analyses of Punjabi mentioned above are based on data produced by a few speakers reading out word lists. To date, there is

no analysis of Punjabi vowels produced in fluent speech. Moreover, no detailed acoustic investigation of vowels in Pakistani Punjabi has been carried out so far. This study aims to fill these gaps by analysing vowels produced in fluent speech by speakers of Pakistani Punjabi (Lahori variety). We also investigate the importance of different acoustic features in the automatic classification of Punjabi vowels. The details of our analysis are as follows.

2. EXPERIMENT

2.1. Participants

Ten male[1] speakers of Punjabi were recorded for this experiment. The participants could not reliably state their dialect of Punjabi. Hence, the geographical area where they spent their early years was used as a proxy for dialect. All the participants had grown up in Greater Lahore and surrounding regions in Pakistan. One speaker's data was excluded as he audibly spoke a different dialect (Multani).

2.2. Data collection

Participants were shown an animation video (4:28 minutes) on YouTube[2]. At the end of the video, they were asked to retell the story in their own words providing as much detail as they could remember. The data was recorded using the university license of the Zoom software[3] (sampling rate: 32KHz, 16 Bit). The data consisted of 26:18 minutes of speech.

2.3. Data annotation

Every speaker's recording was divided into inter-pausal units (IPU) separated by a pause of at least 150ms. Given the lack of consensus on the inventory of diphthongs in Punjabi [4, 5], we excluded all instances of consecutive vowels as well as vowels preceded/followed by approximants. When vowels were preceded by a plosive, boundaries were placed after the burst or the end of aspiration for voiced and voiceless stops respectively. For vowels preceded by a sibilant, boundaries were marked at voice onset for voiceless and at regions of decline in high frequency energy for voiced sibilants. Vowel quality was annotated using the vowel inventory provided by [5].

Vowel boundaries were marked manually by the first (native speaker) and second (no knowledge of Punjabi) authors. The annotators reported uncertainty in the annotation of high front vowels /i ɪ e/, high back vowels /ʊ u o/, and the central low vowels /ə a/. All the annotations by the second author were verified by the first author. In case of inter-annotator disagreement, the formant values of other vowels produced by the same speaker were used to annotate a given vowel.

2.4. Data analysis

A Praat [8] script was used to extract vowel duration as well as the first three formants and f0 measured in the middle of vowels. The formant values were Bark normalized using the "PhonR" package in R [9]. As the data consists of fluent speech, we used z-transformed (by vowel) vowel duration.

Automatic classification was carried out using the random forest algorithm implemented in the Scikit-learn toolkit [10] in Python. The data was randomly divided into training (70%) and test (30%) sets. The model included z-transformed vowel duration (durZ) along with f0 and the first three Bark-normalized formant frequencies. Moreover, permutation based feature importance[4] was calculated on the held-out test set. This helped investigate the contribution of each acoustic feature to the power of the model. The results are discussed in the following section.

3. RESULTS & DISCUSSION

3.1. Frequency of vowels

In total, 3,182 oral vowels were found in our corpus of semi-spontaneous speech. Table 1 presents the frequency and percentage of each vowel. It shows that front vowels occurred more frequently (44%) in the data compared with the back vowels (18%). The high frequency of central vowels is largely due to /a/ that covers a quarter of the data. Moreover, the open-mid back vowel /ɔ/ was found very rarely in this corpus. This raises questions about the status of this vowel in Pakistani Punjabi as it might be an allophone of another back vowel. However, our corpus is based on the narration of a specific story which is not phonetically balanced. Future studies based on other texts and speech genres can shed further light on the frequency of this vowel.

3.2. Formant analysis

Fig. 1 illustrates the Bark-normalized F1 and F2 values for Punjabi oral vowels. Each vowel in the figure presents an individual data point for a given vowel type. The figure shows that Bark-normalized F1 and F2 capture the distinction between front and back vowels on one hand and between high and low vowels on the other. However, there is a high degree of overlap between front vowels /i ɪ e/ and between

i	ɪ	e	ɛ	ʊ	u	o	ɔ	ə	a
242	378	657	141	137	129	280	33	388	797
8%	12%	21%	4%	4%	4%	9%	1%	12%	25%

Table 1: Vowel frequency & percentage.

back vowels /u ʊ o ɔ/. In fact, the ellipse of /u/ is placed entirely within the ellipse for /ʊ/. The overlap is reminiscent of the allophonic realization of /ɪ e/ and /ʊ o/ reported by [2]. Furthermore, there is considerable variation in vowel quality as we find instances of schwa with low F1 and some fronted productions of /u/ and /a/. As our data consists of fluent speech, there is bound to be a high rate of coarticulation in the realization of vowels.

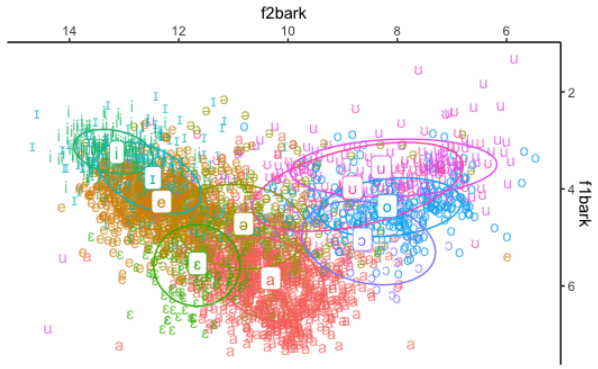


Figure 1: Bark normalized F1 and F2. The text boxes show mean formant frequencies.

Figure 1 differs only slightly from the vowel inventory offered by [5]. In our data, /e/ is realized with a higher F1 (close-mid) in comparison with [5]'s data. The most prominent difference may be observed in the quality of schwa produced as a central vowel in our data, whereas [5] had reported it to have a higher F1, in close approximation with /a/. This difference in vowel quality may be attributed to speech styles. As we have analysed fluent speech, the vowels are produced in different syllable structures (simplex and different types of complex syllables) placed at different positions in a word. In comparison, the data reported by [5] was based on vowels produced in CVC syllables in a word list. Moreover, the vowels in our data are produced in varying prosodic contexts that influence their quality. Finally, the difference between our and [5]'s results may also be attributed to dialectal variation. However, [5] claimed that the Layallpuri dialect analysed in their study is very similar to the Lahori variety spoken by our participants. Hence, the difference may be better explained by style and prosody based differences instead of dialect.

3.3. Classification report

The report for the random forest classification is given in Table 2[5]. It was computed using the true vowel labels in the held-out test set compared with the labels predicted by the model. The overall classification metrics show that the model performed fairly well. Table 2 further illustrates that the model's performance varied for different vowels. The metrics for the classification of /a/ are quite high, whereas our model made no predictions for /ɔ/ because only 12 data points were included in the model for this vowel. It could be argued that the size of the data set for each vowel included in the model is the main predictor of the classifier's performance. While its importance – unsurprisingly – cannot be denied, we argue that the performance of the classifier does not solely result from the number of data points available for each vowel in the model. Consider the classification of /ɛ/, /ʊ/, and /u/. Although the data sets for these vowels are similar in size (cf. Table 2), their precision, recall, and F1-scores differ greatly. Furthermore, the data support for schwa is high, yet the model makes better predictions for /ɛ/ than it does for schwa. In fact, the metrics are generally low for the lax vowels /ɪ ʊ ə/ compared with the peripheral ones. In future research, we aim to include other acoustic features to analyse the tense/lax distinction in Punjabi vowels.

Vowel	Precision[6]	Recall[7]	F1-score[8]	Support[9]
i	0.69	0.64	0.67	81
ɪ	0.59	0.46	0.52	126
e	0.65	0.85	0.73	178
ɛ	0.65	0.47	0.54	47
ʊ	0.76	0.33	0.46	40
u	0.71	0.71	0.71	38
o	0.67	0.76	0.71	87
ɔ	0.00	0.00	0.00	12
ə	0.45	0.42	0.43	97
a	0.83	0.90	0.86	230
Overall	0.66	0.68	0.66	

Table 2: Classification report.

The metrics given in Table 2 only provide a partial picture, as we do not know the false positives

predicted by the classifier. This is illustrated in the confusion matrix shown in Table 3. The x-axis here presents the true vowel labels from the test set, while the vowel labels on the y-axis are predicted by the classifier. The confusion matrix is a good reflection of the overlap observed in F1 and F2 values plotted in Fig. 1. The front vowels are frequently confused with each other, whereas the distinction between front-back and high-low vowels is maintained more successfully. Schwa is an exception to that as it was incorrectly predicted as /a/ as well as /e/.

	i	ɪ	e	ɛ	ʊ	u	o	ɔ	ə	a
i	52	19	10							
ɪ	16	58	43	1			1		7	
e	6	8	151	2		1	1		9	
ɛ		1	5	22					7	12
ʊ		3			13	3	13		6	2
u	1		1		2	27	7			
o			1		2	6	66		6	6
ɔ						1	5	0	1	5
ə		9	20	4		1	5		41	18
a		1	2	5					15	207

Table 3: True labels are presented on x-axis and the labels predicted by the classifier are given in the first column. True positives are shown in dark cells and frequent false negatives in lighter cells.

Taken together, Table 2 and 3 illustrate the need for investigating the phoneme/allophone distinction for Punjabi vowels. As mentioned earlier, [2] claimed that the vowel pairs /ɪ e/ and /ʊ o/ are allophonic. The output of the classification algorithm indicates this further, as the recall and F1-score for /ɪ/ are low compared with these metrics for /e/. Similarly, the recall and F1-score for /ʊ/ are lower than the comparative metrics for /o/. Furthermore, the confusion matrix shows that /e/ is the most frequent false negative prediction for /ɪ/, and /ʊ/ was frequently predicted as /o/. This indicates that there is considerable overlap in the formant frequencies of these vowel pairs. Another overlap observed in our data is between /ə e/ and /ə a/. It's intriguing that schwa is misclassified with either a close-mid or an open vowel. This aspect of schwa in Punjabi has never been reported before. The differing quality of schwa is not surprising as it is considered the "default" vowel in Punjabi. It is not written orthographically [11] and is claimed to be inherently associated with every consonant. [1] reported that the quality of this inherent vowel may vary. Our data provides first acoustic evidence of this change in the quality of schwa in Pakistani Punjabi. Further research is needed to investigate this variation.

Fig. 2 illustrates the importance of different acoustic features when it comes to the classification of vowels in our data. It shows that Bark-normalized F1 and F2 play the most important role followed by z-normalized vowel duration. The lower importance of duration is not surprising as it is influenced by factors such as syllable type and structure, position in a word, and lexical stress [12]. As our data comprises fluent speech, we can not control for these variables. For future analyses, we aim to include syllable and stress level information to investigate the vowel quantity in Punjabi. As the decrease in mean accuracy is also affected by speakers, it is indicative of speaker-based variation in the classification of these vowels.

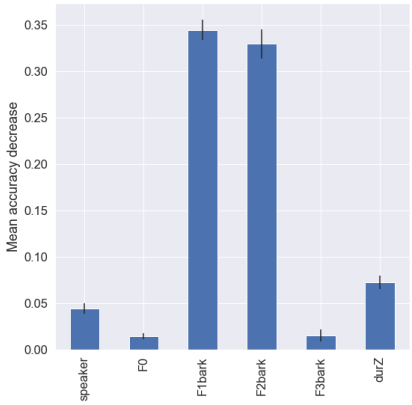


Figure 2: Importance of acoustic features. Bigger decrease in accuracy indicates higher importance.

4. CONCLUSION & FUTURE WORK

Overall, our results confirm the vowel inventory reported by [2, 5]. However, the output of automatic classification indicates that some vowels, hitherto considered as phonemes, may be allophonic in Punjabi. This highlights the importance of using machine learning algorithms to investigate acoustic phenomena. In future, we plan to conduct a perception experiment regarding the identification and discrimination of vowels in Punjabi. Furthermore, we have shown that F1 and F2 play the most important role in the classification of Punjabi vowels. However, these acoustic features led to poor predictions for the automatic classification of lax vowels. Future research should include other variables to analyse the tense/lax distinction in Punjabi vowels.

This is the first detailed analysis of vowels in Pakistani Punjabi. Despite being spoken by millions, research on this variety is lacking. Our methods and findings can be used to inform acoustic analyses of other South Asian languages.

ACKNOWLEDGEMENT

Many thanks go to Fritz Seebauer for the informed discussions on the selection and interpretation of the classification algorithm.

5. REFERENCES

[1] V. Goyal and G. S. Lehal, "Comparative study of Hindi and Punjabi language scripts," *Nepalese Linguistics*, vol. 23, pp. 67–82, 2008.

[2] N. Karamat, "Phonemic inventory of Punjabi," 2012, retrieved from Center for Language Engineering and Technology: http://www.cle.org.pk/information/people/nayyarakaramat.html.

[3] S. Lata, P. Verma, and S. Kaur, "Acoustic characteristics of schwa vowel in Punjabi," in *Proceedings of 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, Gurugram, India, 2018.

[4] O. Maxwell and J. Fletcher, "Acoustic and durational properties of Indian English vowels," *World Englishes*, vol. 28, no. 1, pp. 52–69, 2009.

[5] Q. Hussain, M. Proctor, M. Harvey, and K. Demuth, "Punjabi (Lyallpuri variety)," *Journal of the International Phonetic Association*, vol. 50, no. 2, pp. 282–297, 2020.

[6] P. Singh and K. Dutta, "Formant analysis of Punjabi non-nasalized vowel phonemes," in *Proceedings of International Conference on Computational Intelligence and Communication Systems*, 2011.

[7] V. Narang and D. Misra, "Redefining acoustic space in language contact situation: Case of Hindi and Punjabi in Delhi," *Interdisciplinary Journal of Linguistics*, vol. 5, pp. 13–24, 2012.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program, [v. 6.0.56]," 2013, available at http://www.praat.org/ [retrieved 20.11.2017].

[9] R Core Team, *R: A language and environment for statistical computing [v. 4.2.0]*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] P. Singh and G. S. Lehal, "A rule based schwa deletion algorithm for Punjabi TTS system," in *Proceddings of International Conference on Information Systems for Indian Languages*. Springer, 2011, pp. 98–103.

[12] S. G. Nooteboom and G. J. N. Doodeman, "Production and perception of vowel length in spoken sentences," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 276–287, 1980.

[13] L. Yu and N. Zhou, "Survey of imbalanced data methodologies," 2021, preprint at arXiv.

[1] For our preliminary study, we decided to analyse only male speakers' data. This allowed us to compare our findings with those of [5] who reported data from one male speaker. We plan to include female speakers in future works.

[2] https://www.youtube.com/watch?v=J_t2Z6_fuvU

[3] https://www.zoom.us/

[4] For details on permutation based feature importance, see https://scikit-learn.org/stable/modules/permutation_importance.html##permutation-importance

[5] As the vowel frequencies in our data are highly skewed, this is reflected in the number of data points for each vowel (support) included in the model. We decided against using the under/oversampling of vowels in our data as [13] has shown that the F-score for random forest classification is not affected significantly by imbalanced categories.

[6] Ratio of true positives & false positives

[7] Ratio of true positives & false negatives

[8] Weighted harmonic mean of precision & recall

[9] Number of observations for each vowel type

Non-uniform cue-trading: Differential effects of surprisal on pause usage and pause duration in German

Ivan Yuen, Omnia Ibrahim, Bistra Andreeva, Bernd Möbius

Department of Language Science and Technology, Universität des Saarlandes
ivyuen@lst.uni-saarland.de; omnia@lst.uni-saarland.de; andreeva@lst.uni-saarland.de; moebius@lst.uni-saarland.de

ABSTRACT

Pause occurrence is conditional on contextual (un)predictability (in terms of surprisal) [10, 11], and so is the acoustic implementation of duration at multiple linguistic levels. Although these cues (i.e., pause usage/pause duration and syllable duration) are subject to the influence of the same factor, it is not clear how they are related to one another. A recent study in [1] using pause duration to define prosodic boundary strength reported a more pronounced surprisal effect on syllable duration, hinting at a trading relationship. The current study aimed to directly test for trading relationships among pause usage, pause duration and syllable duration in different surprisal contexts, analysing German radio news in the DIRNDL corpus. No trading relationship was observed between pause usage and surprisal, or between pause usage and syllable duration. However, a trading relationship was found between the durations of a pause and a syllable for accented items.

Keywords: surprisal, pause, prosodic boundary, information status, cue-trading

1. INTRODUCTION

Pause serves multiple linguistic or extra-linguistic functions. For instance, it can be used to delimit hierarchical linguistic units [e.g., 24, 27], and is often interpreted as an index to reflect difficulties in word selection [e.g, 10, 11], speech production [e.g., 17] and advance planning [e.g., 14, 15], although it can also express extralinguistic information such as the degree of commitment affecting the recognition of emotions [e.g., 25].

Early research analysing spontaneous speech data has identified predictability as a factor influencing the occurrence of a pause. For instance, Goldman-Eisler [10, 11] found that a pause tended to occur at the juncture (i.e., in the context of) of a lexically-frequent preceding word and a lexically-infrequent following word, compared to their counterparts without any pause. This suggests the usage of pause might be conditioned by contextual predictability. However, using lexical word frequency to define predictability could have confounded semantic and structural factors.

Evidence for the influence of contextual predictability has been accumulated in recent years on the implementation of acoustic cues (e.g., temporal/durational and spectral) at various linguistic levels [phrase: e.g., 2; word: e.g., 21; syllable: e.g., 3, 4, 13; segmental: e.g., 8, 18]. Estimated from trained language models, measures of contextual predictability can go beyond lexical frequency to capture the conditional probability of a target linguistic unit (operationalized in the current study as surprisal).

Although the effect of contextual predictability has been independently investigated on the incidence of a pause and the acoustic realization of duration, no study has attempted to bring them together and examine whether and how these two cues could be related to one another, when both are known to be subject to the influence of contextual predictability.

According to the Smooth Signal Redundancy hypothesis, predictability affects the acoustic realization of duration, with short duration for predictable information, in order to avoid an abrupt surge in information during the transmission of signals [e.g., 3]. Although Aylett and Turk [3] argued that prosody could mostly account for the predictability effect, Baker and Bradlow [5] provided evidence for other co-existing factors, such as speech styles and information structure that could also modify acoustic cues.

Further support of the latter came from a recent study showing the combined effects of syllable-based predictability (defined as surprisal) and Lombard style on syllable duration in German [13], and from an analysis of word-final syllable duration in a German radio corpus in [1]. In the latter, syllable duration increased from low-surprisal to high-surprisal syllables, but the surprisal-induced durational adjustment was more pronounced when the syllable occurred before a weak rather than a strong intonational phrase boundary (IP). That study operationalized the strength of the IP boundary in terms of pause duration: short pause duration implied weak IP, and long pause duration strong IP. The reported interaction of surprisal and IP boundary strength suggests a possible trading relationship.

However, that study in [1] did not examine cases at an intermediate phrase (ip) boundary. Moreover, the measured syllable duration did not control for the number of syllables in a word. Since a word-final syllable in polysyllabic words is less likely to be stressed or accented than in monosyllabic words, this could have biased the measured syllable duration. Furthermore, other factors such as information structure (e.g., information status), which is known to affect duration [e.g., 5], were not factored in.

If pause provides additional time for a speaker to prepare and encode what to say next, as assumed in previous research [e.g., 10, 11], it is more likely for a pause to occur in less predictable contexts (on the assumption that speech planning is not completed before speaking). If a pause is present, this raises the second question as to whether its presence might attenuate the surprisal-induced modification of syllable duration to reduce the abrupt surge in information during signal transmission, resulting in a possible trading relationship between pause usage or pause duration and syllable duration. The present study aimed at testing the effect of contextual predictability on the absence vs. presence of pause, and the trading relationship between the pause duration and the contextual predictability-induced syllable duration, while taking into consideration the number of syllables, prosodic boundary types, information status and accenting of the host word.

To investigate any trading relationships between pause (in terms of usage and duration) and word-final syllable duration in different surprisal contexts, at different prosodic boundary types and with different information status, three hypotheses were posited: (1) more incidence of pauses in a less than in a more predictable context; (2) shorter word-final syllable duration in the presence of a pause than in its absence because the incidence of a pause and syllable duration share the same burden of minimizing any abrupt surge in information arising from a high-surprisal syllable during signal transmission; (3) pause duration negatively varies with surprisal-induced syllable duration when a pause is present.

In addition, we also expected high surprisal contexts, an intonational phrase boundary and new information status to exhibit longer syllable duration than low surprisal contexts, an intermediate phrase boundary and given information status, respectively.

2. METHOD

Analysis was based on data from the DIRNDL corpus (Discourse Information Radio News Database for Linguistic analysis), a collection of 5-hour news recordings in German from 9 speakers (5M, 4F). The corpus contained orthographic transcription, labelled information structure (e.g., lexical information status) and syntactic constituents [9]. In addition, prosodic information covering pitch accent types and prosodic boundary were annotated in the GToBI(S) framework [19].

2.1. Data

The analysed data consisted of word-final syllables in monosyllabic and polysyllabic words immediately preceding an intermediate phrase (-) or intonational phrase (%) boundary in the DIRNDL corpus, resulting in a total of 2490 items.

2.2. Language Modelling

To estimate language-based surprisal values for word-final syllables in the DIRNDL corpus, language models were first constructed from the deWaC (**de**utsche **W**eb **a**s **C**orpus) corpus [7]. The corpus consisted of web-crawled data totalling about 1.7 billion word tokens and 8 million word types from different genres, for example, newspaper articles and chat messages. During the pre-processing stage German Festival was used to remove unnecessary or duplicate document information from the raw data [20]. The normalized data were then divided into a training set (80%) and a test set (20%). The best-performing language model was based on syllable-level trigrams, after training and evaluating different language models using the SRILM toolkit [23] with Witten-Bell smoothing [26]. Surprisal values were derived from the best-performing language model and defined as the conditional probability of a unit, given two preceding units including syllable and word junctures in (1) where S = surprisal and P = probability [12].

(1) $$S(unit) = -log_2 P(unit|unit_{-1}, unit_{-2})$$

2.3. Descriptive statistics of the data

Information status	Boundary	Pause (no) n	Pause (no) Mean (SD)	Pause (yes) n	Pause (yes) Mean (SD)
given	-	302	229 (75)	35	284 (97)
	%	9	277 (101)	140	249 (83)
new	-	989	243 (88)	74	296 (99)
	%	35	290 (102)	906	279 (106)

Table 1: Mean word-final syllable duration grouped by information status (given vs. new),

prosodic boundary types (intermediate [-] vs. intonational [%]) and pause occurrence (no vs. yes).

Table 1 summarizes the mean final syllable duration preceding an intermediate (-) or intonational (%) phrase boundary with lexically given or new information status when a pause is absent or present.

2.4. Statistical analysis

Syllable duration was extracted from the data using a custom Python script for analysis in R [22], using the lme4 package [6]. Surprisal values were log-transformed to adjust for skewness and then mean-centred. The statistical model included accenting (no vs. yes), information status (given vs. new), boundary type (intermediate vs. intonational), pause occurrence (absence vs. presence) and surprisal as predictors. Random structure of the model included by-speaker and by-item intercepts and by-speaker slope for predictors (subject to model convergence without overfitting). Significance testing for effects was evaluated using *F* values of type III anova which implemented Satterthwaite approximation for degrees of freedom [16]. The structure of the model was syllable duration ~ pause occurrence + boundary + information status + accenting + surprisal + information status * boundary + information status * boundary * accenting + information status * boundary * pause occurrence + boundary * pause occurrence * accenting + (pause occurrence + information status | speaker) + (1 | item).

3. RESULTS

To address hypothesis (1) regarding whether contextual predictability affects the incidence of pause, we first divided surprisal values into 2 groups with approximately equal number of observations in each (n=1246 with low surprisal and n=1244 with high surprisal); and conducted a two-proportions Z-test to compare the observed proportions of pause usage in low- vs. high-surprisal groups. The result revealed equal proportion of pause usage in low- vs. high-surprisal groups, with 562 pauses in the former and 593 in the latter (chi-squared = 1.5, df=1, p=.21).

To address hypothesis (2) as to whether a trading relationship exists between the incidence of pause and syllable duration, the lmer model results yielded significant effects of pause occurrence (F = 5.5, df = 1, p = .02*), information status (F = 8.92, df = 1, p = .004**), surprisal (F = 67.35, df = 1, p < .0001***), and accenting (F = 54.25, df = 1, p < .0001***), with two 2-way interactions: information status * boundary (F = 4.59, df = 1, p = .03*) and pause occurrence * boundary (F = 6.28, df = 1, p = .01*). No other effects or interactions reached statistical significance.

Counter to our prediction, the results did not exhibit a trading relationship between the incidence of a pause and the acoustic implementation of word-final syllable duration (Figure 1). In addition, the effect of pause occurrence on syllable duration was conditional on the types of the contiguous prosodic boundary. Word-final syllable duration was longer when a pause occurred at the intermediate phrase boundary (-) than when no pause was present (t = -4.2, df = 51, p =.0001***). Such a pause-related durational effect was not observed at the intonational phrase boundary (t= -.18, df = 258, p =.86).

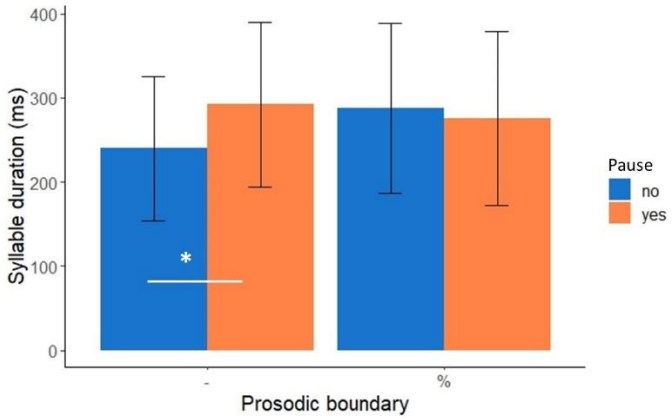


Figure 1: Mean syllable duration grouped by pause occurrence and prosodic boundaries with +/-1 SD.

Consistent with previous reports in [1], we observed a significant effect of surprisal. Information status interacted with prosodic boundary type to influence the acoustic realization of word-final syllable duration (Figure 2), with longer duration for items with lexically new than given information at an intonational phrase boundary (t=-3, df=367.3, p =.003**). Such a durational adjustment was not observed at an intermediate phrase boundary (t=-1.1, df=77.8, p =.27).

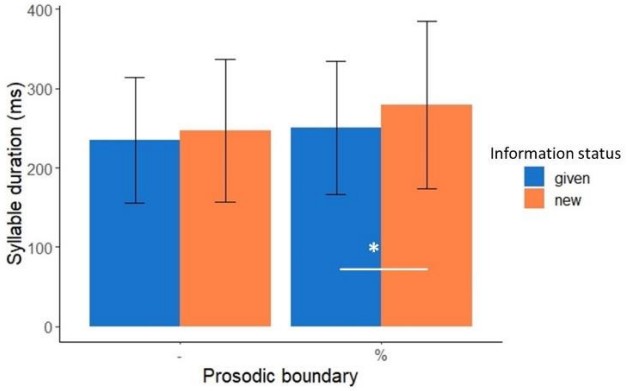


Figure 2: Mean syllable duration grouped by information status and prosodic boundaries with +/-1 SD.

To test for the existence of a trading relationship between the duration of a pause and that of word-final syllable when pause is present as stated in hypothesis (3), a series of correlations (Pearson) were carried out. Since accenting was a significant predictor in the lmer model results, correlations were conducted separately for accented and unaccented items. Figures 3 and 4 illustrate the correlation patterns for accented and unaccented items respectively. Given the relatively few tokens at an intermediate phrase boundary (-), correlations were carried out at an intonational phrase boundary. The results indicated a significant negative correlation between syllable and pause duration for accented items with new information ($t = -2.1$, $df = 209$, $p = .04$*). No other significant correlations were found.

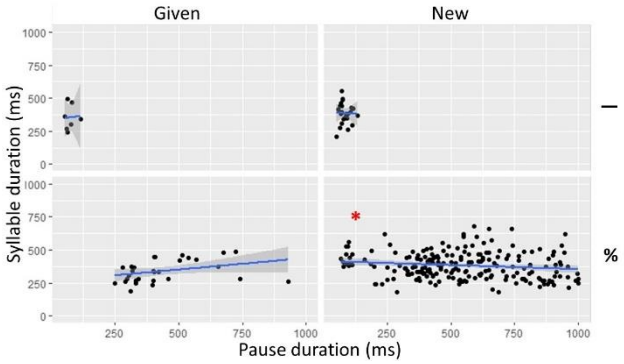


Figure 3: Scatterplots correlating pause duration (x-axis) and syllable duration (y-axis) among accented items, grouped by information status and prosodic boundaries.

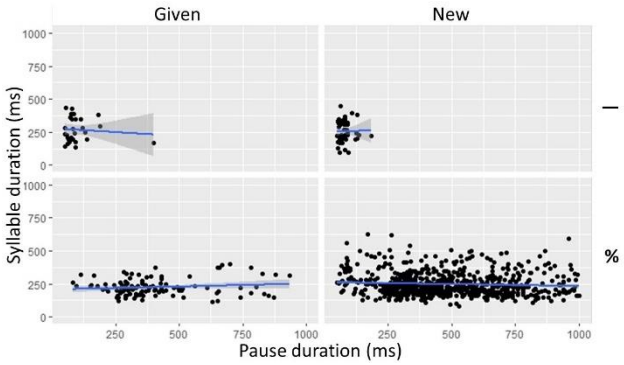


Figure 4: Scatterplots correlating pause duration (x-axis) and syllable duration (y-axis) among unaccented items, grouped by information status and prosodic boundaries.

4. DISCUSSION

Our results did not support hypothesis (1) revealing no trading relationship between the incidence of a pause and contextual predictability. This suggests that the incidence of a pause might not be related to contextual predictability in the current study of radio news which is not spontaneous speech.

Similarly, our results did not support hypothesis (2) either. There was no evidence for a trading relationship between the incidence of pause and word-final syllable durations. Word-final syllable durations were not shorter when a pause is present than when it is absent. Instead, syllable durations were long when a pause is present. Yet this pattern was found at an intermediate phrase boundary, not at an intonational phrase boundary. Since an intermediate phrase is part of an intonational phrase, more phonological planning might be needed before an intonational phrase is attained. This suggests that an intermediate phrase boundary might index a less certain location than an intonational phrase boundary. It is at such a less certain location that syllable duration was longer and the incidence of a pause more likely. This pattern (i.e., syllable duration and the incidence of a pause) might reflect the incremental nature of speech planning. That is, the phonological and phonetic specifications of a prosodic unit might not have been completed before speaking begins.

However, our results provided some evidence for hypothesis (3) that a trading relationship exists between the durations of a pause and a syllable when a pause is present for items at an intonational phrase boundary. When the duration of a word-final syllable is short, an accompanying pause duration is long. Such a trading relationship was observed for accented items, not unaccented items. The asymmetry could have arisen because the syllable duration without accenting is generally shorter than that with accenting. This durational difference might have constrained the manifestation of the trading relationship between syllable and pause duration.

5. CONCLUSION

In short, these results indicate that contextual predictability in terms of surprisal affects syllable and pause duration in a compensatory manner (i.e., in the form of a trading relationship) when a pause is present, although it does not affect pause usage (i.e., the incidence of pause). The existence of a trading relationship between pause and syllable durations suggests that speakers use these two temporal cues efficiently to alleviate any sudden surge in information density during transmission, without being redundant.

6. ACKNOWLEDGEMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 232722074-SFB1102.

7. REFERENCES

[1] Andreeva, B., Möbius, B., & Whang, J. 2020. Effects of surprisal and boundary strength on phrase-final lengthening. *Proceedings in 10th International Conference on Speech Prosody*.

[2] Arnon, I., & Cohen Priva, U. 2013. More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56, 349-371.

[3] Aylett, M., & Turk, A. 2004. The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence duration in spontaneous speech. *Language and Speech*, 47, 31-56.

[4] Aylett, M., & Turk, A. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei, *Journal of the Acoustical Society of America*, 119(5), 3048-3058.

[5] Baker, R. E., & Bradlow, A. R. 2009. Variability in word duration as a function of probability, speech style and prosody. *Language and Speech*, 52(4), 391-413.

[6] Bates, D., Mächler, M., Bolker B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1-48.

[7] Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora, *Language Resources and Evaluation*, 43, 209-226.

[8] Brandt, E., Möbius, B., & Andreeva, B. 2021. Dynamic formant trajectories in German read speech: Impact of predictability and prominence. *Frontiers in Communication*, 6, doi: 10.3389/fcomm.2021.643528.

[9] Eckart, K., Riester, A., Schweitzer, K. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In C. Chiarcos, S. Nordhoff & S. Hellmann (Eds.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, 65-75.

[10] Goldman-Eisler, F. 1958a. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology* 10, 96-106.

[11] Goldman-Eisler, F. 1958b. The predictability of words in context and the length of pauses in speech. *Language and Speech*, 1, 226-231.

[12] Hale, J. 2016. Information-theoretical complexity metrics. *Language and Linguistic Compass*, 10, 397-412.

[13] Ibrahim, O., Yuen, I., van Os, M., Andreeva, B., & Möbius, B. 2022. The combined effects of contextual predictability and noise on the acoustic realization of German syllables. *Journal of the Acoustical Society of America*, 162(2), 911-920. Doi:10.1121/10.0013413

[14] Krivokapić, J. Styler, W., Parrell, B. 2020. Pause postures: the relationship between articulation and cognitive processes during pauses. *Journal of Phonetics*, 79, doi: 100953.

[15] Krivokapić, J., Styler, W., Byrd, D. 2022. The role of speech planning in the articulation of pauses. *Journal of the Acoustical Society of America*, 151, 402-413.

[16] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest: Tests in linear mixed effects models: R package version 2.0-33

[17] Maclay, H., Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15, 19-44.

[18] Malisz, Z., Brandt, E., Möbius, B., Yoon, M. O., Andreeva, B. 2018. Dimensions of segmental variability: interaction of prosody and surprisal in six languages. *Frontiers in Communication/Language Sciences*, 3, 1-18.

[19] Mayer, J. 1995. Transcription of German Intonation – the Stuggart system. Institue of Natural Language Processing. *University of Stuggart Tech. Rep*. doi: doi:https://www.ims.uni-stuttgart.de/documents/arbeitsgruppen/ehemalig/ep-dogil/STGTsystem.pdf

[20] Möhler, G., Schweitzer, A., Breitenbücher, M., Barbisch, M. 2000. IMS German Festival (version 1.2-os). University of Stuggart: Institute für Maschinelle Sprachverarbeitung (IMS) retrieved 2020-01-02.

[21] Piantadosi, S. T., Tily, H. J. Gibson, E. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108, 3526-3529.

[22] R Core Team. 2022. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

[23] Stolcke, A. 2002. SRILM – an extensible language modeling toolkit. *Proceedings of Interspeech*, *Denver, Colorado.* vol. 2, 901-904.

[24] Swerts, M. 1997. Prosodic features at discourse boundaries of different strength. Journal of the Acoustical Society of America, 101, 514-521.

[25] Tisljár-Szabó, E., Pléh, C. 2014. Ascribing emotions depending on pause length in native and foreign language speech. *Speech Communication*, 56, 35-48.

[26] Witten, I. H., Bell, T. C. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37, 1085-1094.

[27] Yang, L. 2004. Duration and pauses as cues to discourse boundaries in speech. In Proceedings of Speech Prosody, pp. 35-48

VARIATION IN PERCEPTION AND PRODUCTION OF /s/-/ʃ/ IN ENGLISH

Roger Yu-Hsiang Lo[1], Suyuan Liu[1], Charlotte Vaughn[2, 3], Michael McAuliffe[4], and Molly Babel[1]

[1]Linguistics, University of British Columbia, Canada, [2]Language Science Center, University of Maryland, United States, [3]Linguistics, University of Oregon, United States, [4]Linguistics, McGill University, Canada

roger.lo@ubc.ca, suyuan.liu@ubc.ca, {cvaughn@umd.edu, cvaughn@uoregon.edu}, michaelemcauliffe@gmail.com, molly.babel@ubc.ca

ABSTRACT

In this exploratory study, we examine the relationship between perception and production in /s/ and /ʃ/ in North American English. To this end, peak ERB_N trajectories fitted with functional PCA characterize the fricative space on an individual level for 121 participants. Mahalanobis distances quantify the degree of category overlap and distance. Participants also completed a categorization task on /s/-/ʃ/ continua with four minimal pairs, the data of which were fit to a mixed-effects logistic regression. Individual logistic slopes are compared with the by-individual Mahalanobis distance, suggesting a subtle link between production and perception. An additional analysis suggests that the strength of the perception-production link is moderated by participants' multi-/monolingual status.

Keywords: sibilant fricatives, production, perception, phonetic variation

1. INTRODUCTION

This project began as a theoretical inquiry about perceptual learning with /s/ and /ʃ/. In discovering that listener samples from two universities had surprisingly different categorization thresholds for the /s/-/ʃ/ contrast, we decided to look at production as a source of the perception differences. That led to the current data, which are exploratory in nature.

Community production patterns often mirror community perception patterns. Spanish-speaking individuals, for example, have a /b/-/p/ perceptual category boundary that reflects the voice onset time (VOT) distribution in Spanish productions, exhibiting a lower VOT crossover point than English-speaking individuals, whose own threshold reflects English productions [1, 2]. On an individual level, however, the relationship between production and perception is more variable. While there is supporting evidence for some type of connection between perception and production, it is not a straightforward reflex [3, 4, 5, 6]. We structure our exploration of perceptual category boundary variation and acoustic-phonetic production variation in the context of this body of literature.

We focus on /s/ and /ʃ/ in North American English, first characterizing the degree of acoustic-auditory overlap on a by-talker basis using functional Principal Components Analysis (FPCA) on fricative trajectories. We use a large number of voices (n = 121) producing a large number of real words (n = 40). We consider the relationship between one's own realization of the /s/-/ʃ/ contrast in production and one's categorization function in recognition of /s/-/ʃ/ minimal pairs. Our focus on word recognition, as opposed to lower-level perceptual acuity [7], connects to individuals' production-perception link at the word level [8, 9]. Because our investigation is exploratory in nature, we take the opportunity to consider how bi-/multilingualism may affect the link between perception and production, theorizing about the impacts of multilingualism on the perception and production relationship in Section 3. The nature of our data set ultimately allows the larger project to also serve as an opportunity to replicate the seminal work of [10].

2. EXPERIMENTS

2.1. Participants

Participants recruited at the University of British Columbia (UBC) and the University of Oregon (UO) were given partial course credit or cash for their time. Data from 121 participants are included in the current analysis. Participants had heterogeneous language backgrounds that are are representative of the speech communities under study, and all acquired English before the age of five.

2.2. Production: Materials and Methods

First, participants read 252 single words, including 100 /s/-words, 47 /ʃ/-words, and 105 filler words not containing sibilant sounds. Words containing

sibilants were either bi- or tri-syllabic. Of the 100 /s/-words, 50 words had /stɹ/ clusters, and 50 were /sV/ words. Of the 47 /ʃ/-words (all /ʃV/), 20 had /ʃ/ word-initially and 27 had /ʃ/ word-medially. Participants had the option to repeat words if they wished, and only the final repetition was saved. To match the perception data, only items with fricatives in initial position and with following vowels (e.g., not the /stɹ/ clusters) are analyzed in this paper ($n = 40$).

Production tokens were discarded if they contained incorrect pronunciations or stress, or any disfluencies. Productions were force aligned using the Montreal Forced Aligner [11]. Following alignment, word and fricative boundaries were hand-corrected to ensure accuracy.

2.3. Perception: Materials and Methods

Perception stimuli came from a previous study [12]. An 11-step continua between monosyllabic /s/-/ʃ/ minimal pairs (*sack-shack*, *sigh-shy*, *sin-shin*, *sock-shock*) was created using Tandem-STRAIGHT [13]. A white college-aged English-L1 male from UBC produced the endpoints. Pretests were conducted at both testing locations (UBC and UO) to find the center portion of each continuum. The categorization for listeners at location UBC was shifted a half to a full step more towards /s/ compared to listeners at location UO. A 7-step continuum centered around each location's midpoint was presented to listeners at that location. Each step of each continuum was repeated 7 times, giving a total of 196 trials (7 steps $\times$ 7 repetitions $\times$ 4 continua) per participant. For each trial, the listener categorized the word as either the /s/-word (e.g., *sin*) or the /ʃ/-word (e.g., *shin*) of the minimal pair.

2.4. Analysis and Results

The goal of the analysis is to derive metrics to quantify the degree of contrast between /s/ and /ʃ/ in both production and perception, and to examine the correlation across modalities. Mahalanobis distance [14] was used for production and the logistic beta-coefficient for perception to quantify the degree of contrast. It should be noted that, despite analyses on production and perception data being described separately, the degree of contrast across the two modalities was assessed in a single Bayesian model to account for uncertainty surrounding estimates. All analyses were performed in R [15], with Bayesian models fitted using `CmdStanR` [16].

2.4.1. Production

Using the script developed in [17], a trajectory of peak ERB_N numbers, henceforth peak ERB_N, which represent frequencies with the greatest amount of excitation on the multitaper psycho-acoustic spectrum calculated over a sequence of 20-ms windows, was extracted over the duration of each fricative with a 10-ms stride.[1]

Tokens with fricative duration shorter than 60 ms and those with within-token standard deviation in peak ERB_N beyond the 0.95 quantile were removed. After removal, 5,199 tokens entered the analysis.

We analyzed peak ERB_N trajectories with FPCA, following the steps laid out in [18]. FPCA offers a way to succinctly quantify global variation in peak ERB_N. In essence, FPCA treats each peak ERB_N contour as a composition of a grand mean curve and a small number of principal component (PC) curves (each of which encodes a different deformation of the mean curve) weighted by corresponding PC loadings, which can be studied using conventional statistical tests. Here we focus on the first three PC components (i.e., PC1, PC2, and PC3), which jointly explain 96.5% of the variance and whose effects on the grand mean peak ERB_N curve are shown in Figure 1. PC1 chiefly controls the height of overall trajectory, PC2 coordinates the extent of dipping/rising at both ends of the curve, and PC3 alters the direction of curvature. PC1 alone accounts for 89.3% of the variance and serves to bipartite the two fricatives, though to different degrees across participants, as shown in Figure 2**A**. To more holistically capture this degree of contrast, the Mahalanobis distance, which can be understood as a generalization of Cohen's *d* [19] to higher dimensions, was estimated for each individual based on the loadings of their first three PCs. The distribution of individual (posterior mean) Mahalanobis distances is shown in the top histogram of Figure 2**C**, and individuals with the largest as well as smallest Mahalanobis distance are also featured in Figure 2**A**. These results indicate that individuals do come with different degrees of contrast, mainly in terms of the height proximity of peak ERB_N trajectories of the two fricatives.

2.4.2. Perception

We excluded perceptions trials where no response was registered. The responses of the remaining trials were modeled with a mixture model: one component was a mixed-effects logistic regression that had **step** as the fixed effect as well as by-participant and by-word random intercepts and slopes, and the other

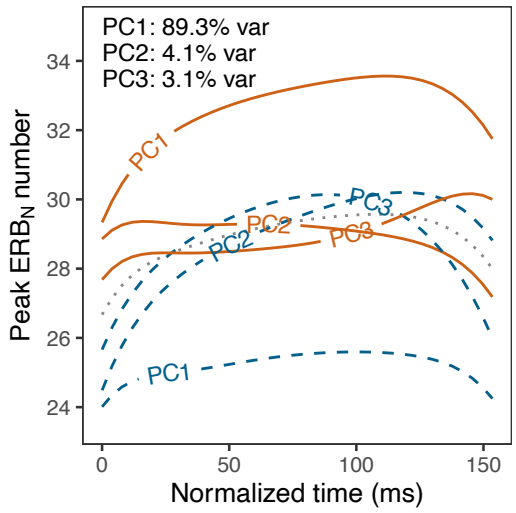


Figure 1: PCs of peak ERB_N trajectories. The dotted line represents the grand mean curve $\mu(t)$ while the solid/dashed curves are obtained by adding (orange) or subtracting (blue) from $\mu(t)$ the curve $\sigma(\text{PC}_i \text{ loadings}) \cdot \text{PC}_i$, where σ denotes the standard deviation, and $i = 1, 2, 3$.

was sheer randomness due to the participant making accidental button-push errors. That is, each response had a chance, γ, of being generated by a random process, and a probability of $1 - \gamma$ that it came from the logistic function of the **step** variable.

The only independent variable—**step**—was treated as a continuous variable and z-transformed with respect to the original manipulation steps (e.g., Step4 was consistently mapped to 0, regardless of listeners). The default level for the response was /s/-word (i.e., /s/-word responses were coded with 0, and /ʃ/-word responses were coded with 1), so a positive $\beta_{\textbf{step}}$ means that larger step values elicit more /ʃ/-word responses. All individuals demonstrate a clear sigmoid function, suggesting a robust perceptual contrast for /s/ and /ʃ/. However, as shown in the right histogram of Figure 2**C** and highlighted in Figure 2**B**, individuals also vary with respect to the slope of their sigmoid function. Overall, these results speak to the existence of cross-listener variation in the "crispness" of the boundary between the two perceptually robust fricatives [20].

2.4.3. Production-Perception Link

Finally, we regressed individual logistic slopes in perception on individual Mahalanobis distances in production to examine the correlation across the modalities. An initial analysis of all participants revealed a positive, but negligible, relationship ($\text{mean}_{\text{corr.}} = .053$, 95% credible interval (CrI) = $[-.049, .186]$, $p(\text{corr.} > 0) = .85$), as depicted in Figure 2**C**. An exploratory analysis separating monolingual and bi-/multilingual individuals found a consistent positive relationship for the English monolingual group ($\text{mean}_{\text{corr.}} = .118$, 95% CrI = $[-.008, .254]$, $p(\text{corr.} > 0) = .96$) and a more variable pattern for the bi/multilingual individuals ($\text{mean}_{\text{corr.}} = -.010$, 95% CrI = $[-.135, .112]$, $p(\text{corr.} < 0) = .57$). The resulting correlations, separated by language background, are visualized in the scatter plot of Figure 3.

3. DISCUSSION AND CONCLUSION

Our findings echo prior work demonstrating that individuals vary in the distinctiveness of /s/ and /ʃ/ in North American English [10]. Talkers exhibited considerable variation in the degree of acoustic-auditory contrast in production, ranging from highly separable /s/ and /ʃ/ categories and those with considerable overlap. Similarly, in perception, while all individuals perceived a contrast, there was variation in the slope of the curve, indicating variation in the perceptual distinctiveness of the two categories. We used these data to explore the within-individual production-perception link. That there is a relationship between perception and production is at the heart of many models of sound-based knowledge (e.g., exemplar-based models [21] and motor-theory [22]), in addition to being a core mechanism for contemporary theorizing about sound change [3]. Considering the dataset as a whole, we did not find an overall strong correlation between production and perception.

We divided our participants into two groups, separating individuals who are monolingual and bi-/multilinguals. The initial observed lack of a strong relationship holds for the bilingual group and a suggestive link emerges for the monolinguals. A weak or nonexistent correlation between bilinguals' perception and production was also attested in [23]. Specifically, that study found a strong correlation between English monolinguals' production and perception, but the relationship did not hold for their Spanish-English bilingual group. The bilingual group produced the contrasts in production, indicating phonetic sensitivity in that modality; [23] reason that the conflicting phonological status across English and Spanish contorts the perception space. Similarly, while [24] finds that Spanish-Catalan bilinguals maintain the Catalan mid-vowel contrasts in production, their listeners find the distinction more challenging in perception and may not robustly encode the contrasts at the lexical level; see also [25].

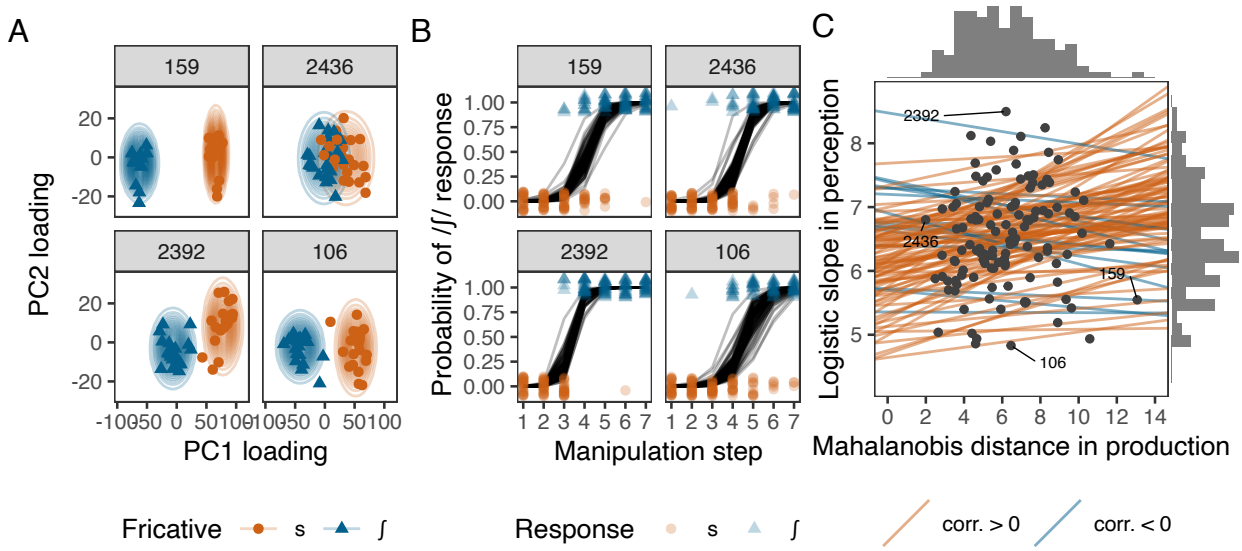


Figure 2: Results of production and perception analyses. **(A)** Distributions of PC loadings along the first two PCs. Each circle/triangle represents a token. The fitted multivariate Gaussian distributions for estimating Mahalanobis distance are shown as density contours. **(B)** Response distributions along the manipulation step continuum. Each jittered dot/triangle marks a response, and the line bundles visualize uncertainty associated with fitted logistic curves. The top row in both figures shows the individuals with the most peripheral degree of contrast in production, and the bottom row shows those with the most peripheral degree of contrast in perception. **(C)** Degree of contrast across production and perception on an individual level. The dots represent the posterior means for individual participants. The solid lines come from linear regressions fitted to 100 posterior draws, to show the direction and uncertainty of the correlation.

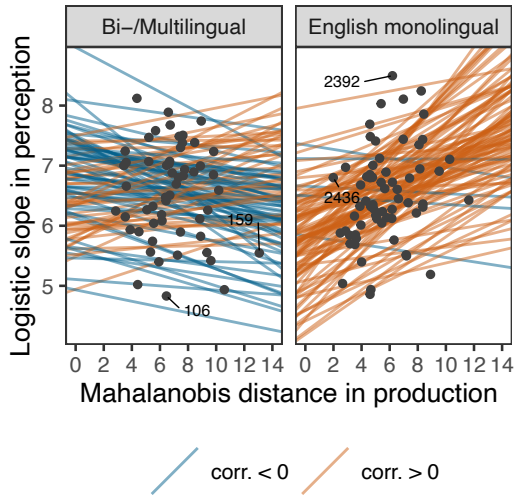


Figure 3: Degree of contrast across production and perception on an individual level with panels by language background.

Why might the perception-production link be absent for bi-/multilingual speakers? It is not due to an across-the-board lack of a relationship in bilinguals [9, 8]. Crosslinguistic influence in bilingual linguistic systems may affect both production and perception [26]. How observable that influence is varies as a function of language dominance and language mode (e.g., [2]). As alluded to above, however, while early bilinguals may maintain distinctions in production and phonetic level discrimination, bilinguals may experience competition in categorization and recognition that masks the cross-modal link or impedes the encoding of phonetic detail in the lexicon. Additional theorizing about the nature of the perception-production link for bilinguals is necessary, and it may shed light on the variable nature of the perception-production connection in monolingual listeners as well (e.g., [27]).

To summarize, this project observed individual variation in both production and perception of North American English /s/ and /ʃ/, showing a positive correlation between the two modalities for monolingual, but not bi-/multi-lingual speakers. Our future work will explore the consequences of this production variation in perception and recognition for listeners, as a conceptual replication of [10].

4. ACKNOWLEDGMENTS

Thank you to Stephanie Chung, Yu Cen Gao, Sam Elliott, Celina Maldonado, Omar Ortiz, and Blair

Prater for their assistance with this work. Funding for this work has come from a SSHRC Grant to Molly Babel.

5. REFERENCES

[1] A. S. Abramson and L. Lisker, "Voice-timing perception in spanish word-initial stops," *JPhon*, vol. 1, no. 1, pp. 1–8, 1973.

[2] J. V. Casillas and M. Simonet, "Perceptual categorization and bilingual language modes: Assessing the *double phonemic boundary* in early and late bilinguals," *JPhon*, vol. 71, pp. 51–64, 2018.

[3] P. S. Beddor, A. W. Coetzee, W. Styler, K. B. McGowan, and J. E. Boland, "The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change," *Language*, vol. 94, no. 4, pp. 931–968, 2018.

[4] L. S. P. Cheng, M. Babel, and Y. Yao, "Production and perception across three Hong Kong Cantonese consonant mergers: Community-and individual-level perspectives," *LabPHon*, vol. 13, no. 1, 2022.

[5] J. Kuang and A. Cui, "Relative cue weighting in production and perception of an ongoing sound change in Southern Yi," *JPhon*, vol. 71, pp. 194–214, 2018.

[6] C. C. Voeten, "Individual differences in the adoption of sound change," *Language and Speech*, vol. 64, no. 3, pp. 705–741, 2021.

[7] J. S. Perkell, F. H. Guenther, H. Lane, M. L. Matthies, E. Stockmann, M. Tiede, and M. Zandipour, "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," *JASA*, vol. 116, no. 4, pp. 2338–2344, 2004.

[8] S. Cheung and M. Babel, "The own-voice benefit for word recognition in early bilinguals," *Frontiers in Psychology*, vol. 13, pp. 1–14, 2022.

[9] N. A. Eger and E. Reinisch, "The impact of one's own voice and production skills on word recognition in a second language," *JEXP: LMC*, vol. 45, no. 3, pp. 552–571, 2019.

[10] R. S. Newman, S. A. Clouse, and J. L. Burnham, "The perceptual consequences of within-talker variability in fricative production," *JASA*, vol. 109, no. 3, pp. 1181–1196, 2001.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *INTERSPEECH 2017*, 2017, pp. 498–502.

[12] M. McAuliffe and M. Babel, "Stimulus-directed attention attenuates lexically-guided perceptual learning," *JASA*, vol. 140, no. 3, pp. 1727–1738, 2016.

[13] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3933–3936.

[14] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 49–55, 1936, p. 1.

[15] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[16] J. Gabry and R. Češnovar, *cmdstanr: R interface to 'CmdStan'*, 2021, https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

[17] P. F. Reidy, "Spectral dynamics of sibilant fricatives are contrastive and language specific," *JASA*, vol. 140, no. 4, pp. 2518–2529, 2016.

[18] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *JPhon*, vol. 49, pp. 16–40, 2015.

[19] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.

[20] G. S. Morrison, "Logistic regression modelling for first- and second-language perception data," in *Segmental and prosodic issues in Romance phonology*, P. Prieto, J. Mascaró, and M.-J. Solé, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 219–236.

[21] M. Walsh, B. Möbius, T. Wade, and H. Schütze, "Multilevel exemplar theory," *Cognitive Science*, vol. 34, no. 4, pp. 537–582, 2010.

[22] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[23] P. Piccinini and A. Arvaniti, "Dominance, mode, and individual variation in bilingual speech production and perception," *Linguistic Approaches to Bilingualism*, vol. 9, no. 4-5, pp. 628–658, 2019.

[24] M. Amengual, "The perception and production of language-specific mid-vowel contrasts: Shifting the focus to the bilingual individual in early language input conditions," *International Journal of Bilingualism*, vol. 20, no. 2, pp. 133–152, 2016.

[25] R. Soo and P. J. Monahan, "Phonetic and lexical encoding of tone in Cantonese heritage speakers," *Language and Speech*, 2023.

[26] M. Fricke, J. F. Kroll, and P. E. Paola, "Phonetic variation in bilingual speech: A lens for studying the production-comprehension link," *Journal of Memory and Language*, vol. 89, pp. 110–137, 2016.

[27] A.-F. Pinget, R. Kager, and H. V. de Velde, "Linking variation in perception and production in sound change: Evidence from Dutch obstruent devoicing," *Language and Speech*, vol. 63, no. 3, pp. 660–685, 2020.

[1] In [17], 17 equal-distance peak ERB_N were extracted for each trajectory. We also performed analyses on the 17-point data and obtained very similar results.

UNRELEASED PLOSIVE CONSONANTS IN THAI: AN ACOUSTIC STUDY

Paula Alejandra CANO CÓRDOBA[1], Thi Thuy Hien TRAN[1], Nathalie VALLÉE[1], Christophe SAVARIAUX[1], Silvain GERBER[1], Nicha YAMLAMAI[2], Morgane ROSSIGNOL[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[2] Faculty of Arts, Univ. Silpakorn, Bangkok, Thailand
{paula.cano-cordoba, thi-thuy-hien.tran, nathalie.vallee, christophe.savariaux}@gipsa-lab.grenoble-inp.fr;
yamlamai_n@su.ac.th; morgane.rossignol@etu.univ-grenoble-alpes.fr

ABSTRACT

Stops are consonants usually produced with an audible noise generated by the sudden release of the complete closure of the vocal tract. However, published studies have reported that in Thai, syllable-final unvoiced stops are frequently produced without an audible release. Our aim is to provide acoustic details of such consonantal productions. By analysing acoustic data from 10 native Thai speakers and by focusing on vowel-consonant transitions, we propose a set of acoustic cues on which listeners might rely on to discriminate and categorize the unreleased stop allophones. Our study also investigates whether acoustic properties of non-released stops differ depending on whether they are at the end of monosyllables or at the end of the penult syllable of disyllabic compounds (end-of-word coda *vs* word-internal syllable coda). Our findings show consistent differences in stop durations, formant transitions, and intensity curves according to place of articulation, within-word position and speaker gender.

Keywords: unreleased final stop, Thai, syllable boundary, VC transition, acoustic cues

1. INTRODUCTION

Plosive consonants occur in all languages [1], [2]. They belong to the category of stops which are produced with a total blockage of the airstream in the vocal tract. Because of the complete closure of the vocal tract and the increase in pressure behind the closure, the sudden release of the airstream generates an abrupt and salient noise [3] which is called burst [4]. The burst contains acoustic information that participates to the consonant identification [5], [6]. Nevertheless, in some languages such as Korean [7], Vietnamese [8] and Thai [9] [10], there is no audible release when plosives are syllable ending consonants. The syllable-final position is particularly subject to articulatory weakening and acoustic change or lenition [11]. Some articulatory weakening processes result in phonological neutralizations leading to a restriction of syllable-final consonantal inventories. This is the case in some East or Southeast Asian languages, as for example in Korean or Vietnamese. In Korean, the three-way manner contrast /t/ (lenis) ~ /t*/ (fortis) ~ /t^h/ (aspirated) is neutralized in coda position and has the same surface phonetic representation [t̚], an unvoiced and unreleased stop [12]. In Vietnamese, [t̚] is an allophone of /t/ and /d/ whose contrast is no longer maintained at syllable end [13]. Thai is phonologically classified as a monosyllabic language, but it is also partially polysyllabic lexically since 35% of Thai words are two-syllable semantic compounds, and 25% are words longer than 2 syllables [14]. Thai has 21 phonemic consonants. Among them there is a set of 8 plosives that can all appear in syllable-initial position /p, p^h, b, t, t^h, d, k, k^h/ contrasting 3 places of articulation (bilabial, coronal, velar) and 3 laryngeal modes (voicing (except for the velar place), non-voicing and aspiration) [15]. However, due to a neutralization process which gives rise to onset-coda asymmetries in Thai [16], only 3 surface phonetic plosives [p̚, t̚, k̚] described as unvoiced and unreleased are found in syllable final position [10].

Several studies have shown interest in unreleased stops mainly because native speakers of these languages are able to easily identify places of articulation despite the absence of burst [8], [17], [18]. This suggests that other acoustic cues than audible release provide robust information for successful perception and categorization of syllable-final plosives. For example in Vietnamese, [19] show that spectral features of the second half of vowel duration contain information on coda's place of articulation, which can guide listeners' processes of identification of unreleased stops. [20] observed in Thai, like [19] in Vietnamese, that vowel duration is influenced by the place of articulation of the syllable-final plosive. In a cross-language perception study, [18] showed that Thai speakers perform better than native English speakers in the identification of Thai and Korean unreleased plosives. Furthermore, in Vietnamese, acoustic features of the syllable's rhyme shows differences depending on whether the unreleased plosive is an end-of-word coda or a word-internal syllable coda [19].

The purpose of our study is to highlight acoustic cues of unreleased plosives in Thai in examining more specifically duration and spectral characteristics of vowel-consonant (VC) transitions as in previous studies cited above.

2. HYPOTHESES

Based on results of published studies mentioned in Section 1, we suggest that acoustic differences in VC transitions in Thai, where C is an unreleased plosive, existed (1) according to the place of articulation of the plosive, and (2) depending on whether the unreleased stop is in word-medial (i.e. coda of the first syllable of a dissyllabic word) or in word-final position (in case of monosyllabic word). We target the first three formants ($F_{1\text{-}3}$), the fundamental frequency (f_0), intensity (I), and the acoustic duration of the rhyme.

3. METHODOLOGY

3.1. Corpus and speakers

Thirty-four Thai words, 17 C_1aC_2# monosyllables and 17 $C_1aC_2.C_3VC$# dissyllables (the dot indicates the syllable boundary), were selected such as each target syllable occurred as a monosyllabic word and as the first syllable in a dissyllabic word. All target syllables carried the high rising tone. For compounds' second syllable, tones are chosen (high rising for syllables with final /p t k/, mid falling for those with final sonorants) to minimise the tonal frequency gap. The low vowel /a/ was chosen for better contrast with the target stops. C_2 could be /p/, /t/ or /k/. Consonants in C_3 were chosen voiced and could be /l/, /d/, /j/, /b/ or /w/ to aid C_2 segmentation, except for ซับซ้อน [sap1.sɔ:n1] *'complicated'* (no words with a voiced C_3 were found with [sap1] in the Thai lexicon, due to the high rising tone constraint). Finally, the 34 selected words were inserted into the carrier sentence พูดว่า __ ดีๆ [pʰu:tꜛ wa:ꜛ __ di:ꜜ di:ꜜ] *'Say __ carefully'*.

Data from five male and five female native Thai speakers were analysed. The participants were from the Bangkok area and spoke the central Thai dialect. At the time of the study, they were 20 years old and were all third-year students in the French Department at Thammasat University (Rangsit Campus). During the experiment, the participants were seated in front of a 22" screen on which the sentences were presented to them one by one. They were instructed to read them aloud at a normal and fluent rate. The recording took place in a sound proof-room, with a Marantz PMD 670 recorder, an AKG C-1000S microphone. The corpus was digitized at 44.1 kHz on 16 bits mono in .wav format. The results presented below are from 34 target words ∗ 3 repetitions ∗ 10 speakers giving 1,020 occurrences in C_2 of the syllable-final target plosives (300 /p/, 360 /t/, 360 /k/).

3.2. Measurements and analysis

The acoustic signal was manually segmented and annotated with Praat [21] in terms of phonemic, phonetic and sub-phonetic labels. When C_2 did not correspond to a production of the phoneme, it was discarded. In total, 9 realizations were excluded (4 for /k/ and 5 for /t/). Moreover, we detected a burst in 19.7% of the plosives produced, of which 9.7% were in word-final position, and 10% were in word-internal syllable coda. These productions were not included in our analyses.

Measurements of the following acoustic parameters were performed automatically using scripts we wrote in Praat software:

- Duration of the coda C_2 and duration of the preceding vowel /a/;
- Temporal evolutions of the first three formants (ΔF_1, ΔF_2, ΔF_3), of the fundamental frequency (Δf_0), and of the intensity (ΔI) measured from the time points of 40% and 50% of the whole vowel duration to the time points of 60%, 70%, 80%, 90% of the vowel length (which corresponds to the acoustic transition between /a/ and the stop closure).

Before proceeding to the statistical analyses, all duration measurements were normalized to the speaker's speech rate (in syllables per minute) following the method used by [19].

3.3 Statistical Analyses

Statistical analyses were performed using several mixed linear models (LMMs) from R software to estimate, on one side, the effects of consonant place, time-point, burst, and word type on ΔF_1, ΔF_2, ΔF_3, Δf_0, and ΔI mean values for each gender, and, on the other side, the effect of consonant place, burst and word type on duration, without distinction between gender. The *lme* function of the *nlme* package was used to construct LMMs. Once the model was established, contrast analyses were performed with the *glht* function of the *multcomp* package, according to the method presented by [22]. Multiple comparisons were performed for the dependent variables 'formants', 'f_0', and 'intensity'. We targeted the following differences:

- Between consonants (/p/, /t/, /k/) for each combination (word type - time point);
- Between the two-word types (simple and compound) for each combination (consonant - time point);
- Between two successive time points (e.g. between 50% and 60%, 60% and 70%, etc.) for each combination (consonant - word type).

4. RESULTS

In terms of statistical significance, the factor place of articulation has the strongest effect on the dependant variables, except on Δf_0. This important and general

result is in agreement with what had been reported in Vietnamese by [19] and is consistent with the general proposal that in case of unreleased stops, the perceptual-acoustic cues for identifying the place of articulation that enables listeners to categorize the consonant are located in the transient part between the vowel and the unreleased plosive [23].

4.1. Acoustic Cues for Articulation Places of Unreleased Final Consonants

An effect of place of articulation is present in segment duration regardless of word-type: [k˺] is in average longer (21.3ms in C_2#, 17.9ms in $C_2.C_3$) than [p˺] (15.8ms in C_2#, 15.2ms in $C_2.C_3$) and [t˺] (19.8ms in C_2#, 15.8ms in $C_2.C_3$). However, significant differences are only found for /k/ *vs* /p/ in C_2# [z=3.5; p<0.01] and for /k/ *vs* /t/ in $C_2.C_3$ [z=3.3; p<0.05]. These observations are similar to those made on Vietnamese by [19] who found a significantly longer duration of [k˺] relative to [p˺] in monosyllables, and to [t˺] in dissyllables. However, unlike [19], no significant differences in vowel duration were found in relation to the place of articulation of coda consonants in Thai.

Also, formant transitions show a consonant articulation place effect. F_1 values go down more for /t/ than for the bilabial and velar stops. The difference is found significant in monosyllables for the male speakers (/t/ *vs* /k/ from 80% [z=4.1; p<0.01] to 90% [z=3.9; p<0.05] and /t/ *vs* /p/ at 90% [z=4.2; p<0.01]), and in dissyllables for the female speakers (/t/ *vs* /k/ from 50% [z=3.6; p<0.05] to 90% [z=6; p<0.01]).

Three specific directions of F_2 transitions, which are consistent between speaker genders and word-types, are noted according to the stop place of articulation (Fig. 1 and 2). Differences are significant for /p/ *vs* {/t/, /k/}: /p/ *vs* /t/ from 60% and /p/ *vs* /k/ from 70% of the vowel duration, for both female and male speakers (p<0.01) and whatever the stop position in coda (C_2# or $C_2.C_3$) (p<0.01). Concerning /t/ *vs* /k/ differences, a significant F_2 deviation is found in monosyllables from 70% of the vowel duration for the male [z=4.3; p<0.01] and female speakers [z=-4.6; p<0.01], whereas in compound words, the deviation of F_2 is significantly different only for the female speakers at 90% of the vowel duration [z=-4.2; p<0.01]. F_2 transition for /k/ shows relatively stable values independently of speakers and word-types. However, it should be noted that we have not observed the expected theoretical ascending F_2 transition for /a/ with a rather steeper slope in velar context as proposed by [24] from synthetic speech.

Regarding F_3 transitions, significant differences are present between /t/ *vs* {/p/, /k/} for both female and male speakers, but with different parts of evolution depending on word-type. For female speakers, as proposed by [24] with synthetic speech, an ascending F_3 evolution in the second half of the vowel duration is noted for the coronal stop (from 3 to 93Hz), significantly attested from 60% until the end of the vowel in C_2# and from 80% in $C_2.C_3$. For male speakers, ΔF_3 of /t/ has a slight increase (C_2#: from 6 to 37Hz; $C_2.C_3$: from 6 to 26Hz). ΔF_3 of /t/ differs significantly from /p/ from 60% of the vowel duration independently of word-type (p<0.01), and deviates from /k/ remarkably from 60% to 80% in C_2# (p<0.05), but only at 60% in $C_2.C_3$ position [z=-3.5, p<0.05]. A difference in F_3 evolution between /p/ and /k/ is significantly attested at only 90% of the vowel duration for male speakers in compound words.

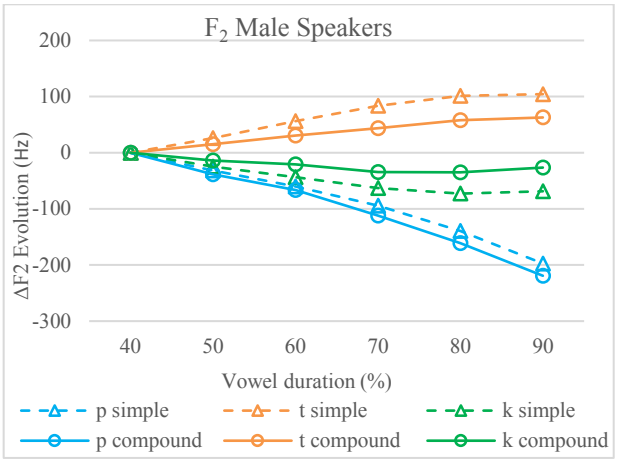


Figure 1: Temporal evolution of ΔF_2 estimated from 40% to 90% of the vowel duration (expressed in Hz) for male speakers as a function of word type (simple or compound).

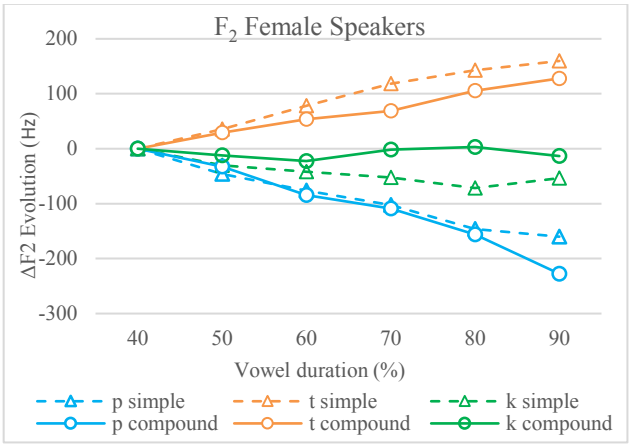


Figure 2: Temporal evolution of ΔF_2 estimated from 40% to 90% of the vowel duration (expressed in Hz) for female speakers as a function of word type (simple or compound).

No significant effect of place of articulation was found on Δf_0 values for both genders. This result is inconsistent with the observation of [19] on Vietnamese in which a significantly increase in Δf_0

from 70% until the end of the vowel was found before a non-released velar coda.

Our results also show that place of articulation impacts the temporal evolution of intensity. For all speakers, intensity significantly slopes down more for /k/ than for /p/ ($p<0.01$) and /t/ (♀: $p<0.05$; ♂: $p<0.01$) in $C_2.C_3$ from 70% of the vowel duration, and in C_2# for ♀ from 80% ($p<0.01$) and for ♂ only at 90% ($p<0.01$). For both word-types, the drop of ΔI for /k/ is significant from 40% until vowel endings ($p<0.01$). These results for ΔI are in agreement with what was observed in Vietnamese by [19].

4.2. Effect of Word Type on Final Stop Production

Analyses of our data show that the unreleased coronal stop /t/ is significantly longer in C_2# (20ms on average) than in $C_2.C_3$ (12ms) [$z=-3.6$, $p<0.01$], regardless of the speakers' gender. /k/ and /p/ are also longer in word-final position C_2# but the difference according to word types is not statistically proven (respectively $p=0.6$ and $p=1$). The duration of /a/ is also significantly different depending on whether the consonant is in C_2# or in $C_2.C_3$: /a_p/ (14.5ms *vs* 11.9ms, $z=-3$, $p<0.05$), /a_t/ (15.8ms *vs* 11.6ms, $z=-5$, $p<0.01$), or /a_k/ (16.2ms *vs* 12.8ms, $z=-3.8$, $p<0.01$). These findings partially confirm results of a previous two-speaker study on Thai [20], in which difference in consonant duration as a function of word type was found for stops.

An effect of word-type is found on F_1 transition but is different according to speaker gender. We observed in male speakers that, relative to 40% of the vowel duration (reference point), ΔF_1 of /p/ declines significantly more at 60% (-13Hz) in single words than in compound words (-9.6Hz) [$z=-3.7$, $p<0.05$], whereas in female speakers, ΔF_1 of /t/ decreases further in compounds from 50% (-13Hz) to 70% (-70Hz) of the vowel duration than in simple words (respectively from 1Hz to -49Hz) [$z=-4.2$; $p<0.01$]. These results are partially consistent with [19]s' findings for Vietnamese, which show that ΔF_1 drops significantly when the unreleased stop is an inner final syllable ending ($C_2.C_3$) rather than a word ending (C_2#). Unlike to F_1 transition, we found no effect of word-type on F_2 and F_3 transitions.

Our statistical analyses show significant differences for Δf_0 as a function of word type but only for female speakers. Indeed, Δf_0 curve decreases during the entire second half of the vowel duration for /t/, more in compound (from -1.7Hz to -6Hz) than in simple words (from -1.3Hz to -2Hz) ($p<0.01$). For /p/, Δf_0 curve is downward in compound words (from -1Hz to -2.5Hz) whereas it is slightly upward in monosyllables (from -1Hz to -0.06Hz). This Δf_0 difference is shown to be significant for /p/ ($p<0.01$). These results for the female speakers are consistent with those found in Vietnamese [19] which also reveal differences in the f_0 evolution from 60% until the end of the vowel depending on whether it is a simple word or a first syllable in a compound word. In terms of intensity, we did not observe an effect of word type on temporal evolution of ΔI that is partly inconsistent with what was observed in Vietnamese for which the intensity drop at 90% of the vowel duration is less important before a word-final stop than before a syllable-final within-word stop [19].

5. CONCLUSION

Knowing that burst helps listeners to discriminate the plosives of a system [25], and that native speakers are able to correctly identify a consonant even in the absence of its audible release [8] [17] [18], this study investigated acoustic cues of unreleased final stops in Thai in order to highlight perceptual-acoustic properties that allow listeners to categorize the unreleased allophones of stops. Based on a previous study [19], we also investigated effects of gender and of word type (simple monosyllables *vs* dissyllabic compounds) on acoustic cues. We analysed, as in [19], the duration of three target stops /p, t, k/ and of the preceding vowel /a/, as well as the temporal evolutions of I, f_0, and $F_{1\text{-}3}$ by considering place of articulation, word type and gender as independent variables.

Our findings confirm that differences in duration of unreleased stops and in other acoustic parameters of VC transitions (i.e. F_1, F_2, I, and in a lesser extent F_3) according to place of articulation are present in Thai. F_2 transition show three specific temporal evolutions depending on the articulation place (rising for coronal, flat for velar and downward for labial), thus partially confirming Delattre's work [24] in synthetic speech. Most of these results on Thai are consistent with those of [8] [19] for Vietnamese and provide information to suggest what acoustic parameter is useful for the categorization of unreleased stops and what is not. However, to be conclusive, they will need to be verified by perceptual tests. More broadly, the studies on relevant acoustic cues to the discrimination of unreleased stops contribute to validate the important role of acoustic transitions in speech perception (see e.g. [26] [27]).

Another result of this study concerns significant effects of word type on the production of syllable rhyme: the realization of a number of acoustic parameters (consonant duration, vowel duration, F_1, f_0) varies as a function of word-type, suggesting that it could probably be the result of different degrees of coarticulation between intra-word *vs* inter-word syllables in this language.

6. ACKNOWLEDGEMENTS

This study was conducted within the research framework of the SCALA project (*Stop Consonants in Asian LAnguages*) (reference ANR 15 IDEX 02).

REFERENCES

[1] N. Vallée, L. J. Boe, J. L. Schwartz, P. Badin, and C. Abry, "The Weight of Phonetic Substance in the Structure of Sound Inventories," *ZAS Papers in Linguistics*, vol. 28, pp. 145–168, Jan. 2002, doi: 10.21248/zaspil.28.2002.163.

[2] I. Maddieson, *Patterns of sounds*. in Cambridge studies in speech science and communication. Cambridge ; New York: Cambridge University Press, 1984.

[3] F. Argod-Dutard, *Éléments de phonétique appliquée: prononciation et orthographe en français moderne et dans l'histoire de la langue ; aspects prosodiques et métriques,* Réimpr. Paris: Armand Colin/Masson, 1996.

[4] Calliope, *La parole et son traitement automatique*. Paris Milan Barcelone Mexico: Masson, 1989.

[5] C. Meunier, Phonétique acoustique, in Les dysarthries, P. Auzou, Ed., Solal, 2007, pp. 164–173.

[6] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Oxford: Blackwell, 1996.

[7] K. Tsukada, A Cross-linguistic Acoustic Comparison of Unreleased Word-final Stops: Korean and Thai. In *Proceedings of INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004, pp. 1293–1296. doi: 10.21437/Interspeech.2004-76.

[8] T. T. H. Tran, *"Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde : de l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE,"* Thèses, Université de Grenoble, 2011. Accessed: Jun. 03, 2021. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01568326

[9] A. S. Abramson, Word-Final Stops in Thai. In *Tai Phonetics and Phonology*, J. G. Harris and R. B. Noss, Eds., Bangkok: Central Institute of English Language, 1972, pp. 1–7. [Online]. Available: http://purl.org/sealang/abramson1972word.pdf

[10] M. R. K. Tingsabadh and A. S. Abramson, Thai. *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 24–28, Jun. 1993, doi: 10.1017/S0025100300004746.

[11] J. Kingston, Lenition. In *Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology*, Somerville, MA, 2008, pp. 1–31.

[12] T. Cho, S.-A. Jun, and P. Ladefoged. Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, vol. 30, no. 2, pp. 193–228, Apr. 2002, doi: 10.1006/jpho.2001.0153.

[13] T. T. Doan. *Ngữ âm tiếng Việt* (Vietnamese Phonetics) Hanoi National University Publishing House, pp. 99–148, 1999.

[14] I. Rousset, *"Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles,"* Thèses, Université Stendhal - Grenoble III, 2004. Accessed: Mar. 17, 2021. [Online]. Available: https://tel.archives-ouvertes.fr/tel-00250154

[15] A. S. Abramson and K. Tingsabadh. Thai Final Stops: Cross-Language Perception. *PHO*, vol. 56, no. 3–4, pp. 111–122, 1999, doi: 10.1159/000028446.

[16] S.-C. Rhee. Onset-Coda Asymmetries in Mishmi and Other Southeast Asian Languages. In *Papers from the Seventh Annual Meeting of the Southeast Asian Linguistics Society*, Tempe, Arizona, 2003, pp. 137–152.

[17] R. Iwata, H. Hirose, S. Niimi, M. Sawashima, and S. Horiguchi, Syllable final stops in East Asian languages: Southern Chinese, Thai, and Korean. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, pp. 621–624, 1990.

[18] K. Tsukada, "Cross-language perception of final stops in Thai and English: A comparison of native and non-native listeners. In *Proceedings of the 10th Australian international conference on speech science and technology*, Sydney, Dec. 2004, pp. 563–568.

[19] T. T. H. Tran, N. Vallée, and L. Granjon. Effects of Word Position on the Acoustic Realization of Vietnamese Final Consonants. *Phonetica*, vol. 76, no. 1, pp. 1–30, 2019, doi: 10.1159/000485103.

[20] N. Yamlamai and T. Tran. Effet de la position de la syllabe sur la réalisation acoustique des consonnes finales du thaï. In *Actes des XXXIIe Journées d'Études sur la Parole*, Aix-en-Provence, France, Jun. 2018, pp. 151–159. doi: 10.21437/JEP.2018-18.

[21] P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program]. Version 6.2.23. Oct. 08, 2022 [Online]. Available: http://www.praat.org/

[22] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous Inference in General Parametric Models. *Biom. J.*, vol. 50, no. 3, pp. 346–363, Jun. 2008, doi: 10.1002/bimj.200810425.

[23] X. H. Cao, *Phonologie et linéarité: réflexions critiques sur les postulats de la phonologie contemporaine*. Paris, France: Selaf, 1985.

[24] P. C. Delattre, A. M. Liberman, and F. S. Cooper. Acoustic Loci and Transitional Cues for Consonants. *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, Jul. 1955, doi: 10.1121/1.1908024.

[25] W. Wang. Transition and Release as Perceptual Cues for Final Plosives. *Journal of speech and hearing research*, vol. 2, pp. 66–73, Apr. 1959, doi: 10.1044/jshr.0201.66.

[26] T. A. X. Tran, V. S. Nguyen, E. Castelli, and R. Carré. Production and perception of pseudo-V1CV2 outside the vowel triangle: speech illusion effects. In *Proceedings of INTERSPEECH 2013*, Lyon, France, Aug. 2013, pp. 407–411. doi: 10.21437/Interspeech.2013-124.

[27] R. Carré, P. Divenyi, and M. Mrayati. *Speech: A dynamic process*. De Gruyter, 2017. doi: 10.1515/9781501502019.

VOICE ONSET TIME (VOT) IN KURDISH EFL LEARNERS

Lureen Ibrahim Naser

Department of English Language, College of Basic Education, University of Duhok, Kurdistan Region, Iraq.

lureen.nasar@uod.ac

ABSTRACT

This paper examines the acquisition of English stops by Bahdini Kurdish learners. Kurdish utilizes a three-way category of voiced, voiceless aspirated and unaspirated stops. Kurdish voiced stops are pre-voiced, while those of English have a short-lag VOT, characteristic of Kurdish voiceless unaspirated stops. Kurdish also has voiceless aspirated stops with shorter VOT than those of English. These differences require learners to restructure their existing categories which is more challenging than creating new categories [1, 2].

The sample consisted of two monolingual control groups of (40) English and Kurdish speakers and two study groups of (60) beginner and advanced Kurdish EFL learners. Two lists of monosyllabic words are constructed, in English and Kurdish, with initial prevocalic stops providing 4140 tokens for analysis. Results show that learners lengthened their VOT towards English as they progress from beginner to advanced stage. Their VOT patterns, however, remain intermediate between both languages.

KEYWORDS: Kurdish EFL Learners, VOT, English stops, Kurdish stops.

1. INTRODUCTION

There are several acoustic signals in speech that help us to detect a foreign accent. Voice Onset Time (VOT) in stops has long been recognized as a significant feature in distinguishing accented speech and an important parameter susceptible to measurement and analysis.

In second language (L2) acquisition, learners' productions in the target language (TL) are constrained by the perceptual experience of their first language (L1), according to the Perceptual Assimilation Model-L2 (PAM-L2) [3]. The perception of learners is adjusted constantly for both L1 and L2. The Speech Learning Model (SLM) [2] postulates that learners create phonetic categories which include long-term memory representations created through exposure to the acoustic-phonetic cues of the language. The SLM hypothesizes that, based on the input learners receive, they start developing sub systems for the language they are trying to learn to encompass the new realization rules of the acquired language. Learners use this subsystem when they are in the TL processing mode. However, learners commonly deviate from the phonetic norms of the TL. The ability to form new phonetic categories decreases due to the fully developed and stabilized phonetic system of the L1, especially with limited L2 input [1]. When both L1 and TL share similar but not identical phonetic categories, it is harder for learners to establish or even adjust or expand those of their L1 to accommodate these slight changes [2]. Thus, small differences may be more problematic for learners who need to re-structure previously established phonetic categories to accommodate the features of the new language.

An example that may support this hypothesis would be the case of this study concerning the acquisition of English stops by Kurdish learners. Both languages have voiced stops, however, there are different acoustic properties attributed to the voiced stops of each language with reference to VOT. English voiced stops are produced with short lag VOT, characteristic of Bahdini Kurdish voiceless unaspirated stops, while Kurdish voiced stops are produced with voicing lead, similar to some other Indo-Iranian languages [4]. These different phonetic details are expected to be challenging for Kurdish learners and may make it harder for them to achieve native-like VOT. Unlike English, Kurdish contrasts voiceless aspirated and unaspirated stops. So, the voiceless stops of both languages are also recorded to check if they occupy different areas along the VOT continuum. Accordingly, the aim of this paper is to answer the following questions:

- Do Kurdish EFL learners produce English voiced stops with a negative VOT characteristic of Kurdish or a short lag VOT characteristic of English?
- Can they adjust their two categories of voiceless stops (aspirated and unaspirated) to that of one English phonological category of voiceless stops?
- Does the proficiency level of learners affect their performance?
- Are the variables of laryngeal state, place of articulation, vowel height or word duration significant?

2. METHODOLOGY

2.1. Stimuli

Tow lists of monosyllabic words were constructed in English and Kurdish. The phonological voiced category of stops was represented by initial /b/, /d/ and /g/ and the phonological voiceless category of stops was represented by initial aspirated and unaspirated (only for Kurdish) /p/, /t/ and /k/. The lists included two example words for each stop, one followed by a high vowel and one by a low vowel.

2.2. Participants

Four groups of participants were chosen for this study. Two control groups of (30) monolingual adult native speakers of Bahdini Kurdish and (10) monolingual adult native speakers of English. They were aged between 20-50 years. The experimental groups were (60) adult Kurdish EFL learners from two proficiency levels. Both groups were university students. The participants had all studied English but differed in the amount of academic input they received in their study starting from one year to four years. They were aged between 20-30 years of age. All groups had equal numbers of males and females.

2.3. Acoustic procedures

Monolingual English and Kurdish participants were asked to read a list of words with initial voiced and voiceless stops, constructed by the researcher, each in their native language. The experimental groups of Kurdish EFL learners were asked to read the same list read by English native speakers. Each word was repeated (3) times, which provided 3 tokens for each word. Participants were asked to repeat a word when pronounced incorrectly. Productions were recorded using a portable Zoom H1n digital audio recorder with a built-in microphone. All the recordings were in uncompressed WAV format, sampled at 44.1 kHz with 16 bits per sample. They were later coded and segmented in Audacity into separate WAV files, one file for each word in the list to be ready for analysis by *Dr.VOT* software [5], which is a package for automatic measurement of VOT. All the resulting acoustic measurements of VOT were visually reviewed and manually corrected in *Praat* [6].

2.4. Statistical procedures

Statistical analyses were carried out using R [7]. To investigate the effect of a group of factors on VOT, the LME model, using the *lme4* package [8] and the *lmerTest* package [9], was adopted. Each studied group had a different LME model. Fixed Factors in the model included Place of Articulation (POA) (bilabials, dental, velar), Laryngeal State (voiceless aspirated, voiceless unaspirated and voiced), Vowel Height and carrier word Duration and a by-Speaker random factor with an alpha level $p = .05$. A post-hoc Tukey's test [10] was carried out to show the direction of difference between the three different levels POA. Another post-hoc Dunnett's test, using *DescTools* package [11], was used to test statistical differences between all studied groups and to conclude whether VOT values produced by any of the research groups, Kurdish Learners G1 and G2, are closer to Kurdish or English.

3. RESULTS

Table 1 shows the mean (M), standard deviation (Std.D) and range (R) of VOT values of word initial stops for all studied groups. Voiced stops are characterized by lead for all groups except for English /d/ and /g/, which have a short lag. Kurdish voicing lead, however, tends to be shorter in Kurdish than the other groups. Kurdish voiceless unaspirated stops are produced with a short lag. Voiceless aspirated stops, on the other hand, are all produced with a long lag. Voiceless stops show a pattern across all groups in which bilabials have the shortest VOT, followed by dentals then velars, which is the universal pattern [12]. This is also evident in the results of the post-hoc Tukey test.

Kurdish						
	/b/	**/d/**	**/g/**	**/kʰ/**	**/pʰ/**	**/tʰ/**
Mean	-114	-119	-114	86	57	62
Std.D	32	34	33	21	15	20
Range	-198:-51	-214:-52	-209:-42	50:155	36:103	31:121
				/k/	**/p/**	**/t/**
Mean				25	11	16
Std.D				5	3	6
Range				14:39	5:21	5:31
English						
	/b/	**/d/**	**/g/**	**/kʰ/**	**/pʰ/**	**/tʰ/**
Mean	-12	16	21	120	90	99
Std.D	36	13	17	24	21	27
Range	-99:12	-54:24	-56:38	50:159	50:120	31:145
KEFL1						
	/b/	**/d/**	**/g/**	**/kʰ/**	**/pʰ/**	**/tʰ/**
Mean	-108	-109	-103	91	62	70
Std.D	36	38	41	26	24	26
Range	-197:11	-203:19	-196:23	35:157	23:118	20:136
KEFL2						
	/b/	**/d/**	**/g/**	**/kʰ/**	**/pʰ/**	**/tʰ/**
Mean	-82	-72	-63	88	55	68
Std.D	35	46	48	21	19	20
Range	-170:11	-178:23	-172:38	33:137	20:108	27:126

Table 1: Range (R), Mean (M) and Standard Deviation (Std. D) of VOT values of the stop categories for all studied groups measured in milliseconds (ms).

To conclude whether VOT values produced by any of the research groups, Kurdish Learners G1 and G2, are closer to Kurdish or English, a post-hoc Dunnett's test was used. It showed significant differences between all four groups. The results of the test concluded that Kurdish Learners G1's (KEFL1) mean VOTs are closer to Kurdish (-3.73, p = .574) than to English (-71.8, p < .001). The Kurdish Learners G2's (KEFL2) mean VOTs, however, reveal change towards English (-57.9, p < .001) and away from Kurdish (-10.15, p = .006). Based on this, we can conclude that Kurdish learners have developed towards English, but still produce English voiced stops with notably long negative VOT.

3.1. VOT and laryngeal state

Laryngeal state of initial stops showed a significant effect on their VOT in all studied groups (F (2, 1568) = 51.34; p < .001) for Kurdish, (F (1, 267) = 7.86; p = .005) for English, (F (1, 998) = 160.99; p < .001) for Kurdish learners G1 and (F (1, 1014) = 145.09; p < .001) for Kurdish learners G2. The two laryngeal categories of voiced and voiceless states had significantly different VOT values, (p < .001) for all studied groups with voiceless stops having longer VOT values than voiced ones, as evident in Figure 1.

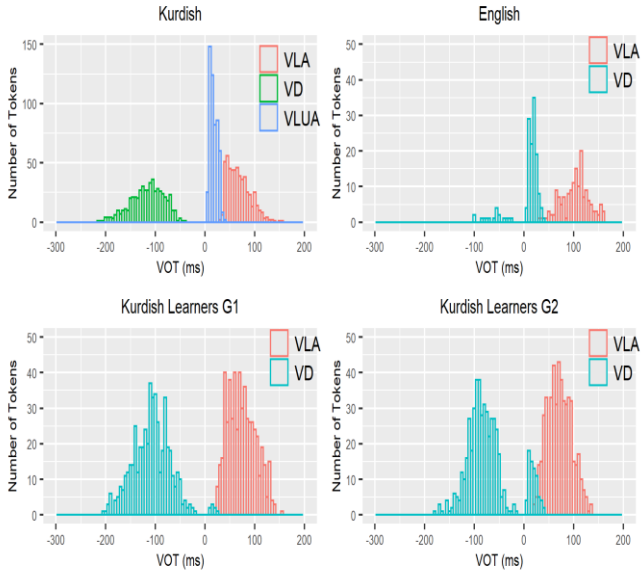


Figure 1: Histogram of the distribution of VOT values of stop categories for all studied groups.

3.2. Place of articulation

Place of articulation of the initial stops showed a significant effect on VOT only for the English and Kurdish G2 groups (F (2, 1557) = 5.26; p = .005) for Kurdish, (F (2, 267) = 1.52; p = 221) for English, (F (2, 996) = 12.59; p < .001) for Kurdish learners G1 and (F (2, 1007) = 1.19; p = .304) for Kurdish learners G2.

The distinction in VOT was evident between the three places of articulation. Bilabials and Velars differed from each other significantly in all of the studied groups (p < .001), except for the English group (p = .072) and Kurdish learner G2 group (p = .262). However, bilabials didn't show such high significant difference in VOT with dentals except for the Kurdish (p = .003) and Kurdish learners G1 group (p = .002). English and Kurdish learners G2 and Kurdish had the following values respectively; (p = .377) and (p = .025). These results are visualized in Figure 2.

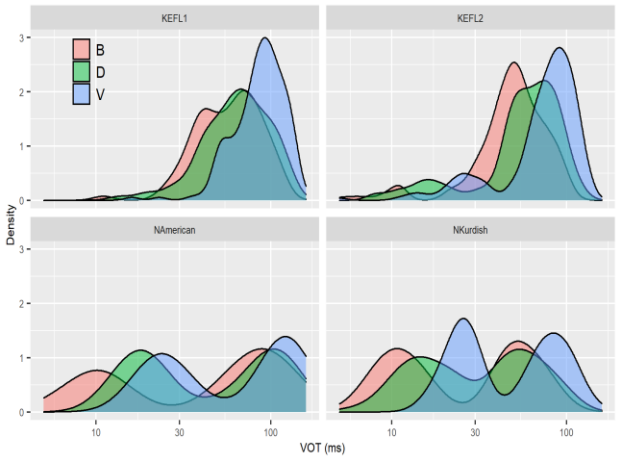


Figure 2: Density plots of the distribution of VOT values based on POA for all groups.

To assess the direction of difference between the three places of articulation, a post-hoc Tukey test [11] was performed. Mean VOT values of voiceless stops show a pattern across all groups in which bilabials have the shortest VOT, followed by dentals then velars.

3.3. Speech rate (Duration)

Duration of word carrier had a statistically significant effect on the VOT of Kurdish, Kurdish learners G1 and Kurdish learners G2 groups (F (1, 468) = 53.71; p <.001) for Kurdish, (F (1, 863) = 38.80; p =<.001) for G1 and (F (1, 1027) = 33.06; p =<.001) for G2. The VOT of the English group, however, didn't show a statistically significant effect of word duration (F (1, 273) = 0.30; p =.582).

3.4. Post vowel height

Height of the following vowel only showed a significant effect on VOT of initial stops for native Kurdish but not for the other studied groups with: F (1, 1558) = 8.64; p =.003 for Kurdish, F (1, 266) = 2.25; p = .134 for English, F (1, 994) = 0.54; p = .464 for Kurdish learners G1 and F (1, 1010) = 0.03; p = .863 for Kurdish learners G2.

As displayed in Figure 3, the distinction in VOT between high and low vowels was only statistically significant for the Kurdish and learners G1 groups with a $p < .001$, $p = .337$, $p = .011$, $p = .564$ for Kurdish, English, learners G1 and G2 respectively.

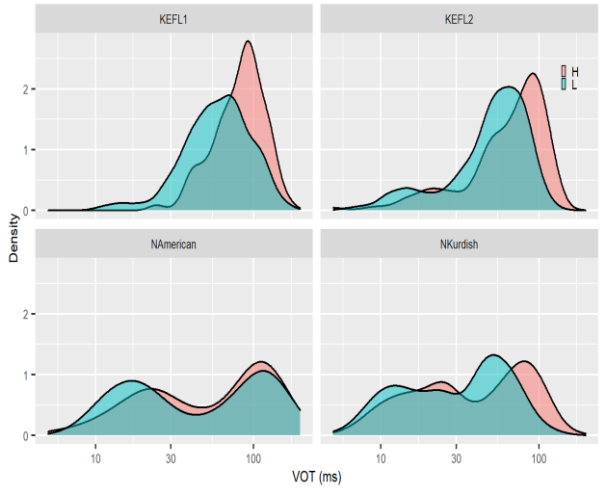


Figure 3: Density plots of the distribution of VOT based on vocalic context for all groups.

4. SUMMARY AND DISCUSSION

Extensive research has been carried out to investigate the role of VOT in L2 stop productions. Learners of varied native language backgrounds exhibit different measures of TL VOT. This paper examines how L1 Kurdish EFL learners produce VOT for the voicing contrast between English voiced and voiceless stops. Kurdish contrasts pre-voiced voiced stops, short-lag unaspirated stops, and long-lag aspirated stops, unlike English which contrasts long-lag voiceless stops with short-lag voiced stops. The three-way voicing contrast in L1 Kurdish is mapped onto a two-way contrast in L2 English, in which all three VOT categories are employed at the level of phonetic realization.

Results show that learners produce pre-voicing for English voiced stops, though less than in Kurdish. Long-lag VOT for voiceless aspirated stops in Kurdish and in EFL learners tends to be shorter than that of voiceless aspirated stops of English by native speakers. Kurdish EFL learners transfer their voiced stop lead VOT values to the TL and tend to produce English voiced stops with a negative VOT, especially in lower proficiency levels. This transfer, however, tends to be less for more advanced levels, whose VOT productions of voiced stops involve both negative and positive values. VOT values of the L1 unaspirated voiceless stop category (short lag) overlap with the L2 voiced stop category. Kurdish speakers have to learn to produce the voiced phoneme with an optional short-lag VOT variant.

Learning an L2 phoneme which is only phonetically different from one in L1, would only require learners to adjust their existing phonetic categories to embrace the subtle changes of the new L2 variant. In the case of this study, although Kurdish EFL learners show progress towards the target language, they still produce English voiced stops with notably long negative VOT values. They also produce English voiceless aspirated stops with less aspiration than the English group. This supports Flege's [2] view that small differences can be more challenging for learners. The inability to produce TL stops with accurate VOT leads to retaining an amount of foreign accent in their speech. It may be due to the following reasons, which represent general common concepts among L2 speech theories.

- L2 learning took place in a foreign language acquisition environment (FLA) where the L1 is the predominant language and is not used outside the classroom. This supports the notion that the ability to form new phonetic categories decreases with limited L2 input.
- Structure of L1 of learners has an effect on learning that of an L2, although phonetic or phonological similarities do not guarantee ease of learning [3]. According to the SLM [2], these learners had already created phonetic categories for their native Kurdish voiced stops with negative VOT values. When they try to learn English, they will only need to adjust that of their native language to include the subtle phonetic details of English voiced stops with a short lag. This is challenging, since the phonetic details of both languages exist in the same phonological space; meaning that voicing lead and short lag are both associated with voiced stops.
- The level of learners (beginner or advanced) determines the degree of effect of certain variables on their L2 production. Advanced learners, for instance, are expected to be able to distinguish phonetically similar structures between L1 and L2 more easily than beginners. Such similarity may create confusion in novice learners. In this study, advanced learners have shown some progress by producing English voiced stops with shorter negative VOT values, compared to the long negative VOT values produced by beginners.

5. ACKNOWLEDGMENTS

The author would like to thank Professor Jane Stuart-Smith at Glasgow University Laboratory of Phonetics (GULP) for her valuable feedback on earlier versions of the paper. Thanks to Dr. Clara Cohen for her statistical advice on R.

6. REFERENCES

1] Flege, J. E. 2003. Assessing constraints on second-language segmental production and perception. Phonetics and phonology in language comprehension and production: Differences and similarities, 6, 319-355.

[2] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. Speech perception and linguistic experience: Issues in cross-language research, 92, 233-277.

[3] Best, C. T., & Tyler, M. D. 2007. Nonnative and second-language speech. Language experience in second language speech learning: In honor of James Emil Flege, 17, 13.

[4] Hussain, Q. (2018). A typological study of voice onset time (VOT) in Indo-Iranian languages. *Journal of Phonetics* 71, 284–305.

[5] Shrem, Y., Goldrick, M., & Keshet, J. 2019. Dr. VOT: Measuring positive and negative voice onset time in the wild. Proc. Interspeech. arXiv preprint arXiv:1910.13255. https://arxiv.org/pdf/1910.13255.pdf

[6] Boersma, P. and Weenink, D. 2018. "Praat: Doing phonetics by computer [computer program]," Version 6.0.39. URL http://www.praat.org/.

[7] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[9] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. 2017. LmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13 (R lmerTest package: version 2.0-36).

[10] Keselman, H. J., & Rogan, J. C. 1977. The Tukey multiple comparison test: 1953–1976. Psychological Bulletin, (845), 1050.

[11] Andri et mult. al. S 2022. *DescTools: Tools for Descriptive Statistics*. R package version 0.99.45, https://cran.r-project.org/package=DescTools.

[12] Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 27, 207–229.

Is focalization the keystone connecting vowel perception and production?

Linda Polka[1], Matthew Masapollo[2], Lucie Ménard[3], Lucile Rapin[3] and Jean-Luc Schwartz[4]

[1]McGill University, Canada [2]University of Florida, USA [3]UQAM, Canada [4]GIPSA-lab Grenoble, France
linda.polka@mcgill.ca; mmasapollo@phhp.ufl.edu; menard.lucie@uqam.ca; lucilerapin@gmail.com;
jean-luc.schwartz@gipsa-lab.gernoble-inp.fr

ABSTRACT

Vowel perception is often asymmetric such that discriminating the same vowel contrast is easier in one direction compared to the reverse direction. According to one account, asymmetries reveal a perceptual bias favoring more extreme vowel gestures which give rise to salient acoustic signals due to increased formant convergence (*focalization*). Consistent with this articulatory account are findings that analogous asymmetries are elicited with silent lip-read videos of vowels, as well as other disembodied, animated displays that simulate the kinematic and configural properties of articulatory mouth gestures. However, our understanding of the relations between vowel constrictions, perceptual asymmetries, and formant patterns is incomplete. Here, focusing on the French /e/-/ø/ contrast, we examine whether formant proximities are related to perceptual asymmetries, and whether there are firm links between formant convergence and tongue constriction degree and lip aperture. We introduce the notion that focalization may serve as the *keystone* connecting vowel perception and production.

Keywords: vowels, perception, production, focalization, electromagnetic articulography

1. INTRODUCTION

Considerable research over the past 30 years has revealed that vowel perception by human adults and infants is often asymmetric: discriminating the same vowel contrast is easier in one direction compared to the same change presented in the reverse direction (for reviews and meta-analytic findings, see [1, 2]). This perceptual pattern is schematized in Figure 1, which shows a standard *F1-F2* plot of vowel contrasts that have been examined in infant vowel discrimination studies with arrows indicating the direction of change that was easier to discriminate.

According to the Natural Referent Vowel (NRV) framework [1], these asymmetries reveal a universal perceptual bias toward acoustic signals produced with more extreme vowel gestures, which act as *natural referent* vowels. Specifically, it is argued that vowels generated by more extreme gestures display heightened acoustic and perceptual salience due to well-defined spectral prominences formed by the convergence of adjacent formants (*focalization*; see also [3]). As Figure 1 illustrates, directional asymmetries can frequently be predicted from *F1-F2* vowel space location because, at least for oral vowels, focalization increases and reaches maximum levels at the corners of the vowel space. Critically, however, focalization and peripherality are distinct phonetic properties. Focalization considers the first three to four formants and their spectral position relative to each other, not just the vowel position within a simple *F1-F2* space.

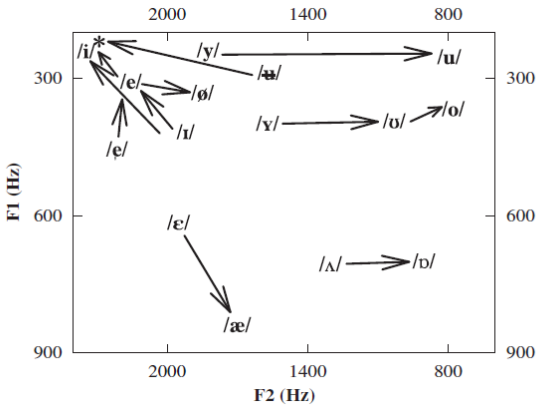


Figure 1: Schematic illustration of phonetic space based on the first two formant frequencies. Vowel contrasts showing asymmetries in infant vowel perception are plotted; arrows point in the direction that is easier to discriminate (adapted from Polka & Bohn, 2011).

The /e/-/ø/ vowel contrast is an example where perceptual asymmetries predicted by peripherality versus focality do not align, and thus, provides an ideal opportunity to test the NRV claim that articulatory gestures – acoustically specified by formant convergence – underlie perceptual asymmetries. Prior studies have examined asymmetries for the /e/-/ø/ vowel contrast in several languages including German [4], Dutch [5], and Danish [1]. In each study, the test stimuli were

multiple tokens produced by one talker; formant values (averaged across tokens) are shown below in Figure 2. In all three studies, *F2* frequencies were higher for /e/ than /ø/. *F1* frequencies were similar for both vowels in the Danish and German productions and were lower for /e/ compared to /ø/ in Dutch. As shown in Figure 2, /e/ is more peripheral than /ø/ in *F1-F2* space in the German and Danish productions, but not in the Dutch productions. When these formant data were recently examined from the perspective of focalization [6], a consistent pattern emerged. As shown in Table 1 below, for each talker *F1* and *F2* frequencies were much closer in /ø/ compared to /e/; *F1* and *F3* distances showed the same pattern. These differences were consistent and robust; in each language there was no overlap in the magnitude of these formant distances between /e/ and /ø/. Some differences in *F2* and *F3* convergence were also noted, but in each language, they were much smaller, overlapped in magnitude across vowels, and were inconsistent in direction (e.g., more *F2-F3* convergence for /ø/ in Danish; for /e/ in Dutch).

These acoustic patterns are intriguing considering perceptual findings (see Table 2). In the Danish and the Dutch study, an asymmetry was observed showing better performance in the /e/ to /ø/ direction; this pattern was not predicted based on peripherality in either Danish or Dutch but aligns well with predictions based on focality in both languages. No asymmetry was observed in the German study; this may be because the acoustic differences between vowels were larger in this study. Overall, these findings bolster NRV by suggesting that asymmetries are associated with differences in focalization, whereas a peripherality-based prediction fails.

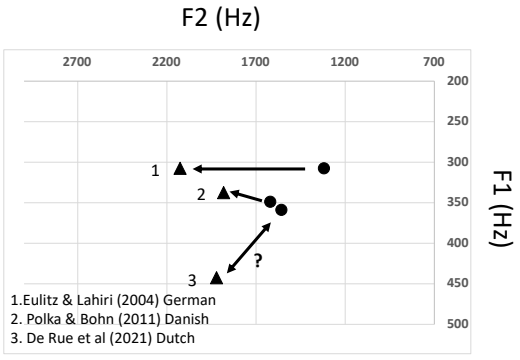


Figure 2: *F1-F2* plot of vowel stimuli from three studies that assessed perceptual asymmetries for /e/-/ø/. Mean values (across multiple tokens) are plotted for /e/ (triangles) and /ø/ (circles). An arrow connects the vowel pair to show the peripherality prediction asymmetry.

The NRV framework further posits that the processes underlying vowel perception asymmetries are sensitive to articulatory information, not focal acoustic patterns *per se* [7]. Especially consistent with this account are recent findings that asymmetries, comparable to those observed with acoustic vowel stimuli, are elicited with silent lip-read videos of vowels [7,8], as well as other disembodied, animated displays that simulate the kinematic and configural properties of orofacial speech movements [9]. Thus, in this view, focalization serves as the *keystone* (or lynchpin) connecting perceptual and articulatory patterns: perceivers directly extract information about distal vowel constrictions from proximal formant patterns. However, this account does not provide firm evidence that vowel constrictions are or are not recoverable from formant parameters, as a comprehensive understanding of the complex relations between vowel constrictions, formant patterns, and perceptual asymmetries is still being built [3,7].

		Danish (n=1) Polka & Bohn (2011)	German (n=1) Eulitz & Lahiri (2004)	Dutch (n=1) de Rue et al (2021)
F1-F2 (Hz)	e	1541	1817	1474
	Ø	**1271**	**1009**	**1198**
	difference	270	808	276
F1-F3 (Hz)	e	2239	2442	1980
	Ø	**1935**	**1709**	**1813**
	difference	304	733	167
F2-F3 (Hz)	e	698	**625**	**506**
	Ø	**664**	700	615
	difference	34*	75	109*

* values overlap

Table 1: Formant convergence analysis of vowel stimuli from three studies that assessed perceptual asymmetries for /e/-/ø/. Shaded values indicate the more focal vowel within the pair.

	Polka & Bohn (2011)	Eulitz & Lahiri (2004)	de Rue et al (2021)
Language	Danish	German	Dutch
Peripherality prediction (Figure 2)	/Ø/ to /e/ easier	/Ø/ to /e/ easier	?
Focality prediction (Table 1)	/e/ to /Ø/ easier	/e/ to /Ø/ easier	/e/ to /Ø/ easier
Subjects	Danish infants	German adults	Dutch adults
Perceptual task	conditioned headturn	MMN oddball paradigm	MMN oddball paradigm
Finding	/e/ to /Ø/ is easier	no asymmetry	/e/ to /Ø/ is easier

Table 2: Summary of acoustic and perceptual findings across three studies that assessed perceptual asymmetries for /e/-/ø/.

While the findings reported in Polka *et al.* [6] are compatible with NRV, their interpretation is post hoc and based on a small sample of vowels produced by one talker in each language (3-4 tokens per talker). The current research provides a stronger empirical test of the core NRV assumptions regarding asymmetries by examining the articulatory and

acoustic properties of the /e/-/ø/ contrast in Canadian French with a broader sample of talkers, using state-of-the-art electro-magnetic articulography (or EMA; [10,11]). We addressed two specific research questions. First, does focalization provide a more direct, principled account for perceptual asymmetries than vowel space peripherality? Second, is there a relationship between degree of articulatory constriction and formant proximity for the /e/-/ø/ contrast? In this paper, we report findings that address the first research question and outline how they guide ongoing analyses to address the second question.

2. METHODS

2.1. Participants

We acquired data from twenty native speakers of Canadian French (mean age: 28 years, *SD* ± 4.9 years; range, 19 to 39 years; all male). None reported a history of a speech, language, or hearing disorder.

2.2. Stimuli

Talkers produced randomized blocks of all ten Canadian French vowels (/i y u e ø o ɛ œ ɔ a). To control the vocal tract configuration at the onset of stimulus production, each vowel was embedded in the carrier phrase, "V as WORD," where V was one of ten vowels and WORD was a CVCV word with /p/ as C and the target vowel V in both vowel positions. For example: «a comme papa» («a as papa»).

2.3. Recordings

Speech movements were recorded using an AG500 EMA system (Carstens Medizinelektronik GmbH, Bovenden, Germany). Dynamic, articulatory sensors were affixed on the midsagittal surfaces of the upper lip, lower lip, jaw, tongue-tip, tongue-blade, and tongue-dorsum. Static, head reference sensors were affixed on the gingiva above the upper left and right incisors and behind each ear on the left and right mastoid processes. Participants were recorded while producing ten repetitions of each vowel, resulting in 100 tokens collected per participant. Stimuli were orthographically cued randomly. Simultaneous audio (with a sampling rate of 44.1 kHz) was recorded using an omnidirectional condenser head-worn microphone (Audio-Technica BP892) and the EMA sensor signals (with a sampling rate of 250 Hz) were recorded using Carsten's `CS5RECORDER` and `CS5VIEWER`.

2.4. Acoustic Analyses

This work is based on the analysis of 10 productions of /e/ and /ø/ by each talker (200 tokens per vowel and 400 vowels in total). To avoid coarticulation effects, only the first (isolated) vowel in the sentence «V as CVCV» was analysed. Speech signals were down sampled to 22.05 kHz (half the sampling frequency). Vowel onset and vowel offset were manually identified based on the appearance and disappearance of the first two formant frequencies on the spectrogram, respectively, using Praat [12]. The first three formants were extracted at vowel midpoint using Linear Predictive coding.

2.5. Articulatory Kinematic Analyses

The raw kinematic data first underwent a series of standardized pre-processing steps to rotate and translate each sensor position signal to the occlusal plane and then, using information from the head reference sensors, were corrected for head motion artifacts. The acoustic and kinematic movement signals were then computed, synchronized, and visualized together using the MATLAB-based toolbox *Mview* [13,14], developed by Mark Tiede at Haskins Laboratories. Ongoing analyses are based on the sensors located on the lips and tongue-dorsum because they provide the most direct information about vowel constrictions. Vertical lip separation (the Euclidean distance between the sensors affixed on the upper- and lower-lip sensors), lip protrusion (relative to the upper incisors), and the vertical displacement of the tongue-dorsum (relative to the upper incisors) at vowel midpoint are being extracted from the movement time series for each vowel token.

3. RESULTS

F1 and *F2* formant frequencies (averaged across tokens) are plotted in Figure 3 below for each of the 20 talkers. For every talker, *F2* frequencies were higher for /e/ than /ø/. As well, for every talker, vowel height was higher (*F1* frequencies lower) for /e/ compared to /ø/. These data confirm that, in Canadian French, /e/ is a more peripheral vowel than /ø/ in the standard *F1-F2* vowel space. The vowels were also analyzed to assess differences in formant distance or focalization. Findings are summarized in Table 3. These data clearly show that *F1* and *F2* frequencies are closer in productions of /ø/ compared to /e/; *F1* and *F3* formant distances show the same pattern. Both patterns are robust and evident in all 20 talkers (χ = 20; p <.0001). Some differences in *F2* and *F3* convergence were noted. Group means show that /e/ is more focal than /ø/ with respect to *F2-F3;* however, the *F2-F3* difference (on average) across vowels was less than half of the differences observed for *F1-F2* and *F1-F3*. Overall, this pattern is weak and inconsistent; it was evident for only 13 out of 20 talkers (χ = 1.8; p =.179). The other 7 talkers show the reverse patten (more focal *F2-F3* for /ø/) or no difference across the vowels. The formant proximity (focalization) patterns are shown schematically in

Figure 4 below. The dotted arrows highlight the more focal formant pattern for /ø/ compared to /e/, especially with respect to *F1-F2* and *F1-F3* formant proximity.

4. DISCUSSION

4.1. Acoustic-Perceptual Relations

Here we examined the Canadian French /e/-/ø/ contrast, which is articulatorily distinguished by lip-rounding (compression and protrusion). To confirm the focalization patterns observed for this vowel pair in prior studies, we recorded and analyzed a multi-talker sample of Canadian French /e/ and /ø/ vowel productions. Our acoustic findings align with the earlier (single talker) analyses reported in Danish, Dutch, and German. In our French corpus, /e/ was consistently more peripheral than /ø/ in *F1*/*F2* space. French /ø/ was also more focal than French /e/ with respect to *F1-F2* and *F1- F3* proximity. These formant patterns were quite robust and uniform across talkers. Like prior studies, weak and inconsistent differences in *F2-F3* proximity were noted. These data provide further evidence that when perceptual asymmetry predictions based on peripherality and focalization conflict, focalization is the winner. These findings support the core NRV claims that perceptual asymmetries reflect a sensitivity to articulatory patterns that are specified by focalization patterns.

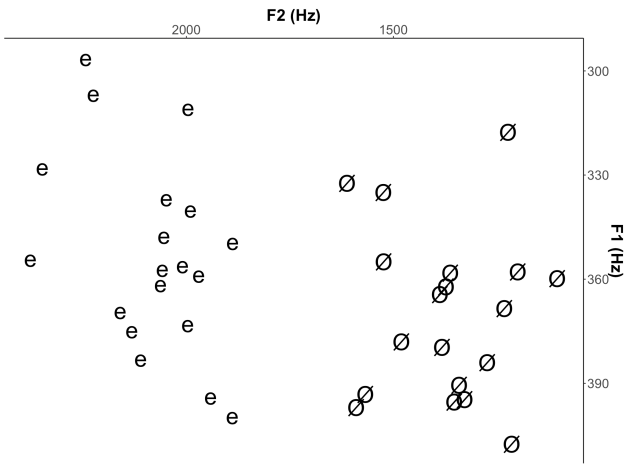


Figure 3: F1/F2 plot of French /e/ and /ø/ productions. Each symbol corresponds to the mean values (across 10 tokens) for 1 talker.

Location in the vowel space (peripherality) often predicts asymmetries for contrasts involving vowel height or front/back dimensions. However, this approach does not work for contrasts that involve lip-rounding, whereas a prediction based on formant convergence is more successful (see also [15]). These findings support the NRV claim that focalization provides a more principled account for perceptual asymmetries. Further research with other lip-rounding contrasts is needed to firmly establish this position.

		French (n=20)
F1-F2 (Hz)	e	1742
	Ø	**1019**
	difference	713
F1-F3 (Hz)	e	2335
	Ø	**1839**
	difference	509
F2-F3 (Hz)	e	588
	Ø	821
	difference	200

Table 3: Formant convergence analysis of French /e/-/ ø / vowels produced by 20 talkers. Shaded values indicate the more focal vowel within the pair.

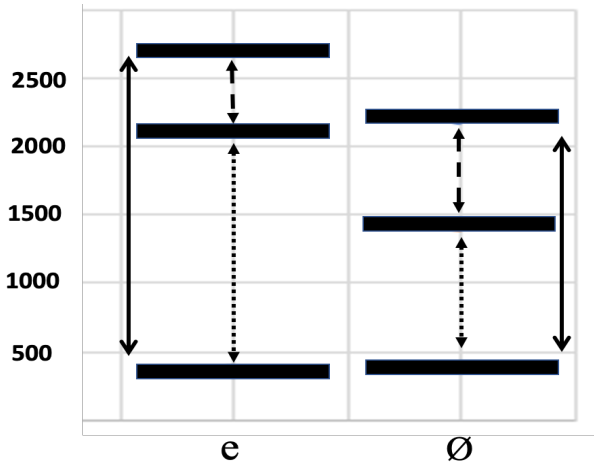


Figure 4: Schematic highlighting formant convergence patterns observed for French /e/ and /Ø/. The formant values shown correspond to mean values across 20 talkers. Lines connected with arrows indicate the close formant proximity for /ø/ compared to /e/, especially for *F1-F2* (dotted lines) and *F1-F3* (solid line).

4.2. Articulatory-Acoustic Relations

To further substantiate the NRV account, the next step is to examine how vowel constrictions are tied to focalization. In one prior study [7], image processing of 2-D video images of a model talker's face during vowel production showed more lip compression was associated with more focal *F1-F2* patterns. Ongoing analyses with the current EMA recordings are exploring how the lips and tongue-dorsum work in concert to alter the overall spectrum of /e/ and /ø/. We predict that the constriction degree of the tongue-dorsum and lips (horizontal and vertical aperture) will be greater during productions of /ø/ compared to /e/, and the extent of constriction will be correlated with the proximity of formants,with lip-rounding showing a larger effect on lowering *F3*.

5. REFERENCES

[1] L. Polka and O-S. Bohn, "Natural referent vowel (NRV) framework: An emerging view of early phonetic development," *Journal of Phonetics,* vol. 39, pp. 467-478, 2011.

[2] S. Tsuji and A. Cristia, "Which acoustic and phonological factors shape infants' vowel discrimination? Exploiting natural variation in InPhonDB," In Proc. of Interspeech, 2017*,* pp. 2108-2112.

[3] J.L. Schwartz, C. Abry, L.J. Boë, L. Ménard and N. Vallée, "Asymmetries in vowel perception, in the context of the Dispersion-Focalization Theory," *Speech Communication,* vol. 45, pp. 425–434. 2005.

[4] C. Eulitz and A. Lahiri, "Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition," *Journal of Cognitive Neuroscience*, vol.164, pp. 577-583, 2004.

[5] N. de Rue, T. Snijders and P. Fikkert, (2021). "Contrast and conflict in Dutch Vowels," *Frontiers in Human Neuroscience*, vol. 15, 2021. [Online serial] Available: doi: 103389/fnhum.2021.629648.

[6] L. Polka, M. Masapollo and O-S. Bohn, "Predicting asymmetries in vowel perception: formant convergence succeeds where peripherality fails," presented at the 181st Meeting of the Acoustical Society of America, Seattle, USA, 2021.

[7] M. Masapollo, L. Polka and L. Ménard,. "A universal bias in adult vowel perception - By ear or by eye," *Cognition*, vol. 166, pp. 358-370, 2017.

[8] M. Masapollo and F.H. Guenther, "Engaging the articulators enhances perception of concordant visible speech movements." *Journal of Speech, Language, and Hearing Sciences*, vol 62, No.10, pp.3679-3688, 2019.

[9] M. Masapollo, L. Polka, L. Ménard, L. Franklin, M. Tiede and J.L. Morgan, "Asymmetries in unimodal visual vowel perception: The roles of oral-facial kinematics, orientation and configuration," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 44, No.7, pp. 1103-1118, 2018.

[10] J.S. Perkell, M.H. Cohen, M.A. Svirksy, M.L. Matthies, I. Garabieta and M.T.T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America,* Vol.92, No.6, pp. 3078–3096, 1992.

[11] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology*, Vol.12, No.6, 2021.

[12] P. Boersma and D. Weenink, (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/.

[13] M. Tiede, "MVIEW: Software for visualization and analysis of concurrently recorded movement data," Haskins Laboratory report, New Haven, CT. 2005.

[14] M. Tiede, "MVIEW: Multi-channel visualization application for displaying dynamic sensor movement," Haskins Laboratory report, New Haven, CT. 2010.

[15] C.T. Best and A. Faber, "Developmental increase in infants' discrimination of nonnative vowels that assimilate to a single native vowel," presented at the International Conference on Infant Studies, Brighton, UK, 2000.

6. ACKNOWLEDGEMENTS

The work reported here was supported by the Natural Sciences and Engineering Research Council of Canada under grant numbers 105397 (L.Polka) and 312395 (L. Ménard).

FUNDAMENTAL FREQUENCY NORMALIZATION AND STATISTICAL POWER: AN ASSESSMENT OF 15 NORMALIZING TECHNIQUES

Jérémy Genette[1], Jo Verhoeven[1,2], Steven Gillis[1]

[1] Antwerp University, [2] City, University London
jeremy.genette@uantwerpen.be; jo.verhoeven@city.ac.uk; steven.gillis@uantwerpen.be

ABSTRACT

The effect of 15 normalization procedures on the power of statistical tests is assessed. The difference in the F0 in Hz between the high and low vowels (N=1471), i.e. the intrinsic vowel pitch, produced by 47 native speakers of Dutch was assessed by means of a t-test. The same test was applied to the F0 values normalized by means of 15 conventional procedures. The power of the tests was registered to assess the effect of the normalization procedures. The results show that statistical power is influenced by applying normalization, but the difference in power of the procedures is levelled out as a function of the number of observations. The effects of the normalization procedures are interpreted in terms of how the variability of the data is accounted for.

Keywords: normalization, methodology, statistics, fundamental frequency, intrinsic vowel pitch.

1. INTRODUCTION

In addition to linguistically relevant information, speech also conveys information about the speaker, such as their sociolinguistic and physiological characteristics [1]. The latter can be studied for their own sake, such as the effect of physiological differences on acoustic differences [2]. But in other areas of research, those characteristics are considered as noise. For instance, in studying F0, individual anatomical differences – whether or not related to gender differences – are treated as noise which researchers typically try to eliminate by applying normalization.

Similarly to the normalization of vowel formants, the normalization of F0 consists of (non-) linearly rescaling raw frequency values in order to level out anatomical differences between speakers. Research on normalization techniques have studied how well they preserve (sociolinguistic) variation while eliminating idiosyncratic physiological noise [3], the degree of overlap between normalized vowel spaces [4], and their ability to reduce the within-vowel category variability while enhancing the between-vowel category variability [5].

However, the *a posteriori* rescaling of the original frequency measurements in Hz can lead to a rescaling of the variability in the data. As such, normalization potentially affects hypothesis testing and the power of statistical tests, i.e. the probability of correctly rejecting the null hypothesis. Research in genetics [6] and neurology [7] has revealed the impact of normalization procedures on statistical power. However, to the best of our knowledge, F0 normalization has not yet been investigated in this respect.

The main objective of this paper is to study the effect of F0 normalization on statistical power. It addresses the question whether normalization affects the statistical power of statistical tests. For this purpose, 15 conventional normalization procedures were applied to the same dataset, which consisted of F0 measurements of 1471 vowels produced by 47 speakers. This corpus was selected because it contains all monophthong vowels of Dutch produced by many speakers in carefully controlled consonantal contexts. The difference in F0 between the high and low vowels, i.e. the so-called “intrinsic vowel pitch”, was then tested on both the unnormalized and normalized data. The power of each statistical test was computed and compared.

2. EXPERIMENT

2.1. Participants

Audio recordings were made of 90 Belgian Dutch speaking children imitating Dutch (non-)words. The mean chronological age of the children was 6 years (min. 5, max. 7). They all attended their first year of primary school in their region of birth.

2.2. Speech materials

Each child imitated 36 monosyllabic CVC stimuli. The vowel nucleus of each stimulus contained one of the 12 monophthongs of Belgian Standard Dutch, i.e. /i, ʏ, ɪ, ɛ, ɑ, ɔ, u, yː, eː, øː, aː, oː/ [8]. Each vowel occurred in three consonantal contexts: (i) /p_t/, (ii) /l_t/ and (iii) /t_r/, thus yielding phonotactically well-formed (non-)words.

The stimuli were read by a trained phonetician and the recordings of these were presented aurally to the participants who were asked to imitate her speech samples. Children's speech was recorded in a quiet room on a TASCAM DAT recorder by means of a head-mounted MicroMic II. The audio files were converted to WAV files by means of a TASCAM US 428 Digital Control Surface. The recording sessions with the children yielded a total of 7,985 speech samples.

Children's imitations were perceptually assessed by 6 expert listeners who identified all speech samples in which the vowels were correct renditions of the target vowels. From these, a subset of samples was selected according to the following criteria. The speech samples containing the corner vowels /i, u, a/ were selected from children who had produced each corner vowel at least twice and who had produced each of the 12 vowels at least once. The final data selection consisted of a total of 1,471 vowels produced by 47 children. For each of those 47 speakers, one randomly selected vowel per vowel category (N=564) was retained because of the requirements of some of the normalization techniques.

2.3. Acoustic analysis

The F0, F1, F2 and F3 of all the vowels were measured using a Python script through the Parselmouth API [9] of PRAAT [10]. F0 was determined using the standard autocorrelation algorithm. The maximum number of candidates was set to 15, the silence threshold to 0.03, the voicing threshold to 0.45, the octave cost to 0.01, the octave-jump cost to 0.35 and the voiced/unvoiced cost to 0.14. To minimise the number of potential octave jumps, the pitch floor and ceiling were set at 175 Hz and 425 Hz respectively after visual inspection of the pitch data with PRAAT's standard parameters. PRAAT's "Kill octave jumps" function was also applied. In addition, the F1, F2 and F3 of the vowels were measured. The number of formants was set to 5 and the formant maximum was set to 5500 Hz. Formant measurements were needed since one normalization technique (as described in [11]) requires them as variables.

2.4. Normalization procedures

The F0 normalization procedures used in the present study are listed in Table 1. The procedures were taken from published research on acoustic vowel normalization [3, 4, 5, 12] and in research dealing specifically with intrinsic vowel pitch [13, 14, 15].

Vowel normalization procedures require information about different vowels (i.e., vowel-extrinsic normalization) or only information about the vowel to be normalized (i.e., vowel-intrinsic normalization) [3]. Four types of F0 normalization techniques can be distinguished:

- rescaling methods transform the physical Hz scale to a perceptually relevant scale. Although they do not perform proper normalization [16], some scholars (e.g. [3, 15]) consider them as potential auditory normalization techniques, hence they were included in the present study;
- range normalization procedures normalize each vowel with respect to all the vowels of individual speakers [12];
- centroid normalization procedures normalize values relatively to a central tendency of the total distribution per participant [12];
- log-mean normalization techniques subtract a reference value from the log-transformed Hz value [12].

UNNORMALIZED FREQUENCY	
HZ	base condition
SCALE CONVERSION – VOWEL-INTRINSIC	
BARK	Bark-conversion of the Hz scale [17]
MEL	Mel-conversion of the Hz scale [18]
ERB	ERB-conversion of the Hz scale [19]
LN	Natural logarithmic conversion of the Hz scale [20]
ST-1	Semitone-conversion (ref. 1Hz) of the Hz scale [15]
ST-50	Semitone-conversion (ref. 50Hz) of the Hz scale [21]
ST-100	Semitone-conversion (ref.100Hz) of the Hz scale [22]
RANGE NORMALIZATION – VOWEL-EXTRINSIC	
GERSTMAN	Range normalization [23]
H & H	Rescaling between 0 to 1 [14]
LCE	Linear compression/expansion method (cf. [5])
CENTROID NORMALIZATION – VOWEL-EXTRINSIC	
LOBANOV	Z-score transformation [5]
TO MEDIAN	Normalization to median [13]
TO IQR	Normalization to IQR [13]
LOG-MEAN NORMALIZATION – VOWEL-EXTRINSIC	
NEAREY1	Single log-mean procedure [11]
NEAREY2	Shared log-mean procedure [11]

Table 1: Selected normalization procedures divided by types according to [3] and [12].

2.5. Statistical analysis

The group difference between the high and low vowels was investigated by means of paired t-tests. The *pwr.t.test* function of the *pwr* package [24] in R [25] was used to compute the power of the t-test. Its effect size was expressed as Cohen's *d* [26], i.e., the difference between the mean F0 of the high and low vowels divided by the pooled standard deviation for the two groups. The significance level was set to 0.05. The power was computed as a function of the

number of vowels per sample. For this purpose, the sample size was increased stepwise from 1 to 100 vowels.

3. RESULTS

The power of the t-tests which aimed to detect a group difference between the F0 of high and low vowels is presented in Fig. 1 as a function of sample size and normalization technique. Table 2 provides the sample sizes required per sample (i.e., separately for both high and low vowels) to attain a statistical power equal to 0.8, which is often considered a good power level [28], and 0.99, which is close to the maximal power.

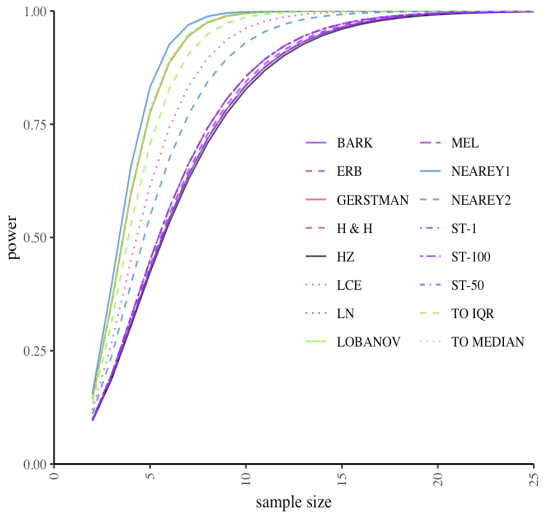


Figure 1: Results of the power analysis as a function of sample size per sample and normalization procedure.

Norm.	0.8	0.99	Norm.	0.8	0.99
HZ	9.5	19.3	GERSTMAN	5.2	9.1
BARK	9.3	19	H & H	5.2	9.1
ERB	9.1	18.5	LCE	6.6	12.5
LN	8.9	17.8	LOBANOV	5.2	9.1
MEL	9.3	18.9	TO MEDIAN	5.1	9
ST-100	8.9	18	TO IQR	5.7	10.5
ST-1	8.9	18	NEAREY1	4.8	8.2
ST-50	8.9	18	NEAREY2	7.4	14.3

Table 2: Sample size per sample needed to reach statistical power of 0.8 and 0.99 as a function of normalization procedure.

As can be seen in Fig. 1, the power curves of the different types of normalization clearly differ. The majority of the normalization techniques reach maximal power with a sample size close to 20. The unnormalized data in Hz (HZ) lead to one of the lowest power curves. Slightly higher power curves are achieved by scale conversion techniques (BARK, ERB, LN, MEL, ST-1, ST-100 and ST-50). It should be noted that the power curves of the three semitone conversions perfectly overlap. Range normalization procedures (GERSTMAN and H & H) result in higher power curves still. Relying on the same principle but differing in multiplication factor only, GERSTMAN and H & H exhibit the same power curves. The centroid normalization techniques (LOBANOV, TO MEDIAN and TO IQR) consistently lead to higher power curves: LOBANOV and TO MEDIAN yield the highest power curves of the three centroid normalization procedures. As far as log-mean normalizations (NEAREY1 and NEAREY2) are concerned, NEAREY1 represents one of the highest power curves. On the contrary, NEAREY2 exhibits a lower power curve.

Fig. 2 indicates that, relative to the unnormalized data, the difference in means/pooled *SD* ratio is very slightly modified by scale conversions. On the contrary, the log-mean, range and centroid normalization techniques clearly reduce the pooled *SD* relative to the difference in means.

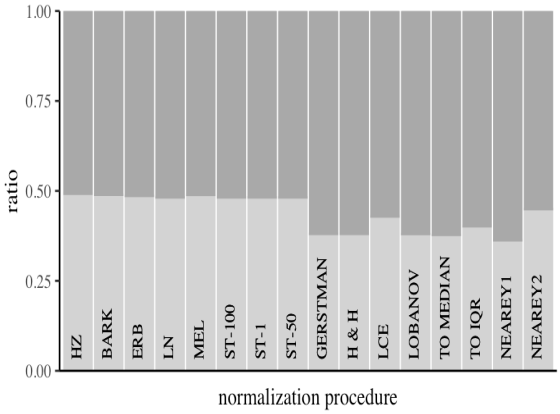


Figure 2: Ratio between the difference between the average F0 of high and low vowels (dark grey) and the pooled SD (light grey) as a function of normalization procedure.

4. DISCUSSION

This study aimed to investigate the effect of 15 normalization techniques on statistical power. 1,471 vowels from 47 speakers were analysed for F0. The inferential test that constituted the benchmark for this investigation was a paired t-test assessing the difference in F0 between high and low vowels.

The most important conclusion from this study is that data normalization influences the power of a statistical test. The results show that some normalization techniques achieve higher power with smaller data sets than others. The unnormalized and scale-converted normalized data tend to require larger sample sizes to achieve higher

power than log-mean, centroid, and range normalized measurements.

The differences between the normalization techniques can be explained by how the data are rescaled. Specifically, if the variability of the data is rescaled to a certain extent and if the subsequent effect size is rescaled to the exact same extent, the power is kept unchanged. If both the difference in means and the pooled *SD* are rescaled proportionally, the ratio between them remains unchanged. Consequently, the power of a statistical test based on normalized data is not affected. However, it is clear that some normalization techniques can significantly affect the difference in means/pooled *SD* ratio and as a consequence, statistical power.

The differences between the scale conversion techniques as opposed to the log-mean, range or centroid normalization techniques can be explained in terms of whether the effect size is rescaled to the same extent as the variability of the data, see Fig. 2.

Scale conversions [15, 17, 18, 19, 20, 21, 22] non-linearly transform the data so that only perceptually relevant differences are reflected on the normalized scale. However, larger differences between high and low vowels due to a higher average F0 remain larger on the normalized scale. The variability of the data is therefore not affected much.

Range [5, 14, 23], centroid [5, 13] and log-mean [11] normalization methods use per-subject reference levels so that the variability between subjects is reduced. The centroid, range or log-mean reference thus help to reduce the per-speaker variability. As such, with respect to the reduced pooled *SD*, the difference in means is proportionally more important, hence the increased power. This is why a given power level can be achieved with fewer data than with unnormalized measurements, despite the data being the same.

This means that normalizing F0 values can reduce the possibility that the null hypothesis is erroneously rejected. To put it another way, the increased power indicates that, while keeping effect size constant, the *a posteriori* reduced variability in normalized data can lead to a more confident rejection of the null hypothesis than what untransformed data would suggest. This should therefore be a matter of careful consideration when analysing the results of inferential tests. However, as the number of observations per sample increases, the statistical power is less affected. It should also be noted that a lower significance level, a lower effect size or increased variability in the data set is expected to increase the effect of the normalization procedures.

5. CONCLUSIONS

In the case of a large enough sample size, the choice of an adequate normalization procedure is only a matter of choosing a normalization that best levels out physiological variation while preserving phonemic and sociolinguistic information. However, if smaller sample sizes are used, the effect of normalizing on power should be considered because log-mean, range and centroid normalizations reduce variability and consequently enhance the statistical power as compared to unnormalized or perceptually rescaled values. Even if similar results are expected, a natural progression of this work is to analyse the effect of formant normalization methods on statistical power.

6. ACKNOWLEDGMENTS

Thanks are due to the participants of this study. The research reported in this paper was supported by a research grant from the Research Foundation – Flanders (FWO) [grant G004321N].

7. REFERENCES

[1] Ladefoged, P., Broadbent, D. E. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29(1), 98–104.

[2] van der Harst, S. 2011. *The vowel space paradox: A sociophonetic study on Dutch*. LOT.

[3] Adank, P., Smits, R., Hout, R. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America* 116(5), 3099–3107.

[4] Flynn, N. 2011. Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series* 2(11), 1–28.

[5] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49(2), 606–608.

[6] Qiu, X., Wu, H., Hu, R. 2013. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics* 14, 124.

[7] Tustison, N. J., Avants, B. B., Cook, P. A., Kim, J., Whyte, J., Gee, J. C., Stone, J. R. 2012. Logical circularity in voxel–based analysis: Normalization strategy may induce statistical bias. *Human Brain Mapping* 35(3), 745–759.

[8] Verhoeven, J. 2005. Belgian standard Dutch. *Journal of the International Phonetic Association* 35(2), 243–247.

[9] Jadoul, Y., Thompson, B., Boer, B. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71, 1–15.

[10] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot. International* 5(9), 341–345.

[11] Nearey, T. M. 1978. *Phonetic feature systems for vowels*. Indiana University Linguistics Club.
[12] Heeringa, W., Velde, H. 2018. The implementation of methods in Visible Vowels. *Vignette of the R package 'visvow'*.
[13] Chen, W.–R., Whalen, D. H., Tiede, M. K. 2021. A dual mechanism for intrinsic f0. *Journal of Phonetics* 87, 101063.
[14] Hoole, P., Honda, K. 2011. Automaticity vs. feature–enhancement in the control of segmental F0. In: Clements, G. N., Ridouane R. (eds.), *Where do phonological features come from*. John Benjamins Publishing Company, 131–171.
[15] Whalen, D. H., Levitt, A. G. 1995. The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23(3), 349–366.
[16] Clopper, C. G. 2009. Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass* 3(6), 1430–1442.
[17] Traunmüller, H. 1981. Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America* 69(5), 1465–1475.
[18] Stevens, S. S., Volkmann, J. 1940. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology* 53(3), 329–353.
[19] Glasberg, B. R., Moore, B. C. 1990. Derivation of auditory filter shapes from notched–noise data. *Hearing Research* 47, 103–138.
[20] Miller, J. D. 1989. Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America* 85(5), 2114–2134.
[21] Rietveld, A.C.M., Van Heuven, V. J. 2009. *Algemene fonetiek*. Coutinho.
[22] Fant, G., Kruckenberg, A., Gustafson, K., Liljencrants, J. 2002. A new approach to intonation analysis and synthesis of Swedish. *Speech prosody 2002*, 283–286.
[23] Gerstman, L. 1968. Classification of self–normalized vowels. *IEEE Transactions on Audio and Electroacoustics* 16(1), 78–80.
[24] Champely, S. 2020. *pwr: Basic functions for power analysis*. https://CRAN.R–project.org/package=pwr
[25] R Development Core Team 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
[26] Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
[27] Winter, B. 2019. *Statistics for linguists: An introduction using R*. Routledge.

VOWEL PRODUCTION CHANGES UNDER NOISE WITH CONSIDERATION OF LOW-ORDER FORMANT MASKING

Yasufumi Uezu, Masato Akagi, Masashi Unoki

Japan Advanced Institute of Science and Technology
{y-uezu, akagi, unoki}@jaist.ac.jp

ABSTRACT

There have been numerous studies on the effects of noise on speech production. However, it is not well understood how noise, which partially obscures the amplitude spectrum of speech, affects the lower-order formants (F1 and F2) that are crucial for vowel perception and production. This study aims to investigate the effects of noise type and noise level on the vocal intensity, amplitude and frequency of F1 and F2 during a speech task of Japanese vowels /a/ and /i/ by a solitary speaker under noise conditions. The noise was designed to mask either both formants, only F1, or only F2. Our findings showed that the amplitude and frequency of the vowel formants varied systematically with the type and level of noise, which we attributed mostly to the Lombard effect and formant masking. However, some results could not be explained by these factors, suggesting that auditory spectral representation may offer further insight.

Keywords: Vowel production, Lower-order formants, Masking noise, Lombard effect

1. INTRODUCTION

Sensory feedback, such as auditory feedback, plays a crucial role in speech motor control in the process of speech production. The Lombard effect [1–4], which refers to a reflexive increase in vocal intensity and fundamental frequency (f_o) in the presence of noise, has been well-known as a phenomenon in speech production under noisy conditions.

There have been numerous studies on the Lombard effect (i.e. [5–15]) that have used different types of broadband noise (e.g., white, speech-shaped, and bubble) and various sound pressure levels (SPLs). Some studies have also investigated the effects of specific types of noise present in real-world environments [16–19], as well as the impact of noise type and noise level on speech production using broadband and filtered noise [20, 21].

Lu and Cooke [20] investigated the effects of various noise types on vocal intensity, f_o, spectral center of gravity, and first formant (F1) frequency using a sentence reading task in the presence of noise. Five types of noise (broadband speech-shaped, two low-pass filtered, and two high-pass filtered) were presented to participants through headphones at 89 dB SPL. The results showed that all of these acoustic parameters were significantly increased in all conditions compared to the quiet condition. Stowe and Golob [21] examined the effects of various types and levels of noise on vocal intensity, f_o, and duration using a picture naming task in the presence of noise. Three types of noise (broadband, notched, and bandpass) were presented to participants through headphones at 75 or 90 dB SPL. The results showed that broadband significantly increased intensity, duration, and f_o, while notched had no effect, and bandpass increased intensity and duration to a lesser extent than broadband, but had no effect on f_o.

These researches have shown that noise that partially obscures the amplitude spectrum of speech can have varying effects on speech production depending on its type and level. However, it is not well understood how noise that partially masks the amplitude spectrum of speech affects the lower formants (F1 and F2), which are an essential factor in vowel perception and production. Additionally, there have been few studies on changes in formant amplitude in speech under noise, which should be considered in conjunction with formant frequency.

The purpose of this study is to investigate the effects of noise type and level on vocal intensity, formant amplitude, and formant frequency during a speech task involving the production of Japanese vowels by a solitary speaker in the presence of noise. The noise was designed to mask one or both of the F1 and F2 on the amplitude spectrum. We expected that formant amplitude would increase with increasing vocal intensity, regardless of the vowel type. In contrast, we hypothesized that the shift in formant frequency would be influenced by the type and level of the noise, as well as the type of vowel being produced.

2. EXPERIMENT

2.1. Participants and apparatus

Ten adult male native Japanese speakers participated in the experiment (mean 25.2 ± 3.79, 22-35 years

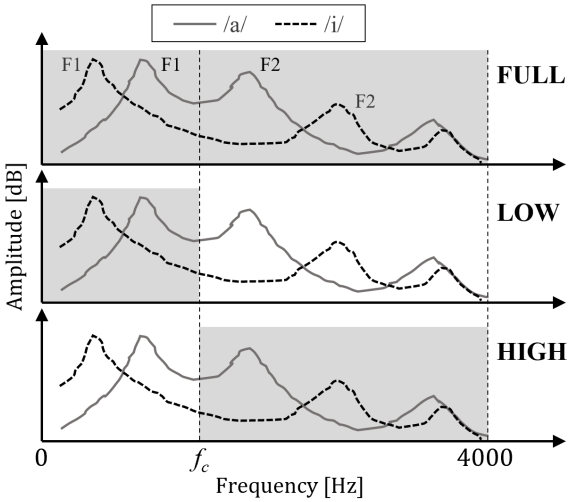


Figure 1: The amplitude spectrum of each noise type, the amplitude spectral envelopes of the vowels /a/ and /i/, and the formants (F1 and F2).

old). Prior to the experiment, all participants were screened to ensure that their speech and hearing were within normal range.

The experimental setup included a headset microphone (DPA d:fine), audio interface (Focusrite Clarett+ 4pre), mixer (YAMAHA MG10XUF), headphones (Sennheiser HD280Pro mk2), and laptop PC (Lenovo ThinkPad X1). A Z-weighted sound level meter (B&K Type2250) connected to an artificial ear (B&K Type4153) was used to calibrate the experimental setup such that a 1-kHz pure tone input to the headset microphone at 94 dB SPL was output through the headphones at 80 dB SPL. The sampling frequency was 16 kHz and the bit rate was 16 bits.

2.2. Stimuli

The back vowel /a/ and the front vowel /i/ were chosen as the speech targets among the Japanese vowels. The noise presented to the participants was generated through the following process: First, broadband (0 – 4 kHz) white noises NF75 and NF85 output through headphones with sound pressure level of 75 and 85 dB were created. Next, NL75 and NL85 were generated from NF75 and NF85 by processing them with a low-pass (0 – f_c kHz) filter, and NH75 and NH85 were generated by processing them with a high-pass (f_c – 4 kHz) filter. The cutoff frequency (f_c) for both filters were set to 925 Hz, which was the median of the mean F1 frequency and the mean F2 frequency of the vowel /a/ pronounced by 153 adult male Japanese native speakers [22]. Fig. 1 shows the amplitude spectrum of each noise type, the amplitude spectral envelope of the vowels /a/ and /i/, and the lower formants (F1 and F2). As

Table 1: Information on the total noise created, where L_N is the actual SPL of each noise presented through headphones.

Label	Noise type	Noise level	L_N(dB)
NF75	FULL	75	75.0
NF85	FULL	85	85.0
NL75	LOW	75	71.5
NL85	LOW	85	81.5
NH75	HIGH	75	73.0
NH85	HIGH	85	83.0

depicted in Fig. 1, on the amplitude spectrum, the noises of LOW and HIGH types were designed to mask only F1 and F2, respectively. Table 1 shows the information (label, noise type, and noise level) on the total noise created and the L_N is the actual SPL of each noise presented by the headphone.

2.3. Procedure

The sequence of the trials was determined by combining the order in which the random block method was applied to all combinations of the two vowel conditions (/a/ and /i/) and the seven noise presentation conditions (six noise and no-noise) with the order of their counterbalances. Thus, all participants performed all condition combinations twice each (a total of 28 trials). The experiment was set up with a 4-second trial duration and a 3-second interval between trials. During each trial, a speech target was displayed in hiragana (“あ” or “い”) on the laptop PC monitor, while noise was presented through the headphones, except for the trials in the no-noise condition.

The experiment was conducted in a soundproof room (background noise level $L_A < 30$ dB). Participants were instructed to sit in front of the monitor, wear a headset microphone and headphones, and vocalize the vowel for as long as possible, starting immediately after the hiragana was displayed on the monitor and ending immediately after the display ended. The speech sounds recorded by the headset microphone was provided with real-time auditory feedback through headphones. The speech sound, presentation noise, and headphone output sound for each trial were recorded simultaneously on the laptop PC through the audio interface.

2.4. Data processing

We extracted the interval between one and two seconds after vocalization onset from the recorded speech and used it as speech data. The following acoustic parameters were obtained from the speech data for each trial: The F1 and F2 frequencies were

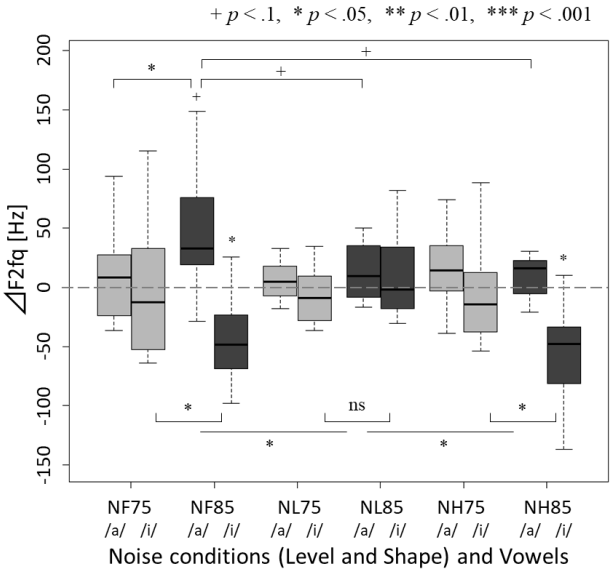


Figure 2: Changes in the F2 frequency $\Delta F2_{\mathrm{fq}}$ of the vowels /a/ and /i/ for each noise condition.

estimated using LPC analysis with an assumption of a pulse train [23] implemented as "To Formant (robust)..." in Praat [24], where the number of formants was set to 5, the formant ceiling was set to 5000 Hz, and the window length was set to 0.02 seconds. The vocal intensity in dB SPL and amplitudes of F1 and F2 in dB were calculated using MATLAB, where the window length was set to 0.02 seconds. To obtain F1 and F2 amplitudes, the cepstrum of the speech data was calculated, and peak picking around estimated F1 and F2 frequencies of the low-order cepstrum components was applied.

The changes in all acoustic parameters with and without noise (vocal intensity ΔL_{sp}, F1 amplitude $\Delta F1_{\mathrm{pk}}$, F2 amplitude $\Delta F2_{\mathrm{pk}}$, F1 frequency $\Delta F1_{\mathrm{fq}}$, and F2 frequency $\Delta F2_{\mathrm{fq}}$) were determined through the following procedure. First, the within-subject averages of the acoustic parameters were calculated for each vowel-noise condition. Next, the difference in the acoustic parameters between each noise condition and the no-noise condition was calculated.

2.5. Statistical analysis

A one-sample t-test was conducted on the change in each acoustic parameter of speech with and without noise, with p-values corrected using the Holm method (see Table 2). Based on the finding that the Lombard effect increases the vocal intensity, the ΔL_{sp}, $\Delta F1_{\mathrm{pk}}$, and $\Delta F2_{\mathrm{pk}}$ for amplitude were one-tailed tests with a change > 0.

To compare the change in each acoustic parameter across noise conditions, a repeated measures analysis of variance (RM-ANOVA) with 3 (noise type) $\times$ 2 (noise level) $\times$ 2 (vowel) was performed. If the assumption of sphericity was rejected according to Mauchly's test, the degrees of freedom were adjusted using Chi-Muller's epsilon. The p-values for multiple comparisons were corrected using the Shaffer method.

Table 2: Results of a one-sample t-test for the amount of change in acoustic parameters of speech with and without noise.

Noise	ΔL_{sp}	$\Delta F1_{\mathrm{pk}}$	$\Delta F2_{\mathrm{pk}}$	$\Delta F1_{\mathrm{fq}}$	$\Delta F2_{\mathrm{fq}}$
NF75	**	**	*	*	ns
NF85	*	*	*	***	ns
NL75	ns	ns	ns	ns	ns
NL85	*	*	*	ns	ns
NH75	*	*	+	ns	ns
NH85	ns	+	ns	ns	ns

3. RESULTS

3.1. Vocal intensity

In Table 2, the change in vocal intensity, ΔL_{sp}, was significant in all noises except NL75 and NH85. RM-ANOVA showed that the type $\times$ level interaction was significant ($F(1.35, 12.17) = 7.14, p < .05$), and the simple main effect of level on FULL and LOW types was significant or tended to be significant (FULL: $F(1,9) = 4.03, p = .07$; LOW: $F(1,9) = 17.11, p < .01$).

3.2. Formant amplitude

The changes in F1 amplitude $\Delta F1_{\mathrm{pk}}$ and F2 amplitude $\Delta F2_{\mathrm{pk}}$ exhibited a similar trend to ΔL_{sp} in Table 2. The RM-ANOVA results indicated that the type $\times$ level interaction tended to be significant for $\Delta F1_{\mathrm{pk}}$ ($F(1.2, 10.8) = 4.51, p = .05$) and the simple main effect of level on FULL and LOW type tended to be significant or had a significant (FULL: $F(1,9) = 3.92, p = .09$; LOW: $F(1,9) = 23.18, p < .001$). The type $\times$ level interaction ($F(1.51, 13.55) = 4.56, p < .05$) was significant for $\Delta F2_{\mathrm{pk}}$, and the simple main effect of level on LOW type was significant ($F(1,9) = 17.98, p < .01$). These suggest that formant amplitude increases in many noise conditions, and that it increases more with noise level when the noise type is LOW.

3.3. Formant frequencies

The change of F1 frequency $\Delta F1_{\mathrm{fq}}$ was significant for NF75 ($t[9] = 3.34, p < .01$) and NF85 ($t[9] = 6.51, p < .001$) in Table 2, while the RM-ANOVA results indicated that the simple main effect of level in FULL was significant ($F(1,9) = 11.94, p < .01$). This suggests that the F1 frequency shifts upward

exclusively in the FULL type condition, irrespective of vowel type, and that the higher the noise level, the more pronounced the upward shift becomes.

Fig. 2 illustrates the change in the F2 frequency $\Delta F2_{\mathrm{fq}}$ of the vowels /a/ and /i/ for each noise condition. In Table 2, $\Delta F2_{\mathrm{fq}}$ was not significant in any noise condition, but the RM-ANOVA results indicated that the interaction of type $\times$ level $\times$ vowel exhibited a significant trend ($F(1.34, 12.09) = 3.24, p < .1$), thus prompting the conduct of a 3 (type) $\times$ 2 (level) RM-ANOVA for each vowel.

The type $\times$ level interaction was significant ($F(1.31, 11.79) = 5.38, p < .05$) for $\Delta F2_{\mathrm{fq}}$ in the vowel /a/. The simple main effect of level was also significant when the noise was FULL type ($F(1,9) = 6.23, p < .05$). Additionally, the simple main effect of type was significant when the noise level was 85 ($F(1.57, 14.14) = 5.37, p < .05$), with significant trends between the FULL-LOW and FULL-HIGH conditions, respectively. This suggests that the F2 frequency in the vowel /a/ shifts upward when the noise exhibits a FULL type at a high level.

In contrast, the main effects of type ($F(1,9) = 6.89, p < .05$) and level ($F(1,9) = 8.50, p < .05$) were significant for $\Delta F2_{\mathrm{fq}}$ in the vowel /i/. Multiple comparisons on type indicated that it was significant between the HIGH-LOW ($t[9] = 3.34, p < .05$) and FULL-LOW ($t[9] = 2.28, p < .05$) conditions, respectively. This suggests that when the noise level is higher and the spectral shape of the noise includes the F2 frequency, the F2 frequency in the vowel /i/ shifts downward.

4. DISCUSSION AND SUMMARY

In this study, we investigated changes in the vowel production of a solitary speaker under various noise conditions. The results showed that the amplitude and frequency of the formants of the vowel speech sounds systematically varied according to the type and level of noise, suggesting that humans control not only vocal intensity but also articulation in response to various noise environments. However, since sustained vowel utterances may not fully reflect actual speech processes, future work should examine speech motor control during communication with others in noisy environments similar to those in the present study.

As expected, our results showed an increase in the amplitude of low-order formants and an increase in the vocal intensity in many noise conditions. These results indicated that the Lombard effect consistently enhances the amplitude of lower-order formants and vocal intensity. Contrary to our expectations, we found that the changes in the low-order formant frequency appear dissimilar in both F1 and F2.

Under FULL noise conditions, the F1 frequency consistently shifted upward in frequency regardless of vowel type. Additionally, as the noise level increased, the F1 frequency shift also increased. As shown in Fig. 1, FULL noise effectively masks vowel speech over a wide range of frequencies, and the higher the level, the stronger the masking effect. Therefore, it is suggested that speakers would have needed to further enhancement the signal-to-noise ratio of auditory feedback for speech under FULL noise conditions, which would have resulted in increased mouth opening and lower tongue position, leading to an upward shift in the F1 frequency.

Lu and Cooke [20] showed that the F1 frequency was higher under noisy conditions compared to quiet ones, regardless of the type of noise, but this finding was not entirely congruent with our results. This discrepancy may be due to various factors such as the speech target and the type and level of noise, as well as the presence or absence of female speakers and differences in formant estimation techniques.

The direction of the F2 frequency shift depended on the vowel and noise type. The F2 frequency of the vowel /a/ only shifted upward when the FULL noise was presented at a high level. Conversely, the F2 frequency of the vowel /i/ only shifted downward when the FULL or HIGH noises were presented at high levels. These results suggest that the noise which effectively masks the F2 of the vowel causes a shift in the F2 frequency when it is input to the auditory system at high level.

It is noteworthy that the F2 amplitude increased under high level of LOW noise, despite the noise not obstructing F2, as depicted in Fig. 1. This increase in F2 amplitude was not observed at low level of LOW noise. This finding suggests the limitations of explanations based on amplitude spectra, which may be hinted at by the spectral representation (i.e., excitation patterns) in the auditory peripheral system. Further research is needed to investigate the relationship between excitation patterns and formant masking.

5. ACKNOWLEDGMENTS

This work was supported by SCOPE (201605002) from the Ministry of Internal Affairs and Communications and by a Grant-in-Aid for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233).

6. REFERENCES

[1] É. Lombard, "Le signe de l'elevation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[2] H. Lane and B. Tranel, "The Lombard Sign and the Role of Hearing in Speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, Dec. 1971.

[3] J. Luo, S. R. Hage, and C. F. Moss, "The Lombard Effect: From Acoustics to Neural Mechanisms," *Trends in Neurosciences*, vol. 41, no. 12, pp. 938–949, Dec. 2018.

[4] S. Uma Maheswari, A. Shahina, and A. Nayeemulla Khan, "Understanding Lombard speech: A review of compensation techniques towards improving speech based recognition systems," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2495–2523, Apr. 2021.

[5] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[6] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech communication*, vol. 20, no. 1-2, pp. 13–22, 1996.

[7] M. Garnier, N. Henrich, and D. Dubois, "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 3, pp. 588–608, Jun. 2010.

[8] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[9] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Computer Speech & Language*, vol. 28, no. 2, pp. 580–597, 2014.

[10] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles," *Computer Speech & Language*, vol. 28, no. 2, pp. 629–647, Mar. 2014.

[11] S. Meekings, S. Evans, N. Lavan, D. Boebinger, K. Krieger-Redwood, M. Cooke, and S. K. Scott, "Distinct neural systems recruited when speech production is modulated by different masking sounds," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 8–19, Jul. 2016.

[12] J. Šimko, Š. Beňuš, and M. Vainio, "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 151–162, Jan. 2016.

[13] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Communication*, vol. 100, pp. 58–68, Jun. 2018.

[14] K. Marcoux and M. Ernestus, "Pitch in native and non-native Lombard speech," in *19th International Congress of Phonetic Sciences (ICPhS 2019)*. Australasian Speech Science and Technology Association Inc., 2019, pp. 2605–2609.

[15] C. Castro, P. Prado, V. M. Espinoza, A. Testart, D. Marfull, R. Manriquez, C. E. Stepp, D. D. Mehta, R. E. Hillman, and M. Zañartu, "Lombard effect in individuals with nonphonotraumatic vocal hyperfunction: Impact on acoustic, aerodynamic, and vocal fold vibratory parameters," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 8, pp. 2881–2895, 2022.

[16] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1988, pp. 331–334.

[17] C. E. Mokbel and G. F. Chollet, "Automatic word recognition in cars," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 346–356, 1995.

[18] J. Lee, H. Ali, A. Ziaei, and J. H. Hansen, "Analysis of speech and language communication for cochlear implant users in noisy lombard conditions," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5132–5136.

[19] A. J. Gully, P. Foulkes, P. French, P. Harrison, and V. Hughes, "The Lombard effect in MRI noise," in *Proc. of the 19th International Congress of Phonetic Sciences*, 2019, pp. 800–804.

[20] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1495–1499, Sep. 2009.

[21] L. M. Stowe and E. J. Golob, "Evidence that the Lombard effect is frequency-specific in humans," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 640–647, Jul. 2013.

[22] T. Hirahara and R. Akahane-Yamada, "Acoustic characteristics of Japanese vowels," in *Proc. 18th ICA*, 2004, pp. 3387–3290.

[23] C. H. Lee, "On robust linear prediction of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 5, pp. 642–650, 1988.

[24] P. Boersma, "Praat: Doing phonetics by computer," *http://www. praat. org/*, 2006.

SPEAKER-SPECIFIC ANTICIPATORY LABIAL COARTICULATION IN FRENCH

Fanny Guitard-Ivent[1,2], Louise Wohmann-Bruzzo[1], Nicolas Audibert[1], Cécile Fougeron[1]

[1]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS/Sorbonne Nouvelle)
[2]Laboratoire de NeuroPsychoLinguistique, Université de Toulouse, Toulouse, France
fanny.guitard-ivent@univ-tlse2.fr; {louise.wohmann, nicolas.audibert, cecile.fougeron}@sorbonne-nouvelle.fr

ABSTRACT

This study examines inter-speaker variability in anticipatory labial coarticulation in French. Production of /sy/ sequences by fourteen French speakers originating from two databases, are compared to unrounded counterparts (/si/-/se/), in terms of spectral lowering measured dynamically over the /s/ duration. Results show speaker-specific coarticulatory patterns: (a) in the degree of labial anticipation measured by the lowering of the center of gravity of /s/ in the rounded context; (b) in the span of anticipation with speakers who anticipate the /y/ much earlier than others; and (c) in the variability of the coarticulatory patterns across recording sessions, with some speakers being more stable in their labial anticipation than others. Results are discussed in relation to individual preferences in the coordination of speech units, which implementation in the speech signal may index information about the speaker.

Keywords: labial anticipation, coarticulation, speaker variability, fricatives, spectral center of gravity.

1. INTRODUCTION

Although we are used to segmenting speech into well-defined units, such as phonemes, syllables, etc., and representing them in a regular and more or less identical manner, we know that these units vary to a greater or lesser extent not only from one individual to another, but also from one production to another of the same individual. Inter-individual variability is usually related to speaker-specific physiology, but also to individual production strategies [1, 2, 3, 4].

In the context of a larger project looking at individual properties in the speech signal, the present paper focuses on speaker-specific patterns of coarticulation. Inter-speaker variability in coarticulation has been reported in various languages and for different types of contextual effects. For instance, individual differences have been observed for vowel nasalization in CVN and NVN sequences in American English [5], for lingual Consonant-to-Vowel coarticulation in Dutch [6] and in French [7], for labial V-to-C coarticulation in French [8], and for lingual anticipatory V-to-V coarticulation [9]. In their study with 10 French speakers producing non-sense iCy sequences (with non labial Cs) [8] showed that some speakers anticipate the protrusion gesture of the /y/ much earlier in the consonant than others.

The present study aims to investigate further this speaker-specific pattern of anticipatory labial coarticulation in French, but in continuous and meaningful speech, and by looking at its acoustic consequences. Indeed, if coarticulation can be used to perceptually differentiate speakers as suggested by [10], it needs to have acoustic consequences. For this reason, we investigate /sy/ sequences, where anticipation of the labial gesture of the vowel can be tracked dynamically on the spectral characteristic of the fricative noise. Several studies have shown that fricatives exhibit high inter-speaker variability [11, 12, 13]. Based on the fact that there is speaker-specific information both in /s/ and, presumably, in labial anticipation, this sequence is particularly interesting.

Speaker-specific patterns will be defined by looking at different aspects: the amount of contextual acoustic difference between a [s] produced in a rounded context compared to an unrounded context; the span of anticipation as measured by dynamic acoustic variation over the /s/; but also by comparing speakers in terms of variability of their coarticulatory patterns across recording sessions.

2. METHOD

In order to study coarticulation on the production of 14 speakers in total, we have merged here productions extracted from two different corpora of read speech. The first corpus (corpus A) is extracted from the French PATAFreq database [14] that included 6 female speakers (F01 to F07) and 3 male speakers (H01-03) recorded in 8 to 10 sessions, over a two-month period. Each session includes the reading of three short texts, which contain 6 words

presenting a [sy] syllable and 6 words containing a [se] syllable (unfortunately no /si/ were present in the texts). In total, this corpus includes 24 to 60 /sy/ per speakers and 24 to 62 /se/ per speaker. The second corpus (corpus B) includes the production of five French female speakers (AE, AN, AP, EDF, ES) recorded in 5 sessions over two weeks [15]. In each session, speakers read a set of sentences, among which two contained [si] and [sy] sequences, and these sentences were repeated 9 times in a random order. In total, this corpus contains 135 /sy/ per speakers and 108 to 134 /si/ per speaker.

In each sound file, the fricative /s/ and the following vowel were segmented manually. The spectral Centre of Gravity (COG) was computed over the /s/ on 10 equally spaced points with a 10 millisecond Hanning window centered on the target point, and filtered to retain frequencies ranging from 350Hz to 18kHz.

COG trajectories were analyzed using GAMMs on R with the package mgcv [16]. Our analyses include autoregressive error models (AR1 models) that take into account the dependence between neighboring points within the same COG trajectory. Estimates of the difference between the rounded and the unrounded contexts were computed using the function get_difference of package itsadug [17], at each time point. Using the same method, pairwise differences between speakers (all sessions pooled) give a measure of inter-speaker variability, and pairwise differences between sessions of a given speaker give a measure of intra-speaker variability.

3. RESULTS

3.1 Is coarticulation speaker dependent?

COG trajectories in rounded and unrounded V2 contexts were modeled for each speaker separately, taking into account the duration of /s/ and including a random smooth per repetition.

All 14 per-speaker models account for 93 to 95% of the variance in our data and indicate an effect of V2 on the shape of the COG trajectories of /s/ as well as on the average COG frequency of these trajectories, except for one male speaker H03. For this speaker, the COG of /s/ was not significantly lowered in the rounded vs. unrounded context. For all 13 other speakers, as expected, the COG of /s/ is significantly lowered and its trajectory is affected by the rounding of the following vowel.

Figure 1 highlights the differences found between speakers in (a) the degree of anticipatory labial coarticulation, displayed as the magnitude of the difference in COG between the two V2 contexts, with darker shades showing more lowering of COG in the rounded V2 context, but also (b) in terms of span of anticipation, illustrated by the proportion of /s/ which is significantly different in the two contexts as indicated by the dark red line overlaid on the x-axis. The difference is considered significant when the 95% credible interval of the difference does not include value 0.

Speakers can be separated into subgroups based on differences in COG trajectories between the unrounded and rounded context. Speakers F02, F05, F07, H01, H02, AN, EDF, ES, F01 show a striking difference between the two contexts both in magnitude of the acoustic difference and in the span of anticipation. AE and F03 show an early start of anticipation but, if the acoustic difference is on the largest part of the /s/, it is not that strong in magnitude, like F04 for which in addition the effects of coarticulation appear much later. AP and H03 present little acoustic differences, which appear only at the end of the /s/.

3.2. Are speakers who strongly coarticulate stable in their coarticulatory patterns across sessions?

To measure whether the variability in coarticulatory pattern is speaker dependent, we selected the 9 speakers who coarticulate the most (see Figure 2) and modeled the trajectories of their /s/ in /y/ context as a function of the speaker and the recording session (8 sessions for speakers of corpus 1 and 5 for those of corpus 2). Our models explain 92% of the variance for corpus 1 and 95% for corpus 2.

In order to focus on intra- versus inter-speaker variability, we report the results of the pairwise comparisons only (see section 2). Regarding individual variability, some speakers are more variable than others across sessions. While speakers like EDF and F07 are very consistent in their coarticulatory patterns, other speakers like F02 and H02 are much more variable. Moreover, variability patterns do not depend on the overall degree of coarticulation: for instance speaker F07 is both a strong and stable coarticulator, while speaker F02 is a strong and variable coarticulator.

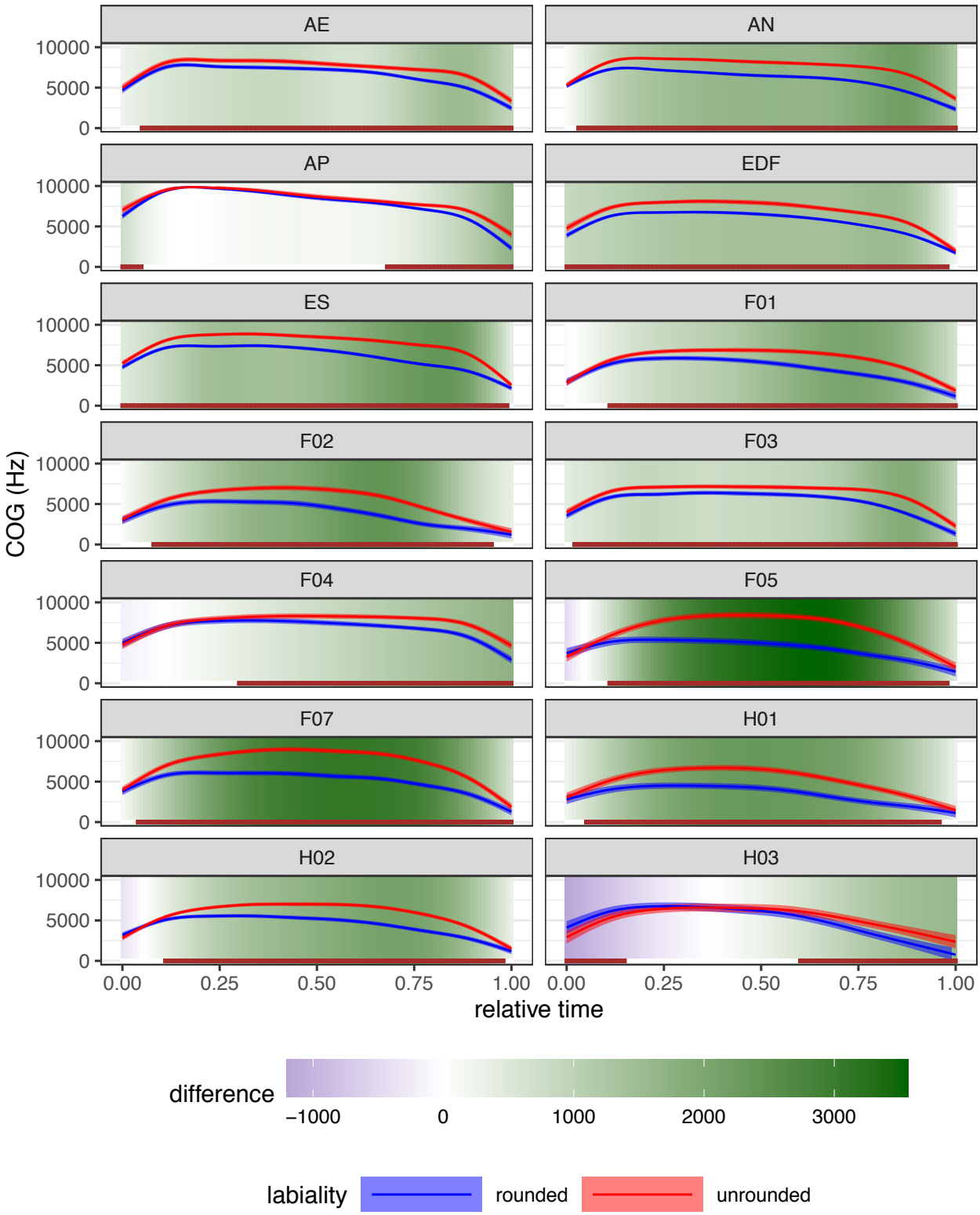


Figure 1: COG trajectory of /s/ estimated by GAM models as a function of V2 (rounded in blue and unrounded in red) for each speaker. The background color quantifies the difference between the two trajectories. Highlighting of the x-axis in dark red indicates that the difference is significant.

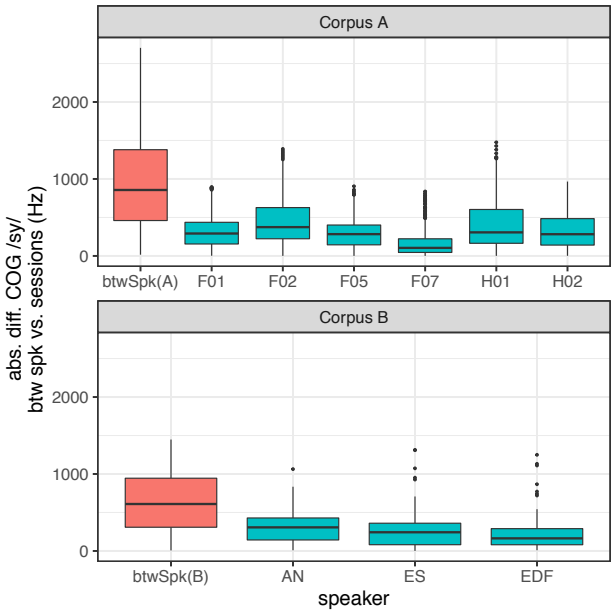


Figure 2: Distribution of absolute pairwise differences between speakers (all sessions pooled, in red) as a measure of inter-speaker variability, and between sessions of a given speaker (in blue) as a measure of intra-speaker variability, for both corpora.

4. DISCUSSION AND CONCLUSION

In this study we confirm with 14 speakers, numerous tokens per speaker, and in continuous (read) speech production, that the anticipation of the labial gesture of /y/ in a preceding /s/ is speaker-specific. Thanks to the dynamical tracking of the spectral characteristics of the fricative noise, we can better characterize these individual coarticulatory patterns. Indeed, speakers differ both in terms of the amount of spectral lowering of the fricative noise induced by the anticipation of the labial gesture of /y/, and in terms of the shape of the COG lowering trajectory indicating how much of the /s/ duration is affected by the anticipation of this labial gesture. While most of the speakers show a large amount of spectral lowering and an early anticipation, quite a few speakers show very little acoustic cues of coarticulation (H03 & AP), and one speaker appears to anticipate the labial gesture later than the others (F04).

Moreover, our comparison of the same speakers over multiple recording sessions separated by several days, showed that for most of the speakers showing labial anticipation, there is little variation in the way they coarticulate across sessions. These results mirror that of [18] who observed C-to-V coarticulation across 300 tokens of words produced by the same speaker recorded in three sessions. The vowel acoustic at its mid point and the formant trajectories showed a striking stability across repetitions, suggesting that both the vowel targets and the trajectories out and to the surrounding consonants are controlled and planned by the speaker. Our results and that of [11], who showed large individual variation in the COG trajectories of fricatives (especially for /s/) depending on the surrounding vowels, further support individual preferences in the coordination of speech units, whose implementation in the speech signal may index speaker-specific information.

Individual patterns of coarticulation, like other types of speaker-specific articulatory aspects, could be attributed to physiological and anatomical characteristics of the speaker (e.g. [13]), but also to speaker-specific strategies to attain a desirable speech rate. In a study of 246 speakers, [9] showed that anticipatory Vowel-to-Vowel coarticulation is stronger for speakers with the fastest speech rates, but only for younger speakers. We did not measure speech rate on each of the sentences produced in our data, however, post hoc observation of the degree of lowering of the COG of /s/ as a function of its duration for each speaker does not seem to indicate a link between the two. Indeed, speakers seem to be constant in their degree of coarticulation regardless of the duration of /s/. Moreover, our models take into account the duration of the consonant and for the majority of them (9 / 14) this parameter is not significant. In other studies with the speakers included in corpus A [19, 20] we found that speaker F07, who present the most stable coarticulatory pattern, is quite variable in speech rate across sessions, while speaker F02, who show more variable coarticulation, has a stable speech rate across sessions. Future studies are thus desirable to understand better how these individual patterns of labial coarticulation relate to articulation rate and to other speaker-specific articulatory and coarticulatory behaviors.

5. ACKNOWLEDGEMENTS

This study was funded by the subside CRSII5_202228 (ChaSpeePro) of the Swiss National Science Foundation, and by the programs ANR-10-LABX-0083 (Labex EFL) & ANR-17-CE39-0016 (VoxCrim).

5. REFERENCES

[1] Atkinson, A. 1976. Inter- and intraspeaker variability in fundamental voice frequency. *J. Acoust. Soc. Am.* 60, 440-445.

[2] Smith, R., Hawkins, S. 2012. Production and perception of speaker-specific phonetic detail at word boundaries. *J. of Phonetics*. 40, 213-233.

[3] Serrurier, A., Badin P., Lamalle L., Neuschaefer-Rube C. 2019. Characterization of inter-speaker articulatory variability: A two-level multi-speaker modeling approach based on MRI data. *J. Acoust. Soc. Am.* 145 (4), 2149-2170.

[4] Toda, M. 2009. Étude articulatoire et acoustique des fricatives sibilantes. PhD dissertation. Université de la Sorbonne Nouvelle, Paris.

[5] Zellou, G. 2017. Individual differences in the production of nasal coarticulation and perceptual compensation. *J. of Phonetics*. 61, 13–29.

[6] van den Heuvel, H., Cranen, B., & Rietveld, T. 1996. Speaker variability in the coarticulation of /a, i, u/. *Speech communication*. 18(2), 113-130.

[7] Guitard-Ivent, F. 2018. Coarticulation C-à-V en français : interaction avec le type de voyelle, la position prosodique et le style de parole. PhD dissertation. Université de la Sorbonne Nouvelle, Paris.

[8] Robert, V., Wrobel-Dautcourt, B., Laprie, Y., & Bonneau, A. 2005. Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for French. In: *5th Conference on Auditory-Visual Speech Processing-AVSP*.

[9] D'Alessandro, D., Fougeron, C. 2021. Changes in Anticipatory VtoV Coarticulation in French during Adulthood. *Languages*. 6(4), 181.

[10] Zellou, G. 2022. Coarticulation in Phonology. In: Cambridge Elements in Phonology Series. Cambridge: Cambridge University Press.

[11] Carralero-Fernandez, S. 2022. Speaker Specific Information in the Acoustic Characteristics of English Fricatives. *Int. J. Forensic Linguistics.* 3, 105-115.

[12] Gordon, M., Barthmaier, P. & Sands, K. 2002 A cross-linguistic acoustic study of voiceless fricatives. *J. Int. Phonetic Association*. 32, 141-171.

[13] Fuchs, S. & Toda M. 2010. Do differences in male versus female /s/ reflect biological or sociophonetic factors?. In Fuchs, Zygis, Toda (eds) *Turbulent sounds: An interdisciplinary guide*. Mouton de Gruyter.

[14] Fougeron, C., Audibert, N., Gendrot, C., Chardenon, E., & Wohmann, L. 2022. PATATRA and PATAFreq: two French databases for the documentation of within-speaker variability in speech. *Proceedings of 13th LREC*, 1939-1944.

[15] D'Alessandro, D. 2022. Interspeaker variations in V-to-V coarticulation: effects of Motor Speech Disorders, age, speech tempo changes, and boundary type. PhD dissertation. Université de la Sorbonne Nouvelle, Paris.

[16] Wood, S. 2011. Fast stable restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 73(1), 3–36.

[17] van Rij, J., Wieling, M., Baayen, R. & van Rijn, H. 2022. itsadug: Interpreting time series and autocorrelated data using GAMMs. R Package, version 2.4.1

[18] Whalen, D. H., & Chen, W. R. 2019. Variability and central tendencies in speech production. *Frontiers in Communication*, 4, 49.

[19] Fougeron, C., Audibert N. 2022. Variabilité intra-individuelle en parole lue. In: *32e Journées d'Études sur la Parole*. Noirmoutier, France

[20] Audibert, N., Fougeron, C. 2022 Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population. In: *Proc. Interspeech 2022*, 4755-4759.

VIABLE SIGNAL PERIODICITIES IN SPEECH RHYTHM

Jessica Campbell, Dani Byrd, Louis Goldstein

University of Southern California, United States
jac95339@usc.edu, dbyrd@usc.edu, louisgol@usc.edu

ABSTRACT

Despite well-known rhythmicities in spoken language, neither strict nor absolute isochrony is present; linguistic structures are not located at equal intervals. An alternative periodicity-based approach to linguistic rhythm instead considers quasicyclic or oscillatory properties of the speech signal—in articulation or acoustics—over extended stretches of speech. Additionally, while periodicity in speech acoustics has been implicated in neural entrainment, it is unknown if periodicity in the articulatory domain is also present in a way that could promote intelligibility. In the present study, the stability of an articulatory spatiotemporal modulation function calculated from kinematic point tracking data and a correlated acoustic modulation function calculated from MFCCs is compared to the stability of amplitude envelope modulation, which is shown to reflect vowel onsets and lexical stresses. Results demonstrate that both types of modulation functions are comparable in stability to the amplitude envelope, indicating their potential viability for speech perception and/or production processing.

Keywords: Rhythm, Modulation, Amplitude Envelope, Articulation, Speech Timing

1. INTRODUCTION

While language exhibits rhythmic characteristics, strict or absolute isochrony is not present. An analytic approach that leverages the stability of quasi-cyclic or oscillatory properties of the speech signal—either in production or acoustics—over longer stretches of speech can provide a means for defining rhythm as durational regularity. Such a global (i.e., long timescale) approach to modulation can be calculated irrespective of language, speaking task, or speaker and may be influenced by or reflect linguistic structures [7], [9], [17], [14]. We call such an approach a "periodicity approach."

Models of rhythm incorporating a periodicity approach vary regarding which specific oscillatory properties they index. In the acoustic domain, these oscillatory properties tend to match frequencies at which neural entrainment occurs [12], [1], [8], [14], [9] and thus may aid in speech perception [5], [16], [3]. The most common periodicity approach uses the fluctuating speech acoustic amplitude envelope (e.g., [17], [14], [11], [4]). In Tilsen and Arvaniti's [17] acoustic amplitude mode decomposition method, a series of filters is applied to the speech signal to extract the quasi-cyclic fluctuations of amplitude during the speech stream. The amplitude envelope can then be decomposed using a sifting process into a series of *intrinsic mode functions*, which reflect amplitude oscillations at increasingly longer timescales [17]. The first (i.e., shortest timescale) of these functions roughly corresponds to the syllable-level timescale, and the second to the stress-foot-level time scale; however, Tilsen and Arvaniti [17, p. 631] establish this correspondence only qualitatively, stating that they "observ[ed] it to hold true in the majority of cases…examined by inspection."

Oganian and Chang [14], in contrast, calculate the derivative of the positive amplitude envelope, so that instead of fluctuations in amplitude, the derived global signal captures fluctuation in instantaneous amplitude change. The instants of peak change in amplitude, they argue, are likely involved in speech perception, because they found that these peak "events" produce time-locked response in the human superior temporal gyrus [14]. Periodicity approaches' relevance to neural events may aid in their interpretability in terms of the cognitive activities involved in speech production and perception (e.g., [7]).

While speech amplitude envelope methods of tracking global periodicity capture oscillations of the acoustic signal, another method for indexing speech rhythm is feasible—the spatiotemporal modulation function approach [6]. This approach can capture rhythmicity in either an articulatory or an acoustic signal, quantifying the oscillatory properties of the rate of global change in the vocal tract or in the acoustic spectrum over time. We speculate additionally that neural responses can entrain to the modulation function (in addition to or alternatively to the amplitude envelope), which would potentially implicate the modulation signal in production, perception, and learning processes. That said, for entrainment to be possible, the modulation pulse frequency must fall in an appropriate and sufficiently stable frequency range across and within speakers. Specifically, for viability in linguistic neurocognitive processes, the appropriate frequency range would be expected to be in or near the theta band range, 4-8Hz [15]. In sum, temporal stability in articulatory and

concomitant acoustic modulation pulse rate is predicted across speakers such that the average frequency of modulation is expected to be resistant to variation, as other periodicity approaches have been found to be. The present study uses articulatory and acoustic modulation data to assess this prediction.

2. METHODS

Readings of the "Grandfather Passage" [2] by nine speakers were sourced from the X-Ray Microbeam (XRMB) Corpus [19]. Two analyses were conducted. The first compares the frequencies and temporal stability of the articulatory and acoustic modulation pulses to those of peaks in the differentiated amplitude envelope. The second analysis explores linguistic drivers of this periodicity by comparing the frequencies of acoustic vowel and stressed vowel onsets to the first and second intrinsic mode functions of the amplitude envelope, respectively.

2.1. Derived variables

The articulatory modulation function was calculated to index instantaneous change of the global articulatory posture over time as captured by pellet positioning on the articulators, and the acoustic modulation function was calculated to index instantaneous spectral change over time as captured by MFCCs. These calculations were conducted following [6], which also used XRMB data. Specifically, the articulatory modulation function was calculated as the squared Euclidean distance between the 14-dimensional vector of XRMB marker positions at successive time samples. To calculate the acoustic modulation function, the difference across successive frames of the first 13 MFCC parameters were squared and summed. Each function was then smoothed using a low-pass 12 Hz-cutoff, ninth-order Butterworth filter. "Pulses" are defined as peaks in a modulation function and were calculated by identifying local maxima of the function.

The acoustic amplitude envelope was calculated following [17] for each speech analysis interval (see below). Specifically, the acoustic signal was bandpass filtered using a fourth order Butterworth filter with cutoffs [400, 4000] Hz to determine vocalic energy, then lowpass filtered with a fourth order 10Hz Butterworth filter. The function was then normalized by subtracting the mean and dividing by the maximum of the function and windowed using a Tukey window. To make the signal comparable to the modulation functions, the amplitude envelope was differentiated and then filtered again with a fifth order 12Hz lowpass Butterworth filter. To evaluate the correlation between the intrinsic mode functions and driving linguistic structures posited by [17], the non-differentiated function was sifted using Matlab's *emd* function into two intrinsic mode functions (S. Tilsen, personal communication, February 6, 2022), which were also filtered with a fifth order 12 Hz lowpass filter. Peaks in the differentiated amplitude envelope, as well as the first two intrinsic functions of the non-differentiated amplitude envelope, were designated as local maxima in the function.

Means and variability in frequency were calculated for both articulatory and acoustic modulation pulses and peaks in the differentiated amplitude envelope for each analysis interval (defined below). The reciprocal of the periods between timepoints provided frequencies for each timeseries. The mean frequencies of the first and second intrinsic mode functions of the amplitude envelope and of acoustic vowel onsets and stressed vowel onsets (defined below) were likewise calculated.

2.2. Analysis intervals & defining linguistic timepoints

Specific windows of speech analysis were determined based on phrasal groupings by identifying potential phrase boundaries in the passage. However, since not all speakers produced a pause at every one of these boundaries, only boundaries occurring with a pause greater than 100ms were used for segmenting the passage into analysis intervals. (Pauses due to disfluencies were not used to define analysis intervals.)

To determine the timepoints of linguistic structures, each of the passage's syllables in the acoustic signal was annotated as having either primary stress ("stressed") or non-primary stress ("unstressed"), consulting the CMU Pronouncing Dictionary [18] in cases of ambiguity. The acoustic onset of the vowel in each syllable was selected as the syllable's timepoint of occurrence.

While phrase boundaries were used to demarcate analysis intervals, some participants produced (ungrammatical) filled or silent pauses interior to phrases as defined here. Pulse and peak intervals during pauses longer than 100ms (as marked by the Penn Force Aligner [20]) were excluded in data cleaning, and vowel onset and stressed vowel onset interval measures spanning such pauses were also excluded.

2.3. Statistical analysis

To compare the vowel onset timeseries and the first intrinsic mode function of the amplitude envelope as well as the stressed vowel onset timeseries and the second intrinsic mode function, two linear mixed effects models were fitted using the R package *lme4*, predicting frequency by timeseries (vowel onset and IMF1 for the first model; stressed vowel onset and

IMF2 for the second) with random intercepts grouped by speaker. Timeseries in each model was contrast coded, and frequency was centered and rescaled.

To calculate *within* speaker variation of the two modulation pulses and peaks in the differentiated amplitude envelope, the coefficient of variation (standard deviation normalized by mean) was calculated for the frequencies occurring in each speaker's analysis intervals; the mean of these coefficients of variation provided the average variability for each individual speaker. Differences between coefficients of variation of the three timeseries were tested using two modified signed-likelihood ratio tests (MSLRT) for equality of coefficients of variation [10] (R package *cvequality*, Version 0.1.3; [13]). Two tests for each speaker are relevant to our predictions; these compared for each speaker (i) the coefficients of variation for articulatory modulation pulses and the differentiated amplitude envelope and (ii) the coefficients of variation for acoustic modulation pulses and the differentiated amplitude envelope. Bonferroni corrections were applied to account for the 18 pairwise comparisons; p values were therefore considered significant at $p \leq 0.003$.

To calculate *between* speaker variation for the same timeseries, for all speakers the coefficient of variation was calculated across the frequency means of the two modulation functions and the amplitude envelope timeseries. Mixed effects models for the frequencies of the three timeseries were fitted separately on all speakers' pooled data using the R package *lme4*, predicted by random intercepts grouped by speaker. Residuals from each model were squared and divided by the mean frequency of the timeseries with all speakers pooled, and two linear models—the first comparing articulatory modulation pulses and the differentiated amplitude envelope, and the second comparing acoustic modulation pulses and the differentiated amplitude envelope—were fitted to predict residuals by timeseries.

3. RESULTS

The mean frequency of articulatory modulation pulses was 8.18Hz (SD = 2.31), and the mean frequency of acoustic modulation pulses was 9.12 (SD = 3.03). These signals were highly periodic; means for each speaker ranged from 7.76Hz to 8.49Hz (articulatory) and 8.25Hz to 9.66Hz (acoustic). The frequency range of modulation pulses was generally higher than the frequency range of the differentiated amplitude envelope; the envelope mean was 7.19Hz (SD = 2.47), with a range of speaker means from 6.89Hz to 7.65Hz. NB: The auditory cortex is most sensitive to modulations at frequencies between 2 and 8Hz [7]; the mean frequencies of each timeseries falls within or very close to this range.

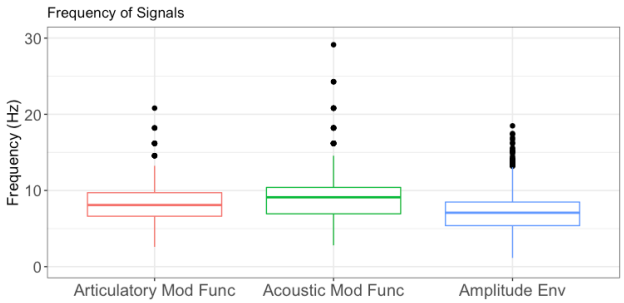


Figure 1: Frequencies of three periodic signals (each significantly different from the others $p < 0.001$). One point (48.5Hz) removed from middle boxplot for space.

3.1. Linguistic drivers of amplitude oscillation

Before turning to the variability of these signals, we assess the previously alluded to correlation [17] between the frequencies of the first two intrinsic mode functions of the amplitude envelope and linguistic structures. The first and second intrinsic mode function of the (undifferentiated) amplitude envelope were compared respectively to the frequencies of the vowel onsets of each syllable and the vowel onsets of each stressed syllable. These frequencies were expected to be non-distinct for the first intrinsic mode function (IMF1) and vowel onsets and for the second intrinsic mode function (IMF2) and stressed vowel onsets. Our results indicate no difference between the frequencies of IMF1 and vowel onsets. The mean frequency of IMF1 with all speakers pooled was 6.52Hz (SD = 3.15); the mean frequency of vowel onsets was 6.32Hz (SD = 4.49). A linear mixed effects model predicting frequency by timeseries with random intercepts grouped by speaker showed that there was no significant effect of timeseries (β = -0.027, SE(β) = 0.025, $t(1683)$ = -1.07, p = 0.28) Similarly, no difference was found between the frequencies of IMF2 and stressed vowel onsets. With speakers pooled, the mean frequency of IMF2 was 2.64 (SD = 0.89), and the mean frequency of stressed vowel onsets was 2.78 (SD = 1.50). A linear mixed effects model predicting frequency by timeseries with random intercepts grouped by speaker showed that there was no significant effect of timeseries (β = 0.061, SE(β) = 0.038, $t(695.4)$ = 1.59, p = 0.11). Therefore, Tilsen and Arvaniti's [17] qualitative observation that the first two intrinsic mode functions of the amplitude envelope correspond to linguistic structures is generally supported.

3.2. Variability of periodic signals

We turn now to variability of the frequencies of modulation pulses and how it compares to that of the

amplitude peaks, considering this both within and across speakers. The mean coefficient of variation within speakers was 28.1% for articulatory modulation pulses, 31.9% for acoustic modulation pulses, and 34.4% for differentiated amplitude envelope peaks. Figure 2 shows variation within each speaker.

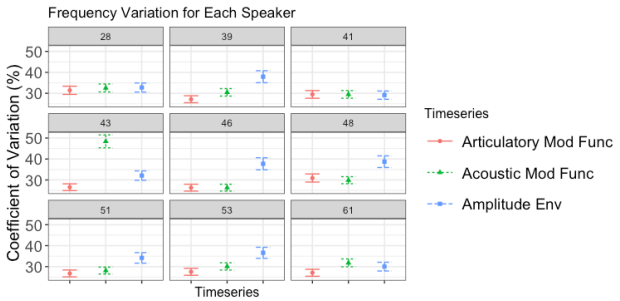


Figure 2: Coefficient of variation of the frequencies of three periodic signals. Error bars represent standard error of coefficient of variation.

Three speakers (speakers 39, 46, & 53) showed significantly lower variation in articulatory modulation pulse frequencies than differentiated amplitude envelope peak frequencies (p = 0.001, 0.0002, and 0.002, respectively); all others were not significantly different. One speaker (speaker 43) showed significantly lower variation in amplitude envelope peak frequencies than in acoustic modulation pulse frequencies, and another (speaker 46) showed significantly greater variation ($p < 0.001$ for these speakers); the rest were not significantly different. In sum, most speakers did not have any significant difference in variation between either type of modulation pulse and peaks in the differentiated amplitude envelope.

To determine between-speaker variability, the mean frequency of each timeseries was calculated for each speaker. The coefficient of variation was then calculated across these nine frequencies for each timeseries. These values were 3.41% for articulatory modulation pulses, 4.97% for acoustic modulation pulses, and 3.30% for differentiated amplitude envelope peaks. Fig. 3 shows variation with all speakers pooled; note that this figure shows higher coefficients of variation than the numbers stated above because it includes within-speaker variation, as well.

Models for between-speaker variability yielded significantly lower variability between speakers for articulatory modulation pulse frequencies than for differentiated amplitude envelope pulse frequencies (β = 0.20, SE(β) = 0.06, t(2476) = 3.56, $p < 0.001$), and no significant difference in variability between speakers for acoustic modulation pulse frequencies than for differentiated amplitude envelope pulse frequencies (β = -0.14, SE(β) = 0.15, t(2631) = -0.95, p = 0.34). The between- and within-speaker variability results indicate the stability of both articulatory and acoustic modulation pulses.

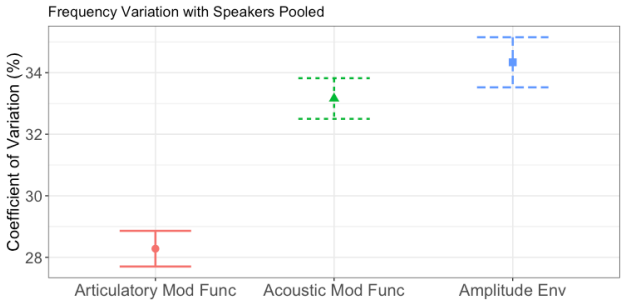


Figure 3: Coefficient of variation of the frequencies of three periodic signals with all speakers pooled. Error bars represent standard error of coefficient of variation.

4. DISCUSSION AND CONCLUSION

The goals of this study were to assess the articulatory and acoustic global modulation functions as an index of rhythmicity in the context of implications for spoken language production and perception processes by comparing them to the speech acoustic amplitude envelope periodicity. Further evaluated were claims that derived functions from the amplitude envelope capture specific levels of linguistic structure.

The modulation pulses were highly periodic in both articulation and acoustics, and their frequencies, like those of the amplitude envelope, fell within or slightly above the range at which the auditory cortex is reported to best entrain with speech [15]. Further, the first two intrinsic mode functions of the amplitude envelope matched the frequency of vowel onsets and stressed vowel onsets, respectively. The frequency of articulatory and acoustic modulation pulses were found to be comparably robust against within-speaker and between-speaker variation when compared with peaks in the differentiated amplitude envelope. The stability of these signals contributes to predictability and could enhance intelligibility and perceptual processing, in addition to foundational internal processes of speech planning and motor control that are thought to require 'binding' of the articulatory and acoustic dynamics [6].

Given that all three periodic signals are extracted from the same speech stream, whether in articulation or acoustics, each may be a window to a common underlying rhythmic phenomenon in speech. Further, given that the amplitude envelope was shown to reflect linguistic structural information of different timescales, the signals may be viable links between abstract linguistic elements and concrete oscillatory acoustic and articulatory patterns in the speech signal. Neural entrainment may then facilitate speech perception by efficiently estimating linguistic structures during listening processes.

5. REFERENCES

[1] Assaneo, M. F., & Poeppel, D. 2018. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, 4(2), 1–9.

[2] Darley, F. L., Aronson, A. E., Brown, J. R. 1975. *Motor Speech Disorders*. W. B. Saunders Co.

[3] Drullman, R., Festen, J. M., & Plomp, R. 1994. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5), 2670–2680.

[4] Frota, S., Vigário, M., Cruz, M., Hohl, F., & Braun, B. 2022. Amplitude envelope modulations across languages reflect prosody. *Speech Prosody 2022*, 688–692).

[5] Ghitza, O., & Greenberg, S. 2009. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66, 113–126.

[6] Goldstein, L. 2019. The role of temporal modulation in sensorimotor interaction. *Frontiers in Psychology*, 10, 1–12.

[7] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. 2003. Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31, 465–485.

[8] Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, 11(12), 1–14.

[9] Keitel, A., Gross, J., & Kayser, C. 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, 16(3), 1–19.

[10] Krishnamoorthy, K., Lee, M. 2014. Improved tests for the equality of normal coefficients of variation. *Computational Statistics*, 29(1-2), 215–232. https://link.springer.com/article/10.1007/s00180-013-0445-2

[11] Leong, V., Stone, M. A., Turner, R. E., & Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America*, 136(1), 366–381.

[12] Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., & Chauvel, P. 2004. Temporal envelope processing in the human left and right auditory cortices. *Cerebral Cortex*, 14(7), 731–740.

[13] Marwick, B., Krishnamoorthy, K. 2019. cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.1.3. Retrieved: https://github.com/benmarwick/cvequality, on 02/13/2022

[14] Oganian, Y., Chang, E. F. 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, 5(11), 1–13.

[15] Poeppel, D., Assaneo, M. F. 2020. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6), 322–334.

[16] Rosen, S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences*, 336(1278), 367–373.

[17] Tilsen, S., Arvaniti, A. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1), 628–639.

[18] Weide, R.L. 1998. The Carnegie Mellon pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[19] Westbury, J. R., Turner, G., Dembowski, J. 1994. *X-ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin.

[20] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5).

Amplitude envelopes as a means to study length contrasts: evidence from L1 and L2 Italian

Marieke Einfeldt, Bettina Braun

Department of Linguistics, University of Konstanz, Germany
marieke.einfeldt@uni-konstanz.de, bettina.braun@uni-konstanz.de

ABSTRACT

Previous research has shown that amplitude envelopes can distinguish between rhythmically different languages and between words differing in phonemic vowel length. We use this method to study the realization of consonantal length contrasts for native Italian and advanced German learners of Italian. Consonant length is phonemic in Italian. Phonetically, geminates are longer than singletons and the vowels preceding geminates are shorter than those preceding singletons. German does not have a phonemic consonantal length contrast. We extracted amplitude envelopes from Italian word pairs differing in consonant length for natives and learners. The results of generalized additive modelling showed higher power between 2.3 and 7.4Hz for geminates vs. singletons for both groups. However, native speakers realized the contrast with larger power differences between 2.3 and 3.1Hz while the learners showed larger differences between 7.9 and 8.4Hz. A comparison to German vowel length data suggests that these differences are due to cross-linguistic influence.

Keywords: Amplitude envelopes, consonantal length, cross-linguistic influence, Italian, GAMMs.

1. INTRODUCTION

The analysis of amplitude envelopes has become a widely used method in the speech sciences, language acquisition and neurolinguistics [1-6]. Amplitude envelopes track the amplitude distribution over an utterance. They hence represent the part of the signal that is relevant to convey speech rhythm [7]. Furthermore, the method is easy to apply without demanding manual annotation. Despite the increasingly wide-spread usage across disciplines, there is little research on *which aspects* of the speech signal influence the amplitude envelopes *in what way*. Cross-linguistic research has shown that a stress-timed language (German) led to lower power around 2Hz and between 7 and 10Hz than a more syllable-timed language (Brazilian Portuguese, cf. [1]). Here we compare amplitude envelope differences across two consonantal length conditions in Italian disyllabic words produced by Italian native speakers (Experiment 1) and advanced German learners of Italian (Experiment 2).

For the analysis of amplitude envelopes, the extraction of the modulation frequencies from the speech signal is necessary. This can be done with several procedures [8]: One approach is to filter the sound into a number of frequency bands; the filtering is spaced either in equal steps on the cochlea or logarithm-based. The frequency range usually chosen is between 100 and 8,000 or 10,000Hz respectively. In a next step, the high-frequency components are removed (low-pass filtering up to ~10Hz or Hilbert transform), so that the low-frequency amplitude envelopes remain (so-called narrowband envelopes). These narrowband envelopes are summed up and the modulation frequencies are derived by Fourier analysis. With this method, the final outcome received are spectra, which represent power distribution across frequency.

There are several ways on how to relate amplitude envelope modulation frequencies to linguistic units, such as the syllable rate [e.g., 8 for a review, 9] or to word or phrasal prosody. [4]'s study showed that there are the following clusters of energy: (1) word stress/stressed syllable rate (~2Hz), (2) and the syllable rate (~5Hz). Also [10] found timescales that could be related to certain clusters of energy: (1) the phrasal (0.6-1.3Hz), (2) the word (1.8-3Hz), (3) the syllable (2.8-4.8Hz), and the phoneme (>8Hz) scale.

In this paper we build on data from German vowel length contrast [11]. Their results showed that target words with a short vowel (e.g., ['mɪtʰə]) had a higher power in a small frequency band just below 2Hz and between 5 and 8.5Hz than words with a long vowel (e.g., ['miːtʰə]). Here, we test the Italian consonantal length contrast in L1 speakers of Italian and in proficient German learners of Italian (L2).

2. EXPERIMENT 1

Italian differentiates between short consonants, so-called singletons, and long consonants, so-called geminates. This length contrast can result in minimal pairs such as *fato* /fato/ 'fate' and *fatto* /fat:o/ 'fact', which phonologically only differ in the length of the medial consonant [12]. Vowels preceding singletons have phonetically longer duration than vowels preceding geminates, but this vowel duration contrast

is not phonemic. [6] have further shown that the initial consonants are lengthened in words with an upcoming geminate. Given the vowel length contrast data in [11] and the allophonic decrease in vowel duration before geminates (creating a similarity to the short vowel condition in the German data), we predict higher energy in geminates than singletons in the range between 5 and 8Hz for native speakers of Italian, similar to the German vowel length contrast. Since duration differences in the Italian consonantal length contrast span a wider range (including the preceding vowel), we further hypothesize that the Italian consonantal length contrast results in power differences in a larger frequency range than the German vocalic length contrast.

2.1. Methods

The experiment was a self-paced reading task.

2.1.1 Participants

Six L1 speakers of Italian participated in the experiment (age range: 30-45 years). All participants lived in Germany during the time of testing and had not acquired a foreign language before 6 years of age.

2.1.2 Materials

We selected 16 disyllabic word pairs that differed in the length of the medial consonant. Eleven were minimal pairs and five pairs had the same consonant and preceding vowel but differed in other segments (e.g., /pinːa/ vs. /mina/, /ramo/ vs. /gamːa/). The target consonants were /p/ – /pː/, /t/ – /tː/, /l/ – /lː/, /n/ – /nː/, /m/ – /mː/, /z/ – /sː/. All target words were trochaic. The frequency of the words was on average similar for long and short consonants (SUBTLEX [13] mean-log-frequency score for "geminate words" was 6.6 (SD = 2.8) vs. 6.6 (SD = 3.5) for "singleton words"). We further selected 10 tri- and disyllabic words as distractors. The target words were embedded in a carrier sentence, such that the target word was accented: *Era "palla" che ho detto*, (English: It was "ball" that I said). The written sentence was accompanied by a picture of the item.

2.1.3 Procedure

Recording. The participants sat in front of a computer screen with a PowerPoint-presentation. The test sentences appeared one-by-one. Participants were asked to read the test sentences out loud at normal speed. The recordings were done in a quiet room at the University of Konstanz or at the participant's home. The recording device (Olympus Linear PCM Recorder LS-11/LS-5, 44.1 kHz/16 bit) was placed on the table next to the participant.

Extraction of the amplitude envelope modulation spectra. The start and end of target words were segmented using standard segmentation criteria [14] and the words were saved as separate wav-files, using *Praat* [15]. We analyzed the wideband amplitude envelopes of the productions, following the procedure in [9]. First, we extracted the narrowband amplitude envelopes. For this analysis we used a script developed by He and Dellwo [16, 17]. The speech signal was first down-sampled to 22050Hz and then filtered into nine frequency bands in the range from 100–10,000Hz, which are equidistant on the cochlear map. The cutoff frequencies were 100.5Hz, 250.7Hz, 458.6Hz, 748.8Hz, 1,159.0Hz, 1,449.0Hz, 2,619.8Hz, 3,954.2Hz, 6,121.8Hz and 10,000.8Hz. To remove high-frequency components, the amplitude envelopes were low-pass filtered (< 10Hz). These narrowband amplitude envelopes were then added to compute the wideband amplitude envelope, which were spectrally analyzed in 100 0.1-Hz steps, resulting in 19200 data points (6 speakers x 2 length conditions x 16 items x 100 frequency bands).

Statistical modeling. To model the effect of length across frequency bands, we used generalized additive mixed models, GAMMs [6, 18-22]. They are well-suited to pinpoint in which frequency bands differences occur, taking into account non-linear relationships and auto-correlation [23, 24]. The response variable was log-normalized power. We modelled non-linear dependencies of *length* over frequency bands using smooth functions s(*fband_Hz*, by = *length,* bs='tp', k = 20*)*). These smooth functions include a pre-specified number of base functions of different shapes, e.g., linear and parabolic functions of different complexity [e.g., 24]. *Length* was further added as fixed (parametric) effect. Smooths for *speakers* and *items* (random intercept and over frequency bands) were also included (e.g., s(speaker, fband_Hz, by='re')). For model fitting, we employed the R package *mgcv* [20, 25]; the package *itsadug* was used to plot the model results [26]. The model was corrected for auto-correlation in the data using a correlation parameter, determined by the acf_resid() function (package *itsadug* [26]). We use the function gam.check() to check whether the number of smooth functions (k) and the smoother (thin plate regression, 'tp') were adequate and adjusted if necessary. The final model included the scat-linking function.

2.2. Results

Figure 1 shows the averaged estimated effects of the factor *length*. It suggests that the long consonants have higher power between 2 and 8Hz. To determine whether these effects are significant, we analysed the differences across conditions.

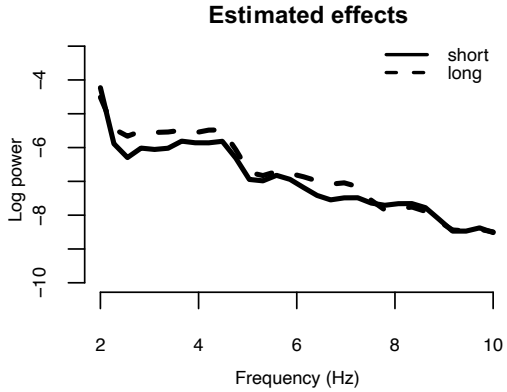


Figure 1: Estimated effects of the factor *length*.

Figure 2 shows the differences between geminates and singletons for the Italian native speakers (L1). Positive values indicate higher power for geminates than singletons; the difference is significant when the grey band (95% confidence interval) excludes 0. Given that the words were very short (on average 401 ms, SD = 80 ms), we only included the interval between 2 and 10Hz in the modelling. Figure 2 indicates that words with geminates had a considerably higher power between 2.2 and 7.4Hz with a short non-significant stretch between 5.3 to 5.8Hz and a slightly lower power in the small range between 7.8 and 8.1Hz.

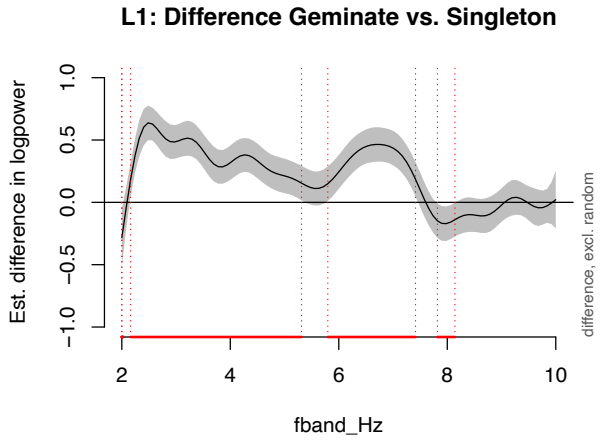


Figure 2: Effect of length for native speakers.

The power differences span a longer frequency interval for the Italian consonantal length contrast than for the German vowel length contrast (between 5 and 8.5Hz, cf. [11]). For direct statistical comparison, we combined the datasets and included a smooth term for the interaction between *language* and *length* (for details on modelling see section 3.2). This interaction model showed significant cross-linguistic power differences between 2.3 and 3.6Hz: In this frequency band the Italian power difference was significantly larger than the German one.

2.3. Discussion

The Italian consonantal length contrast is evident in a large frequency band in the amplitude envelope modulation spectra. Compared to the German vowel length contrast, the Italian consonantal length contrast resulted in significantly larger power differences between 2.3 and 3.6Hz. The most likely reason for these cross-linguistic differences is that the Italian length contrast is distributed over a longer temporal interval, including the preceding vowel and the consonant. Another, albeit less likely, option is that the contrast is larger in Italian due to the use of quasi-minimal pairs. We consider this explanation less likely because the number of quasi minimal pairs was small (fewer than 1/3 of the word pairs).

3. EXPERIMENT 2

Learners of Italian have been shown to produce the consonantal length contrast less clearly/robustly than native speakers [27, 28]. Further, the L2 acquisition of other length related aspects, such as general rhythm, is difficult and prone to L1 transfer [29]. Experiment 2 tested whether amplitude modulation spectra can bring out potentially smaller differences between groups, e.g., deviation from L1 due to transfer in the L2 productions. We predict that the difference between geminates and singletons is less pronounced in German L2 learners of Italian, i.e., differences in power are expected to be smaller and possibly limited to a narrower frequency band. If L2 speakers fail to adequately realize the Italian length contrast (consonantal and vowel duration difference), we particularly predict differences between L1 speakers and learners in the frequency band from 2.3 to 3.6Hz (in which differences between the Italian consonantal length contrast and the German vowel length contrast occurred).

3.1 Methods

Ten German L2 speakers of Italian (20–35 years) were recorded. All L2 speakers studied Italian at university. Their self-rated proficiency on a 6-point-scale (1 = beginner to 6 = native-like) was on average 4. All participants lived in Germany during the time of testing. Materials and procedure were identical to Experiment 1. The L2 speakers additionally completed a background questionnaire including information on proficiency and language use.

3.2 Results

Figure 3 shows the power difference between geminates and singletons for the learners. Geminates resulted in a significantly higher power in the frequency range between 2.3 and 8.7Hz with more

non-significant interruptions (grey confidence interval including 0) than the Italian natives (cf. Figure 2).

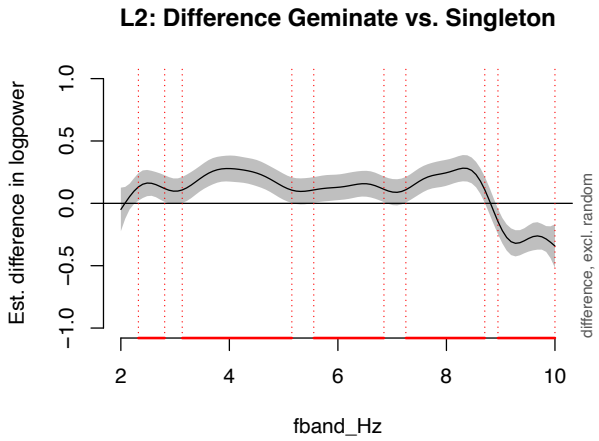


Figure 3: Effect of length contrast for L2 learners.

Furthermore, geminates had significantly lower power in the range from 8.9 and 10Hz, resulting in a biphasic pattern that was absent in this frequency area for the L1 speakers.

To test whether the power differences across language groups are significant, we combined the data from L1 (Experiment 1) and L2 speakers (Experiment 2) and tested for an interaction between *language group* and *length*. To this end we included a factor smooth for the interaction of *language group and length* over frequency bands, s(*fband*, by = *language_length,* bs ='tp', k = 20) and the interaction as parametric (fixed) effect. Otherwise, the model was fitted as described above. The model including the interaction smooth was subsequently compared to a simpler model with separate smooth terms for *language group* and *length,* but without the interaction smooth, using the function CompareML() [26]. The comparison confirmed that the inclusion of the smooth term significantly improved the fit of the model in terms of Maximum Likelihood ($p < 0.003$). To investigate the interaction in detail, we fitted binary smooths (see [24, 30]). The results confirmed the visual comparison of the difference curves between Figures 2 and 3: L1 speakers have significantly higher power in low frequency bands, in particular between 2.3 and 3.1Hz. Furthermore, the L2 speakers produced a larger power difference for the contrast than the L1 speakers at 7.9-8.4Hz.

3.3 Discussion

The results indicate that the L2 speakers produce the consonantal length contrast similar to L1 speakers, with higher power for words containing geminates across most frequency bands from 2.3 to 7.4Hz (intersection between L1 and L2 frequency bands that showed significant differences or differences that approached significance). However, the L1 speakers have higher power difference between 2.3 and 3.1Hz than the L2 speakers, while the L2 speakers show larger power differences between 7.9 and 8.4Hz compared to the L1 speakers. The differences in the higher frequency band are similar to the power differences of the German vowel length contrast [11] (where words with a short vowel had higher power between 5 and 8.5Hz) and could indicate cross-linguistic influence.

4. GENERAL DISCUSSION

Given that the extraction of amplitude envelopes does not necessitate manual segmentation and that amplitude envelopes capture more than just the duration of segments, they provide an efficient means to operationalize the length contrast. This is particularly appealing for phonological contrasts that are distributed over a number of segments, such as Italian gemination (with cues in consonant duration and duration in preceding segments). Our data further show that amplitude envelopes can capture similarities and dissimilarities between L1 and L2 speakers: In terms of similarities, learners produced the length contrast in a similar fashion as the L1 speakers. There was a larger power for geminates than for singletons in the frequency range from 2.3 to 7.4Hz. In terms of dissimilarities, the power differences were smaller for L2 than for L1 speakers in the frequency band from 2.3 to 3.1Hz. On the contrary, learners produced a larger contrast than L1 speakers from 7.9 to 8.4Hz. Interestingly, the smaller effect in the lower frequency band and the larger effect in the higher frequency band mirrors what German speakers do when producing the German vowel length contrast (cf. [11]). Therefore, this difference might be explained in terms of transfer from the L1 (German) to the L2 (Italian).

5. CONCLUSION

Amplitude envelopes seem a reliable and efficient method to compare the production of length contrasts in L1 and L2. They track fine-grained differences between advanced learners and native speakers and are susceptible to transfer effects.

6. ACKNOWLEDGEMENTS

We thank Friederike Hohl for preparing the data, and Tanja Kupisch and Mechtild Tronnier for help with constructing the stimuli. We thank Cesko Voeten for discussion on the function gam.check(), linking functions and smoothing options.

7. REFERENCES

[1] Frota, S., Vigário, M., Cruz, M., Hohl, F., Braun, B. 2022. Amplitude envelope modulations across languages reflect prosody. *Proc. Speech Prosody*, Lisboa, 688-692.

[2] Poeppel, D. 2014. The neuroanatomic and neurophysiological infrastructure for speech and language. *Current Opinion in Neurobiology* 8, 142-149.

[3] Goswami, U. 2019. Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences* 1453, 1, 67-78.

[4] Leong, V., Goswami, U. 2015. Acoustic-Emergent Phonology in the Amplitude Envelope of Child-Directed Speech. *PLoS ONE* 10, 12, e0144411.

[5] Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., Poeppel, D. 2019. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience* 22, 4, 627-632.

[6] Gross, J. Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S. 2013. Speech Rhythms and Multiplexed Oscillatory SensoryCoding in the Human Brain. *PLoS Biologogy* 11, 12, e1001752.

[7] Arvaniti, A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66, 46-63.

[8] Poeppel, D., Assaneo, M. F. 2020. Speech rhythms and their neural foundations. *Nature Review Neuroscience* 21, 322–334.

[9] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A. A. 2009. The natural statistics of audiovisual speech. *PLOS Computational Biology* 5, 7, e1000436.

[10] Keitel, A., Gross, J., Kayser, C. 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology* 16, 3, e2004473.

[11] Hohl, F., Behrens-Zemek, H., Braun, B. 2022. Amplitude envelopes as a means to quantify vowel length contrasts. Presented at: *Phonetik und Phonologie im deutschsprachigen Raum (P&P)*, Bielefeld, Germany.

[12] Payne, E. 2005. Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association* 35, 2, 153-181.

[13] Crepaldi, D., Mandera, P., Keuleers, E., Brysbaert, M. *Subtlex-it: A new frequency list for Italian based on movie subtitles*. [Online]. Available: http://crr.ugent.be/subtlex-it/.

[14] Turk, A. E., Nakai, S., Sugahara, M. 2006. Acoustic segment durations in prosodic research: A practical guide. In: S. Sudhoff, D. Lenertová,, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, J. Schließer (eds), *Methods in Empirical Prosody Research*. Walter de Gruyter, 2006.

[15] Boersma, P., Weenink, D. 2016. *Praat: doing phonetics by computer*.[Online]. Available: https://www.fon.hum.uva.nl/praat/

[16] He, L., Dellwo, V. 2017. Amplitude envelope kinematics of speech signal: parameter extraction and applications. In: J. Trouvain, I. Steiner, and B. Möbius (eds), *Proc. Elektronische Sprachsignalverarbeitung 2017*, Dresden.

[17] He, L., Dellwo, V. 2016. A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform. *Proc. Interspeech 2016*, San Francisco, 530-534.

[18] Wood, S. N., Saefken, B. 2016. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111, 1548-1575.

[19] Wood, S. N. 2017. *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*.

[20] Wood, S. N. 2006. *Generalized additive models: an introduction with R*. Chapman & Hall/CRC Press.

[21] Wieling, M., Margaretha, E., Nerbonne, J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40, 2, 307-314.

[22] Zahner, K., Kutscheid, S., Braun, B. 2019. Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics* 74, 75-95.

[23] Baayen, R. H., van Rij, J., de Cat, C., Wood, S. N., Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In: D. Speelman, K. Heylen, and D. Geeraerts (eds), *Mixed effects regression models in linguistics*, Springer, 2018, pp. 49-69.

[24] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70, 86-116.

[25] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, 1, 3-36.

[26] van Rij, J., Wieling, M., Baayen, R. H., van Rijn, H. 2017. *itsadug: Interpreting time series and autocorrelated data using GAMMs*.

[27] Harada, T. 2006. The acquisition of single and geminate stops by English-speaking children in a Japanese immersion program. *Studies in Second Language Acquisition* 28, 4, 601-632.

[28] Kabak, B., Reckziegel, T., Braun, B. 2011. Timing of second language singletons and geminates. *Proc. International Congress of Phonetic Sciences*, Hong Kong, 994-997.

[29] Li, A., Post, B. 2014. L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German Learners of English. *Studies in Second Language Acquisition* 36, 2, 257-281.

[30] Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., Braun, B. 2022. Three Kinds of Rising-Falling Contours in German wh-Questions: Evidence From Form and Function. *Frontiers in Communication* 7, 838955.

A PERCEPTUAL AND ACOUSTIC STUDY OF MELODY IN WHISPERED CZECH WORDS

Adléta Hanžlová, Tomáš Bořil

Institute of Phonetics, Charles University in Prague
adleta.hanzlova@gmail.com, tomas.boril@ff.cuni.cz

ABSTRACT

The perception of melody in speech depends mainly on the fundamental frequency (f_0) which reflects vocal fold oscillation speed. Whisper is defined by the absence of phonation and therefore the lack of f_0. Intended melody in whisper, however, seems to be discernible regardless.

This paper presents a perception experiment assessing the discernibility of melody in whispered Czech words and words sung in whisper, which proved that melody in whisper in certain cases can in fact be discerned, along with an acoustical analysis of the effect of intended melody in whisper on formant frequencies, formant to formant ratios, center of gravity (CoG) and spectral slope. The parameters affected by intended melody in whispered speech were F2 and CoG of stop-band filtered signal with main formant bandwidths removed. In words sung in whisper, the affected parameters were F2, F3, F2:F1 and F3:F2 ratios, CoG and spectral slope.

Keywords: whisper, melody, absence of phonation, acoustic correlates of intonation, pitch perception

1. INTRODUCTION

Melody is an integral suprasegmental aspect of speech and is mainly dependent on the fundamental frequency (f_0) which reflects vocal fold oscillation. Whisper is defined by the absence of phonation and therefore the lack of f_0.

Notwithstanding claims denying the discernibility of intended melody in whisper (see polemics in [1, 2, 3]), melody has been repeatedly shown to be an aspect perceived in whisper despite the lack of f_0 [1, 3, 4, 5, 6]. Experiments using whispered material from different languages have linked perceived melody in whisper to formant frequencies [7, 8, 9, 10]. Perception experiments [5] and analysis by synthesis [10] suggest that rather than formants alone, the acoustical cues to melody in whisper may be formant-to-formant ratios. Other acoustical properties shown to be linked to melody in whisper are spectral slope [5] and center of gravity (CoG) [11] which changes with intended melody more noticeably in whisper than in modal speech.

The intelligibility of both segmental and suprasegmental aspects of speech in whisper does not have to be based in the same mechanisms as in modal phonation, as can be seen in the worsened abilities of automatic speech recognition models trained on modal phonation when given recordings of whispered speech [12]. Melody in whisper may be discernible on the basis of secondary correlates similar to modal phonation, or by means of compensatory correlates not as prominent in modal phonation [6].

In the last decade, melody in whisper has been studied on French [5] and Dutch [6] words and isolated syllables read by native speakers of Dutch [11]. Recent research focuses on synthesis and automatic recognition of whisper [13, 14], the dominant language of this field being English. Studies of whisper in Czech have so far focused mainly on voicing [15, 16, 17].

This study focuses on the perception and acoustics of melody in whisper in a Czech language environment. A perception experiment was conducted to assess the discernibility of melody in whispered words and words sung in whisper without engaging the vocal cords. An acoustical analysis followed, inspecting the interaction of CoG, spectral slope and formants with intended melody.

2. METHOD

2.1. Material

3 Czech, mainly onomatopoeic, words with CVːCVː and 1 with CVːCV structure repeating the same syllable ([baːba, jɛːjɛː, laːlaː, joːjoː]) were chosen as target words for recording speech material.

For speech, every target word was set in the sentence “Řekl [target] anebo [modified target],” where modified target stands for the target word with two short vowels. Each sentence was realized with 4 different melodic contours as either a statement or question (did he say [target] or [modified target] /

he said [target] or [modified target]). Speakers were asked to repeat the sentences in a shadowing task. Template recordings were in modal phonation with melodic contours manipulated in Praat [18] for the f_0 of vowels in target words to match frequencies of tones to form musical intervals as shown in figure 1.

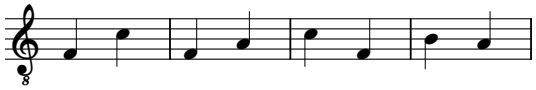


Figure 1: Musical intervals melodic contours of template target words were manipulated to match.

For words sung in whisper, only two target words ([la:la:, jo:jo:]) were used, but each was realized with 8 different melodic contours matched to musical intervals depicted in Figure 2. During recording, the speakers heard through headphones a piano track made in GarageBand [19] playing the target intervals with a metronome. Each interval was played twice, the speakers first listened and then sung along with the track.

Recordings were made in a sound-proof booth using a condenser microphone. Both types of material were recorded first in modal phonation and then in whisper. The templates stayed the same for both types of phonation. Speech material was recorded first, after a short break, the recording resumed with singing. The order of recorded items was randomized for each speaker.

4 female speakers aged 20–24 participated in the recording. All reported having basic musical education and experience with solo or choral singing. Each speaker was recorded in one session which including a break lasted around 25 minutes.

Target words were then extracted from all recordings, labeled using the *align interval* option in Praat and phone boundaries manually corrected according to [20]. Only recordings of whispered target words were used in the following perception experiment and acoustical analysis.

2.2. Perception experiment

A 2AFC perception experiment was conducted online using PsyToolkit software [21, 22] to assess the discernibility of melody in whisper.

Some recorded target words showed signs of vocal fold engagement and were left out from the perception experiment. This elimination resulted in 54 whispered speech and 53 whisper-sung stimuli. Each of these two sets of stimuli was put in a separate experiment. However, both setups were the same and they were embedded in one experiment

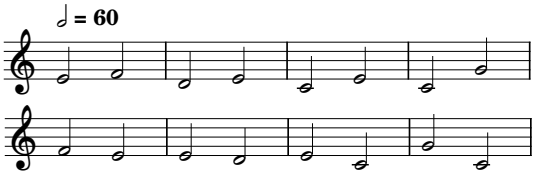


Figure 2: Musical intervals played as template when recording sung and whisper-sung words.

session, starting with words sung in whisper and continuing with whispered speech. To each set of stimuli, filler stimuli were added to motivate the listeners. For whispered speech, 12 stimuli, preliminarily judged as having recognizable melody, with the lower of the two target musical tones low-pass filtered at 6300 Hz (where visible formant structure ended in the spectrogram) were used to create filler stimuli. For whisper-sung stimuli, 9 additional stimuli recorded during a preliminary session with prominent spectral changes were added.

In both experiments, listeners were asked to determine whether they hear a fall or rise in melody by pressing a corresponding button on the screen. Each stimulus was preceded by a beep and could be replayed once. The order of stimuli in each experiment was randomized for each participant. Each experiment begun with a training session (comprising of 6 of the stimuli) and included a short pause in the middle.

1 Slovak and 32 Czech voluntary respondents (13 male) aged 17–63 (median 24) participated in the experiments. They were advised to wear headphones and complete the experiments in a quiet room. The whole experiment session lasted 13–29 minutes (median 19) based on each respondent's pace.

Results of these experiments were analyzed in R [23]. Success rates were calculated as the mean of individual respondents' correct answer rates for each experiment in general as well as for subcategories within them (intended melody and pitch difference). Using binomial tests included in Hmisc package [24], confidence intervals at $\alpha = 0.05$ with Bonferroni correction matched to the number of subgroups were added to each success rate.

2.3. Acoustical analysis

Using a Praat script, the middle third of each vowel was marked based on the labeled whisper recordings. Some recordings showed slight signs of vocal chord engagement. If the spectrogram of a vowel had visible periods in modal speech frequency ranges, a narrow frequency range around a prominent spectral peak was stop-band filtered. The same procedure was applied for low-frequency noises (under 100 Hz)

caused by heavy vehicles passing the studio during recording. Within the middle third of each vowel, mean formant frequencies [Hz], center of gravity (CoG, [Hz]) and spectral slope [dB] were obtained using a Praat script.

Formants were analyzed using the *burg* method between 0 and 5.5 kHz with maximum number of peaks set to 5. Additionally, ratios of mean formant frequencies (F2:F1, F3:F2, F3:F1) were calculated in R. CoG was measured from each vowel's spectrum with default Praat settings. For recordings of whispered speech, a second CoG value was measured after using a stop-band filter between 1 and 6.3 kHz, removing main formant structure visible in the spectrogram. Spectral slope was obtained from LTAS spectra of the vowels using Praat's *get slope* function (method *energy*, bandwidths 50 Hz–2 kHz and 2–8 kHz).

Phone duration as a parameter was omitted in this analysis, as it was likely affected by recording speech via a shadowing task as well as singing with a metronome. Even though duration may be one of the distinguishing factors of pitch in tone languages [25], it was not proven to be a major acoustic cue to melody in whisper recorded in a controlled environment [11].

The effects of each acoustical parameter on melody in whisper were analyzed using linear mixed-effects models in R [23] with lme4 [26] and emmeans [27] packages. The dependent variable for each model was one of the aforementioned acoustical parameters (F1–F3, formant ratios, CoG, spectral slope). Intended pitch movement (rise/fall) of a word in interaction with vowel position within the word (in first/second syllable) was used as a fixed factor (further referred to as intended melody). Random factors encompassed speakers and target words with random intercepts. Formant and CoG values were logarithmed to prevent heteroscedasticity. Statistical significance of the fixed factors' effects was determined using likelihood ratio tests.

3. RESULTS AND DISCUSSION

3.1. Discernibility of melody in whisper

Respondents were able to discern the intended melody in 52–83 % (median 72 %) for whispered speech and 66–94 % (median 83 %) for words sung in whisper.

As shown in Figure 3, intended melody was discernible with statistical significance in words sung in whisper, but the success rate in whispered speech was significant only in stimuli with falling

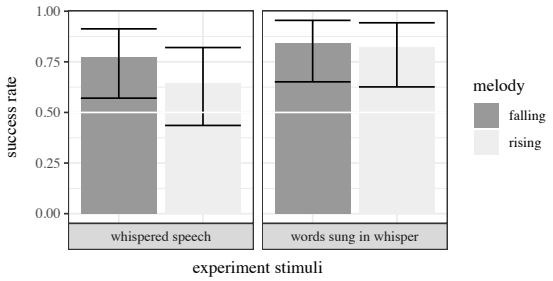


Figure 3: Success rates of perception experiments by intended melody with confidence intervals at $\alpha = 0.05$ with Bonferroni correction for $n = 2$.

melody. Greater ease of discerning falling rather than rising melody may be caused by the fact that whispered melody seems to be more easily discernible when the higher of the two tones matches with the position of word stress [6]. Due to the nature of the shadowing task used while recording and Czech stress being word-initial, the speakers may have been placing prominent word-initial stress on target words, which would make it correspond to higher tones in words with falling melody.

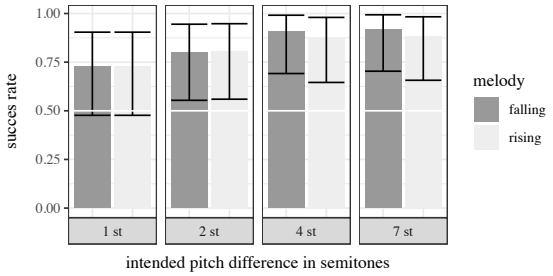


Figure 4: Success rates of discerning melody of whisper-sung words by intended melody and pitch difference with confidence intervals at $\alpha = 0.05$ with Bonferroni correction for $n = 8$.

In addition to falling melody being discerned more easily, greater differences in intended pitch of the syllables within a word lead to higher success rates, as can be seen in Figure 4, which shows success rates of discerning individual melodic contours in words sung in whisper.

3.2. Formants and formant ratios

In whispered speech, only F2 was significantly affected by intended melody ($\alpha = 0.05$, $p = 0.046$). For words sung in whisper, intended melody significantly affected F2 ($p < 0.001$), F3 ($p = 0.008$) as well as F2:F1 ($p < 0.001$) and F3:F2 ($p = 0.012$) ratios. The effect of intended melody on formant ratios in words sung in whisper points to the

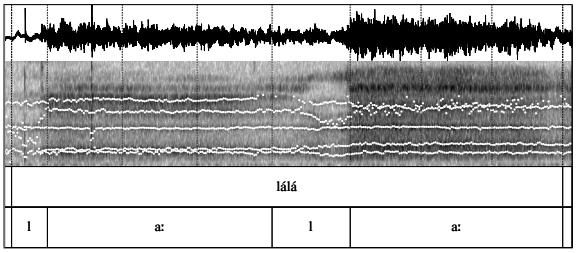


Figure 5: [la:la:] sung in whisper with a rising melody with an intended pitch difference of 7 st. Dotted lines denote formants. The spectrogram frequency range is 0–8 kHz, time step 5 ms.

movement of F2 being more prominent than those of F1 or F3. One realization of such F2 movement can be observed in Figure 5, which shows the spectrogram of a stimulus recognized by 100 % of respondents as having a rising melody. This is in line with the results of I. B. Thomas [8] who determined F2 to be a direct correlate of pitch in whisper. More recent research has also established formants as a correlate of melody in whisper. Perception experiments using synthesized vowels pointed to the movement of F2 corresponding to perceived melody with the effect being stronger when F1 was shifted as well [10]. A study of whispered Dutch then reported a statistically significant effect of melody on the first three formants [11]. F1 was, however, not proven to be significantly affected by intended melody in the present study.

The movement of F2 and its interaction with F1 and F3 was more significant in words sung in whisper than in whispered speech. A possible explanation lies in the characteristics of regular singing and speech. While intonation in speech is based on relative pitch movements, sung melody is intended to match specific musical notes. Speech may also be more precise in maintaining vowel quality, preventing formants from shifting to a degree where formant relations would interfere with those of a different vowel. It should also be noted that the recorded vowels sung in whisper were longer in duration than vowels in whispered speech, so speakers had more time to reach the intended target pitch.

3.3. Center of gravity

Center of gravity was significantly affected by intended melody only in words sung in whisper ($\alpha = 0.05$, $p = 0.002$). In whispered speech, intended melody significantly affected CoG of the signal after using a stop-band filter removing the main formant bandwidths ($p = 0.022$). This corresponds to [7] stating high-frequency spectral noise as a correlate of melody in some whispered vowels, as well as newer studies which found the main acoustical cues to melody in whisper likely to be in frequency ranges above 1.5 kHz [5].

3.4. Spectral slope

Perception experiments with spectral slope manipulations proved that spectral slope can affect perceived pitch [28]. Spectral slope was also found to be a correlate of intended melody in whisper [11], which is in line with the hypothesis that if higher pitch requires higher effort, it will also show a less negative spectral slope [6].

In the present study, the effect of intended melody in whisper on spectral slope was statistically significant only in words sung in whisper ($\alpha = 0.05$, $p = 0.0069$), but not in whispered speech ($p = 0.24$). A possible explanation may lie in the setup of the recording. Speakers were asked to repeat the heard melody in both tasks, it is however likely that while they attempted to be precise in whisper-singing, they repeated the speech materials in their own voice range which was more comfortable for them and therefore required less effort.

4. CONCLUSION

This paper presented an insight into the production and perception of melody in whisper in a Czech-language environment. Two-syllable words realized with different intended melodic contours were recorded and used as data for this study. A perception experiment showed that intended melody can be more easily discerned with greater difference in intended pitch of the two syllables. Falling melody was also more successfully recognized than rising melody.

Several acoustical parameters were then analyzed as possible cues to intended melody in whisper using linear mixed-effects models. In whispered speech, F2 and spectral slope of stop-band filtered signal removing visible formant structure were significantly affected by intended melody. In words sung in whisper, acoustical parameters significantly affected by intended melody were F2, F3, F2:F1 and F3:F2 ratios, which point to greater movements of F2 in comparison to surrounding formants, as well as center of gravity and spectral slope.

5. ACKNOWLEDGMENTS

This work was supported by the Cooperatio research program of Charles University.

6. REFERENCES

[1] F. Giet, *Zur Tonität nordchinesischer Mundarten*, ser. Anthropos-Institut Sankt Augustin: Studia Instituti Anthropos. Verlag der Missionsdruckerei St. Gabriel, 1950.

[2] G. Panconcelli-Calzia, "Das Flüstern in seiner physio-pathologischen und linguistischen Bedeutung," *Lingua*, vol. 4, pp. 369–378, Jan. 1954.

[3] F. Giet, "Kann man in einer Tonsprache flüstern?" *Lingua*, vol. 5, pp. 372–381, Jan. 1955.

[4] M. Kloster Jensen, "Recognition of word tones in whispered speech," *WORD*, vol. 14, no. 2-3, pp. 187–196, 1958.

[5] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2026–2040, Apr. 2014.

[6] W. F. L. Heeren and V. J. van Heuven, "The interaction of lexical and phrasal prosody in whispered speech," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3272–3289, Dec. 2014.

[7] W. Meyer-Eppler, "Realization of Prosodic Features in Whispered Speech," *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 104–106, jan 1957.

[8] I. B. Thomas, "Perceived Pitch of Whispered Vowels," *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 468–470, Aug. 1969.

[9] J. Fónagy, "Accent et intonation dans la parole chuchotée," *Phonetica*, vol. 20, no. 2-4, pp. 177–192, 1969.

[10] M. Higashikawa and F. D. Minifie, "Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 3, pp. 583–591, 1999.

[11] W. F. L. Heeren, "Vocalic correlates of pitch in whispered versus normal speech," *The Journal of the Acoustical Society of America*, vol. 138, no. 6, pp. 3800–3810, Dec. 2015.

[12] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, Feb. 2005.

[13] Z. Raeesy, K. Gillespie, C. Ma, T. Drugman, J. Gu, R. Maas, A. Rastrow, and B. Hoffmeister, "LSTM-based whisper detection," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 139–144.

[14] K. Phapatanaburi, W. Pathonsuwan, L. Wang, P. Anchuen, T. Jumphoo, P. Buayai, M. Uthansakul, and P. Uthansakul, "Whispered speech detection using glottal flow-based features," *Symmetry*, vol. 14, no. 4, p. 777, 2022.

[15] P. Machač and P. Šturm, "The phonological contrast of voicing in whispered Czech and its phonetic correlates – A preliminary study," *20th Czech-German Workshop - Speech Processing*, pp. 34–43, 2010.

[16] R. Skarnitzl, P. Šturm, and P. Machač, "The phonological voicing contrast in Czech: an EPG study of phonated and whispered fricatives." in *14th Annual Conference of the International Speech Communication Association*, 2013, pp. 3191–3195.

[17] M. Svatošová and T. Bořil, "Duration as a cue for phonological voicing contrast in whispered Czech," in *Proc. 20th ICPhS*, 2023. [in print].

[18] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]. Version 6.1.56*, 2021. [Online]. Available: http://www.praat.org/

[19] Apple Inc, *GarageBand for Mac. Version 10.4.6*, 2022. [Online]. Available: https://www.apple.com/mac/garageband/

[20] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Epocha, 2010.

[21] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[22] ——, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] F. E. Harrell Jr, *Hmisc: Harrell Miscellaneous. Version 4.7-0*, 2022. [Online]. Available: https://CRAN.R-project.org/package=Hmisc

[25] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," *Language and Speech*, vol. 47, no. 2, pp. 109–138, 2004.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[27] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[28] J. Kuang and M. Liberman, "The effect of spectral slope on pitch perception," in *16th Annual Conference of the International Speech Communication Association*, 2015.

PROMINENCE EFFECTS ON DUTCH LOW VOWELS

Anna Bruggeman

Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University
anna.m.bruggeman@protonmail.com

ABSTRACT

This paper investigates the acoustic realisation of the Dutch low vowels /a(ː)/ and /ɑ/ based on a large corpus of speech recordings (the CGN corpus). Variation in the realisation of these vowels has been observed, specifically of cases where lax /ɑ/ appears to substitute for tense /a/ in non-prominent positions, but no large-scale acoustic investigation comparing the formant distributions of these vowels has yet been carried out. The present study looks at F1 and F2 of ca. 1 million low vowels to investigate the role of prominence (4 different levels of stress) in conditioning variation in vowel quality. Results show that secondary stressed /a/ is indeed realised much like /ɑ/, i.e. further back and higher in the vowel space. This holds generally for both Netherlandic and Belgian Dutch. Unstressed /a/ on the other hand approaches /ə/ in the Netherlands, but not in Belgium.

Keywords: vowel quality, prominence, variation, Dutch

1. INTRODUCTION

1.1. Variation and alternation in Dutch vowels

The Dutch vowel system has 12 monophthongs (both Netherlandic and Flemish), with schwa counting as an additional phonological vowel for Netherlandic Dutch. Monophthongs are typically described as having a phonological length distinction, phonetically realised reliably as a tense/lax quality opposition, yielding the following 5 pairs: /i/-/ɪ/, /y/-/ʏ/, /e(ː)/-/ɛ/, /o(ː)/-/ɔ/, /a(ː)/-/ɑ/, and unpaired /u/, /ø/, and /ə/ [1, 2, 3].

This paper is concerned with known variation in the vowel quality (F1, F2) of members of pairs across the tense/lax divide, with a focus on the low vowel pair (/a/-/ɑ/). There are at least two different views in the literature on the nature of this variation.

The first sees the phenomenon as phonological substitution: long/tense vowels may be realised as their short/lax counterparts, e.g. the lexical vowel /a/ in words like /ˌka.bi.ˈnɛt/ may be realised as [ɑ] ('vowel shortening' in [1], 'lexical vowel reduction' in [4]). According to [1], this particularly targets word-initial vowels, but can apply to all non-primary stressed syllables alike. The second view considers variation to reflect a general reduction process that targets only unstressed vowels ('vowel reduction' in [1], 'acoustic vowel reduction' in [4]).

Some confusion nevertheless remains about which phenomenon applies to which vowels. For words like /ba.ˈnan/, realisational variants ranging from [a] (expected) to [ɑ] and [ə] have been noted for the first (unstressed) vowel [4], and this particular word is listed as an example both of 'vowel reduction' (to schwa) and 'vowel shortening' (substitution) [1]. As unstressed vowels may be targeted by either of the two processes it is not clear at this point that the phenomena are entirely distinct.

Until now, production studies have not directly addressed tense/lax variation in detail, with the exception of [2, 5] on Belgian varieties of Dutch, which revealed a great degree of overlap between realisations of /a/ and /ɑ/ in Antwerp Dutch. On the perception side, there is evidence that /a/ may be mistaken for /ɑ/, at least when the vowel is presented excised from a word in isolation ([6], and cf. [7]).

1.2. Research aim

In order to provide acoustic substance to the aforementioned observations and claims about variation across the tense/lax opposition, the present work documents the realisational distributions of the low vowel pair /a/-/ɑ/ based on corpus data. The focus is on low vowels because i) these do not diphthongise, in contrast to Dutch mid vowels (in the Netherlands), and ii) observations about reduction and substitution suggest that alternations are readily observable, in contrast to what is the case for high vowels [1].

The main aim of this paper is therefore descriptive: What are the distributions of /a/ and /ɑ/ vowels as a function of prominence, and more specifically stress status?

2. METHODOLOGY

2.1. Speech materials

2.1.1. Corpus extraction & preprocessing

This study uses the *Corpus Gesproken Nederlands* [8], excluding only those parts that consist of (live sports) commentaries (folders *comp-i* and *comp-l*), which were characterised by considerable background noise. F1 and F2 were measured at three timepoints (25%, 50% and 75% into the vowel) for all Dutch vowels, for the first three speakers in each sound file, based on the automatically generated transcriptions provided with the corpus. Vowels followed by /r, l/ were categorically excluded from analysis due to known coarticulatory effects [9]. Lobanov-normalisation [10] was performed on a by-speaker basis to determine the relative position of each vowel within the speaker's vowel space. Vowels with F1/F2 values more than 3 SD away from the mean for that vowel category were excluded from further analysis. This method yielded a total of over 8 million vowel tokens.

2.1.2. Identification of target items with low vowels

The tense/lax contrast in Dutch is correlated with syllable structure. Tense variants tend to be restricted to closed syllables, whereas lax variants may also occur in open syllables [11]. Identical syllabic environments for /a/ and /ɑ/ are thus only found for monosyllabic words (e.g. *kap* /kɑp/ and *kaap* /kap/), where polysyllabic minimal pairs are assumed to have different syllabification (e.g. *kapen* /ka.pən/ but *kappen* /kap.pən/ with an ambisyllabic intervocalic consonant). The present investigation uses two kinds of target words (full list available online, see link for scripts below):

1. a (near-)complete set of Dutch monosyllabic words of the structure (C)(C)(C)V(C)(C)(C) where V is either /ɑ/ or /a/ (N=830);
2. a set of polysyllabic words (2–5 syllables), containing /ɑ/ or /a/ in the initial syllable (N=284).

Target words were identified based on the orthographic occurrence of 'a' and 'aa' in initial syllables of items occurring on the lemma lists of the CGN. Words in which the target vowel was followed by a tautosyllabic /r/ or /l/ were again not included. Phonological vowel category (/ɑ/ or /a/) was manually determined and a number of words with non-target vowels were excluded.

The vowels in the monosyllabic set carry lexical stress by definition. The expectation is that vowels in this set will be maximally distinct if prominence plays a role in determining distributions. In contrast, prominence status is varied (as a variable of interest) for the vowels in the polysyllabic set. Three levels are identified: i) primary stress, ii) secondary stress (assigned based on the principle of alternating rhythm following [1, 11]), and iii) unstressed (stress-adjacent). A distinction is made between secondary stressed and unstressed vowels to investigate the possibility that these may be subject to different types of variation (see Section 1.1), and because a distinction was not made in an earlier study looking at reduction [6]. Example words for all 4 types of prominence status are given in Table 1. Based on these target words, the corpus yielded a total of over 1 million initial low vowels for analysis (ca. 424k /a/ and 684k /ɑ/).

	tense /a/	lax /ɑ/
stressed monosyll.	/ˈχaf/	/ˈχɑf/
stressed polysyll.	/ˈpa.χi.na/	/ˈbɑk.la.va/
secondary stressed	/ˌka.bi.ˈnɛt/	/ˌkɑn.ta.ˈrɛl/
unstressed	/ma.ˈho.ni/	/ɑt.ˈtɛn.si/

Table 1: Example target words with /a/ and /ɑ/ as a function of prominence (stress status).

2.2. Statistics

Analysis was performed in R [12]. Linear-mixed regression models were fitted with *lme4* [13] on Lobanov-normalised F1 and F2 as a function of various predictors. Main effects were assessed through LRTs between full and null models for the predictors tested, and multiple comparisons (Tukey-corrected) are from *emmeans* [14]. R^2 values are generated with the package *MuMIN* [15]. Extracted data and scripts are available here: https://github.com/ambrug/Dutch_low_vowels.

3. RESULTS

3.1. Low vowels within the Dutch vowel space

Figure 1 shows the full vowel space for Netherlandic Dutch and Belgian Dutch, based on all monophthongs. Labelling is kept as it is in the CGN, with I E A O Y reflecting lax/short variants. For the tense mid vowels /e ø o/ (CGN labels: e 2 o), F1 and F2 are based on the measurement at 25% of the vowel duration as these are known to be realised as narrow closing diphthongs in Netherlandic Dutch [3] (see also method used in [16]). The Netherlands data are based on means from >2400 speakers and

>5 million vowel tokens, and the Belgium data on >1200 speakers and >3 million tokens.

These overall results show that /ɑ/ ('A' in Fig. 1) is realised further back and higher than /a/ ('a') in both varieties, in line with formant values reported in [2, 5, 16, 17, 18] (although cf. [19, 20] for some varieties spoken in Belgium). As the labels shown here represent a broad phonetic transcription, however, the apparent distinctiveness of the categories in part reflects the mere existence of acoustically different clusters, rather than differences as a function of phonological vowel status. In the following, the analysis will focus on the tense/lax distinction based on manually determined phonological status instead.

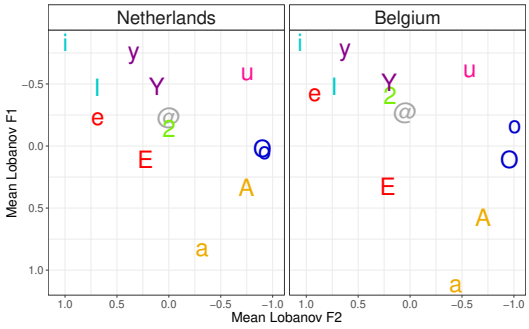


Figure 1: Vowel space showing (phonological) monophthongs for speakers from the Netherlands and Belgium. Labels reflect CGN annotations.

3.2. Realisation of low vowels as a function of stress status

Regression models were run for each formant and region separately, using mean values per speaker. Fixed effects were interacting `vowel` (ə, a, ɑ) and `stress` (4 levels, see Table 1). A random intercept was added for `speaker`. Conditional R^2 for these models was around 0.7 for the Netherlands, and around 0.75 for the Belgian data. Descriptive mean F1 and F2 for low vowels as a function of stress status are shown Figure 2, with schwa (>2 million tokens) plotted for reference as well.

For both Belgian and Netherlandic Dutch there were interactions between `vowel` and `stress` for both F1 and F2, with most pairwise comparisons, within- and between-vowel, significantly different from each other (at $p < 0.05$). Focussing on /a/, which shows to be most variable as a function of stress, the following comparisons are of particular interest for Netherlandic Dutch:

- secondary stressed /a/ is no different in F1 from unstressed /ɑ/ (est= 0.05, SE=0.05);
- secondary stressed /a/ is no different in F2 from /ɑ/ when it i) also carries secondary stress (est=0.10, SE=0.04) or ii) is unstressed (est=0.07, SE=0.03);
- unstressed /a/ is different from /ə/ in terms of both F1 (est=0.12, SE=0.01, p<0.0001) and F2 (est=-0.25, SE=0.01, p<0.0001).

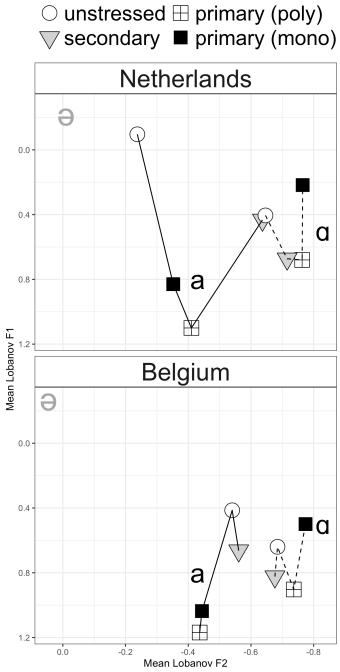


Figure 2: Mean F1 and F2 (Lob-normalised) of low vowels and schwa as a function of stress.

In sum, secondary stressed /a/ is retracted relative to its own stressed and unstressed counterparts, being realised with F2 similar to that of /ɑ/ (although not quite as far back as stressed /ɑ/ in monosyllables). When F1 and F2 are taken into account together, secondary stressed /a/ in words like /ˌka.bi.ˈnɛt/ is indistinguishable from /ɑ/ when it is unstressed in words like /ɑt.ˈtɛn.si/. Unstressed /a/ on the other hand is realised more centrally, although it remains distinct from /ə/.

For Belgian Dutch, the following tense/lax comparisons are most relevant:

- secondary stressed /a/ is no different in F1 from /ɑ/ when it is i) primary stressed and monosyllabic (est=0.13, SE=0.06) ii) secondary stressed (est=-0.21, SE=0.08) and iii) unstressed (est=0.02, SE=0.06);
- secondary stressed /a/ does not differ in F2 from secondary stressed or unstressed /ɑ/ (in both cases est=0.13, SE=0.05), nor from polysyllabic stressed /ɑ/ (est=0.17, SE=0.05);

Belgian Dutch thus exhibits similar retraction patterns as Netherlandic Dutch: secondary stressed /a/ overlaps in values with most cases of /ɑ/.

3.3. Antwerp vowels

To investigate claims about a merger in the Antwerp area, all low vowels spoken by Antwerp speakers (defined here as those who received their education in the Antwerp area, with *eduPlace* codes B-201–206 in the CGN) are considered separately (N=55 speakers and N=216 tokens). Figure 3 shows the mean values for /a/ and /ɑ/ ([ə] plotted for reference).

Regression results predicting F1 and F2 for /a/ and /ɑ/ reveal that F1 differs between `vowel`, but only for several comparisons involving primary stressed vowels. F2, however, is distinct only for the comparison between the most prominent lax vowel and the least prominent tense vowel: primary stressed monosyllabic /ɑ/ and unstressed /a/ (est=-0.32, SE=0.07, p<0.001). The latter, tense vowel can be seen to be realised more centrally, although not as much so as its Netherlandic counterpart. In short, this confirms [2, 5]'s observation that, at least in terms of F2, /ɑ/ and /a/ in Antwerp are to a large extent overlapping.

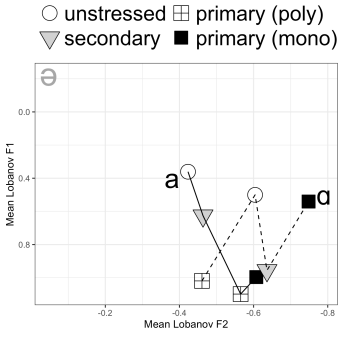


Figure 3: Mean F1 and F2 (Lob-normalised) of low vowels and schwa for speakers from Antwerp.

3.4. Other factors affecting F2 of tense /a/

Beyond prominence, many other factors may of course play a role in determining formant values. In order to assess what else may affect F2 for /a/, an exploratory random forest was run on all Lobanov-normalised F2 values for this vowel, using word-specific speaker means (package *ranger* [21], *mtry*=3, 2000 trees). The variable importance of the predictors is shown in Figure 4 and suggests that there is much speaker-dependent variation, beyond even more categorical speaker characteristics (such as place of education, year of birth). The variable *CGN folder* reflects different speech styles in the CGN, suggesting that there is considerable context-dependent variation. Most interestingly, usage-based factors such as word frequency (*Lg10WF*, based on the SUBTLEX corpus [22]) or other word-specific effects (*word*) do not rank highly here, counter to the general expectations about greater variation for more frequently used items [1, 23, 24]. The low ranking of *stress* (the 4 levels used in this study) might reflect that it makes a small contribution on its own, and rather derives its effect from interaction with other factors. Future work could look into this in more detail.

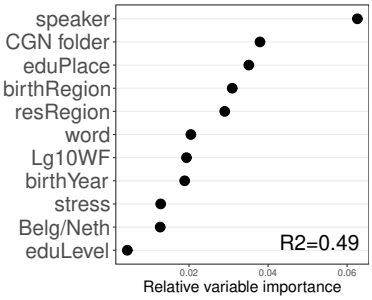


Figure 4: Variable importance for random forest on F2 for tense /a/.

4. SUMMARY AND DISCUSSION

The above results show that vowel quality of Dutch low vowels indeed varies as a function of prominence. Especially tense /a/ was affected: when secondary stressed, it is retracted, occupying a region of the vowel space also used by /ɑ/.

The present data also bear on several other claims in the literature. For Netherlandic Dutch, data are in line with the observation that "[v]owels in open syllables …reduce easier than vowels in closed syllables" [1]: unstressed tense /a/, which always occurs in an open syllable (in contrast to /ɑ/), was realised toward the centre of the vowel space. These results are highly suggestive of the possibility that unstressed and secondary stressed /a/ are subject to variation of a different nature, at least in Netherlandic Dutch, with the former exhibiting patterns in line with reduction/centralisation, and the latter towards a merger with /ɑ/, also observed in Belgium. In Antwerp, /a/ and /ɑ/ generally exhibited similar F2 values, corroborating results by [5].

Finally, results presented here only describe averaged production patterns, and no claims are made with respect to perceptual distinctiveness of low vowels. While there is considerable overlap in formant values for low vowels, leading to non-distinct distributions under some prominence conditions, low vowels might well remain distinct in terms of duration, cf. [3, 16, 25].

5. REFERENCES

[1] G. Booij, *The phonology of Dutch*. Oxford: Oxford University Press, 1999.

[2] J. Verhoeven, "Belgian Standard Dutch," *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 243–247, 2005.

[3] C. Gussenhoven, "Dutch," *Journal of the International Phonetic Association*, vol. 22, no. 1-2, pp. 45–47, 1992.

[4] D. R. van Bergem, "Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress," *Speech communication*, vol. 16, no. 4, pp. 329–358, 1995.

[5] J. Verhoeven and C. Van Bael, "Akoestische kenmerken van de Nederlandse klinkers in drie Vlaamse regio's," *Taal en tongval*, vol. 54, pp. 1–23, 2002.

[6] D. R. Van Bergem, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class," *Speech Communication*, vol. 12, no. 1, pp. 1–23, 1993.

[7] A. C. Rietveld and F. J. Koopmans-van Beinum, "Vowel reduction and stress," *Speech communication*, vol. 6, no. 3, pp. 217–229, 1987.

[8] "Het Corpus Gesproken Nederlands (Spoken Dutch Corpus)," Data set, 2014. [Online]. Available: http://hdl.handle.net/10032/tm-a2-k6

[9] B. Collins and I. M. Mees, *The sounds of English and Dutch*, 5th ed. Leiden: Brill, 2003 [1984].

[10] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, pp. 606–608, 1971.

[11] C. Gussenhoven, "Wat is de beste transcriptie voor het Nederlands?" *Nederlandse Taalkunde*, vol. 12, no. 4, pp. 331–350, 2007.

[12] R Core Team, "R," Software, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[13] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4," R package, 2015.

[14] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Emmeans: Estimated marginal means, aka least-squares means," R package, 2018.

[15] K. Barton, "Mumin: multi-model inference," R package, 2009.

[16] P. Adank, R. Van Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1729–1738, 2004.

[17] L. C. Pols, H. R. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1093–1101, 1973.

[18] F. J. Koopmans-Van Beinum, "The role of focus words in natural and in synthetic continuous speech: Acoustic aspects," *Speech Communication*, vol. 11, no. 4-5, pp. 439–452, 1992.

[19] J. Verhoeven, "The Belgian Limburg dialect of Hamont," *Journal of the International Phonetic Association*, vol. 37, no. 2, pp. 219–225, 2007.

[20] E. Blancquaert, *Praktische uitspraakleer van de Nederlandse taal*, 8th ed. De Sikkel, 1969.

[21] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017.

[22] E. Keuleers, M. Brysbaert, and B. New, "SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles," *Behavior research methods*, vol. 42, no. 3, pp. 643–650, 2010.

[23] J. Bybee, "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change," *Language variation and change*, vol. 14, no. 3, pp. 261–290, 2002.

[24] J. Pierrehumbert, *Exemplar dynamics: Word frequency, lenition and contrast*. Amsterdam: John Benjamins, 2001.

[25] T. Rietveld, J. Kerkhoff, and C. Gussenhoven, "Word prosodic structure and vowel duration in Dutch," *Journal of Phonetics*, vol. 32, no. 3, pp. 349–371, 2004.

ACOUSTIC ABSEMENT IN DETAIL: QUANTIFYING ACOUSTIC DIFFERENCES ACROSS TIME-SERIES REPRESENTATIONS OF SPEECH DATA

Matthew C. Kelley

Department of Linguistics, University of Washington
mattck@uw.edu

ABSTRACT

The speech signal is a consummate example of time-series data. The acoustics of the signal change over time, sometimes dramatically. Yet, the most common type of comparison we perform in phonetics is between instantaneous acoustic measurements, such as formant values. In the present paper, I discuss the concept of absement as a quantification of differences between two time-series. I then provide an experimental example of absement applied to phonetic analysis for human and/or computer speech recognition. The experiment is a template-based speech recognition task, using dynamic time warping to compare the acoustics between recordings of isolated words. A recognition accuracy of 57.9% was achieved. The results of the experiment are discussed in terms of using absement as a tool, as well as the implications of using acoustics-only models of spoken word recognition with the word as the smallest discrete linguistic unit.

Keywords: spoken word recognition, speech technology, acoustics, acoustic absement

1. INTRODUCTION

The field of phonetics is replete with time-series and sequential data. Whether analyzing a waveform, spectrogram, or transcription, there is a natural and prescribed order to the sub-elements of the sequence—e.g., samples, spectra, or segments, respectively. Despite the sequential nature of so many of the types of data that we analyze, we often factor out time-bound aspects and instead focus on instantaneous measurements of a quantity, such as formant values or intensity at a given point in time. This choice is certainly convenient, and it is all but mandated by many of the statistical methods we employ in speech science.

Nevertheless, the field stands to benefit from incorporating more methods of analysis that can account for time-series data. In the present paper, I focus on the quantity of absement as a concrete example of a measurement that reflects the temporal nature of speech data. In the present paper, I present a simplified speech recognition experiment using dynamic time warping. This task itself is a classical example of early speech recognition techniques, and the purpose is to demonstrate a use for absement for phonetic analysis. However, it also bears some relation to models of spoken word recognition, which I will also discuss.

1.1. Theoretical background

Before formally introducing absement, it is necessary to discuss distance. Distance itself is a familiar notion, often quantifying how far apart two objects are in physical space. An analogous example from phonetics would be how far apart two formant values are from each other. There are infinite methods to calculate distance, but in mechanics, distance is defined as the magnitude of a displacement vector, which is the same as the formula for Euclidean distance. Euclidean distance itself is well-attested in phonetics, such as when comparing the centroids of an F1-by-F2 space for two vowel categories [1]. Formally, the Euclidean distance d between two vectors x and y of length k is

$$(1)\quad d(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{k} |\chi_i - \psi_i|^2}\,,$$

where χ_i and ψ_i are the scalar elements within x and y at index i, respectively.

If the distance between two objects is summed or integrated over time, the result is a quantity referred to in [2] and [3] as "absement" [ˈæbsəmn̩t], a blend of *absence* and *displacement*. This is the time-distance product form of absement, though a time-displacement product form also exists [2]. Absement is a measure of prolonged distance. For time-series data, absement can be thought of as how well two objects matched each other over time.

A high absement value could be caused by two objects being apart over a long period of time, a moment of extreme distance, or a combination thereof. Formally, the absement between two time-series X and Y is

$$(2) \quad a(X,Y) = \int_0^T d(x_t, y_t)\,\mathrm{d}t\,,$$

where x_t and y_t are vectors representing the value of X and Y at time t, and T is the total length of time over which absement is being calculated.

Absement naturally accounts for the acoustic events that happen over a span of time, in a way that an instantaneous measure cannot. We do employ methods in phonetics that work well with time-series like the Fourier transform. But, these methods are often used to obtain an instantaneous measurement like a formant or fundamental frequency, rather than as the actual object of analysis. Tools like absement provide an opportunity to incorporate more of the time-varying nature of speech into our formal analyses.

1.2. Dynamic time warping as absement

Perhaps the most common calculation of absement in phonetics has been dynamic time warping. Dynamic time warping is a function that compares two time-series and yields an overall difference or cost value. When the comparison is performed, distance is calculated between each pairwise combination of time steps in the two signals. Then, dynamic programming is used to find a nonlinear warping path indicating which time steps should be compared between the two signals to minimize the overall computed difference between them. This discrete sum over time is roughly analogous to a rectangle method of approximation to the Riemann integral of distance. An example of the distance values calculated over time between *afternoons* and *affection* with dynamic time warping is shown in Figure 1.

The nonlinear warping path is crucial for several reasons. First, it allows for time series of different lengths to be compared, and the durations of any two productions are rarely the same. Second, it allows elasticity in the comparisons, which is important since word and segment durations vary between different productions in speech without resulting in wildly different percepts. Finally, this process allows for similar regions of each signal to be compared, permitting, for example, vowels in one signal to be compared to vowels in the other, and not necessarily to consonants.

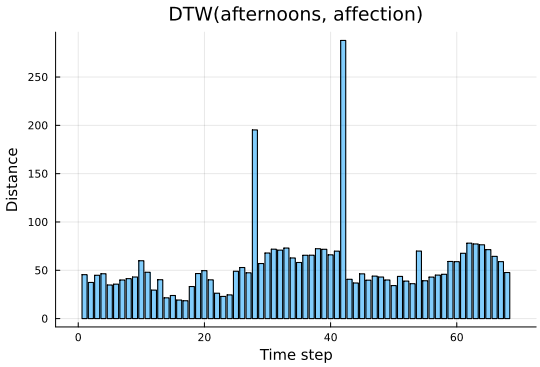


Figure 1: Distance values over time calculated from dynamic time warping between *afternoons* and *affection*. When a time point in *afternoons* was stretched over multiple time points in *affection*, those values were summed. Summing the area contained in each column will yield the dynamic time warping cost qua absement.

The shortcoming, however, is that the nonlinear warping path somewhat perturbs the natural analogy to absement from mechanics. While dynamic time warping does sum distance over time, the movement through time is not linear or uniform, and some moments in time would have multiple distance measurements associated with them. Regardless, I maintain that it is useful to refer to the output of dynamic time warping as absement to differentiate it from distance between two instantaneous measurements.

Dynamic time warping as absement has found recent explicit use in spoken word recognition [4], and previous research has also employed dynamic time warping without invoking absement [5]. Dynamic time warping dates back to the 1970s [6, 7], when it was developed to serve as the computational engine undergirding template-based automatic speech recognition. The field of automatic speech recognition has by and large moved away from dynamic time warping, first to hidden Markov models paired with Gaussian mixtures, of which dynamic time warping is a special case [8]. Modern speech recognition makes prevalent use of deep neural network models. Yet dynamic time warping has persisted as a useful method in data science [9, 10] and cognitive science [11, 12].

2. METHODS

The speech recognition experiment takes the form of an isolated word recognition task. The task is performed over a lexicon of 1,000 words and serves to demonstrate how absement can be a useful concept in phonetics.

2.1. Materials

The audio data came from the Massive Auditory Lexical Decision (MALD) project from Tucker et al. [13]. The project comprises responses to over 26,000 different English words in an auditory lexical decision task. The stimuli for the initial experiment were from a young male speaker of Canadian English who was trained in phonetics. Two other speakers were also recorded for auditory stimuli. These speakers were a young female speaker and an older male speaker of Canadian English. More recording information is given in [13].

I randomly sampled 1,000 of the words from the project that were recorded by all speakers. I converted each recording to a mel frequency cepstral coefficient (MFCC)-by-time representation using `MFCC.jl` v0.3.3 [14] in `Julia` v1.8.2 [15]. The window length was 25 ms with an advance of 10 ms. 13 coefficients were calculated, and the first coefficient was replaced with log energy, as is standard in automatic speech recognition.

I then used dynamic barycenter averaging [16] to create an average between the young female speaker and the older male speaker's recordings for each word. The averaging was performed using the phonetic sequence averaging interface in `Phonetics.jl` v0.1.2 [17], which relies on `DynamicAxisWarping.jl` v0.4.12 [18]. In the averaging process, I randomly selected which of the two recordings of each word would serve as the basic template for the initial average sequence in the dynamic barycenter averaging process.

The MFCC representation of the young male speaker was used to simulate a listener hearing someone speaking. The averages between the younger female and older male speakers' word productions were used as a set of acoustic templates that the listener would discriminate between based on the incoming acoustic signal. That is, the averaged productions were used as an acoustic representation of a small lexicon. I note that "representation" here is used in a general sense of one thing that stands in for another, and not necessarily as a cognitive model of language. Dynamic time warping in this sense can be approximately thought of as quantifying how different the spectrograms are between the young male speaker's productions and the spectrograms averaged between the young female and older male speakers' productions.

2.2. Analysis

Absement between each of the young male speaker's recordings and each of the average recordings was calculated using dynamic time warping with `DynamicAxisWarping.jl`. The distance function was Euclidean distance, and no warping radius was used. It is important to recall that absement can be increased both by duration and distance. While it is reasonable for duration to affect the absement value to some degree in spoken word recognition, the dynamic time warping calculation is such that shorter words tend to have smaller absement values on average, which biases recognition to short words.

To overcome this undesirable behavior, some of the length discrepancy must be factored out. However, care must be used when performing this kind of factorization because simply dividing by the length in time transforms the absement value into average distance, which is no longer absement *per se*. This relationship is analogous to how dividing distance by time yields average speed. A manual search of scaling functions suggested that dividing the absement values by the square root of the length of the averaged template helped normalize away some of the undesirable effect of duration on the absement values without completely destroying the ability to interpret the values as absement.

The scaled absement values for each word were then sorted, and the top ten words in the lexicon with the lowest absement values for a given word were recorded. The word with the lowest absement value was taken as the word that was ultimately recognized given the acoustic input.

3. RESULTS AND DISCUSSION

Of the 1,000 words compared, 57.9% were identified correctly based on the scaled absement values, and 87.9% were in the top ten. These results compare favorably to some recent models of spoken word recognition using naive discriminative learning. Arnold et al. [19] reported an accuracy of 25.2% on a 1,000 word recognition task, and Shafei-Bajestan and Baayen [20] reported an average best accuracy of 11.72% correct on their various tasks for clean speech.

The higher recognition accuracy in the present data must be considered with some caveats. The main limitation is that a 1,000 word lexicon is too limited to be a good reflection of a proficient speaker's knowledge. Additionally, as more words are introduced into the lexicon, the accuracy will almost assuredly fall. Discrimination tasks generally become harder when there are more

items to discriminate between, due to the curse of dimensionality. Indeed, both naive discriminative learning models used a much fuller set of over 10,000 lexomes in their lexicons, regardless of task. Their lexicons were also based on corpora of connected conversational speech, so they were ultimately performing a harder task than the one I performed here.

Regardless of any potential inflation in the accuracy I reported, I want to highlight that it would be much more difficult to recognize words using comparisons of instantaneous measurements. An acoustic measurement taken at a specific point in time is an infinitesimal portion of the acoustic signal and is insufficient to account for the communicative purpose of speech. Previous research has highlighted the need to situate phonetic measurements and cues in the communicative system they are used in [21, 22], discriminating between different words (or forms) and their associated meanings as in [23]. Incorporating tools like absement into our methodologies is one way to better account for the temporal structure of the speech signal.

It is also important to note that the model I have presented in the present paper does not make use of any sublexical unit like a segment, phone, or phoneme. The only possible sublexical units in the method I used are subsets of the MFCC matrices that represent the words in the lexicon. In the context of the present paper, I do not wish to interpret this as either evidence for or against the ontological or psychological reality of traditional sublexical units in linguistics. However, some previous models of spoken word recognition have proposed eschewing these types of units in favor of wholesale spoken word discrimination based solely on acoustic information [19, 20, 23, 24].

I do think that it will be worthwhile to explore this type of spoken word recognition model further. But, we should also keep in mind that this type of word recognition directly from acoustic data hails from a venerable idea from the early days of modern speech technology. It has also been reiterated at various points throughout the field's history. Exemplar theory models based on whole-word acoustics are one example [25]. Goldinger [26] also argued that many linguistic units measured in speech perception tasks emerge based on experimental task. Making explicit use of absement between acoustic representations may help push these lines of thinking forward to either present more convincing accounts of or eliminate certain hypotheses about spoken word recognition.

Absement is a general concept and can be applied across a wide variety of tasks. It could, for example, be used to quantify the differences between vowel formant trajectories. It has indeed already seen similar such use via dynamic time warping on pitch contours [27]. Virtually any comparison between a series of acoustic measurements made over time is amenable to being treated with absement.

There is also more to be said about the use of dynamic time warping to instantiate and measure acoustic absement between two words. Because dynamic time warping has a propensity to allow duration to have an outsized effect on the output value, and because it can have certain time points repeat in the warping path, it is an imperfect representation of the concept of absement. The resolution I used here of scaling the absement values for better recognition accuracy is tantamount to having a constant weight function within the integral of distance. It may be worth exploring more sophisticated weight functions in the future or finding algorithms for calculating absement that do not require weighting.

The code used to perform the analysis and a table of the results is available as supplementary material at https://doi.org/10.5281/zenodo.7823844 or on GitHub at https://github.com/maetshju/absement_in_detail_code.

4. CONCLUSION

Absement is not an entirely new concept in phonetics. However, having a label for the type of time-series comparisons I have exhibited in the present paper should make working with this concept easier. Such a label may also nudge the idea of time-series analysis closer to the fore of our minds. Although, I am sure that none among us in phonetics truly need reminding of the time-varying nature of speech.

5. REFERENCES

[1] T. Kendall and V. Fridland, "Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift," *Journal of Phonetics*, vol. 40, no. 2, pp. 289–306, Mar. 2012.

[2] S. Mann, R. Janzen, and M. Post, "Hydraulophone design considerations: Absement, displacement, and velocity-sensitive music keyboard in which each key is a water jet," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06. New York, NY, USA: Association for Computing Machinery, Oct. 2006, pp. 519–528.

[3] S. Mann, M. L. Hao, M. Tsai, M. Hafezi, A. Azad, and F. Keramatimoezabad, "Effectiveness

of Integral Kinesiology Feedback for Fitness-Based Games," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*, Aug. 2018, pp. 1–9.

[4] M. C. Kelley and B. V. Tucker, "Using acoustic distance and acoustic absement to quantify lexical competition," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 1367–1379, Feb. 2022.

[5] R. Bennett, K. Tang, and J. A. Sian, "Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan)," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 9, no. 1, p. 9, May 2018.

[6] H. Sakoe and S. Chiba, "A similarity evaluation of speech patterns by dynamic programming," in *Nat. Meeting of Institute of Electronic Communications Engineers of Japan*, 1970, p. 136.

[7] ——, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[8] B.-H. Juang, "On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition—A Unified View," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 7, pp. 1213–1243, 1984.

[9] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: Association for Computing Machinery, Aug. 2012, pp. 262–270.

[10] R. Wu and E. J. Keogh, "FastDTW is Approximate and Generally Slower Than the Algorithm it Approximates," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3779–3785, Aug. 2022.

[11] T. Schatz, N. H. Feldman, S. Goldwater, X.-N. Cao, and E. Dupoux, "Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input," *Proceedings of the National Academy of Sciences*, vol. 118, no. 7, p. e2001844118, Feb. 2021.

[12] J. Millet and E. Dunbar, "Do self-supervised speech models develop human-like perception biases?" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7591–7605.

[13] B. V. Tucker, D. Brenner, D. K. Danielson, M. C. Kelley, F. Nenadić, and M. Sims, "The Massive Auditory Lexical Decision (MALD) database," *Behavior Research Methods*, vol. 51, no. 3, pp. 1187–1204, Jun. 2019.

[14] D. van Leeuwen, "MFCC.jl," 2022. [Online]. Available: https://github.com/JuliaDSP/MFCC.jl

[15] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, Jan. 2017.

[16] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, Mar. 2011.

[17] M. C. Kelley, "Phonetics.jl," 2022. [Online]. Available: https://github.com/maetshju/Phonetics.jl

[18] F. Bagge Carlson, "DynamicAxisWarping.jl," 2022. [Online]. Available: https://github.com/baggepinnen/DynamicAxisWarping.jl

[19] D. Arnold, F. Tomaschek, K. Sering, F. Lopez, and R. H. Baayen, "Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit," *PLOS ONE*, vol. 12, no. 4, p. e0174623, Apr. 2017.

[20] E. Shafaei-Bajestan and R. H. Baayen, "Wide learning for auditory comprehension," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 966–970.

[21] C. Redmon and A. Jongman, "From interfaces to system embedding: Phonetic contrasts in the lexicon," Nov. 2022.

[22] C. H. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," Ph.D. dissertation, University of Kansas, 2020.

[23] R. H. Baayen, Y.-Y. Chuang, E. Shafaei-Bajestan, and J. P. Blevins, "The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning," *Complexity*, vol. 2019, 2019.

[24] E. Shafaei-Bajestan, M. Moradipour-Tari, P. Uhrig, and R. H. Baayen, "LDL-AURIS: A computational model, grounded in error-driven learning, for the comprehension of single spoken words," *Language, Cognition and Neuroscience*, vol. 0, no. 0, pp. 1–28, Jul. 2021.

[25] K. Johnson, "The auditory/perceptual basis for speech segmentation," Ohio State University Department of Linguistics, Working Paper, Jul. 1997.

[26] S. D. Goldinger and T. Azuma, "Puzzle-solving science: The Quixotic quest for units in speech perception," *Journal of Phonetics*, vol. 31, no. 3, pp. 305–320, Jul. 2003.

[27] D. R. McCloy, "Prosody, intelligibility and familiarity in speech perception," Ph.D. dissertation, University of Washington, Jul. 2013.

THE IMPACT OF THE INTENSITY RATIO BETWEEN VOWELS AND VOICELESS PLOSIVES ON THE INTELLIGIBILITY OF SUNG TEXT

Allan Vurma[a], Einar Meister[b], Lya Meister[b], Jaan Ross[a], Marju Raju[a], Veeda Kala[a], Tuuri Dede[a]

[a] Estonian Academy of Music and Theatre, Tallinn, Estonia
[b] Tallinn University of Technology, Estonia
allan.vurma@eamt.ee; einar.meister@taltech.ee; lya.meister@taltech.ee; jaan.ross@gmail.com;
marju.raju@eamt.ee; veeda.kala@eamt.ee; tuurielo@gmail.com

ABSTRACT

In classical singing the intelligibility of the sung text is often less clear than in speaking. Our study aims to evaluate the hypothesis that in singing, compared to speaking, the intensity of voiceless plosives increases less than the intensity of vowels. The second goal is to determine whether pronouncing plosive bursts with greater intensity improves their intelligibility. Analysis of the performances of five Italian arias from the romantic period by five classically trained singers revealed that, in singing, the intensity of the vowels increased by 14.2 dB compared to the same sounds in speaking, whereas the intensity of the plosive bursts increased by only 7.1 dB. In the perception test with 60 participants, pronouncing plosive bursts more forcefully improved their recognition, but mainly only when the stimuli were presented with room reverberation and/or with pink noise imitating the masking sounds of the accompaniment.

Keywords: intelligibility, plosives, intensity, reverberation, masking

1. INTRODUCTION

In singing, it is the vowels that mainly determine the general sonority of the voice, while to secure text intelligibility a much greater functional load is on consonants [8, 17]. Research addressing consonants in singing is scarce. In the literature we can find some points of view expressed by voice teachers on how consonants should be pronounced to improve text intelligibility. These opinions, however, often contradict each other. For example, Ware [17], Christy [4], Vennard [16], Melton [11], and Sharonova [13] claim that consonants in singing should be pronounced with exaggerated power, while Fuchs [9], Marshall [10], and Brown [3] have stated that singers, on the contrary, should avoid this.

In the present study we focus on researching the intensity balance between vowels and voiceless plosives. While the intensity of vowels mainly depends on the level of subglottal air-pressure created by the lungs, the air-pressure to produce plosive bursts is created in the mouth cavity by rapidly decreasing its size, e.g., by quickly raising the larynx. The breathing system related to the lungs may not play a direct role here [1]. Therefore, the relative independence of the systems which determine the air pressure to produce vowels and the air pressure to produce /k/, /p/, and /t/ bursts may provide a theoretical basis to explain why, when changing the overall voice intensity, the intensity of vowels and voiceless plosives could change by a different amount.

In terms of text intelligibility, not only the plosive burst but also the glide of the formant frequencies from the *locus* of the plosive to the adjacent vowel may also play a role in providing a clue [6, 7]. Additional factors which could worsen the sung text intelligibility are the reverberation of the concert or opera hall and the sounds from the accompaniment and/or ensemble partner(s), which can mask the singer's voice [12].

This study consists of two stages. In the first stage, our aim is to find out whether or not in singing, compared to speaking, the intensity of voiceless plosives increases less than the intensity of vowels. The goal of the second stage is to investigate whether producing more intense plosive bursts improves their recognition.

2. STAGE I

2.1. Method I – Acoustic analysis and comparison of sung and read texts

Five singers (native Estonian speakers) were asked (1) to sing an Italian aria from a romantic period opera and (2) to read the text of the same aria. The recordings (with SPL calibration @30cm) were carried out in a low-reverberation recording studio using Soprano 1.0.20 software and a DPA SC4061-FM omnidirectional microphone (distance from the corner of the singer's mouth about 3 cm). Subsequently, the recordings were segmented into the level of single vowels and consonants (including plosive stops and bursts) using Praat [2]. The intensities of the vowels and plosive bursts were measured with a custom Praat script.

2.2. Results I

In Figure 1 we can see that all the speech sounds investigated were produced more intensely on average in singing than in reading. The average intensity of vowels was 94.5 dB (*SD* = 6.2 dB) and that of voiceless plosives 71.5 dB (*SD* = 6.3 dB) in singing, compared to 80.3 dB (*SD* = 5 dB) and 64.4 dB (*SD* = 5.3 dB) in speaking. The increase in intensity was 14.2 dB on average in the case of vowels but only 7.1 dB in the case of voiceless plosives. This difference is statistically significant ($t = 13.1$, $p < 0.001$).

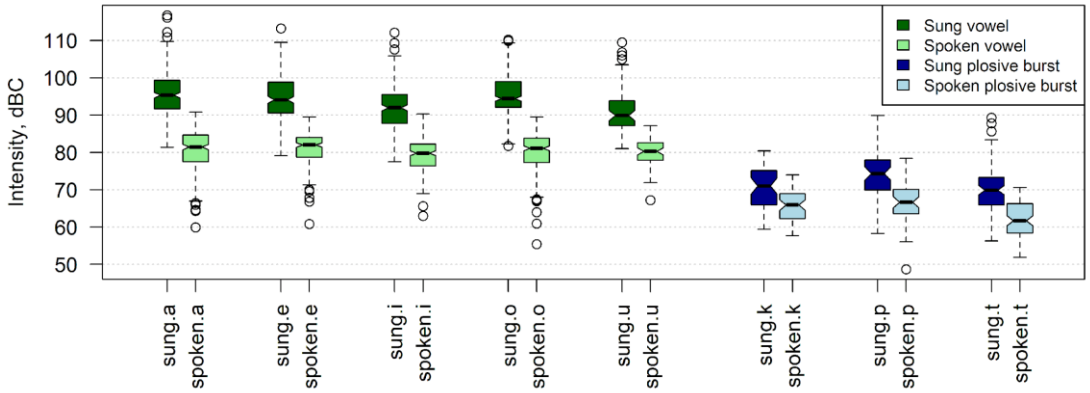


Figure 1. The boxplots of the intensities of sung and spoken vowels and voiceless plosives on an absolute scale (dBC weighting was used in the calibration).

2.3. Are the speech sounds intensity relationships similar in the case of Estonian and Italian singers?

The question may arise whether the aforementioned intensity relationships could depend on the mother tongue of the singer. To investigate that possibility, we also analysed three performances of Donizetti, Puccini and Verdi opera arias by three Italian singers recorded *a cappella* in a low-reverberation studio for another study in Bologna, the data of which has been made freely available in the internet repository (https://zenodo.org/record/3628247#.Y61fEHZBy5c accessed on 29.12.2022).

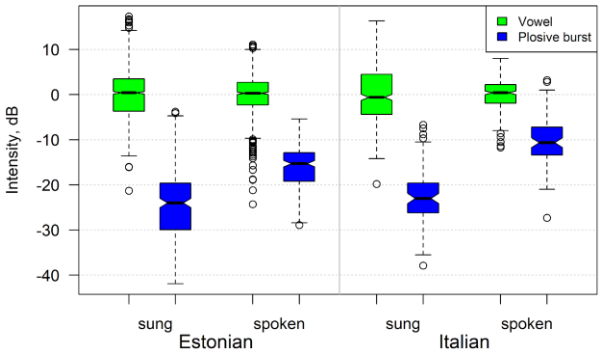


Figure 2. The intensity relationships between vowels and voiceless plosives in the case of Estonian and Italian vocalists.

To rate the intensity relationship between vowels and voiceless plosives in the case of reading, we investigated some random spoken excerpts in Italian (with great probability featuring native Italian speakers) from the internet (www.newsinslowitalian.com, accessed on 10.10.2022), the original purpose of which was to teach the Italian language.

In Figure 2, we can see that all the intensity relationships between vowels and voiceless plosives and between singing and reading are very similar in the case of Estonian and Italian singers and speakers. Therefore, we have good reason to believe that the issue addressed in our study is universal and does not depend to any great extent on the language background of the singer or speaker.

3. STAGE II

3.1. Method II – Perception tests

3.1.1. Stimuli

A classically trained mezzo-soprano sang the /a-k-a/, /a-p-a/ and /a-t-a/ sequences in a slow tempo on G4 (*f0* = 392 Hz) several times, varying each time the intensity of the plosive burst. Versions with a weak, intermediate and strong burst were selected as the basis for the stimuli for the perception test. Additional versions were created in which the plosive bursts were removed and replaced with silence. On the basis of these 12 stimuli (3 plosives × 4 burst magnitudes), we produced five more series of stimuli with the addition of reverberation (to imitate the room acoustics) and/or pink noise (to imitate the masking sounds from accompanying instruments). For this purpose, we used the *Praat Vocal Toolkit* [5]. For the reverberation, we applied the script's pre-programmed choice, "church", with two different reverberation rates (25% and 50%). We may estimate

that the reverberation time T60 of the acoustics modelled by the script was about five seconds. The complete paradigm of stimuli used for our perception test was 3 plosives × 4 burst intensities × 6 acoustic conditions = 72 stimuli.

3.1.2. Participants and the procedure

Altogether, 60 people (21 males, 38 females, one of another gender; aged between 11 and 74, average age 41) including the authors of this paper took the perception test. We tried to engage people over a wide age range, as could be the case in an audience at a random opera performance. The mother tongue of 55 of the participants was Estonian, but there were also participants whose mother tongues were Turkish, Chinese, Portugal, Russian and Spanish. We used the online platform *PsyToolkit* [14, 15] to administer the stimuli. The participants had to use their own laptop or mobile phone with headphones in quiet surroundings at a time convenient to them to run the test. For each participant, a unique order of stimuli was generated. Each stimulus was played only once, and the participants had to enter the consonant they heard between the vowels in the dedicated box on the screen. If recognition was not possible, the participant could enter a question mark in the box. It took between 7 and 24 minutes (14 minutes on average) to complete the test.

3.2. Results II

In Figure 3 we can see that if we look at the responses pooled over all six acoustic conditions the proportion of "correct" replies systematically improved with stronger plosive bursts only in the case of /k/, reaching almost 100% at the strongest burst (burst_3). According to the chi-square test, the dependence of the proportion between "correct" and "incorrect" answers from the intensity of the plosive burst was statistically significant ($\chi2 = 404.6, p < 0.001$).

In the case of /p/, the proportion of "correct" identifications first increased with a more intense plosive burst but already reached its maximum level at medium burst intensity (burst_2) to become worse again at the strongest burst (burst_3). Also, in the case of /p/, the dependence of the proportion of "correct" and "incorrect" identifications on the burst intensity was statistically significant both over the extent of the whole curve ($\chi2 = 135.9, p < 0.001$) and when the replies to only the two stimuli burst_3 and burst_2 ($\chi2 = 8.1, p = 0.004$) were compared, indicating that the decay of the curve at burst_3 was not random.

In the case of /t/, the rate of "correct" recognitions (about 80%) did not depend to any great extent on the intensity of the burst and fell significantly to 65% only in the case of the stimuli where the burst was completely removed.

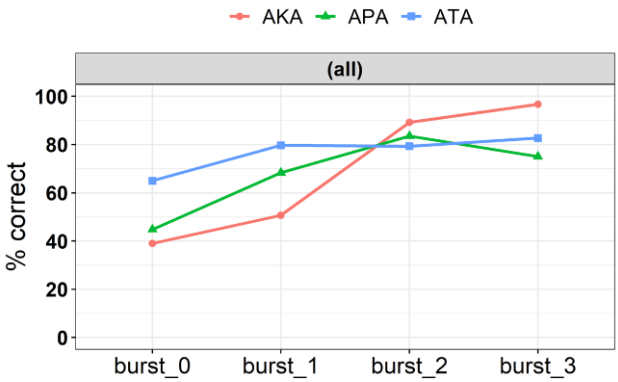


Figure 3. The dependence of the rate of "correct" identifications of plosives on the intensity of the plosive burst (burst_0 – burst removed, burst_1 – weak, burst_2 – medium, burst_3 – strong). The data is pooled over six acoustic conditions.

Despite the decline in the "correct" identifications also in the case of other stimuli with the plosive burst removed, the rate of "correct" identification never fell below the notional 33% borderline which distinguishes mere random guesses. Therefore, we may conclude that besides the plosive bursts, some other types of clues also play a role in identifying the voiceless plosives, and that the absence of the burst (or its complete masking by the accompaniment instruments) need not make the plosive impossible to identify (although the probability of "correct" identification diminishes).

Figure 4 describes the percentage of "correct" identifications split by six acoustic conditions. When the reverberation and/or pink noise were not added to the stimuli, their identification remained high (above 80%) regardless of the plosive burst intensity (the chi-square test was not able to show any statistically significant difference). The only exception is the stimuli with the removed burst (burst_0), in the case of which the level of "correct" identifications of /k/ fell to 22% (which is below the chance level of 33%). The decline in identification for /p/ and /t/ here was far smaller, but the difference in the proportion of "correct" and "incorrect" responses between the stimuli burst_0 and burst_1 was statistically significant in the case of all three plosives (respectively $\chi2 = 63, p < 0.000$ in the case of /k/, $\chi2 = 4.2, p = 0.04$ in the case of /p/ and $\chi2 = 7, p = 0.008$ in the case of /t/).

In the case of the stimuli with added 50% reverberation and with pink noise the recognition generally improved with a stronger burst, although in the case of /p/ there was an optimum burst intensity for the recognition at burst_2.

In the case of /k/, if 50% reverberation, or 25% reverberation and pink noise were added to the stimulus, the recognition improved somewhat if the burst was removed (burst_0) compared with the stimuli with the weakest burst (burst_1). We may speculate that here two clues (the plosive burst and the formant transition from the *locus* of the plosive to the adjacent vowel) are competing. Masking from reverberation and the accompaniment can disturb the clues and create ambivalence. If the plosive burst is absent, the clue from the formant transition becomes dominant and the ambiguity decreases.

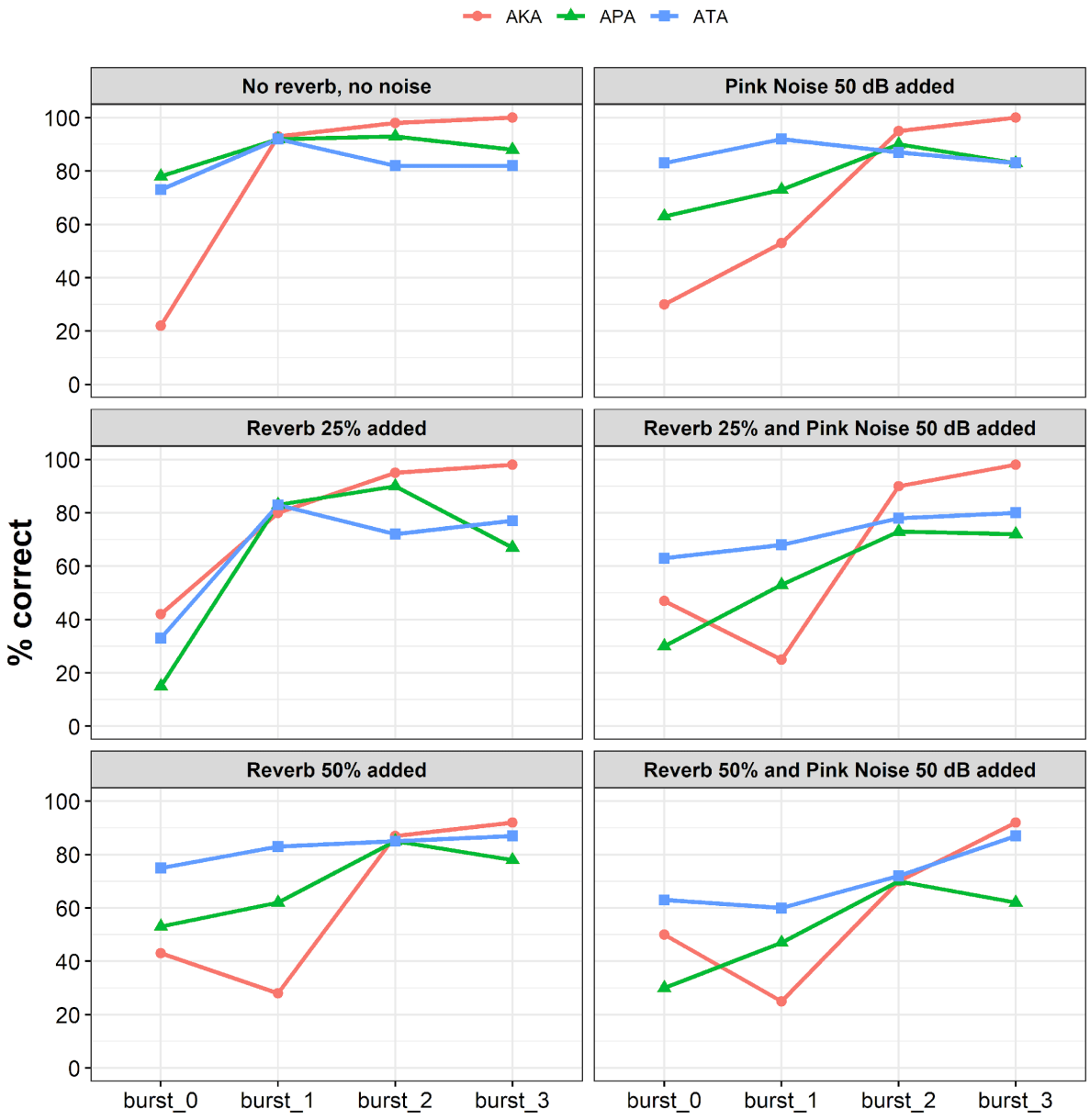


Figure 4. The dependence of the rate of "correct" identifications of plosives on the intensity of the plosive bursts at six acoustic conditions.

4. CONCLUSIONS

The results of the first stage of our study reveal that in operatic singing, when compared to speaking, the balance of intensity between vowels and voiceless plosive bursts changes in favour of the vowels, bringing a relative deterioration in the intensity of the plosives. The results of the second stage show that singing in acoustic conditions with low or absent reverberation and without masking from the accompaniment the intensity of plosive bursts is not critical to text intelligibility – a stronger burst may not improve the intelligibility. However, the total absence of the burst lowers the recognition, especially in the case of /k/. In the case of stronger reverberation and when singing with an accompaniment (or ensemble partners), more intense plosive bursts may improve recognition, although an optimum intensity level of the burst may exist (especially for the /p/) at which the intelligibility is the best.

5. ACKNOWLEDGEMENTS

This study was supported by the Estonian Research Council grant PRG1552.

6. REFERENCES

[1] Behrman, A. 2018. *Speech and voice science. Third edition.* San Diego, CA: Plural Publishing Inc.

[2] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3., retrieved 15 November 2022 from http://www.praat.org//

[3] Brown, R. M. 1946. *The singing voice.* New York, NY: Macmillan Co.

[4] Christy, V. A. 1967. *Expressive singing, Volume 1, Basic principles, A Text for school or studio class or private teaching*. Dubuque, WC: C. Brown Publishing Co.

[5] Corretge, R. 2012-2023. Praat Vocal Toolkit. https://www.praatvocaltoolkit.com (Last viewed Dec 08, 2022).

[6] Deme, A. 2014. Intelligibility of sung vowels: the effect of consonantal context and the onset of voicing. *Journal of Voice,* 28, 4, 523.e19–523.e25

[7] Eek, A, Meister, E. 1996. Acoustics and perception of Estonian word-initial plosives [in Estonian]. *Keel ja* Kirjandus, 4, 241−253.

[8] Fletcher, H. 1929. *Speech and hearing*. New York, NY: D. Van Nostrand.

[9] Fuchs, V. 1964. *The art of singing and voice technique*. New York, NY: London House and Maxwell.

[10] Marshall, M. 1956. Is exaggeration required for good English diction? *Diapason*, 47, 18.

[11] Melton, J. 1953. Do you put the words across? An interview by Annabel Comfort. *Etude*, June, 15.

[12] Meyer, J. 2009. *Acoustics and the performance of music. Manual for acousticians, audio engineers, musicians, architects and musical instrument makers. 5th edition.* New York, NY: Springer.

[13] Sharonova, S. 1947. Diction. *The NATS bulletin*, 3, 4.

[14] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096–1104.

[15] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24–31.

[16] Vennard, W. 1967. *Singing: The mechanism and the technic*. New York, NY: Carl Fisher.

[17] Ware, C. 1998. *Basics of vocal pedagogy: the foundations and process of singing*. Boston, MA: McGraw-Hill.

ACOUSTIC PROPERTIES OF BILABIAL TRILLS IN MANGBETU

Kenneth S. Olson

SIL International
ken_olson@sil.org

ABSTRACT

Mangbetu (D. R. Congo) has three phonemes with bilabial trill releases: /pᴮ, bᴮ, ᵐbᴮ/. This study adds to previous work on Mangbetu by including VOT and more thorough stop closure duration measurements, in addition to trill vibration frequency measurements.

The key factor distinguishing /pᴮ/ from /bᴮ/ and /ᵐbᴮ/ is the absence of voicing during the stop closure of /pᴮ/. The stop closures of /bᴮ/ and /ᵐbᴮ/ are either entirely voiced or voicing terminates near the end of closure. The VOT of /pᴮ/ is short enough that trill oscillations occur after the start of voicing. Hence, transcribing the release of /pᴮ/ as voiced [ᴮ] is warranted.

As in other languages, the oral stop closure of /ᵐbᴮ/ is short, and the failure rate of bilabial trilling increases before front vowels.

Compared to previous studies, the stop closure is longer and the frequency rate of trilling is slower, suggesting careful speech.

Keywords: bilabial trill, VOT, release, Central Sudanic, Mangbetu.

1. INTRODUCTION

Bilabial trills are speech sounds found in about 70 languages. Clusters of languages with bilabial trills are found in Papua New Guinea, Vanuatu, Cameroon, China, the Democratic Republic of the Congo (DRC), South Sudan, and Brazil. Individual languages with bilabial trills are found in Indonesia, India, and Georgia [1, 2].

Unlike [r] and [ʀ], bilabial trills are not attested as unitary phonemes. Rather, they occur in complex phonemes where they pattern as the release of a stop. The preceding stop that initiates the trill is usually bilabial, e.g. /pᴮ, bᴮ, ᵐbᴮ/, but it can also be coronal, e.g. /tᴾ/ [3, 4], or labial-velar, e.g. /kpᴮ, gbᴮ, ᵑgbᴮ/. The one exception to this pattern is found in China, where bilabial trills pattern as allophones of compressed or fricative vowels following bilabial and/or coronal stops, e.g. [bʙ, dʙ] [2].

In northeastern DRC and South Sudan, bilabial trills are attested in five Central Sudanic languages (Mangbetu, Lombi, Asoa, Baka, and Morokodo), one Bantu language (Lika), and one Ubangian language (Dongo) [1]. Cases involving labial-velar stop closures are reported in Lombi, Baka, Morokodo, and Lika [5, 6, 7, 8, 9, 10].

Mangbetu (ISO 639-3 code: mdj) is spoken by over 650,000 people in northeastern DRC in and around the city of Isiro [11]. Several studies have documented bilabial trills in Mangbetu [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. In addition, Demolin has studied the acoustics of the sounds [16, 18, 19].

The Mangbetu consonant inventory includes three phonemes with bilabial trill releases: /pᴮ, bᴮ, ᵐbᴮ/. The following data set demonstrates contrast: [nó.pʙù] ‘to bring out from within’, [nó.bʙù] ‘to fan’, [nó.mbʙù] ‘to enclose’, [ná.pú] ‘ashes’, [nó.bù] ‘to feed a fire’, [né.mbó] ‘have him tour (3sg hortative)’, [né.mbʙó] ‘clay pot’ [21, 22].

Demolin reported on the frequency rate of trilling and included some measurements of stop closure duration. Besides replicating these measurements, my study builds on Demolin’s work by providing more complete stop closure duration information, as well as measurements of voice onset time (VOT).

2. METHODOLOGY

2.1. Subject

An approximately 35-year-old educated male native speaker of Mangbetu (Meegye dialect) participated in this experiment. The subject lives in Isiro, DRC. He has completed some graduate school. Besides Mangbetu, he also speaks French and Lingala. His primary occupation is as a primary school director.

2.2. Procedures

We recorded a list of 73 words and short phrases spoken in isolation. These words were chosen to include bilabial trill and labiodental flap data.

The wordlist was recorded during two sessions: (1) on March 6, 2004 (items 1–43) and (2) March 27, 2004 (items 44–73). The recordings were made at the ACATBA center (Association Centrafricaine pour la Traduction de la Bible et l’Alphabétisation) in Bangui, Central African Republic. The subject was seated and read prompts on sheets of paper in orthographic transcription, which he wrote.

During the first recording session, the subject produced the gloss for each word in French and then the corresponding word in Mangbetu once. During

the second session, the subject produced the word in Mangbetu once followed by the French gloss. This part of the wordlist was recorded twice.

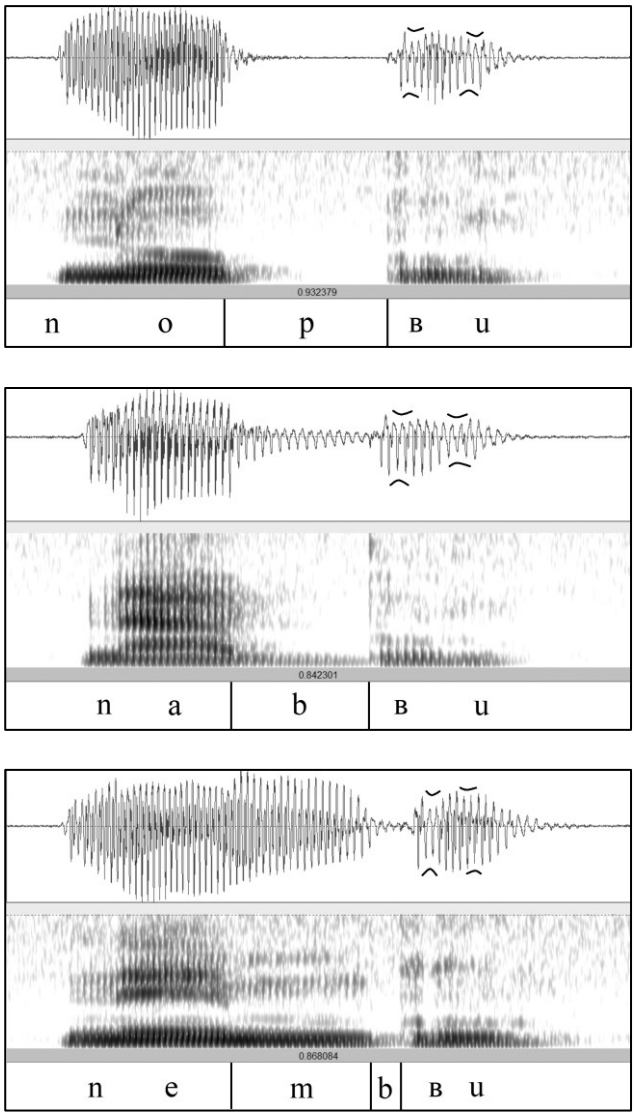


Figure 1: Waveforms and spectrograms of the words [nó.pʁù] 'to bring out from within', [ná.bʁù] 'corpse', and [né.mʁù] 'hip'.

The recording was made with a Marantz PMD 420 monaural cassette recorder and an Audio-Technica ATM 33a condenser microphone. The recording was digitized at the International Linguistics Center in Dallas, Texas in March 2005. The audio cassettes were played on a Marantz PMD 221 analog cassette recorder, and the recording was digitized onto a standard Windows XP computer using a Tascam US-122 USB Audio/Midi interface and CoolEdit 2000 for audio capture. The digitization was done at a 48 kHz sampling rate and a 24-bit quantization. The recordings were stored in non-compressed WAV format. Both a presentation form [22] and an archived form [23] of the recordings are available.

2.3. Analysis

Waveforms and spectrograms were examined using Praat [24]. Default parameters were used, except that the dynamic range of spectrograms was set at 45 dB.

First, the closure duration of the stop preceding the trill release was measured. For prenasalized stops, both the nasal closure (indicated by nasal formants) and the oral closure (indicated by a significant drop in signal amplitude) were measured.

Second, the number of trill periods [25] for each token was counted. For the speaker, the modal number of trill periods was one. In some tokens, a trill was expected but was not observed in the waveform or spectrogram. Rangelov [26, 27] refers to these as *failed trills*, as opposed to *successful trills*.

Third, trill period duration was measured. The first trill period was defined as the time from the release of the stop closure to the release of the first trill closure (or partial closure), indicated by an increase in signal amplitude [25]. Subsequent trill periods were defined as the time from the release of the trill closure to the release of the following trill closure.

Fourth, VOT was measured. This was defined as the time from the release of the stop closure to the beginning of vocal fold vibration, indicated by periodicity in the waveform and the start of vertical striations in the spectrogram. One outlier was removed.

Sample waveforms and spectrograms are given for /p^B, b^B, $^mb^B$/ in Figure 1. A phonetic transcription is provided below each spectrogram, with segments demarcated where appropriate. Arcs were added to the waveform to indicate the (partial) trill closures.

3. RESULTS

3.1. Stop closure duration

The mean duration of the stop closures for /p^B, b^B, $^mb^B$/ are 193 ms, 168 ms, and 190 ms, respectively, as shown in Table 1.

Phoneme	n	Closure duration ($\bar{x}$)	s.d.
/p^B/	43	193 ms	27.3
/b^B/	13	168 ms	25.0
/$^mb^B$/	18	190 ms	31.7

Table 1: Stop closure duration measurements for /p^B, b^B, $^mb^B$/. n = number of tokens, $\bar{x}$ = sample mean, s.d. = standard deviation.

These durations are generally longer than those found in previous studies [25, 26, 28]. I consider this to be suggestive of careful speech. Rangelov [26] notes that there is substantial interspeaker variation for these measures in Ahamb. I expect that examining

the speech of additional speakers in Mangbetu would likely show similar variation.

The stop closure durations of the three phonemes—in which /pᴮ/ and /ᵐbᴮ/ are similar and /bᴮ/ is somewhat shorter—harmonize with previous studies in two respects. First, Rangelov found similar stop closure durations in Ahamb for /pᴮ/ and /ᵐbᴮ/. (/pᴮ/ was slightly longer.) Second, similar patterns have been noted for the stops /p, b, ᵐb/. Lisker [29] observed that the closure duration of English /b/ is shorter than that of /p/ in intervocalic position. Maddieson [30] found that the closure duration of /p/ and /ᵐb/ are similar for Fijian. Olson [31] noted that the closure duration of /b/ is shorter than that of /ᵐb/ in Mono (DRC).

In other words, stops with bilabial trill releases appear to pattern similarly to plain bilabial stops. More work needs to be done to establish that this apparent trend occurs consistently crosslinguistically.

For /ᵐbᴮ/, the mean duration of the nasal portion of the stop closure was 148 ms (s.d. = 28.1, n = 18) and the mean duration of the oral portion of the stop closure was 41.8 ms (s.d. = 13.7, n = 18).

This short oral closure for /ᵐbᴮ/ is well attested [25, 26, 28], and a short oral closure has been documented for /ᵐb/ as well [30, 31]. In other words, the short oral closure of /ᵐbᴮ/ does not appear to be novel but is rather consistent with prenasalized stops in general [32].

For /pᴮ/, there is a rapid attenuation of voicing at the start of the oral stop closure, such that most of the closure is voiceless. Voicing does not start again until after the release of the closure.

For /bᴮ/ and /ᵐbᴮ/, voicing is maintained throughout closure in a little more than half of the tokens. In the rest of the tokens, voicing slowly attenuates until it terminates roughly 4/5 of the way through the closure. Figure 2 shows a waveform and spectrogram where this is the case for /bᴮ/.

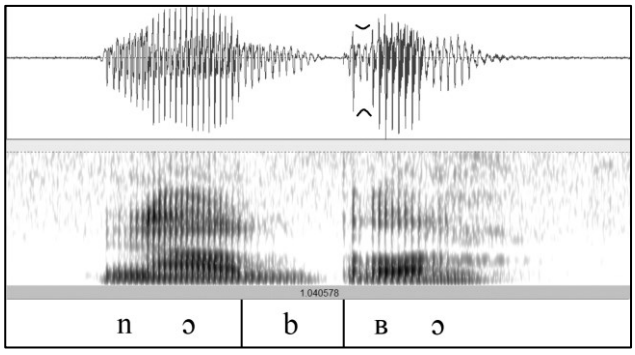


Figure 2: Waveform and spectrogram of the word [nɔ́.bʙɔ̀] ‘to carry’.

It appears then, that the absence of voicing during the production of the stop closure of /pᴮ/ is a strong correlate distinguishing it from /bᴮ/ and /ᵐbᴮ/.

3.2. Number of trill periods

The number of trill periods varied from zero to three in my data. Out of 74 tokens, 36 had one trill period, 15 had two trill periods, and 2 had three trill periods. Hence, the modal number of trill periods was one. This is the same number that Maddieson [25] observed for Ngwe. The data are somewhat limited, but it appears that crosslinguistically two modal trill periods is most common, e.g. Kurti, Na?ahai, Uripiv, Ahamb [25, 26].

The remaining 15 tokens had zero trill periods, i.e. these were failed trills. Previous studies [3, 25, 26, 33] have shown that bilabial trills often do not occur when they are expected to do so, and this is also the case for Mangbetu.

Table 2 shows the number of trill periods broken down by phoneme. In my data, /bᴮ/ and /ᵐbᴮ/ occur only before back and low vowels, whereas /pᴮ/ occurs before front, back, and low vowels. To ensure a like-for-like comparison, I removed the count of tokens in which /pᴮ/ occurs before front vowels from Table 2.

Of the three phonemes, /ᵐbᴮ/ has the highest percentage of successful bilabial trills (94%). It also had the highest number of tokens with two or more trill periods (39%). This highlights the robustness of this particular phoneme [25, 32].

On the other hand, /pᴮ/ has the lowest percentage of successful trills (75%).

Phoneme	n	0 t.p.	1 t.p.	2 t.p.	3 t.p.
/pᴮ/	12	3	5	3	1
/bᴮ/	13	1	10	2	0
/ᵐbᴮ/	18	1	10	6	1

Table 2: Number of trill periods per token for /pᴮ, bᴮ, ᵐbᴮ/. Tokens with a following front vowel are not included in the count.

Another trend emerges if we examine all the tokens of /pᴮ/, including those with a following front vowel. Table 3 shows the number of trill periods per token broken down by whether the following vowel is front, back, or /a/.

The success rate of the trill release of /pᴮ/ is very high when followed by a back vowel (100%). On the other hand, the success rate drops considerably before front (48%) and low (25%) vowels. In other words, bilabial trilling is more successful before back vowels than before other vowels. This harmonizes with other studies that have shown that bilabial trilling is most robust before back vowels [25].

Voicing was not inhibited during bilabial trilling in Mangbetu, contra what is sometimes observed for apical trills [25]. However, as noted in Section 3.1, voicing was inhibited during the preceding stop closure in some tokens.

Following V	n	0 t.p.	1 t.p	2 t.p.	3 t.p.
Front V	31	16	11	4	0
/a/	4	3	1	0	0
Back V	8	0	4	3	1

Table 3: Number of trill periods per token for /pᴮ/, before front vowels, back vowels, and /a/.

3.3. Trill period duration and frequency rate of trilling

The duration of the first trill period for /pᴮ, bᴮ, ᵐbᴮ/ is shown in Table 4. The corresponding frequency rates of trilling are shown as well.

Phon.	n	t.p. (x̄)	s.d.	freq. (x̄)	s.d.
/pᴮ/	24	46.4 ms	4.7	21.7 Hz	2.2
/bᴮ/	12	52.4 ms	4.4	19.2 Hz	1.6
/ᵐbᴮ/	178	56.3 ms	5.2	17.9 Hz	1.7

Table 4: Duration and frequency rate of trilling of first trill period for /pᴮ, bᴮ, ᵐbᴮ/.

These durations are longer than those measured in previous studies [16, 18, 25, 26]. Consequently, the frequency rates of trilling are shorter than previously observed. I consider this to be indicative of careful speech on the part of the Mangbetu speaker.

There is a general decrease of frequency rate as one goes from /pᴮ/ to /bᴮ/ to /ᵐbᴮ/. This harmonizes with what Rangelov [26] found in Ahamb, where the frequency rate of /pᴮ/ is higher than that of /ᵐbᴮ/. Further research is necessary to determine if this is a crosslinguistic trend.

3.4. Voice onset time

All cases of /pᴮ/ have a measurable VOT. Importantly, the start of voicing preceded the start of the first trill closure in all but two tokens of /pᴮ/. This suggests that the transcription of the trill release as voiced [ᴮ] is warranted.

For a little over half of the tokens of /bᴮ/ and /ᵐbᴮ/, there is no break in voicing between the closure and the trill. Consequently, there was no principled way to measure VOT for these tokens.

For the remaining tokens, voicing attenuated gradually until it ceased completely, typically about 4/5 of the way through the stop closure. In those cases, I was able to measure the delay between release of closure and the commencement of voicing during the first trill period. The results for these measurements are shown in Table 5.

Phoneme	n	VOT (x̄)	s.d.
/pᴮ/	42	23.2 ms	10.1
/bᴮ/	7	17.7 ms	10.5
/ᵐbᴮ/	6	17.4 ms	8.32

Table 5: Voice onset time duration measurements for /pᴮ, bᴮ, ᵐbᴮ/.

4. DISCUSSION

As discussed in Section 3.4, the trill release of /pᴮ/ is nearly always voiced. This not only leads us to transcribe the release with a raised small cap b [ᴮ] to indicate voicing, but it raises a question concerning the practice of calling this phoneme a "voiceless bilabial trill," since the voicelessness occurs during the stop closure, not during the trill release.

It would be hasty to generalize that all cases crosslinguistically of a voiceless stop followed by a bilabial trill release involve a voiced release. Coupe [4] and Rangelov [26] include spectrograms of such phonemes that clearly indicate voicelessness during the trill release. My study indicates that voicing during the trill release is a feature worth examining more carefully in future studies.

There is much room for future work on bilabial trills in Mangbetu. This includes:

- Examining the behavior of /bᴮ/ and /ᵐbᴮ/ before front vowels,
- Comparing the stop closure duration of the three trill phonemes with their non-trill counterparts /p, b, ᵐb/,
- Comparing the behavior of /pᴮ, bᴮ, ᵐbᴮ/ with the Mangbetu coronal trills /tʳ, dʳ, ⁿdʳ, r/,
- Using earbuds to measure more precisely the nasal stop closure of /ᵐbᴮ/ [26, 34],
- Studying more speakers of Mangbetu; Ladefoged [35] recommends recording at least six speakers of each sex,
- Video recordings to examine lip oscillation.

6. ACKNOWLEDGMENTS

Thanks to David Mbiri Koogibho for sharing his language with me and for his friendship, to two anonymous reviewers for helpful comments, to ACATBA for providing recording space, to the Laboratoire Parole et Langage for assistance in archiving the recordings, and to SIL Francophone Africa for funding.

7. REFERENCES

[1] Keating, C. A. 2007. *A Cross-Linguistic Overview of Bilabial Trills*. Graduate Institute of Applied Linguistics M.A. Thesis.

[2] Olson, K. S. 2022. The non-existence of the plain bilabial trill phoneme. *Proc. 96th Ling. Soc. Amer.* Washington, 5239. DOI: 10.3765/plsa.v7i1.5239.

[3] Ladefoged, P., Everett, D. 1996. The status of phonetic rarities. *Lg.* 72, 794–800. DOI: 10.2307/416103.

[4] Coupe, A. R. 2015. Prestopped bilabial trills in Sangtam. *Proc. 18th ICPhS* Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0734.pdf

[5] Kutsch Lojenga, C. 2013. Exotic consonants in Congolese languages. Paper presented at: The Phonetics and Phonology of Sub-Saharan Languages, Johannesburg, 7–10 July 2013.

[6] Parker, K. H. 1985. Baka phonology. *Occasional Papers in the Study of Sudanese Languages* 4, 63–85. https://www.sil.org/resources/archives/35934.

[7] Sampson, D. L. 1997. Update on Baka phonology and orthography, as of 1996. *Occasional Papers in the Study of Sudanese Languages* 7, 114–120. https://www.sil.org/resources/archives/35985.

[8] Persson, J. 2001. The Morokodo language. Ms. SIL International.

[9] Persson, J. 2004. Bongo-Bagirmi languages in Sudan. *Occasional Papers in the Study of Sudanese Languages* 9, 77–84. https://www.sil.org/resources/archives/35968.

[10] de Wit, G. 2010. Lika phonologie. *SIL Electronic Working Papers* 2010-002. https://www.sil.org/resources/archives/7818.

[11] Eberhard, D. M., Simons, G. F., Fennig, C. D. 2023. *Ethnologue*. SIL International. www.ethnologue.com.

[12] Larochette, J. 1958. *Grammaire des dialectes mangbetu et medje suivi d'un manuel de conversation et d'un lexique*. Musée Royal du Congo Belge.

[13] Tucker, A. N., Bryan, M. A. 1966. *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa*. Oxford University Press.

[14] Thomas, J. M.-C., Bouquiaux, L., Cloarec-Heiss, F. 1976. *Initiation à la phonétique: Phonétique articulatoire et phonétique distinctive*. Presses Universitaires de France.

[15] Demolin, D. 1988. Some problems of phonological reconstruction in Central Sudanic. *Belgian J. Linguistics* 3, 53–95. DOI: 10.1075/bjl.3.06dem.

[16] Demolin, D. 1990. Les trilles bilabiales du mangbetu. *Pholia* 5, 67–90. http://www.ddl.cnrs.fr/Download/Pholia/Pholia_N-5.pdf.

[17] Demolin, D. 1991. L'analyse des segments, de la syllabe et des tons dans un jeu de langage mangbetu. *Langages* 25, 30–50. DOI: 10.3406/lgge.1991.1800.

[18] Demolin, D. 1992. *Le mangbetu: Étude phonétique et phonologique*. Université Libre de Bruxelles dissertation.

[19] Demolin, D. 2013. The phonetics of bilabial trills in Mangbetu. Paper presented at: The Phonetics and Phonology of Sub-Saharan Languages, Johannesburg, 7–10 July 2013.

[20] McKee, R. G. 1991. The interpretation of consonants with semi-vowel release in Meje (Zaire) stems. In: Rottland, F., Omondi, L. N. (eds), *Proceedings of the Third Nilo-Saharan Linguistics Colloquium*. H. Buske, 181–195.

[21] McKee, R. G. 2007. Concerning Meegye and Mangbetu's bilabial trills. In: Payne, D., Reh, M. (eds), *Advances in Nilo-Saharan Linguistics. Proceedings of the 8th Nilo-Saharan Linguistics Colloquium, University of Hamburg, August 22–25 2001*. Rüdiger Köppe, 181–189.

[22] Olson, K. S., Mbiri, D. 2013. Labial vibrants in Mangbetu. *Afrikanistik Online 2013*. https://nbn-resolving.de/urn:nbn:de:0009-10-38517.

[23] Olson, K. S., Mbiri, D., Olson, R. E., Kuntz, J., Simons, G. F. 2013. Labial vibrants in Mangbetu: Archival form. *ORTOLANG*. https://hdl.handle.net/11403/sldr000829/v1.

[24] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer [Computer program]. Version 6.2.21. www.praat.org.

[25] Maddieson, I. 1989. Aerodynamic constraints on sound change: The case of bilabial trills. *UCLA Working Papers in Phonetics* 72, 91–115. https://escholarship.org/uc/item/52014440.

[26] Rangelov, T. 2019. The bilabial trills of Ahamb (Vanuatu): Acoustic and articulatory properties. *Proc. 19th ICPhS* Melbourne, 1292–1296. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1341.pdf.

[27] Rangelov, T., Walworth, M., Barbour, J. 2023. A multifaceted approach to understanding unexpected sound change: The bilabial trills of Vanuatu's Malekula Island. *Diachronica*. DOI: 10.1075/dia.21051.ran.

[28] Olson, K. S., Meynadier, Y. 2015. On Medumba bilabial trills and vowels. *Proc. 18th ICPhS* Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0522.pdf

[29] Lisker, L. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Lg.* 33, 42–49. DOI: 10.2307/410949.

[30] Maddieson, I. 1989. Prenasalized stops and speech timing. *J. Int. Phonetic Assn.* 19, 57–66. DOI: 10.1017/S0025100300003856.

[31] Olson, K. S. 2005. *The Phonology of Mono*. SIL International and the University of Texas at Arlington.

[32] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Blackwell.

[33] Yoder, B. 2010. Prenasalization and trilled release of two consonants in Nias. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session 50*. DOI: 10.31356/silwp.vol50.03.

[34] Stewart, J., Kohlberger, M. 2017. Earbuds: A new method for measuring nasality in the field. *Language Documentation & Conservation* 11, 49–80. http://hdl.handle.net/10125/24724.

[35] Ladefoged, P. 2003. *Phonetic Data Analysis*. Blackwell.

THE EFFECT OF STRESS AND SPEECH RATE ON VOWEL QUALITY IN SPONTANEOUS CENTRAL MEXICAN SPANISH

Laura Montoya Hernandez[1], Scott James Perry[2], Benjamin V. Tucker[2,3]

[1]Universidad del Valle, [2]University of Alberta, [3]Northern Arizona University
montoya.laura@correounivalle.edu.co, sperry1@ualberta.ca, benjamin.tucker@nau.edu

ABSTRACT

This study explores the influence of stress and speech rate on vowel quality in spontaneous speech produced by 13 Central Mexican Spanish speakers. Results show that stress more consistently influences F1, and the effect of speech rate is inconsistent, affecting the quality of some vowels and not others. We also find that, for some vowels, the effect of speech rate on vowel centralization is stronger for unstressed vowels than for stressed vowels. Interpreted in the context of overall formant variability, these results show that Spanish vowels undergo phonetic, but not phonological vowel reduction. As our stress results more closely mirror studies on spontaneous Spanish rather than those conducted using laboratory speech, it may be that patterns of phonetic vowel reduction in Spanish are not consistent across speech styles.

Keywords: Spanish, vowel space, spontaneous speech, lexical stress

1. INTRODUCTION

For most varieties of Spanish there is a symmetrical 5-vowel system /i,e,a,o,u/ that is generally assumed to lack phonological vowel reduction in unstressed syllables [1]. Although lexically contrastive vowels may not neutralize in unstressed syllables, studies on many varieties of Spanish have found acoustic differences between stressed and unstressed vowels in terms of duration and vowel quality [2, 3, 4, 5, 6]. Differences in vowel duration are more consistent across the vowel space while vowel quality is sometimes reported to only differ by stress in specific vowels [4, 5]. The primary goal of the present paper is to analyze the spectral qualities of stressed and unstressed vowels in spontaneously produced, conversational speech in Mexican Spanish, building on the modest body of literature investigating phonetic variability in spontaneous Spanish. The second goal is to investigate the effect of speech rate on vowel production and explore how speech rate may or may not interact with stress.

Unstressed vowels in Spanish are shorter than stressed vowels, a robust finding across many studies and varieties [2, 3, 4, 6], even if the difference may be smaller compared to other languages [5]. We did not investigate duration in the present paper due to limitations on boundary placement during forced-alignment [7]. However, we are investigating the effect of speech rate, which we assume impacts vowel quality at least partially due to decreased vowel duration as speech rate increases. As pointed out by Lindblom [8], spectral 'targets' of vowels may not be reached at faster speech rates due to 'articulatory undershoot', as there is less time to articulate sounds. In Spanish, faster speech rates while reading have been associated with increased centralization (F1 and F2), i.e. an overall compression of the vowel space [2].

As mentioned previously, the findings concerning the effect of stress on vowel quality show less consistent patterns across different studies. Nadeu [2] found stress-driven F1 differences for /a/ and F2 differences for /i,a,o/, with stress differences for other vowels not reaching statistical significance. In contrast, Torreira & Ernestus [5] reported that the effect of stress on Spanish vowel quality was largely restricted to movement in F1, with unstressed vowels having lower F1 values. In a study including spontaneous speech from Mexican Spanish speakers [4], differences between stressed and unstressed vowels were larger for F1 than F2, although this was not modelled explicitly. The differences reported above are not surprising, as we expect differences between laboratory and spontaneous speech [7].

While previous findings are somewhat inconsistent, there are many potential reasons for these differences. Some differences are undoubtedly due to dialectal differences, which are attested [4]. Other potential sources are different elicitation methods, and researcher degrees of freedom, as similar studies have employed drastically different approaches to analyzing changes in vowel formants [5, 4]. It is also important to note that some studies claim that vowels in stressed and unstressed vowels

are the same based on a non-significant result, confusing the absence of evidence with evidence of absence. As the differences in vowel duration are fairly robust, from an articulatory perspective, there is likely some difference on average between the quality of stressed and unstressed vowels [8].

2. METHOD

The data reported in this paper was recorded as part of a larger research project. Below, we only describe the methods directly relevant to the present paper, omitting other tasks completed by participants.

2.1. Speakers

The speech analyzed for the present study consists of recordings in Spanish of L1 Spanish/L2 English speakers from Querétaro, México (N=13, 9 females and 4 males).

2.2. Materials and procedure

Participants were recorded in a sound booth while speaking to a relative or friend by phone for 15 minutes wearing a head-mounted microphone. The speech produced by the participants in the booth was orthographically transcribed by a native Spanish speaker. We used these transcripts to generate a pronunciation dictionary using a function of the Montreal Forced Aligner (MFA; [9]).

The dictionary was reviewed by the native speaker and a phonetically-trained non-native speaker of Spanish to check that the pronunciations were accurate and consistent with the dialect under analysis. During the transcription process, pronunciation variants noted were added to the dictionary. We then used the MFA to align the data at the word and phone levels using only our data, as we had more than one hour of speech required to produce desirable alignment.

2.3. Measurement

In order to extract spectral information from vowels in our data, we first determined formant tracker settings for each speaker manually by looking at a small number of tokens of each vowel. Using this information, we extracted the formant measures (at multiple time-points for each vowel), duration, previous/following segments, word, and speech rate using a custom script in Praat [10]. We calculated speech rate for each utterance by taking the number of syllables divided by the duration.

As we expected a certain amount of measurement error due to both formant tracking and the forced alignment process, we elected to use a robust method of calculating multivariate outliers - the minimum covariance determinant. For each speaker, the outliers for each vowel were calculated using raw formant values at each of 30, 40, 50, 60, and 70 percent of the vowels' duration. We then calculated a multivariate normal distribution for each vowel by taking a number of subsamples of the data and determining the ellipsis that minimized the covariance structure. This prevents measurement errors from skewing the ellipsis (as would happen with Mahalanobis distance). We calculated this using `rrcov` (v 1.6-0, [11]), and vowels that fell in the 2.5% tail of the resulting chi-square distribution were removed from our analysis. This removed 22.97% of the data, leaving us with 26,292 tokens. We then removed tokens from utterances that were measured to have speech rates more than three standard deviations below the average speech rate, which left us with 25,925 vowel tokens for our analysis. The distribution of these tokens was as follows: /i/ - 10.5% (8.4% stressed), /e/ - 34.1% (21.9% stressed), /a/ - 29.3% (15.7% stressed), /o/ - 22.1% (10.8% stressed), /u/ - 3.8% (2.7% stressed).

2.4. Statistical Analyses

All statistical analyses were carried out in `R` (v 4.1.1; [12]). The formant normalization process, using the `phonTools` (v 0.2-2.1; [13]) implementation of the Nearey Intrinsic [14] vowel normalization procedure, was done to minimize variation in spectral properties between speakers due to anatomical differences. This method was chosen as it minimized within-category variation compared to other vowel normalization methods.

All models fit to address our research questions were linear mixed-effect models fit using `lme4` (v 1.1-29; [15]). The dependent variables were normalized F1 and F2 values at vowel midpoint. The independent variables were stress, vowel, and speech rate. Stress was a two-level factor (stressed vs. unstressed) treatment-coded into the regression model with 'stressed' as the reference level. Vowel was a five-level factor (a,e,i,o,u) treatment-coded with 'a' as the reference level. Speech rate was the average speech rate of the utterance being produced and was log-transformed before being centered and scaled. All models had varying intercepts for speaker, previous segment, and following segment.

The first set of models explored the effect of stress on formant values at vowel midpoint, and included interactions between stress and vowel, with speech rate as a covariate. Both the model for F1 and F2

included random slopes for stress by participant. We then further explored the effect of speech rate by including a three-way interaction between vowel, stress, and speech rate. The model for F1 included random slopes for stress and speech rate, and the model for F2 included a random slope for stress. Interactions were interpreted with the aid of the package `emmeans` (v 1.6.3; [16]).

3. RESULTS

For the models investigating stress, the interaction between stress and vowel significantly contributed to model fit, as evidenced by likelihood ratio test, for both the model predicting F1 values ($\chi^2(4) = 35.18, p < 0.0001$) and F2 values ($\chi^2(4) = 40.52, p < 0.0001$). Figure 1 plots the average formant values for each vowel by stress condition. Stress did not have a statistically significant effect on F1 of /i/, but all other vowels had significantly lower F1 values in the unstressed position: /a/ ($\hat{\beta} = -0.028, SE = 0.004, p < 0.001$), /e/ ($\hat{\beta} = -0.012, SE = 0.004, p = 0.004$), /o/ ($\hat{\beta} = -0.009, SE = 0.004, p = 0.045$), /u/ ($\hat{\beta} = -0.035, SE = 0.009, p < 0.001$). For F2 values, stress did not significantly affect the F2 for the /e, i, u/ vowels. Only the /o/ vowel ($\hat{\beta} = 0.018, SE = 0.006, p = 0.003$) and the /a/ vowel ($\hat{\beta} = 0.021, SE = 0.006, p < 0.001$) showed evidence of higher F2 values in unstressed vowels.

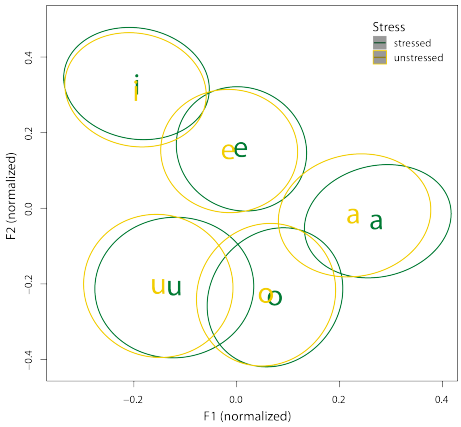


Figure 1: Normalized mean F1 and F2 values for stressed (green) and unstressed (yellow) vowels in Central Mexican Spanish. Ellipses represent one SD.

For the model which further explored the role of speech rate on vowel quality, the three-way interaction between stress, vowel, and speech rate was significant for both the F1 ($\chi^2(4) = 21.409, p < 0.001$) and F2 ($\chi^2(4) = 11.435, p < 0.05$).

First, we checked if speech rate was predictive of F1 values in the context of the vowel:stress interaction. The F1 values of the /e,i,o/ vowels were not significantly predicted by speech rate regardless of stress. However, both stressed /a/ ($\hat{\beta} = -0.016, SE = 0.003, p < 0.001$) and unstressed /a/ ($\hat{\beta} = -0.026, SE = 0.003, p < 0.001$) had significantly lower values of F1 as speech rate increased. For the /u/ vowel, increased speech rate led to higher values of F1 for stressed ($\hat{\beta} = 0.013, SE = 0.005, p = 0.023$) and unstressed ($\hat{\beta} = 0.050, SE = 0.008, p < 0.001$) vowels. The interaction between stress and speech rate reached statistical significance for both the /a/ ($\hat{\beta} = 0.010, SE = 0.003, p = 0.006$) and /u/ ($\hat{\beta} = -0.037, SE = 0.010, p < 0.001$) vowels. The three-way interaction for the model predicting normalized F1 values is visualized in Figure 2.

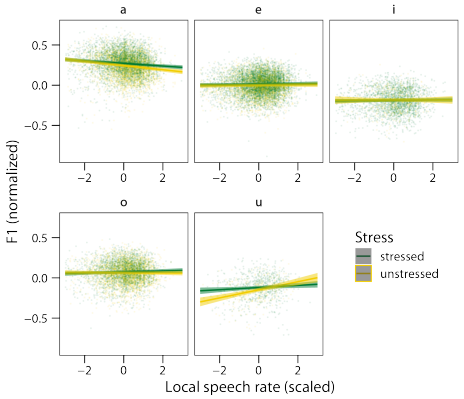


Figure 2: F1 variation of stress (green) and unstressed (yellow) vowels as speech rate increases.

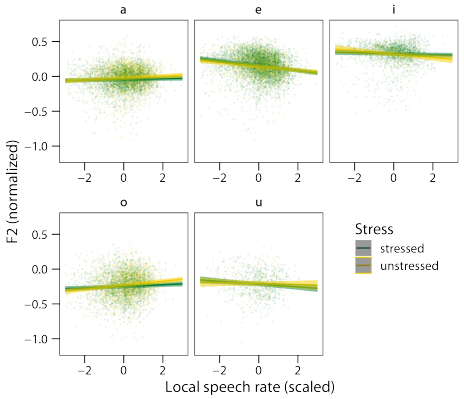


Figure 3: F2 variation of stress (green) and unstressed (yellow) vowels as speech rate increases.

For F2, we find that speech rate is predictive for stressed /e/ ($\hat{\beta} = -0.035, SE = 0.002, p < 0.001$), /i/ ($\hat{\beta} = -0.008, SE = 0.003, p = 0.0297$),

/o/ ($\hat{\beta} = 0.011, SE = 0.003, p < 0.001$), and /u/ ($\hat{\beta} = -0.020, SE = 0.006, p = 0.003$). For unstressed vowels, speech rate was predictive of F2 for /a/ ($\hat{\beta} = 0.013, SE = 0.003, p < 0.001$), /e/ ($\hat{\beta} = -0.028, SE = 0.003, p < 0.001$), /i/ ($\hat{\beta} = -0.024, SE = 0.008, p = 0.003$), and /o/ ($\hat{\beta} = 0.028, SE = 0.003, p < 0.001$). The interaction between stress and speech rate was only statistically significant for the /o/ vowel ($\hat{\beta} = -0.017, SE = 0.004, p < 0.001$). The three-way interaction from the model predicting normalized F2 values is visualized in Figure 3.

4. DISCUSSION

In the current paper, we analyzed a corpus of spontaneous Central Mexican Spanish using automated methods to investigate how stress and speech rate influence vowel quality. We began by analyzing the effect of stress conditional on each vowel to compare these results with previous studies. Our second analysis concerned whether or not the effect of speech rate was the same across stressed and unstressed vowels.

During the initial analysis of the effect of stress, we found that stress is related to changes in F1 in all vowels except for /i/, while stress was only related to increased F2 values for /o/ and /a/. This finding of stress in Spanish being more related to vowel raising rather than centralization is similar to results found for spontaneous Peninsular Spanish by [5] and Mexican Spanish by [4]. Comparing our results concerning F2 with previous research, it seems that changes in F2 by stress are inconsistent across studies, with different vowels having a significant effect of stress for F2 values. Speech rate was only related to changes in F1 values for /a/ and /u/ vowels, with faster speech rates associated with increased centralization. For these two vowels, we also found stronger effects of speech rate for unstressed vowels, which showed more raising with increased speech rate compared to stressed vowels.

For F2, speech rate either had no significant effect on formants or had the predicted effect of leading to more centralized vowels at higher speech rates. The exception is the stressed /u/ vowel, which became more peripheral at higher speech rates. While this is contrary to our predictions, a number of studies have argued that stressed syllables tend to be produced with more articulatory effort [17]. The overall change in the average F2 is small compared to the overall variation in formant values for this vowel. For most vowels, speech rate had a similar effect on F2 for stressed and unstressed vowels, with the exception of /o/, where centralization of unstressed vowels at faster speech rates increased.

While we must be cautious interpreting a statistically insignificant interaction, speech rate seems to have a relatively similar effect on vowel quality in both stressed and unstressed syllables. When there is a difference, it comes in the form of increased centralization for unstressed vowels. This is particularly interesting for the F1 of the /u/ vowel, where stress and speech rate seem to have opposing effects, with unstressed /u/ being raised in contrast to a lowering effect from increased speech rate.

In some cases, our results match previous literature [18] and [2] such as some unstressed vowels like /i,e/ tend to lower their F2 values while the rest seem to increase it. Most formant movement we observe is the lowering of F1 for all the unstressed vowels except /i/, which is the highest vowel on average. It's worth noting here that /u/ having lower average F1 values than /i/ was also reported for other Spanish varieties [2], so in a sense, /u/ had more room to raise than /i/.

We think it is important to acknowledge that we conducted an analysis of vowel quality at the vowel midpoint. We also conducted a non-linear regression on the formant trajectories between 30-70% of vowel duration and found that some vowels did show a pattern of movement over the course of the vowel. These trends were small compared to the overall variation in formant values, so we did not pursue this further due to space constraints. Future research could examine potential vowel inherent spectral changes [19] in Spanish vowels.

5. CONCLUSION

The present study has reported shifts in vowel quality according to stress and speech rate in spontaneous Spanish from Central Mexico. We find that stress has a more consistent effect on F1 than F2, echoing findings from work on spontaneous Spanish from other dialects. We also find that the effect of speech rate is inconsistent across the vowel inventory and that for some vowels there is a stronger effect of speech rate for unstressed vowels. Comparing this study with previous work on Spanish vowels, it seems that the difference between stressed and unstressed vowels may be modulated by speech style, although it is difficult to make these comparisons across different dialects. As such, future research employing a within-speaker design examining the effect of register and elicitation method would be ideal for documenting these differences in Spanish.

6. ACKNOWLEDGMENTS

The data analyzed for this study was part of National Science Foundation grant 1022313 to Natasha Warner and Miquel Simonet. We thank the Mitacs organization for supporting the first author through a Globalink Research Internship and the Social Sciences and Humanities Research Council for supporting the second author in the form of a CGS-D scholarship.

7. REFERENCES

[1] J. I. Hualde, *The Sounds of Spanish*. Cambridge, UK: Cambridge University Press, 2005.

[2] M. Nadeu, "Stress- and speech rate-induced vowel quality variation in Catalan and Spanish," *Journal of Phonetics*, vol. 46, pp. 1–22, 2014.

[3] M. Ortega-Llebaria and P. Prieto, "Acoustic correlates of stress in Central Catalan and Castilian Spanish," *Language and Speech*, vol. 54, no. 1, pp. 73–97, 2011.

[4] F. Santiago and P. Mairano, "The role of lexical stress on vowel space and duration in two varieties of Spanish," in *9th International Conference on Speech Prosody*. Poznan, Poland: ISCA, 2018, pp. 453–457.

[5] F. Torreira and M. Ernestus, "Vowel elision in casual French: the case of vowel /e/ in the word c'etait," *Journal of Phonetics*, vol. 39, pp. 50–58, 2011.

[6] A. M. Delforge, "Unstressed vowel reduction in Andean Spanish," in *Selected proceedings of the 3rd conference on laboratory approahces to Spanish phonology*, L. Colantoni and J. Steele, Eds. Somerville, MA: Cascadilla Proceedings Project, 2008, pp. 107–124.

[7] B. V. Tucker and Y. Mukai, *Spontaneous Speech*, ser. Elements in Phonetics. Cambidge University Press, 2023.

[8] B. Lindblom, "Spectrographic study of vowel reduction," *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of the annual conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[10] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2019. [Online]. Available: www.praat.org

[11] V. Todorov and P. Filzmoser, "An Object-Oriented Framework for Robust Multivariate Analysis," *Journal of Statistical Software*, vol. 32, no. 3, pp. 1–47, 2009.

[12] R. C. Team, "R: A language and environment for statistical computin," Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org/

[13] S. Barreda, "phonTools: Functions for phonetics in R." 2015.

[14] T. M. Nearey, *Phonetic feature systems for vowels*. Indiana University Linguistics Club, 1978.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] R. V. Lenth, "emmeans: Estimated marginal means, aka least-squares means," 2021.

[17] T. Gay, "Physiological and acoustic correlates of perceived stress," *Language and Speech*, vol. 21, no. 4, pp. 347–353, 1978.

[18] R. E. Ronquest, "An acoustic examination of unstressed vowel reduction in heritage Spanish," in *Selected proceedings of the 15th Hispanic linguistics symposium*. Somerville, MA: Cascadilla Proceedings Project, 2013, pp. 151–171.

[19] T. M. Nearey and P. F. Assmann, "Modeling the role of inherent spectral change in vowel identification," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1297–1308, 1986.

MEASURING AND MODELLING THE DURATION OF INTERVOCALIC ALVEOLAR TAPS IN PENINSULAR SPANISH

Scott James Perry[1], Matthew C. Kelley[2], Benjamin V. Tucker[3,1]

[1]University of Alberta, [2]University of Washington, [3]Northern Arizona University
sperry1@ualberta.ca, mattck@uw.edu, benjamin.tucker@nau.edu

ABSTRACT

Factors predicting acoustic variation in the production of Spanish taps have yet to be investigated outside of their relationship to the tap-trill contrast. The present study models the duration of alveolar taps with occlusions visible on a spectrogram in spontaneous Spanish and compares durations measured by automated methods to hand-placed boundaries. We model tap duration from the Nijmegen Corpus of Casual Spanish [1] with a combination of lexical, phonetic, and phonological predictors. Results indicate a high degree of uncertainty regarding the relationship between most of our predictors and tap duration. However, we are confident that faster speech rates are associated with decreased duration. Our automated measurements show deviations from hand-measured duration, indicating a need to evaluate the performance of the automated methods in future research.

Keywords: Spanish, phonetic variation, acoustic measurement, speech production

1. INTRODUCTION

Studies analyzing the duration of Spanish alveolar taps (hereafter taps) have generally focused on comparing productions across speaker groups [2], investigating acoustic correlates of the tap-trill contrast [3, 4] or both [5]. The present study's primary goal is to augment our understanding of variation in tap production. We model tap duration with phonetic, phonological, lexical, and predictability-related factors. Our secondary goal is to compare tap durations based on experimenter markup to three automated methods that measure duration with minimal researcher markup and evaluate the methods' impact on model estimates.

In research on other languages, various predictors have been associated with changes in duration at the segmental, syllabic, and word levels. Increased frequency and predictability have been associated with decreased duration [6, 7]. Duration differences by phonetic factors such as phonetic environment and speech rate have been attested [8, 9]. Lexical stress and pitch accents have also been associated with changes in stop closure duration [10].

When measuring the duration of segments without an apparent onset or offset, which includes many Spanish taps [3, 4], it is desirable to have a measurement method that applies to most realizations. One alternative to human markup includes using boundaries placed by an acoustic model through forced alignment, although this method has some known issues [11]. Another option is to use intensity to measure duration, which can be done in various ways [12, 8, 13]. Before applying these methods to Spanish taps, we believe it is important to compare them to hand measurement for tokens with visual cues to onset and offset, where experimenter markup can be consistent.

2. METHOD

The data and the script documenting the analysis are available through the University of Alberta Education and Research Archive here: https://doi.org/10.7939/r3-5k3f-4t63

2.1. Data coding and measurement

We analyzed a random 10% sample of intervocalic taps from the Nijmegen Corpus of Casual Spanish [1], containing 20 conversations between groups of three university students from Madrid, Spain. Of the 2,606 hand-coded taps, 1,312 had occlusions visible in a spectrogram ('True' or 'Approximant' taps following previous studies [3, 4, 2, 5, 14]).

Each tap's duration was measured in four ways using a script in Praat (v 6.1.47; [15]). The first method was the manual placement of boundaries, carried out by the first author. For stop-like taps, the onset was placed at the beginning of the stop closure, and the offset was placed at the onset of periodic energy after the burst release. For approximant taps, onset and offset were placed where the spectrogram abruptly changed intensity. The second method used force-aligned boundaries from the Montreal Forced

Aligner [16] trained on the corpus under analysis. The third method took the tap onset and offset as the midway points between the maximum intensities of the surrounding vowels and the minimum intensity during the tap [8]. The final method placed the onset and offset of the tap at the largest absolute values of the spline-smoothed intensity slope in and out of the tap, using methods from [12, 13].

When using automated measurements, outliers that are likely measurement errors are removed based on domain knowledge. To fairly compare methods, we removed observations that we judged to have impossible values in any of the measurement methods. Spanish taps have average durations below 50 ms [3, 4], and virtually all taps reported in [5] were under 150 ms. Therefore, we removed nine observations with values of 200 ms or more and seven with negative values. This left a total of 1,296 tokens for statistical analysis.

2.2. Statistical analysis

To analyze tap duration, we fit four hierarchical Bayesian models (one for each measurement method) with lognormal likelihoods using `brms` (v 2.18.0; [17]) in `R` (v 4.1.1; [18]). Priors were weakly-informative in the context of Spanish taps, and their assumptions were assessed through prior predictive simulation. Following best practices [19], we checked our priors' influence on the posterior by using wider priors, which resulted in identical posteriors upon visual inspection.

The population-level predictors for our models appear below. Group-level effects were varying intercepts by speaker and correlated varying slopes for all population-level effects by speaker. Unigram and bigram frequencies from the corpus under analysis were added to counts from the Spanish OpenSubtitles corpus [20], and conditional probabilities derived therefrom. We extracted speech rate (syllables/second), surrounding vowels, and lexical stress from force-aligned text grids. Word length and content/function status were derived from the corpus data file.

The following predictors were used in the models:

- **Unigram freq.** Log unigram frequency for word containing tap
- **Bigram freq. prev.** Log bigram frequency for word and previous word
- **Bigram freq. fol.** Log bigram frequency for word and following word
- **Cond. prob. prev.** Conditional probability of the word based on previous word
- **Cond. prob. fol.** Conditional probability of the word based on following word
- **Number of syllables** Word length in number of syllables (log-transformed)
- **Function word** Sum-coded, difference between function (0.5) vs. content words (-0.5)
- **Local speech rate** Log average speech rate of the utterance containing tap (syllables/s)
- **Prev. vowel** Preceding vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Fol. vowel** Following vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Prev. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) previous vowel
- **Fol. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) following vowel

To measure the similarity between posterior distributions from the manual model to those from the automated methods, we calculated overlap in the population-level posteriors using the `overlap()` function from package `bayestestR` (v 0.11.0; [21]). We entered the overlap values as the dependent variable in a hierarchical Beta regression predicting posterior overlap with the manual model by method. We included varying intercepts for predictors with varying slopes by method.

3. RESULTS

For complete model summaries, the reader is directed to the materials hosted in the repository, which include the saved models for convenience.

3.1. Descriptive comparison of methods

Figure 1 contains visualizations that compare manual duration to the three automated methods. In the top row (A, B, & C) are scatter plots between manual measurements on the y-axis and each automated method on the x-axis. Marginal density distributions are placed along the edge of the plots. All automated methods were weakly correlated with manual measurements. In Figure 1 D, we plot the density distributions of the difference between the manual measurement and the automated methods, subtracting the automated duration from the manual measurement for each token. The red dotted line at zero is where the methods had the same value.

3.2. Factors affecting duration

For the model of manual durations, we report in Table 1 the percentages of each posterior that fell below, within, and above a Region of Practical Equivalence (ROPE). A ROPE establishes a minimum effect size the researchers consider practically different from zero. We chose a ROPE

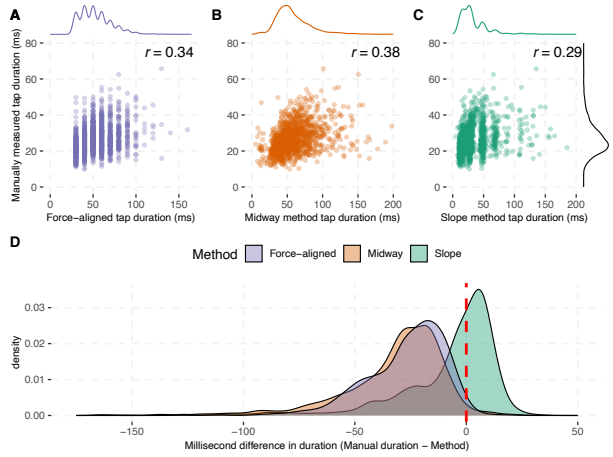


Figure 1: Scatter plots for comparisons between the manually measured durations and the durations taken from forced alignment (A), the midway method (B), and the slope method (C). The density distribution of manual measurements are along the right-hand y-axis of C. Pearson's correlation coefficients are printed on the scatterplot for each comparison. D displays density distributions of by-token differences.

of -0.05 to 0.05 for log-scale duration, which for our model states that if the total effect of a predictor is less than $\approx$ 1.2ms, then we consider the effect to be negligible. This approach allows us to consider the evidence from our model in terms of the probability of both the existence and direction of an effect. An effect below the ROPE is associated with shorter taps, and an effect above the ROPE signals an association with longer taps. For example, our models suggest we cannot be sure if function words contain shorter taps than content words, as roughly 68% of the posterior is within the ROPE, and 32% is below it. We interpret this as a 2/3 chance that there is no difference in tap duration for function words and a 1/3 chance that function words contain shorter taps. The posterior distribution did not extend above the ROPE, meaning we are confident function words don't contain longer taps.

3.3. Modelling differences & posterior overlap

In Figure 2, we plot the population-level posteriors from the models fit to durations from the four methods. The mean of the posterior is plotted by shape, and the two-stage lines visualize the most probable 80% and 95% of each posterior. The posteriors for the same predictor from the four models show variable levels of overlap. For some predictors (e.g., previous bigram frequency), all methods have similar posterior distributions. For others, there are larger differences between the methods, sometimes in a way that could change model interpretation as compared to hand-measured duration or compared to other automated methods.

Table 1: Percentages of posteriors from the model of manually-measured duration that fall below, within, and above the established Region of Practical Equivalence. The ROPE range of (-0.05, 0.05) was divided by the range of continuous predictors to evaluate the total effect size.

Predictor	% Below/Within/Above ROPE
Unigram freq.	2.8 / 26.1 / 71.0
Bigram freq. prev.	56.6 / 42.5 / 0.9
Bigram freq. fol.	17.3 / 76.2 / 6.5
Cond. prob. prev.	70.1 / 7.8 / 22.2
Cond. prob. fol.	91.8 / 7.2 / 1.0
Number of syllables	56.3 / 39.1 / 4.6
Function word	32.4 / 67.6 / 0.0
Local speech rate	98.9 / 1.1 / 0.0
Prev. vowel /i/	0.0 / 10.0 / 90.0
Prev. vowel /a/	4.7 / 95.3 / 0.0
Prev. vowel /o/	23.8 / 76.1 / 0.1
Prev. vowel /u/	22.4 / 73.3 / 4.3
Fol. vowel /i/	0.0 / 0.0 / 100
Fol. vowel /a/	0.7 / 93.0 / 6.3
Fol. vowel /o/	0.0 / 40.2 / 59.7
Fol. vowel /u/	83.1 / 15.7 / 1.3
Prev. stress	14.4 / 83.9 / 1.7
Fol. stress	5.3 / 93.4 / 1.3

The posteriors from the Beta regression estimating overlap as well as group-level standard deviations among the predictors are plotted by method in Figure 3. We cannot be confident that any automated method overlaps more or less with our manual model, although the Slope method had the highest estimated overlap (Figure 3 A). We are confident the Slope method showed less variability by predictor than the other methods (Figure 3 B).

4. DISCUSSION & CONCLUSIONS

The present study measured tap durations in conversational Spanish and modelled this duration with several predictors. As many taps lack visible occlusions, we wanted to evaluate alternative methods of measuring duration that could be applied to more variable realizations. In predicting manually-measured taps, most predictors are highly uncertain regarding the presence or direction of an effect, although we can rule certain patterns out. Measurements from all automated methods correlated similarly with manual measurements, and the Slope method had the lowest absolute error. The force-aligned durations were multiples of 10ms with a floor at 30ms. The midway

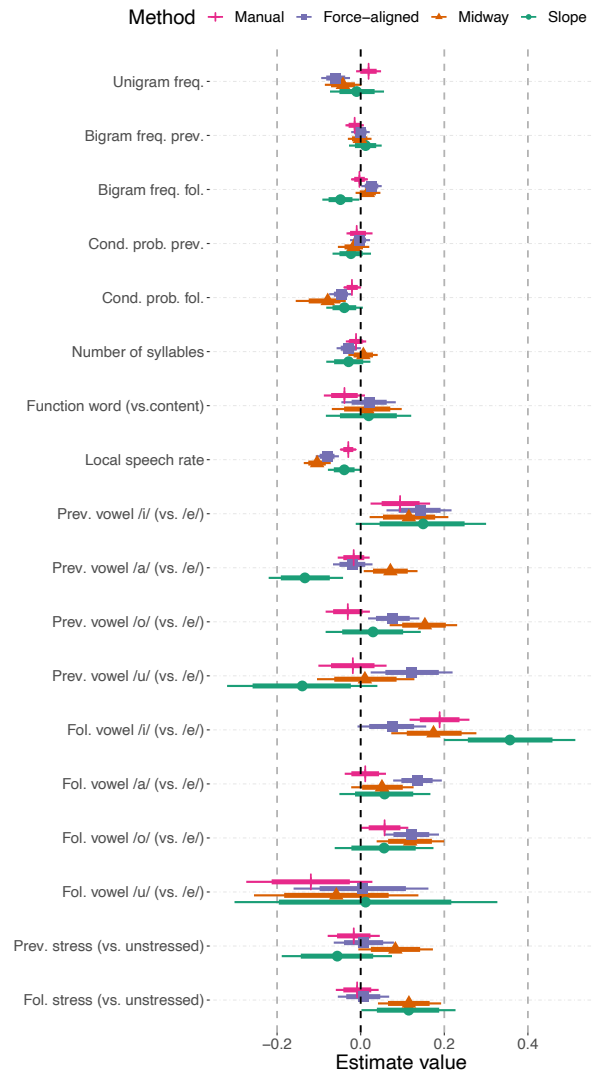


Figure 2: Posteriors for population-level effects models of tap duration calculated by four methods: manually-placed boundaries, force-aligned boundaries, using the midway points of the intensity curve, and using the intensity slope. The thick line corresponds to 80% of the posterior, and the extended, thin line corresponds to 95%.

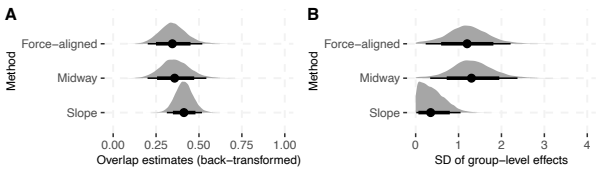


Figure 3: Posterior distributions of estimated overlap with model estimates from manual durations. The thick lines correspond to 80% of the posterior and the extended thin line 95%. 'A' plots predicted overlap, zero meaning posteriors do not overlap and one meaning posteriors are identical. 'B' plots the standard deviations of group-level effects (variation across predictors).

method had a similarly-shaped distribution to manual measurements but with longer values, while the Slope method had clusters at multiple modes.

In our model of manually-measured tap duration, we can only generalize results to taps that have a visible presence in a spectrogram. We only claim with confidence that increased speech rate is associated with shorter taps, and that taps are longer before /i/ than before /e/. Other specific effects show uncertainty regarding the presence of a meaningful association, but effects in specific directions are incompatible with our model. For some effects, large percentages of the posterior fall within the ROPE, indicating the most likely interpretation should be that these predictors are not associated with changes in tap duration.

When comparing estimates from our four models, the patterns of similarity and difference varied depending on the predictor. For some, like speech rate, the effects from all models are reliably negative, but some methods overestimate the effect's size, which is likely due to overestimating tap duration overall. If we consider our manual model the gold standard, other methods make both Type S errors (getting the direction wrong) and Type M errors (wrong effect magnitude). This assumption is reasonable, as retrodictive checks showed the manual model fit the data well, while other models did not. Several predictors with posteriors centered around zero in the manual model show reliably negative or positive effects in one or more automated methods (e.g., Bigram freq. fol, Fol. stress). A potential explanation is that some variables are related to intensity changes independent of duration.

From our model of posterior overlap, all automated methods had less than 60% overlap with the manual model for an average predictor. This result is not reassuring, although the slope method, which may have slightly more overlap, also showed less variation across our predictors, possibly due to having wider posteriors than the other methods. Based on these results, we cannot recommend these automated methods be used to measure Spanish taps. We also must question the reliability of measuring segment duration using intensity more generally, and recommend researchers evaluate their measurement methods as standard practice.

Hand-correcting data is costly, but building knowledge on results skewed by measurement error will be more so. Sharing hand-corrected data publicly will allow for data to be used by the wider research community to answer research questions and generate informative priors that allow them to use their data more efficiently.

5. ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada in the form of a CGS-D scholarship awarded to the first author.

6. REFERENCES

[1] F. Torreira and M. Ernestus, "The Nijmegen corpus of casual Spanish," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Paris: European Language Resources Association (ELRA), 2010, pp. 2981–2985.

[2] N. C. Henriksen, "Acoustic analysis of the rhotic contrast in Chicagoland Spanish: An intergenerational study," *Linguistic Approaches to Bilingualism*, vol. 5, no. 3, pp. 285–321, 2015.

[3] E. Willis and T. Bradley, "Contrast maintenance of taps and trills in Dominican Spanish: Data and analysis," in *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology*, L. Colantoni and J. Steele, Eds. Somerville, MA: Cascadilla Proceedings Project, 2008, pp. 87–100.

[4] T. G. Bradley and E. W. Willis, "Rhotic variation and contrast in Veracruz Mexican Spanish," *Estudios de fonética experimental*, no. 21, pp. 43–74, 2012.

[5] M. Amengual, "Acoustic correlates of the Spanish tap-trill contrast: Heritage and L2 Spanish speakers," *Heritage Language Journal*, vol. 13, no. 2, pp. 88–112, 2016.

[6] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[7] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[8] N. Warner and B. V. Tucker, "Phonetic variability of stops and flaps in spontaneous and careful speech," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1606–1617, 2011.

[9] U. Cohen Priva, "Informativity affects consonant duration and deletion rates," *Laboratory Phonology*, vol. 6, no. 2, pp. 243–278, 2015.

[10] T. Cho and J. M. McQueen, "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," *Journal of Phonetics*, vol. 33, no. 2, pp. 121–157, 2005.

[11] B. V. Tucker and Y. Mukai, *Spontaneous Speech*, ser. Elements in Phonetics. Cambidge University Press, 2023.

[12] J. Kingston, "Lenition," in *Proceedings of the 3rd conference on laboratory approaches to Spanish phonology*, L. Colantoni and J. Steele, Eds. Somerville, MA: Cascadilla, 2008, pp. 1–31.

[13] J. Katz and G. Pitzanti, "The phonetics and phonology of lenition: A Campidanese Sardinian case study," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 10, no. 1, pp. 1–40, 2019.

[14] J. Y. Kim and G. Repiso-Puigdelliura, "Deconstructing heritage language dominance: Effects of proficiency, use, and input on heritage speakers production of the Spanish alveolar tap," *Phonetica*, vol. 77, no. 1, pp. 55–80, 2020.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." 2022. [Online]. Available: http://www.praat.org/

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of the annual conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[17] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[18] R. C. Team, "R: A language and environment for statistical computing," Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org/

[19] J. K. Kruschke, "Bayesian analysis reporting guidelines," *Nature Human Behaviour*, vol. 5, no. 10, pp. 1282–1291, 2021.

[20] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 923–929.

[21] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework." *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019.

IMPROVED FORMANT FREQUENCY MEASUREMENTS IN PRAAT

David Weenink

University of Amsterdam
David.Weenink@uva.nl

ABSTRACT

We present a way for calculating and visualising different formant frequency measurements of the same sound. This will enable a researcher to easily pick, for any interval in the sound, a desired analysis from the available alternatives. To achieve this, we first calculate the alternatives by performing multiple formant frequency analyses of the same sound with different formant ceilings. By varying the ceiling we can adapt the frequency interval of the spectrum to the fixed number of formant peaks that are used to model it. The results of these multiple formant analyses are collected in a FormantPath data structure. A FormantPathEditor makes it possible to view and edit the available formant options. Optionally a corresponding TextGrid can also be shown in the FormantPathEditor. Both the FormantPath as well as the FormantPathEditor were recently added to the sound processing program Praat[1].

Keywords: Formant frequency measurement, variable ceiling, visualisation, trajectory smoothness.

1. INTRODUCTION

Formant frequency measurements have a long tradition in phonetic research and are considered to be important for determining vowel quality. However, formant frequency values are notoriously difficult to measure automatically. Most of the time they are measured by performing some type of linear predictive analysis (lpc), sometimes accompanied by manual intervention.

The *source-filter* model of speech production on which lpc is based assumes that the speech signal is the result of filtering a source signal by the filter characteristics of the vocal tract. The number of filters to model the vocal tract is fixed during the lpc analysis and has to be chosen beforehand. Reasonably good results are obtained when the model and the data fit well. If model and data do not fit well, poor formant frequency estimates may result. For example, if the number of filters is chosen larger then the actual number of formants present in the sound, spurious formant tracks will appear. On the other hand, if the number of filters was chosen too low, then some formants might not be measured at all, or, two or more formants might be averaged which would result in formant frequency estimates at the wrong frequency. Many of the deficiencies of lpc analysis have been known for a long time, see for example [2] or more recently [3]. Often after the measurements a lot of post-processing has to be done. For example, [4] mention their tedious work to measure the acoustic characteristics of 12 American English vowels in h-V-d syllables spoken by 45 men, 48 women and 46 children. They used various linear predictive analyses with different number of coefficients to estimate the formants and very often based their final decision on the steadiness of the formant contour and sometimes also on the experimenter's acoustical phonetic knowledge about the identity of a particular vowel. In another example [5], the authors mention that in their estimation of vowel formant frequencies of Northern and Southern Dutch they had to manually adjust the results in 20 to 25% of the cases. Many more examples could be cited. They all show that a fully automatic perfect formant frequency measurement does not seem to exist and therefore different ad hoc procedures had to be used to improve upon the results obtained by automatic measurements.

In this article we build on a promising lpc-based method that was presented in [6] to overcome some of the deficiencies of standard linear predictive formant frequency analysis, notably the possible mismatch between the number of analysis filters, i.e. the number of lpc coefficients, and the number of formants present in the sound. Essentially our method, instead of performing one analysis with fixed parameters on a sound, performs a multiple of analyses each with a different analysis parameter setting. For each of these analyses a kind of stress value is calculated which can function as an overall smoothness measure of the formant tracks. The analysis parameter that we vary between the different analyses is the so called *formant ceiling* which is the maximum frequency of the formant search range. Because a formant can be

associated with a peak in the spectrum, lpc analysis is equivalent to the matching of a fixed number of peaks to the frequency spectrum of a sound. Because we cannot know beforehand the optimal width of the frequency interval on which a fixed number of formant peaks fits best, we just have to use multiple interval widths. This is equivalent to the use of multiple ceilings in the formant analysis. We have named the data structure that contains all the different formant analyses a *FormantPath*. The visualisation and manipulation of a FormantPath is taken care of in a *FormantPathEditor*.

In the following sections we describe the FormantPath and the FormantPathEditor in somewhat more detail and we show an example of its use.

2. THE FORMANTPATH

A FormantPath object in Praat maintains a path through a collection of Formant objects, where each Formant is the result of a formant frequency analysis of the same sound but with a different setting of the formant ceiling parameter. These Formant objects all have the same time domain, the same time at the first frame, the same time step and the same number of frames.

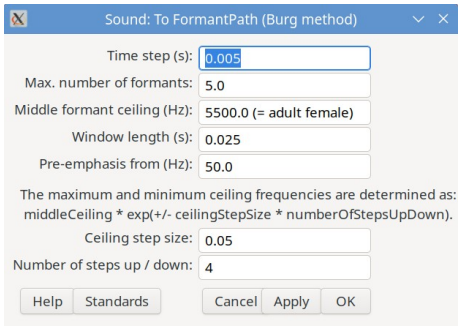


Figure 1: The form to obtain a FormantPath object from a Sound object.

To obtain a FormantPath object from a selected sound we use the **To FormantPath (burg)...** command. Fig. 1 shows the associated form. The first five entries in the form are almost equivalent to the entries in the default **To Formant (burg)** analysis. Two parameters, *Ceiling step size* and *Number of steps up / down* are extra. The last parameter determines how many analysis will be done. For the default number in the form above, 4, there will be 4 analyses with a ceiling higher than the *Middle formant ceiling*, 4 analyses with a ceiling lower than the *Middle formant ceiling* and one analysis with exactly this ceiling, making nine analyses in total. The *Ceiling step size* defines how much the increase or decrease in formant ceiling frequency between two successive analyses will be. The ceiling frequency for the kth step up will be calculated as $middleFormantCeiling.\exp(k \cdot ceilingStepSize)$.

For the form given above which shows the default middle frequency of 5500 Hz for a female speaker, a ceiling step size of 0.05 and 4 steps up and down, this would result in the following rounded nine values for the formant ceiling in hertz: 4503, 4734, 4977, 5232, 5500, 5782, 6078, 6390 and 6718. The top ceiling for a female voice was calculated as $5500\exp(4 \cdot 0.05)$ and this resulted in 6718 Hz; for a male speaker with a middle ceiling of 5000 Hz its top ceiling would be $5000\exp(4 \cdot 0.05)$ which rounds to 6107 Hz. If you would want fifteen analyses instead of the previous nine but with approximately the same top and bottom ceilings you calculate the desired step size by solving $\exp(7 \cdot stepSize) = \exp(4 \cdot 0.05)$ for *stepSize* which gives $stepSize = \frac{4}{7} \cdot 0.05$.

What effect do these different ceilings have on the analyses? Suppose we have a sound with a sampling frequency of 44100 Hz. In the formant analysis, starting with the lowest ceiling, 4503.0 Hz, the sound will first be downsampled to the new sampling frequency of 9006 Hz (two times the ceiling). Pre-emphasis takes place, after which the Burg method of lpc analysis with ten coefficients will be performed and formant frequencies and bandwidths calculated and stored in a Formant object. This Formant object will be collected in the FormantPath object. Next this whole procedure is repeated but now with a 4734 Hz ceiling. This means first downsampling of the original sound to a 9467.8 Hz sampling frequency and then applying the other steps we just described. This will go on until we have collected nine Formant object in the FormantPath.

Since we have guaranteed that all Formants have the same sampling this implies that for each frame number we have nine different analysis frames to our disposal. The FormantPath data structure offers the possibility to assign to each frame number one of the nine available analysis frames. A simple *path interval tier* handles this association. In the default case there is only one interval that has the index of the middle ceiling in the Formant collection as its label (which happens to be "5" here. Following the label indices in this tier would constitute a *path* in time along possibly different Formant objects.

3. THE FORMANTPATHEDITOR

The FormantPathEditor enables the user to view and edit the path taken along the different Formant objects. The editor appears when we select the Sound, the FormantPath and an optional TextGrid together and choose **View&Edit** from Praat's dynamic menu. We use the publicly available Hillenbrand et al.[4] corpus with carefully measured formant frequency values, the original sound files as well as listener evaluations of the vowels, to demonstrate this editor's use.

Fig. 2 shows the FormantPath of the /had/ sound, spoken by female speaker 25 (file w25ah). The FormantPath was calculated with the default values of Fig. 1. The left part of the editor shows the sound oscillogram with a selection on top, the sound's broad-band spectrogram that is overlaid with formant tracks and a TextGrid. The view range of the spectrogram was set to 6800 Hz which is just a little bit higher than the top ceiling used in the analysis. The TextGrid was produced by a simple Praat script that converted all corpus data into TextGrids with one interval tier and one point tier. The interval tier has an interval labeled with the vowel /a/. The two time points in the point tier correspond to the most stable position in the vowel according to two different judges that Hillenbrand et al. used. The figure also shows that the vowel interval has been selected because it is masked by a rectangular shaded area.

The right hand side of the editor is what makes it very different from other editors. The nine panels in this *selection viewer* show the formant tracks from each of the nine different Formant objects in the FormantPath object. In each panel the tracks with even numbers, F_2 and F_4, are drawn in a somewhat different colour than the uneven ones, F_1, F_3 and F_5, to be better distinguishable. In all the panels time runs horizontally from left to right starting at time 0.151 s and ending at 0.581 s; these times correspond to the start and ending time of the selected part of the sound on the left. The vertical frequency scale is linear and tops at 6800 Hz, just like the scale of the spectrogram. Above each panel its ceiling frequency is displayed. The top-left panel shows the lowest ceiling, 4503 Hz, while the bottom-right panel shows the highest ceiling, 6718 Hz. Each frame's formant frequency is displayed as a dot while its bandwidth is drawn as a vertical line centered at its frequency. This might help us in getting a better impression of the analysis results since well defined formants have small bandwidths and, therefore, show shorter vertical lines.

Fig. 2 hints that the panel in the middle has been selected because it shows a different background shading. The panel selection can be changed by clicking with the mouse in any of the other eight panels. This panel selection has some consequences. In the first place a new interval is created in the FormantPath's path tier with the corresponding start and end times and with the selected index as a label. The formant dots displayed on the spectrogram always correspond to the formants of the Formant object at the corresponding index in the FormantPath. This means that selecting one of the panels will also automatically update the dots displayed on the spectrogram. As a second consequence, the selected ceiling of an interval will be shown in the spectrogram area as a fat horizontal line with the ceiling frequency above it. The fact that in the figure there is just one straight horizontal line at 5500 Hz means that we didn't make any selection yet that differed from the default.

Fig. 2 also shows that the default analysis for this particular sound might not be the optimal choice: although the first formant track displayed in the spectrogram area seems to be a rather smooth, some values deviate clearly. Closer inspection reveals that most deviates have a frequency lower than the value of the point in the track. From the middle panel we also note that these deviates have extreme large bandwidths as is shown by the long vertical lines centered at their frequency values. Actually it also shows that the dot in the "first formant track" corresponds to the second formant in this deviate frame. It would be better to completely ignore these deviates. The more because they influence the formant frequency statistics of the interval. If only one value were necessary to represent the interval then taking the median of the values on the interval instead of the mean would be necessary to cure for these outliers. However, if we are interested in the more dynamic aspects of formants, problems might occur.

Inspection of the other panels shows that better alternatives exist. For example, the panel with a ceiling of 6078 Hz at the bottom left. The tracks are much smoother than the ones in the panel at the middle. The start of the interval shows some irregularities because of the influence of the preceding /h/ sound. At reference point 1 the formant values with the 6078 Hz ceiling would be $F_1 = 1000$ Hz and $F_2 = 1580$ Hz. Our values are not to far off from the $F_1 = 1053$ Hz and $F_2 = 1677$ Hz as reported by Hillenbrand et al. Approximately 50 Hz off for the first formant and 100 Hz off for the second

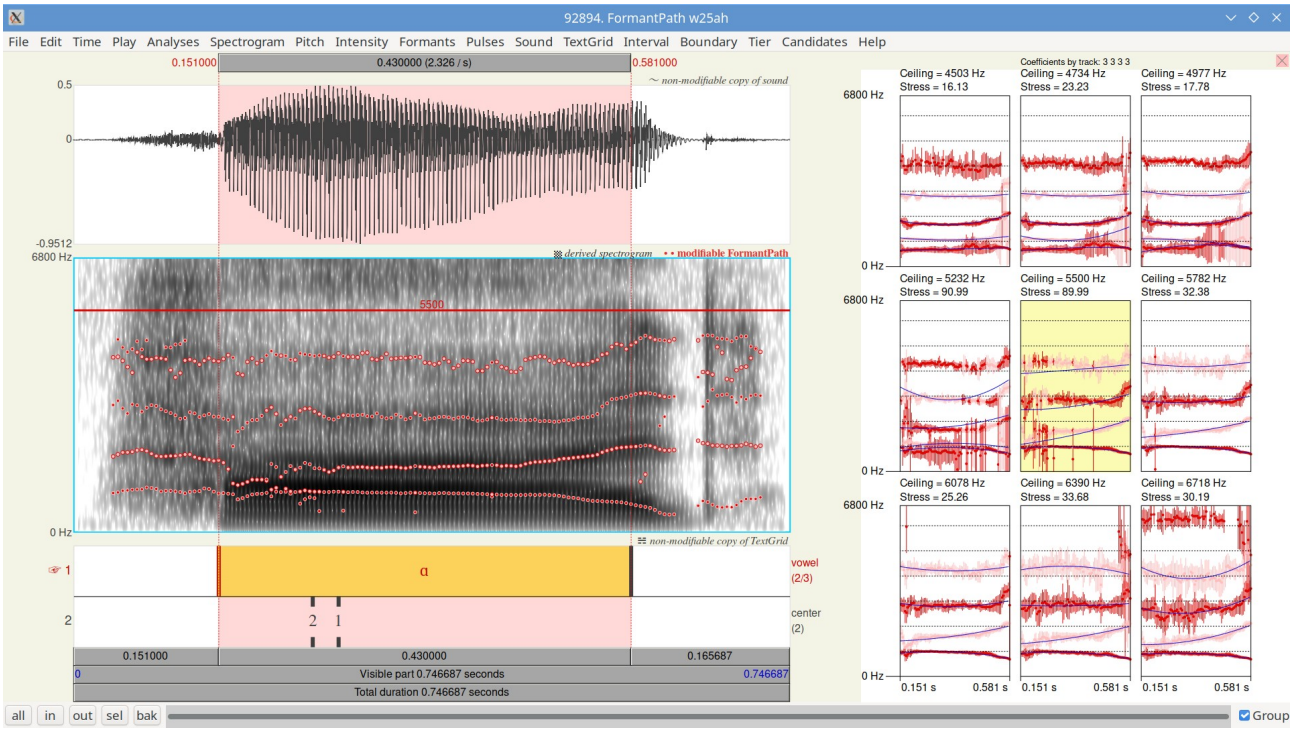


Figure 2: The FormantPathEditor showing the analyses for the /a/ sound of file w25ah from the Hillenbrand corpus.

is not too bad with an lpc-based analysis.

The panel at the top-left shows an interesting example of a formant split: too many peaks have to be fitted on too small a frequency interval. Although its stress value calculated from [6] is the lowest of the nine panels its first three formant frequency values are not realistic for a female /a/ sound.

4. AFTER EDITING

After leaving the editor the FormantPath object contains the changes in the path that have been made. The formant frequency values that would result from following the path can be accessed by using the **Extract Formant** method on the selected FormantPath. A new Formant object will be created and it can be queried just like any other Formant object.

5. SUMMARY

We have described only some of the possibilities of a new method to substantially improve the formant frequency analysis of a sound. However, the new method cannot cope with one of the fundamental limits of every analysis based on linear prediction: the attraction to strong harmonics in the speech spectrum. Despite this deficiency, linear predictive analysis still is a very good practice to get reasonably accurate formant frequency measurements in a fast and reproducible way as this article tried to demonstrate.

6. REFERENCES

[1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," Version 6.3.03, 2023, available from http://www.praat.org/.

[2] J. D. Markel and A. H. Gray, Jr., *Linear prediction of speech*. Springer Verlag, Berlin, 1976.

[3] D. Whalen, W.-R. Chen, C. Shadle, and S. A. Fulop, "Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986)," *Journal of the Acoustical Society of America*, vol. 152, pp. 933–941, 2022.

[4] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[5] P. Adank, R. Van Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *Journal of the Acoustical Society of America*, vol. 116, pp. 1729–1738, 2004.

[6] D. Weenink, "Improved formant frequency measurements of short segments," *Proceedings of the ICPhS XVII 2015*, 2015.

Constructing a short reading passage with phonemic and phonetic coverage for Thai: Traimit passage

Chutamanee Onsuwan[1,2], Kanwara Bairerkdee[1], Juthatip Duangmal[1], Nawasri Chonmahatrakul[1], Atita Amornlaksananon[3], Pimchanok Denmuenwong [1]

[1] Department of Linguistics, Thammasat University, Thailand
[2] Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand
[3]Faculty of Liberal Arts, Rajamangala University of Technology Phra Nakhon, Thailand
consuwan@tu.ac.th

ABSTRACT

This work describes main characteristics of an alternative Thai reading passage, Traimit, as compared with existing passages (i.e., the translated North Wind and the Sun). Traimit was designed to meet a set of criteria which could be beneficial for Thai linguistics and speech research and possibly a potential tool for quick speech assessment. The 102-word passage is short, contemporary, and easily read. It contains a complete Thai phonemic inventory (including clusters and diphthongs) and takes into account five lexical tones in various rhyme units. There are mono- and multisyllabic words as well as utterances/sentences in various length. Traimit has neutral contents with more than 40% of words being "high frequency" and a small set of words repeated to allow for observation of production consistency. Measurements of a subset of vowels from two Thai adults reading the passage revealed its promise of extracting the vowel phonemes from the running speech.

Keywords: Thai, phoneme, vowel, reading passage, speech data

1. BACKGROUND

Standard reading passages/texts (in various languages) have proven to be useful materials for speech science researchers and clinicians as they provide quick and structured samples of the individual's speech and characteristics [1]. Situated within natural reading contexts, they were believed by some to be "closet to spontaneous speech compared to isolated words or sentences" [1].

While some passages were constructed for a general research purpose such as the (English) North Wind and the Sun [2][3], others have been designed and tailored for specific tasks such as the (English) Caterpillar passage for assessment of motor speech disorders [4]. Although a great variety of passages and descriptions/comparisons of their features have existed in many languages, it is only recently that general consensus guidelines have been proposed [1].

Passage features that have often been discussed are phonemic inventory, inclusion of clusters, presence of vowel-to-vowel coarticulation, phonetic balance, passage length and topic, mean length of utterance, word familiarity, reading level, prosodic variations, word length and (syllable) complexity, and word repetition [1][4]. Apart from phonemic coverage (i.e., covering all phonemes), allophonic coverage (i.e., covering many positional contexts—word positions, lexical stress, and proximity to other phonemes) has been analysed and quantified [5].

Following ratings and selections by experts in the international Delphi consensus study, a minimal set of criteria for standard reading passages for an overall speech and voice assessment in adolescents and adults was put forward [1]. The criteria (ranked from highest to lowest) include (1) complete phonemic inventory (2) phonetic balance (3) inclusion of various intonation patterns (4) control of the phonetic complexity of words (5) covering the positions of the phonemes in the words (as well as their vowel/ consonant context) (6) inclusion of consonant clusters (7) repeated presence of some words (8) inclusion of various sentence lengths [1].

Various forms of short reading passages have been around for English and Indo-European languages [1][4][5]. To our knowledge, a few reading passages (see Section 2) have been available for Thai, one for a general research propose and the other two for a small clinical survey. Traimit passage was designed mainly for speech and linguistics research practice. Our goal was to create a relatively short, simple, contemporary and well-structured text, with neutral contents and natural sounding. It should contain all phonemes and capture different rhyme units to allow for study of tonal production/variation in checked and unchecked syllables—one of important aspects of Thai [6][7].

With further evaluations and modification, it may become an effective tool for both clinical and research purposes. Nevertheless, it should not be used

as a reading proficiency test and certainly not recommended for illiterate populations or young children.

2. THAI READIND PASSAGES

Four short reading passages in Thai script along with broad phonemic transcription (according to the conventions given in [8]) are provided (hence no reflections of allophonic variations). Only English translation of the Traimit passage is given. As Thai writing system is usually written with no space between words and a space is often (but not always) used to indicate an end of a phrase, a sentence, or between separate items in a list, here a " | " indicates a space in typed written passage and often corresponds to a break or pause between words and utterances/ sentences in real production. Importantly, main characteristics and limitations of each passage are discussed.

2.1. North Wind and the Sun passage translated to Thai (NW-Th) [8]

The (original) North Wind and the Sun passage is a well-known English standard text (113 words) which has allowed a great number of researchers to compare and describe pronunciation of different English varieties despite absence of some English phonemes [2][3]. In JIPA Illustrations, the International Phonetic Association has encouraged researchers from different languages to capture and document language phoneme diversity using the translated versions of North Wind and the Sun passage [2][3]. For each language in the JIPA Illustrations, there are descriptions of sound inventory and sets of minimal pairs followed by the translated North Wind and the Sun passage, which could be considered as a supplement [8]. However, there are many shortcomings with the Thai version (NW-Th) as in many translated versions in other languages [2]. The NW-Th has 118 words and 172 syllables. Since there are high repetitions of words (26 words get repeated, some more than three times), it fails to capture 14 Thai phonemes (see Section 3).

Thai script: ขณะที่ลมเหนือและพระอาทิตย์กำลังเถียงกันว่าใครจะมีพลังมากกว่ากัน|ก็มีนักเดินทางผู้หนึ่งเดินผ่านมา|ใส่เสื้อกันหนาว|ลมเหนือและพระอาทิตย์จึงตกลงกันว่า|ใครที่สามารถทำให้นักเดินทางผู้นี้ถอดเสื้อกันหนาวออกได้สำเร็จก่อน|จะถือว่าเป็นผู้ที่มีพลังมากกว่า|และแล้ว|ลมเหนือก็กระพือพัดอย่างสุดแรง|แต่ยิ่งพัดแรงมากขึ้นเพียงใด|นักเดินทางก็ยิ่งดึงเสื้อกันหนาวให้กระชับกับตัวมากขึ้นเพียงนั้น|และในที่สุดลมเหนือก็เลิกล้มความพยายาม|จากนั้นพระอาทิตย์จึงสาดแสงอันร้อนแรงออกมา|นักเดินทางก็ถอดเสื้อกันหนาวออกทันที|ในที่สุดลมเหนือจึงจำต้องยอมรับว่าพระอาทิตย์มีพลังมากกว่าตน

(A " | " corresponds to a space in the typed passage in Thai script (p. 27) of [8])

Transcription:[kʰàʔ.nàʔ.tʰîː.lom.nŭa.lɛ́ʔ.pʰráʔ.ʔaː.tʰít.kam.laŋ.tʰiăŋ.kan.wâː.kʰraj.tɕàʔ.miː.pʰáʔ.laŋ.mâːk.kwàː.kan|kɔ̀ː.miː.nák.kʰon.súk.son.lêːn.sɔ̀ː.ŋ.hăː|nŭː.tɕɛ̆w.tòk.tôn.máː.j.rɔ́ː.ŋ.hâː.j.ŋɔ̀ː.ŋɛː|jaː.j.tɕʰím.wâː.jêː.tɕàp.kôn.kòp.jaŋ.mâj.pʰaŋ]

2.2. Rama-I (RM-I) and Rama-II passages (RM-II) [9]

The RM-I and RM-II passages were constructed for a different purpose from the NW-Th. They were designed for an investigation of speech defects in Thai school children with more focus on consonants than vowels [9]. Many vowel phonemes were not taken into account even when the two passages were combined (see Section 3). The RM-I put an emphasis on initial and final consonants whereas the RM-II on the initial cluster set (see Section 3) [9]. The RM-I contains 64 words and 100 syllables.

Thai script: ฝนฟ้ามีวี่แววว่าจะตก|ยายฉิมชวนหนูแจ๋วแจวเรือไปหากำนันแฉ่งที่บางบ่อ|หนูแจ๋วพอกหน้าปะแป้งด้วยดินสอพองดูงามดี|พอถึงที่ฝูงหมาติดตามต้อนหน้าต้อนหลัง|ป้าเยื้อนเอะอะเอ็ดตะโรแล้วยิ้มเบิกบาน|เจ้าแก้วหลานรักรื่มาหักทาย|ผู้ใหญ่พูดคุยหัวเราะขบขัน|สองคนซุกซนเล่นซ่อนหา|หนูแจ๋วตกต้นไม้ร้องไห้งอแง|ยายฉิมว่าแย่จับก้นกบยังไม่พัง

Transcription:[fŏn.fáː.miː.wîː.wɛː.wâː.tɕàʔ.tòk|jaː.j.tɕʰĭm.tɕʰuan.nŭː.tɕɛ̆w.tɕɛː.w.ruaˑ.paj.hăː.kam.nan.tɕʰèŋ.tʰîː.baː.ŋ.bɔ̀ː|nŭː.tɕɛ̆w.pʰɔ̂ː.k.nâː.pàʔ.pêː.ŋ.dûaj.din.sɔ̌ː.pʰɔː.ŋ.duː.ŋaː.m.diː|pʰɔː.tʰŭŋ.tʰîː.fŭː.ŋ.măː.tìt.taː.m.tôː.n.nâː.tôː.n.lăŋ|pâː.juan.ʔèʔ.ʔàʔ.ʔèt.tàʔ.roː.lɛ́ː.w.jím.bɤ̀ː.k.baː.n|tɕâw.kɛ̂ː.w.lăː.n.rák.rîː.maː.tʰák.tʰaː.j|pʰûː.jàj.pʰûː.t.kʰuj.hŭa.rɔ́ʔ.kʰòp.kʰăn|sɔ̌ː.ŋ.kʰon.súk.son.lêːn.sɔ̂ː.n.hăː|nŭː.tɕɛ̆w.tòk.tôn.máː.j.rɔ́ː.ŋ.hâː.j.ŋɔː.ŋɛː|jaː.j.tɕʰím.wâː.jêː.tɕàp.kôn.kòp.jaŋ.mâj.pʰaŋ]

The RM-II has 61 words and 72 syllables.

Thai script: หนูแจ๋วกินกล้วยเพลิน|เผลอถูกเพื่อนผลักลงคลองกลางกรุงดิ้นขลุกขลัก|พระไม่เห็นใครก็คว้าแขนไว้มับ|ครูปลอบแจ๋ว|แจ๋วกราบพระ|ครูตรวจตราพบบาดแผลกว้างกว่านิ้ว|จึงใช้ผ้าโปร่งปรับเป็นผ้าพันแผล|แล้วส่งหมออนามัยกลางตรอก|แจ๋วไม่เจ็บตัวเปล่า|คราวนี้จำเลยขรึมโดนกวดร้องฮักฮัก

Transcription:[nŭː.tɕɛ̆w.kin.klûaj.pʰlɤ̌ː.nˑ|pʰlɤ̌ː.tʰùː.kˑ.pʰûan.pʰlàk.loŋ.kʰlɔː.ŋ.klaː.ŋ.kruŋ.dîn.kʰlùk.kʰlàk|pʰráʔ.mâj.hĕn.kʰraj.kɔ̂ː.kʰwáː.kʰĕː.n.wáj.máp|kʰruː.plɔ̀ː.p.tɕɛ̆w|tɕɛ̆w.kràː.p.pʰráʔ|kʰruː.t rùat.traː.pʰóp.bàː.t.pʰlĕː.kwâː.ŋ.kwàː.níw|tɕɯŋ.tɕʰáj.pʰâː.prò̀ː.ŋ.pràp.pen.pʰâː.pʰan.pʰlĕː.|lɛ́ː.w.sòŋ.mɔ̀ː.ʔàʔ.naː.maj.klaː.ŋ.trɔ̀ː.k|tɕɛ̆w.mâj.tɕèp.tua.plàː.w|kʰraː.w.níː.tɕam.lɤ̌ː.j.kʰrŭm.doː.n.kùat.rɔ́ː.ŋ.hák.hák]

2.3. Traimit passage (TM)

The TM was created primarily for speech and linguistics research practice, but has also met a set of criteria—complete phonemic inventory, control of the phonetic complexity of words, covering a number of positions of the phonemes in the words (especially the rhymes), inclusion of consonant clusters, repeated

presence of some words and inclusion of various sentence lengths which could potentially be useful as a material for quick speech assessment [1].

We believe that a text adapted from traditional stories and tales would yield high repetitions of words while a text deliberately created from minimal pairs would make it unnatural. The TM was a random text found in a school textbook and (subsequently) heavily modified to cover all Thai phonemes—at least one instance of each phoneme (see Section 3) with a reasonable amount of pauses across the passage.

2.3.1. Words and utterances

It contains 102 mono- and multi syllable words and 139 syllables. Of those, 69 were content words, 29 function words and 4 proper nouns. Moreover, 47.6% of words are "high frequency" (occurring more than 3,000 times) and 34% of words occurring between 100-2,900 times based on Thai corpus data [10]. The TM is composed of 25 utterances/sentences in various length (1 to 10 syllables per utterance; average of 5.6 syllables per utterance). The text contains 5 minimal sets, 14 minimal pairs, and 8 near-minimal pairs.

2.3.2. Repetitions of words

The TM has low repetitions of words to allow for observation of production consistency. Thirteen words get repeated once and one word five times.

2.3.3. Rhyme units

One important aspect of the TM is that it could be useful for a research investigation on tonal production and variation across various checked and unchecked syllables [6] in natural speech. Table 1 summarizes the TM rhyme-tone distribution, shaded cells show three only gaps in the distribution. A "*" indicates a combination that is deemed as 'rare occurrence' and "**" as 'no occurrence' in the language [7].

	mid	low	falling	high	rising
short V + stop	0**	18	0*	21	0**
short V + nasal	14	2	2	0	5
short V + glide	6	1	1	0	1
long V + stop	0**	5	5	1*	0**
long V + nasal	15	1	5	4	5
long V + glide	1	0	3	4	1
long V + no coda	7	4	4	1	2

Table 1: Rhyme-tone distribution in the TM.

Thai script: โรงเรียนไตรมิตรรามอินทราของฉันและเธอ|เป็นโรงเรียนเล็กเล็กริมคลองบางชื่อ|อยู่ห่างจากหมู่บ้านสองแคว|ประมาณหนึ่งกิโลเมตร|พวกเรามีสวนติดป่าดง|เราฝึกปลูกพืชผักสวนครัว|และผลไม้หลายชนิด|เช่น|ต้นกะเพรา|มะเขือเปราะ|ผักกระเฉด|กล้วยน้ำว้า|เงาะ|ลูกพลับ|และมะเฟืองหวาน|โดยฉันกับเพื่อนเพื่อน|จะช่วยกันรดน้ำ|พรวนดินและใส่ปุ๋ย|จนผักผลไม้งอกงามอยู่เสมอ|ในตอนเช้าและตอนเย็น|น้องน้องทุกคน|ช่วยกันเก็บขยะที่เลอะเทอะ|กวาดใบไม้แห้งที่ร่วงหล่น|ลงบนถนนรอบอาคารเรียน|และหน้าเสาธงจนสะอาด

Transcription:[ro:ŋ.rian.traj.mít.ra:m.ʔin.tʰra:.kʰɔ̌:ŋ.tɕʰăn.lɛ́ʔ.tʰɤ̂:|pen.ro:ŋ.rian.lék.lék.rim.kʰlɔ:ŋ.ba:ŋ.sûː|jù:.hà:ŋ.tɕà:k.mù:.bâ:n.sɔ̌:ŋ.kʰwɛ:|prà?.ma:n.nùŋ.kì?.lo:.mé:t|pʰûak.raw.mi:.sǔan.tìt.pà:.doŋ|raw.fùk.plù:k.pʰûː.tpʰàk.sǔan.kʰrua|lɛ́ʔ.pʰŏn.lá?.má:j.lă:j.tɕʰá?.nít|tɕʰên|tôn.kà?.pʰraw|má?.kʰúa.prɔ̀ʔ|pʰàk.krà?.tɕʰè:t|klûaj.ná:m.wá:|ŋɔ́ʔ|lûːk.pʰláp|lɛ́ʔ.má?.fuaŋ.wă:n|do:j.tɕʰăn.kàp.pʰûan.pʰûan|tɕà?.tɕʰûaj.kan.rót.ná:m|pʰruan.din.lɛ́?.sàj.pŭj|tɕon.pʰàk.pʰŏn.lá?.má:j.ŋɔ̂:k.ŋa:m.jù:.sà?.mɤ̌:|naj.tɔ:n.tɕʰá:w.lɛ́?.tɔ:n.jen|nɔ́:ŋ.nɔ́:ŋ.tʰúk.kʰon|tɕʰûaj.kan.kèp.kʰà?.jà?.tʰîː.lɤ́?.tʰɤ́?|kwà:t.baj.má:j.hêːŋ.tʰîː.rûaŋ.lòn|loŋ.bon.tʰà?.nŏn.rɔ̂:p.ʔa:.kʰa:n.rian|lɛ́?.nâː.săw.tʰoŋ.tɕon.sà?.ʔà:t]

English translation: **Our Traimit Ram-Intra School is a small school located near Bang-Sue Canal and is about one kilometre from Song-Kwae Village. We have a garden near the forest. We practice growing plants, vegetables, and many kinds of fruit trees such as basil, eggplant, water spinach, banana, rambutan, persimmon, and sweet star apple. My friends and I always water, loosen the soil, and fertilize vegetables and fruit trees until they grow well. In the morning and evening, all students help picking up dirty trash and sweeping fallen dry leaves around the school building and the area in front of an in-ground flagpole.**

3. PHONEME DISTRIBUTIONS

Table 2 summarizes all phoneme distributions (i.e., initial consonant, initial cluster, vowel, lexical tone, and final consonant) across the four passages, A "✓" indicates a complete distribution while "×" marks the incomplete. The TM gets all "✓" marks.

Phoneme distribution of 21 initial consonants and 12 initial clusters from the four passages is given in Table 3, of nine finals in Table 4, of 18 monophthongs and three diphthongs in Table 5 and that of five tones in Table 6. It is important to note that the TM has covered all of the Thai phonemes.

	initial C	initial cluster	monoph thong	diph thong	tone	final C
NW-Th	×	×	×	✓	✓	✓
RM-I	✓	×	×	×	✓	✓
RM-II	×	×	×	×	✓	✓
TM	✓	✓	✓	✓	✓	✓

Table 2: Summary of phoneme distributions across four passages.

initial C	NW -Th	RM -I	RM -II	TM
p	1	4	1	3
pʰ	13	6	5	9
b	0	4	1	4
t	5	8	1	4
tʰ	20	5	1	7
d	8	4	2	3
k	15	4	3	6
kʰ	3	4	1	5
ʔ	8	3	1	3
tɕ	7	7	7	4
tɕʰ	1	4	1	8
f	0	3	0	2
s	12	5	1	8
h	2	4	3	2
m	12	5	5	11
n	20	6	4	9
ŋ	0	3	0	3

initial C	NW -Th	RM -I	RM -II	TM
r	6	6	1	12
l	17	4	3	16
j	6	7	0	4
w	4	4	1	2
pr	0	0	2	2
pʰr	4	0	2	2
pl	0	0	2	1
pʰl	0	0	5	1
tr	0	0	3	1
tʰr	0	0	0	1
kr	0	0	2	1
kʰr	4	0	5	1
kl	0	0	3	1
kʰl	0	0	3	1
kw	3	0	2	1
kʰw	1	0	1	1

Table 3: Initial consonant distribution.

final C	NW -Th	RMI	RM -II	TM
p	3	3	6	4
t	14	2	3	9
k	15	7	7	12
ʔ	18	7	3	25
m	13	6	2	5

final C	NW -Th	RM -I	RM -II	TM
n	23	16	9	32
ŋ	25	11	10	16
j	9	10	8	13
w	5	8	8	5
no coda	47	30	16	18

Table 4: Final consonant distribution.

vowel	NW -Th	RM -I	RM -II	TM
i	6	5	3	7
iː	11	6	1	3
e	2	3	3	6
eː	0	0	0	2
ɛ	4	4	4	6
ɛː	6	7	4	2
a	50	17	17	31
aː	36	20	13	23
ɯ	7	1	2	2
ɯː	2	0	0	2
ɤ	0	0	0	2

vowel	NW -Th	RM -I	RM -II	TM
ɤː	6	1	3	2
u	3	2	2	2
uː	3	7	4	5
o	9	9	3	13
oː	0	1	2	4
ɔ	1	1	0	2
ɔː	13	11	6	9
ia	3	0	0	3
ɯa	9	2	1	4
ua	1	3	4	9

Table 5: Monophthong and diphthong distribution.

tone	NW -Th	RM -I	RM -II	TM
mid	64	30	21	43
low	28	16	18	31
falling	36	25	9	20
high	30	9	12	31
rising	14	20	12	14

Table 6: Lexical tone distribution.

4. VOWEL MEASUREMENTS FROM THE TRAIMIT PASSAGE

Recordings of the TM passage have been made by 134 Thai adults (43 healthy controls, 45 adults with clinical depression and 46 adults with other psychiatric conditions) as part of spoken data profile for each individual in an ongoing research project "Artificial intelligence for screening of depression and other psychiatric conditions with speech features analysis from a Thai psychiatric screening test" [11][12]. The recordings were made with a designed mobile application onto a smartphone in a quiet room at Siriraj Hospital (Bangkok) at a sampling rate of 16.0 kHz.

To demonstrate its utility in extracting the vowel phonemes, the first two formants of a subset of vowels from two healthy adults (a 22- year old female (Figure 1) and a 27 year-old male (Figure 2)) reading the passage were measured at the vowel midpoint using PRAAT Version 6.0.05. One token per vowel was measured for the 18 monophthongs of each speaker. The words were [tìt mīː kèp tɕhèːt lɛ́ʔ hêːŋ phàk tɕàːk fùk phúːt thɤ́ʔ thɤ̄ː thúk lûːk lōŋ lōː ŋɔ́ʔ and ŋɔ̂ːk]. Measurements from more than one token form each vowel would be ideal, but as can be seen in Figures 1 and 2, the measurements from just one token for each vowel could produce a quick phonetic depiction of the speaker's vowel space.

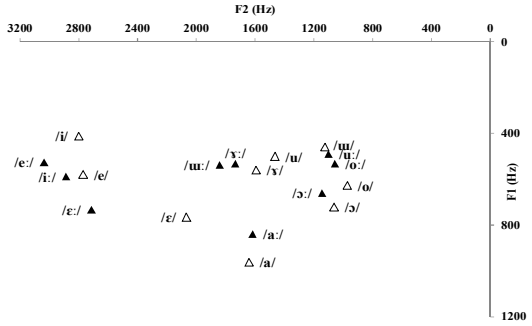


Figure 1: Plot of the first two formants for the Thai female reading the TM passage.

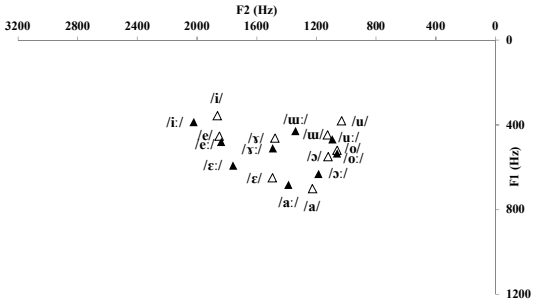


Figure 2: Plot of the first two formants for the Thai male reading the TM passage.

5. CONCLUSION AND FUTURE DIRECTION

As shortcomings of the Thai reading passages were discussed, the TM has been shown to address some of the limitations and could be useful for Thai linguists and speech researchers. With further evaluations and adjustments, it could be a good candidate for a speech assessing task. Being relatively short and simple, the TM may be used in conjunction with designed word list(s) allowing for supplementary phonetic descriptions in controlled phonetics study.

There are still areas of improvement for the TM: modification to allow for various prosodic variations [1][4], incorporation of more words of increasing complexity [1][4], and inclusion of words to fill in the three rhyme-tone gaps in Table 1. Future work should include further evaluations in extracting other Thai phonemes (especially the tones) and evaluations of its level of phonetic balance [13]. Preliminary examination of the recordings in Section 4 has suggested that the word แคว [kʰwɛː], which has contributed to many instances of pronunciation error, should be changed to [kʰwɛ́ːn] แคว้น and that caution should be taken for analysis of segmental duration as the TM was not controlled for pattern of stressed and unstressed words across an utterance/sentence.

6. ACKNOWLEDGEMENTS

We thank Thanaporn Anansiripinyo for her role involving an earlier version of the TM passage and Kankamol Jaisin, M.D., Keerati Pattanaseri, M.D., Juthawadee Lortrakul, M.D., for their role in data collection process at Siriraj Hospital (Bangkok).

7. REFERENCES

[1] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., & Woisard, V. 2022. Criteria for creating new standard reading passages for the assessment of speech and voice: A Delphi consensus study. *Clinical Linguistics & Phonetics*, 1-20.

[2] Baird, L., Evans, N., & Greenhill, S. J. 2022. Blowing in the wind: Using 'North Wind and the Sun' texts to sample phoneme inventories. *Journal of the International Phonetic Association*, Vol.52, no.3, 453-494.

[3] Deterding, D. 2006. The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, Vol.36, no.2, 187-196.

[4] Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., & Russell, S. 2013. "The caterpillar": a novel reading passage for assessment of motor speech disorders. *American Journal of Speech-language Pathology*, Vol.22, no.1, 1–9. https://doi.org/10.1044/1058-0360(2012/11-0134)

[5] Gurevich, N., & Kim, H. 2022. Examination of Consonantal Phonetic Coverage in Standard Reading Passages. *Perspectives of the ASHA Special Interest Groups*, Vol.7, no.5, 1573-1582.

[6] Gedney, W.J. 1989. A checklist for determining tones in Tai dialects. In R.J. Bickner, J. Hartmann, T.J. Hudak and P. Peyasantiwong (Ed.), *Selected Papers on Comparative Tai Studies*, 191-206. Center for South and Southeast Asian studies, the University of Michigan.

[7] Gandour, J. 1974. Consonant types and tone in Siamese. *Journal of Phonetics*, Vol2, no.4, 337-350.

[8] Tingsabadh, M. K., & Abramson, A. S. 1993. Thai. *Journal of the International Phonetic Association*, Vol.23, no.1, 24-28.

[9] Sindermsuk, D. 1986. *The survey of speech defects among prathom 4 students in Mitsampan school group*. MA. dissertation, Mahidol University.

[10] Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T. & Wutiwiwatchai, C. 2009, October. BEST 2009: Thai word segmentation software contest. In 2009 *Eighth International Symposium on Natural Language Processing* (pp. 83-88). IEEE.

[11] Klangpornkun, N., Ruangritchai, M., Munthuli, A., Onsuwan, C., Jaisin, K., Pattanaseri, K., Lortrakul, J., Thanakulakkarachai, P., Anansiripinyo, T., Amornlaksananon, A., Laohawee, S., & Tantibundhit, C. 2021, November. Classification of Depression and Other Psychiatric Conditions Using Speech Features Extracted from a Thai Psychiatric and Verbal Screening Test. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 651-656). IEEE.

[12] Munthuli, A., Pooprasert, P., Klangpornkun, N., Phienphanich, P., Onsuwan, C., Jaisin, K., Pattanaseri, K., Lortrakul, J., & Tantibundhit, C. (to appear). Classification and analysis of text transcription from Thai Depression Assessment Tasks among patients with depression. *PLoS ONE*.

[13] Munthuli, A., Sirimujalin, P., Tantibundhit, C., Kosawat, K., & Onsuwan, C. 2013. A corpus-based study of phoneme distribution in Thai. In *10th International Symposium on Natural Language Processing* (pp. 114-121).

Infant Directed Speech expands F1/F2 space for /i a u/: But what happens to central vowels /ɜ ʊ/ in Australian English?

Antonia Goetz[1], Eylem Altuntas[1], Nicole Swinfield[1], Denis Burnham[1], & Catherine T. Best[1]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University, Sydney, Australia
a.goetz@westernsydney.edu.au

ABSTRACT

Relative to their adult-directed speech (ADS), mothers expand the triangle circumscribed by the reference vowels (/i a u/) in F1/F2 space when interacting with their infants. This vowel space expansion in infant-directed speech (IDS) is thought to support word learning and phonological development. But what about non-reference (i.e., 'central') vowels such as /ɜ ʊ/? Are they also located differently in F1/F2 space in IDS and ADS? We compared /ɜ ʊ/ in Australian English, in which they fall posterior to the vowel triangle, produced by 15 mothers addressing their 4-month-old versus an adult. We extracted F1 and F2 values and calculated the Euclidean distance between the vowels for IDS and ADS. The distance was significantly *shorter* in IDS than ADS, rather than longer as with the /i a u/ triangle. F2 (but not F1) was also lower in IDS than ADS, suggesting more posterior tongue placement for /ɜ ʊ/ in IDS.

Keywords: infant-directed speech, central vowels, speech production, vowel hyperarticulation

1. INTRODUCTION

When people talk to infants, they intuitively use a speech register called infant-directed speech (IDS), which is different from the speech used in communication with other adults (adult-directed speech, ADS). The acoustic features of IDS relative to ADS include higher average fundamental frequency (f0) [1], greater variation in f0 [1, 2] and vowel hyperarticulation [3], which refers to expansion of the triangle circumscribed by /i a u/ in F1/F2 acoustic space in IDS relative to ADS. The phonological reference vowels, /i a u/, mark the extremes of the two primary dimensions of vowel systems: height and anterior-posterior position of the tongue body. All languages with three or more vowels have /i a u/; numerous languages have only /i a u/.

The altered acoustic features of IDS are posited to serve multiple functions in communication with infants. The higher f0 of IDS, and its associated positive vocal affect, help to regulate infants' emotional states [4]; and /i a u/ hyperarticulation in IDS is thought to facilitate early lexical development [5]. For example, mothers' /i a u/ vowel hyperarticulation in communication with their 9-month-old infants is a reliable predictor of expressive vocabulary size at 19 months of age [3], and hyperarticulated /i a u/ are linked to facilitation of infants' word recognition also at 19 months [6]. While it has been hypothesised that caregivers hyperarticulate vowels in order to clarify the acoustic-phonetic features of the vowels [7], /i a u/ hyperarticulation has not always been found in IDS in languages other than English [8-10], and the underlying mechanism and functions of IDS remain unclear [11, 12] (but see, [13]). The specific unresolved issue our study addresses is how *non*-reference, i.e., *central,* vowels are produced in IDS versus ADS, and whether any differences results are associated with their phonological status as non-reference/central in the vowel system, or with their position in F1/F2 space. For this purpose, we investigated the production of the central vowels /ɜ ʊ/ in IDS and ADS in Australian English.

Previous studies with Australian English (AusE) have found that the /i a u/ vowel triangle defined by the vowels in *sheep* (/i/), *shark* (produced as /a/ in AusE, a non-rhotic accent) and *shoes* (/u/) is indeed expanded in IDS compared to ADS [3, 14]. However, in AusE /u/ is articulated in a much more fronted (more anterior) position (realised phonetically as [ʉ]), and /a/ is slightly more fronted, than they are in most other varieties of English (e.g., Received Pronunciation [RP] British English), as illustrated in Figure 1. That is, two of the three reference vowels, /u/ and /a/, are not articulated at the most extreme posterior positions possible in the acoustic vowel space, as they are in RP British English. Consequently, the AusE vowel triangle has become highly truncated in the anterior-posterior (F2) dimension, i.e., compressed into the front half of the oral portion of the vocal tract (see [15]). And in relation to that truncation of the reference vowel triangle in F1/F2 space, the central vowels /ɜ ʊ/ (the vowels in *shirt* and *chook*, respectively) actually fall outside rather than inside the triangle In acoustic space, as they do in RP British. So, these two *phonologically* central (non-reference) vowels are *acoustically* more posterior than AusE /u/ and /a/. This situation is more extreme for /ʊ/ than for /ɜ/ but

even the latter is slightly posterior to the /u/-/a/ edge of the reference vowel triangle instead of being squarely within the triangle. The open question is: What happens to the phonologically central but acoustically peripheral vowels of AusE that are *outside* the vowel triangle? For instance, are these vowels also hyperarticulated, i.e., spread further apart as are the reference vowels) or hypo*articulated*, i.e., more spatially congested in IDS compared to ADS, or are they simply unaffected by the change of register?

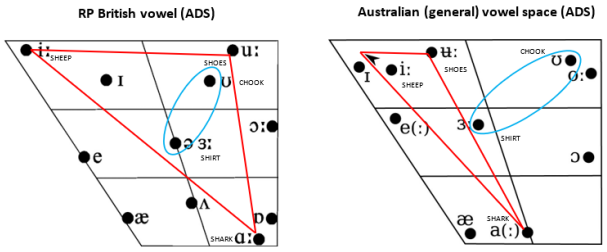


Figure 1: Vowel triangles (red) defined by /i a u/ in F1/F2 space and 'central' vowels /ɜ ʊ/ (turquoise) in British English (left, [16] and AusE (right, [17]. The vertical dimension is vowel height (F1, reverse scale – high F1 = low vowels); the horizontal dimension is anterior-posterior tongue position (F2, reverse scale – high F2 = front vowels).

As part of the [11] project, this study aimed to compare the acoustic features of two AusE vowels in *shirt* (/ɜ/) and *chook* (/ʊ/)[1] produced in IDS and ADS. As we have noted, in AusE these vowels fall outside of the traditional phonologically defined vowel triangle. This situation allows us to compare the acoustic features of a phonologically central vowel (/ʊ/) that is phonetically more extreme in F1/F2 space than the high back reference vowel /u/, to those of another phonologically central vowel (/ɜ/) that is closer to the centre of F1/F2 space but still not inside the reference vowel triangle (see Fig. 1).

Only one previous study has compared IDS and ADS for a vowel that falls *inside* in the vowel triangle in American English [11]. That study compared an anterior non-reference vowel, /ɪ/, to the closest reference vowel /i/, and found the distance between them to be shorter in IDS than in ADS, even though the reference vowels were hyperarticulated [12]. However, based on those results for American English we cannot draw conclusions about hyper- or hypoarticulation of the 'central' vowels in AusE, given that in AusE the phonologically 'central' vowels fall outside of the vowel triangle (Fig. 1).

2. METHOD

2.1. Participants

Fifteen mother-infant dyads participated in this study as part of a longitudinal project. All mothers were native speakers of AusE, and all infants (8 female) were acquiring only AusE as their first language with no more than 4 hours per week of exposure to any language other than English. No infants had any reported health issues, and none were at risk for any language or cognitive disabilities.

Infants participated in an auditory-visual speech perception experiment at the ages of 4, 7 and 10 months, and at each age IDS and ADS recordings were conducted after the perception experiment. Only the IDS session at 4 months (M_{age} = 143 days, range 3.95 – 5.62 months) is included in this report. However, as mothers produced less target words in ADS than IDS, the ADS recordings for all three ages were analysed here to increase the number of ADS productions of the target words in ADS (see *2.2. Procedure*). Infants received a diploma and toy. Caregivers gave informed consent.

2.2. Procedure

For the IDS and ADS sessions, mothers and infants were assigned to separate rooms connected via an audio-visual double screen set-up. This allowed communication between the mother (wearing a microphone-fitted headset (DPA) and infant (or experimenter, for the ADS sessions) in real time with the additional advantage that both participants' visual expressions could be recorded. Only the audio recordings are reported here.

In the IDS session, mothers were provided with five toy objects to prompt elicitations of multiple productions of the target words *sheep*, *shoes*, *shark*, *shirt,* and *chook* (see labels in Fig. 1). Before the IDS sessions, mothers were presented with video-recorded instructions spoken by a native speaker of AusE to ensure that they were introduced to the target words with Australian-English pronunciation. The instructions included directions for the mothers to play naturally with their infant by using the five target-named objects, and to aim to use each word multiple times. Immediately following the IDS recording sessions, the experimenter interviewed the mother in the same double video set-up to elicit productions of the target words in communication with another adult. The IDS and ADS sessions lasted between 10 and 15 minutes in total.

2.3. Acoustic analysis

This report concerns the audio productions of the two central vowels (/ɜ/ and /ʊ/) as in AusE *shirt* and *chook*. As noted above, these vowels are of interest because they fall outside of the phonologically-defined vowel triangle in AusE. All audio files were annotated and acoustically analysed using PRAAT [19]. First, we annotated the onset and offset of the

words and vowels manually using visual and auditory judgments. Segments that contained acoustic artefacts (e.g., vocalisation by the infant, or heavy glottalization) or sung speech to the infant were excluded from analyses. The mean number of extracted tokens per speaker for each vowel in each register (IDS, ADS) is shown in Table 1.

Table 1: Mean numbers per mother (*standard deviation*) of extracted target vowels used for the acoustic analysis of IDS and ADS.

Vowel	IDS	ADS
shirt /ɜ/	8.9 *(4.2)*	6.3 *(2.4)*
chook /ʊ/	13.9 *(7.3)*	7.4 *(1.9)*

Second, we used a PRAAT script to extract the values of the first formant (F1), second formant (F2), and third formant (F3) at 40% and 80% from the vowel onset to identify the steady-state nucleus. Third, we averaged all articulations of each vowel in each register per mother and calculated the Euclidean distance (*ed*) between the mean position of the /ɜ/ and /ʊ/ vowels in each register using the equation:

$$ed = \sqrt{F1_{/ɜ/} - F1_{/ʊ/})^2 + (F2_{/ɜ/} - F2_{/ʊ/})^2}$$

Lastly, we calculated the Pillai score [19] to measure the overlap between two vowel cluster. The Pillai score value range between 0 (indicting almost perfect overlap) and 1 (no overlap).

3. RESULTS

Statistical analysis was performed using R [21] and lme4 [22], and plots were created by using ggplot2 [22]. Figure 2 shows the averaged F1 and F2 values per mother for each Vowel (*shirt* /ɜ/ and *chook* /ʊ/) in each speech Register (ADS, IDS).

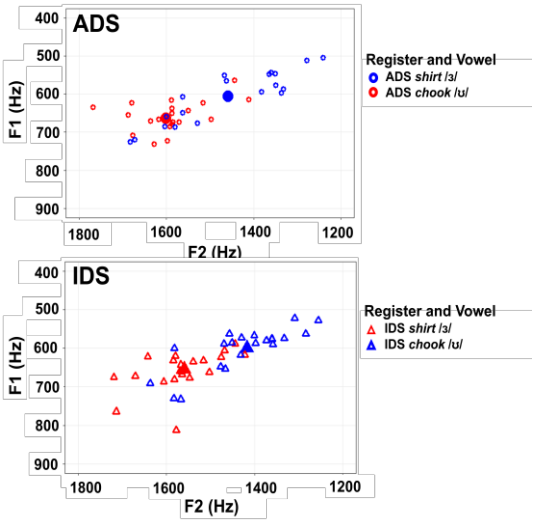


Figure 2: Mean F1 and F2 (Hz) of each mother for each Vowel in each Register (ADS above, IDS below). Filled markers indicate mean values for each vowel in each register.

We performed statistical comparisons with two mixed effect models to compare (i) the Euclidean distance, (ii) the Pillai score and (iii) the F1 and F2 values between the two vowels in ADS and in IDS.

For the first analysis, the dependent variable was Euclidean distance. Register (IDS coded as +0.5 vs. ADS coded at -0.5) was entered as a fixed effect and Mother as a random effect in the model. The results revealed that the Euclidean distance between the /ɜ/ and /ʊ/ vowels was significantly shorter in IDS than ADS (β *(SE)* = -83.26 (15.78), t = -5.276, $p < .001$), see Figure 3.

For the second set of analysis we extracted the Pillai score, by using the manova() function in R. The Pillai score was higher for ADS (Mean = 0.32, SD = 0.20) than IDS (Mean = 0.21, SD = 0.23).

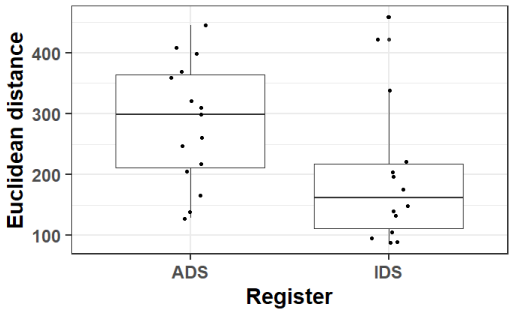


Figure 3: Boxplot of the Euclidean distances between the vowels *shirt* /ɜ/ and *chook* /ʊ/ in ADS and IDS. Dots represent the Euclidean distance per mother in each Register, averaged over tokens.

For the third analyses the dependent variables were F1 and F2 values in Hz with Register (IDS coded as +0.5 vs. ADS coded at -0.5) interacting with Vowel (*shirt* coded as 0.5, *chook* coded as -0.5) as fixed factors, and Mother as a random effect.

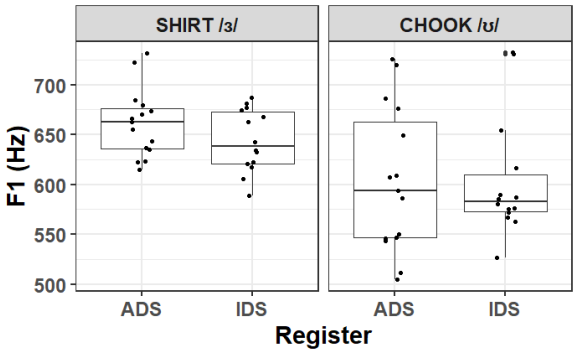


Figure 4: F1 (Hz) for the two Vowels (*shirt* /ɜ/ on the left and *chook* /ʊ/ on the right).

The F1 analysis (see Figure 4) showed neither a significant main effect of Register (β *(SE)* = -4.328 (6.370), t = -0.689, p = .500), nor a significant interaction with Vowel (β *(SE)* = -4.472 (6.338), t = -0.706, p =.484).For the F2 analysis, however, there

was a main effect of Register; F2 was significantly lower in IDS than ADS (β *(SE)* = -28.363 (11.472), t = -2.472, p = .018, see Figure 5), but still showed no significant interaction with Vowel (β *(SE)* = -4.755 (11.392), t = -0.417, p = .679). These F2 findings indicate that there is a more posterior tongue placement for both vowels in IDS than ADS, i.e., both vowels are positioned farther beyond the reference vowel triangle in IDS than ADS.

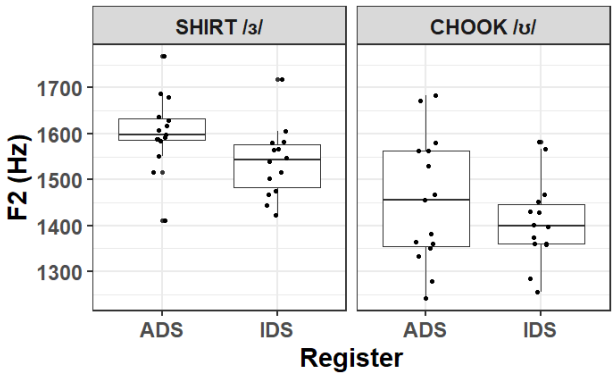


Figure 5: F2 (Hz) for the two Vowels (*shirt* /ɜ/ on the left and *chook* /ʊ/ on the right).

4. DISCUSSION

In contrast to documented expansion of the F1/F2 space circumscribed by the phonological reference vowels /i a u/ in IDS compared to ADS, little is known about whether mothers also change the acoustic features of non-reference vowels when speaking to their infants versus to an adult, and if so, then in what direction (expansion, contraction) such changes might occur. The present study investigated mothers' IDS versus ADS articulation of vowels that function phonologically as non-reference vowels, but that fall outside rather than inside the reference vowel triangle in their phonetic details, in AusE.

We found that Australian mothers produced the non-reference/central vowels /ɜ/ and /ʊ/ with a shorter Euclidean distance and with more overlap in IDS than in ADS, suggesting that these vowels are hypo- rather than hyper-articulated relative to each other in IDS. In addition, lower F2 values in both vowels in IDS than ADS indicate that these vowels are produced with a more posterior tongue placement in IDS than in ADS.

Previous studies with AusE mothers have shown that they reliably expand their vowel space in communication with their infants, via hyperarticulation of the /i a u/ reference vowels. However, results regarding hyperarticulation are not entirely consistent across languages. While some studies have found that reference vowels are hyperarticulated [3, 9], others have found that the reference vowels may be hypo*articulated* in languages other than English [8-10] in IDS compared to ADS. Even in English hyperarticulation is also lacking under certain external conditions such as experimental manipulation of the infant's ability to hear their mother's voice [24], or with inherent factors such as infants' familial risk for dyslexia [14].

In this study we have shown that while the reference vowels, /i, a, u/, are hyperarticulated in AusE, the distance between two non-reference vowels is hypoarticulated. This is consistent with [12] which found that, even when the reference vowels are hyperarticulated, the distance between the non-reference central vowel /ɪ/ that falls *inside* the reference vowel triangle and the reference vowel /i/ was reduced in IDS, i.e., hypo*articulated*, compared to ADS. The co-occurrence of *hyper*articulation of reference vowels and *hypo*articulation of a non-reference vowel could suggest that mothers apply both acoustic strategies to support infant linguistic development. Our results with two non-reference vowels that fall *outside* the reference vowel triangle suggest some additional insights. While these two non-reference vowels, which fall outside the /i a u/ F1/F2 triangle in AusE, are hypo-articulated in relation to each other, they are both articulated farther posterior in IDS than ADS. It is possible that the more posterior tongue placement increases the distance between these and the reference vowels, i.e., the difference between these non-reference vowels and the reference vowel triangle may, in contrast to [14] be *hyper*articulated in AusE. If so, that could facilitate infants' ability to differentiate between reference and non-reference vowels in this variety of English.

These observations raise unanswered questions as to (i) whether hypo- or hyperarticulation of non-reference vowels with respect reference vowels depends on their phonetic position inside or outside the reference vowel triangle, and (ii) whether IDS modifications of reference and non-reference vowels, have didactic value for infant language learning. So, future research is required: (i) to compare productions of /ɜ/ and /ʊ/ to those of /i a u/, as well as to productions of other non-referent vowels that fall *inside* the /i a u/ triangle in AusE (e.g., /ɪ/ as in *chip*); and (ii) to examine how the degree of reference vowel triangle expansion *and* modifications of non-reference vowels in a mother's speech to her infant at an early age (e.g., 4 months) may predict her infant's vocabulary [as per 4], and word recognition [as per 7] at 19 months of age. In addition, given that the two extant studies with non-referent vowels investigated different English varieties, IDS versus ADS articulations of the *shirt* and *chook* vowels should also be examined in other English varieties in which these vowels fall *inside* the reference vowel triangle (e.g., American, RP British).

5. REFERENCES

[1] R. P. Cooper and R. N. Aslin, "Developmental differences in infant attention to the spectral properties of infant-directed speech," *Child development,* vol. 65, no. 6, pp. 1663-1677, 1994.

[2] A. Fernald, "Intonation and communicative intent in mothers' speech to infants: is the melody the message?," *Child development,* pp. 1497-1510, 1989.

[3] M. Kalashnikova and D. Burnham, "Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions," *Journal of child language,* vol. 45, no. 5, pp. 1035-1053, 2018.

[4] M. Papoušek, "Communication in early infancy: An arena of intersubjective learning," *Infant Behavior and Development,* vol. 30, no. 2, pp. 258-266, 2007.

[5] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience,* vol. 5, no. 11, pp. 831-843, 2004.

[6] J. Y. Song, K. Demuth, and J. Morgan, "Effects of the acoustic properties of infant-directed speech on infant word recognition," The Journal of the Acoustical Society of America, vol. 128, no. 1, pp. 389-400, 2010.

[7] H. M. Liu, P. K. Kuhl, and F. M. Tsao, "An association between mothers' speech clarity and infants' speech discrimination skills," Developmental science, vol. 6, no. 3, pp. F1-F10, 2003.

[8] K. T. Englund and D. M. Behne, "Infant directed speech in natural interaction—Norwegian vowel quantity and quality," Journal of psycholinguistic research, vol. 34, no. 3, pp. 259-280, 2005.

[9] T. Benders, "Nature's distributional-learning experiment," Citeseer, 2013.

[10] N. X. Rattanasone, D. Burnham, and R. G. Reilly, "Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age," Journal of Phonetics, vol. 41, no. 5, pp. 332-343, 2013.

[11] Burnham, D., Best, C. (Cutler, A., Johnson, E., Kalashnikova, M.) Origins of Phonology and Lexicon: Abstract representations before 6 months Australian Research Council Discovery Grant, DP00025567.

[12] A. Cristia and A. Seidl, "The hyperarticulation hypothesis of infant-directed speech," Journal of child language, vol. 41, no. 4, pp. 913-934, 2014.

[13] B. McMurray, K. A. Kovack-Lesh, D. Goodwin, and W. McEchron, "Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?," Cognition, vol. 129, no. 2, pp. 362-378, 2013.

[14] M. Kalashnikova, C. Carignan, and D. Burnham, "The origins of babytalk: smiling, teaching or social convergence?," Royal Society open science, vol. 4, no. 8, p. 170306, 2017.

[15] M. Kalashnikova, U. Goswami, and D. Burnham, "Mothers speak differently to infants at-risk for dyslexia," Developmental science, vol. 21, no. 1, p. e12487, 2018.

[16] C. Kitamura, R. Panneton, and C. T. Best, "The development of language constancy: Attention to native versus nonnative accents," Child Development, vol. 84, no. 5, pp. 1686-1700, 2013.

[17] P. Roach, "British English: received pronunciation," Journal of the International Phonetic Association, vol. 34, no. 2, pp. 239-245, 2004.

[18] F. Cox and J. Fletcher, Australian English pronunciation and transcription. Cambridge University Press, 2017.

[19] P. Boersma, "Praat: doing phonetics by computer (Version 5.1. 05)," http://www. praat. org/, 2009.

[20] Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. Journal of Phonetics, 34(4), 458–484.

[21] R Core Team, R: The R project for statistical computing. [Online]. Available: https://www.r-project.org, 2022.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," J. Stat. Soft., vol. 67, no. 1, pp. 1–48, Oct. 2015.

[23] H. Wickham, "Programming with ggplot2," in ggplot2: Springer, 2016, pp. 241-253.

[24] C. Lam and C. Kitamura, "Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation," Developmental Science, vol. 15, no. 2, pp. 212-221, 2012.

[1] Note that /r/ is not rhoticized in Australian English. The word *chook* is a common informal name for 'chicken' in Australian English.

Contribution of wavelet-based features to identification of overlapping Azerbaijani vowels using machine learning

Payam Ghaffarvand-Mokari

University of Eastern Finland, School of Humanities
payam.ghaffarvand.mokari@uef.fi

ABSTRACT

While vowels are generally found to be distinct based on their first two or three formants, Azerbaijani mid-front rounded /œ/, high-back unround /ɯ/, and high-front rounded /y/ are found to be largely overlapping. The present study examines the contribution of wavelet-based decompositional features to semi-automatic identification of these vowels using random forest and neural networks. From a data sample of 607 vowel tokens produced by eight male speakers, first three formants and six Shannon entropy wavelet packets were extracted and served as the training features for the classifiers. Results showed that the addition of the wavelet-based features to formant information improved the classification accuracies by 8–11% across the classifiers. The highest classification accuracy was 83.1% achieved by random forest.

Keywords: Azerbaijani, vowels, wavelet packets, formants, machine learning.

1. INTRODUCTION

The identification of robust acoustic parameters for the classification of speech sounds is crucial for developments of automatic speech recognition systems, as well as for basic research in different areas of phonetics and phonology. Vowels are especially important for automatic speech processing as they contain useful information for speech and speaker identification/verification, clinical assessment of speech/voice disorders, and for forensic applications. Therefore, a better understanding of important features in classification of speech sounds across the languages is important.

First noted by Householder in 1972 [1], Azerbaijani mid-front rounded /œ/ and high-back unround /ɯ/ are found to be largely overlapping on a F1–F2 space [1, 2]. To a lesser extent, the high-front rounded /y/ is also found to be overlapping with the high-back unround /ɯ/ [2]. This study aims to investigate whether further inclusion of wavelet-based decompositional features for training of the classification models improves the classification of these phonologically distant but acoustically similar vowels.

Wavelet decomposition is a mathematical technique that involves breaking down a signal into different frequency components using wavelets, which are small waves that are localized in both time and frequency domains.

1.1. Background

Researchers have studied different sets of features in classification of vowels of different languages. For instance, Krocil et al. [4] have used F1, F2, and zero-crossing rate (ZCR) in classification of Czech vowels. Their results from a heuristic classifier, revealed the classification accuracy ranged from 54.5% to 97.5% for different vowels.

For Hindi vowels, Biswas et al. [5] used different combination of F1, F2, F3, Gammatone Frequency Cepstral Coefficients (GFCC), and Mel-Frequency Cepstral Coefficients (MFCC) and found that using formants together with GFCC outperformed other combinations of features in different noisy conditions.

Classification of American English vowels in noisy and noise-free conditions has been evaluated using MFCCs and formants [6]. It was found that the classification accuracies based on the formant features and MFCCs were almost the same.

Korkmaz et al. [7] studied the classification of eight Turkish vowels using a feature set optimized by Genetic Algorithm. They used formants, energy, ZCR, MFCCs, and Wavelet Decomposition Shannon Entropy features as their input. Their results from a data set based on vowel tokens uttered by 10 male speakers showed that the feature vector optimized by Genetic Algorithm method reached 100% classification accuracy.

Among the limited number of studies on classification of Azerbaijani vowels are the studies by Ghaffarvand-Mokari and Werner [2] and Imamverdiyev and Sukhostat [3]. Ghaffarvand-Mokari and Werner [2] found that the classification accuracies using simple Linear Discriminant Analysis (LDA) with first two formants as predictors were very low for Azerbaijani /œ/-/ɯ/ vowels. They found that further inclusion of the third formant to the LDA models, improves the classifications accuracy but still the largest confusion is between /œ/-/ɯ/

vowels. The /y/ vowel was also found to be overlapping with /ɯ/. Imamverdiyev and Sukhostat [3] evaluated classification of Azerbaijani vowels based on MFCCs using Support Vector Machine (SVM). Their findings show that /œ/, /ɯ/, /y/, and /e/ vowels were poorly classified (with accuracy rates of 45.5% – 60%) compared to other vowels (82.6% – 95.7%).

1.2. Present study

The present paper aims at evaluating the contribution of Wavelet Decomposition Shannon Entropy features in identification of the /œ/, /ɯ/, and /y/ vowels using random forest and neural networks as classifiers.

The findings would help identification of a subset of acoustic features that carry most of the information for efficient classification of these vowels.

2. METHODS

2.1. Speech data

A part of previously recorded corpora for studying Azerbaijani fricative [8] was used as the speech data for the present study. The corpora were recorded from speakers of southern Azerbaijani dialect. Recorded data from 8 male speakers were used who had relatively balanced number of reliable tokens across /ɯ/, /œ/, and /y/ vowels. Recordings used for this study had the vowels in V position of consonant-vowel-consonant (C_1VC_2) contexts. The C_1 varied over nine different fricative consonants and C_2 was kept identical. The target words were embedded in a fixed carrier phrase “burdɑki _ kælmæsidi” (here is the _ word). Each target vowel was repeated 3 times, yielding a total of 648 tokens per speaker (3 vowels × 9 different C_1 consonants × 3 repetitions).

After removing unreliable and wrong pronunciations, a total of 607 tokens were submitted for the analyses. Recordings were done using a condenser microphone (Blue Bluebird, USA) microphone with a sound card (RME Fireface 800, Germany) connected to a PC computer through a controller card (VIA VT6307, Taiwan) in a sound-attenuated booth. Recordings were saved as wav files with sampling frequency of 44.1 kHz and 16-bit quantization. Tokens were segmented and labeled manually based on the waveforms and spectrograms. To minimize coarticulation effects from neighboring consonants, the middle 50% of the vowel duration were extracted and used for the analyses.

2.2. Feature extractions

The vowel parts of the signals were labelled manually in Praat [9] and formants were calculated using a Praat script with Burg algorithm with range of 0-5,000 Hz for five formants. Fig. 1 shows the spectrograms of the middle parts of the vowels uttered by a male speaker. F1, F2 and F3 values were then measured for each vowel by taking an average of the 50% of the vowel centred around the midpoint.

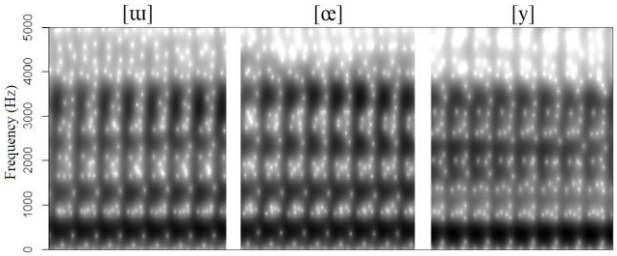


Figure 1: Spectrograms of the middle 50 ms of /ɯ/, /œ/, and /y/ vowels in a h-vowel-r context uttered by a male speaker.

Following the method used in [7], wavelet transform on the selected parts of the vowels were applied to obtain decompositional features in MATLAB. The ‘wavelet transform’ divides signal into sub-signals with low and high frequencies. A Daubechies-filtered 5-depth wavelet packet decomposition was used. Six features selected by considering d1-d5 and a5 sub-signals (Fig. 2).

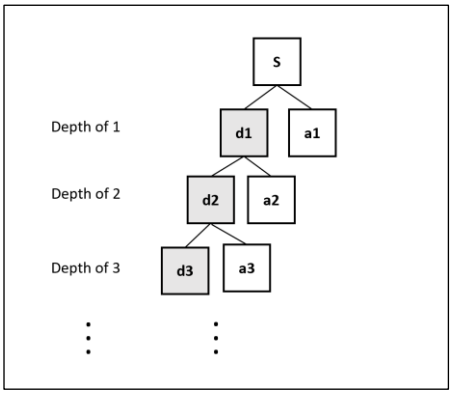


Figure 2: The structure of wavelet packet decomposition. Shaded boxes indicate selected sub-signals.

After having sub-signals, the Shannon entropies of them were calculated. Finally, Shannon entropy values of sub-signals resulted in the original signal’s decompositional features.

2.3. Classifiers

Random forest [10] and neural networks [11] were used to predict the vowels based on the formants and wavelet-based features.

Random forest with 300 decision trees were used to classify the vowels and to analyse the importance of each variable in correctly identifying the vowels.

A feed-forward neural network was used which included input, hidden, and output layers. In this network, the information moves in only one forward direction from the input layer. The neurons in each layer are connected to the neurons of the subsequent layer. Ten hidden layers used for the classifications. The input layer has one neuron for each variable and the output layer has one neuron for each predicted vowel.

The classifiers were trained on a randomly chosen 70% of the data and evaluated on the remaining 30% of the data. The training and testing data had a balanced number of the predicted vowels. Ten replications of each classification were done. Each replication had a randomly partitioned data into training and test subsets.

3. RESULTS

The /œ/ and /ɯ/ vowels, and to some extent /y/, were found to be overlapping on a F1–F2 space (Fig. 3).

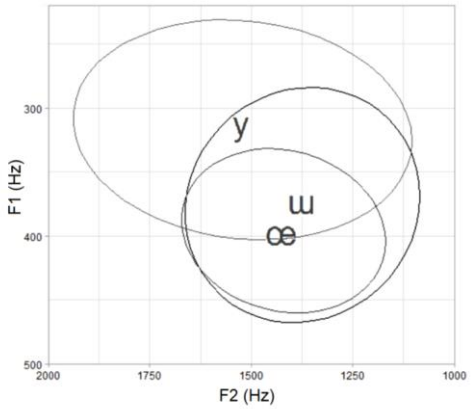


Figure 3: Distribution of the /œ/, /ɯ/, and /y/ vowels in a F1 × F2 (Hz) space with ellipses representing one standard deviations from the mean.

Results from 10 replications of the classifications using random forest and neural networks showed accuracy rates did not vary much between the classifiers for the models with F1 and F2 as predictors, but for the models with all three formants and those with formants and wavelet-based features, random forest resulted in higher accuracies (Fig. 4). The mean accuracy with F1 and F2 as the input variables was 63.6% (95% CI [60.8–66.4]) and 64.7% (95% CI [62.3–67.1]) for neural networks and random forest, respectively. After inclusion of F3, mean accuracy rate reached 69% (95% CI [67.1–70.9]) for neural networks and to 72% (95% CI [70–74.1]) for random forest. Finally, the inclusion of wavelet-based features further improved the mean accuracy for neural networks to 77.1% (95% CI [74.6-79.5]) and for random forest to 83.1% (95% CI [81.1–85.1]) with a maximum of 87.8% accuracy in one of the replications.

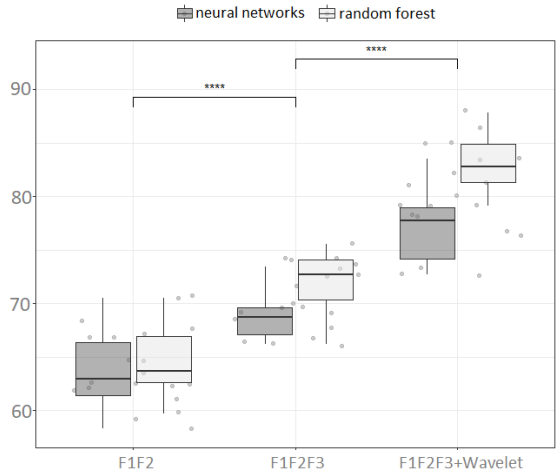


Figure 4: The performance (% of accuracy) of the random forest and neural networks as classifiers in identification of /œ/, /ɯ/, and /y/ based on ten replications with different sets of predictors.

The accuracies across the sets of features were compared by a series of Mann-Whitney Wilcoxon tests. As indicated in Fig. 4 by the three asterisks (***), the accuracy of F1F2 set was significantly different from F1F2F3 set for both classifiers (all $p < 0.001$). The accuracy of F1F2F3 set was also significantly different compared to F1F2F3 + wavelet-based features set for both classifiers (all $p < 0.001$).

3.1. Separate analyses for each vowel

Further, the accuracy of identifications was also analysed separately for vowels to explore how different sets of features contribute to correct identification of each vowel. Mean identification accuracies for /œ/, /ɯ/, and /y/ with random forest as the classifier are presented in Fig 5. The /y/ vowel was identified with mean accuracy of 81.5% (95% CI [78.3–85]) based on F1 and F2. The inclusion of F3 didn't result in noticeable improvement for the identification of /y/ (M = 82%, 95% CI [77.8–86.8]), yet further inclusion of the wavelet-based features resulted in better identification accuracy (M = 89.6%, 95% CI [86.6–92.5]).

For /ɯ/ vowel, the mean identification accuracy with F1 and F2 was 62.6% (95% CI [60.1–65]) and after inclusion of F3, it reached 71.7% (95% CI [69.4–74]). Further inclusion of the wavelet-based features resulted in mean accuracy of 81.8% (95% CI [79.4–84.1]).

For /œ/ vowel, the correct identification rate was close to a chance level[1] with F1 and F2 as predictors (M = 46.1%, 95% CI [41–51.3]). Inclusion of F3 to the models resulted in noticeable improvements in the

identification rates (M = 60.6%, 95% CI [55.5–65.7]). The inclusion of wavelet-based features also substantially improved the identification of /œ/ (M = 78.7, 95% CI [73.9–83.6]).

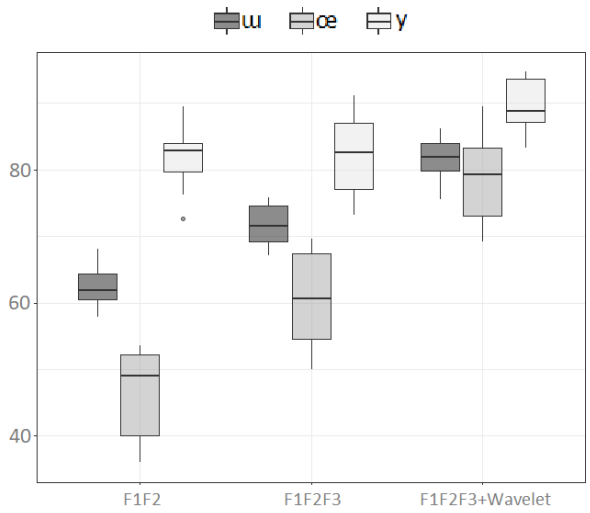


Figure 5: Identification accuracy (%) of /œ/, /ɯ/, and /y/ vowels using random forest as classifier based on ten replications.

3.2. Important features for classifications

To find out which features were the most important ones in identification of the vowels, an estimate of the importance of each predictor for random forest models were calculated. Fig. 6 shows the ranked importance of the features in terms of the mean decrease in the accuracy of the model when excluding a predictor. By a large extent F1 was the most important predictor, followed by F3, F2, d3, and a5.

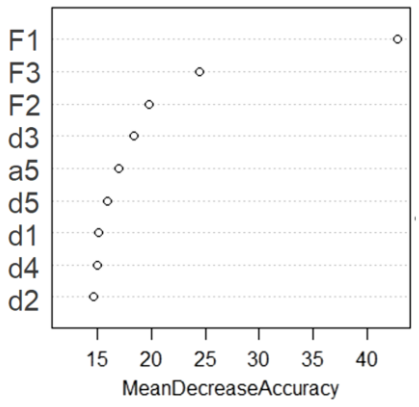


Figure 6: The mean decrease in the accuracy (%) of the random forest model when excluding a predictor (a high decrease means that the variable has salient predictive power)

4. DISCUSSION

This paper was set out to investigate whether formants and wavelet based acoustic cues contain enough information to correctly classify Azerbaijani overlapping /œ/, /ɯ/, and /y/ vowels. Further, two different computational classifiers – random forest and feed-forward neural networks – were evaluated in classifying the vowels using different sets of acoustic cues.

It was found that the accuracy of classifying the vowels based on F1 and F2 was quite low especially for /œ/ vowel. This is in line with findings of [2] that have found large overlap in classification of these vowels based on F1 and F2. Additionally, in line with the findings of [2], further inclusion of F3 to the predictors improved the overall classification accuracy, but it did not have noticeable effect on correct identification of /y/ vowel (Fig 5). This indicates that F3 is an important predictor for /œ/-/ɯ/ classification but not for /y/.

As for the main aim of this paper, wavelet-based features found to improve the classification accuracies by 8.1%–11.1% for random forest. This is in line with the findings Korkmaz et al. [7] which shows five wavelet-based features were selected by Genetic Algorithm as important features in the classification of Turkish vowels. Overall, wavelet-based features seem to contain useful information for classification of the Azerbaijani /œ/-/ɯ/-/y/ vowels. However, they do not contain enough information for optimal classifications. Since wavelet transforms captures both short-term and long-term variations in the signal's characteristics, it is possible that they capture (part of) the information important for listeners in discrimination these overlapping vowels.

Finally, the classification accuracies using random forest were higher than for neural networks. Since neural networks normally need larger data for better trainings of the models, the observed results are possibly due to relatively small size of the data set in this study. Random forests are found to be appropriate for analysis of small data sets and for taking the possible correlation of variables into account [12].

This study has some limitations, probably the most important being the relatively small sample size and that it only analyses data from male speakers. Future studies should investigate the classification of overlapping vowels using larger sample sizes and including vowels produced by female speakers, as well as employing other feature sets and classification algorithms. Future studies should also consider inclusion of all other Azerbaijani vowels and different word contexts in the classifications. Finally, subsequent studies could explore whether there are differences in the acoustic classification outcomes for the vowels of northern and southern Azerbaijani dialects.

5. REFERENCES

[1] Householder, F. 1972. Vowel overlap in Azerbaijani. In: Valdman, A. (eds), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. De Gruyter Mouton, 229–230.

[2] Ghaffarvand-Mokari, P., Werner, S. 2015. An acoustic description of spectral and temporal characteristics of Azerbaijani vowels. *Poznan Studies in Contemporary Linguistics*, 52(3), 503–518.

[3] Imamverdiyev, Y., Sukhostat, L. 2011. SVM based recognition of Azerbaijani vowels. In *5th International Conference on Application of Information and Communication Technologies (AICT)* IEEE, 1–4.

[4] Krocil, J., Machacek, Z., Koziorek, J., Martinek, R., Nedoma, J., Fajkus, M. 2018. Improved method of heuristic classification of vowels from an acoustic signal. *Turkish Journal of Electrical Engineering and Computer Sciences,* 26(6), 2900–2914.

[5] Biswas, A. S. T. I. K., Sahu, P. K., Bhowmick, A. N. I. R. B. A. N., & Chandra, M. (2014). Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition. *WSEAS Trans Syst*, *13*, 130–143.

[6] De Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., & Bourlard, H. 2004. Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America,* 116(3), 1781–1792.

[7] Korkmaz, Y., Boyacı, A., Tuncer, T. 2019. Turkish vowel classification based on acoustical and decompositional features optimized by Genetic Algorithm. *Applied Acoustics*, 154, 28–35.

[8] Ghaffarvand-Mokari, P., Mahdinezhad Sardhaei, N. 2020. Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives. *The Journal of the Acoustical Society of America*, 147(3), EL228–EL234.

[9] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

[10] Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5–32.

[11] Parks, R. W., Levine, D. S., Long, D. L. 1998. *Computational Neuroscience Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience*. MIT, Cambridge, MA.

[12] Tagliamonte, S. A., Baayen, R. H. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change*, 24(2), 135–178.

[1] The baseline was measured as the accuracy of a model that makes completely random guesses. The random baseline was the square of the proportion of each vowel in the data.

RIME CHANGES AND MERGERS IN BEIJING RETROFLEX SUFFIXATION: AN ACOUSTIC STUDY

Yutong Wang[1], Mitko Sabev[2]

[1]University of Cambridge, UK, [2]Saarland University, Germany
yw590@cantab.ac.uk, msabev@lst.uni-saarland.de

ABSTRACT

While Mandarin retroflex suffixation and resultant rime mergers have been approached from phonological and impressionistic phonetic angles, empirical studies remain scarce and inconclusive. This investigation is a step towards filling that gap. 10 Beijing Mandarin speakers were recorded and MANOVAs were performed for the effects of suffixation and rime identity on F_1 and F_2 frequencies of nuclear vowels. Results are reported for rimes that were previously claimed to merge: those with high, mid unrounded, and low nuclei. Overall, suffixed nuclei are retracted, while high vowels also tend to be lowered and low vowels to be raised. Our findings confirm complete acoustic neutralisation of the nuclear [a] rimes, where monophthongal rimes merge with those ending in [i] and [n]. On the other hand, suffixed [ie, y, u, ei] remain acoustically distinct within their merger groups. The differences, however, are small and likely to be perceptually undetectable, pointing to incomplete neutralisation.

Keywords: Beijing Mandarin, retroflex suffixation, vowel merger, incomplete neutralisation

1. INTRODUCTION

A body of phonological literature is devoted to Beijing retroflex suffixation (BRS) and the intricate rime mergers it leads to [1–4]. Generally, the retroflex suffix, [ɚ] or [ɻ], is directly attached to simple monophthongal rimes, and [ə] is inserted after front vowels [i, y, e]. In complex rimes (diphthongs and rimes with nasal codas), the second element of the rime is deleted if it is front or alveolar ([i] or [n]). If it is back (rounded) or velar ([u] or [ŋ]), it is not deleted completely, but leaves a trace of rounding or nasalisation, respectively. Within feature models such as [5, 6], this would mean that coronal post-nuclear segments are completely deleted, while dorsal ones are not.

There have been few empirical studies of the acoustic properties of nuclear vowels in suffixed vs unsuffixed rimes. [7] first visualised and compared the dynamic trajectories of F_1 and F_2 frequencies in different rimes, and their perceptual discrimination. The study confirmed that rimes in triplets such as [a - ai - an] merged after suffixation, and also found that suffixed [i] and [ie] did not merge while [ɤ] and [ei] did. However, the stimuli did not cover the full range of rimes, as only highly frequent suffixable lexemes were selected. Moreover, the tokens carried different tones, the effects of which on BRS realisation were not considered. [8]'s acoustic and articulatory study reached the same conclusion on the merger of triplets. However, here the results suggested that the front [e] in [ie] and [ye] did not contribute to maintaining contrast with the simple rimes [i] and [y], respectively, as it was dropped in suffixation. What is also different from [7]'s earlier findings is that [8] reports that [ɤ] did not merge with [ei] and [ən].

[9] postulated a 'Simultaneity of Compatible Articulations' principle which eliminates all segments incompatible with [+retroflex], although some features can be preserved. The loss of certain features, together with rime coalescence, results in various cases of rime merger, as in the triplet [a - ai - an], which in suffixation retain only the nuclear vowel of the rime. The phonological literature does not as a rule discuss the degree of overlap between different suffixed rimes, but rather assumes complete neutralisation and addresses issues such as whether suffixation results in overall rime rhotacisation or rhotic attachment with an intervening [ə].

In this paper, we report the results of a production experiment and address spectral changes in suffixed rimes, as well as their resultant acoustic neutralisation, from a quantitative perspective. Here we focus only on rimes that potentially merge with other rimes, and leave out of the discussion [uo] and rimes ending in [ŋ, u], which remain contrastive.

2. METHODS

2.1. Participants, procedure and recording

Ten participants (three male and seven female, aged 19–26), volunteered to participate in the production task in return for nominal payment. The test items

were displayed to the participants in carrier sentences on a computer screen, one pair at a time, in Chinese characters with a parallel *pinyin* transcription, in a quasi-random order. A MixPre-6M multitrack recorder with a Sennheiser ME64/K6 condenser microphone was used, recordings were digitised at a sampling rate of 48,000 Hz and stored as uncompressed single-channel WAV files.

2.2. Test items

Beijing Mandarin phonology allows 34 types of rime (excluding the suffix itself, [ɚ]). In a pilot production test, a native speaker (aged 23) produced all possible onset-rime combinations (398 in total), which were analysed acoustically to facilitate the selection of test items to be used in the actual experiment. To control for lexical tone, all selected items were in tone 55, which is a high level tone.

The test items consisted of 47 actual monosyllabic words and 89 phonologically well-formed monosyllabic non-words. The test items were shown in one of two alternative carrier sentences: (a) [wo^{214} ʈʂʐ^{55}tau _____ ʂʐ^{51} ʂən^{35}mɤ i^{51}sɻ] 'I know what _____ means', or (b) [wo^{214} pu^{51} ʈʂʐ^{55}tau _____ ʂʐ^{51} ʂən^{35}mɤ i^{51}sɻ] 'I don't know what _____ means'. Actual monosyllabic words were embedded in sentence (a), and non-words in sentence (b). To stimulate natural production, a question was visually presented, which was to be answered with sentence (a) or (b): 'Do you know what _____ means?'. The subjects were encouraged to speak in a casual style.

2.3. Analysis

Target word rimes were manually segmented in [10] based on the synchronised wideband spectrogram (Gaussian window shape, 0.004 s window length, 0.0001 s time step), waveform and auditory inspection; boundaries were determined by the presence of clear formant structure and a sharp change in intensity. In total, 2960 rime tokens were analysed. For each vowel, the first two formant frequencies were measured at the point of maximum F_1 frequency, using a Praat script [11].

Formant frequencies were then normalised using Lobanov's formant-intrinsic, vowel-extrinsic, speaker-intrinsic method [12]. Outliers, defined as values outside IQR by 1.5 times IQR, were removed (6.89% for F_1', 5.71% for F_2').

Vowel space areas and centroids in Figure 1 were computed using R package [13]. MANOVAs were performed on each unsuffixed–suffixed pair to measure the effect of suffixation on vowel quality. A second series of MANOVAs was used to quantify contrastiveness in potentially merging rime groups. In groups where significant differences emerged, additional comparisons were computed on subsets of the rimes, with Bonferroni correction for multiple testing applied to the significance level α. Pillai's trace (PT) was used as a metric of overall spectral difference. It is a common MANOVA statistic that has been used to quantify vowel merger in a growing body of work [14–24]. PT ranges from 0 to 1, with higher values indicating greater separation, and lower values greater overlap. Symmetric differences between the probability density functions (PDF) of each formant frequency ($\Delta F_1', \Delta F_2'$) in suffixed vs unsuffixed rimes were computed using R package [25].

3. RESULTS

An initial MANOVA was performed on all tokens with the normalised F_1', F_2' frequencies as response variables, and rime, suffixation, carrier sentence type (a or b), and speaker as predictor variables. Rime and suffixation both had a significant effect ($p < 0.0001$), while sentence type ($p = 0.7366$) and speaker ($p = 1$) did not, indicating that the use of alternative carrier sentences did not affect the results, and that the normalisation was highly effective.

Figure 1 shows the mean formant frequencies of nuclear vowels in the rimes under investigation. The area of the vowel space, defined as the area of the convex hulls enclosing the nuclear vowels in the $F_1' \times F_2'$ frequency space, was found to be 3.35 for unsuffixed and 1.16 for suffixed rimes, meaning that the suffixed space area is just over a third of the area of the unsuffixed space. Suffixation therefore results

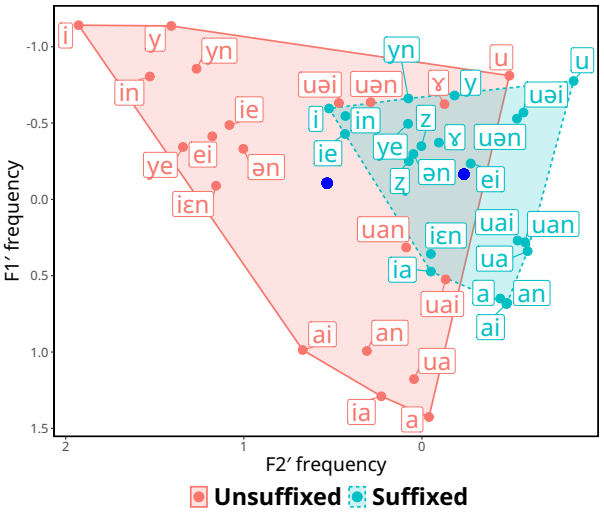


Figure 1: Mean F_1' and F_2' frequencies of rime nuclei; blue dots: vowel space centroids.

in a substantial contraction of the vowel space. The centroids of the two vowel spaces, plotted as blue dots, indicate that suffixation causes overall retraction, which is also observed in individual rimes. In addition, high unsuffixed nuclei are lowered, while low nuclei are raised with suffixation.

Table 1 reports the results of MANOVA comparisons, as well as symmetric differences between PDFs of the F_1' and F_2' frequencies of suffixed and unsuffixed nuclei. MANOVAs were significant for all nuclei except [u]. The overall change is small in [ɤ, ie, uai], with Pillai's trace values under 0.30. Moderate change is found in [uəi, uən, a, ia, ua], where PT ranges between 0.30 and 0.59. The remaining eleven nuclei all exhibit strong change, with PT scores above 0.60. Within the high front nuclei, F_1' frequency increases while F_2' decreases, thus resulting in mid-centralisation; for [yn] the change in F_2' is considerably stronger. In all mid nuclei except [ɤ] there is substantial decrease in F_2' (less pronounced in [ie]), whereas changes in F_1' are smaller. The monophthong [ɤ] shows appreciable increase in F_1' and little change in F_2' frequency. The low nuclei undergo reduction in both formant frequencies, with a stronger change in F_2' in most cases (not in [a, ia]).

The merger patterns of suffixed rimes are displayed as two-dimensional $F_1' \times F_2'$ PDFs in Figures 2 and 3. The results of MANOVAs comparing the nuclei of merging rimes are given in Table 2.

Rime	p	PT	$\Delta F_1'$	$\Delta F_2'$
i	0.0000	0.78	0.91	0.74
in	0.0000	0.68	0.64	0.77
y	0.0000	0.88	0.71	0.96
yn	0.0000	0.68	0.32	0.75
u	0.6553			
ɤ	0.0000	0.26	0.44	0.07
ie	0.0002	0.22	0.18	0.42
ye	0.0000	0.62	0.16	0.76
ei	0.0000	0.74	0.28	0.85
ən	0.0000	0.70	0.19	0.73
uəi	0.0000	0.33	0.22	0.59
uən	0.0000	0.31	0.20	0.58
a	0.0000	0.36	0.46	0.43
ia	0.0000	0.45	0.56	0.35
ua	0.0000	0.46	0.53	0.62
ai	0.0000	0.75	0.30	0.92
an	0.0000	0.62	0.23	0.76
iɛn	0.0000	0.73	0.39	0.64
uai	0.0013	0.17	0.18	0.50
uan	0.0000	0.69	0.14	0.79

Table 1: MANOVA comparisons of suffixed and unsuffixed nuclei; PT: Pillai's trace; Δ: symmetric difference between PDFs.

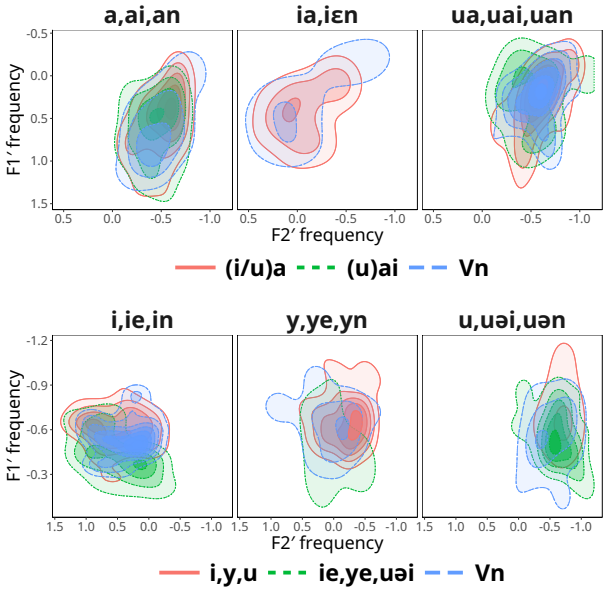


Figure 2: Suffixed rime mergers ($F_1' \times F_2'$ PDFs).

The top panels of Figure 2 show nearly total overlap in the rime groups with nuclear [a], and indeed the MANOVAs (Table 2) identified no significant differences in any of these groups: each of suffixed [a - ai - an], [ia - iɛn], and [ua - uai - uan] are merged. A somewhat lesser degree of overlap is visible in the bottom panels of Figure 2, and the corresponding MANOVAs indicate that one rime in each group is significantly different from the rest: [ie] in [i - ie - in], [y] in [y - ye - yn] and [u] in [u - uəi - uən]. There is strong overlap in the rime group in Figure 3, except that suffixed [ei] spans only about half of the F_2' range for the group. Indeed, the MANOVAs show that only [ei] is significantly distinct in this group.

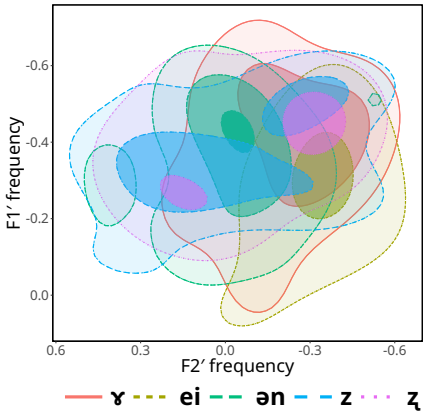


Figure 3: Suffixed [ɤ, ei, ən, z, ʐ] (PDFs).

Comparison	p	$\gtrless$	α	PT
a, ai, an	0.8982	$>$	0.050	
ia, iɛn	0.2684	$>$	0.050	
ua, uai, uan	0.6489	$>$	0.050	
i, ie, in	0.0000	$<$	0.050	0.23
i, in	0.1410	$>$	0.025	
ie, {i+in}	0.0000	$<$	0.025	0.21
y, ye, yn	0.0009	$<$	0.050	0.17
ye, yn	0.1526	$>$	0.025	
y, {ye+yn}	0.0009	$<$	0.025	0.12
u, uəi, uən	0.0001	$<$	0.050	0.22
uəi, uən	0.6042	$>$	0.025	
u, {uəi+uən}	0.0000	$<$	0.025	0.22
ɤ, ei, ən, ʐ̩, z̩	0.0004	$<$	0.050	0.16
ɤ, ən, ʐ̩, z̩	0.3519	$>$	0.025	
ei, {ɤ+ən+ʐ̩+z̩}	0.0000	$<$	0.025	0.12

Table 2: MANOVA comparisons of suffixed rimes; $p \leq \alpha$: significant; PT: Pillai's trace.

All significant differences, however, are small (PT $\leq$ 23) and thus may not be perceptually salient.

4. DISCUSSION AND CONCLUSIONS

Our results demonstrate that Beijing Mandarin rimes undergo a variety of significant spectral changes in retroflex suffixation, and that certain rime groups are completely merged after suffixation, while in other instances neutralisation is incomplete.

In most cases our findings are consistent with earlier phonetic treatments of BRS. In the high front nuclear vowels, viz. [i] and [y], suffixation brings about substantial changes in F_1' and F_2' frequencies in the direction of mid-centralisation. In mid (unrounded) nuclei, on the other hand, spectral changes are considerably smaller. In [uəi, uən], the pre-nuclear glide [u] appears to have an effect on the nuclei, which have lower F_2' frequencies than [əi, ən] in general, while suffixation results in additional retraction. A range of F_2' frequencies are found in the suffixed mid range of the vowel space (cf. [ie, in] vs [uəi, uən] in Figure 1), which refutes the frequent assumption of many phonological accounts that all non-low nuclei merge in an undifferentiated [ə]. In the low nuclei, we find considerable raising in most rimes, as well as comparable degrees of retraction. Pillai's trace for suffixed vs unsuffixed [uai] is exceptionally small for a low nucleus (0.18). Overall, the comparison of unsuffixed vs suffixed rimes reveals a pattern of global retraction, as well as moderate lowering and raising, respectively, for high and low nuclei. Suffixation also leads to dramatic shrinkage of the spectral vowel space, in terms of both formant frequencies.

As regards the neutralisation of contrast between rimes, our findings confirm complete merger in the rime groups with low nuclei, as MANOVAs yielded non-significant results.

The rime groups [i - ie - in], [y - ye - yn] and [u - uəi - uən] also tend to merge, although there is one rime in each which remains statistically distinct from the rest, namely [ie, y, u]. [ie] has a slightly higher F_1' frequency than the rest of its group when suffixed, while [y] and [u] have lower F_2' frequencies than the rest, which may be caused by greater retraction or rounding; suffixed [u] also has lower F_1' frequency than [uəi, uən]. These findings are at odds with the earlier claim that these groups are also fully merged [8]. Empirical studies such as [7, 26] found [ie, ye] to be distinct, while our results corroborate this only for the former rime.

A long-standing dispute concerning suffixed mid unrounded nuclei is whether [ɤ] merges with the rest of the group, [ei - ən - ʐ̩ - z̩]. Our results indicate that it does, which is consistent with [7]. We also found that, in this group, suffixed [ei] remains distinct from the rest, which is at variance with what [8] reported.

It should be noted, however, that since Pillai's traces for the rimes that were identified as distinct within their groups are relatively low, ranging from 0.12 to 0.22, these contrasts are likely to be perceptually undetectable: in all likelihood, we are dealing with cases of incomplete neutralisation, where statistically significant acoustic differences are not large enough to be perceptible.

Participants in this production experiment were asked to read out items in both their strongly contrastive (unsuffixed) and potentially merged (suffixed) forms, which may lead one to speculate whether the elicitation design could have given rise to untypically hyperarticulated renditions of the suffixed rimes. Such hyperarticulation, however, appears to be unlikely, for only four of the rimes emerged as significantly different in their groups, while the majority were clearly merged.

While the experimental results reported here contribute to closing a number of gaps in the literature on Beijing retroflex suffixation, many aspects of the process remain to be examined in the future. These include the effect of lexical tone, utterance-level prosody (e.g. focus, position in a phrase), word frequency, gender, speaking style, interpersonal variation, as well as how acoustic results are related to perception. [7], for example, pointed out that suffixed [i] and [ie] are more accurately perceptually distinguished in some tones than in others. The reasons for such interactions between lexical tone and rime merger remain largely understudied [4].

5. ACKNOWLEDGEMENTS

We are grateful to Beeke Muhlack for writing the *F1-max* Praat script [11], and to our two anonymous reviewers for helpful questions and suggestions.

6. REFERENCES

[1] Y. Lin, "Autosegmental treatment of segmental processes in Chinese phonology," Ph.D. dissertation, University of Texas at Austin, 1989.

[2] S. Duanmu, "A formal study of syllables, tones, stress and domain in Chinese language," Ph.D. dissertation, MIT, 1990.

[3] ——, *The Phonology of Standard Chinese (Phonology of the world's languages)*, 1st ed. Oxford: Oxford University Press, 2000.

[4] ——, *The Phonology of Standard Chinese (Phonology of the world's languages).*, 2nd ed. Oxford: Oxford University Press, 2007.

[5] G. N. Clements and E. V. Hume, "The Internal Organization of Speech Sounds," in *The Handbook of Phonological Theory*, J. A. Goldsmith, Ed. Oxford: Blackwell, 1995, pp. 245–306.

[6] A. Lahiri and H. Reetz, "Distinctive features: Phonological underspecification in representation and processing," *Journal of Phonetics*, vol. 38, pp. 44–59, 2010.

[7] L. Wang and N. He, "*Beijing Erhuayun de Tingbian Shiyan he Shengxue Fenxi* (the perception experiments and acoustic analyses on Beijing erhua)," in *Beijing Yuyin Shiyanlu (The Handbook of Beijing Phonetic Experiments)*, T. Lin and L. Wang, Eds. Beijing: Peking University Press, 1985, pp. 27–72.

[8] W. Lee, "A phonetic study of the erhua rimes in Beijing Mandarin," in *Proceedings of Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 1093–1096.

[9] Y. Chao, *A Grammar of Spoken Chinese*. Berkeley, Calif.; London: University of California Press, 1986.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/Computerprogramversion6.0., 2015.

[11] B. Muhlack, "F1-max," Praat script, 2022.

[12] B. Lobanov, "Classifications of Russian vowels spoken by different listeners," *Journal of the Acoustical Society of American*, vol. 49, pp. 606–608, 1971.

[13] R. Hijmans, C. Karney, E. Williams, and C. Vennes, "*geosphere*: Spherical Trigonometry," https://CRAN.R-project.org/package=geosphere, 2022.

[14] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *Journal of Phonetics*, vol. 34, pp. 458–484, 2006.

[15] L. Hall-Lew, "Improved representation of variance in measures of vowel merger," *Proceedings of Meetings on Acoustics*, vol. 9, no. 1, pp. 1–10, 2010.

[16] M. Sloos, "The reversal of the Bären–Beeren merger in Austrian Standard German," *Mental Lexicon*, vol. 8, no. 3, pp. 353–371, 2013.

[17] J. Nycz and L. Hall-Lew, "Best practices in measuring vowel merger," *Proceedings of Meetings on Acoustics: The Acoustical Society of America through the American Institute of Physics*, vol. 20, no. 1, pp. 1–19, 2014.

[18] M. Amengual and P. Chamorro, "The Effects of Language Dominance in the Perception and Production of the Galician Mid Vowel Contrasts," *Phonetica*, vol. 72, no. 4, pp. 207–236, 2015.

[19] M. E. L. Renwick and R. M. Olsen, "Analyzing dialect variation in historical speech corpora," *Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 406–421, 2017.

[20] M. Sabev and E. Payne, "A cross-varietal continuum of unstressed vowel reduction: evidence from Bulgarian and Turkish," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc, 2019, pp. 1164–1168.

[21] C. Diskin, D. Loakes, R. Billington, H. Stoakes, S. Gonzalez, and S. Kirkham, "The /el–æl/ merger in Australian English: acoustic and articulatory insights," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc, 2019, pp. 1764–1768.

[22] M. Sabev, "Spectral and Durational Unstressed Vowel Reduction: an acoustic study of monolingual and bilingual speakers of Bulgarian and Turkish," Ph.D. dissertation, University of Oxford, 2020.

[23] ——, "Unstressed vowel reduction and contrast neutralisation in Western and Eastern Bulgarian: a current appraisal," *Journal of Phonetics*, forthcoming.

[24] J. A. Stanley and B. Sneller, "Sample size matters in calculating Pillai scores," *Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 54–67, 2023.

[25] M. Pastore, "Overlapping: a R package for estimating overlapping in empirical distributions," *The Journal of Open Source Software*, vol. 3(32), p. 1023, 2018.

[26] Y. Wang, "Contrast neutralisation in Beijing retroflex suffixation: an acoustic study," MPhil dissertation, University of Cambridge, 2022.

LOW DENSITY ASSAMESE ALVEOLAR STOPS SHOW GREATER COARTICULATORY RESISTANCE COMPARED TO LABIALS AND VELARS

Abhilasha Thakuria[1], Pamir Gogoi[2], Indranil Dutta[1]

School of Languages and Linguistics, Jadavpur University[1], Living Tongues Institute of Endangered Languages[2]

abhilashathakuria1@gmail.com, pamirgogoi11@gmail.com, indranildutta.lnl@jadavpuruniversity.in

ABSTRACT

Coarticulatory Resistance (CR) in consonants is affected by density of segments in an inventory. Sparse consonantal systems allow consonants to coarticulate freely with adjoining vowels. Assamese, an Indo-Aryan (IA) language, contains alveolar stops that result from a historical merger between the dental and retroflex stops of the IA language family. Due to the low density of contrast in the coronal place of articulation (POA), we expect Assamese alveolar stops to exhibit greater acoustic variation. Slopes from first order F2 Locus Equations suggest that alveolars in a VCV context exhibit greater CR compared to velars and labials. In fact, slopes drawn between F2 at 25% and F2 at 5% of the vowel are also found to be significantly different between alveolars, labials and velars. We argue that consonantal systems seek articulatory equilibrium by selecting CR patterns that are governed by intrinsic articulatory complexity and motor demands, not just density.

Keywords: Coarticulatory resistance, Locus Equations, Contrast, Assamese

1. LOCUS EQUATIONS, COARTICULATORY RESISTANCE AND DENSITY OF CONTRASTS

1.1. Locus Equations

Systematic acoustic variation in stop place of articulation, irrespective of the effect of contextual vowel was first observed by [1] using locus equations (LE). Formant loci were found to be invariant acoustic cues to place of articulation, especially in the case of coronal stops where F2 transition shapes do not provide cues to place of articulation unambiguously [2]. Locus equations are linear regressions of F2 frequency at the vowel midpoint and onset. It has been found that locus equations are not just an indicator of the stop place of articulation but they also predict the degree of coarticulation in VCV sequences ([3],[4],[5]. It has been employed to predict coarticulation in both linguistic and pathological variation in speech [6], [7], [8]. First-order locus equation is represented by the formula

$$F2_{onset} = k \times F2_{vowel} + c \quad (1)$$

where k is the slope and c is the y-intercept [9].

In 1 above, k, the slope of the linear equation is also known to provide a measure of coarticulatory resistance [10]. Steeper slopes correspond to lower coarticulatory resistance, while shallower slopes correspond to higher coarticulatory resistance.

1.2. Slopes of Locus Equations as a measure of Coarticulatory Resistance

Coarticulatory resistance, as measured through slopes of the First-order locus equations, provide a unified measure with which it is possible to understand how nature of contrast, size, shape and density of phonetic inventories [11, 12] affect coarticulation and by extension, variable coarticulatory resistance. Neighborhood frequencies also affect coarticulatory patterns such that higher neighborhood frequencies lead to higher degree of coarticulation [13]. Measuring coarticulation of Catalan consonants by the standard deviation of F2 values at the onset of the consonant, in relation to its following vowels has also been a useful method to quantify extent of coarticulation [14]. A low standard deviation would imply high coarticulatory resistance, because the onset would be less affected by the following vowel [14]. This relation between coarticulatory resistance and locus equation slopes has also been shown for American English consonants [15]. A steeper slope of locus equations as a result of higher degree of coarticulation was also seen in computer simulated speech [16],[9].

1.3. Density of Contrast

Coarticulatory resistance can be affected by articulatory motor constraints and by density of contrast in the language's segmental inventory. [4] observed that alveolars in a VCV context exhibit greater coarticulatory resistance compared to velars and labials in Swedish. [11][12] studied the extent of coarticulatory resistance in CV syllables between languages that differed in the density of contrast. The results showed that vowels with a higher neighborhood density of contrast had a higher resistance to anticipatory coarticulation than vowels with a low density of contrast. Coarticulatory propensity, resistance, therefore is a function of articulatory motor constraints, in terms of articulatory demands, degree of overlapping gestures, size, shape of the phonetic inventory, and density of contrasts.

2. DENSITY OF CONTRAST IN ASSAMESE

We investigate the relative effect of these factors by analyzing coarticulatory resistance in Assamese, an Indo-Aryan (IA) language spoken in India. Assamese has a low density of contrast in the coronal place of articulation when compared with other IA languages. The dental-retroflex contrast in IA languages, historically merged into an alveolar stop in Assamese[17]. Therefore, based on the results of [11][12], Assamese should exhibit lower coarticulatory resistance in the alveolar stops, as it does not have to maintain the categorical distinction between dental and retroflex stops, which is a feature in most IA languages. On the other hand, higher coarticulatory resistance in alveolar stops compared to velar and bilabials, would indicate that relative articulatory complexity and motor demands mitigate or at least are a greater determiner of coarticulatory resistance in Assamese.The effect of preceding and following vowels on syllable-initial and syllable-final /t/, /t^h/, /n/, /d/, and /d^h/ were studied by [18]. It was found that while these segments are distinctly denti-alveolar in the presence of vowels which are high and front, they were apico-alveolar when the preceding or following vowel is a back vowel. [19]'s data shows the voiceless alveolar plosive acts like the English t. Here, it is suggested that the alveolar may be due to the influence of the Bodo languages in which dental consonants are pronounced as semi-cerebrals as in English. In our study, we employ first-order locus equations in the V_1CV_2 context, where V_1 is [a] and V_2 that varies between [a, i, u]. Deriving locus equations from F2 values at 15%, 25%, and 50% we investigate the nature of the slopes at various temporal landmarks of the vowel. This is motivated by at least one study by [20] that shows that locus equation slopes as early as 25% in to the vowel are sufficient to cue place of articulation in American English. In the following section we describe the methods in greater detail.

3. METHOD

3.1. Participants and Stimuli

Speech samples from five male and five female native speakers of Assamese were analyzed for this study. The stimuli consisted of words in [...aCV_2...] context, where the C is a labial, alveolar or velar stop, while the following V_2 is one of the corner vowels /a, i, u/. The words were presented in a sentence frame of 'mɔi tat _ buli lik^{h}a dek^{h}ilu.' The word list consisted of two words from each VCV combination, making a list of 36 words (6 Consonants $\times$ 3 Vowels $\times$ 2 unique words). For instance ʤapa, ata, aka. The list was read 4 times by each participant. The first iteration was not analysed for this study.

3.2. Procedure and Analysis

Recording was conducted in a sound attenuated booth with a head mounted microphone connected to a solid state recorder. The sentences were presented to the participants on a computer screen, in a random manner and recorded at a sampling frequency of 44100 Hz. The data was manually annotated, carefully marking each phoneme boundary. A PRAAT script was used to generate the F2 frequency at 5%, 15%, 25% and 50% of the vowel. The F2 values generated were then speaker normalized using Lobanov Vowel normalization method [21]. These normalized F2 values where fitted into 1. For each individual speaker and for each place of articulation, the coefficient value of the slope was obtained by running a linear regression model. We then perform ANOVA followed by a Tukey HSD test to analyze the statistical significance of the differences in the coefficient values across place of articulation. The following ANOVA model was used for the statistical analysis:

$$(2) \quad Coefficient_{vowelpoint} \sim POA + Sex$$

In 2 above, *vowel point* is the coefficient value at either 15%, 25% or 50% of V_2.

4. RESULTS

The overall results show that velars have the steepest slope, followed by bilabial and then by alveolars. The slope of each place of articulation determines the degree of coarticulatory resistance, that is, a steeper slope indicates that there is a greater change at the beginning of the vowel for per unit change in the middle. The slopes measured at 15%, 25% and 50% are presented in section 4.1, 4.2 and 4.3 respectively.

4.1. CR at 15% of the vowel

In Figure 1, normalized F2 values at 5% of the vowel are plotted against F2 values at 15% of the vowel. We observe that velars and labials have steeper slopes compared to alveolars. The slope values of the regression models at 15% of the vowel plotted for each place of articulation in Figure 2 correspond to the observation in 1 by showing higher values for velars. The velar slope is significantly steeper compared to the alveolar ($p < 0.001$, 95% C.I. = [0.063, 0.206]). No significant difference between alveolar and bilabial slopes is found at 15% of the vowel.

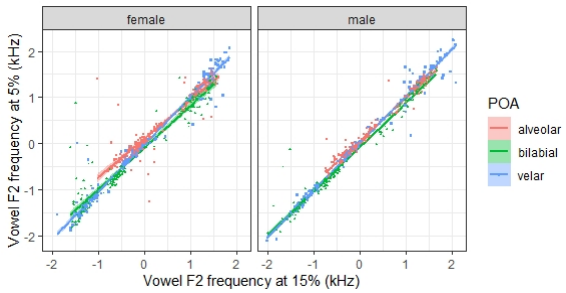


Figure 1: Linear regression of F2V and F2C at 15%

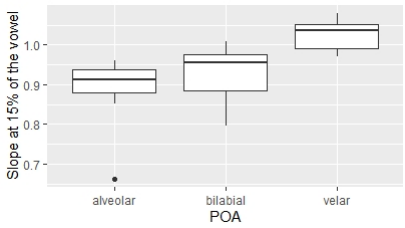


Figure 2: Coefficients at 15% of the vowel

4.2. CR at 25% of the vowel

The F2 values at 25% of the vowel plotted in Figure 3 shows similar pattern as observed at 15% of the vowel, but the difference between the regression slopes for alveolar with velar and labial increases. The coefficients of the F2 slope plotted in Figure 4 show lowest values for alveolars and highest values for velar. Statistical analysis shows a significant difference between all the three place of articulation categories - bilabial-alveolar $p < 0.05$, 95% C.I. = [0.188, 0.025]; alveolar-velar $p < 0.005$, 95% C.I. = [0.167, 0.33]; bilabial-velar, $p < 0.05$, 95% C.I. = [0.059, 0.222]. These results are similar to [20] who found similar early stability in the place of articulation determinant function of the slopes than what has been proposed in standard locus equation models in [4, 22]. The order of the slopes velar > labial > alveolar correspond to the inverse relationship that slopes have with coarticulatory resistance, namely, alveolars greater coarticulatory resistance compared to labials and velars, in that order.

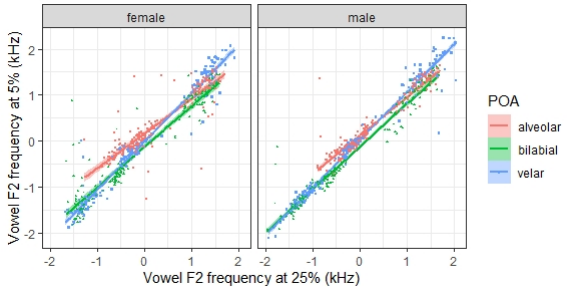


Figure 3: Linear regression of F2V and F2C at 25%

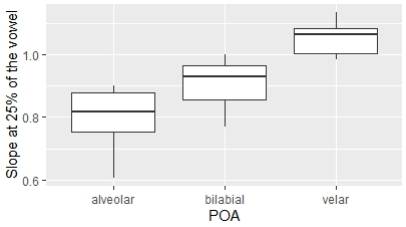


Figure 4: Coefficients at 25% of the vowel

4.3. CR at 50% of the vowel

F2 values at the vowel midpoint are plotted in Figure 5. We observe that the slope for alveolar is shallower compared to the slopes with F2 measures at 15% and 25%. The same is observed in Figure 6 where the coefficient of velar slopes are higher than both bilabials and alveolars; coefficients of bilabial slopes are higher than alveolars. Statistical analysis show significant difference between all the places of articulation (bilabial-alveolar $p < 0.001$, 95% C.I. = [0.085, 0.242]; velar-alveolar $p < 0.001$, 95% C.I. = [0.271, 0.428]; velar-bilabial $p < 0.001$, 95% C.I. = [0.106, 0.264]).

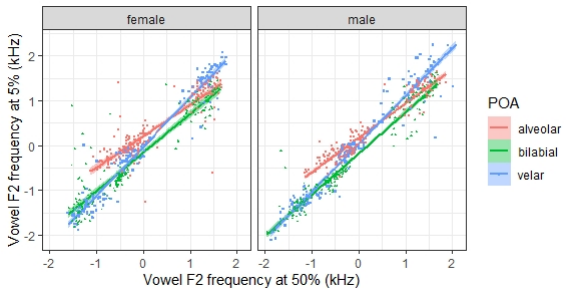


Figure 5: Linear regression of F2V and F2C at 50%

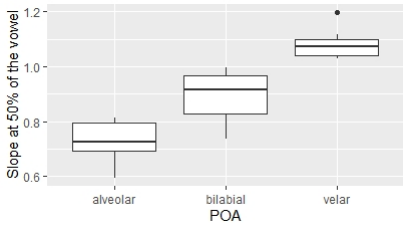


Figure 6: Coefficients at 50% of the vowel

Comparing the slopes at 15%, 25% and 50% of the vowel, we observe a gradual increase in the difference between the slopes for each place of articulation. Significant differences in F2 slopes between the three places of articulation are found between alveolar, labial and velar at 25% and 50% of the following vowel, with the alveolars showing the lowest slope values. These results suggest that the alveolars resist coarticulation to the greatest degree in Assamese, in spite of a purported merger of the Indo-Aryan dental-retroflex contrast in Assamese. While there is no reason to think that the merger would imply an articulatory variation between these two places of contrast and a subsequent articulatory stabilization of the dental-retroflex contrast into an alveolar, there is perhaps a greater likelihood that historically, the Assamese substratum mapped an already existing alveolar to the Indo-Aryan dentals and retroflexes. Alternatively, we could interpret our results to reflect the intrinsic articulatory complexity of the alveolars, compared to the velars and labials in terms of greater motor demands on the tongue tip and dorsum. The latter interpretation is conceivable even with the possible historical merger of the dental-retroflex in the inherited Indo-Aryan substrate of Assamese.

5. DISCUSSION

We examine locus equation slopes as a function of both place of articulation discrimination and coarticulatory resistance in Assamese - a modern Indo-Aryan language in which a historical merger of the Indo-Aryan dental-retroflex contrast has been reported. Our results show that the Assamese alveolars are discriminated quite well when compared to the labials and velars. The Assamese alveolars also show shallower slopes at 25% and 50% of the vowel, suggesting greater coarticulatory resistance. Our results correspond to the coarticulatory pattern seen in an earlier study on coarticulatory resistance of American English consonants, which found that bilabials and velars have lower coarticulatory resistance compared to alveolars [15]. Although coarticulatory resistance is known to be affected by the density of contrast in the segmental inventory of a language, we do not observe any correlation between coarticulatory resistance in Assamese and its sparse presence in the coronal place of constriction compared to its related languages. The possible explanation provided for the low coarticulatory resistance of the labial consonant is that the tongue gestures needed for the vowel can be overlapped substantially with the lip gestures due to the relative independence of these articulators. The low coarticulatory resistance of the velar has been been seen in a number of studies. For instance, it has been observed that the place of articulation of the velar /g/ varies in response to the frontness of the contextual vowel [23].

It has also been demonstrated that the locus equation slope for /b/ is steep, where as the slope for alveolar is flatter [24]. The order of coarticulatory resistance previously seen across place of articulation of stops is labial>velar>alveolar [14]. Another explanation suggested by [5] for a considerably steep slope for velars in some languages is due to the lack of a similar consonant with a place of articulation close to the velar leading to greater confusability [5]. This explanation doesn't quite hold for Assamese due to the presence of a voiceless velar fricative; the only Indo-Aryan language that possesses such a consonant. Given the greater coarticulatory resistance of the Assamese alveolars, we claim that the historical merger of the Indo-Aryan dental-retroflex contrast stabilized into an articulatorily complex configuration as seen in most alveolars. In that sense, the Assamese alveolars could either be conceived of as Bodo type alveolars or strictly governed by motor constraints and not determined by the density or sparsity of the coronal place of articulation. In further research, we will extend our model to include the nasals, the velar and the coronal fricatives, that are available in the inventory of Assamese in addition to fitting Bayesian regression models to tease out any subject-wise variation.

6. REFERENCES

[1] B. Lindblom, "Spectrographic study of vowel reduction," *The journal of the Acoustical society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[2] P. Delattre, F. S. Cooper, A. M. Liberman, and L. Gerstman, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 26, no. 1, pp. 137–137, 1954.

[3] D. Krull, "Acoustic properties as predictors of perceptual responses: A study of swedish voiced stops," Ph.D. dissertation, Department of Linguistics, University of Stockholm, 1988.

[4] ——, "Consonant-vowel coarticulation in spontaneous speech and in reference words," *Speech Transmission Laboratory Quarterly Progress Status Reports, Royal Institute of Technology, Stockholm, Sweden*, vol. 1, pp. 101–105, 1989.

[5] C. A. Fowler, "Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation," *Perception & Psychophysics*, vol. 55, no. 6, pp. 597–610, 1994.

[6] H. M. Sussman, D. Fruchter, J. Hilbert, and J. Sirosh, "Linear correlates in the speech signal: The orderly output constraint," *Behavioral and brain sciences*, vol. 21, no. 2, pp. 241–259, 1998.

[7] M. Robb and M. Blomgren, "Analysis of f2 transitions in the speech of stutterers and nonstutterers," *Journal of fluency disorders*, vol. 22, no. 1, pp. 1–16, 1997.

[8] M. Baillargeon, A. McLeod, D. E. Metz, N. Schiavetti, and R. L. Whitehead, "Preservation of second formant transitions during simultaneous communication:: A locus equation perspective," *Journal of communication disorders*, vol. 35, no. 1, pp. 51–62, 2002.

[9] S. Chennoukh, R. Carré, and B. Lindblom, "Locus equations in the light of articulatory modeling," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2380–2389, 1997.

[10] K. Iskarous, C. A. Fowler, and D. H. Whalen, "Locus equations are an acoustic expression of articulator synergy," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2021–2032, 2010.

[11] S. Manuel, "Coarticulation: Cross-language studies: relating language-particular coarticulation patterns to other language-particular facts," 1999.

[12] ——, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *Journal of the Acoustical Society of America*, vol. 88, pp. 1286–1298, 1990.

[13] R. Scarborough, "Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation," *Journal of Phonetics*, vol. 41, no. 6, pp. 491–508, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447013000557

[14] D. Recasens, "Coarticulatory patterns and degrees of coarticulatory resistance in catalan cv sequences," *Language and Speech*, vol. 28, no. 2, pp. 97–114, 1985.

[15] C. A. Fowler and L. Brancazio, "Coarticulation resistance of american english consonants and its effects on transconsonantal vowel-to-vowel coarticulation," *Language and Speech*, vol. 43, no. 1, pp. 1–41, 2000.

[16] B. Lindblom and H. M. Sussman, "Articulatory and acoustic bases of locus equations," *FONETIK 2004*, p. 8, 2004.

[17] C. P. Masica, *The Indo-Aryan Languages*. Cambridge University Press, 1993.

[18] S. Mahanta, "Assamese," *Journal of the International Phonetic Association*, vol. 42, no. 2, pp. 217–224, 2012.

[19] B. Kakati and G. Goswami, *Assamese, Its Formation and Development: A Scientific Treatise on the History and Philology of the Assamese Language, Being a Thesis Approved for the Ph.D. Degree of the Calcutta University in 1935*. LBS Publication, 1995. [Online]. Available: https://books.google.co.in/books?id=OTSrmQEACAAJ

[20] A. E. Rhone and A. Jongman, "Modified locus equations categorize stop place in a perceptually realistic time frame," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. EL487–EL491, 2012.

[21] B. M. Lobanov, "Classification of russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, 1971.

[22] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews, "An investigation of locus equations as a source of relational invariance for stop place categorization," *The Journal of the Acoustical Society of America*, vol. 90, no. 3, pp. 1309–1325, 1991.

[23] S. E. Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[24] B. Lindblom and H. M. Sussman, "Dissecting coarticulation: How locus equations happen," *Journal of Phonetics*, vol. 40, no. 1, pp. 1–19, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447011000933

THE EFFECTS OF VOICED STOPS ON ADJACENT VOWEL DURATION IN JAPANESE

Shoji Ishibashi[1], Shonosuke Koya[1&2]

[1]Fukuoka University Graduate School, [2] JSPS Research Fellowship for Young Scientists
ld220501@cis.fukuoka-u.ac.jp, ld220502@cis.fukuoka-u.ac.jp

ABSTRACT

Vowel duration is longer before voiced than voiceless consonants (voicing effect) and it is known as universal phenomenon across languages. However, it is still controversial whether the effect is consistently seen in languages which have vowel length contrast (long vs. short). Furthermore, while the effects of consonants on preceding vowels have often been examined, there have been few studies of those on following vowels. To address these issues, the current study conducted a production experiment in which native speakers of Japanese uttered disyllabic nonce words (CV_1CV_2 or CV_1:CV_2). Statistical analysis suggests three main results as follows: (i) a voicing effect is observed regardless of preceding vowel length; (ii) duration of the following vowel is likely to be longer when the preceding consonant is voiced than when it is voiceless; (iii) V_2 duration tends to be shorter when V_1 is phonologically long than when it is phonologically short.

Keywords: voicing effect, vowel length, Japanese, speech production

1. INTRODUCTION

A voicing distinction is one of the most common phonetic contrasts in languages, and various studies have explored voiced contrasts for decades [18, 11]. With respect to the relationship between voicing contrast and vowel duration, it is widely known that vowels followed by voiced obstruents are durationally longer than when followed by voiceless obstruents in various languages (French, Russian, Korean [4] and English [8]). In English, for example, the vowel duration in *tab* is longer than that in *tap*. This phenomenon has been called the 'pre-fortis clipping' [17] or 'voicing effect' [14] (hereafter, voicing effect). Since the effect is confirmed in several languages, it can be a cross-linguistically common tendency.

While the voicing effect is confirmed in several languages, it is also reported that a language-specific phonological system may reduce the effect. In [12], the voicing effect is not observed in Czech and the study implied that the effect may blur the phonological distinction of vowel length as Czech has short and long vowels phonemically (i.e., the voicing effect is not seen or is extremely weak in a language which has vowel length contrasts). A similar tendency is observed in Saudi Arabic, a language with phonological vowel length contrast [6].

Although several studies showed example cases in which the voicing effect may be reduced by a language-specific phonological system, there are also some studies which have different results. For instance, [1] reported that the voicing effect was confirmed in Lebanese Arabic even though it has a phonemic vowel length contrast. Thus, it is still unclear whether the degree of voicing effect is influenced by a language-specific system. To clarify this question, the voicing effect should be explored using more languages with phonological vowel length contrasts.

Japanese can be a suitable language to elucidate the relationship between the voicing effect and vowel length contrast. Japanese has phonemic vowel length distinction (e.g., /i/ 'stomach' vs. /i:/ 'good'), and the voicing effect in Japanese is controversial [15, 20]. [15] analyzed a set of four minimal pairs of words (e.g., [k**a**ka] and [k**a**ga]) uttered by ten native speakers of Japanese and reported that vowels are produced with 20% longer duration before a voiced obstruent than before a voiceless obstruent. On the other hand, [20] examined the voicing effect using the Corpus of Spontaneous Japanese and the results showed that the voicing effect was not observed. Additionally, while [15] examined the voicing effect on only short vowels, [20] examined the voicing effect on short and long vowels, and [20] reported that the voicing effect was not observed regardless of the vowel length. Even though these two studies have methodological differences, it has not yet been fully verified whether voicing effect can be reliably observed in Japanese.

In addition to the voiced obstruent effect on a preceding vowel, the effect on a following vowel is also confirmed in several languages. [13] examined minimal pairs of disyllabic words (e.g., $C_1V_1C_2V_2$; [tat**a**] and [tad**a**]) and reported that the vowel duration after voiced obstruents is longer than after voiceless ones (hereafter, called post-voicing effect) in Japanese. [1] conducted a production experiment with disyllabic nonce words and a similar phenomenon is observed in Lebanese Arabic. Moreover, the data in

[5] implies that Italian also has the post-voicing effect. Considering these reports, the post-voicing effect might be a cross-linguistically common tendency as well as the voicing effect. However, the following points should be noted. Since [1] did not consider the vowel length contrast in their analysis, it is possible that the post-voicing effect is observed only when the vowel in the first syllable is short, but not when long and vice versa. Similarly, [13] did not examine the post-voicing effect when V_1 was phonologically long. Thus, whether the post-voicing effect is observed regardless of the phonological length of the preceding vowel is still a controversial issue and should be examined in more detail.

We also need to refer to the tendency that vowel duration after geminates or a long vowel tends to be shorter than after singletons or short vowel in Japanese [3, 7, 10]. [10] conducted a production experiment using disyllabic words ($C_1V(:)_1C_2V_2$) and reported that vowel duration in the syllable following the long vowel tends to be shorter than following the short one. As [10] used only voiceless stops for word-internal consonants in their stimuli, however, it should also be explored with voiced stops.

Based on the results of the previous studies, the present study addressed the following three research questions by conducting a production experiment in which speakers uttered disyllabic nonce words ($C_1V(:)_1C_2V_2$):

(1) Is the voicing effect observed in not only short vowels but also long vowels?
(2) Is the post-voicing effect observed regardless of the phonological length of the preceding vowel?
(3) Is the V_2 duration shortened when V_1 is a long vowel regardless of the voicing of the word-internal consonant?

2. METHOD

2.1. Participants

Our ten participants, five male and five female, were all native speakers of Japanese with a mean age of 22.2 (range: 18-28). The experiment was conducted at the Phonetic Laboratory of Fukuoka University. All participants were first introduced to the purpose of the experiment and the procedure before signing a consent form. Demographic data, such as age, sex, and birthplace, were described before the experiment started. Participants received monetary compensation.

2.2. Procedure

Materials used in the present study were two-syllable words of the form '$pV_1(:)CV_2$' (V_1 = a, e, o; C = p, t, k, b, d, g; V_2 = a, e, o); the total number of words was 108 (3 vowel qualities of V_1 (/a/, /e/, /o/) × 2 vowel lengths of V_1 (short, long) × 3 types of C_2 (labials, alveolar, velar) × 3 vowel qualities of V_2 (/a/, /e/, /o/). These words were written in *katakana* and were embedded in a carrier phrase /karewa ___ to itta/ ("he said ___"). The participants read the 108 words within a carrier phrase five times, and a total of 5400 tokens (108 words × 5 repetitions × 10 participants) were recorded using a digital audio recorder (Sony PCM-D100) in a soundproof room. 52 tokens from the entire data sets were discarded due to errors of utterances; thus, a total of 5348 tokes were used in the present analysis.

2.3. Analysis

Five acoustic measures were made for each token: the VOT of C_1 (**p**V(V)CV), V_1 duration (p**V(V)**CV), the C_2 closure duration (pV(V)**C**V), the VOT of C_2 (pV(V)**C**V), and the V_2 duration (pV(V)C**V**). The durations were segmented by the visual inspection of the waveform and spectrogram with reference to periodicity and formant energy bands using Praat (ver. 6.1.51) speech analysis software [2], following the segmentation criteria used in [9] and [10]. The criteria are as follows: the VOT of C_1 and C_2 were measured from the onset of the release burst of C_1 and C_2 to the onset of the voicing of V_1 and V_2, respectively. The durations of V_1 and V_2 were determined as the points where the formants abruptly appeared, and the offset of V_1 and V_2 was determined to be the points where the formants abruptly ceased. The C_2 closure duration was measured between the offset of V_1 and the onset of the release burst of C_2.

3. RESULT

Word example	C_1 VOT	V_1	C_2 closure	C_2 VOT	V_2
papo	30.4	58.4	70.6	19.1	69.0
pabo	36.9	76.3	41.2	18.4	83.5
paapo	28.8	173.7	79.7	22.8	58.4
paabo	35.3	201.9	47.1	16.1	72.0

Table 1: Mean duration (ms) for C_1 VOT, V_1, C_2 closure, C_2 VOT, and V_2

Table 1 summarizes the mean duration of C_1 VOT, V_1, C_2 closure, C_2 VOT, and V_2. The mean duration of V_1 and V_2 is illustrated in Figure 1 and 2 above. These show that the voicing effect and post-voicing effect is likely to be observed.

In order to determine (i) whether or not V_1 duration systematically varies depending on the voicing of C_2, (ii) whether or not V_2 duration varies depending on the voicing of C_2 and (iii) whether or not V_2 duration varies depending on the V_1 length, a mixed ANOVA was conducted separately for V_1 and V_2 duration. The fixed factors were V_1 length (phonologically short vowel, long vowel), voicing of

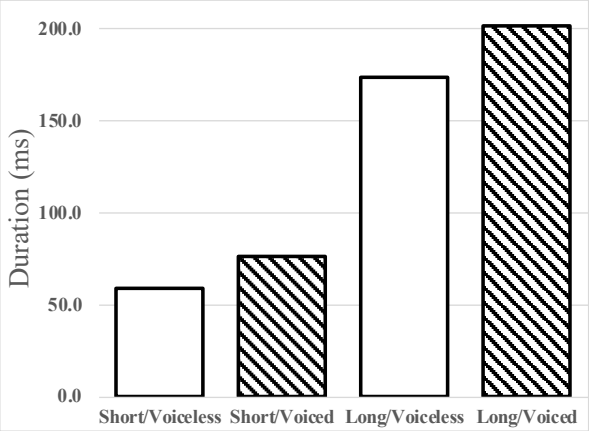


Figure 1: Mean duration of V_1 for each length and voicing of C_2

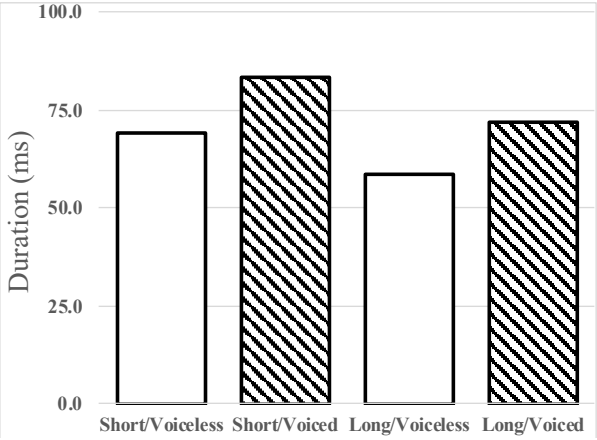


Figure 2: Mean duration of V_2 for each length and voicing of C_2

C_2 (voiceless, voiced), V_1 quality (/a/, /e/, /o/), C_2 types (labial (/p/, /b/), alveolar (/t/, /d/), velar (/k/, /g/)), and V_2 quality (/a/, /e/, /o/); random factors were participants and repetitions. Results for interactions other than V_1 length × C_2 voicing are omitted due to space limitations.

3.1. V_1 duration

A mixed ANOVA was conducted with V_1 duration as the dependent variable and V_1 length, V_1 quality, voicing of C_2, C_2 types, and V_2 quality as fixed factors; random factors were participants and repetitions. All main effects except for V_2 quality were significant, and interaction between C_2 voicing and V_1 length were significant (F (1, 5301) = 23.581, $p < 0.01$). Since the interaction between V_1 length and C_2 voicing were significant, statistical analyses were conducted for each vowel length (short and long vowel).

The V_1 (short vowel) duration is statistically longer when C_2 is voiced than when C_2 is voiceless (C_2 voicing: F (1, 2656) = 2444.457, $p < 0.01$).

The V_1 (long vowel) duration is statistically longer when C_2 is voiced than when C_2 is voiceless (C_2 voicing: F (1, 2650) = 1466.972, $p < 0.01$).

3.2. V_2 duration

A mixed ANOVA was conducted with V_2 duration as the dependent variable and V_1 length, V_1 quality, voicing of C_2, C_2 types, and V_2 quality as fixed factors; random factors were participants and repetitions. All main effects were significant, and interaction between C_2 voicing and V_1 length were significant (F (1, 5301) = 11.605, $p < 0.01$). Since the interaction between V_1 length and C_2 voicing was significant, statistical analyses were conducted for each vowel length (short and long vowel).

3.2.1. Vowel duration after voiceless/voiced obstruents

A mixed ANOVA was conducted with V_2 duration (with short V_1) as the dependent variable and voicing of C_2, V_1 quality, C_2 types, and V_2 quality as the fixed factors; random factors were participants and repetitions. The V_2 duration (with short V_1) is statistically longer when C_2 is voiced than when C_2 is voiceless (C_2 voicing: F (1, 2656) = 1688.961, $p <$ 0.01).

A mixed ANOVA was conducted with V_2 duration (with long V_1) as the dependent variable and voicing of C_2, V_1 quality, C_2 types, and V_2 quality as the fixed factors; random factors were participants and repetitions. The V_2 duration (with long V_1) is statistically longer when C_2 is voiced than when C_2 is voiceless (C_2 voicing: F (1, 2650) = 1381.640, $p <$ 0.01).

3.2.2. Vowel in the syllable after a phonologically short/long vowel

A mixed ANOVA was conducted with V_2 duration (with voiceless C_2) as the dependent variable and V_1 length, V_1 quality, C_2 types, and V_2 quality as the fixed factors; random factors were participants and repetitions. The V_2 duration is statistically shorter when V_1 length is long than when V_1 length is short (V_1 length: F (1, 2656) = 789.771, $p < 0.01$).

A mixed ANOVA was conducted with V_2 duration (with voiced C_2) as the dependent variable and V_1 length, V_1 quality, C_2 types, and V_2 quality as the fixed factors; random factors were participants and repetitions. The V_2 duration is statistically shorter when V_1 length is long than when V_1 length is short (V_1 length: F (1, 2650) = 1087.816, $p < 0.01$).

3.3. Summary of the results

The results of the present study can be summarized as follows:

(i) Vowel duration is longer when the following consonant is voiced than when it is voiceless regardless of the vowel length (voicing effect).

(ii) Vowel duration is longer when the preceding consonant is voiced than when it is voiceless regardless of the preceding syllable's vowel length (post-voicing effect).

(iii) In disyllabic words ($C_1V_1C_2V_2$), vowel duration is shorter when the phonological length of V_1 is long than when V_1 is short regardless of the voicing of C_2.

4. DISCUSSION AND CONCLUSION

In order to examine the effects of voiced stops on adjacent vowel duration, the present study conducted a production experiment in which ten native speakers of Japanese uttered disyllabic words.

The first question of this study was whether the voicing effect is seen in Japanese. The results showed that the voicing effect is observed in Japanese regardless of the vowel length. In the discussion of the voicing effect in Japanese, the reports by previous studies were contradicted. While [15] indicated that the voicing effect is seen in Japanese, [20] said that it is not. In light of the above, our result seems to correspond to the result of [15]. This means our result contradicts the implication of [12], i.e., [12] implied that the voicing effect is not seen or is extremely weak in a language which has vowel length contrast, but our result showed the voicing effect in Japanese which does have this contrast.

Although the results of the present study and [20] are contradicted, it should be noted that there is a methodological difference between [20], a study using a corpus, and the present study, which is an experimental study. Given the difference in the characteristics of corpus and experimental data [19, 16], methodological differences may cause such conflicting results between the [20] and the present study, and, thus, it is not necessarily true that the result of [20] is not valid. Further study is required.

Our second interest was the voicing effect on following vowel (post-voicing effect). The results showed that the post-voicing effect was observed in Japanese regardless of the preceding vowel length. In light of the post-voicing effect, [13] reported that the vowel duration after voiced obstruents is longer than after voiceless ones. However, the research was limited to only the case when the preceding vowel was phonologically short. The current study confirms that the post-voicing effect is seen regardless of the phonological length of the preceding vowel. To clarify to what extent the post-voicing effect is a phonetic tendency, other languages should be explored. Furthermore, although we only used the vowel length contrast on V_1, a relationship between V_2 length contrast and the post-voicing effect is also interesting. These questions should be addressed in future studies.

The third of our questions is the relationship between a long vowel and the duration of following vowel. In conducting a production experiment in which participants uttered disyllabic words ($C_1V_1(:)C_2V_2$), [10] reported that vowel duration in the syllable following a long vowel is likely to be shorter than that following a short vowel. However, [10]'s research only used voiceless consonants as C_2. The current study investigated if the tendency is seen when C_2 is voiced. The results show that V_2 duration tends to be shorter when V_1 is phonologically long than when V_1 is phonologically short regardless of the voicing of C_2. The results reinforce the relationship between a long vowel and the following vowel's duration, which is suggested by [10]. In order to elucidate the generality of this phenomenon, experimental research on other languages which have phonologically long vowels is required.

ACKNOWLEDGMENTS

Many thanks to Professor Howe Stephen Mark, Professor Eiji Yamada, Associate Professor Hajime Takeyasu and two anonymous referees for their helpful comments on earlier versions. This work was partly supported by JSPS KAKENHI Grant Number JP 22J22198. The survey and experiments for this paper were also supported by the Phonetics Laboratory of Fukuoka University. All remaining errors are ours.

REFERENCES

[1] Al-Tamimi, J., Khattab, G. 2018. Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops. *Journal of Phonetics*. 71, 306–325.

[2] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer (Version 6.1.54). Amsterdam. Retrieved from http://www.fon.hum.uva.nl/praat/.

[3] Campbell, N. 1999. A study of Japanese speech timing from the syllable perspective. *Journal of the Phonetic Society of Japan*. 3, 29–39.

[4] Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*. 22, 129–159.

[5] Esposito, A., Benedetto, M-G. D. 1999. Acoustical and perceptual study of gemination in Italian stops. *J. Acoust. Soc. Am*. 106, 2051–2062.

[6] Flege, J. E. 1979. Phonetic Interference in Second Language Acquisition. PhD. dissertation, Indiana University, Bloomington.
[7] Han, M. S. 1994. Acoustic manifestations of mora timing in Japanese. *J. Acoust. Soc. Am*. 96, 73–82.
[8] House, A. S. 1961. On vowel duration in English. *J. Acoust. Soc. Am*. 33, 1174–1178.
[9] Hussain, Q., Shinohara, S. 2019. Partial devoicing of voiced geminate stops in Tokyo Japanese. *J. Acoust. Soc. Am*. 145, 149–163.
[10] Ishibashi, S. 2022. Effects of adjacent segments on the duration of the vowel and consonant within a disyllabic word in Japanese. *Phonological Studies* 25, 11–20.
[11] Keating, P., Linker. W., Huffman. M. 1983. Patterns in allophone distribution for voiced and voiceless stops. *Journal of phonetics* 11, 277–290.
[12] Keating, P. 1985. Universal phonetics and the organization of grammars. In: Victoria F. A. (ed), *Phonetic linguistics: Essays in honor of Peter Ladefoged*. Academic Press, 115–132.
[13] Kubozono, H. 2005. Nihongo ni okeru jikanseigyo no syoso (in Japanese). Inritsu ni tyakumoku sita onseigengojohosyori no kodoka (A report of MEXT Grants-in-Aid for Scientific Research). Retrieved from http://www.govo.t.u-Tokyo.ac.jp/tokutei_pub/.
[14] Mitleb, F. M. 1984. Voicing effect on vowel duration is not an absolute universal. *Journal of Phonetics* 12, 23–27.
[15] Port, R. F., Dalby, J., O'Dell, M. 1987. Evidence for mora timing in Japanese. *J. Acoust. Soc. Am*. 81, 1574–1585.
[16] Wagner, P., Trouvain, J., Zimmerer, F. 2015. In defense of stylistic diversity in speech research. *Journal of Phonetics*. 48, 1–12.
[17] Wells, J. C. 1990. Syllabification and allophony. In: Ramsaran S. (ed), *Studies in the pronunciation of English: A commemorative volume in honour of A. C. Gimson*. Routledge, 76–86.
[18] Williams, L. 1977. The voicing contrast in Spanish. Journal of phonetics. 5, 169–184.
[19] Xu, Y. 2010. In defense of lab speech. *Journal of phonetics*. 38, 329–336.
[20] Yoneyama, K., Kitahara, M. 2014. Voicing effect on vowel duration: corpus analyses of Japanese infants and adults, and production data of English learners. *Journal of the Phonetic Society of Japan*. 18, 30–39.

ACOUSTIC PROPERTIES OF THE FOUR-WAY LARYNGEAL CONTRAST IN BENGALI INFANT DIRECTED SPEECH

Jahnavi Narkar

University of California, Los Angeles
jnarkar@ucla.edu

ABSTRACT

This paper analyzes the acoustic properties of the four-way laryngeal contrast in Infant Directed Speech (IDS), comparing it with Adult Directed Speech (ADS). Since IDS has been demonstrated to be hyper-articulated, its acoustic manifestation is expected to contain enhanced cues to the four-way contrast. I investigated VOT, PVI and H1*–H2* of prevocalic stops and affricates in the speech of 10 Bangladeshi Bengali speakers. The results revealed an asymmetry in IDS – the VOT of voiced-unaspirated and voiced-aspirated segments was longer in IDS, but there was no difference in the VOT of voiceless-aspirated and voiceless-unaspirated segments between the two registers; PVI of voiced-aspirated segments, but not that of voiceless-aspirated segments, was longer in IDS. Finally, H1*–H2* did not differ by register, although the expected effect of category was found. Viewed against studies on similar languages, these results suggest that language-specific phonetic grammars modulate the acoustic manifestation of IDS.

Keywords: Infant directed speech, laryngeal contrast, acoustic cues, VOT

1. INTRODUCTION

Eastern Bengali (Bangladesh), also known as Bangla (henceforth referred to simply as *Bengali*), employs a four-way system of laryngeal contrasts that utilizes both voicing and aspiration. This four-way contrast is illustrated in Table 1 where 'T' denotes voiceless-unaspirated stops and affricates; 'Th' denotes voiceless-aspirated stops and affricates; 'D' denotes voiced-unaspirated stops and affricates; 'Dh' denotes voiced-aspirated stops and affricates.

Cross-linguistically, voicing contrasts have been typically characterized using voice onset time (VOT) which has been proposed to be the primary cue to voicing in languages with two or three-way distinctions [1]. However, VOT is not sufficient to distinguish the four-way contrast of languages like Bengali [2]. In addition to VOT, several other acoustic cues have been found to contribute to the four-way contrast, such as onset f0, burst frequency and measures of spectral tilt [3, 4, 5, 6].

Category	Segments	Example
T	[p, t̪, tɕ, ʈ, k]	[t̪ana] (*drawn*)
Th	[pʰ, t̪ʰ, tɕʰ, ʈʰ, kʰ]	[t̪ʰana] (*police station*)
D	[b, d̪, dʑ, ɖ, g]	[d̪ana] (*grain*)
Dh	[bʱ, d̪ʱ, dʑʱ, ɖʱ, gʱ]	[d̪ʱana] (*paddy*)

Table 1: The Bengali four-way contrast.

In this paper, I tease apart the primary cues to this contrast in Bengali by comparing the acoustic properties of infant-directed speech (IDS) and adult-directed speech (ADS). IDS has been thought to be hyper-articulated to facilitate phonetic learning [7]. This is generally true for vowels which show greater separation between categories, but the results on consonants are mixed. Stops in languages that have a two-way voicing contrast have been found to be hyper-articulated [8, 9], hypo-articulated [10] and no different from each other [11] in terms of VOT. In these cases, where there is a single primary cue to voicing, hyper-articulation may involve lengthening the VOT of both stop categories such that the relative separation between them remains comparable. Conversely, if aspirated stops are produced with longer VOT, and unaspirated stops with comparable or shorter VOT, the two categories are more separated. I investigate the former case whereby VOT is lengthened, but it must be pointed out that this only involves enhancing the primary cue associated with the contrast, not necessarily increasing separation between categories.

The only study on IDS in a language with the four-way contrast is [12] which found that Nepali stops are hypo-articulated in IDS in terms of VOT. In addition to VOT, I measured two acoustic cues that have been proposed to be adequate for distinguishing this complex contrast – PVI (pre-vocalic interval) and H1*–H2*. Speakers are expected to enhance primary cues in hyper-articulated registers like IDS.

Therefore, if they are primary, VOT and PVI, which are durational cues, are expected to be lengthened in IDS, and H1*–H2*, which is a measure of breathy voice, is expected to be greater in IDS. In addition to providing a description of the acoustic properties of IDS, the results also have a bearing on the realization and representation of the four-way contrast, and on the role of IDS in phonetic acquisition.

2. METHOD

2.1. Materials

Data collected by [13] for their investigation of the intonational phonology of Bengali IDS were used in this study. They recorded IDS and ADS speech of 10 native speakers of Bengali reading the "North Wind and Sun" fable translated into Bengali, from [14]. For the ADS speech, subjects were asked to "read (the fable) at a comfortable pace." For the simulated IDS speech, subjects were asked to read the same passage as if speaking to their 4-5 month-old infant. Each speaker produced the IDS and ADS version three times each.

Target segments from the speech samples were coded as [p, b, $b^{ɦ}$, t̪, $t̪^{h}$, d̪, $d̪^{ɦ}$, tɕ, $tɕ^{h}$, dʑ, ʈ, k, k^{h}, g]. Measures for /p^{h}/ were not extracted since the speakers consistently produced these as [f]. There were no instances of /$ʈ^{h}$, ɖ, $ɖ^{ɦ}$, $g^{ɦ}$, $dʑ^{ɦ}$/ in the fable. Each pre-vocalic segment of interest was included. Since VOT is closely related to speech rate, both global speech rate (at the utterance level) and local speech rate (local to the target phone) were checked. Both rates were significantly slower in IDS. Therefore, the two temporal cues, VOT and PVI, are expected to be longer in IDS.

2.2. Acoustic Analyses

2.2.1. VOT

Category	VOT	PVI	H1*–H2*
T	2999	2999	1804
Th	518	518	403
D	1429	1429	1410
Dh	453	453	435

Table 2: Tokens included in the analyses.

VOT was measured in Praat [15] as the temporal interval between the beginning of the release burst and the onset of quasi-periodicity. For voiced stops and affricates (voicing lead), it was measured as the duration between the onset of voicing and the stop burst. For voiceless stops (voicing lag), VOT was measured as the duration between the burst and the onset of voicing. Tokens that were erroneous or had noise or interruptions were excluded. Table 2 shows the total number of tokens included in the final analyses.

2.2.2. Pre-vocalic interval (PVI)

Following [5], I measured PVI in Praat [15] as the temporal interval between the beginning of the release burst and the onset of breathiness in the following vowel as indicated by the clear onset of a dark F2 in the spectrogram or an increase in amplitude in the waveform. Notice that since PVI is a durational measure of the breathy portion of the stops, the crucial categories are the two aspirated ones – Dh and Th.

2.2.3. H1–H2**

The difference in amplitudes of the first and second harmonics, corrected for formant frequencies and bandwidths, H1*–H2*, was measured in Voicesauce [16] at the onset of the following vowel as previous studies have shown that difference in breathiness rapidly declines after the onset [5, 6]. The affricates were excluded from the analysis of H1*–H2* as there were no instances of breathy-voiced [$dʑ^{ɦ}$], and [tɕ], [$tɕ^{h}$] and [dʑ] were expected to have greater H1*–*H2* from the fricative portion following the stop burst. Therefore, the contribution of the breathy portion following voiced-aspirated stops would be obscured due to the greater H1*–H2* in the three other laryngeal categories driven by the affricates. To account for differences between speakers, H1*–H2* was z-scored by speaker to reduce variation. Tokens with absolute z-scores greater than 3 were taken to be outliers and excluded from the analysis (see Table 2). While the number of tokens is not distributed evenly across the laryngeal categories, they do reflect the frequency of segments belonging to the four categories in the lexicon.

2.3. Statistical Analyses

All statistical analyses were conducted using mixed effects regression models using the `lmer` function in the `lme4` package [17] in R [18]. To compare the effect of register on VOT, separate linear mixed effects models were run for each category. Each optimal model was obtained by applying the `step` function, which performs backward elimination of non-significant effects, to a fully specified model. This fully specified model contained all the fixed

factors, their interactions, and random intercepts for speaker and word. Since VOT is known to vary with place of articulation [19] and prosodic position [20], the models of VOT and PVI included these as fixed factors in addition to register. Since there was no reason to expect an effect of place or position on breathiness, these factors were not included in the H1*–H2* model.

3. RESULTS

3.1. VOT

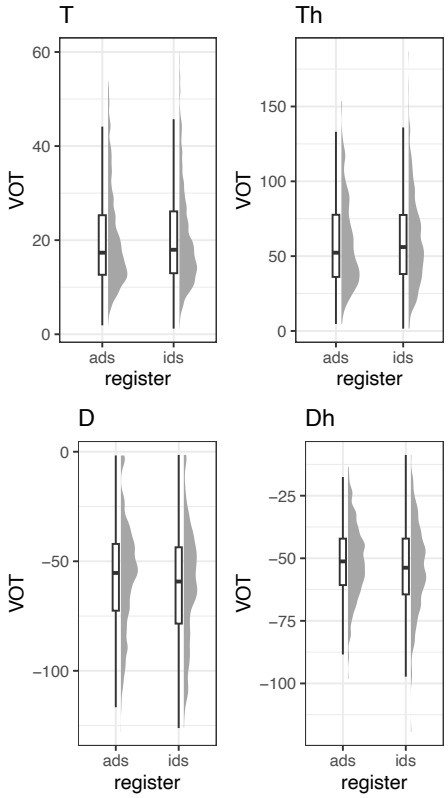


Figure 1: VOT by register.

Fig. 1 shows the VOT for each laryngeal category by register. Note that since the VOT of voiced stops is negative, for the Dh and D categories, greater absolute values of VOT (lower on the y-axis) correspond to longer durations. For the T category, there was no effect of register on VOT, which is in line with studies on English and Spanish which found that voiceless-unaspirated stops were not hyper-articulated in IDS even when voiceless-aspirated and prevoiced stops were [9]. Additionally, the expected result of place was found – [tɕ] had significantly longer VOT than [p]($\beta = -21.85, t(159) = -18.99, p < 0.001$), [t̪]($\beta = -24.06, t(85) = -19.18, p < .001$), [ʈ]($\beta = -24.93, t(320) = -20.26, p < .001$) and [k] ($\beta = -7.58, t(66) = -5.38, p < .001$).

For the Th category, there was no effect of register on VOT, contra [8, 9, 12]. However, the expected effects of place and word position were found – [tɕʰ] had significantly longer VOT than [kʰ] ($\beta = -32.71, t(11) = -9.71, p < .001$) and [t̪ʰ] ($\beta = -49.7, t(10) = -14.22, p < .001$); VOT was longer in word-initial position than in word-medial position ($\beta = -8.44, t(12) = -2.48, p < .05$). Thus, this model shows that voiceless-aspirated stops in Bengali IDS are neither hyper- nor hypo-articulated.

For the D category, the optimal mixed effects model found a significant effect of register on VOT ($\beta = -4, t(1232) = -3.94, p < .001$) such that voiced stops in IDS had significantly longer negative VOT compared to ADS. There was no effect of place, but the expected effect of prosodic position on VOT was found ($\beta = 20.97, t(22) = 2.57, p < .05$). Finally, for the Dh category, there were significant effects of register ($\beta = -2.75, t(424) = -2, p < .05$) and place ($\beta = 22.73, t(6) = 3.4, p < .05$), but no effect of prosodic position. Thus, both voiced-unaspirated and voiced-aspirated stops and affricates in Bengali IDS were hyper-articulated in terms of VOT compared to ADS. The effect of prosodic position on the VOT of Th and D stops also serves to extend results from previous studies investigating the effect of prosodic position on VOT in other languages [20] to Bengali.

3.2. Pre-vocalic interval (PVI)

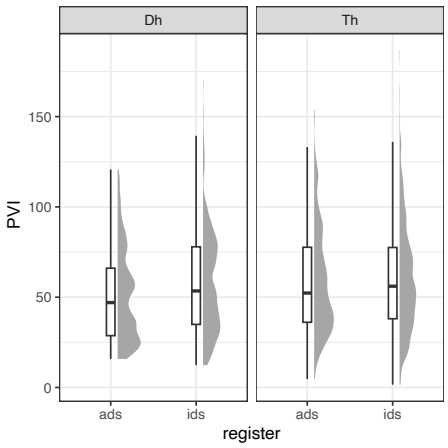


Figure 2: PVI by register.

Fig. 2 shows the PVI of the Dh and Th categories by register. For Dh, the linear mixed effects

model showed a significant effect of register ($\beta = 5.66, t(398) = 2.67, p < .01$) – stops in IDS had longer PVI than in ADS. This implies that, like VOT, the aspiration portion of voiced aspirated stops is also exaggerated via lengthening in IDS. By contrast, the PVI of the Th segments was not affected by register. Thus, voiced-aspirated stops in IDS were hyper-articulated in terms of PVI while voiceless-aspirated stops were not hyper-articulated and had comparable PVI in both registers.

3.3. H1*–H2*

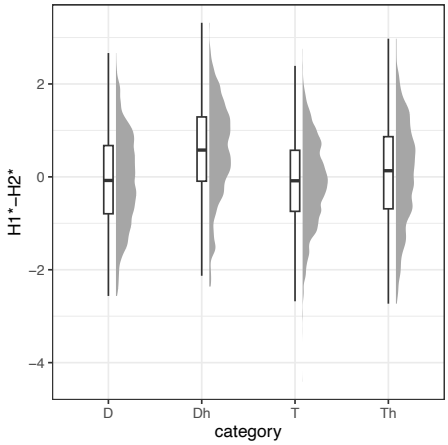


Figure 3: H1*–H2* by laryngeal category.

The linear mixed effects model did not find an effect of register on H1*–H2*. That is, IDS was not found to be breathier than ADS for any category. Since no effect of register was found, the data were combined into a single model to check for the effect of category on H1*–H2*. Fig. 3 shows this H1*–H2* by category with the IDS and ADS combined. This figure shows the expected effect of category – Dh is breathier than the rest of the laryngeal categories, and Th is slightly breathier than D and T. The full mixed effects model confirmed this. The only significant effect was that of category – Dh ($\beta = .93, t(118) = 7.76, p < .001$) and Th ($\beta = .17, t(482) = 2.38, p < .05$) were significantly breathier than T. H1*–H2* of T and D were comparable. The effect of category on H1*–H2* confirms previous findings on other languages with the four-way contrast [4, 5, 6]. However, the lack of effect of register on H1*–H2* is in opposition to findings from [21] which found vowels in Japanese IDS were significantly breathier than in ADS. Breathy voice might, thus, be a language-specific property of Japanese IDS rather than a general property of IDS. It is possible that the H1*–H2* at vowel midpoint is breathier than at vowel onset in Bengali IDS, but this possibility is not explored here and is left as an avenue for future research.

4. CONCLUSION

In summary, the acoustic manifestation of Bengali IDS revealed an asymmetry among the laryngeal categories in terms of VOT and PVI. In terms of primary acoustic cues to the four-way contrast, these results suggest that voiced-aspirated stops are cued by VOT and PVI since both of these were exaggerated in the hyper-articulated register. Similarly, voiced-unaspirated stops and affricates are cued by VOT. However, voiceless-aspirated stops and affricates, which are thought to be adequately cued by positive VOT alone, were not exaggerated in IDS. Finally, breathy voice was not exaggerated in IDS. This suggests that IDS may not be hyper-articulated across the board and that language-specific phonetic grammars may modulate the acoustic manifestation of different speech styles.

These results can also be viewed through the lens of phonological specification. Assuming that phonetic cues associated with specified features are exaggerated in hyper-articulated registers [22], these results support the representation of the Dh category with the features [voice] and [spread], the D category with [voice], and the T category as unspecified. However, since the VOT of the Th category was not exaggerated in IDS, this also supports the specification of this category with a feature other than [spread] (or none at all). Given that this result is problematic for most theories of laryngeal representation, the acoustic properties of registers like IDS or slow speech cannot always adequately comment directly on questions of phonological representation. The relationship between phonetic realization and phonological specification can be abstract, and evidence for featural representation must be *phonological*, not just phonetic.

The acoustic properties of IDS are language-dependent and hierarchically ordered. In some languages, vowels, but not consonants, are hyper-articulated [5]. This paper shows that there are asymmetries even within a single class of consonants – caregivers do not hyper-articulate equally across the board. This raises questions regarding the timeline of acquisition of this contrast and suggests that IDS may not always facilitate phonetic learning as it does not contain uniformly enhanced cues to contrasts in the target language.

5. REFERENCES

[1] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[2] G. N. Clements and R. Khatiwada, "Phonetic realization of contrastively aspirated affricates in Nepali," in *Proc. 16th ICPhS*, 2007, pp. 629–632.

[3] M. K. Rami, J. Kalinowski, A. Stuart, and M. P. Rastatter, "Voice onset times and burst frequencies of four velar stop consonants in Gujarati," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3736–3738, 1999.

[4] I. Dutta, *Four-way stop contrasts in Hindi: An acoustic study of voicing, fundamental frequency and spectral tilt*. University of Illinois at Urbana-Champaign, 2007.

[5] K. H. Berkson, "Capturing breathy voice: Durational measures of oral stops in Marathi," *Kansas Working Papers in Linguistics*, 2012.

[6] J. Schertz and S. Khan, "Acoustic cues in production and perception of the four-way stop laryngeal contrast in Hindi and Urdu," *J. Phon.*, vol. 81, 2020.

[7] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, L. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[8] U. Sundberg, "Consonant specification in infant-directed speech. Some preliminary results from a study of voice onset time in speech to one-year-olds," *Working papers/Lund University, Department of Linguistics and Phonetics*, vol. 49, pp. 148–151, 2001.

[9] M. S. Fish, A. García-Sierra, N. Ramírez-Esparza, and P. K. Kuhl, "Infant-directed speech in English and Spanish: Assessments of monolingual and bilingual caregiver VOT," *J. Phon.*, vol. 63, pp. 19–34, 2017.

[10] U. Sundberg and F. Lacerda, "Voice onset time in speech to infants and adults," *Phonetica*, vol. 56, no. 3-4, pp. 186–199, 1999.

[11] K. T. Englund and D. M. Behne, "Infant directed speech in natural interaction—Norwegian vowel quantity and quality," *Journal of psycholinguistic research*, vol. 34, no. 3, pp. 259–280, 2005.

[12] T. Benders, S. Pokharel, and K. Demuth, "Hypo-articulation of the four-way voicing contrast in Nepali infant-directed speech," *Language Learning and Development*, vol. 15, no. 3, pp. 232–254, 2019.

[13] K. Yu, S. D. Khan, and M. Sundara, "Intonational phonology in Bengali and English Infant-directed speech," *Proceedings of Speech Prosody 7*, pp. 1130–1134, 2014.

[14] S. D. Khan, "Bengali (Bangladeshi standard)," *J. IPA*, pp. 221–225, 2010.

[15] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]. Version 6.3.03," retrieved 17 December 2022 from http://www.praat.org/, 2022.

[16] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *Proc. 17th ICPhS*, 2011, pp. 1846–1849.

[17] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using eigen and s4," 2015, r package version 1.1-7. [Online]. Available: http://CRAN.R-project.org/package=lme4

[18] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2021.

[19] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *J. Phon.*, vol. 27, no. 2, pp. 207–229, 1999.

[20] T. Cho and P. Keating, "Effects of initial position versus prominence in English," *J. Phon.*, vol. 37, no. 4, pp. 466–485, 2009.

[21] K. Miyazawa, T. Shinya, A. Martin, H. Kikuchi, and R. Mazuka, "Vowels in infant-directed speech: More breathy and more variable, but not clearer," *Cognition*, vol. 166, pp. 84–93, 2017.

[22] J. Beckman, P. Helgason, B. McMurray, and C. Ringen, "Rate effects on Swedish VOT: Evidence for phonological overspecification," *J. Phon.*, vol. 39, no. 1, pp. 39–49, 2011.

Spectral implications of codec compression on voiceless fricatives

Krestina V. Christensen

Aarhus University & University of York
kvc@cc.au.dk & kvc507@york.ac.uk

ABSTRACT

This study investigates voiceless fricatives that have been subject to codec compression, as codec compressed speech is commonplace in people's lives, and knowledge is still limited on how the compression affects the acoustics of consonants, here [s, f, fʲ, θ, ʃ] specifically. The paper presents a study of read speech from 30 male speakers of English, compressed with three codecs (AMR, MP3 and Opus) at a single bit rate per codec. For each fricative, CoG, SD, and skewness were measured and compared between the baseline uncompressed PCM-WAV and the codec compressed versions. The findings indicate significant lowering of CoG and SD following codec compression, with segment and codec dependent tendencies. Skewness is likewise affected, but less so than the other measures. This has implications for phonetics when e.g. using codec compressed speech as data, but especially in socio- and forensic phonetics with possible diffusion of sound changes, and speaker comparisons.

Keywords: Codec compression, spectral measures, fricatives, acoustics

1. INTRODUCTION

It is well known that a number of technical effects are apparent in analogue speech transmission, e.g. the limited bandwidth in telephone transmission [1]. However, digital transmission introduces a number of additional variables including codec compression. The codecs are collections of algorithms. These are designed to identify and translate (encode) speech into a numerical representation, which can be sent over a network and finally decode it back to an audible pressure wave [2].

Three aspects in the transmission process and codec compression with AMR-WB and Opus (i.e. speech-specific codecs) are key: a) the digital conversion; b) data carrying capacity and quality, and c) Voice Activity Detection (VAD) and its included voicing parameters. MP3 is a perceptual codec not designed to identify speech, and does not include the VAD. Now, the digital conversion or digitisation of the analogue signal is the translation of the pressure wave into a numerical representation. This digitisation process samples the analogue speech signal at regular intervals, determined by the sampling frequency, and represents the amplitude of the signal at each sample using a pre-determined set of linearly spaced amplitude values [3]. Secondly, the level of acoustic detail (i.e. quality) is determined by the data carrying capacity and the bit rate (i.e. the number of bits that can be sent over a digital network per second). This in turn is also expressed in the available bandwidth for the transmission. All of the investigated codecs are *lossy*, which entails that not all the information in the original pressure wave is captured in the codec-compressed signal [4, 5, 6]. The choice of information to include in the numerical representation is based on the redundancy of speech and identification of regularity (i.e. periodicity/ voicing), as the overall goal is to detect speech and exclude noise (i.e. irregular sound waves) [7, 8]. Thus, codec compression is potentially problematic for voiceless fricatives due to their aperiodic and noise-like acoustic composition. If fricatives are indeed altered by the codec compression this renders the use of digitally transmitted or encoded speech problematic in a range of fields such as sociolinguistics, forensic phonetics and phonetic analysis in general. Therefore, in the absence of relevant research, this study aims to establish a baseline for the spectral implications of codec compression on voiceless fricatives. It does so by investigating [s, f, fʲ, θ, ʃ] as produced by 30 male native speakers of English recorded in a phonetically balanced reading. [fʲ] is included as a separate phoneme as it is used as a distinct label by the forced aligner e.g. in the word e.g. in initial position of the word *furiously*. It is predicted that Centre of Gravity (CoG) and Standard Deviation (SD) will be lowered for all the segments following codec compression partially due to the limited signal bandwidth. However, the lowering is expected to be more prominent for [f], [fʲ], and [θ] considering the relatively lower intensity of the friction than in the sibilants. Skewness is expected to show effects of the codec compression, but with a less clear patterning than CoG and SD. In addition, codec dependent tendencies are expected due to the technical differences between the codecs.

2. METHODOLOGY

2.1. Corpus and participants

The *You Came to Die?!* corpus [9] was used for this study. It included 30 male native speakers of English aged between 18 and 41 reading an approximately 10 minute phonetically balanced passage recorded in studio quality. They spoke five different accents of English with six speakers of each accent: Australian (AUS), New Zealand (NZL), London (LON), Newcastle (NCL), and York (YRK). Female speakers are reported to be more affected by codec compression than male speakers [10], and as the aim is a baseline, male speakers were chosen.

2.2. Segmentation

The files were forced aligned with the Montreal Forced Aligner (MFA) with a customised version of the English (UK) MFA dictionary [11]. The files were then manually corrected in Praat textgrids [12]. These corrections included boundary corrections, where the alignment included e.g. non-speech elements, voicing or wrong transcription.

2.3 Sound files and codec compression

The original 44.1 kHz recordings were down-sampled to 16 kHz. This was done to single out the effects of the codec compression rather than the limitations in bandwidth. The 16 kHz WAV files were codec compressed with AMR-WB [13], MP3 [14], and Opus [15]. A different bit rate was selected for each codec, which represented a typical or average to low quality for each [16, 17, 6]. This was 12.65 kbps for AMR-WB [13], 32 kbps for MP3 [14], and 24 kbps for Opus [15]. This provides a baseline for further research. Spectral analysis was conducted on the 16 kHz WAV files and the three codec compressed files to determine the extent of any changes to the frequency profile of noise-like signals, with particular focus on attenuation at higher frequencies.

2.4 Data extraction

The recordings elicited the fricatives [s, f, fʲ, θ, ʃ] in varying segmental contexts and initial, medial, and final position. This yielded a final baseline dataset with 5104 tokens of [ʃ], 5156 of [θ], 13668 of [f], 960 of [fʲ], and 29336 of [s], and the same number of tokens in each of the three codec compressed versions.

The spectral measures of CoG, SD, and skewness were obtained from the central 20ms frame using multi-taper analysis [18] implemented in MATLAB [19]. Multi-taper analysis requires no additional windowing, and no pre-emphasis was applied. During preliminary examination of the data, it was found that a number of segments had CoG values below 1 kHz, which is unexpected for these fricatives [20]. These segments were excluded from the main dataset along with their equivalent segments in the other codec compressed or baseline files. The 1 kHz threshold was set as the exclusion criterion because a CoG value less than 1 kHz might reflect a strong influence of the mains hum (at around 50 Hz), or a substantial reduction in intensity. An audio file of white noise was also subject to the same encoding processes as the speech samples to estimate the actual upper-frequency limit of the speech encoding.

2.5 Statistical analysis

To provide an overall summary of the spectral measures for each segment, the mean values and difference between these values (i.e. baseline vs. codec compression) for all three spectral measures (i.e. CoG, SD and skewness) in the baseline and each codec compression were calculated.

To examine the influence and interaction of the variables a statistical analysis was performed in R [21, 22] using mixed effects modelling [23] and ANOVAs [23]. The mixed effects models were lmer based with CoG, SD, and skewness as dependent variables.The maximal and optimal model, which was used for the analysis was the following:

Best fitted model = lmer (Spectral measure ~ format * segmentLabel + pre + post + (1|speaker) + (1|word) + (1|duration), data = data)

This model was achieved by comparing a range of models for each spectral measure and codec based on a baseline model predicted to be most accurate from the data. The model was tested with and without a number of variables for each codec and spectral measure with varying fixed and random effects as well as slopes and interactions to ensure the accuracy of the model,. This included the following independent variables: Format (6 levels (2 per model): baseline 16 kHz, and one of the following AMR-WB, MP3, or Opus), Segment (5 levels: [s, f, fj, θ, ʃ]) Speaker (30 levels: individual speakers), word position (with three levels: initial, medial, and final), preceding sound labelled *pre* (with 21 levels) and following sound labelled *post* (with 60 levels). The evaluation of fit was done by inspection of residuals, deviance and Akaike Information Criterion (AIC) [25]. Finally, Tukey adjusted post-hoc tests were done using Emmeans to extract the significance values for the interactions from the models as well as

plot the linear predictions [26]. Thus, this resulted in 4 models in total, one per measure.

3. RESULTS

Similar trends are found across codecs; however, the magnitude of these trends is to some extent codec and segment specific. Figure 1 shows spectrographic representations that illustrate how the codec compression limits the intensity and upper frequency limit to varying degrees in the word *fluff* as produced by one York speaker.

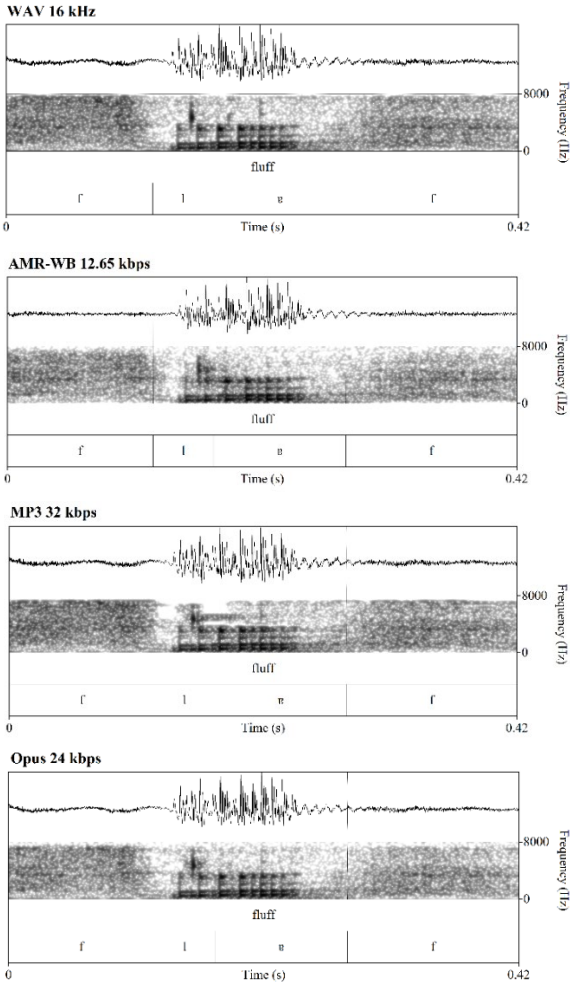


Figure 1: *fluff* produced by one York accented speaker in WAV and all codec compressions.

The white noise analysis indicated that the codecs did not have the same upper frequency limit (i.e. the point where the codec-compressed files do not follow the flat frequency profile of the white noise). For AMR-WB the estimated cut-off frequency i.e. the point when the frequencies clearly start sloping downwards was 5600 Hz, for MP3 7100 Hz, and for Opus 7000 Hz (Figure 2 below). Summary statistics were extracted for each spectral moment in each codec compression. The differences in mean values between the original and codec compressed recordings are illustrated in Table 2, with the arrows indicating the directionality of the average change.

seg	codec	bitrate (kbps)	CoG (Hz)	SD (Hz)	Skew.	Kurt.
[f]	AMR	12.65	306↓	192↓	0.06↓	0.13↓
[f]	MP3	32	128↓	119↓	0.09↓	0.19↓
[f]	Opus	24	415↓	206↓	-0.10↑	-0.35↑
[f̟]	AMR	12.65	219↓	137↓	0.11↓	0.32↓
[f̟]	MP3	32	105↓	112↓	0.12↓	0.28↓
[f̟]	Opus	24	342↓	195↓	-0.02↑	-0.31↑
[s]	AMR	12.65	349↓	50↓	0.13↓	0.07 ↓
[s]	MP3	32	169↓	70↓	0.27↓	-0.19↑
[s]	Opus	24	300↓	65↓	0.33↓	-0.72↑
[ʃ]	AMR	12.65	11↓	-10↑	0.04↓	0.67↓
[ʃ]	MP3	32	18↓	37↓	0.25↓	1.17↓
[ʃ]	Opus	24	121↓	120↓	0.24↓	-0.31↑
[θ]	AMR	12.65	423↓	219↓	-0.02↑	0.08↓
[θ]	MP3	32	180↓	132↓	0.06↓	0.14↓
[θ]	Opus	24	498↓	197↓	-018↑	-0.36↑

Table 2: differences in mean values between baseline (WAV) and codec compression in Hz. Arrows indicate the direction of the change.

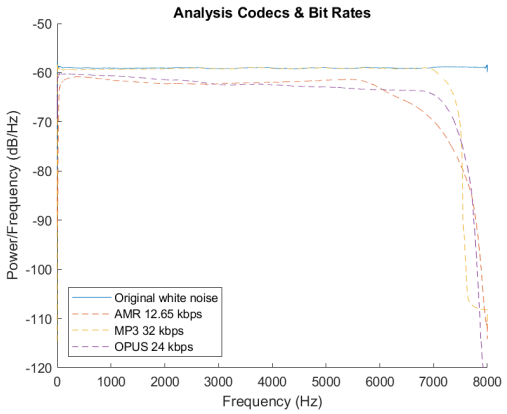


Figure 2: Frequency spectra of original white noise signal and codec compressed versions.

3.1 AMR

The AMR-WB compression generally lowered the spectral measures for all the segments based on the mean values and significantly so [p<.0001]. This is the case for all fricatives apart from [ʃ], where the biggest change is seen for CoG with a lowering of 19 Hz. For CoG the changes vary from between 218 Hz [f̟] to 422 Hz [θ] from WAV to AMR-WB. The changes in SD vary from 137 Hz [f̟] to 218 Hz [θ]. The changes to skewness are limited, but significant apart from [ʃ]. The biggest change for skewness was 0.13 found for [s]. The AMR-WB compression makes [f] and [θ] more alike in terms of CoG and skewness.

3.2 MP3

The MP3 compression lowers all spectral values for all segments. This is significant for most measures, with p<.0001, although not for [ʃ]. [ʃ] stays almost unaffected by the MP3 compression, with the biggest change being for SD at 36 Hz. For the remaining segments, the biggest changes in CoG is found for [s], with a lowering of 349 Hz. [f], [θ], and [fʲ] are lowered with between 104 Hz and 179 Hz. For SD [s] is similar to [ʃ], with less than 100 Hz change, whereas the other segments show changes between 112 Hz and 131 Hz. Skewness shows smaller changes with the biggest change observed for [s] at 0.27. The lowering found for both [fʲ] and [θ] are significant (p<0.05). Overall, the relation between the segments' spectral values remain stable.

3.3 Opus

The Opus compression lowers CoG and SD for all segments [p<.0001], but shows a mixed pattern for skewness. Opus is the only codec with an effect greater than 100 Hz on [ʃ] for both CoG and SD. The most substantial lowering of CoG is found for [θ], with 497 Hz, whereas the smallest change is found for [s] with 300 Hz. [f] is lowered by 415 Hz and [fʲ] by 342 Hz. In comparison, the effect on SD is smaller, with a minimum change for [s] at 63 Hz and the remaining segments changing between 119 Hz ([ʃ]) and 205 Hz ([f]). Skewness is lowered for [ʃ] (0.24) and [s] (0.33) [p<.0001]. The remaining tokens are increased, [f] and [θ] significantly with ~0.1 [p<.0001]. [fʲ] was not significantly affected. Overall, [f] and [θ] become almost identical in terms of CoG and SD, and more similar for skewness.

3.4 Rejected tokens

The number of tokens rejected were 61 pairs for WAV-AMR-WB, 18 pairs for WAV-MP3, and 31 pairs for WAV-Opus. None of these tokens were [fʲ] and [ʃ]. Analysis of these tokens is beyond the scope of the current paper. However, most rejected tokens in the AMR-WB were found to have their CoG value above the 1 kHz cut-off in the corresponding WAV file. In Opus and MP3, the CoG mean values for the rejected tokens were all below the 1 kHz cut-off in the WAV baseline apart from [f] in MP3. The MP3 compression shows mixed segment dependent patterns, whereas the Opus codec appears to lower the tokens below 1 kHz similarly to the tokens in the main dataset.

4. DISCUSSION

The study aimed to investigate the acoustic effects of codec compression with AMR-WB, MP3 and Opus on the voiceless fricatives [f], [fʲ], [θ], [s], [ʃ].

It is clear from the white noise analysis that the codecs introduce different cut-offs. This is potentially correlated with the specific bit rates that were applied (and will be investigated further in future). The lowest limit is found for AMR, which also has the most tokens changing CoG from above to below 1 kHz following compression. However, the biggest effects on spectral measures (excluding rejected tokens) is found for Opus, which has a higher upper frequency limit. This, suggests an effect of the available signal bandwidth, but also that the technical elements e.g VAD are influential as well [4]. Thus, the results suggest that the spectral changes are to some extent codec and segment dependent and must be accounted for individually.

As predicted, [f] and [θ] were more affected by the compression than [s] and [ʃ]. However, both [fʲ] and especially [ʃ] presented less to no effect of the codec compression. It is interesting to note that [f] and [θ] become more alike in both AMR-WB and Opus. These results have implications for any acoustic or auditory analysis of codec compressed speech in a number of scenarios.

Firstly, these results follow previous research on digitally transmitted speech in acoustic phonetics as it also cautions its use as primary data [27, 28], unless the goal of the research is, as here, to investigate the technical implications. This is due to the significant effects found for almost every spectral measure in each codec compression.

Secondly, the changes in spectral information potentially make sounds less distinct or not directly comparable to a higher quality recording. Thus, in forensic phonetics, the analysts need to be aware of these effects when analysing and interpreting codec compressed recordings. Whether this has an auditory and/or perceptual effect is to be assessed. If such an effect is found, it means bias could affect the forensic phonetician in their conclusions. Moreover, the amount of codec compressed speech today have potential implications fir sound change and diffusion of variants stemming from codec compression. In a socio-phonetic perspective, when judging the effects, it is important to bear in mind that the baseline WAV here is already down-sampled to 16 kHz. Finally, this study used completely controlled conditions on male speakers without any live transmission or background noise. With these factors added and using female speakers, the effects are likely to be more prominent e.g. when the codec identifies the fricatives as noise and eliminates them from transmission.

5. REFERENCES

[1] Künzel, H. J. 2001. "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies". *Forensic Linguistics* 8, 1, 80–99.

[2] Zölzer, Udo. 2008. *Digital Audio Signal Processing*. U.K.: Wiley

[3] Ladefoged, P. 1996. *Elements of acoustic phonetics.* Chicago: University of Chicago Press.

[4] 3GPP. 2022. "Technical Specification Group Services and System Aspects; Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Voice Activity Detector (VAD)". *3GPP TS 26.194*. (version 16.0.0). France: Valbonne

[5] Marc Gayer, Markus Lohwasser and Manfred Lutzky. 2004. "Implementing MPEG Advanced Audio Coding and Layer-3 encoders on 32-bit and 16-bit fixed-point processors". *Revision 1.11*. Germany: Fraunhofer Institute for Integrated Circuits IIS.

[6] Valin, JM., Vos K., & Terriberry T. 2012. "Definition of the Opus Audio Codec" *report*. retrieved 06 August 2022 https://www.rfc-editor.org/info/rfc6716

[7] Guillemin, Bernard J., & Watson, Catherine I. 2006. "Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification". *In 11th Australian International Conference on Speech Science & Technology*, 483–88. New Zealand: Auckland University.

[8] Herre, Jürgen & Dick, Sascha. 2019. "Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review", *Applied Sciences*. 9 (14). Doi: 10.3390/app9142854

[9] Best, C., Shaw, J., Docherty, G., Evans, B., Foulkes, P., & Hay, J. 2012-2015. *The You Came to Die?! corpus*. ARC Discovery Project DP120104596

[10] Siegert, Ingo, and Oliver Niebuhr. 2021. 'Speech Signal Compression Deteriorates Acoustic Cues to Perceived Speaker Charisma'. In Tagungsband Der 32. Konferenz, 1–10.

[11] McAuliffe, Michael, Michaela Socolof, Elias Stengel-Eskin, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. *Montreal Forced Aligner* [Computer program]. (Version 1.0.0), retrieved May 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[12] Boersma, P., Weenink, D. 1992-2021. *Praat: doing phonetics by computer*. 6.0.25

[13] 3GPP, 2007-03. "Technical Specification Group Services and System Aspects, ANSI-C code for the Adaptive Multi Rate - Wideband (AMR-WB) speech codec*". TS 26.173*. (version 7.0.0). France: Valbonne

[14] FFmpeg. 2022. "FFmpeg" *64-bit static Windows 4.4.1-essentials_build.* Retrieved October 2022.

[15] Xiph.Org Foundation. 2022. "opusenc/opusdec". *opus-tools 0.2 using libopus 1.3.1*. Retrieved October 2022.

[16] 3GPP. 2022. "Performance characterization of the Adaptive Multi-Rate Wideband (AMR-WB) speech codec". *TR 26.976*. (version 17.0.0). France: Valbonne

[17] Triton. 2022. "Choosing audio bitrate settings". *https://tritondigitalcommunity.force.com*. Accessed 19th of December 2022.

[18] D. Thomson. 1982. "Spectrum estimation and harmonic analysis," *Proceedings of IEEE*, Vol. 70, No. 9

[19] MATLAB. 2010. (version 7.10.0 (R2010a)). Natick, Massachusetts: The MathWorks Inc.

[20] Shadle, Christine Helen. 1986. "The acoustics of fricative consonants" *The Journal of the Acoustical Society of America*. 79 (2). Doi: 10.1121/1.393552

[21] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

[22] RStudio Team. 2019. RStudio Team (2019). *RStudio: Integrated Development for R.*, (version 1.2.5033). Boston, MA: RStudio Inc. http://www.rstudio.com/

[23] Kuznetsova, A., PB Brockhoff, and RHB Christensen. 2017. "LmerTest Package: Tests in Linear Mixed Effects Models". *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13

[24] Girden, E. R.. 1992. *ANOVA: Repeated measures*. Sage'

[25] H. Akaike. 1974 "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19(6): 716-723. doi: 10.1109/TAC.1974.1100705

[26] Lenth, Russell. 2020. *Emmeans*: *Estimated Marginal Means, Aka Least-Squares Means* (version 1.4.6). R package https://CRAN.R-project.org/package=emmeans

[27] Siegert, Ingo & Niebuhr, Oliver "Speech Signal Compression Deteriorates Acoustic Cues to Perceived Speaker Charisma" *in Tagungsband der 32. Konferenz on Elektronische Sprachsignalverarbeitung.* Germany: Berlin

[28] Leeman, Adrian et. al. 2020. "Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing". *Linguistics Vanguard.* 6 (s3). Doi: 10.1515/lingvan-2020-0061

TEMPORAL VS. AREA-SUM FORMULAE OF VOWEL NASALITY IN SIMULATED AND NASOMETRIC CORPORA

Michael Dow

Université de Montréal
michael.dow@umontreal.ca

ABSTRACT

Nasometric and aerodynamic methods are common in the study of vowel nasality, but their various associated "global" formulae, while correlating strongly, diverge along factors like vowel height. In this study, nasometric data for 14,000 artificial vowels were synthesized according to 14 types (nasal-oral crossover, relative slopes, etc.), and a pre-existing corpus of 4,319 French vowels was reinterpreted along the same lines. A comparison of temporal and energy-based formulae on these corpora reveals a distinct sinusoidal relationship depending on crossover point and phonological context (regressive vs. progressive). Maximal differences of 15 percentage points are noted. Energy-based values are higher for vowels with longer nasal phases, while the temporal formula provides greater values for vowels with shorter phases. These trends are robust in both corpora, except for post-nasal French /i/, with greater values according to the energy-based formula, regardless of nasal phase length.

Keywords: Vowel nasality, acoustics, simulation study, French

1. INTRODUCTION

A large variety of ways to quantify vowel nasality exists, differing not only in function of the type of data used (i.e., acoustic, aerodynamic or articulatory) but also of the nature of the research question. This study focuses on a comparison of so-called "global" expressions of vowel nasality, that is, an expression of the relative portion of a vowel which can be considered nasal or nasalized.

Two specific types of formula are considered here, dubbed *temporal* and *area-sum*. The former expresses vowel nasality as a function of the relative number of points measured (and by extension, relative duration) whose nasality exceeds an arbitrary threshold, while the latter eschews temporal information in favour of proportionate areas (nasal to total) underneath a differential curve. These formulae are most readily applicable to "split-level" data, i.e., any data source with simultaneous oral and nasal data expressed in the same units. Here we focus on acoustic data, specifically nasometry, and thus measurements of energy.

Previous studies employing the two types of measurements find that they correlate strongly but are sensitive to vowel height and phonological context (pre- vs. post-nasal) [1, 2]. However, both studies are based in real-world data from French and are thus potentially limited by language-specific constraints. In order to abstract away from these potential issues and approach the topic from a more purely mathematical point of view, this study follows up by comparing one of these corpora [2] with a simulated and more extensive corpus. This paper presents preliminary results and descriptive statistics from this study, with a focus on contextual nasalization, namely regressive (VN) and progressive (NV).

These findings have potential implications of accuracy for any empirical study quantifying vowel nasality, provided oral and nasal activity are measured separately. Additionally, these findings may further enlighten us about the complex, causal relationship between these two tracts and how it is exploited in human language.

2. LITERATURE REVIEW

Many types of instruments can used for measuring nasality, and many readings correlating with nasality can be extracted from the acoustic output (see [3] for a recent example), though not all lend themselves to the same sort of higher-level quantification. Some methods express nasality as the onset of a singular activity or correlate and/or in terms of the combined signal (see [4, 5] for extensive overviews). Meanwhile, split-level methods, which we now turn to, use the relationship between two causally-related indices of nasality and orality in tandem.

First, nasalance is a common measurement defined as the sum of nasal energy over the sum of total energy for a specific interval [6], be it single segment or an entire passage (e.g., the Rainbow Passage, [7]). While this can be used to

quantify vowel nasality (e.g., [8]), it is often used clinically to assess hyponasality or hypernasality (e.g., [9]) and/or for establishing language-specific baselines (e.g., [10]). This formula, however, envisions nasality essentially as a continuum from a completely oral vowel to a completely nasal consonant. That is, any segment's nasalance can only approach 100 as the sum of its oral energy approaches 0. This somewhat complicates a direct comparison among vowels, especially given arguably inherent differences in baseline oral energy along the lines of aperture and centrality (e.g., [11]).

Nasalance can be adopted, however, on a point-by-point basis to arrive at a segment-level score of nasalization. This is exemplified by Rochet & Rochet [12] in their study of regressive nasalization in French and English. An arbitrary threshold of nasalance is used (commonly 50+) to determine whether a given point is sufficiently nasalized, and the sum of these points is taken over the total number of points to express a percentage (see also [13, 14]).

Finally, Dow [15] proposed an area-sum based formula for split-level data, called the Differential Energy Ratio (DER). Essentially, the DER calculates the proportion of nasal-dominant differential energy (nasal minus oral), defined with respect to a similar threshold, to total differential energy (cf. §3.3 for explanation).

A purported advantage of the DER is its greater granularity. The degree of difference between oral and nasal energy is directly incorporated into the final expression of nasality, as opposed to the binary transformation of the nasalance-based approach. On one hand, it is unclear to what extent, if any, this level of detail is exploited in production or perception in human speech. On the other hand, studies have shown that the otherwise significant correlation between the two types of formulae for the same dataset starts to break down for certain vowel qualities in certain phonological contexts. Dow [1] shows that French high vowels in regressive nasalization contexts are judged on average as 13% more nasal by the DER, up to approximately 25%. The cause of this is argued to be the greater susceptibility of high vowels to spontaneous nasalization, given their high intraoral pressure [1]. Meanwhile, Dow [2] finds a similar, if less pronounced, relationship on high vowels and /ø/ in post-nasal contexts.

These previous studies both used data from French and are thus necessarily limited to the vowel inventory of French, as well as its phonotactics (e.g., virtually no nasal vowels before nasal consonants) and other language-specific constraints of controlled vs. mechanical nasalization (e.g., only negligible contextual nasalization of non-high vowels). For these reasons, this study expands upon this previous work by adding a synthesized corpus.

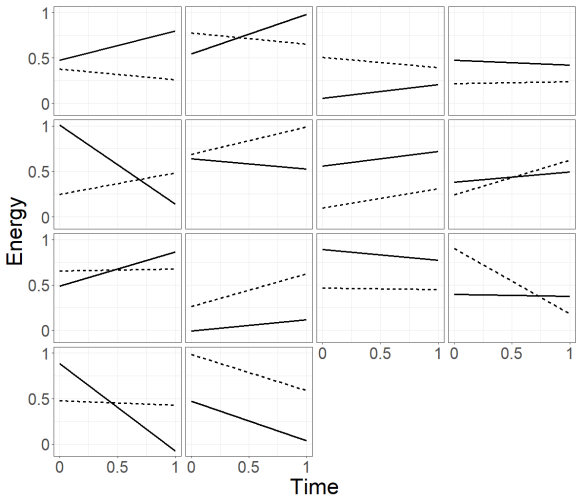


Figure 1: Illustration of all 14 vowel types. Solid line = nasal energy, dashed line = oral energy.

3. METHODOLOGY

3.1. Simulated corpus

A corpus of artificial nasometric data was first constructed in R [16]. Each vowel consisted of two series of simultaneous readings, representing oral and nasal energy, taken at equidistant intervals. A random number of intervals was chosen for each vowel between 7 and 20, essentially approximating differing durations. The initial time stamp for each vowel was a randomly generated number between 0 and 1. (Throughout, the command `runif()` was used to generate random numbers.)

Fourteen types of vowels (simplified illustration in Fig. 1) were predetermined according to every logically possible permutation of the relationships in energy (a, b) between the initial and final points within each set of readings, (c) between the initial point of each set, and (d) between the final point of each set. (Four variables with two outcomes each yields 16, with two impossible types excluded for incompatibility between slope requirements and crossover behaviour.) Essentially, this translates into the sign of the slope of each set (whether energy increases or decreases) and the intercept or lack thereof between the two sets (whether the readings of one always surpass those of the other or not).

The necessary energy relationships for each type were achieved by generating four random points between 0 and 1, sorting them, and assigning their indices as prescribed. For instance, in type 1 (first

panel of Fig. 1), the initial point of the nasal readings was the third (i.e., second highest) of these points, while the final point was the fourth (i.e., highest). The initial and final oral readings of this same type were the second and first of these points, respectively.

After ensuring this relationship, the energy readings themselves were generated for each set by randomly producing n numbers between 0 and 1 (where n = predetermined intervals) between these two points. These points were then sorted for each set of energy readings according to its type (e.g., nasal increasing and oral decreasing for type 1). This was performed an arbitrary number of 1,000 times for each vowel type, yielding 14,000 artificial vowel tokens in the corpus.

3.2. Nasometric corpus

The real-world data used in this study were the same as in [15, 2]. The twenty participants (14 men, 6 women, mean age = 43.4 years) were native speakers of French from France. No significant differences were found in previous analyses along the lines of department of origin, sex or age.

Speakers read aloud a list of French expressions containing one of eleven target vowels (i.e., /a, e, ø, o, i, y, u, ɑ̃, ɛ̃, ɔ̃, œ̃/) either preceded or followed by a nasal or oral consonant, within the limits of French phonotactics. This yielded 72 total expressions. This list was read three times, each in random order, into a Glottal Enterprises hand-held nasometer (NAS-1 SEP Clinic), recorded by Praat in stereo at a sampling rate of 44.1kHz. One vowel was removed for participant error, yielding a total of 4,319 vowels.

Energy readings were extracted from each channel (corresponding to the nasal and oral microphones) at 5 ms intervals. Each channel's energy readings were then min-max normalized separately within each speaker, within each target (see [2] for reasoning and consequences of normalization).

Energy readings were then generalized. First, a linear regression was performed on each set of readings for each vowel. The intercept and slope of each regression were extracted, and the final point of each regression was estimated. Points were then randomly generated for each set (oral and nasal) and sorted according to its regression's slope. Each vowel was then classified by type (cf. §3.1). This procedure of was performed to homogenize data sources and diminish the effect of rare outliers (due to abnormally close oral-nasal readings).

Finally, the temporal (x-axis) intersection of the randomly generated oral and nasal energy readings was performed using the `line.line.intersection()` function [17] and expressed as a percentage over the total number of points for each vowel. This essentially estimates the percentage of a vowel's duration at which one sort of energy surpasses the other (if at all). This was also performed for the simulated dataset.

3.3. Formulae

Two formulae of "global" vowel nasality (i.e., the nasality of the vowel as expressed by a single number) were applied to the vowels of both corpora, namely the Differential Energy Ratio (DER) and a nasalance-based ratio along the lines of Rochet & Rochet [12]. The latter formula does not appear to have a standard name in the literature and is hereafter abbreviated NBR.

First, the DER was calculated in accordance with [15, 2]. At each point, nasal energy (y) was subtracted from oral energy (x) to provide differential energy. The absolute value of the sum of all negative differential energy was then divided by this same number plus the sum of all positive differential energy. This number was then multiplied by 100 to resemble a percentage. This is expressed by the formula in (1), where i is any measured point until the end of the vowel in question.

(1)

$$DER = 100 \times \frac{|\sum_i min(x_i - y_i, 0)|}{|\sum_i min(x_i - y_i, 0)| + \sum_i max(x_i - y_i, 0)}$$

The NBR was then calculated in the following way, as shown in (2). At each point N, the nasalance was calculated by dividing nasal energy (x) over total energy (x + y). The sum of points whose nasalance was equal to or above 0.5 was then divided by the total number of points. This number was then again multiplied by 100.

$$(2) \quad NBR = 100 \times \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{x_i + y_i} \geq 0.5$$

4. RESULTS

4.1. Control vowels

We examine control vowels first, that is, vowels without intersecting oral and nasal energy lines for the simulated dataset or phonemically oral vowels in non-nasal contexts for the French data. DER and NBR scores for control vowels in the simulated dataset unsurprisingly showed perfect agreement,

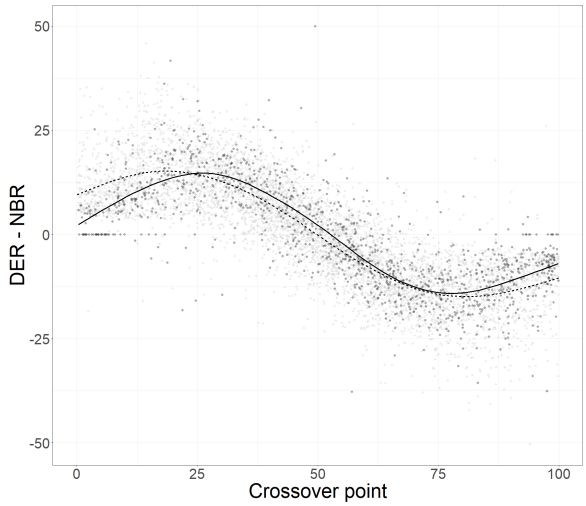


Figure 2: DER – NBR vs. crossover point. Synthesized data = darker dots, dashed line; French data = lighter triangles, solid line.

both being on average 0 or 100, depending on which type of energy dominated. On average, in the French dataset, nasal vowels have an average DER of 85.2 and NBR of 84. Oral vowels in non-nasal contexts have an average DER of 2.5 and NBR of 2.85. Essentially, both formulae agree and function as expected on these ceiling- and floor-effect tokens.

4.2. Contextual nasalization

In contextually-nasalized vowels, the NBR proved to have a linear relation with crossover point (the estimated point at which nasal energy overtakes oral), with minimal difference between corpora. This is somewhat intuitive, given the definition of the formula; the relationship, however, is imperfect, as estimated crossover points occur invariably happens between measured points, while the NBR necessarily references the first point after that crossover point. Meanwhile, the DER shows a more distinct curve towards the limits of the crossover point, such that vowels with a crossover point towards their beginning or end show floor and ceiling effects (e.g., vowels nasalized in their last quarter approach 0 DER).

Figure 2 visualizes the difference of the two (DER minus NBR) as a function of crossover point (y-axis and x-axis, respectively). In the interest of space, the crossover point of progressive nasalization is visualized with respect to the right edge, as both types of nasalization proved to be horizontal mirror images. (For instance, a vowel whose oral energy overtakes nasal energy at 25% of the vowel is given a "crossover point" of 75%.) A distinct sinusoidal relationship can be seen, such that vowels with a longer nasal phase (whether nasalized early for VN or denasalized late for NV) are judged as more nasal by the DER (max conditional means of 14.8 points for the real dataset at 26% and 15.3 at 18.5% for the synthesized dataset). Meanwhile, vowels with a shorter nasal phase (late-nasalized VN or early-denasalized NV) are judged as more nasal by the NBR to a nearly equal degree (min conditional means of -14.2 points at 78.1% for the real dataset and -14.9 at 81.2% for the synthesized dataset).

Finally, in the French data, an anomaly which bears further investigation is the behaviour of post-nasal /i/, which showed a higher difference between the DER and NBR in the 25-50% and the 50%-75% ranges than all other qualities combined. In the second quarter, /i/ tokens had on average a difference of 6.9 points, indicating the DER is higher than the NBR, versus -10.4 for all vowels combined. In the third quarter, /i/ had an average difference of 13 points, versus the 6.7 points of other qualities.

5. DISCUSSION & CONCLUSION

In this study, the DER and the NBR differed on contextually-nasalized vowels as a function of crossover point, such that higher nasality leads to higher differences in favour of the DER, and lower nasality in favour of the NBR. For instance, if nasal energy overtakes oral early within a pre-nasal vowel (VN), that vowel is likely to be judged as more nasal by the DER; the same is true for post-nasal (NV) vowels whose oral energy overtakes nasal relatively late. That this effect was found in both corpora suggests this difference is more inherent to the formulae themselves. It is surprising that the temporal component still appears to be crucial, given the attempts of the DER to divorce itself from the temporal and to model degrees of energy difference. The results of post-nasal French /i/ do suggest, however, that languages may still exploit degree of nasality (and/or velocity of change) beyond temporal properties.

In the future, changepoint detection may give us greater nuance and thus fill a gap observed in real-world data and provide a more accurate "stress test" for these formulae. Furthermore, and more crucially, perceptual research is needed to determine whether one formula more accurately reflects natural speech. Examining these questions may help us better understand vowel nasality in general, as well as providing mathematically accurate benchmarks upon which to characterize phonetic or phonological processes in language, as well as performance in clinical settings (with respect to hypo- and hypernasality, for instance).

6. REFERENCES

[1] M. Dow, "Temporal vs. area-sum measurements of vowel nasality," 2016, Annual meeting of the Linguistic Society of America.

[2] ——, "A phonetic-phonological study of vowel height and nasal coarticulation in French," *Journal of French Language Studies*, vol. 30, no. 3, pp. 239–274, 2020.

[3] C. Carignan, "A practical method of estimating the time-varying degree of vowel nasalization from acoustic features," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 911–922, 2021.

[4] R. A. Krakow and M. K. Huffman, "Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory: An introduction," in *Phonetics and Phonology, vol. 5: Nasals, Nasalization and the Velum*, M. K. Huffman and R. A. Krakow, Eds. Academic Press, 1993, pp. 3–59.

[5] V. Delvaux, *Les voyelles nasales du français : Aérodynamique, articulation, acoustique et perception*. Bruxelles: Peter Lang, 2012.

[6] S. G. Fletcher, ""Nasalance" vs. listner judgements of nasality," *The Cleft Palate Journal*, vol. 13, pp. 31–44, 1976.

[7] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, pp. 127–127, 1960.

[8] N. Audibert and A. Amelot, "Comparison of nasalance measurements from accelerometers and microphones and preliminary development of novel features," in *INTERSPEECH 2011–12th Annual Conference of the International Speech Communication Association*, 2011, pp. 2825–2828.

[9] M. A. Hardin, D. Van Demark, H. L. Morris, and M. M. Payne, "Correspondence between nasalance scores and listener judgments of hypernasality and hyponasality," *The Cleft palate-craniofacial journal*, vol. 29, no. 4, pp. 346–351, 1992.

[10] T. Sweeney, D. Sell, and M. O'Regan, "Nasalance scores for normal-speaking Irish children," *The Cleft palate-craniofacial journal*, vol. 41, no. 2, pp. 168–174, 2004.

[11] I. Lehiste, *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

[12] A. P. Rochet and B. L. Rochet, "The effect of vowel height on patterns of assimilation nasality in French and English," in *Pr. 12th ICPhS*, Aix-en-Provence, 1991, pp. 54–57.

[13] B. L. Rochet and F. Yanmei, "Acoustic measurements of vocalic nasality in Mandarin Chinese," *Canadian Acoustics*, vol. 20, no. 3, pp. 53–54, 1992.

[14] A. P. Rochet and B. L. Rochet, "Patterns of assimilation nasality in English as a function of vowel height," in *Pr. 14th ICPhS*, vol. 14, 1999, pp. 699–702.

[15] M. Dow, "Contrast and markedness among nasal(ized) vowels: A phonetic-phonological study of French and Vimeu Picard," Ph.D. dissertation, Indiana University, 2014.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[17] D. C. Sterratt, D. Lyngholm, D. J. Willshaw, and I. D. Thompson, "Standard anatomical and visual space for the mouse retina: computational reconstruction and transformation of flattened retinae with the retistruct package," *PLoS Computational Biology*, vol. 9, no. 2, p. e1002921, 2013.

EVOLUTION OF VOICES IN FRENCH AUDIOVISUAL MEDIA ACROSS GENDERS AND AGE IN A DIACHRONIC PERSPECTIVE.

Albert Rilliard[1,2], David Doukhan[3], Rémi Uro[1,3], Simon Devauchelle[3]

[1] Université Paris Saclay, CNRS, LISN, Orsay, France; [2] Federal University of Rio de Janeiro, Brazil; [3] French National Institute of Audiovisual, Paris, France
albert.rilliard@lisn.upsaclay.fr, {ddoukhan,ruro,sdevauchelle}@ina.fr

ABSTRACT

We present a diachronic acoustic analysis of the voice of 1023 speakers from French media archives. The speakers are spread across 32 categories based on four periods (years 1955/56, 1975/76, 1995/96, 2015/16), four age groups (20-35; 36-50; 51-65, >65), and two genders. The fundamental frequency (F_0) and the first four formants (F1-4) were estimated. Procedures used to ensure the quality of these estimations on heterogeneous data are described. From each speaker's F_0 distribution, the base-F_0 value was calculated to estimate the register. Average vocal tract length was estimated from formant frequencies. Base-F_0 and vocal tract length were fit by linear mixed models to evaluate how they may have changed across time periods and genders, corrected for age effects. Results show an effect of the period with a tendency to lower voices, independently of gender. A lowering of pitch is observed with age for female but not male speakers.

Keywords: Gender, Diachrony, Vocal Tract Resonance, Vocal register, Broadcast speech

1. INTRODUCTION

Vocal characteristics are an important part of identity, with our voices indicating our gender and other social constructs [1]. Voice obviously changes with age, a lot during our development until adulthood [2], but also later with aging voices [3]. Social constructs, and the acoustic characteristics linked to them, may change with cultures – as was described by van Bezooijen for Japanese and Dutch female voice [4]: it is thus essential to study these characteristics in varied cultural and language contexts (and it typically fits this ICPhS theme). Some studies also claimed that voices may have changed across time (for populations of comparable age); this was claimed for female voices, for example in Australia [5].

We present here a first acoustic analysis of a corpus of voices extracted from French broadcast archives in a diachronic perspective, trying to balance the selection of speakers according to their age and gender. A total of 1023 speakers were selected, and samples of their vocal production were extracted from the original archive. This corpus aims to extract acoustic cues from a large sample, representative of female and male voices presented in French media, to describe them and their possible changes with time and age. Some studies have described diachronic changes of voices across genders, but these studies are few, may have varying conclusions or population sampling, and address populations from different cultural backgrounds. [5] addresses young adult women's voices, with two points in time – while [6] studied changes in male voices over five time periods, focusing on media anchors: comparatively few speakers were included (four for the first three periods), with possible longitudinal effects. For French, [7] selected short samples of voices of anonymous individuals from media archives (mean duration of 5 s.), comparing F_0 of males and females across 70 years. They report complex patterns of F_0 changes across time, with possible opposite tendencies for genders after the 1960s, but don't control for age, a factor known to induce changes in voice F_0 [8]. Vocal characteristics are also affected by vocal effort [9], which has substantial effects on F_0 [10] and on formants [11].

Long-term acoustic measures give reliable information on voice characteristics. [12] showed the long-term average spectrum stabilizes for articulatory changes with about 10 seconds of voiced speech. [13] compared several measures linked to register and found the base-F_0 [14] was faster to stabilize (with less than ten s. of voiced speech) than mean or median F_0. A study [15] linked to the estimation of the vocal tract length (VTL) from formants showed the method proposed in [16] was the most rapidly stable, even on a reduced dataset. The literature reports that gender is perceived notably through two acoustic cues – F_0 and supraglottal resonances [17, 18].

The remaining parts of the paper present estimates

of F_0 and VTL made on a corpus collected following the semi-automatic strategy proposed by [19] to gather speech samples from speakers across 60 years period, found in TV and Radio archives of the French National Institute of Audiovisual (INA) (speakers belong to different categories of age and gender). The acoustic measures related to the speakers' voice pitch are fitted by linear mixed models evaluating potential changes across time for both genders and controlling for age.

2. METHODS

2.1. INA's diachronic corpus

INA's diachronic corpus contains speakers of both genders, spread across four age categories (20-35, 36-50, 51-65, over 65 years old) and four time periods (1955-56, 1975-76, 1995-96, and 2015-16). With an initial goal of gathering at least 30 individuals for each of these 32 categories of age, gender, and time period, samples of voice for 1023 speakers were collected (see Table 1). Some of these categories being much less present in media (typically women), finding target speakers from all age categories in the earliest periods was challenging. It was thus mandatory to increase the targeted period for the 1955-56 and 1975-76 periods by considering the years between 1954-1957 and 1974-1977 (note the additional years only represent a small part of the selected speakers so that these periods will be referred to with the originally targeted years for simplicity). The collection was done thanks to INA's archivists identifying specific women and men within the four age categories and featuring in programs from the four given targeted time periods. Following the method described in [19], the programs featuring target speakers were submitted to a diarization process, and the ID corresponding to each target speaker was then hand-picked. These speakers' samples were then selected to keep samples with minimal noise or background music and remove silences.

These archive extracts were submitted to procedures to discard excerpts with adverse characteristics. `LIUM_SpkDiarization` [20] was used to reject segments associated with a telephone quality. Speakers with less than 10 seconds of valid pitch estimates were rejected (the filtering protocol is detailed in section 2.3). Table 1 presents the distribution of the speakers. Voice samples from 1023 speakers were obtained from 878 radio or TV programs (the voice of 85 persons was collected from more than one program). The median duration of valid acoustic features per speaker was 125 seconds. The amount of unique speakers in the 32 categories varies between 15 and 74.

	20-35		36-50		51-65		>65	
	F	M	F	M	F	M	F	M
1954-57	17	41	22	74	18	50	17	15
1974-77	18	17	23	41	28	39	20	26
1995-96	32	31	32	46	29	47	29	35
2015-16	30	31	30	52	28	48	29	31

Table 1: Number of speakers per time period (rows), age group, and gender in the corpus

2.2. Voicing, F_0 and Formant estimation

To estimate robust F_0 and formant measurements from widely different materials (due to heterogeneity in recording and archival conditions), the voicing decision, and then F_0 and formant estimations, were made by concurrent algorithms. The `Spleeter` source separation framework was used to separate voice from other phenomena (music or noise) [21]. The acoustic features were then estimated twice, from both the original and speech-separated signals. For voicing and F_0 estimation, [22] showed good performances of `Praat` auto-correlation (ac) algorithm; it also shows that these estimations could be improved in noisy conditions by combining `REAPER`'s voicing estimation with `FCN-F0`'s F_0 estimation (based on neural-network). We thus used `Praat` [23] for estimating voicing and F_0 (ac algorithm with a 65-650 Hz F_0 range) and for estimating the first four formants (using the recommended settings for female and male voices: 5 formants for a respective ceiling frequency of 5.5 or 5kHz). `REAPER` algorithm [24] was used with two distinct settings (default and Hilbert transform) to obtain two other voicing estimates. `FCN-F0` was used with default pitch range (30-1000 Hz), and Viterbi smoothing to estimate F_0 [25].

2.3. Acoustic features filtering

The frames detected as voiced by the six estimations (`Praat` and `REAPER` with two sets of parameters on the raw and separated signals) were kept. From these frames, those where the four F_0 estimations (by `Praat` and `FCN-F0` on raw and separated signals) were below a 20% gross error rate threshold were kept. This strategy allowed us to keep 52.2% of the signal's original frames instead of 61.1% obtained with `Praat`'s `ac` algorithm. From these frames, the formant estimates made on both signals (raw and source-separated) were compared to keep only

frames with four formant estimates, with variations below a 20% gross-error-rate threshold (keeping about 96% of the preceding frames) and 50.2% of the signal's original frames.

2.4. Long-term features

For each speaker, the valid samples (cf. above section) were grouped in chunks of at least 10s; i.e., for a 36s extract, three chunks of 12s were used. From each of these chunks, two long-term estimates were made: the speaker's base-F_0 (defined as the seventh decile of the F_0 distribution on the chunk [14, 13]), and the speaker's vocal tract length (equal to the median of the length estimations made on each frame based on the first four formants, using the equation proposed in [16, 15]). Let's make it clear the estimation of the "vocal tract length" is a way to evaluate the tendency one speaker has to produce higher or lower resonances (formants) as a result of their articulatory habits; the relation to actual vocal tract length is undoubtedly more complex [26]. For each speaker, these two long-term estimates (base-F_0 and vocal tract length: VTL) were estimated for each chunk of voiced samples. The median number of chunks was 14 by speaker, ranging from 1 to 175.

2.5. Statistical analysis

The base-F_0 and the estimated VTL were fitted with two linear mixed models that took as fixed effects the actual age (in years) of the speaker, the time period of the recording (four levels: 1955-56, 1975-76, 1995-96, 2015-16), and the speaker's gender – and as random effects the speaker for which the measure was done, and the media program from which the corresponding speaker's chunk was extracted (the program factor was nested in the speaker factor). Following [27], a maximal model was fit (using R's `lme4` library [28]) that included all the interactions between the fixed factors. This model was then submitted to a simplification procedure to remove non-significant terms and reach a minimal adequate model, one for F_0 and one for VTL. These two models are used here to describe the results of the variation of the base-F_0 and VTL across age, gender, and time period.

3. RESULTS

3.1. Base-F_0

The minimal model fitted on base-F_0 was based on the three main fixed factors (`Age`, `Period` and `Gender`), with the `Age:Period` and `Age:Gender` interactions, plus the original random structure (`Program` nested in `Speaker`). An estimation of the conditional and marginal coefficients of determination (using [29]) for this model showed the model explained about 90% of the variance, with fixed factors responsible for 58% of this. Figure 1 presents the effect of the `Age:Gender` interaction on base-F_0: there is a major difference in F_0 across genders (about 9 st), and F_0 decreased with age for female voices. Figure 2 presents the `Age:Period` interaction: while Base-F_0 was increasing with Age in the 1950s, it showed a tendency to decrease with age at later periods (1995-96 and 2015-16). Note that this second interaction has a much smaller effect size than the `Age:Gender` one.

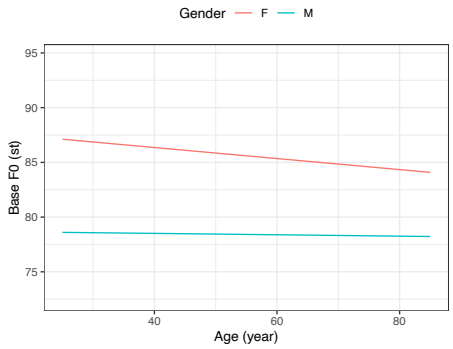


Figure 1: Fit of base-F_0 (in semitones) from the Speaker Age and Gender.

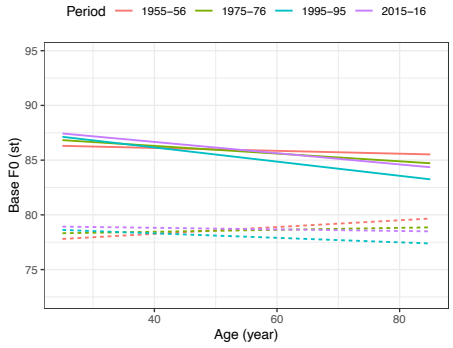


Figure 2: Fit of base-F_0 (semitones) for speaker Age by Period and separated by Gender (males: dashed lines) to show the relative effect of the Period slopes on voices of each gender.

3.2. Vocal tract length

The minimal model fitted on VTL was based on the three fixed factors without interaction and keeping the original random structure. The conditional and

marginal coefficient of determination for this model showed a much smaller part of the variance was explained by the factors, with the complete model explaining about 38% and fixed factors 31%. The `Gender` had the largest effect size, with a mean difference of estimated VTL of 1.7 cm between females and males. `Period` was the next important factor, with a modest but significant increase in VTL of 0.13 and 0.2 cm for the 1995-96 and 2015-16 periods, compared to 1955-56. `Age` showed a significant but small (0.03 cm for ten years) increasing slope of estimated VTL with age.

4. DISCUSSION & CONCLUSION

This study intends to bring discussions and gather remarks and advice from the community by presenting preliminary results on a complex topic related to the perception of voice quality, and importantly voice pitch, across gender, age, and time periods in France. As the material used to extract the stimuli could not, for obvious reasons, be of homogeneous quality (recording conditions, signal quality, etc.), there are a series of limitations related to this approach. A problem is linked with the estimation of vocal tract resonances, as highlighted, e.g., by [26]. The quality of archive acoustic quality further complicates this: we had to implement a series of checks to assert we didn't process voices with telephone quality or other deterioration linked to, e.g., compression.

The question of the formant estimation algorithm's parameters is essential, as it may lead to varying results. The data presented here are based on the default parameters recommended by `Praat` for each gender, but this is a potential bias for estimating VTL in speakers with non-standard characteristics. We also tested the parameters recommended by [16] (estimate six formants for a 5.5kHz ceiling frequency): this option led to very different results – and a complex interaction between `Gender`, `Age`, and `Period` with some cases of female speakers having longer estimated VTL than male ones. It was thus not kept here, but questions remain on the estimation of VTL. While distinct settings are generally recommended for the analysis of male and female voices (frequency range for F_0 estimation, frequency ceiling for formats), the use of these a priori settings for the investigation of gender characteristics is questionable since voices, recording, and archival strategies may have changed over the last 60 years – and because it may hide some non-standard features.

Conversely, measures of F_0 seem much more robust and show trends that confirm the literature, with an effect of age on female voice [3, 30], and a clear gender difference for gender representation in French media. One finding of this study is linked to a change in the evolution of base-F_0 with age across time periods: while it tends to increase in the 1950s for males or stay almost constant for females, steeper slopes were observed in the two later periods (1995-96, and 2015-16). This evolution is not dependent on gender (the interaction was not significant), but the slopes are different for each gender, as shown by figure 2. This decreasing tendency raises questions on the use of voice in public displays (varying social use [31] or changing health conditions?), with lowered voices for older speakers across time periods.

5. ACKNOWLEDGEMENTS

This work has been funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012). The authors are deeply indebted to INA's archivists Laëtitia Larcher and Anissa-Claire Adgharouamane and who selected the target speakers and also wished to thank Luc Ardaillon and Jean-Hugues Chenot for their valuable advice.

6. REFERENCES

[1] D. C. Sergeant and G. F. Welch, "Gender differences in long-term average spectra of children's singing voices," *Journal of Voice*, vol. 23, no. 3, pp. 319–336, May 2009.

[2] M. Fouquet, K. Pisanski, N. Mathevon, and D. Reby, "Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood," *Royal Society Open Science*, vol. 3, no. 10, p. 160395, Oct 2016.

[3] A. Russell, L. Penny, and C. Pemberton, "Speaking fundamental frequency changes over time in women: A longitudinal study," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 101–109, Feb 1995.

[4] R. van Bezooijen, "Sociocultural aspects of pitch differences between japanese and dutch women," *Language and Speech*, vol. 38, no. 3, pp. 253–265, 1995.

[5] C. Pemberton, P. McCormack, and A. Russell, "Have women's voices lowered across time? a cross sectional study of australian women's voices," *Journal of Voice*, vol. 12, no. 2, pp. 208–213, Jan 1998.

[6] Y. Zou, Y. Wang, and W. He, "Diachronic contrastive analysis on read speech in broadcast news: Evidence from pitch and duration," in *2012 8th International Symposium on Chinese Spoken*

Language Processing. IEEE, 2012, pp. 291–295.
[7] A. Suire and M. Barkat-Defradas, "Evolution of human pitch: Preliminary analyses in the french population using ina audiovisual archives of vox pops," in *2020 IASA-FIAT/IFTA Joint Conference*, 2020.
[8] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, Aug 2011.
[9] M. Berg, M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, and T. Berger, "The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors," *Journal of Voice*, vol. 31, no. 2, pp. 257.e13–257.e24, Mar 2017.
[10] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, May 1992.
[11] A. Rilliard, C. d'Alessandro, and M. Evrard, "Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 109–122, Jan 2018.
[12] A. Löfqvist and B. Mandersson, "Long-time average spectrum of speech and voice analysis," *Folia Phoniatrica et Logopaedica*, vol. 39, no. 5, pp. 221–229, 1987.
[13] P. Arantes and A. Eriksson, "Temporal stability of long term measures of fundamental frequency," in *Speech Prosody 2014*. ISCA, May 2014, pp. 1149–1152.
[14] H. Traunmüller and A. Eriksson, "The perceptual evaluation of f_0 excursions in speech as evidenced in liveliness estimations," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1905–1915, 1995.
[15] K. Johnson, "Vocal tract length normalization," *UC Berkeley Phonology Lab Annual Reports*, vol. 14, 2018.
[16] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLOS ONE*, vol. 10, no. 7, p. e0132193, Jul 2015.
[17] K. Pisanski and D. Rendall, "The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2201–2212, Apr 2011.
[18] Y. Leung, J. Oates, S.-P. Chan, and V. Papp, "Associations between speaking fundamental frequency, vowel formant frequencies, and listener perceptions of speaker gender and vocal femininity-masculinity," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2600–2622, Jul 2021.
[19] R. Uro, D. Doukhan, A. Rilliard, L. Larcher, A.-C. Adgharouamane, M. Tahon, and A. Laurent, "A semi-automatic approach to create large gender- and age-balanced speaker corpora: Usefulness of speaker diarization & identification," in *13th Language Resources and Evaluation Conference*, 2022, pp. 3271–3280.
[20] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.
[21] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.
[22] R. Vaysse, C. Astésano, and J. Farinas, "Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 3091–3101, 2022.
[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.09)," 2020. [Online]. Available: http://www.praat.org
[24] D. Talkin, "Reaper: Robust epoch and pitch estimator," 2015. [Online]. Available: https://github.com/google/REAPER
[25] L. Ardaillon and A. Roebel, "Fully-Convolutional Network for Pitch Estimation of Speech Signals," in *Proc. Interspeech 2019*, 2019, pp. 2005–2009.
[26] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.
[27] S. T. Gries, *Statistics for linguistics with R: a practical introduction*, 3rd ed., ser. De Gruyter Mouton textbook. Berlin Boston: de Gruyter Mouton, 2021.
[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[29] K. Bartoń, *MuMIn: Multi-Model Inference*, 2022, r package version 1.47.1. [Online]. Available: https://CRAN.R-project.org/package=MuMIn
[30] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, and T. Yamasoba, "Quantitative analysis of digital videokymography: A preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers," *Journal of Voice*, vol. 29, no. 1, pp. 109–119, Jan 2015.
[31] P. Boula de Mareüil, A. Rilliard, and A. Allauzen, "A diachronic study of initial stress and other prosodic features in the french news announcer style: corpus-based measurements and perceptual experiments," *Language and Speech*, vol. 55, no. 2, pp. 263–293, 2012.

DIFFERENCES IN ADULT SPEECH: DOES BOY-DIRECTED AND GIRL-DIRECTED SPEECH EXIST

Tori Keeping, Tyler Bonnell, Cassandra Robertson & Fangfang Li

University of Lethbridge
tori.keeping@uleth.ca, tyler.bonnell@uleth.ca, cassandra.robertson@uleth.ca & fangfang.li@uleth.ca

ABSTRACT

It has been demonstrated that boys and girls speak differently on a phonetic level; however, the question of whether this is directly a result of the input they receive from adults is yet to be determined. The current study aims to explore potential differences in how adults speak to boy versus girl children. Forty university students participated in a speech production study in which they were asked to read a short passage to different gendered images of babies. Acoustic analysis was conducted on their recorded speech, with measures including the centroid frequency of the sound /s/, as well as f0 and intensity of the read passages. No robust differences were found in any of the acoustic parameters examined, suggesting that the ways adults converse with boy versus girl babies may not be substantially different.

Keywords: gendered speech production, acoustics, phonetics

1. INTRODUCTION

Speech production between males and females has been shown to differ in several aspects, resulting in gendered patterns in adult speech. Differences exist in the distinct stylistic choices that males and females use in speech such that females tend to use more affiliative speech and less assertive speech [1] and use more intensifiers on average [2] than men. In addition, women are more tentative (more uncertain) when discussing stereotypically masculine topics in intergroup contexts and vice versa [3]. Further, females tend to have a higher pitched and breathier voice [4, 5], greater speech intelligibility [6] and larger acoustic vowel space size than men [7]. Moreover, females have been found to produce clearer speech [8] which is exemplified through making greater vowel acoustic changes such as hyperarticulated vowels [9] and have a slower articulation rate compared to males [10]. A large part of these differences are due to anatomical differences which include males having a longer vocal tract [4] and having larger and thicker vocal folds than women, resulting in a lower pitch in males [7]. In addition, adult females were shown to produce a more anterior variant of /s/ [11] and longer closure time for stop consonants than males in German [12] and some varieties of English [13, 14]. In Contrast, males have longer closure time for aspirated consonants in Korean [15].

Similar gendered patterns have also been observed in the speech of young English-speaking children. Children have been shown to speak with distinctive speech styles depending on their gender in various languages [16, 17] and that boys and girls have different average pitch values [18]. Further, girls aged 6-14 showed longer closure time for stop consonants [14] as well as a more anterior variant of /s/ than same aged boys [11] in English. However, boys have a longer closure time for aspirated consonants than girls in Korean [19].

One potential mechanism that may account for the early emergence of gendered speech in children is through role modelling or imitation of the same-sex adults in their environment. As mentioned previously, men and women exhibit different ranges of pitch which children are able to recognize and imitate. According to [20], when imitating different genders, both boys and girls, aged 6-10 years old, raised their pitch for females and lowered their pitch for males. Importantly, adults were shown to be able to identify the gender of children as young as 2.5 years of age above chance based on the child's speech [21]. The fact that gender-specific speech patterns can be detected before puberty suggests that gendered speech is not solely due to sex dimorphism, and instead is learned through social learning such as role modelling or imitation [22].

Another possible mechanism contributing to these gendered differences in children's speech may be due to different input that they receive depending on their gender. That is, adults may alter their speech based on the sex of the child they are speaking to. Adults have been shown to accommodate their conversational partners to facilitate speech communication with people of varying gender [23], age [24], attractiveness [25] and language background [26]. In particular, they modify their speech style when addressing young children, such as using shorter and simpler words and speaking with a clearer and higher pitch as compared to their speech to other adults [27].

Moreover, the type and style of conversations of mothers and fathers varies based on their child's gender [28]. Mothers talk to their daughters more than their sons [29] as well as ask more questions to their daughters than their sons [30]. Further, adults utilize different types of words with male versus female children. For example, parents use more spatial words (describing the size, shape and spatial properties of spaces and objects) when talking to boys than when talking to girls [31].

With specific regard to phonetic details of individual speech sounds that adults produce when speaking to children, evidence is sparse. One notable study is Foulkes and Docherty's study [32]. This study found differences in /t/ production in mothers when speaking to boys or girls. More specifically, they discovered that mothers of boys equally used standard (plain) and nonstandard (laryngealized) variants of /t/ in word-medial position whereas mothers of girls used plain [t] twice as frequently as the laryngealized form in word-medial positions in their speech [32]. That is, mothers use the standard production of /t/ more frequently with their daughters than with their sons. Although Foulkes and Docherty [32] were able to find this gender specific speaking pattern in mothers, other evidence surrounding these patterns in relation to phonetic level differences is lacking. Thus, it still remains a question as to whether differences in input between boys and girls at a phonetic level are a true mechanism of the emergence of gendered speech patterns that emerge in childhood.

The purpose of the current study seeks to further investigate whether adults alter their speech depending on whether they are conversing with boy or girl children. Specifically, this study aims to examine the phonetic details in adult speech focusing on their voice characteristics and the sound of /s/, the two aspects of speech that young children have shown early gender-specific patterns.

2. METHODS

2.1. Study design

University students (N = 40, Females = 26) were recruited through the Psychology subject pool to participate in a speech production study. Participants were asked to read the shortened version of "The North Wind and the Sun" passage [33] to photos of babies of different genders aged one to two years old. The photos were edited such that pairs of baby girls and boys shared the same face to eliminate any confounds of facial features. There are a total of three sets of baby photos (n = 6) presented in randomized order to each participant. Participants were given a scenario in which they imagined themselves as a volunteer at a fictional dayhome and were asked to read passages to the babies at that dayhome.

During the task, each baby was introduced by a name that clearly indicated their gender as well as a sentence describing their gendered play behaviour. For example, "This is Sophia, she really likes to play with dolls." Participants were then asked to greet each baby by name and read them the passage from an illustrated booklet. Their speech was recorded for later acoustic analysis. Finally, participants filled out two questionnaires, one of which assessed basic demographic information and experience with children. The other questionnaire used in [34] was adopted to gather information on participants' communication and interaction style with children.

It should be noted that similar studies were conducted using photos of older-aged children; however, for the purpose of this paper, only the results of the babies will be reported.

2.2. Procedure analyses

Speech data from the 40 participants was analyzed using Praat [35]. Three acoustic parameters were measured: the centroid frequency of the sound /s/ in 9 tokens that began with /s/, the average f0 and the intensity. Each instance of /s/ in word initial position included in the passage was annotated and centroid frequency (m1, the first spectral moment in moments analysis, [36]) values of the /s/ snippet were calculated from the middle 40-ms window extracted from the frication portion of the /s/ sound. In addition, average f0 and RMS intensity of the reading passages as well as the greetings were also measured. It is to be noted that some adults produce /s/ with a [ʃ] variant in words such as "stronger", and those instances were eliminated from further analysis.

3. RESULTS

Analysis of the m1 values of /s/ were performed with a mixed effects Bayesian model using the "brms" package in R [37]. This model estimates the probability of m1 differences when individual adults are addressing male or female babies where the dependent variable is the m1 values, and the fixed effects are speaker gender (male vs. female), baby gender (male vs. female), the interaction between baby gender and speaker gender and presentation order of the photos (1st to 6th presentation).

The model in the analysis has random slopes for baby gender and presentation order that vary with individual speakers, as well as random slopes for speaker gender and baby gender that vary with individual words. The model also has a random

intercept of face ID (baby faces 1-3), participant ID and individual words.

(1) brm(m1 ~ babyGender * speakerGender + babyPresOrder + (1+babyPresOrder+babyGender|ParticipantID) + (1|faceID) + (1+babyGender+speakerGender|word)

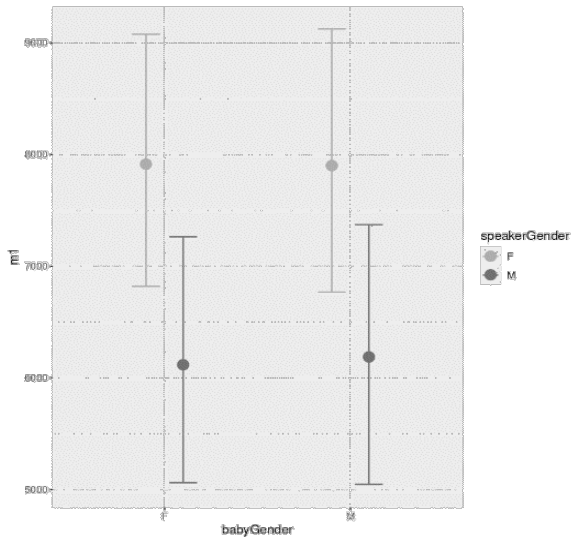


Figure 1: Conditional effects plot from the mixed effects Bayesian model output which reflects m1 values for male and female speakers when they are addressing either male or female babies.

This model ran using 4 chains and 2000 iterations. It converged well with all R^s equalling 1.00. A posterior predictive check (pp_check) was assessed as well, which showed that the model fits well with the data (figure 2).

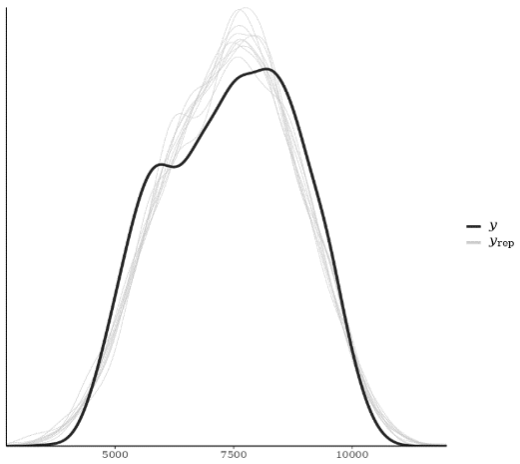


Figure 2. A posterior predictive check of the model where the light grey line represents predicted data, and the dark grey line indicates actual data. M1 values are represented on the x axis.

Population level effects are reported in table 1 which show that baby gender was not significant (-15.04, 95% CI: -353.31, 394.14). Significance is determined in Bayesian statistics through the credible intervals. When the upper and lower credible intervals are either both positive or both negative (i.e. do not cross zero), then the variable in the model is said to be significant [38]. Speaker gender was the only significant factor in this model as the credible intervals do not cross zero. Similar Bayesian models (4 chains, 2000 iterations) were constructed to analyze pitch and intensity of each greeting and passage respectively, in which all R^s converged at less than 1.02. Results did not meet significance for passage pitch (-0.30, 95% CI: -1.20, 0.59) or passage intensity (-0.02, 95% CI: -.017, 0.12) when examining the effect of baby gender. However, baby gender was significant for greeting pitch (5.90, 95% CI: 0.68, 11.09) and greeting intensity (0.84, 95% CI: 0.27, 1.41), suggesting that there may be a difference in how adults greet male babies versus female babies. Despite this significance, the effect size is small and therefore is not substantial evidence for concretely saying adults speak to male versus female babies differently.

Parameter	Estimate	Est. error	Lower 95% CI	Upper 95% CI
Intercept	7940.23	545.57	6842.33	9097.02
Baby gender (M)	-15.04	182.12	-353.31	394.14
Speaker gender (M)	-1799.08	306.97	-2384.12	-1187.98
Presentation order	-3.65	9.96	-23.36	15.46
Interaction	81.62	70.41	-53.24	219.13

Table 1: Population level effects of the current model that assessed baby gender, speaker gender, the interaction between speaker gender and baby gender and presentation order.

4. Discussion

The current study aims to investigate whether adults speak differently to boy babies than they do to girl babies, and whether those differences, if they exist, could be detected at a phonetic level. The results of this study show no robust differences in the acoustic parameters measured in adult speech to babies. Several elements in the current study may provide an explanation for this outcome. First, the repetition within the study may have induced a fatigue effect that could have diminished any results, had there been any evidence in the acoustic recordings. Second, the study was not conducted in a naturalistic environment with adults interacting with real babies. Due to the recent pandemic, we were not able to recruit parent and baby participants and conduct testing in their homes. The alternative design of reading to baby photos in a lab setting is not an authentic interaction. The artificial interaction with children may have prevented us from finding the potential gendered differences in adult speech when they address babies of different genders.

Similar findings were reported by a recent study [29] that compared the ways of word and utterance use in which mothers and fathers talked to babies aged 6-18 months old in their homes. They found no difference in the use of noun type, noun token, and utterance type when examining the speech produced by mothers and fathers when speaking to boys versus girls. These findings collectively suggest that adults may not produce sounds and words in substantially different ways when speaking to baby boys versus. baby girls. It remains a possibility that such input differences could occur later in a child's development after they become a true conversational partner. Ongoing research in our lab is currently testing this hypothesis.

5. Conclusion

Multiple mechanisms may exist for the emergence of gendered speech patterns in childhood. This study aimed to explore one such mechanism, namely, whether there are any input differences at the phonetic level when adults address children of different genders. That is, whether girl-directed vs. boy-directed speech exists. Although no substantial support for this stance has been found in the current study, further research is ongoing to investigate the extent to which adults may modify their speech depending on a child's gender. Thus far, there is inconclusive evidence suggesting phonetic differences in speech when adults address boy vs girl children. Further investigation surrounding this topic would be beneficial to allow for more definitive conclusions.

6. References

[1] Leaper, C., Ayres, M. M. 2007. A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *PSPR*, 11(4), 328-363.

[2] Fuchs, R. 2017. Do women use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics,* 22(3), 345-374.

[3] Palomares, N. A. 2009. Women are sort of more tentative than men aren't they? How men and women use tentative language differently, similarly, and counterstereotypically as a function of gender salience. *Communication Research*, 36(4), 538-560.

[4] Simpson, A. P. 2009. Phonetic differences between male and female speech. *Language and Linguistics Compass* 3, 621-640.

[5] Suire, A., Tognetti, A., Durand, V., Raymond, M., Barkat-Defradas, M. 2020. Speech acoustic features: A comparison of gay men, heterosexual men, and heterosexual women. *Archives of Sexual Behaviour* 49, 2575-2583.

[6] Kwon, H. 2010. Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. *The Journal of Advanced Prosthodontics* 2, 71-76.

[7] Weirich, M., Simpson, A. P. 2018. Gender identity is indexed and perceived in speech. *PLos One* 13.

[8] Albuquerque, L., Valente, A. R. S., Teixeira, A., Oliveira, C., Figueiredo, D. 2021. Age and gender effects in European Portuguese spontaneous speech. *Loquens* 8.

[9] Bradlow, A. R., Kraus, N., Hayes, E. 2003. Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46, 80-97.

[10] Cangi, M. E., Isildar, A., Tekin, A., Sarac, A. B. 2020. A preliminary study of normative speech rate values of Turkish speaking adults. *ENT Updates* 10, 381-389.

[11] Fox., R. A., Nissen, S. L. 2005. Sex-related acoustic changes in voiceless English fricatives. *Journal of Speech, Language, and Hearing Research* 48, 753-765.

[12] Fuchs, S., Toda, M. 2010. Do differences in male versus female /s/ reflect biological factors or sociophonetic ones? In S. Fuchs, M. Toda, M. Zygis (Eds.), *An interdisciplinary guide to turbulent sounds* (pp. 281–302). Berlin: Mouton de Gruyter.

[13] Morris, R. J., McCrea, C. R., Herring, K. D. 2008. Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, 36(2), 308-317.

[14] Whiteside, S. P., Henry, L., Dobbin, R. 2004. Sex differences in voice onset time: A developmental study of phonetic context effects in British English. *The Journal of the Acoustical Society of America* 116, 1179-1183.

[15] Oh, E. 2011. Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, 39, 59-67.

[16] Ladegaard, H. J. Bleses, D. 2003. Gender differences in young children's speech: The acquisition of sociolinguistic competence. *International Journal of Applied Linguistics* 13, 222-233.
[17] Leaper, C., Smith, T. E. 2004. A meta-analytic review of gender variations in children's language use: Talkativeness, affiliative speech, and assertive speech. *Developmental Psychology*, 40(6), 993-1027.
[18] Nicollas, R., Garrel, R., Ouaknine, M., Giovanni, A., Nazarian, B., Triglia, J. M. 2008. Normal voice in children between 6 and 12 years of age: Database and nonlinear analysis. *Journal of Voice* 22, 671-675.
[19] Lee, S., Iverson, G. K. 2008. Development of stop consonants in Korean. *Korean Linguistics*, 14, 21-39.
[20] Cartei, V., Garnham, A., Oakhill, J., Banerjee, R., Roberts, L., Reby, D. 2019. Children can control the expression of masculinity and femininity through the voice. *Royal Society Open Science* 6.
[21] Fung, P., Schertz, J., Johnson, E. K. 2021. The development of gendered speech in children: Insights from adult L1 and L2 perceptions. *JASA Express Letters* 1, 1-6.
[22] Munson, B., Lackas, N., Koeppe, K. 2022. Individual differences in development of gendered speech in preschool children: Evidence from a longitudinal study. *Journal of Speech, Language, and Hearing Research* 65, 1311-1330.
[23] Fitzpatrick, M. A., Mulac, A., India, K. 1995. Gender preferential language use in spouse and stranger interaction. *Journal of Language and Social Psychology* 14, 18-39.
[24] Cockrell, M. D. 2020. Provider-patient interaction: Exploring elderspeak in simulated preclinical chiropractic student encounters. *Gerontology & Geriatric Medicine*, 6, 1-6.
[25] Hughes, S. M., Farley, S. D., Rhodes, B. C. 2010. Vocal and physiological changes in response to the physical attractiveness of conversational partners. *Journal of Nonverbal Behavior*, 34, 155-167.
[26] Rothermich, K., Harris, H. L., Sewell, K., Bobb, S. C. 2019. Listener impressions of foreigner-directed speech: A systematic review. *Speech Communication* 112, 22-29.
[27] Fritche, R., Shattuck-Hufnagel, S., Yung Song, J. 2021. Do adults produce phonetic variants of /t/ less often in speech to children? *Journal of Phonetics* 87.
[28] Kokkinaki, T., Vasdekis, V. G. S., Devouche, E. 2020. Maternal and paternal infant-directed speech to girls and boys: An exploratory study. *European Journal of Developmental Psychology* 17, 379-414.
[29] Dailey, S., Bergelson, E. 2022. Talking to talkers: Infants' talk status, but not their gender, is related to language input. *Child Development*, 1-19.
[30] Clearfield, M. W., Nelson, N. M. 2006. Sex differences in mothers' speech and play behaviour with 6-, 9-, and 14-month-old infants. *Sex Roles* 54, 127-137.
[31] Pruden, S. M., Levine, S. C. 2017. Parents' spatial language mediates a sex difference in preschoolers' spatial-language use. *Psychological Science* 28, 1583–1596.
[32] Foulkes, P., Docherty, G. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34, 409-438.
[33] Deterding, D. 2006. The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association* 36, 187-196.
[34] Kennison, S. M., Byrd-Craven, J. 2015. Gender differences in beliefs about infant-directed speech: The role of family dynamics. *Child Development Research* 2015, 1-6.
[35] Boersma, P., Weenink, D. 2001. PRAAT, a system for doing phonetics by computer. *Glot international* 5. 341-345.
[36] Lathrop-Marshall, H., Keyser, . M. B., Jhingree, S., Giduz, N., Bocklage, C., Couldwell, S., … Jacox, L. A. 2022. Orthognathic speech pathology: Impacts of class III malocclusion on speech. *European Journal of Orthodontics* 44, 340-351.
[37] Bürkner, P. 2018. Advanced bayesian multilevel modeling with the R package brms. *The R Journal* 10, 395–411.
[38] Hespanhol, L., Vallio, C. S., Costa, L. M., Sarahiotto, B. T. 2019. Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, 23(4), 290-301.

NEUTRALIZATION AND SECONDARY ACOUSTIC CUES OF VOICING CONTRAST: A TOHOKU AND TOKYO JAPANESE PRODUCTION EXPERIMENT

Chuyu Huang[1], Sanae Matsui[2], Naoya Watabe[3], Hiroto Noguchi[2, 4], Ayako Hashimoto[5], Ai Mizoguchi[6, 7], Mafuyu Kitahara[2]

[1]Nagoya Gakuin University [2]Sophia University [3]The University of Tokyo [4]Tokyo Medical and Dental University [5]Tokyo Kasei Gakuin University [6]Maebashi Institute of Technology [7]National Institute for Japanese Language and Linguistics

huang@ngu.ac.jp, sanaematsui1107@gmail.com, watabe.naoya.2821@goo.jp, noguchih425@gmail.com, hassy@kasei-gakuin.ac.jp, aimizoguchi@maebashi-it.ac.jp, mafuyu@sophia.ac.jp

ABSTRACT

This study investigates the phonological voicing contrast of plosives in two Japanese dialects: Tohoku Japanese, which is known for its intervocalic voicing of voiceless plosives, and Tokyo Japanese. Despite some previous research on this topic, the extent to which various acoustic properties, such as voice onset time (VOT) and fundamental frequency (F0) (1) covary with intervocalic voicing in Tohoku Japanese and (2) compare to Tokyo Japanese remains unclear. Thus, we conducted online production experiments to examine the voicing contrast in Tohoku Japanese. Our results indicate that the voicing contrast in Tohoku Japanese is incomplete; moreover, other acoustic cues such as F0 and preceding vowel duration, vary between Tohoku and Tokyo Japanese. These findings suggest that while secondary cues, i.e. duration and onset F0 may play an active role in realizing the intervocalic voicing contrast, cue utilization may differ between these two dialects.

Keywords: VOT, neutralization, Tohoku dialects, secondary cues, onset F0.

1. INTRODUCTION

Tohoku Japanese is a group of dialects of Japanese spoken mainly in the North-Eastern regions of Japan, in which voiceless plosives /t, k/ have been typically realized as voiced [d, ɡ] in intervocalic environments [1–4]. The phonologically voiced plosives /d, ɡ/ in some of these dialects were realized as nasalized plosives; however, in the variations where /d, ɡ/ remain oral plosives, voiceless and voiced plosives may be merged due to the intervocalic voicing.

In addition to the internal change of the Tohoku dialect, Tokyo Japanese, the dominant dialect in Japan, also influences the realization of Tohoku plosives. Both dialects may share common undergoing changes. Tokyo Japanese has undergone a significant change wherein initial voiced consonants are realized with non-negative voice onset times (VOTs) [5]. On the other hand, Tohoku Japanese also exhibits a gradual change, with positive VOTs observed in initial voiced segments. Moreover, the younger generation in Tohoku shows a stronger tendency towards positive VOTs in initial voiced segments [6].

VOT, however, does not seem to be the only acoustic cue used for distinguishing the voicing contrast. It has been noted that Tokyo Japanese speakers utilize other phonetic cues in voicing contrast, such as F0 of the following vowel, to realize or detect the word-initial voicing contrast despite the partial neutralization of the primary VOT cue [7, 8]. Meanwhile, research of Tohoku dialects that focused on VOT and other secondary cues also showed that word-medial voiceless plosives might be fully or partially neutralized with voiced plosives in terms of VOT [4, 9]. [4] investigated the VOT of word-initial and word-medial plosives and found no significant difference between voiced and voiceless plosives in the word-medial environment. In our follow-up study [9], VOT measurement was refined to include fully pre-voiced tokens and confirmed partial neutralization in VOT; however, other acoustic cues, including F0 and duration, do not seem to play an active role in voicing contrast realization. The phonological voicing and VOT polarity of both dialects are summarized in Table 1.

	Word Initial	Word Medial
Tokyo	Voiced: −/+ Voiceless: +	Voiced: − Voiceless: +
Tohoku	Voiced: −/+ Voiceless: +	Voiced: − Voiceless: −/+

Table 1: Intervocalic voicing and initial devoicing in the two dialects (−, +: VOT polarity).

The following limitations of the previous studies can be addressed: (1) Most previous research data are based on spontaneous speech, which enhances the difficulty of evaluating any possible effect resulting from the phonological environments, including following vowels, accentual types, and lexical strata, primarily when the data set includes F0. (2) A wide

variety of Tohoku dialects were targeted, which could be an issue since each dialect may have a different extent of intervocalic voicing. Let alone gender and age were also reported as important factors of intervocalic voicing [10, 11]. An experiment is necessary to assess the above factors appropriately. (3) The influence of Tokyo Japanese is challenging to evaluate under the flow of language standardization. A framework within which the realization of both Tokyo and Tohoku dialects can be adequately evaluated is needed.

To investigate the role of VOT and other acoustic cues in the realization of plosives in the Tokyo and Tohoku dialects of Japanese, we conducted a production experiment using a psycholinguistic factorial design that controlled for accentual types and lexical strata. Given the diversity within Tohoku dialects, only speakers from the Yamagata prefecture were included in the experiment.

2. METHODS

2.1. Experiment

To examine the realization of VOT and other acoustic cues in the plosive voicing contrast in two groups of Japanese speakers, we conducted an entirely remote production experiment using *jsPsych* [12], through *Cognition.run* (https://cognition.run). All the participants were recruited online via a web recruiting service (*Crowdworks*; https://crowdworks.jp) and completed the experiment remotely using their own computers with microphones and either speakers or headphones. To control for the potential influence of other dialects, especially the standard Japanese dialect, on Tohoku speakers, tutorials were conducted in the dialects specific to each group. The tutorials were recorded by a native male speaker from Yamagata and a native female speaker from Tokyo.

2.1.1. Participants

The participants included 10 Yamagata-dialect speakers who were born and raised in Yamagata Prefecture, and 20 Tokyo-dialect speakers, born and raised in the Tokyo-Metropolitan area including Tokyo, Saitama, Chiba, and Kanagawa. All participants completed a consent form before beginning the experiment. The data of one speaker from Yamagata and two speakers from the Tokyo-Metropolitan area with incomplete experiments were excluded from the analysis. As a result, the data of 9 Yamagata-dialect speakers (female: 4, male: 5; average age: 38.44 years, *SD* of age: 9.89) and 18 Tokyo-dialect speakers (female: 7, male: 11; average age: 37.22 years; *SD* of age: 10.60) were analyzed.

2.1.2. Experimental design

This experiment contains two within- and one between-subject factor, all of each with two levels as follows: *voicing* (voiced/voiceless), *word position* (initial/medial), and *group* (Tohoku/Tokyo). Four conditions were created as shown in Table 2.

	Voiced	Voiceless
Initial	***g**oma* "sesami"	***k**izu* "scar"
Medial	*fu**d**e* "pen"	*ha**t**o* "pigeon"

Table 2: Examples in each condition. Italic bold characters indicate the target consonants. Bilabial stops have been excluded due to the distinctive phonological contrast present in Japanese ([h, b] rather than [p, b]).

Each within-subject condition contained 16 words with identical numbers of target segments and vowels. A total of 64 items were used in the experiment. Words of foreign stratum and those with clear morphological boundaries were not used due to the possible effects on intervocalic voicing [13]. Items were presented in the form of pictures. Each item consisted of a picture of an older adult man attempting to identify an object, followed by the target word and a carrier sentence in either the Tokyo or Yamagata dialect, as shown in Table 3. Both mean "It should be __. Yeah, it is __." Participants were asked to "tell the senior man what it is" by using the provided carrier sentence after seeing the target word written in Japanese characters (*kanji* and *hiragana*), and a recording session followed. Before submission, each participant checked their recording files. Each sentence contained two repetitions of the target word, and the order of items was fully randomized.

Tokyo group	*... dayo. Sōsō, ... da.*
Tohoku group	*... dabe. Ndanda, ...da.*

Table 3: Carrier sentences used in the experiment.

2.2. Data analysis

Experimental trials were retrieved from *.json* produced on *Cognition.run*. All recorded files were converted from *base64* to a *.wav* format, and we applied auto-segmentation to the recorded files with *Julius* [14]. Five trained phoneticians inspected and, if necessary, corrected the segmentation boundaries. Acoustic properties were measured by *Praat* [15].

VOT values were measured as a duration between the burst and the voicing onset of the following vowel. Fully pre-voiced plosives were annotated following the method proposed in [16] and applied in [9]. F0 values for the vowel following the target plosive were measured. The first F0 detected by *Praat* was used as an onset F0 for the statistical analysis. Time-normalized F0 values measured at each 1/10 point of the entire vowel were obtained for the contour analysis. We corrected apparent F0 measurement

errors by adjusting the pitch settings of *Praat*. F0 values were converted from hertz to semitones relative to each speaker's mean F0. We also measured the duration of preceding vowels and excluded all segment labels overlapping silent pauses.

Three statistical models were applied to evaluate the effect of each acoustic cue, including VOT, F0, and duration. Not only the VOT but the duration was scaled by each participant to avoid any possible effects of speech rate. The statistical analysis was carried out with the Linear Mixed-Effects Model (LME), using *lme4* [17] and *lmerTest* [18] on *RStudio* (*R*: 4.1.2 [19]; *RStudio*: 2022.07.2 [20]). Pairwise comparisons were conducted using *emmeans* [21]. *Word position* was excluded in the statistical model for vowel duration since there was no preceding vowel in word-initial conditions. Factor coding is shown in Table 4.

	Voicing	*Word Pos.*	*Group*
0	Voiced	Initial	Tohoku
1	Voiceless	Medial	Tokyo

Table 4: Factor coding in the LME models.

Other predictors used as fixed factors in the statistical model were *age* (coded as numeric), *sex* (coded as factor), and *accent* (whether the accent nucleus is on the syllable of the target segment or not). Random factors include the intercepts of *participant, item, sentence position* (the first or second reading), and *segment duration* related to speech rate, which may affect the realization of intervocalic voicing [22, 23]. In the model evaluating the effect of F0 and the duration of the following vowel, *vowel type* was added as a random factor. The optimal models were determined using backward selection.

3. RESULTS

3.1. Tohoku and Tokyo speakers' VOT differences

Table 5 and Figure 1 show the results of each condition. A significant three-way interaction of *voicing*, *word position*, and *group* indicates that the interaction between voicing and position varies in the two speaker groups.

	Estimates	*SE*	*t value*	*Pr(>|t|)*
(Intercept)	−.193	.305	−.633	.532
voicing (voi)	1.210	.084	14.425	<.001
wordpos	−1.020	.086	−11.796	<.001
group	−.302	.080	−3.754	<.001
age	.003	.007	.396	.696
sex	.150	.132	1.137	.267
accent	−.004	.049	−.084	.933
voi:wordpos	−.487	.120	−4.055	<.001
voi:group	−.023	.065	−.352	.724
wordpos:group	.374	.068	5.543	<.001
voi:wordpos:group	.461	.094	4.910	<.001

Table 5: The LME results where VOTs are normalized.

In a sub-analysis, Tohoku speakers' VOT was significantly shorter in the word-medial voiceless condition than in the initial condition (β: 1.507, *SE*: .085, *z ratio*: 17.697, $p < .001$). However, another pairwise comparison suggested that the VOT was still significantly longer than the voiced medial condition (β: −.722, *SE*: .086, *z ratio*: −8.405, $p < .001$), indicating incomplete intervocalic neutralization. No significant difference was detected in age, sex, and accent in the VOT duration of each condition. There was, however, a significant difference between the Tohoku and Tokyo groups in the initial devoicing of voiced consonants (β: .325, *SE*: .080, *z ratio*: 4.030, $p < .005$). The extent of non-negative VOT values observed in the Tohoku group was even greater than in the Tokyo group.

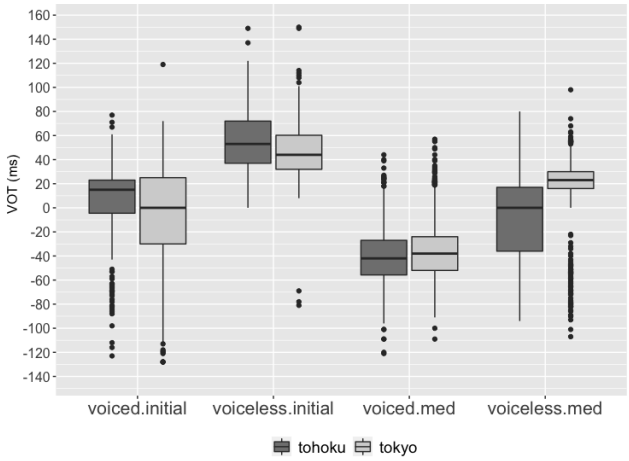


Figure 1: VOT (ms) of each experimental condition.

3.2. Results of F0 and VOT

Table 6 shows the result of F0. The three-way interaction of *voicing*, *position*, and *group* was also observed. The interaction of F0 with two within-subject factors differed in the speaker group.

	Estimates	*SE*	*t value*	*Pr(>|t|)*
(Intercept)	−2.215	1.078	−2.054	.156
voicing (voi)	2.818	.541	5.208	<.001
wordpos	2.331	.543	4.294	<.001
group	.005	.221	.024	.980
age	.002	.011	.143	.887
sex	.102	.210	.875	.390
accent	.695	.282	2.460	.015
voi:wordpos	−2.119	.765	−2.769	.008
voi:group	−.348	.245	−1.421	.155
wordpos:group	.123	.229	.543	.587
voi:wordpos:group	1.311	.332	3.951	<.001

Table 6: The LME results of the onset F0.

Further pairwise comparisons indicated that the initial voiced condition in both groups triggered a lower F0 than the voiceless initial condition ($p < .001$).

As previous studies suggested, voicing contrast in both groups involved a significant F0 change.

In the Tohoku group, voiceless medial consonants provoked a significantly lower onset F0 than voiceless initial consonants (p = .002) but did not significantly differ from voiced medial consonants (p = 1.000) in the onset region, followed by a steady descending by time as shown in Figure 2.

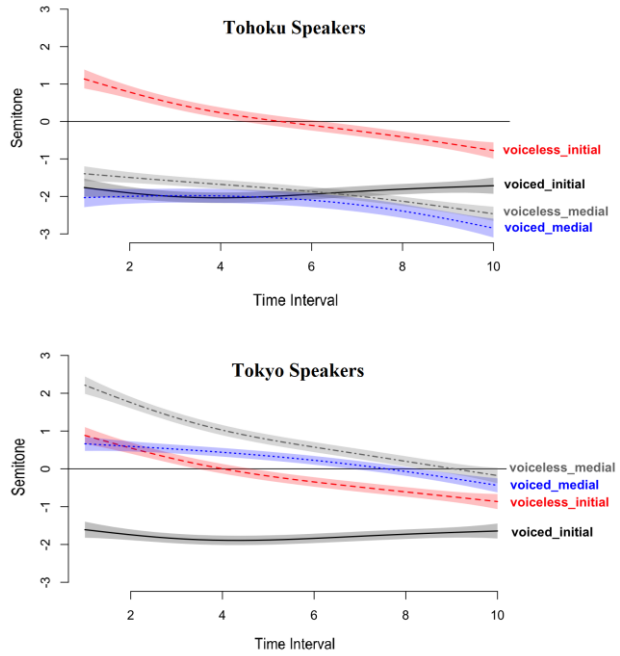


Figure 2: Plots of VOT and F0 in semitone of both speaker groups in which smooths are estimated from a fitted GAMM model using *mgcv* [24].

3.3. Durational cue of the preceding vowel

The duration result revealed a significant interaction between *voicing* and *group* (β: −.302, *SE*: .087, *t value*: −3.478, $p < .001$). The effects of *age*, *sex*, and *accent* were all of no significance. In a pairwise comparison, the duration of the preceding vowel before the voiceless plosives was slightly shorter than voiced conditions in both speaker groups (Tokyo: $p < .001$; Tohoku: $p < .001$).

4. DISCUSSION

The results confirmed the intervocalic voicing neutralization in the Tohoku dialect, which has been reported by previous studies like [4]; however, the neutralization is incomplete, which supports prior analyses [9]. The devoicing of word-initial voiced plosives was observed in both the Tokyo and Tohoku dialects. Positive VOTs in the initial voiced segments were more pronounced in Tohoku Japanese than in Tokyo Japanese, which is consistent with the results of a previous study [6] where the Yamagata dialect was categorized as a "coastal dialect of the Sea of Japan" with a higher frequency of positive VOTs among Tohoku dialects. However, no generational differences were observed in our study.

Incomplete neutralization may leave some other utilizable phonetic cues in speech [25, 26]. Following [9], this study measured the onset F0 of the following vowel and the duration of the preceding short vowels. Despite broad overlapping, the onset F0 showed a difference between the initial voiced and voiceless condition in the Tokyo group, which successfully duplicated the result in previous studies, and between the medial voiceless and voiced condition. Tokyo speakers may realize a higher onset F0 after voiceless plosives, however, in the Tohoku group, no significance between voiced and voiceless medial conditions was observed.

Durational cues also seemed to be realized, another important finding of this study. Similar to earlier studies on English [27, 28], vowels preceding voiceless segments were shorter than those preceding voiced segments in Japanese. Note that the voicing here is phonological following the experimental design. Although intervocalic voicing and initial devoicing changed the VOTs of the segments, both groups show a clear tendency for duration change.

Future research should address incomplete neutralization issues by conducting perception experiments to determine whether the identified secondary cues are also involved in the perception.

5. CONCLUSION

In this study, we utilized an online production experiment to examine intervocalic voicing and initial devoicing in Tohoku and Tokyo Japanese. The results empirically demonstrated that intervocalic voicing neutralization is incomplete in Tohoku Japanese, as evidenced by the VOT distribution. Additionally, non-negative VOTs were observed in the initial voiced conditions in both dialects. The preceding vowel duration was also found to vary with voicing in both dialects, whereas the onset F0 was found to be distinguished only in Tokyo Japanese. These findings are consistent with previous studies, which have demonstrated that vowels preceding voiceless segments in both Japanese dialects are significantly shorter than those preceding voiced segments, despite the phonological contrast of vowel length.

6. ACKNOWLEDGMENT

This research was supported by JSPS Kakenhi 22K00516.

7. REFERENCES

[1] H. Inoue, "Consonant system of the Tohoku-dialect," *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, vol. 1968, no. 52, pp. 80–98, 1968.

[2] J. Ohashi, *Tohokuhogen onsei-no-kenkyu*. Ohfu, 2002.

[3] A. Hashimoto, "Tohoku-hogen-ni-okeru kagyo tagyo-no-yuseion-ni-tsuite: Tohokuhogen onseichosa-kara (2)," *Tsuda Journal of Language and Culture*, no. 34, pp. 88–102, 2019.

[4] A. Mizoguchi, A. Hashimoto, S. Matsui, S. Imatomi, R. Kobayashi, and M. Kitahara, "Neutralization of Voicing Distinction of Stops in Tohoku Dialects of Japanese: Field Work and Acoustic Measurements," *Interspeech 2020*, pp. 1873–1877, 2020.

[5] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[6] M. Takada, *Nihongo-no gotō-heisaon-no kenkyū: VOT-no kyōjiteki-bunpu-to tsūjiteki-henka [A study of Japanese word-initial glottal stops: synchronic distribution of VOT and diachronic changes]*. Kuroshio Publishers, 2011.

[7] J. Gao and T. Arai, "Plosive (de-) voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery," *Journal of Phonetics*, vol. 77, p. 100932, 2019.

[8] H.-G. Byun, "Acoustic characteristics for Japanese stops in word-initial position: VOT and post-stop f0," *Journal of the Phonetic Society of Japan*, vol. 25, pp. 41–63, 2021.

[9] H. Noguchi *et al.*, "VOT and F0 perturbations for the realization of voicing contrast in Tohoku Japanese," in *Interspeech 2022*, ISCA, Sep. 2022, pp. 3428–3432. doi: 10.21437/Interspeech.2022-587.

[10] M. Takada, "Regional and generational variation of VOT in Japanese word-initial stops," *Papers from the First International Conference on Asian Geolinguistics*, pp. 273–282, 2012.

[11] G. Hiratsuka, "About the Voicing Phenomenon of ka, ta line Consonants Appearing in the Middle and End of Words in the Northern Part of Akita Prefecture," *Shigen : Tokyo University of Foreign Studies Descriptive Linguistic papers*, no. 13, pp. 151–158, 2017.

[12] J. R. De Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behavior research methods*, vol. 47, no. 1, pp. 1–12, 2015.

[13] N. Kibe, "Kagoshima-oyobi-Tohokuhogen-no gochu-kagyo-tagyo-no shiin-ni tsuite," 1990.

[14] T. Kawahara, "Open-source speech recognition software Julius," *JSAI*, vol. 20, no. 1, pp. 41–49, 2005.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer programme], version 6.1. 41." 2021.

[16] A. S. Abramson and D. H. Whalen, "Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of phonetics*, vol. 63, pp. 75–86, 2017.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," *Journal of statistical software*, vol. 82, pp. 1–26, 2017.

[19] R Core Team, "R: a language and environment for statistical computing. Version 4.1.2," *R Foundation for Statistical Computing, Vienna, Austria. Freely available at https://www. r-project. org*, 2021.

[20] RStudio Team, "RStudio: integrated development for R," *Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com*, 2020.

[21] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Package 'emmeans.'" 2019.

[22] J. Magloire and K. P. Green, "A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish," *Phonetica*, vol. 56, no. 3–4, pp. 158–185, 1999.

[23] R. Puggaard-Rode, C. S. Horslund, and H. Jørgensen, "The rarity of intervocalic voicing of stops in Danish spontaneous speech," *Laboratory Phonology*, vol. 13, no. 1, 2022.

[24] S. N. Wood, "Generalized Additive Models: An Introduction with R.,(Chapman and Hall: CRC Press, Boca Raton, FL.)," 2006.

[25] N. Warner, A. Jongman, J. Sereno, and R. Kemps, "Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch," *Journal of phonetics*, vol. 32, no. 2, pp. 251–276, 2004.

[26] C. Farrington, "Incomplete neutralization in African American English: The case of final consonant voicing," *Language Variation and Change*, vol. 30, no. 3, pp. 361–383, 2018.

[27] A. S. House, "On vowel duration in English," *The Journal of the Acoustical Society of America*, vol. 33, no. 9, pp. 1174–1178, 1961.

[28] L. J. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *The Journal of the Acoustical Society of America*, vol. 51, no. 4B, pp. 1296–1303, 1972.

Park or bark? VoT Production of English Bilabial Stops in Arabic Learners of English

XXX

XXX
XXX

ABSTRACT

This preliminary study compares VoT patterns of word-initial English /b/ and /p/ in three groups: Omani undergraduate students, Omani professors and English native speakers (ENS). It also compares VoT patterns of these sounds in two phonetic environments: CV and CCV. Words containing initial /b, p/ were elicited using a reading task. The results revealed a significant difference in VoT patterns of /p/ between students and ENSs, but not between professors and ENSs. Students' /p/ was characterized with shorter lag compared to professors' or ENSs'. Variability in the production of /b/ existed among Omani speakers. While ENSs' VoT productions were characterized with short lag, students' and professors' productions were characterized sometimes with voicing lead and others with short/long lag. VoT values of /b/ and /p/ were not significantly affected by whether the sound was in a CV or CCV syllable. The findings are discussed in relation to speech learning research.

Keywords: VoT, voicing lead, voicing lag, speech learning

1. INTRODUCTION

There has been a growing number of studies examining the acquisition of English stops by Arabic native speakers [1, 2, 3, 4, 5]. The main acoustic dimension used to distinguish Arabic-accented stops from English stops in these studies is voice onset time (VoT). Lisker and Abramson [6, p.422] defined VOT as "the time interval between the burst that marks release and the onset of periodicity that reflects laryngeal vibration". VoT is found to be highly reliable in separating homorganic stop categories despite phonetic differences among languages [6].

The VoT patterns of English and Arabic word-initial stops vary considerably. English /p, t, k/ are normally aspirated while /b, d, g/ are unaspirated. Aspirated English stops are produced in the long-lag region and have positive VoT values. Glottal pulsing, which distinguishes voiced and voiceless sounds, during closure is not necessarily available in the production of English stops. Hence, English word-initial /b, d, g/ are produced in the short-lag region. On the other hand, Arabic /b, d, g/ are normally accompanied by glottal pulsing during closure (i.e. voicing lead) and thus are at the opposite end of the continuum compared to English voiceless aspirated stops [5]. The VoT range of Arabic voiceless stops falls within the range of English unaspirated stops /b, d, g/ [5].

Since Arabic stops, /b, d, t, k/, overlap with those in English, English speakers face no difficulty in accurately identifying them when Arabic learners of English attempt to speak English [1]. However, it is widely acknowledged that Arabs have a persistent problem in the production of acceptable English /p/ due to the absence of this sound from the Arabic phonemic inventory [1, 3, 4]. Arabs tend to either substitute English /p/ with Arabic /b/ or produce it with short lag [3, 4].

The Contrastive Analysis Hypothesis (CAH) assumes that L2 phonemes that have no equivalent counterpart in the L1 would be difficult to learn while those that have an L1 equivalent would be relatively easy to learn [7]. However, Flege [8] argues that the CAH ignored the fact that two 'similar' sounds in two languages may differ vastly at the phonetic level. The Speech Learning Model (SLM) and more recently SLM-r predicts that sounds that are 'similar' in two languages may be harder to learn than sounds that are 'new' [8]. It also proposes that L2 learners gradually discriminate between L1 and L2 phonetic details as they gain more experience using the L2 in everyday life [9]. Based on this, one can predict that Arabs, with sufficient input, will eventually produce a satisfactory English /p/. Word-initial English /b/, on the other hand, might be assimilated to Arabic /b/ and produced with glottal pulsing during the closure period.

Although there has been a rise in the number of instrumental studies interested in examining the acquisition of the phonological/phonetic features of English stops by native Arabic speakers, this area of research remains understudied, especially research that examines Omani Arabic speakers of English. Additionally, most of these studies have looked at

whether native Arabic speakers have acquired the distinction between voiced and voiceless stops in a CV syllable. This study is interested in extending this research to include an examination of the VoT patterns of these stops in a consonant cluster. We predict that Omani speakers of English might find it harder to produce native-like VoT when the stop is in CCV syllable than when it is in a CV syllable. Thus, our study aims to: 1) compare VoT production of English /b/ and /p/ in three groups of speakers: Omani undergraduate students, Omani professors and English native speakers (ENS), 2) examine the effect of the phonetic environment on VoT production.

2. METHOD

2.1. Participants

Three groups of participants were recruited for this study. The Omani Native Arabic speakers fall into two groups. The first group involved nine male and female students who had studied English in Omani public schools from grade one through grade 12, and were specializing in English language at the university level (BA) at the time of the study. All were 20-22 years old and had finished at least 30 credits of English language and content courses. The second group of Omani native speakers involved 6 instructors with PhD in Linguistics or English Language that they obtained from the USA, Canada or the UK. This means that they spent at least 7 years residing in a native English country. All were 37-45 years old and taught English specialized courses at the university BA level. Four native English speakers (British, Australian and American) who were between 33 and 47 years old participated as a control group.

2.2. Material

31 English monosyllabic words containing word-initial /b/ and /p/ were used in this study. The words were randomized in a typed list and embedded in the carrier sentence "I like to say…". 16 of the words contained /p/, six of which contained /p/ preceding a vowel as in *pack, pig, path, peek, page, pad*, while 10 of which contained /p/ in a consonant cluster, preceding a sonorant consonant /r/ or /l/ as in *plank, plate, play, place, plane, prop, prat, prig, pram, proud*. 15 of the words contained /b/, 5 of which contained /b/ before a vowel as in *big, bad, beat, bath, back*, 10 words contained /b/ before a sonorant consonant as in *blue, black, blood, bleed, block, broad, brag, brig, broom, bread* and *break.*

2.3. Procedure

The participants were instructed to read the sentences at a comfortable tempo. They were recorded on Praat, which was digitized at 44.1 kHz, using Microsoft headphones and a microphone in a quiet room at the English Department at Sultan Qaboos University. One British English speaker recorded herself on Praat using a microphone in her home in the UK.

2.4. Acoustic and statistical analyses

526 words were submitted to wideband spectrograms in Praat. For each test word, two temporal intervals were measured by hand to the nearest 5 ms from the spectrogram. VoT of /p/ was measured from the onset of the release burst to the point where the first periodic cycle starts. VoT of pre-voiced productions were measured from the offset of the preceding vowel (indicated by a drop in energy in the region of F1 and F2) to the onset of release burst. A binary judgement for the presence or absence of glottal pulsing during the closure interval of speakers was made [1]. A stop was considered to have been produced with glottal pulsing (pre-voicing) if it showed visible periodic striation throughout most of the closure [1, 10].

Statistical analyses were carried out in R [11]. Linear Mixed Effect Models (LMER) were used to generate statistical evidence for the differences in VoT values among the speakers using lme4 package [12]. For fixed effects, we used group (ENS, student, professor), the following sound and an interaction between group and following sound. For random effects, we used speaker and word as random intercepts.

3. RESULTS

3.1 VoT patterns of /p/

In some students and professors, the production of /p/ was accompanied with glottal pulsing. There were two instances of glottal pulsing in the production of /p/ in professors and 7 instances of glottal pulsing in students' productions. This indicates that English /p/ was assimilated to Arabic /b/ in these productions. These tokens were removed from the rest of the analyses.

LMER revealed a significant difference in mean VoT between students and ENSs ($p<0.01$). ENSs' mean VoT was 54.29 ms ±10.87 (standard error) higher than that of students'. The difference in mean VoT of /p/ between professors and ENSs was not significant ($p>0.05$). ENSs' mean VoT was only 17.02 ms ±11.59 (standard error) higher than that of professors. This indicates that professors' mean VoT of /p/ falls within the range of ENSs'. The mean VoT

of professors was also significantly different from students' ($p<0.01$). Professors' mean VoT was 37.26 ms ±9.87(standard error) higher than that of students. As to the effect of the phonetic environment on VoT patterns of /p/, LMER revealed no significant difference between the mean VoT of /p/ before a vowel and the mean VoT of /p/ before a consonant. Visual inspection of Figure 1 shows that in all three groups of speakers, the mean VoT of /p/ is higher when followed by a sonorant consonant than when followed by a vowel, though this was not significant.

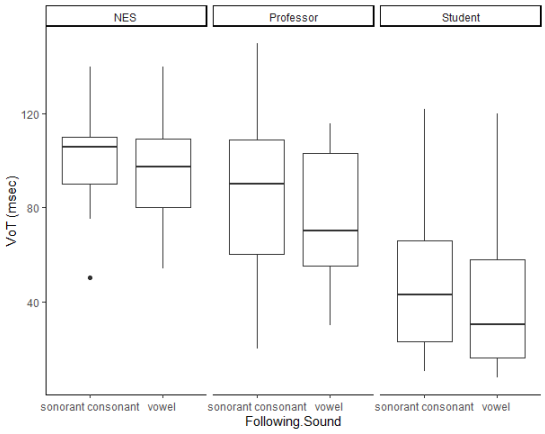


Figure 1. VoT values of /p/ in three groups of speakers distributed based on following sound

It is worth noting that variability in the VoT values of /p/ existed among Omani native speakers. Table 1 shows that the range of VoT values in students' productions is between 7 ms and 122 ms and that in professors is between 20 ms and 150 ms. The minimum VoT value in ENSs is 50 ms. This indicates that some +VoT productions in Omani speakers were within the short lag range, which characterizes English /b/.

	Min	Mean	Max
ENS	50	99	140
Student	7	43	122
Professor	20	81	150

Table 1. Minimum, maximum and mean VoT values of /p/ in three groups of speakers

3.2. VoT patterns of /b/

When measuring VoT productions of /b/, LMER revealed a significant difference between professors and ENS' ($p<0.01$). Professors' mean VoT was 70.65 ms ±22(standard error) lower than that of ENSs'. On the other hand, the mean VoT in students was not statistically different from that of ENSs' ($p>0.05$). These findings might be misleading if we ignored the variability in VoT productions in native Omani speakers. LMER also revealed no significant effect of the following sound on VoT productions of /b/.

The VoT patterns of /b/ varied within and across the three groups (Table 1). While ENSs' productions were mostly characterized with short lag and no voicing lead, students' productions were characterised with voicing lead (20%), short and long lag (80%). Professors' productions were characterised mostly with voicing lead (70%) and to a lesser extent short lag (29%).

/b/	N (%)	Group	Min	Mean	Max
-VoT	50 (20)	Student	-150	-98	-50
	71 (70)	Professor	-130	-97	-60
	59 (100)	ENS	5	14	40
+VoT	116 (80)	Student	8	33	127
	26 (29)	Professor	10	19	40

Table 2. Mean, minimum and maximum VoT values of /b/ in three groups of speakers

When measuring the difference in mean +VoT between the three groups, LMER revealed that ENSs' mean VoT was 21.06 ±7.68 (standard error) lower than that of students ($p<0.01$). On the other hand, ENSs mean VoT was not significantly different from mean VoT of professors ($p>0.05$). This indicates that students' +VoT of /b/ was exaggerated compared to ENSs'. From Table 1, we can notice that the maximum VoT value of /b/ in students is 127 ms, a value that falls within the VoT range of /p/. With regard to the effect of the following sound on +VoT values of /b/, LMER revealed no significant effect of the following sound on mean +VoT within the three groups.

4. DISCUSSION

This study shed light on the VoT patterns of English /b/ and /p/ in the productions of three groups of speakers: Omani undergraduate students, Omani professors and ENSs. It also examined any effects of the phonetic environment in which /b/ and p/ appear.

The study found that Omani professors' VoT productions of /p/ were not different from ENSs'. As ENSs, professors produced their /p/ with long lag (i.e. aspiration) most of the time. This finding is not in line with previous studies [1, 2, 3, 4] that found Arabs'

VoT productions of /p/ fall mostly within the short lag region (below 40 ms). Similar to previous research, however, students' VoT productions were significantly different from those of ENSs and professors'. Mean VoT value of students in all phonetic environments (m= 43 ms) was not near that of ENSs' (m= 99 ms) or professors' (m= 81 ms).

One justification for the native-like production of /p/ in professors is the amount of authentic exposure they have had in the target language. Professors' success in the production of a native-like /p/ compared to students may be in part due to their experience in using English [13]. These speakers spent 7 or more years studying English and linguistics in a native speaking country. Undoubtedly, they might have formed a new phonetic category for /p/ in their phonetic inventory [8, 14]. Hence, this finding does not provide support to the CAH which predicts that L2 learners will face difficulty acquiring sounds that are different from those in their phonemic inventory. Another justification is the type of task the speakers were asked to perform to elicit data for this study. L2 learners are found to perform better in reading tasks than in spontaneous speech because they have more time to access their explicit knowledge [15]. It would be interesting to examine native Omani speakers' productions of /p/ using more naturalistic data-elicitation tasks.

Despite the fact that professors' productions of /p/ were native-like in terms of VoT, it is worth noting that there was also variability in the range of VoT values in professors' productions. Some VoT productions fell within the short lag region and others fell within the long lead region. This points to the importance of considering individual differences and even differences within speakers when making conclusions about L2 acquisition.

The study also found that there was large variability in the VoT productions of /b/ especially among students. Professors' and students' productions were sometimes accompanied with voicing lead and other times short lag but also long lag in students. ENSs' productions of /b/, on the other hand, were accompanied with short lag. This is in line with research that confirmed the absence of prevoicing in word-initial /b/ of ENSs [1, 10]. Overall, the results of the current study showed that the mean VoT of professors' /b/ was significantly lower than that of ENSs' or students'. This is obviously because professors' /b/ was mostly accompanied with voicing lead. Interestingly, this supports the SLM which predicts that L2 learners will find it harder to acquire a sound that is phonetically similar to another sound in their phonetic inventory [8, 9].

Does the previous finding suggest that students were more successful at producing a native-like /b/? It is hard to answer this question with yes just by considering the previous finding. Further analyses of +VoT productions of /b/ revealed that students' mean VoT (m= 33 ms) was significantly higher than that of ENSs' (m=14 ms). Some students' +VoT productions fell within the VoT range of /p/. This might be one reason why students' VoT productions of /b/ were statistically similar to ENSs'. In a study that investigated Saudi learners' acquisition of English stops, Alanazi [4] found that his Arab speakers also produced English /b/ with glottal pulsing, short lag and long lag. Apparently, producing English /b/ with long lead is an influence of the speakers' L1, Arabic [16, 17]. On the other hand, producing /b/ with long lag might be due to the pressure of the reading task the learners had to perform. The learners, being at a developmental stage, probably got confused as to whether they should pronounce /p/ or /b/ as the reading list contained randomized words with either sound word-initially.

The phonetic environment of /p/ and /b/ (CV or CCV) did not have a significant effect on VoT productions in any of the three groups of speakers. The VoT patterns of /p/ preceding a vowel were lower than those of /p/ in a consonant cluster, though this was not significant. This finding is in line with a study by Klatt [18] who found that the VoT of word-initial English voiceless stops was longer before sonorants and high vowels than before mid and low vowels.

Taken all together, the findings of this study confirm that experienced L2 learners are capable of producing native-like sounds that are novel especially when the data-elicitation task stimulates the learners' explicit knowledge. It also suggests that L2 learners may not be able to acquire some phonetic details of L2 sounds that are perceived as similar to L1 sounds. The study calls for the importance of considering variability in L2 productions. The L2 speakers in this study showed large variability in the range of VoT productions of their English /p/ and /b/.

5. REFERENCES

[1] Flege, J. E., & Port, R. 1981. Cross-language phonetic interference: Arabic to English. *Language and speech*, *24*(2), 125-146.

[2] Port, R. F., & Mitleb, F. M. (1983). Segmental features and implementation in acquisition of English by Arabic speakers. *Journal of Phonetics*, *11*(3), 219-229.

[3] Sawsan, A. 2017. An acoustic analysis of the production of word-initial stop/p/by late Arab bilinguals. *Training, Language and Culture*, *1*(4), 68-83.

[4] Alanazi, S. 2018. *The acquisition of English*

stops by Saudi L2 learners (Doctoral dissertation, University of Essex).
[5] Khattab, G. 2002. VOT production in English and Arabic bilingual and monolingual children. *Amsterdam Studies in the Theory and History of Linguistic Sciences Series* 4, 1-38.
[6] Lisker, L and Abramson A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. WORD 20(3), 384-422.
[67] Lado, R. 1964. Linguistics across cultures, University of Michigan press. *Ann Arober, Michingan*.
[8] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277.
[9] Flege, J. E., & Bohn, O. S. 2021. The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3-83.
[10] Abramson, A. S., & Whalen, D. H. 2017. Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of phonetics*, *63*, 75-86.
[11] Team, R. C. 2013. R: A language and environment for statistical computing.
[12] Bates D., Mächler M., Bolker B. & Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.
[13] Piske, T., MacKay, I. R., & Flege, J. E. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of phonetics*, *29*(2), 191-215.
[14] Flege, J. E., Schirru, C., & MacKay, I. R. 2003. Interaction between the native and second language phonetic subsystems. *Speech communication*, *40*(4), 467-491.
[15] Saito, K., & Brajot, F. X. 2013. Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English/ɹ/by late JapaENSe bilinguals. *Bilingualism: Language and Cognition*, *16*(4), 847-863.
[16] Ringbom, H. 1992. On L1 transfer in L2 comprehension and L2 production. *Language learning*, *42*(1), 85-112.
[17] Smith, M. S., & Kellerman, E. 1986. Crosslinguistic influence in second language acquisition: An introduction. *Crosslinguistic Influence in Second Language Acquisition, New York: Pergamon*.
[18] Klatt, D. H. 1975. Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of speech and hearing research*, *18*(4), 686-706.

THE PHONETIC REALIZATION OF INTERVOCALIC OBSTRUENTS IN SUZHOU WU CHINESE

Chunyu Ge, Peggy Mok

The Chinese University of Hong Kong
chunyuge@cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

Suzhou Wu Chinese has a three-way voicing contrast in obstruents and the realization of voiced obstruents is position-dependent. This paper investigated the phonetic realization of intervocalic voiceless unaspirated and voiced obstruents in Suzhou Wu. Acoustic data of disyllabic words from three age groups were collected. Closure duration, proportion of voicing, and partial voicing shape of intervocalic obstruents were analyzed. The results showed that although there were slight differences across age groups in the realization of intervocalic obstruents, all age groups produced only a small portion of voiced obstruents as fully voiced. Moreover, for both voiced and voiceless obstruents, most tokens were realized with the bleed pattern (voicing extending from the previous vowel), which was also consistent across age groups. The results contribute to our understanding of the voicing contrast in obstruents cross-linguistically and further study is needed to explore more factors influencing the realization of voicing.

Keywords: obstruents, phonetic realization, voicing, Wu Chinese

1. INTRODUCTION

The voicing contrast of obstruents can be effectively captured by the time between voice onset and the release burst of the obstruents (VOT), which can distinguish voiced (negative VOT), voiceless unaspirated (short positive VOT) and voiceless aspirated (long positive VOT) obstruents [1, 2]. However, some languages pose challenges to this unidimensional approach. For example, the voiced unaspirated stops and the voiced aspirated stops in Hindi both have negative VOT [3]. Another example is Korean, in which VOT is not sufficient to distinguish lax stops from aspirated stops, at least for younger speakers [4]. Therefore, more fine-grained description of voicing contrast is needed.

Furthermore, to describe the implementation of voicing in more detail is motivated by the fact that the realization of the voicing contrast in obstruents is influenced by various factors. [5] is an early exploration into this direction, which investigated the voicing profiles of stop closures in five languages using the probability of voicing during stop closure. They found that the voicing probability can be influenced by neighbouring sounds, as well as positional and prosodic factors. [6, 7] are more fine-grained in that the author not only analysed the voicing proportion during closure in American English, but also analysed the partial voicing shape of obstruents in intervocalic positions. She classified the partial voicing shape into bleed (voicing extending from previous sounds), prevoicing, hump (voicing strongest in the middle of the closure) and trough (voicing weaker in the middle of the closure than both ends) patterns and found that bleed is the most common implementation pattern of voiced obstruents in American English. Recently, [8] demonstrated that the position-dependent lenition of voiceless stops in Mixetc was modulated by prosodic conditions.

The language investigated in this study, Wu Chinese, has a three-way voicing contrast in obstruents (including stops, fricatives, and affricates) and the implementation of voiced obstruents is position-dependent. Wu Chinese distinguishes voiceless unaspirated, voiceless aspirated, and voiced obstruents [9]. Previous studies have shown that voiced obstruents in Wu Chinese are phonetically voiceless but with breathy voice in the following vowel in isolation and at word-initial positions, but are truly voiced (i.e., with voicing during consonant closure) at word-medial/intervocalic positions. [10] demonstrated with spectrogram that voiced obstruents in Suzhou Wu were not different in VOT from voiceless unaspirated obstruents at word-initial positions, but at intervocalic positions, voiced obstruents were phonetically voiced during closure, while voiceless unaspirated obstruents were voiceless. Consistent results were obtained in Shanghai Wu by [11] using fiberscope, showing that the vocal folds were vibrating during closure for voiced obstruents at intervocalic positions but not at word-initial

positions. Therefore, the phonetic realization of the voicing of voiced obstruents in Wu Chinese is dependent on word position. A limitation of the previous studies is that they were usually based on visual inspections without measurements and they have not investigated the shape of the voicing, which may be more informative. Another background of this study is that our earlier studies [12, 13] showed that voiced obstruents in Suzhou Wu are changing in that younger speakers produce less breathy voice of voiced obstruents in isolation. As our earlier studies concentrated on voiced obstruents in isolation, this study aims to further explore the realization of obstruents at intervocalic positions across age groups and to investigate whether younger speakers also produce voiced obstruents differently from older and middle-aged speakers.

To sum up, this paper investigates fine-grained patterns of the realization of intervocalic obstruents in Suzhou Wu, and also to explore the difference of the realization across age groups. This paper hopes to contribute valuable empirical data to the cross-linguistic patterns of voicing contrast in obstruents.

2. METHOD

To compare the realization of intervocalic obstruents across age groups, native Suzhou Wu speakers of three age groups were recruited to participate in a word reading production experiment. Table 1 shows the number of speakers in each age group and their age range.

Table 1: Number of speakers in three age groups and their age range

Age group	Female	Male	Age range
older	5	5	65-84
middle-aged	10	6	43-54
younger	5	5	21-29

The test words were disyllabic words and the target syllable was the second one. The first syllables were all in CV structure, without any nasal or consonantal coda. The target syllables were in CV or CVN structure, where C included bilabial or alveolar stops /p b t d/ and alveo-palatal affricates /tɕ dʑ/. Due to phonotactic constraints, affricate onsets can only co-occur with rhymes starting with the high front vowel /i/, and the rhyme /in/ was selected for both the voiceless and voiced affricates. Likewise, the rhyme /in/ was also used for bilabial stops. But for alveolar stops, the rhyme /e/ was used due to the absence of suitable minimal pairs with the rhyme /in/. The participants were asked to read the words in isolation. Each word was repeated three times. The sounds were recorded using a Behringer ECM8000 microphone through the Scarlett Solo Studio 2nd Gen audio interface to a laptop, with a sampling frequency of 44,100 Hz.

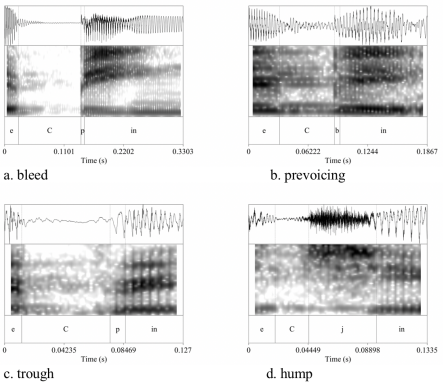


Figure 1: Waveforms and spectrograms of partial voicing shape patterns

The sounds were segmented manually in Praat [14]. The target element is the closure of the obstruent onset of the second syllable. The start of closure was determined as the end of the previous vowel's F2 and the end was the release burst of the stop or affricate. After segmentation, a Praat script was run to get the duration of the closure and the proportion of voicing during the closure using the *fraction of locally unvoiced frames* value available in *Voice Report*. The proportion of voicing was calculated by subtracting the fraction of locally unvoiced frames from one. The pitch ceiling was set at 500 Hz for female speakers and 300 for males, and the floor was at 100 Hz for females and 75 for males. Also, the closure was cut into three equal length intervals, and the proportion of voicing was measured during each interval. The partial voicing shape was determined based on the proportions of voicing of these three intervals (see [6] for more details of the definition). Before determining the partial voicing shape, all tokens which are fully voiceless were excluded. If the proportion of voicing decreased from the first interval to the second and the last, it is a bleed pattern. It is a prevoicing pattern, if the proportion increased from the first to the last interval. If the second interval had the greatest proportion of voicing, it is a hump pattern. Contrarily, it is a trough if the second interval had the smallest proportion of voicing. Figure 1 shows examples of each pattern.

3. RESULTS

3.1. Closure duration

Table 2 shows the mean closure duration of intervocalic obstruents across age groups. The closure duration was analyzed using linear mixed-effects models, with phonological Voicing, Manner of articulation and Age group as fixed effects and also as random slope by speaker. Stops generally had longer closure than affricates (Estimate = -36.01, $p < 0.01$). Voiceless obstruents had longer closure than voiced obstruents (Estimate = 16.39, $p < 0.01$). However, the inclusion of Age group into the model did not significantly improve the model fit ($\chi^2(2) = 0.49$, $p = 0.78$). Therefore, there is no statistically significant difference among older, middle-aged and younger speakers in closure duration of intervocalic obstruents.

Table 2: Mean closure duration (in ms) of intervocalic obstruents across age groups

laryngeal	manner	older	middle	younger
voiced	stop	111	123	118
	affricate	81	88	93
voiceless	stop	140	140	131
	affricate	97	95	104

3.2. Voicing proportion over the total closure duration

The tokens were divided into fully voiced, partially voiced and fully voiceless, according to the proportion of voicing during the entire closure duration. Those with more than 90% voicing were fully voiced, while those with less than 10% were fully voiceless, and partially voiced with voicing ratio in-between, following [6]. Figure 2 shows the percentage of the three realizations of stops and affricates across age groups. Although for both voiced and voiceless obstruents, most tokens were partially voiced, voiced obstruents had more fully voiced tokens than those of voiceless obstruents, and likewise, less fully voiceless tokens than those of voiceless obstruents. The difference between voiced and voiceless obstruents was also slightly smaller for younger speakers than older speakers.

To substantiate the observations, generalized linear mixed-effects models were fitted on the type of overall voicing, which was coded as a binary variable with fully voiced as 1 and partial voicing and fully voiceless as 0. Phonological Voicing and Manner were included as fixed effects and also as random slope by speaker. Closure was also included as a fixed effect. Voiceless obstruents were less likely to be fully voiced than voiced obstruents (Estimate = -0.22, $p < 0.01$), and the longer the closure duration was, the less likely it was to be fully voiced (Estimate = -0.06, $p < 0.01$). But again, the inclusion of Age group did not significantly improve the model fit ($\chi^2(2) = 1.53$, $p = 0.46$). Also, no significant difference was found between stops and affricates ($\chi^2(1) = 1.45$, $p = 0.23$). Therefore, intervocalic voiced obstruents were more likely to be fully voiced than voiceless obstruents, and there was no significant difference among older, middle-aged and younger speakers.

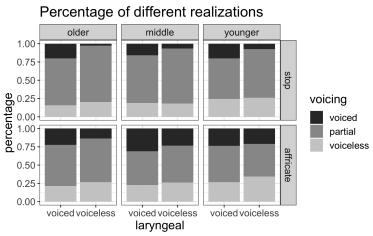


Figure 2: Percentage of fully voiced, partially voiced, and fully voiceless stops and affricates across age groups

3.3. Partial voicing shape

Apart from the overall proportion of voicing, the proportion of voicing was also measured during the three equal length intervals of the closure, in order to investigate the fine pattern of the implementation of voicing. Figure 3 shows the mean proportion of voicing during each interval for voiceless and voiced obstruents across age groups. The data were also analyzed using linear mixed-effects models with phonological Voicing, Manner, Interval, Age group, and their interactions as fixed effects, whereas Manner and Voicing were also included as random slopes by speaker. The effect of these factors was determined by model comparison. The most remarkable effect was interval. The first interval had higher proportion of voicing than both the second and third intervals. Once again, Age group did not improve the model fit significantly ($\chi^2(2) = 1.07$, $p = 0.59$). What is interesting here is that there was a significant interaction between Interval and Voicing ($\chi^2(2) = 10.41$, $p < 0.01$). For all three intervals, voiced obstruents had significantly higher voicing proportion than voiceless obstruents.

The partial voicing shape was also determined using the criteria mentioned in the Method and the proportion of each partial voicing shape for voiced and voiceless obstruents across age groups was shown in Figure 4. It is quite clear that the bleed pattern took up the largest proportion for

both stops and affricates, regardless of their voicing. This observation was consistent among all three age groups. The bleed pattern indicates that the voicing in the closure is extending from the previous vowel. Among the other three patterns, the prevoicing took up the largest proportion, while the hump and trough patterns were extremely rare.

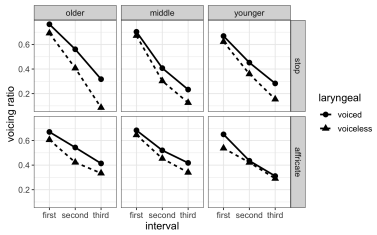


Figure 3: Proportion of voicing by interval of intervocalic stops and affricates across age groups

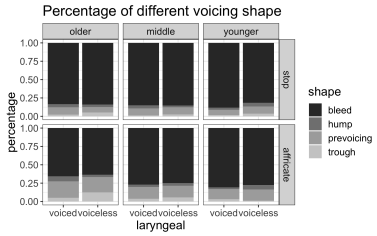


Figure 4: Percentage of different voice shape of intervocalic stops and affricates across age groups

4. DISCUSSION

This paper investigated the implementation of intervocalic voiceless unaspirated and voiced obstruents in Suzhou Wu across age groups. The most striking finding is that only a small portion of intervocalic voiced obstruents were implemented as fully voiced, which is different from the observation of previous studies that intervocalic obstruents in Wu Chinese were usually truly voiced. It is also found that the most frequent partial voicing pattern was the bleed pattern, in which the voicing was extending from the previous vowel. Moreover, younger speakers did not have different realizations from older and middle-aged speakers, neither in closure duration nor in voicing proportion. Furthermore, listeners may be sensitive to the greater proportion of voicing in voiced obstruents, which can be further investigated using a perception experiment.

Although only a small portion of voiced obstruents was realized as fully voiced, it is still premature to draw the conclusion that the voiced obstruents in Suzhou Wu has changed into voiceless even in the intervocalic position. First, the probability of being fully voiced was larger for voiced than for voiceless obstruents and the proportion of voicing was greater for voiced than for voiceless obstruents for all three intervals during closure. Therefore, speakers still make more efforts to maintain voicing in voiced obstruents than in voiceless obstruents. This can be seen in Figure 3, at the third interval, voiced obstruents had larger proportion of voicing than voiceless obstruents and the proportion of voicing of voiceless stops nearly dropped to zero for older and middle-aged speakers. Another finding is that the bleed pattern is the most common realization pattern of both voiceless and voiced obstruents in Suzhou Wu, which is consistent with the findings in American English [6, 7]. This could be evidence that the prevalence of bleed in obstruents is a cross-linguistic trend, which has phonetic groundings. For instance, [15] showed that voicing will continue into closure until the oral pressure builds up to inhibit voicing.

One of the conditions this study failed to control is the prosodic structure of the disyllabic word. As mentioned in the Introduction, prosodic conditions can influence the realization of stops [8]. In Suzhou Wu, as in other Wu dialects, prosodic structure can influence the pattern of tone sandhi [16]. Left-dominant tone sandhi applies in prosodic words and right-dominant tone sandhi applies in prosodic phrases. Left-dominant tone sandhi implies a closer relationship between two syllables than right-dominant tone sandhi. However, the speakers in this study read most of the words with right-dominant tone sandhi. It is possible that the realization of intervocalic obstruents can be influenced by tone sandhi, and they are more likely to be fully voiced only in those disyllabic words which undergo left-dominant tone sandhi. This would be an interesting topic for further study.

5. CONCLUSION

The phonetic realization of the voicing of intervocalic obstruents in Suzhou Wu Chinese was investigated in this paper. Although only a small portion of voiced obstruents was realized as fully voiced, their probability of being fully voiced and their proportion of voicing were greater than voiceless obstruents. Younger speakers had similar realizations of obstruents with older and middle-aged speakers, both in terms of closure duration and voicing. Prosodic factors should be taken into consideration for further investigation of the implementation of voicing in Suzhou Wu obstruents.

6. ACKNOWLEDGEMENTS

This study was supported by the RGC General Research Fund 2019/20 Project Reference 14607619 awarded to the second author and the CUHK Research Fellowship Scheme 2019-20, 2020-21 awarded to the first author.

7. REFERENCES

[1] L. Lisker and A. S. Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," *Word*, vol. 20, no. 3, pp. 384–422, Jan. 1964.

[2] T. Cho, D. H. Whalen, and G. Docherty, "Voice onset time and beyond: Exploring laryngeal contrast in 19 languages," *Journal of Phonetics*, vol. 72, pp. 52–65, Jan. 2019.

[3] J. Schertz and S. Khan, "Acoustic cues in production and perception of the four-way stop laryngeal contrast in Hindi and Urdu," *Journal of Phonetics*, vol. 81, p. 100979, Jul. 2020.

[4] D. J. Silva, "Acoustic evidence for the emergence of tonal contrast in contemporary Korean," *Phonology*, vol. 23, no. 2, pp. 287–308, Aug. 2006.

[5] C. Shih, B. Möbius, and B. Narasimhan, "Contextual effects on consonant voicing profiles: A cross-linguistic study," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 2, San Francisco, 1999, pp. 989–992.

[6] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *Journal of Phonetics*, vol. 54, pp. 35–50, Jan. 2016.

[7] ——, "Phonation and laryngeal specification in American English voiceless obstruents," *Journal of the International Phonetic Association*, vol. 48, no. 3, pp. 331–356, Dec. 2018.

[8] C. DiCanio, W.-R. Chen, J. Benn, J. D. Amith, and R. Castillo García, "Extreme stop allophony in Mixtec spontaneous speech: Data, word prosody, and modelling," *Journal of Phonetics*, vol. 92, p. 101147, May 2022.

[9] J. Cao and I. Maddieson, "An Exploration of Phonation Types in Wu Dialects of Chinese," *Journal of Phonetics*, vol. 20, pp. 77–92, Jan. 1992.

[10] F. Shi, "Acoustic characteristics of voiced obstruents in Suzhou Chinese [in Chinese]," *Studies in Language and Linguistics (Yuyan Yanjiu)*, no. 01, pp. 49–83, 1983.

[11] N. Ren, "Phonation types and stop consonant distinctions-Shanghai Chinese," Ph.D. dissertation, The University of Connecticut, 1992.

[12] C. Ge, W. Xu, W. Gu, and P. Mok, "An electroglottographic study of phonation types in tones of Suzhou Wu Chinese," in *Speech Prosody 2022*. ISCA, May 2022, pp. 570–574.

[13] C. Ge, W. Xu, W. Gu, and P. P. K. Mok, "The change in breathy voice after tone split: A production study of Suzhou Wu Chinese," *Journal of Phonetics*, vol. 98, p. 101239, May 2023.

[14] P. Boersma, "Praat: Doing phonetics with computers," 2020.

[15] J. J. Ohala, "The Origin of Sound Patterns in Vocal Tract Constraints," in *The Production of Speech*, P. F. MacNeilage, Ed. New York, NY: Springer, 1983, pp. 189–216.

[16] X. Li, "Tone sandhi in the Suzhou dialect: Synchronic and diachronic perspectives," *Journal of Chinese Linguistics Monograph Series*, no. 24, pp. 21–32, 2010.

A PRODUCTION STUDY OF KOREAN CONSONANTS

Jeremy Perkins[a], Dahm Lee[b], Seunghun J. Lee[c]

University of Aizu[a], Seoul National University[b], International Christian University[c], IIT Guwahati[c]
jperkins@u-aizu.ac.jp, dahm1021@snu.ac.kr, seunghun@icu.ac.jp

ABSTRACT

A production study of the Korean three-way contrast of tense, lenis and aspirated obstruents was conducted that included all manners. In this paper, psychoacoustic roughness as a measure of laryngeal constriction was used to explore the phonetic acoustic properties of tense obstruents. Because it is a perceptual attribute, it may have stronger correlation with laryngeal constriction than spectral tilt. In addition, the behavior of the Korean laryngeal contrast among the fricatives and affricates was explored.

Tense obstruents correlated with increased roughness and reduced spectral tilt at the following vowel onset. In addition, tense fricatives had longer frication duration than aspirated fricatives; tense stops and affricates had longer closure duration than other obstruents; lenis obstruents had shorter closure duration. Closure and frication duration differences in stops and fricatives, respectively, were larger than in affricates, suggesting they play a more important role in producing the contrast in stops and fricatives.

Keywords: Korean obstruents, laryngeal setting, psychoacoustic roughness

1. INTRODUCTION

Korean has a three-way contrast (lenis vs. tense vs. aspirated) among stops and affricates. The inventory of Korean obstruents is shown in Table 1. Each cell in the table shows the possible laryngeal settings for each combination of place and manner.

Type	Labial	Coronal	Dorsal
Stops	p p’ p^h	t t’ t^h	k k’ k^h
Affricates		tʃ tʃ’ $tʃ^h$	
Fricatives		s’ s	

Table 1: Korean obstruents by place & manner

This contrast shown in Table 1 has received much attention with most of that attention focused on the stop series [1, 2]. Previously, Seoul Korean speakers produced these obstruents with different voice-onset times (VOTs) [3, 4]. However, current speakers produce lenis stops with a large positive VOT, just like the aspirated series. Word-initially, a contrast in f0 of the following vowel has been observed, with lenis obstruents preceding low f0 and aspirated obstruents preceding high f0 [5, 6, 7, 8]. Despite this shift in cues from VOT to f0, the tense series still has a shorter VOT than lenis and aspirated [5]. Some research has found that VOT for lenis stops may have an intermediate value between tense and aspirated [3], and may perceptually not be distinguishable from the other two [9]. Kim et al. [10] showed that cues on the following vowel alone were necessary and sufficient in perception of the lenis category.

While this previous research has mostly focused on stops, there is less research on the status of affricates and fricatives. Some past research suggests affricates pattern with the stops [11]. Unlike stops and affricates, there is only a two-way contrast among the fricatives. In [12], frication duration was found to be longer in tense fricatives, but no f0 difference was found.

Laryngeal setting is also involved in the three-way contrast, and has been measured via spectral tilt. At higher frequencies, intensity drops relatively less in creaky voice than in modal voice [13], resulting in reduced values for spectral tilt measures like H1*−H2* with tense obstruents. On the other hand, lenis stops are often noted to involve breathiness, which are associated with significantly reduced intensity at higher frequencies [14, 12, 6].

The goals of this research are (1) to use psychoacoustic roughness to identify coarticulatory phonation types associated with the laryngeal contrast, in particular with (a) the tense series and (b) among fricatives and affricates; and (2) to assess which cues could be involved in the laryngeal contrast, based on the acoustic findings.

2. METHODS

2.1. Consultants

A production experiment was run with 24 Seoul Korean speakers between the ages of 20 and 27 (14 females). Most speakers spoke some English and all speakers except one reported no problems affecting their speech at the time of the recording.

2.2. Stimuli

A list of 66 CVCV bisyllables was created (most nonce), with all fourteen obstruents represented in the first consonant (C1), which was the focus of the main research question. All vowel positions were filled with the low vowel [a]. With labial and dorsal C1 (p, p', p^h, k, k', k^h), only the three lenis stops (p, t, k) were used in the second consonant (C2), yielding 18 words. With coronal C1 (t, t', t^h), the three aspirated (p^h, t^h, k^h) and the three lenis stops (p, t, k) were used in C2, yielding 48 more words, for a total of 66 words. More coronals were used because the analysis of fricatives is only done among coronals. A separate analysis of the tense series is done with all places though. As such, more coronals were needed to achieve the same relative statistical power among the places of articulation.

2.3. Procedure

Each word was embedded into a Korean carrier sentence, 단어 X 는 무슨 뜻인가요?, meaning "What does the word X mean?" The 66 sentences were arranged in a random order and presented twice via slides displayed on a screen. Five practice items were included. Participants wore a head mounted Shure WH-30 microphone connected to a Tascam MK-2 with an XLR cable. A native Korean-speaking experimenter advanced the slides, noting any pronunciation errors. The resulting audio files were saved as mono wav files at a sampling rate of 44.1 kHz.

All files were automatically segmented and labeled using a Korean forced alignment tool [15]. Further annotation of the audio files was done using Praat [16]. Additional segments were labeled for consonant closure, frication, and aspiration. All boundaries were automatically moved to the nearest zero-crossing.

2.4. Data extraction

Laryngeal constriction (here, spectral tilt and psychoacoustic roughness) at the onset of a following vowel, VOT, closure duration, frication duration, and f0 in the vowel following C1 were measured. VoiceSauce [17] was used to measure spectral tilt, f0 and duration. Formant-normalized spectral tilt was measured via H1*−H2*, H1*−A1*, and H1*−A2*. F0 was normalized within each speaker relative to that speaker's median f0, to account for higher- and lower-pitched voices. F0 and spectral tilt measurements were made during the entire vowel, at 10 ms intervals. F0 was converted to cents (hundredths of a semitone). Duration measures included closure duration for stops and affricates, frication and aspiration duration for affricates and fricatives, and VOT for stops. Release duration (frication plus aspiration) was also included for affricates and fricatives.

Psychoacoustic roughness is defined as a sensory attribute related to rapid changes on the amplitude envelope of a sound (15–300 Hz). It was found to correlate positively with creaky tones in Burmese [18]. As such, it can provide a psychoacoustic measure of creakiness, like loudness or pitch for intensity and f0, respectively. Since it relates to perception, it may relate more closely to phonological differences, so if laryngeal constriction is used as a cue to distinguish tense from lax, roughness may correlate more strongly with this difference than spectral tilt. Roughness was measured in Aspers in the vowel following the first consonant using the Matlab routine from Villegas et al. [18].

2.5. Analysis

F0, spectral tilt and roughness were plotted via smooth scatterplots using geom_smooth with either the gam or loess method automatically chosen in R's ggplot2 library [19, 20].

Linear mixed models were constructed for all duration measures with random intercepts for speaker. Fixed effects for obstruent type, place of articulation and manner were included with mutual interactions, when relevant. Type III ANOVAs with Satterthwaite's method were used to confirm fixed effect significance, with any insignificant effects removed in the final models. The model intercepts are lenis alveolar stops, unless otherwise noted.

3. RESULTS

The combined results are plotted in the following figures by laryngeal setting (tense is light grey or with long dashed lines, aspirated is dark grey or with short dashed lines, and lenis is solid black). Results for VOT are not illustrated as they conformed with

previous findings [5] (effect sizes in linear mixed model: lenis (intercept): 73.0 ms; tense: −64.0 ms; aspirated: −7.5 ms, all differences significant with $p < 0.001$).

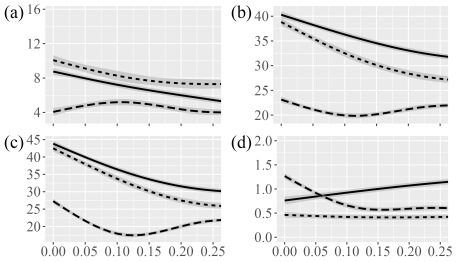


Figure 1: Stops. Spectral tilt in dB, (a) H1*−H2*, (b) H1*−A1*, (c) H1*−A2* and (d) psychoacoustic roughness in Aspers by laryngeal setting in the initial 25% of the vowel (normalized time on the x-axes)

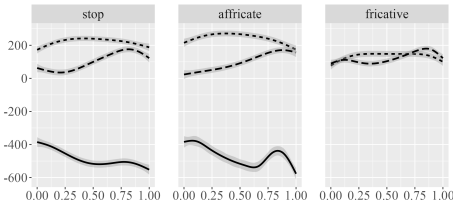


Figure 2: Coronals. f0 in cents relative to speaker median by laryngeal setting and manner (normalized time on the x-axes)

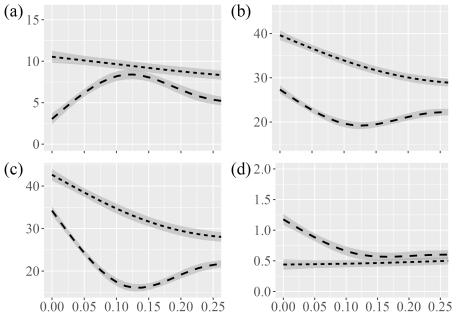


Figure 3: Fricatives. Spectral tilt in dB, (a) H1*−H2*, (b) H1*−A1*, and (c) H1*−A2*, and (d) psychoacoustic roughness in Aspers by laryngeal setting in the initial 25% of the vowel (normalized time on the x-axes)

3.1. Laryngeal constriction and f0 in tense obstruents

Laryngeal setting was measured among the stop series via spectral tilt and psychoacoustic roughness, as shown in Fig. 1. In tense stops, reduced spectral tilt and raised roughness was seen at vowel onset, indicating the presence of coarticulatory creaky phonation. This effect extended throughout the initial 25% of the vowel for H1*−A1*, and H1*−A2*, but only for about the first 5% to 10% for H1*−H2* and roughness. On the other hand, aspirated and lenis stops induced raised spectral tilt, indicative of breathiness. Additionally, roughness increased over the initial 25% of the vowel following a lenis stop. This is likely due to the lowered f0 of the lenis stop (see Fig. 2); roughness has been found to correlate inversely with f0 during modal phonation in Thai for example [21]. Despite this, roughness is actually reduced at vowel onset where coarticulatory effects of the lenis obstruent are expected. Like spectral tilt, roughness results also indicate the presence of laryngeal constriction following tense stops.

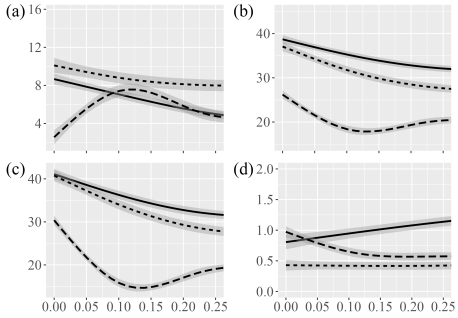


Figure 4: Affricates. Spectral tilt in dB, (a) H1*−H2*, (b) H1*−A1*, and (c) H1*−A2*, and (d) psychoacoustic roughness in Aspers by laryngeal setting in the initial 25% of the vowel (normalized time on the x-axes)

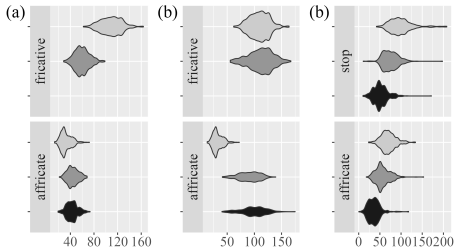


Figure 5: (a) Frication duration, (b) release duration, and (c) closure duration, all in ms, by manner and laryngeal setting

Fricatives and affricates showed the same pattern as the stops: The tense series had reduced spectral tilt difference and increased roughness at the following vowel onset (see Fig. 3 for fricatives and Fig. 4 for affricates). This effect extended throughout the initial 25% of the vowel for H1*−A1*, and H1*−A2*, but only the initial 10% for H1*−H2* and roughness, like the

stops. Regarding f0, there was almost no difference between tense and aspirated fricatives (see Fig. 2). However, the affricates showed the same pattern as the stops, where f0 was lowest following lenis affricates and higher following aspirated and tense affricates, with the tense series having a slightly lower f0 than the aspirated series.

3.2. Duration measures

Fig. 5a shows frication duration by manner and laryngeal setting. The linear mixed model had significant effects for laryngeal setting ($F = 4170$, $p < 0.001$), manner (fricative vs. affricate) ($F = 49,108$, $p < 0.001$) and their interaction ($F = 21,814$, $p < 0.001$). The intercept condition, here [tʃʰ], had a frication duration of 44.0 ms ($t = 29.4$, $p < 0.001$). The tense series had the shortest frication duration (-11.8 ms, $t = -39.9$, $p < 0.001$) and lenis affricates were slightly shorter than aspirated affricates (-1.3 ms, $t = -4.2$, $p < 0.001$). Fricatives had longer frication duration than affricates ($+15.3$ ms, $t = 47.8$, $p < 0.001$). A highly significant interaction effect was seen, with tense fricatives having very long frication durations ($+61.2$ ms, $t = 147.7$, $p < 0.001$).

Fig. 5b shows the results for release duration (frication plus aspiration). Tense obstruents do not include periods of aspiration, unlike the aspirated and lenis series. As a result, the tense series had significantly shorter release durations, which may play an important role in distinguishing the tense series. Aspirated fricatives, meanwhile had very similar release durations to tense fricatives and so release duration is likely less important than frication duration among fricatives.

The results for closure duration are shown in Fig. 5c. A linear mixed model was run only on coronals to test for differences between laryngeal setting and between affricates and stops (manner). The intercept condition, [tʰ], had a closure duration of 74.3 ms ($t = 30.0$, $p < 0.001$). Affricates were found to have shorter closure duration than stops (-17.9 ms, $t = -35.7$, $p < 0.001$). The lenis series had the shortest closure duration (-23.2 ms, $t = -48.1$, $p < 0.001$) and the tense series had the longest ($+23.2$ ms, $t = 51.5$, $p < 0.001$). A significant interaction between laryngeal setting and manner indicated that the relative differences in closure duration for each laryngeal setting were shorter among the affricates ($F = 256.0$, $p < 0.001$). Lenis affricates had relatively longer closure duration ($+2.6$ ms, $t = 3.8$, $p < 0.001$), and tense affricates had relatively shorter closure duration (-10.3 ms, $t = 3.8$, $p < 0.001$).

4. DISCUSSION

Word-initial tense obstruents saw increased roughness and decreased spectral tilt, indicative of laryngeal constriction across all manners and places of articulation. In addition, duration differences were found; tense affricates had longer closure duration, shorter frication duration, and significantly shorter release duration. Meanwhile, tense fricatives had longer frication duration (confirming [12]) and tense stops had longer closure duration.

Regarding the lenis/aspirated distinction, this was traditionally made via VOT, with shorter VOT for lenis obstruents. We explored other cues, including those related to laryngeal setting. Small differences were seen in spectral tilt and roughness, but individual speaker variation was considerable and so these cues are unlikely to be useful. However, regarding closure duration, coronal lenis stops and affricates had consistently shorter closure duration for all speakers (notably, dorsals and labials were less consistent across speakers). Here, future work would be needed to determine if these closure duration differences are perceptually relevant in Korean.

5. CONCLUSION

A production study of the three-way contrast in Korean between tense, lenis and aspirated obstruents confirmed previous research showing an emerging tonal contrast with lenis obstruents preceding low tone and aspirated and tense obstruents preceding high tone. This tonal contrast was seen here for fricatives and affricates as well. In addition to spectral tilt, psychoacoustic roughness also confirmed previous findings that tense obstruents occur with laryngeal constriction, which was seen in tense fricatives and affricates in addition to stops. Among the fricatives, phonation cues in the following vowel and differences in frication duration were found.

VOT is unavailable as a cue among affricates and fricatives, but our results suggested frication duration and release duration (frication plus aspiration) may fill its role among the fricatives and affricates, respectively. Among the stops and affricates, closure duration was shorter for lenis and longer for tense; however, the difference between aspirated and lenis was small enough that its status as a perceptual cue is unclear. Future research on closure duration as a perceptual cue is needed to address this question.

6. ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant-in-Aid for Early Career Scientists Number JP19K13162 and also by the ILCAA joint research project "Phonetic typology from cross-linguistic perspectives (PhonTyp)" (2021-2023).

7. REFERENCES

[1] M. R. Kim and S. Duanmu, ""tense" and "lax" stops in korean," *Journal of East Asian Linguistics*, vol. 13, no. 1, pp. 59–104, 2004.

[2] H. Lee, J. J. Holliday, and E. J. Kong, "Diachronic change and synchronic variation in the korean stop laryngeal contrast," *Language and Linguistics Compass*, vol. 14, no. 7, 2020.

[3] L. Lisker and A. S. Abramson, "Cross-language study of voicing in initial stops: acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[4] C. W. Kim, "On the autonomy of the tensity feature in stop classification," *Word*, vol. 21, pp. 339–359, 1965.

[5] D. J. Silva, "Acoustic evidence for the emergence of tonal contrast in contemporary korean," *Phonology*, vol. 23, pp. 287–308, 2006.

[6] K.-H. Kang and S. G. Guion, "Clear speech production of korean stops: Changing phonetic targets and enhancement strategies," *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3909–3917, December 2008.

[7] J. Kirby, *Origins of sound change: Approaches to phonologization*. Oxford University Press, 2013, ch. The role of probabilistic enhancement in phonologization, pp. 228–246.

[8] H.-Y. Bang, M. Sonderegger, Y. Kang, M. Clayards, and T.-J. Yoon, "The emergence, progress, and impact of sound change in progress in seoul korean: Implications for mechanisms of tonogenesis," *Journal of Phonetics*, vol. 66, pp. 120–144, 2018.

[9] H. Lee and A. Jongman, "Effects of tone on the three-way laryngeal distinction in korean: An acoustic and aerodynamic comparison of the seoul and south kyungsang dialects," *Journal of the Acoustical Society of America*, vol. 42, no. 2, pp. 145–169, August 2012.

[10] M. R. Kim, P. S. Beddor, and J. Horrocks, "The contribution of consonantal and vocalic information to the perception of korean initial stops," *Journal of Phonetics*, vol. 30, pp. 77–100, 2002.

[11] M. Kim, "Correlation between vot and f0 in the perception of korean stops and affricates," in *Eighth International Conference on Spoken Language Processing*, 2004.

[12] T. Cho, S.-A. Jun, and P. Ladefoged, "Acoustic and aerodynamic correlates of korean stops and fricatives," *Journal of Phonetics*, vol. 30, pp. 193–228, 2002.

[13] E. Abberton, "Some laryngographic data for korean stops," *Journal of International Phonetic Association*, vol. 2, pp. 67–78, 1972.

[14] H. Ahn, "Post-release phonatory processes in english and korean: Acoustic correlates and implications for korean phonology," Ph.D. dissertation, The University of Texas at Austin, 1999.

[15] T.-J. Yoon, "Korean forced alignment system," Oct. 2022. [Online]. Available: https://tutorial.tyoon.net

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version xxx)," 2022.

[17] Y.-L. Shue, "The voice source in speech production: Data, analysis and models." Ph.D. dissertation, UCLA, 2010.

[18] J. Villegas, K. Markov, J. Perkins, and S. J. Lee, "Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 355–366, 2020.

[19] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Spring-Verlag, 2016.

[20] R Core Team, *R: A language and environment for statistical computing*. URL: https://www.R-project.org/: Vienna, Austria: R Foundation for Statistical Computing, 2019.

[21] J. Perkins, "Acoustic measurement of laryngeal constriction in thai consonants," in *35th General Meeting of the Phonetics Society of Japan*, September 25-26 2021.

USING MACHINE LEARNING TO MODEL THE THREE-WAY LARYNGEAL CONTRAST IN KOREAN

Jeremy Perkins[a], Yu Yan[b], Dahm Lee[c], Seunghun J. Lee[d]

University of Aizu[a], Ritsumeikan University[b], , Seoul National University[c], International Christian University[d], IIT Guwahati[d]

jperkins@u-aizu.ac.jp, yuyan@fc.ritsumei.ac.jp, dahm1021@snu.ac.kr, seunghun@icu.ac.jp

ABSTRACT

Machine learning via random forest was used to model learning of the laryngeal contrast of Korean obstruents. The model was trained on eight acoustic landmarks: f0, VOT, spectral tilt, psychoacoustic roughness and duration measures of closure and frication. The release duration (frication and aspiration) and aspiration duration of affricates and fricatives were also included. This method identifies which measures form necessary and sufficient conditions for successful machine learning, offering an additional way to identify potential contrastive cues using acoustic data. The results confirmed previous findings where the learning of the lenis-aspirated contrast depends on f0, and that of the tense stops depends on VOT. Release duration was the primary acoustic cue for learning tense affricates, and fricatives were learned using a combination of two measures: frication duration and either spectral tilt or release duration. Our results identify a minimal set of acoustic cues for learning the obstruent contrast in Korean.

Keywords: Korean, laryngeal contrast, machine learning, random forest

1. INTRODUCTION

The three-way laryngeal contrast among obstruents in Seoul Korean has received much past attention as it is a rare exception to the nearly universal two-way voicing contrast seen in most languages [1, 2]. In early research, these obstruents differed in VOT, with tense obstruents having short positive VOT, aspirated obstruents having long positive VOT and lenis obstruents having intermediate VOT [3]. However, among younger speakers currently, VOT differences between aspirated and lenis obstruents have decreased and are no longer clearly distinct. At the same time, f0 differences in the following vowel have supplanted VOT as the primary difference between aspirated and lenis obstruents [4, 5, 6, 7].

This change has provided linguists with an interesting test case then, since a cue shift in progress can be studied. Researchers have asked whether and to what extent f0 and VOT are used as cues in identifying each of the three laryngeal types. Studies that compared younger and older speakers found that younger speakers were more sensitive to f0 than VOT in the lenis-aspirated contrast [8]. Younger Seoul Korean listeners tend to use lowered f0, but not VOT, as a primary cue to recognize lenis obstruents, but tense and aspirated obstruents are recognized based on some combination of f0, VOT and spectral tilt [9, 10, 11].

Regarding the tense series in particular, Korean tense obstruents occur with reduced values of H1*−H2* and H1*−A1* at the onset of a following vowel [12]. However, other spectral tilt measures are also sometimes used (e.g. H1*−A2*, H1*−A3* and H2*−H4* [13, 14, 15]). In addition, recent work has shown that psychoacoustic roughness correlates with creaky voice in Burmese [16]. It's unclear which of these measures correlates most strongly with laryngeal constriction in a given language. Therefore, one of this paper's goals was to assess whether roughness can identify laryngeal constriction associated with the tense series, and also whether there were differences among roughness and the various spectral tilt measures in the relative importance of each measure. This result would be useful to inform future researchers on which of these various measures is most important in the production of laryngeal constriction of tense obstruents.

This paper's goal is to use supervised machine learning for new evidence on which parts of the acoustic signal may be used to learn the three-way laryngeal distinction among obstruents in Seoul Korean. The Random Forest (RF) for language modeling [17], which is a natural extension of the decision tree language models [18], is adopted. The main reason that the RF method was chosen is that RFs yield high accuracy compared with other

machine learning methods such as Support Vector Machine for speech recognition [19]. Although machine learning algorithms differ from human learners (in sensitivity and bias), they provide a way to quantify which and to what degree particular acoustic characteristics differ within a given sound contrast, allowing a glimpse of which features may play a contrastive role in linguistic contrasts.

2. METHOD

In this paper, acoustic data is collected, with a subset of that data used to train a network using the random forest technique. The trained network is then applied to another disjoint subset of the acoustic data to test whether it can accurately identify the three laryngeal obstruent types based on the acoustic data. This method identifies which parts of the acoustic signal are necessary and which are sufficient in learning each binary sound contrast; it can also compare candidate acoustic measures and assess which are more useful in learning a contrast.

2.1. Data collection

Twenty-four Seoul Korean participants (14 females), age 20 to 27 were recorded reading 66 words, all nested in a carrier sentence, 단어 X 는 무슨 뜻인가요?, meaning "What does the word X mean?". The words were bisyllabic CVCV words (most nonce) and the first consonant, which is the locus of the main research question, allowed all possible variations of laryngeal setting, manner and place among the Korean obstruents. The second consonant was one of the lenis or aspirated stops and only the vowel [a] was used. The 66 sentences were presented twice in random order via slides displayed on a screen. Participants wore a head mounted Shure WH-30 microphone connected to a Tascam MK-2 with an XLR cable. The audio files were recorded in mono at a sampling rate of 44.1 kHz. The files were segmented for consonant closure, frication and aspiration with all boundaries moved to the nearest zero-crossing using Praat [20].

F0, formant-normalized spectral tilt and segment duration were extracted from the vocalic portion following the first consonant using VoiceSauce at 10 ms intervals. Psychoacoustic roughness was also measured in the same way using a Matlab routine [16]. For f0, spectral tilt and roughness, only the initial measurement from each word was retained in the analysis in order to get a single measure, maximally adjacent to the consonant. F0 was normalized within each speaker relative to that speaker's median f0, and then converted to cents. Duration measures included closure duration for stops and affricates, frication and aspiration duration for affricates and fricatives, and VOT for stops. Release duration (aspiration plus frication) was also included for affricates and fricatives.

2.2. The Random Forest classification model for the Korean laryngeal contrast

The data was partitioned by manner and RFs were run separately for three subsets, each containing one of the three binary laryngeal contrasts (i.e. (1) aspirated & lenis, (2) aspirated & tense, (3) lenis & tense). Since there is a three-way laryngeal contrast among stops and affricates, but only a two-way contrast among fricatives, this resulted in seven separate models (three for affricates, three for stops, one for fricatives). The training data for each model involved randomly selecting one of the two repetitions from each item from each speaker. The other repetition for that item and speaker was assigned to the test data. As such, the testing and training data were equally balanced with respect to each other both by item and by speaker. All models were run with five layers and 200 nodes per layer.

The full set of acoustic measurements were f0, H1*−H2*, H1*−A1*, H1*−A2*, H1*−A3*, H2*−H4*, psychoacoustic roughness, VOT, closure duration, frication duration, aspiration duration, and release duration. RFs were run containing all possible permutations of these acoustic measurements. However, RFs that included more than one measure of spectral tilt were excluded since these various measures reflect the same physical characteristics (laryngeal constriction). A single spectral tilt measure is chosen from among the five spectral tilt measures to maximize model accuracy in each case. Roughness was only included in models reported in section 3.2.

The acoustic measures also differed for each model, since, for example, closure duration is defined only for affricates and stops, but not for fricatives. Also, since tense affricates and fricatives don't contain aspiration, the four models with tense affricates and fricatives included release duration instead of aspiration duration. RFs can learn the tense series easily based on the fact that frication duration is equal to release duration only for tense, but not for lenis and aspirated obstruents; however this reflects the actual situation in the language and so it was left as is.

A successful model is taken here to be one that achieves an overall accuracy of 95%. Relative importance values (ranging from 0 to 1) that quantify the contribution each measurement made

in each model are reported. A measure or group of measures was treated as sufficient to learn a given contrast if all models that contained those measures were successful. A measure or group of measures was treated as necessary to learn a given contrast if the only models that were successful contained those measures.

3. RESULTS

The RF model testing results based on seven acoustic measures are shown in Fig. 1. In the figure, the fully-loaded RFs with all relevant acoustic measures are arranged along the x-axis with accuracy plotted along the y-axis. Relative importance values are reflected for each acoustic measure via stacked bar plots. All fully-loaded RFs resulted in successful learning.

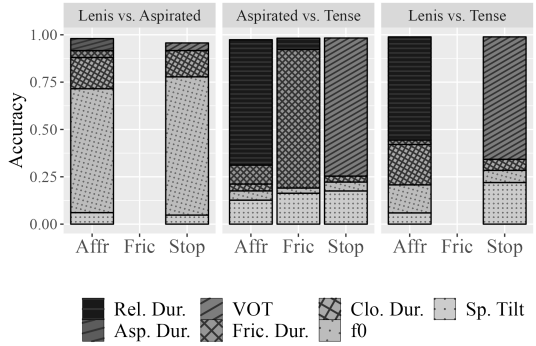


Figure 1: RF model accuracy and importance of each acoustic measure by binary laryngeal contrast and manner

Table 1 shows which acoustic measures were necessary and sufficient in each model. The results generally confirmed past findings [4, 5, 6, 7], while also offering some interesting insights which are discussed in the following subsections.

3.1. Acoustic measure importance

Among both stops and affricates, f0 was necessary for the lenis-aspirated contrast, as expected. However, among the stops, f0 wasn't sufficient on its own, confirming that other secondary cues are also relevant. The most likely secondary cue is closure duration, which had significantly higher importance than VOT and spectral tilt.

Regarding the tense series, previous research has suggested some combination of VOT, spectral tilt and f0 is likely involved for contrasts [9, 10, 11]. Our results showed that for the aspirated-tense contrast, release duration in affricates and VOT in stops were necessary for successful learning.

Table 1: Summary of necessary and sufficient acoustic measures for each contrast

Contrast	Necessary	Sufficient
Len-Asp Affr	f0	f0
Len-Asp Stops	f0	f0, clo. dur. f0, sp. tilt, VOT
Asp-Tns Affr	rel. dur.	rel. dur., any other
Asp-Tns Fric	fric. dur.	fric. dur., rel. dur. fric. dur., sp. tilt
Asp-Tns Stops	VOT	VOT
Len-Tns Affr	none	rel. dur. f0, sp. tilt, clo. dur., fric. dur.
Len-Tns Stops	none	VOT f0, sp. tilt, clo. dur.

Among fricatives, frication duration was necessary and was sufficient when paired with either release duration or spectral tilt. Thus, for affricates, release duration plays the role that VOT does for stops, and for fricatives, frication duration does the same.

In the lenis-tense contrast, no single measure was necessary, but release duration (for affricates) and VOT (for stops) were sufficient on their own. In addition, among the stops, the combination of f0, spectral tilt and closure duration achieved success; among the affricates, the same three measures plus frication duration sufficed. This result suggests primary roles for VOT and release duration, with f0, closure duration and frication duration playing secondary roles. Notably though, spectral tilt was not necessary for successful learning of the tense series. In fact, among the affricates, relative importance values indicated that closure duration plays a larger secondary role in the lenis-tense contrast, and that frication duration and spectral tilt played similarly important secondary roles in the aspirated-tense contrast. However, spectral tilt was by far the most important secondary measure for the tense fricative and stops.

3.2. Spectral tilt and psychoacoustic roughness

Another aim of this paper was to assess which of the various measures of spectral tilt and psychoacoustic roughness are optimal for learning the laryngeal contrast. To make this assessment, models containing the acoustic measures that were sufficient to reach 95% accuracy were selected, with spectral tilt or roughness necessarily included. This resulted in a set of models with minimal acoustic features, at least some of which were sufficient for successful learning, and that differed only on which spectral tilt or roughness measure was included. The resulting

model accuracies are compared in Table 2, and their relative importance values are compared in Table 3. Importance values are taken as more meaningful, since accuracy may be subject to a ceiling effect, and since importance values dispersed more widely.

Table 2: Model accuracy for spectral tilt and roughness measures. The measure with highest accuracy is bold.

Measure	Lenis-Asp		Asp-Tense			Lenis-Tense	
	Aff	Stop	Aff	Fric	Stop	Aff	Stop
Roughness	.967	**.960**	**.965**	.942	**.982**	.944	.907
H1*−H2*	.967	.959	.944	.942	.976	.946	.923
H2*−H4*	.967	.952	.958	.942	.964	.953	.915
H1*−A1*	.951	.952	.963	**.971**	.973	.946	**.958**
H1*−A2*	**.972**	.950	.963	.942	.970	.944	.925
H1*−A3*	.967	.949	.963	.947	.972	**.958**	.931

Table 3: Model importance of spectral tilt and roughness measures. The measure with highest importance is bold.

Measure	Lenis-Asp		Asp-Tense			Lenis-Tense	
	Aff	Stop	Aff	Fric	Stop	Aff	Stop
Roughness	**.135**	**.116**	.160	.178	.152	.060	.142
H1*−H2*	.135	.085	.141	.211	.159	.073	.182
H2*−H4*	.133	.103	.177	.187	.114	.117	.193
H1*−A1*	.101	.085	.218	**.257**	**.278**	**.184**	**.398**
H1*−A2*	.096	.068	.218	.202	.250	.131	.328
H1*−A3*	.096	.074	**.235**	.218	.189	.182	.277

In contrasts involving the tense series, where laryngealization was expected to play an important role, the spectral tilt measures had higher importance. However, in the aspirated-lenis contrasts, where spectral tilt is less important, roughness had higher importance. This result can be explained by the fact that roughness (but not spectral tilt) inversely correlates with f0 (in Thai [21]). Our results showed the same trend: Roughness was higher throughout low-tone vowels following lenis obstruents for most speakers, albeit with a local lowering effect at the consonant boundary.

This line of reasoning also explains why roughness fared extremely poorly in the lenis-tense distinction. Spectral tilt is lowered following tense obstruents due to laryngeal constriction and raised following lenis obstruents due to breathiness, causing diverging effects. However, roughness is raised following tense obstruents due to laryngeal constriction but is also raised due to decreased f0 following lenis obstruents, causing converging effects that make roughness a poor indicator of the contrast. This suggests roughness should only be used for identification of laryngeal constriction between two categories that do not differ in f0.

The aspirated-tense contrast provides a suitable test case to compare roughness and spectral tilt because f0 is the same. The importance values for roughness were relatively closer to spectral tilt, but still lower almost across the board. However, for stops and affricates, the accuracy for the models with roughness were higher than those with spectral tilt, suggesting that roughness may still be useful as a measure of laryngeal constriction. This was not the case for the fricatives however, where H1*−A1* was dominant for both accuracy and importance. Future research is needed to discern why fricatives differ in this way.

Finally, among the spectral tilt measures, H1*−A1* had higher importance values in all of the contrasts involving tense stops and fricatives, indicating that it is generally the best measure of laryngeal constriction among Korean tense obstruents. However, H1*−A3* may be a better measure for the tense affricates. The model with H1*−A3* had the highest importance for the aspirated-tense contrast in affricates. H1*−A3* also arguably outperformed H1*−A1* in distinguishing lenis and tense affricates by virtue of its only marginally lower importance and its higher accuracy. It is unclear why H1*−A3* would be better than H1*−A1* among affricates, but not also fricatives.

4. CONCLUSION

This study used the random forest machine learning method to learn the Korean three-way laryngeal contrast and found that f0 was most important in the lenis-aspirated distinction, and VOT was most important for the tense series, confirming previous findings. Among affricates and fricatives, release duration and frication duration were most important respectively. Spectral tilt was not necessary for learning any of the contrasts, but was found to be a significant secondary measure for most contrasts involving the tense series, as expected. Among the various spectral tilt measures, H1*−A1* was found to be relatively more important than others, although some evidence suggested that H1*−A3* may be optimal for tense affricates. Finally, psychoacoustic roughness was most important in the lenis-aspirated distinction, due to its correlation with f0 rather than laryngealization; as an indicator for laryngealization, comparable results to spectral tilt should only be expected among contrasts where f0 doesn't vary.

5. ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant-in-Aid for Early Career Scientists Number JP19K13162 and also by the ILCAA joint research project “Phonetic typology from cross-linguistic perspectives (PhonTyp)” (2021-2023).

6. REFERENCES

[1] M. R. Kim and S. Duanmu, “"tense" and "lax" stops in korean,” Journal of East Asian Linguistics, vol. 13, no. 1, pp. 59–104, 2004.

[2] C. W. Kim, “On the autonomy of the tensity feature in stop classification,” Word, vol. 21, pp. 339–359, 1965.

[3] L. Lisker and A. S. Abramson, “Cross-language study of voicing in initial stops: acoustical measurements,” Word, vol. 20, pp. 384–422, 1964.

[4] D. J. Silva, “Acoustic evidence for the emergence of tonal contrast in contemporary korean,” Phonology, vol. 23, pp. 287–308, 2006.

[5] J. Wright, “Laryngeal contrast in seoul korean,” Ph.D. dissertation, University of Pennsylvania, 2007.

[6] K.-H. Kang and S. G. Guion, “Clear speech production of korean stops: Changing phonetic targets and enhancement strategies,” Journal of the Acoustical Society of America, vol. 124, no. 6, pp. 3909–3917, December 2008.

[7] Y. Kang, “Voice onset time merger and development of tonal contrast in seoul korean stops: A corpus study,” Journal of Phonetics, vol. 45, pp. 76–90, 2014.

[8] K.-H. Kang, “Generational differences in the perception of korean stops,” Phonetics and Speech Sciences, vol. 2, no. 3, pp. 3–10, 2010.

[9] M. R. Kim, P. S. Beddor, and J. Horrocks, “The contribution of consonantal and vocalic information to the perception of korean initial stops,” Journal of Phonetics, vol. 30, pp. 77–100, 2002.

[10] M. Kim, “Correlation between vot and f0 in the perception of korean stops and affricates,” in Eighth International Conference on Spoken Language Processing, 2004.

[11] J. Schertz, T. Cho, A. Lotto, and N. Warner, “Individual differences in phonetic cue use in production and perception of a non-native sound contrast,” Journal of Phonetics, vol. 52, pp. 183–204, 2015.

[12] T. Cho, S.-A. Jun, and P. Ladefoged, “Acoustic and aerodynamic correlates of korean stops and fricatives,” Journal of Phonetics, vol. 30, pp. 193–228, 2002.

[13] M. Gordon and P. Ladefoged, “Phonation types: a cross-linguistic overview,” Journal of Phonetics, vol. 29, pp. 383–406, 2001.

[14] P. Keating, M. Garellek, and J. Kreiman, “Acoustic properties of different kinds of creaky voice,” in Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, 2015.

[15] M. Garellek, The phonetics of voice, the routledge handbook of phonetics ed. Routledge, 2019, pp. 75–106.

[16] J. Villegas, K. Markov, J. Perkins, and S. J. Lee, “Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 355–366, 2020.

[17] Y. Su, F. Jelinek, and S. Khudanpur, “Large-scale random forest language models for speech recognition,” in Eighth annual conference of the international speech communication association, 2007.

[18] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, “A tree-based statistical language model for natural language speech recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 7, pp. 1001–1008, 1989.

[19] K. H. Oo, “Comparing accuracy between svm, random forest, k-nn text classifier algorithms for detecting syntactic ambiguity in software requirements,” in International Conference on Information Systems and Intelligent Applications. Springer, 2023, pp. 43–58.

[20] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version xxx),” 2022.

[21] J. Perkins, “Acoustic measurement of laryngeal constriction in thai consonants,” in 35th General Meeting of the Phonetics Society of Japan, September 25-26 2021.

THE EFFECT OF PROSE AND POETRY TEXT LAYOUT ON THE ACOUSTICS OF READING ALOUD

Filip Nenadić[1], Benjamin V. Tucker[2,3]

[1]Singidunum University; [2]Northern Arizona University; [3]University of Alberta
filip.nenadic@fmk.edu.rs, benjamin.tucker@nau.edu

ABSTRACT

Previous studies show that the same text is cognitively processed and interpreted differently based on the layout it is presented in – differentiating, for example, prose from poetry. Differences in how these two distinct literary genres are approached could influence the acoustic properties of their oral reading. A recent study by Blohm et al. [1] showed that the same artificially constructed short text presented in a poetic layout (with verses) was read slower and with more frequent, yet shorter pauses. In the present study, we investigate the effect of prose versus poetry text layout on oral reading of ecologically valid short texts written and formatted by published poets. We attempt to replicate the speech production findings from Blohm et al., yet find partly different results. We discuss how these disparate outcomes in oral production of prose and poetry may depend on text characteristics (rhythm and meter, difficulty) and reader characteristics.

Keywords: poetry, genre, oral reading, speech acoustics, prosody

1. INTRODUCTION

Characteristics of speech production vary depending on genre and register. Recent studies of oral reading have examined differences in the way in which poetry and prose is read aloud. The genre-specific hypothesis of reading claims that the first step in any reading is determining the text's genre and adjusting cognitive processing depending on the genre of the text [2, 3, 4]. Studies show that changing the text layout from prose to verses changes readers' genre decision, processing strategies, interpretation of a given text, speech acoustics in oral production [1, 5, 6, 7].

Where speech acoustics of oral reading is concerned, in a case study in Swedish, Fant [8] found that two poems were read more slowly and with a more stable tempo, with higher intensity, and with a higher and less varying F0 when compared to their reformatted prose forms and a separate novel excerpt. Similarly, Tsur [9] focused on case-studies of expert recitations of Shakespeare and Milton, stressing the importance of line breaks and meter on performance.

More recently, Blohm et al. [1] conducted an experimental study in which they constructed short texts in German. The texts consisted of two sentences and the authors altered their layout from prose to poetry. Their participants were instructed to read the texts aloud and their eye-movements were tracked. The participants read texts presented in a poetic layout more slowly as both their articulation rate was slower and the total time spent in silent pauses was longer. The pauses were more frequent and on average shorter when reading a text presented in verse. These differences were also reflected in the recorded eye-movements.

Furthermore, the constructed texts featured a rhythmic pattern in which strong and weak syllables alternated. The authors focused on critical regions consisting of two function words at the beginning of a sentence (which could either be the first or the second sentence in the short text), followed by a post-critical region content word. They observed that the duration ratio between the strong and the weak syllables in the critical region was larger when a text was read in its poetry versus prose format. Such local differences were not noted with regard to ratios of strong/weak syllable intensity or pitch. In the post-critical region, both strong and weak syllables were longer in the poetic layout.

In the present paper, we expand on this line of research by investigating differences in oral readings of prose and poetry in English. Our participants have a similar background to those tested by Blohm et al. [1] and we use a number of the same acoustic measures to investigate their productions of prose and poetry stimuli. However, the key difference is that in our study we did not construct texts to have a particular rhythmic pattern. Instead, we used original short texts written and formatted by young published authors for the purposes of this study. Therefore, the central research question of the present paper is whether text layout (prose versus

poetry) has an impact on oral reading of literary texts even when there is no meter or rhythmic pattern, that is, when free verse is used.

2. METHODS

2.1. Participants

Thirty-six native speakers of English participated in the study (11 male; ages 18-32). All of the participants were recruited from the University of Alberta community. The majority were undergraduate students enrolled in an introductory linguistics course who received course credit for their participation. Participants were recruited as non-experts in literary reading.

2.2. Stimuli

The stimuli were 20 original short texts (word count M = 59.4, SD = 9.13) written by seven young published authors for the purposes of the present study. The authors were instructed to write texts that they deemed acceptable as both short prose and poetry. These texts were further formatted in four layouts by their authors: double-aligned (justified) and left-aligned prose, verses in a single stanza, and verses in two or more stanzas, for a total of 80 different stimuli (see Figure 1 as an example). Three additional texts written by some of the same authors were used as practice stimuli.

Since no other restrictions besides adequacy in both the prose and poetic format were imposed in text creation, the stimuli did not rhyme or have a strict metric or rhythmic structure. Effectively, the poetic layouts can therefore be considered free verse [10, 11].

2.3. Procedure

The participants were given the task to read aloud the texts appearing on a computer screen. Their instruction was to read the texts the way they would read them aloud in company of others or if, for example, a teacher were to ask them to read aloud for the rest of the class. Participants' voice and eye-movements were recorded during reading (although we do not analyze the eye-movement data in the present report).

The participants first read three practice texts. They proceeded with reading the 20 texts, five in each of the four text layouts. The layout in which a certain text would appear was counterbalanced across participants using a Latin square design. The text presentation order was randomized for each participant; different text layouts were not placed in separate blocks.

A short break was offered after 10 texts were read. Upon reading all 20 texts and another break, the participants read the same texts in the same layouts again, but in a new randomized order (although we only analyze the first pass in the present report).

2.4. Data Analysis

One reading of a certain text by one participant constituted a single recording. These recordings were forced aligned using the Penn Forced Aligner [12]. All of the aligned recordings were manually inspected and all recordings that contained an error in reading were excluded.

We could not replicate the analysis Blohm et al. [1] on the local level because our stimuli do not feature the same strong/weak syllable pattern. Instead, we focused on the acoustic measures calculated on the global level. We considered seven global measures of oral reading and calculated them for each of the retained recordings. Two of these seven measures were standard deviation of intensity and fundamental frequency. Both the intensity and the fundamental frequency values were automatically extracted for vowels using a Praat script [13]. The other five global measures were identical to those considered by Blohm et al. [1]:

- **speech rate** – number of syllables divided by the total reading time (pauses included)
- **articulation rate** – number of syllables divided by the total time spent articulating (i.e., with silent pauses excluded)
- **pause rate** – number of syllables divided by number of pauses
- **pause duration** – summed duration of all silent pauses
- **pause proportion** – pause duration divided by the total reading time

Manual inspection of the recordings showed that 39% had at least one reading error. These recordings were excluded for two reasons. First, automatic alignment cannot function properly if changes are introduced in comparison to the expected text. Second, we cannot know what impact a reading error may have on overall reading performance and, specifically, the acoustic measures we estimated. We further excluded four participants that had only six or fewer recordings without errors remaining. The final number of retained recordings was 419 and a chi-square test showed that all four text layouts were equally present among the retained recordings ($\chi^2 = 0.33$, $df = 3$, $p = .95$).

Each of the described measures of oral reading

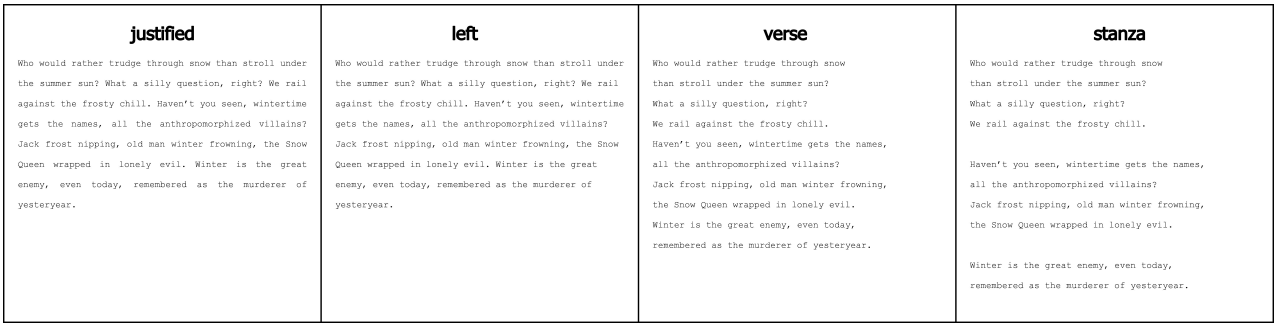


Figure 1: Sample text-stimulus in four layouts.

was used as a dependent variable in a separate linear mixed effects model. Model fitting was performed in R [14] using the lme4 package [15]. We created two sets of models. In one set, the independent variable was text layout with four levels (justified prose, left-aligned prose, verses, verses in stanzas). Post-hoc analyses in these models were performed using the multcomp package [16]. In the second set of models, we only considered the distinction between prose (conflating justified and left-aligned prose) and poetry (conflating verses in one stanza and verses in multiple stanzas). Random intercepts per subject and per text were included in all models. The independent variable (one of the four text layouts or the prose versus poetry distinction) was considered significant if it significantly improved model fit when it was added to a baseline model that included random intercepts only. Model criticism and trimming was performed for all models; model trimming never excluded more than 2.5% of data.

3. RESULTS

We found a significant contribution to model fit in four cases. In the case of speech rate, a significant contribution of the independent variable was registered both when all four layouts (justified prose, left-aligned prose, verses, and stanzas) were considered separately and when a simple prose versus poetry distinction was made. Since it is more informative, we will report on the model that distinguished between the four layouts. We found that texts formatted as left-aligned prose were produced with a higher speech rate than texts formatted as verses ($\beta = 0.35$, $SE = 0.12$, $z = 2.85$) and stanzas ($\beta = 0.40$, $SE = 0.13$, $t = 3.15$). Other pairwise comparisons were not significant (Figure 2).

We also recorded a significant effect of text layout on articulation rate and proportion of pauses, but these effects reached significant levels only when

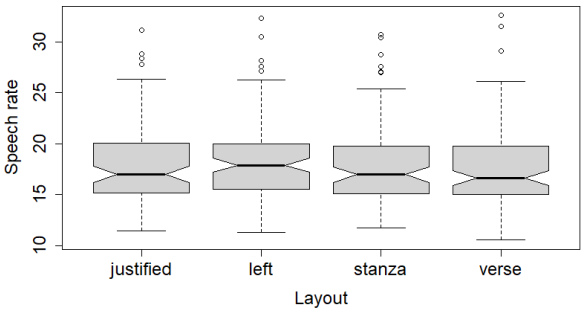


Figure 2

Figure 3: Boxplot of differences in speech rate between the four textual layouts.

we considered the two prose (justified and left-aligned) and the two poetry layouts (verses and stanzas) jointly (Figure 4). Lower articulation rate was registered for poetry ($\beta = -0.20$, $SE = 0.08$, $t = -2.41$). In other words, when they articulated, participants articulated fewer syllables per unit of time. Additionally, participants had a higher proportion of pauses when they read poetry, meaning that they spent more of the total speaking time being quiet ($\beta = 0.005$, $SE = 0.002$, $t = 2.26$).

4. DISCUSSION

The goal of the present paper was to test whether there are differences in acoustic measures of oral productions of the same short literary texts when formatted as prose versus poetry. Studies such as this one serve to inform us about (1) the reading strategies, cognitive processing, and interpretation of different (literary) genres and (2) variability in speech production dependent on the genre/register and purpose of speaking.

The number of potential indicators of performance that can be extracted from a recording of oral reading

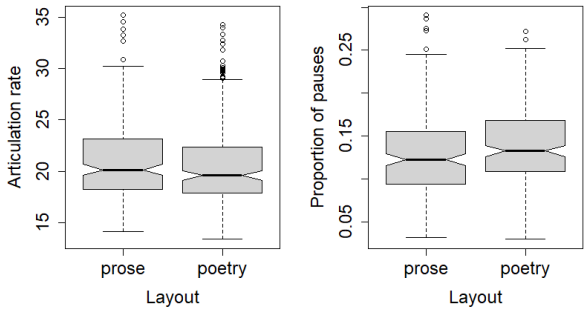


Figure 4

Figure 5: Boxplot of differences in articulation rate and proportion of pauses between prose and poetry.

of a brief text is very large. Different researchers may opt for different operationalizations and these researcher degrees of freedom may have an impact on the results of the study [17]. To avoid this possibility, we utilized the same global measures of oral reading as Blohm et al. [1]. Additionally, we included estimates of standard deviation of fundamental frequency and intensity on the global level, which Blohm et al. [1] analyzed in their local, critical regions.

We find differences in the way the same texts are read when they are presented in the prose versus poetry layout. Our participants generally spent more time reading poetry, as their speech rate was slower. Both lower articulation rate (fewer syllables produced per unit of time when articulating) and higher proportion of pauses (more total reading time spent silent) contributed to this difference in speed. However, we do not find evidence of more frequent pauses, nor differences in pause duration.

Since the participants tested in Blohm et al. [1] and our studies had similar backgrounds, another possible reason could be the differences in the way the stimuli in these two studies were created and, consequently, their characteristics. Our stimuli were written by young published authors for the purposes of the study with instruction to write texts that they believe could be fit as both short prose and a short poem, yielding texts that did not feature a consistent metric or rhythmic structure.

Although periodical repetition of rhythmic patterns is considered a frequent formal constraint of conventional regulated verse, it is not necessarily present in all poetry [10, 11]. Still, the removal of these characteristics may eliminate cues used for expressive poetry reading; free verse is often described as similar to conversational speech. Perhaps our readers found no place for an expressive pause or the need to make longer, more dramatic pauses. Previous studies show that reader appreciation and even ratings of the text as a poem ('poeticity' ratings) dwindle when meter and rhyme were deteriorated or removed from poems when they are expected to be found [18, 19, 20]. If this is true, then future studies should examine which acoustic aspects of expressive oral reading are changed depending on presence or absence of certain text characteristics or poetic devices.

Another possibility is that the stimuli we employed were too long and/or too difficult for our participants. We registered reading errors in almost 40% of recordings, much more than the 11% recorded by Blohm et al. [1]. Studies that tested children's oral reading show that decoding difficulties impact reading prosody by increasing the number and duration of (unexpected) pauses and making the F0 prosodic profile less similar to those registered in proficient, adult speakers [21, 22]. Difficult texts seem to make it more difficult for readers to maintain focus on comprehension, increasing mind wandering [23]; if a reader is allocating cognitive control and resources to decoding, it may be difficult to simultaneously provide an expressive oral reading performance.

This interpretation should be considered with caution, however. It could be that the higher number of errors we record is simply a consequence of a longer text (and therefore more opportunity to make a mistake), not a consequence of text difficulty. The Flesh reading ease score of our texts ranged between 42.4 and 98.2 ($M = 76.19$, $SD = 14.27$), with lower values indicating that a text is more difficult to read. In comparison, text difficulty in the study conducted by Blohm et al. [1] ranged between 40 and 87 ($M = 65.33$, $SD = 8.87$).

Despite these stimuli differences, we cannot be sure that they are the (sole) cause behind partly disparate results. The two studies tested speakers of different languages which may be behind the somewhat disparate results. Cross-linguistic investigation is necessary to expand our understanding of topic. Perhaps the necessary exclusion of recordings reduced our statistical power to capture all the fine differences in relevant acoustic measures, as these effects are often subtle. In the present report, we only considered the first reading of the texts from each participant. A pending analysis of their second pass should lead to an increase in the statistical power, but also potentially yield fewer errors and more expressive reading after the practice participants had with the stimuli.

5. REFERENCES

[1] S. Blohm, S. Versace, S. Methner, V. Wagner, M. Schlesewsky, and W. Menninghaus, "Reading poetry and prose: Eye movements and acoustic evidence," *Discourse Processes*, vol. 59, no. 3, pp. 159–183, 2022.

[2] D. Hanauer, "The genre-specific hypothesis of reading: Reading poetry and encyclopedic items," *Poetics*, vol. 26, no. 2, pp. 63–80, 1998.

[3] R. A. Zwaan, "Some parameters of literary and news comprehension: Effects of discourse-type perspective on reading rate and surface structure representation," *Poetics*, vol. 20, no. 2, pp. 139–156, 1991.

[4] ——, "Effect of genre expectations on text comprehension," *Learning, Memory, and Cognition*, vol. 20, no. 4, pp. 139–156, 1994.

[5] M. Fechino, A. M. Jacobs, and J. Lüdtke, "Following in jakobson and lévi-strauss' footsteps: A neurocognitive poetics investigation of eye movements during the reading of baudelaire's 'les chats'," *Journal of Eye Movement Research*, vol. 13, no. 3, 2020.

[6] D. Hanauer, "Integration of phonetic and graphic features in poetic text categorization judgments," vol. 23, no. 5, pp. 363–380, 1996.

[7] J. Peskin, "The genre of poetry: Secondary school students' conventional expectations and interpretative operations," *English in Education*, vol. 41, no. 3, pp. 20–36, 2007.

[8] G. Fant, *Speech acoustics and phonetics*. Dordrecht: Kluwer Academic, 2004.

[9] R. Tsur, "Poetry reading–rhythmical performance: Triple-encoding, and voice quality: Six case studies," *Thinking Verse*, vol. 2, pp. 88–111, 2012.

[10] N. Fabb, "There is no psychological limit on the duration of metrical lines in performance: Against turner and pöppel," *International Journal of Literary Linguistics*, vol. 2, no. 1, pp. 1–29, 2013.

[11] ——, *What is Poetry?: Language and Memory in the Poems of the World*. Cambridge University Press, 2015.

[12] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics '08*, pp. 5687–5690, 2008.

[13] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2019. [Online]. Available: www.praat.org

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008.

[17] S. Coretta, J. V. Casillas, and T. B. Roettger, "Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses," Nov 2022. [Online]. Available: psyarxiv.com/q8t2k

[18] D. Hanauer, "Reading poetry: An empirical investigation of formalist, stylistic, and conventionalist claims," *Poetics Today*, vol. 19, no. 4, pp. 565–580, 1998.

[19] C. Obermeier, W. Menninghaus, M. Von Koppenfels, T. Raettig, M. Schmidt-Kassow, S. Otterbein, and S. A. Kotz, "Aesthetic and emotional effects of meter and rhyme in poetry," *Frontiers in Psychology*, vol. 4, p. 10, 2013.

[20] W. van Peer, "The measurement of metre: Its cognitive and affective functions," *Poetics*, vol. 19, no. 3, pp. 259–275, 1990.

[21] R. G. Benjamin and P. J. Schwanenflugel, "Text complexity and oral reading prosody in young readers," *Reading Research Quarterly*, vol. 45, no. 4, pp. 388–404, 2010.

[22] P. J. Schwanenflugel, A. M. Hamilton, M. R. Kuhn, J. M. Wisenbaker, and S. A. Stahl, "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[23] S. Feng, S. D'Mello, and A. C. Graesser, "Mind wandering while reading easy and difficult texts," *Psychonomic Bulletin & Review*, vol. 20, no. 3, pp. 586–592, 2013.

WHAT MAKES A GEMINATE? AN ACOUSTIC STUDY OF CONSONANTS IN GARO

Cheman Baira A'gitok[1], Angeliki Athanasopoulou[2]

McGill University[1], University of Calgary[2]
cheman.asangma@mail.mcgill.ca, angeliki.athanasopou@uclagary.ca

ABSTRACT

We investigate the phonetic properties of consonants in Garo, a Sino-Tibetan language, to examine the gemination process that applies to intervocalic consonants in coda position within a word. We compared nasals, liquids, and stops in onset positions, coda positions within a word and word finally, and sequences of identical consonants (concatenated geminates) as produced by four native speakers. We found that only consonants that are allowed in both coda and onset positions geminate. Geminated coda consonants are as long as sequences (concatenated geminates) and 50% (or more) longer than onset consonants (singletons). Geminated nasals and nasal sequences are identical, but geminated stops differ from stop sequences, as the former do not aspirate (short VOT).

Keywords: gemination, length, consonants, syllabification, Garo, Sino-Tibetan

1. INTRODUCTION AND PRESENT STUDY

Garo is a Sino-Tibetan language spoken in Northeast India by around 1.5 million speakers [1]. Most of the work on Garo has been impressionistic descriptions, so we know very little about its phonological and phonetic system. A recent phonological study [2] proposed the existence of surface geminate consonants and explained them as a result of an interaction between morphology and phonology. A'gitok [2] also points out that Garo geminates do not fit with the current geminate consonant typology since geminates are either underlying [3] or due to assimilation [4], but Garo geminates are a consequence of syllabification.

The claims in A'gitok [2], however, are based on the author's language knowledge. We do not know the phonetic properties of these surface geminates and what distinguishes them from singleton consonants. We also do not know if all the consonants geminate, or if there are additional considerations required for the phonological analysis of gemination.

Geminate consonants are typically longer in duration than singleton consonants, but the strength of the difference and the existence of secondary phonetic features depends on the consonant and the language [5–7]. So, although we predict that geminate consonants in Garo are longer than singletons, we do not know if this is true for all and what the ratio of the difference is.

2. GARO LANGUAGE AND GEMINATION

Garo is a monosyllabic agglutinating language, so words are long and morphologically complex. Garo has 17 phonemic consonants: /p, b, t, d, k, g, ʔ, m, n, ŋ, ɾ, s, w, l, t͡s, d͡z/ and 6 phonemic vowels: /i, e, ə, a, o, u/ [2, 6–8]. There are restrictions on the distribution of these sounds however: /b, d, g, ɾ, s, w, t͡s, d͡z/ are limited to onset positions, while /ʔ, ŋ, l/ are limited to the coda positions.

Garo syllables may have up to two onset consonants (e.g., [sk^{h}o] "head") and up to two coda consonants where the second consonant can only be [ʔ] (e.g., [amʔ.na "to search"), but we also find a plain V shape (e.g., [o.a] "open-PRS").

In Garo words, if in a morpheme sequence the first morpheme ends in a closed syllable and the following morpheme starts with an onsetless syllable, the coda consonant of the first morpheme gets geminated [2], as in examples (1, 3). Burling [9] also reports that words like (1) sound identical to (2), but he does not identify a gemination process or discuss the existence of geminate consonants in the language.

(1) /t͡sa**n**-a/ → [t͡s^{h}a**nn**a] "count-PRS" *[t͡s^{h}a**n**a]
(2) /t͡sa**n**-**n**a/ → [t͡s^{h}a**nn**a] "count-INF"
(3) /nok-o/ → [no**kk**o] "house-LOC" *[no**k**^{h}o]
(4) /nok-ko/ → [no**kk**^{h}o] "house-ACC"
(5) /nok-na/ → [no**k**na] "house-DAT"

A'gitok [2] shows that Garo gemination is due to the language's syllabification requirements. He proposes that the right edge of morphemes need to align with the syllables. This alignment requirement prevents the coda consonant from resyllabifying as onset of the following syllable (*[t͡s^{h}a.na] in (1)), but instead, it geminates, preserving the original coda of the morpheme on the left and creating an onset for the morpheme on the right ([t͡s^{h}an.na] in (1)).

Depending on the consonant being geminated, we may find homophones (1-2) or minimal pairs (3-4). Minimal pairs arise in the case of voiceless stops because they are aspirated in onsets (4), but

unaspirated when they geminate (3) or in coda (5). Despite the difference in terms of aspiration, closure duration (CD) might be an important phonetic property, since in most languages, CD and not voice onset time (VOT) is the primary property of geminate stops [5–7].

All the consonants that occur in the coda position geminate, but it is unclear if /ʔ/ geminates but A'gitok [2] suggests that it does not. This description, however, might be because the difference between the geminate and singleton /ʔ/ is not consonant duration (e.g., Maltese [5]). Due to space limitations, we will not discuss /ʔ/ in this paper.

Finally, it is not known if consonants limited to codas, i.e., /ŋ/ and /l/ geminate and what the prosodic domain for gemination is. A'gitok [2] claims that gemination does not occur between words in a phrase, so the likely domain is the phonological word.

The present study is a phonetic investigation of Garo geminate consonants to test if duration differentiates geminates from singletons and consonant sequences (concatenated geminates). We hypothesize that geminate consonants are longer than their singleton counterparts, and equally long as sequences. For stops, we hypothesize that the length of CD is the primary property, but VOT is also relevant due to the aspiration pattern discussed above.

We also test consonants at the end of a word to find if gemination applies across words. We hypothesize that gemination does not apply in this context, and consonants at the end of a word are equally long as singleton consonants within a word.

4. METHOD

Four native speakers of Garo (2 females; 18-25 years) were recruited from Tura in West Garo Hills. All were college educated and reported no speech, hearing, language, or cognitive problems.

The data was recorded in the music room of Hawakhana Baptist Church, Tura. After consent, the participants sat in front of a monitor where the stimuli were projected. Initially, they saw instruction slides and 9 practice items and then continued with the test items. They also had a short break after 30 minutes. The speech data was recorded in Praat [11] at 44.1 KHz using a head mounted Logitech H390 microphone connected to the computer via USB port.

4.2. Stimuli

Consonants in coda (geminate), onset (singleton), and sequence of two identical consonants (concatenated geminates) were recorded in the second syllable of real trisyllabic words and were preceded and followed by a vowel. Since /ŋ/ is not found in onset positions, the onset and sequence conditions were formed with /n/ in onset. Also, since /l/ is not found in the onset and /ɽ/ is not found in the coda position, we combine them into a single comparison. Consonants were also recorded at the end of a disyllabic word (noun) before the adjective [apsanko] "same", which starts with a vowel. We were not able to find any words ending with /l/ to fit with the phrase, so this condition was not tested for the liquids. See Table 1 for examples.

Condition		**Example***
Coda V**C**.V	Nasal	/sa**n**-a-de/→[sa**n**.**n**a.de] "the act of nursing"
	Liquid	/sa**l**-a-ko/→[sa**l**.**l**a.kʰo] "drawings"
	Stop	/go**p**-a-na/→[go**p**.**p**a.na] "because of buring smth"
Onset V.**C**V	Nasal	/sa-**n**a-de/→[sa.**n**a.de] "to serve"
	Liquid	/sa-**ɽ**aŋ-ko/→[sa.**ɽ**aŋ.kʰo] "who (plural)"
	Stop	/go-**p**a-na/→[go.**p**a.na] "to shoot in imitation"
Sequence V**C**.**C**V	Nasal	/sa**n**-**n**a-de/→[sa**n**.**n**a.de] "to nurse"
	Liquid	/sa**l**-**ɽ**aŋ-ko/→[sa**l**.**ɽ**aŋ.kʰo] "days"
	Stop	/go**p**-**p**a-na/→[go**p**.**pʰ**a.na] "to bury smth in imitation"
Word end V**C**# V	Nasal	/meʔ-go**n** # apsanko/ → [meʔ.go**n**#] "the same stirrer"
	Liquid	N/A
	Stop	/pi-to**p** # apsanko/ → [pʰi.tʰo**p**#] "the same cheek"

* . = syllable; - = morpheme; # = word boundaries

Table 1: Examples of stimuli used in elicitation.

The target words were elicited in the carrier sentence (6) presented in a dialogue format after the question "*Is Derang going to think about the word called 'ka'donga' tomorrow?*" Participants had to read both the question and answer of the dialogue.

(6) Carrier sentence:

[m̥hm̥ deraŋ-ŋara X məŋ-gəppa kʰattʰa-ni
no Derang-TOP X call-ADJ word-GEN

gəmən-sa kʰnal-lo t͡sʰant͡sʰi-gen-na-ba]
about-FOC tomorrow-LOC think-FUT-EVI-FACT

"No, Derang is going to think about the word called X tomorrow."

Using Praat [11], we measured the duration of nasals and liquids, and CD and VOT of stops. We run separate mixed-effects analyses in SPSS [12] for each

measurement. In each analysis, the dependent variable was the Duration of the consonant or CD or VOT, the independent variables were Condition (coda, onset, sequence, and word), and Place of Articulation (labial, alveolar, and velar) when relevant, and the random effect was Participant.

5. RESULTS

5.1. Nasal consonants /m/, /n/, /ŋ/

There is a significant main effect of Condition, $F(3, 9.08) = 11.03$, $p = .002$, and of Place of Articulation, $F(2, 6.66) = 9.4$, $p = .011$, on the duration of nasal consonants. There is also a significant interaction between the two factors, $F(5, 110.02) = 3.59$, $p = .005$. Figure 1 summarizes the results.

Bonferroni posthoc tests reveal that in Coda, nasals are 29 ms longer than in Onset ($p < .001$) and 35 ms longer than in Word ($p < .001$), but 15 ms shorter than in Sequence ($p = .023$). Nasals in Onset do not differ statistically from nasals in Word ($p = 1$) and nasals in Sequence are the longest ($p < .001$).

We ran separate mixed-effects analyses for each place of articulation, to explore the significant interaction. We found /m/ and /n/ in Coda do not differ statistically from in Sequence ($p = 1$), but /ŋ/ is 34 ms shorter in Coda than in Sequence ($p = .009$). Also, /ŋ/ in Coda does not differ statistically from in Word ($p = .153$).

With respect to differences between the three places of articulation, we found that /m/ is 19 ms longer than /n/ ($p<.001$), while /ŋ/ does not differ statistically from /m/ ($p=.297$) or /n/ ($p=.067$).

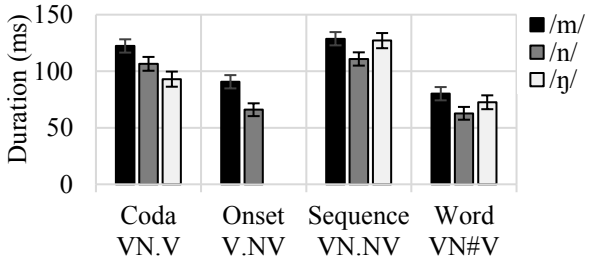


Figure 1: Mean duration of nasal consonants by place of articulation and condition. Error bars = Standard Error

Overall, we found that nasals are the longest in Coda and Sequence and the shortest in Onset and Word. This is compatible with nasals being geminated in coda position. In Coda, however, nasals are shorter than in Sequence, but their difference is only 15 ms and statistically significant only for /ŋ/. Given the length difference of /ŋ/ in Coda and Sequence and the lack of difference in Word, /ŋ/ does not participate in the gemination pattern.

5.2. Liquid consonants /l/, /ɽ/

There is a significant main effect of Condition, $F(2, 5.69) = 26.95$, $p = .001$ on the duration of liquid consonants. Bonferroni posthoc tests reveal that Coda /l/ ($M = 102$ ms, $SE = 5.38$) is 52 ms longer than Onset /ɽ/ ($M = 50$ ms, $SE = 5.38$, $p < .001$), but Coda /l/ does not differ statistically from the Sequence of liquids ($M = 88$ ms, $SE = 5.6$, $p = .255$). Also, the Sequence of liquids is 38 ms longer than Onset /ɽ/ ($p < .001$).

Overall, we found that liquids in Coda are equally long as in Sequence, and liquids are the shortest in Onset. It is hard to draw strong conclusions since there were no liquids in the Word condition and only /l/ is in coda and only /ɽ/ is in onset positions.

5.3. Stop consonants /p/, /t/, /k/

5.3.1. Voice Onset Time (VOT)

There is a significant main effect of Condition, $F(3, 9.38) = 36.59$, $p < .001$, and Place of Articulation, $F(2, 124.61) = 7.94$, $p < .001$, on VOT. The interaction of the two factors is not statistically significant, $F(6, 124.46) = 0.99$, $p = .438$. See Figure 2.

Bonferroni posthoc tests reveal that in Coda, VOT is 37 ms shorter than in Onset ($p < .001$) and 35 ms shorter than in Sequence ($p < .001$), but it does not differ statistically from the VOT in Word ($p = 1$). Also, in Onset, VOT is 40 ms longer than in Word ($p < .001$), but it does not differ statistically from the VOT in Sequence ($p = 1$). Finally, in Word, VOT is 38 ms shorter than in Sequence ($p < .001$).

With respect to VOT differences between the three places of articulation, we found that the VOT of /p/ is 11 ms shorter than /k/ ($p < .001$), but /t/ does not differ statistically from /p/ ($p = .136$) or /k/ ($p = .247$).

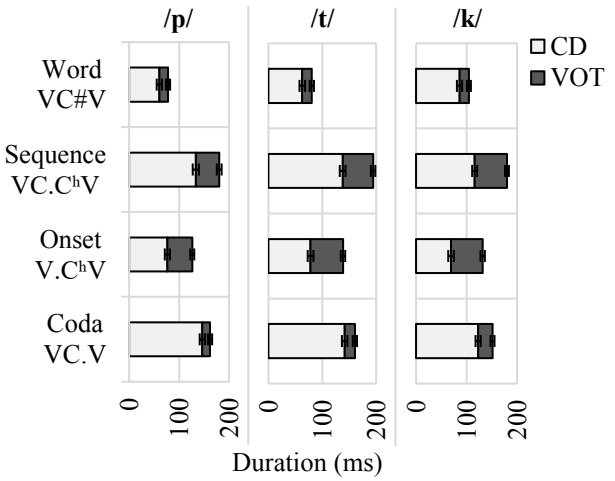


Figure 2: Mean CD and VOT in stops by place of articulation and condition. Error bars = Standard Error

Overall, we found that VOT in Coda is similar to VOT in Word and both are the shortest in comparison to Onset and Sequence. This is consistent with the stops being unaspirated in coda. In Onset and in Sequence, stops have equally long VOT, consistent with being aspirated in onset.

5.3.2. Closure Duration (CD)

There is a significant main effect of Condition, $F(3, 9.63) = 100.78$, $p < .001$ on CD. Place of Articulation was not statistically significant, $F(2, 125.38) = 2.42$, $p = .093$. The interaction between the two factors is statistically significant, $F(6, 125.07) = 6.78$, $p < .001$.

Bonferroni posthoc tests reveal that CD in Coda is 62 ms longer than in Onset ($p < .001$) and 67 ms longer than in Word ($p < .001$), but it does not differ statistically from CD in Sequence ($p = .437$). In Onset, CD is 54 ms shorter than in Sequence ($p < .001$), but it does not differ statistically from CD in Word ($p = 1$). Finally, in Sequence, CD is 59 ms longer than in Word ($p < .001$). When we ran separate mixed-effects analyses for each place of articulation to explore the significant interaction, we found the same results as described here for each stop.

Overall, we found that CD is longest in Coda and Sequence and shortest in Onset and Word. This is compatible with stops being geminated in coda.

5.3.3. CD and VOT

To look at CD and VOT results together, in addition to ms, we calculated the % of the stops' duration occupied by CD and VOT. In Sequence, where there are two stops, the first unaspirated and the second aspirated, there is a combination of long CD (M = 129 ms; 70% of the stop) and long VOT (M = 56 ms; 30% of the stop). In Onset, where stops are aspirated, CD is short (M = 75 ms; 57% of the stop), but similar to Sequence, VOT is long (M = 57 ms; 43% of the stop). Almost half of the stop's duration is its VOT. In Coda, where stops are unaspirated and geminated, similar to Sequence, CD is long (M = 137 ms; 87% of the stop), but VOT is short (M = 21 ms; 13% of the stop). Finally, in Word, stops are in coda position but not in an environment for gemination to occur. Similar to Onset, CD is short (M = 70 ms; 80% of the stop) and a similar to Coda, VOT is also short (M = 18 ms; 20% of the stop).

6. DISCUSSION AND CONCLUSION

The purpose of this paper was to investigate whether duration differentiates geminate consonants from singletons and consonant sequences in Garo and whether gemination applies in phonological domains larger than the phonological word. We tested three consonant classes (stops, nasals, and liquids) in different positions: word coda, word onset, word finally, and in sequences of identical consonants word medially. Stops' VOT and closure duration (CD) were tested separately.

The results confirm A'gitok's [2] description of Garo that intervocalic consonants in coda position, within a word, geminate. We found that these coda consonants are 46%-105% longer than their onset counterparts and very similar in length to consonant sequences. The only exception was the velar /ŋ/, which is no longer in coda than in other positions, indicating that it does not geminate like the other consonants. If [2] is correct and gemination occurs to create an onset for the following onsetless syllable, then the lack of gemination in /ŋ/ can be explained from the fact that /ŋ/ is not allowed in onset positions, but it is only found in coda, for both native and loan words. /l/, on the other hand, seems to geminate despite being found in codas only. In contrast to /ŋ/, we do find /l/ in onset, but only in loanwords, e.g., [lau] "gourd". So, perhaps the phonotactics for /l/ have changed due to the loanwords. We also have to keep in mind that the results for the liquids are not very strong since we are comparing the duration of /l/ in coda to the duration of /ɽ/ in onset. It is possible that there are inherent duration differences to the sounds that make /l/ appear geminate, when it is not.

In stops, we tested both CD and VOT. As expected [5–7], we found that CD is the primary cue of length in stops: coda CD is 83% longer than onset CD and equally long to stop sequences. VOT is also a cue of gemination since the geminated stops do not aspirate and have roughly 60% shorter VOT than aspirated stops (onset).

Our results also indicate that gemination only applies within a phonological word. The consonants we tested at the end of a word before a word that starts with a vowel are 33%-45% shorter than coda consonants within the word and very similar in length to the onset consonants. Word final stops do not geminate or aspirate, which confirms the descriptions [1,9] that stops in Garo aspirate in onset position.

According to A'gitok [2], gemination arises because it is the only way to satisfy the requirement for onsets without violating the requirement for alignment between the right edges of morphemes and syllables. The results of this study indicate however, that additional considerations are required to explain gemination in Garo. A complete analysis needs to address exceptions to gemination, such as /ŋ/, the interaction with the aspiration of stops, and limit the domain to the phonological word.

7. REFERENCES

[1] I. Office of the Registrar General & Census Commissioner, “2011 Census of India,” 2011. [Online]. Available: https://censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf.

[2] C. B. A’gitok, “Garo gemination : A case of faithfulness producing non-moraic geminates,” 2022.

[3] B. Hayes, “Compensatory Lengthening in Moraic Phonology,” *Linguist. Inq.*, vol. 20, no. 2, pp. 253–306, 1989.

[4] S. Kotzor, A. Wetterlin, and A. Lahiri, “Bengali geminates: Processing and representation,” in *The phonetics and phonology of geminate consonants*, H. Kubozono, Ed. Oxford University Press, 2017.

[5] H. Mitterer, “Not all geminates are created equal: Evidence from Maltese glottal consonants,” *J. Phon.*, vol. 66, pp. 28–44, 2018, doi: 10.1016/j.wocn.2017.09.003.

[6] R. Ridouane, “Geminates at the junction of phonetics and phonology,” *Lab. Phonol.*, 2010, doi: 10.1515/9783110224917.1.61.

[7] M. H. Hamzah, J. Fletcher, and J. Hajek, “Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay,” *J. Phon.*, vol. 58, pp. 135–151, 2016, doi: 10.1016/j.wocn.2016.08.002.

[8] C. B. A’gitok, “Non-convergence of pitch and duration: Word-prosody of Garo,” University of Calgary, 2022.

[9] R. Burling, *The Language of the Modhupur Mandi (Garo), Volume 1*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2003.

[10] R. Burling, “Garo,” in *The Sino-Tibetan Languages*, G. Thurgood and R. J. LaPolla, Eds. Routledge, 2017.

[11] P. Boersma and D. Weenink, “Praat: doing phonetics by computer.” 2022, Accessed: Jan. 24, 2022. [Online]. Available: http://www.praat.org/.

[12] IBM SPSS Statistics for Windows, Version 28.0 (released 2021). Armonk, NY: IBM Corp.

A DIACHRONIC STUDY OF VOWEL HARMONY IN FRENCH BROADCAST SPEECH SINCE 1940

Juliusz Cecelewski[1], Cédric Gendrot[1], Martine Adda-Decker[1,2], Philippe Boula de Mareüil[2]

[1]Laboratoire de Phonétique et Phonologie (UMR 7018), CNRS/Sorbonne-Nouvelle, Paris, France,
[2]Université Paris-Saclay, CNRS, LISN, Orsay, France
juliusz.cecelewski@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr,
martine.adda-decker@sorbonne-nouvelle.fr, philippe.boula.de.mareuil@limsi.fr

ABSTRACT

In French, vowel harmony (VH) has traditionally been described as an optional and highly variable process limited to the conversational speech. With the aim of going beyond usual impressionistic descriptions, we propose a quantitative study of the evolution of VH, in the broadcast speech since 1940.

For each of harmonizing vowels (V_1) ∈ {o, ɔ, e, ɛ}, the F1 was measured, according to the aperture of word-final full vowel (V_2), in a 10-hour corpus covering six decades of broadcast news.

Our analyses indicate that VH has been significantly present in French since the 1940s. Furthermore, our data suggests that it has experienced since then a decline in the degree of assimilation. We nevertheless tend to consider that the trend shown may be due to epoch-specific acoustic patterns in combination with a significant diachronic change.

Keywords: acoustic phonetics, diachrony, vowel harmony, French mid vowels

1. INTRODUCTION

Mid-vowels in French have long been described as subject to large contextual variations in quality compared to other vowels, due to several factors including a word-level anticipatory assimilation process involving the partial assimilation of mid-vowels in height to the final vowel, as seen in the contrasting pronunciations of the V_1 in words like *coquine* [k**o**kin] and *cocagne* [k**ɔ**kaɲ]. Grammont remains the first linguist to have theorized this phenomenon in 1914 [10], but the first impressionist descriptions date back to mid-19th century [13]. During the 20th century and until the advent of the first empirical studies, the majority of phonetic and phonological treatises point towards the existence of VH as a common characteristic of French whereas, at the same time, they consider it optional and insist on the idea that it would be limited to conversational speech and absent in careful pronunciation [5,7,10,14,18,19]. A terminological distinction between the diachronic process (*dilation*) and the phonological rule (*vowel harmonization*) is also sometimes postulated [7].

Furthermore, impressionist descriptions differ as to the exact domain of VH: for some researchers it affects only the mid front unrounded vowels /e, ɛ/ [7], whereas, for others, both the front unrounded and rounded /ø, œ/ vowels are concerned [18]. The mid back vowels /o, ɔ/ are least mentioned as being subject to VH [19]. From the 60s, which coincides with the era marked by the generativist paradigm, vowel harmony begins to be incorporated into the phonological system of standard French [5,11].

The first empirical study to have attempted a quantification of VH was published in 2003, with the title: "(...) Is there vowel harmony in French?" [6]. An analysis of a well-controlled corpus composed of a hundred tokens presenting a possible VH context, produced by three speakers, made it possible not only to answer affirmatively the question posed in the title, but also observe that the tendency to harmonization is even stronger in the case of the /o, ɔ/ vowels. However, F2 and not F1 was measured, so the study focused on the degree of retraction of the tongue and not the degree of aperture.

More recently, studies by Turco et al. [20, 21] on possible $V_1C(C)V_2$ harmonic contexts in two large corpora of continuous speech have highlighted a significant interaction between the F1 of the V_1 and several factors, such as the presence or absence of an unpronounced underlying schwa (e.g. *souhaiterez* [swe**t**χe]), the place of articulation of intervocalic consonant, graphic interferences, and prosodic prominence, as well as the speech style. The results were surprising in that the study discovered that the degree of VH was significantly higher in the most conservative journalistic speech, compared to the spontaneous speech.

The present paper aims to shed light on the issue of the highly discordant historical descriptions and seeks to investigate the question of whether VH is empirically present in French throughout the period when many linguists have been doubting its existence. We hypothesize not only that (1) it is systematically present in the formal style throughout the 20th century, but also that (2) the extent of VH tends to decrease. The difference in the degree of

harmony observed by Turco et al. [20] between the formal and spontaneous styles would in this case be a reflection of a diachronic change currently in progress.

2. METHOD

2.1. Corpus

The corpus we used was initially designed for the series of studies by Boula de Mareüil et al. [3,4]. It consisted of 160 archival recordings with an average duration of 3 min 20 s and ranging from 20 seconds to 20 minutes, which amounted to about ten hours of speech in total.

Despite significant variability in speaking style and acoustic conditions, the signal-to-noise ratio estimates were not significantly higher for older documents. The potential impact of background noise on the measurements, calculated in terms of percentages of octave jumps between consecutive vowels and percentage of vowels detected as unvoiced, was also found to be low [4].

The few female voices which featured in the journalistic archives have been excluded from the corpus in order to avoid having to normalise vowel measurements.

Initially, the measurements were conducted by examining each decade independently (cf. **Figure 1** below). **Table 1** demonstrates the structure of the corpus in terms of the duration of each subset and the number of occurrences.

Subsequently, we decided to split the diachronic data into three time chunks: 1940-1959, 1960-1979, and 1980-1999 (cf. **Figures 2**, **3** below). By grouping the data into three periods rather than six, it was possible to better distribute the data and to observe more distinct differences between these time chunks.

2.2. Selecting VH contexts for analysis

The archival recordings were segmented into phonemes using an automatic alignment system with extensively trained context-independent models and a corpus-specific pronunciation dictionary, following the method described in Adda-Decker et al. [1].

No constraints were placed on the set of trigger vowels or the length of the word, as long as it followed the VH structure (C)V_1C(C)V_2(C). Therefore, words with a closed syllable containing /ɔ/ or /ɛ/ such as *soldat* or *berline* were not excluded from the analysis.

The V_1, on which VH, under the influence of V_2, was measured, consisted of either back and rounded /o, ɔ/ or front and unrounded /e, ɛ/ vowels, coded either as distinctive underlying phonemes, i.e. close-mid /e, o/ and open-mid /ɛ, ɔ/ vowels, or as segments underspecified for vowel height, which we refer to thereafter as /E, O/. This approach was adopted to avoid any arbitrary phonological labeling in perceptually weak pretonic position [2,8].

The V_2 were then coded according to their degree of aperture *open-mid* or *close-mid* in order to separate closed or close-mid vowels, /i, y, e, ø, o, u, õ/ from open or open-mid vowels /ɛ, ɛ̃, œ, a, ɑ̃, ɔ/.

The values of the first three formants and F0 of each V_1 were extracted in the center of the segment using Praat [9]. F0 was treated as a two-level categorical variable, with its levels of *high* and *low* corresponding to measurements taken above and below the median calculated for each subset.

	length (min)	#words	#VH tokens
1940s	24	3148	93
1950s	53	8029	216
1960s	204	28365	706
1970s	103	15332	386
1980s	64	9667	223
1990s	92	16807	551
total	**540**	**81348**	**2175**

Table 1: Total duration of the data (min), number of words and VH contexts submitted to analysis.

3. RESULTS

3.1. Diachronic evolution of extent of VH

We started by quantifying the degree of VH according to the aperture of each trigger vowel, separately for each decade. The same technique was utilized for each of the target vowels /o, ɔ, O, e, ɛ, E/, however, we will restrict our presentation to the results pertaining to the archiphonemes /O/ and /E/ which are shown in **Figure 1**.

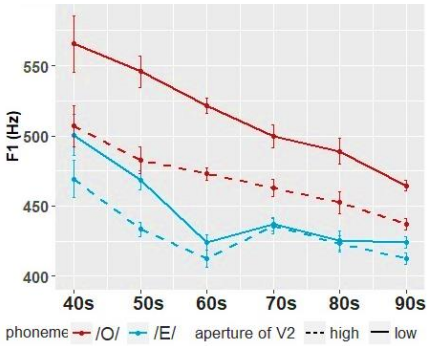


Figure 1: F1 of /O/ and /E/ (mean and St. Err.) according to the V_2 aperture and by decade.

These results come from a two-way ANOVA *F1~aperture_V1+decade*, which revealed a significant effect of the aperture of V_2 and decade on both /O/ and /E/. Furthermore, a significant interaction between these predictors was observed, as demonstrated by a two-way ANOVA interaction *F1~aperture_V1*decade*. The summary statistics are displayed in **Table 2**.

	SS	Df	MS	F	*p*
/O/	aov(F1_V1~aperture_V2+decade)				
aperture_V2	798908	1	798908	104.28	<.001
decade	1143260	5	228652	29.84	<.001
/O/	aov(F1_V1~aperture_V2*decade)				
aperture_V2:decade	135943	5	27189	2.98	<.05
/E/	aov(F1_V1~aperture_V2+decade)				
aperture_V2	69538	1	69538	11.36	<.001
decade	740479	5	148096	24.21	<.001
/E/	aov(F1_V1~aperture_V2*decade)				
aperture_V2:decade	82930	5	16586	2.71	<.05

Table 2: Summary statistics of the effects of the model run on /O/. Formula: *lm(F1_V1 ~ decade * aperture_V2 + decade * f0_height)*

Tukey's Post-Hoc test (α significance level of 0.001) indicated that the F1 differences for V_1 according to the aperture of the V_2 was significant between the following decades, for /O/: 80s-40s, 90s-40s and 90s-50s, and for /E/: 60s-40s and 80s-40s.

Analysis of the data revealed two distinct trends. Throughout the six decades studied, the average value of F1 decreased, indicating more close and centralized mid vowels. Moreover, the degree of VH was also found to diminish between the 40s and 50s and more recent decades.

Nevertheless, we questioned whether there was a genuine diachronic development or if the decrease in the extent of VH was associated with the evolution of speaking style in news presenters over the decades. Indeed, Boula de Mareüil et al. [3,4] demonstrated that mean F0, a prosodic feature which may give an impression of emphatic style, has decreased in French broadcast speech since the 1940s.

3.2. Diachronic evolution of VH with regard to F0

We quantified thus the degree of VH according to F0, *high* vs *low*. In **Figures 2**, **3**, the degree of VH is expressed by the space between the solid and dashed black lines representing average F1 values of V_1 according to the aperture of V_2, without taking into account F0 variability. The solid and dashed coloured lines denote the average variation of F1 of V_1 followed by a *high* or *low* V_2 and, according to F0 height.

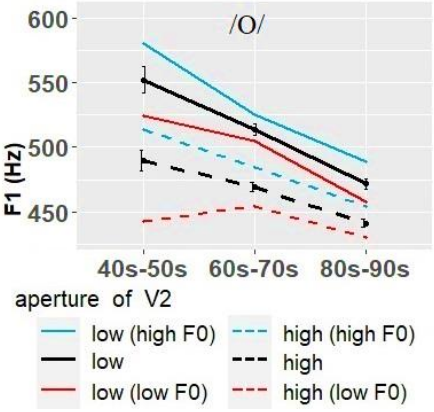


Figure 2: F1 (mean and St. Err.) of V_1 /O/ according to the aperture of V_2; F1 of V_1 according to both aperture of V_2 and the F0 height.

In order to validate these results, we performed a multiple linear regression with interaction: *F1_V1 ~ decade*F0_height + decade*aperture_V2*. The anova() analysis confirmed that the model with interaction was significantly better fitted than the simpler models. The analysis indicated that the three predictors (in two interactions) accounted for 14.6% of the variance (R2=.14, F(8,2064)=44.36, p<.001). Note that the low R^2 values are largely attributable to our inability to account for intra-speaker variability in our corpus.

The results indicated that there was no significant difference in the degree of VH on /O/ between the 40s-50s and the 60s-70s, however, a significant evolution of the degree of VH between the 40s-50s and the 80s-90s was observed. Besides, a significant difference in the F1 value according to the mean pitch was found between the 40s-50s and the 60s-70s, as well as between the 40s-50s and the 80s-90s. The summary statistics of the effects of the model applied to /O/ are presented in **Table 3**.

	Estimate	St. Err	*t*	*p*
40s-50s (Intercept)	491.49	10.88	45.16	<.001
60s-70s	10.95	11.86	0.92	0.355
80s-90s	-31.96	12.06	-2.65	0.008
40s-50s:aperture_V2	-57.44	10.14	-5.66	<.001
60s-70s:aperture_V2	10.98	11.48	0.95	0.339
80s-90s:aperture_V2	26.23	11.93	2.19	0.028
40s-50s:f0_height	84.27	10.87	7.75	<.001
60s-70s:f0_height	-58.21	12.11	-4.80	<.001
80s-90s:f0_height	-56.57	12.58	-4.49	<.001

Table 3: Summary statistics of the effects of the model run on /O/. Formula: *lm(F1_V1 ~ decade * aperture_V2 + decade * f0_height)*.

Subsequently, we modelled the degree of VH on the vowel /E/ according to F0 across three time periods. These results are shown in **Figure 3**.

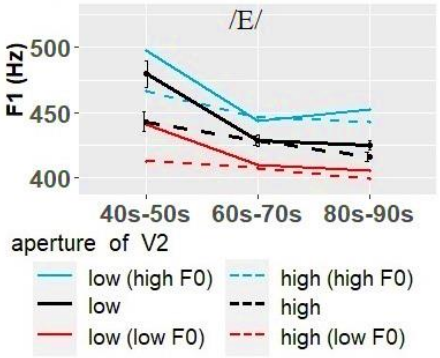


Figure 3: F1 (mean and St. Err) of V_1 /E/ according to the aperture of V_2; F1 of V_1 according to both aperture of V_2 and the F0 height.

A multiple linear regression model with interaction was fitted with the following formula *F1_V1 ~ decade*aperture_V2 + decade*F0_height*. Indeed, the three predictors (in two interactions) explained 12.4% of the variance (R^2=.12, F(8,2104)=37.37, p<.001).

The results indicated a significant evolution in the degree of VH on /E/ between the 40s-50s and the 60s-70s, as well as between the 40s-50s and the 80s-90s. Furthermore, a significant difference in the value of F1 according to F0 was found between the 40s-50s and the 60s-70s, but not between the 40s-50s and the 80s-90s. The summary statistics of the effects of the model applied to /E/ are shown in **Table 4**.

	Estimate	St. Err	*t*	*p*
40s-50s (Intercept)	442.07	7.07	62.48	<.001
60s-70s	-33.80	8.76	-3.85	<.001
80s-90s	-36.05	8.19	-4.40	<.001
40s-50s:aperture_V2	-30.32	7.01	-4.32	<.001
60s-70s:aperture_V2	30.68	9.03	3.39	<.001
80s-90s:aperture_V2	22.55	8.58	2.62	0.008
40s-50s:f0_height	55.30	7.18	7.69	<.001
60s-70s:f0_height	-18.76	9.19	-2.04	0.041
80s-90s:f0_height	-10.13	8.78	-1.15	0.248

Table 4: Summary statistics of the effects of the model run on /E/. Formula: *lm(F1_V1 ~ decade * aperture_V2 + decade * f0_height)*

4. DISCUSSION AND CONCLUSION

The findings of this pilot study validate our first hypothesis (1): VH has been a systematic phenomenon in formal journalistic French since at least 1940, contrary to what most old descriptions postulate.

Additionally, our results suggest that VH reduction is achieved through markedly different patterns for the front unrounded and for the back rounded mid vowels. The evolution of VH effect expressed by average variation of F1 of V_1 is more pronounced for /O/ than for /E/.

Our research also shows that the patterns of mid-close /o/ and mid-open /ɔ/ sounds are slightly different, in that /o/ sounds before a low V_2 have been the type of contexts most affected by the reduction of the extent of vowel harmony over time.

Fagyal et al. [6], in their analysis of a well-controlled corpus, found significant differences in the timbre of the /E/ and /O/ under the influence of the following full vowel, by quantifying the second formant frequencies. They concluded that the tendency for fronting/backing is even stronger for the back rounded mid vowels than for the front unrounded mid vowels. Nguyen & Fagyal [15] highlighted the gradual nature of vowel harmony defined in terms of anteriority/posteriority (F2) and height (F1) modifications of mid vowels.

Turco et al. [21] and our study prioritized vowel openness (F1). However, our results suggest that it is necessary, in order to account for the acoustic differences between front and back mid-vowels, whose affiliation to the cavities is not always clear, to extend the investigation of VH to other parameters, such as, e.g., the difference in amplitude between the first and second formant peaks (A1-A2).

Moreover, our data suggest that, in addition to a decrease in the degree of VH since 1940, which varies in extent depending on the vowel analysed, F0 plays a role in explaining the variability of the degree of VH, potentially in correlation with the so-called emphatic style of journalistic speech in the 1940s and 1950s. The role of F0 warrants further investigation in future research. Other parameters that were not taken into account, such as intra- and inter-speaker variability, lexical and prosodic context, could offer valuable insights into F0 variability and its importance in modeling vowel harmony patterns diachronically. Our corpus did not afford us the opportunity to consider inter-speaker variability, which would have likely enhanced the explanatory power of our statistical model.

Overall, we are inclined to conclude that our second hypothesis (2) postulating a decrease in the degree of VH is supported, although further, more diversified examinations are required to confirm it. The degree of VH has been demonstrated to be significantly lower in more recent decades, which is in line with the findings of Turco et al. [21] indicating that harmony is more prevalent in more conservative journalistic speech, when compared to casual speech. These results are thus consistent with our hypothesis, in that “many diachronic phonological changes mirror synchronic phonetic variations” [22, 16].

The non-discrete nature of VH in French, as well as its sensitivity to a number of parameters, leads us to consider whether we are detecting a genuine phonological sound change or a language-specific phonetic detail that has developed over time. Speaking of VH in French in terms of a phonological process is debatable; indeed, our empirical results do not provide us with sufficient evidence to consider VH, and even less its diachronic evolution, as a functional phonological process that is likely to be modelled as a all-or-nothing change.

Nevertheless, the observed phenomenon can be incorporated into Ohala's paradigm of phonetic sound change [16], which posits several, mainly perceptual, modes of explanation with the aim of modeling an acoustic diachronic change at the subphonemic level.

Indeed, future research should address the extent to which VH can be considered a phonological phenomenon and not a simple case of distant assimilation. A perception experiment could be utilized to assess the extent to which French speakers cognitively exploit VH.

5. ACKNOWLEDGEMENT

This work was partially supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

6. REFERENCES

[1] Adda-Decker, M., Boula de Mareüil, P., Adda, G., & Lamel, L. (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, *46*(2), 119–139.

[2] Boersma, P. (2015). French high-mid vowels are underspecified for height. In: *Proc. of the 18th ICPhS*.

[3] Boula de Mareüil, P., Rilliard, A., & Allauzen, A. (2008). Étude diachronique de l'accent initial au travers d'archives audio. In: *Actes des 27e Journées d'Étude sur la Parole JEP*, pp. 1621-1624.

[4] Boula de Mareüil, P., Rilliard, A., & Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style : Corpus-based Measurements and Perceptual Experiments. *Language and Speech*, *55*(2), 263-293.

[5] Dell, F. (1973). *Les règles et les sons : Introduction à la phonologie générative*. Hermann.

[6] Fagyal, Z., Nguyen, N., & Boula de Mareüil, P. (2003). From dilation to coarticulation : Is there vowel harmony in French? *Studies in Linguistic Sciences*, *32*, 1-21.

[7] Fouché, P. (1956). *Traité de prononciation française*. C. Klincksieck.

[8] Gendrot, C., & Audibert, N. (2019). La distinction /e/ vs /ɛ/ en français standard est-elle maintenue en finale de mot ? Étude sur des corpus de parole journalistique et de parole spontanée. *Langue française*, *203*(3), 53-65.

[9] Gendrot, C., & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels : An automatic analysis of large broadcast news corpora in French and German. In: *Proc. of Interspeech*, pp. 2453-2456.

[10] Grammont, M. (1914). *Traité de phonétique : Avec 179 figures dans le texte*. Delagrave.

[11] Léon, P. (1966). *Prononciation du français standard : Aide-mémoire d'orthoépie*. Librairie Marcel Didier.

[12] Lesaint, M. A. (1850). *Traité complet et méthodique de la prononciation française*. Perthes-Besser et Mauke.

[13] Liénard, J. S., & Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America*, *106*(1), 411-422.

[14] Martinet, A. (1945). *La Prononciation du français contemporain, témoignages recueillis en 1941 dans un camp d'officiers prisonniers*. Droz.

[15] Nguyen, N., & Fagyal, Z. (2008). Acoustic aspects of vowel harmony in French. *Journal of Phonetics*, *36*, 1-27.

[16] Ohala, J., & Jones, C. (1993). The Phonetics of Sound Change. In *Historical Linguistics : Problems and Perspectives*, 237-278.

[17] Ohman, S. E. (1966). Coarticulation in VCV utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151-168.

[18] Straka, G. (1950). Système des voyelles du français moderne. *Bulletin de la Faculté des lettres de Strasbourg*, 220-233.

[19] Tranel, B. (1987). *The Sounds of French : An Introduction*. Cambridge University Press.

[20] Turco, G., Fougeron, C., & Audibert, N. (2016a). Que nous apprennent les gros corpus sur l'harmonie vocalique en français ? In: *Actes des 31e Journées d'Etudes sur la Parole JEP*, 1, pp. 571-579.

[21] Turco, G., Fougeron, C., & Audibert, N. (2016b). The Effects of Prosody on French V-to-V Coarticulation : A Corpus-Based Study. In: *Proc. of Interspeech*, pp. 998-1001.

[22] Vasilescu, I., & Lamel, L. (2019). Synchronic variation and sound change in Romance languages : A corpus-based study of lenition phenomena in Romanian and Spanish. *49th Linguistics Symposium on Romance Languages*. Athens. GA.

VOCAL EFFORT IN HUMAN INTERACTIONS WITH VOICE-AI

Leah Bradshaw[1], Valeriia Perepelytsia[1] and Volker Dellwo[1]
[1]Department of Computational Linguistics, University of Zurich, Switzerland
leah.bradshaw@uzh.ch

ABSTRACT

Speech adaptations occur frequently in the presence of perceived communication barriers. Modern technological advancements have brought with them new interlocutors for human speakers with the introduction of voice-AI assistants. Findings have shown that voice-AI-directed speech is characterised by an increase in vocal effort resulting from the presumed capabilities of these systems for understanding speech. However, studies focus solely on voice-AI assistants which perform speech recognition. In this study, we present an acoustic analysis of speaker interactions with two voice-AI systems with different goals (speech interpretation vs. speaker verification). Using f_0 mean and range as acoustic correlates of vocal effort, we found that speakers show some evidence of increased vocal effort towards voice-AI systems regardless of final task, however, this is enhanced by speech intelligibility goals. This finding is interpreted to suggest that voice-AI-directed speech globally exhibits increased vocal effort, but task plays a clear role in the extent of this.
(149 words)

Keywords: speech communication, vocal effort, voice-AI, speaker accommodation

1. INTRODUCTION

Speech in challenging conditions, i.e., in noisy environments or towards less-proficient interlocutors, is frequently observed to be produced with increased vocal effort, characterised by increased fundamental frequency (f_0) mean and range [e.g., 1-7]. The introduction of voice activated artificially intelligent, or voice-AI, assistants has created a new kind of challenging interlocutor, which speakers assume to require more effortful talk [7–11]. However, human interactions with these systems have thus far only been explored using speech recognition technologies. Contrastively, voice-AI systems can perform numerous tasks which do not process speech in the same way, i.e., speaker verification. It remains to be explored how speakers interact more globally with voice-AI systems and how the specific task plays a role in the amount of vocal effort employed.

In this study, we investigated if speakers employ differing vocal effort towards voice-AI systems according to task. We designed a Wizard of Oz experiment [e.g., 12–14] in which participants interacted with two different mock voice-AI systems; a Speech Recogniser, which aimed to understand their speech; and a Speaker Recogniser, which aimed to identify them from their voice. We present here the findings of an acoustic analysis which explores two acoustic correlates of vocal effort (f_0 mean and range), to examine differences in how speakers interact with the two systems.

1.1. Vocal effort in speech communication

Increased vocal effort occurs frequently in challenging communicative environments, for example, when speakers shift from a neutral speech to shouted speech in the presence of noise [e.g., 1–5] or when the distance between interlocutors is increased [e.g., 6, 15]. One common acoustic correlate associated with this phenomenon in the aforementioned studies, is increased f_0 mean and range. Vocal adaptations corresponding to increases in f_0 mean and range also occur in the presence of challenging interlocutors or to enhance speech intelligibility, i.e., towards children [16–17] or voice-AI systems [7], and are attributed in part to increased vocal effort. However, in these instances speech intelligibility requirements are confounded with challenging interlocutors, thus the influence of each individual factor on vocal effort cannot be isolated.

1.2. Voice-AI-directed speech

Studies have shown numerous interlocutors require speech adaptations for efficient communication. Children, hearing-impaired listeners, and, more recently, machines all pose intelligibility challenges for speakers [16–21]. The introduction of the voice-AI assistant has commanded a new challenging machine interlocutor for humans, which invokes a speech style that is louder [22–23] and contains different vowel formant characteristics [23] compared to human-directed speech. Further, findings have shown increases in f_0 mean and range in voice-AI-directed speech [22] and have attributed these in part to an increase in vocal effort in these interactions [7].

Thus far, voice-AI-directed speech has solely considered interactions with speech recognition technologies, such as Amazon's Alexa or Apple's

Siri, which require interpretation of speech input. Therefore, it is plausible that the previously observed voice-AI-directed speech characteristics are a by-product of the systems requirement for speech intelligibility. However, in fields such as banking, there is an increasing use of voice-AI systems for voice identification/verification purposes, a task which considers the quality of the voice rather than the content of the speech. As such, the question arises whether speakers would employ the same vocal adaptations towards these kinds of voice-AI systems which offer a similarly challenging interlocutor, but without the necessity for speech intelligibility.

1.3. Research hypotheses

This study investigates vocal effort in interactions with two different voice-AI systems, a Speech Recogniser, and a Speaker Recogniser. Previous findings show speakers employ greater vocal effort, measured by an increase in f_0 mean and range, in interactions with speech recogniser voice-AI systems. However, the extent to which this vocal effort relates to the task of speech understanding or to the generally challenging interlocutor remains unknown. We predict that, if vocal effort is employed in voice-AI interactions due to the perceived communicative barriers of the system, vocal effort will be similar in both interactions. However, if speech intelligibility challenges necessitate enhanced vocal effort, we expect lesser vocal effort will be employed towards the Speaker Recogniser.

2. METHODS

2.1. Data collection

39 Swiss German speakers (20 F) completed a Wizard of Oz experiment, an expansion on the experiment design introduced in [24], designed using Gorilla Experiment Builder [25]. Participants spoke to two mock voice-AI systems; a mock speaker recognition system (SpkrRec) with a male text-to-speech voice, Verifico, which aimed to recognise the individual from their voice input; and a mock speech recognition system (SpchRec) with a female text-to-speech voice, Vicky, whose task was to interpret their speech. Participants spoke 34 different prompts to each of the systems which then offered either an immediate correct response, or a misrecognition. Each trial was repeated until the "system" was successful. Responses were assigned randomly by the experiment software, but participants would only ever receive a maximum of misrecognitions per prompt. All speakers spoke Swiss Standard German throughout the experiment [26]. Full details of the experimental procedure can be found in [27].

Prior to the interactions, speakers were recorded reading a list of the same 34 sentences used as prompts in the interaction tasks which were used as baseline productions for each speaker.

2.2. Data analysis

Analysis was completed using only first productions of a sentence, resulting in 102 sentences per speaker. Repetitions of sentences following an error response were excluded so that all sentences were produced following positive feedback (a recognition) or no feedback in the case of the read speech condition.

Acoustic measurements were automatically extracted over each target sentence in each of the three tasks using a Praat script [28]. Fundamental frequency (f_0) minimum and maximum values were extracted for each sentence in logHz, to calculate the f_0 range ($f_{0max} - f_{0min}$). Mean f_0 was calculated at 15 equidistant intervals across each sentence, also in logHz, and averaged to calculate the mean f_0 value. All measures were taken using gender-specific pitch ranges (50-200Hz for males; 75-400Hz for females). All measures were also z-scored prior to statistical analysis to account for individual speaker differences.

2.3. Statistics

Each acoustic measure was subjected to statistical analysis to explore the differences between interactions with each voice-AI system. Mean f_0 and f_0 range were modelled in separate linear mixed effects models with the *lme4* R package [29], with identical model structure: fixed effects of *Task* (SpkrRec, SpchRec, Read), *Gender* (Male, Female), plus by-Sentence and by-Speaker random intercepts.

3. RESULTS

3.1. Mean f_0

Figure 1 contains the distribution of mean f_0 for speakers' productions of each sentence across the three tasks. Model output showed a statistically significant increase in mean f_0 in both tasks, compared to the Reading task (SpchRec: $\beta = 0.553$, SpkrRec: $\beta = 0.1738$, both $p < 0.0001$). The main effect of Gender failed to reach significance. Post-hoc tests to assess the directionality of this relationship were conducted in the form of pairwise t-test comparisons with Bonferroni correction for multiple hypothesis testing. Findings confirmed a statistically significant difference between f_0 mean in the SpchRec and SpkrRec tasks compared to the Reading task, as well as a statistically significant difference between the two voice-AI tasks (all $p < 0.001$).

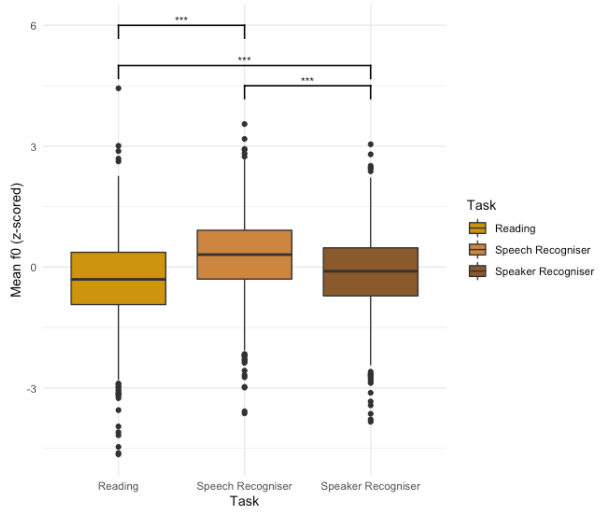


Figure 1: Boxplots showing the median, range, and inter-quartile range of the distribution of scaled mean f_0 values for speakers' production of each utterance by task.

3.2. f_0 range

Figure 2 shows the distribution of f_0 range for speakers' productions of each sentence in each of the three tasks. The model output showed a statistically significant increase in f_0 range in the SpchRec task, compared to the Reading task ($\beta = 0.0895$, $p = 0.002$). However, the difference between the SpkrRec and Reading tasks was not statistically significant. The main effect of Gender also failed to reach significance. Post-hoc analysis of f_0 range by Task was again completed using pairwise t-test comparisons with Bonferroni correction. The output showed a statistically significant difference between the SpchRec and the Reading task ($p < 0.001$), but not between the Reading and SpkrRec tasks ($p = 0.5396$), or the SpchRec and SpkrRec tasks ($p = 0.2573$).

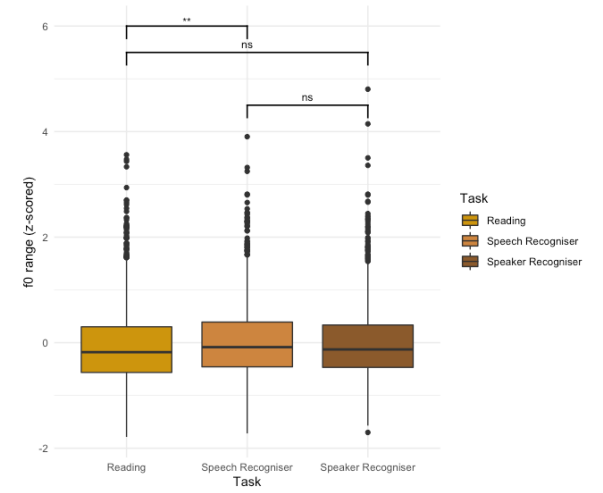


Figure 2: Boxplots showing the median, range, and inter-quartile range for the distribution of scaled f_0 range values for speakers' produced utterances by task.

3.3. Accommodation to voice-AI

We also tested for a potential confound in the finding that mean f_0 was substantially higher in the SpchRec task. Given the voice in this task (Vicky) was a female voice, compared to the male voice (Verifico) in the SpkrRec task, speaker accommodation towards the voice-AI system would increase in average speaker mean f_0, which may account for this finding.

To explore this, we calculated two distances. The distance between the speaker's averaged mean f_0 and the mean f_0 of each voice-AI system *prior* to and *during* the interaction. Mean f_0 for each voice-AI was calculated in the same way as previously stated, and speaker values were averaged across each task. The measures were as follows:

Distance 1: Euclidean distance between the mean f_0 for Vicky/Verifico and each speakers' pre-interaction productions (Reading task) (voice-AI mean f_0 – participants' mean $f_{0\,[Read]}$)

Distance 2: Euclidean distance between the mean f_0 for Vicky/Verifico and each speakers' production in the corresponding interaction task (Vicky mean f_0 – participants' mean $f_{0\,[SpchRec]}$ | Verifico mean f_0 – participants' mean $f_{0\,[SpkrRec]}$)

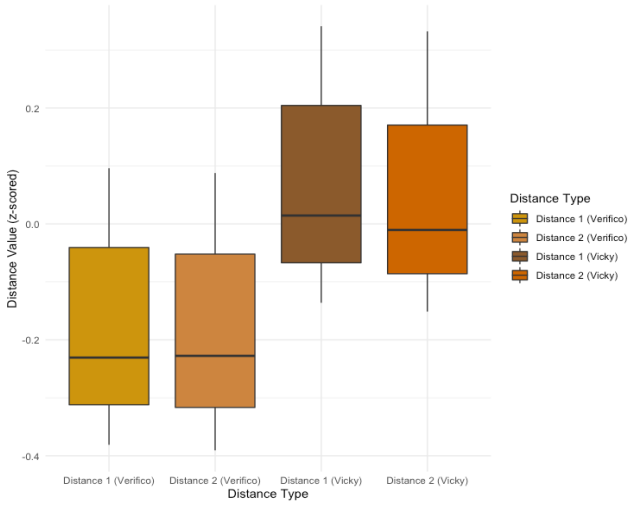


Figure 3: Boxplots showing medians, ranges, and inter-quartile ranges for scaled distances between averaged mean f_0 for each speaker and mean f_0 of the voice-AI systems, both prior and during the interaction.

Figure 3 shows the comparison between *Distance 1* and *Distance 2* for each voice-AI system. We ran a two-way repeated measures ANOVA to test for an interaction between *Time* (e.g., prior or during the interaction with the voice-AI) and *System* (e.g., Vicky or Verifico). The outcome was statistically significant ($p = 0.0067$), suggesting both factors influenced the difference between the distances. To assess accommodation towards each voice-AI system, we conducted further pairwise t-test with Bonferroni correction for repeated measures. Findings showed

that the difference between *Distance 1* and *2* was significant for both voice-AI systems (Vicky: p = 0.000196, Verifico: p = 0.032).

4. DISCUSSION

Overall, we find evidence to suggest that increased vocal effort, measured using f_0 mean and range, is a general characteristic of voice-AI-directed speech. However, we observe differences in the extent of this depending on the task. Namely, we observed a significant increase in both f_0 mean and range in the SpchRec task, but only an increase in f_0 mean in the SpkrRec task. Further, the increase in f_0 mean in the SpchRec task was significantly larger than that observed in the SpkrRec task. Therefore, it is plausible that, while human interactions with voice-AI systems can be generally associated with enhancements in vocal effort, the necessity for speech intelligibility further enhances the amount of vocal effort in speech.

The lesser vocal effort observed in the interactions with the SpkrRec system could be attributed to the lack of requirement for speech intelligibility. Previous findings of increased vocal effort towards challenging interlocutors are confounded with speech intelligibility challenges [e.g., 7, 16–17]. For instance, increased vocal effort observed in speech towards children could arise from the fact that the interlocutor is less proficient, or the desire for speech interpretation. Therefore, the vocal effort exhibited towards the SpkrRec could represent that induced solely by the presence of a challenging interlocutor.

Contrastively, it may be that speakers did not deem the SpkrRec to be as challenging an interlocutor as the SpchRec, due to a lack of prior experience with the system. From post-experiment discussions, it was clear that participants were already familiar with Speech Recognition systems, but not Speaker Recognition systems. Therefore, the greater vocal effort could correspond to prior assumptions that the SpchRec system would be a challenging interlocutor, compared to a lack of prior assumptions for the SpkrRec.

A potential confounding with these findings is the comparison with the monologue speech in the Reading task. Although it is likely that a voice-AI interlocutor would compel more effortful speech, it is possible that this slight increase in vocal effort we observe is due to the presence of the interlocutor in the SpkrRec task. Further research is necessary to explore to the effect of interlocutor presence or absence.

Further consideration was given to possible speaker accommodation towards the voice-AI system which may have commanded the higher f_0 mean in the SpchRec task. However, we observed significant convergence towards both systems and thus cannot solely attribute the increased mean f_0 to speaker accommodation. Further, it is equally not unexpected that participants would converge towards the systems. Previous findings show speaker accommodation is a frequently employed technique for enhancing intelligibility [e.g., 30-31], and speakers have previously exhibited phonetic and prosodic alignment towards machine interlocutors [e.g., 32-34].

Finally, we note that f_0 mean and range only offer an introductory look at vocal effort differences in these two speech styles. Additional research using a more comprehensive set of measures, including for instance, speech intensity/loudness, is necessary to fully examine these differences.

Overall, the above represents a multitude of plausible interpretations for these findings given previous ideas of speech communication and vocal effort. We observed clear differences in speaker f_0 mean and range in interactions with two different kinds of voice-AI systems, however, which factor constitutes to this finding is unclear. Nonetheless, it is conceivable that speakers globally adopt more effortful talk when speaking to voice-AI systems, which is enhanced by speech intelligibility requirements.

5. CONCLUSION

The present study investigated vocal effort in interactions with two voice-AI systems which perform different tasks. Findings showed a tendency for increased vocal effort, captured using f_0 mean and range, in interactions with both of the voice-AI systems compared to the Reading task. However, substantially greater vocal effort was employed when speech intelligibility was required. These findings are interpreted to suggest that voice-AI-directed speech globally exhibits increased vocal effort, but the extent to which this occurs can differ depending on the task. Future research should consider further factors influencing vocal effort, namely, the goal of the interaction and the presence of an interlocutor. Further discussions about the different kinds of voice-AI systems and human interactions with them, would also be beneficial.

6. ACKNOWLEDGMENTS

This research was supported by the Swiss National Science Foundation in the project #185399 “The dynamics of indexical information in speech and its role in speech communication and speaker recognition.”

7. REFERENCES

[1] M. Jessen, O. Koster, and S. Gfroerer, "Influence of vocal effort on average and variability of fundamental frequency," *IJSLL*, vol. 12, no. 2, Art. no. 2, Aug. 2005.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *JASA*, vol. 84, no. 3, pp. 917–928, Sep. 1988.

[3] Z. S. Bond, T. J. Moore, and B. Gable, "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *JASA*, vol. 85, no. 2, pp. 907–912, Feb. 1989.

[4] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[5] S. Koster, "Acoustic-phonetic aspects of Lombard Speech for different text styles," *The Phonetician*, vol. 85, pp. 9–16, 2002.

[6] P. Bottalico, "Speech Adjustments for Room Acoustics and Their Effects on Vocal Effort," *Journal of Voice*, vol. 31, no. 3, p. 392.e1-392.e12, May 2017.

[7] M. Cohn and G. Zellou, "Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments," *Frontiers in Communication*, vol. 6, 2021.

[8] H. H. Clark and G. L. Murphy, "Audience Design in Meaning and Reference," in *Advances in Psychology*, vol. 9, J.-F. Le Ny and W. Kintsch, Eds. North-Holland, 1982, pp. 287–299.

[9] S. Oviatt, G.-A. Levow, E. Moreton, and M. MacEachern, "Modeling global and focal hyperarticulation during human–computer error resolution," *JASA*, vol. 104, no. 5, pp. 3080–3098, Nov. 1998.

[10] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, and A. Brown, "The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers," *Cognition*, vol. 121, no. 1, pp. 41–57, Oct. 2011.

[11] B. R. Cowan, H. P. Branigan, M. Obregón, E. Bugis, and R. Beale, "Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human−computer dialogue," *Int. J. Hum. Comput. Stud.* vol. 83, pp. 27–42, Nov. 2015.

[12] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies — why and how," *Knowledge-Based Systems*, vol. 6, no. 4, pp. 258–266, Dec. 1993.

[13] D. DeVault, J. Mell, and J. Gratch, "Toward Natural Turn-Taking in a Virtual Human Negotiation Agent," in *2015 AAAI Spring Symposium Series*, Mar. 2015.

[14] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *ACM Trans. Comput.-Hum. Interact.*, vol. 11, no. 3, pp. 300–328, Sep. 2004.

[15] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *JASA*, vol. 106, no. 1, pp. 411–422, Jul. 1999.

[16] A. R. Bradlow, N. Kraus, and E. Hayes, "Speaking Clearly for Children With Learning Disabilities," *JSLHR*, vol. 46, no. 1, pp. 80–97, Feb. 2003.

[17] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech," *Speech Commun.*, vol. 49, no. 1, pp. 2–7, Jan. 2007.

[18] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing II," *JSLHR*, vol. 29, no. 4, pp. 434–446, Dec. 1986.

[19] R. Lunsford, S. Oviatt, and A. M. Arthur, "Toward open-microphone engagement for multiparty interactions," in *Proceedings of the 8th international conference on Multimodal interfaces - ICMI '06*, Banff, Alberta, Canada, 2006, p. 273.

[20] D. Burnham, S. Joeffry, and L. Rice, "Computer-and human-directed speech before and after correction," *space*, vol. 6, p. 7, 2010.

[21] C. Mayo, V. Aubanel, and M. Cooke, "Effect of prosodic changes on speech intelligibility," in *Proceedings of INTERSPEECH*, Portland Oregon USA, Sept. 2012.

[22] E. Raveh, I. Steiner, I. Siegert, I. Gessinger, and B. Möbius, "Comparing phonetic changes in computer-directed and human-directed speech," *Elektronische Sprachsignalverarbeitung 2019*, pp. 42–49, 2019.

[23] I. Siegert and J. Krüger, "'Speech Melody and Speech Content Didn't Fit Together'—Differences in Speech Behavior for Device Directed and Human Directed Interactions," in *Advances in Data Science: Methodologies and Applications*, Springer, 2021, pp. 65–95.

[24] V. Dellwo, E. Pellegrino, L. He, and T. Kathiresan, 'The dynamics of indexical information in speech: Can recognizability be controlled by the speaker?', *AUC Philol.*, vol. 2019, no. 2, pp. 57–75, Oct. 2019.

[25] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder', *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020.

[26] I. Hove, *Die Aussprache der Standardsprache in der deutschen Schweiz*. DE GRUYTER, 2002.

[27] V. Perepelytsia, L. Bradshaw and V. Dellwo, "IDEAR: A speech database of identity-marked, clear and read speech,", *Proceedings of ICPhS*, 2023.

[28] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2022. [Online].

[29] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software, 67(1), 1-48, 2015.

[30] R. L. Street and H. Giles, "Speech accommodation theory: A social cognitive approach to language and speech behavior," *Social cognition and communication*, vol. 193226, pp. 193–226, 1982.

[31] C. Gallois, T. Ogay, and H. Giles, "Communication Accommodation Theory," *Communication Accommodation Theory*, p. 28.

[32] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS*, 2003, vol. 3, pp. 833–836.

[33] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connection Science*, vol. 19, no. 2, pp. 131–141, Jun. 2007.

[34] I. Gessinger, B. Möbius, N. Fakhar, E. Raveh, and I. Steiner, 'A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction', presented at *ICPhS*, Melbourne, 2019, pp. 1475–1479.

Acoustic convergence to an unfamiliar language and an unfamiliar accent in shadowed speech

Amy E. Hutchinson[1], Alexis Zhou[2], Yuhyeon Seo[2], and Olga Dmitrieva[2]

[1]Boston University, [2]Purdue University
ahut@bu.edu, atews@purdue.edu, seo86@purdue.edu, odmitrie@purdue.edu

ABSTRACT

The present study investigated the propensity of native speakers of English (two groups of 15 participants each) to spontaneously imitate a model who spoke a different language (Russian) or an accented version of their own language (Russian-accented English). The study consisted of four phases: baseline, exposure, shadowing, and post-test. The change in the voice onset time (VOT) of word-initial voiceless stops from the baseline to shadowing to post-test was assessed across the two groups. The results showed that participants in the Russian condition significantly shortened the VOT of their voiceless stops during the shadowing phase, indicating convergence towards the native speaker model. The values, however, returned to baseline levels in the post-test, indicating that imitation of the phonetic properties of Russian speech did not extend to participants' native English. Participants in the accented English condition did not converge towards the shorter VOT of the model during shadowing or post-test phases.

Keywords: Acoustic convergence, Russian, English, accented speech, shadowing

1. INTRODUCTION

Previous research has demonstrated the ability of talkers to spontaneously adjust the acoustic properties of their speech to sound more like their interlocutor or speech model. This phenomenon is known as phonetic convergence or phonetic accommodation and has been shown to affect a wide variety of acoustic properties including segmental properties, such as voice onset time (VOT) and vowel quality [1]–[5], as well as suprasegmental properties, such as intonation contour and voice quality [4], [6]–[8]. A majority of previous work examined acoustic convergence between two native speakers of the same language, while less is known about convergence to accented speech and/or unfamiliar languages. The present study aims to address this gap by investigating acoustic convergence during shadowing of an unfamiliar accent and an unfamiliar language.

Acoustic convergence is often approached within the framework of the Communication Accommodation Theory (CAT) [9], which postulates that talkers accommodate other talkers in verbal communication in order to reduce the social distance between themselves and their interlocuter. The present study expands the assumptions of the CAT beyond a native language and/or monolingual setting.

Additionally, by examining acoustic convergence to another language or to accented speech, we can begin to assess the potential contribution of long-term convergence to phonetic changes in the speaker's native language (phonetic drift) [10], [11]. For example, could accommodation to a new language or accented speech, such as that of a language instructor, contribute to phonetic drift in a learner's native language (L1)?

While L1 drift in individual second language (L2) learners is believed to be reversal and temporary, it has also been proposed that prolonged contact with accented versions of a language could, over time, lead to permanent sound change in the language itself [12]. This process could also be construed as L1 phonetic drift, but on a larger, societal scale. Assuming that inter-speaker accommodation would be the driving force behind such sound changes, the present study can help us evaluate the viability of this proposal.

In order to compare acoustic convergence to an unfamiliar language and an unfamiliar accent in shadowed speech, the present study examined word-initial voiceless stop shadowing by native speakers of American English who believed they were shadowing either Russian speech or Russian-accented English. Stop voicing was assessed via VOT.

VOT is the primary correlate of voicing in both English and Russian, but different VOT values are used to express phonologically equivalent categories in the two languages. English realizes voiceless stops in word-initial position as voiceless aspirated (long lag VOT >30ms), while Russian realizes voiceless stops as voiceless unaspirated (short lag VOT <30ms) [13], [14]. Therefore, if English speaking participants were to converge towards Russian or Russian-accented speech, we would expect to observe shorter VOT values for their voiceless stops.

2. METHODS

2.1. Participants

A total of 30 native speakers of American English have participated in the current study thus far (22 female and 8 male; mean age 26.63 years)[1]. Participants were self-selected volunteers recruited via flyers on the campus of a large Midwestern university. Twenty-six of the participants indicated low proficiency in at least one language other than English, with Spanish being the most prevalent (16 participants) distantly followed by French (6 participants). No participants indicated that they had knowledge of Russian.

2.2. Stimuli

The present study included three separate sets of stimuli: a baseline set, an exposure/shadowing set, and a post-test set. Each set contained eight CVC target words and 23 CVC fillers. Vowels used in the targets were limited to /ɑ/, /i/, and /ʌ/ and their closest Russian counterparts: /o/, /i/, and /a/. Across stimuli sets, the onset consonants in target words were voiceless stops: /p, t, k/.

The baseline list (Set A) consisted of English target words *pall, pod, posh, pup, tin, top, kill,* and *cod*. The post-test list (Set C) contained English target words *pot, pond, pox, pug, tip, tot, kiss,* and *cob*. The exposure/shadowing stimuli (Set B) was a set of eight English-Russian near-homophones that were recorded as Russian words by a female native speaker of Russian (Table 1). The exposure/shadowing list also contained 23 fillers that were different depending on condition. Fillers in the Russian condition were real Russian words, while fillers in Russian-accented English condition were perceived as English words produced with Russian accent.

	English	Russian	IPA
1	pot	пот (sweat)	[pot]
2	poll	пол (gender)	[pol]
3	pop	поп (priest)	[pop]
4	pun	Пан (Pan)	[pan]
5	tick	тик (teak)	[tik]
6	tock	ток (electric current)	[tok]
7	call	кол (stake)	[kol]
8	kit	кит (whale)	[kit]

Table 1: Set B: English-Russian near-homophones. IPA is representative of the model talker's pronunciation of each word (presented as either Russian or Russian-accented English to participants).

2.3. Procedures

The experiment consisted of four phases: baseline, exposure, shadowing, and post-test. During the baseline phase, all participants completed a word list reading task (baseline Set A) during which 8 English targets and 23 fillers were recorded twice. Items were presented one by one in a randomized order on a screen and participants were instructed to read each word in their normal speaking voice. Items were presented in two blocks and there was a 500 ms ISI between words.

Following the baseline recording, participants were randomly assigned to one of two conditions for the exposure and shadowing portions of the experiment. While the target items were acoustically identical in each condition (English-Russian near-homophones; exposure/shadowing Set B), prior to the start of the exposure phase, participants were explicitly informed that they would hear Russian-accented English words or Russian words. During the exposure phase, participants listened to 8 target items and 23 fillers produced by the model, presented in random order, while images representative of the word meanings were simultaneously displayed. Following the playback of each item, participants responded to a simple multiple-choice question about the sounds they heard (e.g., *what was the first sound of the word?*) by pressing a key on a keyboard. The exposure phase was designed to prepare participants for the shadowing phase, by familiarizing them with the model talker and the target words that they would be producing. Pictures were used to make sure that participants in the accented English condition understood the words that they were to shadow despite the accented pronunciation. For participants in the Russian condition pictures were used to demonstrate that the sequences of sounds they heard in Russian were real lexical items (i.e., had meaning). Questions were used to ensure that the participants maintained focus and to direct their attention to the sound structure of the items that they heard.

In the shadowing phase, participants were aurally presented the same set of words as in the exposure phase (Set B), in a different randomized order and were given 2,000 ms to repeat each word into a microphone as accurately as they could. Each item was presented twice over two blocks.

Following the shadowing portion of the study, participants completed a post-test word-list reading task that was identical in structure to the baseline phase but contained different lexical items (Set C). The post-test phase was conducted to test whether potential acoustic convergence in the shadowing phase persisted over time and generalized to novel English words (that is, words not presented during

exposure/shadowing phase). In addition to generalization, post-test words differed from the baseline words, as previous research has noted that talkers are less likely to converge in items that they already pronounced [15].

Recordings were made in a sound-attenuated booth using an ART Tube MP Project Series preamplifier and a Shure KSM32 Embossed Single-Diaphragm microphone. Audio was captured using Audacity version 2.3.2 at a 44.1 sampling rate. The experimental interface was created using Psycho-Py [16]. All instructions (recorded by the native Russian talker) and exposure/shadowing stimuli were presented to participants via Sennheiser HD 380 Pro headphones.

2.4. Analysis

Word-initial VOT in the target words in the baseline, shadowing, and post-test recordings were measured using Dr. VOT – a machine-learning system based on a recurrent neural network [17]. The pre-trained program automatically measured the time between the release of a stop and the onset of periodic vibration of the following vowel. Each measurement was manually checked for errors prior to statistical analyses and any instances of pre-voicing/negative VOT were excluded from the analysis (8 total tokens; .06% of the total data set). A total of 1,395 tokens (Accented English: 691 tokens; Russian: 704 tokens) were submitted to a mixed-effects linear regression model with voiceless VOT duration as a dependent variable, using R (V. 4.2.1) [18] and the 'lme4' package [19]. The fixed effects of the model included Condition (English: reference, Russian), Phase (Baseline: reference, Shadowing, Post), and their interaction. Additionally, as many participants had knowledge of a foreign language that realized voicing in the same way as Russian (i.e., 'voicing languages' like Spanish, French, etc.), the model included a categorical fixed effect of Voicing Lang Knowledge (No: reference, Yes) and its interaction with Condition. The model also included by-subject random intercepts and slopes for subject by Phase, as well as by-word random intercepts.

3. RESULTS

3.1 Model talker acoustics

The model talker's voiceless VOT duration was first measured in order to confirm that it was within native-like expectations for Russian voiceless stops. On average, the model talker produced their voiceless stops with an average VOT duration of 3.42 ms, which is in line with previous literature on Russian stop voicing duration [14].

3.2 Voiceless stop shadowing by participants

Table 2 summarizes the fixed effects of the mixed-effects linear regression model. In the model, the intercept equals the estimated mean of English VOT produced by the accented English group (β_0 = 89.1, SE = 7.51). The statistical results showed a significant interaction effect of Condition with Phase. Specifically, participants significantly reduced their VOT durations by 20.4 ms on average when shadowing speech they believed to be Russian, compared to the baseline phase (β = -20.4, SE =8.69, t = -2.35, p = .03). A following post-hoc analysis using a pairwise coefficient Tukey adjustment additionally revealed that VOT in the shadowing phase also significantly differed from VOT in the post-test phase (β = 24.20, SE =7.53, t = 3.22, p = .03). By contrast, VOT values were not significantly different between the baseline and post-test phases (p = .39) in the Russian condition. The pairwise contrasts are also visually confirmed by the right panel of Figure 1 that demonstrates the raw VOT values.

Effects	Estimate	*t*	*p*
Intercept	**89.057**	**11.86**	**< .001**
ConditionRussian	4.576	.36	.72
PhaseShadowing	-3.766	-.52	.61
PhasePost	6.68	1.04	.30
VoicingLangKnow	**-20.567**	**-3.31**	**.003**
ConditionRussian:PhaseShadowing	**-20.398**	**-2.34**	**.02**
ConditionRussian:PhasePost	-3.043	-.86	.40
ConditionRussian: VoicingLangKnow	-1.649	-.13	.90

Table 2: Fixed-effects of the mixed-effects linear regression model.

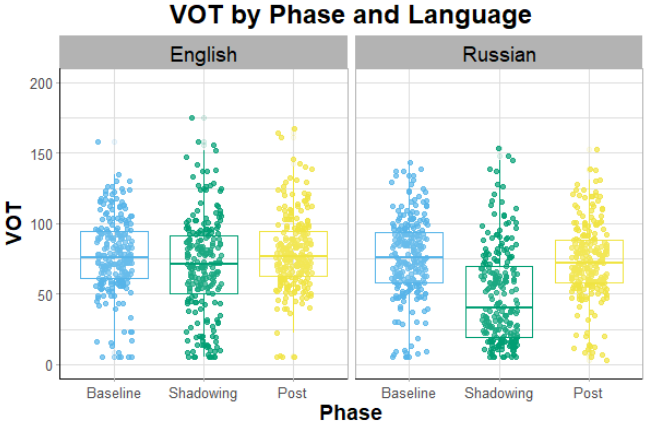


Figure 1: VOT values by Phase and Condition (left: English, right: Russian).

In contrast to the Russian condition, the results of the mixed-effects model indicated that participants'

VOT did not change significantly in either the shadowing or post-test phases compared to the baseline phase in the accented English condition (shadowing; p = .6, post; p = .3) despite a visually apparent tendency towards shorter VOTs in the shadowing phase (Figure 1, left panel). Furthermore, the pairwise post-hoc analysis confirmed that VOT values in the shadowing phase did not differ significantly from those in the post-test phase (p = .8). The results suggest that participants' shadowed speech did not demonstrate convergence towards the short lag VOTs of Russian-accented voiceless stops. In contrast, the results suggest that exposure to Russian speech had a significant effect on shadowed voiceless stops of participants in the Russian condition. The VOT of their voiceless stops in the shadowing phase was significantly shorter than their own baseline recordings, suggesting convergence towards, or imitation of the model speaker. Nevertheless, the effect did not generalize to the English words recorded in the post-test phase.

Finally, results of the mixed-effects model revealed a significant effect of Voicing Lang Knowledge (β = -20.6, SE =6.22, t = -3.31, p = .003), suggesting that participants who had existing knowledge of a language that realized voicing in the same manner as Russian had significantly shorter voiceless VOT across all conditions and phases: 25.1 ms shorter on average than participants without this knowledge.

4. DISCUSSION

The present study found convergence in voiceless VOT towards an unfamiliar language (Russian condition) but not towards an unfamiliar accent (accented English condition). Regarding the CAT, it appears that talkers are willing/able to imitate other languages, but not accented versions of their own, native language. However, the CAT approach may not be completely suitable to the setting of the present experiment. Imitating speech in an unfamiliar language is akin to speaking another language and likely reflects different processes than those involved in adjusting one's speech to accommodate an acoustically different version of one's L1.

The difference between the conditions could be further exacerbated by some of our methodological choices. In the accented English condition, participants may have detected the non-native (shorter) VOT in words like "pot" but were aware of the target phoneme due to the accompanying illustrations. As a result, they may have produced the words with native, English-like pronunciation. Comparatively, in the Russian condition, participants had no knowledge of the phonemic content of the words they shadowed and had to rely on the model's pronunciation in reproducing each item.

Moreover, as participants in the Russian condition were not familiar with the Russian language and were not provided with any orthographic representation of the words they shadowed, it is possible that, at least in some case, Russian voiceless stops were shadowed 'faithfully', that is, with short lag VOT, because they were mis-analyzed by participants as voiced (English short lag VOT is characteristic of voiced stops). This could also explain the lack of carry-over of this imitation effect into participants' English in the post-test phase, since there is no reason to expect that their pronunciation of Russian 'voiced' stops should affect their pronunciation of English voiceless stops. Future analyses of other acoustic correlates of voicing, as well as acoustic properties of other segments, such as vowels, would help paint a more complete picture of acoustic imitation of an unfamiliar language.

At face value, these findings do not lend support to the theory that exposure to accented speech, or to another language, can lead to L1 drift or to sound change in the ambient/dominant language via long-term phonetic convergence. Although participants imitated VOTs of Russian voiceless stops in the shadowing phase, there was no carry-over of this imitation into their English speech in the post-test. Furthermore, in the accented English condition, there was no convergence and no carry-over. That said, evidence of L1 drift was not completely absent from the present study, as results indicated that participants with existing knowledge of voicing languages, regardless of proficiency, produced significantly shorter voiceless VOT than their peers.

While present results do not suggest the tendency for talkers to converge with accented speech, since only one type of accent was examined in the study and the language, we should be cautious about generalizing these findings. It is possible that talkers may be more willing to accommodate certain speakers and accents than others. This may especially be true for Russian, considering the global political climate at the time of research. Therefore, future research should examine other L2 accents.

Moreover, it needs to be acknowledged that in this study participants underwent relatively short exposure in rather artificial conditions, and future work should consider longer exposure in more naturalistic settings in order to corroborate these results.

5. REFERENCES

[1] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, 2012, doi: 10.1016/j.wocn.2011.09.001.

[2] D. Kim and M. Clayards, "Individual differences in the link between perception and production and the mechanisms of phonetic imitation," *Lang. Cogn. Neurosci.*, vol. 34, no. 6, pp. 769–786, 2019.

[3] K. Nielsen, "Specificity and abstractness of VOT imitation," *J. Phon.*, vol. 39, no. 2, pp. 132–142, 2011, doi: 10.1016/j.wocn.2010.12.007.

[4] J. S. Pardo, "Measuring phonetic convergence in speech production," *Front. Psychol.*, vol. 4, 2013, doi: 10.3389/fpsyg.2013.00559.

[5] G. Zellou, R. Scarborough, and K. Nielsen, "Phonetic imitation of coarticulatory vowel nasalization," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3560–3575, 2016, doi: 10.1121/1.4966232.

[6] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Lang. Speech*, vol. 55, no. 2, pp. 231–248, 2012.

[7] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," presented at the INTERSPEECH, 2011.

[8] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Atten. Percept. Psychophys.*, vol. 72, no. 8, pp. 2254–2264, 2010, doi: 10.3758/BF03196699.

[9] H. Giles, *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press, 2016.

[10] C. B. Chang, "A novelty effect in phonetic drift of the native language," *J. Phon.*, vol. 41, no. 6, pp. 520–533, 2013, doi: 10.1016/j.wocn.2013.09.006.

[11] C. B. Chang, "Phonetic drift," in *The Oxford handbook of language attrition*, Oxford University Press, 2019.

[12] C. G. Clopper, "Sound change in the individual: Effects of exposure on cross-dialect speech processing," *Lab. Phonol.*, vol. 5, no. 1, pp. 69–90, 2014, doi: 10.1515/lp-2014-0004.

[13] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964, doi: 10.1080/00437956.1964.11659830.

[14] C. Ringen and V. Kulikov, "Voicing in Russian stops: Cross-linguistic implications," *J. Slav. Linguist.*, vol. 20, no. 2, pp. 269–286, 2012.

[15] S. Dufour and N. Nguyen, "How much imitation is there in a shadowing task?," *Front. Psychol.*, vol. 4, p. 346, 2013.

[16] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, 2019.

[17] Y. Shrem, M. Goldrick, and J. Keshet, "Dr. VOT: Measuring positive and negative voice onset time in the wild," *ArXiv Prepr. ArXiv191013255*, 2019.

[18] R Core Team. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv Prepr. ArXiv14065823*, 2014.

[1] This study did not collect information about participants' race, which we recognize as a limitation considering the effect it has on sociolinguistic variation in the United States.

VARIABILITY IN SWEDISH VOICELESS FRICATIVE CONTRASTS

Carla Wikse Barrow[1], Marcin Włodarczak[1], Mattias Heldner[1] & Sofia Strömbergsson[2]

[1]Department of Linguistics, Stockholm University

[2]Department of Clinical Science, Intervention and Technology, Karolinska Institutet
{carla.w.barrow,wlodarczak,heldner}@ling.su.se, sofia.strombergsson@ki.se

ABSTRACT

Swedish voiceless fricatives vary in their realisation across dialects, as well as across word positions. In the present paper we investigate individual differences in contrast between the voiceless fricatives of twenty adult speakers of Central Standard Swedish. Visual representations and measures of contrast robustness based on spectral moments one and two (spectral centre of gravity and spectral standard deviation) were derived to explore phonetic variation across and within speakers. The described variation in contrast robustness can contribute to a broader understanding of inter- and intra-speaker variation in the realisation of Swedish fricatives.

Keywords: fricatives, acoustics, Swedish

1. INTRODUCTION

Swedish has a relatively rich inventory of voiceless fricatives with five phonemes: labio-dental /f/, alveo-dental sibilant /s/, pre-dorso-alveolar sibilant /ɕ/ (the "tje" sound), velar (or labio-velar) /ɧ/ (the "sje" sound) and glottal /h/ (see e.g. [1, 2, 3]).[1] The fricative system is even denser when allophonic variation in certain dialects is taken into account. The "sje" sound is one of few Swedish consonants that exhibits dialectal variation, with allophones such as prepalatal retroflex sibilant [ʂ], velar [ɧ] and velar [x] [2, 4, 3]. In Central Standard Swedish (henceforth CSS), the average speaker is expected to exhibit (at least) two allophones of the "sje" sound, [ʂ] and [ɧ], described as light and dark "sje" respectively, based on their perceptual qualities [2, 5]. This allophonic variation is, in part, constrained by phonological distribution, such that light "sje" occurs in postvocalic and intervocalic position and dark "sje" in prevocalic position, except in complex onsets [1]. It has been suggested that some speakers use both retroflex [ʂ] and non-retroflex [ʃ], that is, they apply a retroflexion rule where /rs/ sequences are pronounced [ʂ] whereas light "sje" is realised as [ʃ] [5]. Nevertheless, most CSS speakers neutralize this distinction, and the final fricative in "kurs" and "dusch" are indistinguishable [1, 3]. For a more comprehensive description of geographical variation, the reader is referred to [2, 4, 3], and references therein.

Static and dynamic acoustic characteristics of word initial CSS fricatives have previously been described in [6], using spectral moments analysis [7]. The results showed that /s-ɕ-ɧ/ were distinguishable through the first spectral moment (M1), on a group level. However, to the best of our knowledge, individual differences with respect to the contrast between the voiceless fricatives of CSS remains unexplored. Moreover, little is known regarding the acoustic properties and variation of light "sje" /ʂ/ in modern speakers of CSS. In this paper, we visualize and investigate fricative contrast in twenty speakers of CSS. We include both word initial fricatives [f, s, ɕ, ɧ] and word medial and -final [ʂ].

In addition, given that for the English sibilants /s-ʃ/ female speakers were reported to produce slightly larger contrast compared to male speakers [8, 9], we also investigate gender differences. Different measures of contrast robustness have been proposed (see e.g. [9]). In the current paper, we use an overall discriminability measure based on M1 that takes into account distance between fricative categories as well as dispersion within categories [8].

2. METHOD

2.1. Participants

Twenty native speakers of Swedish (10 female) between 18 and 43 years (M = 29.55, SD = 7.59) from the Svealand region took part in a speech production task. All participants provided written informed consent prior to recordings and were compensated for their participation with one movie ticket.

2.2. Task and materials

The speech production task included 28 fricative initial words (see Table 1), two fricative medial words ([kɔʂa] "to cross", [dəʂa] "to shower") and four fricative final words ([kɔʂ] "cross", [dəʂ] "shower", [lɑːʂ] "Lars", [garɑːʂ] "garage") that were repeated four times each, resulting in 136 target words per speaker. The fricative initial words were presented on a screen in writing as well as through headphones, whereas the fricative medial and final words were presented only in writing. The audio prompts were recorded by a female native speaker of Swedish from the Svealand region. The participants were instructed to speak clearly without hyperarticulating.

2.3. Recording

Participants were recorded in the aneachoic chamber in the phonetics lab of Stockholm University using a Brüel & Kjær Type 4189-L-001 microphone and a Motu 8M sound interface. A 44.1 kHz sampling rate and 16 bit quantization were used throughout, although the sounds were downsampled to 22 kHz prior to analysis.

2.4. Annotation and segmentation

The soundfiles were annotated in Praat [10]. The onset of the word initial fricatives was marked at a rapid increase in zero crossing and/or high frequency energy, and the offset was marked at the zero crossing prior to the onset of periodic energy of the following vowel. The segmentation procedure of the word medial and word final fricatives was less detailed. For each fricative, a fricative region of relative stability was determined. The onset and offset of this region were marked conservatively, leaving the less stable edges of the fricative segment outside the boundaries. Clear intrusions of voicing and bursts in the fricative noise were excluded from the fricative region. Therefore, fricative duration and overall intensity cannot reliably be derived from these segments and are not used in the analysis. However, as M1 has been shown to be relatively stable in the central part of voiceless fricatives [6], we believe this segmentation is sufficient to derive a representative spectral centre of gravity. All boundaries were determined through simultaneous inspection of the waveform and spectrogram. A dynamic range of 40 dB and a spectral range of 0-8 kHz were used throughout.

Recordings including disturbances, such as clearing of the throat, or background noise that affected the fricative were removed prior to analysis, leaving a total of 2671 tokens.

2.5. Feature extraction and analysis

A Praat script extracted spectral moments 1 and 2 (centre of gravity and standard deviation) from a 30 ms window Hann Window at centered at 50% of the fricative/fricative region. The spectra were filtered through a Hann pass filter (200-11000 Hz, with 100 Hz smoothing) prior to feature extraction.

The light "sje" showed significant overlap between retroflex [ʂ] and non-retroflex [ʃ] (mean M1 was 3170 Hz (SD= 515) and 3285 Hz (SD= 585), respectively), and are therefore grouped together in the plots and further analyses.

The acoustic features were imported and summarized in R [11]. The phoneme discriminability measure described in [8] was calculated for the fricatives pairs /s-ɕ/, /ɕ-ʂ/ for all speakers. The discriminability measure was calculated as the distance between fricative categories (i.e. difference between the mean M1 for each category) divided by the square root of the mean of the M1 variances of the two categories; $(\mu_{f_1} - \mu_{f_2})/\sqrt{(Var_{f_1} + Var_{f_2})/2}$, were f_1 denotes the first fricative and f_2 the second fricative.

3. RESULTS

Figure 1 shows M1-M2 plots for each speaker. Visual inspection of the plots reveals that /ɧ/, /ɕ/ and /s/ are well separated for all speakers, although some exhibit slight overlap in M1 for /ɕ/ and /s/ (see e.g. the speaker in row five; column three). Both distance between and dispersion within categories varies across speakers and gender, compare for example the speakers in the first three rows in the first column. Velar /ɧ/ is most compact in M1, while category dispersion varies across individuals with respect to the sibilants, although /s/ is often more disperse than /ɕ/.

Labio-dental /f/ is distinguishable for most participants if M2 is included, although the sound is generally most disperse in M1.

For most speakers (e.g. individuals in column two) /ʂ/ overlaps entirely or partly with /ɕ/ in the M1-M2 dimension. However, some individuals show clear separation between the two sounds (e.g. the speakers in row one; column three and row three; column five). For individuals who show separation, [ʂ] has a slightly lower spectral centre of gravity than [ɕ].

To investigate robustness of contrast, a measure of overall discriminability was calculated for the

Vowel context	/f/	/s/	/ɕ/	/ɧ/
[a]	"fall" [fal]	"juice" [saft]	"squabble" [ɕafs]	"chess" [ɧak]
[ɑː]	"plate" [fɑːt]	"auditorium" [sɑːl]	"nagging" [ɕɑːt]	"scarf" [ɧɑːl]
[iː]	"kefir" [fiːl]	"sieve" [siːl]	"wedge" [ɕiːl]	"ski" [ɧiːda]
[eː]	"wrong" [feːl]	"zebra" [seːbra]	"chain" [ɕeːdja]	"spoon" [ɧeːd]
[ɛ]	"five" [fɛm]	"cell" [sɛl]	"cane" [ɕɛp]	"ship" [ɧɛp]
[uː]	"foot" [fuːt]	"sun" [suːl]	"skirt" [ɕuːl]	"lump" [ɧuːk]
[ʉː]	"ugly" [fʉːl]	"sour" [sʉːɾ]	"bull" [ɕʉːɾ]	"sick" [ɧʉːk]

Table 1: The fricative initial targets words used in the repetition task. The words' English gloss is written in quotes and a transcription in brackets.

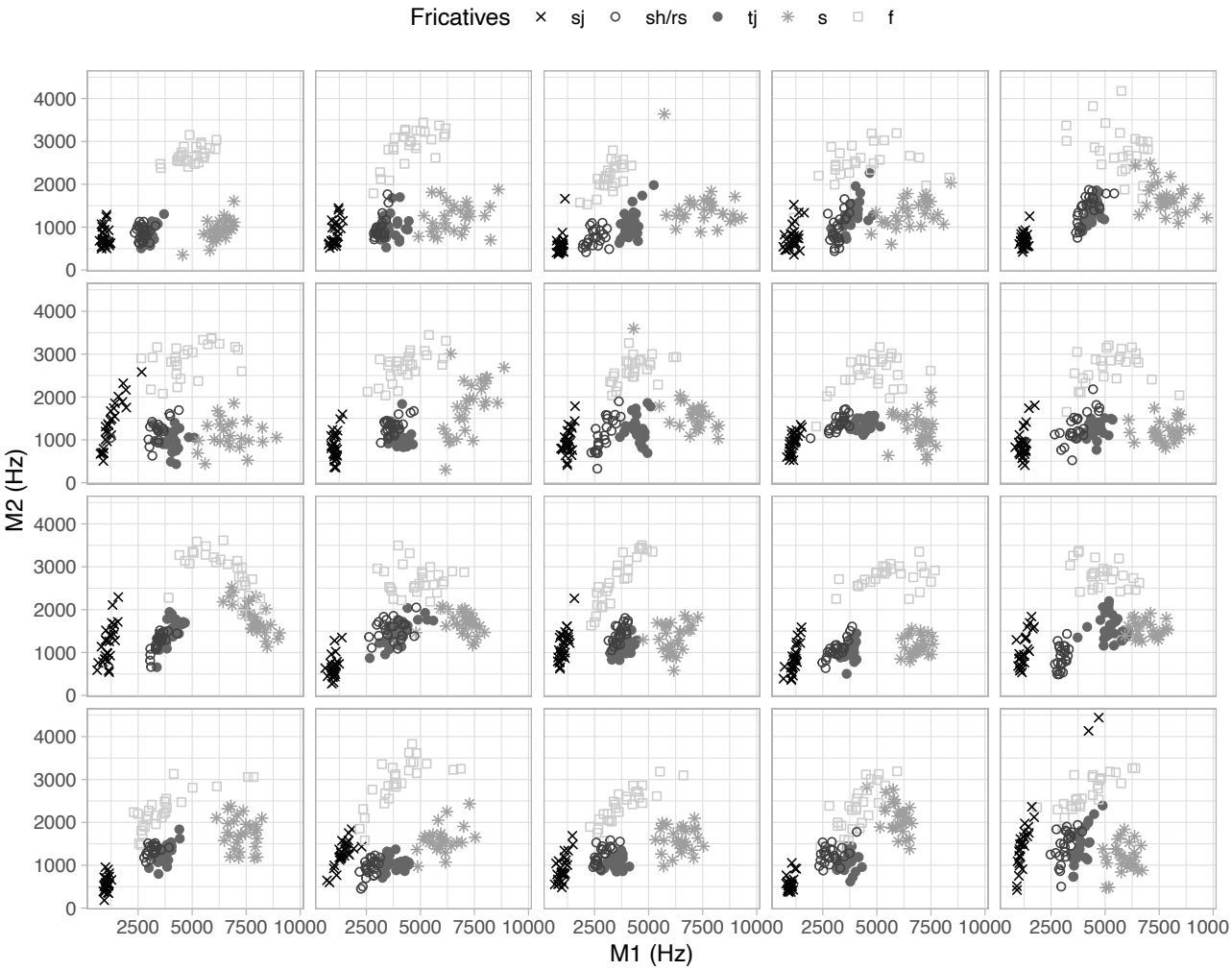


Figure 1: Scatterplot of M1 (center of gravity) and M2 (standard deviation) for Swedish /f/, /s/, /ɕ/ (tj), /ʂ/ (sh/rs) and /ɧ/ (sj), measured at the center of the fricative noise for all speakers. The top two rows show female speakers and the bottom rows show male speakers.

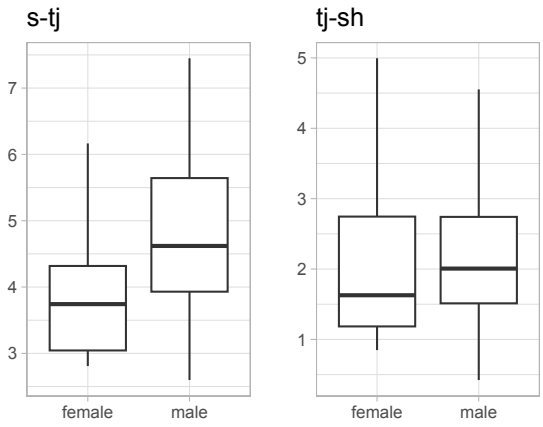


Figure 2: Discriminability in female and male speakers for the fricative pairs /s-ɕ/ (s-tj) and /ɕ-ʂ/(tj-sh)

sibilant pairs [s-ɕ] and [ɕ-ʂ]. Figure 2 shows boxplots of robustness of contrast for female and male speakers. As evident in the figure, male speakers have higher discriminability scores for both fricatives pairs, although the difference is small for [ɕ-ʂ] and the variation substantial for [s-ɕ].

4. DISCUSSION

This study is an acoustic description of Swedish voiceless fricative contrast, that highlights individual differences in variability and fricative discriminability. The results uncover differences in fricative contrast across individuals, both with regards to the dispersion and distance between the voiceless fricatives of SCC. Exploring individual differences is important as patterns which are not evident on a group level emerge. Previous descriptions (see [6]) indicate that Swedish [f] is not readily captured by M1 on a group level, and spans across the /s-ʃ/ distribution of M1 with some overlap in M2 as well. For many speakers in the present investigation, [f] is clearly distinguishable, especially if M2 is included. With respect to the light "sje", [ʂ] is separated from [ɕ] by M1 for some speakers, but most show substantial overlap with the "tje" sound in the M1/M2 dimension.

The discriminability measure shows differences across individuals with respect to contrast robustness in the sibilant fricatives. The gender-related differences described in [8], such that female speakers produced more distinctive categories than male speakers, do not seem to hold for the individuals in this study. In fact, contrary to previous findings, male speakers in the current study showed slightly higher levels of discriminability as compared to females, although the within-group variation was substantial. It should be noted that differences in filtering of the spectra affect the spectral moments calculated thereof, and although Swedish sibilants [s-ɕ] are similar to English [s-ʃ], the investigations in [9, 8] are not directly comparable to this study.

4.1. Limitations and future research

The current study is limited as it involves only two spectral parameters to describe voiceless fricatives of CSS. Including other acoustic cues related to the fricatives spectral representation as well as temporal and intensity properties, and their change over time, may provide insight into patterns that are not evident here.

Moreover, a larger sample including more words and connected speech would be beneficial. Exploring the contrast between light "sje" and the "tje" sound (/ɕ/) in dialects that primarily use the fronter [ʂ] (i.e. Northern Swedish dialects) as well as dialects with affricative realisation of tje (i.e. [tɕ] in Finland Swedish) [2, 3] would also be of interest, to capture variation in contrast and structure of voiceless fricative systems.

Finally, a developmental perspective on fricative variability and contrast would be valuable for research on acquisition of Swedish. The results presented here provide insight into the degree and type of variation that can be expected in adult productions of voiceless fricatives, which are the targets that child productions are often compared to.

5. REFERENCES

[1] T. Riad, *The phonology of Swedish*. Oxford: Oxford University Press, 2014.

[2] P. Lindblad, "Svenskans sje- och tje-ljud i ett allmänfonetiskt perspektiv," Ph.D. dissertation, Lund University, 1980.

[3] G. Bruce, *Vår fonetiska geografi : om svenskans accenter, melodi och uttal*, 1st ed. Lund: Studentlitteratur, 2010.

[4] R. Shosted, "Acoustic characteristics of Swedish dorsal fricatives," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3888, 2008.

[5] C. Garlén, *Svenskans fonologi : [i kontrastiv och typologisk belysning]*, 1st ed. Lund: Studentlitteratur, 1988.

[6] C. Wikse Barrow, M. Włodarczak, L. Thörn, and M. Heldner, "Static and dynamic spectral characteristics of swedish voiceless fricatives," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 2588–2600, 2022.

[7] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-

initial voiceless obstruents: preliminary data," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 115–123, 1988.

[8] R. Romeo, V. Hazan, and M. Pettinato, "Developmental and gender-related trends of intra-talker variability in consonant production," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3781–3792, 2013.

[9] J. J. Holliday, P. F. Reidy, M. E. Beckman, and J. Edwards, "Quantifying the robustness of the english sibilant fricative contrast in children," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 622–637, 2015.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 1994-2021. [Online]. Available: https://praat.org

[11] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[12] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[1] However, /h/ is sometimes described as an approximant [5] and often excluded from acoustic descriptions of fricatives (see e.g. [12]).

Acoustic phonetic classification of primary infant protophones

Eugene H. Buder[1], D. Kimbrough Oller[1]

[1]University of Memphis, Memphis, TN, United States
ehbuder@memphis.edu

ABSTRACT

Infants' utterances are readily classified by researchers and parents into phonatory-based categories such as vocant, squeal, and growl. This classification has proceeded on auditory grounds referencing features such as pitch and voice quality, but there have been no systematic efforts to identify appropriate acoustic phonetic dimensions underlying these vocalization classes: the current study addresses that problem. Fundamental frequency studies have been numerous, and observations regarding non-modal regimes are also longstanding; here those dimensions, reflecting well-recognized features of these speech-like utterances, are integrated. Analysis of a corpus representing utterances produced by 3 infants, in recordings at ~3, 6, and 9 months, reveals a feature set adequately classifying these types relative to human judges: The phonatory protophone classes distinctly cluster in a 3D space defined by f_0 mean, f_0 SD, and a regime-based scale aligned with glottal closure. Classification of infant utterances in these terms reveals distinct development trends.

Keywords: Infant Vocalization, Phonation, Development, Voice Quality

1. INTRODUCTION

Clear phonation-based categories of utterance are so typical in human infancy as likely to be universal, and terms such as squeal, growl, and vocant (vowel-like) have been used in language development literature for at least 50 years [1-3]. While not considered fully phonological and inherently fuzzy, these categories do exhibit distinctive features that can be defined acoustic phonetically, with the greatest elaboration seen in phonatory aspects [4]. Over the same period, acoustic phoneticians have observed evidence in the harmonic structure of infant voices of numerous alternative patterns of vibration [5]. These phonatory patterns have since come to be understood within the framework of non-linear dynamics [6], and dubbed vibratory regimes in this context [7].

The occurrence of vibratory regime shifts during infant vocalization is inherently categorical, yielding categories expected from general non-linear dynamics theory, from harmonic to subharmonic, or even to non-periodicity, e.g. to chaotic vocal fold vibrations or cessation of voicing, all regularly observed in infant phonation [6, 8-10] Systematic analyses of harmonic frequency distributions or amplitudes [11] can be used to fully characterize this aspect of infant voice quality. Furthermore, the arrangement of regimes in such systems is expected to order according to some underlying parameter; in the case of regimes observed in infancy, glottal closure (produced by adduction, aerodynamic force variation, particular configuration of cover/body tissues, etc.) appears to be a prime candidate [12].

In perceptual terms, the most elementary distinction among the earliest protophones is pitch; vocants mid-range, growls lower and squeals higher. Traditionally, many vocal development researchers have been interested in f_0 characteristics [3, 8, 13-15], especially because of anatomical developments of the larynx [16]. Yet virtually none of these studies distinguished amongst protophones, or considered f_0 ambiguities introduced by harmonic regime variations. Another shortcoming of prior reports utilizing f_0 measures is failure to report within utterance variability, overlooking the possible importance of f_0 variability. Finally, intensity is rarely explored in infant vocal development.

In summary, study of the acoustic structure of early protophones is overdue, while acoustic parameter candidates have now been identified: This acoustic phonetic grounding of infant vocal categories helps re-orient basic questions in infant vocal development by operationalizing an acoustically-based utterance typology.

2. METHODS

2.1. Materials

The corpus of 2,312 vocalizations analysed here represents virtually all of the non-vegetative vocalizations heard by trained personnel as either a vocant, growl, or squeal utterance produced by three female infants during pairs of 20-minute lab-based sessions, recorded at approximately 3, 6, and 9 months of age. The infant and caregiver were both fitted with FM wireless microphones with Countryman Associates MEMWF0WNC capsules in vests, configured to minimize friction noise and maintain as consistent mic-mouth distances as possible. Calibration tones were recorded during each

session at known dB levels for use as reference levels for converting RMS voltages in recordings to dB levels comparable across sessions.

2.2. Protophone Coding

Human coding of the targeted vocalization types was conducted in the AACT environment [17] by either the first author or a highly experienced PhD student following procedures outlined in [2], and focusing exclusively on speech-like vocalizations. Segmentations of ongoing vocalizations followed breath-group phrasing principles [18], exclusion of protracted breathy offsets and of vocalizations shorter than 50 ms or effectively at floor amplitudes. Inter-judge Kappa reliability on 20% of the dataset was 0.66, establishing a benchmark for acoustic classification; judge's disagreements in coding were subsequently resolved by consensus coding.

2.3. Acoustic Analyses

2.3.1 fo

The AACT environment implements the ActiveX library of the TF32 program, including implementation of its waveform-correlation based pitch determination algorithm's parameter controls and hand editing tools [19]. Another difficulty with prior f_o literature has been variable results due to different ways of handling the many-octave range of infant phonation.

Here, analysis parameter variation, followed by hand-marking of glottal epochs, was used to yield carefully validated data series ranging from 15 Hz to over 4 kHz. Importantly, in the face of harmonic ambiguities, selection of appropriate f_o was guided by prior regime coding as described below, e.g. subharmonics should not be tracked as they were already coded as such, f_o in chaos and stops is null by definition, etc. At the utterance level, the f_o dataset analysed here ranges from means of 43 Hz (a growl, dominated by pulse register) to 1740 Hz (a squeal, dominated by loft register).

2.3.2 dB

Mean RMR values were extracted at the utterance level and converted to standardized dB with reference to the recorded calibration tones.

2.3.3 Regimes

Regime coding, adopting procedures documented in [7], involved inspection of narrowband spectrograms aided by auditory judgments and inspection of waveforms, in order to classify all segments of each protophone as one of 8 types representing distinct vocal fold vibration patterns: modal, pulse, subharmonics, biphonation, chaos, 'c-stop' (overadduction), 'o-stop' (underadduction), and "hi-modal".

In the absence of any prior reporting on a distinct loft (or 'falsetto') register in infancy, the "hi-modal" code was actually a placeholder for high-f_o segments for further consideration of this possibility. Subsequent research, investigating pitch breaks within infant utterances from modal into loft, identified significant differences across those breaks in the relative amplitudes of the first two harmonics (H1-H2), consistent with variations between modal and falsetto registers known in the human adult literature to involve distinctive dynamics of VF vibration [11]. This work provided threshold f_o and H1-H2 values that were implemented in the current dataset to discriminate loft segments (516 Hz and 1.6 dB respectively), and this register distinction is then incorporated under the general rubric of regime type.

2.3.4 Closure Scale

Regime codes subsegment protophone codes; these can be as short as 50 ms, extend throughout the whole vocalization, or break the vocalization into many units including the possibility of regime repeats. To associate regimes with protophones it was therefore necessary to 'score' vocalizations according to the occurrence of regimes within them, and aggregating their percentages of occurrence within each vocalization served this purpose. This did not, however, solve a 'sparsity' issue: Most vocalizations contained just one regime and none contained them all. A full aggregation of the categorical codes onto some relatively continuous metric addresses this issue.

Aggregation of regimes can be motivated by a principle by which they might be expected to occur along some interpretable scale, and models with parametric control of non-linear dynamic systems' driving forces motivate such principles. Principles of phonation point towards respiratory flow and glottal closure as likely parameters. Observing that over- and under-compressed glottal configurations provides natural poles for a 'closure scale' (while not ruling out flow or other additional parameters), this scaling was explored amongst the entire array of regimes via χ^2 analyses. Ultimately regime associations with protophones provided the scaffolding needed for this exploration.

Details of that analysis are provided in [12], with an emphasis on highly significant +/- standardized deviates within tables associating protophone categories with regime occurrence tallies, as seen in

Table 1 here. Most notably, clear associations emerge for squeal (positively with loft and o-stops), and for growls (positively with subharmonics, pulse, chaos, and c-stops). Aggregating the percentages of those groupings helped to mitigate the sparsity issue (see Table 1 note), and observing that all non-modal regimes were negatively associated with the vocant type, scoring modal as a 0 helped to complete the scale. (Only biphonation is neglected, due to its ambiguous associations but also its sparsity.)

Table 1: Chi-square and standardized deviate statistics assessing the associations of individual regime codes with protophone types.

			Standardized Deviates		
Regime	n[1]	χ^2	growl	vocant	squeal
Loft	296	965	-3.1	-11.0	+26.6
Subharmonics	308	349	+14.8	-8.9	+2.3
Pulse	362	297	+13.9	-8.6	ns
C-Stop	157	110	+8.9	-4.7	ns
Biphonation	106	156	+4.6	-6.7	+9.0
Chaos	90	137	+10.2	-5.3	ns
O-Stop	62	124	ns	-4.3	+10
Modal	1599	42.7	ns	ns	-3.2

[1]n = the number of protophones containing an instance of the regime code.

Scaling the current dataset in this manner, a continuum is obtained that ranges -/+1, and 1614 (70%) of the vocalizations receive a non-0 value. Technically, the values might better be dubbed 'opening' rather than compression due to the polarity of the scale, but the underlying concept remains the same.

3. RESULTS

3.1 Parameter Identification

Exploration of the four candidate parameters identified above was initially conducted by logistic regression analyses with protophone type as the 3-way categorical outcome and candidate parameters as predictors (f_o mean and SD values log transformed to correct for skewness). An optimal model was identified retaining f_o mean, f_o SD, and the closure scale as predictors, obtaining 85% overall classification success. While addition of the dB mean value added a marginally significant increment in log-likelihood, this resulted in no improvement to classification, and it improved χ^2 from 2137 only to 2141. Based on these diagnostics, and conveniently for visualization purposes, a 3D space was created—see Figure 1.

Assessed by Cohen's Kappa against the human coding, the 3-parameter classification yielded 0.67; virtually identical to the reliability with which human coders matched one another, thereby affirming the full adequacy of this model. Subsequent appraisal of Hedge's *g* effect sizes in 2-way contrasts amongst protophone pairs also affirmed that each parameter was operating independently: All three yielded large (>1) effects for distinguishing squeal/vocant and squeal/growl, and while effects remained large for all three in the vocant/growl distinction, the closure scale effect size was twice that of the f_o parameters.

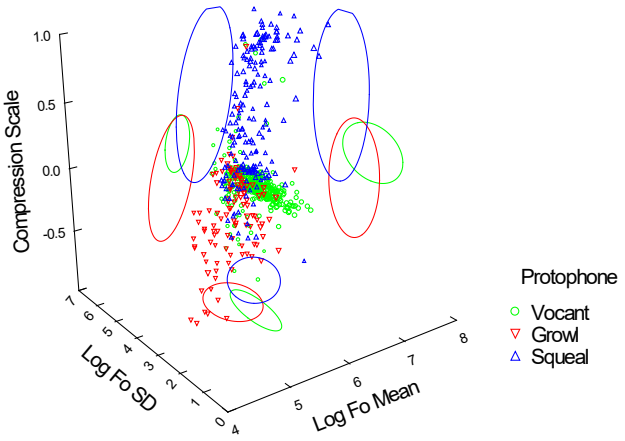


Figure 1: Scatterplot of all vocalizations in the three-dimensional space defined by the logistic regression predictors optimally classifying these vocalizations by protophone type. Ellipses projected onto two-dimensional facets are centered on class means and with axes defined by standard deviations to encompass data with probabilities equal to 0.6827 ("Compression" = "closure").

3.1 Developmental observations

All four parameters were retained for further exploration of developmental trends. Given the small number of infants included, the observations made here are not claimed to generalize. The purpose is rather to demonstrate applicability of the phonetic parameters identified in the proposed protophone classification model, as driven by the sufficient power of this dataset and its sampling strategy. The infants in this dataset were all typically developing and, as among the first to have been recorded, all happened to be female, but should be represented as a random factor in variance

modelling, so a mixed-model analysis was appropriate for examination of age differences in these parameters with protophone codes factored in as well.

Key outcomes from that analysis were: 1) With all other effects accounted for, the three primary f_o- and regime-based parameters significantly accounted for the main effect protophone variation, while dB mean did not; and 2) As a simple main effect, no Age variance was explained; yet, 3) Age × Protophone variance explained was significant, driven by numerous specific interactions in all three primary parameters. Figure 2 displays these effects for two parameters that merit further commentary here.

While, as in prior literature on the first years of life, no age effects were obtained for average vocalization pitch, drops in f_o are observable in *non-squeal* vocalizations, and (as seen by miniscule error bars in vocants) these effects were especially consistent in that protophone type. Examining vocant f_o cross age, small but significant age-difference Cohen's *d* effect sizes on the order of 0.31 are obtained. Furthermore, when vocant values were converted back to Hz the drop in those vocalizations during the 2nd half year of life was 17 Hz (5 %), a period when vocal tract growth would be expected to cause such effects.

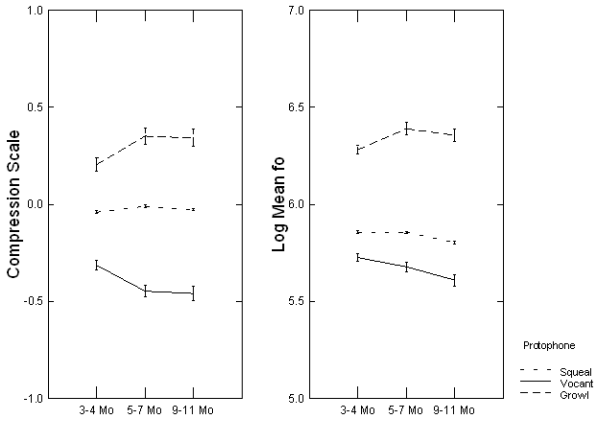


Figure 2: Line charts depicting protophone-specific age effects for two phonetic parameters; see legend for protophone line weights. Error bars are SEs.

The display of closure ("compression") scale values in the left panel of Figure 2 affirm the factors driving Age × Protophone interaction results with this parameter as well: While vocants are steady in this respect, it is interesting to observe the divergence of values occurring from the first to second age periods, as it would appear to be strong affirmation of the awareness among vocal development scientists that infants across the age range considered here appear to engage in vocal play, practicing variations on the phonatory parameters that yield the apparent early protophone categories [1, 20]. It may also evince operation of the respiratory-laryngeal system described recently as among the least well understood "Developmental Functional Modules" underlying emergent stages of early human vocal development [21].

4. DISCUSSION

The present analysis provides a rationale for identifying fairly straightforward acoustic parameters, f_o mean, f_o SD, and the identified "closure scale," as underlying protophone classification but also vocal development itself, with findings that, pending affirmation among more infants, should identify development trends more precisely than prior approaches which specifically neglected protophone or vibratory regime distinctions. Because glottal closure incorporates parameters such as medial compression and aerodynamics, the framework aligns well with other phonatory-based approaches [22, 23].

The outcome that these parameters model the human percept is by design, so the framework they provide for categorizing infant protophones implies that these specific parameters are utilized by the auditory system; this could of course be tested by psychoacoustic research paradigms. And it has always been expected that some acoustically-based model should be able to work well, since human judges have consistently agreed well on the identification of the three phonatory protophone types. What is quite surprising, however, is that, unlike in many cases where human coding has been found to be vastly superior to automated analyses of vocal activity based on acoustic evidence only [2], the present analysis yielded an outcome where the acoustically-based classification approximated its own gold standard (the human coding) precisely.

Having reached that essential benchmark, next steps following this approach should consider automation of spectral harmonic analyses (e.g. [24]) to side-step human coding of regimes, optimally guiding then guiding parameter selection in PDAs for f_o analysis. Such developments would facilitate applications of the model to more infants to affirm the trends observed here, and the resulting larger datasets should then help create improved processing, or at least greater tolerance for error in acoustic input, while still generating significant outcomes. A mapping of f_o and phonatory harmonic structures in infant vocal output should thereby lay groundwork for new approaches to the understanding of vocal development.

Acknowledgment: This work was supported by NIDCD R01001027 to D. K. Oller.

7. REFERENCES

[1] D. K. Oller, *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, p. 428.

[2] D. K. Oller, E. H. Buder, H. Ramsdell, A. S. Warlaumont, L. B. Chorna, and R. Bakeman, "Functional flexibility of infant vocalization and the emergence of language," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 110, pp. 6318-6323, 2013, doi: 10.1073/pnas.1300337110.

[3] M. Z. Laufer and Y. Horii, "Fundamental frequency characteristics of infant non-distress vocalizations during the first twenty-four weeks," *Journal of Child Language,* vol. 4, pp. 171-184, 1977.

[4] E. H. Buder, A. S. Warlaumont, and D. K. Oller, "An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective," in *Comprehensive Perspectives on Speech Sound Development and Disorders: Pathways from Linguistic Theory to Clinical Practice*, B. Peter and A. A. N. MacLeod Eds. Hauppage, NY: Nova Science Publishers, Inc., 2013, pp. 103-134.

[5] P. A. Keating and R. Buhr, "Fundamental frequency in the speech of infants and children," *J Acoust Soc Am,* vol. 63, pp. 567-571, 1978.

[6] W. Mende, H. Herzel, and K. Wermke, "Bifurcations and chaos in newborn infant cries," *Physics Letters A,* vol. 145, pp. 418-424, 1990 1990.

[7] E. H. Buder, L. Chorna, D. K. Oller, and R. Robinson, "Vibratory regime classification of infant phonation," *J Voice,* vol. 22, pp. 553-564, 2008.

[8] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *J Acoust Soc Am,* vol. 72, pp. 353-365, 1982.

[9] R. Buhr and P. Keating, "Spectrographic effects of register shifts in speech production," *The Journal of the Acoustical Society of America* vol. 62, p. 25, 1977.

[10] M. P. Robb and J. H. Saxman, "Acoustic observations in young children's non-cry vocalizations," *J Acoust Soc Am,* vol. 83, pp. 1876-1882, 1988.

[11] E. H. Buder, V. F. McDaniel, E. R. Bene, J. Ladmirault, and D. K. Oller, "Registers in infant phonation," *J Voice,* pp. 1-12, 2018.

[12] E. H. Buder and D. K. Oller, "Glottal closure as a parameter underlying phonatory regime occurrence in infancy," submitted for review.

[13] S. Amano, T. Nakatani, and T. Kondo, "Fundamental frequency of infants' and parents' utterances in longitudinal recordings," *J Acoust Soc Am,* vol. 119, pp. 1636-1647, 2006.

[14] M. P. Robb and J. H. Saxman, "Developmental trends in vocal fundamental frequency of young children," *Journal of Speech, Language, and Hearing Research,* vol. 28, no. 3, pp. 421-427, 1985.

[15] E. Scheiner, K. Hammerschmidt, U. Jurgens, and P. Zwirner, "Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants," *J Voice,* vol. 16, no. 4, p. 509, 2002.

[16] R. D. Kent and H. K. Vorperian, "Development of the craniofacial-oral-laryngeal anatomy: A review," *Journal of Medical Speech-Language Pathology,* vol. 3, no. 3, pp. 145-190, 1995.

[17] *AACT - Action Analysis Coding and Training Software*. (2020). Intelligent Hearing Systems Corp., Miami, FL.

[18] M. P. Lynch, D. K. Oller, M. L. Steffens, and E. H. Buder, "Phrasing in prelinguistic vocalizations," *Developmental Psychobiology,* vol. 1, pp. 3-25, 1995.

[19] *TF32*. (2018). University of Wisconsin-Madison, Madison, WI.

[20] F. J. Koopmans-van Beinum and J. M. van der Stelt, "Early stages in the development of speech movements.," in *Precursors of early speech*, R. Zetterstrom Ed. New York: Stockton Press., 1986, pp. 37-50.

[21] R. D. Kent, "Developmental functional modules in infant vocalizations," *J Speech Lang Hear Res,* vol. 64, no. 5, pp. 1581-1604, May 11 2021, doi: 10.1044/2021_JSLHR-20-00703.

[22] J. H. Esling, S. R. Moisek, A. Benner, and L. Crevier-Buchman, *Voice quality: The laryngeal articulator model*. Cambridge, UK: Cambridge University Press, 2019.

[23] M. P. Robb, F. Yavarzadeh, P. J. Schluter, V. Voit, W. Shehata-Dieler, and K. Wermke, "Laryngeal constriction phenomena in infant vocalizations," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 1, pp. 49-58, 2020.

[24] G. Ramsay, "Multitaper harmonic analysis of infant vocalizations," vol. 144, no. 3, pp. 1767-1768, 2018, doi: 10.1121/1.5067818.

DENTOFACIAL DISHARMONY PATIENTS' SIBILANTS DIFFER FROM CONTROLS' MORE IN SOURCE THAN FILTER PROPERTIES

Madeleine Oakley[1], Auvi Tran[2], Ciana Paye[1], Emma Trudan[1], Timothy Turvey[2], George Blakey[2], David Zajac[2], Jeff Mielke[1], Laura Jacox[2]

[1]North Carolina State University; [2]University of North Carolina Chapel Hill
mjoakley@ncsu.edu; ljacox@live.unc.edu

ABSTRACT

This study uses multitaper spectral analysis to examine the differences in consonants produced by patients who present with different dentofacial disharmonies (DFD) including severe overbites (Class II), underbites (Class III) and anterior open bites. Previous studies have found that patients with these malocclusion types all produce sibilants and plosives with increased spectral center of gravity and increased spectral spread relative to controls. This result is puzzling since some DFD groups differ from controls in opposite ways. To better understand the articulatory basis of these differences, we apply several spectral shape measures and find that all groups of DFD patients produce /s ʃ t tʃ/ with mid-frequency spectral peaks that are less prominent than those of the control group, but peak frequency measures are largely the same across all groups. This indicates that the DFD patients differ more in sibilant noise source than front cavity size.[1]

Keywords: sibilant, multitaper, disordered speech, malocclusion, source-filter

1. INTRODUCTION

The great majority of patients who have dentofacial disharmonies (DFD) present with speech sound disorders [1, 2, 3]. DFD patients present with skeletal and dental discrepancies in the anterior-posterior (AP, horizontal), vertical, and/or transverse dimensions, including Class II overbite, Class III underbite, and anterior open bite (AOB). Previous work examining the relationship between DFDs and speech sound disorders has found an association between malocclusion and speech disorders [4, 1].

Studies examining the acoustic properties of disordered speech have shown that there are differences in spectral properties of consonants between DFD patient groups. Using spectral moment analysis, significant increases were seen in the 1st and 2nd spectral moments (center of gravity and spectral spread) of DFD patients including Class II, III and AOB groups when compared to Class I reference subjects for /k t ʃ s tʃ/ [5, 6]. Patients with AOB presented with the greatest quantitative differences and Class II with the smallest. Among the Class III DFD cohort, the subgroup of Class III AOB patients experienced the greatest differences in the first and second spectral moments compared to the control group, potentially due to the combination of vertical and AP discrepancies [5, 6].

These spectral moment results are surprising as each DFD cohort has unique jaw disproportions and vocal tract anatomy, and it is not predicted that all DFD groups would demonstrate increased first and second spectral moments. One reason for the discrepancy between predicted acoustic differences between groups and articulations may be the use of spectral moment analysis. As discussed by [7], changes in certain moment measurements can be affected by multiple articulatory movements, which may be obscuring differences across DFD cohorts.

Spectral moment analysis typically involves applying the discrete Fourier transform to a waveform. Because the resulting spectra have a large variance, these estimates may have poor precision when applied to fricative waveforms [8]. They are suitable for measuring spectral moments but less so for measuring properties of spectral peaks which may offer more insight into the articulatory basis of the observed differences. Multitaper spectral analysis averages multiple estimators based on the Discrete Fourier Transform and detection of harmonic components despite presence of external noise; this helps to produce a single spectrum minimizing spectral bias, noise, variance, and error [9, 10]. Such spectra are more suitable for more targeted measurements of spectral peaks. Measures such as mid-frequency spectral peak frequency, change in the mid-frequency peak over time, the amplitude difference between the low frequency minimum and the mid-frequency peak, and high frequency slope measures [7]. These acoustic

measurements more closely reflect articulatory changes, such as labiality (which lowers mid-frequency peaks), changes in the size of the front cavity over time, and aerodynamic effects of the decrease in the oral constriction over time.

The current study applies multitaper spectral analysis to speech recordings from pre-surgical DFD populations to evaluate how the severity of Class II, Class III, and AOB malocclusions correlate with speech distortion of consonants. This data provides insight into the physiologic interplay between articulators and explores how jaw disproportions are linked to speech disorders.

2. METHODS

2.1. Participants

One hundred fifty DFD patients with Class II closed bite ($n = 34$), Class III closed bite ($n = 64$), Class I anterior open bite ($n = 7$), Class II anterior open bite ($n = 9$), or Class III anterior open bite ($n = 36$) malocclusions were consecutively enrolled from a DFD clinic. All patients were screened by an orthodontist prior to enrollment based on their malocclusions. Fifty-three reference controls with Class I occlusion and skeletal base were included. All patients were aged 18-40 years old, 115 were female, and 88 were male. Orthodontic and surgical records included occlusal measurements, dental models, photos (intraoral and extraoral), and panoramic and cephalogram radiographs.

2.2. Materials and Procedure

The target sounds /s ʃ t tʃ k/ were selected for analysis based on the prediction that malocclusions affect the production of these consonants. 20 English words containing word-initial /s ʃ t tʃ k/ before the vowels /i u æ ɑ/ were embedded in the carrier phrase "say __ again." All phrases were randomized and presented one at a time. The words were each repeated 3 times for a total of 12 target tokens per phone per speaker. Recordings were made in a sound-attenuated booth while participants wore a head-mounted microphone.

2.3. Analysis

The Montreal Forced Aligner [11] was used to segment and align the recordings using transcripts. Stop and affricate releases were identified based on the high frequency energy and voicing probability [12], roughly following [13]. Spectra were measured from 20 ms windows aligned to the middle of each /s ʃ/ frication interval and to the start of the release of each /t tʃ k/ using the `spectRum` package [14] for `R` [15], with 8 tapers and a bandwidth parameter of 4. Preemphasis was not applied.

This study focuses on two acoustic measurements for each consonant which are designed to access different articulatory parameters, following [16, 7, 17]. The frequency of the mid-frequency spectral peak (measured in the 2-5kHz range for /ʃ tʃ/ and in the 3-7kHz range for /s t/) is expected to be associated with the size of the front cavity between the lingual constriction and the mouth opening. Spectral amplitude difference (the difference in amplitude between the mid-frequency spectral peak and a lower-frequency spectral trough (measured in the 1-2kHz range for /ʃ tʃ/ and otherwise in the 1-3kHz range) is expected to be associated with the noise source itself. /k/ lacks a mid-frequency spectral peak so it was not measured in this way.

Separate mixed-effects linear regression models were run for each consonant with the different acoustic measurements as the dependent variable. Patient group and gender are included as fixed-effects, and word and speaker are included as random effects.

3. RESULTS

The mean spectra for each consonant for each patient group and gender, are shown in Figure 1. The spectra are adjusted so that 0dB corresponds to the low-frequency trough. For /s ʃ t tʃ/, the control group has more prominent spectral peaks, and the frequencies of these peaks appear similar across groups. The /k/ spectra appear similar across patient groups, with the exception of male Class I AOB, which is probably a spurious difference due to the small number of patients in that category ($n = 2$).

3.1. Spectral amplitude difference

Spectral amplitude difference is expected to be correlated with the noise source. As seen in Figure 1, there appears to be a tendency for the control group to produce all of the consonants (except /k/) with a higher amplitude than the other patient groups. This measure was compared across patient groups for each target consonant. Overall, the spectral amplitude difference for /s ʃ t tʃ/ is higher for the control group than for the other patient groups. There is also a tendency for the AOB groups to have a lower spectral amplitude difference compared to non-AOB groups. These findings are represented in Figure 2, divided by gender groups.

Separate regression models for each consonant

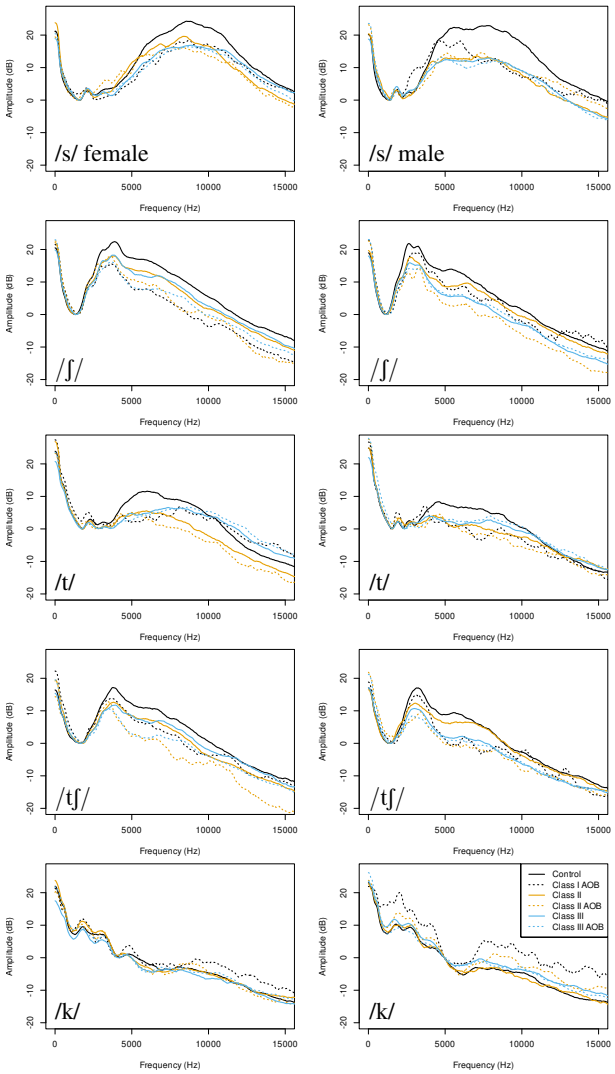


Figure 1: Mean spectra averaged for each patient group for each consonant, by gender (left: female; right: male)

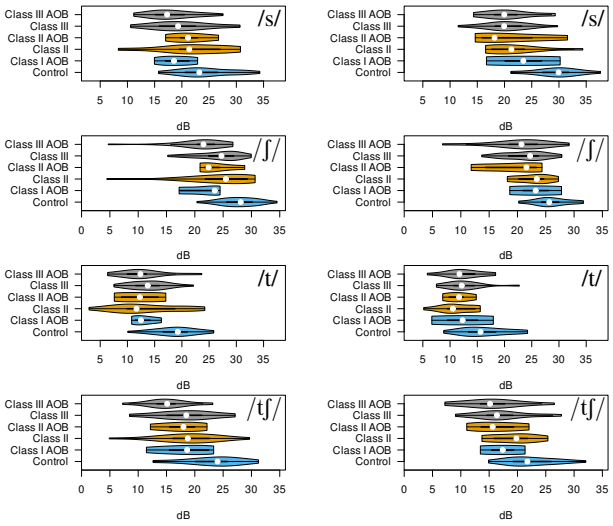


Figure 2: Spectral amplitude difference (dB) measured for each patient group for each consonant with a mid-frequency peak, by gender (left: female; right: male)

Group	/s/		/ʃ/	
Control	25.63	$p = 4.48e\text{-}16^{***}$	27.73	$p < 2e\text{-}16^{***}$
Class I AOB	-6.44	$p = 0.002^{**}$	-6.15	$p = 0.0008^{***}$
Class II	-4.35	$p = 8.37e\text{-}05^{***}$	-3.28	$p = 0.0004^{***}$
Class II AOB	-5.48	$p = 0.001^{**}$	-4.83	$p = 0.0007^{***}$
Class III	-6.54	$p = 4.10e\text{-}11^{***}$	-4.11	$p = 5.41e\text{-}7^{***}$
Class III AOB	-7.45	$p = 1.08e\text{-}10^{***}$	-6.38	$p = 8.17e\text{-}11^{***}$

Group	/t/		/tʃ/	
Control	18.40	$p = 2.92e\text{-}7^{***}$	23.15	$p < 2e\text{-}16^{***}$
Class I AOB	-5.77	$p = 0.001^{**}$	-7.44	$p = 0.0002^{***}$
Class II	-5.15	$p = 1.79e\text{-}8^{***}$	-3.83	$p = 0.0001^{***}$
Class II AOB	-6.04	$p = 1.12e\text{-}5^{***}$	-6.28	$p = 6.77e\text{-}5^{***}$
Class III	-4.28	$p = 5.42e\text{-}8^{***}$	-5.14	$p = 1.48e\text{-}8^{***}$
Class III AOB	-5.13	$p = 2.57e\text{-}8^{***}$	-7.78	$p = 8.60e\text{-}13^{***}$

Table 1: Estimates from linear regression models for each consonant, showing the effect of patient group on spectral amplitude difference (dB)

confirm that each patient group has a significant effect on spectral amplitude difference for /s ʃ t tʃ/. The outputs of the regression models are summarized in Table 1, showing the effect of each patient group on the spectral amplitude difference for each consonant.

3.2. Peak frequency

The frequency of the mid-frequency peak is expected to correlate with front cavity size. This measure was compared across patient groups for all four consonants with mid-frequency peaks, as shown in Figure 3. The control group does not differ in peak frequency value from the different patient groups for the post-alveolar consonants /ʃ/ or /tʃ/. The control group does, however, produce /s/ with a slightly higher frequency spectral peak (measured at the full frequency range) than Class I AOB and Class II patient groups. The control group also produces /t/ with a higher peak frequency than all other patient groups (measured in the lower-mid frequency range). Separate linear regressions for each consonant confirm these observations, showing a significant effect of patient group type for /s t/, but not /ʃ tʃ/. The output of the regressions for /s t/ are summarized in Table 2.

4. DISCUSSION

This study provides detailed quantitative assessments on a large DFD sample representing all major malocclusion classifications. The multitaper analysis shows differences across DFD groups that

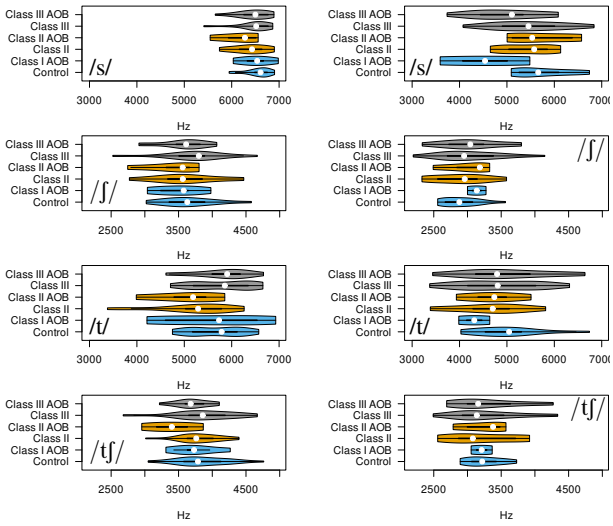


Figure 3: Mid-frequency peak (Hz) measured for each patient group for each consonant with a mid-frequency peak, by gender (left: female; right: male)

Group	/s/		/t/	
Control	8623	$p = 3.66e\text{-}12$***	4265	$p = 5.93e\text{-}14$***
Class I AOB	-993.5	$p = 0.048$*	-957.9	$p = 0.0003$***
Class II	-600.4	$p = 0.018$*	-645.6	$p = 2.82e\text{-}6$***
Class II AOB	-235.9	$p = 0.54$	-546.8	$p = 0.008$**
Class III	-60.16	$p = 0.78$	-514.7	$p = 1.40e\text{-}5$**
Class III AOB	-415.0	$p = 0.1$	-600.4	$p = 1.44e\text{-}5$***

Table 2: Estimates from linear regression models, showing the effect of patient group on peak frequency (Hz) for each consonant with a significant difference

have not previously been observed.

First, the results of this study show that the control group produces /s ʃ t tʃ/ with a higher spectral amplitude difference than any other patient group. As discussed by [7] this measurement is designed to measure the 'sibilance' of fricatives and reflect variations in the noise source at different frequency regions. With the control group having the largest spectral amplitude followed by Class II non-AOB, and then all AOB, this could mean that Class II non-AOB patients are best able to posture so as to produce an alveolar constriction resulting in turbulence most similar to that produced by Class I controls. Most Class II patients present with deficient mandibles in which their tongues are also naturally positioned more posteriorly, which contributes to a more similar constriction location compared to the controls. The AOB cohort has a more difficult time positioning the tongue tip and teeth to produce control-like turbulence. The Class III patients exhibiting a lower amplitude as well can be explained by those patients commonly having prognathic mandible in which their tongues naturally rest more anteriorly. Adaptation for tongue position in Class III patients is difficult, potentially leading to a more anterior constriction location. The results of this study confirm that this measurement does reflect differences in articulatory configuration, as it is predicted that each DFD cohort has different jaw configurations that effect the constriction location of each target consonant.

Second, the control group does *not* produce post-alveolar consonants with a different peak frequency from other patient groups, but *does* produce /s/ with a higher peak frequency than Class I AOB and Class II groups, and *does* produce /t/ with a higher peak frequency than all other patient groups. Postalveolar consonants are produced with lip rounding in English, and it appears that all patient groups are able to achieve a degree of lip rounding that yields peak frequencies similar to controls when lip rounding is a feature of the consonant and/or when the target peak frequency is low. It appears that all patient groups are able to compensate with their lips, which is a movement available to anyone, for the rounded consonants. When lip rounding is not a feature of the consonant, as in /t/, all patient groups have somewhat lower peak frequencies than controls. Class I AOB and Class II have lower peak frequencies for /s/ as well, which is also unrounded. Note that the compensatory behavior that would raise peak frequency is *less* lip rounding. Patient groups do not seem to be using lip unrounding to achieve the higher peak frequency values for unround consonants.

Previous studies [5, 6] found higher center of gravity and spectral spread for various consonants produced by all DFD groups. The fact that center of gravity was observed to be *higher* is likely explained by the flatter spectra produced by Class II, Class III, and all AOB groups. Notably, those previous FFT-based analyses applied preemphasis and calculated spectral moments over a wide frequency range (0-17.64kHz), meaning that a relatively flat spectrum can yield a higher center of gravity than one with a prominent spectral peak under 10kHz. Using multitaper spectra to identify the mid-frequency spectral peak associated with the front cavity resonance made it possible to determine that the main difference between Controls and DFD cohorts is that the controls produce more prominent spectral peaks, indicative of greater sibilance produced at the lingual-alveolar constriction. Differences in peak frequency observed for /t s/ could be attributable to differences in the noise source that cannot be easily compensated for with the lips.

5. REFERENCES

[1] L. I. Black, A. Vahratian, and H. J. Hoffman, "Communication disorders and use of intervention services among children aged 3-17 years: United states, 2012. nchs data brief. number 205." *Centers for Disease Control and Prevention*, 2015.

[2] A. Ocampo-Parra, B. Escobar-Toro, V. Sierra-Alzate, Z. V. Rueda, and M. C. Lema, "Prevalence of dyslalias in 8 to 16 year-old students with anterior open bite in the municipality of envigado, colombia," *BMC oral health*, vol. 15, no. 1, pp. 1–6, 2015.

[3] M. A. Morris, S. K. Meier, J. M. Griffin, M. E. Branda, and S. M. Phelan, "Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey," *Disability and health journal*, vol. 9, no. 1, pp. 140–144, 2016.

[4] R. Feeney, L. Desha, J. Ziviani, and J. M. Nicholson, "Health-related quality-of-life of children with speech and language difficulties: A review of the literature," *International Journal of Speech-Language Pathology*, vol. 14, no. 1, pp. 59–72, 2012.

[5] M. M. B. Keyser, H. Lathrop, S. Jhingree, N. Giduz, C. Bocklage, S. Couldwell, S. Oliver, K. Moss, S. Frazier-Bowers, and C. Phillips, "Impacts of skeletal anterior open bite malocclusion on speech," *FACE*, vol. 3, no. 2, pp. 339–349, 2022.

[6] H. Lathrop-Marshall, M. M. B. Keyser, S. Jhingree, N. Giduz, C. Bocklage, S. Couldwell, H. Edwards, T. Glesener, K. Moss, S. Frazier-Bowers *et al.*, "Orthognathic speech pathology: Impacts of class iii malocclusion on speech," *European Journal of Orthodontics*, vol. 44, no. 3, pp. 340–351, 2022.

[7] L. L. Koenig, C. H. Shadle, J. L. Preston, and C. R. Mooshammer, "Toward improved spectral measures of/s: Results from adolescents," 2013.

[8] P. F. Reidy, "A comparison of spectral estimation methods for the analysis of sibilant fricatives," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. EL248–EL254, 2015.

[9] D. B. Percival and A. T. Walden, *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[10] C. H. Shadle *et al.*, "Acoustics and aerodynamics of fricatives," *The Oxford handbook of laboratory phonology*, pp. 511–526, 2012.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[12] S. Gonzalez and M. Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[13] J. Cronenberg, M. Gubian, J. Harrington, and H. Ruch, "A dynamic model of the change from pre-to post-aspiration in andalusian spanish," *Journal of Phonetics*, vol. 83, 2020.

[14] P. F. Reidy, "spectRum," 2013, r package. [Online]. Available: https://github.com/patrickreidy/spectRum

[15] R Core Team, "R language definition," 2015.

[16] L. Jesus and C. Shadle, "A parametric study of the spectral characteristics of European Portuguese fricatives," vol. 30, pp. 437–464, 2002.

[17] C. H. Shadle, W.-r. Chen, L. L. Koenig, and J. L. Preston, "Acoustic variability of fricatives in normal adults," 2020, poster presented at ISSP 2020, Yale University.

[1] This research was funded by the National Institutes of Health (NIH), specifically the National Institute of Dental and Craniofacial Research (NIDCR), through a K08 award (to Laura Jacox), with a Grant Award Number 1K08DE030235- 01A1 and by the National Science Foundation through grant SMA-1730479 (to Jeff Mielke).

Acoustic analysis of L2 English phonemes for language identification

Samantha Williams, Vincent Hughes, and Paul Foulkes

University of York, York, UK
samantha.williams@york.ac.uk, vincent.hughes@york.ac.uk, paul.foulkes@york.ac.uk

ABSTRACT

This study provides an initial step towards developing a phonetically interpretable method of language identification to determine a speaker's L1 given their L2 English speech. 112 male L2 English speakers representing five L1s were selected from the Speech Accent Archive. F1, F2, and Duration were measured for four monophthongal vowel phonemes (/ɪ/, /i/, /ɛ/, and /ɔ/). These phonetic features were then used to calculate by-phoneme probabilities using the multivariate kernel density method. Findings suggest that for computational analysis, individual segmental acoustics in isolation are unlikely to be sufficient for making good predictions. Additionally, the features that are most effective for identifying a language depend on both the segment and comparison under investigation.

Keywords: L2 English, speech acoustics, language identification, computational analysis

1. INTRODUCTION

Language identification systems, as the name suggests, attempt to automatically identify the native language (L1) of a speaker based on a sample of his or her speech. Such systems have wide usage for commercial purposes (e.g., as a front-end for speech recognition) and for forensic and security purposes (e.g., to identify the language being spoken by a person of interest in an incriminating recording).

Most language identification systems approach this task from either an acoustic-phonetic approach (e.g. [1,2]) exploiting segmental realisations, or a purely acoustic approach which works on more abstract, mathematical representations of the speech signal extracted across an entire recording (such as MFCCs; e.g. [3]).

Current state-of-the-art systems which use Deep Neural Networks (DNN), have been shown to out-perform other approaches when the number of possible L1s is high. For example, previous studies have cited classification rates (CR) of 72% for a 6-way comparison task using DNN [4] and 84% for a 10-way comparison using TDNN [5].

One drawback of purely acoustic systems, and particularly those that utilise deep learning, is that they lack the direct interpretability that segmental approaches provide. Such transparency is essential for use in forensic applications. However, interpretability often means sacrificing performance. For example, [1] provides a recent example of an acoustic-phonetic approach which uses distance measures between segments to determine the L1. The system resulted in a CR of 61% in a 7-way comparison.

Despite the many recent advances in speech technology, language identification based on a speaker's L2 speech remains a difficult and under-researched area. Factors like proficiency, regional variations in the L1, along with other social factors such as attitude towards the target variety, impact L2 realisation and which features are acquired [6-8]. These factors introduce variability within a group, thus making the L2 variety classes difficult to model. Additionally, similar target L2 pronunciations mean between-group differences are expected to be more subtle than between dialects or with language identification based on L1 speech. Whether the L2 varieties should be even be grouped by their L1 and not a finer grained label is a separate issue not discussed in this paper (see [9]).

The present study specifically deals with General American (GA) English as the target L2 of a set of speakers with a wide range of L1s. We aim to explore L2 English phoneme data through the lens of language identification. This presents an initial step towards developing a larger (i.e., based on a wider range of features at different linguistic levels), phonetically interpretable system that can recognise a speaker's L1 based on samples of their L2 English. This approach has forensic applications as a priority.

This paper first explores the distributions of F1, F2, and Duration values for four vowel phonemes, followed by an evaluation of classification rate and probability scores. Finally, we consider whether the phonetic realisations of the different L1 groups are different enough that they can be accurately identified and whether classification performance can be explained in a phonetically interpretable way.

2. METHODS

2.1. Data

The speakers used for this study were 112 male speakers of five varieties of L2 GA English selected from the Speech Accent Archive [10]. The groups are referred to by their L1 as labelled in [10]: Arabic, French, German, Mandarin, Portuguese. Each recording contained approximately 30 seconds of a speaker reading the "Stella" passage. The Montreal Forced Aligner (MFA) [11] was used to segment each sample at the phone level and generate a time aligned Praat TextGrid. The General American acoustic model and grapheme-to-phoneme dictionary were used (english_arpa_us and english_us_ipa_g2p). Therefore, all segment labels are based on the expected phoneme, not the actual realisation. Some alignment error is to be expected but should be sufficient for the purposes of this paper. A study of the performance of the MFA on a subset of this dataset found that 80-93% of boundary placements were within 20 ms and 67-84% were within 10 ms [12]. Taking measurements at the midpoint of the segment mitigates the impact of alignment errors.

2.2. Feature extraction

Using a Praat script [13], F1 and F2 measurements were automatically extracted from a 25 ms midpoint frame for all monophthongal vowel segments contained in the passage, along with duration of the segment. The reference formant values were based on manually corrected values for 2-3 male speakers from each of the varieties.

Four vowel phonemes were selected for this study as a means of focusing on segments that were (i) relatively easy to measure automatically, and (ii) showed high or low variability between varieties based on the manually corrected values. The following four vowels were selected: /ɪ/, /i/, /ɛ/, /ɔ/.

Each feature was z-score normalised, based on the mean across all speakers, to bring the features onto the same scale while maintaining original differences between distributions. This was done to allow for the use of a single smoothing parameter in formula (1) (see below; [14]). Tokens with an absolute z-score above 3 for any of the features were considered measurement errors and removed from the analysis. 97% of the data was retained: 1187 /i/, 893 /ɪ/, 439 /ɛ/, and 556 /ɔ/ tokens.

2.3. Language identification

For within-language comparisons, each language set was randomly split into 80% training and 20% test data using K-fold cross validation (K=5). An adaptation of Aitken and Lucy's [15,16] multivariate kernel density (MVKD) likelihood ratio (LR) formula was used to calculate the probability of a given speaker's data given a model for each language group. This was done via a modified version of Morrison's [17] MATLAB MVKD implementation. MVKD provides a transparent method of comparing

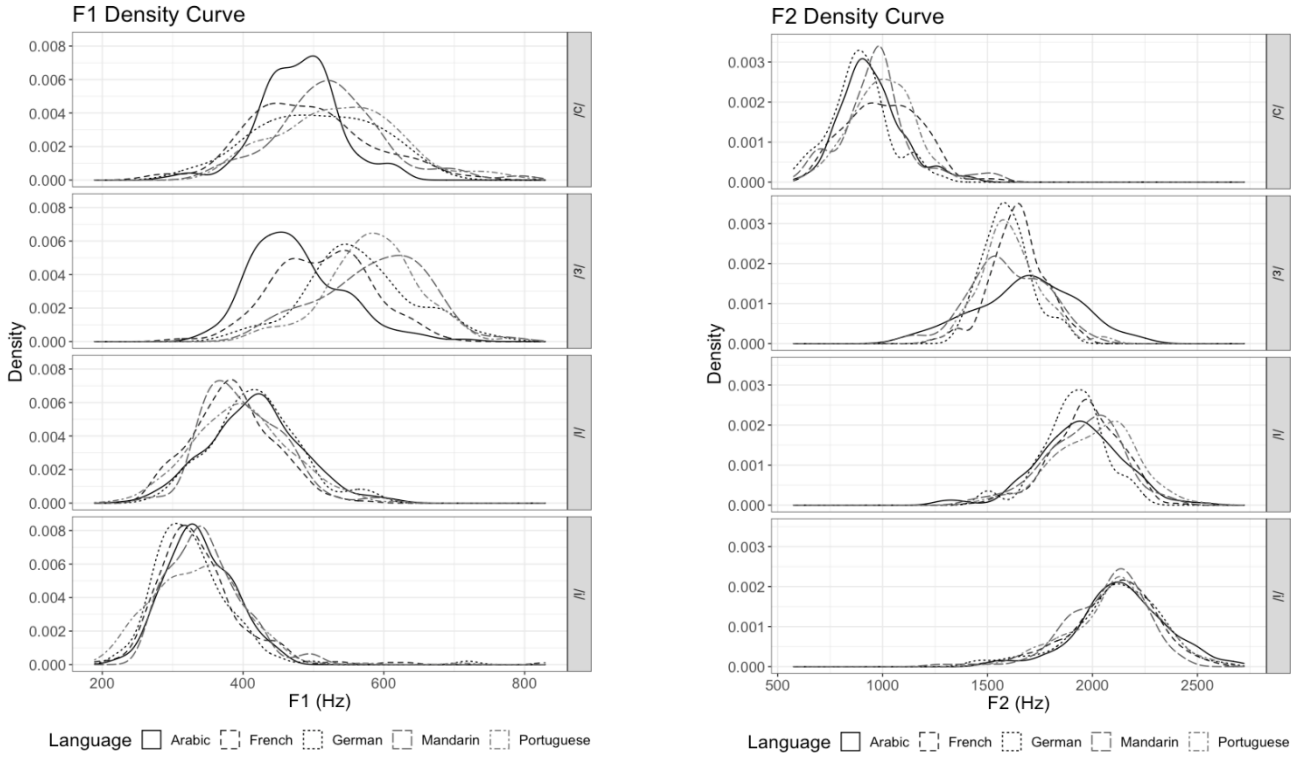


Figure 1: Density distributions for raw F1 and F2 values by vowel and L1.

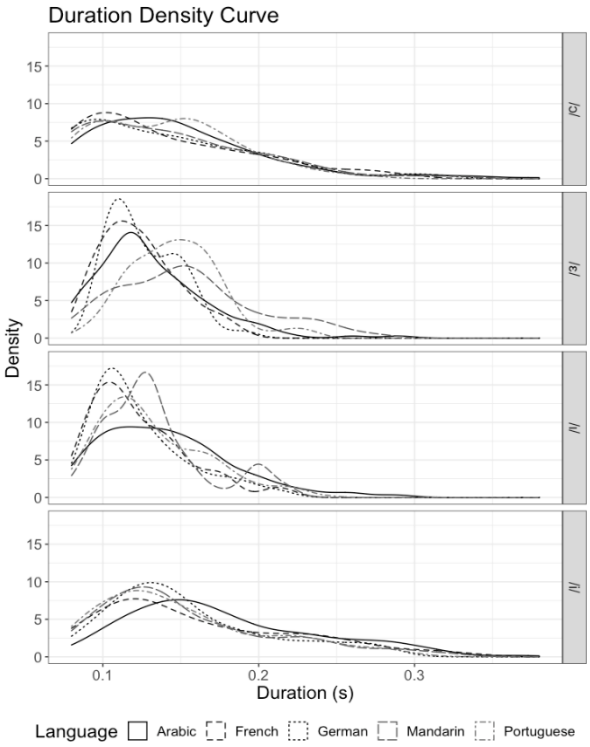


Figure 2: Density distributions for raw Duration values by vowel and L1.

distributions and has been used widely for the purposes of speaker comparison research (e.g., [18,19]). Here, only an adapted version of the denominator of the MVKD LR was used. This involves modelling data for each language group with a multivariate kernel density distribution, made up of equally weighted Gaussians from each speaker in the training data. This also accounts for correlation between features. The mean probability of each feature vector (F1, F2, Duration) from each token for each speaker given each language model is then calculated, as in (1):

$$(1)\ f(z|\mu_1, C, U, H_k) = \int_\theta f(\boldsymbol{z}|\theta_1, U) f(\theta_1|\boldsymbol{\mu}, C) d\theta$$

Where $\boldsymbol{z}$ is the measurements from the speaker, $\boldsymbol{\mu}_1$and C are the within-group mean and covariance, U is the within-speaker covariance and operating under the assumption of hypothesis k (where H_k is the hypothesis that the speaker comes from a given language model).

2.3.1 Evaluation

For each test speaker, probabilities were ranked by language model and then compared against the ground truth to arrive at a conclusion which could be classified in binary terms as correct or incorrect. This was then used to generate an overall classification rate (CR). Additionally, logLRs, which estimate the strength of the evidence, ($\boldsymbol{z}$; i.e., the acoustic measures for a test speaker), were calculated for comparisons between target (H_1) and non-target (H_2) language pairs, as in (2):

$$(2)\ logLR_{1,2} = \frac{P(\boldsymbol{z}|H_1)}{P(\boldsymbol{z}|H_2)}$$

LogLRs are widely used as the output of many automatic speech technologies and are particularly used within the forensic domain. For ease of interpretation, numerical logLRs here are converted to verbal equivalents using the scale in [20].

3. RESULTS

3.1. Feature distributions

Figures 1 and 2 display the distributions of raw F1, F2, and Duration values for the five language groups and four vowels. Larger images and reference formant values are provided as supplementary material [21].

The vowels demonstrate high and low between-language variability (see 2.2). For instance, /i/ shows considerable overlap in the feature space for all selected varieties except Portuguese (for F1) and Arabic (for Duration). However, due to the wider distribution, if a Portuguese speaker's features fall towards the centre of the distribution (around 300Hz along F1 for example), there will always be a higher probability that the speaker belongs to one of the other L1s. This suggests there will be a lot of confusability between the varieties for this segment that will not result in a strong decision in any direction. Conversely, /ɛ/ shows variability across all three features. /ɪ/ displays variability mainly in Duration and F2, while /ɔ/ varies mainly in F1 and F2.

The results predict the performance ranking of the segments, from best to worst, will be: /ɛ/, /ɔ/, /ɪ/, /i/.

3.2. Language identification results

The overall classification rates, where chance is 20%, were: 15% (French), 9% (Arabic), 17% (Mandarin), 11% (Portuguese), and 48% (German). German had a noticeably better classification rate, which is largely attributed to the performance on /ɔ/ and /i/. The performance rankings of each vowel can be found in Table 1. Within a language group, /i/ was most frequently classified correctly, while /ɔ/, /ɪ/ and /ɛ/ tended to be correctly classified less frequently.

Table 2 displays the most common misidentification for each L1 and vowel. In most instances, the probability of the speaker having Mandarin as an L1 was higher than for their correct L1. However, it is worthwhile to note that pairwise logLRs mostly fell between 0 and 2, but more often

between 0 and 1. This indicates a very weak classification, meaning even if the segment was classified incorrectly, the strength of the evidence is limited to moderate on the verbal scale [20].

L1	Vowel			
	1st	2nd	3rd	4th
Arabic	/i/	/ɛ/	/ɔ/	/ɪ/
French	/ɪ/	/ɛ/	/i/	/ɔ/
German	/ɔ/	/i/	/ɪ/	/ɛ/
Mandarin	/i/*	/ɔ/*	/ɪ/	/ɛ/
Portuguese	/i/	/ɪ/	/ɔ/	/ɛ/

Table 1: Vowel phonemes ranked by performance (best to worst) for each L1. * Indicates a tie.

L1	L1 (model)			
	/i/	/ɔ/	/ɪ/	/ɛ/
Arabic (A)	P	G	M	M
French (F)	A/M	G	M	M
German (G)	M	-	M	M
Mandarin (M)	P	G	F	P
Portuguese (P)	M	G	M	M

Table 2: Most common misidentification for each L1 and phoneme label.

Some pairwise comparisons showed greater strength of evidence. For example, speakers with French L1 tended to show logLRs around 3 in support of either hypothesis for /ɛ/. However, higher strength of evidence generally appeared to be speaker-specific rather than patterning with specific L1s.

3. DISCUSSION

The difficulty with classifying L2 varieties is evident from the poor classification rates found in this study. By having the same target L2 English variety, we are minimizing the between-group variability, especially at the phoneme-level.

The performance ranking of the phonemes was not as expected from 3.1. Surprisingly, /ɛ/, which showed the largest differences in Figures 1 and 2 between varieties tended to be the least well-classified. We suspect this is due to where in the feature space a speaker falls. For example, /ɔ/ has very wide distributions for F1 across all varieties. Therefore, despite large differences between the overall distributions, the tails of a given distribution overlap with means of others. Conversely, the vowel with the smallest between-language variation in mean values, /i/, had the best performance for most varieties. The tested varieties all have a phoneme with this label in their native inventory [22-29]. It is possible, however, that /i/ is realised differently across languages, and that such differences are maintained in the L2, in addition to less within-speaker variability. This is reflected in the strength of evidence for pairwise comparisons which often fell into *limited* or *moderate* strength of evidence on the verbal scale [20]. Then again, inclusion of a segment in the speaker's native inventory does not seem to necessarily impact performance. For instance, compare the overall performance of Mandarin and Arabic: 17% and 9% respectively. Of the four vowels in this study, only the phoneme /i/, or one labelled as such, is shared with the native variety, yet they show very different classification rates. There were also a lot more tokens of /i/ than the other phonemes, meaning better representation of the group.

Surprisingly, many of the varieties were confused with Mandarin and German (Table 2). We suspect this has more to do with the relatively smaller number of speakers compared with the other L1s used to form the training set [30]. The German and Mandarin groups contained 12 and 13 speakers respectively, compared with 39 Arabic speakers. Even with only three features, dimensionality could be an issue.

Low strength of evidence was consistent across all comparisons and suggests that while it is possible to correctly identify a speaker's L1 based on their pronunciation of individual vowel phonemes, the classification based on this modelling method is not accurate or consistent enough. However, it does demonstrate issues that will translate to methods using more opaque measurements (such as MFCCs) in higher dimensions, in particular, large within-group variability and substantial overlap between varieties in the phonetic subspace. Also of note is that speakers were not consistent in their performance across segments. This again highlights the challenge of inherent variability of L2 speakers due to differences in factors such as proficiency.

Many other factors could have further affected the results, including alignment error from the MFA, measurement error from Praat formant tracking, and scarcity of data.

4. CONCLUSION

The lack of strong results leads us to conclude that segments in isolation do not provide enough language-specific information to make good predictions. However, some phonemes work better for some languages (and likely some speakers) than others. So, the way forward is to (a) consider a wider range of features at different linguistic levels, and (b) tailor the choice of linguistic features to the language (this may be determined by a combination of knowledge of the languages and through empirical testing). The most effective features for identifying a language, along with the strength of evidence, depend on the pairwise comparison.

5. REFERENCES

[1] Brown, G., Franco-Pedroso, J., & González-Rodríguez, J. 2021. A segmentally informed solution to automatic accent classification and its advantages to forensic applications. *International Journal of Speech, Language and the Law*, 28(2), 201-232.

[2] Kumpf, K., & King, R. W. 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Fifth European Conference on Speech Communication and Technology*.

[3] Bahari, M. H., Saeidi, R., & Van Leeuwen, D. 2013, May). Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7344-7348). IEEE

[4] Upadhyay, R., & Lui, S. 2018, January. Foreign English accent classification using deep belief networks. In *2018 IEEE 12th international conference on semantic computing (ICSC)* (pp. 290-293). IEEE.

[5] Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y. and Xie, L. 2021, June. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6918-6922). IEEE.

[6] Drummond, R. 2013. The Manchester Polish STRUT: Dialect Acquisition in a Second Language. *Journal of English Linguistics*, 41(1), 65–93. https://doi.org/10.1177/0075424212449172

[7] Gut, U. 2009. *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang.

[8] Rindal, U. 2010 Constructing identity with L2: Pronunciation and attitudes among Norwegian learners of English 1. *Journal of Sociolinguistics* 14, no. 2: 240-261.

[9] Markl, N. 2022, June. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 521-534).

[10] Weinberger, S. 2015. *Speech Accent Archive*. http://accent.gmu.edu/index.php.

[11] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502. doi: 10.21437/Interspeech.2017-1386

[12] Williams, S., Foulkes, P., & Hughes, V. 2023. Analysis of forced aligner performance on L2 English speech. [Manuscript submitted for publication] Language and Linguistic Science, University of York.

[13] Harrison, P. 2022 measureAllFormantsTracked.praat v 1.2 (Praat script)

[14] Silverman, B. W. 2018. *Density estimation for statistics and data analysis*. Routledge.

[15] Aitken, C. G. G., & D. Lucy. 2004. Evaluation of Trace Evidence in the Form of Multivariate Data. *Journal of the Royal Statistical Society. Series C, Applied Statistics* 53 (1): 109–22.

[16] Aitken, C., Taroni, F., and Bozza, S. 2020. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd Ed. Wiley.

[17] Morrison, G. S. 2007. Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. http://geoff-morrison.net/#MVKD

[18] Enzinger, E., & Morrison, G. S. 2017. Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277, 30-40.

[19] Morrison, G. S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 532), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

[20] Champod, C. and Evett, I. W. 2000 Commentary on A. P. A. Broeders (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 6(2): 228–41.

[21] https://github.com/sejw/Williams_ICPhS_2023

[22] Bahrani, N., & Modarresi Ghavami, G. 2021. Khuzestani Arabic. *Journal of the International Phonetic Association*, 51(2), 299-313. doi:10.1017/S0025100319000203

[23] Barbosa, P., & Albano, E. 2004. Brazilian Portuguese. *Journal of the International Phonetic Association*, 34(2), 227-232. doi:10.1017/S0025100304001756

[24] Cruz-Ferreira, M. 1995. European Portuguese. *Journal of the International Phonetic Association*, 25(2), 90-94. doi:10.1017/S0025100300005223

[25] Fougeron, C., & Smith, C. 1993. French. *Journal of the International Phonetic Association*,23(2), 73-76. doi:10.1017/S0025100300004874

[26] Kessler, B., Leben, W. R., & Lyovin, A. 2017. *An introduction to the languages of the world*. Oxford University Press.

[27] Kohler, K. 1990. German. *Journal of the International Phonetic Association*, 20(1), 48-50. doi:10.1017/S0025100300004084

[28] Thelwall, R., & Sa'Adeddin, M. 1990. Arabic. *Journal of the International Phonetic Association*, 20(2), 37-39. doi:10.1017/S0025100300004266

[29] Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press. doi:10.1017/CBO9780511753459

[30] Hughes, V. 2017. Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough? *Speech Communication*, 94, 15-29.

Methodological trends in acoustic phonetic analysis

Emily Grabowski

University of California, Berkeley
emily_grabowski@berkeley.edu

ABSTRACT

Acoustic analysis in phonetics presents an opportunity to quantify meaningful differences in speech sounds. Recently, however, methodological research has raised concerns about the degree of variation in acoustic methodology and its impact on generalizability of the results. While researchers often have intuitions of best practices for selecting appropriate features, there is little quantitative investigation of how acoustic features are used in empirical studies. This paper presents the results of a survey of the methodology in a sample of papers from the proceedings of this conference at two time points 20 years apart, with a focus on the target contrast, parameters used, and measurement choices. The results identify themes in acoustic features across contrast types and identifies potential areas for further development of acoustic parameters.

Keywords: methodology; acoustic analysis; phonetics

1. INTRODUCTION

A central part of designing an acoustic analysis is choosing which phonetic parameters to include. Generally, acoustic correlates are established on a per-contrast basis and researchers discuss previous work when determining appropriate contrasts to use. This results in a significant opportunity for variation in the representation of speech. In addition, the appropriate acoustic features to use have been reported to vary along dimensions such as language [1,2] and speaker [3].

Variance alone is not necessarily an issue - language is a complex social phenomenon and will always have significant inherent variation [4]. However, methodological variation may also contribute to issues in replicability, false positives, and other issues that are currently the subject of discussion in the field [5,6,7].

In addition to broader methodological studies, there have been several investigations into the quality of acoustic features. One study on the validity of acoustic features attempted to relate acoustic features to perceptual and acoustic landmarks [8], while another investigated the variance of formant measurements [9].

In addition to directly establishing the quality of a specific measure, there is value in understanding how acoustic methods are used in practice. While there is opportunity for significant methodological variation, it is not clear to what degree that it is exercised in practice. In addition, very little is quantitatively known about how they are used in practice in the broader community. One notable exception is [10], which surveys the presence of various acoustic correlates and contrasts discussed in phonetic descriptions. This survey finds that a handful of correlates (including formants and voice onset time) are used broadly and that there is an imbalance in coverage between consonants and vowels. In this paper, I describe the results of a survey designed to further explore the usage of measurement methods in acoustic phonetics.

2. METHODOLOGY

The data for this study are drawn from the Proceedings of the International Congress of Phonetic Sciences (ICPhS), a major conference that represents a broad cross-section of phonetics and adjacent fields.

The sample contains all proceedings papers for the years 1999 and 2019. The most recent proceedings available at the time of writing is 2019, and 1999 was chosen to be early enough to predate the recent wave of computational tools to aid phonetic analysis. The sample contained 640 papers in 1999 and 792 in 2019, for a total of 1432 papers. Of these, 395 (155 in 1999, 240 in 2019) were determined to be relevant based on the title and the abstract as pertaining to acoustic analysis of a consonantal, vocalic, or suprasegmental contrast.

Each relevant paper was coded for the broad phonetic category under study: sonorants, obstruents, vowels, and suprasegmentals. Studies that included multiple contrasts or contrasts that crossed these categorical boundaries were coded for both categories. Papers were also coded for the acoustic parameter(s) used to measure the contrast and how each parameter was operationalized.

The next section will present several trends in the methods used in each year for three categories: vowels, sonorants, and obstruents. Laryngeal contrasts are not reported here as they constitute a significantly different research paradigm. First, the

rate of coverage of high-level contrasts (vowels, obstruents, sonorants) was calculated as the percentage of papers in the sample for each year, in order to account for trends in the overall popularity of acoustics. All other analyses are reported on a per-contrast basis, where data are normalized to the number of papers in each contrast category that year.

3. RESULTS AND DISCUSSION

In 1991, 22% of the total papers in the proceedings fell within the sample, while in 2019 the rate increased to 32%. This suggests that there is a general increase in the popularity of acoustic measurements over the time period of the study. This could be due to several different factors, including general increased interest in quantitative measurement in phonetics and access to tools that made taking certain acoustic measurements much more straightforward.

The rate of papers covering each general phonetic category is given in Figure 1. Sonorants and obstruents see an overall downward trend between the years in the study, while vowels have an upward trend. Sonorants are also the least commonly reported, while obstruents saw approximately the same number of papers addressing them as vowels in 1999. However in 2019 there is more spread in the distribution of studies across categories, as vowels increased in representation and both sonorants and obstruents decreased.

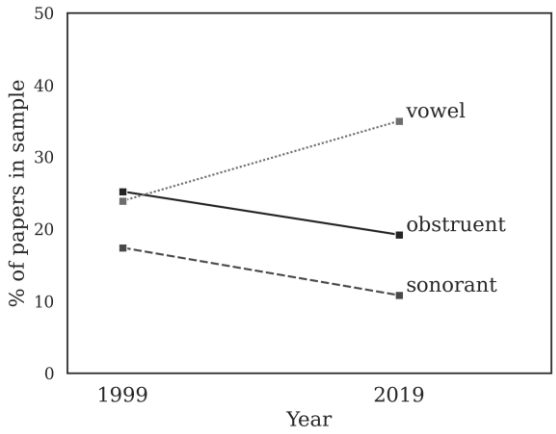


Figure 1: Percentage of papers in each sample addressing major acoustic contrasts.

While some asymmetry in representation of categories can be expected from non-methodological factors such as typological distribution of phonetic categories, those factors wouldn't be expected to influence the trends over time observed in the results of this survey. Rather, these trends over time are more likely influenced by changes in methodological practices in the community. Consonants have a wider range of acoustic variation and the appropriate features must be established for each contrast. On the other hand, measurements for vowels, for example, are quite well understood and there are many tutorials and resources available for automatic or near-automatic analysis of relevant features. This imbalance in accessibility may be helping to encourage a trend towards more vowel analysis and less consonant analysis.

3.1. Usage of qualitative methods of analysis

The trend in phonetics has been to quantify acoustic differences where possible. Quantitative analysis helps to identify the significance and magnitude of differences between sounds and reduces the contribution of perceptual bias to results. Nonetheless, qualitative, impressionistic analysis of a segment is a valuable entry point for investigating contrasts, particularly those that do not have a clearly delimited set of descriptive acoustic features. This suggests that in this sample, a higher rate of qualitative measures for a contrasts indicates a lack of confidence in the ability of existing features to capture the dimensions of the contrast under study.

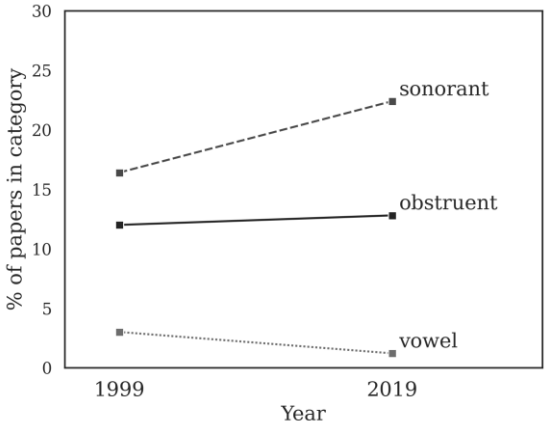


Figure 2: Percentage of papers in each category that utilize qualitative methods

The proportion of papers in each category that used qualitative measures is given in Figure 2. Vowels have the lowest rate of qualitative analyses, followed by obstruents, while sonorants have the highest rate in both years. This pattern aligns with intuitions about the relative confidence in acoustic features that represent each of these categories. Vowels and obstruents do not show much change between samples, but in sonorants there is an increase in the usage of qualitative strategies in 2019. While qualitative analysis is a valuable tool, an increase in rate of this analysis suggests that there is also an increase in uncertainty in the appropriate acoustic parameters to use to represent these segments.

The pattern reported in this section aligns with that identified in the section above: vowels are easier to

analyze quantitatively and sonorants are particularly difficult to analyze given the current set of acoustic features. These trends highlight the potential of establishing features for sonorant consonants for advancing acoustic analysis in phonetics.

3.2. Variance in feature usage and measurement

Aside from overall trends, it is also possible to investigate how the methods and measurements used within each category has changed over time. The distribution of phonetic categories across common types of acoustic features for each year is given in Figure 3, reported as the proportion of papers in each category that utilize a feature type.

Non-vowel analyses have more durational and non-acoustic approaches overall while formants are the most used feature for vowels and sonorants across both years. Formants (in particular formant transitions from the neighboring vowel) were also commonly used for obstruent analyses in 1999 but have decreased in usage. In that category, spectral moments have become the most common acoustic feature type, which appears to be a somewhat recent development.

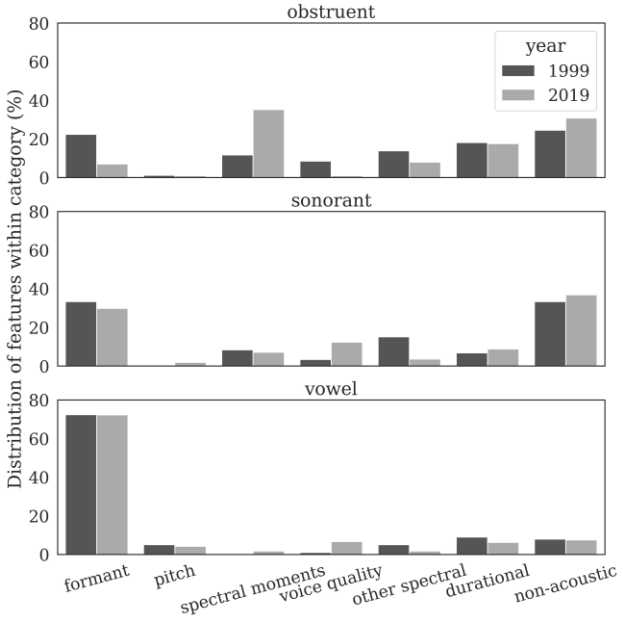


Figure 3: Distribution of sample across acoustic feature categories.

Although there are many possible acoustic parameters that have been proposed for phonetic analysis, in usage there appears to be a convergence on a few common sets of measures. This makes sense given the fact that researchers often build their analysis by reference to previous successful studies on analogous contrasts. While this helps establish a set of best practices for the field, it also makes clear the importance of understanding the limitations of these features given that they can disproportionately impact insights from acoustic analyses. In addition, since the feature used varies by segment type, the proper analysis for contrasts that straddle these categorical boundaries is not immediately clear.

There is also potential variation in the operationalization of these measurements, in particular choices made around where to measure these features within the segment. This is addressed by Figure 4, which gives the distribution of phonetic categories across measurement types reported as proportion of papers in each sample that use each measurement type. Consonants see a move towards point and window (e.g. average value across some portion of the segment) methods, and other more complex statistical formulations, such as formulas, derivatives, etc) are less common in 2019 than in 1999. There is also an increase in trajectories (measurements of multiple time points) used in most categories, which suggests a more general trend towards dynamic measures. There is still some variation between measure types, and it should be noted that there is also significant variation at a finer level of detail within each of these categories, for example where to measure a point within a segment.

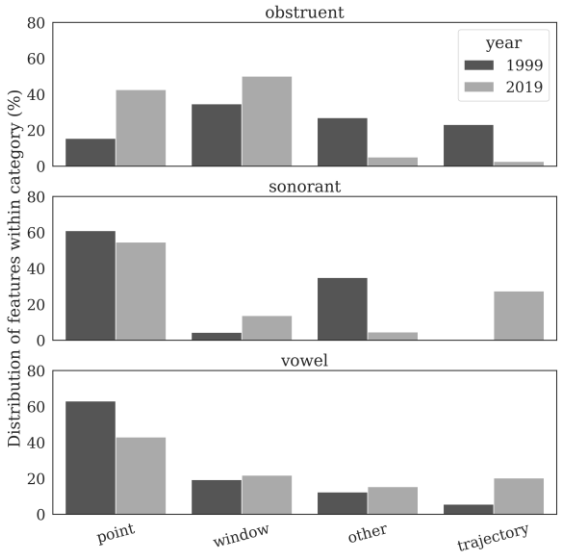


Figure 4: Distribution of sample across measurement strategies.

3.3. Methods across methods of articulation

Up until this point, the phonetic contrasts in this sample were pooled into broader categories. While this is helpful for looking at general trends, it is also informative to consider more fine-grained distinctions. From the convergence on a set of

parameters within the target phonetic categories arises the question of how to analyze data that may cross these boundaries. The properties of this distribution may give insight into the current limitations of acoustic features.

The relative frequency as well as the most used measure for each method of articulation is reported in Figure 5, pooled across years in the study. The diagonal gives the values for studies within that category and the off-diagonal values are those comparing between methods of articulation. The shading of each cell corresponds to the number of papers in that category in the paper. Where there was a tie in the most common method of analysis, both methods are noted.

There is a distinct imbalance in the data in terms of which contrasts are represented, and several potential contrasts have no or very few studies addressing them. While relative frequency may also be affected by phonetic typology, these imbalances may also be influenced by how easy or difficult it is to capture the contrast in numerical terms.

In addition, almost all studies that evaluated values across methods of articulation included either articulatory or qualitative measures in the top methods used to measure them.

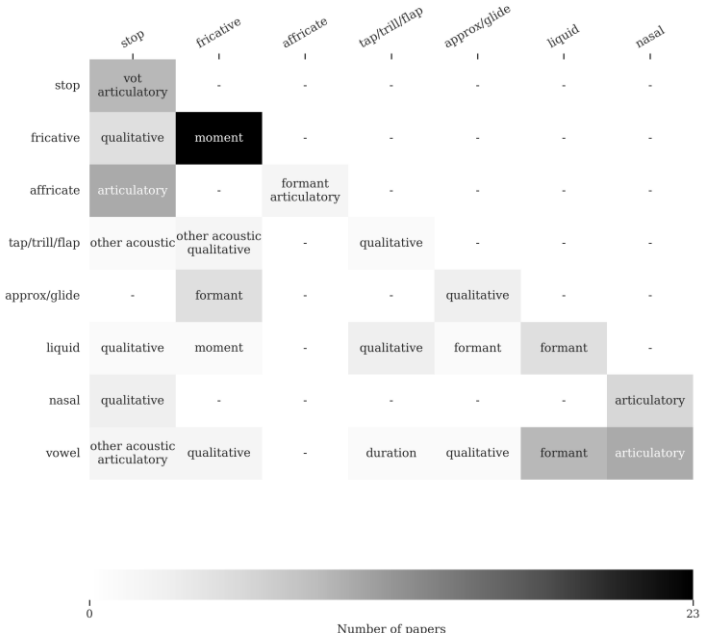


Figure 5: Most common parameters used for comparisons across methods of articulation.

The acoustic measures used most frequently are formants, which are used in both sonorant contrasts and in some obstruent analyses (by way of formant transitions). Formants are a relatively transparent, measurable, and widely-used acoustic measure, and from this analysis it is clear that they have been leveraged to apply to a broad range of phonetic contrasts. However, the interpretation of formant differences still depends on the type of segment being measured, meaning that these measures may not always be directly comparable in different contexts.

The results presented here, including the significant use of non-acoustic measures, suggest that existing measures do not work equally well for all methods of articulation within the general phonetic categories.

4. CONCLUSION

This paper presents some results of a survey of phonetic methodologies in the field. Generally, there is an increase in acoustic studies over the time period in this study. In that time there is an increasing gap in how frequently different categories are represented in the sample, in particular less coverage of consonants and more of vowels. Consonants see more qualitative analyses than vowels, and sonorants in particular see an increase in qualitative measurement between the two study years. There is some convergence on a type of parameter, and to some extent the type of measurement, within each broad phonetic category. However, looking within categories at cross-categorical comparisons found a high rate of non-acoustic measures, suggesting that the common measures for consonants are not richly descriptive enough to capture all contrasts of interest in the study.

These results suggest there is generally less confidence shown in consonant parameters, particularly sonorants: they see a relative decrease in coverage and a higher rate of qualitative description. In addition, consonants see the most non-spectral parameters (such as durational or impressionistic measures) and more shift in parameterizations and measurement methods over the years.

While quantitative analysis has contributed to significant progress in phonetic analysis, this survey demonstrates that there are still questions about best practices in acoustic methodology in these cases, and that falling back on qualitative measures is still quite common particularly for less common contrasts. These observations indicate that there is room for further consideration of acoustic features that can be generalized across contrasts, and further understanding the dynamics of acoustic methodology in phonetics.

5. REFERENCES

[1] T. Cho, D. H. Whalen, and G. Docherty. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages, *Journal of Phonetics* 72, 52–65.

[2] A. Jongman, S. E. Blumstein, and A. Lahiri. 1985. Acoustic properties for dental and alveolar stop consonants: a cross language study, *Journal of Phonetics*, 13(2), 235–251.

[3] K. Amino and T. Arai. 2009. Speaker-dependent characteristics of the nasals, *Forensic Science International*, 185, 21–28.
[4] Grieve, J. 2021. Observation, Experimentation, and Replication in Linguistics. *Linguistics* 59 (5), 1343–56.
[5] Roettger, T. 2019. Researcher Degrees of Freedom in Phonetic Research. *Laboratory Phonology* 10 (1).
[6] Sönning, L., Werner, V. 2021. The Replication Crisis, Scientific Revolutions, and Linguistics. *Linguistics* 59, no 5, 1179–1206.
[7] Winter, B., Grice, M. 2021. Independence and Generalizability in Linguistics. *Linguistics* 59 (5), 1251–77.
[8] Gradoville, M.S. 2011. Validity in measurements of fricative voicing: evidence from Argentine Spanish. *Proc. 5th Conference on Laboratory Approaches to Romance Phonology*, 59-74.
[9] Kendall, T and C. Vaughn. 2015. Measurement variability in vowel formant estimation: a simulation experiment. *Proc 18th ICPhS* Glasgow.
[10] Whalen, D. H., Dicanio, C., and R. Dockum. 2022. Phonetic Documentation in Three Collections: Topics and Evolution. *Journal of IPA* 52, no. 1, 95–121.

IMPACTS OF REMOTE AND TRADITIONAL RECORDING METHODS ON QUALITY OF FORMANT ANALYSIS FOR VOWEL REDUCTION

Jenna T. Conklin

Carleton College
jconklin@carleton.edu

ABSTRACT

Past work on recording quality of remotely-collected speech data has demonstrated that some remote methods are adequate for broad-strokes formant analysis (such as comparing locations of phonemes in the vowel space), but none are suitable for fine-grained formant analysis, such as sociolinguistic work. This study evaluates the adequacy of two remote recording techniques for formant analysis of vowel reduction, which requires an intermediate degree of specificity between these two extremes. It compares lossy recordings captured remotely via Gorilla, lossless smartphone recordings self-administered by remote participants, and traditional recordings created in-person in a sound booth. Results show only minimal variation across recording methods, indicating that remote recording via Gorilla or self-administered smartphone recordings may prove viable options for remote data collection for similar studies, although the convenience of remote recruitment and recording are offset by higher rates of data loss and difficulties in annotation and formant extraction compared to traditional methods.

Keywords: vowel reduction; remote recording; formant analysis

1. INTRODUCTION

Researchers who wish to collect speech data remotely face an array of challenges; variation in compressive algorithms, hardware and environment, and sampling rates conspire to create an unpredictable landscape capable of exerting unequal impacts on the resultant recordings [1]. While it has long been known that variation in recording device and environment can lead to inconsistent acoustic measurements of speech [2], the body of research on the quality of remote recording of speech data has greatly expanded in recent years in response to the COVID-19 pandemic. With regard to the impact of remote data collection on analysis of vowel formants, studies have examined the effect of recording via videoconferencing apps [3]–[5] and varied hardware arrangements [1], [3], [6], [7], largely focusing on the degree of acoustic divergence from traditional recording setups. Many of these studies focused either on large-scale differences, such as the relative arrangement of vowels, or extremely nuanced differences, building a general consensus that analysis of large or relative formant differences can be carried out remotely under the right conditions, while work examining minor shifts will benefit from a traditional recording setting [5], [6]. The current study considers the effect of remote recording on an analysis of vowel reduction, a phenomenon that demands a degree of precision between these two extremes, by comparing data taken in a laboratory setting to that from two remote setups.

1.1. Sources of Variation in Remote Recording

Obstacles to obtaining high-quality recordings remotely are numerous, and only some are within researchers' control. Loss of data due to background noise or microphone placement, for instance, can be offset by providing clear instructions, but even with participants' best efforts, some disruption is likely. One source that can be controlled is the recording format: researchers must choose between lossy audio formats, which reduce file size by selectively deleting data, and lossless formats, which accept larger file sizes in exchange for more faithful recordings. Remote recordings taken online in real time, such as those obtained through Zoom, Skype, Teams, and Gorilla, are lossy, and some distortions to the vowel space relative to traditional recording are to be expected [3]–[5], although [4] was able to correct for these changes through Lobanov normalization. At least one study of remotely-collected speech data quality utilizing lossless recording, wherein participants recorded themselves on a device offline and then uploaded the file, found little difference in formant values from traditional recordings [3], suggesting that offline smartphone recordings might be the most viable alternative to in-person data collection for work considering suitably large effects.

The hardware used for a remote study will also impact the formant values of resultant recordings. [6] examined data recorded using tablets, smartphones, laptops, and a more traditional microphone and found that laptops offered fewer discrepancies than tablets and smartphones, male vowel spaces underwent less distortion than female ones, and the 750-1500 Hz range was especially likely to be distorted, often

affecting the low back vowels. Similarly, [7] compared formant values of Chinese vowels collected simultaneously across seven devices in two different settings (lab and conference room) and found that F1 and F2 values exhibited some systematic differences across devices, with greater variation in F2 compared to F1. These results suggest that remote subjects' personal hardware should be suitable for formant analysis, provided the study is focused on large distinctions or relative arrangements of vowels, but perhaps not for fine-grained sociolinguistic analysis. Researchers should expect some variation to be introduced by the model of the recording device, and should be aware that direct comparison of formant values across recording devices is likely to lead to misleading conclusions.

2. METHODS

2.1. Recording Methods & Subjects

An identical task was administered to two groups of participants, one remote (6M, 4F; M=35.6 yo, SD=10.65), one in-person (2M, 8F; M=20.9 yo, SD=1.96). All subjects were monolingual speakers of Midwestern American English. The in-person group completed the task in a sound-insulated booth wearing Sennheiser HD 380 Pro headphones, the stimuli were presented using PsychoPy [8], and their speech was recorded with a Shure KSM32 cardioid condenser microphone attached to a TubeMP preamp and digitized at 44.1 kHz. Data from the in-person group also served as a control for a separate study on bilingual vowel reduction (currently in preparation).

The remote group was recruited via Prolific [9] and completed the task over Gorilla [10]. One set of recordings was taken using Gorilla's Audio Recording Zone to generate .weba files using whatever microphone participants had for their computer. (.weba is a lossy format created with the OGG Vorbis compression codec [11].) A second, simultaneous recording was created by remote participants on their smartphone; this recording was a lossless .wav file recorded via a free app (Hokusai Audio Editor for Apple users and ASR Voice Recorder for Android owners). (One participant encountered a microphone error of some kind on their computer, and thus contributed only a smartphone recording and no Gorilla recording, leaving nine subjects in the Gorilla group.) Participants adjusted the app settings to at least a 128 kbps bitrate (Apple users selected an even higher-quality 16-bit setting) and a 44.1 kHz sampling rate. They were instructed to place both the smartphone and computer microphone in a stable position 6 – 10 inches from their mouth, or, in the case of a laptop-internal microphone, to sit at a comfortable distance from the screen. When the task was complete, subjects uploaded the smartphone recording to the researcher.

2.2. Task & Procedure

Participants completed a shadowing task in which they heard and repeated a sentence containing a target word with one of five vowels /ɑ, æ, ɛ, ɪ, ʌ/ in stressed or unstressed position. Stimuli were read by a male native speaker of Midwestern American English, recorded in the same manner as the in-person participants described in §2.1. All stressed vowels appeared in monosyllables, and unstressed vowels appeared in disyllables. Each target word had a counterpart in the study with identical structure for the target syllable, but different stress (i.e., bit ~ rabbit, text ~ context). Sixty target words (six for each stressed vowel and six for each unstressed vowel) were included; each subject repeated each target word twice. Target words were embedded in a unique carrier sentence of the form "the word X means…" where the phrase following "means" was relatively consistent in length, rhythm, and complexity.

2.3. Analysis

The beginning and end of the vowel were annotated by hand in Praat [12] and formant values were extracted at vowel midpoint using an LPC-based Praat script that allowed for manual review [13]. If the researcher observed that the automatic formant reading did not match the visual formant, a manual reading was substituted, taken by placing the cursor in the approximate center of the visible formant. Extracted formants were normalized using log-additive regression normalization [14]. The Euclidean distance for each pair of stressed & unstressed target utterances (i.e., each speaker's first repetition of bit ~ rabbit, each speaker's second repetition of bit ~ rabbit , etc.) was calculated to provide a measure of the degree of centralization for unstressed vowels, calculated as shown in (1).

$$(1)\quad EuD = \sqrt{(F1_{V1} - F1_{V2})^2 + (F2_{V1} - F2_{V2})^2}$$

3. RESULTS

3.1. Euclidean Distance

Euclidean distance provides a measure of the distance between two points: in this case, the degree of distance in F1xF2 space between a stressed vowel and its unstressed counterpart (see (1)). A linear mixed-effects model was fit in R [15] to evaluate the degree to which recording method impacted Euclidean

distance using *lme4* [16] and *lmerTest* [17]. The best-fitting model contained a response variable of Euclidean distance, fixed effects for Vowel (containing five levels /ɑ, æ, ɛ, ɪ, ʌ/) and Recording Method (with three levels, Gorilla, Smartphone, and In-Person), as well as an interaction term for Vowel by Recording Method and random intercepts for Subject and Item. Model results showed that the recordings taken via Gorilla differed significantly from those taken in-person (β = -.088, SE = .031, t = -2.865, p < 0.01), while the smartphone recordings did not differ significantly from the in-person data (β = -.047, SE = .030, t = -1.549, p = .131).

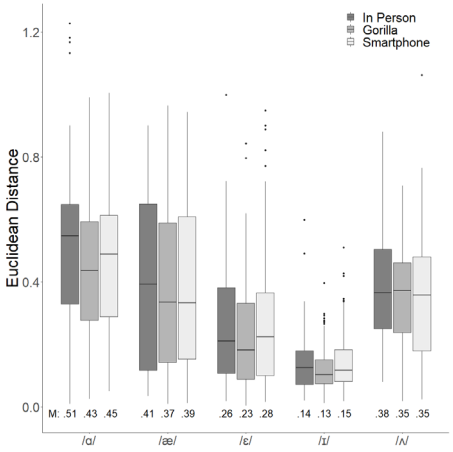


Figure 1: Euclidean distance by Vowel and Recording Method

Figure 1 displays the Euclidean distance of each vowel and recording method. As can be seen, the distribution of Euclidean distance is similar across recording methods for each vowel, but the Gorilla recordings diverge somewhat from the in-person data for vowels /ɑ/ and /ɛ/, while the smartphone recordings track more closely with the in-person data.

3.2. F1

To gain a more complete understanding of the variation in formant values across conditions, normalized formants were also examined. A linear mixed-effects model was fit with normalized F1 as a response variable; the best-fitting model included fixed effects for Vowel, Stress, and Recording Method, interaction terms for Vowel by Recording Method, Stress by Recording Method, and Vowel by Stress, and random intercepts for Subject and Item. Here, the results taken via Gorilla did not differ significantly from the in-person reference data (β = -.019, SE = .019, t = -.978, p = .334), although the uploaded data did (β = -.044, SE = .019, t = -2.306, p < .05). Figure 2 demonstrates that some of the divergence is shared by both the Gorilla and smartphone data (e.g., stressed /æ/) and thus was likely due to between-group differences, while other differences are limited only to the smartphone data and thus must be attributed to distortion introduced by the hardware or user behavior (i.e., microphone placement), as seen in several of the unstressed vowels. The smartphone recordings often exhibit a lower F1 than the Gorilla or in-person data, although this is not consistent across all vowels.

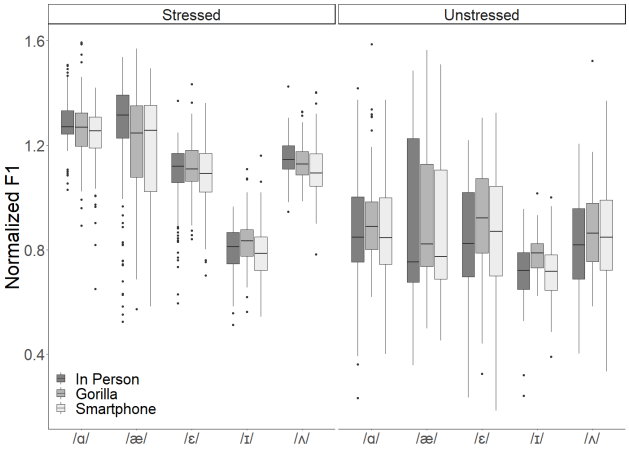


Figure 2: Normalized F1 by Vowel, Stress, and Recording Method

3.3. F2

A third linear mixed-effects model took normalized F2 as a response variable; the best-fitting model also included fixed effects for Vowel, Stress, and Recording Method, all possible interactions among these, and random intercepts for Subject and Item. Both the Gorilla and smartphone recordings differed from the in-person data in F2 (Gorilla: β = .056, SE = .019, t = 3.103, p < .01; smartphone: β = .052, SE = .019, t = 2.873, p < .01). As shown in Figure 3, stressed /ɑ/ and /æ/ had a higher F2 in the remote recordings than the in-person data, while other vowels showed minor shifts in either direction across methods, though F2 was higher in remote recordings for most vowels.

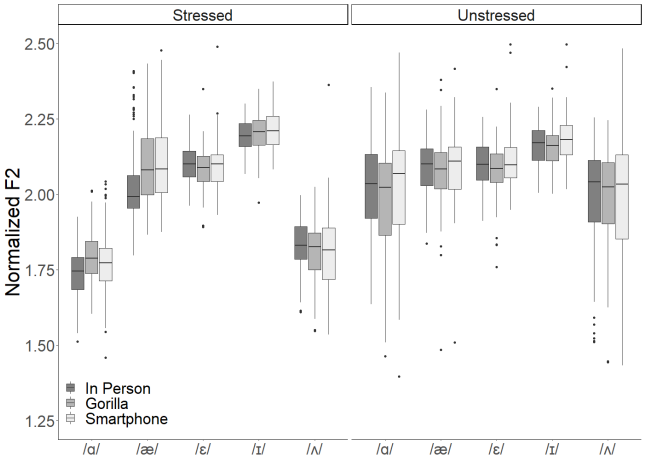


Figure 3: Normalized F2 by Vowel, Stress, and Recording Method

3.4. Pillai scores

Neither Euclidean distance nor direct comparison of formant values provides a measure of the degree of overlap between categories; to understand how distinct a stressed vowel and its unstressed analog are, the Pillai score is informative [18], [19]. Pillai scores closer to 0 indicate a great deal of overlap in the two categories, while scores approaching 1 mark that the two categories have little overlap. Pillai scores comparing each stressed and unstressed vowel for each speaker were calculated and are displayed in Figure 4; additionally, a linear mixed-effects model with a response variable of Pillai score, a random effect of Subject, and fixed effects of Vowel and Recording Method was fit. There was no statistically significant difference between the Pillai scores from the in-person recordings and either the Gorilla (β = -.01, SE = .048, t = -.216, p = .831) or smartphone (β = -.07, SE = .048, t = -1.495, p = .149) recordings. Thus, the degree of overlap between reduced and unreduced instances of the five vowels examined was not meaningfully impacted by switching from a traditional in-person recording setup to either of the two remote recording designs tested.

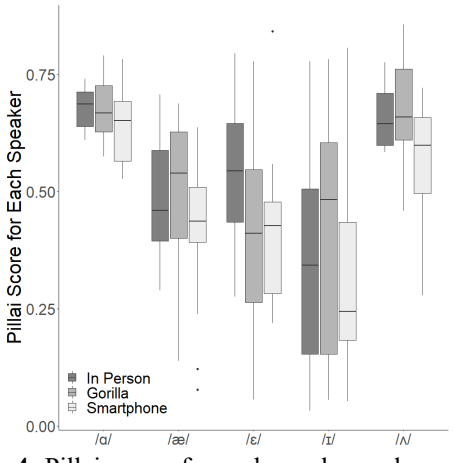


Figure 4: Pillai scores for each speaker and vowel pair

4. DISCUSSION AND CONCLUSIONS

The present study analyzed the reduction of English vowels using three recording methods: one group completed the task in a traditional laboratory setting, and a second group contributed two simultaneous remote recordings, one in a lossy format recorded on a laptop or desktop computer via Gorilla and the second in a lossless format recorded via smartphone. This design intentionally conflated between-speaker differences and variation due to remote participants' individual devices, background noise, and microphone placement to determine whether the results of the study would materially differ based on the researcher's choice to collect data remotely or in person. Analysis of normalized formant values across recording conditions reflected the findings of earlier studies: some variation was detected between the in-person and remote recordings, with the most notable deviance found in low or back vowels. Measures of vowel reduction showed minimal variation across recording methods: no statistically detectable difference emerged when reduction was quantified through Pillai scores, and when Euclidean distance was used, the Gorilla recordings but not the smartphone recordings differed significantly from the traditionally-recorded data. Given that the Gorilla recordings were encoded in lossy .weba files, it is not surprising that they would show greater deviance from the in-person data. The most obvious changes were in low or back vowels, which accords with the findings of [6].

The relative success of the smartphone recording group at replicating the results of an in-person study may be attributed to a handful of causes. First, vowel reduction is both a relative phenomenon, in that any analysis of it must quantify the degree of change of a vowel according to its prosodic prominence, rather than its absolute position in the vowel space. Relative measures are naturally more forgiving than absolute ones when searching for differences across platforms, since direct comparison across recording conditions is not required. Secondly, the scale of difference needed to effect change across groups is relatively large for vowel reduction in comparison to sociolinguistic studies, where the aim of the analysis is to capture the most minute of differences. Finally, it is possible that the normalization procedure served to mask some differences that otherwise would have been salient. (Formants were normalized using a single category for each vowel across recording conditions to ensure the resultant values were on a single scale.) In closing, the present results indicate that remotely-collected lossless smartphone recordings may be a viable alternative for work relying on relative formant values or expecting relatively large shifts in formants across conditions. Recordings taken via Gorilla exhibited greater distortions and should not be used for formant analysis. As others have noted, researchers planning remote data collection should prepare for higher rates of data loss and background noise, and should adjust their recruitment targets accordingly.

5. ACKNOWLEDGEMENTS

Thanks are due to Sophia Chuen for their assistance in collecting and analyzing data and to Josh Weirick and Diego Luna Hernandez for contributing their time and voices to the recording of the audio stimuli. This study was supported by the Headley Fund at Carleton College.

6. REFERENCES

[1] C. Sanker *et al.*, "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," *Language*, vol. 97, no. 4, pp. e360–e382, 2021, doi: 10.1353/lan.2021.0075.

[2] P. De Decker and J. Nycz, "For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?," *Univ. Pa. Work. Pap. Linguist.*, vol. 17, no. 2, Jan. 2011, [Online]. Available: https://repository.upenn.edu/pwpl/vol17/iss2/7

[3] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 3910–3916, Jun. 2021, doi: 10.1121/10.0005132.

[4] J. Calder *et al.*, "Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis," *Linguist. Vanguard*, Feb. 2022, doi: 10.1515/lingvan-2020-0148.

[5] V. Freeman and P. De Decker, "Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps," *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 1211–1223, Feb. 2021, doi: 10.1121/10.0003529.

[6] V. Freeman and P. De Decker, "Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices," *Lang. Linguist. Compass*, vol. 15, Jul. 2021, doi: 10.1111/lnc3.12435.

[7] C. Ge, Y. Xiong, and P. Mok, "How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements," in *Interspeech 2021*, Aug. 2021, pp. 3984–3988. doi: 10.21437/Interspeech.2021-1122.

[8] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019, doi: 10.3758/s13428-018-01193-y.

[9] "Prolific." London, UK, 2022. [Online]. Available: https://www.prolific.co

[10] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[11] "What is a WEBA file?" Accessed: Jan. 06, 2023. [Online]. Available: https://docs.fileformat.com/audio/weba/

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." Nov. 03, 2020. [Online]. Available: http://www.praat.org/

[13] R. Scarborough, "Supervised Formant Reading Script." Jul. 2005.

[14] S. Barreda and T. M. Nearey, "A regression approach to vowel normalization for missing and unbalanced data," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. 500–520, 2018.

[15] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest: Tests in linear mixed effects models." 2015. [Online]. Available: https://cran.r-project.org/web/packages/lmerTest/index.html

[18] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *J. Phon.*, vol. 34, no. 4, pp. 458–484, Oct. 2006, doi: 10.1016/j.wocn.2005.10.001.

[19] J. Nycz and L. Hall-Lew, "Best practices in measuring vowel merger," presented at the 167th Meeting of the Acoustical Society of America, Providence, Rhode Island, 2014, p. 060008. doi: 10.1121/1.4894063.

ACOUSTIC CORRELATES OF THE JAVANESE LIGHT VS. HEAVY DISTINCTION: A LARGE-SCALE CORPUS STUDY

S.C. Angela Xu, Colin Wilson

Johns Hopkins University
angela.xu@jhu.edu, colin.wilson@jhu.edu

ABSTRACT

According to the influential *continuum model* of phonation, only voiced segments can be specified as creaky or breathy. The present study investigated many possible phonetic correlates of the laryngeal contrast in Javanese word-initial prevocalic stop consonants, drawing upon a spoken corpus of more than 180,000 utterances. The results indicate that the laryngeal contrast is cued by voice onset time (VOT) and several acoustic-phonetic properties of the following vowel, including the first formant (F1) in addition to voice source measurements such as H1*-H2* and cepstral peak prominence (CPP). Taken together these findings indicate that Javanese stops can be both voiceless and breathy, supporting a revision of the continuum model in which voicing and other aspects of phonation are decoupled.

Keywords: laryngeal contrast, acoustic correlates, phonation type, Javanese, corpus phonetics.

1. INTRODUCTION

Javanese is a relatively understudied Malayo-Polynesian language of the Austronesian family with approximately 100 million speakers, mostly concentrated on the island of Java in Indonesia [1]. Several dialects of Javanese have been distinguished in the literature. Javanese also has several speech levels, including Krama/Basa (formal) and Ngoko (less formal). This study examines the speech of individuals recorded in Yogyakarta, Indonesia, and focuses on the Ngoko level that is typical of informal settings (e.g., [2]).

The laryngeal contrast in Javanese stop consonants has proven difficult to characterize. The IPA symbols /p t ʈ k/ and /b d ɖ g/ have been used to represent the stops, suggesting a contrast in phonetic voicing. However, previous literature has referred to the two groups of stops in various ways, indicating a contrast that is not, or at least not only, a difference in the presence and timing of vocal fold vibration. Here we follow a tradition of using the terms 'light' vs. 'heavy' ([3, 2, 4, 5]), while others have described the contrast as 'intensive' vs. 'non-intensive' ([6]), 'clear' vs. 'breathy' ([7]), 'tense' vs. 'lax' ([8, 9]), 'stiff' vs. 'slack' ([10, 11, 12]), or 'voiceless' vs. 'voiced' ([13]).

The Javanese laryngeal contrast can be considered in the context of the influential *continuum model* of phonation types, first proposed in [14] and later revised in [10, 15]. Fig. 1 gives a visual schematic of the voicing and phonation terms and their relation with the states of the glottis, as described in [10]. Notice that 'voiceless' refers to lack of vocal fold vibration regardless of glottal aperture. According to this model, classifying a stop as one of the five phonation types (creaky through breathy) implies that it is voiced.

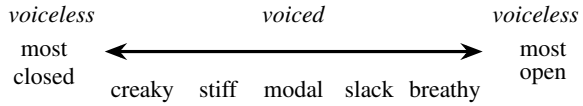


Figure 1: Voicing and phonation types as described in the continuum model.

Problematically for the continuum model, [16] found that glottal stops are frequently (partially) voiced and that the phonetic difference between the so-called 'voiceless' glottal stop and 'voiced' creaky sounds is not one of voicing. The authors hence argued that the model should be modified so that voicing is unspecified for glottal consonants. The laryngeal contrast in Javanese stops may motivate a similar decoupling of voicing and phonation.

The rest of this section reviews previous studies (from the last 20 years) on the acoustic correlates of the Javanese contrast, focusing on cross-linguistically common acoustic-phonetic correlates of voicing and phonation contrasts. For voicing, common cues include a delay of vocal fold vibration in a following sonorant or vibration during closure, as measured by positive and negative voicing onset time (VOT) of stops, as well as the duration and fundamental frequency (f0) of neighboring vowels. For phonation, [17] examined the phonetic space of vowel contrasts with data from 11 different

languages, and concluded that the most informative parameters include the formant-corrected amplitude difference between the first and second harmonics (H1*-H2*), cepstral peak prominence (CPP), the subharmonic-to-harmonic ratio (SHR), and the harmonic-to-noise ratio in the range 0-500 Hz (HNR05), with amplitude difference between the first harmonic and the harmonic closest to the first formant (H1*-A1*), strength of excitement (SoE), and root-mean-squared (RMS) energy also being informative.

[11] recorded 24 words from one Javanese speaker to compare word-initial labial and velar stops /p b k g/ before the vowels /a o u/. Each word was produced twice in isolation. The fundamental frequency (f0), first formant (F1), second formant (F2), and amplitudes of H1, H2, F1, and F2 were measured in post-stop vowels. The results showed negative H1-H2 and H1-F1 values for vowels after heavy consonants (which [11] refers to as 'slack'); this would indicate modal rather than slack or breathy voice. However, H1-F2 values were mostly positive (except for /a/ after light stops) and significantly higher after light stops, which would support the categorization of heavy consonants as slack. F1 was lower after heavy stops for /a o/, and F2 was consistently higher after heavy stops for all three vowels. (See [11] for a review of earlier investigations by Fagan [4] and Hayward [18, 19].)

[8] compared the artication of Javanese light and heavy stops in fiberoptic recordings of two speakers (using data originally collected by [20]). The epiglottal width was significantly greater for heavy ('lax') stops regardless of place of articulation. To the extent that this indicates larynx lowering, it could to some extent explain the longer VOT of heavy stops and their effects on f0, F1, and voice source measures in following vowels.

[12] claimed that the Javanese contrast is one of phonation (i.e., stiff vs. slack) on the basis of acoustic measurements in stop-vowel and stop-lateral-vowel productions from one speaker. Heavy consonants induced lower f0 and formant frequencies, and higher H1-H2 and H1-F2 values, supporting their characterization as slack. These effects varied to some extent across vowel qualities, but importantly persisted through intervening lateral approximants. As [12] suggests, this is consistent with the presence of contrastive phonation on Javanese stops that is spread to the following vowel.

[21] described the stop contrast as 'tense' vs. 'lax' and investigated both the acoustic correlates of this contrast and whether it is neutralized word-finally. Measurements of VOT, f0, F1, F2, the bandwidth of the first formant (B1), H1*-H2*, and CPP were obtained from recordings of 28 speakers who produced words in a carrier sentences. The results showed that f0, F1, and CPP were lower, while B1 was higher, in vowels following heavy ('lax') stops; underlying light and heavy stops were not distinguished at the end of the word. Word-initial heavy stops had negative VOT for some but not all speakers. These findings support the claim that heavy stops are breathier than their light counterparts, a phonation contrast that exists before sonorants but not word-finally (see also [2]).

Lastly, [13], stating the contrast in Javanese stops as one of voicing, recorded a list of words mainly obtained from [22] produced in isolation by one speaker. Measurements were taken for stop VOT, and duration, f0, F1, F2, H1-H2, H1-A1, and H1-A2 of the following vowel. The results showed that in word-initial context both light and heavy stops have positive VOT, suggesting that all of the stops are voiceless, with longer values for heavy /b d g/ than light /p t k/. Among other findings, measurements of F1 in following /i e u/ agreed with previous reports that this is a reliable correlate of the laryngeal contrast.

The present study aimed to provide further evidence about the phonetic nature of the laryngeal stop contrast in Javanese by examining several possible acoustic correlates in a large-scale speech corpus. We first describe the corpus and the measurements that were considered, then report the statistical analysis of each measure, and conclude with a discussion of how the findings bear on phonation in Javanese and cross-linguistically.

2. METHODS

2.1. Corpus description

In this paper, we report acoustic-phonetic analyses of a Javanese speech corpus originally provided by Google for the purpose of training automatic speech recognition systems (henceforth Google ASR; [23]). The corpus was collected by Google in collaboration with the Javanese Literature Department of Universitas Gadjah Mada (UGM) in Yogyakarta, Indonesia, who trained local volunteers to make scripted recordings from native speakers using custom data collection tools.

The written recording prompts were taken from open online resources and used the modern Javanese writing system based on the Latin alphabet. The corpus contains approximately 185,000 utterances from more than 1000 speakers, with an average utterance length of about ten words. Demographic

information, such as dialect of Javanese and additional language background, was not provided. The audio was recorded in 16-bit linear PCM with a sample rate of 16kHz. A pronunciation lexicon containing all words from the corpus with their phonemic transcriptions was generated using a grapheme-to-phoneme (G2P) model.

2.2. Corpus preparation

We selected the 50 speakers who had the most recorded utterances in the data set. After listening to sample utterances from each speaker, we decided to exclude 6 of them for reasons such as significant background noise. For the remaining speakers (44), audio files were aligned to their transcriptions using custom acoustic models trained with the Montreal Forced Aligner (MFA, version 2.0.0b3) ([24]).

2.3. Acoustic analysis

The release of each word-initial, prevocalic stop consonant /p t k b d g/ (omitting rarer /ʈ ɖ/) and the onset of voicing in the following vowel were located with AutoVOT ([25]). The difference between these two time points provides an automatic measure of positive VOT. A separate AutoVOT model was trained for each stop from a minimum of 30 hand-labeled word tokens beginning with the target consonant followed by various vowels. In each of the forced-aligned textgrids, a new tier was added with regions marked within which AutoVOT searched for the point of stop release and the following onset of voicing. The start of each region was at the midpoint of the closure of the word-initial stop and the end was at the midpoint of the following vowel. VOT values for all tokens were extracted using the trained models.

The duration, f0, F1, and F2 of vowels following the word-initial stops were extracted using Praat ([26]) with the FastTrack plugin ([27]). Voice quality measurements, including H1*-H2*, H1*-A1*, CPP, SHR, HNR05, SoE, and RMS energy, were made for the same vowels using VoiceSauce ([28]) with 1 ms frame shifts. FastTrack and VoiceSauce were applied with their standard settings for all speakers. The output of FastTrack provides f0 and formant values for 5 equal-duration time intervals in each token; we averaged the outputs of VoiceSauce in the same intervals and analyze only the central interval here. Frequencies were converted to the mel scale, which is more closely aligned with human auditory perception than traditional frequency values (Hz).

Outliers were identified separately for each acoustic measure. For all measures except vowel formants, the raw data were grouped by speaker and stop consonant; for F1 and F2 the data were grouped by speaker and vowel phoneme (/i e a o u ə/). Within each group, data points 2 standard deviations beyond the mean were taken to be outliers and excluded from data summaries and statistical analyses.

3. RESULTS

Means and standard deviations of F1 and F2 for vowels after word-initial stops are provided in Table 2, with the other measures summarized in Table 1. We first averaged within each speaker and then computed the summary statistics across speakers. Separate linear mixed-effects models were conducted to analyze the influence of the light vs. heavy stop contrast on each measure.

The model for VOT included fixed effects of stop laryngeal specification (light vs. heavy), stop place (labial, coronal, or dorsal), and their interaction, as well as a random intercept and slope of the laryngeal effect for speaker. All fixed factors in this and subsequent models were entered with weighted effect coding. The main effect of laryngeal specification was significant and positive ($\beta_{lar} = 3.85$), indicating that heavy consonants have slightly longer VOT values than light consonants. The main effect of place was also significant, with dorsal stops having significantly longer values ($\beta_{dor} = 9.17$) and, surprisingly, coronals being significantly shorter overall ($\beta_{cor} = -4.50$). The interaction between laryngeal and place distinctions was also significant; post-hoc tests revealed a reliable effect of light vs. heavy effect on the VOT of coronal and dorsal but not labial stops (significance here and throughout is determined by $p < .001$). The same pattern was found in an analysis of log-transformed VOT.

These results are in line with previous reports that the heavy stops /d g/ have longer VOTs than their light counterparts /t k/ ([13]). The lack of a significant difference for heavy /b/ vs. light /p/ could have two explanations. First, labial stops may have frequently had weak bursts that were not detected by AutoVOT. Second, AutoVOT enforces a minimum VOT duration, here set to 5 ms, that may have artificially inflated the measured values for /p/.

The mixed-effects models for the other measures in Table 1 had the same structure. The results indicate a significant effect of light vs. heavy on f0 ($\beta_{lar} = -6.68$), H1*-H2* ($\beta_{lar} = 1.63$), and SHR ($\beta_{lar} = 0.091$) of the following vowel for stops at all three places of articulation. There were also significant effects on log duration and CPP of the

Stop	N	VOT(ms)	Duration(ms)	f0(mel)	H1*-H2*(dB)	CPP(dB)	SHR(dB)
p	4824	9.5 (2.7)	91.2 (17.1)	284.2 (58.4)	4.8 (3.0)	20.8 (2.5)	0.6(0.2)
b	4327	9.7 (2.8)	90.8 (15.9)	275.6 (54.5)	6.7 (3.2)	20.3 (2.3)	0.7 (0.1)
t	5485	11.0 (2.9)	80.0 (14.9)	297.8 (59.8)	4.6 (4.0)	21.9 (2.8)	0.5 (0.3)
d	8464	17.3 (4.6)	86.8 (17.8)	283.9 (57.8)	6.7 (3.0)	20.2 (2.4)	0.7 (0.2)
k	13653	21.4 (4.3)	87.5 (18.2)	288.4 (57.8)	4.4 (3.3)	21.3 (2.2)	0.6 (0.2)
g	1481	31.3 (7.7)	77.4 (15.0)	277.1 (54.5)	7.8 (3.1)	18.6 (2.2)	0.7 (0.1)

Table 1: Acoustic measures of word-initial stops and following vowels.

following vowel, but these held for coronal and dorsal stops only; the same pattern was found for analyses of H1*-A1*, HNR05, SoE, and RMS energy (not shown in the table). The signs of the effects on the voice quality measures are in line with the expectation from previous literature that heavy stops are more breathy than light stops in Javanese. For example, H1*-H2* is larger (i.e., there is greater low-frequency spectral tilt) following heavy stops.

Stops	F1 (mel)	F2 (mel)
	high V	*front V*
/p t k/	539.0 (51.4)	1654.0 (127.1)
/b d g/	496.9 (45.5)	1675.4 (140.8)
	mid V	*central V*
/p t k/	706.2 (62.9)	1366.4 (76.2)
/b d g/	583.6 (51.3)	1413.9 (80.2)
	low V	*back V*
/p t k/	863.7 (78.6)	1038.4 (57.5)
/b d g/	757.0 (69.5)	1205.5 (80.1)

Table 2: F1 and F2 of vowels following word-initial stops.

The separate models for vowel formants included fixed effects of stop laryngeal specification, vowel height for F1 and vowel backness for F2 (as in Table 2), and their interaction. Each model had a random intercept and slope of the laryngeal effect for speaker. F1 was significantly lower after heavy stops ($\beta_{lar} = -60.3$), consistent with larynx lowering ([8]) or other pharyngeal expansions. The laryngeal and height factors interacted significantly, indicating that low and mid vowels show a larger light vs. heavy difference than high vowels, but post-hoc tests established that the laryngeal effect was significant at all three vowel heights.

F2 was significantly higher after heavy stops ($\beta_{lar} = 47.3$), perhaps suggesting some degree of vowel fronting (see also [11]). The laryngeal and vowel backness factors interacted significantly, with non-front vowels showing larger effects of light vs. heavy than front vowels, but the effect was nevertheless significant and in the same direction for all levels of backness.

4. CONCLUSION

This study investigated the phonetic nature of the laryngeal contrast in Javanese by measuring many possible acoustic-phonetic correlates in more than 36,000 tokens of word-initial stops and following vowels. The results broadly support previous findings that stops of both series have positive VOT in this context, that VOT values are larger for heavy than for light stops (with the possible exception of labials), that voice source measures such as H1*-H2* indicate breathiness on vowels following heavy stops, and that vowels have lower F1 and higher F2 in the same context (see [11, 8, 12, 21, 13]).

According to the continuum model ([14, 10, 15]), only voiced segments can be breathy. Our findings, like those of many previous acoustic studies of the laryngeal stop contrast in Javanese, appear to contradict this claim. Both the light stops /p t k/ and the heavy stops /b d g/ have positive VOT values, consistent with the hypothesis that they are all phonetically voiceless. At the same time, there were also significant differences in voice source measurements such as H1*-H2* and SHR within following vowels, and these are consistent with the hypothesis that the heavy series is breathy.

Along the lines of [16], we suggest that phonation specifications such as slack or breathy are not strictly confined to voiced segments. Indeed, voiceless aspirated stops require the glottis to be abducted, and if this gesture persists after stop release it could naturally induce a breathier voice quality on the following vowel. Because measurements other than VOT were taken in the central interval of each vowel, extension of [+spread glottis] or a related feature from word-initial heavy stops to following vowels is a plausible phonological process in the sound system of Javanese.

ACKNOWLEDGMENTS

We would like to thank two anonymous ICPhS reviewers, Jane Li, Paul Smolensky, Kwan Srijomkwan, and Alan Zhou for helpful comments. This work was partially supported by NSF grant BCS-1941593 to CW.

5. REFERENCES

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 25th ed. Dallas, Texas: SIL International, 2022, online version: http://www.ethnologue.com.

[2] K. M. Dudas, "The phonology and morphology of Javanese," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1976.

[3] E. C. Horne, *Beginning Javanese*. New Haven: Yale University Press, 1961.

[4] J. L. Fagan, "Javanese intervocalic stop phonemes: The light/heavy distinction," in *Studies in Austronesian Linguistics*, R. McGinn, Ed. Athens, Ohio: Ohio University Press, 1988, pp. 173–202.

[5] K. M. Hayward and K. M. Muljono, "The dental/alveolar contrast in Javanese," *Bulletin of the School of Oriental and African Studies*, vol. 54, pp. 126–144, 1991.

[6] E. M. Uhlenbeck, "The structure of the Javanese morpheme," *Lingua*, vol. 2, pp. 239–270, 1949.

[7] J. C. Catford, *Fundamental Problems in Phonetics*. Bloomington: Indiana University Press, 1977.

[8] M. Brunelle, "The role of larynx height in the Javanese tense∼lax stop contrast," in *Austronesian Contributions to Linguistic Theory: Selected Proceedings of AFLA*, R. Mercado, E. Potsdam, and L. Travis, Eds. John Benjamins, 2010, pp. 7–24.

[9] R. H. Sumukti, "Javanese morphology and morphophonemics," Ph.D. dissertation, Cornell University, Ithaca, New York, 1971.

[10] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Oxford: Blackwell, 1996.

[11] E. Thurgood, "Phonation types in Javanese," *Oceanic Linguistics*, vol. 43, pp. 277–295, 2004.

[12] M. Matthews, "An acoustic investigation of Javanese stop consonant clusters," in *The Proceedings of the 21st Meeting of the Austronesian Formal Linguistics Association*, C. S. Amber Camp, Yuko Otsuka and N. Tanaka, Eds., 2015, pp. 201–217.

[13] M. J. Kenstowicz, "Phonetic correlates of the Javanese voicing contrast in stop consonants," *NUSA: Linguistic studies of languages in and around Indonesia*, pp. 1–37, 2021.

[14] P. Ladefoged, *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago, 1971.

[15] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, pp. 383–406, 2001.

[16] M. Garellek, Y. Chai, Y. Huang, and M. Van Doren, "Voicing of glottal consonants and non-modal vowels," *Journal of the International Phonetic Association*, vol. 94, pp. 1–28, 2021.

[17] P. Keating, J. Kuang, M. Garellek, C. M. Esposito, and S. Khan, "A cross-language acoustic space for vocalic phonation distinctions," *Language*, To appear.

[18] K. M. Hayward, "/p/ vs. /b/ in Javanese: Some preliminary data," in *SOAS Working Papers in Linguistics and Phonetics 3*, pp. 1–99.

[19] ——, "/p/ vs. /b/ in Javanese: The role of the vocal folds," in *SOAS Working Papers in Linguistics and Phonetics 5*, pp. 1–11.

[20] K. M. Hayward, D. Grafield-Davies, B. J. Howard, J. Latif, and R. Allen, "Javanese stop consonants: the role of the vocal folds. School of Oriental and African Studies [video]." 1994.

[21] S. Seyfarth, J. Vander Klok, and M. Garellek, "Acoustics of the tense-lax stop contrast in Semarang Javanese," *The Journal of the Acoustical Society of America*, vol. 142, no. 4_Supplement, pp. 2581–2581, 10 2017.

[22] E. C. Horne, *Javanese-English Dictionary*. Yale University Press, 1974.

[23] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, Gurugram, India, Aug. 2018, pp. 52–55.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[25] J. Keshet, M. Sonderegger, and T. Knowles, "AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction," 2014, [Computer program].

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2022, [Computer program].

[27] S. Barreda, "Fast Track: Fast (nearly) automatic formant-tracking using Praat," *Linguistics Vanguard*, vol. 7, 2021.

[28] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011, pp. 1846–1849.

WHAT FRENCH SPEAKERS' NASAL VOWELS TELL US ABOUT ANTICIPATORY NASAL COARTICULATION

Francesco Rodriquez[1], Marianne Pouplier[1], Roy Alderton[1], Justin J.H. Lo[2], Bronwen G. Evans[2], Eva Reinisch[3], Christopher Carignan[2]

[1]Iinstitute for Phonetics and Speech Processing (IPS), LMU Munich; [2]Dept. of Speech, Hearing and Phonetic Sciences, University College London; [3]Acoustic Research Institute Austrian Academy of Sciences, Vienna
{f.rodriquez|pouplier}@phonetik.uni-muenchen.de, roy.alderton@city.ac.uk,
{justin.lo|bronwen.evans|c.carignan}@ucl.ac.uk, Eva.Reinisch@oeaw.ac.at

ABSTRACT

This paper examines the timing of anticipatory nasalization in French and relates it to individual speakers' realizations of phonologically nasal vowels. We explore the possibility that coarticulation in VN sequences is more extensive for speakers who only differentiate nasal and oral vowels to a limited extent based on nasality. Nasal intensity was measured in a time window preceding a nasal consonant or oral control segment (VN/VC) as well as on speakers' nasal and oral vowels (Ṽ/V). The results suggest that speakers for whom Ṽ and V differ less in nasality display earlier and more variable coarticulatory timing in VN sequences. Possibly, in speakers for whom nasality is less informative as a cue for the V/Ṽ contrast, anticipatory nasalization is less likely to mask said contrast and thus coarticulation can be more extensive and variable. The results contribute to our understanding of how speaker-specific manifestations of phonological contrast shape coarticulatory behavior.

Keywords: coarticulation, nasalization, speaker variability, contrast, French

1. INTRODUCTION

It is often argued that phonological contrast acts as an important constraining force on coarticulation [1], ensuring that underlyingly distinctive features are not conflated with each other. For instance, in French, where there is a V/Ṽ contrast, rates of anticipatory nasalization are relatively low when compared to English, where there is no V/Ṽ contrast [2, 3]. The rationale is that extensively contextually nasalized oral vowels in French might lead to listeners confusing them with a phonologically nasal vowel.

This view receives support from the fact that anticipatory nasal coarticulation in French is especially constrained in the (mid)-low vowel space, where oral and nasal phonemes are crowded, as opposed to the high part of the vowel system, where there are no nasal vowel phonemes (e.g. [4, 5, 6, 7]). That is, anticipatory nasal coarticulation is most constrained in contexts where contrast is most at stake. While this contrast hypothesis captures these particular facts about French nasal coarticulation, it cannot account for the considerable inter-speaker variability in coarticulatory timing reported in studies like [5, 7].

While examining nasal coarticulation typically involves nasality as a single parameter, nasal vowels and speech sounds in general differ along multiple phonetic dimensions. Also, the importance that a phonetic parameter has in contrasting two speech sounds can vary between speakers (so-called 'cue-weighting': see [8] for a review). Contrary to the traditional conception of nasal vowels as oral vowels with a [+nasal] feature [9], it was found that French nasal vowels' oral articulations are shifted with respect to their supposed oral vowel counterparts (e.g., [10, 11, 12]). The French V/Ṽ contrast is thus not maintained by nasality alone but by vowel quality as well (among other dimensions). Furthermore, the use of nasality in signalling the V/Ṽ contrast is not uniform across speakers of French as shown in [13]: in some speakers the difference between V and Ṽ was not detectable based on velum height alone, but only by also considering oral articulations. Beddor [14] suggests that inter-speaker variability in coarticulation is part of a complex production-perception dynamic, stemming from speaker-specific configurations of contrast signalling. If this is the case, it stands to reason that the way a speaker signals the V/Ṽ contrast may be in a systematic relationship with that speaker's coarticulatory behavior in VN sequences. This possibility presupposes that there is a principled relationship between phonological contrast and coarticulatory behavior. While such a relationship has been elusive at a language level [e.g., 15, 16], the current paper asks whether such a relationship may be evident at the individual level.

When investigating the role of contrast in coarticulatory behavior, one must thus take into account the degree to which the examined phonetic parameter (nasality) actually participates in maintaining the contrast (V/Ṽ) in an individual speaker's contrast system. This in turn might affect how much contrast can be expected to constrain coarticulation among individual speakers. For

instance, contextually nasalized oral vowels might not put the V/Ṽ contrast at stake in a speaker who does not express the contrast primarily by means of nasality but rather by other phonetic cues (e.g., vowel quality). Consequently, anticipatory nasal coarticulation could be more variable and extensive in such a speaker.

The goal of this paper is therefore to quantify the timing of French anticipatory nasal coarticulation and its variability as well as to relate them to the weight nasality possibly has in individual French speakers' V/Ṽ contrast. In an extension of Beddor's hypothesis for speaker-specific cue-trading for English [14], one can make the prediction that speakers time nasal coarticulation earlier and more variably if their V/Ṽ contrast is marked less by differences in nasality. Instead, speakers whose V/Ṽ contrast is heavily marked by differences in nasality, are predicted to time coarticulation later and more consistently.

2. METHOD

2.1. Speakers

Native speakers of Northern Metropolitan French (French hereafter) were recorded in Germany or the UK while completing a reading task. We aim to collect data from 30 speakers: here, we present data from 24 speakers for whom analysis is complete.

2.2. Material

Two sets of stimuli were created: 15 real-word minimal pairs that are contrasted by a nasal/oral consonant, which appears in word-initial, word-medial or word-final position ('consonant corpus' hereafter; see Table 1), and 12 minimal pairs that are contrasted by a nasal/oral vowel ('vowel corpus' hereafter; see Table 2). These were embedded in the carrier phrase '*Je dis à Cléo* **X** *samedi*.' (I told Cleo X on Saturday), where X is the target word.

initial	medial	final
mère, père [mɛʁ] [pɛʁ] mother, father	*ligneur, liqueur* [liɲœʁ] [likœʁ] eyeliner, liquor	*l'homme, lotte* [lɔm] [lɔt] man, burbot

Table 1: Consonant corpus: example minimal pairs for each word position.

The defined region of interest (ROI) spans from the left edge of [e] in *Cléo* to the right edge of the nasal target/oral control segment. Each token was repeated three times. The two stimulus sets were joined, randomized for each speaker, and presented in three blocks.

nasal	oral
lin, long, lent [lɛ̃] [lɔ̃] [lɑ̃] flax, long, slow	*lait, lotte, là* [lɛ] [lɔt] [la] milk, burbot, there

Table 2: Vowel corpus: example stimuli.

2.3. Data recording and pre-processing

We collected acoustic data with a nasalance device (*Glottal Enterprise*), which records nasal and oral signals by means of two microphones separated by an acoustic baffle. The audio files were segmented automatically [17]. Using *Praat* [18], segment boundaries within the ROI were manually corrected and a band-pass filter (80-10,000 Hz) was applied to the recordings. Intensity values within the ROI were extracted from the nasal channel and mean-corrected for total intensity of both channels on a token-by-token basis. In the consonant corpus, nasality was measured from the start of the ROI to the onset of the nasal target/oral control consonant. In the vowel corpus, nasality was measured on the nasal target/oral control vowel. Tokens with a prosodic break before the target word were excluded. One speaker was excluded from analysis due to an overabundance of prosodic breaks. The final consonant corpus comprises 1042 tokens, the vowel corpus 763 tokens.

2.4. Quantifying coarticulatory timing in VN

In the consonant corpus, where we contrast VN and VC sequences, a mean nasal intensity curve was computed across the three VC repetitions of a pair. The resulting signal was used as a pair-specific 'oral' baseline and subtracted from the corresponding VN signal, yielding a 'difference signal' per token. Within a pair, signals were trimmed to the length of the shortest signal. The difference signal computation is exemplified in Fig. 1: the averaged 'oral' baseline signal for *latte* [lat] (top: solid) is subtracted from the signal of a repetition of *l'âme* [lam] (top: dashed), resulting in a difference signal (bottom: dotted).

Using the *sicegar* package [19] in *R* [20], sigmoid functions were fitted to the time- and magnitude-normalized difference signals of the consonant corpus. Fig. 1 (bottom: solid) shows an example of a fitted sigmoid of a difference token of *l'âme* [lam]. The x intercept of the inflection point's tangent is then used as a proxy for coarticulatory timing, which we call the 'divergence point'. Each difference signal was fitted twice: once using a single sigmoid fit and once using a double sigmoid fit. Whichever fit resulted in a better accuracy score (AIC score) was chosen to calculate the token's divergence point in normalized time. Normalized time values

were then transformed to timing values in seconds. A 5% RMS threshold was applied to exclude tokens for which there was no good fit, leaving 938 tokens for analysis.

The mean and standard deviation (SD) of the divergence points of every speaker were then computed to obtain speaker-specific values representing timing of coarticulation (mean) as well as variability of timing (SD). While this means that a speaker's entire data is collapsed into two values, this enables a direct comparison with values obtained from their nasal vowel productions.

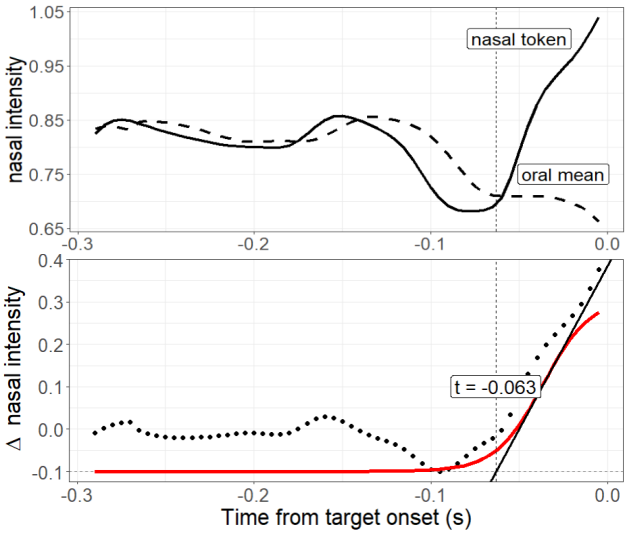


Figure 1: Top: Nasal amplitude signal of a single nasal token of *l'âme* (solid) compared to the mean oral baseline signal of three repetitions of *latte* (dashed). Bottom: Nasal difference signal of a token of *l'âme* (dotted) and its fitted sigmoid (solid). The black solid line marks the inflection point of the sigmoid's slope. The drawn slope's x intercept marks the onset of coarticulatory divergence (*t*).

2.5. Quantifying nasality in the V/Ṽ contrast

To obtain a Ṽ difference measure for nasal tokens in the vowel corpus, an 'oral' baseline – computed by averaging the mean-corrected nasal amplitude across all oral control vowels of a speaker – was subtracted from the mean nasality of a nasal Ṽ token. Using a single speaker-specific oral baseline value avoids the issue of having to pair nasal vowels with supposed oral 'counterparts' (see [20] for a critique). To obtain a single representative value of Ṽ nasality of a speaker, all the pair difference values for a speaker were collapsed into a single average value of V/Ṽ difference. A speaker's V/Ṽ difference is interpreted here as the importance nasality has in expressing the V/Ṽ contrast in that speaker relative to other speakers.

Potential relationships between the speaker-specific values for the extent of anticipatory coarticulation (consonant corpus: mean and SD of divergence point) and V/Ṽ difference (vowel corpus) were analyzed by computing a Pearson correlation.

3. RESULTS

Figure 2 shows the distribution of the onset of coarticulation (divergence point) in seconds for all speakers for the consonant corpus. Speakers are ordered from most (bottom) to least (top) extensive coarticulation.

There is overall a strong negative correlation between the means and SDs of the divergence values (Pearson's R = -0.84, $p < 0.001$). This means that speakers who time the onset of anticipatory coarticulation more consistently (e.g., speaker 25 and 11) tend to have a smaller extent of anticipatory coarticulation compared with more variable speakers (e.g., speakers 1 and 4).

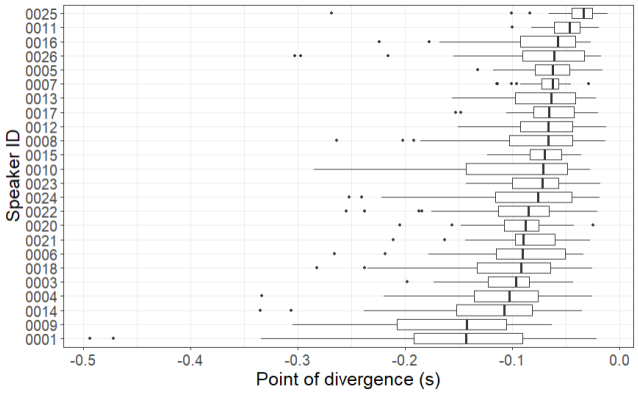


Figure 2: Speakers (y axis = speaker IDs) ranked from 'earliest' (bottom) to 'latest' onset of coarticulation (top).

Looking at speakers' vowel productions suggests that those speakers who time coarticulation early and variably in VN contexts (e.g., speaker 1 and 4) tend to display smaller V/Ṽ differences than more consistent speakers (e.g., speaker 11). In Fig. 3 all speakers are ranked from smallest (bottom) to biggest (top) nasal V/Ṽ difference.

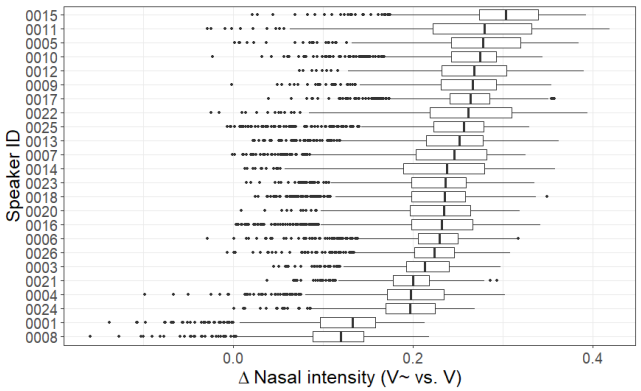


Figure 3: Speakers (y axis = speaker IDs) ranked from smallest (bottom) to biggest (top) nasal V/Ṽ difference.

When mapping the SD of speakers' divergence points onto their average V/Ṽ difference (Fig. 4), a strong negative correlation is apparent (Pearson's R = -0.53, $p \sim 0.008$). This means that in general, speakers who time nasal coarticulation more consistently tend to have a greater average V/Ṽ difference.

A moderate nonsignificant positive correlation was found when mapping speakers' mean divergence points onto their mean nasal V/Ṽ differences (Fig. 5) (Pearson's R = 0.33, p ~ 0.121). This means that speakers with greater mean nasal V/Ṽ differences tend to limit the extent of coarticulation.

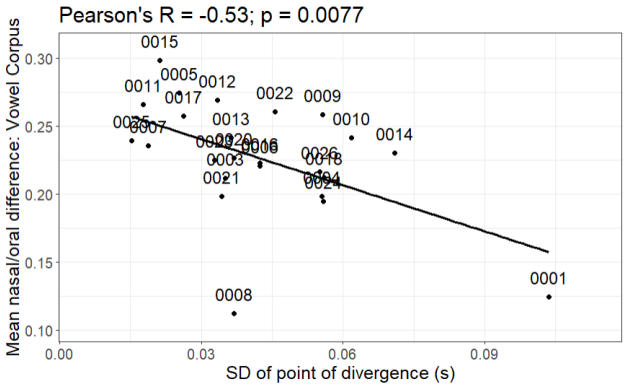


Figure 4: SD of divergence points (x axis) mapped onto mean nasal V/Ṽ difference (y axis). Each data point represents a speaker.

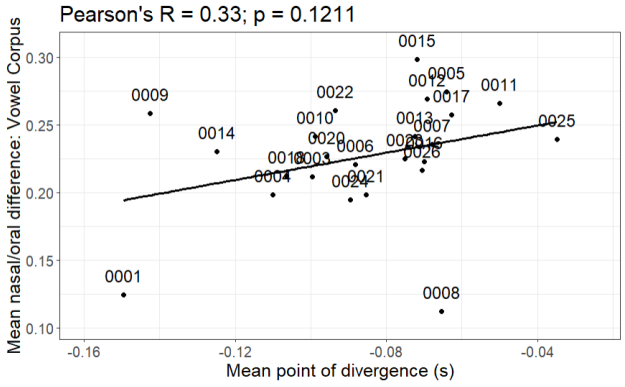


Figure 5: Mean of divergence points (x axis) mapped onto mean nasal V/Ṽ difference (y axis). Each data point represents a speaker.

5. DISCUSSION

The goal of this paper was to quantify the timing of French anticipatory nasal coarticulation in VN sequences and its variability. Furthermore, we aimed to relate coarticulatory timing to the degree to which individual speakers' V/Ṽ contrasts are maintained by means of nasality. The results partially support the hypothesis we initially formulated based on the contrast hypothesis: Less involvement of nasality in the expression of the V/Ṽ contrast does indeed correlate with more variable nasal coarticulatory timing and generally an earlier onset of nasal coarticulation, although the latter is not significant. Here it is argued that the reason for these correlations is that a contextually nasalized V in an VN context cannot put a speaker's V/Ṽ contrast at stake if said contrast only weakly depends on nasality in the first place.

This line of argumentation, however, implies that these speakers might rely more extensively on other informative phonetic cues (such as vowel quality) to maintain the V/Ṽ distinction. In future studies we will test this assumption by measuring spectral differences between V and Ṽ (e.g., formant measures) in the same speakers that participated in this study. We would then expect to find a trade-off between the difference in V/Ṽ nasality and the difference in vowel quality between V and Ṽ.

In summary, the results presented here challenge a language-level account of the role of contrast in shaping coarticulatory behavior, according to which the mere presence of a phonological contrast constrains coarticulation across the board. Our results rather support a speaker-level account: The more a phonetic parameter (nasality) participates in maintaining a phonological contrast (Ṽ/V) in an individual speaker's contrast system, the more coarticulation is constrained in order to preserve that contrast. Nevertheless, this does not exclude other known sources of variability in French nasal coarticulation such as phonotactic restrictions [22] or vowel height and duration [6], the effects of which should be re-examined in tandem with individual speakers' contrast systems.

6. ACKNOWLEDGEMENTS

Work supported by DFG grant PO 1269/5-1 to M. Pouplier and C. Carignan. Katharina Neubert was of invaluable help during data collection and segmentation.

7. REFERENCES

[1] Manuel, S. 1990. The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *J. Acoust. Soc. Am*. 88, 1286–1298.

[2] Tranel, B. 1978. *The sounds of French: an introduction*. Cmabridge University Press.

[3] Valdman, A. 1993. *Bien entendu!: Introduction à la pronunciation française.* Pearson College Division.

[4] Delvaux, V., Demolin, D., Harmegnies, B., Soquet, A. 2008. The aerodynamics of nasalization in French. *J. Phon.* 36(4), 578–606.

[5] Basset, P., Amelot, A., Vaissière, J., Roubeau, B. 2001. Nasal airflow in French spontaneous speech. *J. Int. Phon. Ass.* 31(1), 87–99.

[6] Dow, M. 2020. A phonetic-phonological study of vowel height and nasal coarticulation in French. *J. French Lang. Stud.* 30(3), 239–274.
[7] Rodriquez, F., Pouplier, M., Alderton, R., Carignan, C., Lo, J.J.H, Evans, B.G., Reinisch, E. 2022. Individual variability and contrast in French anticipatory nasal coarticulation [poster]. *18. Phonetik & Phonologie im deutschsprachigen Raum*, October 6th-7th, Bielefeld, Germany.
[8] Schertz, J., Claire, E.J. 2020. Phonetic cue-weighting in perception and production. *WIREs Cog. Sci.* 11(2).
[9] Chomsky, N., Halle, M. 1968. *The Sound Pattern of English*. Harper & Row.
[10] Zerling, J.-P. 1984. Phénomènes de nasalité et de nasalisation vocaliques: Etude cinéradiographique pour deux locuteurs. *Travaux de L'Institut de Phonétique de Strasbourg* 16, 241–266.
[11] Bothorel, A., Simon, P., Wioland, F., Zerling, J.-P. 1986. Cinéradiographie des voyelles et des consonnes du français. *Travaux de L'Institut de Phonétique de Strasbourg* 18, 1296.
[12] Carignan, C., Shosted, R., Fu, M., Liang, Z.-P., Sutton, B. 2015. A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *J. Phon.* 50, 34-51.
[13] Engwall, O., Delvaux, V., Metens, T. 2006. Interspeaker viariation in the articulation of nasal vowels. In: Yehia, H., Demolin, D., Laboissière, R. *Proceedings of the 7th ISSP*, 3-10.
[14] Beddor, P.S., McGowan, K.B., Boland, J.E., Coetzee, A.W., Brasher, A. 2013. The time course of perception of coarticulation. *J. Acoust. Soc. Am.* 133, 2350–2366.
[15] Mok, P. 2012. Does vowel inventory density affect vowel-to-vowel coarticulation? *Language and Speech* 56(2), 191–209.
[16] Brkan, A. 2018. *Étude comparative des phénomènes de coarticulation nasale en anglais américain, bosnien, français, norvégien et ourdou* [PhD thesis]. Université Sorbonne Paris Cité.
[17] Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., Pörner, N. 2016. BAS speech science web services. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, paper id 668.
[18] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer* [computer program]. http://www.praat.org
[19] Caglar, M. U., Teufel, A.I., Wilke, C.O. 2018. Sicegar: R Package for Sigmoidal and Double-Sigmoidal Curve Fitting. *PeerJ 6*, p. e4251, 2018, doi: 10.7717/peerj.4251.
[20] *R: A Language and Environment for Statistical Computing*. 2021. R Foundation for Statistical Computing, Vienna, Austria. [Online]. https://www.R-project.org
[21] Shosted, R. 2015. Nasal vowels are not [+nasal] oral vowels. In: Smith, J., Ihsane, T. *Proceedings of the 42nd Linguistic Symposium on Romance Languages.*
[22] Desmeules-Trudel, F., Brunelle, M. 2018. Phonotactic restrictions condition the realization of vowel nasality and nasal coarticulation: Duration and airflow measurements in Québécois French and Brazilian Portuguese. *J. Phon.* 69, 43–61.

SPIRANTIZATION OF WORD-FINAL PLOSIVES IN STANDARD DUTCH

Alexandra Pfiffner & Julian Martinez-Garcia

University of California, Berkeley
apfiffner@berkeley.edu, julianm@berkeley.edu

ABSTRACT

Lenition is a cover term for a set of processes involving reduction or weakening of segments in similar phonological positions [1]. Many studies of plosive lenition have quantified the reduction in terms of consonant duration and intensity. However, this may not give a full picture in cases of "loss lenition," which generally occur in perceptually weak positions [2], such as word-finally. This study reports patterns of word-final spirantization in Standard Dutch from a corpus of word lists. We quantify lenition by duration and the presence/absence and timing relations of acoustic cues, including a complete closure, a transient burst, and a period of frication. Additionally, we investigate sociolinguistic factors that condition this variable lenition and find spirantization significantly more often in male speakers versus female speakers, and in younger versus older speakers.

Keywords: spirantization, Dutch, obstruents, sociophonetics

1. INTRODUCTION

Plosive lenition is realized in many forms, including (but not limited to) weakening to a fricative or approximant ("spirantization"), affrication, flapping, debuccalization, and deletion [3, 4]. Katz [2] describes two subsets that are phonetically and phonologically different: CONTINUITY LENITION and LOSS LENITION. CONTINUITY LENITION refers to intervocalic consonant lenition that results in shorter durations and/or higher intensities, which has been frequently reported in the literature [1, 5, 6, 7]. On the other hand, LOSS LENITION refers to perceptually weak positions undergoing a loss of features, gestures, and/or a contrast, such as coda consonant debuccalization in Andalusian Spanish [8, 9]. Phonological position is frequently described as a predicting factor of lenition, as well as stress [7], the openness of the preceding vowel [10], and speech rate [1]. Lenition can also be socially conditioned; for example, a sociophonetic study of stop lenition in Concepción, Chile, revealed that variable lenition is a social phenomenon, primarily driven by younger male speakers [5]. In the present study, we examine a case of variable (loss) lenition of word-final plosives in Standard Dutch.

1.1. Dutch word-final plosives

Dutch contrasts five plosives: /p, b, t, d, k/, with /g/ as a marginal phoneme seen in loanwords [11]. Historically, Dutch contrasted voicing in all phonological positions, but eventually underwent word-final neutralization. However, this neutralization has been found to be "incomplete," as small but statistically significant differences remain in cues to underlying voicing, such as closure and burst durations, and the preceding vowel duration [12, 13, 14, 15]. Furthermore, native speakers can use these cues in word recognition to make lexical distinctions at a rate better than chance [12, 13, 15], such as with underlying minimal pairs /ʊɛt/ 'law' and /ʊɛd/ 'bet.' However, a recent study of word-final voicing cues suggests that the incomplete neutralization is heading towards completion, as the burst duration cue has completely neutralized, and the closure duration cue was only significant in perception, but not production [15].

Mitterer and Ernestus [16] have reported that word-final /t/ can be variably lenited, especially in consonant clusters. In a corpus study of casual speech utterances, they identified four realization patterns of word-final /t/. The first type was a canonical plosive, with a closure, transient, and frication noise. The second type lacked a closure and transient, but had frication. Similarly, the third type had only frication, though at a lower amplitude and with a lower spectral center of gravity. The final realization was a complete deletion. Their analysis then focused on fully deleted tokens, leaving the other realizations of lenition unstudied.

Word-final /t/ is the only segment previously discussed in the literature, but it is not the only plosive to lenite in Standard Dutch. The present study adds to previous work by analyzing word-final lenition in /t, d, k/, giving a more robust view of this phenomenon. (Word-final /p, b/ is extremely rare in Dutch, and thus often left out of word-final stop analyses [15]. For this same reason, we do not

analyze word-final /p, b/.) If lenition is found to occur in /t/ but not /d/, then perhaps the incomplete neutralization described above is changing into a new contrast, where /t/ surfaces as [t̬], and /d/ surfaces as [t]. Furthermore, we examine lenition in careful speech, where it is less likely to occur [10]. We give a qualitative and quantitative description of the types of plosive realizations, and then examine possible sociophonetic conditioning factors of this lenition process.

2. METHODS

2.1. Data

Our data come from a corpus of word lists and read speech from 25 native speakers of Standard Dutch, all living in Amsterdam at the time of recording. The speakers formed two age groups: a younger group (n=14, ages 22-29) and an older group (n=11, ages 61-71). Both age groups were balanced for gender. The corpus was originally compiled to study cues to plosive voicing neutralization in word-initial and word-final position, so it includes many tokens of (underlying) minimal pairs. See [15] for more details on the speakers and corpus.

2.2. Analysis

Only word list productions are considered in this study. Audio files were force-aligned using the Montreal Forced Aligner [17] and segment boundaries were hand-corrected in Praat [18]. A total of 1,287 word-final tokens were further assessed for the presence or absence of a closure, transient burst, and frication period. We qualitatively assigned tokens to categories based on these cues and their temporal relations to each other. TYPE 1 was a canonical plosive with a complete closure, transient burst, and frication period. TYPE 2 had a complete closure and a frication period, but no visible or audible burst. TYPE 3 had unclear divisions between subsegments, exhibiting some type of (partial) closure but overlapping frication and transient burst. Finally, TYPE 4 was completely fricated.

Possible social and linguistic conditioning factors of lenition were then analyzed with an ordinal logistic regression model using the polr function of the MASS package [19] in R [20] to predict lenition type. The social factors that were included were gender (levels: female, male) and age group (levels: younger, older). The linguistic variables included plosive (levels = t, d, k), plosive duration, preceding vowel frontness (levels = front, back), and preceding vowel openness (levels = open, mid, near-close, close).

3. RESULTS

3.1. Realizations

A majority of plosives (80%) had canonical realizations, which we will refer to as TYPE 1. An example is shown in Figure 1. Their closures averaged 94 ms, and the post-burst frication periods averaged 178 ms. The average total plosive duration was 272 ms.

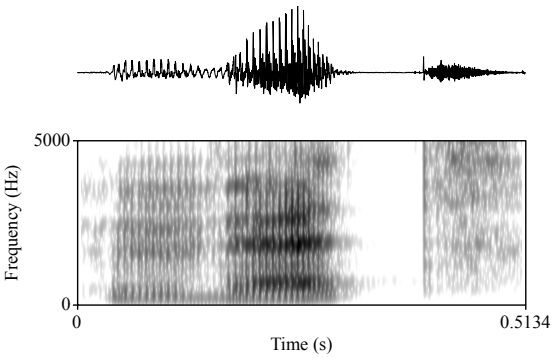


Figure 1: TYPE 1: /vɛt/ 'law' produced with closure, burst, and frication.

On a lenition continuum, TYPE 2 was the least lenited, with closure and frication periods, but no visible or audible burst (Figure 2). Approximately 14% of tokens were classified as this type. They had much shorter closures, averaging 71 ms, but longer frication periods averaging 183 ms. The plosives overall had an average duration of 254 ms.

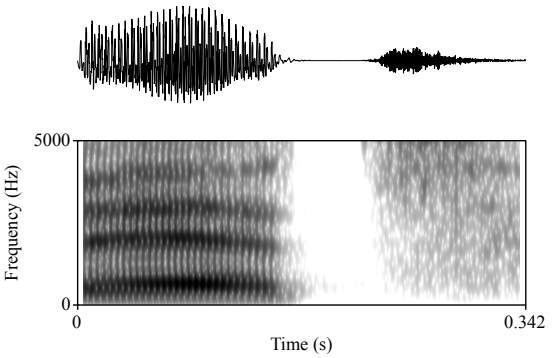


Figure 2: TYPE 2: /eet/ 'eat' produced without a visible or audible burst.

Only 4% of tokens were produced as a TYPE 3, with unclear boundaries between subsegments. Figure 3 shows an example of /boot/ 'boat' with a noisy closure and frication appearing prior to the transient burst spike. The average closure duration for this type was 59 ms, and the average post-burst duration was 195 ms. There were also occasional spectral differences within the frication period, similar to that observed by Mitterer and Ernestus [16]. For example, in Figure 3, there is a difference between the first part of the frication period, with a high amplitude at higher frequencies, and shortly after the burst, with an overall lower amplitude. TYPE 3 realizations had an average overall duration of 255 ms.

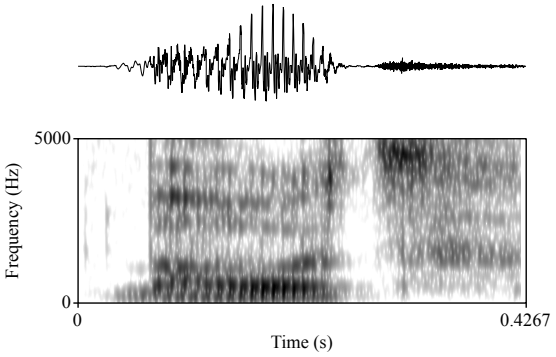


Figure 3: TYPE 3: /boot/ 'boat' produced with frication prior to the transient burst.

Finally, 3% of tokens were fully spirantized (TYPE 4), with no closure or burst, and only a frication period. The average duration of these spirantized variants was the shortest at 212 ms. An example of spirantized /d/ in /eed/ 'oath' is shown in Figure 4.

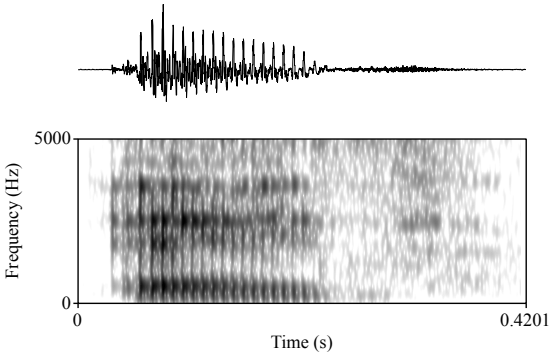


Figure 4: TYPE 4: /eed/ 'oath' completely spirantized.

3.2. Conditioning factors

The results of the ordinal logistic regression model are given in Table 1. The t-value is derived by dividing the regression coefficient by its standard error, and conventionally, if the t-value is greater than +2 or less than -2, it is considered significant. Higher t-values indicate higher levels of confidence in the coefficient as a predictor.

First, age group was significant, with younger speakers more likely to produce lenited productions than older speakers (Figure 5). Furthermore, male speakers, both young and old, were significantly more likely to produce lenited tokens than female speakers. Younger male speakers produced the highest amount of lenited plosives overall. Plosive duration was also significant, decreasing as tokens became more lenited, as outlined in the four types in Section 2.2. Neither openness nor frontness of the preceding vowel was significant.

Factor	Log odds	Std E	t-value
/k/	-14.83	1.338e-07	**-1.108e+08***
/t/	-0.25	0.16	-1.53
duration	-0.01	0.001	**-4.38***
younger	1.54	0.19	**8.22***
m gender	0.52	0.17	**3.13***
front V	0.20	0.17	1.16
mid	0.38	0.21	1.80
near-close	0.34	0.32	1.06
open	0.21	0.22	0.95

Table 1: Output of the ordinal logistic regression model.

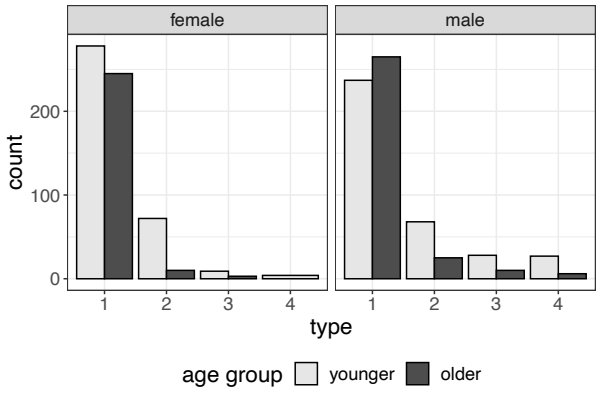


Figure 5: Counts of each lenition type, separated by age group and gender.

There was a significant difference between /d/ and /k/, but not between /d/ and /t/. Overall, /k/ had the fewest lenited realizations, and it had no

instances of complete lenition (TYPE 4). It also had the fewest total tokens due to the structure of the corpus. Notably, both /t/ and /d/ were equally represented and had similar counts of each lenition type. This suggests that the incomplete neutralization in underlying voicing is not becoming a new surface contrast, e.g. [t] versus [t̬].

4. DISCUSSION

All word-final plosives in Standard Dutch can variably lenite in the same general patterns previously reported for /t/ [16]. This lenition also follows the pattern seen in other languages where lenited tokens have shorter durations than non-lenited tokens [1, 5, 6, 7]. Interestingly, the four types of lenition in this study, ordered from a canonical stop to a fricative, had shorter durations at each consecutive step (with one exception of step 2 to step 3, which also differed only by 1 ms). Duration was additionally a significant factor in predicting lenition.

The identity of the plosive was significant, as /k/ lenited less often as its alveolar counterparts. Equal lenition of /t/ and /d/ suggests that the incomplete neutralization is not developing into a new contrast. Instead, all of the word-final plosives are subject to variable lenition.

In terms of social factors, younger speakers and male speakers were significantly more likely to produce lenited tokens. Previous sociophonetic research has shown gender differences in terms of which variants of a sound are used; female speakers tend to use more innovative variants [21], while male speakers are more likely to use non-standard variants [22]. This further provides evidence against the hypothesis that this lenition is a change in progress from a weak incomplete neutralization to a new contrast. Instead, it is more likely that this lenition is simply a stable, non-standard variant in Dutch.

5. CONCLUSION

We present a qualitative and quantitative analysis of word-final plosive lenition in Standard Dutch careful speech. This adds to our knowledge of lenition from previous studies of word-final /t/ deletion in Dutch by quantifying how often each type of lenition occurs as well as the average segmental duration of non-lenited and lenited stops. Furthermore, we present the first sociophonetic analysis of word-final Dutch plosive spirantization, showing that this lenition is both linguistically and socially conditioned. This is a starting ground for future research on Dutch word-final plosive spirantization, which should include additional social and linguistic factors, especially in casual speech, to better understand what (if any) social meaning is attributed to this process.

6. REFERENCES

[1] U. C. Priva and E. Gleason, "The causal structure of lenition: A case for the causal precedence of durational shortening," *Language*, vol. 96, pp. 413–448, 2020.

[2] J. Katz, "Lenition, perception and neutralisation," *Phonology*, vol. 33, pp. 43–85, 2016.

[3] J. Kingston, "Lenition," in *Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology*, L. Colantoni and J. Steele, Eds., 2008, pp. 1–31.

[4] E. Buizza and L. Plug, "Lenition, fortition and the status of plosive affrication: the case of spontaneous RP English /t/," *Phonology*, vol. 29, pp. 1–38, 2012.

[5] B. M. A. Rogers and C. A. Mirisis, "Voiceless stop lenition and reduction as linguistic and social phenomena in Concepción, Chile," *Borealis: An International Journal of Hispanic Linguistics*, vol. 7, pp. 187–215, 2018.

[6] J. Katz and G. Pitzanti, "The phonetics and phonology of lenition: A Campidanese Sardinian case study," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 10, pp. 1–40, 2019.

[7] D. Bouavichith and L. Davidson, "Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech," *Phonetica*, vol. 70, pp. 182–206, 2013.

[8] C. Gerfen and K. C. Hall, "Coda aspiration and incomplete neutralization in Eastern Andalusian Spanish," 2001, unpublished manuscript.

[9] C. Gerfen, "Andalusian codas," *Probus*, vol. 14, no. 2, pp. 247–277, 2002.

[10] R. M. Kirchner, "An effort-based approach to consonant lenition," Ph.D. dissertation, University of California, Los Angeles, 1998.

[11] I. M. Mees and B. Collins, "A phonetic description of the consonant system of Standard Dutch (ABN), fifth revised edition," *Journal of the International Phonetic Association*, vol. 12, no. 1, pp. 2–12, 1982.

[12] N. Warner, A. Jongman, J. Sereno, and R. Kemps, "Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch," *Journal of Phonetics*, vol. 32, pp. 251–276, 2004.

[13] M. Ernestus and R. H. Baayen, "The functionality of incomplete neutralization in Dutch: The case of past-tense formation," in *Laboratory Phonology 8*, L. Goldstein, D. H. Whalen, and C. Best, Eds. Berlin, Germany: Mouton de Gruyter, 2006, pp. 27–49.

[14] ——, "Intraparadigmatic effects on the perception of voice," in *Voicing in Dutch: (De)voicing*

- *phonology, phonetics, and psycholinguistics*, J. van de Weijer and E. J. van der Torre, Eds. Amsterdam, the Netherlands: John Benjamins, 2007, pp. 153–173.
[15] A. M. Pfiffner, "Cue-based features: Modeling change and variation in the voicing contrasts of Minnesotan English, Afrikaans, and Dutch," Ph.D. dissertation, Georgetown University, 2021.
[16] H. Mitterer and M. Ernestus, "Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch," *Journal of Phonetics*, vol. 34, pp. 73–103, 2006.
[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner (Version 0.9.0) [Computer program]," 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/
[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 6.1). [Computer program]," 2019. [Online]. Available: http://www.praat.org
[19] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, and D. Firth, *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, 2023, R package version 7.3-59. [Online]. Available: https://cran.r-project.org/web/packages/MASS/MASS.pdf
[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/
[21] W. Labov, *Principle of linguistic change, Volume 2: Social factors*. Malden, MA: Blackwell Publishers, 2001.
[22] J. L. Fischer, "Social influences on the choice of a linguistic variant," *WORD*, vol. 14, no. 1, pp. 47–56, 1958.

SPONTANEOUS IMITATION OF ENGLISH SIBILANTS BY NATIVE AND NON-NATIVE SPEAKERS

Ivy Hauser, Emily Graham and Xinwen Zhang

University of Texas at Arlington
ivy.hauser@uta.edu

ABSTRACT

This study investigates spontaneous phonetic imitation of the English alveolar sibilant /s/ by monolingual native speakers, bilingual native speakers, and non-native speakers with different language backgrounds in a delayed shadowing task. Participants first produced /s/-initial words, then were exposed to model speech with enhanced spectral mean (SM) on /s/, and finally produced /s/-initial words again post-exposure. All participants increased SM after exposure, converging towards the model talker. There were no significant differences in degree of imitation across the different language backgrounds. These results also provide an example of the starting distance bias associated with the difference-in-distance measure of imitation, as apparent significant differences across language backgrounds are accounted for by differences in starting distance to the model.

Keywords: phonetic imitation, speech acoustics, accommodation, bilingualism, methodology

1. INTRODUCTION

There is a large body of work demonstrating that talkers unconsciously alter their production towards speech to which they are exposed, even in lab settings without explicit instruction to imitate [1, 2]. Spontaneous imitation in native speech is constrained by social attitudes [3] as well as linguistic factors like phonological category boundaries and lexical frequency [2]. These factors also affect imitation in non-native speech, where degree of imitation is affected by belief about the linguistic proficiency of the model talker [4] and phonological representations in the second language (L2) [5, 6].

Linguistic and social distance between interlocutor pairs has been shown to be a factor in degree of convergence, but the direction of the effect is unclear. Some studies found more convergence for same dialect pairs than different dialect/language pairs [7], while other studies found more convergence for different dialect pairs [8, 9]. In addition, degree of convergence for L2 speakers specifically may depend on whether the model talker is a native speaker of the target language. In [10], Polish learners of English were found to converge more towards English vowel duration when the model talker was a native speaker of English. These divergent findings could be due to differences in the type of task; [7] used a conversational task while [8, 9, 10] used shadowing. Ultimately, it is unclear whether imitation is facilitated or inhibited when the model talker / interlocutor is of a different language background than the participant.

The present study builds on this by examining how speakers of different language backgrounds imitate a native English speaker in laboratory task. The study is designed to examine convergence across different language backgrounds, not the influence of native phonology in non-native shadowing, so the stimuli use the alveolar sibilant /s/ which is contrastively present in all participant languages. Participants include native English monolingual speakers, English-dominant native English-Spanish bilingual speakers, and non-native speakers with L1 Spanish, Vietnamese, or Urdu backgrounds. All participants were exposed to enhanced spectral mean (SM) on /s/, and imitation was expected for all groups as none were speakers of languages where enhancing SM on /s/ would threaten phonological contrast (which has been shown to inhibit spontaneous imitation [2]).

If different language backgrounds inhibit phonetic imitation (as in [7]), the native English speakers are predicted to converge to the model talker more than the speakers of other L1 backgrounds. If different language backgrounds facilitate phonetic imitation (as in [8, 9]), the non-native speakers are predicted to converge to the model talker more than the native speakers. It is also possible that the model talker's status as a native speaker of English could facilitate imitation for the non-native speakers independently of the degree of linguistic distance between the model and participants (as in [10]).

This study also provides a test case of the

difference-in-distance (DID) measure often used to quantify phonetic imitation, which has been shown to exhibit a 'starting distance bias' [11, 12]). When using DID, speakers who have greater baseline distance from the model will be found to have converged more than those who have smaller baseline distance from the model. This is the case even in datasets that lack convergence [11]. This is a particular problem in datasets like the one presented here, where differences in baseline starting distance co-vary with other variables of interest.

An alternative to DID is linear combination, which quantifies imitation by modeling post-exposure values as the dependent variable in a linear regression with baseline values and model talker values as predictors [11, 12]. This is different from most DID approaches, which typically model DID as the dependent variable. In a linear combination model, significant effects of both the baseline and model values indicate convergence. The analysis here compares DID and linear combination. Apparent DID differences across language background are accounted for by differences in starting distance to the model and do not appear to reflect systematic differences in degree of convergence based on L1/L2 status.

2. METHODS

The procedure is delayed shadowing with a baseline word-naming task, an exposure (listening) block, and a post-exposure word-naming task with no explicit instruction to imitate. The stimuli were designed to follow [2], who examined spontaneous imitation of enhanced English stops.

2.1. Stimuli

The test words for the baseline word-naming task comprised 40 high frequency [s]-initial words, 40 low frequency [s]-initial words, and 30 sonorant-initial filler words. The listening block contained the 80 [s]-initial words from the baseline production block and 40 sonorant-initial fillers that were different from those in the baseline production. The test words for post-exposure production comprised the baseline list plus 20 novel low frequency [s]-initial words and 20 novel low frequency [z]-initial words (the [z] words are not analyzed here).

All test words had initial stress and no onset clusters. Word frequency data was obtained from CELEX [13]. Thresholds for low and high frequency were below 300 and above 1000 per 17.9 million respectively. Test words were additionally counterbalanced for phonological neighborhood density (data obtained from the English Lexicon Project [14], number of syllables (1-3), and rounding of the following vowel.

The 80 target words and 40 filler words for the listening block were recorded by a phonetically trained female native American English speaker. Recording took place in a sound-attenuated booth at a university lab with a Shure SM35 headworn microphone and Audacity software [15]. Recordings were sampled at a rate of 44.1 kHz with a bit depth of 16.

The model speech with enhanced SM was created by shifting the spectrum of the sibilant noise up using the "shift frequencies" function in Praat [16]. The initial sibilants were segmented using TextGrids, so only the sibilant noise was shifted while the rest of the word was not. The spectra were shifted up 15% of the raw SM value. The manipulated stimuli were evaluated by two native American English speakers and one L2 English speaker for naturalness (to ensure the stimuli did not sound so unnatural that it interfered with participants' perception).

2.2. Participants

Twelve female English speakers between ages of 18-25 were recruited at a large American university. 4 were monolingual native speakers with no early exposure to other languages and no reported proficiency in any other languages, 2 were native bilingual Spanish-English speakers, 2 were non-native L1 Urdu speakers, 2 were non-native L1 Spanish speakers, and 2 were non-native L1 Vietnamese speakers. All non-native speakers reported high English proficiency and high L1 proficiency. Proficiency was determined with self-report on a questionnaire. High English proficiency is additionally verified by the fact that all participants are students at a US university where the language of instruction is English.

2.3. Procedure

The procedure had four blocks: warm-up reading, baseline word-naming, listening (exposure), and post-exposure word-naming. Most participants took 30-40 minutes to complete the study. The recording was done using the same parameters as noted in the above section for the stimulus recording. The stimuli were presented in a random order inside the sound-attenuated booth on an external monitor. Before beginning, participants were told the researchers were interested in how they speak naturally, as if talking to a friend.

The warm-up block for silent reading follows previous work [2, 1] and familiarizes participants with the words used in the baseline word-naming task to reduce potential hyperarticulation of low-frequency words. In the baseline block, the participants were presented with the same words as in the warm-up but they were instructed "to name the word they see by saying it aloud." In the listening exposure block, the participants were asked to listen to the manipulated words in the listening list using headphones. The instructions read, "In this section you will hear words but won't see any on the screen. You do not need to speak, just listen to what you hear." The last block, the post-exposure production task, used the same instructions as the baseline word-naming task, but included the additional novel words.

The recordings were force aligned [17] and all sibilant boundaries were hand edited by the researchers. Time-averaged SM values for each sibilant were extracted using a Praat script [18].

3. RESULTS

This section presents the imitation results by comparing two methods of quantifying imitation: difference-in-distance (DID; e.g. [3]) and linear combination [12, 11]. Figure 1 displays SM across the baseline and post-exposure blocks. On average, participants in each language group increased SM after exposure to the stimuli with enhanced SM. DID was calculated for each token as $|(SM_{baseline} - SM_{model})| - |(SM_{post} - SM_{model})|$ such that positive values indicate convergence and negative values indicate divergence. DID values across each language group are graphed in Figure 2. While all groups show an average positive DID indicating convergence, there is a visual difference in apparent degree of convergence. In particular, the L1 Vietnamese speakers' DID values are higher relative to the other groups.

To estimate the strength of this effect, DID was analyzed as the dependent variable using a basic DID model as in [11]: DID $\sim$ Language Group + Lexical Frequency + (1|speaker) + (1|word). The fixed effect of frequency was included because infrequent words have been shown to exhibit more convergence [1, 2], but there was no significant effect of frequency on DID in the model. There was a significant effect for the L1 Vietnamese group ($\beta = 674.26, p < 0.001^{***}$) indicating significantly higher DID and more convergence relative to the intercept L1 monolingual English group. However, DID is known to exhibit bias according to starting

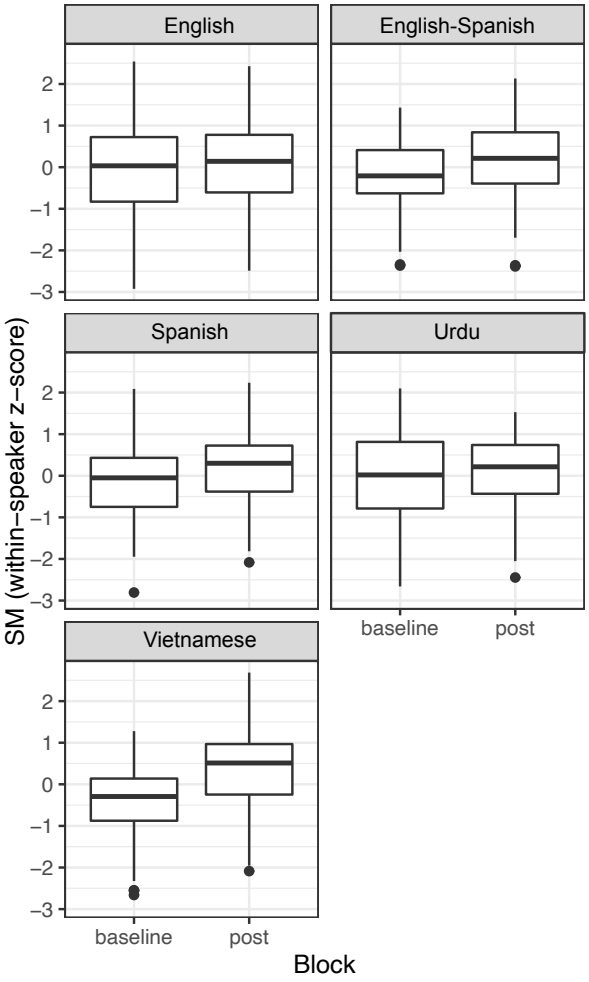


Figure 1: SM across blocks for each language group. Panels are labeled with L1's of participants.

distance and this result does not obtain in a linear combination model which takes baseline SM into account.

Differences in average baseline SM across the language groups are shown in Table 1. The L1 Vietnamese group had the largest DID values, but they also exhibited the lowest average baseline SM. This means that they had more space to converge to the model relative to the other groups. Because of the small number of speakers in each group (2-4), it is not clear whether the differences in baseline SM are due to language background, individual differences, or other factors. However, it is possible to determine whether these baseline SM differences affected the DID results.

Another method of quantifying imitation that takes baseline distance into account is linear combination. A linear combination model following [12, 11] was constructed with post-exposure

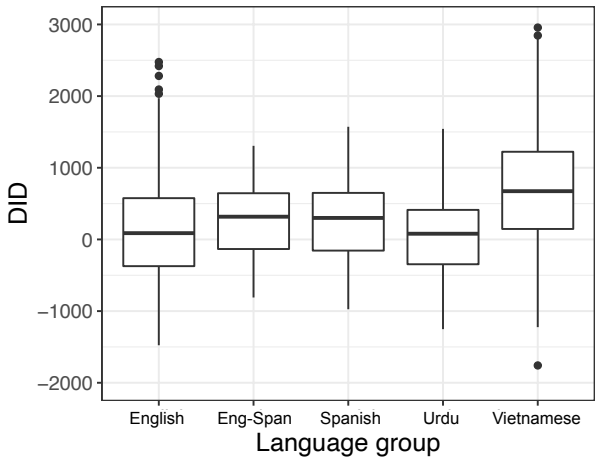


Figure 2: DID for each language group. Positive values indicate convergence.

Language group	Average baseline SM (Hz)
L1 English	7717
L1 English/Spanish	9127
L1 Spanish	7834
L1 Urdu	7791
L1 Vietnamese	6658
Model (enhanced)	10649

Table 1: Baseline SM for each language group and the model speech with enhanced SM.

SM as the dependent variable: Post SM $\sim$ Baseline SM + Model SM + Language Group + Model SM:Language Group + Model SM:Lexical Frequency + (1|speaker) + (1|word). In this model, baseline SM is significant ($\beta = 0.33, p < 0.001^{***}$) indicating that speakers are overall consistent between baseline and post-exposure blocks. Model SM is also significant ($\beta = 0.30, p < 0.001^{***}$) indicating overall convergence to the model talker in the post-exposure block. The interaction between Model SM and Frequency remains insignificant, indicating no significant difference in convergence based on lexical frequency. Unlike in the DID model, none of the Language Group terms or their interactions reached significance. Crucially, the interaction between Model SM and the L1 Vietnamese group is not significant ($\beta = 0.03, p = 0.81$). This means that the degree of convergence for the L1 Vietnamese group does not differ significantly from the intercept L1 English group.

4. DISCUSSION AND CONCLUSION

This study investigated spontaneous imitation of English /s/ across participants with different language backgrounds. After exposure to a native model talker with enhanced SM on /s/, participants in all groups increased their own SM to converge towards the model talker. When analyzing the results using the common DID measure, it appeared that the L1 Vietnamese group converged more than the other language groups. However, when the same results were analyzed using linear combination, there were no significant differences in degree of convergence across groups. This is because the L1 Vietnamese speakers had a lower average baseline SM.

Methodologically, these results provide another test case for the 'starting distance bias' associated with DID in phonetic imitation research. Using DID led to an epiphenomenal significant effect because differences in baseline acoustic values were correlated with other variables of interest (in this case, language background). It is not the case that the L1 Vietnamese speakers imitated the model talker more. Those speakers simply had baseline values that were more distinct from the model, which inflated DID values. These results therefore provide support for methods like linear combination, which take baseline values into account when quantifying degree of convergence.

Empirically, enhancement of the English alveolar sibilant /s/ was imitated across all language backgrounds examined. These results suggest that imitation is not necessarily inhibited or facilitated by a mismatch in L1 status between the participant and model talker when the relevant phonological category structure is similar across the two languages.

5. REFERENCES

[1] S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access." *Psychological Review*, vol. 105, no. 2, p. 251, 1998.

[2] K. Nielsen, "Specificity and abstractness of VOT imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, 2011.

[3] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *Journal of Phonetics*, vol. 40, no. 1, pp. 177–189, 2012.

[4] F. Jiang and S. Kennison, "The Impact of L2 English Learners' Belief about an Interlocutor's English Proficiency on L2 Phonetic Accommodation," *Journal of Psycholinguistic Research*, vol. 51, no. 1, pp. 217–234, 2022.

[5] J. E. Flege and W. Eefting, "Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation," *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 729–740, 1988.
[6] M. Llompart and E. Reinisch, "Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts," *Language and Speech*, vol. 62, no. 3, pp. 594–622, 2019.
[7] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," *Laboratory Phonology*, vol. 2, no. 1, pp. 125–156, 2011.
[8] M. Babel, "Dialect convergence and divergence in New Zealand English," *Language in Society*, vol. 39, pp. 437–456, 2010.
[9] A. Walker and K. Campbell-Kibler, "Repeat what after whom? exploring variable selectivity in a cross-dialectal shadowing task," *Frontiers in Psychology*, vol. 6, p. 546, 2015.
[10] M. Zając and A. Rojczyk, "Imitation of English vowel duration upon exposure to native and non-native speech," *Poznan Studies in Contemporary Linguistics*, vol. 50, no. 4, pp. 495–514, 2014.
[11] B. MacLeod, "Problems in the difference-in-distance measure of phonetic imitation," *Journal of Phonetics*, vol. 87, p. 101058, 2021.
[12] U. C. Priva and C. Sanker, "Limitations of difference-in-difference for measuring convergence," *Laboratory Phonology*, vol. 10, no. 1, 2019.
[13] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The celex lexical database, release 2," 1996.
[14] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English lexicon project," *Behavior research methods*, vol. 39, no. 3, pp. 445–459, 2007.
[15] Audacity Team, *Audacity(R): Free Audio Editor and Recorder*, http://audacity.sourceforge.net/, 1999-2021.
[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5:9/10, pp. 341–345, 2001.
[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner [computer program] version 0.9.0, retrieved from http://montrealcorpustools.github.io/montreal-forced-aligner/." 2017.
[18] C. DiCanio, "Time averaging for fricatives 2.0," Praat script published online, 2021.

Measuring Gradient Effects of Alcohol on Speech with Neural Networks' Posterior Probability of Phonological Features

[1]Ratree Wayland, [1, 2]Kevin Tang, [3]Sophia Vellozzi, [1]Fenqi Wang and [3]Rahul Sengupta

[1]Department of Linguistics & [3]Department of Computer, Information Science and Engineering, University of Florida; [2]Department of English Language and Linguistics, Heinrich-Heine-Universität Düsseldorf
ratree@ufl.edu, kevin.tang@hhu.de, s.vellozzi@ufl.edu, fenqi@ufl.edu, rahulseng@ufl.edu

ABSTRACT

Alcohol is known to impair fine articulatory control and movements. In drunken speech, incomplete closure of the vocal tract can result in deaffrication of the English affricate sounds /tʃ/ and /dʒ/, spirantization (fricative-like production) of the stop consonants and palatalization (retraction of place of articulation) of the alveolar fricative /s/ (produced as /ʃ/). Such categorical segmental errors have been well-reported. This study employs a phonologically-informed neural network approach to estimate degrees of deaffrication of /tʃ/ and /dʒ/, spirantization of /t/ and /d/ and place retraction for /s/ in a corpus of intoxicated English speech. Recurrent neural networks were trained to recognize relevant phonological features [anterior], [continuant] and [strident] in a control speech corpus. Their posterior probabilities were computed over the segments produced under intoxication. The results obtained revealed both categorical and gradient errors and, thus, suggested that this new approach could reliably quantify fine-grained errors in intoxicated speech.
Keywords: alcohol, deaffrication, palatalization, retraction, neural network.

1. INTRODUCTION

Alcohol intoxication has been shown to impair cognitive function and production of both suprasegmental and segmental properties of speech [1, 2, 3, 4, 5, 6, 7]. In English, the segmental errors encompass deaffrication of the affricate sounds /tʃ/ and /dʒ/ [5, 7], spirantization (fricative-like production) of the stop consonants [5, 7] and palatalization (retraction of place of articulation) of the alveolar fricative /s/ (produced as /ʃ/) [5, 9]. These errors occur due to impaired ability to control timing and movement of the active articulators, resulting in failure to form a complete closure (deaffrication and spirantization) or to achieve and/or maintain an appropriate degree of opening, at an intended location in the vocal tract (place retraction). Crucially, these errors have been described as categorical through such coarse-grained measures as perceptual judgment [e.g., 8] and phonetic transcription of the affected speech segments [e.g., 2]. However, sub-contrastive or gradient errors, below the level of a segment or feature are likely missed by such techniques due to the listeners' perceptual bias [10]. Further, while acoustic measurements could circumvent perceptual bias, describing categorical rather than gradient phonetic errors have largely been the focus of most acoustic-phonetic studies of intoxicated speech.

The goal of this study is to examine the nature, gradient and categorical errors, at the feature level in intoxicated speech in an English corpus using a neural network model known as Phonet. Inspired by computational approaches using forced alignment to measure surface, gradient phonetic variations, this approach quantifies gradient phonetic variation of deaffrication, spirantization and place retraction from the posterior probability of relevant phonological features, computed directly from the input signals by bidirectional recurrent neural networks. In this study, the relevant phonological features are [anterior], [continuant] and [strident]. These features capture relative location of the oral constriction, amount of airflow impedance and intensity of frication noise, respectively. A categorical error is operationalized as a sign shift of the phonological features (e.g., from [+anterior] to [-anterior]), while gradiency of an error is reflected in the posterior probability values of a phonological feature. A brief description of Phonet is present in § 2 below.

2. PHONET

Phonet [11] is a bi-directional recurrent neural network model, trained to recognize input phones as belonging to different phonological classes defined by phonological features (e.g., anterior, continuant and strident). It is semi-automatic and only requires a segmentally-aligned acoustic corpus (using forced alignment). Input to Phonet is log-energy distributed across triangular Mel filters computed from 25-ms windowed frames of each 0.5 second chunk of the input signal (see [11] for details). Once trained, posterior probabilities for different phonological features of the target segments can be computed by the model. Phonet has been found to be highly accurate in quantifying degree of lenition in Spanish [7, 15, 21, 22] and modelling the speech impairments of patients diagnosed with Parkinson's disease [11].

The architecture of Phonet is described in detail in [11].

Phonet can be customized with different sets of phonological features and acoustic representations. In this study, we focus on the probability of the phonological features [anterior], [strident] and [continuant] to capture retraction of place of articulation, deaffrication, and spirantization of English voiceless and voiced stops /t, d/, voiced and voiceless affricates /tʃ, dʒ/, and fricatives /s, ʃ/. A fricative-like realization due to an incomplete closure of the oral constriction would be associated with a relatively higher [continuant] and [strident] probability while a relatively low [anterior] probability would indicate a more retracted place of articulation.

3. METHODS

3.1. Materials

The target consonants for this study are English stops /t, d/ (Ns = 2,237 and 1,085), affricates /tʃ, dʒ/ (Ns = 144 and 160), and fricatives /s, ʃ/ (Ns = 1,430 and 96) from a corpus of intoxicated speech [7].

3.2. Stimulus recording procedure

The corpus contains recordings of four female, native speakers of British English reading a dialogue naturally (i.e., not in an animate, acting voice). The original text of the dialogue (based on [17]) was edited to ensure that it is gender- and emotionally neutral, void of overly long turns, and representative of the English phonemic inventory (available at [18]). Two separate readings (sober then drunk) across participants, on different days (1-2 months apart) were recorded in a sound-attenuated room at 44.1 kHz sampling rate and 16-bit amplitude resolution in stereo and were then converted to mono using Audacity. The speakers were told not to eat, drink, or use mouthwash 2 hours before each session and not to smoke half hour before each session.

The participants' blood alcohol concentration (BAC) was measured using a breathalyzer [AlcoMate (Macomb Township. MI) Premium AL-7000] at the beginning of the sober session to ensure absence of alcohol in their system. Intoxicated recording session began when BAC reached 0.12% after consumption of vodka or rum, mixed with juices.

3.2. Stimulus pre-processing

The recordings were divided up into utterances which were then manually annotated for any disfluencies. The disfluent utterances (8.5%) were not discarded since their exclusion did not qualitatively change our findings. The utterances were forced aligned using the Montreal Forced Aligner (version: 2.0) [12] with its released pretrained English model.

3.3. Phonet training procedure

Librispeech [16], a large corpus of English audiobooks was used as a representative English speech sample. A subset of the cleaned portion of 360 hours was selected. The corpus was then forced aligned using the Montreal Forced Aligner (version: 2.0) [12]. The phone set was set to IPA. For other parameters default values were used.

Model training with Phonet was performed on an NVIDIA GeForce RTX 3090 GPU using the Keras [19] library. The corpus was randomly split into a train subset (80%) and a test subset (20%) using the Python (Version 3.9) scikit-learn library [14]. Twenty-one Phonet models were trained for 20 phonological classes (consonantal, syllabic, voicing, labial, coronal, dorsal, lateral, nasal, rhotic, anterior, continuant, sonorant, strident, diphthong, high, low, back, round, stress, tense), and pause.

The model was highly accurate showing unweighted average recall (UAR) ranges from 91% (coronal)-98% (pause). Critically, the UARs for the anterior, continuant and strident features are 93%, 92% and 97%, respectively. The model was then applied to our selected word tokens from our forced-aligned intoxicated speech corpus with /t, d, tʃ, dʒ, s, ʃ/. The predictions were computed for 10-ms frames. For a token containing multiple frames, the average prediction from the middle frame(s) was taken as its prediction. Anterior, continuant and strident posterior probabilities obtained for each target consonant are then used for statistical analyses.

3.4. Statistical analyses

All statistical analyses were performed using the lme4 package [13] in R [13]. Contrast coding (-0.5, 0.5) was used for binary categorical variable. Random variables included speaker and word. Two complementary analyses were performed. First, to examine if posterior probabilities could predict drinking status, for each target consonant, a binary logistic regression analysis was performed with the three posterior probabilities (anterior, continuant and strident) as predictors and drinking status (sober and intoxicated) as dependent variable using the *glmer* function. A contrastive or categorical error was inferred when a feature emerged as the significant predictor. Second, to evaluate gradiency of an error, drinking status was entered as predictors in linear regression models (*lmer*) performed to investigate its predictions on the three posterior probability values for each target consonant. Increase or decrease in

posterior probability of a feature indicate degrees of error gradiency. For both analyses, "drunk" was the reference level. Based on previous literature [5, 7], a higher continuant and strident posterior probability is expected for /t, d/ and /tʃ, dʒ/ while a lower anterior probability is expected for the fricative /s/ in the intoxicated relative to the sober condition [5, 9].

4. RESULTS

Results of the binary logistic regressions analyses with anterior, continuant and strident posterior probabilities as predictors and drinking status as the categorical, binary dependent variable are summarized in Table 1.

Cons.	Predictor	Odds Ratios	*P* value
/t/	Anterior	1.11	0.615
	Continuant	**0.43**	**<.001**
	Strident	0.88	0.353
/d/	**Anterior**	**2.14**	**0.046**
	Continuant	**0.46**	**0.009**
	Strident	1.07	0.760
/tʃ/	Anterior	1.88	0.437
	Continuant	**0.18**	**0.023**
	Strident	1.98	0.641
/dʒ/	Anterior	1.17	0.866
	Continuant	**0.01**	**0.015**
	Strident	1.77	0.706
/s/	Anterior	2.97	0.471
	Continuant	0.62	0.473
	Strident	14.35	0.153
/ʃ/	Anterior	0.80	0.784
	Continuant	0.82	0.847
	Strident	1.51	0.954

Table 1: Results of the binary logistic analyses.

From this table, we can see that continuant posterior probability emerged as the only significant predictor of drinking status for /t/, /tʃ/ and /dʒ/. Specifically, as the continuant probability increases, the likelihood that the speakers were sober decreases (odds ratios <0.5). For /d/, both continuant and anterior probabilities are the significant predictors. However, as expected, as the anterior probability increases, the likelihood of the sober status increases (odds ratios >0.5). No significant predictor was found for /s/ [odds ratios = 0.62-14.35, p>.005] and /ʃ/ [odds ratios = 0.80-1.51, p>0.05]. However, the >0.5 odd ratios indicated that as the anterior, continuant and strident probabilities increase, the likelihood of the sober status also increases. These results suggested that categorical errors (i.e., [-continuant] > [+continuant]) occurred under intoxication for /t/, /d/, /tʃ/ and /dʒ/. Additionally, for /d/, a categorical shift from [+anterior] > [-anterior] also occurred. On the other hand, no categorical error was detected for /s/ or /ʃ/.

Tables 3a, b and c summarize the results of the linear mixed-effect regression models with drinking status as predictors (reference level = drunk) and posterior probabilities of the three phonological features as the dependent variables.

The results obtained indicated that a significantly higher anterior probability for /t/, /d/ and /s/ [βs= 0.029, 0.042, 0.013; *p*s≤.001] is predicted for the sober speech relative to the drunken speech, but a non-significant change in anteriority between the two speech conditions is predicted for /tʃ/, /dʒ/, and /ʃ/ [βs =0.009, -0.037, -0.014; *p*s= >.0.05] (Table 2a). β values suggested that sober /tʃ/ is more anterior than drunk /tʃ/ while sober /dʒ/ and /ʃ/ are less anterior than drunk /dʒ/ and /ʃ/. In other words, a shift in place of articulation is significantly greater for /t/, /d/ and /s/ than for /tʃ/, /dʒ/, and /ʃ/ under intoxication.

Predictor	Consonant	β	*P* value
Anterior	**/t/**	**0.03**	**0.001**
	/d /	**0.04**	**<0.001**
	/tʃ/	0.009	0.830
	/dʒ/	-0.037	0.399
	/s/	**0.01**	**0.001**
	/ʃ/	-0.014	0.746

Table 2a: Summary of the linear regression models for anterior probability.

For continuant probability (Table 2b), a significantly lower value is predicted for /t/, /d/, /tʃ/ and /dʒ/ [βs=-0.098, -0.061, -0.102, -0.068; *p*s≤.01], but not for /s/ [$\beta = 0.009, p=0.098$] or /ʃ/ [$\beta = -0.005, p = 0.88$] under the sober condition compared to the drunk condition. β value is positive for /s/, but negative for /ʃ/, indicating that /s/ is more continuant when sober while the opposite is true for /ʃ/. These results suggest that oral constriction for stops, and affricates became significantly less complete under intoxication. In contrast, oral constriction size was not significantly altered for the two [+continuant] consonant, /s/ and /ʃ/, when the speakers became inebriated.

Predictor	Consonant	β	*P* value
Continuant	**/t/**	**-0.098**	**<0.001**
	/d /	**-0.061**	**<0.001**
	/tʃ/	**-0.103**	**0.013**
	/dʒ/	**-0.068**	**0.003**
	/s/	0.009	0.098
	/ʃ/	-0.005	0.880

Table 2b: Summary of the linear regression models for continuant probability.

Finally, Table 2c shows that strident probability for /t/ was predicted to be significantly lower under the sober condition while the opposite is true for /s/. No significant difference [p>.05] was predicted for the remaining consonants. Although statistically non-significant, β values for /d/ [-0.022] and /tʃ/ [-0.017] are negative suggesting less stridency in their production when sober than when drunk while that of /dʒ/ [0.022] is positive and that of /ʃ/ [0.000] equals to 0, suggesting more stridency for /dʒ/ but no change in stridency for /ʃ/ when sober. These results suggest that /t/ is significantly less strident (less turbulent noise) when produced under the sober condition. On the contrary, drunk /s/ is less strident than its sober version. Furthermore, minimal change in degrees of stridency is observed for /d/, /tʃ/ and /dʒ/ while no change is predicted for /ʃ/.

Predictor	Consonant	*β*	*P* value
Strident	**/t/**	**-0.064**	**<0.001**
	/d /	-0.022	0.175
	/tʃ/	-0.017	0.492
	/dʒ/	0.022	0.451
	/s/	**0.014**	**0.001**
	/ʃ/	0.000	0.943

Table 2c: Summary of the linear regression models for strident probability.

5. DISCUSSION AND CONCLUSION

To detect categorical and gradient errors in intoxicated speech, a new computational approach, Phonet, was applied to a corpus of intoxicated English speech. Target error types are deaffrication, spirantization and retraction of place of articulation whose degrees of variation are estimated from posterior probabilities of three phonological features, [anterior], [continuant], [strident].

Binary logistic regression models yielded [continuant] as the significant predictor of drinking state for /t/, /tʃ/ and /dʒ/ while both [continuant] and [anterior] emerged as the significant predictors for /d/. If a binary, categorical shift of a feature is responsible for its significant predictive power, then, these results suggested that categorical errors ([-continuant] > [+continuant]) occurred for /t/, /d/, /tʃ/ and /dʒ/ under intoxication, suggesting that the size of the oral constriction contrastively shifts from a sober to a drunk state. For /d/, a categorical shift from [+anterior] > [-anterior] also occurred in drunken state, indicating that a concurrent and significant amount of place retraction also took place. The fact that /tʃ/ and /dʒ/ are [-anterior] may account for why further place retraction did not occur. In turns, neutralization (loss of contrastivity) in anteriority between /t/ and /tʃ/ could account for why /t/ did not undergo place retraction since they would both be [-anterior] if /t/ retracted. This suggest that articulatory planning may be intact, but the fine-grained motor control is partially lost when intoxicated.

Surprisingly, no categorical error was committed for /s/ or /ʃ/ under intoxication, at least not at the BAC level tested. Previous literature led to an expectation that a categorical shift in place of articulation would occur for /s/ (i.e., [+anterior] > [-anterior]) [5, 9]. It is possible that this error is only attested at a higher BAC level. However, it is worth noting that /s/ and /ʃ/ are both [+continuant, + strident]. It is possible that these shared and redundant features "add additional motoric instructions to enhance the saliency of the jeopardized features" [20, p. 33], namely [anterior] in this case. Nonetheless, the fact that this error has been previously reported suggested a limit of this enhancement effect.

Gradient errors are revealed by linear regression analyses. For instance, a shift in place of articulation was found to be significantly greater for /t/, /d/ and /s/ than for /tʃ/, /dʒ/, and /ʃ/ under intoxication. These results suggest that degree of place retraction in intoxicated speech is constrained by the existing place feature of the affected consonants: [+anterior] consonants will sustain a greater degree of place shift than [-anterior] consonants.

A similar constraint is observed for [continuant] leading to gradiency in continuant error. Continuant posterior probability significantly increased for [-continuant] consonants, /t/, /d/, /tʃ/, and /dʒ/, but not for [+continuant] consonant, /s/ and /ʃ/. This is due, perhaps, to a higher degree of continuance (greater oral aperture) which would lead to reduction of intensity of frication noise (i.e., stridency).

Finally, gradient errors involving [strident] was also found. Sober /t/ is significantly less strident than drunk /t/, suggesting an incomplete closure resulting in stridency. In contrast, a change in stridency was relatively small for /d/, /tʃ/ and /dʒ/. These results suggest that size of oral opening of [-continuant] consonants could increase. In addition, drunk /s/ is significantly less strident than its sober variant, suggesting a further widening of the oral aperture leading to a loss in stridency. However, no change was observed for /ʃ/. This result suggests that oral aperture could further widen for [-anterior, +continuant, +strident], /s/, but not for [+anterior, +continuant, +strident], /ʃ/.

Both categorical and gradient errors are revealed by Phonet, suggesting that it could reliably quantify fine-grained errors in intoxicated speech. Our findings need to be confirmed with more subjects, and can be extended to languages with different contrastive phonological features [7], and compared

with speech by clinical populations, such as Parkinson's disease.

Acknowledgement: This research was supported by NSF-National Science Foundation (SenSE). Grant No. 2037266. We acknowledge the data contributions of Jeremy Grynpas and Andrew Nevins.

6. REFERENCES

[1] Arbuckle, T. Y., Chaikelson, J. S., Gold, D. P. 1994. Social drinking and cognitive functioning revisited: the role of intellectual endowment and psychological distress. *Journal of studies on alcohol*, 55(3), 352-361.

[2] Pisoni, D. B., Martin, C. S. 1989. Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, 13(4), 577-587.

[3] Pisoni, D. B., Johnson, K., Bernacki, R. H. 1991, September. Effects of alcohol on speech. *In Proc. of the Human Factors Society Annual Meeting* (Vol. 35, No. 10, pp. 694-698). Sage CA: Los Angeles, CA: SAGE Publications.

[4] Swartz, B. L. 1992. Resistance of voice onset time variability to intoxication. *Perceptual and motor skills*, 75(2), 415-424.

[5] Lester, L.; Skousen, R. 1974. The Phonology of Drunkenness. In: Bruck, A.; Fox, RA.; LaGaly, MW. (eds.), *Papers from the Parasession on Natural Phonology*. Chicago: Chicago Linguistic Society.

[6] Hollien, H., Dejong, G., Martin, C. A., Schwartz, R., Liljegren, K. 2001. Effects of ethanol intoxication on speech suprasegmentals. *J. Acoust. Soc. Am.*, 110(6), 3198-3206.

[7] Tang, K., Chang, C. B., Green, S., Bao, K. X., Hindley, M., Kim, Y. S., & Nevins, A. 2022. Intoxication and pitch control in tonal and non-tonal language speakers. *JASA Express Letters*, 2(6), 065202.

[8] Zihlmann, U. 2017, August. The Effects of Real and Placebo Alcohol on Deaffrication. In *Interspeech* (pp. 3882-3886).

[9] Johnson, K., Pisoni, D. B., Bernacki, R. H. 1990. Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*, 47(3-4), 215-237.

[10] Frisch, S.A., Wright, R. 2002. The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, 30(2), 139-162.

[11] Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., Nöth, E. 2019. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In *Interspeech*, 549-553.

[12] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. August. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, 498-502.

[13] R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 969.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... , Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.

[15] Wayland, R., Tang, K., Wang, F., Vellozzi, S., and Sengupta, R. 2022. Lenition measures: Neural networks' posterior probability vs. acoustic cues. *Proc. of Meetings on Acoustics*, vol. 50, no. 1, p. 060002.

[16] Panayotov, V., Chen, G., Povey, D., Khudanpur, S. 2015, April. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206-5210. IEEE.

[17] Simon, N. 1986. *The Collected Plays of Neil Simon: Volume 2*. New York, NY: Plume.

[18] Tang, K., Chang, C. B., Green, S., Bao, K. X., Hindley, M., Kim, Y. S., & Nevins, A. 2022, "Materials for Tang et al. (2022)" Available: osf.io/y2r87.

[19] Chollet, F., others. 2015. Keras. GitHub. Retrieved from https://github.com/fchollet/keras.

[20] Keyser, S. J., Stevens, K. N. 2006. Enhancement and overlap in the speech chain. *Language*, 33-63.

[21] Tang, K., Wayland, R., Wang, F., Vellozzi, S., Sengupta, R., & Altmann, L. 2023. From sonority hierarchy to posterior probability as a measure of lenition: The case of Spanish stops. *The Journal of the Acoustical Society of America*, 153(2), 1191-1203.

[22] Wayland, R., Tang, K., Wang, F., Vellozzi, S., & Sengupta, R. (2023). Quantitative Acoustic versus Deep Learning Metrics of Lenition. Languages, 8(2), 98.

DIFFERENT SENSITIVITY TO TEMPORAL AUDITORY FEEDBACK PERTURBATION IN ADULTS WHO STUTTER: SYLLABLE STRUCTURE EFFECTS

Miriam Oschkinat, Philip Hoole

Institute for Phonetics and Speech Processing (IPS), LMU Munich
miriamo@phonetik.uni-muenchen.de

ABSTRACT

Ten adults who stutter (AWS) and ten fluent speakers were tested in three conditions with a real-time temporal auditory feedback perturbation paradigm (Onset, Vowel, and Coda Condition). In each condition, the perturbation target was stretched (either the onset consonant, the vowel, or the coda consonant of the German mono-syllabic word "Schaf", /ʃa:f/, *sheep*). No compensatory shortening for either of the stretched segments was observed, but both groups lengthened the segment(s) following the perturbed segment in the Onset and the Vowel Condition (reactive feedback control). In comparing both groups, the clearest tendency was that in the Onset Condition, AWS showed earlier and stronger responses by lengthening the following vowel. No group differences were observed in the Vowel or the Coda Condition. The data support the idea that persons who stutter use auditory feedback to a greater extent in onsets than fluent speakers and rely less on their internal representation.

Keywords: Temporal auditory feedback perturbation, stuttering, reactive feedback control, adaptation

1. INTRODUCTION

The integration of auditory feedback into the speech production process has proven crucial in the acquisition, planning, and control of fluent speech. In the online control of ongoing speech movements, the auditory feedback is used to compare the produced outcome against a stored representation. If the acoustic outcome does not match the prediction, corrections can be made in the online control with a latency of 120 to 200 ms (online compensation), or in future productions (adaptation). In adults who stutter (AWS), it is assumed that prolongations, blocks or repetitions occur because of a malfunctioning integration of sensory feedback information into the speech production process [1-3]. This assumption is supported by studies that found enhanced fluency in AWS under conditions of delayed auditory feedback [4]. Further, auditory feedback perturbation studies showed that online responses [5] or adaptation [6-8] to shifted vowel formants are weaker in AWS than in fluent speakers, supporting the idea that AWS rely less on auditory feedback and show deficits in integrating the auditory feedback into the speech production process. However, these studies investigated spectral parameters in syllable nuclei (vowels). Since stuttering symptoms occur in onsets rather than in syllable nuclei/codas, it is assumable that the integration of auditory feedback is differently weighted in AWS. [9] suggested that AWS rely too heavily on auditory feedback in onsets as compared to fluent speakers, who rely more on their internal representations. The latter was also suggested by [10-12] since fluent speakers showed no significant reaction to *temporally* perturbed syllable onsets, but to vowels and codas. To date, two studies investigated responses to temporal perturbations in adults who stutter. With unexpected random perturbations (online compensation paradigms), [13] found weaker responses to a time-shifted spectral vowel target (F2 minimum of /u/ in "owe" either accelerated or decelerated) in adults who stutter, [14] did not find a group difference between AWS and fluent speakers when the onset /s/ in the word "steady" was stretched followed by a compression phase later in the syllable. However, since online responses to a perturbation of speech timing are hardly feasible, the current study uses an adaptation paradigm to investigate responses to temporally perturbed auditory feedback in different parts of a syllable in AWS and fluent speakers. We expect different sensitivity to temporal real-time manipulations in AWS compared to fluent speakers with respect to syllable structure.

2. METHODS AND HYPOTHESES

Ten individuals who stutter (mean age 28.8 years, 3 females, mean age at onset of stuttering 5.5) and 10 fluent speakers (mean age 29.6 years, 3 females) were tested with a real-time temporal auditory feedback adaptation paradigm in Matlab using the Audapter software for real-time time-warping [15, 16]. Speakers spoke into a Sennheiser headset microphone and received auditory feedback via E-A-RTone 3A in-ear earphones with foam eartips. In each of three experimental conditions, speakers uttered the German phrase "mein Schaf" (*my sheep*) 95 times. The first 25 trials served as a Baseline with no perturbation,

followed by a Ramp phase with increasing perturbation over 20 trials up to maximum perturbation which was held for another 30 trials (Hold phase). In the last 20 trials, normal feedback was restored (After-effect phase). The perturbation target was an 80 ms sequence of the onset consonant /ʃ/ in "Schaf" in the Onset Condition, the vowel /a:/ in the Vowel Condition, or the coda consonant /f/ in the Coda Condition. The respective perturbation target was triggered by Audapter's online status tracking with heuristic rules searching for certain landmarks in the signal such as the onset of a fricative or a vowel with predefined intensity or pre-emphasized intensity thresholds. The online status tracking was implemented for the purposes of the current study and programmed to trigger late in the respective segment of interest to avoid manipulation of formant transitions. In each condition, the perturbed sequence was stretched by a factor of 1.9 during maximum perturbation in the Hold phase (perturbed sequence stretched by 72 ms to approx. 152 ms). The signal following the perturbed sequence was therefore delayed by the amount of stretching, and the "catch-up" or compression phase was implemented into the silence after the end of the utterance, eventually reverting the signal back to real-time. The implementation of only stretching the signal without compressing an immediately following part had two main goals: Firstly, the expected compensatory response to a stretched signal would be a shortening in production of the respective segment. This shortening would necessarily be adaptive, since an online shortening as response to a stretched signal is physically not possible. Secondly, examining the duration of segments after the perturbation target would indicate adjustments due to reactive feedback control. Based on previous findings, we do not expect adaptive shortening of the stretched onset target, but for the stretched vowel and the stretched coda. Reactive feedback control effects should be observable in lengthened segments after the perturbation target in the Onset and Vowel Condition. If AWS rely more on auditory feedback in onsets, we expect them to show greater reactive feedback control responses in the Onset condition than in the Vowel or the Coda Condition. Further, we expect greater responses of AWS to the Onset condition compared to fluent speakers. Secondly, if AWS show deficits in auditory-motor mapping, we expect AWS to be less efficient in adapting with shortening responses to the stretched perturbed segments than fluent speakers.

3. ANALYSES

Dysfluent and erroneous trials were removed before calculations. However, due to the speech material, almost none of the AWS produced dysfluent trials. All following analyses were performed in RStudio using mostly packages from the *tidyverse* [17]. The data were grouped by participant, condition, segment, and phase; outliers were removed. To estimate the difference between Baseline productions and responses to maximum perturbation during the Hold phase, linear-mixed models [18, 19] were calculated per group (fluent speakers vs. AWS) and condition (Onset, Vowel, and Coda Condition). The last 20 trials of the Baseline and the last 20 trials of the Hold phase were included into calculations. Models were fitted with absolute durations in ms as dependent variable, with phase (Baseline or Hold phase) and segment (onset consonant, vowel, coda consonant) as predictors as well as the interaction between phase and segment. Random effects included by-subject intercepts and random slopes for phase. Per model, the predictor *segment* included the perturbed and following segments. For the Coda Condition the inclusion of segment was dropped since there was only one segment to be analysed.
Conditional R-squared values of the full models were provided by MuMin's *r.squaredGLMM* function [20] and *emmeans'* [21] pairwise comparison revealed the difference between Baseline and Hold phase for the respective segment.

4. RESULTS

The following section reports the conditional R-squared values for the linear mixed models and the estimates along with the standard error (SE), degrees of freedom (df), t-ratio, and p-value as provided by *emmeans*' pairwise comparison for the contrast between Hold phase and Baseline. Positive estimates indicate longer, and negative estimates shorter productions in Hold phase than Baseline (see Fig.1).

4.1. Onset Condition

The linear mixed models revealed that in the Onset Condition, the Hold-Baseline contrast for the perturbed onset consonant was not significant for the group of fluent speakers (R-squared = 0.75, estimate = 9.17 ms, SE = 5.56, df = 22.2, t.ratio = 1.651, p-value = 0.113) but significant for the group of AWS (R-squared = 0.73, estimate = 22.8 ms, SE = 8.15, df = 16.6, t-ratio = 2.80, p-value = 0.0125*). Both groups significantly lengthened the following vowel in production (fluent speakers: estimate = 33.91 ms, SE = 5.52, df = 21.5, t-ratio = 6.15, p-value < .0001***; AWS: estimate = 57 ms, SE = 8.11, df =

16.3, t-ratio = 7.03, p < .0001***). The coda consonant was significantly lengthened by the group of fluent speakers (estimate = 20.58 ms, SE = 5.58, df = 22.5, t-ratio = 3.688, p-value = 0.0012**), but non-significantly lengthened by the group of AWS (estimate = 16 ms, SE = 8.12, df = 16.4, t-ratio = 1.998, p-value = 0.0626).

4.2. Vowel Condition

In the Vowel Condition, the Hold-Baseline contrast for the perturbed vowel was non-significant in both groups (fluent speakers: R-squared = 0.85, estimate = 3.98 ms, SE = 5.02, df = 15.5, t-ratio = 0.791, p-value = 0.44; AWS: R-squared = 0.72, estimate = 15.5 ms, SE = 7.79, df = 11.7, t-ratio = 1.987, p-value = 0.07). This result is rather unexpected (at least for the group of fluent speakers) since previous studies found shortening responses in reaction to a stretched vowel and will be further discussed in section 5. Both groups significantly lengthened the following coda consonant (fluent speakers: estimate = 22.63 ms, SE = 5.06, df = 16, t-ratio = 4.47, p-value = 0.0004***; AWS: estimate = 21 ms, SE = 7.81, df = 11.8, t-ratio = 2.692, p-value = 0.0199*).

4.3. Coda Condition

The contrast between Hold phase and Baseline for the perturbed coda consonant in the Coda Condition was non-significant for the group of fluent speakers (R-squared = 0.92, estimate = -6.88 ms, SE = 4.91, df = 9, t-ratio = -1.402, p-value = 0.194), as well as for the group of AWS (R-squared = 0.82, estimate = 2.35 ms, SE = 4.79, df = 9, t-ratio = 0.491, p-value = 0.635).

4.4. Group comparison of Hold phase productions

For the group comparison, t-tests were calculated between the group of AWS and fluent speakers with mean Hold phase durations relative to the mean of the last 20 Baseline trials per participant, segment, and condition, after checking the data for a normal distribution. The t-tests indicated no significant group differences for any of the segments per condition (see Fig. 2). Since the means are already provided in the preceding section, the statistical details of the t-tests will not be reported.

4.5. Qualitative analyses of Ramp phase and After-effects

The sensitivity to an auditory shift can be determined by examining the magnitude of perturbation that evokes a change in production. In the Ramp phase, the perturbation is gradually increased by 3.6 ms per trial up to a 72 ms stretch in the Hold phase. In Fig. 1, a sensitivity difference between the two groups is observable in the Onset Condition for the vowel and the onset consonant (left panels, magenta stars and yellow crosses): Both the onset consonant and vowel productions in the group of AWS start to lengthen earlier in the Ramp phase and increase linearly, while the productions for the group of fluent speakers change more slowly over trials, showing less steep curves. The coda consonant seems not to differ visually between groups. In the Vowel Condition, however, the group of fluent speakers seems to be more sensitive to the shift since they lengthen the following coda consonant sooner (right at the beginning at the Ramp phase) and to a greater extent

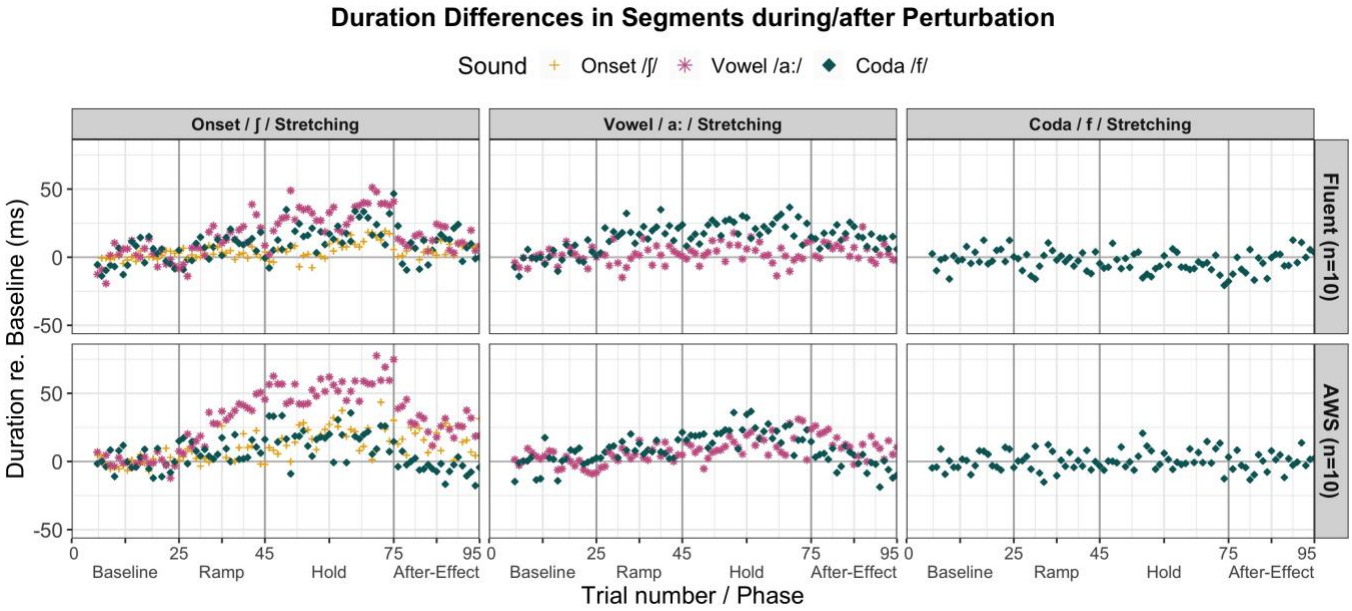


Figure 1: Mean durations (in ms) relative to the Baseline mean per group (upper panels: fluent speakers, lower panels: Adults who Stutter) and condition over the course of the experiment. Onset consonant productions in yellow crosses, vowel productions in magenta stars, and coda consonant productions in green rhombuses.

(middle panels, green rhombuses). Vowel productions themselves seem not to differ much between groups despite a lower general variability in vowel production in the group of AWS (magenta stars). In the Coda Condition, no group differences are observable in the Ramp phase.

The After-effect phase allows for the examination of adaptive behaviour when changes in production from the Hold phase remain. In the Onset Condition, vowel productions show a large drop from the Hold phase to the After-effect phase towards baseline productions in the group of fluent speakers, but remain longer throughout the After-effect phase, at least for the onset consonant and the vowel (upper left panel, magenta stars, yellow crosses). For the group of AWS, the remaining lengthening in the After-effect phase is stronger for vowel and onset consonant compared to the group of fluent speakers, indicating stronger adaptive behaviour. In the Vowel Condition, the group of fluent speakers seems to adapt the lengthened coda consonant (upper middle panel, green rhombuses), while for the group of AWS coda consonant productions revert to Baseline in the After-effect phase (lower middle panel, green rhombuses). Vowel productions are longer than Baseline productions in the After-effect phase for both groups. No differences are observable in the Coda Condition.

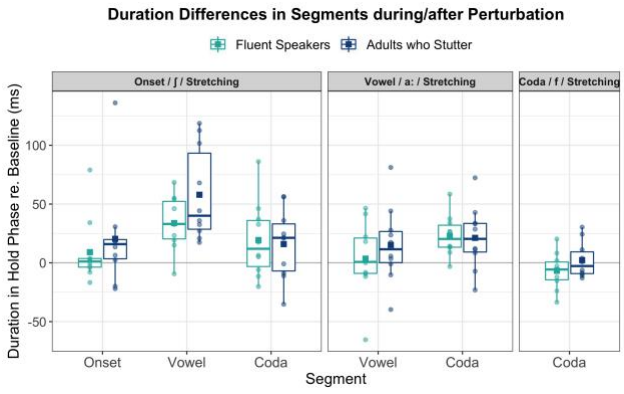


Figure 2: Mean difference in Hold phase relative to the Baseline mean per condition and segment. Left (turquoise) boxplots represent the group of fluent speakers, the right (dark blue) boxplots the group of AWS. Single participants are shown as dots. Boxes correspond to the first and third quartiles, bars represent the median and squares the mean. Whiskers extend from the hinge to the highest/smallest value no further than 1.5*IQR. Data beyond whiskers are outliers.

5. DISCUSSION

The current study investigated differences in responses to an auditorily stretched onset consonant, vowel, and coda consonant of a monosyllabic word in fluent speakers and AWS. Visual examination suggested that AWS show a greater sensitivity to temporally stretched onsets than fluent speakers, visible in earlier reactions in the onset and the following vowel during the Ramp phase. Analyses showed stronger lengthening of the following vowel (reactive feedback control) in AWS than in fluent speakers during maximum perturbation (difference between groups: 23 ms, but non-significant). The lengthening responses remained in the After-effect phase for both groups. In the Vowel Condition, no group compensatorily shortened the vowel in the Hold phase, and responses for the coda consonant were also similar with both groups lengthening the following coda consonant (~22 ms for both groups). The group of fluent speakers seemed to adapt (longer durations in the After-effect phase) while the group of AWS did not. In the Coda Condition, no significant effects were found.

Any group differences in the Hold phase turned out non-significant (section 4.4), presumably due to the currently rather small number of participants and fairly high variability in the data which might be reduced when the full dataset of 20 participants per group is tested/analysed. The findings of the current study support the idea that AWS rely more heavily on auditory feedback in syllable onsets than fluent speakers, and also rely more heavily on auditory feedback in onsets than in nucleus and coda of a syllable. These results are in line with the modelling studies in [9], and represent, to our knowledge, the first empirical demonstration by means of focal perturbation of the model predictions. Further, AWS seem to be able to adapt vowel productions in the Onset Condition, although these adaptive changes are rather an effect resulting from learned reactive feedback control than a direct (compensatory) response to a focal perturbation. The lack of adaptive shortening of the vowel in the Vowel Condition is rather unexpected (at least for the group of fluent speakers), since adaptive shortening for a stretched vowel was previously found in [10]. However, in [10], the perturbation caused a shift in phoneme category from /a/ to /a:/ in the auditory feedback, while the /a:/ in “Schaf” is already long and stretching might not cause an auditory error, or tolerances might be greater towards a longer vowel here.

In summary, the findings of the current study support the assumption of a different weighting of auditory feedback in AWS and fluent speakers with respect to syllable structure. The current investigation thus highlights the importance of taking syllable structure into account when studying the planning and control of speech timing, both in general, as well as in speech fluency impairments such as stuttering.

6. ACKNOWLEDGEMENTS

We would like to thank Nicole Benker, Anton Gadringer, Mona Franke and Charlie Wiltshire for their help in running the tests.

7. REFERENCES

[1] Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., and Wallace, M. E. 2004. "Unstable or Insufficiently Activated Internal Models and Feedback-Biased Motor Control as Sources of Dysfluency: A Theoretical Model of Stuttering," *Contemporary Issues in Communication Science and Disorders,* vol. 31, no. Spring, pp. 105-122, doi: doi:10.1044/cicsd_31_S_105.

[2] Hickok, G., Houde, J., and Rong, F. 2011. "Sensorimotor integration in speech processing: computational basis and neural organization," *Neuron,* vol. 69, no. 3, pp. 407-422.

[3] Harrington, J. 1988. "Stuttering, delayed auditory feedback, and linguistic rhythm," *Journal of Speech, Language, and Hearing Research,* vol. 31, no. 1, pp. 36-47.

[4] Andrade, C. R. F. d. and Juste, F. S. 2011. "Systematic review of delayed auditory feedback effectiveness for stuttering reduction," *Jornal da Sociedade Brasileira de Fonoaudiologia,* vol. 23, pp. 187-191.

[5] Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., and Perkell, J. S. 2012. "Weak responses to auditory feedback perturbation during articulation in persons who stutter: evidence for abnormal auditory-motor transformation," *PloS one,* vol. 7, no. 7, p. e41830.

[6] Daliri, A., Wieland, E. A., Cai, S., Guenther, F. H., and Chang, S. E. 2018. "Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter," *Developmental science,* vol. 21, no. 2, p. e12521.

[7] Kim, K. S. and Max, L. 2021. "Speech auditory-motor adaptation to formant-shifted feedback lacks an explicit component: Reduced adaptation in adults who stutter reflects limitations in implicit sensorimotor learning," *European Journal of Neuroscience,* vol. 53, no. 9, pp. 3093-3108.

[8] Kim, K. S., Daliri, A., Flanagan, J. R., and Max, L. 2020. "Dissociated Development of Speech and Limb Sensorimotor Learning in Stuttering: Speech Auditory-motor Learning is Impaired in Both Children and Adults Who Stutter," *Neuroscience,* vol. 451, pp. 1-21, doi: https://doi.org/10.1016/j.neuroscience.2020.10.014.

[9] Civier, O., Tasko, S. M., and Guenther, F. H. 2010. "Overreliance on auditory feedback may lead to sound/syllable repetitions: Simulations of stuttering and fluency-inducing conditions with a neural model of speech production," *Journal of Fluency Disorders,* vol. 35, no. 3, pp. 246-279, doi: https://doi.org/10.1016/j.jfludis.2010.05.002.

[10] Oschkinat, M. and Hoole, P. 2020. "Compensation to real-time temporal auditory feedback perturbation depends on syllable position," *J. Acoust. Soc. Am.,* vol. 148, no. 3, pp. 1478-1495, doi: 10.1121/10.0001765.

[11] Oschkinat, M. and Hoole, P. 2022. "Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables," *Journal of Phonetics,* vol. 91, p. 101133, doi: https://doi.org/10.1016/j.wocn.2022.101133.

[12] Karlin, R., Naber, C., and Parrell, B. 2021. "Auditory Feedback Is Used for Adaptation and Compensation in Speech Timing," *Journal of Speech, Language, and Hearing Research*, doi: doi:10.1044/2021_JSLHR-21-00021.

[13] Cai, S., Beal, D. S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. 2014. "Impaired timing adjustments in response to time-varying auditory perturbation during connected speech production in persons who stutter," *Brain and language,* vol. 129, pp. 24-29.

[14] Frankford, S. A., Cai, S., Nieto-Castañón, A., and Guenther, F. H. 2023. "Auditory feedback control in adults who stutter during metronome-paced speech I. Timing Perturbation," *Journal of Fluency Disorders,* vol. 75, p. 105943.

[15] Tourville, J. A., Cai, S., and Guenther, F. 2013. "Exploring auditory-motor interactions in normal and disordered speech," *Proceedings of Meetings on Acoustics,* vol. 19, 060180, pp. 1-8, doi: 10.1121/1.4800684.

[16] Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S., 2008. „A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/," in *Proceedings of the 8th ISSP*, 2008, pp. 65-68.

[17] Wickham, H. *et al.* 2019. "Welcome to the tidyverse," *Journal of Open Source Software,* vol. 4(43):1686, pp. 1-6, Art no. 43, doi: 10.21105/joss.01686.

[18] Bates, D., Maechler, M., Biolker, B., and Walker, S. 2015. "Fitting Linear Mixed-Effect Models Using lme4," *Journal of Statistical Software,* vol. 67(1), pp. 1-48, doi: 10.18637/jss.v067.i01.

[19] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software,* vol. 82(13), pp. 1–26, doi: 10.18637/jss.v082.i13.

[20] Bartoń, K. 2022. "MuMIn: Multi-Model Inference," *R package version 1.47.1 Available at https://cran.r-project.org/web/packages/MuMIn/index.html (last viewed: January 4, 2023).*

[21] Lenth, R. 2022. "emmeans: Estimated marginal means, aka least-squares means," *R package version 1.8.0. Available at: https://cran.r-project.org/package=emmeans (last viewed January 4, 2023).*

Auditory Acuity and Response to Altered Auditory Feedback in Children with and without Residual Speech Sound Disorder

Amanda Eads[1], Elaine Kearney[2], Asuka Koda[3], Elaine Hitchcock[4], Douglas Shiller[5], Frank Guenther[6], Tara McAllister[1]

[1]New York University, [2]Queensland University of Technology, [3]Bronx High School of Science, [4]Montclair State University, [5]University of Montreal, [6]Boston University
are326@nyu.edu; elaine.kearney@qut.edu.au; kodaa@bxscience.edu; hitchcocke@mail.montclair.edu; douglas.shiller@umontreal.ca; guenther@bu.edu; tkm214@nyu.edu

ABSTRACT

Children with residual speech sound disorder (RSSD) exhibit speech production patterns that differ from their dialect community and persist beyond the typical developmental window. Differences in auditory acuity or auditory feedback response may contribute to these persisting deviations. To assess this, the present study evaluated auditory feedback response in 31 children aged 9-15 with and without RSSD. Participants heard their speech in near real-time with the first formant frequency shifted upwards by 30%. We also administered an explicit behavioral measure of auditory-perceptual acuity for the phonetic contrast in the perturbation task.

The results indicate a trend toward greater compensation under altered feedback conditions by children with RSSD and a trend towards greater adaptation (i.e., persistence of learned changes after altered feedback was withdrawn) by children without RSSD. There were no significant differences in auditory-perceptual acuity between groups and no significant correlations between auditory-perceptual acuity and degree of compensation or adaptation.

Keywords: altered auditory feedback, sensory acuity, residual speech sound disorder, children

1. INTRODUCTION

Children with speech sound disorder (SSD) exhibit atypical speech patterns that negatively affect intelligibility, posing a barrier to participation in social and academic settings [1]. Delayed speech development typically resolves by 8-9 years old, but between 2-5% of speakers exhibit residual speech sound disorder (RSSD) that persists through adolescence or even adulthood [2].

The causes of these persisting speech deviations are still poorly understood, but it is known that typical speech production requires fine-grained representations of the sensory characteristics of different speech sounds, as well as the ability to update motor plans when sensory feedback indicates a mismatch with the intended target. Previous research shows that children with SSD/RSSD, on average, exhibit weaker speech perception than children with typical production [3, 4]. However, some children with SSD/RSSD perform in the average range on perceptual tasks, even for the contrast(s) they produce in an atypical fashion [3 -5]. It is possible that such children perceive the contrast between their productions and those of others in their environment, but they have difficulty updating their speech-motor routines in response to sensory feedback indicating a discrepancy.

This possibility was supported in previous research investigating compensation for perturbed auditory feedback in children with SSD aged 3-7 years [6]. In perturbation studies, specialized software [7] is used to transform the acoustics of participants' speech in near-real time, such that they hear their own voice with altered auditory-perceptual characteristics. On average, speakers tend to compensate for these real-time perturbations (i.e., they tend to shift their production in the opposing direction). Speakers also typically exhibit some degree of adaptation (i.e., changes in production that persist after the perturbation is removed), which is suggestive of updates to the feedforward plan for the speech sound in question. In a study of young children with and without SSD, Terband et al. [6] found that typically developing children were successful in compensating for perturbed auditory feedback affecting the vowel /e/, whereas children with SSD actually tended to follow the shift rather than compensate for it. This suggested that the children with SSD were successful in detecting the altered feedback but were unable to adjust their motor plans to offset the perturbation.

The present study examined both auditory-perceptual acuity and response to altered auditory feedback in an older population of children/ adolescents with and without RSSD. The children with RSSD specifically exhibited deviations affecting American English /ɹ/, a late-emerging sound that is commonly a target of speech intervention in later childhood. Based on previous literature, we hypothesized that we would observe statistically significant between-group differences in both

auditory-perceptual acuity and compensation/ adaptation in response to perturbed feedback. However, we also expected to observe individual heterogeneity underlying these group-level effects. In the future, we hope that individual differences in performance on such tasks can be leveraged to make customized recommendations for treatment for children with RSSD (e.g., input-oriented intervention if auditory-perceptual acuity is atypical; ultrasound biofeedback intervention if participants have intact perceptual discrimination but have difficulty updating a speech-motor plan in response to sensory feedback). In addition, we investigated the association between auditory-perceptual acuity and compensation/adaptation for altered auditory feedback. Based on previous research [8], we hypothesized that individuals with more acute auditory perception would exhibit a stronger response to perturbation.

2. METHODS

Data collection occurred in two different university laboratory settings with approval from a single external Institutional Review Board (BRANY 18-10-393). Written informed assent and parent permission were obtained for all participants.

2.1. Participants

The data reported in this study come from 31 children aged 9-15 years old who live in New York or New Jersey and speak a rhotic dialect of American English in the home. The children are split into two groups. One group (n=12, 10 male, 2 female) has a history of RSSD and participated in a 16-week treatment study to aid in their development of the English /ɹ/ sound. The second group (n=19, 6 male, 13 female) has no history of RSSD. Parental reports confirm no history of speech-language-hearing difficulties, with the exception of speech production difficulty for the RSSD group. For inclusion, participants were required to pass a bilateral pure-tone hearing screening (20 dB HL at 500, 1000, 200, and 4000 Hz) administered on the day of testing.

2.2. Auditory-Perceptual Acuity Tasks

To assess the participants' auditory-perceptual acuity, an identification and discrimination task focused on the same phonetic context (/hɛd/ - /hæd/) were used in combination. Perceptual stimuli were presented on a laptop computer over Sennheiser HD429 headphones in a sound-shielded booth.

2.2.1 Auditory-Perceptual Identification

Participants completed a perceptual identification task implemented in Experigen [9]. STRAIGHT synthesis [10] was used to generate an 11-step continuum between naturally produced tokens of *head* and *had*, elicited from a typical young adult female talker from the same area as the participants [11]. Participants were instructed to identify each token by clicking the word *head* or *had* in a forced-choice task. After a practice phase featuring the continuum endpoints, participants heard each continuum step 8 times in random order (total 72 trials). A break was provided halfway through.

Performance on the task was analyzed by fitting a logistic function over the number of *head* responses for each step of the continuum. The phoneme boundary is defined as the 50th percentile of probability, where *head* and *had* responses are equally likely [12].

2.2.2 Auditory-Perceptual Discrimination

Participants subsequently completed a gamified AXB discrimination task implemented in MATLAB. A 240-step *head-had* continuum (a superset of the continuum used in the identification task) was used. The starting point was set in an individualized fashion based on the boundary location determined in the identification task.

In a gamified interface, participants heard a standard stimulus (X) and were instructed to identify which of two flanking stimuli (A and B) differed from the central stimulus. The distance between A and B began at half the continuum width and was adjusted in an adaptive staircase design that shifted from 8-down 4-up to 2-down 1-up, with step distance halving after every 3 reversals; the task terminated after 9 reversals. The mean distance across the last three reversals was treated as the just noticeable difference (JND), where a higher JND indicates lower perceptual acuity.

2.3. Altered Auditory Feedback Tasks

2.3.1 Task design and administration

Participants completed reflexive and adaptive altered auditory feedback paradigms during the session. This paper will only report on the adaptive paradigm.

In a protocol modified from Daliri et al. [13], during the adaptive paradigm, Audapter software [7] was used to shift the first formant frequency (F1) values of the target vowel /ɛ/ upwards 30% in near real time. Participants wore a headset microphone and over-ear headphones so they could hear their altered auditory signal in near real time (estimated 27 ms delay [14]). The stimuli included the words *bed*, *head*, and *Ted* along with representative pictures. Participants were asked to produce the word as soon as the picture appeared on the screen and to hold out the vowel sound until the picture disappeared. They

were then given visual feedback regarding their production duration ("Too short", "Just right", or "Too long"). The window for "Just right" was 1100-1500 milliseconds. The stimuli were randomized in blocks of three so that the same stimuli would not be presented twice in a row.

The experimenter introduced the participants to the task during a practice session. The experimenter demonstrated the task expectations and elicited and gave feedback for 4 practice productions of "bed." Participants then repeated "bed" 3 more times during a voice calibration phase, which allowed the experimenter to monitor microphone levels and select the best Linear Predictive Coding (LPC) setting for the participant. Individualized LPC settings were necessary given the developmental age range of our participants [15]. Finally, the adaptive paradigm included a baseline phase of 18 trials, an altered F1 phase of 36 trials, and an aftereffect phase of 18 trials where the perturbation was removed. During the altered F1 phase, the perturbation turned on at voice onset and remained on for the duration of the trial.

2.3.2 Measurement

The results of the altered auditory feedback task were measured in MATLAB using FLVoice software [16]. Two of the authors reviewed both the waveform and spectrogram for all 72 trials per participant to ensure the quality of productions and formant traces. Trials were excluded if they were less than 700ms in duration, included a false start/misproduction, or featured audio distortions in the microphone or headphone signal. LPC settings were adjusted if the initially selected setting yielded unstable formant tracking. A custom MATLAB script extracted the average F1 value in the 350-550ms window for each trial. This time window is expected to capture both within-trial compensation and trial-to-trial adaptation.

3. RESULTS

3.1. Auditory-Perceptual Acuity

The boxplots in Figure 1 represent JND in the auditory discrimination task for each group. The median for the group with RSSD is slightly lower, suggesting a slight trend toward more acute performance in this group of children. However, an independent-samples t-test revealed no significant difference between the groups ($t(28.92) = 0.85$, $p = 0.4$). The boxplots in Figure 1 show two outliers in the non-RSSD group, which could be contributing to the appearance of the slight trend.

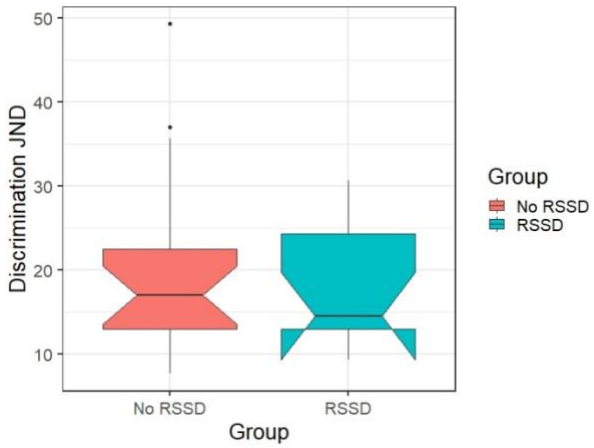


Figure 1: Just Noticeable Difference for each group in the auditory-perceptual discrimination task. Horizontal lines represent group medians, boxes represent the interquartile range, and outliers appear as points.

3.2. Altered Auditory Feedback Response

3.2.1 Altered F1 Phase

The boxplots in Figure 2 represent the average degree of compensation during the altered feedback phase for each group, relative to the baseline phase. The RSSD group has a lower median value than the non-RSSD group, indicating a greater degree of compensation during perturbation, and there is minimal overlap between the notches representing the confidence interval around the median for each group. However, an independent-samples t-test indicated that this difference was not statistically significant ($t(20.88) = 1.06$, $p = 0.3$).

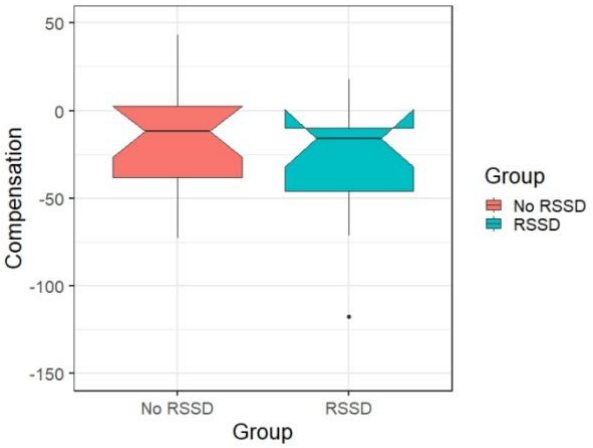


Figure 2: Percent compensation for each group during the altered F1 phase.

3.2.2 Aftereffect Phase

Figure 3 shows each group's average degree of adaptation in the first three trials after the perturbation was removed, relative to the baseline phase. In contrast with the previous plot, Figure 3 shows a greater degree of adaptation for non-RSSD children than children with RSSD. One-sample t-tests were

used to test if the degree of adaptation after the perturbation was withdrawn was significantly different from zero. The difference was significant for the group of children without RSSD ($t(18) = -2.72$, $p = .001$), but not for the group of children with RSSD ($t(11) = -1.04$, $p = .16$). However, the difference between groups was not statistically significant ($t(27.83) = -1.24$, $p = .22$).

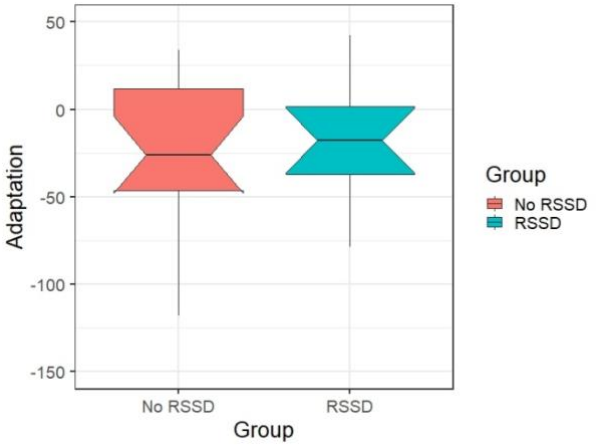


Figure 3: Percent adaptation for each group after withdrawal of the perturbation (aftereffect phase).

3.2. Auditory Acuity and Altered Auditory Feedback Response

Finally, we examined possible correlations between auditory acuity and the auditory feedback responses (percent compensation and adaptation). There was no significant correlation between discrimination JND and auditory feedback response either during the altered F1 phase ($r(29) = 0.18$, $p = 0.35$) or aftereffect phase ($r(29) = 0.05$, $p = 0.78$).

4. DISCUSSION

The results reported here are preliminary and must be interpreted with caution in light of a number of limitations. The number of participants is small, and the group of children with RSSD is roughly half the size of the group without RSSD. At the present time, the study is also limited in that the groups differed in sex breakdown, with a larger percentage of male participants in the RSSD group. We are continuing to recruit children with RSSD in connection with an ongoing clinical trial and will re-evaluate group differences after parity in group size is achieved.

In our preliminary sample, we found no significant difference in auditory acuity between the groups of participants with and without RSSD. While this was contrary to hypothesis, it was not entirely surprising, since we assessed perception using a sound contrast, /ɛ/-/æ/, that participants in the RSSD group did not have trouble articulating. Some previous literature has suggested that perceptual deficits in SSD may be specific to the contrasts that the speakers produce in error [17], although there is a lack of consensus on this point [3, 4].

In these preliminary data, children with RSSD showed a greater degree of compensation during perturbation than children without RSSD, although the difference was not statistically significant. This differs from a previous study of younger children [6], where a significant degree of compensation was observed in children with typical speech, but no compensation was observed in children with SSD. Given that the present study examined older children with a relatively mild speech distortion, it is not entirely surprising that our findings differed relative to [6]. Interestingly, whereas the RSSD group showed a numerically greater degree of compensation while the perturbation was in place, the group without RSSD showed a greater degree of adaptation after the perturbation was withdrawn. One-sample t-tests indicated that the magnitude of adaptation was significantly different from zero for the group without RSSD but not the group with RSSD, although the difference between the groups was not statistically significant. If this result is found to be robust when the complete sample is collected, it could constitute a suggestion that children with RSSD have more difficulty updating feedforward commands than children with typical speech production, and they may tend to rely correspondingly more on feedback in speech-motor control. This is broadly consistent with previous work on a different type of speech sound disorder, Childhood Apraxia of Speech [18].

5. CONCLUSION

While the current results only show group trends rather than statistically significant differences, they point to an interesting potential contrast between compensation during perturbation and adaptation after perturbation is withdrawn. Data collection is ongoing to achieve parity in group size and sex distribution. If the trends observed here are robust when the complete sample is collected, they will suggest a difference in feedback processing versus feedforward updating in children with and without RSSD. Additionally, SimpleDIVA modelling [19] is underway for this dataset, which will provide further insights into speech-motor control in this population.

6. REFERENCES

[1] J. McCormack, S. McLeod, L. McAllister, and L. J. Harrison. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. Int. J. Speech Lang. Pathol., 11(2): 155–170.

[2] Flipsen, P. (2015). Emergence and prevalence of persistent and residual speech errors. Semin. Speech Lang. 36(4):217–223.

[3] Cabbage, K., & Hitchcock, E. (2022). Clinical Considerations for Speech Perception in School-Age Children With Speech Sound Disorders: A Review of the Current Literature. Language, Speech, and Hearing Services in Schools 53(3):768-785.

[4] Hearnshaw, S., Baker, E., & Munro, N. (2019). Speech perception skills of children with speech sound disorders: A systematic review and meta-analysis. Journal of Speech, Language, and Hearing Research, 62(10): 3771-3789.

[5] Cialdella, L., Kabakoff, H., Preston, J. L., Dugan, S., Spencer, C., Boyce, S., Tiede, M., Whalen, D.H., & McAllister, T. (2020). Auditory-perceptual acuity in rhotic misarticulation: baseline characteristics and treatment response. Clinical Linguistics & Phonetics, 35(1): 19-42.

[6] H. Terband, F. van Brenk, and A. van Doornik-van der Zee. (2014).Auditory feedback perturbation in children with developmental speech sound disorders," J. Commun. Disord. 51:64–77.

[7] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization," J. Acoust. Soc. Am. 128(4):2033–2048.

[8] V. M. Villacorta, J. S. Perkell, and F. H. Guenther. (2007).Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J Acoust Soc Am. 122(4): 2306–19. doi: 10.1121/1.2773966.

[9] Becker, Michael and Jonathan Levine. (2013). Experigen – an online experiment platform. Available at http://becker.phonologist.org/experigen.

[10] H. Kawahara, M. Morise, H. Banno, and V. G. Skuk. (2013).Temporally variable multi-aspect N-way morphing based on interference-free speech representations: 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013. doi: 10.1109/APSIPA.2013.6694355.

[11] A. Klaus, D. R. Lametti, D. M. Shiller, and T. McAllister. (2019). Can perceptual training alter the effect of visual biofeedback in speech-motor learning?. J. Acoust. Soc. Am., 145(2):805–817. doi: 10.1121/1.5089218.

[12] Cheng, H-S., Niziolek, C. A., Buchwald, A., & McAllister, T. (2021). Examining the relationship between speech perception, production distinctness, and production variability. Frontiers in Human Neuroscience, 15: 660948.

[13] Daliri, A., Wieland, E. A., Cai, S., Guenther, F. H., & Chang, S. E. (2018). Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter. Developmental science, 21(2), e12521.

[14] Kim, K.S., Wang, H., and Max, L. (2020). It's about time: minimizing hardware and software latencies in speech research with real-time auditory feedback. Journal of Speech Language and Hearing Research, 63 (8): 2522-2534.

[15] Cheung, S. (2020). *Methodological Considerations and Findings of a Vowel Perturbation Study in Typically Developing Children* [Unpublished doctoral dissertation]. University of Toronto.

[16] Nieto-Castanon, A. (2022). GuentherLab FL Voice repository on GitHub. Retrieved from: https://github.com/GuentherLab/FLvoice.git

[17] Rvachew, S., & Jamieson, D. G. (1989). Perception of voiceless fricatives by children with a functional articulation disorder. Journal of Speech and Hearing Disorders, 54(2), 193-208.

[18] Maas, E., Mailend, M. L., & Guenther, F. H. (2015). Feedforward and feedback control in apraxia of speech: Effects of noise masking on vowel production. Journal of Speech, Language, and Hearing Research, 58(2), 185-200.

[19] Kearney, E., Nieto-Castañón, A., Weerathunge, H. R., Falsini, R., Daliri, A., Abur, D., Ballard, K. J., Chang, S.-E., Chao, S.-C., Murray, E. S. H., Scott, T. L., & Guenther, F. H. (2020). A simple 3-parameter model for examining adaptation in speech and voice production. Frontiers in Psychology, 10, 2995. https://doi.org/10.3389/fpsyg.2019.02995

USING COMPUTATIONAL MODELS TO CHARACTERIZE THE ROLE OF MOTOR NOISE IN SPEECH: THE CASE OF AMYOTROPHIC LATERAL SCLEROSIS

Benjamin Parrell[1], Antje Mefferd[2], Sarah Harper[3], Simon Roessig[4], Doris Mücke[5]

[1]Communication Sciences and Disorders, University of Wisconsin-Madison,
[2]Hearing and Speech Sciences, Vanderbilt University Medical Center, [3]Chang Lab, UC San Francisco,
[4]Department of Linguistics, Cornell University, [5]IfL-Phonetics, University of Cologne
bparrell@wisc.edu, doris.muecke@uni-koeln.de

ABSTRACT

Although the speech impairments in neurogenic speech disorders have been characterized using various approaches (kinematic, acoustic, auditory-perceptual), the underlying mechanisms that contribute to the observed symptoms remain poorly understood. A more direct assay of speech motor control is needed to test hypotheses about the underlying cause of the observed changes in speech patterns. Computational models of speech production provide one possibility, as they allow direct manipulation of control parameters.

Here, we present a proof-of-concept of how the Task Dynamic model of speech production can be used to test potential control impairments using an analysis-by-synthesis approach. Specifically, we test how changes in motor noise in speakers with amyotrophic lateral sclerosis may account for altered patterns of spatial variability in this population. This work establishes the potential of exploring how noise and bias in motor control may help uncover the mechanisms underlying speech impairments in this and other disorders.

Keywords: variability, ALS, motor noise, motor control, kinematics

1. INTRODUCTION

One of the major challenges in understanding neurogenic speech disorders has been establishing the precise nature of the underlying motor impairments. Typically, these disorders are characterized by auditory-perceptual speech characteristics, such as the Mayo clinic dysarthria classification system [1]. However, many perceptual speech features overlap across speakers with different underlying etiologies and dysarthria types (e.g., “consonant imprecision”). To better understand these disorders, a growing body of research has focused on the characterization of motor impairment patterns by direct examination of speech acoustics or kinematics rather than listener impression. While this work has promising potential to provide a more precise description of the motor speech impairments in these speakers, acoustic and kinematic analyses can ultimately only address the observable symptoms stemming from the neurologic damage to the speech production system. Alternative approaches are necessary to reveal the pathomechanisms that underlie those symptoms.

One potential way to move beyond description of symptoms to a full characterization of the control impairments in neurogenic disorders is through the use of computational models of speech production. For example, selective impairments in either the feedforward or feedback pathways of the DIVA model have been used to replicate speech impairments seen in stuttering [2] and apraxia of speech [3]. Generally, however, existing simulations replicate gross structural changes in speech (e.g. phoneme repetition or substitution). Here, we present a proof-of-concept for a new approach. Using the Task Dynamics model of speech production [4]–[6] , we modify parameters regulating speech motor control and execution to assess their ability to replicate established gradient patterns of kinematic impairment in a specific speech motor disorder—amyotrophic lateral sclerosis (ALS). Critically, this analysis-by-synthesis approach allows us to test specific hypotheses about control deficits in particular speech disorders rather than relying only on inferences drawn from kinematic patterns.

2. SPEECH IMPAIRMENTS IN AMYOTROPHIC LATERAL SCLEROSIS

ALS is a neurodegenerative disease that results in damage to motor neurons, the nerve cells that convey information from the brain to the muscles. ALS can affect both upper and lower motor neurons, with both spinal and bulbar pathways being affected eventually. This damage results in muscle weakness and loss of motor control broadly, typically resulting in spastic-flaccid dysarthria [1], [7]. Because the impairment in motor control in ALS – impairments in descending pathways from cortex to the orofacial musculature — is relatively simple compared to disorders that involve subcortical structures or cortical damage outside motor neuron projections to the periphery, it

provides a good test case for assessing neurogenic control impairments in speech through the use of computational models.

Although the general neurologic impairment in ALS is well-established, precisely how degeneration of motor neurons affects speech motor control is still unclear. Looking at the broader limb control literature, one intriguing possibility is that damage to motor neurons may results in increased motor variability even when there is no major deficit in muscular strength. Motor noise—variability in muscular contraction—is typically thought to be signal-dependent, such that generating larger forces also results in more variable forces. While this relationship generally holds, there is evidence from limb control that noise depends not only on total force but also on muscular strength [8]. For a given force output, stronger muscles (which recruit more motor neurons firing at lower rates) produce *less* variability than weaker muscles (where fewer motor neurons must fire faster to produce equivalent force output). It follows, then, that loss of motor neurons in ALS would result not only in weakness but also increased variability when producing equivalent forces to healthy controls. Indeed, this seems to be the case when generating isometric force with the fingers, at least when lower motor neurons are affected [9]. Although such direct evidence for this hypothesized relationship is not available in speech, consistent with this idea, the severity of speech symptoms in individuals with ALS is predicted by the variability in kinematic acceleration of the speech articulators [10]. Together, these results suggest that increased motor noise (as measured through the acceleration signal in speech kinematics) may be the driving force behind the high degree of variability in this population.

Given this hypothesis, it would be expected that speech kinematic variability would be higher in ALS patients than in neurotypical controls. There is limited evidence that this may indeed be the case, at least for some speakers [11], [12]. Although speakers with mild ALS produced slightly *less* variable movement patterns than neurotypical controls when speaking at a comfortable rate, their habitual speech rates were slower than those of control speakers. When speakers with ALS were prompted to speak as fast as possible, their speech rates were comparable to the control speakers' habitual rate. However, their speech movement patterns during the fast rate were significantly more variable than those produced by controls during their habitual rate. This finding suggests that speakers with ALS may produce speech at an abnormally slow, habitual rate as a strategy to maximize articulatory control. Indeed, the slowing-down of the speech rate has been frequently discussed as a compensation mechanism to decrease spatial variability across various motor speech impairments as well as aging [13].

However, previous findings on movement pattern variability must be taken with some caution as 1) they used a combined metric of both spatial and temporal variability, the *Spatiotemporal Index* (STI), rather than focusing on the purely spatial effects which could be directly attributed to changes in motor unit recruitment without other changes in control, 2) the sample sizes and number of repetitions from each speaker were relatively small given the heterogenous and idiosyncratic progression of ALS, 3) the increase in STI with speech rate in ALS speakers has not been consistently found [12].

3. MODELLING CONTROL IMPAIRMENTS IN ALS

In the Task Dynamics model [4], the functional unit of control is the speech gesture—movements to create linguistically relevant constrictions of specific sizes at specific locations in the vocal tract. Each gesture is modelled as a point attractor with 2nd order dynamics (essentially, a mass-spring system) that controls the movements of the speech articulators. This model relies on a hierarchical feedback control architecture, where, at each time step, the current positions and velocities of the speech articulators (mobility state) are first used to estimate the size of constrictions (task state). The current task state is then compared against the desired task state (gestural goal) to generate a task-state error, which is fed into the mass-spring dynamic equations to generate a desired change in task state (represented as a velocity given the 2nd order dynamics). This desired task-state velocity is then converted to a "motor command" (mobility-space acceleration) to move the speech articulators in the system.

Signal-dependent motor noise is not currently implemented in the computational implementation of Task Dynamics, Task Dynamics Application (TADA, [5]). However, it is a relatively straightforward modification to add noise to the mobility-space acceleration signal—the output of the control algorithm that is the final movement command to the plant. Essentially, the acceleration signal at each time step is multiplied by a weighting term drawn from a normal distribution, $N(0,1)$:

$$(1) \quad \ddot{u} = J^{-1}(u)\left[\ddot{x} - \dot{J}(u,\dot{u})(u)\right][1 + \alpha N(0,1)]$$

where α is the weighting term that determines the amount of signal dependent noise, $\ddot{u}$ is the mobility-space acceleration or motor command, u is the current mobility-space position, $\dot{u}$ the current mobility space velocity, J is the Jacobean mapping

changes in mobility space to changes in task space, J^{-1} is the pseudo-inverse of this Jacobean, and $\dot{J}$ is its derivative. When α is set to 0, this equation yields the same output as in the standard model.

Our hypothesis is that increased spatial kinematic variability in ALS compared to neurologically healthy control when speaking at the same absolute rate is due to increased signal-dependent noise subsequent to motor neuron generation (due to compensation mechanisms in impaired speech, the fast speech rate of ALS speakers is comparable to the habitual rate of controls). We tested this by evaluating how changes in the weighting term α in Eq. 1 affect the variability of the kinematic output of the TADA model. We varied α from 0 to 0.4 in steps of 0.05. In order to account for speech rate changes, we multiplied the frequency of the planning oscillators which determine the time periods of gestural activation by r = [0.7 0.85 1.15 1.3], where setting r to 1 is equal to the default speech rate of the model, values of r greater than 1 speed up the speech rate, and values of r smaller than 1 slow the speech rate. Our preliminary modelling suggested that modifying this parameter alone does not substantially affect the movement kinematics, as the dynamics of the gestures themselves are still stable across speech rates. It has been observed that the stiffness of speech movements (estimated via the ratio of peak velocity relative to displacement) consistently increases as movement durations become shorter, despite variability across individuals in whether increased speech rate is accomplished though faster movements, smaller movements, or both [14]. To replicate this in TADA, the stiffness parameter of each gesture was multiplied by r, such that fast speech rates resulted in both faster and shorter movements. This change was sufficient to replicate the basic pattern observed in human data, where shorter movements with equivalent amplitudes as longer movements have higher peak velocities.

For each pair of parameters r and α, we generated 20 simulations of the phrase "Buy Bobby a puppy", which has been commonly used to measure STI in both healthy and disordered speech [11], [15], [16]. To measure spatial variability, we analysed the resulting trajectories of the lower lip vertical position parameter in the CASY plant model used in TADA [17], as it gives a good comparison to human data, where the vertical position of the lower lip is often used. Other parameters of the CASY model (jaw angle, tongue body angle and length from the temporomandibular joint) are less directly comparable to human kinematics. No prosodic modulation was applied, resulting in output where each syllable in the phrase as equal duration.

Because we are particularly interested in the effects of signal-dependent motor noise on spatial variability, the frequently-used STI metric is inappropriate given its confounding of both spatial and temporal variability. Instead, we used a metric of the residual spatial variability after temporal variability has been accounted for through non-linear time warping using functional data analysis [18]–[20]. Briefly, this method first linearly normalizes articulatory trajectories to the span [0,1] after each trajectory has been converted from raw positions to z-scores. Trajectories are then transformed into a set of basis functions and are non-linearly aligned to a common reference trajectory, here the barycenter average of all trajectories [21]. The average variability of the aligned trajectories can then be taken as a measure of purely spatial variability, the *Index of Amplitude Variability* (IAV, [18]). Note that the time-warping functions can be used separately to generate a measure of temporal variability, the *Index of Phase Variability* (IPV).

4. RESULTS

To first validate the IAV metric, tongue body trajectories from one control speaker and one speaker with ALS with moderate dysarthria severity were used to calculate both STI and IAV (Figure 1). Each speaker produced 7 (control) or 9 (ALS) repetitions of the phrase "Buy Kaia a kite". The vertical position of the tongue body was tracked using electromagnetic articulography (EMA), trimmed from the time of maximum vertical position of the lower lip during the initial [b] in "buy" to time of the maximum vertical position of the tongue tip during the [t] in "kite". The tongue body was chosen because the signal from the jaw sensor had relatively little movement in the control participant. Results show that the speaker with ALS has a higher STI value than the control speaker, indicating a higher degree of variability. FDA analyses suggest that IAV (and IPV) is higher as well.

Simulated lower lip trajectories for one level of motor noise (α = 0.2) at the default speech rate (r = 1) are plotted in Figure 2 (top panel). Figure 2 additionally shows results from all simulations for STI (middle panel) and IAV (bottom panel). To replicate the relatively low number of kinematic trajectories typically used in studies with patients (typically 5-10), for each pair of α and r values, 8 out of the 20 trajectories were randomly drawn without replacement and used to calculate STI and IAV. This procedure was repeated 20 times for each pair of values to generate a mean and standard deviation.

For both STI and IAV, variability increases with both motor noise and speech rate. These simulations demonstrate that there is a trade-off between speech

rate and the amount of signal dependent noise on the resulting spatial variability. Critically, the increase in spatial variability that would be expected due to higher levels of signal-dependent noise can be countered by reducing the speech rate, which results in slower movements and so, less motor noise. For example, the IAV at a speech rate of 1.15 is 0.095 when α = 0.15 and 0.125 when α = 0.3. However, slowing the rate down to 0.85 results in an IAV of 0.066 when α = 0.3, less than the IAV for α = 0.15 at the original rate. This pattern qualitatively replicates the data in [11], where the ALS patients had lower variability and slower rates than controls, but higher variability than controls when pushed to speak at similar rates.

5. CONCLUSIONS

Our simulations have shown that increased motor noise due to muscular weakness in ALS may provide an explanation for what has been a curious pattern: at least in one study, speakers with ALS show reduced variability compared to controls at their (abnormally slow) habitual rate, but increased variability when speaking at the same rate as controls. However, this proof-of-concept relies on several as yet untested assumptions. First, existing studies report only STI, which conflates spatial and temporal variability. More detailed analyses using metrics such as IAV and IPV are needed to separate spatial and temporal changes in motor speech disorders. Second, increased variability at a fast speech rate in ALS has not been consistently found. Studies with larger samples, and more trials per participant are needed (especially given the high variability found in our simulations using 8 trajectories). Finally, our simulations of rate assume that individuals with ALS show the same increase in stiffness as duration decreases as controls—this assumption should be examined using speech kinematic data from this population. In sum, our results here suggest that analysis-by-synthesis is a promising method for testing hypotheses about the control deficits underlying observable impairments in motor speech disorders.

6. ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) in the collaborative research center SFB1252 Prominence in Language (Project-ID 281511265) and project RO 6767/1-1 (Walter Benjamin program) as well as NIH NIDCD grants R01 DC017091 and R01DC019648.

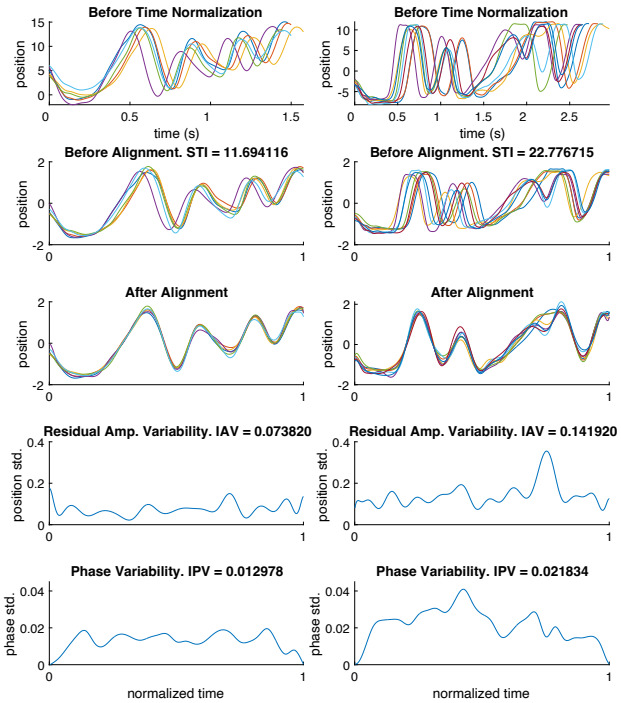


Figure 1: Variability in tongue body movement from a control speaker (left) and a speaker with ALS (right). Top row: original tongue body trajectories, measured with EMA. Second row: z-score and linearly time-alignment trajectories used both as first step of FDA process and to calculate the STI. Third row: non-linearly aligned trajectories. Fourth row: residual spatial variability after time-alignment, used to calculate the IAV. Bottom row: variability of time-warping functions generated in non-linear time alignment process, used to calculate the IPV.

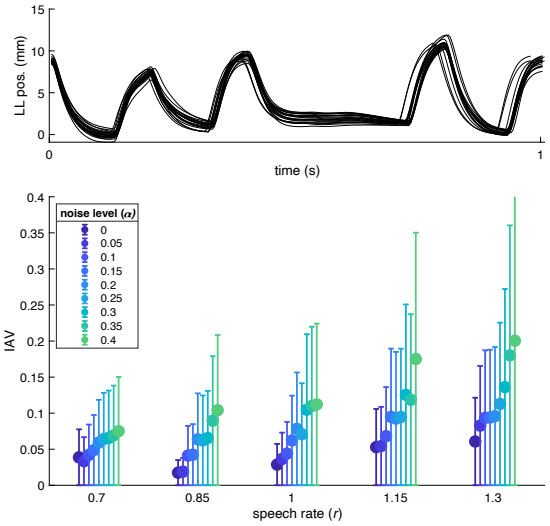


Figure 2: Top panel: Example simulated kinematic trajectories showing the vertical position of the lower lip during 20 repetitions of the phrase “Buy bobby a puppy”, trimmed from the time of maximum vertical position of the lower lip during the [b] in “Buy” to the time of maximum lower lip position during the second [p] in “puppy” (α = 0.1, r = 1). Bottom panel: Mean and standard deviation of IAV across noise levels (colors) and speech rates (on x axis). IAV generally increases with both motor noise and speech rate. However, note the large variability in IAV measures with only 8 repetitions.

7. REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor speech disorders*. Philadelphia: Saunders, 1975.

[2] O. Civier, D. Bullock, L. Max, and F. H. Guenther, "Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation," *Brain Lang.*, vol. 126, no. 3, pp. 263–278, Sep. 2013, doi: 10.1016/j.bandl.2013.05.016.

[3] H. E. Miller and F. H. Guenther, "Modelling speech motor programming and apraxia of speech in the DIVA/GODIVA neurocomputational framework," *Aphasiology*, vol. 0, no. 0, pp. 1–18, May 2020, doi: 10.1080/02687038.2020.1765307.

[4] E. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 333–382, 1989.

[5] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *J. Acoust. Soc. Am.*, vol. 115, p. 2430, 2004.

[6] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: a coupled oscillator model," in *Approaches to phonological complexity*, F. Pellegrino, E. Marisco, I. Chitoran, and C. Coupé, Eds. Berlin/New York: Mouton de Gruyter, 2009, pp. 299–328.

[7] J. R. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*, Third edition. St. Louis, Missouri: Elsevier, 2013.

[8] A. F. de C. Hamilton, K. E. Jones, and D. M. Wolpert, "The scaling of motor noise with muscle strength and motor unit number in humans," *Exp. Brain Res.*, vol. 157, no. 4, pp. 417–430, Aug. 2004, doi: 10.1007/s00221-004-1856-7.

[9] P. K. Kasi *et al.*, "Motor unit firing characteristics in patients with amyotrophic lateral sclerosis," in *2009 IEEE 35th Annual Northeast Bioengineering Conference*, Apr. 2009, pp. 1–2. doi: 10.1109/NEBC.2009.4967680.

[10] K. J. Teplansky *et al.*, "Tongue and Lip Acceleration as a Measure of Speech Decline in Amyotrophic Lateral Sclerosis," *Folia Phoniatr. Logop.*, Jun. 2022, doi: 10.1159/000525514.

[11] A. S. Mefferd, G. L. Pattee, and J. R. Green, "Speaking rate effects on articulatory pattern consistency in talkers with mild ALS," *Clin. Linguist. Phon.*, vol. 28, no. 11, pp. 799–811, Nov. 2014, doi: 10.3109/02699206.2014.908239.

[12] M. Kuruvilla-Dugdale and A. Mefferd, "Spatiotemporal movement variability in ALS: Speaking rate effects on tongue, lower lip, and jaw motor control," *J. Commun. Disord.*, vol. 67, pp. 22–34, May 2017, doi: 10.1016/j.jcomdis.2017.05.002.

[13] T. Thies, A. Hermes, and D. Mücke, "Compensation in Time and Space: Prominence Marking in Aging and Disease," *Languages*, vol. 7, no. 1, p. 21, Jan. 2022, doi: 10.3390/languages7010021.

[14] D. Ostry and K. Munhall, "Control of rate and duration of speech movements," *J. Acoust. Soc. Am.*, 1985.

[15] A. Smith and L. Goffman, "Stability and patterning of speech movement sequences in children and adults," *J. Speech Lang. Hear. Res.*, vol. 41, pp. 18–30, 1998.

[16] A. Smith, L. Goffman, HowardN. Zelaznik, G. Ying, and C. McGillem, "Spatiotemporal stability and patterning of speech movement sequences," *Exp. Brain Res.*, vol. 104, no. 3, Jun. 1995, doi: 10.1007/BF00231983.

[17] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. T. Tiede, and C. P. Browman, "CASY and extensions to the task-dynamic model," in *Proceedings of the 4th Speech Production Seminar*, 1996.

[18] J. C. Lucero, "Comparison of measures of variability of speech movement trajectories using synthetic records.," *J Speech Lang Hear Res*, vol. 48, no. 2, pp. 336–44, 2005, doi: 10.1044/1092-4388(2005/023).

[19] J. Lucero, G. K. Munhall, V. L. Gracco, and O. J. Ramsay, "On the registration of time and the patterning of speech movements," *J. Speech Lang. Hear. Res.*, vol. 40, pp. 1111–1117, 1997.

[20] O. J. Ramsay, G. K. Munhall, V. L. Gracco, and D. Ostry, "Functional data analyses of lip motion," *J. Acoust. Soc. Am.*, vol. 99, pp. 3718–3727, 1996.

[21] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognit.*, vol. 44, no. 3, pp. 678–693, Mar. 2011, doi: 10.1016/j.patcog.2010.09.013.

A REVIEW OF PALATOGRAPHIC MEASUREMENT DEVICES DEVELOPED AT THE TU DRESDEN FROM 2011 TO 2022

Peter Birkholz, Simon Stone, Christoph Wagner, Steffen Kürbis, Alexander Wilbrandt, Mats Bosshammer

Institute of Acoustics and Speech Communication, TU Dresden, Germany
peter.birkholz@tu-dresden.de

ABSTRACT

Electropalatography is an established measurement technique in speech research and therapy. While it captures the time-varying pattern of contact between the tongue and the hard palate in great detail, it provides no information about the tongue position at places where the tongue does not touch the palate. Over the past 12 years, the authors of this paper developed several artificial palate prototypes that include optical distance sensors to capture additional information about the position of the tongue. Here, six of these prototypes are reviewed. At the end, a palate design is proposed that takes into account all the lessons learned from the developmental process.

Keywords: Optical palatography

1. INTRODUCTION

Artificial palates are an established tool to measure various aspects of lingual articulation, e.g., the tongue-palate contact pattern with electropalatography [1, 2, 3] (EPG), the lingual pressure against the hard palate with pressure-palatography (PPG) [4, 5], or the distance between tongue and palate with optical palatography [6, 7, 8] (OPG). These measurement methods are used in phonetic research [9, 10], speech therapy [11, 12], and silent speech interfaces [13, 14], among others.

Due to its commercial availability, EPG is the most widely used palate-based measurement technique. EPG palates have 62 or more electrodes distributed over their surface, each of which registers the presence or absence of contact with the tongue. However, these contact sensors neither provide information about the position of the tongue at places without contact, nor about the lingual pressure at the points of contact. This information is provided by OPG and PPG. OPG uses optical distance sensors to measure the distance from different locations on the palate to the tongue surface, and PPG uses pressure sensors to measure the lingual pressure against the palate. Optical and pressure sensors are both usually larger than EPG electrodes, which limits the number of sensors per palate. Unlike EPG, neither OPG nor PPG have progressed beyond prototype status.

Given the complementary information provided by the three measurement methods, the combination of those on the same pseudo-palate would be useful. This paper reviews the authors' efforts over the past 12 years to achieve this goal. One strategy was to combine a limited number of (relatively large) optical distance sensors in the midsagittal plane with a larger number of surrounding contact sensors. Ever-smaller optical sensors have enabled another strategy, which was to replace contact sensors entirely with optical sensors. Dedicated pressure sensors were not used. Instead, we explored the potential of the optical sensors to measure lingual pressure, which was first proposed in [6]. Since the optical sensors are in the focus of this review and part of all palate models, the following subsection gives an overview of their principle of operation.

1.1. Optical distance sensing

Each distance sensor consists of a light source and a photodetector at a close distance from each other, mounted on the lingual side of the artificial palate. To measure the distance to the tongue, the light source emits a beam of light on the tongue, and the photodetector registers the light reflected from the tongue surface. The recorded light intensity is approximately proportional to the inverse square of the distance [6]. The light source is typically a light-emitting diode (LED) or a laser diode, and the photodetector is a photodiode or a phototransistor. To minimize ambient light interference, these components are usually selected to work with infrared light.

Figure 1a shows an example circuit for the measurement using an LED and a phototransistor. The LED is driven by a *precision* current source (which is essential to avoid measurement drift) with the current $I_{\mathrm{LED}} = V_{\mathrm{CC}}/(2R_1)$ when $V_{\mathrm{in}} = 0$ V (on state), and $I_{\mathrm{LED}} = 0$ A, when $V_{\mathrm{in}} = V_{\mathrm{CC}}$ (off state). At

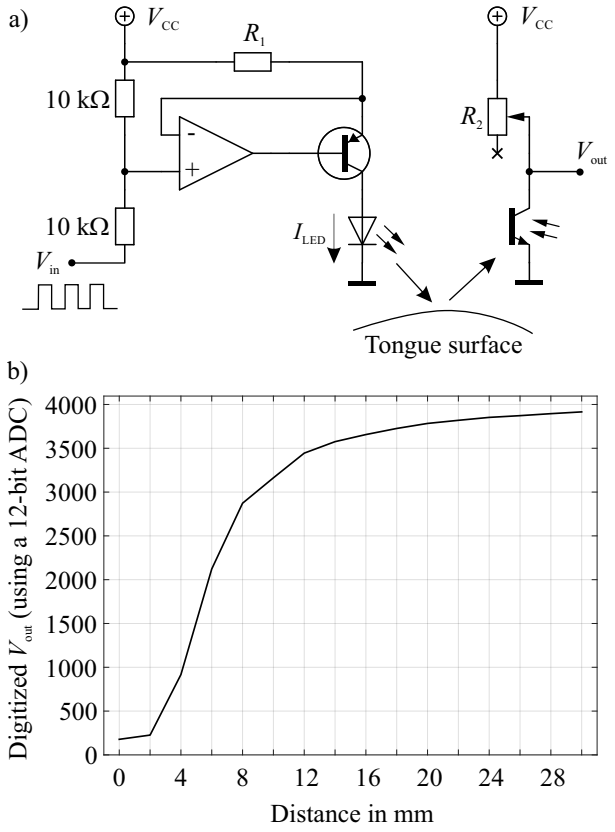


Figure 1: a) Circuit for optical distance sensing. b) Distance-sensing function.

the receiver side, the resistor R_2 converts the current through the phototransistor (which depends on the incident light) into an output voltage V_{out}. Figure 1b shows an exemplary relation between V_{out} and the sensor-tongue distance. This *distance sensing* function was obtained by positioning the tongue at a range of well-defined distances to the sensor (as in [15]) and recording the corresponding values of V_{out}. It is a monotonic function and thus allows to infer the distance from V_{out}. The exact shape of the function depends on R_2, the characteristics of the light source and the photodetector, their spatial distance, and the reflective properties of the tongue, which differ from person to person. This makes some form of calibration of the sensors necessary.

2. PALATOGRAPHY DEVICES

Figure 2 shows six relevant palate models that have been developed at the TU Dresden between 2011 and 2022. They will be discussed in the following.

2.1. Model 1

Model 1 [16] was the first palate that combined contact sensors with distance sensors. Its design is based on the Articulate EPG palate [3], in which the 62 contact sensors (arranged on flexible strips) are embedded between two transparent sheets that are thermoformed over a plaster model of the hard palate. For the optical sensors, we chose units of the type TCND5000 (Vishay Semiconductors) that consist of an LED and a photodiode separated by an optical barrier. Like the contact sensors, the distance sensors were embedded between the thermoformed sheets. Since the sheet covering the optical components reduced the sensing range, the sheet parts around the sensors were cut away. In addition to three distance sensors directed at the tongue, one sensor at the upper incisors for detection of lip movement and one backward-facing sensor for detection of velum movement was included.

The evaluation of this palate showed that the EPG sensors worked as expected, the lingual distance sensors could capture the height of the tongue, and the lip sensor was able to detect lip movements. However, the velum-directed sensor was *not* able to detect differences in velum position, the height of the distance sensors (> 3 mm) partly disturbed the articulation, and the manufacturing time for the palate was quite long.

2.2. Model 2

To have less impact on articulation and make fabrication more effective, Model 2 was made with tailored optical sensors mounted on a strip of flexible printed circuit board (PCB) [15]. This design had no contact sensors. For the distance sensors, different combinations of *separate* LEDs and phototransistors were evaluated. The combination of the LED VSMY2850G and the phototransistor TEMT7100X01 (both by Vishay Semiconductors) with a gap of 0.6 mm between them was found to be the best tradeoff for small size and sufficient measuring range. Important for the good measuring range was the *lens* of the LED to focus the light beam. Four sensors were distributed along the palatal midline, and one sensor at the incisors was directed towards the upper lip. After soldering the optical components to the flexible PCB, their electrical contacts were sealed with modeling resin, and each sensor was individually calibrated as described in [15]. The PCB was then glued to a 0.5 mm thermoformed sheet. Behind the incisors, the sheet around the teeth was cut away. Instead, Adams clasps were used for fixation, which increased wearing comfort. This type of palate, driven by the circuit in Figure 1a, provided reliable measurements of the tongue position and was used for the real-time control of a 2D animated model of

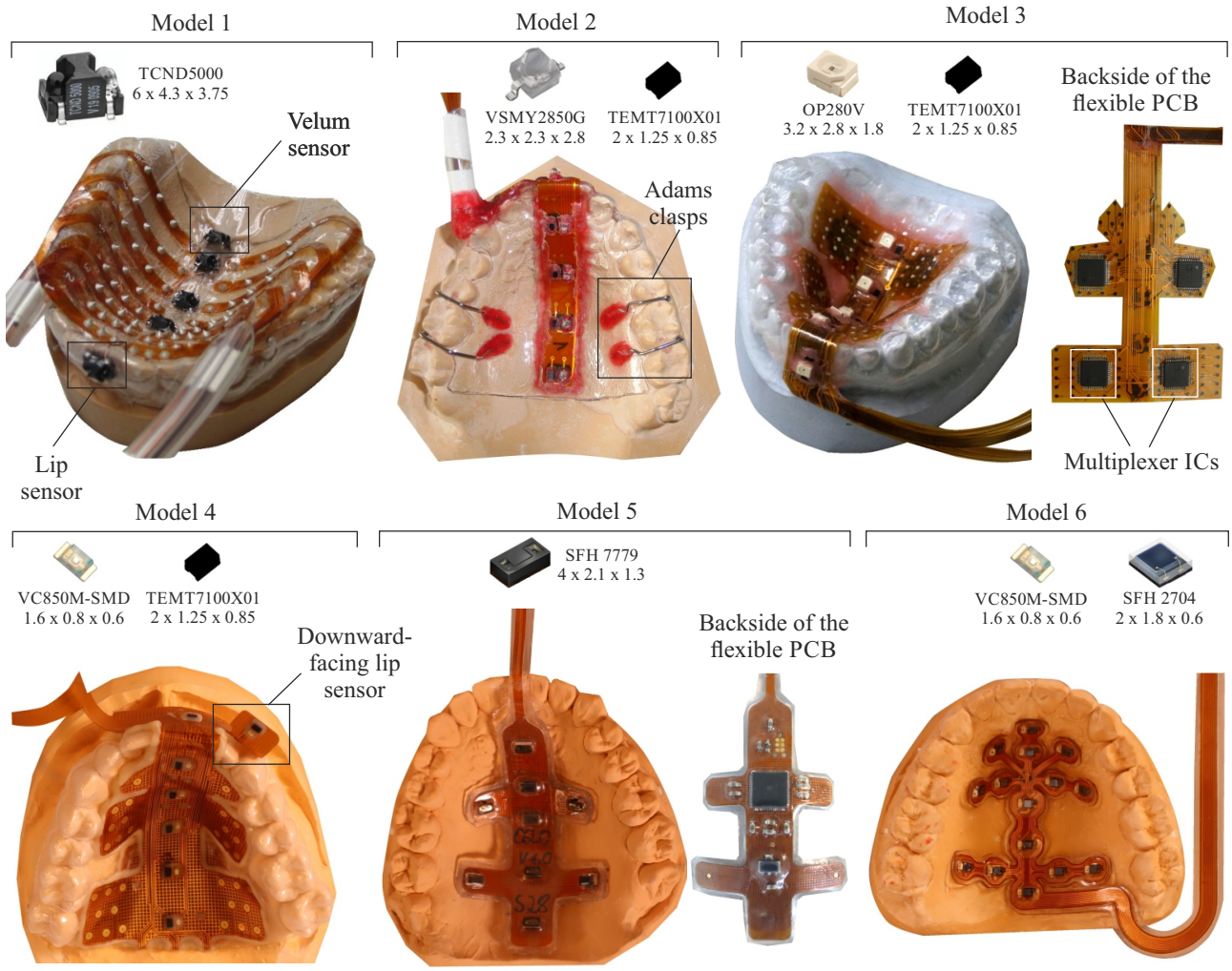


Figure 2: Palate models. For each model, the type and size ($L \times W \times H$ in mm) of the optical components is shown.

the vocal tract [17].

2.3. Model 3

Model 3 is a further development of Model 2 and contains five distance sensors for the tongue, one distance sensors for the lips, and 124 contact sensors [18, 19, 20]. The light source for the distance sensors was changed to the laser diode OP280V (OPTEK Technology Inc.), which has a slightly larger footprint, but a smaller height (which increased wearing comfort) than the LED of Model 2. Compared to the LED, the laser diode generates a narrow light beam without a "bulky" lens and requires substantially less electrical power. The contact sensors were gold-plated electrodes placed on the same PCB as the distance sensors. To reduce the number of wires required to connect the palate with the external control unit, four 32-channel analog multiplexer integrated circuits (IC) at the back side of the PCB were used to multiplex the contact sensors. A denture relining material (GC Reline) was used to glue the PCB to the carrier sheet and to seal the electrical contacts of the optical sensors.

Experiments with this palate model have shown that the combination of distance and contact sensors on the same PCB generally works well. However, some of the units produced broke after a few uses because the solder joints of the multiplexer ICs could break when the palates were bent. With this palate model, a simple in-vivo calibration method of the distance sensors was introduced [18], as well as a method to reduce measurement errors when the tongue surface was not perpendicular to the optical axes of the sensors [20].

2.4. Model 4

Model 4 is a revised version of Model 3 where the laser diode has been substituted with a smaller type (VC850M-SMD by Roithner Lasertechnik GmbH), the number of contact sensors was reduced to 32, and a 2nd optical lip sensor was included facing downwards to the lower lip. With the two lip sensors together it became possible to discriminate lip protrusion from lip opening movements. The lower number of contact sensors reduced the risk of saliva bridges between the electrodes and required only a single 32-channel multiplexer IC on the backside of the PCB, which greatly reduced the risk of defects due to broken solder joints. More than 15 units of this palate model were manufactured (partly still with the previous laser diode) and used, e.g., for silent speech command word recognition [14] and the articulatory analysis of diphthongs [10].

2.5. Model 5

Model 5 contains optical distance sensors only, both along the midsagittal contour and on the sides of the palate, and does not include lip sensors [21]. The sensors are integrated digital sensors (SFH 7779 by Osram) that are equipped with a miniature LED and photodetector and output a 16-bit value that varies with distance. This type was selected based on its ability to measure not only the distance but also the force exerted by the tongue on contact [22]. The circuitry to control the sensors was placed on the back of the flexible PCB. This included an IC to multiplex the I2C bus that is used to communicate with the sensors, because the I2C bus address of each sensor was identical and unchangeable.

The PCB was completely laminated between two flexible transparent 1 mm sheets (Erkoflex by Erkodent). In contrast to Model 1, the additional sheet over the distance sensors hardly affected the distance measurement. Compared to Models 1–4, this model was not anchored to the teeth with Adams clasps or with a carrier sheet formed around the teeth, but with a specially developed adhesive film [23]. The advantage is that no plaster model of the user's hard palate is required to fabricate this model.

Experiments with this model showed that the integrated distance sensors enabled robust measurements, and that the adhesive film was effective in fixing the pseudopalate to the palate for at least 15 min. However, the mechanical stress of attaching the unit to the hard palate sometimes led to (electrical) failure. In addition, the relatively thick thermoforming sheets reduced its flexibility and snugness to the real palate.

2.6. Model 6

Model 6 is a further development of Model 5 where the integrated digital distance sensors have been replaced with sensors made of a separate light source and photodetector, and where all other electronic components have been moved from the palate to the control circuit outside the mouth. The step back to discrete components for the distance sensors was taken because none of the currently available integrated digital sensors fulfills all requirements for optical palatography (easy control of multiple identical sensors, small size, sufficient measurement range and speed, etc.). Each distance sensor consisted of the proven laser diode VC850M-SMD, and a photodiode of type SFH2704 (Osram) with a particularly large active chip area. In practice, however, the photodiode did not provide any advantage over the previously used phototransistor TEMT7100X01. To obtain a sufficient measurement range using these discrete components completely covered under a 0.5 mm thick transparent sheet, small optical barriers had to be installed between the light sources and photodetectors to prevent optical crosstalk. By eliminating the electronic components on the back of the PCB, the pseudopalate was robust, extremely flexible, and conformed very well to the shape of the real palate when glued into the mouth.

3. OUTLOOK

For future palate designs, Model 6 appears to be the best starting point. By using only optical sensors, the entire flexible PCB can be laminated between two sheets, eliminating the need for the time-consuming sealing of optical sensors in combined electro-optical models. The photodiode SFH2704 can be replaced with a phototransistor (TEMT7100X01 or an even smaller type), which would reduce the footprint of the sensors and simplify the control circuitry. Two lip sensors as in Model 4 can easily be added to this design. For more frequent applications, a subject-specific version of Model 6 can be made where the thermoformed sheets extend around the teeth of the user for fixation.

ACKNOWLEDGEMENTS

This research was supported by the DFG (grants BI 1639/1-1, BI 1639/1-2) and by the BMBF (grants 01IS19019B and 16SV7741).

4. REFERENCES

[1] W. Hardcastle, W. Jones, C. Knight, A. Trudgeon, and G. Calder, "New developments in electropalatography: A state-of-the-art report," *Clinical Linguistics and Phonetics*, vol. 3, no. 1, pp. 1–38, 1989.

[2] S. Kelly, A. Main, G. Manley, and C. McLean, "Electropalatography and the Linguagraph system," *Medical Engineering & Physics*, vol. 22, no. 1, pp. 47–58, 2000.

[3] A. A. Wrench, "Advances in EPG palate design," *Advances in Speech-Language Pathology*, vol. 9, no. 1, pp. 3–12, 2007.

[4] R. E. McGlone, W. R. Proffit, and R. L. Christiansen, "Lingual pressures associated with alveolar consonants," *Journal of Speech and Hearing Research*, vol. 10, no. 3, pp. 606–615, 1967.

[5] M. Tiede, J. Perkell, M. Zandiour, M. Matthies, and E. Stockmann, "A new approach to pressure-sensitive palatography using a capacitive sensing device," in *Proc. of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 3149–3152.

[6] C.-K. Chuang and W. S. Wang, "Use of optical distance sensing to track tongue motion," *Journal of Speech and Hearing Research*, vol. 21, pp. 482–496, 1978.

[7] S. G. Fletcher, M. J. McCutcheon, S. C. Smith, and W. H. Smith, "Glossometric measurements in vowel production and modification," *Clinical Linguistics and Phonetics*, vol. 3, no. 4, pp. 359–375, 1989.

[8] A. A. Wrench, A. D. McIntosh, and W. J. Hardcastle, "Optopalatograph: Development of a device for measuring tongue movement in 3D," in *EUROSPEECH '97*, Rhodes, Greece, 1997, pp. 1055–1058.

[9] L. Bombien, C. Mooshammer, P. Hoole, and B. Kühnert, "Prosodic and segmental effects on EPG contact patterns of word-initial German clusters," *Journal of Phonetics*, vol. 38, no. 3, pp. 388–403, 2010.

[10] A.-L. Fietkau, S. Stone, and B. Peter, "Relationship between the acoustic time intervals and tongue movements of German diphthongs," in *Proc. of the Interspeech 2022*, Incheon, Korea, 2022, pp. 734–738.

[11] F. E. Gibbon and S. E. Wood, "Visual feedback therapy with electropalatography (EPG) for speech sound disorders in children," in *Interventions in Speech Sound Disorders*. Paul H. Brookes Publishing Co., Inc., 2010, pp. 509–536.

[12] S. G. Fletcher, P. A. Dagenais, and P. Critz-Crosby, "Teaching vowels to profoundly hearing-impaired speakers using glossometry," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 4, pp. 943–956, 1991.

[13] M. J. Russell, D. Rubin, B. Wigdorowitz, and T. Marwala, "The artificial larynx: A review of current technology and a proposal for future development," in *14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*. Springer, 2008, pp. 160–163.

[14] S. Stone and P. Birkholz, "Cross-speaker silent-speech command word recognition using electro-optical stomatography," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Shanghai, China, 2020, pp. 7849–7853.

[15] P. Birkholz and C. Neuschaefer-Rube, "A new artificial palate design for the optical measurement of tongue and lip movements," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2012*, M. Wolff, Ed. TUDPress, Dresden, 2012, pp. 89–95.

[16] ——, "Combined optical distance sensing and electropalatography to measure articulation," in *Interspeech 2011*, Florence, Italy, 2011, pp. 285–288.

[17] S. Preuß, C. Neuschaefer-Rube, and P. Birkholz, "Real-time control of a 2d animation model of the vocal tract using optopalatography." in *Proc. of the Interspeech 2013*, Lyon, France, 2013, pp. 997–1001.

[18] S. Preuß and P. Birkholz, "Optical sensor calibration for electro-optical stomatography," in *Proc. of the Interspeech 2015*, Dresden, Germany, 2015, pp. 618–622.

[19] S. Stone and P. Birkholz, "Silent-speech command word recognition using electro-optical stomatography," in *Proc. of the Interspeech 2016*, San Francisco, USA, 2016, pp. 2350–2351.

[20] ——, "Angle correction in optopalatographic tongue distance measurements," *IEEE Sensors Journal*, vol. 17, no. 2, pp. 459–468, 2017.

[21] C. Wagner, L. Stappenbeck, H. Wenzel, P. Steiner, B. Lehnert, and P. Birkholz, "Evaluation of a non-personalized optopalatographic device for prospective use in functional post-stroke dysphagia therapy," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 356–365, 2022.

[22] C. Wagner, S. Stone, and P. Birkholz, "Towards combined force and distance sensing using only optical sensors to aid in stroke habilitation," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, A. Berton, U. Haiber, and W. Minker, Eds. TUDPress, Dresden, 2018, pp. 318–325.

[23] C. Wagner, P. Birkholz, L. Stappenbeck, and B. Lehnert, "Formulation of a palatal adhesive strip to secure electronic measuring devices inside the oral cavity," *Laryngo-Rhino-Otologie*, vol. 99, no. S 02, 2020.

PREASPIRATION IN ITALIAN VOICELESS GEMINATE AND SINGLETON STOPS

Angelo Dian, John Hajek, Janet Fletcher

University of Melbourne
a.dian@unimelb.edu.au, j.hajek@unimelb.edu.au, j.fletcher@unimelb.edu.au

ABSTRACT

This study is the first to explore from a cross-regional perspective acoustic phonetic features of preaspiration in both voiceless geminate and singleton stops in Italian, a language for which preaspiration is most typically associated with voiceless geminate stops. Frequency of preaspiration occurrence and duration are investigated in a controlled production experiment involving twelve speakers from two regional areas with different dialect substrata. Results reveal that preaspiration occurs for both geminates and singletons in both regions, with area-specific differences in frequency possibly linked to regional differences in phonetic voicing patterns of intervocalic singletons. We conclude that preaspiration in Italian stops may be best associated with phonetic voicing status, not phonological length.

Keywords: Preaspiration, voiceless stops, gemination, Veneto Italian, Roman Italian.

1. INTRODUCTION

1.1. What is preaspiration?

Preaspiration has been defined broadly as "a period of (usually glottal) friction that occurs between a vocalic and a consonantal interval" [10], as in [V^{h}C]. Initially believed to be a very rare or even extremely rare phenomenon (e.g., [16], [17], [20], [21]), preaspiration has been more recently associated with phonetically voiceless obstruents in a number of languages. [6] reports eighty-six languages across eighteen rather diverse language families that have been claimed to exhibit preaspiration, although [6] also points out that different definitions of the phenomenon are adopted across studies. The definitions of preaspiration mostly vary as to whether: (a) it only concerns voiceless stops or also fricatives and affricates; (b) it is realized as purely glottal [h ɦ] or supraglottal (e.g., [ɸ θ x ç]) friction; (c) it can involve vocal fold vibration alongside glottal (or supraglottal) friction. Where vocal fold vibration is present, preaspiration is generally labelled breathy voice/voicing or breathiness, whereas where only frication noise and no glottal periodicity is detected, it is labelled voiceless (glottal) friction [4], [10]. This latter type of preaspiration, particularly when purely glottal, has been referred to as e.g., "prototypical" [4], "archetypal" [23], "genuine" [17], or "true" [10]. It is this type in particular that is believed to be rare across the world's languages [17].

1.2. Italian

Italian is one of the languages for which preaspiration has been observed. It displays a word-medial phonological contrast between short (or singleton) and long (or geminate) consonants, as in /'fato/ 'fate' vs. /'fat:o/ 'fact' [2]. This contrast concerns the entire stop series including voiceless /p p: t t: k k:/. Italian is mainly spoken in Italy, a linguistically diverse country where different regional varieties of Standard Italian (SI) are found across its various geographic regions. On the one hand, SI is a prestigious, idealized form of Italian, spoken by a small number of educated people [2]. On the other hand, the regional varieties, or "accents" [2], are typically influenced by the local, co-existing Italo-Romance dialects and display distinctive phonetic characteristics, among other things [3]. Despite this fragmentation, the varieties can be broadly grouped into Northern and Central/Southern (cf. [14]). One of the differences between Northern and most Central/Southern varieties is that in the latter intervocalic singleton /p t k/ are frequently realized with some degree of phonetic voicing, as in /la pa'tata/ > [la p̆a't̬a̬t̬a] 'the potato' [3, pp. 55–56], with some exceptions. The most notable exception concerns the Tuscan varieties, where these consonants remain voiceless but are typically spirantized instead (see below).

1.3. Preaspiration in Italian stops

Among the Italian voiceless stop series, optional preaspiration has been extensively reported for geminate /p: t: k:/ [18]–[21], [23]–[25], whereas only one known study has reported it for singleton /p t k/ [22]. Like most earlier acoustic-phonetic studies on Italian preaspiration, [22] focussed on Sienese Italian, a Tuscan (Central) variety that typically spirantizes voiceless singleton /p t k/ as [ɸ θ h] in intervocalic position (as in /la pa'tata/ > [la ɸa'θaθa] 'the potato') [20, p. 341]. The preaspiration of singletons reported

for Sienese was interpreted as a variety-specific weakening of these spirantized stops. In later cross-regional studies this spirantization, as well as the assumed widespread lenition of voiceless singleton stops in other varieties, were considered problematic for a "straightforward duration-based analysis" of preaspiration [18, p. 22]. This possibly led to the decision to overlook singletons, something which was later reinforced by the hypothesis that preaspiration is employed by speakers to reinforce articulatorily [18], [19] or "maximize the perception" [21, p. 60] of Italian geminates. With time, the hypothesis that preaspiration characterised Italian voiceless geminate stops exclusively has become an underlying assumption [25].

1.4. Aims

This study investigates frequency of occurrence and duration of preaspiration in both geminates and singleton voiceless stops in Italian. Given the expected differences in singleton stop realization between Northern and Central/Southern varieties (cf. §1.2), one variety for each broader area is examined – namely, the Italian spoken in the Veneto region in the North-East (Veneto Italian) and the Italian spoken in the city of Rome, in the Centre-South (Roman Italian). As a Northern variety, Veneto Italian is expected to realize all phonologically voiceless stops as phonetically voiceless. Roman Italian, on the other hand, should optionally voice these consonants to different degrees, as previously reported [11]. Therefore, if preaspiration is to be associated with phonetic voicelessness and if it were to occur in singleton stops, it is expected that its frequency of occurrence for singletons would be lower in Roman than in Veneto Italian.

2. METHODS

2.1. Participants

Twelve adult speakers (age range: 32–68) were recruited for the study. Six were from the Veneto and six from Rome, with three males and three females for each area. All the Veneto participants were recorded in the area south-west of Vicenza, in Central Veneto, where they had lived all their lives. The Roman participants were recorded in Melbourne, Australia, although they had all grown up and lived in Rome for most of their lives.

2.2. Materials and procedure

A controlled acoustic phonetic experiment was designed whereby the participants were asked to read out a series of carrier sentences containing a list of real Italian experimental words, shown in Table 1.

Target /p p: t t: k k:/ consonants were embedded in the experimental words and flanked preferably by low vowels or alternatively mid vowels. To test the effect of lexical stress, the consonants either directly followed (post-stress) or preceded (pre-stress) the stressed vowel, as in /'fa**t**o/ vs. /pa'**t**ata/. Post-stress words were disyllabic and pre-stress words trisyllabic paroxytones. The effect of position in the phrase was also tested, with experimental words in either phrase-final (*"Ho detto WORD"*, 'I said WORD') or phrase-medial *("Ho detto WORD prima"*, 'I said WORD before') position. The words were all nuclear accented; to ensure this, the participants were asked to read the sentences as if they were answering a question that placed the focus on the experimental word: *"Che hai detto?/Che hai detto prima?"*, 'What did you say?/What did you say before?'.

Phon	Post-stress		Pre-stress	
	Sing	Gem	Sing	Gem
/p/	*Papa* /ˈ**pa**pa/ 'Pope'	*pappa* /ˈ**pap**pa/ 'mush'	*Papato* /pa**ˈp**ato/ 'Papacy'	*pappato* /p**ap**ˈ**p**ato/ 'gobbled up'
/t/	*fato* /ˈ**fa**to/ 'fate'	*fatto* /ˈ**fat**to/ 'fact'	*patata* /pa**ˈt**ata/ 'potato'	*dettato* /d**et**ˈ**t**ato/ 'dictation'
/k/	*paca* /ˈ**pa**ka/ 'he/she calms'	*pacca* /ˈ**pak**ka/ 'pat'	*pacato* /pa**ˈk**ato/ 'placid'	*paccato* /p**ak**ˈ**k**ato/ 'let down' (p.p.)

Table 1: List of experimental words. /VC(C)/ sequences are in bold.

The carrier sentences were displayed one by one on a computer screen in random order. Each of the twelve experimental words was repeated four times for each of the two position-in-the-phrase conditions by each of the twelve speakers. One token was discarded due to background noise, for a total of 1151 tokens. The Veneto speakers were recorded through a professional, portable solid-state recorder in a quiet, noise-reduced room while the Roman speakers were recorded in a sound-proof recording studio. The recordings for both groups were sampled at 44100 Hz with 16-bit quantization.

2.3. Analysis

The phonetic annotation of the experimental words was carried out manually in EMU-SDMS [26]. Where present, breathy voice in the vowel (BV) and voiceless glottal friction (GF) were identified and labelled separately. In line with [15], the onset of BV was placed where the waveform began to appear more sinusoidal (and/or visibly jagged, cf. [5]) and

the formants more blurred in the spectrogram; its offset coincided with either the onset of GF where present or the onset of stop closure. The onset of GF was placed where increased energy in the higher frequency range of the spectrogram was visually observed along with a clear reduction in amplitude, or cessation altogether, of periodicity as visible from the waveform. An example is shown in Figure 1.

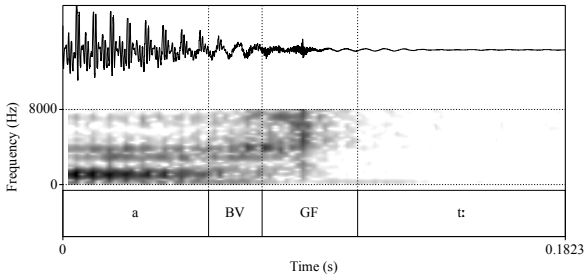


Figure 1: Labelled example of a token /fatto/ displaying both BV and GF, produced by a female Roman speaker.

Previous studies of Italian preaspiration (cf. §1.3) adopted the broader definition of the phenomenon as BV, GF, or both (henceforth, BV+GF). This study also seeks to investigate the distribution patterns of GF alone in line with recent investigations on other languages (e.g., [9]). Thus, the tokens displaying BV+GF and the tokens displaying GF alone were counted separately. The duration of BV+GF was also measured. Tokens displaying no preaspiration were excluded from the durational analysis (cf. [8]).

The data were extracted through emuR [27]. Statistical analyses testing the effect of controlled fixed factors on frequency of occurrence and duration of preaspiration were conducted within the statistical environment R. The fixed factors are as follows:

(a) AREA (Rome/Veneto);
(b) SPEAKER SEX (female/male);
(c) GEMINATION (singleton/geminate);
(d) STRESS CONDITION (post-stress/pre-stress);
(e) POSITION (IN THE PHRASE) (phrase-final/phrase medial);
(f) PLACE OF ARTICULATION (bilabial/denti-alveolar/velar), henceforth PoA.

SPEAKER and WORD were treated as random factors. For frequency of occurrence, two Mixed-Effects Binomial Logistic Regression Models (glmer function, part of [1]) were utilized to predict the effect of all the fixed factors and the interaction of factors (a) and (c) on the probability of occurrence of BV+GF and GF, respectively. Both models had by-SPEAKER and by-WORD varying intercepts. For duration, a Linear Mixed-Effects (LME) model [12] was employed to identify the effect of the fixed factors and their interaction on BV+GF duration. The LME model was at first maximally specified and then step-reduced (through the step function [12]) to obtain the best-fitting model. The final model had by-SPEAKER varying intercepts and varying (d) slopes. The emmeans function [13] was used for post-hoc tests.

3. RESULTS

3.1. BV+GF frequency and duration

Overall, BV+GF was observed in 794/1151 (69%) tokens. It was frequently displayed by both geminates (462/575, 80%) and singletons (332/576, 58%). It was also more frequent in post-stress (451/576, 78%) than in pre-stress (343/575, 60%) tokens. Figure 2 shows that BV+GF frequency increased with more posterior PoA, with 130/384 (34%) bilabial, 308/384 (80%) denti-alveolar, and 356/383 (93%) velar BV+GF tokens. It also shows that for geminates BV+GF occurred with similar frequency across areas, while for singletons it was less frequent for Roman than for Veneto speakers. The results are confirmed by the statistical analysis, with significant main effects of gemination (logit = 2.723, z = 8.601, $p < .001$), PoA (logit = 4.601, z = 12.923, $p < .001$), stress condition (logit = -1.669, z = -6.889, $p < .001$), and a significant interaction between area and gemination (logit = -1.334, -3.520, $p < .001$). Post-hoc tests confirm somewhat greater differences in singleton vs. geminate BV+GF probability of occurrence for the Roman variety (logit = -2.723, z = -8.601, $p < .001$) than the Veneto variety (logit = -1.388, z = -4.445, $p < .001$). Speaker sex and position did not reach statistical significance.

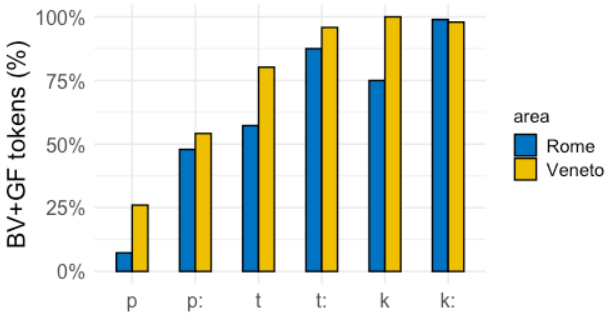


Figure 2: Frequency of occurrence of BV+GF tokens by area, PoA, and gemination types.

As for preaspiration duration, Figure 3 shows no clear effect of gemination on the distribution of non-zero BV+GF values. Overall, higher values were found for post-stress ($\mu = 38$ ms, $\sigma = 26$ ms) than pre-stress tokens ($\mu = 23$ ms, $\sigma = 14$ ms). Furthermore, the values increased with posteriority of PoA: bilabial: $\mu = 18$ ms, $\sigma = 11$ ms; denti-alveolar: $\mu = 25$ ms, $\sigma = 17$ ms; velar: $\mu = 43$ ms, $\sigma = 26$ ms. The statistical analysis only found significant main effects

of stress condition (β = 39.968 ms, t = 3.721, p < .001) and PoA (β = 16.716 ms, t = 2.882, p < .01).

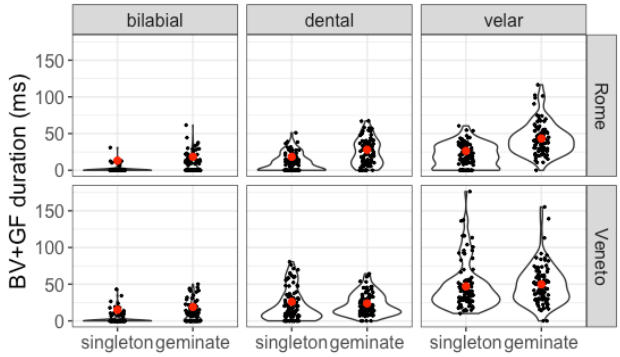


Figure 3: Distribution of BV+GF duration values by area, PoA, and gemination types. Red dots indicate mean non-zero values.

3.2. GF frequency

A clearly different pattern in frequency of occurrence emerges when GF alone is considered. Apart from lower overall frequency (548/1151, 48%), striking cross-regional differences are observed for GF within singletons. As Figure 4 shows, Veneto singleton GF rates of occurrence are comparable to those of geminates (singletons: 152/288, 53%; geminates: 182/288, 63%), while for Rome singletons these are much lower (singletons: 34/288, 12%; geminates: 180/288, 63%). Furthermore, across regions post-stress GF tokens (293/576, 51%) were somewhat more frequent than their pre-stress counterparts (255/575, 44%). Finally, bilabial GF tokens (30/384, 8%) were less frequent than denti-alveolars (233/384, 61%) and velars (295/383, 77%).

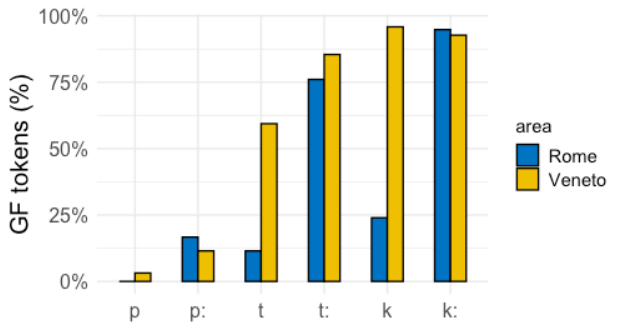


Figure 4: Frequency of occurrence of GF tokens by area, PoA, and gemination types.

Statistically significant main effects of gemination (logit = 4.210, z = 11.246, p < .001), area (logit = 3.293, z = 9.069, p < .001), and a significant interaction between gemination and area (logit = -3.260, z = -8.098, p < .001) were found. This time, post-hoc tests found considerably greater differences in singleton vs. geminate GF probability of occurrence for the Roman variety (logit = -4.210, z = -11.246, p < .001) than the Veneto variety (logit = -0.950, z = -2.937, p < .05). In other words, GF was nearly as likely to occur in Veneto singletons as in geminates, while it was much less likely to occur in Roman singletons than geminates. The model also found a highly significant main effect of PoA (logit = 5.226, z = 13.341, p < .001) and a significant main effect of stress condition (logit = -0.559, z = -2.070, p < .05). Speaker sex and position were not significant.

4. DISCUSSION AND CONCLUSION

The main finding of this study is that preaspiration occurrence in Italian voiceless stops is not confined to geminates as generally thought (but see [22]) – on the contrary, it is also frequent in singletons. As proposed in §1.4, the results also suggest that cross-regional differences may be due to variety-specific variation in the surface realization of underlyingly voiceless singleton stops. Northern varieties of Italian do not typically voice or lenite these stops, and in fact our Northern speakers exhibited comparable: (a) frequency of occurrence and duration of preaspiration; and (b) frequency of occurrence of GF, between singletons and geminates. Conversely, Roman speakers, who are expected to variably voice intervocalic singletons, showed less frequent preaspiration for singletons than for geminates, particularly when only GF was considered. This indicates that preaspiration in our study may be better associated with (phonetic) voicelessness (as alternatively hypothesized by [21]) than phonological length, although future investigations including degree of voicing are necessary to validate this hypothesis. Frequency of GF in singletons appears to be a robust discriminator between varieties (cf. §3.2). Thus, this "prototypical" preaspiration may be more reliably associated with phonetic voicelessness of stops than preaspiration as more broadly defined.

Regarding stress condition and PoA, this study confirms previously reported trends. Post-stress stops display more frequent, longer preaspiration than pre-stress stops, and preaspiration frequency and duration increase with posteriority of PoA [19], [24]. The overall frequency of occurrence of preaspiration as broadly defined, however, is much higher in this study (69%) than in previous ones (ca. 30-40% [18]–[21], [23]–[25]), probably due to all the tokens being in nuclear accented position within the sentence in this study. This explanation would corroborate a proposed association of Italian preaspiration with fortition [18]. Finally, in line with previous investigations [24], speaker sex does not seem to affect the occurrence or duration of preaspiration.

5. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting Linear Mixed-Effects Models using lme4. [Online]. Available: http://arxiv.org/abs/1406.5823

[2] Bertinetto, P.M., Loporcaro, M. 2005. The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association* 35(2), 131–151.

[3] Canepari, L. 1992. *Manuale di pronuncia italiana: con un pronunciario di oltre 30,000 Voci e due audiocassette C45*, 1. ed. Bologna: Zanichelli.

[4] Clayton, I. D. 2010. *On the Natural History of Preaspirated Stops*. PhD dissertation. University of North Carolina, Chapel Hill. Accessed: Sep. 14, 2022. [Online]. Available: https://cdr.lib.unc.edu/concern/dissertations/7m01bk82b?locale=en

[5] Gordeeva, O. B., Scobbie, J. M. 2013. A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality. *Journal of the International Phonetic Association* 43(3) 249–271.

[6] Hejná, M. Forthcoming. On the rarity of pre-aspirated consonants. In: Anderson, C., Kuznetsova, N., Easterday, S. (eds), *Rarities in Phonetics and Phonology: Evolutionary, Structural, Typological and Social Dimensions*. Berlin: Language Science Press.

[7] Hejná, M. 2015. Pre-aspiration in Welsh English: A case study of Aberystwyth. PhD dissertation. University of Manchester. Accessed: Nov. 13, 2021. [Online]. Available: http://rgdoi.net/10.13140/RG.2.1.3485.3842

[8] Hejná, M. 2019. Pre-aspiration and the problem of zeroes: Phonological rule can be variable. In: *The Sign of the V: Papers in Honour of Sten Vikner*, Aarhus University Library, 227–242.

[9] Hejná, M., Kaźmierski, K., Guo, W. 2021. Even Americans pre-aspirate. *English World-Wide* 42(2) 200–226.

[10] Helgason, P. 2002. *Preaspiration in the Nordic Languages: Synchronic and Diachronic Aspects*. PhD dissertation. University of Stockholm.

[11] Hualde, J. I., Nadeu, M. 2012. Lenition and phonemic overlap in Rome Italian. *Phonetica* 68(4) 215–242.

[12] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Soft.* 82(13) 1–26.

[13] Lenth, R.V. et al. 2022. Package 'emmeans'. [Online]. Available: https://cran.r-project.org/web/packages/emmeans/emmeans.pdf

[14] Mairano, P., De Iacovo, V. 2020. Gemination in Northern versus Central and Southern varieties of Italian: A corpus-based investigation. *Language and Speech* 63(3) 608–634.

[15] Morris, J. 2010. Phonetic variation in Northern Wales: Preaspiration. In: Meyerhoff, M., Adachi, C., Daleszynska, A., Strycharz, A. (eds), *Proceedings of the Second Summer School of Sociolinguistics*, University of Edinburgh, 1–16.

[16] Ní Chasaide, A. 1985. *Preaspiration in Phonological Stop Contrasts*. PhD dissertation. University College of North Wales, Bangor.

[17] Silverman, D. 2003. On the rarity of pre-aspirated stops. *Journal of Linguistics* 39(3) 575–598.

[18] Stevens, M. 2011. Consonant length in Italian: Gemination, degemination and preaspiration. In: Alvord, S. M. (ed). *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*, Cascadilla Press, 21–32.

[19] Stevens, M. 2010. How widespread is preaspiration in Italy? *Working Papers* (54), Lund University, Department of Linguistics and Phonetics, 97–102.

[20] Stevens, M., Hajek, J. 2004. Comparing voiced and voiceless geminates in Sienese Italian: What role does preaspiration play? In: *Proceedings of SST 2004*, Sydney, 340–345.

[21] Stevens, M., Hajek, J. 2004. Preaspiration in Sienese Italian and its interaction with stress in /VC:/ sequences. In: *SProSIG 2004,* Nara*,* 57–60.

[22] Stevens, M., Hajek, J. 2005. Spirantization of /p t k/ in Sienese Italian and so-called semi-fricatives. In: *Interspeech 6th 2005*, Lisbon, 2893-2896

[23] Stevens, M., Hajek, J. 2007. Towards a phonetic conspectus of preaspiration. In: *Proc. 16th ICPhS*, Saarbrücken, 429–432.

[24] Stevens, M., Hajek, J. 2010. Preaspirated /pp tt kk/ in standard Italian. A sociophonetic v. phonetic analysis. In: *Proceedings of SST 2010*, Melbourne, 1–4.

[25] Stevens, M., Reubold, U. 2014. Pre-aspiration, quantity, and sound change. *Laboratory Phonology* 5(4) 455–488.

[26] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45 392–410.

[27] Winkelmann, R., Jänsch, K., Cassidy S., Harrington, J. 2021. emuR: Main package of the EMU Speech Database Management System. [Online]. Available: https://cran.r-project.org/web/packages/emuR/emuR.pdf

Pitch-shift Responses to Auditory Perturbation Involve a Mix of Opposing and Following Responses: Evidence from Steady Vowel and Glissando Vocalizations

Li-Hsin Ning
Department of English, National Taiwan Normal University, Taiwan
lihsin@ntnu.edu.tw

ABSTRACT

In this paper, we investigate whether opposing responses and following responses are equally likely to appear in the pitch-shift responses for steady vowels and glissando vocalizations. Two groups of participants were recruited: seven musicians and seven non-musicians. The results show that following responses were not a minority. Musicians and non-musicians produced a mix of opposing and following responses (almost equally likely) for steady vowels and glissando vocalizations. This result suggests that both mechanisms should be activated simultaneously in online voice pitch regulation.

Keywords: audio-vocal control, pitch-shift paradigm, opposing response, following response, glissando

1. INTRODUCTION

Auditory feedback plays an essential role in speech production. When unexpected changes occur in auditory feedback, speakers would make opposing (also called compensatory) responses that go in the opposite direction from the unexpected shifts to reduce the perceived errors. Vocal compensation can appear in steady vowel production [1], word production [2, 3], glissando vocalizations [4], and singing [5]. However, research has shown that opposing responses may not be the majority in the responses to auditory perturbation and that following responses (i.e., responses that go in the same direction as the direction of unexpected shifts) can occupy a considerable proportion of steady vowels and tone word (level tone) productions if we examine the pitch contours on a trial basis [6-8]. It is unclear whether both mechanisms (opposing and following) would be equally likely to appear in glissando vocalizations in which pitch trajectories are not stable. To investigate glissando vocalizations, it would also be interesting to explore how musicians and non-musicians utilize opposing and following mechanisms. If musicians are less affected by pitch perturbation and have reduced compensation [9], we would expect that they would disfavor following responses that generate a larger deviation from the intended pitch.

2. METHOD

2.1. Participants

Fourteen native speakers of Mandarin (20 ~ 30 years old; M = 23, SD = 3.2) were recruited for this study. Half of them were music majors at the time of the experiment. They played piano, cello, violin, flute, or tuba. The other half were non-musicians and had not played music as an amateur in the previous five years. All the participants reported no history of speech or hearing disorders. They passed a hearing screening test (20 dB bilaterally at 250, 500, 750, 1000, 2000, 3000, and 4000 Hz) prior to recording using a MAICO pure-tone audiometer (model MA 25).

2.2. Materials and Procedures

The experiment consisted of three production blocks: upward glissando, downward glissando, and steady vowel. In the upward or downward glissando, participants first listened to a model note, which started with a steady note for 500 ms, was transitioned into an upward or downward glide (100 cents/half second for 2 seconds), and then was held steady again for 500 ms. Participants were instructed to mimic the glissando pattern (without vibrato) using the vowel /a/ within their comfortable pitch range. In the steady-vowel condition, participants were asked to produce a sustained vowel /a/ (for roughly 3 seconds) after hearing the beep sound. Each recording block contained 30 repetitions, for a total of 90 trials (30 repetitions × 3 blocks = 90 trials) per participant.

In each vocalization, participants may have heard a single pitch-shift stimulus presented in the auditory feedback. The voice f_0 was shifted upward, shifted downward, or not shifted (as a control), which were equally likely to occur randomly in a recording block. The pitch-shift stimuli were all 200 ms long and fixed at ±100 cents. They appeared at a random time, 500 ~ 700 ms, after vocal onset. The intertrial delay was 1,000 ms. Participants were asked to *ignore* pitch-shift stimuli and maintain the intended pitch contour. The order of the three recording blocks was counterbalanced across participants.

2.3. Apparatus

The participants sat in a soundproof booth and wore AKG K240 headphones. We placed a standalone microphone, Audio Tech ATR20, in front of them. The voice signal from the microphone was processed (shifted in pitch) in near-real time (14 ~ 20 ms delay) through an Eventide Ultra-Harmonizer (model H7600), which was controlled by Max/MSP (v.8, Cycling 74). To mask bone-conducted auditory feedback, we amplified the voice signal's intensity with a 10-dB gain via a McLELLAND MAR-16P headphone amplifier when presented over the headphones. The vocalizations, pitch-shifted signals, and TTL pulses indicating the onset of pitch-shift stimuli were recorded using a WinDaq DI-720 acquisition device and sampled at 8 kHz per channel in WinDaq Pro.

2.4. Data Preprocessing

The signals recorded in WinDaq were imported into MATLAB (R2020a) and sorted based on the direction of pitch-shift stimuli (upward shift, downward shift, or no shift). Each vocalization was segmented into a 1.2-s long signal that included a 200-ms preshift period, a 200-ms shift period, and an 800-ms postshift period. We then wrote the segmented signals as sound files and estimated the pitch every 10 ms in PRAAT. The pitch values were transformed into cents using the formula cents=1200×log2(f_0/baseline), where the baseline indicated the mean pitch of the preshift period. The f_0 records were then imported back into MATLAB for further processing.

Each segmented vocalization was classified as one of the following four types by comparing it with the averaged pitch contour of the corresponding control trials: opposing response, following response, non-response, or error. If the response changed in the opposite direction to the pitch-shift stimulus, it was tagged as an "opposing" response. If the response changed in the same direction as the pitch-shift stimulus, it was tagged as a "following" response. If the response did not show a clear upward or downward deviation from the averaged control (i.e., less than ±2 SD of the preshift mean), it was tagged as a "non-response." If the response showed an erroneous pitch-tracking result, it was tagged as an "error."

After the classification, we calculated the percentage of opposing responses, following responses, non-responses, and errors under each condition (3 production blocks × 2 stimulus directions) for each participant. Difference waves were obtained for the upward glissando and downward glissando by subtracting the averaged pitch contour of the corresponding control trials from the averaged opposing or following pitch contours at every point (see Fig. 1). We conducted statistical analyses using generalized additive mixed models of the difference waves (cent values) for the upward and downward glissando conditions and the standardized pitch contours (cent values) for the steady-vowel condition.

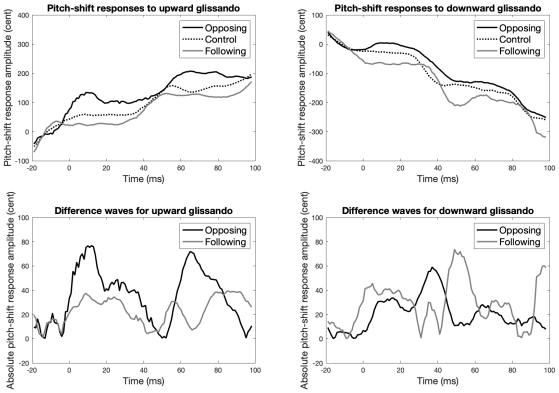


Figure 1: Illustration of averaged opposing responses (dark solid lines), averaged following responses (grey solid lines), and their difference waves (obtained by subtracting the averaged controls displayed in the dotted lines) for one participant in the upward glissando (the left panel) and downward glissando (the right panel).

3. RESULTS

3.1. Proportions of the opposing and following responses

Table 1 displays the mean percentages of opposing responses, following responses, nonresponses, and errors by musicality, production block, and pitch-shift stimulus direction. The results show that our participants produced a mix of opposing and following responses in glissando and steady vowel vocalizations. For musicians, *opposing* responses were more common than following responses in the steady vowel vocalizations whereas opposing and following responses were equally likely to appear in the glissando vocalizations. However, for non-musicians, *following* responses were more common than opposing responses when the pitch-shift stimuli went in an opposite direction from the intended overall pitch trajectories, such as downshift stimuli in the upward glissando (55%) and upshift stimuli in the downward glissando (47%). These findings suggest that following responses, as one response type in the face of auditory perturbation, were not neglectable, which implies that both mechanisms (opposing and following) could be activated simultaneously when we encounter unexpected perturbation.

MUSICALITY	Production block	Pitch-shift stimulus direction	Opposing response	Following response	Non-response	Error
Musician	Glissando up	Upshift	44%	42%	0%	14%
		Downshift	34%	47%	4%	14%
	Glissando down	Upshift	55%	31%	5%	9%
		Downshift	44%	43%	5%	8%
	Steady vowel	Upshift	65%	34%	0%	1%
		Downshift	62%	37%	1%	0%
Non-musician	Glissando up	Upshift	65%	18%	4%	13%
		Downshift	34%	55%	4%	7%
	Glissando down	Upshift	27%	47%	10%	16%
		Downshift	65%	23%	2%	10%
	Steady vowel	Upshift	55%	41%	2%	2%
		Downshift	39%	55%	1%	5%

Table 1: The mean percentages of opposing responses, following responses, nonresponses, and errors by musicality, production block, and pitch-shift stimulus direction.

3.2. Time-varying changes in pitch-shift responses

We used generalized additive mixed models (GAMMs) to account for the time-varying changes in pitch-shift responses. The `bam()` function in the `mgcv` package of R [10]. All the GAMMs involved a factor smooth that accounted for the nonlinear random effects (intercepts and slopes) derived from individual variability across time. This method fits a wiggly curve for each participant, using the same calibration for penalization across all participants. For details of how to create baseline models, conduct model comparisons, and set up parametric predictors (musicality, production block, stimulus direction), please refer to Sun and Shih [11] and Ning [8]. In what follows, we presented two sets of model fitting results: one for *opposing responses* and the other for *following responses*.

3.2.1. Opposing responses

Fig. 2 depicts musicality's effect on the time-varying opposing responses. The results show that non-musicians had *larger* opposing responses than musicians (constant difference) in the upshift of downward glissando (t=2.307, p<.05), the downshift of upward glissando (t=2.278, p<.05), the upshift of upward glissando (t=2.218, p<.05), and the upshift of the steady-vowel (t=–1.899, p<.05) condition. We observed the dynamic changes over time in the upshift of the steady-vowel (F(6.568, 7.694)=3.278, p<.01) condition, in which the divergence between musicians and non-musicians increased significantly from vocal onset to vocal offset.

Fig. 3 depicts the production block effect on the opposing responses. We selected the steady-vowel condition as the reference level. The results demonstrate that opposing responses in the glissando conditions were constantly *larger* than those in the steady-vowel condition. Significant constant height differences appeared in almost all combinations of musicality and stimulus direction: musicians' downshift in downward glissando (t=3.610, p<.001), musicians' upshift in downward glissando (t=0.995, p=.320), musicians' downshift in upward glissando (t=2.733, p<.01), musicians' upshift in upward glissando (t=2.085, p<.05), non-musicians' downshift in downward glissando (t=4.088, p<.001), non-musicians' upshift in downward glissando (t=2.998, p<.01), non-musicians' downshift in upward glissando (t=2.760, p<.01), and non-musicians' upshift in upward glissando (t=2.852, p<.01).

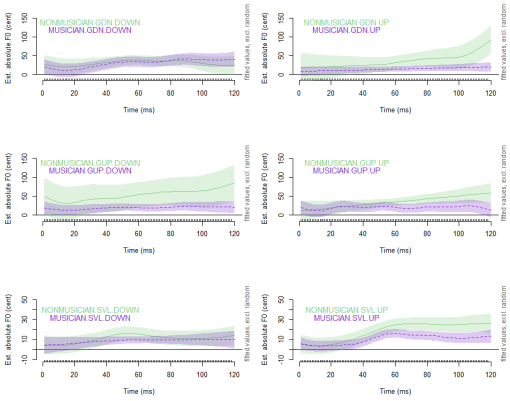


Figure 2: Musicality effect (musicians vs. non-musicians) regarding *opposing responses* at each level of the combination of production block (GDN, GUP, SVL) and stimulus direction (DOWN, UP). The purple curves represent musicians' opposing responses (in absolute values), and the green curves represent non-musicians' opposing responses (in absolute values). The pitch-shift stimuli appeared at 20 ~ 40 ms on the x-axis.

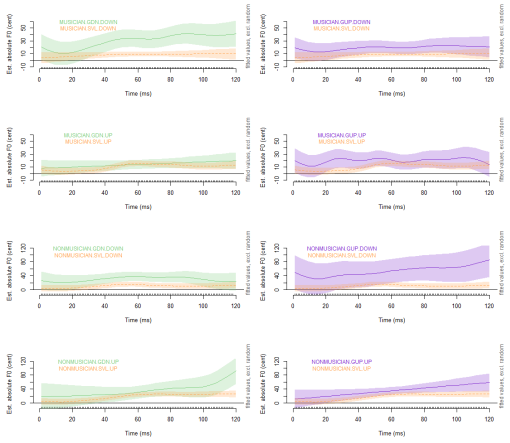


Figure 3: Production block effect (upward glissando, downward glissando, steady vowel) regarding *opposing responses* at each level of the combination of musicality (MUSICIAN, NONMUSICIAN) and stimulus direction (DOWN, UP). The purple curves represent opposing responses (in absolute values) for upward glissando; the green curves represent opposing responses (in absolute values) for downward glissando; the orange curves represent opposing responses (in absolute values) for the steady-vowel condition.

Fig. 4 depicts stimulus direction's effect on the opposing responses. Generally, opposing responses

to downshifts were *larger* than opposing responses to upshifts in the glissando conditions whereas we found the opposite pattern (downshifts *smaller* than upshifts) in the steady-vowel condition.

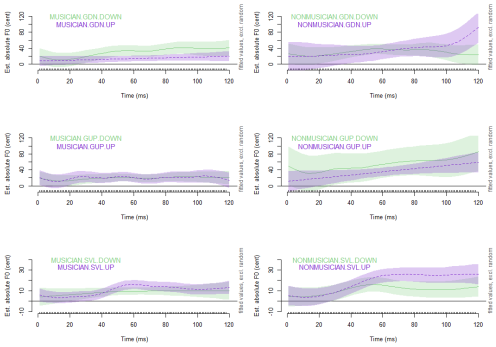


Figure 4: Stimulus direction effect (upshift vs. downshift) regarding *opposing responses* at each level of the combination of musicality (MUSICIAN, NONMUSICIAN) and production block (GDN, GUP, SVL). The purple curves represent opposing responses (in absolute values) for upshift stimuli; the green curves represent opposing responses (in absolute values) for downshift stimuli.

3.2.2. Following responses

Similar to opposing responses, non-musicians also demonstrated *larger* following responses than musicians (see Fig. 5). However, this musicality effect was only significant in the glissando vocalizations: the upshift of downward glissando (t=2.307, p<.05), the downshift of downward glissando (t=2.218, p<.05), and the downshift of upward glissando (t=2.278, p<.05).

Fig. 6 displays the production block effect on the following responses. As in the opposing responses, following responses in the glissando conditions were *larger* than those in the steady-vowel condition (all p<.05).

Finally, for the stimulus direction effect, we found a near-significant pattern in the non-musicians' steady-vowel productions, in which following upshifts led to *larger* responses than following downshifts (t=–1.681, p=.09).

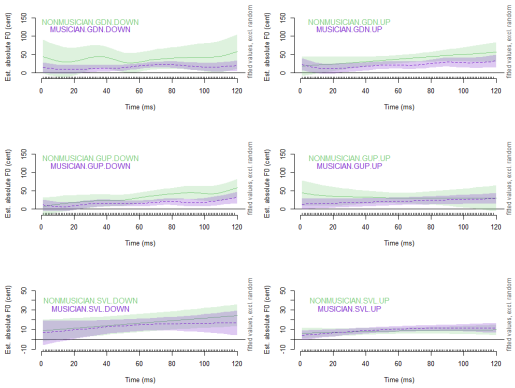


Figure 5: Musicality effect regarding *following responses* at each level of the combination of production block and stimulus direction. The purple curves represent musicians' following responses and the green curves represent non-musicians' following responses.

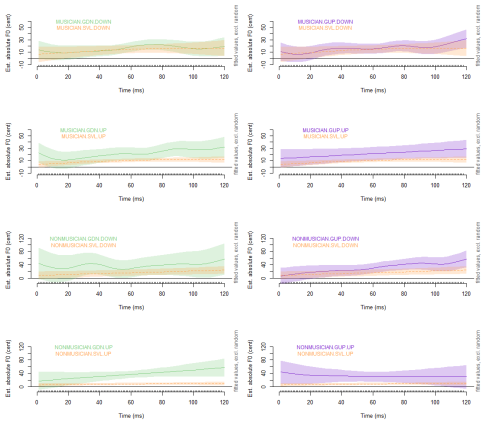


Figure 6: Production block effect regarding *following responses* at each level of the combination of musicality and stimulus direction. The purple curves represent following responses for upward glissando; the green curves represent following responses for downward glissando; the orange curves represent following responses for the steady-vowel condition.

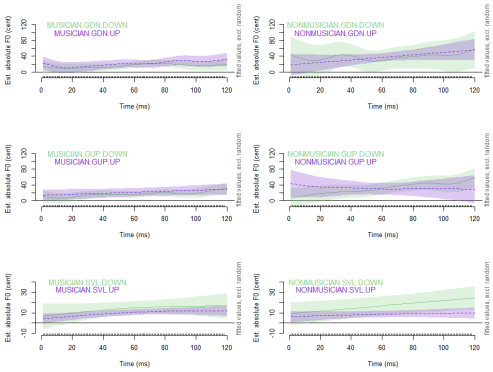


Figure 7: Stimulus direction effect regarding *following responses* at each level of the combination of musicality and production block. The purple curves represent following responses for upshift stimuli; the green curves represent following responses for downshift stimuli.

4. DISCUSSION AND CONCLUSION

Our results show that for musicians and non-musicians, following responses were almost as common as opposing responses when they encountered unexpected perturbations in auditory feedback. This finding suggests the opposing and following mechanisms could be activated simultaneously so that speakers may utilize one of the mechanisms from time to time. Following responses' availability should be independent of vocalization task (vowels and glissandos) and musical background. As the literature suggests, our musicians were also less affected (i.e., smaller responses) by perturbations than non-musicians were. The advantage in vocal pitch control should be associated with many years of musical training. We discovered larger (opposing and following) responses in the scenarios that required more rigid pitch control, such as in glissando vocalizations. Using generalized additive mixed models enables us to discover constant differences and time-varying changes in the pitch contours.

5. REFERENCES

[1] T. A. Burnett, M. B. Freeland, and C. R. Larson, "Voice F0 responses to manipulations in pitch feedback," *Journal of the Acoustical Society of America,* vol. 103, no. 6, pp. 3153-3161, 1998.

[2] S. H. Chen, H. Liu, Y. Xu, and C. R. Larson, "Voice F0 responses to pitch-shifted voice feedback during English speech," *Journal of the Acoustical Society of America,* vol. 121, no. 2, pp. 1157-1163, 2007.

[3] L.-H. Ning, "Pitch-shift responses as an online monitoring mechanism during level tone production," *Journal of the Acoustical Society of America,* vol. 145, no. 4, pp. 2192-2197, 2019, doi: https://doi.org/10.1121/1.5096977.

[4] T. A. Burnett and C. R. Larson, "Early pitch-shift response is active in both steady and dynamic voice pitch control," *Journal of the Acoustical Society of America,* vol. 112, no. 3, pp. 1058-1063, 2002.

[5] U. Natke, T. M. Donath, and K. T. Kalveram, "Control of voice fundamental frequency in speaking versus singing," *Journal of the Acoustical Society of America,* vol. 113, no. 3, pp. 1587-1593, 2003.

[6] R. Behroozmand, O. Korzyukov, L. Sattler, and C. R. Larson, "Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control," *The Journal of the Acoustical Society of America,* vol. 132, no. 4, pp. 2468-2477, 2012, doi: 10.1121/1.4746984.

[7] M. K. Franken, D. J. Acheson, J. M. McQueen, P. Hagoort, and F. Eisner, "Opposing and following responses in sensorimotor speech control: Why responses go both ways," (in eng), *Psychon Bull Rev,* vol. 25, no. 4, pp. 1458-1467, Aug 2018, doi: 10.3758/s13423-018-1494-x.

[8] L.-H. Ning, "Comparison of Involuntary and Volitional Responses to Pitch-Shifted Auditory Feedback: Evidence for Tone Speakers' Flexibility to Switch Between Opposing and Following Responses," *Journal of Speech, Language, and Hearing Research,* vol. 65, no. 6, pp. 2160-2186, 2022, doi: doi:10.1044/2022_JSLHR-21-00597.

[9] J. M. Zarate and R. J. Zatorre, "Experience-dependent neural substrates involved in vocal pitch regulation during singing," *NeuroImage,* vol. 40, pp. 1871-1887, 2008.

[10] S. Wood, *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC, 2006.

[11] Y. Sun and C. Shih, "Boundary-conditioned anticipatory tonal coarticulation in Standard Mandarin," *Journal of Phonetics,* vol. 84, p. 101018, 2021, doi: https://doi.org/10.1016/j.wocn.2020.101018.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to the research assistants, Hao-Yu Wu, Pei-Jiun Chen, and Hsin-Jou Chen, for their invaluable assistance with data collection. Their contributions were crucial to the success of this research project.

AN ARTICULATORY STUDY OF ESTONIAN PALATALIZATION

Anton Malmi, Pärtel Lippus

Institute of Estonian and General Linguistics, University of Tartu, Estonia
anton.malmi@ut.ee, partel.lippus@ut.ee

ABSTRACT

This paper analyzes the effect of palatalization on the articulation and duration of Estonian /s/, /t/, /n/, /l/, and the vowels preceding them in phonologically contrastive word pairs. The experiment with 21 native Estonian participants was carried out using an electromagnetic articulograph (EMA). The results show that in the case of palatalization, the back of the tongue was already raised, and the tongue blade was more fronted during the preceding vowel. The position was maintained through the palatalized consonants. The height of the tongue blade was not affected by palatalization. There was a positive correlation between higher tongue position and longer vowel duration. With consonants, the duration was shorter with higher tongue position, but this effect was inconsistent.

Keywords: articulation, articulography, Estonian, palatalization, Carstens AG501, EMA

1. INTRODUCTION

This paper aims to study the effect of palatalization on the articulatory and temporal properties of Estonian consonants /s/, /t/, /n/, /l/, and the vowels that precede them. Palatalization is a process wherein consonants' place of articulation is altered by a high front vowel or glide /j/ [1]. Estonian has developed many phonologically contrastive pairs in which palatalization is the only distinctive feature, e.g., *pats* [pɑtːs] 'pat' – *pats* [pɑtʲːs] 'braid.' These pairs give a comparable study material to assess the effect of palatalization on articulation.

Studies of different languages have shown that the most persistent descriptive acoustic feature of palatalization is the rise of the F2 value of the vowels preceding consonants [2–9]. This rise in F2 is due to the gesture in which the tongue body rises towards the palate. Studies that have looked at vowel-to-consonant coarticulation [10–13] have found that the following segments are anticipated earlier. The effect of coarticulation might even reach further beyond the syllable boundaries [14, 15].

The rising of the tongue in the vowels have also been observed in the acoustic studies of Estonian consonants. /l/ has higher F2 values with palatalization [3]. The mean spectral energy or the center of gravity of the spectrum (COG) has been measured for plosives and fricatives. COG value can be associated with the frontness (higher frequencies) or backness (lower frequencies) of the tongue [20]. COG is lower for /s/ [16, 17], /t/, and /n/ and higher for /l/ with palatalization in Estonian and in Ocotepec Mixe [17].

Palatographic studies have shown that the ridge in the middle of the tongue was narrower when articulating Estonian palatalized /s/ [18, 19]. This might also explain why the COG values are lower. The place of articulation of /t/ is retracted in Estonian, Russian, and Polish, and the tongue blade and the sides have a wider contact on the alveolar ridge, and the teeth [18–21]. Estonian /n/ has a wider contact area on the post-alveolar region and the teeth [18,19,22]. Estonian /l/ is fronted with palatalization and has a wider contact area on the alveolar ridge and teeth [19, 22, 23].

The raising of the tongue before palatalized consonants has also been shown to lengthen the duration of the vowel in Estonian [3, 9, 24]. This lengthening is used to enhance the perception of palatalization [25]. The duration of Russian /t/ is longer with palatalization [21, 26], as there is a higher degree of aspiration in the release. With /s/, the results vary: [26] found that the duration was shorter with palatalization, and [27] found that it was longer. We could not find any data on the duration of /l/ or /n/. Previous studies have analyzed the duration as a byproduct of the rising of the tongue but have not tried to link it to an articulatory movement.

The present study is a follow-up to [17], where we offer a new perspective on analyzing the dynamic movement trajectories of palatalization. We investigate whether the rising of the tongue can be observed to occur earlier in the vowels and reach further in the consonants than has been previously described. We also supplement the findings in [17] and look at how the rising of the tongue is correlated with the duration of the segments. We will address the following questions:

- How does palatalization affect the position of the tongue of consonants and their preceding vowels? Based on previous research, it will be hypothesized that the position of the tongue will be higher, at least up until the midpoint with palatalization while articulating /s/, /t/, /n/ and retracted while articulating /l/. The tongue is expected to be also higher while producing the preceding vowels. We expect to find a constant movement throughout the vowel instead of a distinguishable rising movement at the end of the vowel.
- If we observe the raising of the tongue with palatalization, we want to know whether this raising can be related to the lengthening or shortening of the segmental durations. As previous studies have pointed out, we expect that the raising of the tongue is causing a longer duration of back vowels and /t/ and possibly /s/.

2. MATERIALS AND METHODS

The data for this study was collected in the Phonetics Laboratory at the University of Tartu, Estonia, from 21 native Estonian speakers (10 male, 11 female). The data consisted of 11 minimal pairs of meaningful words with a CVC: structure, where the target consonant was in the post-vocalic position. The test words were embedded in a medial position of carrier sentences, followed by a comma and a word starting with [mi], e.g. *Külmal hommikul oli maas hall* [hɑlː], *mida oli ilus vaadata.* 'On a cold morning, the land was covered with frost, that was beautiful to look at.' *Minu auto oli hall* [hɑlʲː], *mitte valge.* 'My car was gray, not white.'

The articulatory data were recorded with a Carstens AG501 electromagnetic articulograph, with a 200 Hz sampling rate. We analyzed the data from 3 sensors from the midsagittal plane of the tongue. The sensor on the anteo-dorsum and the sensor on the tongue blade were used to estimate the height of the tongue, and the sensor on the medio-dorsum was used to estimate the frontness or backness of the tongue. Statistical analysis was carried out with the R software [28]. The generalized Additive Mixed Model was used from mgcv package [29] to estimate the effect of palatalization on the height and the frontness or backness of vowels and consonants. The effect of tongue height, vowel, and consonant on the segmental duration values were tested with Linear Mixed Model from lme4 package [30]. In all models, a random intercept for the speaker was included. In the case of trajectory comparisons, post-hoc testing was done with multcomp [31] package using Bonferroni-Holm correction.

3. RESULTS

3.1. The position of the tongue

Figure 1 panels a and b show the data of the words with /l/. During the vowels preceding /l/, the anteo-dorsum was higher ($p < 0.001$) with palatalization. In the case of /ɑ/, also the tongue blade was higher ($p < 0.001$). The tongue was more anterior for both vowels with palatalization ($p < 0.001$).

During the palatalized /l/, the anteo-dorsum was higher in the context of both vowels ($p < 0.001$), the tongue blade was higher in the context of /ɑ/, and the tongue was more anterior ($p < 0.001$) compared to the non-palatalized /l/. The panels c and d of Figure 1 present the results of the words with /n/. The vowels preceding palatalized /n/ were produced with higher anteo-dorsum ($p < 0.001$), and the tongue was more anterior ($p < 0.001$) compared to the non-palatalized context. The tongue blade during /ɑ/ was lower in the palatalized context ($p < 0.001$), but no significant palatalization effect on tongue blade height occurred for /o/.

During palatalized /n/, the anteo-dorsum was higher ($p < 0.001$), and the tongue was more anterior ($p < 0.001$) compared to non-palatalized /n/. Palatalization had no significant effect on the height of the tongue blade. In the case of the vowels preceding palatalized /s/ (Figure 1 panels e and f), the anteo-dorsum was higher ($p < 0.001$), and the tongue was more anterior ($p < 0.001$) compared to the non-palatalized /s/ context. In the palatalized context, the tongue blade was lower for /u/ ($p < 0.001$) but not significantly different for /ɑ/.

The palatalized /s/ was produced with a higher anteo-dorsum and more anterior tongue position ($p < 0.001$). The tongue blade was lower in the context of /u/ ($p < 0.001$), but in the /ɑ/ context, palatalization did not affect the tongue blade during /s/. The words with /t/ are presented in Figure 1 panels g and h. During the vowels preceding palatalized /t/, the anteo-dorsum was higher ($p < 0.001$), and the tongue was more anterior ($p < 0.001$) compared to the non-palatalized context. The tongue blade was lower with palatalization in the case of /o/ ($p = 0.005$), but no palatalization effect occurred in the case of /u/. During palatalized /t/, the anteo-dorsum was higher ($p < 0.001$), the tongue blade was lower in the context of both vowels ($p < 0.001$), and the tongue was more anterior ($p < 0.001$) compared to non-palatalized /t/. In the case of /t/, palatalization did not affect the height of the tongue blade.

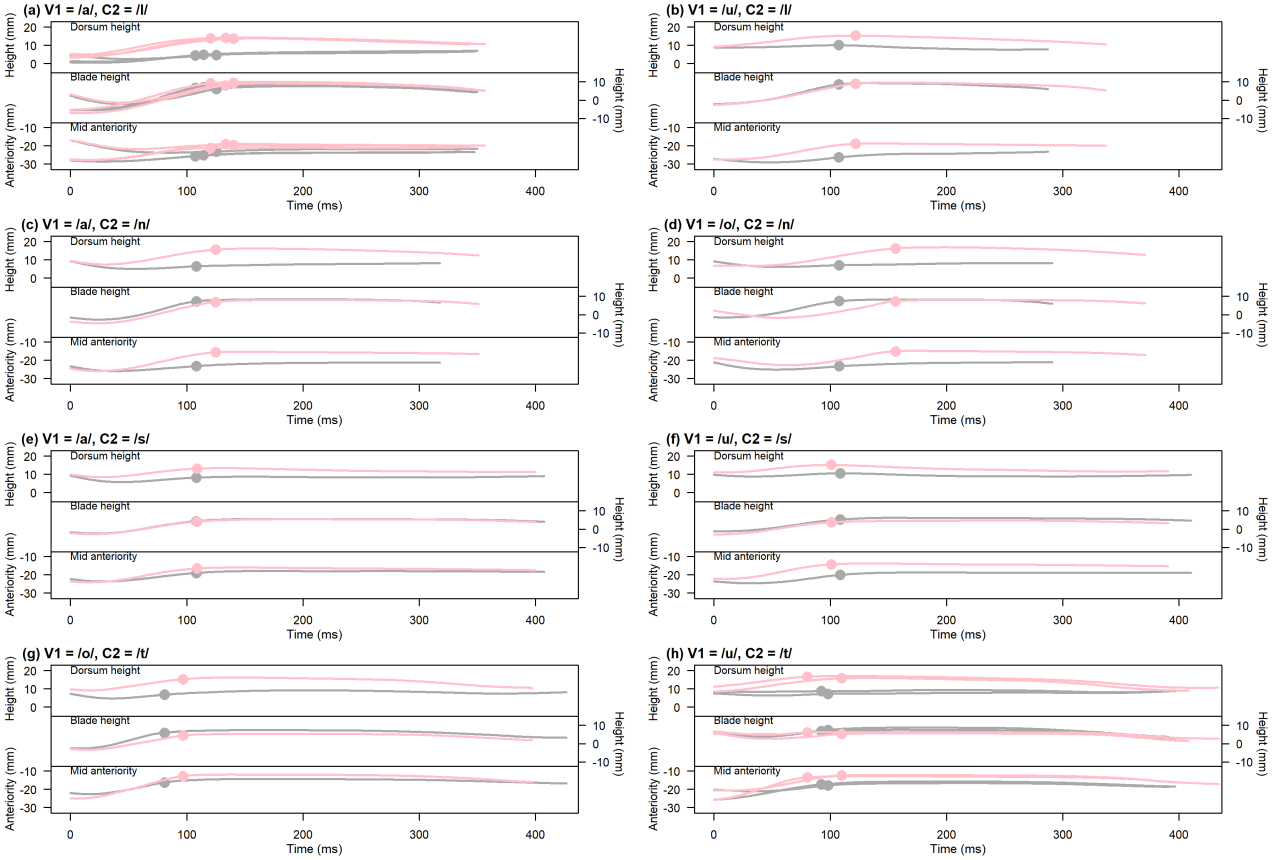


Figure 1: The height of the anteo-dorsum and the tongue blade, and the anteriority of the tongue in millimeters when producing the VC sequence of the words with palatalized (pink) and non-palatalized (gray) consonants. The dot marks the acoustic boundary between the vowel and the consonant. The panels are grouped by V1 and C2. The contours are averaged over the speakers.

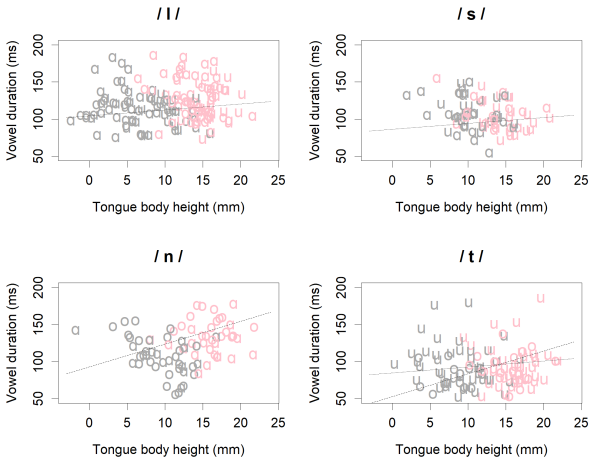


Figure 2: The duration of the vowels preceding non-palatalized (black) and palatalized (pink) consonants (in milliseconds) depending on the height of the tongue (in millimeters).

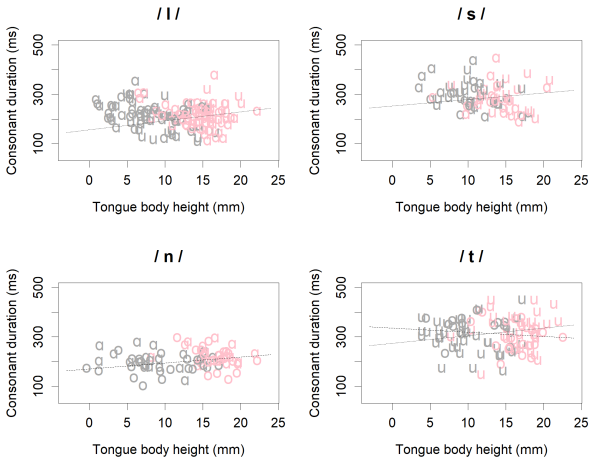


Figure 3: The duration of non-palatalized (black) and palatalized (red) consonants (in milliseconds) in the context of different vowels depending on the height of the tongue (in millimeters).

3.2. Effect of the tongue height on segmental durations

The effect of the tongue height on segmental durations was tested with vowel category and the following consonant using a linear mixed model. As there was a strong correlation between palatalization and tongue height, only the latter was chosen as an independent factor in the model. The data for the vowels are plotted in Figure 2.

The average duration of the vowel was 112 ms. The model showed that with the tongue rising, the vowel duration is longer (β = 1.3 ms/mm, p < 0.001). Compared to (the base value) /ɑ/, the duration of /u/ was not significantly different, but /o/ was shorter (β = -40 ms, p < 0.001). Compared to the (base value) /l/ context, the vowel duration was shorter before /n/ (β = -12 ms, p = 0.003), /s/ (β = -18 ms, p < 0.001) and /t/ (β = -20 ms, p < 0.001). An interaction in the model between the vowel and tongue body movement showed that (compared to /ɑ/) the duration of /o/ was even longer when the tongue was raised (β = 1.8 ms/mm, p = 0.007). Another significant interaction, however, showed that before /n/, the duration of /o/ was longer (β = 32 ms, p < 0.001), neutralizing the strong negative main effect of /o/.

Figure 3 presents the effect of the tongue height on the duration of consonants. The model showed no significant main effect of tongue height. In the case of (base values) /l/ following /ɑ/, the duration was 230 ms, but it was significantly shorter following an /u/ (β = -74 ms, p <0.001). Consonant /n/ was shorter than /l/ (β = -48 ms, p <0.012). Consonants /s/ (β = 96 ms, p < 0.001) and /t/ (β = 117 ms, p <0.001), on the other hand, were longer. Additionally, the model showed two interactions with the tongue height: when following an /u/ and the tongue, the consonant was longer (β = 4.6 ms/mm, p < 0.001), and if the consonant was /n/, and the tongue was higher the duration was also longer (β = 1.4 ms/mm, p = 0.012).

4. DISCUSSION

Based on [18–21], we hypothesized that the place of articulation of consonants would be higher for /s/, /t/, /n/, /l/ and retracted for /t/. The results showed that the tongue was always higher and more fronted while articulating palatalized consonants. Our study did not confirm that the place of articulation of /t/ is retracted with palatalization. Due to the sensors, the tongue may not have enough freedom to move, and thus we did not capture the movements that would occur in more natural circumstances. The results also show that contrary to the claim by [32], the tongue was raised throughout the production of consonants with palatalization, not only in the beginning.

We confirmed the hypothesis based on studies [2, 3, 5–9, 24] that the place of articulation of the vowel preceding palatalized consonants is also higher. Our results showed a constant movement throughout the vowel where the tongue was higher than in the non-palatalized counterpart. In the speech, the following segments are anticipated and adjusted for [10–13]. As palatalized consonant demands a higher secondary place of articulation, the tongue is already raised anticipatorily from the beginning of the vowel.

Based on previous research [3,9,24], we expected that the durations of the vowels and consonants /t/ and /s/ are longer. We found that with the tongue raising, the duration of the vowels that preceded palatalized consonants was lengthened. This lengthening of the vowel is used to enhance the perception of palatalization and to give the listener time to process upcoming information [25]. For consonants, we could not find that the tongue raising affected the duration of /s/ or /t/. The duration of /l/ and /n/ were shorter with palatalization. Fricatives and plosives might be harder to produce with sensors on the tongue, and a temporal compromise is made in the vowels.

5. CONCLUSIONS

This paper studied the effect of palatalization on the articulation and duration of the Estonian /s/, /t/, /n/, /l/, and the preceding vowels using an electromagnetic articulograph (EMA). The results showed that the back of the tongue was always higher, and the tongue was more fronted when articulating palatalized consonants. This position was reached already during the preceding vowels. The height of the tongue blade was not affected by palatalization. With the raising of the tongue, the duration of the vowels that preceded palatalized consonants were longer. The duration of /s/ and /t/ were not affected by the tongue raising, but /l/ and /n/ were shorter.

6. ACKNOWLEDGEMENTS

The paper is dedicated to the 105th anniversary of the Republic of Estonia.

7. REFERENCES

[1] N. Bateman, "A crosslinguistic investigation of Palatalization," Ph.D. dissertation, University of California, San Diego, 2007.

[2] M. Ní Chiosáin and J. Padgett, "An acoustic and perceptual study of Connemara Irish palatalization," *Journal of the International Phonetic Association*, vol. 42, no. 02, pp. 171–191, 2012.

[3] I. Lehiste, "Palatalization in Estonian: Some acoustic observations," in *Estonian Poetry and Language: Studies in honor of Ants Oras*, V. K/ oresaar and A. Rannit, Eds., vol. 136-162. Stockholm: Vaba Eesti, 1965.

[4] G. Liiv, "Some experiments on the effect of vowel-consonant transitions upon the perception of palatalization in Estonian," *Soviet Fenno-Ugric Studies*, vol. 1, pp. 33–36, 1965.

[5] M. Derkach, G. Fant, and A. de Serpa-Leitao, "Phoneme coarticulation in Russian hard and soft VCV-utterances with voiceless fricatives," *STL-QPSR*, vol. 11, no. 2-3, pp. 1–7, 1970.

[6] S. M. Howie, "Formant transitions of Russian palatalized and nonpalatalized syllables," *IULC Working Papers*, vol. 1, no. 1, 2001.

[7] M. M. Vihman, "Palatalization in Russian and Estonian," *Project on Linguistic Analysis Reports*, vol. 1, pp. V1–V32, 1967.

[8] H. Kim, "A Phonetically Based Account of Phonological Stop Assibilation," *Phonology*, vol. 18, no. 1, pp. 81–108, 2011.

[9] P. Teras and K. Pajusalu, "Palatalization and prepalatalization in Estonian spontaneous speech," *Keel ja kirjandus*, vol. 57, no. 4, pp. 257–269, 2014.

[10] B. E. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, vol. 35, pp. 1773–1781, 1963.

[11] E. Farnetani and D. Recasens, "Coarticulation and connected speech processes," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2013, pp. 316–351.

[12] L. Lisker, "Segment duration, voicing and the syllable," in *Syllables and segments*, A. Bell and J. B. Hooper, Eds. Amsterdam: North-Holland, 1978, pp. 133–142.

[13] R. Hammarberg, "The metaphysics of coarticulation," *Journal of Phonetics*, vol. 4, pp. 353–363, 1976.

[14] S. E. G. Öhman, "Coarticulation in VCV utterances: spectrographic measurements." *JASA*, vol. 39, no. September, pp. 151–168, 1966.

[15] Y. Xu, "Syllable is a synchronization mechanism that makes human speech possible," *PsyArXiv*, 2020.

[16] S. Hamann and H. Avelino, "An acoustic study of plain and palatalized sibilants in Ocotepec Mixe," in *ICPhS 16*, J. Touvain and W. J. Barry, Eds. Saarbrücken: Saarland University, 2007, pp. 949–952.

[17] A. Malmi and P. Lippus, "Keele asend eesti palatalisatsioonis," *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, vol. 10, no. 1, pp. 105–128, 2019.

[18] L. Kutser, "Eesti konsonantide palatogrammist," 1935.

[19] P. Ariste, *Katselisfoneetilisi tähelepanekuid : mit einem Referat: Experimentalphonetische Beobachtungen*. Tartu: Acta Universitatis Tartuensis (Dorpatensis), 1943.

[20] M. Żygis and B. Pompino-Marschall, "The articulation of secondarily palatalized coronals in Polish," in *ICPhS Proceedings*, San Fransisco, 1999, pp. 1987–1900.

[21] A. Kochetov, *Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization*, new york ed. Routledge, 2002.

[22] E. Meister and S. Werner, "Comparing palatography patterns of Estonian consonants across time," in *ICPhS 18*. Glasgow, Scotland: University of Glasgow, 2015, pp. 1–5.

[23] A. Eek, "Articulation of the estonian sonorant consonants. III. Palatalized [n] and [l]," *Eesti NSV Teaduste Akadeemia Toimetised. Ühiskonnateadused*, vol. 20, no. 2, pp. 173–191, 1971.

[24] G. Liiv, "Preliminary remarks on the acoustic cues for palatalization in Estonian," *Phonetics*, vol. 13, pp. 59–64, 1965.

[25] M. Ordin, "Palatalization and temporal organisation of CVC clusters in Russian," *Russian Linguistics*, vol. 34, no. 1, pp. 57–65, 2010.

[26] K. Bolla, *A conspectus of Russian speech sounds*. Cologne: Bölau Verlag, 1981.

[27] A. Kochetov and M. Radisic, "Latent consonant harmony in Russian: Experimental evidence for Agreement by Correspondence," in *Proceedings of the 17th Formal Approaches to Slavic Linguistics (FASL) meeting*, M. Babyonyshev, D. Kavitskaya, and J. Reich, Eds. Ann Arbor, MI: Jindřich Tomanâs Michigan Slavic Publications, 2009, pp. 111–130.

[28] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2021. [Online]. Available: https://www.r-project.org/

[29] S. N. Wood, *Generalized additive models. An introduction with R*, 2nd ed. New York: Chapman and Hall/CRC, 2017.

[30] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[31] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous Inference in General Parametric Models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008.

[32] A. Eek, "Observations in Estonian palatalization: an articulatory study," *Estonian Papers in Phonetics*, vol. 2, pp. 9–17, 1973.

LATERAL AND CENTRAL TONGUE BRACING COORDINATION IN TYPICAL SPEAKERS AND COCHLEAR IMPLANT USERS

Marko Liker

Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
mliker@ffzg.hr

ABSTRACT

This paper addresses two research questions: 1. how are lateral and central bracing coordinated in typical Croatian speakers as opposed to prelingually deaf cochlear implant (CI) users, and 2. how are bracing patterns related to hearing status and intelligibility in CI users.

Electropalatographic (EPG) and acoustic recordings of eight typical adult speakers and three CI users with different hearing status and speech intelligibility scores were included in this investigation. They were recorded in a communicative situation which facilitated spontaneous speech. Bracing patterns were analysed during Croatian words /ʃlapama/, which were repeated four times by each speaker. Central and lateral bracing were under competing demands during ʃl-sequences, while subsequent non-lingual sequences did not constrain the tongue directly.

The analysis showed that the coordination and coarticulation between central and lateral bracing differed between typical and atypical speakers. CI users with different hearing status and speech intelligibility scores differed in their bracing patterns.

Keywords: tongue bracing, cochlear implant (CI), coarticulation, electropalatography (EPG), Croatian.

1. INTRODUCTION

Tongue bracing has been described as active mechanical support of the tongue whereby the tongue is in contact with a rigid vocal tract surface during speech [6]. Initial research of this phenomenon focused on lateral bracing in a specific set of sounds, mostly from English [5, 11], while other languages, such as Argentinian and Cuban Spanish [9], Croatian [12] and German [4], received less attention in this respect.

Subsequent research showed that human tongues seek to be actively braced all the time during speech [2, 6] and that bracing could potentially constitute one of the universal postures during speech [14]. These investigations emphasized that bracing was not only a characteristic of one specific set of speech sounds, but rather that it was a pervasive and active strategy during speech. They also showed that bracing could be lateral and central - default bracing posture should be lateral and if lateral bracing was blocked (e.g. during lateral sound productions) or hard to achieve (e.g. during open vowels) then central bracing would take over.

There are several competing and complementary explanations of the purpose of such tongue movement strategy during speech: it is important for supporting the tongue during complex sound productions [18], for deceleration of the tongue during stop productions [4], for establishing proper aerodynamic and acoustic conditions [8] and for somatosensory feedback [6, 17].

Instrumental kinematic techniques such as imaging and palatographic techniques are often used in investigations of the bracing patterns. EPG seemed suitable for the present investigation, because EPG was designed to capture the timing of tongue-to-palate contact patterns during speech [19].

While previous research has recently focused on bracing as pervasive and active strategy in typical speech [6, 14, 20], the pervasiveness and the coordination of bracing patterns in atypical speech has remained largely uninvestigated. Since it is hypothesized that bracing is important for the achievement of acoustic goals and for somatosensory feedback [6, 8, 17, 18], bracing in the speech of persons with hearing impairments is of particular interest. Prelingually deaf CI users require special attention in this respect, because previously published studies showed that their coarticulation and coordination strategies are more atypical when compared with their postlingually deaf counterparts [7, 16]. Therefore, this paper addresses the following research questions: 1. how are lateral and central bracing coordinated in typical Croatian speakers as opposed to their prelingually deaf cochlear implant counterparts, and 2. how are bracing patterns related to hearing status and intelligibility in atypical speakers. EPG recordings of quasi-spontaneous speech are used to test the coordination and coarticulation between lateral and central bracing in words with ʃl-sequences followed by non-lingual sound sequences. Quasi-spontaneous speech provides an opportunity to analyse bracing patterns in a more

natural communicative situation than reading sentence lists. Words with ʃl-sequences offer the opportunity to analyse coordination between lateral (constriction gesture in postalveolar fricative /ʃ/) and central bracing (apical gesture in alveolar lateral approximant /l/) when they are under competing demands. And finally, non-lingual sound sequence (/apama/ in word /ʃlapama/) provides the opportunity to investigate the pervasiveness of bracing in sequences where tongue is not the primary articulator.

2. METHOD

2.1. Participants and speech material

Speech material was obtained from the CROCO corpus [13] and two groups of speakers were included in this study: typical and atypical group. The group of typical participants comprised eight speakers (four female: T1, T2, T7, T8; four male: T3, T4, T5, T6) of Standard Croatian with no speech and hearing impairments, ranging in age between 21 and 24 (mean 22.4). Their speech was rated by five trained phoneticians and those eight speakers received the best scores for speech sound production and overall speech intelligibility among 105 candidates (average score was 6.3 on a scale from 1 to 7). In the second group of participants, three adult prelingually deaf CI users were included. These participants were selected for their different clinical characteristics (table 1) among seven prelingually deaf CI users. They had no other diagnosed impairments.

CI user	A1	A2	A3
Age	24	18	28
Age @ implant.	2	5.3	10
Age @ therapy	2	5.7	3.6
Speech audiogram	90% WR @ 45dB	70% WR @ 65dB	60% WR @ 50dB
N of speech errors	0	22	30
Intelligibility score [3]	8.5/10	5.9/10	3.2/10

Table 1: Clinical data for the three atypical CI speakers (A1, A2, A3). Speech errors were analysed via broad phonetic transcription of the whole recording session by two trained phoneticians.

Speech material was obtained through a dialogue situation, whereby each speaker was asked to describe the path through a maze and read signs at nine check-points marked throughout the path (Fig 1). Participants had to explain to the experimenter where the line was drawn on the map, because the experimenter had the same map, but the line marking the path was not shown. The goal was to explain to the experimenter the position of the line relative to the check-points. Each sign contained a two-syllable or three-syllable word and among these words was the target word for the present investigation ("šlapama" /'ʃlapama/, meaning "with slippers"). Each participant repeated the target word four times.

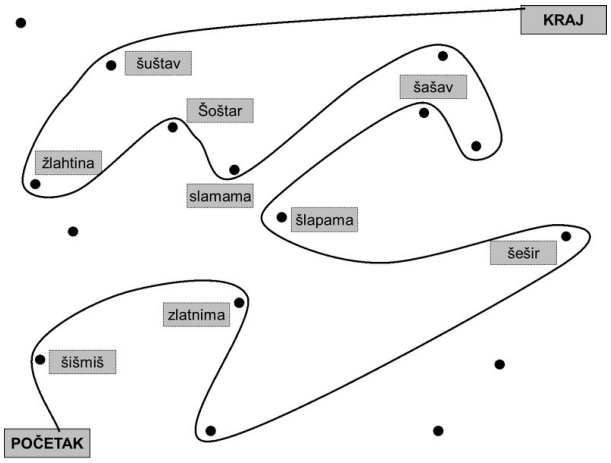


Figure 1: An illustration of the map description exercise for the elicitation of the quasi-spontaneous speech in this investigation.

2.2. Instrumentation and data preparation

EPG and acoustic data were recorded simultaneously using WinEPG system and Articulate Assistant software [1]. EPG data was collected at 100 Hz sampling frequency, while the sampling frequency for the acoustic signal was 44100 Hz. Reading style EPG palate called Mcyam plate [10] was used by all speakers.

Annotation was performed according to well established acoustic criteria. The beginning of the postalveolar fricative /ʃ/ was marked when high frequency noise started. The end of the fricative and the beginning of the dentoalveolar lateral approximant /l/ was marked when high frequency noise stopped and voicing of the approximant started. The end of the approximant and the beginning of non-lingual sequence /apama/ was annotated when the frequency and intensity discontinuity in the harmonic sound was observed.

2.3. EPG indices and data analysis

Central and lateral bracing were analysed at mid-point of each annotated region and also throughout every annotated region at a predetermined number of equally spaced sample points. The analyses were performed at mid-point in order to analyse overall bracing pattern at its least coarticulated point for each annotated region. The analyses were also done during the whole annotated region in order to analyse the bracing patterns dynamics. In order to average productions of different durations and to visualize

them, the duration of each annotated region was normalised via equally spaced sample points. The number of sample points (NSP) was determined according to the following equation:

(1) $$NSP = \frac{t\_min}{10}$$

where *t_min* was the duration of the shortest annotation regions in milliseconds (/ʃ/=119ms in typical speakers and 107ms in CI users; /l/=37ms; /apama/=332ms in typical speakers and 404ms in CI users) and 10 represented the number of milliseconds between successive EPG frames, determined by the EPG sampling frequency (100 Hz).

Central bracing was quantified by measuring the dentoalveolar area weighted index (the number of activated electrodes in the anterior two rows of electrodes divided by the total number of electrodes in dentoalveolar region). Lateral bracing was quantified by measuring the lateral contact area weighted index (the number of activated electrodes in the two outermost columns of electrodes at each side of the palate divided by the total number of electrodes in that region) (Fig 2). The significance of the difference between central and lateral bracing, as well as between speakers was tested via Repeated Measures ANOVA.

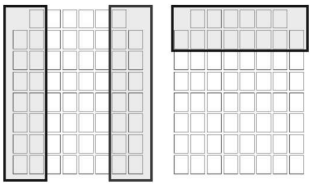


Figure 2: An illustration of the EPG palate areas for the calculation of lateral (left) and central (right) bracing.

3. RESULTS

The results of the central and lateral bracing at mid-point of each annotated region (/ʃ/, /l/, /apama/) show that the coordination of lateral and central bracing is different in typical speakers when compared with atypical speakers (Fig 3). In typical speakers lateral bracing is stronger than central bracing during /ʃ/, while the opposite can be seen during /l/ and finally during the non-lingual sequence /apama/ central bracing equals zero and lateral bracing is equal to the level it showed during /l/. Differences between lateral and central bracing in typical speakers are statistically significant (F(1, 7)=6.927, p=0.03), with the significant interaction between the annotated region and type of bracing (F(2, 6)=38.688, p=0.00).

In atypical speakers the coordination between lateral and central bracing is different from the coordination pattern observed in typical speakers, whereby lateral bracing is higher than central bracing in all annotated regions. The switch between predominantly lateral bracing in /ʃ/ and predominantly central bracing in /l/, which can be observed in typical speakers, is absent in CI users. The difference between lateral and central bracing is not statistically significant in atypical speakers (F(1, 2)=1.757, p=0.32) and there are no significant interactions between the annotated region and bracing type (F(2, 1)=0.295, p=0.79).

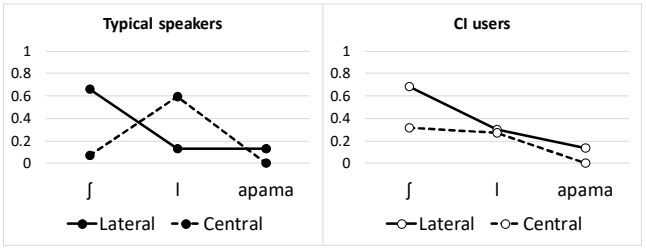


Figure 3: Lateral and central bracing in three annotated regions (/ʃ/, /l/, /apama/) in typical speakers (left chart) and CI users (right chart).

Figures 4 and 5 show central and lateral bracing coordination during normalized durations of each annotated region for each speaker. The results for the typical speakers (Fig 4) show that lateral bracing starts to decrease, while central bracing starts to increase at the end of /ʃ/ or during /l/. Central bracing peaks during or immediately after /l/, while lateral bracing remains low. During non-lingual sequence there are noticeable differences between typical speakers, whereby some speakers show high amount of lateral bracing (e.g. T1, T7) and some speakers have very low lateral bracing (e.g. S3, S5), while some appear to have their tongues completely unbraced during initial portion of the non-lateral sequence (e.g. S4, S6).

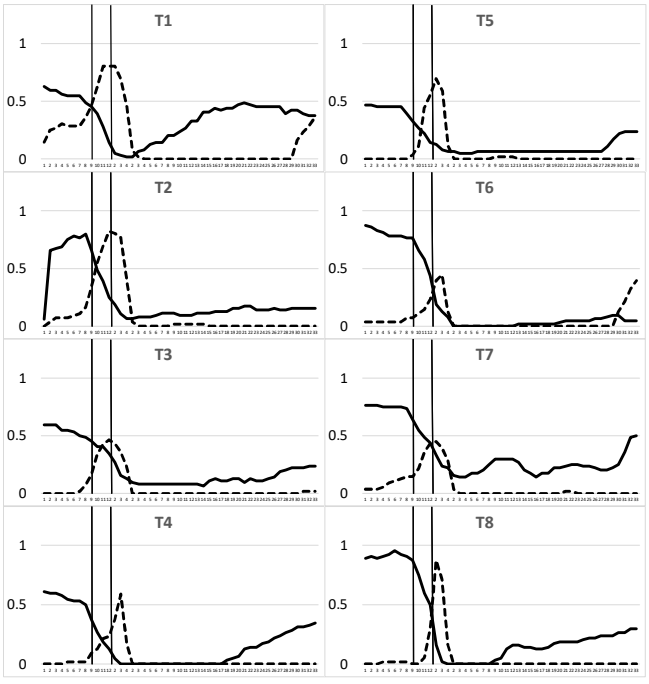


Figure 4: Lateral (solid line) and central (dashed line) bracing during three annotated regions (/ʃ/, /l/, /apama/) at equally spaced sample points in typical speakers (T1-T8). Vertical lines mark annotation boundaries.

The results for the atypical speakers (Fig 5) reveal differences in bracing patterns between the three CI users. Speaker A1 shows patterns which are comparable to those observed in typical speakers, while the other two atypical speakers have noticeably different patterns. In speakers A2 and A3 there is no typical switch between predominantly central and predominantly lateral bracing at the point of transition between the fricative and the lateral approximant. Furthermore, A3 shows very low amount of contact in both lateral and central zone, with almost no difference in lateral pattern throughout the whole word.

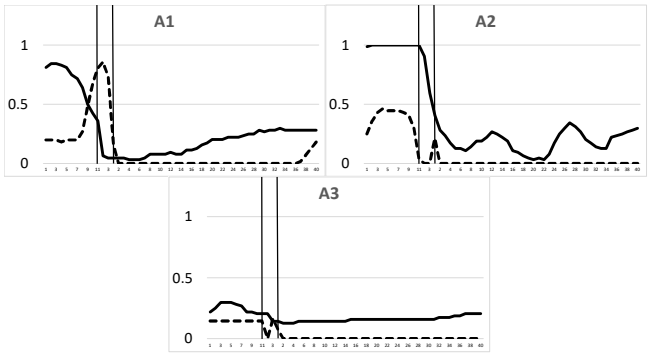


Figure 5: Lateral (solid line) and central (dashed line) bracing during three annotated regions (/ʃ/, /l/, /apama/) at equally spaced sample points in CI users (A1-A3). Vertical lines mark annotation boundaries.

4. DISCUSSION

The results of the present study show that the coordination between central and lateral bracing differs between typical speakers and CI users, whereby typical speakers have significantly different amount of central versus lateral bracing, while this difference is very low and non-significant in CI users. The results also reveal differences in bracing pattern dynamics among CI users and those differences coincide with their different clinical data (table 1).

The results of the analysis of typical speakers are consistent with previously published findings [5, 6, 14]. The analysis shows that typical speakers have their tongues braced throughout the whole target word and that they share similar coordination pattern for central and lateral bracing. Importantly, this investigation shows that when lateral and central bracing are under competing demands, a consistent bracing coordination pattern emerges. At the point of transition between /ʃ/ and /l/, lateral bracing starts to decrease, while central bracing increases. The timing of the switch between lateral and central bracing maxima differs somewhat between speakers. It occurs either at the acoustic boundary between /ʃ/ and /l/ (T1), or during /l/ (T2, T3, T4, T5, T7) or at the acoustic boundary between /l/ and /apama/ (T6, T8). This is consistent with the DAC model of coarticulation [15], which predicts similarly high coarticulation resistance in anterior lingual fricatives and lateral sounds (with fricative gesture showing more resistance in most speakers analysed here). This finding together with the finding that lateral bracing continues even during the non-lingual sequence confirms previous findings from other languages that lateral bracing is primary tongue-bracing strategy [6]. It should be noted that T4 and T6 have their tongues completely unbraced during initial phases of the non-lingual sequence, which does not support the claim that the tongue is always braced. One possible reason for this lack of bracing might be that lateral bracing is not the default form of tongue-bracing for these speakers and that bracing occurs elsewhere during those sections of /apama/ [6].

The analysis of bracing pattern dynamics in CI users builds upon previously documented results and also shows some novel findings. All three CI users show constant and uninterrupted bracing during the whole word. However, the three CI users differ in the coordination of central and lateral bracing and this difference is reflected in their clinical data (see table 1). A3 has highest age at implantation, poorest results of the speech audiogram, lowest intelligibility score and highest number of speech sound errors. She shows very low amount of overall bracing and very little change in the bracing pattern during the transition between the fricative and the lateral approximant. A2 has lower age at implantation and better overall hearing and intelligibility results. Her EPG results show that the overall amount of bracing is comparable to the amount of bracing in typical speakers, but similarly to A3, there is no switch between lateral and central bracing maxima. A1 has earliest age at implantation, best speech audiogram results, highest intelligibility score and no speech sound errors. Her bracing patters are almost identical to those in typical speakers (e.g. T1, T2). Since the coordination between lateral and central bracing might be related to speech intelligibility, these results have important potential implications for speech therapy and should be investigated further.

The limitations of this study should be kept in mind. The number of speakers included in this study is relatively low and some aspects of tongue bracing cannot be captured via EPG [6].

5. ACKNOWLEDGEMENTS

This investigation was funded by the Croatian Science Foundation (the CROCO project, grant number: IP-2016-065367). The help from Vladimir Prašin and the SUVAG Polyclinic in Zagreb is gratefully acknowledged.

6. REFERENCES

[1] Articulate Instruments. (2008). *Articulate Assistant User Guide, ver. 1.17*. Musselburgh: Queen Margaret University.

[2] Bressmann, T., Flowers, H., Wong, W., Irish, J. C. (2010). Coronal view ultrasound imaging of movement in different segments of the tongue during paced recital: Findings from four normal speakers and a speaker with partial glossectomy. *Clinical Linguistics & Phonetics* 24(8), 589–601.

[3] Dobrić, A., Mildner, V., Liker, M. (2019). Identifikacija slušnog statusa govornika na temelju spontanog govora (Identification of the hearing status based on spontaneous speech). In: Liker, M., Vlašć Duić, J. (eds), *Book of Abstracts of the 10th scientific conference with international participation Speech research 2019*, HFD: Zagreb, 34-35.

[4] Fuchs, S., Perrier, P., Geng, C., Mooshammer, C. (2006). What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. In: Harrington, J., Tabain, M. (eds), *Speech production: Models, phonetics processes, and techniques*. New York: Psychology Press, 149-164.

[5] Gibbon, F. E., Yuen, I., Lee, A., & Adams, L. (2007). Normal adult speakers' tongue palate contact patterns for oral and nasal stops. *Advances in Speech-Language Pathology* 9, 82–89.

[6] Gick, B., Allen, B., Roewer-Després, F., Stavness, I. (2017). Speaking tongues are actively braced. *Journal of Speech, Language, and Hearing Research* 60(3), 494-506.

[7] Hardcastle, W., Tjaden, K. (2008). Coarticulation and speech impairment. In: Ball, M.J., Perkins, M.R., Müller, N., Howard, S. (eds), *Handbook of clinical linguistics*. Blackwell publishing, 506-520

[8] Honda, K., Takano, S., Takemoto, H. (2010). Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory. *Journal of Phonetics* 38, 33-43.

[9] Kochetov, A., Colantoni, L. (2011). Coronal place contrasts in Argentine and Cuban Spanish: An electropalatographic study. *Journal of the International Phonetic Association* 41(3), 313–342.

[10] Lee, A., Fujiwara, Y., Liker, M., Yamamoto, I., Takei, Y., Gibbon, F. (2022). Electropalatography (EPG) activities in Japan and the impact of the COVID-19 pandemic on EPG research and therapy: A report of presentations at the 7th EPG Symposium. *International Journal of Language & Communication Disorders* 57(4), 906-917.

[11] Lee, A., Gibbon, F E., Oebels, J. (2015). Lateral bracing of the tongue during the onset phase of alveolar stops: An EPG study, *Clinical Linguistics & Phonetics* 29:3, 236-245.

[12] Liker, M., Gibbon, F.E. (2015). Place of articulation of anterior nasal versus oral stops in Croatian. *Journal of the International Phonetic Association* 45 (1), 35-54.

[13] Liker, M., Vidović Zorić, A., Zharkova, N., Gibbon, F. (2019): Ultrasound analysis of postalveolar and palatal affricates in Croatian: a case of neutralisation, In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proceedings of the 19th International Congress of Phonetic Sciences.* Melbourne, Australia 2019, 3666-3670.

[14] Liu, Y., Tong, F., de Boer, G., Gick, B. (2022). Lateral tongue bracing as a universal postural basis for speech. *Journal of the International Phonetic Association*, 1-16.

[15] Recasens D., Pallarés, M. D., Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America* 102, 544-561.

[16] Sfakianaki, A., Nicolaidis, K., Okalidou, A., Vlahavas, G. (2018). Coarticulatory dynamics in Greek disyllables produced by young adults with and without hearing loss. *Clinical Linguistics & Phonetics* 32 (12), 1162-1184.

[17] Stevens, K. N., Perkell, J. S. (1977). Speech physiology and phonetic features. In: Sawashima, M. S., Cooper, F. (eds), *Dynamic aspect of speech production*. Tokyo: University of Tokyo Press, 323-341.

[18] Stone, M. (1991). Toward a model of three-dimensional tongue movement. *Journal of Phonetics* 19, 309-320.

[19] Stone, M. (2010). Laboratory techniques for investigating speech articulation. In: Hardcastle, W. J., Laver, J., Gibbon, F.E. (eds), *Handbook of phonetic sciences*. Blackwell Publishing Ltd., 7-38.

[20] Ying, J., Shaw, J.A., Carignan, C., Proctor, M., Derrick, D., Best, C.T. (2021). Evidence for active control of tongue lateralization in Australian English /l/. *Journal of Phonetics* 86, 101039, 1-24.

COARTICULATORY RESISTANCE AND AGGRESSION IN PITCH-ACCENTED VOWELS UNDER THE EFFECT OF VOWEL-TO-VOWEL COARTICULATION

Andrea Deme[1,2], Márton Bartók[2], Tamás Gábor Csapó[4,2], Tekla Etelka Gráczi[4,2], Kornélia Juhász[1,3,2], Alexandra Markó[2]

[1]Eötvös Loránd University, Hungary; [2]MTA-ELTE "Lendület" Lingual Articulation Research Group, Hungary; [3]Hungarian Research Centre for Linguistics, Hungary; [4]Budapest University of Technology and Economics
deme.andrea@btk.elte.hu, bartokmarton@gmail.com, csapot@tmit.bme.hu, graczi.tekla.etelka@nytud.hu, juhasz.kornelia@nytud.hu, marko.alexandra.phd@gmail.com

ABSTRACT

In the present study we tested if pitch accent increased coarticulatory resistance and aggression in vowels under the influence of vowel-to-vowel coarticulation, using both acoustic and articulatory measures. We studied F_2 and horizontal tongue positions (EMA) in /i/ and /u/ produced by 9 Hungarian females in non-words. Target vowels occurred either in neutral (e.g., first syllable /i/ in /pipipipi/) or in coarticulating (e.g., first syllable /i/ in /pipupupu/) contexts. We analysed differences of coarticulated and non-coarticulated tokens (quality shift), as well as dispersion of tokens across contexts (production range).

Results showed that pitch accent did not increase coarticulatory resistance and aggression in the tested vowels, which contradicts previous findings showing increased resistance in English and Hungarian. We assume that these results may be explained by language specific effects.

Keywords: vowel-to-vowel coarticulation, EMA, coarticulatory aggression, coarticulatory resistance

1. INTRODUCTION

Vowels interact in running speech even through an intervening consonant; this is termed vowel-to-vowel coarticulation [1, 2]. This interaction may be influenced by several factors. For instance, prosodic prominence is claimed to bring about segmental strengthening that leads to increased coarticulatory resistance and aggression [3, 4, 5]. As a result, vowels in accented syllables are assumed to resist coarticulatory effects more efficiently, and to exert a stronger influence on their (close or even transconsonantal) neighbours. Previous results on the effect of prominence on vowel-to-vowel coarticulation are, in part, inconclusive. (These were obtained for non-words in English, and either for acoustics or for articulation, but not for both at the same time.) These results showed increased resistance in accented syllables in acoustics [3], and in articulation [6], but regarding aggression, they did not confirm the hypothesis that increased coarticulatory resistance also leads to increase coarticulatory aggression in vowels, at least not in articulation [6].

In Hungarian, previous studies tested the effect of pitch accent on coarticulatory resistance against vowel-to-vowel effects in real words, both in acoustic and articulatory data which were gathered and analysed in parallel [7, 8]. These found that pitch accent increased coarticulatory aggression against vowel-to-vowel coarticulation in target vowels. However, the effect of pitch accent on coarticulatory aggression was not studied so far, as due to fixed first syllable stress/accent in Hungarian [9], it was not feasible to test in real words.

Following up on previous research, in the present study, we addressed the question if pitch-accent increased coarticulatory resistance and aggression in vowels against the effect of a transconsonantal vowel in nonsense words in Hungarian, and if resistance and aggression could be detected either in articulation, or in acoustics, or in both domains at the same time.

2. METHODS

We analysed simultaneously recorded acoustic and articulatory data of 9 adult female speakers (aged 25,2±5,9 years) producing /i/ and /u/ in /ˈpVpVpVpV/ shaped nonsense words (similar to e.g., Cho 2004, where data from CVCVC shaped non-words were used). Target vowels were embedded either in neutral (e.g., /ˈ**pi**p<u>i</u>pipi/) or in coarticulating (e.g., /ˈ**pi**p<u>u</u>pupu/) contexts (where bold denotes target vowels, underlining denotes trigger vowels, target vowels occur in pitch-accented syllables, and are under the effect of anticipatory coarticulation). Speakers were asked to produce the tested nonsense words as individual utterances (i.e., like "sentences"). Therefore, we could analyse vowels in pitch-accented (first) and unaccented (second and third) syllables. Note that prominence production was confirmed on the basis of auditory perception of the authors: all tested occurrences were perceived as bearing prominence on the first syllable of the test sequences.

Since /i/ and /u/ differ mainly on the dimensions of F_2 and horizontal tongue position, F_2 frequency and horizontal tongue positions were obtained from audio and Electromagnetic Articulography (EMA) recordings. Horizontal tongue positions were represented by *x*-axis displacement of the mean of two tongue body sensors on the sagittal plane. These sensors were placed on the backmost point on the speakers' tongue, and approx. 1 cm before that.

From these articulatory and acoustic data, we calculated and tested two distinct measures. First, we calculated differences of coarticulated and non-coarticulated/neutrally positioned tokens in the acoustic and articulatory spaces (Fig. 1, left), i.e., *distances* or *quality shift*, similarly to [3, 6, 7, 8]. For this purpose, we calculated distances of these tokens in F_2 in acoustics, and in horizontal tongue displacement measured on the *x*-axis in articulation. Second, we analysed *dispersion* (or *variation*) of vowel tokens *across contexts*, i.e., *production range* as in [7, 8] (Fig. 1, right). This means that we calculated relative standard deviation of F_2, and tongue position data pooled across vowel contexts (i.e., across neutral and coarticulating contexts), in the two production domains, respectively.

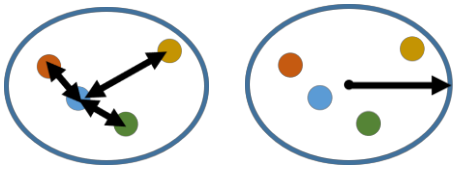


Figure 1: Derived and tested measures: distances or quality shift (left), and across context dispersion or production range (right).

Due to fixed first syllable stress/accent in Hungarian, it was not possible to test coarticulatory resistance and aggression in identical conditions. Coarticulatory resistance, i.e., the effect of pitch accent on target vowels, was tested using target vowels occurring in the first versus the second syllable, and in anticipatory coarticulation using unaccented trigger vowels exclusively (Table 1). Coarticulatory aggression was tested using target vowels under the effect of carryover coarticulation and occurring in unaccented second versus third syllables (Table 2). Note, however, that in this latter case, target vowel data are used for the purposes of drawing conclusions regarding trigger vowels indirectly, and these trigger vowels were also positioned in the first and second syllables. Therefore, all of the presented data reflect tendencies in vowels produced in the first (accented) versus the second (unaccented) syllable, in the same non-words, and this lets us to test if resistance and aggression occur at the same time.

Data were analysed using mixed models in R [10] and the lme4 [11], lmerTest [12], and lsmeans [13] packages. All models included *vowel quality* as a fixed factor, and they also included *target vowel accent* when we tested resistance, and *trigger vowel accent* when we tested aggression. Random intercepts and slopes for speakers and factors were also included in each case if they improved model fit. We report *p* values obtained using the Satterthwaite approximation of the degrees of freedom.

		Trigger /i/ (unacc.)	**Trigger /u/ (unacc.)**
Target /i/	*unacc.*	*pi**p**ipipi* (N)	*pi**p**ipupu* (C)
	acc.	*pi**p**ipipi* (N)	*pi**p**upupu* (C)
Target /u/	*unacc.*	*pup**u**pipi* (C)	*pup**u**pupu* (N)
	acc.	*p**u**pipipi* (C)	*p**u**pupupu* (N)

Table 1: Material used to test coarticulatory resistance in anticipatory coarticulation (bold: target vowel; underline: trigger vowel; C = coarticulating context; N = neutral context)

	Trigger /i/		**Trigger /u/**	
	acc.	*unacc.*	*acc.*	*unacc.*
Target /i/ (unacc.)	*pi**p**ipipi* (N)	*pi**p**ipipi* (N)	*pu**p**ipipi* (C)	*pupu**p**ipi* (C)
Target /u/ (unacc.)	*pi**p**upupu* (C)	*pi**p**upupu* (C)	*pu**p**upupu* (N)	*pupu**p**upu* (N)

Table 2: Material used to test coarticulatory aggression in carryover coarticulation (bold: target vowel; underline: trigger vowel; C = coarticulating context; N = neutral context)

3. RESULTS

Since the main question of the study is if vowels exhibited increased coarticulatory resistance and aggression at the same time (i.e., if vowels in accented position show both), we present data on the two tested measures in pairs: figures on the left show resistance, while figures of the right show aggression of the same vowels (i.e., same vowel qualities in the same syllables, in the same words), both as a function of pitch-accent. First, we report production range, and then quality shift data, and we always start with acoustics, and then carry on with data on articulation.

3.1. Production range

As for coarticulatory resistance (Fig. 2, left), data of *production range* in acoustics showed only the effect of target vowel quality ($F(1, 36) = 6.88$; $p < .05$), as /u/ was clearly more variable than /i/, but no effect of pitch accent was found. As for coarticulatory aggression, we found an interaction effect of vowel quality and trigger vowel accent ($F(1, 27) = 6.05$; $p < .05$) (Fig. 2, right), resulting from a main effect

of vowel quality being dependent on accent: /u/ was more variable (i.e., /i/ was more aggressive) only in those cases, where trigger vowels occurred in pitch-accented syllables (the main effect of trigger vowel accent was not significant).

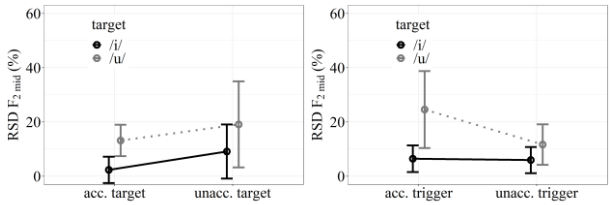


Figure 2: Production range in acoustics: coarticulatory resistance (left), coarticulatory aggression (right).

In articulatory data, we found a significant main effect of vowel quality for resistance exclusively ($F(1, 24) = 83.29$; $p < .001$) (Fig. 3, left), similar to what we observed in acoustics. As for aggression (Fig. 3, right), we found no effect of pitch accent either, only the significant main effect of vowel quality ($F(1, 32) = 60.34$; $p < .001$). To sum up, the measure of production range did not show that pitch-accent increased coarticulatory resistance and aggression in vowels under vowel-to-vowel coarticulation.

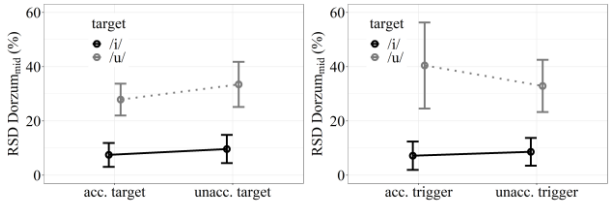


Figure 3: Production range in articulation: coarticulatory resistance (left), coarticulatory aggression (right).

3.2. Quality shift

Data on quality shift for /i/ and /u/ are basically capturing centralisation: if distances of coarticulated and neutrally positioned vowels are close to 0, there is no difference between the (articulatory or acoustic) quality of these tokens. However, if values are positive in the case of /u/, and negative in the case of /i/, it reflects tokens in coarticulating contexts being pulled towards the middle of the acoustic or articulatory space; in other words, to be more centralised. Absolute values of data reflect the magnitude of shift in quality, i.e., the magnitude of coarticulatory effects.

As for acoustics, we found the main effect of pitch-accent to be significant ($F(1, 28) = 4.22$; $p < .05$) (Fig. 4, left). Unexpectedly, however, in part, this was the result of /i/ being more centralised under the effect of vowel-to-vowel coarticulation in accented syllables than in unaccented syllables, while in unaccented syllables we found no centralisation (i.e., quality of /i/ was identical in coarticulating and neutral positions). Data on /u/ showed tendencies that may be considered increased resistance under the effect of pitch-accent: /u/ tokens in unaccented syllables showed basically no centralisation, while /u/ tokens in accented syllables (which were assumed to be less prone to coarticulatory effects) showed more peripheral realisations in coarticulating contexts than in neutral contexts, i.e., dissimilation (which is reflected by negative values in accented syllables).

As for coarticulatory aggression (Fig. 4, right), statistical analysis revealed that only the effect of vowel quality is significant ($F(1, 29) = 46.35$; $p < .001$), because /i/ exerted greater effects on /u/ than vice versa.

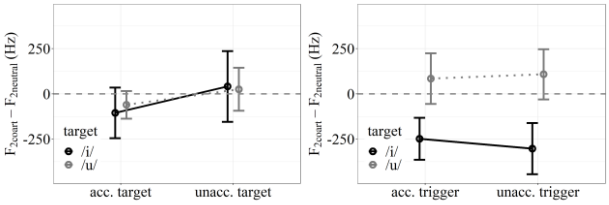


Figure 4: Quality shift in acoustics: coarticulatory resistance (left), coarticulatory aggression (right).

Lastly, we turn to quality shift data in the articulatory domain. With respect to resistance, again, only vowel quality showed a significant effect ($F(1, 33) = 4.65$; $p < .05$) (Fig. 5, left), as under the effect of vowel-to-vowel coarticulation, /u/ showed some degree of centralisation in unaccented syllables, and somewhat more peripheral realisation in accented syllables, but /i/ tokens were basically identical in coarticulating and in neutral contexts irrespective of prosodic prominence. With respect to coarticulatory aggression, we found no significant effect of pitch accent, but the effect of vowel quality ($F(1, 8) = 14{,}03$; $p < .05$) (Fig. 5, right). These data also showed that /i/ exerted a stronger effect of centralisation in /u/ than vice versa.

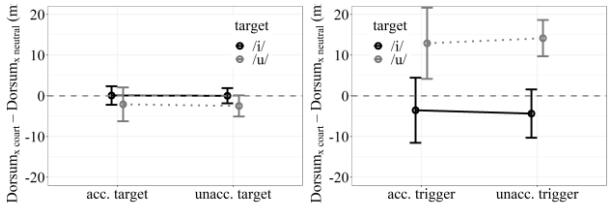


Figure 5: Quality shift in articulation: coarticulatory resistance (left), coarticulatory aggression (right).

4. DISCUSSION

We can conclude that transconsonantal vowels interacting exclusively on the backness dimension, and produced by Hungarian speakers, did not show clear effects of sentence level prominence, i.e., pitch-accent, either in articulation or in acoustics with

respect to their resistance against coarticulatory effects, and to their coarticulatory aggression. We found some support for accent effects, as /u/ showed increased resistance in articulation in pitch-accented syllables, but trends in /i/ went against expectations, and they showed greater resistance in unaccented syllables than in pitch-accented syllables. Based on these results, we can conclude that prosodic prominence did not facilitate segmental strengthening in the tested non-word-shaped tokens.

Present results contradict earlier findings that showed greater coarticulatory resistance in accented vowels in real words in Hungarian [7, 8], and in non-words in English [3, 6]. This could lead us to conclude that production of non-words in the present study blocked possible coarticulatory effects to a greater extent. This can be supported by previous results revealing that non-words may be hyperarticulated [14, 15], and are thus, intrinsically, less prone to coarticulatory effects [15, 16]. We should also note, however, that previous results demonstrating increased coarticulatory resistance in English were obtained using non-words, suggesting that resistance caused by hyperarticulation (if any), may be overruled by prosodic strengthening.

We assume that limited effects of pitch-accent found in Hungarian data in non-words may partly be due to the fact that, as opposed to English, prosody is claimed to play a limited role in prominence marking in Hungarian [17]. Conclusively, segmental strengthening exerted by prosodic means may also be limited in this language. Lack of strengthening effect is also supported by recent empirical findings: a previous study, in which pitch-accented vowels were compared to other, unstressed vowels of the same non-sense word in Hungarian showed no (clear) evidence of the effect of prosodic strengthening in accented vowels either in articulation (horizontal tongue position), or in acoustics (F_2) [18].

Further, it is also important to note that in the present study, for the purposes of controlling tested effects, and have a clear alignment of articulatory and acoustic measures, we tested coarticulation between /i/ and /u/, which is limited to only one dimension, backness. By contrast, [6] tested more widely spaced vowels, /i/ and /a/, differing both in openness and backness, where one expects the most extreme coarticulatory interaction.

Lastly, fixed first syllable stress/accent placement in Hungarian also limited the creation of the linguistic material used in the present study, thus we should also enumerate possible ways these features of the linguistic material could reduce observable coarticulatory effects.

On the one hand, coarticulatory resistance was tested in anticipatory coarticulation involving unaccented trigger vowels exclusively. Although studies for several languages demonstrated that carryover effects may exceed those of anticipatory [19], which would mean that we find limited effects in the present case, previous results for Hungarian demonstrated no clear dominance of directionality in vowel-to-vowel coarticulation [20]. According to these, only /i/ in acoustics, and /u/ in articulation showed that effects of carryover coarticulation may exceed those of anticipatory in Hungarian, but probable consequences of this cannot be seen in the present data. As we found no effect of pitch-accent on coarticulatory aggression in the present analysis, we can also safely assume that the lack of accent in trigger vowels did not diminish particularly the observable coarticulatory interactions.

On the other hand, coarticulatory aggression was tested in carryover effects, using unaccented target vowels. As mentioned, possible effects of direction of coarticulation for Hungarian are not straightforward; nevertheless, previous results suggest they would have facilitated coarticulatory effects in /i/ in acoustics, and in /u/ in articulation. In other words, direction effects would have facilitated higher aggression in /u/ triggers in acoustics, and in /i/ triggers in articulation. But this is also not seen in present results. Lastly, based on previous findings, one might expect that the lack of accent on target vowels facilitates coarticulatory interactions, hence it is an ideal testbed to reveal if accent enhances coarticulatory aggression. Lack of effect of pitch-accent on increasing coarticulatory aggression is then found despite favourable conditions.

To conclude, data on non-words in Hungarian did not show that coarticulatory resistance and aggression are increased in syllables that are the most prominent in the utterance, which supports the assumption that positions that are intended to be produced as prominent, do not universally result in segmental strengthening.

Acknowledgements

The work was supported by the Bolyai János Research Scholarship of the Hungarian Academy of Sciences (A.D., T.G.Cs.), ÚNKP-22-3 (K.J.), ÚNKP-21-5 (A.D.), ÚNKP-22-5 (T.G.Cs.). We are grateful to Ákos Buza, Valéria Krepsz, István Mrázik, Zsófia Puzder, Zsófia Weidl, Anne Hermes, Doris Mücke, and Theo Klinker for their help in gathering data.

5. REFERENCES

[1] Öhman, S. E. G. 1966. Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39, 151–168.

[2] Magen, H. 1997. The extent of vowel-to-vowel coarticulation in English. *J. Phon.* 25, 187–205.

[3] Fowler, C. S. 1981. Production and perception of coarticulation among stressed and unstressed vowels. *J. Speech Hear. Res.* 24, 127–139.
[4] de Jong, K. J. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* 97, 491–504.
[5] de Jong, K., Beckman, M. E., Edwards, J. 1993. The interplay between prosodic structure and coarticulation. *Lang Speech* 36, 197–212.
[6] Cho, T. 2004. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *J. Phon.* 32, 141–176.
[7] Deme, A., Bartók, M., Gráczi, T. E., Csapó, T.G., Markó, A. 2019. V-to-V Coarticulation induced acoustic and articulatory variability of vowels: the effect of pitch-accent. In: Gernot, K, Zdravko, K. (eds.) *Proc. 20th Interspeech 2019.* International Speech Communication Association (ISCA), 3317–3321.
[8] Deme, A., Bartók, M., Gráczi, T. E., Csapó, T.G., Markó, A., submitted. The effect of pitch-accent on the acoustic and articulatory variability of vowels: a real-word EMA study.
[9] Siptár, P. Törkenczy, M. 2000/2007. *The phonology of Hungarian.* Oxford University Press.
[10] R Core Team 2018. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.
[11] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67: 1–48.
[12] Kuznetsova, A., Brockhoff, P. B., Christensen, R., Haubo B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26.
[13] Lenth, R. V. 2016. Least-squares means: the R package lsmeans. *J. Stat. Softw.* 69, 1–33.
[14] Munson, B., Solomon, N. P. 2004. The effect of phonological neighborhood density on vowel articulation. *J. Speech Hear. Res.* 47. 1048–1058.
[15] Scarborough, R. 2012. Lexical similarity and speech production: Neighbourhoods for nonwords. *Lingua* 122, 164–176.
[16] de Jong, K. J. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* 97: 491–504.
[17] Mády, K., Kleber, F. 2010. Variation of pitch accent patterns in Hungarian. *Proc. of Speech Pros. 2010.* http://speechprosody2010.illinois.edu/papers/100924.pdf
[18] Markó, A., Juhász, K., Bartók, ., Csapó, T. G., Gráczi, T. E., Deme, A. 2022. Magyar magánhangzók artikulációs és akusztikai jellemzői a fonetikai pozíció függvényében álszavakban. *Általános Nyelvészeti Tanulmányok* 34, 51–80.
[19] Sharf, D. J., Ohde, R. N. 1981. Psychological, acoustic, and perceptual aspects of coarticulation: Implications for the remediation of articulatory disorders. In: Lass, N. J. (ed.): *Speech and language: Advances in basic research and practice.* Vol. 5. Academic Press, 253–247.
[20] Deme, A., Bartók, M., Csapó, T. G., Gráczi, T. E., Juhász, K., Markó, A. 2022. A magánhangzók centralizációja és produkciós homogenitása az előrefelé és hátrafelé ható magánhangzók közti koartikulációban – artikulációs és akusztikai adatok. *Általános Nyelvészeti Tanulmányok* 34, 15–50.

TOWARDS A MATERIAL POINT METHOD VOCAL TRACT SIMULATION

Scott R. Moisik

Nanyang Technological University
scott.moisik@ntu.edu.sg

ABSTRACT

The material point method (MPM) is a powerful numerical approach to the simulation of deformable bodies. It benefits from not requiring an explicit computational mesh to discretize a body, and this allows for myriad simulation possibilities including large deformations, fracture, phase change, mixture, and interactions between different materials, even with different phases (e.g., solid-fluid interaction). This paper reports the preliminary work on applying the MPM with the goal of creating a comprehensive 3D simulation environment for studying vocal tract biomechanics, aerodynamics, and acoustics (and the interaction of these).

Keywords: 3D vocal tract simulation, material point method

1. INTRODUCTION

The vocal tract is a complex collection of biological tissues. The main approach to three-dimensional computational simulation of vocal tract structures has been via the finite element method (FEM; e.g., [1]–[4]). A FEM simulation is principally defined by a collection of computational meshes, which provide a spatial discretization of any deformable bodies being simulated. Bodies that are very stiff, like bony structures, can be treated as rigid bodies (possessing only six rigid degrees of freedom, three translational and three rotational modes, and thus lacking deformation modes). This is an approximation given that even very stiff structures of the vocal tract can still undergo small deformation, but the loads experienced during speech production are small enough that these deformations can be entirely neglected. While the FEM approach benefits from possessing excellent accuracy in numerically solving the equations of motion over complex domains and from being relatively light-weight in computational burden, the strong dependency on computational meshes presents serious challenges. In particular, vocal tract structures like the tongue can undergo large deformations that excessively distort elements, reducing accuracy and stability (see [5]). Developing meshes with good computational properties is not a trivial matter and automated algorithms do not always produce useful results. Thus, there is a desire for numerical techniques that do away with the strong dependence on a computational mesh, such as the material point method (MPM).

1.1. Material point method (MPM): Background

The MPM was originally proposed by Sulsky et al. [6] as an extension of the particle-in-cell method, a well-known technique for fluid dynamics simulation [7], to the simulation of solid bodies. The MPM shares conceptual DNA with other fluid dynamics methods, such as smoothed-particle hydrodynamics [8]. It uses stateful particles (or packets/volumes) of mass to discretize a material of interest (be it a simple fluid like water or even more exotic matter like plasma or an entire star) and integration of key properties represented by the distribution of such particles is performed via kernel or shape functions.

One of the benefits of the MPM is that, because materials transfer their momentum to a shared background grid, it is possible to effortlessly simulate contacts and collisions between deformable bodies. There is no need for explicit contact detection and collision response handling, as communication between disjoint bodies (and even different parts of the same body in self-collision), occurs when the momenta of material points from interacting bodies are accumulated on the background grid. The down-side to this is that the resolution of the background grid dictates the resolution of the contact interface, with coarser grids leaving a small separation between bodies coming into contact. Another major challenge with shared background grids is that interaction between materials is sticky: that is, the ordinary formulation of the MPM imposes an obligatory no-slip boundary condition on interacting bodies. The practical consequence of this is that separate materials cannot simply slide past each other, but rather experience infinite friction and thus exhibit a sticking behaviour when coming into close contact.

2. MPM VOCAL TRACT SIMULATION

In this work, the development of a simulation environment is described that is intended for simulation of vocal tract structure biomechanics and interaction with aerodynamic and acoustic flows. Currently, the MPM is the sole approach used (although it is possible to mix the MPM with other

numerical approaches, including finite elements, as in [9]). Progress towards the goal of a comprehensive general-purpose solution for vocal tract simulation is underway but certainly not finished, and the focus here is to report on steps taken towards the biomechanical component while also describing some proof-of-concept examples that use the MPM for simulating vocal tract aerodynamics and acoustics.

2.2. Simulation of rigid bodies

While the MPM is a strong choice for simulating deformable solids and fluids, like in FEM simulation, stiff bodies can pose numerical stability issues if the appropriate integration scheme is not used. This means that a strategy is required to effectively handle rigid bodies, like the mandible and hyoid bone.

Rigid bodies can be modelled separately via traditional methods for simulating rigid body mechanics (e.g., as might be used in robotics), and this method has been used successfully with the MPM (see, e.g., [10]). Another possibility is to apply a shape correction to a deformable body that limits the body to purely rigid modes of displacement. While the literature contains no record of such an approach being used in the context of the MPM, it has been pursued for another particle-based method known as position based dynamics (PBD) [11]. After a body is updated, its material points are projected to the positions of the same body if it were to only undergo purely rigid transformation. These projected positions are established via Procrustes registration and once the projections are made, a velocity correction is applied to conserve momentum.

2.3. Simulation of muscle contraction

The approach to muscle simulation used here follows the work of Blemker and colleagues (see, e.g., [12]–[14]). These researchers developed a numerical model that approaches muscle tissue as a fiber-reinforced composite embedded within finite-element volumes representing muscle bellies. There is no research on muscle simulation using the MPM, but the concept of fiber-reinforced materials has been explored in [15] with the aim to simulate anisotropic material damage. The implementation here modifies the constitutive law for chains of particles forming muscle fibers to employ the muscle model described by Blemker.

3. DEMONSTRATIONS

To illustrate the capabilities of the vocal tract simulation environment under development, several different demonstrations are presented below that feature different aspects of what is possible with the MPM approach.

3.1. Deformable body simulation

In Fig. 1, Jell-O-like blocks demonstrate the utility of the shared-background grid in automatically handling contact and collision. As the red (top) block shows in its separating bounce in the final (rightmost) frame, despite "sticky" interaction inherent to basic MPM, the objects can still separate if their material properties, constitutive model, and momenta favour pulling local parts of material away in regions of proximity.

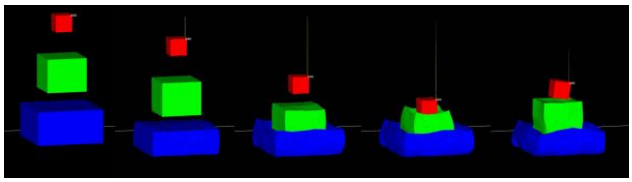


Figure 1: Falling Jell-O-like blocks. A sequence of frames showing the evolution of a simulation where three deformable blocks are dropped and collide into each other demonstrating automatic collision handling on a shared background grid.

3.2. The larynx-hyoid-jaw-tongue model

Currently the lower hull of the vocal tract (Fig. 2), complete with a detailed representation of the larynx (Fig. 3), has been developed. The near term goal is to expand this set of structures to include the lips, maxilla, soft palate, and pharynx. Drawing inspiration from [2], an even more comprehensive set that includes the trachea and cervical spine can also be envisioned.

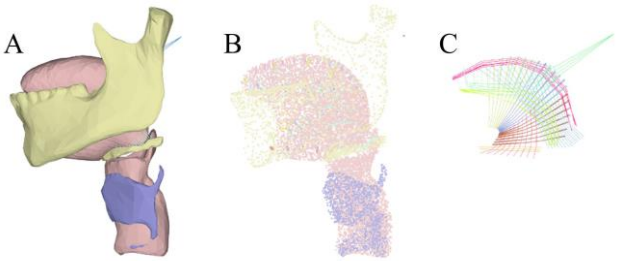


Figure 2: "Lower hull" vocal tract model featuring jaw-tongue-hyoid-larynx complex with laryngeal mucosa. The interior of surface meshes (A) are sampled to identify material points (B). Bodies like the tongue with its lingual muscles (C) are treated as fiber-reinforced composites.

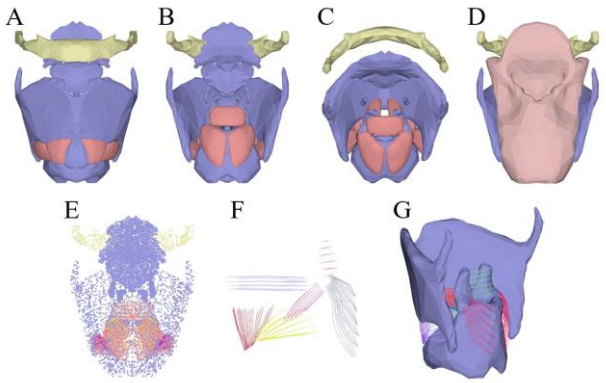


Figure 3: Various views of the larynx model. Surface meshes include the laryngeal cartilages, major internal muscles, and mucosa (A-D). The interior meshes are sampled to generate material points (E), and muscle portions are reinforced by muscle fibers (F-G).

3.2.1. Coupled rigid and deformable bodies

In Fig. 4, the coupling between the tongue and the jaw and the tongue and larynx (via the hyoid bone) is accomplished using invisible springs that simulate the various tendinous and ligamentous connections. More detailed simulation of these structures with material points is also possible. Rigidity of structures like the jaw is enforced using the projection method inspired by PBD described above in Section 2.2.

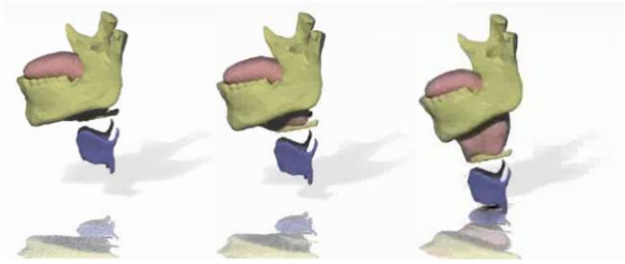


Figure 4: Coupled rigid and deformable bodies simulated under the influence of gravity: (from left-to-right) the larynx-hyoid-jaw-tongue complex starts at rest with the jaw pinned at its condyles and then falls to a resting state when the thyroid collides with the floor.

3.2.2. Muscle simulation

Fig. 5 and Fig. 6 illustrate simulation of tongue and laryngeal muscle action, both using an implementation of the Blemker-style fiber-reinforced muscle model [13]. The sticky (high-friction) interaction between tongue and jaw limits some of the tongue's mobility. Tendons can be implemented using extensibility constraints described in [15] to correct some issues, such as the excessive posterior displacement of the tongue volume in the region of the mental spines in response to genioglossus posterior activity.

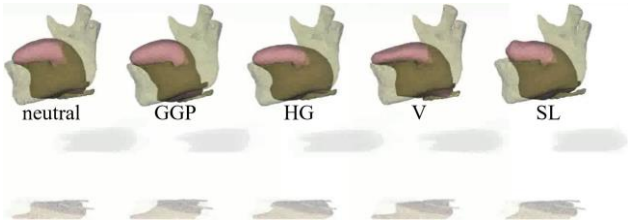


Figure 5: Tongue muscle action: neutral = no muscles active; GGP = genioglossus posterior; HG = hyoglossus; V = vertical; SL = superior longitudinal.

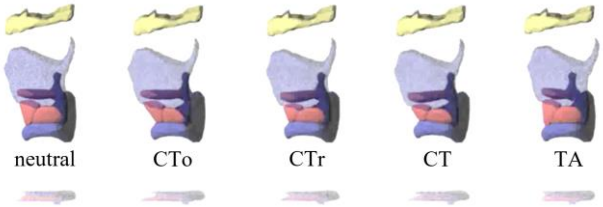


Figure 6: Larynx muscle action: neutral = no muscles active; CTo = cricothyroid pars oblique; CTr = cricothyroid pars recta; CT = cricothyroid (full muscle); TA = (internal) thyroarytenoid.

3.4. Explorations of vocal tract aerodynamics

With the right constitutive model, it is possible to simulate almost any material, including air (see, e.g., [16]). Preliminary tests on aerodynamic and acoustic simulations show promise for simulating phenomena such as ejectives (Fig. 7) and air filled membranes (Fig. 8); with the latter, one possible application is for simulation of the laryngeal air sac of the gibbon.

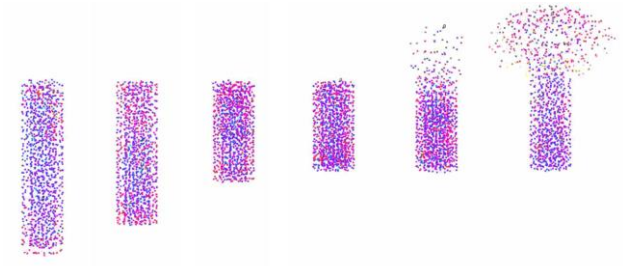


Figure 7: Proof-of-concept ejective simulation. The sequence shows (from left to right) an initially closed vocal tract tube being vertically squashed to emulate (much exaggerated) larynx raising followed by release of the 'oral' stricture at the other end of the tube.

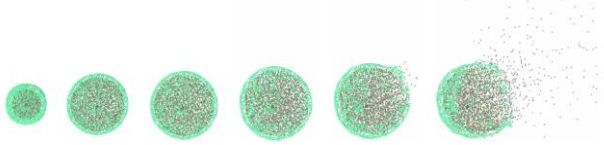


Figure 8: Balloon pop. The sequence shows (from left to right) a membrane filled with compressed air rapidly expanding and then bursting.

3.5. Explorations of vocal tract acoustics

Beyond aerodynamic simulations, the MPM may hold promise for detailed simulation of acoustic and even aeroacoustic phenomena. A proof-of-concept test was made using a constraint surface in the form of the vocal tract in [i] configuration (Fig. 9). The air was subjected to a periodic velocity field near the glottis to simulate the glottal volume velocity. Velocity waves can be seen to bounce around the vocal tract, suggesting transmission of acoustic energy.

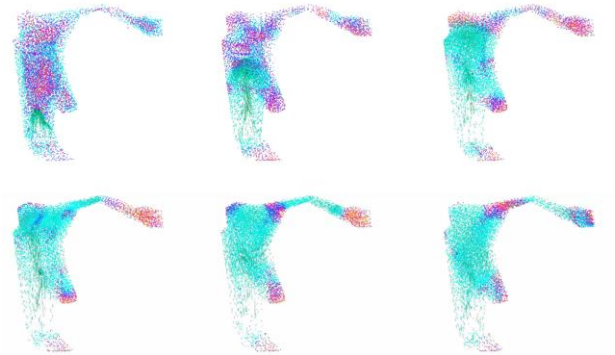


Figure 9: Proof-of-concept [i] acoustics. A 3D vocal tract constraint is filled with air subjected to a glottal volume velocity field. The transmission of wave energy can be seen as the system evolves over time (left to right, top to bottom). Material points are coloured by their speed.

3.6. Fluid-solid interaction: An unconstricted epiglottal trill

The ability to simulate fluid-solid interaction using air within the vocal tract would allow for simulation of a wide range of speech-relevant phenomena involving vibration. As a demonstration, an 'unconstricted' epiglottal trill was simulated using air and the laryngeal mucosa (Fig. 10). Epiglottal trills normally feature strong constriction of the epilarynx, but this proof-of-concept model only uses the mucosa in its neutral state without the laryngeal muscle activity required for constriction. It was necessary to employ separate grids with an explicit collision model to avoid the "sticky" interaction problem.

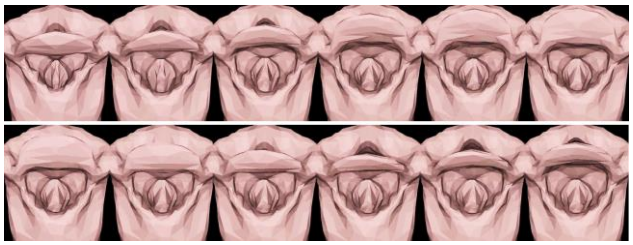


Figure 10: "Unconstricted" epiglottal trill. A sequence of frames (from left to right, top to bottom) showing the laryngeal mucosa interacting with air (not visualized) resulting in vibration of the epiglottis.

4. LIMITATIONS AND FUTURE DIRECTIONS

While the initial explorations of various types of vocal tract simulation are promising, a number of limitations need to be overcome before the model can be used to explore real problems in vocal tract behaviour during speech production (and related areas like singing, swallowing, and so forth). Most important is that the approach to fluid-solid interaction currently employed uses dual and independent background grids to separately update the solid and fluid (air) materials with a collision model to allow for information to flow between the materials. This is not ideal because it undercuts the ability of the MPM to allow for this communication to happen implicitly without any external contact and collision handling. One approach to solving this problem has been proposed by Fang et al. [17] called interface quadrature MPM (IQ-MPM), which employs an air-like "ghost matrix" (analogous to a biological matrix that surrounds and supports living cells). This allows for strong two-way interaction between materials that crucially enables slip boundary conditions, avoiding the sticky-contact problem inherent to 'vanilla' MPM. Another key difficulty is the simulation of very stiff materials. Currently the model only employs explicit integration using the forward Euler numerical scheme to update the equations of motion. The use of an implicit integrator that employs a backward Euler scheme can be implemented to allow for much larger time steps and with much improved numerical stability [18].

5. REFERENCES

[1] F. Alipour, D. A. Berry, and I. R. Titze, "A finite-element model of vocal-fold vibration," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3003–3012, Nov. 2000, doi: 10.1121/1.1324678.

[2] P. Anderson *et al.*, "Chapter 20 - FRANK: A Hybrid 3D Biomechanical Model of the Head and Neck," in *Biomechanics of Living Organs*, Y. Payan and J. Ohayon, Eds. Oxford: Academic Press, 2017, pp. 413–447. doi: 10.1016/B978-0-12-804009-6.00020-1.

[3] S. Deguchi, Y. Kawahara, and S. Takahashi, "Cooperative regulation of vocal fold morphology and stress by the cricothyroid and thyroarytenoid muscles," *J. Voice*, vol. 25, no. 6, pp. e255–e263, Nov. 2011.

[4] J. E. Lloyd, I. Stavness, and S. S. Fels, "ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," in *Soft tissue biomechanical modeling for computer assisted surgery*, Y. Payan, Ed. Berlin, Germany: Springer, 2012, pp. 355–394.

[5] S. Garg and M. Pant, "Meshfree Methods: A Comprehensive Review of Applications," *Int. J.*

Comput. Methods, vol. 15, no. 04, p. 1830001, Jun. 2018, doi: 10.1142/S0219876218300015.

[6] D. Sulsky, S.-J. Zhou, and H. L. Schreyer, "Application of a particle-in-cell method to solid mechanics," *Comput. Phys. Commun.*, vol. 87, no. 1, pp. 236–252, May 1995, doi: 10.1016/0010-4655(94)00170-7.

[7] F. H. Harlow, "The particle-in-cell method for numerical solution of problems in fluid dynamics," Los Alamos Scientific Lab., N. Mex., LADC-5288, Mar. 1962. doi: https://doi.org/10.2172/4769185.

[8] R. A. Gingold and J. J. Monaghan, "Smoothed particle hydrodynamics: theory and application to non-spherical stars," *Mon. Not. R. Astron. Soc.*, vol. 181, no. 3, pp. 375–389, Dec. 1977, doi: 10.1093/mnras/181.3.375.

[9] C. Jiang, C. Schroeder, A. Selle, J. Teran, and A. Stomakhin, "The affine particle-in-cell method," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–10, Jul. 2015, doi: 10.1145/2766996.

[10] Y. Hu *et al.*, "A moving least squares material point method with displacement discontinuity and two-way rigid body coupling," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Aug. 2018, doi: 10.1145/3197517.3201293.

[11] M. Müller, B. Heidelberger, M. Teschner, and M. Gross, "Meshless deformations based on shape matching," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 471–478, Jul. 2005, doi: 10.1145/1073204.1073216.

[12] S. S. Blemker and S. L. Delp, "Three-Dimensional Representation of Complex Muscle Architectures and Geometries," *Ann. Biomed. Eng.*, vol. 33, no. 5, pp. 661–673, May 2005, doi: 10.1007/s10439-005-1433-7.

[13] S. S. Blemker, P. M. Pinsky, and S. L. Delp, "A 3D model of muscle reveals the causes of nonuniform strains in the biceps brachii," *J. Biomech.*, vol. 38, no. 4, pp. 657–665, Apr. 2005, doi: 10.1016/j.jbiomech.2004.04.009.

[14] M. Stone *et al.*, "Structure and variability in human tongue muscle anatomy," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 5, pp. 499–507, Sep. 2018, doi: 10.1080/21681163.2016.1162752.

[15] J. Wolper *et al.*, "AnisoMPM: animating anisotropic damage mechanics," *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020, doi: 10.1145/3386569.3392428.

[16] W. Hu and Z. Chen, "Model-based simulation of the synergistic effects of blast and fragmentation on a concrete wall using the MPM," *Int. J. Impact Eng.*, vol. 32, no. 12, pp. 2066–2096, Dec. 2006, doi: 10.1016/j.ijimpeng.2005.05.004.

[17] Y. Fang *et al.*, "IQ-MPM: an interface quadrature material point method for non-sticky strongly two-way coupled nonlinear solids and fluids," *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020, doi: 10.1145/3386569.3392438.

[18] T. F. Gast, C. Schroeder, A. Stomakhin, C. Jiang, and J. M. Teran, "Optimization Integrator for Large Time Steps," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 10, pp. 1103–1115, Oct. 2015, doi: 10.1109/TVCG.2015.2459687.

CO-REGISTRATION OF SIMULTANEOUS HIGH-SPEED ULTRASOUND AND ELECTROMAGNETIC ARTICULOGRAPHY FOR SPEECH PRODUCTION RESEARCH

Sam Kirkham[1], Patrycja Strycharczuk[2], Emily Gorman[1], Takayuki Nagamine[1], Alan Wrench[3]

[1]Lancaster University, [2]University of Manchester, [3]Queen Margaret University
{s.kirkham, e.gorman, t.nagamine}@lancaster.ac.uk, patrycja.strycharczuk@manchester.ac.uk,
awrench@articulateinstruments.com

ABSTRACT

We outline a method for collecting simultaneous high-speed ultrasound and electromagnetic articulography (EMA) data using widely-available commercial hardware and software. Previous research demonstrates the utility of a simultaneous EMA-ultrasound set-up and reports the magnitude of probe rotation and displacement over an experimental session. We build upon this research by (1) combining systems with higher temporal and spatial accuracy than those previously used; (2) using a more principled method for temporal synchronisation; (3) reporting the effects of ultrasound on the accuracy of EMA tracking; (4) tracking probe movement using an alternative stabilisation method. Our system shows a high degree of temporal and spatial accuracy and can be easily implemented by other researchers.

Keywords: Electromagnetic articulography, ultrasound tongue imaging, speech articulation, co-registration

1. INTRODUCTION

In this study we report the design and evaluation of an approach to recording simultaneous ultrasound and electromagnetic articulography (EMA). Ultrasound provides a way of imaging the majority of the tongue surface from tip to root, and is a relatively non-invasive and cost-effective approach to imaging speech articulation. Ultrasound frame rates are lower with video-based capture (∼30 Hz), but modern digital systems are capable of 200+ Hz frame rates (although 80–100 Hz is more common given the settings used for most speech imaging). Ultrasound typically involves stabilisation of the probe relative to the head [1, 2] or tracking of head motion for post-hoc correction [3]. EMA differs from ultrasound in that it tracks sensors attached to specific fleshpoints such as the tongue and the lips, but it has very high temporal resolution (up to 1250 Hz but commonly downsampled to 250 Hz). A further advantage of EMA is that it can easily capture movement in three-dimensions and simultaneously capture movement in sagittal planes, e.g. by tracking lateral edges of the tongue [4, 5]. There exist a wide range of established timing measures for EMA data [6], while such measures for ultrasound are still largely in development [7].

Combining the high temporal accuracy and precise point-tracking of EMA with the rich spatial information of ultrasound is a promising avenue for a range of experimental paradigms in speech research. To this end, [8, 9] report the results of simultaneous EMA-ultrasound experiments based on a new probe stabilisation method. They find excellent stabilisation of the probe relative to the head, with transducer rotation limited to 1.25° and translation limited to 2.5 mm, which is within the tolerances of 2–4 mm translation and rotation of less than ∼5° outlined by [3]. In terms of empirical and theoretical research, [4] combine EMA and midsagittal ultrasound to analyse coda /l/ darkening and tongue lateralisation in New Zealand English.

There remain challenges, however, to conducting simultaneous EMA-ultrasound studies. These include accurate and principled temporal synchronisation between the two systems, ensuring no effects of the ultrasound hardware on EMA tracking accuracy, and minimising the effect of probe movement. In this study, we build upon previous work, outlining an approach to synchronisation between systems, quantifying the effect of ultrasound on EMA tracking, and reporting probe motion using an alternative probe stabilisation method.

2. EXPERIMENTAL SET-UP

2.1. Hardware

The hardware used in this study is a 16-channel Carstens AG501 electromagnetic articulograph and

a Telemed MicrUS scanner. The AG501 system records three positional coordinates and two angular coordinates at 1250 Hz for each sensor, and emits a synchronisation pulse every 4 ms for audio-EMA synchronisation. The Telemed ultrasound scanner emits nanosecond-level synchronisation pulses, which are converted to millisecond-level pulses using an Articulate Instruments Pulse Stretch unit. Each ultrasound frame is tagged with the corresponding sychronisation time, providing frame-level accuracy in temporal synchronisation between audio and ultrasound signals. The ultrasound system in this set-up uses a 64-element microconvex probe with 20 mm radius and 2 MHz frequency.

In our experiments, the acoustic speech signal is recorded using a Beyerdynamic Opus 55 omnidirectional microphone attached to the ultrasound headset, which is pre-amplified using a Grace m101 pre-amplifier. The signal is then split from the pre-amplifier's two outputs and recorded simultaneously onto both the ultrasound computer (via an Adlink DAQ-2213 analog-digital converter) and EMA computer (via an Alesis iO2 audio interface).

We stabilise the ultrasound probe relative to the speaker's head using the Ultrafit headset [10] produced by Articulate Instruments, which is entirely made out of plastic and non-metallic fabric (including replacement nylon bolts). The headset must be non-metallic for simultaneous EMA data collection, as metallic objects in or near the magnetic field significantly degrade the accuracy of EMA's position estimation.

2.2. Synchronisation

Synchronisation between EMA and ultrasound is achieved using synchronisation pulse signals from each system, which are already used for audio-ultrasound and audio-EMA synchronisation respectively. First of all, each system synchronises data separately at the time of recording, giving us a set of synchronised audio-ultrasound recordings and a set of synchronised audio-EMA recordings. Note that the audio-ultrasound and audio-EMA data are not synchronised with each other at this point, as there will be slight variation in the onset of recording and clock times between the two systems.

In order to precisely synchronise the EMA and ultrasound data, we record the AG501's TTL synchronisation pulse onto the computer recording ultrasound data as an extra audio track alongside the acoustic data and ultrasound sync signal. The EMA sync signal goes low 80 ms before recording begins and then sends a pulse every 4ms. This means we are able to align the edge of the synchronisation pulse that marks the onset of the EMA recording with the beginning of the EMA position file. As we describe below, all of the above procedures are carried out by the Articulate Assistant Advanced software using the Ultrasound and EMA modules, which only require the user to specify the relevant synchronisation channels and all synchronisation is then handled without further intervention from the user.

2.3. Software

This set-up uses the Articulate Assistant Advanced (AAA) software [11]. The software has a number of features that facilitate a simultaneous ultrasound-EMA experiment, which are as follows. First, it allows a network connection to the Carstens AG501's control server, which means that AAA can handle prompt presentation and remotely trigger the AG501 recording, meaning that the experimenter only has to operate one computer while recording data. Second, it automatically synchronises the ultrasound and audio data. Third, it allows EMA position data recorded using the AG501 to be imported and also keeps a record of which AAA recording corresponds to which EMA file. Finally, it automatically aligns the recorded EMA data with the audio-ultrasound, based on the EMA synchronisation pulse that is recorded during the experiment (as detailed in Section 2.2). This allows us to obtain synchronised audio, high-speed ultrasound and EMA data without any additional pre-processing steps.

2.4. Validation experiments

In order to examine the accuracy and validity of our system, we report the results of two controlled tests. The first examines the effects of the ultrasound hardware and headset on the accuracy of EMA tracking. The second aims to quantify the amount of ultrasound probe movement relative to the head over the course of a typical EMA session using the UltraFit headset. The data and specific approach used for each experiment are described in the respective sections that follow.

3. EFFECTS OF ULTRASOUND ON EMA ACCURACY

We first test the effects of recording simultaneous ultrasound and EMA on the accuracy of the EMA position calculation. While we anticipate that

the plastic ultrasound headset should not influence EMA tracking, we are nonetheless required to keep the ultrasound scanner placed 60 cm behind the centre of the tracking field in order appropriately position the probe on the speaker, so we test whether this nearby hardware has any effect on the accuracy of EMA tracking. In order to test this, we recorded six calibration sessions (three with ultrasound, three without ultrasound) in order to factor out any random differences due to the effect of calibration session, although calibration statistics were near identical within each set of three runs. The ultrasound condition involves the ultrasound probe and headset inside the EMA field, with the scanner hardware placed on the rear of the EMA stand, approximately 60 cm from the centre of the field. Calibration involves the rotation of EMA sensors placed on a purpose-built 'circal' device that moves the EMA sensors around a series of pre-determined rotations, allowing us to compare the ideal positions with the actual positions.

condition	Δz (sd)	σz (sd)
no ultrasound	0.661 (0.086)	0.148 (0.025)
with ultrasound	0.670 (0.090)	0.150 (0.026)

Table 1: Δz and σz statistics from calibration runs without ultrasound equipment and with ultrasound equipment placed in the EMA field. Values represent averages across 16 sensor channels for three different runs per condition, with standard deviations in brackets.

We evaluate the accuracy of calibration using two measures: (1) Δz, which is the peak-to-peak deviation of the z-coordinates from each channel during calibration; and (2) σz, the associated standard deviation. In both cases, smaller numbers represents more accurate calibration, and σz should remain below 0.25 mm. Table 1 shows only very minor differences between conditions, suggesting that the effect of the EMA hardware is near-zero and certainly too small to have a noticeable effect on measurements. In addition to this, all channels showed σz values between 0.11–0.21 mm, which are below the recommended threshold of 0.25 mm for accurate calibration.

4. TRACKING ULTRASOUND PROBE MOTION

We also evaluated translation and rotation of the ultrasound probe during a real-world experiment of vowels in British English. We recorded six female speakers of northern British English producing a large number of phrases of the form 'she says X', 'she says X dearly' and 'she says X eagerly', where X was a bV or bVd word. 58 phrases were produced in a block and each block was repeated 5 times, producing 1740 tokens across the 6 speakers. One speaker (F05) had to have the ultrasound probe re-positioned after 1 block due to improper fitting and subsequently recorded the full 5 blocks after the probe was moved; we only use the final 5 blocks for this speaker. A standard set of EMA sensors was used for each speaker and we corrected for head motion using two sensors behind the ears, one on the bridge of the nose, and one on the upper incisors.

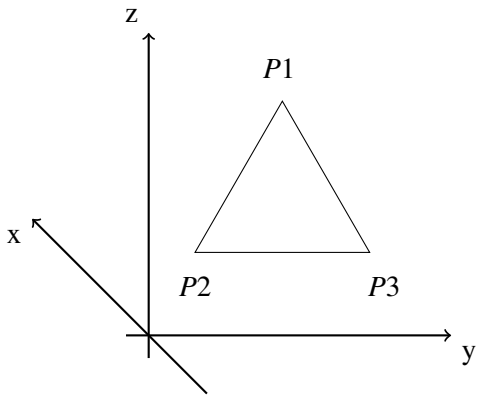


Figure 1: Schematic of rigid triangle attached to the ultrasound headset for tracking probe movement. $\{P1, P2, P3\}$ represent the corners at which EMA sensors were attached. The axes are oriented from the speaker's perspective, with $P2$ located on the speaker's left-hand side and $P3$ located on the speaker's right-hand side. In this coordinate system, x represents anterio-posterior motion, y represents lateral motion, and z represents vertical motion.

We tracked probe motion using a plastic triangle attached to the ultrasound headset, with EMA sensors fixed to each corner of the triangle. The triangle was custom designed and 3D printed to be retrofitted to the Articulate Assistant UltraFit headset. The STL file for the 3D printed triangle is available at the following link: https://github.com/samkirkham/ultrafit-ema. The triangle screws onto the headset using existing probe holder bolts and a small notch on the rear of the triangle prevents rotation around the bolt. We label the sensors on the triangle as $\{P1, P2, P3\}$, each of which has the (x, y, z) coordinates visualised in Figure 1. Note that any references to left-right orientation are from the perspective of the speaker.

As the plastic triangle is displaced from the actual ultrasound probe origin, we measured the distance in x and z dimensions between the top sensor on the headset triangle ($P1$) and the centre of the ultrasound

transducer array, and subtracted these values from the sensor coordinates. We then defined a virtual origin at the probe (reference object) and calculated the rotations required to align the reference object with the measured probe sensors. This was achieved by computing the rotation matrix that would align the position of the probe with the reference position for each data sample, which gives us the rotations around the respective x, y, z axes. We implemented this procedure in Python using the SciPy [12] spatial.transform sub-module, whereby we calculate the rotation between vectors using the algorithm in [13]. The final output is the displacement of the ultrasound probe in x, y, z dimensions, as well as the rotations around these dimensions: roll, pitch and yaw. For comparability with previous research [8, 9], we report the standard deviation of probe displacement and rotation across the experimental session for each speaker, as well as the 95% variance interval (1.98 x SD) for each measurement (note that for the rotations we calculated circular standard deviations). These values are shown in Table 2.

speaker	x	y	z	roll	pitch	yaw
F01	1.26	0.29	0.80	0.22	0.61	0.29
F02	1.39	0.82	0.86	0.68	0.93	0.39
F03	1.76	1.12	1.29	0.91	1.09	0.41
F04	1.21	0.36	0.70	0.38	0.71	1.00
F05	1.49	1.85	1.78	1.39	0.59	0.52
F06	0.66	0.73	0.80	0.52	0.34	0.89
mean	1.29	0.86	1.04	0.68	0.71	0.58
95%	2.56	1.71	2.06	1.35	1.41	1.15

Table 2: Standard deviation of displacement (mm) and circular standard deviation of rotation (degrees) of the ultrasound probe across all recordings for each speaker.

Table 2 shows mean translation of 0.86–1.29 mm and mean rotations of 0.58–0.71°, with maximum 95% variance values of 2.56 mm and 1.41°. This is comparable to the metal Articulate Instruments headset, which shows up to 1.8 mm translation and 1.3° rotation [2], and is well within the HOCUS tolerances of 2–4mm translation and 5–7° rotation [3]. Our statistics are also highly comparable to those in [8, 9], who report maximum values of 2.5 mm and 1.25°. We note, however, that our experimental materials used in this test are phonetically very similar to one another, so it would be worthwhile to conduct comparable tests using more phonetically diverse materials, which could potentially induce a wider range of probe movements.

5. CONCLUSIONS

We reported a method for simultaneous EMA-ultrasound data collection using widely-available hardware and software. Our system shows very high accuracy in temporal synchronisation, with minimal-to-no influence of the ultrasound scanner on EMA tracking accuracy, and ultrasound probe movement is well within accepted tolerances. This set-up can be implemented easily by other researchers who have access to the same commercial hardware and software, meaning that the experimental set-up can be replicated across different laboratories.

In future research, we aim to (1) use the known rotations and translations from this analysis to correct the ultrasound splines for probe movement, thereby locating both modalities in a common reference frame; (2) evaluate the comparability of articulatory timing measures based on EMA versus ultrasound data.

ACKNOWLEDGEMENTS

This research was funded by AHRC grant AH/S011900/1.

6. REFERENCES

[1] M. Stone and E. P. Davis, “A head and transducer support system for making ultrasound images of tongue/jaw movement,” *Journal of the Acoustical Society of America*, vol. 98, no. 6, pp. 3107–3112, 1995.

[2] J. M. Scobbie, A. A. Wrench, and M. L. van der Linden, “Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement,” *Proceedings of the 8th International Seminar on Speech Production*, pp. 373–376, 2008.

[3] D. Whalen, K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-LeHouilier, and E. Vatikiotis-Bateson, “The Haskins Optically Corrected Ultrasound System (HOCUS),” *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 3, pp. 543–553, 2005.

[4] P. Strycharczuk, D. Derrick, and J. A. Shaw, “Locating de-lateralization in the pathway of sound changes affecting coda /l/,” *Laboratory Phonology*, vol. 11, no. 1, p. 21, 2020.

[5] J. Ying, J. A. Shaw, C. Carignan, M. Proctor, D. Derrick, and C. T. Best, “Evidence for active control of tongue lateralization in Australian English /l/,” *Journal of Phonetics*, vol. 86, no. 101039, pp. 1–22, 2021.

[6] S. Marin and M. Pouplier, “Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling

model," *Motor Control*, vol. 14, no. 3, pp. 380–407, 2010.

[7] P. Strycharczuk and J. M. Scobbie, "Velocity measures in ultrasound data: Gestural timing of post-vocalic/l/ in English," *Proceedings of the XVIII International Congress of Phonetic Sciences*, pp. 1–5, 2015.

[8] D. Derrick, C. T. Best, and R. Fiasson, "Non-metallic ultrasound probe holder for co-collection and co-registration with EMA," *Proceedings of the 19th International Congress of Phonetic Sciences*, pp. 1–5, 2015.

[9] D. Derrick, C. Carignan, W.-R. Chen, M. Shujau, and C. T. Best, "Three-dimensional printable ultrasound transducer stabilization system," *Journal of the Acoustical Society of America*, vol. 144, no. 5, pp. EL392–EL398, 2018.

[10] L. Spreafico, M. Pucher, and A. Matosova, "UltraFit: A speaker-friendly headset for ultrasound recordings in speech science," *Proceedings of Interspeech 2018*, pp. 1–4, 2018.

[11] A. Wrench, "Articulate Assistant Advanced, v. 220.04," 2022.

[12] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[13] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica*, vol. A32, no. 5, pp. 922–923, 1976.

5-MINUTE FORMANT ADAPTATION TASK IN DUTCH CHILDREN

Teja Rebernik[1], Thomas B. Tienkamp[1], Katharina M. Polsterer[1], Raoul Buurke[1], Vera Hukker[1], Masha Medvedeva[2], Mara van der Ploeg[1], Iris Schepers[1], Hedwig Sekeres[1], Wietse de Vries[1], Defne Abur[1], Roel Jonkers[1], Aude Noiray[3,4], Martijn Wieling[1,4]

[1]University of Groningen, [2]University of Leiden, [3]Université PSL, [4]Haskins Laboratories
t.rebernik@rug.nl, t.b.tienkamp@rug.nl, k.m.polsterer@student.rug.nl, raoul.buurke@rug.nl, v.hukker@rug.nl, m.medvedeva@law.leidenuniv.nl, a.m.van.der.ploeg@rug.nl, i.schepers@rug.nl, h.g.sekeres@rug.nl, wietse.de.vries@rug.nl, d.abur@rug.nl, r.jonkers@rug.nl, aude.noiray@ens.psl.eu, m.b.wieling@rug.nl

ABSTRACT

The study investigates whether a short formant perturbation experiment elicits an adaptive response under less controlled experimental circumstances. 30 Dutch children were recruited and tested at a festival. They were asked to produce four target words containing an open-mid front rounded vowel /ɛ/ while we manipulated their feedback so that they would hear /ɪ/ for a period of 16 trials. Despite the short adaptation paradigm, our results show that children significantly changed their vowel productions in response to the perturbation. This suggests that long and monotonous experimental paradigms might not always be necessary, especially with populations that have a shorter attention span.

Keywords: formant adaptation, child speech, speech motor control

1. INTRODUCTION

When speaking, we rely on both feedback and feedforward speech motor control mechanisms, producing sounds in line with a known target and issuing immediate corrections if the produced sound does not correspond to this target (see e.g., [1] for one formal description of the speech motor control system). In the auditory domain, we know what a certain target is supposed to sound like, and we will correct for any mismatches between what we expected to hear and what we actually heard. The auditory *feedback* control subsystem plays a crucial role in child language acquisition, while the auditory *feedforward* control subsystem is more important for adult speakers. Despite this, it is currently unclear how precisely the use of auditory feedback changes throughout childhood.

Formant shifting paradigms (pioneered by Houde and Jordan [2]) can tell us more about articulatory-motor control in children and how they use auditory feedback during speech. Prior studies have shown that children do compensate as a response to formant manipulations but do not do so before the age of 2.5 years [3]. Especially younger children tend to be more variable in their productions compared to adult speakers [4, 3], and at least one study has found a differential effect in F_1 and F_2, with a greater compensatory change in F_1 for younger children, but in F_2 for older children [5]. Additionally, a study with a large sample of 244 either Dutch- or English-speaking children [6] showed stronger compensation in literate compared to pre-literate children.

A recent review by Coughler and colleagues [7] found that the majority of formant perturbation studies in children focus on English native speakers, and rarely examine a broad range of age groups. Furthermore, the experiments are rather long (with a median of 100 trials). This longer duration may not be ideal for children who often have a relatively short attention span.

The goal of our study was therefore to determine whether a short, five-minute, experiment including several target words (instead of the frequently used single target word) is also able to elicit vowel formant adaptation, even when carried out under less-than-ideal experimental circumstances with many potential distractors (see below). We additionally wished to investigate whether age plays a role in how much adaptation is observed. Based on prior studies, we expected the children to exhibit adult-like adaptive responses when exposed to perturbed F_1 and F_2 (e.g., [3]). In addition, based on prior studies, we expected younger children to adapt more than older children (e.g., [6]).

2. METHOD

Prior to conducting the study, ethical approval was obtained from the Research Ethics Review

Committee of the Faculty of Arts, University of Groningen (CETO; approval number 82182577). Before participating, parents and children read the information letter, signed the consent form, and filled out a demographics questionnaire together. In the case of younger children (under 12), only the parents did this.

2.1. Participants

Participants were recruited at *Blagenparadijs* - the children's area at one of the biggest music festivals in the Netherlands, *Zwarte Cross*. During three consecutive days, 37 children were recruited, out of which 30 were included in the study. Unfortunately, five children had to be excluded as their consent forms appeared to be incomplete. Additionally, two children were excluded as the data pre-processing stage revealed that their vowel productions were too short to take reliable measurements. The age range of all included children was between 6 and 14 (16 male, and 14 female; see Table 1). All children were native Dutch speakers (1 Dutch-Frisian bilingual). Most children (19 out of 33) were from the province of Gelderland, as that was also where the festival was taking place. None reported any hearing problems, but two reported having dyslexia and an additional two reported to experience stuttering.

Age (in years)	N (F)	N (M)
6	1	-
7	1	1
8	2	5
9	4	4
10	3	3
11	1	2
12	1	-
13	1	-
14	-	1

Table 1: Participants included in the study, separated by sex (M = male, F = female)

2.2. Study design

2.2.1. *Procedure and equipment*

Data collection took place at the festival grounds, inside a mobile laboratory with a sound-dampened booth (dampened to -40dB[1]). The children first completed a short (offline) practice round during which they were asked to practice reading target words. This ensured familiarity with the target words before the experiment began. Children were then positioned in front of a monitor, wearing Sennheiser HD280 Pro headphones to which a Shure WH20XLR headset microphone was attached. Parents generally stayed outside of the van, but were allowed to accompany the child if the child indicated that this would significantly increase their comfort. The microphone and headphones were attached to a Focusrite Scarlett Solo soundcard, which was connected to a HP Z2 Mini G5 Workstation PC.

It is worthwhile to note that the experimental setting itself was controlled — i.e., we tested in a sound booth with high-quality equipment —, but the experimental conditions were not as ideal. This was largely due to the many distractors related to the festival itself: while the children were eager to participate, their attention span was short due to other activities at Blagenparadijs. However, since they were recruited on the spot, this also meant that we were able to include children who otherwise would not have had an opportunity to participate in an experiment at all.

2.2.2. *Formant shifting and target words*

Formants were shifted in real time using the Audapter software (v2.1.012 [8, 9]). We employed a very short experimental paradigm: the participants began with eight trials of veridical auditory feedback ('START' phase), followed by eight trials during which the formants were gradually shifted until reaching the maximum perturbation ('RAMP' phase). The participants then produced 16 trials at the maximum perturbation of a 20% decrease in F_1 and a 15% increase in F_2 ('STAY' phase), followed by a sudden return to veridical feedback ('END') phase for the final eight trials. The experiment thus consisted of only 40 trials (see Figure 1), which is one of the shortest reported (but see Daliri and colleagues [10] who successfully carried out a 30-trial experiment).

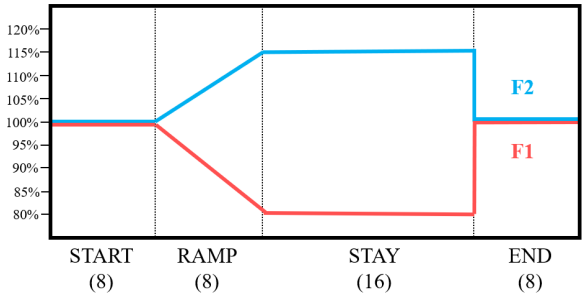


Figure 1: Formant shifting experiment consisting of 40 trials.

Target words in use were four Dutch words with the open-mid front unrounded vowel /ɛ/, namely *pet*, [pɛt] ('cap'); *bed*, [bɛt] ('bed'); *dek*, [dɛk] ('deck'); and *bek*, [bɛk] ('beak'). In all cases, the formant shift resulted in another semantically meaningful word: *pit*, [pɪt] ('(fruit) kernel'), *bid*, [bɪt] ('pray'), *dik*, [dɪk] ('fat') and *bic*, [bɪk] (a well-known brand of pens). The frequency of the first three target words and their shifted counterparts is approximately the same (as determined by the SUBTLEX corpus [11]). The fourth target word ('bek') is much more frequent than its shifted counterpart ('bic'). However, also note that while these are high-frequency words for adults, they might not be as well-known by (especially younger) children.

2.3. Analysis

2.3.1. Pre-processing

Raw F_1 and F_2 values were extracted after being automatically detected by Audapter. Mean F_1 and F_2 in a 50ms window (from 30ms after the start of the vowel until 80 ms after the start of the vowel) were calculated for each trial per participant. Average F_1 and F_2 values per trial were then normalized per participant by subtracting the mean of all START trials from each individual trial and dividing by the standard deviation of all START trials. This allowed us to assess how much each individual trial produced by a participant differed from their (average) trials produced during the START phase. In our analysis, we treated trials as a continuum from 1 to 40, and did not separate them into phases.

2.3.2. Statistical analysis

To assess whether children changed their vowel productions during our short formant perturbation paradigm, we used generalized additive mixed modelling (GAMM; [12]). This approach is well-suited for analyzing formant perturbation data, as it does not assume a linear relationship between two variables.

The statistical analysis was done in R version 4.2.2. [13]. We fit GAMMs using the `bam` function of the *mgcv* package version 1.8-41 [12], following the model-fitting procedure outlined by [14]. We interpreted and visualized the results using the *itsadug* package version 2.4.1 [15]. As the data was not normally distributed and showed heavy tails, we fitted a model using the scaled-t distribution (`family = "scat"`), in line with [14].

For our hypothesis-testing model (in which we aimed to assess whether the children adjusted their vowel productions based on the altered feedback), we used the normalized formant value as the dependent variable and trial number as independent variable. Specifically, we included a (non-linear) smooth over trial, separated by formant type (F_1 and F_2). We included the maximal random effects structure supported by the data, which included a factor smooth over trial (per formant type) for each participant and a factor smooth over trial (per formant type) for each target word, accounting for variation in the non-linear patterns across trials per participant and target word, respectively. In our exploratory analysis, we additionally assessed the possible effect of age and sex.

3. RESULTS

The hypothesis-testing model showed a significant non-linear effect of trial ($p < .001$) for both formants. Children therefore adapted in the expected manner, by increasing their F_1 in response to the 20% F_2 decrease and by decreasing their F_2 in response to the 15% F_1 increase. The pattern of F_1 and F_2 can be seen in Figure 2.

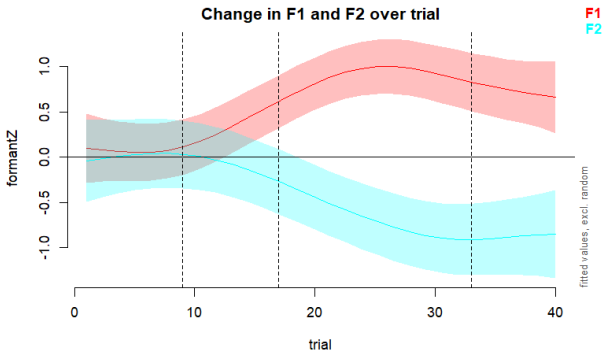


Figure 2: Change in F_1 and F_2 over trial. Shaded areas correspond to the 95% confidence intervals.

We continued with the exploratory analysis, evaluating the effect of age and sex, as indicated above. We first assessed the effect of sex, separated by formant type (F_1 and F_2). The results indicated no significant difference between male and female participants ($p > 0.1$ for both F_1 and F_2). We then assessed whether or not age showed a significant (potentially non-linear) effect per formant type. This revealed no effect of age ($p > 0.1$ for both F_1 and F_2). Finally, we assessed whether age and trial interacted by including a tensor product interaction for trial and age, separated by formant type. This

likewise showed no significant effect of age ($p > 0.1$ for both formants).

4. DISCUSSION

Our study investigated adaptive responses to formant perturbations in a group of 30 Dutch children, aged between 6 and 14. Our goal was to determine whether children adapt to a short experimental paradigm of five minutes when tested at festival grounds. In addition, we investigated whether age is a predictor in how much they adapt (per formant). Our analysis, using generalized additive mixed modeling, showed a significant change in F_1 and F_2 depending on trial ($p < .001$). However, neither sex nor age yielded statistically significant results.

In line with prior studies (e.g., [6], we would have expected some differences in age, especially between the younger participants (aged 6 or younger) and the older participants (aged 9 or older). While plots for F_2 per year did seem to show a trend towards a larger decrease in F_2 for older participants (i.e., more adaptation), this was not statistically significant ($p > 0.5$). Unfortunately, there was a large imbalance in the age distribution of our participants (we only included a single 6-year-old and a single 14-year-old, for example), making any age comparison unreliable.

Our study faces several limitations. First, as stated above, while we tried to recruit children in a broad age range, most children were aged between 8 and 11. Second, while all children could read, not all felt equally comfortable reading during the experiment (although they did complete a short offline training session beforehand), which might have affected the results. Finally, while a short paradigm does work in eliciting adaptive responses, fewer productions can be problematic if any data needs to be discarded. In our case, we had to exclude two participants because there were not enough trials of sufficient length available.

Despite the variability present in children's speech, our results suggest that future studies may employ a similarly shortened paradigm to investigate responses in large(r) groups of children, including teens (ages 10 and up). While testing at large events is not always ideal, it ensures that children of different ages and backgrounds are more easily recruited.

[1] For more information on the mobile laboratory, SPRAAKLAB, please see [16].

5. REFERENCES

[1] J. A. Tourville and F. H. Guenther, "The diva model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, 2011.

[2] J. H. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, 1998.

[3] E. MacDonald, E. Johnson, J. Forsythe, P. Plante, and K. Munhall, "Children's development of self-regulation in speech production," *Current Biology*, vol. 22, no. 2, pp. 113–117, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982211013340

[4] F. van Brenk and H. Terband, "Compensatory and adaptive responses to real-time formant shifts in adults and children," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2261–2270, 2020.

[5] S. T. Cheung, K. Thompson, J. L. Chen, Y. Yunusova, and D. S. Beal, "Response patterns to vowel formant perturbations in children," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2647–2654, 2021.

[6] M. R. van den Bunt, M. A. Groen, S. Frost, A. Lau, J. L. Preston, V. L. Gracco, K. R. Pugh, and L. T. W. Verhoeven, "Sensorimotor control of speech and children's reading ability," *Scientific Studies of Reading*, vol. 22, no. 6, 2018.

[7] C. Coughler, K. Quinn de Launay, D. Purcell, J. Oram Cardy, and D. Beal, "Pediatric responses to fundamental and formant frequency altered auditory feedback: A scoping review," *Frontiers in Human Neuroscience*, vol. 16, 2022.

[8] J. A. Tourville, S. Cai, and F. H. Guenther, "Exploring auditory-motor interactions in normal and disordered speech," in *Proceedings of the 165th Meeting of the Acoustical Society of America, Montreal, Quebec, Canada*, 2013.

[9] S. Cai, M. Boucek, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "A system for online dynamic perturbation of formant frequencies and results from perturbation of the mandarin triphthong /iau/." in *Proceedings of the 8th International Seminar on Speech Production, Strasbourg, France,*, 2008, pp. 65–68.

[10] A. Daliri, E. A. Wieland, S. Cai, F. H. Guenther, and S.-E. Chang, "Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter," *Developmental Science*, vol. 21, no. 2, p. e12521, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12521

[11] E. Keuleers, M. Brysbaert, and B. New, "Subtlex-nl: A new measure for dutch word frequency based on film subtitles," *Behaviour Research Methods*, vol. 42, no. 3, pp. 643–650, 2010.

[12] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Taylor & Francis, 2017.

[13] R Core Team, *R: A Language and Environment for*

Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[14] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447017301377

[15] J. Van Rij, M. Wieling, R. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2017. [Online]. Available: https://cran.r-project.org/web/packages/itsadug

[16] M. Wieling, T. Rebernik, and J. Jacobi, "Spraaklab: a mobile laboratory for collecting speech production data," in *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, 2023.

What does successful L2 vowel acquisition depend on? A conceptual replication

Jurriaan Witteman[1], Ekaterina Karaseva, Niels O. Schiller[1] & James M. McQueen[2]

[1]Leiden University, [2]Radboud University
j.witteman@hum.leidenuniv.nl

ABSTRACT

It has been suggested that individual variation in vowel compactness of the native language (L1) and the distance between L1 vowels and vowels in the second language (L2) predict successful L2 vowel acquisition. Moreover, general articulatory skills have been proposed to account for variation in vowel compactness. In the present work, we conceptually replicate a previous study to test these hypotheses with a large sample size, a new language pair and a new vowel pair. We find evidence that individual variation in L1 vowel compactness has opposing effects for two different vowels. We do not find evidence that individual variation in L1 compactness is explained by general articulatory skills. We conclude that the results found previously might be specific to sub-groups of L2 learners and/or specific sub-sets of vowel pairs.

Keywords: phonetics, speech production, vowels, bilingualism, compactness

1. INTRODUCTION

Acquiring native-like production of a second language (L2) is difficult when the native language (L1) has already been acquired. A basic assumption of all speech learning models is that L2 acquisition is difficult because the neural system has been optimized to process L1 sounds. This can interfere with the processing of L2 sounds.

Processing of L2 sounds is particularly difficult for L2 sounds that do not exist in the L1 but are very similar to an L1 sound. In this case, the L2 learner may simply substitute the L1 neighbouring sound for the two L2 sounds. Therefore, the L2 sounds may deviate from the target (native speaker) sounds. Consider the English vowels /u:/ (as in GOOSE) and /ʊ/ (FOOT). In Dutch, these vowels do not exist, but the similar vowel /u/ (as in MOE 'tired') does. Both English vowels are close to the Dutch vowel, but /u:/ is closer to /u/ than /ʊ/ (Figure 1). Indeed, it is known that Dutch speakers have difficulty contrasting /ʊ/ and /u:/ because they substitute Dutch /u/ for both English vowels [1].

According to the revised Speech Learning Model (SLM-r) [2], when an L2 sound is substituted by an L1 sound, the resulting L2 sound will sound more native-like when the L1 substitute is a better exemplar of the native target sound. Since Dutch /u/ is a better exemplar for English /u:/ than /ʊ/, the corresponding L2 /u:/ is expected to sound more native-like than the L2 /ʊ/.

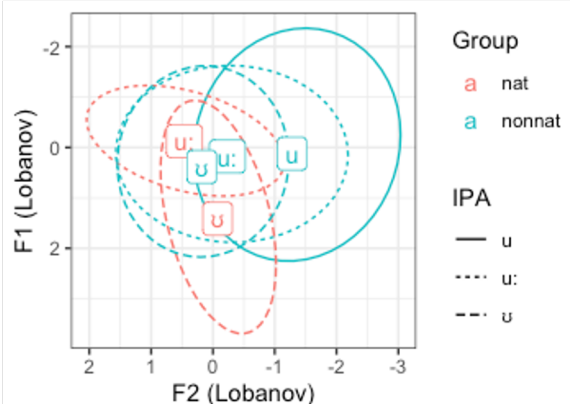


Figure 1. Vowel plot for English FOOT (/ʊ/) and GOOSE (/u:/) and Dutch MOE (/u/) for 6 native (nat) and 123 Dutch (nonnat) speakers of English. Ellipses are based on 0.67 SD's.

L2 vowel acquisition success is highly variable, with some individuals sounding more native-like than others. What might be causes of this individual variation? One hypothesis that can be derived from the SLM-r, is that if the Dutch vowel /u/ for a given individual is closer to the target English vowel in the individual's inter-phonological vowel space, the English token will be produced in a more native-like way.

A previous study [3] suggests an additional factor that could explain variation in L2 vowel production acquisition – variation in how precise individuals produce vowels in the *L1*. Some individuals show more token-to-token variation for the same L1 vowel than others. For instance, in Figure 2 below, vowel plots are shown for two individuals – one individual (left) has a relatively large token-to-token variation (the vowel is relatively 'non-compact') for Dutch /u/ while the individual on the right shows relatively little variation (the vowel is relatively 'compact').

Evidence was found [3] for the hypothesis that individuals with relatively compact L1 vowels produced L2 vowels more like natives than individuals with less compact L1 vowels. According

to [3], better L2 production performance with more compact L1 vowels can be explained by the idea that a more compact L1 vowel space leaves more 'blank space' available for the accommodation of L2 vowels.

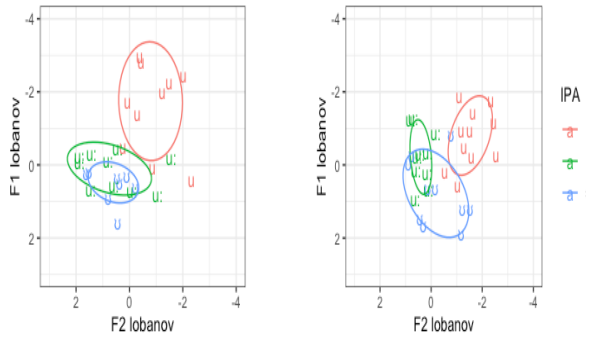


Figure 2. Vowel plot for English FOOT (/ʊ/) and GOOSE (/u:/) and Dutch MOE (/u/) for a relatively imprecise speaker of Dutch /u/ (left) and a relatively precise speaker (right). Ellipses are based on 0.67 SD's.

In line with the SLM-r above, the authors of [3] additionally found that for individuals whose L1 vowel was closer to the target L2 sound, L2 vowel production was more native-like. This result can be explained by the hypothesis that when the L1 competitor vowel in a given individual is a better exemplar of the L2 target vowel, the substitution will result in a more native-like target.
Last, an interaction was found, i.e., as the L1 competitor vowel came closer to the target native vowel for the L2, the influence of L1 vowel compactness on L2 target nativeness became larger. This result was explained as follows: when the distance between an L1 and L2 vowel is smaller, having a more compact L1 vowel is more beneficial in creating 'blank space' for the new L2 vowel then when the distance is large.

Regarding the *cause* of individual variation in L1 compactness, various factors have been proposed. One such factor suggested by [3] is that individual variation in L1 vowel compactness is explained by 'general articulatory skill' in how precisely individuals produce sounds. If true, then we would expect individuals who produce more compact *L1* vowels also to produce more compact *L2* vowels.
In the present study, we conceptually replicate [3] with a much larger sample size (almost 10 times larger), a new language pair (Dutch L1, English L2) and a new vowel pair (English /u:/-/ʊ/). From the literature discussed above we derive the following predictions:

1. Individuals with a more compact L1 /u/ vowel produce the L2 English vowels more accurately (smaller Mahalanobis distance to the native English vowels) than individuals with a less compact L1 /u/ vowel.
2. Individuals whose L1 /u/ is closer to the native speaker target L2 vowel /u:/ or /ʊ/ (smaller Mahalanobis distance) produce that L2 vowel more accurately (smaller Mahalanobis distance to the native speaker vowels).
3. When the distance between the L1 vowel and the English native speaker vowel is smaller, L1 vowel compactness will have a larger negative association with the distance between the L2 vowel and the native speakers than when the distance between the L1 vowel and the native English vowel is large.
4. There is a positive association between the compactness of L1 /u/ and the mean compactness of the L2 English vowels.

2. METHODS

2.1 Participants

A total of n = 123 (91 female) advanced Dutch (L1) learners of English (L2) were included from the first recording session of a longitudinal project investigating individual variation in L2 acquisition. All these participants were university students, had Dutch as their L1, had on average 6 years of high school education in English but had not received extensive explicit instruction in English phonetics and had not experienced immersion in an English-speaking environment. Additionally, six (3 female) native speakers of British (n = 3) and American (n = 3) English were included. All native speakers were monolingual speakers of English, grew up in an English-speaking country and used the contrast FOOT – GOOSE in their native variant of English. None of the participants reported any speech production or speech perception disorders.

2.2. Procedure

The study was approved by the Faculty's ethics committee at Leiden University and all participants gave informed consent. Participants were reimbursed 5-20 euro's for participation in the session (that ranged between 30-180 minutes). All participants read out the *North Wind and the Sun* (NWS) passage three times from a computer screen in a dimly lit sound attenuated booth. Dutch speakers additionally read out loud the Dutch version of NWS three times. Recordings were made using a RØDE NTG2 directional condenser microphone with appropriate

amplification and at 16 bit resolution and a sampling frequency of 44.1 kHz.

2.3 Data analysis

All vowels were manually segmented by students who had received training in phonetics and extensive training in segmenting vowels using a segmentation manual. Vowels were segmented using visual inspection of the spectrogram and oscillogram and by listening to the vowel. F1 and F2 were extracted from vowels that did not contain creaky voice or presence of noise or interruption of the formant tracks. Mean formant frequencies were extracted using a PRAAT [4] script that used the Burg algorithm to calculate the mean F1 and F2 over the middle 50% of the vowel with a time step of 0.01 seconds, a window length of 0.025 seconds and a pre-emphasis of 50 Hz and a formant ceiling of 5,500 Hz.

Formant values that were more than 2 SD's away from the participant's vowel mean F1 or F2 were excluded. Vowel compactness was then calculated by speaker for every vowel as reported in [3]:

(1) $CS_v = SD_{F_1} \times SD_{F_2} \times \pi$

Compactness scores that were more than 2 SD's away from the mean compactness score were excluded. Mahalanobis distances (MD) between a token and a reference distribution were be calculated using the *mahalanobis* function from the *stats* package in R [5]. MDs that were more than 2 SD's away from the mean MD were excluded.

Categorical predictors were mean centered and continuous predictors were Z-normalized to promote convergence and to interpret interactions. For all models, the *lmer* function from the *lme4* package was used. The maximum random-effects structure was obtained by starting with random intercepts for Participant and Word and adding random slopes that increase model fit based on the likelihood ratio test using the *anova* function in R until the model did not converge. Statistical significance of main effects and interactions was tested using the *lmerTest* package.

For analysis involving more than one predictor and interactions, if an interaction is significant, the nature of the interaction was inspected using the *emmeans* function in R.

3. RESULTS

Descriptive statistics concerning MDs are given in Table 1 below. Note that Dutch /u/ indeed is closer to English /u:/ than English /ʊ/.

Table 1

Vowel pair	MD mean (SD)
/u/ - /ʊ/	9.33 (3.06)
/u/ - /u:/	8.37 (3.95)
L2 /ʊ/ - native /ʊ/	3.83 (4.18)
L2 /u:/ - native /u:/	4.65 (4.43)

To test hypotheses 1-3, a multilevel regression analysis was performed with the MD between the native distribution of tokens for a given English vowel and an L2 vowel token as dependent and (1) Z-normalized CSv for /u/, (2) the mean Z-normalized MD between /u/ and the native vowel, and (3) Vowel (/ʊ/ and /u:/) as predictors and Participant and Word as random intercepts and a by-participant random slope for the Vowel factor.

Contrary to expectations, there was (1) no main effect of vowel (2), no main effect of the MD between the L1 vowel /u/ and the native English vowels on production accuracy, (3) no main effect of L1 CSv on production accuracy, and (4) no interaction between L1 vowel compactness and the MD between L1 vowel /u/ and native vowels on production accuracy. However, there was a significant two-way interaction between Vowel and Compactness. In Figure 3 below, it can be observed that while for /ʊ/ there was indeed a positive relationship between L1 compactness and L2 production accuracy, for /u:/ the relationship was *negative*. However, when the interaction was followed up, the negative association for /u:/ did not reach statistical significance (β = -.22, 95 CI = [-0.72, 0.28], T(121) = -.87, p = 0.38) while the positive association for /ʊ/ approached statistical significance (β = 0.41, 95 CI = [-0.01, 0.839], T(113)= 1.93 , p = 0.06)

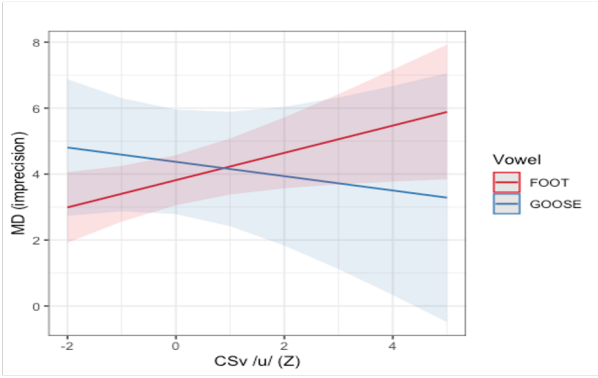


Figure 3. Interaction between Vowel and L1 compactness for L2 production accuracy.

Finally, the Pearson correlation between L1 vowel compactness and L2 vowel compactness was small (r = -0.03) suggesting no linear association between precision of L1 and L2 vowel production (see Figure 4).

4. DISCUSSION

In the present study, a conceptual replication was performed on the speech production part of [3] with an almost 10 times larger sample size (original study: n = 14, present study: n = 123). In contrast to [3], we did not find main effects of compactness or L1-L2 distance on L2 production accuracy, nor an interaction between L1-L2 distance and L1 vowel compactness on the accuracy of L2 production. We did find an interaction between Vowel and Compactness, with the expected positive association between L1 compactness and L2 production for /ʊ/ but a negative association for /u:/. However, neither simple main effect was statistically significant. Finally, as in [3], we did not find a positive association between L1 and L2 vowel compactness.

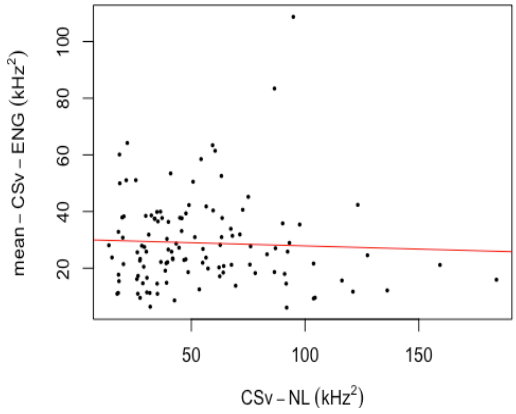


Figure 4. Vowel compactness of the L1 /u/ vowel and the mean compactness of the two L2 English vowels

Several explanations are possible for the lack of correspondence of the results with [3]. First, effects of L1 vowel compactness on L2 vowel production acquisition might be particularly prominent in subgroups of learners who produce L1 vowels with high variability – the L1 vowel compactness scores in the current study (mean: 54.1 kHz^2) were nearly five times smaller than in [3] (mean: 259 kHz^2). Perhaps due a floor effect, therefore, no effect of compactness could be observed. Since the compactness of L1 production was high in the present study, perhaps the effect of L1-L2 distance on L2 production was also smaller than in [3].

A second (non-mutually exclusive) explanation is that effects of L1 vowel compactness and distance of L1 to native target vowels in the to-be-acquired language are particularly prominent among non-advanced learners and less so in more advanced learners such as those in the present study. Perhaps vowel compactness and L1-L2 distance play a role only at the start of L2 vowel acquisition (when the L1 is relatively dominant).

Third, these effects might be limited to specific vowel pairs – in the present work, MD's between native and L2 vowel productions were larger than in [3], suggesting that the vowel contrast in the present study was more difficult than in [3] and that for relatively difficult vowel contrasts the role of L1 compactness might be smaller than for easier vowel contrasts. On the other hand, the difference in direction of the compactness effect for /ʊ/ than /u:/ suggests that the positive effect of L1 compactness found in [3] might particularly apply to L2 vowels for which the L1 competitor vowel is, at the same time, close in the inter-phonological vowel space but also a bad exemplar for the target vowel (i.e., /u/ for /ʊ/).

Fourth, given the small sample size of [3], the effects found might have been inflated [6] and the current larger study might provide more precise estimates.

In the present work, similarly to [3], we did not find a positive linear association between L1 compactness and L2 compactness. This suggests that it is unlikely that general articulatory skills explain variation in the accuracy of vowel production.

Future high-powered (preferably longitudinal) studies using various vowel-contrasts, language combinations and L2 learners in different stages of L2 acquisition are needed, however, to reach more definite conclusions.

5. REFERENCES

[1] Collins, B. & Mees, I. M. (2003). *The phonetics of English and Dutch.* Leiden – Boston.

[2] Flege, J., & Bohn, O. (2021). *The Revised Speech Learning Model (SLM-r).* In R. Wayland (Ed.), Second Language Speech Learning: Theoretical and Empirical Progress (pp. 3-83). Cambridge University Press.

[3] Kartushina, N. & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology,* 5, 1246.

[4] Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.02, retrieved 29 November 2022 from http://www.praat.org/

[5] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[6] Button, K. S. *et al.* (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience,* 14, 365-376.

TALKING CHORALLY ALTERS SPEECH RHYTHM AND INDUCES FLUENCY IN PEOPLE WHO STUTTER, BUT ARE THESE THINGS CONNECTED?

Sophie Meekings[1], Lotte Eijk[1], Santosh Maruthy[2] & Sophie Scott[3]

[1]University of York, [2]All-India Institute of Speech and Hearing, [3] University College London
sophie.meekings@york.ac.uk, lotte.eijk@york.ac.uk, santoshm@aiishmysore.in, sophiescott@ucl.ac.uk

ABSTRACT

When people who stutter speak chorally (in unison with another talker), their speech becomes more fluent. One hypothesis is that stuttering results from difficulty generating rhythmic speech gestures, and choral speech induces fluency by providing an external 'pace-setter' for speech rhythm.

In this study, 20 participants who stutter read aloud, alone and chorally with an experimenter. We measured fluency (percent syllables stuttered) and used envelope modulation spectral (EMS) analysis to derive measures of speech rhythm. Participants stuttered significantly less in the choral condition compared to solo speech, and their speech rhythm also changed: linear discriminant analysis (LDA) identified a combination of 6 EMS metrics which, together, reliably differentiated between choral and solo speech.

However, while some individual EMS metrics correlated with stuttering frequency, the specific rhythmic signature of choral speech identified by the LDA did not predict fluency, suggesting a complex relationship between speech rhythm and stuttering in this task.

Keywords: stuttering, choral speech, speech rhythm

1. INTRODUCTION

Stuttering is a neurological speech production disorder that affects an estimated 1% of adults worldwide [1]. People who stutter (PWS) experience disruptions to their speech in the form of involuntary prolongation, repetition, and blocking of speech sounds. The disorder is highly variable- people vary in how severely they are affected [2], and their disfluency can fluctuate in seemingly unpredictable ways [3]. However, research has identified a handful of situations that temporarily cause a reduction in stuttering in nearly all PWS, including speaking in time with a metronome [4], talking with pitch-shifted or delayed auditory feedback [5], and speaking in unison with another talker (choral speech) [6]. Understanding why these conditions induce fluency may provide insight into the aetiology of the disorder. Here, we focus on choral speech, which has previously been identified as more effective at reducing stuttering than other fluency-inducing manipulations [7], [8].

Since many of these manipulations affect the rate and rhythm of speech [6], it has been hypothesised that stuttering results from a deficit in the initiation and control of rhythmic speech movement [9], [10]. Thus, manipulations such as choral speech and metronome-timed speech induce fluency by providing an external signal that PWS are able to use to time their speech gestures [11].

Envelope Modulation Spectral (EMS) analysis has emerged as a measure of speech rhythm that is well suited to the analysis of disordered speech as it is not affected by typical features of atypical speech, such as prolonged pauses, and can be applied automatically with no assumptions about linguistic content [12]. In this analysis, the amplitude envelope is extracted from the full-band signal and selected frequency bands, low-pass filtered, and Fourier transformed to quantify the amplitude modulation rates that dominate in the signal and in different frequency bands. Various metrics can be extracted from the resulting power spectra that represent different characteristics of speech rhythm.

Liss et al. [12] successfully used discriminant function analysis with EMS to create rhythmic profiles that reliably distinguished different subtypes of dysarthria. These authors identified six metrics that provide information on different aspects of the speech signal, provided in Table 1 below.

Peak frequency	The frequency in Hz of the peak in the spectrum with the greatest amplitude. This identifies the dominant rate of modulation in the signal.
Peak amplitude	The amplitude of the peak frequency (normalised by dividing by the overall amplitude of the spectrum). This indicates how much the rhythm is dominated by the peak frequency.
Energy 3-6 Hz	The sum of energy between 3-6 Hz, normalised. Energy in this band is correlated with intelligibility [14] and captures syllable durations [15].
Energy 0-4 Hz	The sum of energy between 0-4 Hz, normalised. Energy above and below 4Hz was identified by [12] as important predictors of dysarthria type.
Energy 4-10 Hz	The sum of energy between 4-10 Hz, normalised.
Energy Ratio	Energy 0-4 Hz divided by Energy 4-10 Hz.

Table 1. EMS predictors of interest identified by Liss et al., and their meaning.

Applying this approach to stuttering, Dechamma & Maruthy [13] used these six metrics to analyse speech rhythm during choral and solo speech. They found that, when PWS spoke chorally, they stuttered less and their mean peak frequency was higher than in the solo reading condition, while mean peak amplitude was lower during choral speech than solo

reading. This suggests that choral speech does change some aspects of speech rhythm, but it is not yet clear whether these changes cause a reduction in stuttering frequency, or merely co-occur with it. We aim to build on this existing work, using the metrics and discriminant analysis approach identified by Liss et al. [12] to further explore the questions raised by Dechamma & Maruthy [13] by directly testing the relationship between speech rhythm and stuttering frequency.

To examine this, we use linear discriminant analysis to identify the combination of EMS metrics that make up the 'rhythmic signature' of choral speech, and then test whether these metrics also predict stuttering frequency.

2. METHODS

2.1. Participants

21 adults who self-identified as people who stutter took part. They underwent a hearing test using an Amplivox 116 Screening Audiometer with DD45 earphones. Normal hearing was defined as a four-frequency pure tone average of less than 20dB in each ear. One participant's threshold exceeded this level and they were excluded for this reason. Thus, 20 PWS (8 female, 13 male) took part in the study. All participants were adult native British English speakers (mean age: 40 years, s.d. 12 years).

2.2. Task

Participants were asked to speak spontaneously for three minutes, read a passage aloud on their own, and then read a second passage in synchrony with an experimenter. They sat in a soundproofed booth facing a RODE NT1-A one-inch cardoid condenser microphone and wearing Beyerdynamic DT770 Pro circumaural headphones. The microphone was connected to a Windows computer via a Fireface UC high-speed USB audio interface. Their voices were recorded at 44100Hz with 16-bit quantisation using Adobe Audacity 3.0.

For the choral speech task, the experimenter spoke with the participant from outside the booth into an AKG 190E cardoid dynamic microphone, and heard through AKG K240 Studio on-ear headphones. The participant and experimenter were unable to see each other, ensuring that the pair could only use auditory cues to synchronise with each other.

Experiment materials were taken from Riley's Stuttering Severity Instrument IV (SSI-IV) [16]. The order of the read passages was counterbalanced across participants.

2.3. Measures of stuttering severity

Participants' fluency was assessed by a rater who had not been involved in the original experiment and was naive to the experimental manipulation. They performed the evaluation using the methods described in [16]. This gave four measures:

Stuttering frequency: The number of stuttering incidents divided by the total number of syllables uttered.

Stuttering duration: The average length of the three longest stuttering events timed to the nearest tenth of a second.

Naturalness: Perceived speech naturalness, on a scale from 1 (highly natural sounding speech) to 9 (highly unnatural sounding speech).

Baseline severity score: A score out of 46, calculated by taking ratings of stuttering frequency, duration, and physical tics in the spontaneous and solo speech conditions, converting to scale scores and adding them together. As it is rarely possible to produce spontaneous speech chorally, this score can only be calculated for the non-choral condition.

2.4. Envelope Modulation Spectrum metrics

Measures of speech rhythm in the solo and choral reading conditions were derived from the Envelope Modulation Spectrum of the whole signal using a MATLAB script following the methods described in [12], [13]

The full signal was filtered into seven octave bands centred around 125, 250, 500, 2000, 4000, and 8000 Hz. The amplitude envelope was extracted from each of the octave bands and the full-band signal, half-wave rectified and low-pass filtered at 30Hz using a fourth-order Butterworth filter, before being downsampled (to 80Hz, mean subtracted). The power spectrum of each down-sampled envelope was calculated using the Goertzel algorithm and converted to decibels for frequencies up to 10 Hz.

For each of the seven octave bands, and the full-band signal, the MATLAB script calculated six EMS metrics, as described in Table 1: peak frequency, peak amplitude, energy 3-6 Hz, energy 0-4 Hz, energy 4-10 Hz and energy ratio.

4. ANALYSIS

Analysis was carried out in R 4.2.2 [17] using the tidyverse [18], MASS [19], caret [20], klaR [21] and lme4 [22] packages.

4.1. Linear discriminant analysis

The EMS analysis produced 48 variables (six different metrics for each of seven octave bands and the full signal). To identify which, if any, of these variables best characterised the rhythmic differences between choral and solo speech, we ran stepwise linear discriminant analysis. To meet the assumption of equal variance, variables were scaled to have a mean of 0 and a standard deviation of 1.

We identified collinear variables with a pair-wise absolute correlation of 0.8 or above, and removed the variable with the largest mean absolute correlation for each pair. This reduced the predictor set to 25 variables.

The remaining variables were entered into stepwise forward variable selection. Beginning with the variable that most separated the conditions (choral

vs solo), at each step, the variable that best minimised Wilk's lambda was selected, and included in the model if the F-statistic for the change was significant ($p<0.05$). If, after adding a new predictor, any variable no longer significantly contributed to the model ($p<0.2$), it was removed.

This process identified six predictors, which were entered into a linear discriminant analysis to find a weighted linear combination of the predictor variables capable of classifying observations into either the choral or solo speech condition.

The analysis was run across the whole dataset to generate discriminant scores for data visualisation and further analysis, and with leave-one-out cross-validation to estimate the model accuracy. Results for both analyses are reported as the percentage of observations successfully classified.

4.2. Linear mixed-effect regression

We used model comparison with linear mixed effects regression to evaluate main effects of condition on stuttering and speech rhythm, as well as the relationship between stuttering frequency and speech rhythm.

Models were constructed for stuttering frequency, duration, naturalness, mean peak frequency and mean peak amplitude with each variable modelled as a function of condition (contrast-coded as 0.5, -0.5). Then, to explore the relationship between stuttering and speech rhythm, we modelled stuttering frequency (percentage of syllables stuttered) as a function of the linear combination of variables identified by the LDA.

This model was intended to discern whether the specific changes in speech rhythm caused by choral speech also predict fluency. To look at the relationship between speech rhythm and stuttering more generally, we modelled stuttering frequency as a function of peak frequency and peak amplitude (collapsed across all bands). All models included random intercepts for each subject.

These results were compared to the null model (with condition taken out), using a likelihood ratio test.

(1) H1: DV ~ IV + (1 | Subject)
(2) H0: DV ~ 1 + (1 | Subject)

5. RESULTS

5.1. Effect of choral speech on fluency

Assessment for stuttering severity found a wide spread of SSI-IV scores at baseline, ranging from 6 to 44 out of a possible 46 (mean 23.6, s.d. 10.9). However, when participants spoke chorally, they all converged on a relatively fluent speaking style, regardless of baseline severity.

Statistical analysis revealed a significant effect of condition on stuttering frequency ($\beta = -3.96$, SE = 1.6, $\chi^2(1) = 5.9$, $p = 0.015$) and duration ($\beta = -3.71$, SE = 1.49, $\chi^2(1) = 6.07$, $p = 0.01$): participants stuttered significantly less, and the duration of their stutters were shorter, during choral reading compared to solo reading. However, there was no significant effect of condition on perceived naturalness ($\beta = 0.7$, SE = 0.62, $\chi^2(1) = 1.46$, $p = 0.23$); participants' choral speech did not sound any more natural than when they spoke on their own (Figure 1).

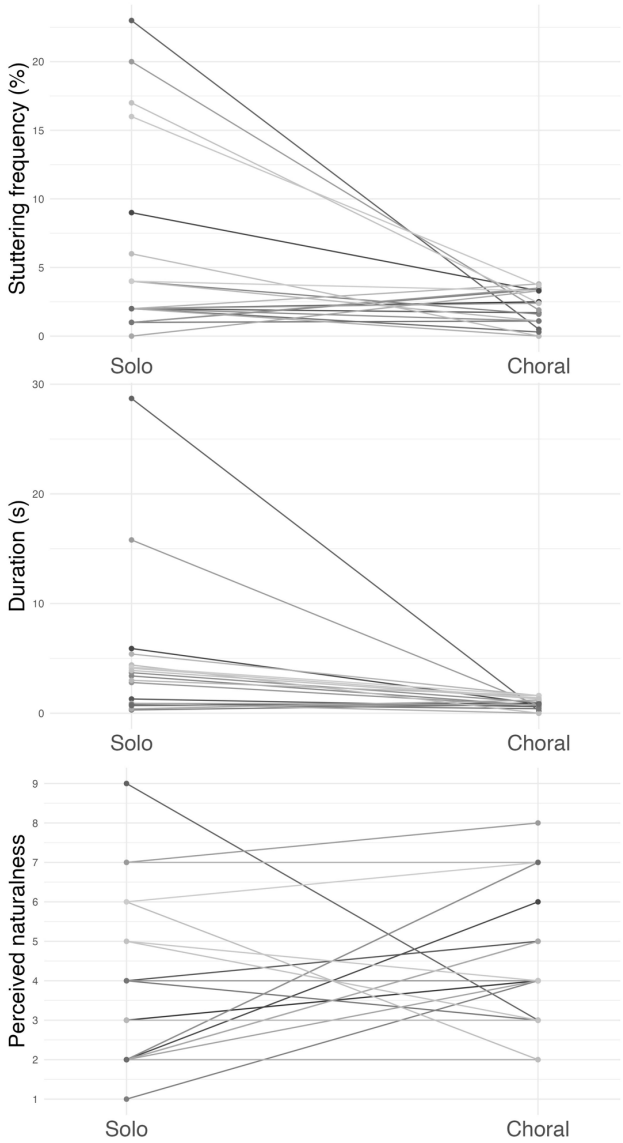


Figure 1: Individual scores for duration, stuttering frequency and perceived naturalness mapped between the solo and choral speaking conditions

5.2. Effects of choral speech on speech rhythm

There was a significant main effect of condition on peak frequency ($\beta = 0.22$, SE = 0.58, $\chi^2(1) = 13.67$, $p = 0.0002$); peak frequency was higher in choral speech than in solo speech. However, there was no significant effect of condition on peak amplitude ($\beta = 0.39$, SE = 2.96, $\chi^2(1) = 0.0176$, $p = 0.89$).

Stepwise variable selection identified six predictors: peak frequency in three octave bands (125, 2000 and 1000 Hz), energy between 4-10 Hz in the 4kHz octave band, full band energy between 3-6Hz and energy ratio in the 8kHz octave band, with

peak frequency in the 125Hz band as the most important predictor. Results are shown in Table 2.

Metric	Band	Λ	F	p	F diff	p diff
Peak frequency	**125**	0.88	5.38	0.026	5.38	0.026
Energy 4-10 Hz	**4000**	0.82	4.04	0.026	2.50	0.122
Peak frequency	**2000**	0.78	3.38	0.029	1.87	0.180
Energy 3-6 Hz	**Full band**	0.70	3.78	0.012	4.12	0.050
Energy ratio	**8000**	0.64	3.85	0.007	3.17	0.084
Peak frequency	**1000**	0.59	3.80	0.005	2.64	0.113

Table 2: Predictors identified by stepwise variable selection, with F-statistics and p-values for the significance of the overall model, and the difference between models when the predictor is included vs excluded.

Using these variables, the linear discriminant analysis was able to predict which condition each speech sample belonged to with 80% accuracy (leave-one-out cross-validation: 67.5%). Figure 2 shows the linear discriminant scores compared to the actual classifications. Four observations (out of 20) from each condition were mis-classified.

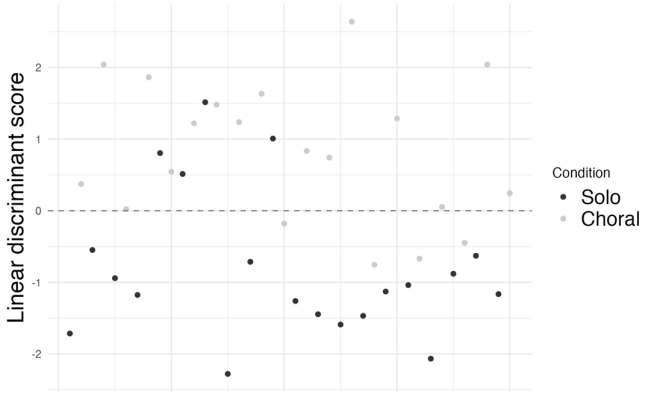


Figure 3: Performance of the linear discriminant. Observations above the dotted line were classified as choral speech; those below the line were classified as solo speech. The colour of the data points shows the actual class of the observation.

5.3. Association between fluency and rhythm metrics

There was a significant effect of peak frequency on stuttering frequency (β = -1.25, SE = 0.35, $\chi^2(1)$ = 12.367 , $p < 0.001$): peak frequency reduced as stuttering frequency increased. However, peak amplitude did not significantly predict stuttering frequency (β = -0.01, SE = 0.01, $\chi^2(1)$ = 3.11 , p = 0.08).

The linear combination of six variables identified by the LDA did not significantly predict stuttering frequency ($\chi^2(6)$= 7.93, p = 0.24), nor were any of the individual components significant. Estimates, standard errors, t-statistics and p-values for each of the 6 variables are given in Table 3.

Metric	Band	β	SE	df	t	p
Peak frequency	**125**	-6.16	4.55	24.9	-1.36	0.188
Energy 4-10 Hz	**4000**	-0.01	0.20	32.8	-0.03	0.978
Peak frequency	**2000**	-1.28	1.81	31.1	-0.71	0.483
Energy 3-6 Hz	**Full band**	-23.3	64.9	28.1	-0.36	0.722
Energy ratio	**8000**	0.01	0.10	30.8	0.05	0.958
Peak frequency	**1000**	-1.99	1.36	32.0	-1.47	0.152

Table 3: Output of linear mixed effects analysis. Degrees of freedom estimated using the Satterthwaite approximation.

6. DISCUSSION

In keeping with previous studies, we found that choral speech resulted in changes to speech rhythm and stuttering behaviour, compared to solo reading. Speaking chorally reduced stuttering frequency and duration in our participants, regardless of baseline severity. However, there was no effect on perceived speech naturalness, perhaps because the rhythmic changes caused by choral speech sound as unnatural as stuttering.

Our attempt to characterise the rhythmic profile of choral speech using EMS metrics met with some success. While we found no relationship between peak amplitude and condition or stuttering frequency, choral speech did result in an increase in mean peak frequency compared to solo speech, consistent with [13], and mean peak frequency was inversely correlated with stuttering frequency.

Linear discriminant analysis identified six EMS metrics that, in combination, significantly predicted which condition (solo or choral) an observation belonged to. Despite the association between overall peak frequency and stuttering frequency, this combination of metrics did not significantly predict stuttering severity compared to the null hypothesis.

Overall, it appears that choral speech and induced fluency are both related to increases in the dominant rate of modulation (i.e., peak frequency), but choral speech is further characterised by more subtle changes in speech rhythm that are not correlated with stuttering frequency. As the sample size was relatively small, these results should be interpreted with caution. A further drawback of our study is that we did not test a control group of typical speakers to contrast with PWS. Future work could test the hypothesis that PWS converge on typical speech rhythms during choral speech by comparing the ability of discriminant function analysis to separate PWS and controls when they speak alone with when they speak chorally.

7. REFERENCES

[1] E. Yairi and N. Ambrose, 'Epidemiology of stuttering: 21st century advances', *J. Fluen. Disord.*, vol. 38, no. 2, pp. 66–87, Jun. 2013, doi: 10.1016/j.jfludis.2012.11.002.

[2] R. I. Lanyon, 'The Measurement of Stuttering Severity', *J. Speech Hear. Res.*, vol. 10, no. 4, pp. 836–843, Dec. 1967, doi: 10.1044/jshr.1004.836.

[3] O. Bloodstein, 'A Rating Scale Study Of Conditions Under Which Stuttering Is Reduced Or Absent', *J. Speech Hear. Disord.*, vol. 15, no. 1, p. 29, Mar. 1950, doi: 10.1044/jshd.1501.29.

[4] R. Hanna and S. Morris, 'Stuttering, Speech Rate, and the Metronome Effect', *Percept. Mot. Skills*, vol. 44, no. 2, pp. 452–454, Apr. 1977, doi: 10.2466/pms.1977.44.2.452.

[5] J. Kalinowski, J. Armson, A. Stuart, and V. L. Gracco, 'Effects of Alterations in Auditory Feedback and Speech Rate on Stuttering Frequency', *Lang. Speech*, vol. 36, no. 1, pp. 1–16, 1993, doi: 10.1177/002383099303600101.

[6] G. Andrews, P. M. Howie, M. Dozsa, and B. E. Guitar, 'Stuttering: speech pattern characteristics under fluency-inducing conditions.', *J. Speech Hear. Res.*, vol. 25, no. 2, pp. 208–16, Jun. 1982.

[7] W. Johnson and L. Rosen, 'Studies in the psychology of stuttering: Effect of certain changes in speech pattern upon frequency of stuttering', *J. Speech Hear. Res.*, vol. 2, pp. 105–109, 1937.

[8] M. Kiefte and J. Armson, 'Dissecting choral speech: Properties of the accompanist critical to stuttering reduction', *J. Commun. Disord.*, vol. 41, no. 1, pp. 33–48, 2008, doi: 10.1016/j.jcomdis.2007.03.002.

[9] E. R. Brayton and E. G. Conture, 'Effects of Noise and Rhythmic Stimulation on the Speech of Stutterers', *J. Speech Lang. Hear. Res.*, vol. 21, no. 2, p. 285, Jun. 1978, doi: 10.1044/jshr.2102.285.

[10] A. C. Etchell, B. W. Johnson, and P. F. Sowman, 'Behavioral and multimodal neuroimaging evidence for a deficit in brain timing networks in stuttering: a hypothesis and theory', *Front. Hum. Neurosci.*, vol. 8, 2014, Accessed: Jan. 05, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00467

[11] M. E. Wingate, 'Sound and Pattern in "Artificial" Fluency', *J. Speech Hear. Res.*, vol. 12, no. 4, pp. 677–686, Dec. 1969, doi: 10.1044/jshr.1204.677.

[12] J. M. Liss, S. LeGendre, and A. J. Lotto, 'Discriminating Dysarthria Type From Envelope Modulation Spectra', *J. Speech Lang. Hear. Res.*, vol. 53, no. 5, pp. 1246–1255, Oct. 2010, doi: 10.1044/1092-4388(2010/09-0121).

[13] D. Dechamma and S. Maruthy, 'Envelope modulation spectral (EMS) analyses of solo reading and choral reading conditions suggest changes in speech rhythm in adults who stutter', *J. Fluen. Disord.*, vol. 58, pp. 47–60, Dec. 2018, doi: 10.1016/j.jfludis.2018.09.002.

[14] T. Houtgast and H. J. M. Steeneken, 'A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria', *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985, doi: 10.1121/1.392224.

[15] T. Arai and S. Greenberg, 'The temporal properties of spoken Japanese are similar to those of English', in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, Sep. 1997, pp. 1011–1014. doi: 10.21437/Eurospeech.1997-355.

[16] G. D. Riley, 'A Stuttering Severity Instrument for Children and Adults', *J. Speech Hear. Disord.*, vol. 37, no. 3, p. 314, 1972, doi: 10.1044/jshd.3703.314.

[17] R. C. Team, 'R: A language and environment for statistical computing (Version 4.0. 5)[Computer software]', *R Found. Stat. Comput.*, 2021.

[18] H. Wickham *et al.*, 'Welcome to the Tidyverse', *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, Nov. 2019, doi: 10.21105/joss.01686.

[19] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

[20] M. Kuhn, 'Building Predictive Models in R Using the caret Package', *J. Stat. Softw.*, vol. 28, pp. 1–26, Nov. 2008, doi: 10.18637/jss.v028.i05.

[21] C. Weihs, U. Ligges, K. Luebke, and N. Raabe, 'klaR Analyzing German Business Cycles', in *Data Analysis and Decision Support*, D. Baier, R. Decker, and L. Schmidt-Thieme, Eds. Berlin, Heidelberg: Springer, 2005, pp. 335–343. doi: 10.1007/3-540-28397-8_36.

[22] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, 'Fitting linear mixed-effects models using lme4', *J. Stat. Softw.*, vol. 67, no. 1, Oct. 2015, doi: 10.18637/jss.v067.i01.

SYLLABLE POSITION EFFECTS WITH FRENCH AND SPANISH /l/

Laura Colantoni, Alexei Kochetov, Jeffrey Steele

University of Toronto
laura.colantoni@utoronto.ca; al.kochetov@utoronto.ca; jeffrey.steele@utoronto.ca

ABSTRACT

Gradient reduction in coda consonant articulation occurs cross-linguistically. We investigate here the possibility of lesser constriction in French and Spanish lateral codas compared to onsets, conditioned by syllable context and language variety. Anterior and posterior contact in syllable-initial and -final singleton /l/ and in labial-/l/ or /l/-labial clusters produced by four French (2 each from France and Quebec) and seven Spanish speakers (5 Argentine, 1 Cuban, 1 Peninsular) was measured in isolated words and carrier sentences using electro-palatography (EPG). Reduction in anterior contact occured only with the Quebec French speakers. A greater degree of syllable-final posterior contact reduction was observed, consistently for Quebec but not European French speakers and for all Spanish speakers. These findings reveal gradient onset-coda /l/ asymmetries in both languages with Quebec French differing from the other Romance varieties studied in degree of weakening. Furthermore, at least for Spanish, coda /l/ lenition differs from that of other consonants.

Keywords: syllable position, gestural reduction, lateral, French, Spanish.

1. INTRODUCTION

1.1. Syllable position effects

Consonant articulations are known to be subject to gradient reduction in syllable-final position [5, 20]. This phenomenon, referred to as gestural 'syllable position effects' [5], has been attributed to the different ways onset and coda consonant gestures are coordinated with following versus preceding vowels – in-phase (synchronously) versus out-of-phase (sequentially), respectively [4, 23]. While typically variable and non-perceptible, these small-scale differences can develop diachronically into larger-scale positional allophonic variation. One of the most-studied examples of this phenomenon is English /l/ allophony, where the syllable-initial 'clear' lateral is produced with a full alveolar closure (and little or no velarization), while the syllable-final 'dark' variant shows a weaker, or even absent, alveolar closure (and considerable velarization/retraction of the tongue back) [5, 15, 36, 38]. Much less articulatory work has been done on positionally conditioned lateral variation in Romance languages, with cross-language articulatory studies pointing to relatively much smaller-scale differences as well as language- or dialect-particular patterns [14, 30, 32, 33]. This contrasts with abundant theoretical and dialectological research that documents that French and Spanish coda /l/ may vocalize, be elided or fully assimilate to a following consonant (see 1.2 & 1.3).

We examine here possible onset-coda lateral constriction asymmetries in these two languages, both of which are described as having a clear /l/ in onsets but which differ in the way in which singleton /l/ and /l/ in clusters is realized (rhotacization and vocalization in Spanish varieties but not in French; [17]). Specifically, we use electropalatography (EPG) to investigate differences in the amount of linguopalatal contact for onset and coda lateral variants. Based on previous research, we hypothesize that /l/ in both languages will exhibit reduction in contact syllable-finally conditioned by syllable type (larger reduction in word-medial clusters than in singleton word-final codas) and dialect (more reduction in Quebec than European French and in Caribbean than non-Caribbean Spanish).

1.2. French /l/

The French lateral has a wide positional distribution, occurring in both onsets (e.g., *lapin* 'rabbit', *aller* 'go') and codas (e.g., *final* 'final') including in consonant clusters (e.g., word-initial: *classer* 'sort'; word-medial: *conflit* 'conflict', *filtrer* 'filter'; word-final: *film* 'film'). Articulatory studies have shown that /l/ is primarily apico-alveolar [6, 9, 34, 37]. Little variability in the articulation of the French lateral is typically reported [12]. In terms of place of articulation, there is no English-type allophonic variation with French /l/ being described as 'clear' or 'light' (e.g., [39]) and acoustic analyses revealing no major effects of syllable or word position on F2 ([1, 6, 25, 31]). However, effects of flanking consonants and vowels on the degree of anteriority have been reported including relatively fronted realizations before a high front vowel or next to alveodental /t d n/ [9, 25, 31, 34] and effects of the anteriority of the preceding vowel on mean F2 [1]. Finally, weakening has been reported including deletion before /j/ (e.g., *milieu* /miljø/, [mijø] 'milieu'; [39, 40]),

intervocalically in fast speech (e.g., *allez* /ale/, /ae/ 'go'; [40]), and syllable-finally in functional words and obstruent-liquid clusters (e.g., *il* /il/ [i] 'he'; *table* /tabl/ [tab] 'table'; [11]).

1.3. Spanish /l/

The Spanish lateral also has a wide distribution. Singleton /l/ is found in word-initial and -medial onsets (e.g., *lana* 'wool'; *sala* 'room') and word-final codas (e.g., *sal* 'salt'). Obstruent-lateral and lateral-consonant clusters occur in syllable onset (e.g., *blanco* 'white') and coda (e.g., *caldo* 'broth'), respectively, with the latter being restricted to word-medial position. As opposed to English, coda /l/ is not velarized [24, 31] but assimilates in place to a following coronal consonant with full assimilation resulting in gemination in Caribbean dialects [16, 17]. Coda /l/ realization varies widely across Spanish dialects including possible rhotacized and vocalized variants [2, 17]. Obstruent-lateral clusters are usually realized as such, however, medial epenthetic vowels have been reported, albeit at lower frequencies than in obstruent-rhotic clusters [7]. Finally, laterals followed by a high vowel (e.g., *utensilio* 'utensil') palatalize in some dialects [35, 44], although such articulations differ from that of the language's palatal lateral [22]. Acoustic studies reveal that the lateral has a more constricted articulation in word-initial than in word-medial and -final positions and that it is more fronted in tonic and initial positions when compared to unstressed final position [26]. Weakening in final position has been confirmed by articulatory studies [28] and has been shown to be gradient, with greater weakening in word-medial codas than in absolute word-final position. In onsets, the Spanish lateral tends to be alveolar [22, 29]. Place of constriction is affected by flanking vowels with more retracted articulations in back-vowel contexts [22].

2. METHOD

2.1. Speakers

The data come from a corpus of previously collected EPG recordings [18]. The current sample includes 11 speakers – four for French and seven for Spanish. All but one (SPA5) were female. The French speakers were 25–29 years old, the Spanish speakers 23–42 years. Among the French speakers, two were from France (FRF1: Cherbourg, FRF2: Clermont-Ferrand) and two from Quebec, Canada (FR1Q1: Chicoutimi, FR2Q2: St-Jean-sur Richelieu). The Spanish participants were from Buenos Aires, Argentina (SPA1-5), Havana, Cuba (SPC1), and Madrid, Spain (SPS1). All participants lived in Canada at the time of testing (mean years of residence: European French 10; Spanish 5) and spoke English as a second language. However, they all continued to speak their L1 at home, at work (the majority were teachers/translators of their native language), and in social situations.

2.2. Materials

Target /l/ appeared syllable-initially and -finally as a singleton consonant or in a cluster with a labial consonant. As the materials were drawn from an existing corpus of EPG recordings, the number of items was not fully balanced across conditions. For French, the list consisted of 20 lexical items: 9 words with singleton onset /l/ (e.g., *lapin* 'rabbit', *collègue* 'colleague'), 7 words with /l/ in an onset cluster (e.g., *flocon* 'flake', *meubler* 'furnish'), 3 words with singleton coda /l/ (e.g., *final* 'final'), and a single word with coda /l/ in a cluster (*illisible* 'illegible'). For Spanish, the list consisted of 27 lexical items: 9 words with singleton onset /l/ (e.g., *laca* 'lacquer', *absoluto* 'absolute'), 9 words with /l/ in an onset cluster (e.g., *flaco* 'thin', *Biblia* 'Bible'), 5 words with a singleton coda /l/ (e.g., *final* 'final'), and 4 words with a coda /l/ cluster (e.g., *calmante* 'calming'). Each word was produced in isolation and in a carrier phrase (French: *Dis__encore une fois* 'Say__again'; Spanish: *Digo__otra vez* 'I say__again'). The French utterances were produced 4-6 times (839 total tokens; 210 tokens per speaker on average). The Spanish carrier phrase utterances were produced 9 times, while isolated words were produced twice (1479 total tokens; 211 tokens per speaker on average). Adjacent vowel context and stress were not controlled for but were examined in the statistical analysis.

2.3. Instrumentation and analysis

EPG is a common method for investigating contact patterns in coronal consonants including laterals [28, 32, 36]. The recordings for the study were made using the WinEPG system [43] at a sampling rate of 100 Hz. A custom-made palate with 62 electrodes was made for each participant. Those for French participants were of the Articulate model [42] whereas those for the Spanish speakers were of the older Reading-style. While the former palate can have somewhat better coverage of dental and velar places, both devices are similar at measuring contact differences within places [19]. Both artificial palates have a grid of 62 electrodes that can be represented with 8 columns and 8 rows (with the first row containing only 6 electrodes). The first four rows correspond to the denti-alveolar region where /l/ is typically produced [13].

The data were annotated based on the waveform and spectrogram using the Articulate Assistant

software [43] with boundaries for /l/ marked at the onset and offset of the closure. For each token, linguopalatal contact values ('1' or '0' for each electrode) were automatically extracted from the frame of maximum contact. Following [28], we examined two variables, namely, the amount of contact in the first and last four rows of the palate, Qa4 and Qp4 (Quotient of maximum activation over the anterior or posterior regions of the palate, from 0.00 to 1.00), calculated as the number of contacts activated divided by the total number of contacts in the region (30 and 32, respectively).

The data were analysed using linear mixed effects regression (LMER) models implemented with the lme4 package [3] using R [27] separately for Qa4 and Qp4. Fixed factors were Language (French, Spanish), Position (onset, coda), Type (isolated word, carrier sentence), and Structure (singleton, cluster); random intercepts were included for Vowel Context (front, back), Word, and Speaker. In cases of non-convergence, the model was simplified by removing Vowel Context. For each analysis, likelihood ratio tests were used to compare the full model to a nested model excluding the factor of interest, employing the Anova() function of the lmerTest package [21]. Pairwise comparisons and posthoc tests (with a Bonferroni correction for multiple comparisons) were performed using the phia package [10].

3. RESULTS

3.1. Positional differences in anterior contact: Between-language comparison

To examine patterns in the amount of contact in the anterior part of the palate (Qa4), we submitted the data to an LMER model containing interactions of the full set of fixed factors (Language, Position, Type, Structure). The model produced significant interactions of Language and Position (p <.05), Language and Type (p <.05), Type and Structure (p <.05), and Position and Type (p <.0001). The top panel of Figure 1 illustrates the variation in Qa4 as a function of these factors. Given the significant interactions with Type, we performed separate analyses for isolated words and those in carrier phrases. The first model produced significant effects of Structure (more contact in singleton /l/ than in clusters; p <.05) and a significant Language and Position interaction. The latter was due to Spanish speakers producing more contact in the anterior region than French speakers, however, only in codas (p <.05). There were no other significant differences. The model for the carrier phrases produced a significant effect of Position (p <.0001) and a significant interaction of Language and Position (p <.05). A posthoc test showed that positional differences were limited to French speakers, who produced /l/ with less contact in the anterior region in codas than onsets (p <.0001).

3.2. Positional differences in posterior contact: Between-language comparison

A full model for the amount of posterior contact (Qp4) produced a 4-way interaction of Language, Position, Type, and Structure (p <.0001). Qp4 differences among these factors are illustrated in the bottom panel of Figure 1. As with Qa4, we performed two separate analyses by Type. The model for isolated words produced a significant effect of Position (p <.0001): /l/ was produced with less contact in the posterior region in codas. There were no other significant effects or interactions, indicating that positional differences held for both language groups and both syllable structure conditions. The model for carrier phrases resulted in significant effects of Position and Structure (both p <.0001). Again, the lateral was realized with less contact in codas than in onsets; in addition, /l/ was produced with less contact in clusters than as a singleton consonant. The lack of interactions with Language suggest that these differences held for both groups.

3.3. Individual positional differences in Qa4 and Qp4

As discussed above, the French group showed the expected positional effect in both the anterior (in carrier phrases) and posterior portions of the palate, while onset-coda asymmetries were observed for the Spanish group in the posterior portion alone. A closer examination of individual results, however, showed that the French positional effect was due to the Quebec speakers. Specifically, FRQ1 and FRQ2 exhibited relatively large-scale coda reduction in anterior (Qa4 0.21 and 0.15, respectively) and posterior contact (Qp4 0.18 and 0.16). In contrast, European French speakers FRF1 and FRF2 produced no such differences in Qa4 (in fact, negative -0.03 and -0.08) and relatively small differences in Qp4 (0.07 and 0.08). Individual models testing for Position, summarized in Table 1, revealed that differences were significant for the Quebec speakers but not for their European counterparts. The same analyses for the Spanish group showed relative uniformity. There were no significant positional effects in anterior contact for any of these speakers (with overall differences ranging from -0.02 to 0.05, approaching significance for SPA2 and SPS1). At the same time, these speakers showed the expected significant positional differences in posterior contact, albeit not as high in magnitude (ranging from 0.06 to 0.11) as for the Quebec French speakers.

Figure 1: Boxplot of anterior (Qa4; top) and posterior contact (Qp4; bottom) by Position and Type by Language group (FR = French, SP = Spanish).

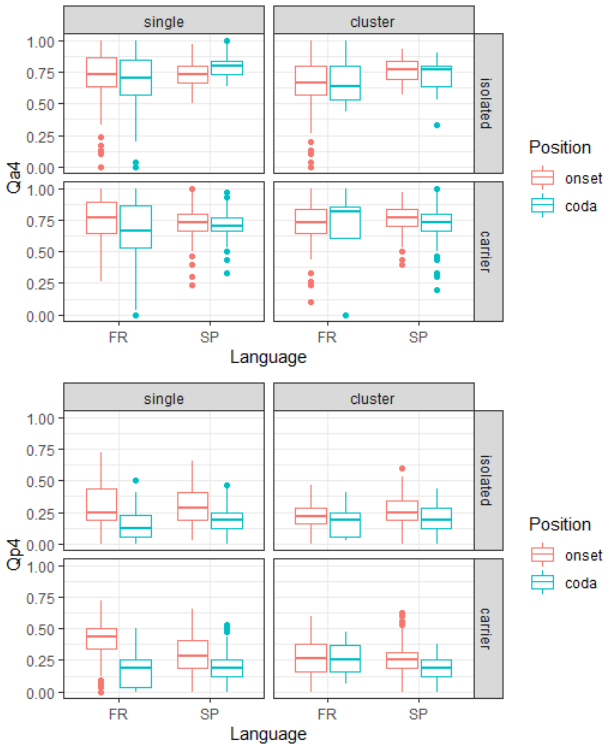


Table 1: A summary of individual model comparisons for Position in anterior (Qa4) and posterior contact (Qp4); '***' <0.001, '**' <0.01, '*' <0.05, 'n.s.' = not significant.

Speaker	Position	
	Qa4	Qp4
FRF1	n.s.	n.s.
FRF2	n.s.	n.s.
FRQ1	*** (onset > coda)	** (onset > coda)
FRQ2	** (onset > coda)	*** (onset > coda)
SPA1	n.s.	* (onset > coda)
SPA2	n.s.	* (onset > coda)
SPA3	n.s.	* (onset > coda)
SPA4	n.s.	* (onset > coda)
SPA5	n.s.	** (onset > coda)
SPC1	n.s.	** (onset > coda)
SPS1	n.s.	** (onset > coda)

4. DISCUSSION

For the Quebec French and non-European varieties of Spanish but not the European French speakers, the present findings are consistent with the proposed cross-linguistic existence of gradient gestural syllable effects as well as with language-specific patterns of coda consonant reduction. The absence of differences among our Spanish speakers is surprising given previous reports of full assimilation of laterals to following consonants in Caribbean varieties [16] and the overall tendency for higher rates of coda consonant reduction in Argentine and Cuban Spanish vis-à-vis Peninsular Spanish [17]. The present study also reveals the possible effects of speech type (isolated versus contextualized), phonetic context (singleton versus clusters), and language variety. The effect of type is in keeping with much previous research that has regularly shown greater lenition in more fluent, contextualized speech [e.g., 41]. The finding of less posterior contact in clusters than with singleton /l/ in carrier phrases is consistent with reports of shorter duration and partial devoicing of laterals in voiceless-obstruent-lateral clusters [8, 25]. In the future, a more balanced dataset would allow to determine whether such differences are related to obstruent voicing or rather syllabic structure.

5. REFERENCES

[1] Ahn, M. 2016. A comparative study of word-final laterals: General American vs. Metropolitan French. *Language Research* 52(2), 197–212.

[2] Alba, O. 1979. Análisis fonológico de /r/ y /l/ implosivas en un dialecto rural dominicano. *Boletín de la Academia Puertorriqueña de la lengua española* 7, 1–18.

[3] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Grothendieck, G. 2017. lme4 package, version 1.1–13 [Computer software].

[4] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

[5] Browman, C. P., Goldstein, L. 1995. Gestural syllable position effects in American English. In: Bell-Berti, F., Raphale, L. J. (eds), *Producing Speech: Contemporary Issues*. AIP Press, 19–33.

[6] Chafcouloff, M. 1972. Recherches sur la structure acoustique du /l/ et ses correlations articulatoires. *Travaux de l'Institut de phonétique d'Aix* 1, 101–111.

[7] Colantoni, L., Steele, J. 2005. Phonetically-driven epenthesis asymmetries in French and Spanish obstruent-liquid clusters. In: Guess, R., Rubin, E. (eds), *Theoretical and Experimental Approaches to Romance Linguistics*. John Benjamins, 77–96.

[8] Colantoni, L., Steele, J. 2011. Synchronic evidence of diachronic change: Voicing and duration in French and Spanish stop-liquid clusters. *Canadian Journal of Linguistics* 56(2), 147–177.

[9] Dart, S. N. 1998. Comparing French and English coronal consonant articulation. *Journal of Phonetics* 26, 71–94.

[10] De Rosario-Martinez, H. 2015. Package 'phia', available at https://github.com/heliosdrm/phia.

[11] Detey, S., Durand. J., Laks, B., Lyche, C. (eds) 2016. *Varieties of Spoken French*. Oxford University Press.

[12] Fagyal, Z., Kibbee, D., Jenkins, F. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

[13] Gibbon, F., Nicolaidis, K. 1999. Palatography. In: Hardcastle, W., Hewlett, N. (eds), *Coarticulation: Data, Theory and Techniques.* Cambridge University Press, 229–245.
[14] Gick, B., Campbell, F., Oh, S., Tamburri-Watt, L. 2006. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics* 34(1), 49–72.
[15] Giles, S. B., Moll, K.L. 1975. Cinefluorographic study of selected allophones of English /l/. *Phonetica* 31, 206–227.
[16] Guitart, J. 1976. *Markedness and a Cuban Dialect of Spanish*. Georgetown University Press.
[17] Hualde, J. I. 2015 *Los sonidos del español*. Cambridge University Press.
[18] Kochetov, A., Colantoni, L., Steele, J. 2017a. The Cross-Language Articulatory Database (CLAD). University of Toronto, http://clad.chass.utoronto.ca/.
[19] Kochetov, A., Colantoni, L., Steele, J. 2017b. A comparison of Articulate and Reading EPG palates: Capturing place/manner contrasts. Poster presented at the *11th International Seminar on Speech Production (ISSP 2017)*, October 16-19, 2017, Tianjin, China.
[20] Krakow, R. A. 1989. *The Articulatory Organization of Syllables: A Kinematic Analysis of Labial and Velar Gestures.* Ph.D. dissertation, Yale.
[21] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82, 1–26.
[22] Martínez Celdrán, E., Fernández Planas, A. M. 2013. *Manual de fonética Española*. Ariel.
[23] Nam H., Goldstein, L., Saltzman, E. 2009. Self-organization of syllable structure: a coupled oscillator model. In: Pellegrino, F., Marisco, E., Chitoran, I., Coupé, C. (eds), *Approaches to Phonological complexity*. Mouton de Gruyter, 299–328.
[24] Navarro Tomás, T. 1972. *Manual de pronunciación española*. Consejo Superior de Investigaciones Cientificas.
[25] O'Shaughnessy, D. 1982. A study of French spectral patterns for synthesis. *Journal of Phonetics* 10, 377–399.
[26] Quilis, A. 1981. *Fonética acústica de la lengua española*. Editorial Gredos.
[27] R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. www.R-project.org [computer program]
[28] Ramsammy, M. 2021. Contextual reduction of word-final /l/in Spanish: An EPG study. *Journal of the International Phonetic Association* 1–28.
[29] Recasens, D. 1986. *Estudis de fonètica experimental del català oriental central*. PAM.
[30] Recasens, D. 2004. Darkness in [l] as a scalar phonetic property: Implications for phonology and articulatory control. *Clinical Linguistics & Phonetics* 18(6-8), 593–603.
[31] Recasens, D. 2012. A cross-language acoustic study of initial and final allophones of /l/. *Speech Communication* 54, 368–383.
[32] Recasens, D., Espinosa, A. 2005. Articulatory, positional and coarticulatory characteristics for clear /l/ and dark /l/: Evidence from two Catalan dialects. *Journal of the International Phonetic Association* 35(1), 1–25.
[33] Recasens, D., Fontdevila, J., Pallarès, M. D. 1996. Linguopalatal coarticulation and alveolarpalatal correlations for velarized and non-velarized /l/. *Journal of Phonetics* 24, 165–185.
[34] Rochette, C. E. (1973) *Les groupes de consonnes en français: Étude de l'enchainement articulatoire à l'aide de la radiocinématographie et de l'oscillographie*. Klincksieck and Les Presses de l'Université Laval.
[35] Rost Bagudanch, A. 2014. An overview of yeismo: A process finished or under construction? *Revista Internacional de Linguistica Iberoamericana* 12(1), 141–163.
[36] Scobbie, J. M., Pouplier, M. 2010. The role of syllable structure in external sandhi: An EPG study of vocalisation and retraction in word-final English /l/. *Journal of Phonetics* 38(2), 240–259.
[37] Simon, P. (1967) *Les consonnes françaises (mouvements et positions articulatoires à la lumière de la radiocinématographie).* Klincksieck
[38] Sproat, R., Fujimura, O. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21, 291–311.
[39] Tranel, B. 1987. *The Sounds of French : An Introduction*. Cambridge University Press.
[40] Walter, H. 1982. *Enquête phonologique et variétés regionals du français*. Presses universitaires de France.
[41] Warner, N., Tucker, B. V. 2011. Phonetic variability in stops and flaps in sponteaneous and careful speech. *The Journal of the Acoustical Society of America* 130(3), 1606–1617.
[42] Wrench, A. 2007. Advances in EPG palate design. *Advances in Speech-Language Pathology* 9, 3–12.
[43] Wrench, A. A., Gibbon, F. E., McNeill, A. M., Wood, S. E. 2002. An EPG therapy protocol for remediation and assessment of articulation disorders. In: Hansen, J. H. L., Pellom, B. (eds), *Proc. of the 7th International Conference on Spoken Language Processing*, Denver, CO, 965–968.
[44] Zampaulo, A. 2019. *Palatal Sound Change in the Romance Languages: Diachronic and Syncrhonic perspectives.* Oxford University Press.

OPTIMAL CONTROL THEORY OF SPEECH PRODUCTION USING PROBABILISTIC ARTICULATORY-ACOUSTIC MODELS

Benjamin Elie[1], Juraj Šimko[2], Alice Turk[1]
[1]Linguistics and English Language, School of PPLS, The University of Edinburgh, United Kingdom
[2]University of Helsinki
benjamin.elie@ed.ac.uk, juraj.simko@helsinki.fi, a.turk@ed.ac.uk

ABSTRACT

This paper introduces the concept of using probabilistic articulatory-acoustic models in optimization-based models of speech articulatory planning. In these models, speech articulatory movements are assumed to be performed such that they satisfy conflicting task requirements, such as least effort and intelligibility. Our model is used to account for intelligibility by computing the posterior probability of a vowel given a vector of formant values. These models can be trained on formant values and associated vowel labels extracted from available audio corpora. We present a minimal example using a model of the American-English vocalic system trained on formant values extracted from the TIMIT database. A preliminary experiment illustrates the interest of the approach by reproducing vowel centralization when least effort is weighted more highly than intelligibility.

Keywords: Articulatory planning; Speech production; Optimal Control Theory;

1. INTRODUCTION

Modeling speech production is a challenging task as several aspects must be considered. Speech communication varies depending on context, and involves precise coordination of several articulators in a short amount of time. One approach to understand these mechanisms is to predict and reproduce them via computational modeling of speech articulatory planning. This approach consists of predicting movements of speech articulators to be produced by a speaker given a string of words and a set of task requirements for the utterance (speech rate, prosodic structure...).

Different theoretical issues related to speech articulatory planning include speech variability [1], dynamic modeling of speech articulatory trajectories [2–4], the nature of coordination and speech goals [5–7], and speech timing patterns [8–10]. Optimal Control Theory [11, 12] is a promising approach for tackling these issues, as it allows models of articulatory planning to account for multiple factors that affect the speech signal. Optimal Control Theory approaches assume that human purposeful movements reflect the optimal balance of costs such as those of not meeting task requirements, and movement costs, such as effort. These task requirements and movement costs are modeled mathematically as a composite multi-objective cost function to minimize. For speech, this involves an efficient trade-off between production costs (e.g. effort) and the costs of not being understood (the task requirement of intelligibility). This has been proposed by Lindblom in his Hyper- and Hypoarticulation (H&H) theory of speech [1]. According to Lindblom, speech variation occurs because of different adjustments of the trade-off between maximal intelligibility, resulting in hyperarticulation, and minimal articulatory effort, resulting in hypoarticulation.

Following this approach, this paper focuses on the mathematical modeling of intelligibility in the objective function. Prior literature has proposed to model the intelligibility cost as the distance from targets [8, 13]. In Embodied Task Dynamics (ETD) [8], the intelligibility cost is a linear function of the distance from invariant canonical targets for vowels. Although this linear function allows a qualitative approximation of intelligibility (the closer to the target, the more intelligibility), it does not provide a realistic approximation of the mapping between intelligibility and distance from the target, as there is no reason why this should be a linear function. For stop consonants, the cost is a binary function as these sounds require well-defined, binary articulatory conditions (a stop is produced in the presence of an occlusion, and isn't if there isn't a full occlusion). This approach of classifying sounds in terms of binary "meeting the target" vs. "not meeting the target" has been extended to vowels in DIVA [14], although in DIVA targets are defined as regions in acoustic space, instead of as a single set of formants or formant pattern. DIVA allows the target region areas to be manipulated to account for particular contexts. DIVA's approach requires precise modeling of the target regions for various speech styles, which may be difficult as it requires a lot of data and learning steps to achieve. Additionally, the DIVA approach suggests that vowels inside the region boundaries are equally good, which is a strong assumption.

In order to overcome these issues, in this paper we propose a novel approach to speech targets and intelligibility based on probabilistic articulatory-acoustic models. The idea is to consider the intelligibility function as the probability of a target speech sound to be recognized as a function of an articulatory configuration. On the assumption that human perception of vowels is based on the statistical distribution of produced vowels in the acoustic space, our approach provides a more realistic approximation of intelligibility during speech communication in a principled way. In addition, our approach offers a more flexible way to

account for different languages and speech variation. The probabilistic model can be modified for specific languages as long as labeled corpora are available for each language. Speech variation can be modeled by varying the weights assigned to each component of the cost function (task requirements and movement costs). This paper presents an example of such a model used for American-English vowels, built using a Gaussian Mixture Model (GMM) trained on formant values extracted from an audio corpus [15, 16].

The structure of the paper outlines the main contributions of the paper. Section 2 introduces the minimal OCT-based model used in this paper to test our approach. Section 3 details an example of a probabilistic model built following our approach, including the presentation of the corpus used for the training data, and the characteristics of the model components. Section 4 presents a short preliminary experiment aiming at illustrating the interest of the approach. In this experiment, we evaluate the impact of the weight assigned to the least effort requirement on the position of the optimized vowel in the formant space.

2. THE OPTIMAL CONTROL MODEL APPLIED TO SPEECH PRODUCTION

This section details the articulatory model and the composite objective function to minimize.

2.1. Multi-task objective function

Although many OCT-based models include objective tasks other than intelligibility and effort, e.g., utterance brevity [8], we will use a simplified model that tests intelligibility only against articulatory effort, *i.e.*, our cost function contains only two elements, one for intelligibility, and one for articulatory effort. Consequently, the composite objective function used in this paper is as follows:

$$(1)\quad C(\mathbf{x}) = \alpha_E E(\mathbf{x}) + \alpha_I (1 - I(\mathbf{x})),$$

where $C(\mathbf{x})$, $E(\mathbf{x})$, and $I(\mathbf{x})$ are the cost function, the effort cost (to be minimized), and the intelligibility cost (to be maximized), respectively, which are function of the model parameter vector $\mathbf{x}$. The trade-off between effort and parsing requirements is adjusted via the weights α_E and α_I applied to the effort and the intelligibility costs.

2.2. Articulatory model

The vector $\mathbf{x}$ contains the parameters of an articulatory model that describe the position of the speech articulators and the geometry of the vocal tract at a given time instant. In theory, it should also contain dynamic information about the timecourse of these static parameters using a dynamic model [2, 3, 8, 17, 18]. We made the choice of using solely a static model (i.e. the Maeda model [19]) in order to focus on a single aspect of our model, namely the intelligibility cost at a given time instant. Note that this model is fully compatible with a dynamic model.

The Maeda model used in this paper [19] generates midsagittal shapes of the vocal tract using 7 independent articulatory parameters, as described in Figure 1. The articulatory parameters are the principal components that explain most of the observed variance in articulatory data. They are expressed in terms of standard deviations above or below the mean value, where the mean value (i.e. 0) corresponds to a neutral position. The vector $\mathbf{x}$ form Eq. 1 contains the values of the seven parameters, where each value is contained between -3 and +3.

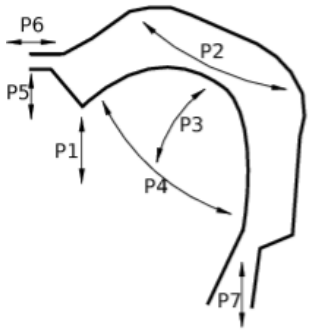


Figure 1: The Maeda articulatory model and its 7 components. P1 controls the jaw position, P2 and P3 control the position and the height of the tongue dorsum, respectively, P4 the position of the tongue tip, P5 and P6 the lip aperture and protrusion, respectively, and P7 the larynx height. Figure extracted from [20].

2.3. Articulatory effort

In this paper, we consider articulatory effort as a function of the distance between successive articulatory configurations in the articulatory space spanned by the parameters of the Maeda model, as suggested in [21]. We will consider only two successive configurations, where the starting configuration is the neutral position configuration $\mathbf{x}_0$, assumed to be the vocal tract configuration at rest. In the Maeda model, this corresponds to the null vector, $\mathbf{x}_0 = \mathbf{0}$. The articulatory effort is then the square of this Euclidean distance, which, in this case, is simply the square of the Euclidean norm of $\mathbf{x}$. Additionally, we normalize the effort costs such that it is between 0 and 1. The Maeda model contains 7 parameters, whose values are between -3 and +3. The maximal squared Euclidean norm of the vector $E_{\max}$ then equals to 63. Consequently,

$$(2)\quad E(\mathbf{x}) = \frac{1}{E_{\max}} ||\mathbf{x}||^2 = \frac{1}{63} ||\mathbf{x}||^2.$$

2.4. Intelligibility cost

Considering a vowel v and a vector of formants $\mathbf{f}$ produced by the articulatory vector $\mathbf{x}$, the intelligibility $I(\mathbf{x})$ is the posterior probability of v given $\mathbf{f}$, hence

$$(3)\quad I(\mathbf{x}) = P(v|\mathbf{f}).$$

Computing $\mathbf{f}$ from $\mathbf{x}$ is done using the Maeda model presented in Section 2.2. Estimating the probability is done using a Gaussian Mixture Model (GMM) presented in the next section.

3. THE PROBABILISTIC ACOUSTIC MODEL

The probabilistic acoustic model is used to estimate the likelihood of a target phone to be produced (or recognized by the listener) given an articulatory configuration. In this paper, we propose a corpus-based formant-to-probability GMM that returns monophthong vowel likelihood within a set of vowels given a formant pattern.

3.1. Corpus

We used the Vocal Tract Resonance (VTR) Corpus [15], which contains manually extracted formant trajectories of 538 utterances from the TIMIT database [16], uttered by 186 speakers of American English.

3.2. Preprocessing

The values of the 4 first formants at the mid-point of each of the 5526 analyzed monophthong vowels in the VTR corpus have been extracted, resulting in a 5526×4 matrix. We then merged some vowels to form one group. This includes the three following groups: /ə/-like vowels ax, axr, and ax-h, merging into a single ax class, /ɪ/-like vowels ix and ih merging into a single ih class, and /u/-like vowels uw and ux merging into a single ux class. Consequently, it results in 11 vowel classes.

Vocal Tract Length Normalization (VTLN) has been applied to formant values [22] to remove speakers' anatomic discrepancies. VTLN consists of estimating the length of each speaker's vocal tract from their formant patterns, and then multiplying the formant values of each speaker by a single factor such that they correspond to the formants of a virtual speaker having a reference vocal tract length L_{ref}. We chose L_{ref} so that it corresponds to the length of the vocal tract in the neutral configuration of the Maeda model, namely $L_{\text{ref}} = 16.27$cm.

We also generated synthetic formant vectors that lie outside the convex hull of the observed formants to simulate non-vocalic sounds. This is done to prevent the optimization process to produce unrealistic formant patterns. 96286 formant vectors have been generated this way, and classified in 57 different classes. We chose to use several components for non-vocalic sounds because the global distribution of non-vocalic sounds in the formant space cannot be fitted with a single Gaussian. We therefore chose to use 57 Gaussian distributions to model the overall distribution of non-vocalic sounds, as we found 57 to be a good trade-off between a good fit of the overall distribution and a reasonably small amount of components. The model then contains $57 + 11 = 68$ components. Finally, minority classes are randomly oversampled to create balanced data [23], yielding a total number of 239020 vectors.

3.3. Parameters of the American English FtP model

The 68-component GMM is fitted on the data using the iterative Expectation-Maximization (EM) algorithm. The model is initialized from prior information on vowel labels, namely each vowel observation is connected to its label. Figure 2 shows a projection of the individual probability function in the $F1-F2$ space for the 6 following vowels: aa (/ɑ/), ah (/ʌ/), ao (/ɔ/), eh (/ɛ/), iy (/i/), and ux (/u/). The centres of the vowel distributions correspond to typical positions for each vowel, as reported in the literature [24]. It also shows overlap between vowels, which is in line with previous observations [25, 26].

4. EXPERIMENTS

This section presents a preliminary experiment designed to illustrate the usefulness of our approach for Optimal Control Theory-based models. In this experiment, we optimize the production of several vowels for different ratios between the weights α_E and α_I assigned to the effort and intelligibility costs, respectively. The idea is to analyze the movement of the produced vowels inside the $F1-F2$ vowel space when this weight ratio varies. We hypothesize a centralization of the vowel space with an increase in the effort weight. That is, giving more penalty to articulatory effort should constrain the optimized articulatory vector to have a smaller norm, namely to be closer to the neutral position.

In order to conduct this experiment, we consider 4 vowels, namely aa, eh, iy, and ux (/ɑ/, /ɛ/, /i/, /u/, respectively). For each vowel, we run the optimization for different values of the effort weight α_E, and we keep $\alpha_I = 1$. For each optimization run, the initial solution is the neutral position $\mathbf{x}_0 = \mathbf{0}$.

Fig. 3 shows the position in the $F1-F2$ space of the produced vowels for different values of the effort weight. As expected, increasing the weight assigned to the effort cost results in vowel centralization: their position in formant space converges towards a central position, corresponding to the formants of the neutral configuration of the Maeda model. As a consequence, the volume of the vocalic space becomes smaller as the effort weight increases, as highlighted by the convex hull shown in Fig. 3. Eventually, when the weight assigned to the effort cost becomes too large, the returned solution is always the neutral vocal tract configuration, and the vowels are all at the same position in the vowel space.

Interestingly, the trajectory of vowels in the formant space when varying the effort weight can exhibit large gaps. This is clearly visible for aa in Fig. 3: The returned solution for aa slowly moves towards the center for $\alpha_E < 6$ and then quickly reaches the center as soon $\alpha_E > 6$. One possible explanation is that our model predicts a larger effort to produce canonical aa than other vowels. Indeed, $E(\mathbf{x}) \simeq 0.55$ for the "optimal" aa, while it is 0.4, 0.12, and 0,10 for eh, ux, iy, respectively. As a consequence, the threshold above which the effort cost weight is too large to allow movement is lower for aa than for other vowels.

These results illustrate the interest of our approach for OCT models and also highlight possible future directions for improving our model. Firstly, this paper introduces a purely static model that considers articulatory effort as a distance between the final articulatory configuration and the neutral configuration. In real speech, articulatory effort to produce a phoneme depends on the previous

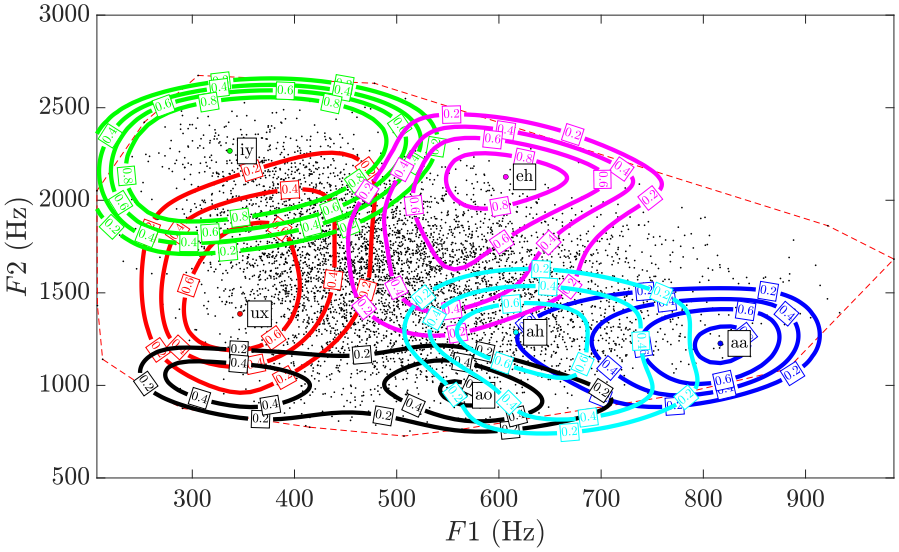


Figure 2: A projection of the individual probability functions for 6 vowels (aa, ah, ao, eh, iy, ux) in the $F1-F2$ space. For each vowel, F3 and F4 are taken as the vowel's mean value, and the probability $P(v|\mathbf{f})$ is computed for various values of $F1$ and $F2$.

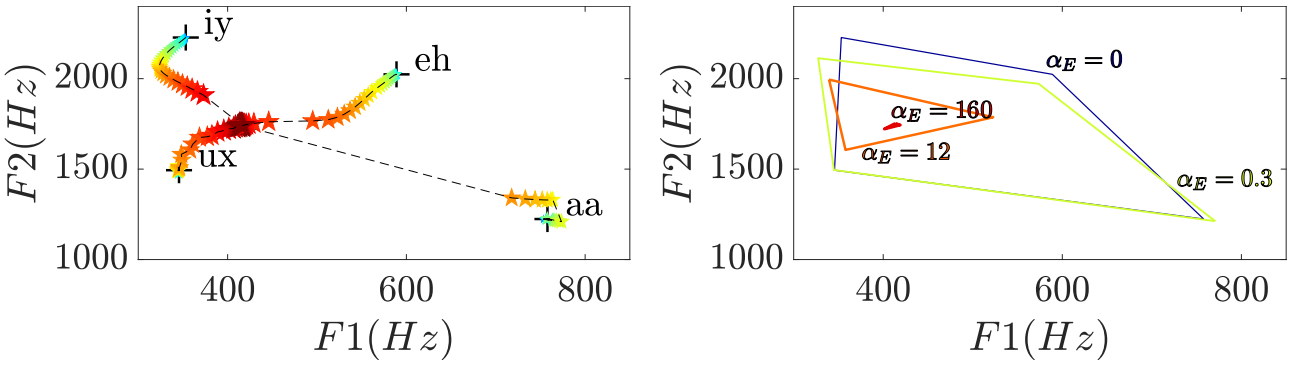


Figure 3: The left plot represents the position of optimized vowels aa, eh, iy, and ux (/a/, /ɛ/, /i/, /u/, respectively), in the $F1-F2$ space for different values of the effort weight. The cross marker '+' represents the position for $\alpha_E = 0$. The marker size and color differ according to the value of the effort weight. The larger the weight, the larger the marker. The right plot represents the corresponding convex hulls for some values of effort weight ($\alpha_E = 0$, $\alpha_E = 0.3$, $\alpha_E = 12$, and $\alpha_E = 160$).

phone, on the timecourse of articulatory trajectories as well as on the mass of the articulators. The next step in our approach is therefore to use a dynamic model to account for effort. Secondly, our model of articulatory effort considers the movement of each articulator to be equally effortful, which may not be true. Indeed, some articulators may require more effort to move than others (eg. jaw vs. tongue tip). In order to take this into account, another improvement would be to weight the static parameters with different coefficients. We believe that this approach might potentially mitigate the issues related to discrepancies among different vowels (including the "jump" for aa vowel described above).

5. CONCLUSION

This paper has presented a probabilistic model to be used in optimal control theory-based models of speech articulatory planning. It consists of computing the posterior probability of a vowel given the values of its first four formants. This approach results in a non-linear mapping between the distance of an articulatory configuration from a canonical target and the intelligibility of the produced target vowel.

This paper illustrates this concept by using this approach in a minimal static OCT model that considers only articulatory effort and the intelligibility as conflicting tasks. Our results show that this model reproduces the expected reduction and centralization of the vocalic space when least effort is required.

Another interest of this approach is also to be very flexible as models only require formant values and labeled vowels as training data. Thus, models that account for any language, dialect, or accented variation of any language can be easily computed, assuming an appropriate audio corpus is available.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the European Research Council (ERC) under the European Unionâs Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A. Turk; Grant agreement No. 101019847).

7. REFERENCES

[1] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.

[2] E. Saltzman, "Task dynamic coordination of the speech articulators: A preliminary model," *Experimental brain research series*, vol. 15, pp. 129–144, 1986.

[3] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[4] T. Sorensen and A. Gafos, "The gesture as an autonomous nonlinear dynamical system," *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[5] D. N. Lee, "Guiding movement by coupling taus," *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[6] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, 1986.

[7] C. P. Browman, L. Goldstein *et al.*, "Dynamics and articulatory phonology," *Mind as motion: Explorations in the dynamics of cognition*, vol. 175, p. 194, 1995.

[8] J. Simko and F. Cummins, "Embodied task dynamics," *Psychological review*, vol. 117, no. 4, p. 1229, 2010.

[9] S. Tilsen, "Selection and coordination: The articulatory basis for the emergence of phonological structure," *Journal of Phonetics*, vol. 55, pp. 53–77, 2016.

[10] A. Turk and S. Shattuck-Hufnagel, "Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production," *Frontiers in Psychology*, p. 2952, 2020.

[11] E. Todorov, "Optimal control theory," *Bayesian brain: probabilistic approaches to neural coding*, pp. 268–298, 2006.

[12] R. Shadmehr and J. W. Krakauer, "A computational neuroanatomy for motor control," *Experimental brain research*, vol. 185, no. 3, pp. 359–381, 2008.

[13] B. Parrell and A. C. Lammert, "Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives," *Frontiers in Psychology*, vol. 10, p. 2251, 2019.

[14] F. H. Guenther, "Neural control of speech movements," *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 209–239, 2003.

[15] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[16] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[17] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[18] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[19] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[20] A. Toutios and S. S. Narayanan, "Articulatory synthesis of french connected speech from ema data." in *INTERSPEECH*, 2013, pp. 2738–2742.

[21] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2144–2162, 2011.

[22] K. Johnson, "Vocal tract length normalization," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[24] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[25] A. Jongman, M. Fourakis, and J. A. Sereno, "The acoustic vowel space of modern greek and german," *Language and speech*, vol. 32, no. 3, pp. 221–248, 1989.

[26] M. C. Kelley and B. V. Tucker, "A comparison of four vowel overlap measures," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 137–145, 2020.

Being looked at in the face attenuates articulatory movements in both sighted and blind speakers

Haruka Saito[a], Paméla Trudeau-Fisette, Camille Vidou, Lucie Ménard[b]

Université du Québec à Montréal
saito.haruka@courrier.uqam.ca[a], menard.lucie@uqam.ca[b]

ABSTRACT

Many studies have shown that looking at the speaker's face improves speech perception, whereas what being looked at does to the speaker's speech production remains unclear. The current study asked both sighted and congenital blind speakers to produce vowels in face-to-face and audio-only conditions (both in a quiet and noisy background) while measuring the tongue and lip movements using electromagnetic articulography. We found that sighted speakers showed attenuated movements of the lips in the face-to-face conditions, while blind speakers showed attenuated movements of the tongue in the same conditions. These results suggest that (a) being looked at in the face may lead to attenuated, rather than enhanced, articulatory movements, and (b) this phenomenon may not only be associated with the speaker's attempt to adjust visual speech cues, but also with other communicative factors.

Keywords: Speech production, audio-visual, electromagnetic articulography, blind speakers

1. INTRODUCTION

Visual information is an integral part of speech communication. Prior literature has consistently found that looking at the interlocutor's face, along with their speech, improves auditory detection [1] and speech intelligibility in noise both for listeners with normal hearing [2] and with hearing impairment [3], as well as enhances learning of foreign sounds [4].

What remains unclear is to what extent speakers are actively aware and make use of visual speech cues to optimize intelligibility. Considering that auditory and visual cues are usually in sync—a clearly pronounced audio tends to be produced by clearly defined articulatory movements, and vice versa—it is possible that the benefits of visual cues are a fortunate by-product of speech production for listeners, but not an active strategy of speakers. A few recent studies, however, suggested that speakers may be capable of adapting their auditory and visual speech cues separately in a face-to-face condition by selectively enhancing visible articulatory movements: speakers showed increased lip movements when performing a face-to-face communicative task in noise, despite that their auditory productions in the same condition were somewhat attenuated (i.e., decreased intensity and/or F0) compared to an audio-only condition [5, 6].

To further investigate whether and how speakers adapt their speech output when they know that visual modality is available for their interlocutor, the current study asked both sighted and congenital blind speakers to produce sentences in both face-to-face and audio-only conditions. Because blind speakers have never been exposed to the visual domain of speech, any changes that they would make in a face-to-face situation are less likely associated with their attempt to adapt visual cues. By comparing them and sighted speakers, we aimed to investigate whether speakers' speech adaptations are more likely to be associated with visual adaptation strategies or with other factors specific to a face-to-face context. In this paper, we focus on articulatory data and will not present acoustic data, which is to be discussed elsewhere [7].

2. METHODS

2.1. Participants

Nine sighted (age M=33.2 years; Male 6, Female 4) and ten congenitally blind (age M=37.6; Male 6, Female 3) native speakers of Canadian French were recruited in Montreal, Canada. All speakers had normal hearing and reported no speech or language impairments. The blind speakers never had any visual perception of light or movement, classified as class 3, 4, or 5 in the International Disease Classification of the World Health Organization (WHO).

2.2. Task and procedures

Participants were asked to produce six Canadian French vowels /a, i, u, y, ɛ, e/ (note that only the first four were used in the current analysis) embedded in a carrier sentence, 'Monsieur /pVp/ est parti' (*Mister /pVp/ is gone)*, in four different conditions: audio-only, audio-only with noise, face-to-face, face-to-face with noise. In all conditions, participants were asked to produce the sentences so that an interlocuter (= experimenter) could understand them, with the only difference being that in the audio-only conditions, the interlocuter was standing behind the participant

(about one meter behind their left shoulder) and the participant was explicitly told that the person could not see their face; in the face-to-face conditions, their interlocuter was standing in front of them (about 1.5 meters away) and the participant was explicitly told that the person was looking at them. In addition, in the conditions with noise, the participant wore earbuds to listen to white noise (60 dB) and was informed that their perceiver was listening to the same noise. All participants produced 192 vowel samples in total (6 vowels x 8 repetitions x 4 conditions).

2.3. Data recordings

Articulatory data was obtained using Carstens AG501 (electromagnetic articulography: EMA) with a sampling rate of 250 Hz. Sensors were placed on upper lip (UL), lower lip (LL), lower incisors (JAW), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD), in addition to three reference sensors to track head movements (two mastoids and upper incisors). We also recorded each participant's bite plane using a solid plate with three sensors on it. Simultaneously, acoustic data was recorded via an Audio-Technica microphone (BP892) at 44100 Hz.

2.4. Data analysis

The EMA position data was transformed to account for head movements using Matlab scripts (courtesy of Mark Tiede, Haskins Laboratories) with reference to three reference sensors and the bite plane. The position data was then extracted at the middle point of each target vowel, determined based on landmarks the acoustic signal. To normalize speakers' physiological differences in size, all sensor positions were z-scored within speaker (across all six experimental sensors and across all three dimensions).

Statistical analysis was performed using the lme4 [8] and emmeans [9] packages in R [10].

2.4.1. Tongue between-vowel distances

To examine whether speakers used a wider, or narrower, 'vowel space' of the tongue as a function of face-to-face and/or noise conditions, between-vowel distances were calculated using the following procedure: first, for each participant, the 'center' of each of the three corner vowel categories (/a/, /i/, and /u/) were calculated by averaging all the samples for the vowel category for each of the three dimensions (x, y, z); the Euclidean distance in the 3D space between each sample of the corner vowels and the 'center' of the other two vowel categories were then calculated (ex., if the sample is /a/, the distance between the x, y, z coordinates of that particular sample of /a/ and the 'center' of the speaker's /i/ category was calculated, as well as between that particular /a/ and the 'center' of the speaker's /u/ category); the two distances were then averaged to have one distance for each sample. This way, we could evaluate how far apart each sample of the corner vowels was from the other two corner vowels, with longer distances representing a wider vowel space. This analysis was done for TT, TB and TD sensors separately, although this paper will only present the results for TD.

2.4.2. Lip aperture

Vertical lip aperture, the Euclidian distance in the 3D space between UL and LL, was calculated for the low vowel /a/, assuming that larger lip aperture represents more salient visual cues for the vowel. One sighted participant was excluded from this analysis due to a measurement failure of UL and LL.

2.4.3. Lip protrusion

Lip protrusion for the rounded vowels /u/ and /y/ was measured as the Euclidian distance in the 3D space between LL and JAW. One sighted participant was excluded from this analysis due to a measurement failure of LL.

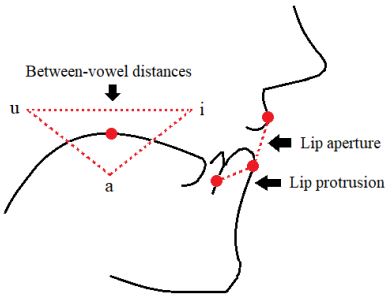


Figure 1: Illustration for EMA measurements

3. RESULTS

3.1. Tongue between-vowel distances

Figure 1 shows the average Euclidian distances between corner vowels /a, i, u/. A linear mixed effects model was built with Face (Audio or Face, sum-coded), Noise (Quiet or Noise, sum-coded), and Group (Sighted or Blind, sum-coded) and a three-way and all possible two-way interactions among the three as fixed effects and random intercepts by Participant and by Vowel. A significant main effect of Face (β=-.008, t=-3.23, p=.001) and an interaction between Face and Group (β=-.012, t=-4.71, p<.001) were found. Tukey's post-hoc test revealed that the blind group showed less tongue movements in the face-to-face condition compared to the audio-only condition

(p<.001 for the quiet, p=.003 for the noisy background). There was no such difference for the sighted group.

Another significant effect found was a main effect of Noise (β=-0.014, t=5.79, p<.001), which suggests that both groups used a larger vowel space when exposed to noise. No significant interactions between Noise and other fixed effects were found.

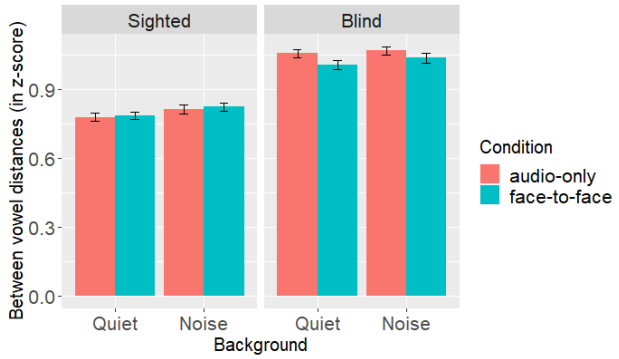


Figure 2: Tongue between-vowel distances of TD for the corner vowels /a, i, u/

3.2. Lip aperture

Figure 3 shows the average vertical lip aperture for /a/ for both groups. A linear mixed model was built using the same fixed and random effects as 3.1. The results show both a significant main effect of Face (β=-.013, t=-2.59, p=.012) and an interaction between Face and Group (β=-.023, t=4.18, p<.001). Tukey's post-hoc test revealed that the sighted group used smaller lip aperture when face-to-face compared to the audio-only condition in the quiet background (p<.001) but not in the noisy background (p=.44). There was no other significant difference with respect to the face-to-face vs. audio-only condition.

As for the effect of Noise, we found a significant main effect of Noise (β=.039, t=-7.22, p<.000) and no interaction with other fixed effects. Again, both groups tended to use larger lip aperture in a noisy background.

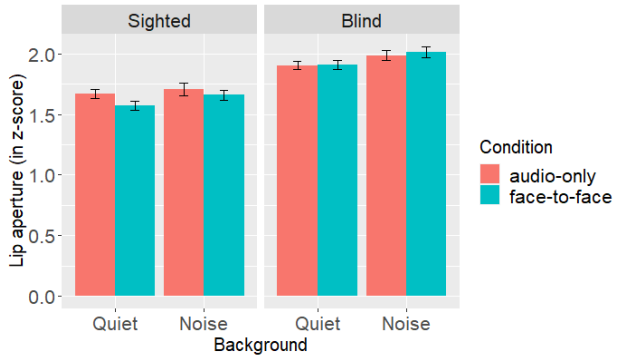


Figure 3: Lip aperture for the low vowel /a/

3.3. Lip protrusion

Figure 4 illustrates the average lip protrusion for /u/ and /y/ for both groups. A linear mixed model was built using the same fixed and random effects as 3.1 and 3.2. Again, we found both a significant main effect of Face (β=-.275, t=-8.19, p<.000) and an interaction between Face and Group (β=.254, t=7.57, p<.000). Tukey's post-hoc test found that, both in the quiet and noisy background, the sighted group showed less lip protrusion in the face-to-face condition compared to the audio-only condition (p<.001 for both the quiet and the noisy background), whereas the same difference was not found for the blind group.

As for Noise, we did not find either main effect of Noise or interaction between Noise and Group, which suggests that participants did not change the extent of lip protrusion as a function of noise.

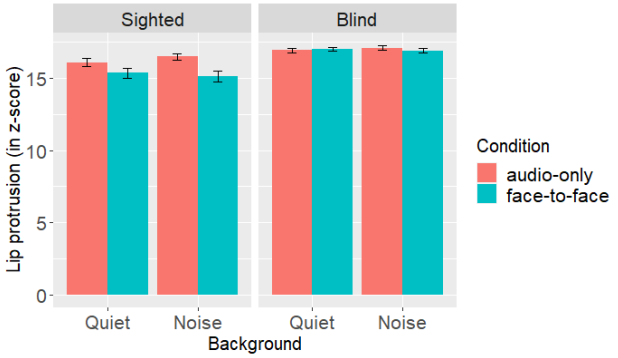


Figure 4: Lip protrusion for the rounded vowels /u, y/

4. DISCUSSION

The current study investigated how sighted and blind speakers may adapt their articulatory movements when they were aware that their interlocuter could utilize visual speech cues. We found that sighted speakers tended to use smaller, rather than larger, visual cues (lip aperture and protrusion) in face-to-face conditions. Blind speakers were also affected by face-to-face conditions: they showed less distinct vowel contrasts in their tongue movements.

Previous studies found that sighted speakers tended to enhance visible articulatory movements in face-to-face conditions and argued that (at least some) speakers could make active use of visual modality to enhance their speech intelligibility [5, 6]. Our findings were opposite: our sighted speakers attenuated the visual cues precisely when they knew their interlocuter could watch their face. Adding one more (i.e., visual) modality to speech communication means that listeners would have an additional source of information, which would in turn decrease the informational burden on one modality. Thus, it seems plausible if speakers make less effort in a face-to-face

situation than in an audio-only situation, since they are probably more confident that their speech would be understood by their listeners due to additional visual information.

In previous studies, speakers did show attenuated speech in the acoustic domain, but enhanced in the visual domain [5, 6] in a face-to-face situation. We speculate that this was because the task used in these studies, where participants must convey minimal word pairs to the interlocutor under the pressure of clarification in a noisy background, had a higher communicative load than the current study. These studies have shown that speakers *can* enhance visual speech cues under a certain condition specifically designed to make speakers rely more on visual modality. We presume that, without that kind of specific condition, simply being looked at in the face more likely attenuates, rather than enhances, speech production in the articulatory aspect as well.

Furthermore, the finding that blind speakers, despite their complete inexperience in the visual domain of speech, also attenuated their speech production in face-to-face conditions led us to consider *why* speakers may make less effort when being looked at during speech production. In the above discussion, we speculated that it was because sighted speakers were aware that additional visual modality would help them understood by their interlocuter, but this does not entirely apply to blind speakers. For this group of speakers, a face-to-face situation should be less defined by the presence of visual modality, but more by other communicative factors. For example, knowing that they have the interlocuter's gaze signals that they most likely have the person's full attention, which would decrease the need for enhanced speech. Another factor in a face-to-face situation might be the need to compose facial expressions (and blind individuals are known to share some universal, albeit different in some contexts, facial expressions as sighted individuals [11]), which would constrain certain articulatory movements as speakers need to use part of muscles to control facial features. These factors are not directly associated with the speaker's attempt to adjust visual speech cues, and if any of these might be part of the motivation for blind speakers to adapt their speech production, why are they not for sighted speakers?

In the current study, sighted and blind speakers differed in terms of which articulator they adapted: sighted speakers decreased visible (lip) movements while blind speakers less visible (tongue) movements. One therefore can argue that the adaptation strategy used by the two groups of speakers were fundamentally different, the former attempting to control visible speech cues and the latter motivated by other communicative factors. However, it has been reported that sighted and blind speakers use different articulators to produce similar acoustic results: the sighted use more lip contrasts and the blind use more tongue contrasts when they produce contrastive or clear speech [12, 13]. Thus, we speculate that the two groups in the current study may have aimed for similar acoustic results (i.e., attenuated speech) due to similar reasons—be it due to having the interlocutor's attention, composing facial expressions, or other factors specific to the face-to-face condition—and even sighted speakers may not have actively attempted to adapt visual speech cues in the face-to-face conditions.

Finally, although the above discussion describes that speakers "attenuated" articulatory movements compared to the baseline of the audio-only condition, one can argue that it was the other way round: speakers may have made an extra effort in audio-only situations considering, in reality, speech production should occur face-to-face more often. We chose to use the audio-only condition as a reference because it is arguably the most common way to collect speech production data in this field of study (i.e., experimenters either do not look at the speaker or do not report where their gaze was during a recording session). This leads to the question of whether non face-to-face conditions should be the standard for data collection, and this study suggested that the awareness of gaze does affect speech production.

5. CONCLUSION

Visual speech cues are a significant source of information for listeners. This, however, does not necessarily mean that speakers enhance their speech cues whenever visual modality is available. The current study found that sighted speakers tended to attenuate, rather than enhance, visible articulatory movements in face-to-face situations. Considering that we observed attenuated articulatory movements in congenital blind speakers as well (albeit it was in an invisible articulator), we assume that this attenuating effect may be more related to general factors in face-to-face situations, rather than specifically to the attempt by sighted speakers to actively control visual cues.

6. REFERENCES

[1] Grant, KW., Seitz, PF, 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108(3), 1197–1208.

[2] Schwartz, JL., Berthommier, F., Savariaux, C. 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93(2), B69–B78.

[3] Bergeson, TR., Pisoni, DB., Davis, RAO. 2003. A Longitudinal Study of Audiovisual Speech Perception by Children with Hearing Loss Who have Cochlear Implants. Volta Rev 103(4). 347–370.
[4] Hazan, V., Sennema, A., Iba, M., Faulkner, A. 2005. Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. Speech Communication 47(3), 360–378.
[5] Fitzpatrick, M., Kim, J. Davis, C. 2015. The effect of seeing the interlocutor on auditory and visual speech production in noise. Speech Communication 74, 37–51.
[6] Garnier, M., Ménard, L., Alexandre, B. 2018. Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? J. Acoust. Soc. Am. 144(2), 1059–1074.
[7] Trudeau-Fisette, P., Vidou, C., Uribe, C., Ménard, L., in Preparation. Acoustic impacts of congenital blindness on speech production strategies linked to Lombard speech and audiovisual interaction.
[8] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-23. Retrieved from http://CRAN.R-project.org/package=lme4
[9] Length, RV. 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.2. Retrieved from https://CRAN.R-project.org/package=emmeans
[10] R Core Team. 2015. R: A Language and Environment for Statistical Computing (Version 4.0.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/
[11] Valente, D., Theurel, A., Gentaz, E. 2018. The role of visual experience in the production of emotional facial expressions by blind people: a review. Psychonomic Bulletin & Review 25, 483–497.
[12] Ménard, L., Leclerc, A., Tiede, M. 2014. “Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults,” J. Speech Lang. Hear. Res. 57, 793–804.
[13] Ménard, L., Trudeau-Fisette, P., Côté, D., Turgeon, C. 2016. Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers,” PloS One 11, e0160088.

EJECTIVES IN GEORGIAN. A REAL-TIME MRI ANALYSIS OF VERTICAL LARYNX MOVEMENT

Nato Sulaberidze [1], Erika Brandt [2], Phil Hoole [3], Martin Krämer [4], Jürgen R. Reichenbach [4], Adrian P. Simpson [1]

[1] Friedrich Schiller University Jena, [2] State Criminal Police Office Berlin, [3] Ludwig Maximilian University Munich, [4] Jena University Hospital

nato.sulaberidze@uni-jena.de, erika.brandt@polizei.berlin.de, hoole@phonetik.uni-muenchen.de, martin.kraemer@med.uni-jena.de, juergen.reichenbach@med.uni-jena.de

ABSTRACT

Vertical larynx movement is often claimed to be responsible for the intraoral pressure increase in ejectives. However, the epiphenomenal production of these sounds in German shows no motivation for such laryngeal involvement and appears to be driven by pulmonic airflow. It is plausible that in a language with phonological ejectives, these stops could also be produced, at least in part, by pulmonic airflow and no larynx involvement. The present analysis aims to investigate this by examining vertical larynx movement in the production of ejectives in Georgian. As part of a larger study, real-time MRI recordings of Georgian speakers are examined. Significant larynx height and movement shape differences between ejectives and their non-ejective congeners (voiceless aspirated and voiced plosives) are observed. Ejectives showed a greater average larynx height than control plosives. Large laryngeal displacement was observed only in the sentence-initial stops, regardless of the stop category.

Keywords: Ejectives, Georgian, vertical larynx movement, larynx height, real-time MRI

1. INTRODUCTION

The textbook description of an ejective is an obstruent (plosive, fricative) produced with an airstream mechanism involving a closed glottis and an upward movement of the larynx that reduces the volume of the supraglottal cavity. If the velum is raised and there is a complete closure or a stricture of close approximation somewhere in the oral cavity, intraoral pressure increases fuelling the plosive release or the turbulent airflow for a fricative [1]. In German, the release of word-final plosives preceding glottalised vowel-initial syllables typically exhibit the auditory and acoustic characteristics of ejective stops [2]. However, in these epiphenomenal ejectives, arising from the temporal overlap of oral and glottal closure in this context, there is no motivation for any vertical larynx movement [3]. Instead, there is evidence that the air pressure fuelling the release is created by a pulmonic airstream flowing into the closed oral cavity *before* the glottis is closed. With the glottis closed or configured for a creaky voice, the release acquires the auditory and acoustic characteristics of an ejective [2, 4]. Accounting for the production mechanisms behind epiphenomenal ejectives in German has two interesting consequences. First, it provides a plausible path for the emergence of ejectives in a language in which they are not (yet) part of the regular phonological system [5]. Second, although we have just described the production of pulmonically fuelled ejective-sounding plosives in German, there is no reason why languages with ejectives as part of their regular phonological inventory should not also exploit a similar method of production, i.e. use a pulmonic airflow during the stop closure for intraoral pressure build-up before closing the glottis prior to stop release. Indeed, this account has been entertained in an earlier study modelling airflow in ejectives [6].

The present study explores the second of these possibilities by using real-time magnetic resonance imaging (rtMRI) to examine whether there is evidence for vertical larynx movement during the production of ejectives in Georgian, a language in which they make up part of the regular phonological inventory [7].

2. METHOD

As part of a larger study comparing the production of ejectives in Georgian, English, and German using dual-channel electrolaryngography and intraoral pressure measurement [3], a subset of 5 Georgian and 7 German speakers were recorded producing

target stops in a range of contexts relevant to each language using rtMRI. In the current study, we focus on the analysis of the Georgian data.

2.1. Speech material

In Georgian, ejectives stand in regular phonological contrast to voiced and voiceless aspirated stops [7]. To allow for comparison between ejectives and their pulmonic congeners, Georgian speech material comprised words (near-minimal pairs) embedded in a sentential context containing either ejectives or their voiceless aspirated and voiced congeners (controls) at bilabial /p' p b/, alveolar /t' t d/ and velar /k' k g/ places of articulation. During the recordings, speech material was projected onto a screen in the MRI device. Subjects first read a short text and then the short sentences, each presented individually and in randomised order. For the present analysis, only the sentence material was analysed. In the elicited material, vocalic context and the position of the obstruents in the sentence and word were controlled. Target and control stops occurred at three different places: (a) utterance-/sentence-initial (e.g. პალატა /**p'**ala'ta/ 'hospital room') and in the second word of the sentence, either in (b) word-initial (სუფთა პანელი /supta **p'**aneli/ 'clean panel') or (c) word-medial position, at the beginning of the second syllable (e.g. წვერის საპარსი /ʦ'veris sa**p'**arsi/ 'beard razor'). Stops occurred in the vocalic context with open and close vowels depending on the sentence context, as follows: Ca & Ci (sentence-initial), aCa & iCi (intervocalic). In total, the sentence material contained 162 target words (3 places of articulation $\times$ 3 manners of articulation $\times$ 3 sentence contexts $\times$ 2 vowel contexts $\times$ 3 elicitations).

2.2. Subjects

Five female native speakers of Georgian ($n = 5$, aged 25–30 years, $M = 27(2)$) were recorded. They were students at the University of Jena. One of the subjects had to be excluded because images were found to be off the midsagittal plane obscuring the tracheal airway needed for visually establishing the lower edge of the larynx. The remaining four subjects gave Tbilisi as the place where they had lived the longest.

2.3. Experiment Setup

Subjects were recorded at the Jena University Hospital using a clinical 3T whole-body MRI scanner (Magnetom PRISMA fit, Siemens Healthineers, Erlangen, Germany) with the 64-channel head coil supplied by the manufacturer. For sound recording, a Sennheiser MO2000 optical microphone was attached to the head coil at 3 to 4 cm from the subject's mouth. For data acquisition, a single slice 2D radial Fast Low Angle-Shot measurement sequence [8] was used using a golden angle temporal ordering [9] and a radial undersampling factor of 7. Parameters: Echo time: 1.2 ms, fat saturation: off, flip angle: 8°, matrix size: 96 x 96, slice thickness: 6 mm, readouts per fully sampled k-space: 150, readouts per undersampled k-space: 21, repetition time: 2.7 ms, spatial resolution: 2.7 x 2.7 mm^2. The slice was aligned in a midsagittal orientation with a field of view size covering the entire head and parts of the neck. Image reconstruction was performed offline using conjugate gradient Sensitivity Encoding (CS-SENSE) [10] and sliding-window reconstruction [11] using a frame-to-frame temporal shift of 19 ms that resulted in a reconstructed frame rate of 53 frames per second.

2.4. Data processing

The audio data has been manually synchronised with the MRI images with the Computer Vision Toolbox in Matlab [12]. To remove MRI noise in the recorded signals, the MRI Speech Denoising Toolbox was used [13]. The audio data were automatically segmented and labelled using WebMaus tools [14] and manually adjusted in *Praat* [15]. The acoustically visible release point of the plosives served as the reference time point for the following analysis. The MRI image processing was done with Matlab [12] to extract the vertical position of the lower edge of the larynx over time. Image registration to a reference image was performed to control for possible movements of the speaker's head during the recording session and to obtain a vertical orientation of the trachea immediately below the larynx. Images with neutral laryngeal and head positions during the production of the word-initial [i] sound in the utterance "Sie fahren zur IAA nach Frankfurt", spoken by all subjects at the beginning of the recordings, were chosen as reference. The analysis was based on the lower edge of the laryngeal tissue to avoid the impact of the epiglottis in the signal. A simple edge-detection procedure extracted the vertical position at the transition from air in the trachea (lower intensity) to laryngeal tissue (higher intensity).

A contour of 15 temporally equidistant points representing the position of the lower edge of the larynx was extracted for each of the target

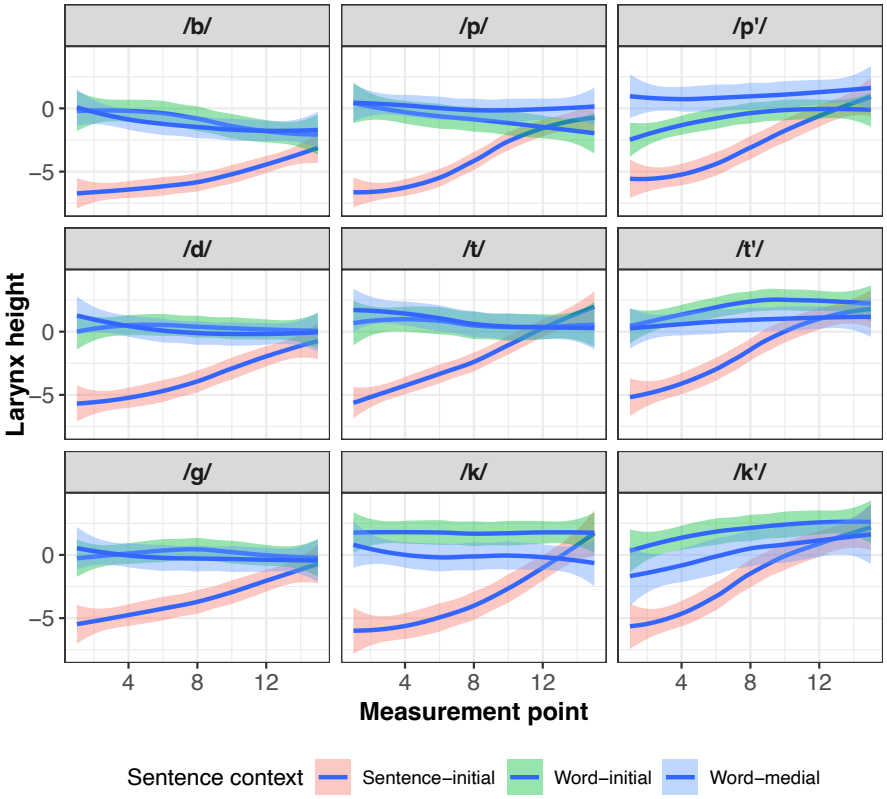


Figure 1: Smoothed larynx lower edge contours of target ejective (right), voiced (left) and voiceless aspirated stops (centre) in the three utterance contexts.

and control plosives for a fixed time range of 300 milliseconds centred around the stop release established in the acoustic record. Larynx contours containing outliers related to non-speech movements and measurement errors (NaN values) caused by the larynx not being visible in the recording were removed from the data set.

2.5. Data analysis

We used generalized additive mixed-models (GAMMs) for statistical analysis based on the mgcv [16] and itsadug [17] R packages [18]. GAMMs allow statistical analysis of nonlinear regressions. Therefore, they are specifically interesting for phonetic data which is collected in a time series and changes dynamically. We followed the modelling procedure given in [19].

3. RESULTS

Figure 1 displays the smoothed larynx lower edge contours of target ejective (right), voiced (left) and voiceless aspirated stops (centre) in the three utterance contexts. Immediately apparent from the plots is the substantial influence of utterance context for all the plosives, with the utterance-initial plosives exhibiting the largest larynx displacement regardless of the airstream mechanism they are being driven by. Although it is difficult to interpret the exact differences in laryngeal height between the ejective, voiceless and voiced plosives from the figure shown, statistical analysis with GAMMs provides more precise information on these differences.

Listing 1: Structure of GAMM for larynx height

```
#main effects on larynx height
m = gam(LH ~ Plosive + Context +
#smooth term plosive ordered
      s(MP, by=PlosiveO) +
#smooth term context ordered
      s(MP, by=ContextO) +
#Non-linear random effect
      s(MP, Speaker, bs="fs",m=1),
#Restricted maximum likelihood
      method='REML', rho=rhoval
#Control for auto-correlation
      AR.start=geo$start.event,
#Data specification
      data=geo)
```

The structure of the GAMM for larynx height (LH) is given in listing 1 with the fixed effects plosive category (levels: ejective, voiceless, voiced) and sentence context (levels: sentence-initial, word-initial, word-medial) and additional smooth effects for measurement point (MP) by ordered plosive and ordered context as well as a non-linear random effect which accounts for non-linear differences over measurement point with respect to each speaker. Ordered by-term factors are used to investigate the difference between the individual levels in trajectory shape. Following [19] we control for auto-correlation (see listing 1). Its value is set to a score that is present at a lag of 1.

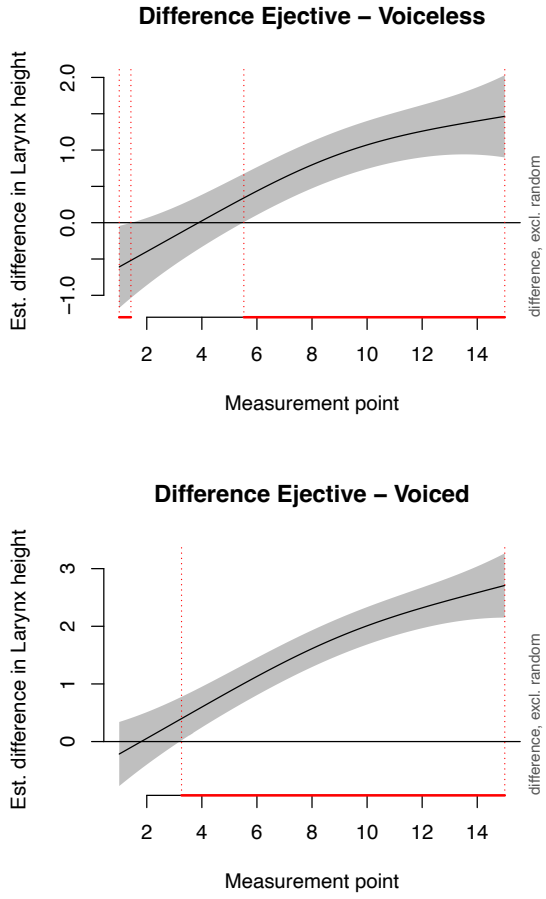


Figure 2: Difference smooth in Larynx height for ejectives and voicelsess (top) and ejectives and voiced stops (bottom) with 95% confidence intervals (shaded band). Significant differences are marked in red.

For model comparisons we use Akaike Information Criterion [20]. The nonlinear patterns are modelled with the default smooth type 'thin plate regression spline' [21]. The basis dimension (k) was set to 10 by default for this model. The final model explains 75.4% of data variability for larynx height in Georgian. On average, larynx height for voiced and voiceless Georgian plosives is significantly lower than for Georgian ejectives. Word-initial and word-medial plosives have significantly higher average larynx height than plosives in the sentence-initial position. With regard to the smooth terms, we find significant differences in the shape of the larynx lower edge contours between ejectives and both voiced and voiceless plosives. The trajectory shape for larynx height in sentence-initial plosives is also significantly different from word-medial and word-initial plosives. The random effect for differences over measurement point for each speaker also adds significantly to model performance. Figure 2 shows the estimated smooth differences in the shapes of the vertical movement of the larynx lower edge in ejectives compared to voiceless (top) and voiced stops (bottom). Red marks in the figure indicate the phases where the differences in the shape of the laryngeal trajectory are significant. The figure also distinguishes between the specific differences in larynx height values in the target and control stops, as follows: If the estimated difference values are above zero, then the laryngeal height values in the target stops (ejectives) are higher than in the conditions such as voiced and voiceless aspirated and vice versa, values below zero indicate lower larynx height in target stops compared to voiced or voiceless aspirated plosives respectively.

4. DISCUSSION

We have been able to register upward vertical larynx movement in the present study. However, the most remarkable factor contributing to differences in larynx movement is utterance context with the largest upward displacement in sentence-initial position, regardless of the plosive type. In a comparison of the target ejectives with voiceless aspirated and voiced plosives more nuanced differences did arise. In our data, Georgian ejectives show on average significantly higher larynx position than the other stop categories. However, based on the trajectory observations of laryngeal movements during the production of ejective stops compared to the other plosive categories, the suggestion that laryngeal movements in ejectives are pronounced enough to be associated with the increase in intraoral pressure, remains questionable. This can be considered as an indication, that ejective stop releases in many word- and sentence-internal contexts are being driven by pressure built up by a pulmonic rather than a glottalic airstream.

5. ACKNOWLEDGEMENTS

This study is funded by the German Research Council (grant: SI743/8-2). We would like to thank the Georgian subjects for their participation.

6. REFERENCES

[1] J. Catford, *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press, 1977.

[2] A. P. Simpson, "Ejectives in English and German – linguistic, sociophonetic, interactional, epiphenomenal?" in *Advances in sociophonetics*, C. Celata and S. Calamai, Eds. Amsterdam: Benjamins, 2014, pp. 187–202.

[3] E. Brandt and A. Simpson, "The production of ejectives in German and Georgian," *Journal of Phonetics*, vol. 89, p. 101111, 11 2021.

[4] A. P. Simpson, "Acoustic and auditory correlates of non-pulmonic sound production in German," *Journal of the International Phonetic Association*, vol. 37, no. 2, pp. 173–182, 2007.

[5] J. J. Ohala, "Emergent stops," in *Proc. 4th Seoul International Conf. on Linguistics [SICOL]*. Seoul: Linguistic Society of Korea, 1997, pp. 84–91.

[6] J. C. Kingston, "The phonetics and phonology of the timing of oral and glottal events," Ph.D. dissertation, University of California, Berkeley, 1984.

[7] R. Shosted and V. Chikovani, "Standard Georgian," *Journal of the International Phonetic Association*, vol. 36, pp. 255–264, 2006.

[8] M. Krämer, M. Weirich, K.-H. Herrmann, A. P. Simpson, and J. R. Reichenbach, "Assessing intrinsic velum height in vowels using time-resolved MRI," in *Proceedings 25. Annual Meeting International Society for Magnetic Resonance in Medicine*, vol. 25, Honolulu, Hawaii, 2017, p. 1511. [Online]. Available: http://archive.ismrm.org/2017/1511.html

[9] S. Winkelmann, T. Schaeffter, T. Koehler, H. Eggers, and O. Doessel, "An optimal radial profile order based on the golden ratio for time-resolved MRI," *IEEE Transactions on Medical Imaging*, vol. 26, no. 1, pp. 68–76, 2007.

[10] K. P. Pruessmann, M. Weiger, P. Börnert, and P. Boesiger, "Advances in sensitivity encoding with arbitrary k-space trajectories," *Magnetic Resonance in Medicine*, vol. 46, no. 4, pp. 638–651, 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1241

[11] S. J. Riederer, T. Tasciyan, F. Farzaneh, J. N. Lee, R. C. Wright, and R. J. Herfkens, "Mr fluoroscopy: Technical feasibility," *Magnetic Resonance in Medicine*, vol. 8, no. 1, pp. 1–15, 1988. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910080102

[12] MATLAB, *version 9.7.0.1319299 (R2019b)*. Natick, Massachusetts: The MathWorks Inc., 2019.

[13] C. Vaz, "MRI speech denoising toolbox," https://github.com/colinvaz/mri-speech-denoising.

[14] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326-347, 2017.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2022. [Online]. Available: https://www.praat.org

[16] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[17] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using GAMMs," 2022, r package version 2.4.1.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[19] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447017301377

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[21] S. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society:Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003.

THE RETROFLEX LATERAL [ɭ] AS AN ALLOPHONE FOR /l/ IN STANDARD AUSTRIAN GERMAN

Jan Luttenberger[1], Eva Reinisch[1]

[1]Acoustics Research Institute, Austrian Academy of Sciences
Jan.Luttenberger@oeaw.ac.at, Eva.Reinisch@oeaw.ac.at

ABSTRACT

Standard Austrian German is assumed to have a single lateral phoneme /l/ realized as alveolar lateral [l]. However, a few studies report a retroflex articulation after back vowels in speakers from the city of Vienna. Since those reports are almost exclusively based on auditory assessment, this pilot study sets out to investigate the presence of retroflex laterals via formant analysis and Ultrasound Tongue Imaging (UTI). Six speakers from the cities of Vienna, Graz and Salzburg were recorded reading sentences containing target words with /l/ in word-final position in 9 different vowel contexts. Acoustic measures showed lowered F2 and F3 in the lateral after back vowels for all speakers. However, UTI results showed that only 3 speakers realize /l/ with a tongue-tip gesture typical for retroflex articulation after back vowels and /ɑː/. The remaining speakers realize alveolar [l] in all utterances. Consequences regarding the relation between acoustics and articulation are discussed.

Keywords: speech production, laterals, Austrian German, ultrasound tongue imaging

1. INTRODUCTION

Laterals are known as a diverse sound class in terms of acoustics and articulation [1]. Even for languages with a single lateral phoneme, often a variety of realizations are reported. One such case is German, including Austrian German, for which a single alveolar lateral /l/ is assumed [2, 3]. Based on the existing literature, /l/ after front vowels is realized as the "default" pronunciation [l], but after back vowels and /ɑ/ it has been suggested that it can be articulated as a retroflex [ɭ]. The present paper reports on a pilot study based on acoustic formant measurements and ultrasound tongue imaging (UTI) investigating the realization of postvocalic /l/ in Standard Austrian German (SAG).

A first report of a retroflex manner of articulation for /l/ after back vowels and labial sounds in Austria dates back to the beginning of the 20th century [4] for speakers in Vienna. A similar description is given in [5] for the speech of homeless people in Vienna recorded in the 1970s. Yet, these reports are solely based on auditory judgements. The question hence arises whether realizations of /l/ in certain vowel contexts can really be counted as retroflex, and specifically, whether such realizations are still found in present-day SAG as spoken by young and middle-aged speakers (<40 years of age). It also remains to be shown whether retroflex articulation is exclusive for SAG speakers in Vienna.

With regard to the acoustics and articulation of retroflex laterals, descriptions are mostly given for non-European languages in which a phonemic opposition between different laterals is found. As for their acoustic properties, alveolar laterals without any secondary articulation such as velarization or palatalization typically show an F2 around 1500-2000 Hz and F3 around 2500 Hz [6]. In contrast, a retroflex articulation tends to lower F3 such that it is close to F2 [7]. Similar values were reported in an acoustic and articulatory study on Tamil, a language with a phonemic contrast between dental /l̪/ and retroflex /ɭ/ [8]: Tamil /l̪/ is reported to have an F2 of around 1200 Hz and F3 around 2400 Hz; F2 for /ɭ/ is around 1460 Hz and F3 around 1800 Hz.

Articulatorily, laterals feature a lingual closure whilst leaving a lateral opening so air can still escape. The gesture for [l] is similar to alveolar plosives; retroflex [ɭ] is described to exhibit a rolling gesture along the alveolar ridge from back to front [8].

Next to those descriptions of the acoustics and articulation of laterals in general and retroflex manner of articulation in particular, a number of studies showed that the exact realization of laterals often depends on phonological context (e.g., [9]) and positional factors [10]; also, social conditioning has been reported [11]. However, studies on retroflex lateral allophony in European languages are rare.

The present study therefore sets out to describe the acoustics and articulation of the lateral phoneme /l/, and specifically the occurrence of allophonic retroflex articulation [ɭ] in young speakers from three cities in Austria (i.e., Vienna, Graz, Salzburg). As mentioned above, previous studies based on auditory descriptions of SAG suggest retroflex allophones in some phonological contexts in speakers of Vienna. The addition of speakers from two other regional centers allows us to test whether retroflexion is found in different local varieties of SAG or whether it is limited to speakers from Vienna. Overall, most

Austrians are assumed to be competent in a form of intended (regional) standard as well as a local dialect. Switching and shifting between varieties according to social and pragmatic circumstances is common [12], [13]. Since the present study set out to focus on SAG, the potential for varietal switching was minimized by using a sentence reading task [14].

2. METHODS

2.1. Participants

Six speakers were recorded acoustically and articulatorily using UTI. They were native speakers of Austrian German, between 26 and 38 years of age, one female and one male speaker from each of the cities of Vienna, Graz and Salzburg. They all grew up and had received a university degree in their city of residence. Unfortunately, the UTI recordings of the female speaker from Salzburg turned out to be unusable as the visual image of alveolar region appeared to be occluded by the lower jaw. Therefore, data from only 5 speakers will be reported. Recordings of additional speakers are planned.

2.2. Materials and Procedure

Nine German words were selected in which the lateral phoneme occurred word-finally in different vowel contexts (see Table 1). Preceding context covered the front vowels /iː/ and /ɛ/, the back vowels /uː/ and /oː/, the open vowel /ɑː/, the front rounded vowels /yː/ and /øː/ and the diphthongs /aɛ̯/ and /ɑɔ̯/. A following context was avoided by embedding the targets in carrier sentences in sentence-final position. For each target a semantically neutral and a semantically meaningful context sentence was constructed. The target words analyzed in this study were part of a larger set of 36 words serving different purposes.

Target	Translation	IPA
viel	*much*	/fiːl/
hell	*bright*	/hɛl/
steil	*steep*	/ʃtaɛ̯l/
fahl	*pallid*	/fɑːl/
Stuhl	*chair*	/ʃtuːl/
hohl	*hollow*	/hoːl/
Maul	*mouth (animals)*	/maɔ̯l/
Öl	*oil*	/øːl/
kühl	*cool*	/kyːl/

Table 1: List of items. *Target* gives the standard orthography; *Translation* the English equivalent; *IPA* the phonological transcription for SAG.

Recordings were made in soundproof booths at local research institutions. AAA software [15] was used for stimulus presentation and UTI recordings. A Telemed MC4-2R20S-3 ultrasound probe was fixed with an Ultrafit headset underneath the chin of the speaker to ensure the stability of the probe during the recording [16]. For the audio recordings a Røde M2 condenser microphone was placed in front of the participants and connected via a Scarlett Audio Interface to a laptop. Participants were asked to read the sentences off the laptop screen: first all targets in the neutral, then the semantically meaningful sentences. Both sets of sentences had to be read twice, resulting in four recordings per target. The recordings were followed by an interview to collect additional information about language use and the social background of the speakers, which is not reported here.

2.3. Data Preparation and Analyses

The phonetic realization of /l/ was annotated in AAA software [15] for the UTI data and in STx [17] for the acoustic analyses. Segment boundaries were marked by auditory judgment and visual inspection of the spectrogram. For the UTI recordings, splines marking the tongue contour were created using the batch processing function of AAA at every available frame of the sentences. For each item, after rectifying erroneous splines, a mean spline at the midpoint of the lateral over all 4 repetitions of the words was computed using the AAA workspace, exported as cartesian coordinates and visualized in R [18] [19].

For the acoustic data, the first three formants of the laterals were tracked automatically using the formant tracker of STx and manually corrected if necessary. Formant measurements in Hz were taken at 20 points evenly spread out across the duration of the /l/ and exported into a data table. The 5 middle points were used for analyses. Since F2, F3 and their relative position have been used most commonly to describe retroflexion, F2, F3 and the difference F3-F2 will be used for acoustic analyses.

3. RESULTS

3.1. Articulatory Data

The mean tongue splines of the UTI data were visualized to compare the position of the front part of the tongue. Retroflex tongue gestures should be discernible from alveolar ones by a higher and more posterior position of the tongue tip [8]. Fig. 1 shows the tongue contours for the 9 items averaged over the 4 repetitions for each speaker. Note that each item refers to a unique vowel context.

The male speakers from Vienna and Salzburg show very similar tongue contours for all 9 items. The tongue tip extends forward and is relatively low compared to the back of the tongue, suggesting a

contact with the tongue blade in the alveolar region. The female speaker from Vienna and the two speakers from Graz show a higher and more posterior tongue tip position after /uː/, /oː/, /ɑ̯ɔ̯/ and /ɑː/ compared to the remaining contexts. This could suggest contact of the tongue tip in the postalveolar to palatal region. Critically, this equals the description of retroflex [ɭ] in Tamil [8].

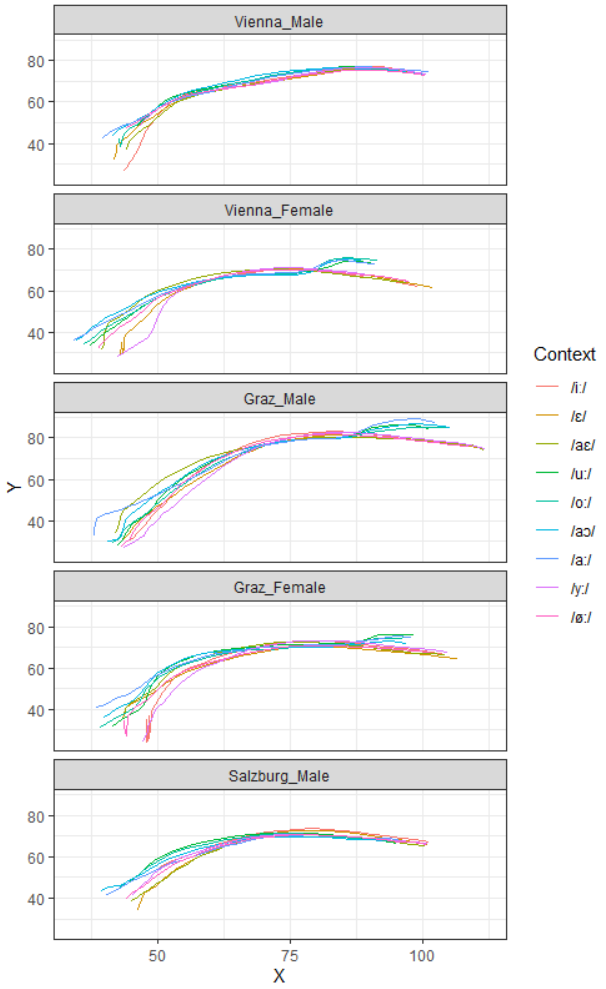


Figure 1: Mean tongue splines per vowel context for each speaker. X and Y refer to the cartesian coordinates as extracted from AAA. The tongue tip is on the right.

3.2. Acoustic Data

For a first assessment of the effect of phonological context on the laterals' formant values, items were grouped into 4 context conditions according to the frontness and rounding of the preceding vowels. /iː/, /ɛ/ and /aɛ̯/ were classified as front unrounded vowels, /yː/ and /øː/ as front rounded, /uː/, /oː/ and /ɑ̯ɔ̯/ as back and by default rounded. /ɑː/ was taken as back vowel without lip rounding. Analyses were conducted by first fitting a linear mixed-effects model with the difference between F3 and F2 in Hz as the dependent variable, Context Group as a fixed factor and Participant as a random factor. The model was then used to calculate pairwise comparisons of the context conditions using the emmeans() and pairs() functions of the emmeans package [20] in R [19]. Degrees of freedom were based on the Kenward-Roger method. P-values were adjusted by the Tukey method comparing 4 estimates. Table 2 shows the results.

Contrast	esti-mate	t.ratio	p
/ɑː/ - back	128	2.43	0.076
/ɑː/ - front	-349	-6.61	< 0.001
/ɑː/ - front round	-53	-0.92	0.794
back - front	-478	-12.6	< 0.001
back - front round	-181	-4.09	0.004
front - front round	296	6.68	< 0.001

Table 2: Statistics for the comparison of formant values (F3-F2 difference) of the laterals between the four types of contexts.

It appears that the formant values, and specifically the difference between F3 and F2, which, if low is typically taken as a measure of retroflexion, differ greatly between context conditions. Interpreting the estimates, it appears that the F3-F2 difference in the lateral is smallest after back (by default rounded) vowels followed by front rounded vowels. Laterals following /ɑː/ tend to pattern with rounded vowels.

Since, however, the articulatory data showed substantial differences between speakers in whether their tongue gestures suggest a retroflex articulation or not, also the acoustic data were inspected on an individual basis. In analogy to the tongue spline data, Fig. 2 illustrates F2 and F3 (in Hz) for /l/ in all items for each speaker. The straight solid line represents the hypothetical line where F2 and F3 would be equal.

Generally, after front (unrounded) vowels, where /l/ is articulated as alveolar by all speakers, the lateral expectedly shows relatively high F2 and F3 values and a rather large F3-F2 difference. In contrast, after back vowels, F2 and F3 as well as their difference are low for /l/ for all speakers and all contexts regardless of retroflex or alveolar articulation. Formant values of /l/ after front rounded vowels, where /l/ is always articulated alveolar, behave similarly to the back vowel context in the male speakers of Vienna and Salzburg. In contrast, in the female speakers they are closer to the front vowel and /ɑː/ contexts. This could explain the lack of difference in F3-F2 values between front rounded vowels and /ɑː/ in the statistical analyses. Finally, in the /ɑː/ context, when /l/ is retroflex in the speakers from Graz and the female speaker from Vienna, a high absolute F2 goes with a small F3-F2 difference. In the remaining two

speakers, alveolar [l] after /ɑː/ patterns with front vowels.

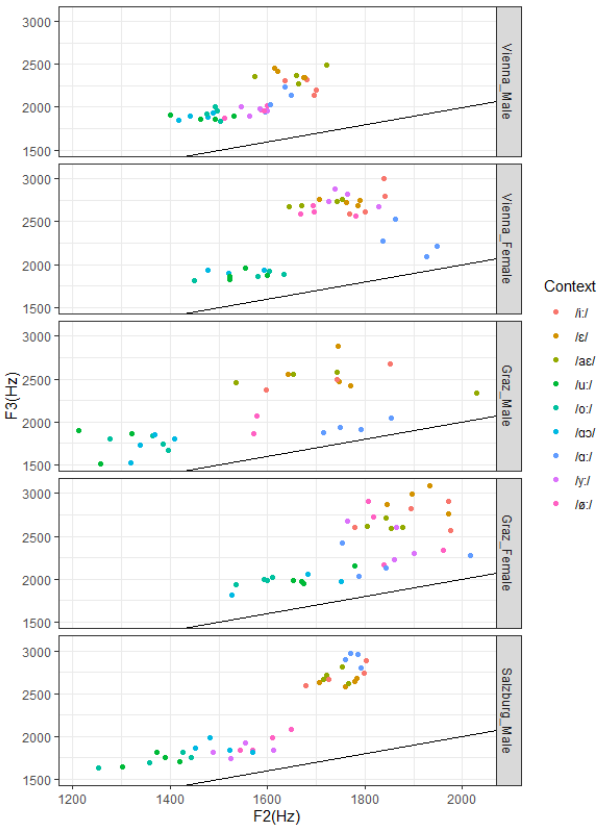


Figure 2: F2 and F3 values in Hz for all items/ vowel contexts per speaker. The grey line indicates the values where the formants would be equal.

3. GENERAL DISCUSSION

The aim of this study was to describe the articulatory and acoustic properties of /l/ in Standard Austrian German in word-final position and to investigate the potential occurrence of retroflex allophones in this position. Lateral productions in 9 vowel contexts from speakers from 3 cities were analyzed.

The articulatory analysis via ultrasound tongue imaging showed that three of five speakers articulated /l/ after back vowels and /ɑː/ as retroflex lateral [ɭ]. This was indicated by a high and relatively posterior tongue tip position. This contrasts with the alveolar articulation of /l/ after front vowels and front rounded vowels that show a relatively low and anterior tongue tip position. In the remaining two speakers, all laterals were found to be articulated as alveolar without any apparent gesture for retroflexion. The acoustic description encompassing F2, F3 and the difference F3-F2 shows less clear-cut results with individual patterns not fully matching expectations based on the articulatory data.

According to the literature, retroflex laterals are characterized by F2 and F3 being in close proximity to each other [8]. For the female speaker from Vienna and both speakers from Graz this assumption was confirmed: After back vowels and /ɑː/ when retroflex tongue gestures are produced, the difference F3-F2 (around 300 to 400 Hz) is much smaller compared to front and front rounded vowel contexts (roughly 600-1000 Hz). Additionally, after back vowels, absolute F2 and F3 values are lower compared to the /ɑː/ context. Curiously, the remaining male speakers from Vienna and Salzburg who according to the articulatory measures only produce alveolar laterals match the other speakers in their acoustic patterns of F2 and F3. Specifically, after front rounded and back vowels their *alveolar* laterals show a difference in F3-F2 similar to the *retroflex* laterals produced by the other three speakers in the same position.

These results may seem puzzling. Despite the difference in articulation visible in the UTI, alveolar and retroflex laterals showed acoustic similarities after back vowels in all speakers. A possible explanation would be that back vowels in German are articulated with the lips rounded, hence coarticulation may cause /l/ after back vowels to also be produced with rounded lips. Lip-rounding tends to lower F2 and F3 [7] and would explain the lower F3-F2 difference in /l/ after back vowels for the alveolar articulations. It would also explain the lower absolute F2 and F3 values in retroflex articulations after back vowels compared to the retroflexes after /ɑː/. However, a lowering effect of preceding lip rounding is absent after front rounded vowels in the female speaker of Vienna and less pronounced in the speakers from Graz. As a methodological consequence, future studies might need to additionally track lip rounding so as to disentangle the origin of lowered F2 and F3 in laterals after back vowels. Note however, that differences in articulatory gestures leading to the same acoustic and even perceptual consequences have been reported for other phones such as retroflexed vs. bunched rhotics in English [7].

From a sociolinguistic perspective, we can add to the existing descriptions of laterals in SAG that retroflexion, first of all, can be found in articulatory gestures as shown by the UTI data. Moreover, it is not exclusive for speakers from Vienna, since retroflex tongue gestures could be shown for the two speakers from Graz. Recording additional participants will show whether the differences found in this pilot study are speaker-specific or whether broader patterns dependent on region, gender or other social factors can be observed.

4. ACKNOWLEDGEMENTS

This study was supported by the Austrian Science Fund Grant F60 Special Research Programme 'German in Austria. Variation – Contact – Perception' (2016–2023) awarded to Alexandra N. Lenz.

5. REFERENCES

[1] Ladefoged, P. and Maddieson, I. 2008. *The sounds of the world's languages*. Blackwell.
[2] Kohler, K. 1990. German. *Journal of the International Phonetic Association*, 20, 1, 48–50.
[3] Moosmüller, S., Schmid, C., Brandstätter, J. 2015. Standard Austrian German. *Journal of the International Phonetic Association*, 45, 3, 339–348.
[4] Luick, K. 1904. *Deutsche Lautlehre: Mit besonderer Berücksichtigung der Sprechweise Wiens und der österreichischen Alpenländer.* ÖBV Pädagogischer Verlag.
[5] Moosmüller, S., Schmid, C., Kasess, C. H. 2016. Alveolar and Velarized Laterals in Albanian and in the Viennese Dialect. *Language and speech*, 59, 4, 488–515.
[6] Recasens, D. 2012. A cross-language acoustic study of initial and final allophones of /l/. *Speech Communication*, 54, 3, 368–383.
[7] Stevens, K. N. 1998. *Acoustic phonetics.* MIT Press.
[8] Narayanan, S., Byrd, D., Kaun, A. 1999. Geometry, kinematics, and acoustics of Tamil liquid consonants. *The Journal of the Acoustical Society of America*, 106, 4, 1993–2007.
[9] Tabain, M., Kochetov, A. 2018. Acoustic Realization and Inventory Size: Kannada and Malayalam Alveolar/Retroflex Laterals and /ɻ/. *Phonetica*, 75, 2, 85–109.
[10] Tabain, M., Butcher, A., Breen, G., Beare, R. 2016. An acoustic study of multiple lateral consonants in three Central Australian languages. *The Journal of the Acoustical Society of America*, 139, 1, 361–372.
[11] Kirkham, S. 2017. Ethnicity and phonetic variation in Sheffield English liquids. *Journal of the International Phonetic Association*, 47, 1, 17–35.
[12] Moosmüller, S. 1991. *Hochsprache und Dialekt in Österreich: Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck.* Böhlau Verlag.
[13] Soukup, B. 2009. *Dialect use as interaction strategy: A sociolinguistic study of contextualization, speech perception, and language attitudes in Austria*. Braumüller.
[14] Fanta-Jende, J. 2021. Situational Effects on Intra-individual Variation in German – Reflexes of Middle High German ei in Austrian Speech Repertoires. In: Werth, A., Bülow, L., Pfenninger, S., Schiegg, M. (eds), *Intra-individual variation in language*. De Gruyter Mouton, 87–126.
[15] Articulate Instruments Ltd. 2012. *Articulate Assistant Advanced User Guide: Version 2.14.* Articulate Instruments Ltd.
[16] Pucher, M., Klingler, N., Luttenberger, J., Spreafico, L. 2020. Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings. *Speech Communication* 123, 83–97.
[17] Noll, A., Stuefer, J., Klingler, N., Leykum, H., Lozo, C., Luttenberger, J., Pucher, M., Schmid, C., Sound Tools eXtended (STx) 5.0—A Powerful Sound Analysis Tool Optimized for Speech. *Proc. Interspeech 2019 - Show&Tell, Graz, Austria.* 2370–2371.
[18] Hadley, W. 2016. *Ggplot2: Elegant graphics for data analysis*. Springer.
[19] R Core Team. 2021. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.
[20] Length, R. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.2.

ASSESSING KINEMATIC RELATIONS WITH HIGH SPEECH RATE RESOLUTION DATA

Stephan R. Kuberski and Adamantios I. Gafos

University of Potsdam, Department of Linguistics and Cognitive Sciences
kuberski@uni-potsdam.de

ABSTRACT

We employ a metronome-driven paradigm of repeated syllable production with systematic speech rate control to assess two important kinematic relations: peak velocity is proportional to movement amplitude, with a proportionality modulated by speech rate, and peak velocity is proportional to the average velocity of a movement. The resulting data show substantially higher kinematic variation than previous works devoted to the same relations. Statistical evaluation of the two relations yields very strong evidence for their presence in the data. Whereas slopes of the first relation are in broad agreement with earlier reports, verified here across a much wider range of speech rates, slopes of the second relation show a rate dependency which has not been previously reported. We consider these results in the context of other work linking slope values to the dynamics underlying the speech movements.

Keywords: articulatory phonetics, speech production, kinematic relations, speech rate

1. INTRODUCTION

In understanding the principles that govern human movements (in general, and speech movements in particular) a large body of previous work has sought to uncover the existence of relations among kinematic parameters such as amplitude (A), peak velocity (v), and duration (T) of movements. In speech, two relations in particular have been at the focus of previous studies: first, peak velocity is proportional to movement amplitude ($v \propto A$), with a proportionality modulated by stress, tone, and speech rate [1–4], and second, peak velocity is proportional to the average velocity of a movement ($v \propto A/T$) [1, 5–7].

Yet evidence for these two relations has remained somewhat limited. A prerequisite to a thorough assessment of any relation is high variation in the parameters to be related. Previous works devoted to the study of the two mentioned relations typically elicited speech movements by conventional verbal instruction (e.g., “speak slow” or “speak fast”). As we demonstrate below, the resulting data exhibit limited kinematic variation. For the present assessment of the two relations, we employed a metronome-driven paradigm of repeated syllable production with systematic, high-range, and high-resolution speech rate control to elicit articulatory movements of much higher, as demonstrated here, kinematic variation.

In the resulting data set, we verify the presence of the two kinematic relations with very strong statistical evidence. The extensive variability elicited in our paradigm furthermore allows an examination of particular rate dependencies of the slope values implicated in the two relations. Regarding the first relation ($v \propto A$), our results verify what is known from previous work, that is, a linear increase in the slope values with increasing speech rate. Regarding the second relation ($v \propto A/T$), however, our results indicate for the first time a declining trend in the slope values as a function of increasing rate. These results and specifically the validity of the two kinematic relations under the rather wide range of speech rate conditions implies the existence of general principles that organize and constrain the functioning of the speech motor system. In the final part of this paper, we thus turn to consider our results in the context of other work drawing bridges between the slope dependency in the $v \propto A/T$ relation and the dynamics underlying the speech movements.

2. METHODS

Five native speakers of German (three female, two male) were recruited at the authors’ institution to participate in an experiment of repeated syllable production. The speakers were between 18 and 27 years old, without any present or past speech or hearing problems. Participation in the experiment was paid and all experimental procedures were performed in compliance with the relevant laws, guidelines and conditions expressed by the authors’ institution.

During the experiment, speakers were asked to produce sequences of repeated /ka/ or /ta/ syllables in time with an audible metronome. The metronome served as an extrinsic index of the intended rate of syllable production with increasing rates between con-

secutive blocks of four trials. Within a trial, speakers were exposed to the beat of the metronome and begun articulation at a point of their choice. Trial lengths (i.e., the number of metronome ticks in a trial) were programmed to allow for the production of 30 consecutive syllables. The rate of the metronome was set to the values of 90, 150, 210, 300, 390, and 480 beats per minute (bpm), covering the range of very slow (90 bpm), slow (down to 150 bpm), normal (around 300 bpm), and fast speech rates (up to 480 bpm); see [8–11] for what ranges of rates are considered normal in German. The experiment begun with the production of /ka/ syllables across all rates, starting with the slowest, and went on with /ta/ syllables afterward.

Data registration was conducted using Electromagnetic Articulography (Carstens AG501). In addition to sensors on rigid reference locations required for head movement correction (left and right mastoids, nose bridge), we tracked the movement of a sensor placed on the tongue body and another on the tongue tip. To preserve the temporal and spatial resolution of the registered data (0.3 mm RMS at 1250 Hz), we employed a for speech not often (e.g., in [12–14]) but in general human movement science frequently used spline smoothing approach. By means of heptic-order splines with a fixed predicted mean square error criterion [15, 16], we obtained signals with more faithful position and velocity characteristics than by traditional filtering and differentiation techniques (see [17], for a comparison of a spline with a filtering approach in speech).

The resulting tongue body (for /ka/) and tongue tip (for /ta/) trajectories were first segmented into movement cycles by means of local tangential velocity minima in the full three dimensions. These minima then served to determine the principal axis of segments defined by their endpoint-to-endpoint vectors in 3-D (referred to as constriction degree in [18] or reaching axis in earlier work [19, 20]). In projecting the three-dimensional movement data onto the principal axes, we finally obtained one-dimensional representations of movements as typically engaged in studies of speech kinematics [6, 21–23]. In a last step, we applied a 20% peak velocity threshold criterion to the data to avoid potential problems with poorly defined transitions into or away from quasi-steady states [24].

In what follows, we refer to movements of positive velocity (motion towards a constriction) as closing movements and to movements of negative velocity (motion away from a constriction) as opening movements; see Figure 1, showing examples of closing and opening movements involved in the production of /ka/ syllables. Each such movement is a period of an effector's continuous motion from

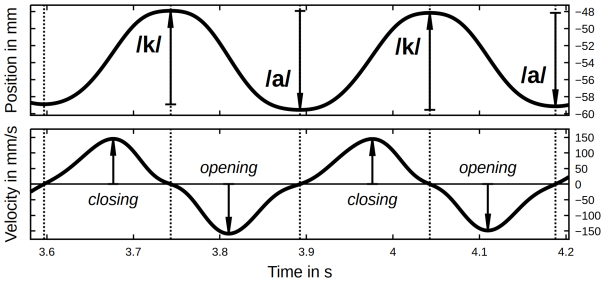


Figure 1: Kinematic trajectories of two /ka/ syllables produced by one of our speakers at normal speech rate. Arrows indicate the kinematic parameters of movement amplitude (top) and peak velocity (bottom).

one quasi-steady state to another, during which the effector's velocity rises from a minimal value up to a maximal value and declines again. Overall, we can identify three temporal landmarks in this period: movement onset, peak velocity, and movement offset. By means of these landmarks, we defined the following kinematic parameters: movement duration (T) as the temporal distance between onset and offset, movement amplitude (A) as the spatial distance between onset and offset, and peak velocity (v) as the value with which velocity peaks within a movement (A and v expressed by the arrows in Figure 1). Across all speakers and metronome rates, we acquired kinematic parameters of 2,470 closing and 2,595 opening movements in the case of /ka/, and 1,948 closing and 2,113 opening movements in the case of /ta/ syllables.

3. RESULTS

We first evaluate the amount of variation elicited in our paradigm by inspecting the range of values in the kinematic parameters defined above. We begin with movement durations, which range from around 45 to 350 ms for /ka/ and from around 35 to 390 ms for /ta/ syllables. Both ranges cover almost one order of magnitude (OM, a tenfold increase in scale), across all speakers and metronome rates.

Figure 2 visualizes movement amplitude and peak velocity as a function of speech rate. Note that, instead of relying on the intended speech rate cued by the metronome, we quantified the actual rate produced by taking the reciprocal value of movement duration (i.e., the abscissae in the figure are in terms of movements per seconds, mps). Data from tongue body movements in /ka/ syllables are shown in the top two panels and data from tongue tip movements in /ta/ syllables are shown in the bottom two panels. As a first result, it becomes clear that, overall, our paradigm successfully elicited movements of a widespread kinematic range: for both syllables,

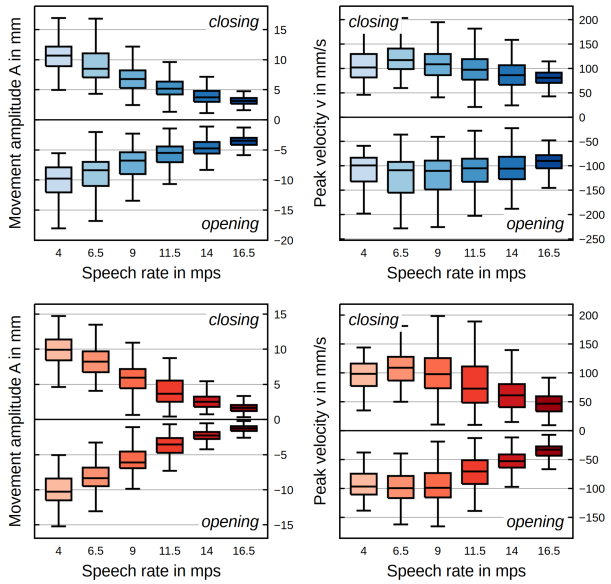


Figure 2: Box plots showing ranges of amplitude (left) and peak velocity values (right) as a function of actual speech rate. Top row shows data of /ka/, bottom row data of /ta/ syllables.

/ka/ and /ta/, the kinematic parameters spread well over the range of an order of magnitude (OM). In particular, for /ka/ syllables, movement amplitudes are in the range of 0.89 to 21 mm (corresponding to 1.4 OM), and peak velocity attains values in the range of 21 to 270 mm/s (1.1 OM); with all ranges given in absolute values. For /ta/ syllables, similar ranges are observed: 0.23 to 16 mm amplitude (1.8 OM), and 7.5 to 210 mm/s peak velocity (1.4 OM).

To facilitate comparison with earlier works [1, 3, 6, 7, 21, 22, 25, 26], we estimated parameter ranges from published tables and graphic representations. Table 1 summarizes the order of magnitudes of these estimates for the ranges of movement duration (T), movement amplitude (A), and peak velocity (v). For example, in a study designed to explore jaw kinematics over a maximally possible range of movement rates, Nelson et al. [25] (second row in the table) reported ranges corresponding to about 0.7 OM in duration, 1.1 OM in amplitude, and 1.0 OM in peak velocity values. Other studies indicate duration ranges in OM of about 0.6 (lower lip) [3, 26], 0.5 (vocal folds) [6], 0.4 (tongue body) [1], and 0.2 (tongue body and tip) [7]; amplitude ranges of about 1.3 (lower lip) [3, 26], 1.1 (tongue body) [21], 0.7 (tongue body) [1], 0.5 (vocal folds) [6], 0.4 (tongue body) [7], and 0.3 (tongue body, tongue tip) [6, 7]; and peak velocity ranges with an estimated OM of about 1.1 (lower lip) [3, 26], 0.7 (tongue body) [21], 0.6 (tongue body) [1], and 0.4 (tongue body, tongue tip) [7]. Overall, then, the kinematic variation achieved by our paradigm (given in the two final rows of Table 1) substantially

Study	Effector	T	A	v
Ostry [21]	tongue body	?	1.1	0.7
Nelson [25]	jaw	0.7	1.1	1.0
Munhall [6]	vocal folds	0.5	0.5	?
	tongue body	?	0.3	?
Ostry [1]	tongue body	0.4	0.7	0.6
Vatikiotis [3, 26]	lower lip	0.6	1.3	1.1
Perkell [7]	lower lip	0.1	0.3	0.3
	tongue body	0.2	0.4	0.4
	tongue tip	0.2	0.3	0.4
present work	tongue body	0.9	1.4	1.1
	tongue tip	1.1	1.8	1.4

Table 1: Comparison of the kinematic variation achieved in different studies. Values which could not be determined from a publication are indicated by a question mark (?). All values in orders of magnitude (OM).

surpasses the majority of ranges in earlier works: order of magnitude values are consistently higher by some tenth of 1 OM. Keeping in mind that an increase of about 0.3 OM corresponds to a doubling of a particular range, it becomes clear that our paradigm elicited movements of almost twice the kinematic variation of previous works in all the kinematic parameters.

We now turn to verify the presence of the two kinematic relations in our data. Figure 3 shows scatter plots of the mutual dependencies between the kinematic parameters as expressed by the two relations, $v \propto A$ (peak velocity is proportional to amplitude, left) and $v \propto A/T$ (peak velocity is proportional to average velocity, right). In the figure, speech rate is color-coded, with fainter shades corresponding to movements of slower, and darker shades corresponding to movements of faster rates. Data from movements of different directions separate into different quadrants of the shown coordinate systems: opening movements reside in the lower left and closing movements in the upper right quadrant (due to our sign convention of the kinematic parameters introduced earlier). Inspection of Figure 3 reveals a clear systematicity, as expected from the two kinematic relations. First, there is a proportionality between peak velocity v and movement amplitude A as expressed by the relation $v \propto A$, with steeper slopes for faster rates (color-coded by different shades). Second, there is a proportionality between peak velocity v and average velocity A/T as expressed by the relation $v \propto A/T$ with a visually indiscernible dependency on speech rate.

To statistically evaluate the empirical evidence for the two kinematic relations, a simple linear regression model without an intercept term was fit to the data. Resulting correlation strengths for /ka/ and /ta/ sylla-

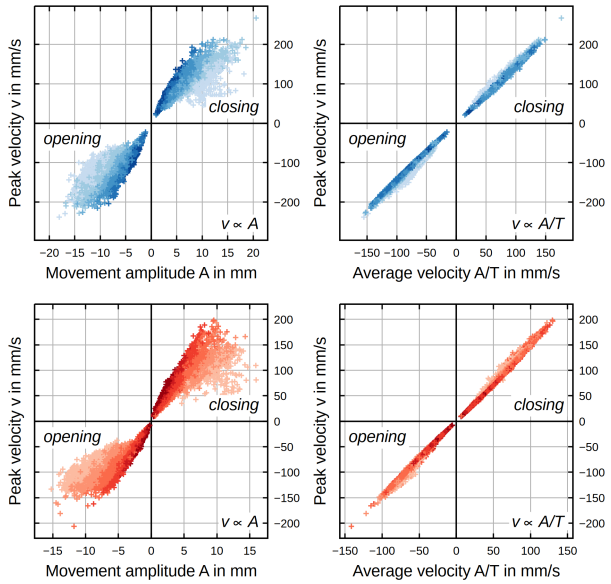


Figure 3: Scatter plots showing the two kinematic relations $v \propto A/T$ (left) and $v \propto A$ (right). Top row shows data of /ka/ syllables, bottom row data of /ta/.

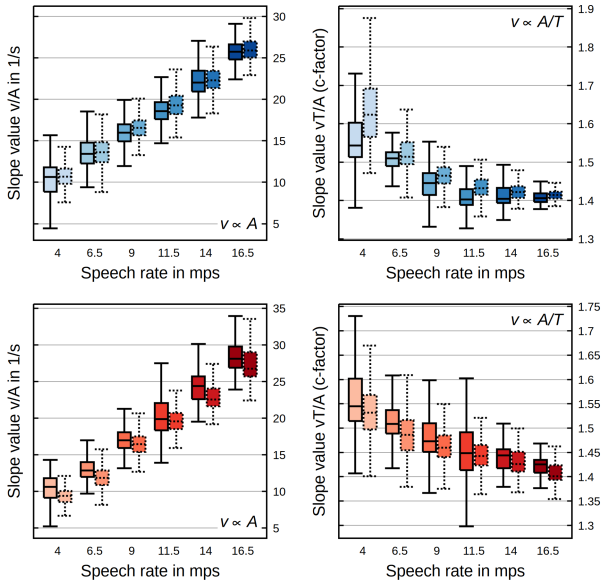


Figure 4: Box plots showing the slope values of the two kinematic relations, $v \propto A$ (left) and $v \propto A/T$ (right). Top row shows data of /ka/ syllables, bottom row data of /ta/. Data of closings (solid) and openings (dotted) are shown side-by-side within each speech rate bin.

bles are in the order of 0.65–0.95 R-squared for $v \propto A$, strictly increasing with speech rate, and 0.95–0.99 for $v \propto A/T$. All p-values reside below 0.001, implying very strong statistical evidence for the presence of the two kinematic relations. Regarding the slopes of the two relations, we evaluated the ratios v/A of each movement for the first, and vT/A for the second relation. Figure 4 then shows the average values as a function of speech rate. For both syllables, these are linearly increasing slopes for the relation $v \propto A$ (left) and declining slopes for the relation $v \propto A/T$ (right). Both speech rate dependencies are significant, with Spearman's rank correlation strengths of 0.99 in case of the first and −0.65 in case of the second relation.

The observed linear increase in the slopes of the $v \propto A$ relation as a function of speech rate is in broad agreement with earlier assessments [1–4], but verified here across a much wider range of speech rates and much higher resolution. Reports on the other $v \propto A/T$ relation [1, 5, 6, 27, 28], however, do not provide any indications of its slope changing with varying speech rate. The literature rather suggests that slope values of the relation remain invariant over a variety of linguistic manipulations including speech rate [1, 5, 6]. However, it must be kept in mind that these studies did not manipulate speech rate to the same extent as in our paradigm.

4. CONCLUSION

In sum, we have employed a paradigm of repeated syllable production with systematic speech rate control to assess two kinematic relations, $v \propto A$ and $v \propto A/T$. The resulting data set shows substantially higher kinematic variation than previous works devoted to the same relations. Statistical evaluation of the two relations yields very strong evidence for their presence in the data. Whereas slopes of the $v \propto A$ relation have been shown to be in broad agreement with earlier reports, slope values of the $v \propto A/T$ relation suggest a rate dependency which has not been previously reported.

The rate dependency of the $v \propto A/T$ relation can be considered in the context of other work which aims to link values of its slope to the dynamics underlying the movements. Specifically, [27] and [5] have pointed out that the relation's slope, referred to as the c-factor, characterizes the shape of a movement's velocity profile (e.g., sinusoidal, trapezoidal, or triangular). In turn, velocity profile shape has served as a diagnostic of the dynamical organization underlying movement. For example, sinusoidal velocity profile shapes are indicative of dynamics corresponding to linear second-order systems, whereas trapezoidal shapes are indicative of dynamics constrained by a minimal impulse of movement [27]. Hence, in future work, we plan to harness the variation in slope values, as uncovered at high resolution in our paradigm, and relate it to the underlying dynamics governing the movements. We further aim to include the less frequently used parameter of peak acceleration which may play out in additional kinematic relations whose slopes, in turn, may provide further information.

ACKNOWLEDGMENT

This work was funded by the German Research Foundation (DFG) under Project ID 317633480, SFB 1287, Project C04. Data collection has been partially funded by the European Research Council (ERC) under Grant agreement ID 249440.

5. REFERENCES

[1] D. J. Ostry and K. G. Munhall, "Control of rate and duration of speech movements," *The Journal of the Acoustical Society of America*, vol. 77, no. 2, 1985.

[2] J. A. S. Kelso, E. Vatikiotis-Bateson, E. L. Saltzman, and B. Kay, "A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling," *The Journal of the Acoustical Society of America*, vol. 77, no. 1, 1985.

[3] E. Vatikiotis-Bateson and J. A. S. Kelso, "Rhythm type and articulatory dynamics in English, French and Japanese," *Journal of Phonetics*, vol. 21, 1993.

[4] B. Kühnert and P. Hoole, "Speaker-specific kinematic properties of alveolar reductions in English and German," *Clinical Linguistics & Phonetics*, vol. 18, no. 6–8, 2004.

[5] K. G. Munhall, "Temporal adjustment in speech motor control: evidence for laryngeal kinematics," PhD thesis, McGill University, Canada, 1984.

[6] K. G. Munhall, D. J. Ostry, and A. Parush, "Characteristics of velocity profiles of speech movements," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 4, 1985.

[7] J. S. Perkell, M. Zandipour, M. L. Matthies, and H. Lane, "Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, 2002.

[8] A. Gerstenberg, S. Fuchs, J. M. Kairet, J. Schröder, and C. Frankenberg, "A cross-linguistic, longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French," in *9th International Conference on Speech Prosody*, 2018.

[9] F. Pellegrino, C. Coupé, and E. Marsico, "A cross-language perspective on speech information rate," *Language*, vol. 87, no. 3, 2011.

[10] F. Pellegrino, J. Farinas, and J.-L. Rouas, "Automatic estimation of speaking rate in multilingual spontaneous speech," in *International Conference on Speech Prosody*, 2004.

[11] V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *15th International Congress of Phonetic Sciences*, 2003.

[12] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, 1995.

[13] J. O. Ramsay, K. G. Munhall, V. L. Gracco, and D. J. Ostry, "Functional data analyses of lip motion," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, 1996.

[14] S. Schötz, J. Frid, A. Löfqvist, and L. Gustafsson, "Functional data analysis of tongue articulation in Gothenburg and Malmöhus Swedish /i:, y:, u-:/," in *Fonetik*. Linköping University, 2013.

[15] H. J. Woltring, "A Fortran package for generalized, cross-validatory spline smoothing and differentiation," *Advances in Engineering Software*, vol. 8, no. 2, 1986.

[16] S. R. Kuberski, "GCV spline smoothing in Matlab," 2023. [Online]. Available: github.com/srkuberski/gcvspl

[17] S. R. Kuberski and A. I. Gafos, "Higher-order kinematic analysis of speech movement data," in *14th International Conference on Bio-inspired Systems and Signal Processing*, 2021, conference poster.

[18] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, 1989.

[19] E. L. Saltzman and J. A. S. Kelso, "Skilled actions: a task-dynamic approach," *Haskins Laboratories Status Report on Speech Research*, vol. SR-76, 1983.

[20] ——, "Skilled actions: a task-dynamic approach." *Psychological Review*, vol. 94, no. 1, 1987.

[21] D. J. Ostry, E. Keller, and A. Parush, "Similarities in the control of the speech articulators and the limbs: kinematics of tongue dorsum movement in speech," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 4, 1983.

[22] D. J. Ostry, J. D. Cooke, and K. G. Munhall, "Velocity curves of human arm and speech movements," *Experimental Brain Research*, vol. 68, no. 1, 1987.

[23] S. G. Adams, G. Weismer, and R. D. Kent, "Speaking rate and speech movement velocity profiles," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, 1993.

[24] P. Hoole, C. Mooshammer, and H. G. Tillmann, "Kinematic analysis of vowel production in German," in *Third International Conference on Spoken Language Processing*, 1994.

[25] W. L. Nelson, J. S. Perkell, and J. R. Westbury, "Mandible movements during increasingly rapid articulations of single syllables: preliminary observations," *The Journal of the Acoustical Society of America*, vol. 75, no. 3, 1984.

[26] E. Vatikiotis-Bateson and J. A. S. Kelso, "Linguistic structure and articulatory dynamics: a cross language study," *Haskins Laboratories Status Report in Speech Research*, vol. SR-103/104, 1990.

[27] W. L. Nelson, "Physical principles for economies of skilled movements," *Biological Cybernetics*, vol. 46, no. 2, 1983.

[28] S. Fuchs, P. Perrier, and M. Hartinger, "A critical evaluation of gestural stiffness estimations in speech production based on a linear second-order model," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, 2011.

Visualising phonetics at the neuromuscular level reveals the discrete pulsatile gated nature of speech

Please write XXX instead of the name(s) of the author(s)

Please write XXX instead of the affiliation(s)
please write XXX instead of the email address(es)

ABSTRACT

Glossometry, lip-video, photoglottography and a new understanding of the compartmental nature of lingual neuromuscular structure, provide the opportunity to study speech production from a motor control perspective. External photoglottography is used to visualise glottal opening/closing; pose-estimation of profile lip-video to visualise lip protrusion and closure; and pose-estimation of ultrasound to visualise the degree of doming/grooving of five independently controlled sectors of the tongue body and an independently controlled tongue blade/tip. A short utterance is analysed to demonstrate the potential of these new tools. Motor commands to agonist/antagonist teams of neuromuscular compartments appear to be initiated at regular discrete intervals of approximately 100ms. The velocity of each compartment contraction appears to be planned such that each phone, comprising a set of motor commands, is timed to reach its target in one or two 100ms periods. Thalamic or oliviocerebellar pulsing neurons are suggested as a mechanism for gating and synchronising motor commands.

Keywords: ultrasound, timing, photoglottography, glossometry, motor control

1. INTRODUCTION

This paper proposes a set of lingual measures which, it is claimed, represent underlying motor commands. In combination with a measure of glottal opening and a measure of lip movement, patterns of movement can be observed. Most apparent is the synchronicity and regularity of the motor commands. This in turn leads to the radical proposal that speech consists of sets of motor commands issued sequentially at discrete intervals regulated by a central "clock".

2. THE COMPARTMENTAL TONGUE

A recent review of the anatomical structure of the tongue [14] indicates the genioglossus muscle may be divided into five independently controlled neuromuscular compartments (Fig 1) which work along with compartments of transversus and verticalis muscles to form five synergistic agonist/antagonist units that control the tongue body. The hypothesis proposes that antagonistic contraction of genioglossus and transversus neuromuscular compartments controls the amount of grooving or doming of each sector independently. This hypothesis is supported by the measurement of distances from EMA-type keypoints on the tongue surface to the origin of the genioglossus muscle which attaches to the mandible via a short tendon. These five points show independent coronal doming or grooving for different vowels and consonants.

A little-known subcompartment of the inferior longitudinal muscle (Oblique longitudinal) arises from the tongue root and extends to the anterior of the tongue body. It is proposed that its purpose is to constrain the rostral/caudal expansion of the genioglossus, thus enhancing and stiffening the bunching of the tongue body.

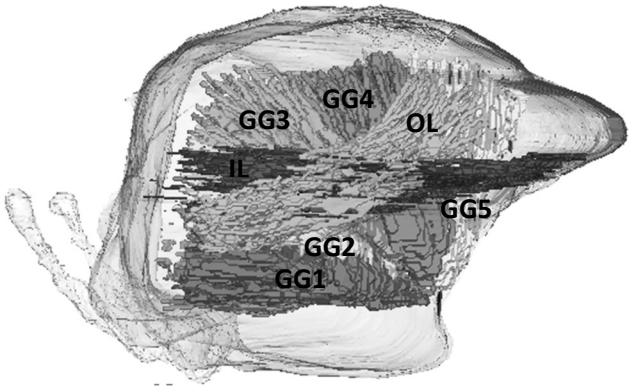


Figure 1: Sagittal view of tongue with 5 independently innervated sectors of genioglossus GG1-5; OL - Oblique longitudinal; IL - inferior longitudinal)

3. METHOD: INSTRUMENTATION AND ANALYSIS

3.1. Ultrasound with DeepLabCut pose estimation

DeepLabCut pose estimation was trained on approximately 2000 hand-labelled ultrasound images of speech and water swallows from five different ultrasound systems, seven different probe geometries, child and adult speakers. 14 keypoints are labelled. 11 points on the tongue, one on hyoid,

one on base of mandible (inferior tubercle) and one on the short tendon (superior tubercle).

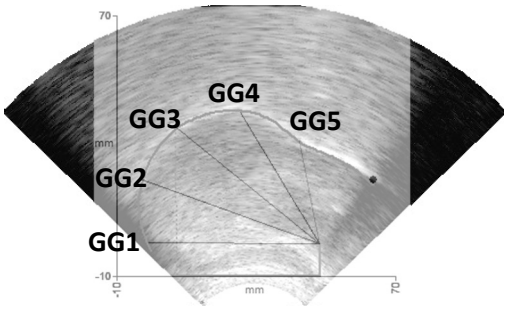


Figure 2: DeepLabCut pose estimation of tongue (11 keypoints) and short tendon with distances GG1-GG5 superimposed

From root to tip the first seven keypoints (Fig 3.) indicate the tongue body and the last four indicate the blade/tip. For the purpose of this study, the tongue body contour from keypoint 1-7 is then interpolated along its length to extract five equally spaced points to represent the five neuromuscular compartments and labelled GG1 (root)-GG5(anterior tongue body) (see Fig 2). Distances from these five points to the origin of the short tendon on the superior tubercle of the mandible are referred to herein as glossometric analysis.

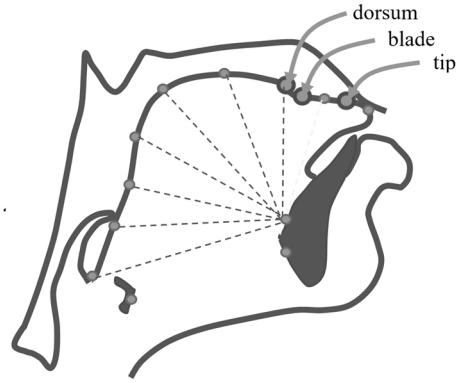


Figure 3: Sagittal diagram of the tongue with 11 pose estimation keypoints marked on the tongue contour, one on the hyoid one on each of the superior and inferior tubercles of the mandible. Three highlighted points are selected for correlation comparison with co-registered EMA data.

Validity of the pose-estimated lingual keypoints was evaluated on a co-registered dataset [8] with electromagnetic articulograph (EMA) dorsum, blade and tip sensors and synchronous ultrasound images (81Hz). Both systems recorded data in mm and the pose estimated data was rotated and translated to match the EMA data. The EMA sensors were correlated with the nearest pose-estimated keypoints (Fig 3.) for around 700 ultrasound frames taken from three phonetically balanced sentences from a single speaker (unseen in the pose-estimation training set). Table 1. Shows the x and y Pearson correlation scores. Y-correlations are high (>0.80)1. X-correlations are lower due to challenges in accurately training the tip position which is not always visible in the ultrasound image.

	tip	blade	dorsum
x	0.46	0.59	0.67
y	0.81	0.88	0.91

Table 1: Pearson correlation scores for tongue tip, blade and dorsum EMA sensors vs closest pose-estimated keypoints derived from co-registered ultrasound images.

3.2 Lip video with DeepLabCut pose estimation

Lip camera images we trained using DeepLabcut. Seven keypoints were labelled: superior and inferior vermillion border of upper and lower lips, commissure and upper and lower keypoints midway between lips and commissure. (see Fig 4.)

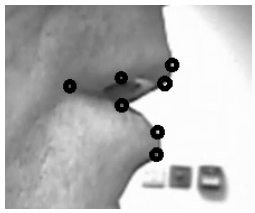


Figure 4: DeepLabCut pose estimation of Upper 4 keypoints and lower lip 4 keypoints

3.3 Photoglottograph

An external photoGlottograph (ePGG) was built similarly to the design by Honda & Maeda [5] and Amelot et al [1] and synchronised with audio, ultrasound and lip camera using sync pulses from the Micro ultrasound system. The ePPG signal is sampled at 2880Hz and reveals the glottal opening gesture with three gross states – open (high value), close (low value) and midway for voicing.

4. RESULTS: MOTOR COMMAND ANALYSIS OF THE PHRASE “THE CHALK”

Figure 5 shows a recording of the phrase “the chalk” /ðəˈtʃɔk/. Different patterns of coronal doming (high values) and grooving (low values) are observed for GG1-GG5. Vertical lines indicate time points where motor commands appear to initiate movement towards a following phonetic target. For example, motor command to transition from ʧ- ɔ starts at the beginning of region D. The glottis takes 1 period to reach its voicing target but GG2, 3 & 4 take 2

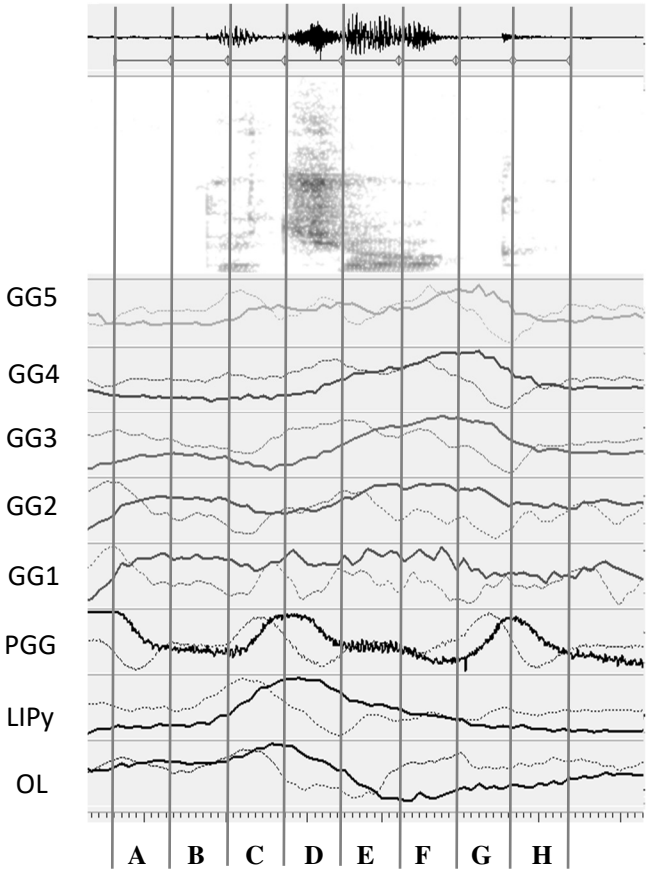


Figure 5: "the chalk" /ðəˈtʃɔk/ Vertical lines represent gating clock with clock periods labelled A-H. GG1-5 Genioglossus compartments; PGG Glotttal opening; LIPy lip commissure raising; OL oblique longitudinal compartment.
A # – ð **B&C** ð – ʧ˒ **D&E** ʧ- ɔ **F** ɔ – k **G&H** k release
Fine dotted lines are contraction velocities [94ms clock]

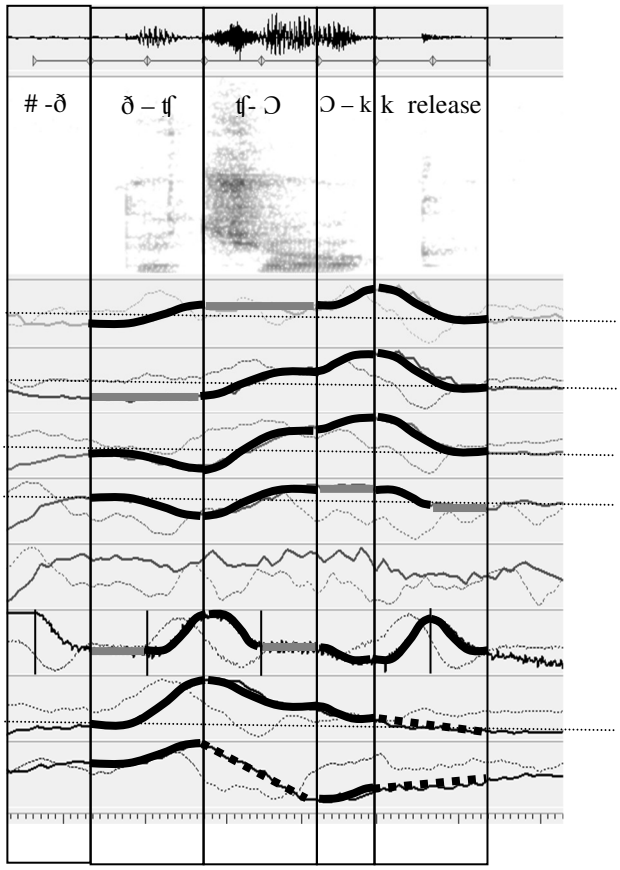


Figure 6: "the chalk" A duplicate of figure 5 but with the motor commands highlighted as sigmoid transitions – black; linear transitions – dotted; or sustained– grey. Horizontal dotted lines represent relaxed neutral positions for the synergistic unit. GG1 signal is too noisy to interpret.

periods to reach their target. This transition can also be observed in the audio spectrogram by the F2 formant movement. A command to retract the lip commissure also takes two periods. By contrast, a single period (F) is required to transition from Ɔ – k (closure target). This involves doming of GG4 &5 and closure of the glottis. Finally, the k closure release takes 2 periods (G&H) involving relaxation of GG1-5 back to a neutral position. The glottis opens in one period for a voiceless release and immediately closes again.

5. CONCLUSIONS

Although it is not prudent to draw any firm conclusions from such a small speech sample, several significant observations can be made.

Motor commands for all the articulators (tongue, lips, glottis) appear to be initiated at regular intervals of approximately 100ms. In this example the period between initiation of movement appears to be 94ms. (vertical lines in Fig 5)

Many of the transitions are sigmoid shape implying a bell-shaped velocity profile. This is most easily observed in the glottal signal where the bell-shaped velocity trace is clearly apparent in the plot (fine dotted line).

The velocity of the transition can be controlled. The GG4 transition for ʧ- ɔ takes ~200ms whereas a similar increase in doming of GG4 is achieved in ~100ms for the ɔ – k closure transition. The contraction velocity is apparently not an inherent property of a neuromuscular compartment synergy but instead can be programmed.

Transitions towards a target are not necessarily initiated at the same instant. In the ð – ʧ˒ transition both tongue and lip trajectories are initiated ~200ms before the target is reached but the glottis remains in voiced position (to realise the schwa) and movement toward the ʧ˒ target is initiated 100ms later.

Some acoustic segments arise as part of the transition towards a subsequent phonetic target. In the ð – ʧ˒ transition, the short central vowel of

"the" emerges as the tongue moves smoothly from dental closure for ð to palatal closure for ʧ.

The example exhibits no evidence of simultaneous initiation of transitions for two consecutive phones. The syllable onset consonant ʧ and the following vowel nucleus are not initiated synchronously. The transition to the ɔ vowel syllable nucleus starts at the beginning of the affrication and takes 200ms to reach the vowel target. It is perhaps easy to see why a vowel transition starting at the beginning of the affricate "segment" would be interpreted as simultaneous initiation. However, the ð – ʧ transition is initiated 200ms before the frication starts and reaches its target at the start of acoustic frication. The subsequent 100ms of frication is a consequence of the slow lingual transition to the vowel and the open glottis.

The majority of the speech signal is transitional. In this example, as soon as a target is reached the transition to the next target is initiated. [Caveat: the authors have viewed other examples where targets are sustained for at least one additional period perhaps for prosodic purposes]

6. DISCUSSION

6.1 Physiological tremor and rythmic neurons

In 1886 Horsely & Schäfer [6] observed the curve of voluntary muscle contraction invariably showed a series of undulations that succeed each other with almost exact regularity at about 10 times a second. This has become known as physiological tremor. This 8-12-Hz physiological tremor is thought to originate from the central nervous system, possibly thalamic or olivocerebellar, rhythmic neurons but the function of these central oscillations remains a subject of debate [11]. One function theory proposes that the tremor may have a role in the timing of voluntary commands [3]. Welsh and Llinás [13] have suggested that discontinuous timing of motor output would simplify the computational demand in comparison to computing commands for every instant of a movement. It has been further suggested that the 8-12Hz oscillation in the alpha wave band provides a gating mechanism that is important for synchronization of pulsatile motor signals, providing functional co-ordination of particular combinations of muscle actions [8][9][12].

Children under 10 have been observed to have a slower tremor frequency of around 6Hz compared with the adult rate of around 10Hz [10].

6.2 The pulsatile nature of movement and speech

If the premise that descending motor commands are gated at discrete intervals of approximately 100ms is accepted, then it follows that the descending motor commands are pulsatile. Put another way, the predominant idea that trajectories of movements are controlled by continuously updated descending commands and continuous feedback throughout the movement is incompatible with the concept of discrete gated commands.

In the example utterance, three of the transitions took two 100ms periods consisting of an initial pulsatile command with the option of a second corrective adjustment to that command 100ms later in the middle of the transition.

6.3 Phonological and phonetic link

This small example reveals the motor plan for the phrase "the chalk" as consisting of five discrete lingual/lip transitions to targets for ð ʧ ɔ k and a final neutral configuration. The schwa between ð and ʧ is efficiently planned to emerge as part of the ð - ʧ transition. These motor command targets map to locations in multidimensional auditory space and the cerebellum is capable of learning this mapping. As proposed by Guenther [4], phonemes may be learnt as a locus in this multidimensional auditory sensory space. More than one set of transitional commands may reach an auditory target within the locus for the desired phonological unit and the appropriate one must be selected so that the transition ends in or passes through that locus. The efficiency of the transition and how close to the centre of the phonemic auditory locus will influence which set of motor commands are selected and sequenced.

7. SUMMARY

A radical, disruptive, new theory of speech production is proposed through observations made using a new set of visualisation tools. Observation of the contraction patterns of 5 independent tongue compartments, the glottis and lips reveals motor commands that appear to be issued synchronously and discretely every ~100ms. An 8-12Hz gating of descending motor commands is proposed, generated by pulsing neurons. This central timekeeper synchronises the activation of all the articulators. It is further observed that these motor commands are programmed to reach their target contractions in either one or two cycles and in combination they specify an articulatory/auditory phone target. The tools to make these observations are widely available to anyone via the AAA software application. It is hoped that these tools can be much more widely used by other research groups to investigate many questions regarding timing, rhythm, stress and variation in speech production from this new perspective.

8. REFERENCES

[1] Amelot, A., Sathiyanarayanan, D., Maeda, S., Honda, K., Crevier-Buchman. L. 2018. Validation of a Non-Invasive System to Observe Glottal Opening and Closing: External Photoglottograph (EPGG), *11th International Conference on Voice Physiology and Biomechanics (ICVPB)* August 1-3, 2018, East Lansing, Michigan.

[2] Garcia-Rill, E., D'Onofrio, S., Luster, B., Mahaffey, S., Urbano, F. J., & Phillips, C. 2016. The 10 Hz Frequency: a fulcrum for transitional brain states. Translational brain rhythmicity, 1(1), 7.

[3] Goodman, D., & Kelso, J. A. S. 1983. Exploring the functional significance of physiological tremor: a biospectroscopic approach. *Experimental Brain Research*, 49(3), 419-431.

[4] Guenther, F. H. 2016. Neural control of speech. Mit Press.

[5] Honda K. and Maeda S. Non-invasive photoelectro-glottography method and device. *U.S. patent* 20100256503A1 (October 7, 2010).

[6] Horsley, V. A. H., & Schäfer, E. A. 1888. I. A record of experiments upon the functions of the cerebral cortex. *Philosophical Transactions of the Royal Society of London.(B.)*, (179), 1-45.

[7] Kirkham, S., Strycharczuk, P., Gorman, E., Nagamine, T., Wrench, A. Co-registration of simultaneous high-speed ultrasound and electromagnetic articulography for speech production research *In Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, Prague, 7-11 August 2023. International Phonetic Association.

[8] Llinás, R. 1991. The noncontinuous nature of movement execution, *in Motor Control: Concepts and Issues,* eds D. Humphrey and H. Freund (New York, NY: Wiley), 223–242.

[9] Llinás, R.R., 2014. The olivo-cerebellar system: a key to understanding the functional significance of intrinsic oscillatory brain properties. *Frontiers in neural circuits*, 7, p.96.

[10] Marshall, J. 1959. Physiological tremor in children. *Journal of Neurology, Neurosurgery, and Psychiatry*, 22(1), 33.

[11] McAuley, J. H., & Marsden, C. D. 2000. Physiological and pathological tremors and rhythmic central motor control. *Brain*, 123(8), 1545-1567.

[12] Welsh, J. P., Lang, E. J., Suglhara, I., & Llinás, R. 1995. Dynamic organization of motor control within the olivocerebellar system. *Nature*, 374(6521), 453-457.

[13] Welsh, J. P., & Llinás, R. 1997. Some organizing principles for the control of movement based on olivocerebellar physiology. *Progress in brain research*, 114, 449-461.

[14] Wrench, A., & Balch-Tomes, J. 2023 The compartmental tongue. *Journal of speech, language, and hearing research*, In Preparation

THE ROLE OF LISTENER FEEDBACK IN PROSODIC CUE PRODUCTION: AN INTERACTIVE TASK

Lahari Chatterjee[1], Kathleen Schneider[1], Isabell Wartenburger[1], Sandra Hanne[1], Andrea Hofmann[1], Outi Tuomainen[1]

[1]University of Potsdam, Cognitive Sciences, Dept. Linguistics
lahari.chatterjee@uni-potsdam.de, kathleen.schneider.1@uni-potsdam.de, isabell.wartenburger@uni-potsdam.de, sandra.hanne@uni-potsdam.de, andrea.hofmann@uni-potsdam.de, outi.tuomainen@uni-potsdam.de

ABSTRACT

The aim of the study is to investigate the situational (in)dependence of production of prosodic boundary cues in an interactive setting. A game-like task was designed in which we tested whether native German speakers, when asked to repeat, change their speaking style while communicating with a confederate listener. In the task, they produced coordinate name sequences in German, with and without grouping of constituents, and received feedback from the listener. If prosody is used for the listener, we expect that after the feedback requesting a repetition of an item, speakers would switch from "casual" to "clear" speaking style.

Preliminary results showed that they use prosodic boundary cues (f0 range, final lengthening and pause) to mark groupings of the constituents but cue productions stayed consistent irrespective of listener feedback. Our results speak in favour of prosody being produced for the speaker and not primarily for the listener, i.e. situationally independent.

Keywords: boundary cue production, situational (in)dependence, listener feedback, clear speech.

1. INTRODUCTION

In dyadic face-to-face conversations, the listener signals understanding/misunderstanding of the message to the speaker via verbal and nonverbal feedback cues (e.g., clarification requests and back channelling) [1,2]. As a result, whenever there is a communication breakdown, speakers often adapt their speech and change from a conversational ("casual") to a more hyper-articulated ("clear") speaking style to increase the intelligibility of their speech [3,4]. These acoustic-phonetic adaptations often include reductions in speaking rate, increases in pause frequency, pause duration and fundamental frequency of speech [4]. Thus, it has been suggested that speech produced in interaction is highly dynamic and dependent on the perceived needs of the listener (i.e., listener-oriented and therefore situationally dependent) [5,6].

However, not all aspects of the speech signal are shown to be adaptive and listener-oriented [7, 8]. For example, a recent study by Huttenlauch et al. [9], investigated the adaptivity of prosodic boundary cues that signal syntactic grouping of three constituents in coordinate name sequences (henceforth coordinates). Speakers were asked to produce coordinates in two conditions (grouping vs no grouping) (see 1 and 2 below) and in five varying contexts corresponding to different virtual listeners introduced via a video recording (Young Adult, Child, Elderly, Non-native, Young Adult in Noise). It was assumed that if prosodic boundary cue production is listener-oriented, speakers will vary their productions between these different contexts.

(1) (Name1 und Name2) und Name3 (grouping)
(2) Name1 und Name2 und Name3 (no grouping)

The results showed, first, smaller f0 range and a decrease in final lengthening and pause durations of Name1; second, an increase in f0 range and final lengthening of Name2 as well as longer pause durations after Name2 in the grouping condition (1), relative to the no grouping condition (2). These findings are in line with the syntax-prosody model proposed by Kentner and Féry [10]. Yet, in terms of the five different contexts, speakers showed only limited adaptations to the different interlocutors. The authors concluded that prosody is not produced with the listener's needs in mind, but it is speaker-oriented and situationally independent [7,8]. However, the study by Huttenlauch et al. [9] relied on virtual listeners who provided no feedback to the speaker. Therefore, it is possible that these "passive" listeners were not real enough to elicit listener-oriented speech adaptations. To account for this, the present study investigates situational independence of prosodic boundary cue production in an interactive game-like setting with a real interlocutor (a confederate listener) who provides real-time feedback on the communicative success (understood/ misunderstood/ repeat request).

We hypothesize that, if prosodic boundary cue production is listener-oriented and situationally dependent, speakers will adapt their speech to the

confederate's feedback and produce casual speech when the confederate listener understands them correctly. However, if the confederate listener requests for a repetition, they would change their speaking style to clear speech. This is expected to result in enhanced prosodic boundary cue productions (e.g., increase in f0 range and final lengthening of constituents and longer pause durations). If, however, prosodic boundary cue production is situationally independent, the speaker would not be influenced by the confederate listener's feedback and prosodic cue productions would remain consistent regardless of whether the listener asked them to repeat.

2. METHOD

2.1. Participants

So far, we recorded 5 adult native speakers of German (male and female, mean age = 26.6. years, *SD* = 5.59; we are aiming for a final sample of 30 speakers). All participants reported normal hearing and no history of language impairments. Informed consent was obtained from the participants and they were remunerated or received course credits for participation. The study was approved by the ethics committee of the University of Potsdam.

2.2. Stimuli

Experimental items were adapted from [9] and [11]. As we needed a bigger number of foil items to test the two speech styles (casual/clear) relative to the listener feedback type (understood/misunderstood/repeat), we added more names in addition to those used in [9, 11] to construct the name sequences. Experimental items included seven coordinate name sequences consisting of three names coordinated by '*und*' (*and* in English). Each name (n=15) used in the sequences was disyllabic and trochaic. Seven of the total of fifteen names, ending in the high front vowel /i/ to reduce glottalisation [9] (Gabi, Manni, Mimmi, Moni, Nelli, Leni, Lilli), were used as Name1 and Name2 for the experimental items. Four names (Manu, Nina, Lola, Lisa) appeared as Name3 in the experimental and filler items, while four additional names (Dora, Lotte, Mira, Sara) served only as filler items.

Each name sequence appeared in two grouping conditions, grouping and no grouping (see examples 3-4).

2.3. Procedure

In the course of the experiment, the speaker sat inside a sound-attenuated audio booth, while the confederate listener (member of the research team) was seated outside the booth. The interlocutors could see and hear each other via headphones/microphones. In order to make the listener's miscomprehension more realistic, the speaker was informed that the confederate listener heard background noise (20-talker BABBLE noise) during the task.

The experiment began with a practice phase consisting of seven example trials where the participants were familiarised with the names at first and then with the two grouping conditions (see examples in 3 and 4). Thereupon, the test phase consisted of 4 blocks of 120 trials with 3 self-paced breaks every 30th trial. During the breaks, in order to make it more realistic, the interlocutors were given a chance to interact with each other and discuss their performance.

In the introduction phase, the speaker was informed that the listener would do an auditory word-picture-matching task. In the test phase, the task began with a question "*Wer kommt?*" (*Who is coming*?), followed by a fixation cross for 1000 ms. Next, the speaker was shown three pictures, each consisting of a visual presentation of three name sequences as well as the corresponding name sequence (see 3-5) written underneath the picture. There was one target picture with the corresponding name sequence written underneath (3) and two other pictures and sequences displaying a different grouping condition with the same names (4), or the same grouping condition with different names (5), respectively. The speaker was then prompted to produce the highlighted target sequence with the intended grouping (either with or without grouping; see examples 3-4).

(3) *(Moni und Lilli) und Lisa*
(4) *Moni und Lilli und Lisa*
(5) *(Lola und Manu) und Leni*

There were three forms of listener feedback: i) the confederate listener understood the speaker correctly and the speaker received feedback on their screen in which a green tick appeared above the target item and the next trial followed; ii) the confederate listener misunderstood the speaker and a red cross appeared above the misunderstood item and the next trial followed; iii) the confederate listener requested a repetition of the item in which case an orange question mark appeared on the screen. The analyses focus on the coordinate sequences produced in the feedback condition iii, where we expect the speaker to change their speaking style from casual to clear when they repeat the item. The task took approximately thirty minutes in total. Both in the practice and the test phase the confederate listener responses were automated to ensure that on approximately 11% of randomly introduced trials, the

confederate listener requests for a repetition. A feedback questionnaire at the end confirmed that the speakers were not aware of this manipulation.

2.4. Annotation and data processing

Out of the total of 140 productions (7 name sequences * 2 grouping conditions (grouping and no grouping) * 2 speaking style (casual and clear) * 5 speakers), 107 productions, recorded at a sampling rate of 44100 Hz (16 bit), were manually annotated using Praat [12]. Due to technical issues during the recording, 33 productions were discarded. Three acoustic cues were measured on Name1 and Name2: f0 range, final lengthening of the last vowel, and pause duration; identical to the criteria employed by Huttenlauch et al. [9]. For f0 range, the range between f0 min and f0 max was measured in semitones formula used for calculation: $12*\log_2(f0_{max}/f0_{min})$ [9]. Further, f0 was time normalised and the resulting contours are plotted in Hz as seen in Figures 1-3. Final lengthening was calculated as the duration of the final vowel (i.e., the last segment) of Name1 and Name2 relative to the total duration of Name1 and Name2 (in %), respectively. Lastly, pause duration was measured as the duration of the pause following Name2 relative to the total duration of the utterance (in %). Preliminary statistical analysis was run on the percentage scores.

First, we report the results for the grouping conditions (grouping vs no grouping). Second, we report the results for the two speaking styles (casual vs clear) in response to the listener feedback. Our dependent variables were f0 range, final lengthening and pause and our independent variables are the two grouping conditions (grouping vs no grouping) and speaking style (casual vs clear). Because in this paper we report preliminary data from only 5 speakers, the statistical analyses focus on the main effects of the grouping condition and speaking style (and not their interaction), for each dependent variable separately (a non-parametric Mann Whitney U test). Results from the linear mixed effect modelling for the full dataset and all comparisons will be reported in the poster.

3. RESULTS

Preliminary results showed a larger f0 range on Name1 in the no grouping condition *(M=7.06, SD= 3.55)* compared to the grouping condition *(M=4.80, SD= 4.85)* (*p<0.001*) (see Figure 1). With regards to speaking style, there was no significant difference (*p*=0.067) (see Figures 2 and 3).

In addition, f0 range on Name2 in the grouping condition was larger (*M=11.51, SD=5.27*) than the f0 range on Name2 (*M=6.77, SD=5.20*) in the no grouping condition (*p<0.001*) (see Figure 1). However, there was no significant difference with regards to the speaking style (*p=0.189*) (see Figures 2 and 3).

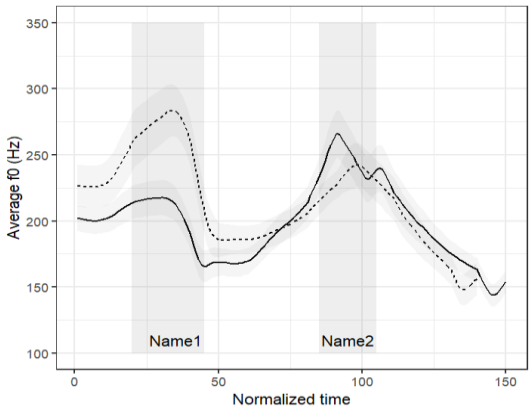


Figure 1: Time normalised f0 (in Hz) contours for the two grouping conditions. Solid line indicates the grouping condition and dashed line the no grouping condition.

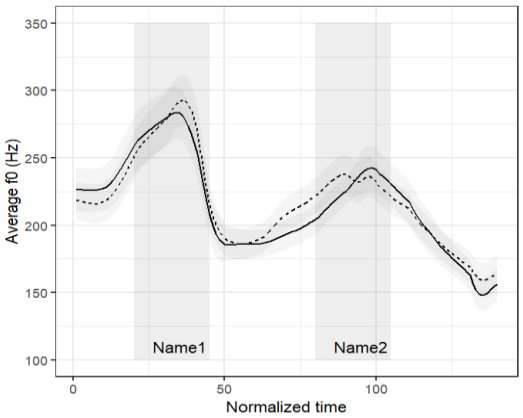


Figure 2: Time normalised f0 (in Hz) contours for the two speaking styles in the no grouping condition. Solid line indicates the casual speaking style and dashed line the clear speaking style.

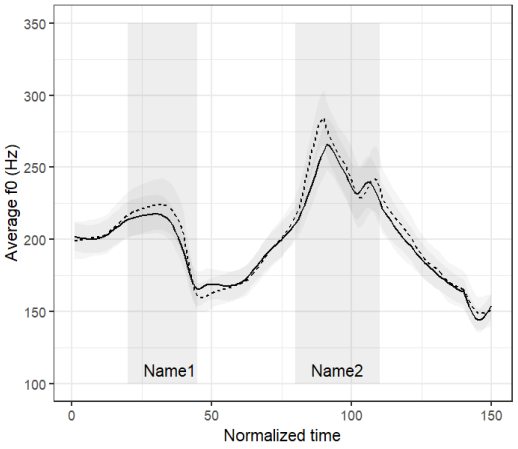


Figure 3: Time normalised f0 (in Hz) contours for the two speaking styles in the grouping condition. Solid line indicates the casual speaking style and dashed line the clear speaking style.

Moreover, the duration of the final segment of Name1 in the grouping condition shortens (*M=31.15, SD=4.85)* as compared to the final segment duration on Name1 (*M=34.90, SD=4.57*) in the no grouping condition (*p<0.001*). But final segment lengthening of Name1 showed no statistically significant difference between speaking styles (*p=0.062*).

Further, the duration of the final segment of Name2 in the grouping condition lengthens (*M=42.20, SD= 5.69*), as compared to final segment duration on Name2 (*M=34.08, SD=5.46*) in the no grouping condition, with *p<0.001*. Relative to the speaking style, we observed a statistically significant effect (*p<0.05)*. The duration of the final lengthening of Name2 in clear speaking style (*M= 39.44, SD=6.60*) is longer than the casual style (*M=36.46, SD=6.91)*.

Lastly, the pause after Name2 was longer (*M=16.49, SD=10.68*) in the grouping condition as compared to the no grouping condition (*M=0.20, SD=0.85*) showing a significant difference *(p<0.001)*. However, no statistically significant difference was found for the speaking style (*p=0.719*).

In sum, results showed statistically significant effects of the grouping condition on the f0 range on Name1 and Name2, final lengthening on Name1 and Name2, relative to the no grouping condition. In addition, there was a significant effect of the grouping condition on the pause duration on Name2. However, we found limited changes in regards to speakers modifying their speaking style as a response to the listener feedback. However, these results need to be validated with more data.

4. DISCUSSION

The aim of the study was to investigate the situational (in)dependence of prosodic boundary cues in an interactive setting. We investigated how speakers modify their speech from casual to clear speaking style after they were requested to repeat an item by the confederate listener. Preliminary results with five speakers showed that the three cue measurements (f0 range, final lengthening and pause) clearly mark the difference between the two grouping conditions indicating a close and stable syntax-prosody link. These results replicate those reported by Huttenlauch et al. [9], Kentner and Féry [10], Petrone et al. [13].

However, with regards to the speaking style, our results showed that there are no changes, or only limited changes in some of the cues (e.g., Name2 final lengthening), in the three acoustic cues between casual and clear speaking styles, thus, partially replicating the results by Huttenlauch et al. [9] obtained with "virtual" interlocutors, now with a "real" interlocutor. It is worth noting that here we reported data from only five speakers and it is, therefore, possible that with a larger sample of speakers we will see listener-oriented adaptations also in (at least some) prosodic boundary cues. As a next step, for the casual-to-clear speaking style modifications, we will run further exploratory analyses for cues other than f0 range, final lengthening and pause, and places other than Name1 and Name2.

As our preliminary data indicates that speakers do not modify their prosodic cue production in response to the listener feedback (apart from one cue), our data support the assumption that prosodic boundary production is independent of the listener. In the context of listener-oriented accounts [5], these results can be somewhat surprising. For example, it is well known that speakers increase their pitch range when speaking to infants and smaller children in contrast to when they are speaking to another adult [14]. However, it is possible that at the prosody-syntax interface, in specific contexts, prosody might serve different functions and could be unaffected by conversational situations as suggested by previous studies [10,15]. Alternatively, it could also be that the speakers were already maximally marking the prosodic boundaries (i.e, at ceiling) in the casual style. Thus, the fact that prosodic cue productions are independent of the interlocutor or the communicative situation, speaks in favour of prosody being situationally independent and mainly produced for the speakers themselves. However, so far, we have tested only a small and preliminary sample, and more data is needed to validate these results.

5. REFERENCES

[1] Krauss, R. M., Garlock, C. M., Bricker, P. D., McMahon, L. E. 1977. The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology* 35(7), 523.

[2] Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R. 2007. Fluid Semantic Back-Channel Feedback in Dialogue: Challenges and Progress. In: Pelachaud, C., Martin, JC., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds) *International Workshop on Intelligent Virtual Agents*. Springer, 154-160.

[3] Buz, E., Tanenhaus, MK., Jaeger, TF. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89, 68-86.

[4] Tuomainen, O. T., & Hazan, V. 2018. Investigating Clear Speech Adaptations in Spontaneous Speech Produced in Communicative Settings. In: M. Gósy, T. E. Grászi (Eds.), *Challenges in analysis and processing of*

spontaneous speech. Budapest, Hungary: MTA Nyelvtudományi Intézet, Budapest.
[5] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In: W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* Dordrecht. Springer, 403–439.
[6] Smiljanić, R., and Bradlow, A. R. 2009. Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass* 3, 236–264.
[7] Clifton, C., Carlson, K., Frazier, L. 2002. Informative prosodic boundaries. *Language and Speech* 45, 87–114.
[8] Speer, S. R., Warren, P., & Schafer, A. J. 2011. Situationally independent prosodic phrasing. *Laboratory Phonology* 2(1), 35–98.
[9] Huttenlauch, C., et al. 2021 Production of prosodic cues in coordinate name sequences addressing varying interlocutors. Laboratory Phonology: *Journal of the Association for Laboratory Phonology* 12(1), 1–31.
[10] Kentner, G., & Féry, C. 2013. A new approach to prosodic grouping. *The Linguistic Review* 30(2), 277–311
[11] De Beer, C., Regenbrecht, F., Huttenlauch, C., Wartenburger, I., Obrig, H., Hanne, S. 2021. Kommunikative Beeinträchtigungen nach rechtshemisphärischer Hirnläsion: Produktion und Perzeption von Prosodie. In Jaecks, P., Hussmann, K., Mantelli, A., Monaco, E., Ruflin, D. (eds). *Aphasie und verwandte Gebiete | Aphasie et domaines associés N°1/2021.*
[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.01.
[13]Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., & Höhle, B. (2017). Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics* 61, 71–92.
[14] Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. 2002. Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior & Development, 24*(4), 372–392.
[15] Kraljic, T., & Brennan, S. E. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50(2), 194–231.

AERODYNAMIC PARAMETERS OF EJECTIVES AND PULMONIC STOPS[1]

Maria-Josep Solé

Universitat Autònoma de Barcelona
mariajosep.sole@uab.cat

ABSTRACT

This paper presents preliminary results of an aerodynamic study of ejective, aspirated and unaspirated stops. Specifically, peak oral pressure, peak oral flow, area of flow and the interaction of each of these with place of articulation were examined. Differences between ejectives and voiceless stops were found in oral pressure (with ejectives having double or triple the value for pulmonic stops), oral airflow and area of flow (with ejectives showing a small and short peak of flow compared to aspirated stops). The rate of oral pressure build-up (slope) varied considerably in ejectives reflecting the less regular raising movement of the larynx. Concerning the effect of place of articulation, oral pressure was higher in velars than anterior places of articulation in ejectives but not in pulmonic stops. Oral airflow tended to be higher in alveolar stops than labials and velars in all stop categories.
Keywords: ejectives, oral pressure, oral flow

1. INTRODUCTION

The aim of this paper is to characterize the aerodynamic parameters of aspirated, unaspirated and ejective stops for linguistic, clinical, and modelling purposes. A second aim is to illustrate how aerodynamic data provide a window into speech articulation. Because aerodynamic properties of speech sounds derive from the action of the articulators (which may compress/rarefy the air and/or impede the airflow), pressure and flow patterns allow us to infer articulatory gestures and their timing in a non-invasive way. For example, oral pressure (P_o) and flow provide information about the degree of glottal constriction in stops [1]. Similarly, the peak of oral flow before (and after) fricatives, particularly /s/, indicates anticipatory glottal opening before the oral constriction is formed [2].

While the aerodynamic characteristics of voiced and voiceless stops are well-described, e.g., [3, 4, 5, 6], there is only a limited amount of aerodynamic data on ejectives, e.g., [7, 8, 9]. This paper focuses on ejectives and compares them to aspirated and unaspirated pulmonic stops. We describe specifically (i) peak oral pressure, (ii) oral flow, (iii) area of oral flow, and (iv) the interaction of each of these with place of articulation in ejective and pulmonic stops. A detailed description of the pressure and flow patterns of ejectives will contribute to a better understanding of production and aerodynamic properties of stops in general.

The articulation and acoustics of ejectives (and implosives) are only relatively well-understood, in part due to (i) their phonetic variation across languages (see cross-language comparisons of the acoustic characteristics of ejectives in [10, 11, 12]), and (ii) their gradual differentiation from pulmonic stops, such that in some languages plain voiceless stops are variably ejective (e.g., Shekgalagari, Ikalanga, Amharic), and plain voiced stops may be realized as implosives. Despite their variation in realization, the category ejective will be used here to describe their aerodynamic properties which, as their articulatory and acoustic properties, may vary in each language.

2. BACKGROUND

We first review the oral pressure and flow patterns of ejectives as described in previous studies and illustrated in our data. Ejective stops are produced with a glottalic egressive airstream mechanism, using the air enclosed between the closed glottis, the closed velum, and a closure in the mouth. A rapid upward movement of the larynx reduces the volume of the oropharyngeal cavity and thus compresses the air within that cavity which is pushed out with a strong burst when the oral constriction is released. This is illustrated in Fig. 1 (left) which shows the sequence [ap'a] produced by a trained phonetician (see section 3.1). When the larynx is raised the air pressure in the oral cavity increases to values as high as 20-30cmH_2O (double or triple the value for pulmonic stops). Because the volume of the oral cavity sealed at both ends is relatively small (ranging from approximately 100cm^3 for labials to 30cm^3 for velars, [13, 14]), a small elevation of the closed glottis makes a large percentage reduction in volume, resulting in high P_o values.

The P_o rise in ejectives is frequently less steady and more variable than for the corresponding plain voiceless stops (with a typical 'hat' shape). This is illustrated in Fig. 2 which shows different P_o pulses of ejectives for the same speaker, some with a relatively steady P_o rise, some with a two-stage increase, or a concave shape (delayed oral pressure

build up) (cf. [9]) —what they all have in common is a high increase in oral pressure before the oral release. A less steady rise in oral pressure reflects the relatively less regular, and variable in time, upward movement of the larynx (which increases air pressure in the oral cavity) compared to the large and steady column of air through the open vocal folds for plain voiceless stops. Note that the oral pressure peak, which typically occurs at closure release for plain stops, may occur before oral release in ejectives (approximately 25ms in Fig. 1) indicating that larynx elevation may be completed a few tens of ms before the stop release. This is corroborated by data on larynx movement —obtained with high-speed video recording— and simultaneous aerodynamic and acoustic data, [15].

Figure 1: Waveform, 0-7 kHz spectrogram, oral pressure (in cmH_2O), and oral flow (in l/sec) for [ap'a] (left) and [ap^{h}a](right). The release of the oral closure is indicated by a vertical bar.

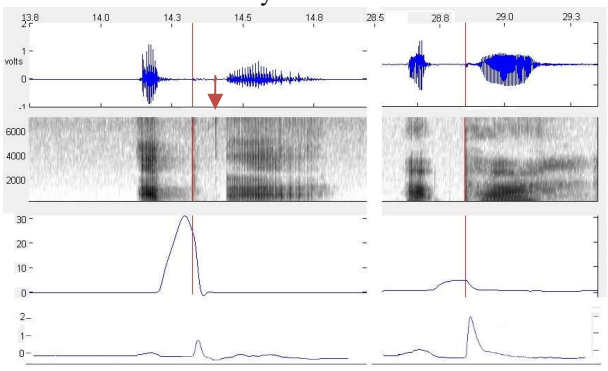
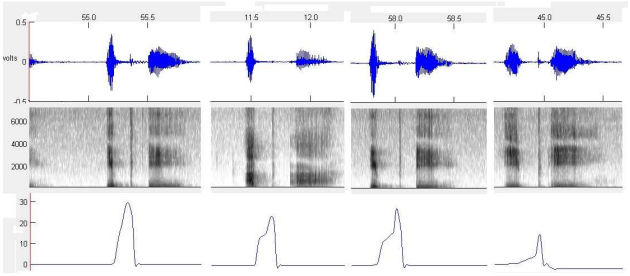


Figure 2: Variation in oral pressure pulses (in cmH_2O) for [ap'a] for a single (male) speaker.



The release of the glottal closure typically lags behind the oral release (glottal lag) or may be synchronic with the oral release (see e.g. [10] for cross-language differences). In Figure 1 (left), the arrow indicates the release of the glottal closure (the vertical bar on the spectrogram at time 14.4) which takes place approximately 80ms after the oral release. When the oral closure is released, the high oral pressure behind the closure creates a high amplitude burst[2] not followed by noise (referred to as low or absent 'post-burst noise amplitude' in the literature) – this is because, differently from plain stops, the glottis remains closed at the point of oral release for ejectives such that there is insufficient rate of flow through the oral constriction to create noise. Figure 1 illustrates the small and short burst of oral flow for ejectives – reflecting the limited air volume enclosed in the oral cavity – as opposed to the large and longer flow for aspirated stops with an open glottis. The time interval between oral release and onset of glottal vibration – VOT with low or inexistent noise– may range between 23ms-80ms in different languages (Ingush, Hausa, Quiche, Navajo, Tigrinya, Amharic and Waima'a, [11, 12]) and is characterized by P_o and oral flow to 0 as the glottis remains closed.

3. METHOD

3.1. Data acquisition

Aerodynamic and acoustic data were obtained for target words read in a carrier sentence by one male and one female phonetician in a laboratory setting. The carrier sentence was 'Say__'. The target consonants – aspirated [p^h t^h k^h], unaspirated [p t k] and ejective stops [p' t' k'] – were preceded and followed by /a/, with stress on the second syllable, e.g. [a'p^{h}a][3]. Six repetitions for each word were recorded, giving 108 tokens in total.

The subjects' productions were recorded using National Instruments PCI-6013 data acquisition hardware and the Matlab Data Acquisition Toolbox (20kHz sample rate per channel and 16 bits/sample). Oral pressure was obtained by a catheter inserted into the pharynx via the nasal cavity and connected to a pressure transducer (Biopac TSD160C). The invasive nature of the procedure limited the number of potential subjects. Oral airflow and nasal airflow were collected with a split Rothenberg mask and a Fleisch pneumotachograph with Biopac TSD160A pressure transducer. The airflow and oral pressure signals were lowpass filtered at 50Hz. Statistical tests were, unless otherwise indicated, conducted on the pressure and flow measurements using one-way ANOVAs in R.

3.2. Measurements

The following measurements were obtained: (1) oral pressure at onset and offset of P_o build-up, (2) duration of pressure build-up, (3) peak oral pressure, (4) peak oral flow, (5) duration of flow, and (6) area under the curve (AUC) of flow in litres[4].

The onset of oral pressure build-up in measurement (1) was defined as the point in time that oral flow drops to the baseline (i.e., complete oral closure) and P_o begins to rise. Measurements (1) and (2) were intended to allow us to calculate the rate of P_o build-up (or slope). However, because the P_o rise varied considerably for ejectives (showing one-stage, two-stage and concave increments, see Fig. 2), the slope was not calculated to avoid averaging across disparate shapes. The AUC, that is, the 2-dimensional

region below the oral airflow curve, that captures the rate of flow over time, was calculated with Matlab.

4. RESULTS

We present the results of the measurements, looking in particular at the behavior of ejectives relative to other voiceless stop types, and at possible place of articulation dependencies concerning pressure and flow values. Results will be presented separately for the female and the male speakers since ANOVAs indicated that the speaker factor interacted with stop type and place in oral pressure and flow values.

4.1. Peak oral pressure

As reported in earlier findings, ejectives have much greater oral pressure than aspirated and unaspirated stops (M= 31.3, 6.5, 4.6, respectively for the female speaker and M= 26.1, 11.15, 9.18 for the male speaker), see Fig. 3.

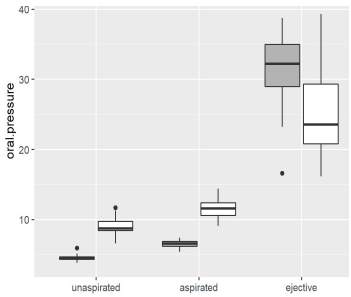
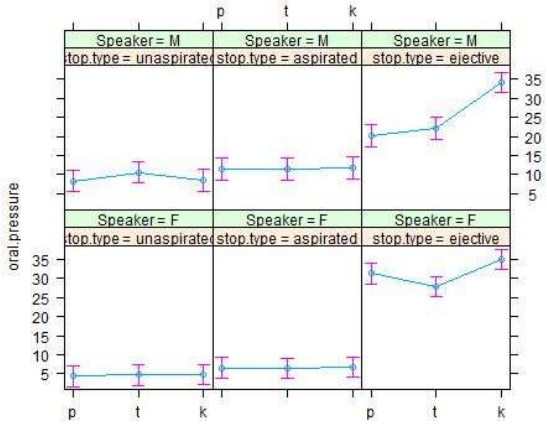


Figure 3: Average peak oral pressure (in cmH_2O) for unaspirated, aspirated and ejective stops for a female (grey) and a male (white) speaker.

The effect of stop type on oral pressure was significant for the male [$F_{(2, 51)}$: 60, p <0.01] and female speaker [$F_{(2, 51)}$: 287.4, p <0.000], with ejectives showing higher P_o values than both aspirated and unaspirated stops for both speakers (p<0.0001 in all cases). The difference between aspirated and unaspirated stops was smaller and did not reach significance for any of the two speakers (p=0.512 and p=0.365 for the male and female speakers, respectively).

Fig. 4 shows oral pressure for the three stop types broken down by place of articulation. Because there was a significant interaction between stop type and place of articulation, separate one-factor ANOVAS were performed. Among ejective stops, oral pressure is higher in velar than in labial (p <0.01) and alveolar stops (p <0.01) for the male speaker. For the female speaker, only the P_o differences between [k'] and [t'] approach, but do not reach, significance (p= 0.176). Labial and alveolar ejectives show similar pressure values for both speakers. Within the unaspirated series, only the alveolar stop shows higher oral pressure values than labial (p= 0.001) or velar stops (p= 0.01) for the male speaker. No effect of place of articulation was found in the aspirated series in any of the two speakers.

Figure 4: Average peak oral pressure for unaspirated, aspirated and ejective stops by place of articulation for a male (M) and female (F) speaker.



4.2. Peak oral flow

As expected, aspirated stops, with a continuous flow of air through the open glottis at release, exhibit a considerably higher peak flow and longer duration of high airflow than ejective and unaspirated stops. Ejectives show a short and small peak of oral flow at release reflecting the limited volume of air enclosed in the oral cavity compared to aspirated stops. The peak flow for ejectives is intermediate between that obtained for aspirated and unaspirated stops (M= 0.536, 1.265, 0.315, respectively for the female speaker and M= 0.833, 2.302, 0.228 for the male speaker), see Fig. 5. The larger peak flow for ejectives than unaspirated stops is presumably due to the higher oral pressure for the former. All the paired comparisons across stop types were significant for both the female and the male speaker (p< 0.000 in all cases except p<0.01 for ejective vs unaspirated for the female speaker).

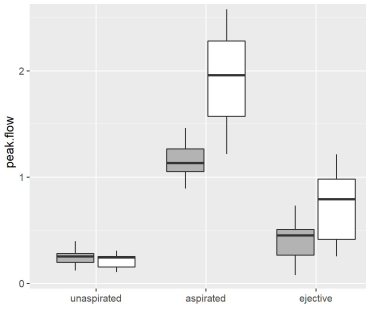
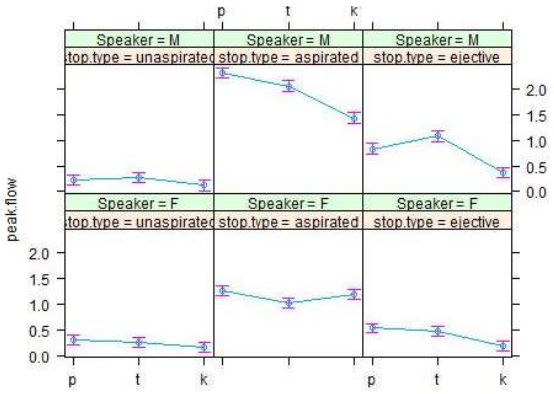


Figure 5: Average peak flow (in l/s) for unaspirated, aspirated and ejective stops for a female (grey) and a male (white) speaker.

The effect of place of articulation patterns rather consistently across stop type for the two speakers (Fig. 6). Peak oral flow is significantly lower for velars than more anterior places of articulation for all stop types (all differences p<0.001, except no difference between /k/ vs /t/ for aspirated stops for the female speaker). Alveolars tend to show significantly higher peak flow values than labials (within ejectives and unaspirated stops (p <0.01) for the male speaker), or similar peak flow values to labials (the rest of paired comparisons).

Figure 6: Average peak flow for unaspirated, aspirated, and ejective stops by place of articulation for a male (M) and female (F) speaker.



4.3. Area of flow

The values for area of flow (i.e., area under the curve) show a very similar pattern to those for peak oral flow both for stop type and place of articulation. The only difference between the two sets of measurements was that while the female speaker had lower peak flow values for aspirates than the male speaker, see Fig. 5, (due to the lower absolute P_o values for the female), the area of flow was comparable for both speakers, Fig. 7, (due to the longer duration of flow for the female compared to the male speaker which offset the lower peak flow). One-way ANOVAs showed that aspirated stops have a significantly larger area of flow, $p<0.0001$ for both speakers, than ejective and unaspirated stops which do not differ between them (M= 0.154, 0.014, 0.008, respectively for the female speaker and M=0.163, 0.021, 0.006 for the male speaker), see Fig. 7.

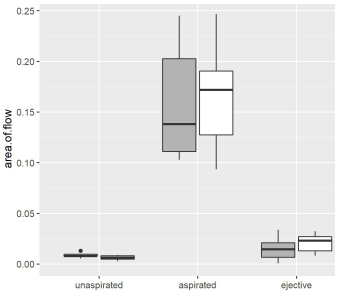


Figure 7: Average area of flow (in litres) for unaspirated, aspirated and ejective stops for a female (grey) and a male (white) speaker.

The effect of place of articulation on area of flow was comparable to that observed for peak oral flow: anterior places of articulation have a significantly larger area of flow than velars for both speakers. For the male speaker, the differences are only significant for labial/alveolar vs velars for all stop types (ejective, unaspirated, aspirated; $p<0.01$ in all cases). For the female speaker the differences are significant for labial vs alveolars/velars ($p<0.01$ in all cases).

5. DISCUSSION

The finding that ejectives show much greater *oral pressure* than pulmonic stops is in accord with the reports of earlier researchers. Because the volume of the oral cavity between the oral and the glottal closure for ejectives is relatively small, a small elevation of the larynx makes a large reduction in volume, resulting in high P_o values. Differences in P_o between aspirated and unaspirated stops may be attributed to differences in glottal resistance during the stop closure [16]. Aspirated stops, with the vocal folds maximally apart, tend to have higher oral pressure than unaspirated stops, with a smaller glottal opening because the vocal folds begin to approximate during the stop closure. The difference in P_o between aspirated and unaspirated stops, however, did not reach significance (as in other studies, e.g., [17]). Concerning place of articulation, the higher P_o expected for velars than more anterior places of articulation – due to a smaller oral cavity volume and faster P_o rise – was only observed in ejectives but not pulmonic stops. In fact, the effect of oral cavity volume on pressure values has not been consistently found by other investigators (e.g., [3]). As suggested in [5], for voiceless stops the fast increase in P_o associated with glottal abduction may simply outweigh supraglottal place effects.

The results show decreasing *peak oral flow* and *area of flow* for aspirated > ejectives > unaspirated stops. Aspirated stops, with a large and continuous flow of air through the open glottis, when released have high peak flow and longer duration of high airflow (and hence larger area of flow). Unaspirated stops have a smaller and more brief airflow peak (and area of flow) than aspirated stops due to the smaller glottal opening: during the stop closure, the vocal folds begin to adduct, and on release there is glottal vibration which offers considerable resistance to transglottal and oral flow. The small and short peak of airflow for ejectives (with a closed glottis) reflects the limited volume of air enclosed in the oral cavity. Peak airflow may vary with place of articulation, with alveolars showing a tendency to have greater airflow than velars and labials, and velars showing the lowest peak airflow, in all stop categories. The higher airflow in alveolars is in accord with previous results for aspirated stops in [18, 19, 20]. The lower peak and area of flow values for velars in all stop categories is unexpected given the usually higher P_o values, and longer VOT, for back compared to anterior stops. The fact that the Rothenberg mask picks up more readily oral flow closer to the transducers (as is the case for anterior places of articulation) – just as in microphone popping– may have played a role.

There is ample room for further investigation of aerodynamic parameters in non-pulmonic stops. Our preliminary results have thrown some light on the production and aerodynamic features of stop types.

5. REFERENCES

[1] Dart, S. 1987. An aerodynamic study of Korean stop consonants. *J. Acoust. Soc. Am.*, Vol. 81, 138–147.

[2] Solé, M.J. 2010. Effects of syllable position on sound change: An aerodynamic study of final fricative weakening. *Journal of Phonetics* 38.2, 289–305.

[3] Subtelny, J. D., Worth, J. H., Sakuda, M. (1966). Intraoral pressure and rate of flow during speech. *J. Speech Hear. Res*. 9, 498–518.

[4] Löfqvist, A. 1992. Acoustic and aerodynamic effects of interarticulatory timing in voiceless consonants. *Language and speech* 35.1, 2, 15–28.

[5] Fuchs, S., L. L. Koenig. 2009. Simultaneous measures of electropalatography and intraoral pressure in selected voiceless lingual consonants and consonant sequences of German. *J. Acoust. Soc. Am.*, Vol. 126.4, 1988–2001.

[6] Solé, M.J. 2018. Articulatory adjustments in initial voiced stops in Spanish, French and English. *Journal of Phonetics* 66, 217–241.

[7] Pinkerton, S. 1986. Quichean (Mayan) glottalized and non-glottalized stops: A phonetic study with implications for phonological universals. In: Ohala, J.J, Jaeger, J. (eds.), *Experimental Phonology.* Orlando: Academic Press, 125–139.

[8] Flemming, E., Ladefoged, P., Thomason, S.G. 2008. Phonetic structures of Montana Salish. *Journal of Phonetics* 36.3, 465–491.

[9] Demolin, D. 2004. Acoustic and aerodynamic characteristics of ejectives in Amharic. *J. Acoust. Soc. Am.* 115.5, 2610.

[10] Lindau, M. 1984. Phonetic differences in glottalic consonants. *Journal of Phonetics*, 12 (2): 147–155.

[11] Warner, N. 1996. Acoustic characteristics of ejectives in Ingush. *Proc. 4th ICSLP*, Philadelphia. 1525–1528.

[12] Seid, Hussein, Rajendran, R., Yegnanarayana, B. 2009. Acoustic characteristics of ejectives in Amharic. *Proc. Interspeech 2009*, Brighton, U.K., 2287-2290.

[13] Catford, J.C. 1977. *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

[14] Story, B.H., Titze, I.R, Hoffman, E.A. 1996. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am*. 100.1, 537-554.

[15] Sprouse, R., Solé, M.J., Ohala J.J. 2010. Tracking laryngeal gestures and their impact on voicing. *Proc. LabPhon 12*, Alburquerque, New Mexico, 127-128. [17] Seid et al 2009,

[16] Löfqvist, A. and McGowan, R.S. 1992. Influence of consonantal environment on voice source aerodynamics. *Journal of Phonetics* 20, 93-100.

[17] Dixit, R.P., Brown, W.S. 1978. Peak magnitudes of subglottal air pressure associated with affricated and nonaffricated stop consonant productions in Hindi. *Journal of Phonetics* 6: 353–365.

[18] Klatt, D.H., Stevens, K.N, and Mead, J. 1968. Studies of articulatory activity and airflow during speech. In: Bouhuys, A. (ed.), *Sound production in man*, Annals of the New York Academy of Sciences 155, art1. 42–55.

[19] Gilbert, H.R. 1977. Oral airflow during stop consonant production. *Folia Phoniatrica* 25, 288–301.

[20] Stathopoulous, E.T. 1980. *Normative airflow study of children and adults using a circumferentially-vented pneumotachograph mask*. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

[1] This study was supported by grants PID2020-117804GB -I00 from the Ministry of Science and Innovation, Spain and 2021 SGR00544 of the Generalitat de Catalunya.

[2] Demolin (2004) reports a higher amplitude burst for ejectives than pulmonic stops, but Seid et al (2009) report the opposite.

[3] As one reviewer pointed out, the results from such metalinguistic task divorced from context and natural realizations should be taken with caution. But because the results are in agreement with previous studies, we feel confident that they reflect ejective realisation in natural languages.

[4] AUC was calculated integrating flow rate (in l/s) with respect to time (s), therefore, the unit for time becomes eliminated.

EFFECTS OF AUTONOMIC AROUSAL AND COGNITIVE LOAD ON SENSORIMOTOR ADAPTATION OF VOICE

Authors: Nicole Tomassi[1], Alyssa Williams[2], Dea Turashvili[3], Cara Stepp[1,2,3,4]

Affiliations: Graduate Program for Neuroscience, Boston University[1], Department of Speech, Language, and Hearing Sciences, Boston University[2], Department of Biomedical Engineering, Boston University[3], Department of Otolaryngology, Boston University School of Medicine[4]
Email addresses: ntomassi@bu.edu

ABSTRACT

Shared neural substrates suggest possible relationships among cognition, autonomic arousal, and speech motor control, but these systems have yet to be studied cohesively. Thus, the purpose of this study was to investigate the effects of cognitive load and autonomic arousal on sensorimotor adaptation of voice. Adults with typical speech (n = 24) were exposed to persistent errors to voice fundamental frequency (f_0) in two cognitive load conditions. Physiological measures of the autonomic nervous system (ANS) were simultaneously recorded to determine if changes in autonomic arousal and cognitive load were associated with the ability to adapt to these errors. Results indicated cognitive load condition as a statistically significant predictor for f_0 responses and ANS arousal changed significantly between conditions. Measures of sensorimotor adaptation, however, showed no differences between conditions.

Keywords: Sensorimotor adaptation, voice, autonomic arousal

1. INTRODUCTION

It is well known that speech production and the autonomic nervous system are related [e.g., 1, 2, 3]; however much remains unknown about the effects of the ANS on the planning and updating processes of speech, known as speech motor control. The ANS controls and regulates physiological functions involving involuntary processes. The sympathetic and parasympathetic divisions are two important branches of peripheral activity that stem from the ANS. The sympathetic branch is responsible for "fight-or-flight" responses and the parasympathetic branch for "rest-and digest" responses [4, 5]. When the ANS is aroused, the sympathetic branch is activated, and the body's physiological functions change in response to the arousing stimuli [4, 5].

The electrodermal system is commonly targeted to measure ANS arousal because it is relatively easy to measure externally. Commercially available technology allows the estimation of electrodermal function non-invasively through wearable sensors. The electrodermal signal measures changes in conductivity due to increases in activity of sweat glands and is a direct measure of sympathetic activation [6]. Skin conductance level and response are the most common measures derived from this signal and increase when the ANS is aroused.

Modulating cognitive load is a common way to elicit ANS arousal. ANS arousal and cognition share anatomical substrates, including the anterior cingulate cortex, as evidenced through functional imaging studies, and thus are integrated mechanistically [7-9]. Cognitive load can be modulated using various tasks employing arithmetic [e.g., 10, 11], distractor interference [e.g., 12], and executive function [e.g., 13, 14] as examples.

Error-based motor integration, or sensorimotor adaptation, is a critical component of goal-directed behavior, including for speech production. From the extensive literature aimed at understanding the neural mechanisms involved in sensorimotor adaptation, three distinct brain regions have surfaced: the anterior cingulate cortex, the cerebellum, and the basal ganglia [see review: 15]. The anterior cingulate cortex and the cerebellum are also important for autonomic regulation and cognitive function; moreover, the basal ganglia play a role in cognitive function as well. The overlapping functions among these brain regions provide theoretical reasoning as to why cognition and ANS arousal may be related to sensorimotor adaptation of speech. Yet, contemporary models of speech motor control do not account for cognitive and autonomic perspectives, leaving a gap in the understanding of the neural mechanisms involved in speech production.

Although models of speech motor control differ, most agree that neural control systems exist to ensure speech is produced accurately and efficiently [see review: 16]. Some models suggest a feedforward control system is used to send (and update when necessary) stored motor programs to speech articulators [17, 18]. Sensorimotor adaptation has been thought to be driven by the feedforward control system and can be targeted experimentally by presenting persistent and expected errors to auditory feedback [e.g., 19]. Sensorimotor adaptation is

captured by observing how the system updates and integrates motor programs to correct for the errors.

Only one prior study has described the behavioral effects of cognitive load, specifically attentional demands, on sensorimotor adaptation [20]. The study applied persistent auditory errors of voice fundamental frequency (f_o) in differing attentional demand conditions, finding significant decreases in adaptation with greater attentional demands. Although this study provided critical evidence of the impact of cognitive load on sensorimotor adaptation of voice, there are some components that need to be addressed to further fill the gap in this area. First, because ANS activity was not recorded in this study, it is unclear if the reported effects were solely due to changes in cognitive processes, or if the effects were based on physiological changes as a result of the simultaneous ANS arousal from the cognitive demands. Second, this study only used about half of the participants in the final analyses because the other half did not show adaptation in the condition without attentional load. For these reasons, the results of this study must be replicated and extended.

The purpose of the current study was to determine the contributions of cognition and autonomic arousal in error-based learning to better understand how these neural mechanisms are integrated. This question was addressed by using sensorimotor adaptation paradigms in adults with typical speech. Persistent perturbations of f_o were compared between low and high cognitive load conditions. Electrodermal estimates of ANS activity were simultaneously collected. It was expected that the electrodermal measures would increase during the high cognitive load conditions, indicating a state of ANS arousal. We hypothesized that the updating processes of the voice would be negatively impacted by cognitive load and ANS arousal, such that adaptive responses to f_o perturbations would decrease with increased cognitive load and increased ANS arousal.

2. METHODS

2.1. Participants

Participants were native speakers of American English (9 cisgender males, 15 cisgender females, M_{age} = 22.76 years, SD_{age} = 2.66 years) with no history of speech, language, hearing, neurological, or autonomic condition and no prior training in singing. Participants were screened to avoid potential confounds affecting ANS function or cognitive ability. Participants were non-smokers and reported no history of drug or alcohol abuse. Participants reported no prior diagnosis of mood or psychological disorders, sleep apnea, low/high blood pressure, dermatological conditions, or hyperhidrosis. No participants were taking prescription medications known affect cognitive function. Participants abstained from the consumption of caffeine, alcohol, and heavy meals, and did not endure physical activity or stressful events three hours prior to the experiment. Participants passed a pure tone hearing screening at 25 DB HL at frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 and a color vision screening with the Ishihara Color Blindness Test [21]. All participants had typical cognitive function as assessed with the Montreal Cognitive Assessment.

2.2. Experimental design

Participants completed four speech tasks (two congruent and two incongruent) consisting of saying the color of the font of single-word stimuli ('red', 'yellow', 'black') and sustaining the midsection vowel of that color (/a/ in black or /ɜ/ in red and yellow) for approximately one second. These stimuli either matched the color of the font in all trials (congruent condition) or matched in some trials, but not in others (incongruent condition; Figure 1). During the incongruent condition, 33% of the trials were incongruent and the remaining 66% were congruent to avoid autonomic habituation. Presentation of the incongruent stimuli was semi-randomized such that incongruent trials were not sequential. In two of the four speech tasks, participants were exposed to sustained, predictable changes in auditory feedback ('shift-on') whereas in the other two tasks participants received unperturbed feedback ('control'). Thus, the four tasks were f_o shift-on under two cognitive load conditions and control under two cognitive load conditions. The order of the tasks was counterbalanced across participants.

During the 'shift-on' tasks, f_o was shifted

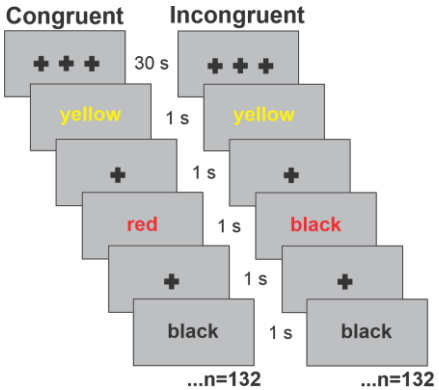


Figure 1: Single-word stimuli. Congruent: font color matches the word in all trials. Incongruent: font color matches the word in some trials, but not others.

downward over the course of four phases. The first

phase (baseline) consisted of 33 trials of unperturbed auditory feedback. The second phase, (ramp) consisted of 33 trials gradually shifted downward by 3.13 cents with each successive trial, reaching 100 cents below the participant's true f_0. The following 33 trials in the hold phase maintained this level of feedback. The last phase (after-effect) consisted of 33 trials in which no perturbation was applied.

2.3. Instrumentation and signal processing

2.3.1. Autonomic Data

Skin conductance is sensitive to the amount of activity in eccrine sweat glands [23] and is assessed through various measures of the electrodermal signal. The electrodermal signal was conditioned using a Biopac MP150 Data Acquisition System (Biopac Systems, Inc., Goleta, CA), collected with two disposable electrodes (EL507, Biopac Systems, Inc., Goleta, CA) and an amplifier (GSR 100C, Biopac Systems, Inc., Goleta, CA), and sampled at 14700 Hz using a National Instrument data acquisition card. Electrodes were attached to the palmar surface of the first and second distal phalange on the right hand. The tonic skin conductance level was amplified with a gain of 10 µS/V via a constant voltage (0.5 V). The phasic skin conductance response was derived from the tonic signal using a second-order Chebyshev high-pass filter with a cut-off frequency of 0.07 Hz.

2.3.2. Adaptation Data

Participants wore a SHURE MX153 omnidirectional microphone positioned 45 degrees from the midline and 7 cm from the lips and Sennheiser HD 280 Pro over-ear headphones. The microphone signal was pre-amplified using a RME Quadmic II and sampled using a RME Fireface UCX at a rate of 44,100 Hz and 32-bit resolution. Auditory feedback was calibrated to +5 dB relative to the microphone [24]. The Eventide Eclipse V4 Harmonizer was used to manipulate f_0.

2.4. Data analysis

2.4.1. Autonomic data analysis

Electrodermal activity was derived from the skin conductance signal and calculated using a custom MATLAB graphical user interface. *Skin conductance response (SCR) count* was used to capture the number of phasic responses. SCR count was calculated for each phase of the task and used in the regression analysis in determining its effects on f_0. Due to high variability in ANS data across participants, the change in SCR count between conditions was calculated for each participant and the average of these changes was used to assess group-level arousal.

2.4.2. Adaptation data analysis

Estimation of f_0 was performed offline using an autocorrelation method via Praat scripts [25]. Analysis of adaptive responses was performed offline using custom MATLAB scripts. The mean f_0 of every trial in each condition was calculated over a time window of 40-120 ms [26] after vowel onset, to avoid coarticulation effects and feedback-based adjustments [27-29], and converted to cents relative to the mean f_0 during the baseline phase. The control conditions were subtracted from the shift conditions to account for prosody and natural variability over time. Adaptive responses were calculated as the mean f_0 during the hold phase after the normalization processes. The average f_0 of each phase for all participants was used in regression analyses.

2.5. Statistical analysis

All statistical analyses were performed in Minitab 21.2. To account for group-level variability in ANS arousal, changes were determined using the average between-condition change in SCR count. One-sample t-tests were used to determine whether between-condition changes in ANS arousal were greater than zero. Overall adaptation was assessed in each condition using the average f_0 responses during the hold and after-effect phases. One sample t-tests were used to determine if f_0 responses during these phases were greater than zero, indicating adaptation. Differences in adaptation between conditions were assessed using two-sample t-tests. A regression analysis was performed to comprehensively assess the effect of cognitive load (congruent vs. incongruent), phase (baseline, ramp, hold, and after-effect), and ANS arousal (SCR count) on f_0 responses.

3. RESULTS

The one-sample t-tests for individual changes in SCR count between cognitive load conditions revealed a statistically significant change. Specifically, the average change in SCR count between conditions was 6.00 peaks (SD = 4.81 peaks); $t(23) = 5.72$, $p < .001$. Group results for f_0 responses are illustrated in Figure 2. One-sample t-tests showed significant f_0 responses greater than zero in the hold phase for both conditions: congruent $t(23) = 2.54$, $p = .010$, incongruent $t(23) = 3.79$, $p = .001$ and in the after-effect phase for the incongruent condition: $t(23) = 1.88$, $p = .037$. One-sample t-tests did not show significant f_0 responses greater than zero in the after-effect phase in the congruent condition: $t(23) = -0.13$,

p = .550 The two-sample t-tests for the changes in adaptation between conditions did not reveal a significant difference in either phase: hold t(46) = -0.81, p = .431, after-effect t(46) = -1.46, p = .076.

Predictor Variables	*df*	*F*	*p*
SCR Count	1	3.58	.063
Phase	3	7.66	<.001*
Condition	1	11.05	.001*
Condition × Phase	3	1.46	ns

Table 1: Results of the regression analyses. * Indicates statistically significant (p-value < .05) predictor variables.

The results of the regression model are described in Table 1. Condition was a significant predictor of f_o responses, F(1, 23) = 11.05, p = .001 with larger responses in the incongruent condition. Phase was found to be a significant predictor in the model, F(3, 23) = 7.66, p < .001 with post-hoc Tukey tests confirming a significantly larger mean f_o during the hold phase, p < .05. SCR count was not a significant predictor of f_o responses, F(1,23) = 3.58, p = .063, nor was the condition and phase interaction, F(3,23) = 1.46, p = .477.

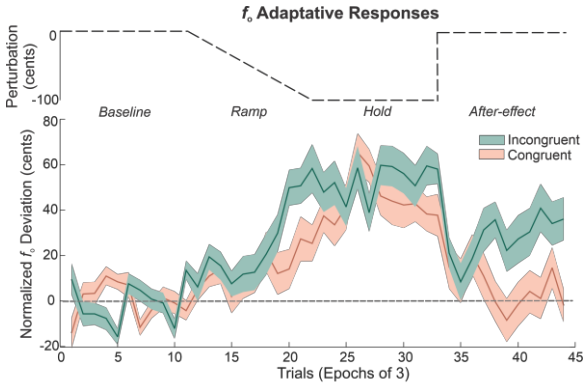


Figure 2: Group f_o responses to altered auditory feedback.

4. DISCUSSION

Our results indicate an increase in ANS arousal between conditions. This shows that the cognitive load task implemented in the current study successfully aroused the ANS. Although SCR count was not a significant predictor of changes in f_o, further investigation is warranted within other systems that rely on sympathetic and parasympathetic activity, such as the cardiovascular system. The f_o responses were significantly greater than zero in the hold phase for both conditions, indicating that participants adapted to the change in auditory feedback; further, the responses were not significantly different between conditions in either phase. Thus, the incongruent load did not diminish adaptation, as predicted, but perhaps, elicited longer-lasting adaptation, relative to the congruent condition, with f_o responses being significantly greater than zero in the after-effect phase. Interestingly, condition was a significant predictor of f_o responses in the regression analysis, likely due to the power of including responses in all phases, rather than the average of only the hold and after-effect phases as with the t-tests.

This result is interesting because it differs from a prior study finding [20]. These findings may be different from the findings of Scheerer, et al. [20] for a number of reasons. First, the current study used data from all participants, regardless of behavior to altered auditory feedback, whereas Scheerer, et al. [20] excluded 14 out of 30 participants for not compensating in the low cognitive load condition. A second reason as to why these results may differ from Scheerer, et al. [20], could be the differences in the cognitive tasks employed. Scheerer, et al. [20] used an attention task, whereas the current study used a response inhibition task. Although attention and response inhibition require skills encompassed under the same umbrella of executive function, the degree of task complexity changes the neural resources allocated to that task. Thus, the ability to adapt vocal motor commands may depend on task complexity. Lastly, these results may differ from Scheerer, et al. [20] due to timing differences: the two conditions in Scheerer, et al. [20] were separated by one week, whereas the conditions in the current study were completed in the same visit. Because the reliability of f_o responses to altered auditory feedback has not been validated, it is unclear if this relatively large break between sessions impacts the study effects.

Surprisingly, from a visual inspection of the data (Figure 2), there seems to be a small difference in adaptation between conditions towards the end of the after-effect phase. The lack of an interaction effect between condition and phase could be explained by the high variability in participant data (congruent after-effect SD = 74, incongruent after-effect SD = 79). Adaptation responses are often variable and thus adds difficulty in capturing statistical changes between conditions. Group variability is a common factor experienced by many research groups working with altered auditory feedback paradigms. One purpose of this study was to determine if other neural systems are impacting the results of these paradigms, with the long-term goal of understanding if these peripheral processes contribute to the variability of responses in both typical and clinical populations. Although we did not find significant changes in adaptation between conditions, more research is necessary to interpret responses with a more wholistic view and to inform current models of speech motor control.

5. REFERENCES

[1] S. Het, N. Rohleder, D. Schoofs, C. Kirschbaum, and O. T. Wolf, "Neuroendocrine and psychometric evaluation of a placebo version of the 'Trier Social Stress Test'," *Psychoneuroendocrinology,* vol. 34, no. 7, pp. 1075-86, Aug 2009.

[2] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'--a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology,* vol. 28, no. 1-2, pp. 76-81, 1993.

[3] T. Heponiemi, N. Ravaja, M. Elovainio, P. Naatanen, and L. Keltikangas-Jarvinen, "Experiencing positive affect and negative affect during stress: relationships to cardiac reactivity and to facial expressions," *Scand J Psychol,* vol. 47, no. 5, pp. 327-37, Oct 2006.

[4] M. Bear, B. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain*. Jones & Bartlett Learning, 2020.

[5] R. W. Hamill, R. E. Shapiro, and M. A. Vizzard, "Peripheral autonomic nervous system," in *Primer on the autonomic nervous system*: Elsevier, 2012, pp. 17-26.

[6] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[7] H. D. Critchley, "Neural mechanisms of autonomic, affective, and cognitive integration," *Journal of comparative neurology,* vol. 493, no. 1, pp. 154-166, 2005.

[8] H. D. Critchley *et al.*, "Human cingulate cortex and autonomic control: converging neuroimaging and clinical evidence," *Brain,* vol. 126, no. 10, pp. 2139-2152, 2003.

[9] H. D. Critchley, J. Tang, D. Glaser, B. Butterworth, and R. J. Dolan, "Anterior cingulate activity during error and autonomic response," *Neuroimage,* vol. 27, no. 4, pp. 885-895, 2005.

[10] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th australian computer-human interaction conference*, 2012, pp. 420-423.

[11] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Transactions on information technology in biomedicine,* vol. 14, no. 2, pp. 410-417, 2009.

[12] J. L. Harbluk, Y. I. Noy, and M. Eizenman, "The impact of cognitive distraction on driver visual behaviour and vehicle control," 2002.

[13] S. Janssen *et al.*, "Validation of the Auditory Stroop Task to increase cognitive load in walking tasks in healthy elderly and persons with Parkinson's disease," *PloS one,* vol. 14, no. 8, 2019.

[14] J. R. Stroop, "Studies of interference in serial verbal reactions," *Journal of experimental psychology,* vol. 18, no. 6, p. 643, 1935.

[15] R. D. Seidler, Y. Kwak, B. W. Fling, and J. A. Bernard, "Neurocognitive mechanisms of error-based motor learning," in *Progress in motor control*: Springer, 2013, pp. 39-60.

[16] B. Parrell, A. C. Lammert, G. Ciccarelli, and T. F. Quatieri, "Current models of speech motor control: A control-theoretic overview of architectures and properties," *J Acoust Soc Am,* vol. 145, no. 3, p. 1456, Mar 2019.

[17] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Lang Cogn Process,* vol. 26, no. 7, pp. 952-981, Jan 1 2011.

[18] B. J. Kroger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," (in English), *Speech Communication,* vol. 51, no. 9, pp. 793-809, Sep 2009.

[19] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science,* vol. 279, no. 5354, pp. 1213-1216, 1998.

[20] N. E. Scheerer, A. K. Tumber, and J. A. Jones, "Attentional demands modulate sensorimotor learning induced by persistent exposure to changes in auditory feedback," *Journal of neurophysiology,* vol. 115, no. 2, pp. 826-832, 2016.

[21] S. Ishihara, *Ishihara's tests for colour-deficiency*. Kanehara, 1996.

[22] M. Berg, M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, and T. Berger, "The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors," (in English), *Journal of Voice,* vol. 31, no. 2, Mar 2017.

[23] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," 2017.

[24] H. R. Weerathunge, D. Abur, N. M. Enos, K. M. Brown, and C. E. Stepp, "Auditory-motor perturbations of voice fundamental frequency: feedback delay and amplification," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 9, pp. 2846-2860, 2020.

[25] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.,* vol. 5, no. 9, pp. 341-345, 2001.

[26] R. A. Lester-Smith *et al.*, "The relation of articulatory and vocal auditory–motor control in typical speakers," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 11, pp. 3628-3642, 2020.

[27] T. A. Burnett and C. R. Larson, "Early pitch-shift response is active in both steady and dynamic voice pitch control," *The Journal of the Acoustical Society of America,* vol. 112, no. 3, pp. 1058-1063, 2002.

[28] C. R. Larson, T. A. Burnett, J. J. Bauer, S. Kiran, and T. C. Hain, "Comparison of voice F 0 responses to pitch-shift onset and offset conditions," *The Journal of the Acoustical Society of America,* vol. 110, no. 6, pp. 2845-2848, 2001.

[29] J. A. Tourville, K. J. Reilly, and F. H. Guenther, "Neural mechanisms underlying auditory feedback control of speech," *Neuroimage,* vol. 39, no. 3, pp. 1429-1443, 2008.

STOP VOICING AND DEVOICING AS ARTICULATORY TASKS: A CROSS-LINGUISTIC RT-MRI STUDY

Yubin Zhang and Louis Goldstein
Department of Linguistics, University of Southern California

ABSTRACT

The current study tests the hypothesis that voicing and devoicing of stop consonants are phonatory tasks that are accomplished by a variable synergy of laryngeal and supraglottal articulators. We investigate the cross-linguistic differences in that synergy using RT-MRI of connected speech. French voiced stops were observed to have larger oral aperture, shorter oral constriction duration, higher velum position, more advanced tongue root, and larger post-constriction cavity than voiceless stops. In English, fewer articulatory dimensions are involved. Voiced stops have more advanced tongue root, lower larynx, and larger post-constriction cavity than voiceless stops. Moreover, by comparing nasals with voiced and voiceless stops, we found that the supraglottal adjustments are primarily employed for the devoicing rather than the voicing task. The results are consistent with the voicing/devoicing goal hypothesis, but languages differ in the articulatory dimensions used for this goal.

Keywords: stop, voicing, articulation, French, English.

1. INTRODUCTION

The stop voicing categories have been described as contrasting in voicing and aspiration [1], [2]. Languages with a two-way voicing contrast, are classified into two types of systems— the 'true voicing' and 'aspirating' languages [1], [2]. The 'true voicing' languages, such as French, Spanish, Russian and Portuguese, have rigorous closure voicing for voiced stops whereas their voiceless stops lack closure voicing and post-closure aspiration. For 'aspirating' languages, such as English and Mandarin, the contrast is considered as primarily in aspiration. These languages all have a category called voiceless aspirated stops with no closure voicing but post-closure aspiration. The other category typically does not have closure voicing and can be described as phonetically voiceless unaspirated stops. Note that while English is described as an 'aspirating' language, the latter category is not the canonical voiceless unaspirated stops. In the literature, it is also called phonologically voiced stops or lax voiceless stops. Closure voicing is generally lacking in domain-initial positions and intervocalic closure voicing is not always robust [1], [3].

The articulatory basis for voicing distinction has been proposed to lie in oral-laryngeal timing in models like articulatory phonology [4], [5] (see Figure 1). For both voiceless unaspirated and aspirated stops, the glottal opening gesture is timed in-phase with oral closing gesture, resulting in the silent closure interval [6]. For voiceless unaspirated stops with short-lag voicing onset time (VOT), the glottal closing gesture occurs before the oral release gesture, whereas for voiceless aspirated stops with long-lag VOT, the onset of the glottal closing gesture occurs near oral release [6], [7]. For phonologically voiced stops, only the oral constriction gesture is supposed to be activated. The presence of vocal fold vibration in the closure interval is not assumed to be specifically controlled. It is regarded as the default glottal state once phonation is initiated and when an active glottal opening-and-closing gesture is absent.

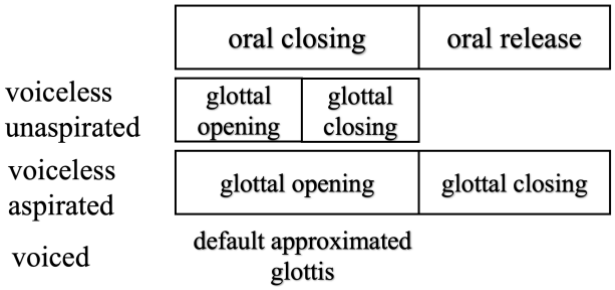


Figure 1: Illustration of the articulatory basis for voicing.

A potential problem with this proposal is that closure voicing does not always occur automatically with adducted vocal folds. Closure voicing is subject to the aerodynamic voicing constraint [8], [9]. With appropriately approximated and tensed vocal folds, a sufficient transglottal pressure difference, i.e., sufficiently larger subglottal pressure than intraoral pressure, is also needed to sustain continuous glottal airflow. However, the vocal tract constriction in stops quickly decreases the transglottal pressure. Without additional adjustments, voicing would cease quickly. Active glottal and supra-glottal articulations, such as vocal folds slackening [10], post-constriction cavity enlargement like larynx lowering and tongue root advancement [8], [11], oral/nasal leakage [12] and constriction duration reduction, [13] have been shown to be possible mechanisms for sustaining voicing [8], [12]. Terminating or inhibiting voicing can also require mechanisms in addition to the glottal opening gesture, like tightly adducting the vocal folds and stretching the vocal folds [7], [14], [15].

To account for these observations, some researchers suggest more global task goals, such as aerodynamic goals [16], [17]. [16] postulate a transglottal pressure task during voiced stop production. For devoicing in voiceless stops, they propose that the glottal opening gesture is the primary task involved. Nevertheless, the authors agree that other gestures like achieving a high F0 task goal by stretching and stiffening the vocal folds may also contribute to devoicing. Thus, [16] further discuss the possibility of postulating a 'voicing/devoicing' task that engages a flexible synergy comprising articulator components like glottal width, glottal tension, total lung force, and post-constriction cavity volume.

The current rt-MRI articulatory study investigates the cross-linguistic differences in the synergies comprising potential phonatory 'voicing/devoicing' tasks by examining the supra-glottal articulations during production of contrasting voiced and voiceless stops in French and English. For 'true voicing' languages like French, multiple articulator components are expected to be involved in achieving the voicing/devoicing tasks. However, for languages like English without rigorous closure voicing, fewer and less consistent articulatory dimensions might be employed in the voicing/devoicing tasks [12].

2. METHODS

The French articulatory data consist of mid-sagittal real-time MRI videos acquired from 5 male and 5 female French speakers aged 29 ± 8 years [18]. The speakers read aloud the speech materials, consisting of 77 manually constructed sentences with an almost-exhaustive coverage of the French speech sounds in various phonetic contexts (see [18] for more details). The MRI images were collected on a Siemens Prisma 3T scanner with a frame rate of 50 frames/s, image size of 136 × 136 and resolution of 1.6 mm × 1.6 mm. The audio was recorded in the MRI scanner with a sampling rate of 16kHz and then noise-cancelled using the algorithm described in [19]. The audio was automatically force-aligned with the transcriptions of words and phonemes using the French speech recognition system Astali. The English articulatory data were taken from the recordings of 7 participants in the publicly available USC-TIMIT real-time MRI database [20]. The participants were native speakers of American English (2 male, and 5 female, aged 30 ± 9 years). They read 460 sentences from MOCHA-TIMIT, a database designed to elicit all the English phonemes in various phonological and prosodic contexts. MRI data were acquired on a Signa Excite HD 1.5T scanner. The image resolution is 2.9 × 2.9 mm. The image size is 68 × 68. The reconstructed frame rate is 23.18 frames/s. The audio was simultaneously recorded at a sampling rate of 20kHz and then noise-cancelled using the method described in [22]. The audio has been force-aligned with text annotations of phonemes, words and sentences using SailAlign [21].

The articulatory analysis was based on stop and nasal tokens in the #_V, V_V, V_# contexts with oral vowels (# indicates utterance boundaries). Nasals were included as a neutral condition against which to evaluate articulator contributions to voicing and devoicing tasks. Velar nasals were not included because French does not have native velar nasals and most English velar nasals are from the 'ing' suffix. We analyzed a total of 1479 and 3910 stop tokens for French and English respectively.

The articulator contours in the MRI videos were tracked using the region-based segmentation method described in [22]. We first selected a reference frame during speaking without close approximation of articulators, and manually constructed an initial template locating the contours of the articulator segments such as tongue, hard palate, velum, lips, jaw, pharyngeal wall, epiglottis, etc. (see Figure 2 top left panel). Then, the initial template was registered to each frame of the MRI video and the air-tissue boundaries of the articulators were extracted using a hierarchical gradient descent procedure.

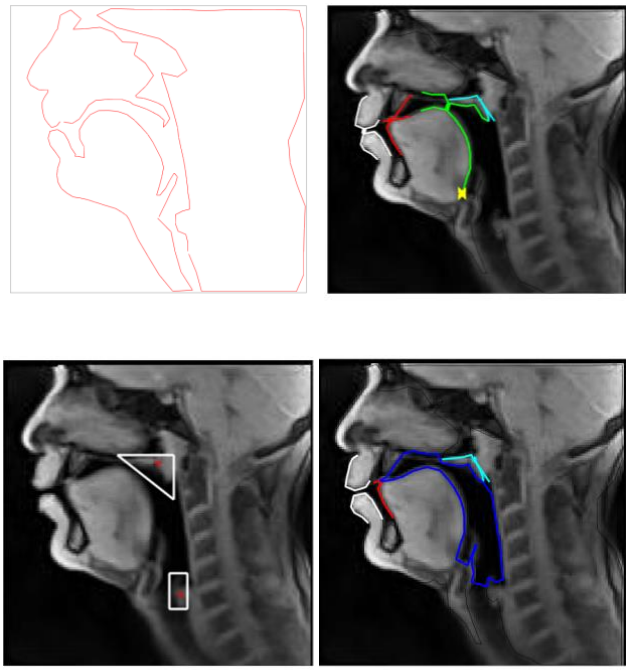


Figure 2: The region-based segmentation and centroid tracking analyses. Top left: the initial template; Top right: labial (white), alveolar (red), velar (green) and velic (cyan) search regions; The displayed frame is one where maximal alveolar constriction is achieved. The straight line denotes the minimal distance between two articulator contours. The yellow cross denotes the tongue root position; Bottom left:

the regions of interest (ROIs, triangle: velum; rectangle: larynx) and tracked centroids; Bottom right: post-constriction cavity mid-sagittal plane (blue).

Four search regions were identified to extract constriction aperture time series, i.e., the minimal distance between two contour lines (Figure 2 top right)— (1) lip constriction: between upper and lower lips, (2) alveolar constriction: between the tongue and alveolar ridge, (3) velar constriction: between the tongue and palate and (4) velic constriction: between the upper boundary of the velum and upper pharyngeal wall. The algorithm generates one single contour for the hard palate and tongue respectively. To separate the alveolar and velar search regions, we manually marked the boundary between the alveolar ridge and hard palate for each speaker by inspecting the reference image. The x-coordinate of the marked boundary point was used to divide the two regions. The generated velum contour was also divided into two parts: the upper and lower contours. The lower velum contour was used for the velar search region whereas the upper velum contour for the velic region. The upper pharyngeal wall was defined as the pharyngeal contour above the lowest velum point.

Horizontal tongue root movement was examined using the x-coordinate of tongue root position over time (Smaller x means more advanced). The tongue root point was defined as the point with the smallest y-coordinate on the tongue contour (the yellow cross in Figure 2 top right panel). Vertical larynx and velum movement trajectories were generated using a centroid-tracking method [23]. The algorithm automatically tracks the intensity-weighted centroid of an object in a user-defined region of interest (ROI). For each speaker, we manually specified a rectangular larynx ROI and a triangular velum ROI, which roughly cover the movement range for the arytenoid cartilage and velum respectively (see Figure 2, bottom left). The y-coordinates of the larynx and velum centroids over time were taken as the time series of vertical larynx and velum movements (larger y means higher position). The time series of mid-sagittal post-constriction cavity area was computed by combining all the articulator contours ranging from the constriction location (the straight line denoting minimal distance between oral articulators) to the middle of arytenoid cartilage (y-coordinate of the larynx centroid) at each time point (see the blue contour in Figure 2, bottom right). All the time series were smoothed with a span of 15 data points using the `rloess` function in `Matlab`.

The maximum constriction and gesture duration of oral gestures were computed using the `findgest` algorithm in `MVIEW` [24]. We searched for a velocity minimum closed to the mid-point of the force-aligned acoustic consonant interval. The aperture size at the velocity minimum was taken as the maximum constriction (MAXC). Then, two velocity peaks (PVEL1 and PVEL2) preceding and following MAXC were identified. Gestural duration was calculated as the difference between gestural onset and offset, which are defined using the 20% threshold criterion (onset: 20% of the range between the velocity minimum preceding PVEL1 and PVEL1; offset: 20% of the range between PVEL2 and following velocity minimum). Labelling errors were corrected manually. Non-oral-constriction articulatory parameters, i.e., velic aperture, velum height, horizontal tongue root position, larynx height, and post-constriction cavity area, were measured at the time point of MAXC.

We fit linear mixed-effects models with factors `voicing` and `place of articulation (POA)` for each articulatory dimension using `lmer()` in the `R` package `lme4` [25]. Their interaction was included if justified by the likelihood ratio test. Two separate series of models were fit—*v=vl* and *v=vl=nas* models. In *v=vl* models, `voicing` has two levels—*voiced* and *voiceless* with *voiceless* as the reference level. In *v=vl=nas* models, *nasal* was included as a reference level for the factor `voicing`. Only bilabial and alveolar data were analyzed because velar nasals were not included. For velum-related and cavity size measures, the nasal condition is no longer a neutral condition. Comparison with the nasal condition is not meaningful for these measures. Thus, only the *v=vl* model is reported in such cases.

	French	**English**
oral aperture	v>vl: **; vl<nas: **; v-nas: n.s.	n.s.
oral constriction duration	v<vl: *; vl-nas: n.s.; v-nas: n.s.	v-vl: n.s.; vl>nas: *; v-nas: n.s.
velic aperture	v-vl: n.s.	v-vl: n.s.
velum height	v>vl: *;	n.s.
horizontal tongue root position	v<vl: *; vl>nas: *; v-nas: n.s.	v<vl: ***; vl>nas: *; v-nas: n.s.
larynx height	v-vl: n.s.; v>nas: ***; vl>nas: *	v<vl: ***; v>nas: *** ; vl>nas: ***
mid-sagittal post-constriction cavity area	v>vl: ***	v>vl: ***

Table 1: The results for the `voicing` effects. Notations: v: voiced; vl: voiceless; nas: nasal; -: non-significant comparison; n.s.: non-significant, * $p \leqslant 0.05$, ** $p < 0.01$, *** $p < 0.001$. `POA` effects are not reported here.

3. RESULTS

The results for the `voicing` effects are shown in Table 1. The *v=vl* models reveal that French voiced stops have larger oral constriction aperture (v>vl), shorter oral constriction duration (v<vl), higher velum (v>vl), more advanced tongue root (v<vl), larger mid-sagittal post-constriction cavity area (v>vl) than voiceless stops. In English, voiced stops only have more advanced tongue root (v<vl), lower larynx (v<vl), and larger mid-sagittal post-constriction cavity area (v>vl) than voiceless stops. In the *v=vl=nas* models, when a significant comparison involving nasals was found, it is generally the nasal-voiceless comparison (e.g., vl<nas for oral aperture in French), except for larynx height (v, vl>nas). This result suggests that `voicing` effect is largely driven by the difference between nasals and voiceless stops.

4. DISCUSSION

The current results show that both languages engage multiple supra-glottal articulatory components in the voicing contrast, consistent with the hypothesis that voicing/devoicing can be more global linguistic motor tasks achieved by a synergy of laryngeal and supraglottal articulatory components [16], [17]. However, the exact task variable for voicing/devoicing remains to be investigated. [16] suggests voicing amplitude whereas [26] focuses more on durational measures like VOT. Moreover, a further way to gain insight on the synergy might be to investigate the compensatory or cooperative behaviors among the articulatory dimensions by more controlled experiments.

We also found that the difference in supraglottal articulations between voiced and voiceless stops is largely driven by the difference between nasals and voiceless stops. Previous research generally assumes that the supra-glottal articulations in stop voicing contrast are employed for sustaining voicing rather than devoicing. Terms like cavity expansion and oral/nasal leakage are frequently used [9], [13], [17]. However, our results suggest that in both languages, devoicing mechanisms, like aperture size reduction and tongue root retraction, seem to be involved more than voicing-sustaining mechanisms, at least for the dimensions where nasal data is meaningful. Most previous studies like [13], [26] did not include nasals as a neutral condition and could not evaluate the articulator contributions to voicing and devoicing. The English data from [9] included nasals for dimensions like tongue root. They found that for tongue root position, voiceless stops and nasals show more similarity than voiced stops and nasals, indicating more involvement of voicing-sustaining mechanisms. However, since they only recruited one speaker and provided no quantitative analysis, the discrepancy remains to be resolved by future studies. Note that nasals do not seem to be a valid neutral condition for larynx height. In both languages, stops have a higher larynx than nasals. This result is consistent with [9], but the reasons remain murky.

For language-specific differences, consistent with our prediction, and previous findings on cross-linguistic voicing/devoicing control [12], French speakers employ more supraglottal articulatory adjustments than English speakers. French speakers engage articulations like adjusting oral aperture size, oral constriction duration, velum height, horizontal tongue root position, and the post-constriction cavity size. English speakers exhibit differences in horizontal tongue root position, larynx height, and post-constriction cavity size, consistent with [8], [27]. Both languages use cavity expansion/shrinkage mechanisms, but the synergies show language-specific differences, i.e., expansion/shrinkage in the tongue root and velar regions for French but in the tongue root and larynx regions for English. Moreover, for aperture size measures, French speakers alter the oral aperture size, but neither French nor English speakers alter the velic aperture size. One explanation is that the effect size of velic leakage is extremely small [12]. The spatial resolution of the rt-MRI images does not allow us to reliably identify small nasal aperture adjustments. It is also likely that unlike the original proposal in articulatory phonology, other dimensions for velum control like velum height are more crucial than the velic constriction gesture.

One limitation of the current study is that voicing-related acoustic measures cannot be reliably analyzed due to the excessive noise induced by MRI recordings, especially for the English dataset. Without the output voicing measure, it is difficult to assess how these articulatory dimensions contribute to the voicing/devoicing tasks in different languages. Previous studies suggest that the appearance of closure voicing is quite variable in English, but relatively robust in French [1], [3]. Thus, one may expect that voicing-sustaining mechanisms rather than devoicing mechanisms differ between these two languages. However, we found no evidence for more voicing-sustaining mechanisms in French than English. Language-specific differences seem to lie primarily in the devoicing aspect. It is unclear how these differences are related to the closure voicing differences observed across languages. One possibility is that other unexamined articulatory dimensions, such as glottal states, tongue body height, timing among gestures, and the respiratory gestures that modulate subglottal/transglottal pressure, play a role.

5. REFERENCES

[1] P. A. Keating, "Phonetic and phonological representation of stop consonant voicing," *Language (Baltim).*, vol. 60, no. 2, p. 286, 1984, doi: 10.2307/413642.

[2] T. Cho, D. H. Whalen, and G. Docherty, "Voice onset time and beyond: Exploring laryngeal contrast in 19 languages," *J. Phon.*, vol. 72, pp. 52–65, Jan. 2019, doi: 10.1016/j.wocn.2018.11.002.

[3] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *J. Phon.*, vol. 54, pp. 35–50, Jan. 2016, doi: 10.1016/j.wocn.2015.09.003.

[4] L. Goldstein and C. P. Browman, "Representation of voicing contrasts using articulatory gestures," *J. Phon.*, vol. 14, no. 2, pp. 339–342, 1986, doi: 10.1016/s0095-4470(19)30662-x.

[5] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonol. Yearb.*, vol. 3, pp. 219–252, 1986, doi: 10.1017/s0952675700000658.

[6] A. Löfqvist, "Laryngeal mechanisms and interarticulator timing in voiceless consonant production," in *Producing Speech: Contemporary Issues for Katherine Safford Harris*, 1995, pp. 99–116.

[7] R. P. Dixit, "Glottal gestures in Hindi plosives," *J. Phon.*, vol. 17, no. 3, pp. 213–237, Jul. 1989, doi: 10.1016/s0095-4470(19)30431-0.

[8] J. R. Westbury, "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1322–1336, 1983, doi: 10.1121/1.389236.

[9] M. Rothenberg, "The breath-stream dynamics of simple-released-plosive production," 1966. doi: 10.1016/0024-3841(71)90101-x.

[10] M. Halle and K. N. Stevens, "A note on laryngeal features," in *MIT Research Laboratory of Electronics Quaterly Progress Report*, 1971, pp. 198–213. doi: 10.1515/9783110871258.45.

[11] F. Bell−Berti, "Control of pharyngeal cavity size for English voiced and voiceless stops," *J. Acoust. Soc. Am.*, vol. 57, no. 2, pp. 456–461, 1975, doi: 10.1121/1.380468.

[12] M. J. Solé, "Articulatory adjustments in initial voiced stops in Spanish, French and English," *J. Phon.*, vol. 66, pp. 217–241, Jan. 2018, doi: 10.1016/j.wocn.2017.10.002.

[13] B. Parrell, "Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials," *Lab. Phonol.*, vol. 2, no. 2, Oct. 2011, doi: 10.1515/labphon.2011.016.

[14] R. Kagaya, "A fiberscopic and acoustic study of the Korean stops, affricates and fricatives," *J. Phon.*, vol. 2, no. 2, pp. 161–180, Mar. 1974, doi: 10.1016/s0095-4470(19)31191-x.

[15] A. Löqvist, T. Baer, N. S. McGarr, and R. S. Story, "The cricothyroid muscle in voicing control," *J. Acoust. Soc. Am.*, vol. 85, no. 3, pp. 1314–1321, 1989, doi: 10.1121/1.397462.

[16] R. S. McGowan and E. L. Saltzman, "Incorporating aerodynamic and laryngeal components into task dynamics," *J. Phon.*, vol. 23, no. 1–2, pp. 255–269, 1995, doi: 10.1016/S0095-4470(95)80047-6.

[17] I. G. Mattingly, "The global character of phonetic gestures," *J. Phon.*, vol. 18, no. 3, pp. 445–452, 1990, doi: 10.1016/s0095-4470(19)30372-9.

[18] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P. A. Vuissoz, "Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers," *Sci. Data*, vol. 8, no. 1, pp. 1–9, Oct. 2021, doi: 10.1038/s41597-021-01041-3.

[19] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012, doi: 10.1109/TASL.2011.2172425.

[20] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1307–1311, 2014, doi: 10.1121/1.4890284.

[21] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign : Robust long speech-text alignment," *Proc. Work. New Tools Methods Very Large Scale Res. Phonetic Sci.*, pp. 28–31, 2011.

[22] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 323–338, Mar. 2009, doi: 10.1109/TMI.2008.928920.

[23] M. Oh and Y. Lee, "ACT: An Automatic Centroid Tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. EL290–EL296, 2018, doi: 10.1121/1.5057367.

[24] M. Tiede, "MVIEW: Multi-channel visualization application for displaying dynamic sensor movements." 2010.

[25] D. Bates, M. Maechler, B. Bolker, and S. Walker, "The lme4 Package. R package version 1.1-25." 2020.

[26] J. R. Westbury and P. A. Keating, "On the naturalness of stop consonant voicing," *J. Linguist.*, vol. 22, no. 1, pp. 145–166, 1986, doi: 10.1017/S0022226700010598.

[27] S. Ahn, "The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese," *J. Phon.*, vol. 71, pp. 451–467, Nov. 2018, doi: 10.1016/j.wocn.2018.10.003.

ARTICULATORY CHARACTERISTICS OF THE JAPANESE /r/: A REAL-TIME MRI STUDY

Kikuo Maekawa
National Institute for Japanese Language and Linguistics
kikuo@ninjal.ac.jp

ABSTRACT

Japanese /r/ is described as a voiced post-alveolar flap in the Handbook of the IPA. This description raises two questions. The validities of the exclusion of the alveolar articulation, and the exclusion of the "tap" articulation as distinguished from the "flap". To clarify these issues, articulation of the Japanese /r/ was analysed using the Real-time MRI Articulatory Movement Database. Regarding the manner of articulation, no "flap" articulation was observed in all 167 samples of the /r/ covering both word-initial and word-medial positions. Regarding the place of articulation, it turned out that the place of articulation of the Japanese /r/ did not cover the post-alveolar region. Moreover, leave-one-out cross-validation revealed that the variation of the place of articulation of the Japanese /r/ could be predicted fairly well by knowing the categories of the adjacent vowels.

Keywords: Japanese /r/, Real-time MRI Articulatory Movement Database, place and manner of articulation

1. INTRODUCTION

In the Handbook of the IPA [1], the Japanese /r/ is described as a "voiced post-alveolar flap" and represented by the phonetic symbol [ɽ] [2]. This description by Okada is problematic at least in two respects ignoring the curious choice of the phonetic symbol. First, concerning the place of articulation (or PoA), its exclusion of the alveolar / dental articulation. Second, concerning the manner of articulation (or MoA), its exclusion of the "tap" articulation as distinguished from the "flap" articulation. Although Okada's description doesn't have wide support from the specialists of Japanese, it is accepted by some non-specialists on the authority of the IPA. This paper examines the PoA and MoA of the Japanese /r/ objectively by using the data of real-time MRI movies.

2. DATA AND METHOD

The data analysed below are all taken from the Real-time MRI Articulatory Movement Database, or rtMRIDB (https://rtmridb.ninjal.ac.jp). This database is a collection of about 26,000 Japanese utterances spoken by 22 native Japanese speakers; it covers materials like mora-unigram, mora-bigram, words containing mora-phonemes, reading aloud of short passages, and so forth. Articulatory movements in the vocal tract are observed in the mid-sagittal plane with a frame rate of about 14 or 27 fps (frames per second). Each frame consists of 256 × 256 pixels (or voxels) with a pixel resolution of 1 mm and a slice thickness of 10 mm. The recording was done using the 3 T MRI scanner (Siemens MAGNETOM Prisma) at the Brain Activity Imaging Center of ATR Promotion Inc., in Kyoto, Japan, in the years 2017–2021.

167 samples of /r/ in the word- (or mora-) initial and word-medial positions were extracted from the rtMRIDB. The initial samples include /ra/, /ri/, /ru/, /re/, /ro/, /rja/, /rju/, /rjo/, and /rje/, while the medial samples include /ara/, /ari/, /aru/, /are/, /aro/, /ira/, /iri/, /iru/, /ire/, /iro/, /ora/, /ori/, /oru/, /ore/, and /oro/. These samples were uttered by seven male speakers of Standard Japanese and recorded in 27 fps. Note that the symbol /rj/ stands for the palatalized version of the /r/.

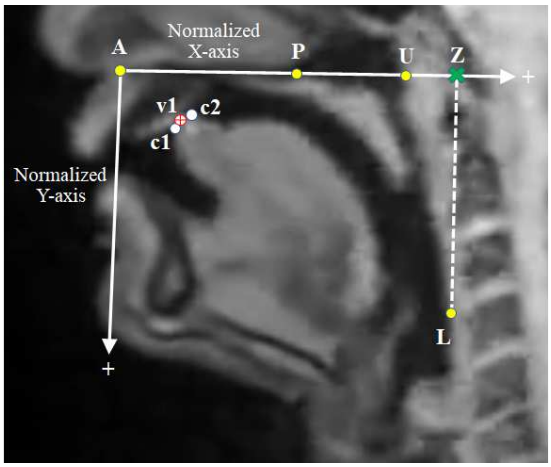


Figure 1: Measurement points and normalized X- and Y-axes.

Figure 1 shows the measurement points and normalized XY-axes used in this study. Points c1 and c2 are the edges (beginning and end) of the vocal tract closure for /r/ (or the constriction of /s/, see 3.1). Point v1 (with a red cross) is the tip of the tongue. A and P stand respectively for the anterior and posterior nasal spines. U is the point where the line connecting A and P reaches the pharyngeal wall. L is the point on the pharyngeal wall corresponding to the boundary between the third and fourth cervical vertebrae. These points were measured for all frames corresponding to the timing of the vocal tract closure by the /r/. In

addition, c1, c2, and v1 are also measured for the frame that precedes the /r/ (the "preceding frame"). Note that the relative location among the c1, c2, and v1 differs depending on the MoA of the /r/ (See 3.2).

The XY-values of the c1, c2, and v1 were normalized as follows. The origin of the normalized coordinate is on the A. The normalized x-axis is the line connecting the A, P, and U. Y-axis is perpendicular to the X-axis. The unit of measurement is the distance between the A and U for the X-axis, and the distance between the L and Z for the Y-axis. Z is the point where the line starting from point L (broken line in Figure 1) intersects perpendicularly with the normalized X-axis. This method is nearly identical to that proposed in [3]; see the literature for the effectiveness of the proposed method.

3. ANALYSIS

3.1. Place of articulation (PoA)

To evaluate the PoA of the /r/, the distributions of the c1 and c2 were compared to those of the word-initial /s/. Japanese /s/ is phonetically realized as a voiceless alveolar fricative [s] before the vowels /a, u, e, o/, but is realized as an alveolo-palatal fricative [ɕ] (or post-alveolar [ʃ]) before the /i/. Phonologically palatalized /s/, represented as /sj/ and appearing before /a/, /o/, and /u/ is also realized as [ɕ] (or [ʃ]). Figure 2 compares the distributions of the c1 and c2 of the plain and palatalized /s/ (the latter including /si/) and /r/. The distribution range of the plain /r/ turns out to be almost identical to that of plain /s/ and more frontal than palatalized /sj/. One more finding about the PoA is that the ranges of distribution are nearly the same in the plain /r/ and palatalized /rj/. This issue will be discussed in section 5.

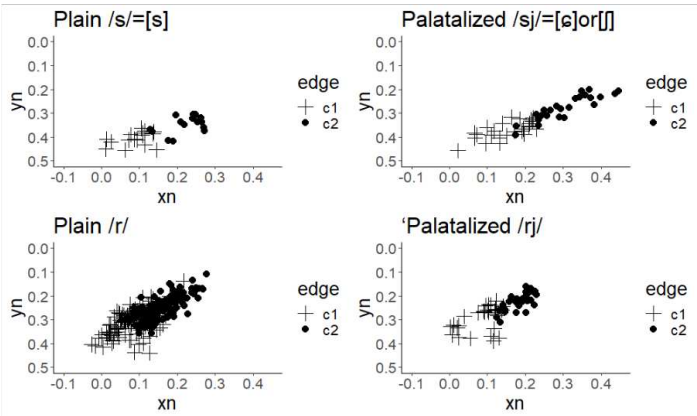


Figure 2: Comparison of the constriction/closure locations between /s/ and /r/. Separate panels are prepared for plain and palatalized /s/ and /r/. "xn" and "yn" stand respectively for "normalized x" and "normalized y" values.

3.2. Manner of articulation (MoA)

The analysis of the MoA is concerned with Okada's distinction between the "tap" and "flap" articulations, which are not distinguished in the IPA chart of pulmonic consonants. In this paper, following the proposals by [4] and [5], the "flap" is defined as the articulation where the vocal tract closure occurs between the palate and the "under-side of the tongue". On the other hand, in the "tap" articulation, the upper surface of the tongue is used for the closure.

Figure 3 compares a typical "flap" /r/ as uttered by a phonetician (panel A) and two variants of "tap" articulation (B and C). Points c1, c2, and v1 are shown in the same manner as in Figure 1. Both the c1 and c2 are to the left of the v1 in the "flap" articulation. In the "tap" articulation, c2 locates always to the right of the v1, but the relative location of the v1 and c1 is variable. Sometimes, c1 is to the right of the v1 (panel B), but there is also the case where the c1 locates to the left of the v1 (panel C).

Figure 4 consists of two histograms showing the distributions of the differences of the X-values between the v1 and c1 (abb. v1-c1), and those between the v1 and c2 (v1-c2), across the initial and medial /r/ positions. The v1-c2 values are positive in

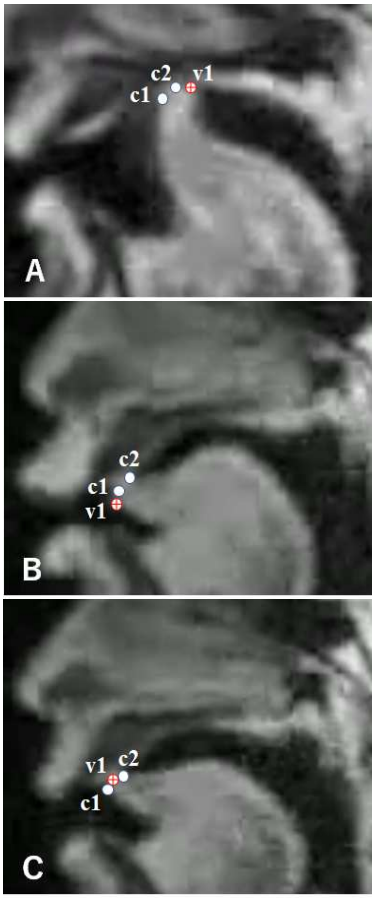


Figure 3: A typical "flap" (A) and two variants of "tap" articulation (B, C).

all samples, indicating that c2 is always to the right of the v1. And the v1-c1 values distribute on both sides of the zero. These properties are the ones expected only in the "tap" articulation like panels B and C of Figure 3.

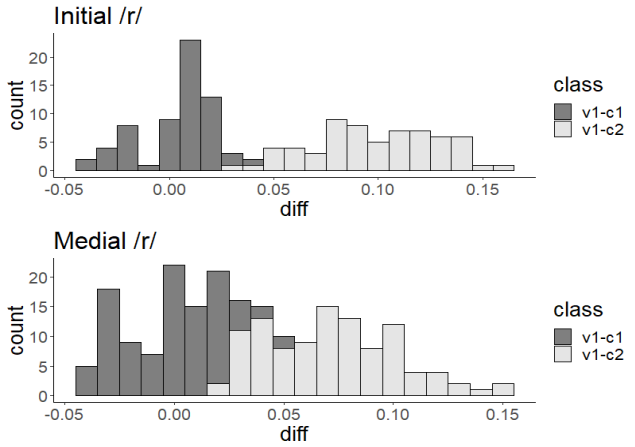


Figure 4: Histograms of v1 – c1 and v1-c1 in initial- and medial-/r/.

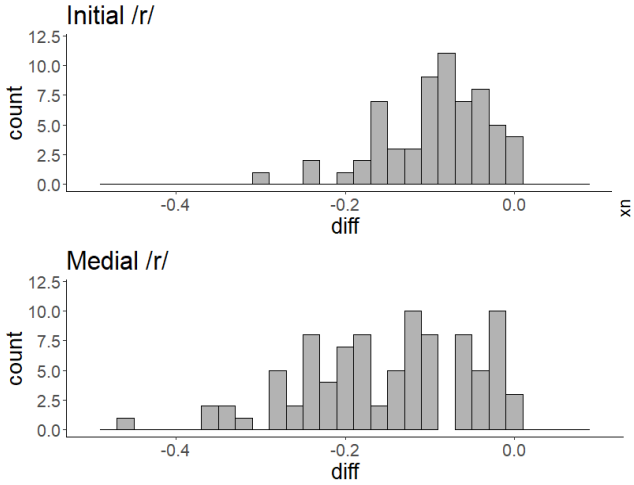


Figure 5: Histograms of the vertical displacement of the tongue tip between the timing of /r/ closure and the preceding frame.

Additional analysis of MoA was conducted concerning the vertical displacement of the tongue tip between the timing of /r/ closure and the preceding frame (see section 2). Figure 5 compares the distributions of the difference between the Y-values of the tongue tip in the frame corresponding to the /r/ closure (v1) and that of the preceding frame (pv1), i.e., v1 – pv1. As can be seen from the figure, all sample values have negative values, meaning that the tongue tip moved upward from the preceding frame to the /r/ in all samples (Note that lower and higher articulator locations correspond respectively to larger and smaller Y-values). Here again, this is what is expected in the “tap” rather than “flap” articulation.

Variables	c1x			c2x		
	Estimate	p		Estimate	p	
Intercept	0.050	0.013	***	0.125	0.000	***
pvwl_a	0.038	0.008	**	0.040	0.003	**
pvwl_o	0.050	0.008	***	0.050	0.000	***
fvwl_e	0.016	0.008		0.019	0.141	
fvwl_u	0.018	0.202		0.014	0.262	
fvwl_a	0.014	0.313		0.017	0.194	
fvwl_o	0.026	0.065	.	0.026	0.049	*
pvwl_a : fvwl_e	-0.023	0.244		-0.022	0.230	
pvwl_o : fvwl_e	-0.014	0.473		-0.015	0.418	
pvwl_a : fvwl_u	-0.007	0.728		0.004	0.812	
pvwl_o : fvwl_u	-0.013	0.500		-0.017	0.350	
pvwl_a : fvwl_a	-0.015	0.457		-0.009	0.628	
pvwl_o : fvwl_a	-0.007	0.727		-0.020	0.277	
pvwl_a : fvwl_o	-0.008	0.668		-0.020	0.280	
pvwl_o : fvwl_o	0.001	0.963		-0.020	0.267	

‘***’ p<.001 ‘**’ p<.01 ‘*’ p<.05 ‘.’ p<.1

Table 1: Summary of the GLMM analyses

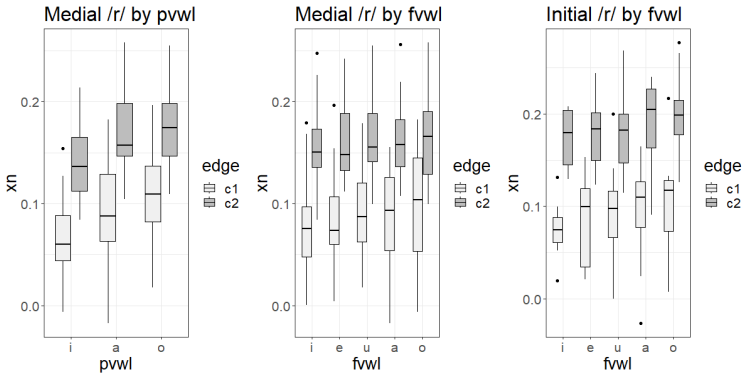


Figure 6: Effects of the preceding and following vowels on the X-values of c1 and c2.

3.3. Variation of the PoA

To examine the effects of adjacent segments on the variation of the PoA (Figure 2), Generalized Linear Mixed-effect Model (GLMM) regression analyses were conducted using the lme4 [6] and lmerTest [7] in R [8]. The X-values of the c1 and c2 of medial /r/ were predicted using the preceding vowels (/i/, /a/, or /o/, abb. “pvwl”), the following vowels (/a/, /i/, /u/, /e/, or /o/, abb. “fvwl”), and all interactions between the pvwl and fvwl. In addition, the subject was used as a random effect on the intercept. The result is summarized in Table 1. The effects of the pvwl are all significant (p<.05), while the effect of the fvwl and all interactions were not significant with the sole exception of the fvwl /o/. Note here the significances were computed between the first member of the factor level (i.e., /i/) and each of the remaining members (i.e., /i/ vs. /a/ and /i/ vs. /o/ in the case of the pvwl). Similar

analysis of the effect of fvwl on the initial /r/ revealed that following /o/ and /a/ had significant effects on both c1x and c2x at the levels of 0.05 or less. Figure 6 is a box-whisker representation of the effects of the pvwl and fvwl on the x values of the medial and initial /r/.

4. MODELING

Analyses in sections 3.1 and 3.3 revealed that the PoA of the Japanese /r/ has a more restricted distribution than Japanese /s/, but the distribution varies by the influence of the adjacent vowels. To know how well it is possible to predict the variation of the PoA by the adjacent vowels, GLMM modelling was conducted. X-values of the c1 and c2 were predicted by four GLMM models and the predictive performance of the models was evaluated by leave-one-out cross-validation (LOOCV).

Model	/r/ positon	LMER models	Corr. coef.	Mean Error
1	Medial	~pvwl + fvwl + edge + (1|subject)	0.898	0.021
2	Medial	~pvwl + edge + (1|subject)	0.897	0.020
3	Medial	~fvwl + edge + (1|subject)	0.845	0.025
4	Initial	~fvwl + edge + (1|subject)	0.933	0.019

Table 2: Summary of the LOOCV

Table 2 summarizes the results. The five columns of the table show respectively the model ID, the position of the target /r/ (either medial or initial), the model used for the prediction (formulae notation is in the lmer format), coefficient of product-moment correlation between the observed and predicted values, and the mean (root-mean-square) prediction error. Note, in the third column, the notation "(1|subject)" means that the subject is used as a random effect, and "edge" is a variable that takes binary values of either 'c1' or 'c2'. This variable was introduced to estimate the c1 and c2 simultaneously.

All models worked well. Especially, with models 1, 2, and 4 it was possible to explain more than 80% of the variance of the c1 and c2 on the X-axes. Figure 7 is the scatter plots of the observed and predicted values in cases of models 2 and 4.

5. DISCUSSION AND CONCLUSION

Sections 3 and 4 provide answers to the questions raised in section 1. As for the MoA, it is mistaken to regard the Japanese /r/ as a "flap" consonant to the exclusion of "tap" articulation. All samples of the /r/ analysed in this study were realized as "taps" (Figures 4, 5). Although there remains some possibility that the /r/ is realized as a "flap" consonant in natural settings, such articulation is unlikely to be observed at least in experimental settings like the recording of the rtMRIDB. It is not completely clear from the literature [2] if Okada really excluded the possibility of "tap" articulation, but his preference to the [ɽ] symbol strongly suggests the flap-only interpretation, since it is widely believed that the lower surface of the tongue tip makes closure in articulating typical retroflex consonants [4,5].

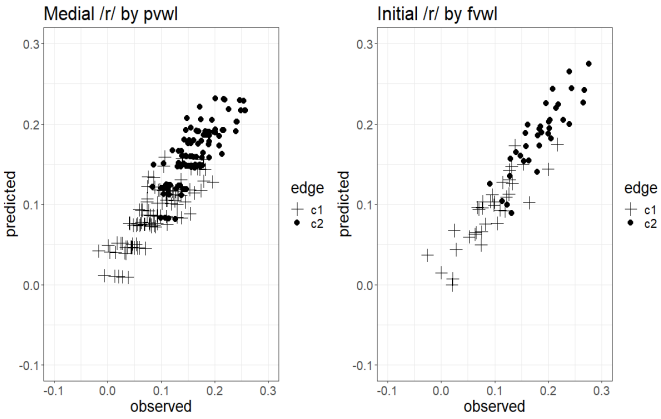


Figure 7: Scatter plots of the predicted and observed X-values of c1 and c2. Models 2 (left) and 4 (right).

As for the PoA, it is clearly mistaken to regard the Japanese /r/ as a post-alveolar sound. /r/ is more frontal than palatalized /s/ (Figure 2). This is contradictory to the description of [2] that regards the Japanese /r/ as a post-alveolar consonant. It is rather an alveolar sound to the exclusion of the post-alveolar articulation. It is, however, unclear if the Japanese /r/ needs to be phonologically specified as an alveolar consonant. As shown by Tables 1, 2, and Figures 6, 7 the precise PoA of the /r/ is almost entirely determined by its adjacent vowels. One possible interpretation is that the /r/ does not have the specification of the PoA as suggested by [9] in somewhat different context. It may be the case that a simple tapping of the tongue tip is superimposed on the global movement of the tongue determined by the adjacent vowels. According to this view, the place of contact between the tongue tip and the palate is determined not by the /r/ but by the adjacent vowel(s). The view that sees consonant articulation as a superimposed event on the articulation of a sequence of vowels is not a novel one; see [10] and its modern version [11], among others.

Moreover, the underspecification view provides a simple explanation for the finding that palatalization does not affect the PoA of the /r/ (Figure 2). Because palatalization is concerned with raising the front of the tongue, it does not affect the movement of the tongue tip and the PoA thereby.

To conclude, putting the issue of phonological underspecification aside, the current description of the Japanese /r/ in the IPA Handbook needs to be revised entirely.

Acknowledgement: This work is supported by the JSPS KAKENHI grants 17H02339 and 20H01265. It is also supported by a grant from the National Institutes for the Humanities.

6. REFERENCES

[1] International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.*" Cambridge. Univ. Press.

[2] Okada, H. (1999). "Japanese". In International Phonetic Association (1999).

[3] Maekawa, K. (2021). "Production of the utterance-final moraic nasal in Japanese: A real-time MRI study". *Journal of the International Phonetic Association*, doi:10.1017/S002510032 1000050

[4] Catford, J.C. (1977). *Fundamental Problems in Phonetics.* Edinburgh Univ. Press.

[5] Ladefoged, P. & I. Maddieson (1996). *The Sounds of the World's Languages*. Blackwell.

[6] Bates D., Mächler M., Bolker B., & Walker S. (2015). "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.

[7] Kuznetsova, A., Brockhoff, P.B., Christensen, R. H. B. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". *Journal of Statistical Software*, 82(13), 1–26. doi:10.18637/jss.v082.i13.

[8] R Core Team (2022). R: A language and environment for statistical computing. URL https://www.R-project.org/.

[9] Mester, R.A. & J. Ito (1989). "Feature predictability and underspecification: Palatal prosody in Japanese mimetics". *Language*, 65 (2), 258-193.

[10] Öhman, S. E. G. (1966). "Coarticulation in VCV utterances: Spectrographic Measurements". *JASA,* 39 (1), 151-168.

[11] Fujimura, O. (2000). "The C/D model and prosodic control of articulatory behavior". *Phonetica*, 57, 128-138.

AN ARTICULATORY STUDY OF THE WORD-INITIAL /stɹ/ CLUSTER IN BRITISH ENGLISH

Mitsuhiro Nakamura

Department of English, Aoyama Gakuin University, Tokyo, Japan
m_nakamura@cl.aoyama.ac.jp

ABSTRACT

This paper explores the articulatory characteristics of the word-initial /stɹ/ cluster in British English. Of specific interest is the phenomenon called /s/-retraction. Using EPG and EMA data collected from the MOCHA-TIMIT database, variable realizations of the /stɹ/ cluster were analyzed for four speakers. The tongue-palate contact pattern of the /s/ in the /stɹ/ cluster (e.g. *street*) was compared with that of the onset singleton /s/ and /ʃ/ (e.g. *seat, sheet*). The temporal extent of anticipatory labial coarticulation was inspected. The results reveal that (i) for one speaker, the contact anteriority index of the /s/ in the /stɹ/ cluster is significantly different from that of the onset singleton /s/, (ii) /t/-affrication (e.g. *street*>*s*[tʃ]*reet*) is found for all the speakers, (iii) the constriction peak of the lower lip gesture is systematically attained at the release of the /t/ in the cluster. These preliminary findings are discussed in terms of speaker-dependent articulatory strategies.

Keywords: /s/-retraction, EPG, EMA, MOCHA-TIMIT

1. INTRODUCTION

The phenomenon of /s/-retraction in the /stɹ/ cluster (e.g. *street* > [ʃ]*treet*) has been investigated in various accents of English: for instance, [3, 11, 13, 14] for American English, [1, 2] for British English, [15] for Australian English, [9, 17] for New Zealand English, [16] for a cross-dialectal analysis. The current study aims to further investigate the articulatory nature of the /s/-retraction phenomenon in British English.

Previous studies have identified that the phonetic basis of /s/-retraction is related to the lower spectral frequency: the alveolar /s/ in the /stɹ/ sequence has been judged to become closer to the postalveolar /ʃ/ due to its lower fricative noise frequency [2, 3, 11]. This measure successfully captures the systematic acoustic variation of /s/ in various contexts. In the current study, an attempt will be made to explore the 'articulatory' retraction of /s/ to /ʃ/ in the /stɹ/ cluster using electropalatographic data from the multi-channel articulatory database, MOCHA-TIMIT.

A previous study of /s/-retraction in British English [2] has proposed that /t/-affrication is the most important factor in modeling /s/-retraction in the /stɹ/ cluster and the rhotic /ɹ/ plays only an indirect role. According to [2], the retraction of /s/ occurs not only in the consonant sequence /stɹ/ (eg. *street*) but also in /stj/ (phonetically [stʃ] as in *student*). Since the consonant sequence /tj/ is affricated, the trigger for /s/-retraction is considered to be the affricate [tʃ]. This proposal effectively supports Lawrence's 'local assimilation' [9] as the mechanisms of /s/-retraction, rather than Shapiro's 'assimilation at a distance' [13].

The mechanism proposed by [2] is appealing but needs further examination. The first point concerns /t/ affrication mentioned above. It has been pointed out by [6, 10, 18] that the /tɹ, dɹ/ and /tj, dj/ clusters are realized as the affricates [tʃ, dʒ], but this tendency has not been substantiated fully by the empirical data. The second point is the coarticulatory effects of the rhotic /ɹ/. A study of the singleton /ɹ/ [5] has shown that the /ɹ/ involves more than one lingual constriction; the articulators are controlled in the order of the lips > the tongue tip/blade > the tongue root at the syllable onset position, and in the order of the tongue root-the lips > the tongue tip/blade at the coda position. Furthermore, long-range coarticulatory effects have been reported for the /l, ɹ/ in Southern British English [19]: adjacent and non-adjacent vowels preceding /ɹ/ have stronger lip rounding and tongue dorsum raising and retraction. It can be asked how the lip gesture is realized in the context of consonant clusters.

The current study aims to (i) provide initial empirical data for the /s/ articulation in the /stɹ/ cluster and (ii) examine the temporal extent of the lip gesture in the production of the /stɹ/ cluster. We will conclude with future directions of this study.

2. METHOD

2.1. Data collection

The speech materials used in this study were drawn from the multichannel articulatory database, MOCHA-TIMIT described by [20, 21]. This database comprises the articulatory and acoustic data of 460 phonetically balanced sentences read by native speakers of English. Data acquisition was made with Carstens Electromagnetic Articulograph (EMA, 500Hz sample rate), Reading Electropalatograph (EPG, 200Hz sample rate), Electrolaryngograph (Lx),

and acoustic recordings. In the EMA data recording, the tongue tip (TT) coil was placed at about 5-10mm, the tongue blade (TB) coil was placed at about 25-40mm, and the tongue dorsum (TD) coil was placed at about 50-70mm from the tongue tip, as measured with the tongue extended. In this study, the EMA data and the EPG contact patterns were analyzed for the selected utterances spoken by four speakers of Standard Southern British English referred to as SE (female), JM (female), SA (male), and AP (male).

For collecting the speech materials, the MOCHA corpus was searched and two sets of data were prepared. Dataset 1 consisted of words with the word-initial /stɹ/ cluster: the target words were preceded by a schwa [ə] or non-coronal consonant. Dataset 2 consisted of words with the word-initial (onset) /s/ and words with the word-initial (onset) /ʃ/. These target words were preceded by a schwa [ə]. All the words collected are content words. The two datasets are summarized in Table 1.

Dataset 1	Dataset 2
street, straightforward, stray, stronghold, strongly	see, seeking, same, save, saw
	shipbuilding, shell, shaving, shocked, shore

Table 1: Target words in dataset 1 and 2.

The position in the sentence of these target words varied. The number of tokens in dataset 1 was 20 in total (5 words x 4 speakers) and that in dataset 2 was 40 in total (10 words x 4 speakers).

2.2. Analysis

The analysis of the EPG and EMA data was performed using a set of programs called EMA tools [21]. The Reading EPG artificial palate has 62 electrodes arranged in eight rows and eight columns. The surface of the palate was divided into two major zones for data analysis: an alveolar zone for rows 1-5 and a palatal zone for rows 6-8. These zones were subdivided into dentalveolar (row 1), alveolar (row 2), postalveolar (rows 3-4), prepalatal (row 5), medio-palatal (rows 6-7), and postpalatal (row 8).

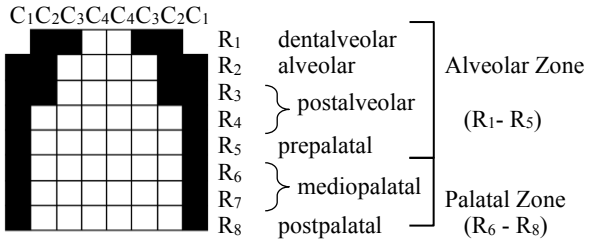


Figure 1: EPG palatogram and two zones.

The configurational characteristics of the EPG contact patterns in the alveolar zone were studied. For the /stɹ/ cluster (dataset 1), three EPG frames were identified. The MAX (linguopalatal narrowing) frame of /s/ was determined by the mid-point of the continuous noise energy, specified by a spectrogram. The two other frames were the first frame showing the complete closure for /t/ and the first frame showing the release of /t/. For the onset /s/ and /ʃ/ (dataset 2), the MAX frame mentioned above was specified.

The contact index method [7, 12] was used to examine the configurational characteristics of the EPG contact patterns. The contact indices were CAa (alveolar contact anteriority index), CCa (alveolar contact centrality index), and Qp (quotient of overall electrode activation at the palatal zone, rows 6-8). The indices CAa and CCa were calculated using the following formula [12].

(1) $CAa=[\log[[1(R_5/8)+9(R_4/8)+81(R_3/8)+739(R_2/8)+4921(R_1/6)+1]]/[\log(5741+1)]$

(2) $CCa=[\log[[1(C_1/8)+11(C_2/10)+121(C_3/10)+1331(C_4/10)]+1]]/[\log(1464+1)]$

The CAa index value increases with alveolar contact fronting. The CCa index was calculated for the four symmetrical columns (C_1 to C_4, C_4 to C_1 in Figure 1) and the CCa value increases as contact becomes more central in the alveolar zone. The Qp index was obtained by averaging all on-electrodes in the palatal zone by the total number of 24 electrodes. This value is related to the degree of tongue dorsum raising.

In order to explore the timing of the lip gesture on the /stɹ/ cluster, the EMA data of the lower lip movement (LLx, LLy) was examined. The time course of the lower lip gesture was identified by the four points (onset, target, release, and release offset) proposed by [8]. These time points were specified at 20% of the tangential velocity peak associated with movement towards, or away from, a target constriction. The total duration was the time interval between the onset and the release offset of a given movement and this was used to inspect the temporal extent of the lip gesture in the production of the /stɹ/ cluster.

The question to be explored in this study was which phase of the /stɹ/ cluster the constriction peak of the lower lip gesture was aligned to. For this purpose, the times of the tangential velocity minimum of the lower lip vertical movement (LLy) and the lingual gestures (the tongue tip, blade, and dorsum) were identified. Based on a qualitative observation, the interaction between the lip gesture and the lingual gesture was studied.

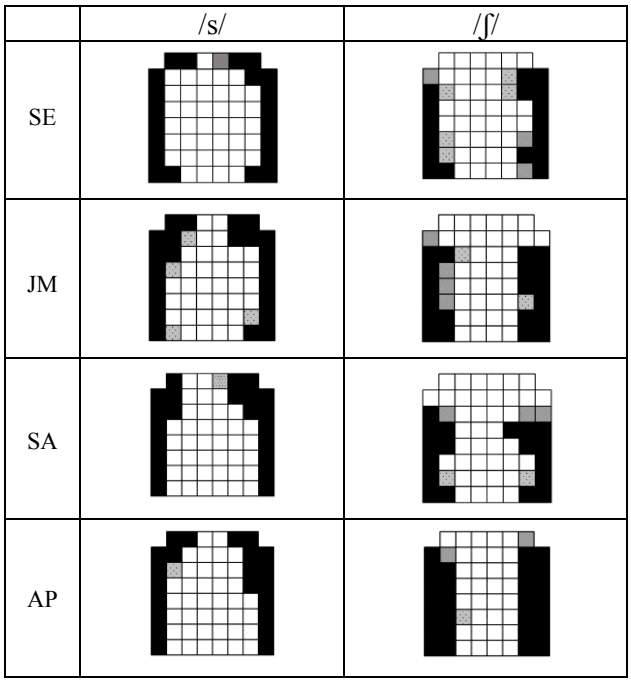


Figure 2: Averaged EPG contact patterns for the word-initial /s/ and /ʃ/. ('black' 100-80%; 'grey' 79-60%; 'dotted' 59-40%; 'white' 39≥)

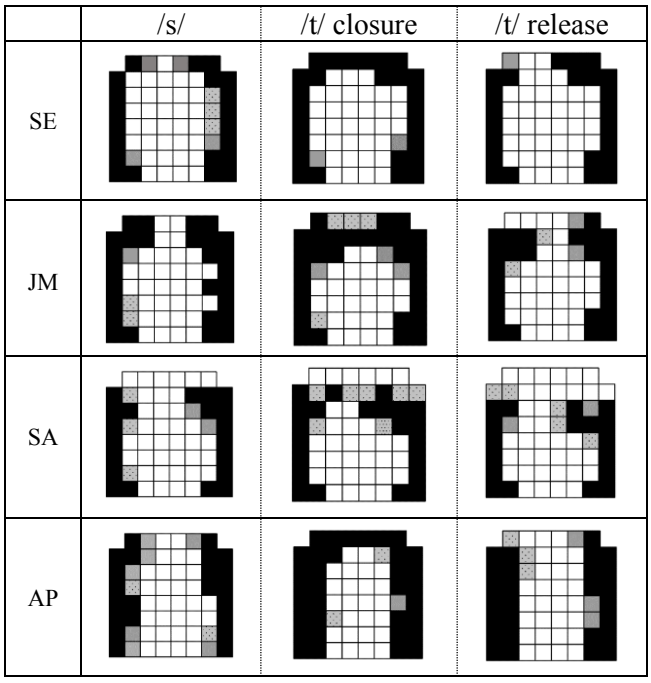


Figure 3: Averaged EPG contact patterns for the word-initial /stɹ/ cluster: /s/ at the MAX point, /t/ closure, and /t/ release. ('black' 100-80%; 'grey' 79-60%; 'dotted' 59-40%; 'white' 39≥)

3. RESULTS AND DISCUSSION

3.1. Linguopalatal contact patterns

Figure 2 above shows the EPG palatograms of the word-initial /s/ and /ʃ/ averaged over the productions of the target words. Figure 3 shows the EPG palatograms of the word-initial /stɹ/ cluster, which illustrate the MAX constriction of /s/, the closure of /t/, and the release of /t/, averaged over the productions of the target words.

The EPG patterns in Figure 2 serve as a basis for comparison. For the word-initial /s/, a narrow constriction is formed in the dentalveolar and alveolar regions (rows 1-2) with accompanying side contact. There are noticeable differences in the constriction formation among the speakers: apical (SE) and apicolaminal (JM, SA, AP) realization. For /ʃ/, the narrow constriction is primarily made in the alveolar and postalveolar regions (rows 2-4). This suggests that the tongue blade is involved in the constriction formation with the tongue tip directed downwards.

speaker	/s/	/ʃ/	/s/ in /stɹ/
SE	0.96 (0.01)	0.65 (0.04)	0.95 (0.02)
JM	0.96 (0.02)	0.54 (0.15)	0.95 (0.01)
SA	0.93 (0.01)	0.41 (0.15)	0.71 (0.01)
AP	0.95 (0.01)	0.77 (0.09)	0.93 (0.02)

Table 2: Mean values of the alveolar contact anteriority index (CAa) for singleton /s/ and /ʃ/, and /s/ in the /stɹ/ cluster. (s.d. in the bracket)

For /s, ʃ/ in the word-initial position and /s/ in the /stɹ/ cluster, the mean values of the contact indices are summarized: CAa in Table 2, CCa in Table 3 , and Qp in Table 4 (next page). To examine the differences between the singleton /s/ and the /s/ in the cluster, a t-test was carried out separately for the four speakers.

An examination of CAa showed a significant decrease only for speaker SA [$t(8)=25.22$, $p<0.01$]: SE [$t(8)=0.92$, $p=0.38$], JM [$t(8)=0.74$, $p=0.47$], and AP [$t(8)=1.93$, $p=0.08$]. Thus, the constriction location in the alveolar zone shifted considerably further back in SA's production.

In accordance with the results of CAa, a comparison of CCa revealed a significant decrease only for speaker SA [$t(8)=3.78$, $p<0.01$]: SE [$t(8)=0.06$, $p=0.95$], JM [$t(8)=0.85$, $p=0.41$], and AP [$t(8)=0.68$, $p=0.51$]. As seen in Figure 3, for SE, JM, AP, there is no substantial divergence between the constriction width for the /s/ in the cluster and that for the word-initial /s/. In contrast, for SA, the constriction width becomes wider at the (post)alveolar region (rows 2-3).

speaker	/s/	/ʃ/	/s/ in /stɹ/
SE	0.61 (0.11)	0.35 (0.08)	0.60 (0.12)
JM	0.57 (0.08)	0.33 (0.14)	0.54 (0.04)
SA	0.70 (0.01)	0.36 (0.09)	0.49 (0.11)
AP	0.48 (0.01)	0.34 (0.05)	0.47 (0.04)

Table 3: Mean values of the alveolar contact centrality index (CCa) for singleton /s/ and /ʃ/, and /s/ in the /stɹ/ cluster. (s.d. in the bracket)

speaker	/s/	/ʃ/	/s/ in /stɹ/
SE	33.33 (4.56)	41.67 (8.74)	46.67 (9.28)
JM	36.67 (7.17)	45.83 (3.73)	39.17 (6.24)
SA	27.50 (3.33)	38.33 (4.86)	34.17 (5.53)
AP	26.67 (2.04)	52.50 (5.65)	39.17 (11.67)

Table 4: Mean values of the quotient of overall electrode activation at the palatal zone, rows 6-8 (Qp) for singleton /s/ and /ʃ/, and /s/ in the /stɹ/ cluster. (s.d. in the bracket)

The results of Qp indicated a significant increase only for SE [t(8)=-2.57, p=0.03]: JM [t(8)=-0.52, p=0.61], SA [t(8)=-2.06, p=0.07], and AP [t(8)=-2.11, p=0.06]. As seen in Table 4, the increase in the palatal contact (rows 6-8) was found for all the speakers, and the significance of SE might be related to the production characteristic specific to the speaker: the apical articulation of /s/.

Based on the results of comparing the EPG indices, /s/-retraction was found for SA (characterized as a retractor) but not for SE, JM, and AP (characterized as non-retractors). Note that all the four speakers pronounced the /t/ in the /stɹ/ cluster as affricated.

It is worth noting that, while the non-retractors showed the EPG contact configuration of the /s/ in the cluster similar to the word-initial singleton /s/, they revealed idiosyncratic patterns at the /t/ closure and release phase. For JM, the contact at the frontmost row (row 1) tended to be lost: the complete closure for /t/ shifted backward (row 2); the constrictive approximation at the release phase of /t/ became similar to that of /ʃ/. For AP, the contact at both sides increased at the closure and release phase of /t/, and the contact configuration at the release phase was almost the same as the word-initial /ʃ/. For SE, the EPG contact at the closure and release phase of /t/ showed no substantial changes to /ʃ/: SE's production of the /stɹ/ cluster can be transcribed as [sts^{w}ɹ]. It is evident from our observation that when the 'articulatory' retraction of /s/ occurs in the /stɹ/ cluster, the affrication of /t/ also takes place, not vice versa.

3.2. Timing of the lower lip movement

Figure 4 (a, b) below shows the EMA recording of the tongue tip (TT), the tongue blade (TB), the tongue dorsum (TD), and the lower lip (LL) in the production of '*street*' in the MOCHA sentence, '*Spring Street is straight ahead.*' For speaker SA (a retractor) and SE (a non-retractor), the raising gesture of the lower lip extends over the whole /stɹ/ cluster (see the box in the LLy panel). The constriction peak of the lip gesture is attained at the release phase of the /t/. After the release of /t/, the lingual gestures (TTx, TBx, TDx) move smoothly to prepare the following /ɹ/. (For SE, no clear TD velocity minimum was observed.)

Such a coarticulatory accommodation of the lower lip was systematically observed across speakers and the /stɹ/ cluster in the target words. This could be considered as the effects of the rhotic /ɹ/. Another possibility is that, if the /stɹ/ sequence is realized phonetically as [stʃɹ], a postalveolar affricate might be accompanied by the lip rounding [4]. To understand these patterns fully, we need to analyze the rhotic gesture in various consonant sequences (e.g. /st/, /tɹ/, /spɹ, skɹ/), and to investigate the acoustic consequences of the lip gesture in the production of the cluster involving the /ɹ/.

4. CONCLUSION

The current study has explored the articulatory characteristics of the /stɹ/ cluster in British English. We have shown that (i) while /t/-affrication is observed for all the speakers, the 'articulatory' retraction of /s/ in the /stɹ/ cluster is found for one speaker only; (ii) the constriction peak of the lower lip gesture is systematically attained at the release phase of the /t/ in the sequence; and (iii) the non-retractors reveal idiosyncratic EPG contact patterns at the closure and release phase of /t/. These results provide evidence for speaker-dependent strategies and variability in the realization of /s/-retraction. Articulatory coordination in various clusters and its acoustic consequences will be substantiated by subsequent research.

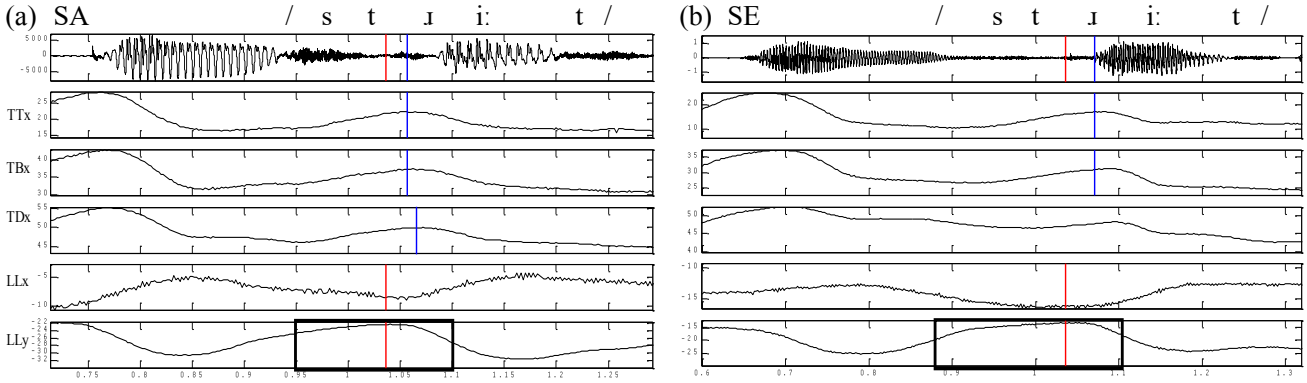


Figure 4: EMA trajectories of TTx, TBx, TDx, LLx, and LLy for (a) speaker SA and (b) speaker SE. (red line=the constriction peak of LL; blue line=the constriction peak of TTx, TBx,TDx)

5. ACKNOWLEDGEMENTS

I would like to thank anonymous ICPhS reviewers for their critical and constructive comments. All remaining errors are of course mine. This research is supported in part by a grant from Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) 2023, Grant no. 21K00487.

6. REFERENCES

[1] Altendorf, U., MacDonald, R., Thielking, N. 2021. S Retraction in the South-East of England: A Snapshot of the Emergence of a Sound Change in Progress. *Anglistik: International Journal of English Studies* 32(1), 45-64.

[2] Bailey, G., Nichola, S., Turton, D., Baranowski, M. 2022. Affrication as the cause of /s/-retraction: Evidence from Manchester English. *Glossa: a Journal of general linguistics* 7(1), 1-30.

[3] Baker, A., Archangeli, D., Mielke, J. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change* 23(3), 347-374.

[4] Brown, G. 1981. Consonant Rounding in British English: The Status of Phonetic Descriptions as Historical Data. In Asher, R.E., Henderson, E.J.A. (eds) *Towards a History of Phonetics*. Edinburgh: Edinburgh University Press, 67-76.

[5] Campbell, F., Gick, B., Wilson, I., Vatikiotis-Bateson, E. 2010. Spatial and Temporal Properties of Gestures in North American English /r/. *Language and Speech* 53(1), 49-69.

[6] Cruttenden, A. 2019. Gimson's *Pronunciation of English.* Eighth Edition. London: Routledge.

[7] Fontdevila, J., Pallares, M.D., Recasens, D. 1994. The contact index method of EPG data reduction. *J. of Phonetics* 22, 141-154.

[8] Gafos, A. 2002. A Grammar of Gestural Coordination. *NLLT* 20, 269-337.

[9] Lawrence, W.P. 2000. /str/→/ʃtr/: Assimilation at a Distance? *American Speech* 75, 82-87.

[10] Lindsey, G. 2019. *English after RP: Standard British Pronunciation Today*. Cham: Palgrave Macmillan.

[11] Mielke, J., Baker, A., Archangeli, D. 2011. Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction. In Fougeron, C., Kuehnert B., D'Imperio, M., Vallee, N. (eds) *Laboratory Phonology 10*. Berlin: De Gruyter Mouton, 699-729.

[12] Recasens, D. Espinosa, A. 2007. An electro-palatographic and acoustic study of affricates and fricatives in two Catalan dialects. *J. of the IPA* 37, 143-172.

[13] Shapiro, M. 1995. A case of distant assimilation: /str/ →/ʃtr/. *American Speec*h 70, 101-107.

[14] Smith, B.J., Mielke, J., Magloughlin, L., Wilbanks, E. 2019. Sound change and coarticulatory variability involving English /ɹ/. *Glossa: a journal of general linguistics* 4(1), 1-51.

[15] Stevens, M., Harrington, J. 2016. The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English. *J. of Phonetics* 58, 158-134.

[16] Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., Thomas, E. 2019. Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. In Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds) *Proc. 19th ICPhS*, Melbourne, 1273-1277.

[17] Warren, P. 2006. /s/-retraction, /t/-deletion and regional variation in New Zealand English /str/ and /stj/ clusters. In Warren, P., Watson, C.I. (eds) *Proc. 11th Australian International Conference on Speech Science and Technology, University of Auckland, New Zealand, December 6-8*, 466-471.

[18] Wells, J. 1982. *Accents of English.* 3 vols. Cambridge: Cambridge University Press.

[19] West, P. 1999. The Extent of Coarticulation of English Liquids: An Acoustic and Articulatory Study. In Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., Bailey, A.C. (eds) *Proc.14th ICPhS*, San Francisco, 1901-1904.

[20] Wrench, A. 1999. MOCHA-TIMIT. https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[21] Wrench, A. 2000. A multi-channel/multi-speaker articulatory database for speech recognition. *Phonus* 5, 1-13.

MARKERLESS TONGUE POSE ESTIMATION IN CROSS-LINGUISTIC SPOKEN AND MIMED SPEECH USING ULTRASOUND

Annslin Maria Lukose, Amelia Gully, George Bailey

University of York, York, UK
annslinmaria46@gmail.com, amelia.gully@york.ac.uk, george.bailey@york.ac.uk

ABSTRACT

Studies of articulatory variation in mimed speech, a proxy for alaryngeal speech, have been restricted to monolingual speakers. This study presents a single-subject case study that investigates cross-linguistic phonological influence of vowels in a 25-year-old female Malayalam-English bilingual speaker, producing spoken and mimed speech. Ultrasound tongue imaging is used to acquire articulatory data, and novel techniques for data analysis are used including pose estimation with DeepLabCut™ and geometric morphometric analysis (GMM) for quantitative measurement of tongue variation. The results reveal significant differences in tongue kinematics across languages and conditions, without direct indication of phonological transfer. However, within-speaker articulatory variation is evident across the independent variables. These findings will aid in the development of silent speech interfaces for voice restoration in bilingual laryngectomy patients.

Keywords: Ultrasound tongue imaging, pose estimation, Geometric morphometrics, mimed speech

1. INTRODUCTION

A bilingual individual is someone who, in their everyday lives, uses two (or more) languages or dialects [11]. Based on the Speech Learning Model (SLM), the sounds comprising the L1 and L2 phonetic subsystems are interconnected, and this interaction occurs because the L1 and L2 sounds exist in a common "phonological space" [10]. As per the revised Speech Learning Model (SLM-r), one reason for the difference in learning outcomes of L1 and L2 is that, since the sounds of L2 are automatically related to the phonetic inventory of L1, L2 sounds are initially replaced by L1 sounds. This would then result in cross-linguistic influence or phonological transfer, which refers to “the ways in which a person’s knowledge of the sound system of one language can affect that person’s perception and production of speech sounds in another language” [13].

Due to their absence of larynxes, laryngectomy patients can only mime speech, and they rely on different voice restoration methods to achieve communication with their alaryngeal speech. Some studies have compared articulation across normal, whispered and mimed speech, for monolingual English [9] [20], and French [4] speakers. The current study is the first investigation to study articulatory variation in spoken and mimed speech conditions in two languages spoken by a single speaker.

With the implementation of machine learning tools, speech can be predicted from articulation in a technique called articulatory-to-acoustic (forward) mapping [5], which was further utilised to develop ‘Silent Speech Interface’ systems (SSI). SSI involves recording motions of the articulators in the absence of audible acoustic signals and automatically synthesising speech based on those movements [6], [7]. Laryngectomy patients are suitable for ultrasound-based SSI since neither glottal excitation nor vocal tract airflow are necessary [7]. Ultrasound tongue imaging (UTI) is a non-invasive and safe technique [3]. Additionally, with regard to cost of equipment, safety, portability and structures visualised, UTI is a more cost-effective technique compared to electromagnetic articulography (EMA) [16].

The objective of the current study is to obtain ultrasound images of the tongue to examine the phonological transfer or cross-linguistic articulatory variation of target vowels in a bilingual speaker speaking in two different conditions, one mimicking alaryngeal speech. An additional goal is to understand within-speaker articulatory variation across languages and conditions. The potential future application of this study is the improvement of ultrasound-based silent speech interfaces for bilingual laryngectomy patients. This paper is a single-subject case study that investigated speech samples collected in mimed speech (approximating alaryngeal speech) and spoken speech conditions from a Malayalam-English bilingual speaker, language variations which are not well studied particularly in this context. Malayalam is a south Indian language with five short and long vowels /ɪ e a o u i: a: o: u:/ [14], and two diphthongs (/ai/ and /au/). Indian English has a vowel system quite similar to the RP variety of British English, with [e] and [o] (monophthongs) replacing the lexical sets FACE and GOAT (diphthongs), respectively [22], which might be due to phonological transfer from L1 to L2.

In order to achieve the research objectives, tongue motion patterns during spoken and mimed production were collected using UTI. This was followed by assessing the cross-linguistic phonological influence of vowels in these speaking conditions and estimating within-speaker articulatory variation using state-of-the-art techniques such as pose estimation of the tongue using DeepLabCut™ (DLC) [21] and geometric morphometric (GMM) analysis. DLC was first implemented on tongue and lip movements by Wrench and Balch-Tomes [23]. Unlike manual labelling or semi-automatic labelling used in Articulate Assistant Advanced version 219.08 [2], DLC is an innovative method for obtaining tongue splines using machine learning, where the tongue contours are automatically landmarked, making numerical measurement of the articulatory kinematics possible.

2. METHODS

2.1. Data collection

A 25-year-old female native Malayalam speaker, whose second language was English, participated in this study.

Four long vowels (/i:/, /e:/, /o:/, and /u:/) at two vowel positions (initial and medial) in both Malayalam and English were investigated for this research, in keeping with similar studies (e.g., [4]). Final vowel position was excluded due to an insufficient number of meaningful words across both languages. Monosyllabic and bisyllabic words with vowels around alveolars and velars were used as stimuli, to ensure that tongue movements were clearly visible in the mimed condition, where audio cues were not available. A clap was also introduced before the target vowel while recording in the mimed speech condition in order to aid with phoneme segmentation during the analysis process. Two words each at initial and medial vowel positions were selected, and three repetitions of each of these words were recorded in both spoken and mimed conditions. A summary of the data set is displayed in Table 1. In total, there were 96 tokens.

Vowels	4: /i:/, /e:/, /o:/ and /u:/
Positions	2: Initial and medial
Words	2 words for each vowel at each position
Conditions	2: Spoken and mimed speech
Repetitions	3 repetitions of each word for each condition

Table 1: A summary of the data set

The midsagittal plane of the tongue for each repetition was captured with ultrasound tongue imaging (UTI), using the software Articulate Assistant Advanced (AAA), version 219.08 [2]. A MC4-2R20S-3 microconvex-array ultrasound transducer with a probe frequency ranging from 2 MHz to 4 MHz was used to collect ultrasound recordings of the tongue. The UltraFit [19] headset was used to hold the ultrasound probe in place. Audio was recorded directly into AAA using a Rode smartLav+ wired lavalier microphone at 22050 Hz sampling frequency.

The participant completed the recordings separately for each language, with a break after the ultrasonic recordings for one language. Considering the comfort of the participant, the probe was adjusted between languages. For each language, every word was recorded three times in the spoken speech followed by three times in the mimed condition.

2.2. Pose estimation

After obtaining the ultrasonic videos of the tongue for each repetition of the words, DeepLabCut™ was used to obtain tongue splines for each recording. An illustration of the pose estimated tongue contour obtained using DLC is shown in Figure 1 (first repetition of the word ‘ache’ in English in the spoken condition). Figure 1 depicts 11 key points along the surface of the tongue corresponding to the vallecula, tongue root (2x), tongue body (2x), tongue dorsum (2x), tongue blade (2x) and tongue tip (2x).

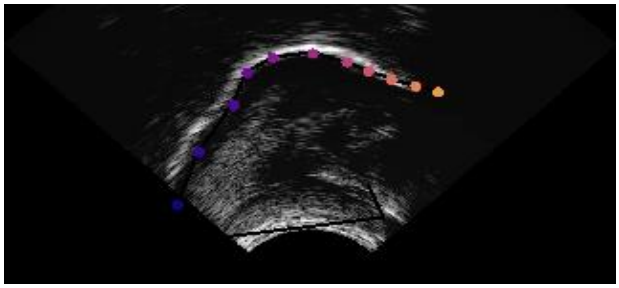


Figure 1: Pose estimated tongue contour for the word ‘ache’ in English in spoken speech condition

2.3. Data analysis

The midpoint method [8] was chosen to measure the target vowels.

In this research study, GMM was applied to the dataset in order to facilitate comparison of tongue shapes. GMM, developed by Bookstein, 1996a (as cited in [1]), is a statistically robust framework for quantifying and comparing shapes. The first step is Generalised Procrustes Analysis (GPA), a superimposition method [1] that minimises

differences in shape [15] by eliminating non-shape variation in the arrangement of landmarks [1] including scale, translation, and rotation. This has the added benefit of normalising any differences in position between recording sessions that may have occurred due to movement of the ultrasound probe. The next step is to quantify the variation in tongue shape across both languages and conditions using principal component analysis (PCA). For GPA-normalised shape data, PCA detects the axes of greatest shape variation in the dataset [15]. The first principal component (PC1) indicates the most shape variation in the data set, the second principal component specifies the second most shape variation in the data set, and so on. PCA has been used to quantify shape in the vocal tract [12], but its application to ultrasound-derived data is new.

3. RESULTS AND DISCUSSION

3.1. Visual analysis

Observations of Figure 2 suggests that the position of the tongue for both conditions vary noticeably across vowels, with obvious tongue height and tongue advancement deviations. For both the languages, most of these variances were seen in vowels /o:/ and /u:/ for both initial vowel position and medial vowel position. A small variation in tongue advancement could be observed for vowels /e:/ and /i:/ in the initial vowel position and medial vowel position, with English tongue splines almost overlapping for both the positions compared to Malayalam tongue postures for these positions.

Although cross-linguistic phonological influence is not evident in these figures, the deviations visible between conditions possibly indicate the reliance on auditory feedback, which is critical in altering the speech motor-control system as speech sounds are acoustically defined [18].

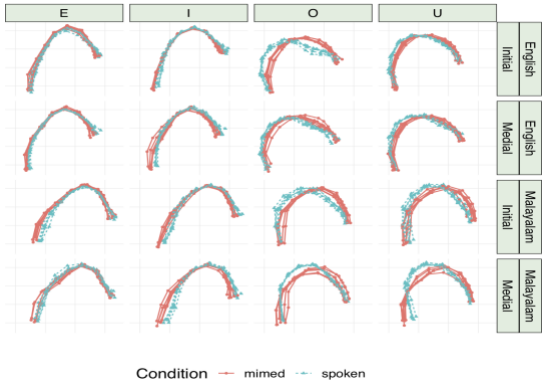


Figure 2: Plot displaying tongue shape variation across mimed and spoken speech conditions in both the languages (for every repetition of the words for each vowel)

3.2. Principal component analysis (PCA)

The obtained principal components scores are depicted in Figure 3 using R [17]. In the figure, each point represents a tongue shape for a single vowel repetition, in spoken speech condition (dots) and mimed speech condition (crosses). Figure 3 illustrates the tongue shapes that correspond to the extremes of the PC axes. The first principal component (PC1) accounted for 65.96% of tongue shape variation (tongue advancement differences), with negative values indicating front vowels and the positive values representing back vowels. The second principal component (PC2) accounted for 18.27% of tongue shape variation (tongue height differences), with positive values showing increased tongue height (elevation) and the negative values displaying reduced tongue height (depression).

Using PCA on shape data permits visualisation of the shape changes observed in the ultrasound data. Noticeable distinctions across vowels are apparent in Figure 3, based on speaking conditions (dots versus crosses) and languages. In comparison to the back vowels, the front vowels exhibit less articulatory variability between languages in both spoken and mimed conditions. On the other hand, back vowels for Malayalam in both the conditions are produced with increased tongue elevation, whereas for English, they are produced with reduced tongue height.

Many speakers of British English and American English use a high central rounded vowel (/ʉ:/) [22]. However, being a native Malayalam speaker, the participant in this study used a high back rounded vowel (/u:/) in Malayalam words, which is seen to be used in English words as well, suggesting the possibility of cross-linguistic phonological influence across languages. Furthermore, lack of auditory feedback might have contributed to the production of a somewhat high central rounded vowel in the mimed speech condition.

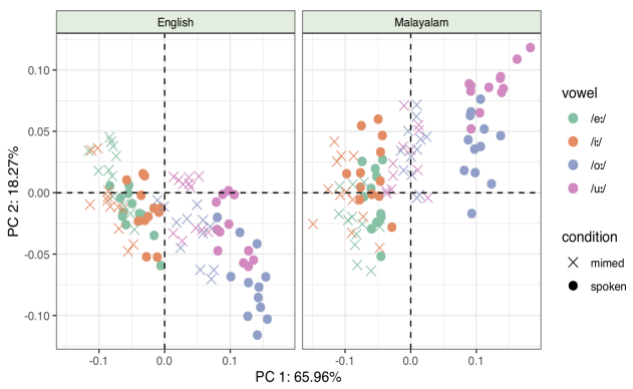


Figure 3: Combined PCA of shapes for four vowels across languages and conditions

A benefit of using GPA-normalised shape data is that differences in shape can be visualised directly, giving physical meaning to the axes in Figure 3.

Figure 4 presents deformation grids for mean, minimum and maximum scores of PC1 and PC2 for both the languages and conditions, illustrating how tongue shape changes throughout the study. From the Procrustes mean shape of the tongue towards the negative PC1 scores, the deformation grid is associated with more protrusion of the anterior part of the tongue and more gap at the posterior part of the tongue (possibly related to the front vowels /e:/ and /i:/). Towards the positive PC1 scores, the deformation grid corresponds to a retraction of the tongue body with less gap at the posterior part of the tongue (possibly related to the back vowels /u:/ and /o:/). In other words, PC1 indicated that the most shape variation in the dataset could be attributed to differences in tongue advancement.

Similarly, from the Procrustes mean shape of the tongue towards the negative PC2 scores, the deformation grid is associated with marginal lowering of the tongue or low tongue height (possibly relating to the low vowels /e:/ and /o:/). Towards the positive PC2 scores, the deformation grid corresponds to unnoticeably minimal raising of the tongue or high tongue height (possibly relating to the back vowels /u:/ and /i:/). Consequently, it could be concluded that PC2 indicated differences in tongue height (second most variation in the data set).

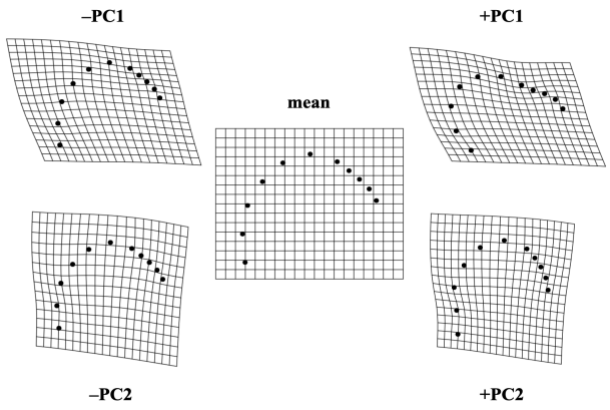


Figure 4: Deformation grids of mean, minimum and maximum scores for PC1 and PC2

Procrustes ANOVA revealed significant differences in the tongue position across conditions (spoken speech condition and mimed speech condition) (p=0.001) and across languages (Indian English and Malayalam) (p=0.001). Additionally, the study revealed a significant difference in tongue shape for different conditions across languages (p=0.018), suggesting that the effect of condition on tongue shape may be language-specific.

These results are indicative of within-speaker articulatory variation across languages and conditions. Although they do not directly indicate phonological transfer across languages, variation in tongue kinematics across independent variables may be suggestive of the influence of auditory feedback. Additionally, this study contributes to a better understanding of mimed speech in a bilingual speaker whose language variations are not well studied and provides information about within-speaker articulatory variation. This could contribute to future development of ultrasound-based SSI, aiming to give voice to bilingual alaryngeal speakers.

4. CONCLUSION

This paper reports the case study of a 25-year-old female Malayalam-English bilingual speaker. The study focused on estimating any phonological transfer of vowels across these languages and two speaking conditions, with mimed speech reflecting the alaryngeal speech. This paper presents a novel combination of state-of-the-art techniques for data collection (ultrasound tongue imaging) and data analysis (pose estimation using DeepLabCut™ and geometric morphometric analysis), with GMM being a new method for ultrasound derived data. The results revealed significant differences in the tongue kinematics across languages and conditions, without direct indication of phonological transfer. However, the understanding of within-speaker variation might help in transferring this knowledge in the development of an ultrasound-based Silent Speech Interface as a voice restoration method for bilingual laryngectomy patients.

7. REFERENCES

[1] Adams, D., Rohlf, F., Slice, D. 2004. Geometric morphometrics: Ten years of progress following the 'revolution'. *Italian Journal of Zoology,* 71(1), 5-16.

[2] Articulate Instruments Ltd. 2022. Articulate Assistant Advanced (Version 219.08) [Computer software]. Edingurgh: Articulate Instruments Ltd. Retrieved April 01, 2022 from http://www.articulateinstruments.com/downloads/

[3] Cleland, J., Wrench, A., Scobbie, J., Semple. S. 2011. Comparing articulatory images: An MRI/Ultrasound tongue image database. *Proceedings of the 9th International Seminar on Speech Production,* 163-170.

[4] Crevier-Buchman, L., Gendrot, C., Denby, B., Pillot-Loiseau, C., Roussel, P., Colazo-Simon, A., Dreyfus, G. 2011. Articulatory strategies for lip and tongue movements in silent versus vocalized speech. *Proceedings of ICPhS XVII*, pp. 1-4.

[5] Csapó, T. G. 2020. Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal

tract. *Proceedings of INTERSPEECH 2020,* 2722-2726.
[6] Csapó, T. G., Tóth, L., Gosztolya, G., Marko, A. 2021. Speech synthesis from text and ultrasound tongue image-based articulatory input. *11th ISCA Speech Synthesis Workshop (SSW 11),* 26-28.
[7] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., Brumberg, J. S. 2010. Silent speech interfaces. *Speech Communication,* 52(2010), 270-287.
[8] Di Paolo, M., Yaeger-Dror, M., Wassink, A. 2010. Analyzing vowels. In: Di Paolo, M., Yaeger-Dror, M. (eds), *Sociophonetics: A student's guide.* Routledge,
[9] Dromey, C., Black, K. 2017. Effects of laryngeal activity on articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 25(12), 2272-2280.
[10] Flege, J. E., Bohn O. 2021. The revised Speech Learning Model. In: Wayland, R. (ed.), *Speech Language Speech Learning, Theoretical and Empirical Progress.* Cambridge University Press.
[11] Grosjean, F. 1997. The bilingual individual. *Interpreting,* 2(1-2), 163-187.
[12] Gully, A. 2021. Quantifying vocal tract shape variation and its acoustic impact: A geometric morphometric approach. *Proceedings of INTERSPEECH 2021,* 3999-4003.
[13] Jarvis, S., Pavlenko, A. 2008. *Crosslinguistic influence in language and cognition.* Taylor & Francis.
[14] Krishnamurti, B. 2003. *The Dravidian Languages.* Cambridge University Press.
[15] Polly, P. 2012. *Procrustes, PCA, and 3D coordinates.* Geometric morphometrics module lecture notes. Indiana University: Department of Earth & Atmospheric Sciences.
[16] Porras, D., Sepúlveda-Sepúlveda, A., Csapó, T. G. 2019. DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging. *International Joint Conference on Neural Networks.*
[17] R Core Team. 2021. R (Version 4.1.2) [Computer software]. Vienna: Bell Laboratories. Retrieved November 1, 2022 from https://www.r-project.org/
[18] Simmonds, A. J., Wise, R. J. S., Leech, R. 2011. Two tongues, one brain: imaging bilingual speech production. *Frontiers in psychology,* 2(166), 1-13.
[19] Spreafico, L., Pucher, M., Matosova, A. 2018. UltraFit: A speaker-friendly headset for ultrasound recordings in speech science. *INTERSPEECH* 2018, 1517-1520.
[20] Teplansky, K., Tsang, B., Wang, J. 2019. Tongue and lip motion patterns in voiced, whispered, and silent vowel production. *Proceedings of ICPhS XIX,* 2831-2835.
[21] The Mathis Group, The Mathis Lab. 2022. DEEPLABCUT (Version 2.2.2) [Computer software]. Switzerland: The Mathis Group, Swiss Federal Institute of Technology Lausanne. Retrieved July 2022 from http://www.mackenziemathislab.org/deeplabcut
[22] Wells, J. 1986. *Accents of English 3: Beyond the British Isles.* Cambridge University Press.
[23] Wrench, A., Balch-Tomes, J. 2022. Beyond the edge: Markerless pose estimation of speech articulator from ultrasound and camera images using DeepLabCut. *Sensors*, 22(3), 1133-1162.

NASAL CODA MERGER IN HANGZHOU MANDARIN: A DYNAMIC ULTRASOUND STUDY

Yin Huang[1], Liang Ma[2], Christopher Carignan[1]

[1]University College London, [2]Fudan University
yin.huang.21@ucl.ac.uk, liangma@fudan.edu.cn, c.carignan@ucl.ac.uk

ABSTRACT

Syllable-final nasals /n/ and /ŋ/ are reported to be neutralised in Mandarin varieties influenced by Wu dialects, among which is Hangzhou Mandarin. Previous articulatory research has shown supporting evidence for neutralisation, but only for single time points. In the current paper, we introduce a dynamic method of analysing ultrasound images, in order to determine whether neutralisation is present throughout the entire temporal interval of nasal coda production. This method quantitatively captures the movement of the tongue during the speech production process, which helps to examine the articulation distinction of different speakers. This study provides conclusive evidence that the place of articulation contrast between alveolar and velar coda nasals is neutralised in Hangzhou Mandarin.

Keywords: Ultrasound, nasal coda, merger, Mandarin varieties.

1. INTRODUCTION

Standard Mandarin allows alveolar nasal /n/ and velar nasal /ŋ/ in the syllable-final coda position. And the contrastive distribution of the /n/ and /ŋ/ carries the important function of distinguishing word meanings [7]. Standard Mandarin has 177 syllables with nasal codas, accounting for 44% of the total 405 regular syllables in Mandarin phonology. However, this contrast is observed to be reduced or neutralised in some regional varieties of Mandarin.

Wu dialects are spoken in southeastern China by more than 85 million people [20]. The phonological system of Wu dialects has vast differences from that of Standard Mandarin, one of which is that Wu dialects allow only a single nasal segment in codas [2,14]. This phoneme is usually transcribed as alveolar nasal /n/. Hangzhou is a major city in the Wu-speaking area, where the local dialect shares the common nasal coda limitation of Wu dialects [19]. Before the establishment of Standard Mandarin as a nationwide lingua franca, Wu dialects used to be the common language of this region. So Mandarin speakers in Wu-speaking areas show some difficulty distinguishing alveolar nasal /n/ and velar nasal /ŋ/, including Hangzhou Mandarin speakers.

From an acoustic perspective, the syllable-final nasals in Standard Mandarin are weakly produced with incomplete oral closure, meaning they have lower intensity than regular nasal consonants and more blurred vowel-nasal boundaries [4]. Thus, determining the place of articulation of the coda nasal by its acoustic pattern becomes rather complicated. As a result, most previous studies rely on perception methods to evaluate the neutralisation of nasal codas in accented Mandarin [9,12,17]. The results from these perceptual studies have suggested that in Mandarin varieties influenced by Wu dialects, the nasal coda /n/ and /ŋ/ are widely neutralised. The nasal coda merger is affected by some social factors, one of which is age, but the merger does exist in all age groups. The Wu dialect speakers tend to pronounce the velar nasal /ŋ/ as alveolar /n/, and their perceptual sensitivity to the /n/-/ŋ/ contrast is much lower than Standard Mandarin speakers.

The only articulation study of nasal coda merger in Wu-accented Mandarin is Faytak et al. [8]. They use a dimensionality reduction technique to process the midsagittal ultrasound data, comparing the tongue shape in the articulation of nasal coda produced by Shanghai Mandarin speakers and Standard Mandarin speakers. The results support the claim that the place contrast of nasal coda in Shanghai Mandarin is reduced in certain vowel environments. In this study, the midpoint of each nasal coda token was taken as a representative "magic moment" for determining tongue shape, thus disregarding possible dynamic factors involved in the neutralisation. However, given that these frames were chosen in an a priori manner, it is not certain whether the tongue shapes at these magic moments are indeed representative of the articulatory reality of the merger. This is due to the fact that, firstly, the vowel-nasal boundary of nasal coda syllables in Mandarin is hard to find, so the segmentation of nasal coda is of great difficulty. Secondly, the production movement of the coda nasal is too rapid for the tongue to achieve a stable position, meaning that a single midpoint frame may not be convincing enough to indicate the place of articulation of a specific nasal coda token.

Therefore, this paper attempts to propose a new approach of ultrasound data analysis that can dynamically track tongue movement to reflect the nasal coda's whole articulation process.

2. METHODS

2.1. Participants

Ultrasound data were collected from three Standard Mandarin speakers (2 female, 1 male, ages 21-23) and three Hangzhou Mandarin speakers (2 female, 1 male, ages 21-23). The control subjects were from Beijing and speak only Standard Mandarin without any dialects or varieties of Mandarin. Native Hangzhou Mandarin speakers were born and raised in Hangzhou city and had never left the Wu dialect-speaking region for more than one year. None of the participants reported any speech deficits or auditory impairments.

Vowel	/n/ coda				/ŋ/ coda			
	Monosyllabic		Disyllabic		Monosyllabic		Disyllabic	
/i/	因	**in**55	会心	xweɪ51.**ɕin**55	英	**iŋ**55	彗星	xweɪ51.**ɕiŋ**55
/ə/	根	**kən**55	出身	ʈʂʰu55.**ʂən**55	耕	**kəŋ**55	出生	ʈʂʰu55.**ʂəŋ**55
/ia/	烟	**ian**55	走线	tsoʊ214.**ɕian**51	央	**iaŋ**55	走向	tsoʊ214.**ɕiaŋ**51
/a/	安	**an**55	一般	i51.**pan**55	肮	**aŋ**35	一帮	i55.**paŋ**55

Table 1: Stimuli.

2.2. Materials

The stimuli consisted of 4 pairs of monosyllabic words and 4 pairs of disyllabic words (with target nasal coda in the second syllable), differing only in their nasal codas. The target syllables were in VN structure or started with a plosive or fricative consonant. To prevent the impact of assimilation, no other nasal sounds appeared in the target words (shown in Table 1).

A carrier sentence was used to present the stimuli in simplified Chinese characters without Pinyin or IPA annotation: __, 我知道怎么读__. ([uɔ ʈʂi.taʊ tsən.mɤ tu] “I know how to pronounce”). The printed-out reading material presented all these sentences in randomised order. For the sake of the participants’ comfort, the recording session of each speaker was controlled to around 40 minutes, which only allowed speakers to repeat the material twice. Therefore, there are 384 tokens in total (16 target words * 6 speakers * 2 sentence positions * 2 repetitions).

2.3. Procedures

The ultrasound experiment was carried out in a sound-attenuated booth of the Fudan University Phonetics Laboratory in Shanghai. Ultrasound images and audio were recorded simultaneously using the Articulate Assistant Advanced system [18]. A Telemed MicrUs portable ultrasound device outfitted with Telemed MC4-2R20S-3 convex probe was used to collect midsagittal ultrasound data with a 90° field of view at a rate of 90 frames per second. Articulate Instruments headsets were used for probe stabilisation. An AKG C544 microphone was mounted on the headset to the right of the speaker's mouth to record 16-bit audio at a sampling rate 44.1kHz.

Before the experiment started, each participant had 30 minutes to familiarise the reading material and adjust their sitting position. The reading list was printed on one page to be conveniently held by subjects. An operator controlled the ultrasound and audio recording.

2.4. Data analysis

The annotation of audio files was made by hand in Praat [1] and the segmentation information was saved to Textgrid scripts. Considering the difficulty of dividing nuclear vowels and the following nasal codas in Mandarin, the target segments were set to include both vowels and syllable-final nasals. The feasibility of this segmentation criterion lies in the fact that each pair of target syllables differ only by their coda nasals, so the distinction of tongue movement in the articulation only comes from the different places of articulation of their coda nasals.

Ultrasound recordings were cut into segments based on the corresponding Textgrid script. Then, these ultrasound video clips were converted to a sequence of continuous images.

For data processing, these ultrasound images were imported into Python [16] as 8-bit numeric matrices. All pixel coordinate information and intensity values from each image were collected for further analysis. The tongue surface in a midsagittal ultrasound image always appears as a bright curve, which is manifested as a group of pixels with greater intensity value than its surrounding in the matrices. Suppose we draw a one-pixel width line crossing roughly perpendicular to the tongue surface (as the white line in Fig.1b) and extract all the intensity values of pixels on the line. In that case, the peak of this group of values indicates where the tongue surface is. This line is referred to as the line of interest in the following.

The alveolar and velar nasals are articulated by lifting parts of the tongue body to touch the alveolar ridge or soft palate, respectively. The tongue undergoes an upwards movement. We mark the bottom point of the line of interest as point A (see Fig.1, b). In this paper, we refer to the distance from point A to the intersection point of the tongue surface and the line of interest (shown as the coloured points

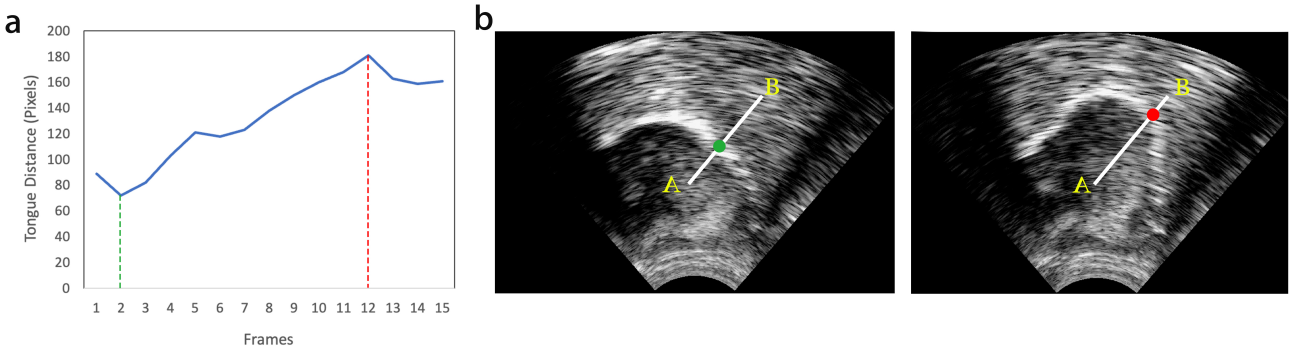


Figure 1: (a) The tongue distance varying over frames on a time series. (b) The line of interest for analysis is denoted between points A and B. The image on the left corresponds to the moment of minimum tongue distance. Likewise, the image on the right corresponds to the moment of maximum tongue distance. The tongue tip is on the left.

in Fig.1, b) as "tongue distance", which indicates the relative position of the tongue surface to point A. The greater the tongue distance, the further the tongue surface is to point A. Therefore, the change of tongue distance over time presents the tongue movement in the direction of the line of interest.

By calculating this tongue distance metric for each ultrasound frame in the target segment, we obtain a dynamic metric of tongue position that demonstrates movement over time (Fig.1, a), which we refer to as the tongue movement trajectory.

The line of interest should be set to explain specific research questions. In this experiment, we are interested in the tongue movement to the alveolar and velar nasal places of articulation. For every participant in this experiment, there are two lines of interest, pointing at the alveolar ridge and soft palate, respectively. Thus, for each participant, we generate two tongue movement trajectories: one for alveolar articulation and one for velar articulation. And to fully track the tongue movement, the line of interest should cover each speaker's whole range of articulation. Taking all the conditions into account, the length of all the lines of interest in this experiment is fixed at 200 pixels while the exact angle and location are adjusted to the physiological characteristics of individuals.

After measuring the movement trajectory data of every token, data cleansing is conducted by filtering out the outliers that have values lesser/greater than 1.5 times the interquartile range. Time normalisation is subsequently applied to all tokens by extracting the tongue distance at decile time points of the target segment. To remove the physical variation among all individuals, the tongue distance data are normalised using max-min normalisation. Each speaker's minimum and maximum is determined by the range of tongue movement throughout their entire data set.

For each speaker, local regression (LOESS smoothing) is performed on the normalised distance data grouped by alveolar and velar nasal coda to display the tongue movement when articulating these two nasal codas.

3. RESULTS

Fig. 2 shows the local regression results of the tongue movement trajectories in alveolar (a) and velar (b) places of articulation for every speaker. The colour-filling area represents the 95% Bayesian confidence interval.

As seen in Fig. 2, the alveolar and velar nasal coda tongue movement trajectories for Standard Mandarin speakers start from comparatively close proximity and go towards separation. And there is very little overlap between the alveolar and velar nasals trajectories, indicating that the tongue undergoes significantly different movements in articulating these two coda nasals. So in Standard Mandarin, the place of articulation of syllable-final /n/ and /ŋ/ are articulatorily distinct.

As a comparison, the two tongue movement trajectories of alveolar and velar coda nasals for Hangzhou Mandarin speakers overlap with each other throughout the articulatory segment, in both alveolar and velar places of articulation. It suggests that the Hangzhou Mandarin speakers display the same dynamic articulatory patterns for both syllable-final alveolar and velar nasals. There is no significant difference between the place of articulation of the alveolar and velar coda nasals in Hangzhou Mandarin, suggesting that the nasal coda is neutralised at both of these places of articulation and throughout the entire VN segments.

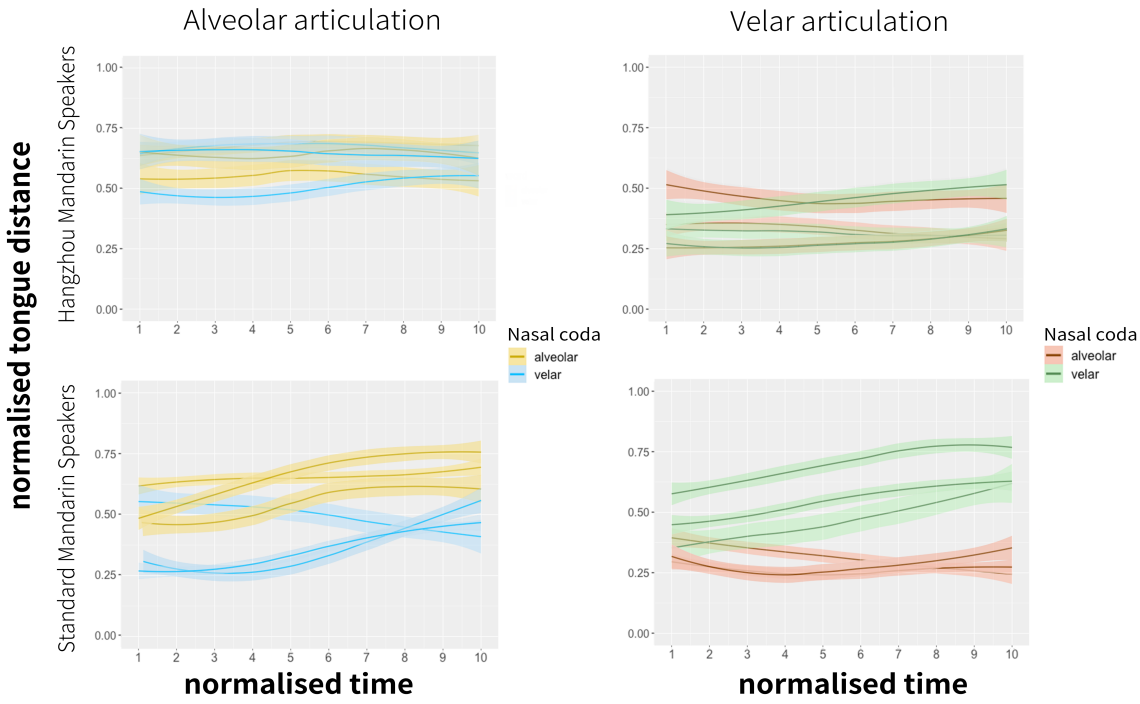


Figure 2: The regressed tongue movement trajectories of the nasal coda articulation for Standard Mandarin speakers (the bottom row) and Hangzhou Mandarin Speakers (the top row) in the place of articulation of alveolar (the left column) and velar (the right column). The x-axes are normalised time, and y-axes are normalised tongue distances.

4. CONCLUSION AND DISCUSSION

In conclusion, this study supports the existence of nasal coda mergers in Hangzhou Mandarin. Compared to other research, this dynamic study provides analysis reflecting tongue movement during the entire articulation process of the VN sequence, rather than the static tongue gesture on a single point.

However, the determination of the merger direction requires additional tongue contour analysis. Although this study does not address this issue, it can serve as a reference for future tongue contour analysis. The tongue movement trajectories provide an ideal method for determining representative frames for generating tongue contours; rather than choosing "magic moment" frames in an a priori fashion, they can be selected in a bottom-up data-driven manner. For example, in this case, to examine whether the tongue shapes of alveolar and velar nasals are merged in Hangzhou Mandarin, the most representative time point should be at the moment when the tongue movement trajectories of alveolar and velar nasals differ most. Images extracted from these time points could be used as representative frames in tongue contour analysis.

Besides, previous studies proposed that the proceeding vowel context may trigger this neutralisation [9,12,17,19]. This new approach allows us to explore this question further, and to give more precise descriptions of the articulation of nasal coda in various vowel contexts.

The method itself also has much room for improvement. For example, in this experiment, the determination of the tongue surface position along the line of interest is merely through finding out the maximum pixel values. However, this pixel value maximum can be influenced by image noise. Further research could improve the stability of automatically identifying the tongue surface locations by investigating alternative methods.

5. REFERENCES

[1] Boersma, P. 2002. Praat, a system for doing phonetics by computer. Glot International, 5.

[2] Chao. Y. (1928). *Studies in the Modern Wu-dialects*. Beijing: Commercial Press.

[3] Chen, C. (1991). The tonal change and diffusion: From the sampling of different age groups in Taipei. *Journal of Chinese Language Teachers Association*,16,69-99.

[4] Chen, M. Y. (2000). Acoustic analysis of simple vowels preceding a nasal in Standard Chinese. *Journal of Phonetics*, 28(1), 43-67.

[5] Chiu, C., Lu, Y., Weng, Y., Jin, S., Weng, W., & Yang, T. (2019). Uncovering syllable-final nasal merging in Taiwan Mandarin: Ultrasonographic investigation of tongue postures and degrees of nasalisation. In Proceedings of *the 19th International Congress of Phonetic Sciences* (pp. 398-402).

[6] Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, 120(1), 407-415.

[7] Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

[8] Faytak, M., Liu, S., & Sundara, M. (2020). Nasal coda neutralization in Shanghai Mandarin: Articulatory and perceptual evidence. *Laboratory Phonology*, 11(1).

[9] Guan, Y. (2019). Nasal coda realization in speech production of Shanghai Mandarin. In Proceedings of the *19th International Congress of Phonetic Sciences*, Melbourne (pp. 467–71).

[10] Hsu, H. (2007) Syllable-final nasal nergers in Taiwan Mandarin-leveled but puzzling. *Concentric: Studies in Linguistics*,33(1), 1–18.

[11] Li,M., Kambhamettu. C, & Stone.M. (2005). Automatic Contour Tracking in Ultrasound Images, *International Journal of Clinical Linguistics and Phonetics*, 19 (6-7), 545-554.

[12] Luo, M., Sloos, M., & van de Weijer, J. (2015). Perception and production of Mandarin nasal codas by Shanghainese speakers. Chinese Accents and Accented Chinese CAAC2014, 16.

[13] Mielke, J. (2015). An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *The Journal of the Acoustical Society of America*, 137(5), 2858-2869.

[14] Qian.N.(1992). *The Research of Contemporary Wu*. Shanghai: Shanghai Education Press

[15] Tse, J. P. (1992) Production and perception of syllable-final [n] and [ŋ] Mandarin Chinese: An experimental study. *Studies in English Literature and Linguistics*,18,143-156.

[16] Van Rossum, G. & Drake, F.L. (2009). Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

[17] Wang, L., Cui, J., & Chen, Y. (2018). Wuxi Speakers' Production and Perception of Coda Nasals in Mandarin. In INTERSPEECH (pp. 2559-2562).

[18] Wrench, A. 2008. *Articulate Assistant Advanced User Guide* version 2.05.

[19] You. R. (2011). Phonetic and Phonological Features of Hangzhou Dialect. *Bulletin of Chinese Linguistics*, 5(1), 129–144.

[20] Zhengzhang, Shangfang, and Wei Zheng. (2015). *Wu dialect*. Oxford: Oxford University Press.

Articulation of the nasal vowels in Shanghai Chinese

Changhe Chen, Jonathan Havenhill

Department of Linguistics, The University of Hong Kong
riverch8@connect.hku.hk, jhavenhill@hku.hk

ABSTRACT

This paper presents an articulatory and acoustic study of the nasal vowels [ãᵑ ɔ̃ᵑ] in Shanghainese. Synchronous ultrasound, audio, nasalance, and EGG recordings from 4 young native speakers were analyzed. Data show that the two nasal vowels in Shanghainese are merged and generally produced with a retracted and lower tongue position compared with the oral vowel [a]. The nasal vowels are significantly nasalized from about half of the duration. Unlike previous studies of Chaoshan Chinese, French, and Yi, a decrease in contact quotient during nasal vowels was not observed. Rather, one speaker is observed to increase the contact quotient (less breathy) for nasal vowels carrying modal rising tone [˧˥]. Given the acoustic similarities of nasality and breathiness, this adjustment may serve to avoid misperception with the breathy rising tone [˨˧]. This study betters our understanding of the nasal vowel articulation in Shanghainese and provides additional data required for cross-linguistic comparisons.

Keywords: nasal vowel, Shanghai Chinese, ultrasound, EGG, nasalance.

1. INTRODUCTION

Existing studies have shown that nasal vowels are both acoustically complex and multidimensional in articulation [1]. Compared to their oral counterparts, nasal vowels show additional nasal formants and anti-formants, broader F1 bandwidth, lower F1 amplitude, and higher spectral tilt (e.g., [2, 3]). These acoustic features can be attributed to velopharyngeal coupling which occurs as a result of lowering the velum. In addition to a lowered velum, nasal vowel production may also involve distinct labial and/or lingual gestures [1, 4-7], breathier voice [1, 8], and different pharyngeal configurations [5, 7]. Such complexity means that formant frequency analysis alone is less reliable for determining the underlying articulatory gestures, necessitating the use of tools like ultrasound tongue imaging to observe the gestures directly.

This study investigates the articulation of nasal vowels in Shanghai Chinese (Shanghainese) from the perspectives of lingual articulation, nasalance, and voice quality. Formant frequency data will also be reported. Shanghainese differs from Shanghai Mandarin, which is a Mandarin dialect spoken in Shanghai. Shanghainese and Shanghai Mandarin are not mutually intelligible. Studies on the sound system of Shanghainese (e.g., [9-15]) are abundant, but comprehensive articulatory and acoustic study of its nasal vowels remains lacking.

The Shanghainese low vowels, both oral and nasal, have been given a range of transcriptions, as shown in Table 1. In Chinese literature, the low central vowel [ä] is usually transcribed with the non-IPA symbol [A], reproduced here. Shanghainese has one low oral vowel, typically transcribed in open syllables as [a], [ɑ], or [A]. Some (but not all) studies report raising of this vowel to [ɐ] in checked syllables. While Shanghainese historically had two low nasal vowel phonemes (/ã/ and /ɔ̃/), recent work reports a merger of the two categories among young speakers.

	In CV	In CV?	In CṼ
Chao [9] (ca. 1910s)	[ɑ]	[A]	[ã], [ɔ̃]
Xu & Tang [10] (ca. 1930s-1950s)	[A]	[A]	[ã], [ɔ̃]
Xu & Tang [10] (ca. 1960s-later)	[A]	[ɐ]	[Ã]
Qian [11] (ca. 1980s)	[A]	[ɐ]	[Ãᵑ] or [Ã] or [ã ɔ̃]
Qian [11] (ca. 2000s)	[A]	[ɐ]	[Ã]
Zee [12] (ca. 1950s)	[a]	[a]	[aᵑ], [ɔᵑ] ([ãᵑ] young speakers)
Chen & Gussenhoven [13] (ca. 1950s-1960s)	[a]	[ɐ]	[ɐ], [ɑ] (nasalized before [ŋ])
Chen [14] (ca. 1970s-1980s)	[a]	[a]	[ã ɑ̃]
Chen [14] (ca. 1990s)	[a]	[ɐ]	[Ã]
Shen [15] (young speakers in 1981)	[a]	[a]	[ã], [ɔ̃]

Table 1: Low vowel transcriptions in previous studies, with approximate speaker birth years.

Previous work regarding the degree to which the nasal vowels are nasalized is insufficient. Only Fung & Lau [16] report nasal airflow data from one Shanghainese speaker and show that the two nasal vowels are fully nasalized throughout their duration.

Given the limited data available, this study will examine the time course of vowel nasalization through nasalance.

Some acoustic features of nasal vowels, such as broader F1 bandwidth and higher spectral tilt, resemble those of breathy voice [17]. Previous studies have found that nasal/nasalized vowels in Chaoshan Chinese [8], French [1], and Yi [18] can be produced with breathier voice, due either to misperception or acoustic enhancement [1, 18]. Interaction of voice quality and nasality is of particular interest for Shanghainese, given that tone [ɻ] is produced with breathy (whispery) voice [19]. Although some studies report the loss of breathy voice on [ɻ] among young speakers [20], this study will also investigate the potential interaction between vowel nasality and breathy voice in syllables bearing the tone [ɻ]. Acoustic measures typically used to detect breathier voicing, e.g., H1-H2 and H1-A3, are unreliable with the presence of nasality, and thus voice quality will be analyzed through electroglottography.

Through a combination of synchronous acoustic, ultrasound, electroglottographic, and nasalance data, this study will investigate:

1. The tongue positions and acoustic features of the Shanghainese vowels [ɐ, ãᵑ, õᵑ] in comparison with [a].
2. The time course of nasalization for the nasal vowels, to determine whether the vowels are fully or partially nasalized.
3. Whether speakers of Shanghainese also use lower contact quotient (breathier voice) in producing the nasal vowels, and how the breathy voice of tone [ɻ] interacts with nasality.

2. METHODS

2.1. Speakers

Data from 4 native speakers of Shanghainese (2 men, 2 women) were analyzed. They are university students in their early twenties and lived in Shanghai prior to age 18. Urban Shanghainese is the primary language spoken in their families, although they are also proficient in Mandarin and English. None of the speakers reported speech or hearing disorders.

2.2. Materials

The word list contained 17 (near) minimal sets (68 words plus 5 additional (C)Vʔ syllables) of the low vowels in (C)(w)V, (C)Vʔ and (C)(w)Ṽ syllables carrying all the Shanghainese tones /˨ ˧ ɻ ˦ ˥/. Tones /˨ ˧ ɻ/ are carried by (C)V, (C)Ṽ and (C)Vŋ syllables, while tones /˦ ˥/ are only carried by (C)Vʔ syllables. The word list also contained 57 fillers with rimes [i u o ɪŋ ʊŋ əŋ]. Sample sets are provided in Table 2.

CV	CVʔ	CṼ
[ka˨] 街 “road”	[kɐʔ˥] 夾 “clip”	[kãᵑ˨] 羹 “thick soup” [kõᵑ˨] 江 “large river”
[tsa˧] 債 “debt”	[tsɐʔ˥] 著 “to wear”	[tsãᵑ˧] 帳 “accounts” [tsõᵑ˧] 壯 “chubby”
[aɻ] 鞋 “shoes”	[ɐʔ˦] 匣 “box”	[ãᵑɻ] 杏 “apricot” [õᵑɻ] 項 “item”

Table 2: Sample (near) minimal sets in the word list.

2.3. Data collection

We use a similar equipment setup to that of Carignan [1]. Synchronous ultrasound, audio, lip video, electroglottograph (EGG) and nasalance recordings were made in a sound-attenuated booth at the University of Hong Kong.

Ultrasound images of the tongue were recorded at a typical rate of 84 frames per second using a SonoSpeech Micro ultrasound system. Participants wore a headset [21] to prevent movement of the cameras and ultrasound transducer. Audio recordings were captured using an Earthworks Ethos supercardioid condenser microphone and a Sound Devices USBPre2 preamplifer. Audio was recorded to a Denon F650R solid state recorder at a 44.1 kHz sampling rate and 16 bit sampling depth. Ultrasound, audio, and lip video recordings were synchronized in Articulate Assistant Advanced (AAA) [22]. EGG recordings were made using a Voce Vista electroglottograph (model: 7050A). Nasalance was recorded using the Glottal Enterprises Nasalance Separator Handle. This system comprises a plate held against the participant’s upper lip to separate the oral and nasal audio signals, rather than a Rothenberg mask which seals the nose and mouth, so high fidelity audio recordings can be made at the same time. Output from the EGG and Nasalance meter, as well as the speech and synchronization audio signals from AAA, were recorded simultaneously on an iMac using Audio Hijack.

Syllables written in simplified Chinese characters were presented in AAA in a unique pseudorandom order for each participant. Participants were instructed to read four successive repetitions of each syllable in isolation with a short pause between repetitions.

2.3. Data analysis

Tongue contours and formant frequencies were extracted at a single point during the steady state portion of the vowel. Tongue contours were tracked in AAA using a pre-trained MobileNet1.0-based neural network in DeepLabCut [23]. Data were analyzed with polar SSANOVA [24].

The nasalance recordings provide two audio channels, including both a nasal channel (A_N) and an oral channel (A_O). To see the time course of vowel nasalization, nasal consonants following the nasal vowels were excluded when present. RMS amplitude values were measured at 20% intervals during the vowel for both channels (A_O and A_N). Nasalance was calculated using the formula $A_N / (A_O + A_N)$, where a higher value indicates a higher degree of nasalization.

For the EGG data, since lower contact quotient (CQ) is related to higher spectral tilt [25], a shared acoustic feature of nasality and breathy voice, CQ was measured in nasal vowels; a lower CQ value indicates breathier voice. The data were obtained using EggWorks [26] (hybrid method [27]). Data within each fifth were averaged in VoiceSauce [28]. Each tone was analyzed separately, given that tones may be produced with different voice qualities. Tone [˩] is produced with glottalization at the end, making it difficult to obtain reliable CQ values, so this tone was excluded. Both nasalance and CQ values were analyzed with SSANOVA. Shading around the splines indicates 95% confidence interval and overlap of the shading suggests no significant difference.

Some data were excluded due to mispronunciation or clipping of the EGG and/or nasalance recordings. The numbers of tokens included are shown in Table 3 (following references).

3. RESULTS

This section presents tongue contours and formant frequencies of the low vowels, as well as nasalance and contact quotient data for the nasal vowels. Although it has been reported that the two nasal vowels [ãᵑ õᵑ] have merged for young speakers [10, 11, 12, 14], we report results for each vowel separately, given that the merger is relatively recent and previous studies are mainly impressionistic.

SSANOVA analysis of tongue position (Figure 1) reveals that, for SH01F, SH01M and SH02F, the oral vowel [a] is higher than [ɐ ãᵑ õᵑ]. For all speakers, the nasal vowels are more retracted than the oral vowels.

Acoustic data reveal that the F1 of [ɐ] tends to be lower than that of [a], although the ellipses for the two vowels still overlap considerably, particularly for SH02F. Consistent with the tongue contours, [ãᵑ õᵑ] tend to have a lower F2 than the oral vowels, with the exception of SH02F. Although the nasal vowels show a lower tongue position than the oral vowels, the F1 of the two nasal vowels usually does not exceed that of [a], as would be predicted by the difference in tongue position. This effect can be attributed to the lowering effect that nasalization has on F1 [29]. In accordance with previous work, this study confirms that the two nasal vowels have merged in production.

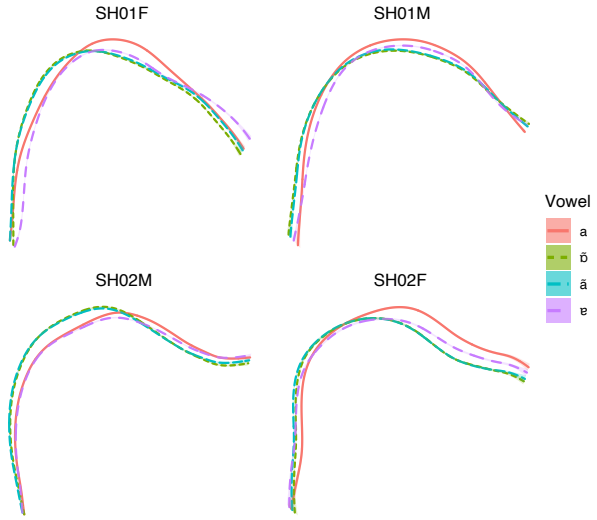


Figure 1: Polar SSANOVA tongue contours. The tongue root is to the left, shading represents 95% CI.

Nasalance data (Figure 3) reveal an increasing degree of nasality throughout the vowel duration. The nasal vowels become significantly nasalized after one fifth (SH02F), or two fifths (SH01M and SH02M) or three fifths (SH01F) of the vowels. This result differs from previous work on French, which shows a high degree of nasalization from the vowel onset [1].

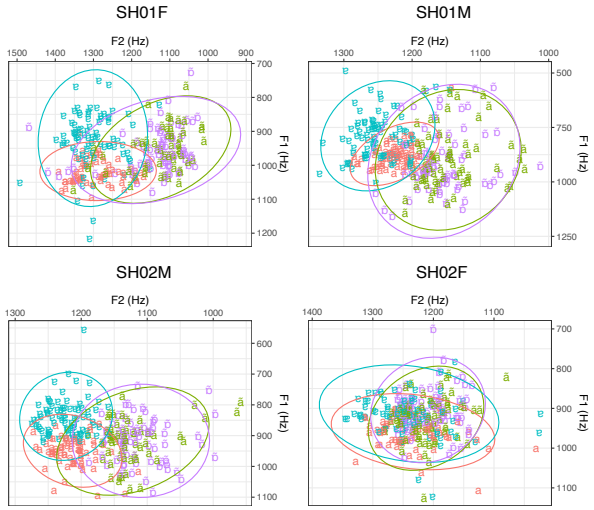


Figure 2: F1 and F2 values of the low vowels.

CQ measurements (Figure 4 and 5) indicate that none of the speakers used a lower CQ (breathier voice) to produce the nasal vowels. On the contrary, SH01M

shows higher CQ values (less breathy) for the two nasal vowels carrying tone [˧] (Figure 4).

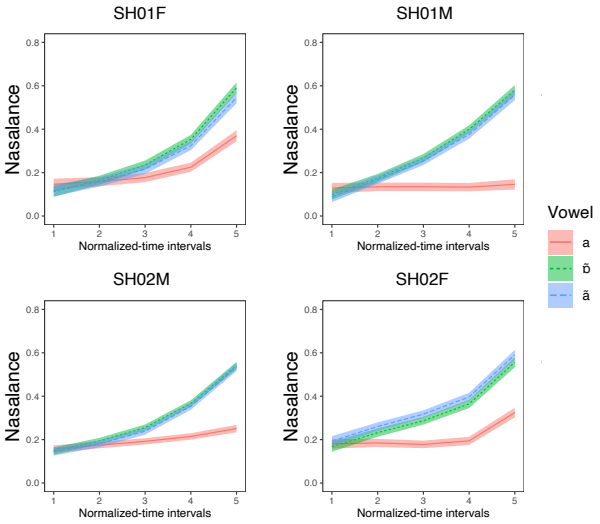


Figure 3: SSANOVAs for nasalance measurements.

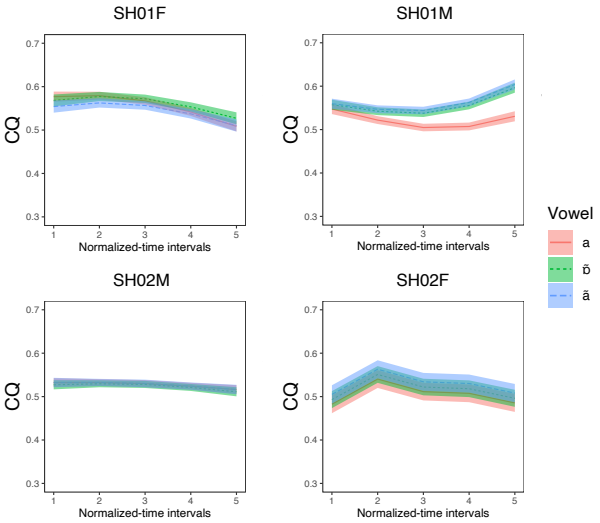


Figure 4: SSANOVAs for CQ values of tone [˧].

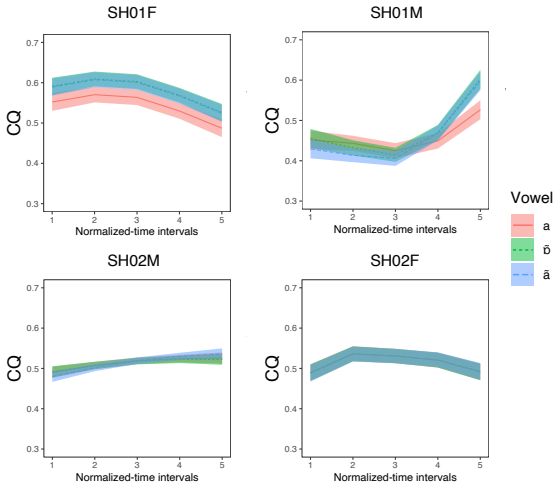


Figure 5: SSANOVAs for CQ values of tone [˧˥].

4. DISCUSSION AND CONCLUSIONS

Ultrasound data show that the nasal vowel is produced with a lower and more retracted tongue position than its oral counterpart [a], and can be transcribed as [ã̠ᵑ]. Retracted tongue position is believed to enhance the acoustic effects of nasalization in French [4]. However, it is not yet clear whether a similar interpretation can also be applied to Shanghainese given that the remaining nasal vowel is the result of merger between [ãᵑ] and [ɔ̃ᵑ] and may partially retain the tongue position used for [ɔ̃ᵑ]. Moreover, the weak coda nasal [ᵑ] may exert coarticulatory influence on the preceding vowel. These possibilities may be confirmed through future studies which more closely examine interspeaker and/or apparent-time variability in terms of the presence/absence of the nasal coda as well as merger/non-merger of the two nasal vowels.

With respect to the vowel height, the lower tongue position used by three speakers counteracts the F1 lowering effect of nasalization on low vowels [29], such that the oral and nasal vowels show similar F1 values. The tongue position used to produce the nasal vowels may therefore also reflect compensation for the F1 lowering effect of nasalization, thereby maintaining an auditorily low quality. While the vowel [ɐ] is produced with a lower tongue height when compared to [a], it tends to have lower F1. This observation suggests that transcription of this vowel as [ɐ] may better reflect the auditory quality of the vowel height rather than the actual tongue height.

Nasalance data show that the two nasal vowels are produced with similar degrees of nasalization, further supporting the conclusion that the two nasal vowels have merged. Unlike what has been observed for nasal vowels in French [1], the nasal vowel of Shanghainese shows only partial nasalization.

One speaker used higher CQ to produce the nasal vowels with tone [˧]. Gao et al. [30] showed that breathy voice can bias the tone perception towards tone [˧˩] in Shanghainese. The use of higher CQ may serve to prevent confusion of the modal rising tone [˧] with the breathy rising tone [˧˩], given the acoustic similarities between the nasality and breathy voice. Perceptual data are needed to verify this hypothesis.

Through the use of multimodal articulatory data, this study shows the multidimensionality of oral-nasal vowel contrast as well as the language- and speaker-specific features of nasal vowel production.

5. REFERENCES

[1] C. Carignan, “Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels,” *Journal of Phonetics*, vol. 63, pp. 87–105, Jul. 2017.

[2] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels," The Journal of the Acoustical Society of America, vol. 77, no. 4, pp. 1560–1575, Apr. 1985.
[3] W. Styler, "On the acoustical features of vowel nasality in English and French," The Journal of the Acoustical Society of America, vol. 142, no. 4, pp. 2469–2482, Oct. 2017.
[4] C. Carignan, "An acoustic and articulatory examination of the 'oral' in 'nasal': The oral articulations of French nasal vowels are not arbitrary," *Journal of Phonetics*, vol. 46, pp. 23–33, Sep. 2014.
[5] C. Carignan, R. K. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, "A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French," *Journal of Phonetics*, vol. 50, pp. 34–51, May 2015.
[6] R. Shosted, C. Carignan, and P. Rong, "Managing the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 455–465, Jan. 2012.
[7] M. Barlaz, R. Shosted, M. Fu, and B. Sutton, "Oropharygneal articulation of phonemic and phonetic nasalization in Brazilian Portuguese," *Journal of Phonetics*, vol. 71, pp. 81–97, Nov. 2018.
[8] C. Chen, "Voice Quality of the Nasal Vowels in Chaoshan Chinese," in *SST 2022*, Canberra, Dec. 2022, pp. 76-80.
[9] Y. R. Chao, 現代吳語的研究 [*Studies of the modern Wu dialects*]. Beijing: Tsing Hua College Research Institute, 1928.
[10] B. Xu and Z. Tang, 上海市區方言志 [*A description of urban Shanghainese*]. Shanghai: Shanghai Education Press, 1988.
[11] N. Qian, 上海語言發展史 [*The history and development of Shanghainese*]. Shanghai: Shanghai People's Publishing House, 2003.
[12] E. Zee, "Shanghai Phonology", in *The Sino-Tibetan Languages*, Routledge, pp. 185-192, 2017.
[13] Y. Chen and C. Gussenhoven, "Shanghai Chinese," *Journal of the International Phonetic Association*, vol. 45, no. 3, pp. 321–337, Dec. 2015.
[14] Z. Chen, "開埠以來上海城市方言語音演變 [*The historical developments of sound system since the beginning of Concession in the Shanghai urban dialect*]", Bulletin of Linguistic Studies, vol. 24, pp. 280-313.
[15] T. Shen, "上海話老派新派的差別 [The linguistic differences between the old and young speakers of Shanghainese]", Dialect, vol. 4, pp. 275-283, 1981.
[16] W. N. Fung and S. L. Lau, "A Physiological Analysis of Nasals and Nasalization in Chinese," in *Eurospeech 2001*, Aalborg, Sep. 2001, pp. 325–328.
[17] T. Arai, "Cue parsing between nasality and breathiness in speech perception," Acoust. Sci. & Tech., vol. 27, no. 5, pp. 298–301, 2006.
[18] M. Garellek, A. Ritchart, and J. Kuang, "Breathy voice during nasality: A cross-linguistic study," *Journal of Phonetics*, vol. 59, pp. 110–121, Nov. 2016.
[19] J. Tian and J. Kuang, "The phonetic properties of the non-modal phonation in Shanghainese," *Journal of the International Phonetic Association*, pp. 1–27, Jul. 2019.
[20] J. Tian and J. Kuang, "Revisiting the register contrast in Shanghai Chinese," in *Proceedings of Tonal Aspects of Languages 2016*, Buffalo, NY, 2016, pp. 147–151.
[21] Articulate Instruments Ltd, *Ultrasound Stabilisation Headset User's Manual: Revision 1.4*. Edinburgh, UK: Articulate Instruments Ltd, 2008.
[22] Articulate Instruments Ltd, *Articulate Assistant Advanced User's Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd, 2012.
[23] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut," *Sensors*, vol. 22, 2022.
[24] J. Mielke, "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2858–2869, May 2015.
[25] E. Holmberg, R. Hillman, J. Perkell, P. Guiod, & S. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice", *Journal of Speech and Hearing Research*, vol. 38, pp. 1212-1223, 1995.
[26] H. Tehrani, "EGGWorks: A Program for Automated Analysis of EGG Signals", available online: https://appsobabble.com/functions/EGGWorks.aspx.
[27] D. M. Howard, "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers", *Journal of Voice*, vol. 9, pp. 163–72, 1995.
[28] Y. Shue, P. Keating, C. Vicenik and K. Yu, "VoiceSauce: a program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 1846–1849.
[29] C. Carignan, "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2588–2601, May 2018.
[30] J. Gao, P. Hallé, and C. Draxler, "Breathy voice and low-register: A case of trading relation in Shanghai Chinese tone perception?" Language and Speech, vol. 63, no. 3, pp. 582–607, Sep. 2020.

		Ultrasound	Nasalance	EGG
01F	[a]	64	56	44
	[ɐ]	80	-	-
	[ãʱ]	56	48	35
	[ɒ̃ʱ]	72	60	37
01M	[a]	67	44	44
	[ɐ]	87	-	-
	[ãʱ]	68	43	44
	[ɒ̃ʱ]	68	45	44
02M	[a]	68	60	36
	[ɐ]	84	-	-
	[ãʱ]	56	60	32
	[ɒ̃ʱ]	68	44	32
02F	[a]	68	60	26
	[ɐ]	80	-	-
	[ãʱ]	68	60	30
	[ɒ̃ʱ]	68	48	33

Table 3: The numbers of tokens included.

LABIALIZATION IN AMAZIGH: ACOUSTIC AND ARTICULATORY MARKING OVER TIME

Philipp Buech, Anne Hermes, Rachid Ridouane

Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/Sorbonne Nouvelle
philipp.buech@sorbonne-nouvelle.fr, anne.hermes@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

ABSTRACT

Labialization is the most widely distributed secondary articulation in the world's languages, however, literature on its phonetic properties is rare. This study provides an analysis of acoustic and articulatory markings of labialization in Tashlhiyt Amazigh. We present speech production data from six native speakers producing a set of plain and labialized dorsal consonants in [aCa] sequences. Formant trajectories of the vowels surrounding the consonants are analyzed as well as articulatory trajectories of the lip aperture and the tongue during the consonantal movements. In the acoustic domain, the Bayesian spline regression analysis reveals a lowering of F2 in the vowels adjacent to labialized consonants. In the articulatory domain, lip protrusion and tongue retraction are the primary articulatory correlates for labialization. Taken together these results suggest that labialization in Tashlhiyt is an instance of labio-velarization.

Keywords: Secondary articulation, labialization, acoustics, articulation, Amazigh

1. INTRODUCTION

Labialization is a secondary articulation that is mainly produced by a protrusion of the lips [1]. It is the most common secondary articulation in the languages of the world and patterns, most often with velar and uvular stops and fricatives [2]. Labialization has been addressed in the phonological literature to some extent, but detailed phonetic investigations of labialization are quite rare. In this study, we provide an analysis of labialization in Tashlhiyt, a language where the entire set of its dorsal stops and fricatives /g, gː, k, kː, q, qː, ʁ, ʁː, χ, χː/ has labialized counterparts /g^w, g^wː, k^w, k^wː, q^w, q^wː, $ʁ^w$, $ʁ^w$ː, $χ^w$, $χ^w$ː/.

1.1. Tashlhiyt Amazigh

Tashlhiyt is one of the three Amazigh varieties spoken in Morocco [3]. Its linguistic system is well-described, including its phonetics and phonology (for a review, see [4, 5]). Labialization is a lexical property of dorsals in Tashlhiyt (e.g., /ik^{w}la/ 'he was colored' vs /ikla/ 'he spent the day', /$ʁ^w$i/ 'here' vs /$ʁ^w$i/ 'catch'). The phonology of labialization in Tashlhiyt has received much scholarly attention [6, 7, 8, 9], but data on the phonetics of the plain/labialized distinction is lacking. The aim of this study is thus to fill this gap, by providing a first analysis of acoustic and articulatory correlates of labialization over the time course of [aCa] sequences in Tashlhiyt.

1.2. The phonetics of labialization

Labialization is mainly produced by the addition of a lip protrusion gesture to a primary gesture [10]. It can be subcategorized into (a) simple labialization that involves an additional lip rounding gesture, and (b) a more complex and more common labio-velarization, produced by the lip protrusion plus a raising of the tongue dorsum [1]. Further, [1] mentions a timing asymmetry, where the location of the labialization gesture concentrates at the release of the primary gesture. For Moroccan Arabic, where labialization is considered a substratum from Amazigh [11, 12], articulatory data of plain and labialized dorsals showed a higher and more protruded lip movement as well as a backward movement of the tongue [13, 14].

Regarding acoustics, the most important information is reported to be on the surrounding vowels, as is the case for secondary articulations in general [1]. Specifically for labialization, a lowered F2 has been reported [13, 14, 15, 16, 17]. Further, [15] reports a lowering of the first formant in Nootka, while [13, 14] observed no significant changes in F1 in general comparing plain and labialized dorsals in Moroccan Arabic.

2. METHODS

2.1. Recordings and speech material

Acoustic and articulatory data were recorded from six native male speakers of Tashlhiyt (32-50 years,

mean=44, sd=7) using the EMA AG 501 (Carstens Medizinelektronik GmbH). Sensors were placed on the upper and lower lips (ULIP, LLIP), the tongue tip (TTIP), the tongue mid (TMID) and the tongue body (TBO). Sensor positions were recorded with a sample rate of 1250 Hz and were filtered afterwards with a Butterworth low pass filter (cutoff frequency: 20 Hz, order 5). The acoustic signal was recorded with a head-mounted microphone at a sample of 44.1 kHz and 16 bit resolution.

The speech material presented here consists of [aCa] sequences that contained plain and labialized consonants, spoken in isolation. These sequences represent the pronunciation of indiviual letters in Tifinagh, the Amazigh alphabet. For instance, the letter <ⵣ>, which corresponds to the Latin letter <z>, is pronounced as [aza]. Consonants differ in place/manner, voicing and length (see Table 1). The stimuli were presented on a computer screen placed in front of the speakers. 271 tokens in total went to the analysis.

	length	
place	singleton	geminate
velar	g, k, g^w, k^w	gː, kː, g^wː, k^wː
uvular	ʁ, χ, $ʁ^w$, $χ^w$	ʁː, χː, $ʁ^w$ː, $χ^w$ː

Table 1: Plain and labialized target consonants.

2.2. Measurements

The acoustic boundaries of the vowels and target consonants were labeled in Praat v6.3.02 [18]. Formant frequencies were measured in steps of 10% across the preceding and following V. In the articulatory domain, the sensor positions of the upper and lower lips (ULIP, LLIP), the tongue tip (TTIP), the tongue mid (TMID) and the tongue body (TBO) in the front-back and low-high dimensions were measured in 10% steps from the beginning to the end of each segment in the [aCa] sequences. Lip aperture (LA) was calculated as the Euclidean distance between ULIP and LLIP. Acoustic measurements were done using Parselmouth v0.4.1 [19], which includes Praat v6.1.38, and a custom script written in Python v3.10.6 [20] was used for the measurements of the sensor positions.

2.3. Statistical analysis

Bayesian spline regression analysis was performed for the acoustic and articulatory trajectories using PYMC v.5.0.1 [21]. Separate models were fitted for the first three formants in the preceding and following vowels and for LA and TTIP, TMID and TBO sensors in the front-back and low-high dimensions. The model structure was the same for all trajectories: TYPE (plain, labialized), PLACE (velar, uvular), VOICE (voiced, voiceless), LENGTH (singleton, geminate) and the interactions of TYPE with PLACE, VOICE and LENGTH were population level effects and an additional spline term for TYPE and Time. SPEAKER was added as random intercept. Weakly informative priors were used for each model. Sampling consisted of four chains with 4000 samples for tuning and 400 samples for draws, resulting in a total of 16000 samples. The HDI+ROPE decision rule [22] was used to reject the null hypotheses. ROPEs were defined as the interval $\pm$0.1*SD for each response variable. Evidence for an effect was determined, if 95% Credible Interval (CI) of the posterior distribution was outside the ROPE.

3. RESULTS

3.1. Formant trajectories

The mean formant trajectories are shown in Fig. 1 (plain as solid lines and labialized as dashed lines). The trajectories of F1 in the labialized productions are slightly lower in both vowels. This effect seems to be stronger in the following V and stronger for uvular Cs. For both vowel positions (ROPEs: prec. V: $\pm$13, foll. V: $\pm$12), the regression results revealed evidence for an effect of PLACE (prec. V: β=135 [123,147], foll. V: β=71 [61, 81]), and the interaction TYPE and PLACE (prec. V: β=-32 [-50, -15], foll. V: β=-33 [-45, -18]) on F1. Only in the following V we found an influence of TYPE (β=-34 [-49, -20]) and VOICE (β=28 [18, 38]).

The trajectories of F2 show a clear lowering in the labialized contexts. This difference appears to be strongest in the first portion of the following V across place, voicing and length and in particular in Vs surrounding velars. Regression results for both Vs (ROPEs both $\pm$27) revealed evidence for F2 being influenced by TYPE (prec. V: β=-259 [-287, -231], foll V: β=-360 [-389, -333]), PLACE (prec. V: β=-273 [-293, -253], foll V: β=-232 [-251, -213]) and the interaction of TYPE and PLACE (prec. V: β=187 [157, 216], foll. V: β=185 [156, 213]).

The trajectories of F3 show very small and less clear differences. We found evidence for an influence of PLACE in both Vs (prec. V: β=191 [173, 209], foll V: β=182 [166, 198]), but in the following V, we observed also an effect of TYPE (β=-63 [-86, -40]) (ROPEs both $\pm$30).

Differences between the plain and labialized

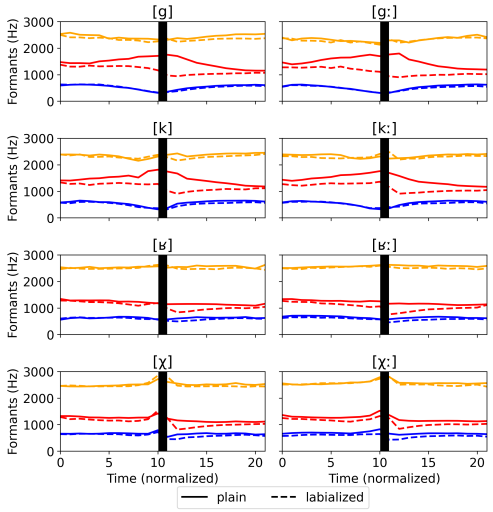


Figure 1: Averaged formant trajectories (F1: blue, F2: red, F3: orange) in surrounding Vs for singletons (left) and gemiantes (right). Black bars mark the C boundary.

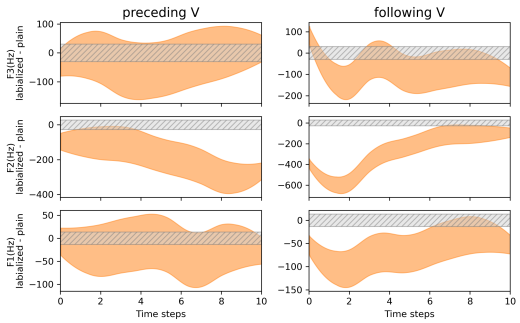


Figure 2: Labialized - plain differences (95% CI of the posterior) by formants and positions. Grey hatched areas indicate the respective ROPEs.

formant trajectories are depicted in Fig. 2. A short portion of F3 at the beginning of the following V and the first half of the F1 trajectory in the following V were outside the respective ROPEs. Strong evidence for a difference between plain and labialized productions in F2 was found for the most part of the surrounding vowels, with the highest difference near the beginning of the following V.

3.2. Articulatory trajectories

The trajectories for the LA are depicted in Fig. 3 and the trajectories for the tongue sensors are shown in Fig. 5. The time points 10 and 20 correspond to the acoustic C segment. Regarding LA, lips were more closed for labialized Cs than for plain Cs. For the labialized productions, the trajectories lowered continuously towards the C's mid as the region of its minimal aperture and raised to the end of the following V. Regression results showed evidence for TYPE (β=-3.2 [-3.5, -2.9]), PLACE (β=1.2 [0.96, 1.4]) and their interaction (β=-1.1 [-1.4, -0.8]) (ROPE $\pm$0.65).

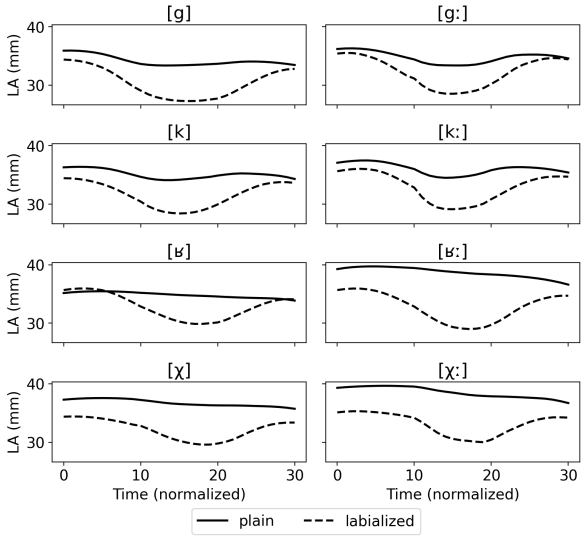


Figure 3: Averaged LA trajectories (plain: solid, labialized: dashed) by consonant (rows) and length (columns).

The difference for LA between labialized and plain trajectories is presented in Fig. 4. The largest portion of the 95% CI of the posterior is outside the ROPE, indicating evidence for a lowered LA in the central region. The highest difference can be found in the mid of C movement with a posterior mean of -6 mm.

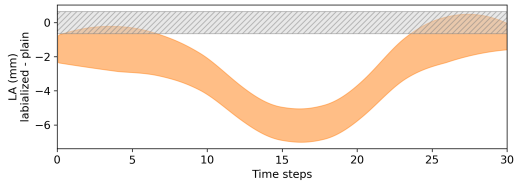


Figure 4: Labialized - plain difference (95% CI of the posterior) for the LA trajectory. The grey hatched area indicates the ROPE.

Coming to the tongue movement, it can be observed from Fig. 5 that the trajectories of labialized productions display a backward movement across place, voicing and length, with the exception of [χ:], where a slight fronting can be observed. The tongue in the low-high dimension may be lowered for labialized C (e.g., [g^w:, k^w]) or (slightly) raised (e.g., [g^w, χ^w]).

In the front-back dimension, ROPEs were set to

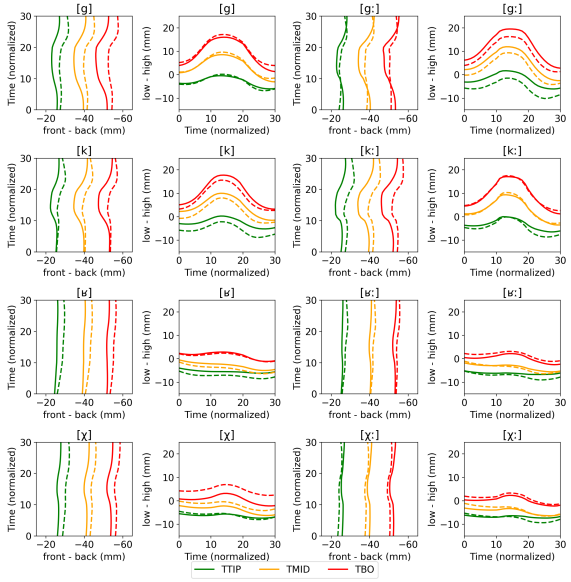


Figure 5: Averaged tongue sensor trajectories of plain (solid) and labialized (plain) productions for each C (rows) and length (singleton: first two columns, geminates: last two columns).

±0.68 for TTIP, ±0.62 for TMID and ±0.71 for TBO. We found evidence for an influence of TYPE on all tongue sensors (TTIP: β=-2.9 [-3.2, -2.7], TMID: β=-3 [-3.2, -2.8], TBO: β=-2.9 [-3.1, -2.6]) and PLACE (TTIP: β=-1.3 [-1.5, -1.2], TMID: β=-2.6 [-2.7, -2.4],TBO: β=-2.8 [-3, -2.7]), and their interaction (TTIP: β=1.2 [0.99, 1.4], TMID: β=1.6 [1.4, 1.8], TBO: β=1.9 [1.7, 2.2]). Regarding the low-high dimension, the regression models showed evidence for an effect of PLACE in every sensor (TTIP: β=-2.7 [-2.9, -2.5]; ROPE ±0.5; TMID: β=-6.5 [-6.7, -6.2], ROPE ±0.74; TBO: β=-8.2 [-8.5, -8,0], ROPE ±0.89). An influence of TYPE was observed for the TTIP only (β=-1.8 [-2.1, -1.6]).

Fig. 6 depicts the differences between the labialized and plain trajectories. The majority of the differences in the front-back dimension of TTIP, TMID and TBO fell outside the ROPEs, indicating a backing of the tongue for labialized productions. The points with the highest differences were towards the acoustic end of the consonant (around time step 20). In the low-high dimension, only the second half of the TTIP was outside the ROPE, indicating evidence for a lowering of the tongue tip in the second half of the C movement.

4. DISCUSSION AND CONCLUSION

This study investigated the acoustic and articulatory marking of labialization over the time course of Tashlhiyt [aCa] sequences.

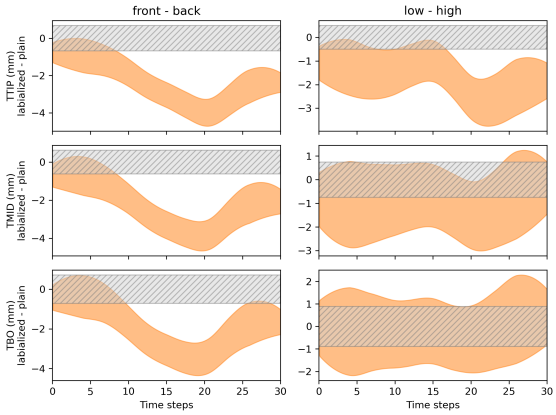


Figure 6: Labialized - plain differences (95% CI of the posterior) by tongue sensors and dimensions. Grey hatched areas indicate the respective ROPEs.

The acoustic results showed a lowering of F2 as the primary acoustic correlate for labialization, similar to the findings in [13, 14] for Moroccan Arabic. The highest difference in F2 was found at the beginning of the following V. However, we also observed a slight lowering of F1 in the following V. Regarding articulation, we found a lip protrusion with the strongest closure of the lips near the mid of the C movement for labialized productions. In addition, a retraction of the tongue during the labialized Cs was observed and this difference was strongest near the acoustic end of the Cs. We also found a slight raising of TTIP in the second half of the C, which may be a result of more flexibility for this articulator since the primary articulator for dorsals is the tongue dorsum.

As the data presented here showed strong evidence for lip protrusion plus a retraction of the tongue, labialization in Tashlhiyt can be considered as labio-velarization [1, 13]. This combination of lip aperture and tongue retraction is most probably what triggers the important lowering of F2 for labialized Cs: Following [23], F2 lowering can be attributed to a backward movement of the tongue and also lip protrusion together with a dorsal constriction.

This preliminary study is a small part of a larger corpus of research. In the future, the study will be expanded to include more items, different word positions and a wider range of vowels. The aim is to provide a more comprehensive understanding of the phonetics of labialization in Tashlhiyt.

5. ACKNOWLEDGEMENTS

This study was supported by the French ANR (ANR-10-LABX-0083, LABEX-EFL).

6. REFERENCES

[1] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Blackwell Publishers, 1996.

[2] I. Maddieson, *Patterns of Sounds*. Cambridge University Press, 1984.

[3] S. Chaker and A. Mettouchi, "Berber," in *Concise Encylopedia of Languages of the World*, K. Brown and S. Ogilvie, Eds. Elsevier, 2009, pp. 152–158.

[4] R. Ridouane, "Leading Issues in Tashlhiyt Phonology," *Language and Linguistics Compass*, vol. 10, pp. 644–660, 2016.

[5] ——, "Tashlhiyt Berber," *Journal of the International Phonetic Association*, vol. 44, no. 2, pp. 207–221, 2014.

[6] N. G. Clements, "lace of articulation in consonants and vowels: A unified theory," *Working Papers of the Cornell Phonetics Laboratory*, vol. 5, pp. 77–123, 1991.

[7] E. Selkirk, "Labial relations," Master's thesis, University of Massachusettes, Amherst, 1993.

[8] J. Alderete and S. Frisch, "issimilation in grammar and the lexicon," in *The Cambridge Handbook of Phonology*, P. de Lacy, Ed. Cambridge University Press, 2007, pp. 379–398.

[9] K. Bensoukas, *Featural dissimaltion in Tashlhit. Avoiding the Repetition of Labial and Round*, ser. Theses and Memoirs. University of Mohammed V, Publications of the Faculty of Letters and Human Sciences, 2014, no. 72.

[10] R. L. Trask, *A Dictionary of Phonetics and Phonology*. Routledge, 1996.

[11] M. Elmedlaoui, "Le substrat berbère en arabe marocain : Un système de contraintes," *Langues et Littératures*, vol. 16, pp. 137–165, 1998.

[12] ——, "L'arabe marocain: Un lexique sémitique inséré sur un fond grammatical berbère," in *Etudes Berbères et Chamito-Sémitiques. Mélanges Offerts à Karl-G. Prasse*. Peeters, 2000, pp. 155–187.

[13] C. Zeroual, J. H. Esling, and P. Hoole, "Ema, endoscopic, ultrasound and acoustic study of two secondary articulations in moroccan arabic. labial-velarisation vs. emphasis," in *Instrumental Studies in Arabic Phonetics*. John Benjamins, 2011, pp. 277–297.

[14] C. Zeroual, J. H. Esling, P. Hoole, and R. Ridouane, "Ultrasound study of moroccan arabic labiovelarization," in *Proc. 17th ICPhS*, 2011, pp. 2272–2275.

[15] S. Rose, "Phonetic aspects of nootka pharyngeals," Master's thesis, University of Victoria, 1979.

[16] J. Stonham and E.-S. Kim, "Labialization in Nuuchahnulth," *Journal of the International Phonetic Association*, vol. 38, no. 1, p. 25–50, 2008.

[17] R. E. Denzer-King, "The acoustics of uvulas in tlingit," Master's thesis, Rutgers University, 2013.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Computer program, 2022.

[19] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[20] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. CreateSpace, 2009.

[21] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *PeerJ Computer Science*, pp. 1–24, 2016.

[22] J. K. Kruschke, "Rejecting or Accepting Parameter Values in Bayersian Estimation," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 2, pp. 270–280, 2018.

[23] B. E. F. Lindblom and J. E. F. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," *Journal of the Acoustical Society of America*, vol. 50, pp. 1166–1179, 1971.

SYNCHRONY AND STABILITY OF ARTICULATORY LANDMARKS IN ENGLISH AND MANDARIN CV SEQUENCES

Benjamin M. Kramer[1], Michael C. Stern[1], Yichen Wang[2], Yuyang Liu[1], Jason A. Shaw[1]

[1]Yale University, [2]Michigan State University
ben.kramer@yale.edu, michael.stern@yale.edu, wangy176@msu.edu, yuyang.liu.yl2472@yale.edu, jason.shaw@yale.edu

ABSTRACT

Theories of articulatory timing differ in whether the onsets or endpoints of movements are coordinated. We assessed these alternatives in word-initial CV sequences in English, a language without lexical tone, and Mandarin, a language with lexical tone, as measured using electromagnetic articulography. We found that, on average, the vowel target and consonant offset were achieved near-synchronously in both English and Mandarin, while the onsets of the consonant and vowel movements showed a temporal lag, a result consistent with endpoint-based coordination. The timing of the near-synchronous landmarks, however, was also more variable than the timing of onsets, suggesting a need for closer examination of synchrony- vs. stability-based metrics of temporal coordination.

Keywords: speech timing, Articulatory Phonology, endpoint-based timing, General Tau theory

1. INTRODUCTION

Languages are known to differ in how articulatory gestures are coordinated (e.g., [1]). Even similar sequences of segments can have different patterns of gestural timing across languages, which can be conditioned by phonological structure, such as syllables [2]–[5]. While the question of how articulatory gestures are coordinated in time has received substantial attention in the Articulatory Phonology (AP) literature [6], competing approaches have also been proposed (e.g., [7], [8]).

A key dimension of variation across theories is whether gestural coordination is based on the *onset* or the *endpoint* of movement. In the coupled oscillator model, as implemented by Nam and Saltzman [9], only the onsets of articulatory gestures are temporally coordinated. Moreover, the onsets of consonant and vowel gestures in a CV syllable are coordinated in-phase, which, in the absence of other influences, predicts temporal synchrony. Such CV synchrony has been reported for English [10] and German [11].

Tone languages, however, tend to show a positive *CV lag*, meaning that the onset of the vowel gesture occurs after the onset of the consonantal gesture. Gao [12] proposed that positive CV lag, first identified in Mandarin, derives from competitive coupling between the consonant gesture, the vowel gesture, and a tonal gesture, which is absent in languages without lexical tone. Similar positive lags have been found in other tonal languages, such as Thai [13] and varieties of Tibetan [14], [15]. There is, however, some conflicting evidence. Liu et al. [8] found consonant and vowel movement onsets to be synchronous (i.e., no CV lag) in Mandarin using a different methodological approach. In fact, methodological variation across studies could account for the difference in whether C and V gestures are found to be synchronous or not; a study of Swedish [16], a pitch accent language, reports a CV lag commensurate with tone languages when using the methods of Mücke et al. [11] but different results when using those of Löfqvist and Gracco [10].

Turk and Shattuck-Hufnagel [7] propose an alternative to onset-based coordination. They theorize that articulatory movement endpoints are involved in temporal coordination in speech, citing greater stability of movement endpoints compared to movement onsets observed across a range of motor functions, including speech [17]. In addition, Shaw and Chen [18] find on-average synchrony of the offset of consonant release movement and the achievement of vowel target in CV syllables in Mandarin. This finding is possibly consistent with endpoint-based models of timing [7], particularly if we consider the movement away from the consonant constriction to be independently controlled [19].

To evaluate onset- and endpoint-based proposals for gestural coordination, we used electromagnetic articulography to record productions of word-initial CV sequences in both English and Mandarin and investigated the synchrony and stability of key intervals bounded by either (1) movement onsets, (2) movement endpoints, or (3) a combination. We considered three landmarks to be movement endpoints: achievement of target of the consonant (Ctar), achievement of target of the vowel (Vtar), and the offset of movement away from consonant constriction (Coff). We investigated four intervals defined by these landmarks, as well as the consonant and vowel movement onsets (Cons, Vons), listed here: (1) Cons–Vons; (2) Vons–Ctar; (3) Ctar–Vtar; and (4) Vtar–Coff. A schematic depiction of the intervals for the syllable /ba/, defined over a lip

aperture (LA) trajectory for the consonant and a tongue dorsum (TD) trajectory in the vertical dimension for the vowel, is provided in Figure 1.

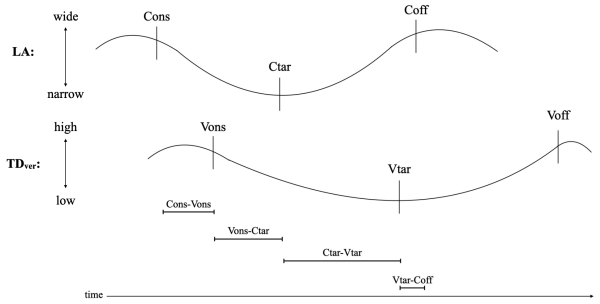


Figure 1: Example diagram of intervals of interest (/ba/).

2. METHODS

2.1 Participants

Acoustic and articulatory data was collected from 12 native speakers of American English (8 female, 4 male; ages 19–28, $M = 20.75$) and 12 native speakers of Mandarin Chinese (7 female, 4 male, 1 non-binary; ages 19–33, $M = 24$). No participants reported speech or hearing impairments, and none of the English speakers spoke or had studied a lexical tone language.

2.2 Materials

We elicited productions of eight word-initial CV sequences in each language, where the initial consonant is bilabial—either /b/ or /m/—and the vowel is either low back /ɑ/ or high front /i/. Target sequences containing the vowel /i/ were immediately preceded by the vowel /ɑ/, and sequences containing the vowel /ɑ/ were immediately preceded by the vowel /i/. This ensured that the magnitude of the tongue body movement towards the target vowel was similar across /i/ and /ɑ/ targets. All Mandarin target syllables bore a falling tone (T4) and were preceded immediately by a low tone (T3). Each target syllable was produced in two carrier sentences, occurring once in an informationally prominent position and once in a less prominent position. To encourage natural speech, each carrier sentence was preceded by a question, which served to provide context for the target sentences.

2.3 Data acquisition

Presentation of materials was controlled using E-Prime. On each trial, an audio recording of a question was played. The question was also displayed in text on the screen for 5000 ms. Participants were instructed to listen to the question and to read aloud the answer that followed.

In total, each participant produced 128 tokens (8 items × 2 carrier sentences × 8 repetitions) across four blocks of 32 items each. Within each repetition block, stimuli were presented in a randomized order.

The NDI Wave Speech Research System was used to record movements of nine sensors attached to the articulators and head at a sampling rate of 100 Hz. High-viscosity PeriAcryl was used to attach three sensors to the tongue: tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), placed ~1 cm, ~3 cm, and ~5 cm from the tip of the tongue. In order to track movements of the jaw, one sensor (lower incisor; LI) was attached to the hard tissue of the gum directly below the left incisor. Two sensors were attached at the vermillion border of the upper lip (UL) and lower lip (LL). Reference sensors were attached on the left and right mastoids, and on the nasion or bridge of the nose. Measurements of the occlusal plane and a midsagittal palate trace were also collected. A Sennheiser shotgun microphone collected acoustic data at a sampling rate of 22,050 Hz.

2.4 Data processing

Articulatory data was rotated to the occlusal plane and corrected for head movement computationally. Articulatory gestures were parsed from sensor trajectories in MVIEW [20]. Gesture *onset* and *offset* were measured as the timepoints at which an articulator's tangential velocity exceeded or sank below, respectively, a 20% threshold of a manually selected velocity peak. *Vowel target* was measured as the timepoint of minimum velocity between the onset and offset, and *consonant target* was measured as the timepoint at which tangential velocity sank below 20% of the onset velocity peak. Gestures associated with /b/ and /m/ were extracted from measurements of lip aperture (LA), calculated as the Euclidean distance between the UL and LL sensors. Gestures associated with the vowel /ɑ/ were parsed from the trajectory of the TD sensor. Gestures associated with the vowel /i/ were parsed from the trajectory of the TB sensor if this sensor was judged to be closest to the palate at the point of maximum /i/ constriction for a participant, and from the TD sensor otherwise.

Out of the 3072 tokens elicited, a total of 556 tokens (18.10%) were eliminated from analysis for the following reasons: data storage failure (18 tokens), failure of the gesture parsing tool to extract the consonant gesture, the vowel gesture, or both (378 tokens); disfluency (5 tokens); or failure of the participant to produce contrastive focus on the informationally prominent syllable, as judged by the experimenters (155 tokens).

3. RESULTS

Distributions of the four temporal intervals of interest are presented in Figure 2. The figures and all subsequent analysis exclude 69 tokens, for which at least one of the four intervals was greater than three standard deviations from the mean. Table 1 presents the mean by-subject duration and standard deviation of each interval.

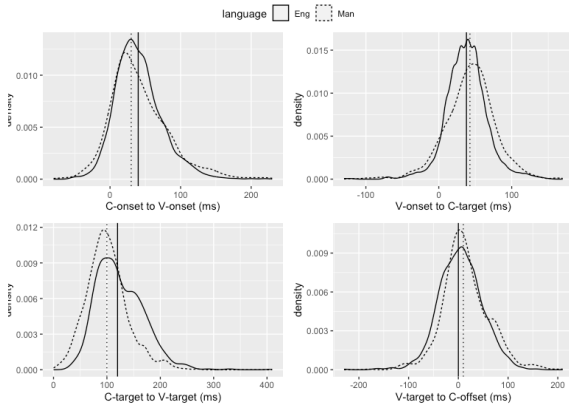


Figure 2: Distribution of key intervals across subjects. Vertical lines indicate the median.

Interval	Mean (ms)	St. Dev. (ms)
Cons-Vons	40.19 (Eng.) 41.17 (Man.)	29.51 (Eng.) 39.59 (Man.)
Vons-Ctar	37.94 (Eng.) 42.85 (Man.)	25.88 (Eng.) 32.38 (Man.)
Ctar-Vtar	123.81 (Eng.) 98.98 (Man.)	40.80 (Eng.) 36.00 (Man.)
Vtar-Coff	4.69 (Eng.) 15.33 (Man.)	42.80 (Eng.) 44.95 (Man.)

Table 1: By-subject mean and standard deviation of key intervals.

First, regarding synchrony, the mean duration of Vtar–Coff is much closer to zero than the other intervals, a pattern that also holds for each subject individually (Figure 3). To further probe this interval, we fit a linear mixed effects model with fixed effects of language, target syllable prominence, and vowel (all sum-coded) and random intercepts for subject and item. The effects of prominence and target vowel were small and not statistically significant (Table 2). The effect of language was significant, with Vtar–Coff longer in Mandarin than in English. The intercept, representing the grand mean, is 10.60 ms, indicating that the consonant offset occurs, on average, 10 ms after the vowel target. Language has a -5.74-ms effect, indicating that the English mean is ~6 ms shorter than the grand mean (i.e., the consonant offset comes just 4 ms after vowel target). Since English and Mandarin have been predicted to differ in Cons–Vons, we also fit a linear mixed effects model with the same structure as the one reported in Table 2 to this interval. Language was not a significant predictor of Cons–Vons duration, as determined by nested model comparison ($\chi^2(1) = .02$, $p = .88$).

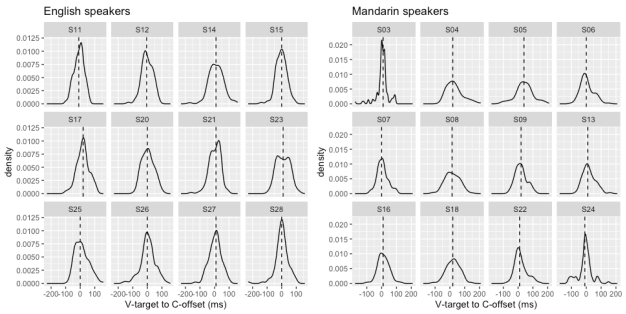


Figure 3: Distribution of Vtar-Coff by subject.

	Estimate	Std. Error	*df*	*t* val.	*p* val.
(Intercept)	10.60	3.54	10.94	3.00	0.01
Language	-5.74	2.09	21.27	-2.75	0.01
Prominence	-0.62	0.88	2418.19	-0.70	0.48
Vowel	-3.67	2.99	5.95	-1.23	0.27

Table 2: Results of linear mixed effects model of Vtar–Coff.

Second, regarding stability, the standard deviations in Table 1 indicate that Vons–Ctar, an interval defined by one onset-based landmark and one endpoint-based landmark, was the most stable. The Vtar–Coff interval, which was the closest to zero, was more variable than the others. Levene's tests confirm that the Vtar–Coff interval is significantly more variable than Cons–Vons ($F = 71.33$, $p < .001$) and Vons–Ctar ($F = 243.18$, $p < .001$); the difference with Ctar–Vtar was not significant ($F = 1.63$, $p = .20$). In addition, Welch's t-tests indicate that by-subject standard deviations for the Vtar–Coff interval are significantly higher than for the other three intervals (every $p < .001$).

4. DISCUSSION

In order to assess proposals for the coordination of consonant and vowel gestures, we evaluated the synchrony and stability of four sets of gestural landmarks, corresponding to either movement onsets or movement endpoints. The landmarks that showed greatest synchrony were Coff (endpoint of consonant release movement) and Vtar (endpoint of vowel

movement to target) [mean duration = 4.69 ms (Eng.), 15.33 ms (Man.)]. The inter-landmark interval that showed the greatest temporal stability was the interval between Vons (onset of vowel movement) and Ctar (endpoint of consonant closure movement) [*SD* = 25.88 ms (Eng.), 32.38 ms (Man.)]. Notably, the landmarks that were most nearly synchronous were different than those that were most stable.

It is also noteworthy that we did not find a significant effect of language on the interval Cons–Vons, since it has been proposed that tonal gestures in languages with lexical tone bring about different patterns of temporal coordination compared to languages without lexical tone [12], [21]. We found a delay (i.e., a CV lag) between Cons and Vons not only in Mandarin (*M* = 41.70), but also in English (*M* = 40.57), a language in which onsets of C and V gestures have been theorized to be synchronized [9]. At least as indicated by the Cons–Vons interval, there was no difference in coordination between languages. Furthermore, the difference between languages was relatively small—only 10.64 ms (approximately one sample of our EMA data)—for the most synchronous interval, Vtar–Coff [mean duration = 4.69 ms (English), 15.33 ms (Mandarin)].

While consonant offset and vowel target may be near-synchronous in English and Mandarin, the question of whether these articulatory landmarks are coordinated (i.e., controlled by the speech production system such that they occur at the same time) remains open. A hypothesized coordination of consonant offset and vowel target represents a departure from the coupled oscillator model of gestural timing, in which only gesture onsets are eligible for coordination, with timing of a trajectory beyond onset simply following from the gesture's dynamic parameters [6], [7]. Gestural alignment patterns referencing targets have, however, featured in other AP research [22]. Phonological control of consonant offset and vowel target is also consistent with the proposal by Turk and Shattuck Hufnagel [7] that movement endpoints—not onsets—are controlled in speech timing, as long as it is assumed that consonant offset is the endpoint of a release gesture (cf. [19]) and vowel target is also the endpoint of a gesture. The interval we found to be most stable, Vons–Ctar, is also bounded by an endpoint (Ctar) on one end. Vons–Ctar stability is consistent with sequential timing (180-degree phasing) in AP, whereby the vowel gesture starts when the consonant gesture achieves its target; this model, however, also predicts temporal synchrony, which was not observed.

As suggested by Turk and Shattuck-Hufnagel [7], a possible theoretical implementation of endpoint-to-endpoint coordination is provided by General Tau theory [23], which allows for two movements to be *tau-coupled*, meaning that continuous, mutual sensory input guides the rate of closure (to a *goal*) such that the goals of each movement are achieved simultaneously. Temporal coordination of gesture endpoints (consonant release and vowel) may also be compatible with an Articulatory-Phonology-style framework if the stiffness parameter is allowed to be modulated (cf. [24]) continuously throughout the duration of the gesture to achieve synchronized gap closure. Synchronizing vowel target achievement with the offset of the consonant release gesture may serve the function of maximizing the acoustic salience of the vowel target by reducing acoustic interference from an ongoing consonantal constriction. This synchronized event may also correspond with the "peakRate" event in the acoustic signal, which is privileged in neural processing of speech [25].

In interpreting these results, it is important to note that we used similarly structured materials, the same method of gesture parsing, and the same methods of analysis across languages, which mitigates the issue of drawing cross-linguistic conclusions without methodological unity [16]. In undertaking a more direct comparison across languages, we found that there is, in fact, less variation in CV timing than expected based on past work. Both English and Mandarin showed a similar CV onset lag. Moreover, in both languages, Vons–Ctar was the most stable interval and Vtar–Coff was the closest to zero, indicating near synchrony of the endpoints of the consonant release movement and the vowel opening movement. Further research should aim to reconcile the metrics of synchrony and temporal stability as indicators of coordination, which paint different pictures in this data.

5. CONCLUSION

To assess theoretical proposals for the temporal coordination of gestures across languages with and without lexical tone, we investigated the synchrony and stability of articulatory landmarks based on both movement onsets and movement endpoints. CV timing was similar across English and Mandarin in several respects. Both languages had a positive CV lag, a pattern thought to be characteristic of languages with lexical tone but not of those without lexical tone. Additionally, the offset of controlled movement of the consonant occurred around the same time as the vowel target in both languages, in line with previous findings for Mandarin CV syllables [18]. These synchronous landmarks, however, were less stable than the others investigated, indicating that more work is required to assess the validity of synchrony and stability as indicators of coordination.

6. REFERENCES

[1] J. A. Shaw, "Micro-prosody," *Lang. Linguist. Compass*, vol. 16, no. 2, pp. 1–21, 2022, doi: 10.1111/lnc3.12449.

[2] J. A. Shaw and A. I. Gafos, "Stochastic time models of syllable structure," *PLoS ONE*, vol. 10, no. 5, pp. 1–36, 2015, doi: 10.1371/journal.pone.0124714.

[3] L. Goldstein, I. Chitoran, and E. Selkirk, "Syllable Structure as Coupled Oscillator Modes: Evidence from Georgian vs. Tashlhiyt Berber," *Proc. XVI Int. Congr. Phon. Sci.*, no. August, pp. 241–244, 2007.

[4] A. Hermes, D. Mücke, and B. Auris, "The variability of syllable patterns in Tashlhiyt Berber and Polish," *J. Phon.*, vol. 64, pp. 127–144, 2017, doi: 10.1016/j.wocn.2017.05.004.

[5] A. I. Gafos, J. Roeser, S. Sotiropoulou, P. Hoole, and C. Zeroual, "Structure in mind, structure in vocal tract," *Nat. Lang. Linguist. Theory*, vol. 38, no. 1, pp. 43–75, Feb. 2020, doi: 10.1007/s11049-019-09445-y.

[6] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, May 1992, doi: 10.1159/000261913.

[7] A. Turk and S. Shattuck-Hufnagel, *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford: Oxford University Press, 2020.

[8] Z. Liu, Y. Xu, and F.-F. Hsieh, "Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics," *J. Phon.*, vol. 90, 2022, doi: 10.1016/j.wocn.2021.101116.

[9] H. Nam and E. Saltzman, "A Competitive, Coupled Oscillator Model of Syllable Structure," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003, pp. 2253–2256.

[10] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1864–1876, 1999, doi: 10.1121/1.426723.

[11] D. Mücke, H. Nam, A. Hermes, and L. Goldstein, "Coupling of tone and constriction gestures in pitch accents," *Consonant Clust. Struct. Complex.*, pp. 205–230, 2012, doi: 10.1515/9781614510772.205.

[12] M. Gao, "Mandarin Tones: An Articulatory Phonology Account," Doctoral dissertation, Yale University, 2008.

[13] R. Karlin and S. Tilsen, "The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai," vol. 060006, no. 2014, p. 060006, 2015, doi: 10.1121/2.0000089.

[14] F. Hu, "Tones are not abstract autosegmentals," *Proc. Int. Conf. Speech Prosody*, vol. 2016-Janua, no. September, pp. 302–306, 2016, doi: 10.21437/speechprosody.2016-62.

[15] C. Geissler, J. A. Shaw, F. Hu, and M. Tiede, "Consistent C-V timing across speakers of diaspora Tibetan with and without lexical tone contrasts," in *Proceedings of the 12th International Seminar on Speech Production*, 2021.

[16] M. Svensson Lundmark, J. Frid, G. Ambrazaitis, and S. Schötz, "Word-initial consonant-vowel coordination in a lexical pitch-accent language," *Phonetica*, vol. 78, no. 5–6, pp. 515–569, 2021, doi: 10.1515/phon-2021-2014.

[17] J. S. Perkell and M. L. Matthies, "Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2911–2925, May 1992, doi: 10.1121/1.403778.

[18] J. A. Shaw and W. R. Chen, "Spatially Conditioned Speech Timing: Evidence and Implications," *Front. Psychol.*, vol. 10, no. December, pp. 1–17, 2019, doi: 10.3389/fpsyg.2019.02726.

[19] H. Nam, "Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structuer," *Lab. Phonol.*, vol. 9, pp. 483–506, 2007.

[20] M. Tiede, "MVIEW: Software for visualization and analysis of concurrently recorded movement data." Haskins Laboratories, New Haven, CT, 2005.

[21] M. Zhang, C. Geissler, and J. Shaw, "Gestural Representations of Tone in Mandarin: Evidence From Timing Alternations," *Int. Congr. Phon. Sci. ICPhS 2019*, no. August, pp. 1803–1807, 2019.

[22] A. I. Gafos, "A grammar of gestural coordination," *Nat. Lang. Linguist. Theory*, vol. 20, no. 2, pp. 269–337, 2002, doi: 10.1023/A:1014942312445.

[23] D. N. Lee, "Guiding Movement by Coupling Taus," *Ecol. Psychol.*, vol. 10, no. 3–4, pp. 221–250, 1998, doi: 10.1080/10407413.1998.9652683.

[24] K. D. Roon, P. Hoole, C. Zeroual, S. Du, and A. I. Gafos, "Stiffness and articulatory overlap in Moroccan Arabic consonant clusters," no. 2004, pp. 1–23, 2021.

[25] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Sci. Adv.*, vol. 5, no. 11, pp. 1–14, 2019, doi: 10.1126/sciadv.aay6279.

THE COARTICULATORY BEHAVIOR OF STANDARD MANDARIN APICAL VOWELS

Sean Foley

University of Southern California
seanfole@usc.edu

ABSTRACT

Although not uncommon, syllabic consonants remain understudied. This study looks at gestural overlap in the apical vowels of Standard Mandarin, which have been analyzed as "syllabic approximants" and "syllabic fricatives". Using ultrasound tongue imaging, coarticulatory effects from adjacent consonants were quantified as a measure of coarticulatory resistance for the two apical vowels in comparison to the three corner vowels of the language. The results show that both apical vowels are considerably less resistant to coarticulation than all three corner vowels, while showing particular susceptibility to these effects in the tongue dorsum. The lack of tongue body stability suggests both segments lack frication noise targets, with their primary tasks involving the tongue tip or tongue blade. The low resistance to coarticulatory effects in both segments stands in contrast to previous studies on gestural overlap in syllabic consonants and is attributed to the nature of their constrictions and their low informativity.

Keywords: coarticulation, apical vowels, Standard Mandarin, ultrasound

1. INTRODUCTION

1.1. Syllabic consonants and gestural overlap

Despite not being particularly rare [1], the phonetics of syllabic consonants is not very well understood, with most studies coming from a few languages, e.g. Tashlhity Berber [2, 3] and Slovak [4]. One of the core claims of models of syllabic structure is that CV syllables are the most stable, as they allow for in-phase coordination between the onset and nucleus [5, 4, 6]. As such, syllables occupied solely by consonants should be disprefered, leading some to hypothesize that syllabic consonants may become more vowel-like when occupying the nucleus [4, 7].

However, studies on articulatory timing and gestural overlap in syllabic consonants in comparison to onset and coda consonants have generally come to the same conclusion: position in the syllable has little effect on the relative timing and gestural coordination of a consonant – phonetically nucleic consonants are indistinguishable from their coda and onset counterparts [3, 4]. Furthermore, these studies have shown that less gestural overlap occurs between consonants when the nucleus is filled by a consonant, as opposed to onset and coda clusters [4, 3, 8]. [7] posited that the lack of gestural overlap in these cases occurs to allow for recoverability of the consonantal gestures.

In one of the few typologies of syllabic consonants, [1] states that diachronically the source of syllabic consonants is almost always a vowel, whether its due to vowel deletion or vowel "consontantalization". According to [1], this latter process is particularly rare, occurring only in Sino-Tibetan languages or Lendu (Central Sudanic) (See [9] for other potential examples). This difference suggests that syllabic consonants that arise from vowel "consonantalization" may differ phonetically from those due to vowel deletion. This study looks at the apical vowels of Standard Mandarin, which are believed to have their origin in high front vowel "apicalization" due to diachronic pressures to maintain sibilant contrasts [10, 11].

1.2. Standard Mandarin apical vowels

In Standard Mandarin (SM), there is a three-way place contrast among sibilants, with the language contrasting dental, alveolo-palatal and retroflex sibilants, e.g. /s ɕ ʂ/. One consequence of this three-way place contrast is the co-occurrence restriction on the high front vowel /i/ following dental and retroflex sibilants, e.g. *si *ʂi. In these contexts, in place of the high front vowel, there occurs two apical segments, [ɿ] and [ʅ], which occur only after sibilants they are homorganic with, e.g. [sɿ] and [ʂʅ] [12]. Therefore, many analyses consider these segments to be allophones of /i/ [11]. Although "apico-dental/alveolar" and "apico-postalveolar" may be potentially more accurate phonetic descriptions of the segments [9], [ɿ] will be referred to as the "dental apical vowel" and [ʅ] as the "retroflex apical vowel" in keeping with the previous literature.

Previous research has shown that both apical vowels are produced with a lingual configuration that closely resembles their onsets [11, 13], though questions remain whether this is due to an assimilatory effect or is inherent to the segments themselves. Despite some impressionist analyses that have analyzed them as "syllabic fricatives" [12, 14], recent evidence suggests that frication noise is present only during the onset of the segments and is likely attributable to the preceding sibilant [15]. As argued in [11], these facts suggest that both segments could best be analyzed as "syllabic approximants". Therefore, given previous research on syllabic consonants, sibilant plus apical vowel sequences could be seen as CC clusters, with a sequence such as [sɿ sa] analyzed as CC.CV. Given these assumptions, the aim of this study is to investigate the coordinating relations and coarticulatory behavior of both apical vowels in comparison to the three corner vowels [i a u] to assess whether or not they pattern with typical vowels in these respects and to assess the nature of their gestural targets.

Assuming that coarticulation is constrained by both the nature of the linguistic task [16] and the hydrostatic properties of the tongue [17], the coarticulatory behavior of the apical vowels should provide insight into their phonetic targets. Viewing the segments as "syllabic fricatives" predicts high resistance to coarticulatory effects in both the tongue tip and tongue body, given the strict demands on both articulators to produce frication noise [18, 19]. In contrast, the "syllabic approximant" view only predicts local resistance in the primary articulator, i.e. the tongue tip or tongue blade.

2. METHOD

2.1. Ultrasound experiment

Seven native speakers of SM took part in the study, all of whom reported no history of hearing or speech disorders. Data from two speakers was excluded due to errors in the placement of the ultrasound probe (SP_03 and SP_06). The five remaining speakers were all aged 18-25 years old. Three speakers were from northern provinces (Liaoning, Shandong, Shaanxi) and two were from central/southern provinces of China (Henan, Jiangsu).

The stimuli used in the study consist of 48 disyllabic nonce words. The use of nonce words was warranted due to the lack of disyllabic words containing the target sequences. The target segments in the first syllable are the four vocalic segments [ɿ ɻ i a u], with three different onsets [s ʂ ɕ]. Due to phonotactic restrictions, each apical vowel occurs only after a homorganic sibilant, while [i] occurs only after [ɕ]. Furthermore, [u] occurs only after [s] and [ʂ]. The second syllable consists of the consonants [t^h s ʂ ɕ x k^h] with [a] serving as the nucleus. These six consonants were chosen to represent a range of both manners and places of articulation. The total disyllabic sequences includes all combinations of the target first syllable and second syllable. All syllables had Tone 1 (high level). 16 disyllabic fillers were also used, containing syllables and tones not present in the target stimuli.

Participants were recorded in a sound-attenuated booth. The stimuli were presented in blocks of five and were randomized so that each target phrase was seen a total of five times across all blocks. Different randomizations were used across all speakers , with only SP_02 and SP_07 seeing the same randomization. The stimuli were presented in the following carrier phrase：我覺得 __ 很好 [wə^{21} tɕyei^{35} də__xən^{35} xau^{213}] "I think__(is) very good". Simultaneous ultrasound video and audio were recorded for each participant. Ultrasound images were recorded using an Articulate Instruments Micro portable ultrasound machine.

2.2. Data analysis

Audio files were analyzed in Praat [20] and segmented using the Montreal Forced Aligner [21]. All alignments were manually checked and corrected. Stop closure and bursts were not manually segmented by the forced aligner and were segmented by hand using the offset of the V_1 formants where closure starts to the C_2 offset just before the stop burst as landmarks. Ultrasound frames corresponding to the target segments were processed in Articulate Assistant Advanced (AAA) [22].

Frames closest to the midpoint of the target C_2 and V_1 tokens were analyzed using Smoothing spline (SS) ANOVAs [23] (Figure 1). The tongue root (TR), tongue dorsum (TD), and tongue tip (TT) were also tracked for all target segments using AAA (Figure 2). This was done by specifying radii in AAA that correspond to each of these parts of the tongue from the 42 radii that compose the ultrasound fan. This was done independently by speaker by observing constrictions of these regions in the ultrasound video. From these five repetitions the mean distance distane from the origin was taken at the midpoint of each target segment.

Coarticulatory resistance (CR) was measured as the degree of variability in tongue contours for a

segment across all C_2 contexts [19]. The two apical vowels [ɿ ʅ] were compared to the three peripheral vowels [i a u]. This was quantified for overall variability in lingual posture and variability in the three tongue regions. Overall variability exhibited by each vocalic segment was measured using the polar coordinates that were the output of the SSANOVAs and by taking the mean variance (σ^2) of the distances across all angles, i.e. six distances per angle from six SSANOVA splines (Figure 1). This measure was found to be highly correlated with taking the mean Euclidean distance of the min and max distance across all angles [24] and fitting a polygon over the SSANOVAs and getting the area of that polygon [19]. Variability by tongue region was quantified by taking the variance of the distances for each region across all contexts. In sum, there is one global measure and three local measures of CR (i.e. for the TR, TD, and TT).

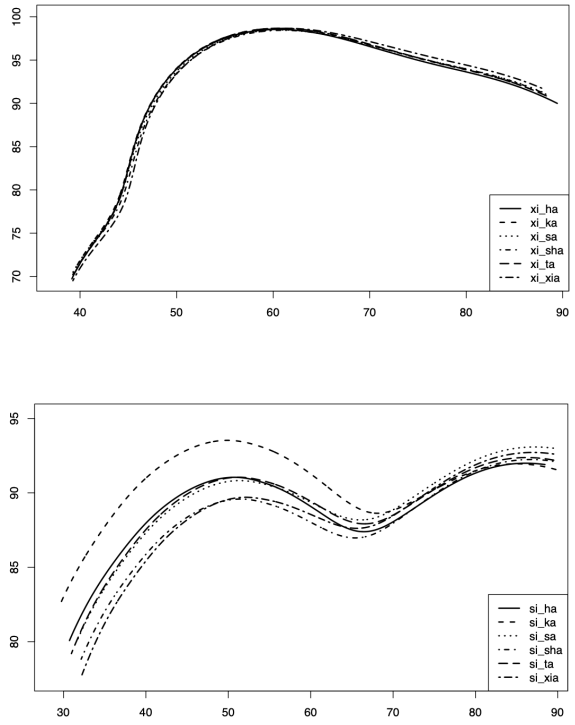


Figure 1: Midsagittal SSANOVA splines for vowels [i] (top) and [ɿ] (bottom) at their midpoints across all target consonant contexts for SP_02. Anterior is to the right.

3. RESULTS

Results are separated into overall tongue variability and variability by the three tongue regions – TR, TD, and TT. To account for across-speaker variability, all results are shown by speaker. As the overall

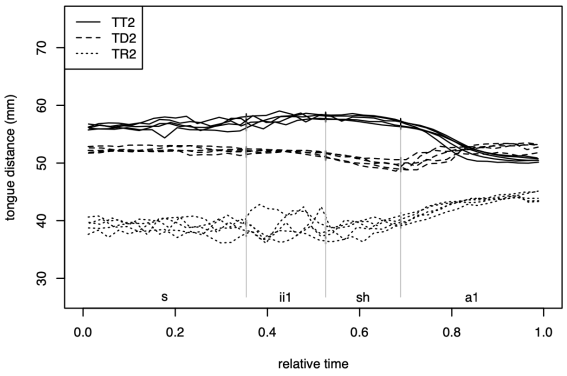


Figure 2: TR, TD, and TT trajectories during "si sha" for SP_04.

and by tongue region variability was measured using variance (σ^2), the ratio of two variances (i.e. one for each segment in a two-way comparison) was taken to get an F-statistic following an F-distribution. F-tests were computed by speaker for all pairwise combinations of the target segments grouped by onset.

3.1. Overall

Figure (3) shows the overall variance for each target vowel segment, separated by the four onset groups, with [i ɿ ʅ] grouped as "homorganic onset" due to each being homorganic with their onsets, i.e. [ɕ s ʂ] respectively. Taller bars indicate greater variance (lower CR) and shorter bars indicate less variance (higher CR). It can be seen for four of the five speakers that the apical vowels consistently show significantly more variability than the three corner vowels ($p < 0.05$). SP_05 behaved differently from the others in this regard, and is indicated as *SP_05. Furthermore, between the two apical vowels, the retroflex segment was consistently more resistant to coarticulatory effects than the dental segment.

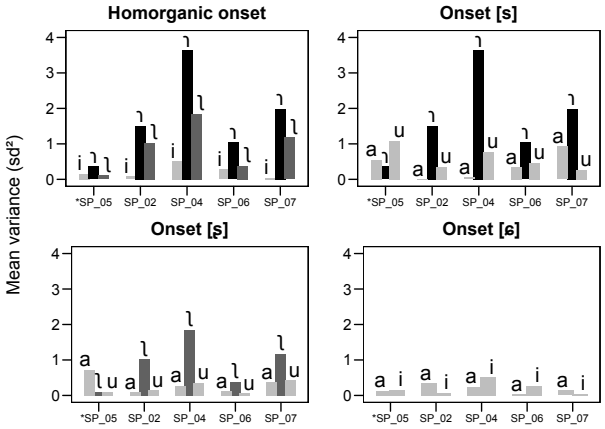


Figure 3: Overall CR for all target segments, organized by shared onset and by speaker.

3.2. By tongue region

Figure (4) shows CR plotted by speaker and by the three tongue regions: TR, TD, and TT, and grouped by onset. A number of observations can be made here. First, four of the five speakers show a tendency for both apical vowels to exhibit greater variability than the three corner vowels in both the TD and TR, with both apical vowels showing much greater variability in the TD than the other two tongue regions. Second, for most speakers, the dental segment [ɻ] has significantly greater TD variance than [i], but this is not always true for [ɻ]. TR variability in both apical vowels is often significantly greater than all three corner vowels, but this is not consistent across speakers. Third, between the two apical vowels, the retroflex segment was consistently more resistant to coarticulatory effects than the dental segment in all three regions. Lastly, linear regression models also showed that for three out of five speakers, the relationship between TT and TD variability is significantly different for the apical vowels compared to the three corner vowels ($p < 0.05$).

4. DISCUSSION

The results show that both apical vowels are considerably less resistant to coarticulatory effects than all three corner vowels. This holds both for overall tongue variability as well as variability in both the TD and TR. While both segments are much more resistant to coarticulatory effects in the TT than other regions, across speakers they were not consistently less resistant to effects in the TT than all three corner vowels. Together, these results suggest that the primary task for both apical vowels is a tongue tip or tongue blade raising gesture, with the lack of tongue body stability in their productions being indicative of both segments not having frication noise targets. This supports previous analyses of the segments as "syllabic approximants".

Previous studies of syllabic consonants have suggested that syllabic consonants permit less gestural overlap to allow for full recoverability of the consonantal gestures [4, 25]. One potential reason for the lack of resistance in the apical vowels is that the information conveyed by the apical vowels is essentially redundant as most of the cues they serve to enhance are already present in the preceding sibilant [11]. In addition, compared to the syllabic consonants of other studies, e.g. /l/ and /r/ in Slovak [4], the constriction tasks of the apical vowels likely require less tongue body activation [26].

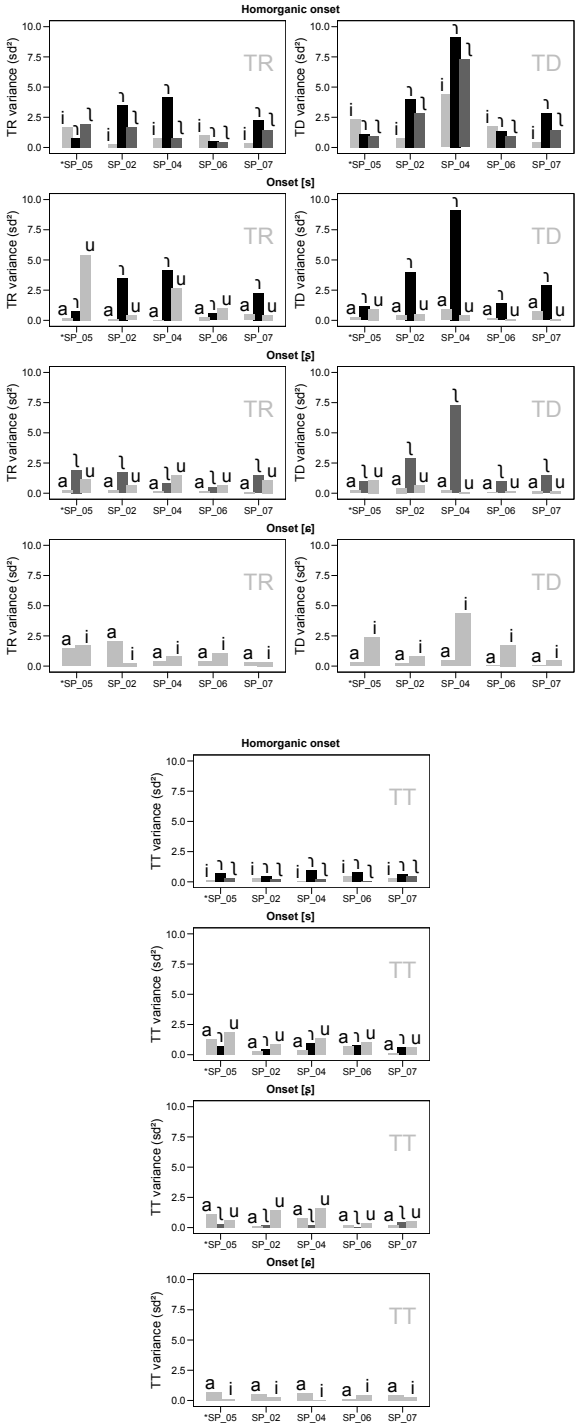


Figure 4: CR by tongue region for all target segments, organized by shared onset and by speaker.

The results are consistent with the view that coarticulatory resistance is highly local [27]. This suggests that rather than viewing segments as "resistant", resistance exists at regions in the vocal tract essential to the task, thus permitting variability in non-task regions [28].

5. ACKNOWLEDGEMENTS

Thanks are due to Jeff Mielke for technical assistance and Elliott Moreton, Jennifer L. Smith, Louis Goldstein, Khalil Iskarous, and members of UNC-Chapel Hill P-Side and USC PhonLunch for helpful comments and discussion.

6. REFERENCES

[1] A. Bell, "Syllabic consonants," in *Universals of human language*, J. Greenberg, Ed. Stanford: Stanford University Press, 1978, vol. 2: Phonology, pp. 153–201.

[2] R. Ridouane, "Syllables without vowels: Phonetic and phonological evidence from tashlhiyt berber," *Phonology*, vol. 25, no. 2, pp. 321–359, 2008.

[3] C. Fougeron and R. Ridouane, "On the phonetic implementation of syllabic consonants and vowel-less syllables in tashlhiyt," *Estudios de Fonética Experimental*, pp. 140–175, 2008.

[4] M. Pouplier and Š. Beňuš, "On the phonetic status of syllabic consonants: Evidence from slovak," *Laboratory phonology*, vol. 2, no. 2, pp. 243–273, 2011.

[5] L. Goldstein, D. Byrd, and E. Saltzman, "The role of vocal tract gestural action units in understanding the evolution of phonology," *Action to language via the mirror neuron system*, pp. 215–249, 2006.

[6] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: A coupled oscillator model," *Approaches to phonological complexity*, vol. 16, pp. 299–328, 2009.

[7] P. Hoole, M. Pouplier, S. Benus, and L. Bombien, "Articulatory coordination in obstruent-sonorant clusters and syllabic consonants: Data and modelling," *Rhotics. New data and perspectives*, pp. 81–98, 2013.

[8] R. A. W. Bladon and A. Al-Bamerni, "Coarticulation resistance in english/l," *Journal of Phonetics*, vol. 4, no. 2, pp. 137–150, 1976.

[9] M. Faytak, "Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese," Ph.D. dissertation, University of California, Berkeley, 2018.

[10] S. Baron, "On the tip of many tongues: Apical vowels across sino-tibetan," in *7th International Conference on Sino-Tibetan Language and Linguistic Studies*, 1974.

[11] S.-I. Lee-Kim, "Revisiting mandarin 'apical vowels' : An articulatory and acoustic study," *Journal of the International Phonetic Association*, vol. 44, no. 3, pp. 261–282, 2014.

[12] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.

[13] M. Faytak and S. Lin, "Articulatory variability and fricative noise in apical vowels." in *ICPhS*, 2015.

[14] A. C. Yu, "Aerodynamic constraints on sound change: The case of syllabic sibilants," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1096–1097, 1999.

[15] M. Faytak and B. Shao, "Degrees of frication in three chinese varieties' fricative vowels," 2022, paper presented at the *LSA Annual Meeting*.

[16] P. A. Keating *et al.*, "The window model of coarticulation: articulatory evidence," *Papers in laboratory phonology I*, vol. 26, pp. 451–470, 1990.

[17] W. M. Kier and K. K. Smith, "Tongues, tentacles and trunks: the biomechanics of movement in muscular-hydrostats," *Zoological journal of the Linnean Society*, vol. 83, no. 4, pp. 307–324, 1985.

[18] M. Tabain, "Variability in fricative production and spectra: Implications for the hyper-and hypo-and quantal theories of speech production," *Language and speech*, vol. 44, no. 1, pp. 57–93, 2001.

[19] D. Recasens and C. Rodríguez, "A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound," *Journal of Phonetics*, vol. 59, pp. 58–75, 2016.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2021.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[22] A. Wrench, "Articulate assistant advanced user guide," *[Computer Program]. Version*, vol. 2, p. 12, 2012.

[23] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[24] N. Zharkova and N. Hewlett, "Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using english /t/ and /a/," *Journal of Phonetics*, vol. 37, no. 2, pp. 248–256, 2009.

[25] I. Chitoran, L. Goldstein, and D. Byrd, "Gestural overlap and recoverability: Articulatory evidence from georgian," *Laboratory phonology*, vol. 7, no. 4-1, pp. 419–447, 2002.

[26] D. Recasens, M. D. Pallarès, and J. Fontdevila, "A model of lingual coarticulation based on articulatory constraints," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 544–561, 1997.

[27] K. Iskarous, C. Mooshammer, P. Hoole, D. Recasens, C. H. Shadle, E. Saltzman, and D. H. Whalen, "The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance," *The Journal of the acoustical society of America*, vol. 134, no. 2, pp. 1271–1282, 2013.

[28] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

QUANTIFYING TONGUE-TIP TO UPPER INCISOR DISTANCE IN AN ULTRASOUND TONGUE IMAGING STUDY OF SWEDISH /i/

Fabienne Westerberg[†], Eleanor Lawson[‡], Jane Stuart-Smith[†]

[†] University of Glasgow

[‡] University of Strathclyde

ABSTRACT

The Swedish /iː/ variant 'Viby-i' has been described as 'thick', 'damped' and 'buzzing', implying the presence of frication. The articulatory basis for its unusual buzzing quality is unknown. We investigate the possibility that frication results from proximity of the tongue tip to the upper incisors during Viby-i production. Ultrasound tongue imaging recordings of word-list speech from 34 Central Swedish speakers was used to study articulatory-acoustic variation across six Swedish long vowels, including Viby-i. Measures of vertical distance between highest point of the tongue and tongue tip indicate that the /iː/ and /yː/ tongue shapes are more apical than the other vowels. /iː/ and /yː/ also have shorter distances between the tongue tip and upper incisors, and acoustic analysis showed greater levels of aperiodicity for /iː/ and /yː/ than for other vowels. However, a correlational analysis did not identify a significant negative relationship between tongue-tip-to-incisor distance and levels of aperiodic noise.

Keywords: Ultrasound Tongue Imaging, vowels, Swedish, Articulatory Phonetics, Sound Change.

1. INTRODUCTION

The Swedish /iː/ variant known as Viby-i (sometimes also called Lidingö-i) has an unusual quality, impressionistically characterised as 'thick', 'damped' and 'buzzing' [1], [2], [3]. Named after Viby parish in Central Sweden, its articulatory nature has long been disputed. [yː] is also found to have a similar buzzing quality [4], but is investigated only articulatorily in this paper. A key acoustic feature of Viby-i, determined from perceptual experiments, appears to be that its F2 is lower than that of /eː/ [5]. However, there is no consensus on how Viby-i is produced, with conflicting impressionistic accounts of the tongue movements involved [6], [7] ,[8], [9]. An ultrasound tongue imaging (UTI) and lip camera study of Viby-i in 2016 [10], identified that many speakers in the study produced Viby-i with an extremely fronted tongue, to the point where the tip could sometimes be seen protruding between the front teeth. A typical visual example is provided in Figure 1 (left), showing front-facing and profile lip-camera views of the midpoint of a Swedish female speaker's prodiction of [iː] and [e] (for comparison). For [iː], but not for [e], the tongue tip can see seen to be in close proximity to the upper incisors.

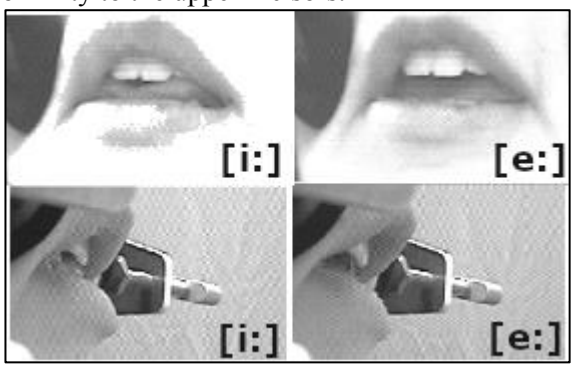


Figure 1. Comparison of tongue-tip postures for /iː/ (left) and /eː/ (right), produced by a female Swedish speaker [10].

We hypothesise that proximity of the tongue tip to the upper incisors is responsible for the 'buzzing', or fricated, quality of some variants of Viby-i, and use articulatory-acoustic methods to investigate aspects of Viby-i production compared to other Swedish long vowels.

2. METHOD

The existing subcorpus of Swedish word-list recordings used in this study was recorded in Central Sweden in 2016. Speakers were recorded in soundproofed studios, or in a quiet room, at the Universities of Gothenburg (12 speakers – 6M/6F), Stockholm (10 – 5M/5F), and Uppsala (8 – 3M/5F). Speakers were audio recorded and their midsagittal tongue movements were recorded with UTI, as they read aloud multiple repetitions of mono- and disyllabic words, containing a subset of the Swedish long vowels. Five long vowels were articulatorily and acoustically measured and analysed alongside /iː/, as a point of comparison, see Table 1.

Table 1: Swedish long vowels included in this study and number of tokens per vowel. N=2790.

vowel	iː	yː	eː	øː	oː	uː

N. tokens	1487	263	265	261	261	253

All /i:/ vowels in the study can be considered to be Viby-i variants, based on the acoustic criterion of having a lower F2 than /e:/.

2.1 Equipment

Recordings of the midsagittal tongue were made using a portable Telemed Echo-Blaster 128 ultrasound machine and a micoconvex probe with a 104º scan radius, and scanning at 67.19 frames per second. The probe was held in place under the chin using a stabilising headset to reduce/eliminate, probe movements in the coronal, axial and sagittal planes.

An Audio-Technica AT831b cardioid lapel microphone, sampling at 44,100 Hertz (16 bit), was clipped to the headset near the participant's mouth.

2.2 Articulatory measures

2.2.1 Occlusal/bite plane reference

The bite plane recording, see [12], is obtained using a medical-grade plastic bite plate (see Figure 2).

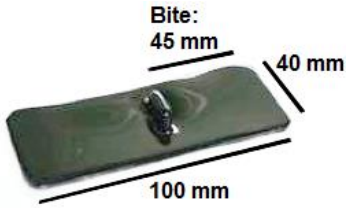


Figure 2. A plastic bite plate.

The bite plate is vacuum-moulded from a standard template with vertical protrusion part-way along its length. Participants are asked to place it inside their mouth, grip it with their molar teeth, with incisors positioned directly before the vertical protrusion. They are instructed to press their tongue flat against the underside of the bite plate, allowing imaging of a flat tongue surface that approximates the speaker's occlusal (bite) plane and helps determine the angle of the ultrasound probe relative to the occlusal plane, which can then be standardised to a 90º angle to improve inter-speaker comparison, see Figure 3.

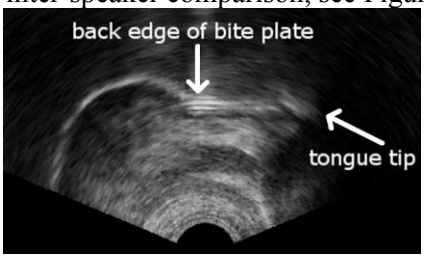


Figure 3. A midsagittal image of the "bite plane".

An additional potential benefit of using the bite plate is that it can help approximate the location of the biting edge of the upper incisors. The back of the bite plate is visible on the ultrasound image as the tongue bulges upwards, see Figure 3. It will be a fixed distance from the biting edge of the upper incisors allowing incisor position to be approximated.

We have estimated this fixed distance to be 45mm with a standard deviation of +/-1mm, based on measurement of 30 bite plates. Using AAA software [11], we fitted a scaled fiducial line, 45mm long, to the imaged bite plane surface, placing the left edge of the line at the point where the tongue's surface bulges upwards. The Cartesian coordinates of the termination of the line will approximate the location of the biting edge of the upper incisors.

2.2.2 Midsagittal tongue-surface measurement

A spline was fitted to each UTI-imaged midsagittal tongue surface, using AAA, at 10% of the vowel duration, as Central Swedish vowels reach their targets early, before moving into variable offglides [4]. Splines were then exported as sets of Cartesian coordinates for further analysis using R [13]. Automatic measures were taken of the highest point of the tongue and tongue tip location, and Lobanov normalised [14], to allow inter-speaker comparison.

2.2.4 Single-point tongue-surface measures

Single-point tongue measures can be used to (1) identify the relative tongue height of the vowels studied, and (2) characterise and quantify tongue shape by subtracting the tongue-tip height from the highest point of the tongue. If these values are identical, or very similar, resulting in a value close to 0, then it can be assumed that the vowel is produced with an apical (tongue-tip-raised) articulation; whereas greater positive vertical distance between the highest point and tongue tip indicates a more canonical convex vocalic tongue shape, with a lowered tongue tip and raised tongue middle, or dorsum.

2.2.3 Tip-to-incisor measure

The tongue-tip-to-biting-edge-of-incisor measure (in mm) was obtained automatically, using R, measuring the Euclidean distance between the (x,y) coordinates for (1) the tongue tip and (2) the first coordinate of the bite-plane fiducial.

2.3 Acoustic measures

Bandpass-filtered zero-crossing rate (bpZCR) quantifies the degree of aperiodicity in an acoustic signal, which may contain voicing [15]. This method filters out the vowels' periodic energy, and measures how often the waveform crosses the zero point in the remaining signal. Higher rates of zero-crossings indicates greater levels of aperiodicity/frication.

First, all frequencies below 1,000 Hz were removed. Filtering was carried out using the 'Hann Band filter' function in Praat [16], set to 100 Hz smoothing. Then the 'Zero Point Process' function was used to identify all (falling and rising) zero-crossings. The average zero-crossing rate was calculated for a 10% window the vowel onset, as acoustic analysis of Central Swedish vowels showed them to be very dynamic and much more variable after the first 10% of the vowel [17].

2.4 Articulatory and articulatory-acoustic analysis

Using spline and fiducial Cartesian coordinates, R was used to automatically quantify (1) the highest point of the tongue, (2) vertical distance between highest point of the tongue and tongue-tip, (3) Euclidean distance between the tongue tip and biting edge of the upper incisors. Praat was used to measure (4) mean bpZCR 10% of the vowel. Finally, using R, we carried out (5) a correlational analysis of measures (3) and (4), hypothesising that a significant negative correlation would be found.

3. RESULTS

3.1 Highest point of the tongue

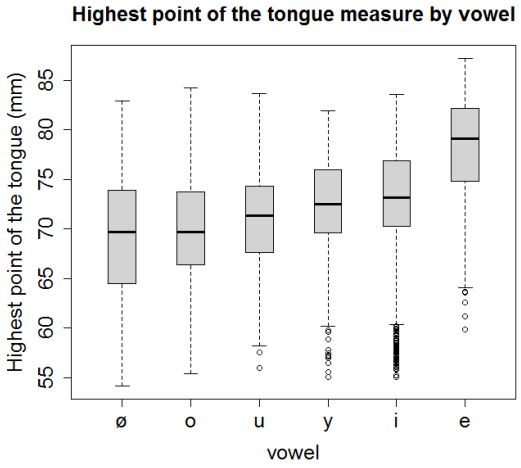


Figure 4. Highest point of the tongue measure (mm – distance from probe surface), by vowel, ordered by mean. N=2790.

Figure 4 shows the highest point of the tongue for all vowels. An ANOVA identified significant variation across these vowels, F=80.99, p<0.001. Figure 4 shows that /e:/, not /i:/, is the highest vowel in articulatory space. Post-hoc Tukey tests showed significant differences between /e:/ - /i:/ and /e:/ - /y:/, p<0.001.

3.1 Highest point of the tongue to tongue-tip height

Figure 5 shows the vertical distance between the tongue tip and the highest point of the tongue. An ANOVA identified significant variation across the vowels studied F=528.9, p<0.001. Post-hoc Tukey tests showed no significant differences between /i:/ and /y:/, which can therefore be said to have the most tip-raised articulation of the vowels studied. The tongue tip was identified as the highest point of the tongue (indicating an apical articulation) for 15% and 13% of /y:/ and /i:/ tokens respectively; whereas the tongue tip was the highest point of the tongue for /ø:/ and /e:/ in only 7% and 4% of tokens respectively, and never for /o:/ and /u:/. Post-hoc Tukey tests found significant differences between /y:/, /i:/ and all other vowels studied at p<0.001.

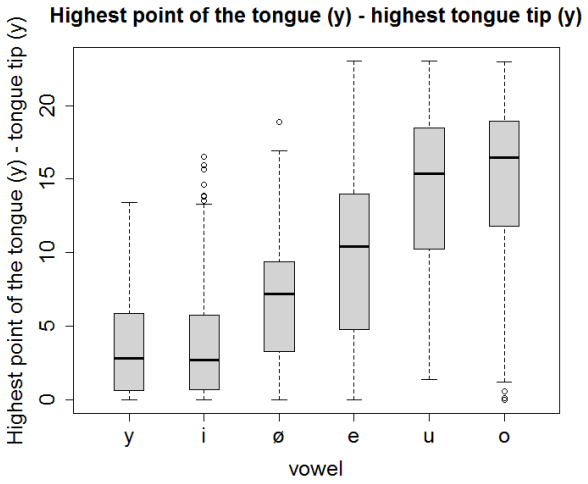


Figure 5. Vertical distance measure between the highest point of the tongue and tongue tip (mm), by vowel, ordered by mean. N=2790.

3.1 Articulatory analysis – tongue-to-teeth measure

Figure 6 shows the tongue-tip-to-biting-edge-of-incisor measure. An ANOVA found significant variation across vowels F=71, p<0.001. Tukey posthoc tests showed no significant differences between /i:/, /y:/ and /ø:/, but significant differences between these vowels and all other vowels at p<0.001. The significant difference between /e/ and /ø/ was at p<0.05.

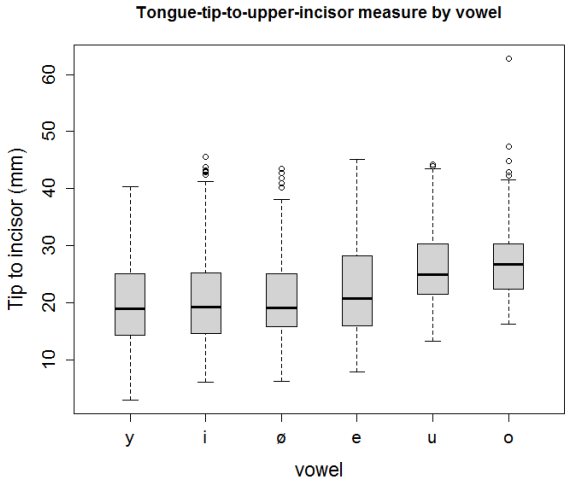


Figure 6. Tongue-tip-to-upper-incisor Euclidean distance (mm), by vowel, ordered by mean. N=2790.

3.2 Acoustic analysis – bandpass-filtered zero-crossing-rate measure

Figure 7 shows the mean bandpass-filtered zero-crossing rate (bpZCR) across the first 10% of the vowel.

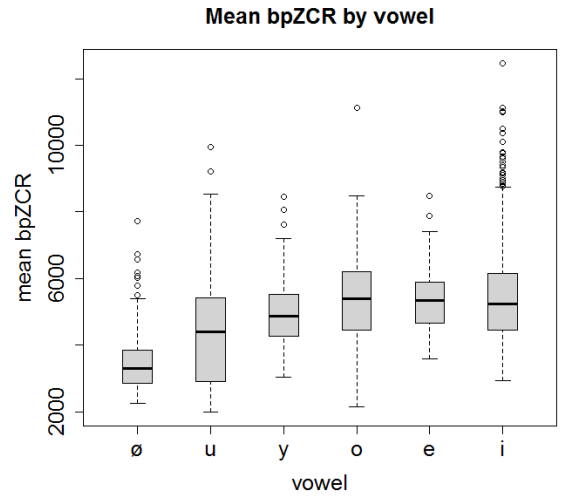


Figure 7. Average bpZCR by vowel, ordered by mean. N=2790.

A higher zero-crossing rate indicates greater levels of aperiodicity in the acoustic signal. An ANOVA found significant differences in teeth-to-incisor measures across vowels F=126.9, p<0.001. Post-hoc Tukey tests showed no significant differences in mean bpZCR between /i:/, /e:/ and /o:/, but significant differences between /i:/ and /y:/, p<0.001, /e:/ and /y:/ p<0.01, and /i:/ and /o:/, p<0.001.

3.3 Acoustic and articulatory correlational analysis

A Pearson's correlation test was carried out on the dependent variables (1) Euclidean distance between tongue tip and biting edge of the upper incisors, and (2) mean bpZCR value across the first 10% of the vowel. There was a significant positive correlation r_P=0.15, p<0.001 (see Figure 8), which was contrary to our hypothesis that a narrower tongue-tip-to-upper-incisor gap would result in greater frication (indicated by bpZCR).

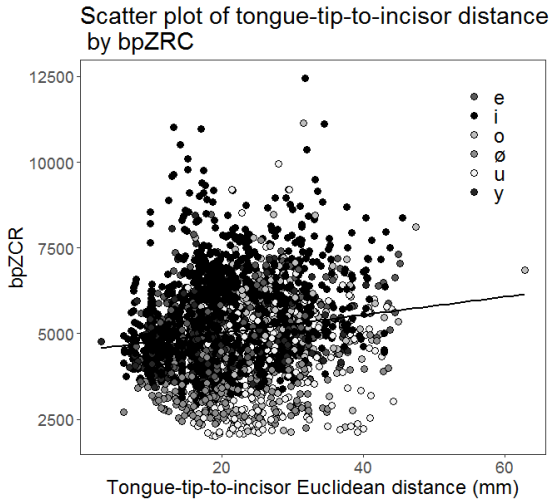


Figure 8. Tongue-tip-to-incisor Euclidean distance (mm) by mean bpZCR (aperiodic noise measure). N=2790.

4. CONCLUSION

This study investigated whether the unusual "buzzy", or fricated, quality of Viby-i could be attributed to the proximity of the tongue tip to the upper incisors.

Our results show that while /e:/ has, on average, a higher tongue-body position in the oral cavity than /i:/, /i:/ has, on average, a higher tongue-tip position, and often an apical tongue shape. We found that /y:/ is also often produced with this articulatory strategy, and, indeed, has also been found to have a similar buzzing quality [4]. /i:/ had the greatest degree of aperiodic energy of the vowels studied, but not significantly greater than /e:/, or /o:/. Our hypothesis of a significant negative relationship between tongue-tip-to-incisor distance and bpZCR value (as a measure of aperiodic noise) was disproved. Despite this negative result, we have established that Viby-i is often produced with a raised tongue tip – an unusual vocalic tongue posture. Additionally, the method of approximating the position of the biting edge of the upper incisors using a bite plate could be of use to future UTI-based speech studies, providing another fixed landmark from the upper surface of the vocal tract with which to compare tongue position.

7. REFERENCES

[1] Engstrand, O., Björnsten, S., Lindblom, B., Bruce, G., & Eriksson, A. (1998). Hur udda är Viby-i? Experimentella och typologiska observationer. Folkmålsstudier, 39, 83–95.

[2] Kotsinas, U.-B. (2007). Ungdomsspråk (3rd ed.). Uppsala: Hallgren & Fallgren.

[3] Schötz, S., Frid, J., & Löfqvist, A. (2011). Exotic vowels in Swedish: An articulographic and acoustic pilot study of /i:/. In Proceedings of the 17th International Congress of Phonetic Sciences (pp. 1766–1769). Hong Kong.

[4] Bruce, G. (2010). Vår Fonetiska Geografi: Om Svenskans Accenter, Melodi och Uttal. Lund: Studentlitteratur.

[5] Björsten, S., & Engstrand, O. (1999). Swedish "damped" /i/ and /y/: Experimental and typological observations. In Proceedings of the 14th International Congress of Phonetic Sciences (pp. 1957–1960). San Francisco.

[6]Lundell, J. A. (1878). Det svenska landsmålsalfabetet. Svenska landsmål och svenskt folkliv, 1(2), 11–158.

[7] Borgström, M. (1913). Askermålets ljudlära. Svenska landsmål och svenskt folkliv, B(11).

[8] Ladefoged, P. & Lindau, M. (1989). Modeling articulatory-acoustics relations: A comment on Stevens' "On the quantal nature of speech". Journal of Phonetics, 17, 99–106.

[9] Noreen, A. (1903). Vårt språk (Vol. 1). Lund: Gleerups.

[10] Westerberg, F. (2016). An Auditory, Acoustic, Articulatory and Sociophonetic Study of Swedish Viby-i (Masters dissertation). University of Glasgow.

[11] Articulate Instruments Ltd. (2019). Articulate Assistant Advanced, version 2.17. Retrieved 2019-03-26,from http://www.articulateinstruments.com/downloads

[12] Lawson, E., Stuart-Smith, J., & Rodger, L. (2019). A comparison of articulatory and acoustic parameters for the GOOSE vowel across British Isles Englishes. The J. Acoust. Soc. Am., 146(6). 4363–4381.

[13]R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved 2019-11 28, from http://www.r-project.org/

[14]Lobanov, B. M. (1971). "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am. 49(2), 606–608.

[15]Gordeeva, O. B., & Scobbie, J. M. (2010). Preaspiration as a correlate of word-final voice in Scottish English fricatives. In S. Fuchs, M. Toda, & M. Zygis (Eds.), Turbulent sounds: An Interdisciplinary Guide (pp. 167–208). New York: Walter de Gruyter.

[16]Boersma, P., & Weenink, D. (2019). Praat: Doing Phonetics by Computer, version 2.0.29. Retrieved 2019-03-26, from http://www.praat.org/

[17] Westerberg, F. (2020) Heavens, what a sound! The acoustics and articulation of Swedish Viby-i. Unpublished PhD thesis, University of Glasgow.

THE ROLE OF FEEDBACK IN LIP-TUBE PERTURBATION OF TAIWAN MANDARIN ROUNDED VOWELS

Jian-zhi Huang[1] & Chenhao Chiu[1,2]

[1]Graduate Institute of Linguistics, National Taiwan University;
[2]Neurobiology and Cognitive Science Center, National Taiwan University
r10142007@ntu.edu.tw; chenhaochiu@ntu.edu.tw

ABSTRACT

Previous lip-tube perturbation experiments have focused on the speaker's compensation capacity, but the role of feedback is less concerned. Previous studies have demonstrated that robust compensation may be induced from multisensory feedback. The present study explores whether somatosensory feedback alone can help with the compensation. The experiments involve lip-tube perturbation on Taiwan Mandarin rounded vowels /u/ and /y/ with auditory-somatosensory feedback and somatosensory feedback only. The acoustic results show that speakers compensate for the lip-tube perturbation in F2 and F3 in both feedback types. Tongue postures do not change significantly when either feedback type is provided in response to the perturbation, suggesting that tongue may not necessarily be a main articulator in contribution to the compensation but other articulators may be much more important. Generally, our results show that somatosensory feedback may play a fundamental role in compensation for lip-tube perturbation.

Keywords: feedback types, somatosensory feedback, lip-tube perturbation, internal model, lip-rounding

1. INTRODUCTION

In speech production, two major hypotheses have focused on two different aspects. One claims that speech targets are acoustically defined and regulated by auditory feedback (e.g., [1]). The other, e.g., Articulatory Phonology [2], posits that speech targets are articulatorily defined and composed of articulatory gestures. During speech production, feedback plays an important role, and the subsystems of the internal model [3]—inverse model and forward model—contribute to different aspects in speech motor control. The inverse model is associated with the muscular composition for a particular motor behaviour whereas the forward model is responsible for predictions, i.e., the expected outcomes of the planned action. The predicted outcome of the movement and the motor commands then will be checked to see if they correspond. If so, the motor commands are issued to the muscles. If there is a mismatch, then an error signal is generated to modify the motor commands. Therefore, sensory feedback used here is considered as closed loop, and sensory feedback monitors intended movements and reduces errors if any. Such a process is like an on-line monitoring system.

Lip-tube perturbation used a lip-tube with 2cm in diameter (to prevent lip-rounding of narrower lip areas) held between the lips with one end against the incisors to perturb lip-rounding. The function of the lip-tube was to create wider lip areas in which is not favourable to lip-rounding gestures. The experimental design aimed to observe whether speakers can still produce the target rounded vowels while lips are prohibited to create narrower lip areas and what kind of articulatory compensations (tongue movements) they made in respond to the lip-tube perturbation. Lip-tube perturbation in previous studies [4, 5] has focused on the compensation capacity and how speakers use the internal model to adapt to the lip-tube perturbation, but the role of feedback in guiding the speakers to use the internal model to reach speech targets has not been carefully examined in previous lip-tube perturbation studies. Several perturbation studies focusing on feedback have demonstrated that without the auditory feedback, somatosensory feedback alone can still constitute speech goal [6], and auditory feedback is not dominant—both auditory feedback and somatosensory feedback can be beneficial to the speakers [7]. However, it is yet to be determined that whether the observed results from previous lip-tube perturbation studies [4, 5] are consequences of reinforcement from multisensory feedback, i.e., speakers are accessible to multiple sensory feedback, so that they might have more accessible resources in which some are potentially beneficial to them. In this case, the auditory-somatosensory feedback received by the speakers in previous lip-tube perturbation studies might somehow bias the results which might simply be the consequence of changed acoustics. Thus, the present study aims to explore whether somatosensory feedback alone can help with the compensation.

2. METHODS

2.1. Participants

Participants were four native Taiwan Mandarin speakers (balanced gender, aged 22–27). The participants did not report any history of auditory abnormality or speech production disorder.

2.2. Stimuli and procedure

The stimuli used in the experiment were two Mandarin rounded vowels /u, y/ with four lexical tones, yielding eight monosyllabic words. The experimental procedure closely followed by [4, 5] with some modifications on the feedback types received by the speakers. Two feedback types were focused in the present study: auditory-somatosensory feedback and somatosensory feedback only. Each feedback type contained three conditions: baseline (10 repetitions), lip-tube perturbation (LP, 40 repetitions), and post-perturbation (PP, 10 repetitions). In lip-tube perturbation, a lip-tube of 2cm in diameter (to prevent lip-rounding of narrower lip areas) and 2cm in length (chosen to avoid lengthening the labial constriction) was held between the lips with one end against the incisors (see Fig. 1. left). During baseline and post-perturbation, no lip-tube was inserted.

The experiment, different from previous studies [4, 5], employed a within-subject design with auditory-somatosensory feedback and with somatosensory feedback only. When speakers receive auditory-somatosensory feedback, they can hear their own voice while at the same time feeling their own lips being perturbed and how they move their own tongues. When speakers receive somatosensory feedback only, their own voice was blocked and thus they were only accessible to the feelings of the perturbed lips and their own tongues. The order of receiving the two feedback types was counterbalanced. In somatosensory feedback only situation, two headphones were attached: the canal type in-ear headphone, playing background noise with *Pastoral Symphony,* and the bone conduction headphone, playing multi-talker babbling (see Fig. 1. right). The volume was adjusted to a level that the speakers could not hear their own voice. The participants were instructed to maintain the words pronounced as consistently and accurately as possible. To observe tongue postures and any tongue movements as articulatory compensations in response to the lip-tube perturbation, a portable ultrasound machine (CGM OPUS 5100) was employed to capture images in real time through a transvaginal electronic curved array probe (CLA 651). Speakers were instructed to sit upright wearing an ultrasound stabilization headset [8]. Acoustics and ultrasound images were recorded simultaneously throughout the experiment.

2.3. Data analysis

In terms of acoustic data, formants (F1, F2, and F3) were obtained at 20 duration-normalized timepoints through Praat[9] and were submitted to generalized additive mixed models (GAMMs) [10]. Ultrasound videos were first automatically traced by DeepEdge [11] for tongue contours and then were manually checked and corrected in GetContours [12]. Each tongue contour with 100 flesh points were then generated. Tongue contours were obtained from the midpoint of the vowel and were submitted to GAMMs [10]. Significant acoustic differences between lip-tube perturbation and baseline across the normalized time are defined as failed compensation. If speakers show successful compensation, there should be less significant differences.

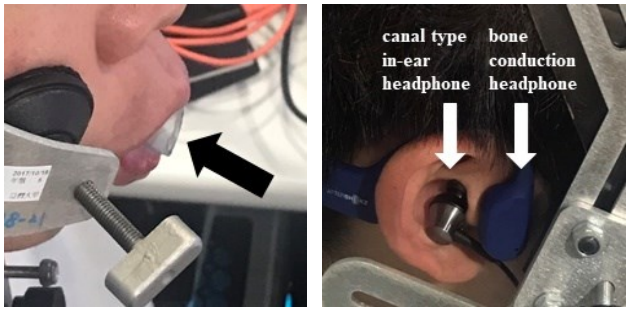


Figure 1: Experimental setup: lip-tube insertion (left) and headphones placement (right).

3. RESULTS

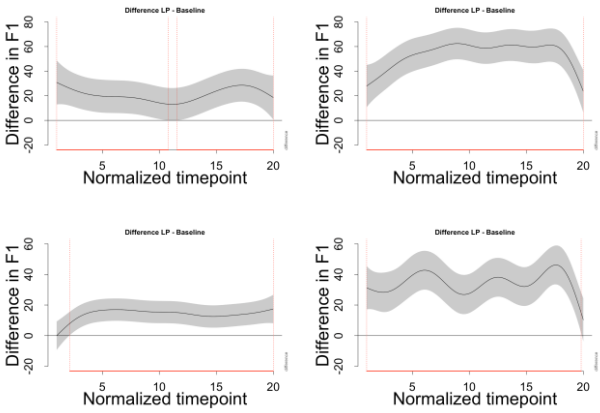


Figure 2: F1 differences of /u/ (upper) and /y/ (lower) between LP and Baseline when participants received auditory-somatosensory feedback (left) and somatosensory feedback only (right). Significant differences were enclosed by red dotted lines.

In both feedback types, significant differences of acoustics in terms of the three formants between post-perturbation (PP) and baseline were not found, suggesting that there is no hypercorrection or after-effect from perturbation. F1 differences between lip-tube perturbation (LP) and baseline for both rounded vowels with auditory-somatosensory feedback and somatosensory feedback only were all significantly different (Fig. 2.), suggesting that speakers cannot compensate for lip-tube perturbation in the dimension of F1 no matter which kinds of feedback are given.

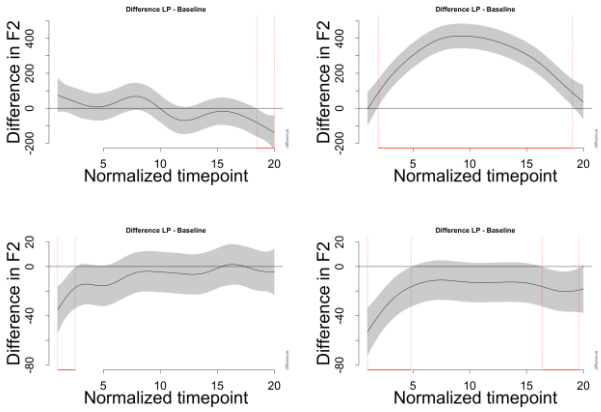


Figure 3: F2 differences of /u/ (upper) and /y/ (lower) between LP and Baseline when participants received auditory-somatosensory feedback (left) and somatosensory feedback only (right). Significant differences were enclosed by red dotted lines.

In terms of F2 differences, in the case of /u/, successful compensation was only found when participants were provided with auditory-somatosensory feedback (Fig. 3. upper left) but not with somatosensory feedback only (Fig. 3. upper right). Differences between LP and baseline revealed that F2 of /u/ produced with auditory-somatosensory feedback was more consistent in general (as evidenced by limited differences in Fig. 3. upper left) while F2 of /u/ produced with somatosensory feedback only showed significant differences enclosed by the red dotted lines (Fig. 3. upper right).

In the case of /y/, successful compensations were found in both auditory-somatosensory feedback (Fig. 3. lower right) and somatosensory feedback only (Fig. 3. lower left). In both cases, speakers tried to maintain F2 even when lip-tube was inserted, thus only a minority part of the trajectory (towards either one end or two) showed significant differences between LP and baseline.

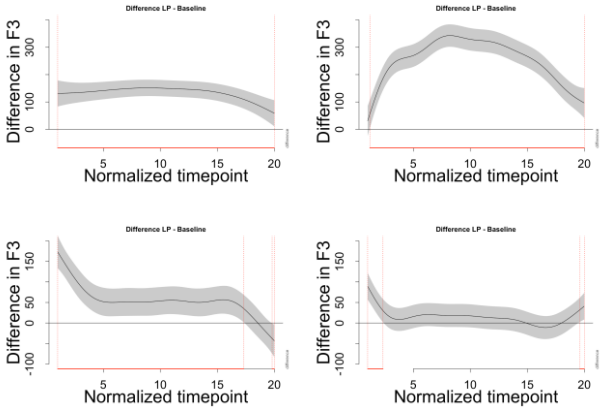


Figure 4: F3 differences of /u/ (upper) and /y/ (lower) between LP and Baseline when participants received auditory-somatosensory feedback (left) and somatosensory feedback only (right). Significant differences were enclosed by red dotted lines.

In terms of F3 differences, in the case of /u/, successful compensation was not found when both feedback types were given to the participants (Fig. 4. upper). F3 differences between lip-tube perturbation (LP) and baseline for /u/ with auditory-somatosensory feedback and somatosensory feedback only were all significantly different, showing that speakers cannot compensate for lip-tube perturbation for /u/ in the dimension of F3 no matter which kinds of feedback are given.

In the case of /y/, successful compensation was not found when receiving auditory-somatosensory feedback (Fig. 4. lower left) but when receiving somatosensory feedback only (Fig. 4. lower right). In Fig. 4. lower left, significant differences enclosed by rot dotted lines showed the majority part of the F3 trajectory was largely modulated. However, in Fig. 4. lower right, only very limited of the F3 trajectory was affected, suggesting the production remained intact.

For the articulatory results, tongue postures of /u/ no matter which kinds of feedback (Fig. 5. and 6.) are given did not change significantly in response to the perturbation (LP) or after the perturbation (PP). Also, the overall tongue postures when receiving auditory-somatosensory feedback (Fig. 5.) are similar to the tongue postures when receiving somatosensory feedback only (Fig. 6.).

Similarly, tongue postures of /y/ no matter which kinds of feedback (Fig. 7. and 8.) are given did not change significantly in response to the perturbation (LP) or after the perturbation (PP), and the overall tongue postures when receiving auditory-somatosensory feedback (Fig. 7.) are similar to the tongue postures when receiving somatosensory feedback only (Fig. 8.).

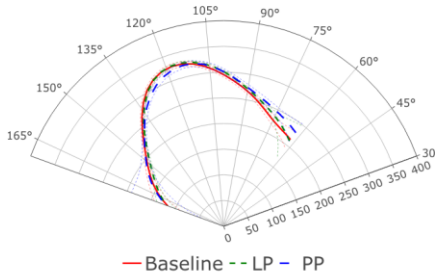


Figure 5: Tongue postures of /u/ when participants received auditory-somatosensory feedback. Tongue tip on the right. LP: lip-tube perturbation; PP: post-perturbation.

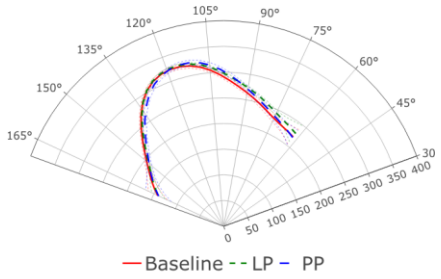


Figure 6: Tongue postures of /u/ when participants received somatosensory feedback only. Tongue tip on the right. LP: lip-tube perturbation; PP: post-perturbation.

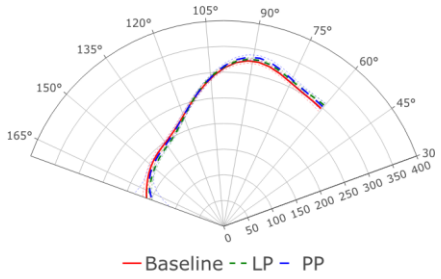


Figure 7: Tongue postures of /y/ when participants received auditory-somatosensory feedback. Tongue tip on the right. LP: lip-tube perturbation; PP: post-perturbation.

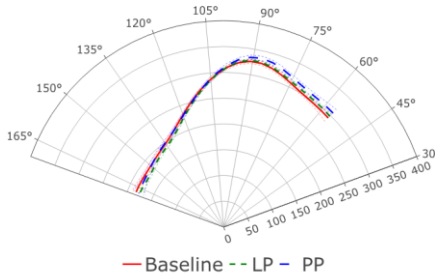


Figure 8: Tongue postures of /y/ when participants received somatosensory feedback only. Tongue tip on the right. LP: lip-tube perturbation; PP: post-perturbation.

4. DISCUSSION

First, we observe that the successful compensations are made in different formant dimensions for the two rounded vowels: F2 for /u/; F2 and F3 for /y/. This may suggest that the compensations made for the two rounded vowels focus on different dimensions.

Second, compensations can be successfully made not only with auditory-somatosensory feedback (F2 for both /u/ and /y/) but also with somatosensory feedback only (F2 and F3 for /y/). This echoes previous findings [6, 7] that somatosensory feedback alone can constitute speech goals. Also, we've found additional evidence for the claim that auditory feedback may not be dominant as in the case of F3 in /y/, where successful compensation was only found in the condition with somatosensory feedback only. This suggests that somatosensory feedback itself may be independent of auditory feedback and influential to perturbation compensation. In response to the two major hypotheses in speech production, the results show that somatosensory feedback alone (F2 and F3 for /y/) can still constitute speech goal and auditory feedback (F2 for /u/; F2 for /y/) might also play an equal role in constituting speech goal, and thus supporting both hypotheses.

Finally, no significant tongue movements were found across the tongue postures in response to the perturbation (LP) or after the perturbation (PP). Backward tongue movements reported in previous studies [4, 5] could be language-specific or due to the methods of dealing with articulatory data, where they all studied French and the articulatory data are not quantified and visualized as tongue postures but numerical transformations. This suggests that tongue may not necessarily be the main articulator in contribution to the compensation of lip-tube perturbation, but other articulators may be of more contribution. Another possible explanation is that lip-rounding may be organized by the coordination of different orofacial muscles which can be adjusted locally to reach the same global effect. Whether the lip-tube perturbation which constrains the lip areas is essentially related to the activation of the orofacial muscles would call for future studies.

5. CONCLUSION

The present study shows that somatosensory feedback itself can be independent of auditory feedback and it may play a fundamental role in compensation for lip-tube perturbation. Also, it does not necessarily mean that either auditory feedback or somatosensory feedback is more dominant, but rather they play in a parallel and relative relationship.

5. REFERENCES

[1] Jones, J. A., Munhall, K. G. 2000. Perceptual calibration of F_0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

[2] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180.

[3] Wolpert, D. M., Ghahramani, Z., Jordan, M. I. 1995. An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

[4] Savariaux, C., Perrier, P., Orliaguet, J. P. 1995. Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428–2442.

[5] Ménard, L., Perrier, P., Aubin, J. 2016. Compensation for a lip-tube perturbation in 4-year-olds: Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *The Journal of the Acoustical Society of America*, *139*(5), 2514–2531.

[6] Tremblay, S., Shiller, D. M., Ostry, D. J. 2003. Somatosensory basis of speech production. *Nature*, *423*(6942), 866–869.

[7] Lametti, D. R., Nasir, S. M., Ostry, D. J. 2012. Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience*, *32*(27), 9351–9358.

[8] Articulate Instruments Ltd. 2008. *Ultrasound Stabilisation Headset User's Manual: Revision 1.4* (Articulate Instruments Ltd, Edinburgh, UK).

[9] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.51.

[10] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116.

[11] Chen, W. R., Tiede, M., Whalen, D. H. 2020. DeepEdge: automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. Paper presented at the 12th International Seminar on Speech Production (ISSP 2020).

[12] Tiede, M., Whalen, D. H. 2015. Getcontours: An interactive tongue surface extraction tool. *Proceedings of Ultrafest VII*.

EFFECT OF INDIVIDUAL VOCAL TRACT GEOMETRY ON THE TONGUE SHAPING FOR AMERICAN ENGLISH /ɹ/

Yijing Lu, Haley Hsu, Louis Goldstein, Asterios Toutios

University of Southern California
yijinglu@usc.edu, haleyhsu@usc.edu, louisgol@usc.edu, toutios@gmail.com

ABSTRACT

Individual differences in the geometry of vocal tract structures have been found to correlate with inter-speaker articulatory variations. However, there has been a lack of agreement in whether the tongue shaping for American English /ɹ/ is influenced by vocal tract geometry. Different ways to quantify /ɹ/ tongue shapes might have contributed to the divergence of previous findings. The current study compares the results of regressing three different tongue shaping measurements on the same set of vocal tract geometry measurements. Three tongue shaping variables show marginally to moderately significant linear relationships with two specific parameters of vocal tract geometry: the degree of mandibular inclination and the horizontal length of the oral cavity. But overall, there is no strong effect of individual vocal tract geometry on the tongue shaping for /ɹ/.

Keywords: articulation, anatomy, tongue shaping, inter-speaker variability, vocal tract MRI

1. INTRODUCTION

The peripheral morphology of the human vocal tract seems to play a key role in speech production. Rigid structures such as the palate, the posterior pharyngeal wall and the mandible delimits the space in which articulation takes place. Hence, the geometric properties of these vocal tract structures are expected to influence the patterns of deformation of the tongue during speech production. Individual speakers appear to adjust their articulatory strategies according to the morphology of vocal tract structures, possibly in order to generate acoustic consequences that are close enough to be perceived as the same phoneme [1].

A number of studies have examined the relationship between vocal tract geometry and inter-speaker articulatory variability in the production of various speech sounds. For example, the vertical displacement of tongue body from /a/ to /i/ was found to be negatively correlated with the pharyngeal distance (the distance from the anterior nasal spine to the pharyngeal wall) [2] and positively correlated with the pharynx length [3]. The apical and laminal variants of /s/ [4] were found to be partially conditioned by the speaker's palatal height [5, 6] and the slope of the anterior palate [7]. As for American English /ɹ/, which exhibits considerable variation in tongue shaping from "bunched" to "retroflex" [8], different findings have been reached as to whether the tongue shapes are influenced by vocal tract morphology. Westbury and colleagues [9] found no relationship between the /ɹ/ tongue shapes and various measurements of oral cavity size, whereas Dediu and Moisik [10] found that /ɹ/ tongue shapes were influenced by the width and height of the hard palate, the overall size of the mouth, and the prominence of the alveolar ridge. Another study [11] found that speakers using retroflex /ɹ/ and bunched /ɹ/ did not show significant difference in their palate doming degrees.

Given that these previous studies have used different quantitative or qualitative methods to characterize /ɹ/ tongue shapes, these methods may have captured different aspects of the tongue shaping, and therefore may have played a role in the divergent findings. In this paper, we examine the effect of vocal tract morphology on three different measurements of /ɹ/ tongue shapes. To do so, we compare the regression results using different tongue shaping measurements as the dependent variable and the same vocal tract geometry measurements as the independent variables.

2. METHOD

2.1. Data

The data analysed in the present study are a subset of the data in a publicly available multi-speaker MRI corpus [12]. Readers are referred to that work for full technical details on data collection and curation. Data from 32 native speakers of American English from various rhotic-dialect regions of the US (16 female, 16 male; mean age = 26; age range = 18–59) were used. The data consisted of midsagittal RT-MRI speech videos showing the vocal tract movements during two repetitions of /ɑɹɑ/ and high-resolution static anatomical T2-weighted MRI of the upper airway.

2.2. Articulatory analysis

A semi-automatic method [13] was used to segment tissues from airway and track the contours of articulators in the RT-MRI frames. The target frames of /ɹ/ that underwent further analysis were frames in which the distance between the tongue and the palate is the shortest, i.e., the palatal constriction degree presents a minimum.

On the selected /ɹ/ frames, 20 gridlines perpendicular to the line connecting the start and end points of the hard palate were superimposed, intersecting the contours of the palate and the tongue (Fig. 1). The distances between intersection points on the palate and the tongue (red lines in Fig. 1) were measured as the aperture function representing the deformation of the area between the tongue and the palate. Skewness and kurtosis of the aperture function were calculated and used as the first two tongue shape measurements in this study. Since skewness and kurtosis reflect the degrees of asymmetry and tail extremity in the shape of a distribution, these two measurements approximately indicate the location and the length of the constriction formed in the tongue-palate area. A negative skewness means that the shape leans towards the right, therefore representing a relatively posterior constriction location. Conversely, a positive skewness represents a relatively anterior constriction location. A larger kurtosis means a greater amount of deviation and therefore a shorter constriction length. On the other hand, a smaller kurtosis means a longer constriction.

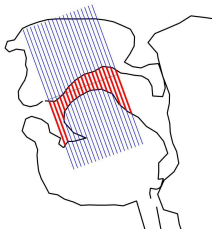


Figure 1: A segmented /ɹ/ frame with 20 superimposed gridlines. The red part of the gridlines constitutes the aperture function that represents the area between the tongue and the hard palate.

Additionally, each /ɹ/ frame was rated as either bunched or retroflex by three human raters. Linear support vector machine (SVM) was used to perform a classification experiment based on the skewness and kurtosis measurements and the human rating results (aggregated using the majority rule). 10 out of 64 /ɹ/ tokens were misclassified. The classification error rate was 15.62%. As shown in Fig. 2, retroflex /ɹ/ and bunched /ɹ/ are largely separated by the decision boundary. The signed distances from each token of /ɹ/ to the decision boundary were calculated as the classification score for each /ɹ/. The classification score is the third measurement of tongue shape in this study, which can be thought of as a measurement combining skewness and kurtosis, indicating a degree of "bunchedness" or "retroflexion".

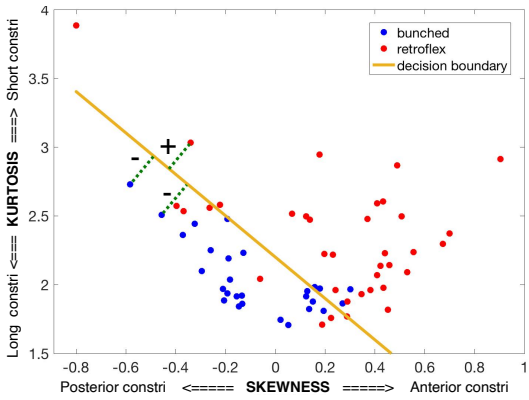


Figure 2: Skewness and kurtosis for each token of /ɹ/, which was colored according to the human rating results. The decision boundary (yellow line) was determined by a linear SVM. The green dotted lines are examples of the distances from /ɹ/ tokens to the decision boundary (classification scores). Tokens above the decision boundary were assigned positive classification scores, and tokens below were assigned negative classification scores.

2.3. Anatomical analysis

Measurements of vocal tract morphology were obtained using the method of [2] to define a geometrical space called Articulatory space (A-space) in the midsagittal vocal tract (Fig. 3). Using the softwere OsiriX Lite v12.0.0 (Pixmeo, Geneva, Switzerland), the anatomical landmarks (i.e., anterior nasal spine, posterior nasal spine, menton) were manually located, based on which the A-space was manually defined for each speaker. The measurements of the geometrical properties of the A-space were directly output from OsiriX. The top border of an A-space is a line that connects the anterior nasal spine (ANS) to the posterior nasal spine (PNS) and eventually stops at the posterior pharyngeal wall. This distance is called pharyngeal distance in [2]. It represents the length of the oral cavity roof. The bottom border is a line parallel to the top boarder, starting from the menton and extending rearward to the posterior pharyngeal wall. The anterior border is the line that connects the ANS to the menton. The posterior border is the outline of the posterior pharyngeal wall between the top and bottom borders. Apart from the four borders, the vertical distance between the top border and the bottom border and the angle between the anterior border and the bottom border were also measured. They respectively represent the lower facial height and the degree of inclination of the mandibular symphysis. Additionally, the ratio of the lower facial height to the

top border was calculated, representing the aspect ratio of the oral cavity. In total, seven anatomical measurements were taken for each speaker: lengths of the top, bottom, front, back borders of the A space, the lower facial height (LFH), the aspect ratio (AR), and the angle of the mandibular inclination (MI).

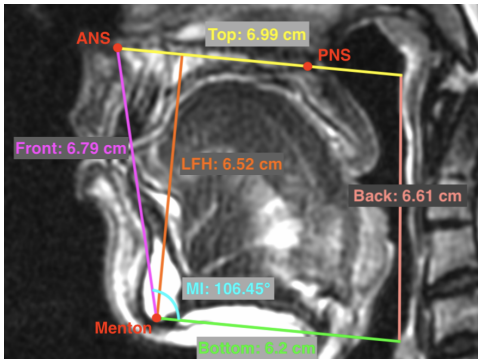


Figure 3: An example of A-space, relevant anatomical landmarks, as well as seven anatomical measurements shown on a T2 weighted midsagittal MRI slice.

3. RESULTS

First, single anatomical features were tested to see if any can form a predictive relationship with either one of the three tongue shape measurements. Simple linear regression models were fitted using each one of the anatomical measurements as the independent variable, and skewness, kurtosis or classification score as the dependent variable. When skewness was the dependent variable, mandibular inclination (MI) was the only predictor yielding a marginally significant relationship (see Fig. 4a, R^2 = 0.05974, $F(1, 62) = 3.939$, $p = 0.0516$). When kurtosis was used as the dependent variable, the length of the top border (top), which approximated the horizontal length of the oral cavity, was the only significant predictor (see Fig. 4b, $R^2 = 0.06722$, $F(1, 62) = 4.468$, $p = 0.03856$). When classification score was used as the dependent variable, mandibular inclination (MI) was the only predictor yielding a marginally significant relationship (see Fig. 4c, R^2 = 0.05873, $F(1, 62) = 3.868$, $p = 0.05369$).

Multiple linear regression models were also fitted to predict each one of the three tongue shape measurements using all the anatomical measurements. None of the models yield significant overall regression. The cumulative R^2 for each model are shown in Table 1.

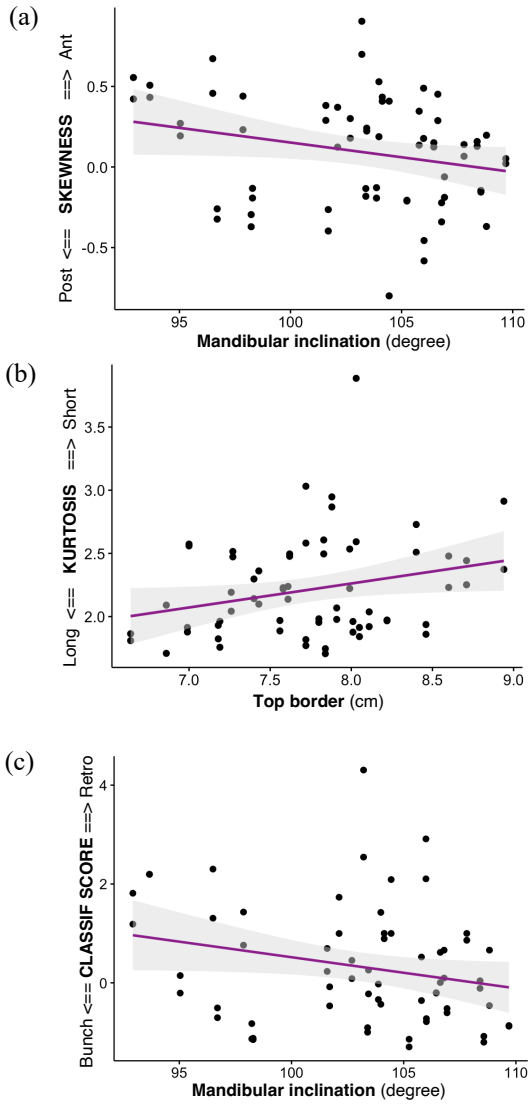


Figure 4: Scatterplots with each dot representing the values of (a) mandibular inclination degree and skewness, (b) top border length and kurtosis (c) mandibular inclination degree and classification score for each token of /ɹ/. The regression line (purple line) and the 95% confidence interval of the regression coefficients (grey area) are superimposed.

DV	R^2
Skewness	0.1033
Kurtosis	0.1532
Classification score	0.1387

Table 1: R^2 for the regression models using different tongue shaping measurements as the dependent variable (DV) and all the anatomical measurements as the independent variables.

4. DISCUSSION

Overall, we did not find a strong effect of vocal tract geometry on the tongue shapes that individual speakers use during the production of American English /ɹ/. For all three quantitative measurements that characterize different aspects of the tongue shape, multiple linear regression models with all the anatomical features as predictors all yield non-significant relationships and low R^2 values (Table 1).

But some marginally significant relationships emerge when certain anatomical features were used individually as the single predictor. Skewness, which characterizes the location of the palatal constriction achieved during the production of /ɹ/, has a weak negative association with the degree of mandibular inclination. A smaller degree of mandibular inclination results in a greater value of skewness, which means a more anterior constriction location. This echoes the findings about /t/ in [2]: English speakers with a smaller degree of mandibular inclination tended to show an apico-dental articulation for /t/, whereas speakers with a relatively larger degree of mandibular inclination tended to use lamino-alveolar /t/. The constriction location for apico-dental /t/ is more anterior compared to lamino-alveolar /t/. For speakers with smaller mandibular inclination, their tongue tends to sit further forward with respect to the palate. Therefore, it is natural for them to make a more anterior coronal constriction.

Kurtosis, which characterizes the length of the palatal constriction, has a positive association with the top border of A-space, which represents the length of the oral cavity roof. A longer oral cavity results in a greater value of kurtosis, which means a shorter constriction length. This is consistent with one of the findings in [10]: the tongue shape for /ɹ/ was significantly related to the length of the anterior mouth. When the mouth is longer, which means the oral cavity is longer, the tongue posture is more likely to have an apical stricture, which means a shorter constriction length. [14] found that for speakers whose palatal constriction degree contributed to F3 lowering more than pharyngeal or labial constriction, they preferred tongue postures with a shorter and more anterior palatal constriction. It is likely that this preference is linked to some morphological feature, i.e., a longer oral cavity. This needs to be further tested.

Classification score, which indicates a degree of "retroflexion" or "bunchedness", was negatively influenced by mandibular inclination. Speakers with a smaller degree of mandibular inclination (prognathic speakers) are more likely to produce "retroflex-flavored" /ɹ/. Prognathic speakers, compared to retrognathic speakers (speakers with a larger degree of mandibular inclination), have a relatively larger sublingual space [2], which may help the elevation of the tongue tip to form a retroflex posture.

In the current study, we characterize the /ɹ/ tongue shapes only using the properties of the palatal constriction. Future work should include measurements of the pharyngeal constriction, which is another factor that can contribute to the difference in tongue shaping. Also, the articulation of /ɹ/ in only one phonetic context was analyzed here. Future research will need to include more phonetic contexts and investigate the interaction between vocal tract morphology and phonetic context in determining the articulatory strategies of individual speakers. If the articulatory variation of /ɹ/ is truly marginally attributable to vocal tract morphology, as suggested by the current study, it is possible that this variation stems from individual differences in the cognitive representations of phonological units [15].

5. REFERENCE

[1] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker Variability in Hard Palate Morphology and Vowel Production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, Dec. 2013.

[2] K. Honda, S. Maeda, M. Hashi, J. Dembowski, and J. R. Westbury, "Human palate and related structures: their articulatory consequences," in *Proceeding of 4th International Conference on Spoken Language Processing*, vol. 2. IEEE, 1996, pp. 784–787.

[3] S. Fuchs, R. Winkler, and P. Perrier, "Do speakers' vocal tract geometries shape their articulatory vowel space?" in *Proceeding of 8th International Seminar on Speech Production*, 2008, pp. 333–336.

[4] N. Dart, *Articulatory and Acoustic Properties of Apical and Laminal Articulations*. University of California, Los Angeles, 1991.

[5] M. Stone, S. Rizk, J. Woo, E. Z. Murano, H. Chen, and J. L. Prince, "Frequency of apical and laminal/s/in normal and postglossectomy patients," *Journal of Medical Speech-Language Pathology*, vol. 20, no. 4, 2012.

[6] D. L. Grimm et al., "The Effects of Palate Features and Glossectomy Surgery on /s/ Production," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, pp. 3417–3425, Dec. 2017.

[7] M. Stone, A. D. Gomez, J. Zhuo, A. L. Tchouaga, and J. L. Prince, "Quantifying tongue tip shape in apical and laminal /s/: contributions of palate shape," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 9, pp. 3149–3159, Sep. 2019.

[8] P. Delattre and D. C. Freeman, "A dialect study of American *r*'s by x-ray motion picture," *Linguistics*, vol. 6, no. 44, 1968.

[9] J. R. Westbury, M. Hashi, and M. J. Lindstrom, "Differences among speakers in lingual articulation for

American English /ɹ/," *Speech Communication*, vol. 26, no. 3, pp.203−226, 1998.
[10] D. Dediu and S. R. Moisik, "Pushes and pulls from below: Anatomical variation, articulation and sound change," *Glossa: a journal of general linguistics*, vol. 4, no. 1, Jan. 2019.
[11] S. Bakst, "Differences in the relationship between palate shape, articulation, and acoustics of American English /r/ and /s/," *UC Berkeley Phonology Lab Annual Report*s, vol. 12, 2016.
[12] Y. Lim et al., "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *Scientific Data*, vol. 8, no. 1, p. 187, Jul. 2021.
[13] E. Bresch and S. Narayanan, "Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, Mar. 2009.
[14] S. Harper, L. Goldstein, and S. Narayanan, "Variability in individual constriction contributions to third formant values in American English /ɹ/," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 3905–3916, Jun. 2020.
[15] S. Harper, "Individual differences in phonetic variability and phonological representation," Ph.D. dissertation, University of Southern California, Los Angeles, 2021.

Individual- and group-level associations between articulatory and acoustic variability for English consonants

Sarah Harper

University of California, San Francisco
sarah.harper@ucsf.edu

ABSTRACT

This study extends previous research probing the relationship between articulatory and acoustic variability by examining the extent to which individual differences in articulatory variability are recoverable from acoustics, and vice versa, for a subset of American English consonants. Articulatory (constriction location, degree, and orientation) and acoustic (formant and spectral) measurements were extracted from tokens of /s/, /ʃ/, /l/, and /ɹ/ for 40 speakers in the Wisconsin X-Ray Microbeam Corpus. Analysis of goodness-of-fit statistics from linear mixed effects models fit to the data suggest that variability observed in articulation is generally recoverable in acoustics, and vice versa, with change in one domain consistently predicted by change in the other. Additionally, the comparison of dispersion metrics from raw and model-predicted values for each speaker indicate that interspeaker differences in variability are to some extent transmissible between articulation and acoustics, although this capacity varies substantially across segments and across measured dimensions.

Keywords: Articulatory-Acoustic Relations, Variability, Individual Differences

1. INTRODUCTION

A substantial body of research on articulatory-acoustic relations in speech production has examined the extent to which variation in articulation is encoded in acoustics (and vice versa). Early research in this area generally suggested that the non-linear and quantal mapping observed between articulation and acoustics [1] would lead to a dissociation of the variability observed in the realization of phonological segments across these modalities. Specifically, it was both predicted and observed that the articulatory patterns used to produce any phonological segment would be more variable than the resulting acoustic signal [2, 3].

Recent work on articulatory-acoustic relations in vowel production, however, has cast doubt upon the existence of a consistently asymmetric relationship between articulatory and acoustic variability, and upon the perspective that variation in articulation is obscured in the acoustic signal more generally. Examining individual differences in the realization of vowel height contrasts, Noiray et al. [4] found that speakers' idiosyncratic articulatory strategies for contrasting front vowel pairs were reflected in acoustics, with speakers who showed articulatory tongue height 'reversals' (i.e., producing a high vowel with a lower tongue position than a mid vowel) demonstrating a comparable reversal in first formant values. Whalen et al. [5] similarly found that articulatory variability was positively correlated with acoustic variability across vowels for speakers of American English, and that speakers differed from one another in whether they exhibited more variability in articulation or acoustics for the same vowel segments.

In this study, we extend previous research on articulatory-acoustic relations by examining whether patterns of variability in consonant production are recoverable between articulation and acoustics. Although studies support the idea that variability is, at least for vowels, relatively comparable in articulation and acoustics [4, 5], the articulatory constrictions used in consonant production could interact with the nonlinearities of vocal tract acoustics in a manner that minimizes acoustic variability (as suggested by [6]) and obscures the transmission of variability between these domains. As such, the extent to which a similarly comparable relationship between articulatory and acoustic variability exists for consonants remains unclear.

2. METHOD

2.1. Corpus and data selection

Articulatory and acoustic data for this study was taken from recordings of 40 American English speakers in the Wisconsin x-ray Microbeam (XRMB) database [7]. The kinematic articulatory data in the XRMB corpus consists of positional trajectories of the movement of small pellets (2.5 mm) attached to the tongue, jaw, and lip and to stable reference points in the mouth (Fig. 1). Acoustic data consists of audio recordings captured synchronously with the kinematic data.

A total of 7,298 tokens of word-initial and -final /s/, /ʃ/, /l/, and /ɹ/ were extracted from sentences read by participants in two of the corpus tasks, a Sentence Reading task and a Prose Passage Reading task. These consonants were chosen for analysis both because they are known to exhibit substantial

variability in articulatory posture across and within speakers in American English ([8]-[12]) and because they are continuants with information about their identity transmitted acoustically throughout their production.

2.2. Articulatory measurement

The gestural movement extremum (**MAXC**) in each token was automatically identified using a modified version of the *findgest* algorithm (originally by Mark Tiede, Haskins Laboratories). Using an acoustic segmentation of the XRMB data created with the Penn Forced Aligner [13], tokens of the consonants of interest were automatically located in the articulatory data by finding articulatory frames corresponding to their acoustically defined segment start and end points. These frames defined a search window over which the algorithm looked for MAXC, defined as the velocity minimum closest to the token's acoustic midpoint. For tokens of /s/, /ʃ/, and /l/, MAXC was extracted using the 2D tangential velocity signal of the lingual pellet closest to the tongue tip (T1). For /ɹ/, MAXC was extracted for the lingual pellet calculated to have the smallest distance from the palate at its velocity minimum (either T1, T2 or T3).

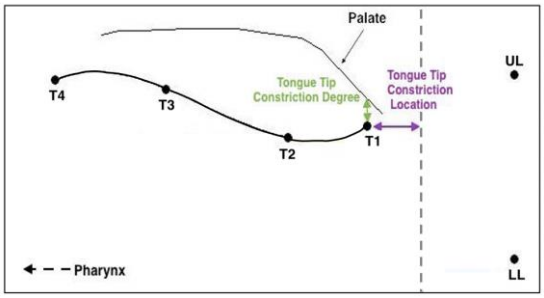


Figure 1: Position of XRMB pellets with palate trace and schematic representation of constriction measurements. Dotted vertical line indicates the origin of the coordinate system used for location measurement.

Positional measurements of the upper and lower lips and the tongue were extracted for all lingual and labial pellets in the XRMB data at MAXC (Fig. 1). Constriction location (**CL**) and degree (**CD**) were logged for the pellet closest to the palate trace for each token (T1 for all tokens of /s/, /ʃ/, and /l/, and either T1, T2, or T3 for /ɹ/). CL was calculated as the x-axis distance of the pellet of interest from the coordinate system origin (at the tip of the maxillary incisors) and normalized by the length of the speaker's vocal tract. CD was calculated as the 2D Euclidean distance between the pellet's coordinate position and the closest point on the speaker's palate trace.

2.3. Acoustic measurement

The acoustic dimensions chosen for analysis differed between the examined liquids (/l/ and /ɹ/) and the fricatives (/s/ and /ʃ/). For the liquids, values of the first four formants (**F1-F4**) were automatically extracted at MAXC or, when MAXC occurred outside of the voicing interval for the target segment, at the closest voiced interval to this timepoint (following [14], [15]). Formant tracking was configured to find five formants below 5000 Hz for male speakers and 5500 Hz for female speakers. For fricatives, the first four spectral moments (**M1-M4**) were calculated using a DFT over a 50 ms Hamming window centered on MAXC. A high-pass filter with a cut-off at 500 Hz was applied to the fricative spectrum prior to spectral moment calculation to exclude spectral energy reflecting vocal fold vibration.

Tokens with one or more formant or spectral moment more than 2.5 standard deviations outside the speaker mean were visually inspected and manually corrected in Praat [16].

2.4. Statistical analysis of articulatory-acoustic relations

To directly examine relationships between articulatory and acoustic dimensions for each segment, a series of linear mixed effects models (LMMs) were fit to the data using the lmerTest package in R [17]. For each consonant of interest, two models were fit with an articulatory dependent variable (CD or CL) and the full set of formant or spectral moment measurements as fixed factors, and four models were fit with an acoustic dependent variable (F1-F4 or M1-M4) and the full set of XRMB pellet positions as fixed factors[1]. Random intercepts for Speaker and Phonetic Context[2] were included in all models. These models were used to evaluate (1) how well variation in articulation or acoustics was explained by the other domain across the group of speakers, and (2) the extent to which individual differences in variability are conveyed between articulation and acoustics.

Group-level encoding of variation between articulation and acoustics was evaluated using marginal and conditional R^2 values ($\mathbf{R^2_M}$ and $\mathbf{R^2_C}$) calculated for each LMM using the R package MuMIn [18]. R^2_M represents the variance explained by the fixed effect structure of a model, while R^2_C represents the variance explained by the fixed and random effect structures combined. An additional metric representing the variance explained by the random effects structure alone ($\mathbf{R^2_R}$) was calculated as the absolute difference between R^2_M and R^2_C.

An analysis comparing patterns of variability in model-predicted values to those observed in raw data was conducted to evaluate how well interspeaker

differences in variability are conveyed between articulation and acoustics. The predicted values for each datapoint in each LMM were calculated using the *predict* function in the lme4 package in R [19]. The interquartile range of each model's predicted values (**IQR_{PRED}**) and the raw XRMB corpus data used as its dependent variable (**IQR_{REAL}**) were calculated separately for each speaker and compared with Spearman's rank-order correlation.

3. RESULTS

3.1. Group-level encoding of variation

R^2_M and R^2_R for all LMMs fit to articulatory data are given in Table 1. R^2_M values are highest for /l/ in the model fit to CL and for /l/ and /ɹ/ in the model fit to CD, suggesting that multi-dimensional acoustic space tends to explain a larger proportion of variance along the examined articulatory dimensions for the liquid consonants. The opposite pattern is observed for R^2_R, with higher values generally observed for the fricatives than for the liquids. This indicates that individual differences in the articulatory-acoustic mapping across speakers have greater explanatory power in the prediction of articulation for the examined fricatives than for the examined liquids.

Table 1: R^2_M and R^2_R values for all LMMs fit to an articulatory dependent variable.

		/s/	/ʃ/	/l/	/ɹ/
R^2_M	**CL**	2%	10%	12%	8%
	CD	3%	10%	16%	22%
R^2_R	**CL**	85%	75%	44%	32%
	CD	56%	59%	15%	44%

R^2_M and R^2_R for all LMMs fit to acoustic data are given in Tables 2 and 3. Inspection of the R^2_M values for these models suggest that multi-dimensional articulatory space explains a higher proportion of variance in acoustic dimensions for /ɹ/ than it does for the other consonants examined (with the high R^2_M in the M1 model for /ʃ/ and F1 for /l/ notable exceptions). R^2_M values are generally higher than they were for the models fit to articulatory dependent variables, indicating that the acoustic variation produced by different articulatory vectors is somewhat more unique than the set of possible articulatory vectors suggested by a particular acoustic signal. This is consistent with the view that similar acoustic signals can result from different vocal tract configurations [20]. R^2_R values are also lower for the fricative consonants in this set of models than they were in the LMMs with articulatory dependent variables, potentially pointing towards lesser predictive power for individual differences in articulatory-to-acoustic mapping than in the acoustic-to-articulatory direction.

Table 2: R^2_M values for all LMMs fit to an acoustic dependent variable.

		/s/	/ʃ/		/l/	/ɹ/
R^2_M	**M1**	17%	31%	**F1**	31%	32%
	M2	13%	21%	**F2**	17%	28%
	M3	8%	17%	**F3**	3%	24%
	M4	3%	18%	**F4**	9%	10%

Table 3: R^2_R values for all LMMs fit to an acoustic dependent variable.

		/s/	/ʃ/		/l/	/ɹ/
R^2_R	**M1**	48%	48%	**F1**	32%	28%
	M2	49%	38%	**F2**	51%	40%
	M3	36%	48%	**F3**	54%	42%
	M4	6%	32%	**F4**	53%	57%

3.2. Encoding of individual differences in variability

Results of the correlation analyses of IQR_{REAL} and IQR_{PRED} for articulatory dimensions suggest IQR_{PRED} is usually a significant predictor of IQR_{REAL}, with a positive relationship consistently observed such that speakers who are more variable in their actual production tend to also have more variability in their predicted data (Fig. 2).

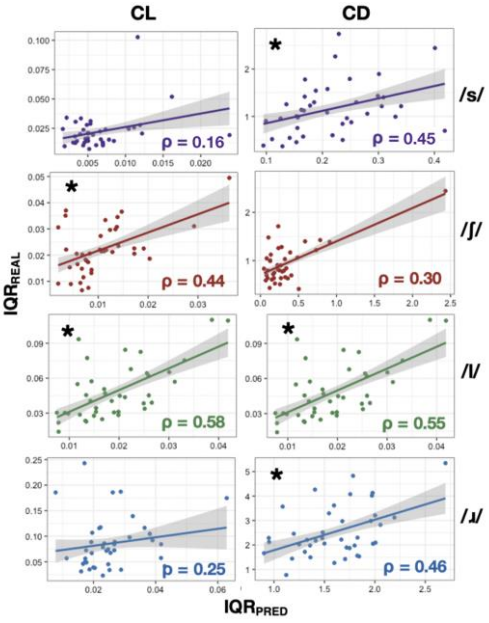


Figure 2: IQR_{REAL} and IQR_{PRED} comparisons across speakers for CD (left) and CL (right) in each consonant. Correlation coefficients are shown in the bottom right corner of each graph. Asterisk in the top left corner indicates significance.

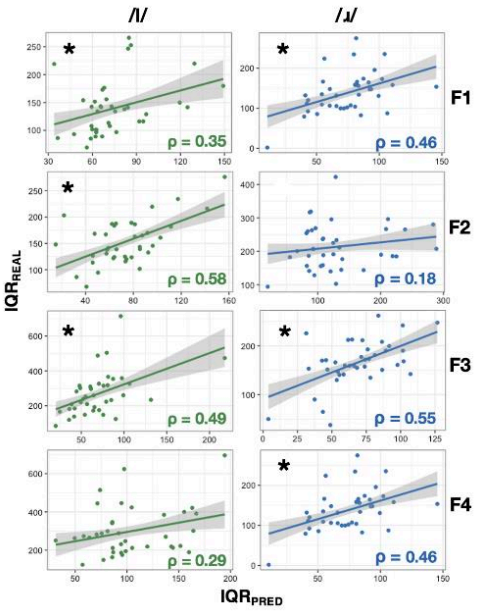


Figure 3: IQR_{REAL} and IQR_{PRED} comparisons across speakers for acoustic dimensions in liquid consonants.

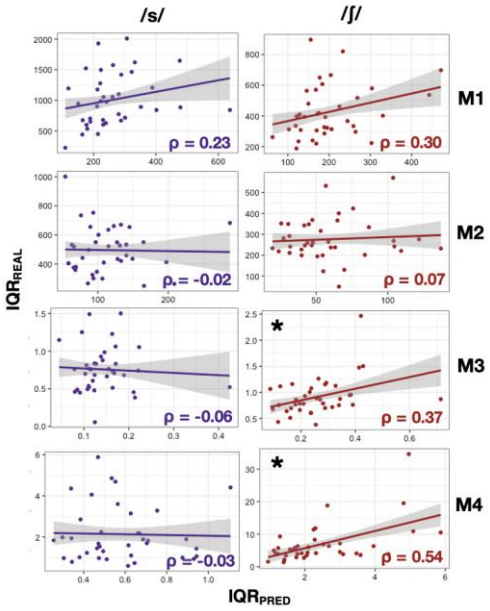


Figure 4: IQR_{REAL} and IQR_{PRED} comparisons across speakers for acoustic dimensions in fricative consonants.

The strength of the relationship between IQR_{PRED} and IQR_{REAL} varies substantially across segments and articulatory dimensions in the examined data, indicating that the recoverability of articulatory variability from the acoustic signal depends on which articulatory dimensions are being examined in which consonant segments. However, the observation that most individual comparisons (5 out of 8) are statistically significant suggests articulatory variability tends to be recovered from the acoustic signal across consonants and articulatory dimensions.

Correlation analyses of IQR_{REAL} and IQR_{PRED} for acoustic dimensions in /l/ and /ɹ/ similarly tend to exhibit a significant positive relationship (Fig. 3), with 6 out of 8 comparisons following this pattern. However, IQR_{PRED} is not as consistently a significant predictor of IQR_{REAL} for the fricative consonants (Fig. 4), with only 2 out of 8 comparisons significant across dimensions for /s/ and /ʃ/. These results indicate that the ability to recover acoustic variability from the articulatory signal may be quite sensitive to the identity of the consonant examined.

4. DISCUSSION

The results of this study confirm that both articulatory and acoustic variability are encoded in the other domain for consonants, although the strength of this relationship depends on the specific consonants and dimensions examined. Statistical models approximating articulatory-acoustic and acoustic-articulatory relationships in American English consonants were frequently able to use the variation observed in one domain to predict change in the other, as demonstrated by the prevalence of R^2_M values accounting for at least 10% of the variation in predicted dimensions. These models were also largely successful at predicting individual differences in variability along specific articulatory or acoustic dimensions, suggesting a fine-tuned ability to predict how components in the acoustic signal reflect changes in articulation and to associate specific changes in the acoustic signal with underlying articulatory actions.

The capacity for variability in articulation to be recovered from acoustics, and vice versa, differs across articulatory and acoustic dimensions and across consonants. However, there is a general tendency for variability to be encoded across articulation and acoustics, in line with recent research on articulatory-acoustic relations in vowels. These results suggest that in natural speech there is less of a dissociation between articulatory and acoustic variability in the production of a single consonant than has been suggested by previous research. This may indicate that the functional consequences of vocal tract nonlinearities on acoustic-articulatory relationships may be smaller than traditionally thought (e.g., [6]), and that considerable information about inter-and intraspeaker patterns of articulatory variability may be accessible to listeners in the acoustic signal. Further research will be necessary to explore this possibility and its potential implications for the cognitive systems underlying production and perception.

5. REFERENCES

[1] Stevens, K. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In: David, E., Denes, P.

(eds), *Human Communication: A Unified View*. McGraw Hill, 51-66.
[2] Johnson, K., Ladefoged, P., & Lindau, M. 1993. Individual differences in vowel production. *J. Acoust. Soc. Am.* 94(2), 701-714.
[3] Guenther, F., Husain, F., Cohen M., & Shinn-Cunningham, B. 1999. Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* 106, 2900-2912.
[4] Noiray, A., Iskarous, K., & Whalen, D. 2014. Variability in English vowels is comparable in articulation and acoustics. *Laboratory Phonology* 5(2), 271-288.
[5] Whalen, D., Chen, W., Tiede, M., & Nam, H. 2018. Variability of articulator positions and formants across nine English vowels. *J.Phon*. 68, 1-14.
[6] Nieto-Castanon, A., Guenther, F., Perkell, J., & Curtin, H. 2005. A modelling investigation of articulatory variability and acoustic stability during American English /r/ production. *J. Acoust. Soc. Am.* 117(5), 3196-3212.
[7] Westbury, J. 1994. *X-ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin, Madison, WI.
[8] Bladon, R., & Nolan, F. 1977. A video-fluorographic investigation of tip and blade alveolars in English. *J.Phon.* 5, 185-193.
[9] Dart, S. 1998. Comparing French and English coronal consonant articulation. *J.Phon.* 26, 71-94.
[10] Delattre, P., & Freeman, D. 1968. A dialect study of American r's by x-ray motion picture. *Linguistics* 44, 29-68.
[11] Lindau, M. 1985. The story of /r/. In: Fromkin, V. (ed), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*. Academic Press, 157-168.
[12] Mielke, J., Baker, A., & Archangeli, D. 2016. Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language* 92, 101-140.
[13] Yuan, J., & Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proc. Acoustics '08*, 5687-5690.
[14] Lin, S., Beddor, P., & Coetzee, A. 2014. Gestural reduction, lexical frequency, and sound change: A study of post-vocalic /l/. *Laboratory Phonology* 5(1), 9-36.
[15] Lawson, E., Stuart-Smith, J., & Scobbie, J. 2018. The role of gesture delaty in coda /r/ weakening: An articulatory, auditory and acoustic study. *J. Acoust. Soc. Am.*143, 1646-1657.
[16] Boersma, P., & Weenink, D. 2020. Praat: doing phonetics by computer (Version 6.1.16) [Computer software]. http://www.praat.org/.
[17] R Core Team. 2021. R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
[18] Barton, K. 2020. MuMIn: Multi-model inference. R package version 1.43.17. https://CRAN.Rproject.org/package=MuMIn.
[19] Bates, D., Maechler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* 67(1), 1-48.
[20] Atal, B., Chang, J., Mathews, M., & Tukey, J. 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.* 63(5), 1535-1555.

[1] The full set of pellet positions was used instead of derived articulatory positional measurements due to the potential for the entire vocal tract, not just the examined dimensions of interest, to affect the acoustic signal. Models using formant and spectral measurements as fixed effects were selected for analysis after observing that they outperformed models with MFCCs in the fixed effects structure.

[2] Phonetic Context defined as the segmental environment, word position (initial or final), and phrasal position (boundary-adjacent or phase-internal) of a token.

LIP POSTURES OF HIGH VOWELS IN TAIWAN MANDARIN

Chenhao Chiu[1, 2], Po-Hsuan Huang[1]

Graduate Institute of Linguistics, National Taiwan University[1]
Neurobiology and Cognitive Science Center, National Taiwan University[2]
chenhaochiu@ntu.edu.tw, r09142003@ntu.edu.tw

ABSTRACT

Taiwan Mandarin contrasts three high vowels with frontness ([i] and [y] being front) and roundedness ([y] and [u] being rounded). In particular, the [round] feature appears to provide rather simplistic interpretations of the lip postures for these high vowels as this feature is associated more with a narrowed aperture than with lip posture. Whether postural differences are articulatorily achieved to distinguish these high vowels from one another remains unanswered. This study investigates the lip postures among high vowels in Taiwan Mandarin, including aperture distance and area, axial ratio, and lip protrusion. Our results show that [u] and [y], though both traditionally labeled as rounded, contrast with each other in lateral distance between the mouth corners, yielding a more circular round posture for [u] and a more laterally compressed posture for [y]. Collectively, our results suggest that high vowels in Taiwan Mandarin are better distinguished along aperture area and lip posture.

Keywords: lip postures, lip aperture, roundedness, high vowels, Taiwan Mandarin.

1. INTRODUCTION

The vowel system in Mandarin constitutes a three-way contrast: [i], [y], and [u] ([3, 5]). These three vowels are distinguished from each other on two dimensions: frontness and roundedness. These two dimensions are associated with two distinctive features, [back] and [round], in phonology term. The high front vowel [i] is assigned with [−back] and [−round], [y] is associated with [−back] but [+round], and [u] is [+back] and [+round]. With these feature assignments, it is suggested that sounds with the same feature value share the same articulatory gestures. For example, [y] and [u] are both assigned with a positive value for the [round] feature, implying that the roundedness for these two sounds is equivocal. However, whether or not these two sounds are identical in terms of their lip postures is not determined or specified by the assignment of their feature values. The +/− values only provide dichotomic categories between sounds; either all or nothing, and nothing in between. Chances are the degree differences with regards to the roundedness is not fully revealed.

In addition, the [round] feature is resulted from two articulatory gestures: lip rounding and lip protrusion. While these two articulatory gestures can be independently controlled, the co-occurrence between them are far from uncommon, which consequently leads to the interpretation that the [round] feature is the combination of lip rounding and lip protrusion. On the other hand, the high front vowel [i] is assigned with a negative value for the [round] feature. The connotation of [−round] is associated with lips not being rounded. How the lips are actually postured is not specified; the lips can be relaxed (as in [ə]), wide open (as in [a]), or bilaterally pulled (as in [i]), etc. In a broader sense, the [−round] feature fails to capture the fact that the lips are pulled bilaterally for [i] such that the lip posture for [i] is suggested to be comparable to those for [a] and [ə] since they are all assigned with the same [−round] feature. As such, the [round] feature appears to provide rather simplistic interpretations of the lip postures for these high vowels. Whether or not subtle differences can be articulatorily achieved to distinguish these high vowels from one another remains unanswered. Phonological feature assignments may not be able to fully characterize how the lips of Mandarin high vowels are articulatorily postured.

With regards to lip rounding, protrusion and vertical compression are the two common parameters to quantify how labial sounds are produced or acquired [3, 8]. These two parameters, however, only characterize two dimensions, namely vertical and depth, out of a 3D space. The horizontal information, that is, how lips are laterally spread, and the aperture area enclosed by the vertical and lateral axes not only provide ample visual cues but also contribute to the generated acoustics. More crucially, labial sounds do not always have the same degrees or types of rounding. Catford [2] identified two types of rounding: endolabial (e.g., [u] or [o]) and exolabial (e.g., [y] or [ø]); the former refers to the rounding of the inner circle of the lips and the latter is associated with the outer rounding of the lips. Endolabial and exolabial sounds contrast each other in two aspects. First, endolabial sounds pull the corners of the mouth close together whereas exolabial sounds do not.

Second, endolabial sounds are usually accompanied by obvious lip protrusion whereas the degree of lip protrusion is more restrained for exolabial sounds [1].

While [i] – [y] differentiation in Mandarin is argued to lie in vertical distance between the lips [7], whether the lateral distance (the distance between the corners of the mouth) may also play a role in differentiating high vowels appears unanswered.

The current study examines if there is any postural difference among the high vowels in Mandarin and whether or not these subtle differences are consistently realized in production.

2. METHODS

2.1. Participants

Eighteen native speakers of Taiwan Mandarin (9 F, mean age = 23.44) participated in the experiment. None of them reported any auditory or visual disabilities. The study was conducted in accordance with ethical guidelines approved by XXX University. All participants were compensated monetarily for their time.

2.2. Apparatus

An iPad Pro (2021, 12.9 inch) was set up in front of participants (at a distance of approximately 0.5m) to capture their lip movements and, at 30 fps in 4k resolution. Four melting beads were attached to the following locations: the nose, the center of the upper vermillion boundary, the center of the lower vermillion boundary, and the right corner of the mouth (Figure 1). These beads were positioned in order to calculate the lip protrusion and lip aperture. A mirror (7 × 5.5 inch) was placed behind the participants at 45° off the midline so that the melting beads can be captured by the camera and measurement of lip aperture and protrusion and be quantified.

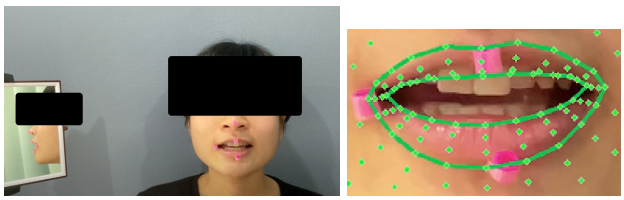


Figure 1: Marker placements for the quantification of lip protrusion and aperture (left), and MediaPipe landmarks on the lips (right).

2.3. Stimuli and procedures

Critical stimuli consisted of the three high vowels in Taiwan Mandarin (i.e., [i], [y], and [u]), produced both in isolation (i.e., monosyllabic) and embedded in disyllabic words. Monosyllabic words were matched with a high level tone (Tone 1): [i1] (‘*one*’), [u1] (‘*house*’), and [y1] (‘*mud*’). Disyllabic words all carried a low-dipping tone (Tone 3), following the same adjective with a falling tone (Tone 4): [ta4 i3], (‘*big ant*’), [ta4 u3] (‘*fifth-grader in college*’), and [ta4 y3], (‘*heavy rain*’). The experiment was designed as a self-paced reading task. Each word was presented via Microsoft PowerPoint, with ten repetitions each.

2.4. Measurements and data analyses

Face detection and tracking were performed by MediaPipe Face Mesh (powered by Google: https://google.github.io/mediapipe/solutions/face_mesh.html). MediaPipe Face Mesh provides 3D information of the 468 landmarks on the face, including the corners of the mouth and the vermillion borders of the upper and lower lips. With these landmarks, four measurements were conducted:

1. *Lateral distance*: the distance between the two corners of the inner lip borders.
2. *Vertical distance*: the maximum distance between the upper and lower lips.
3. *Axial ratio*: the ratio between lateral (numerator) and vertical (denominator) distances, measured from the inner lip borders.
4. *Aperture area*: the area enclosed by the (inner) lip borders.

The degree of lip protrusion was quantified by subtracting the position of the nose from the lips. With the relative positions of the melting beads, another two measurements were performed:

5. *Upper lip protrusion*: the relative positions between the upper lip and nose.
6. *Lower lip protrusion*: the relative positions between the lower lip and nose.

To characterize the overall lip postures, *kinematic stiffness* of the lips was also calculated, following [4, 6]. Multiple MediaPipe landmarks for the upper and lower lips were selected for the quantification of stiffness. To visualize the differences of stiffness between the lip center and corners, data were fitted through the Generalized Additive Mixed Models (GAMMs, [9]).

All measurements were then submitted to linear mixed models (LMM); with each measurement as dependent variable, and VOWEL ([i], [y], [u]) and WORD (isolation vs. disyllabic word) as fixed effects. Random slopes for VOWEL and WORD and random intercept for participant were also included. For all the model constructions, the monosyllabic [y] was set as the reference. For the model of stiffness, we also included LANDMARK (from center to corner) as fixed effects, and monosyllables and disyllables were analyzed through two separate models. All

measurements were first z-transformed before analyses. Significance levels were determined at .05.

3. RESULTS

First, the results showed that the lateral distance for monosyllabic [y] was significantly shorter than that for monosyllabic [i] (β = 1.23, p <.01) but significantly longer than monosyllabic [u] (β = −0.34, p =.01). The shortest distance was found in the production of [u], followed by [y], and then [i] (Figure 2). Less lateral pull (i.e., larger lateral distance) was observed in disyllabic words than in monosyllabic words and this pattern is across-the-broad for all three high vowels, though significant differences between monosyllabic and disyllabic words were found in [y] (β = 0.51, p <.01) and [i], as indicated by the absence of a two-way interaction (β = 0.03, p =0.63).

Jackson and McGowan [7] showed that Mandarin [i] and [y] distinction lies in vertical distance. Our results (Figure 3) echo their findings in that compared with monosyllabic [y], the vertical distance for monosyllabic [i] was significantly larger (β = 1.18, p <.01). No difference was found between monosyllabic [y] and [u] (p =.72). Shorter vertical distances were observed for disyllabic [y] and [u] than their monosyllabic counterparts ([y]: β = 0.33, p =.04; [u]: β = 0.15, p =.09), but the opposite pattern was reported for [i] (β = 0.32, p <.01).

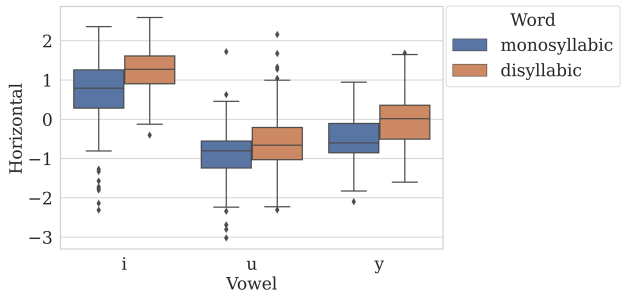


Figure 2: Lateral distance (z-scored) by vowels and conditions.

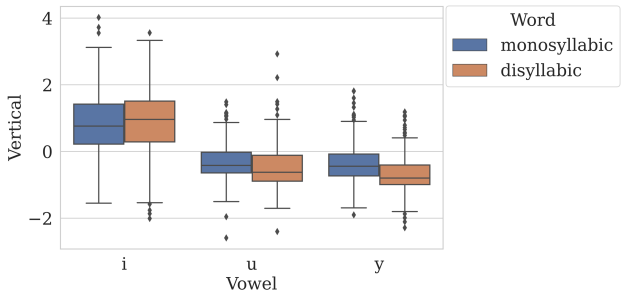


Figure 3: Vertical distance (z-scored) by vowels and conditions.

The results of axial ratios showed that the ratio for monosyllabic [y] were significantly larger than that for [i] (β = −0.81, p =. 01) while the ratios for monosyllabic [y] and [u] did not differ from each other (p = .27). Significant difference between monosyllabic and disyllabic words only lied in [y] (β = 0.56, p <. 01), but not in the other two vowels, as suggested by the VOWEL × WORD interactions ([i]: β = −0.54, p <.01; [u]: β = −0.26, p =.01).

The results of lip aperture area are visualized in Figure 3. The area of monosyllabic words for [y] was significantly smaller than that for [i] (β = 1.42, p <.01) but did not differ from [u] (p =.58). A difference between monosyllabic and disyllabic words was only found in [i], suggested by the significant VOWEL × WORD interaction (β = 0.23, p <.01), but not in the other two vowels ([y]: β = −0.13, p =.25; [u]: β = 0.09, p =.24).

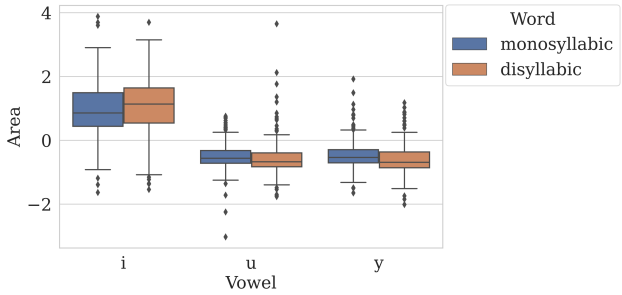


Figure 3: Lip aperture (z-scored) by vowels and conditions.

Considering lip protrusion, the upper lip was more protruded for monosyllabic [y] than for monosyllabic [i] (β = −1.11, p <.01) whereas the protrusion degree did not differ between [y] and [u] (p =.65). While mono and disyllabic [y] did not differ in the degree of lip protrusion (p =.43), interactions between VOWEL and WORD were reported for the other two vowels ([i]: β = −0.21, p =.02; [u]: β = 0.33, p <.01). Similar patterns were also observed in the lower lip protrusion: more protrusion in monosyllabic [y] than in [i] counterpart (β = −0.91, p <.01) and no difference between [y] and [u] (p =.09). No difference between mono and disyllabic [y] and between mono and disyllabic [u] (p =.66, p = .99, respectively), but a VOWEL × WORD interaction for [i] was observed (β = −0.54, p =.02). Results of upper and lower lip protrusion are presented in Figures 4 and 5, respectively.

With regards to the kinematic stiffness, the GAMMs results revealed that for all three vowels, the stiffness of the lip center was higher than those of the corners (Figure 6). This tendency was observed in both mono and di-syllables. The LMM results for monosyllables failed to find significant difference between [y] and [i] (p = .51) and between [y] and [u] (p = .13); the results for disyllables only reported a marginal difference between [y] and [i] (p =.09) but no difference between [y] and [u] (p =.67).

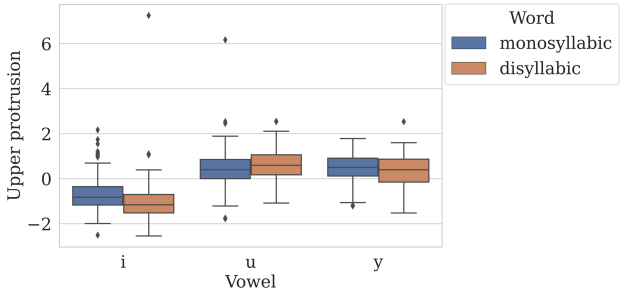


Figure 4: Upper lip protrusion (z-scored) across vowels and conditions.

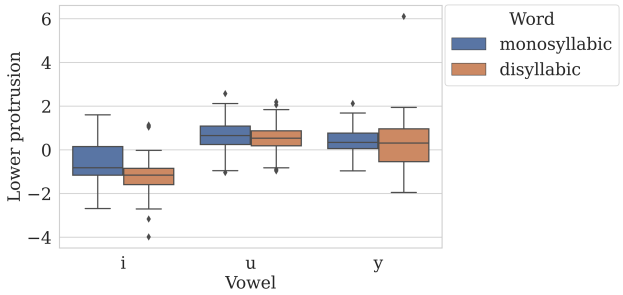


Figure 5: Lower lip protrusion (z-scored) across vowels and conditions.

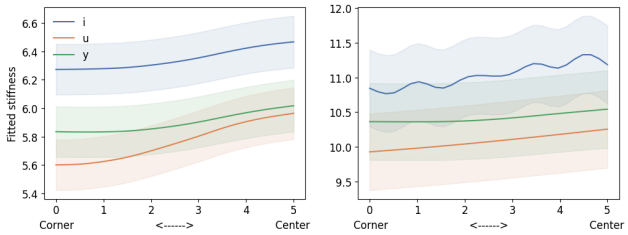


Figure 6: GAMMs predicted fits of kinematic stiffness across different vowels in mono-syllables (left) and di-syllables (right).

4. DISCUSSION

The three vowels in Taiwan Mandarin are traditionally identified through their frontness and roundedness, assuming that segments associated with the same dichotomic value would bear the same articulatory posture. In this vein, the assignment of [+round] for both [y] and [u] leads to an interpretation that the roundedness for these two sounds would be of the same degree. The current study examines the lip postures of the three high vowels in Taiwan Mandarin and compared in what articulatory dimensions they contrast with one another.

Extended from the findings in Jackson and McGowan [7], our results show that the high front vowel [i] is postured significantly different from the other two vowels in all six measurements employed; longer vertical and lateral distances, smaller axial ratios, larger lip aperture, and less lip protrusion. Longer distance both vertically and laterally along with larger lip aperture collectively point to a spread and open lip posture for the high front vowel [i]. These dimensions provide the speakers with visually distinctive cues to tell it from the other two high vowels.

The two “rounded” vowels, [y] and [u], both required shorter lip distances, smaller lip aperture, and more pronounced lip protrusion. According to Catford [2], endolabial [u] and exolabial [y] bear different degrees of lip protrusion: more pronounced for [u] than for [y]. In our results, [u] and [y] exhibited comparable lip protrusion for both upper and lower lips, suggesting that in contrast to the postural difference at the corners of the mouth, the [round] feature may be more associated with lip protrusion since both [u] and [y] are assigned with this feature.

Meanwhile, it is noted that Taiwan Mandarin [y] and [u] contrast with each other only in lateral distances, but not in other measurements. Compared with [u], larger lateral distance along with a larger axial ratio for [y] yield a lip aperture of ellipsis, echoing Catford’s [2] descriptions about the two labial sounds. The [round] feature may not fully capture this postural difference between these two vowels. Despite the fact that [y] and [u] in Taiwan Mandarin only differ in lateral distance, the yielded lip postures can be characterized by other measurements, such as the kinematic stiffness. Our stiffness results demonstrated that while both [y] and [u] had stronger stiffness for the lip center than corners, in general the lip postures for [y] were of slightly stronger stiffness than [u], particularly for the corners of the lips (Figure 6), suggesting that [y] and [u] may be differentiated from each other at the corners of the lips. Additionally, the reported stiffness changes over the lips across the three vowels are not only visually identified but also are associated with physiological accounts. Mayer et al. [10] used 3D face model to simulate lip spreading for [i] and lip compression for [u]. Their results suggest that these lip postures require different groups of muscles while quantal properties are exhibited. Collectively, these observed postural differences for the three high vowels in Taiwan Mandarin ought to be available to the speakers and therefore may serve its function for perceptual identification.

The current study compared perioral postures for the three high vowels in Taiwan Mandarin. While lip aperture characterizes the high front vowel [i], the difference between [u] and [y] did not reside in the degree of roundedness or protrusion, but rather in the postural difference at the corners of the mouth. The characterizations and measurements of these areas would call for future research.

5. REFERENCES

[1] Linker, W. 1982. Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study. *UCLA Working Papers in Phonetics* 56.

[2] Catford, J. C. 1988. *A Practical Introduction to Phonetics*. Oxford: Clarendon Press.

[3] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Oxford, England: Blackwell.

[4] Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. 2002. Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *J. Acous. Soc. Am.* 112(4), 1627-1641.

[5] Lin, Y. H. 2007. *The Sounds of Chinese.* Cambridge University Press.

[6] Van Lieshout, P. H., Bose, A., Square, P. A., & Steele, C. M. 2007. Speech motor control in fluent and dysfluent speech production of an individual with apraxia of speech and Broca's aphasia. *Clinical Linguistics & Phonetics* 21(3), 159-188.

[7] Jackson, M. T.-T., McGowan, R. S. 2012. A study of high front vowels with articulatory data and acoustic simulations. *J. Acoust. Soc. Am.* 131, 3017–3035.

[8] Saito, H. 2016. Lip movements for an unfamiliar vowel: Mandarin front rounded vowel produced by Japanese speakers. *JSLHR* 59(6), S1558-S1565.

[9] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. of Phon.* 70, 86-116.

[10] Mayer, C., Chiu, C., Gick, B. 2021. Biomechanical simulation of lip compression and spreading. *Canadian Acoustics 49*(3), 38 – 39.

Phase relations between the tongue body and the jaw across rate modifications in younger and older speakers

Tabea Thies[1,2], Doris Mücke[2], Angela Ding[3], Antje Mefferd[3]

[1]University of Cologne, Faculty of Medicine and University Hospital Cologne, Department of Neurology
[2]University of Cologne, Faculty of Arts and Humanities, IfL Phonetics
[3]Vanderbilt University Medical Center, Department of Hearing and Speech Sciences
tabea.thies@uk-koeln.de

ABSTRACT

The present study investigated the effects of age and speech rate on intergestural phasing patterns of oral articulators in younger and older speakers across three speech rates. 12 younger and 16 older speakers produced the target word 'Kaia' at a fast, typical and slow speech rate. Phase angles were used to quantify the spatiotemporal relations between the tongue body and the jaw. The timing of the jaw movement was referenced to the tongue body's movement cycle during 'Kai'. The results replicated previous studies on rate effects and showed that younger speakers systematically increased phase angles from fast to typical and further to slow speech. However, in older speakers such increased phase angles were only observed for comparisons of typical and slow speech as phase angles did not differ between fast and typical speech. The findings suggest that intergestural timing patterns may become less flexible with advancing age.

Keywords: phase angle, articulatory timing, aging effects, speech kinematics

1. INTRODUCTION

Speech production is a complex motor task that requires multiple articulators to reach their target positions in a well-timed sequence. Such sequential timing relations are described as patterns of stability and have previously been investigated between the upper lip and the jaw [1]–[4] or the tongue tip and the jaw [5], [6] and described by phase angles on the phase plane. One major goal of these investigations was to identify invariance within speech motor control, as phase angles provide insights into intergestural coordination in the spatiotemporal dimension.

Most studies investigated vowel-consonant-vowel (VCV) sequences with intervocalic labial or alveolar stops or nasals, so that the onset of the consonant with respect to the vocalic jaw cycle was determined. One jaw cycle is defined from the onset of the first vowel to the onset of the second in vowel. Tuller and Kelso investigated regularities in kinematic patterns and reported that the occurrence of lip movement was stable across rate conditions within speakers but variable across speakers [1]. In contrast, Nittrouer and colleagues showed that the phase angle of labial and alveolar movements to the jaw cycle changed due to stress, speech rate and phonetic contrast [3], [5], [6]. As a temporal regularity they observed that the shorter the jaw cycle, the smaller the phase angle and vice versa. They also pointed out that the tongue tip coordination with the jaw exhibited greater variability than the upper lip coordination with the jaw. Shaiman et al. [2] reported rate effects between the upper lip and the jaw in five out of eight speakers and highlighted speaker-specific strategies. In a later study, Shaiman investigated CVC sequences with labial consonants and provided converging evidence that phase angles vary in response to speech rate manipulations [4].

However, so far, tongue tip and lip movements have always been studied with reference to the jaw cycle in symmetrical sound sequences (pseudowords). It remains to be determined if previously reported rate effects on timing patterns can also be observed a) during asymmetrical sound sequences elicited in a sentence repetition task, b) when the tongue body cycle serves as the reference for the jaw, and c) when speakers are of older age. Such knowledge is important to understand the organization of timing patterns between articulators in general and to specify age-related changes on speech motor control.

The present study aims to investigate the intergestural timing between the jaw and tongue body with respect to variation in speech rate (typical, slow, and fast speech). We test whether phasing relations are similar in younger and older speakers, since there is some evidence in the literature that speech motor control is affected by aging [7]–[9]. Phase angles were used to quantify the phasing relations between the tongue body and the jaw. Tongue body cycles were taken as reference movement for the jaw timing.

As shown in most of the previous studies, timing differences due to speech rate changes are expected.

Relative to the typical speech rate, the phase angle is expected to increase during slow speech and decrease during fast speech. No predictions were made for potential aging effects because previous studies have not included age as factor in their analysis (cf. [2]).

2. METHODS

2.1. Participants

12 younger (4 females, 8 males) and 16 older speakers (5 females, 11 males) of American English participated in the study. The mean age was 24 (± 2) years in the younger and 67 (± 8) years in the older group. All participants passed the Mini-Mental State Examination [10] and a pure tone hearing screening to rule out cognitive decline and hearing impairment.

2.2. Speech Material

As part of a larger research project, all participants were asked to produce ten repetitions of the sentence: *Buy Kaia a kite*. For the purpose of the study, we specifically focused on the $C_1V.C_2$ sequence [kaj] of the word *Kaia*, consisting of a stop consonant, a vowel, and an approximant. C_1V is the first stressed syllable, while C_2 is part of the second unstressed syllable. All participants produced the sentence in three conditions: typical, fast, and slow. The typical condition was always elicited first. The fast and slow condition were elicited in a pseudo-randomized order. For fast speech, participants were asked to produce the sentence as fast as possible. For slow speech, participants were asked to produce the sentence at half their typical speaking rate by stretching out the words.

2.3. Data collection

The speech material was recorded with an electromagnetic articulograph (AG 501, Carstens). Small sensors were attached to the articulators to track their movements. One sensor was placed with dental adhesive (Periacryl 90, GluStitch, Inc.) to the sagittal midline of the tongue (approx. 4 cm posterior to the tongue tip) to track the tongue body movement. Another sensor was attached on the central part of the gumline of the lower teeth to track the jaw movement. Three additional sensors were placed on plastic googles that functioned as reference sensors for head correction. The raw data were converted into positional data first using CalcPos software and then head-corrected and rotated into a head-based coordinate system using a biteplane recording and the respective NormPos software provided by Carstens. Data were further processed in SMASH, a MATLAB-based software program [11].

2.4. Data processing

Although the calculation of phase angles in the current study followed closely procedures described in previous studies [1, 8], the tongue body cycle was used as the reference to quantify changes in the spatiotemporal timing of the jaw reaching the first vowel /a/ in the word *Kaia*. Traditionally, the jaw cycle had been used as a reference to quantify timing patterns with the upper lip or tongue tip [2]–[5]. This means that the consonantal closure of the tongue tip or the lip were related to the jaw motions associated with the flanking vowels. In contrast to previous studies, we used the jaw opening motion associated with the vowel and related it to the tongue body motions associated with the flanking consonants in CV.C sequences. We chose this approach because the tongue body exhibited a well-defined lowering and raising movement cycle during the CV.C sequence with reliable landmarks for its onset and offset across all rate conditions.

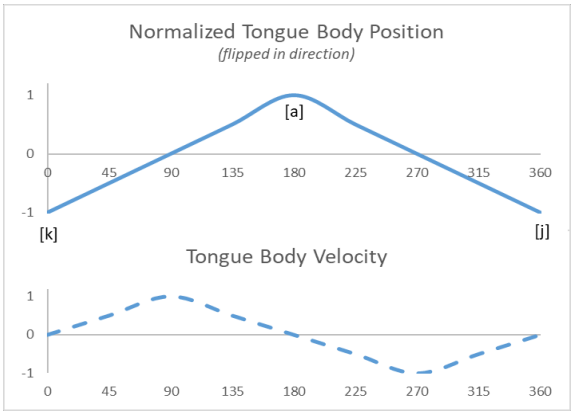


Figure 1: Schematized relationship between tongue body position and related velocity signal along a 360° cycle.

The tongue cycle was defined as the tongue body movement from the constriction [k] to the constriction [j] as it represents a full movement cycle in the closed – open – closed dimension. Positions and associated velocities of the tongue body and jaw movements were taken from the vertical dimension (y-axis). Local maximum and minimum positions were defined by zero crossings of the velocity signal (Figure 1). Please note that movement directions were flipped in their direction as it was necessary for the calculation of normalized tongue body positions and velocities.

Tongue body positions ($TB\ pos_N$) and velocities ($TB\ vel_N$) were normalized across each cycle as described by Kelso and colleagues [12, p. 43,44], so that the tongue cycle was displayed with values ranging from 1 to -1 as in a unit cycle (Figure 2). In Figure 2, the tongue position (pos) is depicted on the

vertical axis, and the tongue movement velocity (vel) on the horizontal axis. Point P (vel, pos) describes landmarks of the tongue body along the phase plane within the unit cycle. The tongue body lowering from [k] to [a] in the word *Kaia* is reflected in the movement from A to C via B. The velocity is zero at A and C, as the tongue body has reached its highest and lowest point. At the midpoint of this movement (point B), the tongue body reaches its local peak velocity. The tongue body raising from [a] to [j] in the word *Kaia* is associated with the movement from point C to point A via point D. Again, at midpoint (point D), the peak velocity is reached.

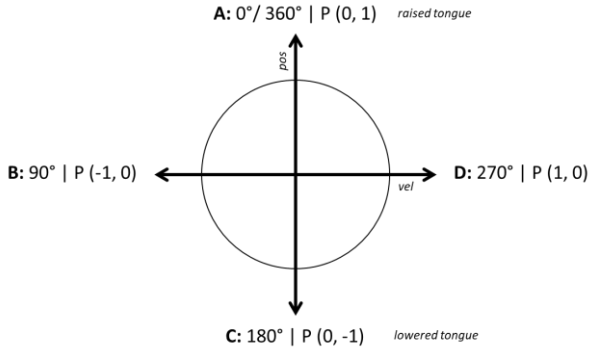


Figure 2: Tongue cycle on a phase plane.

The timing of the jaw reaching its maximum opening position for /a/ in relation to the tongue body cycle was specifically of interest in the current study. Within the tongue body cycle the occurrence of the jaw target for [a] can be defined as an angle on the phase plane, as coordinates within the phase plane can be converted into a phase angle (φ). This allows to determine the timing in degrees of the maximum jaw opening relative to the tongue body cycle.

2.5. Measurements and Statistics

To determine speech rate effects, the tongue body cycle duration was measured. The *cycle duration* refers to the time interval (in ms) between the onset of tongue body lowering from [k] for [a] to the offset of tongue body raising for [j] (both defined as the maximum position in the y-dimension, Figure 1). The phase angle (φ) at the time of the maximum jaw opening was then calculated with the following formula:

$$(1)\ \textit{Phase angle}\ (\varphi) = \arctan\left(\frac{TB\ vel_N}{TB\ pos_N}\right)$$

Values smaller than 180° indicate that the jaw reached the target position for [a] earlier than the tongue body. Angles larger than 180° indicate that the jaw reached the target position for [a] after the tongue body. Values equal to 180° indicate simultaneous arrival of the jaw and tongue body at the target position for the vowel [a].

To verify that participants significantly modulated their speech rate and to determine potential aging effects on speech rate performance, tongue body cycle durations of each participant's 10 productions at each rate condition were submitted to a linear mixed model with task and group as fixed factors and subject as a random factor. Furthermore, to determine rate effects and aging effects on spatiotemporal timing relations of the tongue body and jaw, phase angles for each participant's 10 productions in each speech rate condition were submitted to a linear mixed model. If the main effects or the interaction term were found significant at $p < .05$, pairwise post-hoc analyses with Bonferroni adjustments were completed. Statistical analyses were performed in SPSS (Version 28.0).

3. RESULTS

3.1. Speech Rate

A significant main effect of rate [$F(2,328.6) = 1276.09$, $p < .001$] and group [$F(1,68.93) = 9.63$, $p = .003$] was found. The rate *x* group interaction was not significant. In both groups, the tongue body cycle duration was shortest during fast speech, and longest during slow speech (Table 1). However, the average cycle duration regardless of rate was significantly shorter in younger speakers compared to older speakers (mean difference = 37.18 ms). Although the rate *x* group interaction was not significant, the elicited modifications in cycle duration across all rate conditions tended to be greater in the younger speakers than in the older speakers (Table 1).

	younger	older
fast	200 (7.1)	251 (6.1)
typical	275 (7.2)	322 (6.2)
slow	625 (18.7)	638 (16.4)

Table 1: Mean tongue body cycle duration *(SE)* in ms.

3.2. Timing Relation of the Tongue Body and the Jaw

Averaged jaw phase angles relative to the tongue body cycle are presented in Table 2. While the main effect of group was not significant, the main effect of rate [$F(2, 415.17) = 30.14$, $p < .001$] and the rate *x* group interaction [$F(2, 415.17) = 7.18$, $p < .001$] were significant. Pairwise comparisons of rate revealed that in younger speakers, fast speech was associated with significantly smaller phase angles relative to typical speech (mean difference = -7.47, $p < .001$) and relative to slow speech (mean difference = -11.52, $p < .001$). In addition, slow speech was associated with significantly greater phase angles than typical speech (mean difference = 4.05, $p = .007$).

	younger	older
fast	171 (2.9)	179 (2.0)
typical	178 (2.6)	178 (1.8)
slow	182 (2.5)	185 (2.0)

Table 2: Mean phase angles of the jaw (*SE*) relative to tongue body cycle in degrees.

In older speakers, greater phase angles were elicited in slow speech relative to typical speech (mean difference = 6.32, $p < .001$) and relative to fast speech (mean difference = 5.79, $p < .001$). Fast speech was not associated with significantly different jaw phase angles relative to typical speech in older speakers. No significant between-group comparisons were found at any of the three rate conditions.

3.3. Associations between Tongue Cycle Duration and Jaw Phase Angle across Speech Tasks

No significant correlations were observed between tongue body cycle durations and jaw phase angles across all speakers as well as within each age group (Figure 3). However, fast and typical speech elicited more variable jaw phase angles than slow speech within both groups. In slow speech jaw phase angles hovered around 180°, particularly in older speakers.

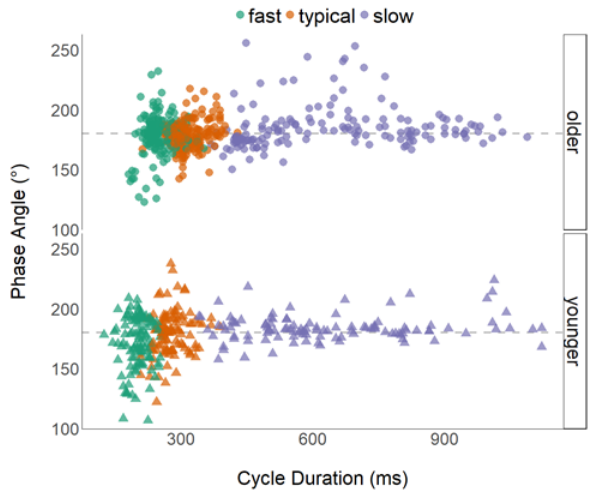


Figure 3: Jaw phase angles as a function of tongue body cycle duration for older speakers (top panel) and younger speakers (bottom panel), speech tasks are color-coded. Grey dashed lines indicate a jaw phase angle of 180°.

4. DISCUSSION

As reported in previous studies [2]–[6], rate effects on phase angles were observed in younger speakers. Phase angles were smaller during fast speech and larger during slow speech relative to typical speech. The significant changes in phase angles from typical speech to fast and slow speech suggest that younger speakers change intergestural timing patterns with task demands. Younger speakers appear to have a lot of articulatory flexibility to accomplish such changes. However, the absence of a significant correlation between cycle duration and phase angles suggests that these changes in intergestural timing are not purely duration-driven in terms of intrinsic consequences of the speech system's dynamical organization. Instead, speakers appear to switch into a different control mode when cued to speak faster or slower.

In older speakers, phase angles changed significantly in response to slow speech but not in response to fast speech relative to typical speech. Nevertheless, cycle durations differed significantly between fast, typical, and slow rates. Older speakers appear to maintain the same timing mode they use for typical speech when being cued to speak faster, although they manage to successfully shorten their cycle durations (Table 1). As phase angles tended to be larger in older adults than younger adults during fast speech (Table 2), older speakers are perhaps not able to move the articulators as independently of each other as younger speakers. This may indicate a less economical articulatory strategy, as the jaw does not move on to the next sound as quickly after it has reached the vocalic target. In addition, significant group differences were found for tongue body cycle durations with older speakers producing longer cycle durations than younger speakers across all speech rates. This finding is in line with previous studies and may suggest a slowing down of the speech system with advancing age [7], [8], [13].

Both age groups change intergestural timing for slow speech. Although the absolute durations during slow speech vary greatly in both age groups (Figure 3), the phase angles are stable [12]. This highlights the task-specific effect, as the variability in phase angles decreases. The observed rate effects on phase angles, particularly in slow speech, could provide a better context to examine intergestural timing in speakers with motor speech disorders as their speaking rate is often abnormally slow [12, 13]. Therefore, the findings of this study are an important first step to understand intergestural timing patterns in speakers with motor speech disorders.

5. CONCLUSION

Although previously reported rate effects on intergestural timing patterns were replicated, the current study showed that younger speakers are more flexible in their intergestural timing responses than older speakers. Furthermore, findings suggest that the cued speech task, but not duration itself, determines the intergestural timing patterns.

6. FUNDING

This work was supported by the visitor program of the Thiemann Foundation, the collaborative research center SFB1252 Prominence in Language (Project-ID 281511265) funded by the German Research Foundation (DFG), and by NIH NIDCD grants R03DC01575 and R01DC019648.

7. REFERENCES

[1] B. Tuller and J. A. S. Kelso, "The timing of articulatory gestures: Evidence for relational invariants," *J. Acoust. Soc. Am.*, vol. 76, no. 4, pp. 1030–1036, Oct. 1984, doi: 10.1121/1.391421.

[2] S. Shaiman, S. G. Adams, and M. D. Z. Kimelman, "Timing relationships of the upper lip and jawacross changes in speaking rate," *J. Phon.*, vol. 23, no. 1–2, pp. 119–128, Jan. 1995, doi: 10.1016/S0095-4470(95)80036-0.

[3] S. Nittrouer, K. Munhall, J. A. S. Kelso, B. Tuller, and K. S. Harris, "Patterns of interarticulator phasing and their relation to linguistic structure," *J. Acoust. Soc. Am.*, vol. 84, no. 5, pp. 1653–1661, Nov. 1988, doi: 10.1121/1.397180.

[4] S. Shaiman, "Articulatory Control of Vowel Length for Contiguous Jaw Cycles: The Effects of Speaking Rate and Phonetic Context," vol. 45, no. 5, pp. 663–675, 2002.

[5] S. Nittrouer, "Phase relations of jaw and tongue tip movements in the production of VCV utterances," *J. Acoust. Soc. Am.*, vol. 90, no. 4, pp. 1806–1815, Oct. 1991, doi: 10.1121/1.401661.

[6] M. Masapollo and S. Nittrouer, "Interarticulator Speech Coordination: Timing Is of the Essence," *J. Speech Lang. Hear. Res.*, vol. 66, no. 3, pp. 901–915, Mar. 2023, doi: 10.1044/2022_JSLHR-22-00594.

[7] T. Thies, A. Hermes, and D. Mücke, "Compensation in Time and Space: Prominence Marking in Aging and Disease," *Languages*, vol. 7, no. 1, p. 21, Jan. 2022, doi: 10.3390/languages7010021.

[8] A. Hermes, J. Mertens, and D. Mücke, "Age-related Effects on Sensorimotor Control of Speech Production," in *Proceedings of Interspeech 2018*, Hyderabad, India: ISCA, 2018, pp. 1526–1530. doi: 10.21437/Interspeech.2018-1233.

[9] J. D. Amerman and M. M. Parnell, "Speech timing strategies in elderly adults," *J. Phon.*, vol. 20, no. 1, pp. 65–76, Jan. 1992, doi: 10.1016/S0095-4470(19)30254-2.

[10] M. F. Folstein and S. E. Folstein, *Mini-Mental State Examination Second Edition | MMSE-2*, 2nd ed. Lutz, FL: Psychological Assessment Resources., 2012. Accessed: Nov. 10, 2022. [Online]. Available: https://www.parinc.com/Products/Pkey/238

[11] J. R. Green, J. Wang, and D. L. Wilson, "SMASH: A Tool for Articulatory Data Processing and Analysis," pp. 1331–1335, 2013.

[12] J. A. S. Kelso, E. L. Saltzman, and B. Tuller, "The dynamical perspective on speech production: data and theory," *J. Phon.*, vol. 14, no. 1, pp. 29–59, Jan. 1986, doi: 10.1016/S0095-4470(19)30608-4.

[13] D. Mücke, T. Thies, J. Mertens, and A. Hermes, "Age-related effects of prosodic prominence in vowel articulation," in *Proceedings of 12th International Seminar on Speech Production*, 2021, pp. 126–129.

[14] A. S. Mefferd, "Effects of Speaking Rate, Loudness, and Clarity Modifications on Kinematic Endpoint Variability," *Clin. Linguist. Phon.*, vol. 33, no. 6, pp. 570–585, 2019, doi: 10.1080/02699206.2019.1566401.

[15] A. S. Mefferd and M. S. Dietrich, "Tongue- and Jaw-Specific Articulatory Underpinnings of Reduced and Enhanced Acoustic Vowel Contrast in Talkers With Parkinson's Disease," *J. Speech Lang. Hear. Res.*, vol. 62, no. 7, pp. 2118–2132, Jul. 2019, doi: 10.1044/2019_JSLHR-S-MSC18-18-0192.

ANTICIPATORY LABIALIZATION IN C[u] CLUSTERS: A CROSS-LINGUISTIC STUDY

Veno Volenec

Concordia University
veno.volenec@concordia.ca

ABSTRACT

This paper presents a cross-linguistic investigation of anticipatory labialization (AL) in C[u] clusters. It addresses two underexplored research questions: (1) Is AL always present in a consonant (C) that is followed by [u] in all languages? (2) Is there a cross-linguistic difference in the temporal unfolding of AL in C[u] clusters?

Lip movements of 10 native speakers of 10 languages were video-recorded during the production of all the C[u] clusters that exist in each language. Participants read isolated words beginning with C[u] (e.g., *rule* [ɹuɬ]) from a screen; this trial was repeated three times. The duration of AL before [u] was determined using OpenFace2.2.

The results showed that: (1) all l0 languages displayed AL in every token of every C; (2) there is a statistically significant difference in the duration of AL in C from C[u] between the 10 languages, ranging from 83.0 ms (Japanese) to 109.7 ms (Brazilian Portuguese).

Keywords: anticipatory labialization, V-to-C coarticulation, cross-linguistic study

1. INTRODUCTION

Anticipatory labialization (AL) is a coarticulatory effect where the lip rounding gesture from a rounded sound, such as [u], is temporally extended in the regressive direction ('right-to-left' if the sounds are transcribed), influencing one or more of the preceding consonants [1, 2]. For example, in the English word *rule* [ɹʷuɬ], the lip rounding gesture originating from the vowel [u] encapsulates the entire production of the consonant [ɹ], so that this consonant, which otherwise does not entail lip rounding (e.g., *read* [ɹid]), becomes rounded or labialized.

Despite the vast experimental and theoretical literature on labial articulation and coarticulation in various languages [10, 11, 12, 13, 14, 15, 16], only a few studies have undertaken a cross-linguistic investigation of AL, and none of them have directly compared AL across more than two languages [3].

Lubker and Gay [4] used electromyography to compare AL in Swedish and American English. The results showed that while AL was present in both languages in all the syllables in which consonants appeared before rounded vowels, AL in Swedish was systematically produced with more extensive and more precise lip-protrusion movements than AL in English. Also, the timing of AL was found to differ in the two languages: for the same number of consonants, the onset of AL occurred earlier for the Swedish than for the English speakers.

Boyce [5] compared patterns of lip rounding in VCV sequences for speakers of English and Turkish. She measured EMG activity in the *orbicularis oris* muscle and found that Turkish speakers produced "plateau" patterns of movement, where lip rounding in a C is approximately the same as in both Vs, whereas in English they produced "trough" patterns, where lip rounding is greater in Vs than in a C. Because the study was done on VCV sequences, it cannot be determined whether the intervocalic C was influenced by the AL of the following vowel or the carryover labialization from the preceding vowel, nor if there are any temporal differences in the labialization of C between the two languages.

Noiray and colleagues [6] found no significant differences in AL in terms of lip protrusion and constriction between American English and Quebec French, while data on its scope and duration were not provided.

Thus, while previous research sheds some light on the similarities and differences in AL between languages, various aspects of it remain insufficiently explored, in particular whether it systematically occurs in all languages with rounded vowels and whether its spatio-temporal properties vary cross-linguistically. With the goal of contributing to this research area, this paper addresses two questions. First, is AL always present in a consonant that is followed by [u] or is it systematically absent in that context in some languages? Second, is there a cross-linguistic difference in the temporal unfolding of AL in C[u] clusters? In order to facilitate direct cross-linguistic comparison, this study only concentrates on AL originating from [u] (because [u] is typologically the most frequent rounded vowel [7]), and it only investigates AL in word-initial C[u] clusters (because

CV syllables are typologically most common [8], and because word-initial C[u] clusters ensure that only anticipatory and not carryover coarticulation is present).

2. METHOD

2.1 Participants

This study included 10 participants, seven of whom were female and three were male. Each person is a native speaker of one of the 10 studied languages: Brazilian Portuguese, Croatian, American English, Quebec French, Italian, Japanese, Jordanian Arabic, Lithuanian, Persian, and Telugu. The crucial criterion for the inclusion of a language in this study was for a language to contain the phoneme /u/ in its vowel inventory. Since the phonetic literature does not indicate that biological and social factors such as gender, age and level of education play a role in the realization of AL [3, 4], these parameters were not tracked.

2.2 Materials

For each of the 10 languages, a list of words was created. The purpose of these lists was to elicit pronunciations that feature AL. The lists contained only words that begin with C[u] clusters. Every particular language's word list covered all of the consonants that may appear word-initially before [u] in that language. Since the consonant inventories differ between languages, the word lists containing the C[u] clusters were also partially different. For example, the Japanese word list contained 10 different words, each beginning with a particular C[u] cluster (such as [nuka], [ɾusu], [t͡suika], etc.), while the Lithuanian word list contained 16 C[u] clusters. Six consonants appeared in all of the languages: [n, t, d, s, ʃ, g]. All labial consonants were deliberately excluded from the study in order to avoid any possible clash between the gestures associated with the labial place of articulation and labial coarticulation.

The word lists were compiled in collaboration with the participants. The investigator consulted the extant phonological descriptions of the 10 languages and compiled a list of all possible C[u] configurations, given the languages' consonant inventories. The participants were then asked to think of a single word in their language that featured each of those clusters. Thus, the word lists contained all of the possible word-initial C[u] clusters, with a single example for every such cluster. Importantly, the topic, the context and the goal of the study were not revealed to the participants during this process. For every language separately, the words were placed into an automated PowerPoint presentation in which a single word appeared on a slide every 6 seconds. The entire word list was presented three times in a row, so that ultimately three tokens per C[u] cluster were obtained. Such presentations were used to elicit pronunciations that feature AL.

2.3 Procedure and data analysis

During the experiment, the participants were seated in front of a computer with a video camera (a 2021 16-inch MacBook Pro). The camera was slightly below their eye level at an approximate distance of 75 cm. They were instructed to read the words of their native language as they appeared on a screen in a natural and spontaneous way. As they were pronouncing the words from an automated PowerPoint presentation, a camera was recording their entire face with a 1920x1080 pixel resolution at 30 frames per second.

In order to determine the duration of AL, the applications OpenFace2.2 and Wondershare Filmora X were used. OpenFace [9] is an automated facial behavior analysis toolkit driven by artificial intelligence. For every frame of a video, it places landmarks (dots) on a person's face and keeps track of the location of those landmarks in a 3D space (Figure 1). Out of 130 facial landmarks, 20 are placed over the lips. All of the video recordings gathered in the experiments were processed in OpenFace, which generated an Excel spreadsheet with 3D coordinates of lip movements during the production of all the C[u] clusters. These data showed the exact moment of the beginning of lip rounding, i.e., the onset on AL.

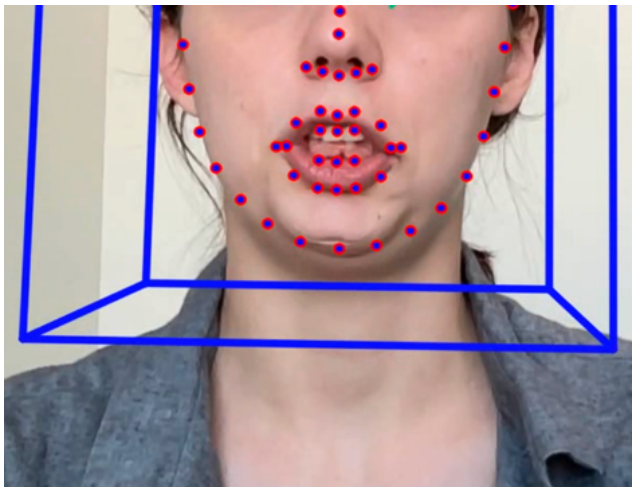


Figure 1: Automated facial landmark detection in OpenFace2.2. The software uses artificial intelligence to keep track of landmarks (blue dots in red circles) as they move in 3D space over time.

The video processing software Wondershare Filmora X was used to determine the ending of AL during the production of the consonants. Audio-visual inspection of the recorded videos, conducted frame by frame, was used to ascertain the beginning of the vowel [u]. Specifically, the beginning of [u] was established on the basis of two criteria: the onset of the auditory perception of [u] and the onset of the formant structure characteristic for the high back rounded vowel. The temporal distance between the onset of AL as determined by OpenFace and the onset of [u] as determined by Wondershare Filmora X was taken as the duration of AL. Thus, these two applications were used in order to determine both the presence versus absence of AL and the duration of AL in C[u] clusters.

Subsequently, the mean duration of AL was calculated for every language separately by summing the durations of AL in all C[u] tokens and dividing that by the number of tokens. One-way ANOVA was used to test if there are statistically significant differences in the duration of AL between the 10 tested languages.

3. RESULTS

All l0 languages displayed anticipatory labialization (AL) in every token of every word-initial consonant that was followed by [u]. There was not a single instance of an absence of AL in any of the tested words. Furthermore, in all of the tested words in all languages, the anticipatory lip rounding gesture preceded all consonantal acoustic cues. In other words, AL could always be observed as the first overt articulatory gesture associated with a C in a C[u] cluster, before any audible signal was emitted by the speaker. For example, in the production of the word *succo* [suk:o] 'juice', the Italian speaker started to round her lips 40 milliseconds before the audible friction associated with [s] began. The instant before [s] became audible is captured in Figure 2, showing that AL preceded the consonant's acoustic cue.

The average duration of AL (expressed in milliseconds) for every language is presented in Table 1. The total average duration of AL in all C[u] clusters calculated across all 10 language is 96.7 ms. The third column shows the respective standard deviations, which are relatively high, but strikingly similar cross-linguistically. Thus, the results show that different languages display very similar variation in the duration of AL.

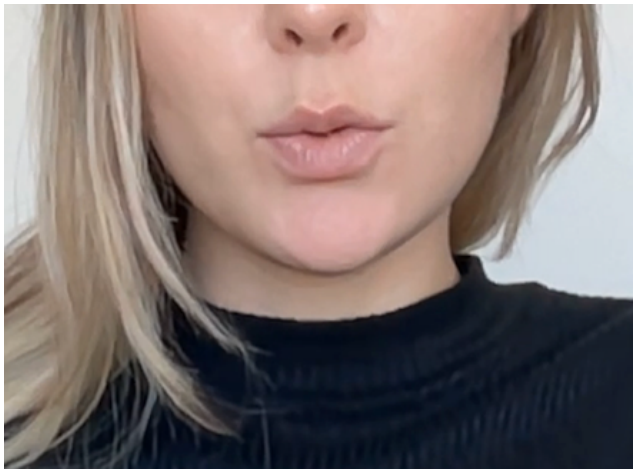


Figure 2: Anticipatory labialization as it is produced 10 ms before the onset of audible friction of [s] in a [su] cluster (in the word *succo* [suk:o] 'juice') by an Italian speaker.

Language	Mean duration of AL (ms)	Standard deviation
Brazilian Portuguese	109.7	18.6
Croatian	89.3	18.5
English	92.3	18.7
French	92.7	19.5
Italian	102.0	19.6
Japanese	86.3	18.7
Jordanian Arabic	104.2	20.1
Lithuanian	96.1	19.1
Persian	97.6	18.7
Telugu	96.7	19.7

Table 1: Mean duration of anticipatory labialization in C[u] clusters in 10 languages.

A one-way ANOVA showed that there is a statistically significant difference in the duration of AL in C from C[u] between the 10 languages: $F = 4.773$, $p = 0.01$. While there are no significant differences in AL between some languages, for example between English and French, some of them do display significant differences, the most prominent difference being between Japanese with the shortest

mean AL of 86.3 ms and Brazilian Portuguese with the longest mean AL of 109.8 ms. The shortest observed instance of AL lasted for 40 ms and it appeared during [n] in the Persian word [nur] 'light'. The longest observed AL lasted for 150 ms during [ʃ] in the Brazilian Portuguese word [ʃuvə] 'rain'.

4. DISCUSSION

The results of this study showed that anticipatory labialization (AL) was always present during the production of a consonant that was followed by a high back rounded vowel [u]: it was consistently observed in all 10 of the tested languages, in all classes of consonants, in all tokens of all the words that were uttered during the experiment. The wide cross-linguistic distribution of AL cannot be attributed to historical factors because AL is ubiquitous in genetically unrelated languages; Telugu is a Dravidian language; Jordanian Arabic is an Afro-Asiatic language, Japanese is a Japonic language, and the rest belong to various sub-families of Indo-European. While previous research showed that it is likely that *particular parameters* of AL are determined by language-specific factors [4, 5, 10, 11, 12, 13, 14, 15, 16], the fact that it systematically appears in genetically unrelated languages with vastly different phonologies suggests that AL *as a general phenomenon* stems from universal (i.e., non-language-specific) phonetic principles. Following [17], [18] and [19], it can be assumed that the main motivation for AL is perceptual: as a result of anticipatory coarticulation, information about an upcoming sound, in this case [u], is available to the listener before that sound is fully articulated, and this prior information may facilitate more accurate perception than would be the case if all acoustic and articulatory cues were confined within the temporal boundaries of that sound.

The results related to the temporal unfolding of AL showed that it cannot be said that AL has a fixed universal duration cross-linguistically, because there is a statistically significant difference in the duration of AL between at least some of the examined languages. These results are consistent with the two previous cross-linguistic studies of AL [4, 5] that also found between-language differences in the spatio-temporal parameters of AL. It is likely that the temporal differences in AL result from the differences in the underlying phonological systems of the languages in question. It has previously been shown that syllable structure, prosody, and the contrastive vs. redundant nature of the feature [±ROUND] can all play a role in the realization of coarticulatory effects [3]. Thus, AL has a different duration in different languages because during speech production the effect of AL is superimposed onto different, language-specific phonological representations. The interaction between the language-specific phonology and the non-language-specific nature of AL explains why on the one hand languages can differ in the average duration of AL (because AL is manifested on different phonological systems) and why on the other hand all of the tested languages display strikingly similar AL variability as indicated by the standard deviations in Table 1 (because the variability of AL is due to idiosyncratic biomechanical factors that are the same irrespective of a speaker's native language).

The main significance of this study is that it provides the first direct cross-linguistic comparison of AL on a relatively large sample of languages, whereas previous studies have focused either on AL in individual languages [10, 11, 12, 13, 14, 15, 16] or on comparing two languages [4, 5, 6]. This work has several limitations, which indicate avenues for future research. First, it would be beneficial to expand the array of compared languages as much as possible in order to increase the reliability of cross-linguistic generalizations about AL. Second, this study has only focused on the temporal aspect of AL, leaving aside its spatial properties. Future cross-linguistic investigations of AL should also aim to determine how languages differ in the shape and extent of lip-rounding during AL. Third, the influence of phonological factors on the realization of AL merits further exploration in order to clarify what exactly it is that, for example, yields a short mean AL in Japanese (86.3 ms) and a long mean AL in Brazilian Portuguese (109.8 ms).

In summary, this paper explored anticipatory labialization in C[u] clusters from a cross-linguistic perspective. It addressed two research questions: is AL always present in a consonant that is followed by [u] in all languages; and is there a cross-linguistic difference in the temporal unfolding of AL in C[u] clusters? The results showed that all l0 of the studied languages displayed AL in every token of every consonant from a C[u] cluster, and that while AL can be seen as a universal phonetic phenomenon, the temporal properties of AL can vary significantly between languages due to their different phonologies.

5. ACKNOWLEDGEMENTS

The writing of this paper was made possible by the selfless support and encouragement of Elizabeth Tobyn, to whom I extend my heartfelt gratitude.

6. REFERENCES

[1] Farnetani, E., Recasens, D. (2013). Coarticulation and connected speech processes. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds.), *The Handbook of Phonetic Sciences*. 316–351. Oxford: Wiley-Blackwell.

[2] Volenec, V. (2015). Coarticulation. In: Davis, J. (ed.) *Phonetics: Fundamentals, Potential Applications and Role in Communicative Disorders*. 47–86. New York: Nova Science Publishers.

[3] Farnetani, E. (1999). Labial coarticulation. In: Hardcastle, W. J., Hewlett, N. (eds.), *Coarticulation: theory, data and techniques*. 144–163. Cambridge: Cambridge University Press.

[4] Lubker, J. F., Gay, T. (1982). Anticipatory labial coarticulation: experimental, biological, and linguistic variables. *Journal of the Acoustical Society of America* 71, 437–448.

[5] Boyce, S. E. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America* 88, 2584–2595.

[6] Noiray, A., Cathiard, M. A., Abry, C., Ménard, L. (2010). Lip rounding anticipatory control: Crosslinguistically lawful and ontogenetically attuned. In: Maassen, B., Van Lieshout, P. (eds.), *Speech Motor Control: New Developments in Basic and Applied Research*. 153–171. Oxford: Oxford University Press.

[7] Ladefoged, P, Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell Publishers.

[8] Gordon, M. (2016). *Phonological Typology*. Oxford: Oxford University Press.

[9] Baltrušaitis, T., Zadeh, A., Lim, Y. C., Morency, L. P. (2018). OpenFace 2.2: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*. URL: [https://github.com/TadasBaltrusaitis/OpenFace].

[10] Benguerel, A. P., Cowan, H. (1974). Coarticulation of upper lip protrusion in French. *Phonetica* 30, 41–55.

[11] Lubker, J. F., McAllister, R., Carlson, P. (1975). Labial co-articulation in Swedish: a preliminary report. In: Fant, C. G. M. (ed.), *Proceedings of the Speech Communication Seminar.* 55–64. Stockholm: Almqvist and Wiksell.

[12] Descout, R., Boë, L. J., Abry, C. (1980). Labialité vocalique et labialité consonantique. 111–126. *Labialité et Phonétique, Publications de l'Université des Langues et Lettres de Grenoble*.

[13] Bell-Berti, F., Harris, K. S. (1979). Anticipatory coarticulation: some implications from a study of lip rounding. *Journal of the Acoustical Society of America* 65, 1268–1270.

[14] Magno-Caldognetto, E., Vagges, K., Ferrigno, G., Busà, M. G. (1992). Lip rounding coarticulation in Italian. *Proceedings of the International Conference on Spoken Language Processing* 92/1, 41–64.

[15] Sussman, H. M., Westbury, J. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research* 24, 16–24.

[16] Perkell, J. S., Matthies, M., (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/: within- and cross-subject variability. *Journal of the Acoustical Society of America* 91, 2911–2925.

[17] Kühnert, B., Nolan, F. (1999). The origin of coarticulation. In: Hardcastle, W. J., Hewlett, N. (eds.), *Coarticulation: theory, data and techniques*. 7–30. Cambridge: Cambridge University Press.

[18] Fowler, C. A. (2005). Parsing coarticulated speech in perception: effects of coarticulation resistance. *Journal of Phonetics* 33, 199–213.

[19] Harrington, J., Hoole, P., Pouplier, M. (2013). New Directions in Speech Production. In: Jones, M. J., Knight, R. (eds.), *The Bloomsbury Companion to Phonetics*. 242–259. London: Bloomsbury.

ARTICULATION OF CANTONESE AND ENGLISH SIBILANTS AMONG BILINGUAL SPEAKERS IN HONG KONG

Jonathan Havenhill, Ming Liu, Tak Wang Li

The University of Hong Kong
jhavenhill@hku.hk, mingliu_7@connect.hku.hk, u3574384@connect.hku.hk

ABSTRACT

English influence has been argued to be responsible for the recent emergence of an allophonic split between alveolar [t͡s, t͡sʰ] and alveolo-palatal [t͡ɕ, t͡ɕʰ] alternants for the affricates of Hong Kong Cantonese (HKC). However, the phonetic similarity and phonological relationship between English and HKC sibilants has not been empirically established. This study uses ultrasound tongue imaging with synchronized audio and lip video to examine the production of English and Cantonese sibilants among native HKC speakers with varying levels of English proficiency. Participants recited a Cantonese word list containing Cantonese /s, t͡s, t͡sʰ/ and an English word list containing /s, ʃ, t͡ʃ, d͡ʒ/. While L1 English, L2 English, and L1 HKC sibilants are largely similar in terms of spectral center of gravity, a high degree of interspeaker variability is observed in articulation, primarily with respect to lip rounding. These findings are considered with respect to their implications for theories of bilingual phonological representation and of contact-induced change.

Keywords: ultrasound tongue imaging, articulation, sibilants, Hong Kong Cantonese, bilingualism

1. INTRODUCTION

Hong Kong is a multilingual environment where both English and Cantonese are in common use and have been in continuous contact for several centuries. Although Hong Kong English (HKE) has received increasing attention as a distinct variety of English [1], English is predominantly learned through schooling as a second language by native speakers of Cantonese [2]. HKE therefore exhibits a number of phonological patterns that derive from the sound system of Hong Kong Cantonese (HKC). These include alternation between [n] and [l], *th*-fronting, and consonant cluster reduction, among others [3]. Substitutions of [s] for [ʃ] (and vice versa) have also been reported. HKC is canonically described as having a single series of alveolar sibilants: the voiceless fricative /s/ as well as aspirated and unaspirated affricates /t͡s, t͡sʰ/ [4]. Given the lack of an English-like alveolar vs. post-alveolar contrast, Chan and Li [5] report that Hong Kong speakers show a merger of English minimal pairs like *save-shave* and *sip-ship* in which both words are produced with [s]. They note that this pattern is reversed before back round vowels, such that "soup" is pronounced as [ʃuːp], although the opposite ([sʊd] for "should") has also been observed [6]. The (typically rounded) English affricates [t͡ʃ, d͡ʒ] are likewise reported to be replaced by [t͡sʰ, t͡s], described as alveolar with lip spreading [5].

However, a handful of Cantonese studies (e.g., [7, 8]) report the emergence of an allophonic split between alveolar [t͡s, t͡sʰ] and alveolo-palatal [t͡ɕ, t͡ɕʰ] or post-alveolar [t͡ʃ, t͡ʃʰ]. Recent acoustic and articulatory data confirm that young HKC speakers use distinct tongue gestures for producing the two allophones, and that the allophones are conditioned by vowel rounding [9]. This change has typically been attributed to English contact, with Cheung [7] asserting that "speakers equate the rounded Cantonese sibilants (basically alveolopalatals) with the English palatoalveolars" and that they "substitute the English palatoalveolars for the usual realization of Cantonese sibilants when these are followed by rounded vowels" (p. 202). Despite this claim, phonetic similarity of the HKC and English sibilants has not been explicitly tested.

Transfer from English to Cantonese is to some extent plausible. Most theories of L2 perception and learning (including PAM/PAM-L2 [10] and SLM-r [11]) assume that L2 learners take as their starting point the phonetic and phonological system of their L1. Phonological acquisition begins with mapping L2 sounds to equivalent L1 sounds according to their phonetic similarity [11]. Such combined L1-L2 categories may subsequently experience phonetic shifts under influence from both L1 and L2 input.

The phonetic properties that comprise such categories, however, have not been fully established. Most studies of L2 production rely on acoustic data, although PAM-L2 and SLM-r differ significantly

with respect to the phonological status of articulatory gestures. Whether the phonetic targets of nonnative speech production are acoustic or articulatory has been examined by Oakley [12], who finds that L2 speakers can adopt a range of distinct articulatory configurations in producing nonnative sounds, and that speakers may target both acoustics and articulation.

This study aims to determine how Cantonese-English bilingual speakers produce both L1 Cantonese and L2 English sibilants. Questions to be addressed include to what extent HKC speakers re-use L1 phonetic targets in producing L2 sibilants; whether L1 and L2 sibilants are similar in acoustics, articulation, or both; and whether native HKC speakers show similarity to native English speakers.

2. METHODS

2.1. Participants

8 L1 speakers of HKC (1 man, 7 women) and 4 native US/UK English (NE) speakers (all women) participated in the study. NE speakers were aged 21-32 years (mean 26.5; SD 3.9). HKC speakers were born between 1996-2004 (mean age 20 years; SD 2.1) and raised in Hong Kong at least through age 18. All began acquiring English between ages 2-6, but show varying levels of English proficiency and use. Proficiency was tested via two pre-screening tasks, C-Test [13] and LexTale [14], and through collection of standardized English exam scores. 4 high and 4 low proficiency speakers were recorded, but results do not clearly pattern according to proficiency, which is not further considered here.

2.2. Materials

HKC participants were first asked to read an HKC wordlist, followed by an English wordlist. The HKC wordlist, adapted from [9], included 72 disyllabic words with target onsets /s t͡s t͡sʰ/ followed by the vowels /i, ɛ, ɐ, a, y, œ, u, ɔ/, as well as 30 filler words with onsets /kʰ, kʰʷ, k, kʷ, w/. Target onsets appeared in the first syllable with either Tone 1 (high level) or Tone 3 (mid level). Each word was presented in the carrier phrase [ŋɔ:˩ sɛ:˧ ____ jɐt̚˧ tsʰi:˩] ("I write ____ one time"), written in traditional Chinese characters. The English wordlist contained 96 words with the onsets /s, ʃ, t͡ʃ, d͡ʒ/ and the vowels /i, ɪ, e, ɛ, u, ʊ, o, ɔ/, appearing in the first syllable of a stress-initial word. Words were embedded in the carrier phrase 'say ____ each time'. NE speakers produced the same English wordlist as HKC speakers.

2.3. Procedure

The experiment was conducted in a sound-attenuated booth at the authors' university. Stimuli were presented to each speaker in a unique pseudo-random order using Articulate Assistant Advanced (AAA) [15]. Each prompt was repeated three times. The HKC wordlist yielded 216 target tokens per speaker (excluding filler items), while the English wordlist generated 288 tokens. The total number of target tokens was 5,184 across 12 participants.

Ultrasound data were collected using an Articulate Instruments SonoSpeech Micro ultrasound system with a 20mm radius 2–4MHz transducer, which captured ultrasound images at an average frame rate of 84 frames per second (fps). Lip video was captured at 60 fps with front-view and side-view cameras. Both the ultrasound transducer and lip cameras were held in place with a stabilizing headset worn by the participant. Audio was recorded at a 48kHz sample rate and 16-bit sample depth on a Denon F650R solid state recorder, using a Sound Devices USBPre2 preamplifier and an Earthworks Ethos cardioid condenser microphone placed 5-10 cm from the side of speaker's mouth. Audio was simultaneously recorded in AAA, which was used to synchronize the acoustic and articulatory recordings. At the beginning and end of each session, an image of the occlusal plane was captured using a plastic biteplate and a palate trace was recorded. For HKC speakers, both word lists were recited in a single session with a short break, but without removal or adjustment of the headset.

2.4. Analysis

Acoustic recordings were segmented using the Montreal Forced Aligner [16] and manually corrected. For the affricates /t͡s, t͡sʰ, t͡ʃ, d͡ʒ/, the first four spectral moments [17] were measured at the midpoint of frication, typically corresponding to approximately 75% of overall constriction duration. For the fricatives /s, ʃ/, spectral measurements were taken at 75% of frication duration, such that all target sibilants were measured at a consistent distance from the vowel onset.

Ultrasound tongue contours were automatically tracked with DeepLabCut [18] using the pre-trained MobileNet1.0-based neural network implemented in AAA [19]. Tongue splines were extracted for each token at a single time point corresponding to the acoustic measurements. Tongue contours were analyzed using polar SSANOVA [20], with splines fit to each combination of language, sibilant, and vowel rounding. Front-view lip video was

also tracked with DeepLabCut using a pre-trained MobileNetV2 network [19]. Rounding was quantified by calculating the degree of horizontal lip opening between the oral commissures [21] for frames corresponding to the acoustic measurements.

3. RESULTS

3.1. Acoustic Results

Combined spectral measurements for all speakers are provided in Figure 1. The patterns observed here are representative of individual speaker results, with only minor variance. Given the reported equivalence between HKC [t͡sʰ, t͡s] and English [t͡ʃ, d͡ʒ] [5, 7], L1-L2 affricate pairs are displayed together for HKC speakers, as are L1 and L2 [s].

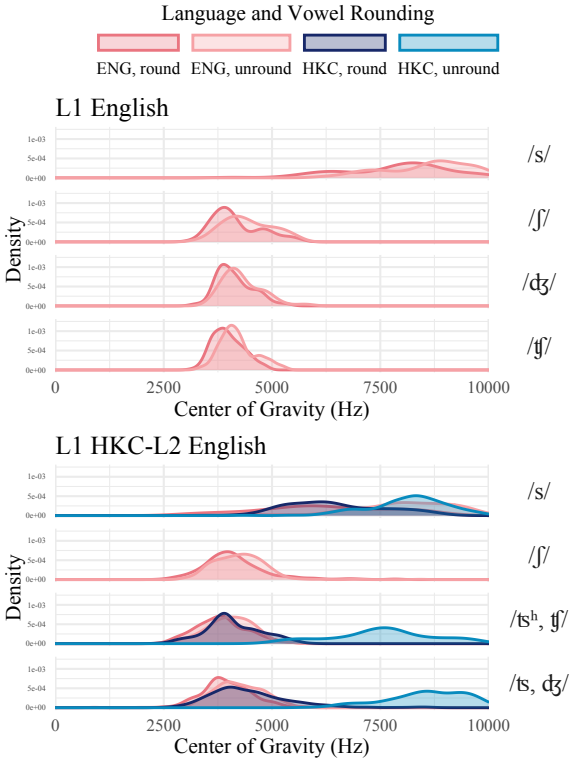


Figure 1: Spectral COG for all participants

In accordance with recent studies [8, 9], HKC speakers show evidence of an allophonic split for their L1 affricates, which differ significantly in COG according to vowel rounding ($p < 0.001$). Although the COG for /s/ (both L1 and L2) also varies significantly by vowel rounding ($p < 0.001$), the effect is much smaller.

With respect to acoustic similarity between L2 English and L1 HKC, it is observed that all three English post-alveolars pattern like the posterior allophones of the HKC affricates that occur in round environments. This is the case regardless of vowel rounding, and the L2 English affricates do not exhibit the same allophonic split observed in HKC. This result suggests an equivalence between the HKC allophones [t͡ʃʰ, t͡ʃ] and the English phonemes /t͡ʃ, d͡ʒ/. The new L2 sound [ʃ] is likewise produced with a low COG similar to that of the [t͡ʃʰ, t͡ʃ] allophones, and does not show an allophonic split.

Finally, the HKC and L2 English affricates, as well as [ʃ], show strong acoustic similarity to the L1 English post-alveolars. HKC and L1 English /s/ show similarly high COG in unround vowel environments, but HKC and L2 English /s/ show lower COG in round environments.

3.2. Ultrasound Results

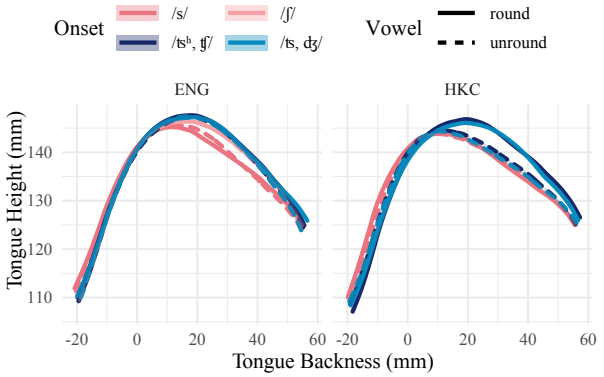


Figure 2: SSANOVA for HKC05. Tongue front to the right.

Two patterns of lingual articulation are observed among the HKC speakers. The predominant pattern, observed for six speakers, is shown in Figure 2. Corroborating the acoustic results, all HKC speakers clearly show an allophonic split for the two affricates. /t͡sʰ, t͡s/ show distinct apical vs. laminal tongue gestures according to vowel rounding. The same apical tongue gesture is also used for [s], although [s] does not vary by vowel rounding. Acoustic differences observed for [s] in round environments can therefore likely be attributed to anticipatory vowel rounding. In L2 English, these speakers show similarity both to their L1 HKC and to the native English speakers. L2 [s] is produced with an apical alveolar tongue gesture, while the three post-alveolars use a laminal post-alveolar gesture similar to that of pre-round HKC [t͡ʃʰ, t͡ʃ]. This pattern is identical to that observed for all native English speakers in this study.

The second pattern is shown in Figure 3. These speakers use similar tongue gestures for their L1 and L2 [s], and for [t͡sʰ, t͡s] in unround environments.

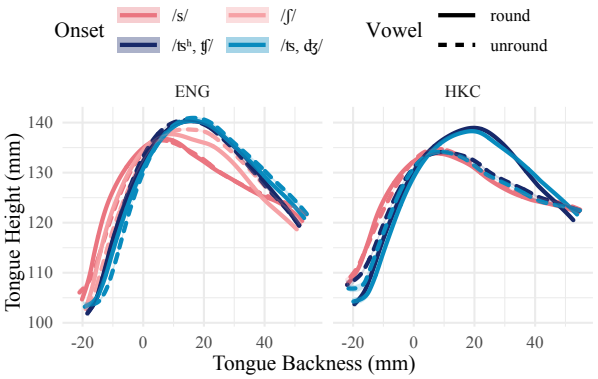


Figure 3: SSANOVA for HKC06.

Likewise, the tongue gestures for L2 [t͡ʃ, d͡ʒ] are similar to L1 [t͡ʃʰ, t͡ʃ] in round environments. Most notably, however, these speakers produce English [ʃ] with a tongue position intermediate to that of [s] and that of [t͡ʃ, d͡ʒ]. Moreover, [ʃ] shows an unexpected split according to vowel rounding, with a higher tongue blade in unround environments.

3.3. Lip Video Results

All native English speakers were consistent in their use of secondary lip rounding on the post-alveolar (but not alveolar) sibilants. For all HKC speakers, the HKC sibilants are variably rounded in accordance with the following vowel. However, the degree of interspeaker variability in rounding among HKC speakers is greater than in either acoustics or tongue position, with several distinct strategies. In L2 English, three speakers show a native English-like pattern with consistent rounding on all post-alveolars and coarticulatory rounding for [s].

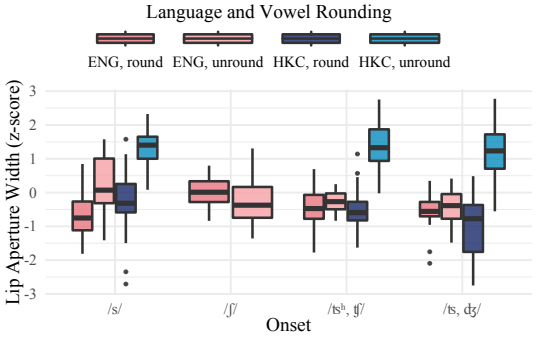


Figure 4: Lip rounding for HKC06

The speaker in Figure 4 (and two others) shows consistent rounding for the English affricates, similar to to the degree of rounding for HKC sibilants in round environments. However, she shows less consistent rounding for for English [s, ʃ], frequently rounding English [s] even in unround environments.

The speaker in Figure 5 (and one other), on the other hand, show the expected coarticulatory rounding for both HKC and English [s], and for the HKC affricates. However, these speakers use significantly less rounding for English post-alveolars, suggesting they have not fully acquired the secondary rounding typical of US/UK English.

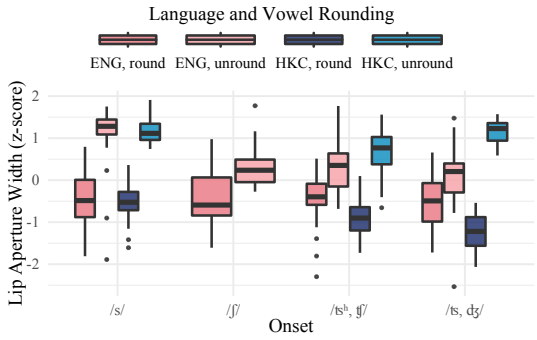


Figure 5: Lip rounding for HKC01

4. DISCUSSION AND CONCLUSION

This study finds substantial similarity between L1 HKC, L2 English, and L1 English sibilants, particularly in spectral center of gravity and tongue position. For the "similar" [11] sounds /t͡sʰ, t͡ʃ/ and /t͡s, d͡ʒ/, nearly all HKC speakers use the same tongue gestures in both their L1 and L2, yielding similar acoustic output for the affricate pairs. For the new sound /ʃ/, most speakers use the same posterior tongue gesture used for /t͡ʃ, d͡ʒ/, but two use novel gestures intermediate to [s] and [t͡ʃ]. For the identical /s/ sounds, speakers use the same tongue gesture in both languages, again with similar COG.

However, substantial interspeaker variation is observed with respect to lip rounding. Thus, like [12], this study finds that speakers do not necessarily use the same *combinations* of articulatory gestures in their L1 and L2. Rather, speakers may recruit individual gestures to achieve an L2 target. Yet these findings also contrast with [12] in that lip gestures were found to be more variable than tongue gestures, perhaps due to their association with coarticulatory vowel rounding in Cantonese and their status as secondary features in English.

Despite the Cantonese-English similarities, it cannot yet be determined to what extent allophonic split for the HKC affricates can be attributed to English influence. Real and apparent-time data (collection of which is underway) are needed to observe the early stages of that split, while sociolinguistic studies are necessary to determine whether the split is associated with Cantonese-English bilingualism. Nevertheless, this study provides phonetic data needed for such future work.

REFERENCES

[1] C. C. M. Sung, "Hong Kong English: Linguistic and sociolinguistic perspectives," *Language and Linguistics Compass*, vol. 9, no. 6, pp. 256–270, 2015.

[2] J. Setter, C. S. Wong, and B. H. Chan, *Hong Kong English*. Edinburgh University Press, 2010.

[3] T. T. N. Hung, "Towards a phonology of Hong Kong English," *World Englishes*, vol. 19, no. 3, pp. 337–356, 2000.

[4] S. Matthews and V. Yip, *Cantonese: A Comprehensive Grammar, Second Edition*. London: Routledge, 2011.

[5] A. Y. Chan and D. C. Li, "English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems," *Language Culture and Curriculum*, vol. 13, no. 1, pp. 67–85, 2000.

[6] K. K. Luke and J. C. Richards, "English in Hong Kong: Functions and status," *English World-Wide*, vol. 3, no. 1, pp. 47–64, 1982.

[7] K. H. Cheung, "The phonology of present-day Cantonese," doctoral dissertation, University of London, 1986.

[8] K.-L. Chan, "The sound change of [ts, tsʰ, s] to [tʃ, tʃʰ, ʃ] in Hong Kong Cantonese," Master's thesis, The Hong Kong University of Science and Technology, Hong Kong, 2007.

[9] P. H. Yeung and J. Havenhill, "Acoustic ambiguity and articulatory re-analysis: Variation and change in the Hong Kong Cantonese sibilants," to appear.

[10] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language experience in second language speech learning: In honor of James Emil Flege*, O.-S. Bohn and M. J. Munro, Eds. Philadelphia, PA: John Benjamins, 2007, pp. 13–34.

[11] J. E. Flege and O.-S. Bohn, "The revised speech learning model (SLM-r)," in *Second Language Speech Learning*, R. Wayland, Ed. Cambridge: Cambridge UP, 2021.

[12] M. Oakley, "Articulating non-native vowel contrasts," doctoral dissertation, Georgetown University, 2021.

[13] T. Eckes and R. Grotjahn, "A closer look at the construct validity of C-tests," *Language Testing*, vol. 23, no. 3, pp. 290–325, 2006.

[14] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[15] Articulate Instruments Ltd., *Articulate Assistant Advanced User Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd., 2012.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[17] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *The Journal of the Acoustical Society of America*, vol. 108, pp. 1252–1263, 2000.

[18] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[19] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut," *Sensors*, vol. 22, 2022.

[20] J. Mielke, "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2858–2869, 2015.

[21] J. Havenhill and Y. Do, "Visual speech perception cues constrain patterns of articulatory variation and sound change," *Frontiers in Psychology*, vol. 9, p. 728, 2018.

THE RELATIONSHIP BETWEEN VOWEL CHANGE AND NASAL LOSS IN THE CHENGDU DIALECT OF MANDARIN CHINESE: EVIDENCE FROM RT-MRI

Sishi Liao, Phil Hoole, Jonathan Harrington

Institute for Phonetics and Speech Processing (IPS), LMU Munich
sishi.liao | hoole | jmh@phonetik.uni-muenchen.de

ABSTRACT

This study aims to examine whether the pre-nasal vowel change that commonly occurs with the phonologization of nasalization, such as the raising and possibly fronting (e.g. Latin 'manus' to French /mɛ̃/), extends beyond the language family and can be generalized to where the nasal trace is lost (e.g. /an/-rime with nasal coda loss in the Chengdu dialect). To do so, real-time MRI was applied to obtain vocal tract aperture signals in the palatal and hyperpharyngeal areas in different rimes (/a, an, i, o/) from four L1 Chengdu speakers. The distances of /a, an/ to anchors /i, o/ were compared in a transformed articulatory space, demonstrating that the vowel in /an/-rime is raised and fronted compared to that in /a/-rime. This concomitant vowel shift leads us to conclude that the universal pre-nasal vowel shift does extend to the nasal loss cases where the VN has become an oral vowel.

Keywords: Chengdu dialect; sound change; vowel raising; vowel fronting; /an/-rime

1. INTRODUCTION

In a vowel-nasal (VN) sequence, the degree and temporal extent of nasalization on the preceding vowel can vary due to anticipatory nasalization [1]; this coarticulatory variation can evolve into sound changes [2], [3] that ultimately result in the reduction of the nasal consonant [4]. As a result, the VN sequence can turn into either a nasal vowel Ṽ [5] or a raised plain oral vowel V̝ [6].

Phonologization of nasalization or nasal loss is commonly accompanied by the pre-nasal vowel height shift, where the vowel is alleged to be higher for low vowels and lower for high vowels, leading to vowel height centralization [7]. This nasal-induced shift in vowel height is thought to result from the intra-speaker interplay between perception and production of speech. Specifically, when low vowels are nasalized, they are perceived as phonetically higher [8], which can lead to tongue dorsum raising [9] during its production.

This nasal-induced vowel height shift is a well-documented phenomenon in phonetics [10] and phonology [11]–[13] but has not been extensively investigated in the VN → V̝ cases [6]. One example of such cases exists in the Chengdu dialect of Mandarin Chinese [14], in which the nasal consonant of /(V)an/-rime has been lost, and the whole rime has become oral. Considering that sound changes often involve the reweighting of multiple cues [15], an unresolved issue is whether the loss of the nasal consonant is associated with changes in vowel quality.

Vowel height/backness are often estimated from the first two formant frequencies in studies of dialect variation [16] and sound changes [17], [18]. However, the relationship between formants and the articulatory data can be non-linear and sometimes ambiguous [19], [20]. To overcome this limitation, researchers have increasingly turned to advanced technology to investigate the configuration of the vocal tract. Techniques such as ultrasound [21], electromagnetic articulography (EMA) [22], and real-time MRI [23]–[25] can provide clearer visual representations of the vocal tract and its movement, allowing for more accurate interpretation of vowel quality in terms of physiological tongue position.

Many studies [24], [26] on vowel and consonant quality go well beyond examining values at specific moments by investigating the dynamic shape of the signal trajectory, using methods such as functional principal component analysis (FPCA) and discrete cosine transform (DCT) [27], [28]. For example, [29] examined the /u/-fronting in the 3D space composed by the first three DCT coefficients of the F2 and compared the relative Euclidean distance of the /u/ to two anchors /a, i/ between younger and older speakers. Another study [30] analyzed /s/-retraction in the sibilant of /str/ by examining the first spectral moment (M_1) trajectory; they compared the first 3 DCT coefficients using the normalized orthogonal projection onto the line connecting the canonical /s/ and /ʃ/ for each speaker.

This study compares the vocal tract constriction in the palatal and hyperpharyngeal regions (relating to tongue height and backness) between the vowels in the historically nasalized /an/-rime, which has become oral [14], and those in /a/-rime followed by a

nasal. The aim is to investigate whether there are changes in vowel quality accompanying the nasal loss in this dialect [14], similar to the vowel height shift in cases where nasality has been preserved [2].

2. METHODOLOGY

2.1. Participants and experiment settings

This study recruited four first-language speakers of the Chengdu dialect (2 females, with an average age of 25.75 years) with no self-reported speech-language disorders. Prompt texts and videos explaining the experiment procedures were sent to the participants in advance, who also signed the consent forms prior to the experiment.

The experiment took place at the Max Planck Institute for Multidisciplinary Sciences in Göttingen, Germany. Real-time MRI data were collected with a 3T MRI system, with the participant in a supine position in the machine and the head placed in a head coil. The monitor showed images of the randomized prompt texts, which were reflected through a set of mirrors to the participants' vision. The texts were split into blocks of PowerPoint slides, with each block consisting of about 16 slides and lasting about one minute; any block with mistakenly pronounced tokens was repeated.

2.2. Speech materials

The speech material included four types of sentences differing in the target segments. Each target segment consisted of a CV_1NV_2 sequence, where C is an onset consonant /p, p^h, f/, V_1 is a nucleus vowel /a, i, o/ with four tones, N is an alveolar nasal /n/, and V_2 is /a/ if V_1N is tautosyllabic, or /ai/ if V_1N is heterosyllabic. The meaning of the sentence was either "she is called <X> aunty" or "she is called <X> grandma", both of which have the same syntactic structure and convey similar semantic meanings. An example set of sentences with C being /p/, and V_1 varying in /a, i, o/ are shown with characters and respective pronunciations in IPA in **Table 1**.

A total of 301 tokens were analyzed, excluding all the mispronounced ones. The target interval in each utterance was segmented and labeled using Praat [32]. In the $CV_1N.V_2$ cases, where the nasal coda was claimed to be lost [14] (realized as $CV_1.V_2$), the target interval began at the voice onset of V_1 and ended at the V_1-V_2 transition; in the $CV_1.NV_2$ cases, the interval also began at the voice onset of V_1 and ended at the transition from V_1 to N, where the nasal damped the intensity and caused the formation of anti-resonance. The reason for /i, o/ being the anchors is that they form the most peripheral contrast in height and backness in the vowel inventory of this dialect, so that /a, an/-rimes could be compared more accurately in both dimensions.

Target segments		IPA of the sentences				
$V_1(N)$-rime		她	叫	**X**	阿	姨
$CV_1(N).V_2$	pa(n).V_2	t^ha	tɕiɑʊ	**pa(n)**	**a**	i
V_1-rime		她	叫	**X**	奶	奶
$CV_1.NV_2$ anchors: /i, o/	pa.nV_2	t^ha	tɕiɑʊ	**pa**	**naɪ**	naɪ
	pi.nV_2	t^ha	tɕiɑʊ	**pi**	**naɪ**	naɪ
	po.nV_2	t^ha	tɕiɑʊ	**po**	**naɪ**	naɪ

Table 1: An example set of sentences (with onset being /p/) in IPA, with target segments in bold and target vowels underlined. The nasal consonant is presumed to be lost in the /an/-rime words.

2.3. Real-time MRI and vocal tract aperture

The real-time MRI data were reconstructed by means of [33] with a frame rate of 50 fps. The MR images were recorded synchronously with noise-suppressed audio so that the time stamps derived from the audio segmentation could be mapped onto the image-derived time series signals.

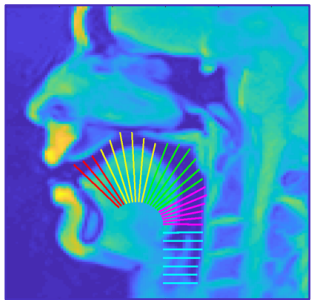


Figure 1: An example of the grid lines used to measure the aperture along the vocal tract (red: alveolar, yellow: palatal, green: velar, purple: hyperpharyngeal, blue: hypopharyngeal)

The analysis of the vocal tract configuration was based on the methods in [24]. The first step of processing the MR images was image registration, which eliminated the influence of any possible slight head movement. Then 28 grid lines were mapped onto the MR image, 20 of which were semi-polar lines extending from the alveolar ridge to the pharynx, 8 of which were horizontal lines from the pharynx to the glottis. The 28 gridlines were divided into five articulatory regions: alveolar, palatal, velar, hyperpharyngeal, and hypopharyngeal. For each articulatory region, the measure of vocal tract constriction was derived based on the average pixel intensity along the gridlines in each region. This measurement in the respective region increases when there is more high-intensity tongue tissue and decreases when there is more low-intensity air; see [25] for more details. For each

speaker, the value of each articulatory signal was normalized to [0, 1] based on the overall minimum and maximum occurring in the whole corpus.

2.4. Discrete Cosine Transform

In each articulatory region, the derived vocal tract displacement signal within the vowel interval was resampled to 1000 data points using the function in the `fdasrsf` package [34] in Python. The first DCT coefficient (*k0*) for each resampled curve was then calculated with the function (1):

$$k_0 = \frac{\sqrt{2}}{N}\sum_{n=0}^{N-1} x(n) \quad (1)$$

To establish a transformed articulatory space, we used the *k0*_hyperpharyngeal and *k0*_palatal values from each vowel interval as the coordinates on the x- and y-axis, respectively. In this two-dimensional space, a token with a more raised and fronted tongue position would locate closer to /i/, and a token with a lowered and more retracted tongue position would locate closer to /o/.

2.5. Orthogonal projection ratio and statistics

To determine the vowel quality differences between /a/-rime and /an/-rime for each speaker, the orthogonal projection ratio (*op* ratio) was calculated with /i, o/ as anchors in the following manner. The centroids of the two anchors were determined on the <*k0*_hyperpharyngeal, *k0*_palatal> space and were then connected. The *op* ratio onto this line was then calculated for each token; the two centroids of /i, o/ have an *op* ratio of ±1. For any token, a positive *op* ratio denotes proximity towards /i/, a negative *op* ratio denotes proximity towards /o/, and an *op* ratio = 0 means this token is of equal distance towards /i/ and /o/.

A linear mixed-effect model was applied to the *op* ratios using the `lmer` function in the `lmerTest` package [35] from `R` to test the proximity of the /a, an/ tokens to the anchors /i, o/. The *op* ratio of each /a, an/ observation was set as the response, the `RIME` as the fixed factor, and the `SPEAKER`, `ONSET`, and `TONE` as the random factors. The model was then applied with ANOVA to test the difference between /a, an/ in this transformed articulatory space.

3. RESULTS

3.1. Movement trajectories and DCT coefficients

The degree of tongue height/backness during the interval of /a, an, o, i/, denoted by the articulatory signal in the palatal/hyperpharyngeal region, was resampled and plotted as a function of time in **Figure 2**. The left panel shows the tongue displacement in the palatal region: the larger the value, the higher the tongue position. The right panel shows the tongue displacement in the hyperpharyngeal region: the larger the value, the more retracted the tongue.

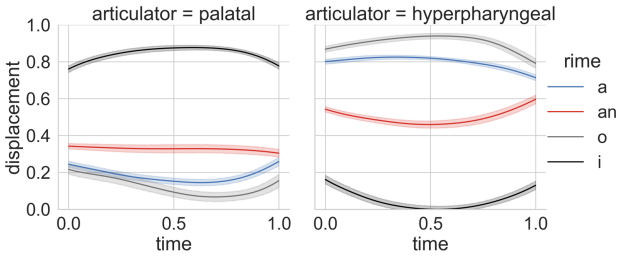


Figure 2: The tongue displacement as a function of time. The larger the value, the higher (left) / retracted (right) the tongue.

In the palatal region, the overall tongue displacement of the vowel in /an/-rime was larger than that in /a/ rime, showing a higher tongue position for /an/ than /a/. In the hyperpharyngeal region, the overall tongue displacement of the vowel in /an/-rime was less than that in /a/-rime, showing a fronter tongue position for /an/ than /a/.

To compare the degrees of tongue height and backness between /a, an/-rimes for each speaker, the DCT coefficient *k0* for both the palatal and hyperpharyngeal regions were plotted in a 2D space in **Figure 3**. Each point represents one utterance token, with the DCT coefficient *k0* of the hyperpharyngeal signal on the x-axis, and the *k0* of the palatal signal on the y-axis.

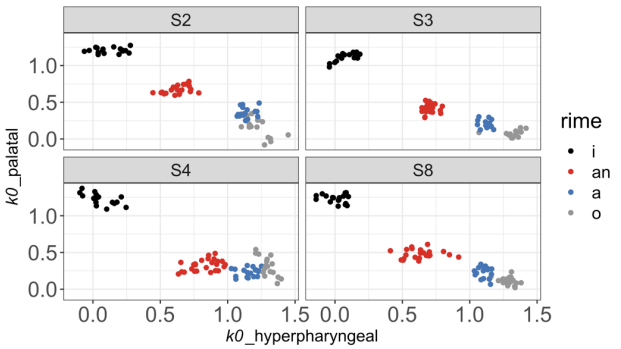


Figure 3: The DCT *k0* space (hyperpharyngeal × palatal) of each utterance token from each speaker, in which speaker S2 and S4 were female.

3.2. Orthogonal projection

The means of /i/- and /o/-rimes in the DCT *k0*_hyperpharyngeal × *k0*_palatal space were calculated and centered at ±1 in the orthogonal projection ratio plot, respectively colored in black and gray. In this transformed articulatory space, the /an/-rime was closer to

/i/ and farther from /o/, compared to the /a/-rime, denoting a more raised and fronted tongue position for /an/-rime. See **Figure 4**.

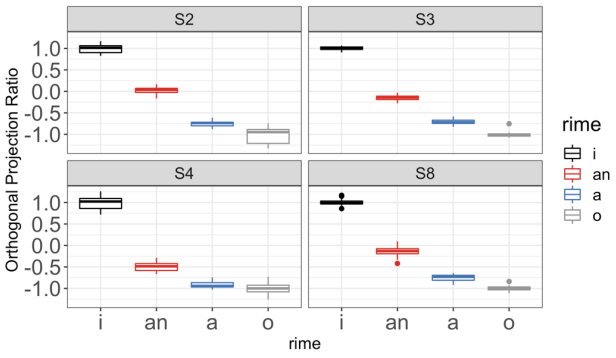


Figure 4: The orthogonal projection ratio in DCT *k0*_hyperpharyngeal × *k0*_palatal space.

3.3. Statistics

If vowels in /an/-rime words in the Chengdu dialect were produced with a more raised and fronted tongue position than that in /a/-rime words, then their *op* ratios should differ in the DCT *k0* space. The random factor `TONE` caused a lack of convergence and was thus not included in the model. The output of the model (**Table 2**) with 162 observations showed that the average *op* ratio was 0.59 higher (SE = 0.07, 95% CI = [0.37, 0.81]) for the vowels in /an/-rime than that in /a/-rime.

	Est.	Std. E	df	t val.	Pr(>|t|)
(Intercept)	-0.78	0.05	3.02	-17.11	0.000 ***
rime.an	0.59	0.07	3.00	8.11	0.004 **

Table 2. The results for the `lmer` model: *op* ratio ~ rime + (rime | speaker) + (1 | onset).

The results of the lmer model showed that there was a significant difference in the *op* ratio between /a/- and /an/-rime ($F[1, 3.00] = 65.75$, $p < 0.01$).

4. DISCUSSION

The result of this study shows that in the Chengdu dialect of Mandarin Chinese, the vowels in /an/-rimes are pronounced with a significantly higher and fronter tongue position than that in /a/-rime in a transformed articulatory space. This finding of the concomitant vowel raising and fronting with the nasal coda loss in /an/-rime [14] (VN → V̰˞) is consistent with other evidence showing vowel height centralization accompanying synchronic vowel nasalization [7].

This sound change does not seem to be derived from language contact because, in the more dominant common language – standard Mandarin, the nasal coda in /an/-rimes has been maintained. It would be necessary for a future investigation on the interaction between vowel quality change and nasal coarticulation, specifically in other Mandarin varieties that may have varying degrees of nasality in the /(V)an/-rimes. Further investigations are also needed to establish the diachronic sequence of vowel raising and nasal loss, i.e., whether the changes in the present study correspond to VN→ V̰˞N→V̰˞ or VN→Ṽ→ V̰˞.

5. ACKNOWLEDGMENT

The first author is supported by the China Scholarship Council (grant No. 202008440473). This study is supported by the InterAccent project, which receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 742289). Website: https://www.phonetik.uni-muenchen.de/Forschung/interaccent/interAccent.html

6. REFERENCES

[1] P. S. Beddor, "A coarticulatory path to sound change," *Language (Baltim).*, vol. 85, no. 4, 2009, doi: 10.1353/lan.0.0165.

[2] J. J. Ohala, "The listener as a source of sound change: An update," in *The Initiation of Sound Change*, Maria-Josep Solé and Daniel Recasens, Eds. John Benjamins, 2012, pp. 21–36.

[3] J. J. Ohala and M. Ohala, "The phonetics of nasal phonology: Theorems and data," *Nasals, Nasalization, and the Velum*, vol. 5, pp. 225–249, 1993, doi: 10.1016/B978-0-12-360380-7.50013-2.

[4] J. J. Ohala and M. G. Busa, "Nasal loss before voiceless fricatives: a perceptually-based sound change," *Riv. Linguist.*, vol. 7, pp. 125–144, 1995.

[5] B. L. Rochet, "The formation and evolution of the French nasal vowels," *Form. Evol. French nasal vowels*, Dec. 1976, doi: 10.1515/9783111328287.

[6] M. Cresci, "VN > V in Camuno: An alternative historical pathway to nasal loss," *Ital. J. Linguist.*, vol. 31, no. 1, pp. 61–92, 2019, doi: 10.26346/1120-2726-132.

[7] P. S. Beddor, "Phonological and Phonetic Effects of Nasalization on Vowel Height," Indiana University, 1983.

[8] R. A. Krakow, P. S. Beddor, L. M. Goldstein, and C. A. Fowler, "Coarticulatory influences on the perceived height of nasal vowels," *J. Acoust. Soc. Am.*, vol. 83, no. 3, p. 1146, Aug. 1988, doi: 10.1121/1.396059.

[9] P. S. Beddor, R. A. Krakow, and L. M. Goldstein, "Perceptual Constraints and Phonological Change: A Study of Nasal Vowel Height," *Phonol. Yearb.*, vol. 3, pp. 197–217, 1986.

[10] T. Arai, "Comparing Tongue Positions of Vowels in Oral and Nasal Contexts," in *INTERSPEECH*, 2005, pp. 1033–1036.

[11] R. Shosted, C. Carignan, and P. Rong, "Managing

the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi," *J. Acoust. Soc. Am.*, vol. 131, no. 1, p. 455, Jan. 2012, doi: 10.1121/1.3665998.

[12] C. Carignan, "Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels," *J. Phon.*, vol. 63, pp. 87–105, 2017, doi: 10.1016/J.WOCN.2017.04.005.

[13] N. Almeida, S. Silva, C. Cunha, and A. Teixeira, "Data-Driven Analysis of European Portuguese Nasal Vowel Dynamics in Bilabial Contexts," *Appl. Sci.*, vol. 12, no. 9, p. 4601, 2022, doi: 10.3390/APP12094601.

[14] S. Liao *et al.*, "Nasal Coda Loss in the Chengdu Dialect of Mandarin: Evidence from RT-MRI," *INTERSPEECH*, pp. 1347–1351, 2022, doi: 10.21437/INTERSPEECH.2022-10775.

[15] J. Kuang and A. Cui, "Relative cue weighting in production and perception of an ongoing sound change in Southern Yi," *J. Phon.*, vol. 71, pp. 194–214, Nov. 2018, doi: 10.1016/J.WOCN.2018.09.002.

[16] J. Riverin-Coutleé, E. Kapia, C. Cunha, and J. Harrington, "Vowels in urban and rural Albanian: The case of the Southern Gheg dialect," *Phonetica*, vol. 79, no. 5, pp. 459–512, Oct. 2022, doi: /10.1515/phon-2022-2025.

[17] J. Harrington, "Evidence for a relationship between synchronic variability and diachronic change in the Queen's annual Christmas broadcasts," *Lab. Phonol.*, no. 9, pp. 125–143, 2007.

[18] J. Riverin-Coutlée and J. Harrington, "Phonetic change over the career: a case study," *Linguist. Vanguard*, 2022, doi: 10.1515/LINGVAN-2021-0122.

[19] P. Hoole, "Articulatory-acoustic Relations in German Vowels," in *14th International Congress of the Phonetic Sciences*, 1999, pp. 2153–2156.

[20] A. B. Ximenes, J. A. Shaw, and C. Carignan, "A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English," *J. Acoust. Soc. Am.*, vol. 142, no. 1, p. 363, Jul. 2017, doi: 10.1121/1.4991346.

[21] C. Carignan, "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels," *J. Acoust. Soc. Am.*, vol. 143, no. 5, p. 2588, May 2018, doi: 10.1121/1.5034760.

[22] T. Rebernik *et al.*, "A review of data collection practices using electromagnetic articulography," *Lab. Phonol.*, vol. 12, no. 1, pp. 1–42, Mar. 2021, doi: 10.5334/LABPHON.237.

[23] C. Carignan *et al.*, "Planting the seed for sound change: Evidence from real-time mri of velum kinematics in German," *Language (Baltim).*, vol. 97, no. 2, pp. 333–364, 2021, doi: 10.1353/LAN.2021.0020.

[24] C. Carignan *et al.*, "Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI," *Lab. Phonol.*, vol. 11, no. 1, pp. 1–26, Mar. 2020, doi: 10.5334/LABPHON.214.

[25] E. Kunay *et al.*, "Vowel height and velum position in German: Insights from a real-time magnetic resonance imaging study," *J. Acoust. Soc. Am.*, vol. 152, no. 6, p. 3483, Dec. 2022, doi: 10.1121/10.0016366.

[26] M. Pouplier, J. Cederbaum, P. Hoole, S. Marin, and S. Greven, "Mixed modeling for irregularly sampled and correlated functional data: Speech science applications," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 935–946, 2017, doi: 10.1121/1.4998555.

[27] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *J. Phon.*, vol. 49, pp. 16–40, Mar. 2015, doi: 10.1016/J.WOCN.2014.10.001.

[28] J. Harrington and F. Schiel, "/u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change," *Language (Baltim).*, vol. 93, no. 2, pp. 414–445, 2017, doi: 10.1353/LAN.2017.0019.

[29] J. Harrington, F. Kleber, and U. Reubold, "Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2825–2835, 2008, doi: 10.1121/1.2897042.

[30] M. Stevens, J. Harrington, F. Schiel, M. Stevens, J. Harrington, and F. Schiel, "Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English," *Glossa a J. Gen. Linguist.*, vol. 4, no. 1, 2019, doi: 10.5334/GJGL.620.

[31] C. Carignan, R. Shosted, C. Shih, and P. Rong, "Compensatory articulation in American English nasalized vowels," *J. Phon.*, vol. 39, no. 4, pp. 668–682, 2011, doi: 10.1016/J.WOCN.2011.07.005.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. [Online]. Available: http://www.praat.org/

[33] M. Uecker, S. Zhang, D. Voit, A. Karaus, K. D. Merboldt, and J. Frahm, "Real-time MRI at a resolution of 20 ms," *NMR Biomed.*, vol. 23, no. 8, pp. 986–994, 2010, doi: 10.1002/NBM.1585.

[34] J. D. Tucker, W. Wu, and A. Srivastava, "Computational Statistics and Data Analysis Generative models for functional data using phase and amplitude separation," *Comput. Stat. Data Anal.*, vol. 61, pp. 50–66, 2013, doi: 10.1016/j.csda.2012.12.001.

[35] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/JSS.V082.I13.

Dynamic articulatory and acoustic features of Hungarian sibilants as a function of phonological voicing

Tekla Etelka Gráczi[1,2], Kornélia Juhász[1,2,3], Tamás Gábor Csapó[2,4], Andrea Deme[2,3], Alexandra Markó[1,2]

[1]Hungarian Research Centre for Linguistics, [2]MTA–ELTE "Lendület" Lingual Articulation Research Group, [3]Eötvös Loránd University, [4]Dept. of Telecomm. and Media Informatics, Budapest Univ. of Technology and Economics,
graczi.tekla.etelka@nytud.hu, juhasz.kornelia@nytud.hu, csapot@tmit.bme.hu, deme.andrea@btk.elte.hu, marko.alexandra.phd@gmail.com

ABSTRACT

It is a phonetic cliché that obstruents' phonological voicing shows various phonetic patterns due to aerodynamic constraints. We analysed articulatory and acoustic features of Hungarian intervocalic sibilants /z, ʒ, s, ʃ/.) Midsagittal Ultrasound Tongue Imaging was used to track tongue root, tongue tip and tongue blade along the total duration of the consonants (83 images per seconds). From the acoustic signal, centre of gravity (CoG), spectral skewness, voicing profile (all at 11 equidistant temporal points), and voiceless part ratio were extracted. Tongue root was advanced in /z, ʒ/ compared to /s, ʃ/; however, in some cases, its backward movement was also found in both voiced and voiceless sibilants, which might appear as compensation for the lowering of the tongue front to maintain the intraoral pressure for intense turbulent frication: CoG did not lower, neither skewness increased considerably at the end of the consonants despite the restart of phonation in the voiced sibilants.

Keywords: tongue root movement, tongue front movement, spectral features, phonation, sibilants.

1. INTRODUCTION

Phonation and obstruent production are contradictory targets: phonation requires low supraglottal air pressure, while the oral seal leads to the increase of intraoral pressure [1, 2]. This contradiction is more enhanced in sibilants, as their aerodynamic target is a high intensity turbulent frication which can be reached by increasing the air pressure behind the oral obstacle.

One of the possible articulatory manoeuvres to maintain phonation is to produce voiced obstruents with an advanced tongue root (TR) compared to their unvoiced counterparts [3, 4, 5]. While the tongue root advancement of voiced consonants is a well described phenomenon, its dynamic features and its interrelation with the spectral features are less documented so far. In a study of Polish and Italian stops, TR moved forward during the preceding vowel for both voiced and voiceless consonants, and it was more advanced for voiced stops in both languages already at the start of the closure, than for their voiceless counterparts [6]. In a comparison of Hungarian /z/ and /s/, [7] found that the TR was more advanced in half of the speakers (6 out of 12) at the start of the consonant duration, and in 75% of them (8 speakers) during the entire duration of the consonant.

The contradictory aims of phonation and obstruent production might also lead to partial devoicing of the voiced obstruents. The presence of TR advancement was not found to vary among speakers as a function of phonetic voicing, i.e., this phenomenon was recognised both in the group of speakers who tended to maintain voicing, and in the group of speakers who devoiced their /z/-realisations [7]. The degree of TR's advancement, however, was found to vary with the ratio of phonation to the consonant duration [8].

The present study raises the question how the TR position changes during the production of intervocalic voiced and voiceless sibilants, and how this change interacts with the contradictory targets of turbulent frication and phonation. We hypothesized that (i) the TR is advanced in voiced targets along the sibilants' total duration to maintain phonation, (ii) the TR position is different in alveolar and postalveolar places of articulation due to the volume difference in the obstacle-vocal fold distance, (iii) tongue front and TR position changes along the sibilants' time interval in a manner to maintain intense turbulent frication.

2. METHODS

Four Hungarian sibilants /s, z, ʃ, ʒ/ were analysed in nonsense words of the structure /l/V_1_V_1/l/, where the analysed sibilants appeared at the onset of the second syllable, and the two neighbouring vowels were identical, either /ɒ, aː, ɛ/ or /i/. Two of these sequences *(lesel, leszel)* are meaningful words in Hungarian. Each item was read out five times by three native female speakers of Hungarian (aged between 35 and 47 ys) in a randomized order. Altogether 4 (C) * 4 (V) * 5 (repetitions) = 80 items per speaker were analysed. The items were recorded together with further nonsense words that are out of the focus of the

current study. The recording was carried out by AAA speech ultrasound system (Articulate Instruments LtD.), which recorded midsagittal tongue contours at a frame rate of 83 image/s. The speech signal was recorded by Beyerdynamic TG H56c omnidirectional condenser microphone.

The segment boundaries were automatically labelled [9] and corrected manually in Praat [10]. The voiceless part ratio (VPR) of the sibilants was retrieved via Praat's voice report function (fraction of locally unvoiced frames; f_0 range was set to 75–300 Hz, for further settings, the standard values were kept). Centre of Gravity (CoG) and spectral skewness were measured automatically unfiltered along the time course of the sibilants at 11 equidistant time points with 0.01 s window. No filtering was used in order to include the effect of the presence or absence of voicing in the results of the spectral measures. The presence or absence of phonation at the same time points was retrieved by the Pitch (ac) function of Praat with the same settings as VPR for all three speakers. Voicing profile is defined by the ratio of phonation present at a time point in the specific consonant [see 11]. While VPR shows how much percent of each consonant realisation's duration is devoiced in percentages, the voicing profile shows how many of all the consonant realisations had maintained/restarted the phonation at each analysed time point. For example, a VPR of 0% means one token with phonation all along its total duration, while a voicing profile of 0% at measurement point 5 means that none of that phoneme's realisations exhibited phonation at this specific point of their duration.

Each tongue contour along the sibilants' duration was drawn manually in the AAA software by the first two authors. Approximately 12.7 tongue contours per sibilant realisations, altogether 3031 tongue contours were manually tracked and analysed.

The statistical analyses were carried out in R [12]. The possible difference of VPR among the consonants and vowel context was analysed using Linear Mixed Models (LMM) [13, 14, 15]. VPR was set as a dependent variable. The most complex model included (PHONOLOGICAL) VOICING, PLACE OF ARTICULATION (henceforth: PoA) and VOWEL (CONTEXT) as fixed factors allowing for their interactions along with random slope on VOWEL, POA and VOICING by SPEAKERS. The model selection (avoiding overfitting, and model simplification by anova()) resulted in the model where the fixed factors were VOICING and VOWEL allowing for interaction, and the random effects included only random intercept by SPEAKER. Pairwise comparisons were carried out via Tukey post hoc tests [16].

Generalised additive mixed models (GAMMs) were run to analyse CoG and skewness differences among the consonants in general, and along their duration [17, 18]. CoG, and skewness served as dependent variables. CONSONANT was set as ordered factor with contrast treatment, and set as parametric term. The smooth terms included TIME, TIME by CONSONANT (ordered factor). Random smooth was included on TIME by ITEM. No further autoregression treatment was needed.

Tongue contours were analysed by polarGAMs, separately for each speaker [19]. Vertical position of the tongue points (*y* dimension of coordinates) was set as dependent variable. The parametric terms contained the ordered factors of CONSONANT with contrast treatments. The smooth terms contained the TONGUE POINT (*x* dimension of the coordinates) by the CONSONANT, the NORMALISED TIME POINT within the consonant independently and by the CONSONANT (ordered factor), and the interaction of the TONGUE POINT and TIME POINT by the CONSONANT (using the te() function). In order to obtain *p*-values for each important consonant pair, two models were run with differently ordered CONSONANT levels. Due to computation capacity no random terms, but AR1 type autoregression correction was used.

The tongue front and root movements were analysed at three points of the estimated tongue contours taken from the TIME POINT * TONGUE POINT interactions of the polarGAMMs. The first column represents the tongue tip. It was not chosen from the edge of the estimates, as the GAMs methods may result in fuzzy data at the edges of the estimated ranges. This "tongue tip" point (TT) was taken from the closest point to the most anterior point of the tongue contour with largest movement excursion during the alveolars (for each speaker). The "tongue blade" point (TB) was selected based on the largest excursion during the postalveolars behind the TT point. The "tongue root" (TR) point was selected based on the heat map results, choosing the most typical point for the movement of the TR's area.

3. RESULTS

3.1. Acoustic data

The VPR of the voiced sibilants showed large variation within speakers, as well as between speakers. Eliminating the outliers, the VPR of spk01's voiced sibilants ranged between 0–75% (/z/), 0–60% (/ʒ/), spk02's between 0–100% (both), spk03's between 97–100% (/z/) and 67–100% (/ʒ/). Spk01's voiceless sibilants had 85–100% IQR of VPR (both), spk02's ones had 100–100% (/s/) and 80–100% (/ʃ/), and spk03's /s, ʃ/ had 100–100%, 67-100%, respectively. The best-fitting model was the one that included PHONOLOGICAL VOICING, VOWEL

CONTEXT and their interaction as fixed effects, and random intercept by SPEAKER (R_m^2=0.309, R_c^2=0.513). This means that the PoA did not considerably affect the VPR. The main effect of PHONOLOGICAL VOICING ($F(1,229)=137.597$, $p<0.01$) and its interaction with VOWEL CONTEXT ($F(3,229)=2.817$, $p=0.04$) had a significant effect on this measure. The post hoc test of the interaction did not reveal any significant effect of the VOWEL CONTEXT on the difference between sibilants of the same phonological voicing.

The mean CoG and the mean skewness were significantly different among the four sibilants according to the parametric terms of the GAMMs (R^2=0.958, and 0.820, respectively). The smooth terms showed significant differences for both measures' change along the time course among all consonants. Plotting the smooth's estimated differences showed that the voiced counterparts' CoG-values were significantly different from appr. the 30% to 100% (both alveolars and postalveolars), the skewness from appr. the 25% to 85% (alveolars) and 70% (postalveolars) of the consonant duration. Figure 1 indicates that the CoG started increasing, and the skewness started decreasing for each speaker's sibilants at the time point where the voicing profile revealed a rapid drop in the phonation possibility, i.e., from which time point the phonation tended to cease in the speaker's pronunciation. While skewness of the voicing counterparts converged after the restart of voicing, CoG values remained significantly different between them.

3.2. Tongue positions and movements

Tongue position, and tongue movement were analysed by polarGAMs (R^2 = 0.855, 0.885, 0.886 for the three speakers respectively). The smooth of the TONGUE POINT averaged across the duration was significantly different between both voicing and PoA counterparts (Table 1, $p \leq 0.02$). This means that the tongue positions in general were different across the consonants. The smooth of the NORMALISED TIME POINT and the tensor of the interaction of TONGUE POINT and NORMALISED TIME POINT can be better understood by investigating the estimated smooths.

Figure 2 introduces three typical points for TT, TB, TR of the tongue contours. The TR's position was more advanced in the alveolars than in the postalveolars; however, it was speaker-specific, whether it was only a difference between the sibilants of the same phonological voicing category or in general. The TR was more advanced in the voiced sibilants during the entire duration in all speakers' pronunciation. Its position, however, changed during the consonant and moved backward for all four sibilants of spk03, and for the postalveolars of spk02. The TR in spk03's /s/ moved into a more anterior position, and in /z/, it did not change its position. In spk01's speech the TR moved backward, and in the last third of the production, the TR moved forward again in all 4 sibilants. The backward movement during the consonants may appear to tighten the pharyngeal region in order to control the intraoral pressure, which is responsible for the intense turbulent noise of the sibilants.

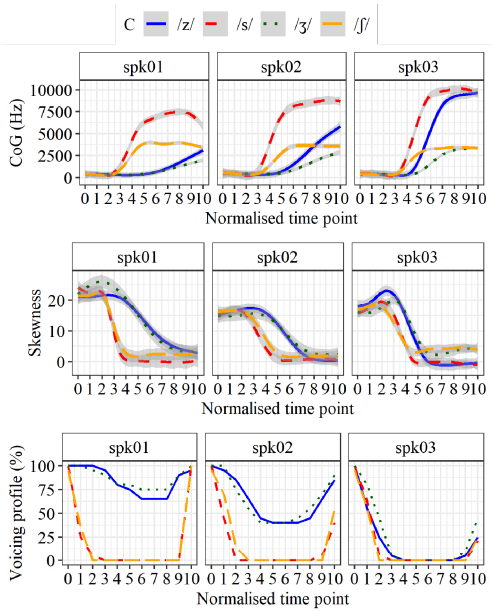


Figure 1: CoG (Hz, top), and spectral skewness (mid) estimated by GAM-smoothing by ggplot2 [20], and voicing profile (bottom) along the duration of the sibilants. (Color online.)

C-pair	F_{spk01}	F_{spk02}	F_{spk03}
/z/ vs. /s/	41.355	29.165	14.197
/ʒ/ vs. /ʃ/	13.998	69.641	3.681
/s/ vs. /ʃ/	44.290	29.165	38.582
/z/ vs. /ʒ/	14.699	7.998	4.861

Table 1: The polarGAM-results.

TT and TB followed similar movement patterns for all sibilants within the speakers. It raised in the first part of the duration and then lowered from the third or half of the consonant duration in spk01's and spk02's pronunciation (except for spk02' /s/). Both points lowered along the consonant duration of spk03. In the present results, spk03 had no significant difference in TT or TB at all, spk02 had significant difference between the postalveolars between 0–40% of the duration, but not at all in alveolars. There was significant difference in the front of the tongue in spk01's alveolars between the 30–100% of the consonant duration, and in 0–80% of that in the postalveolars.

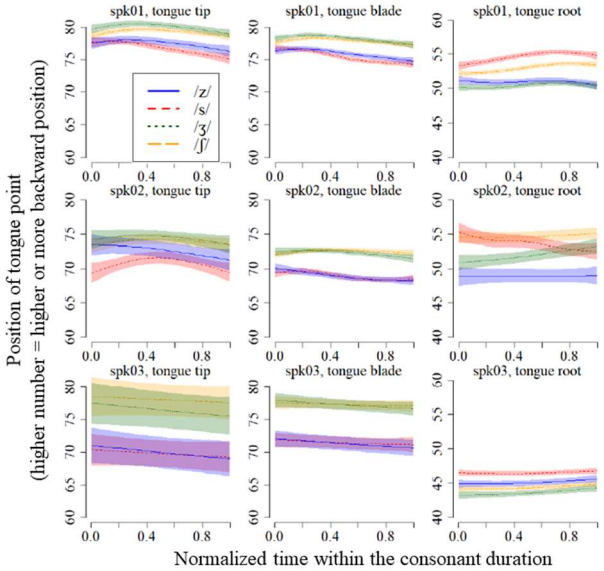


Figure 2: The tongue tip, tongue blade, and tongue root movement along the consonant duration extracted from the polarGAM. (Color online. Color and line type coding identical to Fig. 1.)

4. DISCUSSION AND CONCLUSIONS

The present study investigated dynamic articulatory and acoustic features, and voicing characteristics of four Hungarian sibilants regarding phonological voicing. Tongue root, tongue tip and tongue blade positions and movements were analysed by ultrasound tongue imaging. CoG, and spectral skewness, as well as the voicing profile and voiceless part ratio were also compared across the sibilants.

TR was advanced in voiced sibilants compared to the voiceless ones along the production to maintain phonation. This result contradicts the findings of a previous study on Hungarian /z/ and /s/, where this difference was only found for a part (most) of the speakers. The difference might appear due to the larger variability of phonetic context in the present study (four vowels were used here, while only /i/ context in [7]).

We hypothesised that the TR position was different between the alveolar and the postalveolar places of articulation due to the volume difference between the obstacle-vocal fold distances. In general, TR was found to have a more posterior position for postalveolars than for alveolars, probably as a result of POA differences. However, in some speakers' cases, phonological voicing seemed to override this trend, as voiceless alveolars exhibited a more posterior TR position than voiced alveolars. As TT and TB positions also showed considerable intraspeaker differences, these results may have an interrelation. Analysing English /s/ and /ʃ/, the groove width was found to be longer in the postalveolar sibilant [21]. Groove width difference was also found to vary, however, in a speaker-specific manner between Croatian /ʃ/ and /ʒ/ [22]. Together these findings may indicate that the target of the intense turbulent frication might be reached with somewhat speaker-specific tongue positions, which requires parallel EPG- and UTI-analysis in order to better understand the effect of the place of articulation and voicing together in its speaker-specific behaviour.

The third hypothesis was that tongue front and TR position changes along the sibilants' duration in a manner to maintain the intense turbulent frication. Intense turbulent frication might be reached by higher intraoral pressure behind the obstacle. This might be reached by i) higher tongue front, i.e., a narrower seal, or ii) (within a specific range of seal volume) with the backward movement of TR, i.e., constricting the pharyngeal region to increase the supraglottal pressure. The need of the backward movement might appear as a reason of the tongue fronts lowering found in each speakers' sibilants (transition to the following vowel), but might be also induced (as an additional reason) by the transition to the following vowel. In spk01's pronunciation, a fronting movement appeared in the last 10-20% of the sibilants. In her case, the phonation restarted sooner than in the other two speakers' pronunciation, which might be achieved by this fronting.

Based on the results of the present study, we conclude that the advanced tongue root of the voiced sibilants was found as expected compared to their voiceless counterparts. The tongue root was found to move backward during the duration of both voiced and voiceless sibilants that might appear to maintain intense turbulent frication of sibilants. Probably the relatively frequent occurrence and high ratio of devoicing found in these speakers appeared through the tongue front lowering due to the backward movement.

Due to the large interspeaker variability the analysis is planned to be extended to include a larger number of speakers. As vowel context also seems to have an effect on the data, this factor should also be further analysed in future studies. Naturally, due to the non-linear interrelation of articulation and acoustics [23], and the limited amount of information that midsagittal tongue contours give on the articulation, the spectral consequences of articulatory settings and changes can only be studied with questions necessarily being left open.

5. ACKNOWLEDGEMENTS

The work was supported by the Bolyai János Research Scholarship of the Hungarian Academy of Sciences (A.D., T.G.Cs.), ÚNKP-22-3 (K.J.), ÚNKP-22-5 (A.D., T.G.Cs.).

6. REFERENCES

[1] Stevens, KN. 1998. *Acoustic Phonetics.* MIT Press

[2] Stevens, K. N. 1997. Articulatory–acoustic–auditory relationships. In Hardcastle, W. J., Laver, J. (eds.): *The Handbook of Phonetic Sciences.* Blackwell. 462–506.

[3] Westbury, J. R. 1983 . Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *JASA* 73, 1322–1336.

[4] Ahn, S. 2018 . The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese. *J Phonetics* 71, 451–467. doi:10.1016/j.602wocn.2018.10.003.

[5] Narayanan, S. S., Alwan, A. A., Haker, K. 1995 . An articulatory study of fricative consonants using magnetic resonance imaging. *JASA* 98, 1325–1347.

[6] Coretta, S. 2020 . Longer vowel duration correlates with greater tongue root advancement at vowel offset: Acoustic and articulatory data from Italian and Polish. *JASA* 147, 245-259. https://doi.org/10.1121/10.0000556

[7] Grászi, TE., Csapó, TG., Deme, A., Juhász, K., Markó, A. 2021. Tongue root position in VC sequences with regard to the phonetic realization of obstruent voicing: A preliminary study on Hungarian. In Tiede, M. Whalen, DH., Gracco, V. (eds.), *Proc. of the 12th International Seminar on Speech Production.* New Haven CT , Amerikai Egyesült Államok : Haskins Press. 198–201

[8] Fuchs, S. Brunner, J., Busler, A. 2007 . Temporal and spatial aspects concerning the realizations of the voicing contrast in German alveolar and postalveolar fricatives. *Advances in Speech–Language Pathology* 9, 90–100.

[9] Mihajlik, P., Tüske, Z., Tarján, B., Németh, B. & Fegyó, T. 2010 . Improved recognition of spontaneous Hungarian speech: Morphological and acoustic modeling techniques for a less resourced task. *IEEE Transactions on Audio, Speech and Language Processing* 18, 1588–1600.

[10] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer.* http://www.praat.org/

[11] Shih, C., Möbius, B., Narasimhan, B. 1999. Contextual effects on consonant voicing profiles: A cross-linguistic study. Proc. 14th ICPhS. 989–992.

[12] R Core Team 2022. R: *A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.* https://www.R-project.org/.

[13] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J SS* 67, 1–48.

[14] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J SS* 82, 1–26.

[15] Bartoń, K. 2022. *MuMIn: Multi-Model Inference.* R package version 1.47.1, https://CRAN.R-project.org/package=MuMIn.

[16] Lenth, R. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.8.3, https://CRAN.R-project.org/package=emmeans.

[17] Wood, SN. 2017 *Generalized Additive Models: An Introduction with R 2nd edition* . Chapman and Hall/CRC.

[18] van Rij, J., Wieling, M., Baayen, R., van Rijn, H. 2022. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* R package version 2.4.1.

[19] Coretta S 2022 . *rticulate: Ultrasound Tongue Imaging.* R package version 1.7.3, https://CRAN.R-project.org/package=rticulate

[20] Wickham. H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

[21] Fletcher, S.G., Newman, D.G. 1991. [s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width. *JASA* 89, 850–858.

[22] Liker, M., Gibbon, FE. 2011. Groove width in croatian voiced and voiceless postalveolar fricatives. In Proc. of the XVII International Congress of Phonetics. Hong-Kong, 12–21 August, 2011.

[23] Stevens, KN. 1989. On the quantal nature of speech. *J Phonetics* 17, 3–45.

SPEECH BREATHING DURING PHYSICAL ACTIVITY IS OPPORTUNISTIC

Heather Weston

Humboldt-Universität zu Berlin, Germany
westonhe@hu-berlin.de

ABSTRACT

The ability to speak during diverse physical activities benefits individuals in many ways. What is currently unclear, however, is how individuals modify their respiratory behavior to accommodate both the slow, controlled breaths used for speech and the rapid, deeper breaths required by exercise. While the phonetics literature characterizes the execution of these two tasks as a *competition* between phonatory and respiratory goals, scholarship on neural control of respiration discusses the relationship in terms of *coordination*. To better understand speech breath support during physical activity, the current study presents empirical evidence from respiratory and acoustic recordings of read speech from 48 female speakers at rest and during exercise. Modifications of both the inspiratory and expiratory phases of respiration were observed that suggest that speech breathing during physical activity is opportunistic – linguistic structure is exploited to insert quick inspirations and expiratory airflow is increased to meet the metabolic needs of exercise.

Keywords: speech production, respiratory pattern, talking while walking, dyspnea, human performance

1. INTRODUCTION

This empirical study investigates how speech breath support is affected by simultaneous physical activity. The motivation rests on the following observation: speech and exercise modify the respiratory cycle in opposing ways. Breathing is slowed by speech to accommodate utterances but accelerated by exercise to meet increased metabolic needs. Although speaking during physical activity is part of daily life and crucial for diverse occupations, it is currently unclear *how* speakers negotiate between conflicting linguistic and exercise respiratory patterns. A better understanding of speech breathing during physical activity has both theoretical and practical relevance. Describing how speech sounds are produced under different speaker circumstances lies at the very heart of phonetics, while experimental data can provide insight into real-life speech-breathing situations that pose challenges for diverse groups of speakers, including those with certain respiratory pathologies, occupational voice users, athletes, and users of automatic speech recognition technology.

Breath-taking is commonly quantified with temporal and volumetric measures – the duration/ frequency of each breath and the volume of air inspired. A small body of work has investigated such changes in speech breathing during exercise. For temporal measures, respiratory rate (breaths/minute) has been found to significantly increase [1] by up to 25% [2], suggesting speakers accommodate speech during exercise by producing shorter chunks (i.e., pausing more). But work assessing volumetric changes suggests that this account is incomplete. Two studies [3, 4] measuring the volume of air inspired per minute – the product of breath rate and breath volume – found a significant increase during exercise. But it is unclear whether rate, volume, or both increased; based on also finding fewer syllables per phrase during exercise, [4] concluded that mainly breath rate increased, but stressed the need to pursue this hypothesis by measuring breath rate and volume separately. Work assessing volumetric measures has reported inconsistent results, finding no change in breath volume during exercise [5, 6], a decrease [7], or an increase [2]. These differences may stem from methodological choices: two studies [2, 7] did not measure breath volume directly but inferred changes based on breath noise intensity in the acoustic signal.

One commonality among this scholarship is the conceptualization of speech during exercise as a "competition" between phonatory and non-speech respiration. This is an interesting characterization of two *volitional* activities – speaking and exercising – and one that is absent in the literature on neural control of respiration, which seeks to understand how different respiratory patterns are *coordinated.* While this area is still incompletely understood and different models have been proposed [8], three strands of research add useful perspectives to investigate speech breathing during exercise. First, there is consensus that mammalian breathing comprises three phases: an inspiratory phase, a post-inspiratory phase, and an active expiratory phase incorporated to meet increased metabolic needs [9]. Second, these phases are proposed to be independent oscillators controlled by separate areas [10] of the brain's respiratory rhythm generator, the preBötzinger complex [11]. Third, the "pacemaker" neurons that dictate phase rhythms are highly flexible and can respond to

situational needs on a breath-by-breath basis [12, 13]. This account explains how breathing can be continuously adapted to meet dynamic exercise and linguistic (e.g., utterance length) goals. Related to this is evidence from respiratory physiology showing that breath rate and depth are independent of one another during exercise [14] and other types of stress [15].

Taken together, this research highlights a need for empirical evidence on the timing and depth of speech-breaths during inspiratory and expiratory phases. The present study thus extends previous phonetic work by assessing breath rate, duration, and volume in each phase to take a closer look at the "movable parts" of the breath cycle to see which are adjusted and how.

2. METHODS

2.1 Participants, materials, and equipment

A total of 48 healthy, non-smoker female native German speakers participated in the experiment. One participant was excluded due to a recording error. The average age was 23.6 years (*SD:* 3.8).

Participants read a 126-word passage designed to approximate spontaneous speech: a female German speaker was asked *What do you need to host a party?* Her monologue was transcribed, the disfluencies removed, and punctuation added. Participants were told the text stemmed from spoken speech and read it aloud once for familiarization prior to the experiment.

Exercise was performed on a low-noise stationary bicycle (daum electronic, Germany). Breathing was recorded via respiratory inductance plethysmography (Ambulatory Monitoring, USA), a non-invasive method that captures the expansion/contraction of the rib cage and abdomen as changes in raw voltage using elasticated belts with inductance wires. Thoraco-abdominal displacement is thus used as a proxy for breath volume. Speech was sampled at 22,050 Hz and recorded with a head-mounted microphone (beyerdynamic, Germany) placed 4 cm from the corner of the mouth at a 90° angle.

2.2. Experimental design, conditions, and procedure

The study used a within-participant design with three physical workload conditions: rest (sitting still), low workload, and moderate workload. Condition order was fixed (rest > low > moderate) because the cardiovascular systems may take several hours to recover after exercise [16] and responses are highly individual [17]; if condition order were randomized, it would thus be unclear if the rest condition actually captured baseline respiration values.

The physical workload conditions were defined following guidance by health organizations [e.g., 18] as a percentage of maximal heart rate (HR_{max}): low = 35% and moderate = 65%. These relative values were translated into participant-specific target heart rates (beats per minute) using a standard method in sports science, the Karvonen formula, given in (1).

$$(1) \qquad \text{target HR} = [(HR_{max} - HR_{rest}) \times \%\ \text{intensity}] + HR_{rest}$$

Following [19], HR_{max} was predicted as: 208 − (0.7 × age). HR_{rest} was estimated by having each participant recline for 10 minutes and taking the average HR of minute 11 using a wrist-worn heart rate monitor worn throughout the experiment. The use of age and resting pulse accounts for physiological factors that affect respiratory function and thus makes the physical effort required for each workload comparable across participants, regardless of their age or level of fitness.

To validate the calculated target HRs for each condition, participants were asked to rate their level of perceived exertion using the Borg scale [20]. The average rating for low workload was 9.7 ("very light") and for moderate 13.8 ("somewhat hard").

The procedure was as follows: the text passage was presented on a monitor at eye level and read aloud three times per condition. The rest condition was followed by 4-minute cycling warm-ups to reach the target HR for low and moderate workload, respectively. Heart rate was monitored in real time using a tablet readout and was maintained by adjusting resistance on the bike between trials.

2.3. Preparation and analysis of respiratory data

To smooth motion noise, the respiratory signals were downsampled from 22,050 Hz to 100 Hz and filtered (passband: 1–10 Hz) using a custom MATLAB [21] protocol. The rib cage and abdomen signals were summed using a 2:1 correction [22] to estimate total change in lung volume during speech breathing.

To obtain duration and displacement measures for each breath, inspiratory peaks and expiratory troughs in the summed respiratory signal were automatically detected using MATLAB's findpeaks function [23] and manually corrected. Inspiratory time/displacement was calculated by subtracting the timestamp/voltage of the preceding trough from a given peak. (For expiratory measures: trough − peak).

To aid interpretation of displacement measures, the raw signal in volts was converted to a physiologically functional unit, percentage of maximal displacement (≈ % of vital capacity), using an inspiratory vital capacity maneuver [24]. To assess respiratory differences between rest and exercise conditions, statistical analyses were run in R [25; Version 4.2.2]. Linear mixed-effects models were estimated using the packages lme4 [26], lmerTest [27] and emmeans [28], with condition as the independent variable (three levels: rest, low,

moderate), the outcome for each respiratory variable as the dependent variable, and random slopes and intercepts for speakers (lmer syntax: variable ~ workload + (1 + workload | speaker), data). Prior to analysis the data were log-transformed to account for upper outliers. To account for different numbers of observations (i.e., number of breaths per trial/ participant), differences between conditions were assessed using estimated marginal means (EMMs) – the means taken from the statistical model for each response variable at each level of the predictor variable. The differences between conditions were tested for statistical significance using contrast analysis to assess pairwise differences. The Kenward-Roger method was used to obtain the degrees of freedom needed to compute *p*-values.

3. RESULTS

3.1. Statistical models

EMMs for breath duration and thoracoabdominal displacement are plotted in Figure 1. The original model and the log-transformed model (retransformed to original scale) are plotted together for comparison. Inspiratory and expiratory displacement, and inspiratory time, were found to increase significantly for all workload comparisons (rest vs. low, rest vs. moderate, and low vs. moderate; *p*s < .001) in both models. Expiratory time decreased significantly from low to moderate and from rest to moderate (*ps* < .001), but no significant change was found from rest to low ($p = .151$; log model: $p = .854$). Respiratory rate showed no significant change between rest and low but increased significantly from rest to moderate and from low to moderate (*ps* < .001). Table 1 gives the group means and the log model statistics.

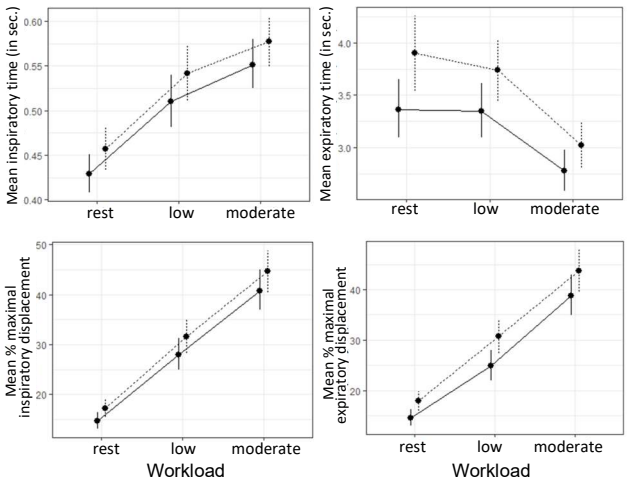


Figure 1: EMMs for log model (*solid line*) and original model (*dashed line*) for breath duration (*top*) and displacement (*bottom*) with increasing physical workload. Respiratory rate shown in Figure 2 (*inset*).

Table 1: Inspiratory/expiratory times and thoraco-abdominal displacement (in % max. displacement) across work conditions: group means with standard deviations; (log) model statistics in gray.

Var.	Cond.	Mean (*SD*)	EMM	*SE*	Conf. level
Insp. time (sec.)	rest	0.44 (0.17)	0.43	0.01	[0.41, 0.45]
	low	0.52 (0.19)	0.51	0.01	[0.48, 0.54]
	mod.	0.56 (0.17)	0.55	0.01	[0.52, 0.58]
Exp. time (sec.)	rest	3.59 (1.90)	3.36	0.14	[3.10, 3.65]
	low	3.51 (1.63)	3.35	0.13	[3.10, 3.62]
	mod.	2.86 (1.19)	2.78	0.10	[2.59, 2.98]
Insp. displ. (%)	rest	16.84 (9.97)	14.69	0.79	[13.2, 16.4]
	low	30.39 (15.25)	27.93	1.54	[25.0, 31.2]
	mod.	43.99 (17.89)	40.73	1.99	[36.9, 44.9]
Exp. displ. (%)	rest	17.48 (11.30)	14.58	0.80	[13.1, 16.3]
	low	29.50 (18.13)	24.83	1.49	[22.0, 28.3]
	mod.	43.07 (18.99)	38.72	2.01	[34.9, 43.0]
Breaths /min.	rest	13.55 (4.0)	13.5	0.55	[12.4, 14.6]
	low	13.87 (3.6)	13.9	0.50	[12.9, 14.9]
	mod.	16.74 (4.0)	16.7	0.54	[15.6, 17.8]

3.2. Further exploration

The statistical analysis showed that speakers adjust breath length, rate *and* depth during exercise, but the measures are affected differently by workload level: at low workload, resting breath rate and expiratory duration can be maintained, and only breath depth is increased; moderate workload shows changes in all measures. These findings are explored here to glean more insight into time–volume relations.

Of note in the inspiratory phase is the significant increase in breath depth with no change in breath rate until moderate workload. However, the histogram in Figure 2 and the higher *SD*s during exercise suggest that breath depth does not increase across the board but rather varies, showing upper outliers but also shallower inspirations (e.g., below mean at rest).

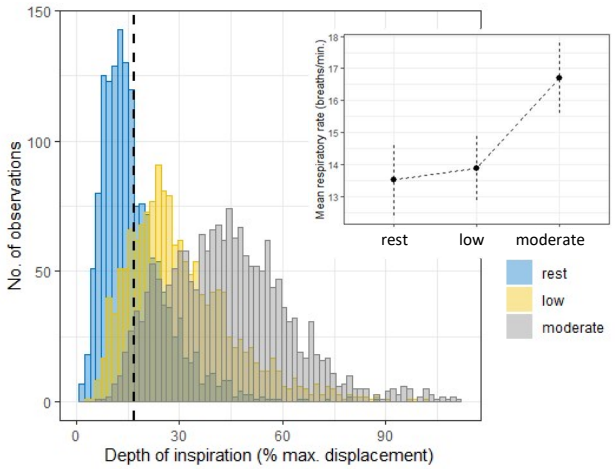


Figure 2: Histogram showing frequency of inspiratory depth; the black dashed line indicates mean depth at rest (16.8%). Inset shows EMMs for mean respiratory rate.

To test this, the non-parametric Fligner-Killeen test was used to evaluate the homogeneity of group variances. The test was significant, X^2 (2, 4860) = 459.4, $p < .001$, indicating that depth of inspiration varied more widely in some conditions than others.

Moving on to the expiratory phase, a striking observation was that speakers seemed to expel more air in a shorter time during exercise. Spearman correlations calculated for each condition showed good correlations between expiratory time and depth at rest ($R = .65$) and low workload ($R = .61$) but weak ($R = .32$) for moderate workload, suggesting that time–volume relations are not dictated by physical activity per se but rather by its intensity.

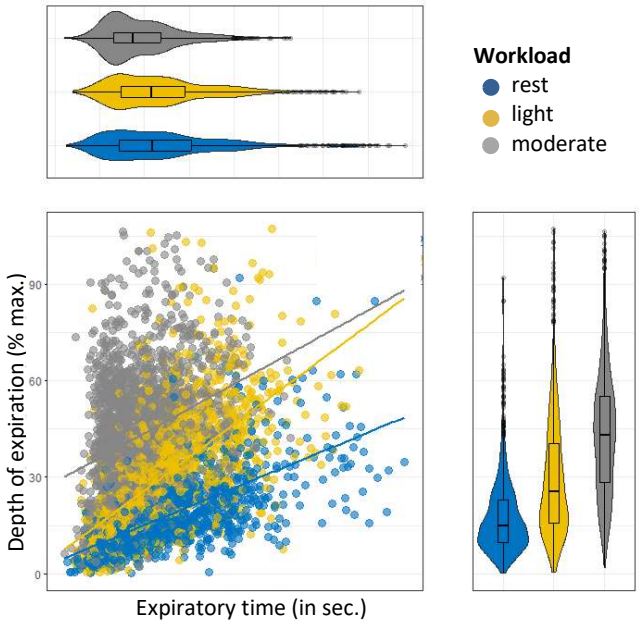


Figure 3: Scatter plot showing relationship of expiratory time to expiratory depth (% max. displacement) with regression lines; wrapped boxplots show the distribution of the expiratory time (*top*) and depth (*right*) data.

4. DISCUSSION

This study found that individuals adjust the timing *and* volume of their breaths to speak during physical activity, but the changes demonstrate more flexibility than previously though. Analyzing breath rate and depth separately revealed two independent – but interrelated – "movable parts" that can be adjusted in different ways to adapt to exercise level. At low workloads, speakers can accommodate their preferred utterance length by increasing breath depth alone. But greater workloads require a further increase in depth and an increase in rate, which may break up speech into smaller chunks. This flexibility may not have been apparent in previous studies because most assessed moderate or heavy workloads. In the present study, the low workload condition is equivalent to brisk walking and the moderate condition to jogging.

Looking at the inspiratory and expiratory phases separately also provided further insights. While inspiratory depth significantly increased with workload, it appeared that variability was also greater. It was rather surprising that small-amplitude breaths were seen at moderate workload – it suggests that breath volumes do not increase across the board. One possible explanation is that speakers become "opportunistic" and insert inspirations where possible. For example, during data labeling, some speakers took small-amplitude in-breaths at phrase breaks that had none in the rest condition (e.g., after the utterance-initial adverb "Additionally,"). Longer syntactic breaks – between sentences – were used for deeper breaths. These observations would fit with neural-control accounts of breathing rhythms being determined on a cycle-to-cycle basis [12, 13].

In the expiratory phase, it was striking that speakers expelled significantly more air compared to rest in the same amount of time (low workload) or less time (moderate workload). Again, timing and volume were adjusted separately to meet the respiratory demands for a given physical workload. But there may be different acoustic consequences for different time–volume relations, such as an increase in airflow during speech or the appearance of audible out-breaths (or both). The former may result in breathier speech, while the latter creates additional noise. Indeed, a widely obtained finding for speech during exercise is the presence of non-phonated expirations during moderate/vigorous activity in [2, 5, 7], with the latter two studies observing that expiratory volume was distributed over both phonated and non-phonated stretches of the expiratory phase. This could explain the weaker time–volume correlation observed in the present study under moderate workload. A next step is thus to further analyze the expiratory phase – most relevant for speech production – by assessing airflow changes using the three-phase breath cycle described in the work on respiratory neural control.

To come back to the question of *how* speakers negotiate between the opposing respiratory patterns required by simultaneous speech and exercise, it seems that situation-responsive, cycle-to-cycle variability is key. Breath-taking during exercise is interpreted here as being *opportunistic*: breaths may be deepened to preserve pause placement, and linguistic structure may be exploited to insert different-sized inspirations. Expiratory airflow may be regulated between phonated and non-phonated sections. These hypotheses must now be tested with further analyses, but the data presented here highlight the respiratory system's flexibility to meet different metabolic and behavioral goals. The guiding question is perhaps not *Which respiratory pattern "wins"?* but *Which time–volume relations occur and when?*

5. ACKNOWLEDGEMENTS

Parts of this work were jointly funded by the French Agence Nationale de la Recherche and the Deutsche Forschungsgemeinschaft through the SALAMMBO project. The author wishes to thank Susanne Fuchs and Amélie Rochet-Capellan for their guidance and support and Theres Weißgerber for her assistance throughout the project. The author also thanks two anonymous reviewers for their valuable suggestions.

6. REFERENCES

[1] Fuchs, S., Reichel, U. D., Rochet-Capellan, A. 2015. Changes in speech and breathing rate while speaking and biking. *Proc. 18th ICPhS Glasgow*.

[2] Trouvain, J., Truong, K. P. 2015. Prosodic characteristics of read speech before and after treadmill running. *Proc. Interspeech 2015 Dresden*, 3700-3704.

[3] Baker, S. E., Hipp, J., Alessio, H. 2008. Ventilation and speech characteristics during submaximal aerobic exercise. *J. Speech Lang. Hear. Res.*, 51(5), 1203–1214.

[4] Ziegler, A., Van Swearingen, J., Jakicic, J. M., Verdolini Abbott, K. 2019. Phonation demonstrates goal dependence under unique vocal intensity and aerobic workload conditions. *J. Speech Lang. Hear. Res.*, 62(8), 2584–2600.

[5] Doust, J. H., Patrick, J. M. 1981. The limitation of exercise ventilation during speech. *Resp. Physiol.*, 46(2), 137–147.

[6] Meckel, Y., Rotstein, A., Inbar, O. 2002. The effects of speech production on physiologic responses during submaximal exercise. *Med. Sci. Sports Exerc.*, 34(8), 1337–1343.

[7] Patil, S., Sangwan, A., Hansen, J. H. L. 2010. Speech under physical stress: A production-based framework. *Proc. 2010 IEEE Cape Town*, 5146–5149.

[8] Del Negro, C. A., Funk, G. D., Feldman, J. L. 2018 Breathing matters. *Nat. Rev. Neurosci.* 19, 351–367.

[9] Richter D. W. & Smith J. C. 2014 Respiratory rhythm generation in vivo. *Physiology* 29(58–71).

[10] Anderson, T. M., Ramirez, J. M. 2017. Respiratory rhythm generation: triple oscillator hypothesis. *F1000Res.*, 6, 139.

[11] Smith, J. C., Ellenberger, H. H., Ballanyi, K., Richter, D. W., Feldman, J. L.(1991. Pre-Bötzinger complex: a brainstem region that may generate respiratory rhythm in mammals. *Science*, 254(5032), 726–729.

[12] Carroll, M. S., & Ramirez, J. M. (2013). Cycle-by-cycle assembly of respiratory network activity is dynamic and stochastic. *J. Neurophysiol.*, 109(2), 296–305.

[13] Ramirez, J. M., & Baertsch, N. A. (2018). The dynamic basis of respiratory rhythm generation: one breath at a time. *Ann. Rev. Neurosci.*, 41, 475–499.

[14] Nicolò, A., Girardi, M., Bazzucchi, I., Felici, F., Sacchetti, M. 2018. Respiratory frequency and tidal volume during exercise: differential control and unbalanced interdependence. *Physiol. Rep.*, 6(21), e13908.

[15] Tipton, M. J., Harper, A., Paton, J. F. R., Costello, J. T. 2017. The human ventilatory response to stress: rate or depth?. *J. Physiol.*, 595(17), 5729–5752.

[16] Romero, S. A., Minson, C. T., Halliwill, J. R. 2017. The cardiovascular system after exercise. *J. Appl. Physiol. (1985)*, 122(4), 925–932.

[17] Travers, G., Kippelen, P., Trangmar, S. J., González-Alonso, J. 2022. Physiological function during exercise and environmental stress in humans—An integrative view of body systems and homeostasis. *Cells*, 11(3), 383.

[18] Physical Activity Guidelines Advisory Committee. 2008. *Physical Activity Guidelines Advisory Committee Report, 2008*. U.S. Department of Health and Human Services, p. D-3. Retrieved from https://health.gov/sites/default/files/2019-10/CommitteeReport_7.pdf, 10 Aug. 2020.

[19] Tanaka, H., K. D. Monahan, K. D., Seals, D. R. 2001. Age-predicted maximal heart rate revisited *J. Am. Coll. Cardiol.,* 37(1), 153–156.

[20] Borg, G. A. 1982. Psychophysical bases of perceived exertion. *Med. Sci. Sports Exerc.*, 14(5), 377–381.

[21] MATLAB. 2010. version 7.10.0 (R2017b). Natick, Massachusetts: The MathWorks Inc.

[22] Banzett, R. B., Mahan, S. T., Garner, D. M., Brughera, A., Loring, S. H. 1995. A simple and reliable method to calibrate respiratory magnetometers and Respitrace. *J. Appl. Physiol.,* 79(6), 2169–2176.

[23] findpeaks. (2007-2016). Natick, Massachusetts: The MathWorks Inc. https://de.mathworks.com/help/signal/ref/findpeaks.html.

[24] Wanger, J., Clausen, J. L., Coates, A., Pedersen, O. F., Brusasco, V., Burgos, F., Casaburi, R., Crapo, R., Enright, P., van der Grinten, C. P. M., Gustafsson, P., Hankinson, J., Jensen, R., Johnson, D., MacIntyre, N., McKay, R., Miller, M. R., Navajas, D., Pellegrino, R., Viegi, G. 2005. Standardisation of the measurement of lung volumes. Eur. Respir. J. 26 (3), 511–522.

[25] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[26] Bates, D., Maechler, M., Bolker, B., Walker, S. 2022. lme4: Linear Mixed-Effects Models Using Eigen and S4. https://github.com/lme4/lme4/.

[27] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.,* 82(13), 1–26.

[28] Lenth, R.V., Love, J., Hervé, M. 2022. Package emmeans: Estimated marginal means, aka least-squares means. https://github.com/rvlenth/emmeans.

ARTICULATORY AND ACOUSTIC ANALYSIS OF COARTICULATED /u/ AND /y/ PRODUCED BY FEMALE AND MALE SPEAKERS OF GERMAN

Lia Saki Bučar Shigemori, Rosa Franzke, Philip Hoole

Institute for Phonetics and Speech Processing (IPS), LMU Munich
lia | rosa.franzke | hoole@phonetik.uni-muenchen.de

ABSTRACT

Acoustic vowel spaces of female and male speakers differ in a non-uniform way, but less is known about whether this difference can be extended to coarticulatory effects. Eleven female and eight male German speakers produced single words containing tense /u/ or /y/ in bilabial, alveolar and velar contexts. Principal component analysis was used to analyse entire ultrasound image frames. For the acoustic analysis F2-values were extracted. At the acoustically determined vowel midpoint a significant fronting effect in alveolar context could be observed for both female and male speakers as a raised F2 in /u/; however in the articulatory data the effect was significant only for females. In /y/, F2 values differed significantly between bilabial and velar contexts for female speakers, while the articulatory data showed no significant effects for consonantal context. The data will be discussed from a sound change perspective and regarding articulatory-acoustic relationships.

Keywords: coarticulation, sex differences, acoustic-articulatory relations, ultrasound, vowels

1. INTRODUCTION

The aim of this study is to compare consonant to vowel coarticulatory effects between female and male speakers for phonological back and front vowels. Vowel spaces of female and male speakers differ in a non-uniform way. In phonological back vowels like /u/, F2 of females and males is similar, but in phonological front vowels like /y/, F2 of females is higher than for males (e.g. Figure 1 in [1]). While this non-uniformity can partially be attributed to physiological differences, different behavioural or sociophonetic explanations (e.g. different acoustic targets [2], or compensation for poorer harmonic sampling of spectral envelopes with higher F0 [3]) have been proposed in addition.

Several studies also reported systematic differences between females and males in speaking style, linking slower speech rate [4], more distinct vowel spaces [5] and greater spectro-temporal variation and consequently a lower degree of coarticulation with female speakers compared to male speakers [6]. It has also been suggested that female and male speakers apply different articulatory strategies due to physiological differences of their oral structures [7]. Although it has been shown that the degree of coarticulation differs for different vowels [8], to our knowledge, no study has compared the degree of coarticulation in different vowels separately for female and male speakers.

In this study, we focus on coarticulated /u/ and /y/ in German, for which it has been shown that greater coarticulatory effects in both articulation and resulting acoustics can be found in /u/ than in /y/ based on data by one female and six male speakers [9]. The results are presented as evidence for the universal preference for /u/ to diachronically change to /y/, rather than /y/ to become /u/. This so called /u/-fronting sound change has also been claimed to be led by females, although the suggested reasons are of sociolinguistic nature [10]. The literature does not link the leadership of sound change and one of its main sources, namely the misinterpretation of speaker-dependent differences in coarticulation. By looking at German, we want to explore whether differences in coarticulatory effect along the back-front dimension can be found between female and male speakers, and whether differences in the acoustic-articulatory relationship can be observed.

Given that a lower degree of coarticulation is reported for female compared to male speakers, it might be counterintuitive to suggest that we would find greater coarticulatory effects for females in our data. However, previous studies mainly looked at F2 locus equation slopes and linked a steeper slope with greater distinction between the consonant and vowel and consequently a lower degree of coarticulation. In the current study, we will instead look at data extracted at vowel midpoints. It is possible that the greater acoustic vowel space of female speakers allows them to convey consonantal context information in their vowel production while still keeping the vowel categories distinct, in contrast

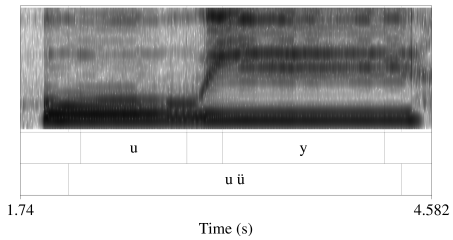


Figure 1: Example of an /u/ to /y/ continuum production by a female speaker. The spectrogram shows a Frequency range from 0 to 4000 Hz.

to male speakers who might be more likely to hypoarticulate.

2. METHODS

Eleven female and eight male German speakers were recorded in a soundproof booth. Simultaneous recordings of acoustic and articulatory data by means of ultrasound were carried out using the AAA-software [11]. The ultrasound probe was attached below the chin using the stabilizer developed by [12], which allows flexible jaw movement. A 5-8MHz, 10mm radius or a 2-4MHz 20mm radius microconvex probe were used depending on the speaker anatomy with adjusted recording settings to ensure good quality imaging but painless sessions. In most cases the frame rate was 81.6Hz. Simultaneous video recordings have been made which can be used to track the movement of the probe in relation to the head, but they have not been evaluated yet. Participants were instructed to read single words presented on a screen one by one. The target words analyzed in this study are shown in Table 1. They consist of nouns with tense vowels /u/ or /y/ occupying the stressed, initial syllable, preceded and followed by either alveolar, bilabial or velar consonants. Four repetitions were recorded per participant. In addition, participants were instructed to produce continua by moving their tongue from /u/ to /y/ or /y/ to /u/ (Example in Figure 1).

	alveolar	bilabial	velar
/u/	T**u**te	B**u**be	K**u**gel K**u**chen
/y/	T**ü**te	B**ü**bchen	K**ü**ken K**ü**gelchen

Table 1: Targetwords used in this study, column-wise for the three consonantal contexts. The target vowels /u/ or /y/ are indicated in bold.

2.1. Data preparation

Acoustic recordings were segmented using WebMAUS [13]. Formant data were extracted using LPC analysis in PRAAT by applying the Burg method with 25 ms Window length. The acoustic data were transformed into an emuDB [14] and segment boundaries and formant data were manually corrected if necessary.

To analyze the ultrasound data principal component analysis (PCA) on entire image frames was carried out using the `prcomp`-function in R [15], similarly to the method described in [16]. Our main interest was the movement of the tongue in the /u/-/y/ dimension. Therefore, the vowel continua served as the reference data, based on which the PC dimensions were calculated separately for each speaker. Ultrasound frames at the acoustically determined midpoint were then extracted and the PC scores were calculated. Upon visual inspection, for each speaker the PC dimension with the lowest number was selected which best separated /u/ and /y/ in word production and reflected a movement in opposite directions for /u/-/y/ continua compared to /y/-/u/ continua.[1] For each speaker the PC values were then scaled from -1 to 1 with reference to the lowest and highest values of the continua productions and of the target vowels, and if required flipped, so that the lower values corresponded to more /u/-like and the higher values to more /y/-like tongue configurations.

2.2. Statistics

To test the effect of vowel, consonantal context and speaker sex on articulation (scaled PC value) and acoustics (F2), linear mixed effects analyses were performed using lme4 [17] in R. CONSONANTAL CONTEXT (three levels: alveolar, bilabial, velar), VOWEL (two levels: /u/, /y/) and SPEAKER SEX (two levels: female, male) were fixed factors, and SPEAKER was a random factor (with by-SPEAKER random intercept for CONSONANTAL CONTEXT and VOWEL). To simplify the analysis the data was further split up by vowel, especially because the consonantal context is expected to have a different effect on front and back vowels. This allowed us to still compare the effect of sex and consonantal context and their interaction, but separately for /u/ and /y/. The significance of each effect was retrieved from the ANOVA type III table with p-values calculated using Satterthwaite's method. Effects were considered to be significant when $p < 0.05$. Post-hoc Tukey-tests were carried out using the emmeans-package [18].

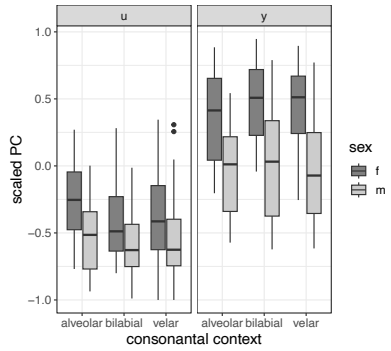


Figure 2: Scaled PC values of ultrasound images at the acoustically determined vowel midpoint, separately for consonantal contexts (alveolar, bilabial and velar), grouped by speaker sex (female in dark grey and male in light grey), left panel for /u/ and right panel for /y/.

3. RESULTS

3.1. Articulatory data

The scaled PC values of the ultrasound frames extracted at the acoustically determined vowel midpoint are illustrated in Figure 2 for the consonantal contexts (alveolar, bilabial, velar) and speaker sex, in the left panel for vowel /u/ and in the right panel for vowel /y/. Keep in mind that the speaker normalization of the PC values was carried out based on the ultrasound frames from the continua productions and the frames extracted at target vowel midpoints. Negative values are more /u/ like and positive values more /y/-like along the /u/-/y/ dimension. Figure 2 confirms this. In addition, /u/ in alveolar context has a slightly higher PC value than for bilabial or velar contexts. It seems that in word productions the difference between /u/ and /y/ was less extreme than when speakers were producing vowel continua, especially for /y/, where the value 1 was not reached in any context, neither by female nor by male speakers.

The full model revealed that the main effects of SPEAKER SEX ($F[1,19.0]=9.4$, $p<0.01$) and VOWEL ($F[1,19.3]=147.0$, $p<0.001$) were significant but the effect of CONSONANTAL CONTEXT ($p=0.8$) was not. In addition, the interactions between SPEAKER SEX and VOWEL ($F[1,19.3]=4.5$, $p<0.05$) and between VOWEL and CONSONANTAL CONTEXT ($F[2, 539.8]=10.2$, $p<0.001$) were significant.

3.1.1. Articulatory data of /u/

The linear mixed effect model for the back vowel /u/ revealed that the main effect of CONSONANTAL CONTEXT was significant ($F[2,19.0]=6.4$, $p<0.01$). There was only a trend effect for SPEAKER SEX ($F[1,19.0]=4.3$, $p=0.05$) and the interaction between CONSONANTAL CONTEXT and SPEAKER SEX was not significant. The post-hoc test revealed that, for females, the difference between /u/ in alveolar and bilabial contexts was significant ($p<0.01$), while the difference between alveolar and velar contexts ($p=0.09$) and between bilabial and velar contexts ($p=0.5$) were not. For male speakers, none of the differences were significant ($p>0.5$). In alveolar context, scaled PC values for females were significantly higher than for males ($p<0.05$), while the scaled PC values for females and males did not differ significantly between bilabial and velar contexts ($p=0.1$).

3.1.2. Articulatory data for /y/

For the front vowel /y/, the model revealed that the effect of SPEAKER SEX on the scaled PC values was significant ($F[1,19.0]=10.8$, $p<0.01$). The effect of CONSONANTAL CONTEXT ($F[2,19.3]=2.8$, $p=0.09$) and the interaction between CONSONANTAL CONTEXT and SPEAKER SEX ($F[2,19.3]=0.09$, $p=0.9$) were not significant. The post-hoc test confirmed that the difference between female and male speakers was significant for all consonantal contexts, while no paired comparisons between consonantal contexts were significant, neither for females nor for males.

3.2. Acoustic data

F2 values extracted at vowel midpoints are illustrated in Figure 3, for the consonantal contexts (alveolar, bilabial, velar) and speaker sex, in the left panel for vowel /u/ and in the right panel for vowel /y/. As expected, F2 values are lower for /u/ than for /y/, and a difference between female and male speakers can be seen for /y/ but not for /u/. In /u/, F2 is slightly higher for the alveolar context than for the bilabial and velar contexts, suggesting a fronting effect due to coarticulaton.

The full model revealed that the effects of VOWEL ($F[1,576.4]=8731.6$, $p<0.001$), CONSONANTAL CONTEXT ($F[2,32.5]=11.2$, $p<0.001$) and SPEAKER SEX ($F[1,19.0]=14.7$, $p<0.05$) were significant. In addition, the interactions between VOWEL and CONSONANTAL CONTEXT ($F[2,576.4]=15.7$, $p<0.001$) and between VOWEL and SPEAKER SEX ($F[1,576.43]=170.8$, $p<0.001$) were significant.

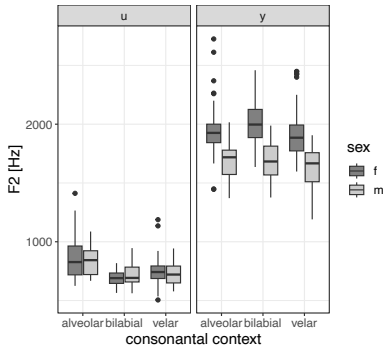


Figure 3: F2 values in Hz at the acoustically determined vowel midpoint, separately for consonantal contexts (alveolar, bilabial and velar), grouped by speaker sex (female in dark grey and male in light grey), left panel for /u/ and right panel for /y/.

3.2.1. Acoustic data for /u/

The model for the F2 values of the vowel /u/ revealed that the main effect of CONSONANTAL CONTEXT ($F[2,29.7]=18.3$, $p<0.001$) was significant, but the effect of SPEAKER SEX ($F[1,19.1]=0.01$, $p=0.9$) and the interaction effect between CONSONANTAL CONTEXT and SPEAKER SEX ($F[2,29.7]=2.3$, $p=0.1$) were not. The post-hoc test revealed that for female speakers, the differences between alveolar and bilabial context ($p<0.001$), alveolar and velar context ($p<0.05$) and bilabial and velar context ($p<0.05$) were all significant, while for male speakers, only the difference between alveolar and velar context ($p<0.05$) was significant, and the differences between alveolar and bilabial context ($p=0.06$) and between bilabial and velar context($p=1$) were not. F2 differences between female and male speakers were not significant regardless of consonantal context.

3.2.2. Acoustic data for /y/

For the F2 values of the vowel /y/, the model revealed that, the main effects CONSONANTAL CONTEXT ($F[2,18.8]=6.9$, $p<0.01$) and SPEAKER SEX ($F[1,19.0]=18.4$, $p<0.001$) were significant, but the interaction between CONSONANTAL CONTEXT and SPEAKER SEX ($F[2,18.8]=0.9$, $p=0.4$) was not. The post-hoc test revealed that F2 values differed significantly only between bilabial and velar contexts for female speakers ($p<0.05$).

4. SUMMARY AND DISCUSSION

The aim of this study was to explore differences in coarticulation between female and male speakers, which might also depend on segmental factors and which might reveal differences in acoustic-articulatory relations between female and male speakers. Although the interaction between CONSONANTAL CONTEXT and SPEAKER SEX was not significant in any of the models, a look at the post-hoc tests revealed that if the main effect of CONSONANTAL CONTEXT was significant, the differences between consonantal contexts were greater for female speakers than for male speakers. The results also corroborate findings from previous literature [9] that the coarticulatory effect in the front-back dimension is greater for /u/ than for /y/.

The articulatory data suggests that male speakers produce less fronted /y/ in words compared to when they are instructed to produce continua. This can be interpreted as evidence that indeed male speakers hypoarticulate more compared to females, in the sense that their articulation is more centralized. In contrast, female speakers produce more distinct vowels but at the same time the within category variation due to consonantal context is also clearer.

A visual inspection of the patterns across speakers revealed that the very low scaled PC values in /y/ for the articulatory data can be attributed to four of the eight male speakers. However, when we compared the F2 values in word productions with F2 values in continua productions separately for each speaker, no such differences could be found. Thus, while some speakers articulated the /u/-/y/ contrast differently in continua than in words, this was not reflected in their acoustic output. One possibility is that for these speakers variation other than only in the front-back dimension was captured in the PC, such as jaw opening or displacement of the probe due to different posture during the continua production. A second possibility is that in continua productions these speakers produced /y/ with a more fronted tongue, which however did not affect F2. In their attempt to reproduce Fant's nomograms with human speakers, [19] report a flattening of the F2 curve for anterior constriction locations for high vowels. They suggest that in the region of high front vowels the tongue body flattens with more anterior constriction location and the length of the back cavity, which F2 is associated with, does not increase.

In this study we showed that gender, vowel type and consonantal context all contribute to acoustic and articulatory variation differently or to a different degree, providing bias which can potentially lead to sound change. To better understand mechanisms of sound change, looking into fine-grained contexts and bringing together what has been shown in previous studies as we tried to do here seems fruitful.

5. ACKNOWLEDGEMENTS

The research was funded by the project SoundAct which has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101053194). We thank Anna Ratzinger for her help with the post-processing, Jessica Siddins for proofreading, and colleagues from the IPS for various inputs at different stages.

6. REFERENCES

[1] A. Simpson and C. Ericsdotter, "Sex-specific differences in f0 and vowel space," in *16th International Congress of Phonetic Sciences*, 2007, pp. 933–936.

[2] G. Fant, "Non-uniform vowel normalization," *STL-QPSR*, vol. 16, no. 2–3, pp. 001–019, 1975.

[3] R. L. Diehl, B. Lindblom, K. A. Hoemeke, and R. P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," *Journal of Phonetics*, vol. 24, pp. 187–208, 1996.

[4] D. Byrd, "Preliminary results on speaker-dependent variation in the TIMIT database," *The Journal of the Acoustical Society of America*, vol. 92, no. 1, pp. 593–596, 1992.

[5] H. Traunmüller, "Paralinguistic variation and invariance in the characteristic frequencies of vowels," *Phonetica*, vol. 45, no. 1, pp. 1–29, 1988.

[6] F. Herrmann, S. P. Cunningham, and S. P. Whiteside, "Speaker sex effects on temporal and spectro-temporal measures of speech," *Journal of the International Phonetic Association*, vol. 44, no. 1, pp. 59–74, 2014.

[7] A. P. Simpson, "Gender-specific articulatory–acoustic relations in vowel sequences," *Journal of Phonetics*, vol. 30, no. 3, pp. 417–435, 2002.

[8] H. van den Heuvel, B. Cranen, and T. Rietveld, "Speaker variability in the coarticulation of /a, i, u/," *Speech communication*, vol. 18, no. 2, pp. 113–130, 1996.

[9] J. Harrington, P. Hoole, F. Kleber, and U. Reubold, "The physiological, acoustic, and perceptual basis of high back vowel fronting: Evidence from German tense and lax vowels," *Journal of Phonetics*, vol. 39, no. 2, pp. 121–131, 2011.

[10] W. Labov, "The intersection of sex and social class in the course of linguistic change," *Language variation and change*, vol. 2, no. 2, pp. 205–254, 1990.

[11] Articulate Instruments Ltd, "Articulate Assistant Advanced user guide: Version 2.14," Edinburgh, UK, 2012.

[12] D. Derrick, C. Carignan, W.-R. Chen, M. Shujau, and C. Best, "Three-dimensional printable ultrasound transducer stabilization system," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, pp. EL392–EL398, 2018.

[13] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017.

[14] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, Sep. 2017.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[16] P. Hoole and M. Pouplier, "Öhman returns: New horizons in the collection and analysis of imaging data in speech production research," *Computer Speech & Language*, vol. 45, pp. 253–277, 2017.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022, R package version 1.8.3. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[19] P. Ladefoged and A. Bladon, "Attempts by human speakers to reproduce Fant's nomograms," *Speech Communication*, vol. 1, no. 3-4, pp. 185–198, 1982.

[1] Sometimes two PC dimensions together seemed to best capture the contrast between /u/ and /y/, however, picking only one PC dimension simplified following speaker normalization and the PC dimension with the lower number explains the greater variance in the data.

THE L~N MERGER IN SOUTHWESTERN MANDARIN: AN ARTICULATORY STUDY

Jing Huang, Kye Shibata, Feng-fan Hsieh, Yueh-chin Chang, Mark Tiede
National Tsing Hua University, Haskins Laboratories
xiaokuidaren94@163.com, kye.shibata@gmail.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw, tiede@haskins.yale.edu

ABSTRACT

This work is an articulatory study of the l~n merger in Southwestern Mandarin. The material includes canonical /l/ and /n/ produced in word-initial position. Results from four speakers show systematically different patterns of Tongue Tip and Body EMA sensors at target for these sounds, even though no consistent difference is found in their formant values. Corresponding coronal ultrasound imaging of /l/ production exhibits both concave and convex tongue shapes within and across speakers. Comparisons of the "lateral angles" between the fitted midsagittal spline and the parasagittal EMA sensor show no significant difference between /l/ and /n/. Taken together, our results suggest that the articulatory reflex of the l~n merger could be due to ongoing loss of posture differentiation at the tongue blade. Additional language-specific features of /l/ production will also be discussed.

Keywords: EMA, ultrasound, lateral, nasal, Southwestern Mandarin

1. INTRODUCTION

The lateral approximant involves the coordination of multiple gestures. In different varieties of English, /l/ production has been reported to involve the following characteristics: tongue tip raising, tongue middle lowering, tongue dorsum retraction, para-sagittal tongue lowering, and jaw-lowering (see [20] for a recent survey). [20]'s EMA results indicate that the lateral channel is formed by tilting the tongue to the left/right side of the oral cavity in Australian English (cf. [4]), that is, asymmetrical lowering of the sides of the tongue.

The lateral approximant is different in Mandarin, at least phonologically speaking. Firstly, /l/ is not possible in coda position in Mandarin. The well-known light /l/ vs. dark /l/ distinction in English (e.g., [5], [11]) is absent in Mandarin. Secondly and more interestingly, it has long been noted that the so-called l~n merger exists in many Southern Sinitic languages, especially in those languages spoken along the Yangtze River, be they Mandarin or non-Mandarin ([8], [3], [10]). To take a famous example, the l~n merger has been completed in Hong Kong Cantonese ([15]). The alveolar nasal onset is merged with the lateral approximant across the broad, resulting in a gap of the alveolar nasal onset. To this end, it is fair to say that the phonology of the lateral is different in Sinitic languages if compared with the more well-known cases in English. The first research question of this study, then, is to explore if there is any language-specific feature of /l/ production, given that the lateral approximant has distinct phonological patterning. The second research question is to explore if there is any phonetic underpinning for the l~n merger. Southwestern Mandarin (SWM) was chosen as the target language to address these research questions. This is primarily because the l~n merger remains a sound change in progress ([17], [21], [22]) in SWM. It is also beneficial if (gain/loss of) lateralization can be investigated through the lens of a phonetically gradient sound change such as the l~n merger.

2. EXPERIMENT

Seven native SWM speakers in their twenties participated in this study. All of them were born and raised in western Hubei, speaking SWM as their dominant language, and have no reported history of speech and/or hearing disorders. The data from four of them (M01~M04) are reported.

The recording materials are comprised of 30 disyllabic words. Here we analyzed and reported six of them, namely, /la.la/, /li.li/, /lu.lu/, /na.na/, /ni.ni/, and /nu.nu/. These are (possible) nicknames. Note also that the stimuli are based on prescriptive transcription. The participants were asked to read a randomized list of the target words from a computer screen at a normal speech rate in a sound-proof recording booth at National Tsing Hua University. The target words were embedded in the carrier phrase, "*p*^{h}*e* A, *pu p*^{h}*e* B", meaning "(I) pat A; do not pat B." Seven repetitions were collected for each token, and the target words in A were analyzed (as they are under focus). We used Carstens's AG501 and a Micro system from Articulate Instruments Ltd. to collect articulatory data. Acoustic data were recorded simultaneously with a sampling rate of 24 kHz. The EMA data were down-sampled to 250 Hz. The ultrasound data were collected using a transducer with a 92° field of view (FOV), set at a depth of 120mm. The frame rate was 65 f.p.s.

Fig. 1 shows the layout of the EMA sensors. In addition to the Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD) sensors along the mid-sagittal line, a parasagittal (PL) sensor is attached to the tongue blade on the speaker's lefthand side.

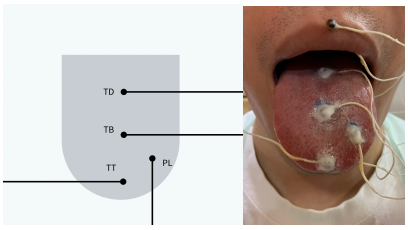


Figure 1: The layout of the EMA sensors

The EMA data were analyzed with the help of Mview [13] and ultrasound images were annotated and analyzed by the AAA software and GetContours [14]. Praat [2] was used for acoustical analysis.

3. RESULTS

The data were first transcribed by a trained phonetician. The "error rate" for /l/ is 13%, while the "error rate" for /n/ is 39%, whereby "error" here means a deviance from the prescriptive pronunciation (e.g., /li/ was transcribed as [ni]). That is to say, the target for the merger is /l/, at least in this study.

3.1. The lateral approximant in SWM

3.1.1. Visualization of the EMA data

In order to characterize the tongue posture in the coronal view, a "pseudo" (virtual) sensor, vTB, on the mid-sagittal surface of the tongue was calculated (cf. [20]). Its position is estimated by taking the point along the fitted spline connecting the mid-sagittal sensors with the shortest Euclidean distance to the PL sensor in the x- and z-dimensions. Fig. 2 illustrates the temporal changes of /l/ in both mid-sagittal and coronal planes. It seems that the TT raising is the most robust lingual movement in /l/ production. We also found variations in the coronal tongue shape: the PL could be raised, lowered, or nearly flat when compared to the vTB position, as shown in Fig. 2.

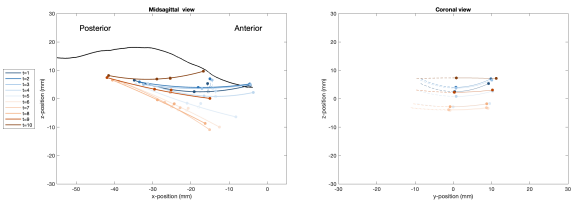


Figure 2: Temporal changes of lingual configurations of [la] of M01 (The speaker is facing right; blue line = starting point and red line = endpoint; midsagittal view on the left and the coronal view on the right side).

3.1.2. Components in /l/ production

It has been claimed in [11] that /l/-production involves a consonant-like coronal component and an intrinsic vowel-like dorsal component (i.e., TT raising and TD retraction) in English. Our results suggest that SWM patterns alike in this regard. We can see in Fig. 3 that the lateral approximant has two gestures: tongue tip raising (TTz) and tongue dorsum fronting (TDx). No significant vertical movement of the TD sensor is observed (TDz), however. The illustrations for /lu/ and /li/were omitted due to space limitations, but the observation still holds.

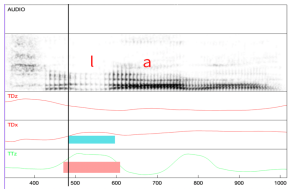


Figure 3: Temporal layout of the horizontal movement of Tongue Dorsum (TDx) sensor and vertical movement of Tongue Tip (TTz) sensor for the word-initial /l/.

3.1.3. Ultrasound results of /l/ in the coronal plane

Following [5] and subsequent work, we present the coronal tongue shapes of /l/ production. We can see from Fig. 4 that there is inter-speaker variation. Both grooved and domed coronal tongue shapes are found (M01 and M02, respectively), all else being equal. The image in Fig. 4 seems to show tilting to the speaker-right side (although it's tilted left in the image due to our POV) of the oral cavity, probably replicating the central claim made in [20], according to which the lateral channel is formed by tilting the tongue to the left/right side. The ultrasound images for /lu/ and /li/were omitted due to space limitations, but the observation still holds. It is equally worth noting that our speakers consistently produced a specific coronal tongue shape (concave or convex) across the tokens in the experiment.

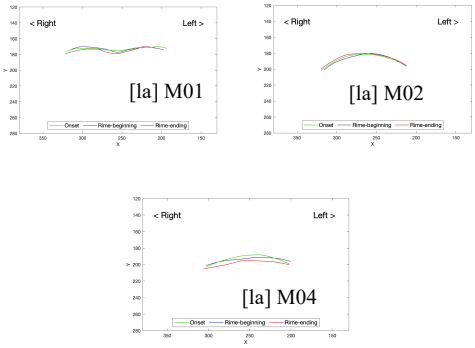


Figure 4: Ultrasound lingual configuration in coronal view of the onset in [la] for M01, M02, and M04.

3.2. The l~n merger

We turn now to the l~n merger. Note again that our results of impressionistic transcription suggest that most tokens were rendered like a lateral approximant. Nevertheless, this study classified the /l/ and /n/ groups according to prescriptive pronunciations.

3.2.1. The lateralization angle

We quantify the degree of lateralization by taking the "lateralization angle" of the tongue. This lateralization angle, θ_L, is defined as

$$(1)\ \theta_L = \tan^{-1}\left(\frac{\alpha\sqrt{d_x^2 + d_z^2}}{d_y}\right)$$

where *dx*, *dy*, and *dz* are the differences in *x*-, *y*-, and *z*-dimensional positions of the vTB and PL, respectively, and α is a coefficient which equals 1 when the PL is below the mid-sagittal spline, and -1 when above, leading to a positive angle for a lowered PL, and a negative angle for a raised PL. This is visualized in Fig. 5. Note again that θ_L is based the "vTB-PL" plane, represented as the shaded plane in the 3D view.

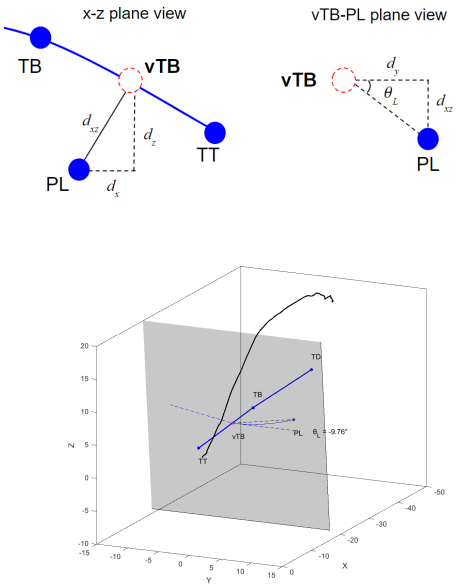


Figure 5: The visualization of the lateralization angle.

Our results show no apparent differences in θ_L between the /l/ and /n/ groups, with both groups having significant inter- and intra-speaker variation (Fig. 6). Both groups had positive lateralization, negative lateralization (or, "anti-lateralization"), and nearly flat configurations.

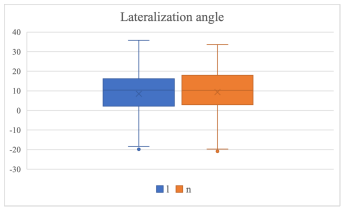


Figure 6: Lateralization angles of /l/ vs. /n/

3.2.2. GAMM results of the EMA data

The Generalized Additive Mixed Models (GAMMs) analyses for the trajectories of the EMA sensors are primarily based on the procedures and suggestions provided in [9] and [18]. The head-corrected data were *z*-transformed and subsequently fed into GAMM models. The models included a by-word smooth function through time to investigate articulatory changes over time and a random smooth to account for variations between all the speakers.

In Fig.7, the Tongue Tip (TT) sensor trajectories in the /l/ vs. /n/ groups were compared. Note that an *x*-axis in red indicates a statistically significant difference. As shown, the TTx, but not the TTz, trajectories are significantly different between the /l/ and /n/ groups.

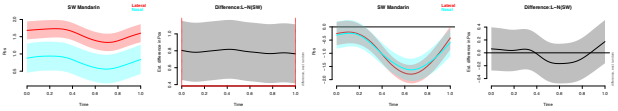


Figure 7: GAMM results of the EMA sensor trajectories of TT (TTx on the left side; TTz on the right side) of the /l/ vs. /n/ groups

The overall results for the three pairs are summarized in Table 1. The trajectories of TT, TB, TD and PL of these pairs are significantly different along either the horizontal (*x*) or the vertical (*z*) dimensions throughout the entire syllable.

Table 1: Prescriptive /l/ vs. /n/: GAMM results.

L vs. N	TT	TB	TD	PL
/la/ vs. /na/	TTx	TBx	TDx	PLx/z
/li/ vs. /ni/	TTz	TBx	n.s.	PLx/z
/lu/ vs. /nu/	TTx/z	TBx/z	TDx/z	PLz

3.2.3. GAMM results of the ultrasound data

Regarding the ultrasound data, the polar coordinates of the tongue contours (comprised of 42 points) were exported from the AAA software. The tongue contours (in the midpoints of the /l/'s or /n/'s) were fed into GAMMs using the *bam* function in *R* from the package *mgcv* [19] for the statistical significance analysis, with the help of [7]'s script adapted to our data by us.

Fig. 8 illustrates the fitted splines of all the tokens of /la/ vs. /na/, /li/ vs. /ni/, and /lu/ vs. /nu/ in the midpoints of the onsets across all the four speakers. The regions of significant difference, indicated by the shaded areas, were based on the outputs of the *plot_diff* function of the *itsadug* package. The GAMM results indicate that the tongue postures are not different in most parts of the tongue; the pair /li/ vs. /ni/ differ only in approximately 50% of the entire fitted splines. To this end, we may say that /l/ and /n/ are not distinguishable in terms of the tongue configuration in the midpoints of the consonantal part.

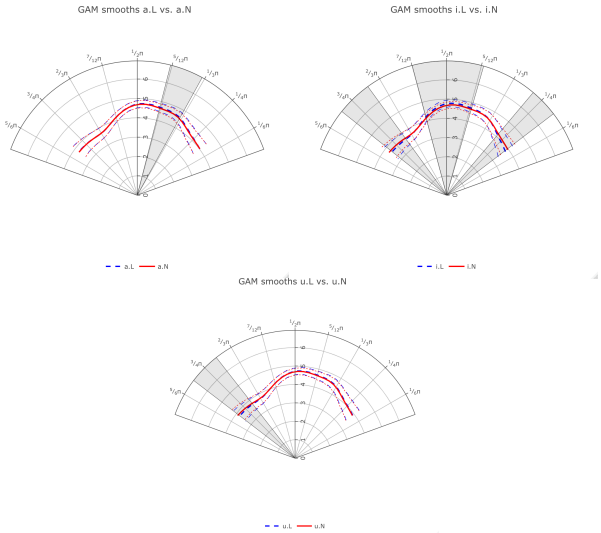


Figure 8: Fitted splines for /la/ vs. /na/, /li/ vs. /ni/, and /lu/ vs. /nu/. Gray areas indicate statistical significance. The speakers (n=4) are facing right.

3.3. The SS ANOVA results of acoustic data

Regarding the acoustic data, the formant values of the consonants were normalized using the Lobanov method ([16]). The SS ANOVA ([6]) results of the comparisons of the F1, F2, and F3 for the three pairs are provided in Table 2, where '√' means statistical significance and "0.5~1", for example, means the two trajectories are significantly different between 0.5 and 1 (consonant duration was normalized to 1).

Table 2: Pair-wise comparisons between /l/ and /n/ groups: SS ANOVA results.

L vs. N	F1	F2	F3
/la/ vs. /na/	√ 0.5~1	n.s.	n.s.
/li/ vs. /ni/	n.s.	n.s.	n.s.
/lu/ vs. /nu/	n.s.	√ 0.6~0.8	n.s.

4. DISCUSSION

Our first research question is concerned with the articulatory characteristics of /l/ in SWM. Firstly, it turned out that /l/ has two components, but they come in different flavors: TTz raising and TDx fronting (but not TD retraction, e.g., [11]). We also found that TT is formed relatively earlier than TD in /l/ production. Secondly, both EMA and ultrasound data confirm that both concave and convex tongue shapes are found in the coronal plane. This is an interesting observation because a domed coronal tongue shape has not been documented at least in /l/ production of English ([1], [12]), to our knowledge. The present results suggest that laterality might be subject to distinct phonetic implementations, especially when /l/ cannot appear in coda position in Mandarin.

Our second research question addresses a relatively understudied issue: the l~n merger. The acoustic and transcription results indicate that there is an ongoing merger, which is also confirmed by the ultrasound results. Inspired by [20], we proposed the lateralization angle of the tongue to quantify the degree of lateralization (1). The results show that there is no apparent difference between the /l/'s and /n/'s, suggesting that these two sounds are not distinguishable in terms of posture differentiation at the tongue blade. We take this finding as a potential articulatory indicator of the l~n merger. Finally, it is equally remarkable that the comparisons of the TT/TB sensor trajectories along the midsagittal planes are significantly different between the two groups. We speculate that the results in Table 1 reflect the kinematic differences between a stop ([n]) and a sonorant ([l]) in the anterior parts of the tongue. In other words, the l~n merger can be treated as a near merger, with the difference of the lateralization angle being lost first.

5. CONCLUSION

The present results suggest that /l/ production may involve language-specific characteristics (i.e., TD fronting in Fig. 3). More importantly, the l~n merger was confirmed not only by the transcription and acoustic results but also by the EMA and ultrasound data. In particular, the EMA data showed loss of posture differentiation at the tongue blade between /l/'s and /n/'s (see the lateralization angles in Figs. 5 and 6). In conclusion, this study contributes to understanding laterality from a novel perspective.

6. ACKNOWLEDGEMENTS

This study was supported by an NSTC research grant (MOST 109-2410-H007-061-). This study was approved by the National Tsing Hua University Research Ethics Committee (10812HS125). The first two authors are listed in alphabetical order as they contribute equally to this paper.

7. REFERENCES

[1] Berkson, K. H., de Jong, K., Lulich, S. M. 2017. Three dimensional ultrasound imaging of pre-and post-vocalic liquid consonants in American English: Preliminary observations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5080-5084.

[2] Boersma, P. 2009. Praat: doing phonetics by computer (Version 5.1.05). http://www. praat. org/.

[3] Cheng, R., Jongman, A., & Sereno, J. A. 2022. Production and Perception Evidence of a Merger: [l] and [n] in Fuzhou Min. *Language and Speech*, 00238309221114433.

[4] Gick, B., Allen, B., Roewer-Després, F., Stavness, I. 2017. Speaking tongues are actively braced. *Journal of Speech, Language, and Hearing Research* 60(3), 494-506.

[5] Giles, S. B., Moll, K. L. 1975. Cinefluorographic study of selected allophones of English /l/. *Phonetica* 31, 206–227.

[6] Gu, C. 2014. Smoothing spline ANOVA models: R package gss. Journal of Statistical Software 58(5), 1-25.

[7] Heyne, M., Derrick, D., Al-Tamimi, J. 2019. Native Language Influence on Brass Instrument Performance: An Application of Generalized Additive Mixed Models (GAMMs) to Midsagittal Ultrasound Images of the Tongue. *Frontiers in psychology* 10. 2597–2597.

[8] Soejima, K., Imagana, H., Kiritani, S. 1990. Alternation between Stop Nasal and (Nasalized) Flap or Lateral. *Ann. Bull. RILP.* 24, 131–144.

[9] Sóskuthy, M. 2021. Evaluating Generalised Additive Mixed Modelling Strategies for Dynamic Speech Analysis. *Journal of phonetics* 84, 101017.

[10] Shi, X. J. 2015. Chengdu hua xiangyin de bihuadu—jian lun /n,l/ bufen de shizhi ji leixing [The nasality degree of sonorant in Chengdu dialect]. *Journal of Chinese Phonetics*, 1, 92-100.

[11] Sproat, R., Fujimura, O. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21, 291–311.

[12] Stone, M., Lundberg, A. 1996. Three-dimensional tongue surface shapes of English consonants and vowels. *The Journal of the Acoustical Society of America*, *99*(6), 3728-3737.

[13] Tiede, M. 2005. MVIEW: software for visualization and analysis of concurrently recorded movement data. *New Haven, CT: Haskins Laboratories*.

[14] Tiede, M., Whalen, D. H. 2015. Getcontours: An interactive tongue surface extraction tool. *Proceedings of Ultrafest VII*.

[15] To, C. K., McLeod, S., Cheung, P. S. 2015. Phonetic variations and sound changes in Hong Kong Cantonese: Diachronic review, synchronic study and implications for speech sound assessment. *Clinical linguistics & phonetics* 29(5), 333-353.

[16] Thomas, E. R., Kendall. T. 2007. NORM: The Vowel Normalisation and Plotting Suite. Online Resource. http://lingtools.uoregon.edu/norm/about_norm1.php.

[17] Wang, L. 2004. *Hanyu Shigao [The history of Chinese]*. Beijng: Zhonghua Shuju.

[18] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70, 86-116.

[19] Wood, S. 2019. Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Available: https://cran.r-project.org/web/packages/mgcv.pdf.

[20] Ying, J., Shaw, J. A., Carignan, C., Proctor, M., Derrick, D., Best, C. T. 2021. Evidence for active control of tongue lateralization in Australian English/l. *Journal of Phonetics* 86, 101039.

[21] Zhang, W. 2007. Alternation of [n] and [l] in Sichuan dialect, Standard Mandarin and English: A single-case study. *Leeds Working in Linguistics and Phonetics* 12, 156-173.

[22] Zhang, W., Levis, J. M. 2021. The southwestern mandarin /n/-/l/ merger: effects on production in Standard Mandarin and English. *Frontiers in Communication*, 171.

COMPARING APPLES TO ORANGES - ASYNCHRONY IN JAW & LIP ARTICULATION OF SYLLABLES

Malin Svensson Lundmark[1] & Donna Erickson[2]

[1]Lund University, [1]University of Southern Denmark, [2]Haskins Laboratories, Yale University
[1]malin.svensson_lundmark@ling.lu.se, [2]ericksondonna2000@gmail.com

ABSTRACT

This study examines jaw and lip aperture characteristics of syllables. High correlations were found between velocity and displacement of opening/ closing of lips and jaw, similar to previous studies. However, lip aperture opening velocity changed as a function of manner and voicing of syllable onset while jaw opening velocity, except for a plosive onset, did not change. In addition, differences were found in timing between jaw opening and lip aperture opening, and the amount of difference varied as a function of syllable type (open/closed) and voicing/ manner of the consonant. Our interpretation of the asynchronous patterns of lip and jaw articulation is that the jaw is the syllable articulator and the lips are the syllable onset/coda articulators. The findings of this study have application to language acquisition as well as to stuttering research.

Keywords: syllable, peak velocity, displacement, jaw articulation, lip aperture

1. INTRODUCTION

This paper is about the separate linguistic functions of jaw and lip articulations. Jaw-lip coordination is a skill that infants need to acquire in order to produce CV(C) syllables. Studies by Green et al. [1] and [2] suggest that infants first acquire jaw articulation before lip articulation, findings supported by fetal studies showing that mandibular movement occurs before lip muscles are developed (e.g. [3], [4]). These findings are in coherence with MacNeilage's Frame-Content theory ([5]) that babies start producing syllable-type jaw movements at the age of about 6 months, about the time they start chewing. Green et al. [2] report that 1- and 2-year-old children have jaw movements similar to those of adults, but their lip movements show considerably more variability than adults. A similar finding was reported on variability in 4-year-olds' lip movements compared with those of 7- year-olds [6]. A suggestion is that children "rely" on jaw (syllable) production, later adding articulation for syllable onsets, and then codas [2].

Research has shown that correlations between velocity and displacement for both the lips and jaw are strongly significant ([7], [8], [9]). But, less work has been done with regard to opening-closing functions of the two articulators, apart from in Gracco ([10]), where examination of voiced and unvoiced bilabials suggested that the lip closing were more related to consonant production, and the jaw opening with vowel production. [11] reported that the lower lip tends to move with the jaw, but upper lip shows more variability. However, lip opening/closing articulation, i.e., "lip aperture", which includes both upper and lower lips, is rarely the focus of studies. Few studies have researched differences in lip or jaw velocity as a function of manner and voicing of syllable onset consonant, or of syllable type (open vs closed syllable); also, very few findings have been reported about differences between opening and closing movements of syllable articulations. This is surprising considering the suggested different purposes of the jaw and the lips in speech, which would yield more of a rhythmic pattern in the jaw cycle as compared to the recurrent nature of the lips revisiting a place of articulation.

In this study we examine (1) the relationship between velocity and displacement of opening/ closing of lips and jaw for syllable productions, (2) the difference between opening and closing velocity/displacement patterns, and (3) the relative timing of peak velocity of lip articulators compared with that of the jaw articulation.

2. METHOD

The material consists of 566 stressed word initial syllables in the target words: /mama/, /papa/, /ma:lar/ and /ba:lar/, as spoken by 18 Swedish speakers. The aggregated dataset is part of a larger corpus of EMA data on 21 Swedish speakers and about 6000 sentences [12].

The word initial syllables have vocalic nuclei of either [a] or [ɑ:]. The syllable onsets are bilabial [m], [p] and [b]. The coda consonants [m] and [p] in /mam/ and /pap/ are bilabial mora-sharing geminate, while the open syllables /ba:/ and /ma:/ are followed by a lateral [l] in the next syllable. All words are produced with the Swedish word accent 2 (a tonal rise in the stressed vowel). The target words are placed in statements preceded by leading questions, to ensure a non-focused elicitation.

2.1. Procedure

Speakers were recorded at 250 Hz with electromagnetic articulography (EMA); a Carstens AG501 at the Lund University Humanities Lab. Audio was recorded simultaneously using an external condenser microphone (a t.bone EM 9600); sampling rate 48 kHz. The speaker read a leading question and a target sentence from a computer screen in a random order, each set of target words appearing eight times.

The first author segmented the acoustic data manually in Praat [13] using ProsodyPro [14]. The textgrid files were used in R [15] as reference time windows for collection of the articulatory data.

2.1.1. Articulatory data and measurements

Articulatory data were collected from six sensors: two placed on the upper and lower lips at the vermilion border, one on the lower incisor (jaw), and three sensors placed on the midline of the tongue. To correct for head movements three additional sensors were used: one behind each ear and one on the nose ridge. All data post-processing was done in the Carstens software or in R, where additional analyses were done (see below). Only the sensor on the lips and the jaw were further analyzed in this study.

Peak velocity and displacement on the lips were measured on lip aperture, which was calculated in R using the three-dimensional Euclidian distance between the upper and the lower lip sensors. Peak velocity was calculated during opening at syllable onset (C1), and closing at syllable coda (C2) (Fig. 1). Displacement was measured at C1 and C2 from the maximum distance between the lips during the vowel.

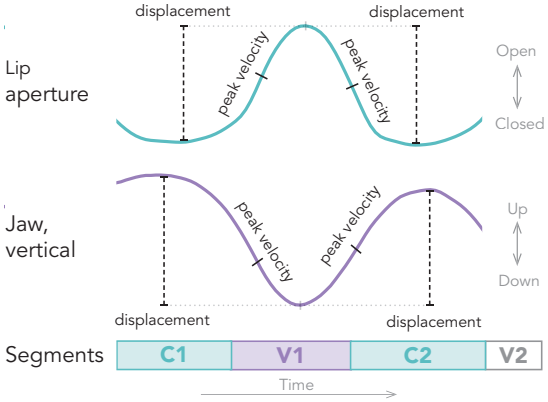


Figure 1: Articulatory measurements and calculations on lip aperture and jaw (example word /mama/ by one speaker).

For the jaw, peak velocity was calculated on the two-dimensional movement (vertical and horizontal) during jaw lowering (opening) and closing (Fig. 1). Displacement was measured as the distance between the highest and lowest vertical position during jaw opening and jaw closing. Notice this is different from studies which measure jaw displacement from the occlusal bite plane (e.g.,[16]). In addition, the relative timing of peak velocity was calculated by subtracting the time of lip peak velocity from jaw peak velocity.

2.3. Statistical analysis

All statistical tests are performed in R [15]. Pearson correlation tests are used to assess the relationship between the two dependent variables peak velocity and displacement. A one-way ANOVA is used to determine whether there is a statistically significant difference between the mean values for the four word initial syllables. As post-hoc tests we use pairwise comparisons with t tests (*pairwise.t.test* in R, which includes the BH adjustment method).

3. RESULTS

3.1. Syllable onset - jaw opening

The main finding from the correlation tests on the jaw is that velocity and displacement have a linear relationship (Fig. 2). The further the jaw moves, the faster it moves. This pattern is seen for all four of the word initial syllables. Thus, the relationship between displacement and velocity is constant, regardless of differences in syllable onset, i.e., /m/, /p/ and /b/.

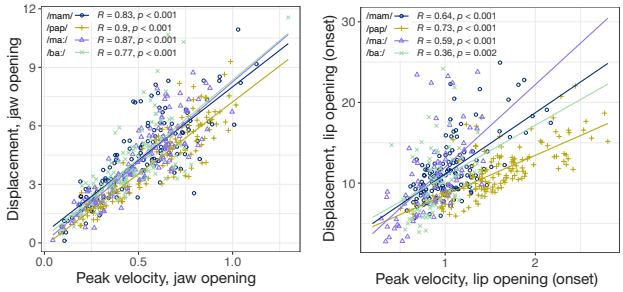


Figure 2: Jaw opening (left) and lip opening (right): displacement (mm) and peak velocity (cm/s).

However, the word initial syllable onsets show some variation in terms of their mean values of peak velocity and displacement, where the jaw appears to be the fastest in /pap/, but moves the longest distance in /mam/ (Table 1). A one-way ANOVA revealed statistically significant differences in jaw peak velocity among the word initial syllable onsets ($F(3,526) = 10.44$, $p < .001$). Pairwise comparisons using t tests with pooled SD found that the mean value of peak velocity was significantly faster for the plosive onset of /pap/ compared to the other word initial syllable onsets; /mam/ ($p < .05$), /ma:/ ($p < .001$) and /ba:/ ($p < .001$) (Table 2). However, for jaw displacement, no statistically significant differences were found between word initial syllable onsets, as determined by the one-way ANOVA ($F(3,526) = 0.511$, $p = .675$).

Dependent variables	Word onsets	Count	Syllable onset				Syllable coda			
			Jaw opening		Lip opening		Jaw closing		Lip closing	
			mean	*sd*	mean	*sd*	mean	*sd*	mean	*sd*
Peak velocity (cm/s)	/mam/	137	0.498	*0.228*	1.06	*0.310*	0.642	*0.246*	1.99	*0.629*
	/pap/	144	0.567	*0.244*	1.56	*0.473*	0.597	*0.247*	1.59	*0.946*
	/ma:/	147	0.468	*0.208*	0.806	*0.287*	0.514	*0.216*	1.37	*0.531*
	/ba:/	138	0.427	*0.186*	0.856	*0.262*	0.438	*0.204*	1.06	*0.351*
Displacement (mm)	/mam/	137	4.24	*2.06*	11.6	*3.68*	4.04	*1.93*	12.2	*3.58*
	/pap/	144	4.06	*1.98*	11.3	*3.17*	3.84	*2.06*	12.5	*3.44*
	/ma:/	147	4.12	*1.89*	10.6	*4.96*	2.01	*1.57*	1.93	*2.03*
	/ba:/	138	3.94	*1.94*	10.3	*4.03*	1.82	*1.42*	1.72	*1.88*

Table 1: Mean and standard deviation of the dependent variables: peak velocity (cm/s) and displacement (mm).

	Peak velocity		Displacement	
	Jaw open	Lip open	Jaw open	Lip open
/pap/ vs /mam/	.012	.000	.760	.570
/pap/ vs /ma:/	.000	.000	.800	.350
/pap/ vs /ba:/	.000	.000	.760	.180
/mam/ vs /ma:/	.241	.000	.760	.180
/mam/ vs /ba:/	.120	.000	.760	.120
/ma:/ vs /ba:/	.141	.220	.760	.570

Table 2: Syllable onset opening patterns of jaw and lips: p-values of pairwise comparisons (ANOVA post-hoc test).

3.2. Syllable onset - lip opening

The relationship between peak velocity and displacement for lip movements, compared to jaw opening, appears to be more affected by factors such as the manner and voicing of the onset consonant and whether the syllable has a coda or not (Fig. 2). Although the correlations are all significant, the r values are more robust for /pap/ (r = .73) and /mam/ (r = .64), moderate for /ma:/ (r = .59), and weak for /ba:/ (r = .36). In addition, velocity differ depending on manner, as well as voicing: during the lip opening, the lips move the fastest in /pap/ (1.56 cm/s, Table 1).

A one-way ANOVA revealed statistically significant differences in peak velocity in the word initial syllable onsets (F(3,562) = 141.9, p < .001). The post-hoc test found that the mean value of lip aperture peak velocity was indeed significantly different between all word initial syllable onsets (p < .001), except for when there was no coda consonant, i.e., between /ba:/ and /ma:/ (p = .22) (Table 2). Only marginally statistically significant differences were found on lip aperture displacement (F(3,429) = 2.298, p = .077).

3.3. Syllable coda - jaw closing

The results on the jaw closing movement are similar to those for jaw opening: as the jaw displacement is larger, so is the movement faster (Fig. 3). However, the strongest relation is for the closed syllables, /mam/ (r = .87) and /pap/ (r = .85). The disyllabic word initial open syllables, i.e. /ma:/ and /ba:/ preceding a syllable with a lateral consonant onset, show a less strong relationship (r = .63 and r = .68, respectively). For these words, the jaw perhaps closes only midway, and thus a shorter distance, whereas for the closed syllables, the jaw closes more completely, a longer distance (Fig. 3).

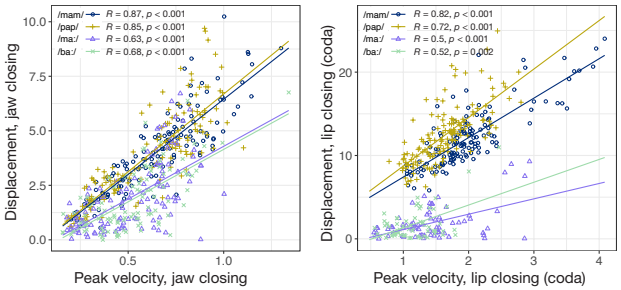


Figure 3: Jaw closing (left) and lip closing (right). Displacement (mm) and peak velocity (cm/s).

The ANOVA confirms that there are statistically significant differences in jaw peak velocity (F(3,562) = 14.35, p < .001), as well as in jaw displacement (F(3,495) = 52.38, p < .001). Interestingly, there is a tendency for the jaw to have the fastest closing for the coda of /mam/ (0.642 cm/s) which is opposite to the pattern of jaw opening (Table 1), where /pap/ shows the fastest jaw opening. Pairwise comparisons using t tests, however, found no statistically significant difference between /pap/ and /mam/ for either peak velocity (p = .123) or displacement of jaw closing (p = .44). Significant differences in jaw closing between the closed syllables (/mam/ and /pap/) and the open syllables (/ma:/ and /ba:/) were found for the mean values of peak velocity and displacement (Table 3).

	Peak velocity		Displacement	
	Jaw close	Lip close	Jaw close	Lip close
/pap/ vs /mam/	.123	.000	.440	.470
/pap/ vs /ma:/	.003	.000	.000	.000
/pap/ vs /ba:/	.000	.000	.000	.000
/mam/ vs /ma:/	.000	.000	.000	.000
/mam/ vs /ba:/	.000	.000	.000	.000
/ma:/ vs /ba:/	.248	.000	.450	.680

Table 3: Syllable coda closing patterns of jaw and lips: p-values of pairwise comparisons (ANOVA post-hoc test).

3.4. Syllable coda - lip closing

Lip closing displays a strong or very strong displacement-peak velocity correlation in /pap/ (r = .72) and in /mam/ (r = .82) (Fig. 3). Moreover, coda /m/ seems to be faster than /p/, while /p/ has a slightly larger lip opening than /m/ (Table 1). As /ma:/ and /ba:/ are open syllables with no syllabic coda, lip displacement/peak velocity is expected to be a by-product of the syllabic or vocalic articulations.

A one-way ANOVA revealed that there are statistically significant differences between the words in both peak velocity ($F(3,562) = 84.94$, $p < .001$), and in lip displacement ($F(3,417) = 360.7$, $p < .001$). A post-hoc test found that the mean value of lip closing peak velocity was significantly different between all word onsets ($p < .001$), while lip closing displacement was only significantly different between the closed and the open syllables (Table 3).

3.5. Relative timing of peak velocity

Regarding the relative timing of peak velocity of the jaw and the lips, a one-way ANOVA reveals statistically significant differences between the words during opening ($F(3,533) = 40.64$, $p < .001$), but not during closing ($F(3,416) = 0.179$, $p = .911$). Post-hoc tests show differences in timing during opening both between the type of syllables (open or closed), as well as between types of syllable onsets/codas, i.e., /mam/ and /pap/ ($p < .001$): in /mam/ the jaw peak velocity is timed with the lips, while in /pap/ the jaw peak velocity is instead about 15 ms after the lip peak velocity. When comparing syllable types, the jaw peak velocity occurs before the lip peak velocity in the open syllables, while after in the closed syllables. This difference is about 30 ms, but we also need to look at inter-speaker differences.

4. DISCUSSION AND CONCLUSION

Similarities and differences are observed for jaw and lip articulation of syllables. For both lip aperture and jaw, we see that the faster the movement, the larger the displacement, a relationship that follows the law of nature. However, for lip aperture opening, velocity varies significantly depending on both manner and voicing, and also whether the syllable is open or closed; but for jaw opening, generally there are no such effects. For coda articulation, lip aperture and jaw are both sensitive to syllable-type combinations, i.e., open-closed or closed-open syllables.

The finding for syllable onsets of essentially no difference in jaw opening velocity/displacement patterns due to initial consonant, yet differences in lip aperture due to manner and voicing of syllable onsets, suggests the two articulators have different functions for speech. These results are in accordance with the findings of Gracco ([10]). Although in [10] /p/ was faster during both lip closing and opening, while we found faster /m/ in lip closing. We suspect this difference between our results to be related to the prosodic characteristics of the syllable (in [10] this was not controlled for) which in our case is always a stressed syllable. Thus, the different functions of the lip and the jaw seem to be dependent on syllable prominence; the specific velocities of different manner of articulation may even rely on the level of jaw displacement. Hence our proposal that the jaw is the syllable articulator while the lips are the segmental articulators. See e.g. work by Svensson Lundmark on segmental articulations [17], and work by Erickson and colleagues ([18] and [19]) which proposes that the jaw is the syllable articulator, where the amount of jaw displacement manifests the degree of syllable prominence, once the intrinsic vowel effects have been factored out ([16] and [20]). This interpretation of the jaw as the syllable articulator fits in with pre- and post-natal studies showing that jaw articulation develops before lip articulation.

About timing differences of peak velocity between jaw-lip aperture, specifically, that timing during opening is affected by both syllable type, i.e, open or closed, and manner and voicing of the initial consonant—as far as we know, this specific combination has not been reported in previous studies. Future work is needed to explore timing of lip-jaw opening/closing, especially, with regard to children's acquisition of syllables; also, how this may relate to stuttering in terms of synchronization of jaw-lip articulation for syllable production.

A final note on our data set, which is limited in that the segmental structure does not entail us to look at the lip opening of both onset and coda (e.g., comparing a release burst in /baba/ to /papa/), or vice versa: lip closing of both coda and onset. We are left with comparing apples to oranges, i.e., two halves of recurrent lip constrictions to the more rhythmic jaw cycle. In order to compare the syllable articulator to the segmental articulators, the correct approach would be to compare the jaw cycle (opening + closing) to two constriction cycles: closing and opening of both the onset and of the coda consonant. Moreover, as the lip opening of the jaw cycle is correlated with the type of vowel (narrow, open, rounded), the lip constrictions movement patterns are dependent on not only the jaw displacement/velocity, but also on the type of vowel. Hence, for vowel production, we suggest a combined approach of the jaw, lip and the tongue to be necessary.

5. ACKNOWLEDGMENTS

This work was supported by an International Postdoc grant from the Swedish Research Council (Grant No. 2021-00334), and by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, 2018–2024; Grant No. 2017-00626). The authors gratefully acknowledge the Lund University Humanities Lab.

6. REFERENCES

[1] Green, J. R, Moore, C. A, Higashikawa, M, Steeve, R.W. 2000. The physiologic development of speech motor control: Lip and jaw coordination. *J. Speech Lang. Hear. Res.* 43, 239–255.

[2] Green, J. R., Moore, C. A., Reilly, K. J. 2002. The Sequential Development of Jaw and Lip Control for Speech, *J. Speech Lang. Hear. Res.* 45.1., 66–79. doi:10.1044/1092-4388(2002/005).

[3] Gasser, R.F. 1967. The development of the facial muscles in man. *American Journal of Anatomy* 120, 357–376.

[4] Humphrey, T. 1964. Some correlations between the appearance of human fetal reflexes and the development of the nervous system. *Progress in Brain Research* 4, 93–135.

[5] MacNeilage, P. F., Davis, B.L. 1990. Acquisition of speech production: The achievement of segmental independence. In: Hardcastle, W. J.; Marchal, A., (eds), *Speech Production and Modeling*. Dordrecht: Kluwer Academic, 55-68.

[6] Sharkey, S. G, Folkins, J. W. 1985. Variability of lip and jaw movements in children and adults: Implications for-the development of speech motor control. *J. Speech Lang. Hear. Res.* 28, 8–15.[PubMed: 3982001].

[7] Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. 1985. A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling, *J. Acoust. Soc. Am*. 77, 266-280.

[8] Munhall, K. G., Ostry, D. J., Parush, A. 1985. Characteristics of velocity profiles of speech movements. *J. Experimental Psychology* 11.4, 457–474.

[9] Kollia, H. B., Gracco, V. L., Harris, K. S. 1995. Articulatory organization of mandibular, labial, and velar movements during speech. *J. Acoust. Soc. Am*. 98, 1313 (1995); doi: 10.1121/1.413468, https://doi.org/10.1121/1.413468

[10] Gracco, V. 1994. Some organizational characteristics of speech movement control. *J. Speech and Hearing Research* 37, 4–17

[11] Van Lieshout, P. H. H. M. 2015. Jaw and lips. In: Redford, M. A. (ed.) *The Handbook of Speech Production*, John Wiley & Sons, Inc. 79–108.

[12] Svensson Lundmark, M. 2020. Articulation in time. Some word-initial segments in Swedish. Lund University

[13] Boersma, P., & Weenink, D. 2018. Praat: Doing phonetics by computer [Computer software]. Version 6.0.37. Retrieved 3 February 2018 from http://www.praat.org/.

[14] Xu, Y. 2013. ProsodyPro—A tool for large-scale systematic prosody analysis. 4. In: Bigi, B. & Hirst, D. (eds.), *Proceedings of tools and resources for the analysis of speech prosody* (TRASP 2013), Aix-en-Provence, France 2013, 7–10. Laboratoire Parole et Langage. http://www2.lplaix.fr/~trasp/Proceedings/TRASP2013_proceedings.pdf.

[15] R Core Team. 2015. R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

[16] Erickson, D. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59, 134–149.

[17] Svensson Lundmark, M. 2023. Rapid movements at segment boundaries. *J. Acoust. Soc. Am.* 153 (3).

[18] Erickson, D. and Kawahara, S. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistic Vanguard* 2, 102–110.

[19] Erickson, D. and Niebuhr, O. 2022. Articulation of Prosody and rhythm. *Nordic Prosody 13*.

[20] Williams, J. C., Erickson, D., Ozaki, Y., Suemitsu, A. Minematsu, N., Fujimura, O. 2013. Neutralizing differences in jaw displacement for English vowels, *Proc. of International Congress of Acoustics*. *POMA 19, 060268.*

INDIVIDUAL DIFFERENCES IN SPEECH PRODUCTION: WHAT IS "PHONETIC SUBSTANCE"?

Keith Johnson

University of California, Berkeley
keithjohnson@berkeley.edu

ABSTRACT

This study explored the talker as a source of phonetic variation. Two patterns emerged in a canonical correlation study of the Wisconsin X-ray Microbeam Database relating individual differences in vocal tract anatomy to differences in the phonetic realization of vowels and consonants. Talkers with a large vocal tracts tended to show large tongue vertical range of motion in vowel production, while those with smaller vocal tracts showed larger jaw vertical range of motion. Talkers who have a more deeply domed palate show more horizontal lower-lip range of motion in vowels, and tend to have more tip down posture in coronal fricatives, while talkers with less deeply domed palate showed larger range of tongue horizontal position in vowels and more tip up fricatives.

Keywords: Speech articulation, vocal tract anatomy, sound change, individual differences.

1. INTRODUCTION

Many phonological patterns emerge historically from phonetically motivated, natural sound changes which are based on a "pool of synchronic phonetic variation" [1, 2]. This study explored one source of phonetic variation, individual differences in vocal tract anatomy, with a focus on the correlated phonetic bias that may result from anatomy.

2. BACKGROUND

The relationship between individual differences in vocal tract anatomy and articulatory phonetic variation has been studied many times [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. For example, the depth of the palate vault has been related to the amount of lip/tongue trading in /u/ [10], the amount of coarticulation in front consonants [11], the articulatory variability of /s/ and /r/ [3, 4], the amount of jaw or tongue movement in a vowel height distinction [8], and the amount of articulatory difference between /s/ and /ʃ/ [13]. Similarly, the length of the oral cavity has been found to relate to the verticality of high vowel variation [15], and the verticality of front consonant coarticulation [11], while the overall length of the vocal tract has been related to speed and magnitude of articulatory movements [9, 12, 6], and to the verticality of vowel [a] to [i] movements [7].

Many of these studies rely on a small number of speakers, and most of them focus on only one or two anatomical or articulatory features. The present study contributes to this literature, and our overall understanding of the pool of phonetic variation, by looking at several dimensions of consonant and vowel articulation and several anatomical features in a mutually constraining statistical model. Using canonical correlation [16, 17], we jointly constrain the search for anatomical and articulatory patterns that are related to each other.

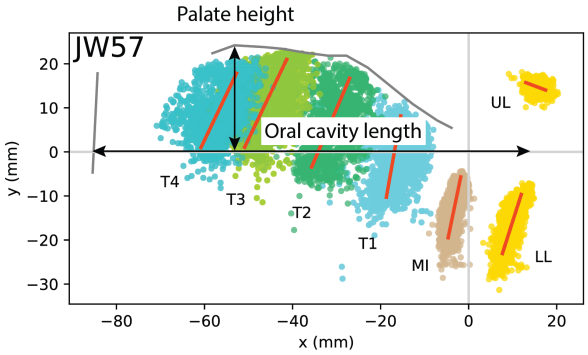


Figure 1: Example data, showing about 1500 stressed vowel midpoint pellet locations for talker JW57. The anatomical measures "palate height" adn "oral cavity length" are illustrated.

3. METHODS

3.1. Dataset

The Wisconsin X-ray Microbeam Database [18, 19] was used in this study. The database (for which textgrids are now available [19]) contains audio, articulatory point tracking data, and time aligned word and phone tags, for 48 dialectically

homogeneous talkers reading word lists, sentences, and short stories. Additionally, a palate-trace was produced for each talker and the approximate location of the pharyngeal wall was determined. Talkers also were recorded in jaw-wagging and tongue protrusion trials. The articulatory data are the x and y locations of gold pellets glued onto the upper lip (UL), lower lip (LL), lower incisor (MI), and four pellets on the tongue (T1-4) in the mid-sagittal plane (see Fig. 1). These were recorded at a sampling rate of 145.64 Hz, and the data were rotated to a standard occlusal coordinate system, with (0,0) at the tip of the upper incisors.

3.2. Anatomical Measures

Table 1 lists the nine measures of vocal tract anatomy that were taken for each talker. Vocal tract length was calculated from vowel formants [20], and pharynx length was VTL minus oral tract length (Fig. 1). Maximum tongue protrusion and maximum jaw opening were measured from recordings in which talkers were asked to perform these motions. The other measurements were taken from the palate trace.

Anatomical measures
Vocal tract length
Oral tract length
Pharyngeal cavity length
Depth of the palate dome
Area under the palate
Area relative to length
Location of palate peak
Maximum tongue protrusion
Maximum jaw opening

Table 1: Anatomical measures.

3.3. Articulatory Parameters

Table 2 lists the articulatory parameters that were measured. The first three are measures of the articulatory vowel space. The input for these measures is a collection of data from the midpoints of all stressed vowels produced by a speaker (see Fig. 1). For each pellet the horizontal and vertical ranges (at the 10th and 90th percentiles) of the cloud of data points were recorded. Thus, there were seven horizontal range measures (one for each pellet) and seven vertical range measures. A measure of the verticality of the data cloud for each pellet was taken from the orientation of the first principal component of variation - the lines in each cloud in Fig. 1. The remaining measures listed in the table were taken from fricative midpoints, grouped by place of articulation into dental [θ, ð], alveolar [s,z], and post-alveolar [ʃ, ʒ , ʧ, ʤ]. Tip posture was measured as the mean difference between T1y and T2y for each place of articulation. The remaining measures of retraction, rounding, fronting etc. were taken by comparing the mean posture in dentals or post-alveolars to that in the alveolar fricatives. For example, the θ blade lowering score for a talker is the difference between the mean vertical location of the T2 pellet in their [s,z] and the mean vertical location in their [θ,ð]. There were 33 articulatory parameters in all.

Articulatory parameters
Horizontal range in vowels
Vertical range in vowels
Verticality of vowel variation
Tip posture
/ʃ/ retraction
/ʃ/ rounding
/θ/ fronting
/θ/ blade lowering
/θ/ jaw opening

Table 2: Articulatory parameters.

3.4. Statistical Analysis

Conceptually, canonical correlation analysis (CCA) finds principal components of variation in one matrix (X, which will be our anatomical measures) that are maximally correlated with principal components of another matrix (Y, which will be our articulatory parameters), and it solves for PCs in the two matrices simultaneously in order to maximize the correlations between patterns of variation in X with patterns of variation in Y.

(1)
$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \rho^2 \begin{pmatrix} C_{XX}+\lambda I & 0 \\ 0 & C_{YY}+\lambda I \end{pmatrix}$$

In practice (Eq. 1, [17]) this is solved as an eigenvalue problem with the covariance matrices for X and Y, C_{XY} and C_{YX}, and their autocorrelation matrices, C_{XX} and C_{YY}. The weight matrix $\mathbf{a}$ defines a pattern in X that is correlated with a pattern in Y which is defined by the weight matrix $\mathbf{b}$. The most highly correlated pattern (i.e. the highest value of ρ) is the first canonical correlation (ρ_{CC1}) and it is achieved with canonical weights $\mathbf{a}_{CC1}$ and $\mathbf{b}_{CC1}$.

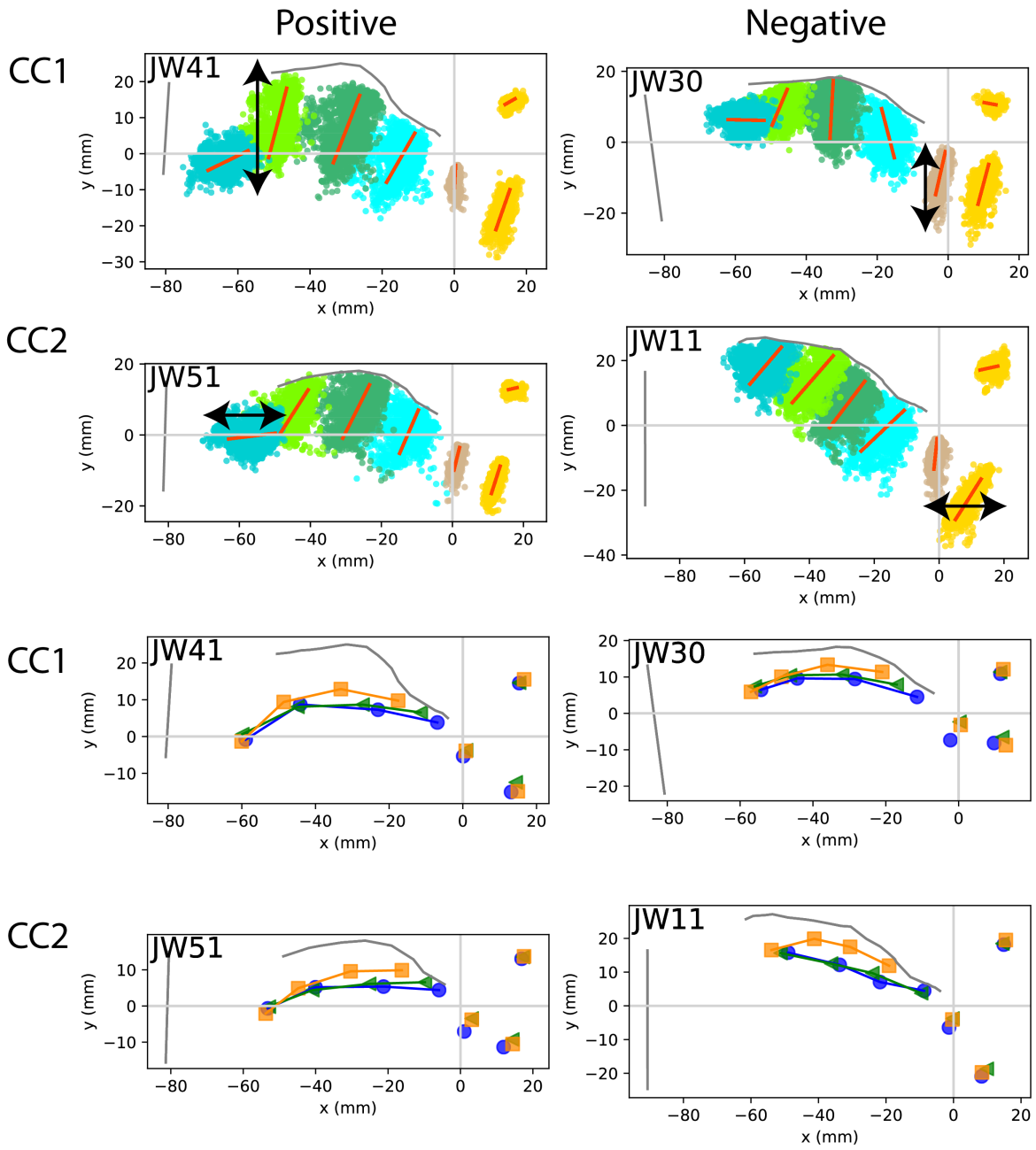


Figure 2: Talkers exemplifying the articulatory patterns discovered in the canonical correlation analysis. The **top row** shows the pattern of vowel articulation that was captured by the first canonical correlation (CC1). Talker JW41 illustrates the pattern for a talker who had a positive score on CC1 and talker JW30 illustrates the pattern for a negative score on CC1. The **second row** illustrates the patterns of vowel variation that were captured by the second canonical correlation (talker JW51 had a positive score on CC2 and talker JW11 had a negative score for CC2). The clouds of points show the midpoint pellet locations for the stressed vowels produced by the talker. Lines drawn in each cloud of points shows the orientation of the first principal component of variation of that cloud. The **bottom four panels** illustrate the patterns found for coronal fricatives for the same example talkers. The orange squares show the average pellet locations for post-alveolar fricatives, the green triangles for alveolar fricatives, and the blue circles for dental fricatives.

In this study, two canonical correlation patterns (CC1 and CC2) between speech anatomy and speech articulation were found. These are patterns in vowel production and coronal fricative production that are correlated with aspects of vocal anatomy.

Imposing L2 regularization (with λI in Eq. 1) constrains the norms of canonical weights **a** and **b** which stabilizes the solution when there are many variables in the matrices X and Y.

This study used a cross-validation procedure which was repeated 100 times and then the most frequently selected analysis parameters were used in the final analysis reported here. Based on this procedure, the regularization parameter λ was set to 0.04, and only the first 2 canonical correlations were kept. The canonical correlations between the anatomical variables and the articulatory parameters were $\rho_{CC1} = 0.81$ for the first component, and $\rho_{CC2} = 0.73$ for the second component.

4. RESULTS

Fig. 2 illustrates the main findings of this study. The first canonical correlation (CC1) was most highly weighted for vocal tract length, and pharyngeal cavity length in the anatomical measures. Among the vowel articulatory features, CC1 had positive weight for tongue body vertical range, and had negative weight for jaw x and y range. This is illustrated by the top row of Fig. 2 by talkers who had large positive or negative scores on CC1. Talker JW41 who has a long vocal tract and a large vertical vowel space for pellets T2 and T3, is compared to talker JW30 who has a shorter vocal tract and more jaw range in vowels than tongue body range. Talkers like these were called "jaw movers" and "tongue movers" in [8]. In coronal consonants CC1 was associated with TH_blade lowering.

The second canonical correlation (CC2) was related to palate doming independent of VTL. In vowels, this component linked shallow palate with larger horizontal range of tongue positions, while deeper palate was association in CC2 with larger horizontal range of lower lip positions. This is illustrated by talkers JW51 and JW11 in the second row in Fig. 2. CC2 also captured variation in coronal consonant articulation; shallow palate was associated with tip up posture, while deep palate was associated with tip down posture for all three places of articulation.

5. CONCLUSION

This analysis of the Wisconsin X-ray Microbeam Database found that the pool of phonetic variation is tied to talker anatomical differences; talkers contribute unique patterns of phonetic variation (and covariation across segments), resulting in talker specific phonetic substance. This means, for example, that patterns of coarticulation are probably talker-specific because the articulatory movements in producing vowel or consonant contrasts differ in terms of which articulator makes the larger movement. Similarly, trading relations between articulatory gestures in producing phonetic contrast appear from these data to be talker-specific. In general, because perfect articulatory imitation may not be possible, we have to conclude that phonetic knowledge is necessarily somewhat abstract.

6. REFERENCES

[1] J. Ohala, "The phonetics of sound change," in *Historical Linguistics: Problems and Perspectives*, C. Jones, Ed. Logman, 1993, pp. 237–278.

[2] A. Garrett and K. Johnson, "Phonetic bias in sound change," in *Origins of Sound Change: Approaches to Phonologization*, A. Yu, Ed. Oxford, 2012.

[3] S. Bakst and S. Lin, "An ultrasound investigation into articulatory variation in American English /r/ and /s/," in *Proc. 18th ICPhS*, Glasgow, 2003.

[4] S. Bakst, "Differences in the relationship between palate shape, articulation, and acoustics of American English /r/ and /s/," *UC Berkeley PhonLab Annual Report*, no. 12, 2016.

[5] J. Brunner, S. Fuchs, and P. Perrier, "On the relationship between palate shape and articulatory behavior," *J. Acoust. Soc. Am.*, vol. 125, pp. 3936–3949, 2009.

[6] S. Fuchs, R. Winkler, and P. Perrier, "Do speakers' vocal tract geometries shape their articulatory vowel space?" in *8th International Seminar on Speech Production, ISSP'08*, Strasbourg, France, 2008, pp. 333–336.

[7] K. Honda, S. Maeda, M. Hashi, J. Dembowski, and J. Westbury, "Human palate and related structures: Their articulatory consequences," in *Proc. ICSLP 2*, 1996, pp. 784–787.

[8] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *J. Acoust. Soc. Am.*, vol. 94, pp. 701–714, 1993.

[9] D. Kuehen and K. Moll, "A cineradiographic study of VC and CV articulatory velocities," *J. Phonetics*, vol. 4, pp. 303–320, 1976.

[10] J. Perkell, M. Matthies, M. Svirsky, and M. Jordan, "Goal-based speech motor control: A theoretical framework and some preliminary data," *J. Phonetics*, vol. 23, pp. 23–35, 1995.

[11] K. Rudy and Y. Yanusova, "The effect of anatomic factors on tongue position variability during consonants," *J. Speech, Lang. Hearing Res.*, vol. 56, pp. 137–149, 2013.

[12] A. Simpson, "Dynamic consequences of difference in male and female vocal tract dimensions," *J.*

Acoust. Soc. Am., vol. 109, pp. 2153–2164, 2001.
[13] M. Weirich and S. Fuchs, "Palate morphology can influence speaker-specific realizations of phonemic contrasts," *J. Speech, Lang. Hearing Res.*, vol. 56, pp. 894–908, 2013.
[14] R. Winkler, S. Fuchs, and P. Perrier, "The relation between differences in vocal tract geometry and articulatory control strategies in the production of French vowels: Evidence from MRI and modeling," in *7th International Seminar on Speech Production ISSP*, Ubatuba, 2006, pp. 509–516.
[15] R. Winkler, S. Fuchs, P. Perrier, and M. Tiede, "Biomechanical tongue models: An approach to studying inter-speaker variability," in *Proc. 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, 2011, pp. 273–276.
[16] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[17] N. Bilenko and J. Gallant, "Pyrcca: Regularized kernel canonical correlation in python and its applications to neuroimaging," *Frontiers in Neuroinformatics*, vol. 10, p. 49, 2016.
[18] J. Westbury, "X-ray Microbeam Speech Production Database User's Handbook, version 1.0," unpublished manuscript, Madison, WI.
[19] Wisconsin X-ray Microbeam Database. [Online]. Available: https://github.com/rsprouse/xray_microbeam_database
[20] A. Lammert and S. Nrayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLoS ONE*, vol. 10, 2015.

MODELING THE TEMPORAL EVOLUTION OF THE VOCAL TRACT SHAPE WITH DEEP LEARNING

Yves Laprie[a], Vinicius Ribeiro[a], Karina Isaeva[b], Justine Leclere[b,c], Jacques Felblinger[b,d], Pierre-André Vuissoz[b]

[a] Université de Lorraine, CNRS, Inria, LORIA, Nancy, F-54000, France

[b] Université de Lorraine, INSERM, U1254, IADI, Nancy, F-54000, France
[c] Service de Médecine Bucco-dentaire, Hôpital Maison Blanche, Reims, F-51100, France

[d] CIC-IT 1433 INSERM CHRU Nancy, F-54000, France
Yves.Laprie@loria.fr

ABSTRACT

This paper overviews our work on the links between coarticulation modeling, approached from the point of view of predicting the vocal tract shape from the phonetic sequence, and the available real-time MRI corpora. Real-time MRI has revolutionized the acquisition of articulatory data through the image quality, the possibility of acquiring and denoising the speech signal, and the possibility of recording corpora containing several thousand sentences. Coarticulation modeling is only possible with the ability to reliably track articulator contours in many images. Tracking techniques using neural networks have provided efficient solutions comparable in reliability to humans. Finally, we show that even if recurrent neural networks trained on these corpora can successfully predict the shape of the vocal tract, it is still necessary to use constraints directly from phonetics to ensure consistency in the prediction.

Keywords: vocal tract shape, articulatory modeling, rt-MRI, RNN, deep learning

1. INTRODUCTION

Modeling coarticulation is a central issue in phonetics which thus raised many works intended to approximate the position of articulators or, equivalently, to predict the constriction locations. One of the great difficulties has always been supporting theories on data in sufficient quantity and quality. For this reason, research has been structured in two directions. The first proposed theories focused mainly on the scope of anticipation of articulatory gestures and then tried to construct small corpora to validate them. Globally, the debate centered on the scope of anticipation, depending on whether it has a fixed duration or, on the contrary, depends on the constraints imposed by the achievement of constrictions in the vocal tract. These approaches have suffered from the handicap of relying on a limited amount of data without the possibility of moving from theory to a numerical model. Farnetani [1] offers a complete overview of approaches from phonetics and phonology.

The second method consisted in proposing strongly constrained numerical models and fitting their parameters from small corpora, which is possible because the introduction of constraints reduces the number of parameters to be fitted. In practice, the Öhman model [2, 3] is probably the most widely used model. Cohen and Massaro's model [4] based on dominance functions was only used for labial modeling but could have easily been extended to other articulators. This model requires five parameters for each articulatory factor and each phoneme, with a total of about 4 000 parameters. Nam et al. [5] determined the parameters of the gestural scores associated to the articulatory phonology proposed by Browman and Goldstein [6] and the TADA task dynamics model [7]. The gestures are assumed to be known and the learning process optimizes two parameters: the onset and offset times of each gesture from a Microbeam X-ray database [8] recorded for 47 speakers who uttered at most 56 sentences. Thus, two parameters for each occurrence of the gesture are optimized to copy the speech signal best. This approach has not been used to generate the shape of the vocal tract but only to carry out automatic speech recognition. Our main idea was to impose a minimum of constraints on the learning process and the global model was therefore learned directly from the data of a speaker who pronounced 700 sentences in French. We now present a larger corpus of 2100 French sentences for one female and one male speaker that will allow

us to extend this approach. The data consisted of the contours of all articulators in the vocal tract, as shown in Figure 1.

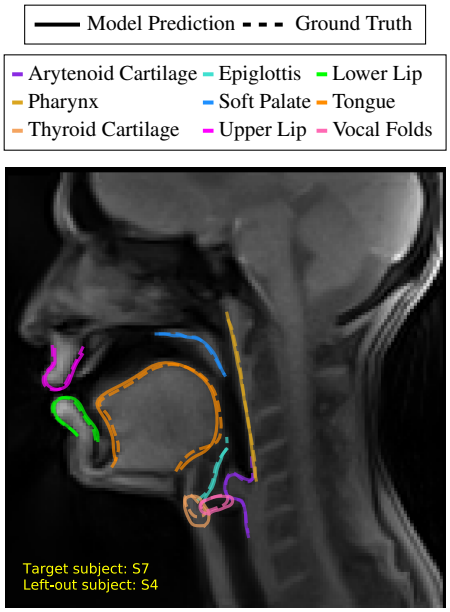


Figure 1: Articulator contours. Authorized reproduction from [9].

2. ARTICULATORY DATA AND ACQUISITION TECHNIQUES

Access to data has always been a major obstacle to the study of coarticulation. Cineradiography was the first source of data, but the hazards to which the subjects were exposed limited the duration of the films to about 30 seconds per year at most and this technique was finally abandoned in the late 1980s for ethical reasons. Then, electromagnetic articulography (EMA) allowed the data acquisition in larger volumes with limited risks.

The strength of EMA is to offer a high sampling frequency of up to 800 HZ but with the limitation of monitoring the position of only a few sensors, what is more with the difficulty of having to glue these sensors for a duration reduced to 45 minutes in the best case. The substantial increase in the recording duration is essential to enable machine learning since it heavily relies upon data availability. For this reason, EMA databases, particularly those acquired by Richmond [10], have been widely used to perform articulatory acoustic inversion experiments. Nevertheless, the limitation of the number of sensors prohibits the determination of tongue root and larynx movements, which play a crucial role in the length of the vocal tract and thus on the acoustic properties of speech.

2.1. MRI specificities w.r.t. speech production

Magnetic resonance imaging and its real-time version significantly evolved the observation of articulatory gestures. It is now possible to acquire excellent quality data at a frequency of 50 Hz, and even recently 80 Hz, with a sufficient spatial resolution (1.4 mm pixels) for our application [11] with the system developed at the Max Planck Institute for Multidisciplinary Sciences by Uecker et al. [12]. The speech signal can be recorded with an optical microphone and be almost entirely denoised using recent techniques [13]. The speech recorded in this way suffers only from the Lombard effect due to machine noise and the supine position.

Despite these very substantial advances, there are still limitations imposed by the MRI acquisition technique itself. The first, called "partial volume effect" [14], is related to the thickness of the mid-sagittal slice. This thickness is of the order of 8 mm to obtain a sufficient signal-to-noise ratio (SNR), hence, a good contrast. Therefore, the volume giving rise to a pixel is $1.4 \times 1.4 \times 8$ mm in our data. In some cases, this volume is only partially occupied by the imaged organs, especially near the edges of the tongue at the groove and the apex level, leading to pixels that are sometimes difficult to classify as part of the tongue or not. The second limitation is related to the MRI acquisition technique. The acquisition is performed in the image's Fourier domain, and a complete image would require as many elementary acquisitions as the number of lines in the image. The real-time acquisition technique is based on acquiring a small number of data – in this case, nine variable rays in Fourier space for the algorithm of Uecker et al. [12] – which reconstructs the complete image with regularization techniques. Each elementary acquisition (one ray in Fourier space) requires a little more than 2 ms, and the fast organs, mainly the apex and the lips, can therefore move while obtaining an image. This effect is particularly noticeable in the case of the apex, which comes into contact with the alveolar or post-alveolar region during articulation of the /l/, /d/, /n/ and /t/ sounds. For the apex, and to a lesser extent for the lips, the two effects are cumulative, so it is always challenging to determine the exact position of the apex and the lips. For this reason, the images cannot be considered as representing the ground truth.

2.2. Acquisition strategy

One challenge is acquiring large corpora of several thousand sentences to have excellent coverage of phonetic contexts for a speaker. For ethical and practical reasons, it is not possible to exceed one hour and 15 minutes in the MRI machine; thus, it is necessary to organize several recording sessions with the same speaker posture to guarantee the homogeneity of the articulatory data concerning the relative positions of the oral and pharyngeal cavities. Therefore, we made a blocking foam perfectly adapted to the MRI antenna and the speaker's head. First, a cast of the back of the head was made in alginate, which was used to create a plaster positive. Next, with a 3D-printed copy of the antenna, we placed the plaster cast in the same position as the speaker's head during the acquisition. We filled the space between the two with expanding foam (FlexFoam-iT III). The foam-blocking helmet (Figure 2d) can be used during the acquisition to ensure the head's position is the same during all sessions. For the last corpus constructed that will be used for future coarticulation models, six recording sessions were required for approximately 2 100 French sentences.

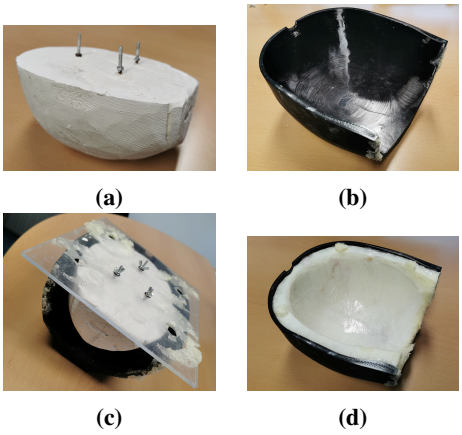


Figure 2: (a) Plaster cast of the head, (b) copy of the antenna, (c) adjustment of the cast in the antenna, (d) blocking foam.

The MRI acquisition parameters are identical to those in [11]. Each acquisition lasts 80 or 90 seconds, allowing the images and the speech signal to be easily processed. Coarticulation modeling requires precise phonetic segmentation because it is essential that the images, and thus the position of the articulators, correspond perfectly to the articulated phoneme. Forced-alignment tools developed for automatic speech recognition (ASR) were used to align the speech signal with the text transcription. The resulting phonetic segmentations were then manually corrected with great care. Coarse errors were corrected, and we separated the closure from the burst stops so that the vocal tract shape prediction could model the relevant closures. This correction work requires about an hour and a half for every 177 acquisitions in the new corpus.

In addition to phonetic segmentation, it is necessary to obtain the position of each articulator. The emergence of deep convolutional networks for image segmentation [15, 16] allowed considerable progress. The system initiated in [17] was extended to cover all articulators from the glottis to the lips in [9], and the mandible and hard-palate positions were estimated using image correlation. We manually delineated the contours in about 2 000 images, which is reasonable compared to the current corpus size of more than 500 000 images. The performance is very close to that of a human with the advantage of providing a remarkable homogeneity in the tracking behavior, unlike a group of human experts which always presents disparities. Due to lack of space, this system is not fully described, but we refer to Ribeiro et al. [9] for a detailed description. Figure 1 gives an illustration of the tracking results.

3. PREDICTING THE VOCAL TRACT SHAPE AND COARTICULATION EFFECTS

One of our primary goals is to model coarticulation and vocal tract shape prediction as independently as possible from constrained numerical models. Modeling can be done at two levels: the geometrical representation of the vocal tract shape and then the prediction of the vocal tract shape. Many articulatory models have been proposed to approximate the shape of the vocal tract. The models can be based either on the use of geometric primitives (lines, circles, and others, in two dimensions, or planes, disks, and others, in three dimensions) [18], or on the analysis of a set of articulatory contours extracted from a corpus of midsagittal images of the vocal tract [19, 20, 21]. For both models, the vocal tract shape is represented by a vector of about ten parameters in the best case, seven for the Maeda model [22] but at the cost of relatively simplified modeling of the vocal tract shape. In the case of models obtained by applying a data analysis technique, a marked geometrical compensation can be generally observed, complicating the choice of the best articulatory vector to represent one shape. Thus, we decided to forego intermediate modeling and use articulatory contours as obtained by the tracking.

The first experiment presented in [23] relies on an deep encoder-decoder neural network built using a bidirectional Gated Recurrent Units (GRU) [24] and a linear decoder with `ReLU` nonlinearities to reconstruct the articulators' shapes. Model training is performed from the contours of 30 90-second acquisitions and the evaluation focuses on the geometric accuracy and realization of four classical articulatory variables: lip aperture (LA), tongue tip constriction degree (TTCD), tongue body constriction degree (TBCD) and velum opening (VEL), which have a significant acoustic impact. The experiments produced very good results despite the limited amount of data available. As we have seen above, the physical origin of the real-time MRI images does not allow to know precisely the position of the apex and the lips, which play an essential role in the realization of complete or partial constrictions of the vocal tract. Also, one of the important points to consider in the vocal tract shape prediction is the model's ability to compensate for a perturbation imposed on an articulator. For these reasons, it would have been preferable to have constraints that better control the degree of a particular constriction. Since each contour is represented by a vector of independent points forcing critical articulators constrictions were not direct since it can lead to very unrealistic shapes [25].

On a second approach, we trained an autoencoder to reduce the articulators' dimensionality, representing each contour by a set of control parameters in the autoencoder latent space [25]. The results are limited to the tongue for simplicity, but the extension to the complete vocal tract is possible. The latent space is analogous to the linear components retained by principal components analysis (PCA). Unlike PCA, an autoencoder does not guarantee the orthogonality, thus, we added a covariance minimization in the feature space to the loss function. We relied on the work done on articulatory model construction [20] to set the size of the latent space to eight. With the autoencoder acting as an implicit articulatory model, we used a similar bidirectional GRU to predict the tongue's principal components as a function of the phoneme sequence to be articulated. The final shape can be reconstructed using the autoencoder's decoder. The advantage of this method is that it allows adding phonetic constraints imposed on the articulatory variables. Taking these constraints into account has a double interest. First, it guarantees that the synthetic vocal tract shapes are compatible with the expected acoustic properties in terms of closure at the place of articulation, which is important in the perspective of articulatory synthesis. Secondly, these constraints are intended to compensate for the limitations of real-time MRI presented in Section 2.1, making it very difficult to detect contacts between the apex or tongue body and the alveolar region or palate. Figure 3 presents the results obtained for the tongue when using the encoder-decoder network from [23] (Figure 3a) and the autoencoder approach from [25] (Figure 3b).

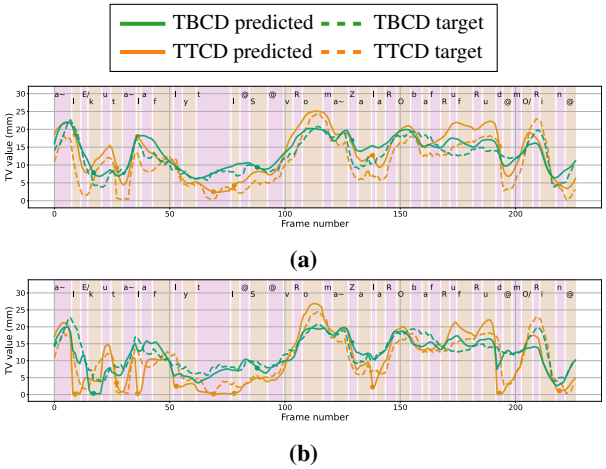


Figure 3: TBCD and TTCD trajectories for the utterance "En écoutant la flûte, le chevreau mangea la robe à froufrous de Maurine". Authorized reproduction from [25].

4. CONCLUSION

As we have shown, the development of real-time MRI and deep learning has profoundly changed the approach to coarticulation. While the number of images was for a long time limited to a few hundred and it was unimaginable to have an automatic tracking of the articulators' contour with a good accuracy, it is now possible to process very large corpora (nearly a million images and more than 2 100 sentences for a single speaker in the case of the last acquired corpus) which allows the use of automatic learning models, often inspired or derived from recurrent neural networks, which are much less constrained than those used in the past. Our past work shows how to design a deep learning-based articulatory synthesizer that allowed us to easily take into account phonetic constraints and will enable the study of articulatory compensation mechanisms.

5. ACKNOWLEDGEMENTS

This research was supported by the French ANR project Full3DTalkingHead, CPER IT2MP, Région Lorraine and FEDER.

6. REFERENCES

[1] E. Farnetani, "Labial coarticulation," in *In Coarticulation: Theory, data and techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge university press, 1999, ch. 8.

[2] S. Öhman, "Coarticulation inVCV utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, vol. 39, no. 1, pp. 151–168, 1966.

[3] ——, "Numerical model of coarticulation," *J. Acoust. Soc. Am.*, vol. 41, pp. 310–320, 1967.

[4] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. Springer-Verlag, 1993.

[5] H. N. V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Epsy-Wilson, and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from natural speech," in *11th Annual Conference of the International Speech Communication Association - INTERSPEECH 2010*, Makuhari, Chiba, Japan, 2010.

[6] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[7] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[8] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, no. S1, pp. S56–S56, 1990.

[9] V. Ribeiro, K. Isaieva, J. Leclere, R. Karpinski, J. Felblinger, P.-A. Vuissoz, and Y. Laprie, "Automatic tracking of vocal tract articulators in real-time magnetic resonance imaging," *Available at SSRN 4192628*.

[10] K. Richmond, "Preliminary inversion mapping results with a new ema corpus," in *10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009*, Brighton, 2009, pp. 2835–2838.

[11] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P.-A. Vuissoz, "Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers," *Scientific Data* , vol. 8, no. 1, p. 258, Oct. 2021.

[12] M. Uecker, S. Zhang, D. Voit, A. Karaus, K. D. Merboldt, and J. Frahm, "Real-time MRI at a resolution of 20 ms," *NMR Biomed*, vol. 23, no. 8, pp. 986–94, Oct. 2010.

[13] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[14] E. Bellon, M. Haacke, P. Coleman, D. Sacco, D. Steiger, and R. Gangarosa, "MR artifacts: A review," *AJR. American journal of roentgenology*, vol. 147, pp. 1271–81, 12 1986.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and P.-A. Vuissoz, "Automatic tongue delineation from mri images with a convolutional neural network approach," *Applied Artificial Intelligence*, vol. 34, no. 14, pp. 1115–1123, 2020.

[18] P. Birkholz and D. Jackel, "A three-dimensional model of the vocal tract for speech synthesis," in *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain*, Aug 2003, pp. 2597–2600.

[19] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2165–2180, 2001.

[20] Y. Laprie and J. Busset, "Construction and evaluation of an articulatory model of the vocal tract," in *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Spain, Aug. 2011.

[21] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated," in *Proc. Interspeech 2021*, 2021, pp. 3325–3329.

[22] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marschal, Eds. Kluwer Academic Publishers, 1990.

[23] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, vol. 141, pp. 1–13, Apr. 2022.

[24] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[25] V. Ribeiro and Y. Laprie, "Autoencoder-Based Tongue Shape Estimation During Continuous Speech," in *Proc. Interspeech 2022*, 2022, pp. 86–90.

MODIFICATION OF EEG BRAIN ACTIVITY INDUCED BY RESPIRATORY MUSCLE REGULATION DURING PHONATION

Said-Iraj Hashemi [1, 2], Mathieu Petieau [1], Guy Cheron [1], Didier Demolin [2], Ana Maria Cebolla [1]

[1]Laboratory of Neurophysiology and Movement Biomechanics (LNMB). Faculty of Motor Skills Sciences | Motricity Sciences Research Center. Interfaculty institutes | UNI, ULB Neuroscience Institute Université Libre de Bruxelles, Brussels, Belgium.
[2]Phonetics and Phonology Laboratory LPP, CNRS-UMR 7018, Sorbonne Nouvelle, Paris, France

Abstract

This paper presents the results of an experiment replicating Ladefoged's (1967) experiment of variations in subglottal pressure and respiratory muscles activity during speech. We set up a methodology in which acoustic and aerodynamic parameters, respiratory muscles, and the electroencephalogram (EEG) activity are for the first time synchronously, registered with one participant producing a suite of [pa] in one long expiration. 4 phases in the regulation of subglottal pressure were identified: (1) the elastic recoil of the inspiratory muscles, (2) the activity the internal intercostals, (3) and (4) the abdominal muscles activations. Interestingly brain oscillations showed identifiable power spectrum modulations of alpha and theta brain rhythms specifically related to the four phases, suggesting specific brain function accompanying expiration phases. The latency of these phases corroborates with the different patterns of EEG activity during the sequences of [pa] syllable production.

1. INTRODUCTION

During the production of speech, there is an increase in the average level of the subglottic pressure (Ps). This pressure returns to its baseline level after the end of the utterance. Slight variations in the average Ps can be observed during a long utterance. In his work published in the book "three areas of experimental phonetics", published in 1967[1], Peter Ladefoged highlighted the different actions of the respiratory muscles responsible for maintaining the Ps during a long utterance. After a deep inspiration, the relaxation pressure is much higher than the required pressure under the vocal folds, the inspiratory muscles relax and pressure decreases in the lungs. When the relaxation (the recoil forces) becomes low, the inspiratory muscles usually stop working and the pressure required for speech is maintained by the work of the expiratory muscles. Toward the end of a long utterance, when the volume of air in the lungs is very low, many expiratory muscles may be needed to maintain pressure at a constant level.

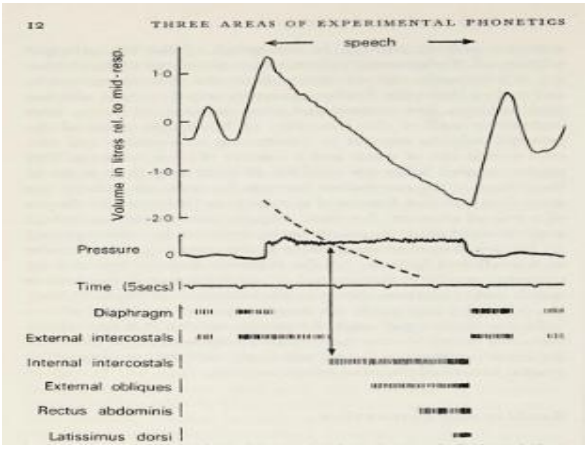


Figure 1. Ladefoged experiment, relating aerodynamic and myologic parameters during a long statement (the subject counted from 1 to 32 without taking a breath)

The lower part of figure 1 illustrates schematically the activity of the respiratory muscles to maintain a constant average pressure. The action of the external intercostals ceases when the volume of air in the lungs is slightly lower than the volume after a normal inspiration. From this point on, expiratory activity is necessary to maintain the pressure under the vocal folds and, therefore, the internal intercostals come into action with progressively increasing intensity. The action of the internal intercostals is complemented by various other muscles, such as the external obliques, rectus abdominis and dorsalis major. These muscle actions could be modulated by the oscillatory activity of the neocortex.

EEG oscillations are direct global electrical brain activity underlying brain function [2] EEG dynamics can be characterized by time-frequency domain measures such as the power spectrum [3] and phase variations of EEG oscillations [4]. For example, in a relaxed awake state, when the eyes are closed, the power of the EEG alpha rhythm (8-13 Hz) increases in the occipital and parietal scalp areas [5]. The increase in alpha power is related to the maintenance of neural network coherence in the absence of visual and sensorimotor information and is considered a marker of cortical inactivity or an indication of active inhibition of sensory information [6]. The increase in the power spectrum (ERS) of theta oscillation in the motor area plays an important function in motor control. In addition, theta oscillation is thought to be phase-locked with movement onset, and theta power is correlated with movement acceleration [7].

The role of brain activity in controlling the mechanisms responsible for phonation remains unknown. Among the rares studies, Lizarazu and all were interested in the adaptation of brain oscillations

to speech rate. Assuming that the delta (<4 Hz) and theta (4-8 Hz) frequency bands drive prosodic and syllabic speech rates respectively. And that the theta band modulates high-frequency neural oscillations in the gamma band (25-40 Hz), which are thought to be crucial for phoneme processing in natural speech, they asked subjects to deliver speech at slowed, normal, and accelerated rates. Their results show that the phase of theta band oscillations in auditory regions adjusts to variations in speech rate. Moreover, they have demonstrated a phase-amplitude coupling between the gamma and theta frequency bands respectively, which follows the speech rate[8].

Another team was interested in the neural mechanisms underlying the coordination and articulation of movements during speech[9]. Participants read aloud 460 sentences from the MOCHA-TIMIT[10] database. The authors focused primarily on the sensorimotor cortex and inferred the articulator movements of the speakers from the acoustics of the speech produced. The main results of their work show that articulatory kinematic trajectories reveal coordinated movements of the tongue, lips, jaw, and larynx and are encoded by the sensorimotor cortex. Finally, we can mention Galgano and colleagues[11] who studied brain generators of voice-related cortical potentials (VRCPs). Participants held their breath for 4 seconds and, after the appearance of a visual signal, they emitted the sound [m] while keeping their mouths closed or exhaled through their nose without emitting a sound. Examination of the spatiotemporal changes in the VRCP -150ms before the onset of phonation and 100ms after the onset of phonation, revealed the involvement of the laryngeal motor cortex, probably responsible for the execution and continuity of phonation. Other sources were also located in the cerebellum.

To our knowledge there is no published study of cortical patterns associated with Ps regulation and the underlying muscle actions during a long utterance. According to Ladefoged's work, during the production of short utterances, Ps regulation occurs autonomously and with the help of recoil forces. However, in the case of a long utterance, this is regulated with the help of the action of the expiratory muscles. In this preliminary work, we revisited the study of Ladefoged 1967, and studied the oscillatory brain patterns underlying the different phases of phonation during expiration and their transition.

2.METHOD

2.1 Subject and protocol

This study was approved by the ethics committee of the Brugmann University Hospital (EC 2022/246). Our participant was a 63-year-old French-speaking male. We replicated the experience of Ladefoged in its "three areas of experimental phonetics" study with a modified statement. Indeed, the protocol consisted in producing the syllable [pa] as long as possible without taking a breath. The participant sat in a chair, took a normal breath, and began to pronounce the syllable [pa] at a self-regulated rate. We recorded 15 sequences of [pa] syllable production.

2.2 Materials

Acoustic parameters were recorded using the Assisted Vocal Assessment System (EVA2) with a sampling frequency of 4400 Hz. We determined the Ps from the intraoral pressure, we also measured the vibration of the vocal folds.
Respiratory muscle activity was recorded with bipolar Delsys EMG TrignoTM surface electrodes on the scalene, external, and internal intercostal, external oblique, and rectus abdominis muscles, with a sampling frequency of 2000 Hz.
EEG signals during [pa] sequences were recorded using the ANT system (sampling rate 2048 Hz), (Netherlands) Neuro with a 128-channel EEG cap comfortably fitted to the participant's head. Electrodes located on both mastoids served as EEG reference. All recording systems were synchronized with each other using an external signal generator sending triggers at the beginning and end of each recording.

3. AERODYNAMIC, EMG, EEG ANALYSIS

3.1 Aerodynamic and EMG analysis

We used the Winpitch LTL software (Martin 2004) to analyze the acoustic parameters of the [pa] syllable production sequences. A visual analysis of the evolution of the estimated Ps allowed us to highlight slight fluctuations during the sequence [pa]. These fluctuations are concomitant with the action of the respiratory muscles. As shown in figure 2, the successive actions of the muscles constitute the different phases of the regulation of the Ps. The elastic recoil of the diaphragm and external intercostal, constitutes phase 1 (P0 and P1). The internal intercostal action constitutes phase 2 (P2). The external oblique action is associated with phase 3 (P3). And finally, the action of the rectus abdominis (P4), at the end of the sequence [pa], constitutes the

last phase of regulation of the Ps.

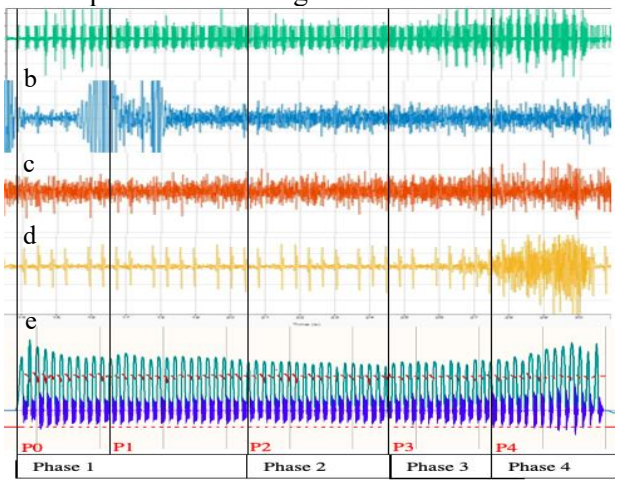


Figure 2 Variation of myologic and aerodynamic parameters during a [pa] sequence. a: Scalene; b: external intercostal; c: external oblique; d: rectus abdominis; e: green curve estimated of the measured Ps; blue curve: vocal folds vibration; red lines: fundamental frequency

3.2 EEG analysis

EEG data was processed using The Math Works, lnc. MATLAB version 2022a and an open-source toolbox EEGLAB v2022.1, Delorme A & Makeig S (2004). After a visual identification and removal of noisy channels, we filtered the data with a 0.5 Hz high-pass filter. The line noise was removed using Zapline plugin [12]. We then performed an independent component analysis (ICA). Components associated with eye movements, heartbeat, channel noise and muscle activity were identified by visual analysis and using ICLabel plugin and were rejected. Only components with at least a 10% probability of containing brain activity were retained. The events related to every phase latency were imported. We spliced the data according to the [pa] sequences and performed time-frequency analyses in the frequency band 1 to 30 Hz on channels F3, C3 and P3 (frontal, central and parietal sites situated on the left at the scalp level). The average duration of the [pa] sequences was 20± 2s. The baseline was therefore defined 2 seconds before the sequences.

4. RESULTS

4.1 Descriptive statistics

The average duration of P1 is 7.54 ± 2.2 s (composed of P0 3.97 ± 1.26s and P1 3.57 ± 0.94), P2 3.94 ± 1.33s, P3 4.11± 1.40s and P4 with 3.64 ± 0.84s. Regarding the duration of scalene muscle contraction during the production of the syllable [pa], a progressive increase was observed. With the minimum time in P1 0.105 ± 0.030ms and the maximum time in P4 0.149 ± 0;054ms.

4.3 Time-frequency analysis

For time-frequency analysis of EEG oscillations, we calculated the baseline normalized spectrogram or ERSP (2s before the P0 event) [3,4]. Event-related spectral perturbation (ERSP) measures variations in the power spectrum of ongoing rhythms at specific time periods and frequency ranges. In ERSP measurements, ERD (event related desynchronization) indicates a reduction in the power spectrum while ERS (event related synchronization) indicates an increase in the power spectrum. ERD/ERS are interpreted as reflecting the reactivity of the brain and are related to specific aspects of information processing in the brain. We also calculated the ITC [13] which measures the phase coherence between ongoing EEG signals at specific time periods and frequency ranges. For the significance level of ERSP and ITC, a bootstrap resampling ($p < 0.05$) was used as a surrogate method.

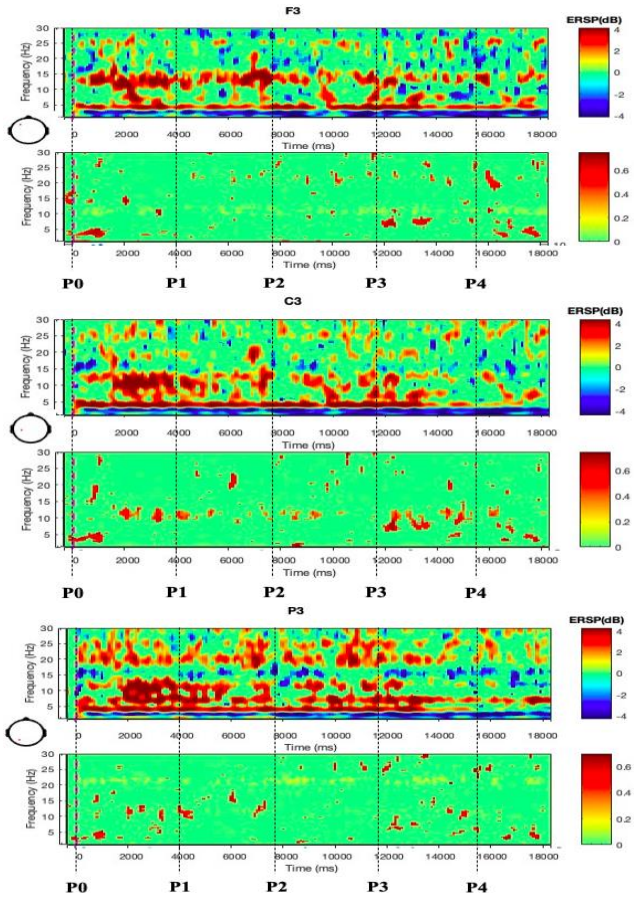


Figure 3 Time-frequency analysis of time-locked [pa] sequences.

Figure 3 shows the average ERSP of time-locked [pa] sequences. The average ERSP spectrogram of F3 shows a clear increase in the power spectrum (event-related synchronization, ERS) at the maxima of 13 Hz and at 7 Hz (ranging from about 4 to 14 Hz) from P0 to P1. Then a large ERS in the 4 Hz to 14 Hz frequency band at 2500ms after P0 and at the end of P1 and P2 is observed. A decrease in spectral

power (ERD) centered at 7 Hz is observed at the end of P3, being followed by an ERS centered at 13 Hz at the beginning of P4. The ERSP on electrode C3, shows similar dynamics as F3, with the particularity of a lower theta ERD at the end of P3 and a lower Alpha ERS at the beginning of P4.

The ERSP of the P3 electrode shows a continuous increase in spectral power (ERS) at 4 Hz until the end of phase 3. And a discontinuous increase in spectral power at 7 Hz at 1800ms until the end of the sequence. We note a ERS centered at 15 Hz in a jerky manner at 0ms and around the latency of each phase. We observe an absence of alpha spectral power from 12600ms, and an ERD in the alpha band at the end of the sequence. Interestingly, variations of theta oscillations accompany variations of alpha oscillations, suggesting the constitution of a unified theta-alpha pattern.

The ITC patterns of F3, C3, and P3 (Figure 3) showed a phase synchronization cluster centered at 4 Hz during the first 1000ms of the sequence. There is a 7-Hz synchronization cluster at 12000 and 13000ms and a 4-Hz synchronization cluster at 17000 and 18000ms in the F3, C3, and P3 electrodes. Finally, we observe a 12 Hz synchronization cluster at 2000 and 3000ms in electrodes C3 and P3.

4.4 Topographical analysis of the ERSP

The topography (128 EEG electrodes) of the ERSP (Figure 4) in the theta frequency band (4-7 Hz) shows an ERS in the central, parietal, and occipital regions at P0. The increase in ERS is centered in the parietal and occipital regions at the end of phase 1. We then observe an increase in ERS, centered in the left lateral and parietal regions, before and during P3 and at the beginning of phase 4. At the end of the sequence, we observe ERS, only in the parietal region at 16s. At 17s, we observe generalized ERS only on the left side of the scalp. Theta spectral power continues to decrease from 18s.

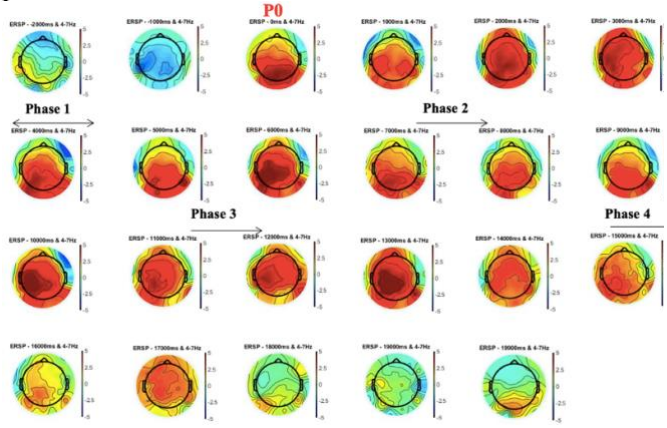


Figure 4 Topography analysis of ERSP sequences of [pa] locked in time. In the Theta frequency band

The topography of the ERSP (Figure 5) in the Alpha frequency band (8 to 13 Hz) shows a localized ERS in the left lateral region at the beginning of phase 1, at 1000ms. We then observe an increase in ERS in all regions of the scalp. At the end of phase 1, we again see a localized ERS increase in the left lateral region. At the intersection of phases 2 and 3, we again observe an increase in spectral power in the left lateral region. Interestingly, the ERSP topography in the theta band showed an ERS in the same region and at the same time. At the beginning of phase 4, there is an ERS in the occipital region and ERD at the end of this phase (at 19s) throughout the scalp and especially in the occipital region.

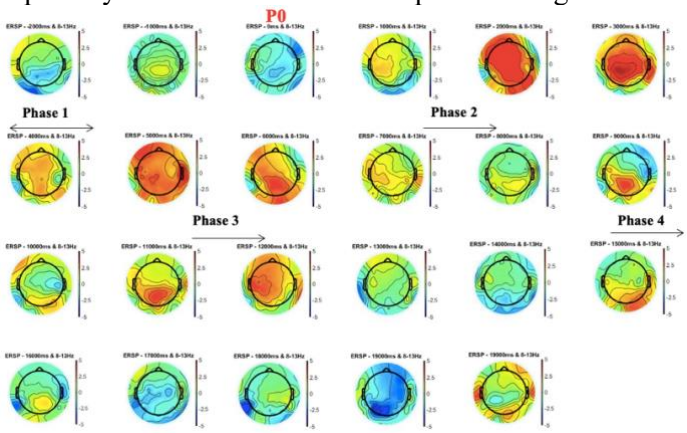


Figure 5 Topography analysis of ERSP sequences of [pa] locked in time. In the Alpha frequency band

5.DISCUSSION

We replicated Ladefoged's (1967) experiment on variations in Ps and respiratory muscle activity during speech. By calculating the ERSP in the time-frequency domain (figure 3) and the topography of the ERSP in the theta and alpha frequency bands (figure 4 and 5), we showed that there are patterns of variation in the spectral power (ERS and ERD) of these frequency bands that underly the phases of the regulation of the Ps. Moreover, the variation of the spectral power of the alpha frequency band is concomitant to the spectral variation of theta band, in the vicinity of the phase transitions.

The strong point of our methodology compared to the study of Louks [14] is the synchronization of the EEG , the electromyography, aerodynamic and the acoustics signals. A more in-depth study of source analysis will allow to investigate the precise regions of the neocortex areas involved in the regulation of Ps seems essential to us. Furthermore, we question the contribution of the cerebral cortex in the transition from autonomous to controlled regulation.

This study is funded by the French National Research Agency (ANR) within the framework of the full 3D talking head project.

6. REFERENCES

[1] Ladefoged P. Three areas of experimental phonetics : stress and respriatory activity, the nature of vowel quality, units in the perception and production of speech. --. London : Oxford U.P.; 1967.
[2] Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. Science 2004;304:1926–9. https://doi.org/10.1126/science.1099745.
[3] Pfurtscheller G, Stancák A, Neuper Ch. Event-related synchronization (ERS) in the alpha band — an electrophysiological correlate of cortical idling: A review. International Journal of Psychophysiology 1996;24:39–46. https://doi.org/10.1016/S0167-8760(96)00066-9.
[4] Makeig S, Debener S, Onton J, Delorme A. Mining event-related brain dynamics. Trends in Cognitive Sciences 2004;8:204–10. https://doi.org/10.1016/j.tics.2004.03.008.
[5] Tudor M, Tudor L, Tudor KI. [Hans Berger (1873-1941)--the history of electroencephalography]. Acta Med Croatica 2005;59:307–13.
[6] Pfurtscheller G, Neuper C, Mohl W. Event-related desynchronization (ERD) during visual processing. International Journal of Psychophysiology 1994;16:147–53. https://doi.org/10.1016/0167-8760(89)90041-X.
[7] Ofori E, Coombes SA, Vaillancourt DE. 3D Cortical electrophysiology of ballistic upper limb movement in humans. NeuroImage 2015;115:30–41. https://doi.org/10.1016/j.neuroimage.2015.04.043.
[8] Lizarazu M, Lallier M, Molinaro N. Phase−amplitude coupling between theta and gamma oscillations adapts to speech rate. Annals of the New York Academy of Sciences 2019;1453:140–52. https://doi.org/10.1111/nyas.14099.
[9] Chartier J, Anumanchipalli GK, Johnson K, Chang EF. Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. Neuron 2018;98:1042-1054.e4. https://doi.org/10.1016/j.neuron.2018.04.031.
[10] MOCHA-TIMIT n.d. https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (accessed April 28, 2023).
[11] Galgano J, Froud K. Evidence of the voice-related cortical potential: an electroencephalographic study. Neuroimage 2008;41:1313–23. https://doi.org/10.1016/j.neuroimage.2008.03.019.
[12] Klug M, Kloosterman NA. Zapline-plus: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG. Hum Brain Mapp 2022;43:2743–58. https://doi.org/10.1002/hbm.25832.
[13] Tallon-Baudry C, Bertrand O, Delpuech C, Pernier J. Stimulus Specificity of Phase-Locked and Non-Phase-Locked 40 Hz Visual Responses in Human. J Neurosci 1996;16:4240–9. https://doi.org/10.1523/JNEUROSCI.16-13-04240.1996.
[14] Loucks TMJ, Poletto CJ, Simonyan K, Reynolds CL, Ludlow CL. Human brain activation during phonation and exhalation: Common volitional control for two upper airway functions. NeuroImage 2007;36:131–43. https://doi.org/10.1016/j.neuroimage.2007.01.049.

SIMILAR AND DIFFERENT TONGUE SURFACE CONTOURS: INTRA-SPEAKER CONTROLS IN ULTRASOUND ANALYSIS

James M Scobbie

Queen Margaret University, Edinburgh, Scotland
jscobbie@qmu.ac.uk

ABSTRACT

Ultrasound studies of speech production analyse differences in dependent variables reflecting the tongue surface's location and shape. Inferential statistics distinguish theoretically-relevant from random effects, somewhat independently of the descriptive size of significant effects. Experimental designs induce measurable dependent changes by manipulating independent variables such as prosody, phonemic target, etc.

This paper presents descriptive statistics quantifying holistically all 15 pairwise differences between six monophthongal long vowel phonemes of one variety of English, comparing these to experimental noise differences attributable to the use of two identical blocks of data collection in sequence. Eight speakers were recorded, using two different ultrasound systems, and analysed in AAA using both edge-tracking and DeepLabCut pose estimation. The smallest phonemic contrast (~2mm) was greater than the experimental noise (~1mm), and was well-evidenced by AAA's t-test of radial difference.

Keywords: methods, ultrasound, articulation, experimental noise.

1. INTRODUCTION

Ultrasound scanning is a safe, accessible, non-intrusive and quick way of imaging the tongue during speech. Ultrasound (US) images are analysed manually or automatically to extract key features relevant to phonetic analysis, typically to derive a mid-sagittal tongue-surface contour. Experimental protocols enable analyses of the properties of these surface contours. In particular, comparisons of two or more groups of contours may, if the between-groups contours can be shown to be different, provide evidence for rejecting a null hypothesis.

There are a number of techniques for comparing two or more groups of tongue-surface contours sampled at single time points (or averaged from multiple time-points). These can be extended to the analysis of dynamic phenomena using time-series, but the focus here will be static configurations. In the analysis of untransformed surface contours, various statistical techniques can determine whether two groups of tongue surface contours are different or not. Popular approaches include SSANOVA e.g. [1] or [2], GAMMs e.g. [3] or [4], or the t-test difference function built into the data capture and analysis software package AAA e.g. [5], [6]. Global properties of contours (topological properties) can also be used, as well as location.

If a localised difference between two groups of contours reaches statistical significance, this does not imply that all aspects of the contours differ. So long as the confidence intervals flanking two mean contours are non-overlapping (for example), a "zone of difference" can be said to have been found. But how small can these zones be? While a threshold for confidence intervals at 95% is widely adopted, there is no general agreement on the minimum length of non-overlap needed to define such a zone of difference. One specific proposal [7] was to classify as significant only zones spanning at least six of AAA's 42 radii in a 135° field of view. This would be about 3cm of tongue surface, and appears rather an arbitrary threshold.

The distance between contours also matters, whether it be a single maximum value, averaged over a zone of statistical difference, or characteristic of some sub-part of the contour [7], [8]. Small distances can be critical in speech production, but these descriptive measures also inform statistical testing. Can a distance between two similar contours of just 1mm along 1cm of contour length be meaningful? Yes, if it is due to the speaker distinguishing between complete closure and close approximation. Probably not, if the difference is between two vowels. What if a mere 1mm difference was found in a midsagittal zone 5cm long?

When it comes to analysing the possible difference between two similar tongue contours, we should provide basic descriptive information, statistically-significant *or not*. We need to ask:

- How far apart are the contours (in a zone of interest)?
- What is the length of any zone of significant difference?

For such information to benefit future statistical modelling, we need descriptive figures on meaningless variation: i.e. *control* data. Experimental noise arises from issues such as instrumental accuracy and

precision, analytic variability, and speaker variation due to (task) effects such as fatigue, confidence, or implicit learning. These can be somewhat addressed by counter-balanced experimental designs, but not always, particularly when a baseline is required.

For concreteness, we have constructed a control study for a specific experiment [9], but the real goal is expository. We aim to exemplify the value of *a priori* meaningless (control) and meaningful descriptive findings, as a general principle. We will quantify: (a) a meaningless task non-difference between identical datasets presented in two blocks; (b) axiomatic contrasts between six vowel phonemes via holistic pairwise comparison; (c) two different ultrasound systems; and (d) two descriptive measures of difference, namely average radial distance difference and mean nearest neighbour distance difference.

2. METHOD

2.1. Protocol

Our protocol is intended to be compatible with a wide range of experimental studies and designs. It was based on studies in which several phonemes are elicited from a single speaker in two distinct prosodies, frames, languages, or affects, with a short rest between blocks. A specific model, [9], was chosen, so that this study is directly a control for [9]; but it also provides norms for Scottish English.

We used speaker-specific randomisations of our materials (see 2.2). There were no fillers or distractors. The materials were 12 words (or pseudowords) that were repeated three times in a block. Each block of 36 randomised words was presented to the speaker twice, with a short pause between blocks (one to two minutes). It was used as a brief rest, for small-talk, and to reset the computer. Speakers drank a little water, looked around, talked to the experimenter, and moved in their chair. They were aware this was the mid-point of the experiment.

2.2. Materials and speakers

The Scottish English materials were tightly structured, following [9]. In Scottish English, there are six "unchecked" or long monophthongal vowels /ieaɔoʉ/ with lexical set incidence: /i/ FLEECE, /e/ FACE, /a/ TRAP & BATH & PALM, /ɔ/ LOT & THOUGHT, /o/ GOAT, /ʉ/ GOOSE & FOOT. /ɔoʉ/ are phonologically rounded. Two C_C contexts were used, in which both C were labial (either /m/ or /p/) so as not to affect lingual behaviour. Thus the wordlist mostly included real words but also pseudowords. There was no carrier phrase.

- peep, pape, pap, pop, pope, poop
- meme, maim, mam, mom, moam, moom

The speakers were judged to have typical Scottish accents. Speakers c1-c4 were undergraduate students, some familiar with phonetics. Speaker c5 was the author, c6 was a colleague of the author, and c7 & c8 were family members of the author. Only c5 was fully aware of the purpose of the study.

2.3. Radial measurement

A typical vowel system for Scottish English is illustrated in Fig. 1, showing mean tongue surface contours labelled for the six vowels. Each mean was calculated from the radius crossing points of each of the 12 tokens, within AAA's fan-shaped analysis grid (not shown). Means are flanked by ± 1 standard deviation (along radii). Two key radii are shown, however: fanlines fn13 and fn33 are the anterior and posterior limits respectively for a common analysis sector of this speaker (c8), containing all contiguous radii in which edge-detection confidence in AAA was above a threshold of 80% for all vowels. For each of the 21 radii within this sector the 15 pairwise differences between each of the six vowels was measured (N=12 for each vowel). The *average* radial difference in distance from the origin is a holistic measure, irrespective of zones of statistical differences between vowels, but trimmed as described.

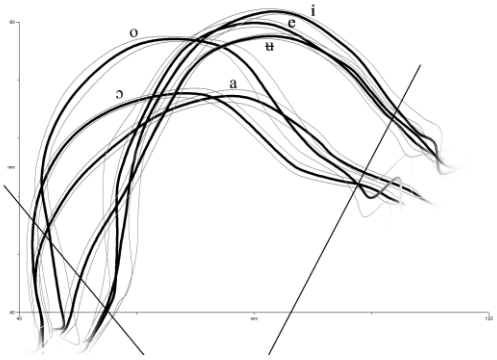


Figure 1: Typical Scottish English vowel space (c8), anterior to right, with a common analysis sector, showing the overall similarity of /i/, /e/ and rounded /ʉ/

Vowel-specific sectors (trimming off all surface contours <80% confidence) were used for within-vowel radial t-testing for Block 1 vs. Block 2 (N=6 for each vowel in each block). The number of contiguous radial fanlines with significant differences was noted.

2.4. Ultrasound system 1

The system used for speakers c1-c4 was similar to that used in [9],[10], comprising an Ultrasonix RP scanner, with an aluminium stabilisation headset [11], and a 5MHz 10mm radius Microconvex probe. AAA software [5] (version 2.19) was set to a 135°

field of view, at 121 frames per second, depth 80mm, with 63 scanlines. Tongue surface contours were created semi-automatically using AAA's edge-tracking procedures, on a fan-shaped grid (as above).

2.5. Ultrasound system 2

The system comprised a Telemed micro system scanner, and an aluminium stabilisation headset [11], with a 2-4MHz 20mm radius convex probe. AAA software [6] (version 220) was set to a 101° field of view, at 81 frames per second, depth 80mm, with 64 scanlines. It was used for speakers c5-c8.

AAA's innovative pose estimation [6], [12] using DeepLabCut (DLC) automatically created tongue surface contours, and AAA's nearest neighbour (NN) function was used to calculate both B1 vs. B2 differences and the pairwise vowel phoneme differences. We then converted the DLC contours to fan-based contours, enabling comparative radial differences to be calculated within AAA.

3. RESULTS, SYSTEM 1

The size of the sectors analysed (in fanlines) was minimally 12 and maximally 24 radii (mean 21, median 23). The locations of the analysis sectors varied, with the most retracted vowels /ɔ/ and /o/ also tending to have the smallest analysis sectors. Typically, the mean radial difference between blocks was below 1mm (Table 1). Only one vowel from one speaker (/i/, from c3) reached a threshold for having a zone of 6 contiguous radii where there was a significant difference in the radial distances.

Table 1: Size of noise (B1 vs. B2) as mean radial difference in distance from fan origin (mm). An asterisk indicates a zone of significant difference.

	c1	c2	c3	c4
i	0.7	0.9	*1.5	0.5
e	0.6	1.0	1.0	0.5
a	0.2	0.6	0.8	0.3
ɔ	0.3	0.7	0.9	0.4
o	1.1	0.9	0.4	1.1
ʉ	0.4	0.4	0.9	0.7

Speaker c3's /i/ vowel appeared to be higher and more anterior in the second block (Fig. 2). In Fig. 2, 12 thick radial lines projecting inwards from the arc indicate both the location of these radii with significant differences (p-value from 0.001 to 0.01) and their sizes (from 0.91mm, fn17, to 2.49mm, fn8). The most anterior radius used in the analysis was fn6 (the rightmost shown). The mean radial distance between these vowels in the zone of significance was 1.83mm, over a midsagittal tongue surface length of 42mm. The total length of confidently-traced tongue surface analysed, spanning from fn6 to fn29 (for c3's /i/), was 77mm. The average radial difference overall was just 1.5mm (Table 1).

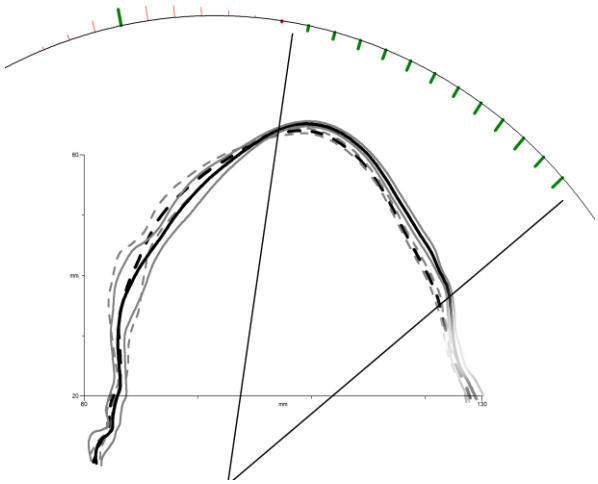


Figure 2: Two significantly different mean-radial tongue curves (speaker c3, /i/), each flanked by ± 1 standard deviation. Dashed /i/ from block 1, solid /i/ from block 2.

Table 2: Pairwise mean radial differences between phonemes (mm). All pairs had some zone of significance (length and width not reported).

	c1	c2	c3	c4	mean
ie	1.4	2.1	2.1	1.2	1.7
ia	12.2	13.2	10.6	11.4	11.8
iɔ	15.4	17.6	13.9	14.3	15.3
io	14.8	16.3	10.8	12.7	13.7
iʉ	5.1	7.3	4.7	3.6	5.2
ea	11.0	11.8	8.4	10.2	10.3
eɔ	14.1	16.2	11.7	13.1	13.8
eo	13.4	14.8	8.6	11.5	12.1
eʉ	3.8	5.6	2.6	2.4	3.6
aɔ	3.2	4.5	3.7	3.4	3.7
ao	4.1	4.5	4.0	5.3	4.5
aʉ	7.3	6.5	6.0	8.2	7.0
ɔo	2.8	2.8	4.0	3.4	3.3
ɔʉ	10.3	10.9	9.4	11.3	10.5
oʉ	9.7	9.3	7.2	10.0	9.0

As well as lacking statistical significance, most of the noise differences reported in Table 1 are tiny in absolute terms. Yet some are close to the smallest pairwise phoneme differences in Table 2, e.g. /i/ vs. /e/. Fronted /ʉ/ (GOOSE & FOOT) has a lingual shape and location very similar to /e/, while /ɔ/ was similar to /o/. All the pairwise differences had zones of significant difference, and though not reported

here, they were large (cf. Fig. 1). The maximum radial pairwise differences and mean difference just within zones of significance were greater than the overall mean pairwise differences in Table 2.

4. RESULTS, SYSTEM 2

The DLC pose estimation surface contours showed greater differences when quantified with AAA's nearest neighbour tool (Table 3 vs. Table 1).

Table 3: Size of noise (B1 vs. B2) as DLC mean nearest neighbour distance (mm).

	c5	c6	c7	c8
i	1.5	0.8	2.3	1.8
e	1.4	1.0	2.0	1.9
a	1.5	1.6	2.2	1.2
ɔ	2.1	2.4	1.7	1.3
o	2.0	1.7	2.6	1.9
ʉ	3.0	1.5	2.5	1.5

For a meaningful comparison, DLC tongue surface contours were therefore converted to contours on a fan grid. Significant difference could be tested. More zones of significance were found (Table 4) for inter-block noise for c5-c8 using system 2) than for c1-c4 using system 1.

Table 4: Size of noise (B1 vs. B2) as mean radial difference in distance from fan origin (mm).

	c5	c6	c7	c8
i	1.4	0.4	1.4	1.0
e	1.1	0.5	1.5	1.7
a	1.1	1.3	*1.9	*0.7
ɔ	*1.4	0.9	1.1	*1.0
o	1.3	1.1	1.4	*1.6
ʉ	2.6	0.9	1.1	1.0

Significant experimental noise was seen in a zone spanning 12 contiguous radii for c7's /a/ (mean 3mm difference). There were two zones of 6 radii for c5's /ɔ/ (together, 2mm). Significant zones for c8 spanned only 6 contiguous radii, and averaged just 1.2mm for /ɔ/, 1.8mm for /a/ and 2.1mm for /o/.

Pairwise matrix analysis (Tables 5 & 6) showed a similar rank order for radial and nearest neighbour measures. As seen in Table 2, the varied sizes of mid-sagittal contrasts are reflected intuitively well.

5. DISCUSSION AND CONCLUSION

Experimental (task) noise was occasionally statistically significant, and any inter-block differences could have been due to changes in speech production or instrumental instability. Combining all results, the small but significant phoneme radial difference of 1.7mm (/i/ vs. /e/) appears appreciably more than the rarely-significant noise in /i/ (0.9mm) and /e/ (1.0mm). This suggests small yet significant results in experiments lacking counter-balancing (e.g. Table 3 in [9]) should be treated with caution, given our current levels of statistical understanding. More control studies, involving a wide range of experimental protocols and systems, will provide a better context for inductive statistical analysis.

Table 5: Pairwise nearest neighbour differences between DLC curves (mm).

	c5	c6	c7	c8	mean
ie	1.8	2.6	2.1	2.3	2.2
ia	10.7	6.5	11.8	10.7	9.9
iɔ	16.3	13.4	13.7	13.5	14.2
io	16.6	12.5	11.3	12.4	13.2
iʉ	8.1	6.2	6.3	2.8	5.8
ea	9.5	4.7	11.1	9.9	8.8
eɔ	15.3	11.8	12.9	12.7	13.2
eo	15.6	10.4	10.5	11.0	11.9
eʉ	6.9	3.8	5.6	2.6	4.7
aɔ	6.5	7.6	3.2	3.7	5.3
ao	7.2	6.5	5.4	5.5	6.1
aʉ	4.8	3.7	6.8	8.8	6.0
ɔo	3.6	6.2	4.4	4.4	4.7
ɔʉ	9.8	9.2	8.4	11.9	9.8
oʉ	9.3	7.0	5.7	11.1	8.3

Table 6: Pairwise mean radial differences between phonemes (mm).

	c5	c6	c7	c8	mean
ie	1.2	2.5	1.1	2.1	1.7
ia	15.1	8.8	12.4	13.1	12.4
iɔ	22.3	17.3	15.3	17.1	18.0
io	22.8	16.3	13.5	16.8	17.4
iʉ	11.7	6.7	7.0	3.6	7.2
ea	14.0	7.2	11.5	12.3	11.3
eɔ	21.1	15.9	14.3	16.1	16.9
eo	21.6	14.0	12.5	15.2	15.8
eʉ	10.5	4.3	5.9	3.5	6.0
aɔ	7.4	8.7	3.4	4.6	6.0
ao	9.1	7.8	5.9	7.3	7.5
aʉ	6.4	4.9	7.1	9.8	7.1
ɔo	3.6	7.5	4.3	5.7	5.3
ɔʉ	11.7	12.8	8.9	14.1	11.9
oʉ	11.2	9.8	6.6	14.4	10.5

5. ACKNOWLEDGEMENTS

For funding support, thanks to ERC Award 101019847 “PlanArt: Planning the Articulation of Spoken Utterances” and Leverhulme Research Fellowship RF-2021-368 “Case studies in ultrasound tongue imaging”.

6. REFERENCES

[1] Davidson, L. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *J. Acoust. Soc. Am.* 120(1), 407–415.

[2] Mielke, J. 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *J. Acoust. Soc. Am.* 137(5), 2858–2869.

[3] Heyne, M. 2016. *The influence of First Language on playing brass instruments: An ultrasound study of Tongan and New Zealand trombonists.* PhD Thesis, University of Canterbury. New Zealand.

[4] Strycharczuk, P., Derrick, D., Shaw, J. 2020. Locating de-lateralization in the pathway of sound changes affecting coda /l/. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1), 21, 1–27.

[5] Articulate Instruments Ltd. 2012. *Articulate Assistant Advanced user guide: version 2.14.* Edinburgh, UK.

[6] Articulate Instruments Ltd. 2022. *Articulate Assistant Advanced, version 220.* http://www.articulateinstruments.com/downloads/

[7] Cleland, J. Scobbie, J.M., Wrench, A.A. 2015. Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics* 29(8–10), 575–597.

[8] Cleland, J. Scobbie, J.M. 2020. The dorsal differentiation of velar from alveolar stops in typically developing children and children with persistent velar fronting. *Journal of Speech, Language, and Hearing Research* 64(6S), 2347–2362.

[9] Scobbie, J.M., Ma, J. 2019. Say again? Individual articulatory strategies for producing a clearly-spoken minimal pair wordlist. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *19th International Congress of Phonetic Sciences, ICPhS 19, Melbourne,* Australia, August 5-9, 2019, *Proceedings.* Australasian Speech Science and Technology Association Inc., 358–362.

[10] Wrench, A.A., Scobbie, J.M. 2016. Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016). *CASL Working Papers*, WP-24. Queen Margaret University.

[11] Articulate Instruments Ltd. 2008. *Ultrasound Stabilisation Headset Users Manual.* Edinburgh, UK.

[12] Wrench, A., Balch-Tomes, J. 2022. Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut. *Sensors* 22(3), 1133.

How speech rate, syllabic complexity and diversity affect the emergence of speech rhythm in speeded syllable repetition

Leonardo Lancia[1], Jinyu Li[2], Cécile Fougeron[3]
[1]Laboratoire Parole et Langage (LPL), [2,3]Laboratoire de Phonétique et Phonologie (LPP)
[1]leonardo.lancia@sorbonne-nouvelle.fr, [2]jinyu.li@sorbonne.nouvelle.fr, [3]cecile.fougeron@sorbonne-nouvelle.fr

ABSTRACT

Speech rhythm has been related to the grouping of nearby syllables, which for several authors depends on features of the language's phonology. However, most evidence of stable relations between rhythmic features and structural features relies on perceptual judgments, which are affected by non-phonological factors like speech rate. We studied how speaker-specific performances and syllable-complexity affect the rhythmic structure emerging in the repetition of the same syllable or of different syllables at maximal speed for 177 French speakers. We found that the degree of coordination between the production of syllables and the production of word-level prominence, determining syllables grouping in a way that reflects French metrical structure, depends on syllabic complexity and on the presence of different syllables in the same sequence. Moreover, speakers reaching higher rates coordinate better syllables and supra-syllabic prominence, supporting the idea that speech rhythm emerges from the interaction between structural features and general coordination principles.

Keywords: Speech rhythm, amplitude modulation, phase coordination, diadochokinetic task.

1. INTRODUCTION

Although there is a lack of agreement on the way to define and characterize speech rhythm, it is unequivocally related to the grouping of phonetic events into larger units. More specifically several authors (see [1 - 4], for early proposals) suggest that speech rhythm is mainly characterized by the grouping of syllables into word-sized structures (e.g. prosodic words, [4]). In a few studies, this theoretical stance led to characterize speech rhythm by measuring the degree of coordination between syllabic activity and activity related to the production of suprasyllabic prominence at the level of the word, as captured by band-pass filtering the amplitude modulation signal at the appropriate frequencies [5, 6]. In this way, it was possible to quantify the similarity between the rhythmic patterns observed in different languages and to reproduce the patterns of cross linguistic similarity that were obtained via human perceptual judgments [6].

In the present work, we adopt this approach to study the rhythmic features emerging in speeded syllable repetition tasks. This choice is motivated by the observation that, when speakers repeatedly produce a sequence of speech sounds at a fast speech rate, we often observe changes in the way speakers coordinate in time consecutive phonetic events both inside a syllable [7, 8] and between consecutive syllables [9, 10]. These results suggest that speech rate pushes the sensorimotor system to explore coordinative strategies underlying different groupings of the speech stream by jumping from one possible coordinative strategy to the other in function of their stability (i.e. unstable coordinative strategies are unlikely to characterize the speaker's behaviour during long time intervals). Speeded repetition paradigms thus provide a powerful experimental tool permitting to study how the relative stability of different groupings of consecutive speech sounds is affected by the interplay between controllable factors. The aim of this study is to investigate how a number of factors that are known to change between rhythmically dissimilar languages affect the degree of coordination between syllables and suprasyllabic prominence in French (in which suprasyllabic prominence depends on accents production). The factors considered are the type of syllables to be repeated, in terms of complexity (CV vs. CCV) and diversity (same syllable vs. different syllables in sequence) and the speech rate reached by the speaker in the task.

2. MATERIAL AND METHODS

The productions of 177 French speakers (from France and Switzerland) in a diadochokinesic (DDK) task were extracted from the MonPaGe_HA database [11]. Speakers were instructed to repeat as fast and as accurately as possible sequences of syllables in a continuous manner in a single breath group. Sequences to be repeated varied in complexity and diversity: (a) repetitive CV syllables involving different places of articulation: /bababa/, /dedede/, /gogogo/, (b) repetitive CCV syllables: /klaklakla/, /tʁatʁatʁa/, (c) sequences of different CV syllables: /badego/.

2.1. Parameterization of the filter's coefficients

To extract the acoustic amplitude modulation (henceforth AM(k) with k belonging to {1,…,K} and K being its length in number of samples) from each sequence of uninterrupted speech, we submit the acoustic signal to the Hilbert transform and compute its amplitude. By band-pass filtering the AM signal we capture its syllabic and supra-syllabic components (henceforth respectively referred to as *syllAM*(k) and *accAM*(k), see Figure 2). Usually, the filter cut-off frequencies are set at values coherent with oscillatory frequencies of ca. 5Hz (for *syllAM*) and 2Hz (for *accAM*) based on empirical observations (e.g., [12]) of cross-linguistic data and on theoretical considerations (c.f. [13]). However, such canonical cut-off frequencies reflect average tendencies and are not appropriate to model deviant behaviours such as that observed in diadochokinetic tasks (in which speakers easily reach speech rates higher than 7hz). Therefore, we adapted the cut-off frequencies to the analysed data. To do that, for each experimental trial, we set the three required cut-off frequencies at the midpoints between the following rates: segments rate, syllables rate, accents rate and the rate of the fastest rhythmic component slower than the accentual one (capturing, for example, modulation of activity due to the production of sentence-level features).

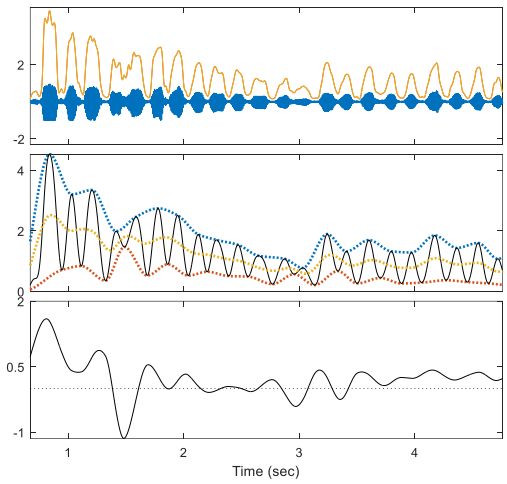


Figure 1: Derivation of accent rate. Topmost panel: waveform and its acoustic amplitude (red). Middle panel, continuous line: *lowPassAM*; dotted lines: maximum minimum and mean envelopes. Bottommost panel: syllabic cycles' extension as derived from *lowPassAM*'s envelopes.

Thanks to a manual segmentation of the syllabic boundaries available for our data, we could readily estimate the syllable and segment rates (via the number of surface segments per syllable). The procedure adopted to determine the accent rate is illustrated in Figure 1. Firstly, we low-pass filter the amplitude modulation signal with a cut-off frequency located halfway between the syllable and segment rates. This operation produces a signal expected to contain both the syllabic and suprasyllabic components of the acoustic energy signal (henceforth *lowPassAM,* see continuous line of the middle panel in Fig. 1). Our modelling strategy is motivated by the observation that speech activity displays a pulse-like shape and by the hypothesis that these pulses are mainly associated with syllables and modulated by one or more levels of suprasyllabic prominence.

To capture the changes of the pulses' amplitude due to suprasyllabic activity, we identify all the peaks and valleys in the *lowPassAM* signal. We then obtain the upper and lower envelope of *lowPassAM* by interpolating separately the values of peaks and valleys. The time-varying distance between these two envelopes minus their time-varying mean (bottommost panel in Fig. 1) is considered as an estimate of the syllabic cycles' extension which, according to our hypothesis, reflects suprasyllabic prominence (accents production). Therefore, we estimate the accent rate by dividing the number of peaks in the obtained signal by the duration of the considered time interval.

The rate of the fastest rhythmic component slower than the accentual component is estimated by submitting the estimated extension of the syllabic cycles to Empirical Mode Decomposition (EMD, [14]). This method returns a small number of independent oscillatory components that summed produce the original signal. The first component captures the fastest oscillation present in the input, which in the case of the signal representing the extension of the syllabic cycles (bottom panel in Fig. 1) is due to accent production. Therefore, by counting the peaks in the second component obtained through EMD, we have an estimate of the fastest possible component slower than the accent rate.

2.2. PLV as an estimate of syllabic and supra-syllabic coordination

The temporal coordination between two signals can be measured by computing the phase locking value (PLV, [15]), inversely related to the variability of the lag between the positions of the two signals in their cycles (as represented by their instantaneous phase values $\phi(k)_{syll}$ and $\phi(k)_{stress}$. Instantaneous phase signals are computed via the application of the Hilbert transform to *syllAM(k)* and *accAM(k)* signals after normalizing their cycles' amplitudes so that each cycle oscillates between -1 and 1 (to meet the requirements of the Hilbert transform [16], see the two bottommost panels of Figure 2).

These values permit computing at each analysis frame the generalized phase difference:

$$(1)\quad \Delta\Phi(k)_{m,n} = m \times \phi(k)_{syll} - n \times \phi(k)_{stress},$$

where m and n are two integers such that $m \times \omega(k)_{syll} = n \times \omega(k)_{stress}$, and $\omega(k)_{syll}$ and $\omega(k)_{stress}$ are the two frequencies of *syllAM* and *accAM* at frame k. The PLV corresponds to the first Fourier moment of the distribution of generalized phase differences observed in a given time window.

$$(2) \quad PLV = \langle \cos \Delta\Phi(k)_{m,n,} \rangle_k^2 + \langle \sin \Delta\Phi(k)_{m,n} \rangle_k^2$$

where the hooks indicate averaging over the time frames (indexed by k) belonging to the considered time window. A window width of 4 seconds is used, while the temporal resolution of the phase signals was set to 1ms. The window duration was chosen because it is the minimal duration of an uninterrupted chunk in our data base. For the same reason, we analysed only the first time-window of each recording. The values of m and n were determined separately for each analysis window by following [15], who compute one PLV for each combination of possible m and n values (where m and n can be chosen among the integer ranging from 1 to 10) and choose the combination that gives the highest PLV.

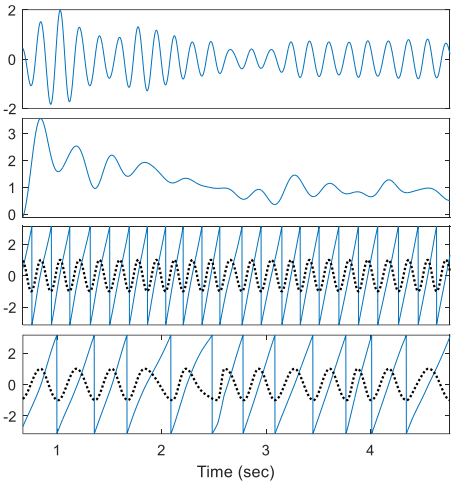


Figure 2: **Amplitude modulations and phases over time**. First panel from top: *syllAM*, second panel: *accAM*, third panel: amplitude normalized *syllAM* (dotted tick line) and phase (continuous thin line); fourth panel: amplitude normalized *accAM* and phase.

A potential issue with the computation of the PLV is due to its dependency on the variability of the oscillatory rates and on the number of cycles of the two signals included in the same analysis window. To factor out these effects from our analysis, for each PLV measurement, we also computed the PLV from 10 bivariate signals obtained by pairing the $\phi(k)_{stress}$ signal used to compute the PLV with 10 different versions of the $\phi(k)_{syll}$ signal obtained by randomly permuting the positions of chunks of instantaneous phase values corresponding to different cycles. These surrogate signals display the same features of the original $\phi(k)_{syll}$ signal but lose all temporal relation with the observed $\phi(k)_{stress}$ signal. The PLV obtained by the observed data is then normalized through division by the average of the PLVs obtained from the analysis of the surrogate data. Note that due to this normalization, a significant degree of coordination is observed when the normalized PLV is significantly higher than one.

3. RESULTS AND DISCUSSION

Figure 3 displays the normalized PLVs obtained from the repetitions of different syllable sequences over speech rate (in syllables per sec).

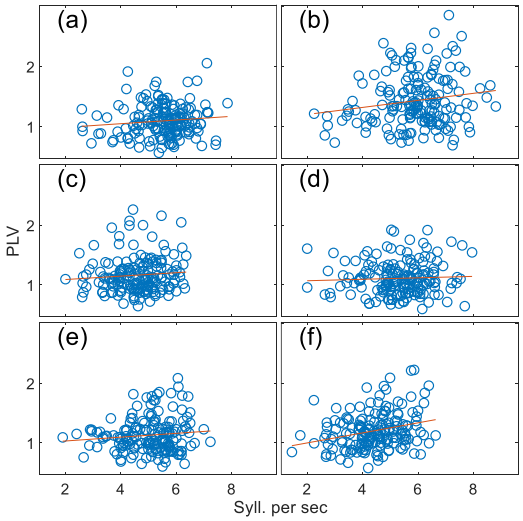


Figure 3: normalized PLV vs syllable rate for each syllable. Panel (a): /bababa/: panel (b): /badego/: panel (c): /klaklakla/; panel (d): /dedede/; panel (e): /gogogo/; panel (f): /tʁatʁatʁa/.

Higher PLVs indicates a stronger coordination between the syllabic and suprasyllabic rates. Pairwise comparisons via t-tests (corrected for multiple comparisons according to the False Discovery Rate criterion [17]) show higher PLVs for /badego/ than for other sequences and higher PLVs for /tʁatʁatʁa/ than for all the CVs (/ba/, /de/, /go/). Finally, PLV is higher during the repetition of /klaklakla/ than during the repetition of /bababa/. In summary, the coordination between the syllabic and supra-syllabic components is lower during the repetition of the same CV syllable than during the repetition of sequences containing different CV syllables or during the repetition of identical syllables with complex structure (CCV).

A linear mixed model with speaker-specific random intercepts and rate-related slopes was used to predict the effects of the syllabic content of the sequences (reference level: /badego/), of syllable rate (centered around its mean), and of their interaction on PLVs. Results show that the coordination of the two rhythmic components was stronger (higher PLV) for /badego/ sequences than for any other sequence (vs.

/bababa/: est.:-0.31, std. err.: 0.04, t val.: -8.39; vs. /klaklakla/: est.:-0.2, std. err.: 0.39 , t val.: -5.42; vs. /dedede/: est.:-0.27, std. err.: 0.036, t val.: -7.91; vs. /gogogo/: est.:-0.26, std. err.: 0.03604, t val.: -7.11; vs. /tʁatʁatʁa/: est.:-0.13, std. err.: 0.04, t val.: -3.22). The simple effect of speech rate (computed on /badego/) had a significant effect on PLV (est.: 0.07, std. err.: 0.023, t val.: 2.856). Interactions between rate and sequence types were not significant, although sequence-specific trends appear in Figure 3.

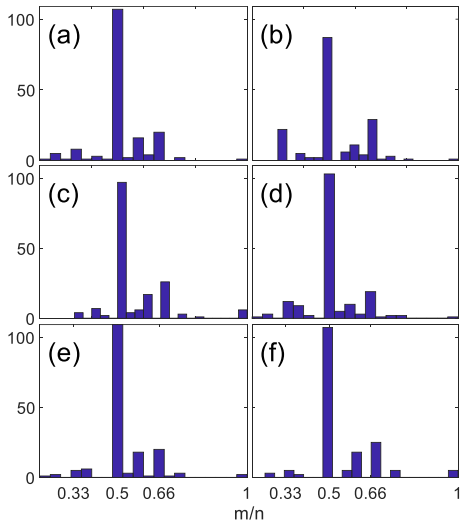


Figure 4: Counts of the optimal m/n ratio values obtained from the PLV analysis for each sequence inversely related to the number of syllables per accent. Same distribution over panels (a to f) as in Figure 3.

Figure 4 displays the distributions of the ratios between the *m* and *n* values estimated to compute the PLVs in Figure 3. As these ratios correspond to the ratios between the frequencies of the syllabic and the supra-syllabic rhythmic components, each value is inversely related to the number of syllables per accent that are produced by one speaker by sequence types. The most frequent ratio across distributions is 0.5. This means that for most of the time, speakers produce two syllables per accent. Most distributions display a peak around 0.66, compatible with the presence of two accents per group of three syllables. Finally, the peak at 0.33 is compatible with the presence of three syllables per accent. Let's note that the number of speakers that adopt these last two frequency ratios, both compatible with a grouping in chunks of three syllables (with either one or 2 accents), increases in the repetition of /badego/. Moreover, the number of speakers producing two accents per 3 syllables (probably an initial and a final accent [18], ratio of .66) increases for /tʁatʁatʁa/ (c) and /klaklakla/ (f) when compared to CV sequences.

To sum up, coordinative strategies resulting in the grouping of three syllables are more frequent in sequences displaying relatively strong degrees of coordination (high PLVs) between syllables and suprasyllabic prominence (/badego/ (b), /klaklakla/ (c) and (f) /tʁatʁatʁa/).

If coordination between rhythmic components is related to the stability of the grouping pattern adopted, a relatively strong degree of coordination is to be expected in repetitions of sequences of different syllables (as /badego/), because these permit anticipation and overlap between syllables, thus favouring grouping. However, the relatively high degrees of coordination observed in the repetition of identical complex CCV syllables may be due to the fact that speakers reduce some CCV syllables in response to a loss of stability at fast speech rate. Coherently with this interpretation, in the production of /badego/ sequences speakers can choose between two ways of chunking the speech stream in groups of three syllables (both one and two accents per group of three syllables are observed) However, in the production of /tʁatʁatʁa/ or /klaklakla/ sequences, only a few speakers produce one accent per groups of three syllables (i.e. the coordinative relation between syllables and accents underlying this grouping strategy becomes unstable).

4. CONCLUSION

This study brings further evidence of the presence of global coordination patterns in speech production. Their role in speech sensorimotor control is quite crucial, as they tie the behaviour of the many different elements composing the vocal apparatus to the production of functional constituents including several syllables. Moreover, speakers that succeed in reaching faster speech rates, as required by the task, better coordinate the production of syllables with the production of accentual prominence. This result reconciles the presence of speech rate differences between languages with the presence of genuine rhythmic differences ([19, 20]), because the latter kind of differences may be induced by the former one. Further work aimed at including other dimensions (e.g. articulator movements and f0) and time scales in the characterization of the rhythmic aspects of speech production within and across languages will benefit from the observational and modelling paradigms introduced in this paper. The proposed analytical approach constitutes a major advancement in the modelling of the rhythmic hierarchy underlying speech, as it requires minimal linguistic information (syllables and segments rates), which often is available, or which can be estimated (e.g. [21, 22]).

5. ACKNOWLEDGMENTS

Research supported by grant ANR-21-CE28-0015-01 (PASDCODE) of the French National Research Agency.

6. REFERENCES

[1] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics, 11*(1), 51-62.

[2] Dauer, R. M. 1987. Phonetic and phonological components of language rhythm. In *Proceedings of the 11th international congress of phonetic sciences* 5, 447-450.

[3] Bertinetto, P. M. 1989. Reflections on the dichotomy 'stress' vs.'syllable-timing'. *Revue de phonétique appliquée, 91*(93), 99-130.

[4] Auer, P. 1993. Is a rhythm-based typology possible. A study of the role of prosody in phonological typology (KontRI Working Paper No. 21). Hamburg: Universität Hamburg.

[5] Leong, V., Stone, M. A., Turner, R. E., & Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America, 136*(1), 366-381.

[6] Lancia, L., Krasovitsky, G., & Stuntebeck, F. 2019. Coordinative patterns underlying cross-linguistic rhythmic differences. *Journal of Phonetics, 72*, 66-80.

[7] Stetson, R. H. 1951. *Motor Phonetics*. North-Holland, Amsterdam, 2nd ed.

[8] Tuller, B., & Kelso, J. S. 1991. The production and perception of syllable structure. *Journal of Speech, Language, and Hearing Research, 34*(3), 501-508.

[9] Rochet-Capellan, A., & Schwartz, J. L. 2007. An articulatory basis for the labial-to-coronal effect:/pata/seems a more stable articulatory pattern than/tapa. *The Journal of the Acoustical Society of America, 121*(6), 3740-3754.

[10] Lancia, L., & Rosenbaum, B. 2018. Coupling relations underlying the production of speech articulator movements and their invariance to speech rate. *Biological cybernetics, 112*(3), 253-276.

[11] Fougeron, Cécile, Véronique Delvaux, Lucie Menard, and Marina Laganaro. 2018. The MonPaGe_HA database for the documentation of spoken French throughout adulthood. Proceedings of the 11th LREC, 2018, Miyazaki, Japan, pp. 4301–6.

[12] Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. 2017. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America, 142*(4), 1976-1989.;

[13] Poeppel, D., & Assaneo, M. F. 2020. Speech rhythms and their neural foundations. *Nature reviews neuroscience, 21*(6), 322-334.

[14] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971), 903-995.

[15] Rosenblum, M., Pikovsky, A., Kurths, J., Schäfer, C., & Tass, P. A. 2001. Phase synchronization: from theory to data analysis. In *Handbook of biological physics* (Vol. 4, pp. 279-321). North-Holland.

[16] Huang, N. E., Wu, Z., Long, S. R., Arnold, K. C., Chen, X., & Blank, K. (2009). On instantaneous frequency. Advances in adaptive data analysis, 1(02), 177-229.

[17] Benjamini, Y., & Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

[18] Jun, S. A., & Fougeron, C. (2002). The Realizations of the Accentual Phrase in French Intonation. *Probus, 14*(special issue on Intonation in the Romance Languages), 147-172.

[19] Arvaniti, A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica, 66*(1-2), 46-63.

[20] White, L., Mattys, S. L., & Wiget, L. 2012) Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*(4), 665-679.

[21] Wang, D., & Narayanan, S. S. 2007. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(8), 2190-2201.;

[22] Hoang, D. T., & Wang, H. C. 2015. Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America, 137*(2), 797-805.

RESPIRATORY AND SUPRALARYNGEAL EFFECTS ON SPEECH BREATHING NOISE ACROSS LOUDNESS CONDITIONS AND SPEAKING TASKS

Susanne Fuchs[1], Laura L. Koenig[1,2,3], Raphael Werner[4]

[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS) Berlin (Germany), [2]Adelphi University, New York (USA), [3]Haskins Labs, New Haven (USA), [4]Universität des Saarlandes, Saarbrücken (Germany)

fuchs@leibniz-zas.de, koenig@haskins.yale.edu, rwerner@lst.uni-saarland.de

ABSTRACT

Respiratory and supralaryngeal actions differ between speech and vegetative breathing, and speech-breathing sounds convey information to listeners. To date, little work has explored how breath noise is shaped by respiratory and supralaryngeal actions. Here, we assess respiratory movement, obtained using inductive plethysmography, and lip apertures, collected using electrostomatography, for eight typical German-speaking women. We varied loudness and speaking task since these can be expected to affect breathing behavior. Participants were recorded while reading, producing a spontaneous narrative, and engaging in an interactive memory game with an interlocutor. The data show widespread effects of speaking task on breath acoustics and respiratory kinematics, with more naturalistic tasks generally yielding louder breath noise, and steeper inhalation slopes. Louder speech is associated with greater oral apertures. Breath intensity correlates with inhalation slope. Lip apertures do not correlate with breath acoustics during reading, but may do so for other tasks.

Keywords: speech breathing, speech kinematics, loudness, task effect

1. INTRODUCTION

Breathing is a natural rhythm that structures the flow of speech temporally [1]. Breath sounds are ubiquitous in spoken language. They mostly occur during inhalation, but can also be produced under physical effort, i.e. forced exhalation. Speech usually involves egressive airflow, but ingressive phonation can also be observed [2] and may convey information about intense emotional states [3]. In general, breath noise may provide listeners with considerable information about the speaker [4] (e.g., Is it a human or a machine? What is the level of physical activity or emotion?) and may also be crucial for management of turn-taking.

In rest breathing, nasal inhalation is the default. However, speech breathing requires faster gas exchange. Previous work has shown that, besides oral-nasal sequences, breath noise can be a consequence of oral, nasal or simultaneous oral-nasal inhalation. In [5], speech was elicited by asking the participant to count, read a paragraph, speak spontaneously about their favorite activity, and do a phone call with a familiar person. The authors also assessed coarticulatory effects of surrounding sounds. The findings showed that speakers typically produce simultaneous oral-nasal breathing in speech. The data also showed a non-significant trend suggesting possible differences in speech breathing patterns as a function of speech task.

Although there have been some studies on articulation during pauses [6, 7, 8, 9] it is neither clear whether inhalations have been produced during these pauses nor what acoustic consequences a breath-related articulation has. Only a few studies have analyzed the relationship between acoustics and articulation of breath noise.

Our work follows up on Werner and colleagues [10]. They compared acoustic characteristics of breath noises to speech sounds with similar acoustic or articulatory characteristics, such as aspiration noise in stops and formants in vowels. The authors also investigated the relationship between acoustic measures of inhalation noise and respiratory kinematics using Inductance Plethysmography. They showed that the Centre of Gravity of breathing noise was similar to the aspiration noise of /k/, but not to that of /t/, nor to the glottal frication noise of /h/. The authors [10] found that breath noises had a higher F1 and slightly higher F2 than /ə/ and thus seemed to involve more open, slightly fronted vocal tract configurations. Results revealed

that inhalation slope, i.e. how fast participants inhale, was positively correlated with the Center of Gravity, F1, and intensity of the breath noise. However, articulatory data on mouth openings were not recorded.

The current study extends previous work in three directions to investigate 1) breath noises in different speech tasks (reading, an interactive game, a narrative); 2) the effect of loudness (speech produced with normal and loud intensity) on breath noise; and 3) the relation between breath noise, inhalation slope, and lip opening.

Exploration of different speech tasks is motivated by [5] and the fact that in a conversation an interlocutor may breathe silently through the nose when listening, but may change her breathing pattern when taking a turn. Breathing while telling a story may add more variability to the acoustic properties of the breath signal because inhalation depth roughly corresponds to the length of the upcoming sentence which may vary considerably in spontaneous speech [11].

Varying loudness as a condition is motivated by the fact that loudness is known to modify breathing behavior [12] as well as acoustics [13] and supralaryngeal articulation [14, 15]. It is, however, unclear whether loudness affects only speech itself, or also enhances voiceless inhalation noise. Finally, using a multimodal instrumental setup permits investigating acoustic-supralaryngeal-respiratory relations directly.

We predict that speaking with higher intensity (in the loud condition) leads to more rapid inhalation and also affects the intensity of the breath noise. However, since there is no phonation in breath noise, the effect might be much smaller than in speech itself. Furthermore, we expect that F1 of the breath noise is related to lip aperture and inhalation slope.

For speech task, our predictions are rather tentative. However, following [5], we suspect that more spontaneous speech tasks may differ from read speech, and in particular may show greater variability in acoustic and kinematic measures.

2. METHODOLOGY

2.1. Experimental design

Two different conditions were experimentally varied: *Speech Task* and *Loudness*. Speech Task comprised the following:

1. Reading sentences (Read) of comparable length varying a target word (sentences were taken from [13]). All utterances began with "Ich mag" (I like) followed by a target word beginning with a bilabial oral or nasal stop (e.g., Mate, Paten, Buesum).

2. Speaking about a summer holiday (Holiday). Participants were instructed: "Do you remember a summer holiday? Please tell us about this holiday. You can also talk about other things. It is important that you speak for at least 2-3 minutes. It may also be longer."

3. Playing the game "I pack my suitcase and take with me ..." (Game). In this game, a participant has to propose one thing that she will take in her suitcase and the next participant needs to repeat the sentence and adds another item to the suitcase. In the end, the suitcase will be full of items and the aim of the game is to memorize everything in the right order. The game ends when one of the participants does not remember everything or puts items in the wrong order. The experimenter served as an interlocutor for the game task.

The reading sentences and the game were repeated twice for each loudness condition. Speech tasks were presented randomly to different participants. All participants produced these tasks first using their normal comfortable loudness and in the second part of the experiment, they were instructed to speak louder, because their interlocutor (the experimenter) was wearing fully ear-covering headphones. In case participants became softer after some time, the interlocutor gave signs that she could not understand and the participant should try to maintain the loudness level.

2.2. Participants

Eight female speakers, all native speakers of German, were recorded. They spoke Standard German (Northern variety). Speakers were selected on the basis of an available custom-made palate and tolerance to wearing this palate. The palates are part of an Electro-Optical Stomatographic System (EOS) [16]. Since the costs for these palates were relatively high, we could not afford to record more participants. Participant ages ranged from 21-34 years and Body Mass Indices were between 19 and 27.

2.3. Equipment

Acoustic data were recorded via two devices: a high-quality Sennheiser condenser microphone that was connected to an amplifier and a PC that simultaneously acquired respiratory and acoustic data. The second acoustic signal was of low quality and was recorded with the internal microphone of the laptop controlling the Electro-Optical

Stomatographic system (see next paragraph). The acoustic data from the two microphones were taken to post-synchronize the data between the PC recording breathing and the PC for the EOS device.

Respiratory kinematics were recorded via Inductance Plethysmography. Two flexible respiratory belts were wrapped around the participant's torso, one at the abdomen and one at the thorax. The EOS device was used to record lip opening during breathing. The EOS system includes a custom-made artificial palate that can record tongue-palatal contact patterns during speech, but additionally senses the mid-sagittal distance from the tongue to the palate, in case there is no contact. Two lip sensors were connected to the EOS system that record lip motions (protrusion and aperture). We will here concentrate on the lip aperture only.

2.4. Data pre-processing and annotation

The on- and offsets of breathing noise were labelled in Praat [17]. We then extracted acoustic parameters from the audio signal using a Praat script. F1, F2, F3 were averaged over the central third of the breath noise, to avoid the influence of potential coarticulation as done in [10], using the Burg algorithm with max. number of formants 5, max. formant 5.5 kHz, window length 0.025 s, and dynamic range 50 Hz. Intensity was averaged over the whole interval. Total lung volume was computed as $2 \times thorax + abdomen$ [18]. We then took the acoustically defined on- and offset of breathing noise (x) and obtained the corresponding respiratory values (y) for these time points. Inhalation slope was calculated as $slope = \Delta y / \Delta x$. It is thus a measure of how quickly speakers inhale.

As noted above, the EOS data processing focused on the lip aperture signal only. Given the poor quality of the associated acoustic signal, we also restricted ourselves to the reading task, where every utterance began with a pair of bilabial consonants (the /m/ in 'Mag' followed by the bilabial-initial target word). These two bilabial closures allowed us to orient to the EOS data and delimit mouth openings that preceded utterances. For these mouth openings, we obtained the aperture as the distance between the lip closure of /m/ and the maximum breath opening. Since lip opening movements during breathing showed considerable spatial and temporal variability, we used overall aperture instead of an opening slope or velocity.

Linear mixed-effect models were run in R [19] using the packages *lme4* [20] and *lmerTest* [21]; graphical exploration was done using *ggplot2* [22]. We describe the statistical models in the respective results section. We used $\alpha = 0.05$ for significance testing.

Our sample includes 1,518 inhalations overall. We excluded cases where inhalations were preceded or followed by laughter because this can coincide with some extreme respiratory dynamics [23] and changes in intensity. We also excluded cases where some obvious spectral changes occurred and perceptually these sounded like oral-to-nasal or nasal-to-oral inhalations. Our final sample consisted of 1,297 breath noises.

3. RESULTS

3.1. Do loudness and speech tasks affect acoustic parameters of breath noise and speech?

To answer this question, we ran the model $lmer(intensity \sim Loudness * Task + (1 + Loudness * Task|Speaker))$. Participants did change their speech intensity between the normal and the loud conditions. On average speech was 7.3 dB louder (t=7.205, p<0.001***) in the loud than in the normal condition. Speech task (Read speech, Game, Holiday) had no effect on this intensity level.

The results look different for speech breathing (Fig. 1): Here we ran the same model as described above, but used breath noise intensity as the dependent measure instead. In the loud condition, breath noise was not significantly louder; only a trend towards higher intensity could be found (t=2.17, p=0.067). However, contrary to speech intensity, an effect of the task was found. Overall, breath noise in the interactive Game task was louder (t=2.51, p=0.039*) than in Read speech and the Holiday task was also louder (t=8.252, p<0.001***) than Read speech.

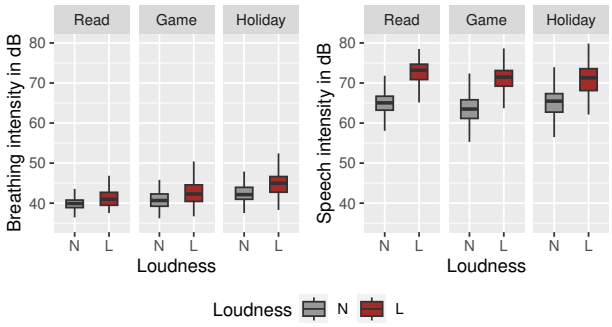


Figure 1: Boxplot for breathing noise intensity (left) and speech intensity (right) in different conditions.

F1 values of the breath noise only showed a trend

towards slightly higher F1s (average 28 Hz) in loud speech (t=2.16, p=0.057) than in normal speech. A trend, but no effect was also found for speech task with F1 being slightly higher in the Game (t=2.07, p=0.075) and Holiday (t=2.001, p= 0.082) tasks than in Read speech. These outputs were obtained using the formula $lmer(F1 \sim Loudness * Task + (1 + Loudness * Task|Speaker))$.

3.2. Do loudness and speech tasks affect breathing slope?

Results for breathing slope are similar to those for intensity, as they only show a trend for loudness (t=2.14, p=0.07), but a main effect for speech task: breath noise in the Holiday task had a steeper slope than in the reading task (t=2.96, p=0.0211*).

3.3. Does loudness affect lip opening?

Finally, the effect of loudness on lip aperture was calculated as follows: $lmer(LipAperture \sim Loudness + (1 + Task|Speaker))$. Since only data for Read speech were analyzed, speech task did not serve as a fixed factor. Loudness affected lip opening, i.e. louder speech was produced with a more open lip configuration during inhalation (t=3.77, p=0.009**).

3.4. Relation between acoustic and articulatory parameters

To what extent are acoustic and articulatory parameters of breath noise correlated? To better understand the relations, we ran the following model: $lmer(intensity \sim slopeCentralized + Loudness * Task + (1 + Loudness * Task|Speaker) + (1 + slopeCentralized|Speaker))$. The predictor Breathing slope was centralized. Results revealed a strong effect (Fig. 2) of breathing slope on intensity of the breath noise (t=6.22, p=0.0005***).

A similar model with F1 as the dependent variable returned no significant effect of breathing slope.

Finally looking at the relation between lip aperture and F1, which was possible only for the Read speech data ($lmer(aperture \sim F1centralized * Loudness + (0 + F1centralized * Loudness|Speaker))$), there was only an effect of Loudness (t=2.55, p=0.048*).

4. DISCUSSION AND CONCLUSION

This study found the following effects: Breath noise intensity differed across speech tasks. It did not vary significantly across loudness conditions, which is probably due to the voiceless nature of the breath noise and speaker-specific realizations of the loudness condition. Similarly, F1 was higher in the Game and Holiday task than in Reading, i.e. it varied according to the task, but only showed an insignificant trend for normal versus loud speech. This was also true for inhalation slope, which showed no effect of Loudness but differences among tasks. A different pattern emerges from the lip aperture data, which were only analyzed for the reading task. In the loud speech condition, speakers opened their mouth wider when producing breath noise than in the normal condition. Thus, speech task had greater effects on acoustic parameters of breath noise than changing between normal and loud speech. In particular, reading had lower breath intensity and F1 values of breath noise and a shallower inhalation slope than one or both other tasks. One reason for this may be that the sentences in the Read task were relatively short and we kept them fairly constant in length. We assume that no deep inhalation was required for this reason and overall, data may be less variable than in the other tasks.

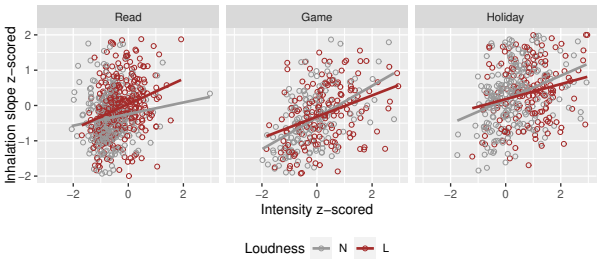


Figure 2: Scatterplot with linear regression lines showing a relation between inhalation slopes (z-scored by speaker) and breathing intensity (z-scored by speaker) in the three different speech tasks. Note, z-scoring is only done for graphical display. Grey colour corresponds to normal speech and dark red to loud speech.

Looking at how acoustic and articulatory measures relate, we could confirm the relation between intensity and inhalation slope as proposed by [10] but could not find a relation between F1 and inhalation slope. We think that no correlation between F1 and lip aperture might be due to nasal coupling in some speakers (cf. [5]). However, data for the other speech tasks need to be investigated before any reliable conclusions can be drawn.

Speech breathing involves the coordinated actions among different articulators and all drivers of respiration. Using multimodal data to study different aspects of this coordination may be a fruitful topic for further investigations.

5. ACKNOWLEDGMENTS

This work was supported by the Leibniz Association. We like to thank our participants and Jörg Dreyer for technical support, Peter Birkholz and Alexander Wilbrandt for advice on the EOS system.

6. REFERENCES

[1] S. Fuchs and A. Rochet-Capellan, "The respiratory foundations of spoken language," *Annual Review of Linguistics*, vol. 7, no. 1, pp. 13–30, 2021.

[2] R. F. Orlikoff, R. J. Baken, and D. H. Kraus, "Acoustic and physiologic characteristics of inspiratory phonation," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1838–1845, 1997.

[3] A. Anikin and D. Reby, "Ingressive phonation conveys arousal in human nonverbal vocalizations," *Bioacoustics*, pp. 1–16, 2022.

[4] M. Kienast and F. Glitza, "Respiratory sounds as an idiosyncratic feature in speaker recognition," in *Proc. 15th ICPhS, Barcelona*, 2003, pp. 1607–1610.

[5] R. A. Lester and J. D. Hoit, "Nasal and oral inspiration during natural speech breathing," *Journal of Speech, Language, and Hearing Research*, vol. 57, pp. 734–742, 2014.

[6] B. Gick, I. Wilson, K. Koch, and C. Cook, "Language-specific articulatory settings: Evidence from inter-utterance rest position," *Phonetica*, vol. 61, no. 4, pp. 220–233, 2004.

[7] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. EL160–EL165, 2009.

[8] O. Rasskazova, C. Mooshammer, and S. Fuchs, "Temporal coordination of articulatory and respiratory events prior to speech initiation." in *Interspeech*, 2019, pp. 884–888.

[9] J. Krivokapić, W. Styler, and D. Byrd, "The role of speech planning in the articulation of pauses," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 402–413, 2022.

[10] R. Werner, S. Fuchs, J. Trouvain, and B. Möbius, "Inhalations in speech: Acoustic and physiological characteristics," in *Interspeech 2021*. ISCA: ISCA, 2021, pp. 3186–3190. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/werner21_interspeech.html

[11] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, "Acoustic and respiratory evidence for utterance planning in German," *Journal of Phonetics*, vol. 41, no. 1, pp. 29–47, 2013.

[12] J. E. Huber, "Effects of utterance length and vocal loudness on speech breathing in older adults," *Respiratory Physiology & Neurobiology*, vol. 164, no. 3, pp. 323–330, 2008.

[13] L. L. Koenig and S. Fuchs, "Vowel formants in normal and loud speech," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 5, pp. 1278–1295, 2019.

[14] R. Schulman, "Articulatory dynamics of loud and normal speech," *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 295–312, 1989.

[15] J. E. Huber and B. Chandrasekaran, "Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults," *Journal of Speech, Language, and Hearing Research*, vol. 49, no. 6, pp. 1368–1379, 2006.

[16] S. Stone and P. Birkholz, "Silent-speech command word recognition using electro-optical stomatography." in *Interspeech*, 2016, pp. 2350–2351.

[17] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2022.

[18] R. B. Banzett, S. T. Mahan, D. M. Garner, A. Brughera, and S. H. Loring, "A simple and reliable method to calibrate respiratory magnetometers and respitrace," *Journal of Applied Physiology*, vol. 79, no. 6, pp. 2169–2176, 1995.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[21] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[22] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[23] M. Filippelli, R. Pellegrino, I. Iandelli, G. Misuri, J. R. Rodarte, R. Duranti, V. Brusasco, and G. Scano, "Respiratory dynamics during laughter," *Journal of Applied Physiology*, vol. 90, no. 4, pp. 1441–1446, 2001.

SPEAKERS COMPENSATE FOR TEMPORAL PERTURBATIONS IN AUDITORY FEEDBACK REGARDLESS OF SYLLABLE STRUCTURE

Robin Karlin[1], Chris Naber[1], & Benjamin Parrell[1,2]

[1]Waisman Center, University of Wisconsin – Madison; [2]Communication Sciences and Disorders, University of Wisconsin – Madison
rkarlin@wisc.edu; cwnaber@wisc.edu; bparrell@wisc.edu

ABSTRACT

It is known that speakers monitor sensory feedback during speech to maintain accurate production. However, it is unclear the extent to which speakers use feedback for timing speech gestures, and how that may interact with phonological structure.

We present an altered auditory feedback study that accelerates and decelerates the formant transition in /aɪ/, testing the hypothesis that speakers use feedforward control to time gestures within a syllable, and auditory feedback to initiate the next syllable. Results suggest that speakers used directionally specific temporal information from auditory feedback: they shortened the perturbed vowel when transitions were accelerated, and slowed down when transitions were decelerated. Shortening was restricted to perturbed segments, while slowing occurred on both perturbed segments and following vocoids, suggesting that some of the slowing may be a non-specific response to feedback delay. Similar results between and within syllables suggest that temporal monitoring in speech has a flat lookahead structure.

Keywords: speech motor control, altered auditory feedback, timing

1. INTRODUCTION

A large body of literature has shown that speakers make use of sensory feedback to ensure accurate speech production. Evidence is particularly robust in the spectral domain, where speakers both compensate online [1]–[4] and adapt their future productions [5], [6] in response to perceived errors.

An outstanding question in the field is how speakers use temporal aspects of auditory feedback in speech production, particularly in the initiation of upcoming gestures. One hypothesis, selection-coordination theory [7], posits that speakers use a mix of feedforward and feedback mechanisms to time the initiation of speech gestures, with the balance of mechanisms based on phonological structures. Selection-coordination theory relies on the concept of "co-selection sets", which are sets of gestures that are selected together and executed using coordinative control (a type of feedforward control). Speakers activate the next co-selection set via sensory feedback that confirms the achievement of the goals of the ongoing co-selection set. Although the scope of gestures in a co-selection set has been suggested to vary based on the context or intent of any given utterance, syllable-sized units are suggested to be the prototypical example of coordinated gestures in a co-selection set, similar to what is used in the DIVA and GODIVA models of speech production [8], [9].

There is some evidence that speakers do attend to temporal aspects of auditory feedback. For example, a large body of literature has shown that delaying auditory feedback alters speech behavior, ranging from overall slowed production at delays from ~50-200 ms, to stuttering-like behavior at larger delays [10]–[13]. However, it is unclear if these responses are the result of compensating for delays (i.e., initiating gestures later because the previous goals were achieved later), or more general motor responses to perceived errors.

Cai et al. [14] explored this question by using bidirectional temporal perturbations, accelerating or decelerating the transition between "I" and "owe" in "I owe you a yoyo". Speakers significantly slowed their speech in the deceleration condition compared to both acceleration and baseline productions, but the acceleration condition did not significantly differ from baseline. Furthermore, the perturbation targeted a region of transition from one word to another, which may not be as specifically timed as gestures within a word. As such, this study did not provide direct evidence that speakers use auditory feedback to guide the timing of upcoming gestures. It also did not address how temporal control of speech interacts with phonological structure.

In the current study, we test the hypothesis suggested in selection-coordination theory that syllable-sized co-selection sets are timed using feedforward mechanisms, and subsequent syllables are timed by auditory feedback. Contrary to prediction, our results indicate that the timing of subsequent gestures responds to temporal changes in auditory feedback both within a single syllable and across syllables. This suggests that speakers use auditory feedback to guide the initiation of subsequent gestures regardless of syllabic affiliation.

2. METHODS

2.1. Participants

Six participants (3 F, 3 M) have participated in this study, of an anticipated 20. Participants ranged in age from 51-61 (median 53) and reported no history of neurological, hearing, or speech disorders. All participants were native speakers of General American English with a diphthongal production of /aɪ/. Participants were monetarily compensated for their time. All procedures were completed in accordance with the IRB at the University of Wisconsin – Madison and the University of California, San Francisco.

2.2. Target phrases

The target vowel for this study was the diphthong /aɪ/. To test the hypothesis that 1) gestural timing across syllable boundaries relies on sensory feedback and 2) gestural timing within a syllable is controlled in a feedforward manner as a single coordinated unit, the vowel was embedded in two target contexts:

1. **buy donuts** [baɪ doʊnəts]: the next consonant follows a word boundary as an onset consonant;
2. **guide boaters [**gaɪd boʊɾəɪz**]:** the next consonant immediately follows the vowel as a coda.

As selection-coordination theory suggests that co-selection sets are formed from highly practiced and frequently co-activated gestures, in this study we use syllable boundaries that are also word boundaries to increase the likelihood that the following gesture is activated differently in each phrase.

Table 1: Summary of phrases used in the experiment.

Phrase	Target	Following gesture
BUY donuts	aɪ	# d
GUIDE boaters	aɪ	d #

2.3. Auditory feedback perturbation

Temporal perturbations were achieved using a novel formant clamp functionality of Audapter [15]. The formant clamp feeds in a predetermined formant trajectory when the onset of the target vowel is detected using online status tracking (OST), and returns to veridical feedback when OST detects the end of the vowel.

Clamped formant trajectories were based on average productions from the most recent 10 unperturbed trials for a specific target word. To change the timing of the diphthong transition, the average formants were warped offline via nonlinear resampling. As the General American production of /aɪ/ has a relatively late transition to [ɪ], the resampling functions for acceleration and deceleration were not symmetrical, but rather calibrated to produce a large temporal difference between the accelerated and decelerated conditions without making the decelerated condition sound monophthongal. The resulting perturbations were thus temporally larger for acceleration than for deceleration. Example perturbations are provided in Figure 1.

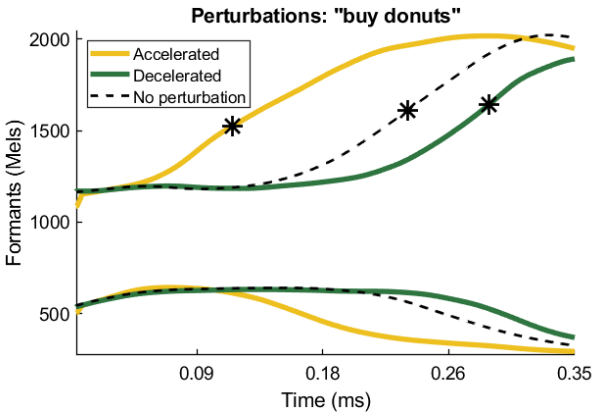


Figure 1: Mean F1 and F2 heard by one participant for "buy donuts", time-normalized to the overall mean duration of "buy". The asterisks on the F2 trajectory indicate point of maximum velocity in F2, where perturbation magnitude is measured.

2.4. Task

In each trial, participants saw the target phrase on the screen and read it out loud. The study consisted of 10 blocks. Participants heard veridical feedback for the first block, which served both as a baseline of production and to establish the mean duration and formant trajectories of each target phrase for formant clamping. In the remaining nine blocks, each target phrase was produced twice with each perturbation condition, and twice with no perturbation. In total, each unique combination of target phrase and perturbation condition was produced 18 times, with the exception of unperturbed trials, which were produced a total of 28 times (10 times during baseline block, 18 times during perturbation blocks). Trials were pseudorandomized with additional distractor phrases such that no two adjacent trials shared either target phrase or perturbation condition.

To facilitate the use of auditory feedback during the target vowel, participants were trained before the experiment to use a moderate speech rate and to place focus on the target word (e.g., "BUY donuts"). During the experiment, participants continued to receive feedback on their speech rate based on the duration of the target vowel as estimated by OST. The goal duration for the target vowel was 300 – 600 ms.

The achieved mean duration of the target diphthongs was 369 and 354 ms for "buy donuts" and "guide boaters", respectively (SD = 36, 33 ms).

2.5. Data processing and analysis

Trials were first segmented with Montreal Forced Aligner [16], and then hand-corrected by the first author, blind to perturbation condition. Transitions between /d/ and /b/ in "guide boaters" were not hand-corrected unless there was a visible release of /d/ or incorrect silent spans were inserted.

The dependent variable is the difference in segment duration between perturbed trials and unperturbed trials (including trials from the baseline block, as well as the unperturbed trials from experimental blocks). Reported estimated means are thus change from the mean of unperturbed trials.

Changes in production are analyzed for the target vowel, the following consonants, and the following vowel. Data was analyzed using linear mixed effects models using the lme4 package in R [17], [18]. Models included random intercepts per participant. Models were built incrementally, starting with the null model, and adding fixed effects of perturbation (acceleration, deceleration, unperturbed), target phrase (buy donuts, guide boaters), and their interaction. Models were compared using the anova function from lmerTest [19]. Post-hoc tests were conducted with the emmeans package [20].

3. RESULTS

Overall results for each perturbation condition are illustrated in Figure 2.

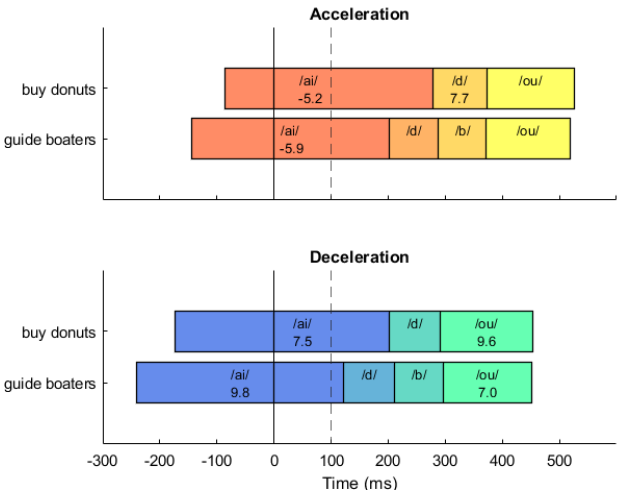


Figure 2: Changes in duration for segments in each phrase. Phrases are zero-aligned to the point of maximum perturbation (solid vertical line). The dashed line indicates the minimum latency necessary to compensate for auditory feedback (100 ms, [21]). All included values are significantly different from unperturbed.

3.1. Duration of /aɪ/

There is a significant effect of perturbation condition on the duration of /aɪ/ ($\chi^2(2) = 30.72$, $p < 0.0001$); compared to the unperturbed condition, /aɪ/ is shorter in the acceleration condition (-5.5 ± 2.1 ms, $p = 0.005$), and longer in the deceleration condition (8.7 ± 2.1 ms, $p = 0.007$). Adding target phrase to the model does not significantly improve model fit ($\chi^2(1) = 0.90$, $p = 0.34$); nor does the interaction between target phrase and perturbation condition ($\chi^2(2) = 0.69$, $p = 0.71$). This indicates that for both "buy donuts" and "guide boaters", speakers shortened the vowel under acceleration, and lengthened the vowel under deceleration. See Figure 3 for an illustration of these changes.

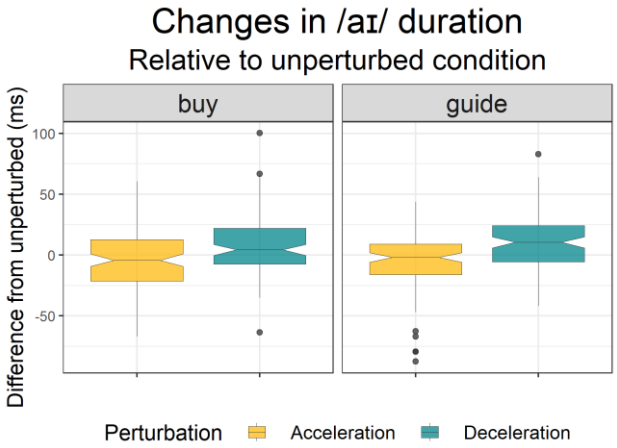


Figure 3: Changes in the produced duration of /aɪ/ under acceleration and deceleration for each phrase.

3.2. Duration of following consonants

Coda consonant of "guide": perturbation condition does not significantly improve model fit ($\chi^2(2) = 1.06$, $p = 0.59$), indicating that the coda /d/ in "guide" does not change in duration under either acceleration or deceleration.

*Onset consonants of following words (**d**onuts, **b**oaters):* Perturbation condition does not improve model fit ($\chi^2(2) = 2.86$, $p = 0.24$), nor does the addition of target word ($\chi^2(1) = 0.52$, $p = 0.47$). However, the interaction between target word and perturbation does improve model fit ($\chi^2(2) = 7.92$, $p = 0.02$). The /d/ in "buy donuts" is significantly longer in the acceleration condition compared to no perturbation (7.7 ± 2.3 ms, $p = 0.03$); no other segments show any significant differences.

3.3. Duration of following vowel /oʊ/

Adding perturbation condition significantly improves model fit ($\chi^2(2) = 43.94$, $p < 0.0001$). Vowels are significantly longer in the deceleration condition (7.8 ± 1.8 ms) compared to both acceleration (-0.5 ± 1.8

ms) and no perturbation (1.0 ± 1.7 ms; both $p < 0.0001$), but acceleration is not significantly different from no perturbation ($p = 0.43$). Target phrase does not significantly improve model fit ($\chi^2(1) = 1.57$, $p = 0.21$), nor does the interaction between target phrase and perturbation ($\chi^2(2) = 0.30$, $p = 0.86$). For both "buy donuts" and "guide boaters", the following stressed vowel is significantly longer in the deceleration condition (buy: 9.6 ± 1.8 ms; guide: 7.0 ± 1.8 ms) compared to both acceleration and no perturbation ($p \leq 0.001$ for all comparisons).

4. DISCUSSION

These preliminary results suggest that speakers use auditory feedback of their own speech to guide the timing of upcoming articulatory gestures, regardless of whether they belong to a different word. In the acceleration condition, the vowels in both "buy donuts" and "guide boaters" were significantly shorter than in unperturbed trials. In addition, the /d/ in "buy donuts" was significantly longer; no other segments showed significant changes. Similarly, in the deceleration condition the /aɪ/ in both "buy donuts" and "guide boaters" was significantly longer than in unperturbed trials. However, increased durations were also found in the /oʊ/ in the following words for both phrases.

Critically, in both acceleration and deceleration conditions, the /aɪ/ in "buy donuts" and "guide boaters" changed similarly. This suggests that speakers advanced the /d/ closure in time regardless of if it was in the same word or in the next word, indicating that the role of auditory feedback in gestural initiation is not affected by phonological boundaries, at least at the word level.

One possible interpretation of this pattern of data is that speech gestures are planned with specific intergestural timing goals that are based in motor predictions. When speakers receive feedback that some part of the plan (e.g., the lingual gesture) has achieved a state at a different time than they predicted, they adjust other gestures so that relative timing remains intact. In the case of "buy donuts" and "guide boaters", the tongue tip gesture adjusts to maintain relative timing with the tongue body: under acceleration, the tongue tip gesture is accelerated to achieve its goal earlier; under deceleration, it is delayed. Because our results using speech acoustics could examine only the achievement of oral closure, it is unclear whether these temporal changes are accomplished by advancing and delaying the initiation of the tongue tip gesture associated with the consonant or by altering the trajectory of an ongoing gesture. Thus, it is unclear if speakers do in fact use auditory feedback to time gestural initiation, as suggested by selection-coordination theory, or if temporal compensation is limited to adjusting ongoing movements (e.g., adjusting the velocity of a movement). More detailed information on this process could shed light on the planning horizon for gestures within and across syllable and/or word boundaries. This question could be resolved in the future through examination of speech kinematics.

Finally, in both the acceleration and deceleration perturbations, speakers lengthened unperturbed segments: in acceleration, the /d/ in "donuts" was lengthened, and in deceleration, the stressed vowel in the following word was lengthened in both phrases. One possibility is that there are two components to the reaction to temporal perturbations: first, a compensatory reaction, which appears in the perturbed vowel, and second, a non-specific response to a large temporal error, which appears later. Slowing down could facilitate the use of sensory feedback for state estimation after an error in temporal prediction is perceived by the speaker. An attempt at motor recovery may be reflected in effects induced by delayed auditory feedback, which at small magnitudes induces lengthened segments and at large delay magnitudes results in disfluencies and articulatory breakdown [11], [12]. It is unclear why only "buy donuts" showed additional lengthening in the acceleration condition, while neither /d/ nor /b/ (or both combined) showed any changes in "guide boaters". As the coda and onset gestures are frequently overlapped both in this dataset (as evidenced by the lack of release burst on /d/) and in English generally [22], one possibility is that there was some gestural elongation that resulted in increased gestural overlap, thus not producing acoustic evidence of elongation. Kinematic data could potentially provide insight on this issue.

5. CONCLUSION

Overall, this study provides evidence that speakers use auditory feedback to guide the timing of upcoming gestures. Crucially, this study indicates that speech gestures within a syllable are not timed exclusively via feedforward mechanisms, as proposed by selection-coordination theory. Instead, the data suggests that speakers always use feedback to maintain desired timing relationships between sequential speech gestures regardless of the syllabic or even word level affiliation of those gestures.

6. ACKNOWLEDGEMENTS

This work was supported in part by NIH grants F32DC019535 (RK), R01 DC017091 (BP), P50HD105353 (Waisman Center).

7. REFERENCES

[1] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech I: Compensation and adaptation," *Journal of Speech, Language, and Hearing Research*, vol. 45, pp. 295–310, 2002.

[2] K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. Johnsrude, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 384–390, 2009.

[3] J. A. Jones and K. G. Munhall, "Perceptual calibration of F0 production: Evidence from feedback perturbation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1246–1251, 2000.

[4] V. M. Villacorta, J. S. Perkell, and F. H. Guenther, "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2306–2319, 2007.

[5] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[6] D. W. Purcell and K. G. Munhall, "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 966–977, 2006.

[7] S. Tilsen, "Selection and coordination: The articulatory basis for the emergence of phonological structure," *Journal of Phonetics*, vol. 55, pp. 53–77, 2016.

[8] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and cognitive processes*, vol. 26, no. 7, pp. 952–981, 2011.

[9] J. W. Bohland, D. Bullock, and F. H. Guenther, "Neural representations and mechanisms for the performance of simple speech sequences," *Journal of cognitive neuroscience*, vol. 22, no. 7, pp. 1504–1529, 2009.

[10] P. Howell and D. J. Powell, "Delayed auditory feedback with delayed sounds varying in duration," *Perception & psychophysics*, vol. 42, no. 2, pp. 166–172, 1987.

[11] K. T. Kalveram and L. Jäncke, "Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition," *Folia Phoniatrica*, vol. 41, no. 1, pp. 30–42, 1989.

[12] J. R. Malloy, D. Nistal, M. Heyne, M. C. Tardif, and J. W. Bohland, "Delayed Auditory Feedback Elicits Specific Patterns of Serial Order Errors in a Paced Syllable Sequence Production Task," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 5, pp. 1800–1821, 2022.

[13] A. J. Yates, "Delayed auditory feedback.," *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[14] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing," *Journal of Neuroscience*, vol. 31, no. 45, pp. 16483–16490, 2011.

[15] S. Cai, M. Boucek, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/," *Proceedings of the 8th ISSP*, pp. 65–68, 2008.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner." Jan. 18, 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[17] D. Bates, M. Maechler, B. Bolker, S. Walker, and others, "lme4: Linear mixed-effects models using Eigen and S4," *R package version*, vol. 1, no. 7, pp. 1–23, 2014.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019. [Online]. Available: https://www.R-project.org/

[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "Package 'lmerTest,'" *R package version*, vol. 2, no. 0, 2015.

[20] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*. 2019. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[21] U. Natke and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 1–8, 2001.

[22] D. Byrd, "Influences on articulatory timing in consonant sequences," *Journal of phonetics*, vol. 24, no. 2, pp. 209–244, 1996.

CV COORDINATION: THE CASE OF ENCHAÎNEMENT AND LIAISON IN FRENCH

Sejin Oh, Cécile Fougeron, Philipp Buech, Anne Hermes

Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle (France)
se-jin.oh@cnrs.fr, cecile.fougeron@sorbonne-nouvelle.fr, philipp.buech@sorbonne-nouvelle.fr, anne.hermes@sorbonne-nouvelle.fr

ABSTRACT

The current study explores how different types of external sandhi in French affect the temporal coordination of gestures. Specifically, we examine the temporal coordination of three different types of CV in French (Onset CV, Enchaînement CV, and Liaison CV) at normal and fast speech rates. The results from this articulatory study revealed that the three different types of CV display the same pattern with respect to their temporal measures (C duration, gestural overlap between C and V, lag between C and V onsets, and lag between C and V targets), at both speech rates. Moreover, the effect of C duration on target-to-target lag is uniform across the different types of CV in French. These results suggest that the CV coordination driven by external sandhi in French exhibits the same coordination (in-phase) as true onset CV.

Keywords: CV coordination, liaison, enchaînement, speech rate, French

1. INTRODUCTION

Syllable structure is hypothesized to be associated with a characteristic pattern of temporal coordination [1-4]. For example, when a gesture occurs in a syllable onset position, it is coordinated in-phase with the following vowel, with the two gestures activated at the same time. In contrast, a coda gesture is hypothesized to be coordinated anti-phase with the preceding vowel, showing a sequential timing between the two gestures. However, a consonant which was an underlyingly coda can also be resyllabified to the following syllable when it is followed by a vowel. Fundamental questions are whether a resyllabified consonant is a 'true' onset, and whether it is shown in the coordination pattern.

In French, there are two different types of resyllabification: 1) an underlying word-final coda consonant can be resyllabified with the following word-initial vowel (Enchaînement CV); 2) a word-final liaison consonant, i.e., a latent consonant surfacing in some French words only when the following word starts with a vowel, can be resyllabified with the following word-initial vowel in what is called 'liaison enchaînée' (Liaison CV). As shown in Table 1, for example, these resyllabification cases and the case with a true word-initial onset consonant (Onset CV) are distinct underlyingly but are said to be homophonous in French.

As with other external sandhi processes, enchaînement and liaison processes have been extensively studied because of their potential cost on word recognition. However, previous perception results have shown that these processes do not generate lexical ambiguities (see e.g., [5, 6]). Acoustic studies have also shown that the phonetic neutralization between forms like the ones presented in Table 1 is incomplete: liaison and enchaînement consonants can be shorter in acoustic duration, the vowel preceding them can be longer, and/or they can preserve some specific allophonic properties associated with their word-final position [5-10]. However, arguments related to the preservation of phonetic contrasts between sequences created by enchaînement or liaison processes and sequences containing underlying onsets, do not rule out the process of resyllabification per se. Indeed, the sequence can preserve some phonetic properties related to its underlying lexical form (with an underlying word-final consonant) and at the same time adopt a 're-syllabified' organization, that could surface in the way the C and the V are temporally coordinated.

Onset	**Enchaînement**	**Liaison**
/CV#**CV**/	/CV.**C#V**/	/CV.**CL#V**/
petit **t**amis /pəti # tami/ [pə.ti.**t**a.mi]	peti**t**e amie /pətit # ami/ [pə.ti.**t**a.mi]	peti**t** ami /pəti # ami/ [pə.ti.**t**a.mi]

Table 1: Examples of three types of CV in French.

An EMA study by [11] examined two types of CVs in Korean, one with a true onset C and the other with a coda C, which turned out to be a resyllabified onset. They reported that there was no difference between the true onset CV and the resyllabified CV, in their coordination with the following vowel and in their stability. However, it might be attributable to

surface ambiguity due to a consistent +/- lag between coordinated landmarks [12]. That is, a lag may cause surface ambiguity between in-phase and anti-phase timing, suggesting that overlap patterns alone may not be enough to determine which landmarks show coordination.

Recently, [13] hypothesized that variation in the duration of the first gesture (G_1), e.g., in CC, influences the G1 onset to G2 onset lag and depends on its coordination relations: for complex segments, such as palatalized /p^j/, the onset of G_2 (the palatal gesture) is temporally coordinated with the onset of G_1 (the labial gesture), while for segment sequences, such as /pj/, the onset of G_2 is temporally coordinated with the offset of G_1. These coordination relations were explored by investigating how the lag between the onset-to-onset varied with G_1 duration. The key finding involved differences between English consonant-glide sequences, e.g., [pj], [bj], [mj], [vj], and Russian palatalized labials, e.g., [p^j], [b^j], [m^j], [f^j], [v^j]. They found that for palatalized consonants in Russian, variation in G_1 duration had little effect on lag, which is consistent with the hypothesized temporal coordination for complex segments. In contrast, for English sequences, as C_1 duration increased, so too did the lag between consonant and glide gestures, showing a strong positive correlation.

Exploiting this temporal diagnosis, the current study aims to understand the temporal coordination of three different CV types (Onset CV, Enchaînement CV, and Liaison CV). The present study asks the following research questions:

- Is there a temporal difference among the three different types of CVs and does this difference hold at different speech rates?
- Do the different types of CVs affect the temporal coordination of gestures?

2. METHODS

2.1. Participants

Three female French native speakers (ages: 27, 42 and 51 years old) who speak Metropolitan French participated in the EMA experiment.

2.2. Speech materials and procedure

Articulatory data was recorded using Electromagnetic Articulography (EMA, AG501) at a sample rate of 1250 Hz and filtered afterwards with a Butterworth low pass filter (cutoff frequency: 20 Hz, order 5). The acoustic signal was recorded with a head-mounted microphone at a sample of 48 kHz and 16-bit resolution. To collect articulatory data, eight sensors were attached to the participants: three sensors midsagittal for tongue movements (tongue tip, blade and dorsum), two for lip movements (upper and lower lips), and two for reference (left/right mastoids). The data was corrected for head movements using the AG501's built-in head-correction procedure.

The data analyzed and presented here were part of a larger study of the temporal organization of gestures in French. Relevant to our purposes here were the three minimal and near-minimal triplets, where each triplet consisted of an Onset CV, Enchaînement CV, and Liaison CV (see Table 2). The target sequences were produced within a carrier sentence (Il y a la carte <X> avant la carte <Y> ici. 'There is an <X> card before the <Y> card here.'), where each sequence presented in the current study was located in <X> in the sentence.

Onset /CV#CV/	**Enchaînement /CV.C#V/**	**Liaison /CV.CL#V/**
un petit tamis	une petite amie	un petit ami
un petit tapis	une petite abyme	un petit habit
un maudit tanneurs	une maudite année	huit années

Table 2: Stimulus items.

To test the stability of our coordination measures against variation in speech rate, participants were asked to read aloud each given sentence at their normal rate and then self-selected fast speaking rates in two successive blocks. Each sentence was produced 7 times per speaker at each rate. Of the 378 tokens collected, 22 tokens (5.8%) were discarded.

2.3. Data analysis

WebMAUS v3.12 [14] was used for acoustic alignments. Afterwards, the onset, target, release, and offset of the consonantal gesture for /t/ (TT sensor) and of the vocalic gesture for /a/ (TD sensor) were automatically identified using a custom script written in Python v3.10.8 [15]. These landmarks were detected within a window of the acoustic boundaries ±100 ms of [t] and of the sequence [ta] for the vocalic gesture, using 20% threshold of the maximum and minimum peak velocity (target, release) and the first zero crossing of the velocity before/after the peak velocities (onset, offset). As shown in Figure 1, the four key temporal intervals computed from these articulatory landmarks were (1) C DURATION = $C_{RELEASE} - C_{TARGET}$ (consonant plateau duration); (2) OVERLAP = $C_{RELEASE} - V_{ONSET}$; (3) ONSET LAG = $V_{ONSET} - C_{ONSET}$; (4) TARGET LAG = $V_{TARGET} - C_{TARGET}$. We also analyzed the correlation between C DURATION and TARGET LAG.

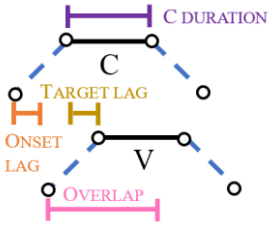


Figure 1: Four key temporal intervals: (1) C DURATION, (2) OVERLAP, (3) ONSET LAG, (4) TARGET LAG

2.4. Statistical models

The *brms* package v2.18.0 [16] in R v4.2.2 [17] was used for the statistical analysis. Bayesian linear mixed models were performed for C DURATION, OVERLAP, ONSET LAG, and TARGET LAG. The model structure was the same for all response variables, that is CV TYPE (Onset CV, Enchaînement CV, and Liaison CV), SPEECH RATE (normal, fast) and their interaction CV TYPE*SPEECH RATE as fixed parameters, and SPEAKER and ITEM as random intercepts. An additional model, for TARGET LAG as the outcome variable, the interaction between C DURATION and CV TYPE as fixed factors and SPEAKER and the ITEM as random intercepts, was performed. All models were run with standard priors and 4 chains with 6,000 samples each from which 3,000 were for warm-up, making a total of 12,000 post-warmup samples. Strong evidence for an effect being non-zero was determined if 0 was not included in the 95% Credible Interval (CI) of the posterior distribution. We report the posterior means (β) and the CI in square brackets.

3. RESULTS

3.1. Temporal differences

Figure 2 shows the distribution of C DURATION (A), OVERLAP (B), ONSET LAG (C), and TARGET LAG (D) across CV TYPE. Normal and fast speech rates are shown on the left and right panels, respectively. The three types of CV exhibit the same pattern for all four temporal measures for both speech rates.

First, the statistical analysis from Bayesian regression models reveals that there was no evidence for C DURATION of Enchaînement CV or Liaison CV being different from that of Onset CV (Enchaînement β=-1.45 [-5.79, 2.89]; Liaison β=-1.71 [-6.03, 2.59]). Moreover, although we found evidence for C DURATION being shorter for a faster speech rate compared to a normal speech rate (β=-10.21 [-13.59, -6.88]), we have no evidence for an effect of the interaction of speech rate and CV type on C duration (Enchaînement_fast β=2.64 [-2.06, 7.33]; Liaison_fast β=3.79 [-0.90, 8.60]). That is, within each speech rate, C DURATION of Enchaînement CV or Liaison CV is not systematically different from that of Onset CV.

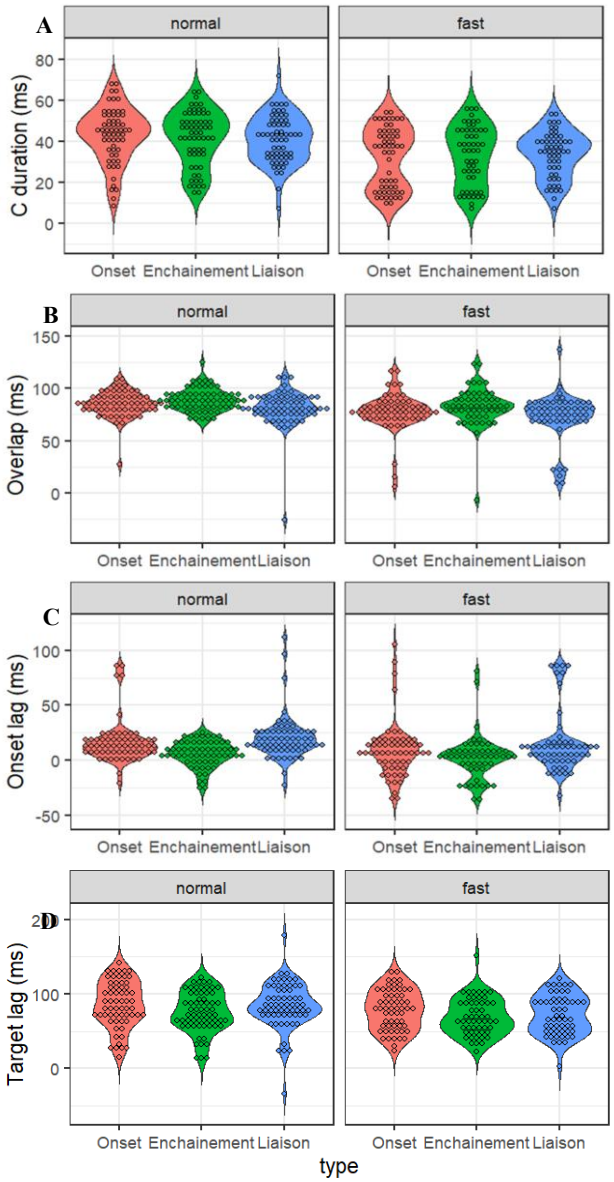


Figure 2: C duration (A); Overlap (B) and Onset lag (C); Target lag (D) at normal speech rate (left column) and at fast speech rate (right column).

We found the exact same pattern for OVERLAP and ONSET LAG. The statistical results reveal that different types of CV are consistent in their OVERLAP (Enchaînement β=3.85 [-4.31, 11.76]; Liaison β=-3.85 [-11.63, 4.04]) and in their ONSET LAG (Enchaînement β=-10.51 [-22.50, 1.54]; Liaison β=4.99 [-6.09, 16.45]). In addition, a faster speech rate also exhibits shorter OVERLAP and ONSET LAG compared to a normal speech rate (OVERLAP β=-7.12 [-13.02, -1.42]; ONSET LAG β=-7.81 [-14.56, -1.14]). However, we have no evidence for an impact of the interaction of speech rate and CV type on OVERLAP (Enchaînement_fast β=-0.22 [-8.07, 8.48];

Liaison_fast β=-1.76 [-10.11, 6.59]), and so is an effect of the interaction of speech rate and CV type on ONSET LAG (Enchaînement_fast β=5.75 [-3.64, 15.19]; Liaison_fast β=4.86 [-4.68, 14.33]).

Regarding TARGET LAG, Enchaînement CV or Liaison CV is not systematically different from that of Onset CV (Enchaînement β=-10.55 [-30.03, 8.68]; Liaison β=-1.61 [-20.48, 17.61]). Also, no evidence was found that normal and fast speech rates were systematically different (β=-6.05 [-15.30, 3.30]). Moreover, we have no evidence for an effect of the interaction of speech rate and CV TYPE on TARGET LAG (Enchaînement_fast β=0.34 [-12.74, 13.52]; Liaison_fast β=-4.56 [-18.04, 8.71]).

3.2. Gestural coordination of CVs

Figure 3 plots the relation between C DURATION (pooled duration at both rates; x-axis) and TARGET LAG (y-axis) across CV TYPE. Changes in C DURATION have little effect on TARGET LAG, showing only a slight upward trend. Notably, we observe the same pattern for all CV TYPE. To assess these results, we fit a series of Bayesian regression models to the data. We found evidence that variation in C DURATION impacts TARGET LAG, but the impact is small (β=0.82 [0.50, 1.15]). Crucially, however, we also found evidence that the way that variation in C DURATION impacts TARGET LAG is uniform across different types of CV in French (Cdur_Enchaînement β=-0.25 [-0.68, 0.18]; Cdur_Liaison β=0.40 [-0.09, 0.89]). To sum up, the results indicate that Enchaînement CV and Liaison CV showed the same temporal coordination as Onset CV.

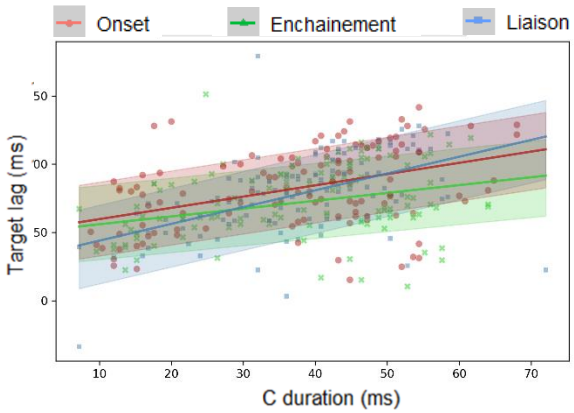


Figure 3: Target lag (ms) by C duration (ms) across CV types.

4. DISCUSSION & CONCLUSION

The present study investigated the temporal coordination between C and V in three different types of CV in French (Onset CV, Enchaînement CV, and Liaison CV) using kinematic data from Electromagnetic Articulography (EMA). We found that all CV types display the same pattern in their C duration, the overlap between C and V, the lag between C and V onsets, and the lag between C and V targets. Although, except for TARGET LAG, we found shorter duration on the other three measures for a faster speech rate compared to a normal speech rate, the interaction between CV types and speech rate on these measures was not significant. This suggests that Enchaînement CV and Liaison CV are not significantly different from Onset CV within each speech rate. Furthermore, we found a small impact of C DURATION on TARGET LAG. As [12] noted, this may be attributable to local variation in speech rate that independently influences both C duration and the lag. Crucially, however, the effect of C DURATION on TARGET LAG is uniform across different types of CV. This suggests that Enchaînement CV as well as Liaison CV in French show the same in-phase timing as a true Onset CV.

The results of the present study provide evidence that resyllabified consonants are also timed in-phase with the following vowel, as Cho and colleagues [11] have found for Korean. This resyllabification across words raises many questions—to be resolved—on the process of inter-gestural coordination. If the coordination between gestures within a syllable is specified in a word-size gestural score, how could it be modified in running speech [18]? Reversely, should this so-called resyllabification be considered to be a simple syllabification, surfacing from the coordination of gestures planned over units spanning several connected words? If so, how can we account for the results of previous studies which reported that speakers, as well as listeners, preserve phonetic contrasts among sequences with lexical onsets and sequences with external sandhi in French (cf., introduction)? One of the more frequently reported acoustic differences is the shorter acoustic duration of enchaînement C and liaison C compared to a true onset C. While acoustic measurements remain to be tested on this data, in the current study where temporal cues are tested on a wide set of articulatory events, we did not replicate this difference: the three types of CV are similar in their constriction duration, as well as their temporal lags and gestural overlap.

Obtaining kinematic data from EMA allowed us to examine the temporal coordination of gestures involving the articulation of external sandhi in French. Moreover, we make use of the concept of coordination to relate speech kinematics to the syllable structure. In future work, we will further investigate the temporal relations of C with the preceding vowel, which was also reported to be one of the acoustic characteristics preserving the contrast between the three sequence types.

5. ACKNOWLEDGEMENT

This research was partly funded by ChaSpeePro (CRSII5_202228) of the Swiss National Science Foundation and by the French ANR (ANR-10-LABX-0083, LABEX-EFL).

6. REFERENCES

[1] Goldstein, L., Byrd, D., & Saltzman, E. 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. *In: Arbib M, editor. From action to language: The mirror neuron system. Cambridge: Cambridge University Press; 2006. pp. 215–249.*

[2] Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. 2009. Coupled oscillator planning model of speech timing and syllable structure. In G. Fant, H. Fujisaki & J. Shen (Eds.) *Frontiers in phonetics and speech science: Festschrift for Wu Zongji* (pp. 239–249). Beijing: Commercial Press.

[3] Nam, H., Goldstein, L., & Saltzman, E. 2009. Self-organization of syllable structure: A coupled oscillator model. *Approaches to phonological complexity* (pp. 297-328). Berlin/New York: Mouton de Gruyter.

[4] Saltzman, E., Nam, H., Krivokapić, J., & Goldstein, L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008), Campinas, Brazil,* 175-184.

[5] Gaskell, M., Spinelli, E., and Meunier, F. 2002. *Perception of resyllabification in French. Memory and Cognition, 3*0, 798–810.

[6] Spinelli, E., McQueen, J., and Cutler, A. 2003. Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233–254.

[7] Rialland, A. 1986. Schwa and syllables in French. *Studies in compensatory lengthening*, *23*, 187.

[8] Fougeron, C., Bagou, O., Stefanuto, M., & Frauenfelder, U. H. 2003. Looking for acoustic cues of resyllabification in French. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPHS)*.

[9] Fougeron, C. 2007. Word boundaries and contrast neutralization in the case of enchaînement in French. *Papers in laboratory phonology IX: Change in phonology*, 609-642.

[10] Nguyen, N., Wauquier-Gravelines, S., Lancia, L., & Tuller, B. 2007. Detection of liaison consonants in speech processing in French: Experimental data and theoretical implications. *In Pilar Prieto. Segmental and Prosodic Issues in Romance Phonology, John Benjamins, pp.3-23, 2007, Current Issues in Linguistic Theory.*

[11] Cho, T., Yoon, Y., & Kim, S. 2014. Effects of prosodic boundary and syllable structure on the temporal realization of CV gestures. *Journal of Phonetics*, *44*, 96-109.

[12] Shaw, J. A., & Gafos A. I. 2015. Stochastic time models of syllable structure. *PloS one, 10*(5), e0124714. doi:10.1371/journal.pone.0124714

[13] Shaw, J. A., Oh, S., Durvasula, K., & Kochetov, A. (2021). Articulatory coordination distinguishes complex segments from segment sequences. *Phonology, 38*(3), 437-477. doi:10.1017/S0952675721000269.

[14] Kisler, T., Reichel, U., & Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326-347.

[15] G. V. Rossum and F. L. Drake. 2009. *Python 3: Reference manual*. United States: SohoBooks.

[16] Bürkner, P.-C. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

[17] R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[18] Kohler, K.J. "Gestural Reorganization in Connected Speech: A Functional Viewpoint on 'Articulatory Phonology'" *Phonetica*, vol. 49, no. 3-4, 1992, pp. 205-211. https://doi.org/10.1159/000261916

THE NATURE OF ACOUSTIC GOALS REFERENCED DURING HYPERARTICULATED SPEECH

Carissa A. Diantoro & Melissa A. Redford

Department of Linguistics, University of Oregon, Eugene, OR 97403-1290 USA
carissad@uoregon.edu & redford@uoregon.edu

ABSTRACT

Hyperarticulation references goals defined in acoustic-perceptual space. The question is whether hyperarticulation enhances contrast between goals in this space or exaggerates salient aspects of the goal without reference to its position in space. To address this question, CVC minimal triplets with English front lax vowels, /ɪ, ɛ, æ/, were elicited from 30 participants during the first and last phase of a 3-phase experiment to map the front lax vowel space. In the middle phase, participants produced only words with /ɛ/ vowels. These were "misheard" by the experimenter either as /ɪ/ or /æ/ vowels. Measurement on corrected repetitions indicated that participants fronted corrected /ɛ/ relative to baseline but did not systematically lower or raise it in response to the misheard vowel quality; corrected /ɛ/ was louder relative to baseline only in the /ɪ/ condition. The mixed results are discussed with reference to the nature of acoustic goals.

Keywords: speech motor goals, speech production, speaking style.

1. INTRODUCTION

Lindblom and colleagues [1] demonstrated that speakers achieve the same formant frequencies for different vowels, /i, u, ɑ, o/, whether their jaw varies naturally with articulation or is fixed at a particular distance relative to the maxilla. Moreover, compensation for the articulatory perturbation is evident as soon as the first glottal pulse of vowel articulation. Lindblom and colleagues interpreted these results to suggest predictive control over speech articulation and so the existence of acoustic-perceptual motor goals. Many other perturbation studies since then have expanded on these findings [2]. Results from auditory feedback perturbation studies provide especially compelling support for the hypothesis that speech articulation is guided by acoustic perceptual motor goals [e.g., 3, 4, 5].

Auditory feedback perturbation studies use an elegant design to test the effect of self-monitoring via perceptual channels on motor goal attainment. We have borrowed and adapted the design here to investigate the effect of feedback from others (i.e., misperception) on production, using the effect to investigate whether acoustic-perceptual motor goals reference discrete phonological representations of sounds, as is typically supposed [e.g., 1, 2, 3, 4], or whether they are dispersed in an extra-linguistic multi-dimensional perceptual space that is used to control articulation, as proposed in [5]. The distinction is important because it has implications for a whole-word production hypothesis [6], which has been offered as a developmentally sensitive alternative to the psycholinguistic hypothesis of phonological encoding during speech production [e.g., 7]. In what follows, we present the logic behind our approach for addressing the study aim.

It is well established that when a listener mishears a speaker, the speaker adjusts their articulation to better communicate that which was intended. In the clear speech literature, this adjustment has come to be known as hyperarticulation [8]. Hyperarticulation is very often supposed to represent a kind of goal maximization [8, 9] and so it has been studied to infer something about the representations that guide speech articulation [e.g., 9, 10, 11]. In clear speech studies, hyperarticulation is interpreted to suggest not only acoustic-perceptual speech motor goals but also that such goals reference discrete phonological representations of sound. For instance, Johnson and colleagues [9] describe a perceptual "hyperspace" effect, linking listener expectations for maximally distinct vowel productions to the speaker's speech motor goals in production. They then explicitly link such goals to discrete phonological representations of sounds, noting, for example, that "the search for acoustic and articulatory correlates of phonological units, in order to be successful, should focus on carefully produced speech (pp. 525-6)."

The linking of phonological units and speech motor control predicts that hyperarticulated speech results in contrast enhancement [e.g., 10, 11], but so far this prediction has not been tested against the prediction from the alternative extra-linguistic hypothesis. Studies specifically designed to investigate contrast enhancement (versus global enhancement) have focused on 2-way contrasts [10, 11]. Yet, the robust evidence for targeted hyperarticulation presented in these studies need not imply linguistic contrast enhancement per se; it could simply reflect the speaker's ability to increase the

acoustic-perceptual distance between discrete speech motor goals in an extra-linguistic acoustic-perceptual space. To distinguish between these alternatives, a more rigorous test of the linguistic contrast enhancement hypothesis is required. The current study provides this test. To do this, we borrow and adapt a study design from the auditory feedback perturbation paradigm to investigate the effect of hyperarticulation on production given a 3-way contrast: Houde and Jordan [3] asked speakers to whisper CVC words with an /ɛ/ vowel. Formant frequencies for this vowel were shifted (or not) and presented over earphones. The shift (= perturbation) moved the /ɛ/ formants either towards the speaker's own /i/ vowel or towards the speaker's own /a/ vowel. Speakers compensated in real-time for the perturbation by shifting their production of /ɛ/ away from either /i/ or /a/, depending on the perturbation. More specifically, if the auditory feedback perturbation suggested that the speaker's /ɛ/ was moved towards the speaker's /i/ and so was confusable with /ɪ/, speakers adjusted their production so that /ɛ/ became more [æ]-like; if the /ɛ/ production was moved in the direction of /a/ and so became confusable with /æ/, it became more [ɪ]-like.

In the current study, we use overt feedback from a listener about the category with which their /ɛ/ production is being confused. If speakers respond as they do in auditory feedback perturbation studies by adjusting intended /ɛ/ away from the misheard vowel, then we will have obtained strong evidence for the connection between speech motor goals and contrastive (i.e., phonological) speech sound representations. If not, then the possibility that such goals are extra-linguistic remains a viable alternative hypothesis.

2. METHODOLOGY

2.1. PARTICIPANTS

Participants were college-aged (18 to 22 years old) adults recruited via word-of-mouth through friend networks. Thirty participants completed the study, all reported normal hearing and typical speech-language development (none had a history of speech-language therapy); a pure-tone hearing screen test confirmed normal hearing.

2.2. MATERIALS

Materials were CVC minimal triplet words with the English front lax vowels, /ɪ, ɛ, æ/. There were 10 triplets. Table 1 lists the minimal triplet words.

/ɪ/	/ɛ/	/æ/
bid	bed	bad
bit	bet	bat
sid	said	sad
sit	set	sat
kid	ked	cad
kit	ket	cat
lid	lead	lad
lit	let	lat
rid	red	rad
writ	rhett	rat

Table 1: List of minimal triplets used in the study.

Each word was paired with a picture for elicitation purposes. Pictures were cartoon-like representations of the objects/events denoted by the CVC word. These were printed in color and laminated to create picture cards. The backside of each card was labelled with the target word. Nonce words were treated as proper names (i.e., Sid, Ket, Rhett) and paired with a cartoon character. Verbs (e.g., sit, said, sat) were paired with events. Adjectives (e.g., red, rad) were paired with objects, where the attribute of the object was highlighted during word-picture training. All picture–word correspondences were easily learned by all participants during a brief training that preceded the elicitation task. If a picture–word correspondence was forgotten during the task, the card was flipped over and shown to the speaker as a written reminder of the correspondence.

2.3. PROCEDURE

The elicitation task took place in adjacent sound-attenuated experimental rooms. The conjoining wall had a window. The experimenter and participant sat facing one another on either side of the window. This set up was used to enhance the plausibility of the mishearing manipulation: Although the participant was able to hear the experimenter over a baby monitor, they could not if the baby monitor was turned off. The participant could therefore easily imagine then how it might be difficult for the experimenter to hear them. The set up also allowed us to collect in person speech data while obeying pandemic-related restrictions on said data collection.

Elicitation occurred in three phases: a mapping phase, a misperception phase, and a remapping phase. The experimenter used the pictures to elicit all 30 words twice in different fixed random orders during the mapping and remapping phase. During the misperception phase, only /ɛ/ words were elicited and the experimenter "misheard" the participant as having produced either the matched [ɪ]-word ("ih" condition) or the matched [æ]-word ("ae" condition). The

participant was obliged to correct the experimenter by repeating the [ɛ]-word. The correction was done twice in a row on each trial (i.e., for each [ɛ]-word); each [ɛ]-word was also elicited twice during the misperception phase, again in random order. As in auditory feedback perturbation studies, the specific misperception manipulation was between subjects: half of the participants were assigned to the "ih" condition and half to the "ae" condition.

Participant speech was digitally recorded with a sampling rate of 44,100 Hz. A lavaliere microphone was attached to the participant's shirt to maintain a fixed mic-to-mouth distance.

2.4. ACOUSTIC MEASURES

Acoustic segmentation and measurement were completed on speech elicited from the 15 participants in the "ih" condition and the 15 in the "ae" condition. Vowels were segmented and labelled for each CVC word by a trained research assistant using Praat [12]. Vowel identity was determined with reference to the intended CVC word, based on the fixed random order used during the picture-based elicitation task. Overall vowel duration was automatically extracted from the segmentations. F1, F2, and vowel amplitude were automatically extracted at 3 equal intervals around vowel midpoint; specifically, at points equal to 40%, 50%, and 60% of the vowel duration. Formant frequency values were based on the automatic formant tracking algorithm used in Praat. The tracks were inspected by hand vowel by vowel against the spectrogram. Tracking parameters were adjusted where necessary (e.g., if F3 was tracked in lieu of F2).

2.4. ANALYSES

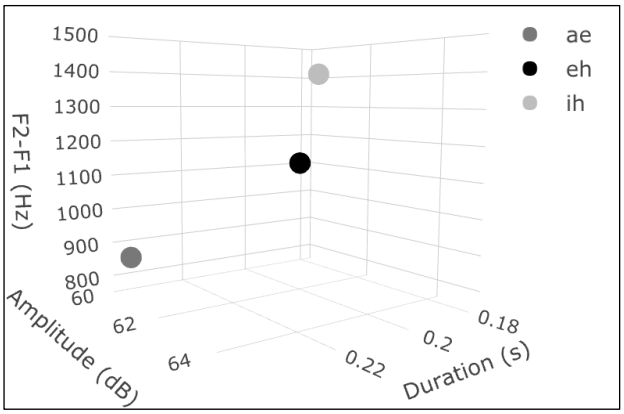
Mixed effects models were built using the lme4 package [13] in R [14] to test for the fixed effects of vowel, elicitation phase, and/or misperception condition while controlling for random effects of word and speaker, where appropriate. P-values were calculated using Satterthwaite's Method [15] with the lmerTest package [16] in R studio. Dependent measures were either raw measurement values or normalized values. Normalized values were used in the critical analyses that tested for the fixed effect of condition ("ih" versus "ae") and manipulation ("correction" versus "control") on /ɛ/ production. To compute these values, F1 and F2 frequencies, vowel amplitude, and vowel duration were each averaged for each word within each phase within each speaker. Word-to-word difference values were then computed by subtracting mapping phase values from misperception phase values (= correction) or from re-mapping phase values (= control).

3. RESULTS

3.1. FRONT LAX VOWEL SPACE

Figure 1 shows the position of the front lax vowels in acoustic space based on elicitations from the mapping phase of the elicitation task. Results from a mixed effects model on raw measurement values confirmed that the vowels differed from one another along all dimensions shown: [F2-F1 model intercept is "ae" at 853.55, SE = 16.95, $p < .001$; "eh" $\beta = 264.03$, SE = 14.60, $p < .001$; "ih" $\beta = 558.40$, SE = 21.48, $p < .001$. Amplitude model intercept is "ae" at 60.01, SE = 1.73, $p < .001$; "eh" $\beta = 0.73$, SE = 0.21, $p < .01$; "ih" $\beta = 0.42$, SE = 0.35, $p = > .05$. Duration model intercept is "ae" at 0.24, SE = 0.008, $p < .001$; "eh" $\beta = -0.05$, SE = 0.003, $p < .001$; "ih" $\beta = -0.06$, SE = 0.004, $p < .001$].

Figure 1: The 3 lax vowel targets in acoustic space. Error



bars indicate 95% confidence interval.

As expected, the vowel in /ɛ/ words occupied a position in acoustic space that was intermediate to the vowels in /ɪ/ and /æ/ words. Also, intended /ɛ/ was closer in acoustic space to intended /ɪ/ compared to /æ/. This relative distance between intended vowels in space was also as expected.

3.2. EFFECT OF MISHEARING ON /Ɛ/ PRODUCTION

Normalized values were used to test for the critical fixed effects of elicitation phase (correction vs. control) and misperception condition ("ih" vs. "ae"). The models thus included only a random slope and intercept for word. Analyses indicated a significant interaction between phase and condition on F1 [$\beta = 22.19$, SE = 6.32, $p < .001$]: F1 was raised in response to misperception, but only in the "ih" condition [$\beta = 25.06$, SE = 4.54, $p < .001$]. The interaction was not significant on F2. Instead, there was only a main effect of phase in both the "ih" condition [$\beta = 62.38$, SE = 10.23, $p < .001$] and "ae" condition [$\beta = 42.75$, SE = 10.99, $p < .001$]. There was no simple effect of misperception condition on /ɛ/. Importantly, whether

/ɛ/ was misheard as /ɪ/ or as /æ/, participants adjusted by raising F2. Misperception also had an effect on /ɛ/ amplitude and duration, but only in the "ih" condition [amplitude: β = 5.63, SE = 0.53, p <.001; duration: β = 0.03, SE = 0.007, p <.001]. The effect of misperception on /ɛ/ can be seen in Figure 2, which shows the position of the participants /ɛ/ vowel in acoustic space as a function of misperception condition; its default position, based on elicitations obtained during the mapping phase of the experiment, is also shown.

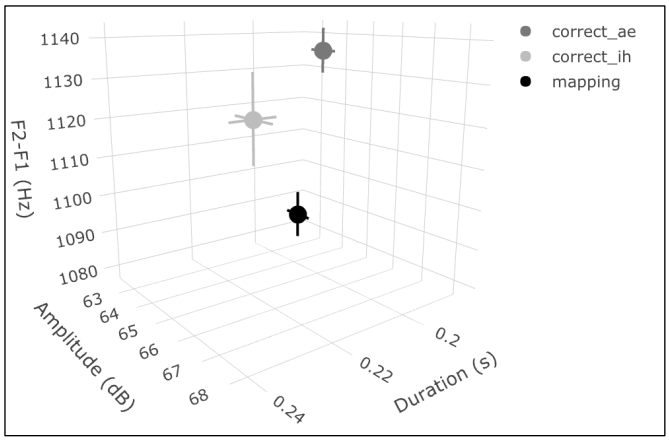


Figure 2: Mid-front lax vowel target shown as a function of misperception condition ("ih" vs. "ae"). The default position of the same vowel is also shown. Error bars indicate 95% confidence intervals.

4. CONCLUSION

Hyperarticulation references goals defined in acoustic-perceptual space. The question addressed in this study was whether this goal incorporates paradigmatic information relevant to linguistic contrast. The results were that, whether the experimenter misheard intended /ɛ/ as /ɪ/ or as /æ/, participants corrected the misperception in the same way – that is, they adjusted their articulation so that the vowels they produced varied in the same direction away from the control vowel along the measured dimensions (vowel formants, amplitude, duration). This result is inconsistent with the prediction of linguistic contrast enhancement. When differences by misperception condition were found, these indicated that /ɪ/ misperception elicited a greater degree of articulatory adjustment to /ɛ/ than /æ/ misperception. This result suggests an implicit awareness that [ɛ] is more apt to be confused with [ɪ] than with [æ] due to the relative positions of these sounds in acoustic space. Note that such a suggestion does not require speech motor goals to reference phonemic contrast; it requires only that the participant note the acoustic similarity between the experimenter's CVC word production due to misperception (e.g., "bid") and the intended CVC word that the participant had produced (e.g., "bed"), and then seek to further differentiate their own production from that of the experimenter's mistaken production.

5. ACKNOWLEDGMENTS

This research was wholly supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) under grant R01HD087452 (PI: Redford). The content is solely the authors' responsibility and does not necessarily reflect the views of NICHD.

6. REFERENCES

[1] Lindblom, B., Lubker, J., Gay, T. 1979. Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *J. Phonetics*, vol. 7, no. 2, pp. 147-161.

[2] Perrier, P., Fuchs, S. 2015. "Motor equivalence in speech production," In: Redford, M. A. (eds), *Handbook of Speech Production*. Wiley, 225–247.

[3] Houde, J. F., Jordan, M. I. 2002. Sensorimotor Adaptation of Speech I: Compensation and Adaptation. *J. Speech Lang. Hear. Res.*, vol. 45, pp. 295-310.

[4] Purcell, D. W., & Munhall, K. G. 2006. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J. Acoust. Soc. Am.*, vol. 120, pp. 966-977.

[5] Davis, M., Redford, M. A. 2019. The emergence of discrete perceptual-motor units in a production model that assumes holistic phonological representations. *Front. Psychol.*, vol. 10, pp. 2121.

[6] Redford, M. A. (2019). Speech production from a developmental perspective. *J. Speech Lang. Hear. Res.*, vol 62, pp. 2946-2962.

[7] Levelt, W. J. 1989. Speaking: From intention to articulation. MIT press.

[8] Lindblom, B. 1990. "Explaining phonetic variation: A sketch of the H&H theory," In: Hardcastle, W.J., Marchal, A. (eds) *Speech Production and Speech Modelling.* Springer, Dordrecht. pp. 403-439.

[9] Johnson, K., Flemming, E., Wright, R. 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, pp. 505-528.

[10] Schertz, J. 2013. Exaggeration of featural contrasts in clarifications of misheard speech in English. *J. Phonetics,* vol. 41, no. 3–4, pp. 249–263.

[11] Wedel, A., Nelson, N., & Sharp, R. 2018. The phonetic specificity of contrastive hyperarticulation in natural speech. *J. Mem. Lang.,* vol 100, 61-88.

[12] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.37, retrieved 16 December 2020.

[13] Bates D, M.chler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.,* vol. 67, no. 1, pp. 1–48.

[14] RStudio Team. 2021. RStudio: Integrated Development Environment for R. RStudio.

[15] Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.*, vol 2, pp. 110-114.
[16] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. J. *Stat. Softw*., vol. 82, pp. 1-26.

MONOLINGUAL AND PLURILINGUAL STRATEGIES IN THE ARTICULATION OF FRENCH R: A CASE STUDY

Andres Felipe Lara, Anne Hermes, Sejin Oh, Claire Pillot-Loiseau

Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle, Paris
andres.lara@sorbonne-nouvelle.fr

ABSTRACT

This case study describes lingual articulatory strategies for French [ʁ] as it pertains to a monolingual, native speaker of French, and two English-Spanish bilinguals, learners of French as their third language (L3). We tracked movements of the tongue using electromagnetic articulography (EMA) in order to investigate inter-speaker differences involving rhotic production in word-initial and final positions. Curve trajectories were compared using generalized additive models (GAMs). Findings indicate active involvement of tongue tip in rhotic articulation, native speaker included. Raising of tongue dorsum in language learners points towards a velar constriction, not uvular. Varying articulation strategies by language learners are interpreted as an indication of language interference. Duration differences and curve complexity as indices of native-like degree of coarticulation were closely linked to the learner's amount of exposure to the target language.

Keywords: EMA, language learner, rhotics, French, coarticulation.

1. INTRODUCTION

Literature on articulatory descriptions of uvular fricatives has been relatively scarce. Moreover, only a few works have attempted to describe the French uvular rhotic [1, 2]. As a result, gestural configurations of these sounds are still poorly understood. The uvular fricative [χ, ʁ] is characterised by the coordination of a tongue body raising gesture and tongue root retraction toward the pharyngeal cavity [3]. Data also suggests that the raising of the tongue body only accounts for 58 percent of uvular constriction. According to Gick et al. [4], velar lowering makes up for the remaining 42 percent. In comparison, tongue tip movements in the articulation of uvular fricatives are minimally described and thus considered to have a minimal role. Nevertheless, ultrasound recordings of uvular fricatives in Upper Sorbian [3] documented less raised tip/blade in /a/ environments. The goal of this study is to describe lingual articulatory strategies for French [ʁ] as it pertains to a monolingual, native speaker of French, and two English-Spanish bilinguals, learners of French as an L3 with varying times of exposure to the target language.

The Spanish phonological inventory has two rhotic phonemes used contrastively in the intervocalic position only: a tap and a trill [5]. Tongue root retraction has been observed for both tap and trill [2, 6]. The level of retraction appears to be more prominent for [ɾ] than for [r] [7]. Since [ɾ] is the only rhotic sound that appears in word-final position in Spanish, a Spanish-dominant bilingual would be expected to retract more in word-final position. The tongue tip place of articulation is described as alveolar but height can vary due to vocalic context. In /a/ environments, the tongue tip is slightly lower but canonically alveolar as opposed to other closed vowels where tongue-tip position is higher and slightly more posterior [7].

The American English /ɹ/ has been studied extensively due to the idiosyncratic nature of its articulatory configurations and a complexity that lies in the coordination of three independent gestures [8-11]. Production of the English rhotic involves retraction of the tongue root towards the pharyngeal wall, raising of the tongue tip/body toward the alveolar ridge/palate (retroflex/bunched), and rounding of the lips. Campbell et al. [12] found that the magnitude of constriction for the tongue root/blade varied according to syllabic position with reduction in tongue body retraction in word-final position and emphasis of tongue root retraction in word-initial position. Thus, an English-dominant bilingual would show less retraction in word-final position.

French /ʁ/ occupies the same phonological category as its Spanish and American English counterparts but articulatory gestures vary depending on the language. A study [13] found that bilinguals, perceived as native speakers in both languages, employed distinct language-specific articulatory settings. In order for that to happen, late learners would have to correctly identify allophones of the same category phonologically and phonetically [14]. A couple of previous studies have concentrated on coarticulation differences in native vs. non-native speakers [15, 16]. Both studies concluded that more experienced learners developed more native-like degrees of coarticulation than less experienced

learners did. Moreover, high frequency words tend to have shorter durations and fewer segments [17, 18]. Words chosen for this case study are considered high-frequency thus resulting in smoother curves, an indication of native-like coarticulation degrees. We expect late learners in this case study to employ French articulatory gestural strategies and coarticulation degrees according to their level of exposure. [19] posits that the quantity and quality of L2 input that learners receive influence the long-term learning of L2 speech, or L3 in this case.

2. METHODOLOGY

Three speakers participated in this case study. One female 52 y.o. native speaker of French from the Paris region with very little exposure to English or Spanish (FR), one male 34 y.o. native speaker of Spanish, exposed to American English before the age of 5 (SP), and one female 30 y.o. native speaker of American English, exposed to Spanish after the age of 10 (EN). The language learners were first exposed to French in a formal setting after the age of 13 but differ in times of naturalistic exposure to the target language. SP has been living in France for six years while EN has only been in France for six months. Language levels were not formally evaluated but it's worth nothing that SP was at a C1 level upon entering the country. EN was less fluent, but still able to carry out a conversation in French.

Participants were asked to repeat the words '*rat*' and '*par*' three times. Words were placed in the carrier sentence '*Ils disent ___ sept fois*' ('They said ___ seven times'). These are two relatively high-frequency words in French with a rhotic consonant in word-initial and word-final position.

Articulatory data was recorded using electromagnetic articulography (EMA; AG 501). Displacement measurements in millimetres were recorded along the horizontal (front-back) and vertical (high-low) axis with four captors placed along the midsagittal plane of the tongue as follows: one captor as far back as possible on the tongue dorsum (TD), two middle captors 1 cm apart, one near the back (TB) and one near the front (TF), one last captor on the tip of the tongue (TT). Two captors were placed on the mastoid processes for spatial reference and in order to correct for deviations in head position during the post-processing stage of the data [20].

Displacement trajectories were visualized and extracted using EMA2WAV [21]. Annotations were done manually on Praat and included gesture offset for the first alveolar consonant preceding the target word and gesture onset for the second alveolar consonant following the target word. Trajectory curves were drawn from measurements taken at 100 points at regular intervals within the target word. The mgcv package [22] was used to analyse for significance amongst speakers. A generalized additive model (GAM) model with a random effect was fit using the formula *captor ~ speaker + s(normalized time, by = speaker) + s(speaker, bs = "re")*. The smooth components are centred around zero with a varying intercept along interpolated normalized times [23, 24]. Variability is shown by overlaying counter plots.

3. RESULTS

Curves were compared by fitting GAMs and partial effects are plotted independently as component effects of the smooth (time) and linear terms (speakers) in the model. This also allowed us to accurately portray rhotic segment trajectories (column 1) and rhotic gesture durations (dotted lines) for each of the speakers based on individual averages. Column 4 shows comparisons between speakers with significant differences marked in red (column 4). Lines correspond to the captors used to measure displacement. Standard errors show 95 percent interval for the mean shape of the effect.

3.1. R in word-initial position

As shown in Fig. 1, TT gesture onset for SP is higher than FR and EN. FR and SP show similar articulatory strategies by lowering both the back (TD and TB) and the front of the tongue (TF and TT). EN shows a different strategy by raising the back of the tongue and then gradually lowering it for the vowel gesture. TF and TT are lowered by all speakers.

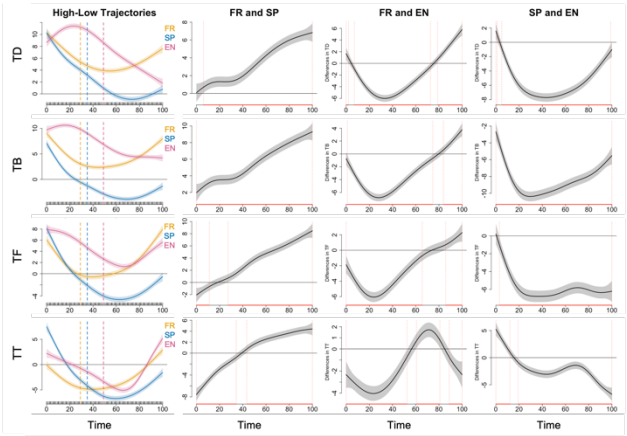


Figure 1: Column 1 shows high-low displacement in mm for the word 'rat'. Dotted lines: start of rhotic gestures per speaker. Columns 2 – 4: significant differences between speakers (in red). Drops in trajectory correspond to tongue lowering.

The lowest point in the curve for TD and TB corresponds to the low vowel gesture at maximum constriction. TD and TB curves for FR show the least amount of change. SP shows the lowest dip amongst the speakers. Curve trajectories appear more complex for EN than for FR and SP.

ANOVA revealed significant values for the effect of speaker on the smooth of the spline for TD, TB, TF, and TT (mean edf = 7.361, p<2e-16). Partial significance between speakers was found at gesture onset for TD. Analysis also showed partial significance for TF differences between FR and SP. TD, TB, and TF rhotic gestures are higher for EN than for FR and SP. Vowel gestures are lower for SP than for FR and EN. There is less of a difference between FR and SP's rhotic gestures as it concerns TD, TB, and TF. See fig 1 column 2.

Front-back displacement for the word 'rat' is plotted in Fig 2. SP and FR show similar articulatory strategies (see curve trajectories for TF and TT). Tongue retraction is present in both SP and FR but gesture transitions are less noticeable for FR. SP advances TD slightly for the vowel gesture. EN's articulation of the rhotic segment differs from the other speakers: the speaker advances the tongue instead of retracting it.

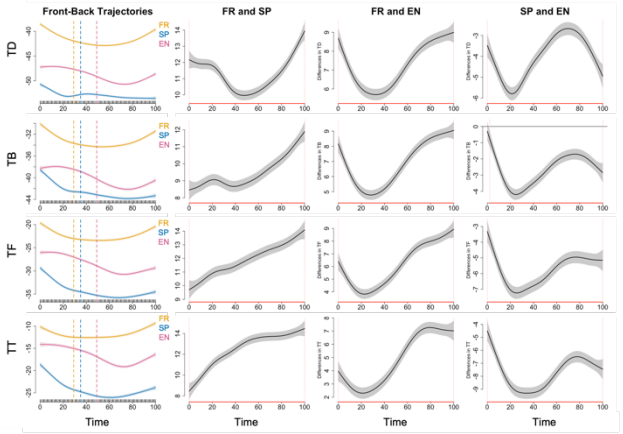


Figure 2: Column 1 shows front-back displacement in mm for the word 'rat'. Dotted lines: end of rhotic gestures per speaker. Columns 2 – 4: significant differences between speakers (in red). Drops in the trajectory correspond to tongue retraction.

Significant effects on the smooth found across speakers for all captors (mean edf = 7.084, p<2e-16). When compared to the other speakers, EN shows significant advancement of the tongue towards the end of the rhotic gesture. SP retracts the tongue further back than EN and FR. The overall tongue position for FR is more advanced than it is for the other two speakers (see column 1 of Fig. 2). Smooth values closest to 0 intercept for FR and EN where TF and TT are concerned. Differences amongst all speakers were more pronounced for the front-back dimension as interpreted from partial significance and values approaching 0 intercept for rhotic gestures in the high-low dimension. SP had the longest average duration for the word "rat" at 343 ms (σ: 32), followed by EN at 307 ms (σ: 21), and FR at 234 ms (σ: 20).

3.2. R in word-final position

An ANOVA revealed significant values for the effect of speaker on the smooth of the spline for TD, TB, TF, and TT (mean edf of 7.361, p<2e-16). Displacement for the rhotic in *'par'* plotted in Fig. 3 coincides with rhotic articulatory configurations for EN and FR in the word 'rat'. EN presents noticeable raising of TD, TB, and TF, followed by a slight fall at gesture offset. FR raises the entirety of the tongue during the rhotic gesture with what seems like a steeper slope for TF and TT.

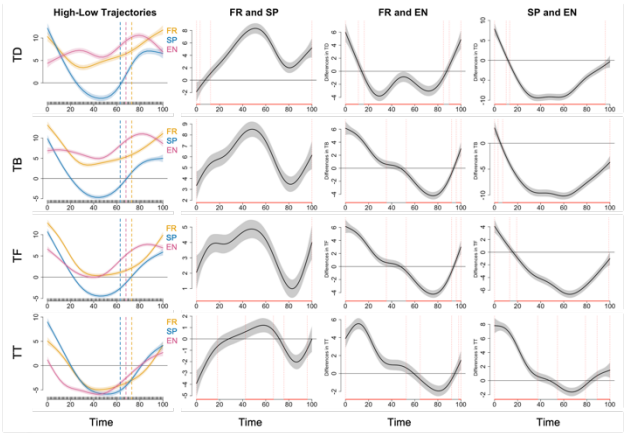


Figure 3: Column 1 shows high-low displacement in mm for the word 'par'. Dotted lines: start of rhotic gestures per speaker. Columns 2 – 4: significant differences between speakers (in red). Drops in trajectory correspond to tongue lowering.

SP raises TD after the vowel gesture and slightly lowers it at the end of the rhotic gesture. TB is also raised but does not appear to fall at gesture offset, it plateaus instead. TF, and TT are all raised and continue to rise until the end. Partial significance for TT was found in all speakers with similar trajectories in column 1. Partial significance at gesture offset was found for TD, TB, and TF between FR and EN.

Front-back displacement in Fig. 4 shows tongue advancement during rhotic articulation for FR and EN. SP shows slight advancement for all four captors but TD and TB undergo a slight fall at gesture offset. Significant differences were found across speakers for the effects of the smooth on the spline (mean edf = 7.091, p<2e-16) for all captors. FR rhotic articulatory gestures are significantly more anterior

than gestures for EN and SP. Tongue gestures are consistently more posterior for SP than for FR and EN for both vowel and rhotic articulation. TB rhotic gestures approach 0 intercept for SP and EN making differences only partially significant. Overall differences between speakers were more prominent where front-back trajectories were concerned. EN had the longest mean duration for the word "par" at 611 ms (σ: 81), followed by SP at 480 ms (σ: 54), and FR at 372 ms (σ: 40).

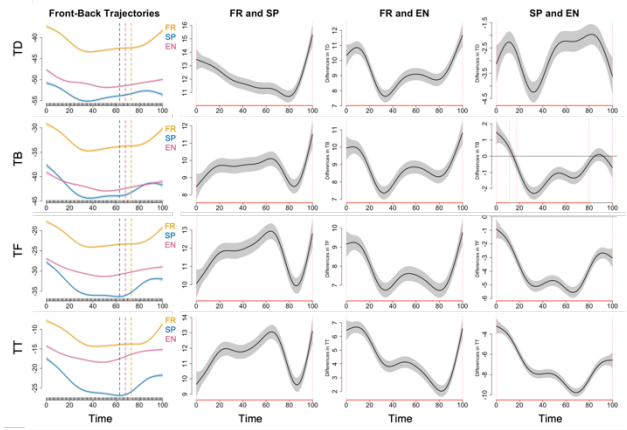


Figure 4: Column 1 shows front-back displacement in mm for the word 'par'. Dotted lines: the start of rhotic gestures per speaker. Columns 2 – 4: significant differences between speakers (in red). Falls in the trajectory correspond to the posteriorization of captor.

4. DISCUSSION

Tongue tip lowering for all speakers suggests an active involvement of the front part of the tongue for rhotic articulation in French but only in initial position. Articulatory studies that describe lingual configurations for [ʁ] either omit tongue tip movement altogether or simply describe it as unsolicited during uvular rhotic production [3]. This lowering could be due to an all-encompassing lowering of the tongue for the vowel gesture. Future studies will describe rhotic production in multiple contexts and in a larger number of participants. Westbury's EMA study [10] did not find any significant links between tongue shape and oral cavity size, or gender, for American /ɹ/. However, no known studies have tested the same effects on French /ʁ/, or uvular rhotics in general. Influence of physiological factors remains inconclusive. Articulatory strategies varied significantly across speakers. EN raised tongue dorsum and body for word-initial and final positions suggesting a lower point of constriction for the alveolar consonant at both ends of the word. SP shows a higher point of articulation in relation to the alveolar boundary but only in word-final position. American English [ɹ] is produced by lowering the tongue dorsum toward the pharynx [25]. Children with [ɹ] distortions raise the back of the tongue toward the velum [26-28]. [29] describe productions of [ɹ] with high tongue body as mergers of oral and pharyngeal constrictions resulting in singular velar constrictions. As per [19], lower time of exposure could mean higher degrees of language interference. This could explain the velarization seen in both EN and SP. EN and SP may have identified the French rhotic as [DORSAL] but, phonetically, it was produced as velar and not uvular. SP only velarized the rhotic gesture in final position. This could be explained by a higher time of exposure to the target language.

SP was expected to retract more in word-final position [7] while EN was thought to retract more in word-initial position [12]. This was not the case. While both SP and EN do in fact articulate gestures further back than FR overall, SP's gestures were less posterior in word-final than in word-initial. Similarly, EN showed less tongue retraction in word-initial than in word-final position. Last but not least, FR produced shorter words and rhotic gestures with qualitatively smoother curves. This suggests a certain level of efficiency when it comes to articulatory effort and, consequently, a certain level of undershoot when it comes to articulatory targets. SP showed similar curve patterns to FR and EN showed the most complex curves by far. Longer durations in word-initial position for SP and word-final position for EN could be interpreted as signs of hyper articulation. Our findings support [15, 16, 30] whereby the learning of L2 coarticulation patterns is defined as a gradual process linked to the amount of exposure to the target language.

5. CONCLUSION

This case study analysed articulatory strategies for French rhotics produced by a native speaker of French and two learners of French as an L3. Lower times of exposure seemed to correlate with higher degrees of language interference as attested by the velarization of French [ʁ] and lower degrees of coarticulation for SP and EN.

Our findings are in line with [14] and go on to highlight the importance of phonetic and phonological relations between the learner's mother tongue and the target language not only when identifying, but equally as important when producing allophones of the same category. This leads us to believe that articulatory strategies in language learners are both phonologically and phonetically driven.

6. ACKNOWLEDGEMENTS

This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL)..

7. REFERENCES

[1] Delattre, P. 1971. Pharyngeal features in the consonants of Arabic, German, Spanish, French, and American English. *Phonetica* 23, 129-155.

[2] Boyce, S.E., Hamilton, S.M., Rivera-Campos, A. 2016. Acquiring rhoticity across languages: An ultrasound study of differentiating tongue movements. *Clinical linguistics & phonetics* 30, 174-201.

[3] Howson, P.J., Kochetov, A. 2020. Lowered F2 observed in uvular rhotics involves a tongue root gesture: Evidence from Upper Sorbian. *J. Acoust. Soc.Am.*, 147, 2845-2857.

[4] Gick, B., Allen, B., Stavness, I., Wilson, I. 2013. Speaking tongues are always braced. *J. Acoust. Soc. Am* 134, 10-1121.

[5] Proctor, M. 2011. Towards a gestural characterization of liquids: Evidence from Spanish and Russian. *Laboratory Phonology* 2, 451-485.

[6] Rivera-Campos, A., Boyce, S. 2013, June. Describing alternative articulations of the Spanish trill /r/ by ultrasound technology. In *Proceedings of Meetings on Acoustics ICA2013* 19. Acoustical Society of America.

[7] Colazo-Simon, A. 2014. Técnica para el estudio cinemático de la coordinación articulatoria de las vibrantes del español: el articulógrafo electromagnético. *Técnica para el estudio cinemático de la coordinación articulatoria de las vibrantes del español: el articulógrafo electromagnético*, 139-150.

[8] Delattre, P., Freeman, D.C., 1968. A dialect study of American r's by x-ray motion picture.

[9] Alwan, A., Narayanan, S., Haker, K. 1997. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *J. Acoust. Soc.Am.* 101, 1078-1089.

[10] Westbury, J.R., Hashi, M., Lindstrom, M.J. 1998. Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication* 26, 203-226.

[11] Harper, S., Goldstein, L., Narayanan, S. 2020. Variability in individual constriction contributions to third formant values in American English /ɹ/. *The J. Acoust. Soc. Am.* 147, 3905-3916.

[12] Campbell, F., Gick, B., Wilson, I., Vatikiotis, E. 2010. Spatial and temporal properties of gestures in North American English /r/. *Language and speech* 53, 49-69.

[13] Wilson, I., Gick, B. 2014. Bilinguals use language-specific articulatory settings. *Journal of Speech, Language, and Hearing Research* 57, 361-373.

[14] Reinisch, E., Juhl, K.I., Llompart, M. 2020. The impact of free allophonic variation on the perception of second language phonological categories. *Frontiers in Communication* 5, 47.

[15] Oh, E. 2008. Coarticulation in non-native speakers of English and French: An acoustic study. *Journal of phonetics* 36, 361-384.

[16] Baumotte, H., Dogil, G. 2008. Coarticulation in non-native speakers of English:/. lv/-sequences in non-proficient vs. proficient learners. *Proc. ExLing* 37-40.

[17] Gahl, S. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84, 474-496.

[18] Bell, A., Brenier, J.M., Gregory, M., Girand, C., Jurafsky, D. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60, 92-111.

[19] Flege, J.E. 2019. A non-critical period for second-language learning. *A sound approach to language matters, Aarhus University. Open access e-book at Aurhus University Library*.

[20] Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., Wieling, M. 2021. A review of data collection practices using electromagnetic articulography. *Laboratory Phonology* 12, 1-42

[21] Buech, P., Roessig, S., Pagel, L., Muecke, D., Hermes, A. 2022. ema2wav: doing articulation by Praat. In *Interspeech* 1352-1356.

[22] Woods, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73, 3-36.

[23] Hastie, T., Tibshirani, R. 1987. Generalized additive models: some applications. *Journal of the American Statistical Association* 82, 371-386.

[24] Coretta, S. 2019. Assessing mid-sagittal tongue contours in polar coordinates using generalised additive (mixed) models.

[25] Preston, J.L., Benway, N.R., Leece, M.C., Hitchcock, E.R., McAllister, T. 2020. Tutorial: Motor-based treatment strategies for /r/ distortions. *Language, speech, and hearing services in schools* 51, 966-980.

[26] Adler-Bock, M., Bernhardt, B.M., Gick, B. Bacsfalvi, P. 2007. The use of ultrasound in remediation of North American English /r/ in 2 adolescents. *American Journal of Speech-Language Pathology* 16, 128-139.

[27] Klein, H.B., Byun, T.M., Davidson, L., Grigos, M.I. 2013. A multidimensional investigation of children's/r/productions: Perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology* 22, 540-553.

[28] McAllister Byun, T., Tiede, M. 2017. Perception-production relations in later development of American English rhotics. *PloS one* 12, e0172022.

[29] Gick, B., Bacsfalvi, P., Bernhardt, B.M., Oh, S., Stolar, S., Wilson, I. 2007, June. A motor differentiation model for liquid substitutions in children's speech. In *Proc. Meetings on Acoustics 153ASA*. 1. Acoustical Society of America, 060003.

[30] d'Apolito, S., Fivela, B.G. 2016. Investigating Speech Gesture Coordination in the Production of L2 Consonant Clusters: the Added Value of Electromagnetic Articulography. In *Conference Proceedings. ICT for Language Learning* 9, 411-419.

NON-PULMONIC INITIATION IN HUMAN BEATBOXING: A REAL-TIME MRI STUDY

Alexis Dehais-Underdown[1], Paul Vignes[1], Lise Crevier-Buchman[1,2], Didier Demolin[1], Pierre-André Vuissoz[3], Karyna Isaieva[3], Marc Fauvel[4], Yves Laprie[5], Jacques Felblinger[3,4]

[1]Laboratoire de Phonétique et Phonologie (UMR 7018 - CNRS/Sorbonne-Nouvelle), [2]Unité Voix, Parole Déglutition, Service ORL et de Chirurgie de la Face et du Cou, Hôpital Foch, [3] IADI, INSERM-U947, Université de Lorraine, Nancy, France, [4] CIC-IT 1433, INSERM, Université de Lorraine & CHRU Nancy, Nancy, France, [5] LORIA (CNRS/Inria), Université de Lorraine, Nancy, France

alexis.dehais-underdown@sorbonne-nouvelle.fr

ABSTRACT

Human Beatboxing (HBB) is a musical technique produced by vocal tract movements. Beatboxers seem to exploit vocal tract capacities to their full extent. Thus, HBB is a good paradigm to study production mechanisms. This study uses aerodynamic and laryngoscopic along with 2D MRI data to analyze non-pulmonic initiation mechanisms during HBB production. The production of 4 professional beatboxers was analyzed while producing glottalic and lingual mechanisms. Results shows that glottalic mechanisms result from a sequence of laryngeal vertical movement, tongue root maneuvers and, sometimes, additional pharyngeal gestures to increase or decrease pressure. Lingual mechanisms were produced by front-back tongue movements to generate egressive or ingressive airflows. A reformulation of Catford model of initiation mechanisms, in terms of initiatory gestures rather than aerodynamic parameters, is proposed.

Keywords: Human Beatboxing, Laryngoscopy, Initiation, rt-MRI, Aerodynamics

1. INTRODUCTION

This study aims to describe Human Beatboxing (HBB) production in a similar way to Proctor *et al.* [1] and Blaylock *et al.* [2]. This study focused on the production of non-pulmonic initiation mechanisms. Contrary to Proctor's and Blaylock's study, this research on HBB production was not originally based on rt-MRI data but on aerodynamic, laryngoscopic and acoustic data. MRI was acquired to confirm our findings in the physiological and acoustic data of 4 professional beatboxers [6]. Both [1] and [2] offer a formal description of HBB production based on MRI. While both studies report a wide range of production mechanisms, Proctor's study focuses on the "paralinguistic use" of production mechanisms found in the world's languages (e.g. [p'] [ts'] [k']) and Blaylock points out that beatboxers not only make use of mechanisms found in phonetic inventories but also mechanisms that are unknown to phonetic inventories (e.g. [↓ɽː] [↓ʙ̥ˡ]). Thus, HBB constitutes a good paradigm to investigate the extent of the human vocal tract capacities.

Initiation is the phase that sets air in motion. Airflow is generated by increasing or decreasing of the volume between the place of initiation and the place of articulation. Table 1 shows Catford's aerodynamic model of initiation [3]. He defines initiation in terms of location (i.e. lungs, larynx, mouth) and directionality of airflow (i.e. egressive and ingressive). Compressive gestures produce positive pressure and generate egressive airflow: *pulmonic pressure*, *glottalic pressure* and *velaric pressure*. Rarefactive gestures produce negative pressure and ingressive airflow: *pulmonic suction*, *glottalic suction* and *velaric suction*. *Velaric* initiation is more commonly referred as *lingual*. Hereafter, *lingual* will be used over *velaric*.

Table 1: Catford's initiation parameters

	Direction	
Location	**Compressive**	**Rarefactive**
Lungs	pulmonic pressure	pulmonic suction
Larynx	glottalic pressure	glottalic suction
Mouth	velaric pressure	velaric suction

Non-pulmonic mechanisms (i.e. *glottalic* and *lingual*) have been extensively documented in the

world's languages. Concerning the production of HBB, it was found that glottalic initiation is composed of a glottal closure, vertical laryngeal movements and tongue root maneuvers to increase or decrease the volume of the lower vocal tract [1, 4, 2, 5]. These studies also report an expansion of the lower vocal tract (i.e. tongue root protraction, laryngeal lowering, pharyngeal expansion) preceding the glottal closure and the laryngeal raising of ejectives. Lingual mechanisms in beatboxing can be produced either by increasing the volume between the anterior constriction and posterior constriction (i.e. click) or by decreasing the volume between the anterior constriction and posterior constriction [1, 2]. Furthermore, an aerodynamic study of [6] reported pressure values between $40hPa$, $100hPa$ for the production of glottalic beatboxed sounds. In comparison, during speech production, intraoral pressure of ejectives ranges from $20hPa$ for labial stops up to $40hPa$ for velar stops [7]. It suggests that beatboxers use a greater degree of compression and expansion to produce glottalic sounds. To further understand initiation and possible differences between beatboxing and speaking, we propose to investigate the production of glottalic and lingual production mechanisms with laryngoscopic and aerodynamic recordings of 4 professional beatboxers and additionnal 2D MRI data of one subject.

2. METHODS

2.1. Protocol and Corpus

The data presented here comes from a larger database of 4 professional beatboxers (3 males, 1 females). Fibroscopic and aerodynamic data were acquired first. To confirm observations from the physiological data (see [6]), a pilot study using real-time Magnetic Resonance (rt-MRI) with one subject was carried out. The subset of the corpus containing only non-pulmonic sounds is given in Table 2. Hereafter, transcriptions with curved brackets indicates sounds produced with lingual mechanism and down arrow indicates ingressive airflow (e.g. {↓ʙ̥}). Sounds were produced in isolation and in *Beatboxing Patterns*. Sounds in isolation were produced 8 times each. This study only report sounds produced in isolation. The analysis focuses on gestures implicated in raising and lowering the pressure.

2.2. Data Acquisition

Laryngoscopic data was acquired by the team's MD with a Xion video-stroboscopic system using a flexible fiberscope with a rate of 25 frames per second. The acoustic signal was acquired through the fiberscope microphone. A 2% Xylocaine anesthesia was administered before the endoscope was inserted through the nose into the pharyngeal cavity above the supraglottic plan. In this session, subjects were not asked to produced lingual mechanisms. Aerodynamic and acoustic signals were acquired with an EVA2 Workstation [8] allowing simultaneous recording of acoustic signal, intraoral pressure (measured in hPa, $1hPa = 1.02cmH_2O$), oral airflow and nasal airflow (measured in dm^3/s). Intraoral pressure (Po) was obtained by inserting a small tube into the pharynx through the nasal cavity. Pressure for clicks is very difficult to measure because one would need to measure it between the two constrictions. Thus, the measured pressure corresponds to the pharyngeal pressure. In the experiment, beatboxers were asked to produce continuous humming while producing lingual mechanisms, so the pressure is 0 and nasal airflow is positive.

Table 2: Corpus of non-pulmonic sounds of our database.

Glottalic	*Stop*	pʼɓʼɸ		tʼ dʼ	
	Affricate		pfʼ	tsʼ tʃʼ	ʞʟ
Lingual	*Stop*	p			
	Affricate		pf	ts	↓kʟ
	Trill	↓ʙ̥			

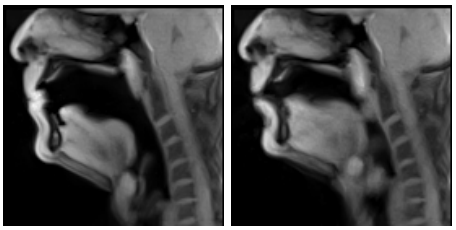
The 2D MRI data was recorded at Nancy Central Regional University Hospital with a Siemens Prisma 3T scanner, Erlangen, Germany. The speaker was in supine position and a Siemens Head/Neck 64 coils was used. For the 2D real-time we used radial RF-spoiled FLASH sequence with TR = 2.22 ms, TE = 1.47 ms, FoV 192x192mm, flip angle = 5°, and slice thickness was 8 mm. Pixel bandwidth was 1670 Hz/pixel. Image size was 136x136, in-plane resolution was 1.6 mm, recorded at 50 fps and reconstructed with a nonlinear inverse technique presented in [9]. Audio was recorded at a sampling frequency of 16 kHz inside the MRI scanner with a FOMRI III optoacoustics fibre-optic microphone (FOMRI III, Optoacoustics Ltd., Mazor, Israel).

3. RESULTS

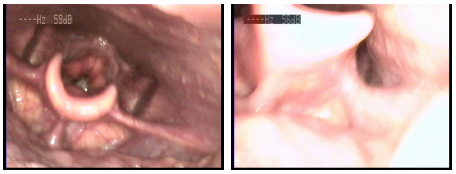
3.1. Glottalic mechanisms

The Figure 1 illustrates the production of [pʼ] with MRI frames as well as laryngoscopic images

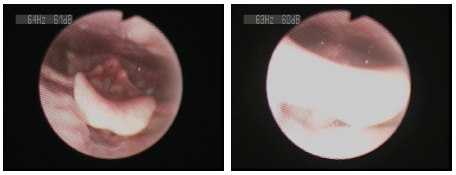
at the beginning and at the end of the initiatory phase. Laryngeal behavior in the laryngoscopic data for glottalic egressive sounds were highly similar across subjects and sounds. The mean pressure for glottalic sounds ranges from 20hPa to 85hPa depending on the subject and the sound. Fibroscopic images show a systematic behavior across subjects and sounds: laryngeal lowering and pharyngo-laryngeal expansion preceding a glottal closure and, simultaneously, laryngeal elevation and tongue root retraction. Two subjects added additional aryepiglottic constriction (i.e. approximation of the arytenoids on top of the root of the epiglottis) and one subject added pharyngeal compression following laryngeal elevation and tongue root retraction. MRI images confirms the systematic tongue root maneuver and laryngeal elevation observed for all subjects. Furthermore, the MRI data suggests that tongue root retraction is voluntarily recruited to decrease the volume of air between the glottis and the lips.



(a) Subject VP



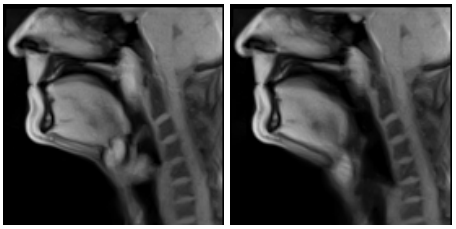
(b) Subject VP



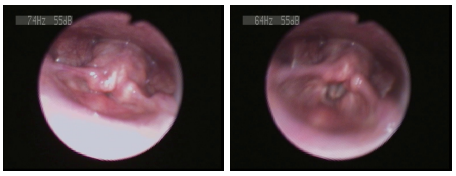
(c) Subject AI

Figure 1: MRI frames (top) and laryngoscopic images (middle and bottom) of the glottalic mechanism produced for [pʾ].

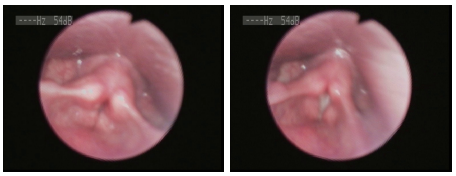
Figure 2 illustrates the production of [ɓʾ]. Once again, individual behaviors were very similar. The mean pressure for [ɓʾ] ranges −5hPa to −25hPa and −50hPa −75hPa for [ɓ] (produced only by 2 subjects). Concerning [dʾ] and [ʁʟ] the pressure signal clipped. The fibroscopic frames show laryngeal raising and pharyngo-laryngeal narrowing preceding a ventricular closure (i.e. a closure of both vocal and ventricular folds) and laryngeal lowering and tongue root protraction. One subject added additional aryepiglottic constriction. 2 subjects added pharyngeal expansion following laryngeal lowering. MRI data shows the same phases confirming once more the active role of tongue root during the initiation.



(a) Subject VP



(b) Subject CJ



(c) Subject GA

Figure 2: MRI frames (top) and laryngoscopic images (middle and bottom) of the glottalic mechanism produced for [ɓ].

3.2. Lingual mechanism

Figure 3 illustrates the egressive lingual mechanism for {pf}. MRI frames shows the formation of two constrictions in the vocal tract. We found anterior closure to be bilabial, labiodental, dental and alveolar. The posterior constriction is formed in the velar or uvular region. To generate airflow for the production of lingual egressive sounds, the tongue dorsum make a frontward movement to compress the volume of air between the posterior and the anterior constriction. For all sounds, except {↓kʟ}, the front constriction was released first. For {↓kʟ} the velar/uvular constriction releases first and was considered as lateral velar based on MRI, acoustic data and the subject's description.

Figure 4 illustrates the the production of {↓ʙ̥ˡ}. Lingual ingressive initiation is well-known for the production of clicks. Here, a similar mechanism occurs: a labial and a velar constriction are created

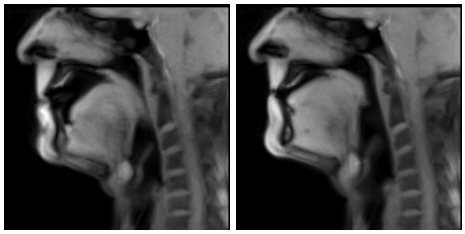


Figure 3: MRI frames of the lingual mechanism produced for {pf} - Subject VP

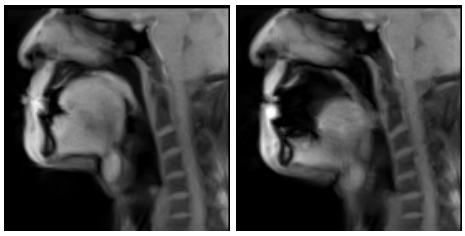


Figure 4: MRI frames of the lingual mechanism produced for {↓ʙ̥ˡ} - Subject VP

and the tongue retracts to expand the volume which results in an ingressive airflow. Ingressive bilabial trills in HBB are lateralized. Beatboxers were unable to produce central ingressive bilabial trilling.

4. DISCUSSION

The data presented here concurs with previous studies of initiation mechanisms during HBB production. For the ejective production, our data support the findings reported in [1, 4, 5, 2], that is expansion of the lower vocal tract previous to compression of the volume of air between the glottis and the place of articulation by laryngeal elevation and tongue root retraction. Also, lingual mechanisms support the findings of [1, 2] concerning tongue maneuver to produce either egressive or ingressive airflow. Lingual trills are not attested in any phonological system but are attested in beatboxing [2, 6]. Additional tongue root maneuvers, in coordination with vertical movement of the larynx, to raise or lower the pressure during the production of glottalic mechanisms has been hypothesized by Kingston [10]. It has been observed on MRI data for ejectives in Tigrinya [11] and implosives in Hendo [12]. Our data confirms that tongue root maneuvers play a key role in raising or lowering the pressure. Pharyngeal expansion or compression may also participate in decreasing or increasing the pressure.

What does it mean for a sound to be "*glottalic*"? Originally, the term was used to differentiate ejectives and implosives from *glottalized consonants* (see Catford's discussion [3], p.247-8, Ch. 5 notes 1, 2). Both "*glottalic*" and "*glottalized*" referred to glottal or laryngeal maneuvers implicated in consonant production. A *glottalic consonant* is now understood (and taught) as a consonant produced with a closed glottis and vertical movements of the larynx. Though, both linguistic and HBB data on ejectives and implosives suggests that this initiation mechanism cannot merely be reduced to vertical movements to raise or lower the pressure. Therefore, we propose to use the term "*laryngeal*" instead of "*glottalic*". Catford's model of initiation is based on aerodynamic parameters (i.e. pressure and airflow direction) to distinguish between initiatory mechanisms. Although the model is motivated, the aerodynamic phase is the result of vocal tract maneuvers to change the pressure. We propose to reformulate Catford's model in terms of initiatory gestures. By analogy to places and modes of articulation, Table 2 proposes a revision of the model with the 3 places of initiation from the original model and 2 modes of initiation. The mode of initiation refers to gestures that either compress or expand the volume between the place of initiation and the place of articulation. The phonetic description takes into account the place of initiation and the direction of the airflow resulting from compression (i.e. egressive) or expansion (i.e. ingressive).

Table 3: Reformulated initiation parameters

Location	Mode	Phonetic description
Lungs	Compression	Pulmonic Egressive
	Expansion	Pulmonic Ingressive
Larynx	Compression	Laryngeal Egressive
	Expansion	Laryngeal Ingressive
Mouth	Compression	Lingual Egressive
	Expansion	Lingual Ingressive

5. CONCLUSION

Human Beatboxing constitutes an original contribution to discuss production mechanisms with non-linguistic data. The reformulation of Catford's model in terms of initiatory gestures is motivated as it allows to make a direct link with constraints on sound patterns (e.g. clicks and the "*back constraint vowel*"). However, in terms of phonetic description, the reformulated model needs to be supported by an acoustic analysis to confirm the relevance of the phonetic features used to differentiate initiation mechanisms.

6. ACKNOWLEDGMENT

This physiological protocol was approved by an ethical committee (RCB-ID n° 2020-A00246-33). The MRI protocol was approved ethical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053, RCB-ID n°CPP EST-III, 08.10.01). The research is supported in France by the Delegation of Research and Innovation of Hôpital Foch and LABEX EFL (Empirical Fundation of Linguistics, ANR-10-LABX-0083) and the ArtSpeech project (ANR-15-CE23-0024) and supported by European founds CPER “IT2MP”, “LCHN” and “FEDER”.

7. REFERENCES

[1] M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan, “Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study,” *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1043–1054, 2013.

[2] R. Blaylock, N. Patil, T. Greer, and S. S. Narayanan, “Sounds of the human vocal tract.” in *INTERSPEECH*, 2017, pp. 2287–2291.

[3] J. C. Catford, *Fundamental problems in phonetics*. Bloomington And London: Indiana University Press, 1977.

[4] T. De Torcy, A. Clouet, C. Pillot-Loiseau, J. Vaissiere, D. Brasnu, and L. Crevier-Buchman, “A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox,” *Logopedics Phoniatrics Vocology*, vol. 39, no. 1, pp. 38–48, 2014.

[5] N. Patil, T. Greer, R. Blaylock, and S. S. Narayanan, “Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging.” in *Interspeech*, 2017, pp. 2277–2281.

[6] A. Dehais-Underdown, P. Vignes, L. Crevier-Buchman, and D. Demolin, “In and out: production mechanisms in human beatboxing,” *Proceedings of Meetings on Acoustics, Vol. 45, 060005*, 2022.

[7] D. Demolin, “Acoustic and aerodynamic characteristics of ejectives in amharic,” *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2610–2610, 2004.

[8] A. Ghio and B. Teston, “Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies,” in *International Conference on Voice Physiology and Biomechanics*, 2004, pp. 55–58.

[9] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, “Real-time mri at a resolution of 20 ms,” *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[10] J. Kingston, “The phonetics and phonology of the timing of oral and glottal events,” Ph.D. dissertation, University of California, Berkeley, 1985.

[11] Z. Hermes, M. Fu, S. Rose, R. Shosted, and B. Sutton, “Representations of place and airstream mechanism: A real-time mri study of tigrinya ejectives,” in *Proceedings of the 15th Conference on Laboratory Phonology, Ithaca, USA*, vol. 133, 2016.

[12] D. Demolin, H. Ngonga-Ke-Mbembe, and A. Soquet, “Phonetic characteristics of an unexploded palatal implosive in hendo,” *Journal of the International Phonetic Association*, vol. 32, no. 1, pp. 1–15, 2002.

SPEED RATE EFFECTS ON HUMAN BEATBOXING: AN AERODYNAMIC STUDY

Alexis Dehais-Underdown[1], Paul Vignes[1], Lise Crevier-Buchman[1,2], Didier Demolin[1]

[1]Laboratoire de Phonétique et Phonologie (UMR 7018 - CNRS/Sorbonne-Nouvelle), [2]Unité Voix, Parole Déglutition, Service ORL et de Chirurgie de la Face et du Cou, Hôpital Foch
alexis.dehais-underdown@sorbonne-nouvelle.fr

ABSTRACT

The present study aims to describe speed rate effects (i.e. tempo) on Human Beatboxing based on aerodynamic data. Intraoral pressure (Po), oral airflow (Oaf), nasal airflow (Naf) and the acoustic signal were simultaneously recorded. The experiment consisted in the production of various Beat Patterns composed of *kick drums*, *hi-hats* and *snare drums*. They were produced at different rates: 90 beats per minute (BPM), 120 BPM and 150 BPM. Based on Po and Oaf, we estimated the relative constriction area. Results show temporal reduction of patterns ($\rho = -0.97$) and sounds ($\rho = -0.37$). Subjects tended to decrease intraoral pressure and increase constriction area as speed rate increased. In few cases, subjects increased intraoral pressure and decreased constriction area as rate increased. Human Beatboxing constitute an original contribution to discuss production mechanisms with non-linguistic data.

Keywords: Experimental Phonetics, Beatboxing, Reduction, Tempo, Aerodynamics

1. INTRODUCTION

The present study aims to describe speed rate (i.e. tempo) effects on Human Beatboxing (HBB) based on aerodynamic data. Speech rate is known to be a factor of reduction affecting supralaryngeal and laryngeal gestures. During fast speech, articulators' velocity is increased and consequently gestures' duration and amplitude are reduced [1, 2, 3, 4, 5]. At fast rates, the interval of time between gestures is shortened, the degree of overlap may increase as well [4, 5]. Speech rate was found to have little or no effects on intraoral pressure and oral airflow peaks [6, 7, 8, 9]. Though, these studies do not account for individual patterns and only investigate pulmonic egressive obstruents. HBB is a musical technique produced with vocal tract movements. A study of [10] found that tempo affects the duration of beatboxed sounds and patterns.

Aerodynamics is a branch of fluid mechanics dealing with the motion of gases. During vocal production (e.g. speech, singing), vocal tract volume continuously changes over time and generates different pressure levels and different types and levels of airflow (i.e. volume velocity). From intraoral pressure and airflow measurements, we can estimate particle velocity and the relative constriction area [11]. *Particle velocity* (v) is the rate of change in particle position. Particle velocity depends on several factors as pressure, air density or gravitational acceleration [12]. Catford [11] offers a simplified equation for phonetic purposes: $v = 412\sqrt{P}$, where P is the measured intraoral pressure. *Relative constriction area* (A) is derived from volume velocity and particle velocity $A = U/v$ where U is the measured oral airflow and v the estimated *particle velocity*.

To investigate speed rate effects on HBB, subjects were asked to produce *Beatboxing Patterns* at 3 different tempo: 90, 120 and 150 beat per minutes (BPM). Tempo is a musical term that refers to the interval of time separating pulsations (e.g. interval of time separating the metronome's beats) and defines how "quick" a musician plays. Low tempo are characterized by long intervals between notes and fast tempo are characterized by shorter intervals. In terms of production, at fast tempo, musician will have less time to reach the targets (e.g. key strike for a pianist). Concerning HBB, we hypothesize that subjects will have less time to reach the targets as tempo increases and we expect intraoral pressure to decrease and constriction area to increase as tempo increases.

2. METHODS

2.1. Data Acquisition

Aerodynamic and acoustic signals were acquired with EVA2 Workstation [13]. EVA2 Workstation allows simultaneous recording of acoustic signal, intraoral pressure (hPa, $1hPa = 1.02cmH_2O$), oral

airflow and nasal airflow (dm^3/s). The acoustic waveform was obtained by EVA2 integrated microphone. Oral Airflow (Oaf) was collected using a flexible silicone mask pressed on the subjects' mouth. Nasal Airflow (Naf) was obtained through a tube placed in the nostril. Intraoral pressure (Po) was obtained inserting a small tube into the pharynx through the nasal cavity. Whenever Po and Oaf clipped, tokens were excluded from the analysis.

2.2. Protocol and Corpus

The data presented here was excerpted from a larger database of aerodynamic data from 4 professional beatboxers: 3 males (VP, CJ, GA) and 1 female (AI). Most of them have 12 or 13 years of training except AI that has only 4 years of training. All of them won at least one french championship and two of them won a world championship. The experiment consisted in the production of various Beat Patterns (BP) produced at 90 BPM, 120 BPM and 150 BPM repeated 4 times each. Tempo was given through a vibrating metronome placed and the subjects' wrist.

The corpus is composed of 11 patterns with the same metrical, rhythmical and melodic structure but with different phonetic structure. Figure 1 gives an example structure of the basic pattern [p' t͡s' ↓k͡ʟ t͡s' p' p' t͡s' ↓k͡ʟ t͡s']. The same pattern may also be produced with a different phonetic structure such as [t' t͡s' t͡ʃ' t͡s' ʈ' ʈ' t͡s' t͡ʃ' t͡s'] for example. On the Figure, the first line shows a 90BPM metronome's pulsation, the second line shows the melodic structure with drums alternations, the third line indicates the position in the metrics and the fourth line gives a transcription of the phonetic structure. The analysis focuses on the pattern illustrated in Figure 1. In this BP, [p'] was produced as an ejective bilabial stop. [t͡s'] was produced as an ejective dental affricate by subjects VP, AI, CJ and as a pulmonic egressive dental affricate [t͡sː] in position 9 by subject GA. [↓k͡ʟ] was produced as a pulmonic ingressive velar lateral affricate by subjects VP, CJ, GA and as an implosive velar stop and a pulmonic ingressive velar lateral fricative [↓k͡ʟ] by subject AI. In total, 48 patterns and 396 tokens were analyzed.

2.3. Aerodynamic measures

For each sound, intraoral pressure (Po) and oral airflow (Oaf) peaks were automatically extracted in Praat. Then, we calculated particle velocity $v = 412\sqrt{P}$, where P is the measured pressure, and relative constriction area $A = U/v$, where U is the measured oral airflow. We also extracted

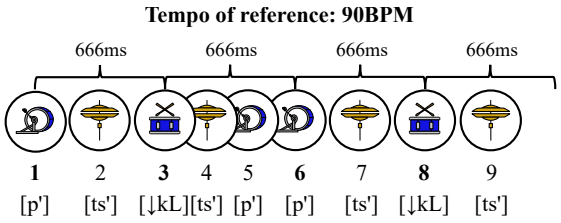


Figure 1: Metrical structure of a Beat Pattern at 90BPM. Line 1: metronome's pulsation, line 2: melodic structure, line 3: metrical position, line 4: phonetic structure

sound duration and patterns duration to assess temporal reduction of beatbox patterns and sounds. Sound duration was measured from the beginning of the occlusion (when Po starts increasing) to the offset of the acoustic signal. Pattern duration was measured from the beginning of the first sound to the end of the last one. Based on patterns duration, we calculated the actual tempo at which subjects produced patterns. Spearman's coefficient (ρ) were calculated in R between temporal and aerodynamic variables. Sounds were analyzed for each participant in each position to investigate individual behaviors and possible effect of sounds and metrical positions regarding speed rate effects.

3. RESULTS

Spearman's coefficients showed very high correlation between the tempo of reference (i.e. 90, 120, 150BPM) and BP duration ($\rho = -0.97$). BP duration decreases when speed rate increases. Correlation between the tempo of reference and sound duration is lower ($\rho = -0.37$). Sounds are less affected than BPs, suggesting inter-gestural intervals decrease more than sounds themselves. There was no correlation between tempo and aerodynamic measures.

3.1. Speed rate effects on pressure

Intraoral pressure changes as a function of speed rate allows us to infer changes of the initiatory gesture (i.e. gesture setting air in motion). A decrease of the pressure would suggest an undershoot of the gesture (i.e. less volumetric reduction) while an increase of the pressure would suggest an overshoot of the gesture (i.e. more volumetric reduction). Figure 2 shows a scatter-plot of intraoral pressure as a function of tempo. There is a tendency of intraoral pressure to decrease at fast rate. Though, some exceptions and non-linear effects are observed.

There is an increase in intraoral pressure as speed rate increases for subject CJ in position 6 ($\rho = 0.52$)

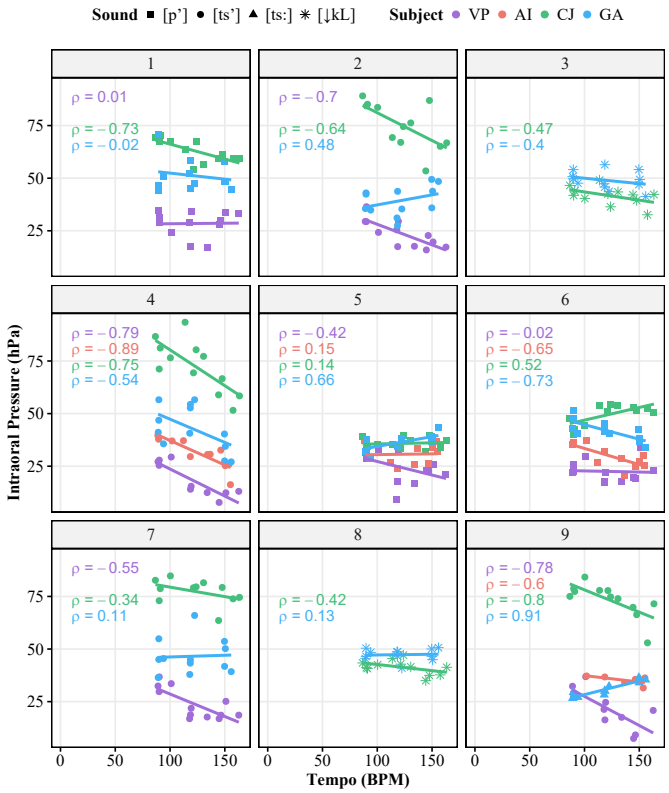


Figure 2: Intraoral pressure as a function of tempo. Each panel represents one position.

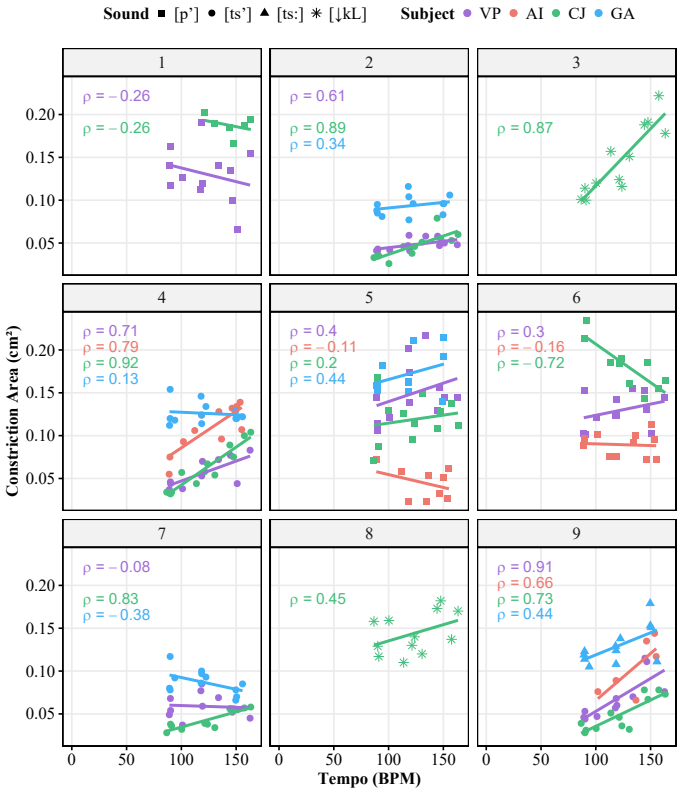


Figure 3: Constriction area as a function of tempo. Each panel represents one position.

and for subject GA in position 5 ($\rho = 0.66$) and 9 ($\rho = 0.73$). This increase suggest that there is a greater volume reduction as tempo increases. In some cases, initiation remains stable as for example subject VP in position 1 and 6.

There are non-linear effects of speed rate on initiation. For example, if we look at the data from CJ in position 6, Po increases between low tempo (i.e. 90BPM) and medium tempo (i.e. 120BPM) but remains stable between medium and fast (i.e. 150BPM). Another non-linear effect can also be observed for subject GA in position 4 where Po increases at 120BPM but decreases at 150BPM.

There is a consistent tendency across subjects in position 4, suggesting position in the metric influences reduction. Indeed, intraoral pressure reduces for all subjects concerning [t͡s'] in position 4 ($\rho < -0.54$). In position 9, there is an important decrease of the pressure for all subjects except for GA who produced [t͡s:] with increasing pressure ($\pm 10 hPa$). Less effect of rate on pressure is observed for [p'] and little effect on pressure is observed for [↓k͡ʟ]. Concerning [p'] in position 1, CJ decreases the pressure about 16hPa at fast rate while for VP and GA, pressure is more or less stable. In position 5, there is a decrease of the pressure for VP. In position 6 we observe a decrease of presure for AI and GA and an increase for CJ.

3.2. Speed rate effects on constriction area

The estimation of relative constriction area allows us to infer changes in the articulation. The tendency observed in the data is an increase of the constriction area as speed rate increases. In few instances, we observe a decrease of the constriction area as for subject VP in position 1, subject CJ in position 6 (i.e. [p']) and GA in position 7 (i.e. [t͡s']). In few cases, constriction area remains stable (e.g. VP position 7, AI position 6, GA position 4). Non-linear effects are also observed concerning speed rate effects on constriction area. For example, [t͡s'] in position 2, GA shows an increase of the constriction area at 120BPM and a decrease from 120 to 150BPM.

Once again, we observe a consistent tendency across subjects in position 4 and 9. Indeed, the data suggest a larger constriction area for all subjects, except GA in position 4 where constriction remains stable as speed rate increases. In initial position, VP and CJ show a decrease of the constriction area. A similar observation is made from CJ's data in position 6. While pressure remained stable for the *inward k-snare* [↓k͡ʟ], data from CJ shows an increase of the constriction area as tempo increases.

4. DISCUSSION

The data showed temporal reduction of beatboxing patterns ($\rho = -0.97$) and sounds ($\rho = -0.37$). Sounds decreased their duration to a lesser extent suggesting a greater effect on intergestural duration. As hypothesized, when rate is increasing, subjects tended to decrease intraoral pressure and increase constriction area. It was found that *hi-hats* produced in position 4 and 9 were the most affected by increased tempo. Contrary to *kick drums* and *snare drums*, *hi-hats* are not produced in phase with the metronome beat therefore, they constitute weaker positions. Furthermore, position 4 forms a complex interval along with the preceding snare and the following kick drum as shown on Figure 1. Thus, it is not surprising to observe reduction in an interval where multiple targets are to be reached (i.e. velar, dental and bilabial).

Our data also shows cases where the pressure is increased and the constriction area is decreased. These data contradicts our initial hypothesis. We think that timing (i.e. synchronization of the production with the metronome), may help explaining this contradiction. Indeed, Figure 2 and 3 shows that subjects are beatboxing more quickly and slowly than the tempo of reference. This suggests that the timing of production differ from the metronome's pulsation. Measuring and quantifying the deviation from the reference of the metronome for each interval would discard or confirm possible effects of timing on HBB production. An acoustic study is needed to investigate if changes are observed in the acoustic output.

5. CONCLUSION

The data presented here shows an effect of tempo on production of HBB. Though, future studies are needed to understand individual adaptation to increasing rate, particularly to explain the contradictions in our data. A study on timing could be relevant to further understand if timing interacts with tempo and if subjects try to compensate tempo and timing effects (i.e. overshoot at fast tempo). Such study would constitute an original contribution to discuss production mechanisms and motor control with non-linguistic data.

6. ACKNOWLEDGMENT

This physiological protocol was approved by an ethical committee (RCB-ID n° 2020-A00246-33). The research is supported by the Delegation of Research and Innovation of the Hôpital Foch and

LABEX EFL (Empirical Fundation of Linguistics, ANR-10-LABX-0083).

7. REFERENCES

[1] B. Lindblom, "Spectrographic study of vowel reduction," *The journal of the Acoustical society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[2] T. Gay, T. Ushijima, H. Hiroset, and F. S. Cooper, "Effect of speaking rate on labial consonant-vowel articulation," *Journal of Phonetics*, vol. 2, no. 1, pp. 47–63, 1974.

[3] D. J. Ostry and K. G. Munhall, "Control of rate and duration of speech movements," *The Journal of the Acoustical Society of America*, vol. 77, no. 2, pp. 640–648, 1985.

[4] K. Munhall and A. Löfqvist, "Gestural aggregation in speech: Laryngeal gestures," *Journal of Phonetics*, vol. 20, no. 1, pp. 111–126, 1992.

[5] D. Byrd and C. C. Tan, "Saying consonant clusters quickly," *Journal of Phonetics*, vol. 24, no. 2, pp. 263–282, 1996.

[6] H. J. Arkebauer, T. J. Hixon, and J. C. Hardy, "Peak intraoral air pressures during speech," *Journal of Speech and hearing Research*, vol. 10, no. 2, pp. 196–208, 1967.

[7] T. Hixon, "Turbulent noise sources for speech," *Folia Phoniatrica et Logopaedica*, vol. 18, no. 3, pp. 168–182, 1966.

[8] W. Brown Jr and R. E. McGlone, "Relation of intraoral air pressure to oral cavity size," *The Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 323–323, 1969.

[9] A. Malécot, "The effect of syllabic rate and loudness on the force of articulation of american stops and fricatives," *Phonetica*, vol. 19, no. 4, pp. 205–216, 1969.

[10] A. Dehais-Underdown, P. Vignes, L. C. Buchman, and D. Demolin, "Human beatboxing: A preliminar study on temporal reduction," in *12th International Seminar on Speech Production*, 2020.

[11] J. C. Catford, *Fundamental problems in phonetics*. Bloomington And London: Indiana University Press, 1977.

[12] C. H. Shadle, "The aerodynamics of speech," *The Handbook of Phonetic Sciences*, pp. 39–80, 2010.

[13] A. Ghio and B. Teston, "Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies," in *International Conference on Voice Physiology and Biomechanics*, 2004, pp. 55–58.

PIAROA VOICELESS STOPS AS PARTIAL UNDERGOERS OF NASAL HARMONY

Jorge Emilio Rosés Labrada[1], Matthew Faytak[2], Tyler Schnoor[1], Myriam Lapierre[3], Lev Michael[4]

[1]University of Alberta [2]University at Buffalo [3]University of Washington [4]University of California, Berkeley
jrosesla@ualberta.ca, faytak@buffalo.edu, tschnoor@ualberta.ca, mylapier@uw.edu, levmichael@berkeley.edu

ABSTRACT

Phonological analyses of long-distance nasal harmony (LDNH) processes classify segments as triggers, undergoers, blockers, or transparent, based primarily on researchers' subjective judgments of nasality. Using novel field instrumental data from Piaroa (a Jodï-Sáliban language with LDNH), we investigate stop duration, voice onset time (VOT), and nasal and oral airflow during the Piaroa voiceless stops in all-oral environments and continuous nasal harmony spans. We find that nasal harmony has no effect on stop duration and VOT, but voiceless stops in nasal contexts exhibit nasal airflow above an oral baseline at the onset of closure, suggesting that voiceless stops in nasal spans are partial undergoers of LDNH. This partial undergoer behavior is consistent with a model of coactivation of antagonistic gestural specifications for velum activity of the voiceless stop undergoer and the nasal harmony span. Similar voiceless obstruent pre-nasalization is likely widespread cross-linguistically but vastly underreported due to its subtle acoustic effects.

Keywords: nasal harmony; voiceless stops; voice onset time; nasal airflow; Piaroa

1. INTRODUCTION

Piaroa (Glottocode: piar1243; ISO 639-3: pid) is a Jodï-Sáliban language spoken along the Middle Orinoco River and its tributaries in an area that straddles the border between Colombia and Venezuela. Like numerous other South American languages [1, p. 268], Piaroa exhibits long-distance nasal harmony (LDNH). In this paper, we investigate the effect of LDNH on Piaroa voiceless stops, specifically examining their overall duration and their voice onset time (VOT), as well as the rate of nasal and oral airflow during stop production.[1]

1.1. Nasal harmony and consonants

In nasal harmony systems, segments can act either as *triggers*, spreading their nasality to other segments; *undergoers*, undergoing nasalization; *blockers*, stopping the spread of nasality; or *transparent*, neither impeding the spread of nasality nor being (phonologically) affected by it. Which (classes of) segments will exhibit which behaviors is, to some degree, language-dependent. The behavior of voiceless obstruents in particular tends to vary between blocking and transparency. For example, voiceless stops are blockers in Warao [2], but have been characterized as (phonologically) transparent in Paraguayan Guaraní [3].

However, there is some evidence that /p t k/ are affected by nasal spreading in Paraguayan Guaraní and are not fully phonetically transparent. Walker [3] finds that the extent of carryover voicing in intervocalic voiceless stops can be affected: the VOT of /p/ and /t/, but not /k/, is significantly longer in nasal harmony spans. These effects could be caused by carryover nasal airflow into the stop closure as demonstrated for Guaraní in [4].

1.2. Piaroa consonants and Piaroa nasal harmony

The consonant inventory for Piaroa consists of 14 oral stops /p t k kʷ pʰ tʰ kʰ p' t' k' k'ʷ ᵑb ᵑd ʔ/, two affricates /ts' ʧ/, three fricatives /s h hʷ/, two nasal stops /m n/, two glides /w j/, and one rhotic /ɾ/ [5], [6]. Krute [5, p. 62] describes a process of LDNH in Piaroa, which is triggered by phonemically nasal vowels in classifier suffixes and targets preceding vowels and voiced consonants (i.e., voiced stops, glides, and the tap). He argues that nasality spreads leftward to these undergoers, which are realized as their corresponding nasalized allophones, up to the first voiceless consonant preceding the trigger. That is, Piaroa voiceless consonants are *blockers* according to this account.

Recently collected primary fieldwork data, however, shows that Piaroa voiceless stops do not act as blockers, but rather appear to be phonologically transparent. In particular, in verb forms with a masculine classifier /-ɑ̃/ or with the durative suffix /-ɑ̃ɾ/ [7], voiceless stops do not impede the leftward spread of nasalization and they appear to be unaffected in their phonetic realization. To test the behavior of all segments (both consonants and vowels) in nasal harmony spans, instrumental phonetic data was collected in May 2022 with 10 Piaroa speakers from the community of Babel (Atures municipality, Amazonas state, Venezuela).

1.3. Research questions and hypothesis

In this paper, we focus on the behavior of Piaroa's plain voiceless stops /p t k kʷ/, comparing their realization in nasal harmony spans with that outside such spans in an oral, non-nasal harmony environment, with an eye towards comparison with previous findings on Paraguayan Guaraní [3], [4]. We first consider the effects of nasality on overall voiceless stop duration and VOT duration In addition, we examine nasal and oral airflow trajectories to evaluate the timing of the velum-raising gesture for voiceless stops and whether voiceless stops behave as transparent segments or undergoers of LDNH. Our research questions are the following:

1. Is there a difference in stop duration or VOT for voiceless stops in a nasal harmony span *vs.* in an oral environment?
2. During the production of plain voiceless stops in a nasal harmony span, is the level of nasal airflow above that within an oral, non-LDNH environment?

2. METHODOLOGY

The data reported here is from two speakers (1F, speaker JPP; 1M, speaker ROS) and was recorded in a quiet room in Puerto Carreño, Colombia.

2.1. Stimuli and data collection

A wordlist that included two target verbs containing each consonant in the inventory was created by the first author with help from Piaroa native speakers. The male future tense conjugation employs a male classifier suffix,[2] which contains a phonemic nasal vowel /-ã̰/ and triggers leftward LDNH; while the female future tense form of the verb has the female classifier suffix /-æhu/, which contains only oral segments and does not trigger LDNH [7]. All target productions were thus future-tense verb forms, which allowed us to record stimulus pairs with the target segments occurring in similar segmental and prosodic contexts but differing in nasality, as illustrated by contrasting the forms in Table 1.

Oral and nasal airflow were collected using a modified Laryngograph® D-800 electroglottograph and a Glottal Enterprises oro-nasal mask. Stimuli were presented in a different random order for each speaker. Speakers were verbally prompted in Spanish by the first author to provide the Piaroa masculine or feminine present-tense of a given verb (to confirm that the intended root had been identified); they then produced five repetitions of the target future-tense form. This resulted in 10 tokens per verb: 5 in the masculine form (i.e. nasal context) and 5 in the feminine form (i.e. oral context). The 10 words selected for this study (Table 1) yielded 420 tokens (210 per speaker, Table 2). The absence of word-medial /k/ in the data is due to an apparent lexical gap.

Target	Inflected FUT forms	Root gloss
p-, -kʷ-	f. pɑ-ˀd-ækʷ-æhu-sæ m. pã̰-ˀn-ãkʷ-ã̰-sã̰ V-1SG-FUT-CLF-1	'sing, pray'
p-, -kʷ-	f. pæ-ˀd-ækʷ-æhu-sæ m. pæ̃-ˀn-ãkʷ-ã̰-sã̰ V-1SG-FUT-CLF-1	'say'
-p-, -kʷ-	f. hæ-ˀd-ep-ækʷ-æhu-sæ m. hæ̃-ˀn-ẽp-ãkʷ-ã̰-sã̰ V-1SG-SUFF-FUT-CLF-1	'ask for'
t-, -kʷ-	f. te-ˀd-æˀd-ækʷ-æhu-sæ m. tẽ-ˀn-ãˀn-ãkʷ-ã̰-sã̰ V-1SG-SUFF-FUT-CLF-1	'open (hand)'
-t-, -kʷ-	f. t͡ʃĩ-teh-ækʷ-æhu-sæ m. t͡ʃĩ-tẽh-ãkʷ-ã̰-sã̰ 1SG-V-FUT-CLF-1	'shine (light)'
-t-, -kʷ-	f. t͡ʃ-ɑˀdit-ækʷ-æhu-sæ m. t͡ʃ-ãˀnĩt-ãkʷ-ã̰-sã̰ 1SG-V-FUT-CLF-1	'work'
k-, -kʷ-	f. ke-ˀd-æˀd-ækʷ-æhu-sæ m. kẽ-ˀn-ãˀn-ãkʷ-ã̰-sã̰ V-1SG-SUFF-FUT-CLF-1	'finish'
k-, -p-, -kʷ-	f. kæ-ˀd-ep-ækʷ-æhu-sæ m. kæ̃-ˀn-ẽp-ãkʷ-ã̰-sã̰ V-1SG-SUFF-FUT-CLF-1	'lift'
kʷ-, -kʷ-	f. kʷæ-ˀd-ækʷ-æhu-sæ m. kʷæ̃-ˀn-ãkʷ-ã̰-sã̰ V-1SG-FUT-CLF-1	'kill/hit'
kʷ-, -kʷ-	f. kʷo-ˀd-ækʷ-æhu-sæ m. kʷõ-ˀn-ãkʷ-ã̰-sã̰ V-1SG-SUFF-FUT-CLF-1	'swim'

Table 1: Target words and segments analyzed here (V = verb root, f. = female form, m. = male form). All tokens translate to "I will V".

Segment	Word-initial	Word-medial
/p/	40	40
/t/	25	35
/k/	40	0
/kʷ/	40	200

Table 2: Number of tokens of each segment by word position and place of articulation.

2.2. Data processing and analysis

Data were segmented manually in Praat [8]. Start of closure and end of release were located using the drop (or rise) in the amplitude of low-frequency audio signal components. In the case of word-initial stops, speaker JPP did not pause between repetitions of stimuli, leading to a clearly identifiable closure

portion. Speaker ROS did pause between stimulus repetitions. To create suitable time-series data for GAMM analysis, the start of closure from ROS's word-initial stops was estimated from oral or nasal airflow (at the cessation of airflow corresponding to exhalation or inhalation). The timing of stop release was determined based on burst energy in the Praat spectrogram or, if this was not sufficient, from the appropriately timed spike in the oral airflow channel. Fig. 1 illustrates the outlined annotation scheme.

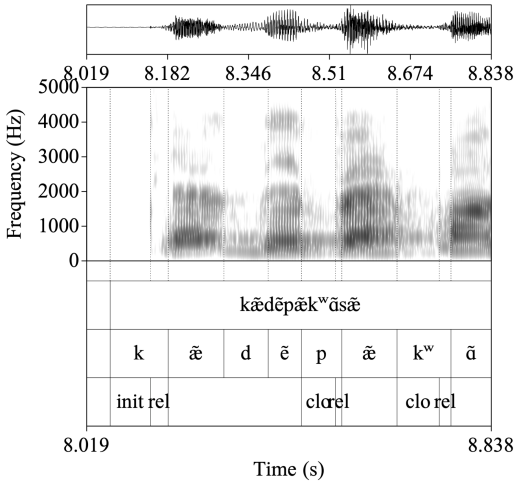


Figure 1: Annotated partial 'lift' male-form token by speaker ROS.

Using these reference points, VOT (release) and the duration of the entire stop (closure+release) were calculated using a Praat script. For 28 of the 420 stop tokens annotated, no clear release phase could be identified. These were excluded from VOT analysis but included in the stop duration analysis. Initial stops were excluded from analysis for stop duration. Speaker ROS's estimated initial closures were not used as a measure of duration and only used to define the time-series data used for GAMM models.

Stop duration and VOT data were submitted to separate linear mixed-effects regressions in R using *lme4* v1.1-31, with *p*-values calculated using *lmerTest* v3.1-3 [9], [10]. Two VOT models were constructed, one for word-medial stops and the other for word-initial stops. The stop duration model excluded all word-initial stops. Models assessed the dependent variable with respect to fixed effects of nasality, stop place, speaker,[3] and their interactions, with random intercepts for word.

Oral and nasal airflow signals were extracted for all target stops using a custom Python script. Airflow signals were smoothed using a Butterworth filter implemented with the *butter* function in the Python *scipy* library (50 Hz cutoff, third-order), downsampled to 200 Hz, and submitted to generalized additive mixed models (GAMMs) built with *mgcv* in R [11]. GAMMs were chosen for their ability to visualize the entire trajectory of airflow over the duration of the consonant, in line with other recent work on time-varying aerodynamic data [12], [13]. Separate GAMMs estimated oral and nasal flow over percent duration as a function of nasality, with factor smooths for speaker and position (initial *vs.* medial) and a random-effect smooth for word. Difference smooths were created using *tidymv* [14].

3. RESULTS

3.1. Stop duration and VOT

In reporting the results of the VOT and duration models, we focus on the presence or absence of main effects of nasality and its interactions with speaker and stop place. We begin with the (medial-only) stop duration model. Here, the main effect of nasality on segment duration fails to reach significance (β=-0.00711, t=-1.054, p=0.294), along with all interactions of nasality with speaker and stop place. Turning to the word-medial VOT model, the main effect of nasality again fails to reach significance (β=.00114, t=0.303, p=0.763), along with all interactions of nasality with speaker and stop place.

In the word-initial VOT model, the main effect of nasality once again fails to reach significance (β=0.000495, t=-0.136, p=0.894). However, the *interaction* of nasality and Speaker ROS reaches significance (β=-0.0165, t=-3.221, p=0.00671), suggesting a slight *shortening* effect of nasality on VOT which is particular to Speaker ROS in initial position. The three-way interaction of nasality, Speaker ROS, and /kʷ/ is also weakly significant (β=0.0172, t=2.372, p=0.0339), suggesting a moderating effect on the VOT of word-initial /kʷ/ specifically in nasal spans.

To summarize, modeling suggests that Piaroa medial stop duration was not affected by nasal harmony spans, in line with Walker's [3] findings for Paraguayan Guaraní, where no effect of nasality on the duration of voiceless stops was observed. However, Piaroa medial stop VOT is likewise not affected, which is *not* in line with Walker's findings: VOT for Paraguayan Guaraní /p/ and /t/ (but not /k/) lengthened within a LDNH span.

3.2. Nasal and oral airflow trajectory

Since VOT and duration data offer indirect evidence for an impact of nasal harmony on voiceless stop production, we now turn to the airflow data for more direct evidence. GAMM fitted smooths for nasal and oral airflow are given in Fig. 2. In nasal harmony spans, compared to all-oral spans, Piaroa plain voiceless stops exhibit elevated nasal airflow from the onset of stop closure, persisting through the first third of the stop's duration. This continued airflow can be

understood as partial nasalization of the voiceless stop due to a delay in the stop's associated velum-raising gesture under the nasal harmony span. Demolin [4] observes that Paraguayan Guaraní voiceless stops in nasal spans exhibit a similar pattern, seemingly as a consequence of a spike in nasal airflow at the end of a nasalized vowel just before stop closure. While preceding vowels are beyond the scope of this paper, this suggests an avenue for further research.

Oral airflow has a much greater variance; in nasal harmony spans, it differs from that during oral, non-harmonized spans mainly at the onset of the stop closure, where fitted airflow in the oral span briefly exceeds fitted airflow in the nasal span. The source of this reduction in oral airflow is less obvious; we speculate that it may be a consequence of globally reduced oral airflow during nasal spans.

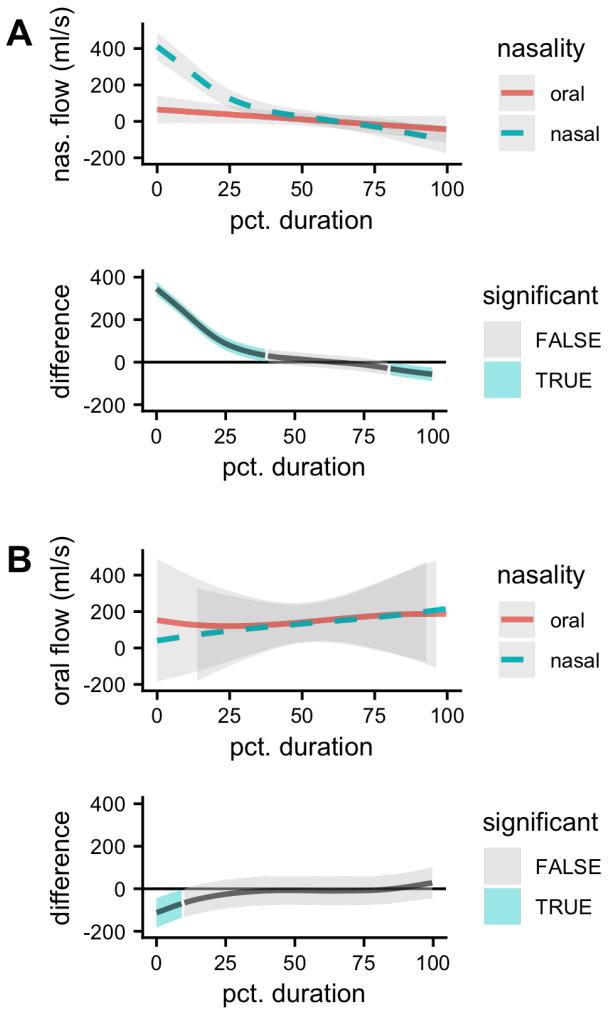


Figure 2: GAMM fits and difference smooths for nasal airflow (A) and oral airflow (B) in oral and nasal conditions.

4. DISCUSSION

While stop duration and VOT data did not suggest that nasal harmony spans affect the production of /p t k kʷ/, our airflow data offer direct evidence that voiceless stops in Piaroa are partial undergoers of nasal harmony, with the initial portion of their closure realized with relatively high nasal airflow (in line with the Paraguayan Guaraní airflow data reported in [4]). This prenasalization process is observed in several other Amazonian indigenous languages. For instance, nasal airflow data for Panãra (Jê) shows that /p t s k/ are prenasalized immediately after phonemically nasal vowels [15]. Lapierre [15, p. 27] discusses additional cases of obstruent prenasalization and suggests that this phenomenon is especially common among Amazonian languages. We speculate that voiceless obstruent prenasalization is underreported in the literature due to its low perceptual salience.

That Piaroa voiceless stops are partial undergoers of LDNH has important implications for models of phonological representation, as it suggests that different portions of a segment (or subsegments) may not behave uniformly as triggers, undergoers, transparent segments or blockers within a LDNH system. In Piaroa, the initial portion of voiceless stop closure behaves as an undergoer, while the latter portion behaves as transparent. Similarly, complex nasal segments in Tupí-Guaraní languages may behave as partial triggers and partial blockers of LDNH [16]. These findings can be captured by formal phonological models of subsegmental representations, such as Q-Theory [15]–[17].

The Piaroa airflow data presented here are equally consistent with an account of partial gestural deactivation under conflicting gestural specifications from the voiceless stop and the nasal harmony span. Per Smith [18], when a segment specified with a velum-raising gesture occurs inside a nasal harmony span specified for a velum-lowering gesture, antagonism between the two gestures arises. Smith proposes that obstruent (partial) transparency (and partial undergoing) in nasal spreading results from favoring the obstruent's velum raising gesture. Specifically, the velum begins to raise during the production of oral constriction for the oral obstruent, and some nasal airflow persists into the stop closure. Upon stop release, the harmony span's velum lowering gesture is reactivated for any adjacent segments in the nasal span that are compatible with a lowered velum gesture. If this gestural account of obstruent transparency within LDNH is correct, we should indeed expect voiceless obstruents in many other languages to show a similar pattern of prenasalization as Piaroa.

5. REFERENCES

[1] L. Campbell, "Typological characteristics of South American indigenous languages," in *The Indigenous languages of South America. A Comprehensive Guide*, L. Campbell and V. M. Grondona, Eds., Berlin; Boston, MA: De Gruyter Mouton, 2012, pp. 259–330.

[2] L. Peng, "Nasal harmony in three South American languages," *International Journal of American Linguistics*, vol. 66, no. 1, pp. 76–97, 2000.

[3] R. Walker, "Guaraní voiceless stops in oral versus nasal contexts: An acoustical study," *Journal of the International Phonetic Association*, vol. 29, no. 1, pp. 63–94, Jun. 1999.

[4] D. Demolin, "The phonetics and phonology of nasal harmony and nasal segments in Guarani," presented at the Nasal 2009, Montpellier, Jan. 2009. [Online]. Available: https://www.researchgate.net/publication/228985362_The_phonetics_and_phonology_of_nasal_harmony_and_nasal_segments_in_Guarani

[5] L. D. Krute, "Piaroa nominal morphosemantics," PhD, Columbia University, 1989.

[6] E. E. Mosonyi, "Elementos gramaticales del idioma piaroa," in *Lenguas indígenas de Colombia: una visión descriptiva*, M. S. González de Pérez and M. L. Rodríguez de Montes, Eds., Santafé de Bogotá: Instituto Caro y Cuervo, 2000, pp. 657–668.

[7] J. E. Rosés Labrada, "The Sáliban languages," in *Amazonian Languages, An International Handbook*, P. Epps and L. Michael, Eds. Berlin and New York: De Gruyter Mouton, forthcoming.

[8] P. W. Boersma, "Praat: doing phonetics by computer [Computer Program] Version 6.1.39." 2022.

[9] D. Bates, M. Maechler, and B. Bolker, "lme4: Linear Mixed-Effects Models using 'Eigen' and S4. R package, version 1.1-31." 2022.

[10] A. Kuznetsova, P. Brockhoff, R. Christensen, and S. Jensen, "lmerTest: Tests in Linear Mixed Effects Models. R package, version 31.-3." 2020.

[11] S. Wood, "Package 'mgcv'. R package, version 1.8-41." 2015.

[12] H. M. Stoakes, J. M. Fletcher, and A. R. Butcher, "Nasal coarticulation in Bininj Kunwok: An aerodynamic analysis," *Journal of the International Phonetic Association*, vol. 50, no. 3, pp. 305–332, Dec. 2020, doi: 10.1017/S0025100318000282.

[13] F. Desmeules-Trudel and M. Brunelle, "Phonotactic restrictions condition the realization of vowel nasality and nasal coarticulation: Duration and airflow measurements in Québécois French and Brazilian Portuguese," *Journal of Phonetics*, vol. 69, pp. 43–61, Jul. 2018, doi: 10.1016/j.wocn.2018.05.001.

[14] S. Coretta, "Tidymv: tidy model visualisation for generalised additive models. R package, version 3.3.2." 2020.

[15] M. Lapierre, "Towards a theory of subsegmental and subfeatural representations: The phonology and typology of nasality," PhD dissertation, University of California, Berkeley, Berkeley, CA, 2021. Accessed: Jan. 02, 2023. [Online]. Available: https://escholarship.org/uc/item/9x90w7c9

[16] M. Lapierre and K. Russell, "Opposite directionality in [+/–F] agreement: The case of Tupí-Guaraní nasal harmony," presented at the 97th Annual Meeting of the Linguistic Society of America, Denver, CO, 2023.

[17] S. S. Shih and S. Inkelas, "Autosegmental aims in Surface-Optimizing Phonology," *Linguistic Inquiry*, vol. 50, no. 1, pp. 137–196, Jan. 2018, doi: 10.1162/ling_a_00304.

[18] C. Smith, "Nasal spreading as defective gestural deactivation," *Proceedings of the 14th Annual Meeting on Phonology*, vol. 2, no. 0, Art. no. 0, Jun. 2016, doi: 10.3765/amp.v2i0.3771.

[19] J. E. Rosés Labrada, "The Piaroa subject marking system and its diachrony," *Journal of Historical Linguistics*, vol. 8, no. 1, pp. 31–58, 2018.

[1] This research was funded by the US National Science Foundation (award #1918064) and carried out under UC Berkeley IRB protocol 2019-06-12311 and University of Alberta REB protocol Pro00075931. We would like to thank the speakers of Piaroa from the community of Babel who participated in this research.

[2] For additional details about the future tense conjugation (including the use of the classifiers /-ɑ̃/ and /-æhu/ as indexes of subject grammatical gender), see [19].

[3] Note that speaker is coded as a fixed effect, rather than as a random effect, in all models owing to the small sample size.

THE LOMBARD EFFECT ON THE VOWEL SPACE OF NORTHERN VIETNAMESE

Giang Le[1], Yan Tang[1,2]

[1]Department of Linguistics, University of Illinois Urbana-Champaign, USA

[2]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, USA

gianghl2@illinois.edu, yty@illinois.edu

ABSTRACT

This study investigates the Lombard effect triggered at different noise levels in Northern Vietnamese, as measured by formant, vowel space area, and vowel dispersion changes. As Lombard speech is generally found more intelligible than speech produced in quiet, we hypothesize that the vowel space area and between-group vowel dispersion would increase in general, in order to maximize vowel-tone contrasts. Acoustic analyses on speech produced in quiet and two noise levels show that contrary to expectation, the vowel space area decreased significantly with the increase of background noise level, as a consequence of the increase of the first formant for most vowels and the decrease of the second formant for non-back vowels. However, the variability within vowel groups decreased, showing better clustering of the same vowel sounds within the vowel space. These findings present a complex picture of the intelligibility of Lombard speech in a tonal language.

Keywords: Lombard speech, Northern Vietnamese, tone, vowel space

1. INTRODUCTION

The Lombard effect is talkers' involuntary, physiological, and vocal responses when speech communication takes place in a noisy environment. It has been extensively studied in the literature. While conclusive findings have been identified and reported for certain features of Lombard speech (e.g. increased duration of sonorous segments [1, 2, 3], increased F0 [1, 4, 3, 5], flattened spectral tilt [3], among others), findings regarding the vowel space have not always been consistent. Some studies reported changes in F1 and F2 for particular vowels of Lombard speech; others found that the vowel space area could either increase or decrease.

Often considered to be a type of clear speech, Lombard speech exhibits acoustic properties that other clear speech registers also do. For example, infant-directed speech (IDS) was found to have increased vowel duration and F0. Similarly, Ping et al. (2017) observed vowel space expansion in both IDS and Lombard speech. The authors explained that in an expanded vowel space, the vowel categories have less overlap and their contrasts are maximized [6]. This is conducive to the infants' perceptual development of phonemic contrasts that IDS helps to facilitate, and better intelligibility to overcome energetic masking in non-ideal listening environments. Smiljanic et al. (2017) also reported that noise-adapted speech and clear speech share similar features, such as a decreased speaking rate and increased energy in 1-3 kHz, sound pressure level (SPL), vowel space area and harmonics-to-noise ratio. Studying the acoustic-phonetic characteristics of speech produced in noise while wearing an oxygen mask [7], Bond et al. (1989) also reported a somewhat increased vowel area.

Other studies, however, reported either no evidence of vowel space expansion [8, 9, 10, 11] or vowel dispersion reduction [12]. Scobbie et al. (2012) remarked that there was a shift in the F1-F2 space due to an F2 increase for the back vowels, but overall the vowel space did not increase in area. Kim et al. (2013) concluded that there was a tendency for the formant values to be produced with less variation, but overall the vowel space was not expanded for speech produced in noise. In a study about speech produced in MRI noise, Gully et al. (2019) observed that the vowel space dispersion decreased, whereby the vowel space dispersion was measured by the mean Euclidean distance of vowel space vertices from the centroid.

In light of these mixed findings, this study analyzed the first (F1) and second (F2) formants, vowel space area formed by four vowels (/a/, /e/, /o/ and /u/), and a vowel space dispersion in Northern Vietnamese Lombard speech. This was to

understand better how the Lombard effect impacts the articulation of vowels in a tonal language.

2. METHOD

2.1. Participants

Six native speakers of the Northern Vietnamese variety (three males, three females) between the ages of 19 and 34 were recruited to participate in this study. A hearing screening showed that all the participants had a normal hearing level.

2.2. Materials

Seventy-eight tokens in vowel, consonant-vowel and consonant-vowel-consonant formants were created for this study. These stimuli were embedded in a carrier sentence *"Toi noi cho ban nghe X bay gio"* (I say X to you now), where X was one of the target tokens. All the stimuli contained one of the three corner vowels /e/, /a/ or /u/. The vowel /i/ was not chosen because of the higher incidences of vulgar words when combined with certain consonants in the stimuli, which could trigger the participants to react in an unexpected manner. /o/ was also elicited as a backup for /u/ for one syllable template due to the same concern. All the syllables were combined with all possible tones. Note that not all the tokens are possible meaningful words in Vietnamese, but the combination of the tones on them is phonotactically possible.

2.3. Procedure

Three recording sessions took place in a sound-treated audio booth. Apart from the control condition, where participants uttered the stimuli *in quiet* ("quiet"), speech-shaped noise was presented to participants at 78 dB SPL ("78-dB SPL") and 90 dB SPL ("90-dB SPL") over a pair of open-back headphones in the other two sessions to elicit Lombard effect of different degrees. The speech-shaped noise was generated to have a representative long-term average spectrum of the Northern Vietnamese variety; it resulted in similar energetic masking on the target speech across frequencies. The participants were instructed to speak the stimuli displayed on a computer screen aloud twice each. While the stimuli were displayed to the participants in groups of the same syllable base, the tone ordering was randomized and the order by which the syllable bases were presented was also randomized. The participants could control how fast they moved through the stimuli via a mouse click. The recordings were saved as WAV files sampled at a rate of 44.1 kHz. In total, 5,136 vowel tokens were collected and analyzed.

2.4. Northern Vietnamese tones

Vietnamese is a tone language where a syllable can carry different pitch patterns, signifying semantic contrasts. Six phonemic tones in Northern Vietnamese can be described as follows.

Tone A1 is a level tone spoken with a modal voice. **Tone A2** is a low to mid-falling tone usually spoken in a modal voice but could also be spoken with a lax or breathy voice [13]. **Tone B1** is a mid-rising tone spoken with a modal voice. **Tone B2** is a mid-falling tone with strong glottalization at the end, or mid-falling *with creakiness*. **Tone C1** is also a falling tone, with a similar F0 contour as tone A2, but with slight laryngealization at the end [13]. Some speakers realize this tone with a mid-falling-rising contour, *similar to the contour of C2*. **Tone C2** is a rising tone with a glottal interrupt in its first half, also known as mid-rising *with creakiness*.

3. RESULTS

3.1. F1

A three-way ANOVA with repeated measures suggested that the effects of noise, vowel, and tone were found significant on the first formant in Hz across the examined sonorous segments and so were the two-way interactions [$\forall p < 0.001$]. Post-hoc comparisons looking at the vowel-noise interaction confirmed a significant increase in F1 for all the vowels from "quiet" to "78-dB SPL" and to "90-dB SPL" [$\forall p < 0.001$], except for the low vowel /a/ (Fig 1). Post-hoc comparisons looking at the tone-noise interaction showed that F1 increased for all tones from "quiet" to "78-dB SPL" [$\forall p < 0.001$] but there was no significant difference between the two noise conditions. The only exception was tone C2, where F1 increased significantly in "90-dB SPL" compared to "78-dB SPL" (Fig 2).

3.2. F2

For F2, a three-way ANOVA with repeated measures also found that noise, vowel, and tone as the main effects, and bi-factor interactions (i.e. noise-vowel, vowel-tone) were significant [$\forall p < 0.001$]. Post-hoc comparisons looking at the noise-vowel interaction showed that F2 increased significantly from "quiet" to "78-dB SPL" and to "90-dB SPL" [$p < 0.001$] for the back vowel /u/ but decreased significantly for

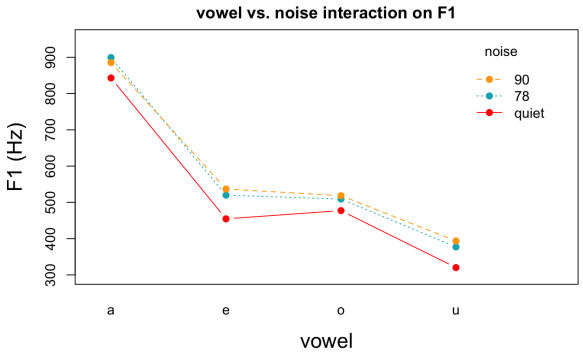


Figure 1: Interaction between vowel and noise on F1

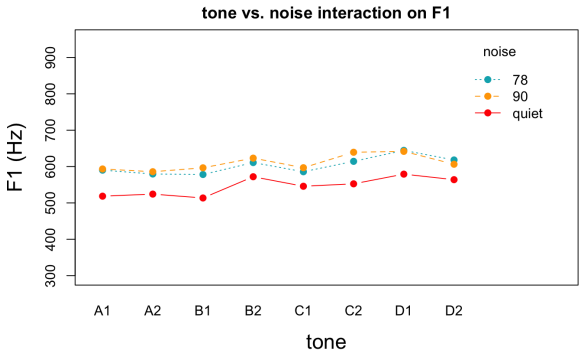


Figure 2: Interaction between tone and noise on F1

vowels /a/ and /e/ (Fig 3). The increase in F2 for a back vowel result agreed with a previous finding by Scobbie et al. (2012).

3.3. Vowel space area and dispersion

The areas of the vowel space by tones and vowels were calculated as the area of the convex hull enveloping all the vowel tokens of each group, using the *phonR* package in R [14]. A two-way ANOVA showed that having vowel and noise as predictors and controlling for tones, noise has a significant main effect on the convex hull area [$p < 0.001$, F=34.257] but vowel has no significant effect; specifically, the area decreased by 116444.3 Hz^2 in "78-dB SPL" compared to "quiet", and 120248.2 Hz^2 in "90-dB SPL" compared to "quiet". There was no significant difference in the vowel space area between the two noise conditions. The same result was found when controlling for vowels and having tone, noise as the main effects (See Fig 4). Post-hoc comparisons looking at the tone-noise interaction showed that while all phonemic tones followed the general pattern, tone C2 only had a significant difference in the vowel space area between "78-dB SPL" compared to "quiet" but not between the other

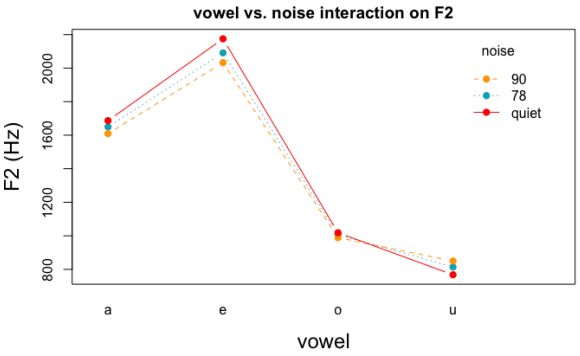


Figure 3: Interaction between vowel and noise on F2

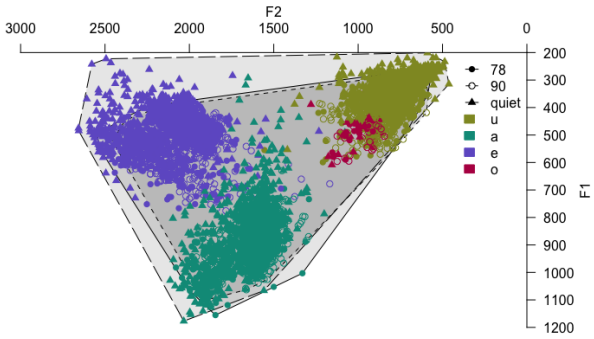


Figure 4: Vowel space area for all vowel tokens

conditions (See Fig 5).

We further examined the within- and between-group dispersions of the F1-F2 (Hz) vowel space area for different vowel groups. In the first measure, three centroids were identified for the three vowel groups, and the within and between-group sums of squares (SS) were calculated (See Table 1). It is evident that both the within and between SS measures decreased going from "quiet" to noise conditions, as an empirical observation. The second measure of dispersion is based on Whitfield and Goberman (2014)'s articulatory-acoustic vowel space's standardized general variance (SGV) [15], which is calculated as the positive p-th root of the determinant of a variance-covariance matrix of F1 and F2 vectors. The SGV can be considered a bivariate standard deviation and thus reflects the degree to which the data spreads out in an F1-F2 space. Our calculated values showed that the SVG decreased from "quiet" (133990.1) to "78-dB SPL" (115028.3), and further to "90-dB SPL" (101029.6).

4. DISCUSSION AND CONCLUSION

As F1 is inversely correlated to vowel height in general, producing a further higher F1 in Lombard

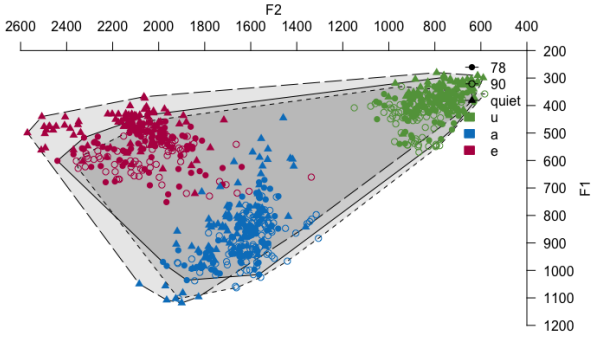


Figure 5: Vowel space area for vowels with tone C2

Table 1: Measures of dispersion within and between clusters

	total within SS	between SS
quiet	63,860,706.00	6.61E+08
78 dB SPL	44,502,705.00	5.66E+08
90 dB SPL	39,729,236.00	4.82E+08

speech could be difficult as a speaker's capacity for hyper-articulation could have reached the ceiling. This may explain why this study observed a significant increase in F1 for all the vowels but the low vowel /a/. Gully et al. (2019) also noted that the increase in F1 in noise is more evident for more closed vowels than for open vowels. As an increase in F1 is associated with a greater degree of jaw-lowering, the results in this study imply varying degrees of hyper-articulation for vowels in Lombard speech.

While F1 for most vowel sounds appeared to increase consistently, F2 seemed less predictable with the pattern changed over vowel backness as exhibited in Fig. 1 and 3. F1 is understood to be correlated with the volume of the back cavity in the vocal tract; hyperarticulation triggered by the Lombard effect could require the jaw to lower further in order to form a larger front cavity, leading to a further contraction of the back cavity, hence the increased F1. Because of the expansion of the front cavity, F2 decreased as a result, especially for non-back vowels such as /a/ and /e/. However, the F2 of back vowels such as /o/ and /u/ appear to be less affected by the change of the vocal configuration in noise, or even be affected conversely, possibly due to the complex acoustic coupling of the two cavities in the vocal tract when different sounds are produced, as shown in Fig. 3. Further investigation is thus warranted. Contrary to some previous findings, we found a decrease in the vowel space area going from quiet to higher noise levels (i.e. "78-dB SPL" and "90-dB SPL") in general. The observed increase in F1 for /e/, /o/ and /u/ but insignificant change for /a/ and a decrease in F2 for /a/ and /e/ but an increase for /u/ together seems adequate enough to shift the centroid of the vowel groups towards the convergence, hence the contraction of the overall vowel space formed by the four vowels as illustrated in Fig. 4. In the meantime, clearer clustering between vowel groups might interfere with the vowel space in the sense that this reduction in variability also leads to a reduced vowel space area produced by speakers in noise. How these two factors interact and what implications they have for the perception of vowels in noise are not yet clear; nonetheless, that reduced vowel space but tighter, more evident clustering might suggest a different type of intelligibility than a broader vowel space might. Additionally, while an expanded vowel space tends to be observed from "clear" speech, such as infant-directed speech or hearing impaired-directed speech, a reduced vowel space can happen in Lombard speech due to a change in speaking rate. Cowley (2020) also reported similar findings obtained here: the vowel space area decreased in Lombard speech. The author speculated that distracting noise could induce a faster speaking rate, leading to reduced articulatory movements and a smaller vowel space area [16]. Our findings also lent indirect support to the dubious correlation between speech's loudness and an expanding vowel space, as observed in [17], where elicited loud speech did not necessarily lead to formant variation and changes in the vowel space.

Our previous research [18] identified limitations to hyper-articulation as noise conditions were manipulated. For example, a reduced Lombard effect following increased noise levels, as measured by the range of F0 values, was detected in tone C2. In this study, we found that different vowels induced different degrees of formant changes in noise: F1 increase was not observed for the low vowel /a/ and F2 changes differed between back vowel and non-back vowels. Given that formants are resonances intensified by filtering effects in the supralaryngeal vocal tract, our findings suggested that filtering effects were impacted by hyper-articulation in varying degrees in Lombard speech. A reduced vowel space suggested that like loudness, an expanded vowel space was not always an automatic response to noise for better intelligibility. The high number of phonemic tonal contrasts in the language calls for greater intelligibility of Lombard speech, which is achieved by tighter clustering rather than increased vowel area.

5. REFERENCES

[1] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers." *Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510 – 524, 1993.

[2] T. D. Hanley and M. D. Steer, "Effect of level of distracting noise upon speaking rate, duration and intensity." *Journal of Speech & Hearing Disorders*, vol. 14, pp. 363 – 368, 1949.

[3] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck, "An acoustic and articulatory study of lombard speech: global effects on the utterance," in *INTERSPEECH*, 2006.

[4] W. Van Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses." *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[5] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1495–1499, 2009. [Online]. Available: https://doi.org/10.1121/1.3179668

[6] T. Ping, R. Nan Xu, I. Yuen, and K. Demuth, "Phonetic enhancement of mandarin vowels and tones: Infant-directed speech and lombard speech." *Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 493 – 503, 2017. [Online]. Available: http://www.library.illinois.edu.proxy2.library.illinois.edu/proxy/go.php?url=https://search-ebscohost-com.proxy2.library.illinois.edu/login.aspx?direct=true&db=mah&AN=124939671&site=eds-live&scope=site

[7] Z. D. Bond, T. Moore, and B. Gable, "Acousticâphonetic characteristics of speech produced in noise and while wearing an oxygen mask," *The Journal of the Acoustical Society of America*, vol. 85, pp. 907–12, 03 1989.

[8] C. Davis and J. Kim, "Is speech produced in noise more distinct and/or consistent?" in *Australasian Speech Science and Technology Association*, 2010.

[9] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," *Computer Speech Language*, vol. 28, no. 2, pp. 629–647, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230813000788

[10] J. Scobbie, J. Ma, and J. White, "The tongue and lips in lombard speech: A pilot study of vowel-space expansion," 09 2012.

[11] J. Kim and C. Davis, "Comparing the consistency and distinctiveness of speech produced in quiet and in noise," *Computer Speech and Language*, vol. 28, no. 2, pp. 598 – 606, 2014, cited by: 16. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84890564861&doi=10.1016%2fj.csl.2013.02.002&partnerID=40&md5=ee7f5403484987afa0fcd3203a051fca

[12] A. J. Gully, P. Foulkes, P. French, P. Harrison, and V. Hughes, "The lombard effect in mri noise," 2019.

[13] V. L. Nguyen and J. Edmondson, "Tones and voice quality in modern northern vietnamese: Instrumental case studies," *Mon-Khmer Studies*, vol. 28, pp. 1–18, 1997.

[14] D. R. McCloy, *phonR: tools for phoneticians and phonologists*, 2016, r package version 1.0-7.

[15] J. Whitfield and A. Goberman, "Articulatory-acoustic vowel space: Application to clear speech in individuals with parkinson disease," *Journal of Communication Disorders*, vol. 51, 09 2014.

[16] C. M. Cowley, "The effects of distracting background audio on speech production," MS Thesis, Brigham Young University, 2020.

[17] L. Koenig and S. Fuchs, "Vowel formants in normal and loud speech," *Journal of Speech Language and Hearing Research*, vol. 62, pp. 1278–1295, 04 2019.

[18] G. Le, C. Shih, and Y. Tang, "Distortion in tone production due to the lombard effect," 2021, 1st International Conference on Tone and Intonation (TAI) 2021 : Tone and Intonation in a globalized, digital world, TAI2021 ; Conference date: 06-12-2021 Through 09-12-2021.

TRIGGERING AND MAINTENANCE OF TONGUE BRACING POSTURE IN MULTI-SYLLABLE SEQUENCES

Annabelle Purnomo[1], Yadong Liu[1], Nicole Ebbutt[1], Bryan Gick[1,2]

[1]Department of Linguistics, University of British Columbia, [2]Haskins Laboratories, New Haven
apurnomo@student.ubc.ca

ABSTRACT

Lateral tongue bracing is a posture that is pervasively maintained in speech across languages, released only for a few select sounds [1]. Previous pilot work with English showed that the lateral tongue maintained a higher position in the bracing-neutral phrase "hubba bubba" when preceded by braced targets, compared to unbraced targets [2]. The present study evaluates a larger number of speakers producing this same sequence with additional analyses. Mean and range results for lateral tongue movement support the view that the lateral bracing posture, once initiated, is maintained through bracing-neutral sounds. No anticipatory (right-to-left) effect was observed.

Keywords: lateral tongue, lingual bracing, speech posture, embodied speech, harmony.

1. INTRODUCTION

Lateral tongue bracing is a lingual posture in which the sides of the tongue are held in contact with the palate and upper molars [3]. Early studies found lateral bracing to be associated with lingual sounds like /s, z, n/ [4]. Gick et al.'s [5] study on English found that lateral tongue bracing was maintained throughout over 97% of continuous speech, released during only laterals and occasional low vowels. Lateral tongue bracing has been found to require dedicated muscle activation [5]. It has been observed in both English-speaking adults and children [6], is robust to perturbation [7] and has been documented across many languages including Akan, Cantonese, English, Korean, Mandarin, and Spanish, suggesting bracing as a universal postural basis for speech [1].

Based on the understanding that postures exhibit interactions over a long distance/time [8], the present study asks how and whether the braced lingual posture is initiated and maintained during long sequences of neutral sounds. A handful of sounds appear neutral to bracing, in that they do not require bracing, but do allow it. These neutral segments include non-lingual consonants, e.g., labials, glottals and the English schwa. This continuation of a tonic movement through neutral sounds is reminiscent of phonetic or phonological harmony processes, sometimes described in terms of spreading of a feature onto non-adjacent segments [9], sometimes passing through neutral segments.

It has long been observed that bracing is required during lingual sounds such as /s, z, n/ [4] for which complete or partial lingual obstruction of air flow is required. These sounds can be thought of as "triggers" for bracing. Bracing has also been found to be consistently released during lateral sounds, in particular the prevocalic English /l/ [5], which can thus be thought of as "opaque" to bracing (i.e., requiring the braced posture to be interrupted). Non-lingual or lingually neutral sounds can be thought of as "transparent" to bracing (i.e., segments that can in principle be produced irrespective of whether the tongue is braced or unbraced).

2. BACKGROUND

In a previous pilot study, [2] presented coronal ultrasound imaging data that observed the tongue movements of eight native North American English speakers. Results showed that the sides of the tongue maintained a higher position during the utterance "hubba bubba" [həbəbəbə] preceded by segments that triggered bracing than by segments that are opaque to bracing. "Hubba bubba" (HB) was chosen as the target utterance as it is made up entirely of transparent (bracing-neutral) sounds ([h], [b], [ə]). However, [2] only analysed lateral tongue height (i.e., no data on tongue medial height were included for comparison), and only measured at the beginning of a HB utterance (i.e., maintenance of bracing over time could not be evaluated). A braced tongue should exhibit a higher tongue position and a smaller range of movement compared to the centre throughout the HB utterance, due to the anchoring of the tongue sides maintained on the palate.

The present study extends [2] by: 1) evaluating a larger number of speakers producing the HB sequence; 2) analysing both lateral and medial tongue movement for mean tongue height and range of displacement; 3) measuring tongue position during the middle syllables of the HB sequence to test maintenance of the bracing posture; and 4) evaluating anticipatory (right-to-left) as well as carryover (left-to-right) effects.

The present study tests the hypothesis that lateral tongue bracing is initiated by a bracing trigger, and is maintained until it is interfered with. This hypothesis predicts that a higher overall height and a smaller movement range of the lateral tongue compared to the medial tongue will be observed during braced vs. unbraced syllables, with this difference being maintained through the middle syllables of the HB sequence; this difference should be initiated by a preceding bracing trigger, but not a following one.

3. METHODS

Identical to Ebbutt et al. [2], an ultrasound probe was used to collect coronal plane images of the posterior region of the tongue. An analysis of the vertical tongue position of the sides of the tongue was done to detect the presence of lateral bracing.

A total of 53 participants took part in the ultrasound study. Participants were paid and recruited via word of mouth or via the Linguistics SONA portal at the University of British Columbia and received course credit for their participation. All participants provided informed consent and completed a language background that reported no speech difficulties. Participants were considered native speakers of English if they acquired it before eight years of age and continued to use it as a primary language at work, school, or home. Participant data were excluded if they were not native speakers of North American English or they did not have clear ultrasound imaging. The ultrasound study analysed a total of 12 participants (3 male and 9 female). Seven speakers were excluded because they were not native speakers of North American English, while 34 participants were excluded due to poor ultrasound imaging quality.

The reading stimuli include four blocks of seven sentences, each containing the lingually-neutral target word "hubba bubba" [həbəbəbə]. Preceding or following the target word were words either with lingual consonants known to require tongue bracing (bracing trigger words: chews, wants, has, eats, chewing gum, needs) or words with the lateral /l/, known to interfere with tongue bracing (bracing opaque words: love, lump, plum, lots). The four blocks were given in randomised order, and participants were given one of sixteen possible combinations of the stimuli.

Examples of stimuli from each block:
Block A: Every Monday, he chews Hubba Bubba chewing gum.
Block B: Every Tuesday, we love Hubba Bubba lots.
Block C: Every Wednesday, they lump Hubba Bubba chewing gum.
Block D: Every Sunday, Chris needs Hubba-bubba chewing gum.

Participants were seated in an ophthalmic chair with a 2-cup rear headrest to stabilise their head. A Manfrotto Magic Arm was attached to position an ALOKA ProSound SSD-5000 ultrasound with a 120-degree convex probe. A microphone stand captured acoustic data from participants and a Focusrite preamplifier was used to process the data. A Canopus ADVC-110 box was used to synchronise the audio and video data. iMovie was used to record audio and ultrasound video data. The microphone was held on a stand 2-3 feet in front of the participants and a music stand stood in front of them to hold the stimuli. The ultrasound probe was positioned against the neck behind the chin to show coronal imaging of the posterior region of the tongue, where bracing contact occurs between the tongue and the rear hard palate and molars. The angle of the probe was determined by asking participants to repeat "Mary had a little lamb", a familiar phrase that elicits repeated bracing and release of the lateral tongue [1]. Participants were asked to read the stimuli three times. The study ended with the participant getting debriefed on the nature of the experiment.

Vertical movement was tracked with videokymography (VKG), a technique for plotting video movement over time. Ultrasound video and audio were converted into a .wav file and annotated in Praat [10]. Pilot results showed greater inter-speaker variation for the first repetition, as speakers got used to the unfamiliar and repetitive stimuli. For second and third repetitions, there was found to be more consistency between speakers. As such, the first readings of the stimuli were omitted from the analysis. Each instance of HB was manually annotated using Praat and the corresponding ultrasound imaging video frames were extracted. ImageJ [11] was used to open each set of image sequences where images were adjusted for brightness and contrast in order to get a clearer image of the tongue surface. Finally, images

were thresholded and converted into VKG to create an image of the tongue height over time for each utterance of HB. VKGs were created from the side of the tongue that showed the clearest ultrasound imaging (Figure 1). Speckle noise appearing anywhere in the ultrasound imaging was manually removed.



Figure 1: A thresholded VKG showing vertical tongue movement of the left lateral tongue during several repetitions of HB preceded by braced (left) and unbraced (right) words.

For each HB production, the height of the tongue was determined by the bottom white pixel in each frame in the VKG. For any frame in which a HB production contained no white pixels, linear interpolation was used to predict vertical position from adjacent frames. Mean height of the tongue as well as the range of vertical movement were calculated across all frames for each HB production. Mean height and range of movement were fitted by linear mixed effects models using the *lme4* [12] package in RStudio [13]. Preceding or following conditions (braced vs. unbraced) and regions of the tongue (medial vs. lateral) were fixed effects, and participants were random effects with both random slope and intercept.

4. RESULTS

	Preceding Condition	Tongue Region	Interactions
Mean	$\chi^2(2) = 81.33$, $p < 0.001$	$\chi^2(2) = 87.59$, $p < 0.001$	$\chi^2(1) = 81.33$, $p < 0.001$
Range	$\chi^2(2) = 13.42$, $p < 0.001$	$\chi^2(2) = 14.31$, $p < 0.001$	$\chi^2(1) = 13.42$, $p < 0.001$

Table 1: Table of normalised mean height and range of movement for preceding condition (braced vs. unbraced), tongue region (medial vs. lateral), and their tongue region & preceding condition interactions.

The linear mixed effects model showed that tongue region and preceding bracing condition have a significant effect on the mean height of the tongue. The interactions show significant results, indicating that the tongue region and preceding bracing condition are interdependent.

The results of a post hoc test using the

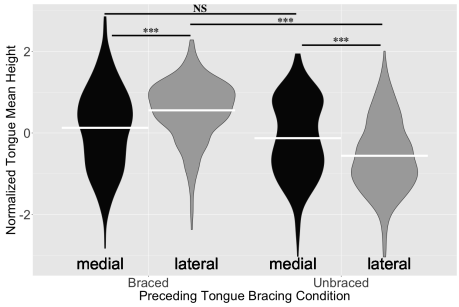


Figure 2: Boxplot comparing normalised mean tongue height of the medial (dark) and lateral (light) regions producing HB with preceding braced vs. unbraced conditions. NS indicates "Not Significant" while *** indicates significance.

emmeans [14] package revealed no difference for the mean height of the medial tongue (t(15.3)= 0.947, p=0.3582) in the preceding braced condition compared to the unbraced condition. This means that the medial tongue remained at a relatively similar height between conditions (see Figure 2). The lateral tongue had a significantly higher mean height (t(15.3)=4.551, p=0.0004) in the preceding braced condition compared to the unbraced condition.

The preceding braced condition showed a significantly lower mean height for medial vs. lateral tongue (t(49.1)=-6.560, p <.0001). In contrast, the preceding unbraced condition had a significantly higher mean medial (t(49.6)=6.554, p <.0001) vs. lateral tongue in the unbraced condition.

	Following Condition	Tongue Region	Interactions
Mean	$\chi^2(2) = 0.8172$, $p = 0.6646$	$\chi^2(2) = 0.812$, $p = 0.6663$	$\chi^2(1) = 0.812$, $p = 0.3675$
Range	$\chi^2(2) = 2.153$, $p = 0.3408$	$\chi^2(2) = 1.939$, $p = 0.3793$	$\chi^2(1) = 1.939$, $p = 0.1638$

Table 2: Table of normalised mean height and range of movement for following condition (braced vs. unbraced), tongue region (medial vs. lateral), and their tongue region & following condition interactions.

The linear mixed effects model showed no significant effect associated with tongue regions and following bracing conditions on the mean height of the tongue. Regardless of following bracing conditions and regions of the tongue, mean height during a HB utterance remains unchanged.

Finally, the results for the range of movement show that the medial tongue shows a similar range in preceding and following braced vs. unbraced conditions ($t(17.5) = 0.212, p = 0.8348$; $t(40.9) = -0.558, p = 0.5796$). No significant difference in

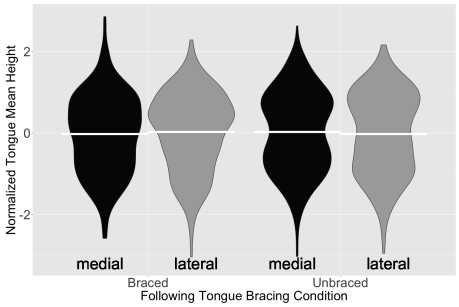


Figure 3: Boxplot comparing normalised mean tongue height of the medial (dark) and lateral (light) tongue producing HB with following braced vs. unbraced conditions. All conditions were found to be not significant.

movement range is observed for the lateral tongue in the preceding and following braced condition compared to the unbraced condition ($t(17.5) = -2.072, p = 0.0533$; $t(40.8) = 1.246, p = 0.2198$), i.e., the sides of the tongue show similar stability across the two conditions. The preceding braced condition shows a significantly higher range of motion for the medial tongue ($t(48.6) = 2.627$, $p = 0.0115$) compared to the lateral tongue; such an effect is not observed in the following braced condition ($t(46.2) = -0.982, p = 0.3312$). The preceding unbraced condition shows a significantly lower range of motion for the medial tongue ($t(49.1) = -2.633, p = 0.0113$) which implies more movement for the lateral tongue when preceded by an unbraced condition. This is also what would be expected if the lateral tongue is not anchored to the palate as during lateral tongue bracing. However, no such an effect is found in the following unbraced condition ($t(45.5) = 0.975, p = 0.3345$).

5. DISCUSSION

The present study investigated the initiation and maintenance of lateral tongue bracing using ultrasound imaging.

Results for mean tongue height show that the medial tongue is consistent irrespective of surrounding bracing triggers. In contrast, the lateral tongue changed height depending on the preceding bracing condition (higher in the preceding braced condition and lower in the preceding unbraced condition). The finding that the lateral tongue maintains a relatively higher position than the medial when preceded by a bracing trigger is consistent with the hypothesis that the lateral bracing posture, once triggered, is maintained throughout neutral syllables. Additionally, the results show that only the preceding braced targets introduce this braced posture, as no significant differences are shown to be associated with a following bracing trigger.

Results for the range of vertical movement of medial tongue were similar to those for mean height, showing that the medial tongue had a similar range of movement across preceding and following bracing conditions. The lateral tongue had a smaller range of vertical movement than the medial tongue when preceded by braced targets, which indicates that the side of the tongue remained more stable.

Given the braced and unbraced behaviour of mean and range of heights, these findings support the hypothesis that tongue bracing triggers a braced tongue posture which is maintained through subsequent bracing-neutral (transparent) syllables. This maintenance of tonic muscle activations across non-adjacent segments is reminiscent of some long-distance phonetic or phonological phenomena such as those relating to harmony processes. These findings advance our understanding of tongue posture and the biomechanics of speech, particularly outside the sagittal plane.

Models of speech planning and production will need to allow for a system in which a single utterance may be implemented in two such dramatically different ways — one with the lateral tongue held in a braced position and one unbraced (e.g., the kinematics and muscle activations for these two sequences would be quite different). We suggest that speech models incorporating posture will be better able to handle such results. An understanding of the fundamentals of posture in speech is important when investigating speech movements, as posture is a critical component of movement, providing a stable substrate upon which intentional movements are built.

Our study has limitations. Due to the nature of ultrasound studies, only tongue position can be measured, not tongue-palate contact. Additionally, only one side of the tongue was analysed and lateral bias exists in tongue movements [15], though lateral bracing is actively maintained for both sides of the tongue [1, 7]. While analysing VKG of individual speakers, distinct patterns across braced and unbraced conditions among individual speakers were observed. Future studies could look into speaker independent patterns and interspeaker variations. Future studies can also consider whether the unbraced posture itself functions as a distinct posture, as similar stability was observed among production of HB between braced and unbraced conditions.

6. REFERENCES

[1] Y. Liu, F. Tong, G. de Boer, and B. Gick, "Lateral tongue bracing as a universal postural basis for speech," *Journal of the International Phonetic Association*, pp. 1–16, 2022.

[2] N. Ebbutt, Y. Liu, A. Purnomo, C. Purnomo, K. Hung, and B. Gick, "Initiation and maintenance of lingual bracing posture," in *Canadian Acoustics*, vol. 50, no. 3, 2022, pp. 96–97.

[3] B. Gick, I. Wilson, and D. Derrick, *Articulatory phonetics*. Malden, MA;Chichester, West Sussex: Wiley-Blackwell, 2013.

[4] S. McLeod, "Australian adults' production of /n/: An epg investigation," *Clinical Linguistics Phonetics*, vol. 20, no. 2-3, pp. 99–107, 2006.

[5] B. Gick, B. Allen, F. Roewer-Després, and I. Stavness, "Speaking tongues are actively braced," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 3, pp. 494–506, 2017.

[6] H. Y. Cheng, B. E. Murdoch, J. V. Goozee, and D. Scott, "Electropalatographic assessment of tongue-to-palate contact patterns and variability in children, adolescents, and adults," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 2, pp. 375–392, 2007.

[7] Y. Liu, S. Luo, M. Łuszczuk, C. Mayer, A. Shamei, G. de Boer, and B. Gick, "Robustness of lateral tongue bracing under bite block perturbation," *Phonetica*, 2023, forthcoming.

[8] J. Massion, A. Alexandrov, and A. Frolov, "Why and how are posture and movement coordinated?" in *Brain Mechanisms for the Integration of Posture and Movement*, ser. Progress in Brain Research. Elsevier, 2004, vol. 143, pp. 13–27, iSSN: 0079-6123. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0079612303430021

[9] G. Hansson, *Theoretical and Typological Issues in Consonant Harmony*. eScholarship, University of California, 2001.

[10] P. Boersma and D. Weenick. (2023) Praat: doing phonetics by computer. [Online]. Available: http://www.praat.org/

[11] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nih image to imagej: 25 years of image analysis," *Nature Methods*, vol. 9, no. 7, pp. 671–675, 2012.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[13] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/

[14] R. V. Lenth, B. Bolker, P. Buerkner, I. Giné-Vázquez, M. Herve, M. Jung, J. Love, F. Miguez, H. Riebl, and H. Singmann, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[15] B. Gick, M. Keough, O. Tkachman, and Y. Liu, "Lateral bias in lingual bracing during speech," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1903–1903, 2018.

SPEECH STABILITY ACROSS TIME: EVIDENCE FROM NORWEGIAN VOWELS IN SPONTANEOUS SPEECH PRODUCTION

Peng Li[1], James Emil Flege[2], Clara D. Martin[3,4], Natalia Kartushina[1]

[1] Center for Multilingualism in Society across the Lifespan, Institute for Linguistics and Scandinavian Studies, University of Oslo, Norway
[2] University of Alabama at Birmingham, USA
[3] Basque Center on Cognition, Brain and Language, Spain
[4] Ikerbasque, Basque Foundation for Science, Spain
peng.li@iln.uio.no; jeflege@uab.edu; c.martin@bcbl.eu; natalia.kartushina@iln.uio.no

ABSTRACT

This study investigated the stability of speech sound production across time. Thirty-four monolingual Norwegian speakers participated in a three-session psycho-socio quiz experiment over three weeks, which was designed to collect spontaneous speech data. Participants had to answer questions orally by choosing either one of three plausible answers or one of two sentences for "I don't know". The "I don't know" sentences contained five target vowels /i, æ, ɑ, o, u/. Using the Pillai score, we analyzed the stability of vowel categories and the distinctness of vowel contrasts across tests. We found that, overall, the tested vowel categories and the three vowel contrasts /i-æ/, /ɑ-o/, and /o-u/ were stable over time. However, the backness-based contrasts /æ-ɑ/ and /u-i/ showed instability. Future research needs to examine the role of phonetic environment, repetition, and prosodic structure on sound variability.

Keywords: stability, vowel, Norwegian, spontaneous speech production, formant analysis

1. INTRODUCTION

Individuals produce speech sounds with considerable variability in both native and nonnative speech [1], which may arise from multiple sources such as dialect [2], the number of distinctive sounds in a given language (i.e., phoneme inventory) [3], and speech style (i.e., elicited vs. spontaneous speech) [4], among others. Recent research has established several lines to study the stability of sound production. Yet, it is unclear whether these differences are stable over time, systemic or sound specific, and whether they generalize to spontaneous speech.

First, very little previous research has evaluated segmental production is stable, that is, sought to determine if individuals produce vowels and consonants the same way in multiple test sessions. Heald and Nusbaum [5] asked American English speakers to produce the English vowels in isolation at three different timepoints on three different days. They found that although the mean values of F0 and F1 changed within one day, there was no significant change in the mean values of the acoustic measures between days. This study showed the stability of the production of vowels in *citation* forms between days. Later, Pierce et al. [6] had native English speakers read English sentences and produce /a/ in isolation for seven consecutive days with three timepoints per day. Their study revealed similar results to Heald and Nusbaum's study. It seems that the time of day had a significant influence on speech production, but speech production is stable on different days at the same time point, as demonstrated by the consistent production of vowels and consonants in repeated readings of scripted passages.

Second, individual differences in speech production can be attributed to other factors [7]. For instance, sound realization might reflect the precision of underlying phonemic perceptual representations [8]. Also, speakers' personalities (e.g., introverted vs. extroverted), mental status (e.g., depression and anxiety), and living habits (e.g., alcohol consumption) may affect the acoustic variability of speech [6]. Research with bilingual speakers found that those who scored higher in L1 intelligibility tended to have more intelligible L2 speech [9]. Likewise, precision in L1 vowel production can affect accuracy in L2 vowel production [10]. Although appealing, these results need to be taken with care, as it is not clear whether these L1 differences in precision are stable, subject-intrinsic, systematic, reflecting higher-order representations, or whether these are momentary modulations in speech production.

Most previous studies measured intra-speaker speech variability using scripted speech production tasks, like paragraph-reading, sentence-reading, and even repetitive production of single vowels [5], [6], focusing participants' attention on speech *per se*. Compared to scripted speech, which often involves a deliberate and careful speaking style, spontaneous and more natural speech production may exhibit greater variability, such as vowel reduction, which can affect the acoustic properties of vowels [11]. A study on the Mixtec language confirmed that vowels

in spontaneous speech are shorter, more contextually assimilated, and show a less dispersed vowel space than in scripted speech [4]. Therefore, caution should be exercised when extrapolating results obtained from elicited speech to spontaneous speech.

Furthermore, the existing methods for eliciting speech samples are lab-based, and the subjects are fully aware that their speech will be acoustically scrutinized. These methods may lead to unintended variations in speech sounds, as participants may deliberately alter their pronunciation in response to the experimental conditions [12]. Therefore, a new technique called Characteristic Speech Production (CSP) was developed to elicit data in which participants concentrate on the meaning of their responses to relevant questions rather than on the pronunciation of specific sounds [13].

This study is the first to use the CSP sampling method to investigate the stability of speech production over multiple days. We developed an online psycho-socio quiz that was administered three times over three weeks to collect samples of spontaneous speech in Norwegian, which is the participants' native language. The target speech sounds were five Norwegian vowels /i, æ, ɑ, o, u/. Note that Norwegian has more vowels, but for the current study, we selected the most frequent ones, three of which were corner vowels /i, æ, u/. We assessed the stability of speech production by analyzing the consistency of vowel categories and the distinctiveness of vowel contrasts using formant analyses. Based on previous research [5], [6] and because the target vowels are distinctive phonemic categories in Norwegian [14], we considered the following hypotheses:

- H1: Vowel categories are stable over time: Norwegian speakers would produce each target vowel in a consistent way across time.
- H2: Distinctness of vowel contrasts is stable over time: Norwegian speakers would show stable vowel pair distinctness across time.

2. METHODS

2.1. Participants

We recruited 34 Norwegian native speakers to participate in this study. Nine participants were excluded: Eight completed only one test, and one returned low-quality recordings across all three tests. The remaining 25 participants (aged 18 – 39, female = 13, male = 12) were analyzed. The participants reported no documented speech disorders or hearing impairment. They all signed a consent form approved by the Norwegian Centre for Research Data. Each of them received a small amount of monetary compensation for their time.

2.2. The psycho-socio quiz

We created 150 questions (50 per testing test) with various degrees of difficulty (see Table 1 for an example):

- *Control questions* (n = 9) required common sense knowledge. E.g., Give the result for 5 +10.
- *Easy questions* (n = 36) required some background knowledge. E.g., Which of these foods is NOT toxic to dogs?
- *Average questions* (n = 45) required basic knowledge in a certain field. E.g., Which is the oldest Norwegian political party?
- *Hard questions* (n = 30) required specific knowledge in a certain field. E.g., What is the atomic number of Uranus?
- *No-concrete answer questions* (n = 30) were questions with no objective or reasonable answer. E.g., Who was Norway's best artist?

Table 1: A trial example from the psycho-socio quiz (English translation is provided for the reader but was not included in the quiz).

	Hvem av disse har verdensrekorden for lengst hår? **"Who among these holds the world record for longest hair?"**
1	*Kate Moss*
2	*Xie Qiuping*
3	*Linda Evangelist*
4	*Jeg vet **i**kke sv**a**ret på spørsmålet h**e**r, men jeg tr**o**r n**o**en **a**ndre k**a**n v**i**te det.* "I don't know the answer to the question here, but I think someone else might know it."
5	***I**ngen k**a**n sv**a**re p**å** dette h**e**r, det **e**r j**o** n**o**e s**o**m **e**r umul**i**g **å** s**i*** "No one can answer this here; it is something that is impossible to say."

On each question, participants were presented with five possible answer choices, including three plausible answers and two long sentences for 'I don't know' (options 4-5 in Table 1). The two long sentences were the target sentences, which were identical across all trials and tests, and contained the target vowels (/i, æ, ɑ, o, u/, indicated in bold in Table 1). Each vowel appeared at least four times over the two sentences. Hard and vague questions triggered as many target sentences as possible to ensure an adequate number of target tokens.

Participants did the experiment on an online survey platform Prolific (https://www.prolific.co/), in a quiet place that they were familiar with (e.g., the living room). The productions were recorded three times (T1, T2, T3) at the same hour of the day, on the same day of the week, over a span of three weeks. At each trial, participants first saw the question on the screen, followed by five options. They had to say one

of the five options/prompts to answer each question. On each day, participants answered 50 unique questions in random order.

2.3. Data coding and analyses

2.3.1. Data screening

We obtained 3,750 recordings from the three tests, each containing one sentence. A Norwegian-speaking research assistant identified 1,557 target sentences. The rest of the utterances (e.g., answers like "Kate Moss" to the example question in Table 1) were discarded. We then excluded 65 recordings due to heavy background noise or poor audio quality. Finally, the remaining 1,492 recordings were used for the acoustic analyses.

2.3.2. Acoustic analyses

We segmented our corpus into words and phones with Montreal Forced Aligner [15]. Since the boundary locations may be inaccurate in spontaneous speech due to deletion and non-canonical pronunciation patterns, we manually checked and corrected all the labels in Praat [16]. In total, we obtained 15,635 tokens. Then, we applied a Praat script bundle to extract the mean values (Hertz) of the first four formants (i.e., F1-F4) from the middle third part of each target vowel [17]. The script bundle extracts formants using the dynamic seeding method [18], taking the vowel formant of Norwegian in previous research as the reference [19], which can minimize errors in formant extraction.

2.3.3. Statistical analyses

To minimize the intrinsic differences across speakers, we converted the Hertz values of F1, F2, and F3 to Bark (Z1, Z2, Z3, respectively). Then, the vowel height was represented by Z3-Z1, and the vowel frontness was represented by Z3-Z2. Fig 1 plots the bark normalized F1 and F2 values with the ellipses representing one standard deviation (SD) for each of the five Norwegian vowels across the three tests.

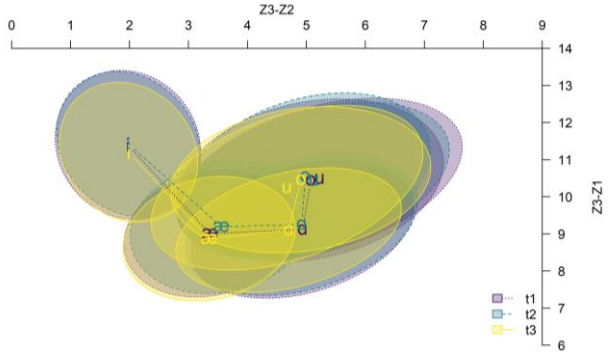


Fig. 1: Five Norwegian vowels plotted by bark normalized vowel height (Z3-Z1) and frontness (Z3-Z2) produced at three tests. The mean is marked by vowel label, and the ellipses show 1 SD distance from the mean.

To evaluate the degree of distinctness between the vowels, we calculated the Pillai score (ranging from 0-1) using MANOVA test [20], taking Z3-Z2 and Z3-Z1 as the dependent variables for each participant.

- *Vowel category stability across tests*: we compared the degree of distinctness of each target vowel within speakers between T1-T2 and between T1-T3 (T1 being the baseline), which yielded 250 Pillai scores (25 participants × 5 vowels × 2 test comparisons).
- *Distinctness of vowel contrasts across tests*: we compared with-in speakers five pairs of vowels (i.e., /i-æ/, /æ-ɑ/, /ɑ-o/, /o-u/, and /u-i/) at each test, yielding 375 Pillai scores (25 participants × 5 vowel pairs × 3 tests).

To address the two hypotheses, we built two generalized linear mixed effects models using the *glmmTMB* package [21] in R [22]. Both models took the Pillai Score as the dependent variable. Model 1 included vowel (/i, æ, ɑ, o, u/), test comparison (T1-T2, and T1-T3), and their interaction as the fixed effects. Model 2 involved test (T1, T2, T3), vowel pair (/i-æ/, /æ-ɑ/, /ɑ-o/, /o-u/, and /u-i/), and their interaction as fixed factors. Participant was added as random intercept in both models. We selected the best random slopes in *buildmer* package [23], which involved a by-participant random slope of test comparison for Model 1 and a by-participant random slope of vowel pair for Model 2. We calculated the significance of the fixed factors with *Anova()* function from the *car* package [24] and tested post-hoc comparisons using *emmeans* package [25] with significance adjusted by false discovery rate for multiple comparisons. All analyses have been preregistered on the OSF (https://osf.io/d3w2k).

3. RESULTS

3.1 The stability of vowel categories over time

Model 1 revealed a significant main effect of vowel, $\chi^2 = 15.93$, $p = .003$. Post-hoc comparisons showed that /u/ had a significantly higher Pillai score than /i/, $t = 3.27$, $p = .011$, /æ/, $t = 2.92$, $p = .032$, /ɑ/, $t = 3.11$, $p = .018$, and /o/, $t = 3.26$, $p = .011$. However, all the vowels showed low Pillai scores (Fig. 2), indicating a considerable within-category overlap over time. More importantly, there was no significant main effect of test comparison, $\chi^2 = 0.21$, $p = .651$, or interaction of Test comparison × Vowel, $\chi^2 = 2.28$, $p = .684$. Hence, the vowels in general were stable over time.

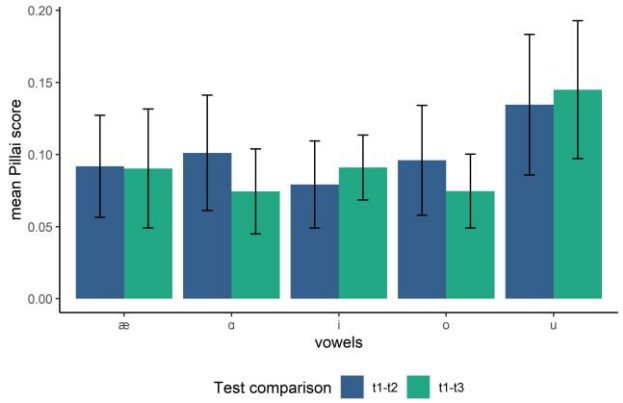


Fig. 2: Mean Pillai score (0 = merge, 1 = separation) of each target vowel compared between T1 and T2, and between T1 and T3. Error bars mark ± 2 *SE*.

3.2 The stability of distinctness in vowel contrasts over time

Model 2 revealed a significant main effect of vowel pair, $\chi^2 = 567.73$, $p < .001$, and a two-way interaction of Vowel pair × Test, $\chi^2 = 21.71$, $p = .005$, suggesting that the degree of distinctness between each of the vowel pairs varied across tests. The Pillai scores of /æ-ɑ/ decreased from T1 to T2, $t = 3.33$, $p = .003$, and from T1 to T3, $t = 3.66$, $p = .001$; the Pillai scores of /u-i/ decreased from T1 to T3, $t = 2.58$, $p = .028$. That is, their distinctness was not stable over time. The other three contrasts showed no effect of test. Hence their distinctness was stable over time. See Fig. 3 for the descriptive data.

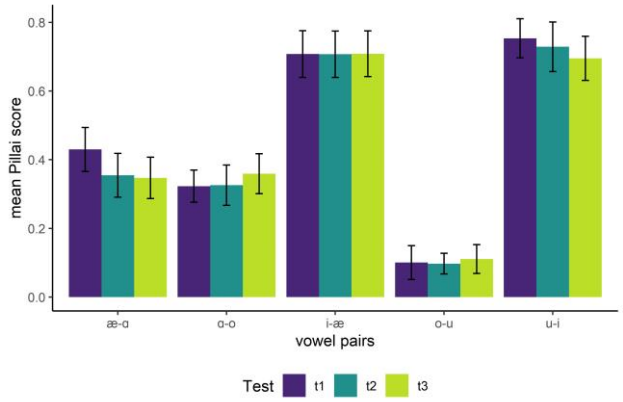


Fig. 3: Mean Pillai score (0 = merge, 1 = separation) compared between each vowel pair across tests. Error bars mark ± 2 *SE*.

4. DISCUSSION AND CONCLUSION

This study investigated the stability in Norwegian vowel production using an innovative speech-eliciting method—Characteristic Speech Production, which instructed the participants to answer meaningful questions in the target language (Norwegian) to elicit the natural production of the target sounds (/i, æ, ɑ, o, u/) without focusing their attention to speech. The test was administered three times over three weeks. The production stability was measured by Pillai score, derived from the formant analyses of the target vowels. We hypothesized that (H1) Norwegian speakers would produce each target vowel in a consistent way over time and (H2) Norwegian speakers would show stable distinctness between the vowel contrasts over time.

Regarding H1, the vowels were stable over time, as revealed by low Pillai scores, suggesting that Norwegian speakers produced the vowels consistently and with overlap across the three tests. Although the five target vowels were produced stably across the three tests, it appeared that, overall, /u/ was more dispersed (with higher Pillai scores) than the other vowels. Because three out of the four words containing /u/ had either a schwa following the target vowel /u/ (i.e., *noe* [nuə], *noen* [nuən]) or a palatal approximant preceding the target vowel (i.e., *jo* [ju]), the phonetic environment may have centralized /u/ and the centralization became more evident through multiple repetitions. In sum, our data supported H1 in terms of stability in individual vowel production.

With respect to H2, the stability of vowel category distinctness was subject to the concrete phonemic contrasts with the two high vowels /u-i/ and the two low vowels /æ-ɑ/ showing closer rapprochement at the follow-up tests compared to the first test.

We interpret our results as follows. First, the repetition across T1 to T3 may have reduced the articulatory efforts for the vowels contrasting mainly in backness, which was reflected by the centralization between the front and back vowel pairs /u-i/ and /æ-ɑ/. Since the two pairs are corner vowels and show clear distinctness, speakers might be less careful in maintaining the distinctness over multiple repetitions. Second, the prosody-segment interaction may have played a role. In spontaneous speech elicitation, the target vowels are inevitably embedded in functional words, which are often unstressed and less prominent on the intonational level. The prosodic structure may have led to vowel reduction, which was a resource of centralization observed in our data. Future research may try to minimize the influence of prosody on speech sound realization.

In conclusion, the current study showed that individual speech sound production was stable in phoneme categories over time. However, some of the category distinctness were reduced. The variability in sound production may be affected by phonetic environment, repetition, and prosodic structure. Future studies in L1 and L2 speech production may consider these factors in the research design.

5. ACKNOWLEDGEMENT

This study was supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265]. CDM was supported by the Spanish Ministry of Economy and Competitiveness [PID2020-113926GB-I00], the Basque Government [PIBA18-29], and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [819093]. We thank Silje Robberstad, Arun Singh, Johannes Sletten, and Sander Vassanyi for their assistance in data collection and analysis.

6. REFERENCES

[1] X. Xie and T. F. Jaeger, "Comparing non-native and native speech: Are L2 productions more variable?," *J. Acoust. Soc. Am.*, vol. 147, no. 5, pp. 3322–3347, May 2020, doi: 10.1121/10.0001141.

[2] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *J. Phon.*, vol. 34, no. 4, pp. 409–438, Oct. 2006, doi: 10.1016/j.wocn.2005.08.002.

[3] S. Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1286–1298, Sep. 1990, doi: 10.1121/1.399705.

[4] C. DiCanio, H. Nam, J. D. Amith, R. C. García, and D. H. Whalen, "Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec," *J. Phon.*, vol. 48, pp. 45–59, Jan. 2015, doi: 10.1016/j.wocn.2014.10.003.

[5] S. L. M. Heald and H. C. Nusbaum, "Variability in vowel production within and between days," *PLOS ONE*, vol. 10, no. 9, p. e0136791, Sep. 2015, doi: 10.1371/journal.pone.0136791.

[6] J. Pierce, K. Tanner, R. Merrill, L. Shnowske, and N. Roy, "Acoustic variability in the healthy female voice within and across days: How much and why?," *J. Speech Lang. Hear. Res.*, vol. 64, pp. 1–17, Jul. 2021, doi: 10.1044/2021_JSLHR-21-00018.

[7] C. C. Heffner and E. B. Myers, "Individual Differences in Phonetic Plasticity Across Native and Nonnative Contexts," *J. Speech Lang. Hear. Res.*, vol. 64, no. 10, pp. 3720–3733, Oct. 2021, doi: 10.1044/2021_JSLHR-21-00004.

[8] J. S. Perkell *et al.*, "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2338–2344, Oct. 2004, doi: 10.1121/1.1787524.

[9] A. R. Bradlow, M. Blasingame, and K. Lee, "Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech," *Lab. Phonol. J. Assoc. Lab. Phonol.*, vol. 9, no. 1, p. 17, Oct. 2018, doi: 10.5334/labphon.137.

[10] N. Kartushina and U. H. Frauenfelder, "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," *Front. Psychol.*, vol. 5, 2014, doi: 10.3389/fpsyg.2014.01246.

[11] C. Meunier and R. Espesser, "Vowel reduction in conversational speech in French: The role of lexical factors," *J. Phon.*, vol. 39, no. 3, pp. 271–278, Jul. 2011, doi: 10.1016/j.wocn.2010.11.008.

[12] J. E. Flege, "New methods for second language (L2) speech research," in *Second language speech learning, Theoretical and empirical progress*, R. Wayland, Ed., Cambridge: Cambridge University Press, 2021, pp. 119–156.

[13] J. E. Flege, "A distributional learning account of L2 speech learning," presented at the 10th International Symposium on the Acquisition of Second Language Speech, Barcelona, Spain, 2022.

[14] G. Kristoffersen, *The Phonology of Norwegian*. Oxford: Oxford University Press, 2000.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2020. [Online]. Available: http://www.praat.org

[17] M. Zhang, "Dynamic seeding scripts bundle," 2022. https://github.com/ZenMule/dynamic_seeding_scripts

[18] W.-R. Chen, D. H. Whalen, and C. H. Shadle, "F0-induced formant measurement errors result in biased variabilities," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. EL360–EL366, May 2019, doi: 10.1121/1.5103195.

[19] A. Rosslund, J. Mayor, G. Óturai, and N. Kartushina, "Parents' hyper-pitch and low vowel category variability in infant-directed speech are associated with 18-month-old toddlers' expressive vocabulary," Dec. 2022, doi: 10.34842/2022.0547.

[20] F. Santiago, "L'accentuation contribue-t-elle à l'acquisition du contraste arrondi vs non-arrondi des voyelles orales en français langue étrangère ?," *Etudes Linguist. Appliquée*, pp. 74–90, 2021.

[21] M. E. Brooks *et al.*, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *R J.*, vol. 9, no. 2, pp. 378–400, 2017.

[22] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.r-project.org/

[23] C. C. Voeten, "buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression." 2021. https://cran.r-project.org/package=buildmer

[24] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*. Thousand Oaks {CA}: Sage, 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[25] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Emmeans: Estimated marginal means, Aka Least-Squares means." R package 1.5.1, 2020. https://cran.r-project.org/package=emmeans

INDIVIDUAL DIFFERENCES IN VOWEL COMPACTNESS PERSIST UNDER INTOXICATION ACROSS FIRST AND SECOND LANGUAGES

Charles B. Chang, Kevin Tang, & Andrew Nevins*

Boston Univ.; Heinrich-Heine-Universität Düsseldorf & Univ. of Florida; Univ. College London
cbchang@post.harvard.edu, Kevin.Tang@hhu.de, a.nevins@ucl.ac.uk

ABSTRACT

Alcohol intoxication facilitates inhibition of one's first language (L1) ego, which may lead to reduced individual differences among second language (L2) speakers under intoxication. This study examined whether, compared to speaking while sober, speaking while intoxicated would reduce individual differences in the acoustic compactness of vowel categories in sequential bilinguals exemplifying diverse L1–L2 pairs (German–English, Korean–English). Vowel compactness in $F_1 \times F_2$ space varied by language (German, Korean, English) and by vowel, and was generally lower in intoxicated compared to sober speech, both across languages and throughout a bilingual's language repertoire. Crucially, however, there was still a wide range in compactness under intoxication; furthermore, individuals with more compact vowels while sober also produced more compact vowels while intoxicated, in both L1 and L2. Taken together, these findings show patterned variability of vowel compactness, suggesting that articulatory precision is an individual-difference dimension that persists across speaking conditions and throughout the repertoire.

Keywords: compactness, individual differences, vowels, bilingualism, alcohol.

1. INTRODUCTION

How does drinking affect one's speech? A growing literature on alcohol intoxication and spoken language production suggests that there may not be a simple answer to this question. On the one hand, several studies have found, consistent with the general decline of motor control under intoxication, that intoxication degrades speech, in terms of more errors, less precise gestural coordination, more consonant lenition, slower speech rate, greater pitch variability, and/or lower intelligibility [1, 2, 3, 4, 5, 6, 7, 8]. On the other hand, studies that have examined intoxication effects on speech in a bilingual's second language (L2), in isolation or in comparison to speech in their first language (L1), have often observed that intoxication affects L2/nonnative speech differently from L1 speech [7, 9]. For example, intoxication improved L1 German speakers' pronunciation in Dutch [10], which can be explained in terms of intoxication making a speaker's "language ego" less L1-biased, facilitating nonnative speech [9].

The possibility of facilitative intoxication raises interesting questions for L2 speech research. For one, if intoxication improves L2 production, might it "level the playing field" among L2 learners, who often show marked individual differences even from the same L1 background [11, 12, 13, 14, 15]? L2 speech research has been increasingly concerned with accounting for individual differences, linking them to factors such as cue-weighting preferences, memory variation, and perceptual and neural differences [16, 17, 18, 19, 20]. In this study, we focused on individual differences in *compactness*, the acoustic consistency of a phonetic category's production [21], thought to reflect the "consistency of [speakers'] phonological-motor mapping" [22, p. 826].

Part of a larger project examining intoxication effects on bilingual speech, the current study asked how individual differences among bilinguals in vowel compactness would be affected by intoxication, across L1 and L2. Based on findings of intoxication degrading L1 but facilitating L2 speech, we hypothesized that intoxication would result in less compact L1 vowels, yet—by enhancing access to L2 phonological-motor mappings—more compact L2 vowels. Given the persistence of variation in compactness across L1 and L2 [21, 22], intoxication-induced changes in compactness were not predicted to reduce individual differences in L1; that is, habitually "less compact" speakers were predicted to become even less compact under intoxication along with "more compact" speakers. In contrast, if intoxication does indeed have an L2-specific facilitation effect (and there is an upper limit on compactness), it could allow less-compact speakers to "catch up"

*We gratefully acknowledge the data contributions and research assistance of Kai Xin Bao, Marissa Carl, Erik Duchnowski, Jenny Geng, Sam Green, Michael Hindley, Young Shin Kim, Peyton Krinsky, Ji Hye Kwon, Wen Jia Liu, Eliana Mugar, Yin Wang, Yifan Wu, and Steven Zhang, as well as the helpful feedback of two anonymous reviewers.

to more-compact speakers in terms of compactness, leading to reduced individual differences in L2. To explore these predictions, we compared the sober and intoxicated speech of bilinguals exemplifying diverse L1–L2 pairs.

2. METHODS

2.1. Participants

Participants comprised two groups of sequential bilinguals: Germans (L1 German, L2 English; $N = 8$; 4f, 4m; $M_{age} = 27.1$ yr) and Koreans (L1 Korean, L2 English; $N = 8$; 8f, 0m; $M_{age} = 27.1$ yr). The groups were born and raised in Germany or South Korea, respectively, and were tested in the UK, where most participants had been living for 1–2 years for study abroad. They reported no history of hearing or speech problems, and their average scores on the IELTS [23] (7.0 band) and LexTALE [24] (60s), which did not differ significantly between groups, suggested that their English proficiency was generally "upper intermediate" (B2) or higher.

2.2. Materials

Speech materials were compiled for each language on the basis of published dialogues (e.g., from a play), which were edited for emotional and gender neutrality, turn length, and archaisms. The German dialogue contained a total of 55 turns; the Korean dialogue 123 turns; and the English dialogue 95 turns. The full materials with source references are available open-access at `osf.io/y2r87/`.

2.3. Procedure

The main task was a dialogue-reading task, which participants completed in a sound-insulated room four times (2 languages $\times$ 2 speaking conditions) across two test sessions spaced no more than 14 days apart. Each dialogue was uttered in two speaking conditions (sober, intoxicated), completed on separate days in the order sober–intoxicated in the Korean group and in counterbalanced order in the German group. Participants were told not to eat, drink, or use mouthwash starting 2 h before each session and not to smoke starting 0.5 h before.

Before beginning the dialogue-reading task, participants were instructed to read the dialogues naturally. Seated in front of a microphone facing the experimenter, the participant went through each dialogue with the experimenter, each person reading one character's lines. The session was audio-recorded at 44.1 kHz with 16-bit resolution in stereo; the recordings were then converted to mono.

Participants' BAC was tested in both conditions using a breathalyzer (AlcoMate Premium AL-7000). First, BAC was measured to ensure that participants had no alcohol in their system. In the intoxicated condition, the target BAC was 0.12%. To reach this BAC, participants consumed, at their own pace, an amount of alcohol (vodka or rum, mixed with juice) determined on the basis of their self-reported weight and BAC charts [25]. BAC was tested 15 min after about 75% of the mixture was consumed and then every 3–5 min until it went over 0.12% and dropped back down to 0.12%; if needed, a small top-up was given from the remaining alcohol. Once BAC had reached 0.12%, participants were taken into the sound-insulated room for the dialogue-reading task.

2.4. Analysis

Forced-aligned TextGrids (produced using the Montreal Forced Aligner [26]) were hand-corrected and then submitted, along with the audio recordings, to acoustic analysis in Praat [27]. For each of the vowel intervals marked in the TextGrids, the analysis extracted both the first and second formants (F_1, F_2) over the middle 50% of the vowel's duration using the default settings of Praat's linear predictive coding method; the maximum formant parameter was set to 5,000 Hz for males and 5,500 Hz for females.

Because German and English, in contrast to Korean, are described as having unstressed vowel quality reduction, the final dataset for statistical analysis included only primary-stressed monophthongs in German and English, as well as the monophthongs of Korean. The F_1 and F_2 measures on this set of vowels were checked for errors separately by participant and by language, using the 3-*SD* criterion for outliers; tokens with one or more errors (comprising 1.2% of the dataset) were removed. Formant measures on the remaining tokens ($N = 64,382$ total; 9,081 in German, 27,450 in Korean, 27,851 in English) were then Nearey-normalized [28] using the `vowels` package [29] in R [30]. The full dataset is available open-access at `osf.io/s9vkw/`.

The normalized formant measures were analyzed both qualitatively and quantitatively. First, we inspected vowel plots to observe variation in vowel compactness among participants and languages and within the repertoire (L1, L2). Then we calculated a metric of vowel area (i.e., the inverse of compactness) in terms of the area of the 95% confidence ellipse around tokens. Finally, we built a linear mixed-effects model on the vowel area data, using the `lmerTest` package [31]. The model included fixed effects for Group (treatment-coded; ref = German),

Condition (treatment-coded; ref = sober), RepLang (treatment-coded; ref = L1), Vowel (sum-coded; all monophthongs including German /i ɪ e ɛ a ɑ aː u ʊ o y ʏ ø œ/, Korean /i e ɛ u o ʌ ɨ/, and English /i ɪ e ɛ æ u ʊ o ɑ ɔ ʌ ɚ/), all interactions among Group, Condition, and RepLang, a Condition × Vowel interaction, as well as random intercepts by Participant.[1]

3. RESULTS

Vowel area varied significantly by language, with tokens of a vowel category generally more diffusely (i.e., less compactly) distributed in the Korean monophthongs than the German monophthongs in sober speech [$\beta = 0.226, t = 4.575, p < .001$]. On the other hand, vowel area was smaller in the German group's L2 English monophthongs as compared to their L1 German [$\beta = -0.122, t = -4.019, p < .001$], and this L2 effect was even stronger in the Korean group [$\beta = -0.138, t = -3.161, p = .002$].[2]

Examination of individual differences in vowel area revealed substantial variation in both groups. Table 1 summarizes the ranges in average vowel area by group, language, and condition. Although the minimum and maximum values within each group are different (consistent with the language effect above), there is a wide range in every case.

Group	Language, Condition	Min	Max
German	German, sober	0.303	0.500
German	German, intoxicated	0.316	0.637
German	English, sober	0.204	0.487
German	English, intoxicated	0.308	0.497
Korean	Korean, sober	0.556	0.885
Korean	Korean, intoxicated	0.565	0.953
Korean	English, sober	0.332	0.483
Korean	English, intoxicated	0.308	0.515

Table 1: Range in average vowel area (min, max values) in Nearey-normalized $F_1 \times F_2$ space.

Examples of individual differences are depicted in Fig. 1–4, which show the vowel spaces for some of the most-compact (G005, K004) and least-compact (G006, K006) participants by language and condition on the same scale, with 95% confidence ellipses around tokens of the same vowel (plots for all participants are at `osf.io/s9vkw/`). Note the marked difference in compactness between G005 and G006 and between K004 and K006, as well as the broad similarity among the four panels in each figure.

As seen in Fig. 1–4, the area for a given vowel tended to be larger in intoxicated than in sober speech for all languages. Thus, instead of showing an L2-specific effect (in the interaction coefficients),

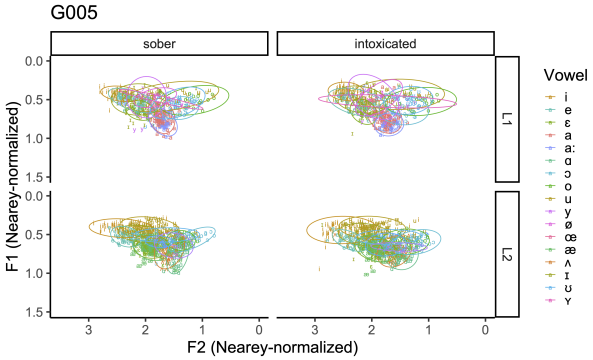


Figure 1: Vowel spaces by language (L1 German, L2 English) and condition in participant G005.

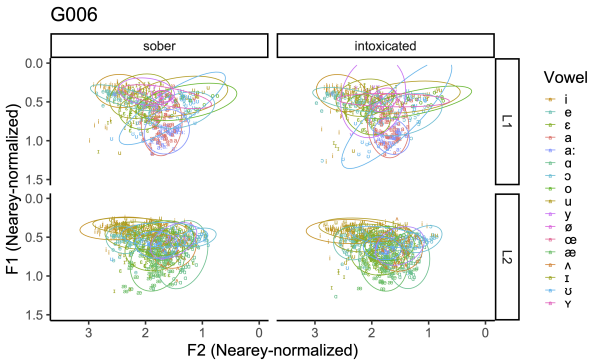


Figure 2: Vowel spaces by language (L1 German, L2 English) and condition in participant G006.

the model showed that the marginal intoxication effect in German [$\beta = 0.049, t = 1.743, p = .082$] did not differ in Korean [$\beta = 0.030, t = 0.625, p = .532$] or L2 English [$|\beta|$s < 0.099, $|t|$s < 1.586, ps $> .1$]. That is, we found no evidence of L2-specific facilitation of vowel production. Rather, the intoxication effect was general, resulting in expanded vowel areas at the group mean level in L1 (Germans: $M_{int} = 0.460$, $M_{sob} = 0.391$; Koreans: $M_{int} = 0.743$, $M_{sob} = 0.677$) and L2 (Germans: $M_{int} = 0.395$, $M_{sob} = 0.316$; Koreans: $M_{int} = 0.414$, $M_{sob} = 0.404$).

In order to further examine the relationship between vowel areas in intoxicated and sober speech at the individual level, we correlated participants' by-language average vowel areas for the intoxicated condition against those for the sober condition. The two sets of values were highly correlated overall [$r(30) = 0.926, t = 13.448, p < .001$], in German [$r(6) = 0.747, t = 2.752, p = .033$], in Korean [$r(6) = 0.902, t = 5.120, p = .002$], and in the German group's English [$r(6) = 0.845, t = 3.866, p = .008$], although not in the Korean group's English [$r(6) = 0.480, t = 1.342, p = .228$]. Thus, in general, bilinguals with more compact vowels while

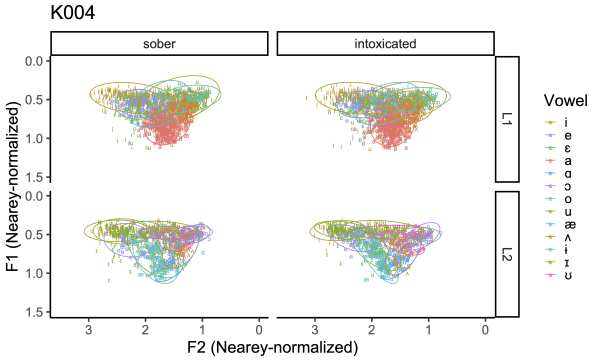


Figure 3: Vowel spaces by language (L1 Korean, L2 English) and condition in participant K004.

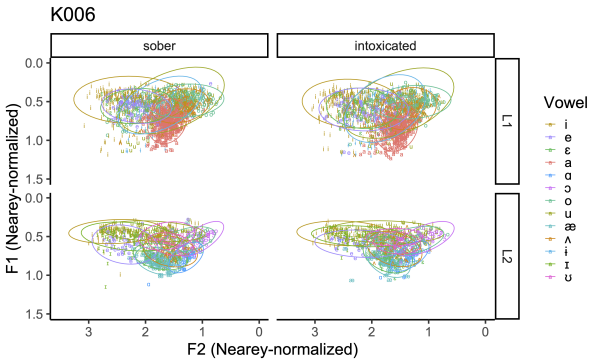


Figure 4: Vowel spaces by language (L1 Korean, L2 English) and condition in participant K006.

sober also produced more compact vowels while intoxicated, whether speaking in L1 or L2.

Finally, there were significant differences among vowels in terms of area, but not in terms of intoxication effects. Compared to the average vowel area over all vowels, the vowels /i ɪ u ʊ/ (in German and English), /y ɑː/ (in German), /i/ (in Korean), and /ɑ/ (in English) were significantly more diffusely distributed, whereas /e ɛ o/ (in German, Korean, and English), /a/ (in German and Korean), /ɔ/ (in German and English), /ʌ/ (in Korean and English), /ø œ/ (in German), and /ɚ/ (in English) were significantly less diffusely distributed (i.e., more compact). Crucially, none of the Condition × Vowel coefficients were significant, meaning that different vowels did not vary significantly in terms of how their vowel area was affected by intoxication.

4. GENERAL DISCUSSION

Taken together, the results of this study do not support the hypothesis of an L2-specific facilitative effect of intoxication on speech production. Contrary to expectation, intoxication decreased vowel compactness in L2 (English), just as it did in L1 for both German and Korean. Intoxication-induced changes in compactness also left a wide range in compactness at the individual level (Table 1). Individual variation in compactness, moreover, tended to be correlated between sober and intoxicated speech, in both L1 and L2. These findings provide evidence that individual differences in compactness persist across speaking conditions and throughout a bilingual's repertoire, consistent with the view that the articulatory precision reflected in compactness is a stable dimension of individual differences.

The contrast between our results and those of [10] invites the question of why [10] observed facilitative effects of intoxication on L2 speech but we did not. There are two possible, and mutually compatible, explanations. First, our L2 speakers may have been more proficient than those in [10]. Second, our study analyzed a specific, objective aspect of segmental production, whereas [10] focused on a subjective measure of overall pronunciation. What we take away from these differences is that the effects of intoxication on speech may be multifaceted and variable over development. For example, intoxication may worsen acoustic *consistency* at the segmental level, but it may not necessarily affect *proximity* to native norms, and it may well improve suprasegmentals that contribute to perceived pronunciation quality, none of which were tested in this study.

While this study found persistent individual differences in the acoustic consistency of vowels, the findings are limited in a few ways that leave open several directions for future research. For one, as mentioned above, we examined only compactness, so we cannot say for sure whether more-compact speakers were also more "on target" vis-a-vis norms than less-compact speakers. Thus, further work investigating central tendencies of L2 phonetic categories in conjunction with their distributions would provide a fuller picture of individual differences in L2 speech. In addition, for reasons of space, we were only able to discuss the range of individual differences in compactness, glossing over potential variation in changes in compactness by vowel. In future research, it would therefore be insightful to examine, for example, whether speakers showing average compactness pattern differently from the most-compact and/or least-compact speakers in terms of intoxication effects on specific vowels. Finally, our results showed suggestive effects of language (German vs. Korean) and language status (L1 vs. L2) on compactness, which could be usefully elaborated in future studies crossing languages with different-size vowel inventories as L1 and L2.

5. REFERENCES

[1] Trojan, F., Kryspin-Exner, K. 1968. The decay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica et Logopaedica* 20, 217–238.

[2] Beam, S. L., Gant, R. W., Mecham, M. J. 1978. Communication deviations in alcoholics; A pilot study. *Journal of Studies on Alcohol and Drugs* 39, 548–551.

[3] Chin, S. B., Pisoni, D. B. 1997. *Alcohol and Speech*. San Diego: Academic Press.

[4] Pisoni, D. B., Martin, C. S. 1989. Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research* 13(4), 577–587.

[5] Johnson, K., Pisoni, D. B., Bernacki, R. H. 1990. Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica* 47, 215–237.

[6] Hollien, H., DeJong, G., Martin, C. A., et al. 2001. Effects of ethanol intoxication on speech suprasegmentals. *JASA* 110, 3198–3206.

[7] Wieling, M., Blankevoort, C., Hukker, V., et al. 2019. The influence of alcohol on L1 vs. L2 pronunciation. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proceedings of the 19th ICPhS*. Canberra: ASSTA Inc., 3622–3626.

[8] Tang, K., Chang, C. B., Green, S., et al. 2022. Intoxication and pitch control in tonal and non-tonal language speakers. *JASA Express Lett.* 2, 065202.

[9] Guiora, A. Z., Beit-Hallahmi, B., Brannon, R. C. L., et al. 1972. The effects of experimentally induced changes in ego states on pronunciation ability in a second language: An exploratory study. *Comprehensive Psychiatry* 13, 421–428.

[10] Renner, F., Kersbergen, I., Field, M., et al. 2018. Dutch courage? Effects of acute alcohol consumption on self-ratings and observer ratings of foreign language skills. *Journal of Psychopharmacology* 32, 116–122.

[11] Dörnyei, Z., Skehan, P. 2003. Individual differences in second language learning. In: Doughty, C., Long, M. (eds), *The Handbook of Second Language Acquisition*. Malden: Blackwell, 589–630.

[12] Chandrasekaran, B., Sampath, P. D., Wong, P. C. M. 2010. Individual variability in cue-weighting and lexical tone learning. *JASA* 128, 456–465.

[13] Perrachione, T. K., Lee, J., Ha, L. Y. Y., et al. 2011. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *JASA* 130, 461–472.

[14] Bowles, A. R., Chang, C. B., Karuzis, V. P. 2016. Pitch ability as an aptitude for tone learning. *Language Learning* 66, 774–808.

[15] Chang, C. B., Kwon, S. 2020. The contributions of crosslinguistic influence and individual differences to nonnative speech perception. *Languages* 5, 49.

[16] Baker, W., Trofimovich, P. 2006. Perceptual paths to accurate production of L2 vowels: The role of individual differences. *IRAL* 44, 231–250.

[17] Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., et al. 2010. Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: Prediction of individual differences. *Cerebral Cortex* 20, 1–12.

[18] Idemaru, K., Holt, L. L., Seltman, H. 2012. Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *JASA* 132, 3950–3964.

[19] Darcy, I., Park, H., Yang, C.-L. 2015. Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences* 40, 63–72.

[20] Schertz, J., Cho, T., Lotto, A. J., et al. 2016. Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, and Psychophysics* 78, 355–367.

[21] Kartushina, N., Frauenfelder, U. H. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology* 5, 1246.

[22] Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., et al. 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *JASA* 138, 817–832.

[23] British Council, IDP: IELTS Australia, Cambridge English. 2023. IELTS in CEFR scale. https://www.ielts.org/about-ielts/ielts-in-cefr-scale.

[24] Lemhöfer, K., Broersma, M. 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44, 325–343.

[25] Miller, W. R., Muñoz, R. F. 1982. *How to Control Your Drinking*. Englewood Cliffs: Prentice Hall.

[26] McAuliffe, M., Socolof, M., et al. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In: *INTERSPEECH-2017*, 498–502.

[27] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. Version 6.2.23, praat.org.

[28] Nearey, T. M. 1977. *Phonetic Feature Systems for Vowels*. PhD thesis, University of Alberta.

[29] Kendall, T., Thomas, E. R. 2018. vowels: Vowel manipulation, normalization, and plotting. R package version 1.2-2. Available from https://cran.r-project.org/web/packages/vowels/.

[30] R Development Core Team, 2022. R: A language and environment for statistical computing. Version 4.2.1, www.r-project.org.

[31] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26.

[1] Random slopes for Condition by Participant were explored but not included because they caused the model to be singular. Results remained the same regardless.

[2] The full model output showing all fixed-effect coefficients is available at `osf.io/s9vkw/`.

TOWARDS SPEAKER-INDEPENDENT ULTRASOUND TONGUE IMAGING-BASED ARTICULATORY-TO-ACOUSTIC MAPPING

XXXXXX

XXXXXXXX

ABSTRACT

In this paper, an articulatory-to-acoustic mapping task is studied, which aims to predict the mel-spectrogram of the audio signals using midsagittal ultrasound tongue images of the vocal tract. Despite of sustainable efforts been made, most previous attempts have been constrained to the speaker-dependent scenario, and the performance is greatly decreased for unseen speakers. Here, a novel approach is proposed for the speaker-independent scenario, using domain-adaptation and adversarial learning. To validate the performance of proposed method, objective evaluation is conducted to demonstrate the effectiveness of the proposed method. The quantitative results indicate our method can achieve superior performance.

Keywords: Ultrasound tongue imaging, articulatory-to-acoustic mapping.

1. INTRODUCTION

In the natural speech production process, the speech signal is highly related to articulation. Understanding the association between the articulation and the speech signals not only can be helpful to improve our understanding of speech production but also can stimulate the theoretical development of speech recognition. There is an increasing trend that aims to build the articulatory-to-acoustic mapping [1] as the applications of the mapping seem to be evident, such as the silent speech interface (SSI) [2] which has a long-term goal to generate the speech using the soundless articulatory information.

Previous attempts employed statistical methods [3] for the inversion between the articulatory movements and speech, while deep learning has begun to dominate this field [4]. For example, [5] explored a two-layer deep neural network for the inversion task, and satisfying performance can be obtained. Despite the sustainable efforts that have been made [6], the mapping performance can be greatly varied for different speakers and most of the previous methods are constrained for the speaker-dependent scenario. In this paper, we present a novel approach towards speaker-independent mapping, which is inspired by the domain adaptation method. Specifically, we explore decoupling the speaker spectral generation task and the speaker recognition task. Leveraging a novel designed loss function, we can improve the performance under the speaker-independent scenarios, through adversarial learning. Our code is available at: https://github.com/xianyi11/Articulatory-to-Acoustic-with-Domain-Adaptation.

2. RELATED WORK

The early studies on articulatory-to-acoustic mapping usually employ low-order spectral representations of speech as the target, such as [7] using 12 coefficients. Later, [8] used other spectral features which has 25-dimensional to handle the articulatory-to-acoustic mapping task. Recently, [9] used a large-scale pre-trained model WaveGlow [10] to predict more detailed representations (80-dimensional mel-spectrogram) from ultrasound tongue images (UTI), resulting in more accurate models. However, to the best of our knowledge, many of previous attempts try to solve the task in a speaker-dependent manner, while the performance is greatly decreased for the speaker independent way (i.e., unseen speaker). There are just a few studies on the speaker-independent articulatory-to-acoustic mapping. Motivation by the shortcomings, we mainly focus on the speaker-independent scenarios.

3. METHODOLOGY

3.1. Overall design

In Figure 1, we present the overall flowchart of the proposed method. Our approach consists of two main parts: the first part is a convolutional neural network for the feature extraction from the UTI,

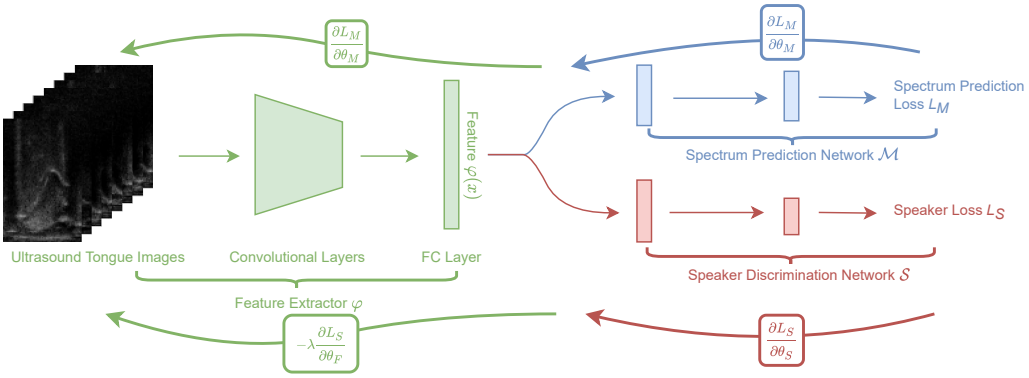


Figure 1: Overall flowchart of the proposed method for the articulatory-to-acoustic mapping using UTI.

and the latter part is designed for the estimation task. To improve the generalization ability across different speakers, we decouple the latter part into two branches, one branch for the generation task and the other is for the speaker discrimination. For the joint training, a novel designed loss is explored to train the network through adversarial learning. Compared to the method without speaker discrimination and adversarial learning, our method can improve the prediction performance under the speaker-independent scenarios. The components will be explained in more detail subsequently.

3.2. Mel-Spectrogram Prediction

The UTI contains the deformation information of the vocal tract during natural speech production. Intuitively, we could predict the sound from the ultrasound image by the high-dimensional non-linear transformation. Thereby, an ultrasound video could be mapped to the continuous speech spectrogram in Mel-scale through the framewise prediction. Specifically, we build the mapping using a multi-stage deep neural network. The predicted Mel-spectrogram $\hat{y}_i$ could be acquired by:

$$(1) \quad \hat{y}_i = u_i^T \varphi(x).$$

where φ is the backbone neural network which is parameterized by θ_φ, and $x \in R^{H\times W}$ is the ultrasound tongue image. The weights of spectrogram prediction network $\mathscr{M}$ can be noted as u, where u_i is the i_{th} weight vector. The loss function applied for this generation task is the mean-squared error (MSE):

$$(2) \quad \mathscr{L}_M = \frac{1}{M}\sum_{i=0}^{M}(y_i - \hat{y}_i)^2,$$

where M is the number of Mel-bins in the Mel-spectrogram, y is the ground-truth of the spectrogram.

3.3. Speaker Discrimination

The speaker information and the speech content are naturally coupled in the ultrasound image and spectrogram. In order to make our system irrelevant or weakly dependent to the speaker information, it is necessary to decouple the speaker information and the speech content. We set up a shallow speaker discrimination network $\mathscr{S}$ which is parameterized by $\theta_\mathscr{S}$ to recognize the speaker from the ultrasound image:

$$(3) \quad \hat{s}_i = v_i^T \varphi(x).$$

where the weights of network $\mathscr{S}$ are denoted as v, and v_i is the i-th weight vector. Note that, during the update of network $\mathscr{S}$, the weights of network $\mathscr{M}$ are fixed, so as to avoid the impact of the cross-modal mapping process. The loss function for learning speaker information is the Cross-Entropy loss:

$$(4) \quad \mathscr{L}_S = -\log\frac{e^{s_i}}{\sum_j^K e^{s_i}},$$

where i is the speaker index, and K is the number of distinguishable speakers which will be explained in detail in Sec.4.

3.4. Adversarial Training

After the network $\mathscr{S}$ converges, the feature $\varphi(x)$ output from the backbone network φ is speaker-dependent, because after the shallow non-linear transformation, the speaker could be recognized. We hope the feature $\varphi(x)$ has better generalization ability across different speakers, rather than close

to the target speaker. We implement this in an adversarial way by decreasing the $-\mathscr{L}_S$ in the speaker discrimination learning process. The final loss for spectral feature prediction is calculated by:

$$\mathscr{L} = \mathscr{L}_M - \lambda \mathscr{L}_S, \quad (5)$$

where λ is the empirical weight which will be investigated in Sec.4.3. The parameters $\theta_{\mathscr{S}}$ of the speaker discrimination network is frozen in the adversarial training.

4. EXPERIMENTAL RESULTS

4.1. Dataset and Implementation Details

The Ultrasuite [11] dataset is used throughout the experiments. These children suffer from speech sound disorders (SSD) to varying degrees. UXTD is the typically developing subset of the Ultrax dataset which contains 58 individuals (31 females and 27 males). The ultrasound framerate is 121.5 FPS. We down-sample the frames by 10 (for every second, we sample 12 frames due to the constraint of the computation resources). For the backbone neural network, we employ ResNet50, which is a well known CNN architectures for the feature extraction. Behind the last convolutional layer, we explore the Batch Norm Fully Connected (BN-FC) [12] structure to get the final 1024-dimension embedding feature. The spectrogram prediction network $\mathscr{M}$ and speaker discrimination network $\mathscr{S}$ are both BN-FC structures. The output dimensions are the number of Mel-bins (80) and the number of speakers. The networks are trained by a stochastic descent optimizer (SGD) with the batch size of 512.

In the experiment, The 66% samples (about 220,468 images) are used for training, and the rest are used for testing. Both the train set and test set contain the data of all 58 individuals (*Base* in table 1). To test the performance of our system on speaker independent data, the partition of the train set and test set is speaker-based. All the data sampled from 38 individuals among all 58 individuals are used for training, and the rest 20 speakersâ data is used for testing (*Sep* in table 1)

4.2. Compared methods

As most of previous attempts focus the mapping using the speaker-dependent manner, we compare the proposed approach with two baselines, *i.e.*, Source-Only model and ST-Adversarial model. Specifically, the Source-Only model is trained without domain-adaptation (no speaker discrimination branch are incorporated into the framework in Fig. 1). This model achieves the articulatory-to-acoustic mapping by the neural network without considering the speaker information. The ST-Adversarial model employs the source and target domain adversarial training inspired by [13]. Specifically, the recognition task is to identify whether the feature belongs to the source or target domain. In our setting, the train set and test set can be regarded as the source and target domain respectively, and the speaker discrimination loss $\mathscr{L}_S$ optimizes a binary classification task ($K = 2$ for Eq.4). In contrast to these, in the proposed ID-Adversarial model, the recognition task is to identify the speaker accurately, and K is the number of speakers during the training for Eq.4.

Table 1: Comparison results in term of mean MSE (Lower is better), SSIM and CW-SSIM (Both metrics ranges between 0 and 1. Higher value denotes better performance, while 1 represents the predicted one is the same as the ground-truth).

		MSE	SSIM	CW-SSIM
Source-Only	Base	1.88	0.73	0.70
	Seq	1.88	0.74	0.73
ST-Adversarial	Base	1.84	0.71	0.68
	Seq	1.78	0.75	0.72
ID-Adversarial	Seq	**1.62**	**0.76**	**0.74**

4.3. Objective Evaluation

Quantitative evaluation is conducted to demonstrate the effectiveness of the proposed method. Three evaluation metrics are used. 1) MSE; 2) Structural Similarity Index (SSIM) between the reconstructed spectrogram and the original spectrogram. 3) Complex Wavelet Structural Similarity Index (CW-SSIM) which is robust to small rotations and translations compared to SSIM. The evaluation results are shown in Table 1. The results show that the adversarial model performs better than the Source-Only model with MSE metric. Note that, the ST-Adversarial model performs even worse than the Source-Only model under the SSIM and CW-SSIM, which indicates that source-target adversarial training may not improve the structural similarity. While the ID-adversarial model is effective in various evaluation metrics.

We analyze the impact of the hyper-parameters λ in Eq.5. The optimal value of λ is about 0.4 to 0.5. We use 0.4 in our following experiments. It is worthwhile noting that λ should be increased cautiously. In order to assign more weights to the speaker discrimination network automatically, the model ignores the main prediction task, which

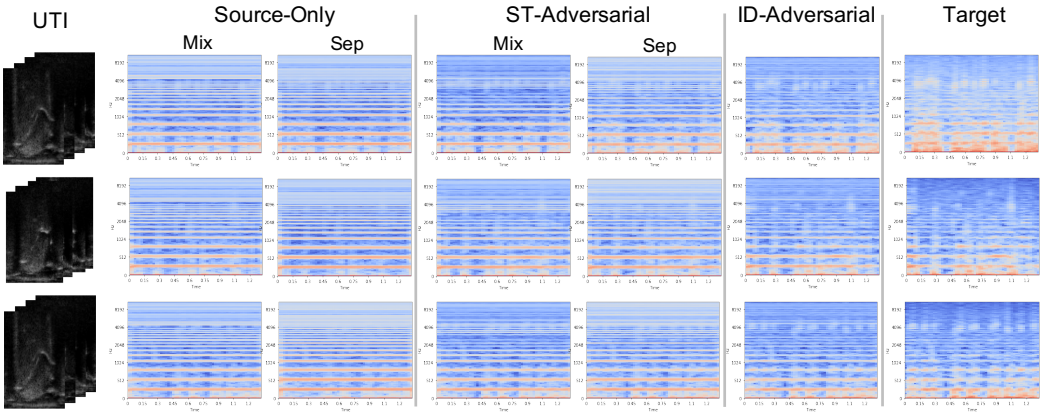


Figure 2: The illustration of generated spectrogram by different models for three different speakers.

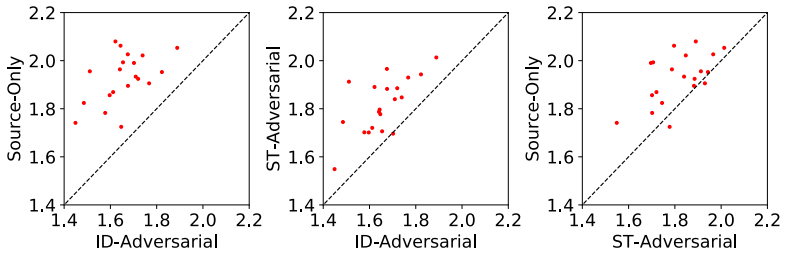


Figure 3: The MSE for 20 speakers with different models. In the middle subplot, all the points are above the diagonal, which indicating that ID-Adversarial model is better than Source-Only model for all the 20 speakers.

makes the performance deteriorate rapidly.

4.4. Qualitative Analyses

We also performed a qualitative analysis to investigate the quality of articulatory-to-acoustic mapping. The generated spectrogram and corresponding target are shown in Fig 2 from three different speakers. We could see that, without domain-adaptation (Source-Only), the spectrogram generated by different speakers are relatively similar. The spectrogram shows multiple horizontal lines along with different frequencies and lacks vocal details. In comparison, the vocal details presented by the ID-Adversarial model are richer.

The formant in the spectrogram is more obvious in the ID-Adversarial model. The formant represents a set of adjacent harmonics which are boosted by resonance in some part of the vocal tract. Obvious formant means that in the subsequent process of generating speech from the spectrogram, the generated speech is clearer. Especially for three different speakers, generated spectrograms are different for the ID-adversarial model.

Fig 3 exhibits the improvement of domain-adaptation approach. It shows the MSE for 20 speakers with different models in the *Sep* setting. In the left subplot, all the points are above the diagonal, which is indicating that the ID-Adversarial model is better than the Source-Only model for all the 20 speakers. We could see that the domain-adaptation approach improves the performance in the speaker-independent scenario.

5. CONCLUSION

This paper proposed a method towards speaker-independent articulatory-to-acoustic mapping, using UTI. Specifically, the domain adaption and adversarial method are applied in our framework, which can decouple the generation and speaker discrimination task. To demonstrate the effectiveness of the proposed method, extensive experiments are conducted. Objective evaluation is conducted to compare the generated spectrograms and ground truth, using three evaluation metrics. The results indicate that our proposed method can achieve superior performance under the speaker-independent scenario. In the future, we plan to conduct subjective listening tests to evaluate the quality of speaker-independent samples.

6. REFERENCES

[1] B. Cao, M. J. Kim, J. R. Wang, J. P. van Santen, T. Mau, and J. Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information." in *INTERSPEECH*, 2018, pp. 3152–3156.

[2] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[3] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded gaussian mixture regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2246–2259, 2015.

[4] H. Li, J. Tao, M. Yang, and B. Liu, "Estimate articulatory mri series from acoustic signal using deep architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4854–4858.

[5] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[6] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded gaussian mixture regressions," in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 2013, pp. 2753–2757.

[7] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "Dnn-based ultrasound-to-speech conversion for a silent speech interface," 2017.

[8] T. G. Csapó, M. S. Al-Radhi, G. Németh, G. Gosztolya, T. Grósz, L. Tóth, and A. Markó, "Ultrasound-based silent speech interface built on a continuous vocoder," *arXiv preprint arXiv:1906.09885*, 2019.

[9] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis," *arXiv preprint arXiv:2008.03152*, 2020.

[10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," 2018.

[11] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," *InterSpeech*, 2018.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

Larynx raising in English word-final ejective stops: A real-time MRI study

Lavinia Price, Marianne Pouplier, Philip Hoole

Institute for Phonetics and Speech Processing (IPS), LMU Munich
{l.price | pouplier | hoole}@phonetik.uni-muenchen.de

ABSTRACT

This study addresses the nature of production mechanisms underlying English ejectives. Using a corpus of real-time MRI data from 27 native British English speakers, it investigates the extent to which English word-final ejective stops are produced by the traditional 'Catfordian' mechanism with active larynx raising. Overall, 8.5% of word-final stops in the dataset are auditorily identified as ejectives. Analysis of larynx position during the ejectives' closure interval reveals a raised position in only 48.6% of cases, suggesting that more than one mechanism is involved in ejective production. Acoustically, these ejectives are significantly distinct from their plain counterparts in the measures applied. Contextual factors known to promote ejective stop production in English are confirmed in our dataset, in that they appear predominantly as allophones of voiceless velar stops in stressed words.

Keywords: ejectives, rt-MRI, airstream mechanism, speech production

1. INTRODUCTION

Ejectives are traditionally defined as the products of a glottalic airstream caused by raising of the larynx during simultaneous constrictions at the glottis and in the oral cavity [3]. Larynx-raising reduces the supraglottal cavity, thereby increasing the intraoral air pressure (IOP) which results in the auditorily distinct quality of these sounds. An ongoing debate surrounds the hypothesis that ejectives may, however, result from a range of different articulatory mechanisms which may or may not involve larynx raising [1, 9, 10, 14, 15, 21, 23, 25]. This in turn means that both a glottalic or pulmonic airstream initiation may be possible for ejectives. How non-glottalic initiation may come about in detail and what the articulatory-acoustic properties of pulmonic ejectives might be remains, however, unknown. The goal of the current paper is to contribute to this debate by investigating larynx position in English ejective production.

1.1. Non-glottalic ejective initiation

Kingston [9] voices doubts regarding the larynx's efficacy in creating the increase in IOP necessary for the high intensity bursts found in some types of ejectives, and suspects the involvement of additional articulators, e.g., tongue root backing. Most recently, Brandt and Simpson [1] compared articulatory mechanisms of Georgian phonological and German epiphenomenal ejectives using dual-channel EGG. Their results indicate no larynx raising at all during the production of German ejectives that occur epiphenomenally when a final stop is overlapped in time by a following glottal stop. In Georgian, they show that larynx activity is variable and that it depends on the position of the ejective within the word or sentence. Combined with IOP measurements they find these results to support Simpson's theory of a type of 'pulmonic' ejective which does not require larynx raising to be initiated [21]. In addition, the traditional tense vs. lax ejective categories, for which differences in vocal fold stiffness and longitudinal tension are proposed to be the articulatory bases [9, 14], are considered too restrictive. The high intra- and inter-language variation found in an increasing number of studies on the acoustic-phonetic properties of ejectives [7, 8, 15, 22, 24, 25], suggests rather that the phonetic realization of ejective sounds may be much more variable than their textbook description allows for [4, 12, 22]. Evidently, there are fundamental gaps in our knowledge regarding the variability of these sounds, and it is against this background that English ejectives represent a particularly interesting case.

1.2. English ejectives

Ejectives are not originally part of the phonological inventory of the English language. Yet, their prevalence in every-day speech is not only well-established, but evidence points to them being phonetic variables at a "prior stage to a more systematic phonological status" [21] (p.191). They have been described to serve a range of interactional, prosodic, and sociophonetic functions, within the English sound system [5, 16, 17, 21]. Structural contexts are known to influence their realization. They appear exclusively word-finally, preferably at the end of phrases or before pauses [17]. It has been proposed that their diachronic emergence in English has most likely been triggered by an internal language source, yet the details of such an account are currently

unclear, as is the nature of how these ejectives are actually produced [21]. These latter points raise questions of broader scientific relevance. The English consonant system is purely pulmonic. What are the circumstances under which sounds of (presumably) glottalic nature develop within this system? What induces an English speaker to actively employ larynx raising during the realization of a stop? In fact, Simpson suspects at least a portion of English ejectives to be produced *without* larynx raising. He suggests that the IOP increase typical for ejectives is caused instead by a pulmonic airstream which continues to emanate through the open glottis after the onset of oral closure. This in turn is, as Simpson suggests, a consequence of the temporal realignment of neighboring glottal and oral articulations as they occur when a glottal stop follows a word-final plosive [21]. Simpson thus provides a pulmonic account of ejective production in English, obviating the puzzle of how glottalic initiation may arise in a purely pulmonic sound system. Yet the extent to which larynx raising actually happens in English ejectives remains poorly understood, due also to the methodological difficulties in observing larynx position. The current study capitalizes on an existing rt-MRI corpus to uncover some of the articulatory and acoustic characteristics of English ejectives. The rt-MRI of a sizable number of speakers offers a unique opportunity to investigate the extent to which ejectives are initiated by vertical larynx movement and determine the acoustic characteristics distinguishing these stops from their pulmonic counterparts.

2. METHODS

2.1. Participants and recordings

This study features rt-MRI data from 27 speakers of Standard Southern British English (13F; mean age=24). Data was collected for an independent study at the Max-Planck-Institute for Biophysical Chemistry (Göttingen, Germany) at 50.05 frames per second with a temporal resolution of 20 ms and an in-plane pixel resolution of 1.4 x 1.4 mm using a 3T MRI system (Magnetom Prisma Fit, Siemens Healthineers, Erlangen, Germany). Synchronized, noise-supressed audio was obtained during the scanning session by means of an Optoacoustics FOMRI III fiber-optic dual-channel microphone (Optoacoustics LTD).

2.2. Materials

The available stimuli consist of 130 lexical items with voiced and voiceless coda stops at three places of articulation (POA), embedded sentence medially into a variety of carrier phrases. Stimuli were produced once per phrasal stress condition (accented vs. deaccented). A total of 5879 tokens were analyzed, 4451 alveolar, 651 bilabial, and 777 velar (items were not balanced for POA, as this corpus stems from an unrelated study on nasalization). Codas were either simplex (/-p, -t, -k, -b, -d, -g/) or /NC/-clusters (/-nt, -nd, -mp, -md, -ŋk, -ŋd/) with a range of preceding vowels and followed by the vowel /ʌ/, e.g., "*Saw **bat** about six*".

The noise-suppressed audio collected during the MRI scanning sessions served as a basis both for stop categorization in the auditory analysis and for the subsequent analysis of acoustic features. The target stop, its closure, and the preceding vowel were manually segmented using Praat [2]. Boundaries marked the interval between the target stop's burst release and the onset of the following /ʌ/.

2.3. Analyses

2.3.1. Auditory categorization

The first question to be addressed concerns the frequency and distribution of ejective realizations. For this, all coda stops in the dataset were auditorily categorized by the first author as either 'ejective' or 'other'. The latter group comprises the subcategories 'pulmonic', 'glottal stop', i.e., the realization of a pure glottal stop with no oral constriction, and 'unclear', i.e., auditorily ambivalent cases.

2.3.2. Acoustic analysis

For this analysis, only tokens from the 'ejective' and 'pulmonic' categories of the 22 speakers who ultimately produced ejectives were retained (n=1622).[1] A range of acoustic measures was selected in order to identify how ejectives may be acoustically distinct from their pulmonic counterparts. Their choice was informed by studies on phonological ejectives in other languages [6, 7, 22, 23, 24, 25]. These measures target the ejective's salient burst quality and the typical interval of acoustic silence caused by the delayed release of the glottal relative to the oral release. *Burst intensity* (dB) was calculated as the mean intensity of the first 40 ms of the burst; *burst intensity difference* (Δintensity) was defined as the difference in mean intensities (dB) between the first 20 ms of the release burst and the following 20 ms interval ($\Delta\text{intensity} = \text{mean intensity}_{\text{phase2}} - \text{mean intensity}_{\text{phase1}}$). This measure is designed to quantify the steep fall-off in intensity that an ejective burst may have in contrast to a pulmonic stop burst. *Break intensity* was the mean intensity (dB) of the silent interval characteristically following an ejective release, quantified here as starting 40 ms after the burst and ending with the onset of the following /ʌ/.

Closure duration (ms) was the interval between closure onset and release of the target consonant. *Break duration* (ms) was calculated as the difference between the release of the target consonant closure and the onset of the following vowel (/ʌ/).

2.3.3. Articulatory analysis

The articulatory movement of the larynx was observed during ejective, pulmonic voiceless and nasal consonants. Larynx movement was quantified by defining a region of interest (ROI) around the bounds of the larynx and tracing the changes in pixel intensity within that region. Larynx displacement (Δlarynx) was calculated as the difference between two positional values extracted from the larynx trace. These values correspond to the position of the lower larynx's edge at the moment of the stop's closure (P1) and burst release (P2), as seen in Figure 1. Positive Δlarynx values mean a raised larynx position at the release relative to the onset of closure, a negative value a lower larynx position.

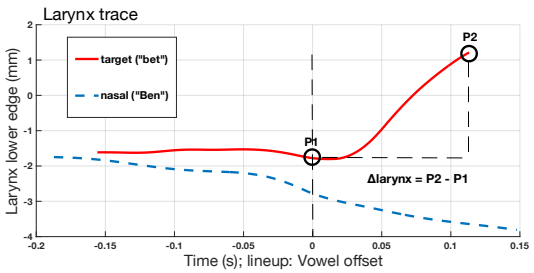


Figure 1: Calculation of larynx displacement using positional values extracted from the ROI's larynx trace. P1=word-final stop closure onset, P2=burst release for plosives/closure offset for nasals.

2.3.4. Statistical analyses

For all continuous articulatory and acoustic measures, linear mixed effects models were conducted using the lmerTest package in R [11, 20]. The five acoustic measures as well as the Δlarynx values were the dependent variables. Duration measures were square root transformed and intensity measures z-score normalized. Stop type (3 levels: ejective, pulmonic voiceless and pulmonic voiced for the acoustic analysis; ejective, pulmonic voiceless and nasal for the articulatory analysis) and POA (3 levels: bilabial, alveolar, velar) were included as fixed factors and a random intercept added for both speaker and lexical item. Pairwise comparisons were conducted using Tukey's post-hoc tests. Significance was evaluated at the $p<.05$ level.

3. RESULTS

3.1. Frequency and structural conditioning

8.5% (n=502) of all word-final stops in the dataset were auditorily identified as ejectives. Five speakers did not produce any.

Place			Voicing		Accent	
bilabial	alveolar	velar	voiceless	voiced	accented	deaccented
11.2%	5.9%	21.4%	12.5%	4.5%	12%	1.2%

Table 1: Percentage of ejectives realized within each phonetic category (for each level of the category 100% is reached adding the non-ejective realizations).

Table 1 shows the influence of phonetic categories on the realization of ejectives. The percentages confirm previous findings that ejectives in English predominantly appear as allophones of originally voiceless, velar stops in accented words [17]. The fact that ejectives replace voiced stops at all, however, is worth noting.

3.2. Acoustic properties of English ejectives

Figure 2 gives the results for the acoustic measures. Following [22], both voiceless and voiced pulmonic stops were considered in the acoustic comparison with ejectives. For all acoustic measures but break duration, main effects of stop type and POA were significant at $p<0.001$ (for burst intensity and closure duration stop type was significant at $p<0.01$). Significant interactions between POA and stop type were found for all measures but closure duration ($p<0.001$).

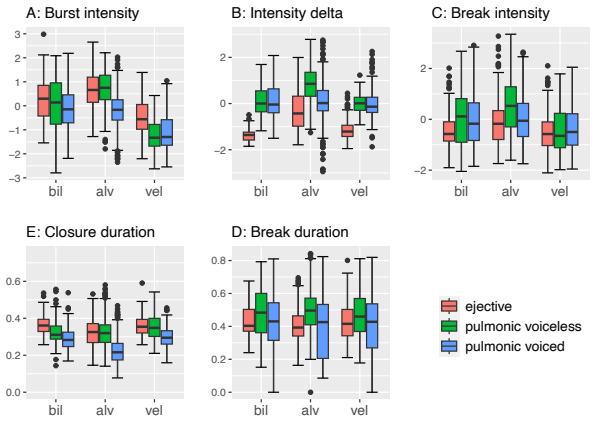


Figure 2: Acoustic measures (A-C z-normalized, D and E square root transformed) by stop type (ejective, pulmonic voiceless, pulmonic voiced) and place of articulation.

Tukey post-hoc tests show that Δintensity is the only measure which significantly distinguishes ejectives from both pulmonic voicing types at all three POAs (p<0.001), as illustrated in Figure 2B. A significant difference in burst intensity (Figure 2A) is found for ejectives and both pulmonic stop types at the velar POA (/k'/vs./k/=p<0.001 and /k'/vs./g/=p<0.05) as well as between alveolar ejectives and pulmonic voiced stops (/t'/vs./d/=p<0.001). Break intensity (Figure 2C) differs significantly for ejectives and pulmonic voiceless stops at both the bilabial (/p'/vs./p/=p<0.001) and alveolar (/t'/vs./t/=p<0.001) POA. Overall, these results suggest that the dynamics of burst intensity are the most consistent acoustic correlate of English ejective production.

3.3. Larynx raising in English ejectives

Figure 3 gives the results for the larynx position measure. It shows a three-way comparison between Δlarynx values calculated for all ejective, pulmonic voiceless and nasal stop tokens in the dataset. Significant main effects of stop type (F[2,30.3] =14.3, p<0.001) and POA (F[2,72]=7.2,p<0.01) were found on Δlarynx values, as well as a significant stop type by POA interaction (F[4,300.1]=17.3,p<0.001). Larynx displacement is significantly different between velar ejectives and both other velar stop types (p<0.001) as well as between bilabial ejectives and their pulmonic counterparts (p<0.05), however, not between any of the three alveolar stop types.

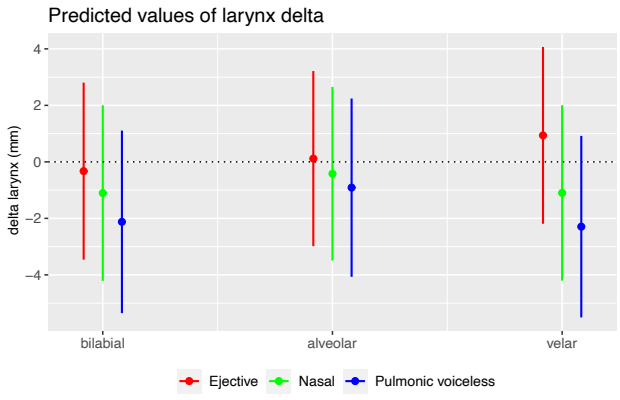


Figure 3: Marginal effects of the three stop types (ejective, nasal and pulmonic voiceless) on Δlarynx values. Positive values indicate raising, negative larynx lowering.

Figure 3 shows a uniform pattern in the predicted Δlarynx values, dependent on the stop type. Across all POAs, the highest values are predicted for ejectives, these are followed by nasals and are lowest for pulmonic voiceless stops. Homogenous larynx behaviour, i.e., larynx lowering, is found for both pulmonic voiceless and nasal stops at all POAs. This is in agreement with previous findings on larynx lowering for nasals in English [19]. Importantly, there is no consistent larynx raising for ejectives across POAs. According to these results, only velar ejective initiation occurs by clear larynx raising in the sense of Catford [3].

4. DISCUSSION AND CONCLUSION

The goal of our study was to investigate larynx raising in English ejectives and to examine their acoustic profile. Articulatorily, we show that ejective initiation is variable in English, they may occur with or without larynx raising. The likelihood of traditional 'Catfordian' initiation, however, is strongly influenced by POA. Velar ejectives are the only ones clearly associated with larynx raising. Our acoustic measures underscore that the most consistent ejective characteristic across POA is burst intensity difference. The two acoustic measures directly targeting burst qualities (burst intensity and Δintensity) are particularly effective at characterizing velar ejectives. This agrees with the impression that /k'/ is the most auditorily salient of the English ejective stops. Combined with the articulatory results, we infer that a larynx raising gesture has a direct effect on ejective burst intensity.

Remarkably, this study finds clear evidence of glottalic initiation in English ejective production. At the same time, it is clear from the results that the Catfordian mechanism does not hold for all of the ejectives in our corpus, and glottalic and (presumably) pulmonic initiation seem to occur side-by-side. It is unclear whether this variation in initiation found here is due to the emergent status of ejectives in English (i.e., a sound change in progress [18]), or whether this may likewise be found among phonological ejectives of other languages. The current study lends further empirical support for the possibility that larynx raising is not the only production mechanism for ejectives [1, 9, 10, 21, 25], even though the exact nature of these 'other' ejective variants remains to be uncovered. It stands to be tested whether these ejective variants are indeed products of a pulmonic airstream mechanism as described by Simpson [1, 21]. Another possibility to be considered is that alternative articulators are acting for the larynx in its role in cavity reduction and IOP increase. It remains to be explored what these articulators may be, other than Kingston's suggestion of tongue root backing [9, 10]. It is conceivable that variations of ejectives, including those outlined above, may co-exist within and across languages to form an articulatory range of ejective production, with pulmonic and glottalic initiations representing two variants among possibly several others.

5. ACKNOWLEDGEMENTS

Work supported by ERC Horizon 2020 grant Interaccent to J. Harrington.

6. REFERENCES

[1] Brandt, E., Simpson, A.P. 2021. The production of ejectives in German and Georgian. *Journal of Phonetics* 89, 101111.
[2] Boersma, P., Weenink, D. 2019. Praat: Doing phonetics by computer [Computer program], version 6.0.50.
[3] Catford, J. 1977. *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press.
[4] Fallon, P. D. 2002. *The Synchronic and Diachronic Phonology of Ejectives*. New York, NY: Routledge.
[5] Gordeeva, O., Scobbie, J. 2013. A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality. *Journal of the International Phonetic Association* 43(3), 249–271.
[6] Gordon, M., Applebaum, A. 2006. Phonetic structures of Turkish Kabardian. *Journal of the International Phonetic Association* 36(2), 159-186.
[7] Grawunder, S., Simpson, A.P., Khalilov, M., 2010. Phonetic characteristics of ejectives – samples from Caucasian languages. In: Fuchs, S., Toda, M. (eds), *Turbulent sounds. An interdisciplinary guide.* De Gruyter Mouton, 209–244.
[8] Ham, S. 2007. Tsilhqut'in ejectives. *2007 Annual Conference of the Canadian Linguistic Association,* University of Victoria.
[9] Kingston, J.C. 1985. *The phonetics and phonology of the timing of oral and glottal events*. Ph.D. thesis, University of California, Berkeley.
[10] Kingston, J.C. 2005. The phonetics of Athabaskan tonogenesis. In: Hargus, S., Rice, K. (eds), *Athabaskan prosody*. Amsterdam: John Benjamins, 137–184.
[11] Kunzetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: tests in linear mixed effect models. *J Stat Softw* 82, 1-26.
[12] Ladefoged, P., 1973. The features of the larynx. *Journal of Phonetics* 1, 73–83.
[13] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Oxford: Blackwell.
[14] Lindau, M. 1982. Phonetic differences in glottalic consonants. *Journal of the Acoustical Society of America* 71, S1. 22.
[15] Lindau, M. 1984. Phonetic differences in glottalic consonants. *Journal of Phonetics* 12, 147–155.
[16] McCarthy, O., Stuart-Smith, J. 2013. Ejectives in Scottish English: A social perspective. *Journal of the International Phonetic Association* 43(3), 273–298.
[17] Ogden, R. 2009. *An introduction to English phonetics*. Edinburgh: Edinburgh University Press.
[18] Ohala, J. J., 1989. Sound change is drawn from a pool of synchronic variation. *Language change: Contributions to the study of its causes*, 173-198.
[19] Riordan, C.J., 1980. Larynx height during English stop consonants. *Journal of Phonetics* 8, 353–360.
[20] RStudio Team (2020). *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
[21] Simpson, A.P. 2014. Ejectives in English and German – linguistic, sociophonetic, interactional, epiphenomenal? In: Celata, C., Calamai, S. (eds), *Advances in Sociophonetics*. Amsterdam: John Benjamins, 187–202.
[22] Vicenik, C. 2010. An acoustic study of Georgian stop consonants. *Journal of the International Phonetic Association* 40(1), 59–92.
[23] Warner, N. 1996. Acoustic characteristics of ejectives in Ingush. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, 1525–1528.
[24] Wysocki, T.M. 2004. *Acoustic analysis of Georgian stop consonants and stop clusters*. Ph.D. thesis, University of Chicago, Chicago.
[25] Wright, R., Hargus, S., Davis, K. 2002. On the categorization of ejectives: Data from Witsuwit'en. *Journal of the International Phonetic Association* 32, 43–77.

[1] The 'unclear' and 'glottal stop' categories were considered irrelevant for the current research question and were removed from further analysis.

The predictability of naturalistic evaluation of all-day recordings for speech and language development

Seunghee Ha[1], Tae-Jin Yoon[2], Jungmin So[3]

[1] Div. of Speech Pathology and Audiology, Hallym University, Korea, shha@hallym.ac.kr
[2]Dept. Of English Language and Literature, Sungshin Women's University, Korea, tyoon@sungshin.ac.kr
[3]Dept. of Computer Science and Engineering, Sogang University, Korea, jso1@sogang.ac.kr

ABSTRACT

This research assessed the predictability of naturalistic evaluation of all-day recordings for speech and language development in children who acquire Korean language. One-day home recordings from 20 Korean children were collected using a Language Environment Analysis (LENA) recorder at 6-8, 12-14, 18-20 months. Both LENA's automated measures and measures from human coding were obtained from recordings at 6-8 and 12-14 months. A number of different words (NDW), consonant inventory, and utterance structure inventory were identified from recordings of 18-20 months. The results showed that conversation turn count (CTC) and canonical babbling ratio (CBR) at 6-8 months could significantly predict NDW at 18-20 months. Also, CTC and CBR at 12-14 months could significantly predict consonant inventory size at 18-20 months. This study highlights that rather than the amount of speech to children or overheard by children, conversational interactions between infants and caregivers positively affect children's later speech and language outcomes.

Keywords: LENA's automated measures, human coding, vocalization, speech & language development.

1. INTRODUCTION

Since Hart & Risley (1995) [1] provided strong evidence that the amount of caregiver speech to children plays a critical role for child cognitive and language development later in life, many studies have focused on the properties of early language environment and their relationship with various aspect of child development. These research efforts have been enhanced in the aspect of research methodology with the advent of automatic speech processing technologies. In particular, the LENA system enables us to collect daylong recordings of a child's language environment and obtain automated analyses of speech data. The LENA system makes it possible to capture children's naturalistic full language environment and overcome methodological limitations associated with language environment research.

The LENA software segments and analyzes the daylong acoustic data, generating automatic estimates related to child vocalization and language environment. The measures include child vocalization count (CVC), adult word count (AWC), and conversation turn count (CTC), which are the most relevant to Hart & Risley (1995)' [1] s measures to reflect the child language environment. Also, the LENA system provides measures that represent child auditory environment such as durations for TV and media, background noise, overlapping speech, and silence. Many studies quantified interactions between children and their caregivers using these LENA automated measures. They made an effort to shed light on the effect of language environment on various aspects of child development and explain individual differences and variability in child development [2-3]. In addition, the research efforts have expanded to the at-risk population, including preterm infants, autism spectrum disorder, and hearing loss [4-6].

Although there is no doubt that LENA automated measures reflect a naturalistic language environment, many researchers from various language backgrounds have tested and discussed the accuracy and reliability of LENA automated measures [7-8]. The research suggests that LENA automated measures have predictability for child language development. According to normative studies of the LENA Research Foundation [9], the segmentation agreement percentages between LENA automated measures and human transcribers were 82% and 76% for adults and children, respectively. This issue related to the validity and reliability of the LENA system is very important so that researchers and clinicians with a variety of language backgrounds can apply the LENA system to their own language settings. As the algorithm used to provide the LENA automated measures was trained on American English, research to verify the reliability of the LENA system in languages other than English should be preceded to adopt the LENA system in their respective languages. For Korean language, the two studies [10-11] evaluated the accuracy of the LENA system applied to Korean. Specifically, McDonald et al. (2021) [10] showed that the correlation between

LENA and human coding for AWC was *r*= .78 or .79 in Korean language, which was identical or similar to that of other studies for other languages [9] as well as Korean [11]. The correlation between LENA and human coding for CVC and CTC was relatively lower (*r*= .34 - .47). However, it fell within ranges found in non-European languages, and Korean language may not lead to more significant error rates than English and other languages. These findings indicate that caution should be needed for research and clinical use of LENA automated measures. For better research and clinical condition, using measures from human coding as a supplementary approach would be ideal.

Furthermore, the LENA automated measures tend to provide only broadly quantified information about the language environment and infant volubility. Specifically, it is difficult to determine to what extent of AWC would involve child-directed speech and to what extent of CVC included speech-like vocalizations or canonical babbling which were considered as more advanced forms of vocalization and precursors to mature phonology. Therefore, many recent studies investigated infant vocalizations utilizing human coding on the LENA recordings [12-13]. Given the importance of the quality of infant-caregiver interactions over the quantity of language input, further investigations are needed to be based on accurate coding of infant-caregiver interactions and as well as valid language environment sampling.

The purpose of this study was to investigate the predictability of measures from LENA recordings reflecting vocal development status and language environment in the prelinguistic stage for later language and speech development. Specifically, this study aimed to explore whether both measures from human coding and the LENA automated measures at 6-8 months and 12-14 months of age predict vocabulary and phonological development at 18-20 months.

2. METHOD

2.1. Participants

Nineteen children in Korean monolingual environments completed all the recording sessions among children recruited for this longitudinal study, and one child will complete the last recording session in January 2023. They consisted of 12 boys and 8 girls. All parents in the study were categorized as middle-class or above and college-educated in two-parent families. All children were recruited when they were between 6 and 8 months of age. According to their parents' reports, none of the children had significant physical, developmental, or hearing problems. This study was approved by the Hallym University Institutional Review Board. Signed informed consent forms were obtained from the children's parents.

2.2. Data Collection

Complete all-day home recordings using LENA recorders were obtained from each child at 3 different time blocks with 6-month intervals. The initial data collection occurred between 6 and 8 months of age, and subsequent data were collected at 12-14 months and 18-20 months of age. Research assistants visited the infants' home, interviewed the primary caregiver regarding the child's general developmental status, and provided detailed instructions on the LENA recorder use. The parents were asked to have their child wear a specially designed vest in a chest pocket with the LENA recorder. The parents were instructed to turn on the LENA recorder as soon as the child awoke in the morning and to maintain the recorder on all day until switching off automatically. Each recording provided up to 12-16 hours of data outside infants' sleep time in their natural environment.

2.3. Data Analysis

2.3.1. LENA automated analysis

Complete LENA recordings were automatically processed with LENA Pro software. Automated measures were obtained for one day-long recording collected at 6-8 months and 12-14 months of age. The measures included (1) AWC (the number of adult words spoken in the child's auditory environment), (2) CTC (the number of conversational turn-taking interactions between the child and adults), (3) CVC (the number of child vocalizations), and (4) the total duration of all segments classified as electronic media exposure. The LENA software supplies data for each measure in terms of both values from the actual recording duration and The 12-hour projected values. The measures were selected from 12-hour projected values as each recording duration was different across sessions and children.

2.3.2. Human coding

In-depth recording analyses on the LENA recordings were conducted based on human coding with the Action Analysis Coding Training (AACT) software. For the human coding, 20 5-minute segments with the child's highest vocalization rate determined by the LENA automated analysis were selected from one-day home recordings of each child.

In this study, an utterance is a vocalization occurring within one breathing cycle and consists of single or several syllables. A syllable in infant vocalization indicates the minimal rhythmic unit of an

utterance, and coders can count the number of syllables in an utterance by listening to "beats" in the rhythm of vocalization during coding [13]. The coders in this study were intensively trained to focus on syllables within an utterance, count the number of syllables, and determine syllable types reliably. They could observe changes in amplitudes of the waveform in the acoustic display of the AACT, by which they could supplement their judgments on the number of syllables. The coders classified syllables as one of four types: true canonical syllables, non-canonical syllables, syllables with glottal consonants only, or syllables with glide consonants only. The coders were trained to mark each syllable as one of the four types in real-time by keystrokes that were automatically recorded in the log of the AACT software. They also were trained to determine utterances of persons other than the child whose voices could be heard on the recordings. They determined whether utterances were directed toward the infant or adults or toward someone other than the infant wearing the record while listening in real-time.

Measures from human coding on the LENA recordings of 6-8 months and 12-14 months included the number of total syllables and the canonical babbling ratio (CBR). CBRs were calculated as the number of true canonical syllables plus syllables with glides by the number of all syllables. Also, the number of utterances directed toward the infant (infant-directed speech [IDS]) was included as a potentially predictive measure.

2.3.3. *Vocabulary and phonological measures*

Vocabulary and phonological measures were analyzed from the LENA recordings obtained at 18-20 months. The coders evaluated vocabulary size from 20 5-minute segments with the child's highest vocalization rate selected from one-day home recordings of each child in the same manner as the LENA recordings of 6-8 and 12-14 months. The coders identified words based on criteria proposed by [14], determinative context, phonetic match, and relationship with other vocalizations and counted the number of different words for each child. The coders then performed broad phonetic transcription on the identified words to determine the consonant inventory and the utterance shapes. The consonant had to occur at least three times in the sample to be considered an "in inventory" consonant for the size of the consonant inventory. In addition, to examine whether children produce various syllable sequences within utterances, the number of different utterance shapes was determined by counting the number of utterance shapes occurring at least three times in the sample.

The two main coders who received training sessions on human coding of vocalization participated in the analysis of recordings from the LENA recordings obtained at 18-20 months of age. They are also involved in the examination of agreements. Approximately 10% of the entire data set was randomly selected and reanalyzed for inter- and intra-coder/transcriber agreements on the vocabulary and phonological measures. Cohen's Kappa coefficients were calculated for inter-coder and intra-coder agreements on the word identification from recordings of 18-20 months, which revealed $\kappa = .752$ and $\kappa = .997$, respectively. The reliability of phonetic transcription was determined on consonants and word shapes. Inter-transcriber and intra-transcriber agreements on consonants were 87.19% and 81.68%, respectively. Inter-transcriber and intra-transcriber agreements on word shapes were 80.07% and 81.80%, respectively.

3. RESULTS

The correlation and linear regression analyses were conducted with measures obtained from 19 children who had completed the longitudinal study. Table 1 presents the results from correlation analyses between speech and language measures at 18-20 months and measures from LENA automated analysis and human coding at 6-8 months and 12-14 months. NDW, consonant inventory, and utterance structure inventory at 18-20 months showed the highest positive correlations with CTCs at 6-8 months and 12-14 months. The amount of IDS also had higher positive correlations with measures from LENA automated analysis and human coding at 6-8 months and 12-14 months.

Linear regressions were conducted to examine the influence of predictor variables on speech-language development at 18-20 months. The regression analyses showed that at 6-8 months, CTC and CBR accounted for 65.8% and 10.9 % of the variance in NDW at 18-20 months, respectively. Also, CTC accounted for 32.8 % and 28.3% of variances in consonant inventory and utterance structure inventory, respectively. At 12-14 months, CTC, AWC, IDS, and CBR significantly accounted for 59.7%, 16.4%, 5.6%, and 5.4% of the variance in NDW at 18-20 months, respectively. CTC and CBR accounted for 38.6% and 16.8% of the variance in consonant inventory, and CTC accounted for 41.8% of the variance in utterance structure inventory at 18-20 months.

Table 1. Pearson coefficients between measurements at 6-8, 12-14months, and speech-language development at 18-

20months

LENA measurement	Speech-language development at 18-20 months		
	NDW	Consonant inventory	Utterance structure inventory
6-8months			
AWC	.525*	.403	.279
CTC	.811**	.573*	.532*
CVC	.795**	.549*	.512*
IDS	.550*	.483*	.452
CBR	.456*	.354	.447
12-14months			
AWC	.356	.338	.396
CTC	.773**	.622**	.646**
CVC	.666**	.487*	.441
IDS	.611**	.347	.590**
CBR	.411	.526*	.421

AWC=adult word counts; CTC=conversational turn count; CVC=child vocalization count; NDW=number of different words;CDS = the amount of child directed speech identified by human coding; CBR = canonical babbling ratio

*p< 05, **p< .01.

4. DISCUSSIONS

The findings provide strong evidence for the predictive validity of the LENA system to analyze language environments with human coding automatically. The results highlight that conversational interactions between infants and caregivers positively affect children's later speech and language outcomes rather than the amount of speech to children or overheard by children. This longitudinal study has significant implications for identifying children at risk for poor speech-language development and providing appropriate early interventions to improve the early home language environment.

5. REFERENCES

[1] Hart, B. and Risley, T.R., 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

[2] Gilkerson, J., Richards, J.A., Warren, S.F., Oller, D.K., Russo, R. and Vohr, B., 2018. Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4).

[3] Wang, Y., Williams, R., Dilley, L. and Houston, D.M., 2020. A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, *57*, p.100921.

[4] Ambrose, S.E., VanDam, M. and Moeller, M.P., 2014. Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and hearing*, *35*(2), p.139.

[5] Caskey, M., Stephens, B., Tucker, R. and Vohr, B., 2014. Adult talk in the NICU with preterm infants and developmental outcomes. *Pediatrics*, *133*(3), pp.e578-e584.

[6] Sulek, R., Smith, J., Bent, C.A., Hudry, K., Trembath, D., Vivanti, G. and Dissanayake, C., 2022. The utility of LENA as an indicator of developmental outcomes for young children with autism. *International Journal of Language & Communication Disorders*, *57*(1), pp.103-111.

[7] Cristia, A., Bulgarelli, F. and Bergelson, E., 2020. Accuracy of the language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), pp.1093-1105.

[8] Gilkerson, J., Zhang, Y., Xu, D., Richards, J.A., Xu, X., Jiang, F., Harnsberger, J. and Topping, K., 2015. Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech, Language, and Hearing Research*, *58*(2), pp.445-452.

[9] Gilkerson, J., Richards, J.A., Warren, S.F., Oller, D.K., Russo, R. and Vohr, B., 2018. Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4).

[10] McDonald, M., Kwon, T., Kim, H., Lee, Y. and Ko, E.S., 2021. Evaluating the Language ENvironment Analysis System for Korean. *Journal of Speech, Language, and Hearing Research*, *64*(3), pp.792-808

[11] Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J.A., Ma, L. and Topping, K., 2016. Effects of feedback on parent–child language with infants and toddlers in Korea. *First Language*, *36*(6), pp.549-569.

[12] Lee, C.C., Jhang, Y., Relyea, G., Chen, L.M. and Oller, D.K., 2018. Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, *50*, pp.140-153.

[13] Ha, S. and Oller, K.D., 2022. Longitudinal Study of Vocal Development and Language Environments in Infants With Cleft Palate. *The Cleft Palate-Craniofacial Journal*, *59*(10), pp.1286-1298.

[14] Vihman, M.M. and McCune, L., 1994. When is a word a word?. *Journal of child language*, *21*(3), pp.517-542.

Canonical Babbling Ratio Extracted from Day-long Audio Recordings: A preliminary report from India

Shoba S. Meera[1], Divya Swaminathan[1], Sri Ranjani V[2], Malavi Srikar[1], Reny Raju[1]

[1]Department of Speech Pathology and Audiology, National Institute of Mental Health and Neurosciences, Bangalore, India, [2]Department of Speech-Language Pathology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India

[1] ssmeera@nimhans.ac.in ssmeeras@gmail.com

ABSTRACT

Presence of canonical babbling (CB) is a crucial vocal developmental milestone in the later part of an infant's first year of life. CB can be measured as a ratio (CBR) – the proportion of canonical syllables to other syllables produced. In this study we present CBR data and consonant inventory from thirteen 6–10-month-old typically developing infants from a culturally and linguistically diverse background. CBRs were computed based on speech samples collected using day-long audio recorders (Language ENvironment Analysis) – a technology being used and reported for the first time in India. All infants above 8 months of age had a CBR of >0.15 indicating they were in the CB stage. Further, all infants in the CB stage produced bilabial sounds, which are some of the earliest consonants present in an infant's consonant inventory. These findings are in-line with literature reported globally, thus generalizing our understanding to previously understudied populations.

Keywords: Canonical babbling ratio, day-long audio recording, infants, Indian languages

1. INTRODUCTION

A critical milestone in an infant's vocal development in the first year of life is the use of adult-like consonant-vowel (CV) combinations with a rapid transition between the CV (e.g., /ba/, /du/)[1], [2]. These CV combinations are called canonical syllables. When canonical syllables are produced together in a repetitive sequence it is called canonical babbling (e.g., /baba/, /dududu/). One way to measure CB is to compute a canonical babbling ratio (CBR). CBR is the proportion of canonical syllables to other syllables produced. A CBR of 0.15 and above has been used to determine if an infant is in the CB stage[3]. Infants produce CB before producing words. CB in typically developing infants is well documented and is present between 6-10 months of age[1], [4], [5]. Some of the earliest sounds infants produce are bilabials and hence CB too comprises bilabial consonant-vowel combinations like /bababa/[6]–[8]. CB and CBR have been shown to predict language outcomes in typically developing infants[3].

Although a perceived notion is that emergence of canonical syllables is independent of the languages children are exposed to, and perhaps the cultures they are brought up in, studies have been primarily conducted in high-income countries (HICs) with largely monolingual samples (see discussion in Cychosz et al., 2021[5]). This bias in our field not only creates serious issues in terms of coverage of human languages and generalizability across human populations (as discussed for instance in Kidd & Garcia, 2022[9]) but also represents an unfair situation, since a majority of the world's infants are growing up in low- and middle-income countries (LMICs). India alone accounts for over 20% of the world's children (https://data.unicef.org/how-many/how-many-children-under-18-are-there-in-india/).

Only recently a paper reported vocal development including canonical syllables from across cultures including some data from LMICs [5]. The study found that canonical syllables were present in most infants by 7 months of age. Further, most infants had a CBR of 0.15 and above by 10 months of age. Although the study measured vocal development in five culturally diverse settings, it did not include a sample of infants from a LMIC that is naturally bi/multilingual, like India. In fact, only a handful among included infants who had exposure to a second language, and no analyses were made to check for potential differences. More broadly, to our knowledge, no previous study has investigated CB development in a naturally bi/multilingual Indian sample, a gap we address in this work.

Most studies that have evaluated early vocal development, including canonical babbling, have used traditional methods like short lab-based[10], [11] recordings or short home recordings[12], [13]. However, these methods do not represent the naturalistic environment of the child, and concern has

been raised about the extent to which observations from short recordings generalize to children's behavior and experiences [14]. A more contemporary approach to measuring CB is using technology that records the entire waking hours of the child e.g., LENA™ (Language ENvironment Analysis)[15], providing a more naturalistic sample of an infant's speech production. These recordings are called day-long audio recordings. Only a few recent studies have used day-long audio recordings to capture and analyse CB[3], [5]. Such technology is relatively new and has seldom been used in LMICs, including India. Also, no such previous study investigated the consonants present in infants' production.

Thus, the current study aimed to add to existing literature on CBR from a culturally and linguistically diverse sample of 6–10-month-old typically developing infants using day-long audio recordings. The study also assessed the consonant inventory that was present in these infants.

2. METHOD

Data from thirteen 6–10-month-old typically developing infants were included in the present study.[i] This sample size is similar to sample sizes found for individual corpora (for instance in Cychosz et al., 2019[5]). Parents provided written consent to participation. The study included infants exposed to one or a combination of the following Indian languages - Hindi, Kannada, Tamil, Malayalam, Telugu, or Indian English.[ii] All infants were screened on two developmental behavioural assessments: (a) The ComDEALL Developmental Checklist (CDDC)[16] which screens for delays in domains such as motor, activities of daily living, language, cognition, social, and emotional skills and (b) The Communication and Symbolic Behavior Scales Developmental Profile – Infant-Toddler Checklist (CSBS – ITC)[17], which rules out potential communication and language delay. Infants were excluded from the study if there was evidence of a genetic condition or syndrome, significant medical or neurological condition affecting development, significant vision or hearing impairment, birth weight <2500 g or gestational age <36 weeks or perinatal brain injury secondary to birth complications.

Canonical babbling was evaluated based on speech samples collected using the LENA digital language processors (DLP)/recorders. Once an infant was recruited to the study, the participating families were given a study kit containing (a) the recorder, (b) a specially designed vest with a front pocket to secure the recorder and, (c) an instruction sheet on how to use the recorder. The researcher also provided instructions on how to use the recorder in-person and over a video call just prior to the start of recording. Families were asked to complete one day of recording when all members of the family were at home and no social events were planned.

2.1. Generating high child vocalization sample to annotate CB

Upon receiving the LENA-DLP/recorder from the family, the recording was uploaded on the LENA Pro 3.4.1 software. The LENA ADEX (Advanced Data EXtractor)[18] program was used to download a .csv file of the 16-hour recording. This file was used to identify three 5-minute segments that contained highest number of child vocalization segments and similarly, three 5-minute segments that contained highest number of adult word segments.[iii] Audio/speech samples from these six segments were compiled into one wave file (total of 30-minutes) that was then annotated using a study specific Graphical User Interface (GUI) developed in Python (Figure-1)

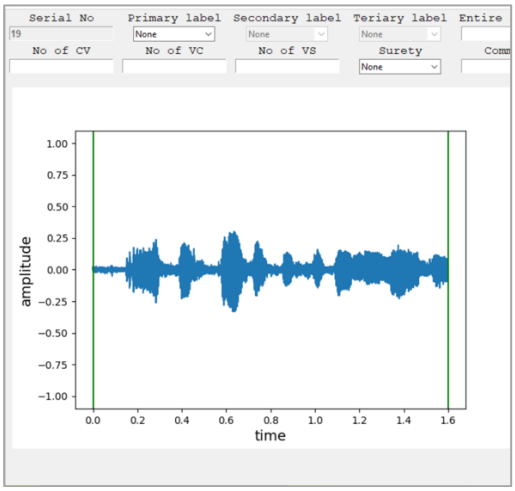


Figure 1: Graphical User Interface (GUI) to annotate CB

2.2. Annotation process:

Two annotators (speech-language pathologists) were trained to tag segments into various primary labels e.g., CH= child segment, FAW= female adult segment, NOIS= noise, etc. Upon successful training (i.e., maintaining an IRR above 0.80 on two successive practice recordings), the annotators annotated a total of 16,260 segments (~390 minutes) in random order.

Inter-rater reliability, Cohen's kappa (k), for CH segments was 0.80 (strong agreement)[19]. Each primary label that was annotated as CH was further annotated into secondary labels of speech-like or non-speech categories. Speech-like category included

vocalizations characterized by the production of consonants and/or vowels. Non-speech included infant cries, raspberries, or squeals. If the secondary label was speech-like, the segment was transcribed. These transcriptions were used to compute CBRs and to describe the consonant inventory. A consonant was considered to be part of the inventory even if it was produced once in the 30-minute sample that was annotated.

3. RESULTS

Data from 13 infants (n=6 females) with a mean age of 8.19 months were analysed. The sample consisted of n=6 (46.15%) monolinguals, n=3 (23.07%) bilinguals and n= 4 (30.76%) multilinguals. Some infants were exposed to languages that were not part of the inclusion criteria (e.g., Marathi). However, these languages were less spoken in the house compared to the primary languages (e.g., Hindi, Kannada). All infants were at, or above par, in their developmental abilities on the CDDC. None of the infants failed the CSBS-ITC indicating no potential risk for language or communication delay. Further, parents reported no concerns in their infant's development. Demographic data of participants and their families are presented in table 1.

Characteristic	Value
	Mean (SD)
Infant age in months	8.19 (1.44)
Maternal age in years	29.92 (2.59)
Paternal age in years	32.46 (1.89)
	n (%)
Infant Sex	
Females	6 (46.15)
Males	7 (53.85)
Maternal Education	
Graduate or post-grad	9 (69.23)
Professional Degree	4 (30.76)
Paternal Education	
Graduate or post-grad	9 (69.23)
Professional Degree	4 (30.76)
Socio Economic Status	
Upper	3 (23.07)
Upper middle	10 (76.92)
Residential area: Urban	13 (100)

Table 1: Demographic data of infant participants and their families

Prior to analysing the number of infants in the CB stage, we ran a Spearman's correlation to determine correlation between age of the infant and total child vocalizations ($r=0.78$, $p<0.001$), and infant age and CBRs ($r=0.64$, $p=0.017$). There was a positive correlation in both cases. See Figure-2

We further found that 10 infants (76.92%) had a CBR of > 0.15 i.e., they were in the CB stage. The three infants (n=1 6mo/female, n=2 7mo/male) who were not yet in the CB stage, had several speech-like vocalizations but no CV combinations. The consonant inventory identified in the 10 infants who had a CBR of >0.15 is presented in Table 2.

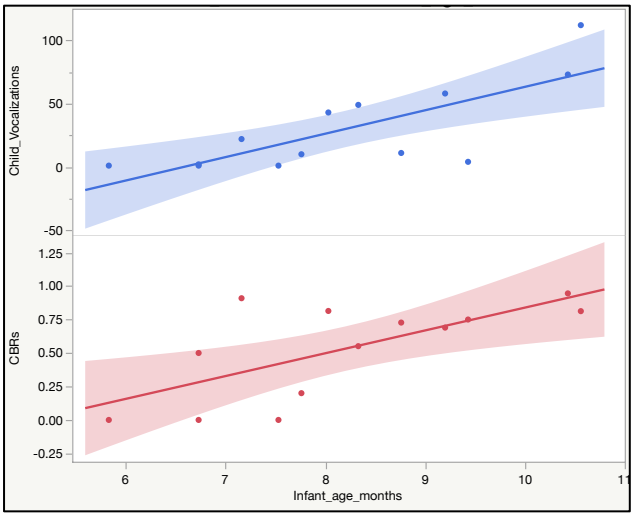


Figure 2: Correlation between infant age and CBRs and child vocalizations

We further found that 10 infants who were in the canonical babbling stage produced bilabial sounds (/m/ as in "mat" 7/10, /b/ as in "bat" 7/10, /p/ as in "pat" 4/10 infants). Other consonants that were present include /j/ as in "yes" (5/10), /n/ as in "no" (4/10), /t/ as in "tap" (4/10), /g/ as in "go" (4/10), /d/ as in "dog" (3/10), /w/ as in "water" (2/10), /h/ as in "hat" (2/10), /k/ as in "king" (2/10), /v/ as in "van" (3/10), /l/ as in "lamp" (1/10).

4. DISCUSSION

The current study from a culturally and linguistically diverse sample from India adds evidence to existing literature on CB – a huge population that is understudied. There was a positive correlation in total child vocalizations and CBRs with infants' age indicating overall increased vocal production with age like previously reported literature[20], [21]. All infants 8 mo and older had a CBR of >0.15 indicating they were in the CB stage. This finding is in line with several other studies that have evaluated CBR to determine the presence of CB stage e.g., Cychosz et al, 2021 [5].

Despite potential concerns for non-generalizability of previous results, this study also converged with previous ones in demonstrating that the first sounds infants produce are bilabial sounds[6]–[8]

CBR was computed from speech samples collected through recorders that collect day-long audio recordings thus providing a more naturalistic sample that accounts for fluctuations in vocalizations within the recording day. Use of day-long audio recorders (LENA) was new for participating families. They asked several questions about the recorder, potential harmful electronic radiation, data storage, etc. Eventually, all families that were approached provided written consent for participation and completed 16 hours of recording.

ID	Age	Language(s)	Consonant inventory
7	7 mo	Kannada	/m/
18	7 mo	Hindi English	/m/, /b/, /p/
9	8 mo	Kannada	/b/, /j/
3	8 mo	Telugu English	/m/, /n/, /b/, /h/ /t/, /k/, /g/, /j/
10	8 mo	Kannada Hindi Marathi English	/m/, /n/, /l/, /g/, /j/
15	9 mo	Telugu	/m/, /t/, /v/
5	9 mo	Hindi Gujarati English	/b/, /w/, /d/, /p/, /h/, /t/, /k/ /g/, /j/
21	9 mo	Hindi Kannada English	/b/
22	10 mo	Malayalam	/m/, /n/, /b/, /d/, /p/, /t/, /g/, /v/, /j/
30	10 mo	Telugu Tamil English	/m/, /n/, /b/, /w/, /d/, /p/, /v/

Table 2: Consonant inventory

These data could be useful for other endeavors: A total of 16,260 segments (~390 minutes) were annotated by two independent rates thus developing an initial pool of well annotated segments that can be used for future studies – e.g., building algorithms to automatically count number of canonical syllables.

In the future, we hope to include a larger sample of infants, more homogeneity within mono/bi/multilingual participants, and representation from different socio-economic strata. An approach that involves citizen scientists [5] may be beneficial for studies such as these, which are labour intensive that make it particularly harder for researchers from LMICs to contribute to literature since resources are limited. That said, there is at present no alternative to careful laboratory annotations to assess infants' consonant inventory, which is one of the strengths of our contribution. More broadly, we hope more diverse researchers contribute similar data, so that together we can start seeing similarities and differences across children as a function of the language(s) they are exposed to and their sociocultural background.

5. REFERENCES

[1] D. K. Oller, "The emergence of the sounds of speech in infancy," in *Child Phonology*, 1980. doi: 10.1016/b978-0-12-770601-6.50011-5.

[2] C. Stoel-Gammon, "D. K., Oller, The emergence of the speech capacity. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. ISBN 0-8058-2629-7.," *J Child Lang*, vol. 30, no. 3, pp. 731–734, Aug. 2003, doi: 10.1017/S0305000903005737.

[3] C. C. Lee, Y. Jhang, G. Relyea, L. mei Chen, and D. K. Oller, "Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings," *Infant Behav Dev*, vol. 50, p. 140, Feb. 2018, doi: 10.1016/J.INFBEH.2017.12.002.

[4] R. E. Stark, "Stages of speech development in the first year of life," *Child Phonology*, pp. 73–92, Jan. 1980, doi: 10.1016/B978-0-12-770601-6.50010-3.

[5] M. Cychosz *et al.*, "Vocal development in a large-scale crosslinguistic corpus," *Dev Sci*, vol. 24, no. 5, p. e13090, Sep. 2021, doi: 10.1111/DESC.13090.

[6] T. Másdóttir, S. McLeod, and K. Crowe, "Icelandic children's acquisition of consonants and consonant clusters," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 5, 2021, doi: 10.1044/2021_JSLHR-20-00463.

[7] C. K. S. To, Cheung, Pamela S.P, McLeod, and Sharynne, "A population study of children's acquisition of Hong Kong Cantonese consonants, vowels, and tones," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, 2013, doi: 10.1044/1092-4388(2012/11-0080).

[8] L. McCune and M. M. Vihman, "Early Phonetic and Lexical Development: A Productivity Approach," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 3, 2001, doi: 10.1044/1092-4388(2001/054).

[9] E. Kidd and R. Garcia, "How diverse is child language acquisition research?," *First Lang*, vol.

42, no. 6, 2022, doi: 10.1177/01427237211066405.
[10] M. R. Talbott, C. A. Nelson, and H. Tager-Flusberg, "Maternal Vocal Feedback to 9-Month-Old Infant Siblings of Children with ASD," *Autism Res*, vol. 9, no. 4, pp. 460–470, Apr. 2016, doi: 10.1002/AUR.1521.
[11] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *J Autism Dev Disord*, vol. 44, no. 10, 2014, doi: 10.1007/s10803-014-2047-4.
[12] B. Hart and T. R. Risley, "Meaningful differences in the everyday experience of young American children," p. 268, 1995.
[13] L. D. Lopez, E. A. Walle, G. M. Pretzer, and A. S. Warlaumont, "Adult responses to infant prelinguistic vocalizations are associated with infant vocabulary: A home observation study," *PLoS One*, vol. 15, no. 11 November, 2020, doi: 10.1371/journal.pone.0242232.
[14] E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, and S. Tor, "Day by day, hour by hour: Naturalistic language input to infants," *Dev Sci*, vol. 22, no. 1, 2019, doi: 10.1111/desc.12715.
[15] J. Gilkerson and J. A. Richards, "The LENA natural language study, LENA Technical Report: LTR-02-2," 2009.
[16] P. Karanth, S. Shaista, and N. Srikanth, "Efficacy of communication DEALL--an indigenous early intervention program for children with autism spectrum disorders," *Indian J Pediatr*, vol. 77, no. 9, pp. 957–962, Sep. 2010, doi: 10.1007/S12098-010-0144-8.
[17] A. Wetherby and B. Prizant, *Communication and symbolic behavior scales: Developmental profile*. 2002. Accessed: Jan. 04, 2023. [Online]. Available: https://psycnet.apa.org/record/2002-17060-000
[18] D. Xu, U. Yapanel, S. Gray, and C. T. Baer, "The LENA(TM) language environment analysis system: The interpreted time segments (ITS) file," *LENA Foundation Technical Report*, no. September 2008, 2008.
[19] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem Med (Zagreb)*, vol. 22, no. 3, 2012, doi: 10.11613/bm.2012.031.
[20] D. K. Oller, R. E. Eilers, R. Urbano, and A. B. Cobo-Lewis, "Development of precursors to speech in infants exposed to two languages," *J Child Lang*, vol. 24, no. 2, 1997, doi: 10.1017/s0305000997003097.
[21] A. S. Warlaumont and H. L. Ramsdell-Hudock, "Detection of total syllables and canonical syllables in infant vocalizations," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-September-2016. doi: 10.21437/Interspeech.2016-1518.

[i] Data presented here are from infants who have participated in a larger study.

[ii] Only infants exposed to these languages were selected since (a) these are languages widely spoken in the study site (b) clinician-researchers/annotators in the lab spoke these languages only (c) these languages were part of the inclusion criteria of the main study these infants were recruited from

[iii] Requirement for the bigger study.

REINFORCING INFANT VOCALISATIONS THROUGH SELF-INITIATED FEEDBACK: A TOOL FOR BASIC RESEARCH AND FOR CLINICAL INTERVENTIONS

Tamar Keren-Portnoy[1], Helena Daffern[2], Rory A. DePaolis[3], Christopher M. M. Cox[1,4], Kenneth I. Brown[2], Florence A. R. Oxley[5], & Mona Kanaan[6]

Affiliations:
[1]Department of Language and Linguistic Science, University of York
[2]School of Physics, Engineering and Technology, University of York
[3]Department of Communication Sciences and Disorders, James Madison University
[4]Interacting Minds Centre, Aarhus University
[5]Department of Education, University of York
[6]Department of Health Sciences, University of York

Email addresses for correspondence: tamar.keren-portnoy@york.ac.uk

ABSTRACT

We introduce a method for reinforcing vocalisations in prelinguistic infants (ages 6-7 months) without recourse to social feedback, through a self-reinforcing app. Study 1 presents findings that show that infants (N = 60; n= 30 in the experimental group) respond to the app's reinforcement by increasing their vocalisations. Study 2 is a randomised clinical trial (data collection ongoing). Infants from low socioeconomic-status families were recruited (N = 96, n = 48 in the intervention group). Families in the intervention group use the app over 3 weeks. Infants' are followed over 9 months until age 16 months. Demographic data for the two groups will be presented. We predict that infants in the intervention group will be more linguistically, but not motorically, advanced than those in the control group by the end of the study. We discuss potential uses of this method for basic as well as clinical research and intervention.

Keywords: Vocalisation, babble, reinforcement, low SES, language development

1. INTRODUCTION

Babble is understood to equip infants with the 'toolkit' required to approach word production [1]. This paper describes a study in which we aimed to encourage babble in infants, to see if that might affect later lexical development. This approach will quantify the relative importance of babbling for typical development and highlight effective clinical interventions for infants at risk for language delay.

The claim that babble is a tool for language learning is based on two types of evidence. The first are findings showing that children's first words share many structural characteristics with their babble ([2], [3]), indicating the continuity between the two behaviours. The second are findings showing that age at consonant mastery (through babble) correlates with the age at first words: children who master consonants in their babble earlier tend to start producing words earlier ([1], [4]). As infants master consonant production through continued practice, these consonants become easily accessible, sounds that the child can target at will and produce relatively effortlessly. These sounds also become especially salient when heard in the input, as produced by others ([5], [6]). As a result, the infant notices and begins to recognise word forms which contain the sounds they are particularly adept at producing, and thus the infant can direct their attention to the meaning or social context associated with those word forms. That leads the infant to be able to produce their first words, with scaffolding of the consonantal sounds they have mastered through babble.

However, to date, the association between age at consonant mastery and at first word production has always been correlational. We have not yet been able to show that the relationship between age at consonant mastery and at the transition to word production is causal. This project aimed to find a way to show such a causal relationship. Our method was to randomly assign infants to groups, one of which will be encouraged to babble more and as a result may master consonants earlier, and the other whose babble will develop at its natural pace, with no special interference. However, it is crucial that the intervention only affects the frequency of babble, without affecting a change in the infant's social world or in the rate of interaction the infant experiences.

To achieve this, we developed an app that responds to infants' vocalisations with moving colourful shapes

that appear on an iPad screen, and remain on it for the duration of the vocalisation [7]. We reasoned that if infants could learn that their vocalisations are causing the shapes to appear, they would vocalise more in order to see more such shapes.

We now describe two studies that utilise this approach. The first (Study 1; [8]) was intended to test whether young infants can learn to use the app. The second (Study 2), which is still ongoing, is intended to test whether infants who use the app over several weeks begin to produce words earlier than those who have not used it.

2. STUDY 1

2.1 Method

Participants were 60 6.5-month-old infants. The infants in the experimental group (n = 30, 19 females, mean age 198 days) received the app described above. Each infant in the control group (n = 30, 12 females, mean age 197 days) was yoked to an infant from the experimental group, and was shown the 'film' created by the sequence of moving shapes that resulted from the experimental group infant's vocalisations.

The infants first played with a non-responsive toy for 5 minutes (non-responsive trial) while the app recorded them without being seen by them, and tallied their vocalisations. Following that, the experimental group played with the app for 5 minutes and the control group watched the videos for 5 minutes (app trial).

We predicted that the infants in the experimental group would increase their vocalisations more when interacting with the app than would the infants in the control group who interacted with a non-responsive film.

2.2 Results

No evidence of a difference was observed between the groups in terms of the number of vocalisations during the app trial, after controlling for each child's baseline level of vocalisations (from the non-responsive trial).

In a second analysis we divided the app session into two halves, each 2.5 minutes long, and calculated the proportion of vocalisations produced in the second half out of the entire 5-minute period. Proportions over 0.5 indicate that the child vocalised more during the second than during the first half. Proportions under 0.5 indicate a reduction in the frequency of vocalising from the first to the second half. In the experimental group 25 out of 30 infants (83%) increased their frequency of vocalising from the first to the second half of the session (see Figure 1). In the control group only 14 out of 29 infants (47%) increased the frequency of vocalising from the first to the second half of the session. We compared the mean proportion of vocalisations in the second half out of the entire 5 minutes in each group to 0.5, the expected proportion under chance. The mean proportion in the control group (M = 0.48) was not significantly different from 0.5 ($t(28) = .0.528$, $p = .6$). However, the mean proportion in the experimental group (M = 0.64) differed significantly from 0.5 ($t(29) = 4.779$, $p < .001$); see Figure 1.

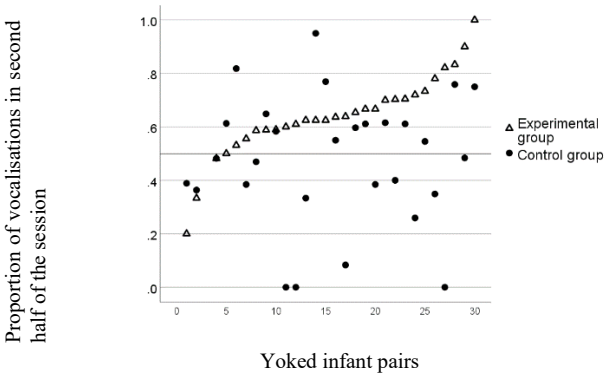


Figure 1: Proportion of vocalisations in the second half of the session in each yoked pair of infants.

In terms of the pairs, in 21 pairs out of 29, the proportion of vocalisations in the second half out of the 5 minutes was higher in the experimental group infant than in the control group infant (see Figure 1: Only in a minority of cases are the filled circles, indicating the member of the yoked pair who was in the control group, higher than the empty triangle, which indicates the experimental member of the yoked pair). A paired t-test showed this difference in proportions to be significant: M(difference) = .16, $t(28) = 3.195$, $p = .003$.

2.3 Discussion of Study 1

The results of Study 1 indicate that the 6.5-month-old infants in the experimental group learned that their vocalisations led to the visual response, and they increased their vocalising while interacting with the responsive app in order to see more visual responses. The fact that the control group didn't increase their vocalisations in the same way indicates that this increase in the experimental group was not in response to seeing the shapes after they appeared, but as a result of the infants, at some level, understanding that their vocalisations cause the shapes to appear (note that for 83% of infants in the control group, more shapes appeared on the screen in the second half of the session than in the first half, but those same infants did not systematically increase the frequency of their vocalisations in response to the rise in the frequency of the shapes). This difference between the

groups was also seen when looking at the yoked pairs, since in the majority of pairs, the infant with the responsive app increased their vocalising more than the infant with the non-responsive app.

We take this result to show that the app can be used to reinforce vocalising in infants, where the incentive to vocalise more is child-generated, and the learning of the relationship between the vocalisations and their external result (the visual scenes on the screen) is done by the infants, with no adult intervention, adult feedback, or social feedback.

This app allows researchers to manipulate the amount of vocalising practice a child experiences. We see this as a novel way of investigating self-generated rewards as part of infant learning. More specifically, we see the use of the app as a novel method that will allow further investigation of the causal relationship between a child's own production experience and practice and their later language development.

But if we can manipulate the amount of vocalising practice that a child experiences, might that not also be harnessed to encourage more vocalising in children who tend not to vocalise much, as a way of improving their future language/lexical prospects?

Study 2 was designed to combine these two aims: to test the causal relationship between vocalising (or babbling) practice and later lexicon development and to do it with a population who we thought would benefit from an intervention delivered through the app: infants from families of low socioeconomic status (SES). Infants from low SES typically vocalise less than those from middle SES homes [9], and typically have smaller lexicons than infants of a similar age from middle or high SES ([10], [11]). Study 2 was run as a clinical trial and was pre-registered (https://www.socialscienceregistry.org/trials/5552).

3. STUDY 2

We recruited 96 families of low SES (49 females, 46 males; one unknown), living in the UK. Ten families out of the 96 have withdrawn. The eligibility criteria for being considered low SES were any one of the following: (1) being considered 'at risk' or vulnerable by the council's children's services, or (2) financial criteria – eligibility for Universal Credit, the government scheme for help with living costs for people with low income, or (3) educational criteria – a requirement that neither parent/carer (in families with more than one carer) would be educated beyond Level 3 on the UK government qualification skills scale (e.g., A-levels), equivalent to 13 years of education, which are currently the number of compulsory years of education in the UK. Eligibility was determined based on families' reports, and no evidence for either financial situation or educational achievements were requested. Families who were recruited to the study were randomly assigned to either the intervention group or the control group. The assignment to groups was done by people not involved in delivering the iPads to families. The intervention group received the app described above. The control group received a reaching app, designed specifically for this study. The aim was for the control group to receive a similar non-language experience to that of the intervention group: both apps allowed the infants to learn to control an external apparatus (the app), an empowering experience for such young infants who have very little control over their own bodies, let alone over external objects. Both involved the child playing with an app with no adult 'teacher' to supply feedback or responses. And both were deemed to be potentially beneficial for skill development – practising vocalising in the one case, and practising reaching for moving objects in the other. The reaching app presented colourful moving shapes on an iPad screen (similar to those used in the vocalising app). When touched, the shapes changed from colour to black and white, grew bigger and then disappeared. The screen would then be black until a new shape appeared. Shapes that were not touched disappeared from the screen after 4-10 seconds. The reaching app, like the vocalising app, recorded the infant's vocalisations and tallied them. In addition, it counted the number of times the infants managed to touch the shapes.

This is a longitudinal study, with infants being followed from age 7-8 months until age 16 months.

3.1 Method

Families were given an iPad with one of the two apps on it, and asked to play with their child with the app twice a day for 5 minutes at a time for 3 weeks, or, if that wasn't possible, then as much as they could handle.

Before starting to play with either app and following the 3 weeks with the app, the families are followed with monthly motoric [12] and language [13] questionnaires, from age 7 or 8 months until age 12 months, and then once more at 16 months. At the start and at the end the families also fill in a questionnaire about their beliefs regarding the importance for infants of being spoken to [14].

3.2 Expected results

To date, all the families who have not withdrawn completed the app stage of the study. Forty-eight infants were randomly assigned to each group (although, due to technical problems and experimenter error, 3 families who were supposed to

receive the reaching app, received the vocalising app instead). By the date of the presentation we will have results from the app stage and from the questionnaires from the early ages. We do not present preliminary results at this stage because we have not planned for an interim analysis. Such analyses would require appropriate advanced adjustments to the statistical methodology utilised including sample size calculations. Therefore, we will only provide analyses comparing the two groups after the data collection is completed.

4. GENERAL DISCUSSION

This paper describes two studies that have made use of a new app intended to encourage vocalising in infants. In Study 1 we have shown that infants increase their vocalisations while they are interacting with the app. In Study 2 we are testing whether interacting with the app repeatedly over several weeks leads to a consistent rise in vocalisations, again – during the interactions with the app. Future studies will be needed to determine whether playing with the app over a period of time will lead to increased vocalisations outside of the interactions with the app, such that infants who play with the app often also vocalise more overall than those who haven't played with the app. However, even if we find this is not the case, as long as infants tend to be particularly vocal while interacting with the app, new avenues for both research and intervention become possible. Researchers will be able to use the app to study the effects of vocal practice in both typically and atypically developing populations, using a true experimental design, with random assignment to groups. Clinicians will be able to make use of the app to encourage more vocalising in infant (or child) populations that tend to vocalise little.

5. ACKOWLEDGMENTS

This work was supported by a C2D2 grant part funded by the Wellcome trust, a University of York Early Commercialisation Fund and an EPSRC Impact Accelerator Award (Study 1), and by a British Academy Grant and a University of York Research Priming Fund (Study 2).

5. REFERENCES

[1] McCune, L., Vihman, M. M. 2001. Early phonetic and lexical development: A productivity approach. *J. of Speech, Language, and Hearing Research, 44,* 670–684.

[2] Oller, D. K., Wieman, L. A., Doyle, W. J., Ross, C. 1976. Infant babbling and speech. *J. of Child Language, 3,* 1-11.

[3] Vihman, M. M., Mcaken, M. A., Miller, R., Simmons, H., Miller, J. 1985. From babbling to speech: A reassessment of the continuity issue. *Language, 61,* 395-443.

[4] McGillion, M. L., Herbert, J. S., Pine, J. M., Vihman, M. M. DePaolis, R. A., Keren-Portnoy, T., Matthews, D. E. 2017. What Paves the Way to Conventional Language? The Predictive Value of Babble, Pointing, and Socioeconomic Status. *Child Development, 88*(1),156-166.

[5] DePaolis, R. A., Vihman, M. M., Keren-Portnoy, T. 2011. Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behaviour and Development, 34,* 590-601.

[6] Majorano, M., Vihman, M. M., DePaolis, R. A. 2014. The Relationship Between Infants' Production Experience and Their Processing of Speech. *Language Learning and Development*, *10*(2), 179–204.

[7] Daffern, H., Keren-Portnoy, T., DePaolis, R. A., Brown, K. I. (2020). BabblePlay: An app for infants, controlled by infants, to improve early language outcomes. *Applied Acoustics, 162,* 107183.

[8] Keren-Portnoy, T., Daffern, H., DePaolis, R. A., Cox, C. M. M., Brown, K. I., Oxley, F. A. R., Kanaan, M. (2021). "Did I just do that?"—Six-month-olds learn the contingency between their vocalizations and a visual reward in 5 minutes. *Infancy, 26*(6), 1057-1075. https://doi.org/10.1111/infa.12433

[9] Oller, D. K., Eilers, R. E., Steffens, M. L., Lynch, M. P., Urbano, R. 1994. Speech-like vocalizations in infancy: an evaluation of potential risk factors. *J. of Child Language*, *21*(1), 33–58.

[10] Furey, J. E. 2011. Production and maternal report of 16- and 18-month-olds' vocabulary in low- and middle-income families. *American J. of Speech-Language Pathology / American Speech-Language-Hearing Association*, *20*(1), 38–46.

[11] Hoff, E. 2003. The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378.

[12] Sparrow, S. S., Cicchetti, D., Balla, D. A. 2005. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II) [Database record]. APA PsycTests. https://doi.org/10.1037/t15164-000

[13] Alcock, K. J., Meints, K., Rowland, C. F. 2017. *UK-CDI Words and Gestures - Preliminary norms and manual.* Retrieved from http://lucid.ac.uk/ukcdi

[14] Kennison, S. M., Byrd-Craven, J. 2015. Gender Differences in Beliefs about Infant-Directed Speech: The Role of Family Dynamics. *Child Development Research*, *2015*. https://doi.org/10.1155/2015/871759

THE DEVELOPMENT OF CANONICAL PROPORTION CONTINUES PAST TODDLERHOOD

Kasia Hitczenko[1], Elika Bergelson[2], Marisa Casillas[3], Heidi Colleran[4], Margaret Cychosz[5], Pauline Grosjean[6], Lisa R. Hamrick[7], Bridgette L. Kelleher[7], Camila Scaff[1,8], Amanda Seidl[7], Sarah Walker[6] & Alejandrina Cristia[1]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, EHESS, CNRS, PSL University

[2]Duke University, [3]University of Chicago, [4]Max Planck Institute for Evolutionary Anthropology,

[5]University of Maryland, [6]University of New South Wales, [7]Purdue University, [8]University of Zurich

kasia.hitczenko@ens.psl.eu

ABSTRACT

A key aspect of the development of speech production, the emergence of vocalizations that combine consonants and vowels, is captured by a measure called *canonical proportion* (CP). Yet this measure has mainly been studied among children under 12 months old learning English. We study CP in naturalistic speech in a cross-linguistic sample of 129 children aged 1 to 72 months. We show that children's CP continues to grow well after 12 months, and that CP development may vary cross-linguistically/culturally. This study has implications for how we conceptualize and monitor the development of speech production and showcases how coarse, semi-automated approaches can be used to study cross-cultural speech development from natural production data.

Keywords: language acquisition; citizen science; child vocalizations; speech production; long-form recordings

1. INTRODUCTION

Children's speech production undergoes rapid development in the first years of life. Young infants only produce "non-canonical" vocalizations, consisting of just a vowel or just a consonant (e.g., "uh" or "mmm"), but at around 6 months, they begin combining consonants and vowels together, with increasingly fast, adult-like, consonant-vowel and vowel-consonant transitions (e.g., "ba" or "up") [1]. This development has traditionally been captured with a measure called "canonical babbling ratio", or the proportion of syllables that contain such transitions, which reaches 15% by 10 months of age among English learners [2]. Less is known about how frequently such canonical transitions occur outside of babbling, when looking at older children, as well as for more diverse populations, leading to two open questions that we aim to address here.

First, beyond 12 months, researchers have typically turned to other aspects of phonological development, leading to an incomplete understanding of the developmental timeline. Recent work has suggested a new definition [3]: "canonical proportion" (CP), defined as the proportion of vocalization sections that have adjacent consonants and vowels (without regards to transition speed). Thus defined, CP can be calculated in all speech-like vocalizations, both from meaningful ones and meaningless babble. At 10 months, CP values thus defined are 15%, matching previous work based only on babbles. By 36 months, CPs are only 40% [3]. While there is no research establishing CP in adults' speech, 40% intuitively feels small, suggesting that this aspect of children's vocalization development may continue beyond 36 months.

By studying CP in older children, we can both better understand the full trajectory of children's phonological development and potentially establish a stable metric of phonological development: one that can be used throughout the first years of life and can be easily calculated from naturalistic recordings collected from highly-diverse communities [3, 4].

Second, the literature on group differences in CP development is small, probably because most researchers relied exclusively on (time-consuming) manual annotation, which offers detailed phonological information about early production, but limits sample sizes. Phonological

properties of the language being acquired impact other aspects of children's vocalizations [5, 6], and the same could be true here. For example, children learning languages with low syllabic complexity (i.e., all syllables are consonant-vowel, e.g., "ma") might learn faster than children learning languages with high syllabic complexity (i.e., languages that allow many different syllabic types, e.g., words like "striped" as well as "pea"), due to fewer targets to learn, more exposure to them, and no need to attempt difficult syllable types, which could lead to non-adult-like performance. As for other group differences, results are mixed, with some research documenting individual and group differences [7] in part as a function of caregiver responsiveness, which likely differs across rural (small-scale, subsistence-level) and urban (industrialized or post-industrial) communities [8]; whereas others report similarities [3]. In sum, more work is needed to establish whether CP varies across children learning typologically diverse languages and/or growing up in diverse communities.

This leads us to our two main research questions: **(Q1)** How does CP develop from toddlerhood to 6 years of age? **(Q2)** Can we observe statistically significant differences in CP development that could be due to diverse languages (e.g., maximum syllable complexity) and/or communities (e.g., rural versus urban)? We address these questions using an innovative method: long-form audio recordings of children's everyday lives analyzed semi-automatically, by combining machine learning algorithms and crowd-sourced labels.

2. METHODS

All code is available at github.com/khitczenko/canprop-by-age_icphs2023.

2.1. Participant Recordings

We study a cross-linguistic sample of 129 children aged 1-72mos, learning English (N=20; 5-19mos) [9, 10, 11, 4], French (N=10; 11-13mos) [12], Quechua/Spanish (N=3; 22-25mos) [13], Tseltal (N=10; 2-36mos) [14, 15], Tsimane' (N=30; 6-71mos) [16, 17], Yélî Dnye (N=41; 1-72mos) [18, 19], or a subset of 12 Austronesian languages, spoken in the Solomon Islands (N=15; 4-48mos) [20]. Children from these 8 different communities wore non-intrusive, light-weight audio recorders in a specialized vest/shirt as they went about a typical day. These recordings provide a large and ecologically-valid sample of children's language development [21].

2.2. Labeling Vocalization Types

We first randomly sampled 100–300 child vocalizations from each long-form recording (5-16h long), using speech technology algorithms (LENA [22, 23, 24] or Voice Type Classifier [25]).[1] For older children, these vocalizations may be meaningful words/phrases, in line with our goal of representing CP in *all* spontaneous child production.

Citizen science is a growing approach used to manage large quantities of scientific data where volunteers assist with research tasks online. This approach allowed us to study many more children than would be feasible otherwise, without compromising label reliability: CP calculated on the basis of the crowd-sourced citizen science labels we use are highly correlated (r = 0.93) with those calculated using traditional in-lab approaches [4].

To obtain citizen science labels, the 100–300 vocalizations from each child were first split into short 400–500ms clips. This step was taken to protect the privacy of those who were recorded because the clips were shared on the Internet via a citizen science platform. Then, 3–5 citizen scientists labeled each clip for vocalization type: Canonical, Non-Canonical, Laughing, Crying, or Junk (i.e., the clip does not contain voices, or there is so much overlap that it is difficult to make out properties of the vocalization).[2] Each clip's category was determined via "majority rule," or excluded if no majority was reached (see [3, 4] for more details).

2.3. Calculating Canonical Proportion

CP was calculated, for each child, as the number of clips that were labeled as Canonical divided by the number of clips that were labeled as Canonical or Non-Canonical [3]: 0 means all of the child's speech-like clips were labeled as Non-Canonical; 1 that all were Canonical.

3. RESULTS

Seeking a response to **(Q1)**, we compared two models: (1) a mixed effects logistic regression predicting the effect of age on CP, controlling for gender and including corpus in the random effects structure and (2) the same model but instead with age-squared as a predictor.

$$(1)\quad CP \sim age_{child} + gender_{child} + (0 + age_{child}|corpus)$$

$$(2)\quad CP \sim age^2_{child} + age_{child} + gender_{child} + (0 + age^2_{child} + age_{child}|corpus)$$



Figure 1: CP as a function of age, colored by corpus (top), syllabic complexity (middle), and rural/urban (bottom). CP continues to increase past toddlerhood, and may differ cross-linguistically/culturally. For subset analyses, we exclude (grayed-out) children in non-shared age ranges.

Fitting a quadratic effect of age allows us to model a non-linear relationship between canonical proportion and age. A log-likelihood chi-squared test comparing the two models (-1121.0 vs. -1029.7, $\chi^2(3) = 182.65$, $p < 0.001$), as well as comparison of AICs (2250.1 vs. 2073.4) and BICs (2261.5 vs. 2093.4), revealed that the quadratic model was a better fit, suggesting that while the measure continues to increase beyond toddlerhood, it may do so at a slowed rate with increasing child age.

Seeking a response to **(Q2)**, we first visually inspected CP by corpus (Figure 1). The corpora seem to differ in how quickly CP grows initially (e.g., it appears to grow faster for children learning Austronesian languages in the Solomon Island corpus, than in Tseltal, than in Tsimane'), as well as when we see a plateau (e.g., Yélî Dnye appears to have plateaued around 40 months of age at a high CP of 0.7-0.8, whereas Tsimane' has not reached this rate even by 72 months). However, assessing these differences statistically is challenging in view of marked differences in sample size (only 3 Quechua learners) and limited variability in age (all French learners are 11-13 months). Therefore, we next concentrated on two dimensions allowing us to pool across corpora, to study potential group differences as a function of language background and community, while controlling for age differences by subsetting to age ranges present in all levels of the factor of interest (1-40mos for syllabic complexity; 6-20mos rural vs urban), and including age in main and interaction terms.

To study the effect of syllabic complexity, we fit the following regression:

$$(3)\quad CP \sim age^2_{child} * syllabic\_complexity + age_{child} * syllabic\_complexity$$

We found a significant main effect of syllabic complexity on CP ($\chi^2(2) = 96.19$, $p < 0.001$), and in interaction with both age ($\chi^2(2) = 169.73$, $p < 0.001$) and age^2 ($\chi^2(2) = 22.72$, $p < 0.001$), meaning that both the rates of learning and plateauing differed by whether children were learning a language with low (i.e., those that only allow consonant-vowel syllables; Solomon, Yélî Dnye), moderate (i.e., those that allow some consonant clusters and/or codas, but nonetheless have substantial restrictions on syllable types; Tsimane', Quechua), or high syllabic complexity (English, French, Tseltal). That being said, the fitted models are visually similar for low/high syllabic complexity languages.

Finally, we fit an analogous regression to (3) to test the relationship between community (urban vs. rural) and CP. We found a significant main effect of community on CP ($\chi^2(1) = 32.05$, $p < 0.001$), but no interaction effects with age ($\chi^2(1) = 0.87$, p=.35) or age^2 ($\chi^2(1) = 0.52$, p=.47), suggesting that children raised in rural environments may have higher CPs than those raised in urban environments, but show similar developmental patterns (i.e., slopes and plateaus).

4. DISCUSSION

The results of this study have broadened our understanding of canonical vocalization development and helped to map its trajectory in ecological settings, across a wider age range, and across diverse populations.

Interestingly, we found that CP continues to increase well into ages where children's vocalizations are thought to be driven by communication goals rather than vocal development. That is, even once children are producing words, which should dictate the consonant-vowel combinations they produce, we continue to see increases in CP. There are three, mutually compatible, possible explanations for this continued increase, which future work should disentangle. First, there could be a pragmatic explanation: as children mature, they begin inhibiting less advanced vocalizations that are not conversationally appropriate, i.e., meaningless non-canonical productions. However, it is unclear how this predicts development beyond the inhibition of babble, and specifically for the increase of canonical transitions. The second explanation is that vocal and/or phonological development is still ongoing. In particular, young children tend to simplify their syllables by e.g., dropping codas; the continued increase in CP could reflect a tendency to do this less and less over time, perhaps as children are able to automate articulatory processes (a speech motor development account), or learn the phonological importance of onsets/codas in their language (a phonological development account).

Finally, the continued increase could arise because children begin speaking faster or producing longer vocalizations, both of which could potentially lead to higher canonical rates in our methodological approach based on 500ms-long chunks. Future work should further validate this method across communities and qualitatively evaluate children's vocalizations to better understand what explains changes in CP across early childhood. Another open question for future work is what CPs are for adults, which would allow us to determine when children have reached adult-like status (there are non-canonical vocalizations, e.g., "yeah" and "hmm", that are appropriate at any age, so CP will likely not stabilize at 100%).

We also found preliminary evidence that CP may not develop universally across children, but may vary across linguistic and/or community settings. However, there were a couple of confounds and theoretical issues that make it premature to draw strong conclusions about what precisely drives these findings. For example, all of the children raised in urban environments were learning phonologically complex languages, and they were all younger than 20 months. Moreover, although our dataset is larger and more diverse than previous approaches, we still only included a couple of languages per syllabic complexity type. Finally, while we observed significant effects of syllabic complexity, the results seemed to be driven by languages with moderate syllabic complexity patterning differently from languages with low and high syllabic complexity (which patterned together), a surprising finding according to the hypothesis that syllabic complexity systematically relates to CP.

That said, while this set of results is ultimately inconclusive, they do raise the possibility that CP may not develop universally across children, which merits further dedicated study. To this end, we have submitted a registered report to undertake a confirmatory study on a larger, more diverse, and more balanced sample. This will also allow us to undertake a more nuanced analysis of the results to consider how other linguistic/phonological factors (e.g., word length, stress), which co-vary with syllabic complexity, may influence CP development.

Overall, these results provide some evidence that canonical vocalization development may be more protracted and potentially more variable across languages and/or settings than previously thought, which should be accounted for when using this measure, saliently in the context of potential applications. CP as studied here is promising because it can be calculated from naturalistic audio recordings collected from highly diverse communities, easily and without onerous and time-intensive transcriptions, and continues developing throughout early childhood. Future work could study its predictive and diagnostic value.

In all, our study shows the promise of coarse, semi-automated approaches towards studying early speech development. Combining long-form recordings, automated approaches, and citizen science approaches, we were able to study a large sample of children from around the world and broaden our understanding of language acquisition.

[1] Except for three of the corpora (Solomon, Yélî Dnye, Tseltal), where a subset of the corpora were labeled by trained, in-lab research assistants.

[2] Citizen scientists were provided with explanations and examples of each category, which they could access at any time and could contact the research team with any clarifications.

5. ACKNOWLEDGMENTS

This work was funded by the Agence Nationale de la Recherche, the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award, and the European Research Council under the European Union's Horizon 2020 research and innovation programme. We thank Nadège Marin, Natalia Kuzminykh, and the Language Acquisition Across Cultures team for their assistance in preparing the data.

6. REFERENCES

[1] D. K. Oller, *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2] A. Nyman, S. Strömbergsson, and A. Lohmander, "Canonical babbling ratio–concurrent and predictive evaluation of the 0.15 criterion," *Journal of Communication Disorders*, vol. 94, p. 106164, 2021.

[3] M. Cychosz, A. Cristia, E. Bergelson, M. Casillas, G. Baudet, A. S. Warlaumont, C. Scaff, L. Yankowitz, and A. Seidl, "Vocal development in a large-scale crosslinguistic corpus," *Developmental Science*, vol. 24, no. 5, p. e13090, 2021.

[4] C. Semenzin, L. Hamrick, A. Seidl, B. L. Kelleher, and A. Cristia, "Describing vocalizations in young children: A big data approach through citizen science annotation," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2401–2416, 2021.

[5] B. de Boysson-Bardies and M. M. Vihman, "Adaptation to language: Evidence from babbling and first words in four languages," *Language*, vol. 67, no. 2, pp. 297–319, 1991.

[6] M. M. Vihman and B. de Boysson-Bardies, "The nature and origins of ambient language influence on infant vocal production and early words," *Phonetica*, vol. 51, no. 1-3, pp. 159–169, 1994.

[7] A. S. Warlaumont, J. A. Richards, J. Gilkerson, and D. K. Oller, "A social feedback loop for speech development and its reduction in autism," *Psychological Science*, vol. 25, no. 7, pp. 1314–1324, 2014.

[8] A. Cristia, "A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations," *Developmental Science*, p. e13265, 2022.

[9] E. Bergelson, "Bergelson SEEDLings Homebank Corpus." [Online]. Available: https://doi.org/10.21415/T5PK6D

[10] E. Bergelson and R. N. Aslin, "Nature and origins of the lexicon in 6-mo-olds," *Proceedings of the National Academy of Sciences*, vol. 114, no. 49, pp. 12 916–12 921, 2017.

[11] E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, and S. Tor, "Day by day, hour by hour: Naturalistic language input to infants," *Developmental Science*, vol. 22, no. 1, p. e12715, 2019.

[12] A. Cristia, "PhonSES: A pilot study to measure socioeconomic status association with infants' word and sound processing," 2021. [Online]. Available: https://gin.g-node.org/LAAC-LSCP/phonSES-public

[13] M. Cychosz, "Cychosz Homebank Corpus," 2018. [Online]. Available: doi:10.21415/YFYW-HE74

[14] M. Casillas, P. Brown, and S. C. Levinson, "Casillas Homebank Corpus," 2017. [Online]. Available: https://doi.org/doi:10.21415/T51X12

[15] ——, "Early language experience in a Tseltal Mayan village," *Child Development*, vol. 91, no. 5, pp. 1819–1835, 2020.

[16] C. Scaff, J. Stieglitz, and A. Cristia, "Tsimane' daylong recordings collected with LENA in 2017-2018," 2018. [Online]. Available: https://doi.org/DOI10.17605/OSF.IO/6NEZA

[17] ——, "Excerpts from daylong recordings of young children learning Tsimane' in Bolivia," 2019. [Online]. Available: https://doi.org/DOI10.17605/OSF.IO/5869Q

[18] A. Cristia and M. Casillas, "LENA recordings in Rossel Island," 2019.

[19] M. Casillas, P. Brown, and S. C. Levinson, "Early language experience in a Papuan community," *Journal of Child Language*, vol. 48, no. 4, pp. 792–814, 2021.

[20] A. Cassar, A. Cristia, P. Grosjean, and S. Walker, "Long-form recordings in the Solomon Islands," 2021.

[21] M. Lavechin, M. de Seyssel, L. Gautheron, E. Dupoux, and A. Cristia, "Reverse engineering language acquisition with child-centered long-form recordings," *Annual Review of Linguistics*, vol. 8, pp. 389–407, 2022.

[22] H. Ganek and A. Eriks-Brophy, "Language ENvironment Analysis (LENA) system investigation of day long recordings in children: A literature review," *Journal of Communication Disorders*, vol. 72, pp. 77–85, 2018.

[23] A. Cristia, F. Bulgarelli, and E. Bergelson, "Accuracy of the language environment analysis system segmentation and metrics: A systematic review," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 4, pp. 1093–1105, 2020.

[24] A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson, "A thorough evaluation of the Language ENvironment Analysis (LENA) system," *Behavior Research Methods*, vol. 53, no. 2, pp. 467–486, 2021.

[25] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," *arXiv preprint arXiv:2005.12656*, 2020.

THE ACQUISITION OF HUNGARIAN VOWELS DIFFERING IN PHONEMIC VOWEL LENGTH

Krisztina Zajdó

Department of Special Education/Speech-Language Therapy,
Széchenyi István University/The University of Győr, Győr, Hungary
zajdo@hotmail.com

ABSTRACT

This study examined spectral properties of five Hungarian vowel pairs with contrasting phonemic vowel lengths in 2;0 and 4;0 years old boys acquiring Hungarian as their native language. Results were obtained by an automated pitch-synchronous bandfilter analysis method that estimates the spectral envelope representation of vowels. Subsequent data reduction was achieved via principal component analysis.

Findings show that 4;0 years olds produce more adult-like vowels due to an increased ability to formulate more peripheral and more rounded vowels. As a result, the vowel space of older children contains more differentiated vowel categories, with more compact vowel spaces corresponding to each vowel. The spectral differentiation of short vs. long vowel pairs is more adult-like in 4;0 years olds than in younger children, but the acquisition of spectrally differentiated vowels with contrasting vowel length is not yet complete at this age.

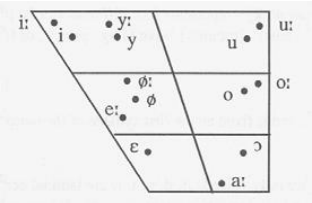
Keywords: vowels, acquisition, children, Hungarian, phonemic vowel length, spectral differentiation.

1. INTRODUCTION

Many languages code differences in meaning by using contrasting phonemic vowel length. In the UPSID317 database, 19.6% of the languages examined apply this differentiation to allow for semantic identification [1]. The standard dialect of Hungarian, a Finno-Ugrian language uses a two-way phonemic vowel length contrast that is characterized by an average duration ratio of 1:2 [2], [3]. Standard Hungarian uses phonemic vowel length as a primary feature to differentiate five of its seven vowel pairs: /i/ vs. /i:/, /o/ vs. /o:/, /ø/ vs. /ø:/, /u/ vs. /u:/ and /y/ vs. /y:/. Further, the members of these vowel pairs are also differentiated by a secondary distinguishing feature of spectral contrast. (Members of the remaining two vowel pairs, /ɔ/ vs. /a:/ and /ɛ/ vs. /e:/, are contrasted by both durational and substantial spectral differences; for the acoustic representation of the Standard Hungarian vowel space, see Fig. 1.)

This study examines the acquisition of spectral differences in the five Hungarian vowel pairs that are primarily differentiated by durational cues and are characterized by subtle spectral differences. The objective was to map out developmental trends in the acquisition of spectral differentiation between phonemically short vs. long vowels.

Figure 1. The vowel space of Standard Hungarian (modified from Szende [4]).



2. METHODOLOGY

This study applied a cross-sectional study design. The study is part of a larger research project that explores vowel acquisition processes in children cross-linguistically; for details of setup and procedures, see [5], [6], [7], [8].

2.1. Participants

Participants included 5-5 boys at the age of 2;0 years and 4;0 years (± 2 weeks from birthday). The children and their caregivers were monolingual native speakers of the standard dialect of Hungarian from the Budapest area, without substantial exposure to other dialects or languages. All children had an uneventful health history.

2.2. Experimental setup and procedures

Participants and their caregivers were instructed to play with stuffed animals in a sound attenuated room (~ 6 m²). Toys had pre-assigned CV(:)1CV(:)1 structured names sawn on their back to help the caregiver recall them. Caregivers attempted to elicit the production of tokens from each child at least five times during a naturalistic interaction. Animal names included /pipi/ vs. /pi:pi:/, /lolo/ vs. /lo:lo:/, /gøgø/vs. /kø:kø:/, /bubu/ vs. /pu:pu:/, and /nyny/ vs. /my:my:/. Tokens with short vs. long target vowels were recorded during two separate sessions that were scheduled for no more than two weeks apart.

2.3. Recording procedure and data selection

The Sound Forge acoustic software (Version 5.0, Built 117) was used to record the speech samples (recording attributes: 32 kHz, 16 bit, mono). Data selection was carried out by a monolingual Hungarian speaker. Only tokens that were produced in relative proximity to the elicitation by caregivers containing two correctly produced vowels were selected perceptually. A second judge rated 10% of the dataset for reliability. Inter-rater agreement for each vowel category exceeded 90%.

2.4. Data analysis

Data analysis was carried out by using an automated pitch-related bandfilter analysis method that allows for an acoustically unbiased measurement of formant measures. The method estimates the spectral envelope representations of vowels. All selection and analysis procedures were carried out by scripts in the Praat acoustic software program [9].

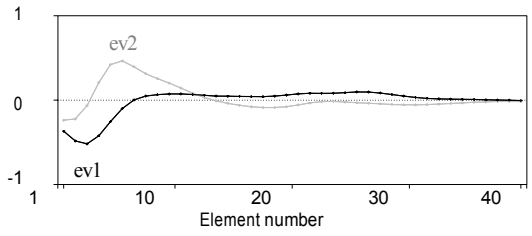
Individual CV(:)$_1$CV(:)$_1$ productions were analyzed as unlabeled tokens. Specified intensity criteria were used to avoid clipped segments and low intensity fragments. To bandfilter the selected part, the F_0 period nearest to the selection was recirculated to generate a sufficiently long signal. Signal filtering was carried out by the application of a Gauss bandfilter, tuned to 0-7kHz in 40 steps of 175 Hz. Bandwidth was set to 1.1 x 425 Hz. The resulting spectral envelope contains spectral intensities of 40 adjacent frequency bins of 175 Hz width. Each measurement represents a point in a 40 dimensional space.

Principal component analysis was used to reduce the data set. This approach has been used successfully to generate vowel production charts in both adults [10], and children [7], [8], [9], [11]. A new space was then generated, with pc1 representing the component with the largest variance and pc2 representing the component with the second largest variance, etc. The first two principal components created a vowel space with pc1 showing resemblance to F2 and pc2 showing resemblance to F1 (see Figures 3A-B to 7A-B in the attachment file).

3. RESULTS

To compare vowel categories in 2;0 and 4;0 years old children, a pc1-pc2 reference plane was created by using data from all 4;0 years old boys. Vowel spectra measures from 687 CV(:)$_1$CV(:)$_1$ tokens that included each of the 14 Hungarian monophthongs from all 5 boys were analyzed to create a pc1-pc2 plane. The plane, which is defined by a set of eigenvectors, includes 4372 measurement points indicated by grey plus signs in the figures. The first eigenvector explains 26.2% of variance, while the second one explains 18.9% of variance (see Figure 2.). Thus, the cumulative percentage of variance explained by the first two eigenvectors is 45.1%. To determine the location of the corner vowels within the vowel space, labels containing the vowels /i(:)/, /u(:)/ and /ɔ/ were analyzed and projected onto the reference plane.

Figure 2. The first two eigenvectors of the Standard Hungarian reference space/plane.



Children's labeled tokens were analyzed according to the target token. To determine the location and dispersion of each vowel category within the reference plane, a randomly selected 60 measurement points from each category were projected onto the pc1-pc2 reference plane. An ellipse was created for each category that represents 1 SD from the mean of the measured variance. Results are shown in Figures 3A-B to 7A-B; "A" figures contain data from 2;0 year-olds; "B" figures are from 4;0 year-olds.

4. DISCUSSION

A number of production characteristics and developmental tendencies are shown by the data. Results are reviewed by comparing the acoustic properties of individual vowel categories both within and between age groups.

4.1. /i/ vs. /i:/

At 2;0 years, boys produce quite centralized vowels. However, phonemically short vowels are more centralized, indicating less stability of production patterns in comparison to phonemically long vowels. In contrast with the long vowels, the majority of short vowels are realized with lower jaw position. Ellipses areas (3.95 vs. 2.59, resp.) show that, during long vowel production children hit a more concentrated target area of the vowel space more consistently. Children have more time to position their articulators to the required position in these cases, and long vowels are thus more adult-like. See Fig. 3A.

At 4;0 years, both vowels are realized with a higher jaw position. Centralization tendencies are diminished. The areas of the ellipses (1.69 vs. 2.31) indicate more consistency in the production of the short vowel. It is reasonable to speculate that long vowel duration results in changing vowel spectra, reflected by a more sizeable ellipse showing 1SD from the mean of measured variance. See Fig. 3B.

Developmentally, vowels produced by older children are more adult like, even though producing the vowels in a front position is still a challenge. The corner vowel position is being hit less frequently, showing difficulties with fronting.

4.2. /o/ vs. /o:/

At 2;0 years, boys realize mid vowels in a low position, indicating an inability to round lips. The majority of short vowels are realized with a lower jaw position. Phonemically long vowels are more centralized, indicating less stability of production patterns in comparison to phonemically short vowels and prominent error patterns (e.g., o > ɔ, o > ɛ). The area of the ellipses (5.93 vs. 4.87, resp.) indicate that, during the production of the long vowels children hit a less concentrated target area of the vowel space with more consistency. However, while producing long vowels, children have difficulties with trying to position and keep their lips rounded, resulting in unstable (changing) production characteristics. Neither the short nor the long vowels are adult-like. See Fig. 4A.

At the age of 4;0, boys produce mid vowels by covering a more sizeable area of the vowel space, as shown by the area of the ellipses (6.41 vs. 5.46, resp.). Many long vowels are produced in more peripheral position, indicating that during a longer duration boys hit

the target area of the vowel space more often. Producing long vowels with higher jaw positioning is a property of the adult vowel system, making production patterns more adult-like in older children. See Fig. 4B.

Developmentally, older children produce more adult-like vowels, due to an ability to a) produce the vowels with higher jaw position, and b) spectrally differentiate the short vs. long vowels. Overall, the production of the vowels in 4;0 years old boys is not yet adult-like, due to an inability to produce all vowels with a back tongue position more consistently.

4.3. /ø/ vs. /ø :/

At 2;0 years of age, boys realize front mid vowels in centralized position, indicating not yet fully developed skills to round lips. The majority of long vowels are realized with a lower jaw position. Prominent error patterns (e.g., ø(:) > o, ø(:) > ɔ, ø(:) > ɛ) show difficulties with a) lip rounding and b) coordinating lip rounding with front tongue position. The areas of the ellipses (6.41 vs. 5.46, resp.) indicate that, during the production of the long vowels children hit a less concentrated target area of the vowel space with less consistency. That is, while producing long vowels, children have difficulties with trying to position their lips into a rounded position, resulting in a unstable (changing) production characteristics. Neither the short nor the long vowels are adult-like. See Fig. 5A.

At 4;0 years, many short vowels are produced in a considerable more frontal position, while long vowels are typically produced in a central position. Additionally, the majority of vowels are produced with a higher jaw position. The areas of the ellipses (4.91 vs. 4.78, resp.) show similar consistency for both short and long vowels in hitting a more restricted area of the vowel space. Higher consistency in aiming at a smaller acoustic sub-space indicates more stable speech production patterns, including a higher ability to round lips. See Fig. 5B.

Developmentally, older children's vowels are more adult-like, due to an ability to produce vowels with higher jaw position and more prominent lip rounding. Further, the spectral differentiation of short vs. long vowels (whereas long vowels are produced with higher jaw position) makes older children's vowels more adult-like. However, the production of long vowels with a less (rather than more) frontal tongue position shows that the differentiation of the two vowels is not carried out in a completely adult-like fashion in 4;0 years olds.

4.4. /u/ vs. /u:/

At 2;0 years, boys realize many high back rounded vowels in a mid back position, clearly showing a non-adult-like ability to coordinate lip rounding with high jaw positioning during vowel production. The majority of long vowels are realized with a lower jaw position, indicating instability of production patterns and a prominent error pattern (u(:) > o). The areas of the ellipses (5.24 vs. 6.50, resp.) show that the longer vowel duration results in more varied spectral characteristics. Clearly, formulating and sustaining lip rounding for a longer duration is a challenge at this age. See Fig. 6A.

At 4;0 years, boys produce the short vowels in a high back position, reflecting relatively more mature skills in lip rounding (ellipse areas are 5.60 vs. 3.70, resp.). Similarly to the 2;0 year-olds, older children are less successful with hitting a more compact area of the vowel space when producing vowels with longer duration. Longer duration appears to be associated with less stable vowel production patterns, indicating non- adult-like abilities to maintain rounding. See Fig. 6B.

Developmentally, older children's vowels are more adult-like, due to an improved ability to round the lips and more success with producing these vowels with a high jaw position. However, spectral differentiation of the members of the vowel pair does not appear adult- like since the long vowels are produced with a lower jaw position and more frontal tongue position.

4.5. /y/ vs. /y:/

At 2;0 years of age, boys appear to centralize both the short and long high front vowels. Even though Hungarian does not have a phonemic schwa sound, the acoustic quality of these vowels is similar to that of a reduced vowel. The areas of the ellipse (4.48 vs. 4.31, resp.) indicate similar production patterns for both vowels. Spectral differentiation is minimal. See Fig. 7A.

At 4;0 years, the majority of vowels are produced in a high central position, indicating strong error patterns (e.g., y(:) > u, y(:) > i). Long vowels have a stronger tendency to be substituted by /u/, showing the challenging nature of coordinating more frontal tongue positioning with lip rounding. The difference in the areas of the ellipses (2.94 vs. 3.83) clearly shows the prominence of this error pattern. See Fig. 7B.

Developmentally, these vowels reflect the most progress towards adult-like vowel production in boys between 2;0 and 4;0 years of age. However, the spectral differentiation of vowels with contrasting phonemic vowel length is not yet adult-like in 4;0 years old boys: the more posterior production of long vowels show vowel production mechanisms in older children that are not yet mature.

5. SUMMARY

Examining the spectral differentiation of vowels with contrasting phonemic vowel length is important when formulating a theory of vowel acquisition since the realization of these spectral differences show the extent of the acquisition in children. Results show that the vowel acquisition process is clearly not completed in 4;0 years old boys acquiring the standard dialect of Hungarian, as reflected by non-adult-like spectral patterns in children's vowels differing in phonemic vowel length.

Overall, vowels produced by 4;0 years old boys are closer to the adult model, due to 1) an increased ability to produce vowels with higher (more closed) jaw position, 2) increased skills for formulating lip rounding, 3) an ability to maintain stable lip rounding for longer durations, and 4) skills to coordinate lip rounding with higher and more frontal tongue positions. In general, the more compact ellipse areas of individual vowel categories in older children are indicative of higher level speech skills used during vowel production.

Figure 3A

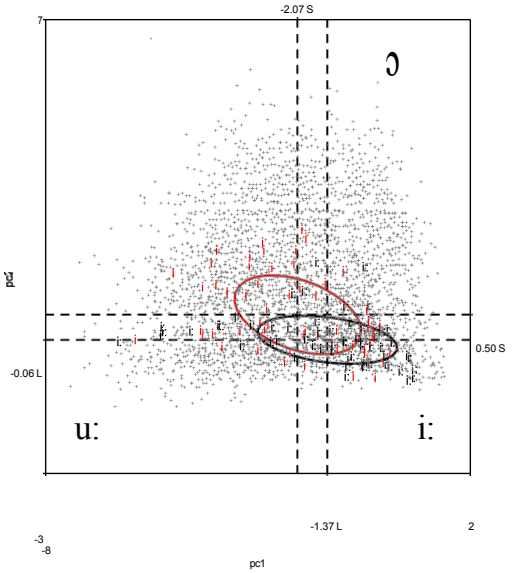


Figure 3B

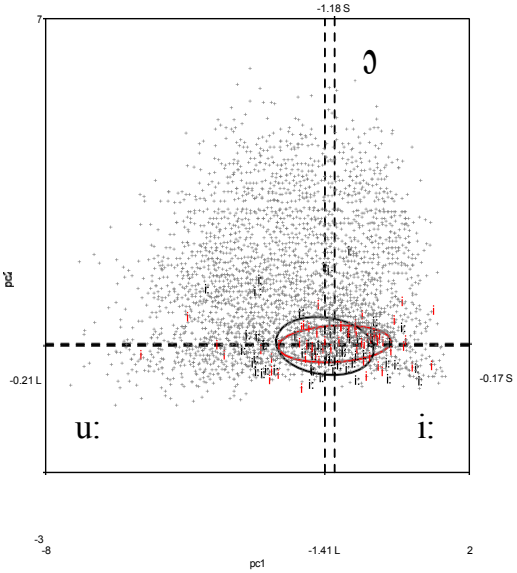


Figure 4A

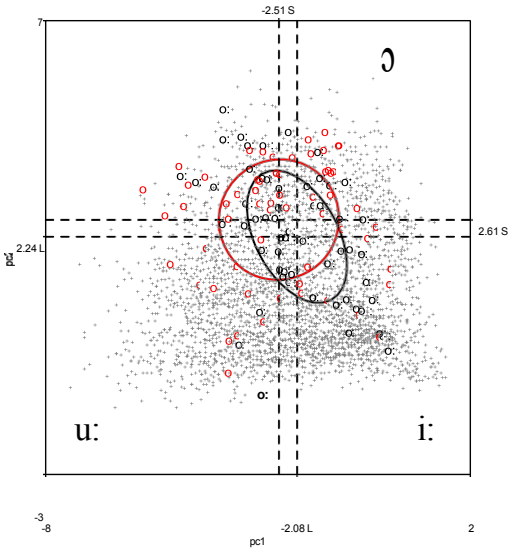


Figure 4B

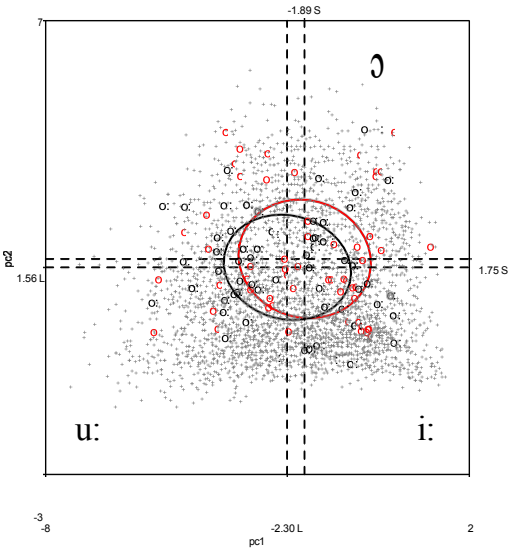


Figure 5A

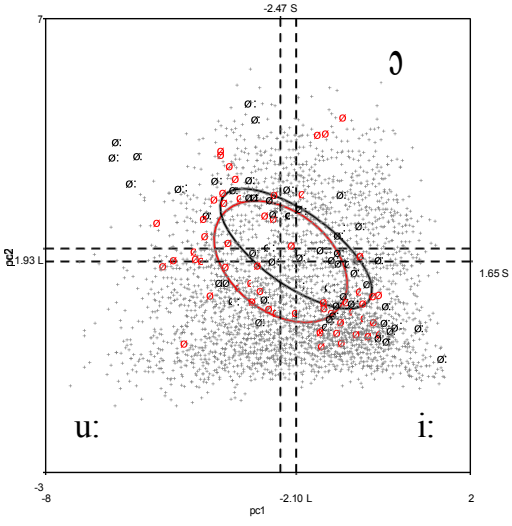
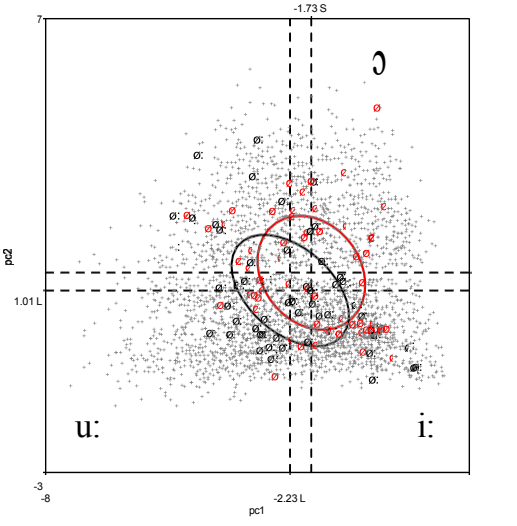


Figure 5B



Vowels of 2;0 years olds (3A: /i/ vs. /i:/, 4A: /o/ vs. /o:/, 5A: /ø/ vs. /ø:/) projected onto the reference plane. Each ellipse represents 1 SD from the mean of measured variance. Red/grey: short; Black: long; S: short; L: long vowels.

Vowels of 4;0 years olds (3B: /i/ vs. /i:/, 4B: /o/ vs. /o:/, 5B: /ø/ vs. /ø:/) projected onto the reference plane. Each ellipse represents 1 SD from the mean of measured variance. Red/grey: short; Black: long; S: short vowels; L: long vowels.

Figure 6A

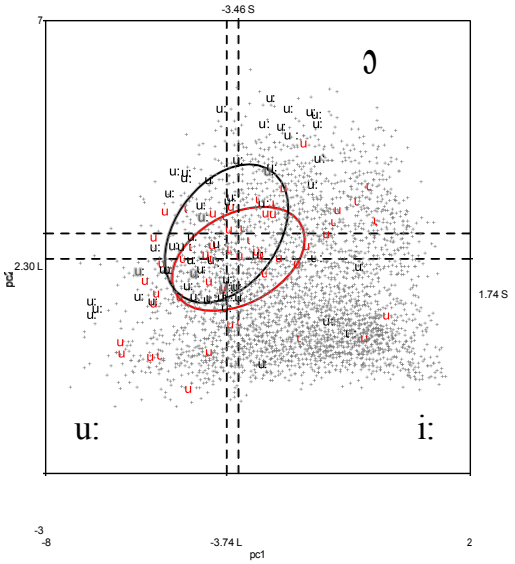


Figure 6B

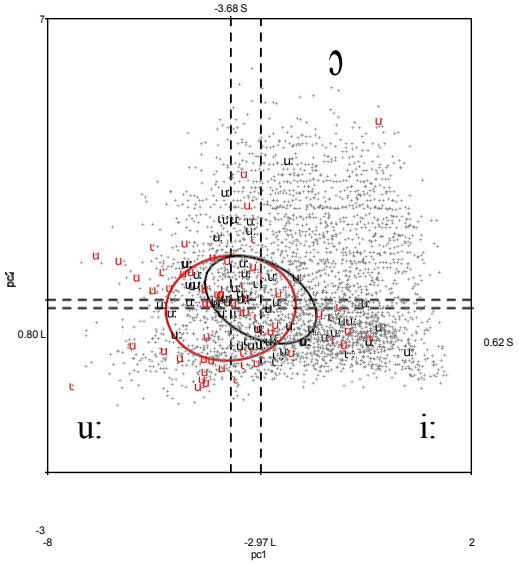


Figure 7A

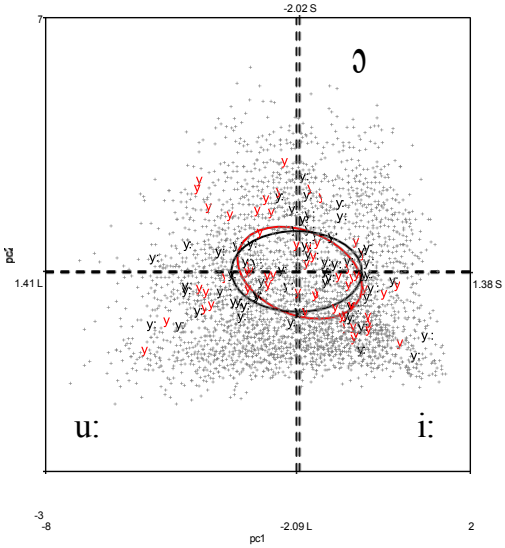
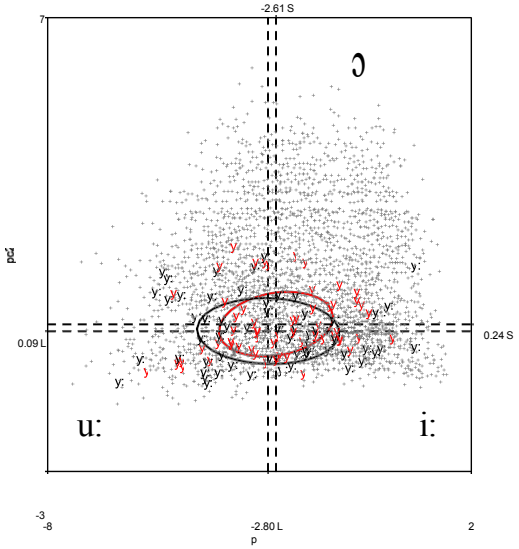


Figure 7B



Vowels of 2;0 years olds (6A: /u/ vs. /u:/, 7A: /y/ vs. /y:/) projected onto the reference plane. Each ellipse represents 1 SD from the mean of measured variance. Red/grey: short vowels; Black: long vowels. S: Short vowels, L: long vowels.

Vowels of 4;0 years olds (6B: /u/ vs. /u:/, 7B: /y/ vs./y:/) projected onto the reference plane. Each ellipse represents 1 SD from the mean of measured variance. Red/grey: short vowels; Black: long vowels. S: short vowels, L: long vowels.

6. REFERENCES

[1] Maddesion, I. 1984. *Patterns of sounds*. Cambridge: CUP.

[2] Tarnóczy, T. 1974. A magánhangzók akusztikai vizsgálatának problémái. *Általános Nyelvészeti Tanulmányok,* 10, 181-196.

[3] Kassai, I. 1979. *Időtartam és kvantitás a magyar nyelvben*. Nyelvtudományi Értekezések, 102. Budapest: Akadémiai.

[4] Szende, T. 1999. Hungarian. In The *Handbook of the International Phonetic Association*. Cambridge, UK: CUP, 104–107.

[5] Zajdó, K., Van der Stelt, J.M., Wempe, T., Pols, L.C.W. 2005. Cross-linguistic comparison of two-year-old children's acoustic vowel spaces: Contrasting Hungarian with Dutch. *Proc. InterSpeech 2005*, Lisbon, 1173-1176.

[6] Van der Stelt, J.M., Zajdó, K., Wempe, T.G. 2005. Investigating the acoustic vowel space in two-year-old children: Results for Dutch and Hungarian. *Speech Communication*, 47, 143-159.

[7] Pols, L.C.W., Lyakso, E., van der Stelt, J.M., Wempe, T., & Zajdó, K. 2006. Vowel data of early speech development in several languages. *Proc. MultiLing 2006*.

[8] Pols, L.C.W. 1977. *Spectral analysis and identification of Dutch vowels in monosyllabic words*. PhD dissertation, Free University of Amsterdam, The Netherlands.

[9] Boersma, P., Weenink, D. 1996. *Praat: A system for doing phonetics by computer*. Inst. Phon. Sci. Univ. of Amsterdam, Report no. 132.

[10] Zajdó, K. 2006. Patterns of vowel space utilization in Hungarian caregiverese addressed to young children: An evaluation of the MIPhI model. In H.C. Yehia, D. Demolin & R. Laboissière (Eds.), *Proc. of the 7th Int. Seminar on Speech Production* (pp. 99-106). Belo Horizonte MG, Brasil: Center for Research on Speech, Acoustics, Language and Music.

[11] Zajdó, K. 2002. *The acquisition of vowels by Hungarian-speaking children aged two to four years: A cross-sectional study*. Unpublished doctoral dissertation, University of Washington, Seattle, WA, USA.

PROSODY OF PRE-FOCAL BACKGROUND DEPENDS ON FOLLOWING FOCUS

Simon Roessig

Department of Linguistics, Cornell University
mail@simonroessig.de

ABSTRACT

This study is concerned with the realization of focus structure by prosodic means in German. Results of a production study are presented that support previous findings about differences between focus types. Crucially, however, the realization of given (background) words in the pre-focal region also depends on the focus type that follows. The data are analyzed with classic scalar measures (F0 maximum and syllable duration) as well as by assessing the evolution of the F0 trajectories over time in the intervals of interest. The two analyses show converging results: Before corrective focus, pre-nuclear words are realized with lower F0 peaks, larger F0 excursions, and shorter duration than before non-corrective narrow focus. In the focal domain, the picture is almost reversed with higher F0 peaks and larger F0 excursions. These results suggest that nuclear *and* pre-nuclear accents contribute to the interpretation of the focus structure of the sentence.

Keywords: prosody, information structure, pre-nuclear, F0, duration

1. INTRODUCTION

In West-Germanic languages, focus is expressed by a prominence on the focused word (or focus exponent) [1], [2] – more specifically by the placement of the nuclear pitch accent, and post-focal deaccentuation or compression [3], [4]. In addition, the realization pattern of the accented syllable is correlated with the *type* of focus. Accented syllables in narrow and corrective focus are realized with higher F0 targets [5], longer durations [6] and more extended articulatory movements [7] compared to their counterparts in broad focus. More recently, it has been shown that there are even differences between narrow and corrective focus [6], [8], [9], although these differences appear to be subtler.

While post-focal deaccentuation and the modulation of nuclear accents in the focal region of a sentence as a function of focus types are well-documented, fewer studies have looked at the pre-focal region. The predominant view is that, as [10] suggested, pre-nuclear accents are optional ("ornamental"), especially when pre-focal. Similarly, [11] posited that pre-nuclear accents normally do not mark information structure but rather are correlates of the metrical structure of the sentence. This view is supported by results from a perceptual learning experiment presented by [12] with English speaking individuals: While younger children pay equal attention to the beginning *and* end of intonation contours, older children and adults seem to have internalized that the end contains the most important information and hence pay more attention to the later part of the contour.

Results from other studies are at odds with the idea that pre-nuclear accents are unaffected by information structure. For German, [5] showed that pre-nuclear accents exhibit lower F0 when they appear pre-focally before a narrow focus compared to pitch accents in the same position in a broad focus sentence (in which case they are part of the focus). Similar results were obtained for Bulgarian by [13]. These findings from production are in line with perception results of [14] for Dutch in which participants judged the excursion of two pitch peaks (pre-nuclear and nuclear) in relation to a focus structure. Compared to broad focus, a smaller pre-nuclear peak was judged as optimal for words preceding corrective focus. The results for German presented in [6] indicate that words in pre-focal position preceding corrective focus are realized with longer durations compared to the same words in broad focus sentences. These studies demonstrate that the pre-nuclear domain can indeed be affected by information structure.

[9], also for German, used speech material that includes narrow and corrective focus. Interestingly, there seemed to be a lower probability for the placement of a pre-nuclear accent before corrective focus than before narrow focus – although in both cases the pre-nuclear domain is given, i.e., pre-focal. This finding suggests that the pre-focal region may actually contain information about the focal region. It raises the question as to how distributed information structure is across the phrase and whether pre-focal elements can contribute to the marking of information structure of the following focal element.

The present study concentrates on this question. In doing so, it takes both the pre-focal and the focal region into account and compares the prosodic patterns of the two regions in sentences with narrow

and corrective focus. A corpus of approximately 1000 productions from 27 speakers is analyzed regarding the realization of F0 and syllable duration. The prosodic patterns are investigated in terms of simple scalar measurements, namely F0 maximum in the word and stressed syllable duration, as well as in terms of the evolution of F0 trajectories over time using generalized additive mixed models (GAMM) – with converging results. The results of the paper contribute to our understanding of the marking of focus structure in German and the dispersion of information in the acoustic signal.

2. METHODS

2.1. Speech material

The speech material analyzed in this study consists of sentences produced with two different focus structures. In order to elicit these focus conditions, question-answer pairs were used. The answers are the analyzed target sentences and were always of the form *Er hat den/die <A> auf die <B> gelegt* ('He put the <A> on the <B>') with two nouns A and B. The questions served as triggers for the focus structure of the answer. In both focus conditions, word A is given and occurs pre-focally, i.e., in the background. The difference lies in the focus type of the focal word B: In the first focus condition, *background-narrow*, word B is in narrow focus. In the second focus condition, *background-corrective,* word B is in corrective focus. The question to elicit *background-narrow* followed the scheme *Wo hat er den/die <A> hingelegt?* ('Where did he put the <A>?'); the question to elicit *background-corrective* followed the scheme *Hat er den/die <A> auf die <C> gelegt?* ('Did he put the <A> on the <C>?') where C is a contrasting alternative referent. Table 1 illustrates the focus conditions with examples. Square brackets and subscript F indicate the focused elements. The full data set comprises two additional focus conditions, *corrective-background* and *broad-broad*, that are not reported in the current paper.

background-narrow	
Question:	*Wo hat er den Hammer hingelegt?* 'Where did he put the hammer?'
Answer:	*Er hat den Hammer [auf die Wohse]$_F$ gelegt.* 'He put the hammer on the Wohse.'
background-corrective	
Question:	*Hat er den Hammer auf die Mahse gelegt?* 'Did he put the hammer on the Mahse?'
Answer:	*Er hat den Hammer auf die [Wohse]$_F$ gelegt.* 'He put the hammer on the Wohse.'

Table 1: Focus conditions with question-answer pairs

As targets for the pre-focal word (A), ten German disyllabic nouns denoting common tools with stress on the first syllable were used: *Amboss* ('anvil'), *Besen* ('broom'), *Bohrer* ('drill'), *Bürste* ('scrub brush'), *Hammer* ('hammer'), *Pinsel* ('paint brush'), *Rolle* ('paint roller'), *Säge* ('saw'), *Schere* ('scissors'), and *Zange* ('pliers').

As targets for the focal word (B), twenty German sounding disyllabic nonce words with a $C_1V_1C_2V_2$ structure were created. All nonce words had stress on the first syllable. C_1 was chosen from the set of {/n/, /m/, /b/, /l/, /v/}, V_1 from {/a/, /o/}, and C_2 from {/n/, /m/, /z/, /l/, /v/}. V_2 was always Schwa. Examples for target word B are *Nahne*, *Mohme*, and *Bahle*.

2.2. Speakers and recordings

27 monolingual native speakers of German (19-35 yrs.; 17 female) were recorded. The subjects were prompted to produce the target utterances by involving them in an interactive game on a computer screen. In the game, their task was to help an animated robot retrieve tools. The robot's questions served as triggers for the focus structure of the answer. A training session with different target words preceded the actual recording session. The recordings were carried out at the University of Cologne using a head-mounted condenser microphone. In addition to the acoustic signal, the articulators' movements were recorded (EMA). This paper only deals with the acoustic data.

2.3. Annotations and measurements

The boundaries of the two target words (words A and B) and their stressed syllables were annotated. Additional segmental annotations were obtained from forced alignment using Kaldi [15] through the Montreal Forced Aligner [16]. Furthermore, the low boundary tone at the end of each sentence was labelled as the last reliable F0 point. Using these annotations, the following measurements were performed. First, the stressed syllable durations of both target words (pre-focal/A and focal/B) were measured. Second, F0 was calculated over the whole sentence using Praat [17] through the Python interface parselmouth [18]. For each speaker, the floor for the F0 calculation was set separately as the F0 value of the lowest L-% boundary tone of that speaker minus 10 Hz. From the F0 track, the maximum in each target word was obtained. Furthermore, time-normalized F0 was extracted in 49 equal time steps over the words in positions A and B, and in 149 equal time steps over the whole sentence. All F0 values in this analysis are expressed in semitones relative to the 5th percentile of the distribution of all L-% boundary tones of the speaker

that produced the utterance. Productions that had a clear phrase boundary between word A and word B were excluded to ensure that the pre-focal word was always pre-nuclear. After this exclusion, the data set comprised 991 recordings. The statistical analyses were carried out in R [19] using the libraries brms [20] for Bayesian regression, mgcv [21], itsadug [22] and tidymv [23] for fitting and visualizing GAMMs, as well as tidyverse [24] and zoo [25] for data processing and plotting. Data and code are available publicly on OSF: https://osf.io/2an3v/.

3. RESULTS

3.1. Analysis of F0 contours

The top panel of Figure 1 presents scatterplots of all F0 points with average contours superimposed. The bottom panel "zooms" into the F0 trajectories by giving the average contours for the pre-focal (left) and focal words (right). The average contours were obtained by averaging over all F0 measures of a point in normalized time. Only those time points for which more than 25% of measures existed entered the calculation of the average contour. The average contours are smoothed by applying a rolling mean with a window size of 3. Comparing the pre-focal and the focal regions, the peak relations are reversed: In the pre-focal region, *background-narrow* (blue) exhibits a higher peak than *background-corrective* (red), while the opposite is true in the focal region.

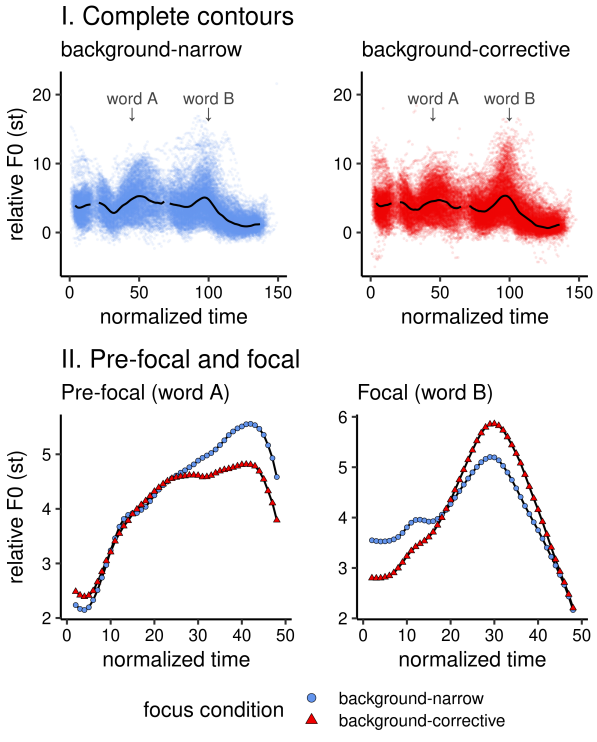


Figure 1: Average contours. Top: Complete sentence. Bottom: pre-focal (word A) and focal (word B) words.

The differences in the contours are assessed more formally by fitting a GAMM to each region – pre-focal and focal. For this analysis, the F0 trajectories are interpolated linearly. As fixed effects, the models include FOCUS CONDITION as a parametric term and smooths over TIME for FOCUS CONDITION. The models were fit such that the smooth over TIME for the condition *background-narrow* represents the reference smooth and the model contains a difference smooth for the condition *background-corrective* in relation to the reference smooth. In addition, random factor smooths per focus condition were included for the individual levels of SPEAKER and TARGET WORD. To account for autocorrelation of the residuals, an AR1 error model was incorporated. The smooths obtained from the GAMMs for the two focus conditions are visualized in the top panel of Figure 2.

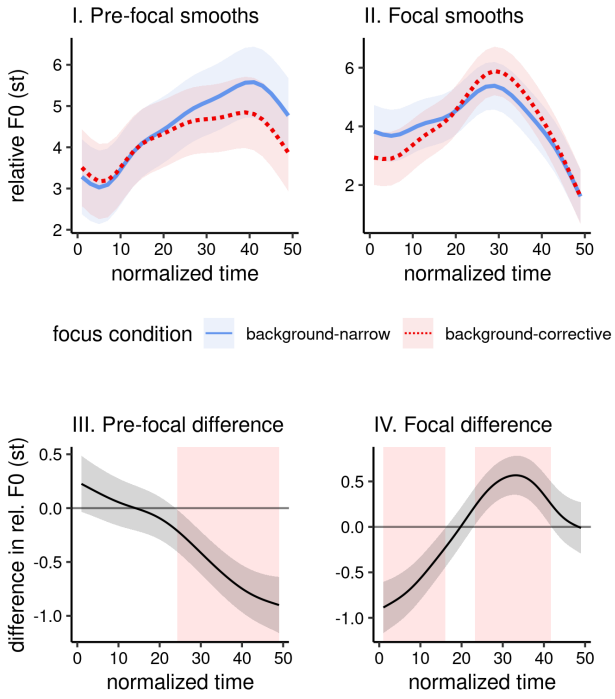


Figure 2: GAMM results for pre-focal (left) and focal regions (right). Top: smooths. Bottom: difference smooths between the two conditions.

Both models have significant difference smooths for *background-corrective*, indicating that the course of F0 over time in this condition is different from *background-narrow*. It should be emphasized that this is not only true for the focal region but also for the pre-focal region – although the information structure remains constant in this region (i.e., background). The difference plots in the bottom panel of Figure 2 visualize where and how *background-corrective* differs from *background-narrow*. In these plots, the red shaded areas indicate the regions of a significant difference. In the pre-focal region (Figure 2 bottom left), the second half of the contour over the

word in *background-corrective* takes a lower course (the difference is negative). In the focal region (Figure 2 bottom right), the start of the contour and the region of the peak are different: the contour of *background-corrective* starts lower (negative difference at the beginning) and reaches a higher peak (positive difference around the peak).

3.2. Analysis of F0 maximum and syllable duration

Figure 3 shows the means and standard errors for F0 maximum and stressed syllable duration. The pre-focal region (purple circles) exhibits a lower F0 maximum and shorter syllable duration in *background-corrective* than in *background-narrow.* The picture is again reversed in the focal region (green triangles).

Bayesian regression models were fit with either F0 MAXIMUM or SYLLABLE DURATION as dependent variable and FOCUS CONDITION as fixed effect. Random intercepts for SPEAKER and TARGET WORD were included as well as by-SPEAKER and by-TARGET WORD random slopes for FOCUS CONDITION. *background-narrow* is the reference level in the models (intercept). The regression coefficient β for *background-corrective* thus indicates the difference between the two focus conditions.

Table 2 lists the estimates β with their 90% credible interval (CI). For the pre-focal region, the table additionally gives Pr(β<0), the probability that β is negative. For the focal region, the table gives Pr(β>0), the probability that β is positive. In the pre-focal region, the models provide strong evidence for lower F0 maxima and shorter syllable durations in *background-corrective* than in *background-narrow* (the estimated β are negative). In the focal region, the models provide strong evidence for higher F0 maxima but not for longer syllable durations in *background-corrective*: here, Pr(β>0) is only 0.88.

Pre-focal (word A)			
Parameter	*β*	*90%-CI*	*Pr(β<0)*
F0 maximum	-0.71	[-0.93 -0.50]	1.00
Syllable duration	-6.96	[-11.01 -2.80]	1.00
Focal (word B)			
Parameter	*β*	*90%-CI*	*Pr(β>0)*
F0 maximum	0.28	[0.05 0.51]	0.98
Syllable duration	2.11	[-0.87 5.13]	0.88

Table 2: Estimates from the Bayesian regression models

4. DISCUSSION

The presented results reveal that not only the focused element is realized differently depending on the focus type; the realization of pre-focal elements depends on the following focus type as well. In the case of syllable duration, the effect is even stronger in the pre-focal region. One interpretation is that flatter, lower F0 and shorter durations in the pre-focal region help to boost the prominence perception of the following focus. In this case, the effect on pre-focal elements would be indirect. A more direct interpretation would be that speakers intend to differentiate background before narrow from background before corrective focus. In both perspectives, the prosodic marking of information structure is distributed across the phrase and not localized in the focus. Hence, this study contributes evidence for the significance of pre-nuclear words. The effects presented here are certainly relatively small. Future research will have to test the perceptual relevance of pre-nuclear accents in information structure marking, a research question to which the findings of [14] and [26] give first positive hints. It is also interesting to investigate the relative scaling of pre-nuclear and nuclear pitch accents in relation to the phenomenon known as the *Gussenhoven-Rietveld effect* [27], [28] – the unexpected finding that raising the pre-nuclear peak boosts the perceived prominence of the nuclear peak. While this effect generally predicts the opposite of the results found in this study, it underlines the importance of taking prominence relations into account.

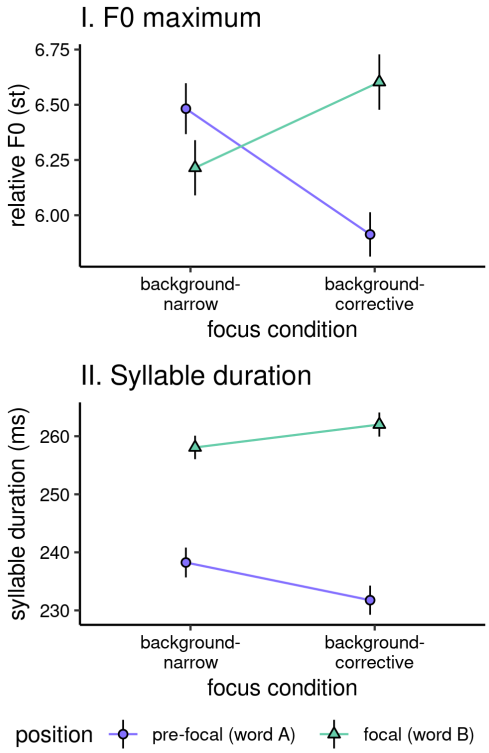


Figure 3: Means (with standard errors). Top: F0 maximum in the word. Bottom: syllable duration.

5. ACKNOWLEDGEMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – RO 6767/1-1 (Walter Benjamin program).

6. REFERENCES

[1] M. Rooth, "Association with focus," Dissertation, University of Massachusetts, Amherst, MA, 1985.

[2] D. R. Ladd, *The structure of intonational meaning: Evidence from English*. Indiana University Press, 1980.

[3] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 2008.

[4] M. Wagner, "Asymmetries in Prosodic Domain Formation," in *Perspectives on Phases*, Cambridge, MA: MITWPL 49, pp. 329–367.

[5] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *J. Phon.*, vol. 36, no. 4, pp. 680–703, Oct. 2008.

[6] F. Kügler, "The role of duration as a phonetic correlate of focus," in *Proc. Speech Prosody 2008*, 2008, pp. 591–594.

[7] D. Mücke and M. Grice, "The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation?," *J. Phon.*, vol. 44, pp. 47–61, 2014.

[8] S. Roessig, D. Mücke, and L. Pagel, "Dimensions of Prosodic Prominence in an Attractor Model," in *Proceedings of Interspeech 2019*, 2019, pp. 2533–2537.

[9] S. Baumann, J. Becker, M. Grice, and D. Mücke, "Tonal and Articulatory Marking of Focus in German," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 1029–1032.

[10] D. Büring, "Intonation, Semantics and Information Structure," in *The Oxford Handbook of Linguistic Interfaces*, G. Ramchand and C. Reiss, Eds. Oxford University Press, 2007.

[11] S. Calhoun, "The centrality of metrical structure in signaling information structure: A probabilistic perspective," *Language*, vol. 86, no. 1, pp. 1–42, 2010.

[12] V. Kapatsinski, P. Olejarczuk, and M. A. Redford, "Perceptual Learning of Intonation Contour Categories in Adults and 9- to 11-Year-Old Children: Adults Are More Narrow-Minded," *Cogn. Sci.*, vol. 41, no. 2, pp. 383–415, Mar. 2017.

[13] B. Andreeva, W. J. Barry, and J. Koreman, "Local and Global Cues in the Prosodic Realization of Broad and Narrow Focus in Bulgarian:," *Phonetica*, vol. 73, no. 3–4, pp. 256–278, Feb. 2017.

[14] H. H. Rump and R. Collier, "Focus Conditions and the Prominence of Pitch-Accented Syllables," *Lang. Speech*, vol. 39, no. 1, pp. 1–17, Jan. 1996.

[15] D. Povey *et al.*, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US, Dec. 2011.

[16] M. McAuliffe, M. Socolof, S. Mihuc, and M. Wagner, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proceedings of INTERSPEECH, 20-24 August, Stockholm, Sweden*, 2017, pp. 498–502.

[17] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[18] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, pp. 1–15, Nov. 2018.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org/

[20] P.-C. Bürkner, "Advanced Bayesian Multilevel Modeling with the R Package brms," *R J.*, vol. 10, no. 1, pp. 395–411, 2018.

[21] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *J. R. Stat. Soc. B*, vol. 73, no. 1, pp. 3–36, 2011.

[22] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs." 2022.

[23] S. Coretta, *tidymv: Tidy Model Visualisation for Generalised Additive Models*. 2022. [Online]. Available: https://CRAN.R-project.org/package=tidymv

[24] H. Wickham *et al.*, "Welcome to the tidyverse," *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019.

[25] A. Zeileis and G. Grothendieck, "zoo: S3 Infrastructure for Regular and Irregular Time Series," *J. Stat. Softw.*, vol. 14, no. 6, pp. 1–27, 2005.

[26] J. Bishop, "Focus projection and prenuclear accents: evidence from lexical processing," *Lang. Cogn. Neurosci.*, vol. 32, no. 2, pp. 236–253, Feb. 2017.

[27] C. Gussenhoven and A. C. M. Rietveld, "Fundamental frequency declination in Dutch: testing three hypotheses," *J. Phon.*, vol. 16, pp. 355–369, 1988.

[28] D. R. Ladd, J. Verhoeven, and K. Jacobs, "Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects," *J. Phon.*, vol. 22, pp. 87–99, 1994.

POEM DECLAMATION IN TWO VARIETIES OF PORTUGUESE: THE LINK BETWEEN PROSODY AND PLEASANTNESS

Plinio A. Barbosa

Dep. of Linguistics, Instituto de Estudos da Linguagem, University of Campinas, Brazil
pbarbosa@unicamp.br

ABSTRACT

In poem declamation, the appropriate use of prosody is essential to cause pleasure. The declamation of three poems in two varieties of Portuguese by ten Brazilian and ten Portuguese speakers balanced for gender is used as a corpus for evaluating the degree of pleasantness by listeners from the same language variety. Six prosodic acoustic parameters related mostly to melody and timing varied largely across the subjects, which is the main source of variability. The evaluation of degree of pleasantness revealed that pause-related and voice quality parameters are the best predictors of pleasantness, but the two varieties differ in the specific parameters and direction of change that increases pleasantness.

Keywords: Poem declamation, Prosody, Portuguese, Pleasantness

1. INTRODUCTION

According to [1], "the evaluation or production of beauty, ugliness, harmony, elegance, fitness or charm is governed by a set of factors such as stimulus symmetry, complexity, novelty, familiarity, artistic style, social status and individual preferences." Elements such as novelty and complexity are associated with prosodic aspects such as the rhythmic structure and the intonation organisation of recited verses, which have effects of surprise and expectation, among others. These effects are coupled together in such a way as to make a performance, such as a recitation, sounds pleasant. Several prosodic parameters seem to have a role in the positive appreciation of a voice. In German, for instance, [2] found a relationship between soft, breathy speech and a lower spectral center of gravity with the growth of positive appreciation by listeners. Also in German, the acoustic measures by [3] seem to confirm these very findings by showing that voices with lower harmonic-to-noise ratios and higher jitter, both related to breathiness, got more pleasant appreciations.

In the specific case of poetry declamation, [4] showed that the realisation of pauses in places not predictable from syntax increased judgments of pleasure, whereas [5] proposed that what characterises the acoustics of poetry declamation are slower speech rates, more frequent pauses, short tone units, relatively equal-length units, low average pitch, and narrow pitch range.

As regards work on European Portuguese (EP), [6] built a Machine Learning system for classifying and predicting the degree of pleasantness of 3-min audio excerpts from prosodic- and voice-quality acoustic parameters. Their results showed that jitter, shimmer, F0 mean, slope of intensity and maximum of F0 slope were the best predictors of pleasantness.

By investigating poetry declamation in Brazilian Portuguese (BP), [7] evaluated the perceptual effects of the production choices of two professional speakers (male and female) reciting the poem "Soneto da Fidelidade" (Sonnet of Fidelity) by Vinícius de Morais. The acoustic analysis of both readings revealed that whispery voice, F0 narrow range and a great number of silent pauses were the production strategies used by the female speaker to express grief and sadness, whereas the male speaker employed much varied intonation patterns to express liveliness. In a follow-up study of the effects of the recitation of the same poem [8] evaluated by eight non-professional speakers, pleasantness was moderately relevant as a subjective parameter related to voice quality and the slope of LTAS.

Working on the acoustic of poetry declamation in both BP and EP, [9] showed that the degree of pleasantness attributed by Brazilian listeners to the recitations in BP was explained by three parameters in order of importance: rate of pause (the lower, the more pleasant), slope of LTAS (the faster the slope, the more pleasant) and pause duration (the longer, the more pleasant) with an additional minor prediction power for F0 median in the case of male reciters. On the other hand, the appreciation of EP recitations by Portuguese listeners depended only on female reciters' pause duration, whereas males' recitations were judged more pleasant due to variation in spectral emphasis, F0 median (the lower, the more pleasant in both cases), and pause

duration (the longer, the more pleasant), in that order of importance.

Based on the reviewed literature and the intuition that poem declamation is a genre in its own that depends less on the content for its expression, our main hypotheses are: H1: There is a cross-variety difference in terms of prosodic-acoustic parameters that explain pleasantness in BP and EP; H2: Pause-related and voice quality parameters are the main parameters for explaining the degree of pleasantness attributed by the listeners, regardless of the theme or the poet; H3: The gender of the reciter affects how pleasantness is evaluated; H4: Rate of pause in BP recitations, and pause duration and spectral emphasis in EP recitations are the most important predictors of pleasantness. The main research question is: What are the acoustic correlates of pleasantness in poetry declamation in BP and EP?

2. METHODOLOGY

The PROS-POIESIS corpus is formed by the declamations of four poems: "Quando vier a primavera" (When Spring comes), with 155 words, and "O amor é uma companhia" (Love is a company), with 111 words, both by Alberto Caeiro (one of Fernando Pessoa's heteronyms, a Portuguese poet), as well as two poems by Adélia Prado, a Brazilian poetess: "Momento" (Moment), with 108 words, and "Tempo" (Time), with 71 words. The four poems were recited by ten Brazilian speakers and ten Portuguese speakers in balanced gender. The set of reciters of the three latter poems was formed by exactly the same persons (called here the Current Set), but this is not the case for the reciters of the first poem, due to the unavailability of the same persons in the moment of the second set of recordings.

Due to the reference of death in "Quando vier a primavera", its valence was labeled as negative, in contrast to the positive one of the second poem "O amor é uma companhia", whose theme is the poet's joy for loving someone. The two poems by Adélia Prado were chosen in order to contrast the positive valence of "Tempo" and the negative valence of "Momento". Not a single reciter has voice professional training and all of them have between 25 and 50 years of age to avoid an age group with effects of vocal aging [10] that usually affect melodic, voice quality and temporal parameters such as speech and pause rate.

Due to the Covid-19 pandemic, the participants themselves used the Easy Voice app in their own cell phones to make all recordings. Because this app allows choosing among different codifications, instructions were given to record all audio files in PCM format (WAV) at a sampling rate of 48 kHz. The author, who is a trained phonetician, further evaluated all audio files. The recordings were resampled at 16 kHz and levelled to the same maximum intensity level at 65 dB.

2.1. Acoustic parameters

The Prosody Descriptor Extractor script for Praat [11] developed by the author [12] was used to extract 22 prosodic-acoustic parameters from a verse-based segmentation of the poems where silent pauses were also segmented. The declamation of each one of the three poems of the Current Set lasted from 31 to 61 s.

For each verse, the prosodic parameters extracted were the following: 12 descriptors of F0: median, semi-amplitude between quartiles (F0SAQ), minimum and maximum, standard-deviations of values and time of F0 local peaks, mean peak rate and mean peak bandwidth, mean and standard-deviations of F0 rates of rises and falls; two intensity descriptors: spectral emphasis [13], and coefficient of variation of total intensity; four voice quality descriptors: Harmonic-to-Noise ratio, HNR (a correlate of breathiness), long-term averaged spectrum slope, computed by the difference in mean energy between the bands 0-1 kHz and 1-4 kHz, jitter and shimmer; and, finally, four temporal descriptors: pause duration and pause rate, speech and articulation rates. Pause rate is studied here by inter-pause interval (IPI) values, because pause rate is the inverse of IPI mean.

2.2. Perception tasks

Two perception tests were carried out: one Likert-scale test for evaluating the degree of pleasantness applied to ten Brazilian listeners and containing declamations by Brazilian reciters, and one Likert-scale test for evaluating the degree of pleasantness applied to ten Portuguese listeners and containing declamations by Portuguese reciters. For that task, each participant listened to the entire declamation of all reciters of s/he variety, which gave 30 declamations to be evaluated per variety (10 reciters x 3 poems). The scale for pleasantness evaluation varied in five degrees from "very unpleasant" ("muito desagradável" in Portuguese, degree 1) to "very pleasant" ("muito agradável" in Portuguese, degree 5) with the neutral response having degree 3. For the perceptual tests, the listeners were selected in the same age range as the reciters to avoid effects of differences in evaluation from younger or older

age groups. In the present study, only the recitations and the evaluations by the listeners of the Current Set are considered for analysis.

2.3. Statistical analyses

To test our hypotheses, three kinds of statistical models were run in the R (R Development Core Team) software. Due to the non-normality of the parameter's distributions, a non-parametric 2-Way ANOVA test, the Scheirer-Ray-Hare (SRH) test, was used for evaluating the mean differences according to two factors: Poet (Adélia Prado vs. Alberto Caeiro) and Valence (positive vs. negative). The models were built on the subsets split according to gender and variety, in each case.

Mixed-effect models, also split by gender and variety, were built in order to evaluate the strength of the two fixed-factors, Poet and Valence, and the random factor, Subject for explaining each parameter' variance. In each case, the rank of the parameter value was used instead, due to the non-normality of the distributions.

The third set of statistical models was based on logistic regression. Two final logistic regression models, one for each variety, corresponding respectively to the two Likert-scale tests presented in the previous section, were retained that predicted the degree of pleasantness from the significant predictor variables (the prosodic parameters). For doing so, the degrees from 1 to 5 were respectively transformed to 0 to 100% in intervals of 5%, allowing the use of a logistic model. Nagelkerke pseudo-correlation measures were used to evaluate the degree of explained variance of these models, which are measures of effect size. In all tests, a 0.05 level of significance was used for decision as regards statistical differences.

3. RESULTS

3.1. Production

The SRH models revealed some significant differences depending on valence and/or poet. Tables 1 and 2 show the significant parameters common to EP and BP: F0 maximum (in Hertz), F0 rise rate (Hertz/frame), F0 fall rate (Hertz/frame), articulation rate (syll./s), pause duration (ms), and IPI (s). When the difference refers to one gender in particular, it is indicated between parentheses (M or F). When the difference refers to valence this is indicated by *n* from negative, or *p* from positive. The same is done in the case of differences related to poet (CA = Alberto Caeiro, AP = Adélia Prado). The mean values are also given.

	BP	
parameter	contrast	p
F0max	n 150 / p 163	0.008 (M)
F0 rise rate	CA 2.2 / AP 2.5	0.0008 (F)
F0 fall rate	CA 3.2 / AP 4.2	0.005 (F)
	n 1.6 / p2.1	0.003 (M)
art. rate	AP.p 5.3 / Others 5.8	0.001
pause dur.	n 476 / p 617	0.05 (M)
IPI	AP.n 2.2 / Others 2.6	0.02 (F)

Table 1: Significantly different parameters for valence and/or poet in common with EP.

Besides the parameters differing in both varieties, there are significant parameters exclusive to a particular variety. In BP, the standard-deviations of F0 peak values ($p = 0.009$) and temporal positions ($p = 0.001$) differ for Valence in the case of female reciters, whereas speech rate is lower for the positive poem by AP (4.2 syll./s vs. 4.5 syll./s for the other poems, $p = 0.01$) in both genders. In EP, F0 mean peak width is higher for the declamations of AP ($p < 0.03$) for male reciters only.

The mixed-effect models revealed that both Poet and Valence and their interaction explain a maximum of only 4% of variance (the highest values are for IPI and pause duration in both BP and EP), considering all significant parameters, whereas the random factor (Subject) explains between 15 and 85% of the total variance, indicating that the subject is the main source of variability. The parameters whose variance is better explained by Subject are melodic (F0 median: around 80%) and voice quality-related (spectral emphasis: 75% in BP and LTAS slope: 72% in EP). Subjects explain between 20 and 30% of the variance of temporal parameters such as speech rate and articulation rate, whereas those of pause duration and IPI relate between 50 (BP, IPI) and 85% (EP, pause duration) with the subjects.

	EP	
parameter	contrast	p
F0max	n 222 / p 231	0.07 (F)
F0 rise rate	CA 2.3 / AP 2.9	0.02 (M)
F0 fall rate	CA 1.4 / AP 2.2	0.0002 (M)
art. rate	AP 6.0 / CA 6.6	10^{-5}
pause dur.	AP 498 / CA 641	0.005 (M)
IPI	n 1.9 / p 2.2	0.003

Table 2: Significantly different parameters for valence or poet in common with BP.

3.2. Perception

The best logistic regression models to predict pleasantness from the acoustic parameters differ across language variety. For BP, in order of importance: IPI (the longer, the more pleasant); spectral emphasis (the lower, the more pleasant); HNR (the higher, the more pleasant); shimmer (the lower, the more pleasant); intensity c.v. (the lower, the more pleasant); and jitter (the higher, the more pleasant). The Nagelkerke pseudo-correlation for this model was 78%. This model does not differ significantly between male and female reciters.

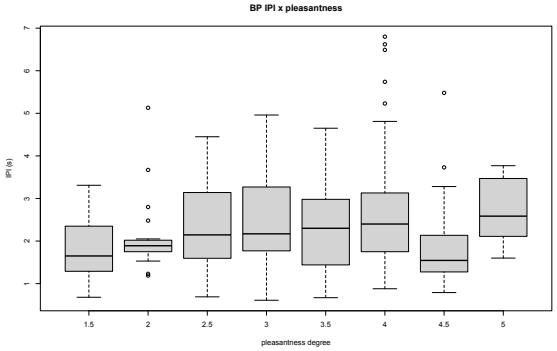


Figure 1: Boxplots of IPI vs. pleasantness in BP.

To illustrate the importance of IPI for pleasantness in BP, see Figure 1. Observe the increase in IPI median for higher values of IPI, with the exception of degree 4.5, which is exclusive for female reciter F2BP evaluations. In the repository given in this paper the reader can appreciate the lowest F0 variability among the female reciters, a feature that could explain her higher degree of pleasantness.

For EP, the models depend on the gender of the reciter. In the case of females, the significant parameters in order of importance are: pause duration (the longer, the more pleasant); rate of pause (the higher, the more pleasant); shimmer (the higher, the more pleasant); LTAS slope (the lower, the more pleasant, which means, less energy in the higher frequencies). For males, only pause duration (the longer, the more pleasant), and LTAS slope (the higher, the more pleasant, which means, higher energy in the higher frequencies was preferred) were significant. The Nagelkerke pseudo-correlations for both models were circa 90%.

4. DISCUSSION

The valence of the poem has a relation to the prosodic parameters. If you look at Tables 1 and 2, F0 maximum (BP males and EP females) and F0 falls (BP males) are higher in positive poems. This could be related to the production of more lively recitations in these poems. Lower IPIs are found in negative poems in at least one poem in both varieties as well. A relation with content cannot be discarded, though, because the recitations in both varieties have higher F0 rises and falls and lower articulation rates for the poems by Adélia Prado.

The findings by [2] and [3] for German are partially the same for Portuguese: in BP, recitations with lower spectral emphasis (less effortful voices) and higher jitter are also preferred. As for HNR, a higher HNR in BP is preferred, though. The role of both jitter and shimmer is confirmed for BP and EP females, as in [6]. Because voice quality parameters were found to be significant for explaining pleasantness in BP, this confirms the findings by [8].

The poetic formula by [5], which proposes that slower speech rates, more frequent pauses, low average pitch, and narrow pitch range, inter alia, are characteristic of poetry, was only related to the fact that more frequent pauses in EP female recitations were more pleasant. The valence and different contents of the recitations presented here do not vary in terms of these parameters as well, which probably could signal that they are not that relevant for present-day poetry declamations.

As regards our hypotheses, H1 to H3 are confirmed and H4 is partially confirmed: although the rate of pause in BP recitations and of pause duration in EP are the most important predictors of pleasantness, spectral emphasis is not a significant predictor in EP but still the second best predictor of pleasantness in BP.

Our findings can contribute as guidelines for coaching aiming at instructing actors, entrepreneurs to have more pleasant voices for the sake of entertaining or persuasion.

5. ACKNOWLEDGMENTS

All stimuli are publicly available for listening at: figshare.com/authors/Plinio_Barbosa/11320902. This research was funded by the CNPq, grant # 302194/2019-3. The study was approved by the Comitê de Ética em Pesquisa of the University of Campinas (protocol # 61430122.9.0000.8142). We thank Ana Rita Valente, Quintino Lopes and Catarina Oliveira in Portugal and Sandra Madureira in Brazil to help with participant recruitment. I also thank Alicia Crochiquia for preparing the online links for the perception tests.

6. REFERENCES

[1] T. Jacobsen, "Beauty and the brain: culture, history and individual differences in aesthetic appreciation," *Journal of Anatomy*, vol. 216, no. 2, pp. 184–191, 2010.

[2] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proceedings of INTERSPEECH 2010*, Chiba, Japan, 2010, pp. 2014–2017.

[3] J. M. Carlsen, M. Franke, L. M. Huttner, and A. Radtke, "How to measure a pleasant voice," in *Proc. of the Conference Phonetics and Phonology in the German Language Area (P&P14)*, Vienna, 2018, pp. 22–26.

[4] S. Postarnak, S. L. Castro, and S. Silva, "The role of linguistic prosody in the responses to recited poetry," *Scientific Study of Literature*, vol. 10, no. 2, pp. 228–249, 2020.

[5] P. B. Byers, "A formula for poetic intonation," *Poetics*, vol. 8, pp. 367–380, 1979.

[6] L. Pinto-Coelho, D. Braga, M. Sales-Dias, and C. Garcia-Mateo, "On the development of an automatic voice pleasantness classification and intensity estimation system," *Computer Speech & Language*, vol. 27, no. 1, pp. 75–88, 2013.

[7] S. Madureira, "Reciting a sonnet: Production strategies and perceptual effects," in *Proc. Speech Prosody Conference 2008*, Campinas, Brazil, 2008, pp. 697–700.

[8] P. Menegon, S. Madureira, and M. Fontes, "Análise de interpretações orais de um poema quanto aos aspectos expressivos," *Intercâmbio*, vol. 49, pp. 113–137, 2021.

[9] P. A. Barbosa, "Pleasantness and wellbeing in poem declamation in European and Brazilian Portuguese depends mostly on pausing and voice quality," *Front. Commun*, 2022, 7: 855177. DOI: 10.3389/fcomm.2022.855177.

[10] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4 - 93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.

[11] P. Boersma and D. Weenink. (2021) Praat: doing phonetics by computer. [Online]. Available: http://www.praat.org/

[12] P. A. Barbosa. (2020) Prosody Descriptor Extractor. [Online]. Available: https://github.com/pabarbosa/prosody-scripts/tree/master/ProsodyDescriptorExtractor/

[13] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," vol. 107, pp. 3438–3451, 2000.

The influence of syllable structure on lexical tones in Croatian: no effect?

Marija Tabain[1], Mate Kapović[2], Matthew Gordon[3], Adele Gregory[1], Richard Beare[4]

[1] La Trobe University, [2] University of Zagreb, [3] University of California Santa Barbara, [4] Monash University and Murdoch Children's Research Institute

m.tabain@latrobe.edu.au, mkapovic@ffzg.hr, mgordon@linguistics.ucsb.edu, a.gregory@latrobe.edu.au, richard.beare@monash.edu

ABSTRACT

Tonal contrasts in many languages are restricted to syllables with overall greater sonority (i.e. syllables containing long vowels or sonorant codas). However, studies to date have tended to focus on languages with relatively restricted syllable structures. In this study we focus on Croatian (Split variety), where both falling and rising lexical pitch accents are possible on all syllable types (unlike many Western South Slavic varieties that restrict tonal contrasts to long vowels), including those containing short vowels and those with an obstruent coda. We examine the timing of pitch peaks and troughs in recordings from 13 speakers who read a list of words varying in lexical pitch accent (rising or falling), vowel duration (long or short) and syllable type (Open, Closed-by-an-Obstruent or Closed-by-a-Sonorant). We find that although relative timing of pitch events varies according to lexical pitch accent, it is remarkably stable across syllable type for a given accent.

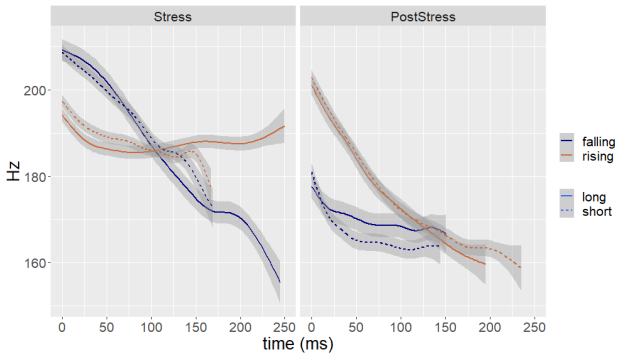
Keywords: lexical pitch accents, timing, syllable structure, Croatian

1. INTRODUCTION

Tonal contrasts in many languages are restricted to syllables with greater overall sonority [1, 2]. Certain languages, e.g. Navajo, thus limit contour tones (falls, rises) to syllables containing a long vowel in the rime, while others, e.g. Kiowa, restrict contours to syllables with either a long vowel or a sonorant coda. These restrictions can be attributed to the relative capacity of different syllables to support f0: syllables with greater sonority provide a better backdrop for tone [2, 3], allowing for crowding of tones that require more time to implement based on number of f0 targets (one, two or more than two) and direction of f0 movement (rising vs. falling tones).

Mitigation of tonal crowding can also manifest itself in gradient fashion through scaling effects, i.e. reducing the f0 distance between high and low targets (undershoot), or temporal shifts, i.e. increasing the distance between tone targets and/or lengthening or adding segmental material on which tones are realized. These strategies are observed not only within syllables but also across longer spans.

This paper explores the role of syllable structure in the realization of tone contrasts in the Split variety of Croatian [4, 5, 6, 7, 8, 9]. As in Standard Croatian, Split Croatian possesses a contrast between a rising and a falling accent on stressed syllables, where the realization of the complete contour often extends into the post-tonic syllable. In the Croatian linguistic orthography, accent type (or "prosodeme") is crossed with vowel length to yield the four categories of short falling, short rising, long falling, and long rising: *pàra* (SF) 'steam', *pàra* (SR) 'money', *Lûka* (LF) 'Luke', *lúka* (LR) 'port'). The falling and rising accent are possible on all syllable types, including those containing short vowels and those with an obstruent coda. The Split variety introduces additional tonal crowding in the form of a "double accent", a f0 fall on the post-tonic syllable after the rising tone. The prosodemes of Split Croatian are shown in Figure 1.



This paper examines the timing of f0 peaks and troughs associated with the prosodemes of Split Croatian in order to assess the impact of syllable structure on temporal aspects of their realization. It is hypothesized that there is greater temporal separation between tone targets in syllable types that are relatively unconducive to supporting a tonal contour, e.g. open syllables containing a short vowel and those containing a coda obstruent.

2. METHOD

2.1. Speakers and recordings

13 speakers (1 male) of the Split variety of Croatian were recorded in March 2022 at the Department of Phonetics recording studio at the Faculty of Humanities and Social Sciences, University of Zagreb, under the supervision of a recording technician and the second author. All speakers were born in Split between 1996 and 2002 (with most born between 1999 and 2002), and had arrived in Zagreb for their university studies between 2015 and 2021. Speakers were recorded using an AKG C414 B-ULS microphone and RME Fireface UFX soundcard. Recordings were saved as mono WAV files using a 44.1 kHz sampling rate and 16 bit bit-depth.

2.2. Stimuli

Speakers read a list of 65 real Croatian words that illustrated the four prosodemes. Words were read in isolation, with no carrier phrase. The words were either stand-alone lexical items, or prosodic words with enclitics. Words contained at least three syllables, and up to six syllables (17 words with three syllables, 24 words with four syllables, 20 words with five syllables and four words with six syllables). The purpose of having longer words was to make sure that there were at least two syllables following the initial stressed syllable, in order to allow for any phrasal tones that may occur in these single-word utterances. In setting up the wordlist, we endeavoured to find words in which the lexical pitch accent occurred in an Open syllable; in a syllable (Closed by a) Sonorant; and in a syllable (Closed by an) Obstruent. (We also included words that contained a syllabic /r/, and these were treated as a separate category.) We based these decisions regarding syllable structure on the Maximum Onset Principle, and on a word-game called *šatrovački govor*, that allows for the reversal of syllables e.g. *go.vnu* > *vnu.go* 'shit'. Morphological structure was also taken into account if there was a transparent prefix+root structure (as for the prosodic word *ìzvadili su ga* 'they took him out', where /iz/ is a prefix meaning 'out' – this lexical pitch accent was therefore classed as being Closed by an Obstruent, even though the sequence /zv/ is permitted in word-onset position, e.g. *zvono* 'bell').

However, in compiling the wordlist, it became apparent that most words have an Open syllable structure on the stressed syllable, given that Croatian is highly permissive in terms of syllable/word onset clusters. For this reason, syllable type is unbalanced across the wordlist:

11 Long Falling (8 Open, 3 Sonorant)
17 Long Rising (16 Open, 1 Sonorant)
17 Short Falling (12 Open, 3 Obstruent, 2 /r/)
20 Short Rising (11 Open, 4 Obstruent, 3 Sonorant, 2 /r/)

We nevertheless elected to keep all of the words for our study, for purposes of greater ecological validity. The wordlist was randomized and each speaker read four repetitions of the list. (One speaker did not produce the word *klȁckali̯su̯se* (Short Falling) 'they were see-sawing', due to the addition of this word to the list after the speaker had been recorded.) If we remove the four words with syllabic /r/, which are not considered in the present study, we have a total of 3172 possible word tokens for analysis (61 x 4 repetitions x 13 speakers, minus four tokens for the speaker who did not read a particular word).

It should be noted that in the present study, we were not able to control for length of the post-tonic vowel, given the other factors we attempted to control for. For this reason both long and short vowels occur in the syllable following the stressed syllable in our word-list.

2.3. Data analysis

Phonetic transcriptions of the words were imported from a spreadsheet and used for preliminary phonetic segmentation with the Munich AUtomatic Segmentation system (MAUS – [10]) pipeline function G2P->MAUS->PHO2SYL. Manual correction of the phonetic MAUS labelling was conducted using the EMU Speech Database Management System [11, 12], interfaced with the R statistical software package [13]. The Snack signal processor [14] was used for calculating formants. The VoiceSauce software was used to extract voicing measures including Straight f0 [15]. Plots were generated using the *ggplot2* package in R [16].

Dominant peaks and troughs in the f0 traces were identified by progressively smoothing the signal until only two roots (zero crossings) of the first derivative of the smoothed signal were present. This strategy eliminated peaks and troughs caused by noise while selecting the strongest peaks/troughs. Smoothing was implemented using the `smooth.spline()` function in *baseR*, which also allows prediction of the derivative, while the *rootsolve* package [17, 18] was used to identify the roots. In cases where a smoothing iteration produced fewer than two roots, the function returned the (e.g. three or four) roots from the previous smoothing iteration. However, if the first smoothing iteration identified two or fewer roots, it returned the root(s) from this first smooth.

We chose to search for two roots as the target number of roots, since we considered it possible to find both a Maximum and a Minimum in each vowel window, based on our examination of pitch traces in a related

study using the same database. The window chosen for analysis was the vowel (in either the Stress or PostStress syllable), with 10 ms added both before and after the vowel in order to maximise the chances of finding a pitch peak or trough in the f0 trace.

Note that in 17 instances, the function failed to converge even after 20 smoothing iterations (this total of 17 instances includes both the Stress and PostStress vowels).

Examination of data returned by the smoothing function showed that most tokens contained either two or three turning points, and a smaller number of tokens contained only one turning point. Very few tokens contained either zero turning points (i.e. a straight line) or four or more turning points: these tokens were discarded. In addition, of the tokens that contained only one turning point, very few of these contained a pitch Minimum as opposed to a pitch Maximum. For this reason, tokens with only a single Minimum were also discarded. Moreover, examination of the data suggested that where three turning points were found, the last turning point was quite late in the token. Given that we had added 10 ms to the end of the vowel window, we decided to discard the third turning point where three turning points were found, and only keep the first two for that particular token.

Finally, we removed any turning points which had an f0 value of less than 100 Hz or greater than 300 Hz, since these were likely to be a result of tracking errors. All of the above filtering procedures left 3000 tokens where the Stress syllable was analyzed, and 2678 tokens where the PostStress syllable was analyzed.

3. RESULTS

For reasons of space, we only present the results for normalized time, which show the timing of the relevant turning point (Minimum (MN) or Maximum (MX)) in the vowel of the syllable (Stress or PostStress).

Figure 1 shows the timing of the Minima and Maxima across the four prosodemes of Split Croatian. Minima are shown on the top row, and Maxima in the bottom row; Stress is shown in the left column, and PostStress in the right column. It can be seen that the Stress syllable Maxima for the Rising accents (red boxes – bottom left) occur later than the Maxima for the Falling accents (blue boxes – also bottom left). This is what we would expect for a Rising versus Falling pitch accent docked onto the Stress syllable. Conversely, the Maxima in the PostStress syllable (bottom right) occur earlier in the Rising accent than in the Falling accent – this is likewise what we would expect given that the Rising accent has a clear high fall on the PostStress syllable in the Split dialect.

Importantly, the variability in the timing of the Maxima differs between the Rising/Falling accents in relation to the syllable type – the Falling accent has less variability in the timing of the Maximum in the Stress syllable, while the Rising accent has less variability in the Maximum in the PostStress syllable. This reflects the importance of the Stress syllable peak for the Falling accent, but of the PostStress syllable peak for the Rising accent.

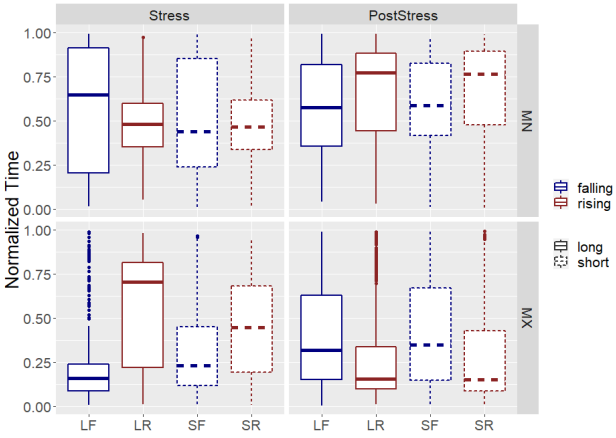


Figure 1: Minima (top row) and Maxima (bottom row) of Stress (left) and PostStress (right) vowels. Rising and Falling pitch accents are shown by colour. Long and short vowels are shown by line type.

Considering now the Minima, we see that the Minimum for the Rising accent tends to occur later in the PostStress syllable, as compared to the Falling accents (top right). This is in line with the high fall that tends to occur on the post-stress syllable of the Rising accent. Turning now to the Minimum in the Stress syllable, we see that although there is less of a clear difference in the means between Rising and Falling accents, there is a clear pattern of less variability in the timing of the Minimum of the Rising accent. This suggests that the timing of a low tone (at around halfway through the Stress vowel duration) is an important identifier of the Rising accent.

To summarize – for the Rising accent, a Minimum is expected at around half-way through the Stress vowel, and a Maximum is expected relatively early in the PostStress vowel. For the Falling accent, a Maximum is expected relatively early in the Stress vowel, while the timing of any Minima or Maxima in the PostStress vowel is less important.

Figure 2 shows the relative timing of Maxima and Minima for the Falling pitch accents according to syllable type. (Note that in all cases, Open~Obstruent~Sonorant refers to the structure of the Stress syllable, not of the PostStress syllable). It can be seen that overall, there are no clear effects of

syllable type on the timing of pitch peaks and troughs. It will be recalled that the timing of the Maximum in the Stress syllable shows little variability for the Falling accent, and this is echoed in Figure 2. It may be noted, however, that the timing of the pitch Maximum on the short vowel is more variable than for the long vowel (bottom left, dotted lines for short vowels) – however, there is no effect of syllable type (Open versus Obstruent). Similarly, there is no difference between Open versus Sonorant for the long vowel (solid lines).

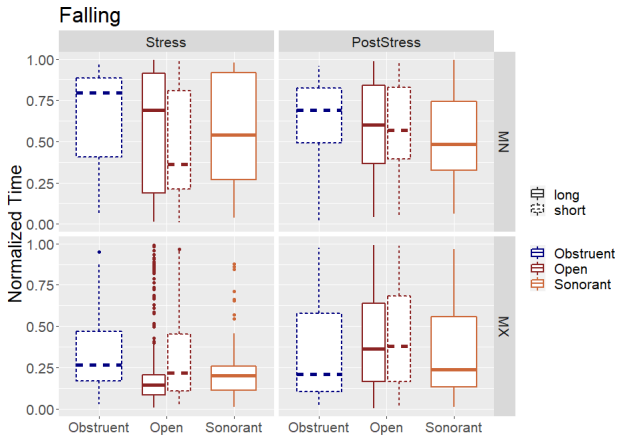


Figure 2: Minima (top row) and Maxima (bottom row) of Stress (left) and PostStress (right) vowels with a Falling pitch accent. Different syllable types are shown by colour. Long and short vowels are shown by line type.

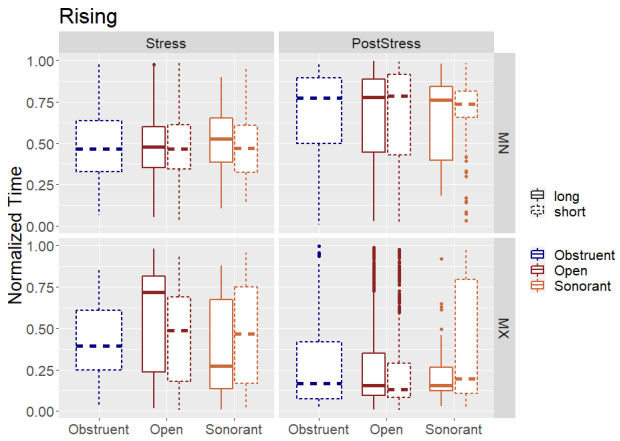


Figure 3: Minima (top row) and Maxima (bottom row) of Stress (left) and PostStress (right) vowels with a Rising pitch accent. Different syllable types are shown by colour. Long and short vowels are shown by line type.

Figure 3 shows the same information as Figure 2, but for Rising pitch accents. It will be recalled that for this pitch accent, the timing of the Minimum in the Stress vowel (top left) and the Maximum in the PostStress vowel (bottom right) are important. However, it is clear that there is neither an effect of vowel duration, nor of syllable type – the Minimum is timed for around halfway through the vowel regardless of any suprasegmental factors, and the Maximum is timed for early in the PostStress vowel, also regardless of any suprasegmental factors.

In summary, the timing of important pitch minima and maxima in this lexical pitch accent language are fixed regardless of syllable type, and perhaps even regardless of vowel duration.

4. DISCUSSION

Analysis of f0 timing and variability indicates a clear division between the rising and falling prosodemes in the importance and location of peaks and troughs. An early f0 peak in the stressed syllable and an early f0 peak in the post-tonic syllable are the most consistent, i.e. least variable, features of the falling and rising prosodemes, respectively. The latter property is a distinguishing characteristic of the Split variety of Croatian and suggests that the contrast between the rising and falling prosodemes may primarily be a contrast between two falling tones, one in the stressed syllable and another in the post-tonic syllable. The timing results thus align with the f0 patterns visually apparent in Figure 1 in that the most salient characteristic of the "rising" prosodeme is the f0 fall in the post-tonic syllable.

The timing of f0 peaks and troughs is not dependent on syllable structure, at least not in a way that suggests sensitivity to the intrinsic ability of different syllable types to support f0 information. The f0 maximum of the falling tone thus does not occur any earlier in a short-voweled syllable that is either open or ends in a coda obstruent. Nor do any peaks or troughs associated with the rising tone reliably shift based on syllable structure.

Overall results suggest that tonal crowding attributed to syllable structure is not alleviated by a temporal shift in f0 targets. Further data would be needed to determine whether tonal crowding is mitigated instead through rescaling of f0 targets. Preliminary results (reflected in Figure 1) indeed suggest that tonal crowding may not be a serious issue if the rising and falling prosodemes both have a single phonological target, a high tone that docks, in the case of the falling prosodeme, on the stressed syllables, and, in the case of the rising tone, on the post-tonic vowel. Investigation of f0 patterns associated with words in different prosodic contexts would be necessary to explore this possibility.

5. ACKNOWLEGEMENTS

We would like to thank our speakers for their time and dedication to language research (Ena Marinković, Ivana Ujević, Jelena Pocedulić, Josipa Papeš, Josipa Teskera, Magdalena Andromak, Mia Mršić, Nuša Vrdoljak, Tea Domić, and four other speakers who did not wish to have their full names published), and Jordan Bićanić for help with the recordings. The ethics protocols for this work were approved by the Office of Research at the University of California, Santa Barbara. Financial support was provided by the School of Humanities and Social Sciences and the Centre for Research on Language Diversity at La Trobe University.

6. REFERENCES

[1] Hyman, L. 1988. Syllable structure constraints on tonal contours. *Linguistique Africaine* 1, 49-60.

[2] Zhang, J. 2002. *The Effects of Duration and Sonority on Contour Tone Distribution: Typological Survey and Formal Analysis.* Routledge.

[3] Gordon, M. 2001. A typology of contour tone restrictions. *Studies in Language* 25, 405-444.

[4] Magner, T. F. "City Dialects in Yugoslavia", *American Contributions to the Eighth International Congress of Slavists* (Zagreb and Ljubljana, September 3-9, 1978). Ohio: Slavica Publishers, Inc., 465–482, 1978.

[5] Kapović, M., *Povijest hrvatske akcentuacije. Fonetika*, Matica hrvatska, Zagreb, 2015.

[6] Godjevac, S., "Transcribing Serbo-Croatian intonation", in S-A. Jun [Ed], *Prosodic Typology*, 146-171, Oxford, 2005.

[7] Browne, E. W. and McCawley, J. D., "Srpskohrvatski akcenat", *Zbornik za filologiju i lingvistiku* VIII:147–151, 1965.

[8] Inkelas, S. and Zec, D. "Serbo-Croatian pitch accent: the interaction of tone, stress, and intonation", *Language*, 64:227-248, 1988.

[9] Rešetar, M., *Die serbokroatische Betonung südwestlicher Mundarten,* Alfred Hölder, K. u K. Hof- und Universitäts-Buchhandler, Wien, 1900.

[10] Kisler, T., Reichel, U. and Schiel, F., "Multilingual processing of speech via web services", *Computer Speech & Language* 45:326–347, 2017.

[11] Winkelmann, R., Harrington, J. and Jänsch K. "EMU-SDMS: Advanced speech database management and analysis in R", *Computer Speech and Language* 45 392-410, 2017.

[12] Winkelmann, Raphael, Jaensch, K., Cassidy, S. and Harrington, J. emuR: Main Package of the EMU Speech Database Management System. R package version 2.0.4, 2019.

[13] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2020.

[14] Sjölander, K. Snack Sound Toolkit, Stockholm: KTH Royal Institute of Technology, Retrieved from http://www.speech.kth.se/snack. 2014.

[15] Vicenik, C., Lin, S., Keating, P. and Shue, Y-L. Online documentation for VoiceSauce. Available at http://www.phonetics.ucla.edu/voicesauce/documentation/index.html, 2020

[16] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2016.

[17] Soetaert K. and P.M.J. Herman (2009). A Practical Guide to Ecological Modelling. Using R as a Simulation Platform. Springer, 372 pp.

[18] Soetaert K. (2009). rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. R-package version 1.6.

Perception of lexical pitch accent and rising shape in South Kyungsang Korean

Hyunjung Joo[1] & Mariapaola D'Imperio[2]

[1]Rutgers University, [2]Aix-Marseille University & Laboratoire Parole et Langage, CNRS
hyunjung.joo@rutgers.edu, mariapaola.dimperio@univ-amu.fr

ABSTRACT

This study investigates how South Kyungsang Korean (SKK) listeners perceive lexical pitch accents depending on three factors: f_0 rise shape, f_0 peak alignment, and segmental duration. Specifically, a two-alternative forced choice task was conducted to see how they categorize different f_0 contours as either H or LH for phrase-medial monosyllabic words. Results showed that f_0 rise shape and segmental duration influenced SKK listeners' categorization, while no effect of peak alignment was observed. Interestingly, they responded to more convex shapes as LH, while more concave shapes, as H. Moreover, shorter duration induced an H categorization, while longer duration was related to LH. Results suggest that SKK listeners use both f_0 shape and segmental duration as important cues for tonal contrast, though f_0 shape shows stronger categorical effect than duration. Thus, f_0 shape information is important to determine phonological representation of pitch accents, as opposed to strict tonal alignment in Autosegmental-Metrical theory.

Keywords: perception, lexical pitch accent, f_0 rise shape, South Kyungsang Korean

1. INTRODUCTION

A great attention has been paid on how to define phonological units of continuous f_0 contour in Intonational Phonology. Broadly, two types of approach have been proposed. The first approach is a configurational approach, where the accentual f_0 contour is considered as a whole, such as rises and falls [5, 11, 22]. That is, f_0 shape in rises and falls would directly reflect intonational categories. The other approach is the Autosegmental Metrical (AM) theory, focusing on local f_0 turning points (targets), such as High and Low [3, 14, 17]. In this approach, a contour is defined with sparse f_0 turning points, which are connected to each other by linear interpolation. However, AM turning points (targets) might not be necessarily enough to characterize an f_0 contour, since they cannot capture f_0 shape differences by limiting to target interpolation [9, 15, 20].

Previous studies showed that several factors such as either f_0 peak alignment or f_0 rise (and fall) shape may play a crucial role in perceiving sentential meaning of f_0 contour in some post-lexical pitch accent languages. For example, in earlier studies, American English listeners have been shown to categorically perceive L*+H (uncertainty) vs. L+H* (assertion) pitch accents depending on the f_0 peak timing [18]. Specifically, they identified earlier peaks as L+H* and later peaks as L*+H. Recently, the Tonal Center of Gravity (TCoG) approach has been proposed to account for both f_0 peak alignment and f_0 contour shape [1, 2]. TCoG calculates the weighted location of the area under the curve of f_0 rising (or falling), being a perceptually important reference.

In fact, recent studies [1, 2] showed that f_0 interpolation shape may influence American English listeners' categorization of L+H* vs. L*+H. That is, listeners responded to more scooped rise shapes as L*+H, while more domed shapes were categorized as L+H*. Moreover, a similar effect was obtained by modifying fall shape within a rise-fall [2]. In similar vein, in French, the rise shape of the LH* accent, being either more concave or more convex, influence the perceptual shift of modality from continuation to question [10]. Specifically, more concave shapes were identified as question, while more convex shapes as continuation.

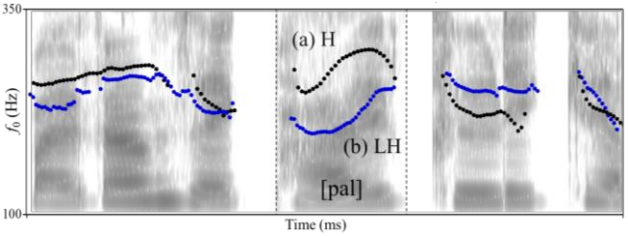


Figure 1: f_0 contour of (a) 'foot' (/pal/ with H tone) and (b) 'shade' (/pal/ with LH tone), embedded in the phrase-medial position.

Unlike pitch accent languages, South Kyungsang Korean (SKK) is a lexical pitch accent language, where f_0 contour is used to distinguish homophone pairs [6, 8, 13, 16]. For monosyllabic words, there are two types of tonal contrast: H and LH[1]. For example, /pal/ with an H tone means 'foot', while /pal/ with a

[1] Several studies used two tonal types, H and LH (or Rising), for monosyllabic words in SKK [6, 8, 13, 16]. Specifically, H tone was described as having a quite higher initial f_0, followed by a low f_0, with a shorter duration. LH tone was described as having a slightly lower initial f_0 and a later f_0 peak alignment (or a low rising contour), with a longer duration.

LH tone means 'shade.' As shown in Fig.1, we can notice that the overall rise shape for H and LH appears to be quite different, and in addition the f_0 peak for H appears to be aligned slightly earlier than the LH peak.

First, note that f_0 height appears to be much higher for H than for LH.[2] Crucially, for the H tone, f_0 rises fast with an increasing slope, showing a concave shape. On the other hand, though both tones are L at the beginning, for the LH tone f_0 curves first downward and then rises with a shallow slope, leading towards a convex rising shape. This may suggest that f_0 rise shape can be an important cue for SKK listeners to differentiate two homophonous lexical items contrasting in lexical pitch accent.

A previous study investigated the effect of factors such as f_0 peak alignment and segmental duration, using only one lexical item in an isolated context [7]. One of the findings was that earlier peaks and shorter duration were identified as H, while later peaks and longer duration were identified as LH. Here, we hypothesize that f_0 rise shape might also affect the H vs. LH contrast in SKK. Therefore, the present study aims to examine how SKK listeners perceive their lexical pitch accent, H vs. LH, depending on three factors: f_0 peak alignment, segmental duration, and f_0 rise shape (concave vs. convex).

2. METHODS

2.1. Stimuli

Target words consisted of three monosyllabic homophone pairs (*gan* [kan] ('taste' and 'liver'), *bam* [pam] ('night' and 'chestnut') *bal* [pal] ('foot' and 'shade')), contrasting by lexical pitch accent (H vs. LH) as shown in Table 1. Test words were embedded in a carrier sentence, *yobeon-eneun mauseulo kkamansaeg ___ keullighaela* [jobʌnænɯn mausɯɾo k'amansæk ___ kɯlɾikhæɾa] 'Click black ___ with the mouse this time', in an IP-medial position. Target words were always preceded by the word *kkamansaeg* [k'amansæk] ('black') to derive the preceding context as HL. 15 repetitions of these sentences were recorded by a female SKK speaker in a sound-proof booth in Seoul, South Korea, using a SHURE KSM44A microphone. In order to clearly induce the tonal contrast, she was asked to assign a contrastive focus[3] on the test words with a normal speech rate. Prosodic rendition was checked by a trained ToBI transcriber.

Word	Tone type	Meaning
/kan/	H	taste
	LH	liver
/pam/	H	night
	LH	chestnut
/pal/	H	foot
	LH	shade

Table 1: Three monosyllabic homophone pairs.

2.1.1. Stimulus resynthesis

Tonal alignment and shape, as well as the duration of the target word were manipulated in Praat [4]. First, the manipulation of f_0 rise shape is illustrated in Fig. 2. As for the target word, the f_0 rise started from the vowel onset (190 Hz), and it ended at the offset of the sonorant coda (290 Hz). In order to create convex and concave f_0 rise shapes, the f_0 value at the midpoint of the rime was increased by 10 Hz steps from the lowest to the highest f_0 level, resulting in 11 steps. The f_0 value of the preceding HL tone ended at the same level (190 Hz) as the vowel onset of the target word. Segmental duration was ambiguous, with mean duration of all H and LH words (290 ms for /kan/, 252 ms for /pal/, and 280 ms for /pam/).

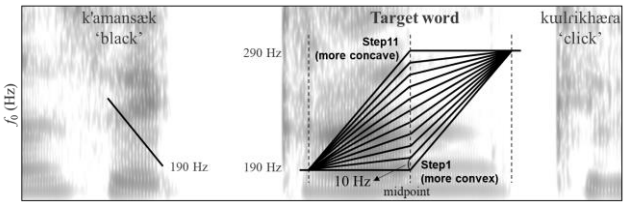


Figure 2: Manipulation of f_0 rise shape.

Second, peak alignment manipulation is illustrated in Fig. 3. Peak alignment was varied from 10% to 100% of the rime, resulting in 10 steps. Each alignment step was increased with 10% of the rime duration (25 ms for /kan/, 23 ms for /pal/, 24 ms for /pam/). The f_0 value of the preceding HL tone ended with the same f_0 value (215 Hz) as at the target word vowel onset. Segmental duration was ambiguous, with the mean duration of all H and LH words (290 ms for /kan/, 252 ms for /pal/, and 280 ms for /pam/), as well as f_0 values for both the L and the H targets (L = 215 Hz; H = 265 Hz).

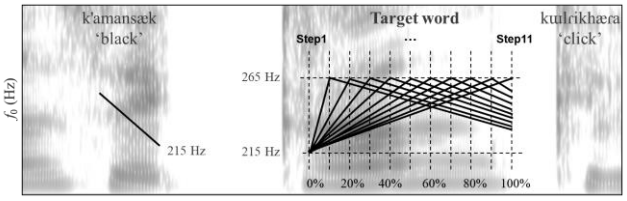


Figure 3: Manipulation of peak alignment.

[2] We haven't factored in f_0 height difference for the categorization of H vs. LH, which will be investigated in a follow-up study. In Fig.1, mean f_0 minimum value was 239 Hz for H and 189 Hz for LH; mean f_0 maximum value (i.e., f_0 peak) was 290 Hz for H and 234 Hz for LH.

[3] In South Kyungsang Korean, only one pitch accent survives in an accentual phrase (AP), while the other pitch accents in the AP become deaccented [13]. Hence, we assigned (a contrastive) focus onto the target word in the AP-medial position to see the lexical pitch accent clearly.

Last, as for the adjustment of segmental duration (Fig. 4), mean duration from the onset of stop release to the offset of sonorant coda was increased by 10 ms, resulting in 11 steps of a continuum. f_0 values of the target word were set to be ambiguous, being the average of all H and LH words (L = 215 Hz; H = 265 Hz). As in the alignment manipulation, the preceding HL tone ended with the same f_0 value (215 Hz) at the target word vowel onset.

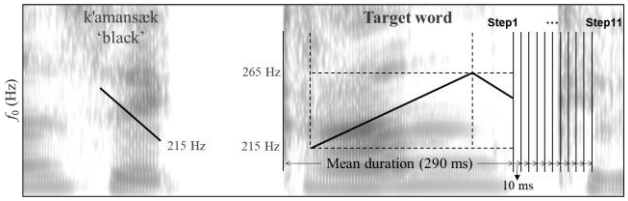


Figure 4: Manipulation of duration.

2.2. Participants

Twenty-five native listeners of South Kyungsang Korean (15 females, 10 males) participated in this study. They were aged from 18 to 27 (mean = 21.6), born and raised in the southern part of the Kyungsang region, mainly Pusan and neighboring cities in South Korea. None of them had hearing problems and they were compensated for their participation.

2.3. Procedure

Participants saw visual stimuli for each target word through a laptop and heard sound stimuli using a headphone (Sony MDR-7509) in a sound-proof booth in Seoul and Ulsan, South Korea. A two-alternative forced choice experiment was carried out using Psychopy [19]. Before the experimental session, the participants had a short training session to become familiarized with the experimental settings. In each trial, they were asked to look at the visual stimuli on the screen upon hearing the sound stimuli and responded by pressing the button of the visual target that matched the target word. Visual stimuli were presented with two different images for a homophone pair (e.g., for the word /pam/, 'night' on the left and 'chestnut' on the right) on the screen. A total of 384 stimuli (3 factors (11 f_0 rise shapes + 10 f_0 peak alignments + 11 durations) x 3 items (/kan/ for 'taste' and 'liver', /pam/ for 'night' and 'chestnut', /pal/ for 'foot' and 'shade') x 4 blocks) were presented. The four blocks were presented in a randomized order, with an interval between the blocks. Among the 25 subjects, the result of one participant was discarded due to overall wrong responses.

2.4. Measurement and statistical analysis

In order to see how SKK listeners perceive f_0 contours depending on rise shape, peak alignment, and segmental duration, response percentage for each step in the continuums were measured. Then, a series of mixed-effects logistic regression models were carried out by using glm package in R [21]. To examine the effects of the three factors on tone identification, H vs. LH response (H coded as 1, LH coded as 0) was used as a dependent variable, while continuum steps were used as a fixed factor. All fixed factors were included as continuous variables. Random intercepts and slopes were set for Subject and Item and *p*-values less than .05 were considered to be significant.

3. RESULTS

3.1. Rise shape

Results showed to what extent SKK listeners categorically perceived H vs. LH depending on f_0 rise shape, either concave or convex, in the f_0 continuum, as shown in Fig. 5. H response percentage for f_0 rise shape showed a curvature pattern similar to a sigmoid function curve. That is, the H response percentage gradually increased from Step 1 to Step 11, which found to be significant in the binomial regression analysis (b = 0.47, z = 7.93, p <.001). The parameter estimates of the model indicated that one step increase in the continuum for Rise shape yielded a change in the log odds of selecting H tone by 0.47, with a positive difference of 19% in the probability of selecting H tone. The category boundary (50% crossover point) was at step 7.03, which was calculated from the simple glmer function (intercept: -2.788; coefficient of Step: 0.397), as shown in the vertical line of Fig. 5. Note that below Step 7 listeners responded as LH (with the rise being more convex), while above Step 7 they responded as H (with the rise being more concave). At the two extremes of the continuum, listeners responded to Step 1 mainly as an LH tone (H response percentage: 0.12), whereas they responded to Step 11 mainly as an H tone (H response percentage: 0.8).

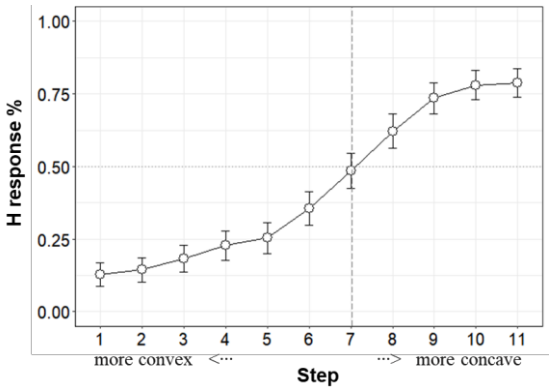


Figure 5: H response percentage for f_0 rise shape.

3.2. Peak alignment

Results showed that SKK listeners did not differentiate H vs. LH depending on f_0 peak alignment, with either an earlier or a later peak as shown in Fig. 6. Even though the H response percentage for f_0 peak alignment appeared to decrease from earlier to later peaks (b = -0.03, z = -7.4, p <.001), H response percentage for all steps were above chance. That is, regardless of the timing of f_0 peak, they perceived all steps in the continuum mainly as H, though more H responses were obtained with earlier peaks.

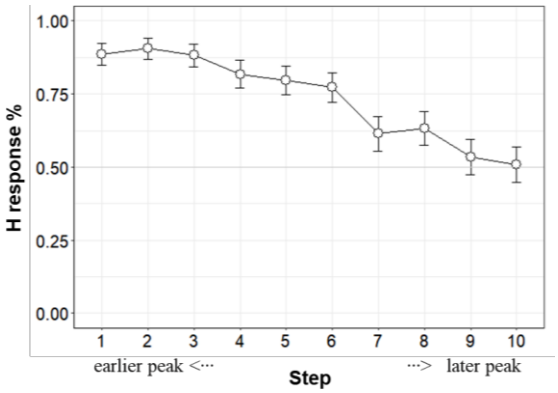


Figure 6: H response percentage for f_0 peak alignment.

3.3. Segmental duration

Results show to what extent SKK listeners categorize H vs. LH depending on segmental duration, either shorter or longer, as shown in Fig. 7. Note that the H response percentage for segmental duration gradually decreased from Step 1 to Step 11 in the continuum, showing a statistical significance of the Step factor (b = -0.02, z = -6.81, p <.001). The parameter estimates of the model indicated that one step increase in the continuum for Duration yielded a change in the log odds of selecting H tone by -0.02, with a negative difference of 5% in the probability of selecting H tone. Specifically, for the two extremes of the continuum, listeners responded to Step 1 (shorter duration) mainly as H tone (H response %: 0.73), whereas they responded to Step 11 (longer duration) mainly as LH tone (H response %: 0.29). The category boundary was at step 4.54, according to the glmer function (intercept: 0.77; coefficient of Step: -0.017), as shown by the vertical line in Fig.7.

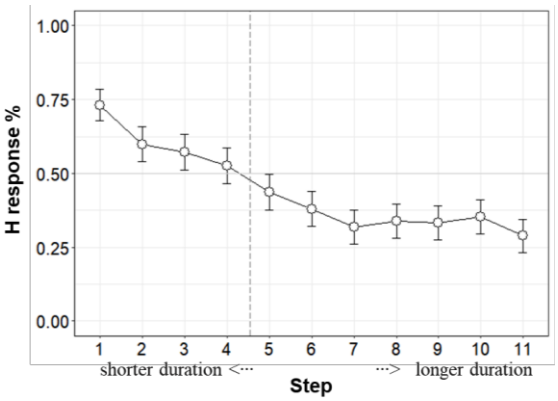


Figure 7: H response percentage for segmental duration.

4. DISCUSSION

This study showed how SKK listeners perceive H vs. LH lexical pitch accents mainly depending on two factors: f_0 rise shape and segmental duration. Crucially, H response percentage showed an effect of both f_0 rise shape and segmental duration. As for the f_0 shape, the H response percentage showed a near-sigmoid curve pattern, with a category boundary between H and LH at Step 7. Note that SKK listeners responded to more convex shape as LH tone while more concave shape as H tone, in an ambiguous duration and f_0 height target. This is consistent with some pitch accent languages in the literature, where tune perception differed depending on the shape of the nuclear pitch accent [1, 2, 10]. This is also in line with the TCoG hypothesis, claiming that the rise (and fall) shape is perceptually crucial to determine pitch accent category in American English [1, 2].

As for the segmental duration, the H response percentage gradually decreased with durational increase, with a category boundary at step 4.54. Specifically, SKK listeners responded to shorter duration as H, while longer duration as LH. This is consistent with previous finding that longer duration increases the identification of an LH tone in SKK [7]. This further confirms the fact that the LH contour tone requires longer duration to be better identified [23]. However, there was no effect of f_0 peak alignment, so that SKK listeners did not differentiate H vs. LH depending on the f_0 peak timing. This is contrary to the previous finding that later peaks were identified as LH tone in SKK, even when the stimuli were presented in isolation [7].

Taken together, both f_0 rise shape and segmental duration influenced tonal perception in SKK. In fact, the overall distribution of the H response was larger for f_0 rise shape (0.12-0.8) than for segmental duration (0.29-0.73), showing a stronger categorical effect. However, further studies are needed to see how these factors interact with each other and which cue is more robust for lexical pitch accent perception in SKK.

Hence, results showed that SKK listeners are mainly sensitive to rise contour shape, in line with what proposed in the configurational approach and TCoG [1, 2, 5, 11, 22] for lexical pitch accents. This also shows that fine-phonetic detail about f_0 shape may be stored in the phonological representation of the pitch accents [1, 2, 12]. However, the absence of a significant effect of f_0 peak alignment supports that it is not sufficient to characterize tunes as a sequence of tone targets. Hence, the target-and-interpolation within AM theory seems rather problematic given the impact of f_0 shape between tone targets [3, 14, 17].

5. CONCLUSION

Our study showed that both f_0 rise shape and segmental duration play a crucial role for SKK listeners to differentiate homophone pairs carrying contrastive lexical pitch accents, i.e. H vs. LH. Specifically, more convex shaped rises as well as longer target words led to a higher LH perception, while more concave shaped and shorter items were mainly identified as H. On the other hand, tonal alignment did not affect lexical accent identification. A potential impact of additional factors such as target f_0 height and its interactions with rise shape and segmental duration need to be examined in further studies.

6. REFERENCES

[1] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, 3(2), 337-383.

[2] Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. 2021. On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics*, 85, 101020.

[3] Beckman, M. E., & Pierrehumbert, J. B. 1986. Intonational structure in Japanese and English. *Phonology*, 3, 255-309.

[4] Boersma, P., & Weenink, D. 2009. Praat: doing phonetics by computer (Version 5.1. 05) [Computer program]. Retrieved May 1, 2009.

[5] Bolinger, D. L. 1951. Intonation: levels versus configurations. *Word*, 7(3), 199-210.

[6] Chang, S. E. 2007. *The phonetics and phonology of South Kyungsang Korean tones.* Unpublished doctoral dissertation, University of Texas, Austin.

[7] Chang, S. E. 2013. Effects of fundamental frequency and duration variation on the perception of South Kyungsang Korean tones. *Language and speech*, 56(2), 211-228.

[8] Do, Y. A., & Kenstowicz, M. 2010. A note on phonological phrasing in South Kyungsang. *Seoul International Conference on Linguistics (SICOL) 2010, June, 23-25, 2010, Korea University, Seoul.*

[9] D'Imperio, M. 2000. *The role of perception in defining tonal targets and their alignment.* Unpublished doctoral dissertation thesis, Ohio State University.

[10] Dorokhova, L., & D'Imperio, M. 2019. Rise dynamics determines tune perception in French: The case of questions and continuations. *In International Congres of Phonetic Science ICPhS2019*.

[11] Hirst, D., & Di Cristo, A. 1998. *A survey of intonation systems*. Intonation systems: A survey of twenty languages, 144.

[12] Kimball, A. E., & Cole, J. 2016. Pitch contour shape matters in memory. In *Proceedings of the International Conference on Speech Prosody*, 8, 1171-75.

[13] Kim, J., & Jun, S. 2009. Prosodic Structure and Focus Prosody of South Kyungsang Korean. *Language Research*, 45.1, 43-66.

[14] Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[15] Ladd, D. R., & Schepman, A. 2003. "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of phonetics*, 31(1), 81-112.

[16] Lee, H., & Zhang, J. 2014. The nominal pitch accent system of South Kyungsang Korean. *Journal of East Asian Linguistics,* 23(1), 71-111.

[17] Pierrehumbert, J. B. 1980. *The phonology and phonetics of English intonation.* Doctoral dissertation, Massachusetts Institute of Technology.

[18] Pierrehumbert, J. B., & Steele, S. 1987. How many rise-fall-rise contours. *In Proceedings of the XIth International Congress of Phonetic Sciences*.

[19] Peirce J. W. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2(10), 1-8.

[20] Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the tobi framework. *In Icslp,* 1, 123-126.

[21] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[22] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech communication*, 46(3-4), 220-251.

[23] Zhang, Jie. 2002. *The effects of duration and sonority on contour tone distribution: A typological survey and formal analysis.* New York: Routledge.

Testing the Locus of Speech-Act Meaning in English Intonation

Thomas Sostarics[1] and Jennifer Cole[1]

[1]Northwestern University, Department of Linguistics
tsostarics@u.northwestern.edu, Jennifer.cole1@northwestern.edu

ABSTRACT

Recent work on rising declaratives proposes a distinction between steep inquisitive rising declaratives and shallow assertive rising declaratives. Yet, it is unclear whether this contrast arises from a phonological distinction of the pitch accent used or a phonetic distinction in the scaling of the boundary tone target. In two perception experiments, we evaluate the contributions of pitch accent and boundary tone in the interpretation of assertive force. In Exp. 1, we find a counterintuitive result for the weighting of pitch accent, which is better understood from the perspective of the Tonal Center of Gravity. This perspective provides a path forward for Exp. 2, which shows no evidence of a contribution from the pitch accent in the interpretation of assertive force. Results speak against a phonological contrast in subtypes of rising declaratives and suggest a need for more narrow investigation in the phonetic domain.

Keywords: intonational meaning, prosody, speech perception, rising declaratives, compositionality

1. INTRODUCTION

The pitch contour from the final stressed syllable to the end of an utterance, the **nuclear tune**, conveys pragmatic meaning in Mainstream U.S. English (MUSE) [1]. In the dominant Autosegmental-Metrical theory of intonational phonology, tunes are made up of high- and low-tone building blocks: **pitch accents** and **edge-tones**, the latter of which are made up of configurations of **phrasal accents** and **boundary tones**. The phonological form of a tune is thus the sum of its parts: typical falling intonation can be described as H*L-L%, a fall (L-L%) from high (H*), and rising intonation can be described as L*H-H%, a rise (H-H%) from low (L*). Tune meaning can be described holistically (e.g., the meaning of "fall" vs. "rise") or compositionally, where pitch accent meaning is typically framed in terms of information structure while edge tones are framed in terms of speech acts or commitment [2,3]. Yet, studies on what pitch accents mean often do not make reference to the edge tones they co-occur with (and vice versa for edge tones). Validating a compositional approach to tune meaning thus requires targeted empirical investigation on whether the meaning conveyed by an individual tone is consistent in the context of other tones, which in turn comprise different tunes.

The contrast between rising and falling nuclear tunes is of particular interest for the phenomenon of **rising declaratives**. While declarative sentences with "default" falling intonation are interpreted as assertions [4], rising intonation changes the interpretation to that of a polar question. However, recent work [5] has proposed a further distinction within rising declaratives, where a steep-rising tune is interpreted as **inquisitive** while a shallow-rising tune is heard as **assertive**. While the two rising tunes differ in the speech act they convey, which would suggest a difference in the edge tones, [5] attributed the contrast to the pitch accent, with L* for the steep rise (in L*H-H%) and H* for the shallow rise (in H*H-H%).

Considering that pitch accents are typically taken to convey referential meaning related to information structure (i.e., focus, givenness), linking the pitch accent to the speech act contrast between the two rises would suggest that the locus of this contrast does not reside solely in the edge-tone configuration. However, despite the claimed contrast in the pitch accent (H* vs. L*), the pitch contours used for both the shallow- and steep-rising stimuli in [5] start from the same initial pitch value. This manipulation suggests that the difference in listeners' interpretation of shallow- and steep-rises was instead due to the phonetic scaling of the edge tones, with lower or higher final pitch. Viewed in this light, the evidence from [5] does not support the claim that the distinction between inquisitive vs. assertive is phonologically encoded through the contrastive intonational feature of the pitch accent. Yet, [5] opens the door to using rising declaratives to more critically consider whether the pitch accent has any bearing on the speech act conveyed by the nuclear tune, or alternatively, whether the locus of this meaning dimension resides solely in the edge tones.

The present paper revisits the form-function mapping between falling vs. rising intonation and assertion vs. question interpretations. We expand the materials from [5] in a two-alternative forced choice task to investigate how the phonetic implementation of rising and falling nuclear tunes relates to proposals regarding the phonological specification of these tunes. Our focus is not narrowly on shallow vs. steep rises, but also on variation in falling tunes. We investigate the form of falling pitch contours with

"gradual" linear slopes (Exp. 1) versus "early" falls that occur immediately after the pitch accent's peak (Exp. 2), a difference we analyze in terms of the Tonal Center of Gravity [6], which provides a common currency for shallow and steep falls as well as rises.

2. EXPERIMENT 1

We adopt the two-alternative forced choice paradigm in [5] to investigate how the accentual and ending pitch of the nuclear pitch contour modulate the probability of a question or assertion interpretation. On each trial, participants listen to a declarative sentence such as *Molly's from Branning*, where *Branning* bears a resynthesized nuclear pitch contour (details below). After listening, participants judge whether the speaker was **telling** them something (=assertion interpretation) or **asking** them something (=question) using the F and J keys on their keyboard.

After making their response, participants count aloud by 2s for 3-5 seconds, starting from a random number presented on the screen. The counting task is added to each trial to prevent participants from comparing pitch contours from trial to trial, as order effects have been previously found for the perception of prominence in rising and falling intonation [7].

Stimuli were created from naturally produced utterances through pitch resynthesis. As base recordings, we use five declarative sentences of the form {Name}'s {determiner/preposition} {noun} such as *Molly's from Branning*. The final noun is disyllabic with word-initial stress and bears the nuclear pitch contour. The first author recorded all utterances in a sound-attenuated booth using H*L-L% (falling) intonation while avoiding phrase-final creak to ensure successful pitch resynthesis.

The resynthesized nuclear pitch contours are linear falls or rises from an **accentual pitch** target (as the cue to pitch accent) to an **ending pitch** target (cueing the edge-tone configuration). Accentual pitch varies from 70Hz to 110Hz in five steps of 10 Hz and is aligned with the end of the stressed syllable of the nuclear word, following [1]. The onglide of the accentual rise/fall begins 50ms before the onset of the nuclear word, while all prenuclear material is held constant at the continuum midpoint (90Hz). Ending pitch varies by ERB-scale differentials based on production data from [8], with five steps between endpoints of -.25 and +2.5 ERBs. These differentials are then added to the lowest accentual pitch (70Hz) to yield five targets between 61Hz and 149Hz. This ending pitch continuum is then crossed with the accentual pitch continuum, shown in Fig. 1. We manipulated the duration of the nuclear word in Praat [9] to maintain a constant duration (thus ensuring equivalent pitch contours when resynthesized) for all utterances.

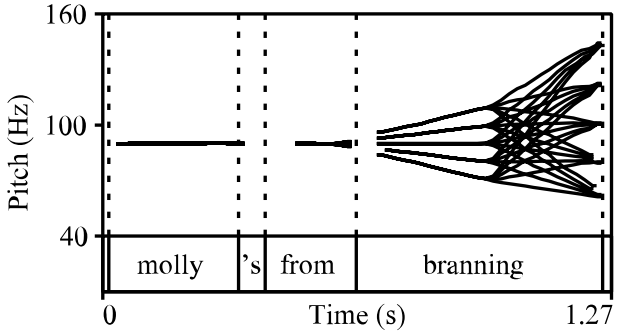


Figure 1: Final 25-step continuum crossing 5 accentual pitch targets and 5 ending pitch targets.

2.1. Experiment 1 Results

We recruited 56 online participants from the Prolific crowdsourcing platform. While participants were instructed to make their decisions as quickly as possible, we omitted trials with reaction times over eight seconds (~3% of the data) from the analysis based on participants' reaction time distributions.

We use Bayesian logistic mixed effects regression to test how accentual pitch, ending pitch, and their interaction (henceforth: the *scaling model*) affect the probability of a *telling* response. We include random intercepts by utterance and by participant and random slopes of accentual pitch, ending pitch, and their interaction by participant. We transformed pitch target values to semitones from 90Hz, which centers the predictors and allows the effects to be interpreted on the semitone scale. The intercept is thus the average log odds of a *telling* response at hypothetical flat pitch at 90Hz. All data, materials, and analyses are available online at osf.io/8hrfv.

Following work from [2,3,4] hypothesizing that differences in the commitment towards a proposition is primarily encoded by the edge-tone configuration and evidence from [5] showing that steeper rises sound more inquisitive than shallow rises, i.e., a higher ending pitch predicts lower likelihood of *telling* responses, *a priori* we predict a negative effect of ending pitch. Under a hypothesis that the pitch accent contributes to the question/assertion contrast [5], where L*H-H% is more inquisitive than H*H-H%, we predict a positive effect of accentual pitch. Alternatively, under a strict compositional account like [2], we would predict no effect of accentual pitch.

Given our bivariate continuum, the proportion of *telling* responses at each combination of accentual and ending pitch are plotted as a heatmap in Fig. 2, with variation in accentual pitch (for the pitch accent) on the horizontal dimension and variation in ending pitch (for the edge tones) on the vertical dimension. An effect of pitch accent would be seen by horizontal

gradation while an effect of edge tone would be seen by vertical gradation.

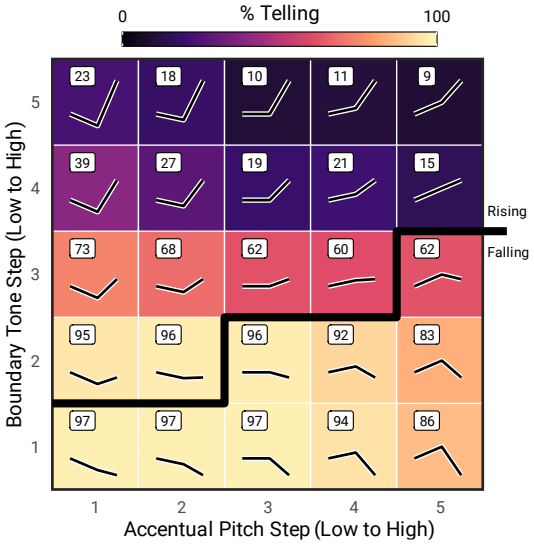


Figure 2: Heatmap of proportion *telling* responses for Exp 1. Schematic F0 contours of continuum steps are shown in each cell, depicting the onglide to the accentual peak across the first syllable and the final rise/fall across the second syllable. Proportions are labelled within each cell.

Our statistical model of the probability of telling responses for Exp. 1 shows a **negative** effect of ending pitch ($\hat{\beta}$ = -0.61, 95% CI [-0.69,-0.53]) and a **weaker negative** effect of accentual pitch ($\hat{\beta}$ = -0.15, CI [-0.23,-0.06]). The model shows no credible evidence of an interaction (CI [-0.01, 0.02]). Overall, there is a bias for *telling* responses, shown most clearly in the middle row of Fig. 2, where responses are closest to chance but still lean towards *telling* responses. This bias is also reflected by a **positive** intercept ($\hat{\beta}$ = 1.98, CI [1.26,2.73]).

2.2. Experiment 1 Discussion

Overall, we find that the accentual pitch and ending pitch are differently weighted as cues to a *telling* interpretation, with the ending pitch cue being four times larger in magnitude than the accentual pitch cue. We also replicate the results of [5], shown by our results in the third column of Fig. 2: steeper rising slopes are more likely to receive inquisitive interpretations. However, rising steps with near-equal pitch excursions, and therefore equal slopes (referencing cells by column-row in Fig. 2, cells 1-3, 3-4, 5-5), differ substantially in the proportion of *telling* responses. Therefore, slope alone, as defined from the pitch accent to the boundary tone, does not capture the range of variation in responses. This same observation suggests that there **is** an effect of pitch accent—yet not in the direction predicted by [5]. This result is counterintuitive: we do not expect **more** *asking* responses as the accentual pitch becomes more like H*, nor do we expect **fewer** *asking* responses with the steepest rise.

The top and bottom rows of Fig. 2 highlight an interesting property: pitch contours that are **overall** "more high" are more likely to receive inquisitive interpretations. This notion of overall highness is captured by the Tonal Center of Gravity (TCoG, [7]), a metric originally proposed to model variation in onglide shapes of bitonal pitch accents. Broadly, TCoG is a 2-dimensional point in the **time** domain (TCoG-T), showing where in time the bulk of high F0 lies, and the **frequency** domain (TCoG-F), showing the overall pitch across a stretch of time. TCoG is calculated by the weighted average of values in one domain weighted by their respective values in the complementary domain (e.g., TCoG-F = Hz values weighted by their timestamps).

In the context of our data and materials shown in Fig. 2, TCoG-F increases both as accentual pitch increases (from left to right in Fig. 2) and as ending pitch increases (from bottom to top). Accordingly, the likelihood of *telling* responses decreases in both directions, hence TCoG-F may provide a univariate lens to better explain the data from Exp. 1. A TCoG perspective also provides an opportunity to return to the simplifying assumption we made in our materials for Exp. 1 where our falling steps are implemented linearly between the accentual/ending pitch targets.

3. EXPERIMENT 2

When making the materials for Exp. 1, we used linear rises **and** falls (henceforth, "gradual" falls), building on [5] to allow for a shared interpretation of slope across both types of contours. However, in the natural recordings using H*L-L% intonation, pitch fell rapidly from the accentual peak, reaching the pitch floor at, on average, 30% of the duration of the second syllable (henceforth, "early" falls). Crucially, TCoG-F for early falls is always lower than that of their gradual counterparts. Based on the results from Exp. 1, the proportion of *telling* responses increases as TCoG-F decreases, and so early falls should be more likely than their gradual counterparts to receive assertive interpretations.

Source files for resynthesis and the continua pitch targets are the same as those described in Exp. 1. However, for falling steps (where ending pitch target < accentual pitch target) we add an additional target at 30% of the duration of the second syllable of the nuclear word. This target uses the same pitch value as the ending pitch target. Rising steps are unchanged. These differences can be seen in the falling steps of the continuum shown in Fig. 3. The paradigm is the same as in Exp. 1, but we added a trial time-out after 8 seconds to improve participant engagement.

3.2. Experiment 2 Results

Fig. 3 shows the heatmap of responses from 54 new participants and the schematic contours for Exp. 2.

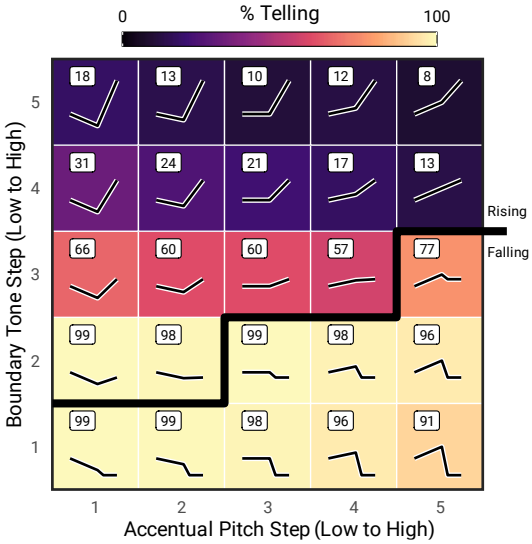


Figure 3: Proportion *telling* responses for Exp. 2 with schematic continuum steps shown in each cell.

We model the results of Exp. 2 using the same statistical model (the scaling model) as in Exp. 1. Again, we find a negative effect of ending pitch ($\hat{\beta}$ = -0.77, CI [-0.88,-0.66]). Importantly, we now find **no credible effect of accentual pitch** ($\hat{\beta}$ = -0.05, CI [-0.13, 0.03]). As predicted, this reduction in the magnitude of the accentual pitch effect is most evident for the falling steps in the lower right quadrant of Fig. 3. Again, we find no credible interaction of accentual and ending pitch. We turn now to modelling our results in terms of TCoG; univariate identification curves are shown in Fig. 4.

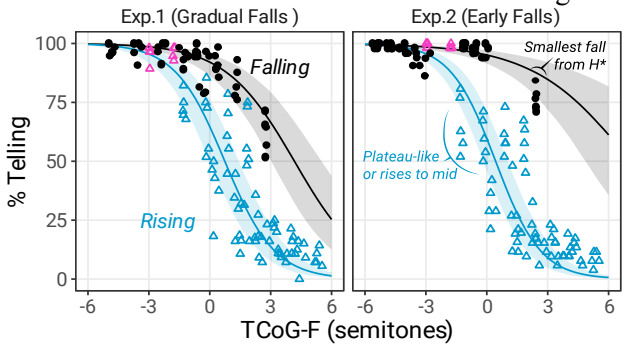


Figure 4: Identification curves with 95% credible interval for TCoG-F, in semitones from 90Hz. Each point is a single utterance from a single rising (blue triangles) or falling (black circles) contour of the 5x5 continuum. The shallowest rises are highlighted in pink.

In Fig. 4, the proportion of *telling* responses for all falling steps (except for cell 5-3 in Fig. 3) are close to ceiling. The scaling model fits the data better compared to a model with TCoG-F as the sole predictor, but it also has more predictors with which to capture variation. Adding a categorical predictor of rising vs. falling shape (coded as +/- .5) and its interaction with TCoG-F allows the identification function to change depending on global tune shape (as in Fig. 4), giving the TCoG-F model a comparable number of parameters to the scaling model. This model shows the expected negative effect of TCoG-F in both experiments (Exp.1 CI: [-0.80,-0.61], Exp2: [-0.77,-0.55]) but a larger difference between the rising/falling groups with the early fall manipulation (Exp.1: [-2.03,-1.67], Exp.2: [-2.83,-2.33]).

Notably, the rising steps with the lowest TCoG-F are the shallow rises (cells 1-2 & 2-2), overlapping with the falling steps' distribution of *telling* response probabilities (in pink at the top of Fig. 4). The interpretation of the most ambiguous steps (row 3) is at chance, with a slight bias towards an assertion interpretation. This bias is likely stems from the declarative syntax and the fact that utterances were presented in isolation with no context. The most ambiguous steps are also most like plateau intonation (ToBI: H*H-L%) which is often used for listing rather than making assertions or questions [2].

4. CONCLUSIONS

In this work we investigated what role, if any, the pitch accent may play in the interpretation of falls and rises within the context of rising declaratives. In Exp. 1 we find a counterintuitive effect of pitch accent when using a simplified implementation of falling steps. A TCoG perspective helps make sense of this finding, predicting that more natural early falls would lower TCoG-F and increase assertion interpretations. This manipulation in Exp. 2 nearly eliminates the previous effect of pitch accent. The TCoG-F model shows a sigmoidal response pattern but requires a supplementary parameter distinguishing the rising/falling shape of the contour to approach performance of the scaling model using accentual and ending pitch. Overall, our results do not support a phonological distinction between inquisitive and assertive rising declaratives based on pitch accent. A phonological distinction may perhaps be maintained by ascribing the contrast to different edge-tone configurations, i.e., L-H% vs. H-H%. Yet, the TCoG results also suggest that the distinction may arise from a probabilistic relation to phonetic gradience such that rising intonation that is overall "more low," as captured by TCoG, is more likely to receive an assertion interpretation. Uncertainty in the interpretation of ambiguous steps points towards the role of context in guiding listeners towards one interpretation over another as well as what other interpretations are available.

5. ACKNOWLEDGMENTS

Thanks to the ProSD Lab at Northwestern and Chun Chan for the experiment implementation. JC's work is supported by NSF BCS-1944773.

5. REFERENCES

[1] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*.
[2] Pierrehumbert, J. and Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication,* vol. 271.
[3] Rudin, D. 2022. Intonational Commitments. *Journal of Semantics* 39(2), 339-383.
[4] Farkas, D., Roelofsen, F. 2017. Division of Labor in the Interpretation of Declaratives and Interrogatives. *Journal of Semantics* 34(2), 237-289.
[5] Jeong, S. 2018. Intonation and Sentence Type Conventions: Two types of Rising Declaratives. *Journal of Semantics* 35(2), 305-356.
[6] Barnes, J., Veilleux, N., Brugos, A., and Shattuck-Hugnagel, S. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3(2), 337-383.
[7] Schiefer, L., Batliner, A. 1991. A ramble round the order effect. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München.*
[8] Steffman, J., Shattuck-Hufnagel, S., Cole, J. 2022. The rise and fall of American English pitch accents: Evidence from an imitation study of rising nuclear tunes. *Proc. Speech Prosody 2022* Lisbon, 857-861.
[9] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.

Correlation of prosodic boundary cues in German

Fabian Schubö[1], *Sabine Zerbian*[1]

[1]University of Stuttgart

fabian.schuboe@ling.uni-stuttgart.de, sabine.zerbian@ifla.uni-stuttgart.de

Abstract

This study investigates how durational and F0 cues expressing intonation phrase boundaries are correlated in German. Experimental speech data is analyzed as to a correlation between pause duration, pre-boundary lengthening, and F0 range. The results reveal a negative correlation between pause duration and pre-boundary lengthening. This observation is compatible with the assumption that the durational cues are engaged in a trading relationship to fill a timing slot associated with the prosodic boundary, as has been proposed for English and Swedish in prior studies. Furthermore, we observed a positive correlation between pause duration and F0 range, which shows that these cues are implemented in congruence at the same level of the prosodic hierarchy. Altogether, our findings suggest that prosodic boundary cues should be investigated in combination rather than in isolation.

Index Terms: prosody, intonation, prosodic phrasing, duration, F0, pause, pre-boundary lengthening, correlation, German

1. Introduction

In speech production, prosodic boundaries are expressed by means of various phonetic correlates. These include durational cues such as pre-boundary lengthening and silent pauses as well as F0 cues such as the excursion of the F0 contour to the top or bottom of the speaker's register at the end of a prosodic phrase (realizing edge tones). While numerous studies investigated prosodic boundary cues in isolation (e.g., [1] for German), only few studies addressed if and how they are correlated (e.g., [2,3] for English).

Some studies provide opposing findings regarding the relationship between pre-boundary lengthening and the presence or duration of silent pauses: For example, [2] found that the amount of pre-boundary lengthening and the duration of a pause are positively correlated in English. That is, the longer the duration of a silent pause was, the longer was the duration of the phrase-final material. Later studies, however, found a negative correlation between these boundary cues in Swedish [4,5] and English [3]. That is, the longer the duration of the silent pause was, the shorter was the duration of the phrase-final material. This effect has been accounted for by means of a trading relationship between the two durational cues: The increased duration of the phrase-final material and the period of silence are both used to fill an abstract timing interval at the prosodic boundary, which can be implemented with differing proportions of each cue [3]. Alternatively, it has been proposed that pre-boundary lengthening and pause duration are not correlated, but that a silent pause only occurs if the pre-boundary material cannot be lengthened any further, which is referred to as the Stretchability Hypothesis [6]. Under this assumption, a silent pause is inserted to fill a timing interval in cases where pre-boundary lengthening has already been maximized.

To our knowledge, correlations between durational and F0 cues for prosodic boundaries have not been investigated in much detail (but see [7]). Different types of F0 contours at prosodic boundaries are often associated with different categories in a prosodic hierarchy, such as the intonation phrase (IP) and the intermediate phrase (ip) in systems of Tone and Break Indices (e.g., [8] with regard to German). The F0 excursion has been observed to be relatively larger at boundaries that are relatively higher in the prosodic hierarchy (e.g., [9] for German). A similar effect has been found for pre-boundary lengthening and pause duration, which are longer at relatively stronger prosodic boundaries (e.g., [3,10]). Taken together these findings suggest a positive correlation of the amount of F0 excursion, the amount of pre-boundary lengthening, and the duration of pauses across different types of prosodic boundaries. It remains unclear, however, if F0 excursion and durational cues are also positively correlated if they occur at the same type of prosodic boundary, such as at the IP level.

The present study investigates how phonetic cues associated with IP boundaries are correlated in German. The investigation is restricted to pause duration, final segment duration (reflecting pre-boundary lengthening), and the expansion of the F0 range on IP-final words (reflecting F0 excursions that realize H% edge tones). Against the background provided above, the present study tests the hypothesis that these cues are positively correlated. With regard to the durational cues, we also test the alternative hypothesis that pause duration and pre-boundary lengthening are negatively correlated. This alternative would be compatible with the assumption that the durational cues are engaged in a trading relationship to fill a timing window, as proposed in [3]. Related to these aspects, the Stretchability Hypothesis [6] is addressed (which predicts that a silent pause occurs only under the condition that the lengthening of the word is maximized). The presence of a correlation between pause duration and pre-boundary lengthening would provide evidence against this hypothesis, as PBL should neither increase nor decrease when a pause is present. The hypotheses were tested on a highly controlled data set, which is described below.

2. Methods

2.1. Data

The data analyzed in this study is a subset of the data collected in a production study on the domain of pre-boundary lengthening in German (see [11] for results). The production study involved the elicitation and audio-recording of read speech in a laboratory setting. The target words were trisyllabic proper names. The names were controlled for the position of main word stress (penultimate vs. antepenultimate) and the

presence/absence of a final coda consonant. The subset of the data employed for the present study includes only those target words that have penultimate stress and an open syllable in final position (e.g., *Ramona*) and were realized with a following IP boundary.

In order to elicit IP boundaries, the target words were embedded in ambiguous lists of the type [N1 or N2 and N3], which can be interpreted as involving a right-branching structure [N1 or [N2 and N3]] or a left-branching structure [[N1 or N2] and N3]. The target words were always the middle name in the list (N2). Prior studies showed that the structural ambiguity of such lists can be resolved by a prosodic phrasing pattern that reflects the respective structure (e.g., [1,12,13]). A right-branching structure typically involves a prosodic boundary after N1 whereas a left-branching structure typically involves a prosodic boundary after N2. The target word (N2) was thus expected to be followed by a prosodic boundary in one of the two conditions, but not in the other. The production study by [1] showed that German speakers produce IP boundaries for the prosodic disambiguation of such lists. In the present study, the lists were embedded in a carrier sentence, as exemplified in (1).

(1) Ich werde Karolin oder Ramona und Peter einladen
I will Karolin or Ramona and Peter invite
'I will invite Karolin or Ramona and Peter.'

Six target words with penultimate stress and CV.CV.CV structure were employed, each having [a] as the final vowel. Twelve sets of sentences of the sort given in (2) were created, henceforth referred to as items. Each target word occurred in two different items. For elicitation, the lists were set in boldface and the respective branching structure was indicated by underlining. The underlining in (2a) indicates a left-branching structure and the one in (2b) indicates a right-branching structure. The sentences were preceded by a short context story. We recorded 24 native speakers of German from the Stuttgart area aged between 18 and 24 years. Each subject produced both types of structures (within-subjects design), which yielded a total of 288 productions for each branching structure (12 items x 24 subjects). As will be further detailed below, we only included those productions that involved an IP boundary after the target word. These were almost exclusively realizations of the left-branching structure (2a).

(2) a. Ich werde **Karolin oder Ramona** und **Peter** einladen.
b. Ich werde **Karolin** oder **Ramona und Peter** einladen.
'I will invite Karolin or Ramona and Peter.'

The elicitation procedure involved a communication task (basically following [1]). Before the recording session, the subject was familiarized with the type of sentences and instructed to produce each sentence in such a way as to communicate the indicated structure to the experimenter. During the recording session, the subject and the experimenter sat at a table separated by a shoulder-high screen. The stimuli were presented to the subject one by one on a display screen in pseudo-randomized order. The subject read each presented stimulus silently and then produced the sentence containing the target word. For each stimulus, the experimenter saw the sentence containing the target word twice on a printed list, marked once with left- and once with right-branching structure. The experimenter listened to the subject's production, decided which of the two structures was expressed, and checked a box next to the respective sentence on the list. The subject did not see the experimenter's decision and no feedback was given. The communication task was meant to make the subjects produce disambiguating prosodic cues more reliably, as prior studies showed that speakers produce prosodic cues for disambiguation in a consistent way only if they are needed for communication (e.g., [14,15]). The recording sessions took place in a sound-attenuated booth at the University of Stuttgart and lasted approximately 45 minutes.

2.2. Analysis

The target words in the recorded productions were analyzed with regard to the duration of the final vowel, the presence and duration of a following pause, the presence and type of edge tones, and F0 range. All annotations and automated procedures were performed using the acoustics analysis software Praat [16]. The boundaries of the word-final vowel, the boundaries of all pauses, and the boundaries of the utterance were manually annotated (following the guidelines in [17]). A pause was defined as a period of silence of at least 20 ms. The presence and type of edge tones, representing the prosodic phrase categories, were annotated based on the GToBI system [8]. For the analysis of F0 range effects, the points of minimum and maximum F0 on the target words and the point of minimum F0 on the last two words of each production were detected by an automated procedure. The resulting duration values for the vowels, pauses, and utterances, the maximum and minimum F0 values (Hz), and the edge tone annotations were extracted by an automated procedure.

In order to neutralize inter-speaker variation, the duration and F0 data were normalized by means of speaker-specific reference points. The duration values for the word-final vowel were normalized by means of z-score transformation. The z-scores were computed based on the subject-specific mean duration and standard deviation of the final vowel in the target words produced without a following prosodic boundary (see also [18]). The following transformation was applied to each vowel duration value: $z=(x-\mu)/\sigma$, where x is the absolute value, μ is the mean, and σ is the standard deviation. The pause duration values were normalized with reference to the articulation rate of the respective utterance, measured as average syllable duration. Relative pause duration was computed by dividing the absolute pause duration by the average syllable duration of the respective utterance (see also [19]). The F0 minimum and maximum values detected on the target word were converted into semitones with reference to the F0 minimum on the last two words of the respective utterance. For each production, the maximum F0 range on the target word was obtained by computing the difference between the respective maximum and minimum semitone value.

For statistical analyses, we employed the software environment R [20], using the Performance Analytics package for correlation analyses and the lme4 package [21] for linear mixed effects (LME) regression models. See below for details.

3. Results

3.1. Prosodic boundary types

As a first step, the productions were classified as to the type of prosodic boundary that followed the target word. Table 1

presents the frequency of prosodic boundary types assumed in the GToBI system for the productions from the left-branching and the right-branching condition, respectively. As expected, the vast majority of instances from the left-branching condition involved an IP boundary after the target word (96%) whereas the vast majority of instances from the right-branching condition did not involve a prosodic boundary in this position (99%). Only three instances involved an ip boundary after the target words. As the present study is restricted to an analysis of phonetic correlates at IP boundaries, all instances not involving such a boundary after the target word are excluded from the subsequent analyses. That is, the data set used in the following includes 279 productions (277 from the left-branching and 2 from the right-branching condition).

Table 1: *Frequency of prosodic boundary type after the target word for each branching structure.*

	IP	ip	None	Total
Left-bran.	277 (96%)	3 (1%)	8 (3%)	288
Right-bran.	2 (1%)	0	286 (99%)	288

The frequency of the types of edge tones in the productions that involved an IP boundary after the target word are given in Table 2. In most instances, the IP boundary was produced with an H-% edge tone (87%). An L-% edge tone occurred only in 3 instances (1%). These instances were excluded from the subsequent analyses of F0 range because the F0 excursion reaches into the opposite direction, which possibly affects the size of the F0 range. Thus, a total of 276 tokens were included in the analyses of F0 range.

Table 2: *Frequency of edge tone types in the production with an IP boundary after the target word.*

H-%	H-H%	L-H%	L-%	Total
243 (87%)	22 (7%)	11 (3%)	3 (1%)	279

The IP boundaries after the target word involve a silent pause in 251 instances (90%). Only 28 productions involved an IP boundary without a silent pause (10%). These productions were excluded from the subsequent analyses of pause duration.

3.2. Correlation analysis

A correlation analysis was performed for each combination of IP boundary cues (relative pause duration, z-score-transformed segment duration, and normalized F0 range in semitones), employing Spearman's rank correlation coefficient. The results of the tests are summarized as a correlation matrix in Figure 1, showing the distribution of each variable on the diagonal, the bivariate scatter plots with a fitted line on the bottom of the diagonal, and the correlation values and significance codes on the top of the diagonal. The tests yielded a positive correlation between pause duration and F0 range ($r(249)=0.39$, $p<0.001$) and a negative correlation between pause duration and segment duration ($r(249)=-0.22$, $p<0.001$). There was no significant effect for a correlation between F0 range and segment duration ($r(274)=-0.017$, $p=0.78$).

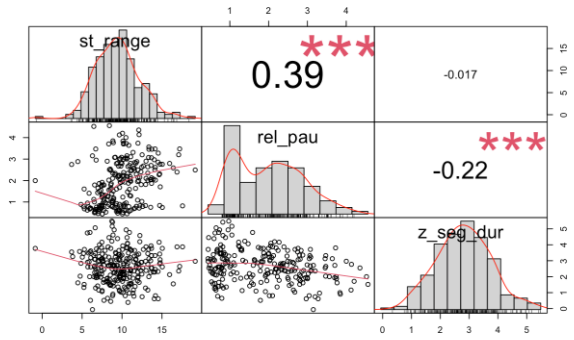


Figure 1: *Correlation matrix for F0 range (st_range), relative pause duration (rel_pau) and z-score-transformed segment duration (z_seg_dur).*

3.3. Short and long pauses

As shown in Figure 2, the histogram for relative pause duration suggests a bi-modal distribution of the data. A Hartigan's dip test yielded a significant effect ($D=0.035$, $p<0.05$), which suggests a deviation from uni-modality. We therefore divided the productions involving pauses into two groups: productions with short pauses and productions with long pauses at the IP boundary. The threshold between long and short pauses was set at the relative duration value 1.5, as pauses near this value have the lowest frequency between the first and the second peak in the distribution. Short pauses occurred in 103 and long pauses in 148 instances.

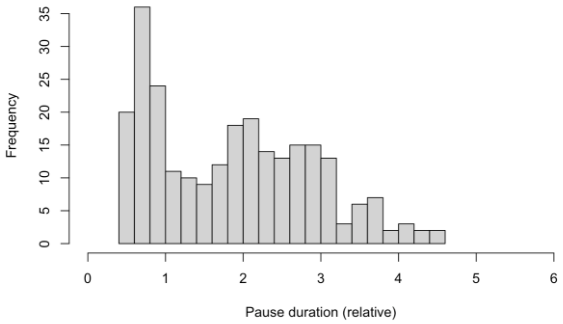


Figure 2: *Histogram for relative pause duration (n=251).*

3.4. Segment duration as a function of pause type

Figure 3 presents the duration data of the IP-final segment (z-score-transformed) by pause type. The boxplots suggest that the segment duration is larger before short pauses than before long pauses. An LME model accounting for segment duration as a function of pause type (levels: short, long) was fitted to the data. The random structure included intercepts for speaker and item. The output is given in Table 3. The model estimated that the duration of IP-final segments is 0.27 points shorter before long pauses than before short pauses. The model was tested against a reduced model without pause type as a fixed factor (thus only involving random factors) by means of a likelihood ratio test, which yielded a significant effect ($\chi^2(1)=4.07$, $p<0.05$). The AIC value of the full model is more than two points lower than

the one of reduced model, which suggests that the full model is significantly better.

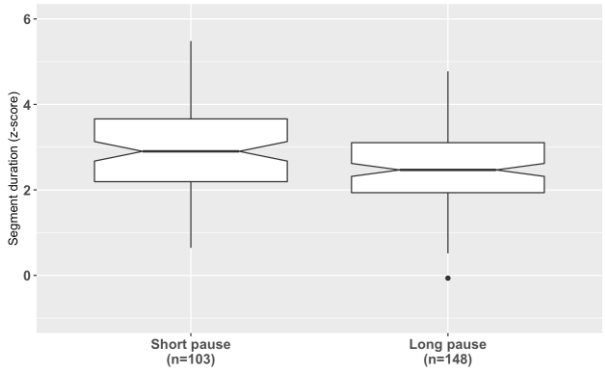


Figure 3: *Duration (z-score) of IP-final segments before short and long pauses.*

Table 3: *Output of an LME model accounting for z-score-transformed IP-final segment duration as a function of pause type.*

	Estimate	Std. Error	t value
Intercept	2.81	0.16	17.82
Pause type *long*	-0.27	0.13	-2.05

3.5. F0 range as a function of pause type

Figure 4 presents the data of the maximum F0 range on the target words (in semitones) by pause type. The boxplots suggest that the F0 range is larger before long pauses than before short pauses. An LME model accounting for F0 range as a function of pause type (levels: short, long) was fitted to the data. The random structure included intercepts for speaker and item. The output is given in Table 4. The model estimated that the F0 range is 1.32 semitones larger on IP-final words before long pauses than before short pauses. The model was tested against a reduced model without pause type as a fixed factor (thus only involving random factors) by means of a likelihood ratio test, which yielded a significant effect ($\chi^2(1)=14.3$, $p<0.001$). The AIC value of the full model is more than two points lower than the one of the reduced model, which suggests that the full model is significantly better.

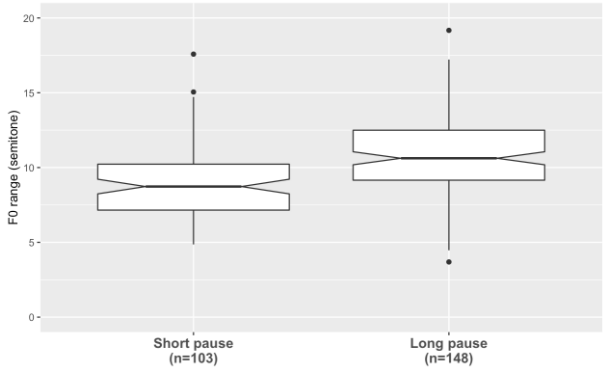


Figure 3: *F0 range (semitones) on target words before short and long pauses.*

Table 4: *Output of an LME model accounting for F0 range (in semitones) as a function of pause type.*

	Estimate	Std. Error	t value
Intercept	9.23	0.42	22.18
Pause type *long*	1.32	0.34	3.87

4. Discussion

The results revealed a negative correlation between pause duration and segment duration (reflecting pre-boundary lengthening) and a positive correlation between pause duration and F0 range (reflecting the degree of excursion in the realization of H% edge tones). There was no correlation between F0 range and pre-boundary lengthening. Pauses were subdivided into short and long pauses based on the frequency distribution of the pause duration data. Pre-boundary lengthening was found to be longer before short pauses than before long pauses while the F0 range was found to be smaller before short pauses than before long pauses.

The findings suggest that pause duration and pre-boundary lengthening are involved in a trading relationship, as found by [3] for English and [4,5] for Swedish. Our findings for German thus contrast with the findings by [2] for English, who observed a positive correlation of the durational cues. Our findings are also incompatible with the Stretchability Hypothesis [6], as a correlation between pre-boundary lengthening and pause duration would be unexpected if the presence of a pause only occurred under the condition that the pre-boundary material cannot be lengthened any further.

The observation that pause duration and F0 range are positively correlated when used for marking IP boundaries shows that a correlation between these cues does not only result from differences between levels in the prosodic hierarchy (e.g., IP vs. ip), but also occurs at prosodic boundaries of the same type. Altogether, the observations from the present study suggest that prosodic boundary cues should be investigated in combination rather than in isolation.

5. Acknowledgements

This study was conducted as part of the research project *Pre-boundary lengthening in a cross-linguistic perspective* at the University of Stuttgart, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 416902968.

6. References

[1] C. Petrone, H. Truckenbrodt, C. Wellmann, J. Holzgrefe-Lang, I. Wartenburger, and B. Höhle, "Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists," *Journal of Phonetics*, vol. 61, pp. 357–366, 2017.

[2] W. E. Cooper and J. Paccia-Cooper, "Syntax and Speech," Harvard University Press, 1980.

[3] F. Ferreira, "Creation of prosody during sentence production," Psychological Review, vol. 100, no. 2, 233–253, 1993.

[4] G. Fant, A. Kruckenberg, and L. Nord, "Prosodic and segmental speaker variations," *Proceedings of the ESCA Workshop on Speaker Characterization in Speech Technology*, pp. 106–120, 1990.

[5] M. Horne, E. Strangert, and Mattias Heldner, "Prosodic boundary strength in Swedish: Final lengthening and silent interval

duration", *Proceedings of the 13th International Congress of Phonetic Sciences*, pp. 170–173, 1995.
[6] E. O. Selkirk, "Phonology and Syntax: The Relation between Sound and Structure," Cambridge: MIT Press, 1984.
[7] G. Kentner, I. Franz, C. A. Knoop, and W. Menninghaus, "The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase," *Journal of Phonetics* 97, 101225, 2023.
[8] M. Grice, S. Baumann, and R. Benzmüller, "German intonation in auto-segmental metrical phonology," In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation*, Oxford: Oxford University Press, pp. 55–83, 2005.
[9] H. Truckenbrodt, "Upstep and embedded register levels," *Phonology* 19, pp. 77–120.
[10] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.
[11] F. Schubö and S. Zerbian, "The patterns of pre-boundary lengthening in German," in F. Schubö, S. Zerbian, S. Hanne and I. Wartenburger (eds.), Prosodic boundary phenomena. Berlin: Language Science Press, to appear.
[12] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.
[13] G. Kentner and C. Féry, "A new approach to prosodic grouping," *The Linguistic Review*, vol. 30, no. 2, pp. 277–311, 2013.
[14] J. Snedeker and J. Trueswell, "Using prosody to avoid ambiguity: effects of speaker awareness and referential context," *Journal of Memory and Language*, vol. 48, pp. 103–130, 2003.
[15] F. Schubö, A. Roth, V. Haase and C. Féry, "Experimental investigations on the prosodic realization of restrictive and appositive relative clauses in German," *Lingua*, vol. 154, pp. 65–86, 2015.
[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", Computer program, version 6.1.05, URL http://www.praat.org/, 2019.
[17] A. E. Turk, S. Nakai, and M. Sugahara, "Acoustic Segments Durations in Prosodic Research: A Practical Guide," in A. Steube (ed.), *Language, Context & Cognition – Methods in Empirical Prosody Research*, pp. 445–472, 2006.
[18] B. Peters, K. J. Kohler and T. Wesener, "Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache," in K. J. Kohler, F. Kleber and B. Peters (eds.), *Prosodic Structures in German Spontaneous Speech* (AIPUK 35a), pp. 143–184, 2005.
[19] P. Hansson, "Perceived boundary strength," *Proceedings of the 7th International Conference on Spoken Language Processing*, 2002.
[20] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org, 2019.
[21] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67(1), pp. 1–48, 2015.

UPTALK AND THE FREQUENCY CODE: HOW GENDER AFFECTS ICONIC ASSOCIATIONS OF PITCH

Sasha Calhoun, Paul Warren, Joy Mills and Jemima Agnew

Te Herenga Waka - Victoria University of Wellington (THW-VUW), New Zealand
firstname.lastname@vuw.ac.nz

ABSTRACT

Meanings of linguistic features are generally taken to be socially constructed. According to the Frequency Code, however, uptalk, involving high, rising pitch, has iconic associations with small body size and female gender, which should influence its affective meanings, e.g., being associated with submissiveness or deference. While uptalk is reported to have some associations consistent with this, the Frequency Code approach assumes culture- and individual-specific ideological links, e.g., between submissiveness and femininity. We used Implicit Association Tests to measure associations between uptalk and each of body size and binary gender. Uptalk was robustly implicitly associated with gender and more weakly with body size. However, the strength and availability of these associations depended on the listeners' gender, gender views and the gender of the voice. We propose physical associations with pitch provide an 'extra-linguistic' basis for meaning, but the salience and availability of these associations differs depending on individuals' beliefs and experience.

Keywords: uptalk, Frequency Code, sociophonetics, iconicity, gender and language.

1. INTRODUCTION

In sociophonetics, social meanings of linguistic features are generally taken to be socially constructed (e.g. [1]). Prosodic features have, however, long been claimed to have universal pre-linguistic associations according to biological codes such as the *Frequency Code*, which links, e.g., high or rising f0 with female gender and small body size [2, 3, 4]. According to this theory, *uptalk* (see [5]), characterised by high rising f0, should have these iconic physical associations, which may lead to affective associations such as submissiveness or friendliness [3, 4, 6]. While this aligns with some proposed meanings of uptalk [5], it does not account for all or for social factors that contribute to establishing such meanings, e.g. listeners who ideologically associate submissiveness with femininity [1, 7]. We seek to reconcile these approaches, proposing the availability of iconic associations of uptalk varies by listeners' experiences and beliefs, particularly about gender [6, 8]. We investigate physical associations of uptalk, and how these vary by listeners' gender and gender beliefs, using Implicit Association Tests (IATs) [9].

Biological codes link physiological properties of pitch production iconically with informational and affective interpretations of pitch in language, which may be phonologized within a language [3, 4]. In many mammalian and avian communication systems, f0 in vocalisations conveys not only physical size but *apparent* size, with, e.g., higher f0 used when acting submissively. Following this, Ohala ([3], p. 327) proposed the Frequency Code: affective meanings such as "deference, politeness, submission, lack of confidence, are signaled by high and/or rising F0". Because of sexual dimorphism, Ohala [3] extended this to human sex differences, as females tend to be smaller than males.

Uptalk, a well-studied feature of many varieties of English, is rising intonation at the end of a declarative [5]. It is found to have many affective associations consistent with the Frequency Code, including deference, uncertainty/lack of confidence, submissiveness and inclusiveness [10, 11, 5, 12, 13, 6]. However, these may be context-dependent: [12] found such associations only in "stereotypical" utterances (about shopping) and [14] found uptalk indexed authority for high status speakers. Further, many of these associations do not match discourse functions of uptalk, e.g. floor-holding [5].

Uptalk is stereotypically associated with women's speech [12]. Many, but not all, studies find uptalk used more frequently by women than men [5]. It attracts significant negative social commentary, so is stigmatised in at least some contexts [5]. The negative affective associations of uptalk above may thus arise through (negative) ideological associations with female gender, whether or not these are seeded or strengthened by iconic pitch associations. To try to unpack this, we consider

listeners' implicit associations of uptalk with physical concepts from the Frequency Code: female gender and small body size. We propose iconic pitch associations provide a shared 'extra-linguistic' basis for affective associations, but the salience and availability of both the physical and affective associations differs according to alignment with the listener's 'world view' and the affordances of the context [1, 15, 7, 16, 6].

In [8], we showed implicit associations between high/low pitch and each of female/male gender and small/large body size using IATs. IATs are a well-established task to measure implicit association strength between paired concepts and attributes [9, 17]. We found stronger associations for male than female listeners and for male than female voices. We argued this is because males, as the historically dominant and privileged gender, have more entrenched ideologies relating to the Frequency Code, so relevant physical associations are more salient. Associations were also stronger with gender than size, which we argued is because cultural stereotypes relating to gender are stronger. Listeners with stronger gender bias (measured in attitude surveys) had stronger pitch associations, although only in 'consistent-first' order (see section 2.3) for some bias measures.

2. METHOD

The IAT experiments aimed to find out if listeners showed implicit associations between uptalk and each of gender and body size, and whether the strength of these associations was affected by listeners' gender, age and experiences and beliefs around gender. The method closely followed [8].

2.1. Participants

Data is reported from 64 female and 67 male participants recruited on Prolific (www.prolific.co). Their median age was 31 years (range 19-72), they had English as their first language and were living in New Zealand or Australia. The study was approved by THW-VUW Human Ethics Committee (No. 29710).

2.2. Materials

Two sets of concept stimuli (gender and size, see [8]) and one set of attribute stimuli (no uptalk/uptalk) were used in the IAT. For gender, twelve names were chosen strongly associated with female (e.g. *Claire*) and male (*Andrew*) gender. For size, photographs of animals (from [18]) were chosen which were either small (e.g. *mouse*) or large (*elephant*).

The stimuli were recorded on discourse markers (DMs), being short phrases which are pragmatically appropriate with either uptalk or falling intonation [19]. The DMs *you know* and *I mean* were used as these are frequent, similar in length and contain only sonorant sounds (important for pitch resynthesis). They were recorded by three female and three male New Zealand English speakers in their thirties and then resynthesised using STRAIGHT in Matlab. Tokens were length normalised to 520ms, with mean F0 values of 195Hz (females) and 115Hz (males). After pretesting, a declination of 1.2 ERB for females and 0.5 ERB for males was imposed over the first 60% of the token. For uptalk, f0 then rose by 8 ERB to the phrase end, while for no uptalk it stayed flat. In a norming study (N=32), participants' accuracy at categorising stimuli as no uptalk or uptalk was 95.2%.

A questionnaire collected basic demographic information and language background, along with responses to statements from established gender surveys that aim to quantify gender attitudes and beliefs. We used five statements from each of the Benevolent Sexism and Hostile Sexism scales [20] used in the New Zealand Attitudes and Values Survey [21], e.g. "Women, compared to men, tend to have greater moral sensibility" and "Women are too easily offended"; six from Social Dominance Orientation, e.g. "Inferior groups should stay in their place" [22]; the five-item Male Norms Inventory [23], e.g. "Boys should prefer to play with trucks rather than dolls"; and two Transgender/Non-Binary Attitude statements adapted from the New Zealand Gender Attitudes Survey [24], e.g. "I would be comfortable with a transgender or non-binary person as a colleague". We used a Likert scale from 1 (strongly disagree) to 7 (strongly agree).

2.3. Design and procedure

The IAT experiments were constructed and run online in PsyToolkit 3.4 [25], following a standard IAT design [9, 26], see Table 1. Participants first learn to classify stimuli from each concept/attribute pair (blocks 1-2), linked to the left ('E') or right ('I') response key. These are then combined. Blocks 3-4 are 'consistent', i.e. the expected concept/attribute pairing on the same response key, e.g. male names and low pitch. The response key for the concept is then reversed (Block 5), and the 'inconsistent' concept/attribute pairing tested (Blocks 6-7). If a participant has an implicit association between concepts and attributes, they should be faster and more accurate in the 'consistent'

(4) than 'inconsistent' (7) blocks.

There were eight IAT versions, crossing two orders of the Gender and Size concepts, each voice gender and two orders of consistent and inconsistent blocks (as effects may be smaller when inconsistent is first, [26]). In inconsistent-consistent order Blocks 1, 3-4 were switched with 5-7. Participants completed the IAT and then the demographic and gender attitude questions.

Table 1: Example sequence of blocks for IAT Experiments. Shows Consistent-First order.

Block	Trials	Type	Left-key response	Right-key response
1	24	Practice	Male names	Female names
2	24	Practice	No uptalk	Uptalk
3	4	Practice	Male + No up	Female + Up
4	48	Consistent	Male + No up	Female + Up
5	36	Practice	Female names	Male names
6	4	Practice	Female + No up	Male + Up
7	48	Inconsistent	No up + Low	Male + Up

2.4. Analysis

Gender attitude responses were analysed through Principal Components Analysis (*prcomp*, with permutation tests in *syndRomics*), which reduced the dimensionality of the 23 questions to 2 significant components. PC1 had positive loadings for all but one question, while PC2 had positive loadings for benevolent sexism, and weaker negative loadings for transgender attitudes and social dominance.

Implicit association strength for each concept-attribute was measured by D-scores [27, 26]. A D-score is a participant's mean RT difference between inconsistent and consistent blocks (i.e. 7 and 4 in Table 1), divided by their SD in these blocks. RTs for incorrect responses were replaced with the participant's mean RT+600ms.

Linear regression models were built in R, with D-score as the dependent. Initial models included three-way interactions of Concept (Gender or Size), Voice Gender (Male or Female voices), Listener Gender, Listener Age, PC1 and PC2, apart from interactions with two or more of Age, PC1 and PC2, as this led to overfitting. The interaction of Order (Consistent-first or Inconsistent-first) and Half (first or second half of the experiment), and listener Handedness (left or right) were also included. Non-significant effects were eliminated using *buildmer*. Model estimates were extracted using *ggeffects*.

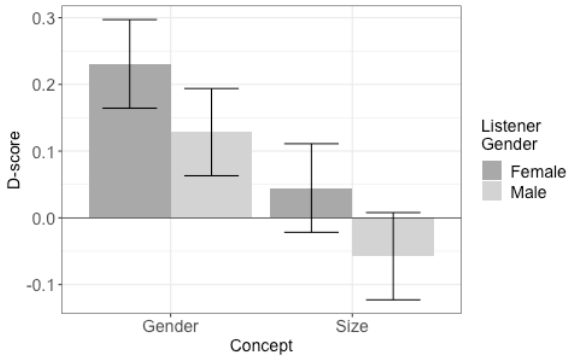


Figure 1: Fitted D-scores by Concept and Listener Gender. Error bars show 95% CIs.

3. RESULTS

The final model contained significant simple effects for Concept ($F(1, 254) = 23.19$, $p<0.001$), Listener Gender ($F(1, 254) = 6.84$, $p=0.009$), PC2 ($F(1, 254) = 6.83$, $p=0.009$) and Order ($F(1, 254) = 30.99$, $p<0.001$) and interactions of Gender:Voice Gender ($F(1, 254) = 4.42$, $p=0.037$) and Gender:PC2 ($F(1, 254) = 3.8$, $p=0.052$). For Order, D-scores showed an implicit association in Consistent-first order (D = 0.19, 95% CI = 0.14, 0.24) but not in Inconsistent-first (D = -0.01, 95% CI = -0.07, 0.04).

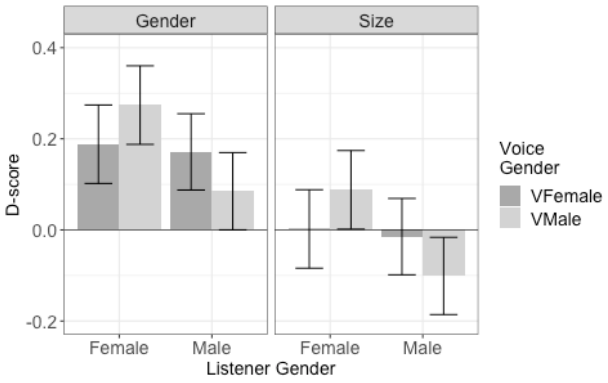


Figure 2: Fitted D-scores by Concept, Listener Gender and Voice Gender. Error bars show 95% CIs.

Remaining effects are presented with the additive effect of Concept, as we are interested in association strength with Gender and Size separately. Comparisons using *emmeans* (fdr method) showed D-scores for Gender are higher than for Size ($t=4.82$, $p<0.001$) and for female listeners than male ($t=2.66$, $p=0.008$). Both genders show the expected association with Gender, e.g. uptalk with female gender, but this is stronger for female listeners (Figure 1). For Size, females show a weak association, while males show a weak reverse association, i.e. uptalk with large body size and/or no uptalk with small.

For Male voices D-scores for female listeners were higher than males (t=3.44, p=0.004). While no other comparisons were significant, listeners of either gender have stronger associations for the opposite gender voice (see Figure 2). In all cases listeners have the expected association of uptalk with female gender. For Size, however, females only show the expected association for male voices, while males show no association with female voices and a reversal with male.

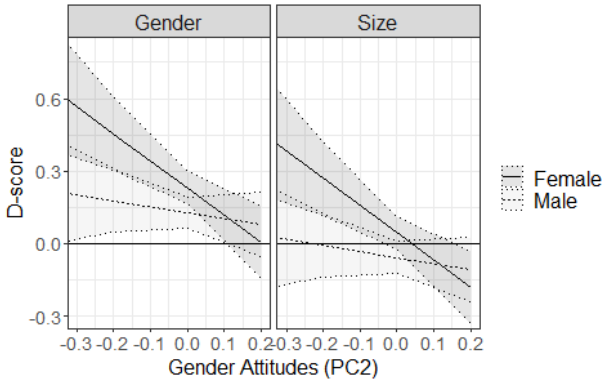


Figure 3: Fitted D-scores by Concept, Listener Gender and Gender Attitudes (PC2). Error bars show 95% CIs.

Slope comparisons in *emtrends* show a significant effect of PC2 for female listeners (t=3.29, p=0.001) but not males (p=0.4). Female listeners with lower PC2 scores have higher D-scores (see Figure 3). All listeners still show expected associations for the Gender concept. For Size, only females with low-to-average PC2 scores show the expected association, while males show a weak reversal. Listeners with low PC2 scores show lower bias on the Benevolent Sexism scale, and more weakly, higher bias on Social Dominance and Transgender Attitudes (see section 2.4).

4. DISCUSSION

We used IATs to explore implicit associations of uptalk predicted by the Frequency Code. Uptalk was robustly associated with female gender, and its absence with male gender, although the strength of the association varied by listener gender, gender bias and voice gender. This is predicted by the Frequency Code, but also follows from a socially constructed association of uptalk with women's speech. Uptalk was more weakly associated with small body size, and its absence with large body size, although only for female listeners, particularly listening to male voices. This provides some support for linking iconic pitch associations to features like uptalk, and particularly for our proposal that the strength and availability of these associations differs according to listeners' experiences and beliefs.

As expected, and matching findings for voice pitch in [8], implicit associations were stronger for the gender concept than body size. This accords with the stronger salience of pitch-related cultural stereotypes for gender than body size. However, opposite to [8], associations were stronger for female than male listeners. We suggest this is because, as stereotypical users, female listeners likely have more experience with uptalk. As uptalk is a phonologized feature, this experience affects association strength more than any gender differences in ideological beliefs re the Frequency Code. Interestingly, only female listeners showed an implicit association between uptalk and size, suggesting this makes wider iconic associations of rising pitch more salient for them, consistent with our proposal.

Both female and male listeners showed stronger associations with voices of the opposite gender. This was not predicted. In [8], stronger associations for voice pitch were found with male voices for both gender groups. We speculate that the stigmatised status of uptalk may make listeners more resistant to associating uptalk with their own gender, making the task more difficult (see also [6]). This may also explain why male listeners listening to male voices apparently show a reverse association for size.

The findings with regard to the gender bias are somewhat ambiguous, as we only found an effect of PC2, which was positively loaded for one measure of gender-related bias (Benevolent Sexism), but negatively loaded for two others (Social Dominance and Transgender Attitudes). It is not clear which of these is more important for interpretation and therefore whether the effect matches our prediction that those with stronger gender bias would show stronger iconic associations. We are still investigating this, but speculate that these explicit measures may not be sufficiently effective for gauging implicit gender bias relevant to the Frequency Code.

IATs show promise to investigate how iconic associations may form part of the cognition and meanings of pitch features. In future work, we plan to use a similar approach to look at affective meanings of uptalk and other pitch features. We believe this has the potential to offer new understandings of how iconic and social sources of meaning interact, by quantifying how differences between listeners affect what they implicitly associate with pitch.

5. ACKNOWLEDGEMENTS

This work was funded by a Faculty Strategic Research Grant from THW-VUW. Thanks to Jen Hay and Abby Walker, and the participants.

6. REFERENCES

[1] A. D'Onofrio and P. Eckert, "Affect and iconicity in phonological variation," *Language in Society*, vol. 50, no. 1, pp. 29–51, 2021.

[2] E. S. Morton, "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," *The American Naturalist*, vol. 111, no. 981, pp. 855–869, 1977.

[3] J. J. Ohala, "The frequency code underlies the sound-symbolic use of voice pitch," in *Sound Symbolism*, L. Hinton, J. Nichols, and J. J. Ohala, Eds. Cambridge University Press, 1994, pp. 325–347.

[4] C. Gussenhoven, "Paralinguistics: Three biological codes," in *The Phonology of Tone and Intonation*. Cambridge University Press, 2004, pp. 71–96.

[5] P. Warren, *Uptalk: The Phenomenon of Rising Intonation*. Cambridge University Press, 2016.

[6] H. White and S. Calhoun, "Mediated iconicity: Effect of age on affective associations of uptalk and creak," in *LabPhon 18*, 2022.

[7] J. Holliday, A. Walker, M. Jung, and E. Cho, "Bringing indexical orders to non-arbitrary meaning: The case of pitch and politeness in english and korean," *Laboratory Phonology*, vol. 14, no. 1, 2023.

[8] S. Calhoun, P. Warren, J. Agnew, and J. Mills, "Gender attitudes affect the strength of the frequency code," in *SST18*, Canberra, Australia, 2022.

[9] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring individual differences in implicit cognition: The Implicit Association Test," *Journal of Personality and Social Psychology*, vol. 74, no. 6, pp. 1464–1480, 1998.

[10] G. R. Guy and J. Vonwiller, "The meaning of an intonation in Australian English," *Australian Journal of Linguistics*, vol. 4, no. 1, pp. 1–17, 1984.

[11] D. Britain, "Linguistic change in intonation: The use of high rising terminals in New Zealand English," *Language Variation and Change*, vol. 4, no. 1, pp. 77–104, 1992.

[12] J. C. Tyler, "Expanding and mapping the indexical field: Rising pitch, the uptalk stereotype, and perceptual variation," *Journal of English Linguistics*, vol. 43, no. 4, pp. 284–310, 2015.

[13] E. Tobin and T. Benders, "Interpretations of Uptalk in Australian English: Low confidence, unfinished speech, and variability within and between listeners," in *SST17*, Sydney, Australia, 2018, pp. 9–12.

[14] C. McLemore, "The interpretation of L* H in English," in *Texas Linguistic Forum-Discourse*, vol. 32. University of Texas, 1991.

[15] P. Eckert, "The limits of meaning: Social indexicality, variation, and the cline of interiority," *Language*, vol. 95, no. 4, pp. 751–776, 2019.

[16] B. Winter, G. E. Oh, I. Hübscher, K. Idemaru, L. Brown, P. Prieto, and S. Grawunder, "Rethinking the frequency code: a meta-analytic review of the role of acoustic body size in communicative phenomena," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 376, no. 1840, p. 20200400, 2021.

[17] B. Kurdi, K. A. Ratliff, and W. A. Cunningham, "Can the Implicit Association Test serve as a valid measure of automatic cognition? A response to Schimmack (2021)," *Perspectives on Psychological Science*, vol. 16, no. 2, pp. 422–434, Mar. 2021.

[18] C. Possidónio, J. Graça, J. Piazza, and M. Prada, "Animal Images Database: Validation of 120 images for human-animal studies," *Animals*, vol. 9, no. 8, p. 475, Aug. 2019.

[19] L. Fung and R. Carter, "Discourse markers and spoken English: Native and learner use in pedagogic settings," *Applied Linguistics*, vol. 28, no. 3, pp. 410–439, 2007.

[20] P. Glick and S. T. Fiske, "The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 491–512, 1996.

[21] C. G. Sibley, "Sampling procedure and sample details for the New Zealand Attitudes and Values Study," New Zealand Attitudes and Values Study, Tech. Rep., 2021. [Online]. Available: http://nzavs.auckland.ac.nz

[22] J. Sidanius and F. Pratto, *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press, 1999.

[23] R. C. McDermott, R. F. Levant, J. H. Hammer, N. C. Borgogna, and D. K. McKelvey, "Development and validation of a five-item Male Role Norms Inventory using bifactor modeling," *Psychology of Men & Masculinities*, vol. 20, no. 4, pp. 467–477, 2019.

[24] National Council of Women of NZ, "New Zealand Gender Attitudes Survey," 2019. [Online]. Available: https://genderequal.nz/ga-survey/

[25] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[26] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, "Understanding and using the Implicit Association Test: I. An improved scoring algorithm." *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 197–216, 2003.

[27] J. Röhner and P. J. Thoss, "A tutorial on how to compute traditional IAT effects with {R}," *The Quantitative Methods for Psychology*, vol. 15, no. 2, pp. 134–147, 2019.

CONTINUATION RISES IN PRE-1974 CYPRIOT GREEK

Joanna Przedlacka[1], Spyros Armostis[2], Mary Baltazani[1], Özlem Ünal-Logačev[3], John Coleman[1]
[1]Phonetics Laboratory, University of Oxford; [2]University of Cyprus; [3]Istanbul Medipol University

{joanna.przedlacka, mary.baltazani, john.coleman}@phon.ox.ac.uk, armosti.spyros@ucy.ac.cy, unaozlem@gmail.com

ABSTRACT

Before the 1974 partition of Cyprus most areas were predominantly Greek-speaking. At the same time there were also many mainly Turkish-speaking settlements and a number of localities with ethnically mixed populations. In this paper we examine the continuation rise intonation tune in archival recordings of the speech of nine Cypriot Greek-speaking males born between 1894 and 1934, from three towns. Using a functional data analysis, we compare their speech patterns to those of their Standard Modern Athenian and mainland Turkish-speaking contemporaries, and to Asia Minor Greek, another Turkish contact dialect in different social circumstances. Our analysis reveals the presence of two patterns in the f_0 curve shape and time alignment of the continuation rise tunes, demonstrating that even the inhabitants of relatively unmixed localities in Cyprus had adopted a Turkish-like intonation pattern in a proportion of their utterances.

Keywords: intonation; mixed populations; Cypriot Greek; Turkish-Greek contact varieties; continuation rise tune; curve fitting.

1. INTRODUCTION

Recent literature ([1-4]) supports the idea that ongoing language contact as experienced by bilingual speakers results in intonational variation and change, giving rise to novel patterns which may combine elements from both contextual languages. Speakers of Asia Minor Greek (AMG), a group of contact dialects originally spoken on the territory of modern Turkey, exhibit a mixture of Greek-like and Turkish-like patterns in continuation rise utterances, both in the f_0 curve shape and the time alignment ([3]). [3] also found that prosodic characteristics of Turkish have persisted in AMG for about a century after the cessation of its contact with Turkish.

Here, we extend the study of continuation rises to Cypriot Greek (CyG), a Greek–Turkish contact variety found in predominantly Greek-speaking communities, while AMG was used in a predominantly Turkish-speaking context. We examine whether the Greek–Turkish sociolinguistic contact in Cyprus resulted in CyG speakers exhibiting a mixture of patterns from the two source languages in the intonation of their continuation rise tunes, as we previously found in AMG. Standard Modern Greek as spoken in Athens (henceforth Athenian) and mainland Turkish are used as controls. We show that CyG continuation rises indeed have two patterns of alignment, one similar to Athenian, the other to Turkish, and we compare the bimodal CyG pattern to that of AMG.

Athenian is the standard variety used in Greece for official communication, in education and in the media. It does not display markedly regional characteristics of any traditional Greek dialect [5], but constitutes an amalgamation of different varieties as a result of internal migration in the latter half of the 20th century [6]. The variety of Turkish examined here is the standard variety as spoken in Istanbul and Western Anatolia. Though not all the Turkish speakers in the corpus come from Istanbul, it was shown in [3] that there were no significant differences in their continuation rise tunes. Cyprus remained under Ottoman rule from 1571 until it was annexed by Britain in 1914 [7], acquiring independence in 1960. When the island was partitioned in 1974, the population was 78.9% Greek and 18.4% Turkish [8]. Very little is known about CyG intonation [9, 10]. Regarding the effect of Greek–Turkish contact on intonation patterns, [11] examined the influence of Cypriot Greek on Cypriot Turkish but we know of no prior work on Turkish influence on CyG.

2. MATERIALS AND METHODS

The data come from 42 male speakers: 8 Greek speakers from Athens, Greece, 20 Turkish speakers from Turkey, 5 speakers of AMG from Cappadocia and 9 CyG speakers from the towns of Morphou, Rizokarpaso and Paphos in Cyprus (Figure 1). Before 1974, the ethnic makeup of these towns was different ([8]): while Paphos had a mixed population (6232 Greek, 2851 Turkish), the other two were mostly inhabited by Greek Cypriots (Morphou 6480 Greek, 123 Turkish; Rizokarpaso 3151 Greek, 2 Turkish). There are reports that the Paphos intonation is perceived as rather distinct from other CyG accents ([12, 13]). The Cypriot speakers were recorded in 1969, eight of them were born between 1894 and 1914 and one in 1934. The Morphou and Rizokarpaso speakers were recorded in interviews, the Paphos informant was recorded narrating a fairy tale.



Figure 1: A contemporary map of Cyprus, showing the locations of speech samples analysed: Paphos, Morphou and Rizokarpaso. Source: Wikimedia Commons

The Athenian data come from popular movies made in 1954–1961 and the Turkish data from radio and TV recordings from 1930–1989 and the Doegen collection (1917). The AMG data are narratives from holdings of Bibliothèque nationale de France (1927–1930). Using spontaneous and semi-spontaneous speech was motivated by reports that contact influences are more evident in less formal speech styles [14].

We analysed 1161 continuation rise tokens (208 Athenian, 479 CyG, 77 AMG, 397 Turkish), varying in length, lexical makeup and syntactic structure. An utterance is defined as a continuation rise if it: 1) is part of a broad focus declarative, 2) is non-final in a speaker's turn, 3) ends in a high boundary optionally followed by a short pause. The Athenian tune is a simple f_0 rise, where a L* nuclear pitch accent typically aligns with the stressed vowel, followed by an H- phrase accent ([15, 16]; Figure 2 top left). The Turkish tune is a rise-fall-rise, where a H* accent is followed by a L H- phrase accent ([17, 18]; Figure 2 top right). Impressionistically, the f_0 movements in Cypriot Greek exhibit two patterns, sometimes with Athenian-like and sometimes with Turkish-like characteristics (Figure 2 bottom left and right).

Native speakers of each language (three of the co-authors) identified the relevant utterances from the corpus of recordings, orthographically transcribed them and translated them into English.

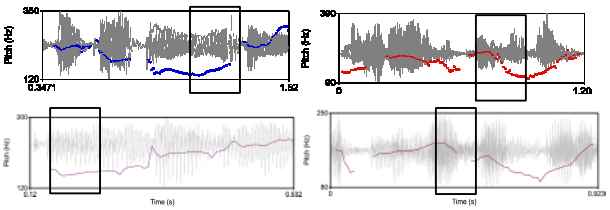


Figure 2: Continuation rise tune examples. Top left: Athenian [triˈada ˈatoma muˈ**i**pane] 'Thirty people told me'. Top right: Turkish [maˈsaja oˈ**tur**madan] 'Before sitting at the table'. Bottom left: Athenian-like Cypriot [iˈ**ka**mari] 'the room'. Bottom right: Turkish-like Cypriot [epluˈ**mi**zando] 'they adorned it'. A rectangle indicates the nuclear vowel, transcribed in bold.

They also located the nuclear vowel in each utterance and manually annotated its beginning and end, using Praat ([19]). We analysed the stretch of the continuation rise over a Region of Interest (ROI), from the beginning of the nuclear vowel to the end of the utterance. Irrespective of any differences in syntactic structure, the same ROI was defined prosodically across the four language varieties for comparability.

2.1. Modelling of f_0 hypotheses and comparisons

For each utterance, f_0 was measured every 10 ms using ESPS *get_f0* [20]. 10th-order polynomials $\hat{f_0} = \Sigma a_n t^n$ for $n = 0, \ldots 10$, were fitted to f_0 contours using the GNU Octave [21] *polyfit* function; pitch errors were inspected and manually corrected. The shape of f_0 contours in the ROI was modelled as 4th-order polynomials $\hat{f_0} = \Sigma a_n t^n$ for $n = 0, \ldots 4$, which were then transformed into orthogonal (Legendre) polynomials $\Sigma\ c_n L_n$ (cf. [22]). The five c_n coefficients capture general shape characteristics of the fitted $\hat{f_0}$ contour: c_0 is the average f_0 height of the contour; c_1 is its slope; c_2 models the shape as a parabola, concave up (or down if the sign is negative); c_3 models the shape as an *N*-like wave with a peak followed by a trough (or the reverse if the sign is negative); and c_4 models the shape as a more complex *M*- or *W*-like wave with more than one peak and trough.

Our hypotheses, informed by visual inspection of the curves and an impressionistic auditory analysis of the annotated CyG tokens are: 1) we expect an influence of Turkish on CyG (as in AMG) to be revealed in the f_0 shape through the similarity in coefficients c_2 and c_3: the Athenian f_0 curve is expected to be a low plateau followed by a final rise, so its c_2 and c_3 coefficients are expected to be near zero, while the Turkish pattern involves a rise-fall-rise f_0 movement, so its c_2 and c_3 coefficients are expected to be positive[1]; 2) based on our findings for AMG we expect two patterns of alignment in the AMG and CyG varieties, one Athenian-like (the L* trough before or within the nuclear vowel) and one Turkish-like, where the L trough is part of the edge tones and occurs after the nuclear vowel. The time of the f_0 minimum (i.e., L) in the Athenian L* H- or Turkish H* L+H- tune was determined by root-finding (i.e. calculating when the first differential of the modelled contour equals zero), using the GNU Octave/Matlab function `real(roots(polyder(a)))`. We define *trough alignment* τ as the difference between the L trough time and the end of the stressed vowel.

Since the AMG and Cypriot data is not unimodal (see Results), many standard parametric statistical models are not appropriate. Therefore, we fitted a 2-

component Gaussian mixture to the AMG and CyG data using the GNU Octave/Matlab `fitgmdist` function, to factor apart the Athenian-like and Turkish-like components of the bimodal AMG and CyG distributions. This function implements an iterative expectation–maximization (EM) algorithm to estimate the parameters of the components. Since the Athenian and Turkish distributions *are* unimodal ([3]), Gaussian distributions were fitted to the Athenian and Turkish data using the GNU Octave/Matlab `normpdf` function. The means of coefficients c_2, c_3 and τ were then compared across the four language varieties using three sets of t-tests.

1) We compared Athenian vs Turkish c_2, c_3 and τ, to check that their patterns are indeed different.

2) Similarly, for AMG and CyG, we tested whether those coefficients are different in the Athenian-like vs. the Turkish-like Gaussian components.

3) We tested whether the Athenian-like components of the CyG and AMG patterns are significantly different from the Athenian controls, and whether the Turkish-like components of the CyG and AMG patterns are significantly different from the Turkish controls. (This is the main question of this paper.)

3. RESULTS

Comparisons between the four varieties revealed that the intonation patterns in CyG are a mixture of Athenian-like and Turkish-like patterns, as we had also found in the AMG contact variety. In the CyG data, there was no significant difference between ethnically mixed and non-mixed towns, so we pool all the Cypriot data together. We present histograms of c_2, c_3 and τ in the four varieties (Figure 3) and the means and standard deviations of the Gaussian distributions (Table 1).

3.1. Continuation rise tune: shape

Mean c_2 is close to 0 in Athenian, indicating a shallower, broader parabolic component than in Turkish (cf. Figure 2), with a more dynamic rise-fall-rise, with a narrower trough than in Athenian and hence a greater c_2. The difference is highly significant ($p < 0.001$). CyG and AMG both show more spread-out c_2 distributions than Athenian or Turkish, and the two components of the Gaussian mixture models have highly significantly different means ($p < 0.001$), even more distinct in CyG than in AMG. In both CyG and AMG, the first component of c_2 is close to 0, similar (but not identical) to Athenian, whereas the second component is >2, similar but not identical to Turkish.

	c_2	c_3	τ (cs)
Athenian	0.19 (1.0)	-0.032 (0.24)	-15.2 (7.5)
Cypriot component 1	-0.01 (0.89)	0.08 (0.19)	-22.5 (7.7)
AMG component 1	0.56 (1.3)	-0.1 (0.25)	-15.2 (9.6)
Cypriot component 2	2.62 (1.19)	0.18 (0.36)	8.1 (10.1)
AMG component 2	3.1 (1.23)	0.36 (0.27)	8.0 (13.8)
Turkish	2.1 (1.7)	0.19 (0.38)	18.0 (11)

Table 1: Means (standard deviations) of c_2, c_3 and τ. For CyG and AMG, means and standard deviations of the two components in the Gaussian mixture models are given separately, for comparison with the single Gaussian pdfs used to model the Athenian and Turkish controls.

As expected, the Turkish-like component of the CyG pattern is *not* significantly different from the Turkish controls in the shape coefficients c_2 and c_3. However, contrary to our expectation, the L trough alignment in the Turkish-like component of CyG *is* significantly different from the Turkish controls ($p < 0.001$). Furthermore, c_2, c_3 and τ of the Athenian-like components of the CyG pattern *are* significantly different from the Athenian controls (for c_2, $p < 0.05$; for c_3 and τ, $p < 0.001$). Somewhat differently from CyG, the Turkish-like component of AMG is highly significantly different from the Turkish controls in c_3 and τ. The parameters of the Athenian-like component of AMG are not significantly different from Athenian except for c_3, which is significantly different only at the $p < 0.05$ level.

In spite of these significant differences between the AMG and CyG components and the Athenian and Turkish controls, it should be noted that he magnitudes of the differences are very small, showing that the two components of CyG and AMG are *similar* to Athenian and Turkish, consistent with the hypothesis that CyG and AMG intonation is a mixture of the two control varieties.

AMG resembles Turkish more than Athenian in contour shape, according to coefficient c_3, which was usually close to 0 in Athenian, i.e., the trough of the contour preceded the peak. In contrast, c_3 was positive in Turkish and AMG component 2, that is, the peak of the contour preceded the trough.

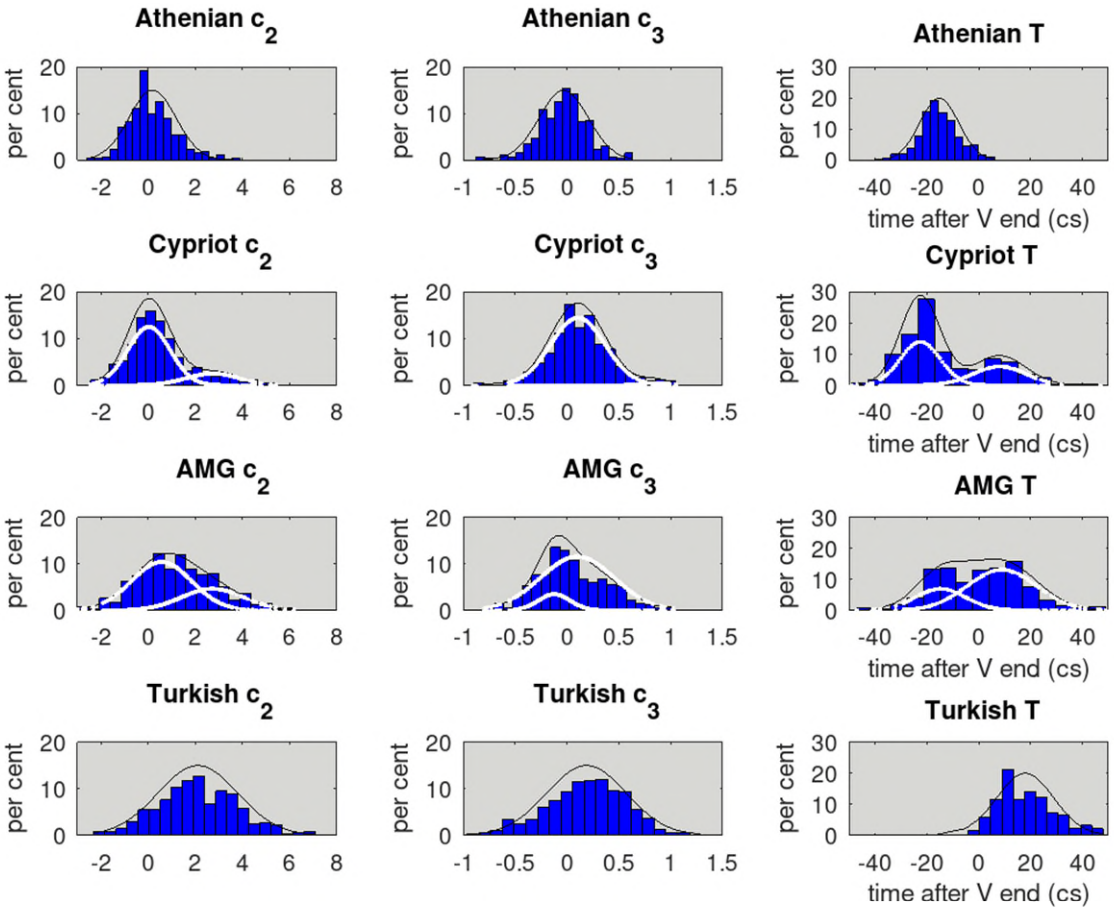


Figure 3: Histograms of *c2*, *c3* and τ (time lag from the nuclear vowel end to the L trough) in continuation rises in Athenian, CyG, AMG and Turkish. Black line: Gaussian distribution fitted to the Athenian and Turkish data, and a 2-component Gaussian mixture fitted to the AMG and Cypriot data. White lines: The two components of the mixture models.

3.2. Continuation rise tune: alignment

The distribution of trough alignment in CyG continuation rises is bimodal, with two peaks, one resembling the Athenian and the other the Turkish distribution. In AMG and CyG, the Athenian-like and Turkish-like Gaussian components of τ are highly significantly different ($p < 0.001$). The Athenian mean is significantly greater i.e. later than the CyG component 1 ($p < 0.001$), by c. 0.7 s, and the Turkish mean is significantly greater than the AMG and CyG component 2 ($p < 0.001$), by c. 0.1 s.

4. DISCUSSION

As reported previously [3, 4], there are differences in the shape and alignment of Athenian and Turkish continuation rises. The nuclear vowel aligns with a trough in Athenian but with a peak in Turkish. In general, CyG continuation rises had variable realisations resembling sometimes the Athenian and sometimes the Turkish intonation patterns. However, the CyG trough in component 1, although similar to Athenian, was aligned significantly earlier than in Athenian, and the trough in component 2 was aligned significantly earlier than in Turkish. The shapes of the two patterns of CyG continuation rises were also similar to (but significantly different from) their Athenian and Turkish controls. These results suggest that the variation in the intonational patterns observed in CyG, as in AMG, can at least in part be accounted for as a result of language contact.

More generally, these results provide further support to previous reports [2-4] that contact between languages from different families which have differences in syntax, morphology and phonology, can influence the intonation systems of the contact variety. Specifics of the intonational phonology and the tune-text alignment seem to be transferred between languages.

5. REFERENCES

[1] Mennen, I. 2004. Bi-directional interference in the intonation of Dutch speakers of Greek. *J. Phon.* 32, 543–563.

[2] Queen, R. 2012. Turkish−German bilinguals and their intonation: Triangulating evidence about contact induced language change. *Language* 88(4):791–816.

[3] Baltazani, M., J. Przedlacka, Ö. Ünal-Logačev, P. Logačev & J. Coleman. 2022. Intonation of Greek in contact with Turkish: a diachronic study. Language Variation and Change, 34(3), 271-303. doi:10.1017/S0954394522000126.

[4] Baltazani, M., Przedlacka, J., Coleman, J. 2020. Intonation of Greek–Turkish contact: a real-time diachronic study. Proceedings of the 10th International Conference on Speech Prosody 2020, 730-734, DOI: 10.21437/SpeechProsody.2020-149.

[5] Trudgill, P. 2003. Modern Greek dialects. A preliminary classification. *Journal of Greek Linguistics* 4:54–64.

[6] Allen, Peter S. 1986. Positive aspects of Greek urbanization: The case of Athens by 1980. *Ekistics* 53(318/319):187–194. ww.jstor.org/stable/43621978.

[7] Borowiec, A. 2000. *Cyprus: a troubled island.* Greenwood Publishing Group. ISBN 978-0-275-96533-4.

[8] Cyprus Ministry of Interior. 1992. The Demographic Structure of Cyprus. Parliamentary Assembly. p. 6. Archived (PDF) from the original on 7 January 2011. Retrieved 22 January 2011.

[9] Grice, M., D. R. Ladd & A. Arvaniti. 2000. On the place of "phrase accents" in intonational phonology. *Phonology* 17:143-185.

[10] Themistocleous, Ch. 2012. Cypriot Greek Nuclear Pitch Accents (*Ta pirinika tonika ipsi tis kypriakis ellinikis*). In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds), Selected papers of the 10th ICGL. Komotini, Greece: Democritus University of Thrace: 796–805.

[11] Kappler, M. 2011. A tale of two languages. Tracing the History of Turkish-Greek Language Contacts. *Türk Dilleri Araştırmaları*, 21.1: 95-130

[12] Christodoulou, Ch. 2015. *The Local Dialect of Northwest Paphos: Phonological Description* [In Greek]. PhD dissertation, Aristotle University of Thessaloniki.

[13] Fotiou, C., Grohmann, K.K. 2022. A Small Island With Big Differences? Folk Perceptions in the Context of Dialect Levelling and Koineization. *Frontiers in communication* 6, article 770088. www.frontiersin.org.

[14] Barnes, H. & J. Michnowicz. 2015. Broad focus declaratives in Veneto-Spanish bilinguals: Peak alignment and language contact. *Studies in Hispanic and Lusophone Linguistics* 8(1):35–57.

[15] Baltazani, M., & Jun, S-A. 1999. Topic and focus intonation in Greek. *Proc 14th ICPhS*, vol. 2, 1305-1308.

[16] Arvaniti, A. & Baltazani, M. 2005. Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora. In S.-A. Jun, (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: OUP. 84–117.

[17] Özge, U., Bozsahin, C. 2010. Intonation in the Grammar of Turkish. Lingua *120 (1)*: 132-175.

[18] Ipek, C., Jun, S-A. 2014. Distinguishing Phrase-Final and Phrase-Medial High Tone on Finally Stressed Words in Turkish. In *Proc. 7th Speech Prosody International Conference*, Dublin, Ireland.

[19] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 8 September 2018 from http://www.praat.org/.

[20] Talkin, David. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Palatal (eds.), Speech Coding and Synthesis, Elsevier Science B.V. 497–518.

[21] Octave community. 2013. GNU Octave 3.7+. Available online at: http://www.gnu.org/software/octave/index.html

[22] Grabe, E., Kochanski, G., Coleman, J. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50(3), 281–310.

[1] We do not expect any differences in c_0, c_1 or c_4 to be linguistically interesting because (i) normalisation of the utterance mean f_0 to 0 semitones means that c_0 will only indicate whether the ROI happens to be above or below the whole utterance mean; (ii) c_1 is generally positive, because the data are all continuation rises of various gradients; (iii) the c_4 term is only included in the model to improve its overall goodness-of-fit to the data, and is not linguistically interpretable. For brevity, we do not examine c_0, c_1 or c_4 further.

TO WHAT EXTENT DO FRENCH LISTENERS PERCEIVE DIFFERENCES IN ACCENT LOCATION WITHIN WORDS? AN EEG INVESTIGATION

Outhmane Rassili, Amandine Michelas, Sophie Dufour

Aix Marseille Université, CNRS, LPL, Aix-en-Provence
outhmane.rassili@univ-amu.fr, amandine.michelas@univ-amu.fr, sophie.dufour@univ-amu.fr

ABSTRACT

In this study, we provided a more in-depth examination of the ability of French listeners to perceive and use the location of accent in a discrimination task. Event-related potentials (ERPs) were recorded while participants performed a same-different task. Different stimuli diverged either in one phoneme (e.g., /**'ʒyʁi**/-/**'ʒyʁõ**/) or in accent location (e.g., /**'ʒyʁi**/-/ʒy**'ʁi**/). Although, participants reached 90% of correct responses, ERPs results indicated that a change in accent location was not detected while a change in one phoneme was detected in two time-windows, the first between 400 and 500 ms, and the second between 500 and 650 ms after target onset. It results that in the early moments of speech processing, stimuli that are phonemically identical but that differ in accent location are perceived as being strictly similar. The good performance observed in behavioral responses was interpreted in relation to attentional processes.

1. INTRODUCTION

Is /**'bo**pelo/ different from /bo**'pe**lo/? The answer is undoubtedly positive for listeners of lexical-stress languages such as Spanish for which accent location is lexically contrastive. However, what about for listeners of languages such as French for which accent doesn't not change the meaning of words[1]? In a well-known study, Dupoux et al. [1] have compared the performance of French and Spanish listeners in the discrimination of accentual patterns. They reported that French listeners had more difficulties than Spanish listeners in discriminating between non-words that differed in accent location (**'bo**pelo, bo**'pe**lo). They also reported that French listeners had more difficulties in discriminating between non-words that differed in accent location (**'bo**pelo, bo**'pe**lo) than between non-words that differed in one phoneme (**'bo**pelo,**'so**pelo), while Spanish listeners performed equally well on accentual and phonemic contrasts. The French listeners difficulties in perceiving accent location has been replicated in several subsequent studies when tested on non-words [e.g., 2-6] and both on Spanish [7] and French words [8]. Although the difficulties of French listeners in distinguishing stimuli that differ in accent location are manifest when their performance is compared to that of Spanish listeners, and perhaps more crucially when their performance is compared to their own performance on phonemic contrasts, it should be noted that in all of the studies, French listeners performed rather well, above the chance level (around 85% of correct responses; [see in 1,8]). Hence, even if their performance was not as high as that of listeners of languages with a lexical stress, and not as high as their performance on phonemic contrasts, French listeners showed a certain capacity to perceive and use accent location to discriminate between stimuli that are phonemically identical.

In this study, we pursued these investigations and provided a more in-depth examination of the ability of French listeners to perceive and use accent location in a discrimination task. More particularly, we conducted an electroencephalographic (EEG) study known to reflect cognitive processes as it unfolds over time[2]. This allowed us to examine the earliest moment at which French listeners perceive the difference in accent location, in contrast to classic behavioral tasks that allow examining only the final product of perception. We thus recorded ERPs while

[1] In French, a word can be accented or not depending on its location within the accentual phrase. For instance, the word *ballon* "ball" receives accent in ([*un petit ba'**LLON***]) "a little ball" because it is in final position within the phrase while it is unaccented in ([*un baLLON **'bleu***]) "a blue ball" because it is not in final position. Note also that the term " accent" was used each time we refer to the acoustic prominence of a syllable.

[2] Although a number of studies have used EEG to examine the representation and perception of accent in listeners of lexical-stress languages (e.g., [9-12]), to the best of our knowledge, only one EEG study [13] has been conducted to examine the French listeners' perception of differences in accent location. However, in this study, EEG measurements were conducted after a training session (with no EEG session before) in which participants learned to distinguish between tri-syllabic stimuli differing in accent location. Here, French listeners' perception was examined without implicit or explicit training which allowed us to examine "basic" perception.

French participants performed a same-different task. They first heard four stimuli that were strictly identical in respect to both their phonemic and accentual patterns, but produced by four different female speakers (e.g., /**'ʒyʁi**/-/**'ʒyʁi** /-/**'ʒyʁi** /-/**'ʒyʁi**/), and then heard a fifth stimulus, the target, always produced by a male speaker (see [14] for the same design with ERPs). The target was either the same as (identical condition; e.g., /**'ʒyʁi**/), or different from the first four stimuli (deviant conditions). In the deviant conditions, the target stimulus differed from the first four stimuli, either in the phonemic (phonemic deviant condition; e.g., /**'ʒyʁɔ̃**/) or in the accentual (accentual deviant condition; e.g., /ʒy**'ʁi**/) pattern. We measured the precise moment(s) at which phonemic and accentual discrimination occurs by comparing the ERPs in the standard vs. deviant conditions.

2. METHODOLOGY

2.1. Participants

20 right-handed French native speakers (15 females) between 18 and 33 years old (mean=22.4, SD=3.6) participated in the experiment. Since English (or German) are taught from middle school in the French educational system, they all reported having knowledge of one of these languages (18 in English and 2 in German). Some of them also reported having knowledge in Italian, Spanish, Russian, Portuguese or Mandarin Chinese, but none spent more than 3 months abroad, and all have been exposed to these languages after 10 years of age during middle school. No participant reported having any neurological, hearing or language impairment. Each participant gave written informed consent prior to the experiment, and they received either a course credit or 20€ for their participation. One participant was removed due to a high error rate (>50%).

2.2. Materials

Two bisyllabic French words *jury* /ʒyʁi/ "jury" and *juron* /ʒyʁɔ̃/ "swearword", differing only on the last phoneme were selected. Five French speakers (four women, one man) produced the two words within carrier sentences of two types. In the first type, target words were accented on the last syllable (e.g., *[On m'avait parlé] [d'un exigeant JU**'RY**] [qu'il était difficile] [de satisfaire]* "I was told about a highly demanding jury that it was difficult to satisfy". Because words bearing primary accent on the first syllable are not found in French, we used sentences in which target words are unaccented (e.g., *[On m'avait parlé] [d'un JURY exi**'geant**] [qu'il était difficile] [de satisfaire]*) but the speakers were explicitly instructed to produce the target words by making the first syllable prominent. To avoid coarticulation effects due to contextualized-speech, the two versions of each word (accented on the first syllable, accented on the second syllable) were then extracted from the carrier sentences and auditorily presented in isolation to each speaker. The speakers were instructed to imitate each version of the words. The recordings were made in a sound-attenuated room, and digitized at a sampling rate of 48,000 kHz. Each word in each of the two versions was then normalized at a level of 60 dB. Acoustic analyses were conducted to ensure that the words were produced with the expected accentual patterns. The length of the two syllables and the f0 movement associated with each syllable were measured (see Table 1).

Speaker	Word accented on the	1st syllable				2nd syllable			
		Syll. dur. (ms)	(Hz)	(Hz)	Pitch slope (%)	Syll. Dur. (ms)	(Hz)	(Hz)	Pitch slope (%)
Male speaker	1st syll.	264	115	153	33	176	98	85	-13
	2nd syll.	163	99	94	-5	275	100	244	146
Female speakers	1st syll.	343	214	304	42	179	224	208	-7
	2nd syll.	182	193	191	-1	286	187	447	141

Table 1. Acoustic properties of the target words in their two accentual patterns for the five speakers. For the 4 female speakers, means are reported.

2.3. Procedure

Stimuli were presented binaurally via headphones in an acoustically shielded room. Trials consisted in five stimuli each separated from the following one by 600 ms of silence. On any given trial, the first four stimuli were strictly identical in respect to both their phonemic and accentual patterns, but produced by the four female speakers. The order of the female speakers was counterbalanced across the trials. The fifth and final stimulus, the target stimulus, always produced by the male speaker, was either the same as (identical condition), or different from the first four stimuli (deviant conditions). In the deviant conditions, the target stimulus differed from the first four stimuli, either in the phonemic (phonemic deviant condition) or in the accentual (accentual deviant condition) pattern. The experimental conditions are illustrated in Table 2.

Condition	Female speakers	Male speaker	Number of trials	Expected response
Identical	/ʒy'**ʁi**/	/ʒy'**ʁi**/	48	Same
	/**'ʒyʁi**/	/**'ʒyʁi**/	48	Same

Accentual	/'ʒyʁi/	/ʒy'**ʁi**/	48	Different
dev.	/ʒy'**ʁi**/	/'ʒyʁi/	48	Different
Phonemic	/ʒy'**ʁɔ̃**/	/ʒy'**ʁi**/	48	Different
dev.	/'ʒyʁɔ̃/	/'ʒyʁi/	48	Different

Table 2: Illustration of the experimental conditions with the *jury* stimuli.

96 trials were presented per condition. So that there was an equal number of trials requiring a "same" and "different" response, 96 filler trials constructed exactly on the same model as the trials used in the identical condition were added, but were not included in the analyses. Each participant heard a total of 384 trials that were presented randomly. They were instructed to respond same when the target word was pronounced both with the same sounds and the same accentuation as the four preceding words, and to respond different when the target word differed from the preceding words either in its sounds or in its accentuation. Different responses were made using their dominant hand on a response box that was placed in front of them. Reaction Times (RTs) were measured from the onset of target stimuli. An inter-trial interval of 3000 ms elapsed between the end of one trial and the beginning of the next. Participants first received 6 practice trials and then three blocks of 128 trials separated by a break. No feedback was given to participants. The experiment lasted 1 h.

3. DATA ACQUISITION & ANALYSIS

The EEG signal was recorded from the scalp with a 64-channel BioSemi Active-Two AD-box. Individual electrodes were adjusted to a stable offset lower than 20 mV, and the EEG signal was digitized at 512 Hz. The EEG epochs, starting at 100 ms before the test stimulus and ending 700 ms after it, were averaged for each experimental condition and for each participant. The EEG data were filtered offline by a bandpass (1–30 Hz), re-referenced offline to the average of left and right mastoids, and corrected by a baseline of 100 ms before the onset of the target stimulus. Epochs were accepted under an artifact rejection criterion of 100 μV. All participants had a number of accepted trials superior to 80 for each experimental condition. For each participant, data from bad channels were interpolated.

An initial inspection of ERP waveforms showed differences between the three conditions at frontocentral sites, and on two negative components. Two topographical sites of 9 electrodes were used for the analyses: left frontocentral (F5, F3, F1, FC5, FC3, FC1, C5, C3, C1) and right frontocentral (F6, F4, F2, FC6, FC4, FC2, C6, C4, C2). To assess the timing of differential ERPs between the conditions, we conducted an ANOVA with the factors Condition (identical, phonemic deviant, accentual deviant) and Hemisphere (left, right). Two time-windows were selected around the peak amplitude of the two negativities identified: 400-500 ms and 500-650 ms. The Greenhouse-Geisser correction was applied and the corrected *p*values are reported.

4. BEHAVIOURAL RESULTS

ANOVAs were also performed on RTs and Error rates for the "different responses". The RT analysis was performed on the correct responses. An inspection of the data revealed that two short RTs (<300 ms) and three extremely long RTs (>10,000 ms) strongly deviated from the distribution, and were thus discarded from the analysis. A significant difference was found between the two deviant conditions on both RTs ($F(1,18)=9.95$, $p<.01$) and Error rates ($F(1,18)=4.44$, $p<.05$). Participants responded slower and with more errors in the accentual (Mean RTs=1158 ms; Mean Error rates=6.80%) than in the phonemic deviant condition (Mean RTs=1044 ms; Mean Error rates =1.32%).

5. ELECTROPHYSIOLOGICAL RESULTS

A main effect of the Condition was observed in the two time-windows with more negative values in the phonemic deviant condition than in both the accentual deviant and the identical conditions (see Table 3). No difference was observed between the accentual deviant and the identical condition. The interaction between Condition and Hemisphere was also significant, and was due to a greater effect of Condition in the right hemisphere in the first time-window, and to a greater effect of Condition in the left hemisphere in the second time-window. Figure 1 displays grand-average waveforms in each condition at FC1 electrode.

	400-500ms	**500-650ms**
Effect of Condition	**p<.001**	**p<.01**
Identical vs. Phon deviant	p<.01	p<.001
Identical vs. Acc deviant	p>.20	p>.20
Phon deviant vs. Acc deviant	p<.01	p<.01
Hemisphere*Condition	**p<.05**	**p<.05**
Left Hemisphere		
Identical vs. Phon deviant	p<.01	p<.001
Identical vs. Acc deviant	p=0.14	p>.20
Phon deviant vs. Acc deviant	p<.05	p<.05
Right Hemisphere		
Identical vs. Phon deviant	p<.01	p<.01
Identical vs. Acc deviant	p>.20	p>.20
Phon deviant vs. Acc deviant	p<.01	p<.01

Table 3. Summary of the ANOVA results.

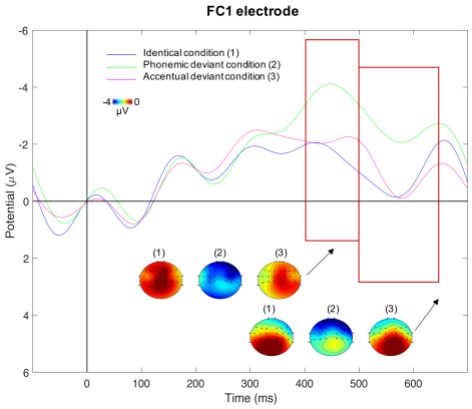


Figure 1. Grand-average waveforms at FC1 electrode in each condition. The maps show the topography of the grand-averaged ERPs for each condition at 400–500ms and 500–650ms.

6. DISCUSSION

ERP results clearly indicated that stimuli that involve a change in accent location are processed differently than stimuli that involve a change in one phoneme. In particular, we observed that a change in one phoneme was detected in two time-windows, respectively between 400-500 ms and between 500-650 ms after target onset, while a change in accent location was not detected. As a result, in the early moments of speech processing, stimuli that are phonemically identical but that differ in accent location are perceived as being strictly similar, and thus no difference between the accentual deviant and the identical conditions has been found. Because in French, accent location contrasts do not exist, our results are in perfect line with all of studies that report that listeners do not detect a phonemic change when the phonemic contrast under investigation does not exist in the native language [14]. For example, using the same design as in the present study, [14] showed that French listeners did not perceive the difference between the dental /da/ and the retroflex /Da/ that do not exist in French. Hence, it appears that when a contrast is not present in the native language, listeners fail to perceive it, at least when we probe processing in the early moments, and this seems to be true for both accentual and phonemic contrasts.

It remains however to explain why our participants succeeded in doing the task with a level of performance that reached 93% for the accentual contrast. As participants were asked to judge whether the target was identical or different from the previous stimuli, we believe that this good performance reflects only decisional processes linked to this particular task, and not unconscious automatic speech perception processes. Moreover, the two types of differences present in the stimuli were explained to the participants during the instructions, and thus we have likely attracted the participants' attention on the differences. Hence, if our reasoning is correct, French listeners would be able to use accent location in order to perform some tasks provided that their attention is directed on the accentual differences. The same conclusion can be drawn from a recent study [13] who examined the ability of French listeners to discriminate accent location contrasts after training. In particular, using ERPs measurement in an active deviant detection task, the authors reported evidence for discrimination on the P3 component known to reflect conscious and attentional processes. However, no evidence was found earlier in the processing, before the P3, and on the N2b component known to reflect automatic and unconscious processes. Further research should thus be done to compare French listeners performance in tasks that attract participant's attention on the differences under investigation, and in tasks that disengage participant's attention from these differences. Despite the good discrimination at a decisional level, behavioral measurements nonetheless revealed more errors and longer RTs in the accentual than in the phonemic deviant condition. This confirms the greater difficulty that French listeners have to use accent location information in comparison to phonemic information in order to perform a discrimination task, likely because it is not usual for French listeners to process accent location.

To conclude, despite a high level of performance in discrimination tasks, French listeners do not perceive the difference between stimuli like /**ˈʒyʁi**/ and /ʒy**ˈʁi**/ during unconscious and automatic speech perception processes. Although discrimination tasks provide important insights regarding conscious and decisional processes, they nonetheless tend to overestimate the ability to perceive and use accentual contrasts. Hence, it seems crucial to test French listeners' ability to perceive and use accent in a wide variety of tasks with experimental paradigms enable to probe both unconscious and conscious processes.

7. REFERENCES

[1] Dupoux, E., Pallier, C., Sebastian, N., Mehler, J. 1997. A destressing "deafness" in French?. *J. of Memory and Language 36*(3), 406-421.

[2] Dupoux, E., Peperkamp, S., and Sebastián-Gallés, N. 2001. A robust method to study stress 'deafness', *J. Acoust. Soc.* 110(3), 1606–1618.

[3] Dupoux, E., Sebastián-Gallés, N., Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': The case of French learners of Spanish. *Cognition 106*(2), 682-706.

[4] Peperkamp, S., Vendelin, I., Dupoux, E. 2010. Perception of predictable stress: A cross-linguistic investigation. *J. of Phonetics 38*(3), 422-430.

[5] Rahmani, H., Rietveld, T., Gussenhoven, C. 2015. Stress “deafness” reveals absence of lexical marking of stress or tone in the adult grammar. *PloS one* 10(12), e0143968.
[6] Carpenter, A. 2015. Phonetic training significantly mitigates the stress 'deafness' of French speakers. *Intern. J. of Linguistics* 7.
[7] Schwab, S., Llisterri, J. 2013. Does training make French speakers more able to identify lexical stress?. *Proceedings of the international symposium on the acquisition of second language speech*, *Concordia working papers in applied linguistics:* 5.
[8] Michelas, A., Esteve-Gibert, N., Dufour, S. 2018. On French listeners’ ability to use stress during spoken word processing. *J. of Cognitive Psychology 30*(2), 198-206.
[9] Domahs, U., Wiese, R., Bornkessel-Schlesewsky, I., & Schlesewsky, M. 2008. The processing of German word stress: evidence for the prosodic hierarchy. *Phonology* 25(1), 1-36.
[10] Molczanow, J., Domahs, U., Knaus, J., & Wiese, R. 2013. The lexical representation of word stress in Russian: Evidence from event-related potentials. *The mental lexicon* 8(2), 164-194.
[11] Zora, H., Schwarz, I.-C., & Heldner, M. 2015. Neural correlates of lexical stress: Mismatch negativity reflects fundamental frequency and intensity. *NeuroReport* 26(13), 791–796.
[12] Lu, S., Correia, S., Jeronimo, R., Vigário, M., & Frota, S. 2016. Revisiting stress 'deafness' in European Portuguese - an ERP study. In J. Barnes, S. Shattuck-Hufnagel, N. Veilleux, & Orgs (Eds.). *Proceedings of speech prosody 2016 Online publication.* ISSNhttps://doi.org/10.21437/SpeechProsody.2016 2333-2042.
[13] Schwab, S., Giroud, N., Meyer, M., & Dellwo, V. 2020. Working memory and not acoustic sensitivity is related to stress processing ability in a foreign language: An ERP study. *J. of Neurolinguistics 55*, 100897.
[14] Dehaene-Lambertz, G. 1997. Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport* 8(4), 919-924.

REALIZATIONS OF THE TRITONAL PITCH ACCENT IN PARAGUAYAN GUARANI

Sun-Ah Jun, Jian-Leat Siah, Marisabel Cabrera, & Hunter Johnson

University of California, Los Angeles
jun@humnet.ucla.edu, jlsiah@ucla.edu, isacabrera@ucla.edu, hunterjohnson@ucla.edu

ABSTRACT

A recent model of intonational phonology of Paraguayan Guarani proposed that Paraguayan Guarani has a tritonal pitch accent, which is typologically rare. The tritonal pitch accent was claimed to be HLH* realized over two syllables, as a falling tone on the pretonic syllable and a rising tone on the tonic syllable. This paper investigates the phonetic realizations of the pitch accent by systematically varying the stress location, the word length, and the inter-stress interval. The overall results corroborated the previous findings, but the alignment of the first H of HLH* was found to vary between the pretonic and the propretonic syllables depending on the stress location and the speaker's age/sex. Additionally, the medial L tone was often realized right after the onset of the tonic syllable, and LH* was more common when stress is word-initial, suggesting H+LH* would be a proper tonal category of the pitch accent.

Keywords: Paraguayan Guarani, tritonal pitch accent, tonal alignment, intonational phonology

1. INTRODUCTION

Paraguayan Guarani (henceforth Guarani) is a Tupi-Guarani language, and together with Spanish, is an official language of Paraguay. Guarani is a highly agglutinative language, and the default word order is Subject-Verb-Object (SVO), though the location of arguments can vary depending on informational properties as well as discourse status and the transitivity of the verb ([1, 2, 3, 4]).

For word prosody, Guarani has lexical stress, typically located at the end of a lexical item. But when a word includes more than one morpheme that has stress, only the last stress is realized ([1, 4, 5, 6]). For example, stress on the verb *ho* 'go' remains the same when the unstressed prospective aspect/modal suffix /-ta/ is added as in [o-hó-ta], (3p-go-PROSP) '(S)he will go'. But, when the stressed, desiderative modal suffix /-sé/ is added to the verb, *ho*, stress shifts to the suffix as in [o-ho-sé] (3-go-DES) '(S)he wants to go'. Stress is not marked orthographically when it falls on the word-final vowel or on a nasalized vowel, which is marked with a tilde (*porã* [põr̃ã] 'well'). Otherwise, stress is orthographically marked by an acute accent on the stressed vowel (e.g., *ñandu* 'spider' vs. *ñandúre* 'spider-at'; *porã* 'well' vs. *porãta* 'well-future'). Acoustically, stressed syllables are longer and stronger in intensity than unstressed syllables.

Several studies [5-10] have examined the intonation pattern of Guarani analyzed in the Autosegmental-metrical framework [11-13], either to investigate how focus is prosodically realized [5-7] or how prosodic constituents are related to morphosyntactic constituents [8-10]. Clopper and Tonhauser [5, 7] proposed that two-word (Subj + Verb) declaratives in Guarani have two tonal contours: a hat contour and a two-peak contour. The hat contour is formed by a rising pitch accent (L*+H), followed by a falling pitch accent (H+L*), and the two peak contour is formed by a sequence of two rising pitch accents. They also found a high plateau between the two pitch accents in each contour type, and explained the high plateau in the hat contour by interpolation between the two adjacent H tones (i.e., L*+H H+L*), but did not provide any explanation for the high plateau in the two-peak contour (i.e., L*+H L*+H). It was suggested that pitch accent categories in Guarani should be further explored. Burdin and collaborators [6] examined focus prosody of a noun phrase (Noun + Adjective) in Guarani and confirmed the two contour types proposed in [5, 7], but claimed that the tone-syllable alignment in the pitch accents was not contrastive, revising the rising pitch accent to "LH" and the falling pitch accent to "HL", i.e., without marking the L tone with a star (*).

Jun and Zubizarreta [8-10] examined intonation of complex sentences in Guarani, which varied in morpho-syntactic structures as well as the length of the words and sentences. They claimed that Guarani has only one pitch accent, a tritonal HLH*, and an Accentual Phrase (AP) whose edges are marked by a H boundary tone. An AP in Guarani typically includes one word, and like other AP languages (e.g., Japanese, Korean, Bengali, Georgian; see [14, 15]), it can include more than one word when the words are short and closely related syntactically and semantically. They further found that, unlike other AP languages, the AP in Guarani also marks a syntactic constituent, i.e., a complex predicate, which includes a Verb and a Low-modifier [8-10, 16, 17], regardless of the number of words in the predicate.

When an AP includes more than one word, the pitch accent is realized on the last stressed syllable of the AP. Since an AP begins and ends with a H tone, the tonal pattern of an AP is /H HLH* Ha/ ('Ha' is an AP-final H boundary tone; 'a' refers to an AP; the AP-initial H is also a boundary tone, so it could be labeled as 'aH'). Jun & Zubizarreta [9] claimed that the HLH* pitch accent is realized as a falling tone on the pretonic syllable and a rising tone (LH) on the tonic syllable. But when the AP is the last AP of an Intonational Phrase (IP) ending with a L% in declaratives, the pitch accent is often realized as a falling tone on the pretonic syllable and a low tone on the tonic syllable. This would correspond to the H+L* pitch accent proposed in [5, 6, 7]. In Jun and Zubizarreta's model, the falling pitch accent is analyzed as an allotone of HLH*, i.e., HLH* can be realized as HLL* when it is immediately followed by a L%. By analyzing the pitch accent being HLH* and APs having H boundary tones, a high plateau between two pitch accents observed in [5, 6, 7] was explained by the interpolation between the H tone at the edge of the tritonal pitch accent and the H boundary tone of an AP, and also between the H boundary tone of adjacent APs.

Thus, these two existing models differ whether a pitch accent is a simple rising tone (LH) or a falling-rising tone (HLH) when the pitch accent is not IP-final. Because the falling-rising pitch accent is realized over two syllables, i.e., falling tone on the pretonic syllable and a rising tone on the tonic syllable, while the rising pitch accent proposed in [5, 6, 7] is realized on the tonic syllable, the main difference between these two models is whether there is a falling tone on the pretonic syllable or not.

The goal of the current paper is to investigate how a pitch accent is realized in Guarani, examining the f0 contours on the pretonic syllables as well as the tonic syllable when the pitch accent is not IP-final.

2. PRESENT STUDY

2.1. Methods and Procedures

The participants included one male (M1 in his late 30s) and three females (F1 in her mid 30s, F2 in her early 50s, F3 in her early 60s). M1 and F1 are from Coronel Oviedo, F2 is from Concepción, and F3 is from Asunción, Paraguay. All speakers are bilingual in Guarani and Spanish. M1 and F1 spoke primarily Guarani growing up and learned Spanish later but continue to use Guarani on a daily basis. The data from M1 and F1 were collected in Paraguay using a digital recorder, and the data from F2 and F3 were collected over *Zoom.*

Each speaker produced 84 declarative sentences varying in word orders and syntactic structures, designed to examine the tone-syllable alignment of the pitch accent. The target words varied in the location of stress and the number of unstressed syllables. The list of sentences was written in Guarani orthography and presented to the speakers in a randomized order in two repeated sets using Microsoft PowerPoint, thus collecting 8 tokens per utterance across all speakers.

To investigate the tone-syllable alignment of the pitch accent contour, two sub-datasets (33 sentences) were examined, where the target word is sentence-medial. The target word had two types of stress location, final or penultimate, and differed in the number of unstressed syllables (s), from zero to four, before the stressed syllable (S), i.e., S, sS, ssS, sssS, ssssS in the final stress condition and Ss, sSs, ssSs, sssSs, ssssSs in the penultimate stress condition.

All the recorded sentences were segmented and annotated in *Praat*. For all the target words, the vocalic intervals were annotated whether a vowel is stressed ('V') or unstressed ('v'). The location of unstressed vowel relative to the stressed vowel was labeled as "v#' (i.e., v1, v2, v3, and v4), with '#' indicating how far away an unstressed vowel is from the stressed vowel in terms of syllable count: '1' being the closest to 'V' and '4' being the furthest. The onset consonant of the stressed syllable, if present, was also annotated with the label 'C'.

For all the annotated intervals, pitch and temporal measurements were obtained automatically using a Praat script. Mean f0 (Hz) values and the time points for the maximum and the minimum f0 of the target vowels were collected. Pitch track errors (e.g., pitch halving and doubling) were manually corrected whenever possible. Pitch track errors that could not be resolved (e.g., missing pitch value) were excluded from further analysis. All f0 values in Hz were then converted to semitone using the formula: F0 (semitone) = $12*(\log_2 (f_1/f_2))$, where f_2 is the mean f0 value in Hz for each speaker. The time point marking the onset of a tonic syllable was also collected. The entire dataset was then imported into R software [18] for statistical analyses and data visualization.

3. RESULTS

Our data confirmed the findings of [9] that Guarani has a tritonal pitch accent, a falling tone immediately followed by a rising tone, with the f0 peak on the stressed syllable. Figure 1 shows an example of the HLH* tritonal pitch accent, with H* realized toward the end of the stressed syllable (word-final for the first two words) and the first H realized at the beginning of the pretonic syllable. The middle L tone is between

the pretonic syllable and the tonic syllable. The word/AP-initial syllable(s) show a high plateau until the first H tone of the tritonal pitch accent, and this high tone can be similar to that of the preceding H* (as in the second word/AP in Figure 1), or can be slightly lower than that of the preceding H* (as in the final word/AP in Figure 1), or can be reset to be higher than that of the preceding H* (as in the second word/AP in Figure 2). Figure 2 shows that H* of the tritonal pitch accent can be downstepped (!H*) when the word is sentence-medial or final. (In the final position, it can also be L* before L%, as in Figure 1). Figure 2 also shows that the f0 on the post-tonic syllable *ra* is higher than the preceding !H* tone. This post-tonic high tone is AP-final Ha, proposed in [9].

Figure 3 shows the mean f0 values (in semitone) for each vowel of the sentence-medial target word (except for initially stressed words) up to the stressed syllable for each speaker when stress is final (top) and penultimate (bottom). Thus, these plots are a rough estimation of f0 contours of the target word. This figure shows that the mean f0 values are similar across v2, v3, and v4, reflecting the high plateau between the beginning of the word and the propretonic syllable (v2). (In the penultimate stress plots (bottom), the mean f0 of the word-initial vowels is higher than that of the following vowel due to the onset consonant in the target word, i.e., microprosody.) It also shows that the mean f0 of the stressed syllable (V) is higher than that of v1 when stress is final than when it is penultimate, which is expected because the star-tone of the sentence-medial pitch accent is often !H*. Also, the mean f0 of v1 can be similar to or slightly higher than that of V when the f0 difference in the falling tone on v1 is small and V carries !H* as shown in Figure 2 (*gua kué*). But for younger female speakers (F1 and F2), the mean f0 on the stressed vowel was substantially lower than that of v1 for some sentences when stress is penultimate (e.g., those in the sSs and ssSs conditions). This is because they sometimes produced L* or L+>H* similar to Spanish prenuclear pitch accent. The influence of Spanish on their Guarani is shown more clearly in the following section where the tone-syllable alignment of the first H of HLH* is shown.

3.1. Alignment of the first H of HLH*

[9] claimed that the tritonal pitch accent starts with a falling tone on the pretonic syllable followed by a rising tone on the tonic syllable. To find out whether the H of the falling (HL) starts on the pretonic syllable (v1) or the propretonic syllable (v2), the mean f0 values of v2 were compared with the maxf0 and min-f0 of v1. If v1 carries a HL tone, a max f0 would occur at v1 onset and a min f0 at v1 offset. If the H of HL

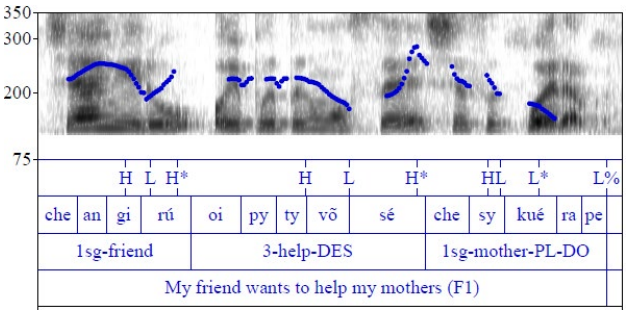


Figure 1: The tonic syllable carries a high tone (H*) when not sentence-final. The pretonic syll begins with H. The L is between the pretonic and the tonic syllable.

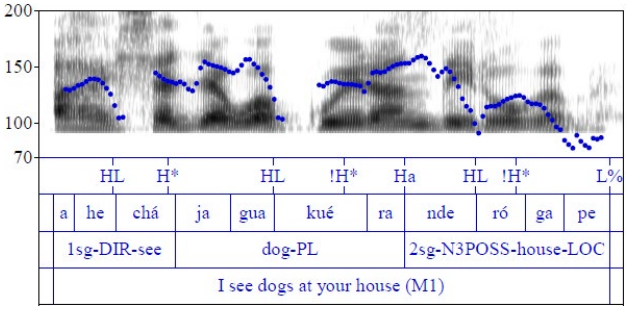


Figure 2: The tritonal pitch accent can be HL!H* in sentence-medial or -final. Ha is higher than !H*

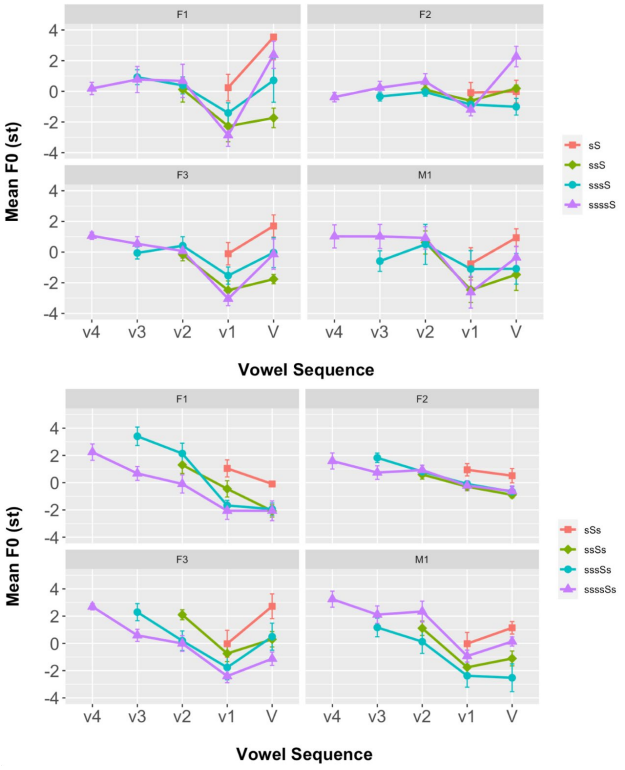


Figure 3: Mean f0 values (in semitone) of each unstressed vowel up to the stressed vowel (V) when stress is final (top) and penultimate (bottom) for each speaker.

starts from v2, v1's maxf0 would be significantly lower than v2 mean, or roughly in the middle of v2 mean and v1's minf0. But if the falling tone starts at

the beginning of v1, as claimed in [9], v2's mean f0 would be similar to or slightly higher than v1's maxf0 (reflecting f0 declination between v2 and v1). Figure 4 shows f0 values (in semitone) on v2_mean, v1_max and v1_min for each speaker when stress is final (top) and penultimate (bottom).

As shown in Figure 4, the alignment of the first H varied across speakers and the stress location. In the final stress condition, v1_max is not significantly different from v2_mean for the speakers F3 and M1, suggesting the falling starts on the pretonic syllable. However, for F1 and F2 (the younger female speakers), the v1_max was significantly lower than v2_mean, suggesting that the falling starts earlier than the pretonic syllable. In the penult stress condition, v1_max was significantly lower than v2_mean for all speakers, but especially so for F1, the youngest female. She sometimes did not produce the first H of the tritonal pitch accent, reducing it to a bitonal pitch accent L+H* or L+>H*, similar to Spanish.

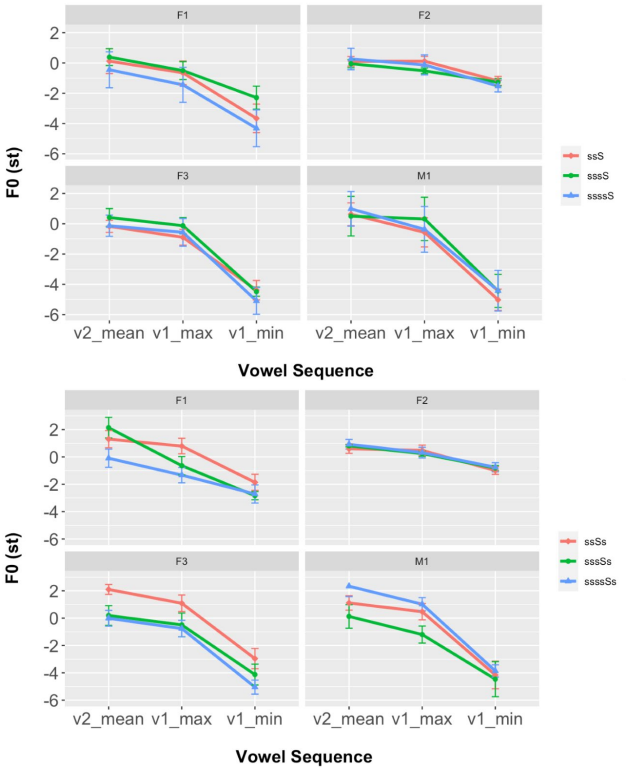


Figure 4: F0 (in semitone) for the v2 mean, the maxf0 of v1, and the minf0 of v1 for each speaker when stress is final (top) and penultimate (bottom).

3.2. Alignment of the medial L of HLH*

The tone-syllable alignment of the medial L tone varied between the pretonic and tonic syllable as claimed in previous studies [5-10], but this variation seemed to be influenced by the intervening consonant between the pretonic and tonic syllables. When the intervening consonant was a sonorant, L was often aligned right after the onset of the tonic syllable. See Figure 5 (those tokens where the stressed syllable had L* are not included). Furthermore, when stress is word-initial, a rising f0 was often observed, reaching its peak at the end of the syllable, suggesting that the medial L tone is a unit with the following H* (LH*), instead of the preceding H (HL). Therefore, a proper tonal category of the tritonal pitch accent would be H+LH*.

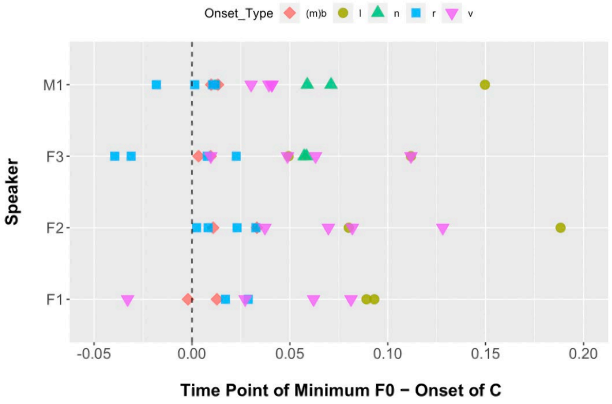


Figure 5: The alignment of the L tone relative to the onset C of the tonic syllable (the vertical dotted line).

4. DISCUSSION AND CONCLUSION

The current data confirmed [9] that Guarani has a tritonal pitch accent, and further demonstrated that the alignment of the first H is variable across speakers and stress location. When stress is word-final, the first H started at the onset of the pretonic syllable, only for the male and the oldest female speakers (M1 and F3), but when stress is penultimate, all speakers tended to start the H at the propretonic syllable. Furthermore, the younger female speakers sometimes produced no initial H and produced a bitonal pitch accent L+H* or L+>H* (delayed f0 peak on the posttonic syllable), which is similar to the prenuclear pitch accent in Spanish. Because all speakers participated in the current study are Spanish-Guarani bilinguals and the elicitation was conducted in Spanish, the influence of Spanish intonation to Guarani intonation might be expected, but given that young female speakers showed more variable alignment of the initial H and more Spanish-like pitch accent, sound change might be currently in progress in Guarani intonation. It is also possible that a H+LH* tritonal pitch accent might be in the process of being simplified to LH* because the HLH tritonal pitch accent is very rare cross-linguistically; So far, the most common tritonal pitch accent type is LHL, found in several varieties of Spanish [19-23], Italian [24], Serbo-Croatian [25, 26], and Brazilian Portuguese [27]. Data from more speakers should be examined to confirm the current findings and suggested interpretations.

5. REFERENCES

[1] Gregores, E., Suárez, J. A. 1967. *A Description of Colloquial Guaraní.* Mouton, The Hague.

[2] Velázquez-Castillo, M. 1995. Noun incorporation and object placement in discourse: The case of Guaraní, In: Downing, P., Noonan, M. (eds), *Word Order in Discourse*, Amsterdam and Philadelphia: John Benjamins, 555–79.

[3] Estigarribia, B. 2020. *A Grammar of Paraguayan Guarani*. UCL press.

[4] Tonhauser, J., Colijn, E. 2010. Word order in Paraguayan Guarani, *International Journal of American Linguistics*. 76(2), 255-288.

[5] Clopper, C. G., Tonhauser, J. 2013. The prosody of focus in Paraguayan Guaraní, *International Journal of Amercian Linguistics*, 79, 219-251.

[6] Burdin, R., Phillips-Bourass, S., Turnbull, R., Yasavul, M., Clopper, C., Tonhauser, J. 2015. Variation in the prosody of focus in head- and head/edge-prominence languages", *Lingua* 165, 254-276.

[7] Clopper, C. G., Tonhauser, J. 2011. On the prosodic coding of focus in Paraguayan Guarani, *Proceedings of the Twenty-eighth West Coast Conference on Formal Linguistics*, 249-257.

[8] Zubizarreta, M.L., Jun, S.-A. 2021. Morpho-syntactic and prosodic cues in Complex Predicates. A Case Study. A talk presented (virtually) at the Workshop 6: Complex predicates across languages: Variation and acquisition, Societas Linguistica Europaea, Athens, Greek.

[9] Jun, S.-A., Zubizarreta, M.L. 2022. Paraguayan Guarani: Tritonal pitch accent and Accentual Phrase. *Proceedings of INTERSPEECH 2022*, Incheon, Korea, 5303-5307.

[10] Zubizarreta, M.L. 2022. The verbal phrase in Paraguayan Guarani: A case study on the role of prosody in linearization, *Languages* 7, 221.

[11] Pierrehumbert, J. 1980. *The Phonology and Phonetics of English intonation*. PhD dissertation. MIT.

[12] Beckman, M., Pierrehumbert, J. 1986. Intonational structure in Japanese and English, *Phonology Yearbook* 3, 15-70.

[13] Ladd, R. D. 1996/2008. *Intonational Phonology*. Cambridge University Press.

[14] Jun, S.-A. (ed.), 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press.

[15] Jun, S.-A. (ed.), 2014. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford University Press.

[16] Zubizarreta, M.L, Pancheva, P. 2017a. A Formal characterization of Person-based alignment: The case of Paraguayan Guarani, *Natural Language and Linguistic Theory* 35, 1161–1204.

[17] Zubizarreta, M.L., Pancheva, P. 2017b. A note on the syntax of possession in Paraguayan Guarani. In: Fernández-Soriano, O., Castroviejo, E., Pérez-Jiménez, I. (eds), *Boundaries, Phases, and Interfaces*, John Benjamins, 85–104.

[18] R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria. Available at http://www.R.project.org/

[19] Prieto, P., D'Imperio, M., Gili Fivela, B. 2005. Pitch accent alignment in Romance: primary secondary associations with metrical structure, *Language and Speech* 48(4), 359-396.

[20] Gabriel, C., Feldhausen, I., Pešková, A., Colantoni, L., Lee, S.-A., Arana, V., Labastía, L. 2010. Argentinian Spanish intonation. In: Prieto, P., Roseano, P. (eds) *Transcription of intonation of the Spanish languages,* 285-317.

[21] Gabriel, C., Feldhausen, I., Pesková, A., Colantoni, L., Lee, S-A., Arana, V., Labastía, L. 2010. Argentinian Spanish intonation. In: Prieto, P., Roseano, P. (eds), *Transcription of Intonation of the Spanish Language*. Lincom Europa, 285-317.

[22] García, M. 2011. *The intonational patterns of the Peruvian Amazonian Spanish (PAS)*, MA Thesis. University of Stony Brook.

[23] Froemming, B., Rao, R. 2021. The tritonal pitch accent in the broad focus declaratives of the Spanish spoken in Cuenca, Ecuador: An acoustic and sociolinguistic analysis, *Estudios de Fonética Experimental (Journal or Experimental Phonetics*), 119-140.

[24] Gili Fivela, B. 2004. *The phonetics and phonology of intonation: The case of Pisa Italian*. Ph.D. Dissertation, Scuola Normale Superiore, Pisa.

[25] Smiljanic, R. 2004. *Lexical, pragmatic, and positional effects on prosody in two dialects of Croatian and Serbian: An Acoustic Study*. London: Routledge.

[26] Yu, K. M. 2008. *The prosody of second position clitics and focus in Zagreb Croatian*, UCLA, M.A. Thesis.

[27] Ferreira, L. 2008. *High initial tones and plateaux in Spanish and Portuguese neutral declaratives: Consequences to the relevance of F0, duration and vowel quality as stress correlates.* PhD Dissertation. University of Illinois at Urbana-Champaign.

Perceived naturalness of accents in noun phrases by mono- and bilingual listeners of Russian

Sabine Zerbian[1] & Kristina Barabashova[2]

University of Stuttgart[1], Saarland University[2]
sabine.zerbian@ifla.uni-stuttgart.de[1], kristina.barabashova1@gmail.com[2]

ABSTRACT

We tested the acceptability of different accent patterns in adjective+noun phrases with contrastive focus on the adjective (A_{CF}+N) by mono- and bilingual listeners of Russian in a perception experiment. Both naturally produced as well as lab-recorded stimuli were presented auditorily, embedded in a discourse context which evokes semantic contrastive focus on the adjective in a noun phrase (e.g., *the* $GREEN_{CF}$ *hat*). Listeners had to evaluate how natural the presented audio stimuli sounded to them. The accent patterns varied with respect to accent on the noun, on the adjective or on both. Moreover, the noun phrases occurred as either the first or second noun phrase in the sentence.

Results show that both mono- and bilingual listeners show a sensitivity to prosody in specific information structural contexts in lab-recorded speech. In natural speech and concerning the position of occurrence, only the bilingual listeners show a preference. Bilingual listeners' more stable judgments across phonetically varied stimuli might be attributed to linguistic experience.

Keywords: accent, perception, Russian, adjective Heritage Russian, noun phrases, focus.

1. INTRODUCTION

In Russian, both word order and pitch accents can be used to indicate information structure in a sentence. Specifically, a pitch accent on a constituent (H*L; indicated by capital letters) can indicate narrow focus on this constituent, as in (1a vs 1b) (Jasinskaja 2014).

(1) Prosodic expression of focus in Russian
(a) Marina slušala MUZYKU.
(b) Marina SLUŠALA muzyku.
'Marina listened to music.'

Within a noun phrase, the location of a pitch accent can also indicate narrow focus (Jasinskaja 2014). For example, a pitch accent on the adjective, as in (2a), indicates narrow contrastive focus on this constituent (word order in split constructions as in (2b) is another option) (Sekerina & Trueswell 2011:282).

(2) Narrow focus in noun phrases in Russian
a. Položite KRASnuju zvezdočku v Poziciju 4.
put red.ACC.FEM star.ACC.FEM in Position 4
b. KRASnuju položite zvezdočku v Poziciju 4.
red.ACC.FEM put star.ACC.FEM in Position 4
'Put the RED star in the Position 4.'

Zerbian et al. (2022) investigated the prosodic realization of contrastive adjective focus in natural speech of mono- and bilingual speakers of Russian, using data from the RUEG corpus (Wiese et al. 2019). The bilingual speakers were heritage speakers of Russian. Heritage languages are "culturally or ethnolinguistically minority languages that develop in a bilingual setting where another sociopolitically majority language is spoken" (Montrul 2016:2).

For monolingual speakers of Russian, a pitch accent on a contrastively focused adjective was tentatively confirmed, at least in informal communicative situations (note though that the number of overall productions was low for this speaker group) (Zerbian et al. 2022).

The data produced by heritage speakers of Russian in the U.S. showed that next to accents on the adjective they more frequently produced double accents (i.e., both on adjective and noun) on A_{CF}+N, both across formal and informal communicative situations (Zerbian et al. 2022).

The current research therefore sets out to investigate by means of naturalness ratings how mono- and bilingual listeners of Russian perceive different accent patterns in A_{CF}+N. Both lab-recorded and naturally produced data were used as stimuli.

The study is motivated by a scarcity of perception studies concerning heritage language prosody. Whereas speech production is often tested in heritage language phonology, perception or comprehension is researched less often (Polinsky & Scontras 2019, footnote 6; but see Sekerina & Trueswell 2011, Zerbian 2012; and the short summary in Polinsky 2018: 158-162).

Prior work on the processing of prosody and word order in mono- and bilingual Russian listeners has attested only a weak effect of contrastive accent (Sekerina & Trueswell 2011) and found instead that it is relevant whether the semantic contrast is linguistically explicit or not, i.e., whether the noun phrase in question constitutes the first or the second

in an utterance (Sekerina & Trueswell 2011). Only in the latter case is the contrast linguistically explicit. Therefore, position of occurrence was also included in our perception study as a potentially relevant factor for perceived naturalness next to the accent pattern itself.

2. RESEARCH QUESTIONS AND HYPOTHESES

The following research questions were investigated:

- RQ1: Do mono- and bilingual listeners rate different accent patterns in A_{CF}+N differently with respect to naturalness?
- RQ2: Does the position (i.e., whether the noun phrase is mentioned first or second) lead to a difference in the ratings of accent patterns?

We expect the rating of A_{CF}+N to depend on …

- Accent location: Adjective accents are rated high on the natural scale by both mono- and bilingual listeners and clearly different in acceptability from accents on nouns.
- Bilingualism of listeners: Mono- and bilingual listeners differ in their naturalness ratings. The prediction based on our findings in production is that double accents (i.e., on both adjective and noun) are rated as more natural more often by bilingual listeners than by monolingual ones.
- Explicit linguistic contrast: Double accents are rated as more natural if they occur on the first A_{CF}+N than on the second A_{CF}+N, because only in the latter is the contrast linguistically explicit.

3. METHODOLOGY

3.1 Method

The research questions were investigated by means of an online experiment in which listeners were asked to evaluate the naturalness of a spoken noun phrase within a given context.

The study consisted of two experimental parts. Experimental part 1 (RQ1) elicited naturalness ratings for A_{CF}+N with differing accent patterns, namely an accent on the adjective, on the noun or on both (double accent). In example (3), the A_{CF}+N in question is *синюю машину* "blue car".

(3) (speaker code: USbi10FR_fsR)
Èt-a bela-ja mašin-a vreza-l-a-s'
this-F.NOM white-F.NOM car-NOM crash-PST-F-REFL
v ***sin-ju-ju*** ***mašin-u.***
into blue-F.ACC car-ACC
'This white car crashed into a blue car.'

Experimental part 2 (RQ2) explored the role of explicit linguistic contrast for naturalness ratings of A_{CF}+Ns as first and second occurrence in an utterance respectively (*Белая машина* and *синюю машину* in (4), in bold).

(4) (speaker code: USbi33FR_isR)
Bela-ja ***mašin-a*** *rezko ostanovi-l-a-s'*
white-F.NOM car-NOM abruptly stop-PST-F-REFL
i vresa-l-a-s' *v* ***sin-juju*** ***mašin-u.***
and crash-PST-F-REFL into blue-F.ACC car-ACC
'The white car stopped and crashed into the blue car.'

3.2 Procedure

The experiment was implemented in Gorilla ®. The participants accessed the experiment online via a link and were presented introductory information on the experiment, including informed consent. In a training session the participants got acquainted with the task. Participants saw a written sentence (in Cyrillic) on the screen in which an adjective followed by a noun was marked in bold and italics. A drawing showing the object described in the noun phrase was placed in bottom-middle position of the slide for illustration purposes. On pressing a button, participants could listen to the noun phrase in bold produced by a speaker. They could repeat listening up to 5 times. On the following slide they had to rate how natural the phrase sounded in this context by assigning a number (1 = not at all natural, 7= very natural). After providing their rating, they were automatically presented with the next item.

In order to make listeners acquainted with the task, the training session progressed from contexts and target nouns being read out, then lab-recorded noun phrases embedded in written contexts, and lastly noun phrases taken from the corpus (thus different speakers) embedded in written contexts. Three examples each were presented during the training session.

After the training session, the experiment started. At the end, participants had to fill in a questionnaire on their linguistic background.

3.3 Stimuli

For both experimental parts we used original productions of A_{CF}+N from the RUEG corpus (see Zerbian et al. 2022 for details), which describe an accident involving two differently coloured cars (which implies a semantic contrastive focus on the adjective).

Only noun phrases were selected in which the adjective was a colour adjective (rather than an inherently deictic adjective such as "other, second"). In addition, only noun phrases were selected which had no hesitation or pronunciation deviation and were of acceptable or good quality (n=30; see below).

Accent on	1st position	2nd position	total
adjective	3	4	7
noun	2	3	5
both	8	7	15

The 27 noun phrases are from 11 different speakers (9 bilingual, 2 monolingual; 2 male, 9 female).

All noun phrases selected have been taken from sentences which contained the two A_{CF}+N phrases that were contrasted, thus the contrast was linguistically explicit in the sentence.

The noun phrases were cut out and presented with the transcribed original contexts, leaving out repetitions, hesitations, and grammatical deviations in order not to create any interference from other linguistic features. Corpus samples were treated for noise, and equalized for intensity (loudness) and tempo, using the digital audio editor in Audacity.

Next to samples from naturally occurring speech, listeners were also presented with lab-produced recordings of A_{CF}+N phrases, mirroring the three accent patterns on adjective, noun and both, in freely created contexts, as in (5).

(5)
Sin-jaja šljap-a mn-e nravit-sja bol'še
blue-F.NOM hat-NOM I-DAT like-PRS.REFL more
čem zelën-aja.
than green-F.NOM
'I like the blue hat more than the green hat.'

The first part contained 14 items from natural speech. In order to increase and adjust the number of data points, some items were repeated up to three times forming the equivalent number of items across three accent types (9 items in each group). In total, the first experiment comprised 27 stimuli (9 x 3 accent types).

The second part contained 22 items from natural speech. Again, to increase the number of data points, all items were repeated up to three times. In total, the second experiment part included 36 items (9 x 2 accent types x 2 positions).

In addition, 30 lab-recorded stimuli (produced by one single female speaker) were included in the first part (5 x 2 context structure x 3 accent patterns). The second part comprised 20 lab-recorded stimuli (5 x 2 accent patterns x 2 positions).

Altogether, each participant gave 113 naturalness judgements (27 + 36 + 30 + 20).

3.3 Participants

Participants were recruited via the platforms Prolific and Mechanical Turk. Requirements specified via the platform included U.S. or Germany as location, Russian as L1, Russian and English or German as fluent languages. Altogether, we recruited 70 participants. Of those, 25 had to be excluded due to not matching the linguistic requirements needed for the study (e.g., using Russian as a language at home). Of the remaining 45 participants, 16 were raised monolingually, 29 were bilingual. In the following, we concentrate on those 12 participants which we consider heritage speakers as they were either born to Russian speaking parents in the U.S. or Germany or immigrated to the respective countries when they were younger than 5 years (17 participants immigrated at > 20 years of age).

Among heritage Russian listeners, four participants belong to the age group 26-35 years, and eight participants were 18-25 years. Monolinguals listeners included seven participants 18-25 years, seven participants 26-35 years, and two participants 51-61 years. Heritage Russian listeners specified the usage of English (6), German (2), and Ukrainian (1) at home. For other purposes, such as work/study and communication with friends, English was selected by 10 participants, German by five participants, and Russian by one participant. Six heritage speakers reported that Russian was a part of their compulsory or additional education. Among monolingual speakers, one participant noted the usage of Chuvash at home. For the purposes of work/study and communication with friends, monolinguals use English (8) and German (2). For 14 monolinguals, Russian was a part of the compulsory or additional education.

4. RESULTS

Fig. 1 shows the naturalness ratings for different accent patterns in the context of contrastive adjective focus for the lab-recorded stimuli by bilingual heritage and monolingual listeners.

The y-axis shows the assigned values (1= not natural, 7= very natural) to the different accent locations on the x-axis (accent on adjective left/blue, double accent middle/grey, accent on noun right/red). Boxes show the quartiles of the data, whiskers show the rest of the distribution, points show outliers. Horizontal lines in the boxes refer to the median (Waskom 2021).

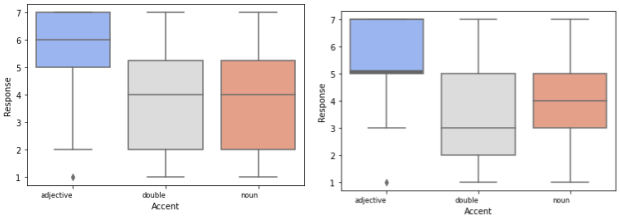


Figure 1: Ratings for lab-recorded stimuli by heritage listeners (left) and monolingual listeners (right)

In the context of contrastive adjective focus, both speaker groups rate an accent on the adjective (heritage: mean=5.4, mono: mean=6.0) as clearly more natural than a double accent (heritage: mean=4.0, mono: mean=3.6) or an accent on the noun (heritage: mean=3.9, mono: mean=4.2).

Figure 2 shows the ratings for the stimuli taken from the corpus.

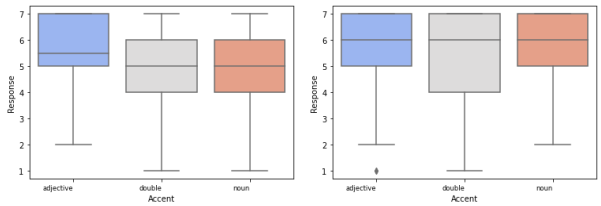


Figure 2: Ratings for natural stimuli by heritage listeners (left) and monolingual listeners (right)

For the heritage listeners, accents on adjectives (mean=5.3) are rated higher on average than double accents (mean=4.9) or accents on nouns (mean=4.8), though less clearly so than in the lab-recorded stimuli.

For monolingual listeners, the difference in naturalness ratings depending on the different accent locations disappeared. All accent patterns are rated as similarly highly natural.

Figures 3 and 4 shows the naturalness ratings depending on whether the noun phrase occurred first or second in a sentence. Figure 3 shows the lab-recorded stimuli, Figure 4 the natural stimuli from the corpus.

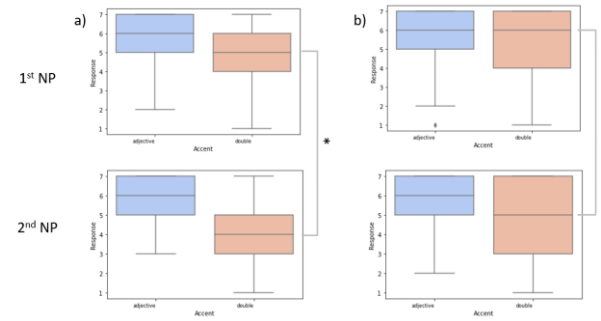


Figure 3: Ratings for pre-recorded stimuli by heritage listeners (left) and monolingual listeners (right)

For both monolingual and heritage listeners, accents on adjectives are judged more natural than double accents both in 1st and 2nd position ($p<0.001$). However, heritage listeners rate double accents significantly more natural in 1st position (mean=4.7) than in 2nd (mean=3.9; $p<0.05$). For monolingual listeners, the position does not matter ($p>0.05$).

The results for natural stimuli are shown in figure 4. In 1st position, double accents (mean=5.3) are as natural as adjective accents (mean=5.3) for heritage listeners ($p>0.05$). However, adjective accents (mean=5.8) are still more natural than double accents (mean=5.2) for monolingual listeners ($p<0.05$). In 2nd position, adjective accents are clearly judged more natural than double accents by both groups ($p<0.001$). Again, double accents are rated higher in 1st position (mean=5.3) than in 2nd (mean=4.5) by heritage listeners only ($p<0.001$).

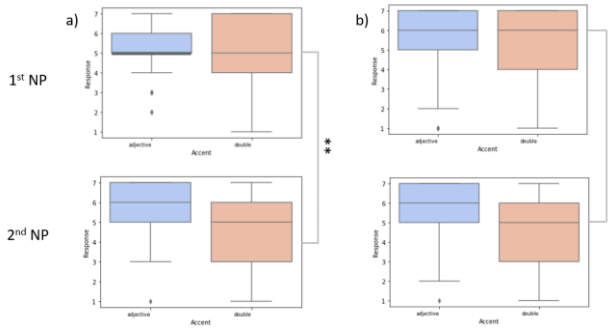


Figure 4: Ratings for natural stimuli by heritage listeners (left) and monolingual listeners (right)

5. DISCUSSION

The results of the lab-recorded stimuli are in line with the literature and our expectations for both listener groups, namely that adjective accent is judged significantly more natural than an accent on the noun or double accent. Thus, both listener groups have shown to be sensitive to accent and its relation to information structure in Russian.

For the stimuli taken from natural productions, all accent patterns sounded equally (highly) natural, at least to monolingual listeners.

The actual naturalness of the data (including noise, lower audio quality, faster speech, less clear rendition, high speaker variability including age, gender, linguistic biographies possibly including traces of accents) might have indirectly led to convergence of judgments in monolingual speakers. This might be even more so due to the generally weak load of accent in Russian (Sekerina & Trueswell 2010). Bilingual heritage listeners, on the other hand, might benefit from a perceptual advantage that leads to 'better-than-native perception' (e.g., Chang 2016). Heritage speakers have more familiarity with variation and accents and might thus adapt more quickly to varied speech.

Last but not least, the majority languages of the heritage listeners were either English or German, two languages which also use variable pitch accent placement for focus in noun phrases. This shared feature in both their languages might have increased the sensitivity to a feature that otherwise has a weaker load in Russian. However, such a confirmatory effect, if it exists, is not equally visible in production.

7. REFERENCES

[1] Jasinskaja, K. 2014. Information structure in Slavic. In: Féry, C., Ishihara, S. (eds), *The Oxford Handbook of Information Structure*, 709-732.

[2] Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Jahns, E., Klotz, M., Krause, T., Labrenz, A., Lüdeling, A., Martynova, M., Neuhaus, K., Pashkova, T., Rizou, V., Tracy, R., Schroeder, C., Szucsich, L., Tsehaye, W., Zerbian, S., Zuban, Y. 2019. RUEG Corpus (Version 0.2.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3236069

[3] Zerbian, S., Böttcher, M., Zuban, Y. 2022. Prosody of contrastive adjectives in mono- and bilingual speakers of English and Russian: a corpus study. *Proceedings of Speech Prosody* 2022, Lisbon.

[4] Montrul, S. 2016. *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.

[5] Polinsky, M., Scontras, G. 2020. Understanding heritage languages. *Bilingualism: Language and Cognition*, 23(1), 4-20. doi:10.1017/S1366728919000245

[6] Sekerina, I., Trueswell, J. 2011. Processing of contrastiveness by heritage Russian bilinguals. *Bilingualism: Language and Cognition* 14(3): 280 – 300. DOI: 10.1017/S1366728910000337

[7] Zerbian, S. 2012. Perception and interpretation of intonational prominence in varieties of South African English. In: Braunmüller, K., Gabriel, C. (eds), *Multilingual Individuals and Multilingual Societies*. Benjamins, Amsterdam. 335-348.

[8] Polinsky, M. 2018. *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

[9] Waskom, M. 2021. Seaborn: Statistical Data Visualization. *Journal of Open Source Software,* 6(60). *https://doi.org/10.21105/joss.03021*

[10] Chang, C. 2016. Bilingual perceptual benefits of experience with a heritage language. *Bilingualism: Language and Cognition*, 19(4), 791-809. doi:10.1017/S1366728914000261

EVALUATING PROSODIC ASPECTS OF ORAL READING PROFICIENCY IN SCHOOLCHILDREN: EFFECTS OF GENDER, GENRE AND GRADE

Simon Wehrle[1] & Christopher Sappok[2]

{[1]I*f*L-Phonetik, [2]Institute of German Language and Literature II}, University of Cologne, Germany
{simon.wehrle, csappok}@uni-koeln.de

ABSTRACT

Oral reading fluency is an essential component of school education. Nevertheless, research on the relevant prosodic correlates is rare and mostly limited to listener ratings or coarse acoustic measurements. We applied an innovative method for characterising intonation styles to a longitudinal data set of oral reading performances by German-speaking schoolchildren. We found that higher listener ratings of oral reading proficiency are correlated with more dynamic intonation styles. Further, both melodicity and listener ratings were found to increase over time for individual speakers. Additionally, we show that male children used a more melodic intonation style and received higher listener ratings than females, and that the type of text stimulus had a clear effect on intonational realisation. Our paper is the first to provide robust empirical evidence for an intonational correlate of oral reading proficiency, and simultaneously corroborates the validity of the method used for measuring intonation styles.

Keywords: oral reading, fluency, intonation style, Wiggliness, L1 acquisition

1. INTRODUCTION

Although research on oral reading fluency has received increased attention in recent years [1, 2], the role of prosody in this regard has not been sufficiently elucidated. Some early examples of research on prosody in read speech can be found in [3, 4] (on adults) and [5] (on children). Two recent meta-analytic studies [6, 7] give an overview of educational research on reading acquisition. The characteristics of and differences between read and spontaneous speech are rarely considered in detail in the studies surveyed (in contrast to e.g. [8]) and, relatedly, the identification of target models for oral reading in children might be questioned in many cases (for more details see [9]). Most importantly for our purposes, the characterisation of prosody in the studies surveyed is typically limited to listener ratings or coarse acoustic-phonetic measurements (e.g. [10]).

We investigated a longitudinal data set of oral reading by German-speaking schoolchildren, using an innovative method for capturing intonation styles [11–13] to investigate the contribution of intonation to perceived oral reading proficiency. We found a clear correlation between a more dynamic intonation style and higher perceptual ratings as well as effects of gender, text stimulus (genre) and age (grade).

2. DATA

2.1. LAUDIO corpus

The present study investigates a subset from a corpus of oral reading by schoolchildren that was recorded at a rural German school and spans four years of learning development. Recording sessions were conducted with one child at a time. Sessions lasted around 15 minutes on average and consisted of multiple readings of various texts. All children were invited to participate again in the following year(s). The final corpus contains recordings of children from German grades 3 to 7 (age range 8;3–13;9).

The final Longitudinal AUDIO (LAUDIO) [14] corpus consists of roughly 1000 recordings. A great variety of text stimuli was used, but two anchor texts were applied in every single recording session, which we will focus on here. These are coded as “DOL”, an entry in a children’s dictionary (53 words in length) and “SCHNEE”, a dialogue in the style of a fable (204 words in length).

2.2. Subsample of highly fluent readings

The subset of recordings investigated here was drawn by means of a large-scale screening procedure involving 270 DOL recordings. Each recording was rated by three university students (51 raters in total) using various scales, most importantly the so-called “NAEP-Fluency-Scale” [15]. This scale distinguishes four fluency levels through very detailed descriptions. The results obtained with this scale were sufficiently reliable (ICC inter-rater-reliability > .8) to allow the identification of a subset of particularly fluent performances.

The definition of “highly fluent” children used here is comparatively conservative, in two ways: 1) whereas for the NAEP scale, a silent familiarization with the text in question usually takes place, we only considered the more demanding *prima vista* recordings (first reading without rehearsal); 2) from

these, only recordings were selected that were judged to be at the highest NAEP level (4) by all raters.

The 26 recordings (by 16 children; 9 female, 7 male) for which this was the case make up the subsample of DOL readings investigated here (representing the uppermost quartile of all readings in terms of proficiency). We also included the corresponding SCHNEE readings recorded by the same speakers in the same session (n = 25; one of the recording was not suitable for prosodic analysis as the speaker suffered from a severe cold at the time).

2.3. Listener ratings

Thirteen students attending a seminar on reading fluency took part in a perception experiment on 156 randomised stimuli from the LAUDIO corpus [9, 14]. We focus here on the ratings of perceived *overall quality* of a given oral reading performance (judged on a 10-point scale), which were found to be sufficiently reliable across listeners (ICC > .8).

3. METHOD

3.1. Intonation style: Wiggliness and Spaciousness

While previous work has examined the LAUDIO corpus regarding the influence on listener ratings of factors such as articulation rate (weak positive correlation) and number of errors (strong negative correlation) [16], our focus here is on an analysis of intonational realisation and its relation to ratings.

Specifically, we follow [11, 12, 17] in using a two-dimensional characterisation of intonation style (Wiggliness and Spaciousness). Wiggliness measures the time-varying dynamics of pitch in the form of slope changes per second (range in this data set 0.5–4.6). Spaciousness captures pitch excursions in the form of the largest f0 rises and falls, measured in semitones (ST; range 1.5–15.6). Spaciousness is closely related to more conventional operationalisations of pitch range, while there is no direct analogue to Wiggliness in other approaches (although there is a close resemblance to the concept of macro-rhythm [18, 19]).

We divided each recording into 6 intervals, the pitch contours of which were carefully manually corrected and smoothed by a previously trained annotator before undergoing automatic processing to extract Wiggliness and Spaciousness values [13]. In total, we were able to analyse 300 intervals (from 26 readings by 16 speakers across 2 stories).

Listener ratings are available for all stories at the recording level; separate listener ratings at the interval level had been collected only for the DOL text at the time of analysis (n = 156 intervals).

3.2 Statistical analysis

We used Bayesian modelling for statistical analysis [20–23]. All code, scripts, model specifications and data frames are available at https://osf.io/axqv9/.

4. RESULTS

4.1 Effect of intonation style on listener ratings

We first examined the correlation between Wiggliness/Spaciousness and listener ratings at the recording (rather than interval) level. Across text material (DOL or SCHNEE), we found a clear correlation between Wiggliness and listener ratings. Bayesian modelling of Wiggliness by rating, with speaker as a random effect, shows this to be a robust effect (δ = 0.2, 95% CI [0.09, 0.3], P (δ > 0) = 1). There is also a strong tendency for a positive correlation between Spaciousness and listener rating, but this effect cannot be assumed to be entirely robust (δ = 0.32, 95% CI [-0.06, 0.72], P (δ > 0) = 0.92). Fig. 1 plots the interplay of listener ratings with Wiggliness and Spaciousness.

A speaker-specific analysis revealed that the positive correlation between Wiggliness (and, to a lesser extent, Spaciousness) and listener ratings held true for all speakers, but also that ratings were strongly influenced by idiosyncratic characteristics. Some speakers always received high ratings (8 or 9) and some speakers always received relatively low ratings (5 or 6), but the correlation with intonation style was still evident within those individual ranges.

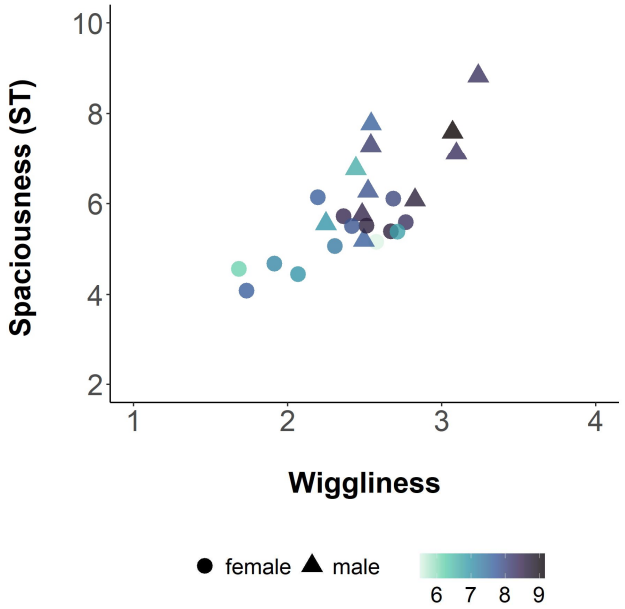


Figure 1: Plot of intonation style and corresponding listener ratings. Spaciousness (in ST) on the y-axis, Wiggliness on the x-axis. Darker shades of blue represent higher listener ratings. Circles represent female speakers, triangles male speakers.

An examination of perceptual ratings and intonation style at the level of intervals broadly confirms the findings from the recording-level analysis, but emphasises the special role of the metric of Wiggliness (capturing pitch dynamics). Only higher Wiggliness values, and not higher Spaciousness values (capturing pitch excursions) were found to be clearly correlated with higher listener ratings. Bayesian modelling confirms a robust positive effect of Wiggliness (δ = 0.15, 95% CI [0.09, 0.22], P ($\delta > 0$) = 1), but not of Spaciousness (δ = 0.06, 95% CI [-0.06, 0.19], P ($\delta > 0$) = 0.81).

4.2 Effects of gender and genre (text material)

The above analysis and the plot in Fig. 1 also point to an effect of gender. We will discuss this here together with the effect of text stimulus, or genre—the SCHNEE dialogue compared with the DOL monologue—on intonation style. Fig. 2 shows Wiggliness and Spaciousness values by gender (male/female) and text type (DOL/SCHNEE).

Turning first to effects of gender, it is clear that both Wiggliness and Spaciousness values tended to be higher for male compared to female speakers. Bayesian modelling confirms this, but also shows that the gender difference was clearer in terms of Spaciousness (δ = -1.2, 95% CI [-1.91, -0.48], P ($\delta > 0$) = 1) compared to Wiggliness (δ = -0.32, 95% CI [-0.65, 0.02], P ($\delta > 0$) = 0.94).

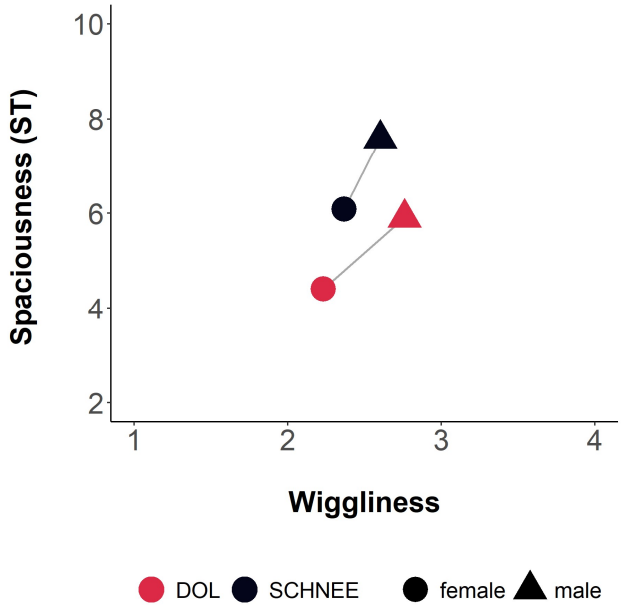


Figure 2: Intonation style by speaker gender and text material. Spaciousness (in ST) on the y-axis, Wiggliness on the x-axis. Circles represent female speakers, triangles male speakers. Red icons represent readings of the DOL monologue, black icons readings of the SCHNEE dialogue.

In contrast, when we focus on a comparison of text materials, it is clear that the intonational difference between readings of the DOL monologue and the SCHNEE dialogue is found in the dimension of Spaciousness only, with Wiggliness (pitch dynamics) not seeming to play any clear role. Bayesian modelling confirms this in unambiguously showing a robust effect for Spaciousness (δ = 1.53, 95% CI [1.1, 1.97], P ($\delta > 0$) = 1) but no effect for Wiggliness (δ = 0.06, 95% CI [-0.11, 0.24], P ($\delta > 0$) = 0.7).

A speaker-specific analysis further confirms this pattern: *all* 15 speakers who produced readings of both stories produced higher Spaciousness values in the SCHNEE dialogue compared to the DOL monologue (with no clear pattern for Wiggliness).

4.3 Effects of grade (individual development by age)

To investigate developmental changes by age, we need to consider data at the level of the individual. Of the 16 individuals in the data set, 7 produced readings at different age grades, with 5 subjects producing readings in two (consecutive) school years, 1 subject in three consecutive school years (aig-03), and 1 subject in four consecutive school years (aig-17).

For all these 7 speakers, listener ratings improved over time (or stayed equivalent, in one case). Concurrently, Wiggliness values increased, representing a more dynamic and varied intonation style. Spaciousness also tended to increase over time, but this correlation was less robust (see section 4.1). Fig. 3 (on the following page) plots listener ratings and Wiggliness values by school year/grade for all 7 speakers that produced longitudinal data. There is a clear overall trend for higher subjective listener ratings and more dynamic intonation styles over time, although some individuals deviated from this pattern. For instance, speaker aig-55 produced near-identical intonation and received near-identical ratings in grades 6 and 7, while speaker aig-17 received higher listener ratings in grade 6 than in grade 5 despite a slight reduction in pitch dynamics.

5. DISCUSSION

We used a two-dimensional characterisation of intonation style to investigate prosodic aspects of oral reading proficiency in schoolchildren. We found that both dimensions are positively correlated with listener ratings, but that a higher degree of Wiggliness (pitch dynamics), in particular, is the most robust prosodic correlate of more highly rated oral reading performances. This is a significant finding for at least two reasons. Our results provide the first suggestion of a reliable intonational correlate of oral reading proficiency (that is supported by thorough prosodic analysis). At the same time, these results corroborate

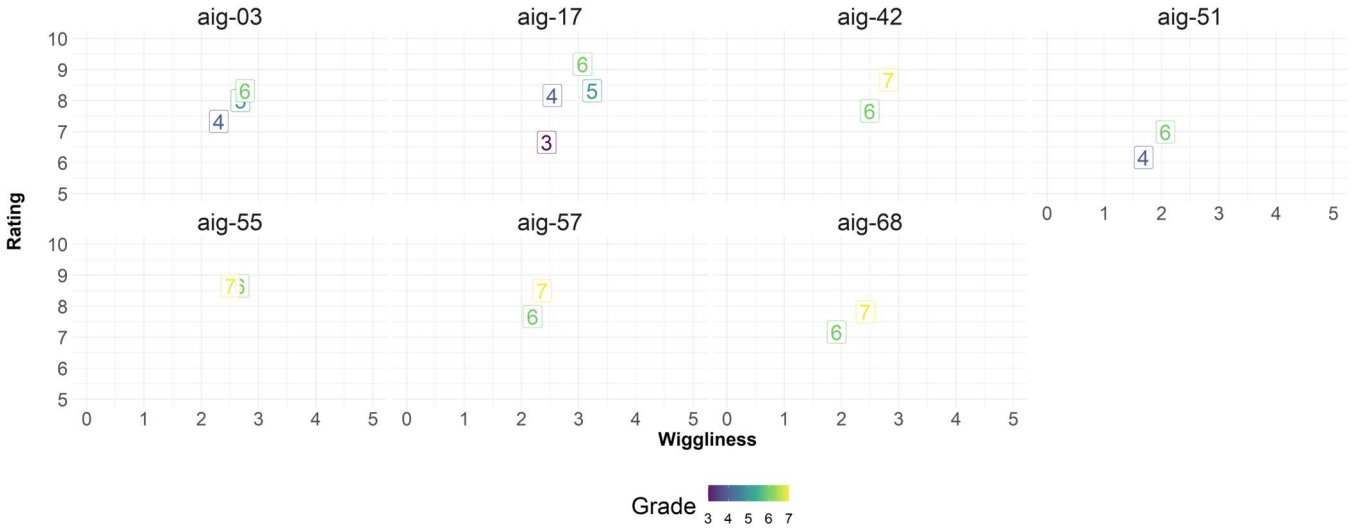


Figure 3: Wiggliness values (on the x-axis) and listener ratings (on the y-axis) by speaker and grade (/school year; higher grades represented with brighter colours). Only speakers that were recorded at more than one time point are included.

the validity of measuring intonation styles along the two dimensions of Wiggliness and Spaciousness: the positive correlations with listener ratings strongly suggest that the metrics used accurately capture acoustic features that guide perceptual judgements.

From a longitudinal perspective, we concordantly observed that both listener ratings and Wiggliness values increased with age. This signifies that schoolchildren are clearly able to develop a more dynamic intonation style, and produce oral readings that are rated to be of a higher quality, in the course of their individual development. All children for whom longitudinal data are available in the current data set showed increases (or equivalent values) in terms of both listener rating and intonation style.

While Wiggliness is thus the most relevant intonational correlate of perceived oral reading proficiency, the complementary measure of Spaciousness was shown to be decisive for effects of gender and genre (text material). Across year groups, males produced more melodic intonation styles than females and concurrently received higher listener ratings. The underlying factors are not clear. Voice change in puberty is not a sufficient explanation, as males differed from females for all year groups (starting at age 8). One speculative interpretation is that we might find males lying at *either* extreme end of the distribution in the full corpus (rather than this subset of particularly fluent speakers), i.e., that males may have produced both the most and the least melodic intonation styles (in terms of Spaciousness).

Differences as a function of genre, or text stimulus, are more easily explained. Spaciousness values were far higher for the SCHNEE compared to the DOL stimulus. The DOL text is not only monologic, it is also, by objective measures, a very difficult text for this age group (even though it was taken from a children's dictionary) [16]. In contrast, the SCHNEE text features a lively dialogue between two characters, a big burly snow man and a cute little hare. Most children clearly distinguished these two characters in terms of Spaciousness. Although this is not entirely surprising [24–26], it does have important implications, as it 1) underlines the importance of elicitation methods and text materials in studies on the production of prosody [8, 27–29] and 2) further corroborates the distinctiveness and complementarity of the Wiggliness and Spaciousness metrics.

Overall, the current work is thus an important contribution to 1) our understanding of intonational aspects of oral reading proficiency and 2) the characterisation and measurement of intonation styles more generally. The most important limitation of the current study is the restricted subset of data, featuring only the most fluent readers. We are in the process of expanding the scope from the current 26 to over 300 recordings in order to verify whether intonation style can also serve as a reliable correlate of oral reading proficiency in the context of less fluent speech.

6. ACKNOWLEDGEMENTS

We would like to give big thanks to Kim Stroeks for her extensive work on data processing.

7. REFERENCES

[1] K. Holle, 'Flüssiges und phrasiertes Lesen (fluency). Lesetheoretische Grundlagen und unterrichtspraktische Hinweise', *Schriftspracherwerb empirisch. Konzepte–Diagnostik–Entwicklung*, pp. 87–119, 2006.

[2] C. Rosebrock and D. Nix, 'Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen

Leseforschung und-didaktik', *Didaktik Deutsch*, no. 20, 2006.
[3] G. Fant and A. Kruckenberg, 'Preliminaries to the study of Swedish prose reading and reading style', *STL-QPSR*, vol. 2, no. 1989, pp. 1–83, 1989.
[4] E. Blaauw, 'The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech', *Speech Communication*, vol. 14, no. 4, pp. 359–375, Sep. 1994, doi: 10.1016/0167-6393(94)90028-0.
[5] R. Cowie, E. Douglas-Cowie, and A. Wichmann, 'Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers', *Language and Speech*, vol. 45, pp. 47–82, 2002, doi: 10.1177/00238309020450010301.
[6] A. P. Wolters, Y.-S. G. Kim, and J. W. Szura, 'Is Reading Prosody Related to Reading Comprehension? A Meta-analysis', *Scientific Studies of Reading*, vol. 26, no. 1, pp. 1–20, Jan. 2022, doi: 10.1080/10888438.2020.1850733.
[7] E. Godde, M.-L. Bosse, and G. Bailly, 'A review of reading prosody acquisition and development', *Read Writ*, vol. 33, no. 2, pp. 399–426, Feb. 2020, doi: 10.1007/s11145-019-09968-1.
[8] L. E. De Ruiter, 'Information status marking in spontaneous vs. read speech in story-telling tasks–Evidence from intonation analysis using GToBI', *Journal of Phonetics*, vol. 48, pp. 29–44, 2015.
[9] C. Sappok, 'Oral reading proficiency and prosody – a perceptual pilot study on especially fluent German students (grade 3 to 7)', in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Repulic, 2023.
[10] M. R. Kuhn, P. J. Schwanenflugel, and E. B. Meisinger, 'Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency', *Reading Research Quarterly*, vol. 45, no. 2, pp. 230–251, 2010, doi: 10.1598/RRQ.45.2.4.
[11] S. Wehrle, F. Cangemi, H. Hanekamp, K. Vogeley, and M. Grice, 'Assessing the Intonation Style of Speakers with Autism Spectrum Disorder', in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 809–813.
[12] S. Wehrle, F. Cangemi, M. Krüger, and M. Grice, 'Somewhere over the spectrum: Between robotic and singsongy intonation', *Il parlato nel contesto naturale. Speech in the natural context*, no. 4, pp. 179–194, Dec. 2018, doi: 10.17469/O2104AISV000010.
[13] S. Wehrle, 'A brief tutorial for using Wiggliness and Spaciousness to measure intonation styles', Jun. 2022, Accessed: Dec. 14, 2022. [Online]. Available: https://osf.io/5e7fd/
[14] C. Sappok and J. Fay, 'Prosodische Aspekte von Leseflüssigkeit messen. Evaluation einer Ratingprozedur mit Audioaufnahmen von DrittklässlerInnen', *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, vol. 23, no. 44, pp. 61–83, 2018.
[15] G. S. Pinnell, J. J. Pikulski, K. K. Wixson, J. R. Campbell, P. B. Gough, and A. S. Beatty, *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4*. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education, 1995.
[16] C. Sappok, 'Exploring Advanced Prosody–eine Best-Practice-Untersuchung zum lauten Lesen in der weiterführenden Schule', *Weiterführende Grundlagenforschung in Lesedidaktik und Leseförderung*, pp. 68–97.
[17] S. Wehrle, 'A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder', PhD Thesis, University of Cologne, Cologne, Germany, 2021.
[18] S.-A. Jun, 'Prosodic typology revisited: Adding macro-rhythm', in *Speech Prosody 2012*, Shanghai, China, 2012.
[19] S.-A. Jun, 'Prosodic typology: By prominence type, word prosody, and macro-rhythm', in *Prosodic Typology II*, Oxford University Press, 2014, pp. 520–539.
[20] P.-C. Bürkner, 'brms: An R package for Bayesian multilevel models using Stan', *Journal of statistical software*, vol. 80, no. 1, pp. 1–28, 2017.
[21] R Core Team, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/
[22] RStudio Team, 'RStudio: Integrated Development Environment for R'. RStudio, PBC, Boston, MA, 2021. [Online]. Available: http://www.rstudio.com/
[23] Stan Development Team, 'Stan Modeling Language Users Guide and Reference Manual, Version 2.29.' 2022. [Online]. Available: https://mc-stan.org
[24] J. J. Ohala, 'Cross-language use of pitch: an ethological view', *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983.
[25] L. Hinton, J. Nichols, and J. J. Ohala, *Sound symbolism*. Cambridge University Press, 2006.
[26] C. Gussenhoven, 'Foundations of intonational meaning: Anatomical and physiological factors', *Topics in Cognitive Science*, vol. 8, no. 2, pp. 425–434, 2016.
[27] A. Janz, 'Navigating Common Ground Using Feedback in Conversation- A Phonetic Analysis', MA thesis, University of Cologne, Cologne, Germany, 2022.
[28] M. Grice, M. Savino, and M. Refice, 'The intonation of questions in Bari Italian: Do speakers replicate their spontaneous speech when reading', *Phonus*, vol. 3, pp. 1–7, 1997.
[29] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, 'Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations', in *CogSci'19*, Cognitive Science Society, 2019, pp. 261–267.

THE INFLUENCE OF EMPATHY AND AUTISTIC-LIKE TRAITS IN PROMINENCE PERCEPTION

Riccardo Orrico, Stella Gryllia, Jiseung Kim, Amalia Arvaniti

Radboud University, Netherlands
riccardo.orrico@ru.nl, stella.gryllia@ru.nl, jiseung.kim@ru.nl, amalia.arvaniti@ru.nl

ABSTRACT

We report an investigation of how individual listeners' characteristics influence the assessment of prominence. Following the RPT paradigm, 82 Standard Southern British English speakers listened to utterances and marked the words that sounded prominent to them. We focused on words produced with either H* or L+H*, which appeared in either contrastive or non-contrastive pragmatic conditions. Both pragmatics (contrastive vs. non-contrastive accent) and phonetics (H* vs. L+H*) systematically guided participants' judgment. L+H*s and contrastive accents were more likely to be marked as prominent, but some listeners prioritized phonetics and others pragmatics, and these differences were related to their empathy and autistic-like traits: the higher a listener's empathy, the more they relied on pragmatics, while the higher the autistic-like traits the smaller the distinction between the two accents. Our results suggest that both acoustic and pragmatic salience are instrumental for prominence assessment, but listeners weigh these cues in different ways.

Keywords: individual characteristics, prominence, pragmatics, Rapid Prosody Transcription, English

1. INTRODUCTION

A growing body of research has been devoted to the study of individual variability, pointing to a heterogeneous nature of language processing. This variability has been linked to individual cognitive styles [1], such as empathy and autistic-like traits. Empathy has been linked to the ability to interpret the interlocutor's intentions and therefore to pragmatic skills [2]. With respect to prosody, empathy has been shown to correlate with the ability to derive meaning from intonation [3, 4]. On the other hand, individuals with more autistic-like traits show poorer pragmatic abilities [2], heightened sensitivity to acoustic cues [1, 5], and poorer high order cognitive ability [6] compared to those with fewer such traits.

Here we present a Rapid Prosody Transcription (RPT) study [7, 8] that investigated how individual differences in empathy and autistic-like traits affect prominence assessment. Previous RPT studies [9, 10, 11] have reported individual variability in RPT responses. The results of [9] and [10] suggest that listeners differ as to how they weigh the available cues when assessing prominence, while [11] reports that individuals with more autistic-like traits show poorer prominence perception. These differences may be due to the broad definition of prominence as the property of standing out in an utterance by means of phonetic, phonological, pragmatic, or lexical cues. Given the interplay of so many co-occurring cues, individuals with different cognitive styles may prioritize different cues when making judgments about prominence.

We examined prominence ratings during RPT as a function of pitch accent type (focusing on H* and L+H*), considering separately each accent's F0 shape and pragmatic function (for details, see 2.2.1). H* and L+H* were chosen because their prominence relates to both their shape and their pragmatics [12]. Further, since prominence perception does not rely exclusively on psychophysical properties [13], we focused on Standard Southern British English (SSBE), to avoid possible regional differences in the accents' function, as function plays a part in how the accents' prominence is assessed (cf. [14]). We hypothesized that neurotypical individuals with different processing styles would employ different strategies in assessing prominence. Processing styles were operationalized as degree of empathy and extent of autistic-like traits as measured by the Empathy Quotient (EQ) and the Autism Quotient (AQ) questionnaires respectively (for details, see 2.3). We predicted that, when assessing prominence, individuals with high EQ would rely more on pragmatic information, while individuals with high AQ would rely more on phonetics.

2. METHODS

2.1. Participants

We used Prolific (https://www.prolific.co/) to recruit 85 native monolingual speakers of SSBE without (self-reported) speech or hearing disorders: they were all brought up in monolingual households and had learned languages other than English through formal instruction only. Participants were paid £18 to complete the experiment. We report results from 82 participants (47 F, 19–50 years of age), after excluding one participant who faced technical

problems and two participants with more than 10% of outlier responses (they selected all words in utterances with 1-5 words, or more than 85% of words in utterances with more than 5 words).

2.2. RPT Stimuli

The stimuli were 86 utterances produced by eight SSBE speakers (5 F). Twenty-two utterances were excerpts from read speech in a reading task, while the remaining 64 were from various spontaneous speech tasks. The total number of words in the 86 utterances was 879. The utterance length was 3–24 words (Mean: 10.2; SD: 4.7).

2.2.1 Categorization of the stimuli

Words with high and rising accents were categorized using phonetic and pragmatic criteria independently of each other. Phonetically, rising accents were annotated as L+H* if a deliberate dip was present at the accented syllable onset, or if the coda preceding a voiceless onset had low F0. Otherwise, accents were annotated as H*. Accents in absolute utterance initial position were classified as H* [15]. Accents followed by uptalk were not included as it would be impossible to separate the effect of the accent from that of the utterance-final rise on prominence assessment.

Pragmatically, words bearing H* or L+H* were classified as *contrastive* or *non-contrastive* within the context of their utterance. Only the orthographic transcripts were used for this categorization to avoid circularity between phonetic shape and pragmatic function in the annotation process. An accent was annotated as *contrastive* when it highlighted one item out of a small set of (mentioned or readily inferred) alternatives in discourse. Focus particles (e.g., *just*, *only*) and negative expressions (e.g., *don't* in *I don't know*) were also marked as contrastive. Finally, items were considered contrastive if they were part of an explicit parallelism in discourse (e.g., in *I want tea as well as coffee*, *tea* and *coffee* would be considered contrastive). All other high and rising accents were considered non-contrastive. This classification cuts across the categories *new*, *given*, *topic*, and *focus*, as it is possible to contrastively accent both new and given items as well as topics and foci, and (phonetic) accent type does not appear to vary systematically based on these pragmatic distinctions [16].

2.3. Additional tests

Participants completed the EQ [17] and AQ [18] questionnaires. The EQ score ranges from 0 to 80; it is based on 40 responses on a 4-point scale and represents the individual's level of emotional and cognitive empathy. The AQ score ranges from 0 to 50; it is based on 50 responses on a 4-point scale and is not a diagnostic test; rather, it positions adults of normal intelligence along a continuum measuring five traits associated with the autism spectrum disorder: social skills, communicative skills, attention to detail, attention switching, and imagination.

2.4. Experimental procedure

The study ran online using ROLEG developed at Radboud University. The self-paced RPT task had 2 practice trials followed by 86 trials in five blocks, giving participants 4 prompts for self-paced breaks. In each trial, participants would first hear an utterance while seeing it in writing on screen; then they heard a second repetition during which they could select which words sounded "highlighted, prominent, important, stressed, or emphasised" to them. The transcript did not include punctuation or capitalization other than apostrophes in contractions and possessives, and capitalization in proper nouns and the pronoun 'I'. Participants were instructed to select as many words as they saw fit but had to select at least one in each stimulus to proceed to the next trial. After the RPT task, participants were given access to the EQ and the AQ questionnaires on PsyToolkit (https://www.psytoolkit.org/, [19, 20]). Participants were asked to complete them within seven days after the RPT task to be remunerated.

2.5. Statistical analysis

The binary RPT responses (word marked as prominent or not) were analysed in three Generalized Linear Mixed-Effects Models (GLMM) using the lme4 package [21] in R [22]. The first model tested the relation between the RPT responses and the accents' phonetic and pragmatic classification. It included as fixed effects PHONETICS (H*, L+H*), PRAGMATICS (contrastive, non-contrastive) and their interaction. The random effects included random intercepts for TALKER, PARTICIPANT and ITEM (the accented word), and by-subject random slopes for ACCENT and PRAGMATICS. Two additional models tested the relations between the RPT responses, the accents' phonetic and pragmatic categorization, and the EQ and AQ scores. The EQ model included as fixed effects PHONETICS, PRAGMATICS, EQ, and the interaction between PRAGMATICS and EQ. The AQ model included PHONETICS, PRAGMATICS, AQ, and the interaction between PHONETICS and AQ. Both models included the random intercepts for SPEAKER, PARTICIPANT, and ITEM as random effects. The effects of EQ and AQ were tested in two separate models to avoid collinearity issues detected in preliminary analysis.

3. RESULTS

3.1. The effect of accent and pragmatics

Details about the data and analysis are available at osf.io/euwn8. The output of the first GLMM showed that both PHONETICS and PRAGMATICS were significant predictors of the binary responses (word selected as prominent or not): words annotated as L+H* or contrastive were more likely to be selected as prominent than those annotated as H* or non-contrastive ($\beta_{\text{PHONETICS: L+H*}} = 1.33$, SE = 0.32, z = 4.11, $p < .0001$; $\beta_{\text{PRAGMATICS: Contr}} = 1.41$, SE = 0.28, z = 5.02, $p < .0001$). The interaction between PHONETICS and PRAGMATICS was not significant (β = -0.34, SE = 0.46, z = -0.74, $p > .4$).

Fig. 1 shows the distribution of p-scores (the percentage of participants selecting a word as prominent) for the four combinations of PHONETICS and PRAGMATICS. The categories *H* non-contrastive* and *L+H* contrastive* peaked towards the low and high end of the p-score distribution respectively, while *H* contrastive* and *L+H* non-contrastive* largely overlapped and showed more variable responses than the first two categories.

3.2. Individual response patterns

Following [9] and [23], we observed individual patterns by inspecting the by-subject random slopes for PHONETICS and PRAGMATICS. This revealed that individual participants prioritized different criteria. To illustrate the three prototypical patterns that emerged, we show in Fig. 2 the random slope values by PHONETICS (x-axis) and PRAGMATICS (coloured dots) for participants P1, P2, and P3; the y-axis shows the predicted probabilities of the GLMM output variable, i.e., whether a word was marked as prominent (1) or not (0). P1 reliably differentiated the pragmatic conditions (represented by the vertical distance between the orange and purple dots); at the same time, this participant showed a less reliable distinction between the two phonetic conditions (represented by the relatively horizontal dashed lines linking the dots across the phonetic conditions, H* and L+H*), suggesting that they relied mainly on pragmatic information to mark prominence. On the other hand, P3 showed a robust distinction between H* and L+H* (as represented by the steeper dashed lines) but exhibited a less robust distinction based on pragmatics, especially in the case of L+H* accents (as represented by the closer vertical distance between the orange and purple dots for L+H*); this suggests they relied mainly on phonetic information to mark prominence. Finally, P2 differentiated the slope values for pragmatics *and* phonetics to a similar extent, indicating that they relied on both criteria relatively uniformly.

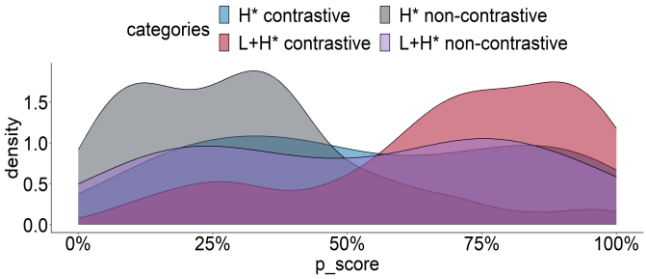


Figure 1: Density plots of p-scores for the four combinations of PHONETICS and PRAGMATICS.

In order to quantify the extent to which a participant relied on a specific criterion (phonetics or pragmatics), we extracted the *slope values* independently for the two variables (PHONETICS and PRAGMATICS) and added the individual slope values to the general slope of the effects predicted by the GLMM. The obtained *values* provide a quantitative measure of the relevance of the two criteria for each participant's responses: the higher the value of a criterion for a given participant, the more they relied on that criterion. Fig. 3 illustrates these *values* and shows that, while the three prototypical patterns can be identified (as shown in the three panels), the individual responses form a continuum extending from those who relied more on pragmatics (left panel; shown in shades of red) to those who relied more on accent (right panel; shown in shades of blue).

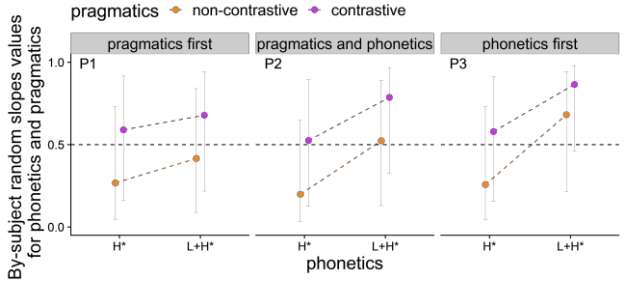


Figure 2: Random slope values for three participants showing prototypical response patterns: those who prioritized pragmatics (P1), those who prioritized phonetics (P3), and those who relied equally on both (P2). Points indicate slope values extracted from the GLMM; vertical lines indicate confidence intervals.

3.2.1 The effect of EQ and AQ

EQ participants' scores ranged from 16 to 70 (Mean = 43.2, SD = 12.3). AQ scores ranged from 1 to 42 (Mean = 19.2, SD = 9.9). Cronbach's Alpha [24] showed high reliability for both (0.95 for EQ; 0.93 for AQ). A Pearson correlation showed that AQ and EQ were negatively correlated (r(80) = -.58, p <.001).

The GLMM output revealed a significant interaction between EQ and PRAGMATICS (β = .006, SE = .003, z = 2.22, $p < .03$), indicating that the higher

a participant's EQ, the more likely they were to mark contrastively accented words as prominent. AQ and PHONETICS also interacted (β = -.008, SE = .004, z = -2.20, $p < .03$), so that the higher a participant's AQ, the more likely they were to mark H*s as prominent, thereby reducing the distinction between the two phonetic conditions. The effects of EQ and AQ are illustrated in Fig. 4.

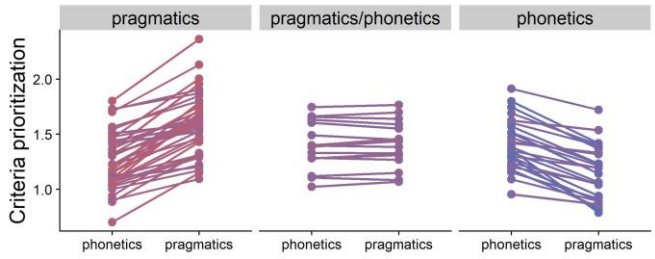


Figure 3: Slope values for PHONETICS and PRAGMATICS for individual responses. Gradience from red (left) to blue (right) reflects a continuum from prioritizing pragmatics to prioritizing phonetics.

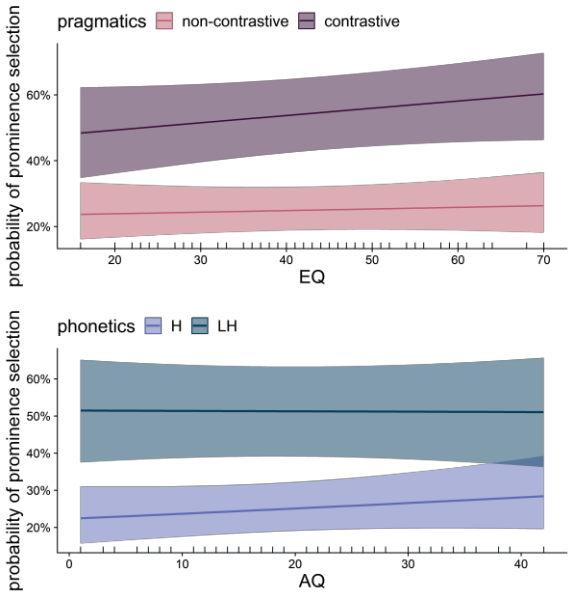


Figure 4: Predicted probabilities of RPT responses for the interaction between PRAGMATICS and EQ (top) and PHONETICS and AQ (bottom).

4. DISCUSSION AND CONCLUSIONS

We tested the relationship between prominence assessment and cognitive styles (operationalized as empathy and autistic-like traits) by considering as predictors the phonetics differences between H* and L+H* accents and the pragmatic information they convey (contrastive or non-contrastive).

The aggregate results (Fig. 1) showed that both phonetic and pragmatic information was exploited when making judgments about prominence, and the two factors worked synergistically in driving these judgments. This suggests that phonetic information was not a sufficient cue to prominence; rather, the role of accent shape was weighed together with other cues within the utterance.

Our results also suggest that these cues were not exploited by the participants in the same way. While both phonetic and pragmatic cues were used by all participants to some extent, many (though not all) individuals prioritized either phonetics or pragmatics to some extent. These different priorities may be the lead cause behind the individual differences in RPT responses reported in [9] and [10]: variation can be explained by individual characteristics like those examined in the present study. Participants with higher EQ were more likely to rely on pragmatics than those with lower EQ, supporting previous findings that empathetic individuals are highly sensitive to pragmatic information [3, 4]. In addition, participants with higher AQ made less reliable distinctions between the phonetic conditions. This finding is in line with previous studies: [11] reported poorer prominence perception for high AQ listeners (though the effect was mainly observed in unaccented words) and linked this result to lower sensitivity to the prosody-meaning mapping. In our study, high AQ listeners showed an increased tendency to mark accented words as prominent, regardless of accent type. This can be interpreted as a tendency to perceive as prominent any prosodically strong word and a poorer integration of F0 shape with context.

Taken together, our results show that individual characteristics related to cognitive styles shape the way individuals process linguistic information and predict the extent to which listeners evaluate cues of different nature (as opposed to relying only on acoustic cues) when making judgments about prominence. Avenues for future research include testing additional individual characteristics that have been argued to play a role in linguistic processing (e.g., musicality [25] and language background [4]). It is also important to investigate whether these individual differences reflect distinct processing styles or simply the way participants interpret the RPT task. Revisiting the definition of prominence by taking into account crucial grammatical and language-specific aspects (cf. [13]) are necessary steps to tackle this question: a rigorous understanding of what prominence is would allow for a thorough investigation of its processing mechanisms.

5. ACKNOWLEDGEMENTS

The financial support of the European Research Council through grant ERC-ADG-835263 (SPRINT) to Amalia Arvaniti is hereby gratefully acknowledged. We thank Na Hu, Katherine Marcoux, and Sofia Sialiaki for their help with the preparation of the experiment and participant recruitment.

6. REFERENCES

[1] Yu, A. C., Zellou, G. 2019. Individual differences in language processing: Phonology. *Annual Review of Linguistics* 5, 131–150.

[2] Baron-Cohen, S. 2004. *The Essential Difference*. Penguin UK.

[3] Esteve-Gibert, N. J., Schafer, A., Hemforth, B., Portes, C., Pozniak, C., D'Imperio, M. 2020. Empathy influences how listeners interpret intonation and meaning when words are ambiguous. *Memory & Cognition* 48(4), 566–580. doi: http://doi.org/10.3758/s13421-019-00990-w

[4] Orrico, R., D'Imperio, M. 2020. Individual empathy levels affect gradual intonation-meaning mapping: The case of biased questions in Salerno Italian. *Laboratory Phonology* 11(1), 1–39.

[5] Yu, A. C. 2010. Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PloS One* 5(8), [Online]. doi: https://doi.org/10.1371/journal.pone.0011950

[6] Stewart, M. E., Ota, M. 2008. Lexical effects on speech perception in individuals with "autistic" traits. *Cognition* 109(1), 157–162.

[7] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1(2), 425–52.

[8] Cole, J., Shattuck-Hufnagel, S., 2016. New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology* 7(1): 8. doi: 10.5334/labphon.29

[9] Baumann, S., Winter, B. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70, 20-38.

[10] Arvaniti, A., Gryllia, S., Zhang, C., Marcoux, K. P. 2022. Disentangling emphasis from pragmatic contrastivity in the English H*~L+ H* contrast. *Proceedings of Speech Prosody 2022*. doi: 10.21437/SpeechProsody.2022-170

[11] Bishop, J., Kuo, G., Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics* 82. 100977.

[12] Ladd, D. R., Morton, R. 1997. The perception of intonational emphasis: continuous or categorical?. *Journal of phonetics* 25(3), 313–342.

[13] Ladd, D. R., Arvaniti, A. 2023. Prosodic prominence across languages. *Annual Review of Linguistics* 9:171–193.

[14] Arvaniti, A., Garding, G. 2007. Dialectal variation in the rising accents of American English. *Papers in Laboratory Phonology,* 9: 547–576.

[15] Brugos, A., Shattuck-Hufnagel, S., Veilleux, N. 2006. *Transcribing Prosodic Structure of Spoken Utterances with ToBI* (MIT Open Courseware). http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm.

[16] Molnár, V. 2002. Contrast from a contrastive perspective. *Language and Computers* 39, 147-162.

[17] Baron-Cohen, S., Wheelwright, S. 2004. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders* 34(2), 163–175.

[18] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. 2001. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* 31(1), 5–17.

[19] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods* 42(4), 1096–1104.

[20] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44(1), 24–31.

[21] Bates, D., Mächler, M., Bolker, B. Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 1(1), 1–48. Available: https://www.jstatsoft.org/v067/i01

[22] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020, [Online]. Available: http://www.r-project.org/.

[23] Drager, K., Hay, J. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change* 24(1), 59–78.

[24] Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: https://doi.org/10.1007/BF02310555

[25] Schön, D., Magne, C., Besson, M. 2004. The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology* 41(3), 341–349.

SO NEAR YET SO FAR: FINE PHONETIC DIFFERENCES IN TONAL ALIGNMENT AND TEMPORAL STRUCTURE IN CLOSE CONTACT VARIETIES

Jasmin Pöhnlein, Felicitas Kleber

Institute for Phonetics and Speech Processing (IPS), LMU Munich
jasmin.poehnlein@phonetik.uni-muenchen.de, kleber@phonetik.uni-muenchen.de

ABSTRACT

This study investigates peak alignment in nuclear rising accents in two generations of West Central Bavarian (WB) speakers compared to Standard German (SG) speakers. WB and SG differ in the segmental timing of vowel plus consonant sequences but younger WB speakers approach the temporal patterns found in SG due to dialect levelling. We analyzed time-normalized, speaker-scaled f0 trajectories extracted from sequences of long vowels plus sonorants in pitch accented trochees in nuclear position from 38 SG as well as 22 older and 22 younger WB speakers. Results from functional linear mixed models indicate differences in peak alignment between SG and WB speakers but not between age groups. However, a subsequent analysis revealed that younger WB and SG speakers share a weakly positive duration-alignment-correlation. No such correlation was found for older WB. Findings are discussed in light of dialect levelling as well as diverging mechanisms to deal with available sonorant material.

Keywords: Tonal alignment, segmental timing, dialect levelling, fine phonetic differences, FLMM.

1. INTRODUCTION

The temporal coordination of a tonal target in the fundamental frequency (f0) contour with the segmental level – known as phonetic alignment [1] – is, amongst others, affected by dialectal background and pragmatic context. Particularly the low beginning (L) but also the high target (H) of non-contrastive prenuclear rises are aligned later with the stressed syllable in Southern compared to Northern German varieties [2-5]. No such regional difference in peak alignment was found for rising pitch accents marking contrast, regardless of whether they occur in prenuclear [3, 6] or nuclear position [6, 7]. A comparison in [8] further suggested the regional differences to emerge mainly in tokens containing phonemically long vowels in the stressed syllable followed by sonorants; i.e., in contexts allowing for a delay given the sufficiently available and advantageous segmental material (e.g., [9, 10]).

While f0 peak delay has been described for many southern varieties of German (e.g., [11, 12]) including Southern Bavarian dialects spoken in Austria [5] and regardless of timing constraints due to the segmental material, not much is known about peak alignment in West Central Bavarian (WB) spoken in south-eastern Germany. Yet, for two reasons this dialect is of particular interest. First, its temporal structure diverges from Standard German (SG; [13, 14]) such that word-medial post-vocalic consonants show clear tendencies of a singleton/geminate contrast (cf. [15], usually referred to as fortis/lenis contrast for obstruents) accompanied by a predictable (allophonic) length of the preceding vowel: Long consonants (including sonorants) follow short vowels and vice versa (e.g., [16, 17]). Second, this WB dialect feature appears to level out (caused by an increasing SG influence, cf. [18]) as suggested by apparent-time studies [19-21] showing that younger WB speakers' fortis consonants are less geminate-like than those of older WB speakers, resulting in shorter durations closer to SG consonants (which does not feature a singleton/geminate contrast). Investigating this dialect offers a rare opportunity to study the effects of segmental structure on peak delay in a variety undergoing a prosodic change in terms of its temporal structure.

The two main aims of this study were thus to test for peak alignment differences between (i) the standard and the WB variety spoken in Bavaria and (ii) two generations of WB dialect speakers. The hypotheses were (i) that WB speakers align non-contrastive nuclear rises later than SG speakers because SG resembles Northern SG more closely than WB and (ii) that younger as opposed to older WB speakers show less peak delay given the levelling out of temporal dialect features in this group.

2. METHODS

2.1. Speech materials and participants

The disyllabic, trochaic target words in Table 1 with phonemically long vowels (V_1) in the stressed, open syllable and sonorants (C_2) in the onset of the unstressed syllable were part of two bigger corpora (cf. [21] for 1) containing, among others, acoustic recordings from 38 SG (aged 19–82, mean 38.0 years,

20 female) as well as 22 younger (aged 18–30, mean 24.8 years, 11 female) and 22 older (aged 49–73, mean 58.0 years, 11 female) WB speakers. While age was balanced within SG speakers, they were not assigned to age groups given their function as control group in the present study. SG speakers were city dwellers with roots in Munich (where Southern SG is spoken) and self-identified as non-dialectal in terms of dialect competence and use. WB speakers were originally from rural areas in the broader vicinity of Munich where WB is still acquired as first variety of German; their WB competence was assessed by native dialect speakers (the experimenters).

Corpus	Target word (item)		Gloss	Tokens (n)
2	*Dame*	/ˈdaː.mə/	'lady'	120
1	*Diener*	/ˈdiː.nɐ/	'servant'	210
2	*Fahne*	/ˈfaː.nə/	'flag'	120
1	*Höhle*	/ˈhøː.lə/	'cave'	210
2	*Krone*	/ˈkʰʁoː.nə/	'crown'	120
2	*Panik*	/ˈpʰaː.nɪkʰ/	'panic'	120

Table 1: Number of tokens per target item (42 speakers × 5 repetitions per item in corpus 1 and 40 speakers × 3 repetitions per item in corpus 2).

All target words are part of the SG and the WB lexicon, respectively. They were embedded in word-specific, but within and across corpora syntactically similar carrier phrases. WB sentences were translations from SG; see (1) for an example with the target word in bold. In all carrier phrases the number of syllables per sentence was kept constant across variety and as constant as possible across sentences (three syllables before and one or two after the target word). The probability for nuclear pitch accents to occur on the target was high (cf. [22, 23]).

(1)	SBJ	MOD	ART	OBJ	V
SG:	*Er*	*will*	*die*	***Dame***	*treffen.*
	/ʔɛʁ	vɪl	diː	ˈdaː.mə	ˈtʰʁɛfən/
WB:	*Ea*	*mog*	*de*	***Dame***	*treffa.*
	/eːa	moːk	de	ˈdaː.mə	ˈtrɛfa/

Gloss: 'He wants to meet the lady.'

Recordings were made using SpeechRecorder [24] at a minimum sampling rate of 44.1 kHz, either in a sound-attenuated booth or at participants' homes (with a condenser microphone) or remote via Wikispeech [25] (with headsets). Sufficient recording quality was carefully monitored. Participants read each carrier phrase prompted on screen in silence and reproduced it from memory after it had disappeared. All target words in Table 1 have a high lexical frequency and were produced at a normal speech rate.

2.2. Data analysis

Recorded utterances were automatically segmented with WebMAUS [26] into words and phones on two interval levels and stored as EMU databases [27, 28]. Relevant segment boundaries, i.e., start and end of (i) the utterance, (ii) the target word, and (iii) the target word's phonemes, were manually corrected. Using a minimal set of GToBI labels [29], tonal targets were marked on an event level based on visual inspection of the f0 trajectory, which was calculated separately for males and females with the *ksvF0*-function of *wrassp* [30]. All of these manual adjustments were done by trained phoneticians. A cascading multi-stage process ensured that ambiguous cases were particularly checked and independently judged by at least two labellers (including one or both authors).

F0 trajectories across $V_1C_2(V_2)$-sequences were computed with *emuR* [31], using *R* [32] (v. 4.2.0) in *Rstudio* [33] (v. 2.3.492). The extracted Hertz (Hz) values per trajectory were linearly normalized to equidistant points in a 0-to-1-time interval, also due to significant between-group differences in phrase-level articulation rate – defined as an utterance's number of canonical syllables per second. Time points with Hz values equal to zero and f0 outliers were removed (cf. [34]). To control for individual pitch level, we calculated speaker-scaled f0 values (*z*-scores) across all voiced frames (e.g., [35]).

Sentences with no nuclear rise on or towards the target word and/or a high phrase-final boundary tone were excluded from all subsequent analyses. Speakers who realized less than 25 % of the sentences with a rise-fall nuclear contour were discarded completely (four SG, one younger and four older WB speakers).

Speaker group	Total (n)	Incl. (n)	BpS (%)
Dialect, older	240	150	0.77
Dialect, younger	240	190	0.83
Standard German	420	296	0.79

Table 2: Percentage of bundles per speaker (BpS) included in the analysis, split by speaker group.

The f0 trajectories across the V_1C_2-sequences (hereafter V+S-sequence) in the remaining utterances (cf. Table 2) were then analyzed with functional linear mixed models (FLMM; [36, 37]). We fitted two separate FLMM using *sparseFLMM* [38] because of the current restrictions to binary (i.e., two-level) fixed factors only. Despite this small methodological drawback, FLMM is particularly suitable for our analyses because it enables the evaluation of entire f0 contours and irregularly sampled functional data. The two models aimed at between-group comparisons of (i) regional background in general (*Variety*: SG vs. WB speakers) and (ii) different dialect generations (*Age*: younger vs. older WB speakers). For both, the dependent variable was the speaker-scaled f0 contour mapped onto the 0-to-1-time interval. *Variety* and *Age* were added as single covariates, respectively.

The binary covariate effects (comparable to fixed effects in *lmer()* of *lme4* [39]) were dummy coded to 0 or 1. Finally, we included *Speaker* and *Item* as crossed random effects (intercepts only, due to model-specific restrictions). FLMM estimations (based on penalized splines) are evaluated on the basis of combined covariate effect and summed effects plots (s. 3.2. for details).

3. RESULTS

3.1. Preanalysis

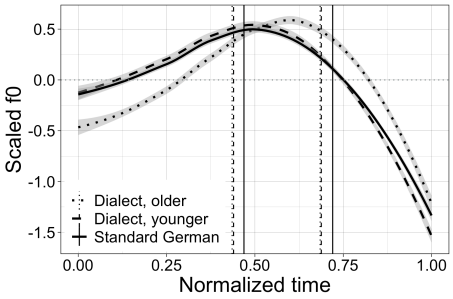


Figure 1: Mean loess-smoothed f0 contours across $V_1C_2V_2$-sequence, corresponding confidence intervals (level = 0.95) and mean segment boundaries per group.

Prior to FLMM, we visually inspected the aggregated f0 trajectories per speaker group (i.e., all three levels within one graph, which is impossible in FLMM) and across the longer $V_1C_2V_2$-sequence. Fig. 1. confirms that, on average, f0 patterns were rise-falls with peaks reached during C_2 (and not later in the sequence). This preliminary observation not only shows the realization of unambiguous nuclear peaks on the target words but particularly justified our focus on the rise within the V_1C_2-sequence in the subsequent FLMM.

$$t_{norm}(bound.) = \frac{t_{orig}(bound.) - t_{orig}(t_{norm}=0)}{t_{orig}(t_{norm}=1) - t_{orig}(t_{norm}=0)}. \quad (2)$$

A further observation is that, while the mean f0 curves of older WB speakers in Fig. 1 clearly show a later peak alignment and steeper f0 rises than those of SG speakers, younger WB speakers seem to pattern with the latter group at first glance. To factor out variation due to potential differences in segmental duration, we calculated the respective mean time-normalized positions of the C_2-on- and -offset with the formula in (2) and superimposed them on the f0 trajectories in Fig. 1 (as alternative to landmark registration; cf. [8]). This further revealed for both WB speaker groups (i.e., regardless of age) and in contrast to SG speakers (i) proportionally shorter V_1 and longer V_2 relative to the entire $V_1C_2V_2$-sequence (s. Table 3 for mean positions) and (ii) a delayed peak alignment in reference to the C_2-onset. Thus, the apparently similar H alignment of both SG and younger WB speakers in Fig. 1 may in fact be caused by temporally diverging segmental structures.

Speaker group	C_2-onset	C_2-offset
Dialect, older	0.44	0.68
Dialect, younger	0.44	0.69
Standard German	0.47	0.72

Table 3: Mean values for time-normalized C_2-onset and C_2-offset positions per speaker group.

3.2. FLMM *Variety*: Standard vs. Dialect speakers

For the between-group comparison of *Variety*, each of the analyzed 636 f0 trajectories consisted on average of 31.44 data points (range: 8-56 data points). The solid line in Fig. 2a shows the mean covariate effect ($f_1(t)$). Significance is claimed at time points at which the point-wise confidence bands (dashed lines) do not include zero. Fig. 2b., in turn, displays the corresponding summed effects curves, calculated by adding the effect curve in plot a. to the reference mean (or 'baseline', $f_0(t)$), which represents the covariate (here: *Variety*) set to zero (0: Standard, 1: Dialect). Both plots need to be assessed together when interpreting potential differences.

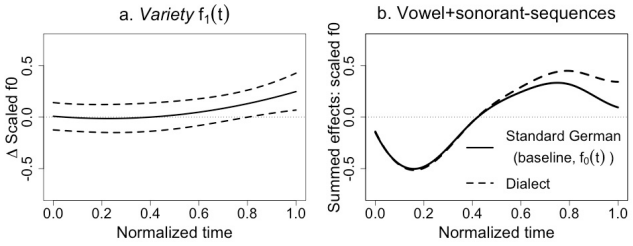


Figure 2: Between-group comparison of *Variety*: covariate effect (a.) and summed effects curves (b.).

In Fig. 2b., the estimated scaled f0 contour for SG speakers (solid baseline: $f_0(t)$) shows a clear peak between time points 0.7 and 0.8. According to Fig. 2a., *Variety* has a rising effect on scaled f0 from around this peak onwards. Comparing the baseline to the model contour for WB speakers (dashed line in Fig. 2b.: $f_0(t) + f_1(t)$) indicates that WB speakers show higher f0 values towards the end of the analyzed V+S-sequence than SG speakers. Together with the time points of C_2-onset and C_2-offset in Table 3, this suggests that dialect speakers realize f0 peaks significantly later than standard speakers.

3.3. FLMM *Age*: Younger vs. older Dialect speakers

The between-group comparison of *Age* among WB speakers consisted of 340 f0 trajectories with a mean of 31.11 data points per curve (range: 8-56 data points). In this analysis, the solid reference mean in Fig. 3b. represents younger WB speakers (0: younger, 1: older). The sum of the baseline and the covariate effect curve of *Age* (Fig. 3a.) is, hence, representative of older WB speakers (dashed line). Again, the baseline shows a clear peak between time points 0.7 and 0.8. The ostensible peak delay in older WB

speakers in Fig. 3b. (and Fig. 1 for that matter) is, however, not systematic as the confidence bands in Fig. 3a. only approach the zero-reference line towards the end of the V+S-sequence without reaching significance. Thus, we do not find significant effects of *Age* on scaled f0 within the dialect group.

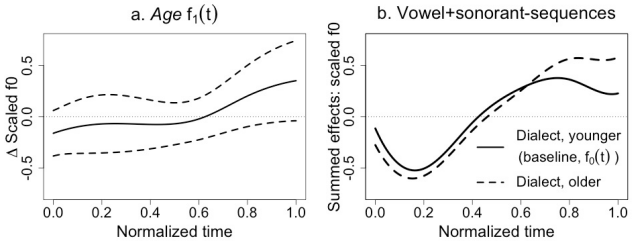


Figure 3: Between-group comparison of *Age*: covariate effect (a.) and summed effects curves (b.).

3.4. Relationship of duration and timing of f0 maxima

In a final step, we examined the relationship of segmental duration and f0 alignment. To this end, we tested for a correlation between the temporal synchronization of the automatically detected f0 maximum with each speaker's loess-smoothed, time-normalized scaled f0 curve and the again longer $V_1C_2V_2$-sequence using Pearson's *r* (cf. [40]). In rarer cases of two peaks, we selected the second (older WB: 9 %; younger WB: 14 %; SG: 16 %).

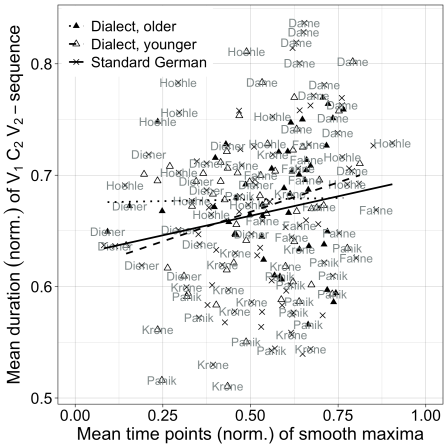


Figure 4: Mean values for both normalized $V_1C_2V_2$-durations and time points of f0 maxima, averaged across repetition. 3 outliers were excluded (x close to 0 or 1).

Commensurate with Fig. 4, younger WB and, to a lesser extent, the SG speakers showed a weakly positive correlation between duration and alignment indicating delayed peaks in longer sonorant sequences. The correlation reached significance in the former group (r = 0.26, df = 56, $p_{one\text{-}tailed}$ = 0.02) and approached it in the latter (r = 0.17, df = 92, $p_{one\text{-}tailed}$ = 0.06). Due to heteroscedastic residual variance, we used rank-based Spearman's ρ as non-parametric test for older WB speakers. Again in line with Fig. 4, no significant correlation was found (ρ = 0.07, $p_{one\text{-}tailed}$ = 0.32), suggesting no effect of duration on alignment in older WB speakers.

4. DISCUSSION AND CONCLUSION

Two main findings arise from this study. First, WB speakers aligned non-contrastive nuclear peaks significantly later than SG speakers. This finding is in line with previous findings on peak delay in southern compared to northern German varieties [2-5] and confirms our first hypothesis. Unlike the SG speakers in [8] who differed regarding the underlying segment (V_1 vs. C_2), the German speakers of this study reached H on average during C_2. This suggests, on the one hand, rather fine phonetic differences in peak alignment between the close contact varieties spoken in Bavaria that differ greatly in register [17] and, on the other, more substantial differences between geographically distant national standard varieties.

Second, although the two WB age groups did not differ in peak alignment, their relationship between the target sequence duration and the timing of the f0 maxima did: Younger but not older WB speakers displayed a weak yet statistically significant positive correlation similar to SG. This result reveals the potential complexity underlying a prosodic change, which apparently not only affects segmental quantity [19-21] but also peak alignment which had been shown to be intertwined in previous studies [22, 23] but without reference to sound change. The correlation differences between the WB age groups are indicative of diverging mechanisms to deal with the available sonorant material, despite the otherwise and perhaps surprisingly similar segmental timing. Younger WB speakers' patterning with SG speakers in this respect can be interpreted as a result of more general trends for dialect levelling in this variety.

Given the identical segmental structure of all analyzed long vowel plus sonorant sequences offering sufficient voiced material for the realization of the intended f0 contours, no right-sided time pressure effects occurred across speaker groups in this study (cf. [8]). The positive correlation found for both younger WB and SG speakers might still reflect a form of truncation by which f0 movements are cut off abruptly under increased time pressure [41] and which has been found to be predominantly used in SG phrase-final falls (cf. [42]). Even though a general pattern to compensate for limited voicing time in WB is yet unknown (but s. [43] on compression in a Southern Bavarian dialect), our results may indicate a non-truncation pattern in older WB. To further address these potential compensation mechanisms as well as their influence on peak alignment, future studies should include target words with short vowels and/or obstruents.

5. ACKNOWLEDGEMENTS

R scripts and data frames are available at Open Data LMU (doi.org/10.5282/ubm/data.378). The study was funded by DFG-DACH grant KL 2697/1-2 “Typology of Vowel and Consonant Quantity in Southern German varieties” awarded to F. Kleber.

6. REFERENCES

[1] D. R. Ladd, *Intonational Phonology*, 2 ed. Cambridge: CUP, 2008.
[2] M. Atterer and D. R. Ladd, "On the phonetics and phonology of “segmental anchoring” of F0: evidence from German," *J. Phon.,* v. 32, n. 2, pp. 177-97, 2004.
[3] B. Braun, "Effects of dialect and context in the realisation of German prenuclear accents," presented at the 16th ICPhS, Saarbücken, Germany, 2007.
[4] D. Mücke, M. Grice, A. Hermes, and J. Becker, "Prenuclear Rises in Northern and Southern German," presented at the 4th Speech Prosody, Campinas, Brazil, 2008.
[5] I. Mennen and D. Chousi, "Prosody in first-generation adult immigrants and second-generation heritage-language users: the timing of prenuclear rising accents," presented at the 9th Speech Prosody, Poznań, Poland, 2018.
[6] D. Mücke, M. Grice, J. Becker, and A. Hermes, "Sources of variation in tonal alignment: Evidence from acoustic and kinematic data," *J. Phon.,* v. 37, n. 3, pp. 321-38, 2009.
[7] D. Mücke, A. Hermes, and S. Tilsen, "Strength and structure: Coupling tones with oral constriction gestures," presented at the 20th Interspeech, Graz, Austria, 2019.
[8] J. Pöhnlein and F. Kleber (in preparation). Allophonic and phonemic differences in nuclear rises in Munich, Viennese and Zurich Standard German.
[9] B. Möbius and M. Jilka, "Effects of syllable structure and nuclear pitch accents on peak alignment: a corpus-based analysis," presented at the 16th ICPhS, Saarbrücken, Germany, 2007.
[10] A. Schepman, R. Lickley, and D. R. Ladd, "Effects of vowel length and “right context” on the alignment of Dutch nuclear accents," *J. Phon.,* v. 34, n. 1, pp. 1-28, 2006.
[11] J. Fitzpatrick-Cole, "The alpine intonation of Bern Swiss German," presented at the 14th ICPhS, SF, CA, 1999.
[12] P. Gilles, *Regionale Prosodie im Deutschen. Variabilität in der Intonation von Abschluss und Weiterweisung*. Berlin: De Gruyter, 2005.
[13] R. Wiese, *The Phonology of German*. NY: OUP, 1996.
[14] M. Jessen, *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins, 1999.
[15] G. Seiler, "On the development of the Bavarian quantity system," *IJGLSA,* v. 10, n. 1, pp. 103-29, 2005.
[16] R. Bannert, *Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage*. Lund: CWK Gleerup, 1976.
[17] P. Wiesinger, "The Central and Southern Bavarian Dialects in Bavaria and Austria," in *The Dialects of Modern Germany. A Linguistic survey*, C. V. J. Russ Ed. Abingdon: Routledge, 1990, pp. 438-519.
[18] P. Trudgill, *Dialects in Contact*. Oxford: Blackwell, 1986.
[19] F. Kleber, "VOT or quantity: what matters more for the voicing contrast in German regional varieties? Results from apparent-time analyses," *J. Phon.,* v. 71, pp. 468-86, 2018.
[20] F. Kleber, "Complementary length in vowel–consonant sequences: Acoustic and perceptual evidence for a sound change in progress in Bavarian German," *JIPA,* v. 50, n. 1, pp. 1-22, 2020.
[21] M. Jochim and F. Kleber, "Fast-Speech-Induced Hypoarticulation Does Not Considerably Affect the Diachronic Reversal of Complementary Length in Central Bavarian," *Lang Speech,* v. 0, n. 0, 2022.
[22] D. R. Ladd, I. Mennen, and A. Schepman, "Phonological conditioning of peak alignment in rising pitch accents in Dutch," *JASA,* v. 107, n. 5, pp. 2685-96, 2000.
[23] D. R. Ladd, A. Schepman, L. White, L. M. Quarmby, and R. Stackhouse, "Structural and dialectal effects on pitch peak alignment in two varieties of British English," *J. Phon.,* v. 37, n. 2, pp. 145-61, 2009.
[24] C. Draxler and K. Jänsch, "SpeechRecorder – a universal platform independent multi-channel audio recording software," presented at the 4th LREC, Lisbon, Portugal, 2004.
[25] C. Draxler and K. Jänsch, "Wikispeech – a content management system for speech databases," presented at the 9th Interspeech, Brisbane, Australia, 2008.
[26] T. Kisler, U. D. Reichel, F. Schiel, C. Draxler, B. Jackl, and N. Pörner, "BAS Speech Science Web Services – an update on current developments," presented at the 10th LREC, Portorož, Slovenia, 2016.
[27] M. Jochim, "Extending the EMU Speech Database Management System: Cloud Hosting, Team Collaboration, Automatic Revision Control," presented at the 18th Interspeech, Stockholm, Sweden, 2017.
[28] R. Winkelmann, J. Harrington, & K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Comput Speech Lang,* v. 45, pp. 392-410, 2017.
[29] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun Ed. Oxford: OUP, 2005, pp. 55-83.
[30] *wrassp: an R wrapper to the ASSP Library. R pkg. v. 1.0.1*. (2021).
[31] *emuR: Main Package of the EMU Speech Database Management System. R pkg. v. 2.3.0*. (2021).
[32] *R: A language and environment for statistical computing*. (2022). R Foundation for Statistical Computing, Vienna, Austria. Available: R-project.org
[33] *RStudio: Integrated Development Environment for R*. (2022). RStudio, PBC, Boston, MA. Available: rstudio.com
[34] L. S. Bučar Shigemori and M. Pouplier, "Realization of F0 on vocalic and consonantal nuclei in Slovak," presented at the 12th ISSP, Providence, RI (virtual), 2020.
[35] J. Kirby, F. Kleber, J. Siddins, and J. Harrington, "Effects of prosodic prominence on obstruent-intrinsic F0 and VOT in German," presented at the 10th Speech Prosody, Tokyo, Japan (virtual), 2020.
[36] J. Cederbaum, M. Pouplier, P. Hoole, and S. Greven, "Functional linear mixed models for irregularly or sparsely sampled data," *Stat Model,* v. 16, n. 1, pp. 67-88, 2016.
[37] M. Pouplier, J. Cederbaum, P. Hoole, S. Marin, and S. Greven, "Mixed modeling for irregularly sampled and correlated functional data: Speech science applications," *JASA,* v. 142, n. 2, pp. 935-46, 2017.
[38] *sparseFLMM: Functional Linear Mixed Models for Irregularly or Sparsely Sampled Data. R pkg. v. 0.4.2*. (2022). Avail.: github.com/alexvolkmann/sparseFLMM
[39] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J Stat Softw,* v. 67, n. 1, pp. 1-48, 2015.
[40] N. Levshina, *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins, 2015.
[41] E. Grabe, "Pitch accent realization in English and German," *J. Phon.,* v. 26, n. 2, pp. 129-43, 1998.
[42] T. V. Rathcke, "How Truncating Are ‘Truncating Languages’? Evidence from Russian and German," *Phonetica,* v. 73, pp. 194-228, 2016.
[43] J. Siddins & I. Mennen, "Pitch Accent Realisation in Austrian German," presented at the 19th ICPhS, Melbourne, Australia, 2019.

PROSODY-SYNTAX INTERFACE AND HUNGARIAN NOUN PHRASES

Corinna Langer and Frank Kügler

Goethe-University Frankfurt, Germany
langer@lingua.uni-frankfurt.de, kuegler@em.uni-frankfurt.de

ABSTRACT

We investigate the prosody-syntax interface of Hungarian noun phrases (NPs) with two types of complex modification: a noun modified by (A) two adjectives or (B) an adjective with an argument. In Hungarian, (narrow, exhaustive) focus is typically marked syntactically in the immediate pre-verbal position. However, [1] found additional prosodic focus marking in modified NPs occupying this position. Replicating [1] with increased complexity of the modification, our results confirm that additional prosodic focus marking is present in syntactically focused NPs. However, the prosodic pattern varies depending on the syntactic structure inside the NP. In Modification A, the prosodic pattern differs between all possible focus domains. In Modification B, the adjective and its argument cluster together forming one modification that mirrors [1]'s results. The variation found in prosody can hence be explained by distinct syntactic structures that match into prosodic structure.

Keywords: Hungarian, prosody, focus, noun phrase, modification

1. INTRODUCTION

Hungarian is a language with a relatively free word order. In the post-verbal domain, it is completely free, while the pre-verbal field is constrained by information structure (IS). This so-called 'discourse-configurationality' (see [2]) is marked in Hungarian by specific positions for topic (sentence-initial) and focus (immediate pre-verbal). Based on recent findings that Hungarian uses additional prosodic focus marking to support or even disambiguate syntactic focus marking (see, e.g., [1], [3]), this study investigates whether the additional prosodic marking is influenced by the syntactic structure of the NP.

We base our study on the IS notions *focus* and *information status*, which have been extensively discussed in the literature on Hungarian (syntax) (e.g., [4], [5]). For focus, we adopt Krifka's [6] definition that focus indicates the selection of an element out of a set of relevant alternatives. Information status concerns the common ground (CG, [7], [8]) of interlocutors. *Given* elements have been previously mentioned. *New* elements, on the other hand, are newly introduced into the CG.

Cross-linguistically, focus is typically marked through prominence and givenness through reduced prominence [9]. Prominence can, for example, be expressed by syntax, prosody or both (see, e.g., [10]). In Hungarian, prominence is typically marked by syntax, i.e. by placing the focused element in the focus position (see, e.g. [2], [4], [11]). The Hungarian prosody supports this syntactic marking: the focused element is assigned the main accent of the sentence, 'reducing' or 'erasing' all following accents. In a neutral (broad focus) sentence without a filled focus position, every content word bears a pitch accent which is downstepped in relation to the preceding one (see, e.g., [2], [11], [12], [13]). Both the main accent on focused elements and 'normal' downstepped accents in neutral sentences are typically realized as falling contours (see, e.g. [14], [15]). However, rising/plateau contours have been observed on, for example, topics (see, e.g., [3]). All accents are word-initial, because Hungarian is a left-headed language (see, e.g. [12], [16]).

In recent years, there has been a growing number of empirical studies investigating the Hungarian IS-syntax-prosody interface (see, e.g. [1], [3], [17], [18]). In particular, [1] found not only supporting but additional prosodic focus marking in sentences where syntactic marking alone can not disambiguate between multiple possible focus domains. Their findings are based on a production study with modified NPs in the syntactic focus position. In these types of sentences, the same linearization leads to multiple possible interpretations. For example, the sentence in (1) can answer each of the three given context questions.

(1) Az idős író-t hívta meg az újságíró.
the old writer-ACC invited VPRT the journalist
'The journalist invited the old writer.'
C1 Who did the journalist invite?
C2 Which writer did the journalist invite?
C3 Which old person did the journalist invite?

As expected, the main accent falls on the focus position with the highest prosodic prominence on

the left-edge of the NP. However, the prosodic pattern inside the NP differs depending on the focus domain: if the focus is on the whole NP, the noun keeps its downstepped accent (as in broad focus), if there is narrow focus on the noun, the accent is enhanced, and if there is narrow focus on the adjective, the noun is deaccented. The accent on the leftmost element is enhanced in all three conditions.

Based on these results, our research questions are: (Q1) Are the patterns found in [1] also used for NPs with a higher complexity? (Q2) Do the patterns differ depending on the syntactic structure of the complex modification (Figure 1)? In our production study, we will look at two types of complex modification that differ in their syntactic structure. In complex modification A (Mod-A), the noun is modified by two adjectives (Adj1 & Adj 2), and in complex modification B (Mod-B), the noun is modified by an adjective (Adj) and its argument (Arg). If prosody is assumed to reflect syntactic structures (e.g., [19]), the prosodic patterns of the two modification-types should differ.

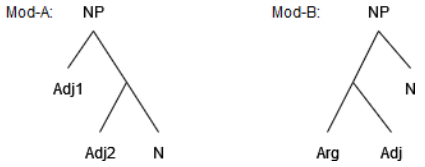


Figure 1: Syntactic structures of the two complex modification types used in the production study

2. PRODUCTION STUDY

2.1. Method

Replicating the method from [1], we conducted a production study with two types of complex modification (mod-types) in the syntactic focus position (Figure 1) as opposed to simple modified nouns in [1]. In Mod-A, a second adjective was added and in Mod-B, the adjective was either *képes* ('able') or *keptelen* ('unable') with a verb as its argument. Examples are shown in (2) (Mod-A) and (3) (Mod-B).

(2) a félénk barna eger-et
the shy brown mouse-ACC
'the shy, brown mouse ...'

(3) az énekel-ni képes eger-et
the sing-INF able mouse-ACC
'the mouse, which can sing ...'

There were five sentences per mod-type, each elicited through contexts ending in a question that prompted one of the five possible focus domains (Table 1), resulting in 50 target sentences in total. All elements in the target sentence were *given* to exclude influences of information status.

	Mod1_F	Mod2_F	Mod_F	N_F	NP_F
Mod-A	Adj1	Adj2	Adj1&2	N	NP
Mod-B	Arg	Adj	Arg&Adj	N	NP

Table 1: Possible focus domains per mod-types

In the experiment, the contexts were presented visually (written form) and auditorily (spoken by a female native speaker). The participants (20 Hungarian native speakers) were then asked to answer the question at the end of the context with the visually presented answer. The order of the contexts (and filler-contexts) was randomized. In total, 50 x 20 = 1000 sentences were recorded.

2.2. Data Analysis

All sentences were manually labeled for word, syllable and vowel with Praat [20]. Using [1]'s Praat scripts, we measured the duration of syllables, f0-minima and maxima per vowel and their position in relation to the syllable. These points were checked manually to prevent calculation errors. At the same time, the number and position of creaky syllables and pauses were extracted. We also measured the overall pitch- and intensity-contours of the NPs at ten equidistant points per syllable. All pitch-values were measured in semitones with speaker-dependent baselines (see, e.g. [21]).

The ten points per syllable were used for generalized additive mixed models (GAMMs - R [22] package mgcv [23])) calculating trajectories of the mean f0/intensity values for each focus condition and windows of significant differences between them. From the f0-minima and -maxima, contour types for every word in the NP and the slope of rising/falling contours were calculated using R scripts. For the contours, we took one semitone as the threshold between a plateau and a rise or fall (see [1]). Words with creaky voice were manually checked and excluded from the slope calculations (240 of 3000 data points), to prevent skewed results. Two datapoints were also excluded from the contour type analysis, because the contour was unclear.

2.3. Results

All sentences show the prosodic pattern of a narrowly focused sentence in Hungarian: the main accent falls on the focus position with reduced prominence on following elements. In the next sections, we will take a closer look at the prosodic patterns inside the target NP.

Intensity and duration do not play a major role in our data. There are small windows of significant differences in the GAMMs for intensity and one ANOVA shows a significant effect of FOCUS on duration ($df = 4$, $f = 4.389$, $p < 0.01$): in Mod-B

the first syllable of the noun is longer in N_F than in all other conditions. However, while these results support our prosodic analysis they all correlate with effects found in the f0-patterns. Thus, due to space limitations, we'll concentrate on pitch. There are significant results in the mean f0-trajectories (GAMMs), the count of contour types (Chi-square tests) and the slopes on falling contours (ANOVAs). There are not enough pauses for a statistic analysis and no effect of voice quality alone (see [1] for a discussion about its interaction with pitch).

In both mod-types, the highest f0 is on the first syllable of the modification. This is in line with [1]'s findings that the highest prosodic prominence falls on the left edge of the focus domain. However, the patterns inside the NP differ between Mod-A and B.

2.3.1. *Complex Modification A*

In Mod-A, the GAMM (Figure 2) shows a significant effect of FOCUS ($p < 0.0001$ all conditions). The pairwise comparisons reveal that Mod1_F (green) has a significantly lower mean-f0 at the second half of Adj1 and Mod2_F (blue) has a significantly higher mean-f0 at the first half of Adj2 compared to all other focus conditions except N_F (red). It is important to note that the difference on, for example, Adj1 can either be explained by a lower number of rising and/or plateau contours in Mod1_F focus or be due to differences in the slopes of falling contours on Adj1 between the focus domains. Thus, we will now take a closer look at these two measurements.

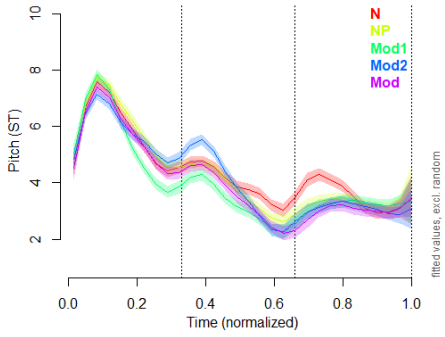


Figure 2: Mod-A: GAMM of the five mean-f0 trajectories (in semitones) per focus in the NP

Table 2 shows the counted contour types on each word in the target NP. The Chi-square test shows a significant correlation between FOCUS and CONTOUR TYPE on Adj2 (X-squared = 24.932, df = 8, p-value < 0.01) and N (X-squared = 20.895, df = 8, p-value < 0.01). Both words have significantly more falling contours when they are narrowly focused and more plateau contours post-focally, and Adj2 has significantly more rising contours when it precedes the narrow focus (N_F).

Focus	Adjective 1			Adjective 2			Noun		
	F	P	R	F	P	R	F	P	R
Mod1_F	89	5	6	67	28	5	49	22	29
Mod2_F	68	13	19	87	12	1	59	21	20
Mod_F	74	7	18	77	18	5	48	34	18
N_F	78	10	12	66	20	14	69	13	18
NP_F	77	9	14	74	19	7	49	24	27

Table 2: Mod-A: Distribution of contour types on each word; F = Fall; P = Plateau; R = Rise

While position (timing) of the f0-peaks does not play a role in our data, their height differs between focus domains, resulting in different slopes of falling contours (Figure 3). The ANOVA shows significant effects of WORD ($df = 1$, $f = 82.07$, $p < 0.001$) and the interaction of WORD and FOCUS ($df = 2$, $f = 4.29$, $p < 0.001$). The post-HOC TukeyHSD test shows no significant difference between the slopes on Adj1 and Adj2 per focus condition. The slopes on Adj1 are significantly steeper than all other accents, except accents on narrowly focused elements. The only difference between focus conditions is on the noun: it is significantly steeper in N_F (red) than in Adj1&2_F (blue&green).

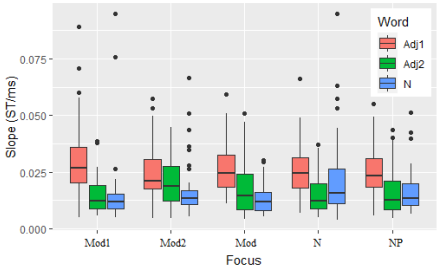


Figure 3: Mod-A: Slopes of the falling contours on the words in the target NP per focus condition

2.3.2. *Complex Modification B*

For Mod-B, the GAMM (Figure 4) shows a significant effect of FOCUS ($p < 0.0001$ in all conditions). The pairwise comparisons reveal that N_F has a significantly higher mean-f0 than all other conditions on the Adj and the noun. NP_F has a significantly higher mean-f0 at the first half of the Arg than Mod2_F and Mod_F.

The Chi-square test for the counted contour types (Table 3) shows significant correlations between FOCUS and CONTOUR TYPE on the adjective (X-squared = 15.742, df = 8, p-value < 0.05) and the noun (X-squared = 21.894, df = 8, p-value < 0.01). N_F has significantly more plateau contours on the Adj and more falling contours on the noun.

Figure 5 shows the slopes of falling contours in Mod-B. The ANOVA shows significant effects of WORD ($df = 2$, $f = 67.09$, $p < 0.001$), FOCUS ($df = 4$, $f = 3.64$, $p < 0.01$) and the interaction of WORD

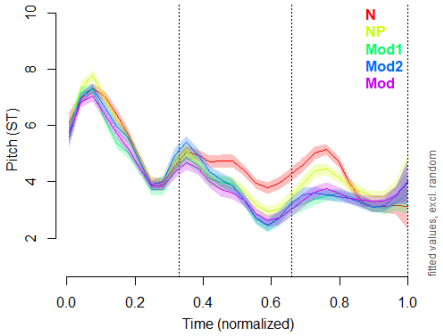


Figure 4: Mod-B: GAMM of the five mean-f0 trajectories (in semitones) per focus in the NP

Focus	Argument			Adjective			Noun		
	F	P	R	F	P	R	F	P	R
Mod1_F	90	2	8	75	17	8	57	23	20
Mod2_F	87	5	8	70	16	14	55	27	18
Mod_F	89	4	7	62	26	12	52	23	25
N_F	86	9	5	52	31	17	78	8	14
NP_F	93	2	5	67	19	14	56	19	24

Table 3: Mod-B: Distribution of contour types on each word; F = Fall; P = Plateau; R = Rise

and FOCUS ($df = 8$, $f = 4.97$, $p < 0.001$). The post-hoc TukeyHSD test shows no significant differences between the slopes on the Arg and Adj. The slopes on Arg are significantly steeper than all other accents except the accents on the noun in N_F and NP_F. On the noun, N_F has a significantly steeper slope than the other focus types except NP_F.

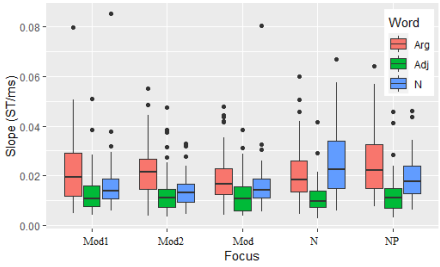


Figure 5: Mod-B: Slopes of the falling contours on the words in the target NP per focus condition

3. DISCUSSION

Regarding our research questions, the data show that (Q1) the patterns found in [1] are also found in NPs with a higher complexity. The exact pattern however differs depending on the syntactic structure of the complex modification (Q2).

In general, the highest prosodic prominence (f0) falls on the leftmost element of the NP (see also [2], [11], [12], [13]). In the majority of the data (77% Mod-A; 89% Mod-B) this element has a falling contour. Similar to [17], no systematic difference between the acoustic details of this contour, i.e. f0-minima, -maxima and slope, was found. While there is some variation in the data (similar to [3]'s results), as in [1], the default-pattern consists of falling contours on (pre-)focal elements and a higher number of plateau contours post-focally. The relatively high number of rises on the noun (Table 2&3) can be explained by a high boundary tone at the end of the NP.

In Mod-A the prosodic pattern differs between all five possible focus domains and follows the patterns found in [1]: the accent on the left edge of the target NP that contains the focus and on narrow focused elements is enhanced (see, e.g., [24] for 'boosting by focus'), post-focal elements have a reduced accent and elements that are part of a broader focus domain, for example in NP_F, are 'normal' (downstepped, see neutral sentences as in [14]). However, not all of the distinctions are as clear as in [1]. This is presumably due to the limited pitch range that speakers can use. The longer a phrase is, the smaller the differences can be. Follow up experiments should test, if the subtle differences between the three accent-types are relevant in perception, or if a more basic distinction between presence or absence of an accent suffices.

Mod-B differs from Mod-A in that the two parts of the modification cluster together. There is no significant difference between Mod1_F, Mod2_F and Mod_F. This is predicted by the syntactic structure where the adjective and its argument are syntactic sisters as opposed to the recursively embedded Adj1&2 in ModA (Figure 1). If they are analyzed as being one single modification, Mod-B perfectly mirrors [1]'s results on simple modification with a single adjective: the left edge and narrowly focused elements have enhanced accents, post-focal elements are reduced and elements that are part of a broader focus domain keep their downstepped accent.

4. CONCLUSION

Our study replicates the results from [1] that prosody does not only support (see, e.g., [3], [11], [12], [13], [17]) but adds to syntactic focus marking in Hungarian. While the main sentence accent, as expected, falls on the syntactic focus position, the prosodic patterns inside the position disambiguate between multiple possible focus domains that syntax alone cannot. However, the exact prosodic pattern is based on syntactic structure (Figure 1). In Mod-B, where adjective and argument syntactically form one modification, the prosodic pattern is the same as for simple modification [1]. In Mod-A, the recursive syntactic structure is mirrored in the prosodic pattern. Thus, predictions from the syntax-prosody interface impact phrase-level prosody in Hungarian.

5. ACKNOWLEDGEMENTS

We are grateful to our student assistant Noémi Ecsedi for her help with creating stimuli, recording and data processing. This work was supported by the DFG (KU2323/4-1). Many thanks to the anonymous reviewer of the 'International Congress of Phonetic Sciences 2023' for the feedback.

6. REFERENCES

[1] C. Langer and F. Kügler, "Focus and Prosodic Cues in Hungarian Noun Phrases," *Proc. 1st International Conference on Tone and Intonation (TAI)*, pp. 219–223, 2021.

[2] K. É. Kiss, *Discourse Configurational Languages*. New York/Oxford: Oxford University Press, 1995.

[3] S. Genzel, I. Shinichiro, and B. Surányi, "The prosodic expression of focus, contrast and givenness: A production study of Hungarian," *Lingua*, vol. 165, pp. 183–204, 2015.

[4] K. É. Kiss, "Topic and focus: Two structural positions associated with logical functions in the left periphery of the Hungarian sentence," *Acta Linguistica Hungarica*, vol. 55.3-4, pp. 287–296, 2008.

[5] L. Hunyadi, "Remarks on the syntax and semantics of topic and focus in Hungarian," *Acta Linguistica Academiae Scientiarum Hungaricae*, vol. 31.1/4, pp. 107–136, 1981.

[6] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica*, vol. 55.3-4, pp. 243–276, 2008.

[7] M. Krifka and R. Musan, Eds., *The expression of information structure*. Berlin: Mouton de Gruyter, 2012.

[8] R. Stalnaker, "Common ground," *Linguistics and Philosophy 25*, 2002.

[9] F. Kügler and S. Calhoun, "Prosodic Encoding of Information Structure. A typological perspective," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 453–467.

[10] M. Zimmermann and C. Féry, Eds., *Information structure. Theoretical, typological, and experimental perspectives*. Oxford: Oxford University Press (Oxford linguistics), 2010.

[11] I. Kenesei, "Adjuncts and Arguments in VP-focus in Hungarian," *Acta Linguistica Hungarica*, vol. 45.1-2, pp. 61–88, 1998.

[12] L. Varga, *Intonation and Stress: Evidence from Hungarian*. New York: Palgrave Macmillan, 2002.

[13] K. Szendrői, "A stress-based approach to the syntax of Hungarian focus," *The Linguistic Review*, vol. 20.1, 2003.

[14] L. Kálmán, "Word order in neutral sentences," in *Approaches to Hungarian*, I. Kenesei, Ed. Szeged: JATEPress, 1985, vol. Vol. 1., pp. 25–37.

[15] ——, "Word order in non-neutral sentences," in *Approaches to Hungarian*, I. Kenesei, Ed. Szeged: JATEPress, 1985, vol. Vol. 1., pp. 13–23.

[16] G. Olaszy, "A beszéd szupraszegmentális szerkezete [Suprasegmental structure of speech]," in *A magyarbeszéd: Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek. [Hungarian speech: Speech research, speech technology and systems]*, G. Olaszy and G. Németh, Eds. Akadémiai Kiadó, 2010.

[17] K. Mády, "Prosodic (non-)realisation of broad, narrow and contrastive focus in Hungarian: a production and a perception study," in *Proc. Interspeech 2015, Dresden*, 2015, pp. 948–952.

[18] K. Mády and F. Kleber, "Variation of pitch accent patterns in Hungarian," in *5th Speech Prosody Conference, Chicago, USA*, 2010.

[19] E. O. Selkirk, "The syntax-phonology interface," in *The Handbook of Phonological Theory*, 2nd ed., J. Goldsmith, J. Riggle, and A. Yu, Eds. Oxford: Blackwell Publishing, 2011, pp. 435–484.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," URL: http://www.praat.org/, 2017.

[21] G. Fant, A. Kruckenberg, K. Gustafson, and J. Liljencrants, "A new approach to intonation analysis and synthesis of Swedish," *TMH-QPSR 44*, pp. 161–164, 2002.

[22] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/, 2017.

[23] S. Wood, "Generalized Additive Models: An Introduction with R," Chapman and Hall/CRC, 2017, 2. edition.

[24] C. Féry and I. Shinichiro, "How focus and givenness shape prosody," in *Information structure. Theoretical, typological, and experimental perspectives*, M. Zimmermann and C. Féry, Eds. Oxford: Oxford University Press (Oxford linguistics), 2010, pp. 36–65.

SOMETIMES LOW REALLY IS JUST THE OPPOSITE OF HIGH: PERCEPTION OF LOW F0 TARGETS BY TONE AND NON-TONE LANGUAGE SPEAKERS

Jonathan Barnes[1], Chris K. C. Lee[1], Alejna Brugos[2], Stefanie Shattuck-Hufnagel[3], & Nanette Veilleux[2]

[1] Boston University, [2] Simmons University, [3] Massachusetts Institute of Technology
jabarnes@bu.edu, chriskcl@bu.edu, alejna.brugos@simmons.edu, sshuf@mit.edu, veilleux@simmons.edu

ABSTRACT

Previous perceptual studies show that High pitch accents with maximum F0 extended in time sound higher than sharp-peaked counterparts with identical maximum F0 to listeners of various language backgrounds. It is not clear, however, whether this 'plateau' effect extends to the perception of Low pitch accents. We report results from two perception tasks involving American English (NE) and Cantonese-English bilingual (BiL) speakers, aimed at uncovering influences of contour shape on perception of Low and High pitch accents in English. We found that while the 'plateau' effect is demonstrated by both groups in the case of high pitch accents, only BiL participants judged a low pitch accent with a temporally extended minimum F0 to be lower than a sharp-turn contour with identical minimum F0. We discuss our findings from the perspective of the language-specific functional load placed on discrimination of Low pitch targets, and an averaging-based model of F0 scaling perception.

Keywords: Intonation, Pitch perception, tone scaling, tonal timing, Tonal Center of Gravity

1 INTRODUCTION

Intonation researchers have learned a great deal over the past several decades about the production and perception of High intonational pitch accents, both within and across languages. Considerably less is known about corresponding aspects of their Low counterparts. There is, in fact, a broad sense in the literature that Low and High tones may behave differently in implementation (e.g., [1], "High is not just the opposite of Low"). The nature of those differences, however, awaits further investigation.

Much attention has been paid recently, for example, to the effect of global F0 contour shape on tone perception. One of the best-known contour shape effects concerns differences between sharp-peaked and "plateau"-shaped High accents ([2]–[7], inter alia). Specifically, it has been demonstrated in a variety of languages that all else equal, a plateau-shaped pitch accent will sound higher in pitch to listeners than an analogous sharp-peaked accent with identical maximum F0 [4], [8], [9] (illustrated schematically in Fig. 1a). One possible explanation for this ([10]–[12]) rests on the assumption that perception of tone scaling involves not precise identification of a single target F0 level (maximum or minimum), but rather, a process of F0 averaging over some domain of interest à la [13], [14]. If this is correct, then we also might expect to find the same pattern for Low accents as for High, i.e. that an L* should sound lower to listeners when the minimum is extended to create a low flat region (a "basin", if you will), than when it is reached only as a sharp fall-rise pattern (a "gully", depicted schematically in Fig. 1b).

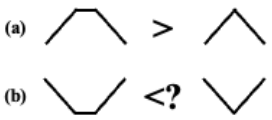


Figure 1: Schematics showing the effect of contour shape on F0 scaling of high (a) and low (b) pitch accents. "$>$" = *sounding higher than*; "$<$" = *sounding lower than*.

We know of only one study investigating potential perceptual differences between low "basins" and "gullies": [15] report on results showing that while Low basin-shaped pitch accents did in fact sound lower than narrow gully analogues, (British English-native) listeners were significantly less adept at discriminating pitches in the lower F0 range, reducing the salience of the plateau effect for Low accents as compared with Highs. Similar results involving reiterant speech and non-speech stimuli, furthermore, lead them to speculate that Low perception may be worse than High for domain-general psychoacoustic reasons. At the same time, they wonder whether speakers of languages with more Low F0 targets might not prove better at its perception, perhaps yielding a different pattern of experimental results.

In this study, we have two goals. The first is to determine whether low basin pitch accents in fact sound lower than sharp gully analogues, as the averaging-based model of F0 scaling perception put forward in [9] (called TCoG-F) predicts. The second is to determine whether the presence or magnitude of the effect could be influenced at all by the language experience of the listeners in question. To do this, we conducted a perception experiment involving Low and High pitch accents of English, judged by two sets of participants: native speakers

of American English, and Cantonese-English bilinguals from Hong Kong. Cantonese is a lexical tone language contrasting several tone categories with a Low F0 component: low level (T6) vs. low fall (T4) vs. low rise (T5). The tone inventory of Cantonese is illustrated in Fig. 2. It should be noted not only that Cantonese contrasts Low tones with different dynamic properties (fall, rise, level), but also that F0 level is one of the most salient acoustic cues, if not the primary cue [16] to the contrast between T6 and T4. (T4 reaches a lower minimum typically than T6.) The English spoken by this population is typically described as influenced in its phonology by its Cantonese substrate ([17], [18]).

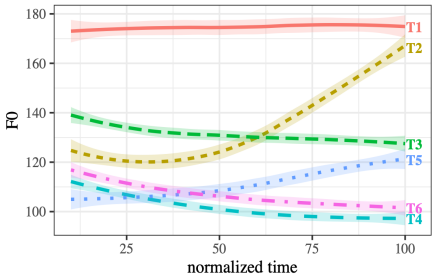


Figure 2: The six lexical tones in Cantonese produced by a male native speaker in citation form.

2 METHOD

Our research questions are thus (1) whether a basin-shaped L*-accented syllable would yield lower F0 scaling judgments relative to an analogous gully-shaped accent (just as high plateaux yield higher judgments relative to sharp peaks), and (2) whether listeners of different linguistic backgrounds might differ in these judgments, depending on the functional load their native languages place on the discrimination of low F0 levels. The study consists of two parts, one focused on replicating prior results involving high F0 plateaux, the other focused on low basins and their narrow gully analogues.

2.1 Stimulus creation

The base stimulus for both parts of the experiment was the sentence "*DAY* might fit", with nuclear accent on *DAY,* produced by a male native speaker of American English, then resynthesized using [19]. In both the **High** and the **Low** judgment tasks, each trial consists of a pair of contours, one we will call the "target", and another serving as the "reference". In the **High** judgment task, both contours take the form H* L-H% (the "uncertainty" contour of [20]). Target contours then bore one of four different shapes realizing the High nuclear pitch accent. In all four, the accented syllable began with a 250 ms F0 rise, from 141 Hz to 200 Hz. This was followed by either (1) an immediate F0 fall, creating a sharp peak coinciding with the end of the accented vowel, (2) a 25-ms plateau at the maximum F0, followed by a fall, (3) a 50-ms plateau, or (4) a 75-ms plateau. To keep the initial rise constant across all targets, rime duration for *DAY* increased in proportion with the duration of the plateaux.

In the **Low** judgment task, all stimuli bore the intonation contour L*+H L-H%, again with *DAY* bearing the nuclear accent. As with the High task, target stimuli took four different shapes. All of them began with a 250 ms fall from 141 Hz to 100 Hz, followed by (1) an immediate rise, creating a narrow gully, (2) a 25 ms low basin, followed by a rise, (3) a 50 ms low basin, or (4) a 75 ms low basin. Low accent contours were thus effectively the mirror image of High accent contours, in terms of shape and alignment of key turning points.

In both tasks, the reference contours that targets were paired with in each trial were identical to targets, except that F0 was flat during the accented syllable *DAY* at one of seven reference F0 levels. In the High task, the highest of these levels was at 200 Hz, identical to the F0 maximum in the High target contour. The remaining six levels formed a continuum declining in half semitone increments. (Reference contours remained perceptually plausible instantiations of H* L-H%.) In the Low task, the lowest reference level was flat throughout the accented syllable at 100 Hz, identical to the minimum in the Low target contour, with remaining reference levels forming a continuum increasing in half semitone increments. In the High task, for each target-reference pair, participants were asked to judge which of the two contours reached a higher pitch during the accented syllable. In the Low task, by contrast, they were asked to judge which of the two accented syllables sounded lower. Fig 3. illustrates these reference levels and target contours.

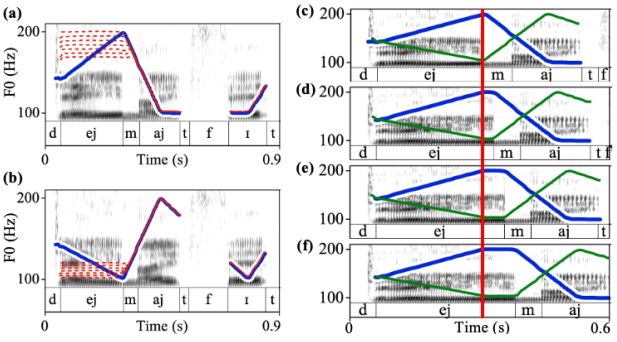


Figure 3: (left) F0 contours superimposed on spectrograms for reference (red) and sharp-turn (blue) target contours, for (a) high and (b) low judgment tasks. (right) The first 600 ms of target contours showing varying plateaux (blue) and basin (green) durations.

2.2 Procedure

The two tasks were presented to the participants in separate blocks (with order counterbalanced across participants). Participants listened to pairs of target contours and corresponding reference levels (order of presentation randomized) and decided which

DAY sounded higher (or lower, depending on the task). 18 additional **control trials** comparing reference levels separated by 2 or 3 continuum steps were included as a measure of participants' accuracy in discriminating pitch levels. There were 130 trials (4 shapes × 7 reference levels × 2 orders × 2 repetitions + 18 controls) in each task, presented in random order.

19 native American English (NE) speakers (13 female, 5 male, 1 non-binary, aged 18-21) and 22 Cantonese-English bilingual (BiL) speakers (13F, 9M, aged 22-33) reporting no speech or hearing deficits participated in the study. Fig. 4 shows accuracy rates in control trials for the two groups of participants. It should be clear that while BiL participants performed well in both tasks, the Low judgment task was challenging for NE participants. Since both groups generally had little difficulty discriminating control trials for the High task, we used participants' accuracy rates in those trials, with a threshold of 70%, as a criterion for inclusion in further analysis. Data from one speaker of each group were discarded.

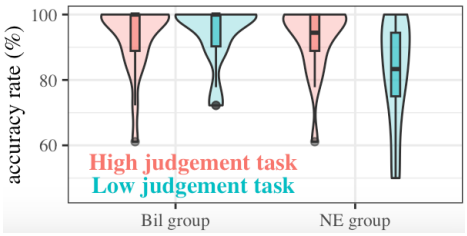


Figure 4: Participants' accuracy rates in control trials.

3 RESULTS

3.1 High judgment task

Fig. 5 displays results of the High judgment task, pooled across subjects separated by group. Lines represent the percentage of trials in which a given shape (*sharp turn*, *25-ms plateau*, *50-ms plateau*, *75-ms plateau*) was judged **higher** than each of its seven reference levels. Two trends are immediately obvious: (1) Across contour shapes, participants tend to judge target *DAY* as higher than the lowest reference levels, but lower than the highest levels. (2) As plateau length increases, participants' 'higher target' judgments decline later, suggesting that longer plateaux result in higher perceived pitch accent scaling (thus replicating previous findings). More interestingly, despite minor differences in the percentage of 'higher target' responses at each step, the two groups of participants display essentially the same response patterns. These observations are supported by mixed-effects logistic regression analysis, using reference level (continuous variable from 0 to 6, with 0 as the lowest level), target contour shape (continuous variable from 0 to 3, with 0 representing sharp turn), and participant group (factor variable with BiL as reference) as fixed effects, and participant and trial included as random intercepts. The best model, based on all-subset selection, contains a main effect of target contour shape (β=.433, z=10.72, p<.001), a main effect of reference level (β=-1.17, z=-12.42, p<.001) and an interaction between reference level and group (β=.193, z=-3.21, p=.001). No main effect of participant group is selected in the best model. This suggests that at least for high F0 target perception, listeners of different linguistic backgrounds are affected by contour shape in qualitatively the same way: high pitch accents with an extended plateaux sound higher than analogous sharp-peak accents with identical maximum F0.

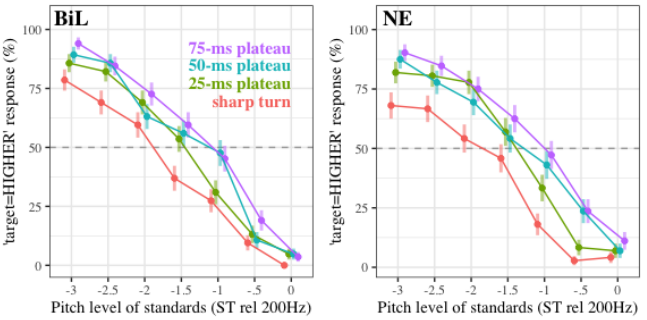


Figure 5: Percentage 'target is HIGHER' judgments for the four contour shapes, as a function of the reference level against which they were compared.

3.2 Low judgment task

Fig. 6 displays results from the Low judgment task, pooled across subjects separated by group. Lines represent the percentage of trials in which a given shape was judged **lower** than each reference level. Unlike in the high task, the two groups differ markedly in terms of how they perceived the scaling of target pitch accents: NE participants' 'lower target' response rate is almost always below 50%, meaning that they tend to perceive target *DAY* as higher in pitch than all but the highest reference levels. Additionally, NE participants show no clear pattern involving the shape or temporal extent of the Low pitch accents. In other words, low-basin contours sounded no lower to NE listeners than sharp-turn or gully-shaped low accents.

The response pattern of the BiL participants, by contrast, is essentially the mirror image of that seen in the High judgment task: (1) Across contour shapes, target *DAY* contours were judged **lower** in pitch than the highest reference contours, but higher than the lowest references. (2) As plateau length increases, participants' 'lower target' judgments increase earlier. Adopting the same statistical analysis reported above, the best model, based on all-subset selection, contains interaction effects between both reference level and group (β=-.498, z=-9.54, p<.001), and contour shape and group (β=-.343, z=-4.57, p<.001). Pairwise comparisons

via model re-leveling show that contour shape has a significant effect on BiL participants' responses (β=.411, z=7.56, p<.001), but not NE participants' responses (β=.068, z=1.31, p=.190).

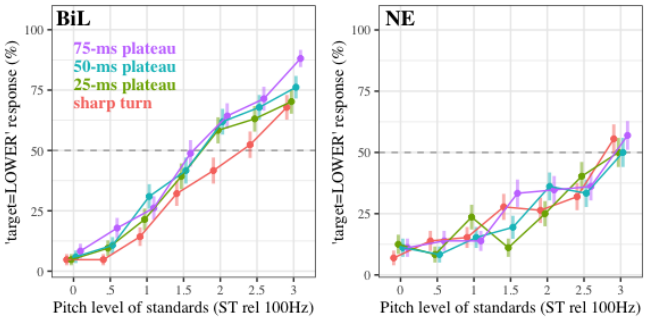


Figure 6: Percentage 'target is LOWER' judgments for the four contour shapes, as a function of the reference level against which they were compared.

4 DISCUSSION AND CONCLUSION

The lack of a "plateau effect" for NE participants in the Low judgment task could be related to their overall poor performance in judging the relative scaling of Low accents. ([15] implicate this same lack of acuity in the attenuation of the plateau effect that they report for Lows.) The fact that BiL listeners show both a plateau effect for Lows, and equivalent performance in discriminating Low and High F0 levels, is consistent with such an account.

It is not immediately obvious why NE listeners consistently rated Low target contours relatively higher in pitch than did their BiL counterparts. Note, however, that if listeners indeed evaluate pitch accent scaling through some form of F0-averaging over accented syllables, as suggested by [10]–[12] and others, then their responses are in fact in the expected direction: Mean F0 levels during accented *DAY* syllables in the low judgment task were as follows: 122.4 Hz (sharp turn), 120.3 Hz (25-ms plateau), 118.2 Hz (50-ms plateau), and 117 Hz (75-ms plateau), while the highest reference level was 119 Hz. NE listener responses effectively mirror this pattern. In this case, however, we must ask both why the BiL listeners failed to show this pattern, and why neither group showed its analogue (with peaks sounding lower than nearly all reference levels) in the High judgment task. See [3] for an explanation of a similar pattern in terms of weighting F0 samples by "recency".

To the question, however, of why BiL listeners performed both better on the discrimination of Low levels, and more symmetrically across the High and Low judgment tasks, we offer an explanation based on differences in the language experience of the two groups. As mentioned above, Cantonese is a lexical tone language that contrasts multiple pitch levels. Low tones realized at different F0 levels are used to signal different lexical meanings (e.g. low level T6 vs. low falling T4). Assessing the relative pitch of F0 targets in the lower portion of the pitch range is thus a frequent and necessary task in the perception of Cantonese.

English does make use of both High and Low targets in its intonation system (H* vs. L*, H- vs. L-, H% vs. L%). Implementation of Lows, however, is such that while High targets undergo a great deal of contextual variation (signaling, e.g., differences in focus type, paralinguistic emphasis, syntactic embedding, presence or absence of downstep, etc.), Low targets are reportedly less variable, most often hovering near the bottom of the speaker's pitch range [21]. Reports of how Low tonal targets across languages behave under differences in focus or emphasis are conflicting at best. (Early reports of lowering, e.g., [22] coexist with reports of no change, or even raising, in other sources–see [23, Sec. 10.4.1] for an overview.) Likewise, in English, there is no L*- analogue to the contrastive accentual downstepping patterns we find for High pitch accents. In fact, when bitonal accents such as L*+H participate in downstepping patterns, it is the peaks of the H, more saliently than the valleys of the lows, that encode the relation [24, p. 226].

Arguably, therefore, for NE speakers, the relative scaling of Low pitch targets is not linguistically meaningful, or at least has dramatically lower functional load than the relative scaling of Highs. This may cause NE listeners to attend to it less in speech, and to perform worse when judging it in experimental contexts. L1 Cantonese speakers, by contrast, must regularly assess the relative scaling of Low pitch targets, and therefore may both attend to it more strongly in speech, and judge it more accurately in experimental contexts. We might hypothesize then that the NE listeners' pattern of results in the Low task represents a non-linguistic processing mode for pitch, while the NE pattern in the High task, and the BiL pattern in both tasks, represents a linguistic processing mode. If this is true, however, then the plateau effect must be a phenomenon specific to linguistic processing of pitch, which may be contradicted by the non-speech results from [15]. Further research will be necessary to resolve these questions.

Jeon and Heinrich [15] suggest that the lower functional load English places on the discrimination of Low F0 targets may be a result of universally weaker domain-general auditory abilities in this area. If this is true, then the performance of our BiL listeners suggests that this deficit can be overcome through experience with one's native language. Alternatively, we might wonder whether in English the causality relation is better seen as reversed: Perhaps the lack of functional load on the discrimination of relative scaling among Lows is in fact the source, rather than a consequence, of listeners' weaker performance in perception.

5 REFERENCES

[1] J. P. Evans, "High is not just the opposite of Low," *J. Phon.*, vol. 51, pp. 1–5, 2015.

[2] J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux, "On the nature of perceptual differences between accentual peaks and plateaux," in *Understanding Prosody*, O. Niebuhr, Ed. Berlin: De Gruyter, 2012, pp. 93–118.

[3] M. D'Imperio, "The role of perception in defining tonal targets and their alignment," PhD dissertation, The Ohio State University, 2000.

[4] J. 't Hart, "F0 stylization in speech: straight lines versus parabolas," *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 3368–3370, 1991.

[5] R.-A. Knight, "The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux," *Lang. Speech*, vol. 51, no. Pt 3, pp. 223–244, 2008.

[6] R.-A. Knight and F. Nolan, "The effect of pitch span on intonational plateaux," *J. Int. Phon. Assoc.*, vol. 36, no. 1, pp. 21–38, 2006.

[7] O. Niebuhr, M. D'Imperio, B. G. Fivela, and F. Cangemi, "Are there 'shapers' and 'aligners'? Individual differences in signalling pitch accent category," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 120–123.

[8] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "Voiceless intervals and perceptual completion in F0 contours: Evidence from scaling perception in American English," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 108–111.

[9] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "Segmental Influences on the Perception of Pitch Accent Scaling in English," in *Proc. SP-2014*, Dublin, Ireland, 2014, pp. 1125–1129.

[10] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "The effect of global F0 contour shape on the perception of tonal timing contrasts in American English intonation," in *Proc. SP-2010*, Chicago, IL, USA, 2010, p. Paper 445.

[11] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology," *Lab. Phonol.*, vol. 3, no. 2, pp. 337–383, 2012.

[12] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity," *J. Phon.*, vol. 85, 101020, 2021.

[13] C. d'Alessandro, S. Rosset, and J. P. Rossi, "The pitch of short-duration fundamental frequency glissandos," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2339–2348, Oct. 1998.

[14] P. Mertens, "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proc. SP-2004*, Nara, Japan, 2004, pp. 549–552.

[15] H.-S. Jeon and A. Heinrich, "Perceptual asymmetry between pitch peaks and valleys," *Speech Commun.*, vol. 140, pp. 109–127, 2022.

[16] J. Gandour, "Perceptual dimensions of tone: Evidence from Cantonese," *J. Chin. Linguist.*, vol. 9, no. 1, pp. 20–36, 1981.

[17] C. Gussenhoven, "On the intonation of tonal varieties of Englishes," in *The Oxford handbook of world Englishes*, M. Filppula, J. Klemola, and D. Sharma, Eds. New York, NY: Oxford University Press, 2017, pp. 569–598.

[18] K. K. Luke, "Phonological re-interpretation: The assignment of Cantonese tones to English words," in *The 9th International Conference on Chinese Linguistics*, 2000.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 6.1.10)," 2020. [Online]. Available: http://www.praat.org/

[20] G. Ward and J. Hirschberg, "Implicating uncertainty: The pragmatics of fall-rise intonation," *Language*, vol. 61, no. 4, pp. 747–776, 1985.

[21] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, M. Aronoff and R. Oehrle, Eds. Cambridge, MA: MIT Press, 1984, pp. 157–223.

[22] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," PhD dissertation, Massachusetts Institute of Technology, 1980.

[23] J. Barnes, H. Mixdorff, and O. Niebuhr, "Phonetic variation in tone and intonation systems," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 125–149.

[24] M. Grice, "Leading tones and downstep in English," *Phonology*, vol. 12, no. 2, pp. 183–233, 1995.

CONTOUR SHAPE AND PROSODIC FOCUS MARKING: THE CASE OF HONG KONG CANTONESE

Chris K. C. Lee & Jonathan Barnes

Linguistics Department, Boston University
chriskcl@bu.edu, jabarnes@bu.edu

ABSTRACT

Hong Kong Cantonese has been reported to show inconsistent, if not absent, pitch-related modulation in response to changes in information structure. However, previous studies focused primarily on F0 target scaling, neglecting other important factors, such as the overall shape of the F0 contour. This study focuses on the High level (T55) and High rising tones (T25), realized between a preceding Low level T22 and a following Low falling T21, and investigates whether Cantonese speakers modulate the contour shapes of the High tones in implementing information-structural contrasts. Exploiting *Tonal Center of Gravity in the time dimension* (TCoG), we confirm that Cantonese speakers hyperarticulate each tone's characteristic shape under focus. When flanked by two Low targets, on-focus T55 have a significantly domier rise, and on-focus T25 a significantly scoopier rise, than their respective broad-focus counterparts, enhancing the contrast between the High tones in terms of TCoG.

Keywords: Prosodic focus marking; Tonal Center of Gravity; Hyperarticulation; Cantonese

1. INTRODUCTION

Information structure is expressed in various ways in the world's languages [1], [2]. Phonetically, constituents with narrow focus are often realized with a wider pitch range, longer duration, and/or higher intensity than those in broad focus, and post-focus constituents with compressed pitch range and intensity (e.g. English [3]–[5], German [6], Mandarin [7], Japanese [8]). Nevertheless, how and whether focus is marked prosodically remains language-specific in nature. In Cantonese[1], longer constituent duration has been argued to be the only consistent cue to narrow focus [9]–[11]. Pitch range of focused constituents is reported to be modulated only inconsistently: while [12], [13] present evidence for expanded pitch range in contrastively focused constituents, [11] reported that in Cantonese on-focus expansion of pitch range is only observed in the two rising tones, while pitch level raising is observed inconsistently in the non-High level tones. [11], [12] also report that post-focus compression (PFC), whether of pitch, intensity, or duration, is non-existent in Cantonese.

The studies on Cantonese cited above, however, all rely on F0 measures such as mean, range, or extremum level within a region of interest. While these measures characterize at least in part the *scaling* of a pitch event, using *only* these measures neglects two other important prosodic properties known to interact with target scaling perceptually: *timing* and *contour shape* of a pitch event [14]–[22]. Concerning the timing of pitch events, [7, p. 86] reports that in Mandarin a rising tone followed by a Low target has *later* F0 peaks when under focus than those not under focus. Timing differences of this kind, however, are also often accompanied by a contour shape difference [16]. In the context of a high level tone, for example, aligning an F0 peak earlier would also mean staying at the max F0 for longer within the tone-bearing syllable, creating a longer High plateau.

Since focus-induced modulation of F0 contour shape of Cantonese tones is hitherto unexplored, this paper aims to investigate

RQ1: whether Cantonese speakers vary the contour shape of High lexical tones in realizing different information structures, and

RQ2: whether tone type affects the direction and magnitude of contour shape modulations induced by focus.

Instead of tracking the timing or scaling of any single "target" point in the F0 contour, this study utilizes the *Tonal Center of Gravity in the time dimension* (**TCoG**), a measure that estimates when a pitch event happens *perceptually*, to characterize the perceptual consequences of a change in contour shape [15]–[17]. TCoG uses weighted F0 averages to generalize about the overall disposition of a pitch event in time. For a High tonal target associated with a given syllable, earlier F0 peaks, higher pre-High F0 valleys, longer High plateaux, and/or domier rises will shift the bulk of the High F0 region, and thus TCoG, leftward. Later F0 peaks, lower pre-High F0 valleys, and/or scoopier rises, by contrast, shift TCoG rightward. Fig. 1 schematizes these generalizations. (See [16] for discussion.)

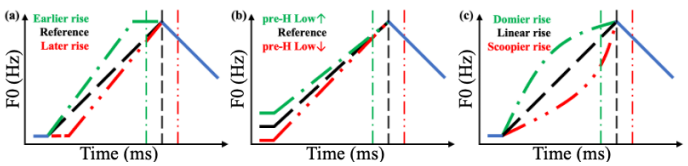


Figure 1: Schematics for the effect of (a) pitch event timing, (b) pre-High F0 valley scaling, and (c) contour shape curvature on TCoG.

Since focus-induced variation in High target timing was reported when it precedes a Low target in a tone language like Mandarin [7], and the applicability of TCoG to characterize the contour shape of a L-H-L tonal sequence is well attested, this paper focuses on Cantonese High tones (T55, T25) surrounded by Low targets (T22, T21).

2. METHOD

In this study, participants read a set of sentences with a natural and context-appropriate tone of voice as a response to some contextualizing questions pre-recorded by a male native Cantonese speaker in his late 20s born and raised in Hong Kong.

2.1. Stimuli

The elicitation materials of the present study are created in such a way that a High pitch target is surrounded by Low targets. The target NPs this study adopts are four trisyllabic personal names with one of the following tone sequences (Fig. 2 shows the pitch tracks of the two tone sequences):

T22-55-21 (*T55 set*): 鄭依明 [tsɛŋ^{22} ji^{55} mɪŋ^{21}],
段威龍 [tyn^{22} wɐj^{55} lʊŋ^{21}]
T22-25-21 (*T25 set*): 范綺玲 [fan^{22} ji^{25} lɪŋ^{21}],
鄧委文 [tɐŋ^{22} wɐj^{25} mɐn^{21}]

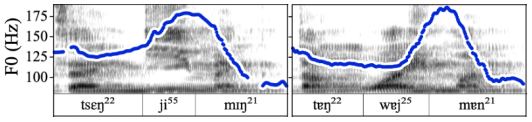


Figure 2: Pitch tracks of a T22-55-21 (left) and a T22-25-21 (right) target NP produced in broad focus condition by a male participant in the current study.

The target NPs are embedded in the carrier sentence NP_1約NP_2過大海 [*NP*$_1$ jœk̄^{3} *NP*$_2$ kʷɔ^{33} taj^{22} hɔj^{25}] (*NP*$_1$ *invited NP*$_2$ *to visit Macau*). Each target NP appears in both NP_1 and NP_2 positions, but in each sentence $NP_1 \neq NP_2$. This generates $4\times(4-1) = 12$ variations of the carrier sentence.

As for the contextualizing questions, they are designed to elicit different contexts such that NP_2 is placed in broad-focus (1), post-focus (2), narrow-focus (3), pre-focus (4) or contrastive-focus conditions (5).

(1) What's new recently?
(2) Who invited NP_2 to visit Macau?
(3) Whom did NP_1 invite to visit Macau?
(4) What did NP_1 invite NP_2 to do?
(5) Did NP_1 invite [siw^{22} jɪŋ^{55} mɪŋ^{21}] to visit Macau?

2.2. Acoustic analysis

For each sentence, only NP_2 was analyzed. Each participant produced 100 trials (4 target NPs × 5 focus conditions × 5 repetitions). Boundaries of the target NPs and syllables therein were annotated in Praat [23]. Acoustic measures, including F0, NP duration (onset of the first syllable excluded), and (root-mean-squared) energy were estimated with VoiceSauce [24] with REAPER [25] integrated for F0 estimation. Default settings were adopted. All measures are *z*-transformed within speaker.

TCoG was used to identify the perceptual reference location in the time dimension for the High target associated with the target NP-medial syllable, in order to investigate the influence of the rising contour shape on the perception of tonal target timing. Its calculation was based on the basic formula proposed in [16], reproduced in (6), where $F0_i$ is the F0 at time t_i relative to the min. F0 of the region of interest.

$$(6)\quad \mathrm{TCoG} = \frac{\Sigma_i F0_i \times t_i}{\Sigma_i F0_i}$$

The start and end points of the region of interest corresponded to the F0 valley before the F0 peak of a target NP and the F0 peak respectively. To compare TCoG values across focus conditions, we took the midpoint of the target NP-medial syllable as a segmental anchor.

3. RESULTS

Data from 13 native Cantonese speakers (six male and seven female university degree holders, aged 24–30 [*M*=27.92], born, raised, and educated in Hong Kong) were analyzed. Among them one speaker merged T25 with the low rising tone completely, so all of their T25 tokens (*N*=50) were discarded. 22 more T25 tokens and three T55 tokens of the other speakers were discarded due to segmental or tonal errors or disfluency. Altogether, 647 tokens in the T55 set and 578 tokens in the T25 set were analyzed.

3.1. The "conventional" acoustic measures

We first consider the more well-studied acoustic correlates of focus marking in Cantonese (NP duration, mean energy) to ensure our participants express information structure in the expected direction. To assess the effect of focus on pitch scaling, we also consider the target NPs' mean F0 and max F0 of the window for TCoG calculation. Effectively, these two measures quantify the pitch scaling of the High pitch target. Fig. 3 shows the group means and standard errors of these values separated by focus condition. Values of broad-focus NPs are represented by red vertical lines. To statistically assess our observations reported below, separate linear mixed effects models (LMEMs) were fitted, using the four measures as dependent variables. We considered FOCUS, TONESEQ (T55 set *vs.* T25 set), and their interaction as potential fixed effects, and included a random intercept of speaker and a random intercept of target NP.[2] Based

on all-subset selection, the best models for all four acoustic measures select FOCUS as the only significant main effect.

Let's start with on-focus marking. Compared to broad-focus NPs, narrow- and contrastive-focus NPs (light and dark green respectively) are realized with significantly longer duration (Narrow-focus [NF]: β=1.20, t=18.79, p<.001; Contrastive-focus [CF]: β=1.21, t=19.01, p<.001) and higher mean energy (CF: β=.225, t=5.48, p<.001; NF: β=.230, t=5.58, p<.001). For pitch scaling measures, our participants realize NPs in different focus conditions with statistically comparable max F0 (ps>.60). As for mean F0, although Fig. 3 suggests that mean F0 is lowered slightly when the NP is under focus, such a difference is statistically not significant (ps>.055). Our participants are hence very comparable to those in previous studies in terms of on-focus marking.

The more interesting finding here concerns PFC where, contrary to previous findings, participants actively compress NP duration (β=-.427, t=-6.73, p<.001), mean energy (β=-.196, t=-4.78, p<.001), and the pitch scaling of the High target (max F0: β=-.524, t=-12.14, p<.001; mean F0: β=-.325, t=-10.22, p<.001) of post-focus NPs relative to broad-focus NPs. In fact, while PFC in general is smaller in magnitude than on-focus expansion, PFC is consistently observed in all the acoustic measures that previous studies have considered.

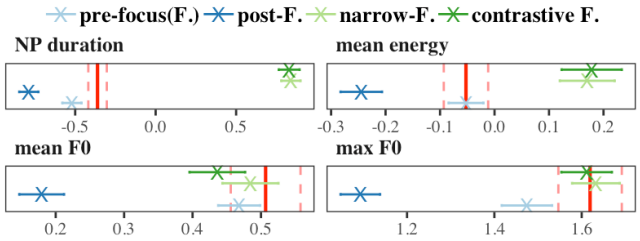


Figure 3: Means of the four "conventional" acoustic correlates for prosodic focus marking in Cantonese. Red solid lines indicate parameter values of broad-focus NP. Error bars and dashed lines indicate SE.

3.2. Contour shape

Apart from the somewhat surprising appearance of PFC in our data, participants otherwise produced focus-related modulations of acoustic properties in expected directions. We now consider the contour shape. Fig. 4 shows time-normalized F0 tracks from the second half of the NP-initial syllable of the target sequence to the first 40% of the NP-final syllable averaged across speakers, separated by tone set. Once again corroborating our findings with respect to PFC of pitch scaling, post-focus NPs (dark-blue dashed line) have a much lower pitch level than NPs in other focus conditions. However, its contour shape is quite comparable to broad-focus NPs (red solid line).

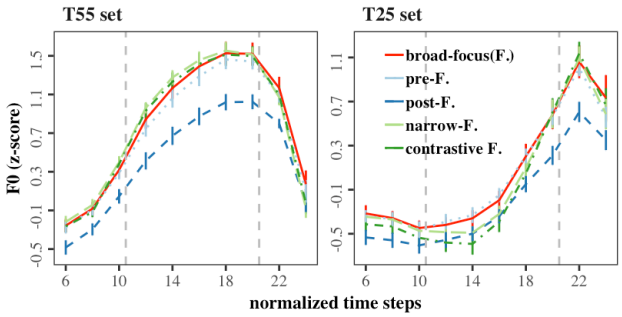


Figure 4: Time-normalized F0 contour of NPs by focus condition and tone set averaged across speakers. Each syllable in NPs is represented by 10 equidistant points.

On the other hand, the contour shapes of narrow- and contrastive-focus NPs (light green dashed and dark green dot-dashed lines respectively) are quite different from the broad-focus variants, and the direction of focus-induced contour shape modulation is tone-sensitive. For the T55 set, on-focus NPs show a steeper rise in F0 from t_n=8 to t_n=14 than broad-focus NPs, and then reach a similar F0 level as broad-focus NPs at t_n=18. This means that on-focus T55 NPs have a *domier* rise than the broad-focus variants. For the T25 set, broad-focus and on-focus NPs have comparable F0 at t_n=10, but the F0 of broad-focus NPs start rising at this same time point, while the F0 rise of on-focus NPs is delayed to t_n=14, where F0 also reaches a lower minimum than broad-focus NPs, and then rise to the same max F0 as broad-focus NPs at t_n=22. This means that on-focus T25 NPs set have a *scoopier* rise than the broad-focus variant. Situating these observations in TCoG terms, we expect that on-focus NPs will show *earlier* TCoG for T55 set, but *later* TCoG for T25 set, when compared to the respective broad-focus variants, effectively enhancing the timing difference between the two High targets. We now turn to verify these predictions.

Fig. 5 shows the TCoG value of NPs relative to the midpoint of the word-medial syllable in different focus conditions and tone sets, averaged across speakers. Values for broad-focus NPs are represented by the red vertical lines. The data is fitted to an LMEM following the same procedure reported in the previous section. The best model contains the interaction effect between FOCUS and TONESEQ. The expected changes in TCoG across focus conditions and tone sets are borne out. First, T25 NPs have significantly *later* TCoG than T55 NPs (β=122.0, t=6.89, p=.020), holding focus condition constant. This suggests that TCoG is useful in characterizing and differentiating T55 and T25 in Cantonese regardless of focus context. Concerning the T55 set, narrow- and contrastive-focus NPs have significantly earlier TCoG than the broad-focus NPs (CF: β=-9.42, t=-3.81, p=.0001; NF: β=-7.14, t=-2.89, p=.004).

Post-focus NPs, on the other hand, have comparable TCoG to broad-focus NPs. As for the T25 set, narrow- and contrastive-focus NPs have significantly later TCoG than broad-focus NPs (CF: β=21.24, t=8.11, p<.001; NF: β=15.81, t=6.06, p<.001). Interestingly, post-focus NPs in this set also have significantly earlier TCoG than broad-focus NPs (β=-6.64, t=-2.55, p=.0109). Finally, comparing the magnitude of focus-induced shift of TCoG, the T25 set exhibits a greater shift in TCoG than the T55 set (CF: 21.24 [T25] *vs.* 9.42 [T55]; NF: 15.81 [T25] *vs.* 7.14 [T55]).

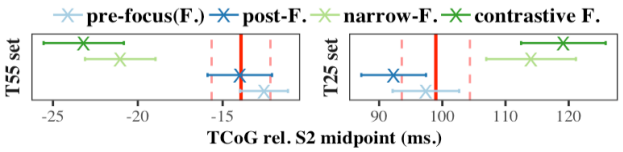


Figure 5: Mean TCoG relative to the midpoint of the NP-medial syllable. Red solid lines indicate group mean and SE of broad-focus NPs. Error bars and dashed lines indicate SE.

4. DISCUSSION

Previous studies argued that in Cantonese, focus is most consistently signaled by durational lengthening and intensity raising of the on-focus constituent. Pitch scaling is at best inconsistently modulated and mostly restricted to rising tones. PFC was also argued to be absent in Cantonese. The current study has gathered production data to demonstrate that our prior knowledge of Cantonese prosodic focus marking is at best incomplete. Based on data from 13 speakers, we find significant durational shortening and lowering of mean energy, max F0 & mean F0 of post-focus constituents relative to broad-focus ones. Our current dataset, however, is not best suited for explaining the discrepancy between our findings and previous ones. Whether it is due to task effects (e.g., adopting elicitation materials that promote *vs.* inhibit tonal coarticulatory variations) or a genuine change in prosodic focus marking strategies in Cantonese requires further investigations.

Turning to our specific research questions, this study presents evidence for the first time that Cantonese speakers modulate the contour shape of lexical tones as a response to information-structural differences. The direction and magnitude of focus-induced modulation of contour shape depend on tone type. In particular, relative to broad-focus target NPs, on-focus T55 shows a leftward shift of TCoG, but on-focus T25 shows a rightward shift. The observed tone-specific direction of change is in line with extant evidence presented in the segmental phonology literature that (sentential) stress leads to localized hyperarticulation (see e.g. [26], [27]). When the lexical tone is under focus, the High level tone (surrounded by two Low tones) is realized with a domier rise, hence more level-tone-like, while a High rising tone is realized with a later but faster (scoopier) rise, hence more rising-tone-like.

However, while hyperarticulated T55 flanked by Low targets adopting a dome-shaped rise seems intuitive, why should T25 be more scooped under focus? Recall that we observe that on-focus T25 reaches a lower F0 before they rise to the maximum F0 later on than T25 in other focus conditions (see also bottom left pane of Fig. 6 below). One potential motivation for the scooping of the T25 rise under focus could then be found in the assumption that this contour tone is best represented as a sequence of two levels, Low and High (see [28], cf. [29]). If this were true, then the scooped rise could serve to enhance simultaneously the later timing for the TCoG of the High tone, and lower scaling for the Low. Note however that the effect of focus on Low targets in our data is not consistent: The Low targets for the NP-medial T25 and NP-final T21 are lowered consistently, but that of the NP-initial T22 is not (see top panel of Fig. 6 below). This inconsistency may itself be serving to enhance contrast among the tones in the system, insofar as T21 is associated with lower F0 targets in general than T22. How Low targets are modulated under focus requires further investigation in the future.

The observation that at least some Low targets in our data (the Low of T25 and T21) are lowered under focus also casts doubts on the argument that pitch scaling is not modulated for narrow focus in Cantonese. To the extent that this lowering effect is true, pitch scaling is then actively modulated for focus purposes in Cantonese. Also related to pitch scaling is the well-documented correlation between contour shape and pitch scaling perception such that plateau-shaped peaks (e.g. green line of Fig. 1a) tend to sound higher than a sharp peak (e.g. black line of Fig. 1a) even if they have identical max. F0 at least to English listeners [30]–[32]. Assuming such findings are generalizable to Cantonese, on-focus T55 should then sound higher, even though max. F0 is not raised relative to broad-focus NPs. Whether the shape-scaling correlations are also true among Cantonese listeners, however, awaits experimental verification.

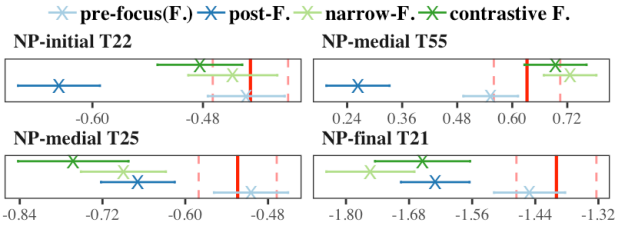


Figure 6: Mean Minimum F0 (z-score) of different tones in the target NPs. Red solid lines indicate values of respective tones in broad-focus condition. Error bars and dashed lines indicate SE.

5. REFERENCES

[1] D. Büring, "Towards a typology of focus realization," in *Information structure: Theoretical, typological, and experimental perspectives*, M. Zimmermann and C. Féry, Eds. Oxford: Oxford University Press, 2010, pp. 177–205.

[2] F. Kügler and S. Calhoun, "Prosodic encoding of information structure: A typological perspective," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 454–467.

[3] W. E. Cooper, S. J. Eady, and P. R. Mueller, "Acoustical aspects of contrastive stress in question-answer contexts," *J. Acoust. Soc. Am.*, vol. 77, no. 6, pp. 2142–2156, 1985.

[4] S. J. Eady and W. E. Cooper, "Speech intonation and focus location in matched statements and questions," *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 402–415, 1986.

[5] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *J. Phon.*, vol. 33, no. 2, pp. 159–197, 2005.

[6] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *J. Phon.*, vol. 36, no. 4, pp. 680–703, 2008.

[7] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *J. Phon.*, vol. 27, no. 1, pp. 55–105, 1999.

[8] S. Ishihara, "Japanese focus prosody revisited: Freeing focus from prosodic phrasing," *Lingua*, vol. 121, no. 13, pp. 1870–1889, 2011.

[9] R. S. Bauer, K.-H. Cheung, P.-M. Cheung, and L. Ng, "Acoustic correlates of focus-stress in Hong Kong Cantonese," in *Papers from the Eleventh Annual Meeting of the Southeast Asian Linguistics Society 2001*, S. Burusphat, Ed. Tempe, AZ: Arizona State University, 2004, pp. 29–49.

[10] W. L. Wu and L. Chung, "Post-focus compression in English-Cantonese bilingual speakers," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 148–151.

[11] W. L. Wu and Y. Xu, "Prosodic Focus in Hong Kong Cantonese without Post-focus Compression," in *Proc. SP-2010*, Chicago, USA, 2010, Paper 040.

[12] W. Gu and T. Lee, "Effects of tone and emphatic focus on F0 contours of Cantonese speech: A comparison with standard Chinese," *Chinese Journal of Phonetics*, vol. 2, pp. 133–147, 2009.

[13] V. C. H. Man, "Focus effects on Cantonese tones: An acoustic study," in *Proc. SP-2002*, Aix-en-Provence, France, 2002, pp. 467–470.

[14] A. Arvaniti, D. R. Ladd, and I. Mennen, "Stability of tonal alignment: the case of Greek prenuclear accents," *J. Phon.*, vol. 26, no. 1, pp. 3–25, 1998.

[15] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "The effect of global F0 contour shape on the perception of tonal timing contrasts in American English intonation," in *Proc. SP-2010*, Chicago, IL, USA, 2010, p. Paper 445.

[16] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology," *Lab. Phonol.*, vol. 3, no. 2, pp. 337–383, 2012.

[17] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity," *J. Phon.*, vol. 85, 101020, 2021.

[18] M. D'Imperio, "The role of perception in defining tonal targets and their alignment," PhD dissertation, The Ohio State University, 2000.

[19] D. R. Ladd, D. Faulkner, H. Faulkner, and A. Schepman, "Constant 'segmental anchoring' of F0 movements under changes in speech rate," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1543–1554, 1999.

[20] D. R. Ladd, *Intonational Phonology*, 2nd ed. Oxford: Oxford University Press, 2008.

[21] O. Niebuhr, M. D'Imperio, B. G. Fivela, and F. Cangemi, "Are there 'shapers' and 'aligners'? Individual differences in signalling pitch accent category," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 120–123.

[22] J. 't Hart, "F0 stylization in speech: straight lines versus parabolas," *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 3368–3370, 1991.

[23] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2020. [Online]. Available: http://www.praat.org/

[24] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 1846–1849.

[25] D. Talkin, "REAPER: Robust and epoch and pitch estimator." 2014. [Online]. Available: https://github.com/google/REAPER

[26] K. J. de Jong, "The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 491–504, 1995.

[27] K. de Jong, M. E. Beckman, and J. Edwards, "The interplay between prosodic structure and coarticulation," *Lang. Speech*, vol. 36, no. 2–3, pp. 197–212, 1993.

[28] J. L. Lee, "The Representation of Contour Tones in Cantonese," in *Proceedings of the Annual Meeting of the Berkeley Linguistics Society 38*, 2012, pp. 272–286.

[29] M. Barrie, "Contour tones and contrast in Chinese languages," *J. East Asian Ling.*, vol. 16, no. 4, pp. 337–362, 2007.

[30] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "Voiceless intervals and perceptual completion in F0 contours: Evidence from scaling perception in American English," in *Proc. ICPhS XVII*, Hong Kong, 2011, pp. 108–111.

[31] J. Barnes, A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel, "Segmental influences on the perception of pitch accent scaling in English," in *Proc. SP-2014*, Dublin, Ireland, May 2014, pp. 1125–1129.

[32] R.-A. Knight, "The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux," *Lang. Speech*, vol. 51, no. Pt 3, pp. 223–244, 2008.

1. Cantonese has six lexical tones: high level (T55), mid level (T33), low level (T22), high rising (T25), low rising (T23), low falling (T21). The short-hand notations are based on Chao's tone numerals, with *1* referring to a speaker's lowest pitch and *5* highest.
2. More complex random effect structures were not considered because such moves resulted in model singularity.

INTERACTIONS OF LEXICAL STRESS, VOWEL LENGTH AND PHARYNGEALISATION IN PALESTINIAN ARABIC

Niamh E. Kelly

Newcastle University
niamh.kelly@newcastle.ac.uk

ABSTRACT

Research on prosody in varieties of Arabic has expanded in recent years, but no work has examined lexical stress in Palestinian Arabic. The current study examined the acoustic correlates of lexical stress in disyllabic and trisyllabic words with penultimate stress and either a phonologically long or short vowel (/a/) in the stressed syllable. The effects of being adjacent to a pharyngealised consonant were also examined. Results indicated that stressed vowels had higher mean f0 and intensity than unstressed vowels. Stressed long vowels had greater duration than unstressed long vowels, but this pattern did not extend to short vowels. Pharyngealisation was found to interact in complex ways with stress and length in terms of its effect on vowel quality. This research provides a phonetic description of lexical stress in Palestinian Arabic, and examines the interaction among stress, phonemic length and pharyngealisation.

Keywords: lexical stress, prosody, acoustics, pharyngealisation, Palestinian Arabic

1. INTRODUCTION

Lexical stress is the prominence of a syllable (or syllables) within a word, and cues to stress include a variety of correlates (see section 1.1), with some cross-linguistic variation [18, 2, 6, 20].

A quantitative phonetic examination of the acoustic correlates of stress in Palestinian Arabic has not yet been conducted. The goal of the current study is to address this gap, thereby adding to typological literature on lexical stress as well as phonetic descriptions of suprasegmental aspects of Arabic varieties. The current study investigates a number of measures of f0, intensity, duration and vowel quality, as well as examining the interaction of stress and length, and the effects of adjacent pharyngealised sounds.

1.1. Cross-linguistic correlates of lexical stress

Research on lexical stress across many languages has found some common correlates of stressed vowels, such as increased duration, higher intensity and more peripheral quality [19, 7]. Higher f0 may also be a correlate of stress, although controlled experiments on English suggest that higher f0 is a correlate of focus (sentence accent) rather than lexical stress [33, 6, 20]. As such, English is described as using primarily duration and vowel quality to express stress [6, 33], while Spanish has been described as using mainly duration and intensity [30].

1.2. Stress and phonemic length in varieties of Arabic

Palestinian Arabic is a variety of Levantine Arabic, a branch of Arabic spoken in Lebanon, Syria, Jordan and Palestine. Arabic is known to be a stress-accent language, similar to English and Dutch [29, 7, 33, 8, 16, 25, 32]. This means that a syllable can be made prominent regardless of whether it bears a pitch accent. That is, f0 alone is not the primary acoustic correlate of prominence. Stress accent languages differ in how they weight the various cues of f0, duration and intensity [15, 32]. Lexical stress in Arabic is described as falling on: first, a word-final super-heavy syllable, in the absence of that, then on a heavy penult, and finally, on the antepenult, whether heavy or light [28, 1, 12, 34]. Long vowels without primary stress have sometimes been described as having secondary stress [29].

Recent research on prosody in Arabic includes work on lexical stress and focus in Jordanian Arabic [16, 17], on the intonation of focus in Egyptian Arabic [24, 23] and one variety of Lebanese Arabic [12, 14], and work on intonation in Emirati Arabic [9]. The correlates of lexical stress in Arabic have been examined in a small number of varieties: in Egyptian, greater duration, mean f0 and intensity were found to correlate with lexical stress [14], and in Jordanian, greater duration and higher F1 were found in stressed vowels [16]. In Moroccan, no acoustic correlates of lexical stress were found [11].

In the Tripoli variety of Lebanese Arabic, spoken in the north of the country, [13] examined lexically stressed syllables with the (short) vowels [a i u] when unaccented, accented and nuclear accented, and found that f0, intensity, duration and quality differed in each of these conditions.

Arabic has a quantity distinction in both vowels and consonants. Long vowels in Jordanian Arabic were found to be about 70 milliseconds (120%) longer than their short counterparts [17].

Syllable structure can interact with duration because Palestinian Arabic has compensatory shortening in closed syllables [36, 22]. The position of a syllable in a word also affects duration, due to final lengthening [5]. As such, these factors will be taken into account in the analysis of duration.

Arabic also has "emphatic" or pharyngealised consonants, which have the acoustic effect of a higher F1 and lower F2 in adjacent vowels [35, 26]. As such, being adjacent to a pharyngealised consonant may have an effect on the same correlates that are used to express stress. In this study, this is examined as another factor, along with length and stress.

In the current study, I investigate a number of acoustic measures relating to f0, intensity, duration and vowel quality, in order to examine how stress, phonemic length and pharyngealisation interact in this variety.

2. METHODS

Data was collected by [22], using a method developed by [21]. Stimuli consisted of the target word and two colloquial frame sentences presented in the Arabic script.

2.1. Participants

Recordings were analyzed of seven female and seven male speakers who were native speakers of Palestinian Arabic, students at Tel Aviv University or University of Haifa aged 21-27, proficient also in Hebrew and English [22].

2.2. Materials

Stimuli examined in the current study were a subset of those collected by [22], specifically those that had the vowel /a/ in the target syllables. All words had penultimate stress. In disyllabic words, some had a long vowel in the stressed syllable, and some had a short vowel in the stressed syllable. The vowel length contrast in trisyllabic words was on the initial syllable. The target words was followed by another word, either monosyllabic or disyllabic. The target words are shown in Table 1.

[mɪnʔu:l X ħe:k] We say X like this.
[ʕam bɪʔu:l X ɪs:a] I am saying X now.

Disyllabic		Trisyllabic
['fadˤəl]	['faːdˤel]	[sa'baqna]
['fasˤəl]	['faːsˤel]	[qa'tˤaʕna]
['ʒabha]	['ʒaːbha]	[da'faʕna]
['ʒablak]	['baːtˤel]	[saː'baqna]
['sˤatˤəl]	['ʔasˤəl]	[qaː'tˤaʕna]
['ħaːsˤel]		[daː'faʕna]

Table 1: Target words

2.3. Measurements and analysis

A variety of acoustic correlates of stress were examined in the current study (in Praat [10]), all on the vowel /a/: mean f0 (semitones (re 1 Hz)), mean intensity (dB), duration (msec), and F1 & F2 (Hz) at the midpoint. Recordings had been labelled previously for segment boundaries as described in [22]. Praat scripts were run to collect measurements.

Linear mixed effects regression tests were run in R [31] using the *lmerTest* [27] package with the independent variables stress (stressed or unstressed), vowel length (long or short) and pharyngealisation (yes or no) and the dependent variables each of the above measures. Token and speaker were included as random intercepts. For mean f0, male and female speakers were examined separately. For the duration measure, as noted above, syllable structure, syllable position and number of syllables could also affect this, but the data was not balanced for these, so they were explored as random factors. For each measure, model comparison was conducted using the *anova* function in R, building up models term by term to find the best one to explain the data. The reference levels were stressed, long and non-pharyngealised.

2.4. Hypotheses

Based on previous research, it is expected that stressed vowels will have longer duration, higher intensity, more peripheral quality, and possibly higher f0 than unstressed syllables. Long vowels are expected to have longer duration than short vowels. Pharyngealisation is expected to be associated with a higher F1 and lower F2 in the adjacent vowel.

3. RESULTS

The total number of tokens was 1715.

For intensity, the best model was one with stress as a fixed effect, where stressed vowels had higher intensity than unstressed vowels (Fig. 1, left panel; Coef.=-1.75, p<0.001) (R code: lmer(Mean.Intensity ~ Stress + (1|Spk) + (1|Token))).

For duration, the best model was one with the fixed factors stress and vowel length with an interaction, with syllable position and syllable structure as added random factors (R code: lmer(Duration ~ Stress*VLength + (1|Spk) + (1|Token) + (1|Syl.Pos) + (1|Syl.Str))). Pairwise tests were conducted using the *emmeans* package in order to determine which factors differed significantly from one another. These results revealed that long vowels were significantly longer than short vowels when both stressed (Coef.=48, p<0.001) and unstressed (Coef.=22, p<0.001). Stressed vowels were longer than unstressed vowels when both were phonemically long (Coef.=38, p<0.001), but there was no significant effect of stress when they were short (Coef.=12, p=0.167) (Fig. 2). Mean duration for unstressed long /a/ was 70 msec, unstressed short /a/ 54 msec and stressed long /a/ 108 msec and stressed short /a/, 59 msec.

For mean f0, the best model for the male speakers was one with stress and pharyngealisation (R code: lmer(MeanF0 ~ Stress + Pharyn + (1|Spk) + (1|Token), data=PAm)), where mean f0 was higher when a vowel was stressed (Coef.=-1.4, p<0.001) or pharyngealised (Coef.=0.6, p<0.01). For the female speakers, the best model included only stress, not pharyngealisation (R code: lmer(MeanF0 ~ Stress + (1|Spk) + (1|Token), data=PAf)). Again, mean f0 was higher in stressed than unstressed vowels for these speakers (Coef.=-1.7, p<0.001) (Fig. 1, right panel).

For F1, the best model was one with stress, vowel length and pharyngealisation and interactions among all three (R code: lmer(F1 ~ Stress*Vlength*Pharyn + (1|Spk) + (1|Token))). The pairwise results showed a significant effect of pharyngealisation only for short vowels (Figure 3), whereby pharyngealised stressed vowels had a higher F1 than non-pharyngealised stressed vowels (Coef.=-152, p<0.001), while pharyngealised unstressed vowels had a lower F1 than non-pharyngealised unstressed vowels (Coef.=102, p<0.001). There was a significant effect of stress also only for short vowels, whereby stressed pharyngealised vowels had a higher F1 than unstressed pharyngealised vowels (Coef.=210, p<0.001) while stressed non-pharyngealised vowels had a lower F1 than unstressed non-pharyngealised vowels (Coef.=-44, p<0.001).

For F2, the best model was one with stress and pharyngealisation and an interaction (R code: lmer(F2 ~ Stress*Pharyn + (1|Spk) + (1|Token))). F2 was lower in pharyngealised vowels than non-pharyngealised vowels (stressed: Coef.=175, p<0.001; unstressed: Coef.=275, p<0.001). F2 was lower in stressed vowels than unstressed vowels when non-pharyngealised (Coef.=-123, p<0.001), but there was no significant effect of stress in pharyngealised vowels (Coef.=-23, p=0.582) (Fig. 4).

Table 2 summarises which factors had a significant effect on each measure.

Measure	Stress	Length	Pharyn.
Intensity	*		
Duration	*	*	
Mean f0	*		M
F1	*		*
F2	*		*

Table 2: Significant effects for each measure (* = significant effect; M = only male speakers)

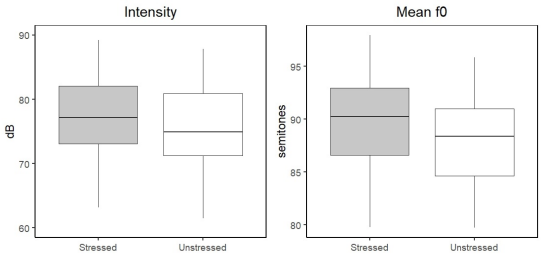


Figure 1: Boxplots showing intensity and mean f0.

4. DISCUSSION

The results found here show a complicated interaction among stress, length and pharyngealisation for the data examined. Both stress and length cause longer duration, however, stress only significantly increases duration in phonemically long vowels. Fig. 2 shows that an unstressed long vowel is similar in duration to a stressed or an unstressed short vowel. This leads to the question of whether this can cause confusion for listeners, and to what extent they use contextual cues to disambiguate these categories. Overall, the

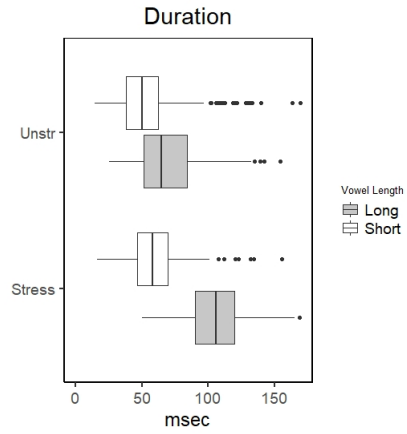


Figure 2: Boxplot showing duration.

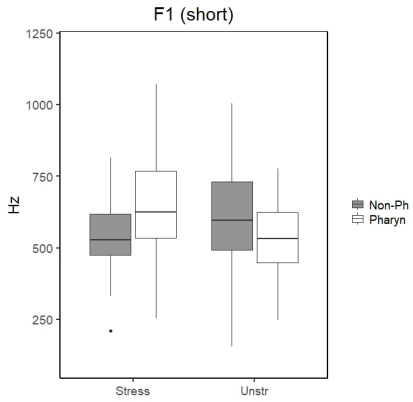


Figure 3: Boxplot showing F1 (short vowels only).

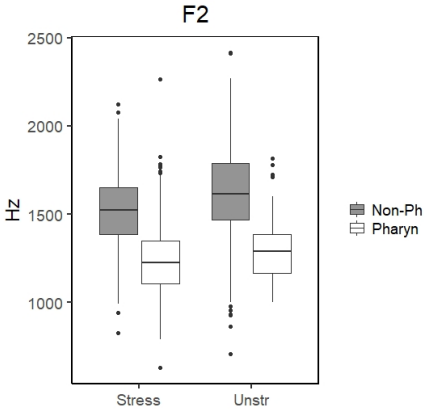


Figure 4: Boxplot showing F2.

correlates of stress for Palestinian Arabic are similar to what has been reported for Egyptian [14].

Pharyngealisation was shown to interact in complex ways with length and stress. First, there was a significant effect of pharyngealisation only on phonemically short vowels. Second, for stressed vowels, pharyngealisation was associated with a higher F1 (lower vowel), while for unstressed vowels, pharyngealisation was associated with a lower F1 (higher vowel). Examining other vowels may shed light on these interactions.

Gender has been reported to interact with pharyngealisation in Arabic, with differing results: in Jordanian, [3] found that male speakers produced stronger cues while [4] found that female speakers produce stronger cues. In this study, f0 was affected by pharyngealisation only among male speakers.

While in some varieties of Arabic, short vowels have been described as being more centralised, in the current study, F2 was not significantly affected by phonemic length. F1 interacted with length in terms of the effects of stress and pharyngealisation only being found on short vowels.

It should also be mentioned that these analyses focused only on the vowel /a/, so further vowels need to be examined in order to understand the broader picture. Similarly, the effects of syllable structure and syllable position are factors that would benefit from more detailed analysis in future work. However, this work is already a contribution to phonetic and acoustic analysis of segmental and suprasegmental patterns in an understudied variety of Arabic and can guide future work into what effects need to be taken into account when analysing lexical stress in Arabic or other languages with stress and phonemic length and/or secondary factors such as pharyngealisation.

5. ACKNOWLEDGEMENTS

The collection of data used in this research was supported by the United States National Science Foundation under Grant 1125164 to Nancy Hall, California State University Long Beach. The recordings were kindly shared by Nancy Hall.

6. REFERENCES

[1] K. Abdul-Karim, “Aspects of the Phonology of Lebanese Arabic,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1980.

[2] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

[3] M. Abudalbuh, “Effects of Gender on the Production of Emphasis in Jordanian Arabic: A Sociophonetic Study,” Master’s thesis, University of Kansas, Kansas, 2010.

[4] M. Al-Masri and A. Jongman, “Acoustic Correlates of Emphasis in Jordanian Arabic: Preliminary

Results," in *Proceedings of the 2003 Texas Linguistics Society Conference*, A. Agwuele, Ed., Somerville, MA, 2004, pp. 96–106.

[5] M. E. Beckman and J. Edwards, "Lengthenings and shortenings and the nature of prosodic constituency," in *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman, Eds. Cambridge University Press, 1990, pp. 179–200.

[6] ——, "Articulatory evidence for differentiating stress categories," in *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, P. A. Keating, Ed. Cambridge University Press, 1994, pp. 7–33.

[7] M. E. Beckman, *Stress and Non-Stress Accent*. Netherlands: Foris, 1986.

[8] T. Benkirane, "Intonation in Western Arabic (Morocco)," in *Intonation Systems: A Survey of Twenty languages*, D. Hirst and A. Di Cristo, Eds. Cambridge: Cambridge University Press, 1998.

[9] A. Blodgett, J. Owens, and T. Rockwood, "An initial account of the intonation of Emirati Arabic," in *16th International Congress of Phonetic Sciences*, Saarbruecken, Germany, 2007, pp. 1137–1140.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.39," http://www.praat.org/, 2018.

[11] A. Bruggeman, N. Louriz, R. Almbark, and S. Hellmuth, "Acoustic correlates of lexical stress in Moroccan Arabic," *Journal of the International Phonetic Association*, vol. 51(3), pp. 425–449, 2021.

[12] D. Chahal, "Modelling the Intonation of Lebanese Arabic Using the Autosegmental-Metrical Framework: A comparison with English," Ph.D. dissertation, University of Melbourne, 2001.

[13] ——, "Phonetic Cues to Prominence in Lebanese Arabic," in *15th International Congress of Phonetic Sciences*, Barcelona, 2003.

[14] D. Chahal and S. Hellmuth, "The intonation of Lebanese and Egyptian Arabic," in *Prosodic Typology Volume II: The Phonology of Intonation and Phrasing*, S. Jun, Ed. Oxford: Oxford University Press, 2014, pp. 365–404.

[15] A. Cutler, "Lexical stress," in *The handbook of speech perception*, D. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, 2005, pp. 264–289.

[16] K. de Jong and B. A. Zawaydeh, "Stress, duration, and intonation in Arabic word-level prosody," *Journal of Phonetics*, vol. 27, pp. 3–22, 1999.

[17] ——, "Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration," *Journal of Phonetics*, vol. 30, pp. 53–75, 2002.

[18] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *Journal of the Acoustical Society of America*, vol. 27, pp. 765–768, 1955.

[19] T. Gay, "Physiological and acoustic correlates of perceived stress," *Language and Speech*, vol. 21(4), pp. 347–353, 1978.

[20] M. Gordon and T. B. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3 (1), 2017.

[21] N. Hall, "Acoustic differences between lexical and epenthetic vowels in Lebanese Arabic," *Journal of Phonetics*, vol. 41(2), pp. 133–143, 2013.

[22] ——, "Phonetic neutralization in Palestinian Arabic vowel shortening, with implications for lexical organization," *Glossa: a journal of general linguistics*, vol. 2(1), pp. 48. 1–23, 2017.

[23] S. Hellmuth, "Focus-Related Pitch Range Manipulation (and peak alignment effects) in Egyptian Arabic," in *Speech Prosody 2006, Third International Conference*, R. Hoffman and H. Mixdorff, Eds., Dresden, Germany, 2006.

[24] ——, "Intonational Pitch Accent Distribution in Egyptian Arabic," Ph.D. dissertation, School of Oriental and African Studies, London, 2006.

[25] L. Hyman, "Word-Prosodic Typology," *Phonology*, vol. 23(2), pp. 225–257, 2006.

[26] A. Jongman, W. Herd, M. Al-Masri, J. Sereno, and S. Combest, "Acoustics and perception of emphasis in Urban Jordanian Arabic," *Journal of Phonetics*, vol. 39(1), pp. 85–95, 2011.

[27] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82 (13), pp. 1–26, 2017.

[28] J. McCarthy, "On stress and syllabification," *Linguistic Inquiry*, vol. 10, pp. 443–466, 1979.

[29] R. T. Nasr, "The predictability of stress in Lebanese Arabic," *Phonetica*, vol. 4(2), pp. 89–94, 1959.

[30] M. Ortega-Llebaría and P. Prieto, "Acoustic Correlates of Stress in Central Catalan and Castilian Spanish," *Language and Speech*, vol. 54, pp. 73–97, 2009.

[31] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: {http://www.R-project.org}

[32] V. Sadeghi, "Word-level prominence in Persian: An Experimental Study," *Language and Speech*, vol. 6(4), pp. 571–596, 2017.

[33] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, vol. 100 (4), pp. 2471–2485, 1996.

[34] J. C. Watson, "Word stress in Arabic," in *The Blackwell Companion to Phonology*. Oxford, UK: Wiley-Blackwell, 2011, pp. 2990–3019.

[35] M. Yeou, "Locus equations and the degree of coarticulation of Arabic consonants," *Phonetica*, vol. 54(3-4), pp. 187–202, 1997.

[36] M. Younes, "On vowel shortening in Palestinian Arabic," in *Perspectives on Arabic Linguistics VII*, M. Eid, Ed. Amsterdam and Philadelphia: Benjamins, 1995, pp. 157–172.

PROSODIC CONVERGENCE ACROSS VARIETIES OF ITALIAN

Michelina Savino[1,3], Francisco Torreira[2] & Martine Grice[3]

[1,3]Dept. of Education, Psychology, Communication, University of Bari, Bari, ITALY
[2]Dept. of Linguistics, McGill University, Montreal, CANADA
[3]IfL-Phonetics, University of Cologne, Cologne, GERMANY
michelina.savino@uniba.it, francisco.torreira@mcgill.ca, martine.grice@uni-koeln.de

ABSTRACT

Polar question intonation in Italian is subject to regional variation. Here we investigate how far speakers of one variety (Bari Italian) converge in the production of such questions when playing a game with a partner from a different variety (Lecce Italian). Although these two varieties each have a number of contours in their repertoire, the distribution of these contours differs considerably. Bari Italian uses mainly Rise-Falls and Rise-Fall-Rises, whereas the Lecce variety uses mainly simple Rises. We show that speakers of Bari Italian produce fewer complex contours and more rises when interacting with a Lecce partner. They thus more frequently select the contour that is typical of the partner's variety. By contrast, an analysis of continuous properties of Rises and Rise-Fall-Rises reveals only minimal adjustments to the phonetic properties of each contour. Convergence thus appears to take place predominantly at the contour selection stage.

Keywords: spontaneous imitation, accommodation, question intonation, task-oriented dialogue, Italian.

1. INTRODUCTION

In research into prosody, sound-meaning mapping has been shown to be probabilistic and distribution-based, depending on linguistic and contextual factors [1], [2]. This is true for the intonation of polar questions in Italian, where a number of different contours are available, with highly different distributions across varieties [3]. Here we are concerned with two varieties, those of Bari (BA) and Lecce (LE), that, although they are in the same geographical region (Apulia), have been shown to have very different distributions. Analysis of task-oriented dialogues (map tasks) in [3] revealed that whereas BA speakers overwhelmingly prefer complex contours (rise-falls and rise-fall-rises) and rarely employ simple rises (98.3% vs. 1.7% respectively), LE speakers make use of simple rises most of the time (85.6%).

In this paper we investigate polar question contours, also from task-oriented dialogues, produced by speakers of Bari Italian when speaking to an interlocutor from the same variety as compared to when speaking to an interlocutor from Lecce. For reference, we also investigate the contours produced by the interlocutors from Lecce. A comparison of these contours and their distributions will tell us how far Bari speakers accommodate their productions according to the provenance of their interlocutor.

Previous studies have shown that speakers of the Bari variety can adjust their question contours towards those of other varieties when explicitly instructed to imitate them, although they tend to produce a blend of the target contour with their own native contour. For example, [4] found that they modified their native peak alignment in question rise-falls when instructed to imitate the rise-fall question contours of a native Neapolitan Italian model speaker, by shifting the peak towards the later Neapolitan contour. As both contours were rise-falls, the imitation involved adjustments to phonetic implementation within this contour shape. In a later study using the same paradigm [5], in which Bari speakers were instructed to imitate questions with the Lecce Italian rising contour, two different adjustments took place. First, they shifted their distributions so as to produce predominantly rises and very few rise-fall(-rise)s. Second, since the Bari inventory contains a further rise conveying non-finality, this contour appeared to influence the phonetic implementation of the imitated rises, such that they appeared to be a blend of the Bari non-final contour and the Lecce question contour.

In the current study, participants took part in a game eliciting spontaneous polar questions. Any imitation of the contours produced by the Bari speakers when interacting with Lecce speakers was not explicitly requested. We explore whether and to what extent Bari speakers converge to the Lecce conversational partners' question intonation by:
a) adjusting the distribution of the contour types;
b) adjusting the phonetic implementation of their contours.

2. METHODS AND MATERIALS

2.1. Elicitation method and experimental setting

To elicit polar questions in a spontaneous interactional setting, pairs of speakers were required to play the popular game "Guess who?". In this game, each

participant receives a board with an array of pictures of characters along with a name written underneath. The game starts with each player selecting a card from a separate pile of cards containing the same pictures as the board. The aim of the game is to guess which card the partner selected, by asking questions about characters' features in order to eliminate candidates until only one is left. Since the answers allowed are either "yes" or "no", questions are necessarily polar.

To introduce variation in the lexical composition of the questions, as well as in the game itself, participants were given five different boards, each with a different set of character types. Once one of the two players has successfully guessed three characters on one of the boards, participants started a new game with another board. The winner was the participant who had accumulated the highest score at the end of the session.

In each recording session, participants sat opposite each other at separate desks. They wore an AKG C520 condenser microphone headset connected to a Marantz PMD 661 digital recorder. A cardboard screen between the two desks prevented eye contact during the game. This restricted the interaction to the audio channel and ensured that players could not see each other's boards. A silent experimenter (the first author) was present during the session but without interfering in the game. All sessions were recorded in a quiet room at the University of Bari, and their average duration was 35 min.

2.2. Participants and recording sessions

Twelve participants, six speakers of Bari Italian (henceforth BA) and six speakers of Lecce Italian (henceforth LE) participated in the game sessions on a voluntary basis. They were all born and living in Bari and Lecce metropolitan areas respectively, aged 22-24. They were also female university students and did not know each other. These parameters were kept constant since it has been shown that gender (e.g. [6]) and familiarity (e.g. [7]) can play a role in speech accommodation. Participants were given one exam credit for participating in the experiment.

BA speakers participated in two different recording sessions:

- BA-LE speaker pair session
- BA-BA speaker pair session

The BA-BA game sessions were all recorded three months after the BA-LE counterparts. This delay between recordings was introduced to avoid possible long-term accommodation effects – as a consequence of interaction with a LE partner – on BA speakers' intonation productions during their exchanges with a same-variety interlocutor. A same-variety dyad condition was not recorded for LE speakers.

2.3. Preparation of speech materials for analysis

All game interactions were orthographically transcribed, and information-seeking polar questions identified based on listening to the recordings, taking the context, including the interlocutor's response, into account. At this stage one of the BA-LE recordings was excluded from further analysis because the spoken accent of the LE partner was judged as not genuinely LE by a native speaker of that variety. Utterances produced with accompanying disfluencies, laughter, or other kind of signal noise generating possible ambiguities in intonation analysis were also discarded.

After this initial screening, 1,439 utterances were included in the analysis: 487 questions produced by BA speakers in the BA-BA sessions, 478 by BA speakers in the BA-LE sessions, and 474 by LE partners in the BA-LE sessions.

3. DATA ANALYSIS

In the speech signal, time intervals corresponding to each question, the nuclear words and component syllables were annotated in Praat [8].

3.1 Phonological analysis of question contours

Polar questions were intonationally annotated by the first author in the Autosegmental-Metrical framework for varieties of Italian [9], [3], [10]. Within this framework, previous work has identified the predominant contour for polar questions in BA as a rising nuclear accent followed by either a low or a rising boundary tone sequence, analysed as L+H*L-L% (a rise-fall) and L+H*L-H% (a rise-fall-rise) respectively. In LE questions, the most widespread F0 contour has low pitch on the accented syllable and a rise at the intonational boundary, analysed as L*L-H% (a rise).

A number of additional polar question contours were identified and sorted as follows into macro-categories. Rises with a fall before the regular rise (transcribed as H+L* L-H%), were treated as a variant of the rise. Rise-falls with a late rise (annotated as L*+H L-L%) were treated as a variant of the regular rise-fall. The few truncated contours on final stressed syllables were treated as simple rises (n=3). The following macro-categories (based on overall F0 shape, as described above) entered into the analysis of the contour distribution**: Rise** (n=637), **Rise-Fall** (n=182)**,** and **Rise-Fall-Rise** (n=620).

3.1.1. Results

Fig. 1 displays the distribution of question contour types produced by all BA speakers across the two conditions (represented by the interlocutor variety in parenthesis) - BA(BA) and BA(LE) - alongside those of their LE conversational partners in the BA-LE sessions - (BA)LE.

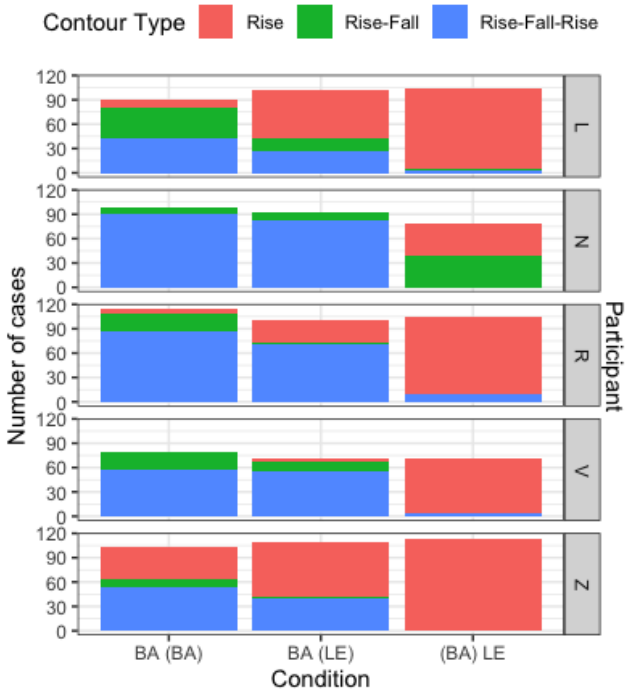


Figure 1: Distribution of contour types for questions produced by all BA speakers in BA(BA) vs BA(LE) sessions, and by all LE speakers in the (BA)LE sessions (BA= Bari Italian, LE=Lecce Italian, on x axis variety of interlocutor in parenthesis). The capital letter on the right indicates the BA speaker in the dyad referred to in the results

For BA speakers (in Fig.1 identified by the letters L, N, R, V, Z), results for the BA(BA) condition confirm the picture provided by past studies in the predominance of Rise-Fall-Rises (68%) and Rise-Falls (20%) over Rises (12%) in polar questions in this variety. For LE speakers, results also confirm the picture from previous studies, with considerably more simple Rises (88%) and very few Rise-Fall-Rises (4%) and Rise-Falls (9%). Crucially, when interacting with a LE interlocutor, BA speakers produced fewer rise-fall-rises (57%) and rise-falls (9%) but more rises (34%). We fitted a mixed-effects logistic regression model to the BA speaker data predicting whether the contour produced was a RISE (i.e., typical of LE) or either of the other two contours (typical of BA). CONDITION was the only fixed predictor (BA(BA) vs. BA(LE)). Random intercepts were fitted for SPEAKER and for the interaction between SPEAKER and CONDITION. The fitted model predicts Rises to be more likely in the BA(LE) condition than in the baseline BA(BA) condition (β =2.6; z=2.7; $p < .01$). This indicates that, when asking questions, BA speakers tended to accommodate intonationally to their LE interlocutors in terms of contour type selection.

Although the distribution of contour types realised by LE speakers in the BA-LE condition shows a predominant use of Rises, roughly corresponding to the distribution reported for LE speakers in native interactional settings [3], one LE speaker (BA speaker N's interlocutor) did make considerable use of Rise-Falls, indicating that this speaker might have accommodated to the BA partner in her contour selection. Interestingly, the BA partner in this dyad is the only one who did not adjust her contour selection, producing no Rises at all.

3.2. Quantitative analysis of question contours

We then examined the phonetic realization of nuclear contours in a subset of the data. In this subset, most combinations of factor levels of interest contained a fair number of observations. This reduced dataset consisted of all Rises and Rise-Fall-Rises featuring a word with penultimate stress spoken in modal voicing (n = 1335; 93% of the complete dataset).

We extracted pitch contours from the acoustic data using the autocorrelation method in Praat set to default parameters except for 'floor' (160 Hz for all speakers except one, 150 Hz), 'ceiling' (400 Hz for all speakers), and 'voicing threshold' (0.85). A more conservative voicing threshold allowed us to exclude microprosodic F0 perturbations occurring at the onset and offset of voiced intervals. We then smoothed and interpolated the extracted contours in Praat using default settings. Finally, we transformed the resulting pitch values to semitones (re 100 Hz), and normalized them relative to each speaker's median, as computed from the complete recordings. Fig. 2 shows the data broken down by SPEAKER and CONDITION as a function of CONTOUR. Note that, although the data for BA speakers N and V were not included in the statistical analyses below, their data are shown in Fig. 2.

We tested the effect of CONDITION on the shape of nuclear contours by fitting mixed-effects linear models predicting F0 scaling of the following tonal landmarks: (1) FINAL F0 in both Rises and Rise-Fall-Rises, (2) the accentual F0 PEAK, and the post-accentual F0 VALLEY in Rise-Fall-Rises. To simplify the interpretation of the models, we recoded CONDITION so as to capture the two main differences of interest in our study: (1) between BA speakers when speaking to a LE or a BA conversational

partner, i.e. BA(LE) vs. BA(BA), and (2) between LE speakers and their BA interlocutors, i.e. (BA)LE and BA (LE). Because of model convergence issues, we could not include interactions between CONDITION and SPEAKER in the models (but see discussion of speaker variation below).

For Rises, FINAL F0 was found to be significantly higher for LE speakers than for their BA interlocutors (Contrast 2: β =1.78; t = 2.6; p < .05), but we observed no difference between conditions BA (LE) and BA (BA) (Contrast 1: β =0.18; t = 7.6; p = .44). Separate linear models fitted to each participant's data confirmed these findings at the individual level (note that participants L and V were excluded from this analysis due to lack of data).

For Rise-Fall-Rises, the mixed-effects model fitted to all participants yielded a statistical difference in F0 VALLEY between BA speakers when speaking to a LE or a BA conversational partner. Rise-Fall-Rises reach slightly lower F0 values in the BA (LE) condition relative to the control BA (BA) condition (Contrast 1: β = -0.24; t=-3.07; p = .005). Linear models separately fitted to each participant revealed however that this effect was exclusively driven by speaker N, as the contrast (BA(LE) vs BA (BA)) did not reach or approach significance in any of the models fitted to the other participants.

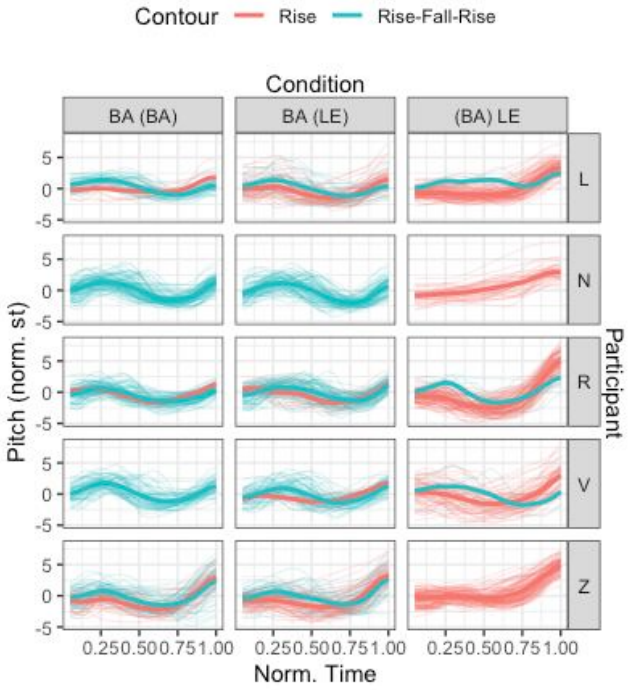


Figure 2: Speaker-normalized pitch (st) as a function of normalized time (1 = duration of nuclear region) for each CONDITION and PARTICIPANT in Rises and Rise-Fall-Rises featuring words with penultimate stress (see text for details).

Analysis of individual participants' data further showed that speaker R's F0 PEAK and FINAL F0 were significantly higher when interacting with her LE than her BA partner (F0 PEAK: β =.42; t = 3.1; p < .005; FINAL F0: β =-0.7; t = 4.52; p < .0001). For BA speakers N and R, therefore, our results point to some phonetic adjustment of the RFR questions contours when exposed to the intonation of their LE conversational partner. Individual analyses also showed, however, that speaker V produces *lower* F0 PEAK (β = -0.54, t = -2.6, p < 0.01), and that speaker N produces *lower* FINAL F0 (β = -0.57, t = -3.13, p < 0.005) when speaking to a LE speaker than when speaking to BA speaker. Together, these results indicate that if speakers make any phonetic adjustments, they are very subtle.

4. DISCUSSION AND CONCLUSION

Results from our explorative study have shown that, in spontaneous game-based dialogue with a LE conversational partner, BA speakers tend to converge to the LE intonation by shifting their native distribution of the question contour types. In particular, they produce more Rises (the predominant question contour in Lecce Italian) and fewer Rise-Falls and Rise-Fall-Rises (the contours with the highest probability in their native variety) when compared to interactions with a same-variety partner. Results indicate that prosodic accommodation is mainly realised at the contour selection stage, since we observed that distribution of contour type was not necessarily accompanied by an adjustment in the phonetic implementation (e.g. of the final rise in Rises and Rise-Fall-Rises).

Importantly, our preliminary results provide further support to the idea that, in prosody, the sound-meaning mapping is probabilistic and distribution-based [1, 2]. In this view, prosodic convergence can be accounted for as shifts in distributions of individual contour types (composed of pitch accent and edge tone sequences). Bari Italian speakers exposed to Lecce Italian question intonation appear to increase the production of Rises not only when they are explicitly instructed to imitate the Lecce Rises uttered by a model speaker, as reported in [5], but also in cases where they are not instructed to pay attention to intonation at all, as in the current study. However, differences in the task – explicit imitation vs spontaneous interaction – implying different conditions in terms of allocation of attention and memory in imitation [11], [12] lead to a greater proportion of Rises in the imitation task than in the conversational setting.

6. ACKNOWLEDGEMENTS

For funding we are grateful to the German Research Foundation (DFG) – Project-ID 281511265 – SFB 1252 “Prominence in Language” (project A01) at the University of Cologne.

7. REFERENCES

[1] Calhoun, S. 2010. The centrality of metrical structure in signalling information structure: A probabilistic perspective. *Language* 86(1), 1–42.

[2] Kurumada, C., Roettger, B. T. 2021. Thinking probabilistically in the study of intonational speech prosody. *WIREs Cognitive Science 13*(1):e1579.

[3] Savino, M. 2012. The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association* 42(1), 23–48.

[4] D’Imperio, M., Cavone, R., Petrone, C. 2014. Phonetic and phonological imitation of intonation in two varieties of Italian. *Frontiers in Psychology* 5, 1–10.

[5] Savino, M., Sbranna, S., Ventura, C., Albert, A., Grice, M. 2022. Imitating intonation in a non-native variety: the influence of the native repertoire. *Proceedings of Speech Prosody 2022*, Lisbon, 739–743.

[6] Pardo, J. 2006. On the phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119 (4), 2382–2993.

[7] Truong, P. K., Heylen, D. K. J. 2012. Measuring prosodic alignment in cooperative task-based conversations. *Proceedings of Interspeech 2012*, Portland, 843–846.

[8] Boersma, P., Weenink, D. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–345.

[9] Grice, M., D’Imperio, M., Savino, M., Avesani, C. 2005. Strategies for intonation labelling across varieties of Italian. In Jun, Sun-Ah (ed.), *Prosodic Typology: the Phonology of Intonation and Phrasing*. Oxford University Press, 362–389.

[10] Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D’Imperio, M., Giordano, R., Marotta, G., Savino, M., Sorianello, P. 2015. Intonational phonology of the regional varieties of Italian. In Frota, S., Prieto, P. (eds.), *Intonation in Romance*. Oxford University Press, 140–197.

[11] Goldinger S.D. 2013. The cognitive basis of spontaneous imitation: evidence from the visual world. *Proceedings of Meetings on Acoustics*, vol. 19, 1–6.

[12] Heath J. 2017. How automatic is phonetic convergence? In *Proceedings of the Linguistic Society of America* 2-35, 1–10.

Rhythmic Similarity hypothesis for cross-cultural vocal emotion recognition

Yachan Liang[1], Jiyoun Choi[2], Mirjam Broersma[1], Martijn Goudbeek[3], Agnieszka Konopka[4]

[1]Centre for Language Studies, Radboud University Nijmegen, The Netherlands

[2]Sookmyung Women's University, Seoul, Korea

[3]Tilburg University, The Netherlands

[4]University of Aberdeen, UK

ABSTRACT

This study examines the newly formulated Rhythmic Similarity hypothesis, which proposes that cross-cultural vocal emotion recognition is more accurate between users of languages with similar rhythmic structures. To disentangle the close relationship between culture and language, this study tested how well American English and French listeners recognized emotions in recordings produced by native speakers of Dutch and Korean. Since French and Korean share similar rhythmic structures, the prediction was that French listeners would outperform American English listeners in recognizing emotions in Korean. However, this prediction was not supported by our data, disconfirming the Rhythmic Similarity hypothesis. Furthermore, emotion recognition accuracy for both listener groups (American English and French) was higher in Dutch than in Korean, supporting the Cultural Proximity [1] and Language Distance hypotheses [2]. Additionally, recognition accuracy was above-chance for all emotions and was affected by arousal, valence, and basicness in ways consistent with previous findings [3, 4, 5].

Keywords: Rhythmic Similarity hypothesis, vocal emotion recognition, cross-cultural, cross-linguistic

1. INTRODUCTION

Previous research has shown that emotion recognition is a product of interactions between universal and culture-/language-specific factors [4, 6]. Listeners from different cultures and with different native languages recognize emotions above chance even in an unknown language, which shows that emotions are conveyed in a similar way across cultures and languages. At the same time, emotion recognition is more accurate when speakers and listeners share the same culture and native language, a phenomenon known as the in-group advantage [7]. However, the specific contributions of culture and language to emotion recognition remain unknown.

Most studies have addressed emotion recognition predominantly either from a cultural perspective or from a linguistic perspective. Disentangling the effect of culture and language has been notoriously difficult. For this reason, two separate theories address the relative contributions of culture and language to emotion recognition: the Cultural Proximity and Language Distance hypotheses.

The Cultural Proximity hypothesis by Elfenbein and Ambady [1] proposes that listeners with relatively similar cultural backgrounds can identify each other's emotional expressions more accurately than those with very different ones. This theory thus likens cultural variation to "dialects" in languages. For instance, to examine the effect of cultural differences on emotion recognition, Laukka et al. [8] tested emotion recognition in English among listeners from five different English-speaking countries (America, Australia, India, Kenya, and Singapore). They found that listeners were more accurate in recognizing emotions produced by speakers from their own culture than from a different one.

The Language Distance hypothesis introduced by Scherer et al. [2] claims that it is easier for listeners to decode emotions produced in a language that is typologically similar to their native language than in a language that is typologically different. In a pioneering study, Scherer et al. [2] investigated the influence of language distance on emotion recognition. They presented 30 pseudo-utterances expressing five basic emotions (anger, fear, joy, sadness, and neutral) produced in German to listeners from nine different countries whose native languages were German, Dutch, English, French, Italian, Spanish, and Indonesian. While recognition accuracy in each listener group was above chance, German listeners had the highest accuracy, with Dutch and English listeners following next, which is consistent with the Language Distance hypothesis. In contrast, Indonesian listeners, whose native language is typologically the least similar to German, had the lowest accuracy. Similarly, Pell et al. [9] tested the recognition of five basic emotions (anger, disgust, fear, joy, and sadness) in pseudo-utterances produced

in Argentine Spanish, Arabic, German, and English, by monolingual listeners of Argentine Spanish. The results showed that Argentine Spanish listeners had above-chance recognition accuracy in all languages, but with the highest accuracy in their native language.

The close relationship between culture and language makes it difficult to disentangle the specific effects of the two on emotion recognition. For example, in Scherer et al.'s [2] study, Dutch and English are not only typologically, but also culturally closer to German than to some of the other languages/cultures in the sample, creating a potential confound. Furthermore, while the Language Distance hypothesis proposes a central role for linguistic similarity, it does not specify which aspects of language typology might play a role. In the present study, we investigate the role of one potentially highly relevant language property, namely rhythmic structure on vocal emotion recognition, with the aim of disentangling the effect of culture and language. We tested emotion recognition in listeners of American English and French exposed to recordings produced by Dutch and Korean speakers. In terms of culture and language, Dutch, American English, and French are related, as they are Indo-European languages spoken by people from Western cultures. Korean, however, is a non-Indo-European language spoken by people from an Asian culture, and is thus unrelated to the other three cultures and languages.

With regard to the rhythmic structure, however, these languages are grouped differently (see Table 1). In Dutch and American English, the foot is the main grouping element in rhythm, and there is lexical stress, i.e., prosodic prominence is used to differentiate word meanings [10, 11]. In French [12] and Korean [13], on the other hand, the phrase is the main rhythmic grouping element, and there is no lexical stress. There are Intonational Phrases (IP) and Accentual Phrases (AP) in French and Korean, and the phrase boundary is typically signaled by a final rising pitch movement and lengthening [12, 14, 15].

Table 1: The relationships between the four languages in terms of culture and language.

	Dutch		Korean	
	Culture	**Language: Overall typology**	**Culture**	**Language: Rhythmic similarity**
American English	Relatively similar	Relatively similar	Dissimilar	Dissimilar
French	Relatively similar	Relatively similar	Dissimilar	Similar

This constellation of languages allows us to investigate whether rhythmic similarities facilitate cross-cultural vocal emotion recognition. Therefore, building upon the Language Distance hypothesis, we propose the Rhythmic Similarity hypothesis, which predicts that cross-linguistic communication of vocal emotion will be more accurate if the languages have similar rhythmic structures than if they have dissimilar rhythmic structures.

This study therefore addresses two questions. First, it examines whether American English and French listeners recognize emotions more accurately in Dutch than in Korean. According to the Cultural Proximity and Language Distance hypotheses, we hypothesize that both groups of listeners will perform better in Dutch than in Korean because American English and French are culturally and linguistically closer to Dutch than Korean (Hypothesis 1). Second, it examines whether French listeners perform better than American English listeners on the Korean recordings. Based on the Rhythmic Similarity hypothesis, we predict that French listeners will outperform American English listeners in Korean since Korean is rhythmically similar to French but not to American English (Hypothesis 2). In addition, we will present a series of analyses comparing recognition accuracy to chance level, and assessing the effects of arousal, valence, and basicness on emotion recognition accuracy.

2. METHOD

2.1. Participants

Participants (referred to as "listeners" below) were twenty-five native American English listeners (19 females, 6 males, age: $M = 20.6$, $SD = 1.94$) who were students at Northwestern University, Chicago, and thirty native French listeners (22 females, 8 males, age: $M = 22.7$, $SD = 4.10$) who were students at the university École Normale Supérieure, Paris. None of the participants reported any speech or hearing problems, or any knowledge of Dutch or Korean. All participants were given either course credits or a small payment as a reward for their participation.

2.2. Materials

As auditory stimuli, we used all the vocal expressions from the Demo/Koremo (Dutch emotion/Korean emotion) corpus collected by Goudbeek and Broersma [16]. To avoid any semantic cues to emotion recognition, the corpus uses a pseudo-phrase [nuto hɔm sɛpikaŋ] which is phonologically legal in both Dutch and Korean, meeting the "stimuli equivalence" requirement proposed by Matsumoto [17]. The pseudo-phrase was produced by professional Dutch and Korean voice actors (referred to as "speakers" below). This corpus includes an equal number of basic and non-basic emotions, which were all used in the present study, contrary to prior studies that have mainly focused on basic emotions like anger, disgust, fear, happiness, sadness, and

surprise [18]. Moreover, the stimuli are balanced on two dimensions that are crucial in the understanding of emotions [19], namely arousal (high-arousal vs. low-arousal) and valence (positive vs. negative; Table 2). The stimuli consisted of 256 portrayals (8 emotions x 2 tokens per emotion x 8 speakers x 2 languages) in total. For more information regarding the corpus and the recording procedure, please refer to Goudbeek and Broersma [16].

Table 2: The eight emotions included in this study in a valence by arousal grid (reproduced from Goudbeek & Broersma, 2010, p.2212), with basic emotions denoted with *.

		Valence	
		Positive	Negative
Arousal	High	Joy*	Anger*
		Pride	Fear*
	Low	Tenderness	Sadness*
		Relief	Irritation

2.3. Procedure

All participants were tested individually in a sound-attenuated room at their university. On each trial, a computer screen in front of the participant showed an "emotion wheel" (Figure 1), with eight emotions written in the participants' native language (English or French), each with four circles in different sizes indicating different intensities. Participants listened to the recordings via high-quality headphones. There was no time constraint for their responses. The experiment was administered using JATOS [20] on a standard laboratory computer and took around 35-45 minutes.

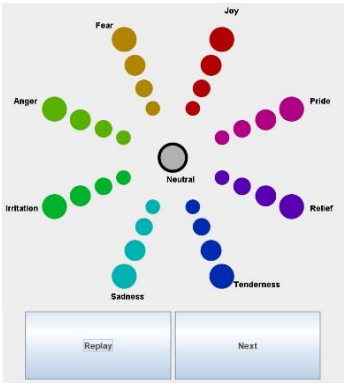


Figure 1: The emotion wheel in English

All participants were given written instructions in their native language, asking them to listen to each recorded stimulus, to choose the emotion they believed the recording expressed from the eight choices shown on the screen, and to indicate the intensity of that emotion. They could also choose Neutral if they felt the recording was not emotional at all. They made their response by clicking on one of the circles on the emotion wheel. They could listen to each recording as often as they preferred. In this study, we only analyzed categorical responses (i.e., recognition accuracy, but not intensity ratings).

The experiment started with a block containing the 128 Korean stimuli, followed by a block containing the 128 Dutch stimuli. Prior to each block, participants were informed about its language, and they performed eight practice trials.

3. RESULTS AND DISCUSSION

Analyses were performed in R [21] with a series of logistic mixed effects models run with the *lme4* package [22]. The models included a combination of the following five factors: Speaker Language (Dutch vs. Korean recordings), Listener Language (American English vs. French listeners), Arousal (high-arousal vs. low-arousal emotions), Valence (positive vs. negative emotions), and Basicness (basic vs. non-basic emotions). We used regression-style contrast coding (i.e., -.5 and .5 contrast codes for the first and second levels of each factor listed above) in all models. The outcome variable was the accuracy of emotion recognition (correct vs. incorrect).

Each model used the maximal random structure justified by the experimental design that allowed convergence. All models initially included random by-participant and by-item intercepts, by-participant slopes for Speaker Language, Arousal, Valence, and Basicness, and by-item slopes for Listener Language. In cases of non-convergence, models were sequentially simplified by dropping the random slope for the variable accounting for the least variance [23] until convergence was reached.

Table 3: Summary of results of the logistic mixed effects model analysis addressing Hypothesis 1 and 2 (only by-item slopes improved model fit).

	Estimates			
	β	Exp(β)	*SE*	*z* value
Model1(Hypothesis 1)				
Intercept	-0.44	0.64	0.10	-4.65***
Speaker Language (*s*)	-0.49	0.61	0.17	-2.83**
Listener Language (*i*)	0.17	1.19	0.10	1.69.
Speaker Language (*s*) x Listener Language (*i*)	-0.03	0.97	0.13	-0.23

"***" $p < .001$, "**" $p < .01$, "*" $p < .05$, "." $.05 < p < .10$

***Notes*: *(s)* and *(i)* indicate the inclusion of by-participant and by-item random slopes respectively.**

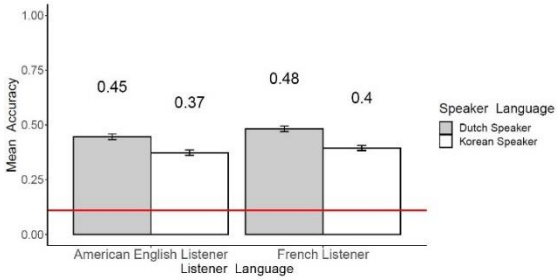


Figure 2: Recognition accuracy for Dutch and Korean recordings by American English and French listeners (by-participant means). Red line indicates chance performance (.11). Error bars are ±1 SE.

Hypothesis 1*: Both listener groups are predicted to have higher recognition accuracy in Dutch compared to Korean.* We tested this hypothesis in a model that included the interaction between Speaker Language and Listener Language (Table 3). The model revealed a main effect of Speaker Language, indicating that, consistent with the Cultural Proximity and Language Distance hypotheses, both American English and French listeners recognized emotions more accurately in Dutch than in Korean (a .16 difference; Figure 2). Additionally, there was a marginally significant main effect of Listener Language, as French listeners performed slightly better than American English listeners overall (.06 higher recognition accuracy).

Hypothesis 2*: French listeners are predicted to outperform American English listeners in Korean recordings.* There was no interaction between Speaker and Listener Language: French listeners did not show higher accuracy than American English listeners when responding to Korean recordings (mean accuracy: .40 vs. .37). Thus, the Rhythmic Similarity hypothesis was not supported by our data.

In addition, we tested for the main effects of three emotion variables (Arousal, Valence, and Basicness), and for their interactions with Speaker Language and Listener Language in separate models (Table 4). The results showed a number of similar patterns as in previous studies [3, 4, 5]: higher accuracy for (i) low-arousal, (ii) negative, and (iii) basic emotions than for (i) high-arousal, (ii) positive, and (iii) non-basic emotions (i.e., three main effects). The analyses also showed a significant two-way interaction between Listener Language and Arousal, as well as Listener Language and Basicness (all *zs* < -2.74), but not with Valence. Importantly, no three-way interactions reached significance.

Table 4: Summary of results of the logistic mixed effects model analyses with Arousal, Valence, and Basicness (all random slopes improved model fit, except by-participant slopes for Speaker Language).

	Estimates			
	β	Exp(*β*)	*SE*	*z* value
Analysis with Arousal				
Intercept	-0.45	0.64	0.10	-4.74***
Arousal (*s*)	0.49	1.62	0.19	2.58**
Speaker Language (*s*) x Listener Language (*i*)	0.003	1.00	0.12	0.03
Listener Language (*i*) x Arousal (*s*)	-0.86	0.42	0.19	-4.44***
Speaker Language (*s*) x Listener Language (*i*) x Arousal(*s*)	-0.12	0.89	0.24	-0.49
Analysis with Valence				
Intercept	-0.45	0.64	0.09	-5.20***
Valence (*s*)	-1.34	0.26	0.16	-8.21***
Speaker Language (*s*) x Listener Language (*i*)	-0.02	0.98	0.13	-0.15
Listener Language (*i*) x Valence (*s*)	0.001	1.00	0.18	0.01
Speaker Language (*s*) x Listener Language (*i*) x Valence (*s*)	0.29	1.34	0.26	1.14
Analysis with Basicness				
Intercept	-0.45	0.64	0.09	-4.94***
Basicness (*s*)	-1.06	0.35	0.17	-6.16***
Speaker Language (*s*) x Listener Language (*i*)	-0.02	0.98	0.13	-0.12
Listener Language (*i*) x Basicness (*s*)	-0.48	0.62	0.17	-2.74**
Speaker Language (*s*) x Listener Language (*i*) x Basicness (*s*)	-0.05	0.95	0.25	-0.20

"***" $p < .001$, "**" $p < .01$, "*" $p < .05$, "." $.05 < p < .10$

Notes: *(s)* and *(i)* indicate the inclusion of by-participant and by-item random slopes respectively.

4. CONCLUSION

The primary goal of this study was to assess the role of language in cross-cultural/linguistic emotion recognition by comparing the performance of American English and French listeners on Dutch and Korean recordings. Building upon the Language Distance hypothesis [2], the newly proposed Rhythmic Similarity hypothesis predicted that cross-cultural (or cross-linguistic) communication of vocal emotion would be facilitated in languages with a shared rhythmic structure. However, this hypothesis was not supported by our data. Even though French and Korean share similar rhythmic structures, with prosodic phrases marked by a final rising pitch and lengthening [12], French listeners were not better at recognizing emotions in Korean than American English listeners were.

A second aim was to investigate the effects of Cultural Proximity and Language Distance on emotion recognition. The results that both groups of listeners recognized vocal emotions more accurately in Dutch, which is both culturally and linguistically closer to American English and French than to Korean, are in line with previous findings that cultural and linguistic similarities improve emotion recognition cross-culturally/linguistically [1, 2, 7].

Additionally, we replicated earlier findings that listeners are capable of recognizing discrete emotions above chance even when they are produced by a speaker from a different culture and language [4]. Furthermore, we found the effects of arousal, valence, and basicness on the recognition of emotion consistent with previous research in this domain [3, 4, 5].

To conclude, our findings corroborate earlier work on the effects of culture and language on vocal communication of emotion. However, since we found no evidence for the Rhythmic Similarity hypothesis, the precise mechanisms by which linguistic similarity contributes to emotion recognition remain elusive.

5. REFERENCES

[1] H. A. Elfenbein and N. Ambady, "Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition," *J. Cross. Cult. Psychol.*, vol. 34, no. 1, pp. 92–110, 2003, doi: 10.1177/0022022102239157.

[2] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross. Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001, doi: 10.1177/0022022101032001009.

[3] Y. Liang, M. Goudbeek, A. Konopka, J. Choi, and M. Broersma, "Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design," *Lang. Speech*.

[4] P. Laukka and H. A. Elfenbein, "Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis," *Emot. Rev.*, vol. 13, no. 1, pp. 3–11, 2021.

[5] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 23, p. E3086, 2015, doi: 10.1073/pnas.1508604112.

[6] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *J.K. Cole (Eds.), Nebraska Symposium on Motivation*, vol. 19. University of Nebraska Press, Lincoln, pp. 207–283, 1972.

[7] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, 2002, doi: 10.1037/0033-2909.128.2.203.

[8] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *J. Pers. Soc. Psychol.*, vol. 111, no. 5, pp. 686–705, 2016, doi: 10.1037/pspi0000066.

[9] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *J. Nonverbal Behav.*, vol. 33, no. 2, pp. 107–120, 2009, doi: 10.1007/s10919-008-0065-7.

[10] C. Gussenhoven, "Transcription of Dutch intonation," *Sun-Ah Jun (Eds.), The phonology of intonation and phrasing*, Oxford: Oxford University Press, 2005, pp. 118–145. doi: 10.1093/acprof.

[11] P. A. Bertan, "Prosodic Typology: On the Dichotomy between Stress-Timed and Syllable-Timed Languages A," *Lang. Des.*, vol. 2, pp. 103–130, 1999.

[12] S. Jun and C. Fougeron, "Realizations of accentual phrase in French intonation," vol. 14, no. 2002, pp. 147–172.

[13] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *J. Phon.*, vol. 40, no. 3, pp. 351–373, 2012, doi: 10.1016/j.wocn.2012.02.003.

[14] S.-A. Jun, "Intonational phonology of Seoul Korean revisited," *Japanese/Korean Linguist. 14*, pp. 15–26, 2006.

[15] J. Kim, C. Davis, and A. Cutler, "Similarity : II . Syllable Rhythm," vol. 51, no. 4, pp. 343–359, 2008.

[16] M. Goudbeek and M. Broersma, "The Demo/Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010).*, 2010, pp. 2211–2215.

[17] D. Matsumoto, "Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence," *Psychol. Bull.*, vol. 128, no. 2, pp. 236–242, 2002, doi: 10.1037/0033-2909.128.2.236.

[18] P. Ekman, "Are there basic emotions?," *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992, doi: 10.1037/0033-295X.99.3.550.

[19] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[20] K. Lange, S. Kühn, and E. Filevich, "'Just another tool for online studies' (JATOS): An easy solution for setup and management of web servers supporting online studies," *PLoS One*, vol. 10, no. 6, pp. 1–14, 2015, doi: 10.1371/journal.pone.0130834.

[21] R Core Team, "R: A language and environment for statistical computing," in *R Foundation for Statistical Computing*, Vienna, 2022. [Online]. Available: https://www.r-project.org

[22] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[23] D. . Barr, R. Levy, C. Scheepers, and H. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.*, vol. 68, no. 3, pp. 255–278, 2013, doi: 10.1016/j.jml.2012.11.001.

REDUNDANCY AND INDIVIDUAL VARIABILITY IN THE PROSODIC MARKING OF INFORMATION STATUS IN GERMAN

Janne Lorenzen[1], Simon Roessig[2], Stefan Baumann[1]

[1]I*f*L Phonetik, University of Cologne, [2]Cornell Phonetics Lab, Cornell University
{janne.lorenzen; simon.roessig; stefan.baumann}@uni-koeln.de

ABSTRACT

Information status in German is encoded via prosodic prominence, such that *new* referents are realized prosodically most prominently, *given* referents least prominently and *accessible* referents lie in between these extremes. Since prominence is multifaceted, encompassing a variety of cues related to timing, spectral properties and the F0 contour, one may ask whether speakers are redundant in their prosodic encoding of pragmatic contrasts or select different cues to focus on.

To address this question, we collected data from 32 participants in an interactive reading task with target words varying in information status. We measured six acoustic correlates of prominence related to timing and the F0 contour. Results indicate that some speakers exploit the full range of these cues. However, most speakers reduce this maximally redundant encoding of prominence by following individual strategies. These are compensatory in nature, prioritizing select cues which nevertheless serve the same prominence-marking function.

Keywords: German, individual differences, information status, prominence, prosody

1. INTRODUCTION

Prosodic prominence is multidimensional, encompassing a variety of cues related to timing, changes in F0 and spectral characteristics of the speech signal [1],[2]. Pragmatic categories differentiated via prominence levels are thus often redundantly encoded [3]. To reduce this redundancy, speakers may only select single cues, which gives way to considerable inter-speaker variability in prominence production. In this paper, we investigate how redundant individual speakers are in their prosodic encoding of information status in German.

Inter-speaker variability has recently gathered much attention in prosodic research. For example, [4] show that German speakers differ in the type and number of prosodic cues they employ to distinguish between broad, narrow and contrastive focus. Similarly, [5] conclude that American English (AE)-speaking individuals differ in how strongly they encode informativity (i.e., focus type, contextual probability, and word frequency) via the F0 contour. [6] finds considerable individual differences in boundary marking by AE speakers. Concomitantly, perception studies observe substantial variability in the cues listeners attend to in the decoding of prosodic prominence [1],[4].

Previous studies on individual differences in prominence production have been concerned predominantly with focus marking, a function that in German is typically strongly encoded via prosodic prominence, e.g., the use of different pitch accent types [7]. The present study however looks at information status. Information status in German is often prosodically marked in that *discourse-new* referents are most prominent, *accessible* referents are less prominent and *given* referents least prominent [8],[9],[10]. Since information status can also be signaled via morphosyntactic cues (i.e., *new* referents are preceded by indefinite articles, *accessible* and *given* referents by definite ones), the *given-new* distinction is arguably prosodically less strongly encoded than focus. This allows for even more variability to emerge, as speakers may choose not to encode this contrast via prosodic cues at all.

This paper investigates how differently speakers behave in their prosodic encoding of information status. We will consider several acoustic cues, allowing us to explore the degree of redundancy in prominence encoding, both at population level and individually.

2. METHODOLOGY

2.1. Reading material

To address these research questions, we collected data from a reading task. Target words were disyllabic with stress on the initial syllable. They were embedded in eight different short stories consisting of four sentences each. The third sentence included the target words in indirect and direct object roles. These were either referentially and lexically *new* or *accessible* through the context (see Fig. 1 for an example of a story with an *accessible* indirect object and a *new* direct object). Each story was devised in four different combinations of *new* and *accessible* target words, but participants saw only one version of each story in a Latin square design.



Anna hatte einen erfolgreichen Tag.

"Anna had a successful day."



Am Mittag sind einige Handwerker von der Baustelle nebenan in ihr Geschäft gekommen.

"At noon, some craftsmen from the construction site next door came into her store."



*Unter anderem hat sie **dem Maler eine Waage** verkauft.*

"Among other things, she sold the painter a scale."



Jetzt ist sie zu Hause und entspannt sich bei schöner Musik.

"Now she is at home and relaxes by listening to beautiful music."

Figure 1: Example story with an *accessible* target word ('painter') followed by a *new* one ('scale').

2.2. Experimental procedure and participants

Recordings were collected remotely via a video call with the participant, the experimenter and a confederate. Participants were asked to complete an interactive task with the confederate, in which they read a short story aloud so that the confederate would be able to memorize the story and sort corresponding picture cards into the correct order. During the task, participants wore headphones and recorded themselves onto their own laptops via a podcasting app (Ennuicastr, https://ecastr.com/), which served to circumvent recording issues caused by unstable internet connections. We recorded 32 native speakers of German (8 male, 24 female, 20-38 years old). Each speaker read one version of each story, resulting in 512 target words (256 utterances), which were always produced in broad focus and with a pitch accent.

2.3. Acoustic parameters

We measured three parameters related to the alignment and scaling of the F0 contour (synchrony, tonal onglide, and Delta F0), two parameters related to timing (syllable and word duration) and one integrating both timing and intensity (mass).

Syllable duration and word duration were measured only in (the accented syllable of) non-phrase-final target words, in order to avoid differences due to final lengthening. Periodic energy mass is a measure from the ProPer toolbox [11] which quantifies the area under the periodic energy curve as an integral of duration and intensity. Note that mass in a syllable is considered relative to the average mass of a syllable in the utterance and is therefore unitless. Tonal onglide [12] was measured as the difference in semitones between the tonal target of a pitch accent and a preceding "non-accent tone", as determined by labels annotated for the data following the DIMA guidelines [13]. L*+H accents were excluded from tonal onglide measures, as the most salient F0 movement arguably occurs *after* the tonal target. Synchrony and Delta F0 are measures from the ProPer toolbox as well. In contrast to tonal onglide, these parameters are not dependent on landmark annotations, characterizing the F0 contour without the need for prior labeling. Synchrony is a measure of alignment calculating the distance in milliseconds between the tonal center of gravity [14] and the center of periodic energy mass of a syllable. It captures the extent and direction of the F0 contour as well as its shape. Synchrony is reported relative to the duration of the syllable in percent. Delta F0 characterizes the difference in F0 (in semitones) between two syllables, measured at the center of mass of the accented syllable and the center of mass of the preceding one.

The parameters are either locally confined to the accented syllable (syllable duration, synchrony, periodic energy mass) or characterize the whole target word (word duration, tonal onglide, Delta F0). All parameters were z-scored for further analysis.

2.4. Statistics

We used Bayesian mixed effects linear regression models to query the role each acoustic parameter plays in distinguishing *new* from *accessible* referents. Each model contained information status (levels: *new*, *accessible*) as the predictor and one of the acoustic parameters as the response variable. In addition, we included random intercepts for word and speaker and by-speaker random slopes for information status.

The models were fitted in R [15] using the Stan modeling language [16] via the *brms* package [17]. Four sampling chains ran for 4000 iterations each with a warm-up period of 2000 iterations, yielding a total of 8000 posterior samples per model. We used a weakly informative, normally distributed prior with a

mean of zero and a standard deviation of ten for the regression coefficients and default priors supplied by *brms* for the remaining parameters. For each model, we report the estimate β and 90% credible intervals under the posterior distributions as well as the posterior probability that β is larger than zero. We judge there to be compelling evidence for the hypothesis that β>0 if zero is not included in the 90% credible interval and the posterior probability Pr(β>0) is larger than 0.95.

To investigate individual differences in the usage of the different acoustic parameters, we further exploit the random slopes estimates. Random slopes capture the differences in the direction and size of an effect between individuals, which makes them ideal for the analysis of speaker-specific behavior. We ran a hierarchical cluster analysis on the random slopes, following [1] who performed a similar analysis on frequentist regression models. This serves to group together speakers who follow similar strategies in their encoding of information status. Rather than considering each of the 32 individuals separately, which quickly becomes convoluted, the amount of the strategies that need to be considered is thus reduced in an objective manner. Furthermore, we tested for correlations between the random slope values associated with the parameters to find out which cues are used to the same end.

3. RESULTS

3.1. Overall results

Parameter	β	90% CI	Pr(β>0)
Tonal onglide	0.21	[0.08; 0.34]	0.99
Synchrony	0.11	[0.01; 0.21]	0.96
Delta F0	0.35	[0.22; 0.47]	1.00
Word duration	0.12	[-0.02; 0.26]	0.92
Syllable duration	0.06	[-0.07; 0.19]	0.76
Periodic energy mass	0.01	[-0.13; 0.14]	0.53

Table 1: General results of the Bayesian models for each z-scored parameter.

To be able to assess the usefulness of investigating individual differences, we will first consider population-level effects. Table 1 shows the results of Bayesian mixed-effects models for every acoustic parameter. Overall, F0-based cues are more likely to distinguish *new* from *accessible* referents than duration-based cues, with higher Delta F0 constituting the strongest cue to newness. Adhering by the decision criteria determined in Section 2.4, changes in tonal onglide, synchrony, and Delta F0 are reliably predicted by information status, while the other parameters fail to meet these criteria.

3.2. Individual and group behavior

Descriptively, the F0-based parameters are also the ones used most consistently across speakers (see Table 2). 22 speakers mark *new* referents with higher tonal onglide values, 21 speakers produce higher synchrony values in *new* referents and 27 speakers use higher Delta F0 in *new* referents. Only 15 speakers use periodic energy mass, which accords with the model results.

Parameter	# of users	mean acc.	mean new	Δ new – acc.
Tonal onglide (in st)	22	0.66	2.61	1.94
Synchrony (in %)	21	1.73	3.43	1.70
Delta F0 (in st)	27	-0.43	0.58	1.01
Word duration (in ms)	19	384	408	24
Syllable duration (in ms)	17	239	261	22
Periodic energy mass (unitless)	15	1.64	1.90	0.27

Table 2: Number of users of a cue and their descriptive means for *new* and *accessible* referents.

Table 2 also summarizes the average values for *new* and *accessible* target words and the difference between the two for users of a parameter. At least for tonal onglide, Delta F0 and the duration measures, these differences likely have a perceptual reality [18],[19], indicating that they are not just the result of random fluctuations but may be used by the listener to distinguish *new* from *accessible* referents.

To query individual differences and similarities, we ran a cluster analysis on the random slopes of the models reported in 3.1. Following [20], a four-cluster solution is deemed the most appropriate as the one with the highest average silhouette width. However, one cluster consisted of just one speaker so we will report on only three clusters. Two other speakers were excluded from this analysis, one because they produced only phrase-final *new* referents so that the effect of newness on duration could not be estimated and another because the recording was clipped and thus unfit for mass measures.

Figure 2 shows the average random slope values and standard errors per cluster for every parameter. A positive value indicates that this parameter increases from the *accessible* referent to the *new* one for this

cluster, negative values indicate that parameter values are lower in *accessible* than in *new* referents.

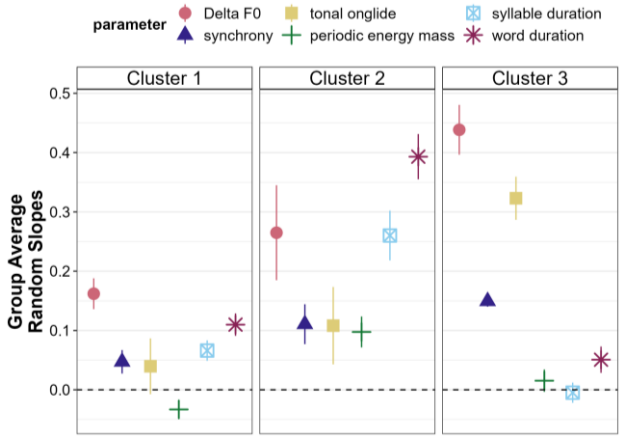


Figure 2: Average random slopes in clusters (speaker groups) with standard errors.

Cluster 1 (11 speakers) includes speakers that produce *new* as opposed to *accessible* referents with higher values in almost all parameters, except for mass. Members of this cluster rely especially on Delta F0, but compared to the other Clusters appear to mark the *new-accessible* contrast relatively weakly. Cluster 2 (4 speakers) is defined by a clearer usage of all surveyed parameters to mark *newness*, particularly word duration. Finally, Cluster 3 (14 speakers) relies on F0-based parameters, especially the ones related to scaling, but less on duration and mass. What is also evident from this plot is that no speaker group behaves exactly as predicted by the overall results in Section 3.1. While Cluster 3 somewhat mirrors the general tendency, Clusters 1 and 2 prefer parameters that are not reliably predicted by information status as per the Bayesian models.

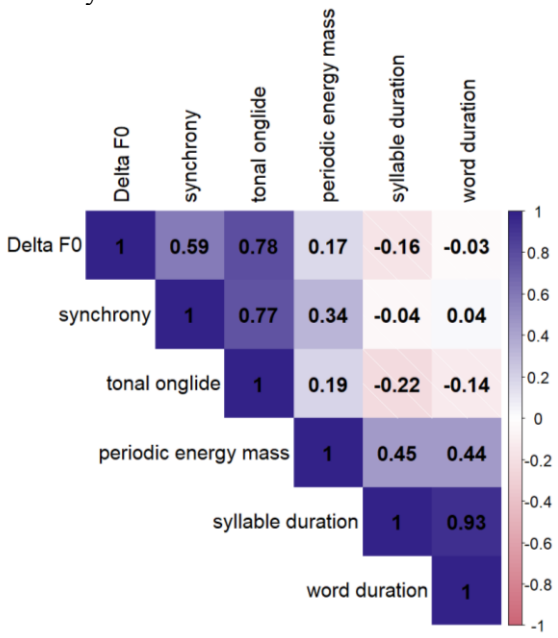


Figure 3: Correlation matrix of random slopes.

We also computed correlations between the random slopes of each model. As evident in Figure 3, tonal onglide correlates fairly strongly with Delta F0 and synchrony, which is unsurprising given that those parameters are meant to capture the shape of the F0 contour. Similarly, syllable duration and word duration are strongly correlated. However, there are no correlations between duration parameters and any of the F0-based cues, indicating that speakers using F0-based correlates tend to not modify duration as a cue to *newness* and vice versa.

4. DISCUSSION

Considering the overall and individual results together, averaging seems to capture the behavior of most individuals quite well, which is reassuring for standard empirical practices. Prosodic cues to prominence, e.g., increased tonal onglide, synchrony and Delta F0, are reliably predicted by *newness* in regression models and are used by the majority of speakers. Other cues, which fail to reach the set reliability thresholds, are concomitantly used by fewer speakers. However, an individual analysis revealed other viable strategies to mark *newness* via prosodic prominence employed by some speakers. For instance, 4 speakers (Cluster 2 in Section 3.2) rely more heavily on longer duration to mark *new* referents as more prominent than *accessible* ones.

Most speakers (18 out of 30, Clusters 2 and 3) mark information status by prioritizing either durational or F0-based cues, which mirrors the finding in [1] that *listeners* tend to focus on either F0-based or durational (plus semantic-syntactic) variables in their processing of prominence. Nevertheless, four speakers (Cluster 2) use all cues investigated here and thus encode information status maximally redundantly and in a cumulative fashion. Thus, redundancy in prominence marking is for the most part compensatory in nature, since it stems from individuals prioritizing different cues serving the same function, but this does not mean that most speakers use a specific prominence cue exclusively.

The robustness of the prosodic encoding of information status has implications for perception: The more redundantly a speaker marks *given* and *new* information in production, the more easily the listener can decode the message [4]. However, it is unclear how the listener deals with contradictory prominence cues, e.g., when a speaker marks *new* referents as more prominent than *accessible* ones via the F0 contour, but at the same time *accessible* referents are longer and thus more prominent in terms of durational cues than *new* ones (or vice versa). A corresponding perception study will reveal the effectiveness of different prominence marking strategies.

5. ACKNOWLEDGEMENTS

We thank Tiffany Zukas for her invaluable help with picture creation and data annotation. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the Collaborative Research Center 1252 “Prominence in Language” (Project-ID 281511265) and project RO 6767/1-1 (Walter Benjamin program).

6. REFERENCES

[1] Baumann, S., Winter, B. 2018. What makes a word prominent? Predicting untrained German listeners’ perceptual judgments. *Journal of Phonetics* 70, 20–38.

[2] Roessig, S., Winter, B., Mücke, D. 2022. Tracing the phonetic space of prosodic focus marking. *Frontiers in Artificial Intelligence* 5.

[3] Winter, B. 2014. Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays* 36, 960–967.

[4] Cangemi, F., Krüger, M., Grice, M. 2015. Listener-specific perception of speaker-specific productions in intonation. In: Fuchs, S., Pape, D., Petrone, C., Perrier, P. (eds.), *Individual differences in speech production and perception*. Peter Lang, 123–145.

[5] Ouyang, I. C., Kaiser, E. 2015. Individual differences in the prosodic encoding of informativity. In: Fuchs, S., Pape, D., Petrone, C., Perrier, P. (eds.), *Individual Differences in Speech Production and Perception*. Peter Lang, 147-189.

[6] Kim, J. 2019. Individual differences in the production of prosodic boundaries in American English. *The Journal of the Acoustical Society of America* 145, 1933.

[7] Baumann, S., Becker, J., Grice, M., Mücke. D. 2007. Tonal and articulatory marking of focus in German. *Proc. 16th ICPhS* Saarbrücken, 1029–1032.

[8] Baumann, S., Grice, M. 2006. The intonation of accessibility. *Journal of Pragmatics* 38, 1636-1657.

[9] Féry, C., Kügler, F. 2008. Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics* 36, 680-703.

[10] Baumann, S., Riester, A. 2013. Coreference, lexical givenness and prosody in German. In: Hartmann, J., Radó, J., Winkler, S. (eds.), *Lingua* 136, 16–37.

[11] Albert, A., Cangemi, F., Grice, M. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. *Proc. 9th International Conference on Speech Prosody*. Poznań, 804-808.

[12] Ritter, S., Grice, M. 2015. The Role of Tonal Onglides in German Nuclear Pitch Accents. *Language and Speech* 58, 114–128.

[13] Kügler, F., Baumann, S., Röhr, C. T. 2022. Deutsche Intonation, Modellierung und Annotation (DIMA) – Richtlinien zur prosodischen Annotation des Deutschen. In: Schwarze, C., Grawunder, S. (eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion*. Narr, 23–54.

[14] Barnes, J., Brugos, A., Veilleux, N. Shattuck-Hufnagel, S. 2021. On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics* 85, 101020.

[15] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[16] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software* 76, 1–32.

[17] Bürkner, P.-C. 2018. Advanced Bayesian multilevel modeling with the R Package brms. *The R Journal* 10, 395–411.

[18] ‘t Hart, J. 1981. Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America* 69, 811-821.

[19] Lehiste, I. 1970. *Suprasegmentals*. M.I.T. Press.

[20] Levshina, N. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins.

PERCEPTUAL ASSIMILATION MODEL FOR SUPRASEGMENTALS PREDICTS CANTONESE TONE SEQUENCE RECALL IN PUNJABI MINORITIES

Doris Yu[1], Veronica Ka Wai Lai[2], William Choi[1]

The University of Hong Kong[1], University of Manitoba[2]
dorisyuu@connect.hku.hk, veronica.lai@umanitoba.ca, willchoi@hku.hk

ABSTRACT

Research has suggested that the Perceptual Assimilation Model for Suprasegmentals (PAM-S) predicts non-native tone discrimination. This study investigates whether PAM-S applies to sequence recall. Twenty-two Punjabi speakers completed a Cantonese tone perceptual assimilation task and a Cantonese tone sequence recall task. In the perceptual assimilation task, Punjabi speakers assimilated the Cantonese level tone contrast (T1-T3) and the contour tone contrast (T4-T5) as single category (SC) and uncategorized-categorized pair without overlap category (UC_{no}), respectively. In the sequence recall task, they recalled T4-T5 (UC_{no}) more accurately than T1-T3 (SC) across different sequence lengths. The findings are consistent with the prediction of PAM-S on sequence recall. Moreover, memory load does not seem to limit its applicability.

Keywords: The Perceptual assimilation model for suprasegmentals, sequence recall, Cantonese tone.

1. INTRODUCTION

About 8% of the population in Hong Kong are non-Chinese ethnic minorities with diverse language backgrounds (e.g., Hindi, Punjabi, and Urdu) [1]. Research has found that non-Chinese ethnic minorities have difficulties learning Cantonese [2-4], particularly in Cantonese tone perception [5]. With the increase in learning needs in Cantonese as a second language [6], however, the perceptual patterns across groups with various native language backgrounds have not been fully explored. Thus, there is a dearth of research examining the perceptual characteristics of various non-Chinese ethnic minorities who learn Cantonese as their second language (L2). Among the native languages non-Chinese ethnic minorities speak, Punjabi is a tone language whereas Urdu and Hindi are not. Additionally, the question of how the native tone system of Punjabi speakers influences their phonological perception has yet to be explored [7]. From the theoretical perspective, previous studies have not examined whether PAM-S applies to the tone sequence recall task which taps into phonological perception [8-12]. Hence, we examined how Punjabi listeners assimilate Cantonese tones and recall Cantonese tone sequences.

PAM-S predicts the discrimination patterns of non-native suprasegmentals based on the assimilation patterns [13, 14]. For perceptual assimilation, listeners' native prosodic systems may or may not categorize non-native suprasegmentals. The assimilation patterns can be classified into five types: assimilation of two non-native categories into two native categories (two-category assimilation, TC), assimilation of two non-native categories into one native category (single-category assimilation, SC), assimilation of two non-native categories into one native category with different assimilation goodness (category goodness difference in assimilation, CG), assimilation of one categorized non-native category and one uncategorized non-native category (uncategorized-categorized, UC), and assimilation of two uncategorized non-native categories (uncategorized-uncategorized, UU) [13, 14].

Regarding the discriminability of the assimilation pairs, PAM-S predicts that non-native speakers most easily discriminate TC; easily discriminate UC with no native category overlap (UC_{no}) and UU with no native category overlap (UU_{no}); and fairly easily discriminate CG [13, 14]. Additionally, non-native speakers poorly discriminate SC, UC with native category overlap (UC_{o}), UU with partial native category overlap (UU_{o}), and UU with the same category overlap (UU_{s}) [13, 14].

Previous studies put forward that the AX tone discrimination task cannot probe into phonological processing even with memory load manipulations [15, 16]. In the AX tone discrimination task, listeners have to store the first stimulus in short-term memory, and use phonetic details to judge whether the second stimulus is the same as the first stimulus [15, 16]. In previous studies, listeners' discrimination accuracy was not influenced by short (500 ms) or long (2000 ms) interstimulus intervals [15, 16]. These findings suggest that AX tone discrimination task under short and long memory loads can uphold the PAM-S at the phonetic level but not at the phonological level [15,16]. In this regard, the tone discrimination task

under memory load manipulations limits the applicability of PAM-S in phonological perception. To extend the applicability of PAM-S at the phonological level, it is necessary to adopt a more robust test of the phonological mode of perception.

Sequence recall task is a potent tool to probe into the phonological mode of perception. In the familiarization session, listeners temporarily encode nonnative speech sounds into short-term phonological representations [8, 12]. In the test session, listeners have to use the phonological representations to efficiently encode and retain the information in short-term memory for recalling the sequence. In addition, using phonetic variability in sequence recall task can prevent participants from using acoustic features, since stimuli produced by different talkers pose more difficulties for listeners to employ acoustic features and listeners thus have to use phonological representations to efficiently encode the information [8-12]. Therefore, the sequence recall task with phonetic variability is a robust short-term memory test for phonological perception.

Cantonese and Punjabi are both tone languages. Punjabi comprises three lexical tones: high falling tone, level tone, and low rising tone [17-19]. Likewise, Cantonese consists of six lexical tones: high level tone (T1), high rising tone (T2), mid-level tone (T3), low falling tone (T4), low rising tone (T5), and low level tone (T6) [20]. Based on the similarities and distinctions between Cantonese and Punjabi lexical tone systems, we formulated a hypothesis on how Punjabi listeners would assimilate Cantonese tones into the Punjabi tone system (see Table 1). To avoid prolonged testing and exhaustion effects, we only employed T1-T3 and T4-T5 pairs in the tone sequence recall task. Since T1, T3, and the Punjabi level tone are level tones, we expected that T1 and T3 would be assimilated as the Punjabi level tone. As such, the T1-T3 contrast is an SC pair. On the other hand, T4 and T5 are contour tones with different pitch contours, i.e., low falling contour and low rising pitch contour, respectively. As the tonal patterns of T5 and the Punjabi low rising tone are similar, we assumed that T5 would be assimilated as the Punjabi low rising tone. Given that no low falling tonal pattern was found in the Punjabi tone system, we expected that T4 would not be assimilated as a single Punjabi tone, but rather it would be equally assimilated as the Punjabi level tone and high falling tone because T4 has a similar tone onset to the Punjabi level tone and a similar tone contour to the high falling tone. For this reason, we hypothesized that T4 and T5 were assimilated to a UC_{no} pair. In short, we suggested that the T1-T3 pair would represent SC pair and T4-T5 would represent as UC_{no} pair (see Table 2), and both tone pairs would yield empirically testable predictions of PAM-S.

If PAM-S fits well for phonological perception in the sequence recall task, the recall should be better for UC_{no} than for SC. Otherwise, Punjabi listeners might not recall UC_{no} better than SC at the phonological level in the sequence recall task.

Tone	Punjabi low-rising	Punjabi level	Punjabi high-falling	Uncate-gorized	None
T1		+			
T2	+				
T3		++			
T4			+		
T5	++				
T6		+			

Table 1: Hypothesis of Cantonese tone assimilation patterns[1]

Tone	T1	T2	T3	T4	T5
T2	TC				
T3	**SC**	TC			
T4	UC_o	UC_{no}	UC_o		
T5	TC	SC	TC	**UC_{no}**	
T6	CG	TC	SC	UC_o	TC

Table 2: Hypothesis of assimilation pairs

Altogether, this study set out to examine whether PAM-S applies to tone sequence recall. In particular, we investigated (i) how Punjabi listeners assimilate the six Cantonese tones into their native tone system and (ii) whether Punjabi listeners perform better on recalling the UC_{no} pair than the SC pair.

2. METHODS

2.1. Subjects

Twenty-two Punjabi speakers (M_{age} = 22.5 years old, SD = 11.7 years) living in Hong Kong participated in the experiment. All subjects were native Punjabi speakers and grew up in Punjabi-speaking households. They identified Cantonese as their second language. They self-reported their Punjabi (M = 4.6, SD = 0.6) and Cantonese proficiencies (M = 2.3, SD = 0.1) on a 5-point Likert scale (1 = naïve, 5 = native), with Punjabi proficiency significantly higher than that of Cantonese ($t(21) = 9.63$, $p < .001$). Punjabi (M = 4.2, SD = 1.1) was their dominant language for communication while Cantonese was seldom used (M = 2.6, SD = 1.1) on a Likert scale (1 = never, 5 = always), and the difference between them attained statistical significance ($t(21) = 4.9$, $p < .001$). Their average length of residency in Hong Kong was 14.3 years (SD = 7.4 years). All subjects reported having normal hearing.

2.2. Cantonese Tone Assimilation Task

Subjects were visually and audibly presented with three Punjabi words (/tʃá/ *tea*, /tʃã/ *enthusiasm*, and /tʃà/ *peep*) with their corresponding tone symbols. To ensure all subjects were familiar with the three Punjabi tone symbols, they had to successfully match the Punjabi words to the tone symbols before starting the test.

The stimuli were 12 Cantonese words (/fu-T1/, /fu-T2/, /fu-T3/, /fu-T4/, /fu-T5 /, /fu-T6 /, /ji-T1/, /ji-T2/, /ji-T3/, /ji-T4/, /ji-T5/, /ji-T6/) naturally produced by two native Cantonese female adults and two native Cantonese male adults. On each trial of the test phase, a Cantonese word was played once with four options (the three Punjabi tone symbols and a "none" option) presented on the screen. Each subject was required to select a sign and indicate the degree of the heard tone that matched the selected option on a seven-point scale (1 = poor, 7 = perfect). The test phase contained 72 trials (12 Cantonese words × 6 repetitions). The reliability of the assimilation task was satisfactory (Cronbach's α = .8).

2.3. Cantonese Tone Sequence Recall Task

The sequence recall task was adopted from [10], with a new set of stimuli. There were two pairs (/ji-T4/ and /ji-T5/, /ji-T1/, and /ji-T3/) with five sequence lengths in this task. Based on the results of the assimilation task, these two pairs could be defined as a UC_{no} pair (/ji-T4/ and /ji-T5/) and an SC pair (/ji-T1/ and /ji-T3/). The stimuli were naturally produced by the same Cantonese-speaking adults as aforesaid. In the familiarization phase, subjects were able to associate the /ji-T1/ and /ji-T5/ associated with key 1, and the /ji-T3/ and /ji-T4/ with key 2.

On each trial, a sequence of two to six Cantonese words (e.g., /ji-T1/, /ji-T3/, /ji-T1/, /ji-T3/) with 80ms inter-stimulus interval was played once. To reduce subjects' reliance on acoustic memory, each sequence was followed by a pure tone with an interval of 80 ms. The order of genders was counterbalanced across tones in each trial. Each subject was required to replicate the sequence of tones by pressing the associated keys (e.g., 1212 for /ji-T1/, /ji-T3/, /ji-T1/, /ji-T3/). One point was given only when the consecutive answer correctly matched the sequence. Feedback was provided only in the practice phase. The test phase included 60 trials (2 assimilation pairs × 5 sequence lengths × 6 trials per sequence length). The reliability of the sequence recall task was high in the UC_{no} pair (α = .90) and the SC pair (α = .87). To eliminate the impact of a carryover effect, the Cantonese tone sequence recall was conducted before the assimilation task.

3. RESULTS

3.1. Cantonese Tone Assimilation Task

We adopt the categorization criteria set by [13, 14]. If a Cantonese tone matches one of the assimilation options with a percentage higher than chance level (25%), then the Cantonese tone is considered to be assimilated as a Punjabi tone or none category [13, 14]. Additionally, the mean of category goodness should also be higher than five for the sake of stringency.

To explore the assimilation patterns of each Cantonese tone by Punjabi speakers, one-way ANOVAs were utilized on each Cantonese tone with assimilation option as the within-subject factor. For T3, Greenhouse-Geisser adjustment was used to correct for a sphericity violation. The Shapiro-Wilk test indicated that five Cantonese tones did not follow normally distributed patterns. Since one-way ANOVA is robust to deviations from normal distribution [21], we presented the results obtained from it. The results indicated that the main effects of the assimilation options were significant for each Cantonese tone: T1, $F(3, 33) = 19.56$, $p < .001$, $\eta_p^2 = .64$; T2, $F(3, 33) = 15.89$, $p < .001$, $\eta_p^2 = .59$; T3, $F(1.42, 15.62) = 20.93$, $p < .001$, $\eta_p^2 = .66$; T4, $F(3, 33) = 26.45$, $p < .001$, $\eta_p^2 = .71$; T5, $F(3, 33) = 11.71$, $p < .001$, $\eta_p^2 = .52$; T6, $F(3, 33) = 35.15$, $p < .001$, $\eta_p^2 = .76$. Levene's test revealed that the homogeneity of variance assumption was not met for two Cantonese tones. In this respect, we used one-way Welch's ANOVA, which yielded consistent results.

It is evident from the results that T1, T3, and T6 had similar assimilation patterns. The three tones were assimilated more frequently as the Punjabi level tone, $ps < .01$, than to the Punjabi high falling tone, the Punjabi low rising tone, and none category. T2 and T5 also showed similar assimilation patterns. Both tones were assimilated more frequently as the Punjabi low rising tone, $ps < .05$, compared with the Punjabi level tone, the Punjabi high falling tone, and none category. T4 and T5 had different assimilation patterns. T4 was assimilated more frequently as the Punjabi high falling and the level tone, $ps < .01$, than as the Punjabi low rising tone and none category. Unlike T4, T5 was assimilated more frequently as the Punjabi low rising tone, $ps < .05$, than to the Punjabi level tone, the Punjabi high falling tone, and none category.

The results were mainly consistent with our assimilation hypothesis. Of interest to our study is that, the T1-T3 pair is assimilated as SC, and the T4-T5 pair is assimilated as UC_{no}. Figure 1 shows the assimilation percentages of four Cantonese tones, and

Table 3 indicates the assimilation patterns obtained from pairwise comparisons.

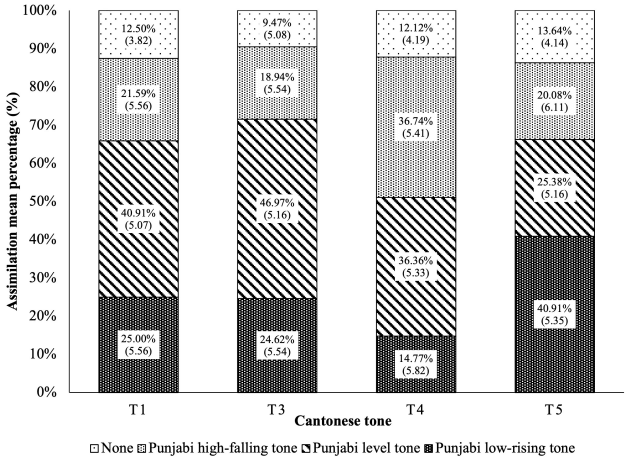


Figure 1: Assimilation percentages of Cantonese T1, T3, T4, and T5 and their means of category-goodness in parentheses

Tone	Punjabi low-rising	Punjabi level	Punjabi high-falling	Uncate-gorized	None
T1		++			
T3		++			
T4		+	+	++	
T5	++				

Table 3: Assimilation patterns of Cantonese T1, T3, T4, and T5

3.2. Cantonese Tone Sequence Recall Task

A two-way ANOVA was conducted on mean accuracies with assimilation pairs (UC_{no} and SC) and sequence length (two, three, four, five, six) being the within-subject factors. As Levene's test showed a violation of the assumption of homogeneity of variance, we performed a square root transformation on the mean accuracy to eliminate the violation. Then, ANOVA was done on the transformed percentage accuracy. The Shapiro-Wilk test of normality showed that the transformed accuracy was normally distributed. Mauchly's test indicated that the assumption of sphericity has been met, $\chi^2 = 6.65$, $df = 9$, $p = 0.844$. The results showed a significant main effect of assimilation pair, $F(1, 21) = 48.84$, $p < .001$, $\eta_p^2 = .70$. Post-hoc Bonferroni tests showed the mean percentage of UC_{no} pair was significantly higher than the SC pair ($p < .001$). We also found the significant main effect of sequence length, $F(4, 84) = 35.21$, $p < .001$, $\eta_p^2 = .63$. Specifically, pairwise comparisons with Bonferroni adjustments showed that the mean accuracy of two-tone sequence length was significantly higher than the three-, four-, five-, and six-tone sequence lengths ($ps < .05$), and three-tone sequence length was significantly higher than the four-, five-, and six-tone sequence lengths ($ps < .01$), but there was no significant difference in four-tone sequence length compared to the five- and six-tone sequence lengths ($ps \geq .074$). The interaction between assimilation pair and sequence length was not statistically significant, $p = .656$. Figure 2 shows the mean accuracy (not transformed) of two assimilation pairs in the sequence recall task.

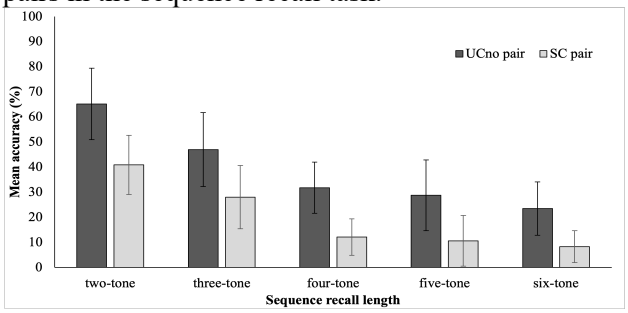


Figure 2: Mean accuracy of the UCno and SC tone pairs in the sequence recall task.

Additionally, we conducted supplementary analyses with outlier removal of the length of residency and found the results were consistent with the assimilation patterns and sequence recall performance.

4. DISCUSSION

The sequence recall performance was consistent with PAM-S' prediction on tone discrimination. In PAM-S, the discriminability of the UC_{no} pair is higher than the SC pair at low and high perceptual levels [13-15]. In this study, Punjabi listeners assimilated Cantonese T1-T3 as an SC pair and T4-T5 as a UC_{no} pair. In the sequence recall task, they exhibited a better performance in recalling T4-T5 (UC_{no}) than T1-T3 (SC), aligning with the prediction of PAM-S. The present study extends the predictive potential of PAM-S for predicting SC and UC_{no} pairs in L2 tone sequence recall. In particular, memory load does not limit the applicability of PAM-S in the tone sequence recall task.

5. CONCLUSION

To sum up, our findings reflect the applicability of PAM-S on tone sequence recall. Particularly, Punjabi speakers recalled the UC_{no} pair significantly better than the SC pair. Given that we only included UC_{no} and SC pairs, further studies with a whole set of assimilation pairs can fully test the applicability of PAM-S to sequence recall. With an increasing number of non-Chinese students with diverse language backgrounds in Hong Kong, more research is needed to assess their performance in perceptual learning of lexical tone in order to test the theories of language perception and inform pedagogical practice.

6. REFERENCES

[1] Census and Statistics Department. 2016. Thematic report: Ethnic minorities. Hong Kong: Census and Statistics Department. https://www.censtatd.gov.hk/en/data/stat_report/product/B1120100/att/B11201002016XXXXB0100.pdf.

[2] Li, D. C. S., Chuk, J. Y. P. 2015. South Asian students' needs for Cantonese and written Chinese in Hong Kong: A linguistic study. *International Journal of Multilingualism* 12(2), 210–224.

[3] Loh, E. K. Y., Tse, S. K. 2012. An investigation of school-based curriculum of effective Chinese character learning for Non-Chinese speaking kindergarteners. *Han-Character and Classical Writing Language Education* 28, 171–195.

[4] Lee, S.-l. 2019. Teaching and learning Cantonese as a second language: Learners' attitudes and learning hurdles. In: Wakefield, J. C. (ed), *Cantonese as a Second Language: Issues, Experiences and Suggestions for Teaching and Learning*. Routledge, 47–66.

[5] Qin, Z., Mok, P. P. K. 2014. Discrimination of Cantonese tones by speakers of tone and non-tone languages. *Kansas Working Papers in Linguistics* 34, 26–42.

[6] Lee, S.-l. 2020. Learning needs of Cantonese as a second language in Hong Kong tertiary education. *Electronic Journal of Foreign Language Teaching* 17(2), 460–478.

[7] Yu, A. C. L., Lee, C. W. T., Lan, C., Mok, P. P. K. 2022. A new system of Cantonese tones? Tone perception and production in Hong Kong South Asian Cantonese. *Language and Speech* 65(3), 625–649.

[8] Dupoux, E., Peperkamp, S., Sebastián-Gallés, N. 2001. A robust method to study stress "deafness". *The Journal of the Acoustical Society of America* 110(3), 1606–1618.

[9] Dupoux, E., Sebastián-Gallés, N., Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': The case of French learners of Spanish. *Cognition* 106(2), 682–706.

[10] Choi, W. 2020. The selectivity of musical advantage: Musicians exhibit perceptual advantage for some but not all Cantonese tones. *Music Perception* 37(5), 423–434.

[11] Kim, H., Tremblay, A. 2021. Korean listeners' processing of suprasegmental lexical contrasts in Korean and English: A cue-based transfer approach. *Journal of Phonetics* 87, 101059.

[12] Choi, W., Chiu, M. M. 2022. Why aren't all Cantonese tones equally confusing to English listeners? *Language and Speech*, 00238309221139789.

[13] So, C. K., Best, C. T. 2014. Phonetic influence on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition* 36(2), 195–221.

[14] So, C. K., Best, C. T. 2010. Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech* 53(2), 273–293.

[15] Chen, J., Best, C. T., Antoniou, M., Kasisopa, B. 2019. Cognitive factors in perception of Thai tones by naïve Mandarin listeners. In: *Proc. 19th ICPhS Melbourne*, 1684–1688.

[16] Chen, J., Antoniou, M., Best, C. T. 2023. Phonological and phonetic contributions to perception of non-native lexical tones by tone language listeners: Effects of memory load and stimulus variability. *Journal of Phonetics* 96, 101199.

[17] Evans, J., Yeh, W. C., Kulkarni, R. 2018. Acoustics of tone in Indian Punjabi. *Transactions of the Philological Society* 116(3), 509–528.

[18] Nara, K. 2016. Acoustic analysis of Punjabi stress and tone (Doabi dialect). *The Journal of the Acoustical Society of America* 139(4), 2215–2215.

[19] Singh, A., Pandey, D., Agrawal, S. S. 2015. Analysis of Punjabi tonemes. In: *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1694–1697.

[20] Bauer, R. S., Benedict, P. K. 1997. *Modern Cantonese Phonology*. Berlin: Walter de Gruyter.

[21] Lix, L. M., Keselman, J. C., Keselman, H. J. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research* 66(4), 579–619.

[1] ++ = firmly assimilated, + = moderately assimilated.

[2] The error bars show 95% confidence intervals.

The Role of Prosody and Beat Gesture in Enhancing Memory for Discourse Information in Mandarin

Mengzhu Yan[1], Sasha Calhoun[2]
[1] Huazhong University of Science and Technology, China [2]Victoria University of Wellington, New Zealand
mengzhu_yan@hust.edu.cn, sasha.calhoun@vuw.ac.nz

ABSTRACT

Both auditory (e.g., prosody) and visual cues (e.g., beat gesture) available in communication are important for listeners to comprehend discourse, given that speech is multimodal. While a vast amount of research has been devoted to investigating the role of prosody in discourse comprehension, it is surprising that relatively little research has been conducted to uncover how visual cues interact with auditory cues given the importance of visual cues in speech processing. This paper examines the roles of prosodic prominence and beat gesture in the memory for discourse information in Mandarin. A dominant role of prosody but not beat gesture was found in facilitating the memory for discourse information. This study contributes significantly to our limited knowledge of multimodal comprehension of focus, and how cues from multilevel sources are integrated.

Keywords: prosody, beat gesture, discourse information, Mandarin

1. INTRODUCTION

It is well-established that discourse information is encoded through various means, including verbal (e.g., lexical and morpho-syntactic) and non-verbal cues (e.g., prosody and gestures) (e.g., [1-7]). Among the non-verbal cues, both prosody from the auditory domain and the co-speech beat gesture from the visual domain have been tested as facilitating listeners' processing of discourse information (e.g., focused words and their alternatives that are relevant for the interpretation of the discourse) (e.g., [8-15]). In speech communication, listeners must integrate information from multiple sources (auditory and visual) to facilitate the processing of discourse, given the communication is usually multimodal. While a large body of previous literature has primarily focused on prosody and Germanic languages, much less attention has been directed to visual cues (e.g., beat gesture) and Mandarin. This paper investigates whether and how the integration of prosody and beat gesture affects memory for discourse information in Mandarin.

Prosody is one of the most important markers to signal focus in many languages, including English and Mandarin (e.g., [1, 7, 16, 17]). Unlike English which marks focus by pitch accenting (e.g., H* or L+H*), Mandarin signals focus prosodically through the expansion of pitch range while maintaining the local F0 contour of each syllable to indicate lexical tones; for example, *Mexican* in the top picture in Figure 1 (the Mandarin sentence and its English translation are given in (1) and (2)) is marked with prosodic prominence, and has an expanded pitch range than that without prosodic prominence in the bottom picture. Similarly, *dessert* in the bottom picture has a larger pitch range than that in the top.

(1) **Chinese:** 但这位评论家因为得了流感，只去了墨西哥(word 1)餐厅，并对那里的甜品(word 2)给予了好评。
English: But he caught the flu, so he only visited the Mexican (word 1) restaurant, where he gave the desserts (word 2) a favorable review.

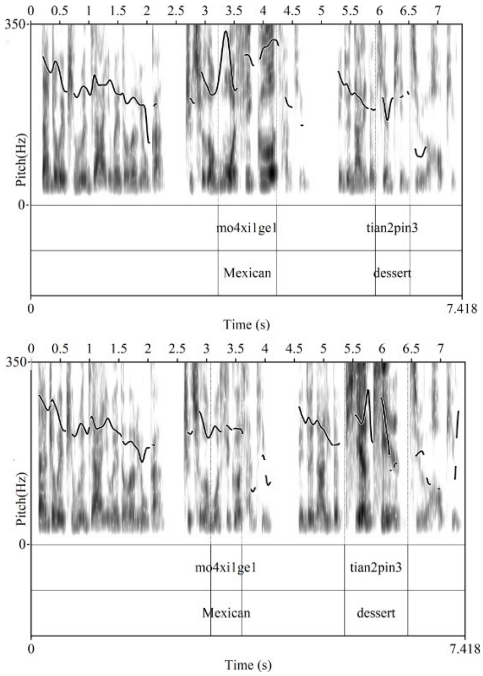


Figure 1: Pitch tracks of sentences with prosodic prominence on the first (top) and second word (bottom)

Despite the realizational differences, prosodic prominence has been shown to be functionally equivalent in the comprehension of focus in English and Mandarin in recent comparative studies [10, 12]. Increasing but yet still limited evidence has shown that prosodic prominence facilitates the immediate activation and memory (with a delay of 10 canonical SVO sentences, approximately 40 seconds) for focused words and focus alternatives in Mandarin [13, 18]. However, it remains unknown whether this facilitative effect on memory for discourse information could last longer (e.g., after 35 long discourses like those shown in Table 1, around 15 mins). Further, most research on the role of prosody in Mandarin has focused on the auditory modality only (see e.g., [12, 13]). Considering that communication often involves both visual and auditory input, it is surprising that little research has shown how visual cues interact with auditory cues in searching for focus, especially in Mandarin.

The seminal McGurk effect [19] has acknowledged the importance of visual input (i.e., lips) to sound cognition, and a growing body of research has established evidence on the key role of visual cues (beat gestures, head nods, eyebrow movements, hand gestures, see [23]) in language comprehension [15, 20-24]. It has been shown that visual cues enhance the perceived prominence of the word, improve the accuracy rate of identifying where the focus is and facilitate interpretation and memory of spoken discourse. Beat gesture, as one of the common types of visual cues, co-occurs with speech and is frequently used to give importance or prominence to discourse information in the visual domain in communication, functioning as a "gestural yellow highlighter" [25], which works in a similar way to prosodic prominence in the auditory domain to focus. More recently, [15] has shown a complex relationship between prosody and beat gesture in English that beat gesture modulates the role of prosody in comprehension, i.e., when the experiment had beat gesture as a condition (Experiment 1), listeners used prosodic cues to remember discourse information only when the information was said with beat gesture; while when no gesture condition was included in the experiment (Experiment 2), prosody was effective in facilitating memory for discourse information. To the best of our knowledge, the effect of beat gesture in long-term memory for discourse information is yet unknown in Mandarin and the present study sets out to test this.

2. THE EXPERIMENT

The experiment aims to investigate whether and how prosody and beat gesture affect the comprehension of discourse information by native Mandarin speakers, adopting the task employed in [14, 15]. In the task, participants watched a sequence of 35 discourses (see Table 1) with critical words marked with prosodic prominence or no prominence, and/or beat gesture, and then were asked about the content of the discourses in the later recognition memory task.

2.1 Participants

Seventy-two native Mandarin speakers aged 17-24 (mean = 19.7, SD = 1.88; 54 females and 18 males) from the student population from Huazhong University of Science and Technology participated in the study for a small gift. Data from two were discarded due to technical errors, leaving 70 participants for the final analysis. The participants reported no reading or hearing difficulties.

2.2 Materials

The materials included 48 test discourses and 22 filler discourses, with each containing a context sentence that introduces two sets of contrastive alternatives (e.g. *Mexican/ Italian* and *mains/ desserts* in Table 1). Most of the test materials were adopted and translated from those used in [14, 15]. The context sentence was followed by a continuation sentence that described a fact using one word from each of the contrastive sets (e.g., *Mexican* and *desserts*). The order of the two critical items that appeared in the context sentence was counterbalanced to avoid potential bias.

For each continuation sentence, 16 versions were created varying by the presence/absence of prosody and/or beat gesture on the first and second critical items, i.e. 2 beat gesture conditions (no beat gesture, beat gesture) * 2 prosody conditions (no prosodic prominence, prosodic prominence) on the first word, * 2 beat gesture conditions * 2 prosody conditions on the second word. The examples of prosody are shown in Figure 1 and beat gesture in Figure 2. The context and continuation sentences were produced by a female native Mandarin speaker and recorded using a SONY HXR-NX100 video camera in a sound-attenuated room. A teleprompter was placed to prompt the speaker with the visual text which marks when to produce prosodic prominence and

use beat gesture. The beat gestures were performed to naturally align with the critical words. Adobe Premiere was subsequently used to edit the frame of the videos, blur the face of the speaker, and export video clips in AVI files. The beat gestures and prosodic prominences were double-checked by five native speakers, ensuring they were produced as intended. The soundtrack was extracted and segmented using *Montreal Forced Aligner* [26]. The acoustic measures (mean F0, duration, and intensity) of the two critical items were extracted using *ProsodyPro* [27] in *Praat* [28] and analyzed in separate linear mixed-effects models using *lme4* [29] in *R* [30], showing that prosodic prominence was produced as intended.

Context	最近，一家新的墨西哥餐厅和一家新的意大利餐厅开业。双方都在等这位美食评论家对他们的主菜和甜品进行评价。 A new Mexican and a new Italian restaurant had recently opened in the city. Both were waiting to hear the comments from the food critic on their mains and desserts.
Continuation	但这位评论家因为得了流感，只去了墨西哥餐厅，并对那里的甜品给予了好评。 But he caught the flu, so he only visited the Mexican restaurant, where he gave the desserts a favorable review.
Test question	最近，一家新的墨西哥餐厅和一家新的意大利餐厅开业。双方都在等这位美食评论家对他们的主菜和甜品进行评价。但这位评论家因为得了流感，只去了_____（1. A. 墨西哥; B. 意大利）餐厅，并对那里的_____（2. A. 主菜; B. 甜品）给予了好评。 A new Mexican and a new Italian restaurant had recently opened in the city. Both were waiting to hear the comments from the food critic on their mains and desserts. But he caught the flu, so he only visited the _____ (1. A. Mexican; B. Italian) restaurant, where he gave the _____ (2. A. mains; B. desserts) a favorable review.

Table 1: Examples of test materials

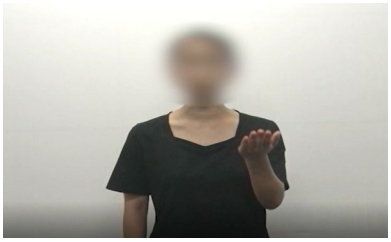


Figure 2: Example of beat gesture

A total of 768 test stimuli (48 items * 16 versions) were constructed using a Latin square design, rendering 16 lists of 48 stimuli with three discourses appearing in the same condition in each list. Each participant was shown only one list. A further 22 filler discourses, following the same structure as the test stimuli, except that the filler discourses had prosodic prominence and beat gesture on non-critical items in the continuation sentences, were added to each list, totaling 70 discourses. The addition of the filler discourses was to prevent participants from figuring out experimental manipulations were exclusively related to the critical items. The 70 discourses were divided into two blocks, with 35 discourses in each block.

2.3 Procedure

The experiment was administered using E-prime 3.0 [31]. Participants were seated at a computer in a quiet lab. They first read written instructions on the computer screen, which informed them to carefully watch the videos and make choices about the content of the videos. Once the participants confirmed that they had understood the instructions, the experiment entered the practice of four discourses followed by a recognition memory test of the content of the discourse (see *test question* in Table 1). The participants could only enter the main experiment if their accuracy was 75% (6/8) or higher, otherwise, they would be asked to redo the practice. In the first block of the main experiment, participants watched a sequence of 35 videos, and then were asked about what happened in the discourses in the test phase. The order of the discourses was randomized within each block for each participant to reduce potential response bias. In the test phase, the same discourses were visually shown in the same order in which it was audio-visually displayed to control for the delay between the display of the video and the presentation of the test question for each trial. Participants pressed "A" or "B" to indicate their choice for the first critical item and then for the second critical item.

Participants could move to the second block of 35 trials after a break of two minutes. The order of the two blocks was counterbalanced. Participants completed an online background questionnaire by scanning a QR code at the end of the experiment. The entire session lasted approximately 40 minutes.

3. RESULTS

A logistic mixed-effects model, a form of generalized linear mixed effects model (GLMER) (family: binomial), was fit using the *lmerTest* package [32] in R to the 6720 responses (48 items *

2 critical words * 70 participants) for the analysis of how the accuracy (coded binomially as 0 [incorrect] or 1 [correct]) of response was predicted by prosodic prominence and/or beat gesture. The initial model included interactions between prosodic prominence, beat gesture and word order (whether it was the first or the second critical word), simple effect of the cTrial (centered position of the trial in the experiments), block order (whether the trial was in the first block or second), as well as a maximal random structure, following [29]. The initial model was reduced by eliminating non-significant effects by model comparison.

The final model, shown in Table 2, indicates a significant effect of prosody, showing that the recognition memory for facts about critical items was better when the words were marked with prosodic prominence, regardless of the presence/absence of beat gesture. The significance of *block order* also shows that participants' memory for the discourse information in the second block was better than in the first block. Figure 3 shows the predicted accuracy in terms of the proportion of correct responses affected by prosody. The effect of gesture was not significant, nor was its interaction with other key factors.

Model: Prosody+BlockOrder+(1+cTrial|Participant) +(1|Item)

	Chisq	Df	P
Prosody	18.27	1	<0.001
BlockOrder	4.21	1	0.04

Table 2: ANOVA table of the final model

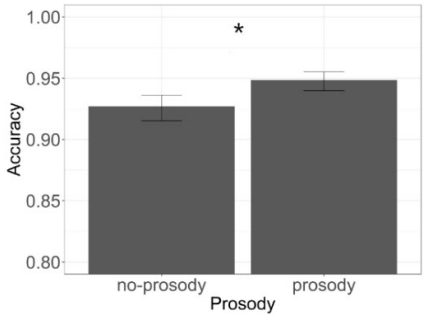


Figure 3: Estimated accuracy of the recognition memory test influenced by prosody, with confidence intervals.

4. DISCUSSION AND CONCLUSION

This research examined whether prosodic prominence and/or beat gesture has a persisting effect on the memory for discourse information in Mandarin. The results have shown a significant role of prosody but not the effect of beat gesture, suggesting differing effectiveness of cues from the auditory and visual domains, although the two types of cues are both believed to highlight important information. We elaborate on the role of prosody and beat gestures below.

Regarding the role of prosody, it has been shown to play an important role in the interpretation of focus location (e.g., [33]), lexical activation and short-term memory of focused words and their alternatives [13, 18] in Mandarin. This present study has added further evidence of the long-lasting or enduring effect of prosody on the memory for discourse information that was mentioned quite a long time ago (approximately 15 minutes), consistent with what has been reported in other languages (e.g., [14]). This beneficial mnemonic effect of prosody is first shown in Mandarin, providing additional support for "equivalent functionality" of prosodic prominence in long-term memory cross-linguistically.

However, beat gesture did not seem to affect the recognition memory for discourse information in Mandarin. One possibility could be the ceiling effect given the average accuracy was high, approximately 90.5%, so there was no room for improvement when adding a gesture cue. One may also argue that it might be possible for participants to learn that gesture was not informative, as they needed to recall the same discourse information whether or not it was gestured; but this is less likely, as prosody enhanced the memory, although participants needed to recall the same information regardless of prosody. Therefore, it is more likely that the gesture effect, if there was one, has been weakened or obscured by the strong prosodic cue in the discourse. Prosodic prominence has been shown in previous studies to have a dominant role in focus processing in Mandarin, and other types of cues (e.g., syntactic clefting) played a rather little role when prosodic prominence was present (e.g., [12, 13, 33]). Further research may use materials that only manipulate the presence or absence of beat gesture with the absence of strong prosodic prominence to see whether beat gesture was effective under no dominant influence of prosody.

Also, given the cross-linguistic differences in the relative role of prosody and other cues in language comprehension (e.g., [7]), it would be worth investigating using the current design languages in which prosody plays little role or visual cues play a bigger role than auditory in language processing to see whether beat gesture would be shown to have a larger influence.

5. ACKNOWLEDGEMENTS

This work is supported by The National Social Science Fund of China (21CYY014). Thanks to Siyuan Fang, Zhangwen Xiong, Weiwei Cen, Jiafei Deng and Fengming Zeng for assisting with materials creation and data collection and for all the participants for taking apart.

6. REFERENCES

[1] Breen, M., Fedorenko, E., Wagner M., Gibson, E. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes* 25(7–9), 1044–1098.

[2] Calhoun, S. 2010. The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language.* 86(1), 1–42.

[3] Féry, C., Ishihara, S. 2016. *The Oxford Handbook of Information Structure*. Oxford University Press.

[4] Féry, c., Krifka, M. 2008. Information structure: Notional distinctions, ways of expression. In: Sterkenburg, P. v., (ed), *Unity and Diversity of Languages*. John Benjamins, 123–136.

[5] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57(3), 396–414.

[6] Krifka, M. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55(3–4), 243–276.

[7] Kügler, F., Calhoun, S. 2020. Prosodic encoding of information structure: A typological perspective. In: Gussenhoven, C., Chen, A., (eds), *The Oxford Handbook of Language Prosody*. Oxford University Press, 454–467.

[8] Braun, B., Tagliapietra, L. 2010. The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes* 25(7–9), 1024–1043.

[9] Husband, E. M., Ferreira, F. 2016. The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience* 31(2), 217–235.

[10] Ip, M. H. K., Cutler, A. 2020. Universals of listening: Equivalent prosodic entrainment in tone and non-tone languages. *Cognition* 202, 104311.

[11] Kember, H., Choi, J., Yu, J., Cutler, A. 2019. The processing of linguistic prominence. *Language and Speech* 64(2), 413-436.

[12] Yan, M., Calhoun, S. 2020. Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology* 11(1), 17.

[13] Yan, M., Calhoun, S. 2019. Priming effects of focus in Mandarin Chinese. *Frontiers in Psychology* 10, 1985.

[14] Fraundorf, S. H., Watson, D. G., Benjamin, A. S. 2010. Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language* 63(3), 367–386.

[15] Morett, L. M., Fraundorf, S. H. 2019. Listeners consider alternative speaker productions in discourse comprehension and memory: Evidence from beat gesture and pitch accenting. *Memory & Cognition.* 47(8), 1515–1530.

[16] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics* 27(1), 55–105.

[17] Chen, S., Wang, B., Xu, Y. 2009. Closely related languages, different ways of realizing focus. In: *Proceedings of Interspeech 2009*, 1007–1010.

[18] Yan, M., Calhoun, S., Warren, P. 2022. The role of prominence in activating focused words and their alternatives in Mandarin: Evidence from lexical priming and recognition memory. *Language and Speech*. doi: 10.1177/00238309221126108

[19] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(5588), 746–748.

[20] Rohrer, P. L., Delais-Roussarie, E., Prieto, P. 2020. Beat gestures for comprehension and recall: Differential effects of language learners and native listeners. *Frontiers in Psychology* 11, 575929.

[21] Krahmer, E. J., Swerts, M. G. J. 2006. Hearing and seeing beats. In: *Proceedings of Speech Prosody 2006*.

[22] Morett, L. M., Fraundorf, S. H., McPartland, J. C. 2021. Eye see what you're saying: Contrastive use of beat gesture and pitch accent affects online interpretation of spoken discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47(9), 1494–1526.

[23] Turk, O. 2020. Gesture, prosody and information structure synchronisation in Turkish. Ph.D. dissertation, Victoria Univ. of Wellington, New Zealand.

[24] Swerts, M., Krahmer, E. 2008. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics* 36(2), 219–238.

[25] McNeill, D. 2006. Gesture and Communication. In: Brown, K. Anderson, A. H., Bauer, L., Berns, M., Hirst, G., Miller, J., (eds), *Encyclopedia of Language and Linguistics (2nd edition)*. Boston: Elsevier, 58–66.

[26] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner (Version 0.9.0).

[27] Xu, Y. 2013. ProsodyPro—A Tool for large-scale systematic prosody analysis. In: *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 7-10.

[28] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. (Version 6.3).

[29] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[30] R Core Team. 2017. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.

[31] Schneider, W., Eschman, A., Zuccolotto, A. 2012. E-Prime User's Guide. Pittsburgh: Psychology Software Tools, Inc. (Version 3.0).

[32] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13), 1–26.

[33] Yan, M., Warren, P., Calhoun, S. 2022. Focus interpretation in L1 and L2: The role of prosodic prominence and clefting. *Applied Psycholinguistics* 43(6), 1275-1303

ACCENT REALIZATION IN SPONTANEOUS SPEECH: GREEK H* AND L+H*

Stella Gryllia, Jiseung Kim, Riccardo Orrico, Na Hu, Amalia Arvaniti

Radboud Universiteit, Netherlands
stella.gryllia@ru.nl, jiseung.kim@ru.nl, orrico.riccardo@ru.nl, na.hu@ru.nl, amalia.arvaniti@ru.nl

ABSTRACT

Greek H* and L+H* contrast, such that the former pitch accent is realized as high pitch indicating new information, while the latter is realized as a "scooped" rise encoding contrastive information. This distinction is robust in controlled, scripted data but has not been tested in unscripted speech. To address this gap, we used Functional Principal Components Analysis (FPCA) to examine 1599 rising accents from unscripted speech elicited from eight native Greek speakers. The coefficients of the resulting PCs were analysed using LMEMs, which showed statistically significant differences between H* and L+H* for PCs 1-4. PC1 captured scaling differences, while PCs 2-4 reflected differences in contour shape, such that they support the analysis of L+H* as a "scooped" rise, and H* as high pitch. These results suggest that the distinction between H* and L+H* is relevant for Greek intonation phonology independently of speech type.

Keywords: intonation, variability, spontaneous speech, Greek, FPCA.

1. INTRODUCTION

English and other Germanic languages have a H*~L+H* accentual contrast [e.g., 1, 2, 3]. In English, H* indicates that the accented item is new in discourse and should become part of the common ground, while L+H* indicates it should be chosen from a closed set of alternatives and added to the common ground. Thus, H* and L+H* typically mark broad and narrow focus respectively. This contrast is robust in perception [4, 5], and in production with controlled, scripted data where contrastivity is critical [3, 6]. However, spontaneous speech has not provided equally strong evidence: [7] found that rising and high accents (L+H* and H* respectively) are present in spontaneous speech but this phonetic difference does not correspond to distinct pragmatic functions. In short, in spontaneous speech, English H* and L+H* are not as distinct as in scripted speech, shedding doubt on the validity of the contrast.

An analogous situation obtains in Greek, in which the descriptions of H* and L+H* are comparable to those for English in terms of phonetic realization and pragmatics [8, 9]. Based largely on qualitative data, these accents have been described as appearing in focal position followed by L-L% edge tones. In this context, H* creates a slightly declining plateau with the preceding accentual peak (if one is present) and is realized as a fall (Fig.1a); L+H* is realized as a rise from a low F0 point followed by a fall (Fig.1b).

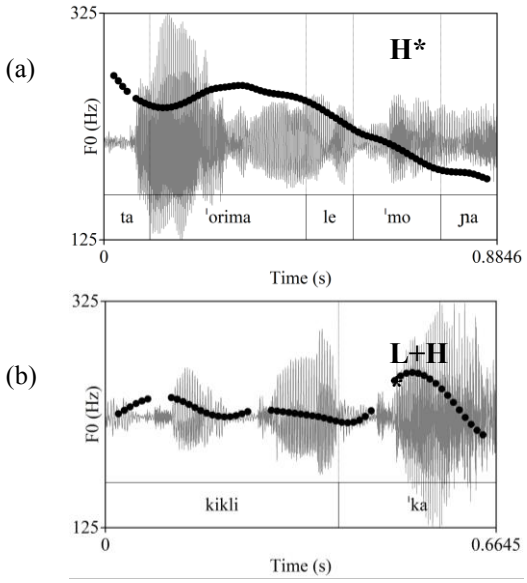


Figure 1: Waveforms and F0 tracks of [ta ˈorima leˈmoɲa] "the ripe lemons" with a H* accent on [leˈmoɲa] (a), and [kikliˈka] "in a circle" with a L+H* accent (b).

Recently, [10] applied Functional Principal Components Analysis (FPCA) on controlled scripted data to quantitatively examine this accentual contrast in Greek. FPCA is a data-driven, dimension-reduction method in which curves (time- and speaker-normalized F0 curves in [10]), are modelled as B-splines. Based on this modelling, FPCA returns the dominant modes of curve variation called Principal Components (PCs; [11, 12]; see 2.3 for details). In line with [8, 9], [10] showed that in Greek H* and L+H* differ in shape and scaling and are used in distinct pragmatic contexts. However, this quantitative evidence comes from scripted data only. Given the findings from English [7], it is reasonable to question whether the Greek accentual distinction is as robust in unscripted speech as it is in scripted speech [8, 9, 10] . We addressed this question by examining the H*~L+H* contrast in unscripted Greek data using FPCA.

2. METHODS

2.1. Participants

Eight native speakers of Standard Greek (4F), 22-27 years old ($\bar{X}$ = 24; SD = 1.7), who reported no speech or hearing disorders, took part. APF01 was bilingual in Greek (dominant) and Albanian (heritage); her data were included as they did not deviate from the average. All participants had learned English and additional languages via formal instruction (Italian, N = 3; French, N = 3; Spanish, N = 2; German, N = 1).

2.2. Materials and procedure

Unscripted speech was elicited by means of five tasks. In three monologues, participants narrated from memory a fable and a news item they had previously read, and told three stories using the app "Story Dice", in which players create a narrative based on icons depicted on six dice. Map tasks were used to elicit dialogues [13]: each speaker took part twice, once as instruction-giver (guiding the follower by using a map marked with a route) and once as instruction-follower (recreating the same route on their map which differed slightly from the instruction-giver's). In addition, participants discussed in pairs the functions of seven unusual objects appearing in a short video. Finally, the participants used the app "Guess who you are" (available in Greek) in groups of four to play an e-version of HedBanz (players place a card on a band on their head and guess the object on the card asking only yes/no questions).

Due to COVID-19 restrictions, the recordings were made using the experimenter's computer and the AVR application on the participants' phones. The recordings were saved as mono .wav files (256 kbps, 44.1 kHz sampling rate), a format that, according to [14], does not pose problems for F0 analysis.

2.3. Annotation and analysis

The data were annotated in Praat [15] after being orthographically transcribed. Trained annotators identified the nuclear accents of intermediate and intonational phrases that ended in L- and L-L% edge tones respectively. The annotators then used phonetic and pragmatic criteria to classify the accents (if high or rising) as H* or L+H*: accents were annotated as L+H* if they showed a rise from a low F0 point and their use was contrastive or corrective in context; high accents without a rise and accents that were not contrastive or corrective were annotated as H*. Finally, the annotators marked the *analysis window*, the interval over which the F0 analysis was to be performed; this included the accented word and its clitics, if any (e.g., του δεινόσαυρου [tu ðiˈnosavru] "the dinosaur.$_{GEN}$" was one analysis window). Forced alignment was then performed using Praat's "Align interval" function with the language set to Greek, and erroneous segment boundaries were manually corrected. The resulting corpus included 1599 accents categorized as H* or L+H*; see Table 1.

	Speaker								
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**Total**
H*	146	87	62	78	204	148	280	224	1229
L+H*	48	50	19	22	68	53	61	49	370
Total	194	137	81	100	272	201	341	273	1599

Table 1: Accent distribution by speaker.

The F0 contours of the analysis windows were extracted using Praat with an octave cost of 0.1 and the following pitch range and time steps (chosen to minimize tracking errors): APF01, APF05, APF07, APF08: 120–500 Hz, 5 ms time step; APM02, APM06: 70–280 Hz, 10 ms time step; APM03, APM04: 70–300 Hz, 10 ms time step. Remaining F0 halvings and doublings were visually identified [16] and manually corrected. Undefined F0 values were interpolated using stine interpolation [17]. The F0 contours were then scaled by speaker.

The next step was *landmark registration* to align the F0 curves around a specific time point and time-normalize them in the process. The onset of the accented vowel was used as landmark. The speaker-normalized and landmark-registered curves were then smoothed using k = 8 and $\lambda = 10^{\wedge}4$ (k refers to the number of spline knots and reflects the degree to which the original curves are faithfully modelled; λ reflects the degree to which smoothing is penalized).

The resulting F0 curves were subjected to FPCA following the procedures in [11, 12]. As mentioned in section 1, FPCA returns the dominant modes of curve variation (PCs). The shape of each F0 curve is treated as a function and receives a coefficient (or *score*) for each PC. Scores can then be used for statistical modelling to uncover differences between linguistic categories, here H* and L+H*.

The scores of the first five PCs were statistically analysed using linear mixed effects models (LMEMs; R: 4.2.1 [18]; lme4: 1.1.26 [19]). In all models, the PC scores were the dependent variable, and ACCENT (H*, L+H*) the independent variable. SPEAKER was included as random intercept and ITEM was nested within SPEAKER. This structure was selected because the data come from unscripted speech and thus each *item* was given a unique id (e.g., when APF01 says [ˈperasa ta ˈorima leˈmoɲa] "I went past the ripe lemons" *lemons* is item 01; when she says [ðeksiˈa ap ta ˈorima leˈmoɲa] "to the right of the ripe lemons" *lemons* is item 02).[1] The base model formula is given in (1); for PC1, R indicated that the model was

"singular", so we simplified the random structure by dropping the random intercept for SPEAKER.

(1) lmer (PC# ~ ACCENT + (1|SPEAKER) + (1|SPEAKER: ITEM), data = data, REML = FALSE)

3. RESULTS

Data and analyses are available at osf.io/kca26. Fig.2 shows the first five PCs, which capture 98.2% of curve variability in this dataset. The black line in each panel represents the smoothed and normalized average F0 curve and is the same in all five panels. The red curves show the changes to this average F0 curve when the coefficient of the depicted PC is up to +1 standard deviation above the mean; the blue curves show the same changes when the coefficient is up to -1 standard deviation. Table 2 presents the output of LMEMs and, for each model, marginal R^2 (an estimate of the proportion of variance explained by fixed factors, here ACCENT) and conditional R^2 (an estimate of the proportion of variance explained by the combined effect of fixed and random factors, here ACCENT, SPEAKER, and ITEM).

Fig.2 shows that PC1 captured 76.4% of the curve variability and that this variability was related to F0 scaling, with higher scores (red curves) leading to higher scaling relative to lower scores (blue curves). As shown in Table 2, PC1 scores for L+H* were significantly higher relative to H* (see also Fig.3a). This suggests that L+H* was scaled higher than H*. Conditional R^2 indicates that 27% of the variance was due to the combined effects of ACCENT, SPEAKER, and ITEM, while marginal R^2 indicates that only 3% of the variance in PC1 scores was related to ACCENT.

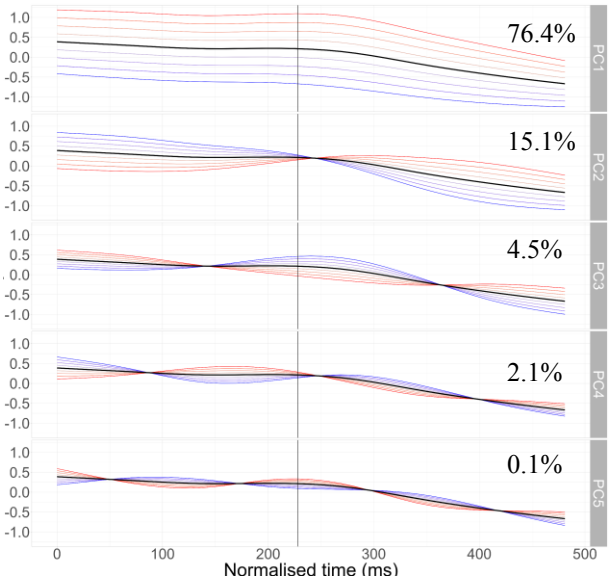


Figure 2: PC1-PC5 of the corpus; the vertical line indicates the onset of the accented syllable; percentages indicate variability captured per PC. For details, see text.

PC2 captured 15.1% of curve variability in the corpus. Fig. 2 suggests that this variability was related to F0 shape: higher PC2 scores resulted in a slightly "scooped" rise-fall, while lower scores resulted in a plateau-fall. The fact that PC2 scores were significantly higher for L+H* than H* (see Table 2 and Fig.3b), suggests that L+H* had a slightly rise-fall shape, while H* was realized as a plateau-fall. The conditional R^2 indicated that 68% of the variance was due to the combined effect of ACCENT, SPEAKER, and ITEM, while the marginal R^2 indicated that 10% of the PC2 variance was due to ACCENT.

PC3 and PC4 captured together 6.6% of curve variability. Fig.2 suggests that this variability was related to curve shape: higher PC3 and PC4 scores resulted in a fall, while lower scores resulted in a rise-fall. PC3 and PC4 scores for L+H* were significantly lower than those for H* (see Table 2 and Fig. 3c and 3d), indicating a difference between a rise-fall vs. a fall respectively. Conditional R^2 indicated that 54% and 64% of the variance (for PC3 and PC4 respectively) was due to the combined effect of ACCENT, SPEAKER, and ITEM, while marginal R^2 showed that just 2% of the variance in PC3 and PC4 was due to ACCENT.

Recapitulating, PC2, PC3, and PC4 captured differences between a rise-fall and a plateau-fall, with the differences among the three PCs relating mostly to the location of the peak in the rise-fall and the beginning of the fall in the plateau. These events can appear before or after the onset of the accented vowel (shown as the vertical black line in Fig.2).

PC1	**Est.**	**SE**	**df**	**t-value**	**Pr(>|t|)**
(Intercept)	-1.31	0.58	1024.9	-2.25	*
Accent: L+H*	7.16	1.05	1571.6	6.82	***
Marginal R^2: 0.03		*Conditional R^2: 0.27*			
PC2	**Est.**	**SE**	**Df**	**t-value**	**Pr(>|t|)**
(Intercept)	-2.04	0.83	8.35	-2.46	*
Accent: L+H*	5.92	0.42	1530.6	14.14	***
Marginal R^2: 0.09		*Conditional R^2: 0.68*			
PC3	**Est.**	**SE**	**Df**	**t-value**	**Pr(>|t|)**
(Intercept)	0.35	0.36	8.71	0.97	0.359
Accent: L+H*	-1.62	0.25	1584.6	-6.44	***
Marginal R^2: 0.02		*Conditional R^2: 0.54*			
PC4	**Est.**	**SE**	**Df**	**t-value**	**Pr(>|t|)**
(Intercept)	0.22	0.29	8.09	0.78	0.49
Accent: L+H*	-0.95	0.17	1533.7	-5.67	***
Marginal R^2: 0.02		*Conditional R^2: 0.64*			
PC5	**Est.**	**SE**	**Df**	**t-value**	**Pr(>|t|)**
(Intercept)	0.07	0.08	10.92	0.85	0.41
Accent: L+H*	-0.07	0.12	1444.62	-0.62	0.54
Marginal R^2: 0.00		*Conditional R^2: 0.70*			

Table 2: Results of LMEMs for PC1, PC2, PC3, PC4 and PC5; $p < .05$ = *, $p < .01$ = **, $p < .001$ = ***; the marginal and conditional R^2 of each model are presented below the relevant model in italics.

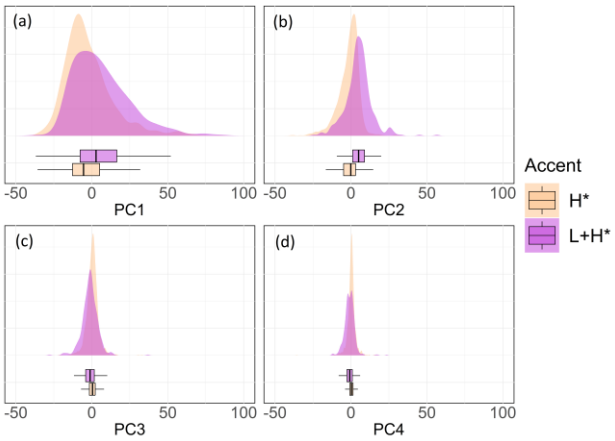


Figure 3: Density and boxplots of PC1, PC2, PC3, and PC4 scores, separately for H* and L+H* accents.

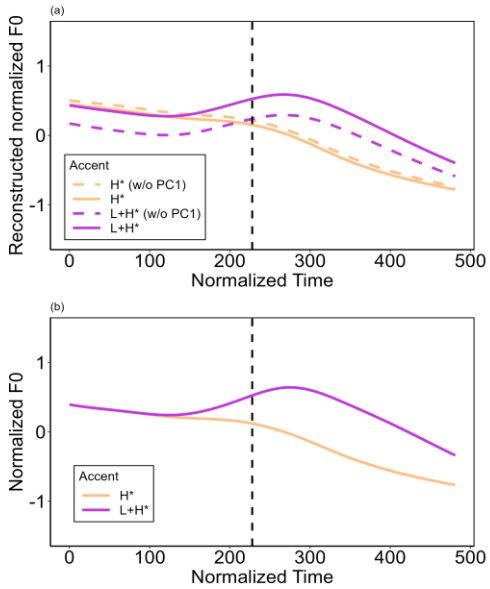


Figure 4: (a): Reconstructed F0 curves by ACCENT using (PC1), PC2, PC3, and PC4, and obtained by incorporating the average values of each PC score, as predicted by the linear model, into the FPCA equation. The equations for the two curves with all four PCs[2] are: for H*, $f(t) = \mu(t) - 1.31 \times PC1(t) - 2.04 \times PC2(t) + 0.35 \times PC3(t) + 0.22 \times PC4(t)$; for L+H*, $f(t) = \mu(t) + 5.85 \times PC1(t) + 3.87 \times PC2(t) - 1.27 \times PC3(t) - 0.73 \times PC4(t)$; (b) smoothed and averaged F0 curves pooled across speakers and tasks. In both plots, the vertical line indicates the onset of the accented syllable.

4. GENERAL DISCUSSION AND CONCLUSIONS

This study showed that H* and L+H* are distinct in Greek unscripted speech. In this dataset, H* was realized as a declining plateau followed by a fall, while L+H* had a slightly "scooped" rise-fall shape (Fig. 4). These findings align with the results reported in [9, 10] for scripted speech, providing supporting evidence that this accentual distinction between H* and L+H* is relevant for Greek intonational phonology.

This conclusion is reached by considering the shape and contribution of the PCs that had significantly different scores for ACCENT, the conditional and marginal R^2 for each LMEM, and the effect of each PC on curve reconstruction (Fig. 4).

As noted, PC1 captured scaling differences but its low conditional and very low marginal R^2 (compared to the other PCs) suggests that these scaling differences do not help distinguish H* from L+H* in Greek. They are most likely due to the changes in pitch level and span that are inevitably present in unscripted speech. This is supported by comparing the reconstructed curves with and without PC1. As shown in Fig. 4a, the curves are identical for H* and show only scaling differences for L+H*, suggesting that the scaling of L+H* is more variable (as indicated by the density plots in Fig. 3a as well). In contrast, while PC2, PC3, and PC4 explained relatively little of the variability in the corpus (21.7% all together), they captured contour shape differences of importance for the accentual contrast. This is reflected in the reconstructed curves which retain the same shape with and without PC1.

Conditional and marginal R^2 exhibited large differences in this analysis. The values of conditional R^2 suggest that, for all PCs, much of the variance was explained by the variables in the random structure. This is likely due to the nature of unscripted speech, in which syllable structure, accented syllable location, and the proximity of other tonal events are not controlled. Despite the extensive and uncontrolled variability in these unscripted data, however, the distinction between H* and L+H* did hold for Greek.

Finally, the use of FPCA followed by statistical modelling has important methodological repercussions for intonation research. Our results show that FPCA is suitable for analysing the intonation of unscripted speech. Most importantly, by using a data-driven, dimension-reduction technique followed by statistical modelling and by taking marginal and conditional R^2 values into account, and considering curve reconstruction, we have shown that it is possible to systematically distinguish variability due to linguistic factors of interest, such as the accentual contrast investigated here, from variability due to other factors, such as speaker-specific and item-specific differences. Further, this methodology allows us to deduce the features that are critical for distinguishing accentual contrasts. Here, for instance, it was shown that scaling differences were not as important as shape for distinguishing H* from L+H*. Taken together, these are helpful steps towards separating linguistic variation from general variability in the study of intonation and facilitate the study of spontaneous data.

5. ACKNOWLEDGEMENTS

The financial support of the European Research Council through grant ERC-ADG-835263 (SPRINT) to Amalia Arvaniti is hereby gratefully acknowledged. We thank our speakers who participated during the COVID-19 pandemic, Eleftheria Politi and Danae Tsinivits for data collection, Dafni Bagioka and Vicky Ioannidou for data annotation, and Katherine Marcoux for data processing in the early stages of this work.

6. REFERENCES

[1] Pierrehumbert, J. B. 1980. *The Phonetics and Phonology of English Intonation*. MIT PhD dissertation.

[2] Pierrehumbert, J. B., Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P. R., Morgan, J., Pollack, M. E. (eds) *Intentions in Communication*. Cambridge: MIT Press, 271–311.

[3] Greif, M., Skopeteas, S. 2021. Correction by Focus: Cleft Constructions and the Cross-Linguistic Variation in Phonological Form. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2021.648478.

[4] Ito, K., Speer, S. R. 2008. Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language* 58, 541–573. doi:10.1016/j.jml.2007.06.013

[5] Kurumada, C., Brown, M., Bibyk, S.A, Pontillo, D., Tanenhaus, M.K. 2014. Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition* 133(2), 335–342. doi:10.1016/j.cognition.2014.05.017.

[6] Kim, J., Arnhold, A. 2022. Prosodic focus-marking in Canadian English. [Conference poster]. 18th Conference on Laboratory Phonology.

[7] Jepson, K., Zhang, C., Lohfink, G., Marcoux, K., Arvaniti, A. 2021. H* and L+H* in English and Greek. [Conference presentation]. 4th Phonetics and Phonology in Europe.

[8] Arvaniti, A., Baltazani, M. 2005. Intonational analysis and prosodic annotation of Greek spoken corpora. In: Sun-Ah Jun (ed) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 84-117.

[9] Georgakopoulos, T., Skopeteas, S. 2010. Projecting vs. interpretational properties of nuclear accents and the phonology of contrastive focus in Greek. *The Linguistic Review* 27, 315-341. doi: 10.1515/tlir/2020.012

[10] Lohfink, G., Katsika, A., Arvaniti, A. 2019. Variability and category overlap in the realization of intonation. *Proc. 19th ICPhS* Melbourne, 701–705. https://assta.org/proceedings/ICPhS2019.

[11] Asano, Y., Gubian, M. 2015. "Excuse meeee!!": (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation. *Speech Communication* 99, 183–200.

[12] Gubian, M., Torreira, F., Boves, L. 2015. Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics* 49, 16–40. doi:10.1016/J.WOCN.2014.10.001.

[13] Anderson, A. H., Bader, M., Gurnan Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S. et al. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 351–366.

[14] Zhang, C., Jepson, K., Lohfink, G., Arvaniti, A. 2021. Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America* 149, 3910–3916. https://doi.org/10/1121/10/0005132.

[15] Boersma, P., Weenink, D. 2022. Praat doing phonetics by computer. [Online]. Available: http:// www.praat.org/.

[16] Baltazani, M., Przedlacka, J., Ünal-Logačev, Ö., Logačev, P., & Coleman, J. 2022. Intonation of Greek in contact with Turkish: a diachronic study. *Language Variation and Change*.

[17] Moritz, S. Bartz-Beielstein, T. 2017. ImputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1), 207-218. https://doi.org/ 10.32614/RJ-2017-009.

[18] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.r-project.org/.

[19] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 1.1, 1–48. [Online]. Available: https://www.jstatsoft.org/v067/i01.

[1] We also ran the analyses with a *crossed* random structure and the picture does not change with respect to the effect of ACCENT, except that R^2 conditional increases (which is expected given the model change).

[2] The equations for the two curves with PC2-PC4s[2] are: for H*, $f(t) = \mu(t) - 2.04 \times PC2(t) + 0.35 \times PC3(t) + 0.22 \times PC4(t)$; for L+H*, $f(t) = \mu(t) + 3.87 \times PC2(t) - 1.27 \times PC3(t) - 0.73 \times PC4(t)$

PROSODY AND WORD ORDER IN MARKING FOCUS WITHIN DJAMBARRPUYŊU NOUN PHRASES

Kathleen Jepson

Institute of Phonetics and Speech Processing (IPS), LMU Munich
k.jepson@phonetik.uni-muenchen.de

ABSTRACT

This paper examines word order and the prosodic realisation of focus within noun phrases in Djambarrpuyŋu, an Australian Indigenous language with free word order.

Within noun phrases, the prosodic marking of focus differs across languages and may include phonological and phonetic means. Languages with greater syntactic flexibility, however, may not prosodically encode information status within noun phrases in the same ways, or at all, though this is under-explored.

Djambarrpuyŋu noun phrases with focus on an adjective, a noun, or the noun phrase were elicited. Analysis of the data showed that focus condition was a predictor of word order within the noun phrase, specifically, focused adjectives occurred before given nouns. But neither phonological (accent distribution, accent type) nor phonetic (f0 peak, RMS amplitude peak) correlates of prominence consistently contributed to distinguishing the focus conditions.

Keywords: prosody, focus, information status, word order, Djambarrpuyŋu

1. INTRODUCTION

This paper examines word order and the prosodic realisation of focus within noun phrases in Djambarrpuyŋu (ISO 639-3: djr), an Australian Indigenous language with free word order, spoken in northern Australia by ~4,000 people (see Fig. 1).

The relationship between information structure and information status and prosody is well established cross-linguistically [1], [2]. Focus, which indicates the presence of alternatives [3], and givenness, which considers how activated or accessible information is in the minds of interlocutors [4], are commonly examined. Focused and new elements are often accented, sometimes with a special pitch accent, and are realised with concomitant phonetic cues to the phonological designation of "accented"; longer duration, increased intensity, and higher fundamental frequency (henceforth f0), resulting in greater phonetic prominence. Given information is acoustically less prominent and may be deaccented [2], [5]. In languages that resist deaccenting (see e.g., [6], [7]), phonetic cues may still encode focus and information status. Languages that allow greater syntactic variability may also rely less on prosodic cues to information structure or status [7], [8].



Figure 1: A map of Australia showing where Djambarrpuyŋu is spoken.

The prosodic realisation of focus within noun phrases (henceforth NPs)—a noun and its semantically-related modifiers such as adjectives—shows considerable cross-linguistic variation [9]. However, the aforementioned trends of focus marking are observed [7], [9], [10]. In some languages such as Italian and Japanese, distribution of intonational accents is not a significant predictor of focus nor information status within NPs [7], [11]. However, phonetic cues may still be employed in these languages. In Japanese NPs, adjectives and nouns have higher f0 range when new information focused than given [11]. Similar NP-internal pitch variability has also been reported for Hungarian, which has a syntactic focus position [12].

Djambarrpuyŋu presents a point of difference from previously investigated languages because it allows free word order, including within noun phrases [13]–[15].[1] Speakers may considerably alter the syntactic structure of an utterance as a means of conveying information structure, and clause initial position has pragmatic importance or is a position of prominence in Australian languages [16]. However, this is not a formal structural expression of focus. Prosodic investigations of Australian languages have found that focused constituents that are fronted may additionally occur in their own intonational phrase, prosodically dislocated from following material, and have the highest f0 (associated with a pitch accent) in the utterance [17]. Narrow focus may also be encoded by a rising pitch accent [18], [19].

Djambarrpuyŋu has fixed stress on the first syllable of the word, and shows similarities to head/edge-prominence languages [20]. Previous prosodic analysis of Djambarrpuyŋu [21] suggests that deaccenting is an uncommon strategy for encoding given information, as reported for other Australian languages [18], [19]. But there is evidence that if syntactic position is maintained (see [22]), deaccenting can be used to encode accessible information in Djambarrpuyŋu [21].

In this paper, word order and prosodic cues to focus are examined within contiguous NPs (i.e., in which target words are adjacent), considering patterns of accentuation and accent type, as well as the acoustic measures of peak f0 and peak root mean square (henceforth RMS) amplitude. This study aims to describe if and how focus is encoded in Djambarrpuyŋu noun phrases by these measures, whether the order of the adjective and noun in NPs varies consistently by focus condition, and how syntactic and prosodic encoding may interact. To investigate these topics, data from a production task eliciting noun phrases were examined.

2. METHODS

2.1. Participants

Thirteen native Djambarrpuyŋu speakers (7 women and 6 men, mean age 45, age range 20-69) were recorded in Milingimbi, N.T., Australia. All participants were familiar with related language varieties, other Aboriginal languages, and Australian English. Participants were paid for their time.

2.2. Stimuli

Five nouns and five colour-term adjectives were selected for creating the materials, and were matched in all combinations resulting in 25 unique picture tiles (see [9]). The target words were selected to be both segmentally and morphologically diverse, and broadly representative of the types of structures observed in Djambarrpuyŋu words; targets words were one to five syllables in length, and included reduplicated forms and compound nouns.

Twelve 2×2 grids were created using the 25 tiles (see e.g., Fig. 2). There were three combination types corresponding to three focus conditions: 1) each tile had a different picture of a different colour resulting in whole noun phrase focus (henceforth NPF); 2) all the pictures were different but of the same colour resulting in focus on the noun (henceforth NF); and 3) all the pictures were the same but of different colours resulting in focus on the adjective (henceforth AF). The design is summarised in Table 1.

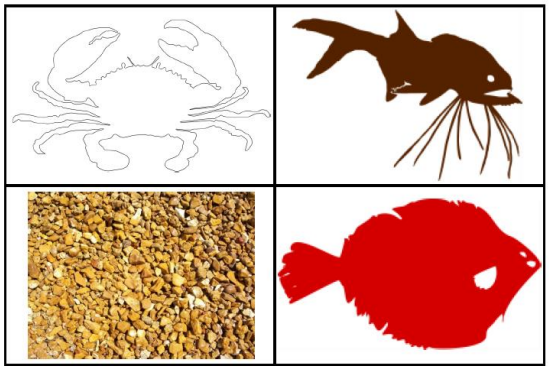


Figure 2: A grid from the NPF condition, using nouns *nyoka* 'crab', *malmuŋu* 'threadfin salmon', *minyirrminyirr* 'gravel', *raŋanŋaniŋ* 'sole', and adjectives *watharr* 'white', *gurrŋan* 'brown', *buthalak* 'yellow', *miku* 'red'. Other target words, not shown, were *buthuruwuŋgan* 'hammer oyster', and *mol* 'black'.

	Focused element	
Condition	Adjective	Noun
NPF	✓	✓
NF		✓
AF	✓	

Table 1: Summary of the focused element in the three focus conditions.

2.3. Procedure

Before commencing the task, participants were shown the materials, and the nouns and colour terms were discussed. The participant and their interlocutor (the author, or fellow participant) were seated across from each other at a table. Each participant completed three blocks of four trials, corresponding to the three focus conditions. In each trial, the participant was shown one of the 2×2 grids on a computer screen not visible to the interlocutor. The interlocutor had a blank, hardcopy 2×2 grid in front of them. Between the participant and interlocutor were five laminated coloured picture tiles that included those on the participant's screen plus an additional tile appropriate for the focus condition. For each tile in the grid (i.e., turn), the participant was asked to described the tile, and instruct the interlocutor where it was located on the grid using directionals "top", "bottom", "left", and "right". Example frame sentences, e.g., (1) were discussed with participants in the instructions; however, participants varied word order at the utterance level, and also used other verbs such as intransitive "to go", though the effect of verb is not considered here (see [23], [24]). Participants were encouraged to use both the noun and adjective words but were free to give instructions however they preferred, resulting in naturalistic speech. There was some variability in the words used (e.g., *buthuruwuŋgan~buthuruwaṯu*), which is not explored

in the current study. The interlocutor had to place the correct tile in the correct position on the blank grid. The trials were presented in three blocks of the NPF-, NF-, and AF-conditions.

(1) *Nyoka watharr nhe dhu rulwaŋdhun*
crab white 2SG FUT put.down
garrwar-lil wiŋ'kuŋu-lil
top-ALL left-ALL
'You will put the white crab to the top left.'

Audio data were collected using a Zoom H6 digital recorder and Countryman H6 headset microphone with a hypercardioid pattern directional capsule covered with a windshield. Recordings were made at 24 bit bit-depth and a 48 kHz sample rate. Recording sessions primarily took place sitting outside, either in the open or on a veranda.

2.4. Data processing and acoustic measures

Audio files were transcribed and prepared in Praat [25], [26], and utterances were forced aligned using WebMAUS [27]. An EMU-SDMS database was created [28], and segmentation was manually corrected. Additional tiers included utterance, target word, orthographic word, and a tone tier for prosodic annotation. Data were labelled prosodically based on visual inspection of the f0 contour and auditory impressions following [21]. F0 (Hz) and RMS amplitude (dB) contours were extracted from the EMU database for the target words, and the f0 and RMS amplitude peaks associated with the accented syllable were calculated.

2.5. Analysis

Data were statistically analysed with generalised linear mixed models (henceforth GLMM) and linear mixed effects models (henceforth LMM) in R using the *lme4* [29] and *emmeans* [30] packages. Fixed effects were WORD ORDER (adjective-noun vs. noun-adjective) or POSITION ORDER (first position vs. second position), FOCUS CONDITION (AF- vs. NF- vs. NPF-condition), and WORD TYPE (adjective vs. noun). Random intercepts were SPEAKER and WORD IDENTITY. Details of models are provided in the results subsections where relevant.

3. RESULTS

A total of 654 utterances were collected: the intended 624 (4 tiles × 4 trials × 3 focus conditions × 13 speakers) plus 30 repetitions, predominantly prompted by production errors. In this study, contiguous NPs are considered. Therefore, utterances in which the target words were not adjacent (126), or where only one target word was produced (88) were excluded from the analysis. A further 13 utterances were excluded as they were repetitions, and nine due to production errors, resulting in 418 utterances for analysis; 836 target words. The distribution of utterances by focus condition was AF = 163, NF = 133, NPF = 122.

3.1. Word order

The relationship between word order and focus condition was tested with a GLMM. Word order was predicted, with a fixed effect of focus condition, and a random intercept for speaker.

Focus condition was a significant predictor of the order of the noun and adjective within NPs ($\chi^2(2) = 73.47$, $p < 0.0001$). Word order patterns were similar for NPF- and NF-conditions, with both orders being almost equally used—adjective-noun order was used in 56% of NF-condition and 54% of NPF-condition utterances. In the AF-condition, the adjective-noun order was far preferred, being used in 85% of utterances (AF ~ NF $p < 0.0001$, AF ~ NPF $p < 0.0001$, NF ~ NPF n.s.).

3.2. Accent distribution and pitch accent type

Almost all target words were accented (n = 828). Accent type was predicted in an LMM with fixed effects of focus condition, position order, and word type, with a three-way interaction, and a random intercept for speaker. No effect was significant in predicting accent type.

While accent distribution could not be analysed statistically as deaccenting was used infrequently (n = 8), deaccenting was employed as a strategy to mark given information by one speaker. This speaker deaccented nouns in the AF-condition in five out of 16 utterances, and strongly preferred the adjective-noun order in all focus conditions.

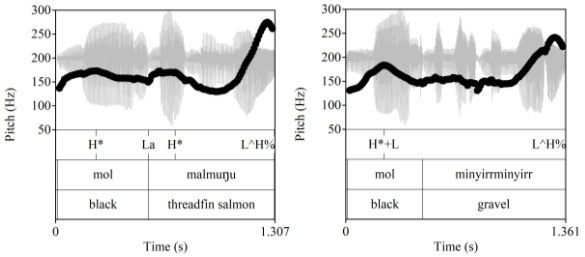


Figure 3: Right: an utterance where the adjective and noun are accented (NPF-condition); Left: a rare example of deaccenting (AF-condition)—the noun *minyirrminyirr* does not have an associated pitch accent.

3.3. F0

F0 peak was predicted in an LMM with fixed effects of focus condition, position order, and word type, with a three-way interaction, and random intercepts for speaker and word identity. The interaction

between focus condition, position order, and word type was significant in predicting f0 peak height ($\chi^2(3) = 11.2$, $p = 0.011$). However, there was only one significant comparison: adjectives in the NPF-condition had a significantly lower f0 peak when they occurred in the second position compared with the first ($\beta = -22.47$, $p < 0.0001$).

While other differences where not significant, there were trends (see raw values in Fig. 4). For example, in the AF-condition, adjectives in either first or second position had higher f0 peak values than nouns.

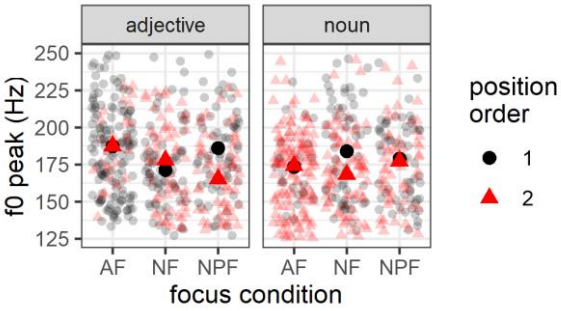


Figure 4: F0 peak values (Hz) and means for adjectives and nouns across focus conditions, grouped by whether they were the first (1) or second (2) word in the NP.

3.4. RMS amplitude

RMS amplitude peak was predicted in an LMM with fixed effects of focus condition, position order, and word type, with a three-way interaction, and random intercepts for speaker and word identity. The interaction between focus condition, position order, and word type was significant ($\chi^2(3) = 31.6$, $p < 0.0001$).

A number of comparisons were significant (see z-score values in Fig. 5; RMS amplitude values were z-scored by speaker for plotting due to strong systematic speaker differences that interacted with position order). Adjectives in second position had significantly lower RMS amplitude peak values when in NF- than AF-conditions ($\beta = -3.33$, $p < 0.001$). Nouns in second position had higher RMS amplitude peak values when in NF- than AF-condition ($\beta = 1.6$, $p < 0.05$). In second position of AF-condition utterances, nouns had significantly lower RMS amplitude peak values than adjectives in that position ($\beta = -3.54$, $p < 0.05$). Adjectives in AF-condition had significantly higher RMS amplitude peaks in second than first position ($\beta = 3.22$, $p < 0.001$).

4. DISCUSSION AND CONCLUSIONS

The aim of this paper was to describe how focus is realised prosodically in Djambarrpuyŋu noun phrases by both phonological and phonetic means, and what

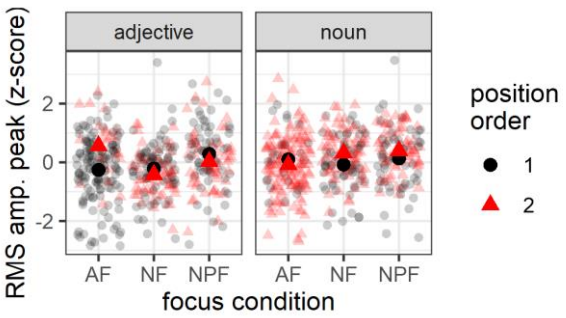


Figure 5: RMS amplitude peak values (z-score) and means for adjectives and nouns across focus conditions, grouped by whether they were the first (1) or second (2) word in the NP.

effect focus condition may have on word order. Specifically, accentuation and accent type, peak f0, and peak RMS amplitude were examined.

Word order significantly varied due to focus condition, with adjective-noun order being more common in the AF-condition than the NF- or NPF-conditions. Neither accentuation nor accent type varied consistently with focus condition. Phonetic prosodic cues did not vary consistently to distinguish between the different focus conditions in these data, and it is not clear whether all the patterns that were observed served to increase the prominence of the focused element c.f. [9]. This aligns Djambarrpuyŋu with languages like K'iche', of Guatemala, for which traditionally investigated phonological and phonetic means of encoding prosodic prominence are not found to reflect focus condition in NPs [9] (see also [31]). However, the current figures and statistical results do suggest that adjectives and nouns are treated differently from one another across focus conditions, paralleling results in [9].

Word order in this study was narrowly considered, only as the order of the target words in contiguous NPs, but discontinuous NPs may shed light on the diversity of ways focus within NPs is encoded in Djambarrpuyŋu. Further, it may be that other methods for encoding givenness assist in distinguishing the focus conditions, such as deletion, and, as mentioned above, deaccenting which while not regularly used, may be employed by some speakers in cases of continuity of word position in combination with givenness, as has been described for English [22].

Investigations are currently underway into discontinuous noun phrases, as well as additional acoustic measures such as duration, dynamic measures of pitch, and other aspects of prosody such as phrasing. However, the current results do not suggest phonological nor phonetic correlates of prominence consistently contribute to distinguishing words' information status within Djambarrpuyŋu NPs in these three focus conditions.

5. ACKNOWLEDGEMENTS

Many thanks to the Yolŋu who participated in this study and in earlier work. Thank you to Elizabeth Räkay Milmilany, Johanna Schirmer, and Rasmus Puggaard-Rode for assistance and advice, and colleagues at the IPS and a reviewer for their comments. This work was supported by the Alexander von Humboldt Foundation, the University of Melbourne and the ARC Centre of Excellence for the Dynamics of Language (CE140100041).

6. REFERENCES

[1] Arvaniti, A. 2022. The autosegmental-metrical model of intonational phonology. In: Barnes, J., Shattuck-Hufnagel, S. (eds), *Prosodic Theory and Practice*. MIT Press, 25–83.

[2] Kügler, F., Calhoun, S. 2020. Prosodic encoding of information structure: A typological perspective. In: Gussenhoven C., Chen, A. (eds), *The Oxford Handbook of Language Prosody*. OUP, 453–467.

[3] Krifka, M. 2008. Basic notions of information structure, *Acta Linguist. Hung.*, 55(3-4), 243–276.

[4] Baumann, S., Grice, M. 2006. The intonation of accessibility, *J. Pragmat.* 38(10), 1636–1657.

[5] Chodroff, E., Cole, J. 2019. The phonological and phonetic encoding of information structure in American English nuclear accents. *Proc. 19th ICPhS* Melbourne, 1570-1574.

[6] Chahal, D., Hellmuth, S. 2014.The intonation of Lebanese and Egyptian Arabic. In: Jun, S.-A. (ed), *Prosodic Typology*, vol. II. OUP, 365–404.

[7] Swerts, M., Krahmer, E., Avesani, C. 2002. Prosodic marking of information status in Dutch and Italian: A comparative analysis. *J. Phon.* 30(4), 629–654.

[8] Ladd, D. R. 2008. *Intonational Phonology*, 2nd edn. CUP.

[9] Burdin, R. S., Phillips-Bourass, S., Turnbull, R., Yasavul, M., Clopper, C. G., Tonhauser, J. 2015. Variation in the prosody of focus in head- and head/edge-prominence languages. *Lingua* 165, 254–276.

[10] Ito K., Speer, S. R. 2006. Using interactive tasks to elicit natural dialogue. In: Sudhoff, S. et al. (eds), *Methods in Empirical Prosody Research*. De Gruyter.

[11] Swerts, M., Taniguchi, M., Katagiri, Y. 2000. Prosodic marking of information status in Tokyo Japanese. *6th ICSLP* Beijing 1, 78–81.

[12] Langer, C., Kügler, F. 2021. Focus and prosodic cues in Hungarian noun phrases. *1st TAI* Sonderborg, 219–223.

[13] Wilkinson, M. 2012. *Djambarrpuyŋu: A Yolŋu Variety of Northern Australia*. LINCOM Europa.

[14] Nordlinger, R. 2014. Constituency and grammatical relations in Australian languages. In: Koch, H., Nordlinger, R (eds), *The Languages and Linguistics of Australia*. Mouton de Gruyter, 215–261.

[15] Louagie, D., Verstraete, J.-C. 2016. Noun phrase constituency in Australian languages: A typological study. *Linguist. Typology* 20(1), 25–80.

[16] Simpson, J., Mushin, I. 2008. Clause-initial position in four Australian languages. In: Mushin, I., Baker, B. (eds), *Discourse and Grammar in Australian Languages*. John Benjamins, 25–57.

[17] Fletcher J., Butcher, A. 2014. Sound patterns of Australian languages. In: Koch, H., Nordlinger, R (eds), *The Languages and Linguistics of Australia.* Mouton de Gruyter, 91–138.

[18] Bishop, J., Fletcher, J. 2005. Intonation in six dialects of Bininj Gun-wok. In: Jun, S.-A. (ed), *Prosodic Typology*. OUP, 331–361.

[19] Fletcher, J. 2014. Intonation and prosody in Dalabon. In: Jun, S.-A. (ed), *Prosodic Typology*, vol. II. OUP, 252–272.

[20] Jun, S-A. 2014. Prosodic typology: By prominence type, word prosody, and macro-rhythm. In: Jun, S.-A. (ed), *Prosodic Typology*, vol. II. OUP, 520–539.

[21] Jepson, K., Fletcher, J. Under review. The intonation of Djambarrpuyŋu.

[22] Terken, J., Hirschberg, J. 1994. Deaccentuation of words representing "given" information: Effects of persistence of grammatical function and surface position. *Lang. Speech* 37(2), 125–145.

[23] Röhr, C. T., Baumann, S., Grice, M. 2015. The effect of verbs on the prosodic marking of information status: Production and perception in German. *Proc. 18th ICPhS* Glasgow.

[24] Fletcher, J., Stoakes, H., Singer, R., Loakes, D. 2016. Intonational correlates of Subject and Object realisation in Mawng (Australian). *Speech Prosody* Boston, 188–192.

[25] Boersma P., Weenink, D. 2022. Praat: Doing phonetics by computer. (v. 6.2.23)

[26] Zhang, C. 2021. Praat (GitHub repository). https://github.com/congzhang365/Praat

[27] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347.

[28] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Comput. Speech Lang.* 45, 392–410. (v. 2.3.0)

[29] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 67(1), 1–48. (v. 1.1-31)

[30] Lenth, R. V. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. (v. 1.8.2)

[31] Kaland, C., Swerts, M., Himmelmann, N., Red and blue bananas: Time-series f0 analysis of contrastively focused noun phrases in Papuan Malay and Dutch. *J. Phon.* 96, 101200.

[1] Discontinuous NPs (in which semantically-related nominal elements are distributed throughout the clause) are also permitted in Djambarrpuyŋu though not explored here.

The perception of parallel encoded emotional and linguistic prosody in Mandarin Chinese

Qianyutong Zhang[1,2], Shanpeng Li[1,3]

[1] MIIT Key Lab for Language Information Processing and Applications (LIPA), School of Foreign Language, Nanjing University of Science and Technology, China
[2]Institute of Linguistics, Shanghai International Studies University, Shanghai, China
[3]Lab of Anhui Language Resources Preservation and Research, Anhui University, China.

ABSTRACT

In speech prosody, emotional prosody (e.g., angry or neutral emotion) and linguistic prosody (e.g., declarative or interrogative intonation) are the two most common prosody functions. Successfully recognizing these different functions are crucial to daily communication. However, the two functions are usually parallel encoded in the same prosody contour, causing perceptual difficulties. In scarce previous studies focusing on the recognition of the two prosodic functions, most research focused on the non-tonal languages. It remains unknown how the two functions interacted when perceived by tonal language speakers. To answer this question, twenty-two native Mandarin speakers were participated to identify the linguistic and emotional information embedded in the one sentence. The results confirmed the interaction between linguistic and emotional function in Mandarin but with different orientations: the interrogative intonation impedes the identification of neutral emotion, but facilitates the angry emotion. On the other side, angry emotion impedes the identification of declarative intonation only.

Keywords: Parallel Perception, Linguistic Function, Emotional Function, Mandarin Chinese

1. INTRODUCTION

Speech communication is an essential part of our daily life, in which prosody is a reliable element for listeners to comprehend the speaker's intention [1]. Speech prosody is mainly used to express linguistic (e.g., declarative or interrogative intonation) and emotional information (e.g., angry or neutral emotion), manifested by the same acoustic parameters such as F0, duration, and intensity [2-4]. Emotional prosody is used to convey the speaker's emotional state. For example, if the speaker is angry, the acoustic features will have higher mean F0 (in most languages), larger intensity, and faster speech rate than emotionlessness (i.e., neutrality) [5]. At the same time, speech prosody also conveys linguistic information such as sentence type. In interrogative sentences, for example, the pitch contour is rising, and the mean F0 is higher than declarative counterparts [6].

According to the PENTA model [7-8], different communicative functions were encoded in parallel in the same prosody. However, previous studies have demonstrated that the decoding of these coinciding prosodic functions appears an interaction effect so that the listeners may have difficulty in identifying each prosody [3, 9-10]. To be specific, linguistic prosody like intonation will affect the recognition of emotion types. For example, Scherer et al. [11], found that the rising intonation of yes/no questions are perceived as more agreeable and polite. On the other hand, emotional prosody also interferes with the perception of linguistic prosody. Pihan et al. [3] revealed that the identification accuracy of interrogative/declarative contrasts was reduced in emotional stimuli because sentences with emotional prosody have greater pitch variability, so the linguistic signals marked by pitch direction become blurred. However, although these studies demonstrated the fact that the existence of one function of prosody does affect the perception of another prosodic function, they didn't reveal the detailed pattern in which how emotions and intonations affect the perception of each other. What's more, previous studies mostly focused on Indo-European languages such as English and German [3,11], thus we don't know whether people who speak Mandarin Chinese have similar perception results.

Despite the scarcity of Chinese studies in this field, the study of Mandarin Chinese is of great significance since it is a tonal language that has some differences from non-tonal languages. For example, in Mandarin Chinese, F0 works not only at the sentence level to convey intonation information similarly with non-tonal languages, but also at the syllable level to distinguish lexical meanings [12]. At the same time, the perception of prosody mainly depends on F0 [13], so the F0 differences may manifest themselves in the prosody perception process of Chinese and English utterances. Hence if the experiment on tonal Mandarin shares similar results with non-tonal English, it will be a better verification for a cross-language commonality that

different prosodic functions have interactions during the perception of speech prosody.

Therefore, the present study aims to explore whether and how different functions of prosody influence Mandarin Chinese speakers' identification when emotional and linguistic prosody are simultaneously perceived. More specific, the current study explore whether and how the existence of speech emotions affects the identification of intonation information to distinguish interrogative and declarative sentences and whether and how intonations affect the perception of emotions by Mandarin Chinese speakers.

2. METHOD

2.1. Participants

Twenty-two native Mandarin Chinese speakers (12 female; 10 male) aged from 19 - 23 yrs. (mean = 21.5, SD = 1.64) participated in this perceptional experiment. All the participants speak Mandarin Chinese as their first language. All of them reported with normal hearing and normal or corrected-to-normal vision. They were provided written informed consent forms and were remunerated for their participation.

2.2. Stimuli

2.2.1. Corpus

The corpus consisted of 130 syntactically similar and semantically neutral sentences (e.g., "小马正在看电视Xiaoma zhengzai kan dianshi." / Mark is watching TV). Each sentence was produced with two different intonations (declarative or interrogative) and two different emotions (angry or neutral). These two factors are fully crossed, resulting in four conditions: (1) declarative intonation superimposed with a neutral emotion; (2) declarative intonation superimposed with an angry emotion; (3) interrogative intonation superimposed with a neutral emotion; (4) interrogative intonation superimposed with angry emotion. In addition, 130 fillers are also recorded, which are different from the target sentences in syntax and emotion.

2.2.2. Recording session

All target sentences were recorded by a female native Mandarin speaker. The speaker was instructed to produce the sentences presented randomly within four conditions and to use normal speech rate and loudness as they were having daily conversations. Before formal recording, the speaker looked through the textual materials to familiarize target sentences. Stimuli were recorded in a soundproof room in Phonetics Lab in Nanjing University of Science and Technology. The sounds were recorded using the cardioid condenser microphone Neumann U87Ai and the audio interface RME Fireface UCX and saved in WAV format with a 44.1 kHz sampling rate at a 16-bit resolution. All the recordings were normalized to amplitude of 70 dB HL (to correct for slight differences in recording levels across recording sessions). Finally, a total of 650 recordings (130 target sentences × 4 speaking styles + 130 filers) were collected.

2.2.3. Validation test

To guarantee the ecological validity of stimuli recorded, a validation test was conducted after recording. Sixty-one native Chinese speakers who did not participate in the recording session were recruited. They were required to rate the degree of emotion using a 5-points Likert scale, from 1 (emotionless) to 5 (extremely angry), and to identify the intonations of the utterances using a two-alternative forced-choice (i.e., "declarative sentence" or "interrogative sentence"). The results of the validation test showed that the mean score of sentences with an angry emotion was 3.73, while that of the neutral emotion was 1.33. The mean accuracy of declarative and interrogative utterances was 99.04% and 88.43% respectively. In consideration of the quantity and quality of the corpus, those sentences scored less than 3 for angry or more than 2 for neutral, and the identification rate of declarative or interrogative utterances less than 75% were excluded for further experiments. Finally, 480 recordings were retained (120 target sentences × 4 conditions).

2.3. Procedures

The perception experiment was conducted in a quiet computer classroom. Each participant was seated in front of a computer monitor and the recording was played through a headset. All the stimuli were randomly presented in the experiment, divided into 2 tasks. In each task, there were 250 trials (30 trials × 4 conditions + 130 fillers). Each trial involved one sentence, and each sentence presented in the task was different.

Before each stimulus, a beep was played to call the subjects' attention. When the audio was playing, a picture of a loudspeaker was displayed at the center of the screen. After each playback, a task was presented. The subjects should respond within a limited time (5 sec). There were two tasks, both of which were three-item forced choice identification tasks: In the emotion task, subjects were asked to

ignore the sentence types and judge the emotion types of each utterance within **neutral** / **angry** / **other** emotion ("other" is set because fillers contain emotions other than angry and neutral); (2) In the intonation task, participants were asked to ignore speech emotion and judge the sentence type within **declarative** / **interrogative** / **other** intonation ("other" is set to balance with that in the emotion task). Subjects were required to press corresponding keys on the computer keyboard. And the sequence of tasks was balanced among the subjects. The subjects had about 10-minute rest between the two tasks.

2.4. Data collection and statistical analysis

Participants' responses were collected via PsychoPy [14]. The percentage accuracy was arcsine-transformed first to satisfy the normal distribution of linear models. All transformed mean accuracy values were analyzed using the linear mixed-effect model in R [15] with the "lmer" function in the "*lme4*" package [16]. The task (emotional vs. intonational task), and condition (neutral declaratives, neutral interrogatives, angry declaratives, and angry interrogatives) were set as fixed factors and the subjects was set as random factor. The p -values of the main and interaction effects were calculated using Satterthwaite approximation from the "ANOVA" function in the "*lmerTest*" package [17]. The observed significant interaction effect was further executed by the "emmeans" function in the "*emmeans*" package [18] to conduct post-hoc pairwise comparisons.

3. RESULTS

As Table 1 shows, the linear mixed-effect model revealed a significant two-way interaction effect between task and condition (F (3, 119) = 7.450, p = 0.0001), and also a significant main effect of task (F (1, 119) = 17.195, p < 0.0001), as well as a significant main effect of condition (F (3, 119) = 26.194, p <0.0001). The post hoc results of the two-way interaction effect were reported in the following sections.

	df1	df2	F	p
Task	1	119	17.195	<.0001
Condition	3	119	26.194	<.0001
Task×Condition	3	119	7.450	.0001

Table 1: The omnibus effect of each main and interaction fixed effect from linear mixed-effect for identification accuracy rate.

3.1. The effect of intonation under emotion task

The Tuckey-HSD post hoc comparison for the two-way effect showed that, for intonation task, the accuracy of interrogative intonation (77.96%) was significantly higher than declarative intonation (49.80%; β = 0.367, SE = 0.085, t = 4.337, p < 0.0001) in angry utterances. On contrast, the accuracy of interrogative intonation (72.04%) was significantly lower than declarative intonation (97.71%; β = - 0.407, SE = 0.085, t = - 4.809, p <0.0001) in neutral utterances (see Fig. 1).

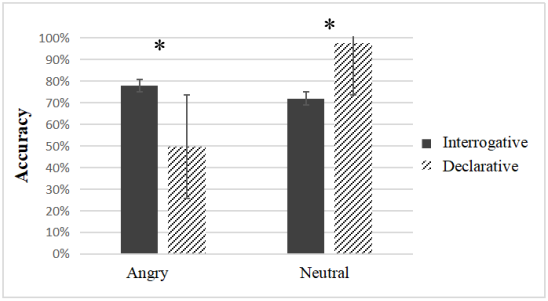


Figure 1: The mean identification accuracy of angry and neutral emotions under different intonations.

3.2. The effect of emotion under intonation task

For the emotion task, there was no significant difference between angry and neutrality in interrogative utterances (86.98% and 79.63% respectively; p = 0.4). However, the identification rate of angry emotion (79.68%) was significantly lower than neutral emotion (100%; β = - 0.346, SE = 0.085, t = -4.090, p < 0.001) in declarative utterances.

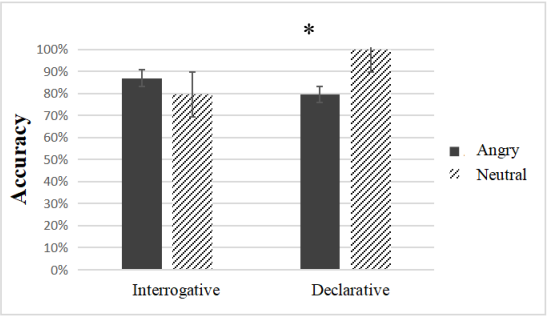


Figure 2: The mean identification accuracy of interrogative and declarative intonations under different emotions.

4. DISCUSSION

The current study explored the perception of different prosodic functions by native Mandarin speakers. The results revealed an interplay effect between linguistic and emotional prosody.

4.1. The influence of linguistic prosody on the perception of emotional prosody

The current study revealed that linguistic prosody (i.e., intonation) affects the perception of emotional prosody in Mandarin: interrogative intonation interferes with the identification of neutral emotion, but seems to improve angry emotion judgment. The role of declarative intonation was just in reverse: the identification of neutral emotion was improved, while that of angry emotion was declined when the carrier sentences were coupled with declarative intonation. This phenomenon might be attributed to the mean F0 and the final F0 movement [19]. Specifically, interrogative intonation has a higher mean F0 and a rising final F0 contour. However, these two prosodic cues are also used to convey emotionally exciting expressions like surprise, happy and anger [5]. Therefore, subjects might be more likely to mistake the emotionally neutral prosody for angry or other exciting emotions when they perceive the sentences with the interrogative intonation.

The accuracy of interrogative intonation was significantly higher than that of declarative intonation when sentences sounded angry, indicating that the interrogative intonation facilitates the perception of angry emotion. Similar to what has discussed above, interrogative intonation and angry emotion share both higher mean F0 and rising F0 contour, thus a superposition effect could be generated to make emotionally angry prosody seemed angrier under interrogative intonations.

4.2. The influence of emotional prosody on the perception of linguistic prosody

The results also revealed an effect of emotional prosody on the perception of declarative intonation but no effect on the identification of interrogative intonation. This is consistent with the results of Pihan et al. [3]. In their research, the discrimination rate of interrogative/declarative contrast was reduced when perceiving stimulus with emotional intonation. They explained that the pitch variability associated with emotion identification interfered with the pitch direction which is related to linguistic prosody perception. In angry sentences with high pitch variability, the stable signal of pitch direction in declarative intonation may have become covered, resulting in participants' lower accuracy of declarative intonation under angry emotion.

As for interrogatives, however, angry or neutral emotions do not affect the perception of interrogative intonation. It seems unconvincing since other results revealed a preliminary conclusion that different prosodic functions who share similar cues will have interactions. Nevertheless, this exceptional case has also been proved by Bryant and Fox Tree [20] that the linguistic function of interrogative prosody is not affected by the emotional function. From a prosody production perspective, McRoberts et al. [21] also proved that the final F0 rise of interrogative intonation does not decrease in emotional sentences. Combined with Pihan et al.'s speculation, it seems that the high pitch variability will not affect the prominently rising pitch direction in interrogative sentences.

Whilst the present study revealed the interaction pattern of emotional and linguistic prosody simultaneously decoded by Mandarin Chinese speakers, there are still some limitations in the current study. First, only declarative/interrogative contrasts were included as instances of linguistic prosody while prosody has many other linguistic functions, like tone, focus, prominence, etc. Similarly, only angry and neutral emotions were used as emotional stimuli in the current study, thus positive emotional valence and other emotional types should be explored in future studies. Second, since we only considered correct responses, it's not clear whether participants are failing to distinguish between emotional and neutral speech, or between anger and some other emotions. This can be further explored in future studies. Finally, the current study didn't control the final lexical tone of Mandarin sentences, though many studies have proved its influence on the identification of the interrogative sentences. Hence more related research is needed to confirm the conclusion of this study in future studies.

5. CONCLUSION

This study examined the interaction of linguistic and emotional prosody during speech perception in Mandarin. The results showed that linguistic intonation can affect the perception of emotional prosody. Specifically, interrogative intonation interferes with the identification of neutral emotion but facilitates the identification of angry emotion. Meanwhile, emotional prosody also affects intonation perception. Angry interferes with the perception of declarative intonation, while different emotions do not affect the perception of interrogative intonation. The present study enriches our knowledge of the synchronic perception of parallel encoded emotional and linguistic prosody by Mandarin Chinese speakers.

6. ACKNOWLEDGEMENTS

This research was supported by the National Social Sciences Foundation of China, No. 19BYY043 and the Fundamental Research Funds for the Central Universities, No. 30921012205.

7. REFERENCES

[1] Hirst, D. 2005. Form and function in the representation of speech prosody. *Speech Communication,* 46(3-4), 334-347.

[2] Eckstein, K., Friederici, A. 2006. It's early: event-related potential evidence for initial interaction of syntax and prosody in speech comprehension. *Journal of cognitive neuroscience*, 18(10), 1696-1711.

[3] Pihan, H., Tabert, M., Assuras, S., Borod, J. 2008. Unattended emotional intonations modulate linguistic prosody processing. *Brain and language*, 105(2), 141-147.

[4] Paulmann, S., Jessen, S., Kotz, S. 2012. It's special the way you say it: An ERP investigation on the temporal dynamics of two types of prosody. *Neuropsychologia*, 50(7), 1609-1620.

[5] Wang, T., Lee, Y., Ma, Q. 2018. Within and Across-Language Comparison of Vocal Emotions in Mandarin and English. *Applied Sciences*, 8(12), 2629.

[6] Yuan, J. 2006. Mechanisms of Question Intonation in Mandarin. *Lecture Notes in Computer Science,* 19–30.

[7] Prom-On, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The journal of the Acoustical Society of America*, 125(1), 405-424.

[8] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech communication*, 46(3-4), 220-251.

[9] Pell, M. 2001. Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, 109(4), 1668-1680.

[10] Zora, H., Rudner, M., Montell, A. 2020. Concurrent affective and linguistic prosody with the same emotional valence elicits a late positive ERP response. *European Journal of Neuroscience*, 51(11), 2236-2249.

[11] Scherer, K., Oshinsky, J. 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1(4), 331-346.

[12] Ross, E., Thompson, R., Yenkosky, J. 1997. Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. *Brain and language*, 56(1), 27-54.

[13] Seddoh, S. 2002. How discrete or independent are "affective prosody" and "linguistic prosody"?. *Aphasiology*, 16(7), 683-692.

[14] Peirce, J., Gray, J., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Lindeløv, J. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51(1), 195-203.

[15] Team, R. 2013. R: A language and environment for statistical computing.

[16] Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting linear mixed-effects models using lme4.

[17] Kuznetsova, A., Brockhoff, P., Christensen, R. 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82, 1-26.

[18] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R package version*, 1(1), 3.

[19] Grandjean, D., Bänziger, T., Scherer, K. 2006. Intonation as an interface between language and affect. *Progress in brain research*, 156, 235-247.

[20] Bryant, G., & Fox Tree, J. 2005. Is there an ironic tone of voice?. *Language and speech*, 48(3), 257-277.

[21] McRoberts, G., Studdert-Kennedy, M., Shankweiler, D. 1995. The role of fundamental frequency in signaling linguistic stress and affect: Evidence for a dissociation. *Perception & Psychophysics,* 57(2), 159-174.

BOUNDARY STRENGTH AND PREDICTABILITY EFFECTS ON DURATIONAL CUES AT TONE SANDHI GROUP BOUNDARIES IN TAIWAN SOUTHERN MIN

Sheng-Fu Wang

Institute of Linguistics, Academia Sinica, Taiwan
sftwang@gate.sinica.edu.tw

ABSTRACT

This study examined syllable duration at the right edge of Tone Sandhi Groups (TSGs) in an 8-hour spontaneous speech corpus in Taiwan Southern Min. TSG boundaries under three prosodic conditions were compared: matching with an intonational phrase boundary (TSG+IP), an intermediate phrase boundary (TSG+ip), and neither (TSG-only). We also examined whether the correlation between syllable duration and predictability (lexical frequency, surprisal, and informativity) was affected differently at these three boundary conditions. Results showed significant penultimate and final lengthening at TSG boundaries in all conditions, with longer pre-boundary syllables in the TSG+IP condition, followed by the TSG+ip and TSG-only conditions. Analyses also revealed that the positive correlation between syllable duration and surprisal/informativity was weaker towards all three types of boundaries. To summarize, TSG boundaries exhibited durational patterns previously reported for IP boundaries, with an incremental strengthening of final lengthening when overlapping with larger prosodic breaks.

Keywords: predictability, tone sandhi, Taiwan Southern Min, duration, boundary strength

1. INTRODUCTION

This study aimed to investigate the interaction between pre-boundary lengthening and probabilistic reduction based on predictability at the boundaries of Tone Sandhi Groups (henceforth TSGs) in Taiwan Southern Min. There are two specific research questions: First, do the pre-boundary durational patterns incrementally change as TSG boundaries match up with higher prosodic breaks? Second, does the interaction between predictability effects and pre-boundary durational variability also differ for TSG boundaries in different conditions?

Tone sandhi in Taiwan Southern Min refers to a phonological process where all syllables except for the rightmost one in a TSG switch their tones. The formation of TSGs has long been tied to syntactic configurations. Even though there are different accounts [4, 12], they share the assumption that right edges to certain syntactic constituents in some conditions trigger a TSG break.

TSG patterns in two example sentences adapted from [12, 4] are shown in (1), where TSG breaks are marked by the # symbol. Example (1a) shows how the right edges of the noun phrase ti^{44}-$\mathrm{ba}ʔ^{2}$ 'pork' and the quantifier phrase tsit^{4} kun^{44} 'one catty' are marked with TSG breaks, and the tone of the first syllable switches in the surface forms ti^{22}-$\mathrm{ba}ʔ^{2}$ and tsit^{2} kun^{44}. Interestingly, as shown in (1b), it is possible to have a whole sentence consisting of only one TSG, given the right configuration, i.e., the exceptionality of pronouns and phrases that are potentially TSGs being governed [12].

(1) a. $\mathrm{ti}^{44\to22}$-$\mathrm{ba}ʔ^{2}$# $\mathrm{tsit}^{4\to2}$ kun^{44}# $\mathrm{sã}^{44\to22}$ $\mathrm{k^{h}o}^{44}$#
pork one catty three dollars
'Pork is three dollars a catty (= 0.6 kg)'

b. $\mathrm{i}^{44\to22}$ $\mathrm{kjoŋ}^{24\to22}$-$\mathrm{kjoŋ}^{24\to22}$ $\mathrm{kjo}^{21\to53}$ $\mathrm{gwa}^{53\to44}$ $\mathrm{ke}^{44\to22}$ $\mathrm{k^{h}wã}^{21\to53}$ $\mathrm{pwã}^{21\to53}$ $\mathrm{tjam}^{53\to44}$-$\mathrm{tsiŋ}^{44\to22}$ $\mathrm{ku}^{53\to44}$ $\mathrm{ts^{h}e}ʔ^{2}$#
he by-force ask I more read half hour long book
'He insisted that I read for another half an hour.'

For the purpose of this discussion, it is important to note that TSGs can be identified by the tonal changes within a sentence, and their formation is likely governed by syntactic principles. Thus, any changes in TSG formation may result in alterations in meaning that reflect differences in syntactic

structure. Examples of this can be seen in (2) from [6, 12]: removing the TSG break by pronouncing the second syllable in mwa^{24}-a^{53} in the base tone results in 'sesame seed' becoming a modifier of the adjective twa^{22}, leading to a change in the phrase's meaning. In summary, TSGs have a distinct phonological definition that likely reflects the syntactic structure.

(2) a. $\mathrm{mwa}^{24\rightarrow22}$-$\mathrm{a}^{53}$# twa^{22}# $\mathrm{e}^{24\rightarrow22}$
sesames seed big 's
$\mathrm{sjo}^{44\rightarrow22}$-$\mathrm{pi\tilde{a}}^{53}$#
bun
'buns with big sesame seeds'

b. $\mathrm{mwa}^{24\rightarrow22}$-$\mathrm{a}^{53\rightarrow44}$ twa^{22}# $\mathrm{e}^{24\rightarrow22}$
sesames seed big 's
$\mathrm{sjo}^{44\rightarrow22}$-$\mathrm{pi\tilde{a}}^{53}$#
bun
'buns as big as sesame seeds' (i.e., tiny buns)

There has been disagreement over whether TSGs are a part of the prosodic hierarchy in Taiwan Southern Min, along with other levels such as syllable, word, intermediate phrase (ip), and intonational phrase (IP). While some studies, such as [16, 9, 10], treat TSGs as a level of prosodic unit below the IP and above the word, other studies and corpora, such as [15, 14, 23], exclude TSGs from these levels of prosodic annotation and treat ip as the level between IP and the word. The separation of TSGs from IP and ip mirrors how they are defined in different terms: As mentioned earlier, TSGs have a clear phonological definition that likely reflects morphosyntactic patterning. On the other hand, IP and ip are mostly defined by potentiality categorical perception of acoustic cues.

It is worth noting that TSG boundaries were shown to be more likely to align with larger prosodic boundaries (i.e., IP and ip, as compared with word and syllable) in a study that directly investigated the relationship between the distribution of TSGs and prosodic units [15]. The authors suggest that this finding indicates that a purely morpho-syntactic approach cannot fully explain the distribution of TSGs in natural language.

To complement the ongoing research on TSG and the prosodic structure in Taiwan Southern Min, we aimed to answer two questions in this study. Firstly, we investigated whether TSGs exhibit pre-boundary lengthening effects similar to those seen in intonational phrases. We went beyond the binary distinction between citation and sandhi tones reported in previous studies [8, 22] by examining syllable duration in final, penultimate, and ante-penultimate positions before a TSG boundary.

Secondly, we investigated whether overlapping with IP and ip boundaries influenced pre-boundary lengthening and its interaction with probabilistic reduction at TSG boundaries. Previous studies have reported that linguistic units with higher lexical frequency or contextual probability tend to be reduced in speech production. One issue related to this correlation is whether the relationship between acoustic cues and predictability is direct or mediated by the prosodic structure [2, 21, 13], and directly examining whether predictability effect covaries with the strength of the prosodic boundary has been used to answer this question [1]. We also explored whether the predictability effects are modulated as a function of a syllable's distance from a TSG boundary, which has been reported as a characteristic of IP boundaries in Taiwan Southern Min [22].

In summary, this study aimed to provide a more detailed description of durational patterning at TSG boundaries. Crucially, we sought to examine whether TSGs themselves exhibit patterns that are observed for higher prosodic units such as the IP, and how overlaps with different levels of prosodic units may change those durational patterns.

2. METHOD

2.1. Speech corpus

Speech data came from an 8-hour spontaneous speech corpus with recordings from 16 speakers. Each speaker contributed approximately 30 minutes of monologue-like speech from sociolinguistic interviews. The corpus included annotations of word segmentation, TSG breaks, and prosodic breaks. A 30-minute sample of the corpus was prosodically annotated by two annotators, who had a high agreement rate of 93.2%. For analysis, we focused on final, penultimate, and ante-penultimate positions in TSGs longer than three syllables. Our dataset included 10,934 TSGs consisting of 46,414 syllables (35,545 words).

2.2. Written corpus and language models

In this study, we estimated informativity and surprisal using trigram language models trained on a written corpus consisting of approximately 4.7 million words (from 6 million syllables). The corpus contained materials from various sources such as fiction, prose, poetry, conversation transcripts, and

academic writings, among others [7]. Trigram language models were trained in both directions using the SRILM toolkit [18, 19] with modified Kesner-Ney smoothing [5] to obtain smoothed bigram and unigram probabilities.

2.3. Variables in analyses

The statistical analyses were performed using mixed effects regression models with the LME4 package [3] in R. Three types of predictability variables were examined: **unigram surprisal** ($-\log P(x)$), which measured lexical frequencies; bigram surprisal ($-\log P(x|context)$), which measured local contextual predictability, and **bigram informativity** ($-\sum_{\text{context}} P(\text{context}|x)\log P(x|\text{context})$), the average of a word's surprisal across all contexts. Bigram surprisal and informativity were measured both in the forward and backward directions [17]. All these variables were log-transformed (base 10) and normalized.

A syllable's position within a TSG (initial/medial, ante-penultimate, penultimate, final) and the boundary's prosodic strength (TSG+IP, TSG+ip, TSG-only) were two other crucial variables. Their interaction was also included in the analysis.

Some controlling variables were included in the analysis: Speech rate (syllable count per second in each TSG), syllable count of the word a syllable occurs, the surface tonal category of a syllable, and each syllable's baseline duration. Following a similar method reported in [20], baseline duration was calculated from the predictions of a linear regression model trained to predict the duration of a syllable based on its component phonemes.

3. RESULTS

3.1. Full model statistics

In the full model, all variables except for unigram surprisal (β = 0.015, p = 0.23) and bigram forward informativity (β = 0.011, p = 0.4) significantly improved the model fit. The other three predictability variables had a positive estimate, indicating that higher surprisal/informativity was associated with longer syllable duration (backward informativity: β = 0.063, p < .001; forward surprisal: β = 0.018, p < .0001; backward surprisal: β = 0.015, p < .001).

There was also a significant interaction between boundary strength and prosodic position. Figure 1 shows the estimated syllable duration as a function of prosodic position and boundary type. Across different boundary types, there was significant final and penultimate lengthening (p < .0001 for all pairs of comparisons). The main difference among boundary types was the duration of the pre-boundary syllable: TSG+IP boundaries had the longest final syllable, followed by TSG+ip and TSG-only boundaries (p < .0001 in all pairs). The penultimate syllables at TSG+IP boundaries were also longer than the TSG-only counterparts (p < .001). In other words, having higher levels of prosodic breaks coinciding with TSG breaks did result in a significant difference in boundary cues.

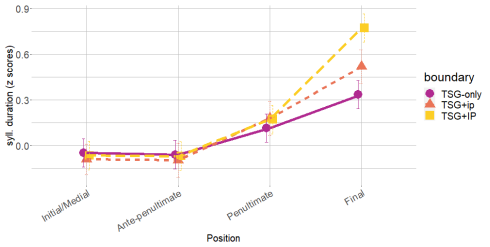


Figure 1: Syllable duration as a function of prosodic position (x axis) and boundary strength (color/shape). Error bars indicate standard errors.

As for the controlling variables, syllable duration was shorter with faster speech rates (β = -0.19, p < .0001), in longer words (β = -0.19, p < .0001), and with more phonological neighbors (β = -0.029, p < .05).

3.2. Predictability effects at different conditions

Next, we fitted a series of regression models that only contained one of the predictability variables so that their effect sizes had an interpretation free of issues such as suppression [24]. In these models, we also included the three-way interaction between the predictability variable, prosody position, and boundary strength. Post-hoc analyses were conducted using the lstrends() function in the lsmeans [11] package in R.

Figure 2 illustrates the interaction between informativity, prosodic position, and boundary strength. The effect of forward informativity (upper panel) was significant in all conditions except the final position for all three types of boundaries (p < .0001). **Backward informativity** (lower panel) was significant under the same conditions (TSG+ip at the ante-penultimate position: p < .001; all other non-final conditions: p < .0001).

The effect of **Unigram surprisal** is shown in the top panel of Figure 3. Similar to informativity, it was a significant factor for all but the final position across three boundary types (TSG-only: p < .0001; TSG+ip: p < .05 at initial/medial, p < .01 at ante-

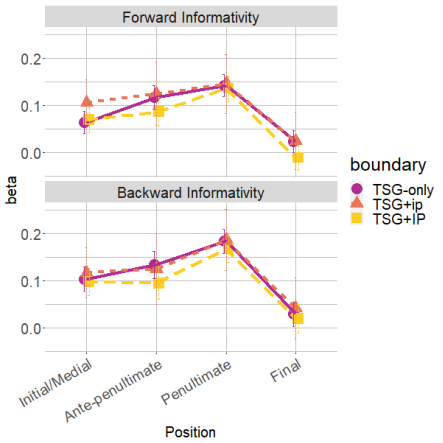


Figure 2: The effect size of **informativity** as a function of direction (panel), prosodic position (x axis) and boundary strength (color/shape). Error bars indicate standard errors.

penultimate, $p < .0001$ at penultimate; TSG+IP: $p < .0001$ at initial/medial and final, $p < .001$ at ante-penultimate).

The effect of **forward surprisal** (middle panel in Figure 3) was only significant at the penultimate position for TSG+ip ($p < .01$) and TSG+IP ($p < .0001$) breaks, but for all but the final position for TSG-only breaks (penultimate & ante-penultimate: $p < .0001$, initial/medial: $p < .001$). Finally, the **backward surprisal** effect (bottom panel) showed a three-way difference between prosodic conditions: significantly positive for all but the final position in the TSG-only condition (penultimate & ante-penultimate: $p < .0001$, initial/medial: $p < .01$), not significant at all in the TSG-ip condition, and significantly positive for all but negative at the final position in the TSG-IP condition (initial/medial: $p < .01$, penultimate/final: $p < .0001$). The negative effect suggests that syllables from words predicted to be frequent at the sentence/utterance-final position were actually lengthened at the IP boundary.

4. DISCUSSION & CONCLUSION

Results show that TSG boundaries exhibit lengthening of the penultimate and final syllables even without the boundary overlapping with an IP or ip boundary. In the overlapping scenarios, the duration of the pre-boundary syllable was significantly longer, especially when overlapping with an IP boundary. While this finding does not directly lead to a conclusion the TSG should be incorporated as a level of prosodic units alongside IP and ip, it does show that a linguistic domain defined purely on phonological terms and likely syntactically motivated may still exhibit similar acoustic cues at least in terms of durational patterns.

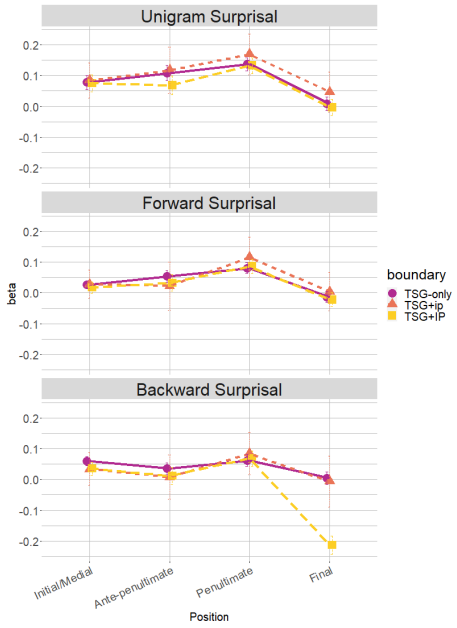


Figure 3: The effect size of **surprisal** as a function of surprisal type (panel), prosodic position (x axis) and boundary strength (color/shape). Error bars indicate standard errors.

TSG-only boundaries were also similar to TSG+IP and TSG+ip boundaries (as well as IP boundaries in general, as shown in [22]) in how the positive correlation between syllable duration and surprisal/informativity usually became neutralized towards a boundary, where final lengthening occurs. As discussed in previous studies, such neutralization is consistent with a view that prosodic modulates the relationship between predictability and acoustic cues, so that when boundary cues are already present, predictability effects become neutralized [21, 13]. Again, the current study shows that this type of interaction can also occur in a linguistic domain other than the IP. As for boundary strength, other than the reversal of backward surprisal's effect at TSG+IP boundaries, overlapping with larger prosodic boundaries did not affect the interaction between predictability and final lengthening at TSG boundaries.

To conclude, this study presents a detailed description of pre-boundary durational variability at TSG breaks in Taiwan Southern Min. We show how TSGs, defined entirely differently from prosodic units such as the IP and ip, still exhibit very similar boundary cues. Future studies with a more comprehensive survey on boundary cues across different types of syntactic units may further shed light on the issue of syntax-prosody interface and the status of TSGs as part of Taiwan Southern Min's prosodic system.

5. REFERENCES

[1] B. Andreeva, B. Möbius, and J. Whang, "Effects of surprisal and boundary strength on phrase-final lengthening," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 146–150.

[2] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and speech*, vol. 47, no. 1, pp. 31–56, 2004.

[3] D. Bates, M. Mächler, B. Bolker, and S. Walker, "2015. fitting linear mixed-effects models using lme4," *J. Stat. Softw*, vol. 67, no. 1, pp. 1–48, 2015.

[4] M. Y. Chen, "The syntax of xiamen tone sandhi," *Phonology*, vol. 4, pp. 109–149, 1987.

[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[6] R. L. CHENG, "Tone sandhi in taiwanese," vol. 6, no. 41, pp. 19–42, 1968. [Online]. Available: https://doi.org/10.1515/ling.1968.6.41.19

[7] U. Iunn, "Taiwanese corpus collection and corpus based syllable/word frequency counts for written taiwanese," 2005.

[8] C.-H. Kuo, "Perception and acoustic correlates of the taiwanese tone sandhi group," Ph.D. dissertation, UCLA, 2013.

[9] G. Kuo, "Perceived prosodic boundaries in taiwanese and their acoustic correlates," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] ——, "Processing prosodic boundaries in natural and filtered speech," in *Proceedings of the 7th International Conference on Speech Prosody*, 2014, pp. 983–986.

[11] R. V. Lenth, "Least-squares means: The R package lsmeans," *Journal of Statistical Software*, vol. 69, no. 1, pp. 1–33, 2016.

[12] J.-w. Lin, "Lexical government and tone group formation in xiamen chinese," *Phonology*, vol. 11, no. 2, pp. 237–275, 1994.

[13] Z. Malisz, E. Brandt, B. Möbius, Y. M. Oh, and B. Andreeva, "Dimensions of segmental variability: Interaction of prosody and surprisal in six languages," *Frontiers in Communication*, vol. 3, p. 25, 2018.

[14] H.-h. Pan and H.-t. Huang, "Lexical propensity and taiwanese min tone sandhi rules," in *Proceedings of the International Conference on Speech Prosody*, vol. 2020, 2020, pp. 518–522.

[15] H.-h. Pan, H.-t. Huang, and S.-r. Lyu, "The occurrence of Taiwanese Min juncture tones before prosodic boundaries and modification marker," in *Proceedings of the 19th International Congress of Phonetic Sciences*, vol. 2019, 2019, pp. 3423–3427.

[16] S.-h. Peng and M. E. Beckman, "Annotation conventions and corpus design in the investigation of spontaneous speech prosody in taiwanese," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[17] S. Seyfarth, "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," *Cognition*, vol. 133, no. 1, pp. 140–155, 2014.

[18] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[19] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proceedings of IEEE automatic speech recognition and understanding workshop*, vol. 5, 2011.

[20] K. Tang and R. Bennett, "Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan)," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 997–1017, 2018.

[21] A. Turk, "Does prosodic constituency signal relative predictability? a smooth signal redundancy hypothesis," *Laboratory phonology*, vol. 1, no. 2, pp. 227–262, 2010.

[22] S.-F. Wang, "The interaction between predictability and pre-boundary lengthening on syllable duration in taiwan southern min," *Phonetica*, vol. 79, no. 4, pp. 315–352, 2022. [Online]. Available: https://doi.org/10.1515/phon-2022-0009

[23] S.-F. Wang and J. Fon, "A Taiwan Southern Min spontaneous speech corpus for discourse prosody," in *The Proceedings of Tools and Resources for the Analysis of Speech Prosody, Aix-en-Provence, France*, 2013, pp. 20–23. [Online]. Available: http://www.lpl-aix.fr/~trasp/Proceedings/20427-trasp2013.pdf

[24] L. H. Wurm and S. A. Fisicaro, "What residualizing predictors in regression analyses does (and what it does not do)," *Journal of memory and language*, vol. 72, pp. 37–48, 2014.

REJECT?! ON THE PROSODY OF NON-ACCEPTANCE

Sophie Repp, Heiko Seeliger

Department of German Language and Literature I, University of Cologne
sophie.repp | heiko.seeliger@uni-koeln.de

ABSTRACT

This production study examines two rejecting speech acts (ReA) in German: plain rejections and rejecting questions (RQs), which encode speaker uncertainty regarding the rejecting act. In comparison to assertions, the ReA show increased prosodic prominence of the nuclear accent but no specialized contours signalling disagreement as observed in other languages [24]. Rejections have a falling final contour. RQs come with two different rising contours known from other questions but may also end in a fall. Both illocutionary meaning components—non-acceptance and uncertainty—have gradient prosodic effects.

Keywords: prosody, rejection, rejecting question, disagreement, German

1. INTRODUCTION

Rejections and certain checking questions express that the speaker *S* does not wish to accept the truth of a proposition *p* which is associated with a preceding utterance by an interlocutor. Rejections express that *S* will not accept *p* into the common ground (CG). Rejecting questions leave room for the possibility that *S* might err, so that it is uncertain if *p* should not be accepted into the CG. To illustrate, when *S* hears, *Benny is thinking about buying raffle tickets*, *S* might conclude that Benny wants to buy raffle tickets (= *p*). *S* might disagree and say, *Benny doesn't want to buy raffle tickets!*, thus rejecting *p*. Alternatively, *S* might hear, *Benny mentioned spending his money on the fair.*, and ask, *Surely, he doesn't want to buy raffle tickets?!*, thereby suggesting to reject *p*.

Non-acceptance or, more generally, disagreement has prosodic reflexes, which have been discussed both in relation to specific intonation contours [24], and in relation to specific utterance types. Rejections have been suggested to come with the so-called *contradiction contour* (CC) [17, 19, 24], which in English is a low rise contour that is preceded by a fall before the final low pitch accent. It has received various transcriptions, e.g. %H (L*) L* L-H% [11]. The CC occurs with full-clause rejections of negative or positive polarity in English [11, 19], and also with rejecting uses of response particles like Catalan *sí* 'yes' [10]. However, as [23] point out, in English the CC not only rejects but also indicates that the addressee should already be aware of the truth of the speaker's utterance. Relatedly, in Catalan the CC can be used to 'state the obvious' [1, 8]. In fact, the CC seems to be generally ambiguous [23].

In German, the prosody of rejection has been studied for utterances with VERUM focus [13], which reject a conjecture or previous assertion by highlighting the truth of the current utterance, and for corrections with narrow focus. VERUM focus is marked by an accent on the finite verb in an overall falling contour in German. The accent has been reported to be (^)H* L- in utterances rejecting negative assertions [32], and L*+H in negative utterances rejecting a conjecture [30]. In corrections, the narrowly focused element is marked by the more frequent occurrence of more prominent pitch accents (L+H*), higher maximum F0, higher F0 excursion and greater duration in comparison to non-corrective narrow focus, and by reduced accentuation in the prenuclear region [3, 21].

Regarding non-accepting questions, previous research has described various *incredulity contours* (IC) in different languages, which have been proposed to mark disagreement [24]. In English the IC is a rise-fall-rise contour (L*+H L-H%) [11, 17, 34], which, depending on phonetic details, may also mark uncertainty [33] or lack of commitment [34] in a more general sense. Dutch uses the same contour to mark incredulity, while Catalan uses a neutral question contour (L* HH%) but with expanded F0 range [7]. In Puerto-Rican Spanish, incredulity is marked by a fall-rise-fall (L*HL%) [2].

In German, non-acceptance in questions has been studied in indignant echo questions, which express incredulity, and in various surprised questions. German echo questions come with the sharply rising final L* H-^H% contour familiar from information-seeking polar questions (which may also end in a fall [14, 16]). In echo questions, the L* accent is on the in-situ *wh*-word [25]. Indignant echo questions show a lower L* target, a steeper and higher rise and longer duration of the *wh*-word, and have higher prenuclear pitch than echo questions that are asked for other reasons (e.g., noisy environment, missing new infor-

mation) [25]. Surprised declarative and polar questions also have a later and higher final rise than their information-seeking counterparts [15] and a slower speaking rate [21].

In this paper we explore the prosody of two rejecting speeach acts (ReA) in German: rejections and rejecting questions (RQs) [26, 27, 29]. Both ReA reject a conjecture that suggests itself as an implication of an utterance by the interlocutor. Rejections plainly reject the conjecture; RQs signals speaker uncertainty regarding the rejecting act. In addition, both ReA indicate that the interlocutor should have been aware that the implied conjecture is false, which is a discourse type that does not license VERUM focus. The meaning components just described are marked transparently by modal particles (MPs). The rejections in our study contain the MP *doch*, which signals a conflict with a previous utterance and reminds the interlocutor that they should be aware of the conflict because the CG contains information that is incompatible with their previous utterance [31]. The RQs contain the MP combination *doch wohl*, which unambiguously marks the illocutionary force of this question type; *wohl* typically signals speaker uncertainty [35].

We compare the ReA with assertions. This three-way comparison enables us to examine the prosodic contributions of (i) the rejecting meaning component combined with the signal that the interlocutor should be aware of the truth, and (ii) the uncertainty (interrogative) meaning component. In a previous prosodic study of RQs in Swedish [28], (negative) RQs were compared only to (negative) rejections, and were found to exhibit the regular characteristics of question marking in Swedish.

2. METHOD

We tested utterances without negation, see (1), which were embedded in dialogue contexts (abbreviated here). There were three conditions: the target utterance was an assertion (As), a rejection (R) or a RQ.

(1) Speaker 2: *[...] A lucky fellow like Benny has many options in the fair, right?* (As) || *It looks like Benny wants to try out all the roundabouts in the fair.* (R, RQ).

Speaker 3 (= participant):

As: Benny will da dann Lose kaufen.
B. wants there then raff.t. buy

R: Benny will doch dann Lose kaufen!

RQ: Benny will doch wohl Lose kaufen?!

'Benny wants to buy raffle tickets (there then / then, as you know / surely).'

In the dialogues, three speakers chat about friends. Speakers 1 and 2 (prerecorded; presented in written and auditory form) set the scene, and speaker 2 either asks a fairly general tag question *p?* (condition As), or utters a supposition which allows the conjecture that a proposition *p* concerning the plans of a friend is false (R, RQ). Speaker 3 (= participant) then reacts to the question *p?* in an utterance containing adverbs instead of MPs (As), rejects the conjecture by declaring the obvious truth of *p* in a *doch*-utterance (R), or suggests the obvious truth of *p* in a *doch-wohl* utterance (RQ); followed by an assertion. All target utterances have broad VP focus, which in the R and RQ conditions is corrective.

18 native speakers of German (5 male; aged 19-32, mean: 25.2) participated in the experiment. They were students of the University of Cologne at the time of the experiment and were either paid for their participation or received course credit.

The recordings were annotated in Praat [4] by trained research assistants (GToBI guidelines [12]). The raw data was imported into R using rPraat [5].

3. RESULTS

Rejections and assertions were realized with falling contours except for one rejection with the IC familiar from English. Most RQs (64%) also had falling contours, 30% were rising, the remainder were mostly incomplete falls after nuclear [L+]H* accents, or late falls (≈ H-L%). There was considerable inter-individual variation in the RQs. 4 participants produced only falling RQs, 1 produced only rising RQs; 13 produced primarily falling, and 5 primarily rising RQs. See Fig. 1 for averaged normalized contours.

In the statistical analysis, we fitted two (generalized) linear mixed models for each dependent variable: one with assertions as baseline, comparing the two ReA separately, and one comparing the two ReA with each other. We report Holm-Bonferroni-adjusted *p*-values. Significance levels are as follows: $^{***} = p < 0.001$, $^{**} = p < 0.01$, $^{*} = p < 0.05$.

Utterance-level measures. The speaking rate was the same in all three speech acts. Mean intensity was higher in rejections than in assertions ($b = 0.5, SE = 0.17, t = 2.9^{*}$), and in RQs than in assertions, but not significantly so. Intensity range was highest in assertions and lowest in RQs, with rejections intermediate. Only the difference between RQs and assertions was significant ($b = -0.92, SE = 0.26, t = -3.5^{**}$). F0 range was higher in both ReA than in assertions (R: $b = 2.2, SE = 0.35, t = 6.2^{***}$; RQ: $b = 2.3, SE = 0.34, t = 6.7^{***}$). Mean F0 was higher in rejections than in assertions

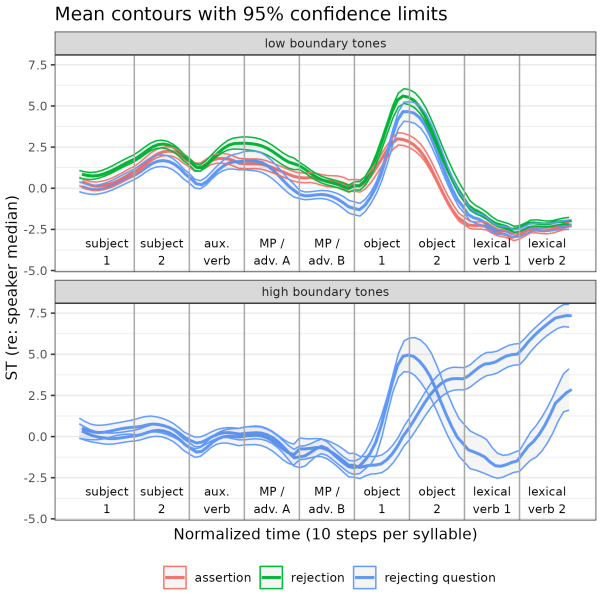


Figure 1: Average contours split by speech act and boundary tone height. High boundary tones are further split into low rises ([L+]H* L-H%) and high rises (L* H-^H%).

($b = 0.66, SE = 0.15, t = 4.4^{**}$), and it was non-significantly higher in RQs than in assertions.

Accent placement and accent types. Fig. 2 shows the distribution of accentuation and accent types across syllables in all speech acts. Regarding accent placement, we found that the subject was accented less often in the ReA than in assertions (R: $b = -1.5, SE = 0.52, z = -2.9^{**}$; RQ: $b = -1.7, SE = 0.52, z = -3.2^{**}$). Since the object was accented extremely often across conditions (As: 100%; R/RQ: > 95%), we could not fit a model. The auxiliary was accented non-significantly more often in the ReA than in assertions. All other syllables were accented too rarely to fit a model.

Regarding accent types, the proportion of L+H* accents (amongst all accents) was higher on the object in the ReA than in the assertions (R: $b = 1.7, SE = 0.3, z = 5.3^{***}$; RQ: $b = 2.1, SE = 0.3, z = 6.6^{***}$). The proportion of L* accents was lower on the subject in the ReA than in assertions (R: $b = -0.9, SE = 0.2, z = -3.6^{***}$; RQs: $b = -1.1, SE = 0.24, z = -4.4^{***}$). L* accents on the object only occurred in RQs. H+!H* only occurred on the object in assertions (usually as the second accent in a hat pattern), while !H* only occurred on the object in the ReA and usually followed a very prominent accent on the auxiliary, i.e. !H* accents were arguably post-nuclear pitch accents.

The rising contours in RQs were unevenly split between those with L* vs. (L+)H* nuclear accents: L* occurred in 52 utterances (48 followed by H-^H%),

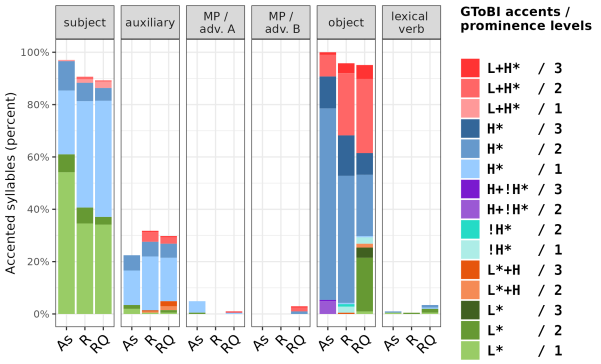


Figure 2: GToBI accent types and DIMA prominence level combinations per syllable and condition

and (L+)H* in 135 (14 followed by L-H%; 11 by one participant). The other (L+)H* utterances either had low boundary tones or incomplete falls.

Syllable/word-level measures. For space reasons we do not report numerical details for these measures here (see `osf.io/xerfg`). Fig. 3 shows the relative duration (foot, i.e. both syllables) and mean intensity for subject, object and lexical verb. Fig. 4 shows the F0 measures. We carried out separate analyses for objects with H* accents vs. objects with L+H* accents. Table 1 presents all significant differences (at least $p < 0.05$) and effect directions.

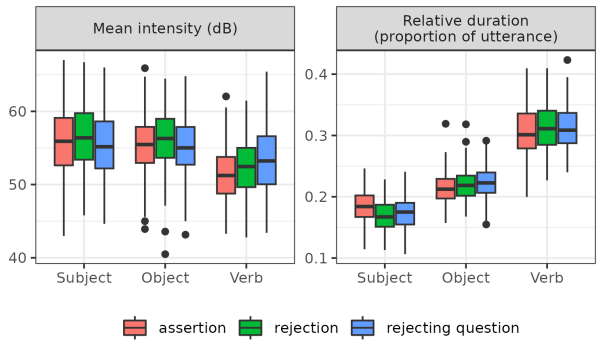


Figure 3: Mean intensity (accented syllable) and relative duration (word) for subject, object and lexical verb.

4. DISCUSSION

Our results indicate that the meaning components rejection and uncertainty/interrogativity have prosodic reflexes in ReA. For the rejections we did not observe a specific CC but there are many differences with assertions. RQs share characteristics with other German questions but also with rejections.

Overall, both ReA show an increased prominence of the nuclear accent on the object in comparison to

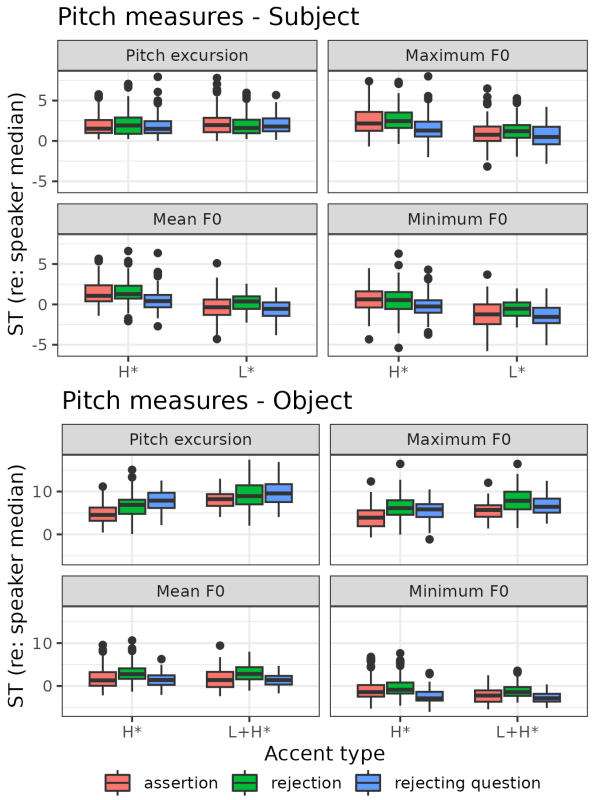


Figure 4: Pitch-related measures for subject and object, split up by speech act and accent type

assertions, with more L+H* accents, higher maximum F0, greater F0 excursion for H* accents and longer duration. These effects are known from narrow object focus in corrections, but since in the current ReA there is no narrow focus but broad VP focus, the effects are more likely to be an illocutionary effect of the rejecting meaning component, as might be the longer duration and higher intensity in the post-nuclear region.

While utterance-level pitch range was higher in both ReA (which arguably is driven by the higher nuclear pitch accent in rejections and the high boundary tone in RQs), mean utterance-level pitch was higher only in rejections. This most likely results from the overall higher prenuclear pitch, which also shows in the higher pitch of L*-accented subjects in rejections. This observations is unfamiliar from work on the prosody of disagreement but accords with findings for emotional arousal [6, 18]: higher mean pitch is associated with higher emotional arousal ([25] on indignant echo questions). A rejection arguably is a face-threatening act, and is thus plausibly associated with higher emotional arousal. Just suggesting a rejection with a RQ might be less face-threatening, so that the emotional arousal is lower than in rejections, as is indicated by the lack of higher prenuclear pitch.

Table 1: Summary of significant differences for syllable/word-level measures of accented subjects (S), of accented objects in falling contours (O), of non-accented verbs (V); for F0 split by accent type

		R:As	RQ:As	RQ:R
S	duration	<	<	>
	intensity	>	-	<
S: H*	$F0_{exc}$	-	-	-
	$F0_{max}$	-	<	<
	$F0_{min}$	-	<	<
	$F0_{mean}$	-	<	<
S: L*	$F0_{exc}$	-	-	-
	$F0_{max}$	>	<	<
	$F0_{min}$	>	-	<
	$F0_{mean}$	>	<	<
O	duration	>	>	>
	intensity	>	-	<
O: H*	$F0_{exc}$	>	>	>
	$F0_{max}$	>	>	-
	$F0_{min}$	-	<	<
	$F0_{mean}$	>	-	<
O: L+H*	$F0_{exc}$	-	-	-
	$F0_{max}$	>	-	<
	$F0_{min}$	>	-	<
	$F0_{mean}$	>	-	<
V	duration	>	>	-
	intensity	>	>	>

RQs are also marked as distinct from rejections by two rising contours: The L* H-^H% contour is familiar from other indignant and surprised questions [15, 25], and from information questions [12, 20]. The (L+)H* L-H% contour has been suggested to mark the expectation of a quick, non-elaborating *yes-no*-answer [22] . Intuitively, this fits the illocution of RQs: the addressee is reminded that they should agree with the speaker. The contour also has been suggested to mark polite offers [9, 12], which is incompatible with the illocution of RQs. Falling RQs show only subtle differences with rejections: lower minimum and mean F0, which might be an effect of the arguably lower emotional arousal of RQs.

In sum, the rejecting and the uncertainty meaning components have many prosodic reflexes, although rejections and RQs are not always clearly distinguished. Overall, the prosody of non-acceptance seems to depend on specific discourse conditions, as is reflected in differences with earlier findings on other types of rejections and disagreeing questions.

Acknowledgements: Funding: DFG Deutsche Forschungsgemeinschaft (German Research Foundation), Project-ID 281511265, SFB 1252. Research assistants: Franziska Busche, Lukas Kurzeja, Karoline Marliani, Marlon Siewert, Sven Weber.

5. REFERENCES

[1] Aguilar, L., De-la Mota, C., Prieto, P. 2009. Cat_ToBI training materials. http://prosodia.upf.edu/cat_tobi/.

[2] Armstrong, M. E. 2017. Accounting for intonational form and function in Puerto Rican Spanish polar questions. *Probus* 29, 1–40.

[3] Baumann, S., Grice, M., Steindamm, S. 2006. Prosodic marking of focus domains – categorical or gradient? Hoffmann, R., Mixdorff, H. (eds), *Proc. Speech Prosody 2006* Dresden, 301–304.

[4] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer.

[5] Bořil, T., Skarnitzl, R. 2016. Tools rPraat and mPraat. Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds), *Proc. Text, Speech, and Dialogue 19* Brno, 367–374.

[6] Bänziger, T., Scherer, K. R. 2005. The role of intonation in emotional expressions. *Speech Commun.* 46, 252–267.

[7] Crespo-Sendra, V., Kaland, C., Swerts, M., Prieto, P. 2013. Perceiving incredulity: The role of intonation and facial gestures. *J. Pragmat.* 47(1), 1–13.

[8] Espinal, M. T., Prieto, P. 2011. Intonational encoding of double negation in Catalan. *J. Pragmat.* 43(9), 2392–2410.

[9] Féry, C. 1993. *German Intonational Patterns*. Tübingen: Niemeyer.

[10] González-Fuente, S., Tubau, S., Espinal, M. T., Prieto, P. 2015. Is there a universal answering strategy for rejecting negative propositions? Typological evidence on the use of prosody and gesture. *Front. Psychol.* 6(899).

[11] Goodhue, D., Wagner, M. 2018. Intonation, yes and no. *Glossa* 3(1), 1–45.

[12] Grice, M., Baumann, S., Benzmüller, R. 2005. German intonation in autosegmental-metrical phonology. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: OUP, 55–83.

[13] Höhle, T. 1992. Über Verum-Fokus im Deutschen. In: Jacobs, J. (ed), *Informationsstruktur und Grammatik*. Westdeutscher Verlag, 139–197.

[14] Kohler, K. J. 2004. Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In: Fant, G., Fujisaki, H., Cao, J., Xu, Y. (eds), *From Traditional Phonology to Modern Speech Processing*. Foreign Language Teaching and Research Press, 205–215.

[15] Kohler, K. J. 2013. From communicative functions to prosodic forms. *Phonetica* 70, 24–65.

[16] Kügler, F. 2003. Do we know the answer? Variation in yes-no-question intonation. In: Fischer, S., van de Vijver, R., Vogel, R. (eds), *Experimental Studies in Linguistics 1*. Potsdam: Potsdam University Press, 9–29.

[17] Ladd, D. R. 1980. *The structure of intonational meaning: Evidence from English*. Indiana UP.

[18] Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., Scherer, K. R. 1985. Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *J. Acoust. Soc. Am.* 78, 435–444.

[19] Liberman, M., Sag, I. 1974. Prosodic form and discourse function. *Proc. CLS 10* Chicago, 416–427.

[20] Michalsky, J. 2017. *Frageintonation im Deutschen. Zur intonatorischen Markierung von Interrogativität und Fragehaltigkeit*. de Gruyter.

[21] Niebuhr, O., Bergherr, J., Huth, S., Lill, C., Neuschulz, J. 2010. Intonationsfragen hinterfragt: die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax. *Z. Dialektol. Ling.*, 304–346.

[22] Peters, J. 2018. Phonological and semantic aspects of German intonation. *Ling. Online* 88(1), 85–107.

[23] Pierrehumbert, J., Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P. R., Morgan, J. L. (eds), *Intentions in Communication*. MIT Press, 271–311.

[24] Prieto, P., Espinal, M. T. 2020. Negation, prosody and gesture. In: Déprez, V., Espinal, M. T. (eds), *The Oxford Handbook of Negation*. OUP, 677–693.

[25] Repp, S., Rosin, L. 2015. The intonation of echo wh-questions. Möller, S., Ney, H., Möbius, B., Nöth, E., Steidl, S. (eds), *Proc. Interspeech 2015*.

[26] Seeliger, H. 2015. "Surely that's not a negative declarative question?" Polar discourses in Swedish, German and English. Csipak, E., Zeijlstra, H. (eds), *Proc. Sinn und Bedeutung* Göttingen, 571–589.

[27] Seeliger, H. 2019. *Swedish and German Rejecting Questions. Experimental Investigations of Question Bias*. PhD thesis Humboldt-Universität zu Berlin.

[28] Seeliger, H., Repp, S. 2017. On the intonation of Swedish rejections and rejecting questions. Abrahamsen, J. E., van Dommelen, W. A., Koreman, J. (eds), *Proc. Nordic Pros. XII* Trondheim, 135–146.

[29] Seeliger, H., Repp, S. 2018. Biased declarative questions in Swedish and German: The syntax of negation meets modal particles (väl and doch wohl). Dimroth, C., Sudhoff, S. (eds), *The Grammatical Realization of Polarity Contrast: Theoretical, empirical, and typological approaches*, 129–172.

[30] Seeliger, H., Repp, S. 2023. Information-structural surprises? Contrast, givenness and (the lack of) accent shift and deaccentuation in non-assertive speech acts. *Laboratory Phonology* 14(1), 1–46.

[31] Thurmair, M. 1989. *Modalpartikeln und ihre Kombinationen*. Max Niemeyer Verlag.

[32] Turco, G., Braun, B., Dimroth, C. 2014. When contrasting polarity, the Dutch use particles, Germans intonation. *J. Pragmat.* 62, 94–106.

[33] Ward, G., Hirschberg, J. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 747–776.

[34] Ward, G., Hirschberg, J. 1986. Reconciling uncertainty with incredulity: A unified account of the L*+H L H% intonational contour. *Paper presented at the LSA Annual Meeting* New York.

[35] Zimmermann, M. 2004. Zum Wohl: Diskurspartikeln als Satztypmodifikatoren. *Ling. Berichte*, 253–286.

CONTRAST OR CONTEXT, THAT IS THE QUESTION

Francesco Cangemi[1], Martine Grice[2], Hae-Sung Jeon[3], Jane Setter[4]

[1,2]University of Cologne, [3]University of Central Lancashire, [4]University of Reading
fcangemi@uni-koeln.de, martine.grice@uni-koeln.de, HJeon1@uclan.ac.uk, j.e.setter@reading.ac.uk

ABSTRACT

In the past decades, most research on question intonation has relied on controlled experimental tasks for the production and perception of syntactically well-formed, pragmatically or syntactically ambiguous, unbiased and genuine requests for new information.

The exploration of a real-life interaction in East Midlands British English shows that such questions are neither frequent nor typical. Prosodic information does undoubtedly contribute to the identification of questions, as shown by comparing questionhood ratings, based on either audio recordings or orthographic transcriptions only. However, compared to the role of syntactic, lexical, contextual, sequential, and interactional information, the role of speech prosody is shown to be marginal.

These findings suggest that prosodic information is difficult to isolate or to manipulate independently. Therefore, question intonation can only be adequately studied by taking into account naturalistic, context-rich, spontaneous behaviour.

Keywords: Question, Intonation, Ecological validity

1. INTRODUCTION

Questions are ubiquitous in human life, from the private negotiations of early childhood to the world-changing endeavours of scientific research. Unsurprisingly, questions are also crucial in linguistics. They are at the crossroads of all areas of analysis (from phonetics to discourse), of all perspectives of inquiry (from typology to variationism), and of all contexts of application (from technology to education). As such, in the last century questions have attracted a large share of linguists' interest and have consequently been studied using a variety of methods. The evidence used for theorising about questions thus includes data from introspection (e.g. linguists' intuitions), from observation (e.g. records of naturally occurring interactions) and from experimentation (e.g. controlled elicitation in production and perception). Naturally, the weighting of these three types of evidence varies across areas, perspectives, and contexts of study.

In recent decades, linguists specifically interested in the phonetic and phonological aspects of questions seem to have favoured the use of evidence based on experimentation. This preference is not only implicitly attested in actual research practices (e.g. in Speech Prosody 2020, 90% of papers on questions relied on controlled elicitation in the laboratory context), but also explicitly voiced in publications on linguistic methodology. For instance, [1] emphasised how experiments allow researchers to control for confounding factors, isolate dimensions of interest, devise pertinent minimal pairs, efficiently elicit many such pairs, and thus explore both the formal and functional aspects of linguistic contrasts.

In this frame, the success of experimental methods is inherently linked to the notion of contrast, which has played a crucial role in modern linguistics since its earliest days [2:166]. However, equally early work suggests that linguistic patterns can only be understood when evaluated within their larger communicative context, including an exploration of everyday interactions [3:5]. This is particularly true for contrasts that involve meaning beyond the lexical level, where context plays a crucial role [4], as in the case of questions. Therefore, it is possible that by enforcing experimental control to maximise contrast, one might impoverish context to the point of altering the very nature of the observed phenomena. As a result, it has been suggested that question intonation and other prosodic phenomena should be studied via the observation of naturalistic, context-rich, spontaneous behaviour [5].

A synthetic approach is offered by [6] who support the integration of different types of evidence. This is akin to the program of cognitive ethology, according to which the "experimental simplification of a real-world situation is a reasonable research tactic when it *follows* careful real-world investigation" [7:337, original emphasis]. By giving temporal precedence to the observation of non-experimentally constrained interactions, the researcher can identify the most frequent and potentially most salient phenomena. These can then be used as cornerstone for further theoretical development to remain grounded in the interactionally relevant facts [8].

In this paper, we therefore set out to explore to what extent the questions typically employed in controlled experimental studies are representative of the questions we observe in real-world interactions. After providing a working definition of such typical questions (§2.1), we introduce the corpus (§2.2) and

the labelling strategies (§2.3) used in this study. We then illustrate the distribution of different types of questions in our corpus (§3.1), the contribution of intonation to raters' confidence (§3.2) and a comparison of ratings based on different strains of information (§3.3). Then we interpret these findings as suggesting that, in real-world interactions, prosodic and non-prosodic information is intertwined in the production and perception of questions.

2. METHOD

2.1. Stipulations

Most studies agree with [9] in considering as typical a certain set of functional and formal features of questions used in production and perception experiments. Functional features include "the speaker emits a request for information" (unlike rhetorical questions) and "the speaker has no bias towards an expected answer" (unlike confirmation requests). Formal features include "the question can be uttered independently" (unlike tag questions) and "the question is syntactically well-formed" (unlike elliptical questions). Given the emphasis on contrastive analyses discussed in §1, these questions are usually compared to less typical items, including those which are not marked by particles or morphosyntactic devices (like declarative questions).

While more nuanced treatments of typicality in questionhood are available [10, 11], the characterisation above seems to fit at least the studies from the Speech Prosody 2020 conference mentioned above. Many such studies featured the production of read speech, with questions such as *Yǒurén chī níngméng me?* (Mandarin Chinese, "Does anyone eat lemons?") [12], *Kas keegi sööb sellerit?* (Estonian, "Does anyone eat celery?") [13], *Che cosa ti volevano servire?* (Salerno Italian, "What did they want to serve you?") [14, 15]. Question sentences were usually uttered after silently reading a contextualisation paragraph, which is assumed to prompt a variety of interpretations (e.g. unbiased new information seeking question or rhetorical question).

In the following, we therefore investigate the occurrence of syntactically well-formed, pragmatically or syntactically ambiguous, unbiased, genuine requests for new information.

2.2. Corpus

We analysed excerpts from an interaction lasting 2 hours and 15 minutes between two native speakers of an East Midlands variety of British English. The participants played an online adaptation of a miniature wargame [16] and discussed issues of game design after the match. *Zoom* [17, 18] was used to record separate audio tracks for the two participants, which in the following will be referred to as L(eft) and R(ight). *Open Broadcaster Software* [19] was used to record the gameplay video feed with the conversation as a single audio track. This video recording was subsequently edited and published by one of the participants [20].

The interaction was natural, in the sense of "'non-experimental', not co-produced with or provoked by the researcher" [21:530]. The first author provided technical assistance with the recordings, but the interaction would have taken place independently of this research. In fact, the two players L and R had independent and intrinsic motivations for the interaction. L and R were expected to play the match as part of an online tournament, and R intended to upload footage of the match to his *YouTube* channel. Since R is one of the co-developers of the wargame, he also intended to gather feedback from L concerning issues of game design and balance. After the recording, participants were asked for permission to use the recording for research purposes. They agreed to this and further noted that the interaction is publicly available on *YouTube*.

From the *Zoom* recording we extracted two excerpts of approximately 15 minutes. The first excerpt [20, 1:18:51-1:35:25] contains a game turn played by L and can be analysed as a structured goal-oriented interaction. Typically, L initiates game actions (e.g. declaring the target of an attack) and R reacts to these actions in a dynamic decision-making process. Both players also exchange other static information (e.g. properties of game pieces as written in the rulebook) and comment on unpredictable events (e.g. virtual dice rolls). The second excerpt [20, 2:04:28-2:19:37] contains a discussion between L and R on issues of game design and can be equated with a casual conversation between friends on a topic of mutual interest. The interaction is slightly asymmetrical, since R probes L's opinions in order to amend the game's rulebook based on this feedback. This difference in the interactional scope of the two excerpts was expected to maximise the range of types of observable questions. The corpus size was kept below 30 minutes to ensure the possibility of a thorough analysis.

2.3. Ratings

A research assistant segmented the audio recordings into interpausal units using *Praat* [22] and provided an orthographic transcription which was revised by the first author. Each excerpt was then segmented into 4 portions, which were each assigned to one of the remaining authors and to a fifth collaborator. All raters were phoneticians trained in prosodic analysis.

In the first annotation stage, the raters provided a coarse annotation by listening to the audio recording, reading the transcription, and flagging interpausal units which contained possible instances of questions. At this stage, we followed [10] and left annotation criteria intentionally vague. Raters were asked to monitor utterances which sounded like a question (i.e. prosodically), had the structure of a question (i.e. syntactically), or functioned like a question (i.e. interactionally).

In the second annotation stage, the interpausal units flagged as potentially containing a question were submitted again to the raters for a finer evaluation. For each item, we extracted the orthographic transcription and the audio signal of the interpausal unit of the pre-item context (up to 1.5 seconds before the item beginning) and of the post-item context (up to 1.5 seconds after the item end). This information was combined into three stimulus versions. Version (a) contained only the orthographic transcription of the item. Version (b) contained both the orthographic transcription and the audio recording of the item. Version (c) contained the orthographic transcription of the context and of the item, the audio recording of the pre-item and of the post-item context, but not of the item itself.

For each item, the three versions were presented for evaluation to three different raters, thus excluding the rater who flagged the item in the first annotation stage. In this way each rater read or listened to every item extracted from the corpus. In the second stage, annotation criteria were made explicit and approximated the definition provided in §2.1 above. Raters were asked to judge whether the item represented a "Core question". This label was used to indicate an utterance in which the speaker "expresses an unbiased knowledge gap concerning new information, assumes that the listener is able and willing to fill this gap, directs the listener's attention to the gap, and invites the listener to fill it with a response". Responses ranged from 1 (definitely not a Core question) to 9 (definitely a Core question), and were collected using a *Praat* script. The annotation files and the scripts for the extraction and presentation of stimuli are available in an online repository [23].

3. RESULTS

3.1. Typical questions are not typical

After transcription and segmentation, the two extracts contained 1147 interpausal units. In the first annotation stage, 140 of such units were rated as potentially containing a question. In the second rating stage, 122 of such units received a score of less than 8 ("the item is a Core question") after being presented either with prosodic information (§2.3b) or with contextual information (§2.3c). This finding suggests that, at least from a distributional point of view, Core questions are not the most typical questions.

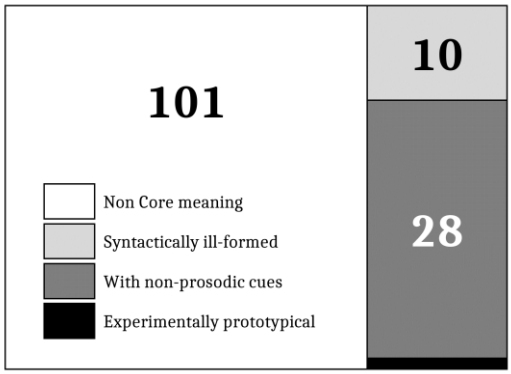


Figure 1: Core questionhood ratings.

We expanded the inclusion criteria to items with a rating of at least 7 ("probably a Core question"), thus filtering out only 101 cases (see Figure 1). Out of the remaining 39 items, 10 were not syntactically well-formed, and thus did not fit the typical features provided in §2.1. Of the remaining 29 items, 28 presented non-prosodic cues to questionhood, such as verb inversion or interrogative words. Only 1 item was rated as a Core question in the absence of non-prosodic marking. This finding raises the issue of how important intonation is in the identification of questionhood in natural interactions.

3.2. Prosodic information is important

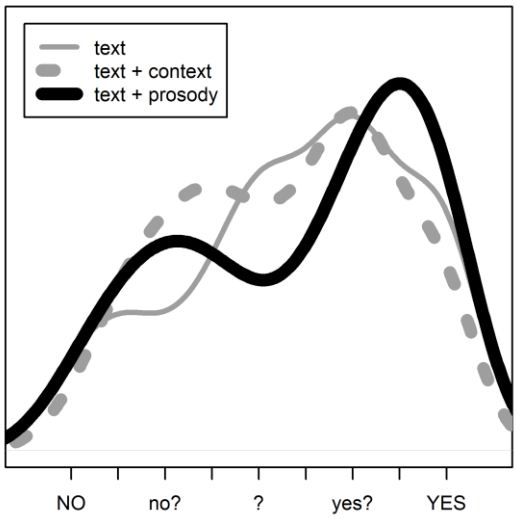


Figure 2: Questionhood ratings by type of stimulus.

We therefore explored how prosodic information contributes to the ratings, by visualising raters' responses separately for the three strands of information. In Figure 2, the thin grey line represents kernel density estimates [24] for responses to stimuli

presented with orthographic transcription (§2.3a), the thick dashed grey line is used for stimuli with contextual information (§2.3c) and the thick black line for stimuli with prosodic information (§2.3b).

Compared to the grey density curves, the black curve for responses to stimuli with prosodic information shows greater separation between its two peaks. In other words, when raters were provided with prosodic information, they were more confident in their judgement of whether an item contains a Core question or not. Combined with the evidence from Figure 1, this finding suggests that prosodic information might not be essential to the identification of questions, but it might still play an important reinforcing role.

3.3. Non-prosodic information is crucial

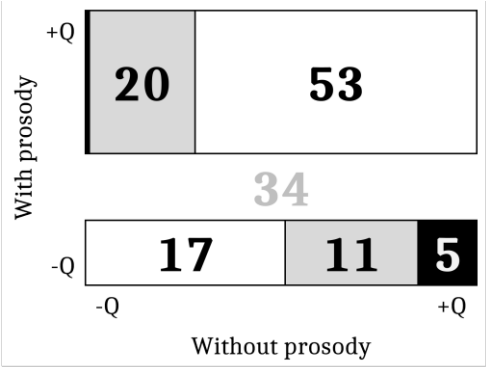


Figure 3: Agreement across questionhood ratings.

In order to further explore the contribution of prosodic information, we analysed the 74 items with positive Core question ratings (responses 7, 8 or 9, shortened to "+Q" below) separately from the 33 items with negative ratings (1, 2 or 3, "-Q"). Conservatively, in this analysis we ignore the 34 items which received intermediate ratings.

Figure 3 shows results along the y-axis for stimuli with prosodic information (§2.3b) and along the x-axis for the others. Out of 74 items rated as +Q based on prosodic information (top row), 53 were also rated as +Q based on textual or contextual information only. Only 1 was rated as -Q. Similarly, but less strikingly, of 33 items rated as -Q based on prosodic information (bottom row), 17 were rated as -Q based on textual and contextual information only, while 5 were rated as +Q. The fact that Core questions can be identified in most cases via textual and contextual cues demonstrates a high overlap between informational dimensions, and suggest that prosodic information might be difficult to isolate or to manipulate as a separate source of information.

4. DISCUSSION

Although the present study used a small set of data from one type of interaction, our findings highlight some potential drawbacks to the experimental study of question intonation. First, Figure 1 shows that the type of questions which are largely featured in experimental research might not necessarily correspond to the type of questions that are prevalent in everyday language use. This finding joins ends with research showing that syntactically well-formed, unbiased requests for new information might not be theoretically [11] or empirically [25] typical.

Second, and perhaps more radically, Figure 3 serves as a reminder that linguistic categories tend to be encoded redundantly [26], i.e. using different strands of information, which are then all used in perception by listeners. This casts a shadow on the assumption that different layers of information can be separated, isolated, manipulated and controlled in an experimental setup. This scepticism is compatible with views recognising the cohesive nature and function of prosody [27], against a mere superposition of segmental and suprasegmental levels [28]. It also echoes the body of work on prosodic constructions, where prosodic and lexical material are not seen as necessarily independent [29, 30].

While recognising the benefits of studying language in a controlled environment, it might thus be necessary to either verify the experimental findings using evidence from different sources [6, 31], or to ground experimentation on previous thorough observation of the phenomena of interest outside of the laboratory [7], where language is produced and perceived in a contextualised and meaningful way. Failure to do so runs the risk of divorcing research methods and goals from the reality in which they should be grounded [32], potentially leading to the atomistic, disconnected and often irreplicable body of findings [33, 34] that currently characterises research on question intonation.

ACKNOWLEDGEMENTS

FC conceptualised the study, collected and analysed the data, and wrote the first manuscript draft. MG, HJ and JS rated the stimuli and revised the manuscript.
We are grateful to Hauke Lindstädt (segmentation, transcription) and Aviad Albert (additional ratings). Funding was obtained from the German Research Foundation (DFG), Project-ID 281511265, CRC 1252 "Prominence in Language".

REFERENCES

[1] Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, *38*(3), 329–336. https://doi.org/10.1016/j.wocn.2010.04.003

[2] de Saussure, F. (1916). *Cours de ling. générale*. Payot.

[3] Spitzer, L. (1921). *Italienische Kriegsgefangenenbriefe. Materialien zu einer Charakteristik der volkstümlichen italienischen Korrespondenz*. Peter Hanstein.

[4] Benveniste, É. (1974). *Problèmes de linguistique générale II*. Gallimard.

[5] Rischel, J. (1992). Formal linguistics and real speech. *Speech Communication*, *11*(4–5), 379–392. https://doi.org/10.1016/0167-6393(92)90043-7

[6] Wagner, P., Trouvain, J., Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. https://doi.org/10.1016/j.wocn.2014.11.001

[7] Kingstone, A., Smilek, D., Eastwood, J. D. (2008). Cognitive Ethology: A new approach for studying human cognition. *British Journal of Psychology*, *99*(3), 317–340. https://doi.org/10.1348/000712607X251243

[8] Albert, S., de Ruiter, J. P. (2018). Improving Human Interaction Research through Ecological Grounding. *Collabra: Psychology*, *4*(1), 24. https://doi.org/10.1525/collabra.132

[9] Haan, J. (2000). *Speaking of Questions. An Exploration of Dutch Question Intonation*. LOT.

[10] de Ruiter, J. P. (2012). *Questions: Formal, Functional and Interactional Perspectives*. Cambridge University Press. https://doi.org/10.1017/CBO9781139045414

[11] Ozerov, P. (2019). This is not an interrogative: The prosody of “wh-questions” in Hebrew and the sources of their questioning and rhetorical interpretations. *Language Sciences*, *72*, https://doi.org/10.1016/j.langsci.2018.12.004

[12] Zahner, K., Xu, M., Chen, Y., Dehé, N., Braun, B. (2020). The prosodic marking of rhetorical questions in Standard Chinese. *Speech Prosody 2020*, 389–393. https://doi.org/10.21437/SpeechProsody.2020-80

[13] Asu, E. L., Sahkai, H., Lippus, P. (2020). The prosody of rhetorical and information-seeking questions in Estonian: Preliminary results. *Speech Pros. 2020*, 381–384. https://doi.org/10.21437/SpeechProsody.2020-78

[14] Orrico, R., D’Imperio, M. (2020). Tonal specification of speaker commitment in Salerno Italian wh-questions. *Speech Prosody 2020*, 361–365. https://doi.org/10.21437/SpeechProsody.2020-74

[15] Prieto, P., Borràs-Comes, J., Roseano, P. (Coords.) (2010-2014). *Interactive Atlas of Romance Intonation*. http://prosodia.upf.edu/iari/

[16] The Ninth Age (2022). *Essence of War*. https://www.the-ninth-age.com/download/quick-play/

[17] Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection. *International Journal of Qualitative Methods*, *18*, 1609406919874596. https://doi.org/10.1177/1609406919874596

[18] Zoom Video Communications Inc. (2021). *Security guide*. https://explore.zoom.us/docs/doc/Zoom-Security-White-Paper.pdf

[19] OBS Studio Contributors (2021). *Open Broadcaster Software*. https://obsproject.com

[20] Agoners (2022). *The 9th Age EoW Advanced Battle: "This Is Why We BetaTest!!"*. https://youtu.be/VTkdztIq0OE

[21] ten Have, P. (2002). Ontology or methodology? Comments on Speer’s `natural’ and `contrived’ data: a sustainable distinction? *Discourse Studies*, *4*(4), 527–530. https://doi.org/10.1177/1461445602004004070l

[22] Boersma, P. Weenink, D. (2022). *Praat: doing phonetics by computer*. https://www.praat.org/

[23] Cangemi, F. (2023). *23-08.* Supplementary material for “Contrast or context, that is the question”. https://osf.it/wp62b

[24] R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

[25] Enfield, N. J., Stivers, T., Levinson, S. C. (2010). Question–response sequences in conversation across ten languages. *Journal of Pragmatics*, *42*(10), 2615–2619. https://doi.org/10.1016/j.pragma.2010.04.001

[26] Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967. https://doi.org/10.1002/bies.201400028

[27] Arvaniti, A. (2016). *Basic notions of prosody*. Talk at Aix Summer School on Prosody. https://aixprosody2016.weebly.com/program.html

[28] Roettger, T., Grice, M. (2019). The tune drives the text - Competing information channels of speech shape phonological systems. *Language Dynamics and Change*, *9*, 265–298. https://doi.org/10.1163/22105832-00902006

[29] Couper-Kuhlen, E., Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on interaction. E. Couper-Kuhlen & M. Selting (eds.), *Prosody in Conversation: Interactional Studies* (pp 11-57). Cambridge University Press.

[30] Ward, N. G. (2019). *Prosodic Patterns in English Conversation*. Cambridge University Press. https://doi.org/10.1017/9781316848265

[31] Beckman, M. E. (1997). A Typology of Spontaneous Speech. Y. Sagisaka, N. Campbell, & N. Higuchi (eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 7–26). Springer US. https://doi.org/10.1007/978-1-4612-2258-3_2

[32] Neisser, U. (1978). Memory: What are the important questions? M. M. Gruneberg, P. E. Morris, & R. N. Sykes (eds.), *Practical aspects of memory* (pp. 3-24). Academic Press.

[33] Savino, E. (2012). The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association, 42*(1), 23-48. https://doi.org/10.1017/S002510031100048X

[34] Garassino, D., D. Dipino, F. Cangemi (2022). Per un approccio multidimensionale allo studio dell’intonazione. In Schmid, S., Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S. & Pellegrino, E. (eds.), Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications. Milano: Officinaventuno. 219-242. https://doi.org/10.17469/O2108AISV000012

IMPLICIT IMITATION OF INTONATION CONTOURS IN WORD SHADOWING

Yu Jin Song, Cynthia G. Clopper

The Ohio State University
song.1693@osu.edu, clopper.1@osu.edu

ABSTRACT

Phonetic imitation of a range of segmental and suprasegmental features has been observed across tasks and instructions, suggesting that this process may be spontaneous and implicit. In contrast, imitation of phonetic features of intonation contours may depend largely on the explicitness of the instructions, especially in non-interactive tasks. In the current study, a word shadowing task without explicit instructions to imitate was used to explore the imitation of intonation contours and their phonetic implementation in American English. The analyses indicate imitation of rising contours, their f0 range, and their alignment with the stressed vowel. However, imitation of falling contours was limited to their alignment with the stressed vowel. These results suggest that the shape of intonation contours may affect their imitation, and that a perceptually more salient contour, such as a rise on an isolated word in American English, may lead to greater imitation.

Keywords: phonetic imitation, intonation, word shadowing, phonological abstraction

1. INTRODUCTION

Humans align their speech with the phonetic realization of the speech of their interlocutor [1, 2]. This phonetic imitation has been observed for a range of segmental features, including vowel quality [2, 3, 4] and voice onset time [5, 6], and for suprasegmental features, including overall f0 [1, 7, 8, 9] and speaking rate [7]. Talkers imitate in conversational settings [7, 10, 11] and in non-interactive tasks, such as word shadowing [1, 8, 12]. In non-interactive tasks, imitation has been observed both with and without explicit instructions to imitate [1, 6] and even with explicit instructions not to imitate [4], suggesting that phonetic imitation may be spontaneous and implicit.

In contrast, imitation of the phonological and phonetic features of intonation contours may depend on the explicitness of the instructions, especially in non-interactive tasks. In interactive tasks, talkers imitate both phonological intonation contours [10, 13] and their phonetic features, such as pitch scaling [14], without instructions to imitate. For example, Lee et al. [14] found imitation of rising boundary tones and of f0 maxima and minima in American English, using a maze navigation task.

In non-interactive tasks, talkers likewise imitate both phonological and phonetic features of intonation contours when they are explicitly asked to imitate [9, 15, 16, 17]. For example, Cole and Shattuck-Hufnagel [16] asked participants to repeat the way the sentence-length utterances were said and observed imitation of phonological features of the intonation contours, including pitch accents and boundary tones, but idiosyncratic imitation of phonetic features of the contours, such as the duration of pauses and glottalization at prosodic boundaries. Similarly, D'Imperio et al. [15] asked Italian participants to imitate as closely as possible the productions of a model talker who spoke an unfamiliar Italian dialect. Along with imitation of phonological features of the contours, D'Imperio et al. [15] found imitation of phonetic features of the contours, such as tonal alignment and pitch scaling. In non-interactive tasks without explicit instructions to imitate, talkers imitate intonation contours [18] and overall f0 [8, 9, cf. 19, 20], but implicit imitation of phonetic features of intonation contours has not been examined. Thus, convergence to phonetic features of intonation contours can be achieved by explicit instructions to imitate [15, 17]. However, imitation of phonological features of intonation contours tends to be more robust than that of phonetic features [16, 17, 21], suggesting that explicit imitation is mediated by abstract phonological representations [22].

In the current study, the goal was to further explore the implicit imitation of intonation contours and their phonetic implementation using a word shadowing task without explicit instructions to imitate. This approach allowed us to determine to what extent talkers spontaneously align their intonation contours with that of a model talker. We predicted that we would observe robust imitation of overall intonation contours [18], but less imitation of the phonetic features of those contours [16, 17, 21].

2. METHODS

2.1. Participants

Word shadowing data were collected from 45 participants (22 female, 23 male) recruited from a

local science museum in Columbus, OH. The participants were all native American English speakers and self-reported no history of speech, hearing, or language disorders. Their ages ranged from 19 to 69 years old (M = 35 years).

2.2. Stimulus materials

The stimulus materials comprised 48 multisyllabic English words, each produced by two female model talkers. The target words were 2-4 syllables long and the stressed syllable in each target word contained one of the following eight vowels: /i ɪ ɛ æ ɑ aɪ oʊ u/, with six words for each vowel. The auditory stimuli were selected from the Indiana Speech Project corpus [23]. One of the model talkers was from Indianapolis, IN, and was 19 years old at the time of recording. She produced all 48 target words with a rising intonation contour. The other model talker was from Fort Wayne, IN, and was 22 years old at the time of recording. She produced all 48 target words with a falling contour.

2.3. Procedure

Participants were seated at an individual testing station with a computer and a headset microphone in a glass-enclosed laboratory inside the science museum. In the first block of the word shadowing task, participants were asked to read aloud the set of words from the computer screen. Each of the 48 words appeared on the computer screen one by one. Each word stayed on the screen for 3.5 s followed by a .5 s blank screen to signal the onset of the next trial. Word order was randomized separately for each participant. This task served as the baseline for the shadowers' intonation contours.

A shadowing task followed in the second block. During shadowing, participants were asked to repeat the same set of words after one of the model talkers. Nineteen participants completed the shadowing task with the model talker who produced rising contours and 26 participants completed the shadowing task with the model talker who produced falling contours. Contour condition (Rise, Fall) was therefore a between-participant variable. In the shadowing block, the words were played one at a time over the headphones. Each auditory stimulus was preceded by a fixation cross which stayed on the screen for .5 s. The cross remained for another 4 s as the word played and the participants produced their repetition. A blank screen followed for 1 s to signal the onset of the next trial. Word order was randomized separately for each participant. Participant utterances were recorded in Audacity. All misread and misheard tokens were discarded prior to analysis. Thirty-nine tokens (2.14%) from the Rise condition and 43 tokens (1.72%) from the Fall condition were excluded.

2.4. Analysis

2.4.1. Contour coding

To assess imitation of the phonological features of the intonation contours, the intonation contours of the target words produced by the model talkers and by the shadowers in both the baseline and shadowing blocks were coded by a team of trained undergraduate research assistants. Each utterance was coded based on auditory and visual inspection as having been produced with one of three nuclear contour categories (rise, fall, plateau) to capture the intonation contour from the stressed syllable to the end of the word. The data from four shadowers (9%) were coded independently by two coders to assess reliability. The coders agreed on 97% of the utterances, suggesting very high reliability for this coding scheme.

Given that the model talkers produced exclusively rising contours in the Rise condition and falling contours in the Fall condition, respectively, imitation of intonation contours was assessed by comparing the proportion of rising contours in the baseline and shadowing blocks in the Rise condition and the proportion of falling contours in the baseline and shadowing blocks in the Fall condition. An increase in these proportions from baseline to shadowing indicates imitation (i.e., the proportions for the shadowers were closer to those for the model talker in the shadowing block than in the baseline block).

2.4.2. Acoustic analysis

To assess imitation of the phonetic features of the rising contours in the Rise condition and of the falling contours in the Fall condition, the f0 range and alignment relative to the stressed syllable of the rises produced in the Rise condition and of the falls produced in the Fall condition were analyzed. For all of the model talkers' utterances and for utterances produced by the shadowers in the baseline and shadowing blocks that were coded as rises in the Rise condition and as falls in the Fall condition, the recordings were analyzed in VoiceSauce [24] to extract f0 values at 1 ms intervals using the STRAIGHT algorithm [25].

The f0 range of the rises was defined as the difference between the minimum f0 value in the stressed vowel and the maximum f0 value between the minimum f0 location and the end of the word. The f0 range of the falls was defined as the difference between the maximum f0 value in the stressed vowel and the minimum f0 value between the maximum f0 location and the end of the word. The minimum and maximum f0 values were inspected for outliers. Values that fell outside of two standard deviations of an individual talker's mean were replaced with the

respective talker f0 maximum or minimum mean. In the Rise condition, 36 minima and 44 maxima (4.64%) were replaced and, in the Fall condition, 42 minima and 46 maxima (3.29%) were replaced. The corrected f0 maxima and minima were used to calculate f0 ranges in semitones to normalize for overall f0 range differences across talkers. Distances in f0 range between the shadowers and the model talkers were defined as the absolute difference in f0 range between the shadower and the model talker, separately for each word in each block [3, 19]. Given that f0 range measures were only available for rises produced in the Rise condition and for falls produced in the Fall condition, different numbers of tokens were included in each of the two blocks for each shadower. A difference-in-distance measure [3, 19] was therefore not possible. Imitation was instead assessed by comparing the f0 range distances in the baseline and shadowing blocks. A decrease in these distances from baseline to shadowing indicates imitation (i.e., the f0 ranges of the shadowers were closer to those of the model talker in the shadowing block than in the baseline block) [3, 19].

Alignment was defined as the time from the onset of the stressed vowel to the f0 minimum in the Rise condition and to the f0 maximum in the Fall condition. Similarly to the f0 range analysis, distances in f0 alignment between the shadowers and the model talkers were defined as the absolute difference in f0 alignment between the shadower and the model talker, separately for each word in each block [3, 19]. A decrease in these distances from baseline to shadowing indicates imitation [3, 19].

3. RESULTS

3.1. Rise condition

Table 1 shows the mean proportions of rising contours in the baseline and shadowing blocks in the Rise condition. To explore the effects of block (baseline vs. shadowing) and shadower gender (female vs. male) on the use of rising contours (rising vs. non-rising) in the Rise condition, a logistic mixed-effects model with the maximal random-effects structure that converged was constructed in R using lme4 [26]. Random effects were random intercepts by shadower and word and random slopes for condition by shadower and for gender by word. The main effect of block was significant ($\chi^2(1) = 10.44$, $p = .001$), confirming imitation of the model talker's rising contours in the shadowing block. The effect of gender and its interaction with block were not significant.

Table 2 shows the mean f0 range distance in semitones between the shadowers and the model talker in the baseline and shadowing blocks in the Rise condition. To explore the effects of block and shadower gender on the f0 range distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for gender by word, was constructed in R using lmerTest [27]. The main effect of block was significant ($\chi^2(1) = 5.09$, $p = .024$), confirming imitation of the model talker's f0 range in rising contours in the shadowing block.

Baseline	Shadowing
.34 (.33)	.63 (.34)

Table 1: Mean (standard deviation) proportion of rising contours by block in the Rise condition.

Baseline	Shadowing
3.08 (1.19)	2.65 (.90)

Table 2: Mean (standard deviation) f0 range distance in semitones between the shadowers and the model talker by block in the Rise condition.

Table 3 shows the mean f0 alignment distance in milliseconds between the shadowers and the model talker in the baseline and shadowing blocks in the Rise condition. To explore the effects of block and shadower gender on the f0 alignment distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word, was constructed. The main effect of block was marginally significant ($\chi^2(1) = 3.71$, $p = .054$), suggesting modest imitation of the model talker's f0 alignment in rising contours in the shadowing block.

Baseline	Shadowing
56 (15)	50 (13)

Table 3: Mean (standard deviation) f0 alignment distance in milliseconds between the shadowers and the model talker by block in the Rise condition.

3.2. Fall condition

The mean proportions of falling contours increased from 0.53 (*SD* = .39) in the baseline block to 0.55 (*SD* = .38) in the shadowing block in the Fall condition. To explore the effects of block and shadower gender on the use of falling contours (falling vs. non-falling), the same logistic mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by shadower, was

constructed, as for the Rise condition. No effects were statistically significant.

The mean f0 range distance between the shadowers and the model talker increased from 2.85 semitones (*SD* = 2.23) in the baseline block to 2.97 semitones (*SD* = 1.04) in the shadowing block. To explore the effects of block and shadower gender on the f0 range distances, the same linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by shadower, was constructed, as for the Rise condition. No effects were statistically significant.

Table 4 shows the mean f0 alignment distance in milliseconds between the shadowers and the model talker in the baseline and shadowing blocks in the Fall condition, separately for female and male shadowers. To explore the effects of block and shadower gender on the f0 alignment distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by word, was constructed. The main effect of block was significant ($\chi^2(1) = 6.73$, $p = .009$), confirming imitation of the model talker's f0 alignment in falling contours in the shadowing block. The interaction between block and gender was also significant ($\chi^2(1) = 7.02$, $p = .008$). Post-hoc pairwise least square means comparisons revealed that the mean f0 alignment distance decreased significantly from the baseline block to the shadowing block for the female shadowers only ($t(208.2) = 3.76$, $p < .001$).

	Baseline	Shadowing
Female	49 (17)	36 (12)
Male	37 (14)	35 (32)

Table 4: Mean (standard deviation) f0 alignment distance in milliseconds between the shadowers and the model talker by block and gender in the Fall condition.

4. DISCUSSION

The results of the word shadowing task reveal a significant increase in the shadowers' use of rising intonation contours when repeating after the model talker in the Rise condition, relative to the shadowers' baseline. The shadowers' rising contours were also acoustically more similar to the model talker's rises in terms of f0 range and alignment in the shadowing block than in the baseline block. In the Fall condition, the female shadowers' falling contours were acoustically more similar to the model talker's falls in terms of f0 alignment in the shadowing block than in the baseline block. Together, these results suggest imitation of both phonological and phonetic features of intonation contours, as in previous work [9, 15, 16, 17, 21]. However, the results from the Fall condition, in which only phonetic imitation was observed, suggest that imitation of intonation contours need not be mediated by abstract phonological representations [cf. 22] and may reflect a phonetically-based perception-production link [1].

The evidence for imitation was more robust overall for the Rise condition than the Fall condition. Given that the primary difference between the two conditions was the shape of the intonation contour produced by the model talker, the salience of the contour may have affected the magnitude of imitation. Previous research [28, 29] has shown that more marked variants are imitated more than less marked variants and phonologically relevant variants or features may be imitated to a greater extent than variation that does not signal a change in meaning [10, 22]. Romera and Elordieta [13] found a similar asymmetry as in the current study in the imitation of L2 Majorcan Spanish intonation contours by L1 Peninsular Spanish speakers. Although both contours that they considered are falls, the intonation contour of interrogative sentences was imitated while that of declarative sentences was not. Romera and Elordieta [13] attributed this asymmetry to the relative perceptual salience of the interrogative contour. The interrogative fall has a relatively greater f0 range and steepness, making the tune more perceptually salient than that of the declarative tune.

Likewise, the difference in imitation of the rising and falling contours in the current study may reflect their relative salience. In American English, rises are associated with various functional meanings, such as listing, uncertainty, or "uptalk," whereas falls are associated with simple declaratives [30, 31]. The various meanings associated with rises in American English may result in the assignment of differential status to the contours, such that rises are perceived as marked relative to falls when they are produced on isolated words as in the current word shadowing task.

In conclusion, the observed patterns of implicit imitation of intonation contours in the word shadowing task in the current study demonstrate that both the phonological and phonetic features of intonation contours can be imitated without explicit instructions to imitate. However, the extent of imitation may be affected by the salience of the contour. In particular, more marked contours are imitated more robustly than less marked contours.

5. ACKNOWLEDGMENTS

We would like to thank Laura Beebe, Emily Behm, Elizabeth Bohinski, Sarah Chilson, Hali Clark,

McKenna Reeher, and Erika Shane for assistance with data collection and analysis. These data were collected at the Language Sciences Research Lab at the Center of Science and Industry in Columbus, OH.

6. REFERENCES

[1] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Rev.* 105, 251-279.

[2] Pardo, J. S., Jay, I. C., Krauss, R. M. 2010. Conversational role influences speech imitation. *Atten. Percept. Psychophys*. 72, 2254-2264.

[3] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon.* 40, 177-189.

[4] Walker, A., Campbell-Kibler, K. 2015. Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Front. Psychol.* 6, 546.

[5] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *J. Phon.* 39, 132-142.

[6] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Percept. Psychophys.* 66, 422-429.

[7] Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., Campbell, N. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. In: *INTERSPEECH 2013, Lyon, France, August 25-29, 2013, Proceedings*. 539-543.

[8] Babel, M., Bulatov, D. 2012. The role of fundamental frequency in phonetic accommodation. *Lang. Speech* 55, 231-248.

[9] Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., Nguyen, N. 2013. Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Front. Psychol.* 4, 422.

[10] Savino, M. 2017. The dynamics of prosodic adaptation between Italian conversational partners. In Botinis, A. (ed), *ExLing 2017, Heraklion, Crete, Greece, June 19-22, 2017, Proceedings*. 93-96.

[11] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* 119, 2382-2393.

[12] Garnier, M., Lamalle, L., Sato, M. 2013. Neural correlates of phonetic convergence and speech imitation. *Front. Psychol.* 4, 600.

[13] Romera, M. Elordieta, G. 2013. Prosodic accommodation in language contact: Spanish intonation in Majorca. *Int. J. Sociol. Lang.* 221, 127-151.

[14] Lee, Y., Gordon, D. S., Parrell, B., Lee, S., Goldstein, L., Byrd, D. 2018. Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task. *PLoS ONE* 13(8), e0201444.

[15] D'Imperio, M., Cavone, R., Petrone, C. 2014. Phonetic and phonological imitation of intonation in two varieties of Italian. *Front. Psychol.* 5, 1-10.

[16] Cole, J., Shattuck-Hufnagel, S. 2011. The phonology and phonetics of perceived prosody: What do listeners imitate? In: *INTERSPEECH 2011, Florence, Italy, August 27-31, 2011, Proceedings*. 969-972.

[17] Petrone, C., D'Alessandro, D., Falk, S. 2021. Working memory differences in prosodic imitation. *J. Phon.* 89, 1-17.

[18] Michelas, A., Nguyen, N. 2011. Uncovering the effect of imitation on tonal patterns of French Accentual Phrases. In: *INTERSPEECH 2011, Florence, Italy, August 27-31, 2011, Proceedings*. 973-976.

[19] Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E. 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *J. Mem. Lang.* 69, 183-195.

[20] Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J. 2017. Phonetic convergence across multiple measures and model talkers. *Atten. Percept. Psychophys.* 79, 637–659.

[21] German, J. S. 2012. Dialect adaptation and two dimensions of tune. In: *Speech Prosody 2012, Shanghai, China, May 22-25, 2012*, *Proceedings*. 430-433.

[22] Mitterer, H., Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109, 168-173.

[23] Clopper, C. G., Carter, A. K., Dillon, C. M., Hernandez, L. R., Pisoni, D. B., Clarke, C. M., Harnsberger, J. D., Herman, R. 2002. The Indiana Speech Project: An overview of the development of a multi-talker multi-dialect speech corpus. In: *Research on Spoken Language Processing Progress Report No. 25*. Speech Research Laboratory, Indiana University, 367-380.

[24] Shue, Y.-L., Keating, P., Vicenik, C., Hu, K. 2011. VoiceSauce: A program for voice analysis. In: Lee, W. S., Zee, E. (eds), *ICPhS 2011, Hong Kong, August 17-21, 2011, Proceedings.* 1846-1849.

[25] Kawahara, H., Cheveigne, A., Patterson, R. D. 1998. An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT-suite. In: *ICSLP 98, Sydney, Australia, November 30-December 4, 1998, Proceedings*. 659-662.

[26] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1-48.

[27] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1-26.

[28] Honorof, D. N., Weihing, J., Fowler, C. A. 2011. Articulatory events are imitated under rapid shadowing. *J. Phon.* 39, 18-38.

[29] Mitterer, H., Müsseler, J. 2013. Regional accent variation in the shadowing task: Evidence for a loose perception–action coupling in speech. *Atten. Percept. Psychophys.* 75, 557-575.

[30] Ladd, R. D. 1980. *The Structure of Intonational Meaning.* Indiana University Press.

[31] Warren, P. 2016. *Uptalk: The Phenomenon of Rising Intonation.* Cambridge University Press.

Enhancement of intonational contrasts in American English

Jennifer Cole[1], Jeremy Steffman[1,2]

[1]Northwestern University, [2]The University of Edinburgh
jennifer.cole1@northwestern.edu, jeremy.steffman@ed.ac.uk

ABSTRACT

We test a subset of intonational contrasts proposed in the Autosegmental-Metrical model for American English for evidence of contrast enhancement in phonologically and phonetically longer vs. shorter intervals. F0 trajectories were assessed from 32 speakers' imitated productions of six tonally distinct tunes, e.g., HHH, HHL. Maximally three tune shapes emerge from clustering analyses of imitated f0 trajectories, each cluster comprising imitations of two phonetically similar but phonologically distinct tunes. We find enhancement of tune contrasts *between* the emergent clusters in measures of f0 differences (RMSD, end f0, center of gravity). There is no evidence of enhancement for phonetically similar tunes grouped *within* the same cluster, though fine-grained phonetic distinctions are detected for these "lost" tune contrasts, suggesting a reanalysis as within-category variation.

Keywords: intonation, contrast, nuclear tune, English

1. INTRODUCTION

The intonation system of English presents a rich variety of pitch patterns in the final region of an intonational phrase. From the rightmost word with 'nuclear' phrasal stress to the end of the phrase, we find monotonically rising or falling patterns that vary in their initial or final f0 values, and more complex patterns with slope that changes e.g., rise-plateau, rise-fall, or rise-fall-rise [1-3] There is a long history of research that relates variation in these nuclear pitch patterns to variation in pragmatic meaning [4, 5], and though there is general agreement on the type of meaning distinctions encoded through intonation in English (e.g., referential alternatives, givenness, epistemic knowledge), there is not yet a consensus about the categorical status of pitch patterns—the number and type of discrete phonological contrasts that are encoded, and their associated meaning functions. This paper investigates the distinctive status of phrase-final pitch patterns in Mainstream American English (MAE) through a phonological lens, to identify distinctions that are phonetically enhanced in phonological contexts that support the full expression of pitch patterns. Informed by work in the segmental domain [6-8], we view enhancement as a phenomenon that serves to increase the perceptual distinctiveness of phonological contrasts. Applying this notion to intonation, we consider the phonetic enhancement of a distinction between two pitch patterns as indicating a representational distinction between phonological categories. This study aims to identify phonological contrasts in the phonetic distinctions among phrase-final pitch patterns, which may inform later work on the discrete and/or gradient meaning functions of intonation.

This investigation focuses on the Autosegmental-Metrical (AM) account of MAE intonation [9, 10], codified in the ToBI annotation system [11], which models phrase-final pitch patterns in terms of three features: the pitch accent marking the word with rightmost ('nuclear') phrasal stress, followed by the phrase accent and boundary tone marking the end of phrasal domains at two levels of prosodic phrasing. These intonational features are specified in terms of the tonal primitives H(igh) and L(ow), and considering only monotonal pitch accents, the system generates eight phonologically distinct "tunes", each comprised of three tone features: HHH, HHL, HLH, HLL, LLL, LLH, LHL, LHH (inclusion of a downstepped High and three bitonal pitch accents further extends the inventory to as many as 24 phonologically distinct nuclear tunes). Each tune generates a distinct pitch trajectory, and while some of these trajectories are robustly distinct (HHH, LLL), others are less so. We focus here on three tune pairs with very similar pitch trajectories that differ only in the final region: {HHH, HHL}, {HLL, HLH} and {LLL, LLH}, as shown in the schematized trajectories of Figure 1, adapted from [10].

Using a tune imitation paradigm, we examine the phonetic implementation of these six tunes to assess the shape and magnitude of f0 distinctions between tune pairs as produced over words that differ in the number of stressed and unstressed syllables in the nuclear region. Prior work with MAE speakers shows poor perceptual discrimination of these tune pairs when presented in 3-syllable words [12], and little or no distinction in the f0 trajectories produced for each pair, again in 3-syllable words [13]. The goal of the present study is to determine whether the predicted contrasts among this set of tunes are enhanced when produced on longer words, and more broadly, how phonetic distinctions in the f0 trajectories of these tunes increase or decrease as a function of metrical

structure—the syllable count and stress pattern in the region of the nuclear tune.

2. METHODS

Stimuli. Participants heard model utterances with resynthesized f0 trajectories representing the six tunes, and imitated the heard tune on each trial, reproducing it in a new sentence. Model utterances were naturally produced by two speakers (one male, one female) in two sentences ("Her name is Marilyn"/ "He answered Jeremy"). The six nuclear tunes were implemented with f0 resynthesis using PSOLA in Praat [14,15], based on straight-line approximations with five target f0 values located in each model speaker's pitch range (Figure 1). The scaling and alignment of resynthesized tunes were based on examples from [10, pp. 391- 401] and online training materials [16].

Participants and procedure. 32 self-reported native speakers of American English were recruited via Prolific and completed the experiment remotely. In a given trial, participants heard the same tune on two stimuli separated by one second. Participants were prompted to reproduce the heard tune on a new target sentence presented orthographically. Target sentences were in one of five metrical conditions, all with initial stress, that varied in the syllable count (1-4) of the final word, with an additional secondary stress in 4b as a potential anchor for the phrase accent [17] (Table 1). There were 120 trials: 6 tunes x 5 conditions x 2 sentences x 2 repetitions. Model gender order (male/female or female/male) and model sentence order were evenly distributed across metrical conditions and tunes and each appeared an equal number of times in the experiment. We measured f0 using STRAIGHT in VoiceSauce [18,19] and computed time-normalized trajectories with 30 samples over each nuclear word. Files containing f0 tracking errors were flagged and removed using an automated algorithm [20]. Differences in the phonetic implementation of tunes were assessed as follows.

Clustering analyses using k-means clustering for longitudinal data [21] were performed on unlabeled

Condition	Sentence
1	He ran with *Moe*
	She lived with *Neil*
2	Her roommate *Nóra*
	His neighbor *Mánny*
3	She gathered *lávender*
	They honored *Mélanie*
4a	He went there *críminally*
	She travelled *mínimally*
4b	They saw the *nóminàtor*
	He was a *lúminàry*

Table 1: Stimuli sentences for the experiment.

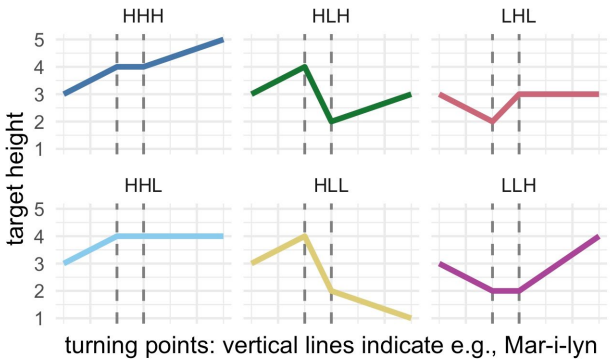


Figure 1: Schema for the model stimuli

imitated f0 trajectories to identify the number and type of distinct trajectories and their relation to the tune labels of the imitated stimuli. We tested clustering solutions with 2-10 clusters, using the Calinski-Harabatz criterion [22] to select the optimal partition of the data as the one with the highest ratio of between- to within-cluster variance. The analysis was carried out on participant mean trajectories—a single mean trajectory for imitations of each of the 6 model tunes, from each participant.

Root-mean squared difference (RMSD). We assessed differences among tunes in phonetic space using RMSD as a measure of the phonetic distance between a pair of f0 trajectories, computed for each speaker over their mean trajectories for each of the six tunes, for all pairwise combinations of tune within metrical condition. Each speaker contributed 75 RMSD values (15 tune pairs by 5 metrical conditions). Pairwise RMSD values were analyzed using mixed-effects Bayesian regression [23]. Metrical condition was modeled as a monotonic effect [24], reflecting our expectation that RMSD would increase with the number of syllables across metrical conditions 1-4. We included random intercepts for speaker, and tune pair, and by-speaker slopes for the fixed effects and interactions. Results are reported using the median posterior estimate and 95% credible intervals (CrI) and the probability of direction metric (pd) [25]. A CrI that excludes zero

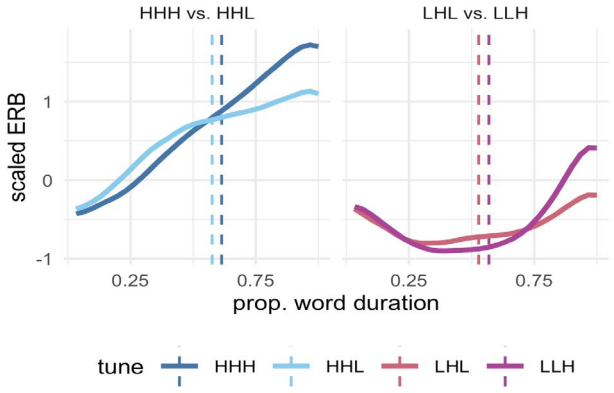


Figure 2: Mean trajectories for four tunes, with mean temporal TCoG, indicated by a dashed vertical line.

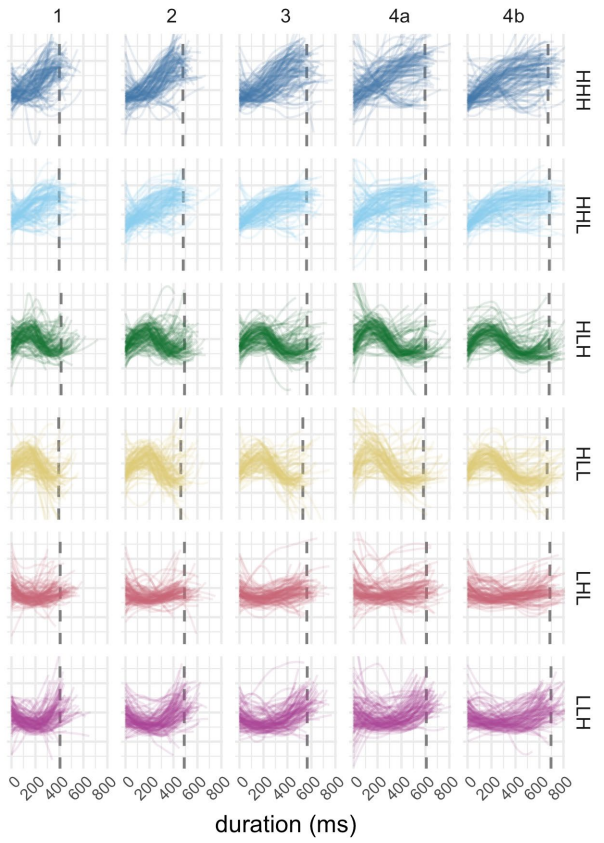


Figure 4: All trajectories for tunes split by metrical structure (columns) and tunes (rows). Vertical line indicates mean duration within each panel.

and pd > 95% are considered as compelling evidence for an effect.

Other parameters. We measured two additional parameters predicted to exhibit enhancement effects: **End f0** and **Tonal Center of Gravity** (TCoG). End f0 was measured in centered ERB values at the end of the nuclear tune interval. For tune pairs of interest, identified based on clustering and described below, we modeled variation in End f0 a s a function of tune (e.g. HHH versus HHL) and metrical condition, and their interaction, with a random intercept for speaker and by-speaker random slopes for the fixed effects and interaction. TCoG is a measure of the location in time of the bulk of the f0 mass over a specified region [26,27]. In our data, TCoG captures subtle differences in the alignment (early/late) and shape (domed/scooped) of f0 rises. We computed TCoG with reference to the end of the first syllable, in milliseconds. Figure 2 shows the mean TCoG for the four tunes we examine using this measure. For tune pairs of interest, defined by the clustering results, we modeled variation in TCoG as a function of tune, metrical structure, and their interaction, with the same random effect structure as for the end f0 model.

Hypotheses and predictions. We hypothesize enhanced contrasts between phonologically distinct tunes across metrical conditions 1-4, manifest in

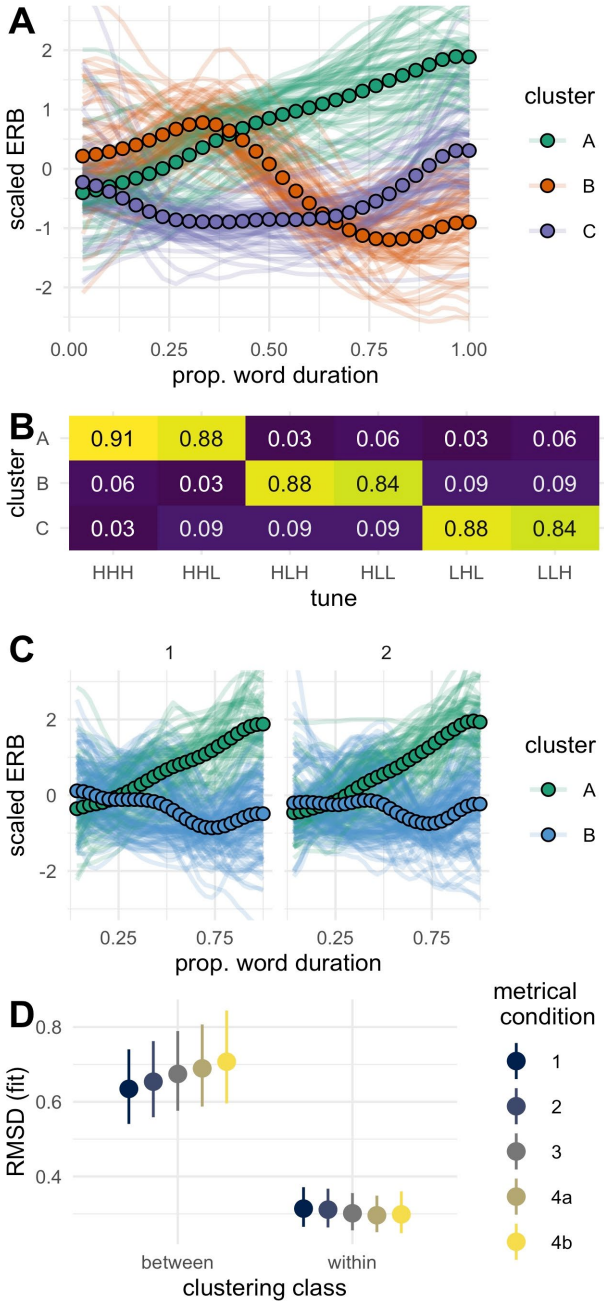


Figure 3: Panel A: Cluster means for three clusters in the partition, Panel B: the proportion of tunes (columns) in each cluster (rows). Panel C: Clustering solutions for 1 (left) and 2 (right) syllable words only. RMSD by cluster class and metrical structure.

enhanced phonetic distinctions with (i) f0 trajectories of imitated tunes that define a greater number of clusters, up to a maximum of 6 clusters in conditions 4a and 4b; (ii) increased RMSD between tune pairs; and (iii) greater differences between tunes in the measures of End f0 and TCoG.

3. RESULTS

Figure 3 shows the raw f0 trajectories from all trials in the experiment, split by metrical condition (columns) and tune (row). As expected, modeling of

the nuclear word duration showed a credible difference across metrical structures (all pairwise pd > 95). Phonetic and phonological lengthening create the potential for contrast enhancement.

Clustering results. The clustering algorithm optimally partitions the data into three clusters, merging imitations of two model tunes. The mean trajectories of the three clusters are shown in Figure 4A, while the mapping of tune to cluster is shown in Figure 4B. Cluster A shows a monotonically rising shape that is composed mainly from imitations of HHH and HHL. Cluster B has a rising-falling shape and comprises imitations of HLH and HLL. Cluster C has a low rising shape and mainly comprises imitations of LHL and LLH. These results suggest poor differentiation of the predicted tune contrasts in each cluster. Additional clustering analyses over subsets of data split by metrical condition yield similar results; conditions 3, 4a and 4b show the same three-way partition of the data as in Figure 4A, while conditions 1 and 2 show only two clusters (Figure 4C): Cluster A with imitations of HHH and HHL, and cluster B with imitations of the remaining four tunes.

RMSD Results. We focus our analysis of RMSD based on results from the clustering analysis, to compare the phonetic distinction between tunes that clustered together with those that clustered apart. Within-cluster pairs are {HHL, HHH}, {HLH, HLL} and {LHL, LLH}; all other pairs are between-cluster (see Figs. 4A,B). This cluster class variable was used as a predictor in the model, interacting with metrical structure. Results show an unsurprising main effect of cluster class: within-cluster tune pairs have a smaller RMSD (β = -0.70, 95%CrI = [-1.01,-0.39], pd = 100). There was no credible main effect of metrical structure, but notably, there was an interaction between clustering class and metrical condition (pd = 99), showing that there is enhancement (larger RMSD across conditions 1-4), *only for tune pairs that are grouped into different clusters,* as shown in Figure 4D: RMSD ascends left to right across conditions, *only* for between-cluster pairs. Pairwise comparisons between metrical conditions confirm credible differences for between-cluster pairs: {1,2} < 3 < {4a,4b} with pd > 95, while within-cluster pairs show only a slightly *smaller* RMSD for 3 versus 2 syllables (pd = 96), an effect opposite to enhancement.

Tune-end f0 results. For each pair of tunes that cluster together, we observe small differences in the End f0 as predicted by the model tunes (see Figure 1): HHL < HHH (β = -0.35, 95%CrI = [-0.50,-0.20], pd = 100), HLL < HLH (β = -0.36, 95%CrI = [-0.54,-0.19], pd = 100), and LHL < LLH (β = 0.27, 95%CrI = [0.08,0.45], pd = 100). Importantly, there was no interaction between tune and metrical structure in any of the three models (pds = 69, 93 and 89 respectively), indicating no evident enhancement effects across metrical conditions. In line with the RMSD results, the End f0 data show a *lack of within-cluster enhancement*. **TCoG results** show a similar pattern. TCoG in HHL is credibly earlier than that in HHH (β = -21, 95% CrI = [-33,-9], pd = 100), and TCoG in LLH is credibly later than LHL (β = 23, 95%CrI = [7,40], pd = 100). See Figure 2 for reference. Crucially however, there was not an interaction between tune and metrical structure (pd for HHH vs. HHL = 55, pd for LHL vs. LLH = 60). In other words, the distinction in TCoG in these tunes did not change systematically across metrical conditions, again showing no within-cluster enhancement.

4. DISCUSSION & CONCLUSION

We tested the hypothesis that phonological contrasts between nuclear tunes are enhanced when produced over longer vs. shorter intervals. Imitated productions of six tunes were assessed for evidence of enhanced contrasts, across five metrical conditions of increasing syllable count and in the presence of an additional, secondary stressed syllable following the nuclear pitch accent. Evidence from the clustering analysis shows maximally three distinct tune shapes, each comprising imitations of two phonetically similar tunes that differ in their tonal specification, with no further increase in the number of tune shapes in longer nuclear intervals. RMSD findings show enhancement of tune contrasts between, but not within, the three emergent tune clusters, with greater RMSD in nuclear intervals of increasing length, but with no extra enhancement in the presence of an additional secondary stress as a potential anchor for the tune-medial phrase accent. Two additional f0 measures, ending f0 and the temporal Tonal Center of Gravity, show expected differences between tunes, but critically, these effects are uniform across metrical conditions, and thus provide no evidence of enhanced contrasts between tunes in the phonetically similar pairs. For the six tunes tested, our findings support an analysis of a three-way phonological contrast in tune shape, with phonetic enhancement in words with longer nuclear intervals. The "lost" tune contrasts involve distinctions in phrase accent and boundary tones: {HHL vs. HHH}, {HLH vs. HLL}, {LHL vs. LLH}. The same tunes are also poorly discriminated in perception [28]. Tunes that cluster together may be better understood as within-category variation. This conclusion calls for reconsideration of categories and gradience in intonational phonology (see [29]), across speech communities and styles.

Acknowledgments: Thanks to Chun Chan and the Prosody & Speech Dynamics Lab at Northwestern U. This project was supported by NSF BCS-1944773.

5. REFERENCES

[1] Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

[2] Palmer, H. E. (1924). *English intonation with systematic exercises*. W. Heffer & Sons Limited.

[3] Pike, K. (1945). *The intonation of American English.* Ann Arbor, MI: University of Michigan Press.

[4] Prieto, P. (2015). Intonational meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 371-381.

[5] Westera, M., Goodhue, D., & Gussenhoven, C. (2020). Meanings of Tones and Tunes. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 442–453).

[6] Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language*, 33-63.

[7] Stevens, K. N., & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language*, 81-106.

[8] Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*(1), 10-19.

[9] Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

[10] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

[11] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. Im S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing*.

[12] Cole, J., Steffman, J., & Tilsen, S. (2022). Shape matters: Machine classification and listeners' perceptual discrimination of American English intonational tunes. *Proceedings of Speech Prosody 2022*, 23-26.

[13] Cole, J., & Steffman, J. The primacy of the rising/non-rising dichotomy in American English intonational tunes. In *Proc. 1st International Conference on Tone and Intonation (TAI)* (pp. 122-126).

[14] Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1., retrieved from http://www.praat.org/.

[15] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5-6), 453-467.

[16] Veilleux, N., Shattuck-Hufnagel S. & Brugos A. *6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI.* January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu.
License: Creative Commons BY-NC-SA.

[17] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2010). Turning points, tonal targets, and the English L-phrase accent. *Language and Cognitive Processes*, 25(7-9), 982-1023.

[18] Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T. & Irino, T. (2005). Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Ninth European Conference on Speech Communication and Technology*.

[19] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011) VoiceSauce: A program for voice analysis. In *Proceedings of ICPhS XVII*, 1846-1849.

[20] Steffman, J., & Cole, J. (2022). An automated method for detecting F 0 measurement jumps based on sample-to-sample differences. *JASA Express Letters*, *2*(11), 115201.

[21] Genolini C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4), 1-34.

[22] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

[23] Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*, 1-28.

[24] Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420-451.

[25] Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

[26] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2), 337-383.

[27] Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics*, *85*, 101020.

[28] Ladd, D. R. (2022). The trouble with ToBI. In Barnes, J., & Shattuck-Hufnagel, S. (Eds.). *Prosodic Theory and Practice*, 247-258. MIT Press.

TEMPORAL DETERMINANTS OF PHRASING IN COORDINATED STRUCTURES IN GREEK

Katerina Nicolaidis, Mary Baltazani

Aristotle University of Thessaloniki, Greece, University of Oxford, UK
knicol@enl.auth.gr, mary.baltazani@phon.ox.ac.uk

ABSTRACT

This study examines the temporal characteristics of utterances, phrases, words and segments in coordinated structures with different phrasings in Greek. Eight speakers (4F, 4M) produced utterances with proper names (henceforth N) coordinated by [ce] 'and' in three phrasing conditions [N+N+N], N+[N+N], and [N+N]+N, in normal and fast speech rates. Ns were disyllabic or trisyllabic with penultimate or antepenultimate stress. Results show that phrasing and speech rate influenced different measures. Preboundary lengthening affected the phrase-final word, with largest lengthening on the phrase-final syllable, but extending up to the stressed syllable and affecting segments to different degrees. Utterance, phrase, word, and segment durations were shorter in fast than in normal rate. Overall, results on the effect of phrasing on duration support previous literature on Greek and other languages showing lengthening that extends beyond the final syllable. Novel findings on temporal patterns are reported by the combined effect of phrasing and speech rate.

Keywords: prosodic phrasing, duration, pre-boundary lengthening, speech rate

1. INTRODUCTION

The syntax-prosody relationship has been established in several languages [1, 2]. Syntactic phrase boundaries can be indicated prosodically through language-specific acoustic correlates: pauses, pre-boundary lengthening, sandhi and f_0 changes are reported as cues in adult speech processing, infant acquisition and speech impairments [1-4].

However, it is still debated whether such acoustic correlates of prosodic boundaries are consistently found in speech [5] and if so, whether they are reliable cues to syntactic units [6]. It is therefore necessary to establish the existence of these prosodic correlates in a specific language before investigating their use as prosodic cues in perception.

Little work has been done on this topic in Greek. So far, there have been reports on the importance of pitch cues, pre-boundary lengthening and segmental sandhi in the parsing of relative clause attachment and other types of structural ambiguities [7, 8, 9], but their presence in speech production is less well understood.

The presence of phrase-final lengthening in speech production has been established in Greek ([10-14]) and it is reported to occur mainly in the final pre-boundary syllable but also, to a smaller degree, in non-final stressed syllables. Cross-linguistically, and in Greek, little is known about the possible extent of pre-boundary lengthening in syllables preceding the stressed one. In [15] (and references therein) it is reported that lengthening decreases the further away a syllable is from a phrasal boundary and that it is mostly rimes rather than onsets that lengthen. One question addressed here is the extent of phrase-final lengthening in Greek, by including syllables preceding the stressed one in tri-syllabic words. The π-gestures theory [16] predicts stronger effects near boundaries without room for discontinuity, i.e., skipping syllables. It is interesting to test these predictions by examining possible variation in the scope of lengthening under different conditions of stress, number of syllables and segmental content.

We also investigate how speakers manipulate the presence and duration of pauses to disambiguate structurally ambiguous phrases, another under-researched question (cf. [14]).

Finally, we look into the effect of speech rate on phrase-final lengthening and pausing. Cross-linguistically there have been reports of modifications in the prosodic organization of phrases at fast rate ([17, 18]). Although segmental duration is in general reduced at fast rate, pre-boundary lengthening still marks strong prosodic boundaries ([17]). These effects have received little attention in the Greek literature (but see [14] for a recent study).

2. METHODOLOGY

Eight native speakers of Standard Modern Greek (4F, 4M), age range 28-57, with no history of hearing or speech problems, produced structures with three proper names coordinated by the conjunction [ce] 'and' in three phrasing conditions: A: [N+N+N], B: N+[N+N], and C: [N+N]+N. Three triads of proper names were used: 1. [ˈmina ce ˈnina ce ˈlina], with disyllabic penultimate stressed Ns; 2. [maˈrina ce meˈlina ce maˈnina], with trisyllabic penultimate stressed Ns; 3. [ˈelena ce ˈartemi ce ˈlaura], with trisyllabic antepenultimate stressed Ns. Each triad was

produced in the three phrasings (A, B, C), two speech rates (comfortable/normal vs. fast), and three repetitions, by eight speakers, resulting in a total of 432 tokens.

Sentences of type A phrasing form a single phrase, while sentences B and C form two phrases. The written sentences were presented to the speakers with brackets indicating the three different phrasings (A, B, C) in random order.

Recordings were carried out in a sound-treated studio with a Røde NT1-A cardioid condenser microphone. Annotations and segmentations were carried out in Praat [19]. Segmentation at utterance, phrase, word and segmental level was carried out.

We hypothesize phrasing will affect all unit durations (i.e., utterance, phrase, word, segment), with longer durations next to stronger (phrase) than weaker (word) boundaries due to pre-boundary lengthening, e.g., longer durations for the first proper name in phrasing B than in A. Furthermore, because of previous reports in the literature [11, 15], we measured all segments in each N to determine if the scope of pre-boundary lengthening extends to syllables preceding the final one. With reference to speech rate, we hypothesize that the magnitude of the durational effects will be reduced at the faster speech rate. Regarding pauses, as the conjunction 'ce' has a voiceless stop in onset position, the closure duration of [c] could not be distinguished from a possible preceding pause. Results are reported for this (possibly combined) interval and compared across phrasings.

Mixed model ANOVAs tested each temporal measure as dependent variable. Note the difference in stress position between triad 3 (antepenult) and triads 1 and 2 (both penult), due to which triad 3 underwent a separate statistical analysis from triads 1 and 2. Independent variables included phrasing (A, B, C), speech rate (normal, fast) as fixed factors, and speaker (1-8) as random factor. For the statistical analysis of triads 1 and 2, the additional fixed factor of interstress interval was included (i.e., the number of unstressed syllables between consecutive stresses: 2 or 3 syllables).

3. RESULTS

3.1. Utterance and word duration

Utterance duration was affected by phrasing ($F(2, 417)=63.49$, $p < 0.001$), with [NNN] utterances significantly shorter than N[NN] or [NN]N ones (Table 1), but no significant differences between the last two. The speech rate main effect was also significant ($F(1, 417)=468.64$, $p < 0.001$); utterances produced at fast speech rate were significantly shorter than those produced at normal rate (mean: 1353 ms vs. 1673 ms). There was a significant speech rate by phrasing interaction ($F(2, 417)=4.46$, $p=0.012$); N[NN] and [NN]N utterances were reduced by 20% in fast speech and NNN ones by 17%.

	Utter	N1	N2	N3	ce1	ce2
NNN	1396	356	342	421	140	139
N[NN]	1585	453	333	424	240	136
[NN]N	1558	332	441	429	140	216

Table 1: Utterance, N1, N2, N3, and conjunction ce1 and ce2 durations (ms) by phrasing; ce durations may include pause durations (see text for details).

Phrasing also affected word duration, with the first N significantly longer ($F(2, 415)=214.05$, $p < 0.001$) in the N[NN] phrasing and the second N ($F(2, 416)=135.66$, $p < 0.001$) in the [NN]N phrasing (N1 & N2 in Table 1) due to pre-boundary lengthening. There was also an effect of speech rate, with both nouns significantly longer in normal (N1: 424 ms; N2: 413 ms) than in fast (N1: 337 ms; N2: 330 ms) speech rate (N1: $F(1, 415)=291.23$, $p < 0.001$; N2: $F(1, 416)=197.38$, $p < 0.001$), but no interactions between phrasing and speech rate. As expected, utterance final Ns were not affected by phrasing (N3 in Table 1) but only by speech rate (normal: 451 ms vs. fast: 398 ms), ($F(1, 414)=84.15$, $p < 0.001$).

Finally, results for the conjunction [ce] durations, which can include pause durations, showed that the first [ce] was significantly longer in the N[NN] phrasing ($F(2, 416)=155.36$, $p < 0.001$) and the second [ce] in the [NN]N phrasing ($F(2, 416)=179.99$, $p < 0.001$) (Table 1). Speech rate affected the duration of the conjunctions (ce1: $F(1,416)=82.75$, $p<0.001$; ce2: $F(1,416)=146.72$, $p<0.001$) with significantly longer durations in normal than fast rate (ce1: normal: 198 ms, fast: 149 ms; ce2: normal: 187 ms, fast: 140 ms).

3.2. Segmental duration

Segmental duration was first examined in triads 1 & 2. Initially, we report segmental duration in the stressed and the post-stressed syllable of the first two Ns, which can occur next to stronger or weaker phrase boundaries. Due to the large number of analyses, most F and p values are included in endnotes. Phrasing had a significant effect on all segments except for the consonant in the stressed syllable of N1[i]. Since the latter was the bilabial nasal in triad 1 and the alveolar tap in triad 2, we ran separate analyses on these consonants and found a significant phrasing effect on /m/ ($F(2,130)=3.43$, $p=0.035$) but not on the tap, suggesting that

preboundary lengthening is conditioned by the segmental make-up of the segmental string.

Figure 1 shows longer durations for all segments of the first N in the phrasing N[NN] (top panel, left), and of the second N in the phrasing [NN]N (top panel, right), i.e., next to a phrasal boundary in both cases, compared to the other phrasings. Segmental durations for both N1 and N2 in the other two phrasings were shorter; post-hoc tests showed no significant differences between them. For both N1 and N2, we observe the greatest lengthening on the final vowel before a phrase boundary (Figure 1 and Table 2).

	strC	strV	final_C	final_V
N1/n	4	32	16	93
N1/f	15	32	41	75
N2/n	9	26	34	99
N2/f	20	23	28	72

Table 2: Increase (%) in segmental duration of N1 in N[NN] phrasing and N2 in [NN]N phrasing in normal (n) and fast (f) speech rates for triads 1&2. % calculated in relation to the means of corresponding Ns in the other phrasings.

We also examined possible durational variation due to differences in the interstress interval, i.e., two unstressed syllables between the stressed ones in triad 1 vs. three unstressed syllables in triad 2. We report results on identical syllables in the triads, i.e., the segments of the final syllable of N1 [na] in [mina] and [marina] and the final syllable of N2 [na] in [nina] and [melina] so as to avoid the influence of inherent durational differences and related contextual effects among different segments. The duration of all segments differed significantly[ii]. Post-hoc tests showed significantly shorter segmental durations in triad 2, which has a longer interstress interval of three syllables. This was also the case for the vowel of the [ce] conjunctions which was significantly shorter in triad 2.

All segment durations were significantly shorter in fast speech[iii]. Differences in the percentage increase of segmental duration in pre-boundary position in the two rates can be seen in Table 2. A larger percentage increase for the final vowel is noted in normal than fast rate of production while greater percentage increase may be seen for some consonants in fast speech. Table 2 also shows greater lengthening effects on the vowel compared to the consonant in each syllable and a reduction in the magnitude of effects as we move away from the phrasal boundary.

Presence of a pause at phrase boundaries was examined by comparing the duration of the prevocalic interval of the conjunction [ce] in the three phrasing conditions. An effect of phrasing and rate was found (endnotes i, iii). For the first [ce], longer duration was found in the N[NN] utterances (193 ms vs. 84 for NNN and 80 for [NN]N) and for the second [ce] in the [NN]N utterances (162 ms vs. 81 for NNN and 78 for N[NN]), as expected. These large differences are interpreted to suggest the presence of a pause. Longer durations in normal than fast rate were also found (ce1: 141 vs 98; ce2: 124 vs. 87).

In order to examine the potential presence of effects beyond the stressed syllable, we ran an analysis on the segmental durations in the first unstressed syllables of N1 and N2 in the trisyllabic words (first syllables in [marina] and [melina]). No phrasing effects were found on the initial consonant. Significant effects were found on the vowel (N1: $F(1,130)=12.40$, $p<0.001$; N2: $F(1,129)=3.35$, $p=0.038$); post-hoc tests showed that duration decreased in the order N[NN]=NNN>[NN]N for N1 and [NN]N=NNN>N[NN] for N2. These results suggest that there was not a preboundary lengthening effect beyond the stressed syllable.

	strV	midC	midV	fin_C	fin_V
N1/n	17	2	13	25	82
N1/f	16	8	11	31	81
N2/n	10	12	18	26	174
N2/f	21	16	20	17	173

Table 3: Increase (%) in segmental duration of N1 in N[NN] phrasing and N2 in [NN]N phrasing in normal (n) and fast (f) speech rates for triad 3. % calculated in relation to the means of corresponding Ns in the other phrasings.

Finally, similar results were found in triad 3. Phrasing had a significant effect on all segments[iv]. Figure 1 shows longer durations of all segments of the first N in the phrasing N[NN] (bottom panel, left) and of the second N in the phrasing [NN]N (bottom panel, right). A significant effect of speech rate was also found on all segments[v] with shorter segmental durations in fast speech. Post-hoc analyses showed that significant lengthening reached the first stressed vowel of N1 only in normal rate. No significant lengthening effects were evident on the consonant of the first post-stressed syllable in either rate while for the vowel of this syllable significant effects were present only for N2. Significant lengthening effects were found systematically on both segments of the last syllable of N1 and N2 in both rates. Table 3 shows similar percentage increase in final vowel duration across speech rates, greater final vowel lengthening for N2, greater effects for vowels than consonants in each syllable and a reduction in the magnitude of effects as we move away from the phrasal boundary. Finally, regarding the prevocalic interval of [ce],

longer duration was found in the N[NN] utterances for the first [ce] (173 ms vs. 100 for NNN and 96 [NN]N) and in the [NN]N utterances for the second [ce] (158 ms vs. 83 for both NNN and N[NN]), suggesting the presence of a pause. Longer durations were present in normal than fast rates (ce1: 137 vs 109; ce2: 124 vs 92).

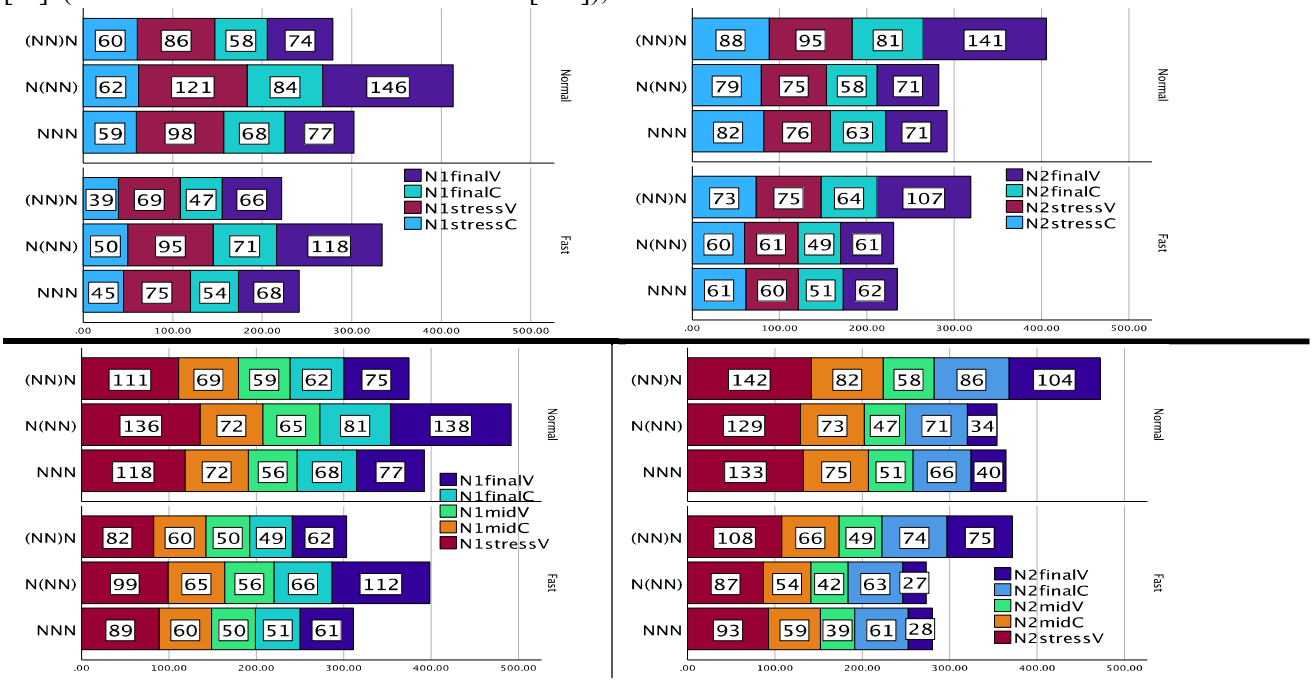


Figure 1: Triads 1, 2 (top) and 3 (bottom): N1 and N2 segmental durations by phrasing and speech rate.

4. DISCUSSION

In agreement with previous literature [10-14], this study shows that although pre-boundary lengthening affects mostly the final syllable and especially the final vowel, it extends to earlier syllables but with gradually reduced degree the further away from the boundary. Lengthening affects vowels to a greater degree than onset consonants in each syllable.

Novel evidence on the scope of pre-boundary lengthening in Greek is also provided: it turns out that pre-boundary lengthening can extend up to the stressed syllable but not to the syllable preceding it. Lengthening effects were evident not only on the stressed penultimate (e.g., on the [mi] in ['mina]) but also on the stressed antepenultimate syllable to some degree, i.e., only for N1 in normal rate (i.e., on the [e] of ['elena]). Thus, variation in the scope of effects was evident in triad 3 words. Interestingly, for N1 in normal rate, effects were not significant for the middle post-tonic syllable (typically a reduced syllable in Greek [20, 21], which may thus constrain lengthening effects) suggesting that effects can skip syllables and be attracted by stress, cf. [11, 17]. For N2, effects were not significant beyond the vowel of the middle post-tonic syllable (i.e., the [e] of ['artemi]). These results suggest that stress and rate modulate lengthening effects. In addition, syllable structure and segmental parameters, as discussed below, may play a role.

Novel findings are reported concerning the relationship between lengthening effects and segmental production constraints. Notably, effects varied among onsets with, for instance, presence of lengthening on a bilabial /m/ in the onset of a stressed syllable but absence of effects on a tap in the same position. Such differences suggest that segmental production requirements constrain lengthening as, for instance, it is not articulatorily possible to prolong a tap without compromising its identity.

Lengthening effects were present in both normal and fast rates of production with the magnitude of effects differing among consonants and vowels in the two rates and between triads 1/2 and 3 (Tables 2 & 3). The results suggest that lengthening still marks prosodic boundaries in fast rates; however, temporal constraints due to rate and stress result in modifications in the prosodic organization of phrases.

Finally, in addition to lengthening, phrasing was found to be marked by pauses as suggested by the results of the prevocalic interval of the conjunction [ce]. Intervals, and thus partially pauses, were also shorter in fast rates of production.

Work is underway that examines other acoustic parameters, i.e., f_0, to explore its contribution to marking prosodic boundaries in both rates of production.

6. REFERENCES

[1] Price, P. J., Ostendorf, M., Shattuck- Hufnagel, S. Fong, C. 1991. The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am*. 90, 2956–2970.

[2] Féry, C., Hörnig, R., Pahaut, S. 2011. Correlates of Phrasing in French and German from an Experiment with Semi-Spontaneous Speech. In Gabriel, C., Lleó, C. (eds), *Intonational Phrasing in Romance and Germanic*. John Benjamins,11–41.

[3] Johnson, E. K., and Seidl, A. 2008. Clause segmentation by 6-monthold infants: a crosslinguistic perspective. *Infancy* 13, 440–455.

[4] Titone, D. A., Koh, C. K., Kjelgaard, M. M., Bruce, S., Speer, S. R., Wingfield, A. 2006. Age-related impairments in the revision of syntactic misanalyses: Effects of prosody. *Lang. Speech* 49, 75–99.

[5] Fernald, A., McRoberts, G. 1996. Prosodic bootstrapping: a critical analysis of the argument and the evidence. In Morgan, J. L., Demuth, K.(eds), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, L. Erlbaum Assoc., 365–388.

[6] Peters, B. 2005. Weiterführende Untersuchungen zu prosodischen Grenzen in deutscher Spontansprache. In Kohler, K. J., Kleber, F., Peters, B. (eds), *Prosodic Structures in German Spontaneous Speech,* IPDS, 203–345.

[7] Papadopoulou, D., Clahsen, H. 2003. Parsing Strategies in L1 and L2 Sentence Processing: A Study of Relative Clause Attachment in Greek. *Studies in Second Language Acquisition* 25, 501–528.

[8] Papangeli A., Marinis, Th. 2010. Επεξεργασία δομικά αμφίσημων προτάσεων στην ελληνική ως Γ1 και ως Γ2 (Processing of ambiguous sentences in L1 and L2 Greek). *Studies in Greek Linguistics* 30, 477–486.

[9] Tserdanelis, G. 2005. *The Role of Segmental Sandhi in the Parsing of Speech: Evidence from Greek*. Ph.D. dissertation, The Ohio State University.

[10] Kainada, E. 2007. Prosodic boundary effects on durations and vowel hiatus in modern Greek. *Proc. 16th ICPhS, Saarbrücken,* 1225–1228.

[11] Kainada, E. 2010. Boundary-related durations in Modern Greek. *3rd ExLing Proceedings*, 69–72.

[12] Katsika, A. (2016) The role of prominence in determining the scope of boundary-related lengthening in Greek. *J. Phon,* 55, 149-181.

[13] Stavropoulou P., Baltazani, M. 2021. The prosody of correction and contrast. *J. Pragmatics.* 171, 76–100.

[14] Baltazani, M., Nicolaidis, K. 2022. Phrasing and speech rate effects on segmental and prosodic variability in Greek. *11th Speech Prosody*, 896–900.

[15] Byrd D, Krivokapić J, Lee S. 2006. How far, how long: on the temporal scope of prosodic boundary effects. *J Acoust Soc Am*. 120(3):1589–99.

[16] Byrd, D., Saltzman, E. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *J. Phon.* 31(2), 149-180.

[17] Fougeron, C., Jun, S.-A. 1998. Rate Effects on French Intonation: Phonetic Realization and Prosodic Organization. *J. Phon.* 26(1) 45–70.

[18] Michelas, A., d'Imperio. M. Durational Cues and Prosodic Phrasing in French: Evidence for the Intermediate Phrase. *Speech Prosody* 2010, paper 881.

[19] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 8 September 2018.

[20] Arvaniti, A. 1991. The Phonetics of Modern Greek Rhythm and its Phonological Implications. PhD dissertation, University of Cambridge.

[21] Baltazani, M. 2007. Prosodic rhythm and the status of vowel reduction in Greek. In *Sel. Papers on Theoretical and Applied Linguistics from the 17th ISTAL*, 1, AUTH, 31–43.

[i] Statistical information for the phrasing factor in triads 1&2.

	N1		N2	
	F (2,267)	p	F (2,267)	p
StrC	3	=0.051	22.6	<0.001
StrV	139.42	<0.001	82.43	<0.001
Final_C	159.04	<0.001	106.19	<0.001
Final_V	368.86	<0.001	282.58	<0.001
Ce_C	116.15	<0.001	155.13	<0.001

[ii] Statistical information for the interstress interval factor in triads 1&2.

	N1		N2	
	F (1,267)	p	F (1,267)	p
Final_C	61.16	<0.001	29.21	<0.001
Final_V	40.27	<0.001	5.27	=0.022
Ce_C	2.67	=0.103	3.61	=0.058
Ce_V	286.25	<0.001	406.59	<0.001

[iii] Statistical information for the speech rate factor in triads 1&2.

	N1		N2	
	F (1,267)	p	F (1,267)	p
StrC	75.81	<0.001	170.13	<0.001
StrV	204.75	<0.001	190.34	<0.001
Final_C	133.10	<0.001	130.07	<0.001
Final_V	50.77	<0.001	61.07	<0.001
Ce_C	38.45	<0.001	81.20	<0.001

[iv] Statistical information for the phrasing factor in triad 3.

.	N1		N2	
	F (2,130)	p	F (2,130)	p
StrV	18.57	<0.001	8.92	<0.001
Mid_C	3.08	=0.049	7.91	=0.001
Mid_V	12.40	<0.001	18.22	<0.001
Final_C	46.19	<0.001	32.60	<0.001
Final_V	143.63	<0.001	114.10	<0.001
Ce_C	77.39	<0.001	61.72	<0.001

[v] Statistical information for the speech rate factor in triad 3.

5.	N1		N2	
	F (1,130)	p	F (1,130)	p
StrV	125.79	<0.001	123.78	<0.001
Mid_C	44.25	<0.001	57.73	<0.001
Mid_V	33.99	<0.001	36.24	<0.001
Final_C	80.10	<0.001	18.29	<0.001
Final_V	32.67	<0.001	21.81	<0.001
Ce_C	25.19	<0.001	26.36	<0.001

F0-based Pairwise Variability Index: A Prosodic Metric for Holistic Language Processing

Alvin Cheng-Hsien Chen

Department of English, National Taiwan Normal University, Taiwan
alvinchen@ntnu.edu.tw

ABSTRACT

This study examined the effectiveness of a pitch variability metric—f0-based Pairwise Variability Index (f0PVI), which measures the pairwise f0 variability of a speech fragment. Analyzing spontaneous speech production, we asked (a) whether multisyllabic words' f0PVI values are smaller than nonwords'; (b) how multisyllabic words' f0PVI values are connected to lexical frequencies; and (c) to what extent the f0PVI values are conditioned on different representative f0 values of syllables. Multisyllabic words above a frequency cut-off in a native corpus were defined, and a subset of them used in a spontaneous speech corpus were particularly identified for the analyses of the relationship between their f0PVI in the speech corpus and their lexical frequencies in the native corpus. Results show f0PVI effectively differentiated words from nonwords and significantly correlated with lexical frequencies. We discuss the role of f0PVI in uncovering speakers' statistical knowledge in language processing.

Keywords: pitch variability, PVI, frequency effects, usage-based, spontaneous speech

1. INTRODUCTION

Human speech, although rich in forms and combinations, is always subject to the limitation of the biomechanism system. Studies have shown that there seems to be an upper limit (i.e., maximum speed) for human speakers to change their pitch voluntarily [1-3]. Along with the economy of efforts suggested by [4], pitch variation should probably be produced at the time when its pragmatic function is the most readily needed in a communicative setting. For example, [5] observed that pitch change in spontaneous speech production is closely connected to the semantic coherence and completeness of the proposition to be articulated. This suggests that significant pitch change often occurs at the semantic completion of the speaker's intended message. One corollary is that the pitch variation of a holistic linguistic unit is kept to the minimum compared to the pitch variation of a complex structure spanning several fundamental units.

In this study, we take on this connection between pitch variability and holistic language processing, and further investigate whether the pitch variability of a lexical unit is connected to its lexical frequency in use. We present a case study from spontaneous speech production in Taiwan Mandarin.

There are three main goals in this study. First, we experiment with a pitch encoding metric, which is based on the idea of duration-based variability, and validate its effectiveness in measuring the pitch-related variability of a speech segment. Specifically, we extend the computation of the Pairwise Variability Index (PVI), which is used to measure cross-linguistic durational variability, to pitch variability by creating the f0-based Pairwise Variability Index (f0PVI). The duration-based PVI was first proposed by [6] to capture the rhythmic pattern of a speech segment (e.g., a phrase) by analyzing every local durational variation of the composing sub-units (e.g., syllables of a phrase). In this study, we build upon the PVI's ability to capture local variability and explore its potential to measure local pitch variability of a speech segment.

Second, we aim to determine the effectiveness of the f0PVI by comparing the f0PVI values of disyllabic words to those of nonwords' (i.e., any two-syllable combinations that span the word boundary). Our hypothesis is that a holistic unit such as a word should consistently exhibit smaller degrees of pitch variability, reflected in lower f0PVI values, compared to a nonword.

Finally, we will investigate the extent to which the f0PVI values of disyllabic words are associated with their frequency of use. This investigation is particularly significant in language processing because the systematic prosodic encodings contributed by distributional/statistical properties of language can provide strong empirical evidence of the development of usage-based native intuition [5, 7, 8].

This study addresses three questions: (a) whether disyllabic words consistently demonstrate lower degrees of pitch variability than disyllabic nonwords; (b) to what extent the pitch variability of disyllabic words are connected to lexical frequencies; (c) to what extent the analysis of the pitch variability metric (f0PVI values) is dependent on the selection of different representative f0 values. To achieve these objectives, we selected disyllabic words above a

frequency threshold in a native corpus of Taiwan Mandarin and identified a subset of these words for analysis of the relationship between their f0PVI in a spontaneous speech corpus and their lexical frequencies in the native corpus.

2. METHOD

2.1. Data

The native corpus used in this study was the Academia Sinica Balanced Corpus of Mandarin Chinese (Sinica Corpus) [9], an eleven-million-word text collection tht covers a wide range of topics and genres. The speech data for pitch analysis came from the Sinica Phone-Aligned Chinese Conversational Speech Database (SPCCSD) [10], which includes approximately 3.5 hours of face-to-face free conversations from 16 speakers (7 males and 9 females; age range: 16–46). The SPCCSD has been time-aligned at the segmental level by the corpus provider, which allows for acoustic and prosodic analysis.

2.2. Disyllabic words

Due to the prevalence of disyllabic words in Mandarin (i.e., words consisting of two Chinese characters, e.g., *tongcai* 同儕 'peer', *shiyie* 事業 'career'), they are ideal candidates for pitch variability analysis. We selected all disyllabic words with frequency values above 5 in the Sinica Corpus as potential units for pitch analysis. Subsequently, a subset of these words was identified in the speech corpus of the SPCCSD for the purpose of examining the relationship between their f0PVI values in the speech corpus and their lexical frequencies in the native corpus.

2.3. Pitch variability metric: f0PVI

In this study, we are concerned with pitch variability of a speech fragment, which refers to the f0 variations of the base units within the speech fragment. In this section, we outline the method used to compute the pitch-related variability of disyllabic words observed in the SPCCSD.

First, we extracted the raw f0 values of each syllable using Praat's autocorrelation-based pitch tracking algorithm, with gender-dependent pitch ranges (75-300 Hz for males, and 100-500 Hz for females). The hertz values were converted into semitones using a 50 Hz base. Second, we computed the pairwise local f0 variability by selecting a representative f0 value from each syllable. To determine a representative f0 value, we explored several alternatives, including (a) the f0 value at the maximum energy (f0dbmax), (b) the maximum f0 value (f0max), (c) the minimum f0 value (f0min), and (d) the mean f0 (f0mean) value of the syllable. The f0dbmax may deserve more explanation. It is important to note that the f0dbmax represents the f0 value at the time point of the maximal decibel value in the syllable. This energy-based extraction of f0 values has been shown to produce more reliable f0 values and reduce the likelihood of tracking errors [5, 11, 12]. For our analysis of the connection between pitch variability and lexical frequencies, we will use the f0dbmax as the basis. However, the other f0 values (i.e., f0max, f0min, and f0mean) will be used for critical comparison of their impact on the f0PVI computation.

The computation of the words' pitch variability was based on the PVI proposed by [6] for the study of the typological rhythmic patterns. This duration-based PVI was first proposed as a quantitative metric to measure the isochrony and/or rhythm of syllables in languages and help determine the timing patterns (e.g., syllable-timing or stress-timing) of languages [13]. The duration-based PVI in [6] calculates the differences of durations between successive pairs of phonological units. For example, given a speech segment of n syllables, we first need to obtain the duration values of all the syllables (i.e., D_k), sum all the durational differences between each successive pair of syllables (i.e., $\sum_{k=2}^{n}|D_k - D_{k-1}|$), and compute the mean difference, as formulated in (1). Because the PVI based on (1) can potentially be scaled up due to the slow speech rates, [13] introduced a normalization mechanism to mitigate the influence of speech rates, as formulated in (2).

$$(1)\ \text{Duration-based PVI} = \frac{\left[\sum_{k=2}^{n}|D_k - D_{k-1}|\right]}{n-1}$$

$$(2)\ \text{Normalized Duration-based PVI} = 100 \times \left[\frac{\sum_{k=2}^{n}\left|\frac{D_k - D_{k-1}}{(D_k + D_{k-1})/2}\right|}{n-1}\right]$$

In this study, we operationalized an f0-based normalized PVI (f0PVI) to quantify the pitch variability of a speech segment. The f0PVI in (3) captures the variability of f0 differences between each successive pair of phonological units (i.e., syllables) in a speech fragment:

$$(3)\ \text{f0PVI} = 100 \times \left[\frac{\sum_{k=2}^{n}\left|\frac{f0_k - f0_{k-1}}{(f0_k + f0_{k-1})/2}\right|}{n-1}\right]$$

In contrast, the earlier duration-based PVI in (2) measures the variability of durational differences between syllables. The key trick was to determine a

representative f0 value for each syllable within the speech segment. In this study, we computed the f0PVI values based on four different f0 values extracted from each syllable of the disyllabic word: f0dbmax, f0max, f0min, and f0mean, which are represented by $f0_k$ in (3). The f0PVI is an intuitive and informative prosodic metric that quantifies the degree of pitch variability in a speech fragment. Higher f0PVI values indicate greater pitch variability (less regularity), while lower f0PVI values suggest a more monotonous f0 variation within the fragment.

2.4. Research design

The working hypothesis is that a holistic unit should be processed more cohesively, thus demonstrating smaller pitch variability (i.e., smaller f0PVI values) when compared to non-holistic units. To test this hypothesis, we carried out three analyses to determine the effectiveness of f0PVI and its connection to holistic language processing. The first two experiments used the f0PVI values computed based on the f0dbmax.

In Analysis 1, we examined the effectiveness of f0PVI in discriminating between disyllabic words and nonwords. Specifically, we compared the f0PVI values of all the disyllabic words with those of disyllabic nonwords (i.e., all nonword character bigrams) observed in the SPCCSD. If f0PVI is indicative of the extent of holistic processing by speakers, we would expect the f0PVI values of disyllabic words to be significantly lower than those of nonword character bigrams.

In Analysis 2, we investigated the relationship between the disyllabic words' f0PVI values and their lexical frequencies. We posited that as the lexical frequency of a disyllabic word increases, it is more likely to be processed holistically, resulting in a smaller f0PVI value.

In Analysis 3, we assessed the potential impact of different representative f0 values used in f0PVI computation on the results of the first two analyses. In particular, we investigated whether using f0max, f0min, or f0mean instead of f0dbmax for computing f0PVI would lead to different generalizations in distinguishing disyllabic words from nonwords and the relationship between f0PVI values and lexical frequencies of disyllabic words.

2.5. Statistical analysis

A linear mixed-effect analysis was used in this study with the f0PVI value as the dependent variable and the disyllabic wordness (in Analysis 1) and (log-transformed) lexical frequency (in Analysis 2) as the experimental fixed effects. Speaker-based intercepts were included as the random variable. In a spontaneous speech setting, pitch variability can arise from several factors, such as segmental structures, syllabic structures, relative position in utterance, speech rates, prosodic phrasing, and information structure. Therefore, we included several essential control factors in the mixed-effect model: (a) the relative position of the word in a speaker turn; (b) whether the word is at the onset/end of an intonation unit or a pause; and (c) the phonemic length of the word. We compared models using likelihood ratio tests to examine the significance of experimental fixed effects.

3. RESULTS

Disyllabic words and nonwords that could not be extracted with reliable f0 values due to speech reduction were removed from analysis. After filtering, we identified 12,824 tokens of disyllabic words in use from the speech corpus of SPCCSD, encompassing 2,246 word types. In SPCCSD, every character bigram spanning the word boundary formed a potential disyllabic nonword, amounting to 12,109 tokens (7,151 types).

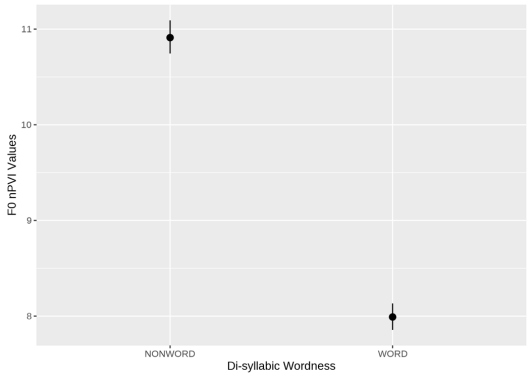


Figure 1: F0PVI means and confidence intervals of disyllabic words and nonwords

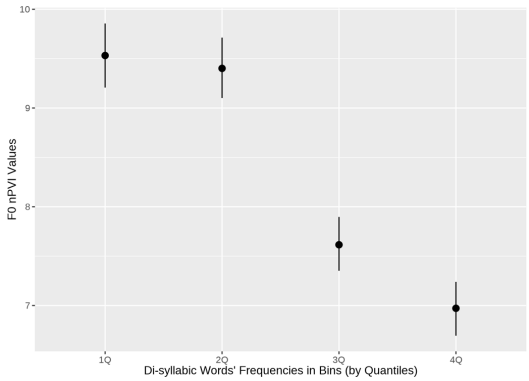


Figure 2: F0PVI means and confidence intervals of disyllabic words by lexical frequency bins

The results of Analysis 1 show that disyllabic words' f0PVI values were significantly lower than nonwords ($\chi^2 = 651.15$, $p < .01$), as shown in Figure 1.

This finding provides evidence that f0PVI can serve as an indicator of segmental integrity and a holistic unit (e.g., a disyllabic word) exhibits consistently lower pitch variability.

In Analysis 2, a significant correlation was observed between di-syllabic words' f0PVI values and their frequency values ($\chi^2 = 215.53, p < .01$). Figure 2 depicts the frequency effects, displaying the f0PVI mean values for disyllabic words divided into four frequency bins based on quantiles. The findings suggest that there is a negative relationship between word frequency and pitch variability, with higher frequency words showing smaller pitch variability.

In the final analysis, we examined whether the choice of representative f0 values for f0PVI computation could impact the results obtained in the previous analyses. Specifically, we assessed whether using f0max, f0min, and f0mean instead of f0dbmax to compute f0PVI values would lead to different generalizations in Analyses 1 and 2. However, the results showed no significant differences, reinforcing the robustness of the findings from the previous two analyses.

4. DISCUSSIONS

Holistic processing is a crucial area of study in psycholinguistics because it has significant implications for the psycholinguistic validity of a linguistic structure. To investigate holistic processing, researchers often examine the durational patterns of linguistic units. Reduction, in particular, is a frequently observed prosodic cue that reflects speakers' holistic processing of entrenched linguistic units [7, 14]. High-frequency words are produced and/or processed faster than low-frequency words, providing evidence for frequency-based effects on duration. This effect has also been observed with multiword units, including idioms [15], binomial phrases [16], collocations [17, 18], and even compositional multiword sequences [19-22].

On the other hand, pitch variation has more often been analyzed as a prosodic cue for discourse disjuncture at various levels [5], and its connection to the idea of holistic processing is relatively underexplored. Although human speech is rich in forms and combinations, it is always subject to the constraints of the biomechanism system.

Particularly relevant to the present study is an upper limit (i.e., maximum speed) for human speakers to voluntarily modulate their pitch [1-3]. Additionally, in accordance with the economy of effort principle proposed by [4], stronger pitch variation should be more likely to be produced only when its pragmatic function is most readily needed in a communicative setting. For example, [5] has shown that pitch change in speech production is closely connected to the semantic coherence and completeness of the proposition to be articulated. Significant pitch change occurs at important semantic boundaries of the speaker's intended message. A corollary of this is that the pitch variability of non-compositional basic linguistic units such as words should be minimized, optimizing communicative needs and the ease of articulatory efforts.

This study has developed an intuitive pitch variability index—f0PVI, which effectively distinguishes between (a) words and nonwords and (b) words at different frequency ranges. Although f0PVI calculation relies on a representative f0 value from the base units (i.e., syllables) of the speech segment (i.e., disyllabic word), our results demonstrate that f0 selection has no significant effect on the strong relationship between f0PVI and lexical frequency.

Our findings are consistent with and extend previous usage-based research by highlighting pitch variability as a strategic cue that reflects speakers' sensitivity to distributional properties in language use [5]. The pitch variability index can serve as a prosodic indicator of the degree of entrenchment of a linguistic unit in language processing. As a lexical unit is repeatedly used, it becomes progressively entrenched to the point of being processed holistically. The inverse relationship between a word's pitch variability in production and its frequency of use provides clear empirical support for this usage-based grammatical competence. In addition, the straightforward and intuitive nature of f0PVI suggests its potential for application in analyzing pitch variation with other more complex linguistic structures.

5. CONCLUSIONS

This study makes significant contributions in three key areas. First, we have introduced and validated a pitch variability metric for characterizing the holistic processing of linguistic units. Second, our focus on pitch-related encodings provides a novel and complementary perspective on the frequency effects observed in usage-based research. Finally, our analysis of spontaneous speech highlights the critical role of speakers' usage-based competence in online speech production and has important implications for psycholinguistic research.

6. ACKNOWLEDGEMENTS

This work was partly supported by National Science and Technology Council, Taiwan (110-2410-H-003-014 and 111-2410-H-003-030).

7. REFERENCES

[1] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America,* vol. 111, pp. 1399-1413, 2002.

[2] J. Sundberg, "Maximum speed of pitch changes in singers and untrained subjects," *Journal of Phonetics,* vol. 7, no. 2, pp. 71-79, 1979.

[3] J. J. Ohala and W. G. Ewan, "Speed of pitch change," *The Journal of the Acoustical Society of America,* vol. 53, no. 1, pp. 345-345, 1973.

[4] B. Lindblom, "Economy of speech gestures," in *The production of speech*, P. F. MacNeilage Ed. New York, NY: Springer, 1983, pp. 217-245.

[5] A. C.-H. Chen and S.-C. Tseng, "Prosodic encoding in Mandarin spontaneous speech: Evidence for clause-based advanced planning in language production," *Journal of Phonetics,* vol. 76, pp. 1-22, 2019, doi: 10.1016/j.wocn.2019.100912.

[6] E. L. Low, E. Grabe, and F. Nolan, "Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English," *Language and Speech,* vol. 43, no. 4, pp. 377-401, 2000.

[7] A. C.-H. Chen, "Durational patterns of recurrent multiword combinations in Mandarin spontaneous speech production," *Language and Speech,* vol. 64, no. 3, pp. 742-767, 2021, doi: 10.1177/0023830920966010.

[8] P. M. S. Lin, "The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus," *International Journal of Corpus Linguistics,* Article vol. 18, pp. 561-588, 2013, doi: 10.1075/ijcl.18.4.05lin.

[9] C.-R. Huang and K.-j. Chen, *Academia Sinica Balanced Corpus of Modern Chinese 4.0*. Taipei, Taiwan: Academia Sinica, 2010.

[10] S.-C. Tseng, "Lexical Coverage in Taiwan Mandarin Conversation," *International Journal of Computational Linguistics and Chinese Language Processing,* vol. 18, no. 1, pp. 1-18, 2013.

[11] D. R. Ladd and C. Johnson, "Metrical factors in the scaling of sentence-initial accent peaks," *Phonetica,* vol. 44, pp. 238-245, 1987, doi: 10.1159/000261801.

[12] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech,* vol. 50, pp. 281-310, 2007, doi: 10.1177/00238309070500030101.

[13] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology 7*, C. Gussenhoven and N. Warner Eds. Hague: Mouton de Gruyter, 2002, pp. 515-546.

[14] N. C. Ellis, "Frequency effects in language processing," *Studies in Second Language Acquisition,* vol. 24, pp. 143-188, 2002, doi: 10.1017.S0272263102002024.

[15] A. Siyanova-Chanturia, K. Conklin, and N. Schmitt, "Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers," *Second Language Research,* vol. 27, pp. 251-272, 2011.

[16] A. Siyanova-Chanturia, K. Conklin, and W. J. B. van Heuven, "Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences," *Journal of Experimental Psychology: Human Learning and Memory,* vol. 37, pp. 776-784, 2011, doi: 10.1037/a0022531.

[17] P. Durrant and A. Doherty, "Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming," *Corpus Linguistics and Linguistic Theory,* vol. 6, pp. 125-155, 2010, doi: 10.1515/cllt.2010.006

[18] S. Sonbul, "Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing," *Bilingualism: Language and Cognition,* vol. 18, pp. 419-437, 2015, doi: 10.1017/S1366728914000674.

[19] I. Arnon and N. Snider, "More than words: Frequency effects for multi-word phrases," *Journal of Memory and Language,* vol. 62, pp. 67-82, 2010, doi: 10.1016/j.jml.2009.09.005.

[20] C. Bannard and D. Matthews, "Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations," *Psychological Science,* vol. 19, pp. 241-248, 2008, doi: 10.1111/j.1467-9280.2008.02075.x.

[21] A. Tremblay and H. Baayen, "Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall," in *Perspectives on formulaic language: Acquisition and communication*, D. Wood Ed. London: The Continuum International Publishing Group, 2010, pp. 151-173.

[22] A. Tremblay, B. Derwing, G. Libben, and C. Westbury, "Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks," *Language Learning,* vol. 61, pp. 569-613, 2011, doi: 10.1111/j.1467-9922.2010.00622.x

THE INTERACTION OF DISCOURSE MARKERS WITH PROSODIC BOUNDARIES IN ESP

Nicolas Ballier[1], Alice Henderson[2], Adrien Méli[3], Taylor Arnold[4]

[1]Université Paris Cité, CLILLAC-ARP & LLF, F-75013, ,[2]UGA, LIDILEM, [3]Université Paris Cité, CLILLAC-ARP, F-75013, [4]University of Richmond
nicolas.ballier@u-paris.fr, alice.henderson@univ-grenoble-alpes.fr, adrienmeli@gmail.com, taylor.arnold@richmond.edu

ABSTRACT

This paper discusses the acoustic and phonetic realisations of the discourse marker *so* in the EIIDA corpus. We examine how the various functional realisations of *so* are reflected in the F_1/F_2 acoustic space. Using manually annotated paratones and automatic prosodic labels (INTSINT) on the EIIDA corpus, we examine the different correlates for the functional realisations of *so*. We use duration, rhythmic measures, paratone boundaries and INTSINT values as potential cues for the detection of functional realisations.

Keywords: English, discourse markers, paratones, prosodic boundaries, INTSINT.

1. INTRODUCTION

Discourse markers (DMs) are optional, prosodically distinctive "sharing devices and intimacy signals in our everyday talk" [1] and appropriate use demonstrates a speaker's pragmatic and communicative competence (e.g., [2] ; [3]). The functions of DMs have been categorized in different ways, e.g., ideational, textual and/or interpersonal for [4], or for [5] referential, structural, cognitive and/or interpersonal. We are particularly interested in how the interpersonal function surfaces even in well-prepared, non-spontaneous conference talks. Few studies have examined the prosodic environment of DMs in detail [6], and EFL textbooks tend to give only broad guidelines, if any. Corpus studies of authentic use would make it possible to improve guidelines, as some studies have done, e.g., *so* and *well* have been the focus of research. Within the various theoretical and methodological approaches used to study DMs, researchers have examined their prosodic environment in an attempt to correlate prosodic properties with different functions. For example, the prosodic contours of *now* were analysed by [7] to distinguish its sentential and discursive functions and [8] distinguished between semantic and pragmatic connectives, noting that the latter "will often be sentence-initial, followed by a pause and uttered with specific intonation contour".

Our study focused on *so*, for which previous analyses have provided "a fairly scattered picture of [its] functional potential" [4]. Moreover, although all of *so*'s functions occur across diverse text types [2], research on *so* has rarely included prosodic features [6]. As Wennerstrom notes "Because the functions of markers are so broad, any and all analyses of markers [...] can teach us something about their role in discourse"[9].

To that end, we analysed four conference talks by native English speakers for the four aspects studied by Wennerstrom (frequency of use, word position, function, intonational profile), as well as vowel quality. First, we identified the prosodic traits of each token, using manually annotated paratones and automatic INTSINT prosodic labels. Then, we categorized the function of each token as either adverbial or DM by using both lexicogrammatical context (e.g., adverbials retain semantic content and tend not to occur initially), and [10] suggested three conditions for DMs. She posited that they must usually be found in initial position, be sequentially dependent elements which bracket units of talk, and have a range of prosodic contours. Our two research questions are: RQ1: Do the intonational forms or the prosodic profiles of *so* distinguish its adverbial uses from those as a DM? RQ2: Does vowel quality reflect the differences in use? Results are discussed in terms of functional uses of *so*, intonational profile and vowel quality.

The rest of the paper is organised as follows: Section 2 summarises previous research on the analysis of *so* as a discourse marker (DM). Section 3 describes the corpus data and the method of automatic prosodic annotation. Section 4 presents the results and discusses the lexical environment of prosodic boundaries, especially the combinations of discourse markers (lexical bundles). Section 5 discusses vowel quality in relation to speech rate and

coalescence and outlines further research.

2. PREVIOUS RESEARCH

In spite of exploratory analyses of the "organization of conversational interaction" [11], researchers insist on the scarcity of research into the prosody of discourse markers [4]. *So* has been analysed as produced as a single unit, as a stand-alone object in interactional contexts [12]. It has also been analysed in opposition [13] or in combination with other markers. Its interaction with other markers such as *because* [14] has been analysed. [15] and [16] have investigated the combinatorial properties of *so*, demonstrating fewer sequencing constraints than previously claimed for discourse markers. The combining properties of *so* have also been addressed in relation to some constructions. [17] describes the complex emergence of the construction of *so to speak/say* and its interactions with parentheticals. Its hedging function is treated on a par with parentheticals like *as it were* and *if you like*. Different speech situations have been addressed, in dialogical or monological contexts such as seminars [18], in native or non-native speech [3], where it is more frequently used as "I mean" [6].

On the semantics-pragmatics interface, *so* has been analysed as an inferential marker [19] expressing result, inference, connection, topic development or action prompting [20]. In a corpus of video-mediated English as a Lingua Franca, [20] have analysed collocates and frequencies of the uses of *so*, insisting on examples where it is used as 'prompting inference'.

Phonetically, according to the *Longman Pronunciation Dictionary*, "There is an occasional weak form" [21] and the *Cambridge English Pronouncing Dictionary* signals that "the weak form is used only rarely, and only in casual speech before adjectives and adverbs (e.g. *not so bad*)". Our hypothesis is that for discourse marker realisations, the full diphthong /əʊ/ is more likely to be used than in intensive or phraseological uses where schwa /ə/ is more likely. The alternative realisations of *so* when it is not a DM are summed up in [6] as 'an adverb of degree or manner' as 'marking purpose', as 'a part of fixed phrases (e.g., *and so on*)' or 'as a pro-form (e.g., *I think so*)'.

3. MATERIAL AND METHODS

3.1. The EIIDA Corpus

We focus on four lectures (two men, two women) with General American pronunciation models. The speaking style can be defined as corresponding to well-prepared conference talks.

The EIIDA corpus is one of the first multilingual spoken corpora of specialized academic language. It was the main deliverable of a 2012-17 project coordinated by Shirley Carter-Thomas and Jeanne-Marie Debaisieux, from the LATTICE (UMR 8094) research group at the Ecole Normale Supérieure and Université Paris 3 Sorbonne Nouvelle.

EIIDA corpus academic presentations were recorded in different conditions, so all audio files were normalized for sample rate (11,000 Hz), format (wav) and channels (mono) using ffmpeg. Volume normalization across files was carried out following the European Broadcasting Union recommendations, by applying a filter implementing the R128 algorithm.

EIIDA can be used to carry out comparative linguistic analyses on written and spoken academic discourse (research articles vs. conference presentations). It is in three languages - English, French, Spanish - thus facilitating the analysis of the impact of writers' / speakers' linguistic culture on these two modes of communication. Furthermore, as the corpus has written and spoken forms of discourse from geochemistry and linguistics, disciplinary comparisons can be established. The whole spoken corpus corresponds to roughly 20 hours of audio recordings (300,000 tokens). Only recordings by native English speakers were used in the current study. The four manually annotated talks of 22, 34, 28 and 31 minutes are taken from a conference in linguistics (The John Swales Conference). But for one recording, they have a similar speech rate (4.22, 4.24 and 4.26 syllables per second, following [22]).

3.2. Automatic Prosodic Labels

The automatic prosodic labels for pitch were obtained using INTSINT (International Transcription System for Intonation), [23] using SPPAS [24] and a routine to split the talk recordings into smaller files. INTSINT consists in 8 labels representing the annotation: T (Top), M (mid), B (bottom), H (Higher), L (Lower), S (Same), U (Upstepped) and D (Downstepped). To operationalise the detection of rhythmic changes, following De Jong & Wempe's algorithm [22], we extracted syllable peaks during the production of *so* and used their algorithm to count syllable peaks during the production of *so*.

3.3. Prosodic Cues

Following [25], we also took into account paratone boundaries, since paratones (roughly defined in that

study as an increase in pitch range at rhetorical junctures to signal topic shift) correlated significantly with the scores received at an exam for Mandarin speakers of English. For the correlates of paratones in the EIIDA corpus, we followed the results in [26] and in [27], revisiting the correlates postulated for paratones in [28], namely focusing on the acceleration of rhythm at the beginning of paratones. We measured duration, with the hypothesis that uses as DM were more likely to be longer, especially compared to phraseological uses such as *and so on*.

3.4. Hypotheses for Speech Cues of DM uses

For the uses of *so* as DM, we expected occurrences to be longer (duration in ms) and to be closer to diphthong realisations (longer trajectories in plots). Since these occurrences may be in the vicinity of paratone initial boundaries, we also expected them to be closer to paratone boundaries (we computed the mean distance to the preceding paratone boundary) and to verify some of its prosodic correlates (mostly falls in paratone intial contexts). For acceleration, we computed the number of syllable nuclei detected by De Jong and Wempe's algorithm [22] half a second before *so*. For intonation patterns, we expected INTSINT pitch targets to match the frequent Fall pattern identified in [27].

4. RESULTS

4.1. Lexical Bundles

In line with [16], we noticed a strong co-occurrence of *so* with other discourse markers such as *all right* and *OK*.

4.2. Duration

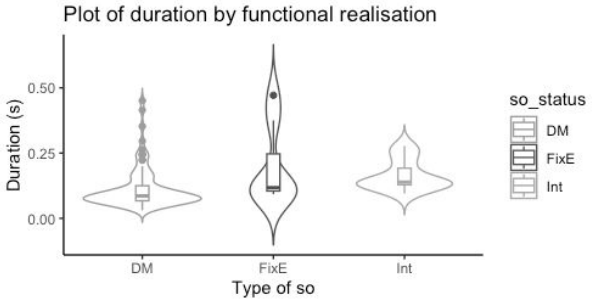


Figure 1: Violin plots of duration for the different realisations of *so* in our EIIDA data

The distribution of duration across types of *so* (see Figure 1) does not correspond to our assumptions for the distribution of duration between DM and Fixed expressions. In other cases, emphatic stress on adverbial *so* lengthened the vowel, so that the duration is not a reliable cue in our data to discriminate DM uses of *so*.

4.3. Acoustic Realisations

Figure 2 plots the diphthong realisations measured at 20% and 80% of the vowel duration of all the occurrences for talk # 2. In this talk, the diphthong is more marked for the intensive uses of *so* (see the black arrow). The other vowel plots similarly show that no complementary distribution between /əʊ/ for DM and /ə/ for adverbial or phraseological intensive uses can be observed. The two European reference

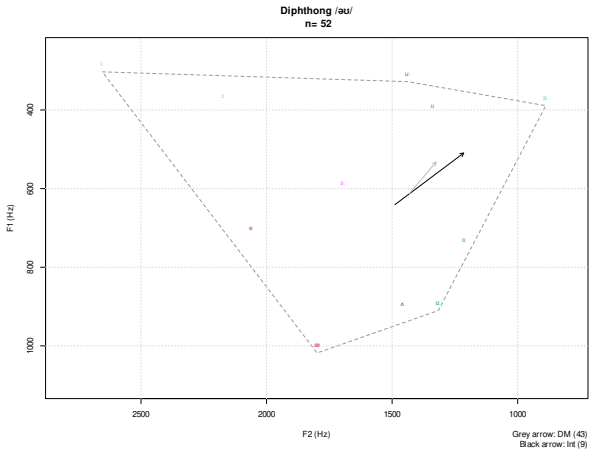


Figure 2: F_1/F_2 values (in Hz) formant plot of *so* in the EIIDA corpus (talk # 02)

pronunciation dictionaries ([21],[29]) do not assign phonetic realisations to categorical realisations for *so*, which they do for other grammatical markers (such as *that*). One way to interpret this result is reference dictionaries are well-advised not to mention categorical distributions for the realisations of *so*.

4.4. Prosodic Cues: INTSINT Labels

With the proviso that our dataset is limited, the INTSINT label associated to the different functional realisations of *so* do seem to indicate a pattern. The low counts and number of INTSINT labels do not lend themselves to parametric tests so we report the raw values for the INTSINT labels in Figure 3.

4.5. Rhythmic Cues: Paratone-like Signals

Taking the INTSINT representation of intonation, DM uses of *so* co-occur with a majority of H and T

	B	D	H	L	M	S	T	U
DM	4	6	9	4	10	3	4	4
FixE	2	1	0	0	0	0	1	3
Int	1	3	1	1	0	0	0	1

Figure 3: Distribution of INTSINT labels for the different realisations of *so* in our EIIDA data

labels, while other uses of *so* mostly co-occur with D , U and L. In other words, DMs tend to be associated with Falls, observed at the beginning of paratones. Not all the files were annotated for paratones, but Figure 4 represents the distance to the preceding initial boundary of the paratone. In spite of outliers, DM uses are earlier in the paratone and, for a fourth of them, used at the beginning of the paratone.

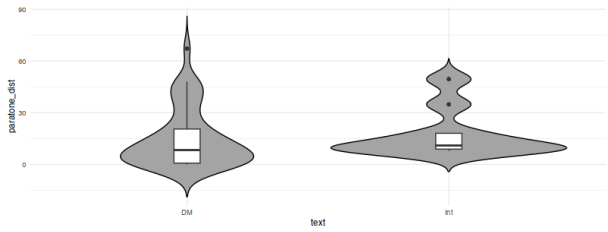


Figure 4: Distance (s) to paratone initial boundaries in talk # 02

5. DISCUSSION

5.1. Connected Speech Processes

While speakers relied sometimes on written notes and speech rate varied from 3.57 syllables per second to 4.26, connected speech processes were observed when investigating the vowel quality of *so*. Several cases of vowel coalescence complexified the analysis for cases like *so at, so we, so actually, so what* and probably account for the duration of some of the phraseological uses (like in *and so on*).

5.2. Prosodic Profiles and Pragmatic Uses

It remains to be seen if more subtle distinctions as listed in [20] are confirmed. They acknowledge no less than 10 pragmatic uses of *so*, some of them are mostly relevant in dialogues such as 'floor-holding device', 'emergence from incipient action', 'Topic developer' and 'Topic sequencer & connection marker Inference'. Some labels could be applied to monological speech such as 'self correction', 'summing up' or 'inferred and explicit result'. Careful manual annotation would be required for further analyses in this direction.

5.3. Further Research

For the detailed semantic interpretations of *so*, [6] notes ambiguous cases in the analysis, which could perhaps be resolved with multimodal cues. Analyzing *so* in multimodal data could be a way to co-cluster criterial features for the detection of DM uses of *so* if specific gestures co-occur with them. A great variability has to be noted as the number of occurrences for *so* ranges from 5 to 53 per lecture in our data, totalling 115 occurrences. These are split into realisations as 95 DM, 10 intensity (adverbial) uses (Int), 7 phraseolocical (fixed) expressions (FixE) and one occurrence as proform. The distribution of the different uses varies across speakers, so that this kind of analysis should be carried out on longer talks (or, for example, on a series of Youtube videos). In any case, the variability of the uses of *so* as a discourse marker is a good candidate to investigate idiosyncrasy. Our prosodic analysis of INTSINT labels could be tested on a larger scale, in a reverse engineering experiment where we would try to predict the status of *so* on the basis of these prosodic properties.

6. CONCLUSION

In this paper, we have analysed the acoustic and prosodic correlates of the uses of *so* in the context of four lectures in linguistics (English for Specific Purposes). For the majority of the occurrences in our data, it seems that prosodic cues are more robust than segmental cues for the automatic detection of the uses of *so*. In his analysis of discourse markers, Zwicky claims that "on the grounds of distribution, prosody, and meaning, discourse markers can be seen to form a class" [30] but does not elaborate on the prosodic properties. It may well be the case that for *so*, prosodic features play a more important role than segmental properties.

7. ACKNOWLEDGEMENTS

We thank the reviewers for comments on an earlier version of this paper. The prosodic annotation of the corpus was funded by a grant to Alice Henderson's lab from CORLI, the French Linguistic Consortium for Corpus, Language and Interaction, via their 2019 call for projects dedicated to corpus delivery (*Appel à Finalisation de corpus*).

8. REFERENCES

[1] Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. 1985. *A comprehensive grammar of the English language*. Longman.

[2] Lam, P. W. 2010. Discourse particles in corpus data and textbooks: The case of well. *Applied Linguistics* 31(2), 260–281.

[3] Müller, S. 2005. *Discourse markers in native and non-native English discourse* volume 138. John Benjamins Publishing.

[4] Buysse, L. 2012. So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics* 44(13), 1764–1782.

[5] Carter, R., Fung, L. 2007. Discourse markers and spoken english: Native and learner use in pedagogic settings. *Applied Linguistics* 28(3), 410–439.

[6] Buysse, L. 2009. So as a marker of elaboration in native and non-native speech. S. Slembrouck, S., Taverniers, M., Herreweghe, M. V. (eds), *From will to well: Studies in linguistics offered to Anne-Marie Simon-Vandenbergen*. Academia Press, 79–91.

[7] Hirschberg, J., Litman, D. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3), 501–530.

[8] Van Dijk, T. A. 1979. Pragmatic connectives. *Journal of Pragmatics* 3(5), 447–456.

[9] Wennerstrom, A. 2001. *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.

[10] Schiffrin, D. 1987. *Discourse markers: Studies in interactional sociolinguistics*. Cambridge University Press.

[11] Local, J., Kelly, J. 1986. Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 185–204.

[12] Raymond, G. 2004. Prompting action: The standalone "so" in ordinary conversation. *Research on Language and Social Interaction* 37(2), 185–218.

[13] Lam, P. W. Y. 2010. Toward a functional framework for discourse particles: A comparison of well and so. *Text & Talk* 30(6), 657–677.

[14] Passot, F. 2007. Circularity and discourse progression in conversational English. In: Celle, A., Huart, R. (eds), *Connectives as discourse landmarks* volume 161. John Benjamins Publishing, 117–134.

[15] Koops, C., Lohmann, A. 2013. Discourse marker sequencing and grammaticalization. *Annual Meeting of the Berkeley Linguistics Society* volume 39, 108–122.

[16] Koops, C., Lohmann, A. 2015. A quantitative approach to the grammaticalization of discourse markers: Evidence from their sequencing behavior. *International Journal of Corpus Linguistics* 20(2), 232–259.

[17] Claridge, C. 2013. The evolution of three pragmatic markers: As it were, so to speak/say and if you like. *Journal of Historical Pragmatics* 14(2), 161–184.

[18] Rendle-Short, J. 2003. "So what does this show us?": Analysis of the discourse marker 'so' in seminar talk. *Australian Review of Applied Linguistics* 26(2), 46–62.

[19] Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31(7), 931–952.

[20] Collet, C., Diemer, S., Brunner, M.-L. 2021. So in video-mediated communication in the expanding circle. *World Englishes* 40(4), 594–610.

[21] Wells, J. 2008. *Longman Pronunciation Dictionary*. London: Pearson Longman.

[22] De Jong, N. H., Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2), 385–390.

[23] Hirst, D. 2006. Phonetic and phonological annotation of speech prosody. In: Savy, R., Crocco, C. (eds), *(a cura di) Analisi prosodica. teorie, modelli e sistemi di annotazione, Atti del II Convegno Nazionale della Società Italiana di Scienze della Voce (AISV)*. Torriana: EDK, 24-29: Università degli Studi di Salerno 30 novembre-2 dicembre 2005.

[24] Bigi, B. 2015. SPPAS - Multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111-112, 54–69.

[25] Wennerstrom, A. 1998. Intonation as cohesion in academic discourse: A study of chinese speakers of english. *Studies in Second Language Acquisition* 20(1), 1–25.

[26] Méli, A., Ballier, N., Falaise, A., Henderson, A. 2021. An experiment in paratone detection in a prosodically annotated EAP spoken corpus. *Proc. Interspeech 2021*, 2616–2620.

[27] Ballier, N., Méli, A., Arnold, T., Henderson, A. 2022. Revisiting paratone prosodic features with the EIIDA corpus. *Proc. Speech Prosody 2022*, 450–454.

[28] Tench, P. 1996. *The intonation systems of English*. Bloomsbury Publishing.

[29] Jones, D. 2011. *Cambridge English pronouncing dictionary*. Cambridge University Press.

[30] Zwicky, A. M. 1985. Clitics and particles. *Language* 61(2), 283–305.

THE EFFECT OF F0 AND POST-FOCAL COMPRESSION ON THE PERCEPTION OF NARROW FOCUS IN HEBREW

Noam Amir[1], Vered Silber-Varod[2], Ori Anavi[1], Ella Deouell[1], Hansjörg Mixdorff[3]

[1]Department of Communication Disorders, Tel Aviv University
[2]Center for AI and Data Science (TAD), Tel Aviv University
[3]Berliner Hochschule für Technik
noama@tauex.tau.ac.il, other email addresses?

ABSTRACT

In this study, we examine the effect of shifts in fundamental frequency (F0) on the perception of narrow focus (NF). F0 was manipulated systematically on the first word in a subject-verb-object (SVO) sentence, and separately on the rest of the sentence, using the Fujisaki model of intonation. The original sentence was taken from an existing Hebrew corpus where it had previously been found to be uttered with narrow focus and measurable post focal compression (PFC). Findings indicate, as expected, that systematically lowering F0 on the initial word reduced recognition of NF. Systematically raising F0 in the rest of the sentence also reduced recognition of NF. Notably, raising F0 in the second part of the sentence reduced recognition rates, even when no F0 manipulation was performed on the first word. This indicates that PFC is not only a production strategy of Hebrew speakers but has a role in perception of NF also.

Keywords: Narrow Focus, Fujisaki Model, Post Focal Compression.

1. INTRODUCTION

Communication is based not only on words, but also on supra-segmental features that allow us to convey meanings beyond the words alone, such as intentions, feelings, irony and humor. Changes in prominence affect supra-segmental features, and make a part of a sentence (e.g., a word) stand out by changes in pitch, duration and intensity [1, 2, 3] and hence modifies the semantic focus. There are two types of focus, according to the length of the unit of the sentence being highlighted: Broad or "default" Focus (BF), which concerns the whole sentence, and Narrow Focus (NF), which emphasizes, for instance, one word within the sentence [4, 5]. The purpose of the focus is to convey new information to the listener [6]. Thus, the type of focus (broad vs. narrow) and its location in the sentence is based on the information the speaker wants to provide or the question that he or she is answering [7, 5].

In most languages, NF is created by increasing fundamental frequency, duration and intensity of the word being emphasized [8, 9, 10, 11, 12, 13]. There is evidence indicating that in some languages, F0 is the most important feature in the recognition of NF (as reviewed in Terken [14]). There is supporting evidence for this in Hebrew as well [15], though more research needs to be done.

Interestingly, it has been shown that in some languages, including Hebrew [17], acoustic changes do not occur only in the focused word but also in the words following it [16, 11]. Post-focal compression (PFC) is the compression in F0 of the units following the prominent word. Studies show that there are languages in which PFC is crucial to the perception of NF [16]. Evidence of that has been shown in a few studies, as Xu [16] reviewed, which found that NF perception worsens when the word being emphasized is at the end of the sentence, since there are no units after it that can undergo compression. Unpublished studies conducted by the present authors showed the same effect in Hebrew [17]. This indicates that there might be an effect of the PFC on the perception of NF in Hebrew, though this has not been directly examined yet. This was therefore the primary goal of the present study.

Stimuli used to assess the perception of NF can be produced by using computer software, which allows isolating and systematically changing the acoustic features in a chosen word in the sentence [3], such as the FujiParaEditor software based on the Fujisaki model in conjunction with Praat software. The Fujisaki model provides a representation of the F0 contour along the sentence using the superposition of the constant base frequency, the phrase components (manipulated by the phrase commands) and the accent components (manipulated by the accent commands). Neutral sentences are characterized with low and wide accent components while NF sentences are characterized with high and narrow accent components. Studies show that the amplitude of the accent commands manipulated by this model is highly correlated with judge's perception of NF [3].

In the present study, we tried to shed light on the

acoustic features contributing to the perception of NF in Hebrew. To this end we employed the Fujisaki model to create systematic changes in F0, separately on the focused word and on the units that follow it.

2. METHODS

A set of 300 three-word target stimuli and 100 non-targeted stimuli were presented to 20 listeners. Their task was to determine whether any one of the three words in each stimulus was emphasized, and to what degree. Further details are presented below.

2.1. Stimuli

The target stimuli were based on a single utterance recorded in a previous production study on narrow focus [17]. The selected utterance achieved a 70% recognition score for focus on the first word. This utterance was a three word sentence in Hebrew: *Xatul nixnas l-a-gan* (cat.SG.M enter.past.SG.M. to-the-garden) "A cat entered the garden". The intonation contour of this utterance was modified using software based on the Fujisaki model of intonation [18]. Analysis of the original intonation contour determined the presence of two accent commands – a strong one on the initial word, and a weaker one on the following two words, as shown in Figure 1.

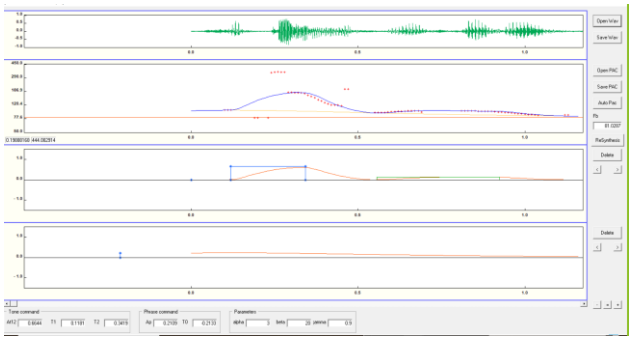


Figure 1: Screenshot from the FujiParaEditor.

The utterance was then resynthesized 30 times as follows: Leaving the second accent command intact, the first accent command was reduced in five increments, in order to lower the pitch peak in the first word in steps of 10 Hz, from its initial value of 176Hz down to 126Hz. Next, the second accent command was raised, to heighten the pitch peak in the second part of the utterance from 100Hz to 140Hz. Once again, six variations were created with six different heights in the first pitch peak. This was repeated five times. Thus, the second pitch peak attained a value of 150Hz in the final iteration. In total this process yielded 30 stimuli, comprising all the different combinations of six different initial peak heights and five different secondary pitch peak heights (Table 1).

Each of the resultant stimuli was presented for judgment 10 times, giving 300 presentations of the target stimuli. To avoid boredom and to prevent the listeners from guessing what manipulations were performed, 100 additional three-word sentences were presented, but not analysed.

Table 1: The different variations of the target stimulus.

F0 peak associated with the first accent component (F0_NF)	F0 peak associated with the second accent component (F0_PFC)
F0_NF1=176Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz
F0_NF2=166Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz
F0_NF3=156Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz
F0_NF4=146Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz
F0_NF5=136Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz
F0_NF6=126Hz	F0_PFC1=100Hz F0_PFC2=110Hz F0_PFC3=120Hz F0_PFC4=130Hz F0_PFC5=140Hz

2.2. Listeners

Twenty female listeners participated in the study (M=27.4, STD=2.5), aged 20 to 31.

2.3. Procedure

Stimuli were presented to the listeners using custom written software employing a graphic user interface (GUI). After hearing each stimulus, the listeners

could mark one of four radio buttons, indicating that word 1, 2 or 3 were emphasized, or that no word was emphasized. If options 1-3 were selected, listeners were asked to rate the degree of emphasis on a 5-point scale from weakly emphasized to highly emphasized (The latter was not further analysed in the present study). Stimuli were presented in a random order. Since each distinct stimulus was presented to each subject 10 times, each stimulus received a recognition score of 0 to 100%, from each individual listener.

3. RESULTS

Plotting overall mean percent-correct scores vs both manipulations, gave a 3D plot shown in Figure 2.

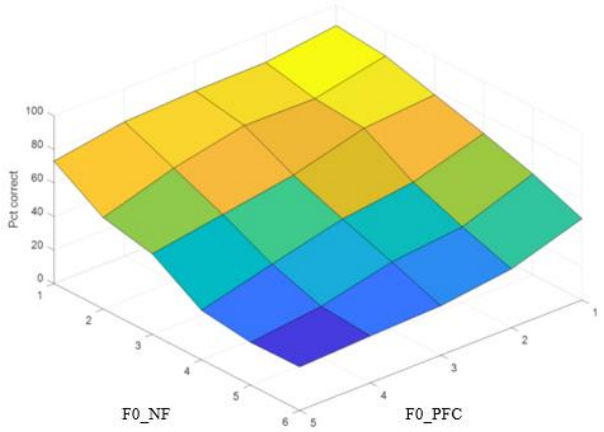


Figure 2: Mean recognition percentages of narrow focus as a function of both manipulations.

The plot clearly suggests that both manipulations affected recognition of narrow focus. It demonstrates that manipulation of the focused word alone has a strong effect, reducing mean recognition from nearly 100% in the unmanipulated condition to about 40% when manipulation was strongest (condition F0_NF6). Manipulation of PFC had a smaller effect, reducing mean recognition to about 70% in the most extreme manipulation (condition F0_PFC5). The plot also suggests that the two manipulations interact, so that combining a manipulation of the focused word with manipulation of PFC reduced recognition more than each manipulation alone. An exploratory ANOVA with repeated measures was therefore conducted with two within-subject factors: narrow focus manipulation (F0_NF, with 1 indicating no manipulation, up to 6 indicating a **reduction** of 50 Hz in main peak height) and post-focal-compression manipulation (F0_PFC) with 1 indicating no manipulation, up to 5 indicating an **increase** of 40 Hz in secondary peak height). Significant main effects were found for both F0_NF ($F(1.58,28.39)=47.11$, $p<0.001$) and F0_PFC ($F(1.46, 26.33)=28.64$, $p<0.001$) and their interaction ($F(20,360)=2.25$, $p=0.002$). This indicates that both factors affected narrow focus recognition significantly. The significant interaction, however, indicates that F0_PFC did not affect recognition in the same manner for different levels of F0_NF. This can be seen in the interaction plot in Figure 3, showing separate lines for the different values of F0_NF, with the x-axis indicating different levels of F0_PFC.

We wished to specifically examine whether PFC did indeed influence recognition scores at all levels of F0_NF, therefore we further conducted six ANOVAs with F0_PFC as a within subject factor, for each level of F0_NF separately. In effect, this is equivalent to examining each line in Figure 3 separately. A highly significant main effect was found in all six, as well as significant linear contrasts, as summarized in tables 2 and 3. This indicates that PFC has a significant effect on recognition of NF, with recognition rates decreasing linearly as PFC diminishes (i.e. as F0 increases on words following the focused word).

Table 2: significances of main effects for the six ANOVAs.

	df	*F*	*p*
F0_NF1	[4,76]	6.46	0.000
F0_NF2	[4,76]	14.43	0.000
F0_NF3	[4,76]	10.34	0.000
F0_NF4	[4,76]	14.12	0.000
F0_NF5	[4,76]	14.40	0.000
F0_NF6	[4,76]	7.04	0.001

Table 3: significances of linear contrasts for the six ANOVAs.

	df	*F*	*p*
F0_NF1	[1,19]	23.65	0.000
F0_NF2	[1,19]	31.38	0.000
F0_NF3	[1,19]	21.88	0.000
F0_NF4	[1,19]	20.40	0.000
F0_NF5	[1,19]	26.48	0.000
F0_NF6	[1,19]	10.76	0.004

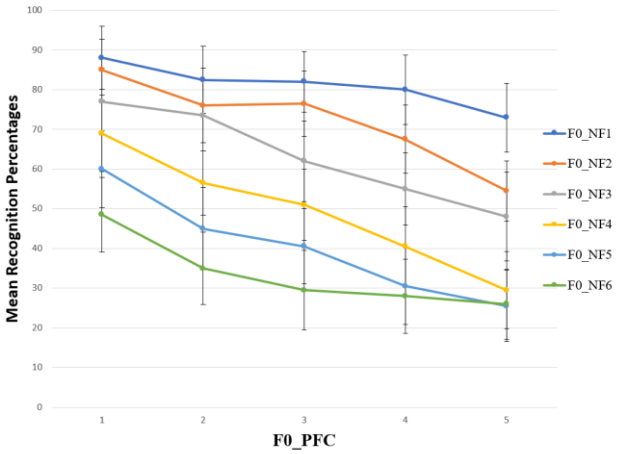


Figure 3: Mean recognition percentages of the narrow focus at the different levels of F0_PFC. Each line represents a different value of F0_NF.

4. DISCUSSION AND CONCLUSIONS

The results of this study highlight the role of F0 in creating the perception of narrow focus in Hebrew. As expected, the height of the pitch peak on the focused word has a major influence on narrow focus perception. PFC, however, influences this perception also. Even when the peak on the focused word is at its highest, reducing PFC (i.e. increasing the secondary pitch peak) reduces perception scores of narrow focus. This indicates that PFC is not only a production strategy, but that listeners rely on this cue also in judging NF.

Furthermore, the results indicate that narrow focus is not a categorical phenomenon in Hebrew. Gradual decrease in pitch on the focused word, and gradual decrease of PFC, resulted in gradual changes of NF recognition scores, in contrast to reports in the literature on other languages [7].

A future publication on data we have already collected [17] will address production strategies, demonstrating that PFC is indeed widely employed by speakers of modern Hebrew.

6. REFERENCES

[1] Raphael, L. J., Borden, G. J., Harris, K. S. 2007. *Speech science primer: Physiology, acoustics, and perception of speech,* 5th ed. Lippincott Williams & Wilkins.

[2] Lehiste I. 1970. *Suprasegmentals*. Cambridge, MIT, Press.

[3] Rao, P., Sanghvi, N., Mixdorff, H., Sabu, K. 2017. Acoustic correlates of focus in Marathi: Production and perception. *Journal of Phonetics*, 65, 110–125.

[4] Vilaplana, E. E., Maidment, J. 1999. Central Catalan declaratives: the relation between focus and downstep. *14th International Congress of Phonetic Sciences*, San Francisco, CA, United States.

[5] Sityaev, D., House, J. 2003. Phonetic and phonological correlates of broad, narrow and contrastive focus in English. *The 15th International Congress of Phonetic Sciences*, Barcelona, Spain.

[6] Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R., Lotts, D. W. 1986. Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech*, 29(3), 233–251.

[7] Ladd Jr, D. R. 1980. *The structure of intonational meaning: Evidence from English*. Indiana University Press.

[8] Breen, M., Fedorenko, E., Wagner, M., Gibson, E. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes*, 25(7–9), 1044–1098.

[9] Hanssen, J. E. G., Peters, J., Gussenhoven, C. 2008. Prosodic effects of focus in Dutch declaratives. *Speech Prosody 4thInternational Conference*, Campinas, Brazil.

[10] Lee, A. Xu, Y. 2012. Revisiting focus prosody in Japanese. *Speech Prosody 6th International Conference*, Shanghi, China.

[11] Mixdorff, H., Amir, N. 2002. The Prosody of Modern Hebrew-A Quantitative Study. *Speech Prosody International Conference*, Aix-en-Provence, France.

[12] Mixdorff, H., Cossio-Mercado, C., Hönemann, A., Gurlekian, J., Evin, D., Torres, H. 2015. Acoustic correlates of perceived syllable prominence in German. *16th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2015), Dresden, Germany.

[13] Xu, Y., Xu, C. X. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, 33(2), 159-197.

[14] Terken, J. 1991. Fundamental frequency and perceived prominence of accented syllables. The *Journal of the Acoustical Society of America*, 89(4), 1768–1776.

[15] Amir, N., Almogi, B. C., Mixdorff, H. 2008. A systematic framework for studying the contribution of F0 and duration to the perception of accented words in Hebrew. *Speech Prosody 4th International Conference*, Campinas, Brazil.

[16] Xu, Y. 2011. Post-focus Compression: Cross-linguistic Distribution and Historical Origin [Paper presentation]. *17th International Congress of Phonetic Sciences* (ICPhS 2011), Hong Kong, China.

[17] Bechar Marom, N., Hechter, N., Tahayu, Y., Amir, N., Silber-Varod, V. 2021. Prosodic focus in Hebrew: A Perception study. Proceedings of Tone and Intonation 2021 (TAI 2021), *The 1st Internaional conference on TAI* (pp. 192–193), Sønderborg, Denmark.

[18] Fujisaki, H. and Hirose, K. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5(4): pp. 233–241.

Gender-related variation of nasality and sound change of denasalization driven by prosodic boundaries in Seoul Korean

Jungah Lee[1], Sahyang Kim[2], Taehong Cho[1]

[1]Hanyang Institute for Phonetics & Cognitive Sciences of Language, Hanyang University (HIPCS);
[2]Hongik University
jlee14@uoregon.edu; sahyang@hongik.ac.kr; tcho@hanyang.ac.kr

ABSTRACT

This study investigates gender-related variation of nasal consonant duration and coarticulatory V-nasalization across two boundary conditions (IP-initial/IP-medial) in Seoul Korean. N-duration and V-nasalization (using A1-P0) of bisyllabic words (e.g., /mami/) were measured from 17 speakers in their 20s (F:8, M:9). IP-initially, N-duration was extremely reduced along with reduced V-nasalization as compared with IP-medial patterns, indicating that position-driven denasalization is underway. There were no gender-related effects on N-duration, but there was a significant Gender x Boundary interaction on V-nasalization. Females nasalized the vowel much more than males, and maintained a similar degree of V-nasalization across boundary conditions. In contrast, males substantially reduced V-nasalization IP-initially, suggesting that the on-going sound change of position-driven denasalization has been led by males, counter to the fact that sound changes are often female-led. The male-led denasalization process can be accounted for by different gender-related social meanings associated with nasality in interaction with boundary-related prosodic structure.

Keywords: nasalization, denasalization, Korean, prosodic boundary, sound change

1. INTRODUCTION

A linguistic message is encoded in a phonetic form that is inevitably variable, driven by various linguistic and non-linguistic factors. Speech variation is often deeply rooted in the phonetics-prosody interface, especially in how phonetic encoding of a linguistic message is fine-tuned by prosodic structure. A low-level phonetic process such as coarticulatory nasalization, phrase-final lengthening or non-contrastive creakiness, often considered to bring about non-contrastive phonetic variation, often turns out to be systematically modulated by prosodic structural factors such as stress hierarchy and the level of prosodic boundary [6, 14, 16]. For example, phonetic variation reflected in the degree of consonantal nasality and its influence on the vowel's coarticulatory nasalization can be controlled by the speaker in a language-specific way in reference to prosodic structure as discussed in various studies (Chinese [25]; French, [13]; English, [9, 10]; and Korean, [7, 17]). The converging evidence from these studies indicates that despite its cross-linguistic variation, the non-contrastive coarticulatory nasalization varies as a function of the strength of prosodic boundary and prominence system. Seoul Korean provides an interesting case because the nasality of the nasal consonant is extremely reduced in a phrase-initial position [17], compared to a nasal consonant that occurs in other languages such as English, French and Chinese [17]. For such an extreme reduction of the nasality, it is generally considered that Seoul Korean undergoes a sound change of 'denasalization' [21, 30, 31], which is taken to have originated from a boundary-induced reduction of nasality in phrase-initial position [17].

In the present study, we continue to explore the nature of the position-driven sound change of denasalization in Seoul Korean in conjunction with effects of gender as a sociolinguistic factor. It is well known that speech production often differs systematically between the genders [11, 23, 26]. In particular, female speakers are more likely to lead a new sound change than male speakers as female speakers tend to prefer to produce more innovative linguistic production [22, 26]. In fact, in Seoul Korean, young female speakers have led a recent on-going sound change, showing a merger of VOT difference between lenis and aspirated stop categories which is localized to phrase-initial position similar to the denasalization process [2, 10, 19]. One might therefore expect that female speakers will also play a leading role in the sound change of densalization process that are underway in Seoul Korean.

We, however, still leave an alternative possibility open. Female speakers are often thought to produce more articulate and clearer speech forms because they are more sensitive to produce standard and socially-upward prestigious linguistic forms [3, 24, 26]. The denasalization process results in a reduced form deviating from the canonical nasal sound. Moreover, a nasalized sound is often preferred in female speech especially in Asian speech communities especially in Japan and China [12, 32]. Thus, it is possible that

female speakers may turn out not to take a lead in the denasalization process.

In the present study, we will test these possibilities by examining how the nasality of the nasal consonant and the degree of vowel nasalization of bisyllabic words (/mami/ or /mima/) is modulated by prosodic boundary and how the position-sensitive denasalization process manifests itself differently between the female versus the male speech in Seoul Korean.

2. METHODOLOGY

2.2. Participants and recording

Seventeen native Seoul Korean speakers (F:8, M: 9) participated in the experiment as part of a large-scale acoustic-articulatory database building project in Korean at Hanyang Institute for Phonetics and Cognitive Science of Language [15]. All speakers were born and raised in Seoul and were students at Hanyang University in Seoul. They were all in their 20s (mean age: 23.4) and were paid for the participation. The acoustic data from the seventeen speakers were collected in a soundproof room with a Tascam HC-P2 digital recorder and SHURE KSN44 condenser microphone at a sampling rate of 44kHz, along with the articulatory data acquisition using Electromagnetic Articulograph (Carstens AG 501). We analyzed the acoustic data only for the purpose of the present study.

2.2. Speech materials and measurement

The HIPCS database corpus included several passages, and we used a passage that contains two bisyllabic words (names) with nasal consonants (/mami/, /mima/). The passage was originally constructed to induce these words to be produced in different prosodic boundary contexts. To elicit an IP boundary context, the target names were preceded by an adverbial phrase, which aided the speaker in inserting a phrase boundary as in (1a); and to elicit an IP-medial context, each target word was used as the second part of a two-word compound noun as in (1b). The whole passage will be available at https://osf.io/ecm6s/.

(1)

a …우리동네에 #미마네라는 빵집이 생겼는데.. (# = IP)
*…ulitoŋnɛɛ # mimanɛlanɯn p*aŋtʃipi sɛŋkjʌtnɯntɛ..*
'…the bakery called 'mima' opened in our neighborhood…'

b …계피#마미빵에 집중하게 되었어 (# = phrase-medial)
*…kjɛpʲi#mamip*aŋɛ tʃiptʃuŋhakɛ twɛʌs*ʌ*
'(they) got to specialize in cinnamon 'mami' bread…'

The participants were asked to read the passages as naturally as possible twice at a normal speech rate and twice at a fast speech rate, and we used the data produced at a normal speech rate. At the normal speech rate, speakers produced the target words most of the time in the intended prosodic boundary contexts, as checked by the authors. In total, 936 target word tokens were obtained to be used for analysis (/mami/, IP-initial, 345, IP-medial, 123;/mima/, IP-initial, 345, IP-medial, 123).

Following [9] and [17], to assess the nasality of the words, we measured the duration of nasal murmur for the consonant (N-duration) and the A1-P0 during the vowel to estimate the degree of vowel nasalization. A1-P0 is the difference in amplitude for the first formant and the amplitude of the spectral nasal peak that was detected around 250-450 Hz. As in [9], A1-P0 values were taken at relative time points (25%, 50%, and 75%) of the vowel duration, using a Praat script [29]. We also measured A1-P0 at absolute time points (20ms, 40ms, and 60ms) from the vowel onset, but the results are not included in the results section partly due to the space limit and partly due to the fact that absolute measures yielded more or less comparable results as relative measures. These measures were taken from both the first and the second syllable of each target word.

2.3. Statistical analyses

All the statistical analyses were performed by running linear mixed-effects models [1] in [27]. The first and the second syllables were analyzed separately. As in (2a), a model included N-duration as a dependent variable with Boundary and Gender as fixed factors (with their interactions). Another model as in (2b) included A1-P0 as a dependent variable with Gender, Boundary, and Timepoints as fixed factors (including their interactions). The factors were sum coded to interpret main effects and to compare conditions with the overall means. The models in (2) were obtained with the maximal random effects structure to the extent the model converged (i.e., random intercept by subject and slopes for fixed factors). Because the target words contained two vowel contexts (/i/ vs. /a/), and because A1-P0 values obtained from the high vowel context could be more error-prone due to the proximity of A1 and P0, we included Vowel Type as a fixed effect in the model, so that the results could be interpreted with a possible vowel height effect taken into account. When there was an interaction between Boundary and Gender, post-hoc t-tests with Tukey's HSD correction were further carried out to examine what would have caused the interaction effect.

(2) a. N-duration ~ Gender*Boundary*Vowel Type + (1 + Gender + Boundary+ Vowel Type|Speaker)
b. A1-P0 ~ Gender*Boundary*Vowel Type + Timepoint+ (1 + Gender + Boundary + Timepoint |Speaker)

3. RESULTS

3.1. Nasal duration

For the first syllable (NVNV), there was a significant effect of Boundary on N-duration (β = - 19.8 , SE = 1.9, t = -10.2, p < 0.001), showing that N-duration was *much shorter* IP-initially than IP- medially (Fig. 1a). There was, however, no significant interaction between Boundary and Gender (β = 5.2, SE = 2.7, t = 1.9, p = 0.07). As shown in Fig.1a, while there was no gender difference in the IP-initial condition, N-duration was significantly longer for females than for males in the IP-medial condition (IP-initial, Gender difference, β = -4.6, SE = 2.56, t = -1.8, p = 0.092; IP-medial, Gender difference, β = 10.6, SE = 3.66, t = 2.9, p = 0.0111).

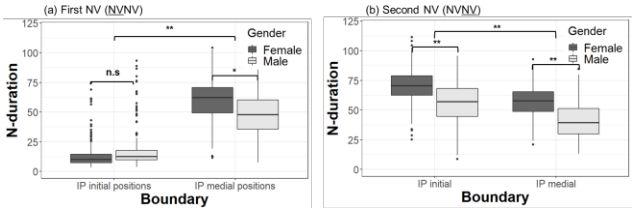


Figure 1. N-duration of first (a) and second (b) syllables across Boundary Positions

As for /n/ in the second syllable (NVNV) of the target word, there were significant effects of Boundary and Gender on N-duration (Boundary, β = 5.77, SE = 0.91, t = 6.3, p < 0.0001; Gender, β = 7.85, SE = 1.61, t = 4.86, p = 0.0001) with no Boundary x Gender interaction (β = 0.81, SE = 0.91, t = 0.89, p = 0.38). N-duration of the second syllable was significantly longer when the target word occurred in the IP-initial condition than when in the IP-medial condition; and it was also significantly longer for female speakers than for male speakers (see Fig.1b). These results indicate that even the non- edge (the second) syllable was influenced by the preceding boundary strength.

3.2. Vowel nasalization (A1-P0)

For the first syllable (NVNV), there was a significant effect of Boundary and Timepoint 75% on A1-P0 (β = 1.08, SE = 0.22, t = 4.8, p < 0.0001; β = -0.7, SE = 0.22, t = -3.17, p = 0.001, respectively) with no Boundary x Timepoint interaction, indicating that the degree of V-nasalization was generally greater in the IP-initial than in the IP-medial condition, and it got smaller later in the vowel. Crucially, however, there was a significant interaction between Gender and Boundary (β = -0.71, SE = 0.22, t = -3.15, p = 0.004) with no further interaction with Timepoint.

As can be seen in Fig.2a, the interaction stemmed from the fact that there was a significant gender difference only in the IP-initial context. Notably, female speakers nasalized the vowel much more than male speakers in the IP- initial context. Put differently, despite the fact that the N-duration itself was extremely reduced in the IP- initial context, female speakers still nasalized the vowel to a larger extent, as much as they did in the IP-medial context (β = -0.005, SE = 0.6, t = -0.009, p = 0.99); whereas the male speakers substantially reduced V-nasalization in the IP-initial position matched with the reduction of N-duration compared to when in the IP-medial position (β = 2.52, SE = 0.56, t = 4.46, p = 0.0005).

As for the V-nasalization of the second syllable, there was no effect of either Boundary or Gender on V-nasalization (β = 0.39, SE = 0.33, t = 1.2, p = 0.2; β = -0.05, SE = 0.6, t = -0.09, p = 0.9, respectively) nor was there an interaction between them (β = 0.39, SE = 0.33, t = 0.11, p = 0.9). But it is noteworthy that the overall nasalization was greater for the second syllable than for the first syllable, despite the fact that the first vowel was flanked by the nasal consonant. It appears that the vowel is nasalized progressively larger into the second syllable in the NVNV context.

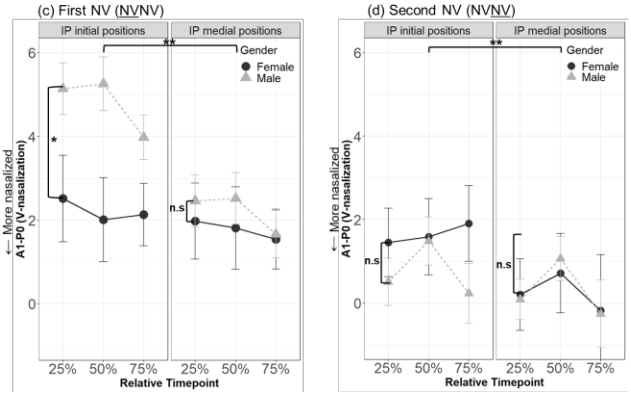


Figure 2. V-nasalization of first (c) and second (d) syllables across Boundary Positions

4. SUMMARY AND DISCUSSION

The present study examined how nasality of NVNV words (/mami/, /mima/) as reflected in N-duration and coarticulatory V-nasalization (A1-P0) would vary as a function of prosodic boundary in Seoul Korean and how the boundary-induced variation would be further conditioned by the gender difference. The results can be recapitulated as follows. First, N- duration of the first syllable (NVNV) was extremely shortened when in the IP-initial position compared to when in the IP-medial position with no further difference due to Gender. It is noteworthy that in a similar study on boundary effects on nasality in Seoul Korean, [17] showed that IP-initial N-duration was around 30 ms (estimated based on Figure 1c of [17]).

In the present study, N-duration was found to be even shorter (ca. 16ms). Although such a comparison would not provide an accurate difference between the two studies because of the different speech materials and experimental conditions, the even more extreme reduction of N-duration implies that denasalization may have progressed over the past several years. It is also important to note that the same word-initial N-duration did not undergo denasalization when the word occurred IP-medially, confirming that the on-going denasalization process must be taken as a position-driven sound change.

Let us now consider a possible origin of the sound change. A relevant question here concerns why to begin with N-duration is generally shorter in the IP-initial than in the IP-medial position, as has been observed across languages (Chinese, [25]; French, [13]; English, [8, 9]; and Korean, [7, 17]). The initial nasal reduction has been considered as a kind of domain-initial articulatory strengthening. It arises with an increased articulatory force at the beginning of a phrase that applies to an oral articulation which may in turn has an effect of elevating the velum [9, 13]. The denasalization process in Seoul Korean then establishes an interesting case of sound change whose origin has stemmed from prosodic-structurally conditioned articulatory strengthening. Given that domain-initial strengthening is more robust in Korean as an edge-prominence language, as discussed in [5, 20], than in other languages such as English as a head-prominence language (see [18] for the typological difference), the left-edge effect of boundary-related strengthening (also known as domain-initial strengthening) appears to have been exaggerated over time to the extent that the nasal murmur of the consonant may not be easily heard by the listener [21, 31]. (The reader is also referred to [10] for a similar prosody account for another on-going sound change of Seoul Korean.)

Most notably, however, our results indicate that such a prosodic-structurally driven sound change is being led by male speakers rather than female speakers. Recall that despite the fact that IP-initial N-duration was extremely reduced in both female and male speech alike, female speakers still nasalized the vowel substantially just like they did in the IP-medial position. In other words, the female speakers still maintain the nasality in the form of V-nasalization. This phenomenon can be understood in phonological terms. For female speakers, the [nasal] feature originally associated with the consonant has spread to the following vowel, but the [nasal] feature itself is being delinked from the consonantal node, leaving a substantial nasalization over the vowel. On the other hand, for male speakers, the denasalization process has progressed much further in such a way that the [nasal] feature is delinked from both the consonant and the vowel.

Another important question is why it is the male speakers that appear to have led the sound change of denasalization in Seoul Korean, especially given that a sound change is often led by female speakers [22]. We suggest that the male-led denasalization process is attributed to possible social meanings associated with phonetic nasality towards which the speech community is more positively inclined when produced by females than by males. In East Asian societies, linguistic forms that can characterize “immature, weak, feminine, friendly, polite, cute and passive” are often preferred in females’ speech. It is often assumed that Japanese and Chinese female speakers tend to employ high-pitched voice and more nasality to produce phonetic forms preferred for the female speech in the speech community they belong to [12, 32]. Since social meanings of linguistic varieties and variants are often formed differently between genders [4, 28], the resulting non-sonorancy of denasalization appears to carry a negative (possibly stigmatized) social meaning attached to female speech, thus being disfavored by female speakers. Another evidence of the female speakers’ preference of nasalization was also found in the present study. Recall that in the IP-medial position, female speakers produced a longer N-duration of the first syllable than male speakers did; and for the second syllable, female speakers produced a longer N-duration than male speakers did in both IP-initial and IP-medial positions.

5. CONCLUSION

The results of the present study elaborate on how prosodic-structurally conditioned phonetic variation of consonantal nasality and coarticulatory vowel nasalization may have developed into a sound change in Seoul Korean. The results also indicate that the on- going sound change is being led by males rather than females. We suggest that the target phonetic form (‘denasalization’) carries differential social meanings as a function of whether it is produced by males versus females, possibly leading to a stigmatized form in the latter case. More broadly, it is underscored that phonetic variation that is essentially modulated by prosodic structure can be further fine-tuned by social factors, an area of research that needs expanding to further our understanding of linguistic- phonetic underpinnings of speech variation in various social contexts.

6. ACKNOWLEDMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A02086884) awarded to TC (PI) and SK (Co-PI).

7. REFERENCES

[1] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

[2] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., & Yoon, T. J. (2018). The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120-144.

[3] Byrd, D. 1994. Relations of sex and dialect to reduction. *Speech communication*, *15*(1-2), 39-54.

[4] Campbell-Kibler, K. 2007. What did you think she'd say?: Expectations and sociolinguistic perception. In *annual conference on New Ways of Analyzing Variation. Philadelphia, Pennsylvania, October*.

[5] Cho, T. (2022). The phonetics-prosody interface and prosodic strengthening in Korean. In *The Cambridge Handbook of Korean Linguistics Cambridge Handbooks in Language and Linguistics* (pp. 248-293). Cambridge University Press Cambridge.

[6] Cho, T. 2022. Linguistic functions of prosody and its phonetic encoding with special reference to Korean. *Japanese/Korean Linguistics*, *29*.

[7] Cho, T., & Keating, P. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of phonetics*, *29*(2), 155-190.

[8] Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466-485.

[9] Cho, T., Kim, D., & Kim, S. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics*, *64*,71-89.

[10] Choi, J., Kim, S., & Cho, T. 2020. An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. *Plos one*, *15*(10), e0240682.

[11] Eckert, P. 1989. The whole woman: Sex and gender differences in variation. *Language variation and change*, *1*(3), 245-267.

[12] Endo, O. 2008. The role of court lady's language in the historical norm construction of Japanese women's language. *Gender & Language*, *2*(1).

[13] Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of phonetics*, *29*(2), 109-135.

[14] Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, *94*, 101155.

[15] Hanyang Institute for Phonetics and Cognitive Sciences of Language. 2022-, Dynamics of Speech Production through Articulatory DB Construction

[16] Iskarous, K., & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. *Journal of Phonetics*, *95*, 101195.

[17] Jang, J., Kim, S., & Cho, T. 2018. Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *The Journal of the Acoustical Society of America*, *144*(1), EL33-EL39.

[18] Jun, S. A. (Ed.). (2005). *Prosodic typology: The phonology of intonation and phrasing*. OUP Oxford.

[19] Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics*, *45*, 76-90.

[20] Keating, P., Cho, T., Fougeron, C., & Hsu, C. S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic interpretation: Papers in laboratory phonology VI*, 143-161.

[21] Kim, Y. S. (2011). *An acoustic, aerodynamic and perceptual investigation of word-initial denasalization in Korean* (Doctoral dissertation, UCL (University College London)).

[22] Labov, W. 1973. *Sociolinguistic patterns* (No. 4). University of Pennsylvania press.

[23] Labov, W. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change*, *2*(2), 205-254.

[24] Labov, W. 2006. *The social stratification of English in New York city*. Cambridge University Press.

[25] Li, H., Kim, S., & Cho, T. (2020). Prosodic structurally conditioned variation of coarticulatory vowel nasalization in Mandarin Chinese: Its language specificity and cross-linguistic generalizability. *The Journal of the Acoustical Society of America*, *148*(3), EL240-EL246.

[26] Oh, E. 2010. Speaker gender and the degree of coarticulation. Eoneo "Language", *35*(3), 743-766.

[27] R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

[28] Strand, E. A. 1999. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology*, *18*(1), 86-100.

[29] Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, *142*(4), 2469-2482.

[30] Yoo, K. & Nolan, F., (2020) "Sampling the progression of domain-initial denasalization in Seoul Korean", *Laboratory Phonology* 11(1): 22. doi: https://doi.org/10.5334/labphon.203

[31] Yoshida, K. (2008). Phonetic implementation of Korean denasalization and its variation related to prosody. *IULC Working Papers*, *8*(1).

[32] Yueh, H. I. S. 2016. *Identity politics and popular culture in Taiwan: A sajiao generation*. Lexington Books

WORD-LEVEL AND SYLLABLE-LEVEL PREDICTABILITY EFFECTS ON SYLLABLE DURATION IN TAIWAN MANDARIN

Sheng-Fu Wang, Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica, Taiwan
sftwang@gate.sinica.edu.tw, tsengsc@gate.sinica.edu.tw

ABSTRACT

This study examined how word-level and syllable-level predictability affects the variability of syllable duration in disyllabic words in Mandarin conversational speech. Specifically, we examined how the correlation between predictability measurements (unigram surprisal, bigram surprisal, and informativity) and syllable duration was affected by final lengthening, and whether word-level predictability effects were uniform across syllables. Our results showed that incorporating both levels of predictability variables improved model fit significantly compared to using only one level. We also found that predictability effects weakened at prosodic boundaries, and that word-level bigram surprisal had a stronger impact on syllables closer to relevant word boundaries, while word-level informativity and unigram surprisal were more predictive of duration for a word's first syllable. Overall, our findings highlight the relevance of different types and levels of statistical information for various aspects of durational variability in Mandarin.

Keywords: predictability, Mandarin, syllable, duration, word

1. INTRODUCTION

Linguistic units that occur more frequently or are more likely to occur in a given context are often reduced or shortened (e.g., [15, 2, 21, 12]). Typically, frequency and probability estimation is conducted at the word level, focusing on how word-level frequency and contextual probabilities affect word duration. Although some recent studies have shown this correlation at the syllable level [18, 22], the relationship between word and sub-word predictability measurements in shaping speech production remains underexplored. To complement these studies, we investigated how word-level and syllable-level predictability affect syllable duration in disyllabic words in Taiwan Mandarin.

This study aimed to provide a more nuanced understanding of how different levels of predictability contribute to predicting syllable duration in spontaneous conversations in Taiwan Mandarin. We investigate whether word-level and syllable-level predictability measurements both contribute to modeling durational variability at the syllable level. Answers to this question may further our understanding of the level(s) of representation relevant in speech production planning, which has been discussed as a potential driving force behind the observed predictability effect [2, 12].

This question is especially relevant for Mandarin, as it is a language with an abundance of standalone monosyllabic morphemes that are strongly associated with specific semantic content. In the context of research on the effects of frequency and contextual probability on phonetic realizations, this characteristic makes it more likely that syllable-level predictability plays an important role in speech planning.

Since the target acoustic variable was syllable duration, we were essentially testing to what extent frequency and probability estimates from a higher level percolate to the durational pattern of a lower-level unit. In addition, we investigated whether word-level predictability effects are uniform across syllables within a word. For example, in the word /tjan^{51} iŋ^{214}/ movie' in the context of the phrase /k^{h}an^{51} # tjan^{51} iŋ^{214}/ watch movies', we examined whether the word-level bigram probability P(movies|watch) predicts the duration of both syllables tjan^{51} and iŋ^{214} in the same way. If word-level predictability influences component syllables equally, it suggests that syllable-level speech planning is closely linked to word-level representations.

Finally, we examined these predictability effects' interactions with final lengthening. This continues a recent thread of studies that look at whether predictability effects are modulated by prosodic effects, which boundary marking is supposedly a part of. The rationale was that the highlighting of less predictable units has been served by boundary marking as part of the overall prosodic structure that mediates the relationship between predictability and

acoustic cues [1, 20, 11]. This research question also motivated the selection of syllable duration as the target, since previous studies on final lengthening mostly focus on syllable position [7, 6].

Three types of predictability measurements were examined: unigram surprisal (i.e., lexical frequency), bigram surprisal (i.e., contextual predictability at specific local contexts), and bigram informativity. Following [13, 14], we define bigram informativity as a word/syllable type's average bigram surprisal as estimated from a corpus.

2. METHOD

2.1. Speech corpus

We extracted durational data of Taiwan Mandarin from the Sinica Phone-aligned Chinese Conversational Speech Database (SPCCSD, Sinica File No. 24T-1031221; [19]). The corpus contains 3.5 hours of recordings with boundary annotation at the phone and syllable levels. Additionally, the corpus has word segmentation and part-of-speech tagging based on the CKIP system [4], which maps to a simplified tag set [10]. For research on discourse and prosody, the corpus also has annotations of prosodic units (PU), which are based on paralinguistic signals such as pause and inhalation, and discourse units (DU), which refer to units containing a predicate and its key arguments.

As previous studies have shown that PU boundaries with a match and mismatch to a DU boundary exhibit different acoustic cues (e.g., [3]), we controlled for such potential effects by only looking at PU boundaries that coincided with a DU boundary. We also limited our inquiry to syllables from disyllabic words so that we could see a clear picture of how word- and syllable-level predictability measurements compare in modeling phonetic variability. Overall, the analysis contained 7420 disyllabic words.

2.2. Written corpus and language models

For language modeling, we used the Academia Sinica Balanced Corpus of Mandarin Chinese [8, 4], which has 11M word tokens and 17.5M syllable/character tokens. Trigram language models were trained with modified Kesner-Ney smoothing [5] at the levels of words and syllables using the SRILM toolkit [16, 17]. In addition to training the model to obtain probabilities given the previous context, models that read sentences from the backward direction are also trained to obtain surprisal and informativity given the following context. This was motivated by research using the same method showing that informativity and surprisal given the following context account for more variances in word duration than informativity and surprisal given the previous context [15].

2.3. Variables in analyses

The analyses included five predictability variables: unigram surprisal ($-\log P(x)$), bigram forward/backward surprisal ($-\log P(x|context)$), and bigram forward/backward informativity ($-\sum_{\text{context}} P(\text{context}|x) \log P(x|\text{context})$), which were calculated for two unit types (word, syllable). All these variables were log-transformed (base 10) and normalized.

To examine the effect of syllable positions within a prosodic unit (PU), we included a predictor with four categories: initial, medial, penultimate, and final. We highlighted the penultimate position on top of the final position since it has been reported that Taiwan Mandarin has a robust disyllabic domain for final lengthening [7, 6].

Three other variables are included: Speech rate (a prosodic unit's syllable count divided by its length in milliseconds), syllable position (a syllable's position within a word), and each syllable's baseline duration. Following [18], baseline duration was the prediction of a linear regression model that predicted syllable duration based on the syllable's tone and segments.

3. RESULTS

3.1. Word- vs. syllable-level predictability

Three multivariate mixed-effects models were fit: A model with only word-level predictability measurements (*Word* model), a model with only syllable-level predictability measurements (*Syllable* model), and a model with both types of variables (*Full* model). Two likelihood ratio tests showed that the *Full* model provided a significantly better fit than the *Word* model [$\chi^2(5) = 197.95$, $p < .00001$] and the *Syllable* model [$\chi^2(5) = 193.45$, $p < .00001$].

Table 1 summarizes the *Full* model. The effect of within-PU position shows a penultimate and final lengthening, with the initial position not significantly different from the medial position ($p < .0001$ for all pairs of comparisons except for initial vs. medial). Most of the predictability variables had a positive estimate, i.e., higher surprisal/informativity was associated with longer syllable duration. It is worth noting that even though a few predictability variables had a negative estimate, it was likely a result of suppression, i.e.,

these variables had strong correlations with variables that predicted the dependent variable better [23]. Additional analyses showed that these variables had positive estimates when they were the only predictability variable in a model (Forward bigram word surp.: 0.15, unigram word surp.: 0.09, backward syll. inf.: 0.12, forward syll. surp.: 0.04).

Table 1: Summary of fixed effects in the mixed-effects model; **W**: word-level; **S**: syllable-level

	β	SE	t	$p(\chi^2)$
(Intercept)	2.02	0.04	49.88	
Baseline dur.	0.24	0.01	21.56	< .0001
Speech Rate	-0.31	0.01	-52.11	< .0001
N. density	-0.05	0.01	-4.10	< .001
W Forward Inf.	0.14	0.01	9.82	< .0001
W Backward Inf.	0.03	0.01	1.90	= .05
W Forward Sur.	-0.05	0.01	-6.55	< .0001
W Backward Sur.	0.07	0.01	5.95	< .0001
W Unigram Sur.	-0.05	0.01	-4.10	< .0001
S Forward Inf.	0.05	0.02	2.80	< .01
S Backward Inf.	-0.03	0.02	-1.91	= .06
S Forward Sur.	-0.04	0.01	-3.89	< .001
S Backward Sur.	0.11	0.01	9.18	< .0001
S Unigram Sur.	0.07	0.02	3.27	< .01
Initial	0.00	0.04	0.10	(< .0001)
Penult.	0.49	0.02	24.35	
Final	0.85	0.03	30.31	
2nd Syll in Word	-0.25	0.02	-11.16	< .0001

3.2. Directionality and boundary effects

For analyses in this section, we ran additional regression models where each model only had a predictability variable [23], and the three-way interaction between the predictability variable and syllable position in a prosodic unit and a word. The *lstrends()* function in the lsmeans [9] package was used for post hoc analyses. We use 'S1' and 'S2' to refer to the 1st and 2nd syllables in a word.

The upper panels in Figure 1 show the effect sizes of word-level bigram surprisal in different positions. Surprisal was only a positive predictor of duration in the medial position in both the forward (S1: β = 0.04, p < .0001; S2: β = 0.02, p < .05) and backward direction (S1: β = 0.07, p < .0001; S2: β = 0.21, p < .0001).

Comparisons of the effect size further show that forward surprisal had a stronger effect in S2 than S1 (β = 0.14, p < .0001), which follows the directionality hypothesis. The opposite trend was observed for backward surprisal, which had a stronger effect on S1, even though the difference did not reach statistical significance.

The lower panels in 1 show the effects of global word-level measurements of predictability. The effect of unigram surprisal was significant for both S1 & S2 within a word in the medial position (S1: β = 0.13, p < .0001; S2: β = 0.07, p < .0001) and for S1 in the penultimate position (β = 0.07, p < .01). The effect of forward informativity was significant at all positions (p < .05 for S1 in the final position and p < .0001 for other positions) except for S2 in the penultimate position. The effect of backward informativity was significant at all positions (p < .0001) except for S2 in the penultimate position and S1 in the final position. In other words, the effect of word-level unigram surprisal was neutralized at the final position similar to bigram surprisal, while the effect of informativity was not.

Comparisons of the effects in S1 and S2 show that unigram surprisal was a significantly or nearly significantly stronger predictor at S1 than S2 at the medial (β = 0.06, p < .0001) and the final position (β = 0.07, p = .05). For forward informativity, the same comparisons were significant in the same direction (medial: β = 0.05, p < .0001; penultimate: β = 0.08, p < .05). Finally, for backward informativity, the same comparison was only significant in the same direction for the penult position (β = 0.11, p < .01). Overall, these comparisons showed that unigram and informativity generally predicted the duration of S1 better.

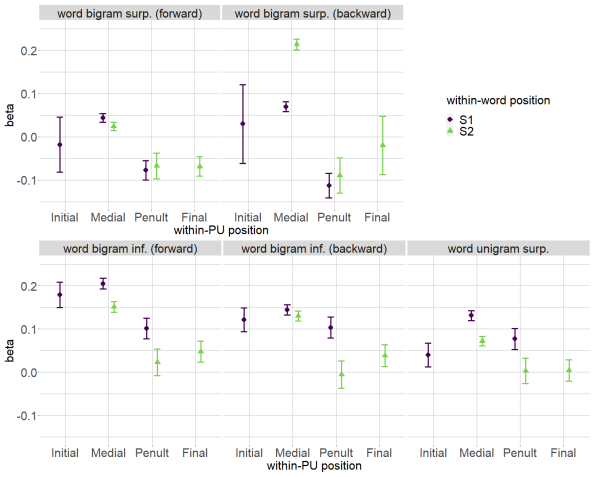


Figure 1: Effect sizes of word-level bigram surprisal (upper panels), informativity, and unigram surprisal (lower panels) at syllable positions within an PU (X axis) and word (color/shape).

Syllable-level bigram surprisal also exhibited a strong directionality effect in terms of its effect size. As shown in the upper panels in Figure 2, at medial and penultimate positions, forward bigram surprisal is a stronger predictor of duration for S2 (medial: p < .0001), whereas backward bigram surprisal was a stronger predictor for S1 (medial: p <

.0001; penult: $p < .001$). In other words, contextual probability conditioned on a syllable within a word had a stronger effect in predicting syllable duration.

On the other hand, as shown in the lower panels in Figure 2, the effect of syllable-level unigram and informativity was either stronger at S1 (backward surprisal at penult: $p < .01$; medial: $p < .0001$) or did not differ between S1 and S2. These two measurements were also a positive predictor of syllable duration at more within-word and within-PU positions (positive estimates with at least $p < .05$ for all except for backward informativity for S2 at the penultimate position) than syllable-level bigram surprisal.

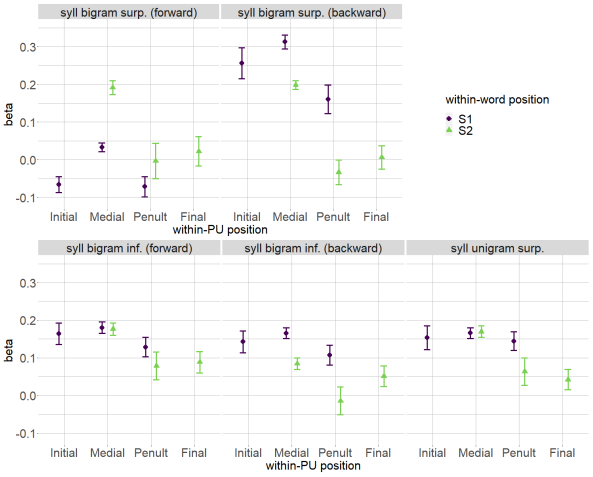


Figure 2: Effect sizes of syllable-level bigram surprisal (upper panels), informativity, and unigram surprisal (lower panels) at syllable positions within an PU (X axis) and word (color/shape).

4. DISCUSSION & CONCLUSION

The results show that having both word-level and syllable-level predictability variables significantly improved the modeling of syllable duration. If the correlation between predictability and durational variability reflects the planning of speech production, this finding suggests that such planning is likely made at both levels of representation. The effect of word-level predictability is especially interesting since it suggests that the phonetic variability at a lower-level unit (syllable) is also affected by word-level statistics.

We further inspected the percolation effect of word-level predictability by examining whether, within a disyllabic word, word-level predictability effects applied uniformly to both syllables. The results show a split between bigram surprisal on one hand and bigram informativity and unigram surprisal on the other. For bigram surprisal, the expected directional effect was found: forward surprisal predicted the duration of the initial syllable better, while backward surprisal predicted the duration of the second syllable better. In other words, word-level bigram surprisal mostly had local effects.

However, word-level informativity and unigram surprisal, which are word-specific measurements that do not vary across a word's occurrences in different local contexts, were more predictive of syllable duration in the initial syllable. It suggests that phonetic variability of a word's first syllable might have been used to signal predictability at the level of word types, i.e., whether the word type is frequent or informative (i.e., often occurring in more predictable contexts).

In addition, we found that syllable-level bigram surprisal is more predictive of syllable duration when conditioned within a word. In other words, even if we focus on how syllable duration and syllable-level predictability correlate, word boundaries are still making an impact, suggesting the crucial role of word-level information (and in terms of data processing for a language like Mandarin Chinese, the role of word segmentation) in the modeling of syllable-level phonetic variability.

Finally, predictability effects (especially for surprisal) were weakened or neutralized at PU boundaries: At the final position, where pre-boundary lengthening was the strongest, and at the initial position, where pitch reset has been found [3]. In other words, there was likely a trade-off between the presence of boundary cues and predictability effects. This finding is consistent with the view that the prosodic structure modulates the relationship between predictability and phonetic variability [1, 20, 11].

To conclude, this study presents a clearer picture of how predictability measurements at different levels affect phonetic variability under different conditions. The results suggest that speakers may be simultaneously tracking two levels of statistical information in speech production. We have further shown that local predictability measurements like surprisal exhibit local effects and are more prone to neutralization in their interaction with prosodic positions, which differs from the behavior of global (i.e., word-type level) measurements such as lexical frequency and informativity.

5. REFERENCES

[1] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic

prominence, and duration in spontaneous speech," *Language and speech*, vol. 47, no. 1, pp. 31–56, 2004.
[2] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational english," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.
[3] A. C.-H. Chen and S.-C. Tseng, "Prosodic encoding in Mandarin spontaneous speech: Evidence for clause-based advanced planning in language production," *Journal of Phonetics*, vol. 76, p. 100912, 2019.
[4] K.-J. Chen, C.-R. Huang, L.-P. Chang, and H.-L. Hsu, "Sinica corpus: Design methodology for balanced corpora," in *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, 1996, pp. 167–176.
[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.
[6] Fon, J, K. Johnson, and S. Chen, "Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech," *Language and speech*, vol. 54, no. 1, pp. 5–32, 2011.
[7] Fon, Y.-J. J, "A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech," Ph.D. dissertation, The Ohio State University, 2002.
[8] C.-R. Huang and K.-j. Chen, "Academia sinica balanced corpus of modern chinese 4.0," *Academia Sinica*, 2010.
[9] R. V. Lenth, "Least-squares means: The R package lsmeans," *Journal of Statistical Software*, vol. 69, no. 1, pp. 1–33, 2016.
[10] Y.-F. Liu and S.-C. Tseng, "Word use and word-level reduction in story-telling speech of chinese-speaking hearing and hard of hearing children," in *(Dis)fluencies in children's speech*, B. Judit, Ed. Akadémiai Kiadó, 2020.
[11] Z. Malisz, E. Brandt, B. Möbius, Y. M. Oh, and B. Andreeva, "Dimensions of segmental variability: Interaction of prosody and surprisal in six languages," *Frontiers in Communication*, vol. 3, p. 25, 2018.
[12] M. Pluymaekers, M. Ernestus, and R. Baayen, "Articulatory planning is continuous and sensitive to informational redundancy," *Phonetica*, vol. 62, no. 2-4, pp. 146–159, 2005.
[13] U. C. Priva, "Sign and signal: Deriving linguistic generalizations from information utility," Ph.D. dissertation, 2012.
[14] ——, "Informativity affects consonant duration and deletion rates," *Laboratory Phonology*, vol. 6, no. 2, pp. 243–278, 2015.
[15] S. Seyfarth, "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," *Cognition*, vol. 133, no. 1, pp. 140–155, 2014.
[16] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.
[17] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proceedings of IEEE automatic speech recognition and understanding workshop*, vol. 5, 2011.
[18] K. Tang and R. Bennett, "Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan)," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 997–1017, 2018.
[19] S.-C. Tseng, "ILAS Chinese spoken language resources," in *Proc. LPSS*, 2019, pp. 13–20.
[20] A. Turk, "Does prosodic constituency signal relative predictability? a smooth signal redundancy hypothesis," *Laboratory phonology*, vol. 1, no. 2, pp. 227–262, 2010.
[21] R. J. Van Son and J. P. Van Santen, "Duration and spectral balance of intervocalic consonants: A case for efficient communication," *Speech Communication*, vol. 47, no. 1-2, pp. 100–123, 2005.
[22] S.-F. Wang, "The interaction between predictability and pre-boundary lengthening on syllable duration in taiwan southern min," *Phonetica*, vol. 79, no. 4, pp. 315–352, 2022. [Online]. Available: https://doi.org/10.1515/phon-2022-0009
[23] L. H. Wurm and S. A. Fisicaro, "What residualizing predictors in regression analyses does (and what it does not do)," *Journal of memory and language*, vol. 72, pp. 37–48, 2014.

BOUNDARY-DRIVEN ACCOUNT FOR DOWNSTEP IN JAPANESE

Kei Furukawa[1] & Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology, Japan
furukawa.kei.fi4@is.naist.jp, s-nakamura@is.naist.jp

ABSTRACT

This study verifies the boundary-driven account for downstep in Japanese: downstep is triggered by phonological boundaries. The assumption widely adopted in the literature is that downstep is triggered only by accents. However, it remains unknown which account is more accurate. We propose that downstep is one example of the phonetic realization of boundary-driven downstep. We define boundary-driven downstep as a phonological mechanism that lowers the pitch of subsequent PPhrases when a PPhrase or a PClause directly dominates two or more PPhrases. Nine native speakers of Tokyo Japanese participated in a production experiment, which paradigmatically compared the F0 heights of a sequence of a final-accented word preceding an unaccented word to a sequence of unaccented words. The results showed that the condition with final accents without particles did not show a large F0 downtrend. Our finding indicates that accents do not directly trigger downstep.

Keywords: downstep, Japanese, prosodic phrasing, boundary-driven downstep, PPhrase

1. INTRODUCTION

1.1. Downstep in Japanese

Japanese is one of the most widely studied languages concerning intonation phonology, and downstep has been actively researched since the late 1960s [1–9]. In Tokyo Japanese, words are classified as either accented (A) or unaccented (U). Traditionally, downstep has been defined as a pitch range compression triggered by lexical pitch accents [2–5, 7–9]. Most theories of Japanese intonation assume two distinct prosodic categories, Minor Phrase (MiP) and Major Phrase (MaP), both of which can theoretically be unified into recursive phonological domains called PPhrases [10–12]. MiP is defined by accent culminativity and initial lowering, and MaP is defined as a domain of downstep [10]. In these two works [11, 12], the domains of initial lowering and downstep are considered recursive PPhrases.

Two primary approaches help identify downstep in Japanese: syntagmatic and paradigmatic [7]. In the former approach, if the F0 peak following an accented word is clearly lower than the preceding F0 peak, it is diagnosed as downstep [6]. On the contrary, the latter paradigmatic approach claims that if the pitch height of X is significantly lower after an accented word (A) than after an unaccented word (U), X is diagnosed as having been downstepped [5, 7, 12]. This approach has been adopted in many studies [2–5, 7–9, 12]. Both diagnoses assume that downstep in Japanese is triggered only by accents.

1.2. Accent-driven vs. boundary-driven accounts

In contrast to previous studies, perhaps downstep is triggered by phonological boundaries, not lexical pitch accents [13]. One work argued that a phonological boundary must be inserted after every accent owing to accent culminativity and anti-lapse constraints [12]. Another study [13] reported that parallel structures trigger the insertion of boundaries, which prompt a step-like F0 downtrend that resembles downstep, even without accents (gray arrows in Figure 1). Conversely, since accents lead to the insertion of phonological boundaries, downstep's direct cause may not be the accent but phonological boundaries [13].

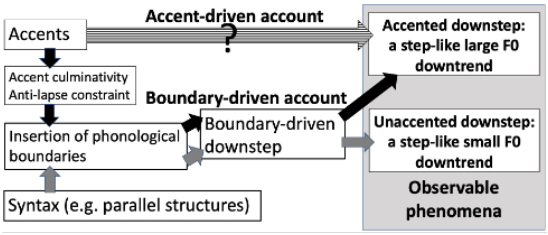


Figure 1: Accent-driven vs. boundary-driven accounts for downstep in Japanese.

In this paper, the accent-driven account argues that downstep is caused directly by accents, as shown in the horizontal striped arrow in Figure 1. In contrast, the boundary-driven account theorizes that downstep is caused by the insertion of phonological boundaries (black arrows in Figure 1).

1.3. Research objectives

This study primarily aims to verify the boundary-driven account for downstep in Japanese: downstep is caused by boundaries rather than directly by accents. To clarify the discussion, this paper redefines the terminology concerning downstep, which is a pitch range compression triggered only by lexical pitch accents, as diagnosed by the paradigmatic approach. Accented downstep is a step-like large F0 downtrend after an accented word. The term merely refers to observable phenomena without presupposing that accented downstep is triggered directly by accents. Unaccented downstep is a step-like small F0 downtrend after an unaccented word [13].

We propose that the large F0 step-like downtrend, which has traditionally been called downstep, is one example of the phonetic realization of boundary-driven downstep. We define boundary-driven downstep as a phonological mechanism that lowers the pitch of subsequent PPhrases when a PPhrase or a PClause directly dominates two or more PPhrases. Boundary-driven downstep phonetically shows a step-like F0 downtrend, unless neutralized by other pitch-rising phenomena. It can phonetically be realized as either accented or unaccented downstep.

2. EXPERIMENT

2.1. Experimental materials

Our experiment paradigmatically compared the sequence of a final-accented word preceding unaccented word to a sequence of unaccented words. Words with an accent on the final syllable are called final-accented or *odaka*-accented words. The difference between final-accented and unaccented words is clearer when a particle follows them. For instance, the final-accented word *hana* (LH*) "flower" and the unaccented word *hana* (LH-) "nose" in Tokyo Japanese do not show an acute pitch fall when pronounced in isolation. However, when followed by a nominative case marker -*ga*, the final-accented *hana-ga* shows the LH*L pattern. In contrast, an unaccented *hana-ga* shows no acute pitch fall, following LHH. Final-accented words in isolation and unaccented words cannot be distinguished in production [14–16] or in perception [17].

In the stimuli, three nouns (N1, N2, and N3) are given in a parallel structure with a conjunction *ya* or middle dots. Region 1 is defined as the area where N1 and the conjunction *ya* or middle dots are combined. The area between N2 and the following conjunction *ya* or the middle dots is defined as Region 2. N1 is final-accented or unaccented; N2 and N3 are unaccented words. One stimuli set is given in Table 1. The stimuli were constructed with two factors: Accent and Particle. The Accent factor is comprised of two levels: [-accent] and [+accent]. At the [-accent] level, N1 is unaccented. At the [+accent] level, in contrast, N1 is final-accented. The second factor is the Particle factor, which has two levels: [-particle] and [+particle]. At the [-particle] level, N1, N2, and N3 form a parallel structure with middle dots. In Japanese, middle dots are not pronounced and are realized as silences or short pauses. In the [+particle] level, N1, N2, and N3 form a parallel structure with a conjunction *ya*. This study assumes that a parallel structure triggers the insertion of phonological boundaries after *ya* or the middle dots [13]. In the [+accent, +particle] condition, a final-accented N1 is followed by *ya*, so the last and the second to last moras are expected to show a H*L pattern. In the [+accent, -particle] condition, on the other hand, the final mora of Region 1 remains high, and no sharp pitch fall is expected to be observed.

Table 1: Sample stimuli used in experiment: Accented moras are underlined.

a. [-accent][-particle]						
item	hana,	mori,	ue	to-itta	kanji-ga	kaitearu
gloss	nose,	forest,	top	such as	chinese character-NOM	written
	'The Chinese characters such as a nose, forests, and top are written there.'					
b. [+accent][-particle]						
item	hana,	mori,	ue	to-itta	kanji-ga	kaitearu
gloss	flower,	forest,	top	such as	chinese character-NOM	written
	'The Chinese characters such as flowers, forests, and top are written there.'					
c. [-accent][+particle]						
item	hana-ya	mori-ya	ue	to-itta	kanji-ga	kaitearu
gloss	nose-and	forest-and	top	such as	chinese character-NOM	written
	'The Chinese characters such as a nose, forests, and top are written there.'					
d. [+accent][+particle]						
item	hana-ya	mori-ya	ue	to-itta	kanji-ga	kaitearu
gloss	flower-and	forest-and	top	such as	chinese character-NOM	written
	'The Chinese characters such as flowers, forests, and top are written there.'					

2.2. Participants

Nine native speakers (four females and five males, mean age 19.6 years, SD 0.96) of Tokyo Japanese from the Kanto area (Tokyo, Kanagawa, Saitama, and Chiba) participated in our experiment as subjects. No participant ever lived outside of Kanto area for more than two years. None reported a history of speech or hearing impairments.

2.3. Procedures and analysis

The recording was conducted in a soundproof booth at the University of Tokyo using a Shure

WH20XLR Dynamic Headset Microphone, which was connected to a Roland QUAD-CAPTURE audio interface; the audio was recorded to a computer at a sampling rate of 44.1 kHz. The stimuli were displayed on a screen one by one in a pseudo-random order. Participants read sentences aloud three times at a normal speech rate that felt natural to them. When they inserted an undesired pause or a mispronunciation while reading a sentence, they repeated it. We recorded a total of 6 items × 4 sentence types × 3 repetitions = 72 tokens. 120 sentences (360 tokens) from other experiments served as fillers. Sound files were annotated using Praat [18] and a script called ProsodyPro [19]. Segmentation between the conjunction and the following unaccented words was based on formants and waveforms. Two of the three repetitions were analyzed. Apparent errors by the algorithm in Praat, such as octave jumps, were manually checked and corrected.

For each utterance, we made the following three measurement variables. The first measurement is R1Fall: the maximum value of the pitch in N1 minus the minimum value of the pitch of the following conjunction *ya*. This measurement is only calculated for the [+particle] level. It is then converted to semitones. The second measurement is R1MaxF0: the normalized F0 maximum in Region 1. To factor out the pitch range differences among speakers, we converted the values of F0 maximum x to normalized values y with reference to two points, using the following formula (1). R_1 was the mean value of the F0 maximum in Regions 1 and 2 across all the data points for the speaker, and R_2 is the mean value of the F0 minimum in Regions 1 and 2 across all the data points for the speaker. This normalization was previously used [7]:

$$(1) \quad y = \frac{x - R_2}{R_1 - R_2}$$

The third variable is R2MaxF0, which is the normalized F0 maximum in Region 2. This variable is also normalized using the formula (1). In the sequence of *hana-ya mori-ya* "flowers and forests" in the [+accent, +particle] condition, for instance, R1MaxF0 is calculated as the normalized maximum value of the pitch in the sequence *hana-ya* "flowers and", while R2MaxF0 is the maximum value of the pitch in the sequence of *mori-ya* "forests and." As for the [+accent, -particle] condition, in the sequence of *hana*, *mori* "flowers, forests," R1MaxF0 is the normalized maximum value of the pitch in the sequence *hana* "flowers", while R2MaxF0 refers to the maximum value of the pitch in the sequence *mori* "forests."

The data were analysed within the linear mixed-effects model (LME) using the lmer function within the lme4 package [20] in R [21], where the subjects and items are random effects. The factor labels of Accent and Particle were cantered to have a mean of 0 and a range of 1. The final models were obtained using backward selection [22].

3. RESULTS AND DISCUSSION

3.1. Results

Our results show two noteworthy findings. First, contrary to the prediction from the accent-driven account, the [+accent/-particle] condition, which has final accents without particles, did not show a large F0 downtrend from N1 to N2 (Figure 2). Second, the small step-like downtrend found in the [+accent/-particle] level is slightly larger than the downtrend in the [-accent/-particle] level.

The results for each variable are presented in Table 2 (for R1Fall), Table 3 (for R1MaxF0), and Table 4 (for R2MaxF0). Figure 2 displays the sample F0 contours of all conditions.

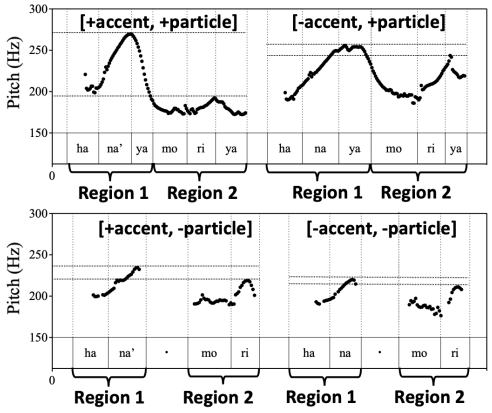


Figure 2: Sample F0 contours of all conditions.

For R1Fall, the main effect of the Accent factor was significant, where the [+accent] level showed a greater F0 fall than the [-accent] level. This suggests that the final accent is realized as a sharp F0 fall in the [+accent, +particle] condition.

Table 2: Results of mixed-effects models for R1Fall.

Predictor	β	t	p
(Intercept)	3.763	7.993	<.001
Accent	4.372	26.260	<.001

Regarding R1MaxF0, the difference between the [-accent, -particle] condition and the [+accent, -particle] condition was not statistically significant. Meanwhile, the [+accent, +particle] condition had a significantly higher value compared to the [-accent,

+particle] condition. These results suggest that pre-low raising causes accentual H* tones to have a higher pitch than phrasal H- tones [23], since a higher F0 peak at Region 1 was only observed when an accent was realized as a sharp F0 fall at the [+particle] level. Moreover, the Particle factor has significant effect on the R1MaxF0 in both [-accent] and [+accent] levels.

Table 3: Results of mixed-effects models for R1MaxF0.

Condition	Predictor	β	t	p
[-accent, -particle]	(Intercept)	0.858	37.17	<.001
vs. [+accent, -particle]	Accent	0.045	0.90	.372
[-accent, +particle]	(Intercept)	1.083	39.895	<.001
vs. [+accent, +particle]	Accent	0.274	8.547	<.001
[-accent, -particle]	(Intercept)	0.891	39.891	<.001
vs. [-accent, +particle]	Particle	0.110	3.734	<.001
[+accent, -particle]	(Intercept)	1.050	28.992	<.001
vs. [+accent, +particle]	Particle	0.339	6.884	<.001

The R2MaxF0 for the [+accent, -particle] condition was significantly smaller than that for the [-accent, -particle] condition. This does not satisfy the requirement for a paradigmatic diagnosis because the F0 difference between the two conditions is quite small, as illustrated in Figure 2. The R2MaxF0 for the [+accent, +particle] condition was also significantly smaller than that for the [-accent, +particle] condition. The [+accent, +particle] condition is diagnosed as under downstep because a large F0 difference satisfies the requirement for paradigmatic diagnosis. The R2MaxF0 for the [-accent, -particle] condition was significantly smaller than that for the [-accent, +particle] condition. Additionally, the R2MaxF0 for the [+accent, +particle] condition was significantly smaller than that for the [+accent, -particle] condition. The observation that the F0 peak of Region 2 for the [+accent, +particle] condition is considerably smaller than that of the [+accent, -particle] condition suggests that the F0 downtrend in the [+accent, -particle] condition should be differentiated from the paradigmatically diagnosed downstep found in the [+accent, +particle] condition.

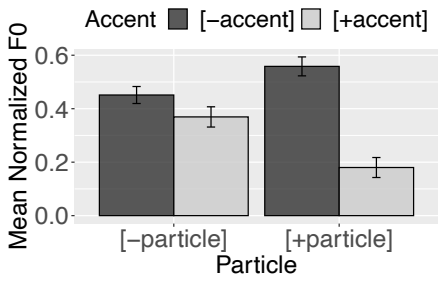


Figure 3: Normalized F0 peak of Region 2: Error bars represent 95% confidence intervals.

Table 4: Results of mixed-effects models for R2MaxF0.

Condition	Predictor	β	t	p
[-accent, -particle]	(Intercept)	0.410	11.497	<.001
vs. [+accent, -particle]	Accent	-0.081	-3.947	<.001
[-accent, +particle]	(Intercept)	0.369	12.919	<.001
vs. [+accent, +particle]	Accent	-0.378	-8.199	<.001
[-accent, -particle]	(Intercept)	0.504	18.758	<.001
vs. [-accent, +particle]	Particle	0.108	4.998	<.001
[+accent, -particle]	(Intercept)	0.275	7.570	<.001
vs. [+accent, +particle]	Particle	-0.189	-4.399	<.01

3.2. Discussion

Our results support the boundary-driven account. Accents do not directly trigger a large F0 fall, which is traditionally referred to as downstep. Two facts indicate that there is no downstep in the [+accent, -particle] condition: 1) the [+accent, -particle] condition did not exhibit a large F0 compression for meeting the requirements for a paradigmatic diagnosis; 2) the [+accent, -particle] condition had a sufficiently higher F0 peak at Region 2 compared to the [+accent, +particle] condition. One might argue that the phonological accents in the [+accent, -particle] condition are deleted rather than unrealized. However, the significant difference between the [-accent, -particle] and [+accent, -particle] conditions refutes the possibility because the Accent factor did create a slightly lower F0 peak at Region 2 in the [+accent, -particle] condition than in the [-accent, -particle] condition.

We believe that boundaries are inserted between N1 and N2 due to the Lapse-L constraint or the parallel structure. These boundaries trigger boundary-driven downstep, which is phonetically realized as accented downstep in the [+accent, +particle] condition and unaccented downstep at the [-accent] level. The slightly larger downtrend than the unaccented downstep at the [+accent, -particle] condition may be due to the unassociated tones of L from the final accent H*L.

A question remains concerning the differences of the F0 fall between AX and UX by a paradigmatic approach [2–5, 7–9, 12]. This study suggests that pre-low raising, phonological phrasing, and spillover effect of accents could be confounding factors, but additional research is necessary to determine their exact role.

4. CONCLUSION

Our experiment results reveal a step-like F0 downtrend that is smaller than downstep, despite the presence of final accents. Thus, our study contends that accented downstep in Japanese is caused by boundaries rather than directly by accents.

5. ACKNOWLEDGEMENTS

Part of this work was supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Number JP21H05054.

6. REFERENCES

[1] J. D. Mccawley, *The phonological component of a grammar of Japanese*. Mouton, 1968.
[2] W. J. Poser, *The phonetics and phonology of tone and intonation in Japanese (Doctoral dissertation)*, 1984.
[3] J. B. Pierrehumbert and M. E. Beckman, *Japanese Tone Structure. Number 15 in Linguistic Inquiry Monographs*, 1988.
[4] H. Kubozono, *The Organization of Japanese Prosody*, 1988.
[5] ——, "Syntactic and rhythmic effects on downstep in japanese," *Phonology*, vol. 6, no. 1, pp. 39–67, May 1989.
[6] E. Selkirk and K. Tateishi, "Syntax and downstep in japanese," in *Interdisciplinary Approaches to Language*. Dordrecht: Springer Netherlands, 1991, pp. 519–543.
[7] S. Ishihara, "Japanese downstep revisited," *Nat. Lang. Linguist. Theory*, vol. 34, no. 4, pp. 1389–1443, Nov. 2016.
[8] M. Hirayama and H. K. Hwang, "Relative clause and downstep in japanese," *Proceedings of the Annual Meetings on Phonology*, vol. 6, no. 0, Jun. 2019.
[9] M. Hirayama, H. K. Hwang, and T. Kato, "Lexical category and downstep in japanese," *Languages*, vol. 7, no. 1, p. 25, Jan. 2022.
[10] S. Ishihara, "14 syntax-phonology interface," handbook of japanese phonetics and phonology," pp. 569–618, 2015.
[11] J. Ito and A. Mester, "Recursive prosodic phrasing in japanese," prosody matters: Essays in honor of elisabeth selkirk," pp. 280–303, 2012.
[12] ——, "Prosodic subcategories in japanese," *Lingua*, vol. 124, pp. 20–40, Jan. 2013.
[13] K. Furukawa and Y. Hirose, "Boundary-driven downstep in japanese," *Proceedings of 19th ICPhS, Canberra: ASSTA*, 2019.
[14] M. Sugito, "Comparison between "high-final" & "level" tone in tokyo accent [a study of final-accented and unaccented disyllabic words in tokyo using dynamic measurement," *Onseigakkai Kaihoo*, vol. 129, pp. 1–4, 1968.
[15] T. J. Vance, "Final accent vs. no accent: Utterance-final neutralization in tokyo japanese," *Journal of Phonetics*, vol. 23, no. 4, pp. 487–499, 1995.
[16] Y. Sugiyama, *Japanese pitch accent: Examination of final-accented and unaccented minimal pairs. Toronto Working Papers in Linguistics*, 2006.
[17] N. Warner, "Japanese final-accented and unaccented phrases," *J. Phon.*, vol. 25, no. 1, pp. 43–60, Jan. 1997.
[18] P. Boersma and D. Weenink, *Praat: A System for Doing Phonetics by Computer" Version 6.0. 39*, 2018.
[19] Y. Xu, "ProsodyPro — a tool for large-scale systematic prosody analysis," in *Tools and Resources for the Analysis of Speech Prosody*. Aix-en-Provence, France: Laboratoire Parole et Langage, France, 2013, pp. 7–10.
[20] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, G. Grothendieck, and P. Green, "Package "lme4," 2015.
[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/
[22] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," Jun. 2015.
[23] A. Lee, S. Prom-On, and Y. Xu, "Pre-Low raising in japanese pitch accent," *Phonetica*, vol. 74, no. 4, pp. 231–246, Mar. 2017.

THE INTERACTION OF INTONATION BOUNDARY AND LEXICAL TONES UNDER TWO FOCUS CONDITIONS IN THE WUZHI DIALECT OF CHINESE

Yuzhu Zhu[1,3], Zhiqiang Li[2], Aijun Li[3,4]
[1]Beijing International Studies University (BISU), China
[2] University of San Francisco, United States
[3]Institute of Linguistics, Chinese Academy of Social Sciences, China
[4]University of Chinese Academy of Social Sciences (UCASS), China
zhuyuzhu@bisu.edu.cn, zqli@usfca.edu, liaj@cass.org.cn

ABSTRACT

This study investigates the mechanism of boundary marking in tone languages through the interaction of intonation boundary and lexical tones under two focus conditions in Wuzhi. We analyze the prosodic realizations of the lexical tones at the right edge of string-identical statements and echo questions, both spoken with a broad or narrow focus that coincides with the nuclear stress on the final prosodic word. Results show that the final f_0 lowering is a robust indicator of statements, resulting in a falling contour in most tonal contexts. Questions display more intricate patterns, in which the H% target drives the surface tonal patterns in the two convex and two falling tones at the intonation boundary. Furthermore, the intonation boundary is signaled by longer duration and higher pitch register in questions than in statements. We conclude with a typological account of boundary marking for the simultaneous realization of tone and intonation in Chinese.

Keywords: intonation boundary, lexical tones, broad and narrow focus, Wuzhi

1. INTRODUCTION

In tone languages, tone and intonation are both expressed by the acoustic property of fundamental frequency (f_0) and its perceptual correlate of pitch. An essential aspect of studying intonation of Chinese boils down to how pitch contours that are associated with lexical tones are manipulated both locally and globally in response to different intonational meanings and focus conditions (See [17] for an overview). Of particular interest in this line of research is how lexical tones in the utterance-final prosodic word interact with the intonation boundary that serves to distinguish utterance types and emotions or affective states of utterances.

In one of the seminal works of tone and intonation in Mandarin Chinese, Chao [3] identified two modes of interaction in which lexical tones and intonational tones can be realized either simultaneously or successively. While the simultaneous addition may occur in any part of the utterance, the successive addition is often observed at the intonation boundary. The two modes of interaction shape terminal pitch contours differently at the boundary. In Mandarin, for example, the boundary tones associated with declarative and interrogative utterances modify the register or slope of the pitch contours at the right edge of the intonation phrase such that they do not alter the canonical shapes of the lexical contours [9, 10, 11]. In other words, the falling tone – Tone 4 in Mandarin – maintains its falling contour in questions, probably with a smaller pitch range, while the rising tone – Tone 2 in Mandarin – also retains its rising contour in questions, but its pitch register is likely elevated and expanded. The successive addition is found in the question intonation of Cantonese when a final rise is added to the pitch contour of the final tone in the utterance [6, 18], or in utterances that convey strong emotions such as disgust or anger in Mandarin [8].

In an attempt to understand how tone and intonation interact at the intonation boundary from a typological perspective, we ask the general question of how the two modes of interaction outlined above play out in different tonal systems other than the relatively well-studied ones in Mandarin or Cantonese. The present study compares the prosodic realizations of lexical tones at the intonation boundary in statements and echo questions (without question particles) under two focus conditions in a Jin dialect. We seek to address the following questions:

1. Are there any differences in the realization of lexical pitch contours at the intonation boundary in string-identical statements and echo questions?
2. How is the realization of pre-boundary lexical contours shaped by the introduction of narrow focus in comparison with broad focus?
3. How do different lexical tone systems respond to boundary marking differently?

Wuzhi is a Jin dialect of over 700,000 speakers in the north-west of Henan province, which borders the vast Mandarin speaking region. It is reported to have five lexical tones: two falling tones, two convex tones, and a checked tone [19]. Jin dialects, unlike Mandarin dialects, retain the checked tones in their lexical tone system, and Wuzhi serves as a typical example. Experimental work on their intonation has been lacking. In this study, we start with the lexical tone

system and the tone sandhi patterns before tackling intonational features, as Wu [11] suggested in his study of intonation in Mandarin. It is important to note that tonal alternations trigged by a phonological process is not intonational. A well-known tone sandhi is the Tone 3 sandhi in Mandarin: when two syllables in Tone 3 are juxtaposed, the first syllable changes to Tone 2 [3]. In this paper we focus on data and results of the intonation study, but will draw on findings from the tonal study too, which will be dealt with in more detail in [20].

2. METHODS

2.1. Materials and speakers

The recording materials consist of three parts: 64 monosyllables, 837 disyllables and 25 sentences. The word lists were constructed based on the historical origins of tonal categories in the Middle Chinese and tonal combinations. 25 disyllabic combinations out of five lexical tones were embedded into a carrier sentence "张军初三飞___" (Zhangjun flied to ___ on the third day of the month), in which "飞" is the verb "fly to". All 25 combinations are city or country names that occur in the sentence-final object position and receive the nuclear stress when the sentence has a broad, neutral focus or when a narrow focus is placed on the object noun [5]. Three mini-dialogues were created similarly for each of the 25 target sentences, as follows:

1. A: 你刚刚说什么？ (What did you just say?)
   B: 张军初三飞英国。(Zhangjun on the third day of the month flied to England.)
2. A: 你刚刚问什么？ (What did you just ask?)
   B: 张军初三飞英国? (Zhangjun on the third day of the month flied to England?)
3. A: 张军初三飞【英国】? (Zhangjun on the third day of the month flied to **England**?)
   B: 张军初三飞【法国】。(Zhangjun on the third day of the month flied to **France**.)

In Dialogue 1, the response to the question receives a broad focus. In Dialogue 2, B repeats the question in response to A. The response in B has the form of an echo question (without question particles) that has the pragmatic effect of expressing surprise, requesting confirmation, or making an indirection negation [14]. An echo question in Chinese have the same form as a statement. In Dialogue 3, the object noun receives a narrow focus in A and B when the focus of the question is on the location noun.

12 speakers of the Wuzhi dialect, six males and females, participated in the experiment. They reported no hearing or speaking difficulties. Their average age is 52 (SD = 15).

2.2. Recording and data extraction

The recording was carried out in the quiet bedrooms at the participants' homes. An in-house recording program X-recorder [13] was used to collect data on a PC laptop computer with a "Sound Blaster Play! 3" external sound card and a Philips SHM 1900 headset. The sampling rate and resolution were set to be 22.05 KHz and 16 bits respectively.

For the monosyllables, 11 speakers read 64 syllables three times, generating 2112 recordings (64×3×11). For the disyllables, 11 speakers read 837 syllables once for a total of 9207 recordings (837×11). 12 speakers were asked to read the three mini-dialogues for each target sentence, producing four utterance types: a statement and question in broad focus (1B and 2B), and a statement and question in narrow focus (3B and 3A). We ended up with 1200 utterances (12 speakers, 25 target sentences, four utterances for each target sentence). One speaker's data were excluded from the analysis because of their limited ability to read Chinese characters.

A Praat-based [1] automatic segmentation and transcription tool was employed to provide segmental annotations for syllabic and subsyllabic elements (initials and finals) [12]. Pitch contour was extracted for each syllable, with spurious pitch cycles manually modified. 11 pitch values were extracted at equal intervals on each syllable final. Durations of syllable initials and finals were also obtained. An example of segmentation and pitch extraction is shown in Fig 1.

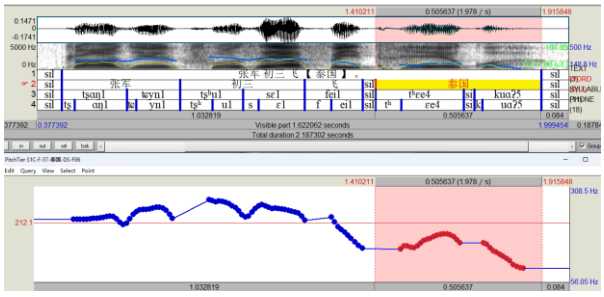


Figure 1: Segmentation, transcription and pitch tracks.

3. RESULTS

3.1. Lexical tone system in Wuzhi

Table 1 provides a summary of tone values on a five-point scale [2] and their corresponding tonal features. Tones 1 and 4 are convex tones that have three that tonal targets MHM or LHL, differentiated by pitch registers. Meanwhile, Tones 2 and 3 are falling tones in low and high registers, respectively. Lastly, Tone 5 is a mid-level checked tone.

Tones	T1	T2	T3	T4	T5

shapes	/343/	/31/	/51/	/242/	/3/
features	MHM	ML	HL	LHL	M

Table 1: Lexical tones and features in Wuzhi.

Fig. 2 illustrates the major variants averaged over 11 speakers. All tones in the citation form exhibit tremendous pitch movement over the course of the duration, even in the checked tone. Comparing the two convex tones, Tone 1 has a relatively longer falling tail while Tone 4 has a much longer rising portion. The two falling tones tend to level off or even go up in pitch toward the end of the syllable in some variants. Most variants conform with tonal shapes in Table 1, especially when they occur in disyllabic combinations. The falling tail is more likely to occur when the tone is in the final position.

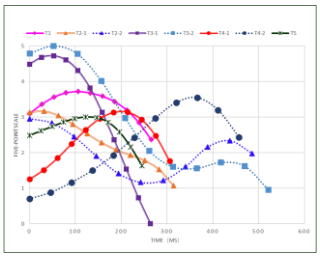


Figure 2: Averaged pitch contours of tones in Wuzhi.

3.2. Disyllabic patterns: Tone 4

The disyllabic patterns are given in Fig. 3. In each of the panels, X is one of the five tones. Since Tone 5 patterns with Tone 1, "X+T5" is omitted. The f_0 plots were generated after averaging the f_0 values at each of the 11 points where measurements were taken.

Tonal contrasts are preserved in the disyllabic patterns. For example, panels labelled "X+T2" and "X+T3" show that the contrast between the Tones 2 and 3 is clearly maintained in the word-final position. Their register distinction is also retained in the word-initial position, as in the panel of "X+T1".

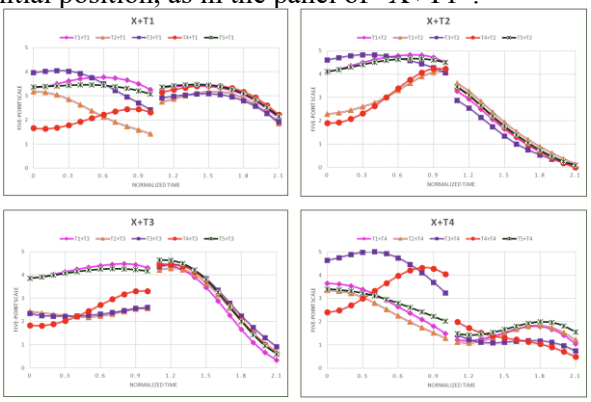


Figure 3: Disyllabic patterns of tones in Wuzhi.

3.3. Effects of intonation boundary on lexical contours

Regarding the effect of intonation boundary on lexical contours, three situations can be identified in statements under two focus conditions. Firstly, Tone 2 and Tone 3 are realized as falling contours before the intonation boundary, regardless of the focus condition. In this context, the contour shape and the register specification are faithfully maintained as in disyllabic combinations. Secondly, Tones 1 and 4 are realized as a low or low fall in the utterance-final position. Both tones are influenced by the ending f_0 of the preceding tone, which suggests the presence of a coarticulatory effect [15]. This means that the tonal target for both tones is most likely L. Thirdly, in Wuzhi, Tone 5 is a checked tone whose duration is generally shorter than long tones. However, its surface contour shape appears to align with that of Tones 1 and 4 in statement intonation under both focus conditions. An example is provided in Fig. 4. The final syllable is in Tone 1 and the penultimate syllable is in one of the five tones labelled as X. In almost all statements produced by the speakers, the final f_0 lowering has been a robust indicator of statement intonation, resulting in a falling contour in most tonal contexts.

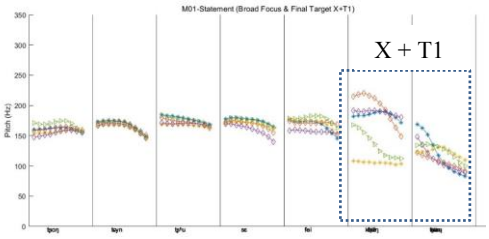


Figure 4: Pitch contours of a statement in broad focus.

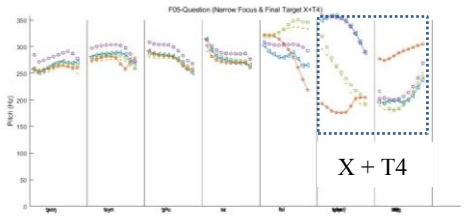


Figure 5: Pitch contours of a question in narrow focus.

Questions display more intricate patterns under the two focus conditions. Tones 2 and 3 are realized with a rising tail toward the end of the syllable, leading to a somewhat concave f_0 configuration in some cases. A very intriguing case is seen in Tones 1, 4 and 5 when the boundary is associated with the question intonation. As Fig. 5 shows, when the final prosodic word was focused, Tone 4 was realized as a rise in the question. The final rise in Tones 1, 4 and 5 appears to be the predominant pattern across speakers.

The pitch register specification participates in the computation of the surface tonal contours, as we see lower rise in Tone 4, but higher rise in Tones 1 and 5.

3.4. Prosodic cues for utterance types and focus conditions

GLMM was conducted on the pre-boundary tones in the two utterance types under two focus conditions in order to determine the prosodic cues that can be used effectively to distinguish statements and questions, and different focus types. It was found that the intonation boundary was signaled by longer duration (Fig. 6a, $p<0.05$) and higher pitch register (Fig. 6b-d, $p<0.05$) in questions than statements. The maxima, minima and average values of f_0 in the final syllable are all significantly higher in questions than statements ($p<0.05$ for all). The pitch range of the final syllable is significantly smaller in questions than statements (Fig. 6e).

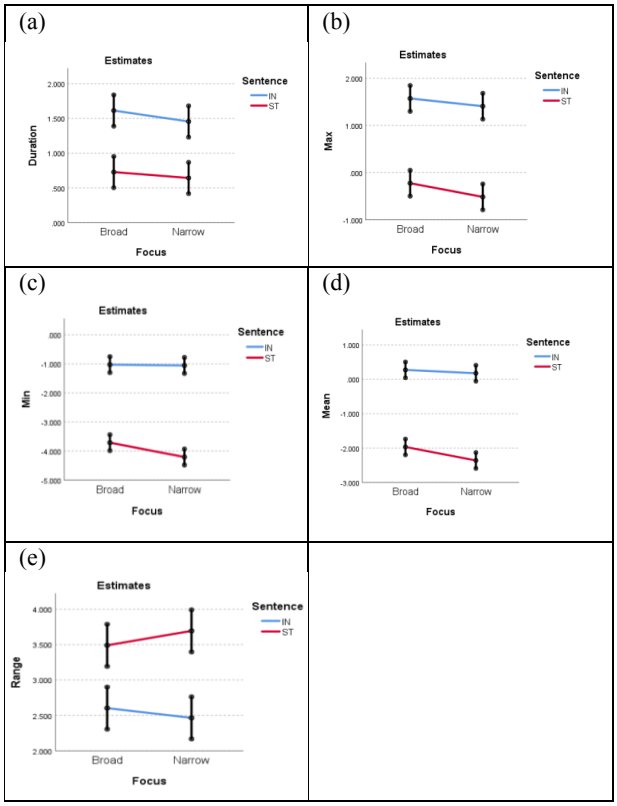


Figure 6. GLMM analysis results for five prosodic cues across focus conditions and utterance types.

The difference in duration under the two focus conditions is not significant (Fig. 6a, $p=0.558$). However, we notice that the domain of the narrow focus is the prosodic word consisting of the final and penultimate syllables, not just the final syllable.

In statements, the pitch range of the final syllable was found to be larger when it was under the narrow focus. However, the opposite seems to be true in questions (Fig. 6e).

Significant interactions between utterance type and focus type were found in f_0 minima ($p=0.027$), and marginal significant interactions were found in average f_0 values ($p=0.056$), but no interactions for other cues ($p>0.05$).

4. DISCUSSION

In tone languages, intonation meanings are expressed by modifying pitch contours associated with lexical tones, especially at the intonation boundary. The boundary marking mechanism in tone languages appears to be subject to both universal and language-specific constraints of intonation. For example, in numerous languages, the question tune ends with a rising pitch movement, represented by the L-H% boundary tone in the AM model of intonation [7]. In languages where word meanings are expressed in tonal contrasts, intonation features typically do not alter tonal identity, as we have seen in Mandarin [9, 10, 11]. Nonetheless, a boundary tone can manifest itself even without a segmental host, such as in the example of Cantonese question intonation without the particle [6, 18].

In Wuzhi, the H% appears to drive the surface tonal patterns at the intonation boundary in the question intonation, and the mechanism here is different from Mandarin and Cantonese. In the case of Tones 1, 4 and to some extent Tone 5, the occurrence of the final rise is only associated with echo questions. The disyllabic tone sandhi data [20] suggest that the convex tone is better represented as R+L. Using Tone 4 as an example, it appears as a rising tone in the first syllable and a low tone in the second. While the H% target in question intonation does not directly manifest itself as it does in Cantonese, it does influence the emergence of the R tone. In the case of the two falling tones, the rising tail toward the end of the syllabic host provides a cue for the H% boundary tone.

The concurrent realization tone and intonation at the intonation boundary can described in terms of simultaneous addition and successive addition [3]. Our anslysis suggests that Wuzhi displays the successive addition pattern in both statements and questions, but the specific mechanism differs from Mandarin and Cantonese.

In conclusion, we have seen that lexical pitch contours at the intonation boundary are modified differently in statements and echo questions, due to the different boundary tones aligned with the boundary. Focus, especially narrow focus, generally modifies pitch contours in the same way as in other tone languages [16].

5. ACKNOWLEDGEMENTS

This work was supported by the Project of Cultural Experts and "Four Batches" of Talents awarded to Aijun Li and the CASS Innovation Project. We would also like to thank Ziyu Xiong, Liang Zhang and Xiaoyan Zhang for their contributions to data collection and comments.

6. REFERENCES

[1] Boersma, P., & Weenink, D. 2017. *Praat: Doing phonetics by computer* (version 6.0.24). Amsterdam: University of Amsterdam.

[2] Chao, Y. R. 1930. A system of 'tone-letters'. *Le Maître Phonétique*, 30, 24–27.

[3] Chao, Y. R. 1933. Tone and Intonation in Chinese. *Bulletin of the Institute of History and Philology Academia Sinica*, 4(2), 121-134.

[4] Chao, Y. R. 1968. *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

[5] Duanmu, S. 2007. *The Phonology of Standard Chinese* (2nd). New York: Oxford University Press.

[6] Ge, C., Li, A. 2019. Intonation of Cantonese interrogative sentences with and without sentence final particle. *Proc. 19th ICPhS* Melbourne, 2455–2459.

[7] Ladd, D. R. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

[8] Li, A., 2013. Successive addition boundary tone of Chinese emotional intonation: production and perception. In: Cheng, C. (ed.), 206-308.

[9] Lin, M. 2006. Interrogative mood and boundary tone in Chinese. *Journal of Chinese Linguistics*. Vol. 4. 364-384.

[10] Lin, M., Li, Z. 2011. Focus and boundary tone in Chinese intonation. *Proc. 17th ICPhS* Hong Kong, 1246–1249.

[11] Wu, Z. 1990. Basic contour patterns of intonation in Mandarin Chinese. In: *A Festschrift for Wang Li*. Beijing: Commercial Press.

[12] Xiong, Z. 2019. xSegmenter: a tool for automatic segmentation and labelling. *Journal of Chinese Phonetics* 11. 27-34.

[13] Xiong, Z. "X-recorder: recording tool for speech corpus," https://mp.weixin.qq.com/s/ewBC2hDg4seuV1YWoPVi-A.

[14] Xu, J. 2016. Linguistic forms of echo questions. In: Eom I., Zhang, W. (eds.), *Language Evolution and Changes in Chinese*. Project on Linguistic Analysis, University of California, 206-215.

[15] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25. 61-83.

[16] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* 27. 55-105.

[17] Xu, Y. 2019. Prosody, tone and intonation. In: Katz, W. F., Assmann, P. F. (eds), *The Routledge handbook of Phonetics*. New York: Routledge, 314–356.

[18] Zhang, L., 2014. Segmentless sentence-final particles in Cantonese: an experimental study. *Studies in Chinese Linguistics*, 35(2), 47–60.

[19] Zhu, Y. 2018. *A Study of Rime Change in the Wuzhi Dialect*. Doctoral dissertation at Tsinghua University.

[20] Zhu, Y., Li, A. Submitted. A phonetic study of disyllabic tones in the Wuzhi dialect.

AN ACOUSTIC STUDY OF V-TO-V COORDINATION WITHIN AND ACROSS WORDS IN FRENCH

Alice Yildiz, Daria D'Alessandro, Sejin Oh, Anne Hermes, Cécile Fougeron

Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/Sorbonne Nouvelle (France)
alice.yildiz@sorbonne-nouvelle.fr, daria.dalessandro@sorbonne-nouvelle.fr, se-jin.oh@cnrs.fr, anne.hermes@sorbonne-nouvelle.fr, cecile.fougeron@sorbonne-nouvelle.fr

ABSTRACT

To better understand coordination beyond the syllable, we investigate inter-syllable coordination in French, depending on the boundary separating the syllables, i.e., a syllable boundary within a word vs. a word boundary in a subject-verb phrase. Patterns of coordination are inferred by acoustical measures of V-to-V anticipatory coarticulation and temporal V-to-V lag. The effect of the intervening consonant (/p/ vs. /ʁ/), as well as the variability in the spectral and lag characteristics across 45 repetitions, is compared across the two boundary types. Coarticulation decreases across-words compared to within-word, but only with the consonant /p/, showing the least articulatory constraint, while token-to-token variability is similar in the two boundary contexts. V-to-V lag is longer across- than within-word. Interestingly, variability in lag duration is found to increase across-words, but only with /p/. Overall, within-word cohesion between syllables translates in a specific pattern of coordination only in the context favoring gestural overlap.

Keywords: anticipatory V-to-V coarticulation, V-to-V duration, inter-syllabic coordination, consonantal constraint, variability

1. INTRODUCTION

Even though a huge amount of work has been done on intergestural coordination, several aspects underlying this process remain poorly understood. One unresolved matter relates to the speech units over which speech gestures are coordinated with each other. Indeed, if a special cohesion is considered to characterize tautosyllabic gestures, more unclear is how speech gestures are organized over units larger than the syllable, especially within a word. In particular, the question is open whether the segments composing a lexical unit present a specific coordination pattern which reflects within-word gestural cohesion; in other terms, whether a lexical word is a unit of gestural organization. Studies that have compared intergestural coordination within and across words by looking at CV coarticulation or overlap in CC sequences reported no effect of word boundary on the degree of overlap [1, 2, 3], or inconsistent results [4, 5, 6]. However, some studies looking at acoustic rhyme lengthening reported longer rhymes at word boundaries than in word-internal position [7, 8]. In a first investigation of anticipatory V-to-V coarticulation across syllables within vs. across words in French, [9] found that four out of five speakers coarticulated more within a word than across words, even in cases where the two words belonged to the same accentual phrase (i.e., they were not separated by a boundary other than a lexical or prosodic word boundary).

In this acoustic study, we tackle further the specificity of word-internal coordination. Thus, we infer coordination between syllables within- and across-words indirectly by spectral cues of anticipatory V-to-V coarticulation, but also by the temporal acoustic lag between the two vowels. Following the assumption that intergestural coordination in a unit where the coordination is planned would result in more stable patterns [10], we test whether spectral cues of coarticulation and temporal lag are more stable within-words, when looking at multiple repetitions of the same tokens by five speakers. Moreover, following the DAC model [11], according to which constraints on the articulation of the consonants should affect V-to-V lingual anticipation, we test our acoustic proxies of coordination in sequences favoring overlap (/VpV/) or disfavoring (/VʁV/) it.

2. METHOD

2.1. Speech material and procedure

Target sequences were /pV$_1$.CV$_2$/ disyllabic items where V$_1$ was /a/, C$_2$ was either /p/ or /ʁ/ and V$_2$ was /i/. The boundary between the two syllables was either a syllable boundary within a word (within-word position, e.g., in *Paris*) or a word boundary between a subject and a verb (across-word position, e.g., in *papa rit beaucoup*, Engl. "Dad laughs a lot"). Target sequences were embedded in meaningful sentences (see Table 1) and recorded in five subsequent sessions. Each session contained the sentences in a pseudo-randomized order, with each target sequence occurring nine times per session, leading to a total of

45 repetitions per speaker. This procedure was followed to collect a large number of repetitions and to test the variability in the production of the target sequences. Five female native French speakers, aged 22 to 28, participated in the recordings.

Sequence	Sentence (example)
/a.pi/ within-word	Quand p**api** l'aura vu, il nous croira. *When Grandpa sees him/her/it, he'll believe us.*
/a#pi/ across-words	Pap**a pi**lote un avion de ligne. *Daddy flies an airliner.*
/a.ʁi/ within-word	Quand P**ari**s est sous la neige, c'est très beau. *When Paris is under the snow, it is very beautiful.*
/a#ʁi/ across-words	Pap**a ri**t beaucoup au cinéma. *Daddy laughs a lot at the movies*

Table 1: Sentences per boundary and consonant types, for the V_2 = /i/ context.

2.2. Measures

A first proxy of coordination was defined as the spectral influence of V_2 /i/ on V_1 /a/ in /api/ and /aʁi/ sequences, with /apa/ and /aʁa/ sequences being used as a control context. Since the anticipation of /i/ on /a/ should translate into a lowering of F1 and an increase in F2, a composite measure of F2-F1 compacity (in Hz), was taken for V_1 /a/ and V_2 /i/. This F2-F1 compacity was used to carry out two analyses:
(a) first, to assess whether there is anticipatory V-to-V coarticulation for both boundary conditions and intervening consonants, F2-F1 compacity of V_1 /a/ followed by V_2 /i/ is compared to the one of V1 /a/ followed by V_2 /a/. A higher compacity is expected for V_1 /a/ when V_2 is /i/.
(b) second, to compare conditions in terms of the degree of coarticulation, a measure of acoustic assimilation of V_1 /a/ to V_2 /i/ is computed within each /aCi/ token as the difference between the compacity of /a/ and /i/ over the compacity of /i/:

(1) *acoustic assimilation index:*

$$\frac{((F2-F1)_{/a/} - (F2-F1)_{/i/}}{(F2-F1)_{/i/}}$$

A higher acoustic assimilation index corresponds to a higher degree of coarticulation.

The second proxy of coordination relies only on temporal information. V-to-V acoustic lag is measured as the distance from the acoustic onset (periodicity in the signal and apparition of formants) of V_1 to the acoustic onset of V_2. This lag covers V_1.C duration and a possible break between the two syllables, even if a pause is not expected within a word, or between a subject and a verb in the across-word condition.

For both the acoustic assimilation index and the V-to-V lag measure, variability across repetitions is measured as the absolute deviation of each token from the grand mean in each condition per speaker.

2.3. Statistical analysis

The data were analyzed with either linear mixed models, used for the analyses on coarticulation, or generalized mixed models, used for the analyses on V-to-V lags and variability. The choice between a linear or generalized model was motivated by data distribution. Models were built in the R environment with the lmer() or glmer() functions of the lme4 package [12]. The best-fitting distribution of the generalized linear mixed model was chosen by comparing AIC values. A gamma distribution was selected for the analyses on variability (log link for acoustic assimilation variability and inverse link for V-to-V lag duration variability) while an inverse Gaussian ($1/\mu^2$ link) distribution was chosen for V-to-V lag.

To test for the presence of coarticulation, intended as a spectral difference between /aCa/ and /aCi/ sequences, a model was run with F2-F1 compacity of V_1 /a/ as dependent variable and BOUNDARY ('within-word' vs. 'across-words'), CONSONANT (/p/ vs. /ʁ/) and V_2 TYPE (/a/ vs. /i/) as fixed effects and their interaction.

To compare conditions and C articulatory constraints, four models were run with *acoustic assimilation index, assimilation index variability, V-to-V lag,* and *lag variability* as dependent variables and BOUNDARY and CONSONANT as fixed effects and their interaction. Random intercepts for SPEAKER and REPETITION were included (except for the models on variability).

The effect of each fixed factor and interaction on the dependent variable was tested either by comparing the model with the given interaction/factor with a model lacking that particular interaction/factor by performing a chi-squared test, or with a simple type III ANOVA analysis. Post-hoc comparisons were performed with the *emmeans* package [13].

3. RESULTS

3.1. F2-F1 compacity

Presence of V_2 to V_1 anticipatory coarticulation is found regardless of consonant or position, as shown by a higher F2-F1 compacity of V_1 /a/ when it is

followed by V_2 /i/ than by V_2 /a/ (as illustrated by the pairwise comparisons in Table 2). However, the magnitude of the difference between the two vocalic contexts depends on consonant type and position (see Table 3).

BOUNDARY	V2	mean (sd)	*p-value*
/p/			
within-word	/a/ /i/	870 (129) 1131 (171)	p <.001
across-words	/a/ /i/	912 (136) 1028 (142)	p <.001
/ʁ/			
within-word	/a/ /i/	673 (164) 869 (140)	p <.001
across-words	/a/ /i/	788 (137) 878 (156)	p <.001

Table 2: F1-F2 compacity of V_1 /a/ according to V_2 for each boundary and consonant condition.

Fixed effects	*p*-value
V_2 TYPE	p <.001
V_2 TYPE*BOUNDARY	p <.001
V_2 TYPE*CONSONANT	p <.001
V_2 TYPE*BOUNDARY* CONSONANT	p =.03

Table 3: Type III ANOVA analysis results on the relevant fixed effects and interactions of the model for F2-F1 compacity measure.

3.2. Acoustic assimilation index

As illustrated in Fig. 1, a higher degree of coarticulation, corresponding to higher values of the acoustic assimilation index, is found within-words than across-words, but only in the /p/ context ($p < .001$) and not in the /ʁ/ context. A summary of the fixed effects and their interaction is given in Table 4.

Fixed effects	*p*-value
BOUNDARY	p <.001
BOUNDARY*CONSONANT	p <.001

Table 4: Type III ANOVA analysis results on the model's relevant fixed effects and interactions for acoustic assimilation index.

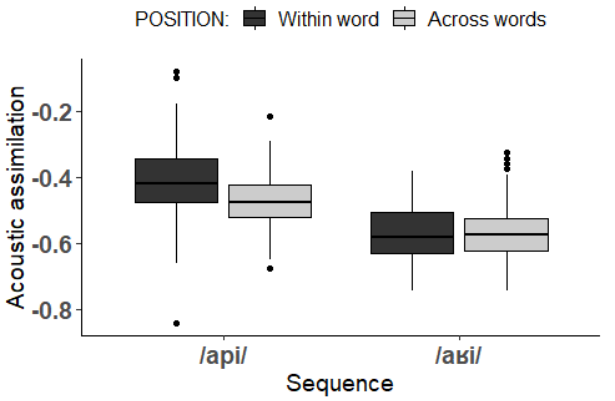


Figure 1: Acoustic assimilation index of the V_2 /i/ on the V_1 /a/ across consonant type (/p/ and /ʁ/) in the two boundary conditions (within-word and across-words). The higher the index, the more there is coarticulation.

3.3. V-to-V lag

V-to-V lag is significantly longer across-words than within-word ($z = 7.571, p < .001$) in both consonantal context: there is no interaction between BOUNDARY and CONSONANT ($X^2(1) = 0.46, p = n.s.$).

3.4. Variability analysis

The variability in acoustic assimilation index does not depend on the boundary condition or consonant type, as shown by the fact that none of the factors contribute significantly to the model.

However, a boundary effect is found for the V-to-V lag measure, but in interaction with the consonant type (BOUNDARY*CONSONANT $p < .001$). Indeed, as illustrated in Fig. 2, V-to-V lag duration is found to be more variable across-words than within-words only in the /p/ context ($z = 6.746, p < .001$), but not in the /ʁ/ context ($z = -1.113, p = n.s.$).

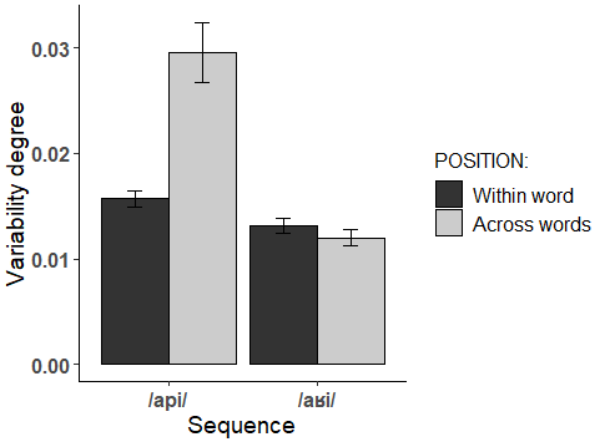


Figure 2: Variability in V-to-V lag (absolute deviation from the grand mean) across boundary conditions per consonant types.

4. DISCUSSION AND CONCLUSION

In this study, we investigated possible acoustic consequences of a difference in coordination between syllables within- and across-words, by looking at both spectral and temporal domains.

The first main result of this study is that there is an anticipation of V_2 in V_1, in terms of spectral assimilation, regardless of the type of intervening consonant and whether this anticipation is observed within or across words, in line with previous studies observing V-to-V coarticulation across one or more words, e.g. [14, 15]. More interestingly, this coarticulation is found to be larger within a word, but only when the consonant separating the two vowels is a labial consonant. This consonant-dependent effect of the boundary condition can be accounted for by the different articulatory constraints in the two types of consonantal context. With an intervening /p/ between the two vowels, the tongue is freer to vary and to adapt its trajectory according to V_2 height. Across /ʁ/, V-to-V coarticulation exists but is much reduced in terms of spectral effects on V_1 (smaller acoustic differences in compacity and lower acoustic assimilation index). The uvular articulation of /ʁ/ entails an adaptation of the V-to-V trajectory. The acoustic consequence of this articulatory constraint is that the contextual effects of /ʁ/ and /i/ on V_1 /a/ are antagonistic: /ʁ/ increases F1 and decreases F2 of the surrounding vowels [16], while /i/ pushes /a/ to a lower F1 and higher F2. With a reduced degree of freedom of V-to-V lingual anticipation in this context, it is possible that differences in coarticulation within- vs. across-words do not surface in the acoustic signal, or do not exist.

Nonetheless, the absence of difference in coarticulation variability shows that the degree of V_2 anticipation is equally stable regardless of the intervening consonant or boundary condition. This stability in the spectral cues of this V-to-V anticipation across repetition is in-line with [10], suggesting that V-to-V anticipation is controlled.

Indication of a higher intersyllabic cohesion within- than across-words comes also from the results on the temporal measures. Indeed, V-to-V lag is consistently longer across a word boundary regardless of the different articulatory constraints imposed by the intervening consonant. This is consistent with previous acoustic studies showing final lengthening at the word boundary (in a sequence composed of two content words, see [8]). Moreover, more variability is found in V-to-V lag duration across word boundaries, again in the freer /p/ context. A more stable timing within-words reinforce the picture of tighter gestural coordination within a lexical unit: for instance, Byrd [4] reported less variability in timing for syllable onset clusters, which are considered to have specific timing relationships, than for clusters spanning a word boundary. However, in this case, the difference within- vs. across-words is not found in /ʁ/ context, for which there is no difference in lag variability depending on position, similar to the pattern found for the measure of anticipatory V-to-V coarticulation. Again, the stronger constraint imposed by the lingual consonant and its coordination with the following vowel in the /ʁV_2/ syllable may hinder a boundary effect.

Taken together, our results provide evidence of a greater cohesion between two syllables belonging to the same lexical word, as suggested by a higher degree of acoustic assimilation (only in /p/ context) and longer lag between V_1 onset and V_2 onset (in both C contexts), and a more stable lag in the /p/ context. Follow-up studies, incorporating articulatory and acoustic data, are on the agenda to further investigate cues of a specific gestural organization within lexical units, as compared to various types of word boundaries in controlled prosodic constituents smaller than the accentual phrase in French.

5. ACKNOWLEDGEMENTS

This study was funded by ChaSpeePro (CRSII5_202228) of the Swiss National Science Foundation and by the program "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL).

6. REFERENCES

[1] Daniloff, R., Moll, K. 1968. Coarticulation of Lip Rounding. *Journal of Speech and Hearing Research*, 11(4), 707-721.

[2] Moll, K. L., Daniloff, R. G. 1971. Investigation of the Timing of Velar Movements during Speech. *The Journal of the Acoustical Society of America*, 50(2B), 678-684.

[3] Lewis, J., Daniloff, R., Hammarberg, R. 1975. Apical coarticulation at juncture boundaries. *Journal of Phonetics*, *3*(1), 1–7.

[4] Byrd, D. 1996. Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, *24*(2), 209-244.

[5] Hardcastle, W. 1985. Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication*, *4*(1–3), 247–263.

[6] Hardcastle, W., Roach, P. 1979. An instrumental investigation of coarticulation in stop consonant sequences. *Current Issues in the Phonetic Sciences*, 531.

[7] Sugahara, M., Turk, A. 2004. Phonetic reflexes of morphological boundaries at a normal speech rate. *Speech prosody 2004*, 353–356.
[8] Turk, A., Shattuck-Hufnagel, S.I. 2000. Word-boundary-related duration patterns in English. *J. Phonetics,* 28, 397-440.
[9] D'Alessandro D., Yildiz A., Fougeron C. 2022. Variation individuelle de la coarticulation en fonction de la frontière prosodique. *XXXIVe Journées d'Études sur la Parole*, Noirmoutier, France.
[10] Whalen, D. H., Chen, W. R. 2019. Variability and Central Tendencies in Speech Production. *Frontiers in Communication*, 4.
[11] Recasens, D., Pallarès, M. D., Fontdevila, J. 1997. A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, *102*(1), 544-561.
[12] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.
[13] Lenth, R. 2022. Estimated Marginal Means, aka Least-Squares Means. *R package version 1.7.3.*
[14] Abry C., Lallouache, T. 1995. Le MEM : Un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français. *Bulletin de la Communication Parlée*, 3, 85-99.
[15] Grosvald, M. 2009. Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics*, *37*(2), 173-188.
[16] Guitard-Ivent, F., Turco, G., Fougeron, C. 2021. Domain-initial effects on C-to-V and V-to-V coarticulation in French: A corpus-based study. *Journal of Phonetics*, *87*, 101057.

Variations of focus prominence in three tone languages

Katrina Kechun Li, Francis Nolan, Brechtje Post

University of Cambridge
kl502@cam.ac.uk, fjn1@cam.ac.uk, bmbp2@cam.ac.uk

ABSTRACT

Prosodic focus involves a diverse range of phonological and phonetic devices, and cross-linguistic variation is often subject to the constraints of the prosodic profile of a language. In tone languages, the functional load of tones limits the scope for focus realisation in terms of discrete categories. The current study investigates whether tone languages can still realise focus in distinctive prosodic patterns by examining three closely related languages: Chengdu dialect (a variety of southwest Mandarin), Changsha dialect (a variety of Xiang) and Guangzhou Cantonese. The analysis of pitch and duration reveals systematic differences in terms of both phonology and phonetic realisation between them. It emerges that the three languages demonstrate various degrees of tonal reduction at pre- and post-focal positions. We relate the variations in sentence-level focal prominence to the languages' lexical-level tonal prosody, and discuss the results in the light of metrical properties of tone languages.

Keywords: prosodic focus, prosodic typology, metrical structure, tone languages, Chinese languages

1. INTRODUCTION

The conveying of information structure, especially focus, has been an important phenomenon in informing typological studies on prosody. Despite the diverse phonological and phonetic devices involved in prosodic focus, the prosodic means available to a language are subject to constraints within its prosodic profile [1]. For example, in languages with lexical stress (e.g. English), focal prominence (or its absence) is typically realised as the application (or non-application) of pitch accents to lexically stressed syllables. This is reflected in the autosegmental-metrical theory (AM) by the association of pitch accents with metrically strong syllables [2]. In contrast, languages lacking lexical stress (e.g. Korean) do not have such an anchor for pitch accents and thus rely more on prosodic phrasing. Nevertheless, there is no straightforward relationship between focus encoding and the lexical prosody of a language, as the phonetic realisation also interacts with other language-specific constraints such as syntactic and semantic features [3]–[5].

In tone languages, prosodic encodings of focus are subject to the constraint of preserving tonal categories, which limits the scope for realising focus in discrete categories. Previous studies have focused on the gradient phonetic effects, primarily pitch modification, which is often accompanied by duration and intensity changes [6]–[8]. However, recent studies also suggest that variations of focus marking in terms of discrete categories might exist even within closely related tonal languages. At the on-focal position, for example, Cantonese and Southern Min have been found to be more limited in f0 modification than standard Mandarin and only consistent in showing a durational change [9][10]. At the post-focal position, tone languages also differ in having post-focal compression or not [11]–[13].

This study extends previous comparative studies by investigating the realisation of prosodic focus in three Chinese languages. Furthermore, we relate the variations in focal prominence to other aspects of their prosodic systems, particularly the reduction in lexical prosody (cf. Standard Mandarin neutral tone). Finally, we discuss the implications of our findings for understanding the underlying metrical organisation of tone languages.

2. METHODOLOGY

Three dialects from distinct Chinese language families were chosen: Chengdu dialect (a variety of southwest Mandarin), Changsha dialect (a variety of Xiang) and Guangzhou Cantonese.

2.1. Materials

Target sentences with SVO structure were constructed in each language. In Chengdu and Changsha, the subject (S) was a name prefix plus a commonly used name, and the monosyllabic verb (V) carried the same tone as the name. The object (O) was also designed to be a disyllabic word with the first syllable carrying the same tone as the preceding name and verb, and the second syllable being a suffix (examples below; the key syllables that were designed to carry the same tone are in bold). The Cantonese dataset, collected at an earlier time and not designed for comparison with other languages, did not control for tone on the object nor the type of onsets. However, it will not affect our ability to

address the main research question, which is how the tonal contour of the same sentence changes in different focus conditions. Two tonal conditions on the key syllables which demonstrated comparable pitch contours across the three dialects are reported here: a maximally dynamic contour reaching the ceiling of the tonal space ('dynamic' group: Chengdu T3[53], Changsha T4[45], Cantonese T2[35]) and a relatively static tonal contour ('static' group: Chengdu T1[45], Changsha T3[42], Cantonese T1[55]). The numbers are tonal representations in Chao number for citation syllables, but Changsha T3[42] undergoes sandhi and changes to [44].

(1) Chengdu:

ɕiaʊ53	**jɪn53**	**mai53**	lə0	pa53	**ji53**	tsɹ0
Name Prefix	Name	Verb	Aspect marker	Classifier	Noun	Suffix
Xiao-yin		bought		a	chair	

(2) Changsha:

ɕiaʊ42	**jɪŋ45**	**mɑi45**	tɑ21	tsɑ34	**iou45**	tsɹ33
Name Prefix	Name	Verb	Aspect marker	Classifier	Noun	Suffix
Xiao-ying		sold		a	grapefruit	

(3) Cantonese:

a33	**jɪŋ35**	**hau35**	tʃɔ35	tʃɐt5-ʃap1	kei35	fɐn55
Name Prefix	Name	Verb	Aspect marker	Number	Numb -er	Noun
A-ying		achieved		seventy	-ish	points

The target sentences were elicited under various focus conditions, including the three reported here (Table 1), which were all induced by wh-questions, but targeted either subject, object or neither (broad focus). The Cantonese dataset did not contain an object focus condition.

Table 1: Focus conditions

Condition	Prompt	Answer
Broad focus	What happened?	Xiao-yin brought a basket.
Subject focus	Who brought a basket?	**Xiao-yin** brought a basket.
Object focus	What did Xiao-yin bring?	Xiao-yin brought **a basket.**

2.2. Recording Procedure

Native speakers of Chengdu (10, f=5), Changsha (8, f=4) and Cantonese (13, f=6) were recorded in a quiet room locally. Participants were presented with an audio recording of the prompt question (recorded by a native speaker) together with the written mini-dialogue, in order to resemble the environment of natural dialogues. Focus and tonal conditions were semi-randomized, with the broad focus condition for each sentence always appearing first. Each sentence was recorded three times (twice in Cantonese).

2.3. Annotation and Analysis

The recordings were segmented using *The Penn Phonetics Lab Forced Aligner* (P2FA) [14]. Although developed for Mandarin Chinese, adapting the dictionaries to fit other Chinese languages yields reasonably good performance given the similar syllable structures. The automatically aligned boundaries for onsets and rhymes were manually corrected. We then extracted duration of the syllable and f0 from ten equidistant time points of the rhyme. A set of *Praat* [15] scripts was developed to facilitate the process (scripts can be retrieved from [16]).

In the analysis, f0 was converted to semitones with each participant's average f0 as their base [17]. We built generalised additive mixed models (GAMM) to analyse the effects of focus, tone (dynamic/static), and language on f0 contours, using R [18] and the *mgcv* package [19]. We also included factor smooth to model the non-linear random effect of the tone and focus conditions for each speaker. The final model was corrected for autocorrelation and fitted with the scaled *t* distribution [20]. Duration was analysed using linear mixed-effects models (LMM), including focus, tone, language, position (Prefix, Name, etc) and their interactions as fixed effects and a random intercept for each speaker.

3. RESULTS

Figure 1 (left) shows the average f0 contours of the dynamic group on the five key syllables. The visualisations can be inspected in two ways. From a paradigmatic perspective, we can compare the f0 contours 'vertically' at the same position, which reveals the effect of different focus conditions. While this may be influenced by different baselines (e.g. a higher overall f0 under broad focus due to all new information), a syntagmatic perspective can reveal how focus changes relative prominence within a focus condition by comparing the pitch level of different positions 'horizontally'. Viewing the f0 contours of Chengdu and Changsha paradigmatically, at the position 'Object(Syl1)', the f0 contour of subject focus condition (blue solid line) is significantly reduced compared to the object focus (orange dashed line) and broad focus condition (grey dotted line), indicating that post-focal material is realised with reduced pitch range. In contrast, the f0 contour in Cantonese barely changes. Similarly, at the position 'Name', the f0 contour of object focus condition also shows some degree of compression compared to that of subject focus and broad focus in Chengdu and Changsha, indicating a 'pre-focal' compression. At both positions, Chengdu has a higher

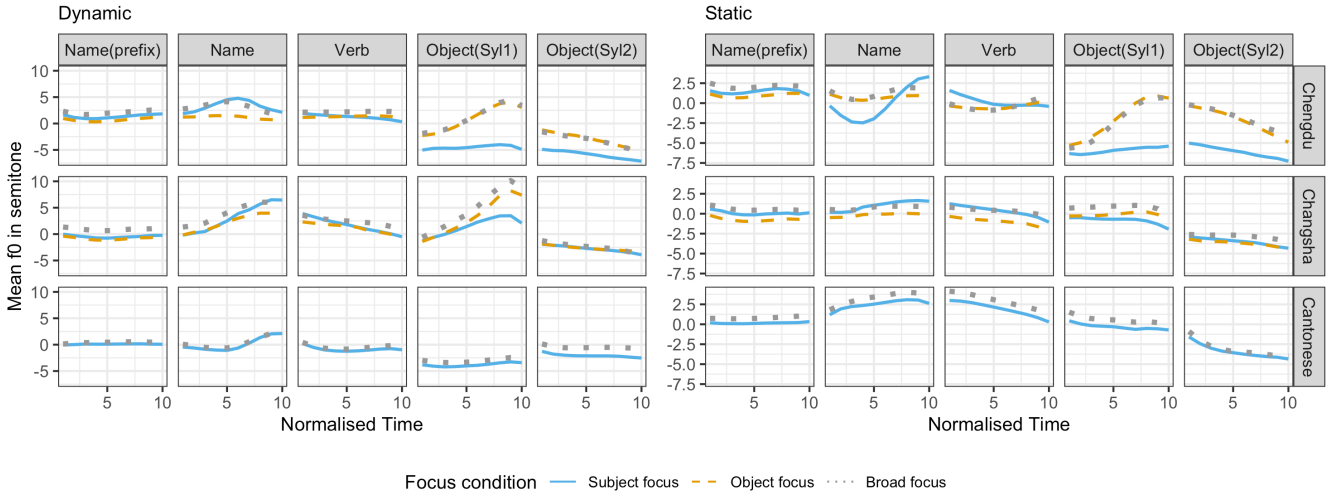


Figure 1: The f0 contours under different focus conditions (left: dynamic group; right: static group).

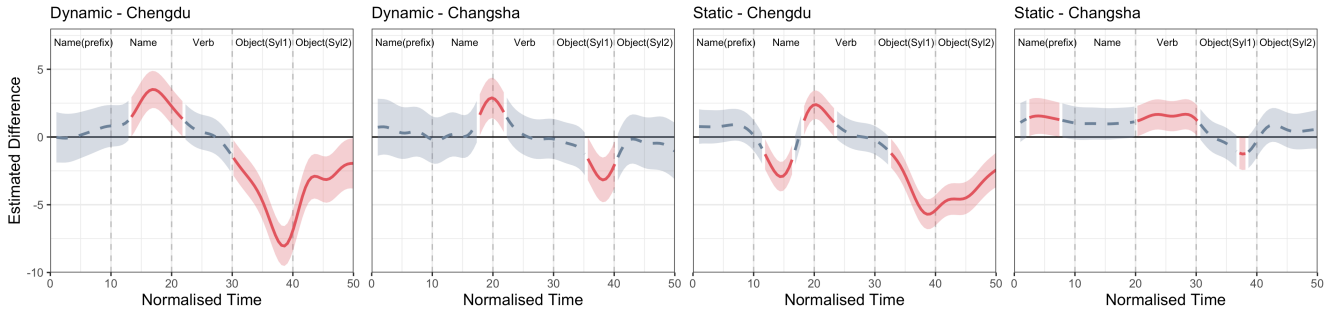


Figure 2: Predicted difference curves (the subject focus minus the object focus condition) in Chengdu and Changsha for the dynamic group (left) and static group (right). The shading displays the 95% CI of the predicted mean difference. The red solid line indicates when the value is significantly different from zero.

degree of compression such that the f0 contour almost becomes flattened, accompanied also by a strong register-lowering effect. The different degrees of compression in Chengdu and Changsha are confirmed using the GAMM model, as shown by the predicted f0 difference between the subject focus and the object focus condition in Figure 2 (left). The significant difference (red solid line) is larger in scale and covers a longer time range in Chengdu.

The f0 findings for the static group tell a similar story as shown in Figure 1 (right; notice the different scale on the y-axis compared to the left). Given that pitch range is less likely to change for level tones, the most noticeable effect from a paradigmatic view is that the slight rising trend on key syllables in Chengdu becomes much more significant under focus. Although there is no such tendency in Changsha, the effect of focus is still evident from a syntagmatic view. The f0 contour of subject focus (blue solid) condition is higher at 'Name' than at 'Object(Syl1)' whereas this tendency is reversed under object focus (orange dashed). For Cantonese, however, the f0 contours remain parallel under different focus conditions, reflecting only a change in baseline. The predicted difference curve generated by the GAMM model (Figure 2 right) confirms that Chengdu involves more f0 changes in marking focal prominence.

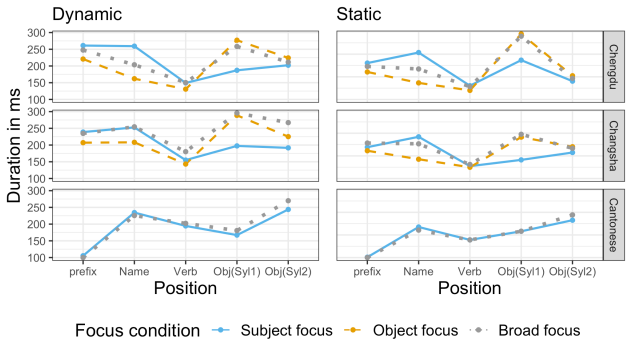


Figure 3: The duration of the dynamic group (left) and static group (right) under different focus conditions.

The duration patterns (Figure 3) mirror the f0 patterns on the syllables. In both Chengdu and Changsha, the blue solid lines of both dynamic and static groups are significantly higher than the orange dashed line at 'Name', and lower at 'Object(Syl1)', indicating a boosting prominence under focus and a reduced prominence elsewhere. The post-hoc pairwise comparisons of different focus conditions on two key syllables (Table 2) confirm the existence of pre-and post-focal durational compression in both languages, and the differences are larger in Chengdu, especially between the subject and object focus, as shown in the last line of Table 2. In contrast, the duration changes are not significant in Cantonese.

Table 2: Post-hoc pairwise comparisons of duration at 'Name' and 'Object(Syl1)' under different focus conditions in Chengdu and Changsha. The number indicates model estimates with p values in brackets (estimates with p<0.0001 are marked in bold).

Focus Condition	Chengdu		Changsha	
	Name	Object	Name	Object
Broad–Subject	**-64.06** (<.0001)	**88.53** (<.0001)	-14.58 (0.3967)	**105.65** (<.0001)
Broad-Object	51.60 (0.0057)	-14.28 (0.3023)	**57.17** (<.0001)	9.69 (0.6640)
Subject-Object	**115.65** (<.0001)	**-102.80** (<.0001)	**71.74** (<.0001)	**-95.97** (<.0001)

4. DISCUSSION

In summary, our findings demonstrate systematic variations in sentence-level compression when marking focal prominence in terms of both f0 contour and duration across the three languages. We showed that both pre-focal and post-focal compressions are possible, despite post-focal compression receiving more attention in previous literature [11]. While Changsha shows a moderate degree of compression with the tonal shapes well maintained, Chengdu exhibits a much larger degree of compression at both pre- and post-focal positions such that the tonal contours are flattened. In contrast, no compression is found in Cantonese. Although the pre-focus condition is not examined in Cantonese, it is likely that this pre-focal compression is also absent since it is often less manifest than the post-focal compression.

These different degrees of sentence-level compression reflect the stability of tonal realisations, which can be related to the phenomena of lexical-level tonal reductions. Specifically, tones can undergo phonetic reduction at unstressed positions in Changsha words, but the tonal neutralisation is incomplete and the original tonal category is partially identifiable [21]–[23]. In Chengdu, in contrast, the tonal neutralisation leads to a complete loss of the original tonal identities and the phonetic realisations entirely depend on the surrounding tonal contexts, similar to Mandarin neutral tones. In Cantonese, however, there is no such lexical tonal reduction.

This apparent correspondence of sentential and lexical level tonal reductions suggests the relevance of underlying metrical structure. Metrical structure captures the relative prominence relationship within a word (e.g. lexical stress) as well as within a phrase or sentence (e.g. relative metrical strength [24, p. 263]). Cross-linguistically, metrical structure can be 'phonologically activated' [25] to different degrees, driving positional (non-)prominence effects on segments or tones. Specifically, languages with a stronger metrical motivation, such as Chengdu, have more freedom to manipulate the tonal contour to express pragmatic functions, as well as showing more stress-like properties on the lexical level. Lexical stress in Chengdu is almost an undisputable property, unlike its debated status in Mandarin, which is evidenced not only by the neutral tone but also the consistent duration lengthening at stressed positions in disyllabic words [26]. Changsha has a weaker metrical motivation compared to Chengdu, followed by Cantonese. For languages like Cantonese where the metrical structure does not allow for drastic changes in relative prominence, there are no phonological phenomena that can be attributed to lexical stress. Alternative means such as syntactic particles are employed for pragmatic functions [27].

Although it is not straightforward to find stable acoustic correlates of metrical structure in tone languages due to the functional loads of tones, the metrical structure at word level has been discussed in other Chinese languages with respect to tonal neutralisation as well as the tone-spreading kind of tone sandhi [28][29]. These languages whose focal prominence has been examined seem to conform to our prediction that their compression at sentence level is in accordance with tonal reduction at word level. The implication is that the dimension of reduction might be particularly useful for characterising prosodic types in tone languages. More studies are needed to further testify to the connection.

5. CONCLUSIONS

The current study provides a comparative analysis of focal prominence by examining the changes that occur in comparable tonal shapes under different focus conditions across languages. The analysis of f0 along with duration reveals that Cantonese, Changsha, and Chengdu show (in this order) increasing levels of compression, not only at post- but also at pre-focal positions. Moreover, sentence-level compression is found to correlate with lexical tonal reduction in these languages, suggesting the relevance of the underlying metrical structure. Languages with a higher degree of activation of the metrical structure have greater flexibility in manipulating the tonal contour at both word and sentence level. At word level, metrical structure can be inferred through phenomena like tonal neutralisation and tone sandhi. At sentence level, focal prominence is particularly informative, as it tends to induce the most obvious alternation of strong and weak. Interpretation of the sentence-level prosodic pattern nevertheless shows the need of combining a paradigmatic and syntagmatic comparison. Finally, this study highlights the importance of prominence reduction in collaboration with prominence enhancement in characterising typological differences in tone languages.

6. ACKNOWLEDGEMENTS

This work was supported by the CSC Cambridge Scholarship and a grant from the University of Cambridge Faculty of Modern and Medieval Languages and Linguistics Fieldwork Fund.

7. REFERENCES

[1] F. Kügler and S. Calhoun, 'Prosodic encoding of information structure: A typological perspective', in *The Oxford Handbook of Language Prosody*, Oxford University Press, 2020. doi: 10.1093/oxfordhb/9780198832232.013.30.

[2] J. B. Pierrehumbert, 'The phonology and phonetics of English intonation', Thesis, Massachusetts Institute of Technology, 1980.

[3] R. S. Burdin, S. Phillips-Bourass, R. Turnbull, M. Yasavul, C. G. Clopper, and J. Tonhauser, 'Variation in the prosody of focus in head- and head/edge-prominence languages', *Lingua*, vol. 165, pp. 254–276, 2015, doi: 10.1016/j.lingua.2014.10.001.

[4] S. Calhoun, 'The interaction of prosody and syntax in Samoan focus marking', *Lingua*, vol. 165, pp. 205–229, 2015, doi: 10.1016/j.lingua.2014.11.007.

[5] F. Hamlaoui, M. Żygis, J. Engelmann, and M. Wagner, 'Acoustic correlates of focus marking in Czech and Polish', *Lang. Speech*, vol. 62, no. 2, pp. 358–377, 2019, doi: 10.1177/0023830918773536.

[6] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, 'Acoustic correlates of information structure', *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 1044–1098, Sep. 2010, doi: 10.1080/01690965.2010.504378.

[7] Y. Xu, 'Effects of tone and focus on the formation and alignment of f0 contours', *J. Phon.*, vol. 27, no. 1, pp. 55–105, 1999, doi: 10.1006/jpho.1999.0086.

[8] Y. Chen and C. Gussenhoven, 'Emphasis and tonal implementation in Standard Chinese', *J. Phon.*, vol. 36, no. 4, pp. 724–746, 2008, doi: 10.1016/j.wocn.2008.06.003.

[9] H. S. H. Fung and P. P. K. Mok, 'Temporal coordination between focus prosody and pointing gestures in Cantonese', *J. Phon.*, vol. 71, pp. 113–125, 2018, doi: 10.1016/j.wocn.2018.07.006.

[10] H.-H. Pan, 'Focus and Taiwanese unchecked tones', in *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, C. Lee, M. Gordon, and D. Büring, Eds. Dordrecht: Springer Netherlands, 2007, pp. 195–213.

[11] Y. Xu, S. Chen, and B. Wang, 'Prosodic focus with and without post-focus compression: A typological divide within the same language family?', *Linguist. Rev.*, vol. 29, no. 1, pp. 131–147, 2012, doi: 10.1515/tlr-2012-0006.

[12] B. Wang, L. Wang, and T. Qadir, 'Prosodic realization of focus in six languages/dialects in China', in *ICPhS XVII*, Hong Kong, 2011, p. 4.

[13] W. L. Wu and Y. Xu, 'Prosodic focus in Hong Kong Cantonese without post-focus compression', in *Proceedings of Speech Prosody 2010*, Chicago, IL, USA, 2010, pp. 1–4.

[14] J. Yuan and M. Liberman, 'Speaker identification on the SCOTUS corpus', *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3878–3878, 2008, doi: 10.1121/1.2935783.

[15] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2019. Accessed: May 26, 2019. [Online]. Available: http://www.praat.org/

[16] K. Li, 'InTone', 2023, https://github.com/kc-li/InTone_Visualisation.

[17] J. Zhang, 'A comparison of tone normalization methods for language variation research', in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 2018.

[18] R Core Team, 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[19] S. N. Wood, 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *J. R. Stat. Soc. B*, vol. 73, no. 1, pp. 3–36, 2011.

[20] M. Wieling, 'Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English', *J. Phon.*, vol. 70, pp. 86–116, 2018, doi: 10.1016/j.wocn.2018.03.002.

[21] C. Guo, 'On the tonal neutralization of bi-syllabic prosodic words of Changsha dialect', *Studies in Prosodic Grammar*, vol. 7, pp. 131–149, 2021.

[22] Q. Zhong, 'Light tones in Changsha dialect', *Fangyan*, vol. 3, pp. 255–264, 2003.

[23] G. Wei, 'On the neutral tone and the atonicity', *Fangyan*, no. 01, pp. 20–29, 2000.

[24] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge: Cambridge University Press, 2008.

[25] L. M. Hyman, 'Do all languages have word accent? or: what's so great about being universal?', *UC Berkeley Phonol. Lab Annu. Rep.*, 2010

[26] C. Liu, K. Li, and F. Nolan, 'On the speech rhythm and stress pattern of tone languages', *Journal of Sichuan University (Philosophy and Social Science Edition)*, no. 3, pp. 151–163, 2022.

[27] P.-P. Lee, *Focus manifestation in Mandarin Chinese and Cantonese: a comparative perspective*. Abingdon; New York: Routledge/Taylor & Francis Group, 2019.

[28] C. Guo and F. Chen, 'Phonetic realizations of metrical structure in tone languages: Evidence from Chinese dialects', *Front. Psychol.*, vol. 13, 2022,

[29] S. Duanmu, 'Metrical structure and tone: Evidence from Mandarin and Shanghai', *J. East Asian Linguist.*, vol. 8, no. 1, pp. 1–38, 1999, doi: 10.1023/A:1008353028173.

MANDARIN QUESTION INTONATION PATTERNS IN THE PRODUCTION OF HUNGARIAN LEARNERS OF CHINESE

Kornélia Juhász & Huba Bartos

Eötvös Loránd University & Hungarian Research Centre for Linguistics
juhasz.kornelia@nytud.hu, bartos@nytud.hu

ABSTRACT

In this acoustic analysis we aim to compare Mandarin Chinese (MC) syntactically marked (trisyllabic) and unmarked (disyllabic) interrogative intonation patterns contrasted with their declarative counterparts in the production of two L2 learners' groups with different level of language experience whose L1 is atonal Hungarian. We hypothesize that the synchronization of tones and intonation in MC poses problems for L2 learners and when the analysed L2 tonal patterns sharply differ from the native L1 pattern, L2 learners favour the L1 pattern. We recorded short dialogues of interrogative and declarative sentences, and the extracted f_0 contours were compared by GAMMs. Our results show that L2 tone sequences containing T1 & T2 pose more difficulties for L2 learners than those with T1 & T4. Our hypothesis regarding the presupposed difficulties of effects of L1 transfer has been partially confirmed, although L2 learner groups with different levels of L2 experience differed in producing L2 interrogative patterns.

Keywords: Mandarin Chinese, tonal and atonal languages, L2 intonation production

1. INTRODUCTION

In this acoustic analysis, we aim to explore how Hungarian learners contrast Mandarin Chinese syntactically marked and unmarked yes-no interrogative intonation patterns to declaratives. In tonal languages, such as Mandarin Chinese (MC), f_0 serves for the realization of both lexical tones and intonation [1]. This means that, on one hand, f_0 modulation is locally dependent primarily on tone values or tonal contexts. On the other hand, these local effects also interact with intonation patterns, yielding the actual f_0 contour [2]. Taking a broader view, in distinguishing MC statement and question intonation patterns, local (e.g., terminal rise on the last (tonal) syllable) and global acoustic cues (raised f_0 over the whole utterance) have been identified concerning f_0 register and f_0 range [3]. Regarding f_0 register, according to Shen's MC intonation model [1], statements display a gradually descending pattern, while unmarked yes-no questions feature a significantly higher f_0 throughout the whole utterance (compared to the declarative contour), complemented by a terminal rise. This terminal rise, in the case of syntactically unmarked ('bare') interrogatives, can be attributed to the absence of lexical/syntactic cues (such as the particle *ma*) of interrogative force, thus prosodic cues are used exclusively to express this force [3, 4, 5]. In the case of syntactically marked ('particle-final') interrogation, an additional particle, the particle *ma*, verbalizes interrogative mood. Lacking an inherent tonal value, *ma* is realized tonally as the prolongation of the contour of the preceding tone: following a rising Tone 2 (T2), the toneless particle has a high target, while following a falling Tone 4 (T4) it has a low target [1, 3].

In contrast to MC, Hungarian is a non-tonal language. While in Hungarian the declarative pattern is realized with a descending contour similar to MC, the prosodic structure of character contour in yes/no questions differs in MC and Hungarian: In the latter, the f_0 contour is characterized by a rising structure followed by a fall (L*HL) [6, 7, 8]. However this character contour only appears at the terminal part of the intonation phase, more precisely, it initiates on the last accented syllable of the utterance [6]. Thus we can define three relevant positional alternants for the realization of yes/no questions [8]: (i) if the final stress group contains only one syllable, then truncation of the final falling phase of the character contour occurs, in this manner the interrogation is realized exclusively with a rising phase (L*H); (ii) If the final stress group contains two syllables, then L* appears on the 1st syllable, and the falling HL is displayed by the 2nd syllable; (iii) if the final stress group contains at least 3 syllables then each target of the character contour has its own syllable to convey. It should be further added that prior to the terminal rising-falling contour, a scale and/or preparatory contour appears, most frequently realized as a half-fall, that is, a slightly decreasing pattern [6: 52].

In this particular case, we are analysing disyllabic broad focus MC utterances, where the first syllable is the subject (他 *tā 'he'* (with high level Tone 1 (T1))), followed by the verb (来 *lái, 'come'* (rising T2) or 去 *qù 'go'* (falling T4)) and in the case of syntactically marked questions, there is a subsequent syllable, the particle 吗 *ma*. We primarily focus on

the verb, because in our examples the verb occurs as the last accented syllable, which is the locus of the terminal rise in bare MC questions, as well as the 'host' defining the pitch value of the particle *ma* in particle-final questions, and in L1 (Hungarian) the last accented syllable is precisely where the character contour is initiated, too.

Additionally, the analysis of the f_0-contour of the 1st syllable is also important, considering the fact that MC elevates the f_0 contour of the whole utterance, while Hungarian does not.

2. HYPOTHESIS

The synchronization of tone and intonation in the production of L2 learners is in the centre of our attention. We chose to use T2 and T4 verbs in our examples because if interrogative and declarative utterances are contrasted, then the interference between the L1 and L2 f_0 curves is expected to be the strongest in these cases. The presupposed interference is induced by the striking disparity of L1 and L2 interrogative and declarative contours with a falling-tone (T4) tonic in questions, and a rising-tone (T2) tonic in statements, and can be attributed to the L1 transfer effect in the L2 acquisition process [9, 10]. In sum we hypothesize that in those cases where the L1 and L2 f_0-patterns differ, L2 learners face difficulties in production, favouring the L1 pattern. The cases, and the presupposed difficulties, are summarized in Table 1. The L1 examples are not analysed in this experiment, just added here as a reference.

	3-syll. syntactically marked interrogation	**2-syll. 'bare' interrogation**	**2-syll. declaration**
L2 T1+T2	Tā lái ma? [tʰa laj ma] H.LH.H *'He comes PRT?'* elevated rising	Tā lái? [tʰa laj] H.LH *'He comes?'* elevated rising	Tā lái. [tʰa laj] H.LH *'He comes.'* rising
L2 T1+T4	Tā qù ma? [tʰa tɕʰy ma] H.HL.L *'He goes PRT?'* elevated falling	Tā qù? [tʰa tɕʰy] H.HL *'He goes?'* elevated falling	Tā qù. [tʰa tɕʰy] H.HL *'He goes.'* falling
L1 pattern	És | jön ma? prep. | L*.HL *'And come-3sg today?'* prep. + rise-fall	És | jön? prep. | L*H *'And come-3sg?'* prep. + rise	És | jön. prep. | HL *'And come-3sg.'* prep.+ fall
Presupposed difficulty	- *ma* particle with H target after *lai,* - elevated character	- producing *qu* with a fall - elevated character	- producing *lai* with a rising pattern

Table 1: The analysed L2 utterances, compared to the exemplary L1 patterns, and the summary of the presupposed difficulties

3. METHOD

We analysed three adult speaker groups (5 female speakers per group): 1. Hungarians with cca. one year language experience of MC: second year Chinese studies undergraduates ('beginners'); 2. Hungarians with 3-4 years of learning MC: Chinese Studies master's programme students ('advanced learners'), and 3. a control group of Chinese natives. We recorded the utterances summarized Table 1., presented as short question–answer dialogues, projected on a screen with both Chinese characters and pinyin transcription. Both 'particle-marked Q + declarative A', and 'declarative + bare echo-Q' utterance-pairs were repeated 5 times. In this manner we recorded a total of 2×150 = 300 utterance pairs, where 'declarative + 2-syll echo-Q' pairs added up to 600 syllables, and pairs with 3-syll 'particle-marked Q + A' added up to 750 syllables (2 sentence types × 2 tones × 5 reps × 15 speakers). Within voiced segments, f_0 was extracted by 5 ms intervals automatically in Praat [11]; since the the onsets of the 2nd syllables of T4-tonic utterances were voiceless segments, they were excluded from the analysis. The extracted f_0 values were converted to semitones (with a reference value of 50 Hz [12] in R [13]) and f_0 curves were analysed by GAMMs [14], using the packages *mgcv* [15] and *itsadug* [16]. We ran 8 models in total: in the case of 'T1+T2' utterances the first two syllables were analysed together since they consisted exclusively of sonorants (after the initial aspirated [tʰ]). Since 'T1+T4' sequences contained a voiceless onset [tɕʰ], we had to analyse the first two syllables separately. The 3rd syllable of interrogatives (the particle *ma*) was analysed in an additional model as well, because it had no counterpart in declaratives. As regards the structure of the models, in each case, f_0 was analysed dependent on the normalized duration of the vocalic section of the syllable, and the models were further complemented by a parametric factor (with contrast treatment), which we defined as an ordered combined variable of speaker group and sentence type. Within the ordered variable, native declaration (Nat D) was chosen as the reference curve, from which difference curves were computed. Additionally, a random smooth function was applied to each f_0-trajectory, respectively. The models were treated for autocorrelation [17].

4. RESULTS

Our results with the estimated f_0 curves are shown in Figure 1., and the parametric coefficients of the GAMMs are summarized in Table 2. Regarding the pairing of a particle-marked 'T1+T2+*ma*' interrogative (I) and declarative answers (D), native speakers produced these questions with an estimated 3 semitones higher f_0 compared to Ds, and thus they produced a significant difference up until the last 12% of the 2nd syllable, from where the two f_0 curves overlapped. As expected, the rising pattern of the 2nd syllable was extended to the 3rd syllable (the *ma*

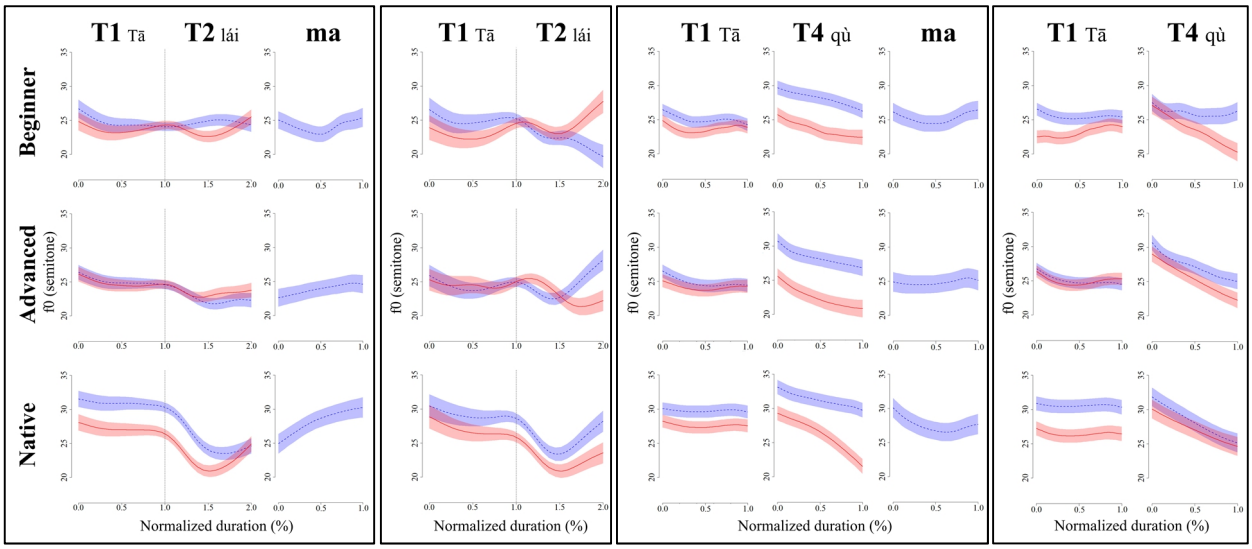


Figure 1: The estimated f_0 curves of the four groups of tri- and disyllabic interrogative (blue dashed curve) + declarative (red solid curve) pairs within the normalized duration of each vocalic section (95% conf. interval)

Synt. Marked T1+T2+*ma* I + D					Bare T1+T2 I + D			Synt. Marked T1+T4+*ma* I + D						Bare T1+T4 I + D				
Para-	TA LAI		MA		Para-	TA LAI		Para-	TA		QU		MA		Para-	TA		QU	
metric	*Est.*	*t*	*Est.*	*t*	metric	*Est.*	*t*	metric	*Est.*	*t*	*Est.*	*t*	*Est.*	*t*	metric	*Est.*	*t*	*Est.*	*t*
D_Nat.	24.4	70.7	-	-	**D_Nat**	24.1	68.6	**D_Nat**	27.6	61.2	26.2	51.4	-	-	**D_Nat**	26.4	53.4	27.2	41.9
I_Nat.	27.6	20.0	28.2	43.8	**I_Nat**	26.8	19.1	**I_Nat**	29.8	12.8	31.3	40.9	27.5	41.6	**I_Nat**	30.6	18.1	28.3	3.5
D_Adv.	23.9	-1.0	-	-	**D_Adv**	23.8	-0.6	**D_Adv**	24.1	-5.4	22.5	-5.0	-	-	**D_Adv**	24.9	-2.4	25.5	-2.0
I_Adv.	23.6	-1.8	23.9	-5.0	**I_Adv**	24.4	0.7	**I_Adv**	24.8	-4.3	28.2	2.8	24.9	-2.9	**I_Adv**	25.0	-2.1	27.3	0.1
D_Beg.	23.7	-1.6	-	-	**D_Beg**	24.0	-0.2	**D_Beg**	23.6	-6.4	23.6	-3.7	-	-	**D_Beg**	23.1	-4.9	23.7	-4.0
I_Beg.	24.6	0.4	24.1	-4.8	**I_Beg**	23.3	-1.5	**I_Beg**	25.0	-4.1	28.1	2.8	25.2	-2.6	**I_Beg**	25.5	-1.5	26.0	-1.4
R^2	96.0%		99.6%		**R^2**	94.5%		**R^2**	97.8%		98.7%		99.8%		**R^2**	97.3%		96.2%	

Table 2: The 8 GAM-models' parametric coefficients (estimated average f_0 *(Est.)*, *t*-value) for the two sentence types (interrogative = I, declarative = D) produced by the three speaker groups (Native = Nat, Advanced = Adv, Beginners = Beg), (*t*-value significant if $|t| \geq 1.96$)

particle), which was thus realized in a relatively high f_0 range. In contrast, both L2 learner groups produced more overlapping f_0 curves in this case. Neither group discriminated between the 1st syllable of Is and Ds to a significant extent. Advanced learners produced the 2nd syllable of Ds with a significantly higher f_0 compared to the 2nd syllable of Is (sign. diff: 140–200% of normalized duration). Beginners produced reverse curve patterns for the two moods: domed for Is and concave for Ds (sign. diff: 28–85% of the 2nd syllable). The particle *ma* in the 3rd syllable was produced by L2 learners in a lower f_0 range compared to natives, with a significantly different shape as well.

In unmarked 2-syllable 'T1+T2' I and D pairs, in natives' production the two sentence types were significantly discriminated for the whole normalized duration with an average difference of 2.7 semitones (st), and with Ds positioned in a lower f_0 range than Is. In contrast, both L2 groups produced intersecting I and D curves, and the estimated average difference throughout the two curves was less than 1 st. Advanced learners made no distinction between the two sentence types in the 1st syllable. Additionally, in interrogatives, T2 in the 2nd syllable featured the native-like rising, however Ds showed a strikingly opposite, falling, pattern. In beginners' production, the 1st syllable of Ds was realized in a significantly lower f_0 range compared to Is, however the difference resulted from the gradually rising pattern of declarative T1, meaning that the high level quality of the T1 was missing. In contrast, the T1 in interrogatives was realized with a level pattern, while the rise of T2 in the subsequent syllable was replaced by a falling pattern.

In pairs of particle-marked 'T1+T4+*ma*' I and D, natives produced Is with significantly higher f_0 compared to Ds for the whole duration of the 1st syllable, with an average difference of 2.2 st. In contrast, L2 groups produced less difference, on average, between the two sentence types in the 1st syllable (beginners: ~1.4 st, advanced learners: < 1 st). Discriminating the falling T4 contours of the two types in the 2nd syllable did not pose difficulties for neither L2 learner groups: similarly to natives (who produced an average of 5.1 st difference), advanced learners featured a difference of 5.7 st, and beginners a difference of 4.5 st on average, between the two sentence types. As for the shape of tonic T4 in declaratives, natives produced a slightly domed falling pattern, but both L2 learner groups showed a concave descending pattern (significantly differing from the native shape), positioned to a lower f_0 range compared to those of natives. The contour of the

particle *ma* – here, too – was realized as a prolongation of the preceding falling pattern, however in this case L2 learners positioned the f_0 curve in a lower f_0 range compared to natives.

In disyllabic 'T1+T4' utterances, native speakers significantly discriminated (bare) Is and Ds for the whole duration of the 1st syllable, producing Is with an average 4.2 st higher than Ds. In a similar manner, beginners also distinguished the curves of the two sentence types on the 1st syllable, producing Is an estimated 2 st higher compared to Ds. In the advanced learners' production, the curves for two sentence types intersect in this phase, as Ds feature a rising, and Is feature a quasi-level, pattern. Moving on to the 2nd syllable, the natives' T4 featured a falling structure with the two curves converging up 72% of the curve. However, both L2 groups produced diverging curves here, overlapping in the first ~20%. Advanced learners produced native-like falling T4 patterns, but beginners featured the interrogative T4 with a level structure.

5. DISCUSSION

In this acoustic analysis we compared how L2 learners contrast 'bare' and particle-marked interrogatives with declaratives in MC, using very short example utterances. Native production featured a clear distinction, with question curves positioned to a higher f_0 compared to declaratives and in the case of particle-marked questions, the f_0-level of the particle *ma* appeared as a prolongation of the preceding tonal curve pattern. Concerning L2 learners' production, we hypothesized that due to the interference of L1 intonation and L2 tonal patterns, having to contrast Is with Ds poses problems for the learners – in particular, we suspected L1 transfer in those cases where the L1 and L2 f_0 patterns were inversely related.

Beginners' interrogative contours showed signs of L1 transfer, but not in the expected manner. We hypothesized that the interrogative character contour would begin on the verb in the 2nd syllable, but the interrogative contour apparently initiated on the 1st syllable in beginners' production of particle-marked 'T1+T2+*ma*' and bare 'T1+T2' questions. Thus in the case of particle-marked questions, the final stress group contains 3 syllables, and each target in the L*HL character contour transferred from L1 has its own syllable to convey. In 'bare' questions, on the other hand, the final stress group contains only 2 syllables, so transferred L* appears on the 1st syllable, and the falling HL is displayed by the 2nd syllable. Beginners' 'T1+T4+*ma*' question patterns also confirm this explanation: Since both L1 and L2 question contours require a H target in the 2nd syllable, the production does not pose any problem and approximates the native pattern. However, 'T1+T4' 'bare' questions partially contradict this hypothesis, since the falling T4 is realized with a level contour by the learners, in defiance of the hypothesized falling patterns concurring in L1 and L2. The level-contour realization of T4 might be attributed to some kind of dissimilation applied in order to avoid the complete overlap of contours of the two sentence types.

Advanced learners often did not differentiate the initial T1 syllable of the two sentence types, and when they did, the distinction did not approximate the native contrast in either disyllabic or trisyllabic utterances. Particle-marked 'T1+T2+*ma*' questions shared similarities with the native pattern, and were exempt from obvious traits of L1 transfer, except for the overlapping sections of the two sentence types. Additionally, in the advanced learners' case, the f_0 range was more compressed than in natives' production. Similarly, the particle *ma* was not produced as high as by natives, just as expected, though it showed a rising pattern. Advanced learners' 'T1+T2' utterances were in proof of our hypothesis: in declaratives the rising curve of T2 is overwritten by the falling pattern of L1 declaratives, while the rise remains intact in interrogatives. It should be noted that the 'bare' questions rarely appear in MC L2 teaching, so in this case the lack of experience on the side of the learners could give rise to direct L1 transfer. Although advanced learners' 'T1+T4+*ma*' and 'T1+T4' questions differed from those by natives both in terms of f_0 level and contour shape, the differentiation of the two sentence types approximated the native contrast. We should further note that the curve of tonic T4 in declaratives differs in its shape from the native pattern, which might again be attributed to L1 transfer [17].

In sum, our results show that Hungarian learners of MC cope better with sequences featuring T4 in the tonic syllable, compared to those with a T2 tonic. Moreover, the two learner groups followed different patterns in establishing a contrast between declaratives and interrogatives. Further testing is desirable with recording spontaneous dialogues and comparing the intonation patterns therein.

6. ACKNOWLEDGEMENT

The research has been supported by ÚNKP-22-3 New National Excellence Program of the Ministry for Innovation and Technology and the National Research, Development and Innovation Fund.

7. REFERENCES

[1] Shen, X. 1990. *The Prosody of Mandarin Chinese*. University of California Press.

[2] Shih, C. L. 2000. A Declination Model of Mandarin Chinese, *Intonation: Analysis, Modelling and Technology* 1, 1–23.

[3] Lee, O. J. 2005. *The prosody of questions in Beijing Mandarin.* Doctoral Dissertation, Ohio State University.

[4] Shen, X. 1989. Interplay of the four citation tones and intonation in Mandarin Chinese, *Journal of Chinese Linguistics* 17, 61–74,

[5] Li, A., Huang, G., Li, Z. 2019. Prosodic Cues in the Interpretation of Echo Questions in Chinese Spoken Dialogues. *Proceedings of APSIPA Annual Summit and Conference 2019* 1056–1061.

[6] Varga, L. 2016. The Intonation of Topic and Comment in the Hungarian Declarative Sentence. *Finno-Ugric Languages and Linguistics* 5/2, 46–77.

[7] Olaszy, G. 2002. The Most Important Prosodic Patterns of Hungarian, *Acta Linguistica Hungarica* 49, 277–306.

[8] Varga L. 2002. The intonation of monosyllabic Hungarian yes-no questions. *Acta Linguistica Hungarica* 49/3–4, 307–320.

[9] Jun, S. A., Oh, M. 2000. Acquisition of Second Language Intonation, *Acoustical Society of America Journal* 107, 73–76.

[10] Ramírez, D., Romero, J. 2005. The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers, *Intercultural Pragmatics* 2/2, 151–168.

[11] Boersma, P., Weenink, D. 2019. *Praat: doing phonetics by computer*. v. 6.1.05, https://www.praat.org/.

[12] Nolan, F. 2003. Intonational equivalence: an experimental evaluation of pitch scales. *Proc. of 15th ICPhS 2003,* 771–774.

[13] R Core Team 2019. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. v. 3.6.1. https://www.R-project.org/.

[14] Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. Chapman and Hall.

[15] Wood. S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* 73/1, 3–36.

[16] van Rij, J., Wieling M., Baayen, R. H., van Rijn, H. 2022. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* R package version 2.4.1.

[17] Juhász K.. in press. Atonális és tonális dallammenetek összehasonlítása. [A comparison of atonal and tonal f_0 patterns] A XVI. Alkalmazott Nyelvészeti Doktoranduszkonferencia Tanulmány-kötete. Nyelvtudományi Kutatóközpont.

THE ABSENCE OF PAUSES IN SPOKEN NARRATIVES AND MEMORY RECALL

Jan Volín and Pavel Šturm

Institute of Phonetics, Charles University, Prague, Czech Republic
jan.volin@ff.cuni.cz, pavel.sturm@ff.cuni.cz

ABSTRACT

An experiment with a memory recall test based on continuous spoken texts was carried out. The primary factor to examine was the absence of pauses. A sample of 24 subjects listened to eight narratives with the task to subsequently fill in the missing words into gaps in a printed version of each of the texts. With pauses defined as cessation of lexical item articulation, the present study tests the hypothesis that spoken texts with the pauses removed will lead to a worse recall of the words contained in them.

The results indicate that pauses have an impact on the effectiveness of language communication, as the lexical content of spoken texts without pauses was more difficult to retrieve relative to texts with the pauses intact. This seems to suggest that apart from speech production requirements, pauses meet beneficially some of the perceptual constraints. Nevertheless, the resulting effect was neither straightforward nor monotonous across texts. Benchmark data for further experimenting are provided.

Keywords: pause, memory, recall, prosodic structure, temporal organization.

1. INTRODUCTION

One of the central questions in linguistics is – or if it is not, perhaps it should be – that of effectiveness of communication. The costs of achieved correspondence between the speaker's intended meaning and the addressee's recovered meaning seem to be an obvious indicator of the effectiveness. However, there is currently no procedure that would allow for measureing them directly. Proxies that are often used instead include the complexity of cerebral processing as reflected in reaction times while fulfilling various behavioural tasks, scaled judgements of the speaker's personality features (e.g., agreeableness, competence) or, finally, various measures of memory performance. This latter one will be focused on in the present study.

Memory recall has been tested prevalently on strings of individual items: numerals, unrelated nouns, non-words, etc. (e.g., [1], [2], [3], [4], [5], [6]). Yet, it is undisputed that natural language communication uses utterances as primary structural units, even if occasionally very short ones. Utterances can be analysed as fields of relations among words [7] and relational links outside the utterance have also been often considered. Lists of isolated items lack this essential feature of everyday language use. The authors in [6] actually state that "…memory for lists of words likely differs from memory for conversational speech in important ways". Therefore, we decided to test the listeners' memory for the contents of spoken texts using continuous strings of utterances in narratives.

It is widely acknowledged that the impact of a message is not attributable solely to the choice of words or syntactic links among them. The prosodic form is seen as an important factor in speech comprehension modelling (e.g., [8], [9], [10], [11]). Interestingly, in [12] an effect of such a crude measure as mean F0 (a correlate of mean pitch termed *key* or *register*) on memory was established with clear, albeit complex results.

Our study focuses on a prosodic event that is also quite salient: the pause. Superficially, pauses are viewed as a response to the plain need to breath, but they actually reflect certain cognitive constrains. First, speakers need pauses during neuro-linguistic planning of their speech unit [13]. Second, and more pertinent to our study, listeners may need pauses to process the contents of utterances and to store information in their memory (see, e.g., [14]).

In [5], for instance, the impact of pause presence on number recollection was investigated. The authors extended their previous research that suggested a beneficial effect of pauses on remembering the contents of numeral strings. Apart from confirming their previous findings, the new experiment showed that longer pauses were better than the shorter ones.

The chief objective of our study is to determine how absence (rather than presence) of pauses affects memory for lexical contents of narratives. Since presence and absence are complementary modes, this wording might deserve explanation. In the previously mentioned study [5], the researchers added pauses into their speech material. We, on the other hand, carefully deleted naturally occurring pauses from our narratives (see Section 2).

Although the null hypothesis would stipulate no difference between our two conditions (*Pause* vs. *NoPause*), we expect words from the spoken texts with the pausing left untouched to be better recalled.

2. METHOD

2.1. Material

Eight extracts from narratives were used for the experiment. They were all obtained from high-quality audiobooks produced in professional studios with experienced actors (4 women + 4 men). The general style could be characterized as fluent speech read out from a well-edited written text. The extracts were selected so that no markedly tense atmosphere nor overly dynamic development of the plot was present. Therefore, we assume that topics of the narratives were of no crucial importance.

As the narrative recordings were made in studios, there were only logical pauses in line with syntactico-semantic structure (guaranteed by the studio director), that is, no abrupt or disturbing dysfluencies.

Mean duration of the extracts was 47.2 seconds with std. dev. of 1.1 seconds. Further descriptors of the sample are displayed in Table 1.

Speaker	DurP+	DurP–	n Wds	n Pss
F1	45.2	37.2	111	13
F2	46.3	37.1	119	14
F3	48.0	41.2	102	19
F4	46.4	34.3	91	15
M1	48.9	35.9	106	17
M2	46.7	38.1	116	15
M3	47.2	36.9	112	23
M4	48.3	37.3	84	18
Mean	**47.2**	**37.2**	**105.1**	**16.8**
Std.dev.	**1.1**	**1.8**	**11.5**	**3.0**

Table 1: Descriptive facts on the narrative extracts used in the test. DurP+ = Duration of the complete extract in seconds; DurP– = Duration with pauses removed (s); n Wds = number of words; n Pss = number of pauses.

Table 1 reveals that although the spoken text durations (whether with or without pauses) were quite balanced across the sample, the number of pauses ranged from 13 to 23. Due to the repeated-measure design of the study, other differences (tempo, liveliness, etc.) were not considered crucial.

Pauses were carefully removed in Adobe Audition CS6 sound editor with special attention to the resulting smooth transitions. In other words, sharp cut-offs or other sound artefacts were not allowed. Note that all other prosodic signals such as phrase-final lengthening (deceleration) were left untouched.

Two counterbalanced versions of a perception test were created in which half of the extracts had naturally occurring pauses and the other half had pauses removed. Listeners were assigned to the two groups randomly.

2.2. Participants

Volunteers from various philological programmes at the Faculty of Arts, Charles University were recruited to participate. There were 18 female and 6 male students (n = 24). Their mean age was 20.4 years and they reported no hearing nor dyslexic problems. In all the cases, the Czech language was their mother tongue. The volunteers expressed their wish to participate in a 'strictly anonymous listening test investigating how people remember individual words in story-telling'. They were blind to the experimental condition and a post-test interview revealed that they found the test interesting but difficult.

2.3. Testing procedure

The participants were invited in groups of 3 or 4 into a quiet room with a quality sound equipment. They were informed that extracts from spoken narratives would be played for them to listen to. After that, they would receive a printed transcript of each narrative with 10 words in each extract missing. Their task was to fill in the missing words into the gaps in the written text. Two spoken texts were presented at a time. Then the printed sheet for the first text would be handed out and, after completion, the second text would be done. The time limit for one text completion was 2 minutes.

At the end of each allotted time, the respondents were instructed to indicate whether they recognized the text by ticking YES or NO at the bottom of their answer sheet. The texts were generally unknown, only exceptionally the participants knew from which book the extracts came.

Before a new pair of extracts was played, the experimenter asked the participants about their feelings and encouraged them for the next stage. The whole testing session took 25 minutes (30 minutes with the arrival, greetings and the instruction).

2.4. Predictability without the sound

Before the actual perception test, the written texts were piloted with 8 different volunteers (of the same social characteristics as the perception test participants) to examine the predictability of the target lexical items. In other words, we tested which of the gaps could be filled in without listening to the spoken narrative.

The results of this pre-test showed that most of the words were utterly unpredictable (63/80 = 78.8%). Eight out of the remaining 17 words were correctly 'guessed' by just one of the subjects. Only three words (3/80 = 3.8%) manifested predictability over 50%, that is, more than half of the subjects filled them in correctly. These results were later taken into account in the course of the analyses (see below).

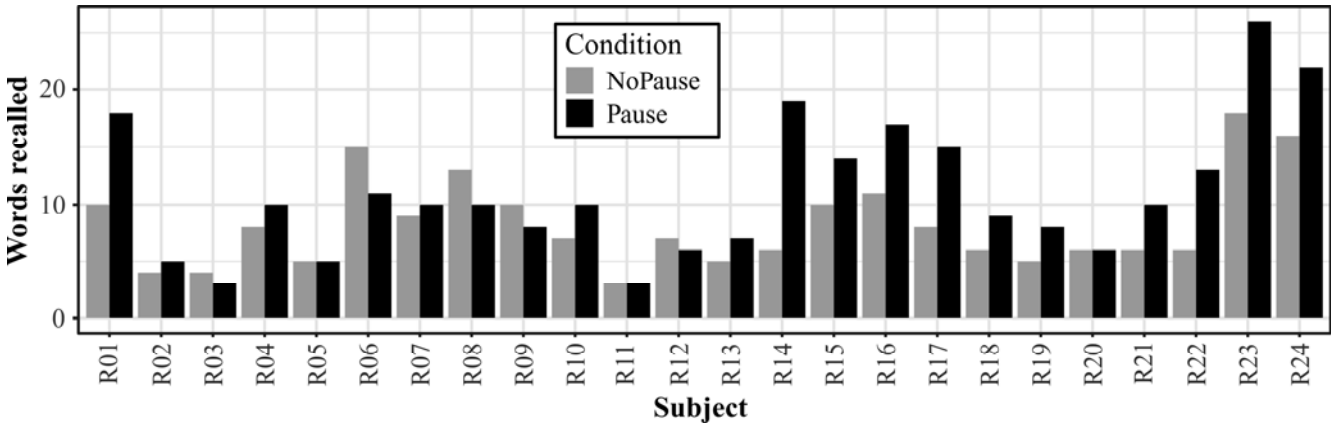


Figure 1: Counts of correctly recalled lexical items by 24 respondents (R01–R24) under two experimental conditions.

3. RESULTS

Due to the intended difficulty of the task, only 463 out of 1920 items (24 subjects × 80 words) were recalled correctly. In terms of percentage, it is 24.1% success rate. When split by CONDITION, the recall rates were 20.6% under the *NoPause* condition and 27.6% under the *Pause* condition (198 against 265 correctly recalled items, respectively). There were 10 words to be recalled per text. On average, 2.47 words were remembered correctly under the *NoPause* condition, whereas under the *Pause* condition it was 3.31 words.

Out of 40 missing words under each condition, an average respondent remembered only 8.25 under the *NoPause* condition, compared to 11.04 under the *Pause* condition. However, Figure 1 shows that individual respondents contributed to the results in a disparate manner. Whereas R23 and R24 are clearly champions of the sample, R03 and R11 had extremely poor recall. Be that as it may, two thirds of the respondents (16 out of 24) performed in the expected direction, that is, they achieved better results under the *Pause* condition. In contrast, there were also 5 subjects who performed in the opposite direction and 3 who had equal performance under both conditions.

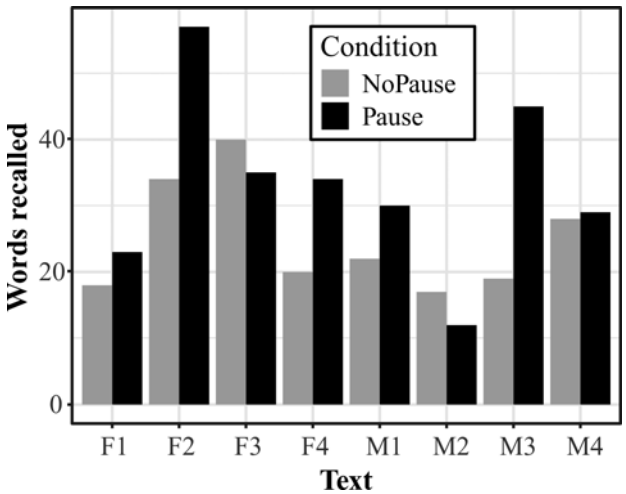


Figure 2: Counts of correctly recalled lexical items in 8 spoken texts under two experimental conditions.

The results can be further broken down according to the individual experimental texts. Figure 2 shows that the effect size varies across the eight texts and two of the texts (F3 and M2) even produced results in an unexpected direction.

To evaluate the difference between *Pause* and *NoPause* conditions statistically, we constructed a generalized linear mixed model with the Poisson distribution. The situation in Figures 1 and 2 suggests that the effect should vary freely across subjects and texts in the model. The count data included 192 rows (8 texts × 24 subjects) specifying the number of observed correct responses (max = 10 in each cell). The following syntax was used in *lme4* package [15]:

```
glmer(nCorr ~ Condition + (1+Condition|Subject) +
(1+Condition|Text), family = poisson, data = data)
```

Comparison of the full model to the reduced model (without the CONDITION predictor) led to statistically significant differences in model fit using likelihood-ratio tests ($\chi^2(1) = 4.94$, $p = 0.026$). The model predicts 1.85 (SE = 0.277) words recalled per text and subject in the *NoPause* condition but 2.41 in the *Pause* condition (SE = 0.370). However, singular fit was returned, indicating an excessively complex structure for the data. Therefore, we removed the by-text and by-subject slopes from the model; the intercept-only model produced almost identical predictions (*NoPause* = 1.85, SE = 0.28; *Pause* = 2.42, 0.359), and CONDITION was a significant predictor ($\chi^2(1) = 8.23$, $p = 0.004$). In conclusion, the effect – an advantage for *Pause* over *NoPause* of approximately 0.5 words per text – holds across subjects and texts.

For each word, we plotted its predictability from the pre-test and its recall in the experiment (Fig. 3a). Although the results varied (especially for zero predictability), there was a clear link between the two variables. The easier it was to guess the word, the better it was recalled after listening, producing a significant correlation ($r = 0.58$, $p < 0.001$). However,

Figure 3b plots the predictability vs. the effect size (*NoPause* condition subtracted from *Pause* condition; positive values mean the effect is in the expected direction). The absence of any significant correlation ($r = 0.15$, $p = 0.172$) confirms that the effect is due to the memory recall experiment manipulations.

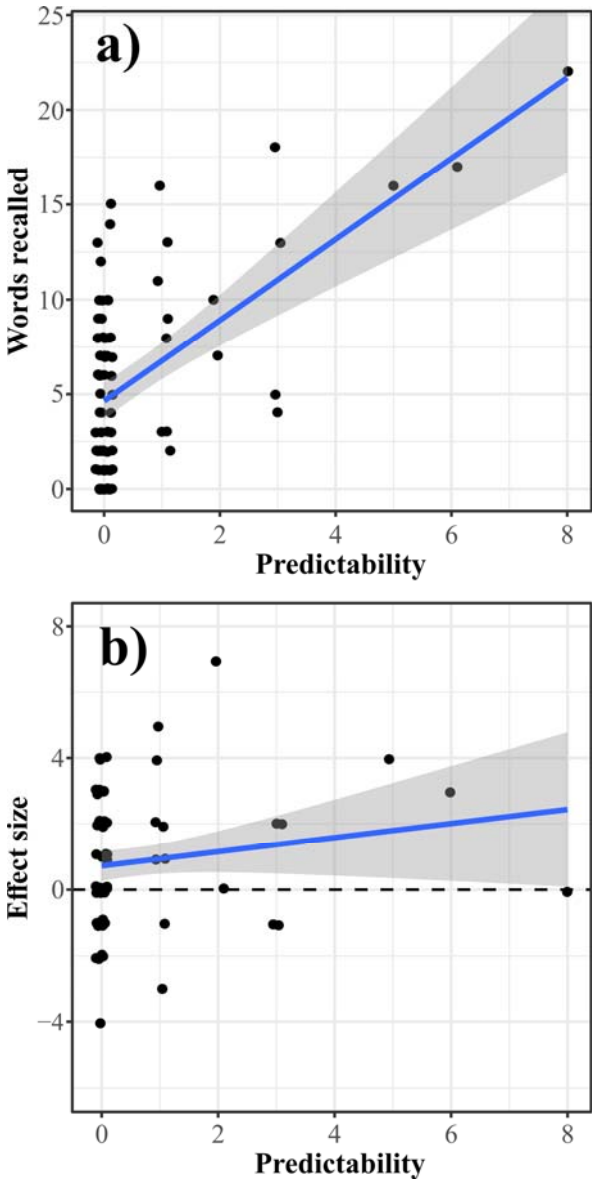


Figure 3: The relationship between predictability of words based on a pre-test and (a) recall in the experiment; (b) manipulation effect size. Jitter was applied to increase visibility of overlaid points.

4. DISCUSSION

The effect that we found may look impressive if we cite 265 correctly recalled items from texts with pauses against 198 items from the same texts without pauses. In reality, the effect is not particularly robust, even though it is statistically significant. It must be noted, however, that the more robust results reported in literature are usually achieved with highly artificial material that is distant from typical communicative language use. We, on the other hand, used narrative texts that were produced by renowned authors and told by experienced storytellers. The contents of the texts were not trimmed in any way, so their communicative purpose is genuine.

The exact underlying cerebral mechanisms responsible for the effect are still unclear. The authors in [6] speak about items that are "processed more deeply by listeners, resulting in improved recall". Such "deep processing" could probably be phrased as "more thorough processing that requires specified time lapse". If the time is unavailable (the speech keeps running), the processing is somehow incomplete. One fact from the past research that has been also corroborated by our experiment, however, seems to be apparent: listeners' memory is influenced by the prosodic structure of speech.

Our future work needs to turn to other prosodic cues as well. Researchers in [4] clearly showed that it is not "any grouping" that is beneficial for speech processing. In their experiment, they compared groupings delimited only by pauses with groupings delimited by clear intonation contours. Groups with melodic characteristics led to better memorability. Similarly, the influence of phrase-final deceleration, together with amplitude and voice quality changes, need to be examined.

It should also be noted that we only tested short-term memory retention. Helfrich and Weidenbecher found in their experiment that voice pitch does not affect immediate text retention, yet has a significant impact on long-term memory [12]. That poses still another research challenge.

Lastly, we are not aware of any study similar to ours on the Czech language. Ultimately, cross-linguistic comparisons in this field that take into account prosodic and syntactic specificities of Czech might be possible.

5. CONCLUSION

Pauses in spoken narratives should not be considered important only for speech production (breathing, neurolinguistics planning) but also for speech perception. The listeners in our experiment clearly benefited from their presence in a recall task. With the pauses absent in otherwise identical spoken texts, the lexical recall was significantly worse. Pauses thus clearly belong to the temporal structure of speech and fulfil important roles with regard to effectiveness of language communication.

6. ACKNOWLEDGEMENTS

The study was carried out with the support of GAČR (Czech Science Foundation), Project 21-14758S.

7. REFERENCES

[1] Baddeley, A. D. 1966. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *The Quarterly Journal of Experim. Psychology* 18(4), 362–365.

[2] Turner, M. L., Engle, R. W. 1989. Is working memory capacity task dependent? *Journal of Memory and Language* 28, 127–154.

[3] Mathy, F., Feldman, J. 2012. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* 122, 346–362.

[4] Savino, M., Bosco, A., Grice, M. 2013. Intonation and positional effects in spoken serial recall. *Proc. of the Annual Meeting of the Cognitive Science Society*, 35, 3360–3365.

[5] Elmers, M., Werner, R., Muhlack, B., Möbius, B., Trouvain, J. 2021. Evaluating the effect of pauses on number recollection in synthesized speech. *Proc. 32nd Conference Elektronische Sprachsignalverarbeitung (ESSV '21)* Berlin, 298–295.

[6] Kimball, A. E., Yiu, L. K., Watson, D. G. 2020. Word recall is affected by surrounding metrical context. *Language, Cognition and Neuroscience,* 35(3), 383–392.

[7] Firbas, J. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press.

[8] Sanderman, A. A., & Collier, R. 1997. Prosodic phrasing and comprehension. *Language and Speech* 40, 391–409.

[9] Cutler, A., Dahan, D., van Donselaar, W. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40, 141–201.

[10] Clifton, C., Carlson, K., Frazier, L. 2002. Informative prosodic boundaries. *Language and Speech* 45, 87–114.

[11] Dahan, D., Ferreira, F. 2019. Language comprehension: Insights from research on spoken language. In: Hagoort, P. (ed), *Human Language: From Genes and Brains to Behavior*, pp. 21–33. MIT Press.

[12] Helfrich, H., Weidenbecher, P. 2011. Impact of voice pitch on text memory *Swiss Journal of Psychology* 70 (2), 85–93.

[13] Hawkins, P.R. 1971. The syntactic location of hesitation pauses. *Language and Speech* 14, 277–288.

[14] Shipstead, Z., Lindsey, D. R., Marshall, R. L., Engle, R. W. 2014. The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language* 72, 116–141.

[15] Bates, D., Maechler, M., Bolker, B. & Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 526 67, 1–48.

THE EFFECT OF PROSODIC PROMINENCE ON DIPHTHONG REDUCTION IN TAIWAN MANDARIN SPONTANEOUS SPEECH –USING /aɪ/ AS AN EXAMPLE

Chieh-Ching Chen[1], Janice Fon[1,2,3]

[1]Graduate Institute of Linguistics, National Taiwan University (NTU), [2]Graduate Institute of Brain and Mind Sciences, NTU, [3]Neurobiology and Cognitive Science Center, NTU
r10142003@ntu.edu.tw, jfon@ntu.edu.tw

ABSTRACT

Vowel reduction is frequently observed in spontaneous speech. Previous studies have shown that acoustic reduction of monophthongs was manifested through formant undershoot and shorter duration. For diphthongs, which differ from monophthongs by having an additional vowel target, little is said about whether and how reduction should take place. This study thus examines how the reduction of the diphthong /aɪ/ is realised in Taiwan Mandarin and how stress influences the reduction pattern. Results showed that /aɪ/ reduction is mostly in the form of monophthongization with two vowel targets merged. Besides, diphthongs without stress were shorter in duration with formant undershoot on both vowel targets. Finally, the reduction of /aɪ/ likely lacks a negative connotation since no gender difference was found.

Keywords: vowel reduction, diphthong, prosodic prominence, spontaneous speech, Taiwan Mandarin.

1. INTRODUCTION

Vowel reduction can be classified as lexical vowel reduction and acoustic vowel reduction [1, 2]. While lexical reduction is a phonological process that can only be observed on particular words (tel**e**graphy [ɛ] vs. tel**e**graphic [ə]), acoustic reduction is a phonetic process in that vowels are realised with shorter duration and formant undershoot [3]. Previous studies showed that reduction appears more frequently in spontaneous speech [4], and prosodic prominence greatly influences the reduction of vowels. For instance, vowels without lexical stress were more likely to be reduced than those with lexical stress in English and Swedish [1, 3]. In languages without a lexical stress distinction, such as French [5], vowels in non-final positions of a word were more reduced [6]. This study thus intends to examine whether a similar relationship between stress and vowel realisation could be found in Taiwan Mandarin, which is prosodically disparate, and if so, what the exact pattern is.

As a tone language, Mandarin assigns one of the four lexical tones to most of its syllables [7]. The tones are realised fully in prosodically strong positions but are reduced in prosodically weak positions [8]. As prosodic strength is manifested through tonal realisation, not vowel reduction itself, as is more commonly found in stress-timed languages like English, it would be interesting to see whether and how vowel reduction is affected by stress.

Moreover, previous studies on vowel reduction mainly focused on monophthongs [1, 2, 3, 6]. How diphthongs are reduced is less clear. Diphthongs differ from monophthongs by having two vowel targets [9], between which is an additional vowel transition [10], whose duration is vowel-dependent and language-specific [11]. Using 18-min spontaneous speech, Su [12] found that 12% of the four diphthongs in Mandarin were monophthongized. However, the study did not examine whether the reduction was caused by prosodic weakening, nor did it look into how reduced diphthongs were realised. Therefore, in this study, we would like to examine the actual realisation of diphthongs when reduced with larger spontaneous speech data.

/aɪ/ was chosen in this study for two reasons. First, /aɪ/ is the most productive diphthong in Taiwan Mandarin [13]. Therefore, its reduction could show us how diphthongs are generally reduced in the language. Secondly, the distance between the two vowel targets is large for formant measurements. We can visualise different states of diphthong reduction by examining /aɪ/. For example, there might be formant undershoot on either of the two vowel targets, which results in /aɪ/ reducing into other diphthongs. On the other hand, /aɪ/ could be monophthongized with the two targets merged as [e], one target dropped as [a], or centralised as [ə].

The aim of this study could thus be specified as whether and how diphthong /aɪ/ is reduced and affected by stress in Taiwan Mandarin. It was expected that /aɪ/ at a lower stress level would have a higher reduction rate, shorter duration, and more reduced spectral results. Also, males and females would be analysed separately since gender difference was observed in acoustic reduction [6]. In our prediction, males would have a higher reduction rate since in stable sociolinguistic stratification, men use

a higher frequency of nonstandard forms than women [14].

2. METHODS

Four hours of monologue recordings including 8 speakers (4M, 4F) of the Taipei dialect were chosen from the Taiwan Mandarin-Min Spontaneous Speech Bilingual Corpus [15]. All speakers were young (20-35) fluent Mandarin-Min bilinguals. Each speaker talked for about 30 minutes.

2.1. Vowel labelling

All realisations of the vowel /aɪ/ were manually labelled and transcribed by two transcribers and inter-labeller consistency was later checked by the first author. Visually identifiable F1 and F2 were used as major cues to label the vowels regardless of voice quality.

2.2. Stress labelling

The stress level of each syllable was labelled primarily based on principles outlined in the Pan-Mandarin Tone and Break Indices (M-ToBI) [8]. There are three levels of stress, S1 to S3. S1 is the lowest level of stress and is used when a tone has completely lost its tonal shape. S2 is the next higher level. It is used when the tone still retains its distinctive contour, even though some of its tonal specifications have been lost. S3 is the highest level of stress. It is used to label tones that are realised with a full-fledged contour. However, necessary modifications were made to accommodate the peculiarities of spontaneous speech [16], as the original version of M-ToBI was solely based on read speech. In addition to tonal realisation, amplitude was considered an additional important cue in stress labelling.

In spontaneous speech, S2 is the most common and could be considered the default stress level. Detailed labelling criteria can be seen in Table 1.

Stress	Tone	Amplitude
S1	loss of original tonal shape	soft
S2	default	default
S3	tone expanded/raised	loud

Table 1: Labelling criteria of stress suggested for spontaneous speech in Chuang [16].

2.3. Measurements

Ten F1 and F2 values were extracted at equal time intervals from each token with Praat [17]. Formant settings were adjusted to accommodate gender differences.

3.RESULTS

In total, 2,835 tokens of /aɪ/ were labelled, in which males had 1,370 tokens and females had 1,465 tokens. Table 2 summarises the distribution of /aɪ/ realisations. Six tokens were excluded from the following analyses due to laughter or noise resulting in unrecognisable formant values. A total of 47 tokens longer than 300 milliseconds were regarded as hesitation [6] and were thus excluded. Five truncated tokens were also excluded since they were out of the domain of this study. In the end, there were 2,777 tokens for later analyses. Detailed distribution is shown in Table 3.

Retained	Monophthongs			Diphthongs		Else
[aɪ]	[e]	[ə]	[a]	[eɪ]	[aə]	-
1,274	968	286	135	87	24	3

Table 2: Distribution of different realisations of /aɪ/.

Gender	Stress	[aɪ]	[eɪ]	[aə]	[e]	[ə]	[a]
M	S1	76	6	5	92	55	24
	S2	466	24	14	396	94	54
	S3	27	1	-	1	-	-
F	S1	62	7	1	86	56	22
	S2	562	47	4	388	81	35
	S3	81	2	-	5	-	-

Table 3: Number of phonetic realisations of /aɪ/ across gender and stress.

3.1. Overall diphthong reduction rate

In the four-hour recording, only 45.89% tokens of /aɪ/ were retained. 50.04% were monophthongized and 4.07% were reduced to other diphthongs. Monophthongization is mainly in the form of merging rather than deletion, as [e] occupied 69.69% of the reduced monophthongs, and 20.59% were centralised as [ə]. Only 9.72% were reduced to [a], in which the second vowel target was dropped.

3.2. The ratio of reduced /aɪ/ across groups

The ratio of /aɪ/ reduction was calculated from the number of neutralised /aɪ/, including monophthongs and diphthongs, divided by the sum of all tokens spoken in the group. Logistic regression was performed to analyse the effect of Stress (3) and Gender (2) on the reduction rate. Table 4 summarises the statistical results. Two main effects Stress1 and Stress3 were significant. Specifically, S3 had more /aɪ/ retained than S2 ($p < .001$), while in contrast, S1 had more /aɪ/ reduced than S2 ($p < .001$). Nevertheless, gender difference was not found (p

= .213). The result of the ratio of retained and neutralised /aɪ/ is presented in Fig. 1.

	Est.	SE	z value	Pr
(Intercept)	-0.00	0.12	0.00	.998
stress1	-1.03	0.16	-6.43	.000***
stress3	2.40	0.40	6.02	.000***
genderM	-0.21	0.17	-1.25	.213
stress1:genderM	0.43	0.22	1.93	.053
stress3:genderM	0.45	0.84	0.54	.590

Table 4: Logistic regression results of the ratio of reduced /aɪ/.

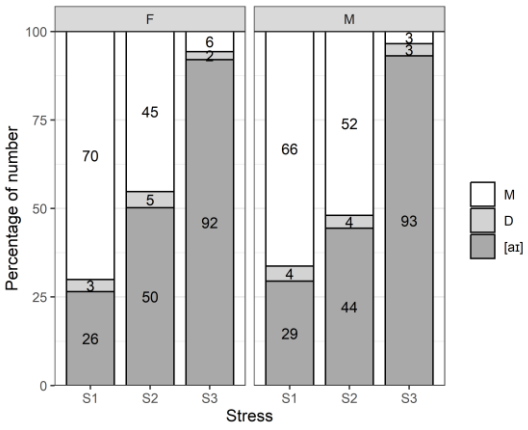


Figure 1: Proportions of neutralised monophthongs (M), neutralised diphthongs (D), and retained /aɪ/ across stress and gender.

3.3. Duration

To investigate whether differences exist across realisations and three levels of stress on duration across groups, a linear mixed effects model (LMM) was performed. Reduction (2) × Stress (3) × Gender (2) were set as independent variables, in which Reduction was defined as whether the realisation was [aɪ] or not. Retained /aɪ/, S2, and female was set as the reference level. Table 5 summarises the result of the main effects. Results showed that reduced /aɪ/ was shorter (p < .001). Besides, S1 was significantly shorter than S2, while S3 was significantly longer than S2. No specific gender effect was found for duration. Fig. 2 presents the mean duration across the three levels of stress and realisations.

	Est.	SE	t value	Pr
Intercept	0.13	0.01	25.44	.000***
Reduction	-0.03	0.00	-15.15	.000***
Stress1	-0.02	0.01	-4.86	.000***
Stress3	0.03	0.00	6.83	.000***
GenderM	0.00	0.01	0.50	.615

Table 5: The results of LMM on duration.

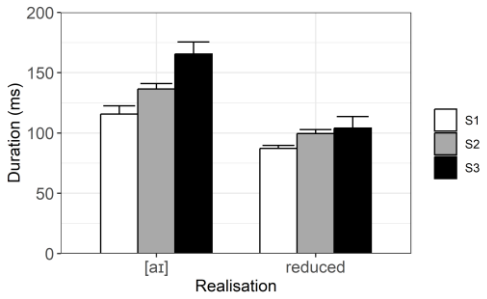


Figure 2: Mean duration across realisations and stress.

3.4. Spectral results

To examine the effect of stress on formant frequencies, F1 and F2 trajectories of retained [aɪ] in S1, S2, and S3 were evaluated. The package lme4 for linear mixed effects models in R was fitted [18]. Stress and Time (t) were the independent variables, and the F1 and F2 values of males and females were the dependent variables. In the analysis, the reference category for Stress was S2.

For both females and males, the interaction of Stress3 × Time and two main effects Stress3 and Time were found in F1. Specifically, S3 was significantly higher than S2 at the beginning; however, as the time went by, S3 went significantly more downwards than S2. Additionally, females had lower and flatter F1 for S1. Tables 6 and 7 present the LMM results of F1 for both genders. On the other hand, results of F2 in Tables 8 and 9 show a main effect Stress1 for females and a Stress3 × Time interaction effect for both males and females. Specifically, S3 went higher than S2 as time went by. Fig. 3. presents the result of formant frequencies across three levels of stress.

	Est.	SE	t value	Pr
Intercept	872.35	30.57	28.54	.000***
Stress1	-50.43	11.22	-4.50	.000***
Stress3	78.78	9.97	7.90	.000***
Time	-38.64	0.66	-58.36	.000***
Stress1:t	8.17	2.10	3.89	.000***
Stress3:t	-15.00	1.87	-8.04	.000***

Table 6: Results of female F1.

	Est.	SE	t value	Pr
Intercept	759.62	22.07	34.41	.000***
Stress1	7.00	7.59	0.92	.356
Stress3	30.09	12.09	2.49	.013*
Time	-29.00	0.53	-54.75	.000***
Stress1:t	0.71	1.41	0.50	.617
Stress3:t	-8.40	2.26	-3.71	.000***

Table 7: Results of male F1.

	Est.	SE	t value	Pr
Intercept	1801.10	15.31	117.63	.000***
Stress1	-34.81	15.43	-2.26	.002*
Stress3	-25.83	13.72	-1.88	.006
Time	50.81	0.91	55.76	.000***
Stress1:t	-2.37	2.89	-0.82	.411
Stress3:t	23.77	2.57	9.26	.000***

Table 8: Results of female F2.

	Est.	SE	t value	Pr
Intercept	1482.03	47.20	31.40	.000***
Stress1	-2.74	12.24	-0.22	.823
Stress3	-27.83	19.51	-1.43	.154
Time	47.78	0.85	55.91	.000***
Stress1:t	-3.60	2.28	-1.58	.114
Stress3:t	23.85	3.65	6.53	.000***

Table 9: Results of male F2.

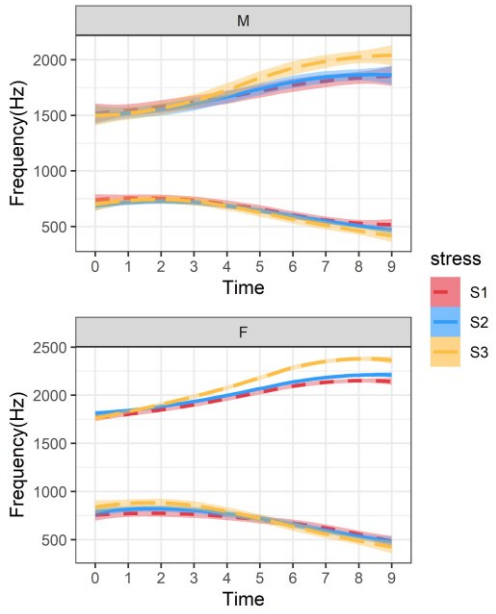


Figure 3: Spectral results of retained [aɪ] across S2 and S3.

4. DISCUSSION

In this study, we found that the diphthong /aɪ/ is prone to be reduced to monophthongs in spontaneous speech, and reduction is mainly caused by target merging, not target drop. Besides, with larger data, we observed a higher monophthongization rate than what was indicated in Su [12]. The difference lies in the fact that we looked at all tokens of /aɪ/ rather than the first 100 occurrences in the speech. Also, our speakers talked for 30 minutes instead of 3 minutes. It is very likely that our result differed from Su because speakers become more relaxed in longer monologues.

Moreover, we found that vowel reduction was correlated with tonal reduction. The effect of stress was manifested in the reduction rate, duration, and formant frequencies. In general, S2 and S1 have higher reduction rates and shorter duration, while in formant frequencies, S2 has more centralised spectral results compared with S3. Our result of unstressed vowels having shorter duration is similar to the result of lexical stress discussed in English, Swedish, and Dutch [1, 2, 3]. Besides, we found that the shorter duration was not only manifested between retained and reduced /aɪ/ but also observed across the three levels of stress within retained /aɪ/ and reduced /aɪ/. Retained /aɪ/ is longer than reduced /aɪ/. S3 is the longest, while S2 is second, and S1 is the shortest.

In spectral results, both F1 and F2 of the retained /aɪ/ were affected by stress. In the result of F1, S2 was lower at the beginning and higher in the later part compared with S3. In other words, in unstressed [aɪ], the tongue position for [a] was higher and [ɪ] was lower. The effect was further highlighted in females' S1, in which the tongue position for [a] was even higher and that for [ɪ] much lower compared with their tokens in S2.

On the other hand, the result of F2 shows that S2 went lower than S3 in the later part of the trajectory, which means that [ɪ] was less fronted when unstressed. Our result thus followed the pattern observed in the reduction of monophthongs [1, 2, 3, 6], in which unstressed vowels were more likely to be centralised.

Interestingly, gender difference was not observed either in terms of reduction rate or duration. Our result differed from what was found in French in that males had more reduced vowel space [6]. However, we only examined /aɪ/ and its reduction in this study. Therefore, it can be inferred that the gender difference in reduction might be manifested in the overall vowel space rather than the mere reduction of /aɪ/. Moreover, since we found that males produce nonstandard forms as frequently as females, it is likely that the reduction of /aɪ/ did not have a negative connotation [14].

5. CONCLUSION

In this study, we examined the reduction of /aɪ/ in terms of possible realisations, reduction rate, duration, and formant frequencies. The diphthong /aɪ/ was more reduced as two vowel targets merged and reduction does not differ across genders. Also, we found that the prosodic prominence in Mandarin has an effect on the /aɪ/ reduction. The effect was not only observed in reduction rate, but also in the durational and spectral results of retained /aɪ/. In sum, diphthongs realised with more reduced tonal shapes tended to have shorter duration and more centralised spectral results.

6. REFERENCES

[1] M. Fourakis, "Tempo, stress, and vowel reduction in American English," *The Journal of the Acoustical society of America,* vol. 90, no. 4, pp. 1816-1827, 1991.

[2] D. R. Van Bergem, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class," *Speech communication,* vol. 12, no. 1, pp. 1-23, 1993.

[3] B. Lindblom, "Spectrographic study of vowel reduction," *The journal of the Acoustical society of America,* vol. 35, no. 11, pp. 1773-1781, 1963.

[4] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech & Language,* vol. 22, no. 2, pp. 171-184, 2008.

[5] D. J. Hirst, A. Di Cristo, and Y. Nishinuma, "Prosodic parameters of French: A cross-language approach," *Contrastive studies of Japanese and other languages series,* pp. 7-20, 2001.

[6] C. Meunier and R. Espesser, "Vowel reduction in conversational speech in French: The role of lexical factors," *Journal of Phonetics,* vol. 39, no. 3, pp. 271-278, 2011.

[7] C. N. Li and S. A. Thompson, "Chinese," in *The world's major languages*: Routledge, 2003, pp. 811-833.

[8] S.-h. Peng, M. K. Chan, C.-y. Tseng, T. Huang, O. J. Lee, and M. E. Beckman, "Towards a Pan-Mandarin system for prosodic transcription," *Prosodic typology: The phonology of intonation and phrasing,* pp. 230-270, 2005.

[9] C. Gussenhoven and H. Jacobs, *Understanding phonology*. Routledge, 2017.

[10] I. Lehiste and G. E. Peterson, "Transitions, glides, and diphthongs," *The journal of The acoustical society of America,* vol. 33, no. 3, pp. 268-277, 1961.

[11] M. Lindau, K. Norlin, and J.-O. Svantesson, "Some cross-linguistic differences in diphthongs," *Journal of the International Phonetic Association,* vol. 20, no. 1, pp. 10-14, 1990.

[12] T.-t. Su, "Using the same methodology to compare reduction and assimilation phenomena in spontaneous French and Taiwanese Mandarin," in *Proceedings of the 15-th International Congress of Phonetic Sciences, Barcelona, Spain*, 2003.

[13] S.-C. Tseng, "Spontaneous Mandarin production: results of a corpus-based study," in *2004 International Symposium on Chinese Spoken Language Processing*, 2004: IEEE, pp. 29-32.

[14] W. Labov, "The intersection of sex and social class in the course of linguistic change," *Language variation and change,* vol. 2, no. 2, pp. 205-254, 1990.

[15] J. Fon, "A preliminary construction of Taiwan Southern Min spontaneous speech corpus," National Science Council technical report [NSC-92-2411-H-003-050-], 2004.

[16] Y.-Y. Chuang and J. Fon, "The effect of prosodic prominence on the realizations of voiceless dental and retroflex sibilants in Taiwan Mandarin spontaneous speech," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 6.1. 51)[Computer software]," ed, 2021.

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823,* 2014.

PROSODIC BOUNDARY STRENGTH AND PHRASE-INITIAL SYLLABLE DURATION

Gerrit Kentner[1,2], Isabelle Franz[2,3], Christine A. Knoop[2], Winfried Menninghaus[2]

[1]Goethe University Frankfurt, [2]Max-Planck Institute for Empirical Aesthetics, [3]HS Gesundheit Bochum
kentner@lingua.uni-frankfurt.de, isabelle.franz@hs-gesundheit.de, christine.knoop@ae.mpg.de, w.m@ae.mpg.de

ABSTRACT

Research on speech timing has produced conflicting results regarding the duration of phrase-initial syllables. The articulation of these syllables has been shown to be subject to *initial strengthening*, suggesting that phrase-initial segments (and by transitivity: phrase-initial syllables) have longer durations than comparable segments in other positions. In contrast, the notion of *anacrusis* holds that phrase-initial syllables are shortened, at least when these syllables are non-prominent. We explore a large sample of German literary prose texts read aloud, with a focus on the duration of phrase-initial syllables. Comparing five predicted levels of boundary strength (level 0: no break; level 1: simple phrase break; level 2: short comma phrase; level 3: long comma phrase; level 4: sentence boundary), we found evidence for initial strengthening and for a higher degree of strengthening after stronger boundaries. In contrast, we don't find systematic evidence for anacrusis.

Keywords: syllable duration; prosodic boundary; anacrusis; initial strengthening.

1. INTRODUCTION

Speech streams are divided into larger constituents (e.g., utterances, sentences) which contain smaller constituents (phrases within sentences, words within phrases, syllables within words). This hierarchical prosodic grouping of constituents underlying the sequence of speech sounds manifests itself in the rendition of prosodic boundaries, with stronger boundaries set between the higher-order constituents, and weaker boundaries set between the smaller embedded constituents. The distribution and varying strength of the prosodic boundaries in spoken texts reflect both coherence and division between speech chunks, and give rise to the temporal rhythm of the text. Prosodic boundaries are marked by a variety of cues [1-3]: The most obvious phonetic expression of a prosodic boundary is a pause between two chunks of speech. Strong boundaries are generally likely to be realized with a pause, and pauses at strong boundaries tend to be longer than pauses at weak boundaries [4]. However, not all prosodic boundaries result in a pause. Apart from pauses, prosodic boundaries are generally marked by a decrease in loudness, by boundary tones (pronounced excursions of the pitch contour preceding a boundary), and by a slow-down in speech rate approaching the boundary (pre-boundary lengthening) [5] (but see [6] for evidence of pre-boundary shortening at stronger prosodic breaks).

The effect of prosodic boundaries on the duration of the following syllable(s) is disputed: The π gesture model holds that phrase-initial segments are subject to the same "prosodically induced local slowing" of articulatory movements as pre-boundary syllables [7]. This slow-down is predicted to be greater the greater the boundary strength is. Other researchers report an articulatory strengthening of phrase-initial segments that manifests itself in an increase in the contact area between active and passive articulators [8]. Both slowing and strengthening (while conceptually quite different) result in a durational increase which – according to most researchers – mainly affects the initial segment, but has been shown to extend to the entire syllable [9].

In contrast to these findings, other researchers suggest a speed-up on pre-ictic phrase-initial syllables in English, i.e., relatively short duration of phrase-initial syllables when these syllables are unaccented. This phenomenon is termed *anacrusis* [10]. We are not aware of studies probing effects of boundary strength on anacrusis. Whereas *initial strengthening* is said to merely affect the initial post-boundary segments, the domain of anacrusis is wider, as it involves the entire pre-ictic stretch, i.e., potentially more than a single syllable. The conflicting results and the different domains motivate a fresh look at how boundary strength and prominence affect the duration of phrase-initial syllables. Here, we explore a large corpus of read aloud German prose texts to shed light on this issue.

2. EXPERIMENT: METHODS

We applied the coding manual by [11] which – on the basis of punctuation and certain syntactic features of written texts – guides the annotation of several degrees of prosodic boundaries. Here, we apply this

tool, together with a corpus of read aloud prose texts, to address the following questions:

- What are the effects of prosodic boundary strength on the duration of post-boundary, phrase-initial syllables?
- What effect does the presence/absence of a pause and pause duration have on post-boundary syllable duration?
- What effect does the degree of prominence of the phrase-initial syllable have on post-boundary syllable duration, and does it interact with boundary strength?

2.1. Recordings, annotation of data

We evaluated recordings of eight professional speakers (four male, four female). At the time of the recordings, all speakers were enrolled in a university program for rhetoric and professional stage reading.

2.1.1. Recordings

Each speaker was seated in a sound-attenuated recording booth with a Neumann U87 Ai Studio microphone placed ~30 cm from their mouth, and read aloud four prose text samples (excerpts from novels by Fontane, Goethe, Kafka, and Kleist). All speakers were familiar with the texts and had diligently prepared the reading before the recordings took place. The excerpts contained roughly ~1500-1800 words each. When slips of the tongue or disfluencies occurred during the recordings, speakers were asked to re-read the affected sentence or paragraph. The faulty segments were later erased and replaced by the corrected renditions, with the original pauses that preceded the affected sentences preserved. In total, the recordings comprised 52697 spoken words in roughly 6 hrs of speech.

2.1.2. Prosodic boundary and prominence prediction

We predicted prosodic boundaries using a coding system for the annotation of prominence and phrasing of written prose texts [11]. For the transitions between words, this system assumes four degrees of prosodic boundary strength: the strongest prosodic boundary stipulated in this annotation protocol is a sentence boundary (level 4, determined by sentence-final punctuation marks). Commas demarcate weaker breaks: Commas preceded by three or more words are assigned level 3 (long comma phrase), whereas short comma phrases (with only one or two preceding words) are assigned level 2 breaks. Level 1 boundaries are word transitions that do not come with punctuation marks but still afford prosodic boundaries due to the syntactic configuration and/or to split up exceedingly long constituents. All other word transitions were assumed to be not conducive to a prosodic break and therefore considered neutral (level 0).

Apart from boundary annotation, the coding system is used for the annotation of predicted syllable prominence. The coding system assumes, apart from further minor distinctions, the following prominence levels: schwa syllables, unstressed syllables with full vowels, lexical stress, phrasal/sentence accent.

2.2. Measurements

The recordings were automatically segmented for words and syllables using Web-MAUS [12]. Using praat [13], we extracted duration values for each annotated segment in the text, i.e., for syllables as well as pauses. In order to correlate post-boundary syllable duration with boundary strength, all word-initial syllables were coded according to the strength of the preceding boundary.

3. RESULTS

In order to evaluate potential effects of initial strengthening and/or anacrusis, we examined the duration of all word-initial syllables and compared those following a neutral word-boundary with syllables following various degrees of prosodic boundaries. In doing so, we consider potential effects of pausing at the boundary, as well as prominence and segmental content of the syllable.

The plot in Figure 1 shows post-boundary syllable durations broken down by boundary strength (level 0 to level 4) and presence/absence of a pause (top panel: no pause at break; bottom panel: post-pausal syllables). This plot does not suggest a systematic increase in syllable duration along the boundary strength scale. Instead, for the phrase-initial syllables not preceded by a pause (top panel), there is a noticeable decrease in duration for boundary strength level 1 (minor phrase break) when compared to neutral boundaries. Only for level 4 breaks, there is an evident durational increase relative to neutral word boundaries. Hardly any change in syllable duration along the boundary strength scale is detectable for post-pausal syllables.

The plot in Figure 2 shows post-boundary syllable durations broken down by boundary strength and prominence (top panel: unaccented syllables, bottom panel: accented syllables). This plot reveals a stepwise increase in syllable duration along the boundary strength scale for unaccented syllables (top panel Figure 2). For accented syllables, the increase appears to be less marked.

To model the effects of boundary strength, pausing and prominence on syllable duration, we fitted a generalised mixed effects model, assuming a gamma distribution with log link to account for the non-normal distribution of syllable duration. Along with the main effects *boundary strength*, *pausing*, and *prominence*, we considered effects of the *number of segments per syllable*, and the interactions between i. *pause duration and boundary strength*, and ii. *prominence and boundary strength*.

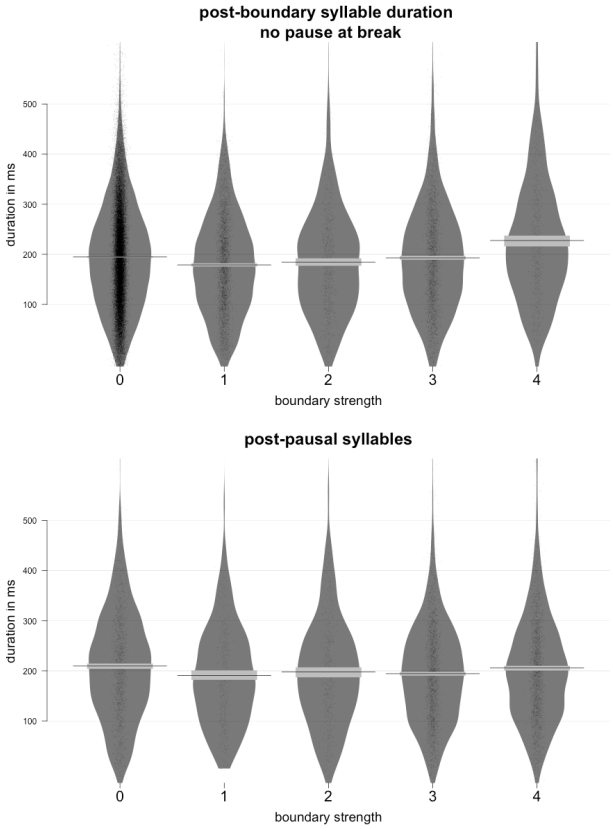


Figure 1: Duration of word-initial syllables broken down by boundary strength (0: neutral word boundary; 1: minor phrase break; 2: small comma phrase; 3: long comma phrase; 4: sentence boundary) and presence/absence of a pause preceding the break (top panel: no pause at break; bottom panel: syllables preceded by a pause).

The model confirms significant effects for boundary strength levels 2-4 on syllable duration (no significant effect of boundary level 1); significant main effects of prominence and number of segments per syllable; a significant negative interaction between pause duration and boundary strength (the effect of boundary strength on post-boundary syllable duration is weakened for post-pausal syllables); a significant negative interaction between boundary strength and prominence (the effect of boundary strength on post-boundary syllable duration is weakened for accented syllables). The coefficients of interest are tabulated in Table 1.

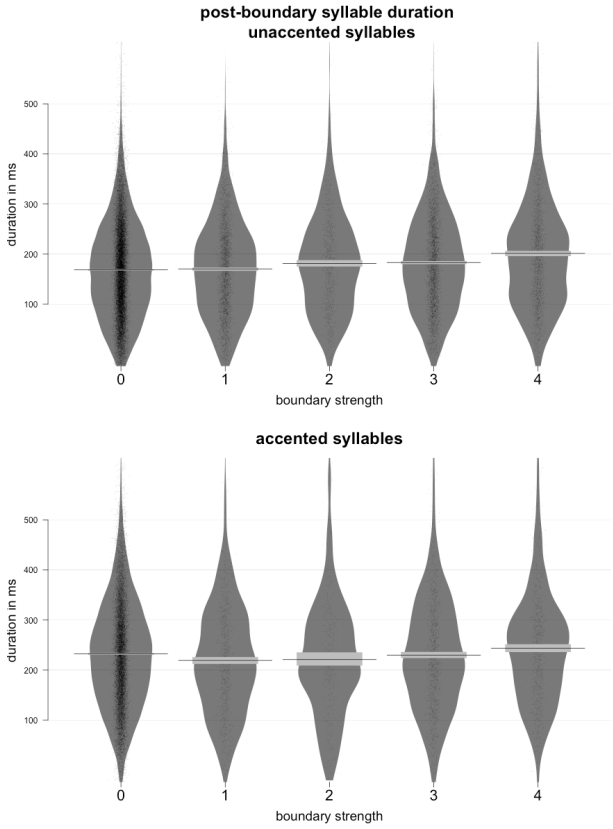


Figure 2: Duration of word-initial syllables broken down by boundary strength and prominence (top panel: unaccented syllables; bottom panel: accented syllables).

Coefficient	Estimate	SE	t-value
Boundary 1	0.007	0.005	1.2
Boundary 2	0.058	0.01	5.7***
Boundary 3	0.062	0.006	10.6***
Boundary 4	0.182	0.01	17.6***
Pause	0.005	0.006	0.9
Accent	0.181	0.004	51.3***
Segments/Syll	0.286	0.002	163.3***
Bndry : Pause	-1.3e-5	2.4e-6	-5.3***
Bndry : Accent	-0.031	0.003	-11.4***

Table 1: Model coefficients of interest (see main text for details).

In sum, whereas inspection of the raw data suggests shortening of phrase-initial syllables in some contexts (level 1 boundaries, see, e.g., upper panel of Figure 1), the statistical model does not corroborate this finding. There is especially no evidence for a shortening of pre-icitic phrase-initial syllables, as predicted in [10] for English. Instead, there is considerable evidence for initial strengthening.

4. DISCUSSION AND CONCLUSION

The results of the analysis reveal longer durations on phrase-initial syllables, providing evidence for initial strengthening. In addition, syllable duration correlates with boundary strength: sentence-initial syllables are longest on average; specifically, they are longer than comma-phrase-initial syllables, which in turn are longer than non-boundary related word-initial syllables. Furthermore, initial syllable duration is subject to an interaction between boundary strength and preceding pause duration, with the generally longer post-pausal syllables showing only little additional effects of boundary strength. Similarly, prominent syllables, which are inherently longer than the non-prominent ones, show little if any additional lengthening along the boundary strength scale. The clearest strengthening/lengthening effect, with increased lengthening along the boundary strength scale, is visible on unaccented syllables (Figure 2, top panel). This latter effect is in conflict with the concept of anacrusis which holds that unaccented or pre-ictic syllables are subject to shortening rather than lengthening.

Finally, we note a number of limitations: First, the current analysis does not provide information about effects on the level of individual segments; the corpus data lacks the control over the segmental content of the syllables. Also, we only investigate acoustic data – the articulatory dynamics leading to the increased duration of phrase-initial syllables is beyond the scope of this study. Finally, the lack of evidence in favour of anacrusis in German should not be taken to invalidate effects of anacrusis found in certain contexts in other languages.

5. REFERENCES

[1] Féry, C. (2017). *Intonation and prosodic structure*. Cambridge University Press.

[2] Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

[3] Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

[4] Gee, J. P., Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15(4), 411-458.

[5] Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B*, *369*(1658), 20130397.

[6] Kentner, G., Franz, I., Knoop, C.A., Menninghaus, W. (2023). The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase. *Journal of Phonetics*, 97, 101225.

[7] Byrd, D., Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2), 149–180.

[8] Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, 10(3), 120–141.

[9] de Souza, R. F. N. (2019). *The Interaction of Domain-initial Effects with Lexical Stress: Acoustic Data from English, Spanish, and Portuguese*, Ph.D. dissertation, The University of New Mexico, USA.

[10] Jassem, W., Hill, D., Witten, I. (1984). Isochrony in English speech: its statistical validity and linguistic relevance. In: Gibbon, D., Richter, H. (eds.), *Intonation, Accent and Rhythm*. Berlin: De Gruyter, 203-225.

[11] Franz, I., Knoop, C. A., Kentner, G., Rothbart, S., Kegel, V., Vasilieva, J., Methner, S., Scharinger, M., Menninghaus, W. (2022). Prosodic phrasing and syllable prominence in spoken prose: A validated coding manual. *OSF Preprints*. https://osf.io/h4sd5.

[12] Kisler, T., Reichel, U., Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326-347.

[13] Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* v 6.1. 21.

PROSODY OF CORRECTIVE FOCUS IN JAPANESE COMPLEX DPs

Le Xuan Chan[1], Rina Furusawa[2], Rin Tsujita[3], Seunghun J. Lee[4]

[1,2,3,4] International Christian University, [1] National University of Singapore, [4] IIT Guwahati
[1] lxlinguistics@gmail.com, [2] furusawar72@gmail.com, [3] applepie846.clock@gmail.com, [4] seunghun@icu.ac.jp

ABSTRACT

This paper reports the focus realization of internal elements within Japanese complex DPs with two modifying adjectives and a head noun, building on previous work on focus by varying modifier type, focus position, and accent patterns. Six native speakers of Japanese were recorded producing accented and unaccented complex DPs with focus on either adjectival modifiers or the head noun. Overall, prosodic prominence was given to focused elements within each DP and is accompanied by pre-focal F0 lowering, though this prominence is realized as an F0 expansion in accented phrases but as an initial F0 rise in unaccented phrases. Accentuation (realization of unaccented words with a pitch fall accent) was also observed in unaccented adjectives. Though this is subject to speaker variation and more data is needed to investigate its correlation with focus, we propose that the accentuation of focused adjectives lends support to the prosodic prominence of focus in Japanese.

Keywords: Japanese, complex DP, prosody, focus, accent change

1. INTRODUCTION

This paper examines the prosody of elements inside a complex DP that has two adjectives modifying the head noun in sentences with two accent types: all accented or all unaccented words.

Focused words in Japanese display f0 expansion, i.e. realization with a higher f0. This focus prosody of Japanese has been well reported in statements and questions of Tokyo Japanese and other dialects (cf. [1]–[6]). For accented phrases, [1] and [4] show evidence of such expansion accompanied by *post*-focal compression in the context of wh-questions and complex DPs with nominal modifiers. Both studies also concur that focus does not block downstep in accented phrases, and hence does not result in phonological rephrasing. [6] also reports a similar F0 expansion, and provides evidence of *pre*-focal lowering in elements preceding focus. Focus prosody, however, has also shown to be variable, based on [7], [8]'s study on telephone numbers recordings.

Lesser work has been done on focus in unaccented phrases. [9] states that like accented phrases, focus in unaccented phrases is manifested as an F0 expansion, providing evidence of MaxF0. However, this claim needs revisiting as F0 values were not standardized between male and female speakers, and further explanation on how this F0 expansion occurs is not provided – If unaccented words are defined by their lack of an F0 peak, how does this F0 expansion take place?

As such, this paper investigates i) whether previous findings of focus in accented phrases also apply to complex DPs with adjectival modifiers, ii) the presence of downstep even in focused words, and iii) how focus is realized phonetically and prosodically in unaccented phrases.

Additionally, we also report variations of underlying accent patterns, specifically the accentuation of unaccented adjectives, which emerged in our data, lending support to the observations in [10] of how unaccented adjectives such as *akai* 'red' can be realized with a falling accent depending on environment and context.

2. METHODS

2.1. Stimuli

The data used for analysis in this study is part of a larger set of data investigating the prosody of sentences with varying accent types and focus position. Each target DP consists of two modifier adjectives (A1/A2) and a head noun (N) in that order, which are either all accented or all unaccented.

	Adj. Modifier		Noun	
Accented	*ooki↓i*	'big'	*me↓gane*	'glasses'
	shiro↓i	'white'	*ni↓motsu*	'luggage'
Unaccented	*omoi*	'heavy'	*iruka*	'dolphin'
	marui	'round'	*mogura*	'mole'

Table 1: Adjectival modifiers and nouns by accent (falling accent marked by a downward arrow ↓).

Focus was varied across morphemes with four conditions: A1-focus, A2-focus, Noun-focus, and a neutral no-focus condition elicited for control. To elicit corrective focus, the target DPs were placed in carrier sentences following a negated DP, as in (1).

(1) *Gakkoo-de, ookii kuroi megane-janakute,*
ookii SHIROI megane *dake hakkiri mieta*
'At school, I clearly saw the **big WHITE glasses** only**,** not the big black glasses.'

No-focus conditions were elicited in the same carrier sentence with the negated DPs omitted. The data consisted of 8 types (4 conditions x 2 accent patterns). Four sentences were constructed for each condition, yielding 32 items.

2.2. Data Collection and Analysis

Speech data was elicited from four male and two female speakers in their early twenties. All participants were speakers of Kanto Japanese and were born and raised in Tokyo.

The stimuli were randomized and advanced sequentially as slides on a computer screen. Visual illustrations were displayed alongside each sentence to ensure that the participants understood each focus condition (Fig. 1). Where there was a negation clause preceding the target clause, speakers were instructed to emphasize the difference between the two clauses. After silently reading each sentence, participants produced each phrase as a reply to an audio prompt asking *nani ga mieta?* 'What did you see?' as if conversing with a friend. Practice trials were conducted and breaks were included to mitigate any fatigue effect throughout the task.

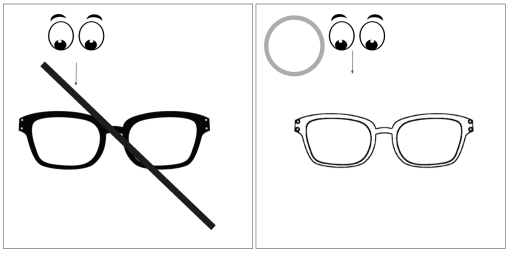


Figure 1: Example of visual A2-focus stimuli as per (1).

All speech data was recorded with a digital recorder Tascam DR-100 MK-III, at 44.1 kHz with 16-bit, attached to head-worn SHURE WH30 unidirectional microphone.

Out of 384 elicited tokens (32 items x 2 repetitions x 6 speakers), 288 were included in the final analysis as corrective focus was not produced by two speakers, JPN004 and 008 (i.e. identical realizations regardless of focus condition). Following manual segmentation in Praat [11], MaxF0, MinF0, and F0 range (MaxF0 – MinF0) from each morpheme in the DP were collected and Z-standardized by speaker, leading to 864 data points (288 tokens x 3 morphemes). Linear mixed-effect models were run with each of these measurements as dependent variables using the lmerTest [12] package in R [13]. Focus condition and morpheme were inserted as fixed effects, while speaker, word, and repetition were inserted as random effects. Accented and unaccented tokens were tested separately, and unaccented tokens showing accentuation (n=91) were excluded from the model (discussed in 3.4).

3. RESULTS

Overall, focused elements are realized with prosodic prominence. The phonetic realization of this prominence, however, differs by accent type. The prosodic prominence assigned to the focus element in both accented and unaccented cases are also accompanied by pre-focal lowering of the preceding element. In no-focus conditions, prosodic prominence is assigned to the leftmost morpheme. We also observed unexpected accentuation of unaccented adjectives in our data.

3.1. Focus in sentences with all accented words

In accented tokens, focus was realized phonetically as an F0 peak expansion. In no-focus conditions, a string of accented words is realized as one phonological phrase, where prominence is given to the DP-initial element and downstep occurs across the other morphemes. In focus conditions, the focused element receives an F0 expansion, rendering F0 values similar to or exceeding the phrase-initial F0 (Fig. 2).

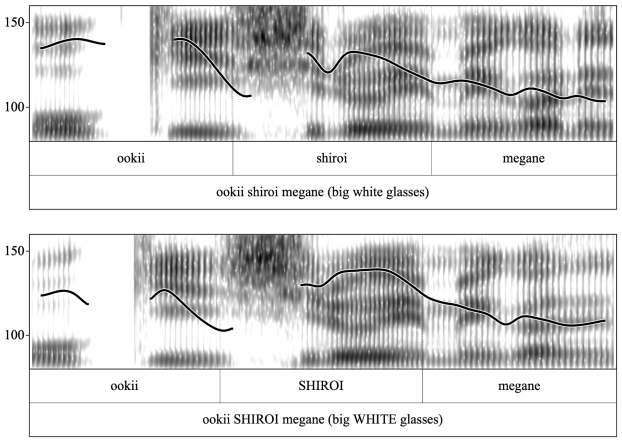


Figure 2: No-focus (top) and A2-focus (bottom) conditions in accented phrases by JPN005.

To measure this, we looked at MaxF0, and found a significant effect of focus on the MaxF0 of each morpheme in the DP ($F=29.83$, $p<.001$). A Tukey pairwise comparison revealed that each focused element was realized with a significantly higher max F0 compared to their neutral counterparts in the no-focus condition (e.g. A2 in A2-focus condition vs A2 in no-focus condition). However, as prominence is assigned to the leftmost element in the no-focus

condition, no significant difference was found between the no-focus and A1-focus conditions.

Morpheme	Mean MaxF0 (Z-score)		df	t	p
	no focus	focus			
A1	0.989	1.08	129	-0.54	0.951
A2	-0.0697	0.65	139	-7.52	<.001
N	-0.524	0.17	136	-6.87	<.001

Table 2: Max F0 (Z-score) by focus condition (accented).

Though MaxF0 differs between focus and no-focus conditions, however, the pattern of decreasing F0 values correlating with morpheme position is still present in the focus conditions, indicating downstep. Even though F0 expansion takes place at each focused morpheme, it is still subject to downstep based on its position in the DP – MaxF0 is highest on the A1 (M=1.47), followed by the A2 (M=0.65), then the N (M=0.17). If focus were independent of downstep, we would expect similar MaxF0 values across all focused morphemes regardless of position. As this is not the case, our data supports the claims made in [1] and [4].

3.2. Focus in sentences with all unaccented words

In unaccented tokens, focus was realized phonetically as a rising F0 contour, indicating the initial rise of a phonological phrase. In neutral conditions, a string of unaccented words is realized as one phonological phrase, where F0 rises at the phrase-initial element and plateaus across the following morphemes. In focus conditions, however, the F0 rise of the phonological phrase is shifted to the focused element (Fig. 3).

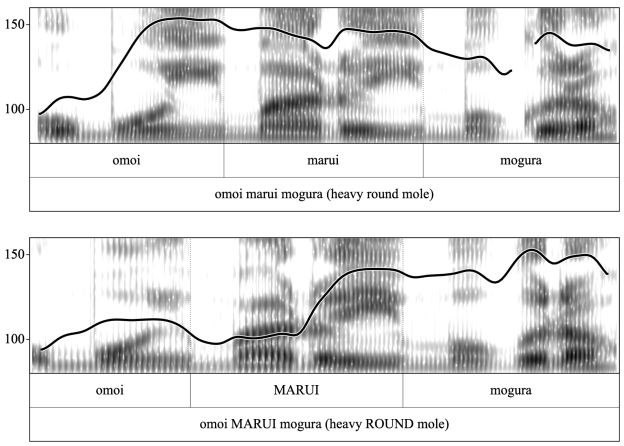


Figure 3: No-focus (top) and A2-focus (bottom) in unaccented phrases by JPN007.

To measure this, we compared the F0 range (difference between the maximum and minimum F0) within each element across focus conditions. A large difference indicates a rising F0 contour, while a smaller difference indicates the lack of F0 rise, i.e. an F0 plateau. As with the accented tokens, the model returned a significant effect of focus on the F0 range of each morpheme (F=12.03, p<.001). Tukey comparisons of focus and no-focus conditions revealed that F0 range was significantly higher for focused tokens, except in A1-focus (Table 3).

Morpheme	Mean F0 range (Z-score)		df	t	p
	no focus	focus			
A1	0.34	0.07	75.7	1.60	0.357
A2	-0.45	0.17	72.1	-4.95	<.001
N	-0.56	0.55	75.8	-6.54	<.001

Table 3: F0 range (Z-score) by focus condition (accented).

3.3. Pre-focal lowering

F0 lowering of the morpheme preceding focus was also observed in both accented and unaccented conditions, supporting [6]. As per Tables 4 and 5, the morpheme preceding the focused morpheme is realized with a significantly lower maximum F0 than their counterparts in the no-focus neutral condition. Here, 'preceding focus' refers to the conditions where the following morpheme is focused – MaxF0 of A1 in the A2-focus condition, and MaxF0 of A2 in the Noun-focus condition.

Morpheme	Mean MaxF0 (Z-score)		df	t	p
	neutral	preceding focus			
A1	0.99	-0.02	127	4.82	<.001
A2	-0.07	-0.59	139	3.54	<.001

Table 4: Max F0 (Z-score) of neutral and preceding focus conditions (accented).

Morpheme	Mean MaxF0 (Z-score)		df	t	p
	neutral	preceding focus			
A1	0.33	-0.33	79.1	5.23	<.001
A2	-0.22	-0.69	76.7	4.99	<.001

Table 5: Max F0 (Z-score) of neutral and preceding focus conditions (unaccented).

3.4. Accentuation of unaccented adjectives

In unaccented tokens, we also observed an unexpected realization of unaccented adjectival modifiers with a falling accent, which we refer to as accentuation. As in Fig. 4, the unaccented adjectives *omoi* 'heavy' and *marui* 'round' are realized with a HL falling pitch contour.

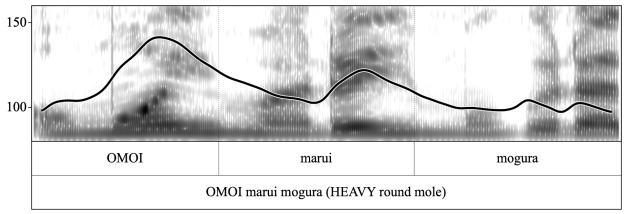


Figure 4: Accentuation of unaccented adjectives in A1-focus condition by JPN009.

This pattern of accentuation, however, varied greatly by speakers – 3 speakers (JPN004, 006, 007) had almost no tokens of accentuation, while the remaining 3 speakers showed varying frequencies and patterns of accentuation: only A1, only A2, or both (Table 6). The presence of accentuation in each morpheme was coded by linguistically trained research assistants, and any inter-rater discrepancies were cross-checked and resolved.

Speaker	Accentuation (no. of tokens)				total
	No accentuation	Only A1	Only A2	Both A1 & A2	
005	4	24	1	3	32
008	1	0	0	31	32
009	13	4	3	12	32

Table 6: Accentuation patterns of JPN005, 008, 009.

Does focus affect accentuation? The data from speakers JPN005 and 009 seem to show a slight trend; when either A1 or A2 is focused, accentuation is more likely to occur on that morpheme (Table 7).

Focus	Accentuation (no. of tokens)				total
	No accentuation	Only A1	Only A2	Both A1 & A2	
no-focus	7	9	0	0	16
A1-focus	0	10	1	5	16
A2-focus	0	7	2	9	16
Noun-focus	10	6	1	7	16

Table 7: Accentuation patterns of JPN005, 009 by focus.

This observation lends support to the notion of prosodic prominence being assigned to the focused element – focused adjectives are more readily accentuated than unfocused adjectives. However, this argument needs further investigation as data from two speakers are insufficient to be conclusive, and as shown, speaker variation seems to be a big factor in determining accentuation.

4. DISCUSSION AND CONCLUSION

Overall, we have shown that though focus in Japanese complex DPs is defined by prosodic prominence, phonetic realizations differ by accent.

4.1. Focus and Downstep in Accented Phrases

The results of accented tokens in our study expands on the findings of [1]-[6] by showing that the principles of focus and downstep observed in DPs with nominal modifiers such as 'noun-*no* noun-*no* noun' also apply to complex DPs with adjectival modifiers. More importantly, by varying focus position, we provide more evidence for the arguments in [1] and [4] that focus does not block downstep and does not trigger a pitch reset. In other words, focused elements are still subject to downstep in an accented phrase. While previous studies such as [1] show this by comparing focus in varying accent sequences, our data provides a clearer picture of downstep taking place in focused elements by varying focus position in the same accent sequence (see Table 2).

4.2. F0 range over MaxF0 in unaccented phrases

Our findings in the unaccented sequences also provide a new way of understanding focus realization in unaccented phrases, setting it apart from accented phrases. Unlike the F0 expansion in accented phrases, focus is realized as an F0 rise that reflects the initial rise of a prosodic domain. Each unaccented phrase takes on an initial F0 rise, and this F0 rise is shifted to each focused element.

In quantifying this, we have shown that F0 range is significant in signaling focus for unaccented phrases, unlike the findings for unaccented phrases in [9]. We propose that MaxF0 is less significant for unaccented phrases as the pitch track plateaus towards the end of the DP after the focused morpheme (see Fig. 3). In this way, our findings show that accented and unaccented phrases pattern differently, indicating that accent has a direct effect on focus realization in Japanese.

4.3. Change in underlying accent pattern

Finally, the accentuation of unaccented adjectives provide evidence for the patterns reported in [10] for Tokyo Japanese. Like [10], accentuation only takes place for adjectives but not nouns, and is subject to speaker variation. Though more data is needed, we also suggest that focus and the need to place prosodic prominence on an unaccented adjective influences accentuation.

Where our data departs from [10], however, is the environment that accentuation takes place in. Though [10] reports that accentuation only takes place at the phrase-final position, our data also shows that non-phrase final adjectival modifiers are also subject to accentuation, perhaps indicating a diachronic trend of allowing accentuation in more phonological environments.

5. ACKNOWLEDGEMENT

This work was partially supported by JSPS KAKENHI Grant #20K00578 (POP), the ILCAA joint research project "Phonetic typology from cross-linguistic perspectives (PhonTyp)" (2021-2023), and the travel grant from the National University of Singapore and the Ministry of Education, Singapore. We thank Honoka Asai, Kotone Sato, Keitaro Mitsuhashi for the data collection, and Michinori Suzuki and Toshio Matsuura for comments on the stimuli.

6. REFERENCES

[1] Kubozono, H. 2007. Focus and intonation in Japanese. *Interdisciplinary studies on information structure* 9, 1–27.

[2] Kubozono, H. 2018. Focus Prosody in Kagoshima Japanese. In: Goedemans, R., Heinz, J., van der Hulst, H. (eds), *The Study of Word Stress and Accent*. Cambridge University Press, 323–345.

[3] Kubozono, H. 2021. *Ippan gengogaku kara mita Nihongo no purosodī: Kagoshima hōgen o chūshin ni (Japanese prosody from general linguistic perspectives)*. Kuroshio Shuppan.

[4] Ishihara, S. 2011. Japanese focus prosody revisited: Freeing focus from prosodic phrasing. *Lingua* 121(13), 1870–1889.

[5] Ishihara, S. 2016. Japanese downstep revisited. *Natural Language Linguistic Theory* 34(4), 1389–1443.

[6] Hwang, H.K. 2011. Distinct Types of Focus and Wh-Question Intonation. *ICPhS 17*, 922–925.

[7] Lee, Y. Nambu, S., Cho, S. 2018. Focus prosody of telephone numbers in Tokyo Japanese. *The Journal of the Acoustical Society of America* 143(5), EL340–EL346.

[8] Lee, Y., Nambu, S., Cho. S. 2019. Dataset of focus prosody in Japanese phone numbers. *Data in Brief* 25, 104–139.

[9] Lee, A., Xu, Y. 2012. Revisiting focus prosody in Japanese. *Proc. Speech Prosody* 2012, 274-277.

[10] Kobayashi, M. 2003. An Examination of Adjective Accent Change in Tokyo Japanese and Factors Influencing the Change. *Journal of the Phonetic Society of Japan* 7(2), 101–113.

[11] Boersma, P., Weenink, D. 2019. Praat: doing phonetics by computer. 2019. Accessed: Dec. 17, 2019. [Online].

[12] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26.

[13] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

REGIONAL VARIATION, ARTICULATION RATE, AND PAUSING PATTERNS IN THREE VARIETIES OF SPANISH

Lorenzo García-Amaya[a], Kelly Kendro[b], & Nicholas Henriksen[a]

[a]University of Michigan & [b]University of Utah
lgarciaa@umich.edu, kelly.kendro@utah.edu, nhenriks@umich.edu

ABSTRACT

This study investigates regional differences in articulation rate and pausing patterns among speakers of three regional varieties of Spanish. We elicited speech data from talkers of two varieties of European Spanish (North-Central Peninsular Spanish and Andalusian Spanish) and one variety of Latin American Spanish (Porteño Spanish). The European groups produced faster articulation rates than the Porteño group; male speakers also spoke faster than female speakers. Regarding pauses, the NCPS speakers produced shorter silent pauses than the Andalusian and Porteño speakers, while both European groups produced shorter filled pauses (e.g, *eh*, *em*) than the Porteño speakers. We additionally found an effect of speaking condition, with all groups demonstrating higher silent-pause rates in sociolinguistic interviews than in a reading passage. Taken together, the findings contribute to a literature demonstrating that regional varieties of Spanish robustly exhibit different patterns of elocution rate and pausing.

Keywords: Articulation rate, silent pauses, filled pauses, Spanish, regional variation

1. INTRODUCTION

Sociophonetic research documents the diverse ways in which temporal measures in a language can be modulated by regional origin, including cross-dialectal comparisons of articulation rate (AR), pausing rate, and pause duration [1-7]. In American English, for instance, southern speakers talk slower and produce longer and more frequent pauses than northern speakers [1, 4], which confirms the social stereotype about a southern "drawl." With respect to French, Schwab and Avanzi [3] show that Swiss-French speakers articulate speech at a slower rate than Belgian- and Parisian-French speakers. The findings of Verhoeven et al. [5] reveal that Dutch speakers from the Netherlands talk 16% faster than those from Belgium. Leeman's [6] analysis of crowdsourced data from nearly 3,000 speakers reveals a west-east divide in Swiss German, with eastern speakers having faster ARs.

In addition to regional origin, social factors such as age, sex, speaking style, and possibly the urban or rural origin of speakers, can significantly influence temporal measures of speech [7]. Many studies converge in showing that men speak more quickly than women [8, 9], as do younger speakers compared to older speakers [10, 11]. Previous work reports inconsistent results with regards to comparative speech rate between free conversation and prepared reading tasks [1, 12].

Pellegrino et al.'s [13] cross-language comparison shows that speakers of Spanish produce the highest number of syllables/second among the Romance languages, and that Spanish is one of the fastest of the world's languages. In addition to speech rate, the authors report that "information rate," which is calculated using speaking rate and information density (e.g., average bits per second of information), is stable across languages. However, in Pellegrino et al. a single regional variety of Spanish was included, and it remains uncertain whether Spanish can be considered a monolith with respect to temporal measures of speech. The findings from Dutch, English, French, and German motivate the need for new research ascertaining whether cross-dialectal variation is present among Spanish speakers' elocution and pausing patterns.

The body of work on speech rate and pausing differences across Spanish dialects is limited, although highly proficient speakers of the language tend to perceive European Spanish as spoken faster than most Latin American varieties [14]. This social stereotype is reflected in recent studies led by Schwab [15] and Santiago and Mairano [16]. The former demonstrated that speakers from Spain articulate speech at a rate 5% faster than those from Costa Rica, while the latter found Spaniards to articulate speech at a rate 9% faster than Mexicans. The results for pauses were less consistent between the two studies: [15] showed that Latin American speakers produce longer pauses than European speakers, while [16] did not find an effect of regional origin on pause usage.

The present study builds upon the previous findings by comparing three varieties of Spanish. We explore whether differences exist among two regional varieties of European Spanish and Porteño Spanish. The research question driving our analysis

is the following: How do extralinguistic variables such as dialect, speaker sex, and task affect rates of articulation and pausing in Spanish?

2. METHODS

2.1. Participants and recording protocol

We recorded speech data from 30 speakers, aged 18 to 25: ten each of North-Central Peninsular Spanish (NCPS), Andalusian Spanish, and Porteño Spanish. The speakers were university students born and raised in Salamanca (Spain), Jerez de la Frontera (Spain), or Buenos Aires (Argentina), respectively. Each group included five female and five male speakers (self-identified). All speakers read aloud the same two-paragraph passage and participated in a twenty-minute sociolinguistic interview with an experimenter, who in each case was a native speaker of the regional variety. All recordings were conducted in the participants' home cities.

2.2. Analysis

The speech from the first ten minutes of each sociolinguistic interview was first orthographically transcribed. We then performed acoustic labelling in Praat [17] which included: measuring the speaking time of all inter-pausal units (i.e., speaking time without pauses); measuring the duration of all silent pauses (SPs), defined as periods of silence longer than 100 milliseconds (ms); and measuring the duration of all filled pauses (FPs), such as *em*, *eh*, or *mm*. This labeling protocol resulted in 4,831 SPs and 768 FPs across all speakers and tasks.

We calculated five measures of speech tempo: ARTICULATION RATE (AR), FP RATE, SP RATE, MEAN FP DURATION, and MEAN SP DURATION. AR was calculated by dividing the total number of syllables produced in all inter-pausal intervals by the total speaking time of the inter-pausal units; AR thus excludes the production of FPs and hesitations. To enable comparisons with previous studies of Spanish regional variation [15-16], we counted the number of expected (or canonical) phonemes in the AR calculation. The one exception was that the word *para* "for" was considered monosyllabic for the Andalusian speakers but bisyllabic for the NCPS and Porteño speakers. We additionally calculated each speaker's mean FP and SP durations (in ms), as well as each speaker's FP RATE and SP RATE. For the last two measures, we divided the number of pauses by their total speaking time within the ten analyzed minutes.

We fitted linear mixed-effects models to the five outcomes using the *lmerTest* package [18] within the statistical software R [19], with the predictors DIALECT, SEX, TASK, and the interactions among them. We followed the top-down stepwise approach in the model-fitting procedure, meaning that we only report main effects if higher-order interactions are not significant [20]. All models included random intercept effects for SPEAKER; although we originally included by-SPEAKER random slopes for TASK, these were removed due to lack of model convergence. To facilitate model-based inference regarding the fixed effects, we computed marginal predicted means based on the fitted models using the R package *ggeffects* [21].

3. RESULTS

For the outcome AR, we uncovered significant effects of both DIALECT ($F(2,144) = 5.808$, $p = .004$) and SEX ($F(1, 144) = 16.500$, $p < .001$). Per Figure 1, we see that the Andalusian speakers produced the highest mean AR value, and that both European groups were faster than the Porteño speakers. Figure 2 shows that male speakers (7.0 syll/sec) additionally spoke faster than female speakers (6.6 syll/sec). Next, we uncovered a significant effect of TASK on SP RATE ($F(1,146) = 44.224$, $p < .001$), plotted in Figure 3. The participants produced more SPs in the sociolinguistic interview (22.4 SPs/min) than in the reading task (17.5 SPs/min). For FP RATE, however, none of the tested effects were significant (no figure is shown).

Turning to MEAN SP DURATION (Figure 4), DIALECT was a significant predictor ($F(2,144) = 8.904$, $p < .001$): the Andalusian and Porteño groups produced the longest SPs on average (559 ms and 562 ms), while the NCPS speakers produced statistically shorter SPs (500 ms). Regarding MEAN FP DURATION, the predictor DIALECT was again significant ($F(2,27) = 9.818$, $p < .001$). Per Figure 5, the Porteño speakers produced the longest FPs on average (472 ms), while the NCPS and Andalusian speakers produced shorter FPs (359 ms and 335 ms, respectively).

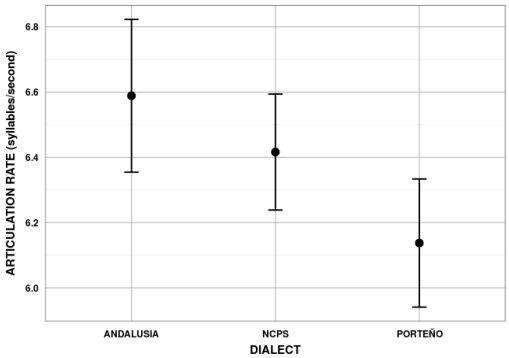


Figure 1: Effect of DIALECT on ARTICULATION RATE.

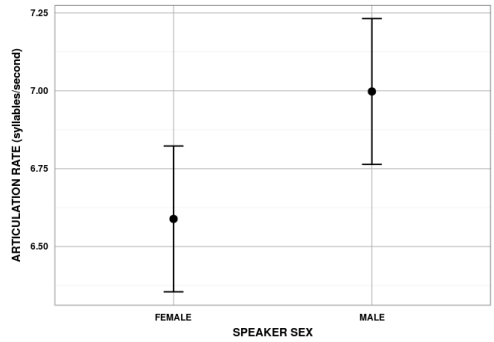


Figure 2: Effect of SEX on ARTICULATION RATE.

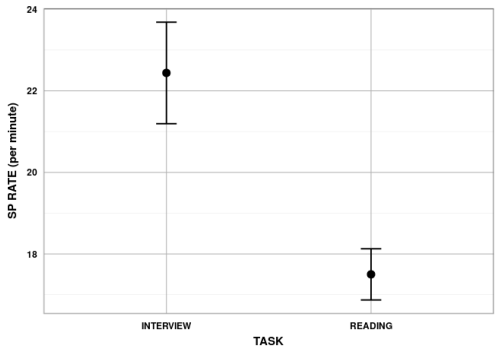


Figure 3: Effect of TASK on SP RATE.

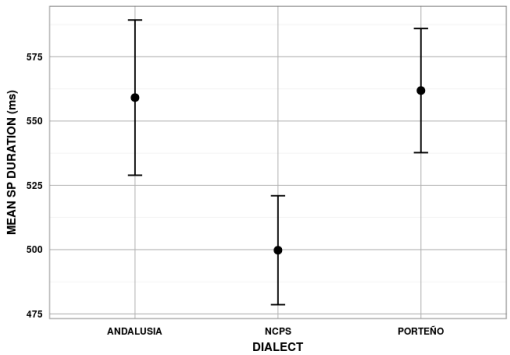


Figure 4: Effect of DIALECT on MEAN SP DURATION.

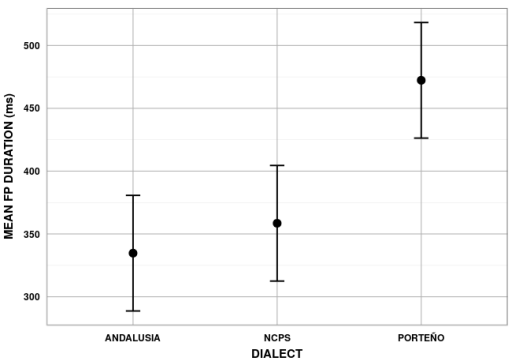


Figure 5: Effect of DIALECT on MEAN FP DURATION.

4. DISCUSSION

In response to our main research question, we found dialect, speaker sex, and task to predict elocution and pausing values in Spanish, with a speaker's dialect being the most common significant predictor. Specifically, we found that the speakers from the two European varieties produced speech using faster ARs than the Latin American speakers from Buenos Aires. Regarding the participants' pausing patterns, the NCPS speakers produced shorter SPs than both the Andalusian and Porteño speakers. The Porteño speakers additionally produced longer FPs than the two European groups (who did not differ from each other in a substantial way). The emergent finding is that the Porteño participants spoke at the slowest rate and produced the longest pauses among the three dialect groups.

These findings underscore the need to evaluate regional varieties of Spanish as separate rather than condensed into a single language (cf. 13]). Similar to Santiago and Mairano [16], while we generally confirmed the social stereotype that speakers of European Spanish have fast elocution rates, we would caution against considering Porteño Spanish as a proxy for all varieties spoken in Latin America. For example, there is a pervasive stereotype that Chilean talkers speak very quickly in comparison to other Latin American varieties [22]. However, Chilean Spanish remains under-investigated in the relevant sociophonetic literature.

It is noteworthy that the Andalusian speakers produced the fastest AR among the three groups, and it is therefore important to consider the role that local processes might play in affecting global measures of speech rate. For American English, the study by Clopper and Smiljanic [1] considered how processes of consonantal reduction contribute to the attested regional differences. For example, New England talkers deleted coda /t/ (e.g., *whi<u>t</u>e li<u>gh</u>t*) and simplified fricative-stop clusters (e.g, *lo<u>st</u>*) more commonly than talkers from other regions. Andalusian Spanish is known to lenite syllable-final /s l θ ɾ/; the deletion of such consonants in phrase-final position likely contributes to the speakers' comparably faster AR as observed in our study. Importantly, studies also demonstrate that Andalusian speakers employ compensatory-lengthening strategies upon leniting consonants in the phrase-medial position, such as post-aspirating phonological /sp st sk/ sequences (e.g., [t^h]) and geminating /ɾl ɾn/ (e.g., [ll nn]) [28]. Such compensatory lengthening strategies allow for more overt speech being produced, which may explain why the Andalusians' AR values remained close to those of the NCPS speakers.

Relatedly, the present findings prompt inquiry into why the Porteño speakers produced the slowest means for AR and the longest pause duration values. Although our study was not intended to directly answer this question — our metrics are global and do not consider the impact of local phonetic processes — [25] and [26] propose that phrase-final lengthening is present in Porteño Spanish, which leads to speakers producing long durations of vowels in the phrase-final position and, presumably, overall slower speech rates. In addition to this, Coats [4] has suggested that elocution differences in a language may reflect historical legacies of language contact. Specific to Porteño Spanish, studies have argued that contemporary prosodic patterns may result from the situation of intense language contact that arose in Argentina between immigrant languages and Spanish at the start of the twentieth century [26]. It is hence possible that the overall slower speech produced by Porteño speakers, which is attributable in part to phrase-final lengthening, was conventionalized during a period of language contact among immigrant communities in Argentina. However, additional work would be necessary to examine this hypothesis.

The findings also revealed new data with respect to cross-dialectal pausing variation. The NCPS speakers produced SPs that were on average 12% shorter than the Andalusian and Porteño speakers. More remarkably, perhaps, the Porteño speakers produced FPs that were between 30% and 35% longer than those of the European Spanish speakers. While Santiago and Mairano [16] did not find a significant effect of DIALECT on pause durations in their comparison between Spaniards and Mexicans, Schwab [15] showed that Costa Rican speakers produce longer SPs than Spaniards. To our knowledge, social stereotypes of regional variation in Spanish do not typically reference pausing patterns as they do in other languages [e.g., 1]. It is thus unclear whether listeners of Spanish are attuned to differences in FP durations along the lines of the data presented here. Note that Quené [27] showed that listeners' threshold for reliably detecting elocution differences is 10%.

This study's main effect of SEX on the speakers' AR values corroborates previous research showing that male speakers produce speech with faster elocution than female speakers [2-3, 6; cf. 1]. Here, the male speakers' mean AR values were 6% faster than those of the female speakers, which mirrors the findings of Verhoeven et al. [5]. Santiago and Mairano [16] argue that such differences are likely not attributable to physical or anatomical factors and are more likely to be rooted in sociolinguistic sources, such as the notion that female speakers are generally more careful talkers than male speakers [23]. This hyper-articulation by female speakers may result in overall slower speech.

This study found an effect of TASK on SP RATE, with speakers tending to produce more SPs in the interview than in the reading task. Combined with the Spanish-speaking participants from [16] producing longer pauses in semi-directed interviews than in a reading passage task, we suggest that speakers' tendencies to produce more and longer pauses in conversational settings may be due to the pragmatic or social functions held by pauses, such as turn-taking or floor-holding [24]. Regarding the lack of effect of TASK on AR observed here, it bears noting that some studies have reported faster AR in free conversation than in read speech [2, 3], while others report the inverse pattern [16]. Santiago and Mairano [16] suggest that these contradicting findings may be due to the use of nonstandardized elicitation methods across studies.

Finally, future research may address limitations of this study by: (i) recruiting a higher number of participants to ensure replicability; (ii) including additional varieties of Latin American Spanish, such as Chilean Spanish; and (iii) taking into account possible linguistic predictors, such as length of utterance and the AR of a preceding utterance, on elocution and pausing metrics [3].

5. CONCLUSION

The present results demonstrated cross-dialect variation with respect to AR and pausing in Spanish, consistent with those of studies from other languages [1-6]. The Andalusian dialect was characterized by the fastest AR, long SPs, and short FPs. The NCPS speakers produced AR values intermediate to the other groups, but notably produced shorter SPs than the Andalusian speakers. The Porteño participants had the overall slowest AR, as well as the highest predicted means for SP and FP duration. While the results for AR are generally consistent with social stereotypes about European Spanish speakers talking faster than Latin American speakers [24], this study revealed novel findings with respect to pausing patterns.

Additional variables such as speaker sex and task were significant in the models: male speakers produced faster AR, and all groups showed higher values for SP rate in the sociolinguistic interview than in the reading passage. Finally, future work would do well to investigate the linguistic sources of the regional differences reported here (e.g., coda-weakening processes in Andalusian Spanish, phrase-final lengthening in Porteño Spanish).

6. REFERENCES

[1] Clopper, C. G., Smiljanic, R. 2015. Regional variation in temporal organization in American English. *Journal of Phonetics, 49*, 1–15.

[2] Jacewicz, E., Fox, R. A., O'Neill, C., Salmons, J. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change*, *21*(2), 233–256.

[3] Schwab, S., Avanzi, M. 2015. Regional variation and articulation rate in French. *Journal of Phonetics*, *48*, 96–105.

[4] Coats, S. 2020. Articulation Rate in American English in a Corpus of YouTube Videos. *Language and Speech 63*(4), 799–831.

[5] Verhoeven, J., De Pauw, G., Kloots, H. 2004. Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech, 47*(3), 297–308.

[6] Leeman, A. 2017. Analyzing geospatial variation in articulation rate using crowdsourced speech data. *Journal of Linguistic Geography, 4*, 76–96.

[7] Hewlett, N., Rendall, M. 1998. Rural versus urban accent as an influence on the rate of speech. *Journal of the International Phonetic Association*, *28*(1-2), 63–71.

[8] Byrd, D. 1992. Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America, 92*(1), 593–596.

[9] Whiteside, S. 1996. Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association, 26*(1), 23–40.

[10] Duchin, S. W., Mysak, E. D. 1987. Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders*, *20*(3), 245–257.

[11] Ramig, L. A. 1983. Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders*, *16*(3), 217–226.

[12] Butcher, A. 1981. Aspects of the speech pause: Phonetic correlates and communication functions. *Arbeitsberichte Kiel*, *15*, 1–233.

[13] Pellegrino, F., Coupé, C., Marsico, E. 2011. A cross-language perspective on speech information rate. *Language*, *87*(3), 539–558.

[14] Morrison, G. S., Escudero, P. 2007. A cross-dialect comparison of Peninsula- and Peruvian-Spanish vowels. In: J. Trouvain, W. J. Barry (eds.), *Proc. 16th ICPhS.* Saarbrücken, 1505–1508.

[15] Schwab, S. 2015. Las variables temporales en el español de Costa Rica y de España: un estudio comparativo. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, *41*(1), 127-139.

[16] Santiago, F., Mairano, P. 2022. Spaniards articulate faster than Mexicans: Temporal patterns in two varieties of Spanish. *Spanish in Context 19*(2).

[17] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program].

[18] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

[19] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

[20] West, B. T., Welch, K. B., Galecki, A. T. 2014. *Linear mixed models: A practical guide using statistical software*. CRC Press.

[21] Lüdecke, (2018). Ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software, 3*(26), 772.

[22] Kendro, K., Henriksen, N., García-Amaya, L. 2022. Spanish Talker Intuition Regarding Rate of Speech Repository (STIRRS). [Online repository]. https://umich.edu/~stirrsrepo/.

[23] Trudgill, P. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, *1*(2), 179–195.

[24] Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

[25] Benet, A., Gabriel, C., Kireva, E., Pešková, A. 2012. Prosodic transfer from Italian to Spanish: Rhythmic properties of L2 speech and Argentinean Porteño. In: Ma, Q., Ding, H., Hirst, D. (eds), *Proc. 6th International Conference on Speech Prosody, 438–441.*

[26] Gabriel, C., Kireva, E. 2014. Prosodic transfer in learner and contact varieties: Speech rhythm and intonation of Buenos Aires Spanish and L2 Castilian Spanish produced by Italian native speakers. *Studies in Second Language Acquisition*, *36*(2), 257–281.

[27] Quené, H. 2004. What is the Just Noticeable Difference for tempo in speech?. *LOT Occasional Series*, *2*, 149-158.

[28] Ruch, H., & Harrington, J. (2014). Synchronic and diachronic factors in the change from pre-aspiration to post-aspiration in Andalusian Spanish. *Journal of Phonetics, 45*, 12–25.

WHAT DOES BERT LEARN ABOUT PROSODY?

Sofoklis Kakouros[a] and Johannah O'Mahony[b]

[a]University of Helsinki, [b]University of Edinburgh
sofoklis.kakouros@helsinki.fi, johannah.o'mahony@ed.ac.uk

ABSTRACT

Language models have become nearly ubiquitous in natural language processing applications achieving state-of-the-art results in many tasks including prosody. As the model design does not define predetermined linguistic targets during training but rather aims at learning generalized representations of the language, analyzing and interpreting the representations that models implicitly capture is important in bridging the gap between interpretability and model performance. Several studies have explored the linguistic information that models capture providing some insights on their representational capacity. However, the current studies have not explored whether prosody is part of the structural information of the language that models learn. In this work, we perform a series of experiments on BERT probing the representations captured at different layers. Our results show that information about prosodic prominence spans across many layers but is mostly focused in middle layers suggesting that BERT relies mostly on syntactic and semantic information.

Keywords: language model, BERT, prosody, prominence, part-of-speech

1. INTRODUCTION

Pre-trained language models (LMs) have achieved high performance on several natural language processing (NLP) tasks such as constituency parsing [1], semantic role labeling, and coreference resolution [2]. Recently, prosody and prosodic phenomena have witnessed increasing attention in natural language applications enabled by pre-trained LMs leading to improved performance [3, 4]. Despite the gains, we do not yet fully understand what enables these models to perform at this level. Their performance suggests that their learning objectives potentially teach the models details about the structure of the language. It remains, however, unclear, what specific prosodic information these models are able to implicitly capture during their pre-training. In this work, we focus on analyzing the prosodic information in one of the most widespread models, BERT (Bidirectional Encoder Representations from Transformers) [5].

Several recent works have investigated the representations learned at different layers of BERT in an attempt to interpret and understand the linguistic information captured by the model. These studies have uncovered that BERT indeed learns various aspects of the language. For instance, [6] showed that BERT layers capture a rich hierarchy of linguistic information spanning from surface features in the lower layers to syntactic in middle layers and semantic in the higher layers. These findings are further supported by a number of other works with the general observation indicating that learned representations vary with increasing network depth, with greater depth typically involving linguistic functions that require larger contextual relationships across the word tokens [2, 7, 8]. However, to the best of our knowledge, BERT has not been examined with respect to its prosodic information.

In general, prosody can be viewed as the characteristics in an utterance that extend individual phonetic segments and encapsulate phonetic and phonological properties that are not due to the choice of individual lexical items, but depend on factors such as their semantic and syntactic relations [9, 10]. These characteristics convey information about the meaning and structure of an utterance. Although prosody is a characteristic of spoken language, the prosodic patterns in speech are connected and interact with their associated sequences of syllables, words, and phrases. Therefore, it is meaningful to assume that some aspects of the prosodic variation can be captured from text alone [11, 3].

In this work, we investigate how prosodic information is linguistically encoded by probing BERT with respect to the prosodic phenomenon of prominence. Prosodic prominence is defined as the subjective impression of a linguistic unit standing out of its context [12, 13, 14]. Given the recent success of BERT in predicting prosodic prominence [3], we attempt to answer the question of what BERT learns about prosody during its pre-training. Does the model rely on general linguistic and

syntactic knowledge for its prosodic predictions or does it have a different pattern of weight allocations across its layers suggesting that the model can capture prosodic information?

We use three datasets with prominence annotations and examine the weights at different layers and compare them with existing findings from the literature on other tasks. To further validate our approach with the literature we also extract part-of-speech tags for our data and examine the learned layer weights. The code to reproduce the results is publicly available at `github.com/skakouros/bert-prosody`. In the next we describe the BERT model architecture, related work, experimental methodology and results.

2. BERT

BERT [5] is a language model based on the Transformer architecture [15] that enables the bidirectional pre-training of representations by jointly conditioning on the left and right context in all layers. This allows the model to learn the entire surrounding context of a word which also means that the same word in different context will have a distinct representation. In contrast, earlier approaches looked at text sequences from left to right or by combining left to right and right to left training. The BERT representations are optimized based on two training objectives: (i) predicting randomly masked words in the input, and (ii) predicting whether the next sentence is the subsequent sentence in the input.

Our experiments are based on the `bert-base-uncased` variant of BERT. The model consists of 12 layers, each with an embedding size of 768, and 12 attention heads.

3. RELATED WORK

Probing network layers to investigate the structural knowledge of the language that a model has captured is an active research area that falls under the topic of neural network interpretability. In recent years there has been an increasing number of studies examining the representations that language models learn. Some works use probing tasks to unveil the linguistic features encoded in neural models [6, 16], others use attribution methods such as Integrated Gradients [17, 18] and some analyze Transformers' attention heads for evidence of linguistic and syntactic phenomena [19]. In this work, we identify each layer's contribution to a specific prosodic task by attaching one trainable weight on each layer of BERT and training a light-weight classification head on top of a frozen pre-trained BERT. This enables us to observe the weight allocations across the layers and compare them with existing findings.

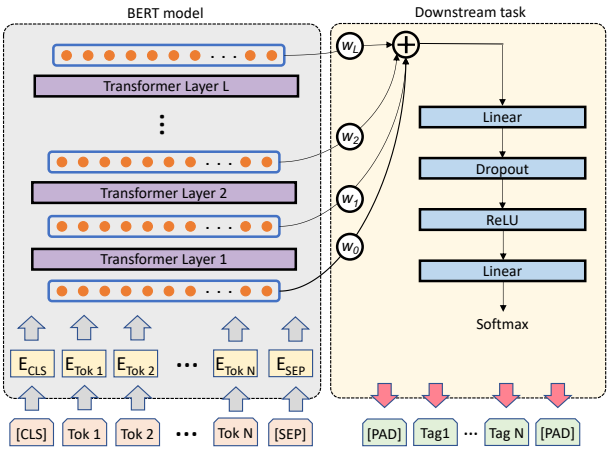


Figure 1: Overview of the experimental setup.

4. LAYER-WISE ANALYSIS

4.1. Layer weights

We obtain the contribution of BERT layers to the prediction task by introducing learnable scalar weights attached to each transformer layer of the model. An overview of the experimental setup can be seen in Fig. 1. We take representations from all transformer layers in the model and collapse them to one via a weighted average. There is one weight for each layer (a total of $L+1$) and all weights are trained jointly with the classification network. The weights are implemented with a learnable vector of size $L+1$, followed by the Softmax function.

4.2. Layer embeddings

To further probe the contribution of individual BERT layers, we extract the embeddings from each layer separately and use them to train a classification head consisting of a single dense layer followed by the Softmax function. For each embedding layer, we then obtain the classification accuracy for the task.

5. EXPERIMENTS

In our experiments we use three datasets: two consisting of read speech and one of spontaneous dialogue speech. These are presented next followed by a description of the experimental setup.

5.1. Data

5.1.1. BURNC

The Boston University Radio News Corpus (BURNC) is a corpus of professionally read news data in American English [20]. The corpus consists of speech from seven speakers (three female). The corpus also contains phonetic alignments, orthographic transcriptions, part-of-speech tags, and prosodic labels. In this work we use the text prompts

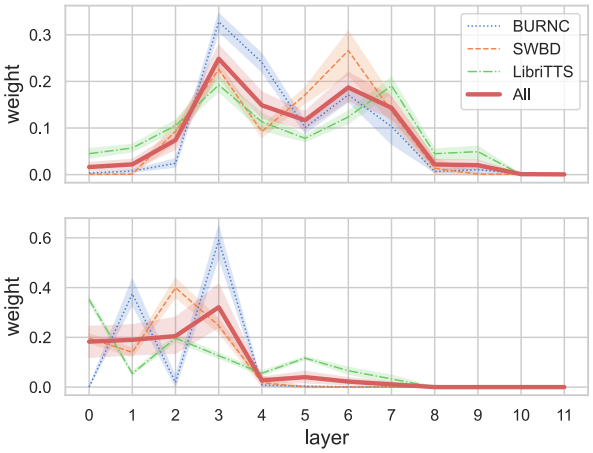


Figure 2: BERT layer weights for prominence (top) and POS (bottom) prediction.

and prosodic labels from the manually labelled part of the corpus (six speakers; approximately 3 h of data). The prosodic labeling system in BURNC is based on the Tones and Breaks Indices (ToBI) labeling convention and includes prosodic phrasing, phrasal prominence, and boundary tones. To obtain a single prominence label, all ToBI pitch accent types (e.g., H*,L*,L*+H) were marked as prominent while the rest as non-prominent.

5.1.2. NXT Switchboard

The NXT Switchboard corpus is a dataset that includes the original Switchboard corpus annotations [21] into one coherent integrated format (NITE XML; NXT) enriched with annotations of prosody and contrast as well as syllable and phone information [22]. The prosody annotations are available for a subset of the data that includes 76 conversations labelled with the ToBI transcribing convention. Similar to BURNC dataset, we use a binary prominence distinction marking words with ToBI accents as prominent and the rest as non-prominent.

5.1.3. LibriTTS

The LibriTTS corpus [23] is a processed (automatically aligned, segmented, and filtered) subset of the original audio and text data of the LibriSpeech corpus [24] that is based on English audiobooks of the LibriVox project. From the corpus we use the *clean* subset that consists of 262.5 hours of read speech from 1230 speakers that were subsequently automatically labelled for prominence in the Helsinki Prosody Corpus (HPC) [3]. From the HPC data, we use the binary prominence tags.

5.2. Experimental Setup

To obtain the layer weights, we train the network (only the classification head and layer weights) with frozen pre-trained BERT for prominence and POS prediction for each dataset separately. For the training we use a 80-15-5 split for train, validation, and test. We run the training for 20 epochs with a batch size of 4 and we repeat each experiment five times. The results are averaged across the independent runs for each task. The network is configured with a learning rate of $5e-5$ and a different learning rate $1e-2$ for the layers weights. This was done in order to allow the model to adjust the weights for the different layers more rapidly. We extract the weights from the model checkpoint with the best development accuracy and average them over the five runs. We repeat the same procedure and setup for POS prediction.

In addition to the frozen pre-trained BERT we also fine-tune the entire model and report the results in Table 2 to compare the overall model performance with both fine-tuned and frozen BERT models. We use the same learning rates and epochs as the previous experiment with frozen BERT.

For POS prediction we used Spacy [25] to extract the part of speech categories. This resulted into 17 discrete classes. POS classes include, for example, adjectives, adpositions, adverbs (see [25] for a complete list of the coarse POS categories).

Finally, to balance the data and enable better comparison, each dataset is post-processed to include one full sentence per sample. As BURNC may include an entire paragraph and Switchboard several sentences per dialogue turn within a sample, we explicitly set sample size to be one sentence. Thus, a batch size of four includes four sentences.

6. RESULTS AND DISCUSSION

We report overall model performance for POS and prominence prediction in Table 2, layer-specific performance in Table 1, and illustrate how layer weights vary with respect to different tasks in Fig. 2. We are interested in examining how different layers contribute to prominence and POS prediction with respect to findings on other tasks that have indicated different linguistic functions associated with different layers. Overall, for prominence we observe widespread distribution of the layer weights while POS appears more focused in the earlier BERT layers. These are presented in more detail in the next sections.

6.1. Prosodic Prominence

For prosodic prominence, the three datasets tested have shown differences in their overall performance. For example, for frozen BERT, prominence prediction accuracy was 87.14% for BURNC, 78.10% for Switchboard, and 82.67% for

Table 1: Layer-wise accuracy for the test set runs for prominence and POS prediction with frozen BERT. Numbers in bold denote the two top results in each task.

Layer	**BURNC**		**SWBD**		**LibriTTS**	
	Prom	*POS*	*Prom*	*POS*	*Prom*	*POS*
0	81.51	93.84	75.32	**94.62**	80.18	89.50
1	83.52	**93.90**	**76.14**	**94.67**	**80.32**	**90.08**
2	85.20	**94.33**	**76.01**	94.02	80.29	**89.53**
3	85.20	93.84	75.47	93.50	**81.08**	89.00
4	**85.26**	93.47	75.25	92.52	80.28	88.40
5	**85.26**	93.17	75.82	91.87	80.00	87.29
6	84.13	92.98	75.71	90.95	79.71	86.12
7	84.46	91.76	75.82	90.11	79.40	84.86
8	83.79	91.03	75.45	88.37	79.05	83.48
9	82.92	89.20	74.71	87.20	78.63	82.36
10	82.59	88.47	74.29	86.64	78.32	81.14
11	83.26	87.19	73.08	82.09	77.27	76.67

LibriTTS. These differences are likely due to the different speaking styles involved in the data, with dialogue speech having the lowest performance. Another interesting observation in the results is that the model performance is degraded for the fine-tuned models. It seems that when fine-tuning the model, generalizability decreases due to overfitting on the idiosyncrasies of the training data. With the frozen model, performance is high and on a par with results reported in the literature. For instance, for LibriTTS [3] report 83.20% and in our setup we obtained 82.67%.

When it comes to layer weights, prosodic prominence has a widespread pattern of weights across BERT layers. Weights span most layers and are primarily focused within layers 2-8 with a peak appearing at layer 3. Interestingly, the pattern of weights appears to be quite similar across the three datasets tested. The spread of the weights suggests that different types of linguistic information are used in the prediction of prominent tokens. One interpretation of the results is that the model relies greatly on surface linguistic features such as POS but also involves syntactic and semantic information with increasing layers (see also [6]). For BURNC, we also observe that the third layer has a high weight for both prominence and POS prediction. It is possible that prominence in BURNC (professionaly read speech) correlated more with POS than audiobooks (LibriTTS) and dialogue speech (SWBD). Another possibility for these differences could be also attributed to the different prominence coding schemes in the three datasets.

6.2. POS

Part-of-speech information seems to be encoded in the early BERT layers with model accuracy being high for both frozen and fine-tuned runs of the experiments and across all datasets tested. Fine-tuning the model leads to improved performance,

Table 2: Accuracy for the test set runs for prominence and POS prediction with frozen and fine-tuned (ft) BERT.

Prominence	**BURNC**	**SWBD**	**LibriTTS**
frozen	87.14	78.10	82.67
ft	85.53	75.95	80.32
POS			
frozen	95.97	97.94	98.49
ft	97.56	98.54	98.96

where for BURNC we get an increase in accuracy from 95.97% with the frozen model to 97.56% with fine-tuning. Switcboard and LibriTTS perform similarly with an increase in performance when the model is fine-tuned.

Layer weights for POS demonstrate a very different pattern when compared to prominence. POS information is found mainly in the lower layers of BERT with weights across the datasets varying but being focused in the early layers of the model. Most of the information seems to come from layers 0-4 which have been shown to encode surface features [6]. This finding is also in agreement with other work that shows maximum POS tagging performance in the lower layers of models with accuracies ranging from 97.2% to 97.4% [7, 2].

7. CONCLUSIONS

In this work, we performed a series of experiments on prosodic prominence to investigate whether prosody is part of the structural information of the language that BERT learns. Our results show that BERT captures information about prosodic prominence through a widespread allocation of weights across its layers reaching high performance. The weight allocations suggest that BERT relies on a variety of linguistic information for its predictions including surface linguistic features such as POS but also involving syntactic and semantic information. In future work, we will explore the same tasks with an extended set of datasets and methodological approaches. In addition to layer weights, we want to include layer integrated gradients as an attribution method in the experiments. We also aim to examine differences between the styles in the datasets, that is, read versus dialogue speech.

8. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland project no. 340125 "Computational Modeling of Prosody in Speech" and Horizon 2020 Marie Sklodowska-Curie grant agreement No 859588. The authors wish to acknowledge CSC - IT Center for Science, Finland, for providing the computational resources.

9. REFERENCES

[1] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," *arXiv preprint arXiv:1805.01052*, 2018.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[3] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proceedings of NoDaLiDa*, 2019.

[4] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "Bert, can he predict contrastive focus? predicting and controlling prominence in neural tts using a language model," *arXiv preprint arXiv:2207.01718*, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[7] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," *arXiv preprint arXiv:1808.08949*, 2018.

[8] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

[9] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[10] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive science*, vol. 40, no. 7, pp. 1739–1774, 2016.

[11] S. Kakouros, J. Pelemans, L. Verwimp, P. Wambacq, and O. Räsänen, "Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence," *Proceedings Interspeech 2016*, vol. 8, pp. 1074–1078, 2016.

[12] S. Kakouros and O. Räsänen, "3pro–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[13] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment*. Springer, 2000, pp. 89–127.

[14] S. Kakouros and O. Räsänen, "Automatic detection of sentence prominence in speech using predictability of word-level acoustic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, D. Das *et al.*, "What do you learn from context? probing for sentence structure in contextualized word representations," *arXiv preprint arXiv:1905.06316*, 2019.

[17] S. Ramnath, P. Nema, D. Sahni, and M. M. Khapra, "Towards interpreting bert for reading comprehension based qa," *arXiv preprint arXiv:2010.08983*, 2020.

[18] A. Nayak and H. P. Timmapathini, "Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by bert," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 2021, pp. 80–85.

[19] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[20] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[21] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[22] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] M. Honnibal and I. Montani, "spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

THE PHONETICS–PHONOLOGY–SYNTAX INTERFACE: A COMPUTATIONAL IMPLEMENTATION

Tina Bögel, Romi Hill, Justin Hofenbitzer, Tianyi Zhao

University of Konstanz
firstname.lastname@uni-konstanz.de

ABSTRACT

This paper introduces a new, computationally implemented end-to-end system for German that takes a speech signal as input, interprets the phonetic data in phonological/prosodic terms, and makes the results available to a linguistically deep computational grammar. The grammar uses the provided information to disambiguate syntactically ambiguous structures, thus reducing overgeneration. A system evaluation showed promising results for this new combination of automatic speech signal analysis and computational grammars, which is a significant step towards a fine-grained linguistic analysis including all grammar modules and hence towards real automatic speech understanding.

Keywords: German, syntactic ambiguities, end-to-end system, automatic speech understanding

1. INTRODUCTION

Systems that allow for automatic speech recognition (ASR) and the identification of prosodic events are often used in phonetic and prosodic research (see, e.g., [1]). For research on German, MAUS [2, 3] is frequently utilized to automatically annotate segments and words. For the identification of prosodic events (e.g., accents or boundaries), available systems include the *Prosodizer*, which assigns pitch accents and boundary tones during speech recognition and synthesis [4, 5], and the prosody module of the Verbmobil system, which integrates word-based annotation and classification of boundaries and accents for German dialogues [6]. [7] trained a number of classifiers on acoustic, phonological, and basic morphosyntactic attributes of German reaching recognition accuracy rates of up to 86% for the occurrence of accents, and 93% for the occurrence of larger boundaries.

Concerning speech synthesis, these approaches went beyond the sole interpretation of acoustic cues and additionally included basic morphosyntactic information (e.g., part-of-speech tags). As this was found to considerably improve accuracy rates, it can be deduced that the algorithms would benefit even more if a fine-grained linguistic analysis of the underlying string were available in the form of syntactic or semantic representations. This information could in turn be associated with specific prosodic events (e.g., grouping or focus). The same is true for ASR which would benefit greatly from the inclusion of deep linguistic information, to allow for real automatic speech understanding (ASU).

While linguistically deep computational grammars (CGs) are available for text-based input and for a number of frameworks (a.o., LFG [8], HPSG [9]), these grammars are unable to process spoken language. As such, these CGs would profit considerably if prosodic information was available for linguistic interpretation.

This paper introduces a new system that bridges the gap between the automatic recognition of prosodic events on the one hand, and CGs on the other. The implementation includes a representation of the speech signal in phonetic and phonological/prosodic terms, where the latter categorical representation enables the CGs to prosodically disambiguate syntactically ambiguous structures. As a consequence, the CG is able to return the correct and linguistically fine-grained representation and thus takes a huge step towards real automatic speech understanding.

2. THE DATA

In the following syntactically ambiguous German sentence, the NP_2 (*der Freundin* 'the friend') can either be part of a genitive structure (meaning a) or constitute a separate dative argument (meaning b).

(1) Sie sahen, dass [der Partner]$_{NP1}$
They saw that the.MASC.NOM partner

[der Freundin]$_{NP2}$ fehlte
the.FEM.GEN/DAT friend was.missing

a) "They saw that the friend's partner was missing."
b) "They saw that the friend missed the partner."

Such syntactically ambiguous structures result in overgeneration, i.e., the CG returns several possible solutions (Figure 1).

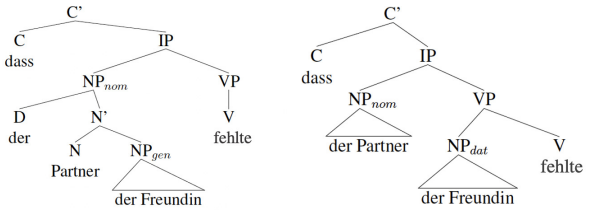


Figure 1: Two syntactic trees resulting from example (1): Genitive structure on the left, dative structure on the right.

These structures can be disambiguated by means of prosody [10] and several studies have shown this to be true for German as well [11, 12]. In a production experiment, [13] showed that speakers place a prosodic phrase boundary between the two NPs in the dative structure in (1), but not in the genitive. Acoustic indicators for the boundary included the lengthening of the last syllable of the first NP ($p < 0.01$), a rise with a following reset in F_0 ($p < 0.01$), and occasionally a pause ($p < 0.05$) between the two NPs in the dative structure.

3. THEORETICAL ANALYSIS

The implementation presented here is couched within the (generative, nontransformational) framework of Lexical-Functional Grammar (LFG) [14, 15]. The modular architecture of LFG proposes different representative structures for separate linguistic aspects (e.g., trees for syntax, see Figure 1), which constrain each other through mathematically well-defined functions. LFG comes with XLE, a state-of-the-art grammar development platform [16, 17], which allows researchers to build industrial strength CGs covering a wide range of languages.[1] However, while these grammars are well-established for syntactic and semantic analyses of texts, they have so far been unable to process spoken language.

This paper follows the theoretical approach to the interface proposed in [18, 13] which discusses the integration of speech signal information into the overall grammar, and develops a representation of the phonetic and prosodic information: the *p-diagram*, a syllable-based representation of the speech signal over time. The prosodic information provided in the p-diagram can subsequently be processed by other modules (e.g., syntax or semantics), thus allowing for a deep linguistic analysis. Under this approach, the input to the grammar consists of a speech signal annotated with syllables. In a first step, each of these syllables receives a vector which, in addition to the related segments, encodes acoustic information relevant to this syllable, e.g., duration or mean F_0. This information is stored at the *signal level* in the p-diagram. Figure 2 shows the p-diagram fragment for the six syllables related to the string *der Partner der Freundin* (ex. (1)), where, e.g., the vector for the first syllable of *partner* has the index S_2, the segments [pa6t] (in SAMPA [19]), the syllable length (0.25s), and a mean F_0 value (181Hz) at the signal level (lower part of Figure 2).

...	...	...	...	...	...		interpretation
PHRASING	-	-	$)_\varphi$	$(_\varphi$	-	-	↓
SEMIT_DIFF	...	-1	6.8	-4.3	-1.9	2.6	
GToBI	-	L*	+H H-	-	L*	+H	
DURATION	0.15	0.25	0.25	0.13	0.31	0.19	signal
FUND. FREQ.	192	181	269	209	188	218	↓
SEGMENTS	[de:6]	[pa6t]	[n6]	[de:6]	[fROYn]	[dIn]	
VECTORINDEX	S_1	S_2	S_3	S_4	S_5	S_6	

Figure 2: The p-diagram's signal level and interpretation level for *der Partner der Freundin* (example (1))

Based on this initial (phonetic) information from the speech signal, a categorical (phonological) *interpretation level* can be derived and added to the p-diagram (upper part of Figure 2), which can include, e.g., differences between adjacent semitones, prosodic phrase boundaries, or a (G)ToBI annotation. For example, a strong rise in F_0 and a following drop (S_2–S_4) and a comparatively long duration on the last (unstressed) syllable of *Partner* (as seen at S_3: [n6]) are strong indicators for a phonological phrase boundary and also justify an L*+H accent. As a result, PHRASING = $)_\varphi$ is added to the syllable's vector at the interpretation level and the L*+H accent is distributed over the associated syllables for the GToBI attribute.

While the p-diagram representation was developed with regard to LFG, it is an encapsulated representation that can be plugged into any modular framework. It is furthermore very adaptable: It could be based on segments or could include any other attribute of interest (e.g., intensity or F_{1-3}). The advantage of the p-diagram is that it allows for a compact and formal representation in terms of ordered vectors with attributes and associated values which makes it accessible to other modules of grammar (see [18] for formal details).

4. COMPUTATIONAL IMPLEMENTATION

The computational implementation follows the theoretical approach to the interface between signal and grammar, as presented in the previous section. Its accuracy is demonstrated by prosodically disambiguating syntactic structures as given in (1).

4.1. Information extraction and preparation

Input to the implementation is a speech signal annotated with (SAMPA) syllables in Praat [20].[2] In a first step, a Praat script extracts information that is relevant for the p-diagram's signal level (Fig. 2): segments, duration, and the mean F_0-value for each syllable vector.

For a more fine-grained analysis of the pitch, the script divides each syllable into five even-spaced subintervals, takes the mean F_0-values of each subinterval and turns the values into semitones, thus effectively normalizing duration and pitch. Each subinterval is also tagged for position within the syllable, either as central or as preceding or following a syllable boundary.

4.2. Interpreting the pitch

In a second step, the raw values from the speech signal are interpreted in terms of categories in order for them to become 'meaningful' for other modules of grammar. In addition to the semitones and the differences between these semitones indicating falls and rises, the implementation also uses residuals of a linear regression calculated based on the pitch values of a given speech signal. The residuals return the distance each value has from this line and are thus a good measure to describe deviations from the average while at the same time including the tendency of the signal's general pitch contour.

Taken together, semitones and residuals allow for the detection of deviations from the norm in the signal, i.e., maximums (H) and minimums (L). To avoid microprosodic effects, the distance between any Hs and Ls has to consist of at least one syllable. Slopes to and from a H/L tone are calculated based on a ratio between the semitones of adjacent Ls and Hs and the distance (the number of subintervals) that lies between them. The resulting values indicate whether the associated slopes are steep or flat.

In order to mark both categories, accent and slope, in one representation, the following system was devised, where each level of L or H is characterised by a particular height and shape of the slopes leading to it (lead) and following it (tail).

Cat.	Max/Min	lead	tail
H4/L4	Max/Min	steep	steep
H3/L3	Max/Min	steep	flat
H2/L2	Max/Min	flat	steep
H1/L1	Max/Min	normal	flat

Table 1: (Part of the) system of pitch accents and slopes in the computational implementation

H4 and L4 thus represent accents where the lead and the tail show a strong rise/fall respectively, while H1 and L1 have a relatively flat lead and tail. L2/L3 and H2/H3 are positioned between these two extremes, with each having a slightly different shape depending on the slopes. These tone values are stored in the interpretation level of the p-diagram (Fig. 2), where they replace the traditional GToBI values in order to facilitate (and simplify) the automatic interpretation by other modules of the grammar.

4.3. Matching against the lexicon

In order to acquire the correct syntactic string, the syllable-based segmental string is matched exhaustively against a lexicon which includes phonological and morphosyntactic material in form of a finite-state transducer [21]. Following [22], the lexicon stores information on the individual segments and the metrical frame, i.e., (in the case of German) the number of syllables, lexical stress and prosodic word status for each word in the p(honological)-form. This p-form is associated with a specific s(yntactic)-form which is then used for the syntactic parse. Figure 3 shows the p-form and s-form for the noun *Freundin*, a trochaic prosodic word, and the determiner *der*, a single, prosodically underspecified syllable.

s-form		**p-form**	
(↑ PRED)	= 'Freundin'	SEGMENTS	/f R OY n d I n/
↑ NUM)	= sg	METR. FRM	$(\prime\sigma\sigma)_\omega$
(↑ GEND)	= fem		
(↑ PRED)	= 'der'	SEGMENTS	/d e 6/
(↑ CASE)	= {gen | dat}	METR. FRM	σ

Figure 3: (Simplified) lexical entries for *der* and *Freundin*

Once the segmental string (*de6.pa6t.n6.de6. fROYn.dIn*) is exhaustively matched against the lexicon, the syntactic string (*der Partner der Freundin*) becomes available for syntactic parsing. In addition, the lexical p-form information can be included in the p-diagram (e.g., information on lexical stress or prosodic word/clitic status).

4.4. The p-diagram

Figure 4 shows an automatically created p-diagram for the string *der Partner der Freundin* based on a speech signal with a dative construction. As discussed in Section 3, each vector includes the segments, the duration, and the mean F_0-value for the associated syllable. The p-diagram also includes the lexical p-form information by

pros_phrase	pp(σ	σ	(σ	σ))pp	pp(σσ	(σ	σ)	(σ	σ))pp
pitch_tones		L2		H4		L2	H1		
lex_stress	–	–	x	–	–	x	–	x	–
F0_mean	225.62	193.49	198.90	267.53	219.35	194.02	213.77	176.27	85.71
duration	0.17	0.16	0.33	0.18	0.14	0.30	0.20	0.28	0.22
segments	das	de:6	pa6t	n6	de:6	fROYn	dIn	fe:l	t@
Vector_index	1	2	3	4	5	6	7	8	9

Figure 4: P-diagram for a dative interpretation of the string *der Partner der Freundin* ('the partner of the friend')

indicating lexically stressed syllables with x and by adding the prosodic unit information to the attribute PROS_PHRASE (prosodic phrasing). While each function word (*dass, der*) is indicated by an underspecified syllable σ, the nouns' prosodic word status is indicated by the syllables within a set of unmarked brackets: (σ σ).

In a next step, the system then calculates different high and low tones (see Section 4.2) and prosodic phrase boundaries $_{pp}(\ )_{pp}$ which can be determined based on the acoustic cues reported in Section 2: F_0 movement, duration, and pauses. At this stage, the placement of prosodic phrase boundaries was only determined by F_0 movement; the estimation whether a particular syllable is longer or shorter than expected is work in progress and will be added shortly to the system.

Figure 4 shows that the system gives a fairly accurate categorical representation of the speech signal. Interesting points of debate are, e.g., the question whether the low tone L2 associated with vector 2 (GToBI: L*), which occurs just before the syllable boundary, should be 'moved' to vector 3 where the syllable carries lexical stress, or whether an additional attribute for 'early' or 'late' L/H tones would be more useful.

4.5. Disambiguation

The syntactic string determined in Section 4.3. is parsed with an LFG-CG for German and returns the expected overgeneration as seen in Figure 1. However, now that the information on prosodic phrase boundaries is available via the p-diagram the grammar is able to prosodically disambiguate the syntactically ambiguous structure. For space reasons, the interested reader is referred to [23] for details of the syntactic implementation.

5. EVALUATION

In order to evaluate the implementation, the recordings from [13] were used to create a 'gold standard'. In an online perception experiment, 32 native German speakers were asked to rate the recordings from the production study on a scale from one to five, where each endpoint corresponded to a particular meaning (1 represented dative and 5 genitive case). All recordings were randomized and assigned to different experimental lists. Participants were asked to listen to each assigned recording and to indicate which meaning they thought was associated with the signal. Each sentence was rated by at least two listeners. Only the sentences that were correctly rated twice (i.e., where the case of the produced sentence matched the case perceived by the listeners) were included in the gold standard and used for the evaluation.

The resulting 72 (59 dative, 13 genitive) recordings were annotated with SAMPA syllables and used as input for the complete implementation. The system was able to determine the correct syntactic structure (dative or genitive) in 55 cases (76,4%) based on the occurrence or absence of a prosodic phrase boundary after the first NP.

6. DISCUSSION AND CONCLUSION

This paper introduced a new end-to-end system, which takes a speech signal annotated with syllables as input, extracts the different acoustic cues, and calculates pitch accents and prosodic phrase boundaries based on this information. The prosodic structure in the resulting representation is used by a computational LFG-Grammar to disambiguate syntactically ambiguous structures. The implementation thus enables these formerly text-only processing grammars to process spoken language as well, and closes the gap between automatic speech recognition and linguistically deep computational grammars. As such, it takes a major step towards real automatic speech understanding.

An initial evaluation of the German system showed promising results which are expected to improve even further once additional factors (e.g., duration) are added to the interpretation. Implementations in two further languages (English and Urdu) with other crucial linguistic phenomena at the interfaces are underway and are expected to challenge and improve other aspects of the system.

7. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with projects BO 3388/3-1 and BO 3388/4-1.

8. REFERENCES

[1] Y. Xu, "ProsodyPro - A tool for large-scale systematic prosody analysis." Laboratoire Parole et Langage, France, 2013.

[2] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech and Language*, vol. 45, pp. 326–347, 2017.

[3] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proceedings of ICPhS*, San Francisco, 1999, pp. 607–610.

[4] N. Braunschweiler, "Automatic detection of prosodic cues," Ph.D. dissertation, University of Konstanz, 2003.

[5] ——, "The prosodizer – automatic prosodic annotations of speech synthesis databases," in *Proceedings of Speech Prosody*, Dresden, Germany, 2006.

[6] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin [a.o.]: Springer, 2000, pp. 106–121.

[7] A. Schweitzer and B. Möbius, "Experiments on automatic prosodic labeling," in *Proceedings of INTERSPEECH*, Brighton, UK, 2009.

[8] M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer, "The parallel grammar project," in *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, 2002.

[9] A. Copestake, *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications, 2002.

[10] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[11] M. Żygis, J. M. T. Jr., C. Petrone, and D. Pfütze, "Acoustic cues of prosodic boundaries in German at different speech rate," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 999–1003.

[12] A. Gollrad, E. Sommerfeld, and F. Kügler, "Prosodic cue weighting in disambiguation: case ambiguity in German," in *Proceedings of Speech Prosody*, Chicago, 2010.

[13] T. Bögel, "German case ambiguities at the interface: production and comprehension," in *Prosody in Syntactic Encoding*, ser. Linguistische Arbeiten, G. Kentner and J. Kremers, Eds. Berlin: De Gruyter, 2020, no. 573, pp. 51–84.

[14] M. Dalrymple, *Lexical Functional Grammar*. San Diego [a.o.]: Academic Press, 2001.

[15] J. Bresnan and R. M. Kaplan, "Lexical-Functional Grammar: a formal system for grammatical representation," in *The Mental Representation of Grammatical Relations*, J. Bresnan, Ed. Cambridge, MA: MIT Press, 1982, pp. 173–281.

[16] M. Butt, T. H. King, M.-E. Niño, and F. Segond, *A Grammar Writer's Cookbook*. Stanford, CA: CSLI, 1999.

[17] R. Crouch, M. Dalrymple, R. M. Kaplan, T. H. King, J. T. Maxwell III, and P. Newman, *XLE documentation*. Palo Alto, CA: Palo Alto Research Center, 2022, online documentation.

[18] T. Bögel, "The syntax–prosody interface in Lexical Functional Grammar," Ph.D. dissertation, University of Konstanz, Konstanz, 2015.

[19] J. C. Wells, "SAMPA computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. Berlin, New York: Mounton de Gruyter, 1997, pp. 684–732.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program, Version 5.3.56]," 2013, available at http://www.praat.org/ [retrieved 15.09.2013].

[21] K. R. Beesley and L. Karttunen, *Finite State Morphology*. Stanford, CA: CSLI Publications, 2003.

[22] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and Brain Sciences*, vol. 22, pp. 1–75, 1999.

[23] T. Bögel, "The prosody-syntax interface: a computational implementation," in *Proceedings of the LFG22 conference*. CSLI Publications/University of Konstanz, 2022.

[24] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database (CD-ROM)," in *Linguistic Data Consortium*. Philadelphia: University of Pennsylvania, 1995.

[1] See the XLE-Web Interface which features a number of different computational LFG grammars that can be used interactively: https://clarino.uib.no/iness/xle-web.

[2] In principle, this step can be automatized as well with the help of MAUS and CELEX [24], but as this paper is concerned with automatic speech understanding, the focus lies on the intersection between signal and grammar.

Macro- and Micro-prosodic Changes in the Duration of Brazilian Portuguese Vowels

João Antônio de Moraes[1,2] Albert Rilliard[1,3]

[1] Federal University of Rio de Janeiro, Brazil; [2] CNPq; [3] Université Paris Saclay, CNRS, LISN, France
jamoraes3@gmail.com; albert.rilliard@lisn.upsaclay.fr

ABSTRACT

This accounts for the effect of intrinsic segmental characteristics on the duration of Brazilian Portuguese oral vowels, compared to durational changes that can be attributed to macro-prosodic features, such as stress location, sentence modality, and syntactic structure. A multi-speakers corpus containing logatomes with systematic variations of vowel quality (seven vowels) embedded in sentences with controlled characteristics was recorded. A semi-automatic procedure allowed to segment and measure the duration of 3036 vowels; their duration was fitted to a series of factors, representing the controlled micro- and macro-prosodic characteristics. This modeling process showed the importance of micro-prosodic changes mainly in the case of stressed vowels (explained by jaw aperture and tongue posteriorization), which decreases with the vocalic reduction on prestress and post-stress positions, depending on macro-prosodic factors. The main macro-prosodic effects are linked to stress, the syllable position, final lengthening, and modality.

Keywords: Vocalic duration, Micro-prosody, Macro-prosody, Brazilian Portuguese.

1. INTRODUCTION

Segmental effects on the fundamental frequency (F0), intensity, and duration are referred to as micro-prosodic effects and may influence the description of macro-prosodic changes; thus the choice often made by prosodists to work on lab speech based on controlled sentences. This paper aims to describe quantitatively the variations induced in the duration of Brazilian Portuguese (BP) vowels by both micro-prosodic factors (i.e., the segmental, intrinsic characteristics of these vowels) compared to macro-prosodic ones (stress position in the word, the word's place in the sentence and the sentence modality).

The vocalic system of BP has seven oral vowels, /i e ɛ a ɔ o u/, that form minimal pairs at stress positions, but it reduces to five opposable qualities at pre-tonic positions with the neutralization of the middle vowels /e/ ~ /ɛ/ and /ɔ/ ~ /o/ [1], although the seven qualities are still produced mainly due to vocalic harmony processes [2]. Post-tonic positions show only three qualities in the word-final position, /i a u/, but may have four or even five qualities (/i e a o u/) in non-final post-tonic syllables (e.g., pen-ultimate syllables of proparoxytonic words) [3].

The effect of vowel quality on segmental duration has been evidenced for Dutch or English [4], [5], or as part of a larger set of factors (e.g., [6]), as an intrinsic duration factor [7], with jaw aperture being the main explication to intrinsic duration changes. Several works have dealt with vocalic duration for Brazilian Portuguese. [8] measured the duration of BP vowels of ten speakers from Rio de Janeiro at different sentence positions and confirmed a lengthening linked to aperture until mid-open vowels – /a/ being shorter than /ɛ/ and /ɔ/; differences along the anteroposterior dimension were observed, but not consistently. Similar results were replicated for another BP dialect, São Paulo [9]. [10] observed the duration of BP oral vowels in 20 speakers from São Paulo for stress position of disyllabic paroxytones. Differing from the two previous studies, they showed an increased duration linked to aperture even for the /a/; they also observed some effect on the duration of the front-back articulatory dimension.

The present article targets the duration of BP oral vowels and is part of a wider project aiming at describing micro-prosodic effects (intrinsic and co-intrinsic) in relation to macro-prosodic changes; to that aim, known factors of variation were controlled while creating the corpus. After presenting the corpus construction and recording, the output of the fit of syllable durations to the controlled factors is given and discussed.

2. METHODS

2.1. Corpus construction

The corpus is based on trisyllabic $CV_1CV_2CV_3$ logatomes (with C being /p/, and V_x on the seven oral vowels of BP /i, e, ɛ, a, ɔ, o, u/) inserted in the carrier sentence "*Ele dirá X [de novo], mas parecerá Y [de novo]*" ("*He will say X [again], but it will sound like Y [again]*"), where X and Y are two different logatomes, and [*de novo*] an optional extension. The sentence may or not propose an extension of the nominal group containing the logatome, and it was produced with two modalities – as an assertion or an interrogation.

The logatomes had the stress at each possible position (oxytone, paroxytone, proparoxytone). They were constructed with the same vowel at each position (thus, 21 logatomes are considered here: 7 vowels x 3 stress positions), with a limitation of the vowel inventory at unstressed syllables (see introduction).

2.2. Recording procedure

Six speakers (3 females and 3 males, mean age 46 years) were recorded in a sound attenuated room, using an omnidirectional measurement DBX RTA PHM919 microphone placed 30cm from their mouth, plugged into a Zoom H5 recorder, with the recording level calibrated by a Brüel & Kjær 4230 sound level calibrator. Speakers were presented with the sentences in random order via a computer interface and had to read them once – unless the experimenter judged the reading unsatisfactory.

2.3. Segmentation and measurements

A forced alignment procedure was applied to the recorded sentences based on their orthographic transcription supplied to the Montreal Forced Aligner [11]. Ad hoc phonetization of logatomes was provided to the program. The segmentation was then hand-corrected using Praat [12] by two experienced phoneticians. Considering two sentence modalities, two logatome positions in the sentence, the eventual extension, three vowels per each of 21 logatomes, and six speakers, it ended with 3036 individual vowels. From this dataset, a Praat script extracted the duration of each segment and other measures not presented here.

2.4. Statistical analysis

The following factors were controlled: the *Logatome position* (**LogP**) – at the syntagm end or not (Boundary or No boundary); the syntagm was at the initial part of the sentence or the end (*Syntagm position* (**SynP**): initial or final); the sentences were produced with a given *Modality* (**Mod**: Assertive or Interrogative). The logatomes also bear a specific *Stress pattern* (**Stress**: Oxitonic, Paroxitonic, Proparoxitonic), and its syllables are at three positions (*Syllable position* (**SylP**): Antepenultimate, penultimate, and ultimate). These five controlled factors are linked with macro-prosodic characteristics.

Then, the set of phonological vowels constitutes the micro-prosodic controlled factor (Vowel quality – **VowQ**). As existing vowels in PB vary according to the syllable position and the word tonal pattern (with syllables bearing the stress or being at pre- or post-stressed positions), different linear mixed models were fitted to the data. A first model was fitted on the cardinal extreme vowels /i a u/, found in all possible levels of the controlled factors (even if reduced at post-stressed positions) to test the effect of all macro-prosodic factors. Then, different mixed linear models were fitted on subsets of syllables, according to their position relative to stress (which determines the number of possible vocalic timbres): one model was made on stressed syllables (7 qualities), one on prestressed vowels (7 qualities with 5 having phonological oppositions), one on post-stress syllables at word-final position (3 qualities), and one on non-final post-stress syllables of proparoxytones (5 qualities).

All models were fitted using the same process (following [13] and using the lme4 library [14] of the R software [15]): we started by fitting a maximal model, with, as fixed factors, the micro-prosodic factor (having a varying number of levels as a function of the selected vocalic qualities, from 3 to 7) and the available macro-prosodic factors, with all interactions, and a random structure with an intercept for the six speakers (**Spk**) and for each produced sentence (**Snt**) nested in the speakers. This maximal model was simplified until reaching a minimal model

adequate for describing the change in vocalic duration (**Dur**, in ms), used as the dependent variable of these models. The following section describes these different models and the main effects observed in each case; ANOVA tables are presented in Tables 1 & 2 for the first two models, with the effect size estimated using the partial ω^2 given by the *effectsize* library [6].

3. RESULTS

3.1. Models of duration

The exact formulas of the minimal models are given in the endnotes for the sake of clarity.

3.1.1. Duration of cardinal vowels

The minimal model[1] was based on all main fixed factors, implied in two four-way interactions linked to macro-prosodic effects (*Stress:SylP:SynP:LogP* and *SylP:SynP:LogP:Mod*), and three-way and two-way interactions mixing micro- and macro-prosodic factors (*VowQ:SynP* and *VowQ:Stress:SylP*). The random structure kept only the *Spk* effect. This model explained about 81% of the variance, with fixed factors responsible for 75% ($R^2_{marginal} = 0.75$; $R^2_{conditional} = 0.81$). From this complex structure, most of the changes in duration are explained by the *Stress*:*SylP* interaction, i.e., if the syllable bears or not the stress. In Figure 1, the durations of stressed vowels are systematically longer than the unstressed ones and above 100ms; most unstressed syllables have a duration between 50 and 75ms. Factors that induce lengthening are syllable position at word end (ultimate) when a syntactic boundary occurs and syntagm at the sentence-final position.

The *SylP:SynP:LogP:Mod* interaction is essentially linked to a further lengthening of sentence-final vowels (i.e., ultimate syllable of the word, for logatome at syntagm end and syntagm at sentence end), for interrogative modality compared to assertive.

The micro-prosodic effect showed longer /a/ and shorter /i/, with /u/ in between (thus an effect of vowels height and backness dimensions) – but this figure is affected by macro-prosodic characteristics: this holds for the longest vowels (i.e., the stressed ones). Conversely, most unstressed vowels, with centered qualities, showed few to no micro-prosodic differences (see latter models for details).

	χ^2	df	*p*	ω^2
VowQ	85.2	2	0.000	0.06
Stress	73.7	2	0.000	0.06
SylP	184.6	2	0.000	0.13
LogP	319.3	1	0.000	0.21
VowQ: SynP	6.1	2	0.048	0.00
SynP:SylP	57.7	2	0.000	0.04
Stress:SylP	3994.6	4	0.000	0.76
SynP:LogP	4.2	1	0.041	0.00
SylP:LogP	149.8	2	0.000	0.11
Mod:LogP	6.2	1	0.013	0.00
VowQ:Stress:SylP	35.0	8	0.000	0.02
SynP:Stress:SylP	12.4	4	0.014	0.00
SynP:SylP: LogP	31.1	2	0.000	0.02
Stress:SylP: LogP	112.3	4	0.000	0.08
SynP:Mod: LogP	7.4	1	0.006	0.00
Stress:SylP:SynP:LogP	11.9	4	0.018	0.00
SylP:SynP:LogP:Mod	13.6	2	0.001	0.00

Table 1: Significant factors for the model fit on cardinal vowels. For each factor or interaction: corresponding χ^2, degree of freedom, *p*-value, and partial ω^2. Only significant effects are reported here.

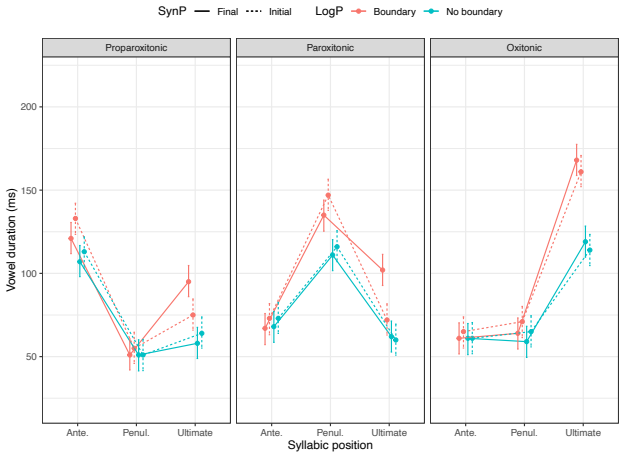


Figure 1: mean fitted vocalic duration for the *Stress:SylP:SynP:LogP* interaction

3.1.2. Duration of stressed vowels

The model on stressed vowels was based on *VowQ* plus the four-way interaction between *SylP*, *LogP*, *SynP*, and *Mod*, keeping the complete random structure. There was thus no interaction between the micro- and macro-prosodic levels: the *VowQ* factor, with seven qualities, was replaced by two factors – for the vowel's "***Height***" (high, mid, low) and "***Backness***" (front, back)– to reach the minimal model[2] (see Table 2); the interaction *Height:*

Backness was not significant. The model had a lower (72%) explanatory power, with fixed factors explaining only 37% of the variance: this shows the importance of stress lengthening, which is not accounted for here (considering only stressed vowels), leaving more importance to random factors.

	χ^2	df	*p*	ω^2
Height	181.5	2	0.000	0.20
Backness	36.4	1	0.000	0.03
SynP	37.7	1	0.000	0.07
SylP	93.7	2	0.000	0.16
Mode	0.4	1	0.505	0.00
LogP	295.5	1	0.000	0.38
SynP: SylP	15.8	2	0.000	0.03
SynP:Mode	13.7	1	0.000	0.03
SylP:Mode	0.4	2	0.833	0.00
SynP:LogP	9.3	1	0.002	0.02
SylP:LogP	43.3	2	0.000	0.08
Mode:LogP	7.7	1	0.005	0.01
SynP:SylP:Mode	2.1	2	0.346	0.00
SynP:SylP:LogP	1.8	2	0.402	0.00
SynP:Mode:LogP	22.6	1	0.000	0.04
SylP:Mode:LogP	1.6	2	0.451	0.00
SynP:SylP:Mode:LogP	6.2	2	0.046	0.00

Table 2: Significant factors for the model fit on stressed vowels. For each factor or interaction: corresponding χ^2, degree of freedom, *p*-value, and partial ω^2.

The 4-way interaction explains classical macro-prosodic lengthening at phrase final boundary, an interaction with modality; details cannot fit here. The two micro-prosodic factors, vowel *Height* and *Backness,* show the importance of both characteristics on the duration change, with a relatively important effect size for *Heigh* (ω^2=0.2; with a 9ms lengthening from high to mid vowels, and a 12ms one from mid to "low" ones, that groups /a, ɛ, ɔ/), and reduced effect size for *Backness* (ω^2=0.03; with a 9ms lengthening for back vowels over front ones). We did not observe a duration difference between low /a/ and low-mid vowels – thus, a three-level *Height* factor was sufficient (and /a/ was processed as a front vowel for *Backness*).

3.1.3. Duration of unstressed vowels

Three models were fitted for the duration of unstressed vowels, which descriptions do not fit here (see formulas in footnotes): one for pretonic vowels[3], one for final post-tonic vowel[4], and one for non-final post-tonic[5]. An interaction between micro- and macro-prosodic factors was observed for pretonic vowels, showing that the micro-prosodic differences no longer apply for macro-prosodic factors that most reduce the vocalic duration. This was also observed for non-final post-tonic vowels, the position showing the shortest vowels, where the five vowel qualities have mostly comparable durations. In the case of final post-tonic vowels lengthened because of their position, the three qualities show the same hierarchy described for cardinal vowels (/a/ > /u/ > /i/).

4. DISCUSSION & CONCLUSION

This work gives a first and rapid analysis of the micro- and macro-prosodic factors that influence oral vowel duration in BP in this dataset. As only intrinsic duration phenomena were considered here, this shall be extended in the future to report on co-intrinsic phenomena and include nasal vowels and diphthongs.

The results mostly back what was already described in the literature on micro- and macro-prosodic effects. It replicates the same results as [8, 9], with /a/ not being the longest vowel but comparable in aperture effect to /ɛ/ and /ɔ/, a finding that differs from [10] results (in our data, /ɔ/, not /a/, was the longest stressed vowel: as a "low" –i.e., within the category of longest vowels in our data, for the vowel Height factor– and back vowel). The effect of posterior articulation was systematic in our data for stressed vowels. This differs from the three previous studies, which do not report systematic lengthening with posterior articulation.

Another difference was linked to using a wide set of macro-prosodic factors, the effect of which has been shown here to interact with the micro-prosodic level. Further investigation may be required to understand when and why this interaction happens and propose an interpretation of the reasons for such behavior, but it seems there is a link between the duration of a vowel and the strength of the micro-prosodic effects, thus they are more obvious on stressed vowels. Vowels with shorter durations tend to be centralized, and this affects the importance of articulatory gestures – thus the reduced micro-prosodic changes in these cases.

5. ACKNOWLEDGMENTS

The authors are deeply grateful to the six speakers that agreed to a considerable effort to record the corpus.

6. REFERENCES

[1] J. M. Câmara, Para o estudo da fonêmica portuguesa. Padrão, 1977.
[2] S. H. Lee, 'Sobre as vogais pré-tônicas no Português Brasileiro', Estud. Lingüíst., vol. XXXV, pp. 166–175, 2006.
[3] L. Bisol, 'Neutralização das átonas', DELTA Doc. Estud. Em Lingüíst. Teórica E Apl., vol. 19, no. 2, pp. 267–276, 2003, doi: 10.1590/S0102-44502003000200002.
[4] S. G. Nooteboom, 'Production and perception of vowel duration in Dutch: A Study of Durational Properties of Vowels in Dutch', University of Utrecht, 1972.
[5] D. H. Klatt, 'Interaction between two factors that influence vowel duration', J. Acoust. Soc. Am., vol. 54, no. 4, pp. 1102–1104, Oct. 1973, doi: 10.1121/1.1914322.
[6] J. P. H. van Santen, 'Contextual effects on vowel duration', Speech Commun., vol. 11, no. 6, pp. 513–546, Dec. 1992, doi: 10.1016/0167-6393(92)90027-5.
[7] A. Di Cristo, De la microprosodie à l'intonosyntaxe, 2 vols. Aix-en-Provence : Marseille: Publications, Université de Provence ; Diffusion, J. Laffitte, 1985.
[8] J. A. De Moraes, 'Durée Intrinsèque des Voyelles du Portugais Brésilien', in Mélanges Léon: Phonetique, Phonostylistique, Linguistique et Litterature, Mélodie., P. Martin, Ed. Toronto, 1992, pp. 367–377.
[9] P. A. Barbosa, 'Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala', in Estudos de prosódia, E. M. Scarpa, Ed. Campinas, SP, Brasil: Editora da UNICAMP, 1999, pp. 21–52.
[10] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, 'A cross-dialect acoustic description of vowels: Brazilian and European Portuguese', J. Acoust. Soc. Am., vol. 126, no. 3, pp. 1379–1393, Sep. 2009, doi: 10.1121/1.3180321.
[11] M. McAuliffe et al., 'MontrealCorpusTools/Montreal-Forced-Aligner: Version 2.0.6'. Zenodo, Aug. 08, 2022. doi: 10.5281/ZENODO.6972621.
[12] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer [Computer program]. Version 6.2.08'. Feb. 05, 2022. [Online]. Available: http://www.praat.org/
[13] S. T. Gries, Statistics for linguistics with R: a practical introduction, 3rd revised edition. Berlin Boston: de Gruyter Mouton, 2021.
[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models Using lme4', J. Stat. Softw., vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.
[15] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/
[16] M. S. Ben-Shachar, D. Lüdecke, and D. Makowski, 'effectsize: Estimation of Effect Size Indices and Standardized Parameters', J. Open Source Softw., vol. 5, no. 56, p. 2815, 2020, doi: 10.21105/joss.02815.

[1] Formula for the model on cardinal vowels:
Dur ~ VowQ + TonP + SylP + SynP + LogP + Mod + VowQ:TonP:SylP + VowQ:SynP + TonP:SylP:SynP:LogP + SylP:SynP:LogP:Mod + (1 | Spk)

[2] Formula for the model on stressed vowels:
Dur ~ Height + Backness + SylP * LogP * SynP * Mod + (1 | Spk/Snt)

[3] Formula for pretonic vowels:
Dur ~ VowQ + SynP + SylP + Mod + LogP + VowQ: SylP + SynP: SylP + SylP: LogP + SynP:Mod + SynP: LogP + Mod: LogP + SynP:Mod: LogP + (1 | Spk/Snt)

[4] Formula for final post-tonic vowels:
Dur ~ VowQ + SynP + Stress + Mod + LogP + SynP:LogP + SynP:Stress + Stress:LogP + SynP:Stress:LogP + SynP:Mod + Mode:LogP + SynP:Mod:LogP + (1 | Spk)

[5] Formula for non-final post-tonic vowels:
Dur ~ VowQ + SynP + LogP + SynP:LogP + (1 | Spk/Snt)

American English Pitch Accent Dynamics: A Minimal Model

Khalil Iskarous[1], Jeremy Steffman[2], Jennifer Cole[3]

[1] University of Southern California, [2]The University of Edinburgh, [3]Northwestern University
kiskarous@usc.edu, jennifer.cole1@northwestern.edu, jeremy.steffman@ed.ac.uk

ABSTRACT

What is the minimal mathematical model that can generate the F0 trajectories for a system of pitch accents? In this work, we propose a nonlinear coupled dynamical systems theory of American English pitch accents with a single basic parameter. As that parameter increases, F0 profiles for different pitch accents are generated. The terms in the differential equation are based on a novel dynamical analysis of a large database of F0 productions in terms of measurements of F0 peak, peak velocity, and the time to achieve peak velocity. We describe the basic dynamical properties of pitch accents in our database and argue for the proposed model as the simplest one that realizes all the major dynamic F0 properties of the pitch accent system. We argue that the proposed model describes both abstract phonological and concrete phonetic aspects of the system.

Keywords: Intonation; Pitch Accents; Dynamical Systems; Fitzhugh-Nagumo Differential Equation.

1. INTRODUCTION

What are the basic scales or dimensions on which the pitch accents of a language could be contrasted? Current level-based theories of prominence contrast pitch accents in terms of the symbols: L, H, and *, indicating static, relative F0 target values (low, high), and their temporal alignment with a phonological landmark [1,2]. An algorithm translates from the symbolic expressions to F0 trajectories. Yet there is a long research line arguing for the inherent temporal gradience and fundamental variability in the expressions of F0-based prominence [3,4,5,6], emphasizing that it is the F0 *configuration*, not just target values, that are used for pitch-based prominence [3,7]. And polynomial coefficients have been used successfully to describe those configurations [8,9,10].

Another approach is the use of dynamical systems theory to predict F0 trajectories by solving a differential equation that expresses the relation between F0 and its derivative(s) *at every moment in time* [11]. Differential equations are valuable since they bridge the abstract and concrete. The relation between the value of a function and its derivatives is abstract as it holds invariantly (or within stochastic limits) even as the value of the function and its derivative change, but it can also explain concrete details of the shape of the function. Thus, differential equation descriptions can be seen as a bridge between abstract level-based descriptions, and more concrete configuration-based descriptions. Moreover, the differential equation has parameters that can be used as a scale for contrast. The greatest hope for a dynamical account of pitch accents is that they potentially allow for observed properties of the phenomenon to *emerge*, rather than be stipulated symbolically or verbally. This is why we believe that a proposal for an underlying differential equation for a pitch system is worthwhile. For general discussions of differential equation-based approaches to phonology and phonetics, the reader can consult [12,13]. In the original dynamical work on F0 shape [11], a linear differential equation was used. In the present work, we propose a new nonlinear dynamical system of differential equations with a single parameter and show that the variation in shape of pitch contours in Mainstream American English (MAE) requires this additional complexity. Nonlinear differential equations have also recently been used [14,15] to describe the processes of tonal selection, as an application of intentional dynamics to accent and meaning [16].

Our goal is to show, through novel quantitative analysis of F0 dynamics in a large database of MAE intonation, that the minimal model required to account for the dynamical properties is a cubic nonlinear system representing the interaction of F0-raising and F0-lowering forces. The specific system we propose is an instance of the Fitzhugh-Nagumo equation, one of the most fundamental equations in mathematical neuroscience [17]. However, we modify these equations in a novel way, to represent discrete, as opposed to rhythmic movement [18]. We will show that this system for describing pitch accents combines phonetics and phonology in a highly organic way, contributing to the solution of the problem of how phonetics and phonology combine in the description of intonation [19].

2. DATA AND METHODS

The pitch accent model we propose is based on data from 130 MAE speakers, aggregated from several imitative speech production experiments [29,30]. In

each experiment, on a given trial, participants heard two model utterances that exemplified a particular tune. F0 in the model utterances was resynthesized based on straight-line approximations from [1]. The participant produced the heard tune on a metrically and syntactically similar sentence (e.g., *heard*: "He answered Jeremy", "Her name is Marilyn"; *produced*: "They honored Melanie"). Here we focus on F0 of the final trisyllabic stress-initial word, which bears the nuclear (phrase-final) pitch accent. Data from 70 speakers producing MAE H*, L+H* and L*+H in all intonational boundary contexts (H-H%, H-L%, L-H%, L-L%) is used, with 48 repetitions per speaker, per accent. Data for L* is taken from two different experiments in which 60 speakers each produced 72 repetitions of L* (and other accents) across all four boundary contexts. F0 was extracted from the nuclear word and time-normalized to 30 samples.

A differential equation describes the functional relation at every moment of time between the function's value, e.g., and the value of its derivatives (e.g., velocity $\frac{dx}{dt}$ and maybe acceleration $\frac{d^2x}{dt^2}$). To determine a differential equation, dynamical properties of the data are examined, and the simplest differential equation accounting for those properties is proposed. The equation can then be solved and predicted solutions can be assessed for their fit to the data. In the study of speech dynamics, the peak velocity (PV) and the time at which peak velocity is reached (TTPV) are of paramount importance in the attempt to induce a differential equation [20,12,21], and they will be two of the measurements we present as clues, along with the error-corrected extremum (max/min) F0 during the pitch accents. Though TTPV is not frequently used as a measure in intonation modeling, it serves to quantify the amount of delay in pitch with respect to some supralaryngeal event like a stressed vowel. Indeed, the notion of delay is one of the earliest innovations of the Autosegmental-Metrical approach to tone, in Goldsmith's introduction of the star diacritic * to signify alignment of tonal to non-tonal tiers [23].

3. EMPIRICAL RESULTS

The top panel of Figure 1 shows mean curves of each pitch accent over the entire data, a qualitative view of the F0 trajectories (a) and F0 velocity (b) for the initial 2/3 of the entire nuclear tune interval (to exclude most of the final region that implements the boundary-marking tones). Theses panels show the extrema of the pitch accents and where they are reached. F0 trajectories are normalized to start at the same value. The trajectories show that these data are highly representative of observations made throughout the last few decades about the pitch accents of MAE: the rises rise to different extents and show different alignments with the stressed syllable. Note that H* reaches a lower F0 peak than the bitonal accents. Since this paper is about the dynamics of F0, it is also important to view the F0 velocity curves (b), which shows the expected earlier rise of L+H* vs. L*+H [2], for instance, by a later achievement of peak velocity for the latter.

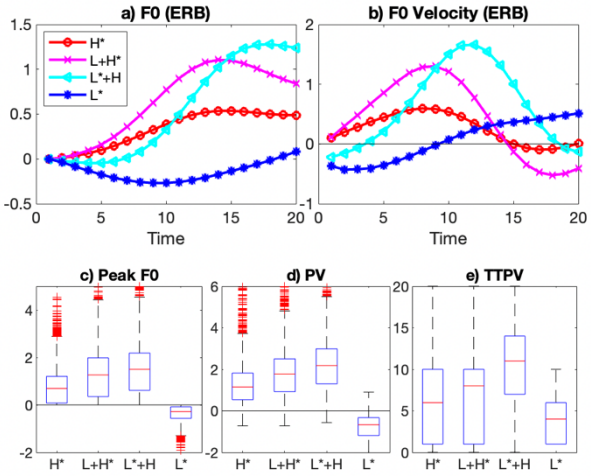


Figure 1: a) Mean F0 trajectories. b) Mean F0 velocity. c) Peak F0, d) Peak velocity, e) Time to peak velocity by pitch accent type.

We do not aim in this paper to provide a statistical determination of the phonetic properties of the AM pitch accent labels. Our goal, rather, is to glean from these dynamical measurements clues as to the relation between F0 and its derivatives. We use the effect size metric *Cohen's d*, which is a measure of the difference in means between two distributions, normalized by pooled standard-deviation [22], to assess the difference between accents in their dynamical F0 measures. To judge magnitude of effect, Cohen [22] argued that a .5 difference is a *medium* effect, and a .8 difference is *large*. Within the rises, for Peak F0, there is a medium effect for H* vs. L+H* (d = -0.58), a large effect for H* vs. L*+H (d = -.77), but there is no effect for L+H* vs. L*+H. For PV, there is a medium effect for H* vs. L+H* (d = -0.50), a large effect for H* vs. L*+H (d = -.81), but there is no effect for L+H* vs. L*+H. For TTPV, there is no effect for H* vs. L+H*, a medium effect for H* vs. L*+H (d = -.55), and a medium effect for L+H* vs. L*+H (d = -.53). L* is distinct from the rises with a large effect for Peak F0 and PV. Considering all three rising accents, we observe significant overlap in the distributions for each F0 measure, resulting in mostly medium effect sizes, signalling a gradient scale of dynamic variation. But despite this variation, the Cohen's d above shows a pattern among the MAE rising pitch accents: *the*

higher F0 goes, the faster it rises, and the later it rises. This will be our first clue to the underlying dynamic. Of course, it's possible to have a model in which Peak F0, PV, and TTPV are independently controlled, but the variation we see in Figure 1 points to joint control of these dynamical measures. Since a dynamical model captures the relation between F0 and its derivatives, it should be able to capture at least some aspects of this covariation, with as few parameters as possible.

Another observation from Figure 1 is that, as has been remarked in the literature, L*+H is often "scooped", falling before it rises. This is crucial to a dynamical model, because for the same value of F0 the system can output a negative velocity (for the falling part of the scooped rise) or a positive velocity (for the rising part). This means that for the dynamic model of pitch accents, velocity is not a single-valued function of F_0, and instead requires two variables, each with its own differential equation, not just one [17]. Another alternative is to use acceleration as part of the model, however, we have found that the resulting model does not capture MAE pitch accent shapes without having several controllable parameters. As will be seen for the proposed pitch accent model, these two variables can be interpreted as the level of excitation of the forces for raising and lowering pitch, respectively.

To summarize, the dynamical clues are: 1) rise later-rise more; 2) the dynamical system needs to have at least two interacting differential equations.

4. DIFFERENTIAL EQUATION MODEL

We first motivate the proposed differential equation to account for the dynamical properties apparent from Figure 1, and then offer a brief interpretation for how it operates in the phonetic and phonological systems. Instead of presenting the entire differential equation directly, we will build it bit by bit from the simplest interesting differential equation, so that the properties gained from each term are clear. The hierarchy of differential equations have polynomial functions on the right-hand side: starting from the simplest linear polynomial and ending with a cubic polynomial. While it's possible to consider more complex, non-polynomial functions, our focus on polynomials is justified by Taylor's Theorem showing that any function can be approximated with a polynomial of arbitrary degree. Also, we consider only 1-parameter differential equations and models with velocity, but no acceleration, as we seek a minimally complex model with the fewest number of controllable parameters.

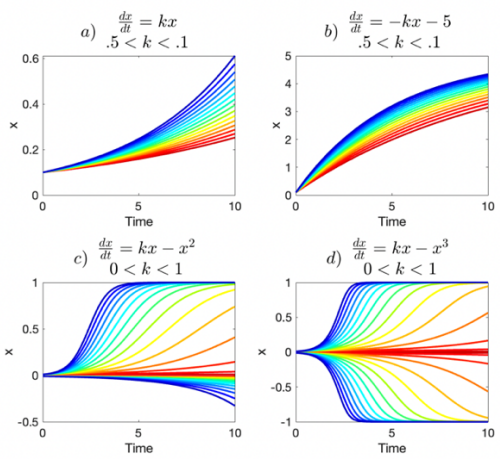


Figure 2: A hierarchy of 1-paramater differential equations. k is swept from a low number (red) to a high number (blue).

Figure 2 shows the solutions of several differential equations with a single parameter *k,* with trajectories generated with different values of *k* within the specified range. In (2a) the slope function $\frac{dF_0}{dt}$ (where x represents F_0) is linear and positively related to the value F0, which yields exponential growth with the rate of growth set by the parameter *k*. The F0 trajectories generated by different *k* values with this model are like those of rising pitch accents, but unlike empirical F0 trajectories, these trajectories rise without bound, so the model fails. In (2b) *k* is negative and a constant is added, which yields bounded growth and generates earlier and later rise patterns, as seen for the different rising accents in the empirical data. But in this model velocity peaks at time 0, where curves are steepest, contrary to the empirical data (Figure 1), so the model fails. For a quadratic polynomial as in (2c), as x approaches 1, x and x^2 balance (become equal in value), which means their difference becomes 0, at which point the trajectory flattens at an equilibrium value. Here it's possible to reach equilibrium with peak velocities at time points later than 0, as in the empirical F0 rises. However, note that the trajectories generated from this model exhibit the pattern "rise-later, rise-less" (trajectories where the rise starts later, rise to a lower F0 peak value), the opposite to the empirical pattern for rises, so this model also fails. The trajectories above 0 are for an initial small positive x. If x starts as a small negative, however, which we show below 0, there is an exponential fall. For a cubic polynomial (d), there is balance of the linear and cubic terms at two equilibria 1 and -1, so there can be stable (bounded) rises and falls as observed in the empirical trajectories. In this model we see the germ of the tonal constructs Low (L) and High (H), emergent here as stable F0 goals at -1 and 1. As already noted, to get to the Low equilibrium requires a negative initial

condition. Also, this model, like that in (2c), generates the pattern of "rise-later rise-less", opposite to the empirical pattern, so again, the model fails.

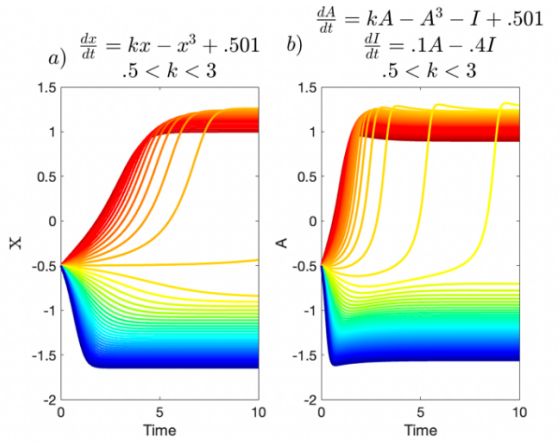


Figure 3: Cubic Models with broken L/H symmetry (a) and, in addition, (b) to an Activator/Inhibitor dynamic.

Figure (3a) shows the result of breaking the symmetry between the high equilibrium at 1 and the low equilibrium at -1 of Figure (2d) by adding a small constant .501. All the curves in (3a) rise from the same baseline -0.5 and the constant 0.501 is chosen to be just large enough to release the system from the initial inhibited value. As *k* varies, both the L and H equilibria are stable. Lower *k* (red) is H*-like, rising early to a low extent, and as *k* rises (orange), we get to intermediate L+H*-like, then L*+H trajectories, before L* for the highest *k*. Therefore, a key dynamic property of the data in Figure 1 *emerges* from adding that constant .501. However, the differential equation in (3a). has a major problem: rises can only rise, there is no ability to fall then rise, as we see in many scooped L*+H.

Our proposed model is in Figure (3b). As discussed below, we interpret A as F0. This model, with the help of the interacting variable I, is also able to generate L*+H scooped rises. There are two variables, an Activator A, and an Inhibitor I. The A equation is the same as in (3a), except that now A's change also depends on the value of the second variable, I. The negative sign on I signifies that I inhibits A. I depends on A, but positively due to the positive coefficient of the A term in the function for $\frac{dI}{dt}$. As the value of A increases (an F0 rise), it increases I, which in turn inhibits A. This Activator-Inhibitor dynamic is fundamental to mathematical biology in general [24] and theoretical neuroscience in particular [25, 26]. The specific instantiation we developed in (3b) is a version of the Fitzhugh-Nagumo equation [17] that we modified via the equation for I so it will not generate oscillatory/rhythmic trajectories, but instead generate trajectories that achieve one of two discrete equilibria, corresponding to the H/L tone targets of the MAE pitch accent system. Note how variation in a single parameter *k*, as it is gradiently swept, generates rises with properties as we observe, as well as falls. Even small effects like the earlier fall for L* vs. the rises is captured. We can therefore say that from the variation of one parameter *emerges* a set of correlated dynamical properties. Note however, that the intonation system of MAE does not need to use every value of *k*. Categories of pitch accents can emerge from the selection of certain *k* values (or regions in *k* space) as the conventional signalling values of the language. Other dialects of English, or different languages, could divide the *k*-scale differently.

We can also interpret A to represent the level of motor factors leading to a rise in F0 (e.g., Cricothyroid muscle) and I to be the level of activation of motor factors leading to a lowering of F0 (e.g., Thyroarytenoid muscle). The coupled system of differential equations represents the internal structure of the motor system interaction through activation and inhibition, as we know muscular systems to be organized since Sherrington's work [27]. We propose that the motor system is phonologically parameterized by *k*, and that values of *k* are specified at least in part by the prominence setting system. Examination of other languages could also reveal the need for different constants in the equation, or even more or fewer terms in the equation, while remaining in the Activator-Inhibitor framework. Therefore, the model is a non-dualistic model where cognition and the motor system interact directly in a language-dependent way.

5. CONCLUSION

We have proposed a coupled nonlinear system of differential equations representing the motor-system governing pitch control, with one phonological parameter whose variation leads to phonetic F0 trajectories representing pitch accents. It's possible that exposure to a dialect of a language leads to a few regions of the scale being more prominent than others, resulting in dialectal variation of the pitch accent system within a language [22]. Finally, though the system we have proposed represents the same motor system that all humans possess, there can be phonologically induced differences in the terms and parameters of the interaction, while maintaining the A/I interaction, leading to different systems of pitch variation in different languages.

Acknowledgments: Project supported by NSF BCS-1944773 to JC and PAC-1246750 to KI.

6. REFERENCES

[1] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT. Distributed 1988, Indiana University Linguistics Club.

[2] Beckman, M., Hirschberg, J., Shattuck-Hufnagel, S. 2005. The original ToBI system and the evolution of the ToBI framework. In: Jun, S-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 9–54.

[3] Bolinger, D.L. 1952. Intonation: Levels Versus Configurations. *Word*. 7, 199–210.

[4] Bolinger, D.L. 1961. *Generality, Gradience, and the All-or-None.* Mouton.

[5] Kohler, K. 2005. Timing and communicative functions of pitch contours. *Phonetica*. 62, 88–105.

[6] Cole, J. 2015. Prosody in Context: A review. *Language, Cognition, and Neuroscience*, 30, 1–31.

[7] Barnes, J., Burgos, A., Veilleux, N, Shattuck-Hufnagel, S. 2021.On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics,* 85, 1–20.

[8] Kochanski, G., Grabe, E., Coleman, J, Rosner, B. 2005. Loudness predicts prominence; fundamental frequency lends little. *JASA,* 118, 1038-1054.

[9] Grabe, E., Kochanski, G., Coleman, J. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech,* 50, 281-310.

[10] Reichel, U. D. 2017. CoPaSul Manual: Contour-based, parametric, and superpositional intonation stylization. *arXiv:1612.04765*.

[11] Prom-on, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *JASA,* 125, 405-424.

[12] Sorenson, T., Gafos, A. 2016. The Gesture as an Autonomous Nonlinear Dynamical System. *Ecological Psychology,* 28, 188-215.

[13] Iskarous, K. 2017. The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics,* 64, 8-20.

[14] Roessig, S., Mücke, D., Grice. M. 2019. The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLOS ONE,* 15, e0231221.

[15] Roessig, S., Mücke, D. 2019. Modeling Dimensions of Prosodic Prominence. *Frontiers in Communication,* 4, 10.3389.

[16] Kelso, J.A.S. 1994. The informational character of self-organized coordination dynamics. *Human Movement Science,* 13, 393-413.

[17] Izhikevich, E. 2010. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting.* MIT.

[18] Hogan, N., Sternad, D. 2007. On rhythmic and discrete movements: reflections, definitions and implications for motor control. *Experimental Brain Research,* 4, 181, 13-30.

[19] Ladd, R. 2022. The Trouble with ToBI. In: Barnes, J. Shattuck-Hufnagel, S. (eds), *Prosodic Theory and Practice*. MIT.

[20] Perrier, P., Abry, C., Keller, E. 1988. Vers une modelisation de la langue. *Bulletin de la Communication,* 2, 181, 45-63.

[21] Iskarous, K., Pouplier, M. 2022. Advancements of phonetics in the 21st century: a critical appraisal of time and space in Articulatory Phonology. *Journal of Phonetics,* 2, 95, 101195.

[22] Cohen, J. 1988. *Statistical Power nalysis for the Behavioral Sciences.* Routlage.

[23] Goldsmith, J. 1974. *English as a tone language*. Distributed in 1981 in D. Goyvaerts (ed). *Phonology in the 1980's. Story-Scientia.*

[24] Turing, A. 1952. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society B,* 237.

[25] Fitzhugh, R. 1961. Impulse and physiological states in models of nerve membrane. *Biophysical Journal,* 1, 445-466.

[26] Amari, S-i. 1977. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics,* 27, 77-87.

[27] Sherrington, C. 1932. Inhibition as a Coordinative Factor. *Nobel Prize Lecture*.

[28] Burdin, R.S., Holliday, N., Reed, P.E. 2022. American English pitch accents in variation: Pushing the boundaries of mainstream American English. *Journal of Phonetics,* 2, 94, 101163.

[29] Steffman, J., Shattuck-Hufnagel, S., Cole, J. 2022. The rise and fall of American English pitch accents: Evidence from an imitation study of rising nuclear tunes. Proceedings of Speech Prosody 2022, Lisbon.

[30] Cole, J., Steffman, J., Shattuck-Hufnagel, S., Tilsen, S. (in press, 2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. Laboratory Phonology.

Differential effects of prosodic boundary on glottalization of word-initial vowels in Korean

Sungwok Hwang[1], Sahyang Kim[2], and Taehong Cho[1]

Hanyang Institute for Phonetics & Cognitive Sciences of Language (HIPCS), Hanyang University[1], Hongik University[2]
sungwokhwang@gmail.com, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study explores how the voice quality of word-initial vowels is modified by prosodic boundary in Korean. Prosodic boundary effects were examined where the boundaries were aligned with syntactic junctures, resolving structural ambiguity of coordinate structures (*Noun1-hako Noun2 animjʌn Noun3*, 'N1-and N2 or N3'). Results showed asymmetrical effects depending on whether the test word was a noun or a function word. Vowels tended to be glottalized *less* for N2 but *more* for *animjʌn* ('or') phrase-initially (at the Intonational-Phrase boundary) than phrase-medially. Notably, however, in the *less* glottalized case, there was much larger preboundary lengthening, compared to the *more* glottalized case. These results indicate that word-initial vowels are not necessarily glottalized to provide an invariant prosodic boundary marker. Rather, the modification of voice quality, we suggest, is modulated systematically by an optimization of the phonetics-prosody interplay that takes into account the relative contribution of available prosodic/segmental cues to signal prosodic structure.

Keywords: prosodic boundaries, glottalization, syntactic junctures, structural ambiguity

1. INTRODUCTION

Glottalization of word-initial vowel occurs in many languages. In general, it occurs either in a form of epenthetic glottal stop or with laryngealized (glottalized) voicing during the vowel [cf. 8]. Even in a language in which a glottal stop has a phonemic status, the vowel-initial allophonic glottalization still occurs often at a prosodic boundary, which may serve as a prosodic boundary marker [5, 12, 13] (See also [7] for further discussion on prominence as an important factor that induces glottalization.). For example, Mitterer et al. [12] suggested that in Maltese a glottal stop occurs either as a phoneme or as an epenthetic IP juncture correlate, such that Maltese listeners more likely perceive the glottalized sound as an epenthetic boundary marker when the preceding temporal structure as reflected in preboundary lengthening supports the percept of prosodic boundary. This study, though it focused on the perceptual aspects, implies that speakers produce various phonetic reflexes of prosodic juncture, whether suprasegmental or segmental, that marshal synergistically into signaling prosodic juncture, which are also reflected in the way the listener exploits the available cues to prosodic structure. Mitterer et al. [13] further revealed that word-initial glottalization could serve as a phonetic reflex of prosodic juncture that is often mapped onto a major syntactic juncture as a means of resolving syntactic ambiguity [e.g., 3, 9, 18].

The role of glottalization, however, is not limited to marking a large prosodic juncture such as an Intonational Phrase. But it may also be used to avoid vowel hiatus either within a word for marking a syllable boundary [e.g., 3] or across a word boundary thus delimiting a lexical boundary within a phrase [e.g., 4, 8]. Moreover, Garellek [7] further suggests that glottalization comes more consistently with prominence rather than with prosodic boundary especially in English, a head-prominence language [cf. 10]. Thus, a question remains open as to how glottalization is used to mark prosodic structure in a given language. In addition, given that the way that the syntax-prosody interface is actually realized may also vary depending on the language's syntactic and prosodic structure [cf. 6], it also remains open how phonetic reflexes of prosody including glottalization may be used in relation to syntactic disambiguation.

The present study uses Korean to explore how prosodic boundaries are realized in terms of temporal expansion and glottalization at a prosodic boundary and how these temporal and segmental reflexes of prosodic juncture are related to signalling syntactic junctures. To examine the syntax-prosody alignment, structurally ambiguous coordinate structures ([A] and [B or C] vs. [A and B] or [C]) will be used, which will provide different prosodic phrasings (with different prosodic boundaries) in relation to syntactic structure. Preboundary lengthening and the pause duration will be examined for the boundary-related temporal expansion, and word-initial vowels will be examined for the boundary-related glottalization that will be estimated by a spectral tilt measure (H1*-H2*) and HNR (Harmonic-Noise Ratio).

2. METHOD

2.1. Participants and speech materials

Seven males and seven females in their 20's who had no residential experience in the area other than Seoul Korea were recruited.

There were five test words: three proper nouns (person's names; /ali/, /aʧi/, /ami/) and two function morphemes (/-hako/ '-and' as a suffix, /animjʌn/ 'or' as a free morpheme). The nouns had the same vowel sequence (/aCi/). Each word appeared in six types of sentences which were answers to a question in a mini dialogue created for the purpose of the present study (see Table 1).

The three nouns (N1, N2, N3 crossed with a factorial combination) were used to create different coordinate structures that are structurally ambiguous:

(1) a. Early Closure: *(N1-hako) (N2 animjʌn N3)*
b. Late Closure: *(N1-hako N2) animjʌn (N3)*
glosses: *-hako* (‘-and’), *animjʌn* (‘or’)

In (1a), a major syntactic juncture comes earlier after *-hako* (‘-and’), which is suffixed to N1 (Early Closure), and in (1b), a major syntactic juncture comes later after N2 and before *animjʌn* (‘or’) (Late Closure). Moreover, as exemplified in Table 1, we used three different question types, which were meant to generate various target (answer) sentences with different information structures. The first question type was *what’s happening?*, so that the answer would be produced with ‘broad’ focus (BF). The second question type was *who’s coming to the party*, so that all three nouns (as person’s names) together would receive ‘narrow’ focus. And the third question type was *Did you say (N1 and N2) or N3 are coming to the party?* In this case, the question was meant to indicate the Late Closure structure (pre-recorded), but the speaker was instructed to correct the sentence to have the Early Closure structure to be produced with ‘contrastive’ focus (see below for more details). Note that the contrast in this case is not the one between specific lexical items, but the one between two different syntactic parsings, so that, for example, a Late Closure parsing was corrected to be an Early Closure parsing by the speaker.

Table 1. An illustration of test sentences. Each content word (/ali/, /atʃi/, /ami/) could be placed in all three locations: Noun1 (N1), Noun2 (N2), Noun3 (N3).

(a) Info= BF	Q: musin ilija “What’s happening?” A: *a N1-hako N2 animjʌn N3-ka ontɛ* “Well, N1, and N2 or N3 are coming.”
(b) Info= NF	Q: ipʌn ʧʰukʧɛ: nuka ontɛ “Who’s coming to the party?” A: a N1-hako N2 animjʌn N3-ka ontɛ “Well, N1, and N2 or N3 are coming.”
(c) Info= CF	Q: mwʌ IP N1-hako N2 IP animjʌn N3-ka ʧʰukʧɛ: ontɛ “Did they say N1 and N2, or N3 are coming to the party?” A: ani N1-hako N2 animjʌn N3-ka ontɛ “No, they said N1, and N2 or N3 will.”

2.2. Procedure

In each trial, the participants were presented with a visual display on the computer screen indicating an ‘intended’ coordinate structure to be used in an answer to the question which was pre-recorded by a female native speaker of Seoul Korean.

Participants heard the question first, and read the corresponding target sentence, guided by the visual cue to the underlying coordinate structure. Fig. 1 showed a case where the question type was *What’s happning?*. Given that N1 alone was inside a circle, while N2 and N3 were grouped together by a circle, the answer was meant to be produced with an Early Closure coordinate structure.

The data were recorded in a soundproof booth with a Tascam HC-P2 digital recorder and a SHURE KSN44 microphone at a sampling rate of 44kHz. After the recording, the authors labelled target words and checked their prosodic boundary conditions. In total, 1512 tokens were collected (2 syntactic structures x 3 focus conditions x 6 word orders x 3 repetitions x 14 speakers), but 1428 tokens were used for analysis, with an exclusion of 84 tokens whose renditions did not sound natural, as confirmed by all three authors.

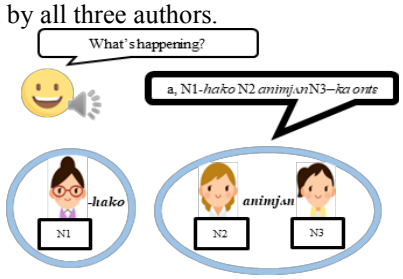


Fig. 1. An illustration of the visual display along with an auditory stimulus as a prompt question which was pre-recorded. This example was used to induce a broad focus with an Early Closure target sentence.

2.3. Measurements and statistical analyses

For N2, /animjʌn/, and N3, word-initial vowel’s glottalization was assessed by H1*-H2* and HNR. In H1*-H2*, H1 and H2 refers to the first and second harmonics in the spectral domain, respectively. H1*-H2* is gained by subtracting the amplitude of H2 from the one of H1. The asterisk (*) means that the value was corrected for formants. Lower H1*-H2* values indicate greater glottalization. HNR refers to the harmonic-to-noise ratio, with lower HNR values indicating non-modal voice quality. Mean values of H1*-H2* and HNR were taken from three timepoints corresponding to three equally-divided portions of the whole vowel length. H1*-H2* and HNR values were extracted by using Voice Sauce [16, 17]. For /-hako/ (‘-and’), N2, and /animjʌn/ (‘or’), the acoustic duration of the last syllable and the temporal gap (pause) between the offset of the last syllable and the onset of the following word were measured using Praat [2].

The statistical analyses were carried out by using the *lme4* [1] packages in R [14] to investigate the influences of prosodic boundary on the four dependent variables: the duration of the last syllable and pause, H1*-H2*, and HNR. The fixed effects were Boundary (Wd vs. IP), Focus (BF vs. NF vs. CF), and Timepoint (1st vs. 2nd vs. 3rd). The underlined categories above were the reference level, and all factors were contrast-coded. The random-effect structure by speaker included random intercepts and random slopes for Boundary, Focus, and their interaction. Random slopes for Timepoint were not involved because the models failed to converge with those slopes. As Timepoint was incorporated as a factor primarily to examine how Boundary would interact with Timepoint, random slopes for Timepoint were excluded for the sake of convergence. The random effect by word order was also included in the model because a single set of the three content words (N1, N2, N3) was considered as one item. The random structure by word order took in only random intercepts because of the convergence issue.

3. RESULTS

3.1. Prosodic phrasing and syntax-prosody mapping

All the sentence tokens reported in the present study could be divided into three prosodic phrasing groups as shown in Fig. 2. Each group showed a consistent syntax-prosody

alignment. The Early Closure construct was produced always with Phrasing Type 1 (100%), with a major syntactic juncture aligned with an IP boundary (#) between /-hako/ and N2. For the Late Closure construct, it was produced mostly with Phrasing Type 2 (81%), but often with Phrasing Type 3 (19%). An invariant phrasing aspect in line with the syntactic juncture was that for the Early Closure, an IP boundary was placed always after an function morpheme (*-hako* 'and') and for the Late Closure, it was placed always before a function morpheme (*animjʌn*, 'or') with an optional boundary after it.

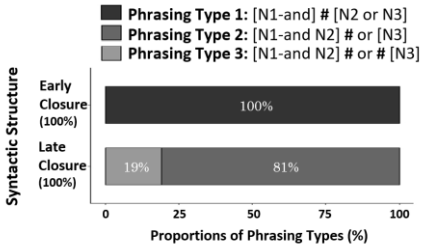


Fig. 2. Distribution of phrasing types for Early and Late Closure.

3.2. The first IP juncture between /-hako/ ('-and') and N2 (*N1-and # N2 or N3*)

Boundary had significant effects on the acoustic duration of the final syllable of /-hako/ (β=84.15, t=17.40, p<0.001) and pause (β=135.97, t=65.85, p<0.001). As can be seen in Fig. 3a-b, both the duration of /ko/ and the pause were significantly longer in IP-final than in IP-medial condition (by ca. 169ms and ca. 260ms, respectively). When these two measures were combined, there was a temporal expansion of approximately 420ms from IP-medial to IP-final condition. For the initial /a/ of N2, there was no clear sign of the difference in glottalization between the IP-initial and the IP-medial position, as can be inferred from Fig. 3c-d. For H1*-H2*, there was no Boundary effect (β=0.30, t=1.33, p=0.207), but Boundary interacted with Timepoint (β=0.52, t=8.29, p<0.001). As shown in Fig. 3(c), the value of the 2nd and 3rd timepoints tended to be higher in IP-higher than in IP-medial position (2nd: β=0.40, t=1.84, p=0.090; 3rd: β=0.80, t=2.15, p=0.051), indicating that, if there is any, the initial vowel of N2 is on a less glottalized (breathier) side. On the other hand, HNR showed a Boundary effect (β=-2.70, t=-7.99, p<0.001), indicating that the word-initial vowel was noisier IP-initially than IP-medially.

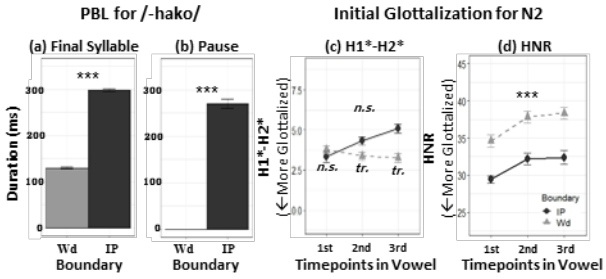


Fig. 3. Effects of boundary at the first IP juncture (*N1-and # N2 or N3*) on PBL for /-hako/ and pause duration (a, b), and glottalization of initial /a/ of N2 (c, d). *n.s.*, p>0.1; *, p<0.05; **, p<0.01; ***, p<0.001. Error bars refer to standard errors.

3.3. The second IP juncture between N2 and /animjʌn/ (*N1-and N2 # or N3*)

There was a significant effect of Boundary on the acoustic duration of the second (final) syllable of N2 (β=67.96, t=15.09, p<0.001) and pause (β=104.47, t=57.32, p<0.001). As shown in Fig. 4a-b, both the duration of the last syllable of N2 before the boundary and the pause were significantly longer at an IP boundary than at an Wd boundary (by ca. 120ms and ca. 210ms, respectively). When the two temporal measures were combined, there was a temporal expansion of approximately 330ms from the Wd boundary to the IP boundary. Compared with the boundary-related temporal expansion (420ms) near /-hako/ ('-and') (preboundary lengthening + pause), this temporal expansion (preboundary lengthening of N2 + pause at an IP) was relatively smaller.

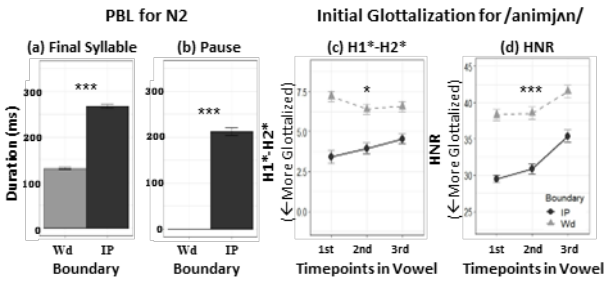


Fig. 4. Effects of boundary at the second IP juncture (*N1-and N2 # or N3*) on preboundary lengthening (PBL) of N2 and pause duration (a, b), and glottalization of initial /a/ of /animjʌn/ ('or') (c, d). *n.s.*, p>0.1; *, p<0.05; **, p<0.01; ***, p<0.001. Error bars refer to standard errors.

However, this time, as can be seen in Fig. 4c-d, there was a clearer indication of variation of glottalization as a function of boundary for the initial /a/ of /animjʌn/ ('or else'). There was a significant effect of Boundary on H1*-H2* (β=-1.40, t=-2.86, p=0.013)—i.e., H1*-H2* was smaller IP-initially than IP-medially (Fig. 4c), indicating that the initial vowel of /animjʌn/ was more glottalized IP-initially. There was also a significant Boundary effect on HNR (β=-3.55, t=-5.92, p<0.001)—i.e., HNR was lower IP-initially than IP-medially (Fig. 4d), indicating that the spectral harmonics were less clear (noisier) IP-initially, in line with the (non-modal) glottalization pattern.

3.4. The third IP juncture between /animjʌn/ and N3 (*N1-and N2 or # N3*)

Boundary had significant effects on the acoustic duration of /mjʌn/ (the final syllable of /animjʌn/) (β=31.99, t=3.38, p=0.008) and pause (β=13.97, t=21.86, p<0.001). As can be seen in Fig. 5a-b, both the duration of /mjʌn/ and the pause were significantly longer at an IP boundary than at an Wd boundary (by ca. 70ms and 30ms, respectively). When these measures were combined, there was a total temporal expansion of approximately 100ms from the Wd to the IP boundary condition. Compared to the other IP junctures (which showed a total temporal expansion of 420~330ms), this IP juncture was marked by a far less temporal expansion. Nevertheless, for the post-boundary initial /a/ of N3, there was a clear variation in glottalization as a function of boundary. For H1*-H2*, there was no significant effect of Boundary but there was a significant interaction between Boundary and Timepoint (with 1st/2nd

timepoint: β=0.66, t=3.049, p=0.002; with 1st/3rd timepoint; β=-1.07, t=-5.00, p<0.001). As can be seen in Fig. 5c-d, the interaction was due to the fact that there was a significant Boundary effect at the first timepoint (the beginning of the vowel) with a lower H1*-H2 (more glottalized) at the IP-initial position (β=-1.29, t=-2.96, p=0.016), while the effect disappeared into the vowel at the following timepoints. For HNR, there was a significant Boundary effect (β=-1.51, t=-3.42, p=0.011), indicating a substantially lower HNR of the vowel in the IP-initial than in the IP-medial position.

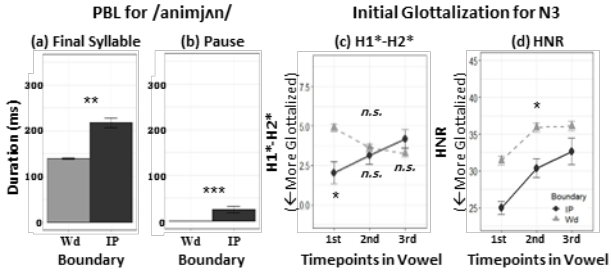


Fig. 5: Effects of boundary at the 'optional' third IP juncture (*N1-and N2 or # N3*) on preboundary lengthening (PBL) of /animjʌn/ and pause duration (a, b), and glottalization of initial /a/ of N3 (c, d). *n.s.*, p>0.1; *, p<0.05; **, p<0.01; ***, p<0.001. Error bars refer to standard errors.

4. SUMMARY AND DISCUSSION

The present study has examined how major syntactic junctures of coordinate structures are aligned with major prosodic boundaries to be used for resolving syntactic ambiguity in Korean and how the major prosodic boundary (IP) is phonetically expressed not only in the suprasegmental dimension of preboundary lengthening and pause duration, but also in the glottalization of post-boundary vowel that is definable in the segmental dimension.

One of the basic findings was that the major syntactic juncture was unequivocally aligned with the major prosodic boundary—i.e., the Intonational Phrase boundary. That is, for an Early Closure construct ([N1-and] # [N2 or N3]), the syntactic juncture ('#') was consistently aligned with an IP boundary; for a late closure construct ([N1-and N2] # or (#) [N3]), the first syntactic juncture '#' before 'or' was consistently aligned with an IP, the second juncture '(#)' was only occasionally (thus *optionally*) aligned with an IP. Such a syntax-prosody mapping means that the major syntactic juncture was expressed by phonetic reflexes of prosodic juncture including a temporal expansion reflected in preboundary lengthening and the pause duration.

But our results further indicated that although the Intonational Phrase may be phonologically defined, governed by the intonational phonology of the language [11], the phonetic reflexes of the same IP boundary could be realized in a gradient fashion. Recall that the magnitude of the temporal expansion associated with an IP boundary was substantially greater for the first IP juncture before N2 of Phrasing Type 1 than for the second IP juncture before 'or' of Phrasing Types 2 and 3. One possible reason may be related to a phonetic declination of phrase-final lengthening. Since the IP of Phrase Type 1 occurs earlier in the utterance, its temporal expansion could possibly be larger compared to when an IP occurs later in the utterance as in Phrasing Types 2 and 3. But it is also possible that preboundary lengthening can be controlled by the speaker as discussed in the preboundary lengthening literature [15, 19]. If so, the differential effects of the phonologically defined IP may well be interpreted to have come about as a consequence of the speaker control. Recall that the occurrence of an IP in [N1-and] # [N2 or N3] is invariantly aligned with a major syntactic juncture. From the listener's point of view, the syntactic parsing is constructed as the speech unfolds over time [e.g., 13, 18]. Thus, the presence or absence of an early IP should play a pivotal role in determining whether the utterance is meant to be an early versus a late closure construct. The speakers may therefore produce the prosodic juncture with a more heightened phonetic clarity through an augmented temporal expansion to clearly signal the critical syntactic juncture. There is yet a third possibility that may not be mutually exclusive from the speaker control account. That is, because the excessive preboundary lengthening before the first IP was observed with the function morpheme *–hako* ('-and'), one might assume that it has to do with the functionality of the morpheme. But the fact that it was not only the preboundary lengthening but also the pause duration that showed a substantial difference takes the support away from the third possibility.

The same account may apply to the differential effects between the second IP and the third IP. Recall that the temporal expansion was much greater for the second IP before 'or' of Phrasing Types 2 and 3 than for the third IP before N3 of Phrasing Type 3. For a late closure construct, the syntactic juncture was consistently aligned with the second IP, and only optionally with the third IP. In this case as well, the speaker could deliberately augment the strength of the prosodic juncture in reference to syntactic structure.

Finally, our results showed that an IP boundary does not always induce a substantial glottalization for the word-initial vowel as compared to a smaller prosodic boundary (i.e., in the phrase-medial position). Interestingly, however, the boundary-related glottalization appears to be correlated with the magnitude of temporal expansion. For the first critical IP for which the most robust temporal expansion was found, no additional boundary-related effect was observed with glottalization. But for the second IP, the following vowel of /animjʌn/ ('or') was substantially more glottalized IP-initially. A similar pattern was observed for the (optional) third IP. These results imply that the glottalization is not invariantly employed in marking a prosodic boundary, but rather, the voice quality (including glottalization) is likely to be modulated by the motor system that takes into account relative contributions of available suprasegmental and segmental (glottalization) cues to signal prosody structure.

In conclusion, our results indicate that prosodic structure that may be constructed based on phonologically determined prosodic categories is fleshed out with the gradient phonetic content in both segmental and suprasegmental dimensions. The phonetic content appears to be the outcome of a fine-tuning of phonetic encoding of prosodic structure in reference to syntactic factors as well as system-driven factors.

5. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

6. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. "Fitting linear mixed-effects models using lme4," J. Stat. Soft. 67(1), 1-48.

[2] Boersma, P., and Weenink, D. 2020. "Praat: Doing phonetics by computer" [computer program], version 6.0.40, http://www.praat.org/ (Last viewed 29 April 2020).

[3] Davidson, L., & Erker, D. 2014. Hiatus resolution in American English: The case against glide insertion. *Language* 90(2), 482-514

[4] Davidson, L. 2021. Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian. *Journal of phonetics* 88, 101075.

[5] Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of phonetics* 24(4), 423-444.

[6] Elfner, E. 2018. The syntax-prosody interface: Current theoretical approaches and outstanding questions. *Linguistics Vanguard* 4(1), 20160081.

[7] Garellek, M. 2014. Voice quality strengthening and glottalization. *Journal of Phonetics* 45(1): 106-113

[8] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94: 101155.

[9] Jun, S.-A. 2009. Prosody in sentence processing. In Li, P., Lee, C., Simpson, G., & Kim, Y. (Eds.), *Handbook of East Asian Psycholinguistics, Part III: Korean Psycholinguistics* (pp.423-432). Cambridge: Cambridge University Press.

[10] Jun, S.-A., editor 2014. *Prosodic Typology II: The phonology of intonation and phrasing*. Oxford University Press.

[11] Ladd, D. R. 2008. *Intonational phonology* (second edition). Cambridge: Cambridge University Press.

[12] Mitterer, H., Kim, S., & Cho, T. 2019. The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language* 108, 1-19.

[13] Mitterer, H., Kim, S., & Cho, T. 2020. The role of segmental information in syntactic processsing through the syntax-prosody interface. *Language and Speech* 64(4), 962–979

[14] R Core Team. 2022. "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, https://www. R-project.org/ (Last viewed 30 November 2022).

[15] Seo, J., Kim, S., Kubozono, H., & Cho, T. 2019. Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter? *Journal of the Acoustical Society of America* 146(3), 1817-1823.

[16] Shue, Yen-Liang 2010. The voice source in speech production: Data, analysis and models. Los Angeles: University of California. http://www.phonetics.ucla.edu/voiceproject/ Publications/shue_dissertation.pdf

[17] Shue, Y.-L., Keating, P., Vicenik, C. & Yu, K. 2011. Voicesauce: A program for voice analysis. In Wai-Sum Lee & Eric Zee (eds.), Proceedings of the 17th International Congress of Phonetic Sciences, 1946–1849. Hong Kong.

[18] Snedeker, J., & Trueswell, J. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language* 48, 103–130.

[19] Turk, A., & Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4), 445–472.

SPEECH RHYTHM METRICS: A TYPOLOGICAL SURVEY

Matthew K. Gordon, Argyro Katsika, Sherry Chien, Jiyoung Jang, Ryan Ka Yau Lai

University of California, Santa Barbara
mgordon@linguistics.ucsb.edu, argyro@ucsb.edu, sherrychien@ucsb.edu, jiyoung@ucsb.edu, kayaulai@ucsb.edu

ABSTRACT

Several hypothesized parameters diagnosing the distinction between syllable-timing and stress-timing were examined for recordings of the *North Wind and Sun* fable for a set of 10 genetically, areally, and structurally diverse languages. Measured properties included the following: duration of interstress intervals, ratio of stressed-to-unstressed vowel and syllable durations, variability in the duration of intervocalic and vocalic intervals, and the standard deviation of consonant and vocalic intervals. Results suggest greater typological diversity than in previous studies and a lack of convergence across diagnostics typically employed to diagnose speech rhythm. This diversity is plausibly due to interlanguage variation in syllable complexity and word length, the latter of which is attributed to differences between languages in their degree of morphological synthesis. Nevertheless, certain parameters suggest rhythmic categories; however, potentially more than just a binary division between syllable- and stress-timing.

Keywords: rhythm, syllable-timed, stress-timed

1. INTRODUCTION

The study of acoustic correlates of rhythm has been a productive area of research since the identification of a perceptual dichotomy between syllable-timed and stress-timed languages [1, 2]. Research has examined a number of potential quantitative metrics for diagnosing the distinction between the two categories but no single property has been demonstrated to reliably distinguish syllable- and stress-timing across a broad sample of languages. More generally, evidence for a typologically robust perceptual dichotomy between syllable- and stress-timed languages is lacking [3]. Rather, languages fall along a continuum capped at either end by canonical prototypes forming the basis for the hypothesized categorical division, e.g., English (stress-timed) vs. Spanish (syllable-timed). Furthermore, the distance between stresses, the "interstress intervals", tend to vary little across rhythmic types with most intervals ranging from 300-700 milliseconds [4].

This paper aims to enrich the research program on rhythm classes in two ways. First, we extend the typological purview of investigation to include a broader sample of areally and genetically diverse languages than have previously been examined. By expanding the breadth of cross-linguistic coverage, we aim to gain more reliable insight into the range of variation in rhythmicity and the factors that contribute to it. Second, we explore the effects of not only syllable structure examined in previous work (e.g., [5, 6, 7]) but also the length of morphological (grammatical) words on commonly adopted acoustic correlates of rhythm. A language's characteristic word length, calculated here in terms of number of syllables per word, potentially exerts a substantial effect on its rhythmic profile due to the one-to-one mapping between morphological words and stresses in many languages, a mapping that creates relatively long interstress intervals for long words.

2. METHODOLOGY

2.1. Languages and Corpus

A set of 10 genetically and areally diverse languages were selected for which there exists a JIPA Illustration containing a recording of the *North Wind and the Sun* fable. This fable has been used to quantity rhythm in previous studies [7] since it provides a comparable sample (in length and style) of phonetically transcribed naturalistic speech across languages. The languages in our sample all are described as having stress-based prosodic systems but differ in two important respects germane to rhythm. They permit differing degrees of syllable complexity and they represent varied morphological subtypes along a continuum ranging from relatively analytic, e.g., Kalasha, to highly agglutinative, e.g., Yine; this morphological variation leads to corresponding variation in word length.

The list of languages, their geographic locations and genetic affiliations, mean syllable complexity (in number of segments per syllable) and mean morphological word length (in number of syllables per word) is provided in Table 1.

2.2. Measurements

The North Wind and the Sun recordings were parsed into segments, syllables, and sentences using text grids in Praat. Surface syllabification conventions (including epenthetic vowels) were followed and

words were parsed following the orthographic breaks; words were thus morphological rather than prosodic. Sentences were divided based on punctuation if marked in the Illustration, and otherwise on the basis of major prosodic breaks. Long vowels and diphthongs were treated as unitary segments while geminate consonants were treated as two segments when split by a syllable boundary.

Language	*Country*	*Family*	*Seg/σ*	*σ/Wd*
Abkhaz (Cwyzhy)	Turkey	N. Caucasian	2.32	2.80
Arrernte (Central)	Australia	Australian	2.08	3.00
Estonian	Estonia	Uralic	2.65	2.00
Gitksan	British Columbia	Tsimshianic	2.74	1.58
Kalasha (Bumburet)	Pakistan	Indo-European	2.19	1.93
Kedayan	Malaysia	Austronesian	2.36	2.14
Nen	New Guinea	Yam	2.33	2.22
Quichua (Salasaca)	Ecuador	Quechuan	2.19	3.07
Seri	Mexico	isolate	2.56	1.53
Yine	Peru	Maipurean	2.41	3.83

Table 1: Languages in the study

Several measurements commonly used to quantify rhythm were extracted. Interstress Intervals (ISI) were measured from the onset of each stressed syllable up to but not including the onset of the next stressed syllable. Primary stress, but not secondary stress, was indicated in the transcriptions for all languages. Thus, only primary stresses were included in the calculation of interstress intervals. Certain texts did not mark stresses for monosyllabic words; these were manually added for content words. In addition, the following measures were calculated, following [7]: the raw Pairwise Variability (rPVI) across Interstress Intervals, the duration ratio of stressed-to-unstressed syllables and vowels, and the standard deviations of the vocalic (ΔV) and consonantal (ΔC) phases of the sentences.

3. RESULTS

3.1. Interstress intervals, word and syllable length

The densest concentration of interstress intervals is similar across languages (Fig. 1), hovering between 300ms and 700ms, with the overall distribution of intervals varying greatly as a function of language. Gitksan, Seri, Abkhaz and Yine tend toward longer interstress intervals, where Yine shows particularly wide variation with several very long intervals. This variability is likely attributed in large part to the relatively large proportion of long words in Yine, which is observed in the interstress intervals by word length in syllables (Fig. 2). Conversely, Arrernte and Quichua show relatively little variation in interstress intervals despite having relatively long words, likely due to the counterbalancing effect of simple syllable structure (Fig. 3).

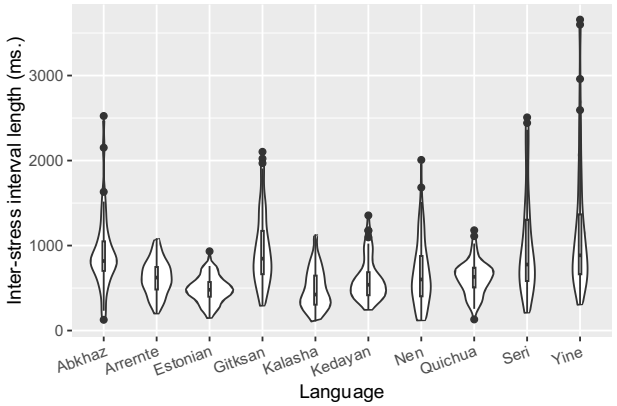


Figure 1: Interstress intervals for examined languages

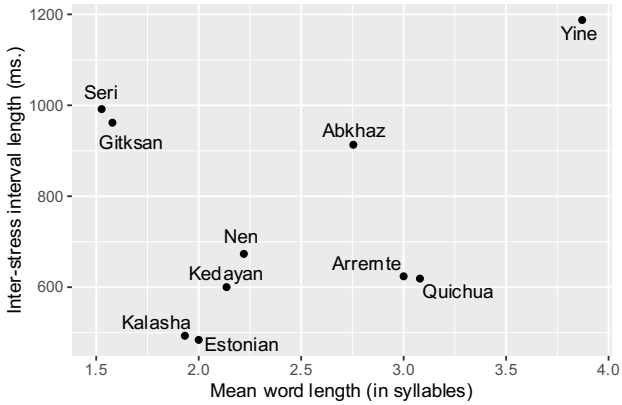


Figure 2: Mean interstress intervals by mean word length

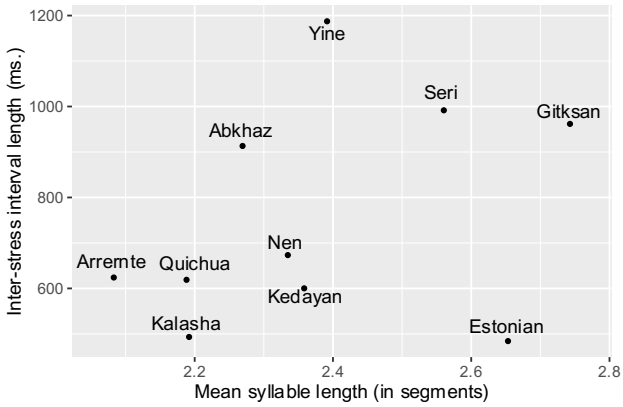


Figure 3: Mean interstress intervals by mean syllable length

The three languages with the characteristically shortest interstress intervals, Estonian, Kalasha and Kedayan, also have relatively short mean word length

in terms of syllable count. Estonian is an interesting case since it has relatively complex syllables but short words (despite being a language with moderately synthetic morphology); the short interstress intervals in Estonian thus suggest a greater weight for syllable count in predicting interstress intervals. Gitksan, though, has a similar level of syllable complexity and word length to Estonian but has much longer interstress intervals.

3.2. Relative duration and PVI measures

The duration ratio between the stressed and unstressed syllable intervals within an ISI is shown in Fig. 4 and between stressed and unstressed vowel intervals in Fig. 5. In both figures the average number of syllables per ISI in each language is at the top.

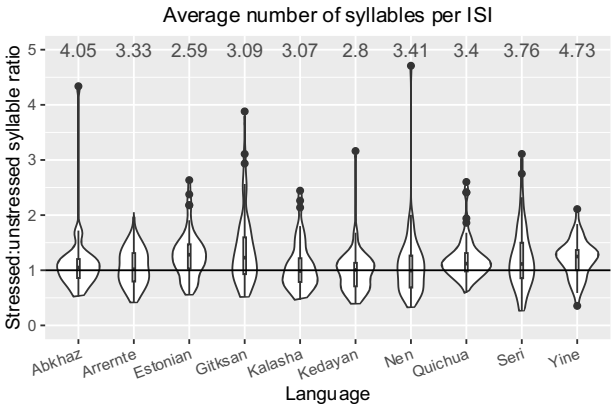


Figure 4: Stressed-to-unstressed syllable duration ratio within interstress intervals

In Fig. 4, most languages have distributions centered just above a one-to-one ratio, indicating that stressed portions of interstress intervals generally comprise more than half of the total duration of the interval. Yine has the longest mean ISI length in number of syllables and also relatively large stressed-to-unstressed interval ratios, and conversely, Kedayan has relatively few syllables per ISI in keeping with also having characteristically short stressed-to-unstressed interval ratios. There is not a consistent relationship, however, between number of syllables in an ISI and the proportion of the ISI that is stressed. Estonian thus has the fewest syllables per ISI but stressed-to-unstressed syllable interval ratios that are among the larger ones for the examined languages. Fig. 5 suggests that the duration ratio observed between stressed and unstressed syllables in Fig. 4 is generally attributed to vowels rather than consonants.

All surveyed languages have many data points reflecting *shorter* stressed than unstressed syllable and vowel intervals within an ISI. Although this finding may seem counterintuitive, it has at least two plausible sources. One is that unstressed syllables are longer due to segmental complexity, i.e., longer consonant clusters in unstressed syllables. Another factor could be final lengthening targeting unstressed domain-final syllables, a possibility given that all of the examined languages have non-final stress either predominantly or exclusively, and final syllables are longer than non-final ones in all languages examined except Quichua.

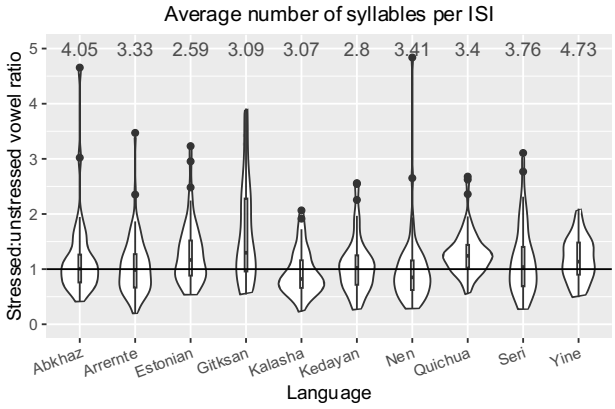


Figure 5: Stressed-to-unstressed vowel duration ratio within interstress intervals

Fig. 6 plots the duration ratio of stressed-to-unstressed syllables for interstress intervals with differing numbers of syllables in order to assess polysyllabic shortening, the tendency for individual segments to shorten in proportion to the number of syllables in an interstress interval [8, 9, 10]. It is difficult to compare languages differing in their word lengths but the clearest tendency toward polysyllabic shortening is observed in Gitksan.

Another diagnostic for the stress-timed vs. syllable-timed distinction is the Pairwise Variability Index, which is greater for stress-timed languages due to their tendency toward unstressed vowel reduction and their greater tolerance of syllable complexity. Fig. 7 plots raw PVI values for intervals of vocalic (i.e., just vowels) and intervocalic (i.e., just consonants) sounds in each language. Gitksan stands out as an outlier in terms of possessing unusually large rPVI values for both vocalic and especially intervocalic interstress intervals. The languages are otherwise well spread across a range of rPVI values for vocalic and intervocalic intervals.

Fig. 8 plots the standard deviation of the duration of consonant and vocalic intervals within each sentence. Gitksan is again a conspicuous outlier in possessing very large ΔC and ΔV values, while other languages again appear to form a continuum, with positive correlation between the two values. Higher vocalic rPVIs tend to be associated with higher ΔV values, and higher intervocalic rPVIs with higher ΔC. For example, Kedayan has high vocalic rPVI and ΔV

and moderate intervocalic rPVI and ΔC, while Yine has relatively high intervocalic rPVI and ΔC and low vocalic rPVI and ΔV.

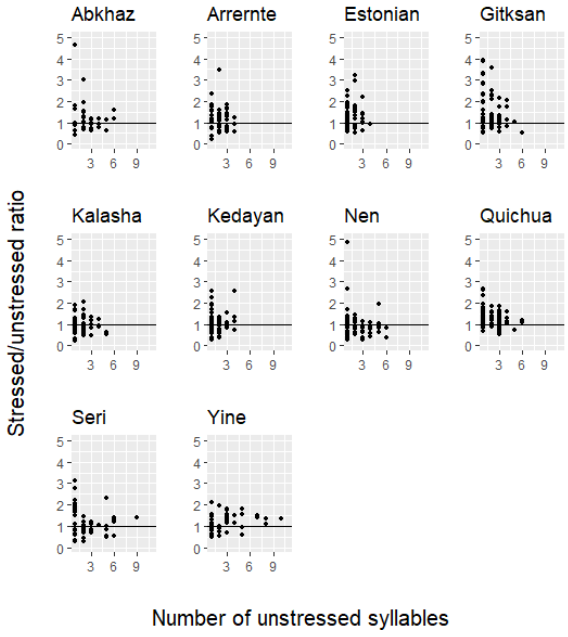


Figure 6: Duration ratio of stressed-to-unstressed syllables by number of syllables in interstress interval

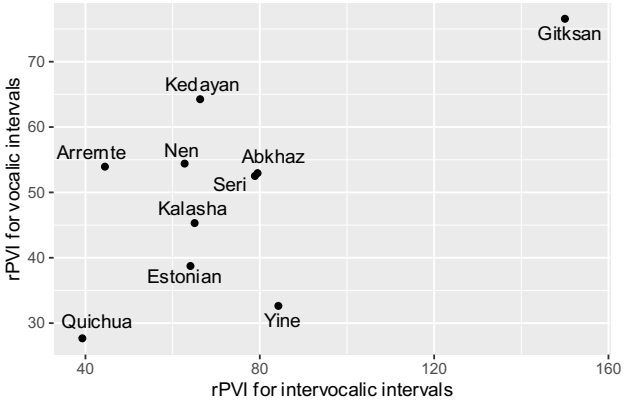


Figure 7: Raw pairwise variability index for intervocalic and vocalic intervals

4. DISCUSSION

Results overall suggest greater divergence between languages in their acoustic diagnostics of rhythm than found in previous research. Interstress interval durations are generally consistent with earlier work [4], although the current study suggests greater variation across and especially within languages, a difference that is likely due to differences in syllable complexity and length of words, the latter of which serves as a proxy for morphological complexity.

Metrics commonly used to quantify the impressionistic dichotomy between stress-timed and syllable-timed languages revealed some distinctions between languages but not consistently convergent ones. Quichua and Gitksan have relatively large durational differences between stressed and unstressed vocalic intervals, suggesting stress-timed rhythm. However, Quichua falls at the opposite end of the continuum from Gitksan in terms of vocalic and intervocalic rPVI, ΔC and ΔV with values that are more consistent with syllable-timing. This divergence may be explained in part by a difference between the two languages in the locus of complexity: Quichua has long words but relatively simple syllable structure whereas the opposite is true of Gitksan. Similar word length, however, does not ensure parallel rhythmic metrics since Yine, which has the longest words in the sample, diverges from Quichua in intervocalic rPVI.

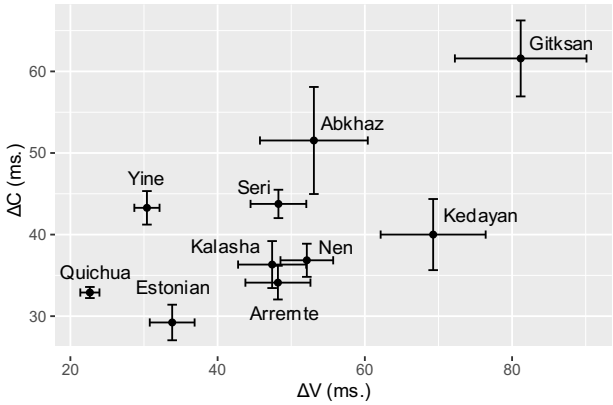


Figure 8: Mean ΔC vs. ΔV values with Student's *t*-test std. errors

The vocalic and intervocalic rPVI, ΔC and ΔV parameters do not show any clear clustering of languages, but a continuum, although there is a correlation between ΔC and ΔV observed. Some languages have low values for all four measures, a single language has extremely high indices for all four, and other languages have intermediate values. While the values for languages with low measures on all four are consistent with those characteristic of syllable-timed languages [7], values associated with the other languages are consistent with a stress-timed designation, suggesting that stress-timing might occupy a larger perceptual space, presumably attributed to the relative homogeneity (compared to the current sample) of languages forming the basis for the original impressionistic classification.

5. SUMMARY

An acoustic study of rhythm metrics in a structurally, genetically, and areally diverse set of languages indicates commensurate heterogeneity in the acoustic diagnostics of rhythm. Two categories, or even two ends of a unidimensional continuum, may be insufficient to capture the rhythmic typology.

6. ACKNOWLEDGMENTS

We thank the students (Gia Ahn, Gabriel Bernal, Jason Galaviz, Yan Lashchev, Katie Lee, Yujie Li, Jade Morton, Ella Peter, Kirra Sebastian, Lily Yin, Alice Zhang, Boyi Zheng, Dingyan Zhou) who assisted with the data analysis as well as the anonymous reviewers who provided feedback on the earlier version of the paper.

7. REFERENCES

[1] Pike, K. L. 1945. *The Intonation of American English*. University of Michigan Press.

[2] Abercrombie, D. 1967. Elements of General Phonetics. Edinburgh University Press.

[3] White, L., Malisz, Z. 2020. Speech rhythm and timing. In: Gussenhoven, C., Chen, A. (eds.), *Oxford Handbook of Language Prosody*. Oxford University Press, 167-182.

[4] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51-62.

[5] Dasher, R., Bolinger D. 1982. On pre-accentual lengthening. *Journal of the International Phonetic Association* 12, 58-71.

[6] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73(3), 265-292.

[7] Grabe, E., Low, E. L. 2002. Durational variability in speech and the Rhythm Class Hypothesis. In: Gussenhoven, C., Warner, N. (eds), *Laboratory Phonology* 7. De Gruyter Mouton, 514-546.

[8] Lindblom, B., Rapp, K. 1973. *Some temporal regularities of spoken Swedish*. Papers from the Institute of Linguistics, University of Stockholm 21.

[9] Meyer, E.A., Gombocz, Z. 1909. *Zur Phonetik der ungarischen Sprache*. Berling.

[10] Kim, H., Cole, J. 2005. The stress foot as a unit of planned timing: evidence from shortening in the prosodic phrase. *Interspeech* 2005, 2365-2368.

HOW DOES FOCUS-INDUCED PROMINENCE INFLUENCE REALIZATION OF EDGE TONES AND SEGMENTAL ANCHORING IN SEOUL KOREAN?

Richard Hatcher[1], Hyunjung Joo[1,2], Sahyang Kim[1,3], & Taehong Cho[1]

[1]Hanyang Institute for Phonetics & Cognitive Sciences of Language (HIPCS), Hanyang University, [2]Rutgers University, [3]Hongik University

richard.j.hatcher.jr@gmail.com, hyunjung.joo@rutgers.edu, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study investigated how focus-induced prominence influences the global F0 contour within an Accentual Phrase (AP) embedded in an Intonational Phrase in Seoul Korean. Since an AP typically contains edge tones (#LH…LH#), it was tested how realization of underlying LH is conditioned by focus. Monosyllabic target words occurred phrase-initially/finally, and the coda sonorancy was varied (*pam* vs. *pap*) to examine whether/how the tone bearing ability would influence tonal realization. Results showed that focus on phrase-initial words induced pitch range expansion (LH), evident in H-peak anchored early in the post-focal word. But phrase-finally focus induced no F0 expansion but slightly delayed L%, showing an interaction with boundary tones. The coda sonorancy yielded local F0 perturbation, but not global F0 modification. Our results suggest that the segmental anchoring and range of phonologically-defined tones are systematically modulated by prominence, illuminating an interplay between tonal and segmental realizations in reference to information structure.

Keywords: edge tones, focus-induced prominence, segmental anchoring, intonation, Korean

1. INTRODUCTION

Linguistic prominence is often realized by the intonational contour. Languages like English and German are classified as head-prominence languages, as linguistic prominence relies on pitch-accented syllables serving as head of the prosodic phrase [15, 14, 9] In these languages, when prominence is licensed by focus, it is realized in the speech signal by expanded pitch range on the focused (and pitch accented) elements and subsequent deaccenting or pitch compression on the post-focal elements [13]. Other languages, such as Korean, lack the accent system and are said to employ prosodic phrasing in order to mark prominence. In such an edge prominence system, focus is often assumed to trigger insertion of a prosodic boundary, with focused words aligned to the left edge of a prosodic phrase [10, 6].

Setting aside the theoretical issue of whether there should be a one-to-one mapping between focus and phrasing in Korean, it is clear that edges of a prosodic phrase are marked by edge tones of some sort in such an edge-prominence language. What is less clear, however, concerns whether and how prominence may influence the tonal realization of both edge tones and non-edge tones in shaping the global tune of an intonational structure in a language. The present study attempts to fill this gap by investigating the tonal manifestation of focus-induced prominence and its influence on the intonational landscape of the Accentual Phrase (AP) embedded in an Intonational Phrase (IP) in Seoul Korean.

The Korean AP is intonationally defined by two rising contours (/#LH…LH#) at its edges, although they may or may not be fully realized depending on the number of syllables within an AP as well as other (non-)linguistic factors. (While the identity of the initial tone is conditioned by the laryngeal status of the AP-initial segment, we will only focus on the initial L tone in this study assigned for sonorants and lenis obstruents.) As shown in Figure 1, an AP is embedded in a larger prosodic phrase, the Intonational Phrase. While the tonal composition of an AP directly contributes to the form of an IP, especially at its left edge, the IP-final boundary tone (marked by %) overrides the right-aligned AP edge tone.

Previous studies of prominence in Korean intonation broadly agree that focus (especially the narrow or contrastive focus) is expressed by phrasing, followed by optional dephrasing of any post-focal materials [5, 12, 4]. But analyses may differ on the precise level of the phrase boundary initiated by focus. There are accounts for the Accentual Phrase [5, 6], the Intermediate Phrase [7, 8], and the Intonational Phrase [4]. In addition,

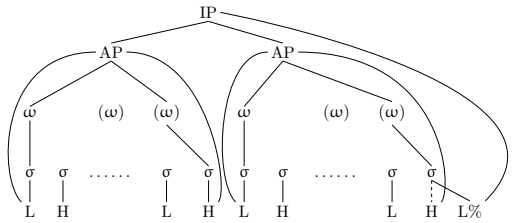


Figure 1: Intonational structure of Seoul Korean following Jun [6]

focus is also described as influencing the phonetic realization of phrasal tones, whereby focused H tones were realized with higher F0 [4].

The purpose of the present study is not to attempt to contribute to these theoretical accounts, but to provide more empirical data to illuminate the nature of the relationship between prominence and tonal realization. We therefore investigate in more detail how focus effects on the F0 contour are constrained by the intonational grammar of Korean. We specifically focus on how focus-induced F0 range expansion is further conditioned by the prosodic position of the focused elements, and how edge tones are anchored with segments and further interact with boundary tones under prominence.

2. METHODS

Data were collected from 14 native speakers of Seoul Korean (7 female, 7 male) in their twenties. Participants were required to have been born and reside in Seoul.

There were two monosyllabic target words varying in the sonority of their coda (*pam* "chestnut; night", *pap* "cooked rice"). Table 1 provides an example set of question-answer pairs for target word *pam*. In order to induce corrective focus, participants were asked to make a contrast between the words in bold in Sentences A (a question) and B (a corrective answer to A). To control for the possible effects of phrasing, target sentences (B in Table 1) were constructed to consist of three words[1] (5 syllables total) with the target word occupying either the phrase-initial or -final position. The phrase position of the test words were varied to investigate how the effects of F0 contour change in different prosodic position.

In the experiment, participants were presented with a mini dialogue via a visual aid provided on the computer screen. The participants were asked to read the target sentence (as Speaker B) in response to the prime sentence presented auditorily as well as visually. Participants were asked to produce the sentence naturally without pausing between the words.

Initial	Focus	A:	*ipʌn tanʌnin* ***kuk*** *twiɛta nwa?*
			Should I put this word behind the soup?
		B:	*ani.#****pam*** *twiɛta nwa.*
			No. Put it behind the **chestnut**.
	No Focus	A:	*ipʌn tanʌnin pam* ***apʰɛta*** *nwa?*
			Should I put this word on the chestnut?
		B:	*ani.#pam* ***twiɛta*** *nwa.*
			No. Put it **behind** the chestnut?
Final	Focus	A:	*ipʌn tanʌnin nʌhĩ ʌnni* ***kʰoŋini?***
			Is this one (for) your sister's bean?
		B:	*ani.#uri ʌnni* ***pam****.#twiɛta nwa.*
			No. Our sister's **chestnut**. Put it under.
	No Focus	A:	*ipʌn tanʌnin nʌhĩ* ***op*a*** *kʰoŋini?*
			Is this (for) your brother's chestnut?
		B:	*ani.#uri* ***ʌnni*** *pam.#twiɛta nwa.*
			No. Our **sister's** chestnut. Put it under.

Table 1: An example set of test words with varying focus and boundary conditions.

The recording was made in a sound-attenuated booth, using a SHURE KSM44 condenser microphone and a Tascam US-4x4 digital recorder at a sampling rate of 48 kHz. In total, 3360 tokens were collected (3 items x 2 boundaries x 2 focus types x 20 repetitions x 14 speakers). From these, 90 tokens were discarded due to incorrect phrasing and an additional 79 tokens were excluded for incorrect or ambiguous focusing.

The acoustic recordings were force-aligned via the Montreal Forced Aligner [16], using a pre-trained Korean model [17]. An additional tier was created in Praat [1] segmenting only the sonorant portions of words capable of bearing F0.

Acoustic measurement of F0 was computed at nine equidistant time-points across the sonorant portion of the word in VoiceSauce [19], using the STRAIGHT algorithm [11]. Prior to statistical analysis, F0 was z-score normalized to address interspeaker differences in pitch range and height. Generalized additive mixed models (GAMM) [21] were constructed using the `bam` function of the *mgcv* package [20] in R [18] adjusting for autocorrelation in the time dimension by specifying a ρ value corresponding to the autocorrelation function at lag $= 1$. To account for the interaction of focus on separate words of the phrase, an interaction factor was created combining factors Word and Focus. This interaction term was then used as fixed effect and as smoothing term. Full random smooths were added for Speaker.[2]

3. RESULTS

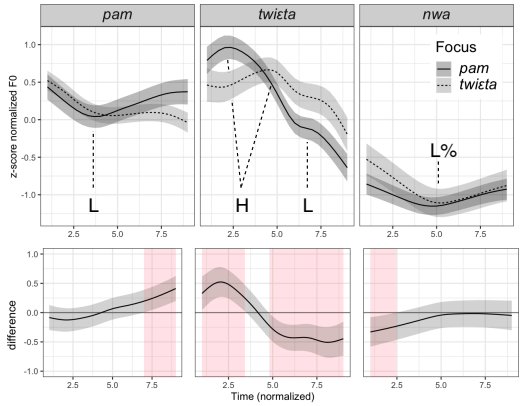
3.1. Phrase-initial Target words

In order to investigate differences between phrase-initial vs -medial focus, a GAMM was fit on the phrase-initial subset of the corpus. The subset yielded a total of 1600 observations and 42,740 data

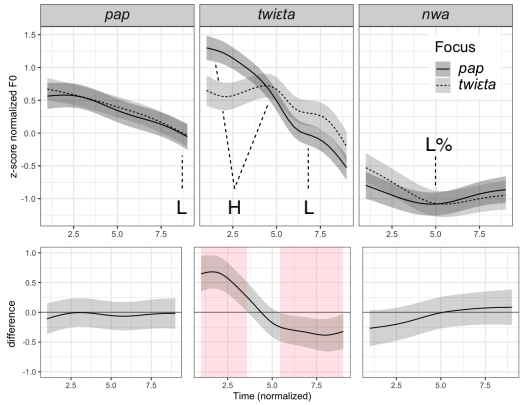
points (28,944 for *pam*, 13,796 for *pap*).

GAMM smooth and difference plots of the F0 contour for phrase-initial target words are shown in Figure 2. These smooth plots were created with the `plot_smooths` function of the *tidymv* R package [3]. The figures are organized such that beneath each GAMM's smooth plot is provided the difference plot indicating which portion of the two curves are significantly different.

The plots for both phrases containing *pam* and *pap* evidence substantial differences in the F0 curve of the phrase-medial word *twiɛta* "behind." For both, phrase-initial focus results in an earlier-anchored, higher-scaled peak in comparison to the corresponding phrase-medial focus. The F0 contour of the initial word evidences only a small difference by focus conditions for *pam*. Phrase-initial *pap*, however, bears the same contour regardless of its prominence status.



(a) *pam*



(b) *pap*

Figure 2: Visualization of the non-linear smooths and difference plots for phrase-initial target word contexts. The pointwise 94%-confidence intervals are shown as grey ribbons. The pink vertical bars represent portions of the curve where the two focus conditions significantly differ.

The difference in the predicted contours of the two target words is perhaps exaggerated by the effects of time-normalization. The sonorant portion of *pam* was approximately twice as long as that of *pap* on average. If we restrict our attention to the first half of the contour for *pam*, the shape of the contour is comparable to that of *pap*, i.e. a fall to a local F0 minimum. The shortened contour of *pap*, however, does not result in a lowering or a shift in the anchoring of the following peak. Instead, the local F0 extrema occur with no explicit rising contour as illustrated in the upper left two panels of Figure 2b.

The effect of the target word's coda on the F0 of the following word is shown in Figure 3. Focused *pap* induces a higher initial F0 value in the following word. This may due to local perturbation resulting from the *post-obstruent tensing* rule in Korean whereby the following lenis obstruent is tensified due to the presence of the preceding obstruent (i.e., /pap twi/ > [pap t*wi]) within the same AP. Tensified consonants yield higher F0 due to their laryngeal status.

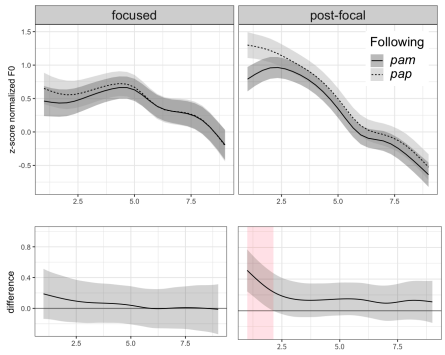


Figure 3: Visualization of non-linear smooths an difference plots of phrase-medial *twiɛta* under focal (left) and post-focal (right) contexts. The pointwise 95%-confidence intervals are shown as grey ribbons. The pink vertical bars represent portions of the curve where the two focus conditions significantly differ.

3.2. Phrase-final Target words

Another GAMM was fit to the subset of the corpus where the test words occupy phrase-final position. In these sentences, focus occurs either on the target word phrase-finally or on the immediately preceding word *ʌnni* "older sister." This subset provided a total of 1,591 observations and 42,948 data points (28,674 for *pam*, 14,274 for *pap*).

GAMM smooth and difference plots of the F0 contour for phrase-final target words are provided in Figure 4. These plot the model's non-linear predictions for the F0 contours for phrase-final (dark line) and phrase-medial (light dotted line) focus. Beneath the smooth plots are situated difference plots illustrating which portion of a word's F0 contour differs significantly by focus.

Phrase-medial focus on *ʌnni* results a substantial F0 rise over the disyllabic word. An effect of prominence-induced scaling is also manifest on the initial word *uri* "our," which displays a lowering effect at its right edge pre-focally perhaps suggesting tonal coarticulation with the following tone.

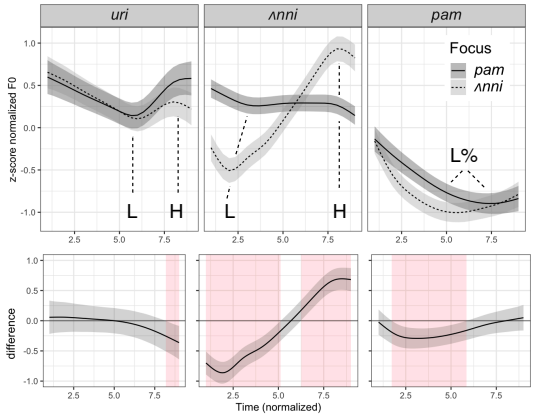
Phrase-final focus differs from initial and medial focus in the absence of the typical rise associated with AP-initial position. Instead, both focal and post-focal phrase-final words exhibit a fall, whose exact shape differs according to focus. For both *pam* and *pap*, the F0 minimum occurs later under focus. This is especially apparent for *pap*, which consists solely of a fall; the trough occurring at the phrase edge.

In contrast to the raising of the peak under initial focus, there appears to be no scaling effect of focus on the phrase-initial trough. For both words regardless of focus, the GAMM predicts an F0 minimum of zero, i.e. a speaker's mean F0 value. The trough of post-focal words, assumed to represent the boundary tone occurs earlier, thereby permitting the possibility of a small rise at the right edge.

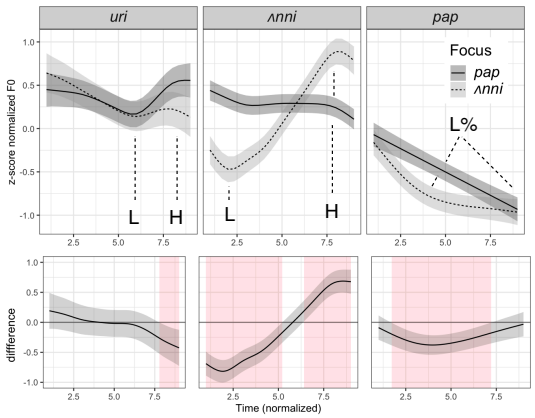
4. DISCUSSION

These results clearly show that the F0 manifestation of focus-induced prominence is sensitive to phrase position. Monosyllabic phrase-initial words incapable of bearing both tones necessary for the AP-initial rise show little effect of prominence on F0. Notably, however, the effect occurs on the post-focal word. The domain for prominence-induced scaling of F0, therefore, is not to the focused word *per se*, but may be the entire AP including any dephrased material. The scaling of the AP-initial L tone was not effected by focus in our data, contrasting with the results from South Kyungsang Korean [2], which was shown to exhibit a "polarizing" of initial pitch accents under focus. Note, however, the prosodic system of South Kyungsang differs crucially in bearing lexically-specified pitch accents as opposed to the phrasally specified tones of Seoul Korean.

Phrase-final focus differs from other positions in lacking an F0 rise. Instead final words evidence a fall anchored late in the word. The late anchoring of the L may perhaps represent the dual association of AP initial L tone and IP-final boundary tone. This behavior suggests a more complex relationship between the boundary tone and AP tones than that of simply overriding.



(a) *pam*



(b) *pap*

Figure 4: GAMM for phrase-final target words (a) *pam* and (b) *pap*. The pointwise 95%-confidence intervals are shown as grey ribbons. The pink vertical bars represent portions of the curve where the two focus conditions significantly differ.

Finally, it is worth noting that coda sonorancy plays a role in determining the shape of the contours investigated above. In both contexts investigated, the F0 contour of *pap* consisted solely of a fall as shown in the top left panels of Figs 2b & 4b. The phrase-initial trough representing the L tone was anchored with the vowel offset, with the sonorant coda of *pam* interpolating to the peak associated with the following syllable. While these effects are attributed to the tone bearing ability of the coda, what remains invariant is the timing of L and H targets regardless of whether there is an obstruent gap or not. Phrase finally, the difference seems to bear on the phonological system, restricting the choice of boundary tone (e.g., better accommodating a complex tone LH% with the sonorant coda.)

In conclusion, the present study has shown that tonal distribution in Seoul Korean is influenced by both its intonational grammar and low-level phonetic differences in consonant sonorancy. Prominence may interact with boundary-induced strengthening and the existence or absence of phrase-final boundary tones.

5. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

6. REFERENCES

[1] P. Boersma and D. Weenick, "Praat: Doing phonetics by computer," Feb. 2021.

[2] T. Cho, D. J. Kim, and S. Kim, "Prosodic strengthening in reference to the lexical pitch accent system in South Kyungsang Korean," *The Linguistic Review*, vol. 36, no. 1, pp. 85–115, Feb. 2019.

[3] S. Coretta, J. v. R. I. functions from itsadug), and M. W. I. functions from itsadug), "Tidymv: Tidy Model Visualisation for Generalised Additive Models," Sep. 2022.

[4] H.-S. Jeon and F. Nolan, "Prosodic Marking of Narrow Focus in Seoul Korean," *Laboratory Phonology*, vol. 8, no. 1, p. 2, Jan. 2017.

[5] S.-A. Jun, *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*, ser. Outstanding Dissertations in Linguistics. New York ; London: Garland Pub, 1996.

[6] ——, "Korean Intonational Phonology and Prosodic Transcription," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 2005.

[7] ——, "The intermediate phrase in Korean: Evidence from Sentence Processing," in *Tones and Tunes: Experimental Studies in Word and Sentence Prosody*, T. Riad and C. Gussenhoven, Eds. De Gruyter Mouton, Feb. 2009, vol. 2.

[8] ——, "Prosodic Markings of Complex NP Focus, Syntax, and the Pre-/Post-focus String," in *West Coast Conference on Formal Linguistics (WCCFL)*. Somerville, MA,: Cascadilla Proceedings Project, 2011, pp. 214–230.

[9] ——, "Prosodic typology: By prominence type, word prosody, and macro-rhythm," in *Prosodic Typology II: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2014, pp. 520–539.

[10] S.-A. Jun and H.-J. Lee, "Phonetic and phonological markers of contrastive focus in Korean," in *5th International Conference on Spoken Language Processing (ICSLP 1998)*. ISCA, Nov. 1998, pp. paper 1087–0.

[11] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[12] H.-S. Kim, S.-A. Jun, H.-J. Lee, and J.-B. Kim, "Argument Structure and Focus Projection in Korean," in *Speech Prosody*, Dresden, Germany, 2006.

[13] F. Kügler and C. Féry, "Postfocal Downstep in German," *Language and Speech*, vol. 60, no. 2, pp. 260–288, Jun. 2017.

[14] D. R. Ladd, *Intonational Phonology*, 2nd ed., ser. Cambridge Studies in Linguistics. Cambridge ; New York: Cambridge University Press, 2008.

[15] M. Liberman and J. B. Pierrehumbert, "Intonational Invariance under Changes in Pitch Range and Length," in *Language Sound Structure: Studies in Phonology*, M. Halle, M. Aronoff, and R. T. Oehrle, Eds. Cambridge, Mass: MIT Press, 1984, pp. 157–233.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 498–502.

[17] M. McAuliffe and M. Sonderegger, "Korean MFA acoustic model v2.0.0a," https://mfa-models.readthedocs.io/acoustic/Korean/Korean MFA acoustic model v2_0_0a.html, Tech. Rep., May 2022.

[18] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2022.

[19] Y.-L. Shue, P. A. Keating, C. Vicenik, and K. Yu, "Voicesauce: A Program for Voice Analysis," in *Proceedings of International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 1846–1849.

[20] S. Wood, "Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation version 1.8-38 from CRAN."

[21] S. N. Wood, *Generalized Additive Models: An Introduction with R*, second edition ed., ser. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press/Taylor & Francis Group, 2017.

[1] This is excluding the initial negation /ani/ 'no,' which is phrased as an independent IP.

[2] The R code for the GAMM construction of the phrase-initial target words is shown below in Listing 1.

Listing 1: R code for phrase-initial GAMM construction

```
# creation of interaction term
ini <- mutate(
  labFoc = interaction(word, focus))
# main effect of focus
bam(scaleF0 ~ labFoc
# smooths by focus
+ s(Time, by=labFoc, bs="cr", k=9)
# random smooths for Speaker
+ s(Time, Speaker, by=labFoc,
    bs="fs", m=1),
data=ini, method="fREML",
rho=r1, AR.start=ini$start.event)
```

PROSODIC PROMINENCE OF SWAHILI TU IN POLAR INTERROGATIVES

Seunghun J. Lee[1, 2], Shigeto Kamano[1], Yuko Abe[3], Kumiko Miyazaki[4]

[1] International Christian University, [2] University of Venda, [3] Lanzhou University,
[4] Global Center for Kiswahili Studies and Advancement
seunghun@icu.ac.jp, shig.kamano@gmail.com, yukoa@lzu.edu.cn, kumikomiyazaki.10@gmail.com

ABSTRACT

Prosodic patterns of the exclusive particle TU in Swahili polar interrogatives are examined. TU is realised with a pitch boost and appears in both sentence-final and non-final positions. Interrogatives overall display sentence-final falling intonation, resulting in a prosodic mismatch with sentence-final TU.

We recorded 30 stimuli sentences with 4 or 5 words varying in the position of TU from six speakers of Zanzibar Swahili. The Praat algorithm was used to extract f0 and the duration of the penultimate syllable of each word as well as that of TU and the ultimate syllable of a final word. Results show that the prosody of interrogatives differs from declaratives by raising f0 of the final word. TU is always associated with a pitch boost regardless of its position. Longer durational cues in sentence-final TU were used by two speakers, showing inter-speaker variation.

Keywords: Sentence Prosody, Exclusive Particle, Interrogatives, Zanzibar Swahili

1. INTRODUCTION

This paper aims to answer two questions related to the yes-no question patterns in Swahili. First, we investigate how sentence intonations are realized in declaratives and interrogatives. Secondly, we further investigate the intonation pattern of sentences including the exclusive particle TU 'only, just' which 1) occurs in a phrase or sentence-final position and 2) bears an inherent high pitch. Our main interest lies in investigating the intonational distinction between statements and corresponding polar questions in sentences that contain the particle TU that is required to display pitch boost.

Swahili is a stress/accent language while many Bantu languages are tonal. Stress in Swahili is accompanied by length and a falling tone and is normally realized on the penultimate syllable of a word or phrase [1], [2]. The intonation of Swahili varies depending on sentence types, but some common features can be used when describing the intonation system of yes-no polar questions [1].

In the descriptions of intonation of questions in Swahili polar questions are marked with stress and a falling tone on the final syllable in [2], whereas a phonetic-based description in [1] mentions both the penultimate and ultimate syllables; (a) the stressed penultimate syllable is realized with a mid, a high-falling, or a mid-rising pitch, (b) the final syllable is associated with a high falling pitch or a low pitch, (c) the final syllable may be longer in the high falling pitch cases. [3] and [4] report that the initial element of an interrogative sentence receives a high pitch. These studies indicate that polar questions may have a falling pitch on the final syllable, but our data suggest that this pattern is not universal non-falling, non-low pitch pattern also marks polar questions (cf. [5]).

A summary of intonation in African tonal languages in [6] demonstrates that most languages use falling intonation with a L boundary tone in polar questions. Swahili is a stress language, and in our data at least, polar questions do not end with a L boundary tone when compared with declaratives.

The rest of the paper is structured as follows. In section 2, methods of data collection are described. The results concerning the intonation of polar questions with and without the exclusive particle TU in section 3 demonstrate the prosodic effect of TU: pitch boost regardless of its position in a sentence. The paper concludes with a discussion about the interplay between intonation of sentence types and pitch realization of a particle.

2. METHODS

2.1. Stimuli

For stimuli, three conditions were considered: No-TU, Non-final-TU, and Final-TU. The base structure of all stimuli sentences is S + V + O. The No-TU condition sentences exclude the TU particle (1a, 2a). Sentences with TU have two conditions: the Non-final-TU condition, S + V + *tu* + O in (1b, 2b), and the Final-TU condition, S + V + O + *tu* (1c, 2c). Each condition contains 5 types of object DP composed of a noun and an adjectival modifier, which differ from one another in the vowel quality of the sentence-final syllable. All conditions had two sentence types: polar interrogatives and declaratives.

Overall, 30 items were created: 3 TU-related conditions x 5 object DP types x 2 sentence types.

(1) Examples of polar interrogative sentences
a. No-TU condition with final [e] vowel
Juma a-li-wa-fuata wa-toto wa-nane?
Juma 3SG-PST-3PL-follow 2-child 2-eight
'Did Juma follow eight children?'

b. Non-Final-TU condition with final [e] vowel
*Juma aliwafuata **tu** watoto wanane*?
'Did Juma just follow eight children?'

c. Final-TU condition with final [e] vowel
*Juma aliwafuata watoto wanane **tu***?
'Did Juma follow only eight children?'

(2) Examples of declarative sentences (for control)
a. *Juma aliwafuata watoto wanane*.
'Juma followed eight children.' (cf. 1a)

b. *Juma aliwafuata **tu** watoto wanane*.
'Juma just followed eight children?' (cf. 1b)

c. *Juma aliwafuata watoto wanane **tu***.
'Juma followed only eight children?' (cf. 1c)

2.2. Participants and Recordings

Six native speakers of Zanzibar Swahili, evenly divided by gender, were asked to read 30 stimuli sentences twice, resulting in 60 tokens per speaker. At the time of the recording, the age of participants ranged between 27 to 49. All participants were compensated for their time.

All recordings were conducted in a quiet room in Zanzibar, Tanzania, using Audacity on a MacBook computer set at 44.1kHz sampling rate and 16 bit depth.

2.3. Annotation and analysis

The recordings were processed and annotated using Praat [7]. The vowel of the target TU and the vowel of the penultimate syllable of each word is annotated as in Figure 1. For example, the [u] vowel in the penultimate syllable of the second word *alikula* 'he/she ate' is marked with 2. The fourth word is further annotated into 4P that stands for the penultimate syllable, and 4U for the ultimate (final) syllable. Three acoustic variables (maximum F0, minimum F0, and duration) from each annotated segment were extracted using the Praat algorithm. This data was then processed using R [8].

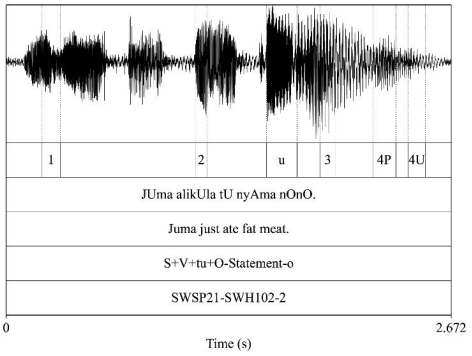


Figure 1: An example of the annotated sentences: *Juma alikula tu nyama nono*. 'Juma just ate fat meet' The annotated vowels are capitalized in Swahili.

3. RESULTS

3.1. Interrogatives without TU: No-TU condition

The prosodic differences between declaratives and interrogatives are reported in this section. In most tokens (n = 39 out of 40) produced by four speakers (SWH102, 103, 104, 106), a relatively higher pitch of the final word is observed in interrogatives than in declaratives. The remaining two speakers did not display any pitch differences, resulting in the identical intonation between the two sentence types.

The higher pitch of a final word has a variable position. Three participants (SWH103, 104, 106) placed an emphasis on the ultimate syllable of the final word. As Figure 2 shows, the ultimate syllable of the fourth word (4U) in interrogatives is produced with a higher pitch compared to the penultimate syllable of that word (4P). The participant SWH102 produced the penultimate syllable in the final word in interrogative sentences with a higher pitch than in declaratives (see Figure 3).

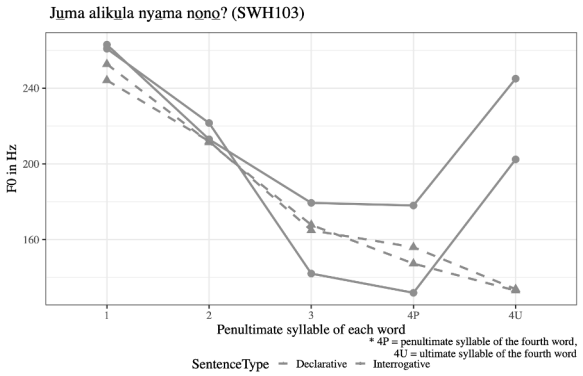


Figure 2: Pitch transition for no-TU sentences of SWH103. Dotted lines are declaratives and solid lines are interrogatives.

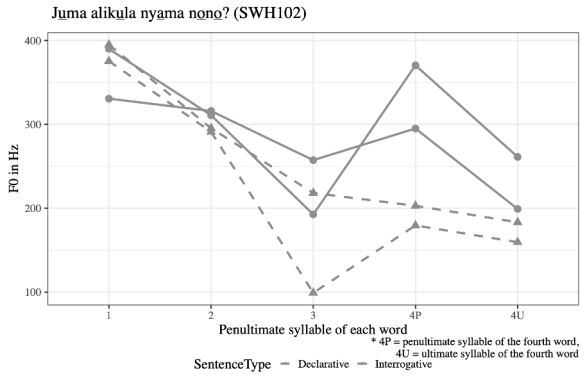


Figure 3: Pitch transition for no-TU sentences of SWH102. Dotted lines are declaratives and solid lines are interrogatives.

The remaining two speakers (SWH101, 105) show intonation patterns, in which interrogatives have a falling intonation on the final word, which does not differ from corresponding declaratives.

3.2. Interrogatives with a non-final TU

As per [9], Swahili TU triggers pitch boost on itself. Excluding the pitch boost of TU, interrogatives in the Non-final TU condition have the same intonation pattern as the no-TU condition. As shown in Figure 4, after the pitch boost triggered by TU, the pitch of penultimate syllable *nyama* in the Non-final-TU condition (solid line) returns to the level comparable to the No-TU condition (dotted line).

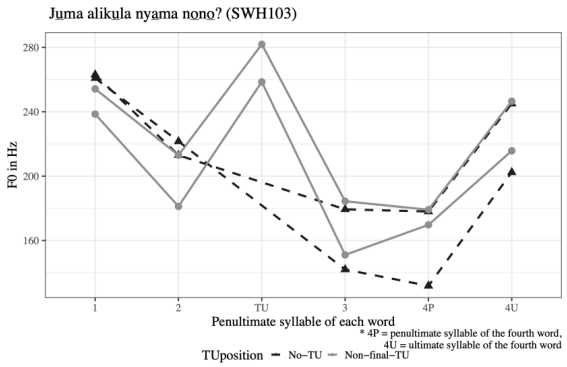


Figure 4: Pitch transition for No-TU and Non-final TU conditions (SWH103): *Juma alikula nyama nono*? 'Did Juma eat fat meet?' (No-TU in dotted lines) vs. *Juma alikula tu nyama nono*? 'Did Juma just eat fat meet?' (Non-final-TU in solid lines).

3.3. Interrogatives with a final TU

Since the prosodic cue for interrogatives in Swahili depends on the stress of the final word, due to the prosodic prominence of TU, it would be difficult to produce/perceive the F0 difference in the final TU between declarative and interrogative. In addition to the F0 difference observed in no-TU polar interrogatives, the duration of the final TU is used by three speakers as an alternative prosodic cue for interrogatives.

3.3.1 F0 differences in the final TU

Using R, we modeled the pitch contour of the final TU vowel with the F0 of 10 equidistant points extracted from the vowel interval. As shown below in Figure 5, the participant SWH102 uses the F0 final TU vowel to differentiate Final-TU interrogatives from Final-TU declaratives.

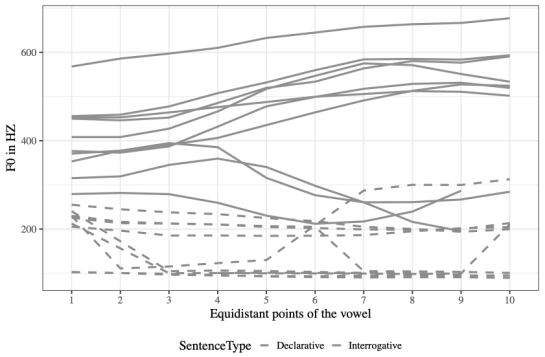


Figure 5: Pitch contours by normalised duration of the vowel of the final TU of SWH102. Dotted lines for declaratives and solid lines for interrogatives.

3.3.2 Durational differences in the final TU

Two speakers (SWH103, 104) utilized the duration of the final TU as an alternative prosodic cue for interrogatives. Figure 6 shows that SWH103 produces the final TU in interrogatives longer than in declaratives.

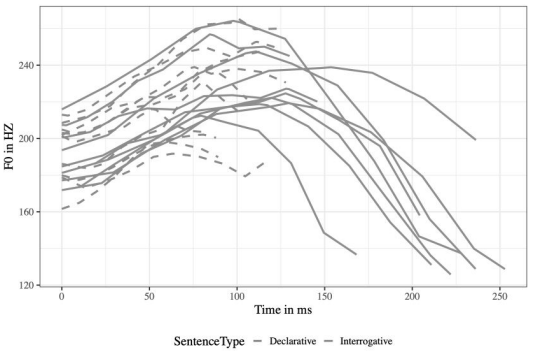


Figure 6: Pitch contours by non-normalised duration of the vowel of the final TU of SWH103. Dotted lines for declaratives and solid lines for interrogatives.

For analysis, the maximum and minimum F0 of each vowel interval was extracted, alongside the F0 of 10 equidistant points throughout the interval. The F0 values are plotted against their relative extracted time (extracted time – start time). This durational

difference is also observed in the Final-TU interrogatives in three tokens out of ten spoken by SWH106.

In addition to the durational difference, a difference in pitch fall is observed in the final TU spoken by SWH103, 104, and 106. The three speakers produce a larger pitch fall in the vowel of the final TU in interrogatives than in declaratives.

4. DISCUSSION

4.1. Two groups in the No-TU condition

The results show that speakers can be divided into two groups based on how they produce the interrogatives without TU. The first group emphasizes the final word, either on the ultimate syllable as in *nonó* (SWH103, 104, 106) or penultimate syllable as in *nóno* (SWH102). Since Swahili is a language that generally places the accent on the penultimate syllable of each word, placing a higher pitch on the penultimate syllable suggests a parsimonious strategy for distinguishing interrogatives from declaratives.

When speakers place the accent on the ultimate syllable instead, the principle of accent assignment in Swahili is violated. It is interesting that a number of speakers utilize this pattern of interrogative intonation. This ultimate-accent pattern for interrogatives might be attributed to the influence of rising-intonation of English interrogatives, but this account is not reliable as the speakers who display this intonation do not belong to the younger group of speakers (age 29, 42, and 49). As such, we suggest that the constraint assigning prominence to the penultimate syllable of a word may not be an absolute requirement in Swahili prosody.

As no participant puts an emphasis on the antepenultimate syllable when the final word consists of three syllables, the prosodic domain where the accent can be realized in Swahili is assumed to be limited to the final two syllables.

The second group consists of speakers who do not differentiate declaratives and interrogatives in terms of prosody (SWH101, 105). This unmarked intonation for interrogatives could stem from sociolinguistic or pragmatic reasons; speakers who use this intonation might ask questions with a context-based approach to avoid using a direct question or a command. Interrogatives of African tonal languages have been reported to have similar prosody to declaratives [6]. Our study shows that a non-tonal African language, Swahili, could also exhibit such a pattern.

4.2. Interrogatives with TU

With regards to TU, we conclude that the intonational characteristics for interrogatives with TU are not obligatory but optional. For those who differentiate interrogatives from declaratives by intonation, the prosodic prominence of TU interferes with the intonational differentiation; since TU is monosyllabic, a speaker cannot put an accent on any other syllable as they can do in *nonó*. Three speakers, that put an accent on an ultimate syllable of the final word for interrogatives without TU, use the duration of TU or its pitch fall as alternative cues for the interrogatives with TU. On the other hand, another speaker, that produces a higher pitch on a penultimate syllable of the final word for interrogatives without TU, uses only F0 even when asking questions with TU. These inter-speaker variations indicate that interrogatives with TU are realized in several ways or sometimes become hidden in terms of prosody.

More discussions on the durational difference and pitch fall are required. These characteristics are observed in tokens spoken by SWH103, 104, and 106, and both characteristics never conflict with each other. When duration is lengthened, pitch is also lowered to relieve any articulatory burden. Vice versa, when pitch is lowered, duration is also lengthened to avoid a sudden pitch fall. At this juncture, it is still unclear as to which constitutes the main attempt and which belongs to the quadratic effect.

5. CONCLUSION

The question prosody of Swahili with or without the exclusive particle TU has been investigated in this paper. We found that (a) interrogatives end with a high boundary tone compared to declaratives, and (b) TU is realised with a pitch boost. When TU is non-final, the intonation patterns of interrogatives are identical except for the pitch boost. When TU is sentence-final, (c) the pitch of TU is higher in interrogatives than in declaratives, and (d) TU is realised longer with a clear falling contour.

6. AKNOWLEDGEMENTS

This work was partially supported by JSPS KAKENHI Grant #20K00578 (POP), JSPS KAKENHI Grant (B) #21KK0005 (JSAntu), JSPS KAKENHI Grant #JP23H00622 (PACroBaT), the ILCAA joint research project “Phonetic typology from cross-linguistic perspectives (PhonTyp)” (2021-2023), and JICUF grant [Spring 2022-2023 | 77].

7. REFERENCES

[1] Polomé C. 1967. *Swahili Language Handbook*. Center for Applied Linguistics.

[2] Ashton, E.O. 1989. *Swahili grammar: including intonation*. Longman.

[3] Maw, J., Kelly, J. 1975. *Intonation in Swahili*. SOAS, University of London.

[4] Maw, J. 1992. *Narrative in Swahili*. SOAS, University of London.

[5] Nassenstein, N. 2015. Kisangani Swahili. LINCOM.

[6] Downing, L.J., Rialland, A. 2017. *Intonation in African Tone Languages*. De Gruyter Mouton.

[7] Boersma, P., Weenink, D., 2018. *Praat: doing phonetics by computer*. Version 6.0.29. Retrieved May 29, 2020.

[8] R Core Team. 2020. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna. [Online]. Available: https://www.R-project.org/

[9] Kamano, S., Abe, Y., Miyazaki, K., Lee, S.J. 2022. Prosodically prominent clitic: the exclusive particle tu in Swahili. Presented at the *Prosody and Grammar Festa* 6.

SOME ASYMMETRICAL PRE- *VERSUS* POST-FOCAL EFFECTS ON ARTICULATORY REALIZATION OF PROMINENCE DISTRIBUTION IN KOREAN

Suyeon Im[1], Sahyang Kim[2] & Taehong Cho[3]

[1]Soongsil University, [2]Hongik University, [3]Hanyang Institute for Phonetics & Cognitive Sciences of Language, Hanyang University
sim@ssu.ac.kr, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study examines how CVC words in non-prominent contexts are kinematically realized in Korean, compared to when focused. Results on lip aperture showed some asymmetric effects between pre-focal and post-focal contexts. While both pre-/post-focal gestures were much more reduced than focal counterparts, the difference tended to be clearer for the pre-focal/focal than the post-focal/focal comparison. There was also some evidence for post-focal gestures being relatively less reduced than pre-focal gestures. We attribute this small asymmetry to differential auditory impacts of non-focal gestures on focal gestures, assuming that gestures are auditory-perceptually more salient with a drastic energy increase from the preceding gesture than the opposite direction. Relatively less reduced post-focal gestures compared to pre-focal gestures suggest that focus-induced accentuation spreads to the following, rather than preceding, gesture. These results imply that although prominence may be phonologically determined, the articulatory realization is fine-tuned, driven by perceptual and articulatory optimization of prominence distribution.

Keywords: prosody, focus, prominence distribution. articulatory kinematics, Korean

1. INTRODUCTION

Prominence is one way of making speech signal perceptually salient to the listener, so that the intended linguistic message is delivered clearly to the listener as intended by the speaker (cf. [1]). In a language like English, prominence is often localized on a lexically stressed syllable, especially when it receives a pitch accent that is post-lexically assigned by the prominence system of the language. Such a prominence effect is most clearly evident when the pitch accent is assigned in conjunction with focus associated with information structure [2, 3, 4, 5]. A language like Korean does not employ lexical stress, so that it may not be localized to a specific stressed syllable, but the notion of prominence itself may still apply to Korean, so that it makes a linguistic unit more salient than its neighbouring units to enhance the perceptual salience of the linguistic message under prominence.

As such, prominence may be defined in a relative term not only in the hierarchy of stress in a given language, but also in terms of relative perceptual salience (e.g., [6, 7]). This suggests that a speech unit (a syllable or a prosodic word) can be perceived salient when the unit itself is hyperaticulated more than neighboring segments. From another angle, this also means that the unit is perceived salient when neighboring units are reduced. Thus, one could infer that such relative salience associated with prominence is employed cross-linguistically by the prominence system of any language, regardless of whether the language employs lexical stress as in English or not as in Korean.

In the present study, we investigate how CVC words are reduced in non-prominent (non-focal) contexts adjacent to the focal context in Korean, relative to when they are focused. That is, we explore how a language which does not employ lexical stress in the prominence system expresses prominence distribution in relative terms and how the results may compare to existing prominence-related data available in other languages such as English (e.g., [3, 4, 6, 8, 9, 10]) that uses a typologically different prominence system. Moreover, we use an Electromagnetic Articulograph, so that we can examine how reduction in non-focal contexts versus hyperarticulation in focal contexts are kinematically expressed in both spatial and temporal dimensions.

A specific question to be addressed is how articulatory reduction of pre-focal versus post-focal words may be similar to or different from each other, as compared to when the same words are focused. Both the pre- and post-focal conditions are expected to induce articulatory reduction, but given the directional asymmetry of pre- versus post-focal conditions, we expect that the degree of reduction will also be asymmetrical. It is hypothesized that the degree of articulatory reduction will be greater in the pre-focal than in the post-focal context, assuming that the auditory-perceptual impacts are likely to be greater when there is a drastic and rapid increase in articulatory force from a non-focal gesture on the following focal gesture rather than the other way around (e.g., [11]). This asymmetrical effect can also be predicted independently by how the domain of

accentual impact is often defined. For example, it has been suggested that accentual impact in English may spread to the right rather than to the left, showing a kind of carry-over effect, although the domain may be defined phonologically (e.g., a foot; [12, 13]).

2. METHOD

2.1. Participants & Speech Materials

As part of a larger project building an articulatory database, twelve speakers of Seoul Korean (6 male and 6 female college students, M_{age} = 23.3 years) participated in an articulatory experiment using an Electromagnetic Articulograph (Carstens AG501). They were instructed to imagine playing a card game in which the participant would play a role to answer the interlocutor's question regarding where to put a card (with a target word written on it) on the board. The participants read provided written sentences in response to question sentences with the pictures of the cards visually presented along on a computer screen. The question sentences were designed to elicit corrective focus in the answer sentences on (a) the target word (focal condition), (b) the preceding word (pre-focal condition) and (c) the following word (post-focal condition). The target words were /pap/ ('rice') and /pam/ ('night'/'chestnut') in a phrase-medial position (in the middle of an Intonational Phrase) with the same preceding and following words. An example of test sentences is shown in Table 1.

Table 1: Example of test sentences. Target words are underlined. Focused words are in bold.

(a) focal
Q: *[ipʌn tanʌnin ʌnni kuk twiɛta nwa]?*
'Do (I) put the word behind the sister's soup this time?'
A: *[ani. ʌnni* ***pap*** *twiɛ]*.
'No. Behind the sister's **rice**. '
(b) pre-focal
Q: *[ipʌn tanʌnin ʌnni pap ap^{h}ɛta nwa]?*
'Do (I) put the word in front of the sister's soup this time?'
A: *[ani. ʌnni pap* ***twiɛ****]*.
'No. **Behind** the sister's rice.'
(c) post-focal
Q: *[ipʌn tanʌnin op*a pap twiɛta nwa]?*
'Do (I) put the word behind the brother's rice this time?'
A: *[ani.* ***ʌnni*** *pap twiɛ]*.
'No. Behind the **sister's** rice.'

2.2. Measurement

Eight sensor coils were attached to the nose (R1), the two lips (L1, L2), the two gums (R2, J), the tongue tip (T1), tongue body (T2), and the tongue dorsum (T3) as illustrated in Figure 1. The lip aperture (i.e., the Euclidian distance between the upper lip and the lower lip) was used to examine kinematic characteristics of lip closing and opening gestures of CVC words with bilabial consonants (/p/ and /m/). Figure 2 illustrates gestural landmarks used for kinematic measurements. The onset (ONS) and the target (TARG) of each (closing or opening) gesture were determined as the time points when the velocity reached 20% of the peak velocity (PKVEL) as shown in Figure 2. Movement duration (ms; DUR), displacement (mm; DISP), and peak velocity (cm/sec; PKVEL) were measured for each movement from the 2043 tokens in the corpus (3 words * 3 focus conditions * 20 repetitions * 12 speakers - 117 excluded tokens whose renditions deviated from intended phrasings and focus distribution), using MVIEW [14].

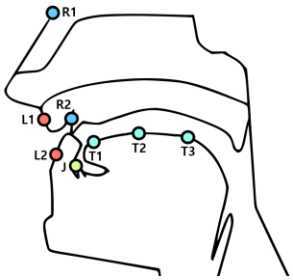


Figure 1: Location of eight sensor coils for articulatory experiment using an Electromagnetic Articulograph (Carstens AG501).

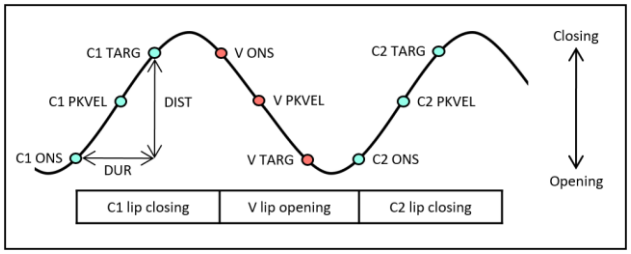


Figure 2: Gestural landmarks for C1 lip closing gesture, V lip opening gesture and C2 closing gesture.

2.3. Statistical Analysis

A linear mixed-effects model was run for each movement (C1 closing, C-to-V opening, C2 closing) with Focus (focal, pre-focal, post-focal) and Word (/pap/, /pam/) as fixed factors including their interaction. (Note that Word was included as a control factor, and given that there was no Focus x Word interaction, we did not report the statistical details of Word effects.) The random structure included by-subject intercept and slope for all the fixed factors. The models were run in R [15] using the lme4 package [16].

3. RESULTS

3.1. C1 (Onset) Lip Closing Gesture

For C1-closing gesture, duration showed no significant difference among focal, pre-focal, and post-focal conditions (pre-focal/focal: β=-1.44, *t*=-1.25, *n.s*; post-focal/focal: β=.48, *t*=.34, *n.s.*). On the other hand, as shown in Figure 3, both displacement and peak velocity measures showed a significant difference between the focal and the pre-focal conditions, such that compared to the focal condition, the pre-focal C1 closing movement was *smaller* in displacement and *slower* in peak velocity (disp., β=-.69, *t*=-3.58, *p*<.01; pkvel., β=-1.35, *t*=-2.86, *p*<.05). But there was no difference between the focal and the post-focal conditions (disp., β=.05, *t*=.20, *n.s.*; pkvel., β=-.26, *t*=-.57, *n.s.*). (Note that the overall duration of C1-closing movement was longer for /pam/ than /pap/, but it did not influence the focus-related effects.) There were no other significant effects or interactions among factors including.

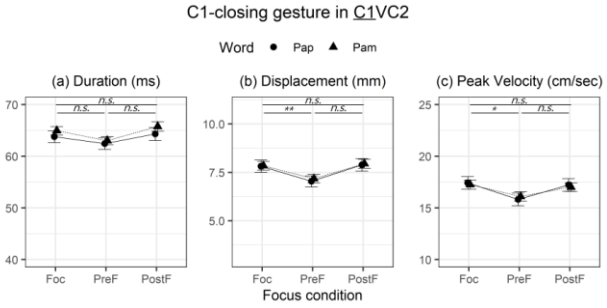


Figure 3: C1 lip-closing duration (left), displacement (middle), and peak velocity (right) plotted by Word (/pap/, /pam/) and Focus condition (focal, pre-focal, post-focal).

3.2. C-to-V Lip Opening Gesture

For C-to-V lip opening gesture, the focal condition significantly differed from both pre- and post-focal conditions in all the measures. As shown in Figure 4, C-to-V opening was longer in duration, larger in displacement and faster in peak velocity in the focal than in the non-focal conditions (focal/pre-focal: duration, β=-18.67, *t*=-8.50; disp., β=-5.23, *t*=-6.22; pkvel., β=-8.36, *t*=-5.93, all at *p*<0.01; focal/post-focal: duration, β=-16.36, *t*=-7.75; disp., β=-4.84, *t*=-6.05, pkvel., β=-7.66, *t*=-6.04, all at *p* < .001). (Note that C-to-V opening gesture was generally longer, larger and faster for /pam/ than /pap/, but it did not influence the focus-related effects.) No other significant effects between conditions were observed.

3.3. C2 (Coda) Lip Closing Gesture

There was a significant difference among the focus

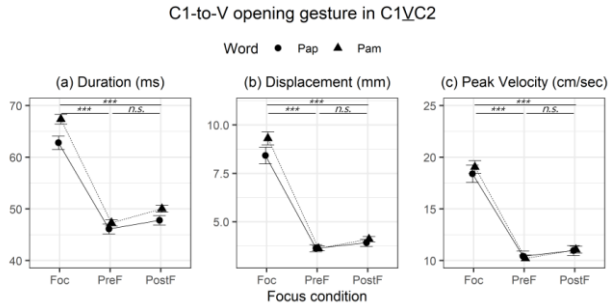


Figure 4: The V lip opening duration (left), displacement (middle), and peak velocity (right) plotted by Word (/pap/, /pam/) and Focus condition (focal, pre-focal, post-focal).

conditions in the C2 closing gesture. C2-closing movement was longer, larger and faster in the focused than in the unfocused condition (focal/pre-focal: duration, β=-9.90, *t*=-4.54; disp., β=-5.21, *t*=-7.20; pkvel, β=-9.44, *t*=-8.43, all at *p*<.001; focal/post-focal: duration, β=-8.06, *t*=-4.84; disp., β=-4.66, *t*=-6.51; pvkel, β=-8.34, *t*=-7.34, all at *p* < .001). Crucially, there was also a significant difference between pre- and post-focal conditions, so that post-focal C2-closing gesture was larger in displacement and faster in peak velocity (disp., β=-8.34, *t*=-7.34, *p* < .001; pkvel, β=-8.34, *t*=-7.34, *p* < .001). (Note that C2-closing gesture was generally longer for /pam/ than /pap/ especially in the focal context, but it did not influence the focus-related effects.) There were no other significant effects or interactions among factors.

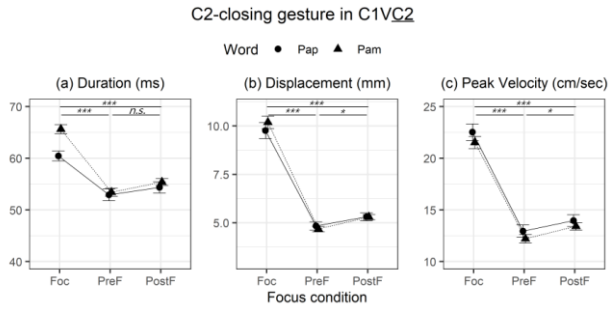


Figure 5: The C2 lip closing duration (left), displacement (middle), and peak velocity (right) plotted by Word (/pap/, /pam/) and Focus condition (focal, pre-focal, post-focal).

4. DISCUSSION

This study examined how articulatory Lip Aperture movements (C1-closing, C-to-V opening, and C2-closing) of C1VC2 words with bilabial consonantal contexts (/pam, pap/) would be realized in non-prominent contexts (pre-focal and post-focal) compared to when they are in the prominent (focal) context in Korean.

A basic finding was that non-focal gestures were much more reduced in both spatial and temporal dimensions than focal gestures that received a

corrective contrastive focus. In other words, focal gestures were hyperarticulated (being larger, longer and faster than non-focal gestures), being 'prominent' above the surrounding non-focal words. This hyperarticulation pattern in Korean is largely consistent with the hyperarticulation pattern generally reported in English [3, 6, 17]. The present study taken together with previous studies demonstrates cross-linguistic similarities—i.e., gestures under focus are hyperarticulated in much the same way across languages, regardless of whether a language is a head-prominence language like English (where prominence is localized to a stressed syllable) or an edge-prominence language like Korean (where prominence is primarily expressed by phrasing) (cf. [18]).

As for the specific research question of how pre-focal versus post-focal effects may differ from each other relative to focal effects, our results indicated that pre-focal gestures tended to be reduced more than post-focal gestures. This asymmetry was evident in two cases. For one thing, C1-closing gesture when in the pre-focal condition was reduced (in displacement and peak velocity) compared to when in the focal condition, whereas the same C1-closing gesture of the post-focal word showed no such reduction. For another, while C2-closing gesture was substantially reduced when in both pre- and post-focal conditions, it was the pre-focal C2-closing gesture that was reduced more (as evident in displacement and peak velocity), relative to the post-focal C2-closing gesture. (Recall that C1-to-V opening gesture was reduced to a similar extent in both pre- and post-focal conditions relative to the focal condition.)

These results indicate that the nature of reduction differs depending on the directionality of prominence distribution—i.e., whether it occurs in the pre-focal or post-focal context. On the one hand, the pre-focal word was reduced as a whole from the beginning C1-onset gesture to the final C2-closing gesture which was immediately adjacent to the focal word. On the other hand, the post-focal word was reduced in a rather progressively gradient way. The beginning (C1-closing gesture) of the post-focal word which was immediately adjacent to the focal element was not reduced at all in any kinematic measure, and the reduction became evident on the following C1-to-V opening and C2-closing gestures. And even the final C2-closing gesture of the post-focal word was reduced relatively less than the final C2-closiong gesture of the pre-focal word. These asymmetrical results can be interpreted in terms of differential auditory-perceptual effects and articulatory propensity of post-focal reduction.

The substantial across-the-board reduction of the pre-focal word is indeed consistent with what has previously been reported in other languages such as English [9, 10, 19] and French [20]. This cross-linguistically similar effect can be taken as suggesting that the speaker plans to reduce the preceding gesture deliberately to enhance the auditory-perceptual impact on the focused word that follows as discussed in [19], which would in turn bring about contrast maximization (e.g., [10]). This possibility is consistent with the general assumption that speech production is auditory-perceptually more salient and impactful with a drastic energy increase from the immediately preceding acoustic event than the opposite direction (see [11] for related discussion on asymmetrical auditory-perceptual effect.).

On the other hand, the post-focal production appears to be subject to the propensity to carry over the accentuation to the right rather than the other way around. The progressively increasing reduction to the post-focal word can be considered as the evidence of 'spill-over' (carry-over) effect as discussed in [12, 13, 21]. Such a rightward spread of articulatory force associated with prominence is consistent with a view that accentuation operates within a phonologically-defined domain of accentuation in English (e.g., a foot structure [12, 13]). But we suggest that the gradual rightward spread of accentuation also reflects the efficiency of the production system—i.e., more effort is required to drastically reduce the articulatory force immediately after the focal articulation than do so gradually. Thus, the observed gradual attenuation of the focus effect on the following non-focal unit can be interpreted as suggesting that the speaker avoids a drastic reduction unless otherwise required by the system, in line with a principle of effort minimization that underlies speech production [22].

5. CONCLUSION

The present study was the first to explore how prominence distribution would be reflected in kinematic terms in Korean, especially with respect to reduction patterns of pre-focal versus post-focal gestures as compared to focal gestures. The general differences in the focal versus the non-focal contexts illuminate that relative prominence is kinematically realized in a form of 'hyperarticulation' in much the same way across languages. Furthermore, the directional asymmetry of pre-focal versus post-focal effects further implies that although prominence may be defined differently in the phonology of a given language, articulation of prominence is fine-tuned by the production system of the language that optimizes prominence distribution taking into account both the listener-oriented auditory-perceptual saliency and the speaker-oriented motor efficiency.

6. REFERENCES

[1] Cangemi, F., Baumann, S. 2020. Integrating phonetics and phonology in the study of linguistic prominence. *J. Phonetics.* 81, 100993.

[2] Cho, T. 2011. Laboratory phonology. In: Kula, N. C., Botma, B., Nasukawa, K. (eds), *The Continuum Companion to Phonology*. Continuum, 343–368.

[3] de Jong, K. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* 97, 491–504.

[4] de Jong, K. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *J. Phonetics*. 32, 493–516.

[5] Fletcher, J. 2010. The prosody of speech: Timing and rhythm. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds). *The Handbook of Phonetic Sciences (2nd ed.)*. Blackwell, 523–602.

[6] Beckman, M. E., Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In: Keating, P. A. (ed), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*. Cambridge University Press, 7–33.

[7] Beckman, M. E., Pierrehumbert, J. 1986. Intonational structure in Japanese and English. *Phonology Yearbook,* 3, 255–309.

[8] Cho, T. 2006. Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. In: Goldstein, L., Whalen D. H., Best, C. T. (eds), *Papers in Laboratory Phonology VIII: Varieties of Phonological Competence (Phonology and Phonetics)*. Mouton de Gruyter, 519–548.

[9] Erickson, D., Lehiste, I. 1995. Contrastive emphasis in elicited dialogue: Durational compensation. *Proc. 13th ICPhS* Stockholm, 4, 352–355.

[10] Jensen, C. 2004. *Stress and accent: Prominence relations in Southern Standard British English.* Ph.D. dissertation, University of Copenhagen, 1–192.

[11] Wright, R. 2004. A review of perceptual cues and cue robustness. *Phonetically Based Phonology*, 34, 57.

[12] Turk, A. E., Sawusch, J. R. 1997. The domain of accentual lengthening in American English. *J. Phonetics,* 25, 25–41.

[13] Turk, A. E., White, L. 1999. Structural influences on accentual lengthening in English. *J. Phonetics,* 27, 171–206.

[14] Tiede, M. 2005. MVIEW: Software for visualization and analysis of concurrently recorded movement data. Haskins Laboratories.

[15] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

[16] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Statistical Software,* 67(1), 1–48.

[17] de Jong, K. 1991. An articulatory study of consonant-induced vowel duration changes in English. *Phonetica,* 48, 1–17.

[18] Jun, S. A. 2006. Intonational phonology of Seoul Korean revisited. *Japanese-Korean Linguistics*, 14, 15–26.

[19] Cho, T., Kim, J., Kim, S. 2013. Preboundary lengthening and preaccentual shortening across syllables in a trisyllabic word in English. *J. Acoust. Soc. Am.* 133(5), 384–390.

[20] Fougeron, C., Jun, S.-A. 1998. Rate effects on French intonation: Prosodic organization and phonetic realization. *J. Phonetics*. 26, 45–69.

[21] Sluijter, A. M. C., van Heuven, V. J. 1995. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica* 52, 71–89.

[22] Flemming, E. 1995. *Auditory representations in phonology.* Ph.D. dissertation, UCLA.

7. ACKNOWLEDGEMETNS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

EXAMINING VARIABILITY IN THE PRODUCTION AND PERCEPTION OF PROSODIC ATTITUDES

Jeesun Kim, Kate Raue and Chris Davis
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University
j.kim@westernsydney.edu.au, k_alimon@hotmail.com, chris.davis@westernsydney.edu.au

ABSTRACT

We tested the extent that expressions of prosodic attitude have been conventionalised by examining their variability and recognition accuracy. We used acoustic and perceptual measures of recordings of 10 speakers (five females) expressing six different prosodic attitudes 16 times (4 trials x 4 sessions). Recognition scores from 46 listeners were better than chance but generally poor. The top 10 acoustic attributes from classification models showed most speakers used some form of F0 variation to convey different attitudes, but few attributes were common across speakers. Productions from speakers whose use of prosody was more consistent across trials, i.e., productions better classified by four-fold cross-validation models, were better recognised. Further, these consistent models led to better classification when applied to the productions of other speakers with relatively high than low recognition scores. We suggest different attitudes can be appropriately signalled by prosodic forms, but these are only partly conventionalised.

Keywords: prosody; prosodic attitudes; expressive speech

1. INTRODUCTION

Speech conveys information through word meaning and word ordering, and via prosody, i.e., the melodies of pitch variation and the rhythm of speech timing. Traditionally, prosodic information has been classified into two broad categories: linguistic and paralinguistic information. Linguistic prosody refers to prosody that serves a linguistic function such as word stress, sentence focus, segmenting the speech into phrases and signalling their broad pragmatic categories. Given this role in language processing, it is presumed that knowledge and use of linguistic prosody has been conventionalised. Paralinguistic prosody, on the other hand, signals such things as a speaker's affect and their intention or attitudes and as such does not have a linguistic function per se. The current study examined the extent to which the knowledge and use of paralinguistic prosody have been conventionalised, specifically focussing on prosodic attitudes.

Traditionally, it has been argued that for attitudes to be appropriately communicated, they need to be presented in context [1]. That is, context, including the speaker and receiver relationship and the accompanying conversation, makes a large contribution to the success of perceiving the attitude [2]. However, more recent research has suggested that the prosodic forms associated with the expression of attitude have been conventionalized and can be appropriately realized even without context [3].

In [3]), the conventionalisation of prosodic forms was tested by determining how similar speaker productions of the different attitudes were, and how well listeners could identify the intended attitudes. To do this, they had four speakers' express 6 different attitudes (criticism, doubt, suggestion, warning and naming) when saying the German word "Bier" (beer). To assess production, the study used a set of 7 acoustic features (e.g., stimulus duration, mean intensity, harmonics-to-noise ratio (HNR), mean fundamental frequency (f0), the difference between offset and onset f0, spectral centre of gravity and the standard deviation of the spectrum). The results of a discriminant analysis showed the correct attitudes were classified with 92% accuracy. This high level of discrimination indicated the distinctiveness of the prosodic patterns that speakers used to express the different intentions. To assess how well the attitudes were recognised, a forced-choice task was used and mean correct recognition measured. Correct recognition accuracy was 82%. This high recognition rate along with the production discrimination result was taken as strong evidence that the prosodic forms for the expression of attitudes had been conventionalised.

An issue with [3], however, is that the speakers used were voice coaches who would have had expert knowledge and use of paralinguistic prosody. Indeed, another study [4] that used a similar method as [3] but tested 2 speakers who had no voice training, found considerable variation in how speakers produced the different attitudes and much lower recognition scores. This result suggests that the findings of [3] may not generalise.

One way that the results of the above two studies could be reconciled is if prosodic attitudes have only been partly conventionalised. That is, the extent of knowledge and use of the conventionalised forms

vary across individuals. Given this, several predictions can be made. First, when a speaker has knowledge of and uses a more conventionalised prosodic pattern, their productions should be more consistent over trials and better recognised than those of speakers who did not have such knowledge. Second, speaker productions that are better recognised should be more like each other than those of speakers that were not well recognised. The current study aimed to test these predictions.

In the current study, we asked 10 speakers with no voice-training to produce the two-word phrase "first class" to express six different attitudes. We first determined how variability productions were by using classification models. As a first step, we identified the top 10 acoustic attributes used in each speaker's trained classification model and examined how common these were across speakers. We then tested the first prediction above by determining the association between classification performance and recognition score. Finally, we tested the second prediction by examining the extent that classification models of speaker productions that were well recognised could classify other well recognised productions compared less well recognised ones.

2. METHOD

2.1. Production study

2.1.1. Participants

Ten native Australian English speakers (5 females, mean age = 25.8 years, range 22-28) participated in the study and given a small reimbursement. The speakers had no voice training.

2.1.2. Recording

Auditory recordings were made using an externally connected lapel microphone, (an AT4033a audio-technica microphone) in 44.1 kHz, 16-bit mono.

2.1.3. Procedure

Each talker was recorded individually. Talkers were seated in a quiet room with a camera positioned directly in front at face level at approximately 0.6 metres distance. Recording was controlled by an operator in a control room who ensured that the participants looked at the camera throughout the capture. The different prosodic attitudes were elicited using the method outlined in [3]. That is, the speaker read short scenarios that described a situation in which she/he would say the carrier phrase 'first-class' to the interlocutor using each of 6 different attitudes (criticism, doubt, longing, suggestion, warning and naming). For each scenario, the speaker freely vocalised until ready to say the target phrase with the intended attitude. In each session the phrase was said four times in each prosodic attitude. Each speaker participated in four sessions that took place at least one day apart.

2.2. Perception study

2.2.1. Participants

Forty-six (38 females) first year students (mean age = 21 years; range = 17-43 years) participated in the study for course credit. Participants reported normal or corrected-to-normal hearing and vision and were native speakers of Australian English.

2.2.3. Procedure

Each participant was tested in a sound-attenuated booth and wore Sennheiser HD 660S headphones. Participants were told that they would hear different speakers (presented one at a time) expressing five different attitudes using the phrase 'first class'. That is, the experiment was blocked and randomised by speaker and the order of trials was randomised within each speaker block. The participant was told that at the start of each speaker block, three audio recordings of the speaker saying the phrase 'first-class' in neutral naming mode would be presented to provide speaker specific calibration. Following this, the remaining five attitudes (criticism, doubt, warning, longing and suggestion) would be presented in random order. After each of these trials, they would see a screen with boxes containing the names of the attitudes. Their task was to choose the attitude they thought was being expressed. Participants were given a practice trial that used audio recordings by a speaker who was not among the ten speakers used for subsequent analysis. The participants were not given feedback on their responses. The display of trials and recording of responses was done by the DMDX system [5].

3. RESULTS

3.1 Behavioural results

The results were analysed by fitting a logistic mixed model to predict percent Correct recognition with the variables Attitude and Speaker as fixed factors (formula: Correct recognition ~ Attitude * Speaker). We used the AFEX R package [6] to obtain p-values; this package uses effect coding. The model included Participant and Item as random effects (formula: list(~1 | Participant, ~1 | Item)). The model's total explanatory power was low to moderate (conditional R^2 = 0.18) with the part related to the fixed effects

alone (marginal R^2) = 0.10. The results of this analysis are shown in Table 1.

	Df	Chisq	Chi Df	P-value
Speaker	48	219.14	4	.0000
Attitude	43	174.50	9	.0000
Attitude x Speaker	16	339.93	36	.0000

Table 1: Results of the regression model for the effect of Speaker, Attitude and their interaction.

There were significant main effects of Attitude and Speaker and a significant interaction between these variables. To illustrate the overall results, mean percent correct recognition scores are plotted by Attitude (Figure 1) and by Speaker (Figure 2).

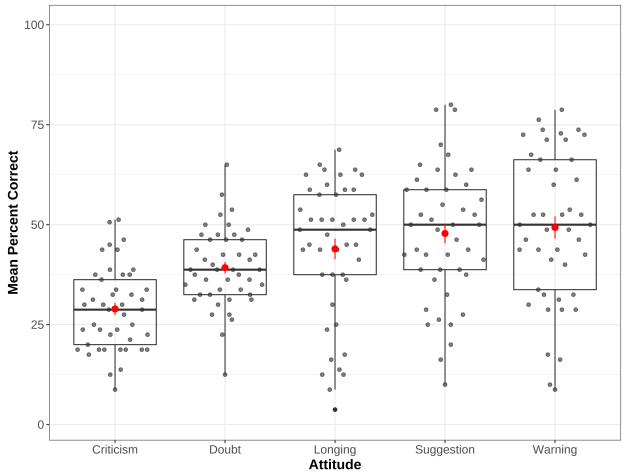


Figure 1: Mean percent correct recognition of each attitude (Tukey-style boxplot).

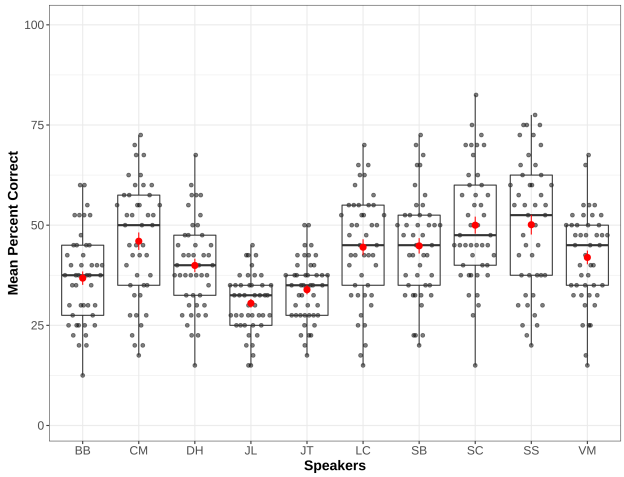


Figure 2: Mean percent correct recognition of each speaker (Tukey-style boxplot).

As can be seen in Figure 1, although recognition levels are better than chance (20%), they are well away from ceiling. The overall percent correct recognition was 41.8% (SE 1.2%). Figure 2 shows that there was considerable across-speaker variability in recognition scores.

3.2 Classification results

3.2.1. acoustic attributes

To represent the auditory productions, we used the 384 acoustic attributes of the 2009 Interspeech emotion challenge [7] via the openSMILE program{ref}. These attributes consisted of three types of low-level descriptors, broadly similar to those used in [3], i.e., zero-crossing rate, root mean square energy, pitch frequency, harmonics-to-noise ratio and 12 mel-frequency cepstral coefficients (numbers 1-12). Also, openSMILE's pitch autocorrelation function was used to calculate a voiced probability measure that represents the probability of a frame representing voiced speech. For each of these descriptors, delta coefficients (i.e., changes in the descriptor over time) were calculated, as well as statistical measures such as the mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range. In addition, linear regression was performed to calculate the offset, slope, and mean square error of the audio signal. Acoustic attributes were calculated for the utterances of each speaker. These data were combined, the results standardised and then each speaker's data was saved separately for classification.

3.2.2. Common acoustic attributes

A statistical classifier was used determine how well the expressed attitude of each speaker could be classified (see 3.2.3) and to determine the top 10 acoustic attributes used in each speaker model. We used the C4.5 classifier (a statistical classifier that uses a decision tree algorithm) with 4-fold, cross validation (see 3.2.3). The C4.5 uses normalized information gain as a splitting criterion and a pruning procedure to mitigate overfitting. We used C4.5 rather than a linear discriminant analysis (as used by [3]) because the acoustic attributes used in classification are clear.

Figure 3 is a chord diagram illustrating the distribution of the connections between each speaker and their top 10 acoustic attributes. The main point of the figure is to show the diversity of the acoustic attributes used in each speaker's classification model. A short code was used, rather than the names of each attribute as these could not be graphed. The attribute that 6 speakers had in common (A13), the quadratic error of linear F0 estimate; 5 speakers had A2 in

common (the probability of voicing); the remaining common features were related to estimates of F0.

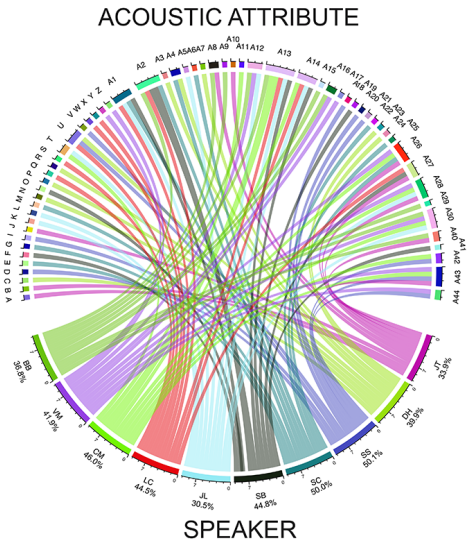


Figure 3: A chord diagram showing the connection between each of Speaker (shown with percent correct) and the top 10 acoustic attributes of 4-fold cross-validated classification models (using arbitrary coding, see text)

3.2.3. Model performance and recognition accuracy

Four-fold cross-validation was used to provide a measure of consistency of the calculated classification model for each speaker's data. That is, the performance of a machine learning model was evaluated by dividing the data into four equal-sized folds, or subsets. Four-folds were used because the data were collected over four sessions at least one day apart. The model was trained on three of the folds and tested on the fourth, and this process was repeated four times, with each fold serving as the test set once. If there is consistency in what properties are used to distinguish the different attitudes, then the model created on three folds should produce a good test result on the remaining fold. This entire process was conducted 10 times (each with a random selection of folds) and the average performance taken. This was done for each speaker.

Pearson correlation was used to determine the strength of the association between the performance of each speaker model and participant perception results. The results are shown in Figure 4. As can be seen the figure there was a significant positive correlation between the model and percent correct recognition ($r = .63$, $p = .0049$). This suggests that those speakers whose prosodic attitudes were better recognised had a more consistent use of auditory attributes.

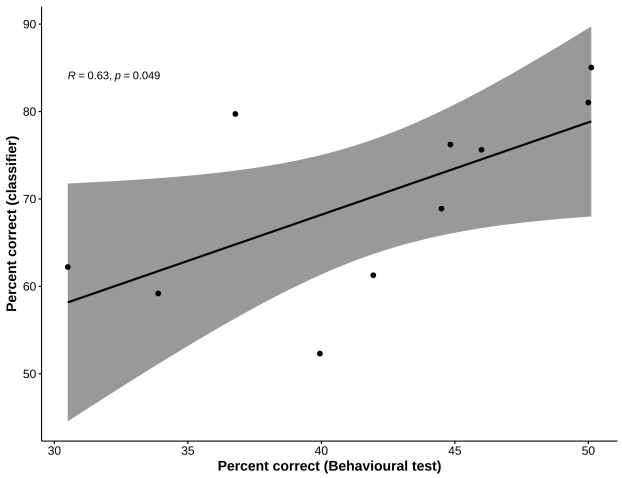


Figure 4: A depiction of the correlation between the behavioural recognition scores and classifier performance.

3.2.4. Model transfer

If prosodic patterns used to express different attitudes were more common (conventional) for speakers whose attitudes were better recognised, then the classification models for these speakers should be more transferable than those of speakers who attitudes were not well recognised. To assess this, we classified the auditory attributes of each speaker (using 10 4-fold cross validated classification runs) using the trained model of each other speaker. We grouped the speakers into those whose attitudes had high or low recognition scores and calculated the mean classification scores for these groups. The results showed that models trained and applied across the high group had significantly better classification (28.9 % correct, SE = 1.9%) than models trained and applied across the low group (19.5% correct, SE = 1.9%), $t = 3.412 < p = 0.0373$.

4. DISCUSSION

Overall, listeners recognised attitudes from prosody better than chance, but well below the level shown in [3]. The productions of speakers whose prosodic attitudes were better recognised than others, tended to be better classified and these classification models were better at classifying other well recognised productions. These results are consistent with some speakers having a partly conventionalised model of the prosodic forms that represent different attitudes. It seems plausible that with training a speaker could refine such a model to make each attitude more distinctive and hence better recognised (a kind of super–normal stimulus), perhaps like the voice coaches in the [3] study. Yet such highly distinctive models do not appear to occur in speakers with no specific training.

5. RFERENCES

[1] Ohala, J.J., 1996. Ethological theory and the expression of emotion in the voice. In: Proceedings of the International Conference on Speech and Language Processing. Vol. 3, Philadelphia, USA, pp. 1812-1815.

[2] Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

[3] Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. Journal of Memory and Language, 88, 70-86.

[4] Kim, J., & Davis, C. (2016). The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes. In INTERSPEECH (pp. 57-61).

[5] Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior research methods, instruments, & computers, 35, 116-124.

[6] Singmann, H., Bolker, B., Westfall, J., Aust , F. & Ben-Shachar, M. (2022). afex: Analysis of Factorial Experiments. https://afex.singmann.science/, https://github.com/singmann/afex

[7] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.

ALIGNMENT OF PROSODIC AND SYNTACTIC JUNCTURES AND VOWEL-INTIAL GLOTTALIZATION IN SYNTACTIC DISAMBIGUATION IN ENGLISH

Jae-Eun Jennifer Shin[1], Sahyang Kim[2] & Taehong Cho[1]

[1]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS), Hanyang University,
[2]Hongik University
jaeeunjshin@gmail.com, sahyang@gmail.com, tcho@hanyang.ac.kr

ABSTRACT

This study examines how ambiguity of coordinate structures is resolved by prosodic junctures in American English, and how vowel-glottalization serves as a prosodic boundary marker. Results indicated that syntactic junctures were unequivocally aligned with Intonational-Phrase boundaries (IP-boundary=#), critically placed before conjunctions ([***Anna***] # *and (#)* [***Annie-or-Angie***]; [***Anna-and-Annie***] # *or (#)* [***Angie***]), and optionally after them. Qualitatively, vowels were substantially glottalized both IP-initially and IP-medially, but spectral tilt measures revealed some boundary effects at the critical juncture, interacting with prominence: Vowels of conjunctions were glottalized more IP-initially (critical IP-boundary) only when pitch-accented, whereas no boundary effect was observed on vowels of nouns (optional IP-boundary). Moreover, both pre- and post-boundary lengthening were greater at the critical than optional IP-boundary. Finally, no prominence-induced glottalization was observed, and the opposite was true in some cases. These results suggest that the phonologically-defined IP category is realized not with the same phonetic content, but with phonetic modulation in reference to syntax and information structure.

Keywords: prosodic boundary, syntactic ambiguity, vowel glottalization, syntax-prosody interface, prominence

1. INTRODUCTION

In spoken language, speakers must present speech through a phonetic output to be understood clearly by the listener. For the phonetic form to arise, speakers go through phonetic encoding which is later decoded by the listener. Prosodic structure is considered to be an essential component of this process which regulates boundary and prominence marking to provide a detailed framework for articulation (see [3] for review). Such prosodic structuring can be further modulated in reference to other linguistic structures, possibly through the syntax-prosody interface that deals with prosodic phrasing and syntactic parsing [e.g., 7,10,12,13]. This interface may become more conspicuous when clearer manifestation of syntactic structure is demanded for resolving syntactic ambiguity that might arise, for example with coordinate structures (e.g., [A] and [B or C] vs. [A and B] or [C]). To resolve syntactic ambiguity, speakers often signify a specific syntactic structure by aligning a major syntactic juncture with a major prosodic juncture such as the one arising with the intonational phrase (IP) boundary.

Evidence of syntactic disambiguation through syntax-prosody mapping has indeed been cumulated in the literature [e.g., 7,10,13,14,18]. The notion of syntax-prosody interface may be related to how syntactic junctures are aligned with prosodic junctures [7,14]. Our view, however, goes beyond a simple mapping between the two junctures of prosody and syntax. Although prosodic boundaries may be phonologically defined, and thus treated in categorical terms, we hypothesize that the syntax-prosody interface can be defined in a much more granular way [3,12]—i.e., both suprasegmental and segmental features associated with a particular prosodic structure may be realized with fine-phonetic details that are tuned to reflect the underlying syntactic structure.

In the present study, we explore this possibility by examining how syntactically ambiguous coordinate structures in American English are resolved through syntax-prosody mapping, and how prosodic-structurally driven phonetic features may be realized in reference to syntactic structure. To this end, we will examine durational measures as well as a voice quality-related segmental feature as phonetic reflexes of prosodic juncture. For the durational features, preboundary lengthening, pause duration, post-boundary (initial) lengthening will be measured; and for the segmental feature, glottalization of vowels in vowel-initial words will be measured.

Glottalization refers to irregular creaky phonation that arise from low open quotient, low airflow and ventricular incursion [4]. Glottalization can be found in various contexts and be facilitated by several factors. [6] explains that glottalization in American English occurs amongst several segments or syllables. In American English, glottalization is most likely found for initial vowels [4,6,8,9]. As [4] suggests, the salient nature of creaky phonation may further enhance contrast amongst other acoustic cues. Therefore, glottalization employment helps with segmental contrasts, and a vowel hiatus occurring across a word boundary is often resolved with glottalization or an epenthetic glottal stop [4,5,9]. Glottalization is also found to occur with a vowel at a larger prosodic boundary [6,8,14] as well as under prominence [6, 8]. Thus, glottalization can serve as markers of both boundary and prominence presumably due to an increase in accompanied phonetic salience, relative to modal voice (cf. [8]). In the present study, we will examine how the degree of glottalization, as measured by spectral tilt measure (H1-H2) and noise-related measures HNR and CPP, may vary with the boundary strength that is mapped onto a major syntactic juncture, and how it may also manifest itself under prominence, especially when it occurs in a pitch-accented context compared to an unaccented context.

2. METHOD

2.1. Participants, materials and procedure

Fourteen native speakers of American English (7 male, 7 female; aged 19-35) participated in the acoustic recording

experiment. They had been living in Seoul Korea for a short-term stay.

Speech materials were constructed with coordinate structures containing [Name1] and [Name2] or [Name3] as in *'Anna and Annie or Angie'* which created different (structurally ambiguous) syntactic structures: *[N1] and [N2 or N3]* (the Early Closure construct) *versus [N1 and N2] or [N3]* (the Late Closure construct). The early closure construct has a major syntactic juncture after 'N1', and the later closure construct after 'N2.' A mini dialogue context was used to induce these two types of syntactic parsing to be produced by the participants in different contexts of information structure. As exemplified in Table 1, the mini dialogue contained a question (pre-recorded) and an answer (as the target sentence). When the question was *What is going on?,* the answer sentence to be provided by the participant had 'broad' focus; and when the question was *WHO will come to the party?,* the answer sentence had 'narrow' focus on all the noun phrases; and when the question was *Did they say (Anna and Annie) or (Angie) will come?*, the answer sentence had 'contrastive' focus, in the sense that the late closure construct meant by the question was corrected to be the early closure construct. Thus, the focus used in the present study did not fall on a particular lexical item, but either on the whole utterance or on a particular coordinate structure. We employed these contexts to induce various prosodic phrasing patterns that could be aligned with an intended syntactic structure. The three names were crossed over the three locations (*N1 and N2 or N3*), yielding 6 combinations, each of which was produced with two different syntactic parsings (Early vs. Late).

Table 1. Speech Materials according to Focus Type. Only the Answer category has been recorded. Early Closure has a syntactic juncture before 'and' whereas Late Closure (LC) has one before 'or'. Both cases have been recorded for all focus types with randomization of names and closure.

	Question	Answer	
Broad Focus	What is going on?	Well, (Anna) and (Annie or Angie) are coming.	**EC**
		Well, (Anna and Annie) or (Angie) are coming.	**LC**
Narrow Focus	WHO will come to the party?	Well, (Anna) and (Annie or Angie) will.	**EC**
		Well, (Anna and Annie) or (Angie) will.	**LC**
Contrastive Focus	Did they say (Anna and Annie) or (Angie) will come?	No. They said, (Anna) and (Annie or Angie) will.	**EC**
	Did they say (Anna) and (Annie or Angie) will come?	No. They said, (Anna and Annie) or (Angie) will.	**LC**

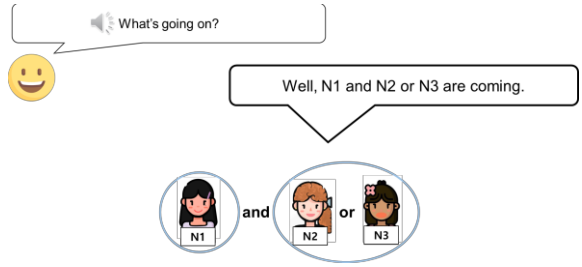


Fig. 1. Speakers were presented the target sentences as shown in this figure. Speakers were to produce the syntactic structure shown in the illustration

In the recording session, two coordinate structures to be produced by the participant were cued by a visual aid as illustrated in Fig. 1. Upon hearing an auditory prominent sentence (pre-recorded by a female speaker), the participant was instructed to produce the written sentence as an answer to a question prompt with a grouping of people as marked by circles displayed on a computer screen.

2.2. Measurements and analyses

Prosodic boundary types and pitch accents of all the obtained tokens were checked by all three authors (two of them were experienced ToBI transcribers), and the data were coded in terms of Boundary (IP, ip, Wd) and pitch accent (pitch accented vs. unaccented. The occurrence of an intermediate phrase (ip) boundary was relatively less frequent, and we excluded the ip tokens from analysis, so that we could observe a clear difference between the IP and the Wd boundaries. Names were generally produced with a pitch accent, so we did not analyse the pitch effects associated with N1, N2, and N3.

To examine glottalization of initial vowels, we used Voice Sauce [16, 17] to measure H1*-H2* (corrected, H1H2c, the difference between amplitudes of the first and the second harmonic) and two noise-related measures CPP (Cepstral Peak Prominence) and HNR (Harmonics to Noise Ratio). For HNR, we used HNR35 in Voice Sause to account for HNR up to 3,500 Hz. According to [9], H1-H2 (difference between amplitude of first and second harmonic) reveals vocal fold constriction with low values indicating creaky voicing. [9] further mentions phonation types can be sufficiently discriminated with additional measurements of CPP. H1H2 and CPP are considered crucial in evaluation. HNR is also a noise-related measurement that is utilized in non-modal voicing [4]. These parameters have been accounted for with Voice Sauce from three time points. The vowel was divided into three equi-interval portions, and these measures were taken from these three portions (Time points 1~3). To examine temporal variation along with the degree of glottalization, we also measured the duration of the phrase-final syllable (preboundary lengthening) and of the phrase-initial syllable (domain-initial/post-boundary lengthening) as well as the pause duration at the prosodic juncture. In the present study, we combined the preboundary lengthening and pause duration to assess the boundary-related lengthening as these two measures are often closely related to each other to signal a prosodic boundary.

A series of linear mixed-effect models were run using lme4 [2] package in R [15] with fixed factors of Boundary (IP, Wd), Closure (EC, LC), and Pitch Accent (accented, unaccented). Dependent variables were H1H2c, HNR35, CPP as well as preboundary lengthening + pause duration and initial duration (i.e., duration of the phrase-initial vowel). Note that we added the focus difference (broad, narrow, contrastive) as a control factor in the models, and do not report their effects in interaction with Boundary and Pitch Accent. Speaker differences have been accounted for with a random intercept for each model.

3. RESULTS

3.1. Syntax-prosody mapping

The results of prosodic transcription indicate that the major syntactic junctures in both the Early closure and Late

closure contexts were consistently aligned with an Intonational Phrase (IP). As shown in Fig. 2, for the Early Closure construct ([N1] # and [N2 or N3]), an IP boundary occurred after N1 without exception; and for the Late Closure construct ([N1 and N2] # or [N3]), an IP boundary occurred after N2 without exception. Thus, these are *critical* IP boundary conditions matched with the syntactic junctures. Along with this obligatory IP juncture, speakers occasionally placed an additional IP boundary elsewhere (12%, 14%) as shown in Fig. 2—e.g., for the Early Closure construct, in addition to the obligatory IP after 'N1', they placed an optional '(#)' IP boundary for about 12% of the time ([N1] # and [N2 (#) or N3]).

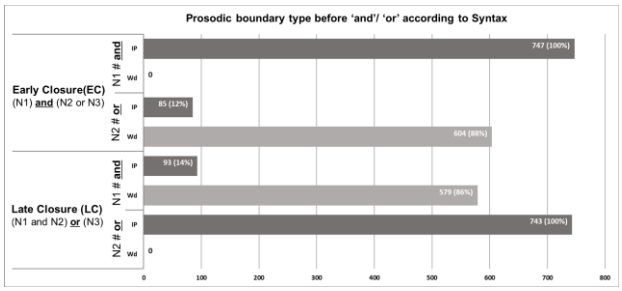


Fig. 2. Distribution of IP boundaries versus Wd boundaries in line with syntactic junctures of two different coordinate structures.

3.2. N2 and N3 (Second, Third Name)

Our qualitative phonetic transcription of the vowels of the names at N2 and N3 (as confirmed by all three authors) indicated that they were all produced with a clear percept of glottalization (creaky phonation) regardless of the boundary conditions. The results of spectral analyses indeed confirmed our auditory impression. There were no significant effects of Boundary at any measured point on any spectral measure: H1H2c, HNR35, and CPP, as can be also inferred from Fig. 3a-c. On the other hand, as shown in Fig. 3d, there was a significant effect of Boundary on preboundary lengthening (N2: β=-53.64, t=-6.12, p<0.001, N3: β=-33.51, t=-4.47, p<0.001). (Note that the pause duration was not included here because these nouns were not followed by any detectable pause even at an IP boundary). There was no boundary effect on the post-boundary syllable duration.

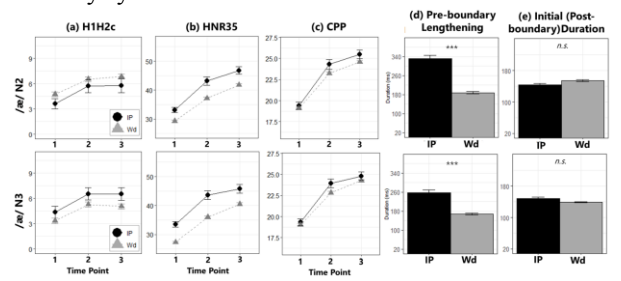


Fig. 3. Boundary effects on H1H2c, HNR35, CPP, Pre-boundary lengthening and Initial (post-boundary) duration for the vowel /æ/ of the second and third names (N2, N3). Error bars represent standard errors. The lower the spectral tilt values, the more glottalized (creakier).

3.3. Boundary effect on 'and' and 'or'

Our initial qualitative inspection of the voice quality of the vowels of 'and' and 'or' also indicated that they were all glottalized regardless of the boundary types. But our initial impression clearly varied as a function of whether these conjunctions were pitch accented or not.

For /æ/ of 'and', there was no Boundary effect on any of the measures, but there were significant interactions between Boundary and Accent for H1H2c, HNR and CPP (H1H2c: β=-2.74, t=-3.42, p<0.001, HNR35: β=10.47, t= 6.68, p<0.001, CPP: β<-0.01, t= 4.62, p<0.001). As shown in Fig. 4a-c (upper panels), 'and' was produced with more glottalization at the IP boundary than at the Wd boundary, especially in the accented condition, as reflected in a clear interaction between Boundary and Accent on H1H2. The other two noise-related measures showed some interaction effects, but not as clearly as H1H2c, indicating that there was no clear-cut boundary effect difference as a function of Pitch Accent. On the other hand, as shown in Fig. 4d, preboundary lengthening (+pause) showed a clear boundary effect (accented: β=308.47, t=5.75, p<0.001, unaccented: β=400.618, t=39.55, p<0.001) with no interaction with Pitch Accent. Interestingly, 'and' also showed a longer vowel duration of itself (initial duration), showing an initial strengthening effect (accented: β=60.55, t=11.03, p<0.001, unaccented: β=31.62, t=2.51, p<0.05),

The vowel of 'or' has also shown significant interaction between Boundary and Pitch Accent for HNR and CPP (HNR35: β=5.4, t=4.43, p<0.001, CPP: β=1.57, t=2.42, p<0.05), and H1H2c at Timepoint 2 (β<-0.01, t= 2.75, p<0.01). As shown in Fig. 4a-c (lower panels), there was a significant Boundary effect on H1H2c in the accented condition, but not in the unaccented conditions. However, results of HNR35 and CPP were not consistent with those of H1H2c (see Fig. 4b-c, lower panels), indicating that the noise-related measures did not provide interpretable results for the boundary effect on glottalization. As for the temporal measure, there was a clear boundary effect on preboundary lengthening (+pause) (accented: β=301.72, t=10.77, p<0.001, unaccented: β=280.47, t=29.23, p<0.001) (Fig. 4d, lower panel). There was also a boundary effect on initial (post-boundary) lengthening but only in the accented condition (β=38.87, t=3.62, p<0.001) (Fig. 4e, lower panel).

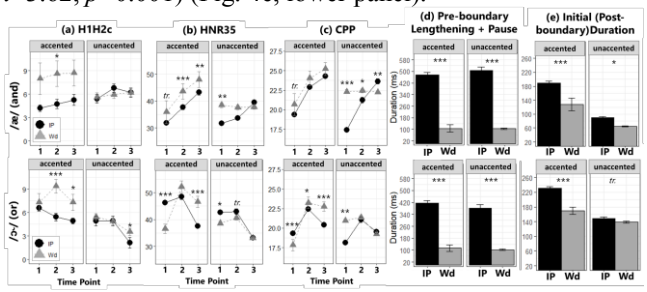


Fig. 4. Boundary effects on H1H2c, HNR35, CPP, Pre-boundary lengthening and Initial (post-boundary) duration for the vowel /æ/ of 'and' and /ɔ-/ of 'or' regarding pitch accent. Error bars represent standard errors. The lower the spectral tilt values, the more glottalized (creakier).

3.4. Prominence Effect on 'and' and 'or'

In the previous section, interactions between Boundary and Pitch Accent were focused on boundary effects. This section focuses on prominence (Pitch Accent) effects that may vary across different boundary conditions. The pitch accent effects are summarized in Fig. 5. For 'and', H1H2c indicated an increased glottalization at the IP boundary, but the opposite was true at the Wd boundary (Fig. 5a, upper panel). HNR35 and CPP did not show any consistent focus

effects, either. For 'or', all the measures indicated that if there was anything, vowels tended to be *less* glottalized in the accented than in the unaccented condition, as can be inferred from Fig. 5a-c (lower panels).

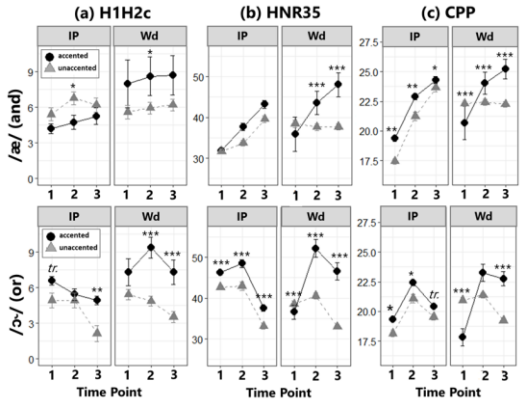


Fig. 5. Prominence effects on H1H2c, HNR35, CPP, Pre-boundary lengthening and Initial (post-boundary) duration for the vowel /æ/ of 'and' and /ɔ-/ of 'or' regarding boundary types. Error bars represent standard errors. The lower the spectral tilt values, the more glottalized (creakier).

4. DISCUSSION AND CONCLUSION

In the present study, we examined how different coordinate structures (*N1 and N2 or N3*) in English are expressed by prosodic boundary marking (or phrasing) to disambiguate otherwise structurally ambiguous surface forms, and how preboundary lengthening and glottalization may serve as phonetic reflexes of prosodic junctures that are mapped onto syntactic junctures.

Results indicated that syntactic junctures were invariantly aligned with IP boundaries. With the experimental task of grouping nouns of the phrase *Anna and Annie or Angie* as guided by a visual aid, speakers consistently produced an IP boundary that was aligned with a major syntactic boundary according to an intended syntactic parsing. For the Early Closure construct, a critical IP boundary was placed consistently (100%) at the major syntactic juncture between *N1* and the conjunction 'and' ([N1] # and [N2 (#) or N3]), and optionally (12%) an additional IP after N2. Similarly, for the Late Closure construct, a critical IP boundary was placed consistently (100%) at the major syntactic juncture between *N2* and the conjunction 'or' ([N1 (#) and N2] # or [N3]), and optionally (14%) an additional IP after N1. The alignment of major syntactic junctures with prosodic junctures are generally in line with the assumptions made in the literature on the Syntax-Prosody interface [10, 13, 14, 18].

Most remarkably, we observed some differential effects of the IP boundary depending on whether the IP boundary was used *critically* at the major syntactic juncture or optionally at a non-critical juncture (with prosody and syntax misalignment). At the critical juncture (with syntax-prosody alignment), a substantial temporal expansion was observed, which came from both preboundary and post-boundary lengthening as well as from the pause duration. Moreover, at this critical juncture, there was also a substantial increase in vowel-initial glottalization (as reflected in H1H2c, HNR and CPP) especially when in the pitch accented condition. On the other hand, at the non-critical prosodic boundary (when it was not aligned with a major syntactic juncture), the boundary effect was quite reduced. Not only did we fail to find the boundary-induced glottalization, but we also observed a reduced magnitude of preboundary lengthening with no detectable pause or post-boundary lengthening.

The phonetic expression of syntactic junctures may be mediated by prosodic boundaries (see [7] for a related discussion), and an IP boundary may be phonologically defined [11]. But, the clear differences in the way that phonetic reflexes of prosodic structure were realized depending on syntactic structures imply that phonetic encoding of prosodic structure makes direct reference to syntax at least to some extent.

Another related point to be made concerns the boundary effect on glottalization in interaction with prominence. Recall that spectral tilt measures revealed some boundary effects at the critical junctures of 'and' and 'or' only when pitch accented. In fact, this goes hand in hand with observations of [6] where phrase-initial syllables were more likely to be found with glottalization when they were also pitch accented. But we found no clear prominence-induced glottalization, and the opposite was true in some cases. This appears to run counter to what has previously been observed in English [8, 9] in which prominence may be an important trigger of glottalization. We do not have a clear account to offer, but our qualitative observations of those word-initial vowels indicated that they were all creaky regardless of the boundary and pitch accent types. In fact, glottalization may also be employed to resolve the vowel hiatus (in addition to the function of marking a larger prosodic boundary) [5]. Since detecting a word boundary in the vowel hiatus context may become harder in the absence of prominence, the salience of glottalization may come into play, counteracting a possible prominence effect.

Finally, our results leave a question unsolved as to whether the observed differential effects in relation to syntax-prosody alignment may have to do with different lexical properties of the target words. At the *critical* juncture, the target words were always conjunctions 'and' and 'or.' These were the ones produced with boundary-induced glottalization and post-boundary lengthening. On the other hand, at the non-critical juncture of syntax-prosody misalignment, the target words were nouns (*Annie, Angie, Anna*), which showed none of these effects. These observed differences may then be due at least in part to initial-strengthening neutralization for content words. [1] states that the lengthening of initial vowels is neutralized since it may otherwise hinder the information of stress within a word. Such prominence may be essential for content words especially for trochaic words as used in the present study, but not for function words which are often susceptible to reduction in various contexts. If so, vowels such as the initial /æ/ for the target word like *Anna* may not be subject to initial strengthening effects that are readily observable for the function words. But, this stress-related account does not explain why there was still a robust preboundary lengthening and a substantial pause at the critical juncture. Thus, our interim conclusion to be made is that our results still support the view that the phonetic encoding of prosodic structure is modulated in reference to syntax.

5. ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

6. REFERENCES

[1] Barnes, J. A. (2001). Domain-initial strengthening and the phonetics and phonology of positional neutralization. *Proceedings of the Northeast Linguistic Society (NELS)*, 32, vol. 1, pp.1-20.

[2] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

[3] Cho, T. 2022. The phonetics-prosody interface and prosodic strengthening in Korean. In S. Cho and J. Whitman (Eds.), *The Cambridge Handbook of Korean Linguistics* (pp. 248-293). Cambridge University Press.

[4] Davidson, L. 2021. The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(3), e1547.

[5] Davidson, L., & Erker, D. (2014). Hiatus resolution in American English: the case against glide insertion. *Language*, 482-514.

[6] Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of phonetics*, 24(4), 423-444.

[7] Elfner, E. 2018. The syntax-prosody interface: Current theoretical approaches and outstanding questions. *Linguistics Vanguard*, 4(1).

[8] Garellek, M. 2014. Voice quality strengthening and glottalization. *Journal of Phonetics*, 45, 106-113.

[9] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, 94, 101155.

[10] Jun, S. A. 2009. Prosody in sentence processing. *Ping Li (General ed.), Handbook of East Asian Psycholinguistics, Part III: Korean Psycholinguistics (C. Lee, G. Simpson, & Y. Kim, Eds.)*. London: Cambridge University Press. 423-43.

[11] Ladd, D. R. 2008. *Intonational phonology*. Cambridge University Press.

[12] Mitterer, H., Kim, S., & Cho, T. 2021. Glottal stops do not constrain lexical access as do oral stops. *PloS one*, 16(11), e0259573.

[13] Mitterer, H., Kim, S., & Cho, T. 2021. The Role of Segmental Information in Syntactic Processing Through the Syntax–Prosody Interface. *Language and Speech*, 64(4), 962-979.

[14] Nespor, M., & Vogel, I. 2012. *Prosodic phonology*. Dordrecht: Foris

[15] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[16] Shue, Y. L. 2010. *The voice source in speech production: data, analysis and models.* Los Angeles: University of Calinfornia

[17] Shue, Y. L., Keating, P., Vicenik, C., & Yu, K. 2011. VoiceSauce: A program for voice analysis. In Wai-Sum Lee & Eric Zee (eds.), Proceedings of the 17th International Congress of Phonetic Sciences, 1846-1849. Hong Kong.

[18] Snedeker, J., & Trueswell, J. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48, 103–130.

Variability in the stress group pattern of Copenhagen Danish with a focus on the Copenhagen Multiethnolect

Nicolai Pharao

Department of Nordic Studies and Linguistics, University of Copenhagen
nicolaip@hum.ku.dk

ABSTRACT

This paper presents an acoustic analysis of the variability in stress group patterns, i.e. the fundamental frequency contour associated with stressed and following unstressed syllables, in Copenhagen Danish with a particular focus on the register known as the Copenhagen multiethnolect. This is a register of Danish spoken by multilingual adolescents. It is prosodically distinct from the register used by monolingual peers. Previous work has shown a difference in vowel durations between the two registers with a near-neutralisation of the vowel length contrast in the multiethnolect. Recordings with 4 speakers from a larger corpus of map tasks containing trisyllabic words with initial stress as place names have been analyzed. All speakers show an expected rise in pitch from the stressed to the first posttonic syllable, but the rise is significantly less steep for speakers of the multiethnolect, which may contribute to the impression of a more syllable-timed rhythm.

Keywords: pitch, multiethnolect, sociophonetic variation, rhythm.

1. INTRODUCTION

Multiethnolects are registers typically found among adolescent speakers in areas where people of several different ethnic minorities live in close contact, typically urban areas. Multiethnolects have been documented in several European cities such as Berlin [1], London [2], and Stockholm [3]. There is also a multiethnolect in Copenhagen, the capital of Denmark. It was first described in [4], where a number of phonetic characteristics were mentioned, notably prosodic features. Segmental characteristics have also been documented for the Copenhagen Multiethnolect, notably a fronting of /s/ and palatalization of /t/ [5]. Most previous studies of this register have relied exclusively on auditory analysis, mainly of sociolinguistic interviews. An exception is [6] which used a small corpus of recordings of map tasks with adolescents from Copenhagen to find acoustic correlates of what [4] describes as the staccato-like rhythm of the Copenhagen Multiethnolect. The present study is based on the same corpus which will therefore be described in more detail below. [6] measured durations of stressed and unstressed short and long vowels in the target words that were included on the maps used in the map task. [6] found that for the vowels /i a u/, there was no difference in duration between phonologically long and short vowels in stressed syllables in the Multiethnolect, whereas the durational distinction was maintained in the speech of monolingual peers. This pattern was mainly due to a shortening of phonological long vowels by the speakers of Copenhagen Multiethnolect. This difference between the two registers, the Copenhagen Multiethnolect and Modern Copenhagen, the register used by monolingual peers, partly explains why the Copenhagen Multiethnolect is perceived as more staccato: the durational contrast between short and long vowels is partly neutralized making consecutive syllables in running speech more similar in this register than in Modern Copenhagen.

The prosody of the Copenhagen Multiethnolect is not only relevant in terms of providing a detailed phonetic description of the register, it has also been shown to be particularly relevant to speakers and listeners of Danish in general. When listeners are asked to provide characteristics about speakers on the basis of short stimuli, examples of recordings with the prosody typical of the Copenhagen Multiethnolect are almost invariably classified with reference to the speaker having an immigrant background even in the absence of segmental, grammatical or lexical cues. The details of the prosody of this register therefore warrants further research.

In the present paper, I use the corpus from [6] for a qualitative study of pitch-stress relations in four male speakers. Specifically, I examine the shape of the tonal contour associated with the prosodic stress group as defined in [7]. In this superpositional model of Danish intonation, Grønnum, on the basis of extensive acoustic studies of fundamental frequency variation in read speech, posits a number of components that are hierarchically ordered in this model of Danish intonation. The smallest unit is the stress group, which is initiated by a stressed syllable and includes all following unstressed syllables until the next stressed syllable. The tonal contour associated with this unit has the same shape regardless of the degree of stress or position in the utterance, although the magnitude of the excursions may vary, according to [7]. In [6] the authors observe

that the stress group pattern in the Copenhagen Multiethnolect while similar in shape to the stress group pattern of Modern Copenhagen, i.e. with a relatively low tone on the stressed syllable followed by a rise to the first posttonic and then a subsequent fall, the rise to the posttonic was not as steep in Copenhagen Multiethnolect as in Modern Copenhagen. This paper provides a qualitative pilot study of this reported difference with a comparison of the stress group patterns in the two registers.

2. METHOD AND MATERIALS

The results presented in this paper come from analysis of the corpus that was also used in [6]. The corpus consists of recordings of map task conducted by 12 pairs of adolescents (age range 13 to 15 years) who were all born and raised in the city of Copenhagen. 6 of the pairs consisted of speakers from monolingual homes and 6 of the pairs consisted of speakers from multilingual homes. All participants had Danish as a first language but the multilingual participants also used either Western Arabic, Farsi, Punjabi or Turkish at home with family members. While multiethnolects are not exclusively used by speakers with multilingual backgrounds, and although being multilingual does not automatically lead speakers to use a multiethnolect register, speakers with this background are more likely to use such a register because they live their everyday lives in areas where it is prevalent. Participants were put in pairs where the conversational partner could be reasonably assumed to use the same register so as to maximise the likelihood of participants using a non-standard register like the Copenhagen Multiethnolect despite the formal setting of being recorded. Participants were recorded directly to compact disc in a quiet room at the school they went to using professional recording equipment from the Linguistics Laboratory at the University of Copenhagen.

2.1 Map design and task

The participants were seated with their backs to each other, ensuring that they could not see each other's maps. One participant in a pair, the follower, had a blank map with only roads and landmarks on it, while the other participant, the guide, could see the names of roads and particular places in a fictional town. The task for the participants was for the guide to enable the follower to place a set of stickers with street names and labels for locations correctly on the blank version of the map. When one set of maps had been completed the two participants switched roles and completed another map task. The participants were told that the purpose of the experiment was to see how people help each other find their way, when they can only speak to each other. In other words, the instructions did not call attention to particular ways of speaking Danish.

The names of roads on locations on the maps were designed to control for a number of microprosodic effects, originally with a specific view to minimising intrinsic durational differences between vowels of different heights and vowel differences as a consequence of number of syllables in a word according to [6]. Thus the target words of the map tasks in this corpus are all words of three or four syllables. What is particularly useful for the purposes of the present study is that a subset of the road and place names contained the same vowel in the stressed and first posttonic syllable, e.g. [ˈminivɑjˀ], [ˈmanavɑjˀ] or had vowels of the same height in these poisitions e.g. [ˈ gůb̥i̥gåæːð̞ə̥]. Since I wanted to study the pitch movement of the stress group, it was desirable to control for intrinsic fundamental frequency as a function of vowel height [7]. We therefore only used the 17 target words where this condition was met.

2.2 Segmentation and measurement

The recordings have been transferred from the original compact discs in lossless wav-format and were thus available for analysis in Praat [8]. Only two pairs of speakers were analysed for the current study, one with two male adolescents speaking in the Modern Copenhagen register and the other pair two male adolescents speaking in the Copenhagen Multiethnolect register. Two recordings are available for each of these four participants, one as guide and one as follower, yielding a total og eight map task recordings to be analysed. In each of these eight recordings, the 17 target words were found in previous transcriptions and each vowel in every token segmented by hand. Each vowel was labelled for whether it was 1 stressed, 2 the first posttonic or 3 the syllable following the first posttonic. For example, in a token of the target word [ˈmanavɑjˀ] the [a] in [ma] was labelled 1, the [a] in [na] was labelled 2 and the [ɑ] in [vɑjˀ] was labelled 3.

Preliminary visual inspection of the raw fundamental frequency contours overall confirmed the expectation that the tonal contour of the stress group in both registers as represented by these four speakers would consist of a relatively low tone on the stressed syllable with a rise to the first posttonic and a subsequent fall. Therefore, and following previous studies of Copenhagen Danish, it was decided to measure the minimum fundamental frequency of

syllables with stress but the maximum fundamental frequency of the first posttonic so as to capture the turning points of the tonal contour. For the syllable following the first posttonic, the average fundamental frequency was measured. All measurements were made in semitones relative to 100 Hz and extracted automatically using the autocorrelation method in the Praat editor with pitch range set for 50 to 400 Hz.

3. RESULTS

Visualization of the measurements for each speaker are presented in Figures 1 and 2. Each column represents a recording, and each pair of columns represent a speaker of the pair. Each row in the figure represents a target word. The x-axis shows the labelling of the vowel, and the y-axis gives the measure in semitones. The colour of the lines represents the time of occurrence in the recording, but this will not be treated in the current study.

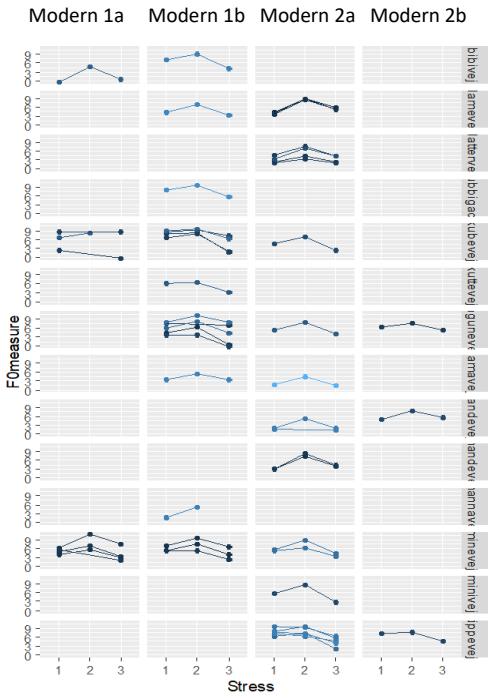


Figure 1: Stress group patterns for the two speakers of Modern Copenhagen.

The contours for the two Modern Copenhagen speakers tend to confirm the expectation that the stress group pattern consists of a low tone on the stressed syllable with a rise to the first posttonic and a subsequent fall. As can be seen, reliable measures were not always obtained for each vowel resulting in contours with less than three anchor points. What is also apparent is that not all stress groups exhibit the expected pattern. Looking at column number two (representing recording Modern 1b) we see in rows 7 and 12 indications that these words (*lagunevej* and *minevej* respectively) may occasionally exhibit a flat or even slightly falling contour from the stressed to the first posttonic. The same can be seen in the bottom row of column 4 (*tippevej*) which represents the other speaker of Modern Copenhagen.

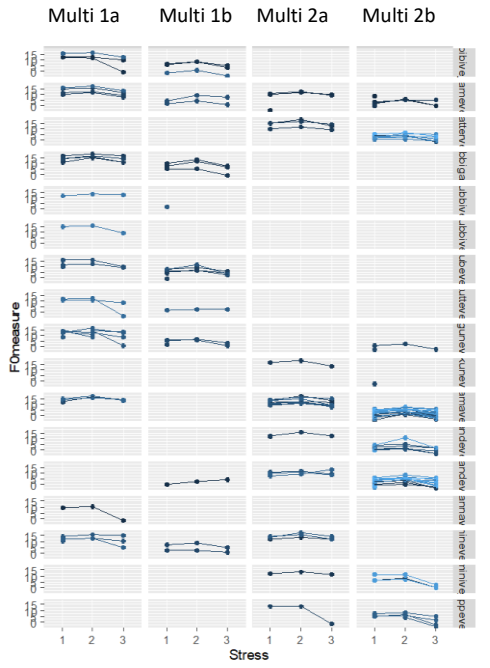


Figure 2: Stress group patterns for the two speakers of Copenhagen Multiethnolect.

The contours for the two Copenhagen Multiethnolect speakers given in Figure 2 tend to show quite a different pattern from this found in Modern Copenhagen. Almost all of the contours appear to be nearly level when it comes to the change in pitch from the stressed to the first posttonic. Note however, that the range in semitones is larger for Figure 2 than Figure 1. This is because one of the Speakers of Copenhagen Multiethnolect had higher general pitch than the other three. It is therefore helpful to zoom in on a selected few words to make a clearer comparison. This is done in Figures 3 and 4.

3.1 Individual words

Figures 3 and 4 have the same structure as Figures 1 and 2. Note that the ranges are still different on the y-axis, since all recordings were analysed in semitones relative to 100 Hz. This also means that the visual differences in magnitude cannot be directly compared, as a smaller difference will appear larger for the Modern Copenhagen speakers.

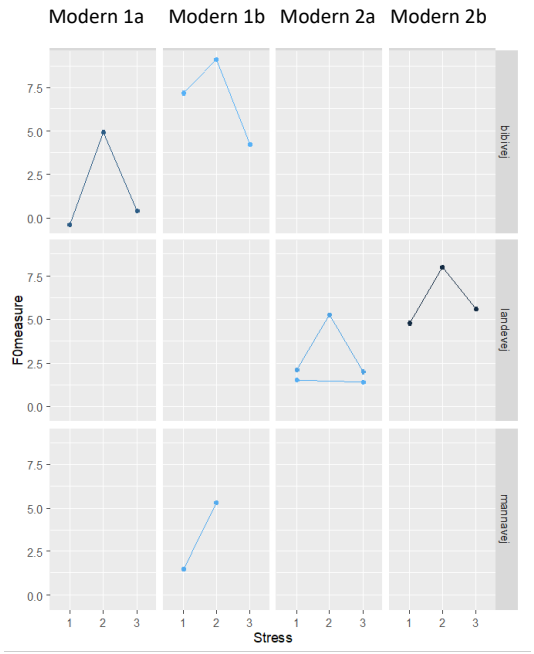


Figure 3: Stress group patterns for the words [ˈbi̯bi̯vɑjˀ], [ˈlanəvɑjˀ] and [ˈmanavɑjˀ] for the two speakers of Modern Copenhagen.

Zooming in on these three words, where the segmental structure controls for microprosodic effects on the measurements, the expected stress group pattern becomes clearer for the two Modern Copenhagen speakers: there is generally a steep rise of about 3 to 5 semitones (although again there are exceptions). An average difference of 2.9 semtiones is found for these two speakers (standard deviation of 2.2 st).

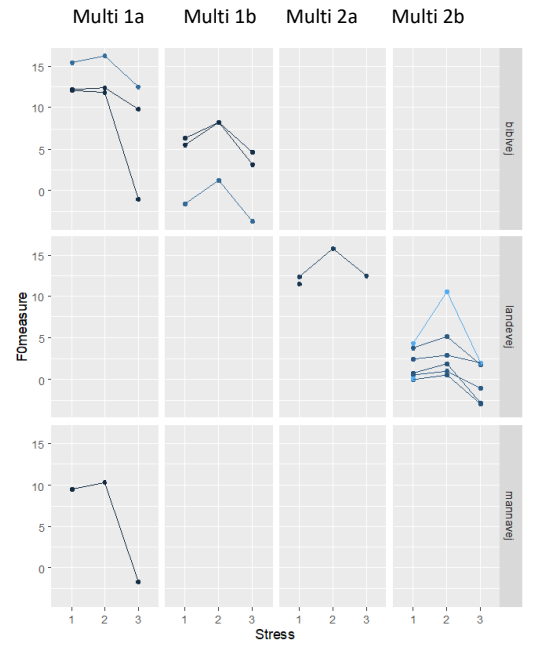


Figure 4: Stress group patterns the words [ˈbi̯bi̯vɑjˀ], [ˈlanəvɑjˀ] and [ˈmanavɑjˀ] for the two speakers of Copenhagen Multiethnolect.

For the individual words for the Copenhagen Multiethnolect speakers we see that while the contour is not flat, the rise from stressed to first posttonic indeed tends to be much less steep for this pair. While it may comprise a span of approximately 3 semitones, the tendency is for the slope to be between 0.5 and 2 semitones. An average difference of 0.5 semitones is found for these two speakers (standard deviation 2.4 st).

4. DISCUSSION

The analyses presented here only provide a qualitative look at the tonal stress group patterns in this pilot study of the differences between the two registers. While it is clear that further analysis of all of the recordings in the corpus is needed before more substantial claims on a solid quantitative foundation can be made, the pilot study also shows interesting findings with respect to the prosody of the Copenhagen Multiethnolect. There is evidence for the hypothesis that the rise in pitch from the stressed to the first posttonic tends to be less steep in the Copenhagen Multiethnolect compared to Modern Copenhagen. There was also greater variability of the tonal stress group pattern in Copenhagen Danish than what has been found in read speech (cf. [7]), since speakers of both registers produced tokens with no difference between the stressed and first posttonic. This is in spite of the fact that the analyses rest only on target words controlled for the intrinsic effect of vowel height. This could be due to the fact that these recordings are of spontaneous speech, where one might expect greater variability in general. However, a study of read speech in a number of different dialects of Jutlandic Danish [9] also indicates that the tonal stress group pattern may be more variable than reported in [7]. Furthermore, the variability appeared to be greater in the Copenhagen Multiethnolect with the two speakers analysed here showing a tendency to vacillate between the pattern found for the Modern Copenhagen of the two age matched monolingual peers and the less steep contour that appears more prevalent in the multiethnolect. If this greater variability emerges as a robust difference between the two registers, it would suggest that the perceived staccato-like rhythm is related to what [10] refers to as macro-rhythm: the stability and regularity of tonal units in the phrase. In this framework, the Multiethnolect speakers studied here could be classified as having a less macro-rhythmic pattern, because of the shallower pitch rise and the greater variability in the tonal contour associated with the stress group. Further analyses will help to clarify and substantiate the observations made here.

5. REFERENCES

[1] Weirich M., Jannedy S., Schüppenhauer G. 2020. The Social Meaning of Contextualized Sibilant Alternations in Berlin German. *Frontiers in Psychology*. 2020 Oct 29, 11, DOI=10.3389/fpsyg.2020.566174.

[2] Cheshire, J. 2020. Taking the longer view: Explaining Multicultural London English and Multicultural Paris French. *Journal of Sociolinguistics* 24: 308-327.

[3] Young, N.J. 2021. The sociolectal and stylistic variability of rhythm in Stockholm. *Language & Speech* (ETAP4 Special Issue on Sociolinguistic Variation in Prosody), 1–37.

[4] Quist, P. 2000. Ny københavnsk 'multietnolekt'. Om sprogbrug blandt unge i sprogligt og kulturelt heterogene miljøer. *Danske Talesprog* 1, 143-212.

[5] Maegaard, M. 2010. Linguistic practice and stereotypes among Copenhagen adolescents. In: Quist, P. & Svendsen, B.A. (eds), *Multilingual urban Scandinavia: New linguistic practices.* Multilingual Matters, Bristol,189-206.

[6] Pharao, N. & Hansen, G.F. 2010. Prosody in the Copenhagen Multiethnolect In: Quist, P. & Svendsen, B.A. (eds), *Multilingual Urban Scandinavia: New Linguistic Practices.* Multilingual Matters, Bristol, 79-95.

[7] Grønnum, N. 2022. Modeling Danish Intonation. In: Barnes, J. & Shattuck-Hufnagel, S. (eds), *Prosodic Theory and Practice.* MIT Press, Cambridge MA & London, 85-116.

[8] Boersma, P. & Weenink, D. 2022. Praat: doing phonetics by computer [Computer program].

[9] Jespersen, A.B., Šturm, P. & Hejná, M. 2021. On the feasibility of the Danish model of intonational description: Phonetic evidence from Jutlandic Danish. *Proceedings of Interspeech 2021.* 2611-2615.

[10] Jun, S.-A. 2012. Prosodic typology revisited: adding macro-rhythm. *Proceedings of Speech Prosody 2012*, 535-538.

PROSODIC ALIGNMENT IN DIFFERENT CONVERSATIONAL FEEDBACK FUNCTIONS

Carol Figueroa[1,2], Štefan Beňuš[3,4], Gabriel Skantze[1,5]

[1]Furhat Robotics;[2]Aix-Marseille University;[3]Institute of Informatics, Slovak Academy of Sciences;[4]Constantine the Philosopher University in Nitra;[5]KTH Royal Institute of Technology
carol@furhatrobotics.com, sbenus@ukf.sk, skantze@kth.se

ABSTRACT

Prosodic alignment is a phenomenon where the interlocutors' speaking style converge along various (para-)linguistic dimensions. Previous work has suggested that in terms of pitch, local alignment exists between backchannels and the preceding utterance of the interlocutor. In this paper, we propose a new operationalization of local prosodic alignment and investigate whether such alignment exists between short feedback utterances and the preceding, as well as the following, utterance of the interlocutor. Furthermore, we investigate whether this alignment differs between different feedback functions, including continuers, agreement, disagreement, yes/no responses, disapproval, non-understanding, sympathy, and mild/strong surprise. From the Switchboard corpus, we analyzed 2,118 instances of short feedback utterances with the mentioned 10 communicative feedback functions. Although we find significant results, they are much weaker than what has been reported in previous work.

Keywords: prosody, feedback, alignment, backchannels, entrainment

1. INTRODUCTION

The phenomenon where interlocutors become similar in their speaking styles is often referred to as alignment, entrainment or convergence [1]. Interlocutors may align their speaking styles among different dimensions, such as phonetic [2, 3, 4], acoustic-prosodic [5, 6, 7], lexical [8, 9, 10], and syntactic [11, 12]. Acoustic-prosodic alignment can be measured on a global or local level [7]. Global alignment refers to changes that take place over a longer time period (e.g., if speakers gradually converge towards a similar pitch level), whereas local alignment refers to local changes taking place in the vicinity of turn-exchanges. An example of the latter would be similarities of features between the beginning of an utterance and the ending of the preceding utterance of the interlocutor.

Previous work has found that the pitch similarity between backchannels (e.g., *yeah*, *mhm*) and the interlocutor's preceding utterance is greater than the pitch similarity of other turn-shifts [13]. They posit that it is due to this similarity that backchannels are perceived as unobtrusive in conversation. A similar analysis was done by [14], who concluded that pitch similarity between backchannels and the interlocutor's preceding utterance can be seen as an indication of entrainment or alignment between the speakers. However, it is not clear that this is a valid conclusion from this form of analysis. Heldner et al. [13] also found the pitch of backchannels to be higher in general, and it is well known that a rising pitch can serve as a backchannel-inviting cue [15], which could also explain the finding. Thus, we do not know whether the speakers truly align, in the sense that the listener changes the pitch level of the backchannel to 'tune-in' to the pitch level of the preceding utterance.

In this paper, we investigate local prosodic alignment between short feedback utterances and the preceding, as well as the following, utterance of the interlocutor. Following [16], we use a more strict operationalization of local alignment using Pearson's correlation, where we expect correlation between the prosodic features of the two interlocutors. The prosodic features we examine are pitch, pitch slope, and intensity. We also investigate whether local prosodic alignment differs for 10 different feedback communicative functions, namely continuers, agreement, disagreement, yes/no responses, disapproval, non-understanding, sympathy, and mild/strong surprise [17].

The results of our analyses can inform researchers who are interested in implementing feedback in spoken dialogue systems. There has been interest in implementing human alignment behavior in dialogue systems [18, 19]. If there is alignment for certain feedback functions in human-human interactions, this alignment can be implemented in spoken dialogue systems and can contribute to the

systems generating feedback in a more human-like manner. Our results can also be useful for better recognition of the user's feedback function by the system depending on whether the user aligns their prosodic features in their feedback to the system.

2. METHOD

2.1. Corpus and feedback functions

The short feedback utterances were extracted from Switchboard [20], which is a corpus consisting of about 2,500 telephone calls between 500 native speakers of American English. The dyadic pairs did not know each other and spoke between 3-10 minutes. Although the speakers were given a topic to discuss, they sometimes changed the topic of the conversation. The telephone conversations were recorded in two separate channels. The recordings were also transcribed and word level time-alignments were provided.

Feedback Function	Count
(C) Continue	1004
(U) Non-Understanding	61
(A) Agree	413
(D) Disagree	43
(Y) Yes response	57
(N) No response	107
(S) Sympathy	83
(MS) Mild Surprise	99
(SS) Strong Surprise	186
(Ds) Disapproval	65

Table 1: Feedback communicative functions and count used in analysis.

In order to annotate short feedback utterances, we first searched for lexical tokens that could potentially be instances of feedback; we used the list of lexical tokens provided in [17]. We then listened to these lexical tokens in context and assigned them one of the 10 feedback communicative functions presented in [17]. In total, we annotated 2,118 instances of feedback. The breakdown of the counts per feedback function are listed in Table 1.

2.2. Extraction of prosodic features

We used Parselmouth [21] to extract pitch (F0 Hz) and intensity (dB) every 10 ms from the speech signal. F0 values were first transformed to log scale and then z-score normalized. Intensity values were also z-score normalized. We z-score normalized the features by speaker using each speaker's mean and standard deviation computed from their entire conversation.

In order to obtain features for the preceding 500 ms of the interlocutor, we first extracted features from a window of 2000 ms preceding the start of the feedback. From this list of features we searched for the last voiced frame, then from the last voiced frame we took the preceding 49 frames. Concretely, this means that the preceding 500 ms of the interlocutor consists of the last voiced frame and the preceding 49 frames, 50 frames in total. For the following 500 ms of the interlocutor, we first extracted features from a window of 2000 ms following the end of the feedback. From this list of features we searched for the first voiced frame, then from the first voiced frame we took the following 49 frames, in total 50 frames.

2.3. Measuring local prosodic alignment

We measure local alignment (i.e. local synchrony [7]), as the Pearson's correlation coefficient of prosodic features between the preceding/following utterance of the interlocutor and the short feedback utterance. We decided to use Pearson's correlation to measure local alignment since it captures the following patterns:

- Positive correlation: prosodic features move up or down together, positive alignment
- Negative correlation: prosodic features move in opposite direction, negative alignment

We will interpret the correlations in the following way:

- Correlation between the listener's feedback and the preceding utterance of the interlocutor indicates that the listener aligns to the interlocutor.
- Correlation between the listener's feedback and the following utterance of the interlocutor indicates that the interlocutor aligns to the listener's feedback.
- Correlation between the preceding and the following utterance of the interlocutor does not indicate any alignment between the interlocutors, but the feedback can be seen as less obtrusive.

3. RESULTS AND DISCUSSION

In this section we report the Pearson correlation coefficient of the prosodic features between different feedback functions and the 500 ms preceding and following utterance of the interlocutor. Since we make 30 comparisons for each prosodic feature (10 functions * 3 correlations), a Bonferroni correction

should be applied to avoid Type I errors. In this case, correlations with $p < .0016$ ($\alpha = .05$) can be considered significant. However, since Bonferroni is a very conservative correction, we also report correlation coefficients which could be considered marginally significant ($p < .05$ and $p < .01$) and which could be interesting to explore in future work. We denote the strengths of correlation coefficients as: strong correlation ($r > .5$), moderate correlation ($.3 < r < .5$), and weak correlation ($0 < r < .3$).

We also conducted the analyses by excluding feedback that contained overlapping speech. However, this did not change the observed patterns in the results. Therefore, we report the results using the original 2,118 annotations of short feedback tokens.

3.1. Mean pitch

Figure 1 shows the significant results of the Pearson's correlations of mean pitch between the feedback functions and the preceding, and following 500 ms utterance of the interlocutor.

3.1.1. Feedback vs. preceding utterance

Using Bonferroni correction, we do not find any significant results indicating that the listener aligns their mean pitch of their feedback to the mean pitch of the preceding utterance of the interlocutor. Feedback function (C) *continue* has the same communicative function as backchannels. Heldner et al. [13] reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than other turn-shifts. Similarly, [14] reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than smooth switches, and that this similarity can be seen as an indication of entrainment or alignment. The differences in results could be either due to different methodologies or to the type of corpus. Both [13] and [14] used the same task-oriented corpus where preceding utterances of the interlocutor had a rising pitch [15] and the pitch of backchannels were also found to have a higher pitch in general [13]. It would be interesting to see future work adopting our methodology, to see if the listener 'tunes-in' to the pitch level of the interlocutor in task-oriented conversations.

3.1.2. Feedback vs. following utterance

We only find marginally significant results for feedback functions (A) *agree*, (S) *sympathy*, and (N) *no response*. Feedback functions (A), and (N) have

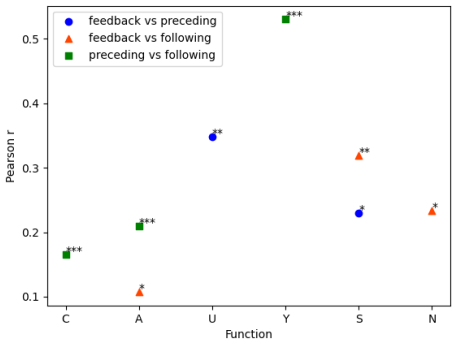


Figure 1: Pearson's correlation coefficients between the mean pitch of feedback functions and preceding/following 500 ms of the interlocutor. $*p < .05$, $**p < .01$, $***p < .001$

positive weak correlations, while (S) is moderately correlated. In these cases we see evidence that the interlocutor may align to the listener.

3.1.3. Preceding utterance vs following utterance

We find significant evidence that the mean pitch of functions (C) *continue*, (A) *agree*, and (Y) *yes response* can be seen as less obtrusive to the conversation. There is a weak positive correlation between the mean pitch of the preceding and following utterances of feedback function (C) ($r = 0.165$, $p < .001$). For the preceding and following utterance of feedback function (A) there is significant positive correlation ($r = 0.209$, $p < .001$). For the preceding and following utterance of feedback function (Y) there is significant and strong positive correlation ($r = 0.53$, $p < .001$). In the case of feedback function (Y), after the listener gives a *yes response*, the interlocutor may continue with the same pitch in a follow-up question or they continue speaking in the case where the listener responded to a rhetorical question.

3.2. Mean intensity

Figure 2 shows the significant results of the Pearson's correlation of mean intensity between the feedback functions and the preceding, and following 500 ms utterance of the interlocutor.

3.2.1. Feedback vs. preceding utterance

We find significant evidence that the listener may align their intensity to that of the interlocutor's for feedback functions (C) *continue* and (A) *agree*. Although we find significant positive correlation between the mean intensity of feedback function (C) and the preceding utterance of the interlocutor, the

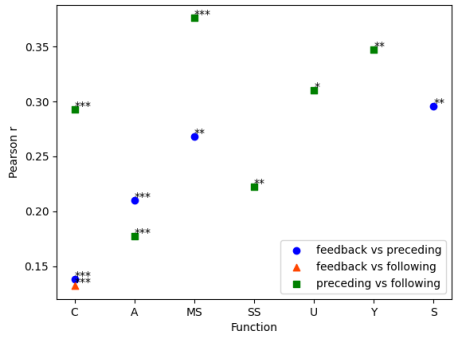


Figure 2: Pearson's correlation coefficients between the mean intensity of feedback functions and preceding/following 500 ms of the interlocutor. $*p < .05$, $**p < .01$, $***p < .001$

correlation coefficient is fairly weak ($r = 0.138$, $p < .001$). The correlation coefficient between feedback function (A) and the preceding utterance of the interlocutor is also significant but weak ($r = 0.21$, $p < .001$). For both function (C) and (A), we can speculate that the listener does not want to take the floor and will therefore 'tune-in' to the intensity of the interlocutor.

3.2.2. Feedback vs. following utterance

We only find evidence that the interlocutor may align their intensity to the listener's feedback function (C) *continue*. However, the positive correlation is weak ($r = 0.132$, $p < .001$). For feedback function (C) we now see the following alignment pattern: interlocutor A says something, listener B then produces a feedback function (C) which aligns its intensity to interlocutor A, then interlocutor A continues speaking and aligns its intensity to the preceding feedback function (C) of listener B.

3.2.3. Preceding utterance vs. following utterance

We find significant evidence that the intensity of the feedback functions (C) *continue*, (A) *agree*, and (MS) *mild surprise* may be perceived as less obtrusive to the conversation. The correlation coefficient between the preceding and following utterance for function (A) is positive and weak ($r = 0.177$, $p < .001$). The correlation between the preceding and following utterance of (C) is also weak, however the results are close to showing moderate correlation ($r = 0.293$, $p < .001$). The preceding utterance and following utterance of feedback function (MS) *mild surprise* are positive and moderately correlated ($r = 0.376$, $p < .001$). It is not surprising that these feedback functions are perceived as less obtrusive since the listener does not try to take the floor.

3.3. Pitch slope

Unlike mean pitch and intensity, we did not find any significant or marginally significant results across all the feedback function categories. This could be due to the way pitch slope was calculated. For example, in [15] they were computed by fitting least-squares linear regression models to the f0 values.

3.4. Limitations and future work

Our analysis does not account for temporal information that global alignment captures; there may be differences with how feedback functions align in the beginning of the conversation compared to the end of the conversation. Future work should also explore the difference in task-based conversations compared to spontaneous conversations and how this affects alignment. Different normalization methods for speech features might also affect the results of alignment. Levitan et al. [14] showed that latency can negatively affect local entrainment. Future work should investigate the difference between the end time of the preceding utterance of the interlocutor and the start of the feedback, as well as the difference between the end time of the feedback and the start time of the following utterance for all feedback functions.

4. CONCLUSION

We find that listeners align their intensity for feedback functions (C) *continue* and (A) *agree* to the intensity of the preceding utterance of the interlocutor. In terms of intensity for feedback function (C) we also find a pattern of alignment: when interlocutor A says something, listener B aligns the intensity of their feedback to that of interlocutor A, then interlocutor A continues speaking and aligns their intensity to the feedback of listener B. This can be useful for spoken dialogue designers who would like to implement human-like feedback. Our results also show that in terms of pitch and intensity feedback functions (C) *continue*, (A) *agree* can be perceived as less obtrusive in conversation.

5. ACKNOWLEDGEMENTS

This work was funded by the European Union's Horizon 2020 research and innovation program

under the Marie Skłodowska Curie grant agreement No 859588 and in part by the Slovak Granting Agency grant VEGA 2/0165/21, Slovak Research and Development Agency grant APVV-21-0373.

6. REFERENCES

[1] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics*, vol. 94, p. 101173, 2022.

[2] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Communication*, pp. 50–66, 2022.

[3] A. J. Olmstead, N. Viswanathan, T. Cowan, and K. Yang, "Phonetic adaptation in interlocutors with mismatched language backgrounds: A case for a phonetic synergy account," *Journal of Phonetics*, vol. 87, p. 101054, 2021.

[4] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[5] S. I. Levitan, J. Xiang, and J. Hirschberg, "Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue," in *Proc. Speech Prosody 2018*, 2018, pp. 532–536.

[6] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 325–334.

[7] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech 2011*, 2011, pp. 3081–3084.

[8] J. Kruyt, Š. Beňuš, C. Faget, C. Lançon, and M. Champagne-Lavau, "Prosodic and lexical entrainment in adults with and without schizophrenia," in *Proc. Speech Prosody 2022*, 2022, pp. 125–129.

[9] A. Weise, V. Silber-Varod, A. Lerner, J. Hirschberg, and R. Levitan, "Talk to me with left, right, and angles": Lexical entrainment in spoken Hebrew dialogue," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 292–299.

[10] H. Friedberg, D. Litman, and S. B. F. Paletz, "Lexical entrainment and success in student engineering groups," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 404–409.

[11] M. L. Allen, S. Haywood, G. Rajendran, and H. Branigan, "Evidence for syntactic alignment in children with autism," *Developmental Science*, vol. 14, no. 3, pp. 540–548, 2011.

[12] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[13] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proc. Interspeech 2010*, 2010, pp. 3054–3057.

[14] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *2015 AAAI Spring Symposium Series*, 2015.

[15] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proc. Interspeech 2009*, 2009, pp. 1019–1022.

[16] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, 2011, pp. 113–117.

[17] C. Figueroa, A. Adigwe, M. Ochs, and G. Skantze, "Annotation of communicative functions of short feedback tokens in switchboard," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1849–1859.

[18] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 111–118.

[19] R. Levitan, Š. Beňuš, R. H. Galvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Proc. Interspeech 2016*, 2016, pp. 1166–1170.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[21] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

EXTRACTION OF EMOTIONAL PROSODY FROM TELEPHONE CALLS TO HOSPITAL EMERGENCY DEPARTMENTS

Dongjun Wei[1], Mohamed Embarki[1], Oussama Barakat[2], Coralie Vaucherey[1,2], Thibaut Desmettres[2,3], Tania Marx[2,4] et Stephan Robert[5]
[1]ELLIADD EA 4661, université de Franche-Comté, Besançon
[2]Laboratoire de Nanomédecine, Imagerie et Thérapeutique EA 4662, université de Franche-Comté, Besançon
[3]Hopitaux universitaires Genève, université de Genève [4]Centre Hospitalier Régional Universitaire de Besançon
[5]Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud, Yverdon-les-Bains
dongjun.wei@edu.univ-fcomte.fr, mohamed.embarki@univ-fcomte.fr, oussama.barakat@univ-fcomte.fr, coralie.vaucherey@edu.univ-fcomte.fr, thibaut.desmettre@univ-fcomte.fr, tania.marx@univ-fcomte.fr, stephan.robert@heig-vd.ch

ABSTRACT

This study is based on the exploitation of telephone calls to hospital emergency departments in order to train artificial intelligence tools for the automatic recognition of vocal emotions for the purpose of improving emergency medical regulation. The main objective is to extract the most characteristic acoustic cues of vocal emotions. Very limited samples were selected from the audio databases of two University Hospitals in the Jura Arc (CHRU of Besançon in France and CHUV of Lausanne in Switzerland). Only extracts containing linguistic vocal emotions with a negative polarity (angst, anger, embarrassment and sadness) were retained for the study. These extracts were segmented and labeled, and their acoustic parameters (*F0*, intensity and duration) were measured manually. The preliminary results show that *F0* modulations seem more accurate to discriminate the four vocal emotions.

Keywords: vocal emotion, *F0*, emergency medical regulation, artificial intelligence

1. INTRODUCTION

Automatic Speech Recognition (ASR) has made significant progress both in terms of approaches and models [1], speaker recognition has also reached important steps [2]. Speech conveys both linguistic and non-linguistic information, such as emotions. The human ability to recognize emotions from verbal vocalizations has been repeatedly demonstrated [3]. Also, from non-linguistic vocalizations listeners are able to identify different types of emotion [4]. Apart from the lexicon, which can easily be attached to a type of emotion, prosody has been described as an important source in the expression of emotions [5]. Based on prosodic features, listeners have the ability to identify different types of emotions not only within their culture, but also cross-culturally [6,7]. Automatic emotion recognition has been developed for various applications [8]. A growing need for vocal emotion recognition systems is needed in the field of emergency medical regulation in order to be able to quickly identify the seriousness of the situation, direct towards the right course of care and/or engage the appropriate means of transport.

2. VOCAL EMOTIONS

Emotions cannot be reduced to a single process but are part of a complex set of processes [9]. They are grouped into two categories, basic emotions and extended emotions [10, 11]. Both are expressed by the modulation of linguistic (verbal) and/or non-linguistic (laughing, crying, screaming, moaning) vocalizations [12]. Vocal emotion research has focused more on basic emotions and their acoustic properties which are of 3 types: prosodic, spectral and voice quality. During strong emotional states, such as anger, significant acoustic modulations are visible, speech becomes higher pitched, shorter and more intense in auditory terms, faster and with less pauses. During a moderate emotional state, such as sadness, the speaker's voice becomes deeper, less intense, and the tone deeper, the rate slower, with more pauses. Ultimately, any modification of the emotional state is conveyed by modulations of the acoustic properties, such as the fundamental frequency (*F0*), the intensity and the duration [13]. Among all these cues, *F0* is generally recognized to discriminate more accurately vocal emotions.

In order to design an automatic vocal emotion recognition system suitable for the regulation of medical emergencies, it is necessary to better understand their acoustic properties.

3. METHODS

The audio data used in this study come from telephone calls to the medical emergency services of two hospitals, the CHRU of Besançon (France) and the CHUV of Lausanne (Switzerland). The telephone calls used are balanced in number (16 calls from each hospital) and gender (16 women and 16 men). The calls were first anonymized before processing. Two native French evaluators were recruited to label negative valence vocal emotions (angst, anger, embarrassment, fear and sadness). The labeling of vocal emotions was largely based on the subjective feelings of the labelers, however a defining explanation had been provided to them in order to better distinguish the callers emotions:

1. angst: anticipation of future threats
2. anger: reaction to the fact of not being respected, of not having one's needs met.
3. Embarrassment: reactions when one's behavior deviates from and falls below normative social standards.
4. fear: reaction to a real or imminent threat in the present.
5. sadness: reaction following the experience of an irreparable loss.

For each labeled extract, the unanimity of the two evaluators was required. In case of disagreement between

the two, a third evaluator was asked and if the disagreement persisted, the extract was removed from the sample. The extracts were labeled in Praat [14] and at the end of this step, it was observed that the evaluators frequently confused angst with fear. The choice was made to group the two vocal emotions within the angst category. The acoustic properties of the extracts selected in this study were measured manually using Praat. Non-linguistic vocal emotions were excluded from the study, and the linguistic vocal emotions retained all correspond to complete utterances, *i.e.* a total of 406 utterances.

The *F0* measurements targeted the four inflection points of each utterance (onset, offset, maximum and minimum) as well as the *F0* average. The conversion formula is 12*Log2(*F0* target/*F0* reference)=tonal difference (semitone) and 12 semitone=1 octave, which is converted to a chromatic gap when the *F0* gap is involved. The same points were measured for the intensity. Duration measurements targeted articulation rate and speech rate.

4. RESULTS

The abbreviations used in all tables are listed first by category as follows:

1) Ax.= angst, Ag=anger, E.=embarrassment, S.=sadness
2) Max.= maximum, Min.= minimum, Ons.=Onset, Off.=Offset, Avg.=average
3) W.=woman, M.=man
4) Fr.=France, Sw.=Switzerland
5) *F0*A=*F0* average, AI=intensity average, ARA= articulation rate average, ASR= speech rate average
6) SD=standard deviation, ST=semitone, ID=intensity deviation, DD=duration deviation

406 linguistic vocal emotions were classified according to the type of emotion (angst, anger, embarrassment and sadness), gender of the patient and his/her country：274 French extracts and 132 Swiss extracts were analyzed. Table 1 shows a wide variation in the recognition of the 4 emotions by the evaluators: with 84% for angst and 11%, 4% and 1% for embarrassment, sadness and anger respectively. Angst vocal emotion ~~has~~ obtained the highest recognition rate, while embarrassment, sadness and anger ~~have~~ had a lower recognition rate. This suggests that although embarrassment, sadness and anger vocal emotions are serious negative emotions, they are less common due to their occurrence in phone calls to medical emergency departments.

	Ax.		Ag.		E.		S.		
	M	W	M.	W	M	W	M	W	
All	342		5		43		16		406
	197	145	2	3	38	5	13	3	
Fr.	214		5		41		14		274
	127	87	2	3	36	5	12	2	
Sw.	128		0		2		2		132
	70	58	0	0	2	0	1	1	

Table 1: Number of linguistic vocal emotions classified by type of emotion, gender and country.

The following tables present acoustic measurements. Tables 2, 3 and 4 relates to *F0* measurements; Tables 5, 6 and 7 to intensity measurements and Tables 8 and 9 to duration measurements. Cells in the tables without data are marked in grey.

Table 2 deals with the *F0*A (Hz) and semitone difference (ST) of the four labelled emotions. ST 1 (2.3.4) represents gendered tonal difference in the same country, while ST5 represents tonal difference between patients of the same gender across countries.

		Fr.		Sw.		ST5
		*F0*A	SD	*F0*A	SD	
Ax.	W.	231	32	232	30	0
	M.	151	20	142	24	1
	ST1	7		8		
Ag.	W.	226	22			
	M.	143	10			
	ST2	8				
E.	W.	224	53			
	M.	133	19	108	19	4
	ST3	10				
S.	W.	232	40	331		6
	M.	131	16	101		5
	ST4	10		20		

Table 2: *F0*A (Hz), SD and ST in French and Swiss patients in four vocal emotions.

Table 2 reflects the following general patterns: 1. Women in different countries had higher *F0* average in angst and sadness vocal emotions and lower ones in embarrassment and anger, whereas men had higher ones in angst and anger vocal emotions and lower ones in embarrassment and sadness. 2. As far as angst vocal emotion is concerned, *F0* average and its differences were narrower across countries.

Table 3 deals with *F0* average (Hz) and tonal difference (ST) for inflection points for different callers in angst and anger vocal emotions. ST1 (2, 3 and 4) indicates tonal difference of the same gender across countries. ST5 (6) indicates gendered tonal difference of the same country. ST7 (8) relates to the tonal difference between *F0* onset/offset and *F0* maximum/minimum of the same gender in the same country.

		Ax.			Ag.		
		W.	M.	ST5	W.	M.	ST6
Ons.	Fr.	224	142	8	228	142	8
	Sw.	234	138	9			
	ST1	1	1				
Off.	Fr.	226	159	6	181	120	7
	Sw.	224	141	8			
	ST2	0	2				
ST7	Fr.	0	2		4	3	
	Sw.	8	0				
Max.	Fr.	360	237	7	345	220	8
	Sw.	323	209	8			
	ST3	2	2				
Min.	Fr.	137	108	4	155	73	13
	Sw.	162	99	9			
	ST4	2	1				
ST8	Fr.	17	14		14	19	
	Sw.	12	13				

Table 3: *F0*A (Hz) and ST at four inflection points in French and Swiss callers in angst and anger vocal emotions.

Table 3 shows the following general patterns: In angst vocal emotion, *F0* onset and offset average were similar for French women, whereas *F0* onset average was higher than the offset for Swiss women (8 semitones). French men *F0* onset average was lower than the offset (2 semitones); Swiss men *F0* onset and offset average was similar. Differences between *F0* maximum and minimum average

for men and women in both countries were not less than 1 octave.

Table 4 deals with *F0* average (Hz) with tonal difference (ST) at four inflection points in embarrassment and sadness vocal emotions.

		E.			S.		
		W.	M.	ID5	W.	M.	ID6
Ons.	Fr.	220	133	9	263	138	11
	Sw.		103		305	106	18
	ST1		4		3	6	
Off.	Fr.	205	138	7	217	111	12
	Sw.		118		447	89	28
	ST2		3		13	4	
ST7	Fr.	1	1		3	4	
	Sw.		2		7	3	
Max.	Fr.	328	217	7	291	194	7
	Sw.		142		463	115	24
	ST3		7		8	9	
Min.	Fr.	142	93	7	132	94	6
	Sw.		88		254	80	20
	ST4		1		11	3	
ST8	Fr.	15	15		14	13	
	Sw.		8		10	6	

Table 4: *F0*A (Hz) and ST at four inflection points in French and Swiss patients in embarrassment and sadness vocal emotions.

Table 4 reflects the following general pattern: 1. In ~~the~~ embarrassment vocal emotion, women's *F0* onset average in both countries was higher than the offset, 1 and 8 semitones for French and Swiss women respectively. In both countries, men's *F0* onset average was lower than ~~at~~ the offset (no less than 1 semitone). 2. In both countries, *F0* onset average for sadness vocal emotion was lower than the offset for men and women (not less than 3 semitones). The difference between *F0* maximum and minimum average for both French men and women was more than 1 octave~~,~~ while it was within 1 octave for both Swiss men and women.

Table 5 presents the intensity average (dB) and intensity deviation (ID) in different countries in the four vocal emotions. ID1 (2.3.4) represents the intensity average differences across genders in the same country, while ID5 represents the difference for the same gender across countries.

		Fr.		Sw.		ID5
		AI	SD	AI	SD	
Ax.	W.	74	3	75	7	1
	M.	71	5	72	5	1
	ID1	3		3		
Ag.	W.	74	3			
	M.	73	4			
	ID2	1				
E.	W.	71	4			
	M.	70	6	73	2	3
	ID 3	1				
S.	W.	74	3	80		6
	M.	74	2	69		5
	ID 4	0		11		

Table 5: intensity average (dB), standard deviation and intensity deviation (ID) for French and Swiss patients in the four vocal emotions.

The following general pattern is reflected in Table 5: intensity average in the four vocal emotions were very close. Table 6 deals with intensity average and intensity deviation (ID) at four inflection points in angst and anger vocal emotions. ID7 (8) relates to the difference in intensity deviation between the onset/offset and the maximum/minimum for the same sex in the same country.

		Ax.			Ag.		
		W.	M.	ID5	W.	M.	ID6
Ons.	Fr.	67	67	0	65	66	1
	Sw.	71	67	3			
	ID1	4	0				
Off.	Fr.	62	62	0	64	63	1
	Sw.	64	61	3			
	ID2	2	1				
ID7	Fr.	5	5				
	Sw.	7	6				
Max.	Fr.	82	82	0	87	81	6
	Sw.	84	81	3			
	ID3	2	1				
Min.	Fr.	28	28	0	25	32	7
	Sw.	35	31	4			
	ID4	7	7				
ID8	Fr.	54	54				
	Sw.	49	50				

Table 6: intensity average (dB) and intensity deviation (ID) at four inflection points in French and Swiss patients in angst and anger vocal emotions.

Table 6 shows the following general pattern: for angst and anger vocal emotions, onset and offset intensity average were not very different from patient to patient, as was intensity average at the maximum and minimum.

Table 7 deals with intensity average (dB) and intensity deviation (ID) at the four inflection points in embarrassment and sadness vocal emotions.

		E.			S.		
		W.	M.	ID5	W.	M.	ID6
Ons.	Fr.	67	68	1	70	69	1
	Sw.		71		72	75	3
	ID1		3		2	6	
Off.	Fr.	55	61	6	75	64	11
	Sw.		62		62	56	6
	ID2		1		13	8	
ID7	Fr.		7		5	5	
	Sw.		8		10	9	
Max.	Fr.	80	80	0	82	83	1
	Sw.		82		88	78	11
	ID3		2		6	5	
Min.	Fr.	26	27	1	72	29	43
	Sw.		41		46	26	20
	ID4		13		26	3	
ID8	Fr.		53		10	54	
	Sw.		41		42	52	

Table 7: intensity average (dB) and intensity deviation (ID) at four inflection points in French and Swiss patients with embarrassment and sadness vocal emotions.

The following general pattern is reflected in Table 7: as for the two previous vocal emotions, onset and offset intensity average in embarrassment and sadness were not very pronounced, and the same was true for intensity average at the maximum and the minimum.

Tables 8 and 9 concern the speech rate and articulation rate in the two countries in each of the four vocal emotions. DD1 (2.3.4) indicates the duration difference in the same gender in different countries, whereas DD5 (6) indicates gendered duration difference in the same country.

	Fr.				DD5
	SRA	SD	ARA	SD	

Ax.	W.	5	1	5	1	0
	M.	5	2	5	1	0
	DD1	0		0		
Ag.	W.	5	0	6	0	1
	M.	3	0	4	0	1
	DD2	2		2		
E.	W.	5	1	6	1	1
	M.	5	1	5	1	0
	DD3	0		1		
S.	W.	5	2	5	1	0
	M.	4	2	5	2	1
	DD4	1		0		

Table 8: duration average (syl/s), standard deviation and duration difference (syl/s) in French patients in four vocal emotions.

		Sw.				DD5
		SAR	SD	ARA	SD	
Ax.	W.	5	1	5	1	0
	M.	5	2	6	2	1
	DD1	0		1		
Ag.	W.					
	M.					
	DD2					
E.	W.					
	M.	4	2	4	2	0
	DD3					
S.	W.	4		4		0
	M.	4		5		0
	DD4	0		1		

Table 9: duration average (syl/s), standard deviation and duration difference (syl/s) in Swiss patients in four vocal emotions.

The following general pattern is reflected in Tables 8 and 9: the differences in speech rate and articulation rate between callers were not very significant in the four vocal emotions.

5. DISCUSSION

Regardless of gender or country, data presented here showed that linguistic vocal emotions were discriminated mainly by *F0* modulations, intensity and duration showed minor distinctions. *F0* differences of vocal emotions can be broadly summarized as follows:

5.1 *F0* average

5.1.1 Same gender in the same country

French women highest vocal emotion *F0* average was, in descending order, sadness, angst, anger and embarrassment, whereas French men highest *F0* average was anxiety, followed by anger, embarrassment and finally sadness. The highest *F0* average among Swiss women was for sadness, followed by anxiety (as a reminder, data were missing for anger and embarrassment). Among Swiss men, the highest F0 average was for anxiety, followed by embarrassment and finally sadness (data were missing for anger).

5.1.2 Gender differences in the same country

F0 average was consistently higher for French women than for men across the four vocal emotions, with the largest tonal difference in sadness and embarrassment and the smallest in angst and anger. For Swiss women, *F0* average was consistently higher than for men in all four vocal emotions, with the largest tonal difference in sadness, which remained above 1 octave, and the smallest in angst, which remained within 1 octave.

5.1.3 Same gender across countries

F0 average was higher for Swiss women than for French in sadness, with tonal difference within 1 octave, while in angst, the difference was not significant. *F0* average was higher for French men than for Swiss in angst, embarrassment and sadness, but the tonal difference remained within 1 octave.

5.2. *F0* average at four inflection points

French women *F0* onset average was generally higher in anger, embarrassment and sadness vocal emotions than the offset, with a tonal difference of 1 octave or less, while the difference was not significant in angst vocal emotion. In contrast, *F0* maximum average was 1 to 2 octaves higher than *F0* minimum average. French men *F0* onset average in angst and embarrassment vocal emotions was generally 3 semitones lower than the offset, whereas in anger and sadness vocal emotions the onset was 4 semitones higher than the offset. Swiss men *F0* onset average was generally 2 semitones lower than the offset in angst and embarrassment vocal emotions whereas it was 3 semitones higher than the offset in sadness.

5.2.1 Gender difference in the same country

French women *F0* onset and offset average in the four vocal emotions was consistently higher than for men, however tonal difference remained within 1 octave for all of them. The same was true for *F0* maximum and minimum average. Swiss women onset and offset *F0* average in angst and sadness vocal emotions were consistently higher than those of men, with a tonal difference of approximately 1 octave or less in angst and more than 1 octave in sadness. The same was true for *F0* maximum and minimum average.

5.2.2 Same gender across countries

French women *F0* onset average in angst and sadness vocal emotions was 3 semitones lower than that of Swiss women –It was the opposite for *F0* offset average, with a small tonal difference in angst vocal emotion and a tonal difference of more than 1 octave in sadness. French women *F0* maximum and minimum average was lower than that of Swiss women, tonal difference was within 1 octave. The situation was reversed for men in both countries, tonal difference was also within 1 octave.

Thus, vocal emotion acoustic properties in both countries showed different *F0* patterns, and vary very little in intensity and duration. Swiss Results were more complex and lack sufficient data support for some vocal emotions (anger and embarrassment), either because the prevalence of the Swiss accent compared to the French accent influences the labelers' judgment, the relatively small number of overall emotion samples, or because the corpus is composed of purely natural speech, which differs markedly from previous studies based on a large corpus of simulated emotions.

6. CONCLUSION

The results show that, among the 3 acoustic properties analyzed which are *F0*, intensity and duration, the modulations of *F0* are the most representative of the change in linguistic vocal emotion, and more particularly the modulations in *F0* contour and *F0* average. On the other hand, the intensity and the duration distinguish only very weakly the linguistic vocal emotions. Therefore, in the design of automatic voice emotion recognition system, we can assume that the data extraction of *F0* remains the basic acoustic property for natural linguistic voice emotion recognition.

7. REFERENCES

[1] Ainsworth, W.A. 1999. Some approaches to automatic speech recognition. In: Hardcastle, W., Laver, J. (eds), *The Handbook of Phonetic Sciences*. Blackwell, 721-743.

[2] Nolan, F. 1999. Speaker recognition and Forensic phonetics. *The Handbook of Phonetic Sciences*. Blackwell, 744-767.

[3] Juslin, P.N., Laukka, P. 2003. Communication of Emotions in Vocal Expression and Music PerFormance: Different Channels, Same Code? *Psychological Bulletin* 129, 770-814.

[4] Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. 2016. The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion* 16, 117–128.

[5] Keltner, D., Tracy, J., Sauter, D. A., Cordaro, D. C., & McNeil, G. 2016. Expression of emotion. In L. F. Barrett & M. Lewis (eds.), *Handbook of emotions*, GuilFord Press, 467-482.

[6] Cowen, A.S., Elfenbein H.A. Laukka, P. & Keltner, D. 2019. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist* 74, 698-712.

[7] Cowen, A.S., Laukka, P., Elfenbein H.A., Liu, R. & Keltner, D. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour* 3, 369-382.

[8] Yanga, N., Deyb, N., Sherrattc, R.S. & Shia, F. 2020. Recognize basic emotional statesin speech by machine learning techniques using mel-frequency cepstral coefficient features. *Journal of Intelligent & Fuzzy Systems* 39, 1925-1936.

[9] Wager, Tor D., Krishnan A., Hitchcock, E. 2018. How are emotions organized in the brain? *The nature of emotion: Fundamental questions*. OxFord, 112-118.

[10] Ekman, P. 1972. Universal and cultural differences in facial expression of emotions. In Cole, J. (ed.), *Nebraska symposium on motivation*. Lincoln; 207-283.

[11] Plutchik, R., Kellerman, H. (eds.). 1980. Emotion: theory, research, and experience. *Theories of emotion*. Academic Press, Vol. 1, 3-33.

[12] Bryant, G.A. 2021. The evolution of human vocal evolution. *Emotion Review* 13, 25-33.

[13] Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227–256.

[14] Boersma, P. (2017). Praat, a system For doing phonetics by computer. *Glot International* 5:9/10, 341-345.

THE DOMAIN OF BOUNDARY-RELATED LENGTHENING IN TOKYO JAPANESE, A MORA-TIMED LANGUAGE

Sherry Chien, Karen Tsai, Argyro Katsika

University of California, Santa Barbara
sherrychien@ucsb.edu, karentsai@ucsb.edu, akatsika@linguistics.ucsb.edu

ABSTRACT

Although boundary-related lengthening is well-established, it is unclear whether the stretch of speech affected is determined by a grammatical domain (e.g., mora, syllable, word) or whether it corresponds to a fixed interval at the boundary. Here, we use Electromagnetic Articulography (EMA) to address this issue in Tokyo Japanese, a mora-timed language. Disyllabic initial-accented words with all possible combinations of moraic structure in each syllable were tested in phrase-final and phrase-medial positions. Analyses of consonant gestures in each syllable's onset show that boundary-related lengthening extends beyond the phrase-final mora or syllabic rhyme to the onset of the phrase-final syllable when the latter is light (CV), but not when it is heavy (CVC or CVV). No evidence for boundary-related lengthening was found in the onset of the penultimate syllable, suggesting that the phrase-final word is not the domain of the effect. Results are discussed within the framework of Articulatory Phonology.

Keywords: boundary lengthening, moraic structure, Tokyo Japanese, articulation, Articulatory Phonology

1. INTRODUCTION

Phrase-final lengthening is the phenomenon in which acoustic and articulatory durations are longer in phrase-final positions as compared to their phrase-medial counterparts (e.g., acoustics: [1]–[5]; articulation: [6]–[11]). The effect is so widely attested in numerous languages and language varieties that it is considered a language universal [12, 13]. Yet, the scope of the effect, i.e., the stretch of speech affected, remains an understudied and ambiguous issue. An important parameter that stays unknown is whether the scope of the effect is determined by a grammatical domain. For instance, there is evidence suggesting that the effect can extend over several pre-boundary segments, with the exact domain of the effect varying, possibly on a language-specific manner, from the phrase-final syllable rhyme [1, 14, 15] to the phrase-final foot [4] or even word [16]. Alternatively, the effect may not scope over a grammatically defined domain, but instead over a fixed interval at the boundary [17]. Furthermore, it seems that the scope and degree of the effect is further fine-tuned by several factors that also affect phonetic duration, such as segment type, vowel length, syllable structure as well as language-specific features. For instance, in English, phrase-final lengthening has been found to apply to the rhyme of the phrase-final syllable [5], while other studies have suggested that the effect begins earlier, extending to the onset of the final syllable in cases when its vowel is reduced or is not a diphthong [1]. Another factor that emerges from the literature as playing an important role in determining the scope of the effect is position of lexical stress and/or phrase-level pitch accent [10] (See also [4], [5]; but see [18]). Recent work in Japanese shows interactions of the effect with lexical pitch accent as well [19].

Here, we also turn to Japanese. While the majority of studies have looked into boundary-related effects in stress languages like English and Dutch, where the most relevant phonological unit is mainly the syllable, mora-timed languages like Japanese can offer an interesting perspective on the domain of phrase-final lengthening, since the candidate domains of mora, syllable and word can be disentangled from each other. Previous findings on this issue in Japanese has been limited and inconclusive. An acoustic study by Shepherd [20] has found that phrase-final positions showed proportionally greater lengthening in short vowels than long vowels, and took this to mean that the domain of final lengthening in Japanese is one mora. On the other hand, studies from Seo et al [21, 22] suggest that the scope of boundary lengthening is better explained by syllabic rather than moraic structure, as lengthening on final CV syllables was comparable to that of the combined effect of lengthening in the two moras in the rhyme of a final CVN syllable. Moreover, Seo et al's [22] study showed that the degree of phrase-final lengthening was greater when the penultimate syllable had a moraic nasal. Finally, the number of segments in the rhyme was found to affect the degree of lengthening, as the scope of boundary was found to be more limited in words that have more segments (e.g., CVN.CVN) compared to those that have fewer segments (e.g., CV.CV).

In this study, we examine the interaction between moraic structure and phrase-final lengthening in Tokyo Japanese via Electromagnetic Articulography

(EMA). Disyllabic initial-accented words with all possible combinations of moraic structure in each syllable (see Table 1) are tested in phrase-final and phrase-medial positions. If the domain of phrase-final lengthening is the syllable, we expect to see the final syllable to lengthen in test disyllabic words with various moraic structures [21, 22]. If the domain of phrase-final lengthening is the mora, we expect to see lengthening limited to the final mora across the different syllable weights [20]. If the scope is domain-agnostic, effects should be seen on a specific interval at the end of the phrase (e.g., [17]). If lengthening is attracted to heavy non-final syllables, long vowels and geminate consonants should have the same effect as moraic nasals in Seo et al's [21, 22] research.

2. METHOD

2.1. Participants and recording apparatus

Participants were 6 native speakers of Tokyo Japanese (1 male and 5 females, M_{age} = 22 years, range 20-25 years). All participants could read Japanese orthography and reported no speech, hearing, or vision problems. They were not aware of the purpose of the experiment and were compensated for their participation.

Data were collected via the AG501 three-dimensional electromagnetic transduction device (Carstens Medizinelektronik). Five sensors were attached to tongue dorsum, tongue body center, tongue tip, upper lip, and lower lip. An additional five sensors were attached as reference points: upper incisor, lower incisor, nose, and left and right ears. Audio recordings were performed simultaneously with the kinematic recordings with a Sennheiser shotgun microphone set at a sampling rate of 16 kHz.

2.2. Experimental design and procedure

Test words were 11 disyllabic initial-accented words with every possible combination of moraic structure in the first and second syllable, as shown in Table 1.

All test words were embedded in frame sentences in a phrase-medial (PhM) control position or in a phrase-final (PhF) test condition, as illustrated in Table 2. Frame sentences were controlled for length (number of moras) and pitch accent of the following verb, which was varied to create sensical utterances. Short dialogues were used to create contexts that prompted frame sentences embedded with the target words as responses.

All sentences and stimuli were presented in Japanese orthography on a computer screen placed about 3 feet away from the participant. Context sentences were read silently and target sentences aloud. Altogether, there were 2 phrasal positions, 10 test words (manipulating moraic structure), 9 repetitions, and 6 speakers for a total of 1188 utterances collected.

Mora Count	Moraic Structure	Syllabic Weight	Test Word	Gloss
2	CV.CV	LL	ma*mi	'name'
2	CV.CV	LL	na*mi	'name'
3	CV.CVV	LH	ma*ni:	'money'
3	CV.CVN	LH	bi*naɴ	'handsome man'
3	CVV.CV	HL	bi:*mu	'beam'
3	CVN.CV	HL	baɴ*bi	'southern area'
3	CVC.CV	HL	bag*gu	'bag'
4	CVN.CVV	HH	meɴ*bo:	'cotton swab'
4	CVV.CVV	HH	ba:*bi:	'Barbie'
4	CVV.CVN	HH	ma:*biɴ	'name'
4	CVC.CVN	HH	mak*kuɴ	'name'

Table 1: Test words by moraic structure and resulting syllable weight. L = light syllable, H = heavy syllable

Phrasal Position	Test Sentence
Phrase-medial (PhM)	[hoɴto: ni **na*mi** makasi̥ta?]IP really Nami defeated? 'Really (you) defeated Nami?'
Phrase-final (PhF)	[hoɴto: ni **na*mi** ?]IP [makasi̥ta?]IP really Nami defeated? 'Really Nami? (You) defeated (her)?'

Table 2: Sample frame sentences for test word na*mi.

2.3. Data analysis

Kinematic labels for all consonants (C) were made using a semiautomatic procedure that identifies constriction gestures (Tiede, Haskins Laboratories). Labial C gestures (/m/) were labelled on the lip aperture tract, coronal C gestures (/t, d, n/) on the tongue tip vertical displacement tract and dorsal C gestures (/k, g/) on the tongue dorsum vertical displacement tract.

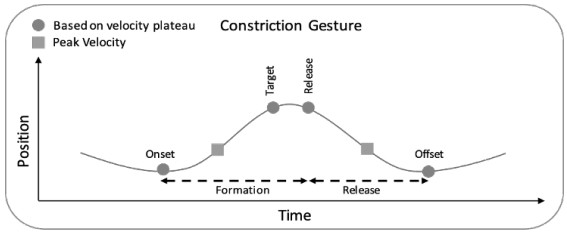


Figure 1: Schematic representation of a gestures and labelled timepoints.

The labelling procedure detected the following timepoints in each test C gesture using velocity criteria: onset, target, release, and offset of the gesture. The timepoints at peak velocity of both the formation and the release phase were also labelled. Labels were used here to calculate the duration of the formation and the release phases of each test constriction gesture. Formation duration was defined as the interval between the onset and release of the gesture, and release duration as the interval between the release and the offset of the gesture, as shown in Figure 1.

The moraic nasals /ɴ/ that precede a bilabial stop /b/ in the words such as /baɴ*bi/ and /meɴ*bo:/ are segmented as one single gesture (i.e., /ɴb/), as moraic nasal /ɴ/ in Japanese assimilates to the following stop in terms of place of articulation [23]. Consonants that are realized as geminates, such as the velar stops /g/ and /k/ in /bag*gu/ and /mak*kuɴ/, respectively, are also segmented as one single gesture. This is because both /ɴb/ and geminates appear as a single constriction.

The constriction gestures for each onset consonant of the disyllabic test words were analyzed. We refer to the onset consonants of the first and second syllable as C1 and C2 respectively. Separate linear mixed effects models with formation duration and release duration as response variables were fitted for each test C gesture using the lmerTest package [24] in R [25]. Random effects included random intercepts by speaker and word. The three fixed factors included were Boundary (phrase-medial vs. phrase-final, referred to as PhM and PhF respectively), First Syllable's Moraic Structure and Second Syllable's Moraic Structure (levels for two last factors: light [CV], heavy with a long vowel [CVV], heavy with a moraic coda [CVC]).

The drop1 function in R [25] was used to determine the optimal model. Random Effects Principle Components Analysis (rePCA) was used to determine the best random effect structure, and emmeans with Holm correction [26] were used to derive pairwise comparisons (α =0.05).

3. RESULTS

Results are reported for each test C gesture separately. C2 gesture is presented first since it is closer to the IP boundary, and thus more likely to be affected by it.

3.1. C2 formation duration

Statistical analyses detected an interaction effect between boundary and second syllable's moraic structure on the formation duration of C2 gesture ($F(2)= 20.279$, $p<0.0001$). Post-hoc pairwise comparisons clarified that boundary had a lengthening effect on the C gesture in the onset of the phrase-final syllable (C2 gesture) only when that syllable was light (CV) ($p<0.001$), and not when it was heavy either with a long vowel (CVV) or a moraic coda (CVC) (see Figure 2). This indicated that phrase-final lengthening can extend beyond the phrase-final mora or syllabic rhyme to the onset of the phrase-final syllable, but only when the syllable is light. As Figure 2 illustrates, this pattern might arise from a ceiling effect on C gestures in onsets of heavy syllables.

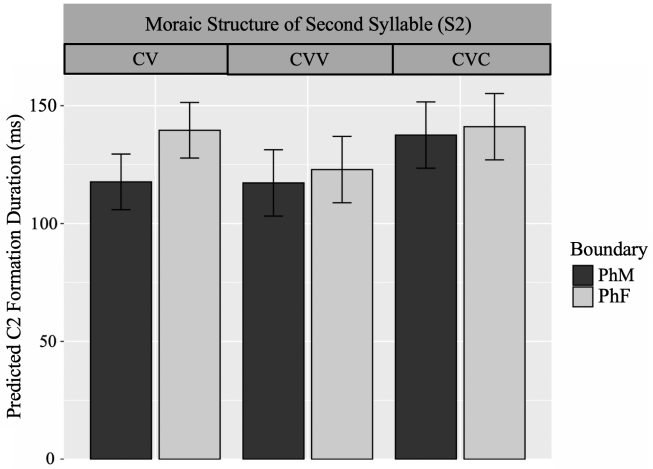


Figure 2: Predicted formation duration (ms; with standard error) of C2 gesture as a function of Boundary (PhM, PhF) by moraic structure (CV, CVV, CVC) of the second syllable' (S2).

3.2. C2 release duration

The release duration of C2 gesture showed an interesting three-way interaction of Boundary with the moraic structure of both the first and the second syllable ($F(4)= 6.2144$, $p<0.0001$).

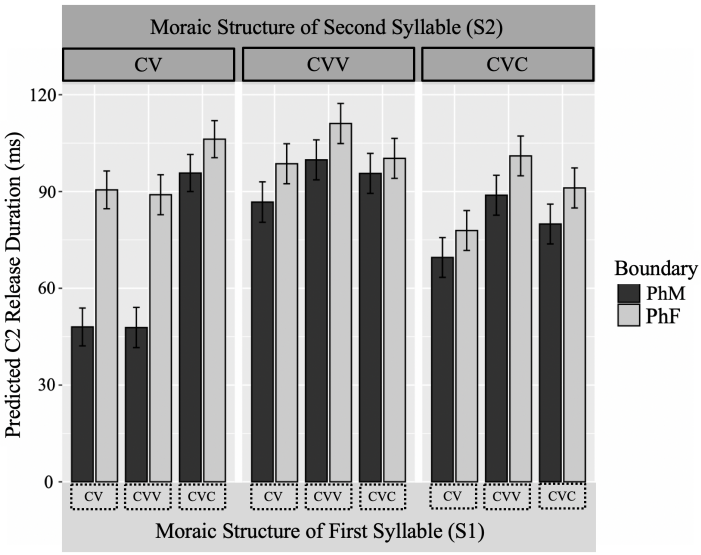


Figure 3: Predicted release duration (ms; with standard error) of C2 gesture as a function of boundary (PhM, PhF) by moraic structure (CV, CVV, CVC) of first (S1) and second (S2) syllable.

According to the pairwise comparisons, C2 release underwent phrase-final lengthening only in words with CV.CV or CVV.CV structures, in which the second syllable is light and the first syllable is not closed with a moraic coda (p<0.0001 for both comparisons). Thus, similarly to C2 formation, phrase-final lengthening extends from the right edge of the IP to C2 release only when the final syllable is light, with the following exception: this pattern does not hold in cases of CVC penultimate syllables. Once again, as Figure 3 indicates, these patterns seem to arise from a ceiling effect, since phrase-medial C gestures in CV.CV and CVV.CV are much shorter than their counterparts in the other moraic structures (p<0.0001 for all comparisons).

3.3. C1 formation duration

No effect of Boundary was detected on the formation duration of C1 gesture.

3.4. C1 release duration

A three-way interaction effect among the fixed factors of Boundary, First Syllable's Moraic Structure and Second Syllable's Moraic Structure was detected for C1 release (F(4)=2.4901, p=0.04). Pairwise comparisons showed that the source of this interaction was not the phrase-final lengthening effect. As illustrated in Figure 4, C1 release was not lengthened phrase-finally in any of the combinations of moraic structure for the two syllables of the phrase-final words.

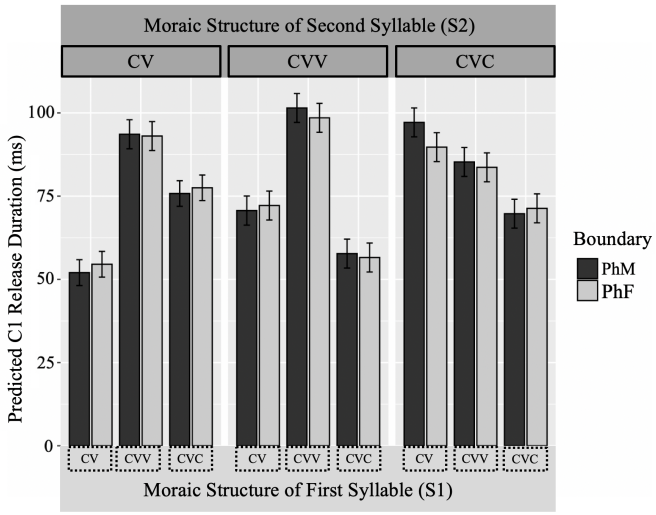


Figure 4: Predicted release duration (ms; with standard error) of C1 gesture as a function of boundary (PhM, PhF) by moraic structure (CV, CVV, CVC) of first (S1) and second (S2) syllable.

4. DISCUSSION

In order to examine the scope of boundary lengthening in Tokyo Japanese, this study investigated the durational profile of C gestures in the two syllabic onsets of initial-accented disyllabic words. No systematic boundary-related lengthening was found in the onset consonant of the penultimate syllable, suggesting that the phrase-final word is not the domain of the effect in this language, as was proposed for German in [16].

For the word-final syllable, boundary effects were found up to the onset of the phrase-final syllable, but only when that final syllable was light. Absence of phrase-final lengthening on C onsets of heavy syllables (CVV and CVC) is presumably due to a ceiling effect. These results in combination suggest that the domain of phrase-final lengthening in Tokyo Japanese is either the phrase-final mora, being in agreement with [20], or the phrase-final rhyme, suggested in e.g., [5]. Next steps of our analysis will assess acoustic durations of all C and V components of phrase-final syllables in an attempt to differentiate between these two alternative accounts.

We further propose that in order to better understand the presence of phrase-final lengthening on the onset of the word-final syllable, one should consider the fact that the test words used here are initially-accented. Lexical pitch accent has been found to affect the scope of phrase-final lengthening, such that words with non-final pitch accent present a larger scope of the effect than words with final pitch accent (e.g., [19], see also [21]). An account of these patterns could be provided from within the framework of Articulatory Phonology (e.g., [27]), in which prosodic effects at boundaries are instantiated by prosodic modulation gestures, called π-gestures, that control the spatio-temporal profile of the C and V constriction gestures that overlap with them [17]. These π-gestures have been proposed to have a dual coordination in stress languages [10]: one with the phrase-final vowel (V) gesture and one with the prosodic modulation gesture, called μ-gesture, that gives rise to stress [28]. Based on [19], Tokyo Japanese presents a similar coordination system for their π-gestures: these are coordinated both with the lexical pitch accent (when present) and the phrase-final mora/vowel gesture. When accent is initial, phrase-final lengthening is initiated earlier, extending towards the onset of the final syllable. In light syllables, lengthening actually reaches the onset C gesture, but in heavy syllables, which are longer due to their moraic structure, π-gesture does not get to overlap with the onset C gesture, and thus the latter does not get affected by it.

5. ACKNOWLEDGEMENTS

This work was supported by NSF Grant 1551428 and funds from the UCSB Academic Senate.

6. REFERENCES

[1] D. K. Oller, "The effect of position in utterance on speech segment duration in English," *J. Acoust. Soc. Am.*, vol. 54, no. 5, pp. 1235–1247, Nov. 1973, doi: 10.1121/1.1914393.

[2] R. Berkovits, "Durational Effects in Final Lengthening, Gapping, and Contrastive Stress," *Lang. Speech*, vol. 37, no. 3, pp. 237–250, Jul. 1994, doi: 10.1177/002383099403700302.

[3] T. Cambier-Langeveld, "The Domain of Final Lengthening in the Production of Dutch," *Linguist. Neth.*, vol. 14, no. 1, pp. 13–24, Jan. 1997, doi: 10.1075/avt.14.04cam.

[4] L. White, "English speech timing : a domain and locus approach," Doctoral dissertation, University of Edinburgh, 2002. Accessed: Jan. 06, 2023. [Online]. Available: http://hdl.handle.net/1842/23256

[5] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *J. Phon.*, vol. 35, no. 4, pp. 445–472, Oct. 2007, doi: 10.1016/j.wocn.2006.12.001.

[6] J. Edwards, M. E. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 369–382, Jan. 1991, doi: 10.1121/1.400674.

[7] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *J. Phon.*, vol. 26, no. 2, pp. 173–199, Apr. 1998, doi: 10.1006/jpho.1998.0071.

[8] D. Byrd, S. Lee, D. Riggs, and J. Adams, "Interacting effects of syllable and phrase position on consonant articulation," *J. Acoust. Soc. Am.*, vol. 118, no. 6, pp. 3860–3873, Dec. 2005, doi: 10.1121/1.2130950.

[9] D. Byrd, J. Krivokapić, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1589–1599, Sep. 2006, doi: 10.1121/1.2217135.

[10] A. Katsika, "The role of prominence in determining the scope of boundary-related lengthening in Greek," *J. Phon.*, vol. 55, pp. 149–181, Mar. 2016, doi: 10.1016/j.wocn.2015.12.003.

[11] S. Kim, J. Jang, and T. Cho, "Articulatory characteristics of preboundary lengthening in interaction with prominence on tri-syllabic words in American English," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. EL362–EL368, Oct. 2017, doi: 10.1121/1.5005132.

[12] J. Vaissière, "Language-Independent Prosodic Features," in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd, Eds. Berlin, Heidelberg: Springer, 1983, pp. 53–66. doi: 10.1007/978-3-642-69103-4_5.

[13] M. D. Tyler and A. Cutler, "Cross-language differences in cue use for speech segmentation," *J. Acoust. Soc. Am.*, vol. 126, no. 1, pp. 367–376, Jul. 2009, doi: 10.1121/1.3129127.

[14] L. H. Nakatani, K. D. O'Connor, and C. H. Aston, "Prosodic Aspects of American English Speech Rhythm," *Phonetica*, vol. 38, no. 1–3, pp. 84–105, Jan. 1981, doi: 10.1159/000260016.

[15] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1707–1717, Mar. 1992, doi: 10.1121/1.402450.

[16] K. J. Kohler, "Prosodic Boundary Signals in German," *Phonetica*, vol. 40, no. 2, pp. 89–134, Mar. 1983, doi: 10.1159/000261685.

[17] D. Byrd and E. Saltzman, "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening," *J. Phon.*, vol. 31, no. 2, pp. 149–180, Apr. 2003, doi: 10.1016/S0095-4470(02)00085-2.

[18] D. Byrd and D. Riggs, "Locality interactions with prominence in determining the scope of phrasal lengthening," *J. Int. Phon. Assoc.*, vol. 38, no. 2, pp. 187–202, Aug. 2008, doi: 10.1017/S0025100308003460.

[19] K. Tsai and A. Katsika, "Pitch accent and phrase boundaries: Kinematic evidence from Japanese," in *Speech Prosody 2020*, May 2020, pp. 215–219. doi: 10.21437/SpeechProsody.2020-44.

[20] M. A. Shepherd, "The scope and effects of preboundary prosodic lengthening in Japanese," *USC Work. Pap. Linguist.*, vol. 4, pp. 1–14, 2008.

[21] J. Seo, S. Kim, H. Kubozono, and T. Cho, "Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter?," *J. Acoust. Soc. Am.*, vol. 146, no. 3, pp. 1817–1823, Sep. 2019, doi: 10.1121/1.5122191.

[22] J. Seo, S. Kim, H. Kubozono, and T. Cho, "Interaction between rhythmic structure and preboundary lengthening in Japanese," in *PProceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019, pp. 637–641.

[23] T. J. Vance, *The Sounds of Japanese with Audio CD*. Cambridge University Press, 2008.

[24] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, Dec. 2017, doi: 10.18637/jss.v082.i13.

[25] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria., 2022. [Online]. Available: https://www.R-project.org/

[26] R. V. Lenth *et al.*, "emmeans: Estimated Marginal Means, aka Least-Squares Means." Dec. 06, 2022. Accessed: Jan. 07, 2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[27] Browman, C.P., Goldstein, L.M. 1992. Articulatory phonology: An overview. *Phonetica* 45, 155-180

[28] Saltzman, E., Nam, H., Krivokapić, J., Goldstein, L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proc. of the Speech Prosody Conference 2008*, 175-184.

FUNCTIONS OF AMFO'AN (METO) PROSODY: A PRELIMINARY ACOUSTIC STUDY

ABSTRACT

This paper presents a preliminary study of phrase-level prosody in Amfo'an, an understudied variety of Meto (Austronesian) spoken in Eastern Indonesia. It investigates the role of prosody in demarcating phrases and its possible role highlighting important elements within the phrase. It also examines the possible interaction of prosody with word stress, for which acoustic evidence has previously been demonstrated. This is done by acoustic analysis of F0 movements in the phrase. Results indicate that prosody primarily serves a demarcating function in Amfo'an, and stress was not found have a significant effect on F0 movements. Based on these findings, Amfo'an may be considered an example of a language where 'intonation ignores stress'.

Keywords: Austronesian, Indonesia, prosody, F0, word stress

1. INTRODUCTION

Two primary functions of phrase prosody are typically distinguished: demarcating boundaries between phrases and marking highlighting important elements within phrases [1].

This paper presents a preliminary study of the functions of prosody in Amfo'an, which is a variety of Meto (Austronesian) spoken in Eastern Indonesia. This is done by acoustic analysis of of F0 movements, especially in the final two syllable of the phrase. This focus on final two syllables of the phrase is because in Austronesian languages of Eastern Indonesia, the main prosodic events typically occur at the right phrase boundary [2].

F0 is investigated because it is typically considered to be the important cue for phrase prosody cross-linguistically (e.g. [3]). Following the methodology of [1], the form the shape of F0 movements in different phrase positions is investigated, as well as the role of prosody in demarcating phrases and its possible role highlighting important elements within the phrase.

In this study, I also investigate the possible interplay of F0 movements and word stress. Amfo'an has previously been demonstrated to have acoustic evidence for stress on the penultimate syllable [4]. The primary acoustic correlates of stress in Amfo'an are spectral tilt, vowel quality and the ratio of perceptible pitch rises and falls, with penultimate (stressed) syllables having a higher ratio of rises to falls than ultimate (unstressed) syllables.

In languages with word stress, highlighted words are often marked by F0 movements on or near a syllable which bears word stress, often referred to as *pitch accents* [5]. Models of intonation often assume that this will be the case (e.g. [3]). There are, however, counter-examples such as Kuot [6], where there is no evidence of interaction of F0 and word stress.

2. METHOD

2.1. Data collection

Data for this study comes from spontaneous narrative collected by the author in 2017 in Lelogama, West Timor. Recordings were made using a zoom H1 recorder.

2.2. Participants

All participants are native speakers of Amfo'an, ranging in age from 17-71 at the time of recording. There were four male and three female participants.

2.3. Data selection and preparation

The data was transcribed by a native speaker of Meto, and segmented in into intonation units for transcription with the aid of the transcriber. All words and syllables were then segmented by the author using Praat textgrids.

A subset of the syllables (N = 2834) were selected for analysis. Syllables in utterances that were interrupted (especially by background noise) or cut off were omitted. Words containing sequences of two vowels were omitted because they can be realised as either one or two syllables (either as a long vowel, phonetic diphthong or sequence of two distinct vowels).

The position of each syllable in a phrase was labeled according to a five- way distinction (first, second, medial, pre-final or final). Note that for all statistical tests involving the five-way position distinction (see §2.5), this was performed on syllables taken from phrases that were minimum syllables in length. This was to avoids ambiguous

labels for phrases less than five syllables.

Syllables were also labeled for word class; either content (nouns, verbs, numbers, adverbs) or function words (pronouns, demonstratives, conjunctions). Syllables were also labeled for stress, based in on the findings of [4]. Syllables were also annotated to indicate whether they occurred before a silent pause. Two kinds of pauses were distinguished; within-phrase pauses and between-phrase pauses.

2.4. Measurements

F0 measures were taken using Praat [7] using scripts written by Constantijn Kaland, available at https://osf.io/cr4vn. For each syllable, the F0 minimum and maximum in semitones (henceforth ST) and their timestamps were extracted. Then, for each syllable the size of the F0 movement was calculated by subtracting the F0 minimum from the F0 maximum. F0 movements were also labeled as either rises, falls or rise-falls. Movements where the minimum occurs before maximum were labeled as rises, and movements where the maximum occurs before minimum were labeled as falls. Pitch movements on either side of the max were labeled as rise-falls. Movements were only labeled if they were above the threshold $_{gthr}$(ST/s) = .16/T2 which is a measure of the speed of frequency change in semitones per seconds for a given F0 movement duration, T. This threshold is indicative of whether pitch movements are likely to be perceivable [8, 32].

2.5. Statistical analysis

Statistical analyses were conducted using R [9]. The *lme4* and *afex* packages [10, 11] were used to carry out Linear mixed models fit by maximum likelihood and chi-square tests. The *multcomp* package [12] was used to carry out pairwise comparisons. Predictors are referred to in this paper using SMALL CAPS.

2.5.1. F0 movement shape

The shape analysis examined the sizes and directions of the F0 movements, especially in the pre-final and final positions of the phrase. It also investigated if F0 movements were affected by word stress. To do this, a LMM was carried out with F0 MOVEMENT as the response and PHRASE POSITION (first, second, medial, pre-final, final) and STRESS (stressed, unstressed) as predictors. Participants and words were random intercepts. Pairwise comparisons Tukey honestly significant difference (HSD) test (Bonferroni corrected) were conducted between PHRASE POSITIONS. Chi-square tests were done on the distribution of rises and falls in the two final syllables in the phrase, testing for the effect of position and stress.

2.5.2. Demarcating function

The demarcating analysis investigated the relationship between F0 movements and pauses in order to determine the extent to which F0 marks prosodic boundaries in Amfo'an,. In particular, it was tested if larger F0 movements signal stronger phrase breaks, as has been previously observed, [13]). This was done by examining the size of F0 movements before within-phrase and between-phrase pauses.

This was done by conducting a LMM on the final two syllables before a pause. F0 MOVEMENT was the response and PAUSE TYPE (within, between) and PHRASE POSITION relative to the pause (pre-final, final) were predictors. Participants and words were random intercepts.

2.5.3. Highlighting function

The possible highlighting function was investigated by examining the possible effect of word class on F0 movements. This is because content words have been predicted to occupy more important positions in prosodic structure than function words [1, 14]. Possible effects of word stress were also examined. Given the acoustic evidence for stress on the penultimate syllable in in Amfo'an, we would expect that the pre-final syllable of the phrase is a prosodically prominent position in Amfo'an.

This was done in three statistical tests. The first was a chi-square test on the distribution of perceptible rises and falls in function and content words. Then, two LMMs were conducted with F0 MOVEMENT as the response with WORD CLASS (function, content), STRESS (unstressed, stressed) and PHRASE POSITION as predictors. In the first LMM, the PHRASE POSITION predictor compared pre-final vs other syllables, and in the second final vs other syllables.

3. RESULTS

3.1. F0 movement shape

Final syllables show larger F0 movements than other positions in the phrase. Significant comparisons were only found for the final position in the phrase (Table 7).

The LMM found no significant affect of stress on

Table 1: Mean F0 movements in ST (SD) in different phrase positions

first	second	medial	pre-final	final
3.74	4.54	3.5	4.78	6.36
(4.78)	(6.52)	(5.44)	(6.93)	(7.02)

Table 2: Pairwise comparisons for different phrase positions

Pairwise comp	β	SE	z	*p*
First-Second	0.86	0.73	1.18	n.s.
First-medial	0.01	0.6	0.01	n.s.
First-pre-final	0.85	0.76	1.11	n.s.
First-final	3.12	0.82	3.80	**=0.001**
Second-medial	-0.85	0.59	-1.43	n.s.
Second-pre-final	-0.01	0.76	-0.01	n.s
Second-final	2.26	0.81	2.81	0.04
Medial-pre-final	0.84	0.61	1.36	n.s.
Medial-final	-3.11	0.66	-4.72	**<0.001**
Pre-final-final	-2.27	0.81	-2.81	**=0.04**

F0 movement in any phrase position (Table 3).

Table 3: LMM results: shape

Predictor	β	SE	t	*p*
Stress	-0.8	0.42	-1.92	n.s.

The distribution of rises, falls and rise falls did not differ significantly between stressed and unstressed syllables (Table 4). Regarding the distribution of rises and falls in the final two syllables in the phrase, the most common pattern is a fall in the pre-final syllable followed by either a rise or fall in the final syllable. A significant affect for position was found (Table 4).

Table 4: Shape analysis, chi-square test results: Perceptible F0 movement

	Rise	Fall	Risefall	χ^2	*p*
Pre-final	130	208	38	204	**<.001**
Final	216	204	74		
Stressed	149	174	63	5.19	n.s.
Unstressed	195	238	48		

3.2. Demarcating

The LMM indicated that PAUSE TYPE significantly affected the size of F0 movements significantly (Table 5) F0 movements were smaller in pre-final (*M*=2.89, *SD*=2.21) and final (*M*=5.28, *SD*=7.51) syllables before within-phrase pauses than they were in pre-final (*M*=4.7, *SD*=6.67) and final (*M*=7.57, *SD*=8.29) syllables before between-phrase pauses occurring between phrases. The LMM also showed a significant effect of PHRASE POSITION (pre-final, final, Table 5). Pre-final syllables typically have smaller F0 movements (*M*=4.53, *SD*=6.41) than final syllables (*M*=7.29, *SD*=8.23).

Table 5: LMM results: demarcating

Predictor	β	SE	t	*p*
Pause type	-2.58	0.69	-3.71	**<0.001**
Phrase position	-2.83	0.42	-6.8	**<0.001**

3.3. Highlighting

There is no evidence that the distribution of rises and falls differs significantly between content and function words (Table 6).

Table 6: Chi-square test results: Highlighting

	Rise	Fall	Risefall	χ^2	*p*
Content	458	678	176	12.49	n.s.
Function	120	237	33		

Regarding the the size of F0 movements, the LMMs found no significant of WORD CLASS. The effect of PHRASE POSITION was only significant for final syllables, with final syllables having larger F0 movements (*M*=7.48, *SD*=8.23) than syllables in other positions (*M*=4.23, *SD*=6.24). Again, stress did not have a significant affect on the size of F0 movements.

Table 7: LMM results: Highlighting

Predictor	β	SE	t	*p*
Word class	-0.14	0.8	-0.19	n.s.
Phrase position (final vs other)	-3.38	0.94	-3.59	**<0.001**
Stress	-0.57	0.54	-1.05	n.s.
Word class	-0.3	0.83	-0.36	n.s.
Phrase position (pre-final vs other)	0.22	0.8	0.28	n.s.
Stress	0.17	0.52	0.32	n.s.

4. DISCUSSION

The results of the current study indicate that the largest F0 movements in Amfo'an occur in the final syllables of the phrase, indicating that the main

prosodic events in this language occur at the right phrase boundary. In particular, the final syllables differ significantly from other syllables in the size of the F0 movement. This suggests that phrase boundaries in Amfo'an are mainly marked by pitch movements on phrase-final syllables.

F0 movements are larger before between-phrase pauses than before within- phrase pauses. This indicates that stronger phrase breaks are marked by larger F0 movements, as expected [13].

I find no significant effect of word class on F0 movement or on the distribution of rises and falls. I also do not find significant effect of stress on F0 movements, which is not surprising as F0 has not been found to be a correlate of stress in Amfo'an. I find that word stress has a significant affect on the distribution of rises and falls in pre-final and final syllables of the phrase. This is notable because the distribution of rises and falls has been found to be a correlate of stress at the word level [4]. These results provide evidence that stressed syllables do not act as F0 anchoring points in Amfo'an (i.e. pitch accents). Based on these results, Amfo'an may be described as a language where 'intonation ignores stress'.

Based on these findings, it is likely that prosody in Amfo'an does not have a highlighting role, but only a demarcating one. This paper represents the first in-depth study of prosody in an Austronesian language of Timor, and contributes to our understanding prosody in Austronesian languages more broadly. These results are, however preliminary and there are many avenues for further research.

5. REFERENCES

[1] C. Kaland and S. Baumann, "Demarcating and highlighting in Papuan Malay phrase prosody," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2974–2988, 2020.

[2] N. P. Himmelmann and D. Kaufman, "Austronesia," in *The Oxford Handbook of Language Prosody*, C. Gussenhhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 370–383.

[3] D. R. Ladd, *Intonational phonology*, 2nd ed. New York: Cambridge University Press, 2008.

[4] K. Culhane, "Examining acoustic evidence for stress in Amfo'an," Paper presented at the 14th International Austronesian and Papuan Languages and Linguistics Conference (APLL14), Leibniz Centre General Linguistics (ZAS) and Humboldt University, Berlin, Germany, 2022.

[5] T. Roettger and M. Gordon, "Methodological issues in the study of word stress correlates," *Linguistics Vanguard*, vol. 3, no. 1, 2017.

[6] E. Lindström and B. Remijsen, "Aspects of the prosody of Kuot, a language where intonation ignores stress," *Linguistics*, vol. 43, no. 4, pp. 839–870, 2005.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer. version 6.1.42," 2021. [Online]. Available: praat.org

[8] J. 't Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. New York: Cambridge University Press, 1990.

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[10] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[11] H. Singmann, B. Bolker, J. Westfall, F. Aust, M. S. Ben-Shachar, S. Højsgaard, J. Fox, M. A. Lawrence, U. Mertens, J. Love, R. Lenth, and R. H. B. Christensen, *afex: Analysis of Factorial Experiments*, 2021, r package version 1.0-1. [Online]. Available: https://CRAN.R-project.org/package=afex

[12] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008.

[13] J. Choi, "Pause length and speech rate as durational cues for prosody markers," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2395–2395, 2003.

[14] S.-A. Jun, Ed., *Prosodic typology II: The phonology of intonation and phrasing*. Oxford: Oxford University Press, 2014.

THE PERCEPTION OF SURPRISE QUESTIONS IN ESTONIAN

Pärtel Lippus[1], Eva Liina Asu[1], Katrin Leppik[1], Heete Sahkai[2]

[1]Institute of Estonian and General Linguistics, University of Tartu, Estonia

[2]Institute of the Estonian Language, Tallinn, Estonia

partel.lippus@ut.ee, eva.liina.asu-garcia@ut.ee, katrin.leppik@ut.ee, Heete.Sahkai@eki.ee

ABSTRACT

The study examines the perception of surprise questions as compared to information-seeking questions. An online perception experiment was carried out in the PCIbex environment using natural stimuli from 6 different speakers from an earlier production experiment. The task of the participants was to evaluate whether the speaker intended to ask for information or to express surprise. The overall correct identification score for information-seeking questions was 84% and for surprise questions 55%, but there was large variability between speakers and items. Higher perception rate of surprise questions is above all related with a longer duration, a wider pitch range, and a lower mean pitch of the utterance.

Keywords: surprise, non-canonical questions, information-seeking questions, prosody, perception, Estonian

1. INTRODUCTION

This study examines the perception of Estonian surprise questions (SQs) as compared to information-seeking questions (ISQs). The SQs examined in the study are syntactically and lexically canonical wh-interrogatives that have been uttered by the speaker to express surprise. They are compared to string-identical interrogatives that have been uttered in order to request information. SQs differ from ISQs in terms of speech act, as they are not (canonical) questions, as well as in terms of added emotional expressivity, as they express surprise.

Prosodic expression of speech acts is usually associated with the use of phonological pitch accents and edge tones while the expression of emotions has been associated with the gradient use of phonetic realisation, see e.g. [1]. A previous production study [2] showed, however, that Estonian SQs differ from string-identical ISQs by their longer duration, lower pitch, wider pitch range, and a more frequent use of creaky voice quality, but not in terms of pitch accent types or boundary tones. A similar production study examining string-identical SQs and ISQs in French [3] identified both phonological and phonetic differences between the two categories: ISQs were shorter and produced with a faster tempo, and contained a greater proportion of rising final contours. A following perception study [4] confirmed that prosodic cues were sufficient for distinguishing the two conditions without any context provided, but the correct perception rate of ISQs was significantly higher than that of SQs (82% vs. 69%).

The present study addresses two research questions: (1) Which acoustic features correlate most with the perception of surprise? (2) Does the context of surprise influence the recognition rate of SQs? The broader aim of the study is to investigate which factors contribute to the perception of surprise and speech acts.

2. MATERIALS AND METHOD

2.1. Stimuli

The stimuli of the perception test were recorded as part of a production study [2] and included 24 interrogative sentences: 12 with the question word *mis* 'what' (e.g. *Mis loom see on?* 'What animal is this?') and 12 with the question word *mida*, partitive case of *mis* 'what' (e.g. *Mida sa teed?* 'What are you doing?'). Two contexts were created for each sentence in order to elicit an information-seeking reading and a surprise reading of the sentence. The contexts for eliciting ISQs prompted the participant to imagine a situation that would require asking for information from a knowledgeable addressee. The contexts for eliciting SQs, on the other hand, prompted the participant to act surprised at an unexpected situation (surprise caused by the unprepared mind of the speaker) or a counterexpectational situation (surprise caused by a contrary expectation of the speaker). A subset of the contexts included incongruity (untypicality of the

object of surprise) as an additional cause of surprise, and contexts of another subset were intended to elicit disapproval in addition to surprise. The purpose of the different context types was to test the effect of the context of surprise on the expression and perception of surprise (for a full list of test sentences and elicitation contexts see https://osf.io/b4fek).

The design of the perception test followed largely that of a similar test by [4]. The stimuli consisted of 240 unique utterances produced by six randomly chosen female participants of the production study. An equal number of utterances was randomly chosen from each speaker (40 per speaker). The stimuli were distributed between four lists of 72 so that each list contained 36 ISQs and 36 SQs. Twelve interrogative sentences in each list were uttered once as a SQ and twice as an ISQ by different speakers, and 12 sentences were uttered once as an ISQ and twice as a SQ. There were eight stimuli from each speaker (in total 48 stimuli) that appeared in two lists but no repetitions by the same speaker of the same sentence were included in one list.

2.2. Perception test

The perception test was carried out online in the PCIbex environment [5]. 52 native speakers of Estonian (44 female, 7 male, 1 non-binary; 24 to 69 years of age, mean age 42) participated in the test. All but one participant had self-reported normal hearing but as her responses did not differ from the rest of the group they were not excluded. Two participants reported no knowledge of languages other than Estonian (on B2 level or above); 26 reported the knowledge of one, 13 of two, and 10 of more than two other languages.

During the test, upon hearing each stimulus the participants had to decide whether the speaker's intention was to ask for information or express surprise. The test was preceded by a short practice session consisting of four stimuli where the intention of the speaker was indicated (two requests for information and two expressions of surprise).

Due to technical reasons the lists did not receive an equal number of listeners. The total number of listeners per each list was as follows: A 12, B 14, C 11, D 15. Considering the design of the experiment each stimulus received, thus, at least 11 responses and some stimuli up to 29 responses.

2.3. Analysis

The results of the perception test were analysed in *R* [6, 7]. Logistic regressions from the *lme4* package [8] were used to assess the effects of the context type and the acoustic factors on the perception of surprise. The effects of the acoustic features on the perception test ratings were further evaluated with a Random Forest model using the package *randomForest* [9].

3. RESULTS

The overall average results are shown in Figure 1. The information-seeking questions were recognised (correctly) as questions asking for information in 84% of the cases but the surprise questions were recognised as expressing surprise only in 55% of the cases, which is just above the chance level. In the following subsections, the results are presented in relation to the effect of the context type and the acoustic features of the surprise condition stimuli.

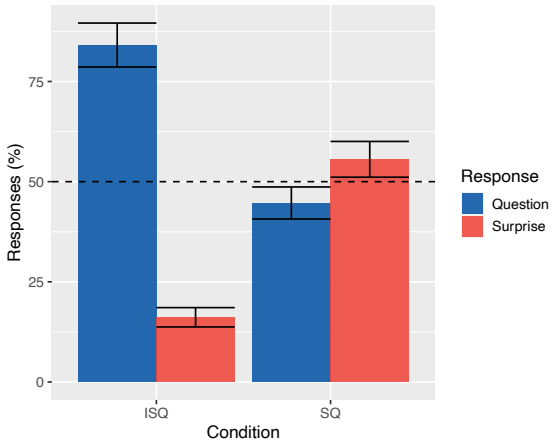


Figure 1: Recognition of ISQs and SQs as asking for information (blue) or expressing surprise (red).

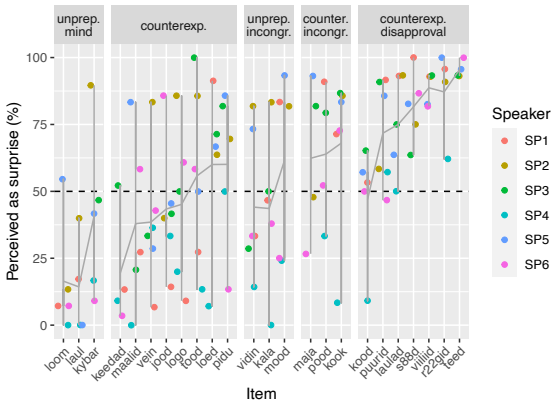


Figure 2: The frequency of surprise condition items perceived as expressing surprise grouped by the context type.

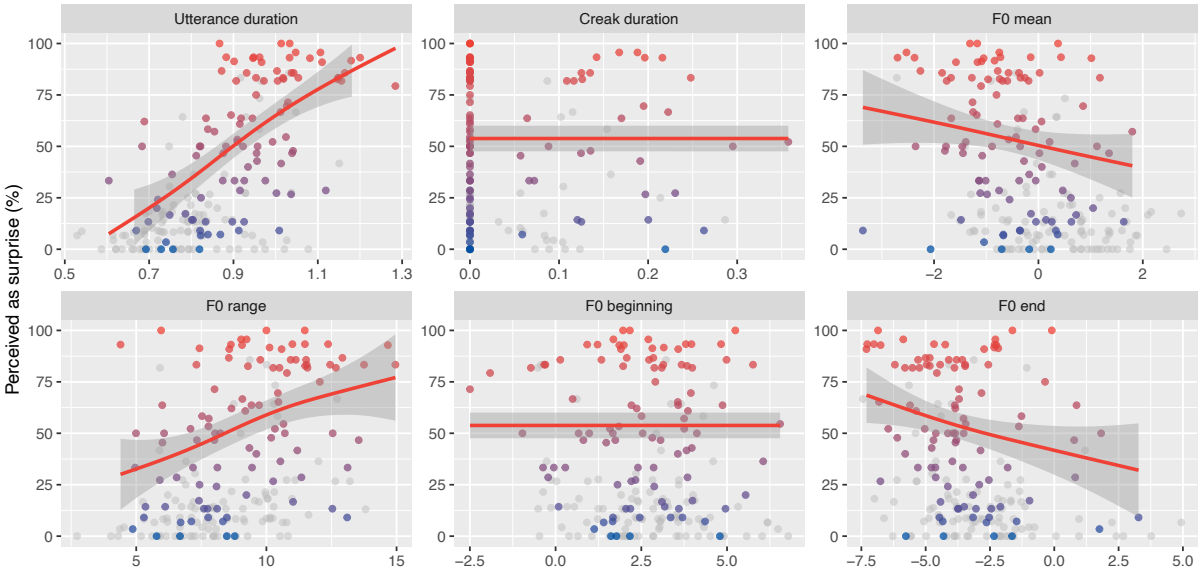


Figure 3: Perception of surprise as a function of acoustic features of the surprise condition stimuli (ranging from blue to red according to the rate of perceived surprise; duration in seconds, pitch in semitones re speaker's mean). The question condition stimuli are shown in grey in the background.

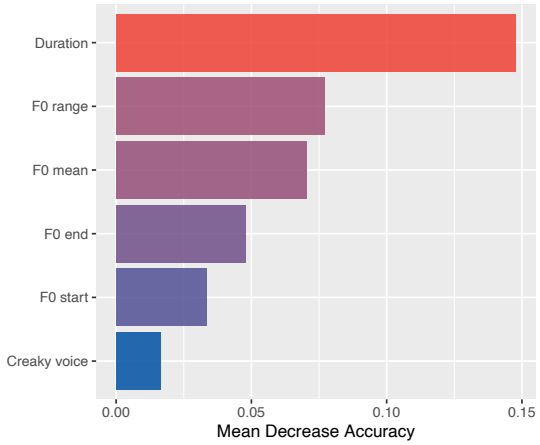


Figure 4: Mean decrease of accuracy of the acoustic factors in the random forest model.

3.1. Context of surprise

There was a large speaker- and item-dependent variation in the perception of SQs. Figure 2 shows the frequency of surprise condition items perceived as expressing surprise grouped by the context type. It can be seen that some stimuli were rarely identified as expressing surprise (e.g. items *loom*, *laul*, *keedad*) while there were stimuli that were in most of the cases recognised as expressing surprise (e.g. items *teed*, *viilid*). It can also been seen that the stimuli by some speakers were less well recognised as expressing surprise, e.g. in particular those of SP4.

Figure 2 also shows that the recognition of surprise is affected by the context type that was coded in the context eliciting the utterances used for the stimuli. The perception of surprise was the lowest in the case of *unprepared mind* (28%), followed by *counterexpectation* (45%), *unprepared mind combined with incongruity* (47%), and *counterexpectation combined with incongruity* (66%). The perception of surprise was the highest and significantly above chance level in the case of *counterexpectation combined with disapproval* (75%).

A post-hoc test of a logistic regression model (with stimulus and participant as random intercepts) showed that the difference was significant only between the two categories with lower recognition rate (*unprepared mind* and *counterexpectation*) on the one hand and the category with the highest recognition rate (*counterexpectation combined with disapproval*) on the other hand ($p < 0.001$).

3.2. Acoustic features

Figure 3 shows the perception of surprise questions as a function of different acoustic features of the stimuli such as utterance duration, F0 range, F0 mean, utterance-initial and utterance-final F0 and the duration of creaky voice quality. There is a

strong positive correlation between the perception rate of surprise and the duration of the utterance as well as the F0 range of the stimuli, and a negative correlation between the perception rate of surprise and the F0 mean and the utterance-final F0. There was no correlation between the perception rate and utterance-initial F0 or the duration of creaky voice. A logistic model showed only a significant effect of utterance duration (β = 8.34, z = 7.85, $p <$ 0.001) and F0 range (β = 0.09, z = 2.03, p = 0.042) indicating a strong intercorrelation between the different measures of F0.

The effect of the acoustic features on predicting the perception of surprise was tested with a random forest model. The data was split into two subsets: 80% of the data was used for training the model and 20% for testing. The model was trained with the six acoustic factors with 500 trees and 2 variables at each split. The OOB estimation rate of the model was 19.5%. The variable importance is presented in Figure 4. The most important factor is the total duration of the utterance, which is followed by the F0 range and F0 mean. Utterance-final F0 had a stronger effect than utterance-initial F0. The duration of creaky voice was the weakest of the six factors. On the testing sample the prediction accuracy of the model was 81% (error rate 15% for question and 28% for surprise).

4. DISCUSSION

This study investigated the perception of information-seeking (ISQs) and surprise questions (SQs) in Estonian using natural stimuli produced by six speakers. ISQs were more reliably recognised than SQs (at 84%), and nearly half of the SQs were perceived as ISQs. A similar outcome (albeit a higher recognition rate of SQs) was obtained for French by [4], who suggest that the interpretation of the stimuli as questions may have been supported by their interrogative syntax and absence of context.

By far the strongest acoustic cue correlating with the perception of surprise was the longer duration of the utterance. Longer duration was also identified as the main phonetic correlate of French SQs [3] and has been found to correlate with non-canonical questions more generally, in particular rhetorical questions [10, 11] and exclamatory questions [12, 13, 14, 15]. The next strongest correlates of perceived surprise were a wider F0 range and lower mean F0. A wider F0 range also distinguishes Estonian exclamations from questions [12] and has been associated with the expression of surprise [16, 17] as well as other emotions [18, 19]. The lower mean pitch of Estonian SQs as compared to ISQs signals a non-questioning speech act [2] and is characteristic of Estonian statements [20, 21, 22], rhetorical questions [10] and exclamations [12] as compared to questions.

Yet another characteristic acoustic feature of SQs that was identified in the production study [2] - a longer duration of creaky voice - did not correlate with the perception of surprise. This can be explained by the fact that there is a lot of individual variation and the present study used the stimuli from only six randomly chosen speakers.

Regarding the effect of different possible causes of surprise (the unprepared mind of the speaker, a contrary expectation by the speaker, and the untypicality of the object of surprise), and the combination of surprise with another emotion/attitude (disapproval), the results suggest that surprise alone is perceptually relatively similar to a request for information. This is especially obvious when caused only by the unprepared mind or counterexpectation on the part of the speaker, while surprise in combination with disapproval is perceptually distinct from a request for information. Such a result might imply that when the expression of surprise is accompanied by the expression of a judgment (for example, that a situation is untypical or disapprovable), the resulting speech act is more clearly perceptually non-questioning. In future work, we intend to examine in more detail the correlation of the acoustic cues of SQs with the different contexts of surprise.

5. CONCLUSION

It can be concluded that the acoustic feature that is most robustly related to the perception of surprise questions in Estonian is the longer duration of the utterance. Additionally, such pitch-related features as a wider F0 range and a lower mean F0 contribute to the accurate perception of surprise questions. Regarding the effect of the context of surprise, surprise questions seem to be perceptually more distinct from information-seeking questions when they also express a judgment.

6. ACKNOWLEDGEMENTS

We are very grateful to Maud Pélissier for her valuable advice on experimental design. The study was supported by the projects EKKD10 and EKKD114 (the Estonian Ministry of Education and Research) and the basic governmental financing of the Institute of the Estonian Language by the Estonian Ministry of Education and Research.

7. REFERENCES

[1] P. Prieto, "Intonational meaning: Intonational meaning," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 4, pp. 371–381, Jul. 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/wcs.1352

[2] E. L. Asu, H. Sahkai, and P. Lippus, "The prosody of surprise questions in Estonian," *Journal of Linguistics*, pp. 1–21, Mar. 2023. [Online]. Available: https://doi.org/10.1017/S0022226723000014

[3] A. Celle and M. Pélissier, "Surprise questions in spoken French," *Linguistics Vanguard*, vol. 8, no. s2, pp. 287–302, Jan. 2022. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/lingvan-2020-0109/html

[4] ——, "C'est quoi and qu'est-ce que as surprise questions: a perception study," in *Presentation given at the PHC Parrot final workshop*, 2022-10-28.

[5] J. Zehr and F. Schwarz, "PennController for Internet Based Experiments (IBEX)," Aug. 2022, publisher: Open Science Framework. [Online]. Available: https://osf.io/md832/

[6] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[7] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[8] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[9] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[10] H. Sahkai, E. L. Asu, and P. Lippus, "Prosodic characteristics of canonical and non-canonical questions in Estonian," in *Proc. Speech Prosody 2022*, 2022, pp. 135–139. [Online]. Available: https://www.doi.org/10.21437/SpeechProsody.2022-28

[11] B. Braun, N. Dehé, J. Neitsch, D. Wochner, and K. Zahner, "The prosody of rhetorical and information-seeking questions in German," *Language and Speech*, vol. 62, no. 4, pp. 779–807, Dec. 2019. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0023830918816351

[12] H. Sahkai, E. L. Asu, and P. Lippus, "Prosodic characteristics of exclamatives and questions in Estonian," in *Proc. 1st International Conference on Tone and Intonation (TAI)*, 2021, pp. 41–45. [Online]. Available: https://www.doi.org/10.21437/TAI.2021-9

[13] S. Repp, "The Prosody of *Wh* -exclamatives and *Wh* -questions in German: Speech Act Differences, Information Structure, and Sex of Speaker," *Language and Speech*, vol. 63, no. 2, pp. 306–361, Jun. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0023830919846147

[14] S. Repp and H. Seeliger, "Prosodic Prominence in Polar Questions and Exclamatives," *Frontiers in Communication*, vol. 5, p. 53, Aug. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomm.2020.00053/full

[15] J. Rett and B. Sturman, "Prosodically marked mirativity," in *Proceedings of the 37th West Coast Conference on Formal Linguistics*, D. K. E. Reisinger and M. Huijsmans, Eds. Somerville, MA, USA: Cascadilla Proceedings Project, 2021, publisher: The University of British Columbia Version Number: 1. [Online]. Available: http://www.lingref.com/cpp/wccfl/37/paper3511.pdf

[16] C. Gunlogson, *True to form: rising and falling declaratives as questions in English*, ser. Outstanding dissertations in linguistics. New York: Routledge, 2003.

[17] C. Lai, "Perceiving surprise on cue words: prosody and semantics interact on right and really," in *Proc. Interspeech 2009*, 2009, pp. 1963–1966.

[18] S. Repp and L. Rosin, "The intonation of echo wh-questions," in *Proc. Interspeech 2015*, 2015, pp. 938–942.

[19] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, no. 3-4, pp. 252–267, Jul. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639305000890

[20] E. L. Asu, "Downtrends in different types of questions in Estonian," in *Proceedings of the Speech Prosody 2002 conference*, B. Bel and I. Marlien, Eds. Aix-en-Provence: Laboratoire Parole et Langage, 2002, pp. 143–146.

[21] ——, "The phonetics and phonology of Estonian intonation," Ph.D. thesis, University of Cambridge, 2004.

[22] K. Vende, "Phonetic conditioning factors of pitch in Estonian vowels," in *Estonian Papers in Phonetics*, Tallinn, 1973, pp. 46–85.

Oral reading proficiency and prosody – a perceptual pilot study on especially fluent German students (grade 3 to 7)

Christopher Sappok

Institute of German Language and Literature II, University of Cologne, Germany
csappok@uni-koeln.de

ABSTRACT

Phonetic research on children's read speech is rare compared to educational research. In this pilot study, the established educational NAEP-scale [1] for perceptual ratings of reading fluency is juxtaposed to several phonetic, mainly prosodic, parameters. From a greater corpus we focus on recordings in which performance was identified as maximally fluent on the NAEP scale by university student raters with maximal reliability, arriving at a subsample of 50 recordings of 15 subjects reading two texts. We found a considerable amount of systematic variance in a number of our partly newly introduced perception- and measure-based variables, which are thus relevant for quantifying the developmental stage of *oral reading proficiency*, establishing a more advanced notion of fluency. (Note that all variance found is evidence for a ceiling effect of the NAEP fluency scale.) The findings pave the way for a phonetically sound target modelling of oral reading proficiency development.

Keywords: oral reading, read speech, fluency, prosody, German children

1. INTRODUCTION

In our context, several lines of research need to be distinguished, namely phonetic research on natural speech and on read speech and educational research on reading acquisition. In phonetics, one often has to rely on written stimuli in order to control for content, even when the eventual objective is maximally natural speech. Read speech as an objective in its own right is underrepresented. Earlier examples are [2] on Swedish and [3] on Dutch grown up readers. An early example on children's read speech is [4] (American English), who see prosody as a *space* and read speech as a *variety*: "The target is in effect to define a prosodic space, with dimensions corresponding to the relevant variables, and different varieties occupying different regions of the space" (other varieties mentioned in [4]: expression of emotion, functional distinctions, social varieties, and clinical conditions). In order to arrive at such an (intersubjective or even interlingual) space we need to understand intrasubjective relations between read and spontaneous speech as well as within subject reading skill development. Our approach focuses on development. An approach focusing read vs. spontaneous speech with German children is [5], who points out the problem of the target model relevant with all efforts to evaluate developmental stage: "child language researchers interested in phonetic and phonological acquisition need to be very careful when collecting adult speech data to determine the target model" [5]. This issue leads to educational research, based on the term *fluency*. Two recent meta analytic studies [6, 7] give an overview, together covering 43 studies from 1995 to 2019. Even though both meta studies have the term *prosody* in their titles, a majority of the included studies does not fully follow phonetic standards on the matter, e. g., in [8] a pitch target model for evaluating early elementary school children is coarsely calculated from measures in adult readings of the same text; in [9] the target model for reading acquisition is coarsely equated with spontaneous speech. Further, most studies are limited to elementary school contexts, leaving blank the "teenage" period between elementary and adult oral reading. On the other hand, this line provides a definition based, well established rating procedure. An agreed upon fluency definition is subsumed in [9], naming the following dimensions: *accuracy*, *automaticity*, and *oral reading prosody*. The underlying concepts are mirrored in the following instruction for listeners to detect fully developed fluency used by the National Center for Education Statistics (NAEP) of the United States: "Reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation" [1]. With its mixture of relevant but very different facets this formulation represents level 4 on the 4-level-NAEP fluency scale. In our recent work we found strong hints that this scale produces a pronounced ceiling effect, when applied (in translated form) to longitudinal audio stimuli from grade 3 to 7 [10, 11]. To quantify the extent of this ceiling effect, the present study investigates mainly recordings identified as NAEP-level 4, and uses alternative hypothetically relevant parameters to disentangle the very different features mentioned "in one go" in the level-description above. We

will present arguments for the parameters' relevance as well as evidence from descriptive and correlational analysis. Our main research question is concerned with perception: Which rating scales reliably capture oral reading proficiency? Only when this is sufficiently clear we are ready to fully tackle the next question: Which relations may be established between the according perceptual parameters and parameters that can be measured in the signal? [12]

2. SAMPLE

2.1. The DOL-recordings of the LAUDIO-corpus

The LAUDIO-corpus (= Longitudinal AUDIO) is a collection of audio files from German pupils reading aloud. The underlying subject sample consists of 4 German elementary school classrooms (2-times 3rd grade, 2-times 4th grade, N_{LAUDIO} = 61), all of the same "Inklusive Grundschule" (desired standard in Germany), in a rural part of the Federal State of Rheinland-Pfalz. The design was to record all children and then again once a year, 4 recording rounds altogether, the biographical windows covered being grade 3 to 6 for the original 3rd graders and grade 4 to 7 for the original 4th graders. Rounds were held in winter, the first being winter 2014/2015, the last 2017/2018. (Note that thus these windows are almost the very last windows that were totally unaffected by the Covid pandemic.) In each round, children were recorded in sessions of roughly 15 minutes on being asked to read aloud to the experimenter (author) a variety of text stimuli, each prima vista and secunda vista. Two anchor texts were used in every session from the first round on: DOL, an entry of 53 words on the lemma "Dolmetscher" (interpreter) in a youth's lexicon, estimated grade level: 6; and SCHNEE, a fable-like story ("The hare and the snowman"), 191 words, mainly dialogue, estimated grade level: 3. The validity of the subject sample is limited by factors such as school and school type, state, city-country-gap etc., and attrition.

2.2. Subsample selection

Many of the LAUDIO recordings are too poor to be included in a setting suitable for the perceptive isolation of distinct prosodic parameters. These were identified in a screening procedure involving 247 recordings divided into 17 subsets á 14 or 15 files and 51 university student raters (BA and MA in teaching contexts, but non expert in phonetics). Each rater was randomly placed at a headphone plus computer station displaying a spreadsheet plus one of the audio subsets and given a paper including the NAEP scale. After a 15-minute introductory session, raters self activated the files repeatedly and noted their NAEP level evaluations (further details in [10]) in the spreadsheet. This way each subset was rated by three raters. Inter rater reliability was computed via intra class correlation (ICC) analysis using the R-package *irr* ([13], model = twoway, type = consistency, unit = average). The median of the subsets' ICC coefficients was > .8 ("excellent" [14], details in [10]). Now, based on the resulting rating means per recording, two subsamples DOL+ (fluent) and DOL++ (especially fluent) were selected using only prima vista recordings. For DOL+ we eliminated subjects with unsuccessful biographies, i.e. those of which no recording with NAEP fluency > 3 is available at all, arriving at 144 recordings from N_{DOL+} = 31 subjects. Further, for DOL++ all recordings were eliminated that had not been rated NAEP fluency = 4 by all three raters (*post hoc* ICC coefficient = 1). This left us with 26 recordings from N_{DOL++} = 16 subjects. Since all recordings are specified with respect to subject, year and vista, there is a SCHNEE-twin for almost every one of the DOL recordings. Based on the selection process, we assumed these twin SCHNEE recordings to represent equivalent developmental stage and included them in the analyses presented here. For economic reasons, with the SCHNEE recordings we used only the sequence of the first 120 of its 191 words in perception and measure analysis.

3. METHODS

3.1. Dependent variable: overall proficiency [C]

There is no rating scale for advanced fluency as established as the NAEP fluency scale for earlier fluency. In our approach to this *desideratum* we waived explicating any hypothetically relevant prosodic features because there is so little research on post elementary oral reading. Posing specific listening citeria would have been not hypothetical but speculative. Instead we developed a holistic scenario of the rater being part of a jury of a fictious school's oral reading championship, confronted with a 10-point scale as commonly known from popular contexts such as figure skating. The only hint was that the championship was *omnium contra omnem* to keep the raters from integrating perceived voice age in their ratings ("pardon young effect"). The championship scale [C] was introduced and successfully evaluated in [11]. Due to the largeness of the + samples, the C scale was applied to the DOL+ sample and the SCHNEE+ sample with only three fixed expert raters (including the author). This time the rating session was conducted in a fully automated computer setup prepared with the ExperimentMFC functions (including randomisation of audio stimulus presentation) provided by praat [15]. ICC coefficients were > .8 with the DOL+ and > .7

with SCHNEE+ sample ([13], model as above, smaller reliability with SCHNEE can be attributed to specifics of the text).

3.2. Basic independent variables: deviation from text [SMS, E] and duration [T]

We count all text deviating sequences and obviously unplanned pauses > 1s (otherwise regardless of actual length) as occurrences of special mixup sequences [SMS]. In these cases readers are quite obviously confused (or "mixed up") and cognitively busy regaining control. Often, longer duration of such a sequence represents thoroughness of self monitoring rather than lack of proficiency. As errors proper [E] we count deviations on the word level (i.e. deviation sequences < 1s, otherwise regardless of actual length and of self correction). With respect to duration, the measures established in educational research are words/minute (too coarse on our opinion) or correct words/min (even coarser, because this mixes the dimensions accuracy and automaticity mentioned above [9]). In our context we capture the latent trait of habitual tempo (details see [11]) by erasing from the signal all SMS sequences and all inter phrasal pauses regardless of length and dividing the result's duration by the number of speech syllables, arriving at speech rate or habitual tempo T [syllables/s].

3.3. Perceptual independent variables: phrasing [P] and rhythm [R]

There is no room here to discuss phrasing and rhythm sophisticatedly from a phonetic perspective. From an educational perspective it is important to note that prosodic concepts should be comprehendable for children. Otherwise it would be impossible to explain to them what they should work on when according deficits are detected in their performance. Based on this consideration we developed a dichotomous specification of "expressiveness" [1]: clarity of phrasing [P] and diversity of rhythm [R], with rhythm subsuming overall liveliness and phrase level accentuation. For both P and R, 4-level Likert scales were applied to the DOL+ sample and the SCHNEE+ sample in the same ExperimentMFC based expert rater setup as with the C scale (see above). With DOL+, ICC coefficients were > .8 for P and R, with SCHNEE+, ICC coefficients were > .7 for P and R.

4. RESULTS

4.1. The DOL+ subsample

The present paper is focused on especially fluent children, i.e. the ++ subsamples, but note that ++ is a subset of +, meaning that the ++ perceptual values for C, P and R were obtained in the according + reference fields. Within such listening reference fields, certain self normalization tendencies need to be taken into account (see [16] for our work on the normalization of rating data). Tab. 1 gives a descriptive overview of DOL+ as it is the reference field most relevant here.

	mean	sd	min	max
NAEP	3.70	0.33	3.00	4.00
C	7.34	1.10	4.00	9.67
P	2.81	0.64	1.00	4.00
R	2.94	0.67	1.33	4.00
E	3.36	2.26	0.00	10.00
SMS	0.30	0.63	0.00	3.00
T	4.05	0.73	2.31	5.79
AGE	11.17	1.38	8.24	13.74
GRADE	5.21	1.25	3.00	7.00

Tab. 1: Descriptive overview of the DOL+ subsample (144 recordings of 31 subjects reading a quite difficult text of 53 words). C represents overall oral reading proficiency.

NAEP mean and sd indicate skewedness towards 4 which is first evidence of a ceiling effect of this scale when applied to post elementary school contexts (see GRADE). C values display that raters rather make use of the upper half of the 10-point C scale when confronted with recordings from the upper third of the NAEP spectrum. P and R appear to behave in similar ways. E is high with respect to text stimulus length = 53 words, the range is broad. SMS shows the same tendency but seems less important.

4.2. The DOL++ and SCHNEE++ subsamples

	mean	sd	min	max
C	7.77	0.98	5.00	9.33
P	2.92	0.66	1.00	4.00
R	3.09	0.65	1.33	4.00
E	2.44	1.76	0.00	6.00
SMS	0.16	0.37	0.00	1.00
T	4.50	0.59	3.55	5.79
AGE	11.67	1.31	8.24	13.29
GRADE	5.84	1.14	3.00	7.00

Tab. 2: Descriptive overview of the DOL++ subsample (25 recordings of 16 subjects). Note that NAEP = 4 with all recordings.

In the first step, DOL++ description (Tab. 2) is compared with DOL+ description (Tab. 1). C mean is only slightly higher with DOL++ than it is with DOL+, showing that ++ is not so ++ after all. This

points at another problem that appears to have occurred with our method of selecting DOL++: Presumably NAEP = 4 has been attributed to young readers more freely, taking account of perceived voice age, thus making the NAEP scale a referential one not only with respect to the rater's reference field but also with this "pardon young"-effect. Note that with the C scale this effect is checked with the scenario description of a fictious school holding an internal oral reading championship *omnium contra omnem.*

In the second step DOL++ description (Tab. 2) is compared to SCHNEE++ description (Tab. 3). C mean doesn't seem to take account of text difficulty. This is an argument for calculating means of the two (our best estimation for the proficiency latent trait in connection with individual developmental stage) without any normalisation [16]. Note that the better part of the SCHNEE text stimulus sequence considered consists of dialogue, prepared so that speaker identity had to be marked by the reader with prosodic means only. This "reported speech" component goes beyond the mere genre differentiation of factual text (DOL) vs. narrative text (SCHNEE), when it comes to oral reading prosody, giving a greater "playground" for prosodic variation. Accordingly the P and R means are higher with SCHNEE++, showing no self normalisation effect as with C.

	mean	sd	min	max
C	7.65	0.94	6.00	9.67
P	3.40	0.54	2.00	4.00
R	3.39	0.47	2.33	4.00
E	2.36	2.25	0.00	9.00
SMS	0.28	0.54	0.00	2.00
T	4.97	0.55	4.09	6.02
AGE	11.67	1.31	8.24	13.29
GRADE	5.84	1.14	3.00	7.00

Tab. 3: Descriptive overview of the SCHNEE++ subsample (N = 25 recordings of 16 subjects reading a quite easy text sequence of 120 words).

The E value of SCHNEE is remarkable with respect to the fact that in our operationalisation errors are simply counted and not set in proportion to either time or number of words. A comparison of, say, percentage of words done wrong, yields 4,6% with DOL vs. 2% with SCHNEE. Here text difficulty shows its face at last and the same holds for SMS and T.

In the third step means of the DOL++ and the SCHNEE++ values (except E and SMS) are calculated. The results no longer represent features of recordings but of developmental states of oral reading proficiency and its hypothetical components. With the text deviation measures E and SMS we simply sum up DOL++ and SCHNEE++ values. The outcome delivers no new information so we can directly proceed to correlation analysis.

5. DISCUSSION

The main research question finishing our introduction is: Which rating scales reliably capture oral reading proficiency? We presented adequate answers in the form of the newly introduced rating scales C, P and R. It could be shown that C well dissolves the ceiling effect of the NAEP fluency scale.

	C	P	R	E	SMS	T	AGE	GRADE
C	1							
P	.68	1						
R	.54	.72	1					
E	-.42	-.28	-.49	1				
SMS	-.48	-.45	-.48	.19	1			
T	.16	-.33	-.12	.07	.09	1		
AGE	.19	-.21	-.20	.27	.13	.60	1	
GRADE	.33	-.08	-.04	.14	.05	.59	.97	1

Tab. 4: Correlation matrix of the DOL++ and SCHNEE++ unified values (representing 25 child-timepoint combinations).

Tab. 4 shows that P and R capture distinct relevant prosodic aspects of oral reading that contribute a considerable deal to C. Deviation from text needs to be reconsidered in future research: The missing correlation E ~ SMS shows that the rationale that led to their differentiation seems to hold (see 3.2). The most prominent difference between elementary *fluency* and teenage *proficiency* concerns T: once a threshold of, say, T = 4 to 4.5 is passed, it is irrelevant. With respect to AGE and GRADE we tend to conclude that oral reading proficiency is – probably to a much greater extent as, say, mathematical proficiency – more a matter of individual talent than of age or learning advancement. The second question about relations between perceptual parameters and parameters that can be measured in the signal can only now that a perceptual foundation is established be envisaged. As elaborated in [12] we see considerable involvement of two measures of intonational style (*wiggliness* and *spaciousness* of the pitch contour) with C. Maybe in future research we can succeed with bringing perception and signal based values even better in line, rendering costly rating procedures unnecessary.

For educational research we can state that more openness to interdisciplinary perspectives would definitely be promising.

6. REFERENCES

[1] Daane, M.C., Campbell, J.R., Grigg, W. S., Goodman, M.J., Oranje, A. 2005. Fourth-grade students reading aloud: NAEP 2002 special study of oral reading. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington, D.C: Government Printing Office.

[2] Fant, G., Kruckenberg, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR, 30(2)*, 001-080.

[3] Blaauw, E. 1994. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication, 14(4)*, 359–375.

[4] Cowie, R., Douglas-Cowie, E., Wichmann, A. 2002. Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers. *Language and Speech, 45(1)*, 47–82.

[5] De Ruiter, L. E. 2015. Information status marking in spontaneous vs. read speech in story-telling tasks – Evidence from intonation analysis using GToBI, *Journal of Phonetics, 48*, 29–44.

[6] Wolters, A.P., Young-Skuk G.K., Szura, J.W. 2020. Is Reading Prosody Related to Reading Comprehension? A Meta-analysis. *Scientific Studies of Reading, 26(1)*, 1-20.

[7] Godde, E., Bosse, M.L., Bailly, G. 2020. A review of reading prosody acquisition and development. *Reading and Writing, 33(2)*, 399-426.

[8] Miller, J., Schwanenflugel, P. J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98(4)*, 839–853.

[9] Kuhn, M. R., Schwanenflugel, P. J., Meisinger, E. B. 2010. Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45(2)*, 230–251.

[10] Sappok, C, Linnemann, M., Stephany, S. 2020: Leseflüssigkeit – Prosodie – Leseverstehen. Eine Longitudinalstudie zur Entwicklung der Leseflüssigkeit von Jahrgangstufe 3 bis 7. In: Rautenberg, I. (ed): *Evidenzbasierte Forschung zum Schriftspracherwerb*. Baltmannsweiler: Schneider Verlag Hohengehren, 175-209.

[11] Sappok, C. 2021: Exploring Advanced Prosody – eine Best-Practice-Untersuchung zum lauten Lesen in der weiterführenden Schule. In: Gailberger, S., Sappok, C. (eds.): Weiterführende Grundlagenforschung in der empirischen Leseforschung und Lesedidaktik. Theorie – Empirie – Didaktik. SLLD-B 1, 69-81. https://omp.ub.rub.de/index.php/SLLD/catalog/view/189/167/1101.

[12] Wehrle, S., Sappok, C. 2023. Evaluating prosodic aspects of oral reading proficiency in schoolchildren: effects of gender, genre and grade. *Proc. 20th ICPhS*, Prague.

[13] Gamer, M., Lemon, J., Singh, I. 2019. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1 https://CRAN.R-project.org/package=irr

[14] Hallgren, K. A. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol, 8(1)*, 23-34.

[15] Boersma, P., Weenink, D. (2020). Praat: doing phonetics by computer [Computerprogram]. Version 6.1.16. http://www.praat.org/.

[16] Sappok, C., Arnold, D. 2012. More on the Normalization of Syllable Prominence Ratings. In: *Proc. 13th Interspeech*, Portland, Oregon, USA.

CONTRAST AND GIVENNESS IN BIASED DECLARATIVE QUESTIONS

Sophie Repp, Heiko Seeliger

Department of German Language and Literature I, University of Cologne
sophie.repp | heiko.seeliger@uni-koeln.de

ABSTRACT

The prosodic marking of information structure (IS) in German polar questions with an interrogative syntax has been observed to differ from that in assertions, which might be due to different final contours (falling, rising), different semantic-pragmatic restrictions, or a general inertness of questions to IS. In a production study, we investigated whether German declarative questions show similar deviations. We tested transitive structures where the object and the lexical verb were either both given, both new, or one was contrastive while the other was given, and found a strong adherence to default prosody across conditions (inertness). The association of contrast with high, and givenness with low, prosodic prominence familiar from assertions could be observed but it was modulated by speech-act-specific characteristics. Overall, there was substantial inter-individual variation.

Keywords: prosody, biased declarative question, givenness, contrast, focus

1. INTRODUCTION

In intonation languages like German, information structure (IS) influences the prosodic prominence of the corresponding expression. In assertions, givenness is associated with prominence reduction in terms of deaccentuation, the choice of less prominent accent types, shorter duration or lower pitch (F0) [1, 4, 5, 6]. Narrow non-contrastive focus and contrastive focus are associated with a gradual increase in prominence, relative to newness, for instance in terms of fewer H+!H* accents and more L+H* accents, greater intensity, longer duration and higher mean and maximum F0 for narrow vs. broad focus [2, 11, 12, 13, 15]. Whether or not there are IS effects in the prenuclear region is controversial: in sentences with corrective focus there may be deaccentuation and reduced gradient prominence [3], yet [19] suggest that there is at most slight compression. Prenuclear prominence reduction arguably increases the prominence difference with a contrastive element, as does post-focal compression. [19, 32].

Non-assertive speech acts have been found to show deviations from these prosody-IS associations [24, 25, 29]. To wit, different final contours—falling vs. rising—, which are linked to (but not determined by) the speech act type, may interact with IS marking: In rising polar questions with an interrogative syntax (PQs), contrast-associated prominence increase in the final part of the utterance has been observed to be realized by lower, rather than higher F0, arguably because lower F0 is a prominent deviation from a rising baseline [20, 24, 29]. Another reflex of contrast marking in PQs is a prominence decrease in the prenuclear region by deaccentuation [23, 29]. In other languages, there are other differences between questions and assertions, e.g. Russian questions do not seem to show the subtle prominence distinctions for different focus types known from assertions, and have been suggested to be IS-inert [22]. Then again, in Swedish, questions seem to show regular contrast marking [27].

Turning to givenness marking, German PQs seem to lack prominence reduction in certain IS configurations, compared to assertions. In all-given assertions where the truth of the denoted proposition is highlighted, the nuclear accent is not realized in its default position but occurs on the finite verb, indicating VERUM focus [17, 31]. The rest of the utterance is deaccented. In questions, VERUM signals either a previous speaker belief [26] or impatience [7, 8]. If such signals are not appropriate in a discourse, there is no deaccentuation in all-given PQs, and the nuclear accent occurs in its default position, which has been argued to ensure the headedness of the intonational phrase [29]. Interestingly, [29] found no phonetic reduction in all-given PQs, either.

This paper aims at broadening the empirical picture of IS marking in *yes/no*-questions by studying DQs in German. DQs come with a contextual *bias*: their declarative syntax signals that there is contextual evidence supporting the truth of the questioned proposition [16, 28, 30, 33]. The speaker uses a DQ to double-check the evidence, possibly because of a conflicting previous belief and concomitant amazement. [23] show that the latter aspect, if implemented as a surprised vs. matter-of-fact attitude of the speaker, plays a role for the prosody of DQs in terms of declination, speaking rate and voice quality.

[23] also observe that DQs have fewer prenuclear accents than comparable assertions, which they adduce to the function of DQs: DQs tend to contain mostly given material because speakers double-check the truth of a proposition that has just been asserted or implicated. However, DQs are actually quite flexible regarding their IS. The information in a DQ can be all given (if all the information was just mentioned) but it can also be all new (if the proposition denoted by the DQ can be plausibly deduced from the context but was not mentioned). It is also possible that some of the information is given whereas other information is in contrastive focus. Consider the DQ *Jana wants to hunt deer?*. A speaker might use this DQ to double-check that it is deer that Jana wants to hunt because a plausible alternative was mentioned in the context, or, alternatively, that it is hunting that Jana wants to involve the deer in rather than for instance examining it. Thus, the speaker is considering contrastive alternatives so that the material that is double-checked is in contrastive focus.

In this paper we present a production study on matter-of-fact DQs investigating the prosodic reflexes of contrast and givenness vs. newness. We hypothesized that DQs would show assertion-like IS-prosody associations as well as deviations. Specifically, we expected, on the one hand, increased prominence for contrast (possibly with prenuclear reduction, certainly with post-nuclear reduction), and decreased prominence for givenness (similar to assertions). On the other hand, we expected contrast in the final part of the utterance to be marked by lower rather than higher F0 due to a rising baseline (similar to PQs). For all-given DQs, we expected reduced or even a lack of givenness marking and thus default accentuation: in matter-of-fact DQs VERUM is inappropriate because there is no strong speaker belief or impatience. In view of [29]'s findings for all-given PQs, there might not even be phonetic prominence reduction for all-given DQs in comparison to DQs with some new material. In other words, DQs with given vs. new material might be prosodically identical, suggesting some IS-inertness.

2. METHOD

The materials of our study consisted of dialogues between three speakers who chat about various friends. The first two speakers (prerecorded; presented in written and auditory form) set the scene, and speaker 2 utters a speculation which may be taken to suggest the truth of a proposition *p* concerning the expected action of a friend. Speaker 3 (= participant) then asks a DQ, *p?*, which double-checks the truth of *p*, e.g., *Jana will da dann Rehe jagen?* (lit. `Jana wants.to there then deer hunt`; 'Jana wants to hunt deer there then?'; Structure: subject-auxiliary-adverb-adverb-object-verb). The DQ is followed by an assertion, also uttered by speaker 3.

There were 12 dialogues in 4 IS conditions each, disambiguated by the left and right context of the target DQ. Condition [$O_N V_N$]: object and verb of the DQ are new, the follow-up assertion is neutral, e.g. *That sounds interesting*; [$O_G V_G$]: object and verb are given in the left context, the follow-up is neutral; [$O_C V_G$]: contrastively focused object and given verb, with a focus alternative for the object (e.g. *Hasen* 'hare') in the left context and in the follow-up; [$O_G V_C$]: given object and contrastive verb, with a focus alternative for the verb (e.g. *erforschen*, 'examine') in the left context and the follow-up. In all conditions, the subject and the auxiliary are given. In addition to the experimental items, there were 4 filler dialogues in 4 sentence types each (*wh*-question, *wh*-exclamative, negative declarative, polar question). All participants saw all items.

24 native speakers of German (2 male; aged 19-30, mean: 24.4) participated in the experiment. They were students of the University of Cologne and were paid for their participation or received course credit. Most of the participants (16) were from North Rhine-Westphalia. The recordings were annotated in Praat [9] by trained research assistants (GToBI guidelines [14]).

3. RESULTS

The DQs produced by the participants overwhelmingly ended with a high or rising boundary tone: 94%. There were 845 high rises (L* H-^H%) (79% of rising contours) and 226 low rises (L+H* L-H%) (21%). 1.8% of DQs had a low boundary tone, while 4.3% had other contours, e.g. (mid-)high, level plateaus (≈ H-%) or incomplete falls.

For the analysis of the dependent variables reported below we fitted one model (using `lmerTest` [18]) comparing all four factor levels using treatment coding with the $O_N V_N$ condition as the baseline for every dependent variable, and one model that compared $O_G V_G$ to $O_C V_G$ and to $O_G V_C$. We also investigated utterance-level measures, which for space reasons we only mention in passing. Absolute F0 values (i.e. mean, maximum and minimum) were converted to semitones relative to each speaker's median F0 value. Relative F0 values (excursion) are given in semitones. The raw data was imported into R using rPraat [10].

3.1. Accent placement and accent type

Fig. 1 shows the distribution of accents and accent types for the accentable syllable of each word by condition. Since the object was accented extremely often across conditions, a model could not be fitted. Yet we can observe that there was *some* deaccentuation of given objects in the presence of a contrastive verb: 86.6% object accentuation in O_GV_C vs. 97.5% in O_NV_N, 99% in O_GV_G, 100% in O_CV_G. The verb was accented in 9.5% of O_NV_N utterances, 43.3% in O_GV_C, 6% in O_GV_G, and 2.5% in O_CV_G. The increase in O_GV_C compared to O_NV_N is significant ($b = 2.91, SE = 0.3, z = 9.4, p < 0.001$), as is the decrease in O_CV_G vs. O_NV_N ($b = -1.6, SE = 0.47, z = -3.4, p < 0.01$). There were no significant differences concerning the presence of a prenuclear accent on the subject or any other elements.

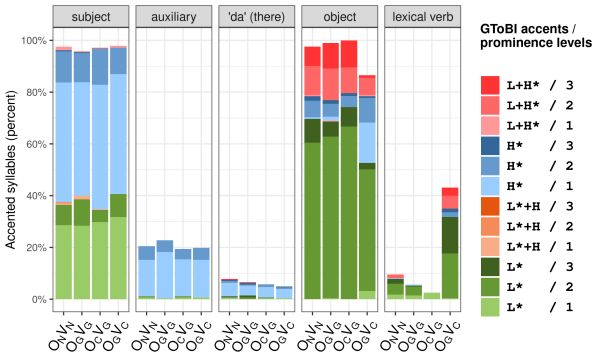


Figure 1: GToBI accent types [14] and prominence levels [21] for the accentable syllable of each word by experimental condition

For the accent type on the object we analyzed the proportion of L* among all accents, and the proportion of L+H* among accents with high starred-tones (L+H*, H*). The proportion of L* on the object was smaller in O_GV_C than in O_NV_N and than in O_GV_G (O_NV_N: $b = -1.3, SE = 0.27, z = -4.7, p < 0.001$ || O_GV_G: $b = -0.5, SE = 0.13, z = -3.9, p < 0.001$). The proportion of L+H* was smaller in O_GV_C than in O_NV_N and than in O_GV_G ($b = -2.2, SE = 0.42, z = -5.2, p < 0.001$ || $b = -1.2, SE = 0.22, z = -5.7, p < 0.001$).

3.2. Acoustic measures for accented syllables of object and lexical verb

For the **F0-related measures** of the object, we pooled objects with high starred tones (H*, L+H*), as they were characterized by comparable contour shapes. L* accents were treated separately but they showed no F0-related significant differences. For objects with high starred tones, mean F0 was lower in O_GV_C than in O_NV_N and than in O_GV_G ($b = -1.42, SE = 0.25, t = -5.7, p < 0.001$ || $b = -0.8, SE = 0.12, t = -6.7, p < 0.001$). Maximum F0 was also lower in O_GV_C than in O_NV_N and O_GV_G ($b = -2.4, SE = 0.37, t = -6.5, p < 0.001$ || $b = -1.4, SE = 0.2, t = -6.9, p < 0.001$). Minimum F0 did not show significant differences after correction for multiple comparisons. F0 excursion on the accented syllable of the object showed the same pattern as mean and maximum F0: it was smaller in O_GV_C than in O_NV_N and O_GV_G ($b = -2, SE = 0.42, t = -4.8, p < 0.001$ || $b = -1.15, SE = 0.3, t = -3.7, p < 0.01$). Thus, objects with high starred tones showed reduced prominence in terms of F0 mean, maximum and excursion, if the object was given and the verb was contrastive in comparison to the new and the all-given conditions. The latter two conditions neither differed from each other nor from O_CV_G. Overall, there thus were significant differences for accented objects only for one of the nuclear contours—(L+)H* L-H%—, and these effects were prominence-reducing effects that verb contrast had on the object.

For the F0-related measures of the verb, we pooled all accented verbs regardless of accent type. The only significant difference with a reasonable sample size is that the F0 excursion was larger in O_GV_C than in O_NV_N ($b = 1.3, SE = 0.4, t = 2.9, p < 0.05$). Maximum F0 on the verb was significantly lower in O_CV_G than in O_NV_N but the O_CV_G condition had a sample size of 7.

For the **duration analysis**, we follow [2] and take the foot as the relevant domain for IS marking (= entire object; entire verb). Also, since the utterance-level speech rate was significantly higher in the contrastive conditions, we analyzed relative duration, i.e. the proportion of the utterance taken up by the relevant foot. The relative duration of accented objects was longer in O_CV_G than in O_NV_N ($b = 0.006, SE = 0.002, t = 3.1, p < 0.01$), and it was shorter in O_GV_C than in O_NV_N and O_GV_G ($b = -0.007, SE = 0.002, t = -4, p < 0.001$ || $b = -0.004, SE = 0.002, t = -3.3, p < 0.01$).

Syllable-level mean intensity of objects was non-significantly lower in O_GV_C than in O_NV_N. Maximum intensity was significantly lower in O_GV_C than in O_NV_N ($b = -0.7, SE = 0.27, t = -2.6, p < 0.05$). There were no differences on the verb.

3.3. Pre- and postnuclear regions

Prenuclear region. For the F0 of the accented syllables of the subject (separate models for H* and L*) and the auxiliary (only models for H*) there were no significant differences. Neither were there differ-

ences for duration or intensity.

Postnuclear region. To (further) investigate the influence of IS on the postnuclear region, we investigated F0, duration and intensity across the whole verb when it was *unaccented* (for F0: in separate models split up by object accent type (L* vs. [L+]H*)). For F0 and duration, there were no significant differences (verbs tended to be longer in O_GV_C than in O_NV_N before multiple comparison adjustment). For intensity, we found that verbs were less loud in O_CV_G than in O_NV_N ($b = -0.8, SE = 0.2, t = -4.3, p < 0.001$). Thus there was post-focal reduction only in terms of intensity.

3.4. Inter-individual variation

There was substantial inter-individual variation in response to the experimental manipulations. While contrast on the verb led some speakers to shift the nuclear accent to the verb with attendant deaccentuation of the object, others kept the nuclear accent in its default position. The 'keepers' did not use more subtle, possibly continuous means of prosodic contrast marking: in extreme cases, they produced identical-sounding contours in all conditions. Interestingly, 'shifters' produced more clear-cut distinctions between objects in O_CV_N vs. O_NV_N, even though the classification of a shifter is based on the O_GV_C condition, where the object is *given*. In other words, their behavior regarding a fairly 'strong' experimental manipulation (location of the nuclear accent) predicted their behavior regarding a more subtle manipulation (new information vs. contrast on the object). There was no indication that the inter-individual variation was associated with regional provenance.

4. DISCUSSION

Overall, we found for DQs IS-prosody associations familiar from assertions, and we found deviations from these associations familiar from PQs. The most striking deviation is perhaps the quite pervasive reduced sensitivity to IS. As predicted, there was no prominence reduction in all-given DQs in comparison to DQs with a new object and verb. Both types of DQs had default prosody with the object as the carrier of the nuclear accent and there were no phonetic differences. The lack of accent shift plausibly results from the matter-of-fact nature of the DQs we tested: VERUM marking is not appropriate. It is nevertheless remarkable that there is no prominence reduction at all, which replicates the findings for PQs [29].

Regarding the contrastive conditions, we found substantial IS effects only for DQs with a given object and a contrastive verb. For DQs with a contrastive object and a given verb the effects were very small. Thus, the only IS constellation that led to a substantial departure from the default contour is the one where the nuclear accent may shift from its default position—the object—elsewhere: to the lexical verb (at least in a subpopulation of the speakers). This is a licit shift because the verb is contrastive. Still, note that speakers also produced double-accent structures with an accent on object and verb. In both cases, we have a clear case of contrast marking on the verb: new and given verbs are hardly ever accented.

The prominence reduction on the given object in utterances with a contrastive verb (less prominent accent types, phonetic measures) is difficult to interpret. It might 'directly' mark givenness on the object, but it might also be due to prenuclear reduction before the prominent contrastive verb (or a combination of both). Since all-given DQs seem to be inert in terms of IS marking, the observed reduction for given objects before contrastive verbs in comparison to new objects before new verbs might indeed be due to givenness marking on the object. However, we also observed that other prenuclear elements, which were always given—the subject and the auxiliary—showed no differences between the conditions. We therefore assume that the prominence reduction on a given object which is adjacent to a contrastive verb more likely serves to boost the prominence of the verb.

The finding that in DQs with a contrastive object the IS effects are very small, and there is hardly any post-focal reduction, which even in view of the general lack of accents on the verb in the new and all-given conditions might still have materialized clearer in the phonetic measures, seems to support the view that DQs generally have a tendency towards IS-inertness. Like [22] for Russian PQs, we did not find a difference between focus types.

As for the expected difference of prominence increase by higher vs. lower F0 depending on a falling vs. rising final pitch contour, the results are somewhat confirmative. While we did not find phonetic effects of contrast for either L* or (L+)H* accents, it is a new finding that L*—an accent that in assertions is considered not to be very prominent [5]—is compatible with contrastive focus: if it occurs within a rising final contour. That is, contrast is not inherently linked to high pitch or to L+H*, specifically.

In sum, DQs seem to display a reduced sensitivity to IS, which, however, is speaker-dependent. Future research directly comparing IS in assertions and *yes/no*-questions within speakers must show if the observed reluctance of some participants to prosodically mark IS in questions can be replicated.

Acknowledgements: Funding: DFG Deutsche Forschungsgemeinschaft (German Research Foundation), Project-ID 281511265, SFB 1252. Research assistants: Franziska Busche, Lukas Kurzeja, Karoline Marliani, Marlon Siewert, Sven Weber.

5. REFERENCES

[1] Baumann, S. 2006. *The Intonation of Givenness – Evidence from German*. Niemeyer.

[2] Baumann, S., Becker, J., Grice, M., Mücke, D. 2007. Tonal and articulatory marking of focus in German. Trouvain, J., Barry, W. J. (eds), *Proc. 16th ICPhS* Saarbrücken, 1029–1032.

[3] Baumann, S., Grice, M., Steindamm, S. 2006. Prosodic marking of focus domains – categorical or gradient? Hoffmann, R., Mixdorff, H. (eds), *Proc. Speech Prosody 2006* Dresden, 301–304.

[4] Baumann, S., Riester, A. 2013. Coreference, lexical givenness and prosody in German. *Lingua*, 16–37.

[5] Baumann, S., Röhr, C. 2015. The perceptual prominence of pitch accent types in German. *Proc. 18th ICPhS* Glasgow, Paper number 298, pp. 1–5.

[6] Baumann, S., Röhr, C. T., Grice, M. 2015. Prosodische (De-)Kodierung des Informationsstatus im Deutschen. *Z. für Sprachw.* 34(1), 1–42.

[7] Beltrema, A., Meertens, E., Romero, M. 2018. Decomposing cornering effects: an experimental study. Sauerland, U., Solt, S. (eds), *Proc. Sinn und Bedeutung 22* volume 1 Berlin, 175–190.

[8] Biezma, M. 2009. Alternative vs polar questions: the cornering effect. Cormany, E., Ito, S., Lutz, D. (eds), *Proc. SALT 19* Columbus, 37–54.

[9] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer.

[10] Bořil, T., Skarnitzl, R. 2016. Tools rPraat and mPraat. Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds), *Proc. Text, Speech, and Dialogue 19* Brno, 367–374.

[11] Braun, B. 2006. Phonetics and phonology of thematic contrast in German. *Lang. Speech* 49, 451–493.

[12] Braun, B., Tagliapetra, L. 2010. The role of contrastive intonation contours in the retrieval of contextual alternatives. *Lang. Cogn. Process* 25(7-9), 1024–1043.

[13] Breen, M., Fedorenko, E., Wagner, M., Gibson, E. 2010. Acoustic correlates of information structure. *Lang. Cogn. Process* 25(7-9), 1044–1098.

[14] Grice, M., Baumann, S., Benzmüller, R. 2005. German intonation in autosegmental-metrical phonology. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: OUP, 55–83.

[15] Grice, M., Ritter, S., Niemann, H., Roettger, T. B. 2017. Integrating the discreteness and continuity of intonational categories. *J. Phon.* 64, 90–107.

[16] Gunlogson, C. 2003. *True to Form: Rising and Falling Declaratives as Questions in English*. Routledge.

[17] Höhle, T. 1992. Über Verum-Fokus im Deutschen. In: Jacobs, J. (ed), *Informationsstruktur und Grammatik*. Westdeutscher Verlag, 139–197.

[18] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82(13), 1–26.

[19] Kügler, F., Féry, C. 2017. Postfocal downstep in German. *Lang. Speech* 60(2).

[20] Kügler, F., Genzel, S. 2012. On the prosodic expression of pragmatic prominence: The case of pitch register lowering in Akan. *Lang. Speech* 55, 331–359.

[21] Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., Peters, J., Ritter, S., Röhr, C., Schweitzer, A., Schweitzer, K., Wagner, P. 2015. DIMA – Annotation guidelines for German intonation. *Proc. 18th ICPhS* Glasgow, Nb 317, pp. 1–5.

[22] Meyer, R., Mleinek, I. 2006. How prosody signals force and focus – a study of pitch accents in Russian yes-no questions. *J. Pragmat.* 38(10), 1615–1635.

[23] Niebuhr, O., Bergherr, J., Huth, S., Lill, C., Neuschulz, J. 2010. Intonationsfragen hinterfragt: die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax. *Z. Dialektol. Ling.*, 304–346.

[24] Repp, S. 2020. The prosody of wh-exclamatives and wh-questions in German: Speech act differences, information structure, and sex of speaker. *Lang. Speech* 63(2), 306–361.

[25] Repp, S., Seeliger, H. 2020. Prosodic prominence in polar questions and exclamatives. *Front. Commun.* 5(53), 1–26.

[26] Romero, M., Han, C.-h. 2004. On negative yes/no questions. *Linguist. Philos.* 27(5), 609–658.

[27] Seeliger, H., Repp, S. 2017. On the intonation of Swedish rejections and rejecting questions. Abrahamsen, J. E., van Dommelen, W. A., Koreman, J. (eds), *Proc. Nordic Pros. XII* Trondheim, 135–146.

[28] Seeliger, H., Repp, S. 2018. Biased declarative questions in Swedish and German: The syntax of negation meets modal particles (väl and doch wohl). Dimroth, C., Sudhoff, S. (eds), *The Grammatical Realization of Polarity Contrast: Theoretical, empirical, and typological approaches*, 129–172.

[29] Seeliger, H., Repp, S. 2023. Information-structural surprises? Contrast, givenness and (the lack of) accent shift and deaccentuation in non-assertive speech acts. *Laboratory Phonology* 14(1), 1–46.

[30] Trinh, T. 2014. How to ask the obvious - a presuppositional account of evidential bias in English yes/no questions. *MIT Working Papers in Linguistics* 71, 227–249.

[31] Turco, G., Braun, B., Dimroth, C. 2014. When contrasting polarity, the Dutch use particles, Germans intonation. *J. Pragmat.* 62, 94–106.

[32] Xu, Y., Xu, C. X. 2005. Phonetic realization of focus in English. *J. Phon.* 33(2), 159–197.

[33] Šafářová, M. 2006. *Rises and falls: studies in the semantics and pragmatics of intonation*. PhD thesis University of Amsterdam.

THE EFFECT OF (NON-)NATIVE LANGUAGE AND TASK COMPLEXITY ON SPEECH ENTRAINMENT

Lucia Mareková[1], Joanna Kruyt[2,3], Štefan Beňuš[1,2]

[1]Constantine the Philosopher University in Nitra, Slovakia
[2]Institute of Informatics, Slovak Academy of Sciences in Bratislava, Slovakia
[3]Slovak Technical University, FIIT in Bratislava, Slovakia
lucia.marekova@ukf.sk, joanna.kruyt@savba.sk, sbenus@ukf.sk

ABSTRACT

Speech entrainment (also known as alignment or accommodation) has been documented on various linguistic and paralinguistic levels but often with complex and sometimes conflicting outcomes. Hence, the understanding of the mechanism(s) underlying this complex behaviour in dyadic conversations is still limited. In an effort to increase this understanding, the current study tests the effect of language (L1 Slovak vs. L2 English) and task complexity (easy, medium, difficult) on local and global entrainment in f0 and intensity in a within-subject design. Pairs of undergraduates played a collaborative game of giving directions structured into three levels gradually rising in complexity. The results corroborate other recent findings in yielding complex and unexpected patterns, particularly failing to show the assumed link between cognitively easier tasks and greater entrainment.

Keywords: entrainment, task complexity, L1 vs. L2, cognitive load.

1. INTRODUCTION

Entrainment (also referred to as alignment, convergence or accommodation) is the tendency of communicative partners to get closer to each other in verbal or non-verbal aspects of their communication. The patterns of entrainment in speech and language characteristics have been extensively analysed in native oral interactions, bringing insights on the cognitive aspects of inter-personal entrainment on various linguistic and para-linguistic levels, such as semantics [1], syntax [2, 3, 4], prosody, and phonetic realizations [5, 6, 7], or focusing on diverse social and psychological factors underlying entrainment [e.g. 8, 9]. These studies show that entrainment is a complex phenomenon occurring on multiple levels of communication, depending on diverse underlying factors whose character is still not well understood. Given this complexity and inconclusive results of previous studies, various questions regarding the patterns of speech entrainment on multiple levels of communication are pending a satisfactory address.

For instance, entrainment on the interface of native (L1) and non-native (L2) is not well understood due to conflicting findings. On the one hand, L2 speakers were likely to imitate L1 speakers when sharing certain social identities [10], or they converged in vowel pronunciation towards those about whom they assumed were L1 speakers, while diverged from those about whom they assumed were L2 speakers [11]. On the other hand, [12] in a longitudinal study found that native English speakers converged in rhythm to L2 English speakers, while L2 speakers did not converge. A similar pattern was shown in [13]: L1 English speakers aligned towards Spanish-accented speakers in a shadowing task. Yet, [14] showed greater entrainment in matched language dyads than in mismatched (L1 vs. L2) dyads.

It is believed that metalinguistic factors play a significant role in these complex outcomes of L1 vs. L2 entrainment studies. For instance, the level of entrainment in L1-L2 dialogues may be influenced by the L2 speaker´s metalinguistic awareness of L1 speaker´s knowledge of the language [15], or by a set of other L2-specific issues such as insecurity in pronunciation patterns preventing the L2 speaker to lexically converge toward the L1 speaker, or linguistic similarity of interlocutors´ L1 and L2 affecting sensitivity to automatic priming [16].

To the best of our knowledge, the analysis of the patterns of entrainment on the interface of L1 vs. L2 as a within-subjects factor in semi-spontaneous conversations warrants further exploration. There are three relevant questions in this regard. First, is the speech entrainment language dependent? Second, do the same speakers align differently when interacting in their L1 compared to interacting in their L2? Third, is the level of entrainment dependent on the cognitive demand of the conversation? For example, acoustic analysis in [17] showed that entrainment was more significant in easier conditions with fewer errors and shorter completion time. A holistic analysis in [17] brought the same results.

Entrainment covers a wide range of behaviours and literature offers many methods for assessing it; different methods presumably tap into different aspect of this behaviour. One of several dimensions

that differentiates the methods is locality: some try to capture entrainment globally at the level of a conversation, for example comparing two halves of each dialogue, and some evaluate speakers' adjustments more locally at each turn-exchange, e.g. [18]. Previous studies showed that this dimension might show differences among languages and capture individual differences, e.g. [19,20].

In the present study, we test the relationship of entrainment in L1 (Slovak) vs. L2 (English) spoken dyadic communication between undergraduate students playing an interactive map game of giving directions sequenced into three levels of task complexity: easy, medium and hard.

Regarding the language effect on entrainment, theoretical accounts lead to contradictory predictions. On the one hand, the patterns of entrainment may be similar for L1 as for L2, mainly if we consider its dependence on various social mechanisms (dominance and roles), which are in our design the same in L1 as in L2. On the other hand, speakers´ perception of each other's L2 proficiency may cause differences in how the same speakers entrain towards each other in L1 vs. in L2 due to changes in social dynamics. For the task complexity effect on entrainment, we expect more entrainment in the easy and medium tasks than in the hard task, which is based on previous findings [17].

Understanding entrainment in L2 spoken interactions and its links to L1 speech entrainment has potential for 1) better understanding and ultimately formally modelling the cognitive nature of this process, and 2) developing communicative techniques for L2 learning informed by the patterns and functions of speech entrainment since more and more interpersonal spoken interactions takes place in L2. Specifically, our main hypothesis is that we observe more entrainment in L1 and easier tasks compared to L2 and harder tasks since lower cognitive demands might facilitate managing and navigating social space and entrainment as a means for doing that [17]. Alternatively, more prevalent entrainment observed for complex tasks and L2 might support the theories advocating an automatic, possibly priming-based, account of entrainment [21].

2. METHODS

Fifteen pairs (8 females, 3 males, 4 mixed) of native Slovak undergraduates studying English took part in the experiment. The subjects used a standardized test to self-evaluate their L2 proficiency, with results ranging from the B2 to C1 levels.

Each pair completed six collaborative communicative tasks involving giving directions based on an information gap. The students were randomly assigned the role of the guided person or the guide and kept the roles for all tasks. The main factors under investigation were language, L1 Slovak (SK) and L2 English (EN), and task complexity (easy, medium, difficult). The easy task, top of Fig. 1, included giving directions while the route was indicated. In the medium task, middle of Fig. 1, the route was not indicated, but the number of elements in the map and possibilities of route choice was low. The map of the difficult task, bottom of Fig. 1, included multiple elements and afforded several possibilities for the subjects to complete the task so that they had to make several collaborative decisions. This design resulted in 90 conversations (15 pairs x 3 complexity levels x 2 languages). The tasks were blocked by the language, i.e. all three complexity levels together, but the order of the language (SK first vs. EN first) and the order of the tasks within a language block were randomised.



Figure 1: Maps used for three complexity levels: easy (top), medium (middle), and difficult (bottom).

Our task design brought together a variety of previous approaches. Firstly, following [22], our tasks contain three basic elements: there is some kind of a gap, subjects rely on their own resources (linguistic and non-linguistic), and there is an apparent communicative outcome. Next, we distinguished task complexity from task conditions/communicative stress: gradually rising task complexity is not determined by planning time, time limits for completing the tasks, or the number of

participants [23,24]. Task complexity is rather created by manipulating the character of the tasks: In line with [24], our tasks differed in the number of elements occurring in the map and by the reasoning demand (in the difficult task, time labels were included urging the student to pick the fastest route and provide arguments to support their choice). The difference is also in the prior knowledge (in line with [23]): In the easy task, the students knew the area well, as it is the centre of a city in which their university is located. The medium task is situated in the same city but near the suburbs. The difficult task is situated in a distant city. The factor of prior knowledge was verified by a post-test questionnaire.

We used an automatic tool available in Praat to separate speaking time from silent time. The silence threshold was set to -25dB and minimum silence duration to 100ms. Manual correction of the silent vs. speaking intervals boundaries for samples with consistent errors was done. We then employed Praat for extracting the raw values for f0 medium, f0 range, and mean intensity from each speaking interval. We minimised the effect of outliers and octave jumps in f0 extraction by initially setting the floor and ceiling by the speaker biological sex and following the two-pass approach in [25] for floor/ceiling adjustment.

We analysed global, task-based entrainment following [26] that compares value means extracted from the initial and final thirds of the conversation respectively (using IPU numbers for the splitting into thirds). The local entrainment was assessed at each turn-exchange using the approach in [7] that operationalizes entrainment with local proximity, convergence, and synchrony; see both papers for definitions and formulas. Our implementation of both global and local methods is described in [27].

With values for global and local entrainment we then employed standard statistical modelling using linear mixed effects in R (package lmerTest) exploring the effect of LANGUAGE (L1 Slovak, L2 English), task COMPLEXITY (easy, medium, difficult), and their interaction on f0 median, f0 range, and intensity specifying random slopes for the dyad and the order of the language/tasks. We first explored the *anova* function on the model for the main effects and then the function *emmeans* in the package of the same name for the post-hoc comparisons in case the anova results returned significant values.

3. RESULTS

The effect of language and task complexity on the global entrainment in f0 median, f0 range and intensity is illustrated in Fig. 2. Positive values indicate entrainment and negative disentrainment. Pooling over the two factors, f0 range displays weak entrainment ($t(89) = 2.36$, $p = 0.02$) and intensity weak disentrainment ($t(98) = -2.21$, $p = 0.03$). Linear mixed effects modelling shows that neither the effects of LANGUAGE (SK vs. EN), COMPLEXITY, nor their interaction, on entrainment in f0 are significant. For intensity, there is a marginal interaction between LANGUAGE and COMPLEXITY ($F = 3.1$, $p = 0.05$), which is related to the fact that in EN, subjects dis-entrain more in medium tasks than in difficult tasks, while in SK it is vice versa.

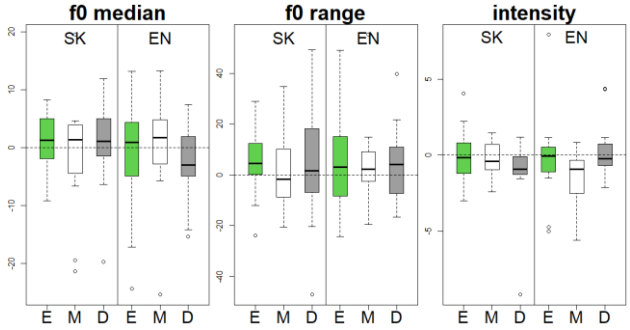


Figure 2: Global entrainment in f0 median (left), f0 range (middle), and intensity (right) divided by Language (leftmost 3 bars in each plot correspond to L1 Slovak, rightmost to L2 English), and Complexity (E[asy] - green, M[edium] - white, and D[ifficult] - grey).

The results pertaining to the local measures of entrainment are summarised in Table 1. For each of the three types of local entrainment (proximity, convergence, synchrony) and the three features (f0 median, f0 range, intensity), the table shows in the leftmost cell of each of the three columns if significant entrainment (ent) or disentrainment (dis) occurred when pooling over both factors and using a one-sample t-test. The rightmost cells show the significant patterns regarding LANGUAGE and COMPLEXITY.

	Proximity		Convergence		Synchrony	
f0 med	ent***	–	–	E_{EN}>M,D_{EN} M,D_{SK}>E_{SK}	dis***	–
f0 range	ent***	E_{EN}>E_{SK}	ent*	–	–	–
intensity	ent*	–	–	all>E_{SK}	dis***	–

Table 1: Local measures of entrainment with proximity, convergence, and synchrony. Task complexity in caps (E[asy], M[edium], D[ifficult]) and language (SK, EN) in subscript. "***" correspond to $p < 0.001$, and "*" to $p < 0.05$. See text for a detailed description.

In local proximity assessed on f0 range, the interaction between COMPLEXITY and LANGUAGE

returned a significant effect ($F = 3.84$, $p = 0.027$). The post-hoc exploration suggests that entrainment in English easy conversations was greater than in Slovak easy ones. Local proximity on f0 median and intensity did not show any significant effects of the two factors or their interaction despite the overall entrainment when both of these factors were pooled.

For local convergence in f0 median, dialogues in English were more entrained than in Slovak ($F = 10.13$, $p = 0.002$) the interaction with task complexity was also highly significant ($F = 11.17$, $p < 0.001$) stemming primarily from the difference between languages in the easy tasks ($F = 5.6$, $p < 0.001$): entrainment occurred in English and disentrainment in Slovak, which contrasted with the lack of (dis)entrainment in medium and difficult tasks.

No significant result was found in local convergence in f0 range. In intensity, however, a tendency for COMPLEXITY ($F = 3.0$, $p = 0.056$) and the significant interaction with LANGUAGE ($F = 6.43$, $p = 0.003$) were found. The pattern is partially similar to the one in f0 median: The disentrainment in Slovak easy tasks was significantly greater than the lack of entrainment in all other conditions.

Local synchrony yielded no significant effects of LANGUAGE or COMPLEXITY on the three dependent variables despite overall disentrainment in f0 median and intensity.

4. DISCUSSION AND CONCLUSION

The current study analysed the effect of language, native vs. non-native, and task complexity on local and global prosodic entrainment in f0 and intensity. The results do not yield a robust overall pattern and provide only several indicative systematic findings.

First, the effects of (non-)native language and task complexity are relatively small. We used one measure for global entrainment and three for the local one on three prosodic features. This gives 12 potential chances for these factors to affect inter-personal acoustic-prosodic entrainment. Yet, only 4 measures showed a significant effect (global intensity, local proximity f0 range, local convergence f0 median and intensity). Thus, we conclude that the patterns of speech entrainment tend to work similarly irrespectively of the language (native or non-native) or task complexity. This result might also suggest that in a learning environment, strategies positively influencing the link between speech entrainment and positive social aspects (social attractiveness, trust, task effectiveness, etc.) might provide comparable outcomes irrespective of the language of the interaction or the complexity of the task at hand.

Second, if the effect of language/complexity occurred, the easier tasks and tasks in the L1showed less entrainment (or showed disentrainment) than more difficult tasks and tasks in the L2. This was the case for three of the four significant patterns (local proximity f0 range, local convergence f0 median and intensity) where the easy tasks in Slovak showed the least entrainment. Hence, we might speculate that acoustic-prosodic entrainment in task-based spoken interactions is associated with coping with tasks demanding greater cognitive resources. The alternative possibility, that tasks with low cognitive demands free up resources for speech entrainment being used in managing and navigating social space, received little support despite findings of [17] where easier tasks yielded greater entrainment than more complex ones. Naturally, this speculation requires further rigorous testing.

Third, while the effect of language/complexity described above is relatively consistent across the selected entrainment measures, the assessment of entrainment itself is much less consistent. While f0 range showed entrainment both globally and as local proximity and convergence, intensity showed disentrainment globally and as local synchrony but entrainment as local proximity. This low consistency across entrainment measures corroborates recent studies, e.g. [28, 29], and supports efforts at researching how particular methods capture potentially different types of entrainment along potentially varied dimensions.

The sequencing of the tasks within one language was randomised as we wanted to prevent a possible uncontrollable effect of the task order. This might be a limitation of the study since sequencing tasks from simple to complex might be more relevant as the easier tasks gradually prepare the subjects for the harder tasks. Such sequencing is reasonable from a methodological point of view as a tool for increasing students´ oral competence. On the other hand, [23] showed that sequencing tasks from simple to complex increases accuracy, while sequencing tasks from complex to simple increases fluency. By randomising the order of the tasks, we aimed to prevent these diverse effects on subjects´ oral production. Furthermore, our descriptive analyses did not reveal any substantive systematic patterns regarding the order of the maps in our data.

6. ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588 and in part by the Slovak Granting Agency grant VEGA 2/0165/21, Slovak Research and Development Agency grant APVV-21-0373, and the Slovak University Granting Agency.

7. REFERENCES

[1] Boland, J. 2019. Conversation Transition Times: Working Memory & Conversational Alignment. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 159-165.

[2] Lee, J., Man, G., Ferreira, V., Gruberg, N. 2019. Aligning sentence structures in dialogue: evidence from aphasia. *Language, Cognition and Neuroscience* 34(6), 720-735.

[3] Branigan, H. P., Pickering M. J., Cleland, A. A. 2000. Syntactic coordination in dialogue. *Cognition* 75, B13-B25.

[4] Levelt W. J. M., Kelter, S. 1982. Surface form and memory in question answering. *Cognitive Psychology* 14, 78-106.

[5] Giles H., Coupland N., Coupland J. 1991. Accommodation theory: Communication, context, and consequence. In: Giles H. Coupland N. Coupland J, (eds). Contexts of accommodation: Developments in applied sociolinguistics. Cambridge University Press, Cambridge, 1–68.

[6] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119, 2382-2393.

[7] Levitan, R., Hirschberg, J. B. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech 2011, 12th Annual Conference of the International Speech Communication Association*, 3081-3084

[8] Denby, J., Yurovsky, D. 2019. Parents' linguistic alignment predicts children's language development. *Cogsci*, 1627–1632.

[9] Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., Brown, A. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121(1), 41-57.

[10] Zając, M. 2013. Phonetic imitation of vowel duration in L2 speech. *Research in Language* 11(1), 19-29.

[11] Jiang, F., Kennison, S. 2022. The Impact of L2 English Learners' Belief about an Interlocutor's English Proficiency on L2 Phonetic Accommodation. *J Psycholinguist Res* 51, 217–234.

[12] Quené, H., Orr, R. 2014. Long-term convergence of speech rhythm in L1 and L2 English. *Proceedings of the International Conference on Speech Prosody*, 342-345.

[13] Lewandowski, E. M., Nygaard, L. C. 2018. Vocal alignment to native and non-native speakers of English. *Journal of the Acoustical Society of America*, 144, 620.

[14] Kim, M., Horton, W. S., Bradlow, A. R. 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance, *Lab. Phonol.* 2(1), 125–156.

[15] Bortfeld, H., Brennan, S. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes* 23, 119-147

[16] Costa, A., Pickering, M. J., Sorace, A. 2008. Alignment in second language dialogue. *Language and Cognitive Processes* 23(4), 528-556.

[17] Abel, J., Babel, M. 2017. Cognitive Load Reduces Perceived Linguistic Convergence Between Dyads. *Language and Speech* 60(3), 479-502.

[18] Wynn, C. J., Borrie, S. A. 2022. Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics* 94, 101173.

[19] Levitan, R., Beňuš, Š., Gravano, A., Hirschberg, J. 2015. Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. Proceedings of SIGDIAL, 325-334.

[20] Menshikova, A., Kocharov, D., Kachkovskaia, T. 2020. Phonetic Entrainment in Cooperative Dialogues: A Case of Russian. Proc. Interspeech 2020, 4148-4152.

[21] Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences* 36(4), 329-347.

[22] Ellis, R., Skehan, P., Li, S., Shintani, N., Lambert, C. 2020. Task-Based Language Teaching: Theory and Practice. Cambridge University Press.

[23] Robinson, P. 2001. Task complexity, task difficulty, and task production: Exploring Interaction in a Componential framework. *Applied Linguistics* 22, 27-57.

[24] Robinson, P. 2005. Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching* 43, 1-32.

[25] De Looze, C., Rauzy, S. 2009. Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. In *Interspeech 2009, 10th Annual Conference of the International Speech Communication Association.*

[26] Lehnert-LeHouillier, H., Terrazas, S., Sandoval, S. 2020. Prosodic entrainment in conversations of verbal children and teens on the autism spectrum. *Frontiers in Psychology*, 2718.

[27] Kruyt, J., de Jong, D., D'Ausilio, A., Beňuš, Š. Submitted. Replicability in prosodic entrainment research: a methodological comparison.

[28] Weise, A., Levitan, R. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (Short Papers), pp. 297-302.

[29] Weise, A., Levitan, S. I., Hirschberg, J., Levitan, R. 2019. Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication* 115, 78-87.

Perceptual Evaluation of Natural and Synthesized Prosody in Chinese Learners' English Production

Ying Chen, Yan Li, Wentao Xiao, Li Liu

School of Foreign Studies, Nanjing University of Science and Technology
ychen@njust.edu.cn

ABSTRACT

The prosody of English production by Mandarin Chinese learners was modeled and synthesized in terms of learners' L2 experience and compared with the production of native speakers of English. Both original speech and synthesized speech were perceptually evaluated by another group of native speakers of English to identify the focus status and rate the speech naturalness. The results reveal that natural speech was recognized and rated better than synthesized speech, more-experienced learners' speech better than less-experienced learners' speech, and narrow-focus sentences better than neutral-focus sentences. Among narrow foci, initial focus and medial focus were identified more accurately than final focus. Lexical stress of focused words showed a minor effect and interacted with focus status and speaker group. The overall results of the two perceptual evaluation tasks suggest a positive correlation between the accuracy of focus status identification and the speech naturalness rating in the English prosody produced by Chinese learners.

Keywords: L2 English, prosodic focus, synthesized prosody, speech naturalness, perceptual evaluation

1. INTRODUCTION

Focus refers to the emphasized word(s) in a sentence. Except for syntactic control, focus can be encoded phonetically with an increase of duration, pitch (F0) and intensity on the focused words, known as in-focus expansion, and a decrease of F0 and intensity, known as post-focus compression (PFC). Studies on bilingual production of prosodic focus have indicated that learners use duration and intensity more than F0 and in-focus expansion more than PFC to code focus [3, 4]. Even though Mandarin and English share the prosodic pattern with in-focus expansion of F0, intensity and duration and PFC of F0 and intensity, the L2 Mandarin production of L1 English learners and the L2 English production of L1 Mandarin learners reveal that learners have more difficulties in using F0 to code focus in L2, especially PFC of F0 [5].

F0, as the most difficult acoustical parameter for L2 learners to code focus, requires a theoretical model to explain its complexity in the prosodic realization of a sentence. The Parallel Encoding and Target Approximation (PENTA) model [6] is plausibly applicable to this. According to PENTA, target approximation parameters, including local pitch target, local pitch range, articulatory strength and duration, encode communicative functions in parallel and are sequentially and asymptotically realized by the surface F0. A quantitative target approximation (qTA) model [8] was proposed to use three model parameters to control the F0 trajectory of each syllable—target slope (m) and target height (b) specify the form of the pitch target and the rate or strength of target approximation (λ) indicates how rapidly a pitch target is approached. The equation of surface F0 is expressed as below:

$$f_0(t) = (mt + b) + (c_1 + c_2t + c_3t^2)\, e^{-\lambda t} \quad (1)$$

The transient coefficients, c_1, c_2, and c_3, are calculated based on the initial F0 dynamic state and the pitch target of the specified syllable:

$$c_1 = f_0(0) - b \quad (2)$$

$$c_2 = f'_0(0) + c_1\lambda - m \quad (3)$$

$$c_3 = (f''_0(0) + 2c_2\lambda - c_1\lambda^2)/2 \quad (4)$$

After a series revision of qTA model, a research tool PENTAtrainer2 [9], as a data-driven system, was developed to learn, model and synthesize the prosody of natural speech on the basis of hierarchically layered functional annotations.

The current study adopts PENTAtrainer2 to model and synthesize English sentences with prosodic focus produced by Chinese learners in terms of their L2 English experience. Native speakers of American English were recruited for a perception experiment to evaluate Chinese learners' original speech and group-synthesized speech in order to explore three research questions: (1) Can learners' acoustic realization of focus be recognized as well as native speakers' production by native listeners? (2) Can native listeners recognize focus in

the synthesized speech modelled by speakers' language experience? (3) How natural do native listeners find learners' original speech and its synthesized version compared with native speakers' production?

2. METHODS

2.1. The corpora and synthesis

Three corpora with 300 sentences in each produced respectively by native speakers of American English (AE), senior Chinese students (SC) and freshman Chinese students (FC), five females and five males in each group as presented in [12], were analysed and modelled. Table 1 shows the target sentences with their prompt questions.

Table 1: Prompt questions and answers.

Neutral focus	Question	What's the news?
	Answer	See initial, medial, final focus sentences below.
Initial focus	Question	Who may marry Ray?
	Answer	LEE/ NIna/ MElanie/ MaRIE/ RaMOna may marry Ray.
Medial focus	Question	What may Lee do to Norman?
	Answer	Lee may leave/ MArry/ NOminate/ reMIND/ reMEMber Norman.
Final focus	Question	Who may Ray marry?
	Answer	Ray may marry Lee/ NIna/ MElanie/ MaRIE/ RaMOna.

The natural speech productions were acoustically analysed by ProsodyPro [10], a Praat script for systematic prosody analysis. Syllable boundary labels were exported to PENTAtrainer2 [9] for annotation. Three functional layers were annotated: word stress (*A*, *Ab*, *aB*, *Abc*, and *aBc*), syllable position (word-final, sentence-final, non-final and monosyllabic) and focus condition (pre-focus, in-focus and post-focus) (cf. [8, 9]). The acoustic modelling was done by normalizing F0 heights across all speakers in each group. The many-to-one synthesis was done by imposing the acoustic model of the ten members in the same speaker group on each individual speaker's original speech.

2.2. The perceptual experiment

The mean standard deviation was calculated for the F0 values of all ten evenly spaced points in each syllable produced by all the 10 speakers in each group. The natural and synthesized speech of one male speaker and one female speaker, who displayed the median standard deviations in each group, was selected as the stimuli for the perceptual experiment. Thus 360 sentences (30 sentences * 2 speech types * 3 speaker groups * 2 speaker genders) were run and randomized by Praat ExperimentMFC. A forced choice identification task requested 24 native speakers of American English to choose the "emphasized" word (S, V and O respectively for subject, verb and object) or to judge that no word (N) was "emphasized" in the sentence. Meanwhile, the listeners were also required to rate the naturalness of the speech with a 1-5 scale (1 for poor and 5 for good). The experiment was conducted in a sound-attenuate booth, using a Lenovo desktop computer and a Sony MDR7506 dynamic headphone.

3. RESULTS

3.1. Focus status identification

Mixed-effects logistic regression model was applied to predict the response of focus status for the 360 sentences taking listener (n = 24) as a random factor, and speaker group [three levels: (1) AE, (2) SC, (3) FC], speech type [two levels: (1) natural, (2) synthesized), focus status [four levels: (1) neutral, (2) initial, (3) medial, (4) final] and word stress [five levels: (1) *A*, (2) *Ab*, (3) *aB*, (4) *Abc*, (5) *aBc*] as fixed factors. The first level of each factor was set as the reference level in the regression analysis. Results are shown in Table 2.

Table 2: Mixed-effects logistic regression results of focus status identification.

	β	***SE***	***z***	***p***
(Intercept)	-0.096	0.113	-0.849	0.396
SC group	0.347	0.078	4.446	0.000***
FC group	-0.283	0.072	-3.918	0.000***
Syn. speech	-0.117	0.052	-2.265	0.024*
Initial focus	1.852	0.077	24.012	0.000***
Medial focus	1.524	0.071	21.263	0.000***
Final focus	0.643	0.064	10.077	0.000***
Ab stress	-0.037	0.075	-0.491	0.624
aB stress	-0.189	0.075	-2.516	0.012*
Abc stress	-0.267	0.075	-3.559	0.000***
aBc stress	-0.125	0.075	-1.656	0.098

Surprisingly, the accuracy of focus status identification for the SC group was higher than that for the AE group; unsurprisingly, the accuracy of focus status identification for the FC group was lower than that for the AE group.

Results of the focus status identification task, by speaker group, are illustrated in conditional inference trees produced using the "party" package in R (version 3.5.0) [11] in Figures 1-3. The numbers on the branches correspond to the numbers of the levels in the fixed factors as in the above-

mentioned text. Bars in Figures 1-3 represent the percent accuracy of focus identification. The more robust the factor, the higher position in the tree branches. Therefore, focus type was the most significant factor in the results of focus status identification across speaker group.

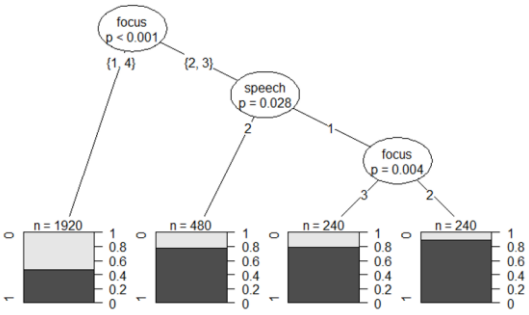


Figure 1: Focus status identification for the AE group.

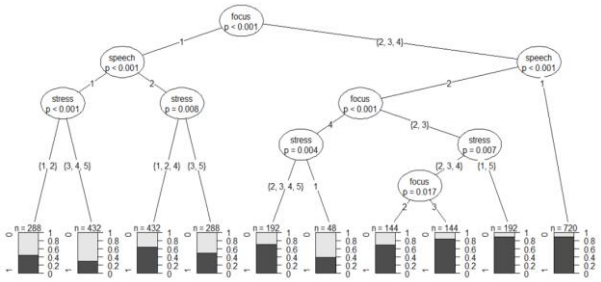


Figure 2: Focus status identification for the SC group.

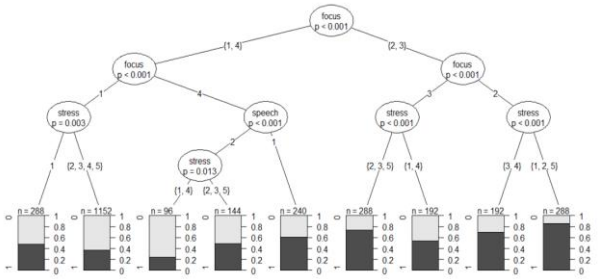


Figure 3: Focus status identification for the FC group.

3.2. Speech naturalness rating

Speech naturalness rating scores in 1-5 scales were transformed into *z* scores in terms of sentence in order to normalize subjects' individual differences. Mixed-effects linear regression model was employed to analyse native English speakers' rating to the learners' original speech and the synthesized speech based on their acoustical models by speaker group. The *z* score of the ratings was set as dependent variable. The random factor and the fixed factors in the independent variables were the same as in the logistic regression analysis for the focus status identification. The first level of each factor was again set as the reference level. The results of mixed-effects linear regression analysis are displayed in Table 3 (df = 26.67, 8606).

Table 3: Mixed-effects linear regression results of speech naturalness rating.

	β	*SE*	*t*	*p*
(Intercept)	0.702	0.089	7.877	0.000***
SC group	-0.641	0.021	-31.069	0.000***
FC group	-0.915	0.021	-44.379	0.000***
Syn. speech	-0.513	0.017	-30.435	0.000***
Initial focus	0.137	0.024	5.733	0.000***
Medial focus	0.056	0.024	2.371	0.018*
Final focus	0.091	0.024	3.812	0.000***
Ab stress	0.051	0.027	1.926	0.054
aB stress	0.081	0.027	3.051	0.002**
Abc stress	-0.092	0.027	-3.443	0.001**
aBc stress	0.086	0.027	3.239	0.001**

Since both SC group and FC group showed lower speech naturalness than the AE group, another mixed-effects linear regression was conducted taking SC group as the reference level and found that the focus status identification result of the FC group was significantly different from that of the SC group [β = -0.35, SE = 0.026, t = -13.31, $p < 0.001$].

The results of speech naturalness ratings are illustrated in Figure 4. The conditional inference tree shows that speaker group was the most significant factor and followed by speech type.

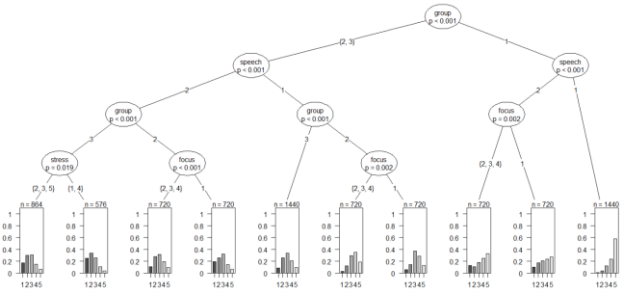


Figure 4: Speech naturalness rating. X- and Y-axes respectively represent the rating scale and percent.

4. FINDINGS AND DISCUSSION

The statistical results of focus status identification indicate that the production of the SC group was recognized better than that of the AE group and the AE group better than the FC group (see Table 2). The overall results answer the first research question that learners' focus production can be recognized by native listeners and more experienced L2 learners' production was recognized more accurately. Nonetheless, the reasons for the surprising result that the SC learners' focus production was recognized more accurately than the AE native speakers' production can be analyzed as follows. The low focus-identification accuracy for the AE group's production mainly lies in the neutral and final foci (see Figure 1). This could be due to

the fact that the boundary tone is located in the last content word of an intonation unit with either neutral focus or final focus [13, 14] and thus final focus did not demonstrate more robust prosodic prominence than neutral focus on the last tonic word in an intonation unit in English. The AE group, as native speakers, might have not intentionally produced final focus more salient than neutral focus, resulting in native listeners' perceptual confusion between final focus and neutral focus. The SC group, as advanced learners, might have learnt to specially emphasize on final focus, made it more prominent than neutral focus, which was at least reflected in duration in [12], and thus final focus was recognized more easily and accurately than neutral focus in their English production (see Figure 2).

The results of identifying focus statuses that were produced by the AE group also indicate that focus status is the most robust factor and followed by speech type (see Figure 1). The identification accuracy of initial and medial foci was significantly higher than that of neutral and final foci. This may be attributed to that initial and medial foci are auditorily more prominent than final and neutral foci. The acoustic analysis showed that both initial and medial foci were produced concomitantly with post-focus compress (PFC) of F0 and intensity in the original speech [12] and PFC was found to facilitate identifying focus status [2, 15]. Regarding the effect of speech type, original speech was identified more accurately than synthesized speech, suggesting no synthesis tool is perfect even though this corpus was native speech. Initial focus was identified more accurately than medial focus in the original speech, suggesting initial focus is more identifiable than medial focus. This could be again due to the effect of more words bearing PFC in the sentence with initial focus than in the sentence with medial focus and PFC works as an acoustic cue for recognizing focused words.

Similar to those of the AE group, the results of identifying focus statuses that were produced by the SC group indicate that focus is the most robust factor and followed by speech (see Figure 2). The identification accuracy of narrow focus, i.e., initial, medial and final foci, was significantly higher than that of broad focus, i.e., neutral focus. This result reflects that the SC group produced more robust final focus than neutral focus compared with the production of the AE group. Neutral focus in original speech was identified less accurately than in synthesized speech whereas narrow foci in synthesized speech less accurately than in original speech. In the synthesized speech with narrow foci, initial focus and medial focus were identified more accurately than final focus. These findings suggest that PENTAtrainer2 may have moderated the surface F0 of L2 learners' speech with certain focus statuses so that they became more identifiable. Lexical stress of the focused words is also a significant factor in focus status identification. However, no clear pattern was found across stress type and its effect interacted with that of speech type and focus status.

In the results of identifying focus status produced by the FC group, focus was the most robust factor again (see Figure 3). Overall, neutral and final foci were clustered together, significantly different from the cluster of initial and medial foci. Although the accuracy was lower in the results of the FC group than that of the AE group, initial and medial foci were again identified more accurately then final and neural foci. However, final focus was identified better than neural focus and initial focus better than medial focus. These results reveal that even less-experienced learners produced narrow focus more prominent than neutral focus and PFC might have also played a role in the identification of initial focus. Speech type only affected the identification of final focus produced by the FC group. Lexical stress of focused words was also a significant factor in focus status identification but again did not demonstrate clear patterns.

The above analyses answer the second research question that native listeners were able to recognize focus in the speech synthesized based on speaker-group modelling but the accuracy was not as high as that in the original speech though the effect of speech type interacted with the effects of focus status and lexical stress of the focused words.

Figure 4 illustrates that speaker group is the most salient factor of the speech naturalness rating. The AE group's production was rated higher than the production of the SC and FC groups and the SC group's production higher than the FC group's production in both original speech and synthesized speech. These results answer the third research question of the comparison between learners' and native speakers' production. The original speech was generally rated more naturally than the synthesized speech, again suggesting no synthesis tool is perfect. Sentences with initial, medial and final foci were rated more naturally than those with neutral focus even though the audio tokens were presented to the raters with no discourse context. Lexical stress was a minor effective factor in the FC group's synthesized speech.

5. ACKNOWLEDGMENT

This work was supported by the National Social Science Foundation of China (approval number 19BYY0430).

6. REFERENCES

[1] W. Wu and Xu. Y, "Prosodic focus in Hong Kong Cantonese without post-focus compression," *Proc Speech Prosody 100040*, pp. 1–4, 2010.

[2] Y. Xu, S. Chen, and B. Wang, "Prosodic focus with and without post-focus compression: A typological divide within the same language family?" *The Linguistic Review*, vol. 29, no. 1, pp. 131–147, 2012.

[3] Y. Chen, Y. Xu, and S. Guion-Anderson, "Prosodic realization of
focus in bilingual production of Southern Min and Mandarin," *Phonetica*, vol. 71, no. 4, pp. 249–270, 2014.

[4] W. Wu and L. Chung, "Post-focus compression in English Cantonese bilingual speakers," *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, pp. 148–151, 2011.

[5] Y. Chen, "The hierarchies of phonetic realization of focus in second language speech," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 2160–2160, 2016.

[6] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[7] Y. Xu and Q. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319–337, 2001.

[8] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, pp. 405–424, 2009.

[9] Y. Xu and S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[10] Y. Xu, "ProsodyPro.praat," Version 5.5.2. http://www.phon.ucl. ac.uk/ home/yi/ProsodyPro/, 2014.

[11] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 307, pp. 1–11, 2008.

[12] Y. Chen, "Post-focus compression in English by Mandarin learners," *Proceedings of the 18th International Congress of Phonetic Sciences*, vol. 287, pp. 1–5, 2015.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004.

[14] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2006.

[15] Y. Xu, C. Xu, and X. Sun, "On the temporal domain of focus," *Proceedings of International Conference on Speech Prosody*, pp. 81–84, 2004.

[16] Y. Xu, A. Lee, S. Prom-on, and F. Liu, "Explaining the PENTA model: A reply to Arvaniti and Ladd," *Phonology* 32: 505-535, 2015.

[17] Y. Xu, S. Prom-on, and F. Liu, The PENTA model: Concepts, use and implications. In *Prosodic Theory and Practice*. S. Shattuck-Hufnagel and J. Barnes (eds.). Cambridge: The MIT Press. pp. 377-407, 2022.

Macro- and Micro-rhythm in L2 English: Exploration and Refinement of Measures

Ortega-Llebaria, M, Silva Jr, L, Nagao, J.
University of Pittsburgh, State University of Paraiba, Gifu Shotoku Gakuen University
mao61@pitt.edu, leonidas.silvajr@servidor.uepb.edu.br, junnagao@gifu.shotoku.ac.jp

ABSTRACT

This paper constitutes a pioneering attempt to explore the pitch-based macro-rhythm measures proposed by Jun [4], namely macroR_Var and macroR_Freq, in a large L2-speech database and relate them to micro-rhythm measures based on vowel and consonant interval durations. It also proposes a refinement of Jun's macroR_Freq. Eleven 2.5-minute TED talks in L1 English and their corresponding imitations by 11 Japanese English-learners were analyzed using semi-automated labeling and scripts. After intensive practice, Japanese speakers successfully imitated pitch contours, reaching target-like macroR_Var scores while failing at imitating macroR_Freq scores and the duration variation of vowel and consonant intervals as illustrated by several Varco and nPVI micro-rhythm measures. Relevantly, the newly proposed macroR_Freq not only was sensitive to L1-L2 speech differences but was also modulated by several micro-rhythm measures, illustrating for the first-time the nature of the micro-rhythm and macro-rhythm interaction in the Japanese L1-English L2 interface. Possible cognitive and linguistic based explanations are discussed.

Keywords:
L2 rhythm, Macro-rhythm, L2 acquisition, L2 Prosody, EFL

1. INTRODUCTION

Research in L2 rhythm has been largely based on duration-based micro-rhythm measures (microR), e.g., VarcoV, VarcoC, nPVI_V [1], [2], which compute duration variations of the vowel and consonant intervals within an utterance. These measures capture cross-linguistic differences in syllabic structure and the expression of word prominence, which differ between traditional rhythm-classes (for a critical review, see [3] and [4]). For example, stress-timed languages like English –where stressed syllables have long full vowels, unstressed syllables have short reduced vowels, and syllables have complex onsets and codas – tend to have more duration variation of vowel and consonant intervals than syllable-timed languages like Spanish and mora-timed languages like Japanese – which have simpler syllable structures and no vowel reduction. As a result, English obtains lower %V, and higher VarcoV and VarcoC scores than Japanese and Spanish [5].

A newer set of rhythm measures based on the repetition of F0 patterns was proposed by Jun [6]. These measures, referred to as macro rhythm (macroR), capture patterns of F0 repetition above the word level. For example, macroR computes the distance between F0 peaks and F0 valleys within an utterance. So far, macro-rhythm measures have been tested across different L1 and speaking styles, e.g., [7], [8], but, as far as we know, not yet in L2.

While there is a body of research on the imitation of L1 sentence prosody (cf. [17], [18] among others), to our knowledge, this paper is the first attempt to explore macroR measures in a large L2-speech database of Japanese learners of English. The L1 Japanese – L2 English context is especially interesting as the two languages largely differ in their microR and macroR strengths. Japanese[1] has a weaker microR than English since not all content words in Japanese have pitch accents while all content words in English have stress which is cued not only by pitch but also by duration. With regards to macroR, it is predicted to be stronger in Japanese than English [6] because English has a larger inventory of pitch-accents than Japanese increasing variation. In contrast, Accentual Phrases in Japanese have a fixed hat-shaped pitch-contour that repeats along the utterance. Thus, Japanese-accented English is expected to display a weaker microR and a stronger macroR than L1 English.

To analyze this large amount of L2-speech data we used webMAUS forced-aligner [9] and two scripts for Praat [10], [11] that automatically post-processed the alignment into new phonetic units, corrected F0 tracking errors, and computed macroR and microR measures, and pauses. Moreover, a new macroR frequency index, based on the ratio of the F0 peaks and F0 variance (macroR_Freq_F0var), is proposed (see 2.3.1.).

The research questions we are addressing are the following:

[1] Are microR measures sensitive to Japanese accented English as to convey the expected weaker microR in comparison to L1 English?

[2] Are macroR measures sensitive to Japanese accented English as to convey the expected stronger macroR in comprison to L1 English?

[3] Is there any relation between macroR measures and microR measures in Japanese-accented English?

2. METHODOLOGY

2.1. Participants and Task.

Eleven Japanese students of English enrolled at Gifu Shotoku Gakuen University, Japan, participated in this study. They were 18-21 years old, 4 males and 7 females. They participated in English Recitation Contest where they recite a 2.5-minute TED talk or English speech of their choice in front of a team of judges and a large audience. They rehearsed for an average of 2 months with feedback from their teachers and peers to faithfully reproduce the content and emotion of the original talk. Altogether, the 11 imitations along with the corresponding original TED talks resulted in a 55-minute speech sample (31-minute L2 English, and 26-minute L1 English).

2.2. Data processing

After transcribing the original TED talks, the students' imitations, and editing recordings for noise (e.g., cutting off applauses), transcriptions and sound files were input into the aligner to obtain TextGrids which segmented words into phonemes. When needed, TextGrid labels were manually adjusted. Following protocols in [19], each TED talk was divided into 10 to 15 chunks, i.e., utterances between pauses that conveyed a coherent meaning and ended with a final (non-continuant) boundary tone. These TextGrids were input into the scripts VVUnitAligner and SpeechRhythmExtractor which output for each chunk the measures described in the section below.

2.3. Measures

2.3.1. Micro-Rhythm Measures

Micro-Rhythm (microR) in this paper refers to the measures based on the duration of the consonant and the vowel intervals within a chunk, e.g., percV calculates the proportion of vowel intervals within each chunk. As explained in the Introduction, these duration-based measures capture cross linguistic differences in syllable structure and word prominence, e.g., in languages with simpler syllabic structures like Japanese, percV scores around 50%, while in languages with complex onset and codas like English values below 50% are expected. The variation coefficient of vowels and consonants (VarcoV and VarcoC respectively) are calculated by the ratio of the standard deviation of vowel or consonant intervals and the interval mean value. The higher the Varco, the more variable the duration of the analyzed intervals [2]. This variation tends to be higher in languages like English, where lexical stress is cued by duration, in comparison to languages like Japanese, where lexical pitch-accents are not cued by duration. The Normalized Pairwise Variability Index for vowels and consonants (nPVI-V and nPVI-C respectively) is computed by the ratio of the differences between consecutive vowel or consonant intervals and the sum of the duration of the respective intervals [1].

2.3.2. Macro-Rhythm Measures

Jun [6] defined Macro-Rhythm (macroR) as the phrase-medial tonal rhythm conveyed by the repetition of F0 patterns. The more regular this repetition is, the stronger macroR becomes. She proposed two macroR measures to capture it, namely, macroR_Var as in (1) and macroR_Freq as in (2).

$$(1)\quad MacR\_Var = rSD + fSD + pSD + vSD$$

$$(2)\quad MacroR\_Freq = \frac{number\ of\ F0\ peaks}{number\ of\ Prosodic\ Words}$$

MacroR_Var captures the F0 contour of a chunk. Larger scores in (1) come from larger variations of rising (rSD) and falling (fSD) F0 excursions as well as larger duration variations between F0 peak-to-peak (pSD) and valley-to-valley (vSD). Thus a higher score in MacroR_Var indexes more variation, and therefore, weaker macroR. MacroR_Freq captures the synchronization of the F0 peaks in an utterance with prosodic words. A score close to one in (2) comes from having one F0 peak per each Prosodic Word conveying a strong macroR. Thus, while both measures capture macroR, macroR_Var focuses on the chunk F0 contour without reference to the segmental content that carries it, whereas macroR_Freq refers to the synchronization of F0 peaks with prosodic words.

The SpeechRhythmExtractor script [11] used in this paper implements macroR_Freq as a function of the F0 variance within a chunk as in (3) below. While keeping the synchronization interpretation, i.e., larger F0 variation requires a larger number of F0 peaks to convey strong macroR, this implementation becomes fully-automated as it avoids the need to define and count prosodic words.

$$(3)\quad MacroR\_Freq\_f0Var = \frac{number\ of\ F0\ peaks}{F0\ variance}$$

[6] defines macro-rhythm as being the phrase-medial tonal rhythm, and [7] points out that its strength is determined by the number of F0 alternations between peaks and valleys within a phrase. The use of the F0 variance can bring forward a closer look on how the alternations of the peaks and valleys are represented

within a phrase, and wether there is (some) level of regularity for the F0 alternations. By these means, F0 variance showed consistent differences between L2 speech rhythm studies, such as [19].

2.3.3. Speech Rate Measures

Because microR and macroR are affected by speech rate, speech rate is measured to justify the use of normalization measures. Speech rate is computed as the number of syllables per second, and articulation rate as the number of syllables per second, discarding silences. Mean duration and standard deviation of pauses (Pause SD), as well as pause rate, are also computed.

2.4. Statistical Analysis

To assess the sensitivity of the above microR and macroR measures to Japanese-accented English, linear mixed models were computed with R (lme4 package) with each measure as the dependent variable, language (L1 English, L2 English) as the predictor, and participants, TED talk, and chunk as the random factors. To assess the effect of microR on macroR, mixed models were run in R with macroR measures as the dependent variable, microR measures and language as the fixed factors, and participant and TED talk as the random factors.

3. RESULTS

3.1. Pauses, Speech, and Articulation Rates

On the one hand, Japanese speakers closely imitated pause location and duration from the target TED talks (β = -2.78, t = -0.116, p = .90). On the other hand, Japanese speakers inserted additional pauses which led to a significantly higher quantity of pauses (β = 56.15, t = -2.07, p = .03), and, consequently, consistent differences in pause rate between groups (β = 0.01, t = 5.08, p < .001). The speech rate of the imitations was significantly slower than that of the original TED talks (β = -0.50, t = -5.11, p < .001). Moreover, the articulation rate in the imitations was also significantly slower (β = -0.44, t = -6.33, p < .001) as shown in Table 1.

Descriptors	English	Pause SD	Pause rate	Speech rate	Articulation rate
Mean	L1	4.71	0.41	3.83	4.86
	L2	4.12	0.49	3.35	4.46
Median	L1	3.66	0.41	3.79	4.84
	L2	4.09	0.50	3.38	4.46
SD	L1	3.46	0.16	1.05	0.89
	L2	1.96	0.13	0.86	0.64

Table 1. Standard deviation of pauses (Pause SD), Pause rate, Speech Rate, and Articulation Rate in the original talks (L1) and their imitations (L2).

3.2. MacroR and MicroR Measures

Given that L2 speech- and articulation-rate were significantly slower than that of L1, normalized measures, e.g., Varco, nPVI, were analyzed. With regards to macroR, no statistically significant differences between L1 and L2 were obtained for pitch rises (β = 0.03, t = 0.09, p = .92), pitch falls (β = 0.40 t = 1.45 p = 0.14), F0 peak-to-peak distances (β = 0.45 t = 0.34, p = .73), and F0 valley-to-valley distances (β = -0.12, t = -0.07, p = .94) indicating that as a group, Japanese speakers accurately imitated these F0 targets. Consequently, macroR_Var showed no significant differences between L1 and L2 (β = 0.59, t = 0.23, p = .81), as well as macroR_Freq (β = 0.02, t = 0.88, p = .38). However, L2 speakers' macroR_Freq_F0var scores were significantly higher than L1 speakers' (β = 0.06, t = 4.31 p < .001) indicating that Japanese speakers obtained less F0 variability and stronger macroR than those of the original TED samples. The macroR measures from equations macroR_Var, macroR_Freq, and macroR_Freq_F0var are shown in Table 2.

Descriptors	English	MacroR_Var	MacroR_Freq	MacroR_Freq_F0var
Mean	L1	53.50	0.33	0.34
	L2	54.90	0.35	0.40
Median	L1	47.50	0.27	0.33
	L2	52.50	0.29	0.37
SD	L1	35.90	0.29	0.11
	L2	28.60	0.24	0.17

Table 2. MacroR Measures in the original talks (L1) and their imitations (L2).

In contrast to macroR, all microR measures (see Table 3) reached significant differences between groups, showing that Japanese speakers were not as successful at perceiving (and then, imitating) the vowel- and consonant-duration intervals of the original talks. More precisely, %V was significantly higher in the L2 imitation (β < .001, t = 2.07, p = .03) while its counterpart %C was lower (β = -0.01, t = -2.07, p = 0.03) indicating that vowel-intervals in the same chunk occupy more time in L2 than in L1. VarcoV indicates that the duration variation of vowel-intervals in a chunk is larger in the L1 original talk (β = -0.04, t = -2.85, p = .004) while VarcoC shows that for consonant-intervals, it is larger in the L2 imitations (β = 0.04, t = 2.45, p = .01). When duration variation is measured in adjacent intervals in nPVI measures, L2 obtains significantly higher scores than L1 for vowels (β = 0.07, t = 5.97, p < .001), consonants (β = 0.04, t = 4.13, p < .001), vowels or consonants (β < 0.001, t = 7.02, p < .001). This larger

L2 variation between adjacent units is likely related to the frequent inclusion of short epenthetic vowels to break consonant clusters as well as to the slower L2 articulation rate [9, 15, 16].

	English	%V	%C	VarcoV	VarcoC	nPVI-V	nPVI-C	nPVI-VC
Mean	L1	48.40	51,60	0.68	0.61	0.59	0.60	0.62
	L2	50.00	50,00	0.63	0.65	0.67	0.65	0.68
Median	L1	48.00	52,00	0.65	0.58	0.59	0.60	0.62
	L2	51.00	49,00	0.60	0.63	0.67	0.64	0.68
SD	L1	6.94	6,94	0.16	0.13	0.12	0.10	0.09
	L2	8.15	8,15	0.13	0.15	0.11	0.09	0.08

Table 3. MicroR Measures in the original talks (L1) and their imitations (L2).

3.4. MacroR and MicroR Interactions

Linear mixed models with macroR_Freq_F0var as the dependent variable, language (L2 English and L1 English) and microR variables as fixed effects, and participants and TED talks as random effects were run to explore which microR factors affected macroR. No single fixed effect reached statistical significance, only the following interactions did, namely, language with VarcoV ($\beta = 0.18$, $t = 1.97$, $p = 0.04$), with nPVI-C ($\beta = 0.35$, $t = 2.30$, $p = 0.02$), and with nPVI-VC ($\beta = 0.37$, $t = 2.14$, $p = 0.03$). Figure 1 illustrates the interaction between the nPVI-VC and language showing no correlation for L1 speakers. However, for L2 speakers, reducing nPVI scores towards L1-targets correlates with reducing macroR_Freq_F0var towards L1-targets. Similar effects are observed in the language* nPVI-C interaction. Thus, the higher the nPVI values in L2 – which indicate an effort to increase duration variability at the segmental level to reach L1 duration targets, the higher, and therefore the more regular, macroR_Freq_F0var becomes, missing to match the irregular macroR_Freq L1 index.

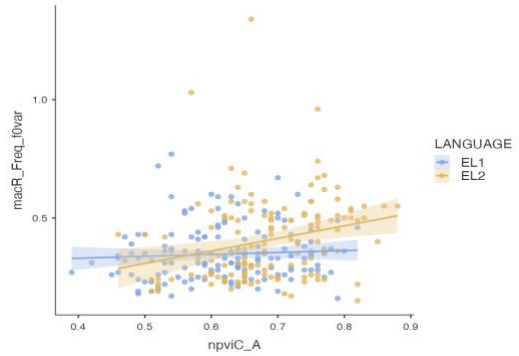


Figure 1. Effect of nPVI-VC on macroR_Freq_F0var by speakers of English as L1 or as L2

4. DISCUSSION AND CONCLUSION

Both speech rate, pauses, and microR results were coherent with previous L2 literature. L2 English imitations of L1 English TED talks were produced at slower speech and articulation rates, and were inserted more pauses than in L1 illustrating the higher cognitive demands of speech planning in L2 e.g., [12]. As for microR, the higher proportion of vowels and smaller variation of vowel-duration intervals in the L2 imitations showed that Japanese speakers transferred to English some of their L1 patterns, e.g., the simpler syllabic structure of Japanese and the lack of duration cues to word prominence. Similar results have been obtained for Japanese speakers of English in previous research, e.g., [13, 14]. The consistently higher L2 nPVI scores, which were also present in [9], were related to the insertion of short epenthetic vowels in complex consonant clusters and to the longer articulation rate in L2 which worked together to increase nPVI scores. Altogether, these microR results answer affirmatively research question 1, namely microR scores captured that microR in Japanese-accented English was not as regular or as strong as in L1 English.

MacroR results to L2 English constitute our main contribution and address research question 2. On the one hand, L1-L2 differences on macroR_Var were non-significant, showing that L2 speakers were highly successful in imitating the L1 pitch contours of the analyzed chunks. On the other hand, L2 macroR-Freq_F0var scores were significantly higher than in L1 due to Japanese speakers' production of F0 peaks that were not present in the original TED talk. These higher values indicated a stronger macroR_Freq index than in L1, answering affirmatively research question 2. Altogether, these macroR_Var and macroR_Freq scores indicate that while Japanese speakers of English produced a sentence F0 contour very similar to that of L1, synchronizing F0 peaks with smaller domains such as prosodic words was still a challenge. In other words, the de-accentuation rate of prosodic words is higher in L1 English.

Finally, research question 3, namely, whether microR interacts with macroR is answered affirmatively so that approximating the larger duration variation of microR in L1 English leads Japanese speakers of English to miss the irregular macroR_Freq of L1 English. Two possible explanations are proposed. The first explanation is based on the deployment of cognitive resources. The higher the effort to imitate vowel and consonant duration intervals, the less resources are directed to process F0 variation causing to miss the irregular macroR of L1 English [12, 15]. In other words, the more cognitive resources are deployed to achieve the strong English microR, the least resources are devoted to achieving the irregular and weak English macroR preventing Japanese speakers from reducing their initial Japanese-like strong macroR.

The second explanation is more linguistically oriented. The learning of the weaker English

macroR_Freq by Japanese speakers is contingent on learning duration cues in words. The hypothesis is that Japanese speakers' deaccentuation of English words will occur more frequently, and therefore, match the weaker English MacroR_Freq, as they use duration to express prominence. To test this hypothesis, future research needs to elucidate what exactly in the vowel interval durations (e.g., English stress, content versus function words) is conditioning macroR_Freq to Japanese learners of English.

7. REFERENCES

[1] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(1982), 515-546.

[2] Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for delta C. In Karnowski, P.& Szigeti, I. (eds), *Language and Language Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*. Frankfurt am Main, Germany: Peter Lang Publishing Group, 231-241.

[3] Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46-63.

[4] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351-373.

[5] Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), pp. 265-292.

[6] Jun, S. A. (2014). Prosodic typology: By prominence type, word prosody, and macro rhythm. *Prosodic typology II: The phonology of intonation and phrasing*, pp. 520-539.

[7] Prechtel, C. (2020). Quantifying Macro-rhythm in English and Spanish: A Comparison of Tonal Rhythm Strength. University of California, Los Angeles.

[8] Ordin, M., & Polyanskaya, L. (2015). Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *The Journal of the Acoustical Society of America*, *138*(2), pp. 533-544.

[9] Kisler, T., Reichel, U., Schiel F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*, pp. 326-347.

[10] Silva Jr., L. (2022a). VVunitAligner for webMAUS.https://github.com/leonidasjr/VVunitAlignerCode_webMAUS.

[11] Silva Jr., L., Barbosa, P.(2022b). SpeechRhythmExtractor.https://github.com/leonidasjr/SpeechRhythmCode.

[12] Järvinen, K., & Laukkanen, A-M. (2015). Vocal Loading in Speaking a Foreign Language. *Folia Phoniatrica et Logopaedica*, 67, pp. 1-7.

[13] Nagao, J., & Ortega-Llebaria, M. (2022). Micro-and macro-rhythm measures in English and Japanese as first and second languages. In *1st International Conference on Tone and Intonation (TAI). Sonderborg, Denmark*.

[14] Grenon, I. & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In H. Chan, H. Jacob, & E. Kapia, eds., Proceedings of the 32nd Boston Conference on Language Development. Somerville, MA: Cascadilla Press, pp. 155–166.

[15] Krivokapić, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weihrich, D. Pape & P. Perrier (Eds.), *Speech planning and dynamics*. (pp. 157-190). Peter Lang.

[16] Igarashi, Y. (2014). Typology of intonational phrasing in Japanese dialects. *Prosodic typology*, *2*, pp. 464-492.

[17] Cole, J., & Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate?. In *Twelfth Annual Conference of the International Speech Communication Association,* pp. 969-972.

[18] Petrone, C., D'Alessandro, D., & Falk, S. (2021). Working memory differences in prosodic imitation. *Journal of Phonetics*, *89*(3), pp. 101100. 10.1016/j.wocn.2021.101100

[19] Silva, C., & Arantes, P. (2021). Quantitative analysis of fundamental frequency in Spanish (L2) and Brazilian Portuguese (L1) evidence of learning and language attrition. *Journal of Speech Sciences*, 10, pp. e021003.

[1] Japanese in this paper refers to the Tokyo Japanese variety, i.e., the prestige Japanese variety used in the media and higher education, which is described as a [+accent], [+multiwordAP] by Igarashi in [16].

THE EDGE TONES INVENTORY IN STANDARD RUSSIAN PROSODY: ARE PHRASE ACCENTS REQUIRED?

Pavel Duryagin

Ca' Foscari University of Venice, Italy
pavel.duryagin@unive.it

ABSTRACT

This paper discusses the structure and inventory of edge tones in the intonational phonology of Standard Russian. Based on the qualitative analysis of four Russian tunes elicited in experimental conditions, we argue that the position of postnuclear pitch targets in these tonal configurations can be better modelled if the notion of phrase accent is employed. Three configurations of edge tones are discussed: H-L%, H-H% and L-L%. According to the data, Russian phrase accents tend to demonstrate boundary-seeking behaviour when a long postnuclear stretch of syllables is available. Either edge tone truncation or vowel lengthening are attested when postnuclear segmental material is lacking. The choice of tune-text negotiation strategy is presumably determined by the need to preserve distinctive pragmatic meaning (in the case of H-L%) and by the alignment within the nucleus (in the case of L-L%).

Keywords: intonation, autosegmental metrical model, phrase accent, truncation, Russian

1. INTRODUCTION

A comprehensive prosodic description of a given language within the autosegmental metrical (AM) model of intonation should define the tones associated with constituent heads (*pitch accents*) and those associated with phrasal boundaries (*boundary tones* and *phrase accents*, collectively known as *edge tones*) [1, p. 28]. Although the need to include phrase accents in the analyses of well-studied prosodic systems has been debated extensively in the literature [2]–[4], the edge tones inventory for Standard Russian remains largely underdiscussed. Previous AM-based studies of Russian have predominantly remained agnostic regarding the issue and either included the default phrase accents identical to boundary tones in their labelling systems [5], [6] or excluded phrase accents altogether due to the lack of empirical evidence for their presence in Russian [7]. One particularly explicit approach to this issue is present in Odé's proposal for labelling Russian intonation, which posits no phrase accent in Russian and the sole L% boundary tone contrasted with the boundary not marked by pitch [8]; that is, in Odé's model the nuclear pitch level is treated as a default for the postnuclear stretch.

One source of such a lack of attention to the Russian edge tones inventory is the paucity of common tunes containing non-prominence-lending postnuclear changes in pitch. In most frequent Russian tonal configurations, the level of the final nuclear tonal target of the phrase is sustained throughout the postnuclear string. Additionally, until recently, no tunes were attested in Standard Russian containing a pitch turning point aligned with the phrase-final unstressed syllable (a typologically common manifestation of boundary tones associated with the final syllable of intonation phrase; IP).

In this paper, we analyse four Russian tonal configurations contradicting the general tendency to sustain the nuclear pitch level throughout the postnuclear string. This exploratory study aims to address three main research questions:

1) Can all postnuclear pitch movements in Russian be modelled by straightforward interpolation from the nuclear level to the level of the IP-final syllable?

2) Do postnuclear turning points in Russian demonstrate boundary-seeking or stress-seeking behaviour typical of the phrase accents?

3) How do postnuclear pitch events adjust under the time pressure in the vicinity of the IP boundary?

These questions are addressed below based on the data recorded under experimental conditions.

2. DATA COLLECTION

The data presented in this paper were elicited from 12 native Russian speakers (nine female, three male, mean age 26.3). Most were exchange students or temporary visitors at Ca' Foscari University of Venice, except for one speaker who had resided in Venice for four years but lives in a Russian family and exclusively speaks Russian at home.

The elicitation procedure included reading aloud short reactions (2–3 phrases) to contextualised situations. The participants were permitted to familiarise themselves with the contexts and the stimuli before reproducing them. Four different tunes were elicited, corresponding to the pragmatic meanings presented below in sections 3.1–3.4.

In addition to the contexts, the syllabic structure of stimuli was manipulated to observe the tune-text

negotiation strategies employed by the speakers in the vicinity of the IP boundary (see Table 1). The three syllabic conditions are hereafter referred to as 6S (6 postnuclear syllables), 2S (two postnuclear syllables) and 0S (no postnuclear syllables; the syllable bearing nuclear prominence is IP-final), respectively.

Stimuli (polar question, see sect. 3.3)	Syllables
(1) *I Golovanovu uvolili?* ‘And Golovanova-Acc. they fired?’	6
(2) *I Golovanovu?* ‘And Golovanova-Acc. (they fired)?’	2
(3) *I golova?* ‘And (your) head (hurts)?’	0

Table 1: The syllabic structure of the elicited stimuli; the number of syllables following the nuclear accent is indicated in the right column. Golovanova (with the third syllable stressed) is a common Russian surname.

Four repetitions of each stimulus in each condition were elicited to ensure that the speakers produced the contours consistently. Only the first repetitions of the stimuli are discussed in this paper. A total of 144 tokens (12 speakers, four contexts and three syllabic conditions) were segmented and labelled manually in *Praat* [9]; all contours, both original and time-normalised in [10], were inspected visually to address the research questions.

3. RESULTS AND DISCUSSION

3.1 The qualitative implicature tonal configuration

The first tonal configuration discussed here is that of a pragmatically loaded statement:

(1) *Ne tol’ko melkikh soshek uvol’nyayut. I Golovanovu uvolili!* ‘It’s not that they only fire small fry. Golovanova was fired, too’.

When produced with a complex rising-falling-rising contour, this phrase conveys a qualitative implicature regarding the person named Golovanova. The listener is not merely informed of the fact that this individual was fired but also that they are not expected to be treated this way. This uncommon tune was mentioned in [11] and first studied experimentally in [12]. An example of its typical phonetic realisation is given in Fig. 1: under the 6S condition (*left*), when a long string of postnuclear segments is available, the nuclear pitch accent (preliminarily labelled here as H+L*) is followed by a high plateau starting at the right boundary of the prominent word. The plateau spans the prestressed string of syllables in the IP’s final word and is followed by an abrupt fall aligned with the IP’s last metrically strong syllable.

One interpretation of this non-prominent falling pitch movement could be a separate postnuclear pitch accent assigned to the phrase-final verb *uvolili*. However, the data elicited under the 2S and 0S conditions indicates that the postnuclear plateau and the fall are preserved even when the IP's final word or syllable bears the H+L* nuclear accent (Fig. 1, *centre* and *right*). We suggest that the observed postnuclear pitch movements can be analysed as a combination of a boundary tone L% preceded by a phrase accent H-. Instead of being associated with singular targets, both edge tones show signs of secondary association. In 6S, the L% is spread/copied from the phrase boundary in all post-stressed syllables of the IP-final word. Conversely, the H- is associated with the last metrically strong syllable of the IP and spread/copied to the preceding syllables up to the right boundary of the nuclear-accented word. When nuclear prominence is assigned to the final word or syllable in the IP (2S and 0S, respectively), both edge tones are aligned with the IP-final syllable. Importantly, in the 0S condition, the edge tones cannot be truncated because, in this case, the contour becomes homophonous to the regular narrow focus statement H+L* (L-L%) and thus loses the distinctive qualitative implicature. The need to fit three tones in a single syllable is met by vowel lengthening.

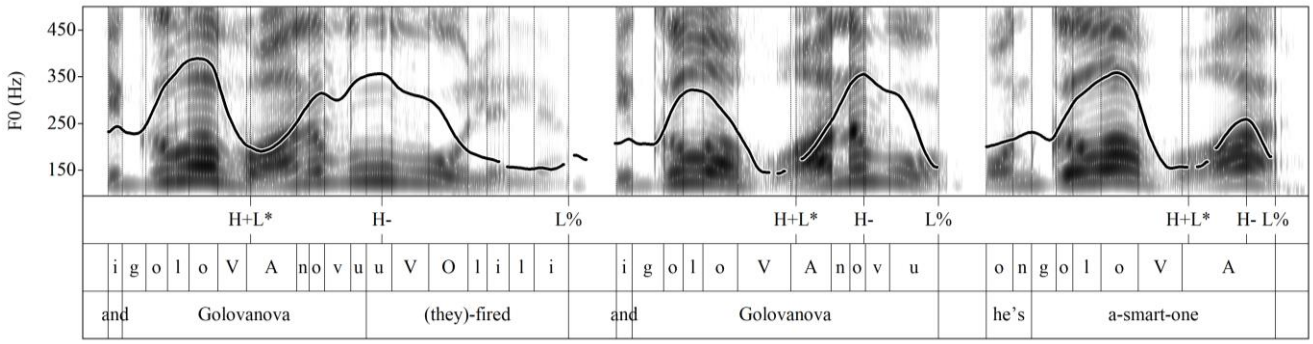


Figure 1: Pitch contours of the qualitative implicature tune produced by speaker 9 in three conditions: 6S (*left*), 2S (*centre*) and 0S (*right*). The tokens were recorded separately and further concatenated in *Praat* for ease of presentation. The segments within the stressed syllables are shown in uppercase.

3.2 The comparative question tonal configuration

A logical comparison to the tune presented in 3.1 is the comparative question contour, labelled IK-4 (*fourth intonational construction*) in Bryzgunova's traditional description of Russian intonation [13]:

(2) *Tak, ty otdala svoy klyuch Ivanovu. A Golovanova komu dala?* 'So, you gave your key to Ivanov. And Golovanova to whom gave [it]?'

Within the AM framework, the IK-4 nuclear accent is usually analysed as monotonal L*; the optional high target preceding the L* is interpreted as an independent prenuclear H* since it can be aligned several syllables before the nucleus. The comparative question contour is the only Russian tune that has been specifically studied concerning postnuclear pitch movements. Knyazev and Savelieva [14] reported a large variability of apparently different postnuclear movements in IK-4, including a gradual rise in pitch to the IP-final syllable and a high plateau reached a few syllables earlier. It should be noted, however, that their study only controlled for the total number of postnuclear syllables, while the effects arising from the presence of a word boundary in the postnuclear string were not investigated separately.

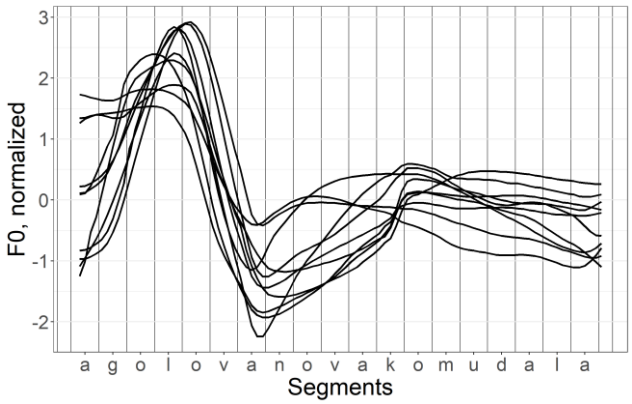


Figure 2: Time-normalised contours of comparative questions produced under the 6S condition ('*And Golovanova to whom gave (it)?*').

Our data reveals that the subjects produced long postnuclear strings more consistently than the speakers in [14]. As the time-normalised contours in Fig. 2 demonstrate, the postnuclear elbow was most frequently aligned with the voiced onset of the first postnuclear constituent *komu*, 'to whom', followed by a high plateau. Rarely the pitch maximum was reached immediately after the syllable bearing the NA, and, importantly, no instances of gradual pitch rise to the IP-final syllable were attested. That is, the high tonal target at the beginning of the postnuclear plateau demonstrates boundary-seeking behaviour similar to the behaviour of H- reported in 3.1. By analogy with 3.1, we suggest that this edge configuration can be analysed as a combination of H- and H%, with the same patterns of secondary association for both phrasal and boundary tones as those proposed for H- and L% in 3.1.

Under the 2S condition, the pitch targets for both H- and H% are aligned with the IP-final syllable. As for the 0S, the vowel lengthening strategy was employed by most speakers to fit the pitch accent L* and the edge configuration H-H% within the phrase-final syllable, a known feature of IK-4 [13].

3.3 The polar question tonal configuration

In 3.1 and 3.2, two tonal configurations containing postnuclear high targets were discussed. However, the prosodic inventory of Russian includes a frequent tune with the opposite type of pitch movement: a nuclear accent culminating at a high level and followed by a postnuclear low target. This tonal configuration is the default tune marking yes/no questions in Russian. After the nuclear L*+H, the low tone is reached early in the postnuclear syllable string and is sustained until the end of the IP.

This pattern bears similarity to the English H* L-L% that originally served as a primary argument supporting the necessity of an additional level of edge tones in the prosodic phonology of this language [15]. Likewise, in Russian polar questions, postnuclear pitch movements cannot be modelled by mere interpolation from the nuclear H to the boundary L% (see Fig. 3, *left*). Before quantitative results are available, we limit ourselves to pointing out that if it is found that the elbows observed in long postnuclear syllable strings do demonstrate a tendency to align with the right edge of the nuclear word, this can be taken as a further argument in favour of analysing edge configuration in Russian yes/no questions as containing two separate elements: L- and L%.

3.4 The mirative corrective focus tonal configuration

The final part of this section presents one more postnuclear falling movement that has not been previously studied experimentally:

(3) *Akh vot ono chto! Akh Golovanovu uvolili! A ya-to dumal, chto Ivanova...* 'Oh, that's how it is! Oh, [it was] Golovanova who was fired! I thought it was Ivanov…'.

In this corrective focus statement (3), in addition to the standard focal information about the incorrectness of the alternatives (normally marked by the H+L*pitch accent [16]), the pragmatic meaning of incredulity can be expressed in Russian using the monotonal H* nuclear accent. Although the resulting tonal configuration bears similarity to the polar question tune in 3.3, the two tunes demonstrate consistent differences in the placement of prenuclear

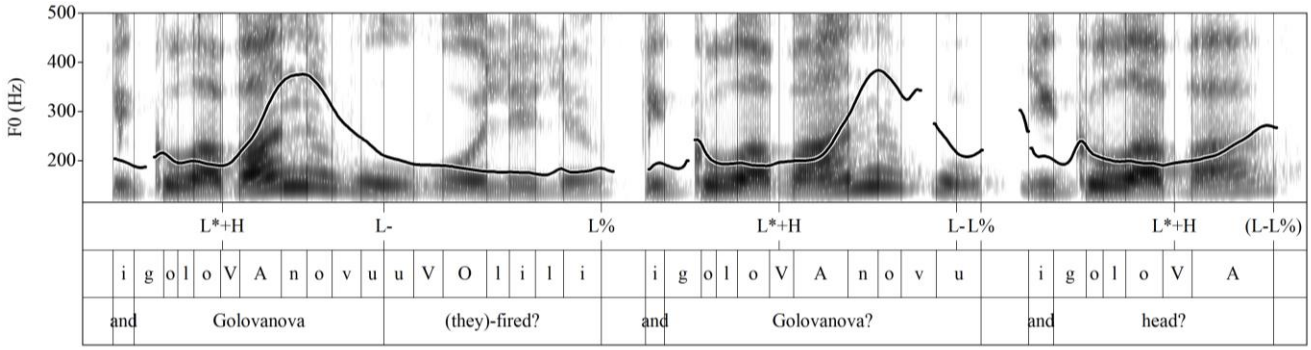


Figure 3: Pitch contours of polar questions produced by speaker 4. The truncated edge tones are shown in brackets.

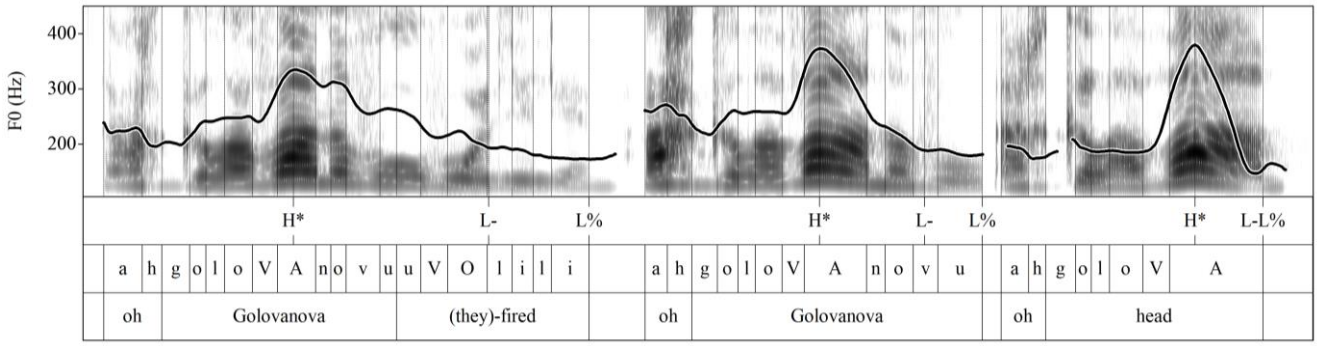


Figure 4: Pitch contours of mirative corrective focus statements produced by speaker 2.

and nuclear pitch targets (Fig. 4). First, the mirative corrective focus nucleus lacks an obligatory low nuclear target (the pitch can rise gradually from the base tone to the H* throughout the prenuclear syllable stretch, without a low dip in the onset of the nucleus considered a typical feature of L*+H). Second, the peak for H* is aligned with the centre of the stressed syllable, while L*+H is produced with peak delay.

Regarding the postnuclear movements, the alignment of the low elbow in mirative corrective focus statements was less consistent than in the polar question tune. For some subjects, such as speaker 2 (see Fig. 4. *left*), no boundary-seeking behaviour of the phrase accent was attested, while for several subjects, the postnuclear part of the contour was identical to that attested in 3.3 for polar questions.

Finally, one difference between the realisation of edge configurations in 3.3 and 3.4 was highly consistent in our data. Two different strategies were employed for the nuclear L*+H and H* under the 0S condition. On the one hand, in polar questions (Fig. 3, *right*), a categorical truncation of postnuclear tones, a documented feature of Russian prosody [7], was observed consistently. On the other hand, in the mirative corrective focus tune, no truncation was attested (Fig. 4, *right*); the nuclear H* and the low target fitted within the lengthened IP-final stressed vowel. These differences can be explained by the earlier alignment of a high nuclear target in the mirative H* compared to the delayed target for the trailing high tone in L*+H.

4. CONCLUSIONS

In this exploratory study, four Russian tonal configurations containing non-prominence-lending postnuclear tonal movements were briefly presented. Based on preliminary qualitative analysis of these tunes, elicited from 12 speakers, we argue that having two distinct levels of edge tones (the phrase accent, in addition to the boundary tone level) allows us to more faithfully model the features of postnuclear pitch movements attested in Russian. First and foremost, a phrase accent differing from the boundary tone helps capture the complex structure of the qualitative implicature tune. In addition, the tendencies observed in the alignment of elbows in polar and comparative question tunes can be approached successfully using the mechanisms of secondary association typical of phrase accents.

One issue not addressed in this paper is the metrical structure implications of employing phrase accents in Russian. Since [17], AM-based analyses often assign a secondary delimitative function to phrase accents: the role of a terminal tone within the intermediate phrase. Quantitative studies are required to clarify this issue for Russian; however, if they confirm the word-boundary-seeking tendencies of Russian phrase accents, including a level of phrasing lower than the IP in the prosodic phonology of this language would be a natural corollary.

5. REFERENCES

[1] A. Arvaniti, "The autosegmental metrical model of intonational phonology," in *Prosodic Theory and Practice*, S. Shattuck-Hufnagel and J. Barnes, Eds. MIT Press, 2022, pp. 25–63.

[2] M. Grice, D. R. Ladd, and A. Arvaniti, "On the Place of Phrase Accents in Intonational Phonology," *Phonology*, vol. 17, no. 2, pp. 143–185, 2000.

[3] C. Gussenhoven, "The boundary tones are coming: On the nonperipheral realization of boundary tones," in *Papers in Laboratory Phonology V: Acquisition and the lexicon*, M. B. Broe and J. B. Pierrehumbert, Eds. Cambridge University Press, 2000, pp. 132–151.

[4] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Turning points, tonal targets, and the English L- phrase accent," *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 982–1023, 2010.

[5] Y. Igarashi, "Russian Interrogatives and Intonational Categories," in *The Discourse Potential of Underspecified Structures*, A. Steube, Ed. De Gruyter, 2009, pp. 227–270.

[6] P. V. Duryagin, "Intonacija častnogo voprosa v russkom jazyke: èksperimental′noe issledovanie istočnikov variativnosti," *Rus. Jaz. v naučnom osveščenii*, vol. 41, no. 1, pp. 137–177, 2021.

[7] T. Rathcke, "How Truncating Are 'Truncating Languages'? Evidence from Russian and German," *Phonetica*, vol. 73, no. 3–4, pp. 194–228, 2017.

[8] C. Odé, "Transcription of Russian intonation ToRI, an interactive research tool and learning module on the internet," *Stud. Slav. Gen. Linguist.*, no. 34, pp. 431–449, 2008.

[9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022.

[10] P. Arantes, "parantes/time-normalized-f0: Version 2.0 of time-normalized-f0 Praat script." 2022.

[11] S. V. Kodzasov, *Issledovanija v oblasti russkoj prosodii*. Moscow: Jazyki slavjanskix kul′tur, 2009.

[12] P. V. Duryagin, "Prosodija konstrukcij «Èto/zdes′/tut tebe/vam ne X»: opyt èlicitacii dannyx metodom imitacii fragmentov mul′timedijnogo korpusa," in *Analiz razgovornoj russkoj reči (AR3-2021): Trudy devjatogo meždisciplinarnogo seminara*, D. A. Kočarov and P. A. Skrelin, Eds. Saint Petersburg: Skifija-print, 2021, pp. 24–30.

[13] E. A. Bryzgunova, "Intonacija," in *Russkaja grammatika, Vol. I*, N. Švedova, Ed. Nauka, 1980, pp. 96–120.

[14] S. V. Knyazev and K. A. Savelyeva, "Fonetičeskoe i fonologičeskoe opisanie v intonacii (IK-4 v russkom jazyke: postcentr ili pograničnyj ton?)," *Rus. Jaz. v naučnom osveščenii*, vol. 2, pp. 41–64, 2021.

[15] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," MIT, 1980.

[16] P. Duryagin and S. Knyazev, "Prosodic diversity in Standard Russian: pitch alignment in Central and Northern varieties," *Russ. Linguist.*, vol. 46, no. 2, pp. 55–75, 2022.

[17] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonol. Yearb.*, vol. 3, pp. 255–309, 1986.

EVIDENCE FOR PREPLANNED AND ADAPTIVE F0 CONTROL

Seung-Eun Kim

Northwestern University
seungeun.kim@northwestern.edu

ABSTRACT

This study investigates how speakers plan and adjust F0 in relation to the anticipated length of an utterance. Utterance length was manipulated prior to response initiation by cueing utterances with one, two, or three subject noun phrases (NPs) with visual stimuli. Furthermore, on some trials, a novel experimental manipulation was tested in which additional stimuli were presented upon detection of utterance initiation, thereby increasing the utterance length after participants began production. There are two main findings: (i) participants raised the F0 ceiling and floor and broadened the span of the initial NP when they had to produce longer utterances; (ii) participants reduced the amount of F0 ceiling compression from the first to the second NP to adapt to the changes in utterance length. Both findings suggest that speakers aim to maintain a sufficient F0 space while speaking, by taking into account the anticipated length of the utterance.

Keywords: F0 control, F0 preplanning, adaptive control, utterance length, delayed stimuli

1. INTRODUCTION

Many studies have investigated whether speakers vary utterance-initial F0 according to sentence length, aiming to find evidence for preplanning of F0 parameters. The hypothesis that was tested is that speakers would raise the initial F0 in longer utterances to avoid hitting the bottom of their F0 range before the utterance ends. A correlation between utterance length and initial F0 peak was examined in a variety of languages, and the studies have found mixed results. For instance, a significant effect of utterance length on the initial F0 peak was found in [1], [2], [3], [4], while no correlations were observed in [5], [6], [7], [8]. The results of the studies were inconsistent even within the same language (e.g. English: [1] vs. [5]).

The current study revisits the issue of F0 preplanning by examining the effects of utterance length on the overall F0 control, specifically by analyzing F0 peaks, valleys, and ranges. This contrasts with previous studies which almost exclusively examined F0 peaks. The goal of the current analysis therefore is to investigate how speakers control F0 register (F0 space) that is composed of F0 ceiling, floor, and span before utterance initiation (rather than just the ceiling), which may provide better insights on F0 preplanning.

The other question that is examined in this study is whether speakers can adjust F0 when utterance length changes after the start of production – i.e. adaptive F0 control. For this purpose, a novel experimental paradigm was developed in which the stimuli that cue the parts of the utterance are delayed until after participants initiate production. Under this condition, participants have to quickly adapt to changes in the length and content of the utterance.

As far as I am aware of, no studies have examined how speakers adapt F0 control to externally cued changes in the utterance length/content that occur after utterance initiation. These lexical/phrasal perturbations may be similar to or different from other sorts of perturbations made in the previous studies. For instance, [9] and [10] found evidence that speakers can incorporate a segment into an ongoing utterance that was missing before response onset but was cued after response initiation. A number of studies also found that speakers can adapt their F0 control to pitch-shifted auditory feedback (e.g. [11], [12], [13], [14], [15], [16]). In these studies, participants produced a vowel, syllable, or sentence, while hearing back their production with F0 perturbations – i.e. pitch of their voice was either raised or lowered compared to the original production. Participants were sensitive to the pitch-altered feedback and in general showed compensatory responses.

In sum, this study examines speakers' preplanned and adaptive F0 control through variations in utterance length. Participants produced sentences with one, two, or three subject noun phrases (NPs), which were cued by visual stimuli. For the multiple-NP sentences, two conditions were tested: one in which the stimuli for the non-initial NPs were delayed, and the other in which no stimuli were delayed. Various F0 parameters – peaks, valleys, and ranges – were examined for the length effects.

The following are the hypotheses and predictions.
Hypothesis 1. Preplanned F0 control: Speakers will make a pre-utterance F0 plan according to the initial utterance length. *Predictions*: The values of the F0 peaks and valleys will be higher, and ranges will be larger in longer utterances.
Hypothesis 2. Adaptive F0 control: Speakers will adapt F0 control in response to changes in utterance length. *Predictions*: The differences in the values of the F0 peaks, valleys, and ranges between the first and second NP will be smaller (less F0 register compression) in the condition where the stimuli are delayed compared to the condition without delay.

2. METHODS

2.1. Participants and experiment design

13 native speakers of English (7M, 6F) participated in the experiment. Participants produced sentences with one, two, or three subject NPs. The lexical and phonological content of the NPs was controlled as monosyllabic numeral {*eight, nine*} + monosyllabic color {*red, green, blue*} + disyllabic animal {*llamas, rhinos, weasels*}. In sentences with multiple NPs, animals were always unique, although numerals and colors could be repeated. Participants were instructed to connect NPs with the conjunction "*and*". An example sentence with three NPs was "*Nine green rhinos and eight red weasels and eight blue llamas live in the zoo*". In the experiment, NPs were cued with visual stimuli as in Figure 1.

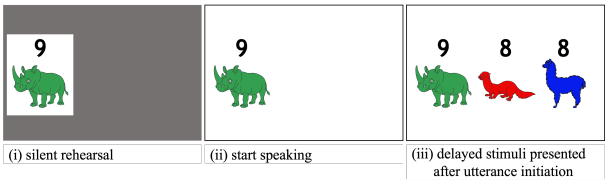


Figure 1: Presentation of a single trial.

For sentences with multiple NPs, a condition was tested in which the visual stimuli that cue non-initial phrases were delayed until after utterance initiation was detected. In this condition, participants saw only one stimulus before production, and the rest of the stimuli appeared immediately after they initiated an utterance. This condition is referred to as the *delayed stimuli* (DS) condition, as opposed to the *no-delayed stimuli* (NS) condition, in which all stimuli appeared before production. Thus, a total of five conditions were tested in the experiment – i.e. three lengths × two delayed conditions (1NS, 2NS, 2DS, 3NS, 3DS); cf. 1DS could not be tested. In each block, there were 30 trials: each of the 2NS, 2DS, 3NS, and 3DS conditions appeared five times, and the 1NS condition appeared 10 times, all of which were randomized from trial to trial. Each experimental session had nine blocks.

The target sentences were cued as in Figure 1. In each trial, the initial stimuli appeared with a grey background (Figure 1-(i)). Participants were instructed to silently rehearse the sentence during this phase. The grey background lasted for 2.7s if there was a single stimulus, 4.4s for two stimuli, and 6.1s for three stimuli; these durations were determined based on the average utterance durations in the pilot experiments, which were conducted on three English speakers. After the given periods, the background automatically changed to white, which cued participants to start speaking (Figure 1-(ii)). In the DS conditions (2DS, 3DS), the stimuli that cued non-initial NPs were presented immediately after utterance initiation was detected (Figure 1-(iii)).

Participants were instructed not to emphasize any words in the utterance. This was to prevent them from putting contrastive focus on the parts of the NPs, which may affect natural intonation of the utterance. Participants performed 24 practice trials with balanced exposure of all conditions and words before the experiment session. This was to familiarize them with the task and to ensure that they produce sentences without any focus.

2.2. Data collection and processing

Acoustic signals were recorded at a sampling rate of 22050 Hz. Word and phone-level segmentations were conducted using Kaldi speech recognition toolkit ([17]). For each participant, ten trials, which included at least one instance of each numeral, color, and animal, were manually labelled and used to train monophone HMMs. A forced alignment was then conducted on the remaining trials. F0 data were extracted in Praat with participant-specific pitch floor and ceiling, and then a smoothed and interpolated F0 contour was generated for each trial. With the segmented data, F0 of the subject phrase was extracted for the analyses.

Due to the novelty of the delayed stimuli design, participants produced disfluencies such as hesitations or speech errors on some trials. These trials were identified algorithmically based on word and between-word interval durations. Specifically, for each word and interval, a mixed-effects linear model was fit to the durations with the experimental condition as fixed effect and the participant as random intercept. At each location, datapoints whose standardized absolute residuals were larger

than 3.09 (the 0.1/99.9 percentiles of a normal distribution) were considered to be duration outliers, and trials with these outliers were excluded from the analyses. After excluding these trials, there were two participants for whom the exclusions constituted more than 20% of their data. This suggests that these participants did not conform to the task instructions, and thus, their data were excluded. In sum, among 2970 trials (270 trials $\times$ 11 participants), 265 trials (8.9%) were identified to contain duration-based outliers and excluded.

For those trials that had delayed stimuli, the timepoint of the presentation of the delayed stimuli and the onset of the utterance determined by the forced alignment were compared to ensure that the delayed stimuli appeared after utterance was initiated. Three trials (0.1%) were identified as errors, as the beginning of the utterance followed the presentation of the delayed stimuli. Nine trials (0.3%) that had problems in recording were also discarded. This left a total of 2693 trials (90.7%) for the analyses.

2.3. Data analysis

To ensure that the F0 patterns that are analyzed are similar across participants, a global analysis was first conducted to compare participant-level F0 patterns. An average time-warped F0 contour was obtained for each participant and condition and was qualitatively assessed. Out of 11 participants, seven of them had similar intonation patterns, in which an F0 valley occurred at the numeral, an F0 peak at the color, and another F0 valley at the second syllable of the animal. The current analyses thus focused on the F0 trajectories of these seven participants, and the F0 values of the landmarks – F0 valleys and peaks – and the ranges between them (F0 rises/falls) were measured.

For F0 preplanning, F0 values of the landmarks and rises/falls of the initial NP were analyzed using linear mixed-effects models. For each measurement, a mixed-effects model was fit to F0 values with the fixed effect of utterance length (i.e. number of subject NPs) and the random intercept of participants.

For adaptive control, stepwise regressions were conducted on the differences of F0 values of the landmarks and ranges between the first and the second NP (e.g. peak differences between NP1 and NP2). A mixed-effects model with the fixed effects of utterance length and delay, their interactions, and the random intercept of participants was fit to each of the difference measures. The maximal model was subsequently compared to the model that lacked the interaction term and the models that lacked either of the fixed effects. As delayed stimuli were presented shortly after response onset, their effects were likely to emerge during NP1 and NP2.

3. RESULTS

The results found evidence for both preplanned and adaptive F0 control. First, it was found that the participants started with a higher F0 peak and valley as well as a wider F0 range, when they were presented with more initial NPs, providing evidence for F0 preplanning. Not only the F0 peaks, but also valleys and ranges varied with utterance length.

For this analysis, the five experimental conditions were grouped according to the number of initial stimuli; thus, 1NS, 2DS, and 3DS were coded as 1Pi (one initial stimulus), and 2NS and 3NS were coded as 2Pi and 3Pi, respectively. This alternative coding is motivated by the finding that the 1Pi conditions (1NS, 2DS, 3DS) showed a similar F0 pattern in

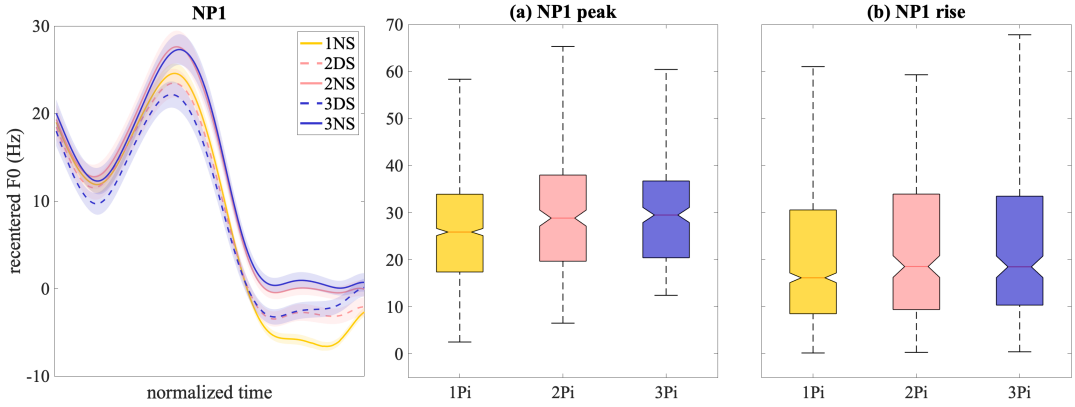


Figure 2: Average time-warped F0 contours of NP1 and the distributions of (a) peaks and (b) rises. For visualization, F0 (y-axis) was recentered within each participant using their global F0 mean.

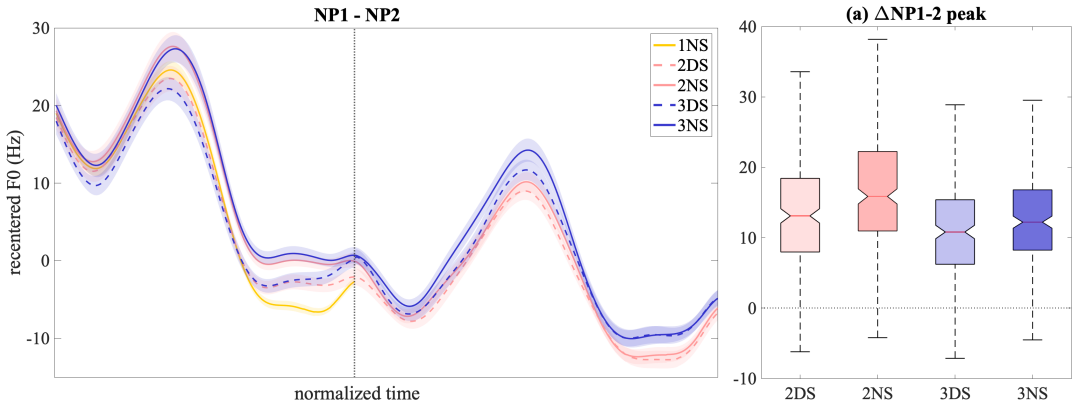


Figure 3: Average time-warped contours of NP1 and NP2 and the distributions of (a) peak differences. The vertical dotted line in the left panel represents NP boundary (i.e. the length of the time-warped NP).

NP1, which can be found in the leftmost panel of Figure 2. The regression coefficients of the 2Pi and 3Pi compared to 1Pi were 1.87 Hz and 2.07 Hz for F0 valleys ($p < 0.01$), 3.94 Hz and 5.06 Hz for F0 peaks ($p < 0.001$), and 2.11 Hz and 2.64 Hz for F0 rises ($p < 0.01$). Differences between conditions are also shown in Figure 2-(a)/(b). The difference between 2Pi and 3Pi was significant only at the peak and was minimal, as they differed only by 1.31 Hz ($p < 0.01$).

Second, participants lowered F0 peaks to a *lesser* extent when they saw delayed stimuli. Participants in general lowered F0 peaks from NP1 to NP2, as shown in the left panel of Figure 3. However, the amount of peak compression differed between DS and NS conditions such that the compression was larger in the NS conditions than the DS conditions. This is shown in Figure 3-(a) and confirmed by the statistical analysis, in which the regression coefficient of NS conditions (compared to DS) was 2.78 Hz for the F0 peak differences between NP1 and NP2 ($p < 0.001$). This suggests that as participants saw delayed stimuli, they adapted to the increase in the utterance length, by reducing the amount of F0 peak lowering (i.e. F0 register compression) that they would otherwise have done.

4. DISCUSSION AND CONCLUSION

This study examined whether speakers preplan F0 according to the initial utterance length, and moreover, whether they adapt F0 control in response to the unanticipated changes in length. The results found evidence for both preplanned and adaptive control: first, participants produced higher F0 peaks and valleys and wider F0 ranges when they had to produce longer utterances; second, when additional phrases were presented, they reduced the amount of F0 peak lowering to accommodate length changes.

One likely motivation behind both of these patterns is to establish a sufficient F0 space for production of upcoming phrases, given an expected declination of F0 across the utterance. When participants are producing a long utterance, they start from a higher F0 ceiling/floor as well as a wider F0 span, to avoid a circumstance in which the declination leads F0 floor to be too low or F0 span to be too narrow. When the length of the utterance changes after they start production, they quickly adjust the amount of F0 ceiling compression to make more F0 space for delayed phrases. The results thus suggest a strong tendency of speakers to produce F0 within their habitual, comfortable range and to establish sufficient F0 space throughout the utterance.

The present findings contribute to the inconsistent literature on F0 preplanning, by showing that the speakers factor utterance length into their F0 control prior to utterance initiation. It further informs us that it is not just the F0 peak, but other F0 variables such as valleys and ranges that are influenced by utterance length. Moreover, the adaptive responses suggest that speakers continuously estimate the remaining length of the utterance as well as the relation of their current F0 from the available F0 space and adjust their F0 control if necessary.

Overall, this study found robust evidence for F0 preplanning and adaptive control. Some possible future directions are whether similar results are observed when length is manipulated differently (e.g. adding sentences, words rather than NPs), whether there are individual differences in preplanned and adaptive F0 control, and whether the production differences observed in this study are perceptible by listeners.

5. REFERENCES

[1] W. E. Cooper and J. M. Sorensen, *Fundamental frequency in sentence production*. Springer Science & Business Media, 1981.

[2] G. Bruce, "Developing the swedish intonation model," *Working papers/Lund University, Department of Linguistics and Phonetics*, vol. 22, 1982.

[3] C. Shih, "A declination model of mandarin chinese," in *Intonation*. Springer, 2000, pp. 243–268.

[4] Y. O. Laniran and G. N. Clements, "Downstep and high raising: interacting factors in yoruba tone production," *Journal of phonetics*, vol. 31, no. 2, pp. 203–250, 2003.

[5] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language sound structure*. MIT Press, 1984.

[6] R. van den Berg, C. Gussenhoven, and T. Rietveld, "Downstep in dutch: Implications for a model," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press, 1992, pp. 335–359.

[7] P. Prieto, C. Shih, and H. Nibert, "Pitch downtrend in spanish," *Journal of Phonetics*, vol. 24, no. 4, pp. 445–473, 1996.

[8] B. Connell, "Tone, utterance length and fo scaling," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[9] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, no. 1, pp. 3–35, 1990.

[10] S. Tilsen, "Selection and coordination of articulatory gestures in temporally constrained production," *Journal of Phonetics*, vol. 44, pp. 26–46, 2014.

[11] T. A. Burnett, M. B. Freedland, C. R. Larson, and T. C. Hain, "Voice f0 responses to manipulations in pitch feedback," *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3153–3161, 1998.

[12] J. A. Jones and K. G. Munhall, "Perceptual calibration of f0 production: Evidence from feedback perturbation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1246–1251, 2000.

[13] T. M. Donath, U. Natke, and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on voice f0 contours in syllables," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 357–366, 2002.

[14] U. Natke and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 577–584, 2001.

[15] S. H. Chen, H. Liu, Y. Xu, and C. R. Larson, "Voice f0 responses to pitch-shifted voice feedback during english speech," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1157–1163, 2007.

[16] R. Patel, C. Niziolek, K. Reilly, and F. H. Guenther, "Prosodic adaptations to pitch perturbation in running speech," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1051–1059, 2011.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

INVESTIGATIONS OF F0 CONTROL: PITCH TARGETS VS. PITCH REGISTER

Seung-Eun Kim[1], Sam Tilsen[2]

[1]Northwestern University, [2]Cornell University
seungeun.kim@northwestern.edu, tilsen@cornell.edu

ABSTRACT

This study examines speakers' control of F0 by evaluating *target*- and *register*-control hypotheses. The *target*-control hypothesis holds that speakers adjust individual pitch targets to produce variations in F0, while the *register*-control hypothesis holds that speakers adjust pitch register (in which the pitch targets are defined) to vary F0. These hypotheses are evaluated with empirical F0 trajectories, examining correlations between F0 peaks and valleys. The results found that the peaks and valleys are not independently controlled, and that they are positively correlated. This suggests that speakers may control F0 by adjusting pitch register rather than by changing targets, and with a greater extent of manipulation of register ceiling or floor compared to span.

Keywords: F0 control, pitch targets, pitch register, correlation

1. INTRODUCTION

This study is motivated by the insight that there are two ways in which the control of F0 can be accomplished. One is that the speakers adjust individual *pitch targets*, and the other is that they adjust *pitch register*, understood here as a sensorimotor representation of the F0 control space in which the targets are defined. We refer to the first possibility as the *target*-control hypothesis and the second as the *register*-control hypothesis, and we aim to assess which one better accounts for empirical patterns in F0 trajectories over multi-phrase utterances.

A pitch target is defined here as a cognitive representation of F0 that speakers want to achieve while speaking. Speakers are assumed to implement a series of F0 goals or targets during production, and the vocal control system has parameters that determine these values. Theories of intonation have conceptualized the notion of pitch target in various ways. For instance, in the Autosegmental-Metrical intonation phonology (e.g. [1], [2], [3], [4]), speakers are considered to aim for distinctive pitch levels, and these targets are realized as peaks and valleys in surface F0 contours. The PENTA model proposed by [5] holds pitch targets to be either levels or rises/falls, which however are the underlying targets that do not necessarily map to the surface peaks and valleys. In some computational F0 models (e.g. [6], [7], [8]), pitch targets are defined with parameters that specify target values, forms/shapes, and durations.

Pitch register is commonly referred to as the range of F0 values (F0 space) that speakers can produce and utilize at a given time in an utterance. It is defined by a combination of at least two of the following three parameters: ceiling, floor, and span. According to [4], the notion of pitch register has been introduced in the F0 literature under different names, such as "tonal space" [9] in resemblance to vowel space, "tonal level frame" [10], "transform space" [3], or "grid" [11]. While all these terms represent the notion of space or range, some models lack a full specification of range, and instead specify a rough position within the range that pitch targets are superimposed on (e.g. [7], [8]).

Figure 1 illustrates the two different (yet not mutually exclusive) control strategies. Under the (i) *target*-control hypothesis, variation in the values of the F0 peaks and valleys within an utterance is directly manipulated by speakers. As shown in Figure 1-(i), to produce an F0 contour with two different F0 peaks, speakers would have two distinct high (H) pitch targets in mind, for example one at the top of the current pitch range (1.0) and the other at about 60% of the range (0.6). In contrast, under the (ii) *register*-control hypothesis, F0 variation arises from changes in register: speakers adjust the tonal space, while the targets remain constant. This is reflected in Figure 1-(ii): although targets are same across prosodic units (both are 1.0), the register shift results in different surface F0 peaks.

Evidence for these hypotheses is evaluated using F0 trajectories extracted from the experimental data. The experiment elicited utterances with one, two, or three subject noun phrases (NPs). The F0 contours

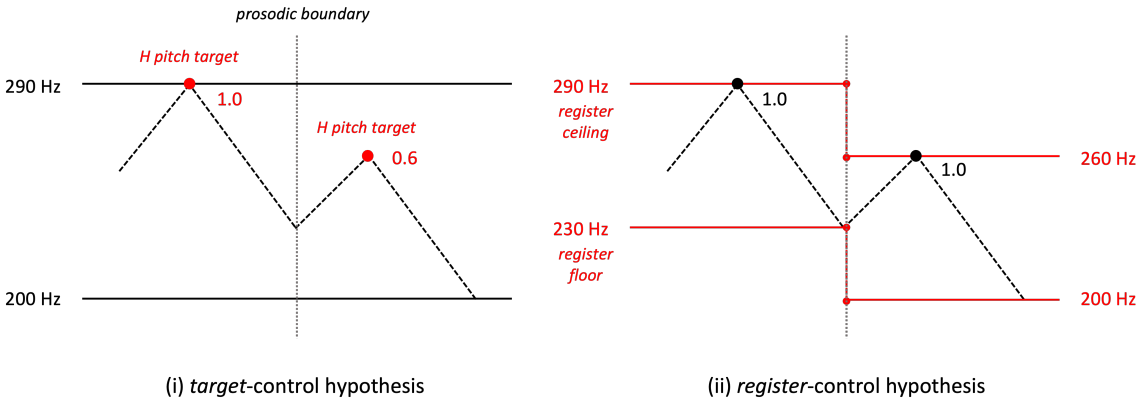


Figure 1: Schematic illustrations of F0 control hypotheses. The black dashed line represents a schematic F0 contour, the solid horizontal lines show pitch register ceiling and floor, and the dots at the peaks indicate high (H) pitch targets. A hypothetical prosodic boundary (prosodic word or phrase) is indicated as a vertical dotted line. The main F0 parameter that leads to variations in F0 under each hypothesis is marked in red. Arbitrary parameter values are provided for H targets and register ceiling/floor.

that we analyze had an F0 valley, a peak, and another valley at each NP. This may be represented in the Autosegmental-Metrical framework as an L+H*-L sequence, although nothing in our analysis depends on the validity of this representation.

Since the control parameters cannot be observed directly, F0 peaks and valleys are assumed to reflect high (H) and low (L) pitch targets, and F0 ranges to reflect register span. Although surface F0 peaks, valleys, and ranges may not exactly reflect internal representations of pitch targets or register, these surface measures are likely to be highly correlated with the underlying control parameters. We note that this assumption is most tenable in the moderate-to-slow rate of speech of this study, where target undershoot is less likely to occur than in fast speech.

The analysis we conduct below examines the correlations between F0 peaks and valleys of each NP. Specifically, if correlation coefficients are relatively close to 0, it suggests that the F0 peaks and valleys are controlled independently, which follows more directly from the *target*-control hypothesis. In contrast, if the correlation coefficients are relatively far from 0, it supports the *register*-control hypothesis, because variation in register will result in non-independent (i.e. correlated) changes of peaks and valleys.

Furthermore, if we observe relatively strong peak-valley correlations, it is important to examine the signs of those correlations, because these allow for further inferences on which register parameter(s) – ceiling, floor, span – are the ones that are manipulated. Our interpretation of the correlation sign is premised on the assumption that the H and L targets are represented as fixed proportions of the register (F0 space) as in Figure 1, and that they are near the (normalized) ceiling and floor of the register. Under these assumptions, a *negative* correlation (the peaks and valleys tend to vary in the opposite directions) suggests that the speakers vary the register span to a greater extent than the register ceiling or floor. On the contrary, a *positive* correlation suggests that the speakers vary floor or ceiling more than span. It should be noted that there are more complicated ways for combinations of both span and floor/ceiling changes to result in positive or negative correlations, and our interpretations are limited to the simplest possibilities.

2. METHODS

2.1. Experiment design

A production experiment was conducted which elicited sentences of varying lengths. Specifically, the length of the subject phrase was varied so that it was composed of one, two, or three conjoined noun phrases (NPs). Each NP was comprised of a numeral (*eight, nine*), color (*red, green, blue*), and animal (*llamas, rhinos, weasels*). In cases of multiple-NP subject phrases, animals were always unique, while numerals and colors could be repeated. An example for the three NP sentences was "*Nine green rhinos and eight red weasels and eight blue llamas live in the zoo.*" All NPs were cued with visual stimuli.

For sentences with two and three NPs, a condition was tested in which the stimuli cueing the non-initial NPs were presented immediately after participants initiated an utterance rather than at the beginning of the trial. In these conditions, participants had

to quickly incorporate the newly presented NPs into their ongoing utterance. A total of five experimental conditions – 3DS (three subject NPs with delayed stimuli), 3NS (three NPs without delayed stimuli), 2DS, 2NS, and 1NS – were thus tested in the experiment, and the conditions were randomized from trial to trial. Each of the 3DS, 3NS, 2DS, and 2NS conditions appeared 45 times, and the 1NS condition appeared 90 times in each experimental session, which resulted in a total of 270 trials. 13 native speakers of English participated in the experiment. Note that the distinction between DS vs. NS conditions is indeed not necessarily relevant to the control hypotheses that we examine in this study; they exist to test hypotheses that are not considered here.

2.2. Data and measurements

Data were collected at a sampling rate of 22050 Hz. Acoustic segmentation was conducted using the Kaldi speech recognition toolkit ([12]). For each participant, ten trials, which included at least one instance of each numeral, color, and animal, were manually labelled to train monophone HMMs, and the rest of the trials were forced aligned.

Before conducting analyses, trials with potential disfluencies or with problems in data collection were removed. Trials with potential disfluencies were first algorithmically identified. Specifically, for each word and between-word silence interval, a mixed-effects linear regression model was fit to the word or interval durations with experiment condition as a fixed effect and participant as a random intercept. Trials with extreme durations were considered to contain disfluencies and excluded from subsequent analyses (265/2970 trials: 8.9%). Data from two participants were excluded due to a high proportion of hesitation or speech errors (i.e. more than 20% of trials). A small number of trials with problems in recording or stimuli presentation (i.e. delayed stimuli presented before utterance initiation) were also identified and discarded (12/2970 trials: 0.4%). This altogether left 2693 trials in total.

For the remaining trials, F0 data were extracted in Praat. A smoothed and interpolated F0 contour was then generated for each trial. To ensure that the accentual patterns in the analysis were comparable across participants, an average time-warped F0 contour was generated for each participant and condition and was qualitatively compared. Among 11 participants, seven of them showed similar intonation patterns in the subject phrase; specifically, they produced an F0 valley, F0 peak, and another F0 valley at each NP (which presumably constitutes a phonological phrase). F0 values of these landmarks of the seven participants were measured and subject to analyses.

2.3. Data analysis

The linear correlations between F0 peaks and valleys were calculated for each participant and NP. In each of these cases, we analyzed correlation between (i) preceding valley and peak, and correlation between (ii) peak and following valley. These correlations were calculated separately in DS and NS conditions. This is because variations in peaks and valleys may differ by whether all stimuli were presented at the beginning of the trial (fully planned before utterance initiation) vs. when some of them were delayed (had to change utterance plan after production). Note, however, that how the F0 values of peaks and valleys differ between DS vs. NS conditions is not directly relevant to the goal of the current study and is not analyzed here.

Thus, we obtained four correlation coefficients for each combination of participant and NP – i.e. two types of correlations (preceding valley & peak, peak & following valley) were calculated for each of the DS and NS trials. There were seven participants, and the subject phrase had maximum of three NPs, which resulted in a total of 84 correlation coefficient values.

3. RESULTS

The analysis showed that the peaks and valleys are highly correlated, providing evidence for the *register*-control hypothesis. In the majority of cases (i.e. in all participants/NPs, between peaks and preceding/following valleys), the correlation coefficients were positive and far from 0. Figure 2 shows the coefficients calculated with preceding valley and peak (a)/(b), and peak and following valley (c)/(d) at NP1 and NP2. Except for a handful of cases (e.g. (a) PA01, (b) PA02, (d) PA06), correlation coefficients were positive and relatively far from zero.

In fact, out of 84 correlation coefficients we have obtained, only five of them were negative. In addition, the absolute coefficient values were above 0.2 in 67 cases, showing that most of them were far from 0.

4. DISCUSSION AND CONCLUSION

This study aimed to assess evidence for two ways in which the control of F0 can be accomplished – *target*-control vs. *register*-control. We examined the

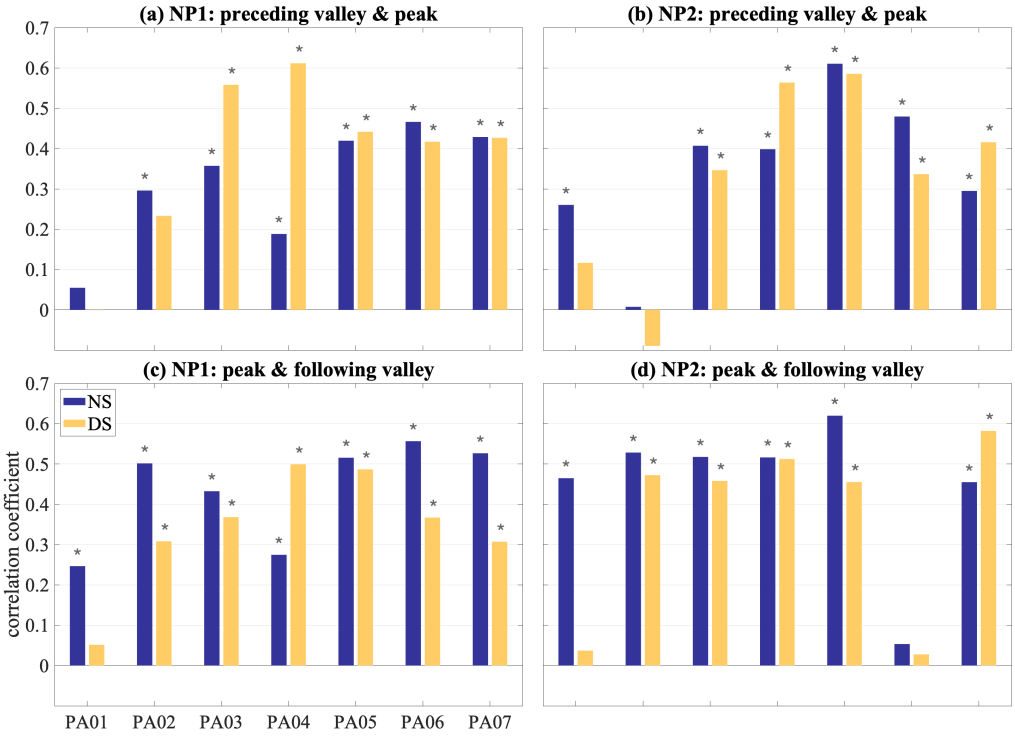


Figure 2: Correlation coefficients (a), (b): between preceding valley and peak, (c), (d): between peak and following valley at NP1 and NP2. An asterisk indicates a significant correlation ($p < 0.05$). The average number of data points used to calculate the correlation was 158 (NS) and 71 (DS) for NP1, and it was 73 (NS) and 70 (DS) for NP2.

correlations between F0 peaks and valleys at each NP, and the correlation coefficients in most cases were far from 0. This finding suggests that F0 peaks and valleys are not independently controlled but are correlated, providing evidence for the *register*-control hypothesis. A plausible interpretation of the correlation is that for a given utterance, speakers have a set of invariant cognitive representations of high and low pitch targets, and they control pitch register to realize the abstract representation into different F0 peaks and valleys (Figure 1-(ii)).

Furthermore, the analysis found that the F0 peaks and valleys are positively correlated in most cases (i.e. coefficients > 0). The simplest interpretation of this finding is that speakers manipulate either ceiling or floor to a greater extent than they manipulate span, at least in the current experimental task. The reasoning behind this interpretation is as follows: H and L pitch targets are likely to be located near the register ceiling and floor, respectively. If span is the primary register parameter that speakers manipulate, the contractions/expansions of span across NPs will lead to opposing changes in peaks and valleys, i.e. negative correlations. Conversely, if either the floor or ceiling is the primarily manipulated parameter, F0 peaks and valleys will tend to be mutually raised or lowered, i.e. positive correlations. We note that these inferences assume that *register*-control is as parsimonious as possible, in that speakers manipulate primarily just one of the three register parameters. Note that in the case of manipulating span, one or both of the edge parameters (ceiling and/or floor) would change indirectly. Conversely, in the case of manipulating the edge parameters, indirect span changes may occur.

Another caveat regarding the interpretation of our results is that there is a more complicated account of the empirical patterns that is consistent with *target*-control. Specifically, speakers might control H and L pitch targets in a correlated manner, not because of the register per se, but due to some other unknown mechanisms. Yet, we hold that the *register*-control interpretation is preferrable, since it does not require an additional mechanism for correlating H and L pitch targets. Moreover, there are a wide variety of phonological patterns (such as downstep and post-focus compression) which can be well understood as instances of *register*-control (e.g. [13], [14], [15]). Thus, although both *target*- and *register*-control can generate correlations of peaks and valleys, *register*-control is the simpler one and also has external motivations.

5. REFERENCES

[1] J. Pierrehumbert, "The phonology and phonetics of english intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[2] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language sound structure*. MIT Press, 1984.

[3] J. Pierrehumbert and M. E. Beckman, *Japanese Tone Structure*. MIT Press, 1988.

[4] D. R. Ladd, *Intonational Phonology (Second Edition)*. Cambridge University Press, 2008.

[5] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech communication*, vol. 46, no. 3-4, pp. 220–251, 2005.

[6] G. Kochanski and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, no. 3-4, pp. 311–352, 2003.

[7] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*. Springer, 1983, pp. 39–55.

[8] ——, "Prosody, information, and modeling-with emphasis on tonal features of speech," in *Workshop on Spoken Language Processing*, 2003.

[9] D. R. Ladd, "An introduction to intonational phonology," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press, 1992, pp. 321–334.

[10] G. N. Clements, "The description of terraced-level tone languages," *Language*, pp. 536–558, 1979.

[11] E. Gårding, "A generative model of intonation," in *Prosody: Models and measurements*. Springer, 1983, pp. 11–25.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[13] G. N. Clements, "The status of register in intonation theory," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press, 1990, pp. 58–71.

[14] D. R. Ladd, "Phonological features of intonational peaks," *Language*, pp. 721–759, 1983.

[15] ——, "Metrical representation of pitch register," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press, 1990, pp. 35–57.

INTONATIONAL PREFERENCES FOR LEXICAL CONTRAST AND VERUM FOCUS

Christine T. Röhr, Martine Grice, Stefan Baumann

I*f*L Phonetik, University of Cologne, Germany
christine.roehr@uni-koeln.de, martine.grice@uni-koeln.de, stefan.baumann@uni-koeln.de

ABSTRACT

This paper explores the intonational marking and functional relatedness of lexical contrast and verum focus in German. Lexical contrast, such as "She teaches" vs. "She learns" is commonly expressed by an L+H* accent, although L*+H is also attested. Verum focus highlights the truth of a proposition, such as "She does learn", by an accent, reportedly H*, on the finite verb. In a rating study, we investigated the appropriateness of H*, L+H* and L*+H on the two focus types. Listeners preferred L+H* on lexical contrast and L*+H on verum focus. Interestingly, in both conditions L+H* was preferred over H*. Hence, both lexical contrast and verum focus require prominent accents, i.e. rising accents with a large pitch excursion, but can also be realized by two distinct intonation patterns. These results contribute to the debate on the nature of verum focus suggesting a relatedness to contrastive focus.

Keywords: focus, contrast, verum, intonation, rises

1. INTRODUCTION

In West Germanic languages words in narrow and contrastive focus have been found to involve greater prosodic prominence than words in broad and non-contrastive focus, respectively (see [1] for an overview). This prominence results from both the realisation of different pitch accent types and continuous modulations of pitch and segmental durations. In this paper we aim to investigate the intonational marking of lexical contrast and "verum focus" in German with a view to contributing to the debate about their possible relatedness.

Contrastivity is often related to narrow focus (which refers to the size of the focus domain, i.e. a single syntactic constituent [2]), but requires the presence of focus alternatives in the common ground, either explicitly or implicitly (cf. [3, 4]). The typical use of contrastive focus is corrective [4], i.e. when the proposition in the context differs from the proposition in the target sentence. While lexical contrast focuses on a lexical category (noun, adjective, verb), verum focus is widely understood as, inter alia, focusing on a functional category – a semantic operator VERUM, which is invisible on the phonetic surface [5]. According to Höhle [5], verum focus is expressed by a nuclear accent on the finite full verb or auxiliary (alternatively a complementizer or relative pronoun) in order to mark a proposition as true. The verb is supposed to be associated with the VERUM operator, which becomes prominent when the verb is accented. Crucially, verum focus is generally assumed to be non-contrastive (cf. [5]). However, Lohnstein [6, 7] re-analyzed verum focus and argues that it is a sub-type of contrastive focus which reduces the alternatives to sentence mood functions in the discourse situation. Höhle [5] and Lohnstein [6, 7], inter alia, clearly define the verum accent as focus, but there are also some approaches that treat it as phonological realisation of a conversational operator that is not linked to focus (see [8, 9] for discussion).

In German, the most common (GToBI [10]) accent type marking lexical contrast reported in the literature is a medial-late peak accent L+H* [11, 12, 13, 14, 15, 16, 17], although a late peak accent L*+H is also possible. For verum focus Turco et al. [18, 19] report a medial-peaked high-falling nuclear pitch accent (H* L-/H+L*) on the finite verb. However, their test stimuli involved the auxiliary *hat* "has", with a short vowel flanked by voiceless consonants, leading to possible truncation of the pitch contour (e.g. [20, 21]), making it difficult to determine the accent type. Observations from spontaneous speech (*Fokus-DB* [22]) reveal that different accent types can be used to express verum focus, and that there is some overlap with those used for contrastive focus. However, accent types with a late peak (L*+H) tend to be used more often to mark verum focus than lexical contrast.

To shed light on the intonational form and functional relatedness of lexical contrast and verum focus, we tested the perceptual appropriateness of three types of accentual rise, differing in the timing and the extent of the rise – namely H*, L+H* and L*+H – as markers of these focus types in German.

2. METHOD

2.1. Stimuli and Conditions

In a web-based perception task, we tested the appropriateness of five different target sentences,

each realized with three different nuclear accent types on the focus exponent, i.e. the *finite* verb, in relation to contexts eliciting lexical contrast or verum focus.

The target sentences (1) all consist of a sentence-initial pronominal subject and a finite verb in second position followed by a verbal complement. A *full* verb occupies the position of the finite verb in sentences (1a-d), while sentence (1e) has a *modal* verb in this position. In the latter case, the predicate is jointly composed of the modal verb and a full verb, which is placed at the end of the sentence in the infinitive. The finite verbs in second clause position serve as target words marked by different nuclear accent types. They are all disyllabic with primary stress on the first syllable, have a simple segmental structure and are predominantly voiced to provide a continuous f0 trajectory. The verbal complements consist of either two or three syllables to avoid truncation or compression of the f0 contour on the preceding nuclear accented target word.

(1)		**pronoun**	**finite verb**	**verb. compl.**	**infinitive**
	a.	*Ich*	***lebe***	*auf Rügen.*	
		"I	live	on Rügen."	
	b.	*Ich*	***lerne***	*Spanisch.*	
		"I	learn	Spanish."	
	c.	*Ich*	***nähe***	*mein Kostüm.*	
		"I	sew	my costume."	
	d.	*Ich*	***wohne***	*in Köln.*	
		"I	reside	in Cologne."	
	e.	*Ich*	***wollte***	*den Hund*	*füttern.*
		"I	wanted	the dog	to feed."

The target sentences were read by a trained 23-year-old male phonetician (native German speaker), recorded in a sound-attenuated booth. The speaker was instructed to produce each of the five target sentences naturally and in a neutral mode with three rising nuclear pitch accent types (according to GToBI [10]) on the finite verbs: A high accent with a medial peak and a small pitch excursion on the stressed syllable (H*), a high accent with a medial-late peak and a steep rise on the stressed syllable, i.e. with a higher scaling and larger pitch excursion (L+H*), and a low accent with a late, steeply rising f0 movement whose peak is reached after the stressed syllable (L*+H). The perceptual prominence of these accents has been attested to increase from H* over L*+H to L+H* [23]. In order to keep the prosodic variability of the test sentences to a minimum they all begin and end with a low intonation contour. Furthermore, we controlled both the perceptual and acoustic equivalence of the respective accent types on the target words. No adjustments of the recorded sentences were made, except for an equalization of the sound level. Figure 1 shows the individual f0 traces of all test sentences superimposed.

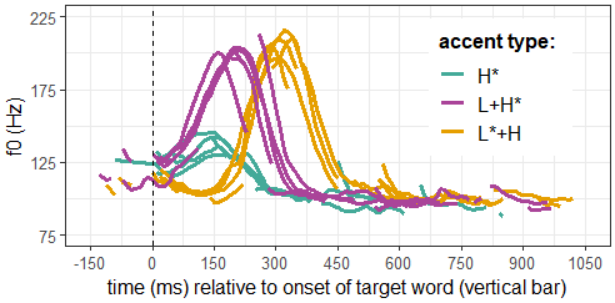


Figure 1: Individual f0 traces of all test stimuli temporally aligned at onset of target word (vertical bar).

In the perception study, each of the 15 test sentences (5 lexically different sentences * 3 accent types) was rated in relation to two different contexts (presented orthographically) eliciting either a contrastive focus (2) or a verum focus (3) reading on the test sentences (capitals reflect implicit prosodic highlighting).

(2) **Lexical contrast (context):**

a. *Du MACHST doch URLAUB auf Rügen.*
"You TAKE HOLIDAYS on Rügen."

b. *Du UNTERRICHTEST doch Spanisch.*
"You TEACH Spanish."

c. *Du KAUFST doch dein Kostüm.*
"You BUY your costume."

d. *Du ARBEITEST doch in Köln.*
"You WORK in Cologne."

e. *Du MUSSTEST doch den Hund füttern.*
"You HAD to feed the dog."

(3) **Verum focus (context):**

a. *Du LEBST doch gar nicht auf Rügen.*
"You don't LIVE on Rügen."

b. *Du LERNST doch gar kein Spanisch.*
"You don't LEARN Spanish."

c. *Du NÄHST dein Kostüm doch gar nicht.*
"You don't SEW your costume."

d. *Du WOHNST doch gar nicht in Köln.*
"You don't RESIDE in Cologne."

e. *Du WOLLTEST den Hund doch gar nicht füttern.*
"You didn't WANT to feed the dog."

The combination of context and target sentences results in short dialogues, e.g. (4), in which the target sentence (underlined) is the answer (B) to a preceding assertion (A). The assertions' finite verbs were written in capital letters indicating the intended position of a focal accent (implicit prosody) to facilitate the different intended focus readings on the target sentences.

(4) a. **Lexical contrast (experimental condition):**

A: *Du KAUFST doch dein Kostüm.*
B: *Nein. Ich nähe mein Kostüm.*

b. **Verum focus (experimental condition):**

A: *Du NÄHST dein Kostüm doch gar nicht.*
B: *Doch! Ich nähe mein Kostüm.*

In the lexical contrast conditions, e.g. (4a), it is the finite verb that is being contrasted. Updating the

lexical value of the finite verb results in a contrastive focus structure (B) with an accent on this verb. The denial of the assertion is additionally expressed by the negation *Nein* preceding the target sentence.

In the verum focus conditions, e.g. (4b), the assertion (A) implies a strong presupposition concerning the negative truth value of the target sentence's proposition. This leads to a verum focus structure in the target sentence (B) with an accent on the finite verb updating the truth value of the whole proposition. The affirmative answer of the negated proposition is additionally enhanced by the particle *Doch* (meaning roughly "Yes, on the contrary!") preceding the target sentence.

2.2. Procedure

The study was web-based, implemented with the *SoSci Survey* software and made available to participants at www.soscisurvey.de [24].

The participants' task was to evaluate how well the melody of a test stimulus matched a corresponding context. They indicated their judgements by choosing one value on a five-point Likert scale with discrete verbal references. That is, each point of the "appropriateness" scale has a nominal label representing a possible response ranging from totally inappropriate to very appropriate (Fig. 2). The distance between each value is supposed to be approximately equal. The responses were measured as a numerical variable ranging from one to five (there is no absolute zero point). Hence, higher ratings reflect a higher degree of perceived appropriateness (see Fig. 2, bottom).

The evaluation was carried out for each test stimulus on a separate page. Both types of context containing the same target sentence were displayed simultaneously, each with its own scale for evaluation (see Fig. 2, top). While the context and target sentences were presented orthographically, only the target sentence was presented acoustically. Participants had control over when and how often to play a stimulus.

To avoid systematic order effects by the presentation of the contexts, two experimental lists were used with a different order of context presentation. Each participant saw only one of the two lists. Moreover, each stimulus was rated twice. Therefore, the set of stimuli was presented for a second time after all stimuli had been rated once. Stimuli were presented in random order within each set. Accordingly, for each participant 60 evaluations (5 target sentences * 3 accent types * 2 contexts * 2 repetitions) entered the analysis. At the beginning of the study, participants were familiarized with the experimental procedure by means of a short practice part including five stimuli.

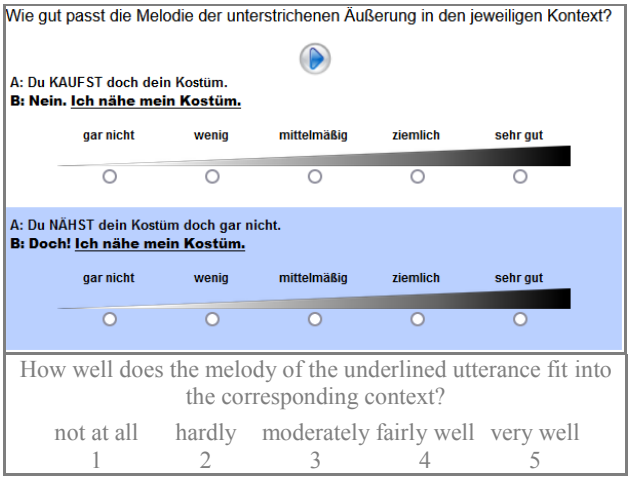


Figure 2: Example of the appropriateness rating task (top) with English translation of the task and labels (bottom).

2.3. Participants and Statistical Analysis

Seventy-seven native speakers of German (66% female, 32% male; 9% bilingual) participated in the study. They were first-year students at the IfL Phonetik at the University of Cologne aged between 19 and 36 years (mean = 22.2 years, SD = 3.2).

Since the elicited evaluations are ordinal ratings, we calculated mixed effects ordinal logistic regression models on the appropriateness ratings using the *clmm()* function from the "ordinal" package [25] in R [26]. In total, 4620 observations (77 participants * 60 evaluations) entered the analyses. The statistical models included FOCUS (lexical contrast, verum focus) and ACCENT (H*, L+H*, L*+H) as fixed factors and assumed random intercepts and slopes for FOCUS and ACCENT by participants and items (five target words).

3. RESULTS

The distributions of the ratings on the five-point Likert scale (Fig. 3) clearly show differences in the perceptual appropriateness of different accent types as markers of lexical contrast and verum focus.

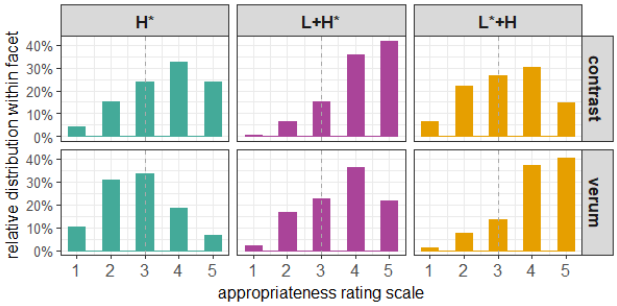


Figure 3: Relative distribution of appropriateness ratings on five-point Likert scale per condition.

Likelihood ratio tests on the appropriateness ratings reveal a significant effect of ACCENT ($\chi^2 = 13.044$, $p < 0.01$) as well as an interaction of ACCENT and FOCUS ($\chi^2 = 604.01$, $p < 0.001$). The mean rating scores (Fig. 4, left panel) display the main differences at a glance: H* and L+H* accents receive higher scores when marking lexical contrast, while L*+H accents receive higher scores when marking verum focus. Moreover, results reveal the following prosodic preferences for the two focus types (indicating a decrease in appropriateness from left to right).

Lexical contrast:

L+H* (mean = 4.10, SD = 0.95)	>	H* (mean = 3.57, SD = 1.14)	>	L*+H (mean = 3.26, SD = 1.15)

Verum focus:

L*+H (mean = 4.08, SD = 0.97)	>	L+H* (mean = 3.59, SD = 1.08)	>	H* (mean = 2.81, SD = 1.07)

Pairwise comparisons confirm that the ratings for the three accent types significantly differ from each other overall (with $p < 0.001$ each), but also within and across focus conditions (with $p < 0.05$ for contrast/H* vs. contrast/L*+H and for contrast/H* vs. verum/L*+H and with $p < 0.001$ for the other comparisons), except for the comparisons of contrast/H* vs. verum/L+H*, contrast/L+H* vs. verum/L*+H and contrast/L*+H vs. verum/L+H* that do not differ significantly.

Hence, L+H* accents prove to be appropriate prosodic markers of both focus types, being best suited to mark lexical contrast. However, in the verum focus condition L*+H accents are clearly preferred as prosodic markers, while they are least adequate to mark lexical contrast. The perceptual scores for H* accents take an intermediate position: They are rated as fairly appropriate when marking lexical contrast but are, by far, least adequate as markers of verum focus.

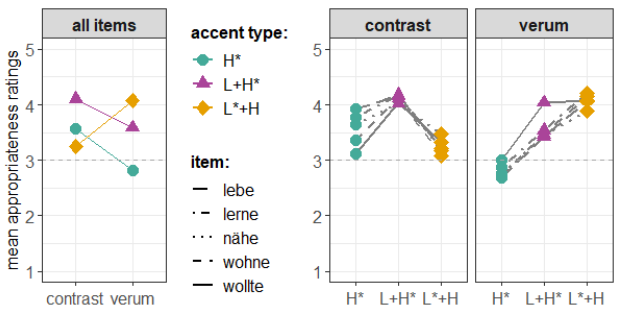


Figure 4: Mean appropriateness ratings for all stimuli (left panel) and per item/target word (right panels).

The right panels in Figure 4 show that the overall preferences in prosodic marking of the two focus types hold for all target sentences apart from a few exceptions. For instance, in target sentence (1e) containing the modal verb (*wollte*, solid line) L+H* and L*+H accents are rated equally appropriate as markers of verum focus. Interestingly, H* accents show a high variation in appropriateness for the different target sentences when marking lexical contrast but not when marking verum focus.

4. DISCUSSION AND CONCLUSIONS

In this study we investigate the perceptual appropriateness of intonational rises as prosodic markers of lexical contrast and verum focus. Results reveal a preference for L+H* over H* and L*+H accents for lexical contrast and a preference for L*+H over L+H* and H* accents for verum focus. Interestingly, in both conditions, L+H* was consistently preferred over H*, indicating that both lexical contrast and verum focus require prominent accents, but also that the mapping of function to accent type is not one-to-one.

Although both H* and L+H* involve a rising onglide (excursion up to the accented syllable), L+H* involves a larger onglide than H*. The rising onglide has been found to be a perceptual cue to contrast [16, 17, 27, 28], with a larger onglide being perceived as more prominent [23], in turn increasing the likelihood of a contrastive interpretation. The present study reveals that both lexical contrast and verum focus can be produced with this cue to contrast, indicating that the two focus types could indeed be related.

However, while the preference of L+H* over H* can be primarily attributed to a difference in f0 *scaling* (and prominence), the difference between L*+H and the other two accents applies additionally to the dimension of f0 *timing*, i.e. the rise of the accent in L*+H is later than in H* and L+H* (but, in our stimuli, the pitch excursion in L+H* and L*+H are comparable). The timing difference in L*+H appears to function as a marker of verum focus, more clearly than of lexical contrasts (see also [17, 29] for timing differences in focus marking). In fact, the use of L*+H (or a late peak accent) may express "a personal affective evaluation to a contrast" [30] made by the speaker, in the case of verum focus in the form of a protest or explicit objection to what has been claimed before.

To conclude, both types of focus investigated here show aspects of contrastive intonation, with differences in pitch scaling and timing mapped onto three distinct accent categories. Our perception data suggest that listeners use these prosodic dimensions to detect functional differences in the marking of lexical contrast and verum focus.

5. ACKNOWLEDGEMENTS

This research has been funded by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation) – Project-ID 281511265 – SFB 1252 "Prominence in Language" (project A01) at the University of Cologne. We would like to thank Horst Lohnstein (University of Wuppertal, Germany) for inspiring us to work on verum focus and discussion of the stimuli as part of a former collaborative project also funded by the *DFG* (Project-ID 144877651).

6. REFERENCES

[1] Bishop, J. 2012. Information Structural Expectations in the Perception of Prosodic Prominence. In: Elordieta, G., Prieto, P. (eds), *Prosody and Meaning* (Interface Explorations). Mouton de Gruyter, 239–270.

[2] Ladd, D. R. 1980. *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press.

[3] Rooth, M. 1992. A Theory of Focus Interpretation. *Natural Language Semantics* 1, 75–116.

[4] Krifka, M. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica* 55, 243–276.

[5] Höhle, T. 1992. Über Verum-Fokus im Deutschen. In: Jacobs, J. (ed), *Informationsstruktur und Grammatik* (Linguistische Berichte, Sonderheft 4). Westdeutscher Verlag, 112–141.

[6] Lohnstein, H. 2012. Verumfokus – Satzmodus – Wahrheit. In: Lohnstein, H., Blühdorn, H. (eds), *Wahrheit – Fokus – Negation* (Linguistische Berichte, Sonderheft 18). Buske, 31–67.

[7] Lohnstein, H. 2016. Verum focus. In: Féry, C., Ishihara, S. (eds), *The Oxford Handbook of Information Structure*. Oxford University Press, 290–313.

[8] Gutzmann, D. 2012. Verum – Fokus – Verum-Fokus? Fokus-basierte und lexikalische Ansätze. In: Lohnstein, H., Blühdorn, H. (eds), *Wahrheit – Fokus – Negation* (Linguistische Berichte, Sonderheft 18). Buske, 67–103.

[9] Gutzmann, D., Hartmann, K., Matthewson, L. 2020. Verum Focus is Verum, not Focus: Cross-linguistic Evidence. *Glossa: a journal of general linguistics* 5(1): 51, 1–48.

[10] Grice, M., Baumann, S., Benzmüller, R. 2005. German Intonation in Autosegmental-Metrical Phonology. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 55–83.

[11] Féry, C. 1993. *German Intonational Patterns.* Niemeyer.

[12] Alter, K., Mleinek, I., Rohe, T., Steube A., Umbach, C. 2001. Kontrastprosodie in Sprachproduktion und –perzeption. In: Steube, A., Umbach, C. (eds), *Kontrast – lexikalisch, semantisch, intonatorisch*. Linguistische Arbeitsberichte 77, 59–79.

[13] Steube, A. 2001. Correction by Contrastive Focus. *Theoretical Linguistics* 27, 215–249.

[14] Baumann, S., M. Grice & S. Steindamm (2006). Prosodic Marking of Focus Domains - Categorical or Gradient?. *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden. 301–304.

[15] Hermes, A., Becker, J., Mücke, D., Baumann, S., Grice, M. 2008. Articulatory Gestures and Focus Marking in German. *Proc. 4th Speech Prosody* Campinas, Brasil, 457–460.

[16] Ritter, S., Grice, M. 2015. The Role of Tonal Onglides in German Nuclear Pitch Accents. *Language and Speech* 58(1), 114–128.

[17] Grice, M., Ritter, S., Niemann, H., Roettger, T. 2017. Integrating the Discreteness and Continuity of Intonational Categories. *Journal of Phonetics* 60, 90–107.

[18] Turco, G., Dimroth, C., Braun, B. 2012. Intonational Means to Mark Verum Focus in German and French. *Language and Speech* 56(4), 461–491.

[19] Turco, G., Braun, B., Dimroth, C. 2014. When Contrasting Polarity, the Dutch use Particles, Germans Intonation. *Journal of Pragmatics* 62, 94–106.

[20] Grabe, E. 1998. Pitch Accent Realization in English and German. *Journal of Phonetics* 26, 129-143.

[21] Rathcke, T. 2013. On the Neutralizing Status of Truncation in Intonation: A Perception Study of Boundary Tones in German and Russian. *Journal of Phonetics* 41(3-4), 172–185.

[22] Lohnstein, H., Stommel, H. 2023. Fokus-DB [language database]. http://www.fokus-db.de/

[23] Baumann, S., Röhr, C. T. 2015. The Perceptual Prominence of Pitch Accent Types in German. *Proc. 18th ICPhS* Glasgow, UK, vol. 298, 1–5.

[24] Leiner, D. J. 2019. SoSci Survey (Version 3.1.06) [Computer software]. https://www.soscisurvey.de

[25] Christensen, R. H. B. 2022. ordinal—Regression Models for Ordinal Data (R package version 2022.11-16). https://CRAN.R-project.org/package=ordinal

[26] R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/

[27] Kügler F., Gollrad, A. 2011. Production and Perception of Contrast in German. *Proc. 17th ICPhS* Hong Kong, 1154–1157.

[28] Roessig S., Mücke, D., Grice, M. 2019. The Dynamics of Intonation: Categorical and Continuous Variation in an Attractor-based Model. *PLoS ONE* 14(5), e0216859.

[29] Kügler, F., Genzel, S. 2016. Effects of Information Structure, Syllable Structure, and Voicing on Nuclear Falling Pitch Accents in German. *Proc. 8th Speech Prosody*, Boston, USA, 678–682.

[30] Kohler, K. J. 2006. What is Emphasis and How is it Coded? *Proc. 3rd Speech Prosody* Dresden, Germany, 748–751.

PROSODY IN CONTACT: POLAR QUESTIONS IN CYPRIOT VARIETIES OF GREEK AND TURKISH

Elinor Payne[a], Spyros Armostis[b], Eftychia Lombardo[b], Rainer Simon[c], Mary Baltazani[a]

[a]University of Oxford, [b]University of Cyprus, [c]Austrian Institute of Technology
elinor.payne@phon.ox.ac.uk, armosti.spyros@ucy.ac.cy, lombardo.e@unic.ac.cy, Rainer.Simon@ait.ac.at, mary.baltazani@phon.ox.ac.uk

ABSTRACT

Cypriot varieties of Greek (CYG) and Turkish (CYT) have been in close contact for many centuries, leading to mutual influence (e.g. lexical borrowings), despite belonging to different language families, and being typologically distinct. Little empirical research has been conducted on prosodic influence between contact languages. Some analyses of CYT polar questions report similarities with CYG, but do not give quantitative results. Using quantitative modelling of intonational contours, we compare CYG polar questions with those in CYT (geographically close), and with those in Athenian Greek (ATG) (phylogenetically close). We confirm previous research showing CYG questions are phonologically very similar to ATG, albeit with phonetic differences. We also find these patterns in CYG speakers' L2 English. We then compare CYT questions with those in Istanbul Turkish (IST) and find, in contrast, both Greek-like variants (presumably due to contact with CYG), and more Turkish-like variants, with distribution likely conditioned by sociolinguistic factors.

Keywords: Cypriot Greek, Cypriot Turkish, intonation, polar questions, convergence

1. INTRODUCTION

Within the area of contact linguistics, research on prosodic contact has received relatively less attention than other aspects of language, especially where intonation is concerned. This is partly due to certain challenges particular to prosody, rendering it arguably more elusive to analysis, e.g. lack of orthographic representation, which in turn means lack of historical textual evidence, and until recently a lack of quantitative methods for capturing more gradient properties of intonational contours.

However, recent methodological advances mean that we can derive more robust forms of evidence that ongoing language contact results in intonational variation and change in varieties spoken in multilingual societies. Contact may result in either phonological or phonetic prosodic transfer [1], influencing the category of phonological tonal events (pitch accents and edge tones) or their phonetic realization. For example, [2] show the influence of Italian early-peak H tone alignment on Buenos Aires Spanish, and [3] shows contact effects of Quechua on Cuzco Spanish. [4] report novel patterns arising from contact between Basque and Spanish. [5] documents codeswitching in the polar questions (PQs) of German-Turkish bilinguals, and [6] report that speakers of Asia Minor Greek exhibit a blend of Greek-like and Turkish-like patterns in continuation rise utterances.

1.1. Cyprus as a prosodic melting pot?

The linguistic picture in contemporary Cyprus is dense and complex, resulting from multiple linguistic contact situations owing to a combination of geo-political factors both of considerable historical depth (e.g. Ottoman, Venetian, British rule, settlement of Maronites, Armenians) and pertaining to recent times (internal population displacement around the events of 1974). In addition to Cypriot varieties of Greek and Turkish, which each exist in a diglossic relationship with standard varieties, there are long-standing minority communities of Cypriot Arabic, Western Armenian and Kurbetcha, and high proficiency in English. As such it makes an ideal case study for investigating the existence and dynamics of prosodic convergence within a multilingual society.

1.2. Polar questions in Greek vs Turkish varieties

We focus on PQs because despite notable differences between ATG (nucleus L* plus H-L% edge tones) and IST (nucleus H* plus either L-L% or L-H% edge tones, though reports for the latter vary and more research is needed), impressionistic accounts [7] report similarities between CYG and CYT.

A further distinguishing feature is the use of the question particle {mI} in IST. According to [7: 72], this attaches to the question focus, and if there is no specific focus, post-clitically to the finite verb (normally phrase-final). [8: 161] states that, where the particle is used, the intonation is identical to that of statements (both use an L tone finally). When the particle is final, "the verb in front of the {mI} particle has the highest tone. However, the [position of] the

question particle [...] can change depending on the focus word, and the highest tone moves from verb to focus word." [9] adds that there is only a fall (i.e. L tone) on the particle if post-focal, and otherwise there is a rise. CYT PQs are reported as differing from IST, in both intonation and reduced use of the question particle [7], but empirical research is scant.

ATG and CYG PQs are reported to have the same structure, L* NPA, H- L% but a difference in the alignment of the H phrase accent when focus is early in the phrase, in which case the ATG H aligns with the last stressed syllable but the CYG one aligns with the final two syllables (for ATG: [10-13]; for CYG: [14-17]).

The ATG PQ tune also transfers into L2 English PQs by ATG intermediate learners ([18]) who produce the full set of Greek tonal events, including tonal alignment, speech rate, pitch span and pitch level. We expect similar transfer in L2 English PQs by CYG speakers. In this paper we compare PQs produced by speakers of ATG, CYG, L2 English produced by CYG speakers, IST and CYT.

2. METHOD

2.1. Participants, materials and procedure

We elicited and analysed productions of PQs from the varieties of interest: 7 speakers of Athenian Greek (ATG; 6F, 1M), 7 speakers of Cypriot Greek (CYG; 5F, 2M) from Nicosia, 4 speakers of Cypriot Turkish (CYT; 3F, 1M) from Nicosia, 1 speaker of Istanbul Turkish (IST; F), and English as spoken by 2 of the Cypriot Greek speakers (CYG-Eng; 1F, 1M). Participants were aged 18-30, and completed a detailed sociolinguistic questionnaire eliciting background information.

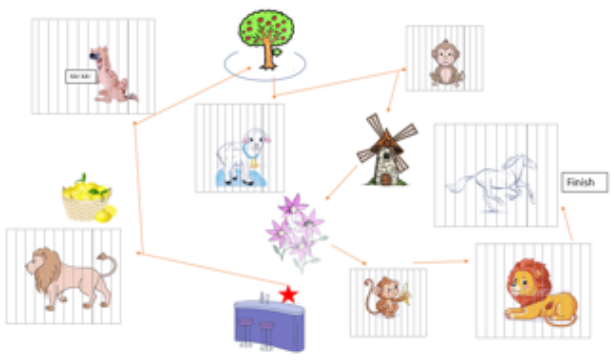


Figure 1: Zoo map.

PQs were elicited through a map task based on the premise that the participant and experimenter (native speakers of their language variety) had lost each other on a visit to the zoo, and were communicating via mobile phone. Both had a map of the zoo, but only the experimenter's map showed the orange path (Fig. 1). Participants were instructed to ask yes-no questions to find the experimenter, e.g. "*Do I go towards the lion?*".

2.3. Data segmentation and labelling

For each language variety, all instances of PQs were identified, checked for naturalness, orthographically annotated and translated into English by a native speaker (359 tokens in total). Judgments were made on the positioning of the nuclear focus, dividing the extracted PQ phrases into two categories: early and final nuclear focus. Consonants and vowels were segmented and labelled for the nuclear stressed syllable and any post-nuclear syllables. We analysed PQs over a Region of Interest (ROI), from the beginning of the nuclear vowel to the end of the utterance. The same ROI was defined across all language varieties for maximal comparability.

The results revealed two patterns in the CYT PQs, one resembling the Greek pattern and one resembling the Turkish pattern. We labelled these two patterns as CYT-G and CYT-T. Inspection of the questionnaire revealed the two participants with the Turkish-like tune had recently studied in Turkey.

For L2 English, judgments about naturalness and location of focus are more difficult to make, since a) the intended target is not certain, and b) proficiency is variable. Thus, these were examined and agreed upon by a consortium of annotators.

2.4. f_0 modelling, hypotheses and comparisons

For each utterance f_0 was measured every 10 ms using ESPS *get_f0* [19]. 10th-order polynomials $\hat{f_0} = \Sigma a_n t^n$ for $n = 0, \ldots 10$, were fitted to f_0 contours using the GNU Octave [20] *polyfit* function; pitch errors were inspected and manually corrected. The shape of f_0 contours in the ROI was modelled as 4th-order polynomials $\hat{f_0} = \Sigma a_n t^n$ for $n = 0, \ldots 4$, which were then transformed into orthogonal (Legendre) polynomials $\Sigma\ c_n L_n$ (cf. [6]). The five c_n coefficients capture general shape characteristics of the fitted $\hat{f_0}$ contour: c_0 is the average f_0 height of the contour; c_1 is its slope; c_2 models the shape as a parabola, concave up (or down if the sign is negative); c_3 models the shape as an *N*-like wave with a peak followed by a trough (or the reverse if the sign is negative); and c_4 models the shape as a more complex *M*- or *W*-like wave with more than one peak and trough.

The time of the f_0 maximum was determined by root-finding (i.e. calculating when the first differential of the modelled contour equals zero), using the GNU Octave/Matlab function *real(roots(polyder(a)))*. We define *peak alignment* as

the difference between the H peak time and the end of the nuclear vowel.

For the comparison across language varieties, a separate univariate Anova test was used for each of 6 parameters as dependent variable (the alignment of the H peak as well as the shape of f_0 contours in the ROI through the five c_n coefficients) and language as the independent variable with 6 levels, one per language variety.

Our hypotheses were: 1a) CYG questions resemble ATG questions phonologically (i.e., share the L* NPA, H- L% tune) but 1b) differ phonetically in the alignment of the H phrase accent (see 1.2); 2) the CYG question pattern transfers onto the L2 English of CYG speakers; 3) CYG questions influence CYT in: (a) the alignment of the H phrase accent, so the f_0 maximum (i.e., H) will occur *after the end* of the nuclear vowel as in Greek and not *within* the nuclear vowel as in Turkish; and (b) the f_0 shape, revealed mainly through coefficient c_3: the Greek-like pattern is a trough followed by a peak (i.e., a L* NPA followed by H-L% edge tones) so its c_3 coefficient is expected to be negative, while the Turkish-like pattern is the reverse, so its c_3 coefficient is expected to be positive.

3. ANALYSIS AND RESULTS

Figure 2 gives representative examples of PQs as produced by speakers of the six varieties under investigation. Red boxes show the location of the nuclear stressed vowel. In addition, the black boxes for the two examples in the top row highlight the phonetic difference between ATG and CYG.

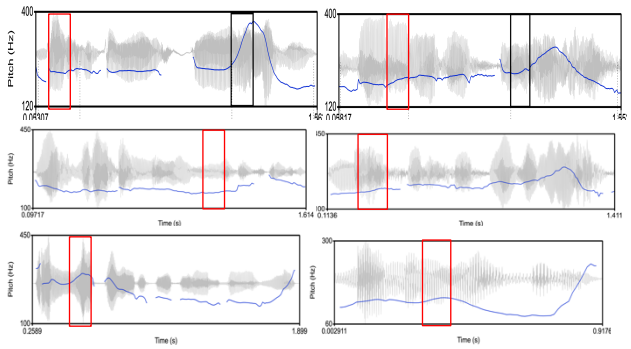


Figure 2: Examples of PQs for top: ATG, CYG middle: CYT-G, CYG-Eng; bottom: IST, CYT-T

ATG	CYG
Mporo na pao sto alogo? *Can I go to the horse?*	Mporo na pao sto alogon? *Can I go to the horse?*
CYT-G	CYG-E
Zizzirodan maymuna gidecem? *Should I go to the monkey from the cricket?*	Now can I go to the mill?
IST	CYT-T
Ayıya doğru bir yol var mı? *Is there a way to the bear?*	Maymunun yanında mın? *Are you next to the monkey?*

Table 1: Utterances shown in Figure 2.

Comparisons between ATG and CYG confirmed hypothesis 1a), that ATG and CYG share the L* NPA and H-L% edge tones, and 1b) have different alignment of the H phrase accent, which falls later in CYG (see Fig. 2 top panel; $F(1, 222) = 4.2$, $p = 0.042$). The results also suggest differences between ATG and CYG in the placement of focus, with CYG (and CYG-Eng) opting more commonly for final focus, and early focus being more common in ATG (ATG: 79 early and 28 final focus tokens vs. CYG: 28 early and 98 late tokens).

Figure 3 illustrates this phonological similarity between ATG and CYG: the L* for both these varieties of Greek occurs *within* the nuclear vowel (i.e., before the end of the stressed vowel, indicated by zero on the vertical axis) and the H- phrase accent occurs *after* the nuclear vowel (i.e., after the end of the stressed vowel indicated by zero on the horizontal axis). Note that in some tokens the H occurs before the end of the stressed vowel in Fig. 3. These correspond to words with final stress where all tones (L*, H-, L%) are crowded within the same vowel, due to lack of further segmental material.

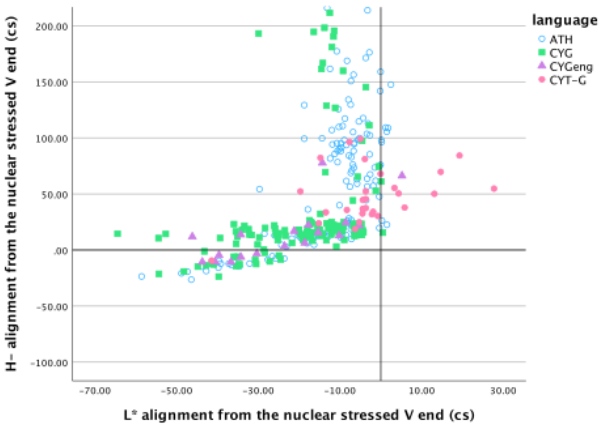


Figure 3: The alignment of the L* nuclear pitch accent and of the H- phrase accent, both in relation to the end of the nuclear vowel (indicated by 0).

In support of hypothesis 2, the results also show transfer of the Greek pattern to L2 English for CYG speakers (i.e., no difference in the tone alignment or shape), see Fig. 3.

Hypothesis 3 was also supported, as the influence of CYG on CYT PQ patterns was quantitatively established, while in addition the importance of sociolinguistic factors was brought to light. Notably, two of the four speakers (indicated as CYT-T, orange squares in Figure 4) produced questions with a Turkish-like tune, while the remaining two (CYT-G) produced Greek-like questions (pink circles in Figure 4). We note also that when speakers produce the Turkish-like tune, the question particle is optional, but when they use the Greek-like tune, the particle is never used. These patterns in CYG-Eng and CYT-G, constitute evidence for *phonetic* transfer; with regard to *phonological* interpretation, further investigation will be needed, e.g. through perceptual experiments.

As can be seen in Figure 4, both the alignment of the H tone ($F(5, 353) = 6.9$, $p < 0.001$) and the f_0 shape, as shown through coefficient c_3 ($F(5, 353) = 23.2$, $p < 0.001$), contribute towards the differentiation between the language varieties. As the nuclear vowel aligns with a trough in ATG but with a peak in IST, the H tone aligns *after* the end of this vowel in the Greek-like tune (CYT-G), indicated by 0 in the y-axis in Figure 4, but *before* it in the Turkish-like tune (CYT-T). The varieties displaying the Greek-like pattern (ATG, CYG, CYG-Eng, CYT-G) have a negative c_3 coefficient in the x-axis, i.e. the trough precedes the peak, while those with the Turkish-like pattern (IST, CYT-T) have a positive one, that is, the peak of the contour precedes the trough.

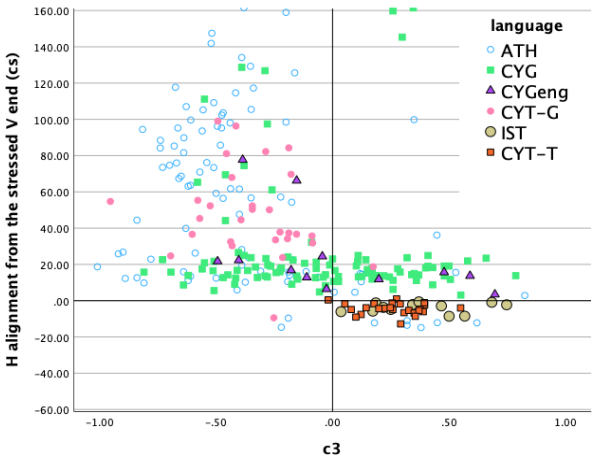


Figure 4: Coefficient c3 plotted against H tone alignment (time lag from the end of the nuclear vowel, indicated by 0 in the y-axis, to the H peak).

4. DISCUSSION AND NEXT STEPS

The results confirm previous findings that CYG PQ intonation is *phonologically* the same as ATG, but with different *phonetic* alignment (evident when focus is early in the phrase). These properties of CYG are also transferred into some speakers' productions of L2 English.

The study also reveals, for the small sample analysed so far, a bimodal distribution of contour types for CYT, with some tokens showing Turkish-like intonation, and others showing Greek-like intonation (at least phonetically). The presence of such Greek-like structures points strongly to some degree of prosodic convergence on the island, in line with [7]. Further data will enable us to determine how far such productions resemble CYG with regard to their precise phonetic *alignment*, and where they sit in the wider range of Greek and Turkish varieties under examination. This will help establish the nature of the outcome of contact on the island, and whether we can identify a cluster of innovations that are intermediate between the two 'source' languages. This in turn raises questions about the phonological status of these variants and their sociolinguistic distribution, to be investigated through perceptual experimentation. An important factor in analysing this will be to compare closely with non-Cypriot varieties of Turkish.

A further step will be to deepen this investigation with a larger pool of CYT speakers, including older speakers (who may have been less exposed to non-Cypriot varieties of Turkish), and speakers in other parts of northern Cyprus (e.g. Morphou and Kyrenia, where many displaced CYT-speakers from the south settled). Further investigation of CYG will also incorporate older speakers and geographical variation, with particular focus on the city and area around Paphos, where impressionistic accounts report distinct intonation and where many displaced CYG-speakers from northern Cyprus settled. Further evaluation of the nature of convergence will be conducted through perceptual testing and comparison with a third axis of contact, namely with Cypriot Arabic, an endangered language showing extensive influence of CYG, and examination of other intonational patterns such as continuation rises.

Finally, this study forms part of a larger project digitally mapping prosodic convergence in the wider linguistic area of the Eastern Mediterranean (https://mappingprosody.phon.ox.ac.uk/) and the extent to which historical and contemporary migrations help shape prosody.

7. REFERENCES

[1] Mennen, I. 2006. Phonetic and phonological influences in non-native intonation: an overview for language teachers. *Working Papers Queen Margaret University College* 9, 1–18.

[2] Colantoni, L., Gurlekian, J. 2004. Convergence and intonation: Historical evidence from Buenos Aires Spanish. *Bilingualism: Language and Cognition* 7(2), 107–119.

[3] O'Rourke, E. 2012. The realization of contrastive focus in Peruvian Spanish intonation. *Lingua* 122(5), 494–510.

[4] Elordieta, G., Calleja, N. 2005. Microvariation in Accentual Alignment in Basque Spanish. *Language and Speech* 48(4), 397–439.

[5] Queen, R. 2012. Turkish-German bilinguals and their intonation: Triangulating evidence about contact induced language change. *Language* 88(4), 791–816.

[6] Baltazani, M., Przedlacka, J., Ünal-Logačev, Ö., Logačev, P. & Coleman, J. 2023. Intonation of Greek in contact with Turkish: a diachronic study. *Language variation and change* 34(3), 271–303. doi:10.1017/S0954394522000126

[7] İmer, K., Çelebi, N. 2006. The intonation of Turkish Cypriot dialect: a contrastive and sociolinguistic interpretation. *International Journal of the Sociology of Language,* 181, 69–82.

[8] Demircan, Ö. 2001. *Turkish Prosody*. Yıldız Teknik University.

[9] Kamali, B. 2015. Information structure in Turkish yes/no questions. In: Zeyrek, D., Sağın Şimşek, C., Ataş, U., Rehbein, J. (eds.), *Ankara Papers in Turkish and Turkic Linguistics: Proceedings of the 16th International Conference on Turkish Linguistics*, 27–39. Wiesbaden: Harrassowitz Verlag.

[10] Arvaniti, A., Ladd, D. R., Mennen, I. 2006. Tonal Association and Tonal Alignment: Evidence from Greek Polar Questions and Contrastive Statements. *Language and Speech* 49(4), 421–450.

[11] Baltazani, M., Jun, S.-A. 1999. Topic and focus intonation in Greek. *Proceedings of the XIVth ICPhS*, vol. 2, 1305–1308.

[12] Baltazani, M. 2007. Intonation of polar questions and the location of nuclear stress in Greek. In: Gussenhoven, C., Riad, T. (eds.), *Volume 2: Experimental Studies in Word and Sentence Prosody*. De Gruyter Mouton, 387–406.

[13] Ladd, D. R. 1996. *Intonational phonology* (Cambridge Studies in Linguistics 79). Cambridge University Press.

[14] Arvaniti, A. 1998. Phrase accents revisited: comparative evidence from Standard and Cypriot Greek. *Proceedings of the 5th International Conference on Spoken Language Processing*, vol. 7. Sydney. Australia, 2883–2886.

[15] Grice, M., Ladd, R., Arvaniti, A. 2000. On the place of phrase accents in intonational phonology. *Phonology* 17(2), 143–185

[16] Themistocleous, C. 2011. *Prosody and information structure in Athenian and Cypriot Greek* (PhD Thesis). National and Kapodistrian University of Athens [In Greek].

[17] Themistocleous, C., Tsiplakou, S. 2013. High Rising Terminals in Cypriot Greek: Charting 'urban' intonation. In: Janse, M., Joseph, B., Ralli, A., Bagriacik, M. (eds.), *MGDLT 5: Proceedings of the 5th International Conference on Modern Greek Dialects and Linguistic Theory*. Patras: University of Patras, 445–562.

[18] Kainada, E., Lengeris, A. 2015. Native language influences on the production of second-language prosody. *Journal of the International Phonetic Association* 45(3), 269–287.

[19] Talkin, D. 1995. A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B., Palatal, K. K. (eds.), *Speech Coding and Synthesis*. Elsevier Science B.V., 497–518.

[20] Octave community. 2013. GNU Octave 3.7+. Available online at: http://www.gnu.org/software/octave/index.html

PROSODIC PREDICTORS OF TEMPORAL STRUCTURE IN CONVERSATIONAL TURN-TAKING

Kathryn Franich[1,2]

[1]Harvard University and [2]University of Delaware
kfranich@fas.harvard.edu

ABSTRACT

Fluid conversation depends on conversation partners' ability to make predictions about one another's speech in order to forecast turn ends and prepare upcoming turns. One model used to explain this process of temporal prediction is the coupled oscillator model of turn-taking [1]. A generalization that the model captures is the relative scarcity of interruption in turn-taking, as it predicts partners' turns should be counter-phased to one another, with minimal pause time between turns. However, in naturalistic conversation, turns are often delayed, rather than occurring in perfect succession. We hypothesize that these delays are not of arbitrary duration, but are structured in their timing, just as between turns with immediate transitions. We demonstrate that relative timing of prosodic events occurring at *turn ends* is key to modelling pause duration between turns, providing evidence that inter-turn pauses exist in a temporal trading relation with the final syllable and prosodic word of immediately preceding turn.

Keywords: Turn-taking, prosody, conversation, speech timing.

1. INTRODUCTION

Research on timing in conversational turn-taking supports the idea that transitions between turns generally take place so rapidly that conversation partners must be able to make *predictions* about turn ends in order to time their own turns appropriately. For example, Levinson & Torreira [2] point out that the average time to prepare and execute an utterance is around 600 ms, but the average pause time between speaker turns is much shorter, closer to 200 ms. The authors take this to mean that speakers must be projecting the end of their partner's turn well before it actually ends. Research has shown that several cues are predictors of turn ends, including falling or rising intonation patterns, segment or syllable lengthening, lower intensity, lexical cues, syntactic structure, and utterance completion, among others [3]-[9]; however, the specific way in which conversation partners time their turns in the context of these cues remains unclear. One influential model that has been proposed to capture conversational patterns is the coupled oscillator model of turn-taking developed by Wilson & Wilson [1]. The model proposes that turns are timed to cycles of *readiness* based on speakers' syllable rate. Readiness is at its highest towards the end of the syllable and at its lowest in the middle of a syllable. Counter-phasing of listener-speaker syllable oscillations can be used to explain the fact that speakers rarely interrupt one another, rather waiting until their conversation partner is finished speaking to initiate their own speech. The fact that pause durations between turns tend to be relatively short—on the order of 100-300 ms—can be explained by a high frequency syllable oscillator which governs turn-taking.

1.1. Rhythmicity, turn latency, and turn ends

Despite a trend toward immediate turn transitions, turn latencies often exceed the duration of a single syllable, sometimes by quite a lot. Given that speakers need to be able to make turn-taking predictions even when their conversation partner delays a response, we might hypothesize that delayed turns still display a temporal structure that is consistent with the coupled oscillator model. One possibility is that longer turn latencies constitute several cycles of the syllable-level oscillator proposed by Wilson & Wilson. However, in languages like English where there is relatively high variability in the timing of syllables [10]-[12], the timing of syllables alone may not allow for precise predictions. Prior research on speech timing in turn-taking has found that variability in timing between stressed syllables is reduced at turn transitions between conversation partners, consistent with the idea of a greater degree of foot-based isochrony at turn transitions [13]. Greater isochrony at a turn end may help the listener to develop more fine-grained predictions about speech timing of their conversation partner, and hence to plan their own speech in a more consistent way. Additional research by Shattuck-Hufnagel and Turk [14] has identified the phrase-final syllable and the main stress syllable of the final prosodic word of a phrase as sites for final lengthening, indicating that this process—which also occurs at turn ends—is not limited to a single prosodic event. Thus, it may be that larger turn-final

constituents, such as the final prosodic word or foot, would be more stably timed events around which to plan turns. In this paper, we explore how well the timing of various turn-final prosodic structures can predict latency of a conversation partner's turn, with the goal of modelling in greater detail the temporal planning process of turn-taking.

2. METHOD

Five pairs (10 subjects total) of English speakers from the northeast and mid-Atlantic regions of the United States participated in a game of "Twenty Questions" in which participants took turns thinking of a person who was mutually known to the pair and having their partner ask yes/no questions until they could identify the person. In order to ensure that participants would have a sufficient number of mutual acquaintances to engage in the task, pairs of participants who already knew one another well were recruited. Pairs included three married couples (one male and one female each), one pair of female friends, and one pair of sisters (avg. age = 45 years). Participants within pairs had all known one another for at least five years.

Participants wore head-mounted microphones and were video and audio recorded in a room in the Phonetics and Phonology lab at the University of Delaware playing the game while sitting in chairs and facing one another. Partners took turns thinking of a person and guessing. Each pair played the game for approximately 15 minutes, during which time an average of 80 turns, or 'inter-pausal units' [3] (40 per person) were elicited. An example exchange is in (1).

(1) Partner 1: *Does this person live in Bóston?*
Partner 2: *Yes.*
Partner 1: *Is it a fámily member?*
Partner 2: *No.*

2.1. Data coding

Questions were coded in Praat [15] on a series of TextGrid tiers with intervals marked for phrase duration, duration between the onset of the final pitch-accented word (marked with acute accents in (1)) and the end of the phrase, duration of the final word, duration of the final foot, and duration of the final syllable, as well as duration of the pause between the end of the final word of the question and the beginning of the 'yes' or 'no' response. Response to the question was also coded, though data for turns of both responses is pooled in the present work due to sample size. In order to control for response structure as much as possible, data were trimmed so that only turns with responses of 'yes' or 'no' were included. Turns that exceeded 2 standard deviations in duration from the mean were excluded from the analysis, as many of these responses involved clear uncertainty on the part of the partner answering. Exchanges containing a disfluency on the part of either speaker were also excluded. The final dataset contained 370 turns.

3. RESULTS

A linear mixed effects model was used to evaluate how well inter-turn pause duration was predicted by the various prosodic variables, including 1) duration of the preceding inter-pausal unit (the question, which was usually a single phrase, e.g. "Is he a family member?"), 2) time from final pitch accent of the question to the end of the phrase, 3) duration of the phrase-final foot, and 4) duration of the phrase-final syllable. Two-way interactions between all variables and speech rate (syllables per second) were also included, and number of syllables in the phrase was also included as a fixed effect. A by-subject random slope for speech rate was included in the model. All continuous predictors were mean-centered.

As expected based on previous work [16] there was a significant negative relationship between pause duration and speech rate (β=-100.21, t=-4.32; p<0.001) and a positive relationship between pause duration and number of syllables in the phrase (β=-75.378, t=3.35; p<0.01). Among the prosodic variables, only the duration of the question-final word predicted inter-turn pause duration to a significant degree. Despite the overall trend toward shorter pauses at higher speech rates, there was a *negative* relationship between final word duration and inter-turn pause duration (β=-35.73; t=-2.10; p<0.05). In other words, as the duration of the final word increased, pause duration decreased (Fig. 1).

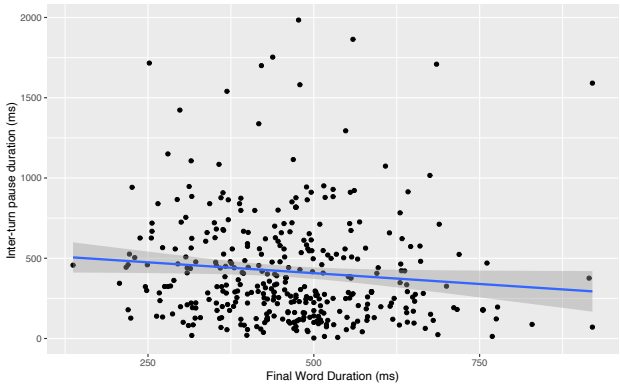


Figure 1: Duration of question-final word negatively predicts inter-turn pause duration

This result is consistent with a complementary (or 'trading') relationship between the final word and inter-turn pause, and may suggest that pauses constitute part of a larger planning unit which also includes the phrase-final prosodic word. Such a

pattern is reminiscent of findings on within-subject pause durations between sentences [18,19], where pause duration has been found to be in a complementary relationship with duration of phrase-final feet. Specifically, Fant & Kruckenberg report a multimodal distribution in pause durations in Swedish, such that the duration of a pause plus final lengthening equate to integer multiples of subjects' average foot (or inter-stress interval) durations.

Looking more closely at by-subject distributions of inter-turn pause latencies in our data (Fig. 2), many participants also displayed more than a single peak in latency—indeed, a Shapiro-Wilk normality test revealed that the data were not normally distributed (p<.001). Individual subject means for pause duration were also highly variable, ranging from 190 ms to 550 ms. Pairs of partners tended to have similar means—both subjects in Pair 3 (P3), in particular, had shorter latencies than the rest of the participants—though partners did not always pattern together, as demonstrated, for example, by P1.

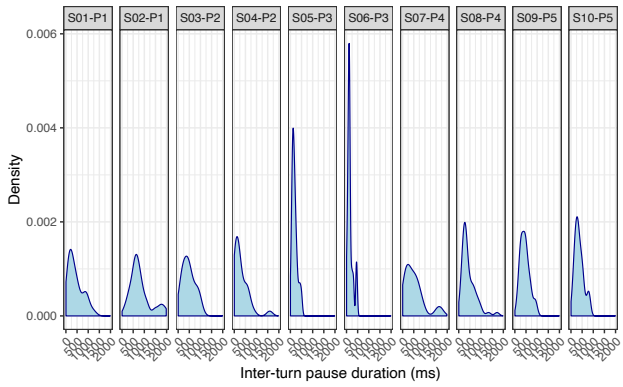


Figure 2: By-subject density plots of pause duration

Given these patterns of variation in turn latency and the potential for a complementary relationship between inter-turn pause (ITP) duration and question-final word duration, we hypothesized, similar to Fant & Kruckenberg, that inter-turn intervals might be more efficiently modelled as a *proportion* of the duration of the question-final word and the inter-turn pause interval added together (henceforth 'word+pause'). An equation deriving this measure (labelled Prop_Word) is provided in (2).

(2) Prop_Word = ITP Dur / (Final Word Dur + ITP Dur)

Specifically, we hypothesized that the duration of inter-turn pause intervals would exist in a quantal relationship to the word+pause duration, such that pause durations would cluster around lower order fractions, e.g. 1/3 or 1/2, of the word+pause interval.

For comparison, we calculated similar values for question-final syllable+pause ('Prop_Syllable'), question-final foot+pause ('Prop_Foot'), and final pitch accent to question end+pause ('Prop_PA'). Given the multimodal nature of the data, we fit the data to mixtures of Gaussian distributions with different number of components, using a parametric bootstrap of log-likelihood ratio statistics to evaluate the optimal number of components for the data for each measure. For all variables, bootstrap results revealed 2 components as the optimal number, indicating that pause proportions overall showed a bimodal distribution in the data.

To better understand the role of different prosodic structures in conditioning inter-turn pause durations, we fit the data for each prosodic pause proportion measure to three different 2-component models. Following a procedure from [20], in the first model (Mod1), means, mixing proportions, and standard deviations were chosen automatically in order to maximize model fit to the data. In the second model (Mod2), means were set to .25 and .5, to explore the goodness of fit of a model where pause duration was 1/4 or 1/2 of the word+pause duration.[1] In the third (Mod3), means were set to .33 and .67 to explore the goodness of fit of a model where pause duration was 1/3 or 2/3 of word+pause duration. Differences in log likelihood were then compared between the unrestricted model and the two more restricted models to evaluate which of the latter two models better captured patterns in our data. The same process was repeated for other prosodic variables; results for Models 1-3 are presented in Table 1.

	Log Likelihood				
	Mod1 (Unr.)	Mod2 (.25,.5)/ (.5,.75)	Mod3 (.33,.67)	Diff Mod1 vs. Mod2	Diff Mod 1 vs. Mod 3
Prop_ Word	102.02	99.67	94.38	2.34	7.64
Prop_ Syll	71.92	71.35	59.99	0.57	11.93
Prop_ Foot	98.23	95.41	89.08	2.82	9.15
Prop_ PA	89.25	68.66	66.48	20.59	22.77

Table 1: Log likelihoods of three mixture models fit to proportional measures of four prosodic variables

For all prosodic variables, the difference between Model 2 and Model 1 was smallest, indicating Model 2 was a better fit to the data than Model 3. Mod2 turned out to provide a very close fit to the Prop_Syllable data, with a difference in log likelihood to the best fit model of only 0.57. Mod2 also provided a close fit to the data for both Prop_Word and Prop_Foot, with differences in log likelihood from the best fit model at 2.34 and 2.82, respectively. Mod2 for Prop_PA was a considerably

worse fit to the data. Results for Mod2 for Prop_Syllable and Prop_Word are plotted in Figure 3. Histograms represent raw data, and gray density curves represent model-estimated components.

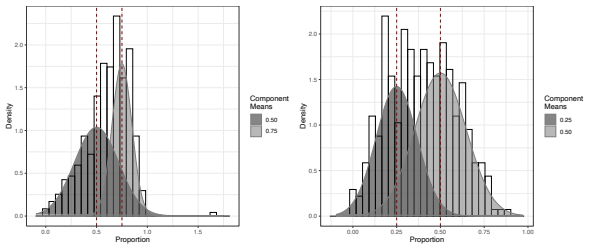


Figure 3: Results of Mod2 fit to Proportion_Syllable and Proportion_Word measures

In sum, it appears that inter-turn pause durations can be modelled efficiently as a proportion of a single unit spanning the pause and either the turn-final syllable or turn-final word. Specifically, it seems the pause tends to occupy either one-quarter or one-half of the duration spanning the final prosodic word and the pause, and one half or three-quarters of the duration spanning the final syllable and the pause.

5. DISCUSSION

Our results show some patterns which are consistent with Wilson & Wilson's coupled oscillator model of turn-taking, including that inter-turn pause duration was found to shorten overall as speech rate (measured in syllables per second) of the preceding utterance increased. This is in line with the idea of a syllable-level oscillator which regulates both speech rate and pause duration. The picture is complicated by the fact that the duration of the phrase-final prosodic word is negatively related to pause duration. One explanation for this pattern, in line with Wilson & Wilson's proposal, is that listeners plan their next turn based on the average syllable rate of their interlocutor's speech. As they anticipate the end of their partner's turn, they plan their pause based on a certain number of silent cycles of the syllable-level oscillator, which operates at this average frequency. This state of affairs predicts that, in the presence of more phrase-final lengthening (leading the oscillation rate of the partner's speech to be slowed relative to the average for the utterance), the inter-turn pause duration will be relatively shorter, since lengthening will cause the phrase-final word to extend further into the timing window projected by the listener for the pause. Under this account, the listener is essentially ignoring the presence of phrase-final lengthening for the purposes of pause planning.

The results of our mixture modelling procedure would suggest that pause duration is much more structured in its timing relative to phrase-final prosodic constituents, however. Specifically, our results suggest that pause time between turns tends to cluster around lower-order fractions of temporal units comprising turn-final prosodic constituents and the pause itself. Specifically, proportions tend to cluster around ¼, ½, and ¾ of the duration from the start of a turn-final syllable+pause or prosodic word+pause.

These results suggest that listeners are more sensitive to the durations of phrase-final constituents themselves, rather than simply ignoring the lengthening that applies to these constituents when planning pause duration. If this is the case, an account of the negative relationship between phrase-final word duration and pause duration could be that listeners plan for a certain number of oscillatory 'beats' based on the duration of the final prosodic word: if the word is shorter, they incorporate more silent beats, and if it is longer, they incorporate fewer. This may reflect the need for less turn-planning time where the phrase-final word is longer and affords the listener more syllables during which to plan their own speech. This proposal is in line with findings from Griffin [21], who shows that speakers utilize the time during articulation of their own speech to plan an upcoming word: in uttering pairs of nouns, for example, less silent planning time is needed at the start of the utterance if the first noun has a greater number of syllables during which the participant can plan articulation of the second noun. As previously mentioned, evidence suggests that listeners must plan their own turns during conversation while their interlocutor is finishing speaking [2,22]; if the listener's processing of the final word is mostly accomplished during the first syllable [23,24], then less attention needs to be paid to the remainder of the word, and this time can be used for planning the upcoming utterance. Of course, some pauses in our corpus were very long, and these longer durations are not likely to reflect solely lexical processing time. In these cases, we propose that participants are incorporating additional oscillatory 'beats' in order to consider their response to their partner's question, but still answer it within a predictable timeframe.

Our findings are in line with work by Fant and Kruckenberg [18] for speaker-internal pause timing, where the authors found evidence for complementary timing between phrase-final prosodic constituents and within-speaker pause durations, as well as a proportional relationship between pause durations and prosodic constituents like the foot. This work highlighted the importance of pauses in the temporal planning of speech. Our findings add to this body of research, and suggest that prosodic constraints on pausing behavior at the level of the individual are also active at the interpersonal level during conversational turn-taking.

6. ACKNOWLEDGEMENTS

Thank you to the study participants, to Karee Garvin for helpful feedback and insights, and Nicole Taylor for assistance with data collection. This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are the researcher's own.

7. REFERENCES

[1] Wilson, M. and Wilson, T.P. 2005. An oscillator model of the timing in turn-taking? *Psych. Bul. & Rev.* 12(6), 957–968.

[2] Levinson, S.C. and Torreira, F. 2015. Timing in turn-taking and its implications for processing models of language. *Front. Psych.*, 6, 731.

[3] Gravano, A. and Hirschberg, J. Turn-taking cues in task-oriented dialogue. 2011. *Comp. Speech & Lang*, 25, 601–634

[4] Corps, R.E., Crossley, A., Gambi, C., and Pickering, M.J. 2018. Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, 175, 77–95,

[5] Ford, C.E., and Thompson, S.A. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the projection of turn completion. In: Schegloff, E.A. and Thompson, S.A. (eds), *Interaction and Grammar*. Cambridge: Cambridge University Press, 135-184.

[6] Ferrer, L., Shriberg, E., and Stolcke, A. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003. Proceedings. (ICASSP '03).

[7] M. Atterer, T. Baumann, and D. Schlangen. 2008. Towards incremental end-of-utterance detection in dialogue systems. In *Proceedings of Coling*, Manchester, UK.

[8] Cutler, A. and Pearson, M. 1986. On the analysis of prosodic turn-taking cues. In Johns-Lewis, C. (ed), *Intonation and discourse*. London: Croom Helm, 139–155

[9] De Ruiter, J.P., Mitterer, H., and Enfield, N.J. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82, 515–535.

[10] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[11] Dauer, R.M. 1983. Stress-timing and syllable-timing re-analysed. *J. Phon.,* 11 51−62.

[12] Grabe, E. and Low, E.L. 2002. Durational variability in speech and the Rhythm Class Hypothesis. In: Gussenhoven, C. and Warner, N. (eds), *LabPhon 7*. Berlin: Mouton de Gruyter.

[13] Mooney, S. and Sullivan, G.C. 2015. Investigating an acoustic measure of perceived isochrony in conversation: Preliminary notes on the role of rhythm in turn transitions. *Selected Papers from New Ways of Analyzing Variation (NWAV)* 43, 129–135.

[14] Turk, A.E. and Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words. *J. Phon.* 35, 4, 445–472,

[15] Boersma, P. and Weenink, D. (2022*). Praat: doing phonetics by computer* [Computer program]. Version 6.3.03. http://www.praat.org/

[16] Schegloff, E.A. 1996. Turn organization: one intersection of grammar and interaction. In: Ochs, E., Schegloff, E.A. and Thompson, S.A. *Interaction and Grammar*. Cambridge: Cambridge University Press, 52–133.

[17] O'Dell, M.L., and Nieminen, T. 1999. Coupled oscillator model of speech rhythm. In: Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. (eds), *Proceedings of the XIVth ICPhS*, 2, 1075−1078.

[18] Fant, G. and Kruckenberg, A. 1996. On the quantal nature of speech timing. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA.

[19] Lea, W.A. *Trends in Speech Recognition*. Prentice Hall, Inc., 1980.

[20] Franich, K. 2021. Metrical prominence asymmetries in Medumba, a Grassfields Bantu language. *Language*, 97(2):365-402.

[21] Griffin, Z.M. 2003. A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psych. Bul. & Rev.* 10, 603-609.

[22] Sacks, H., Schegloff, E., and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language* 50, 696-735.

[23] Cutler, A. & Carter, D.M, (1987). The predominance of strong initial syllables in the English vocabulary. *Comp, Sp. Lang.*, 2, 133-142.

[24] Culler, A, & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Exp. Psy.: Hum. Perc. Perf.,* 14, 113-121.

[1] For final syllables, a slightly different model (Mod2a) was constructed to evaluate the fit of a model with component means at .5 and .75, since syllable duration is relatively shorter than our other measures and therefore pause duration may have constituted a relatively larger proportion of the syllable+pause interval.

LANGUAGE REDUNDANCY EFFECTS ON F0: A PRELIMINARY CONTROLLED STUDY

Cong Zhang[1], Catherine Lai[2], Ricardo Napoleão de Souza[2], Alice Turk[2], Tina Bögel[3]

[1]Newcastle University, UK [2]University of Edinburgh, UK [3]University of Konstanz, Germany
cong.zhang@newcastle.ac.uk; c.lai@ed.ac.uk; a.turk@ed.ac.uk;
R.N.deSouza@ed.ac.uk; Tina.Boegel@uni-konstanz.de

ABSTRACT

Previous research suggests that words with a high level of language redundancy (i.e. recognition likelihood from familiarity and predictability based on syntactic, pragmatic, and semantic factors) have reduced acoustic salience, such as shorter duration and reduced vowels. The Smooth Signal Redundancy Hypothesis proposes that acoustic salience is controlled via prosodic structure, and makes the prediction that parameters such as fundamental frequency should also be affected by language redundancy. This study investigates the relationship of F0 with lexical frequency, together with bigram (verb-adjective or adjective-noun) frequency and the ratio between these two bigram frequencies. Results from a carefully controlled experiment with quadruplets of minimal pairs suggests that language redundancy can affect fundamental frequency in English.

Keywords: Smooth Signal Redundancy Hypothesis, f0, frequency effects, prosodic structure, English

1. INTRODUCTION

In order to achieve efficient communication, speakers manipulate acoustic salience based on how redundant given linguistic units are in discourse [1]. Language redundancy encompasses a range of pragmatic, lexical, syntactic, and semantic factors. Acoustic salience refers to segment duration and spectral properties associated with hyperarticulation, as well as to acoustic properties associated with more salient prominence markers and stronger prosodic boundaries, including f0. Language redundancy correlates inversely with acoustic salience, so that the more redundant the linguistic item, the less acoustically salient it is, and vice versa [1, 2, 3, 4]. For instance, a speaker might highlight a rare or unexpected word (i.e., less redundant) by making it longer and/or by making it stand out in the speech stream through more pronounced boundary markings [5]. Since acoustic salience varies inversely with language redundancy, the information conveyed in an utterance is distributed more evenly, thereby maximising its likelihood of recognition. The Smooth Signal Redundancy Hypothesis [1, 5]proposes that prosodic structure is used to control acoustic salience.

While previous research suggests robust effects of language redundancy on duration measures [1, 6, 7], much less is known about its local effects on f0 (but see [4]). The current study tests the extent to which speakers manipulate f0 on strings of words that differ in their language redundancy profile. Given that f0 movement is one of the acoustic cues English speakers use to cue prosodic prominence via intonation contours (e.g., [8]), the Smooth Signal Redundancy Hypothesis (SSRH) predicts that redundancy should affect f0 in systematic ways.

In a study of spontaneous American English, [4] found that contextual plausibility (i.e., a measure of redundancy) affected f0 values as predicted by the SSRH: lower redundancy yielded overall higher f0 values. On the other hand, discourse mention and focus status showed less clear results, suggesting that redundancy might affect f0 differently from duration. Unlike the current study, however, lexical frequency measures were not manipulated in that investigation.

Using data obtained under tightly controlled experimental conditions, this study investigates the relationship between acoustic salience measures (prosodic prominence, boundary tone) and three measures of language redundancy: lexical frequency, bigram frequency (verb-adjective, adjective-noun), and the ratio between those two bigram frequencies.

2. METHODS

2.1. Materials

The materials used in this study were originally designed to examine the relationship between duration and language redundancy in [7]. Recordings were made in a sound-treated studio at the University of Edinburgh, at a sampling rate of 44.1 kHz, 16-bit.

A total of 23 participants (14F) participated in the reading task. All participants were students at the University of Edinburgh and received compensation.

Fourteen sets of well-balanced quadruplets were created for the study such as those exemplified in (1). Each utterance in a set included a Verb-Adjective-Noun (V-A-N) sequence. Each participant was randomly presented with the target sentences twice.

Three frequency measures were used as indicators for (non-acoustic) language redundancy:

a. Lexical frequency: The verbs and the nouns selected had either frequent (f) or infrequent (i) usages. The lexical frequencies were obtained from the WebCelex's Cobuild corpus [9]. For verbs, f corresponds to a raw number of occurrences above 2000, whereas i indicates fewer than 200. For nouns, f corresponds to raw frequencies over 3000, whereas i indicates raw frequencies below 100. A substantial buffer zone (V: 200-2000; N: 100-3000) was deliberately selected to ensure that the two categories were distant enough in their lexical frequencies. Frequent and infrequent verbs in each set both shared an identical rhyme, e.g. *make* vs *rake* in (1). On the other hand, frequent and infrequent nouns shared the same onsets and nuclei, e.g. *fields* vs *fiefs* in (1). The adjective between the verb and the noun was the same for all four in a quadruplet. In the examples below, frequency codes are labelled as subscripts on the verbs and the nouns.

(1) A quadruplet with frequent(f) and infrequent(i) verbs and nouns

ff: Whatever you **make**$_f$ **clean fields**$_f$ should be a priority
fi: Whatever you **make**$_f$ **clean fiefs**$_i$ should be a priority
if: Whatever you **rake**$_i$ **clean fields**$_f$ should be a priority
ii: Whatever you **rake**$_i$ **clean fiefs**$_i$ should be a priority

All sentences in the quadruplets were syntactically ambiguous. As a result, the adjective could be prosodically grouped with either the preceding verb (VA%N) or with the following noun (V%AN). No punctuation was given in the experiment to indicate prosodic phrasing to allow for participants' unbiased boundary placement. In the ICE-GB corpus [9] and the Brown corpus [7], the V%AN structure is much more frequent (ca. 77%) than the VA%N counterpart (ca. 4%).

b. Bigram frequency: Word bigram frequency was obtained through Google for the VA sequences and the AN sequences respectively. The bigram frequencies were categorized as "`high`" for frequencies above 60% of the total range (9 to 45 million occurrences), and "`low`" for data below 40%. A 20% buffer zone was kept to avoid borderline frequency categories.

c. Ratio of bigram frequencies: The ratio of the two bigram frequencies was calculated by dividing the bigram frequency of VA by the bigram frequency of AN, i.e. RATIO-BI = $Freq_{(VA)}$ /$Freq_{(AN)}$. The methods for normalizing and parting data followed the bigram frequencies'. The frequency ratio was coded with the higher value between $Freq_{(VA)}$ and $Freq_{(AN)}$.

2.2. Analysis

Participants were free to choose the parsing that they preferred for each utterance, for instance by placing the prosodic boundary either between V and AN (V%AN) or between the VA and N (VA%N). We only analyzed the utterances where all four utterances and both repetitions in a set were produced with the same prosodic phrasing. A total of 344 utterances (produced by 11 speakers, from 11 quadruplets) were then selected for the final dataset.

2.2.1. Data annotation

Segments in all utterances were automatically labelled using a customised version of the Montreal Forced Aligner [10]. Sonorant segments in target words (V, A, N) of each utterance were labelled on a different tier and used for data extraction. The intonation of all utterances was labelled following the original Tone and Break Indices [11] to inspect pitch accents, phrase-tones and boundary tones used in the production.

2.2.2. F0 extraction

Three analyses were conducted to examine language redundancy effects on f0. All f0 values were extracted using Praat [12] with a range of 75 Hz to 400 Hz using "Get mean" function in Praat.

Sonorant intervals of each target word (e.g. /meɪ/ in "make", /liːn/ in "clean") were divided into three equal portions for the analyses (henceforth *initial third*, *second third*, *final third*). Our test words were monosyllabic, and when pre-boundary, any pitch accents, phrase-tone, and boundary tone would all occur on the same syllable. Our rationale for dividing the syllable into three was that the f0 of the initial third might correspond more closely to pitch-accent-related f0, and the f0 of the last third might correspond more closely to boundary-related f0. We acknowledge that dividing a syllable into three parts based on f0 is a simplification and does not capture the full complexity of intonation patterns. However, we are using this division to provide some preliminary insights.

2.2.3. *Statistical analysis*

The data were divided into two datasets according to their prosodic grouping, i.e. V%AN (296 utterances in total) and VA%N (48 utterances in total). The data were analyzed using linear mixed effect models [13] in R [14]. The base model only included the predicted factors and their interactions, i.e. the lexical frequencies of V and N and their interaction (V-FREQ*N-FREQ). The interaction term was removed when insignificant and if a reduced model was a better fit (i.e. with a lower AIC). The duration of sonorant interval (SONORANT-DUR) was also added to test whether the f0 and duration are related. A full model includes all relevant factors as the main effects, including the lexical frequencies of V and N and their interaction (V-FREQ*N-FREQ), the duration of sonorant interval (SONORANT-DUR), the bigram frequency for VA and AN respectively (VA-FREQ, AN-FREQ), as well as the ratio between VA and AN (RATIO-BI). SPEAKER and the index of QUADRUPLET were used as random intercepts; when models failed to converge, QUADRUPLET was removed. The dependent variable in the current analysis is MEAN F0. We tested the mean f0 of the initial, second, and final third of the verb and the noun respectively in different statistical models.

3. RESULTS

Tune types did not seem to be associated with word frequency. In the V%AN dataset, the V and the N mainly had a !H* pitch accent, with a falling or a flat boundary. The A mostly had a H* pitch accent. The most popular tune was V [!H* + H-L%] + A [H*] + N [!H* + H-L%], which was used in 74 utterances (30% of the dataset). This tune had a flat contour for both three target words, making the entire utterance fairly monotonous and flat. The VA%N dataset was much smaller and the most used tune was V [!H*] + A [!H* H-L%] + N [H* + H-L%]. A small percentage of the tunes end with a H% boundary or edge tone, including 28.5% in V, 6.4% in A, and 4.7% in N across both datasets. These instances were excluded from the analysis to keep the analysis unified and the predictions consistent.

If SSRH holds for f0, we would expect a higher f0 for infrequent words when the pitch accents and boundary tones are high, and lower f0 when the tones are low. Our prediction for !H* was unclear.

The following results are based on the best-fit models. None of the full models turned out to be the best-fit model so the bigram frequency measures, including VA-FREQ, AN-FREQ and RATIO-BI, were not significant predictors for the dependent variables.

For the VA%N dataset, the statistical analysis showed that none of the frequency effects was significant, possibly due to the small number of instances in this dataset (48 instances). Therefore we will not report further about this dataset.

3.1. Mean f0 of initial third

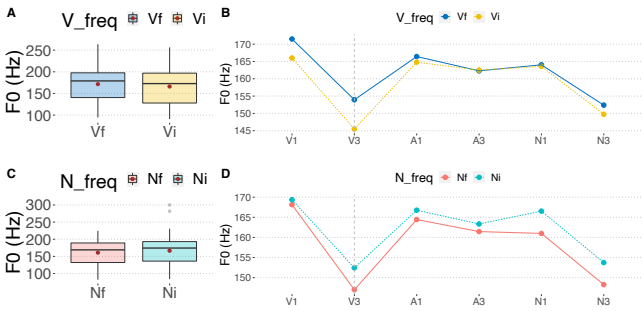


Figure 1: A & C: Mean f0 of the 1st third of V for each V-FREQ group (A) and for each N-FREQ group (C). Red dots: mean f0. **B & D**: Schematized lines by V-FREQ (B) and N-FREQ(D). V1/A1/N1: mean f0 of the initial third in V/A/N; V3/A3/N3: mean f0 of the final third in V/A/N.

In the V%AN dataset, the results showed two significant predictors on the initial third of the verb and the initial third of the noun. For the initial third of the V, the mean f0 was lower when the V was infrequent than when it was frequent (β = -7.989, SE = 3.106, t = -2.572, $p < 0.05$), as shown in Fig.1A. For the initial third of N, the mean f0 was significantly higher when the noun frequency was low ($\beta = 4.342$, SE = 2.091, t = 2.076, $p < 0.05$), as shown in Fig.1C. Fig.1B divides the data into two groups by V-FREQ, showing a difference of around 5 Hz in the verb (V1). Similarly, Fig.1D divides the data into two groups by N-FREQ, and exhibits a difference of approximately 5 Hz in the noun (N1). Both results indicated a link between the F0 value for the first third of the sonorant interval and word frequency, and thus provided certain insights for SSRH. However, because the pitch accent on V1 was downstepped, we did not have a clear hypothesis for the direction of the effect. The effect of mean f0 for N was in line with our expectations, i.e. higher f0 on a H* when N was infrequent.

3.2. Mean f0 of final third

Following the duration analysis in [7], in which the effect of V-FREQ led to a significant increase of 15 ms in the pre-boundary V coda when the verb was infrequent, we hypothesize that pre-boundary f0 is also affected by the frequency of the pre-boundary

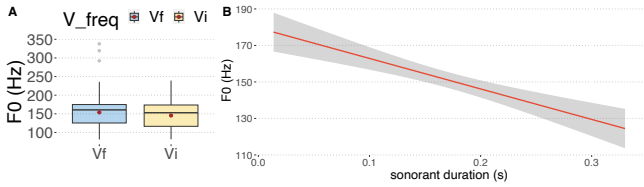


Figure 2: **A**: Mean f0 of the final third of V for each V-FREQ group. Red dots show the mean f0 difference of each condition. **B**: Regression line between the mean f0 and the duration of sonorant segment in the final third of N V-FREQ. Grey area indicates 95% confidence interval.

word.

In V%AN, the results show significant changes brought by the frequency measures and other acoustic measures on the V and the N respectively. For the final third of V, the mean f0 was significantly lower when V-FREQ was infrequent (β = -8.448, SE = 3.062, t = -2.759, $p < 0.01$), as shown in Fig. 2A. Fig.1B illustrates that when V is infrequent, the boundary tone indicated by V3 on the yellow dotted line is approximately 8 Hz lower than in the frequent V condition (V3 on the blue solid line). This is consistent with the SSRH prediction of a stronger prosodic boundary, i.e. an even lower boundary tone, after a less frequent word. For the final third of N, the mean f0 was also lower as the duration of the sonorant interval increased (β = -90.536, SE = 33.441, t = -2.707, $p < 0.01$), as in Fig. 2B. This clear inverse correlation between temporal information and f0 demonstrates further support for SSRH.

3.3. Global analysis

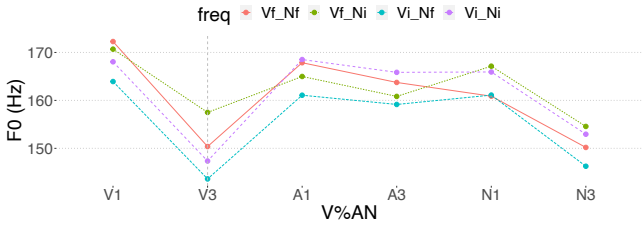


Figure 3: Schematized lines for V%AN dataset by verb-noun frequencies. V1/A1/N1: mean f0 of the initial third in V/A/N; V3/A3/N3: mean f0 of the final third in V/A/N. Grey dashed vertical lines indicate prosodic boundaries.

Fig. 3 shows schematized lines using the mean f0 values of the initial and final thirds of V, A, and N. Different lines demonstrate different combinations of verb frequencies and noun frequencies.

The results from the V%AN dataset showed a mixed picture. While both the final third of V (V3) and the initial third of N (N1) were inversely correlated with frequency measures, i.e. the lower the V-FREQ, the higher the f0 values, the mean f0 of the initial third of V (V1) changes positively with the frequency measures, i.e. the f0 was lower when the V-FREQ was low. There are several possibilities that can explain V1's pattern. The first cause for the infrequent V (Vi) being lower in the initial third (V1) than the frequent ones (Vf) may be that the low boundary tone in the final third of the V (V3) has brought down the f0 of the entire word since the words are quite short in this dataset. An additional analysis of the mean f0 of the second third of the verb showed that the f0 of V was lower for the infrequent verb by around 5 Hz than the frequent verbs (β = -5.033, SE = 2.250, t = -2.237, $p < 0.05$), consistent with the results from V1 and V3. This also suggests that the extra acoustic salience is indicated more by a much lower boundary tone than a higher pitch accent. Moreover, as mentioned previously, V having a !H* pitch accent further complicated the issue. When the pitch accent is a H*, the prediction is that the mean f0 will be higher to increase prominence[4]; however, the prediction is less clear when the pitch accent is a !H* since its F0 value may be more easily influenced by other surrounding tones. Another contributing factor may be the general prosodic composition in English. In Fig. 3B, the Vf_Nf line (red solid line) is a typical intonation contour for a (!)H* + H-L% + H* + H-L%, i.e. a high pitch accent with a falling boundary and another pitch accent with a falling boundary. In contrast, Vf_Ni seems more likely to have a narrow focus (L+H*) on the infrequent noun. Last but not least, the f0 contour on the verb was also affected by the upcoming noun: the mean f0 of the second third of the verb was also influenced by N-FREQ (β = 4.473, SE = 2.250, t = 1.988, $p < 0.05$), showing that the second third of V was lower in f0 when the frequency of the upcoming noun was higher. This highlights the need for more understanding of global phrasal patterns and long-distance coarticulation in intonation.

4. DISCUSSION AND CONCLUSIONS

In summary, this study offers useful preliminary observations regarding the relationship between frequency measures and f0 in experimental English data. In line with [4], the current results provide some tentative support for the *Smooth Signal Redundancy Hypothesis* and suggest that language redundancy does affect f0 in some ways. It also confirms that different acoustic salience measures, such as duration and f0, influence each other in a complementary manner. However, further insights might emerge with an improved design.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from AHRC-DFG Grant No. AH/W010801/1, to A. Turk , T. Bögel and C. Lai.

6. REFERENCES

[1] Aylett, Matthew and Turk, Alice, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[2] Lindblom, Björn, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990, pp. 403–439.

[3] Bell, Alan, Brenier, Jason, Gregory, Michelle, Girand, Cynthia, and Jurafsky, Dan, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[4] Turnbull, Rory, "The role of predictability in intonational variability," *Language and Speech*, vol. 60, no. 1, pp. 123–153, 2017.

[5] A. Turk, "Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis," *Laboratory Phonology*, vol. 1, no. 2, pp. 227–262, 2010. [Online]. Available: https://doi.org/10.1515/labphon.2010.012

[6] Aylett, Matthew and Turk, Alice, "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3048–3058., 2006.

[7] T. Bögel and A. Turk, "Frequency effects and prosodic boundary strength," in *International Congress of Phonetic Sciences: ICPhS2019*, 2019, pp. 1014–1018.

[8] Ladd, D. Robert, *Intonational Phonology*. Cambridge University Press, 2008.

[9] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "WebCelex," 2001, online resource: http://celex.mpi.nl/.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[11] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The Original ToBI System and the Evolution of the ToBI Framework," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 2005.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. Version 6.2.23," 2022, online resource: http://www.praat.org/.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, pp. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v067i01

[14] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2022. [Online]. Available: http://www.r-project.org/

PREDICTING PITCH ACCENT PATTERNS IN OSAKA JAPANESE USING A MACHINE TRANSLATION METHOD WITH A TRANSFORMER

Hiroto Noguchi

Tokyo Medical and Dental University/Sophia University
noguchih425@gmail.com

ABSTRACT

This study proposes a transformer-based neural machine translation method to predict accent patterns in Osaka Japanese. Unlike Tokyo Japanese, which involves only pitch fall, Osaka Japanese also employs pitch rise. Accent patterns are therefore not treated categorically as a classification task but rather as a translation task. Although the data used in this study consist only of kana characters, it was possible to predict accent patterns to some degree of accuracy, at least far beyond the level of chance. To process imbalanced data, oversampling was used to avoid the influence of frequent accent patterns. Despite using traditional recurrent neural networks, attention in the transformer model allows for some interpretability of the training results. In addition, this method will enable the consideration of accent variants in the future, as translation is not a one-to-one correspondence. The prediction accuracy can also be evaluated using BLEU scores, used to evaluate translations.

Keywords: pitch accent, Osaka Japanese, machine learning, attention, transformer

1. INTRODUCTION

In linguistics, accent patterns in Tokyo dialects have been particularly widely studied. Machine learning is also used in natural language processing for those patterns of words that cannot be processed with accent dictionaries. Traditionally, researchers have extracted features, but the use of surface form readings is now becoming the mainstream method. To the best of my knowledge, however, no studies have used machine learning to predict accent patterns in Osaka Japanese. One reason is that Osaka Japanese accent patterns have more elements with which to predict whether a word is low-beginning or high-beginning (i.e., the position of pitch rise and pitch fall if they are present), while Tokyo Japanese requires only the position of pitch fall if present. This study attempts to make such predictions using a neural machine model by replacing the input sentence and the output sentence with each *kana* letter that forms the word and the string of letters representing the high pitch and the low pitch.

2. PREVIOUS STUDIES

2.1. Accent patterns in Japanese

2.1.1. Tokyo Japanese

In Tokyo Japanese, only n+1 accent patterns are possible for n-mora words. The following are examples adapted from [1]. Words start with a low pitch, as in (1b–d), due to initial lowering (initial rise) unless the first syllable is not accented, as in (1a), or heavy [2]. What should be predicted is the position where an accent falls.

(1)	a. i'noti(+ga)	'life+NOM'	**HL**L(L)
	b. koko'ro(+ga)	'heart+NOM'	L**HL**(L)
	c. atama'(+ga)	'head+NOM'	LH**H(L)**
	d. miyako(+ga)	'city+NOM'	LHH(H)

Nouns are the main focus of research into the prediction of accent patterns. For loanwords, it has been shown that accents are more likely to be placed on the antepenultimate mora [3], as in (2a), and that the Latin accent rule can also be applied, as in (2a), and more words are predictable [4], as in (2b, c).

(2)	a. bi.ta'min	'vitamin'	LHLL
	b. ma'a.ga.rin	'margarine'	HLLLL
	c. a'.ma.zon	'Amazon'	HLLL

For compound words, the accent is either retained in the first element (N1) or the second element (N2) or shifted to a new position. When the accent is shifted, the last syllable of N1 is accented for words with short N2 and the initial syllable of N2 for words with long N2 [1]. In regard to other parts of speech, such as verbs and adjectives, the locus of accents does not matter. They are divided into two groups: accented words and unaccented ones [5]. In any case, whether a word starts high and where the pitch rises are not lexical matters in Tokyo Japanese, but predictable. The only information that is needed to predict the accent pattern is where the pitch falls, although native and Sino-Japanese simple words have not been extensively studied, as researchers have assumed that they are simply stored in the lexicon.

2.1.2. Osaka Japanese

Osaka Japanese or Kansai Japanese has two registers: the high register and the low register [6]. The following are examples of Osaka Japanese adapted from [7]. Some words are high-beginning, as in (3a–c), and the others are low-beginning, as in (3d, e). Unlike Tokyo Japanese, more than one word-initial low tone is allowed in Osaka Japanese.

(3) a. kodomo(+ga) 'child+NOM' **H**HH(H)
b. i'noti(+ga) 'life+NOM' **H**LL(L)
c. kimi'ra(+ga) 'you+NOM' **H**HL(L)
d. suzume(+ga) 'sparrow+NOM' **L**LL(H)
e. hata'ke(+ga) 'field+NOM' **L**HL(L)

Even when the position of pitch fall is identical to that of Tokyo Japanese, the tonal pattern up to that point can be different, as in (4) [7]. Compound words have the same register as N1, as in (5) [8]. Verbs are accented or unaccented, and adjectives have only one accent pattern. Some words exhibit accent variants, and generational changes have been reported [7]–[9]. However, native and Sino-Japanese words have also not been thoroughly examined in Osaka Japanese. With respect to the two registers, they are also believed to be under the veil of our mental lexicon.

(4) a. piani'suto 'pianist' **HH**HLL
b. erebe'etaa 'elevator' **LL**HLLL

(5) a. na'tsu (**H**L) + yasumi 'summer vacation' **H**H+HLL
b. haru (**L**H) + yasumi 'spring vacation' **L**L+HLL

2.2. Machine learning and accent patterns in Japanese

In [10], the researchers used an SVM to predict the accent pattern of personal names. In [11], a statistical method was employed to predict the pattern when accents are joined together. A method using LSTM and attention was proposed in [12] to develop a dictionary without burdening annotators with having to input audio and transcribe the accent patterns. In [13], a neural network was used to build an accent dictionary of millions of words based on word readings. In [14], a statistical machine learning method was used to compare the effectiveness of word readings and features found in linguistics using words registered in a dictionary. The researcher in [15] tested the ability of statistical machine learning methods to predict words already in the dictionary based on the distributed representation of words. All of the studies mentioned above are concerned with Tokyo Japanese. To the best of my knowledge, however, there are no studies addressing accent patterns of other dialects in Japanese employing machine learning.

3. METHODS

To investigate whether the words registered in a dictionary of Osaka Japanese are simply memorized (i.e., whether they are entirely random and without regularity), a machine translation model using a transformer was trained by inputting word readings, and the model was evaluated.

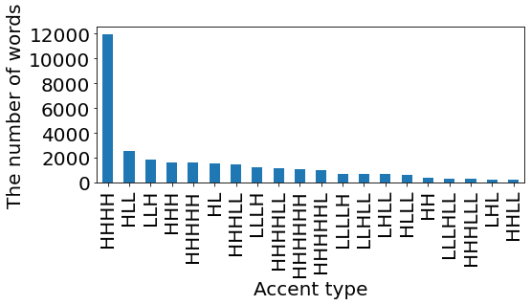
3.1. Data

[16] was used for verification. The dictionary contained 65,928 headwords, and for each headword, the accent patterns uttered by six speakers were recorded. The speakers included three from the older generation and three from the younger one [17]. Since pronunciation variants are beyond the scope of this study, only the 33,162 words with the same accent pattern for all speakers were extracted. Examples of words and accent patterns are shown in (6). Each word was divided into moras.

(6) a. ロンブン 'article' L L L H
b. カク 'write' L H
c. ヤキュー 'baseball' L L H

Figure 1 is a bar chart of the 20 most frequent accent patterns, indicating that the data are imbalanced.

Figure 1: The bar plot of the 20 most frequent accent patterns.



The data were shuffled and divided into the training set, the validation set, and the test set in the proportions of 60%, 20%, and 20%, respectively. Oversampling was applied to the training set using [18] to eliminate the problem of imbalanced data. Failure to resample unbalanced data may result in models predicting the most frequently occurring class.

3.2. Procedures

[19] was used for the translation model in Python. The tutorial codes are models for translating from Portuguese (input) to English (output), and so they were rewritten where necessary. To predict accent patterns, Portuguese must be rewritten as word readings (input) and English as accent patterns (output). Tokenizers were recreated to encode the space-separated moras and accent patterns in (6) into unique numbers. The encoding was performed as follows and padded for length alignment as in (6'). In other words, each letter is encoded into a unique numerical value, and the resulting strings are padded with zeroes to ensure that they are of equal length.

(6') a. [115, 117, 89, 117, 0, 0, 0, 0, 0]
[5, 5, 5, 4, 0, 0, 0, 0]
b. [7, 17, 0, 0, 0, 0, 0, 0, 0]
[5, 4, 0, 0, 0, 0, 0, 0]
c. [105, 11, 119, 0, 0, 0, 0, 0, 0]
[5, 5, 4, 0, 0, 0, 0, 0]

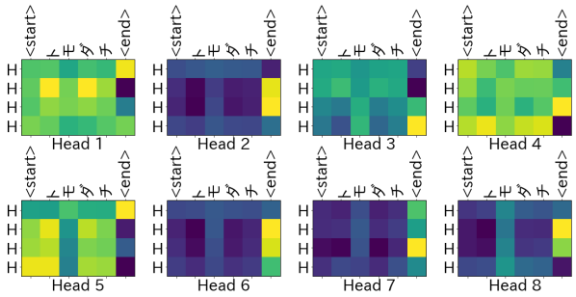
The data were converted to the same data type as the original code, and the rest of the code was run. The results were decoded from numerical values to strings with the tokenizers created earlier. The model used is as presented in [20]. The values of hyperparameters are smaller than those in the paper, as shown in the tutorial [19].

4. RESULTS

Of the 6,633 words in the test set, 3,930 were correctly predicted, giving an accuracy rate of 59.24%. The BLEU score, which originally measures the similarity between a candidate translation and one or more reference translations based on n-gram overlap, was calculated using NLTK [21]. The score obtained was 0.6047, with the range of the score being from 0 to 1, where 1 represents a perfect match between the candidate and references. However, since there were extremely short strings, a smoothing function was used. While the score alone may not be a useful indicator, as it is not a normal translation model, it can still be used to compare predictions made by other models.

The attention heads of the model for one of the words in the test set are visualized in Figure 2. <start> and <end> symbols represent the beginning and end of a word. Attention heads are components of transformer-based language models. Each attention head decides which parts of the input are important for a given output. In the following heatmaps, the brighter the color, the stronger the relationship. For example, Head 1 indicates that the first and third moras contribute to the second mora being H. In the case of translation, when a large corpus is used for learning, the cells where the corresponding words intersect with each other are brightened.

Figure 2: Visualized attention weights in the case of "トモダチ(tomodachi)" (friend).



5. DISCUSSION

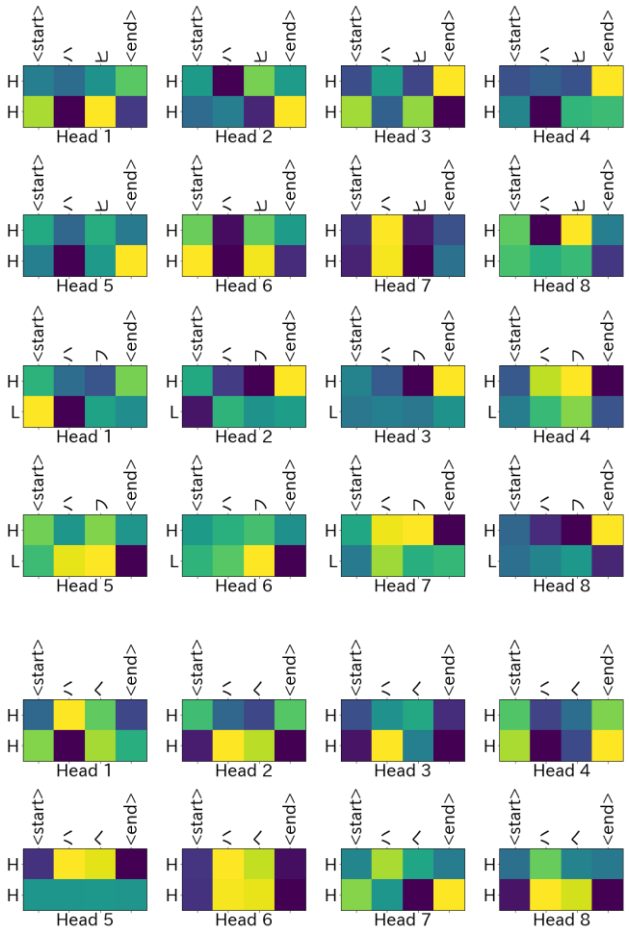
The model performance is relatively high, considering that only readings are given and that the position of pitch rise must also be predicted in Osaka Japanese, unlike Tokyo Japanese. Because of oversampling, the model did not learn accent pattern frequency. Assuming random selection, the correct response rate would be $1/n$ for n types of accent patterns. The fact that the accuracy rate is lower than that during the learning process can be attributed to overlearning. This is most likely due to the fact that oversampling has resulted in more duplicates of the same data in classes with less data.

Although visualization is possible in deep learning when using attention, it is difficult to interpret what the model has learned. However, neural models, due to their structure inspired by the human brain, exhibit a much greater degree of human-like behavior than traditional machine learning models. The fact that simply giving the sequences of segments to models as they are, rather than using the features that have been claimed in linguistics, gives better predictability with the model suggests that linguists will have difficulty building models employing the methods used in the past [14]. It is possible, however, to use neural networks to build models by classifying the data by frequency and other factors. This allows researchers to examine which data would make it easier to draw analogies for learners. In education, neural network models could be used to analyze which words should be taught first in order to facilitate the acquisition of accent patterns.

Moreover, although there are lexical and accidental gaps in the human vocabulary, the model can be used to predict accent patterns of words that do not exist as in Figure 3 and determine if they match human judgments. It can also be used to determine if

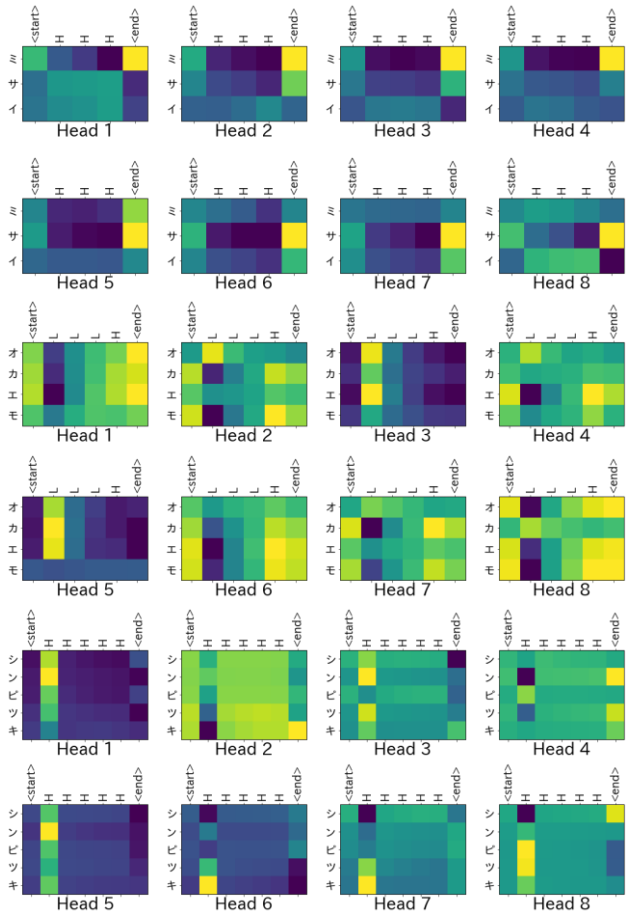
neural networks really mimic the way humans process language. Language could serve as a non-invasive way for researchers to understand their cognitive processes.

Figure 3: Visualized attention weights in the cases of “ハヒ(hahi)”, “ハフ(hafu)”, and “ハヘ(hahe)”.



Furthermore, since the model used in this study is a translation model, it is interesting that the reverse direction of translation is also possible. The model was constructed in the opposite direction, with the accent pattern as input and the *kana* letter string as output. Although this model cannot be used for prediction due to the large number of classes of strings, it can be used, as shown in Figure 4, to determine which strings are most likely to have a certain accent pattern.

Figure 4: Visualized attention weights in the cases of HHH, LLLH, and HHHHH. (“ミサイ (misai)”, “オカエモ (okaemo)”, “シンピツキ(shinpitsuki)”).



It is interesting that the time has come when models can be built that, through fine-tuning, can approximate human-like outputs, even if they cannot be described by simple rules or constraints. In the present case, oversampling was applied to avoid overfitting (“overgeneralization” in human language learning). However, the impact of the data, including frequency and variants, which were excluded in this study, on the learning process must be explored in the future.

6. SUMMARY

This study used a neural machine translation model to convert word pronunciations into accent patterns. Although many issues remain, it is worth using transformers for research on accent patterns of words registered in accent dictionaries, as large language models are surpassing human capabilities.

7. REFERENCES

[1] S. Kawahara, "The phonology of Japanese accent," in *Handbook of Japanese phonetics and phonology*, De Gruyter Mouton, 2015, pp. 445–492.

[2] M. Ota, "L1 phonology: phonological development," in *Handbook of Japanese phonetics and phonology*, Mouton de Gruyter, 2015, pp. 681–717.

[3] J. D. McCawley, *The phonological component of a grammar of Japanese*. Mouton, 1968.

[4] H. Kubozono, "Syllable and accent in Japanese–Evidence from loanword accentuation," *Onsei-Gakkai-Kaiho*, vol. 211, pp. 71–82, 1996.

[5] T. J. Vance, *The sounds of Japanese with audio CD*. Cambridge University Press, 2008.

[6] L. Labrune, *The phonology of Japanese*. OUP Oxford, 2012.

[7] Matsumori A., Nitta T., Kibe N., and Nakai Y., *Nihongo akusento nyumon [An introduction to Japanese accent]*. Sanseido, 2012.

[8] H. Kubozono, *Ippangengogaku kara mita nihongo no purosodhi [Japanese phosody from general linguistic perspectives]*. Kurosio, 2021.

[9] Y. Takeda, "Class 5 bimoraic nouns in accents of Osaka dialect: Case study of data on readings by speakers of three generations," *Handai Nihongo Kenkyu*, vol. 21, pp. 109–127, 2009.

[10] H. Nakajima, M. Nagata, H. Asano, and M. Abe, "Estimating Japanese person name accent from mora sequence using support vector machines," *IEICE(D)*, vol. 88, no. 3, pp. 480–488, 2005.

[11] R. Kuroiwa, N. Minematsu, Y. Den, and K. Hirose, "Accent labeling of a large-scale database by a single labeler and its use in statistical learning of accent sandhi," *Tech. Rep. IEICE*, vol. 106, no. 614, pp. 31–36, 2007.

[12] A. Bruguier, H. Zen, and A. Arkhangorodsky, "Sequence-to-sequence neural network model with 2D attention for learning Japanese pitch accents," in *Interspeech 2018*, ISCA, Sep. 2018, pp. 1284–1287. doi: 10.21437/Interspeech.2018-1381.

[13] H. Tachibana and Y. Katayama, "Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries," Sep. 2020, doi: 10.1109/ICASSP40776.2020.9054081.

[14] H. Noguchi, "Kikaigakushu wo mochiita tokyohogen tanjungo akusentopatan no yosokukanosei [The Predictability of Accent Patterns of Simplex Words in Tokyo Japanese Using Machine Learning]," presented at the Spring Meeting 2022 (PhSJ), Jun. 04, 2022.

[15] H. Noguchi, "Word meaning and accent patterns in Tokyo Japanese: An examination using Word2Vec word embedding," presented at the Phonology Forum 2022, Aug. 24, 2022.

[16] Sugito M., *CD-ROM accent dictionary of spoken Osaka and Tokyo Japanese*. Maruzen, 1995.

[17] M. Sugito, "CD-ROM accent dictionary of spoken Osaka and Tokyo Japanese," *J. Phon. Soc.*, vol. 4, no. 2, pp. 36–43, 2000, doi: 10.24467/onseikenkyu.4.2_36.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[19] "Neural machine translation with a transformer and Keras," *TensorFlow*. https://www.tensorflow.org/text/tutorials/transformer (accessed Jan. 03, 2023).

[20] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

DO WORDS SING THEIR OWN TUNES? WORD-SPECIFIC PITCH REALIZATIONS IN MANDARIN AND ENGLISH

Yu-Ying Chuang[1], R. Harald Baayen[1], and Melanie J. Bell[2]

University of Tübingen[1]; Anglia Ruskin University[2]
yu-ying.chuang@uni-tuebingen.de; harald.baayen@uni-tuebingen.de; melanie.bell@aru.ac.uk

ABSTRACT

This study investigates the pitch realizations of disyllabic words in Mandarin and English spontaneous speech. From two spoken corpora, we obtained F0 measurements for Mandarin words with the Rise-Fall tonal pattern, and for English words with left stress. Analyses were conducted on the pitch contours of these words, using Generalized Additive Mixed Models (GAMMs). Pitch excursions in Mandarin were generally larger, and had narrower confidence intervals compared to English, consistent with Mandarin being a tone language. Surprisingly, for both languages, word identity accounts for a considerable amount of variance in F0, independently of other covariates such as speaker sex and speech rate. The GAMMs reveal that words have their own characteristic pitch contours, just as they have their own characteristic segmental realizations. A word's specific F0 contour probably reflects the contexts and context-sensitive senses that are most common and characteristic for that word.

Keywords: Word-specific pitch, Mandarin tone, English lexical stress, F0 contour, disyllabic words

1. INTRODUCTION

In language, pitch is multi-functional. Pitch can reflect speakers' emotions and direct listeners' attention to the important parts of utterances.[1] At the lexical level, in tone languages, it serves to distinguish word meanings, and in non-tone languages, it contributes to perceived stress. However, speakers produce far more variation in pitch than would be predicted by discrete categories such as tone or stress [1, 2]. Various theories have been put forward to account for the observed pitch variation in laboratory speech, and many factors have been shown to affect pitch. Nevertheless, the way in which these factors interact in spontaneous speech production is not well understood.

The present study uses corpus data to investigate the pitch realizations of disyllabic Mandarin and English words in spontaneous speech. We focused on Mandarin words with the Rise-Fall (RF) tonal pattern (e.g., 學校 *xuéxiào* 'school') and English words with initial stressed syllable (left-stressed, e.g., ***hea**vy*). We modeled the pitch contours of our corpus tokens using generalized additive mixed models (GAMM, [3]). GAMMs allow us to model F0 as a non-linear function of time across an utterance, while also including other predictors known to affect pitch, such as speech rate and speaker sex. They can thus capture fine-grained pitch undulations in the realization of words.

Our central question is whether, in spontaneous speech and across languages, pitch contours vary consistently with word. That is to say, do individual word types have their own unique 'pitch signatures' just as they have their individual segmental makeup? In what follows, we first present the analyses and results for Mandarin. We then report our results for English. In the General Discussion, we reflect on the theoretical implications of our findings.

2. MANDARIN

The realizations of Mandarin tones have been widely studied using experimental methods in laboratory settings. It has been found that tonal realizations are subject to a variety of factors, ranging from influences at the local level such as syllable structure and consonant type, to utterance-level effects such as focus or intonation (see [4] for a review). It is therefore usually the case that actual tonal realizations deviate substantially from the canonical contours, and might even completely lose the expected shapes in connected speech [5]. Specifically for the RF pattern, previous studies based on lab speech show that its actual realization is a dipping contour followed by a delayed fall [6]. In an extreme case of severe contraction, neither rising nor falling is preserved [7].

2.1. Data

Instead of using lab-controlled experimentation, the present study investigates the realizations of the

RF contour in spontaneous speech. From the Taiwan Mandarin spontaneous speech corpus [8], we selected 51 RF words, each with at least 20 tokens. After removing tokens with pitch tracking error, we were left with a total of 3860 tokens. For each token, we took F0 measurements every 15 ms. The timepoints of these measurements were subsequently transformed into normalized timepoints from 0 to 1 (TIME). We also calculated the local speech rate for each token (SR), defined as the number of syllables per second within a time window of four words to the right and to the left of the target word. In addition, we coded the tonal context of each token (CONT), i.e., the tone type of the preceding and following words, as well as the token's normalized position in the respective utterance (POS).

2.2. Analyses

We fitted a GAMM to F0, with the following additive components: (1) a main effect of SEX, to take into account that on average, females have higher F0, (2) two TPRS[2] smooths modeling pitch as a function of TIME, one per SEX, (3) two TPRS smooths for SR, one per SEX, (4) a TPRS smooth for POS, (5) two tensor-product interaction smooths for TIME by SR, one per SEX, (6) a tensor-product interaction smooth for TIME by POS, (7) a random-effect factor smooth for TIME by SPEAKER, and (8) a random-effect factor smooth for TIME by CONT:

```
F0 ~ SEX + s(TIME, by=SEX) +
     s(SR, by=SEX) + s(POS) +
     ti(TIME, SR, by=SEX) +
     ti(TIME, POS) +
     s(TIME, SPEAKER, bs="fs", m=1) +
     s(TIME, CONT, bs="fs", m=1)
```

To address our research question, we added to these control variables our main variable of interest, namely a factor smooth for word type (WORD). With this variable included, the addition of further word-based control variables such as consonant types, vowel types and syllable structure invariably led to unacceptably high collinearity, rendering model parameters difficult to interpret [9]. We therefore excluded these additional variables from the analysis.

2.3. Results

Figure 1 presents the partial effects of the main-effect smooths modeling F0 as a function of TIME, for females (left) and males (right). Similar to results based on lab speech, the contours in spontaneous

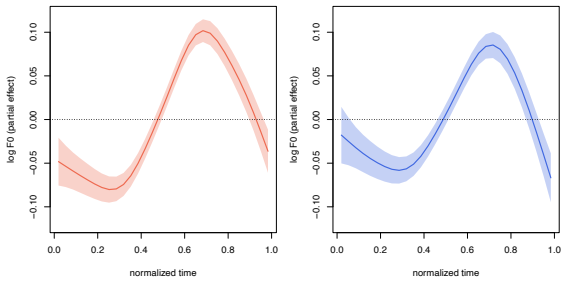


Figure 1: Partial effects of the smooths for F0 as a function of normalized time for females (left) and males (right).

speech are characterized by an initial shallow fall, followed by a long rise, and finally a fall. The partial effect for males has somewhat reduced minima and maxima, except for the final fall, which ends lower. These contours are fine-tuned by the other control variables in the baseline model.

Inclusion of the by-WORD factor smooth improved the model fit by no less than 10,250 AIC units. This result implies that how an RF pitch contour is realized is co-determined by the identity of the word. To ensure that this improvement was not an artefact, we randomized the word labels, fitted GAMMs with and without WORD as predictor, recorded the difference in AIC, and repeated this process 100 times to produce a distribution of model fit improvements in AIC units. The mean of this distribution was 1443, and the SD was 224. It is therefore very unlikely that the improvement of model fit with true word labels (10,250 units) is a statistical artifact. Spontaneously produced RF words in Mandarin thus emerge as having individual pitch signatures, just as they have their own segmental signatures.

Figure 2 presents the (partial) pitch contours that result from adding, for nine words, the partial effect of WORD to the base contour for females. (Contours for male speakers are similar, and are thus not shown.) For most words, the fall-rise-fall pattern is preserved. However, there are substantial differences in how F0 varies with time. While some words have a pronounced initial fall, e.g., (c) and (d), this fall can be severely muted, e.g., (f) and (i). Even though all these words are supposed to be realized with a fall-rise-fall pattern theoretically derived from an underlying RF tone specification, the details of actual phonetic realizations differ consistently on a word-by-word basis.

3. ENGLISH

In English, pitch contributes to the perception of lexical stress and plays an important role at the

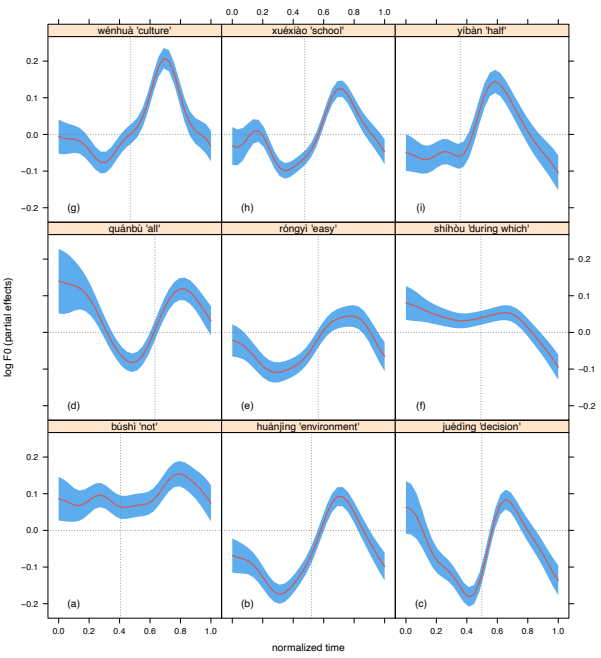


Figure 2: Predicted (partial) contours for a sample of nine Mandarin words. Vertical dotted lines indicate average syllable boundaries.

discourse level, for example to distinguish between questions and statements or between old and new information. It has been suggested that English has between four and six types of meaningful pitch accents [10, 11, 12], largely identified by the alignment of F0 peaks and valleys with segmental material [13]. However more recent research has shown that the alignment of peaks shows considerable variation between speakers, for the same speaker on different occasions and with the segmental composition of words [14, 15]. The relationship between pitch contour and function is therefore probabilistic rather than categorical [16].

3.1. Data and analyses

The English data were obtained from the Buckeye Corpus [17]. We selected the 72 left-stressed disyllabic words that have at least 20 occurrences in this corpus, producing a total of 4,724 tokens. As with Mandarin, F0 measurements were taken across the entire pitch contour, and measurement timepoints were normalized by duration. Local speech rate and utterance position were also calculated. Instead of the tonal context variable used for Mandarin, we coded, for each token, the stress level of the immediately preceding and following syllables (no stress, primary stress, or secondary stress, according to the CMU Pronouncing Dictionary[3]). We fitted a GAMM

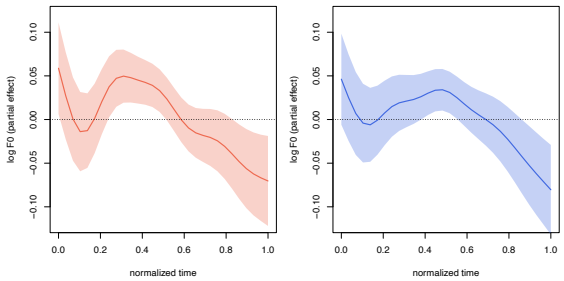


Figure 3: Predicted partial contours of English left-stressed disyllabic words for females (left) and males (right).

to the English F0 data, using the same model specification that was used for the Mandarin data.

3.2. Results

The partial pitch contours predicted for the English words are shown in Figure 3, differentiated by SEX. As with the Mandarin data, the female and male contours are similar, with the contours for females varying slightly more in magnitude than those for males. However, compared to Mandarin, the predicted contours come with wider confidence intervals, indicating more variability in the realization of pitch. Although the contours feature a small dip at the beginning followed by a gradual fall, only the falling part should be considered, given that the confidence interval of the initial fall-rise always contains 0. The pitch contour of left-stressed words therefore emerges as initially basically level, followed by a small rise and then a long shallow fall. Inclusion of a by-word factor smooth resulted in a substantial increase in goodness of fit (by 36,180 AIC units). We carried out the same randomization experiment as for Mandarin, obtaining a distribution with $M = 13,504$ and $SD = 1,292$, confirming that the random effect of WORD is unlikely to be a statistical artefact.

Figure 4 presents the predicted (partial) contours for nine English words. Compared with the pitch realizations of Mandarin words, the contours for the English words appear to be more idiosyncratic. For some words, e.g., (c) and (e), a falling pitch is obviously present, but for others, e.g., (d) and (f), some kind of plateau can be observed. The contour of the word 'only' (h) is to a large extent a flat line, while the contour for 'business' (b), has two rises.

4. GENERAL DISCUSSION

In this study we investigated pitch variation in Mandarin and English words. For both languages, word identity is a strong predictor, accounting for

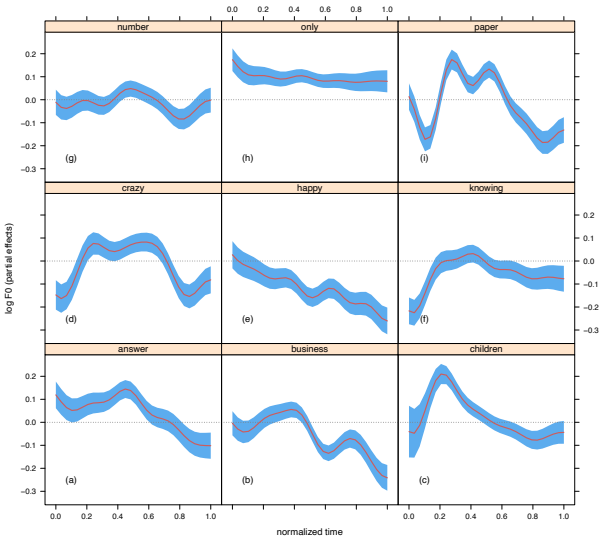


Figure 4: Predicted partial contours for a sample of nine English words. The contours were obtained by summing the base contour for females and the word-specific partial contours.

a significant amount of variance in F0, even for words with theoretically the same underlying pitch contour. This result suggests that there is by-word systematicity in pitch realization, and that in both languages, words appear to have their own pitch signatures.

What differs between the two languages is that in Mandarin, the signatures are 'variations on a theme' provided by a sine-wave like tone contour, whereas in English, word-specific signatures are more idiosyncratic, and show less resemblance to the general pitch contour, at least for disyllabic left-stressed words. This cross-linguistic difference may be due to the fact that compared to Mandarin tones, stress in English is more loosely tied to pitch, and that English words' actual pitch realizations are largely dependent on utterance intonation. However, in further analyses with additional discourse-based variables, we have so far not found a single predictor or a combination of predictors that can completely replace the by-word effect.

Despite our results, it might be thought that the evidence for word-specific pitch contours will evaporate once all variables known to co-determine pitch have been taken into account. But this argument runs into at least three problems. Firstly, such a hypothesis would be difficult to test, since many of the factors co-determining pitch are highly correlated, leading to substantial collinearity and concomitant lack of interpretational transparency. Secondly, the argument is logically self-defeating since the very large number of factors known to influence pitch effectively define the word type and possibly even the specific token. Thirdly, this line of reasoning supposes a straightforward causal link between the full array of words' other properties (lexical, syntactical, and discourse-related, but excluding pitch) and the observed word-specific pitch signatures. However, such a unidirectional causal explanation seems unlikely, since pitch is an integral part of every token heard.

Especially in the light of recent advances in artificial intelligence that use statistical patterns across the speech signal as a whole, it makes sense to consider the present findings in a broader perspective, where the word as a fundamental unit plays a much more modest role. We note that words' forms are subject to immense variation, not only at the level of segments and syllables, for which reduction is widespread [18], but also at subsegmental levels [19]. Furthermore, we note that what a word actually means is also highly dependent on the context in which it is used [20, 21, 22]. It follows that the word as a theoretical and cultural construct suggests far more constancy and unity than is actually present in spoken language. This conclusion is supported by studies reporting a much tighter link between fine-grained phonetic detail and shades of meaning than is generally assumed to exist [23, 24, 25].

The finding of word-specific pitch contours therefore does not entail or suggest that a given word is always realized in exactly the same way. Just as there is variation in the realization of a Mandarin tone type or an English stress pattern, we assume there will be variation in the exact realization of word types, reflecting large numbers of parameters including e.g., lexis, syntax, discourse, pragmatics, and emotion that jointly determine not only pitch, but also segmental realization, assimilation, and co-articulation. What the GAMM provides is a best estimate of the pitch contour of a word averaged across all its usages. We therefore hypothesize that the word-specific pitch signatures revealed by our models are the pitch contours with which words are most likely to be realized. They might be seen as reflections of, or pointers to, words' most typical patterns of use.

In conclusion, we note that from a learning perspective word-specific pitch contours make words more discriminable from one another and might hence facilitate learning. At a practical level, they could therefore have useful application in language teaching, where it may be helpful to draw attention to words' most characteristic contours.

5. REFERENCES

[1] A. T. Ho, "The acoustic variation of Mandarin tones," *Phonetica*, vol. 33, no. 5, pp. 353–367, 1976.

[2] I. Plag, G. Kunter, and M. Schramm, "Acoustic correlates of primary and secondary stress in north american english," *Journal of Phonetics*, vol. 39, no. 3, pp. 362–374, 2011.

[3] S. N. Wood, *Generalized Additive Models*. New York: Chapman & Hall/CRC, 2017.

[4] Y. Xu, "Sources of tonal variations in connected speech," *Journal of Chinese Linguistics Monograph Series*, pp. 1–31, 2001.

[5] C. Shih, "Tone and intonation in mandarin," *Working Papers, Cornell Phonetics Laboratory*, vol. 3, pp. 83–109, 1988.

[6] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[7] C. Cheng and Y. Xu, "Mechanism of disyllabic tonal reduction in Taiwan Mandarin," *Language and speech*, vol. 58, no. 3, pp. 281–314, 2015.

[8] J. Fon, "A preliminary construction of Taiwan Southern Min spontaneous speech corpus," National Science Council, Taipei, Taiwan, Tech. Rep. NSC-92-2411-H-003-050-, 2004.

[9] F. Tomaschek, P. Hendrix, and R. H. Baayen, "Strategies for addressing collinearity in multivariate linguistic data," *Journal of Phonetics*, vol. 71, pp. 249–267, 2018.

[10] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," 1997, http://www.cs.columbia.edu/~agus/tobi/labelling_guide_v3.pdf.

[11] E. Grabe, B. Post, and F. Nolan, "Modeling intonational variation in english," *The IViE system. Paper presented to the Proceedings of Prosody*, 2000.

[12] L. C. Dilley and C. C. Heffner, "The role of f0 alignment in distinguishing intonation categories: evidence from american english," *Journal of Speech Sciences*, vol. 3, no. 1, pp. 3–67, 2013.

[13] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology," *Laboratory Phonology*, vol. 3, no. 2, pp. 337–383, 2012.

[14] S. Calhoun, "The centrality of metrical structure in signaling information structure: A probabilistic perspective," *Language*, pp. 1–42, 2010.

[15] J. Cole and S. Shattuck-Hufnagel, "New methods for prosodic transcription: Capturing variability as a source of information," *Laboratory Phonology*, vol. 7, no. 1, 2016.

[16] C. Kurumada and T. B. Roettger, "Thinking probabilistically in the study of intonational speech prosody," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 13, no. 1, p. e1579, 2022.

[17] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," 2007, columbus, OH: Department of Psychology, Ohio State University (Distributor). [Online]. Available: www.buckeyecorpus.osu.edu

[18] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*. Tokyo, Japan: The National International Institute for Japanese Language, 2004, pp. 29–54.

[19] Y. Yao, "Understanding VOT variation in spontaneous speech," *UC Berkeley PhonLab Annual Report*, vol. 5, no. 5, 2009.

[20] J. L. Elman, "On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon," *Cognitive science*, vol. 33, no. 4, pp. 547–582, 2009.

[21] J. R. Firth, *Selected papers of J R Firth, 1952-59*. Indiana University Press, 1968.

[22] T. Landauer and S. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[23] S. Gahl, "Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech," *Language*, vol. 84, no. 3, pp. 474–496, 2008.

[24] K. K. Drager, "Sociophonetic variation and the lemma," *Journal of Phonetics*, vol. 39, no. 4, pp. 694–707, 2011.

[25] S. Gahl and H. Baayen, "Time and thyme again: Connecting spoken word duration to models of the mental lexicon," *arXiv*, 2022. [Online]. Available: https://osf.io/2bd3r/

[26] S. N. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003.

[1] This research was funded by ERC, project SUBLIMINAL (No. 101054902).

[2] TPRS refers to thin plate regression splines [26].

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

PROSODIC CUES OF DISTINGUISHING NEUTRAL AND NON-NEUTRAL YES/NO QUESTIONS IN HUNGARIAN: THE ACOUSTICS OF SURPRISE

Katalin Mády, Uwe D. Reichel, Anna Kohári, Cecília Sarolta Molnár, Ádám Szalontai

Hungarian Research Centre for Linguistics
mady.katalin|uwe.reichel|kohari.anna|molnar.cecilia|szalontai.adam@nytud.hu

ABSTRACT

Hungarian information-seeking yes/no questions are often realised with a single rising-falling f0 contour throughout the utterance. A common way to indicate surprise in Hungarian is to put an emphasis on more than one lexical unit, which creates multiple rise-fall patterns in questions. This paper compares f0 contours of accent groups (AGs) in non-neutral yes/no questions expressing surprise with neutral information-seeking questions, both consisting of a single intonation phrase including at least two AGs and thus multiple f0 contours. Acoustic analysis was based on f0 parameters of the AG that best represent the difference according to a clustering-based incremental feature selection procedure. Non-neutral AGs were characterised by lower f0, but larger f0 range than those appearing in neutral questions. This is in line with previous perception experiments showing that surprise is encoded in f0 parameters in Hungarian.

Keywords: prosody, yes/no questions, surprise, feature selection, Hungarian

1. INTRODUCTION

Speakers can express their attitude or pragmatic bias by several means. A well-known example is non-neutral word order in exclamations with *wh*-words that differ from the neutral word order for interrogatives, e.g. *How large the conference room is!* vs. *How large is the conference room?* Another efficient way is to utilise prosody divergent from the neutral or canonical pattern. Additionally, bias in attitude can be marked by lexical items, e.g. discourse particles.

An interesting case for speaker bias is when the speaker utters a yes/no question that is not an information-seeking interrogative in its function, rather, it expresses that the information the speaker receives goes against their previous expectations. Reasons can be surprise, disbelief, incredulity or disapproving. According to the formal description in English, these questions have rising intonation similar to neutral yes/no questions, but they keep the declarative word order. Due to the additional pragmatic information (speaker attitude or bias), they are categorised as rising declaratives (see [1] for an overview).

The way Hungarian marks yes/no questions is different from English due to the fact that this sentence type is only marked by prosody, not by word order, i.e. syntax. Alternatively, the morphological marker *-e* attached to the predicate can be used, but it is restricted to certain regional and stylistic varieties and specific contexts and will not be discussed further in this paper.

The intonational phonology of neutral yes/no questions is described among others by Ladd [2], referred to as Eastern European question intonation. A neutral yes/no question with broad focus, i.e. with no specific emphasis on any constituent, is typically realised by a rising-falling intonation starting with a low pitch accent on the initial syllable of the first lexical unit, followed by a rise until the penultimate syllable and a fall on the last syllable. In terms of intonational phonology [2, 3], the tonal pattern is modelled as given in (1). (In colloquial speech, it is common to use a definite article before proper names in most Hungarian varieties.)

(1)
Meghívták a Melindát a moziba?
L* H- L%
invited-they the Melinda+ACC the cinema-to
'Was Melinda invited to the cinema?'

The tonal pattern is not sensitive to the presence of word stress in the lexical units following the last (here: only) accented word. The f0 maximum is expected on the penultimate syllable if it is preceded by the low pitch-accented one. When the penultima is accented itself, it carries the low pitch accent, and the rise-fall sequence is realised on the last syllable. If the last syllable is pitch-accented, the rise-fall is truncated into a single rise.

In Hungarian, the rise-fall pattern characterising neutral yes/no questions is also used in biased

questions with functions similar to the English rising declarative pattern. The most influential paper on the intonation pattern for Hungarian incredulous questions is by Varga [3] who models multiple rise-fall patterns as independent intonation phrases (IPs) containing the pattern L*HL% in each contour [3]. This is illustrated by the following sentence.

(2)

Meghívták	a	Melindát	a	moziba?
L* H- L%	L* H-		L%	L* H- L%
invited-they	the	Melinda+ACC	the	cinema-to

'Melinda was invited to the cinema?'

While this approach clearly has its benefits, i.e. it can account for the complex boundary tone specific to the yes/no question pattern, it leaves other questions open. One is that the rise-fall contours are not independent in their f0 values: there is an overall declination resulting in descending f0 maxima on the penultimate syllable of each prosodic unit. Another counterargument against assuming independent IPs is that it is highly unnatural to insert a pause before the accented L* syllable, as was also pointed out by Varga. We assume instead that the multiple f0 contours can be regarded as lower-level prosodic phrases, i.e. accentual phrases (AP) or intermediate phrases (ip). APs are part of the Hungarian prosody, as was shown by [4], while the existence of ips has not been investigated so far. We will leave the structural description of the above pattern aside for the present investigation and use accent groups (AG) as the domain of prosodic investigation (see Section 2 below). A further argument against regarding AGs as being equivalent to independent IPs is that it is possible to produce the same sentence with two non-final rises and a final rise-fall rather than with three rise-falls. This is in fact the only option if the definite article *a* is dropped before the second and third pitch-accented lexical unit. Since the phonological status of the high tones is unclear, the - sign signalling a phrase accent in the intermediate phrase is omitted here, while the utterance is regarded as a single IP.

(3)

Meghívták	Melindát	moziba?
L* H	L* H	L* H L%

'Melinda was invited to (the) cinema?'

No matter which tonal model is preferred in Hungarian neutral and non-neutral yes/no questions, there is agreement that surprise/incredulity is expressed by the presence of multiple accents in utterances that could be produced with a single pitch accent on the first lexical unit in the neutral question like in (1). Marking surprise by an intonation pattern different from neutral information-seeking questions was also observed in French [5]: surprise questions end less often with a rising contour, but their mean fundamental frequency (f0) and pitch range does not differ from neutral questions. In Estonian, surprise is manifested in the following acoustic parameters [6]: surprise questions are produced with higher initial pitch and a larger pitch range, but with lower mean f0 over the entire utterance and with more creaky voice.

For Hungarian, empirical results on the intonation of yes/no questions expressing surprise are available via a perception experiment by [7, 8]. Five gradually manipulated f0 curves represented a single five-syllabic word. Participants in the first experiment were asked to decide whether the utterance conveyed a request for confirmation or surprise; and in the second one, whether it expressed a question or surprise. The perception of surprise was elicited most reliably if the first part of the rise-fall was elevated, i.e. the pitch accent was realised with higher f0 – this is in line with findings for Estonian surprise questions. Contours with a rise-fall were more likely to be identified as confirmation- or information-seeking questions.

These results give only indirect hints to the acoustic characteristics of yes/no questions expressing surprise as opposed to the more broadly defined category of neutral (information- or confirmation-seeking) questions. Thus, it remains to clarify the following issues:

- Is there any formal difference between the f0 contours in neutral and non-neutral questions? Do they differ in shape or in f0 range?
- Do neutral questions have a more pronounced rise-fall f0 pattern than those expressing surprise similarly to previous perceptual evidence?

2. DATA

The dataset is based on the Budapest Games Corpus [9] that was developed in a similar manner to the Columbia Games Corpus [10]. The task-oriented dialogues induced a large variety of sentence types with manifold pragmatic functions. A total of 525 yes/no questions were found. These were split into two main categories: neutral and non-neutral. The neutral category contained questions ($n = 461$) whose primary function was seeking information or confirmation. The non-neutral category contained questions ($n = 64$) that expressed surprise or disbelief when speakers were confronted with information contrary to their expectations.

Questions with multiple f0 contours for which the context did not provide evidence for expressing surprise (e.g. no previous contradicting expectation could be detected) were categorised as neutral. Interlabeller agreement between two expert labellers was 100% for the categories neutral vs. non-neutral on a subset of 21 questions. IPs were segmented into accent groups (AG) by labelling the accented syllable, the non-accented parts and potential pauses.

Given that neutral questions with a single accent are often realised with a rising-falling f0 contour spanning over a longer sequence like in Example (1), a direct comparison between non-neutral and neutral questions seems reasonable if only neutral questions with multiple AGs are considered. Thus, the final data set was limited to IPs consisting of at least two AGs, leaving us with 233 questions (neutral: 195, non-neutral: 38) containing 370 non-final and 228 final AGs altogether.

3. METHODS

3.1. Feature extraction

F0 was extracted by autocorrelation (Praat 6.0.37 [11], sample rate 100 Hz; allowed f0 range between 50 and 400 Hz; default settings). Voiceless parts and f0 outliers were bridged by linear interpolation. Outliers were defined separately for each file as deviating more than three times the standard deviation from the f0 mean. The contour was then smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a speaker-dependent base value b. b was set to the f0 median below the 5th percentile of an utterance and served to normalise f0 with respect to its overall level.

For prosodic feature extraction we applied the CoPaSul software version 1.0.3 [12] that decomposes the intonation of an utterance into a global intonation phrase and a local accent group component. For this study, the global component consisted of an f0 base-, mid- and topline that were fitted to the $[0, 1]$ time-normalised f0 contour in the intonation phrases (IPs). This robust fitting procedure is described and discussed in [13]. The midline was then subtracted from the f0 contour in order to remove the global component, and 3rd order polynomials were fitted to the $[-1, 1]$ time-normalised residual contour within the AG segments. These polynomial coefficients represent the local contour shapes on the AG level. Additionally, we extracted f0 residual summary statistics like mean and standard deviation, as well as f0 register and Gestalt features for the AGs. Register features measure AG-related aspects of local f0 level and range from a linear base-, mid-, and topline fit. Range is represented by a linear fit to the pointwise distances between top- and baseline. Gestalt features compare these register features between AG and IP in order to quantify how much and in what way an AG "pops out" of the underlying IP (see [12] for further details). Our feature pool comprised 32 features in total.

3.2. Feature selection

Our goal was to determine those features that best represent the AG characteristics in neutral and non-neutral questions for IP-final and non-final AGs separately. The statistic testing of such a large amount of features, 32 in our case, for each of the two conditions *IP-final* and *non-final* considerably increases the risk of type 1 errors. Therefore, we decided for an alternative approach which is clustering-based incremental feature selection. We applied this approach separately for the IP-final and non-final AGs.

Starting with zero features we iteratively added the feature that maximally increases the agreement between the clustering of the AG feature vectors and their classes neutral vs non-neutral. We applied the k-means algorithm with two clusters, a *k-means++* initialisation, and a constant random seed. As a scoring function that measures the amount of agreement between the clustering and the AG types we used the adjusted mutual information score (AMI) [14] between cluster IDs and question type. It measures the amount of uncertainty reduction about one of these variables if the values of the other variable are known. The AMI further corrects for chance-level agreement. As stopping criterion we defined the absence of any feature that would further increase the AMI score. Both clustering and AMI calculation were done with the Python *scikit-learn* package version 0.24.2 [15].

4. RESULTS

Figure 1 shows the feature selection result for the non-final and final AGs within the IP. For non-final AGs the closest agreement between AG features and question types was achieved with the features *ml_c0, rng_m*, and *rng_d_fin*. For final AGs the features maximising the agreement were *c0, ml_m, m, c3* and *rng_c0*. These features are described in Table 1.

The violin plots in Figures 2 and 3 for non-final and final AGs show how neutral and non-neutral AGs differ with respect to the selected features.

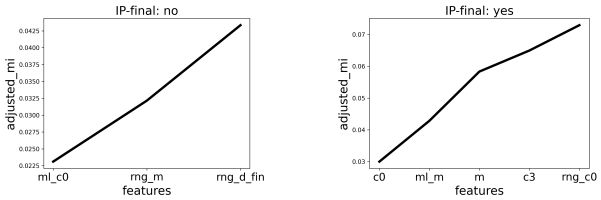


Figure 1: Feature selection results for non-final (left) and final AGs (right). See Table 1 for a description of these features.

Feature	Description
c0	offset coefficient of the local contour stylisation, that corresponds to the stylised f0 residual value at the midpoint of the AG
c3	hyperbolic polynomial coefficient of the local contour stylisation
m	arithmetic mean of the local f0 residual contour
ml_c0	offset coefficient (start level) of the linear AG midline fit
ml_m	arithmetic mean f0 of the AG midline
rng_c0	offset coefficient (start range) of the linear range fit to the pointwise distances between base- and topline in the AG
rng_d_fin	difference between AG and IP range at the end of the AG
rng_m	arithmetic mean of fitted range line in the AG

Table 1: Description of the AG intonation features resulting from incremental feature selection.

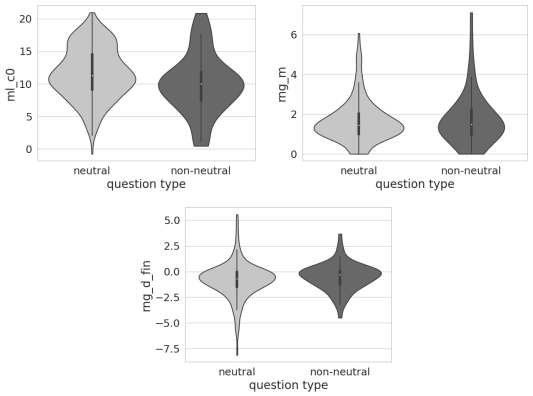


Figure 2: Distributions of selected features for **non-final** neutral and non-neutral AGs. See Table 1 for a description of these features.

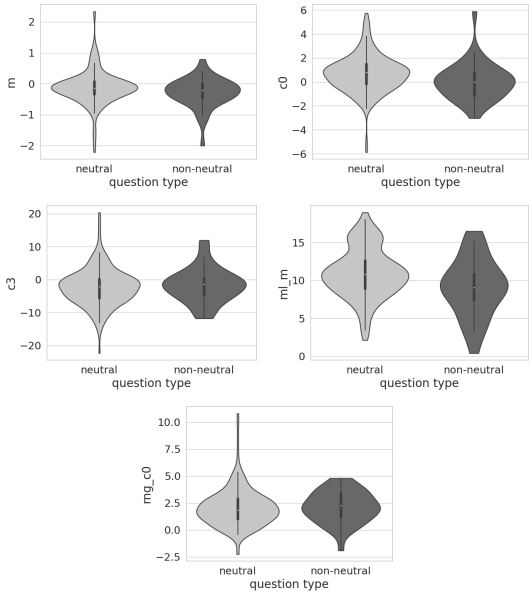


Figure 3: Distributions of selected features for **final** neutral and non-neutral AGs. See Table 1 for a description of these features.

5. DISCUSSION

Register can be expressed in terms of f0 *level* and *range* [16]. As Figures 2 and 3 show, this distinction is of relevance to describe both non-final and final, neutral and non-neutral AGs. Generally, non-neutral AGs can be characterised by a **lower f0 level** (in terms of f0 mean *m*, local contour offset coefficient *c0*, and midline mean *ml_m*), but **larger f0 range** (in terms of range offset coefficient *rng_c0*, and range mean *rng_m*). The difference might result from the fact that unlike in the perception experiment, our utterances contained at least two f0 contours. Furthermore, non-neutral AG intonation shows a **weaker downtrend tendency**, i.e. a smaller f0 slope which is expressed in terms of less negative hyperbolic coefficient values *c3*, which causes a less negative local f0 contour trend, and in terms of a larger range at the end of the AG *rng_d_fin*. These results only partly confirm the findings of [7, 8], according to whom questions perceived to indicate surprise have a larger f0 range and a *higher* f0 onset. At the same time, they are in line with findings for Estonian surprise questions.

According to the feature selection procedure, parameters representing the f0 contour shape such as *c2* were not chosen as contributing to the disctinction between the neutral and non-neutral category. This might be a result of the inhomogeneous contour shapes present in spontaneous speech.

6. ACKNOWLEDGEMENTS

This study was funded by the National Research, Development and Innovation Office, grants NKFIH K 135038 and PD 134775. We would also like to thank Beáta Gyuris and Hans-Martin Gärtner for their help to better understand the theoretical framework and the contextual interpretation of non-neutral questions.

7. REFERENCES

[1] B. Gyuris, "Thoughts on the semantics and pragmatics of rising declaratives in English and rise-fall declaratives in Hungarian," in *$K + K = 120$: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*, B. Gyuris, K. Mády, and G. Recski, Eds. MTA Research Institute for Linguistics, 2019, pp. 247–280, http://www.nytud.hu/kk120/www_print/index.html.

[2] D. R. Ladd, *Intonational phonology, 2nd ed.* Cambridge: Cambridge University Press, 2008.

[3] L. Varga, "Boundary tones and the lack of intermediate phrase in Hungarian," *The Even Yearbook*, vol. 9, pp. 1–27, 2010.

[4] Š. Beňuš, U. D. Reichel, and K. Mády, "Modeling accentual phrase intonation in Slovak and Hungarian," in *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014.*, L. Veselovská and M. Janebová, Eds. Olomouc: Palacký University, 2014, pp. 677–689.

[5] A. Celle and M. Pélissier, "Surprise questions in spoken French," *Linguistics Vanguard: a Multimodal Journal for the Language Sciences*, vol. 8, no. 2, pp. 287–302, 2022. [Online]. Available: https://hal.science/hal-03133878

[6] H. Sahkai, E. L. Asu, and P. Lippus, "Prosodic characteristics of canonical and non-canonical questions in Estonian," in *Proc. Speech Prosody, Lisbon*, 2022, pp. 135–139.

[7] A. Kiss and A. Szalontai, "The form and meaning of hungarian confirmative and echo declarative questions," in *15. International Conference on the Structure of Hungarian*, Pécs, 2021, https://icsh15.netlify.app/.

[8] A. Kiss, "A magyar deklaratív kérdések karakterdallamának vizsgálata percepciós kísérlettel," in *Általános Nyelvészeti Tanulmányok*, K. Mády and A. Markó, Eds. Budapest: Akadémiai Kiadó, 2022, no. XXXIV, pp. 169–194.

[9] K. Mády, A. Kohári, U. D. Reichel, A. Szalontai, and P. Mihajlik, "The budapest games corpus," in *Beszédkutatás – Speech Research Conference*, T. E. Gráczi, V. Horváth, K. Juhász, A. Kohári, V. Krepsz, and K. Mády, Eds., Budapest, 2023, pp. 75–77.

[10] A. Gravano, v. Beňuš, H. Chávez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of 'okay'," in *Proc. 45th Annual Meeting of Association of Computational Linguistics*, Prague, 2007, pp. 800–807.

[11] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[12] U. Reichel, *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*, RIL, MTA, Budapest, Hungary, 2016, https://arxiv.org/abs/1612.04765.

[13] U. Reichel and K. Mády, "Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech," in *Elektronische Sprachsignalverarbeitung 2013*, ser. Studientexte zur Sprachkommunikation, P. Wagner, Ed. Dresden, Germany: TUDpress, 2013, vol. 65, pp. 223–230.

[14] N. Xuan, V. Julien, S. Wales, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] T. Rietveld and P. Vermillion, "Cues for perceived pitch register," *Phonetica*, vol. 60, pp. 261–272, 2003.

Explicit Rules or Implicit Imitation: a Comparative Study of Two Approaches to Teaching English Prosody

Xiaodan Zhang, Joaquín Romero

xiaodan.zhang@urv.cat , joaquin.romero@urv.cat
Rovira i Virgili University (Spain)

ABSTRACT

This study aims to explore the effect of explicit and implicit instructions on improving English pronunciation as far as stress production is concerned. 28 tertiary-level Chinese EFL learners were divided equally into two groups and attended a pronunciation training course instructed using the two approaches in question respectively. The pre-test and post-test, involving an identical controlled reading task, were conducted before and after the experiment. Participants' recordings were analyzed acoustically on PRAAT. Readings of intensity peaks and pitch averages for each syllable at word-level and for each word at sentence-level were obtained. These values were then processed in order to obtain the difference in amplitude and pitch between stressed and unstressed syllables/words. The statistical results generated from a series of repeated measures ANOVAs revealed that while both groups improved their pronunciation in stress production, the improvement was more pronounced in the explicit group.

Keywords: English stress, implicit approach, explicit approach, prosody teaching

1. INTRODUCTION

Improper English prosody is one primary cause of foreign accent which can lead to unintelligibility of an EFL learner's speech [1]. Early research on pronunciation focused almost exclusively on segmental instruction. However, in more recent studies, a paradigm shift has been witnessed from a focus on segmentals to an increased emphasis on suprasegmental aspects and prosody [2].

Several studies have attempted to prove the effectiveness of introducing rhythm instruction within the English L2 class to improve EFL learners' prosodic skills [3] [4]. Nevertheless, there is still a scarcity of research exploring the efficacy of prosody instruction, especially to Chinese EFL learners whose L1 has a completely different rhythm from English.

Therefore, the current research intends to investigate if Chinese EFL learners' pronunciation could be refined by a period of training as far as word stress and sentence stress are concerned. Explicit vs. implicit instruction methods are investigated in terms of their efficacy in improving EFL learners' pronunciation. Both approaches have been widely researched, and positive effects have been observed. We aim to further explore whether improvements vary between the two approaches after the same period of training. The following two major research questions are formulated:

RQ1. Does prosody training help Chinese EFL learners with their production of English stress?

RQ2. Which approach (explicit vs. implicit) is more effective in improving Chinese EFL learners' production of English stress?

2. LITERATURE REVIEW

The issue of whether to teach pronunciation implicitly or explicitly has long been a matter of controversy. According to Ellis [5], implicit instruction advocates unconscious and automatic learning, which is similar to the process of first language acquisition; explicit instruction, on the other hand, constitutes direct intervention and provides descriptions and explanations of the rules to facilitate learning.

Abundant studies carried out in a variety of EFL/ESL contexts have taken side for the explicit approach and obtained positive results generated from explicit teaching [3] [6] [7] [8]. However, there are other studies that have favored an implicit instruction of pronunciation [9] [10] [11], taking it for granted that such a mode of instruction could help learners enhance their English pronunciation by arousing their unconscious knowledge.

Meanwhile, a more balanced view is put forward by [12] that both instructions, explicit and implicit, have equally significant effects on pronunciation learning. "It might be the input, practice, and feedback included in pronunciation instruction, rather than the explicit phonetics lessons, that are most facilitative of improvement in pronunciation" [12].

Despite the existing controversy in the literature, it is observed that most associated research is confined within the segmental level, very few studies have extended the research scope to suprasegmentals in terms of word stress. Particularly, sentence stress instruction and its efficacy has barely been covered up to date.

3. METHOD

3.1. Participants

28 tertiary-level Chinese EFL learners (20F, 8M) aged from 21 to 25 were selected to participate in the study. Every participant had reached the equivalent level of English B1-B2 (based on CEFR standard) prior to the experiment, according to the English official exams they had taken (IELTS, TOEFL, CET6,TEM 4). Most of the participants were language major students or graduates (English & Spanish) and all of them claimed unfamiliarity to English prosody and lack of rhythm in their speech, according to the questionnaire collected before the experimental treatment. Participants were equally divided into two groups (explicit vs implicit) and attended a pronunciation training course instructed by different approaches.

3.2. Procedures &Materials

For each group, the training course consisted of 8 modules of approximately 30 minutes each, with a frequency of twice per week. The training was completely done remotely via ZOOM due to the Covid restrictions.

The pronunciation course was designed to instruct placement of word stress and sentence stress. Following the same syllabus, however, the teaching approaches and materials varied between the two experimental groups. The explicit group was exposed to the pronunciation rules explicitly. For instance, students were instructed with different types of sentence stress, namely, sentence-end tonic stress, stress shift for new information, contrastive or emphatic purpose. Meanwhile, the implicit group was provided with the same practice materials used in the course accompanied with native speaker recordings to imitate, which the other group did not have access to. Thus, the implicit group was expected to infer the underlying rules themselves without explicit instruction from the researcher.

All participants took a pre-test before the treatment and the identical test upon completion of the training as the post-test. Both tests comprised a controlled reading task of a series of words and sentences/dialogues. The selected words contain at least two syllables so that the stress difference between syllables could be identified. The following 8 target words chosen: *curiosity*, *maintain*, *analysis*, *satisfactory*, *agriculture*, *analyze*, *cooperate* and *immigrant*.

In terms of sentence/dialog reading, different types of sentence stress were covered. For instance, Figure 1 presents the sentences involving contrastive stress in the pre/post-test.

1. He is a doctor and I am a nurse.
2. I love playing badminton rather than watching it.
3. Is he traveling here? No, he is living here.
4. I prefer a car to a bike.
5. He cleaned the room, not me.
6. It is cold in the morning and hot in the afternoon.

Figure 1: Target sentences with contrastive stress in the Pre/Post test

Given the restrictions imposed by the remote research conditions, the recordings were done on the participants' own mobile devices in a required noise-free environment then transferred to the researcher.

3.3. Data analysis

Pre-test and post-test data produced from only 7 participants from the explicit group and 10 from the implicit group were adopted for acoustic and statistical analysis as the rest failed to complete 90% of the 8 sessions. We set a high attendance standard for data collection so that the effect of course implementation could be verified with higher confidence. All the recordings were then transcribed and analyzed acoustically using the PRAAT software. Due to time and space limitation, this paper only presents the result of a selection of the controlled reading data from the pre- and post-test, particularly related to word stress and contrastive stress.

As far as word stress is concerned, each syllable in the word was identified in the acoustic signal and a reading of the intensity peak and pitch average was obtained. These values were then processed in order to obtain differences in amplitude and pitch respectively between the syllable with the primary stress and a randomly selected unstressed syllable from the same word. Eventually, these values of difference were compared statistically between pre- and post-test for each group.

In terms of contrastive stress, the procedure was slightly different. Each word in the sample sentences was segmented rather than the syllables, and so were the intensity peak and pitch average readings. Just as in the word stress analysis, differences in amplitude and pitch were likewise calculated between the target contrastive stress and another random content word within the same sentence.

Regarding the statistical method, a series of repeated measures ANOVA tests were performed to compare the pre- and post- difference as well as across groups.

4. RESULT

4.1 Statistical analysis of words

Table 1 presents the words' intensity difference for each group between the two time points by means and SD. Both groups have witnessed an increase in their values in the post-stage compared with the pre-stage.

Time	Group	Mean	SD	N
Post	Explicit	1.623	3.886	56
	Implicit	0.825	3.190	80
Pre	Explicit	−0.277	2.738	56
	Implicit	0.548	3.752	80

Table 1: Word intensity Mean and SD statistics for two groups in two stages.

The improvement was more pronounced in the explicit group: M =-0.28, SD = 2.74 in the pre-test and M=1.62, SD =3.89 in the post-test. The implicit group, on the other hand, obtained M =0.55, SD = 3.75 in the pre-test, and M = 0.83, SD =3.19 in the post-test conditions. Figure 2 illustrates these differences.

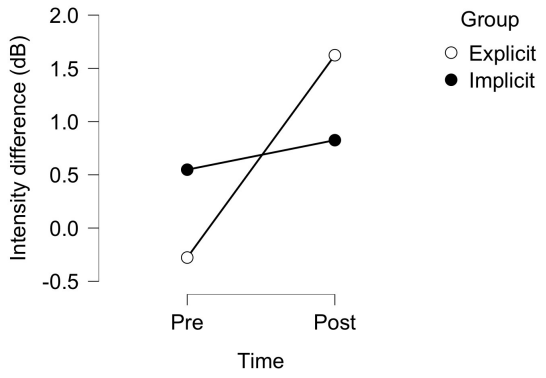


Figure 2: Plot of word intensity difference across groups in two stages.

The results of the post-hoc tests indicate no significant difference in the pre-test stage for the two groups (t=−1.60, p=0.33), which means the participants started at a similar level before entering the experiment. However, the explicit group witnessed a significant difference between the pre and post stage (t=−3.86, p<.05), whereas the implicit did not (t=−0.48, p=0.63).

In terms of the pitch analysis, both groups have shown a higher pitch difference in the post stage. In the same vein, the explicit group revealed a bigger difference between pre- and post- test, as presented in Table 2 and Figure 3 (Explicit: pre M=8.82 vs post M=21.60; Implicit: pre M=14.54 vs post M=23.68).

Time	Group	Mean	SD	N
Post	Explicit	21.599	36.688	56
	Implicit	23.679	39.566	80
Pre	Explicit	8.822	29.966	56
	Implicit	14.543	37.263	80

Table 2: Word pitch Mean and SD statistics across two groups in two stages.

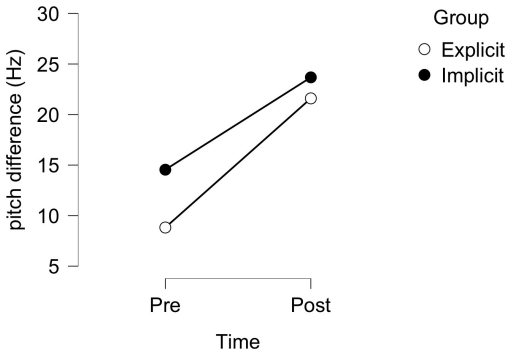


Figure 3: Plot of word pitch difference across two groups in two stages.

Slightly different from the statistics for intensity, the post-hoc tests indicate a significant improvement in pitch difference between pre and post-stage in both groups (Explicit: t=−3.48, p<.05; Implicit: t=−2.63, p<.05).

4.2 Statistical analysis of contrastive sentence stress

Regarding contrastive sentence stress, Figure 4 presents the intensity difference between the pre and post stages in the two groups. As seen in graph, both groups have experienced an increase in intensity difference in the post stage. A greater contrast was seen in the explicit group (pre M=-0.28, SD = 2.74 vs post M=1.65, SD = 4.02) than the implicit group (pre M=-0.13, SD = 4.56 vs post M=1.01, SD = 4.48), seen from the descriptive statistics in Table 3. However, the improvement was significant in both groups between pre and post, as indicated by the results of the post-hoc tests (Explicit: t=−4.20, p<.05; Implicit: t=−3.17, p<.05).

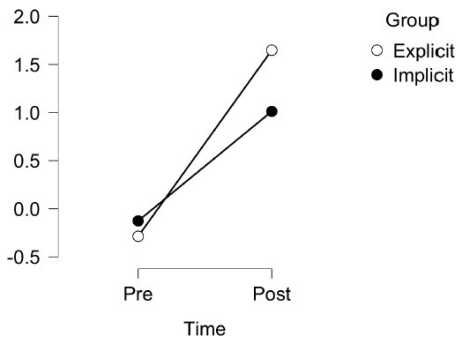


Figure 4: Plot of intensity difference for contrastive sentence stress across groups in two stages.

Time	Group	Mean	SD	N
Post	Explicit	1.647	4.016	84
	Implicit	1.011	4.481	120
Pre	Explicit	−0.286	3.063	84
	Implicit	−0.126	4.560	120

Table 3: Intensity Mean and SD statistics for contrastive stress across two groups in two stages.

Taking a look at the pitch statistics for contrastive sentence stress, Figure 5 shows a very similar pattern to the intensity graph shown above. Once again, the explicit group revealed a bigger difference than the implicit group (Explicit: pre M= -3.27, SD = 41.36 vs post M=13.61, SD = 64.00; Implicit: pre M= -1.84, SD = 44.90 vs post M=9.44, SD = 47.15), seen from Table 4 descriptive figures.

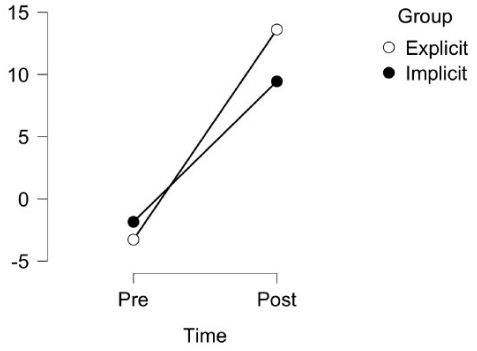


Figure 5: Plot of pitch difference for sentence contrastive stress across groups in two stages

Time	Group	Mean	SD	N
Post	Explicit	13.611	64.004	84
	Implicit	9.444	47.146	120
Pre	Explicit	−3.272	41.357	84
	Implicit	−1.838	44.896	120

Table 4: Pitch Mean and SD statistics for contrastive stress across two groups in two stages

Yet, the significant difference was only obtained in the explicit group between the pre and post ($t=-3.18$, $p<.05$), whereas the implicit group did not show a significant improvement ($t=-2.22$, $p= 0.13$).

5. DISCUSSION & CONCLUSION

In answer to RQ1, the above results for both intensity and pitch measurements have revealed positive increases from the pre stage to the post, which indicates that participants were showing awareness to place more stress on the target words. It seems our prosody training has worked efficiently on participants from both groups, despite the fact the progress in some cases was not statistically significant.

Particularly, the explicit group demonstrated significant improvement in the post stage for word stress and contrastive sentence stress respectively, supported by the statistics of both pitch cues and intensity cues. Nevertheless, the differences in the implicit group were not consistently significant. For instance, regarding word stress, only the pitch cue showed a significant difference and progress represented by the intensity cue was insignificant. Yet, sentence stress exhibited the opposite trend. Significance was detected in intensity values rather than pitch.

Despite the customary belief that stressed syllables are usually produced with relatively higher F0, greater intensity, and longer duration compared to unstressed syllables [13], inconsistency among these measurements is not rare to trace in acoustic research, especially when external factors are involved, e.g. noise, pitch accent, which can cause misrepresentation on intensity and F0 cues. Besides, there has been a lack of consensus on the hierarchy of the acoustic correlates of English lexical stress. Most believe the primary cue for stress in English in both natural and synthesized speech is relative pitch prominence, followed by length and amplitude [14] [15] [16]. Thus, although pitch and intensity values did not always correspond with each other in terms of significance, the tendency of showing improvement was demonstrated.

Above all, to answer RQ2, it seems the treatment in the explicit group has generated better results to improve participants' production of word stress and sentence stress. Such a finding concurs with the existing literature advocating explicit pronunciation instructions. With the help of explicit rules, learners are more likely to notice language features and build up a greater awareness to pronounce the stress strongly. On the other hand, the implicit approach, despite its visible potency, did not benefit learners to the same extent. This could be explained by the connectionist theories that implicit knowledge requires exposure to massive amounts of input and acquisition is a slow, organic, elaborate process [17].

Our study only provided a short-term training and thenative recordings for imitation practice could not meet the criterion of massive amounts of input. Therefore, it appears logical to assume that the implicit group could not achieve as notable a progression as the explicit group. However, as Ellis [18] suggested, the explicit knowledge resulted from explicit learning can assist the processes involved for implicit learning, and thus we anticipate that a mix of the two methods might yield a double effect in a positive way. Future studies that incorporate a larger sample for a longer period of prosody training could add more credence to the efficacy of the two approaches in this study.

6. REFERENCES

[1] Hahn, L. D. 2004. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

[2] Busà, M. G. 2012. The role of prosody in pronunciation teaching: A growing appreciation. In: Busà, M. G., Stella, A (eds.), *Methodological Perspectives on Second Language Prosody*. Papers from ML2P 2012, 101-106. Padua: Cleup.

[3] Quesada Vázquez, L., Romero, J. 2018. The improvement of Spanish/Catalan EFL students' prosody by means of explicit rhythm instruction. Proc. ISAPh 2018 International Symposium on Applied Phonetics, 104-109, DOI: 10.21437/ISAPh.2018-19.

[4] Tsiartsioni, E. 2011. Can pronunciation be taught? Teaching English speech rhythm to Greek students. In Eliza Kitis, Nikolas Lavidas, Nina Tpointzi and Tasos Tsangalidis (eds.), *ISTAL* 19, 447-458.

[5] Ellis, N. C. 2008. Implicit and explicit knowledge about language. *Encyclopedia of Language and Education, 6*, 1-13.

[6] Couper, G. 2003. The value of an explicit pronunciation syllabus in ESOL teaching. *Prospect, 18*(3), 53-70

[7] Koike, Y. 2014. Explicit pronunciation instruction: Teaching suprasegmentals to Japanese learners of English. In Sonda, N., Krause. A. (eds.), *JALT 2013 Conference Proceedings*. 361-374. Tokyo: JALT.

[8] Naeini, B., Adni, Z. 2017. The impact of explicit vs. implicit instruction on pronunciation intelligibility among Iranian EFL learners. *Journal of Studies in Learning and Teaching English, 6*(2), 103-123.

[9] Krashen, S. 1981. *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon.

[10] Minhong, Y., Ailun, L. 2006. The effect of implicit learning on the learning of a second language pronunciation. *Psychological Science, 3*, 26.

[11]Papachristou, V. 2011. Explicit vs. implicit pronunciation teaching to Greek children: The case of the acquisition of English vowels. *Selected papers from the 19th ISTAL*, 371-381.

[12] Kissling, E. M. 2013. Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal, 97*(3), 720-744.

[13] Goffman, L., Malin, C. 1999. Metrical effects on speech movements in children and adults. *Journal of Speech, Language, and Hearing Research, 42*(4), 1003-1015.

[14] Bolinger, D. L. 1958. A theory of pitch accent in English. *Word, 14*(2-3), 109-149.

[15] Beckman, M. E. 1986. *Stress and Non-Stress Accent*. Dordrecht, the Netherlands: Foris Publications.

[16] Roach, P. 2009. *English Phonetics and Phonology: A Practical Course*. Cambridge University Press.

[17] Ellis, R. 1989. *Understanding Second Language Acquisition* (Vol. 31). Oxford: Oxford university press.

[18] Ellis, N. 1994. Introduction: implicit and explicit language learning–an overview. In Ellis. N (ed.). *Implicit and Explicit Learning of Languages*. San Diego: Academic Press.

Acknowledgements This work was supported in part by grant PID2020-117804GB-I00 from the Spanish Ministry of Science and Innovation.

BOUNDARY PERCEPTION AS INTERPLAY BY ACOUSTIC CUES, SYNTACTIC CONSTITUENCY, AND PROMINENCE

Jianjing Kuang, May Pik Yu Chan, Nari Rhee

University of Pennsylvania
kuangj@sas.upenn.edu, pikyu@sas.upenn.edu, nrhee@sas.upenn.edu

ABSTRACT

Boundary perception is a complex process related to multiple factors, such as acoustic boundary cues, syntactic constituency, and prominence, however, empirical studies are still lacking to explore how they work together in determining listeners' boundary perception; and it is unclear whether these factors carry the same weights in different languages. This study aims to address these questions by investigating the boundary perception rates of utterances from continuous corpora in Mandarin and English. Boundary and prominence ratings were collected with a crowd-sourcing perception experiment. The relative strength of the syntactic boundary of both the left and right sides of the constituents was extracted from the syntactic parsing annotations. A wide range of acoustic cues of both prosodic domain-final and domain-initial positions were examined. Results confirmed that boundary perception is shaped by complex interactions among acoustic cues, syntactic parsing, and prominence perception. More importantly, there is cross-linguistic variation between Mandarin and English.

Keywords: speech prosody, boundary perception, prominence, syntax-prosody interface

1. INTRODUCTION

Being able to parse speech into meaningful junctures is an important step in speech processing by human listeners. Boundary perception is a rather complex process that is related to several levels of factors, such as prosodic cues, syntactic structure, and information structure. However, there are still relatively few empirical studies to explore how these factors work together to determine the location and strength of the perceived boundaries. This study intends to provide important evidence for a better understanding of the effects of acoustic boundary cues, syntactic constituency, and prominence on boundary perception.

It has been well-established that prosodic boundaries correlate with a number of acoustic cues. For example, domain-final positions can often be indicated by pauses, final lengthening [1, 2, 3, 4], pitch declination and final lowering [5, 6, 7, 8], creakier voice quality [4, 9, 10, 11]; while domain-initial positions are correlated with domain-initial strengthening and pitch reset [12, 13]. However, although these acoustic cues are widely used among languages, their relative cue weighting may vary across languages.

Prosodic phrasing has a tight link with syntactic parsing. Studies have shown prosodic boundaries are useful for locating syntactic boundaries [14, 15], and resolving syntactically ambiguous sentences [16, 17]. However, the effects of syntax on prosodic boundary perception are much less well understood, though [18, 19] suggest that syntactic cues significantly influence the detection of prosodic boundaries. It is also not clear how the left and right edges of the syntactic constituents align with the prosodic domains (c.f.[20]) in different languages.

Moreover, prominence and boundary as the two important prosodic aspects have an interdependent relationship. Prominence has effects on boundary perception, and vice versa. For example, phrase-final position appears to be a preferred position for prominence perception. Some studies [21, 22] found that phrase-final words are more likely to be perceived to be prominent; with similar effects being found for Spanish, French, and English [21], despite the three languages being quite different in terms of their phrasal prosody. It remains to be seen whether this domain-final effect is universal across other languages as well. When resolving syntactically ambiguous sentences, English speakers treat the prominent words as the end of the syntactic phrase [17]; by contrast, Mandarin listeners parse prominence as the beginning of the phrase [23]. Therefore, although the interdependence between prominence and boundary appears to be common among languages, it is possible that the direction of the alignment between prominence and boundary varies across languages. More cross-linguistic empirical studies are needed to clarify these issues.

This study aims to address these questions

by investigating the boundary perception rates of utterances from continuous corpora in Mandarin and English. Boundary and prominence ratings were collected with a crowd-sourcing perception experiment. The relative strength of the syntactic boundary of both the left and right sides of the constituents was extracted from the syntactic parsing annotations. A wide range of acoustic cues of both prosodic domain-final and domain-initial positions were examined.

2. METHODS

2.1. Participants and materials

Sentences from syntactically-parsed read speech corpora were used as the stimuli for the experiments for both Mandarin and English. For English, a female native speaker's reading of Jane Austen's Emma was obtained from LibriVox [24]. We chose a female speaker's recording of Volume II Chapter 10 of the book for our study, a chapter also in the 2nd edition of The Penn Parsed Corpus of Modern British English (PPCMBE2) [25], which contains Penn Treebank-style annotated brackets. For Mandarin, readings of news articles by one male and one female speaker were retrieved from the Chinese Tree Bank speech corpus [26]. The text of this corpus is based on Chinese Tree Bank 9.0 [27], which included news article texts parsed and annotated with Penn Treebank-style labeled brackets.

A total of 76 participants were recruited for the perception study, including 47 native speakers of English (18-25 years; 29 female) and 29 native speakers of Mandarin (18-35 years; 18 female). All were recruited from the university student community and completed the experiment for partial course credit. From the aforementioned corpora, sentences of around 30 seconds or less were selected for the study (24 for English, 22 for Mandarin). Participants listened to the sentences in their native language and were asked to identify and annotate prominent words and prosodic boundaries. There were up to 382 potential boundaries and 408 potential prominent words in the English condition, and 458 potential boundaries and 480 potential prominent words in the Chinese condition that participants could rate.

2.2. Rapid Prosody Transcription Task

We conducted a listening experiment where participants provided ratings of prominences and boundaries as in the Rapid Prosody Transcription task used in [18, 28], and the boundary detection task in [19]. Participants heard selected sentences one at a time while reading the transcription that was displayed simultaneously. Once the participant entered a trial, the audio started automatically, though participants could replay the audio as many times as they wished. Participants were asked to select where they "think there are boundaries" and where they "think there are prominent words" within the sentence. Participants could not select the beginning or end of the sentence as boundary markers, though any word in the sentence could be selected as a prominent word. The experiment was administered on Qualtrics.

Before the experiment, participants first heard three sample recordings involving the same sentence read with different prosodic focus and structures as examples. This was to familiarize participants with the experimental interface. No further explanations for what constitutes a "boundary" or "prominent" word were given. For both prominence and boundaries, the response rate was calculated by the response count divided by the maximum number of corresponding prominence/boundary responses in a given language condition.

2.3. Syntactic parsing and acoustic measurements

Syntactic annotations following the Penn Treebank guidelines were available for both the Chinese Treebank 9.0 [27] and Penn Parsed Corpus of Modern British English (PPCMBE2) [25], where syntactic parsings for the chosen sentences were extracted. The numbers of left and right brackets between each pair of consecutive words were used as a proxy for the depth of the syntactic structure, capturing the additive strength of syntactic boundaries at the left and right edges of the constituent structure. By the design of the annotation guidelines, each word is wrapped by at least one left and one right bracket. To normalize the varied sentence lengths, the number of left or right brackets was divided by the maximum number of left or right brackets in the sentence.

The chosen sentences were aligned using the HMM-based Mandarin forced-aligner [29] and the Penn Phonetics Lab Forced Aligner [30] for Mandarin and English respectively. We took a range of acoustic measurements in Praat, including: (1) Whether there is a pause at the boundary junction as determined by whether silent portions were identified by the corresponding aligners; (2) the average syllable duration of each word; (3) the minimum and maximum fundamental frequency (F0), (4) the minimum and maximum sound

pressure level (SPL), (5) the alpha ratio (the level difference between the 1k-5kHz region and the 50-1kHz region), (6) L1-L0 (the level difference between the F1 region (defined as 300-800 Hz) and the F0 region (defined as 0-300 Hz), and (7) Cepstral Peak Prominence-Smoothed (CPPS). (3)-(7) were taken from voiced segments of the word. (5)-(7) as measures of voice quality are discussed in [31].

The data was then restructured such that each observation corresponded to a boundary, and acoustic measurements of the pre-/post-boundary words were reformalized as follows: The continuous acoustic measurements were first z-scored by speaker, then a difference in acoustic measure between pre- and post-boundary word was calculated except for (1) and (2). To capture the effects of phrase initial strengthening, the minimum of the pre-boundary word was subtracted from the maximum of the post-boundary word for F0 (3) and SPL (4) measurements. We also subtracted the acoustic measurements' mean value of the pre-boundary word from the post-boundary word for measures (5)-(7). We coded pause (1) as a binary grouped variable instead of a continuous measure as pause thresholds were defined arbitrarily by the aligners. Furthermore, the effect of pause was unlikely to be linear as most words do not have pauses between them (86% for English, 79% for Mandarin).

3. RESULTS AND DISCUSSION

We evaluate the effect of various variables on boundary perception using a linear mixed effects model in R in the lme4 package. Boundary perception ratings as a dependent variable were included in the model as the number of boundary responses for each token divided by the maximum number of boundary responses in a given language (numeric). We included the main effects of the proportion of pre- and post-boundary prominence responses in each given language (numeric), the pre- and post-boundary word's average syllable duration, F0, SPL, alpha, L1-L0 and CPPS differences (numeric), the proportion of left and right brackets (numeric), presence of pause (0 or 1) and language (2 levels: English [reference level], Mandarin). Two-way interactions between pause and every other main effect, as well as between language and every other main effect were included. And three-way interactions between pause, language, and all the acoustic, syntactic, and prominence perception variables were also included in the model, as were random intercepts for talkers. We computed p-values via Satterthwaite's degrees of freedom method using the lmerTest package.

As expected, a number of acoustic cues contribute to the boundary perception; importantly, the effects of acoustic cues are modulated by pause and language, as there are significant two-way interactions between acoustic cues (post-boundary syllable duration ($\beta = 8.33, p = 0.004$), F0 difference ($\beta = 12.35, p < 0.0001$)) and pause, and three-way interactions between language, acoustic cues (pre-boundary syllable duration ($\beta = -16.56, p < 0.0001$), F0 difference ($\beta = -8.80, p = 0.008$), SPL difference ($\beta = 7.00, p = 0.015$), L1-L0 difference ($\beta = 7.59, p = 0.029$) and pause. As illustrated in Figure 1, when there is no pause, longer pre-boundary syllable duration is related to higher rates of perceived boundaries for both languages; but when there is a pause, unlike English listeners, Mandarin listeners expect shorter pre-boundary syllables. As shown in Figure 2, for both languages, reset effects are generally stronger when there is a pause, but the effect size is greater for English listeners. In general, the presence of pauses is the dominant acoustic cue for Mandarin listeners, as exemplified by significant interaction effects between pause and language ($\beta = 62.18, p < 0.0001$).

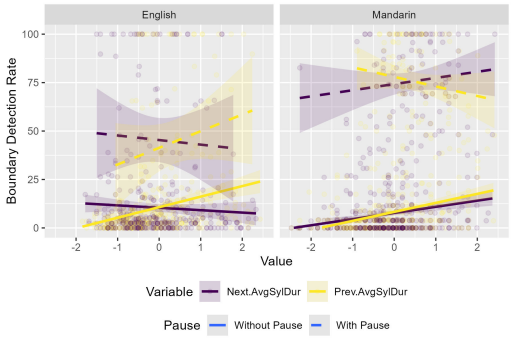


Figure 1: The relationship between average syllable duration and boundary perception.

As for syntactic constituency, significant main effects were found for both left ($\beta = 9.77, p = 0.013$) and right ($\beta = 32.76, p < 0.0001$) syntactic brackets, and for both languages, right brackets have a stronger effect. Therefore, both Mandarin and English align the right edge of syntactic boundaries with prosodic boundaries. Moreover, the syntactic effects are modulated by pause and language, indicated by two-way interactions between pause and brackets (left: $\beta = 31.18, p = 0.012$, right: $\beta = 29.04, p = 0.002$), between brackets and language (right: $\beta = -23.27, p < 0.0001$), as well as

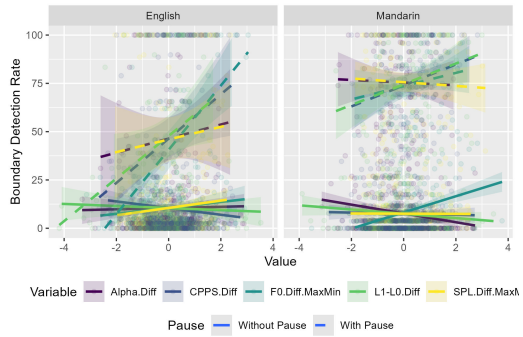


Figure 2: The relationship between other acoustic measures and boundary perception.

three-way interactions among brackets, pause and language (left: $\beta = -36.87, p = 0.011$). As illustrated in Figure 3, the effect of right syntactic brackets is stronger for English listeners, and for Mandarin listeners, syntactic boundaries largely co-vary with pause. Overall, syntactic cues play more important roles in detecting prosodic boundaries for English listeners, and pause plays a more important role for Mandarin listeners.

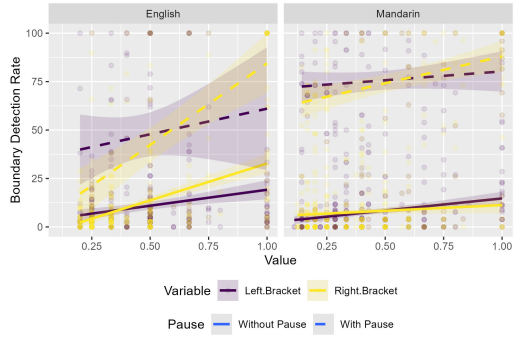


Figure 3: The relationship between syntactic brackets and boundary perception.

Finally, Figure 4 illustrates the effects of perceived pre-boundary prominence and post-boundary prominence on boundary perception. With English as the reference level, significant main effects of both prominence perception of the previous ($\beta = 0.19, p < 0.0001$) and following word ($\beta = 0.09, p = 0.009$) were found. There were also significant interactions between the pre-boundary prominence and language ($\beta = -0.22, p < 0.0001$). These results indicate that, for English listeners, prominence perception of the pre-boundary word has a positive effect on boundary perception, the effect of which appears stronger when accompanied by a pause. And there is a much weaker effect of the post-boundary word on boundary perception. Mandarin patterns differently from English, showing much weaker effects of prominence perception. More importantly, post-boundary prominence instead of pre-boundary prominence is positively correlated with boundary perception, showing the opposite direction from English. In other words, prominence marks the end of the domain for English listeners, whereas signals the beginning of the domain for Mandarin. This result is consistent with [23, 17] with a different experiment paradigm.

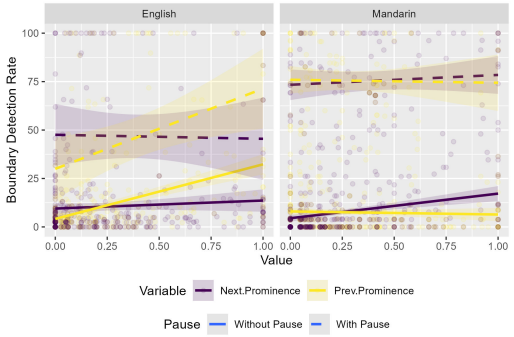


Figure 4: The relationship between promiennce and boundary perception separated by presence of pause.

4. CONCLUSION

This study examined the effects of acoustic boundary cues, syntactic constituency, and perceived prominence on boundary perception by investigating the boundary perception rates of utterances from continuous corpora in Mandarin and English. Results confirmed that boundary perception is shaped by complex interactions among acoustic cues, syntactic parsing, and prominence perception. Importantly, how these factors contribute to prosodic phrasing is language specific, as we show Mandarin and English pattern differently at various levels: 1) Pause is the determinant cue for Mandarin listeners, and final lengthening is not expected when a pause is present; but for English, final lengthening and reset effects are stronger when there is a pause. 2) English overall has stronger syntactic effects than Mandarin; syntactic boundaries largely co-vary with pauses in Mandarin. 3) Perceived prominence marks the right boundary for English, but indicates the left boundary for Mandarin. These findings open up more questions about the interface between prosody and other linguistic processing, such as how information structure and syntactic heading are encoded in prosody for different languages.

5. REFERENCES

[1] S.-A. Jun, *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press on Demand, 2007, vol. 1.

[2] D. R. Scott, "Duration as a cue to the perception of a phrase boundary," *The Journal of the Acoustical Society of America*, vol. 71, no. 4, pp. 996–1007, 1982.

[3] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[4] S. Chavarria, T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of ip and IP boundary levels: Comparison of l-and ll% in the switchboard corpus," in *Speech Prosody 2004, International Conference*, 2004.

[5] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[6] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology*, vol. 3, pp. 255–309, 1986.

[7] B. Connell and D. R. Ladd, "Aspects of pitch realisation in Yoruba," *Phonology*, vol. 7, no. 1, pp. 1–29, 1990.

[8] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[9] E. Bird and M. Garellek, "Dynamics of voice quality over the course of the English utterance," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 2406–2410.

[10] J. Slifka, "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of voice*, vol. 20, no. 2, pp. 171–186, 2006.

[11] J. Kuang, "The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL509–EL515, 2018.

[12] P. Keating, T. Cho, C. Fougeron, and C.-S. Hsu, "Domain-initial articulatory strengthening in four languages," *Phonetic interpretation: Papers in laboratory phonology VI*, pp. 143–161, 2004.

[13] T. Cho, J. M. McQueen, and E. A. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," *Journal of Phonetics*, vol. 35, no. 2, pp. 210–243, 2007.

[14] W. E. Cooper and J. Paccia-Cooper, *Syntax and Speech*. Harvard University Press, 1980.

[15] D. Watson and E. Gibson, "The relationship between intonational phrasing and syntactic structure in language production," *Language and Cognitive Processes*, vol. 19, no. 6, pp. 713–755, 2004.

[16] A. J. Schafer, S. R. Speer, P. Warren, and S. D. White, "Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task," *Approaches to studying world-situated language use*, pp. 209–225, 2005.

[17] S.-A. Jun and J. Bishop, "Priming implicit prosody: prosodic boundaries and individual differences," *Language and speech*, vol. 58, no. 4, pp. 459–473, 2015.

[18] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[19] A. Buxó-Lugo and D. G. Watson, "Evidence for the influence of syntax on prosodic parsing," *Journal of Memory and Language*, vol. 90, pp. 1–13, 2016.

[20] E. Selkirk and T. Shen, "Prosodic domains in Shanghai Chinese," *The phonology-syntax connection*, vol. 313, p. 337, 1990.

[21] J. Cole, J. I. Hualde, C. L. Smith, C. Eager, T. Mahrt, and R. N. de Souza, "Sound, structure and meaning: The bases of prominence ratings in English, french and Spanish," *Journal of Phonetics*, vol. 75, pp. 113–147, 2019.

[22] J. Bishop, G. Kuo, and B. Kim, "Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription," *Journal of Phonetics*, vol. 82, p. 100977, 2020.

[23] J. Kuang, "Prosodic grouping and relative clause disambiguation in Mandarin," in *Eleventh annual conference of the international speech communication association*, 2010.

[24] H. McGuire, "Librivox," Aug 2005. [Online]. Available: https://librivox.org/

[25] A. Kroch, B. Santorini, and A. Diertani, "The Penn Parsed Corpus of Modern British English (PPCMBE2)," *Philadelphia: Department of Linguistics, University of Pennsylvania.*, 2010.

[26] J. Kuang, M. P. Y. Chan, N. Rhee, M. Liberman, and H. Ding, "The mapping between syntactic and prosodic phrasing in English and Mandarin," *Proc. Interspeech 2022*, pp. 3443–3447, 2022.

[27] N. Xue, X. Zhang, Z. Jiang, M. Palmer, F. Xia, F.-D. Chiou, and M. Chang, "Chinese Treebank 9.0 LDC2016T13," *Philadelphia: Linguistic Data Consortium*, 2016.

[28] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," *Computer Speech & Language*, vol. 45, pp. 300–325, 2017.

[29] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2539–2543.

[30] J. Yuan, M. Liberman *et al.*, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[31] A. E. da Silva Antonetti, V. V. Ribeiro, A. G. Brasolotto, and K. C. A. Silverio, "Effects of performance time of the voiced high-frequency oscillation and lax vox technique in vocally healthy subjects," *Journal of Voice*, 2020.

BETWEEN TONE AND STRESS: ACOUSTIC EVIDENCE OF LEXICAL PROMINENCE IN SIRIANO

XXX

XXX
XXX

ABSTRACT

This paper provides the first phonetic study of lexical prominence in Siriano, an Eastern-Tukanoan language spoken in Colombia and Brazil, whose prosodic features are poorly understood. The language, along with its sister languages in the language family, is reported to have contrastive high (H) and low tone (L), but there is no consensus among researchers whether this is due to stress or lexical tone. The current study provides acoustic evidence indicating that certain syllables are more prominent than others within a word and shows that stress and lexical tone are independent properties of the system. While primary stress and lexical tone do not influence each other, voiceless obstruent onsets attract stress (aka onset-sensitive stress), the only factor influencing stress placement (default stress = 1st syllable, no codas allowed, no phonemic vowel length).

Keywords: Siriano, onset-sensitive stress, lexical tone, prosodic features, Eastern-Tukanoan language family

1. INTRODUCTION

The prosodic features of the Eastern-Tukanoan language family (ET languages) are still rather poorly understood. For instance, there is a long-time debate about what the contrastive pitch variation, high (H) and low (L), in lexical items in ET languages are: lexical tones [1, 2, 3], stress [4], or pitch accent [5].

Siriano, an ET language whose speakers mainly live along the Paca river and Caño Viña in Colombia and Brazil [6], is an understudied language with only less than 500 native speakers. Our current study provides acoustic evidence for Siriano being a language with with a prosodic system that contains both stress and lexical tone.

Theoretically, natural languages are divided into two categories in terms of quantity-sensitivity [7], which means whether or not a language makes use of syllable weight to determine the location of stress. Stress assignment in quantity-sensitive languages is sensitive to the structure of the syllable rime (nucleus and coda) and is usually claimed to disregard the structure of onsets [8]. Natural languages may differ in what kind of syllables they call heavy. For example, Walangama put its primary stress on CVV syllable, rather than CV or CVC syllables. Heavy syllables in this respect are the ones that contain a long vowel. By contrast, Latin prefers to assign stress on CVV and CVC, and treats CV as the light one [9]. Quantity-insensitive languages typically have fixed stress assignment position, and stress are placed on the same position in every word, such as Czech. In Siriano, codas are not permitted, and long vowels are not phonemic, which means only V and CV are allowed syllables. According to mainstream metrical theory [7], both syllable types thus have phonologically equal syllable weight.

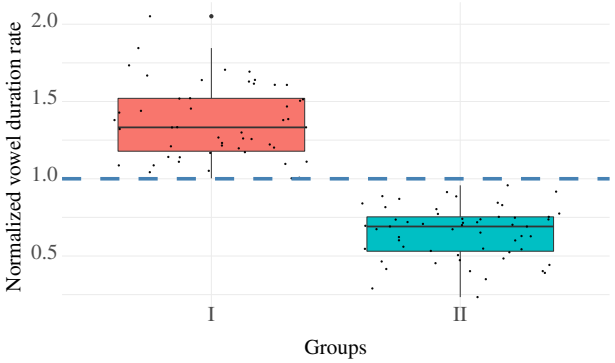


Figure 1: Relative vowel duration in disyllabic tokens with LH lexical melody.

On the other hand, we observed from elicited data that in certain disyllabic words, one syllable seems to have a longer vowel duration than the other. We divided the duration (ms) of the vowel in the first syllable by the duration of the vowel in the second syllable to normalize each word, and thus get a ratio δ. If $\delta > 1$, the first syllable has a longer vowel than the second. The result in Figure 1 shows that $\delta = 1.0$ (the blue dot line in the figure) happens to separate all disyllabic words with the same lexical melody (LH) into two groups. This indicates that in some words (GROUP I), the first vowel is phonetically longer, while other words (GROUP II) generally

have a longer second vowel. This essentially rules out the case that this pitch variation is only caused by lexical tones, as lexical tones are believed not to lead to vowel duration difference [10]. Further, the onsets of the second syllable in the GROUP II words are restricted to voiceless obstruents, namely [p^h, t^h, k^h, s, h, ʔ] in Siriano. We hereby have our research questions for this study: *Does this phonetic difference in vowel duration indicate the existence of stress?* and *Does the quality of onset consonant play a role in lexical prominence in Siriano?*

In the rest of this paper, I will first introduce the methodology, including the dataset and the experiment in Section 2, then I give the results and the corresponding discussions in Section 3. Lastly, I conclude in Section 4.

2. METHODOLOGY

2.1. Speaker backgrounds and the dataset

The materials for this study are drawn from the recording of spoken Siriano that were collected as part of Wilson Silva's original fieldwork conducted in 2018. The recordings were taken as conversations between Silva and a male native speaker of Siriano. The demographic information of this speaker is unknown. The original recording consists of four parts: isolated words, phrases, sentences, and monologues. The speech time is not controlled, and the speaker's speech rate varies throughout the conversations. In order to study lexical prominence, we therefore only include the isolated word and phrase recordings, because the sentence and monologue are affected by the speaker's listing intonation. Each isolated word is repeated twice or more by the speaker. The phrases are built up with these isolated words.

2.2. Annotation and measurements

All tokens were annotated by segment by hand via Praat [11], according to the available Siriano grammar description [6, 12] and a Siriano-Spanish dictionary [13]. Since codas are not allowed, we follow the three criteria, as shown below, to determine the syllabic or lexical boundary. We look for

- a sudden change of the formant values;
- a change of the shape of the periodicity;
- a voicing offset.

Tone, either L or H, on each syllable is determined based on the F0 analysis from ProsodyPro [14], in which eight F0 values are extracted in each normalized syllabic interval; F0 sample rate is 100 Hz. The results for each acoustic parameter are subsequently subjected to statistical analysis in R [15]: a Welch's *t*-test is conducted when comparing between different groups, given that each group has different amounts of tokens. Our dataset shows that all the tokens must have at least one H tone. For all disyllabic tokens, we have 3 possible lexical melodies: LH, HH and HL, while in all trisyllabic tokens, we have 7 possible lexical melodies: only LLL is not found. The count of each lexical melody is shown in Figure 2. Since we only have plenty of data with particular lexical melodies, we put our focus on disyllabic words with LH lexical melody, and trisyllabic words with LHH, LHL and HLH lexical melodies. [1]

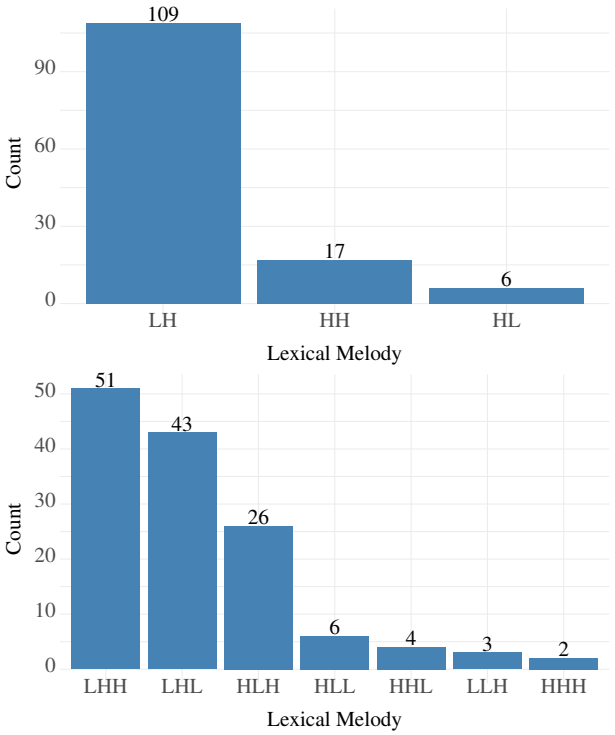


Figure 2: The count of each lexical melody based on lexical tone in the dataset for the current study. Upper: the count of the 3 possible lexical melodies of all disyllabic tokens; Lower: the count of the 7 possible lexical melodies of all trisyllabic tokens.

2.3. Procedure

All the data with certain lexical melody types are classified into groups based on the onset of the non-initial syllables. For disyllabic words with LH, GROUP II has all the words whose second syllable has voiceless obstruents as onset only, while the rest belong to GROUP I. For trisyllabic words, they have two more groups, as shown below:

- GROUP I: Second and third syllables do not have voiceless obstruent onset;
- GROUP II: Onset of the second syllable is restricted to voiceless obstruents;
- GROUP III: Onset of the third syllable is restricted to voiceless obstruents;
- GROUP IV: Onset of the second and the third syllables are restricted to voiceless obstruents.

The count of each group with particular lexical melody is given in Table 1.

	LHH	LHL	HLH
I	27	24	26
II	16	15	-
III	4	4	-
IV	4	-	-

Table 1: The count of each group with respect to each lexical melody.

3. RESULTS

F0, or fundamental frequency, is believed to be the most important acoustic correlate of lexical prominence, followed by vowel duration, vowel quality, etc. [9, 16], we have our comparisons of F0 between syllables with the same tone from different groups.

3.1. Disyllabic tokens: LH

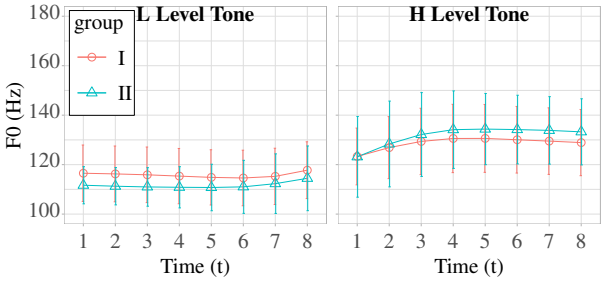


Figure 3: The F0 visualization of disyllabic tokens with LH melody.

The averaged time-normalized F0 traces (with 8 equidistant points) for disyllabic tokens with LH lexical melody are given in Figure 3. The error bars here reflect p-value of 0.05 (same below). The two panels represent the two syllables. Along with Figure 1, the first syllable (left panel) from the tokens in GROUP I is reliably different from that in GROUP II in that the first syllable has greater duration and higher F0 values on average (F0 values: $t = -7.2747$, $p < .001$), even though the first syllables of the tokens from both groups all bear L level tone. Conversely, the second syllables (right panel) in the tokens from GROUP II tend to have greater duration, and higher F0 values on average, even though all tokens have H level tone on the second syllable regardless of group. The comparison between the two groups in the interval, $4 \leq t \leq 8$ when values have stabilized, shows that GROUP II has significantly higher F0 values than GROUP I ($t = 10.663$, $p < .001$).

If a greater vowel duration and a higher pitch indicate metrical stress, then tokens in GROUP I have their primary stress on the first syllable, and stress is placed on the second syllable when voiceless obstruents are found to be onsets of the second syllable in disyllabic tokens.

3.2. Trisyllabic tokens: LHH

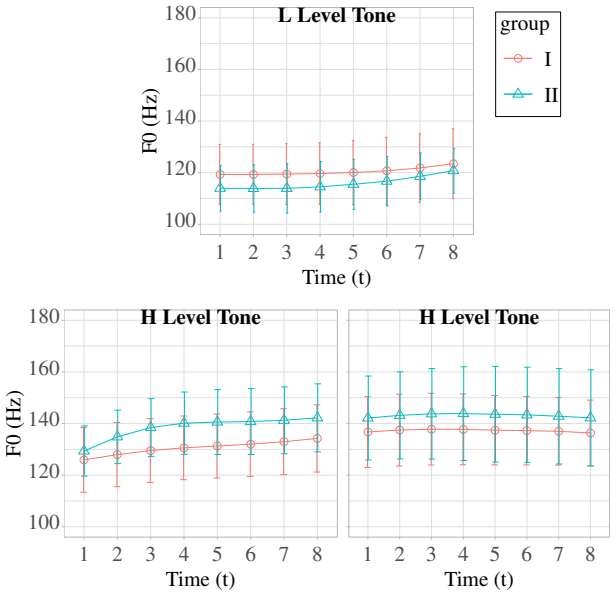


Figure 4: The F0 visualization of trisyllabic tokens with LHH melody.

Figure 4 contains average F0 traces for trisyllabic tokens with lexical melody LHH. The first syllable with level tone L is given in the top panel, while the second and third syllables with level tone H are given in the bottom ones. We can observe the same pattern here as the one we get in disyllabic tokens with LH. The first syllables bear L, under which circumstance GROUP I has higher mean F0 value than GROUP II. The reverse is also true for the second syllable. The third syllable shows the same difference as the second. In addition, in the second syllable of GROUP II, the F0 curve reaches the peak earlier at $t = 3$, while the F0 curve of GROUP I keeps climbing until the end of the second syllable,

indicating that the primary stress is on the second syllable when voiceless obstruents are in the onset position.

3.3. Trisyllabic tokens: LHL

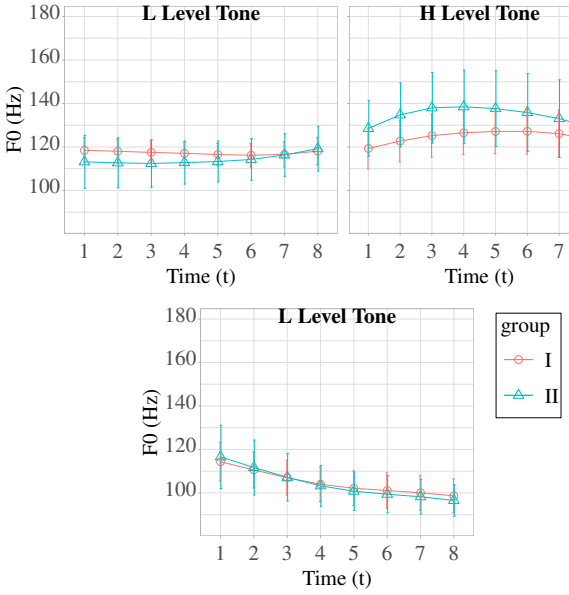


Figure 5: The F0 visualization of trisyllabic tokens with LHL melody.

Figure 5 contains average F0 traces for trisyllabic tokens with lexical melody LHL. First two syllables (top panel), with L and H tone respectively, again show the same pattern as the previous two cases. An early rise in the first syllable at $t = 6$ and a peak in the second syllable at $t = 3$ are observed as well. However, the third syllable (bottom panel) seems unexpected. The two groups (I and II) have overlapping mean F0 values ($t = 0.165$, $p = .8715$), showing no difference statistically. Compared to the first syllable with same L tone, a one-way ANOVA test shows that both curves are significantly lower than the unstressed L tone (first syllable) in GROUP II ($F = 9.01$, $p < .001$). Additional measurements, with GROUP III and IV show that when voiceless obstruents are in onset position of the third syllable, the mean F0 values of the four groups also show no significant difference ($F = 1.03$, $p = .375$), which indicates that these voiceless obstruents do not attract stress to the third syllable.

Among all 43 trisyllabic tokens with LHL, 41 tokens are bimorphemic, such as a nominal root (disyllabic) with a plural or a classifier suffix (monosyllabic). We think that in the metrical structure of a trisyllabic word with LHL, phonologically, the first two syllables from a nominal root form a foot, while the third syllable from a suffix is not a part of a prosodic word, and thus it is extrametrical, or not parsed. Consequently, when a voiceless obstruent is in the onset position of the third syllable, a stress will never be attracted out of the prosodic word.

3.4. Trisyllabic tokens: HLH

Figure 6 shows the mean F0 curves for trisyllabic words with HLH lexical melody. As shown in Table 1, tokens with HLH all belong to GROUP I, in which the two H tones, shown in left panel, are significantly different from each other ($t = -6.3938$, $p < .001$). This essentially tells us that the third syllable is primarily stressed.

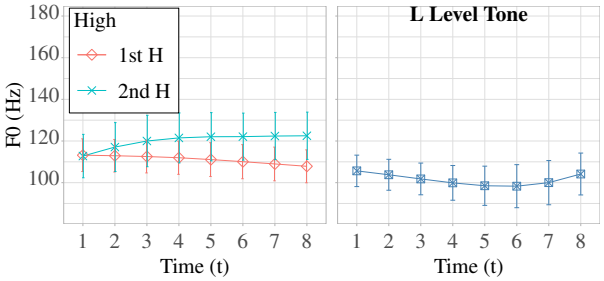


Figure 6: The F0 visualization of trisyllabic tokens with HLH melody.

4. CONCLUSION

We investigated the acoustic evidence for the existence of stress in Siriano, which is a separate system from lexical tone. Stress placement is not affected by the lexical tone type, in that both L and H tone can be stressed. The stressed lexical level tone essentially has higher pitch, compared to the unstressed one. L tone does not switch to H tone when it is stressed. Siriano is a trochaic language, with its primary stress assigned on the first syllable in a word by default except words with HLH lexical melody. A stressed syllable within a word tends to have a longer vowel duration phonetically. Our experiments show that the only effective factor influencing stress placement is the onset consonants. In particular, within a word when a voiceless obstruent in Siriano is in the onset position of the second syllable, stress is attracted to this syllable. But these consonants lose their power when they are as onsets in the third syllable of the words with LHL lexical melody. We hereby propose that Siriano is one of the languages that have onset-sensitive stress system.

5. REFERENCES

[1] E. Gomez-Imbert, "Morphologie et phonologie barasana: approche non-linéaire," Ph.D. dissertation, Paris 8, 1997.

[2] ——, "More on the tone versus pitch accent typology: Evidence from barasana and other eastern tukanoan languages," *Cross-Linguistic Studies of tonal phenomena: Tonogenesis, Japanese accentology, and other topics*, pp. 369–412, 2001.

[3] E. Gomez-Imbert and M. Kenstowicz, "Barasana tone and accent," *International Journal of American Linguistics*, vol. 66, no. 4, pp. 419–463, 2000.

[4] L. M. Hyman, "Tone and/or accent. elements of tone, stress and intonation, ed. by donna jo napoli, 1-20," 1978.

[5] J. Barnes, "Autosegments with three-way lexical contrasts in Tuyuca," *International Journal of American Linguistics*, vol. 62, no. 1, pp. 31–58, 1996.

[6] T. Ni, "The Phonology and Morphology of Siriano: A Grammar Sketch," Master's thesis, The University of Arizona, 2021.

[7] B. Hayes, *Metrical stress theory: Principles and case studies*. University of Chicago Press, 1995.

[8] M. Halle and J.-R. Vergnaud, "Three dimensional phonology," *Journal of Linguistic Research*, vol. 1, pp. 83–105, 1980.

[9] R. W. N. Goedemans, "Weightless segments - a phonetic and phonological study concerning the metrical irrelevance of syllable onsets," Ph.D. dissertation, Leiden University, 1998.

[10] I. Lehiste, "Suprasegmentals." 1970.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[12] L. Criswell and B. Brandrup, "Un bosquejo fonológico y grammatical del siriano," in *Lenguas indígenas de Colombia: una visión descriptiva*, M. S. González de Pérez and M. L. Rodríguez de Montes, Eds. Bogotá: Instituto Caro y Cuervo, 2000, pp. 395–418.

[13] B. A. c. Brandrup, *Vocabulario siriano-español*. Lomalinda: ILV/MG, 1980.

[14] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis." Laboratoire Parole et Langage, France, 2013.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[16] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[1] The non-accessibility of the data with particular lexical melodies could be the result that, for example, LH is preferred in disyllabic words in this language.

The Phonetics of Downtrends in Chokri

Tulika Gogoi, Sekholu Tetseo, Amalesh Gope,

Department of Linguistics and Language Technology, Tezpur University, India

tulikagogoi303@gmail.com, sekholu7@gmail.com, amalesh@tezu.ac.in

ABSTRACT

This paper investigates f0 downtrend processes in Chokri, a Tibeto-Burman language spoken in Nagaland, India. Three types of downtrends- temporal declination, final lowering, and automatic downstep, are considered in this study. The mathematically fitted f0s were used to examine the dependency of f0 on different tonal sequences of syllables (viz., all high, all mid, all low). The results confirm a general time-dependent declination in all high, mid, and low tones sequences. Furthermore, the fitted equations confirm that all high and low tone sequences follow a declining linear trend with an adjusted R-squared value of 0.79 and 0.87, respectively. In contrast, the mid-tone sequence follows an exponential decay, with a final lowering. The language indicates the presence of an automatic downstep when there is an intervening L tone between two H tone sequences.

Keywords: downtrends, prosody, declination, downstep, mathematical modeling.

1. INTRODUCTION

Researchers working on intonation in tonal and non-tonal languages have analyzed downtrend patterns, i.e., the gradual lowering of f0 from the starting towards the end of utterances. This implies that the pitch of a tone occurring later in the utterance will be lower than that of the same tone occurring at the beginning. Although debatable, such progressive pitch descent is attributed to phonetic effects [1]. Downtrends include (i) declination- the time-dependent lowering of f0, (ii) downstep- the lowering of f0 of H tones induced by the intervention of an L tone; and (iii) final lowering [2]. In intonation-only languages, studies on downtrends are focused chiefly on declination. However, downtrend analysis in many African tone languages reports the presence of downstep, a process involving H and L tones occurring alternatively, wherein the later H in an HLH sequence has a lower f0 due to the influence of the intervening L tone [1].

Declination refers to the global downward trend of f0 contour stretched over phrases or utterances. [3, 4, 5, 6, 7, 8] It is the "gradual modification (throughout a phrase or utterance) of the phonetic backdrop against which the phonologically specified F0 targets are scaled" [9]. Generally considered a phonetic effect, declination is considered universal in declarative utterances across languages. However, many researchers observed the absence of declination in some tonal languages, which has been one of the exciting findings from the works on intonational downtrends. For example, declination and final lowering are not found in Yoruba for sentences with all-H and all-M tones [9]. In Mambila, Tone 1, the highest tone, shows little tendency for downtrends [1]. While the declination effect can be seen for all tones, in tone languages, they are clearer through analysis of sentences containing words with the same lexical tones, like all H or all L sequences.

Final lowering is the more abrupt and excessive lowering of f0 at the end of speech domains like phrases or utterances [2, 6, 10]. Welmers [11] reports the presence of final lowering in several discrete level tone languages of Africa.

Another type of downtrend frequently reported in African tone languages concerns alternate H and L tone sequences. H tone following an L tone is realized with a lowered F0 compared to the H tone preceding an L tone. This process is called automatic downstep or downdrift [8]. Non-automatic downstep, on the other hand, is the lowering of an H tone without a conditioning L tone being physically present. A floating L tone triggers such a downtrend. This study investigates how different types of (f0) downtrends affect sentence prosody in Chokri.

2. THE LANGUAGE: CHOKRI

Chokri is a Tibeto Burman language primarily spoken in the Phek district of Nagaland, North-East India. The language is spoken by the Chokri-mi or Chokri subgroup of the Chakhesang tribe. The 2011 Census Reports of India estimates a total of 111,062 Chokri speakers. Bielenberg and Nienu [12] report the presence of 39 consonants, six vowels, and four tones in the language. However, our study shows that the segmental phoneme inventory of Chokri consists of 33 consonants, 7 vowels, and 5 lexical tones (four level tones- Extra High, High, Mid, and Low, and a contour (mid-rising) tone) [13].

3. EXPERIMENTAL PROCEDURE

We recorded speech data from 5 native Chokri speakers (three male, two female, and aged between 18 and 55 years) in the form of scripted sentences. The subjects were shown each sentence on a

computer screen and instructed to produce them naturally. Each speaker repeated the entire dataset five times, with a considerable gap between each repetition. Individual sentence files were segmented at the syllable level using PRAAT (version 6.0.43) [14]. The mean f0 and time-normalized f0 values were extracted using the ProsodyPro [15]. The resulting values were plotted as line graphs for visual inspection. A one-way repeated measure ANOVA (RM ANOVA) was performed to verify the statistical significance of the observed f0 differences.

The corpus used for this study consisted of three datasets containing 58 sentences. Following is the structure of the dataset used for the experiment:

A. All high-tone (henceforth H) sequences
   All low-tone (henceforth L) sequences
   All mid-tone (henceforth M) sequences
B. H tone sequences with an intervening L

4. RESULTS

4.1. Declination

Production experiments were conducted to examine the trends of declination in all L, H, and M tone sequences in Chokri. Results show that the (phonetic) declination process is present in this language for all three tone sequences, viz., H, L, and M tone sequences. However, the rate of decline is not constant for all syllables. Time-normalized average f0 values calculated for each speaker showed that declination is steepest on the first syllable of the utterance. The first drop in pitch is realized during the production of the very first syllable. The rate of declination then becomes progressively slower as the utterance progresses. While the phonological specification of all the syllables is the same (High), the phonetic value of f0 of the syllables occurring late in utterance is less compared to the ones in the beginning. Figures 1-4 represent the time-normalized average f0 trend for male and female speakers for H and L tones. The error bars depict the standard error of the aggregated data.

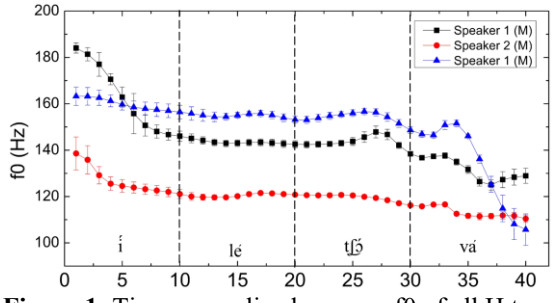


Figure 1: Time-normalized average f0 of all H tone sequence /í lé ʧɔ́vá/ *I am cooking* by 3 male speakers.

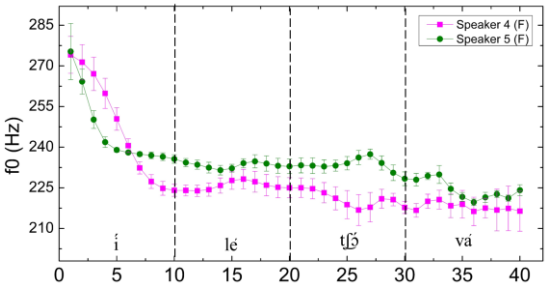


Figure 2: Time-normalized average f0 of all H tone sequence /í lé ʧɔ́vá/ *I am cooking* by 2 female speakers.

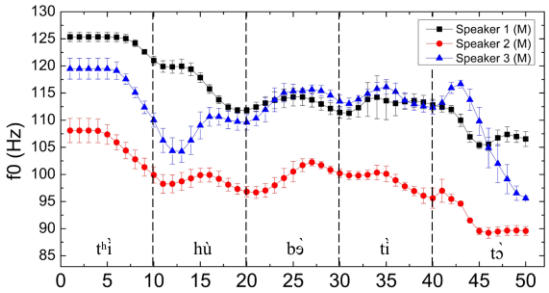


Figure 3: Time-normalized average f0 of all L tone sequence /tʰì hù bɔ̀ tì tɔ̀/ *(I) will boil some meat and eat* by 3 male speakers.

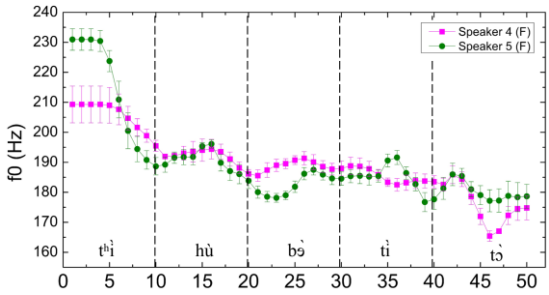


Figure 4: Time-normalized average f0 of all L tone sequence /tʰì hù bɔ̀ tì tɔ̀/ *(I) will boil some meat and eat* by 2 female speakers.

4.2. Final Lowering

Final lowering is not observed in the sentences featuring the sequences of all H and L tones across speakers' repetitions considered in this study. However, the sentences featuring the sequences of all M tones indicate a final lowering in Chokri.

4.3. Automatic Downstep

We also examined sentences where an intervening L occurs in an all-H sequence to explore the possibility of automatic downstep in Chokri. Results indicate that the intervening L might influence the preceding and the following H tone syllables. The visual observations of each speaker's time-normalized contours across repetitions indicate that an H tone appearing after an L has relatively lower f0 values than the H tones preceding the L (Figure 5). The precise nature of automatic downstep is discussed in section 5.

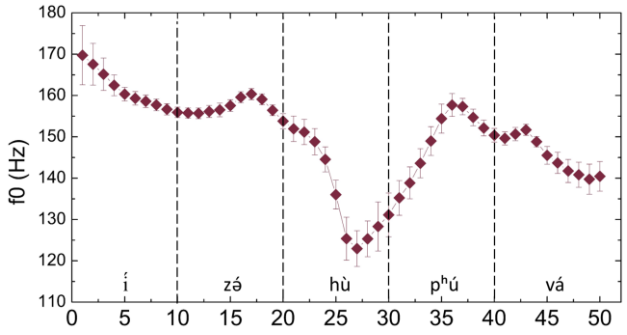


Figure 5: Time-normalized average f0 of all tokens of the sentence / í zə́ hù pʰú vá/ *I am searching for some water* by all 5 speakers.

5. DISCUSSIONS

To examine whether the declination patterns (Figures 1 - 4, for examples) observed in all H, M, and L tone sequences are significant, we conducted one-way repeated measured ANOVA. The results confirm that f0 differences observed in each syllable are significantly different in the sequence of all H ($F[3, 12] = 66.7$, $p < 0.0001$*), all M ($F[5, 20] = 30.83$, $p < 0.0001$*), and all L ($F[3, 12] = 69.52$, $p < 0.0001$*), respectively. A subsequent post hoc Bonferroni test confirmed a significant interaction between each syllable in the sequence of all H, M, and L tones, respectively.

Different mathematical equations, viz., (i) linear equations (for the sentences with all H and L tone sequences) and (ii) exponential decay (for the sentences with all M tone sequences), are individually fitted to examine the slopes of declination (Figures 6 - 8). The following formulas for (a) a linear equation with $y = m*x + c$, where m is the slope and c is the intercept, and (b) an exponential decay equation with $y = a*\exp(-x/b) + d$, where a, b, and d are the constants, are incorporated to calculate the slopes. In figures 6 – 8, the Y-axis shows the f0 range, and the X-axis shows the number of syllables present in the target sentence.

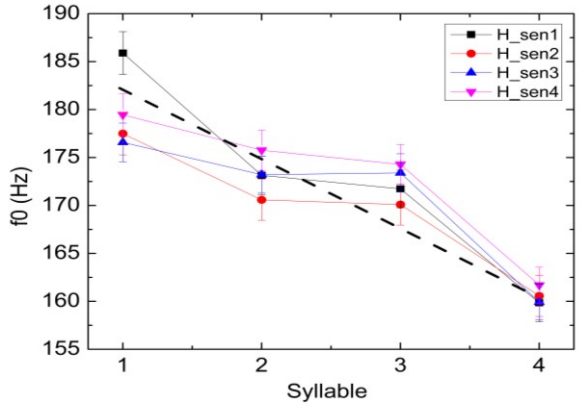


Figure 6: Time-normalized average f0 contours of all the sentences featuring all H tone sequences produced by all the subjects. The dashed solid line depicts the trends of declination.

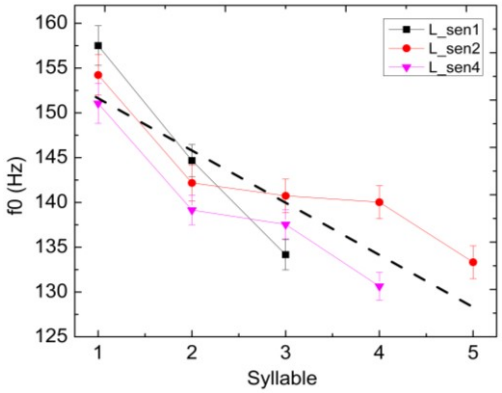


Figure 7: Time-normalized average f0 contours of all the sentences featuring all L tone sequences produced by all the subjects. The dashed solid line depicts the trends of declination.

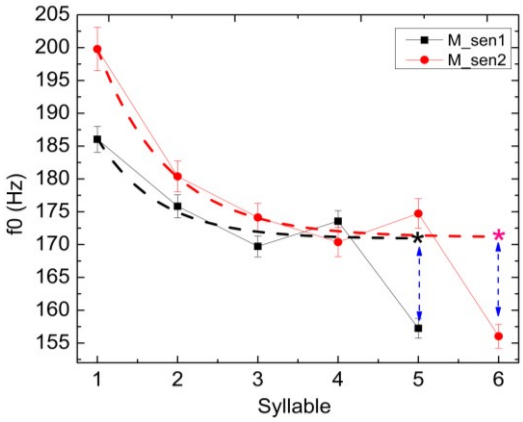


Figure 8: Time-normalized average f0 contours of all the sentences featuring all M tone sequences produced by all the subjects. The corresponding dashed lines exhibit the exponential decay fitted lines. The blue dashed line shows the distance between the fitted lines and the (actual) final point.

The visual observations of the f0 contours shown in Figures 6 – 8 could be misleading. The observation of the final syllable in all these contours may seem like a final lowering. However, the quantitative interpretation confirms that it is not the case. Only a linear line could be fitted for the sentences featuring the sequences of all H and L tones, respectively. The f0 range depicted in the Y-axis varies from ~188 to ~161 Hz (for the syllables with all H tone sequences) and ~157 to ~130 Hz (for the syllables with all L tone sequences). It implies that the f0 range between the successive syllables is relatively low (only 27 Hz) for an exponential decay to fit in. Furthermore, the f0 value of a particular syllable is relatively closer when compared to the preceding and the following syllables (of the target syllable). The adjusted R-squared values for the linear fit of all the sentences featuring all H tone sequences are 0.88, 0.85, 0.65, and 0.77, respectively. Similarly,

the adjusted R-squared values for the linear fit of all the sentences featuring all L tone sequences are 0.99, 0.77, and 0.84, respectively.

In contrast, the linear fit equation cannot the applied to the f0 contours of the sentences featuring all M tone syllable sequences (due to higher R-squared values). The R-squared values increased from 0.46 to 0.81 for M_sen1 and 0.47 to 0.92 for M_sen2. The increased R-square values confirm that the fitted equation is more justified for an exponential decay than a linear fit. When the exponential curves are extrapolated, it is also observed that the final point is relatively lower than the fitted point (~15 Hz), implying a *final lowering* in the syllable sequences featuring all M tones in Chokri.

	Fitted value	Std Error
M_sen1_d	170.87427	1.87488
M_sen1_a	57.52609	46.00498
M_sen1_b	0.75476	0.45813
M_sen2_d	171.10861	1.41236
M_sen2_a	90.08204	29.68016
M_sen2_b	0.87493	0.22661

Table 1: Fitted values for the exponential decay equation applied on sentences featuring all M tone sequences.

Table 1 tabulates the fitted values for the exponential decay, where the decay constant (b) is dependent on the syllable no., i.e., the value of "b" changes from 0.75 to 0.87, when the number of syllables in the sentences featuring all M tone sequences increases from 5 to 6.

It is to be noted that the number of syllables in the sentences featuring all H tone sequences is fixed at 4. The ranges of the slope values (m) observed in all four sentences featuring all H tone sequences are -7.85 ± 1.62 Hz, -5.12 ± 1.18Hz, -5.13 ± 1.98Hz, and -5.69 ± 1.69 Hz, respectively. The negative values confirm the presence of declination in Chokri. On average, the slope value (i.e., the change of the f0 value with respect to the syllable number) is observed to be a constant value of 5.94 ± 1.61 Hz.

The parameters of the mathematical fitted equations are exploited to examine the interactions between the (increased) syllable numbers and declination. The number of syllables in the sentences featuring all L tone sequences varies from 3 to 5. The range of slope values in the sentences containing all L tone sequences is -11.55 ± 0.71Hz (three syllables), -5.85 ± 1.39Hz (four syllables), and -4.21 ± 1.09 Hz (five syllables), respectively. The trend of slope values indicates an inversely proportionate correlation with the number of syllables present in a sentence, i.e., the slope values decrease with the increased number of syllables in a sentence (from -11.55 to -4.21 Hz). The sample size concerning the number of syllables featuring different tonal sequences is relatively small (n = 0 for all H tone sequences, 2 for all M tone sequences, and 3 for all L tone sequences). It is, therefore, beyond the scope of this approach to confirm the exact dependency of the syllable number on declination patterns.

Finally, to examine the effect of an intervening L tone on the preceding H (denoted as H_p) and following H (denoted as H_f) tone sequences, we have considered three different tonal sequences in this study, viz., HLHH, HHLH, and HHLHH. Figure 9 shows the time-normalized average f0 contours (averaged across all the repetitions produced by all the subjects) for the sentences with HLHH, HHLH, and HHLHH tonal sequences. The visual observations of the contours confirm that the intervening L tone, by and large, reduces the f0 of the H_f tone. The f0 of H_p also gets reduced, most likely as a result of the declination effect. A stiff fall on the intervening L syllable, followed by a linear f0 contour, could be observed in the HLHH and HHLH tonal sequences, indicating the presence of an automatic downstep in Chokri.

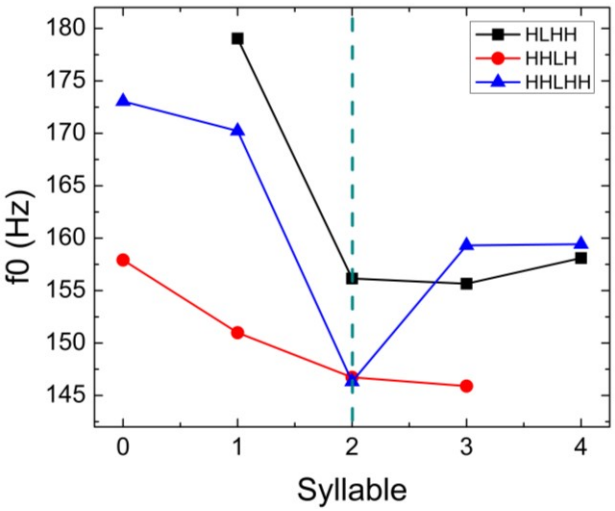


Figure 9: Time normalized average f0 for different syllable numbers. The green dashed line is the reference line to distinguish the preceding and following H tones.

5. CONCLUSIONS

This study shows the nature of f0 downtrends in Chokri. Temporal declination is present in *like level tone* sequences (H, L, and M) with statistically significant f0 differences among syllables. Final lowering is found only in all M sequences. Furthermore, this paper also confirms the presence of automatic downsteps in the language.

7. REFERENCES

[1] Connell, B. 2001. Downdrift, Downstep, and Declination. Typology of African Prosodic Systems Workshop Bielefeld University, Germany

[2] Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

[3] Connell, B. 2006. Four tones and downtrend: a preliminary report on pitch realization in Mambila, a language with four level tones, in *27th ACAL*, University of Florida, Gainesville, 29-31 Mar 1996 and *the LAGB*, University of Sussex, Brighton, 11-13 Apr 1996.

[4] Cohen, A., & ' t Hart, J. 1967. On the anatomy of intonation. *Lingua, 19* 177–192.

[5] Ladd, R. 1984. Declination: A review and some hypotheses. *Phonology Yearbook*, 1:53–74,.

[6] Liberman M &. Pierrehumbert J. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language Sound Structure*, pages 157–233. M.I.T. Press, Cambridge, Massachusetts.

[7] Pierrehumbert, J & Beckman M. 1988. *Japanese Tone Structure*. The MIT Press, Cambridge, Massachussetts,

[8] Stewart, J.M. 1983. 'Downstep and Floating Low Tones in Adioukrou.' *Journal of African Languages and Linguistics*, Vol 5, 57-78.

[9] Connell, B. & Ladd R. 1990. Aspects of pitch realization in Yoruba. *Phonology* 7: 1-30

[10] Truckenbrodt, H. 2004. Final lowering in non-final position. Journal of Phonetics. 32. 313-348. 10.1016/j.wocn.2003.11.001.

[11] Welmers, W. 1973. *African Language Structures*. Berkeley: University of California Press.

[12] Bielenberg, B., Nienu, Z. 2001. Chokri(Phek Dialect): Phonetics and Phonology. *Linguistics of the Tibeto Burman Area*. Vol 24.2, 85-122

[13] Tetseo, S., Gope, A. 2021. Tonal Contrast and Grammatical Tone in Chokri. [Paper Presentation] 43rd International Conference of Linguistic Society of India. CIIL-Mysore

[14] Boersma. P., Weenik D 2012. Praat: doing phonetics by computer (Version 5.3.04_win64) [Computer programme].

[15] Xu, Yi. 2013. ProsodyPro - A tool for large-scale systematic prosody analysis. Proceedings of the TRASP Conference. 7-10.

[16] Becker, L.A. 1979. 'Once more on the nature of downdrift.' *Studies in African Linguistics*, Vol 10. 3, 233–245.

[17] Strik H & Boves L.1995. Downtrend in f0 and psb. *Journal of Phonetics*, 23:203–220.

French Rhythm in Read Speech by French Natives and Chinese L2 Learners

Lei XI, Rachid Ridouane
Laboratoire de Phonétique et Phonologie (UMR 7018, CNRS/Sorbonne Nouvelle, Paris)
lei.xi@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

ABSTRACT

This study examines rhythmic patterns in French sentences read by 10 native French speakers and 21 Chinese learners of French (with different levels of proficiency). We used 9 metrics, both normalized and non-normalized, to compare the rhythmic patterns between natives and learners, and between the 3 groups of learners.

Overall, results from normalized metrics suggest that the rhythmic patterns of natives and learners are rather similar. Indeed, while raw metrics, %V, ΔV, ΔC, and rPVI-C, show significant differences both between natives and learners, and among learners, normalized metrics, VarcoC, VarcoV, and nPVI-V, show no such differences. These results suggest that traditional metrics may not capture changes over stages of rhythm acquisition, especially when L1 and L2 belong to the same rhythm typology. In addition, the CCI (Control/Compensation Index), well-distinguished among learners, shows that as acquisition progresses, learners gain a better timing control ability, and rhythmic pattern becomes more stable.

Keywords: rhythm metrics, L2 rhythm production, French, Mandarin Chinese.

1. INTRODUCTION

Speech rhythm of world languages has traditionally been classified into stress-timed (e.g., English, Dutch), syllable-timed (e.g., French [1], Chinese [2]), and mora-timed (e.g., Japanese) [3, 4, 5]. However, empirical evidence to assess this idealized surface isochrony in different languages is rather weak [6]. Departing from this traditional view, a series of metrics has been proposed in different studies for quantifying rhythmic phenomena based on vocalic and intervocalic interval durations [1, 5, 7]. Even though this quantitative approach has also been criticized [among others 8, 9], it is still a main approach used in the study of cross-linguistic speech rhythm and L2 speech rhythm [10]. In this study, 9 metrics, both normalized and non-normalized, are used to investigate the rhythmic patterns in French sentences produced by French native speakers and 3 groups of Chinese learners of French with different levels of proficiency.

Having a strong command of rhythmic patterns that are characteristic of a native speaker's speech is crucial for improving the intelligibility and reducing the foreign accent of a second language speaker [11, 12, 13, 14]. The ability to produce and perceive the rhythms of speech is therefore of great importance for learning a second language. In recent years, there has been a growing interest in understanding the temporal aspects of non-native speech. Factors such as the type of materials used, the demands of the task, the rate and style of speech, can all impact the rhythmic patterns of second language speech [10, 15]. While previous research has often compared the rhythmic patterns of native speakers to those of second language learners, there has been less focus on how the rhythmic production of a second language changes as the learner progresses.

There have been relatively few studies on Chinese learners of French and their rhythmic patterns. In one study [16], native Chinese speakers learning French as a second language were compared to native French speakers using two metrics, %V and VarcoV, and it was found that the two groups had virtually identical values. Another study [17] found that only the value of %V significantly differed between Chinese learners and native French speakers when examining six different metrics. These studies suggest that Chinese learners may be able to achieve native-like rhythms in French, potentially due to the shared rhythmic patterns between French and Chinese as syllable-timed languages.

This new study aims to further explore these findings by conducting a production experiment with additional rhythm metrics and involving Chinese learners at various proficiency levels. The research aims to answer two questions: 1) How do normalized and non-normalized rhythm metrics capture similarities and differences between French speech produced by learners and natives, and 2) Do Chinese learners' performances vary based on their proficiency levels?

2. METHOD

2.1. Speakers

Ten adult native speakers of French (5 males and 5 females, mean age: 31.3, SD: 8,37) and 21 Chinese learners participated in the production experiment. None of the participants reported any hearing

impairments or language impairments. The 21 learners completed a LEAP questionnaire [18] and provided information about their personal backgrounds and language proficiency. Based on their self-reported proficiency in French, the Chinese learners were divided into three groups (as shown in Table 1).

Level of Proficiency	Gender (M: male, F: female)	Mean Age (SD)	Mean Learning Time (year)
High	3M, 4F	27.00 (2.82)	8.6
Intermediate	2M, 5F	27.14 (2.91)	4
Low	2M, 5F	26.29 (4.15)	1.3

Table 1: Information about the Chinese learners.

2.2. Task and procedure

In this study, participants were asked to read 20 French sentences as part of a speech production task. These sentences were selected to contain a range of syllable lengths (from 7 to 27 syllables) and a variety of pauses and breaks. Each participant was asked to read each sentence at a normal speech rate and with neutral emotion, and to repeat any misread or disfluent sentences before moving on to the next. Each sentence was repeated twice, resulting in a total of 1240 sentences (400 produced by native speakers and 840 produced by Chinese learners). The production data was collected in a sound-proof room at the Laboratoire de Phonétique et Phonologie in Paris, and participants were recorded individually using the Pro Tools program and an AKG C520L wireless head-worn microphone (with the microphone positioned 3 cm away from the mouth). The data was recorded at a 44.1kHz sampling rate and 16-bit quantization.

2.3. Segmentation and measurements

The speech samples were automatically segmented using SPPAS [19] and then manually corrected by the authors. The data was then annotated in Praat [20] in two steps: 1) phonetic segmentation of the sentences into phonemes, and 2) classification of individual phonemes as vowels or consonants. The annotation and classification techniques from [15] were used. Pauses were marked as "#" and excluded from metric calculations. A total of nine rhythm metrics were used in this study (as shown in Table 2). In addition to traditional metrics (such as ΔV, ΔC, %V [5]; VarcoV, VarcoC [7], rPVI, nPVI [1]), values for CCI-V and CCI-C [21] were also obtained. The CCI index (Control/Compensation Index), which is a model for formally representing the rhythmic tendencies of natural languages, can be used to calculate the level of compression (lengthening or shortening) in language production. According to this model, "controlling" languages (such as syllable- or mora-timed languages) should have similar tendencies of vocalic and intervocalic duration fluctuations; while "compensating" languages (such as stress-timed languages) should fluctuate more in vocalic than in intervocalic intervals [22].

Metrics	Description
%V	Proportion of vocalic intervals duration to the duration of the utterance.
Δ	Standard deviation of vocalic (ΔV) or intervocalic (ΔC) intervals duration.
Varco	Coefficient of variation of vocalic (VarcoV) or intervocalic (VarcoC) intervals, i.e., ΔV or ΔC divided by the mean.
rPVI	Raw pairwise variability index for intervocalic intervals, i.e., mean of differences between successive segment intervals.
nPVI	Normalized rPVI for vocalic intervals divided by their sum and multiplied by 100.
CCI	rPVI relativized to the number of segments composing successive vocalic (CCI-V) or intervocalic (CCI-C) intervals.

Table 2: The rhythm metrics examined in this study (with traditional metrics in bold).

A total of 1240 textgrid files of segmented sentences were proceeded in Correlatore [23]. The nine metrics were calculated for each sentence and compared among the four groups.

3. RESULTS

The results for the global raw and normalized metrics [5, 7] are presented first, followed by the pairwise metrics [1, 21].

3.1. Global metrics

3.1.1. ΔC, ΔV and %

The results for the unnormalized variability metrics ΔV and ΔC are depicted in Fig. 1.

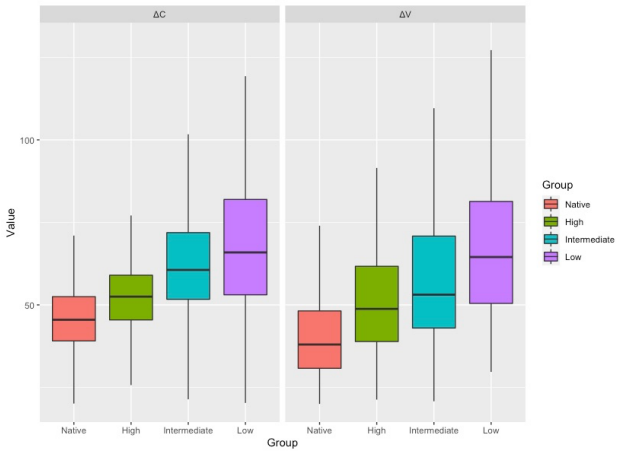


Figure 1: Means and standard errors of the metrics ΔC and ΔV for the four groups.

These results indicate that the learner groups exhibit larger values of ΔV and ΔC than the native group, with proficiency level showing an inverse relationship with Δ . One-way ANOVAs were conducted separately for each metric value, with group of proficiency as the between-subjects factor and the metric value as the dependent variable. The statistical tests revealed significant differences between learners and native speakers for ΔV (F-value = 122.7, $p < 0.001$) and ΔC (F-value = 118, $p<0.001$). To determine whether the metrics differ between the high, intermediate, and low proficiency groups, Tukey post-hoc tests were applied, which showed significant differences between the groups (pairwise comparisons for ΔV and ΔC between high vs intermediate, intermediate vs low, and high vs low: all at $p<0.001$).

The results for %V are presented in Fig. 2. It can be observed that the mean value of %V is larger in low proficiency learners than in intermediate-high proficiency learners (high: 48.40; intermediate: 48.36; low: 50.74). When compared to native speakers (mean %V: 46.39), post-hoc pairwise tests revealed that %V values between learners and natives are significantly different, with the exception of %V ($p = 0.99$) in the intermediate vs high proficiency groups.

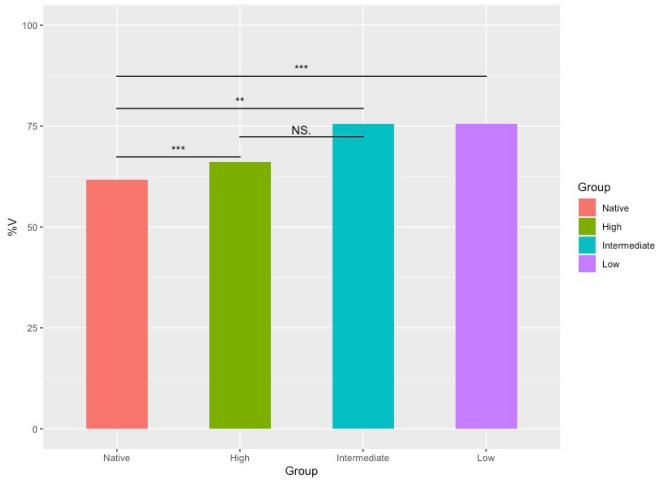


Figure 2: %V values of the four groups and post-hoc tests results.

3.1.2. VarcoV and VarcoC

The normalized variability metrics are presented in Figure 3. The results showed that both native and learner groups had virtually similar values of VarcoV and VarcoC. The mean values of the metrics before and after normalization, shown in Table 3, indicate that the patterns produced by learners were comparable to those of native speakers after accounting for speech rate. The raw and normalized metrics also demonstrated a significant effect of speech rate on learner productions.

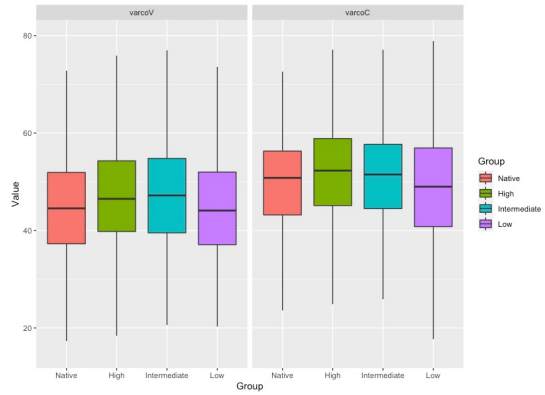


Figure 3: Means and standard errors of the metrics VarcoV and VarcoC for the four groups.

Metric	Proficiency			
	Native	High	Intermediate	Low
ΔV	39.90	52.01	57.69	70.48
VarcoV	45.67	47.46	47.36	44.91
ΔC	46.87	56.83	62.03	74.51
VarcoC	50.17	52.00	51.41	50.19

Table 3: Mean values of raw and normalized metrics.

Tukey post-hoc comparisons (Table 4) showed that there was no significant difference in the durational variability of vocalic/intervocalic intervals between learners and native speakers.

Metric	Significance of contrasts		
	Native-High	Native-Intermediate	Native-Low
VarcoV	0.15	0.19	0.92
VarcoC	0.11	0.35	0.98

Table 4: Post-hoc tests for Varco comparison between groups.

3.2. Pairwise metrics

3.2.1. rPVI & nPVI

The pairwise metrics results for nPVI-V and rPVI-C are shown in Fig. 4.

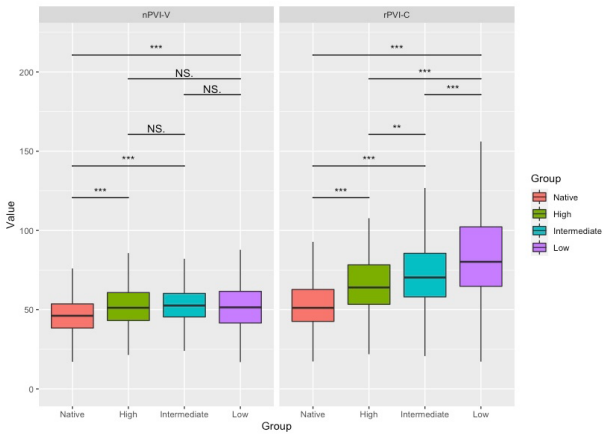


Figure 4: PVI values of the four groups and post-hoc tests results.

Native speakers displayed less variation in nPVI-V values compared to learners. There were no significant differences among the three learner groups (as shown in Fig. 4 on the left), indicating similar performances in successive vocalic intervals after controlling for speech rate. On the other hand, rPVI-C values indicated that lower proficiency was associated with greater temporal variability in successive intervocalic interval productions.

3.2.2. CCI

The average values of CCI-C and CCI-V are depicted in Fig. 5. Native speakers and learners are separated by the bisector. The productions of native speakers and advanced learners are close to the bisecting line, indicating similar controlling tendencies in vocalic and intervocalic durational fluctuations.

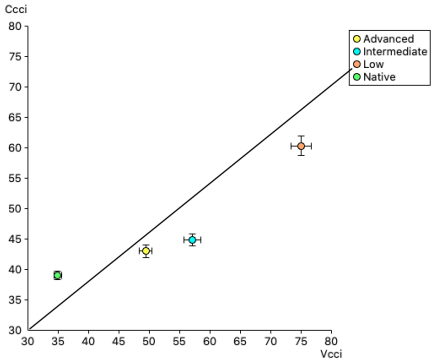


Figure 5: Distribution of CCI values for four groups of subjects.

Native speakers displayed the greatest degree of control, as indicated by the smallest values of CCI-V and CCI-C. In contrast, values for intermediate and low proficiency learners were farther from the bisecting line, indicating a pattern that is more compensating than controlling during speech articulation. Low proficiency learners showed much more fluctuation in both vocalic and intervocalic intervals compared to the other learners.

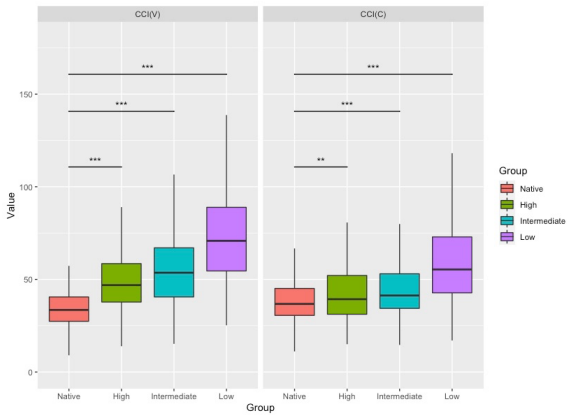


Figure 6: CCI values of the four groups and post-hoc tests results.

The statistical analysis reveals significant differences between natives and learners (Fig. 6): for CCI-V: F-value=226.3, $p<0.01$; for CCI-C: F-value=77.11, $p<0.01$.

4. DISCUSSION AND CONCLUSION

This study compared the rhythmic patterns of French read speech between native speakers and three groups of Chinese learners with low, intermediate, and high levels of proficiency using nine rhythm metrics.

The results for non-rate-normalized metrics indicated differences among all groups. However, these metrics may not be reliable indicators of differences due to their sensitivity to speech rate [15]. In L2 speech, slower speech rates often result in the realization of more and longer pauses, and thus the formation of more shorter prosodic units. Given that prosodic units are often accompanied by segmental lengthening at boundaries, the insertion of more prosodic boundaries leads to more lengthening, which may contribute to increased variation in duration and larger raw durational metrics [24, 25]. This tendency was more evident in the low proficiency group than in the advanced group. Other L2 rhythm metric studies (e.g. [24, 26]) have also reported similar tendencies in raw metrics. Previous research [15, 27] has demonstrated that %V, VarcoV, and nPVI-V are more robust to variation in speech rate and are effective at discriminating cross-linguistic speech rhythm. In the present study, the VarcoV and VarcoC values for learners in the three proficiency groups approximated the target pattern when the metrics were normalized for speech rate. This may be due to the shared phonological properties of French and Chinese, which are classified as syllable-timed languages: 1) the prevalence of the same CV syllable structure, and 2) the rarity of instances of vocalic reduction [28].

While traditional methods of measuring rhythm have difficulty in accurately assessing the development of rhythm in Chinese speakers learning French at different stages of acquisition, the CCI index does offer some insight into their progress. However, it is not sufficient for fully understanding all issues related to rhythm acquisition. One particular challenge for Chinese learners is their ability to correctly group rhythms, as non-native speakers often have less fluency in their speech and tend to divide an utterance into smaller chunks, resulting in more variability in duration at the segment level.

This research is part of a larger project that also investigates final lengthening and F0 movement. Future publications will provide additional findings on Chinese learners' progress in acquiring French speech rhythm.

5. ACKNOWLEDGEMENTS

This study was supported by the French ANR (ANR-10-LABX-0083, Labex EFL).

6. REFERENCES

[1] Grabe, E., Low, E. 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. *Papers in laboratory phonology* 7, 515-546.

[2] Lin, H., Wang, Q. 2007. Mandarin Rhythm: An Acoustic Study. *Journal of Chinese Language and Computing* 17(3), 127-140.

[3] Pike, K. 1945. *The Intonation of American English*. University of Michigan Press.

[4] Abercrombie, D. 1967. *Elements of general phonetics*. Edinburgh University Press.

[5] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265-292.

[6] Turk, A., Shattuck-Hufnagel, S. 2013. What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology* 4(1), 93-118.

[7] Dellwo, V. 2006. Rhythm and Speech Rate: A Variation Coefficient for deltaC. In: Karnowski, P., Szigeti, I., *Language and language-processing*. Peter Lang, 231-241.

[8] Arvaniti, A. 2012. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40, 351-373.

[9] Arvaniti, A. 2009. Rhythm, Timing and the Timing of Rhythm. *Phonetica* 66, 46-63.

[10] Gut, U. 2012. Rhythm in L2 speech. *Speech and Language Technology* 14/15, 83-94.

[11] Kolly, M., de Mareüil, P., Leemann, A., Dellwo, V. 2017. Listeners use temporal information to identify French- and English-accented speech. *Speech Communication* 86, 121-134.

[12] Kolly, M., Dellwo, V. 2014. Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition. *Journal of Phonetics* 42, 12-33.

[13] Polyanskaya, L., Ordin, M., Busa, M. 2017. Relative Salience of Speech Rhythm and Speech Rate on Perceived Foreign Accent in a Second Language. *Language and Speech* 60(3), 333-355.

[14] Polyanskaya, L., Ordin, M., Ulbrich, C. 2013. Contribution of Timing Patterns into Perceived Foreign Accent. *Elektronische Sprachsignalverarbeitung* 71-79.

[15] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., Mattys, S. 2010. How stable are acoustic metrics of contrastive speech rhythm? *Journal Acoustical Society of America* 127(3), 1559-1569.

[16] Gabriel, C., Stahnke, J., Thulke, J. 2014. On the acquisition of French speech rhythm in a multilingual classroom: Evidence from linguistic and extra-linguistic data. *Actes de Congrès Mondial de Linguistique Française*, 1267-1283.

[17] Pillot-Loiseau, C., Xie, H. 2018. Transfert rythmique du chinois mandarin au français dans l'apprentissage du Français Langue Étrangère : acoustique et perception. *Actes de Congrès Mondial de Linguistique Française* 46, 09001.

[18] Marian, V., Blumenfeld, H., Kaushanskaya, M. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research* 50, 940-967.

[19] Bigi, B. 2015. SPPAS - MULTI-LINGUAL APPROACHES TO THE AUTOMATIC ANNOTATION OF SPEECH. *The Phonetician* 111(112), 54-69.

[20] Boersma, P., Weenink, D. 2022. *PRAAT: doing phonetics by computer*.

[21] Bertinetto, P., Bertini, C. 2008. On modeling the rhythm of natural languages. *Proc. of the 4th International Conference on Speech Prosody* Campinas, 427-430.

[22] Droua-Hamdani, G., Selouani, S., Alotaibi, Y. 2014. Rhythm analysis in Arabic L2 speech. *Proc. of the 7th International Conference on Speech Prosody* Dublin, 1007-1011.

[23] Mairano, P., Romano, A. 2010. Un confronto tra diverse metriche ritmiche usando Correlatore. *La dimensione temporale del parlato (Proc. of the V National AISV Congress)* University of Zurich, 79-100.

[24] Gu, W., Hirose, K. 2014. Rhythmic Patterns in Native and Nonnative Mandarin Speech. *Proc. of the 7th International Conference on Speech Prosody* Dublin, 592-596.

[25] Yazawa, K., Kondo, M. 2022. A Comparison of Rhythm Metrics for L2 Speech. *Proc. of the 10th International Conference on Speech Prosody* Lisbon, 332-336.

[26] Ordin, M., Polyanskaya, L., Ulbrich, C. 2011. Acquisition of Timing Patterns in Second Language. *Proc. Interspeech* Florence, 1129-1132

[27] White, L., Mattys, S. 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35, 501-522.

[28] Dauer, R. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51-62.

The Phonetics of Prosodic Marking of Focus in Sylheti

Tulika Gogoi, Amalesh Gope

Department of Linguistics and Language Technology, Tezpur University, India

tulikagogoi303@gmail.com, amalesh@tezu.ac.in

ABSTRACT

This paper investigates the prosodic properties of the focus marking system in Sylheti. Sylheti exhibits two-way tonal contrasts. The prosodic means used for focus marking in tone languages include pitch register modification, durational changes, and phrasing. A production experiment was designed to examine the prosodic properties of focus. Results confirm that focus in Sylheti is indicated by pitch register modification of focused words with underlying L and H lexical tones. In-situ focus, the contrastive and corrective focus on the object, are consistently marked by a lowered f0 scaling on the target word. This marks a deviation from the 'effort code,' a phenomenon also observed in Akan. Additionally, significant modifications of pitch register in pre-focal words have also been observed on in-situ focus on the object and contrastive focus. An overall reduced duration on the focused constituents (including the pre-focal and post-focal constituents) is observed compared to their broad-focus counterparts.

Keywords: prosody, in-situ focus, contrastive focus, corrective focus, pitch register.

1. INTRODUCTION

In most languages of the world, intonation serves the function of conveying different post-lexical meanings. Apart from marking sentence types and dividing stretches of speech into smaller units, intonation also serves to encode focus or prominence in languages. However, many works on intonation in tonal languages indicate that tonal languages generally avoid complex intonation systems [1, 2, 3, 4]. It is crucial to examine whether they employ any prosodic means to mark focus to understand the extent to which intonation is used in tonal languages. Some of the prosodic cues that are reported to be employed in tonal languages include pitch register modification [5, 6, 7, 8, 9], lengthening or durational changes [8, 9], alteration of intensity, or post-focal compression [10, 11]. Mandarin uses register expansion for focus marking, leading to higher scaling of H and lower scaling of L tone [9]. Akan employs pitch register lowering as a prosodic means for marking focus [11]. Mandarin also uses post-focal compression (henceforth, PFC) for encoding focus. PFC is also present in Bodo, where it indicates in-situ focus. [7]. This paper addresses how the phonetic properties of focus are prosodically marked in Sylheti.

Some portions of an utterance can be focused on or highlighted to signal newness or contrastivity, which requires the hearer's special attention. A language may employ various phonological, lexical, morphological, and grammatical means for marking focus. The present paper examines the prosodic means used in Sylheti to encode different types of focus on words in the object positions. Analysis of in situ focus on the object (i.e., the focus that does not involve any syntactic movement of the object, often realized in wh- answers), contrastive focus, and corrective focus in Sylheti show that focus marking strategies differ from those primarily seen in other (tonal) languages. It involves lowering the pitch register on the target words in contrast to the general expectation of focus attracting a higher pitch. This can be interpreted as a deviation from one of the biological codes, 'the effort code,' a phenomenon also reported to be employed in Akan [6]. The pre-focal constituents exhibit some modifications in Sylheti, while no consistent changes were observed in the pitch levels of post-focal domains. Pre-focal words in sentences with in-situ focus and corrective focus on the object have lowered pitch, while in contrastive focus, they are realized with a higher pitch than their broad focus counterparts.

2. LANGUAGE UNDER STUDY: SYLHETI

Sylheti is generally considered to be one of the varieties of eastern Bangla [12]. It is typically spoken in the Sylheti district of northeastern Bangladesh. In India, Sylheti speakers inhabit two northeastern states- the Barack Valley region of Assam and northern Tripura (districts such as Kumarghat, Dharamnagar, and Kailashar). The distinct phonetic and phonological properties distinguish Sylheti from Standard colloquial Bangla [12, 13, 14, 15]. One such property is the application

of obstruent weakening, which has significantly reduced and restructured the language's phoneme inventory [14, 15]. Such a reduction has consequently created a two-way tonal contrast, high and low [12, 13, 16]. This reduction can be attributed primarily to the loss of (underlying) breathy voice contrasts among the consonants and partially to the processes of spirantization and deaffrication. In a recent study, Raychoudhury and Mahanta [17] reported that the loss of aspiration in the voiceless segments occurring in (different) onset positions, viz., the first or the second syllable of a disyllabic word, might trigger a three-way lexical contrast in Sylheti. Like most other Indo-Aryan languages, Sylheti follows an SOV word order.

3. EXPERIMENTAL PROCEDURE

Five native speakers of Sylheti from the Dharmanagar district of north Tripura were asked to produce scripted sentences. The recordings were digitized at a sampling frequency of 44.1 kHz and 32-bit resolution. In the experiment, the sequences of target words with underlying H and L lexical tones were embedded in carrier frames. Sentences with in-situ focus on objects were elicited as a response to a wh-question. An example of such a sentence frame is listed below-

Q: Subject QW Verb_J?

In-situ focus: Subject $[Y]_{\text{narrow foc}}\text{Verb}_J$

Sentences with contrastive focus were elicited as embedded sentences, in which the object of the matrix clause was in contrast with that of the embedded clause:

Subject [X] Verb_i BUT Subject $[Y]_{\textbf{contfoc}}\text{Verb}_J$

The corrective focus has been elicited by placing the target words as the answers to yes/no questions with the following sentence frame:

Q: Subject [X] Verb_J?

Corrective Focus: NO, Subject $[Y]_{\textbf{correfoc}}\text{Verb}_J$

Here, Verb_i and Verb_J are replaced with two different verbs. The target word replaces [Y] and is realized with any of three kinds of focus under consideration.

The individual sound files of each sentence are segmented at word leven using PRAAT (version 6.0.43) [18]. The acoustic components, viz., mean f0, duration, and intensity, are considered in this study. The values for all three parameters are extracted using the ProsodyPro [19]. The resulting values are compared between the sentences carrying in-situ focus, contrastive focus, and corrective focus with their broad focus counterparts. Each subject repeated the dataset containing different sentence types six times with a considerable pause between each repetition. The second author supplied questions on in-situ, contrastive, and corrective focus. The best five repetitions are considered for the analysis. A total of 1300 sentence tokens are considered in this study (13 sentences x 4 focus types x 5 repetitions x 5 speakers).

4. RESULTS

4.1. Focus and F0: In-situ focus on object

The results of the production experiment show that Sylheti's in-situ focus is prosodically marked through pitch register lowering. The target words specified with both (underlying) L and H lexical tones surfaced with a lower f0 than their broad focus counterparts. This recurs consistently in all the sequences of lexically H or L tone words that are examined in this study.

Apart from the pitch lowering of the target word, the f0 of the pre-focal words is also observed to be lowered in all the combinations examined in this study. On the other hand, the pitch in the post-focal domain remains neutral.

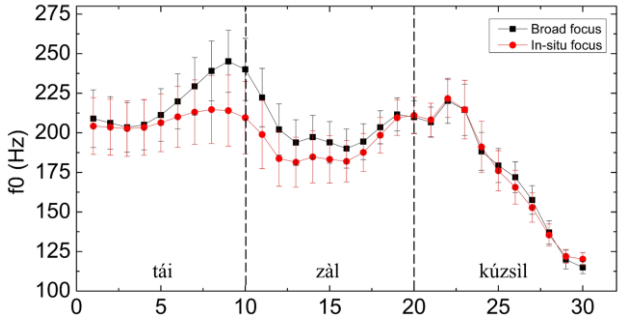


Figure 1: Time-normalized f0 contours averaged across all the tokens produced by all the speakers for the sentence [tái zàl kúzsìl] *she looked for a net*.

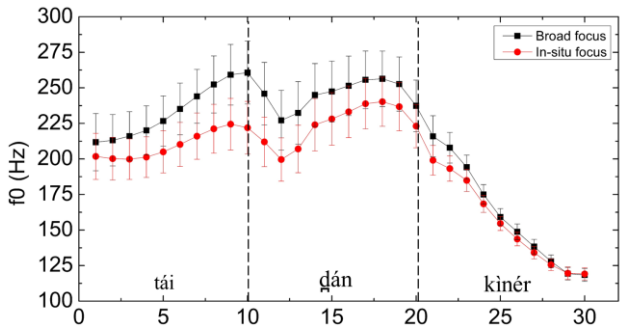


Figure 2: Time-normalized f0 contours averaged across all the tokens produced by all the speakers for the sentence [tái ḍán kìnér] *she is buying rice*.

Figures 1 – 2 show the representation of f0 contours of broad focus (black color) and in-situ focus (red color). The error bars depict the standard error of data post-aggregation. The target words that carry the in-situ focus, viz the objects [zàl] (in Figure 1) and [ḏán] (in Figure 2), are realized with lowered pitch. The pre-focal elements, the subject word [tái] in both cases undergo pitch lowering as well.

4.2. Focus and F0: Contrastive Focus and Corrective Focus

Contrastive focus in Sylheti is marked with pitch register lowering on the target word with underlying L or H lexical tones. While the lexical tonal specifications of target words are retained, the effect of contrastive focus is manifested as changes in f0 scaling on the target words, which is lower than their broad focus counterparts. Interestingly, the words preceding the target words have a higher pitch compared to the corresponding broad-focus sentences (Figure 3). Pitch register compression is observed on the post-focal domains in all the sentences representing different tonal sequences.

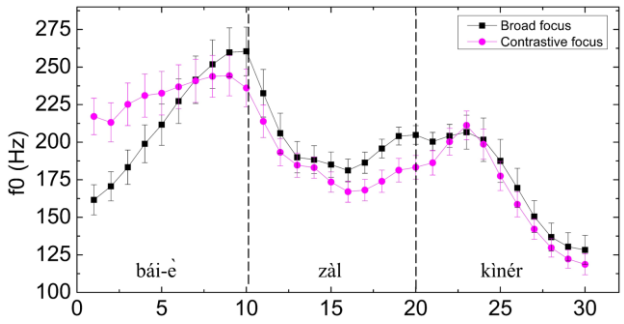


Figure 3: Time-normalized f0 contours representing broad focus (in black) and contrastive focus (in purple); the f0 contours are averaged across all the tokens produced by all the speakers for the sentence [bái-è zàl kìnér] *brother is buying a net*.

Corrective focus, too, is realized through pitch register lowering on the target words. However, pitch levels of both pre-focus and post-focal domains are realized with a lowered f0 compared to their broad-focus counterparts (Figure 4). This holds true for all the combinations of tonal sequences considered in this study.

4.3. Focus and Duration

We compared the duration of focused constituents featuring in the context of in-situ focus, contrastive focus, and corrective focus with their broad focus counterparts.

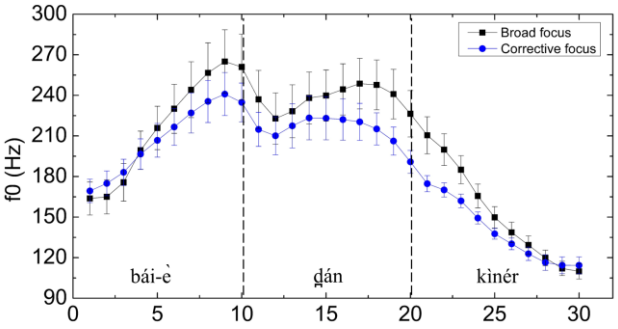


Figure 4: Time-normalized f0 contours representing broad focus (in black) and corrective focus (in blue); the f0 contours are averaged across all the tokens produced by all the speakers for the sentence [bái-è dán kìnér] *brother is buying rice*.

Results indicate that duration is considerably reduced on the pre-focal element compared to their broad focus counterparts (Figures 5 – 7). This pattern is consistently observed in all the sentences featuring different (underlying) tonal sequences (viz., HHH, HLH, LHL, LLL, and so on). A one-way *repeated measure ANOVA* (RM ANOVA) was conducted to examine the effect of the duration on the independent variable (focus types) for the pre-focal word (word occurring on the subject position), the focused word (word occurring on the *object* position), and post-focal word (word occurring on the *verb* position). Results indicate a significant reduction of the pre-focal words occurring in the in-situ context compared to their broad-focus counterparts $F[1, 300] = 11.648$, $p = 0.000731$. However, the durational reduction for the in-situ focus compared to the broad focus observed on the *object* ($F[1, 300] = 2.696$, $p = 0.102$) and the *verb* ($F[1, 300] = 0.449$, $p = 0.503$) is not statistically significant (Figure 5).

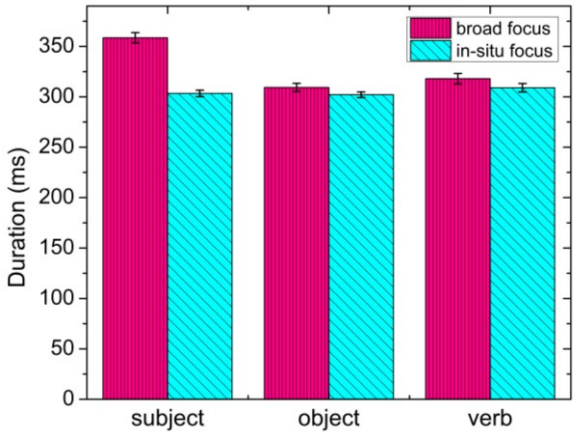


Figure 5: Represents the average duration of the constituent words of the sentence with HHH tonal sequences.

In Sylheti, contrary to the general expectation, the duration of words carrying both contrastive ($F[1, 299] = 9.77$, $p = 0.002$) and corrective focus ($F[1, 299] = 20.693$, $p < 0.0001$) is observed to be significantly lowered compared to their broad focus counterparts. Furthermore, this duration reduction is not limited to the constituent in focus. In sentences where the object receives contrastive focus, both pre-focal ($F[1, 299] = 8.033$ $p = 0.005$) and post-focal ($F[1, 329] = 1.653$, $p = 0.009$) elements are realized with a significantly shorter duration. Durational changes induced by corrective focus are similar to those of contrastive focus. The pre-focus ($F[1, 299] = 12.097$, $p = 0.00058$) and post-focus positions ($F[1, 300] = 11.248$, $p < 0.0001$) undergo an overall reduction when they occur in corrective focus context compared to their broad focus counterparts.

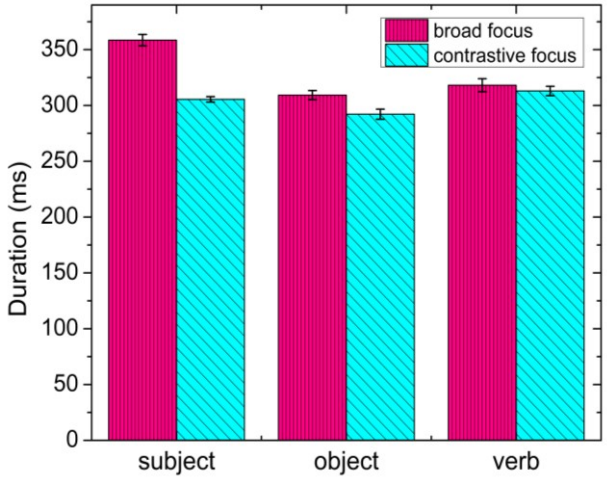


Figure 6: Represents the average duration of the constituent words of the sentence with LHL tonal sequences.

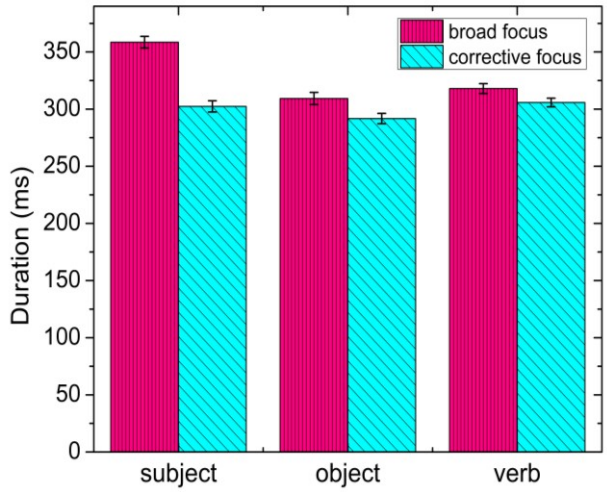


Figure 7: Represents the average duration of the constituent words of the sentence with LLL tonal sequences.

5. DISCUSSION

Results from the experiments conducted for this paper show that Sylheti speakers do employ prosodic means to mark prominence. Pitch register modification is a primary means for prosodic marking of focus in the language. This allows encoding prominence in utterances while preserving the underlying lexical tones of the constituent words. On the other hand, the duration of the pre-focal elements in sentences carrying focused constituents is considerably reduced for both in-situ, corrective, and contrastive focus.

The lowering of the pitch on the target word and reduction of the duration of the target word as well as post and pre-focal elements for encoding focus, suggests a deviation from the expected effects of focus as suggested by the effort code, one of the three biological codes of intonation [8]. The effort code assumes that parts of utterances are made more prominent by increased effort in production by the speaker. This entails that the pitch register on the constituent will be raised when it receives focus. However, Sylheti shows a pattern quite the opposite to it. Instead of a raised pitch register, constituents in focus exhibit a lowering of f0. Moreover, focus-induced lengthening is also not seen in the language. This further exemplifies Kügler & Genzel's [10] observation that the prosodic marking of focus does not involve deviation from the neutral register only in a particular direction.

6. CONCLUSION

Our experimental study has thus identified two primary prosodic markers of focus in Sylheti- pitch register modification and changes in terms of duration. Pitch register lowering on the target words is consistently used to encode focus in the language. Focus-induced pitch modification is limited to the focused word and extends to pre-focal and post-focal domains. Durational changes also serve as cues to focus; the in-situ, contrastive, and corrective focus induce a reduction in duration on the pre-focal element as well as the focus constituents. One of the significant findings of this paper is the observed deviation from the effort code of intonation, which is exhibited through the lowering of f0 and a reduction in duration values of focused constituents. While this study involved the investigation of prosodic properties of focus on object constituents, future research on focus prosody on other locations and morphological focus markers will be required to understand further the scope of prosodic focus marking in the language.

7. REFERENCES

[1] Gussenhoven, C.2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

[2] Yip, M. 2002. *Tone*. Cambridge: CambridgeUniversity Press.

[3] Cruttenden, A.1997. *Intonation*. 2nd edition. Cambridge: Cambridge University Press

[4]Ladd, R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[5] Inkelas, S, Leben, W. 1990. Where phonology and phonetics intersect: the case of Hausa intonation. Papers in Laboratory Phonology I: Between the grammar and the physics of speech:17-34.

[6] Kügler, F, Genzel, S.2012. On the prosodic expression of pragmatic prominence–The case of pitch register lowering in Akan. *Language and Speech 55(3):331-359.*

[7] Mahanta, S. Das, K., Gope, A. 2016. On the Phonetics and Phonology of Focus marking in Boro. Proceedings of the Annual Meetings on Phonology. 3. 10.3765/amp.v3i0.3668

[8] Pan, H. 2007. Focus and Taiwanese unchecked tones. In C. Lee, M. Gordon, & D. Buring (Eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation (pp. 195–213). Dordrecht: Springer.*

[9] Xu, Yi. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of phonetics* 27(1): 55-105.

[10] Downing, L. 2008. Focus and prominence in Chichewa, Chitumbuka, and Durban Zulu. ZAS Papers in Linguistics 49, 47–65.

[11] Kügler, F. 2017. Tone and intonation in Akan. In L. J. Downing, & A. Rialland (Eds) *Intonation in African Tone Languages*(pp 24:89). De Gruyter Mouton

[12] Gope, A. 2016. The Phonetics and Phonology of Sylheti Tonogenesis. Ph.D. Dissertation. Indian Institue of Technology, Guwahati.

[13] Gope, A, and Mahanta, S. 2014. Lexical Tone in Sylheti. Paper presented at 4th International Symposium on Tonal Aspects of Languages, Nijmegen, The Netherlands, May 13–16; pp. 10–14.

[14] Gope, A, Mahanta, S. 2015. An Acoustic Analysis of Sylheti Phonemes. Paper presented at 18th International Congress of Phonetic Sciences, Glasgow, UK, August 10–14.

[15] Gope, A. 2018. The Phoneme Inventory of Sylheti: Acoustic Evidences. Journal of Advanced Linguistic Studies 7: 7–37. Bahri Publications.

[16] Gope, A. 2021. The Phonetics of Tone and Voice Quality Interactions in Sylheti. Languages 6: 154.

[17] Raychoudhury, P, Mahanta S. 2020. The Three Way Tonal System of Sylheti. Paper presented at 10th International Conference on Speech Prosody, Tokyo, Japan, May 25–28; pp. 503–7.

[18] Boersma. P., Weenik D 2012. Praat: doing phonetics by computer (Version 5.3.04_win64) [Computer programme].

[19] Xu, Yi. 2013. ProsodyPro - A tool for large-scale systematic prosody analysis. Proceedings of the TRASP Conference. 7-10.

The phonetics and phonology of corrective focus in Xiangshan Wu Chinese: interaction with disyllabic tone sandhi

Yibing Shi, Brechtje Post, Francis Nolan

University of Cambridge
ys538@cam.ac.uk, bmbp2@cam.ac.uk, fjn1@cam.ac.uk

ABSTRACT

Exploring the interaction between tones and prosodic focus has been an important topic in prosodic studies in tonal languages, as f_0 multi-tasks, both cueing prosodic focus and realizing tonal contrasts. This study provides novel data on the phonetic realisations of corrective focus and its interaction with tone sandhi in Xiangshan dialect, an under-documented Chinese dialect spoken in the Northern Wu region in China. Thirty-eight disyllabic lexical compounds and phrases were selected for this study. Corrective focus was elicited on either the whole disyllable or one of the two syllables. Results show that prosodic focus can be realised via gradient acoustic cues: raised intensity and lengthened duration. Meanwhile, focus can also be optionally realised by phonological re-phrasing as evidenced in tone sandhi changes. The cue-trading relations between tone sandhi changes and gradient acoustic cue changes are also investigated.

Keywords: prosodic focus, tone sandhi, Xiangshan Chinese

1. INTRODUCTION

Prosodic focus, which serves to emphasise a part of the utterance, is typically realised through enhancement of variable gradient acoustic cues on the focused items, including expanded pitch range, lengthened duration, and increased intensity [1],[2]. Alternatively, focus can result in phonological re-phrasing, as evidenced by research across different languages, e.g., Korean [3] and Bengali [4]. In tonal languages, the investigation of focus becomes especially compelling as pitch potentially competes to signal prosodic focus, lexical tonal contrasts, as well as tone sandhi patterns, which are categorical and systematic tonal changes in certain environments.

While rather consistent and similar gradient cue changes have been found in several tonal languages, such as duration lengthening in Shanghai Chinese [5] and Standard Chinese [6], the picture of the interaction between focus and tone sandhi remains largely complex and unclear. Take Shanghai Chinese for example: although many studies witnessed no drastic effects of focus on tone sandhi patterns [7], some accounts reported that focus can block tone sandhi, realising both syllables as their underlying tones [8], [9], or it could occasionally produce other specific tone sandhi patterns, e.g., realising the second syllable in a disyllable as a high falling tone [10].

The current study is the first investigation of the realisations of prosodic focus in Xiangshan Wu Chinese, an under-documented Chinese dialect that is mainly spoken in Xiangshan County, Zhejiang Province in a Southeastern part of China. It has six lexical tones, of which four are non-checked tones (HH, HL, LHL, LH) and the remaining two are checked tones that are only associated with syllables ending with a glottal stop (Hq and LHq) [11].

The following research questions are addressed in the study: (1) To what extent is focus realised in Xiangshan dialect using gradient acoustic cues i.e., intensity and duration? (2) How does focus interact with tone sandhi? Does it block, change, or preserve sandhi patterns? (3) If there is a sandhi change under focus, do other acoustic cues still change, and if so, at a larger or a smaller magnitude?

2. METHODOLOGY

2.1. Participants

Eight Xiangshan native speakers were recruited and recorded by the first author (4 female, age range: 47-53, mean age: 50). All of them were born and raised in Xiangshan and speak Xiangshan as their major language in their daily life. No hearing and speaking impairments were reported among them.

2.2. Materials and procedure

Thirty-eight disyllabic tokens were chosen for this study, with the first syllable bearing the tones LHL or HH and the second syllable bearing any of the four non-checked tones, i.e., HH, HL, LHL, LH. Three morphosyntactic structures were covered in the test materials: lexical compounds, Modifier-Head phrases, and Verb-Object phrases.

The test tokens were presented to the speakers in a Microsoft PowerPoint presentation, with each slide showing only one test token. In the first session of the experiment, to elicit the non-focused citation realizations of each test token, the speakers were asked to read the test tokens one by one as naturally as they could.

Corrective focus was elicited on three different domains in the disyllables: the first syllable, the second syllable, or the whole disyllable. For each test token, a sentence containing a disyllable which differs from the target token in the first syllable, the second syllable, or the whole disyllable, was pre-recorded by a native male Xiangshan speaker aged 53. The pre-recorded audio file was played to the speakers while the target token was shown on a laptop screen. In this way, speakers were prompted to correct the speech using the target tokens after listening to the recordings. The scenario is summarised in Table 1, with the symbol 'σ' representing 'syllable' and the bold and italic letters showing the domain of focus. In total, there were 1,216 tokens elicited (8 speakers * 38 tokens * 4 focus conditions).

Table 1: Corrective focus elicitation scenario

Focus	Recording	Desired output
σ_1	These words are CB.	No, they are ***A***B.
σ_2	These words are AD.	No, they are A***B***.
$\sigma_1\sigma_2$	These words are EF.	No, they are ***AB***.

Prior to the actual recording of the focused items, all speakers completed a practice series with three tokens that were not part of the actual experiment, in order to familiarise themselves with the self-managed recording procedure. The participants were recorded in a small quiet room located in Xiangshan, using an H4n Zoom recorder at a sampling rate of 44.1k Hz.

2.3. Data Analysis

All the acoustic data were manually segmented and annotated using Praat [12]. For each test tokens, the rhymes – which include onset glides, vowels, and final nasals – of both syllables were manually labelled. Duration, mean intensity, and mean f_0 values were extracted using the ProsodyPro Praat script [13]. f_0 values at ten equidistant measurement points in each rhyme were also obtained using the same script. The first normalised time point in each rhyme was discarded to eliminate possible f_0 perturbations.

All f_0 values were log-z-score normalised for each speaker based on the given speaker's mean f_0 value. Duration values were also normalised to neutralise the potential final lengthening effect of the final syllables as opposed to the first syllables. To achieve this, actual duration values were divided by mean duration values in the same position (either the first or the second syllable) within each speaker.

The normalised duration ratio (σ_1/σ_2), mean intensity differences ($\sigma_1 - \sigma_2$), and normalised mean f_0 differences ($\sigma_1 - \sigma_2$) between the two syllable rhymes were calculated. The advantage of using the relative values over the absolute ones is to better capture the relative changes between the two syllables, as prominence is argued to represent a relation between elements in an utterance rather than an absolute threshold [14].

The tone contour patterns of each token across different focus conditions were first compared and labelled manually by a phonetically trained native Xiangshan speaker. Tokens with phonetically and perceptually different tone contours were labelled with different tone sandhi categories according to the actual realisations of tone contours.

Linear mixed-effects models were used to compare the duration ratio and mean intensity differences under different focus conditions, using the *Lme4* package [15] in R [16].

3. RESULTS

3.1. Gradient acoustic cues

Linear mixed-effects models were constructed to analyse the relations between gradient acoustic cues (i.e., mean intensity differences and normalised duration ratio) and focus conditions. FOCUSCONDITION (broad, first, second, disyllable) was taken as the fixed effect, and SPEAKER and ITEM were held as the random effects with random intercepts. All the 1,216 tokens were analysed.

The overall duration ratio (σ_1/σ_2) averaged across all the speakers and morphosyntactic structures in four focus conditions are given in Figure 1. The broad focus condition serves as the baseline for observing the durational effect induced by focus. Linear mixed-effects analysis revealed that corrective focus on the first syllable significantly increases the duration ratio (Estimate = 0.06, t = 2.93, $p < 0.01$**) and that on the second syllable significantly decreases the duration ratio (Estimate = -0.08, t = -3.98, $p < 0.001$***). It indicates that corrective focus lengthens the relative duration of the focused syllable. Corrective focus on the whole disyllable, however, does not affect the duration ratio significantly compared to the broad focus condition (Estimate = -0.03, t = -1.66, p = .10).

The mean intensity difference ($\sigma_1 - \sigma_2$) patterns for all the tokens and speakers are depicted in Figure 2.

Similarly, the relative intensity of the focused syllable is significantly raised when either the first or the second syllable is focused (first: Estimate = 1.02, t = 3.4, p < 0.001***; second: Estimate = -1.15, t = -3.88, p < 0.001***). When the whole disyllable is focused, however, the intensity difference between the two syllables does not differ significantly from the broad focus condition (Estimate = -0.33, t = -1.1, p = .27).

Figure 1 Duration ratio (σ_1/σ_2) by focus condition

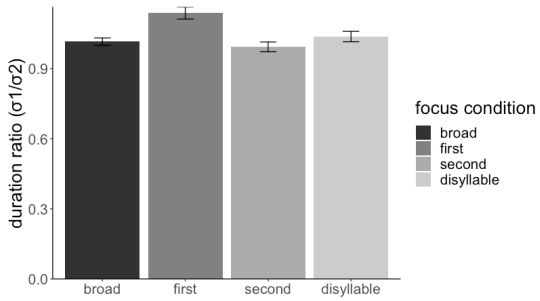
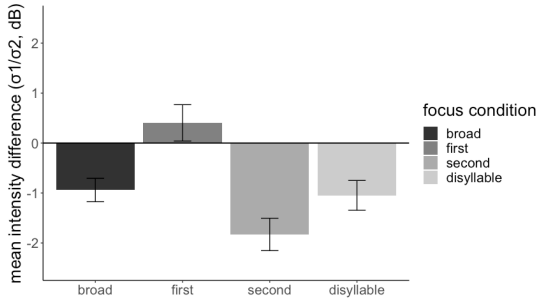


Figure 2 Mean intensity difference (σ_1 - σ_2) by focus condition



3.2. Tone sandhi patterns

Due to the space limit, only the tone sandhi patterns for lexical compounds and Modifier-Head phrases with LHL-σ_2 combinations, as well as their interaction with corrective focus, are reported in this paper. Eight disyllables produced by each speaker were selected, covering all the LHL-σ_2 tone combinations and focus conditions, i.e., 256 tokens in total (8 tokens * 4 focus conditions * 8 speakers).

Figure 3 shows the three sandhi patterns found for those tokens when there is no corrective focus present. Across all speakers and tokens, sandhi1 pattern is the predominant pattern for the LHL-H, LHL-LHL, and LHL-LH tone combinations (78.7% of those combinations), and only appears a few times for LHL-HL. Sandhi2 pattern is found for all except the LHL-HL underlying tone combination, and only 5.7% of all the tokens show this pattern. Sandhi3 pattern is exclusively found in the LHL-HL combination and is the predominant one for this combination (76.5%).

In the current experiment, there are four possible outcomes under focus: broad sandhi on the disyllable (BS, 73%), focus realising syllable 2 sandhi as a rise (σ_2R, 13%), focus realising syllable 2 sandhi as a fall (σ_2F, 9.9%), and the focused syllable carrying its underlying tone (UT, 4.1%). Tokens under the BS type show the sandhi patterns they adopt in the broad focus condition, suggesting that focus has no effect on sandhi patterns for this type. Second syllables of tokens in the σ_2R and σ_2F categories exhibit different sandhi tones when they are under corrective focus, as opposed to when they are under broad focus. The disyllables under the σ_2R type, which show a sandhi2 pattern (L-LML) under broad focus, exhibits a sandhi1 pattern (L-LM) under corrective focus regardless of the focus position. The opposite is observed for σ_2F type, as can be seen in Figure 4.

Figure 3 Tone sandhi patterns (broad sandhi—BS) for LHL-σ_2 disyllables under broad focus

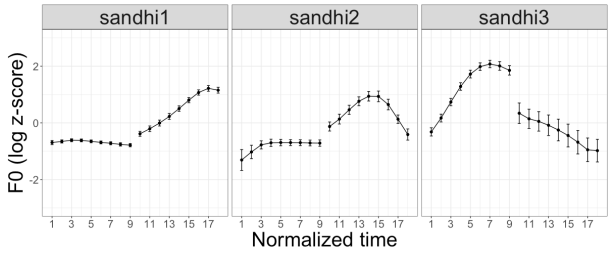
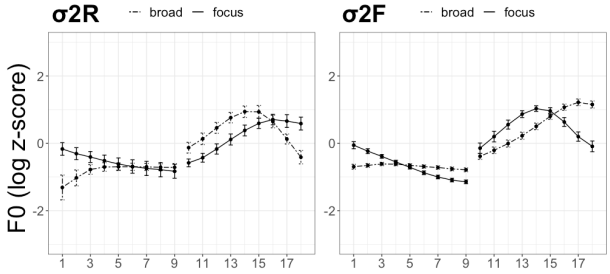


Figure 4 Focus realising σ_2 sandhi as a rise or a fall for LHL-σ_2 disyllables under corrective focus



The fourth type (UT) realises the focused syllable with its underlying tone (Table 2). For example, as shown in the highlighted row in Table 2, a token with underlying LHL-H tones exhibits a sandhi1 pattern (L-LH) under broad focus. However, when the second syllable receives corrective focus, it is realised with its underlying tone H, while the first tone still preserves its broad focus sandhi tone L. This type (with only one exceptional token where both syllables are realised with their underlying tones) is not found when the whole disyllable is focused.

Table 2 Focused syllable carrying its underlying tone under corrective focus for LHL-σ_2 disyllables

Focus	UR tones	Broad sandhi	Focus sandhi
σ_1	LHL-HL	LH-ML	**LHL**-ML
σ_1	LHL-H	L-LH	**LHL**-LH
σ_2	LHL-HL	LH-ML	L-**HL**
σ_2	LHL-H	L-LH	L-**H**

3.3. Tone sandhi changes and gradient acoustic cues

To analyse the cue-trading relations between tone sandhi changes and gradient acoustic cue changes, linear mixed-effects models were constructed with mean intensity difference and normalised duration ratio as the dependent measures, respectively. The fixed factors were FOCUSCONDITION (broad, first, second, disyllable), SANDHITYPE (BS, σ_2R, σ_2R, UT), and their interaction; SPEAKER and ITEM were the random factors with random intercepts. Models with only main effects were compared to the same models with interaction through Likelihood Ratio Tests, where the p-value is derived. Post hoc Tukey HSD tests were conducted when interactions were significant.

The results of mean intensity differences exhibit a significant interaction between FOCUSCONDITION and SANDHITYPE ($\chi2(8) = 21.428$, $p = 0.006$**). Among all the pairwise comparisons, the UT sandhi type displays significantly lower mean intensity differences ($\sigma_1 - \sigma_2$) compared to the other three sandhi types (σ_2R, σ_2F, BS), as shown in the post hoc tests results (σ_2F – UT: Estimate = 6.103, t = 3.233, p = 0.008**; σ_2R – UT: Estimate = 5.359, t = -2.605, p = 0.049*; BS – UT: Estimate = 5.85, t = 3.734, p = 0.002**).

For normalised duration ratio, no statistical significance is found for the main effect of SANDHITYPE ($\chi2(3) = 5.301$, $p = .151$) or the interaction between FOCUSCONDITION and SANDHITYPE ($\chi2(8) = 11.07$, $p = 0.198$).

4. DISCUSSION AND CONCLUSION

The current study has established that corrective focus in Xiangshan Chinese can be realised both phonetically and phonologically. Phonetically speaking, the focused monosyllable within a disyllable domain is generally characterised by a greater relative intensity and lengthened relative duration compared to that under broad focus. This finding aligns with commonly reported prosodic ways of encoding focus in previous research [1], [17], [5], [6]. Corrective focus on the whole disyllable, however, does not induce too much change on the intensity differences and duration ratio between the two syllables.

On the other hand, corrective focus in Xiangshan Chinese can affect phonological tone sandhi patterns. Among disyllabic lexical compounds and Modifier-Head phrases with LHL-initial underlying tone combinations, three divergent tone sandhi change types have been identified under corrective focus: (a) a L-LHL broad sandhi pattern is realised as a L-LH pattern; (b) a L-LH broad sandhi pattern is realised as a L-LHL pattern; (c) the focused syllable is realised as its underlying tone, while the non-focused one maintains its broad sandhi tone. It should be noted that the two sandhi patterns L-LHL and L-LH in types (a) and (b) are not evenly distributed in either broad focus or corrective focus conditions: six out of eight speakers predominantly use the L-LH pattern under broad focus, and switch to the L-LHL pattern under corrective focus; the remaining two speakers show the opposite patterning. It suggests that the two patterns may not be free variants that are randomly produced by the speaker, but are rather associated with the presence or absence of the corrective focus. More data need to be investigated before a reliable account can be made for such sandhi differences.

More interestingly, this study found that focus can potentially change the phonological phrasing, blocking the sandhi for the focused syllable in Xiangshan Chinese. This blocking strategy has been found in other Chinese languages with tone sandhi phenomena, but is usually reported to block sandhi for both syllables [8], [9]. In the present study, Xiangshan Chinese tends, when one syllable is focused, to apply broad sandhi to the underlying disyllable, determining the shape of the unfocused syllable, and then to realise the focused syllable with its underlying tone. As far as the authors know, this has not been found in any other Chinese dialect.

Moreover, this study discovers that the focused second syllable which carries its underlying tone under focus has a significantly higher relative intensity compared to the other three sandhi types. This further confirms an existing correlation between the phonetic and phonological strategies.

To sum up, this study shows that corrective focus in Xiangshan Wu Chinese induces a higher relative intensity and lengthened duration on the focused syllable. Tone sandhi patterns can be realised differently under corrective focus, and in some cases focus potentially blocks the sandhi for the focused syllable while having no effect on the sandhi of the non-focused syllable. There is a potential cue-trading relation between sandhi and gradient acoustic cues.

5. REFERENCES

[1] W. E. Cooper, S. J. Eady, and P. R. Mueller, 'Acoustical aspects of contrastive stress in question–answer contexts', *J. Acoust. Soc. Am.*, vol. 77, no. 6, pp. 2142–2156, Jun. 1985, doi: 10.1121/1.392372.

[2] S. Baumann, M. Grice, and S. Steindamm, 'Prosodic marking of focus domains-categorical or gradient', in *Proceedings of speech prosody*, Dresden Germany, 2006, pp. 301–304.

[3] S.-A. Jun and H.-J. Lee, 'Phonetic and phonological markers of contrastive focus in Korean', in *5th International Conference on Spoken Language Processing (ICSLP 1998)*, ISCA, Nov. 1998, p. paper 1087-0. doi: 10.21437/ICSLP.1998-151.

[4] E. Selkirk, 'Bengali Intonation Revisited: An Optimality Theoretic Analysis in which FOCUS Stress Prominence Drives FOCUS Phrasing', in *Topic and Focus*, C. Lee, M. Gordon, and D. Büring, Eds., in Studies in Linguistics and Philosophy, vol. 82. Dordrecht: Springer Netherlands, 2007, pp. 215–244. doi: 10.1007/978-1-4020-4796-1_12.

[5] Y. Chen, 'Durational adjustment under corrective focus in Standard Chinese', *J. Phon.*, vol. 34, no. 2, pp. 176–201, Apr. 2006, doi: 10.1016/j.wocn.2005.05.002.

[6] Y. Chen and C. Gussenhoven, 'Emphasis and tonal implementation in Standard Chinese', *J. Phon.*, vol. 36, no. 4, pp. 724–746, Oct. 2008, doi: 10.1016/j.wocn.2008.06.003.

[7] B. Ling and J. Liang, 'The nature of left- and right-dominant sandhi in Shanghai Chinese—Evidence from the effects of speech rate and focus conditions', *Lingua*, vol. 218, pp. 38–53, Jan. 2019, doi: 10.1016/j.lingua.2018.02.004.

[8] M. Y. Chen, *Tone Sandhi: Patterns across Chinese Dialects*, 1st ed. Cambridge University Press, 2000. doi: 10.1017/CBO9780511486364.

[9] E. Selkirk and T. Shen, 'Prosodic domains in Shanghai Chinese', in *The Phonology-Syntax Connection*, Chicago: University of Chicago Press, 1990, pp. 313–337.

[10] J. Tian and J. Kuang, 'The phonetic realization of contrastive focus in Shanghainese', in *Speech Prosody 2020*, ISCA, May 2020, pp. 265–269. doi: 10.21437/SpeechProsody.2020-54.

[11] Y. Shi, 'Tones and Disyllabic Tone Sandhi in Xiangshan Chinese Dialect', University of Cambridge, 2020.

[12] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2018.

[13] Y. Xu, 'ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis', in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France, 2013, pp. 7–10.

[14] J. Terken and D. Hermes, 'The Perception of Prosodic Prominence', in *Prosody: Theory and Experiment*, M. Horne, Ed., in Text, Speech and Language Technology, vol. 14. Dordrecht: Springer Netherlands, 2000, pp. 89–127. doi: 10.1007/978-94-015-9413-4_5.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models using lme4'. arXiv, Jun. 23, 2014. Accessed: Jan. 03, 2023. [Online]. Available: http://arxiv.org/abs/1406.5823

[16] R Core Team, 'R: A language and environment for statistical computing'. Vienna, Austria, 2018. [Online]. Available: https://www.r-project.org/

[17] S. Baumann, M. Grice, and S. Steindamm, 'Prosodic Marking of Focus Domains - Categorical or Gradient?'.

Methods for PoLaR Explorations with Machine Learning: Grammatical Analysis of Intonation without Grammatical Labels

Nanette Veilleux, Byron Ahn, Alejna Brugos, Sunwoo Jeong, & Stefanie Shattuck-Hufnagel

Simmons University, Princeton University, Simmons University, Seoul National University, MIT

ABSTRACT

This study provides a proof-of-concept for a new method for analyzing intonational form and meaning, demonstrated by analysis of mirative utterances in American English. Here, K-means clustering using measures derived from PoLaR labels (i.e., TCoG) revealed emergent clusters of pitch accents that are suggestive of familiar phonological categories (e.g., MAE_ToBI H* and L+H*). A Random Forest analysis then classified utterance-level meaning based on measures from both smaller granularity (about clusters and acoustics) was subsequently (related to individual pitch accents) and larger granularity (e.g., global f0 information), showing >85% correct categorization of exclamative vs filler sentences.

This work has implications for how to model mappings between prosody and meaning, especially where existing phonological categories alone don't identify semantic/pragmatic categories.

Keywords: intonation, methodology, phonetics-phonology interface, form-meaning mapping, machine learning

1. INTRODUCTION

Intonation is known to convey a wide range of meanings, but exploring intonational form-meaning mapping has been challenging (e.g., [1] and [2] for some recent overviews, and [3], [4], [5], and [6] for some critical junctures in the development of the theoretical landscape). This challenge stems in part from the persistent indeterminacy regarding the relevant units of analysis, both on the form side (i.e., which phonetic and phonological aspects of tunes signal systematic meaning differences?) and on the meaning side (i.e., what types of meanings are conventionally encoded by tunes?)

On the form side, using phonological categories alone can miss important details, such as gradient variation in f0 slope that may generate incremental shifts in meaning. But unpacking these categories into global acoustic measures and treating them indiscriminately (in, e.g., machine learning), may miss the key generalization that linguistically meaningful prosodic features are often localized.

We address this methodological hurdle, with a case study on the intonation of mirativity in mainstream American English (henceforth MAE). Mirativity can be defined intuitively as an expression of speaker surprise and a perceived violation of speaker expectations (regarding a proposition). Manifesting in exclamatives like (1), it can be marked by certain particles (e.g., '*Wow*!'), or by designated syntactic configurations (e.g., the wh-fronting *without* subj-aux inversion in (1); compare this to (2), a non-exclamative), and most relevant here, intonation.

(1) (Wow!) How believable Theodore is!
[*exclamative, conveying mirativity*]

(2) How is Theodore believable?
[*non-exclamative, not conveying mirativity*]

Regarding the intonational correlates of mirativity, previous work [7] identifies certain (phonologically defined) pitch accents as its primary prosodic cue, but also points out that additional gradient cues may be at play. Building on this, we have developed a method to clarify the aspects of intonation associated with meanings of mirativity, by annotating the corpus data from that paper for some of its acoustic characteristics, and submitting that phonologically-informed acoustic information to machine learning.

More specifically, we use PoLaR ([8], [9]) to identify relevant acoustic cues in phonologically-defined regions (e.g., pitch accents), and submit the resulting labels and related measures (e.g., tonal center of gravity; TCoG [10]) to k-means clustering, thereby bundling accent-related measures in a form that can be converted to utterance-level information (in the form of, e.g., each accent cluster's rate of occurrence).

Utterance level features (including information about labels, clusters and acoustics) were subsequently submitted to a Random Forest. The results produced over 85% correct categorization of a balanced sample of over 250 exclamative vs. filler sentences. This approach has more general implications for analyzing intonational meaning, and establishes a method that is extendable to other prosody-meaning mappings where existing phonological categories alone do not distinguish semantic/pragmatic categories.

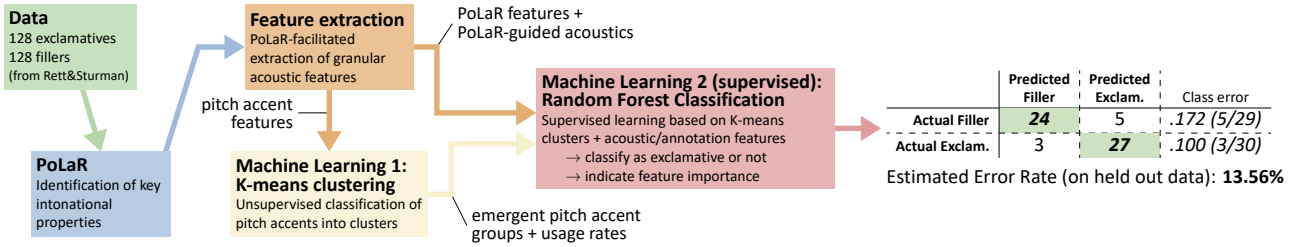


Figure 1: Overall methodology and final results of analysis

2. METHODS

The dataset used in this study is a corpus of 256 utterances collected by [7], in which two MAE speakers read scripts, half of which were exclamatives (e.g., '*Wow, is that nice*!') and the other half not (e.g., '*Is that nice?*'), occurring in four different syntactic frames (declarative, subject-aux inversion, fronted WH-phrase, definite nominal). This dataset was analyzed following the flowchart in Fig.1. The recordings were first force-aligned ([11]) and then PoLaR-labelled in Praat ([12]), as illustrated in the lower half of Fig.2.

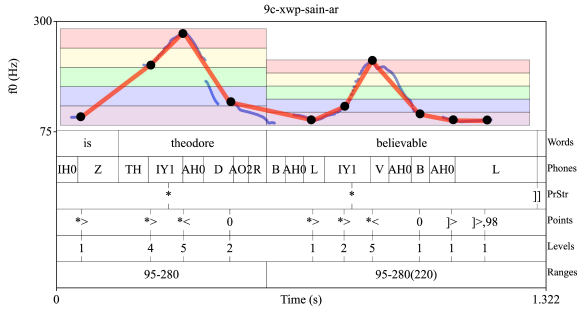


Figure 2: An annotated recording from the corpus

The f0 visualization in the top half of Fig.2 has been marked up to show some key aspects of PoLaR labelling, described more fully in [8, 9]. The dots are at coordinates of (time, f0), where the time value depends on the timing of labels on the Points tier, and the f0 coordinate is either taken from special Points tier labels ("comma override values") or (as is default) from the f0 value calculated by Praat (shown in blue). Interpolating between these dots creates a straight line approximation of the f0 (shown as a red line). PoLaR labels also include a Ranges tier, which defines the *local* f0 floor/ceiling, which can change over an utterance (or even within a phrase). Each range defined by the floor/ceiling is divided into evenly-spaced quintiles (shown as colored bands); the quintile in which the (time, f0) coordinates of a Points label occur is translated into a numerical value (1-5), which is transcribed on the Levels tier. Levels labels thus encode scaled pitch that is normalized relative to the local intonational context. In addition to these three tiers annotating f0 properties, the PrStr (Prosodic Structure) tier contains minimal phonological labels that indicate perceived prominences (*) and boundaries (]).

The annotation process involved three pairs of labellers. The two members of a pair each labeled alone, then compared their labels, and discussed disagreements to generate consensus labels. The first pair labeled according to the basic PoLaR annotation guidelines [9], and the second two pairs used the advanced annotation guidelines, which (among other things) augment basic Points labels to indicate whether Points-defined f0 movements are related to prominences and/or boundaries as labeled on the PrStr tier.

For each of these labeled recordings, a variety of features were subsequently extracted for use in analysis. Certain f0 attributes (such as maximum, minimum, average) were calculated both in raw values and z-score normalized by speaker. PoLaR-labelled TextGrids facilitated extraction and calculation of additional features, including (1) Features based on PoLaR labels themselves (e.g., counts of phrasal prominences and boundaries), (2) direct measures such as timing and (normalized) pitch values, and (3) derived measures such as slope between certain PrStr-associated Points labels and Tonal Center of Gravity ([10]) relative to local f0 ranges (i.e. PoLaR Ranges labels).

Machine Learning 1: pitch accent type clusters

The machine learning modeling of this data takes place in two sequential stages: unsupervised clustering of pitch accents, and supervised random forest categorization of utterances's mirativity. The reason for this two-stage approach is two fold. First, even though ToBI-type pitch accents can signal semantic/pragmatic differences (e.g., [4]), we suspect that they may be too broad for capturing all the relevant distinctions; so, we used unsupervised learning to capture distinctions without bundling them into these umbrella phonological categories. The second issue is at the heart of the difficulties of using machine learning in prosodic form-meaning mapping: prosodic events occur at a more local level (e.g., a syllable or word) than meaning events (e.g.,

an utterance). This work is an example of classifying each utterance as one of two categories (exclamative vs filler) and using the characteristics of the (usually multiple) pitch accents in each utterance. If the utterances in each category had been more parallel, then the unique pitch accent on the target words could have been used. Here, we use the percentage of pitch accents belonging to each stage-one cluster in the final random forest categorization model.

An unsupervised k-means clustering algorithm ([13], [14]) was used to model the pitch accent types, resulting in 3 clusters. Though this was a linguistically informed question (based on [7]), the k-means algorithm automatically determines the number of clusters and the feature values associated with each cluster. After systematic exploration of various combinations of intonational features (mentioned above) to the clustering algorithm, we choose the feature set that produced clusters with the best Sum of Squares characteristics. As a result, the two pitch accent measures used in the clustering algorithm were Tonal Center of Gravity (TCoG) measures, which have also been previously shown to differentiate pitch accents ([10]). Specifically, TCoG was measured over the pitch-accent's rise, and we submitted two relativized (and subsequently z-scored) values: the time of the TCoG relative to the vowel center (tcogT), and the frequency of the TCoG relative to the Range min/max (tcogF). The resulting three clusters are shown as different shapes/colors in Fig.3, with each cluster's centroid annotated in black.

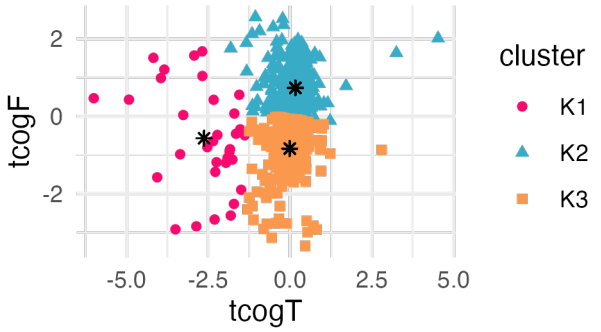


Figure 3: Results of K-Means clustering

These three clusters are suggestive of MAE_ToBI labels: K1 as H* with a preceding high target, K3 as H* without a preceding high target, and K2 as L+H*.

Machine Learning 2: classification as exclamative vs. filler

The ultimate goal of the machine learning model is to determine if exclamatives can be categorized separately from the filler sentences and what features contribute to this separation. In this stage, a supervised random forest model ([13], [15]) was used to classify exclamatives and filler sentences. The particle "wow" was excised from exclamative input utterances. In addition to the rate at which each pitch accent cluster occurred in an utterance, other acoustic and semantic features served as input to this classification, listed below. (Data analysis materials can be found at [16].)

Direct acoustic measures and derived features

- Changes in f0 (max, min, average, delta): raw and z-score normalized by speaker
- Tonal Centers of Gravity (time, frequency)

PoLaR label features

- Measures of f0 for the utterance: timing and (normalized) pitch values for turning points in f0, (local) f0 Ranges, location of prominences and boundaries
- Counts of prominences and phrase boundaries, raw and as a ratio of number of content words

Semantic features:

- Semantic type (content vs function word) for the word containing the maximum f0

3. RESULTS

A random forest was trained on a random sample of 77% of the data (197 utterances: 99 fillers and 98 exclamatives from a set of 256 equally distributed utterances.). When this model was tested on the remaining 23% of the data (59 utterances, 29 filler, 30 Excl.), the resulting classification has a 86% accuracy rate [95% CI : (0.7502, 0.9396)]. The confusion matrix shown at the end of the flowchart in Figure 1 further details this model's output. In examining the mis-classification by class, fillers tend to be misclassified as exclamatives more often. However, given the limitations of the data set size, these values are sensitive to the random selection of the test set.

In addition, Random Forest models allow the predictors to be ranked according to importance (which tends not to vary significantly between the random selections of training/test sets), as shown in Fig.4 for the present model.

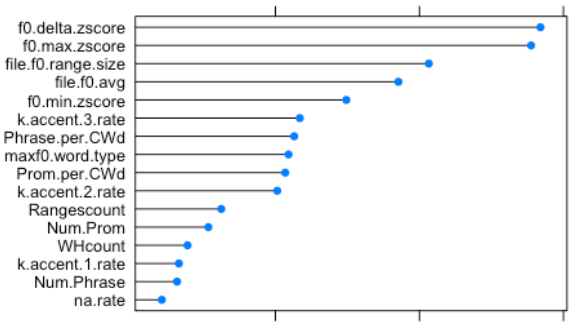


Figure 4: The relative importance of factors in the Random Forest classification of exclamatives or not

Each utterances' z-scored f0 max-min and f0 max were ranked highest. Accent types still play a role despite the use of related acoustic measures.

4. DISCUSSION

First, we discuss the broad findings of this study. This work reaffirms the idea that semantic mirativity is marked intonationally. Moreover, the specific findings are strikingly similar to [7]'s, suggesting that their analysis does not depend on the use of MAE_ToBI labels or adoption of its grammar. While the rate of occurrence of a particular pitch accent type (here: k-means cluster) is not as important as other acoustic characteristics for classification in this model (see Fig.4), the features that do rank high in importance (e.g., f0.delta.zscore) may be the very cues that identify L+H* apart from other accents — again consistent with [7]'s analysis. On the other hand, the meaning of mirativity may arise *not just* from (cues to) L+H* alone, but rather from some constellation of intonational characteristics, including categories of intonational phonology (such as L+H*) and other gradient components of intonation.

Methodologically, the main contribution here is a proof-of-concept machine learning analysis that is on par with established qualitative methods for intonational meaning, since our results are broadly consistent with the findings in [7]. Moreover, unlike a model that uses only global acoustics, the use of PoLaR labels has the advantage of enabling the targeting of acoustics from smaller, *phonologically-relevant* domains, thereby allowing for a clearer characterization of the intonational form of MAE exclamatives.

Combining PoLaR annotation with ML techniques opens many new avenues for pursuing form-meaning mapping research, in areas where it's not (yet) clear what the categories of form are. This is facilitated by PoLaR's capacity to capture linguistically-informed acoustic measures without presupposing a particular set of phonological categories. In fact, this methodology can help identify intonational categories by revealing which aspects of form map onto particular meanings. In addition, for cases where there are no discrete phonological categories (i.e., if the meanings and forms are not grammatically structured; e.g., for so-called "paralinguistic" uses of intonation), PoLaR labels can potentially identify which dimensions of acoustic form are relevant for signalling particular meanings.

Beyond using ML techniques alongside PoLaR, using PoLaR for intonational research is itself advantageous, as it does not require the same extensive training or experience that other labelling systems might. Instead, labelers are able to identify appropriate regions for collecting the salient acoustic measurements that feed into the statistical and machine learning analyses. Utterance-level acoustic measures (e.g., f0 min/max/average) are insufficiently targeted, as the most critical acoustic values are often localized in specific phonologically-relevant regions. On the other hand, strictly phonological annotation systems run the risk of ignoring key patterns in the acoustics that may be of interest in conveying meaning (for discussion on this point, see [8]). In contrast, PoLaR identifies phonologically-informed acoustic measures which can be input into ML models of intonational form-meaning relationships, to discover potential meaning-bearing aspects of the f0 signal that are not directly related to phonological contrasts.

Turning now to an analysis of how to formally model intonational meaning, these results are consistent with a model in which L+H* is the marker of mirativity. Despite this, caution should be exercised in modelling this relationship with a conventional and categorical one-to-one mapping between L+H* form and mirative meaning. Instead, we speculate that it may be more fitting to propose a many-to-many mapping between form and meaning, with multiple intonational cues (related to or partly consisting of L+H*) marking multiple possible interpretations. This is especially plausible since L+H* has been argued to mark other meanings, such as contrastive focus. Additionally, determining whether these ML algorithms reflect the intonational factors that matter for interpretation of utterances by human listeners, will require extensive study of human intonation perception.

5. CONCLUSIONS

The key contribution of this paper is its demonstration of a new methodology for exploring intonational form and meaning. We are confident in the applicability/usefulness of this methodology, but due to practical limitations such as the size of the corpus, we do not yet draw strong conclusions about the phonology of exclamatives in MAE and what in the semantics maps onto the relevant phonology. Coupling this methodology with a deeper investigation into the production and (human) perception of these intonational variables (e.g., with minimal pairs that differ in terms of the intonational dimensions identified here) may be able to illuminate form-meaning relationships *as well as* the phonetics-phonology interface.

6. REFERENCES

[1] H. Truckenbrodt, "Semantics of intonation," in *Semantics: An International Handbook of Natural Language Meaning*, K. V. Heusinger, C. Maienborn, and P. Portner, Eds. De Gruyter, 2012, vol. 3, pp. 2039–2069.

[2] D. Büring, *Intonation and Meaning*. Oxford: Oxford University Press, 2016.

[3] G. Ward and J. Hirschberg, "Implicating uncertainty: The pragmatics of fall-rise intonation," *Language*, vol. 61, pp. 747–776, 1985.

[4] J. Pierrehumbert and J. B. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," pp. 271–311, 1990.

[5] M. Steedman, "Information structure and syntax-phonology interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, 2000.

[6] N. Constant, "English rise-fall-rise: A study in the semantics and pragmatics of intonation," *Linguistics and Philosophy*, vol. 35, no. 5, pp. 407–442, 2012.

[7] J. Rett and B. Sturman, "Prosodically marked mirativity," in *Proceedings of the 37th West Coast Conference on Formal Linguistics*, D. Reisinger and M. Huijsmans, Eds. Somerville, MA: Cascadilla Proceedings Project, 2021, pp. 1–20.

[8] B. Ahn, N. Veilleux, and S. Shattuck-Hufnagel, "Annotating prosody with PoLaR: Conventions for a decompositional annotation system," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 1302–1306.

[9] B. Ahn, N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, "PoLaR annotation guidelines (version 1.0)," Available at https://osf.io/usbx5/, 2021.

[10] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology," *Laboratory Phonology*, vol. 3, no. 2, pp. 337–383, 2012.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner [Computer program] (version 1.0.1)," Retrieved from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/, Apr. 2019.

[12] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]," Version 6.2.09, retrieved from http://www.praat.org/, Feb. 2022.

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022, iSBN 3-900051-07-0.

[14] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R package for determining the relevant number of clusters in a data set," *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014.

[15] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[16] B. Ahn, A. Brugos, S. Jeong, S. Shattuck-Hufnagel, and N. Veilleux, "A Case Study for PoLaR Explorations with Machine Learning," Available at https://osf.io/h8n6f/, 2023.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, and 2042748. We would like to thank Jessica Rett and Beth Sturman for consulting and sharing their corpus. We are also indebted to the undergraduate members of our research team: Ismah Ahmed, Samin Charepoo, Julia Hartnett, Michael Malvone, Tal Schaeffer, Sreeniketh Vogoti, and Megan Willis, who did the bulk of the PoLaR labelling, as well as Shirley Fong, Madeline Guettler, and Sofia Hirschman, who were invaluable in the development of the machine learning models.

THE KINEMATIC PROPERTIES OF PROSODIC BOUNDARIES IN CONVERSATIONAL TURN-TAKING

Ruaridh Purse[1] and Jelena Krivokapić[1, 2]

[1]University of Michigan, [2]Haskins Laboratories
rupurse@umich.edu, jelenak@umich.edu

ABSTRACT

Studies on the kinematic properties of read speech have established an effect of lengthening on gestures at prosodic boundaries, which increases in strength according to the hierarchical position of the boundary. However, the effect of turn-position—whether a boundary occurs in the middle or at the end of a conversational turn—on these properties is underexplored. In an electromagnetic articulography semi-spontaneous speech study, pairs of participants (with one speaker in the magnetometer per dyad) asked each other questions to collaboratively solve a puzzle. We compare targetwords in phrase-medial position, phrase-final turn-medial position, and phrase-final turn-final position, in both questions and answers, exploring two competing hypotheses: (1) turn-final boundaries exhibit a stronger lengthening effect owing to a higher hierarchical position, and (2) lengthening is reduced in turn-final position since there is no imminent speech to plan, thus no additional lengthening is required to accommodate that planning time.

Keywords: Turn-taking, final lengthening, prosody, kinematics, planning

1. INTRODUCTION

Articulatory gestures at prosodic boundaries are longer than equivalent phrase-medial gestures [1,2,3,4]. This lengthening effect is larger at hierarchically higher prosodic boundaries (those delimiting higher prosodic units) and smaller and hierarchically lower boundaries [1,3,4]. Research on the scope of this lengthening effect finds that it extends from the phrase edge. Gestures and acoustic segments closest to the edge (e.g., the last gesture of a phrase-final word) are most strongly and reliably affected. Gestures and acoustic segments further away show less or no lengthening relative to phrase-medial gestures [5,6,7,8,9,10].

Studies describing the kinematic properties of prosodic boundaries in this way, robust as they are, have been almost exclusively limited to read speech monologues. While we do not expect read speech and spontaneous speech to be fundamentally different, some kinematic properties may differ depending on context: Specifically, final lengthening may differ depending on whether speakers need to plan an upcoming utterance. Indeed, longer or structurally more complex utterances are preceded by longer pauses than shorter and/or less complex utterances [11,12,13,14]. This has been related to the speakers' need to plan the upcoming utterance, with longer pauses allowing the speaker more time to plan upcoming speech. Since pauses are part of prosodic boundaries, and speakers plan speech continuously, it is likely that speech planning will also affect final lengthening.

The present study examines final lengthening in turn-taking. Despite the broad interest in the timing properties of turn-taking in the field of Conversation Analysis [15,15], the temporal properties of boundaries at turn-ends have not been examined extensively [cf. 17] and the fine-grained kinematic properties of prosodic boundaries at turn-ends remain unexplored. We propose two competing hypotheses about how phrase-final lengthening at turn-final prosodic boundaries compares to final lengthening at turn-medial boundaries if, indeed, these conditions are found to differ at all.

(1) Turn-final prosodic boundaries are hierarchically higher than turn-medial and therefore induce more lengthening in nearby gestures.

(2) Planning loads are lighter turn-finally than turn-medially and therefore turn-final gestures exhibit less planning-induced lengthening compared to phrase-final turn-medial gestures.

We also explore the turn-medial lengthening effect by comparing phrase-final turn-medial gestures to phrase-medial gestures (preceding a word boundary).

2. METHODS

2.1. Procedure

2.1.1. Participants

Six pairs of participants, native speakers of American English, participated in the study. They were naïve to the purposes of the study. In each dyad, synchronized acoustic and kinematic recordings were taken from one participant (participant A) using a Carstens AG501 Articulograph. The other participant (participant B) acted as an interlocutor.

2.1.2. Stimuli and experimental design

There are three conditions (phrase-medial, phrase-final turn-medial, phrase-final turn final) and one targetword, *mima* (pronounced ['mimə]).

The targetwords were elicited using a modified game of tangrams, with each participant solving up to 10 tangrams. For each trial, both participants were shown the same two tangram puzzles, with the outlines of their component shapes visible. For one of these tangrams, its component shapes were colored and labelled; labels for triangles contained the targetword *mima* and labels for quadrangles contained the targetword *nana*. The shapes making up the other visible tangram were left blank. Each targetword was embedded in a phrase consisting of the definite article, a color word, and a targetword, for example "the Lava Nana", “the Big Mocha Mima”). Participants were instructed how to produce the targetwords (['mimə]) and ['nɑnə]). To ensure that the targetwords do not have a pitch accent, they were also instructed to place emphasis on the color words.

The labelled tangram seen by participant A matched the blank tangram seen by participant B and vice versa, such that each participant had access to information about the location of shapes in their interlocutor’s uncompleted tangram (Fig. 1). Each participant was prompted to find the location of the labelled shapes in their blank tangram by asking their interlocutor questions.

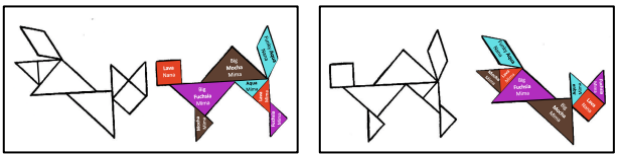


Figure 1: Example tangrams seen by participant A (left) and participant B (right) for a given trial.

Participants took turns asking and answering questions about the position of one named shape relative to another. On each turn, a participant could choose to ask one question or two questions, ensuring that sentences occur in both turn-medial and turn-final condition. To ensure that the targetwords occurred in all boundary conditions (phrase-medial, phrase-final turn-medial, phrase-final turn final), participants were asked to produce the questions and statements in a specific format. For example, for the two-question sequence, the template was: "Is the Aqua Mima to the right of the Big Mocha Mima? And is the Lava Nana right above the Fuchsia Mima?" (with the first *mima* being phrase-medial, the second *mima* being phrase-final turn-medial, and the third *mima* being phrase-final turn-medial). The answers had the parallel format (Table 1).

One question	Is the **Aqua** Mima$_1$ to the right of the Big **Mocha** Mima$_3$?
Two questions	Is the **Aqua** Mima$_1$ to the right of the Big **Mocha** Mima$_2$? And is the **Lava** Nana above the **Fuchsia** Mima$_3$?
One-sentence answer	Yes, the **Aqua** Mima$_1$ is to the right of the Big **Mocha** Mima$_3$.
Two-sentence answer	Yes, the **Aqua** Mima$_1$ is to the right of the Big **Mocha** Mima$_2$. But no, the **Lava** Nana is not above the **Fuchsia** Mima$_3$.

Table 1: Template for sentences. Bolding indicates targeted pitch accents. The indices at the targetwords indicate the boundary condition: targetword$_1$ indicates phrase-medial, targetword$_2$ indicates phrase-final turn-medial, and targetword$_3$ indicates phrase-final turn-final.

2.2. Analysis

2.1.1. Data labelling

Targetword tokens were labelled for their phrase position, turn position, and whether they were used as part of a question or an answer. Targetwords in the middle of a question or answer were labelled “phrase-medial” and act as a control condition. Targetwords at the end of a whole question or answer turn were labelled both “phrase-final” and “turn-final”, and targetwords at the end of a question or answer in a two-question or two-answer turn were labelled both “phrase-final” and “turn-medial”.

Kinematic data were semi-automatically labelled in MATLAB using the *findgest* function in Mview (custom software written by Mark Tiede at Haskins Laboratories, New Haven, CT). Using velocity criteria, we label gesture onset, peak velocity of the closing movement, nucleus onset, maximum constriction (velocity minimum), nucleus offset, peak velocity of the constriction opening movement, and gesture offset. For the targetword *mima*, the lip aperture trajectory was used, for *nana*, the vertical tongue tip movement was used. The gestures for both consonants were labelled (Fig. 2).

From these measures we calculate the following variables for each consonant gesture:

- Closing movement: from gesture onset to nucleus offset
- Opening movement: from nucleus offset to gesture offset

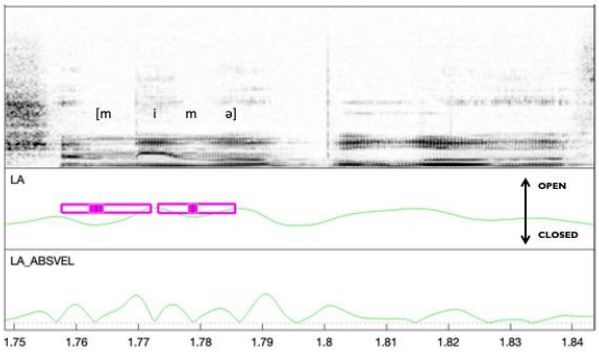


Figure 2: Sample token with the word *mima*, in the context of "Is the big fuchsia mima below…" The rectangles show labelled vocal tract gestures. The whole rectangle is the gesture, the filled part the gesture nucleus, the dashed line the time of maximum constriction. LA: the lip aperture trajectory (for the consonant [m]).

To ascertain that targetwords were produced in the intended boundary position, pauses after targetwords were also labelled according to acoustic cues. Phrase-medial targetwords followed by a pause longer than 100ms were excluded from the analysis [19] as these could not be ascertained to not be phrase-final targetwords. Targetwords preceded by pauses were also excluded from the analysis, as these might be phrase-initial, and therefore subject to phrase-initial lengthening. Targetwords produced with disfluencies or a pitch accent (as noted by a trained ToBI labeller) were also excluded.

2.1.2. Statistical analysis

Analysis is limited to instances of the targetword *mima* in the four speakers analysed to date, owing to the scarcity of instances of the targetowrd *nana* following exclusions.

	P. Medial		T. Medial		T. Final	
	Q.	A.	Q.	A.	Q.	A.
M1A	22	40	10	16	19	23
F3A	67	53	11	13	37	23
F4A	35	62	16	21	24	29
F5A	25	45	11	19	15	15

Table 2: Per-speaker token counts by phrase/turn position and sentence type (question/answer).

Separate mixed-effects linear regression models were built predicting variance in duration for each of the four extracted movements: C1 closing, C1 opening, C2 closing, and C2 opening. Each model contained an interaction between phrase position and sentence type (sum-coded) and a random intercept for speaker. Additional linear regressions were used to model individual speaker behavior.

3. RESULTS

Aggregating across speakers and sentence types (Fig. 3), the closing movements are longer on average in phrase-final turn-medial position than in phrase-medial position for both C1 [7.1ms, p<.05] and C2 [3.9ms, p<.05]. However, C2 opening movements—which lie closest to the prosodic boundary—are not significantly longer in phrase-final turn-medial position than in phrase-medial position. This pattern also holds for two out of four speakers individually (M1A, F4A), while the other two speakers do not exhibit significant final lengthening on any measures.

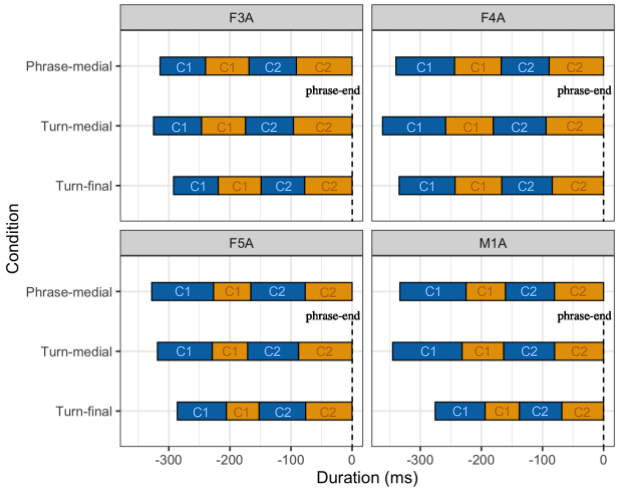


Figure 3: Average duration of C1 and C2 closing (blue) and opening (orange) movements.

Contrary to our first hypothesis, we also do not find more pronounced lengthening in turn-final position compared to phrase-final turn-medial position. In fact, in aggregated data, phrase-final turn-medial movements are longer compared to turn-final movements across the board: for C1 closing [14.7ms, p<.001], C1 opening [5ms, p<.001], C2 closing [8ms, p<.001] and C2 opening [10.6ms, p<.001]. Every individual speaker also produced at least two of these four movements significantly shorter [p<.05] in turn-final position, with C2 closing and opening at least trending towards significantly shorter turn-finally for all four individuals [p<.08].

3.1 Kinematic properties of different sentence types

Comparing the kinematic properties of prosodic boundaries in questions and answers separately (Fig. 4), we find significant duration differences in all measures between sentence types for phrase-medial targetwords in the aggregate data. Responses show shorter C1 closing [5.4ms, p<.01], C1 opening [4.7ms p<.001], and C2 closing [4.2ms p<.001]. However, phrase-medial targetwords in responses also show a longer C2 opening [6.3ms p<.001] than in questions. The only interaction effect that emerges between sentence type and phrase/turn position shows that there is no such effect of longer C2 opening in questions in turn-final position.

Looking at individuals, it emerges that one speaker (F5A) *does* produce a C2 opening that is significantly longer in phrase-final, turn-medial position than phrase-medial position, even though this effect is not evident in the aggregate data, but only in questions [11.5ms, p<.001].

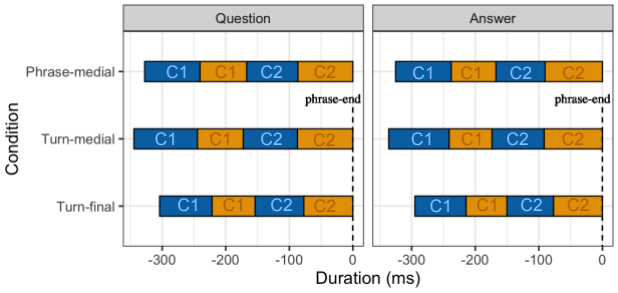


Figure 4: Average aggregate duration of C1 and C2 closing (blue) and opening (orange) movements.

The effect whereby turn-final gestures are shorter than turn-medial phrase-final gestures holds significantly for most measurements across all individuals in both sentence types.

4. DISCUSSION

4.1. Reduced planning load turn-finally

Across the board, turn-final gestures are reliably shorter than phrase-final turn-medial gestures. Thus our hypothesis that a greater lengthening effect may be found here due to a hierarchically stronger prosodic boundary was not borne out. This finding also contrasts with previous studies associating turn-ends with a lengthening effect [15, 19].

The results support our second hypothesis that the effect of final lengthening would be reduced in turn-final position, relative to turn-medial prosodic boundaries, because of a difference in planning load. In turn-medial positions, additional lengthening may occur in order to afford a speaker more time to plan upcoming speech in the same turn. In turn-final position, there is no need for this additional lengthening since there is no more speech left to plan in that turn.

4.2. Simultaneity of lengthening effects

The effect of planning may also explain the pattern of results found for the comparison between phrase-medial and phrase-final turn-medial gestures. That is, the highest planning load may be found in phrase-medial positions, which are relatively early in a sentence, as speakers continuously formulate their questions and answers as they produce them. Thus, the need for more planning time might lead to lengthening. Even though phrase-medial tokens with following pauses greater than 100ms were excluded, their original presence in the dataset supports that planning loads were high in phrase-medial position [see also 20,21]. As such, the effects of structural prosodic lengthening may be obscured by planning-induced lengthening in phrase-medial position, which reduces the difference in gesture duration that might otherwise be found between phrase-medial and phrase-final movement.

The notion that planning-induced lengthening may obscure structural lengthening is also relevant for the turn-final position. We cannot rule out that an underlying effect of structural prosodic lengthening is present in turn-final position, only that it is not large enough to be evident under the current conditions. Put differently, the sources of increased or reduced effects of final lengthening in turn-final position that we illustrated in our original two hypotheses are not necessarily mutually exclusive. It may be true that turn-final phrase boundaries, since they are hierarchically higher, induce a stronger effect of structural lengthening. However, this effect will only be evident when the difference in speech planning load between turn-medial and turn-final position is not so large as to induce an even greater reduction in planning-induced lengthening turn-finally.

5. CONCLUSIONS

Gestures at turn-final prosodic boundaries are produced with shorter durations than gestures before turn-medial prosodic boundaries. This phenomenon is explained by the fact that a turn-final position is associated with a lighter planning load than turn-medial positions.

Heavy speech planning loads in phrase-medial position may also obscure phrase-final turn-medial lengthening relative to these tokens. Properly accounting for the role of performance factors like planning is crucial as we attempt to understand kinematic properties and the representation of prosodic boundaries.

6. ACKNOWLEDGMENTS

Thank you to Wyatt Barnes, Ana-Patricia Collazo, Benjamin Hermen Paige Marshner, Lucas White and Brennan Yaghmour for data labelling, and to attendees of ASA183, and the members of the University of Michigan Phondi and Prosody groups for their feedback. The study is supported by the Associate Professor Fund of the University of Michigan.

7. REFERENCES

[1] Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. JPhon, 26(2), 173-199.

[2] Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. JPhon, 29 (2), 155–190.

[3] Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. JASA, 101(6), 3728-3740.

[4] Krivokapic, J., & Byrd, D. (2012). Prosodic boundary strength: An articulatory and perceptual study. JPhon, 40 (3), 430–442.

[5] Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. JPhon, 31(2), 149-180.

[6] Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects

[7] Oller, K. D. (1973). The effect of position in utterance on speech segment duration in English. JASA 54, 1235—1247.

[8] Berkovits, R. (1993). Progressive utterance-final lengthening in syllables with final fricatives. Lang. Speech 26, 89—98.

[9] Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. Journal of Phonetics, 44, 62-82.

[10] Katsika, A. (2016). The role of prominence in determining the scope of boundary-related lengthening in Greek. Journal of phonetics, 55, 149-181.

[11] Ferreira, F. (1991). “Effects of length and syntactic complexity on initiation times for prepared utterances,” J. Mem. Lang. 30, 210–233.

[12] Ferreira, F., and Swets, B. (2002). “How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums,” J. Mem. Lang. 46, 57–84.

[13] Krivokapic´, J. (2012). “Prosodic planning in speech production,” Speech planning dynamics in Speech planning and dynamics (Peter Lang, Frankfurt, Berlin, Bern, Bruxelles, New York, Oxford, Vienna), pp. 157–190.

[14] Watson, D., and Gibson, E. (2004). “The relationship between intonational phrasing and syntactic structure in language production,” Lang. Cogn. Proc. 19, 713–755.

[15] Duncan, S., Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. J. Personal. Soc. Psychol. 23, 283–292.

[16] Cowley, S. J. (1998). Of Timing, Turn-Taking, and Conversations. J. Psycholing. Research 27, 541–571.

[17] Rühlemann, C., & Gries, S. T. (2020). Speakers advance-project turn completion by slowing down: A multifactorial corpus analysis. Journal of Phonetics, 80,100976.

[18] Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with" articulatory" pauses. Language and Speech, 26(3), 203-214.

[19] Local, J., & Walker, G. (2012). How phonetic features project more talk. JIPA, 42 (3), 255–280.

[20] Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. J. Mem. and Lang., 60(1), 92-111.

[21] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. JASA, 113(2), 1001-1024.

THE PROSODY OF HIGH AND LOW AFFECTIVE ENGAGEMENT IN POLISH AND GERMAN PARLIAMENTARY SPEECHES

Maciej Karpiński, Katarzyna Klessa, Brygida Sawicka-Stępińska, Hanna Kasperek

Adam Mickiewicz University, Poznań
{maciej.karpinski | katarzyna.klessa | brygida.sawicka-stepinska | hanna.kasperek}@amu.edu.pl

ABSTRACT

Parliamentary speeches tend to stick to predetermined topics and structures but they are often characterised by a high degree of emotionality and expressivity. Meant to sound convincing and reliable, they still raise emotions in the audience. In the era of international parliamentarism, it is important to understand the role of (para)linguistic and cultural settings in the perception of such speeches. This study is based on a multimodal corpus of parliamentary speeches from the German Bundestag and Polish Sejm, including multilayer annotations of speech and gestures. We inspect prosodic features of specific chunks of the recordings marked as speakers' "high" or "low" affective engagement areas. For the above engagement areas, the variability of pitch and timing features of the utterances is analysed. Our results show that despite progressing internationalisation and collaboration of politicians, certain differences in the paralinguistic prosody of German and Polish speeches remain significant.

Keywords: paralinguistic prosody, affective engagement, Polish, German, parliamentary speech

1. INTRODUCTION

Parliamentary debates have the potential to influence the direction of political and socio-economic changes. While the content of the speeches is of obvious importance, the way they are delivered may be at least equally significant [1]. Their persuasive power lies not only in the selection of words and in the argument structure but also in their prosodic characteristics [2, 3, 4, 5], with a substantial contribution from attitudinal, affective prosody [6, 7, 8]. Prosodic symptoms of emotional arousal typically occur in significant, ideologically engaged debates. However, speakers definitely differ in their abilities to control and hide emotions, as well as in the very strategy of showing or not showing them. Some act and fake emotions more consciously, while others are honestly engaged. Finally, their default, everyday individual speaking styles may differ, being perceived as generally more or less emotional. Therefore, the analysis of affective engagement may not only be important for correlating it with, e.g., persuasive power, but also helpful in decoding speakers' intentions and judging their credibility. In the present study, selected differences in the prosodic characteristics of Polish and German parliamentary speeches are explored. The areas of high and low affective engagement are marked manually on the basis of audio-visual scrutiny by Polish and German expert teams [9, 10]. Selected pitch- and duration-related parameters are measured for the Polish and German speakers in the areas of high and low affective engagement and hypothesized as cues for identifying these areas. We explore them using lineaer mixed effects regression models with language and affective engagement as fixed factors, and speaker as random effects. The results are discussed in the context of public political communication and multimodal analyses.

2. STUDY MATERIAL

The material under study comes from a multimodal corpus of German and Polish parliamentary speeches, developed within the MuMo Stance project (see Acknowledgement). The corpus comprises audio and video recordings of speeches delivered by the members of parliament during the 2020 budgetary debates, along with multi-tier, time-aligned annotation and segmentation on the level of the phrase, word, syllable, and phone. For speech transcription and segmentation, orthographic transcripts of parliamentary speeches available in the archives of the Polish Sejm and German Bundestag were used. They were time-aligned to inter-pausal units and manually adjusted by experienced annotators as they often lack phonetically important details (like repetitions or information on fillers). The inter-pausal units were used as the input for automatic segmentation into words, syllables, and phones. The segmentation of the Polish material was performed using the ANNPRO desktop module of CLARIN-

PL tools [11], while the German material was segmented using WebMaus [12]. In both cases, manual corrections of segment boundary positions were introduced based on visual inspection of spectrograms, auditory inspection and guidelines specified, for example, in [13]. Speakers' behaviour is described based on perception judgements, following the expressive movement framework approach [14]. The framework involves annotation of audio-video recorded utterances in terms of the areas of speakers' high and low affective engagement (HAE and LAE, respectively). Such areas are manually annotated using ELAN [15] in a three-fold procedure. First, each speech is divided into interactive units, i.e., sequences of speech interrupted either by applause or by any audible or visible interjections which receive a reaction from the speaker. Then, within each unit, HAE and LAE areas are distinguished on the basis of the perceived behaviour expressive intensity [16, 17]. As speakers vary in their ways and strategies of emotional expression, the annotators must gain an understanding of a speaker's unique style, as well as their personal benchmark of expression. For this reason, affective engagement areas were determined individually for each speaker, assessed according to their overall performance and forms of expression presented during the speech. The greater the expressive resources that are mobilized, the greater the level of affective engagement, and thus such passages are recognised as HAE areas; the remaining parts are described as LAE areas. The third step of the expressive movement description consists of tandem sessions, during which final version of each annotation is calibrated. Should discrepancies occur, they are discussed during a meeting of the annotators involved. The entire corpus was inspected for the occurrence of HAE and LAE areas. For the present study, only the speeches of male speakers containing both HAE and LAE areas were selected from the budgetary debate material. As a result, seven German and nine Polish speeches meeting the above criteria have been used. Polish speeches tend to be shorter (from 67s to 1567s, mean=454, sd=480 for Polish and from 236s to 977s, mean=575, sd=291 for German), while the proportion of LAE and HAE is almost equal (16:84 for Polish and 17:83 for German, respectively).

3. METHODS AND ANALYSES

In the present study, we analyse how selected prosodic characteristics of HAE and LAE areas differ between Polish and German parliamentary speeches. The following prosodic parameters were taken into account as potentially influenced by the affective state of the speaker: mean pitch frequency, pitch frequency standard deviation [18, 19], mean intensity, intensity standard deviation [20, 21], vocalic, consonantal, and syllabic nPVI [e.g., [22, 23, 24]], and TGA duration slope and intercept linear regression values [25]. For quantitative analyses, all the HAE and LAE areas were sampled using a 5 second time window which gave the total of 654 samples (537 from LAE and 117 from HAE areas) for German and 505 (424 from LAE and 81 from HAE areas) for Polish. For each sample, the prosodic parameters were extracted or calculated using Annotation Pro plugins [26]. Pitch frequency values were extracted from pitch-smoothed [27] signals using Praat's [28] standard autocorrelation method with the frequency range from 75Hz to 350Hz. The upper limit was selected on the basis of previous studies of pitch in male parliamentary speakers e.g., [29]. A custom-made plugin was used to extract mean pitch and standard deviation values from the samples. Intensity values were extracted using standard Praat algorithm with the mean energy averaging method, and imported into Annotation Pro. However, they turned out to systematically differ between the languages in a way that may have resulted from some differences between recording setups in the two parliaments. While potentially useful for other purposes, here they were excluded from further analyses. The TGA slope and intercept values were calculated with the Annotation Pro+TGA plugin [30] for the entire affective engagement areas and then inherited by the samples taken from those areas. Table 1 provides basic descriptive statistics for the variables under study. Linear mixed effects regression models (lmar in R) were fitted for each variable with language and affective engagement, their interactions as fixed factors and speaker as random effects (random intercept and slope for affective engagement), cf. [31]. P-values were adjusted using the Benjamin-Hochberg correction [32]. The results of analyses with $p<0.05$ are listed in Table 2. In spite of their apparent significance, most of the main effects may be disputable as they may result from systematic differences between the two languages or from strong variation among individual speakers, as their number is relatively limited. However, the main interactions of the mean pitch value, mean syllabic nPVI, and mean vocalic nPVI interactions, may provide an insight into some relevant phenomena. The remaining factors did not have a significant effect ($p>0.07$). As shown in Fig. 1, the distribution

of the mean pitch values in samples from LAE and HAE areas differs between the languages, and a clear difference can be spotted between LAE and HAE values for the German language. In Fig. 2, the distribution of the syllabic nPVI is shown, and Fig. 3 illustrates the data on the vocalic nPVI from the LAE and HAE in German and Polish speakers. The distribution of the syllabic nPVI values within languages does not differ much between LAE and HAE but there is a significant difference between the languages. The same applies to the vocalic nPVI where the distributions are relatively similar within languages but differ between them, with lower means for Polish and more dispersion of the values for the LAE samples. It may be partially due to the phonological duration present in German but absent in Polish. Nevertheless, at this stage, one cannot exclude the impact resulting from factors related to different traditions of public speaking in Germany and in Poland. Higher TGA mean slope values are observed for Polish than in German. For both languages, the slope is lower in HAE which means less deceleration in interpausal time groups than in LAE. Basic descriptive statistics of the explored parameters in the material under study are listed in Table 1.

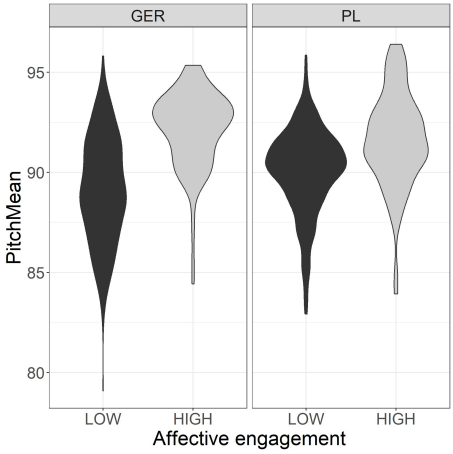


Figure 1: Mean pitch values in the LAE and HAE in German and Polish parliamentary speeches.

4. CONCLUSIONS

Political speeches are a peculiar instance of language use where speakers address a very wide, often physically absent audience. They tend to be well-controlled and are often pre-prepared but still emotional. However, the expression of emotions may be influenced or channeled by the social and cultural factors, including, e.g., the parliamentary

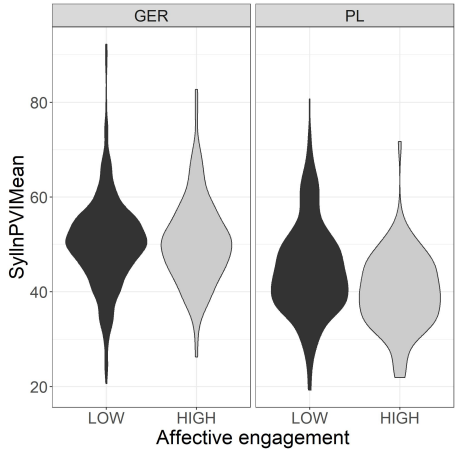


Figure 2: Mean syllabic nPVI values in the LAE and HAE in German and Polish parliamentary speeches.

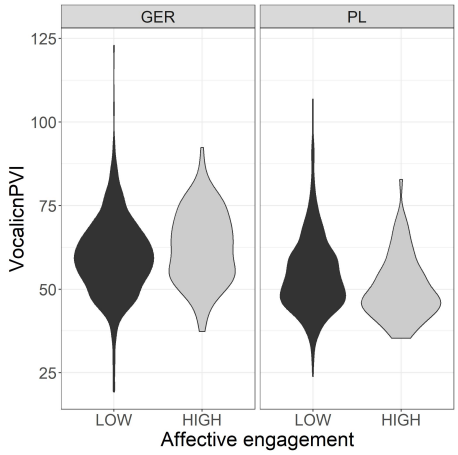


Figure 3: Mean vocalic nPVI values in the LAE and HAE in German and Polish parliamentary speeches.

rules of conduct. Therefore, because of its peculiar characteristics and contextual factors, this variety still requires exploration in spite of the huge body of already available studies on emotional speech prosody. In the present paper, some differences are shown in how the states of high and low affective engagement are reflected in the prosody of German and Polish parliamentary speakers. As expected, the language itself, along with its culture-related, paralinguistic component, turns out to be a very strong factor. While the meaning of the main effects may be disputable, it may be clarified with a larger group of participants. The main interactions of the mean pitch value, mean syllabic nPVI, and mean vocalic nPVI, seem to point to a promising direction for further research on both

		GERMAN		POLISH	
	Affective engagement	Mean	Standard deviation	Mean	Standard deviation
Vocalic nPVI	LAE	59.80	11.47	50.33	11.27
	HAE	62.67	10.80	54.41	9.38l
Consonantal nPVI	LAE	45.91	10.52	48.43	8.56
	HAE	44.88	11.57	48.53	7.45
Syllabic nPVI	LAE	50.34	9.29	44.67	9.74
	HAE	50.52	9.11	40.04	8.00
Mean Pitch (ST)	LAE	89.15	2.62	90.09	2.12
	HAE	92.17	1.90	91.64	2.36
SD of Pitch (ST)	LAE	2.42	0.94	2.68	0.87
	HAE	2.15	0.79	2.80	0.83
TGA Mean Slope	LAE	0.02	0.08	0.09	0.16
	HAE	0.00	0.05	0.05	0.10
TGA Mean Intercept	LAE	195.82	45.38	166.20	44.36
	HAE	221.28	44.20	167.69	36.14

Table 1: Basic descriptive statistics of the explored parameters in the material under study.

	analysis	p-value	adjusted p-value	DenDF	F
1	PitchMean_main_AffEng	0.000	0.000	15.70	85.92
2	PitchMean_interaction	0.04606	0.04606	15.70	4.69
3	SlopeMean_main_AffEng	0.04076	0.0439	8.57	5.79
4	SlopeMean_main_Language	0.00039	0.00109	14.16	21.30
5	InterceptMean_main_AffEng	0.0199	0.02567	12.9	7.05
6	InterceptMean_main_Language	9.00E-04	0.00193	13.71	17.76
7	SyllnPVIMean_main_AffEng	0.02017	0.02567	1091.89	5.41
8	SyllnPVIMean_main_Language	9.00E-05	0.00041	13.58	30.22
9	SyllnPVIMean_interaction	0.03534	0.04123	1091.89	4.44
10	VocalicnPVI_main_Language	0.00097	0.00193	11.12	19.72
11	VocalicnPVI_interaction	0.00857	0.01333	78.55	7.27

Table 2: The significance of interactions and main effects (only the cases of p<0.05 are listed). AffEng stands for Affective Engagement of the speaker, labelled as "low" or "high".

individual differences (including gender or political orientation) and more general tendencies regarding the roles of the prosodic parameters as cues to the identification of low and high affective engagement. Our research will include the gestural component of the speeches and facial expression as well [33]. Finally, a cross-cultural testing of the perception of the prosodic characteristics will be carried out [34].

5. ACKNOWLEDGEMENTS

MuMo Stance project is funded by the Polish National Science Fundation (NCN 2018/31/G/HS2/03633) and Deutsche Forschungsgemeinschaft. We are extremely grateful to Bettina Braun for her invaluable advise on data analysis.

6. REFERENCES

[1] I. Poggi, F. D'Errico, L. Vincze, and A. Vinciarelli, Multimodal Communication in Political Speech Shaping Minds and Social Action: Int. Workshop, Political Speech 2010. Springer, 2013, vol. 7688.

[2] P. Touati, "Prosodic Aspects of Political Rhetoric," in ESCA Workshop on Prosody, 1993.

[3] J. B. Hirschberg and A. Rosenberg, "Acoustic/Prosodic and Lexical Correlates of Charismatic Speech," Columbia Academic Commons, 2005.

[4] G. Kišiček, "Persuasive power of prosodic features," Argumentation and Advocacy, vol. 54, no. 4, pp. 345–350, 2018.

[5] J. Michalsky, H. Schoormann, and T. Schultze, "Towards the prosody of persuasion in competitive negotiation. The relationship between f0 and negotiation success in same sex sales tasks," in Interspeech, 2019, pp. 311–315.

[6] F. D'Errico, R. Signorello, D. Demolin, and

I. Poggi, "The Perception of Charisma from Voice: A Cross-Cultural Study," in Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE, 2013, pp. 552–557.
[7] R. L. Mitchell and E. D. Ross, "Attitudinal prosody: What we know and directions for future study," Neuroscience & Biobehavioral Reviews, vol. 37, no. 3, pp. 471–479, 2013.
[8] E. Strangert, "Prosody in public speech: analyses of a news announcement and a political interview," in 9th European Conference on Speech Communication and Technology, 2005.
[9] H. Kappelhoff and C. Müller, "Embodied meaning construction: Multimodal metaphor and expressive movement in speech, gesture, and feature film," Metaphor and the social world, vol. 1, no. 2, pp. 121–153, 2011.
[10] C. Müller and H. Kappelhoff, "Cinematic metaphor," in Cinematic Metaphor. De Gruyter, 2018.
[11] K. Klessa, D. Koržinek, B. Sawicka-Stępińska, and H. Kasperek, "ANNPRO: A Desktop Module for Automatic Segmentation and Transcription," in Language and Technology Conference. Springer, 2022, pp. 65–77.
[12] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," Computer Speech & Language, vol. 45, pp. 326–347, 2017.
[13] P. Machač and R. Skarnitzl, Principles of Phonetic Segmentation. Epocha, 2009.
[14] D. Horst, F. Boll, C. Schmitt, and C. Müller, "Gesture as interactive expressive movement: Inter-affectivity in face-to-face communication," in Body - Language - Communication: An International Handbook on Multimodality in Human Interaction, vol. 38.2. De Gruyter Mouton, 2014, pp. 2112–2126.
[15] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in 5th LREC, 2006, pp. 1556–1559.
[16] C. Müller, Metaphors Dead and Alive, Sleeping and Waking: A Dynamic View. Univ. of Chicago Press, 2008.
[17] C. Müller and S. Tag, "The Dynamics of Metaphor: Foregrounding and Activating Metaphoricity in Conversational Interaction," Cognitive Semiotics, vol. 6, no. s1, pp. 85–120, 2010.
[18] J. Vroomen, R. Collier, and S. J. Mozziconacci, "Duration and intonation in emotional speech." in Eurospeech, 1993.
[19] B. J. Dietrich, M. Hayes, and D. Z. Oâbrien, "Pitch Perfect: [v]ocal Pitch and the Emotional Intensity of Congressional Speech," American Political Science Review, vol. 113, no. 4, pp. 941–962, 2019.
[20] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," JASA, vol. 93, no. 2, pp. 1097–1108, 1993.
[21] C. Pereira, "Dimensions of emotional meaning in speech," in ISCA ITRW Speech and Emotion, 2000.
[22] L. E. Low, E. Grabe, and F. Nolan, "Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English," Language and Speech, vol. 43, no. 4, pp. 377–401, 2000.
[23] L. Quinto, W. F. Thompson, and F. L. Keating, "Emotional Communication in Speech and Music: The Role of Melodic and Rhythmic Contrasts," Front. Psychol., vol. 4, p. 184, 2013.
[24] M. Goudbeek and M. Broersma, "Rhythm in vocal emotional expressions: the normalized pairwise variability index differentiates emotions across languages," in Bi-Annual Conference of ISRE, Geneva, 2015.
[25] D. Gibbon, "TGA: a web tool for Time Group Analysis," in Proc. of the TRASP Workshop, Aix-en-Provence, 2013, pp. 66–69.
[26] K. Klessa, M. Karpiński, and A. Wagner, "Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features," in Proc. TRASP Workshop, Aix-en-Provence, 2013, pp. 51–54.
[27] "Praat Vocal Toolkit," https://https://www.praatvocaltoolkit.com, visited 5-Jan-23.
[28] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 4.3.14) [computer program]," 2014, (visited 20-Dec-17). [Online]. Available: http://www.praat.org
[29] M. Karpiński, K. Klessa, and E. Jarmołowicz-Nowikow, "High-pitched prominences in the speeches of male Polish members of parliament," in Proc. of the 11th Speech Prosody, Lisbon, 2022, pp. 367–371.
[30] K. Klessa and D. Gibbon, "Annotation Pro+ TGA: automation of speech timing analysis," in Proc. of the 9th International LREC, 2014, pp. 1499–1505.
[31] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," Journal of memory and language, vol. 59, no. 4, pp. 390–412, 2008.
[32] D. Thissen, L. Steinberg, and D. Kuang, "Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons," Journal of Educational and Behavioral Statistics, vol. 27, no. 1, pp. 77–83, 2002.
[33] J. Streeck, "Gesture in Political Communication: A Case Study of the Democratic Presidential Candidates During the 2004 Primary Campaign," Research on Language and Social Interaction, vol. 41, no. 2, pp. 154–186, 2008.
[34] A. Cullen, A. Hines, and N. Harte, "Building a Database of Political Speech: Does Culture Matter in Charisma Annotations?" in Proc. of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 27–31.

PROSODIC REALIZATION OF SYNTACTIC PHRASE AND CLAUSE BOUNDARIES IN TOKYO JAPANESE

Shinichiro Ishihara

Centre for Languages and Literature, Lund University
shinichiro.ishihara@ostas.lu.se

ABSTRACT

This paper presents the results of a production experiment on the syntax–prosody mapping in Tokyo Japanese, whereby the phonetic realization of syntactic phrase and clause boundaries were systematically compared. The data consist of 672 utterances of read speech from fourteen subjects. The results corroborate earlier findings on the mapping of syntactic phrases to phonological phrases, but do not support the claim that clauses are mapped to intonational phrases. Data from a few subjects suggest instead that there may be a mapping of discourse-related domains (such as topic and comment) to intonational phrases.

Keywords: syntax–prosody mapping, Japanese, syntactic phrase, clause, intonational phrase

1. INTRODUCTION

Recently proposed theories of intonation [1, 2, 3, 4, 5] put forward a strong version of the syntax–prosody mapping hypothesis, which states i) that there are, language universally, three prosodic categories—phonological word (ω), phonological phrase (φ) and intonational phrase (ι)—above the foot, and ii) that these three prosodic categories show correspondence to three syntactic categories, words, phrases, and clauses, respectively. Match Theory [5], for example, explicitly hypothesizes that syntactic phrases are mapped to φs and clauses to ιs.

In the case of (Tokyo) Japanese, however, it is still unclear whether there is such a clear-cut distinction between the phrase-level and clause-level syntax–prosody mapping. Previous studies have shown that left edges of syntactic constituents—be it phrasal or clausal—are realized with high f_0-peaks. For example, the left edge of a relative clause has shown to be marked by a high f_0-peak [6, 7]. [8] showed that left edges of syntactic phrases (XPs) are realized with an f_0-boosting phenomenon called *metrical boost* and that metrical boost is larger when there are two syntactic boundaries at one position than when there is only one syntactic boundary. In these studies, however, no distinction was made between phrase and clause boundaries, or between φ and ι.

According to [9], five phonetic cues in (1) distinguish ιs from φs.

(1) a. ι-final f_0-lowering
b. creaky vowels at the end of ι
c. obligatory pauses at the end of ι
d. stronger pitch resets at the onset of ι
e. larger pitch rises at the onset of ι

Using these cues, [9] showed that syntactic clauses are mapped as ιs in Japanese, as clause boundaries are prosodically marked differently from phrase boundaries. However, since the coordinated clauses tested in this study were all main clauses, it still remains to be examined whether embedded clauses, such as clausal objects, exhibit phonetic cues in (1). This point is particularly important because it has already been acknowledged [4] that some languages do not map embedded clauses to ι despite the proposed mapping hypothesis.

Opposite results have also been reported (though with a caveat). In the study of the prosody of embedded clauses [10] and relative clauses [11] containing a single content word, it was found that the embedded/relative clause undergoes downstep[1] relative to the preceding material, indicating that they are contained in a single φ, which entails that these clauses are not mapped to ιs. However, since Japanese is known to be subject to the so-called *rhythmic (a.k.a. binarity) effect* [8, 12], which triggers grouping of two phonological words into a single φ, the one-word embedded/relative clauses may have been grouped together with the preceding word, forming a φ, due to this effect. It remains to be seen if downstep can be observed in longer embedded clauses.

2. EXPERIMENT

2.1. Methodology

A phonetic production experiment was conducted with fourteen native speakers of Japanese (nine females, five males). They were undergraduate or graduate students at a university located in Tokyo

and were from Tokyo or surrounding areas.

Four conditions (named 0xp, 1xp, 2xp, and cp) were tested with 4 lexical items each. The stimuli (target sentences + 192 fillers) were presented on a computer screen one sentence at a time. Each subject read all of the stimuli three times, in three different pseudo-randomized orders. This procedure yielded a total of 672 samples of the target sentences (14 speakers×4 conditions×4 items×3 repetitions).

The recorded target sentences were annotated using Praat [13]. They were first forced aligned for segment and word boundaries using Julius ver. 4.5 [14], Julius Segmentation Kit ver. 4.3.1, and pyjuliusalign ver. 2.0 [15]. Word boundaries were checked and corrected manually. Then, for each word, f_o-maximum and f_o-minima before and after the f_o-maximum were measured, with the help of Praat scripts. f_o-maxima were chosen either from a point that is judged as corresponding to the peak of a lexical pitch accent (H*+L in X-JToBI notation, [16]) or on the BPM at the end of each word. In addition, the presence of a detectable pause after Word1 and Word2 was checked by listening to the data and by visually examining pitch contours.

2.2. Stimuli

The four conditions share the same first four nouns (Word1–Word4) and the verb, which are all lexically accented. The cp condition additionally contained a complementizer (*to*) and a verb. (2)–(5), which are one of the four sets of lexical items used in the study, differ in terms of the number and the type of syntactic boundaries in front of the target word, Word2 (*Naoya*).

(2) No XP boundary (0xp)
[$_{NP}$ *Yuuta-to Naoya-wa*] [$_{VP}$ *imooto-o*
Y.-and N.-TOP sister-ACC
paatii-ni maneita]
party-to invited
'Yuta and Naoya invited their sisters to the party.'

(3) One XP boundary (1xp)
Yuuta-wa [$_{VP}$ *Naoya-o* [$_{NP}$ *imooto-no*
Y.-TOP N.-ACC sister-GEN
paatii-ni] *maneita*]
party-to invited
'Yuta invited Naoya to his sister's party.'

(4) Two XP boundaries (2xp)
Yuuta-wa [$_{VP}$ [$_{NP}$ *Naoya-no imooto-o*]
Y.-TOP N.-GEN sister-ACC
paatii-ni maneita]
party-to invited
'Yuta invited Naoya's sister to the party.'

(5) Clause boundary (cp)
Yuuta-wa [$_{CP}$ *Naoya-ga imooto-o paatii-ni*
Y.-TOP N.-NOM sister-ACC party-to
maneita to] *omotteita*
invited that was.thinking
'Yuta believed that Naoya invited his sister to the party.'

In 0xp (2), there is no XP-boundary between Word1 and Word2 as they form an NP together. In 1xp (3), there is a VP-boundary between Word1 and Word2. In 2xp (4), in addition to the VP-boundary, there is a boundary of the NP containing Word2 and Word3. Lastly, in cp (5), there is an embedded clause boundary. If the number of syntactic boundaries (0xp, 1xp, 2xp) and/or the type of boundary (phrase or clause) affects prosody, the realization of the target word is expected to reflect these effects.

2.3. Predictions

Based on the the previous findings summarized in §1, two predictions can be made as to the realization of the target word (Word2) in the four conditions (0xp, 1xp, 2xp, cp). The first prediction concerns the mapping of syntactic phrases to φs. When comparing 0xp, 1xp, and 2xp, the f_o-peak of Word2 in 0xp is predicted to be lower than that of 1xp and 2xp, because Word2 of 0xp is subject to downstep due to the lack of syntactic boundary to its left [8, 17, 18]. Furthermore, when 1xp and 2xp are compared, it is predicted that the f_o-peak of Word2 in 2xp is realized higher than that in 1xp, because *metrical boost* is stronger when more than one syntactic boundary co-occurs at one position [8].

(6) Predictions 1: 0xp < 1xp < 2xp

The second prediction concerns the mapping of clauses to ιs. According to (1), an ι-boundary shows a higher pitch rises than a φ-boundary. If the embedded clause in cp is mapped to an ι, it is expected that the f_o-peak of Word2 is higher in cp than in 1xp and 2xp.

(7) Predictions 2: 1xp, 2xp < cp

In addition to the expected differences in f_o-peak height, pauses are expected consistently before word2 in cp, as per (1c), if the embedded clause is mapped to ι and preceded by another ι.

3. RESULTS

The pooled data of all the fourteen subjects confirmed Prediction 1 while they did not confirm Prediction 2. Figure 1 shows the normalized means of the f_o-maximum on the target word (Word2) in each condition. 0xp, in which no metrical boost but downstep is expected, shows the lowest f_o-maximum. 1xp and 2xp show higher f_o-peaks compared to 0xp. Furthermore, 2xp is significantly higher than 1xp (2-sided t-test: $t = -3.9389, df = 333.1, p = 9.974\text{e}-05$). In contrast, there is no significant difference between 2xp and cp (2-sided t-test: $t = 0.11212, df = 327.79, p = 0.9108$).

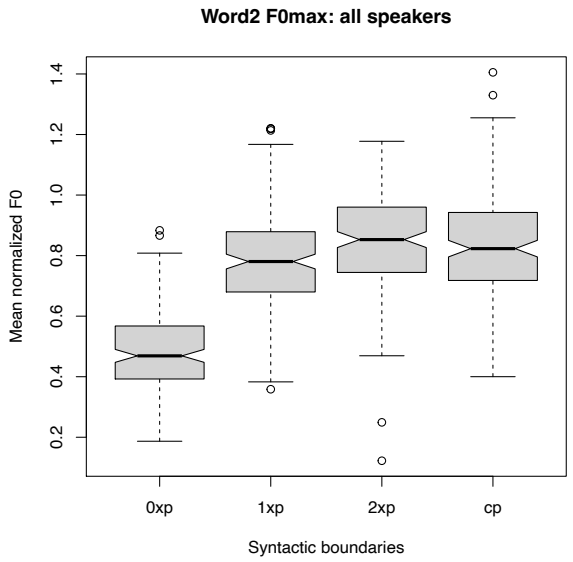


Figure 1: Boxplot of normalized means of f_o-maximum on Word2 for all subjects

Table 1 shows the frequency of detectable pauses after Word1 and Word2. The frequency of pauses after Word1 in cp (41.7%) is comparable to those in 1xp and 2xp (36.3% and 38.7%, respectively), which suggests that there is no sign of ι-boundary at this position. Another point to note is that there are much fewer pauses after Word1 in 0xp (5.4%) and after Word2 in 2xp (17.3%) than other places. This is presumably because these positions are between two coordinated nouns (*Yuuta-to Naoya-wa* 'Yuta-and Naoya-TOP' in (2); *Naoya-no imooto-o* 'Naoya-GEN sister-ACC' in (4)), that is, where there is no syntactic XP boundaries. At these positions, downstep was also observed (as in 0xp in Figure 1).

There were inter-speaker variations that need to be noted. In what follows, speakers are numbered (01–14) and indexed with f/m to indicate the gender.

	Word1		Word2	
0xp	5.4%	(9)	72.6%	(122)
1xp	36.3%	(61)	53.0%	(89)
2xp	38.7%	(65)	17.3%	(29)
cp	41.7%	(70)	33.9%	(57)

Table 1: Percentages (and counts) of pauses detected after Word 1 and Word2 in each condition. The number of total data points in each position and condition is 168.

As for Prediction 1 in (6), while all but one speaker (05m) showed a significant contrast in 0xp < 1xp, a half of the speakers did not show the expected contrast 1xp < 2xp.

(8) a. Significant difference (1xp < 2xp):
Speakers 01f, 03f, 04m, 06f, 07m, 09m, 10f
b. No significant difference (1xp $\not<$ 2xp):
Speakers 02f, 05m, 08m, 11f, 12f, 13f, 14f

Regarding Prediction 2 in (7), i.e, the comparison of 2xp and cp, the subjects can be divided into three groups, as in (9). The first group (9a) did not show any significant difference between 2xp and cp, similarly to the pooled data in Figure 1. The second group (9b) showed a significantly lower f_o-max in cp than in 2xp (2xp > cp), i.e., in the opposite direction to the prediction. Only the last group (9c) showed the expected difference, 2xp < cp. This means that the first two groups, (9a) and (9b), do not confirm the prediction of the clause-level mapping, while group (9c) is in line with it.

(9) a. No significant difference between 2xp and cp ($2\text{xp} \not< \text{cp}, 2\text{xp} \not> \text{cp}$):
Speakers 02f, 03f, 08m, 09m, 13f
b. cp significantly lower than 2xp (2xp > cp):
Speakers 01f, 04m, 06f, 07m, 10f
c. cp higher than 2xp (cp > 2xp)
Speakers 05m, 11f, 12f, 14f

There is a noteworthy correlation between the grouping in (8) (1xp vs. 2xp) and that in (9) (2xp vs. cp): Group (9b) speakers all belong to Group (8a), while Group (9c) speakers belong to (8b).

Sample pitch tracks of cp taken from these two groups illustrate this difference. Pitch tracks from Group (9b) tend to show a lower Word2 f_o-peak in relation to the preceding word (Word1) in cp (Figure 2). This group also showed a similar pattern in 1xp. The Word2 f_o-peak values in 1xp and cp were therefore comparable to each other. Figure 2 shows downstep from Word1 to Word2 followed by the pitch resetting on Word3, which means that Word1 and Word2 are phrased together within a single φ.

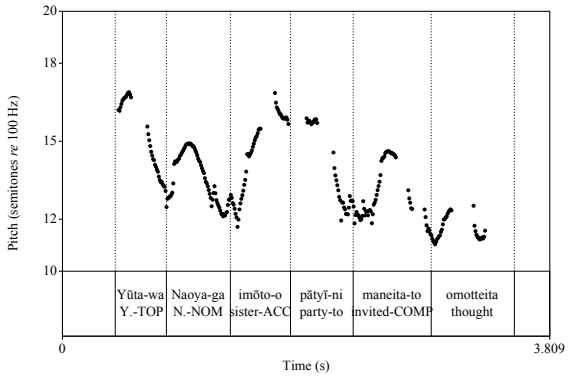


Figure 2: Sample pitch track of cp from a speaker in Group (9b). Word2 is downstepped relative to Word1.

This phrasing is presumably due to the rhythmic effect [8, 12], which can be explained as avoidance of φ containing a single prosodic word. Since both Word1 and Word2 are single-word NPs in 1xp and cp, they are phrased together due to this effect. This phrasing pattern excludes the possibility to have an ι boundary between Word1 and Word2, both in 1xp and cp conditions.

In contrast to Group (9b), Group (9c) speakers showed a high f_0-peak on Word2, as exemplified in Figure 3. Furthermore, these speakers showed a tendency to insert a pause between Word1 and Word2 in cp. Given that pauses are one of the phonetic cues of ι boundary (see (1c)), it can be concluded that there is an ι boundary at this position.

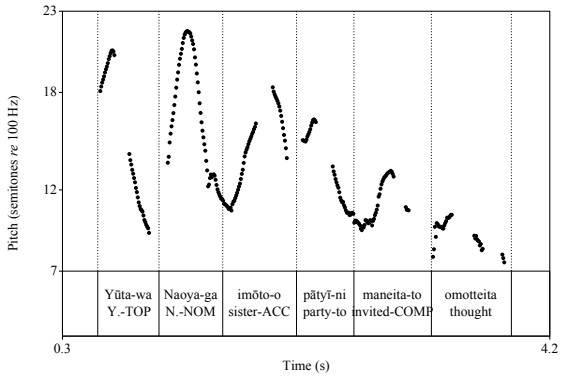


Figure 3: Sample pitch track of cp from a speaker in Group (9c). Word2 shows a high f_0-peak, preceded by a pause.

In their data, however, high f_0-peaks and pauses were observed in other conditions as well. Group (9c) speakers showed a high f_0-peak in 1xp and 2xp as well as cp. Note that none of the Group (9c) speakers showed the contrast 1xp $<$ 2xp, as in (8b).

In fact, the location of pauses seems to be rather correlated with the location of topic marker *wa*. The topic marker *wa* is attached to Word1 in 1xp, 2xp, and cp, and to Word2 in 0xp. Three of the four speakers in Group (9c) show a clear tendency to place a pause after a noun marked with *-wa*. Table 2 shows that pauses are consistently found in all the locations with *-wa* in these speakers' data.

	Word1		Word2	
0xp	0.0%	(0)	100.0%	(36)
1xp	75.0%	(27)	36.1%	(13)
2xp	80.6%	(29)	25.0%	(9)
cp	100.0%	(36)	11.1%	(4)

Table 2: Percentages (and counts) of pauses detected after Word 1 and Word2 in each condition by speakers 11f, 12f, and 14f.

4. DISCUSSION

The results of this study suggest that there is no mapping of embedded clause to ι in Japanese. Although this conclusion is still in line with Match Theory, as claimed by [5], the results also showed a stronger correlation between the location of the topic marker *wa* and the presence of pauses for some speakers, which seems to suggest that ιs may be more directly correlated with discourse-related notions such as topic and comment.

It has been claimed that ιs correlate with illocutionary forces, or speech acts associated with them [19, 20, 21, 22]. There is also a claim that the introduction of topics is a type of speech act [23]. Taking these claims into consideration, Ishihara [24] proposes that, contrary to the syntax–prosody mapping hypothesis presented in §1, ι is not part of the syntax–prosody mapping, and that it is rather part of the discourse–prosody mapping: Speech acts are mapped as ιs in the prosodic representation. Further investigation is needed to test this claim.

Another point to be noted is the rhythmic effect observed in Group (9b) speakers' data. Downstep on Word2 in cp (Figure 2) suggests that two single-word NPs can be grouped into a single φ even when there is a clause boundary in between. This results explains the otherwise puzzing results in [10, 11].

5. ACKNOWLEDGMENT

This study was supported by Swedish Research Council (2018-01539). I sincerely thank Natsumi Goto, Thea Johansson, and Dillen Luis Smit for their help with data annotation, and anonymous reviewers for useful comments. All errors are my own.

6. REFERENCES

[1] J. Itô and A. Mester, "Prosodic adjunction in Japanese compounds," in *Formal Approaches to Japanese Linguistics 4 (FAJL4)*, 2007, pp. 97–111.

[2] ——, "Recursive prosodic phrasing in Japanese," in *Prosody Matters: Essays in Honor of Elisabeth Selkirk*, T. Browsky, S. Kawahara, T. Shinya, and M. Sugahara, Eds. London: Equinox Publishing, 2012, pp. 280–303.

[3] ——, "Prosodic subcategories in Japanese," *Lingua*, vol. 124, no. 1, pp. 20–40, 2013.

[4] E. Selkirk, "On clause and intonational phrase in Japanese: The syntactic grounding of prosodic constituent structure," *Gengo Kenkyu*, vol. 136, pp. 35–74, 2009.

[5] ——, "The syntax-phonology interface," in *The Handbook of Phonological Theory*, 2nd ed., J. Goldsmith, J. Riggle, and A. Yu, Eds. Oxford, UK: Blackwell, 2011, pp. 435–484.

[6] T. Uyeno, H. Hayashibe, and K. Imai, "On pitch contours of declarative, complex sentences in Japanese," *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, vol. 13, pp. 175–187, 1979.

[7] T. Uyeno, H. Hayashibe, K. Imai, H. Imagawa, and S. Kiritani, "Syntactic structure and prosody in Japanese: A study on pitch contours and the pauses at phrase boundaries," *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, vol. 15, pp. 91–108, 1981.

[8] H. Kubozono, *The Organization of Japanese Prosody*. Tokyo: Kurosio Publishers, 1993.

[9] S. Kawahara and T. Shinya, "The intonation of gapping and coordination in Japanese: Evidence for intonational phrase and utterance," *Phonetica*, vol. 65, no. 1–2, pp. 62–105, 2008.

[10] M. Hirayama and H. K. Hwang, "Not all XPs affect prosody in Japanese," in *NELS 46: Proceedings of the North Eastern Linguistics Society*, C. Hammerly and B. Prickett, Eds., 2016, pp. 95–104.

[11] ——, "Relative clause and downstep in Japanese," in *Supplemental Proceedings of the 2018 Annual Meeting on Phonology*, K. Hout, A. Mai, A. McCollum, S. Rose, and M. Zaslansky, Eds. Washington, DC: Linguistic Society of America, 2019.

[12] T. Shinya, E. Selkirk, and S. Kawahara, "Rhythmic boost and recursive minor phrase in Japanese," in *Proceedings of the Second International Conference on Speech Prosody*, 2004, pp. 345–348.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computers [Computer programm]," Version 6.1.53, retrieved 8 September, 2021 from http://www.praat.org, 2021. [Online]. Available: http://www.praat.org

[14] A. Lee and T. Kawahara, "Julius v4.5," 2019.

[15] T. Mahrt, "pyjuliusalign v.2.0," 2019. [Online]. Available: https://github.com/timmahrt/pyJuliusAlign

[16] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: An extended J_ToBI for spontaneous speech," in *Proceedings of The 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, September 16–20 2002, pp. 1545–1548.

[17] E. Selkirk and K. Tateishi, "Syntax and downstep in Japanese," in *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, C. Georgopoulos and R. Ishihara, Eds. Dordrecht: Kluwer Academic Publishers, 1991, pp. 519–543.

[18] S. Ishihara, "Japanese downstep revisited," *Natural Language & Linguistic Theory*, vol. 34, no. 4, pp. 1389–1443, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11049-015-9322-8

[19] E. Selkirk, "Comments on intonational phrasing in English," in *Prosodies: With Special Reference to Iberian Languages*, S. Frota, M. Vigário, and M. J. ao Freitas, Eds. Berlin/New York: Mouton de Gruyter, 2005, pp. 11–59.

[20] G. Güneş, "Constraints on syntax-prosody correspondence: The case of clausal and subclausal parentheticals in Turkish," *Lingua*, vol. 150, pp. 278–314, 2014.

[21] ——, "Deriving prosodic structures," Ph.D. dissertation, University of Groningen, 2015.

[22] H. Truckenbrodt, "Intonation phrases and speech acts," in *Parenthesis and Ellipsis: Cross-Linguistic and Theoretical Perspective*, M. Kluck, D. Ott, and M. de Vries, Eds. Berlin/Boston: De Gruyter Mouton, 2015, pp. 301–349.

[23] C. Ebert, *Quantificational Topics: A Scopal Treatment of Exceptional Wide Scope Phenomena*. Dordrecht: Springer, 2009.

[24] S. Ishihara, "On the (lack of) correspondence between syntactic clauses and intonational phrases," in *Prosody and Prosodic Interfaces*, H. Kubozono, J. Ito, and A. Mester, Eds. Oxford, UK: Oxford University Press, 2022, pp. 420–456.

[1] In Japanese, downstep is a pitch range compression triggered by H*+L lexical pitch accents [18] and takes the φ as its domain. That is, f_0-peaks within a φ are cumulatively lowered relative to the previous one, and at the beginning of the following φ, pitch resetting takes places.

Perceptual learning of novel intonation patterns by Cantonese and Wu speakers

Man Yan Priscilla Lam, Wenwei Xu, Yiya Chen

Leiden University Centre for Linguistics, Leiden University
m.y.p.lam@hum.leidenuniv.nl; w.xu@hum.leidenuniv.nl; yiya.chen@hum.leidenuniv.nl

ABSTRACT

This study explores how tonal language speakers categorize novel intonation patterns. The perceptual intonation learning paradigm constructed in [1] was tested on native speakers of two Sinitic languages, Cantonese and Wu. The results were compared with those from the English native speakers in [1]. No difference was found in the success or rejection rates between the Sinitic language speakers and English speakers in various conditions. However, further analysis of reaction time revealed subtle differences between the speaker groups, such as a longer reaction time observed for English speakers when processing some of the intonation patterns. The results suggest that while the ability to categorize simple intonation contours may be universal among tonal and non-tonal language speakers, listeners' sensitivity to different cues can vary depending on their linguistic background.

Keywords: intonation, perceptual learning, categorization, pitch aptitude

1. INTRODUCTION

1.1. Pitch variation and intonation processing

Pitch variation is a vital part of the speech signal in human languages. Across spoken languages of the world, in any utterance containing sonorant segments, there exists pitch variation. The possible (para)linguistic functions of pitch variation are wide-ranging. In lexical tone, for example, which is estimated to be present in 60–70% of the world's languages [2], pitch variation can occur at the syllable level to mark lexical meaning. In contrast, in intonation, which is a universal feature of all languages [3], pitch variation occurs at the utterance level to convey discourse-level meaning such as sentence mode and information structure [4].

That many communicative functions can be encoded by pitch variation leaves curious implications on how speakers of typologically different languages process pitch signals. For successful speech communication, intonation language speakers need to attend to pitch variation at the utterance level, while tonal language speakers need to additionally attend to pitch variation at a much lower level such as syllables and morphemes. Given the continuously varying nature of pitch in an utterance, what has remained open is how, during intonation processing, listeners of tonal languages extract the relevant pitch cues for discourse-level information.

In intonation languages, a popular view regarding intonation processing is that listeners can parameterize an intonation contour into major intonational events such as pitch accents and boundary tones [4]. Previous studies on English online intonation processing provided evidence corroborating this claim [5,6].

The ability of intonation language listeners to pay attention to major intonational events and use them as cues for discourse comprehension likely makes intonation learning more effective and less cognitively demanding. Kapatsinski and colleagues [1] conducted a perceptual intonation learning experiment with English children and adults, involving stimuli derived from three intonation pattern prototypes: flat, final fall, and double peak. They found that both children and adults were able to learn new intonation contours quickly and formed relatively abstract representations of the contours. Adults were also found to attend more to the end of a contour than the beginning.

To our knowledge, there is no research on intonation learning by speakers of tonal languages. Considering tonal speakers need to constantly detect syllable- or morpheme-level pitch variation to determine lexical meaning, it is reasonable to assume that they use higher temporal resolution than intonation language speakers during pitch perception. Such an assumption is also in line with the PENTA model of intonation [7] where pitch targets are specified per syllable. A possible consequence is that they may find it difficult to only attend to major intonational events when learning temporally extended patterns, and thus perform worse than intonation language speakers in this respect.

1.2 Overview of the present study

In this project, we set out to investigate how tonal language speakers process pitch information to learn intonation patterns and whether their learning patterns differ from those of non-tonal speakers. We adopted the perceptual learning paradigm to address

our research question. In perceptual learning, it is believed that experienced categories could enhance one's future perceptual processing ability. In [1], a perceptual intonation learning experiment was conducted with speakers of English. We aimed to extend this line of research with speakers of two Sinitic languages: Hong Kong Cantonese (Yue; 'HKC') and Kunshan Wu (Wu; 'KS').

Within the Sinitic language family, Cantonese and Wu constitute an interesting pair for comparison due to radical structural differences between their tonal systems. In Cantonese, each syllable, regardless of its position, is specified for a lexical tone. On the contrary, in Wu, there is obligatory tone sandhi in prosodic domains larger than a syllable. In the case of KS, not all syllables in a sandhi domain are specified for tone. Regarding the density of tone specification above the syllable level, HKC and KS represent two opposite ends of the spectrum within the Sinitic languages, thus providing an ideal testing ground for whether the effect of L1 tonality on intonation learning is further modulated by the typological characteristics of the L1 tone system.

2. METHODS

2.1. Experimental design

We adapted the design of the intonation learning experiment in [1] and kept the audio stimuli identical to those published in [1]. Each stimulus is a string of fifteen segmentally identical syllables (/mi/), with an intonation contour superimposed on it. The intonation contours are random distortions of three prototypical pitch contours: "flat" (a monotone), "final fall" (a pitch drop over the last four syllables), and multifeature (hereafter "M"; two peaks at the fourth and eleventh syllable respectively). The distortion level is set to be one, three, or five semitones. An additional type of exemplar is "M" distractors, which are manipulations of the "M" prototype with 1-semitone distortion, such that the first peak, the second peak, or the valley between the two peaks is eliminated; they are referred to as "late peak", "early peak" and "hat" respectively. For details of how the stimuli were constructed, readers are referred to [1].

In the learning task, participants were first exposed to 36 exemplars (generated from low-level distortions of the three prototypical contours). Each intonation category was associated with one of three aliens represented by a name and a picture. In the next phase, different from the original design in [1], participants heard the same training exemplars but had to identify which of the three aliens each exemplar corresponds to. Then, regardless of the correctness of their response, they received feedback with the correct alien displayed and the exemplar replayed. This feedback phase was one novelty of our experiment and enabled us to know the effect of the training before the final test phase. In the test phase, participants were presented with a total of 72 exemplars, which consists of the training exemplars, prototype exemplars, novel exemplars with three levels of distortion, and "M" distractors. In this phase, participants had to either pick one of the three aliens, or choose "None of the above" if they judge the exemplar as not belonging to the three categories.

Another novelty of the current study is that two tasks were added to account for possible extralinguistic effects on intonation learning: an n-back task and a musical background questionnaire developed from Edinburgh Lifetime Musical Experience Questionnaire [8]. They served to examine the participants' working memory capacity [1,9,10] and indirectly estimate their pitch aptitude [11,12,13] respectively. Also, a linguistic background questionnaire developed from LEAP-Questionnaire [14] and the Lexical Test for Advanced Learners of English (LexTALE) [15] were included to serve as indicators of the participants' language background and proficiency in English, which is an additional language of all participants.

The experiment was administered online using the *Gorilla Experiment Builder* (www.gorilla.sc). The working language of the experiment was Chinese except that LexTALE was administered in English.

2.2. Participants

A total of 62 participants completed the experiment (age: 21–32), including 31 HKC speakers (16F, 14M, 1 non-binary gender; age: $M=23.3$, $SD=1.32$) and 31 KS speakers (16F, 15M; age: $M=26.3$, $SD=1.47$). Participants were recruited through online social media platforms, and they received monetary compensation for taking part in the study.

2.3. Analysis

The responses in the intonation learning task were transformed into three types, *success*, *confusion* and *rejection*, for assessment of the participants' performance. *Confusion* corresponds to wrong identification of the intonation categories, while *rejection* includes all the instances of the "None of the above" option chosen. The full dataset included the data collected in this experiment and the data on English native adult speakers published by [1]. Logistic regression and ordinary regression were fitted for the responses and the reaction time, respectively, with the R package *lme4* [16]. The baseline logit model contained a by-subject random intercept and by-speaker random slopes for distortion

level and intonation category. The following fixed effects were included if model fit was improved: distortion level (*1-semitone, 3-semitone, 5-semitone*), intonation category (*flat, final fall, M*) and language group (*English, HKC, KS*). The other effects, including gender, working memory, musical experience, English proficiency, and exposure to English (living in English-speaking countries for at least one year), were only tested on the data collected for this experiment, given that they were not conducted in [1]. Post-hoc comparison was performed where necessary with Holm-Bonferroni correction.

3. RESULTS

3.1 Learner profile

Regarding the participants in this experiment only, no effect was found for gender, musical experience, and exposure to English on their performance based on their success rates (*success* vs. the rest). Working memory showed a significantly positive effect (p = 0.007) for categorizing low-level (1-semitone) distortions. English proficiency showed a marginally significant effect (p = 0.021) in that the participants with higher English proficiency were less likely to correctly identify 3-semitone distortions.

3.2 Categorization performance

Two measures were specifically analysed to compare the participants' performance between language groups. "M" distractors were excluded for this part of analysis.

First, the success rate (*success* vs. the rest) was analysed to directly compare the listeners' ability to identify an exemplar for the correct intonation category rather than make a mistake or judge the exemplar as not belonging to the three learned intonation categories. Both distortion level and category, as well as their interaction, significantly improved the logit model fit. However, including the language group did not improve the model fit, nor did its two-way interaction with distortion level or category respectively.

Then the rejection rate (*rejection* vs. the rest) was analysed as an indirect measure of a listener's ability to attend to multiple features. It has been argued that the more features a listener can attend to at a time, the more likely they can distinguish exemplars with decreasing perturbation distance from the training items [1]. This may be reflected by them judging the exemplar as not belonging to the three learned categories. Distortion level, intonation category and their interactions all significantly improved the logit model fit, but language group and its interactions again did not.

Therefore, the learning results were comparable across the two language groups. As shown in Figure 1, across intonation categories, exemplars were less likely to be correctly identified and more likely to be rejected as the distortions become larger (p < 0.001 within each category between distortion levels). Within low-level (1-semitone) distortions, "final fall" was more likely to be correctly identified than the other categories (p < 0.001), while there was no significant difference between "M" and "flat" (p = 0.250). "Flat" was more likely to be rejected than the other categories (p < 0.001), while there was no significant difference between "final fall" and "M" (p = 0.058). Within high-level (3-semitone or 5-semitone) distortions, there was no significant difference between categories regarding both success and rejection (p > 0.069), except that in 5-semitone distortions, "M" was more likely to be rejected than "final fall" (p = 0.016).

For low-level distortions of "M" and "M" distractors only, the rejection rate was analysed as it reflects the participants' tolerance of the absence of some of the necessary features. Again, language group did not improve the model fit. Across language groups, all three types of distractors were more likely to be rejected than "M" (p < 0.001), among which "late peak" was the least likely to be rejected (p < 0.001), and "hat" the most (p = 0.002 between "hat" and "early peak").

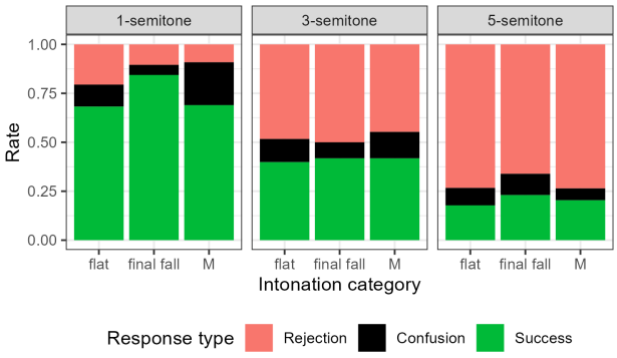


Figure 1: The success, confusion, and rejection rates by intonation category and distortion level.

3.2 Reaction time

For reaction time (RT), the predictor variables response type, distortion level, intonation category, language group, and their interactions were tested in sequence, and a three-way interaction model was first established between the first three variables. Including language group and its interaction with intonation category improved the model fit, which

was also significant in the final model (language: $p < 0.011$; language × category: $p < 0.001$). Post-hoc comparison showed that it was within the "flat" category that the English listeners showed a longer RT than both HKC ($p < 0.001$) and KS ($p = 0.001$) listeners, and within the "M" category that HKC listeners showed a shorter RT than the English listeners ($p = 0.001$). Given that there are up to three-way interactions, raw RTs are visualised in decile plots (mean RT within each decile, see Figure 2) for each intonation category between language groups.

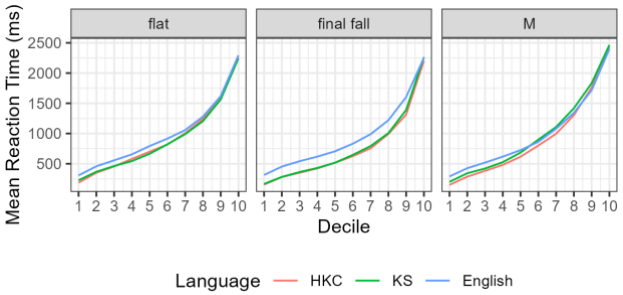


Figure 2: Decile plots of reaction time.

4. DISCUSSION & CONCLUSIONS

The categorization performance and RT patterns presented above have offered complementary perspectives on how tonal language speakers categorize intonation contours compared with non-tonal language speakers.

First, the Sinitic language speakers' performance shows that they have comparable categorization ability to the English speakers, at least for the paradigm and intonation categories tested in this experiment. The same patterns between distortion levels and intonation categories were observed across the language groups: (i) exemplars with lower-level distortions were easier to learn and categorize; (ii) within low-level distortions, "final fall" was an easier category than "flat" and "M". As discussed earlier, it is reasonable to speculate that one's ability to categorize intonation patterns is modulated by the tonality in their L1. The performance results of this study do not corroborate this idea, nor the idea that tonal language speakers, who are believed to use higher temporal resolution for pitch, would find it harder to only attend to major intonational events. However, it is worth noting that all KS and HKC participants in this study were exposed to English since childhood, and some (36%) have lived in an English-speaking country for at least one year. It is possible that the considerable exposure to English has affected the way these tonal language speakers process intonation contours, resembling the behaviour of English native speakers.

Regarding the importance of different features within an intonation contour, the results concerning "M" distractors demonstrate a finality advantage: among the three distractor types, "late peak" is least likely to be rejected, and thus probably deemed more acceptable as a member of "M". This corroborates Kapatsinski and colleagues' [1] findings, for which they proposed the following explanation: since the end of an utterance is often the most informative, it is likely to be associated with pitch peaks; consequently, listeners pay more attention to the end of an intonation contour. Our results suggest that the final prominence effect may be universal across languages.

The conditional effect of English language proficiency is also intriguing. That speakers with higher proficiency perform worse than those with lower proficiency, crucially only in the 3-semitone distortion level but not in the 1-semitone level, suggests that greater English proficiency may reduce intonation category breadth, which means they are better at attending to multiple features.

Whereas there is no apparent difference between the performance of different language groups, their RT reveals some subtle differences. L1 background is shown to have an effect on RT in the categorization of some of the intonation categories: HKC and KS speakers process "final fall" faster than the English speakers, and HKC speakers process "M" faster than the English speakers. Given that this is the only difference between the two tonal language groups, it is unlikely a consequence of different experimental settings. One possible explanation for the Sinitic speakers' processing advantage (compared to English speakers) concerns the language-specific functional load of these intonation contours. For example, in both HKC and KS, dubitative questions can be constructed with sentence-final falling intonation, which potentially leads to a high functional load for similar contours. The higher functional load may, in turn, cause HKC and KS speakers to devote more attention to utterance-final contours. In contrast, in English, a falling intonation typically marks a declarative statement, which can be considered the 'default' setting.

To conclude, this study shows that while language background, specifically L1 tonality, does not necessarily modulate one's intonation categorization performance, it could nonetheless pose a processing advantage for certain intonational contours. Finally, to enable a clearer examination of the L1 tonality effect, it would be worthwhile to conduct further experiments with native speakers of tonal languages who do not have any non-tonal language experience. Future studies should also consider a wider range of intonation contours.

5. ACKNOWLEDGEMENTS

We wish to thank Kapatsinski and colleagues [1] for making available the stimuli and dataset of their study. This research is supported by an NWO VICI grant (vi.c.181.040) to Yiya Chen.

6. REFERENCES

[1] V. Kapatsinski, P. Olejarczuk, and M. A. Redford, "Perceptual Learning of Intonation Contour Categories in Adults and 9-to 11-Year-Old Children: Adults Are More Narrow-Minded," *Cognitive Science*, vol. 41, no. 2, pp. 383–415, 2017.

[2] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[3] C. Gussenhoven, The Phonology of Tone and Intonation. Cambridge: Cambridge University Press, 2004.

[4] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

[5] K. Ito and S. R. Speer, "Anticipatory effects of intonation: Eye movements during instructed visual search," *Journal of Memory and Language*, vol. 58, no. 2, pp. 541–573, 2008.

[6] W. F. L. Heeren, S. A. Bibyk, C. Gunlogson, and M. K. Tanenhaus, "Asking or telling – real-time processing of prosodically distinguished questions and statements," *Language and Speech*, vol. 58, no. 4, pp. 474–501, 2015.

[7] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech communication*, vol. 46, no. 3–4, pp. 220–251, 2005.

[8] J. A. Okely, I. J. Deary, and K. Overy, "The Edinburgh Lifetime Musical Experience Questionnaire (ELMEQ): Responses and non-musical correlates in the Lothian Birth Cohort 1936," *PloS one*, vol. 16, no. 7, p. e0254176, 2021.

[9] T. J. Laméris and B. Post, "The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers," *Second Language Research*, p. 02676583221090068, 2022.

[10] S. Goss, "Exploring variation in nonnative Japanese learners' perception of lexical pitch accent: The roles of processing resources and learning context," *Applied Psycholinguistics*, vol. 41, no. 1, pp. 25–49, 2020.

[11] A. Cooper and Y. Wang, "The influence of linguistic and musical experience on Cantonese word learning," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4756–4769, 2012.

[12] P. C. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Applied Psycholinguistics*, vol. 28, no. 4, pp. 565–585, 2007.

[13] J. A. Alexander, P. C. Wong, and A. R. Bradlow, "Lexical tone perception in musicians and non-musicians," in *Ninth european conference on speech communication and technology*, 2005.

[14] V. Marian, H. K. Blumenfeld, and M. Kaushanskaya, "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp.9 942–947, *2007*.

[15] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Soft.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

Marking prominence: Towards cue-based annotation of prosodic prominence

Alejna Brugos[1], Mara Breen[2], Stefanie Shattuck-Hufnagel[3], Nanette Veilleux[1], Jonathan Barnes[4]

[1]Simmons University, [2]Mount Holyoke College, [3]Massachusetts Institute of Technology, [4]Boston University
alejna.brugos@simmons.edu, mbreen@mtholyoke.edu, sshuf@mit.edu, veilleux@simmons.edu, jabarnes@bu.edu

ABSTRACT

Phrase-level prosodic prominence in American English is understood, in the AM tradition, to be marked by pitch accents. While such prominences are characterized via tonal labels in ToBI (e.g. H*), their cues are not exclusively in the pitch domain: timing, loudness and voice quality are known to contribute to prominence perception. All of these cues occur with a wide degree of variability in naturally produced speech, and this variation may be informative. In this study, we advance towards a system of explicit labelling of individual cues to prosodic structure, here focusing on phrase-level prominence. We examine correlations between the presence of a set of 6 cues to prominence (relating to segment duration, loudness, and non-modal phonation, in addition to f0) and pitch accent labels in a corpus of ToBI-labelled American English speech. Results suggest that tokens with more cues are more likely to receive a pitch accent label.

Keywords: prosodic prominence, ToBI, phonetic cues, prosodic annotation, cue integration

1. INTRODUCTION

Phonological categories, both segmental and prosodic, are realized with great variability in the signal. Evidence from the segmental domain suggests that speakers systematically control aspects of speech below the level of contrastive linguistic constituents, and that listeners are sensitive to this variation [1,2]. Likewise, in the study of prosodic boundary and prominence events, there is well-documented variation in the cues that signal a prosodic category (see [3] for overview).

ToBI (for **To**nes and **B**reak **I**ndices) [4,5] is a phonological prosodic annotation system in which labellers mark prosodic categories of prominence (pitch accents) and grouping (boundaries). In annotating prominences with pitch-accent labels, a labeller may be uncertain as to the tonal identity of the pitch accent (i.e. H or L tonal categories), or whether a word is prominent at all. We hypothesize that this uncertainty might be related to cue density, in the sense that prosodic events evidenced by more cues may be easier to categorize than events with fewer cues.

Additional motivation for labelling individual cues to prosodic categories comes from the fact that informative phonetic variability is lost when considering categorical labels alone. Further, when there is disagreement and uncertainty, it is beneficial to know which aspects of the signal influence the labeller's category assignment. [3] advocate for "the identification of individual cues to the contrastive prosodic elements of an utterance." As argued in [3, 6], an essential goal of prosodic transcription is to identify contrastive linguistic categories of utterances, and not merely the salient aspects of the surface form. Towards this goal, we explore explicitly labelling acoustic cues to prosodic categories. Using a small corpus of ToBI-labelled American English speech we compare candidate prominence cue annotations to independently annotated ToBI pitch-accent labels.

1.1. Cues to prosodic prominence

In spoken American English, prosodic prominences are marked by a number of cues, including f0 excursions, duration changes, loudness, and voice quality (See [7] for an overview of prominence cues in English, & [8] for discussion of prominence cross-linguistically.) [9] found significant duration correlates of prominence (marked with Rapid Prosody Transcription [3]), as well as weaker influences from filtered intensity and little or no correlation with F0. Although the ToBI system focuses on F0 markers of phrase-level prominence (e.g. [5]), [10] reports that loudness predicts prominence, but F0 lends little to its perception. Thus there is disagreement about which acoustic cues correlate with perceived prominence; systematic labelling of cue patterns in conjunction with perceived phonological prominence will contribute to resolving these issues.

Phonetic cues to prominence have been extensively discussed in the literature, along with the role of cues for prominence perception (e.g. [7, 11]) and automatic prominence detection (e.g. [12]). However, we are unaware of any system designed for explicit annotation of prominence cues. The current proposal, like [13], complements categorical prosodic annotation systems.

1.2 Annotation of prosodic categories

ToBI [4,5] is a system of prosodic annotation based largely on the autosegmental-metrical model of prosodic phonology [14], with which users identify and label phonological categories of prosodic events. Tonal events indicate prominence or grouping, categorized as either pitch accents or two levels of edge tone. ToBI annotations are tier-based labels time-aligned to the speech signal. Here we focus on the *tones* tier for labelling pitch-related events of prominence (pitch accents) and phrase-level boundaries (phrase accents and boundary tones), which are placed in conjunction with orthographic text in a *words* tier.

2. METHODS

In the current study, we compare a set of acoustic cues to prosodic prominence with independently labelled MAE (Mainstream American English) ToBI pitch accent labels. We hypothesize that where prominence cues converge, there will be greater likelihood of a perceived phrase-level prominence, operationalized by ToBI pitch accent labels. In cases where only a subset of cues are present, we predict labellers will exhibit more uncertainty and disagreement about pitch accents.

The corpus [15] used to test this prosodic prominence-related cue labelling is the same used in [13] and [16]. It consists of 8 files, comprising roughly 6 minutes of speech: 178 seconds professionally read and 181 seconds of spontaneous speech. The files were produced by 7 individual speakers, roughly balanced between male and female, containing a total of 1076 words (1483 syllables). All files were labelled with MAE ToBI conventions by four expert labellers [16].

The cue annotation used 6 labels described below, taking inspiration from the cue labelling of prosodic boundaries [13], disfluencies [17] , and stuttered speech [18]. Two labels (**proc** and **prv**) are based on timing cues, and are adapted from the disfluency and stuttering annotations [17, 18]. Whereas [13] used a single label (**pr**) to capture segmental lengthening, we here expand this label to better capture lengthening in terms of segment type and position within the syllable (vowel, **prv** vs onset consonant, **proc**), taking on some of the granularity proposed in disfluency annotations of [19] and capturing that prominence-related lengthening primarily affects durations of syllable onsets and nuclei [20]. A new label is tested relating to amplitude (**ampi**), standing in for loudness. Two labels relating to voice quality are used, including irregular pitch periods of syllable beginning with a vowel (**gi**, used similarly to [13]) and a new label intended to capture other amodal phonation (**vq**) that may contribute to perceived prominence, such as localized creakiness or breathiness. All of these are labelled specifically when they are heard in conjunction with a prominence, and when their presence is interpreted as contributing to that prominence perception:

proc: Prolongation of the onset consonant
prv: Prolongation of a vowel
f0: A local f0 event (e.g. peak or valley) on or around a stressed syllable of a word
ampi: A local increase in vowel amplitude
gi: Irregular pitch periods at the beginning a vowel (or other sonorant consonant)
vq: Other amodal voice quality of a vowel

An expert prosodic labeller independently labelled the eight files for both perceived prosodic structure and cues. Labels for phrasal prominence and boundary locations were adapted from PoLaR Prosodic Structure tier labels [21], similar to RPT labels [9]. Tokens perceived as clearly prominent were labelled with a star (*); cases where the labeller was uncertain about a prominence were labelled with a star and a question mark (*?). For each perceived prominence (whether * or *?), the labeller chose from among the 6 proposed prominence cue labels.

Cue labels were placed on a point tier, time-aligned to be within the associated vowel's interval. Labels were used in sequences delimited by a period. Each syllable could hypothetically be labelled with between 0 and 6 cues from this set (e.g. no prominence labels, or only **proc** or **proc.prv.ampi.f0.vq**), but tokens marked with all 6 cues were not expected, as **gi** and **proc** are typically mutually exclusive, as **gi** only rarely occurs with sonorant consonants). All files were annotated

Cue	proc	prv	ampi	f0	gi	vq
Count	373	400	392	368	40	47

Table 1. Number of times each acoustic cue was labelled in the ToBI-annotated corpus.

using Praat TextGrids [22] and labels were automatically extracted for analysis. Of the 1483 syllables in the data set, 591 syllables were labelled as potentially prominent by the cue labeller, with 1 or more prominence cues. Each individual cue appeared between 40 and 400 times (Table 1). Because the number of tokens labelled with the two voice quality cues were relatively few, and no tokens were labelled as both gi and vq, these cues were collapsed into "vq" in the analyses.

3. ANALYSIS & RESULTS

Table 2 shows correspondence between the number of cues for a given token and pitch accent score, which is the number of the four ToBI labellers who indicated a pitch accent label (whether with tones, or the uncertainty marker). Tokens associated with an increased number of cues tend to be labelled as pitch-accented by more labellers.

	Number of labellers indicating * or *?				
	0	1	2	3	4
# of Cues					
0	749	84	19	31	9
1	19	6	16	17	30
2	14	15	26	25	84
3	7	6	9	26	114
4	1	2	2	18	144
5	0	0	0	1	9

Table 2: Number of labellers indicating a pitch accent by number of cues labelled.

We used linear mixed effects regression to determine whether prominence ratings (i.e., ToBI pitch accent labels) increased with the number of acoustic cues. We assigned each word the average prominence rating of the four labellers: any pitch accent label category was coded as 1; an uncertain pitch accent (***?**) was coded as 0.5; no pitch accent was coded as 0. These scores were averaged for each token. Number of cues was a significant predictor of average prominence rating, B = 0.24, SE = 0.004, t = 57.61, p < .01 (Figure 1).

We used linear mixed effect regression to predict the average prominence rating of each token from the five cues. Each cue was a fixed effect in the model; speaker was included as a random effect. The fixed effects are shown in Table 3 and Figure 2.

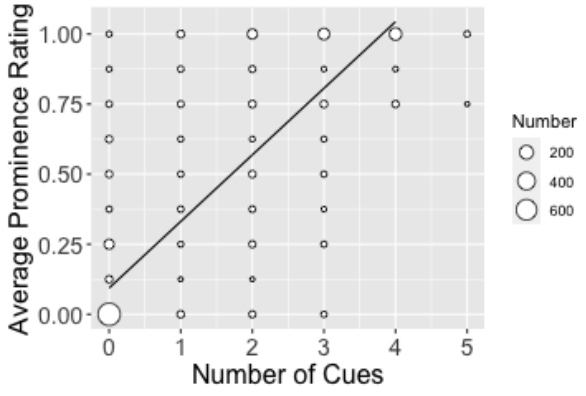


Fig. 1. Average prominence rating by number of cues per token. Dot size indicates number of tokens. For clarity of display, cases with only one token in a cell are excluded.

	Estimate	SE	t	p-value
(Intercept)	0.68	0.02	38.19	0.00
proc	0.26	0.02	13.25	0.00
prv	0.30	0.02	16.27	0.00
f0	0.28	0.02	13.93	0.00
ampi	0.13	0.02	6.83	0.00
vq	0.20	0.03	7.38	0.00

Table 3. Fixed effects in the model predicting average prominence rating from each cue.

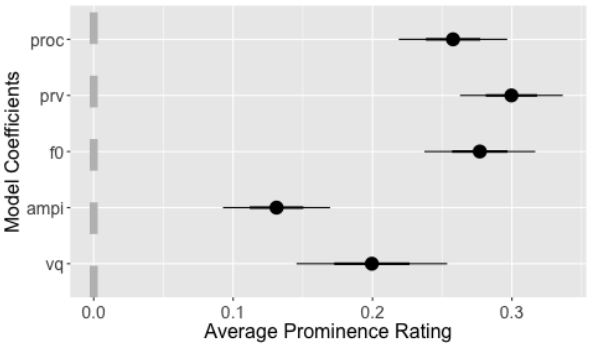


Figure 2. Coefficients for fixed effects in the model predicting average prominence rating from each cue.

The five cues (where **vq** includes the 6th cue, **gi**) independently increased the average prominence rating by ToBI labellers, demonstrating that each cue in the current study can signal pitch accents.

Finally, we assessed the relationship between number of acoustic cues and labeller uncertainty. If any of the ToBI labellers labelled the syllable with *?, we considered it uncertain. Table 4 shows the frequency of uncertainty by the number of cues labelled for each token. A chi-square analysis shows that the proportion of uncertain labels was higher for a small number of cues than for no cues, or many, $X^2(df = 5, N = 1483) = 100.32$, $p < .01$.

		Cue Number					
		0	1	2	3	4	5
Uncertain	0	782	54	101	115	120	7
	1	110	34	63	47	47	3
%		12.3	63.0	62.4	29.0	28.1	30.0

Table 4: Labeller uncertainty cases by cue number.

4. DISCUSSION

We found a strong correlation between the number of annotated acoustic cues and pitch accent labels from four independent ToBI labellers. Moreover, there was greater uncertainty in ToBI labels in cases with a small number of cues than in cases with many or no cues. However, a more informative

picture emerges by examining cases that countered predictions, namely: 1) where the cue labeller perceived a potential prominence and labelled acoustic cues to that prominence, but few or no ToBI labellers marked a pitch accent and 2) where the cue labeller did not mark a prominence, and labelled none of the cues, but one or more ToBI labellers marked a pitch accent. (See figure 3.)

We consider several possible explanations for such divergences. First, the method of capturing cues for only syllables where a prominence was perceived by the labeller entails that tokens not heard as prominent were not labelled for cues. Some cues may have been arguably present but did not meet a threshold of prominence for this listener, but may have been sufficient to cue prominence for another. Further, cues were labelled in a binary way: present or absent. For cases with high agreement of the presence of a pitch accent, but only one or two labelled cues, tokens may have evidenced those cues with a stronger magnitude (e.g. an extra long or loud vowel, or a more extreme f0 excursion) and therefore a single cue led to perceived prominence.

Why label cues at all and not just extract acoustic data? Like [7], we believe that the acoustic characteristics are always interpreted in the context of the prosodic structure: the listener "interprets most fine-grained phonetic variation only after having performed a parse of the signal into a coarser, more discrete sequence of categorical events [7]." Specifically, acoustic features are always context- and speaker-dependent, and must be interpreted with reference to this context. Not only are acoustic values (e.g. duration and amplitude) variable across discourses, they also vary by contextual factors such as speech rate, loudness and other paralinguistic factors. In addition, listeners perceive subtle acoustic variability across a range of listening conditions, including in noise where feature extraction is less reliable. Ultimately, prosodic categories are likely integrated perceptually [23], as suggested for category perception in a wide range of domains, both linguistic [24, 25] and auditory [26].

In this study we focus on the binary presence (or absence) of a given cue perceived by a trained labeller, but that these cues can and do also vary in their magnitude. The lengthening of onset consonants and vowels vary, as well as the size of f0 excursions of pitch accents. Further, directly annotating the cues to prosodic events may shed light on additional sources of ambiguity, such as when a labeller is uncertain as to whether a given cue indicates a prominence event (pitch accent) or a boundary event, or some combination of the two. Still another source of contextual ambiguity is in the case of neighboring strong syllables: labellers must decide which one is prominent–one, the other or both. (And all of these factors likely interact with signal-extrinsic factors [7,9]). Reliability of cues may vary; for example, speakers with creakier voices may not signal boundaries with irregular pitch periods, or such cues may be less salient.

A future direction of investigation would be to combine labelling of cues to prominence and boundary. For example, the voice quality cue **gi** can signal that a vowel-initial word is either phrase-initial or that is pitch accented, or in some cases both [27]. Further, the overlapping of cues of prominence and boundary may itself be informative, highlighting the prominence of certain structural positions [8]. Explicit labelling of prosodic cues can give insight to the relative strength of the instantiations of those categories, reflecting suggested prominence hierarchies [7].

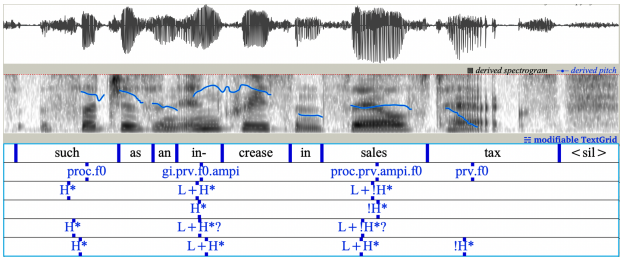


Figure 3: Labelling example, with two tokens ("in-" and "sales") where 4 cue labels were marked and all labellers labelled a pitch accent, and two tokens ("such" and "tax") where 2 cues were marked, but only a subset of labellers labelled a pitch accent. Tiers are syllables, cues, and pitch accent labels from each of four ToBI labellers.

5. CONCLUSIONS

Speakers control a set of acoustic cues with which they may signal prosodic structure. However, not all cues are used equally, and not all speakers use the cues in the same way. Listeners (and prosodic labellers) are sensitive to aspects of the speech signal that are cues to phonological categories, signaling prosodic structure events of phrasing and prominence. This study shows promising support for the meaningful relations between labels intended to reflect the presence of specific cues in the acoustic signal, and categorical/phonological labels of pitch accents as used in ToBI labelling.

6. ACKNOWLEDGEMENTS

This material is based partly upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, and 2042748. We are also grateful to several undergraduate research assistants, whose insights contributed to the development of these labelling methods: Kallie Dimaris, Laena Tieng, and Elaine Wang.

7. REFERENCES

[1] Cole, J., Shattuck-Hufnagel, S. 2018. Quantifying phonetic variation: landmark labelling of imitated utterances. In Cangemi et. al (eds) *Rethinking Reduction.* De Gruyter Mouton, 164-204.

[2] Turk, A., Shattuck-Hufnagel, S. 2014. Timing in talking: what is it used for, and how is it controlled?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130395.

[3] Cole, J., Shattuck-Hufnagel, S. 2016. New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1) 8, 1–29.

[4] Beckman, M., Ayers Elam, G. 1993/1997. Guidelines for ToBI Labelling (Version 3, March 1997) Copyright (1993) the Ohio State University Research Foundation.

[5] Beckman, M., Hirschberg, J., Shattuck Hufnagel, S. 2005. The Original ToBI System and the Evolution of the ToBI Framework. In S. A. Jun (ed.) *Prosodic Typology: the Phonology of Intonation and Phrasing*, 9–54.

[6] Frota, S. 2016. Surface and Structure: Transcribing Intonation within and across Languages. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1): 7. 1–19.

[7] Bishop, J., Kuo, G., Kim, B. 2020. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics*, 82, 100977.

[8] Grice, M., Kügler, F. 2021. Prosodic prominence–a cross-linguistic perspective. *Language and Speech* 64.2 (2021), 253-260.

[9] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), 425-452.

[10] Kochanski, G., Grabe, E., Coleman, J., Rosner, B. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2), 1038-1054.

[11] Ludusan, B., Wagner, P., Wlodarczak, M. 2021. Cue Interaction in the Perception of Prosodic Prominence: The Role of Voice Quality. *Proc. Interspeech 2021*, 1006-1010.

[12] Rosenberg, A. 2010. AutoBI-a tool for automatic ToBI annotation. *Proc Interspeech 2010*, 146-149.

[13] Brugos, A., Breen, M., Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. 2018. Cue-based annotation and analysis of prosodic boundary events. *Proc. Speech Prosody 2018*, 245-249.

[14] Pierrehumbert, J. 1980. The phonology and phonetics of English intonation (Unpublished Ph.D. thesis). Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

[15] Breen, M., Dilley, L. C., Kraemer, J., Gibson, E. 2012. Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8, 277–312.

[16] Dilley, L. C., Breen, M., Brown, M., Gibson, E. A. F. Rhythm and Pitch. LDC2018S04. Web Download. Philadelphia: Linguistic Data Consortium, 2018.

[17] Brugos, A., Langston, A., Shattuck-Hufnagel, S., Veilleux, N. 2019. A cue-based approach to prosodic disfluency annotation. *Proc. of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 3413-3417.

[18] Arbisi-Kelm, T. R. 2006. An intonational analysis of disfluency patterns in stuttering (Dissertation, UCLA).

[19] McDougall, K., Duckworth, M. 2017. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication* 9, 16-27.

[20] Turk, A. E., White, L. 1999. Structural influences on accentual lengthening in English. *Journal of Phonetics,* 27(2), 171-206.

[21] Ahn, B., Veilleux, N., Shattuck-Hufnagel, S., Brugos, A. 2021. PoLaR Annotation Guidelines (version 1.0). Available at https://osf.io/usbx5/. doi: 10.17605/OSF.IO/USBX5.

[22] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from www.praat.org

[23] Brugos, A.M., 2015. *The interaction of pitch and timing in the perception of prosodic grouping* (Doctoral dissertation, Boston University).

[24] Martin, A. E. 2016. Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, 7, 120.

[25] Brown, M., Tanenhaus, M. K., Dilley, L. 2021. Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science*, 13(2), 351-398.

[26] Bregman, A. S. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

[27] Dilley, L., Shattuck Hufnagel, S., Ostendorf, M. 1996. Glottalization of Word Initial Vowels as a Function of Prosodic Structure. *Journal of Phonetics* 24 (November 5, 1996): 423–444.

FOCUS PROSODY IN FIJIAN: IN-SITU FOCUS MARKING

Tsz Ching Mut[1], Candide Simard[2], Apolonia Tamata[2], Albert Lee[1]

[1]The Education University of Hong Kong, [2]University of the South Pacific
stcmut@gmail.com, candide.simard@usp.ac.fj, apolonia.tamata@usp.ac.fj, albertlee@eduhk.hk

ABSTRACT

This paper presents an acoustic analysis of focus prosody in Fijian - a verb-first language. We analysed the pitch (f_0), intensity, and duration data of the same sentence that vary in the position of narrow focus: (i) initial, (ii) medial, and (iii) final focus, compared to the neutral focus. Compared with neutral focus, we found that narrow focus led to a general elevation of the pitch f_0, greater intensity, and longer mean syllable duration. The only exception is for initial focus. Initial focus differs from the other two positions, with shorter mean syllable duration and no significant difference in intensity with neutral focus. We used sentences composed solely of juxtaposed NPs to elicit focus because in a verb-first language, focus is otherwise often marked by word order.

Keywords: speech prosody, focus prosody, Fijian

1. INTRODUCTION

Studying focus prosody of verb-initial languages can be difficult because non-prosodic focus markers such as fronting (i.e. moving the focussed item to the front) are often involved (see review in [1]). For example, in Samoan, fronting can be used to mark (contrastive) focus [2], alongside prosodic markers. This means that comparisons of prosodic cues among focus conditions are not based on otherwise identical utterances – a potential source of confounds. Possibly in part due to this challenge, Fijian focus prosody has yet to be empirically investigated with any systematic production experiment, making it an understudied topic in phonetics to this day.

Fijian is an Austronesian language spoken by about 400,000 as a first language [3] in Fiji. Its basic word order is often considered verb-object-subject (but note alternative accounts such as [3]). Currently, there is no published production study of Fijian focus prosody (except one pilot study [4]).

While not much is known about Fijian focus prosody, researchers have investigated a related language, verb-initial Samoan [2]. It was found that individual speakers used various focus-marking strategies. The initial phonological phrase was always the most prominent. In verb-agent-object sentences, the verb and agent were in the initial phrase. Speakers raised the pitch on the object in object focus, and lowered it in agent focus; although they did not do this consistently. No prosodic marking of focus on the agent was found.

To gain a better understanding of the understudied Fijian language, the current research project investigating the Fijian focus prosody contains two experiments. The first is a picture description task [4], that collects and analyses data on focus marking strategies in natural Fijian speech data... The second is designed, for better experimental control, to circumvent the fronting strategy commonly used to mark focus in verb initial languages, eliciting prosodic focus markers in situ with fronting suppressed. One possible strategy to achieve this is to avoid using natural sentences, e.g. using strings such as phone numbers[1] or, in the present study, a series of items. With such a paradigm, one can answer research questions such as: (i) Is narrow focus different from neutral focus? (ii) Are different focus locations marked differently? (iii) What acoustic cues (e.g. f_0, intensity, duration) are used to mark focus?

2. METHODS

2.1. Participants

Ten students from the University of South Pacific (5 female, 5 male) were recruited in Fiji. They are native speakers of Fijian, speaking English as an L2. They have no (history of) hearing or language impairment.

2.2. Materials

Fijian is a verb-first language where verbal suffixes are applied and focus is often marked by word order [5]. Thus, in the present study, a sentence composed of three adjacent noun phrases (NPs), i.e. *uvi, uto, dalo* 'yam, breadfruit, taro' was designed to elicit prosodic focus. The sentence has four focus conditions based on their positions, namely initial focus, medial focus, final focus (all of them are narrow focus), and neutral focus (Table 1). Regarding the neutral focus, as the order of items is identical to the order of the precursor question, it is in fact an echo answer to the question.

All four focus conditions of the sentence were elicited by a precursor question asked by the interviewer (a linguist and a Fijian native speaker). Altogether, we recorded 120 utterances (1 sentence *

4 focus conditions * 3 repetitions * 10 speakers). All utterances were retained for analysis.

Precursor question	Target sentence	Focus condition
uto, *uto, dalo?*	***Uvi***, *uto, dalo*	narrow (initial)
uvi, ***uvi***, *dalo?*	*Uvi,* ***uto***, *dalo*	narrow (medial)
uvi, uto, ***uto****?*	*Uvi, uto,* ***dalo***	narrow (final)
uvi, uto, dalo?	*Uvi, uto, dalo*	neutral

Table 1: Summary of stimuli used

2.3 Recording Procedure

This was a fruit sequence naming task. Participants were shown three pictures showing yam, breadfruit and taro and were instructed to put the pictures in a specific order. The interviewer then pretended to mishear the picture order and asked the precursor question starting with 'Did you say…?' followed by the sentence targeting different focus conditions (refer to "Precursor question" in table 1). For all focus conditions, the participants were instructed to respond by providing a complete answer, i.e. *sega/io* 'yes/no' plus the target sentence.

2.4 Annotation and measurements

We first chunked the raw sound data into individual utterances, then labelled them by syllable with PRAAT [6]. Vocal pulses were manually checked and rectified. After annotation, we obtained the time-normalised f_o, intensity, and syllable duration with ProsodyPro [7]. As the mean f_o of each participant is similar (84.27 - 98.65Hz), normalisation into semitones was not performed.

3. RESULTS

3.1 F0

We fitted linear mixed effects models to the f_0 data using *lmerTest()* [8]. Post-hoc comparisons were done using *emmeans()* [9]. Intensity (§3.2) and duration (§3.3) data were analysed using the same approach. The main effect of focus on f_0 was significant, $X^2(3) = 508.57$, $p < .001$. Post-hoc test shows that initial, medial and final focus had significantly higher f_0 than neutral focus ($p < .0001$). It means that, regardless of focus locations, a general elevation of f_0 is observed for all narrow focus conditions. The SS ANOVA [10] plot (Figure 1) shows that this difference was significant, and the difference lied in the last word *dalo* 'taro' (the red line indicates word boundaries).

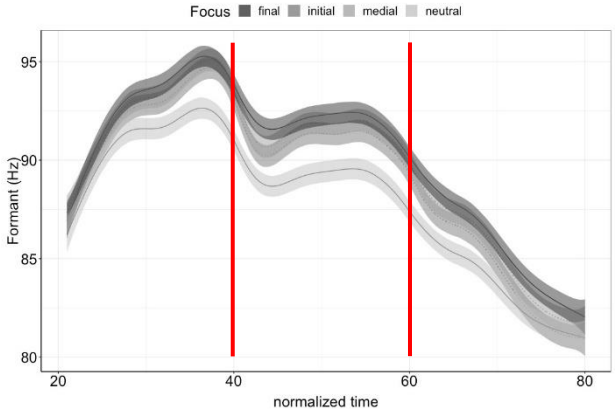


Figure 1: SS ANOVA comparing f_0 (Hz) of different focus conditions

3.2 Intensity

The main effect of focus on intensity was significant, $X^2(3) = 69.829$, $p < .001$ too. The Post-hoc test indicates that intensity of narrow focus (final and medial) is generally greater than neutral focus ($p < .0001$, see Figure 2). The SS ANOVA plot (Figure 2) shows the difference was significant, and the difference lied in the last word *dalo* 'taro'.

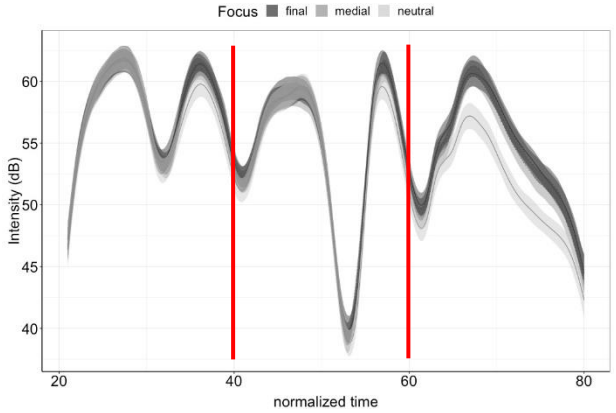


Figure 2: SS ANOVA comparing intensity (dB) of final, medial and neutral focus

3.3 Duration

While there are significant differences between f_0 and intensity of neutral focus and other narrow focuses, the difference in duration is not significant, $X^2(3) = 5.6281$, $p = 0.1312$.

In general, mean syllable duration is longer in narrow focus conditions, except for initial focus, in which syllable duration is shorter than in neutral focus (see Figure 3).

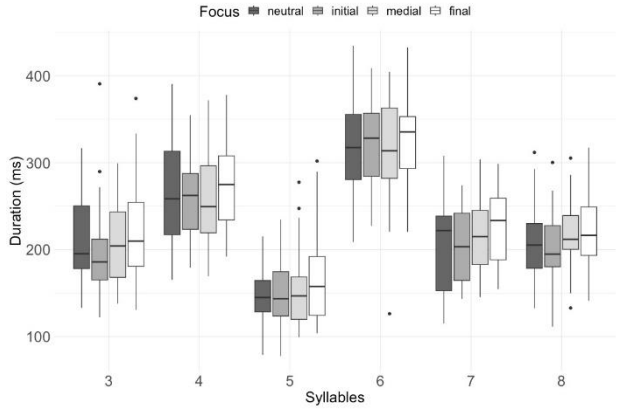


Figure 3: Syllable duration of different focus conditions

4. DISCUSSION

This study set out to identify prosodic focus marking strategies in Fijian. We compared narrow focus with neutral focus at different locations of a sentence, namely in - (i) initial, (ii) medial and (iii) final position. It appears that in situ prosodic focus in Fijian is mainly marked by a general elevation of f_0, as significant f_0 raising was found for all narrow focus conditions, compared to neutral focus. Significant increase in intensity was also observed in the narrow vs. neutral focus, though localised to the final word.

We found that focus conditions significantly affected prosodic focus markers, however, it is likely that Fijian speakers use different ways to mark focus in natural speech. In addition, as the strategies available to speakers are multiple, cross-speaker variability is nearly inevitable.

As a first systematic production study of Fijian focus prosody, this pilot study reports on the correlates of focus of scripted sentences, making comparisons between each condition possible. To gain a better understanding of Fijian focus prosody, further studies are needed, taking account of the various strategies available to speakers, ideally based on natural language samples from daily conversations.

5. CONCLUSION

To conclude, Fijian language seems not to mark prosodic focus making use of the cue of duration. The phonetic cues used to mark focus include a raised f_0 regardless of focus locations, and a raised intensity when the focus element is in the medial or final position.

[1] Phone numbers in Fiji are usually given in English and not in Fijian.

6. REFERENCES

[1] Kügler, F., Calhoun, S. 2020. Prosodic encoding of information structure: A typological perspective. In C. H. M. Gussenhoven & A. Chen (eds.), *The Oxford Handbook of Language Prosody*. Oxford University Press. 454–467

[2] Calhoun, S. 2015. The interaction of prosody and syntax in Samoan focus marking. *Lingua*, 165, 205–229.

[3] Geraghty, P. A. 2006. Fijian. In K. Brown (ed.), *Encyclopedia of Language and Linguistics* (2nd ed.). Elsevier, 465.

[4] Albert Lee, Candide Simard, Apolonia Tamata, Sun Jiaying, Tsz Ching Mut. Submitted. Focus prosody in Fijian: a pilot study. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*. Prague.

[5] Milner, G. B. 1989. On prosodic relations between Fijian bases and verbal suffixes. In J.H.C.S. Davidson (ed.) *South-East Asian Linguistics*. University of London, 59-88.

[6] Boersma, P. P. G., Weenink, D. J. M. 2012. Praat: Doing phonetics by computer.

[7] Xu, Y. 2013. ProsodyPro: a tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10.

[8] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.*, vol. 82, no. 13, 1–26.

[9] Lenth, R. V. 2020. Emmeans: Estimated marginal means, aka least-squares means. [Online]. Available: https://cran.r-project.org/package=emmeans.

[10] Davidson, L. S. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *J. Acoust. Soc. Am.* vol. 120, no. 1, 407–415.

7. ACKNOWLEDGEMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China awarded to AL (Project No. 18600621).

DETECTING ACCENTUAL PHRASE BOUNDARIES IN SEOUL KOREAN USING TONAL AND SEGMENTAL CUES

Seung Suk Lee

University of Massachusetts, Amherst, USA
seungsuklee@umass.edu

ABSTRACT

This paper investigates the informativity of tonal and segmental cues for detecting prosodic constituents boundaries in a prosodically unlabeled spontaneous speech corpus of Seoul Korean. The focus of this paper is Accentual Phrases (AP) that start with a Lenis obstruent. Previous work has found that listeners detect AP boundaries by attending to tonal cues (change in pitch from the previous syllable) and segmental cues (allophonic variation of Lenis). This paper reports that in spontaneous speech, while the segmental cues are more informative than the tonal cue for detecting AP boundaries, the tonal cue is informative for detecting the boundary of prosodic constituents higher than the AP.

Keywords: Seoul Korean Accentual Phrase, K-ToBI, Spontaneous speech corpus, Phonetics-prosody interface

1. INTRODUCTION

This paper investigates the informativity of tonal and segmental cues to the Accentual Phrases (AP) of Seoul Korean in a spontaneous speech corpus [1].

In the ToBI framework, prosodic constituent boundaries are associated with Break Indices (BI) [2]. In the autosegmental-metrical model of Seoul Korean intonational phonology, K-ToBI, there are 3 levels of BIs associated with 3 levels of prosodic units: the Phonological Word (PWd), the Accentual Phrase (AP), and the Intonational Phrase (IP) [3, 4, 5], see Table 1.[1] Previous work has shown that the listeners pay attention to both segmental and tonal cues to determine the presence or absence of these prosodic constituent boundaries [6, 7, 8, 9].

One way to assess the informativity of acoustic cues in perceiving these different boundary types is by formalizing perception as a categorization task where the boundary types are phonological categories [10]. This approach typically proceeds by first getting the data labeled with prosodic transcription to mark each syllable in the data with a BI value. Next, the acoustic properties of the labeled syllables are parameterized with relevant phonetic measurements, such as f0 or duration. Then, the effect of such cues in accounting for the differences between the BI levels is measured e.g., with a regression analysis.

BI level	Prosodic Position
3	IP-initial
2	AP-initial
1	PWd-initial ($[_{\text{PWd}}\sigma$)
0	PWd non-intial ($[_{\text{PWd}}...\sigma$)

Table 1: The BI levels and the associated prosodic positions. PWd non-initial includes both PWd medial and PWd final syllables.

Crucially, though, this approach requires the prerequisite step of first prosodically labeling the data, which becomes more expensive as the size gets larger, especially since it is known that the labelers also differ in their judgment of the type and the presence/absence of the prosodic constituent boundaries [11], which means it is ideal to have multiple labelers. This sets the entry barrier quite high to conduct such research and only a handful of datasets for only a few languages are publicly available (e.g., [12] for Japanese). In the case of Korean, there are few publicly available prosodically transcribed speech corpora, and none of them contains spontaneous speech (see [13] for the importance of investigating spontaneous speech).

For example, the data used in this study come from a spontaneous speech corpus of 40 native Seoul Korean speakers [1], but it is not prosodically transcribed. Consequently, while there are a few phonetic studies conducted on this corpus (e.g., [14]), none are on prosody.

On the other hand, one could also assess the informativity of cues without the prerequisite step of prosodic labels by investigating the distributions of the categories along a cue dimension and inspecting how separable the categories are along that dimension. The informativity of cues can be further explored with a clustering analysis, where each syllable in the dataset is considered a token in a phonetic space, constructed with acoustic cues hypothesized to be primary cues for AP detection

as the dimensions. The clustering analysis can assess the number of clusters that can best account for the variation in the data. If the clusters that emerge from the data can be inferred to be the boundary type categories based on where they lie in the phonetic space, it provides evidence that the cues that construct the space are informative in separating the categories.

This paper reports an initial exploratory analysis of the informativity of the acoustic cues to the AP boundaries of Seoul Korean in spontaneous speech, in this second sense—in the absence of prosodically labeled data.

2. BACKGROUND

Unlike typical prosodic constituents found in other languages, the AP in Korean is argued to not have final lengthening, but instead to be 'tonally marked' [3, 4]. The K-ToBI model suggests that the tonal marking is distinct at the AP junctures so that the listener can pay attention to tonal targets to notice whether or not there is an AP juncture in between the two syllables. According to [4], the initial tone of the AP can be 'L' or 'H'[2] but in this paper, we limit our discussion to the APs that start with a Lenis obstruent and therefore with an initial 'L' tone. APs typically end with an 'Ha' tone [4, 15]. This suggests that in a typical AP juncture, F0 falls from the pre-boundary syllable to the post-boundary syllable. However, F0 also falls from one syllable to another within an AP in longer APs [4]. Previous work has suggested that the gradient nature of the fall matters, as the fall across an AP juncture seems to be larger than the fall within an AP [16], and that listeners are sensitive to the gradient difference in the fall in perceiving AP boundaries [6, 7, 8].

In addition, [3] showed that the AP also serves as a domain that conditions the realization of segments. In particular, AP-initially, Lenis obstruents are acoustically realized as 'strong' (via 'domain-initial strengthening', [17, 18] among others); but optionally realized as voiced and lenited ('acoustically weak') AP-medially ('Lenis medial voicing', [19, 20] among others). The voicing of Lenis can be seen as a type of cross-linguistically attested domain-medial lenition [21], which can bear a demarcative function, as the lenited realization can signal the absence of a boundary or the continuation of a constituent [22]. In Korean, [9] showed that listeners are sensitive to this allophonic variation of Lenis stops and use it in determining the presence/absence of AP boundaries. However, [9] did not test how this segmental cue interacts with the tonal cue as her stimuli were artificially monotonized.

While experimental work has shown listeners are sensitive to these cues, it is yet to be investigated to what extent each is informative in separating the boundary type categories. Also, it is not yet known how the two types of cue interact since existing experimental work has tested the tonal [6, 7, 8] and segmental cues [9] independently, but not together. For listeners to be able to use these cues in spontaneous speech, these categories should be distinguishable in a phonetic space parameterized with these cues. Also, given the emphasis on the importance of tonal marking in the K-ToBI literature [4], I hypothesize that the tonal cue alone might be sufficient for separating the boundary type categories.

3. METHOD

To take a first step towards filling these gaps in the literature, this paper investigated a spontaneous speech corpus of Seoul Korean [1], choosing two speakers as a first pilot: one teenage female and a male in his forties since they were maximally different in terms of their age and gender. The data were also filtered to only include syllables with a Lenis stop onset (7393 out of 27607 syllables, 27%).

Following [16], it was assumed that the tonal marking for the AP is realized as a change in F0 over a disyllabic window. In order to parameterize this tonal cue, F0 was extracted for every 5 ms, from syllables with a Lenis onset and the syllable that came before it. The mean F0 over the preceding syllable was compared with the syllable with the Lenis onset, to determine whether the F0 was falling or rising over the two syllables. Then, the difference between the minimum and the maximum F0 values was taken so as to maximize the fall/rise. The difference was standardized as a fraction of range [16] for each given utterance[3]. This variable, named '$\text{MAX}\Delta_{F0}$', was negative when the F0 change was a fall, and positive when the change was a rise. The 'utterance initial' syllables were excluded since they did not have a previous syllable in an utterance, by definition. The utterance is segmented by a pause in the corpus [1], which indicates that the utterance initial syllables are always IP-initial syllables. However, as the IP boundary is not always preceded by a pause, this meant that some PWd-initial syllables could be both IP-initial and AP-initial. Considering all the possible tonal markings of AP junctures, including AP junctures that are also IP junctures, both rising and falling contours could

occur at an AP juncture, as the pitch resets when an IP starts [5]. As mentioned earlier, the size of the tonal fall is larger across AP boundaries, than within AP boundaries, we expect some syllables to have a more extreme negative value of MAXΔ_{F0} (AP initial) than others (AP non-initial).

The segmental realization of Lenis stops was parameterized using three measurements taken over the Lenis stop closure interval that were reported to be correlated with the lenition of stops: percentage of voiced interval [23], the difference between the maximum and minimum rate of change in intensity as in [21], and the closure duration, which was speech-rate normalized following [9]. Principal Components Analysis was used to combine them into a single variable for comparison with the tonal cue, following [24]. The first component, which accounted for 77% of the variance in the data, was named STRENGTH. A Lenis token that was strong had a positive value of STRENGTH, whereas a Lenis token that was realized as weak had a negative value.

The corpus [1] provided three levels of segmentation: phone, 'Eojeol', and utterance. An 'Eojeol' is an orthographically defined sequence of syllables, that is often loosely equated with the PWd in previous work including the K-ToBI labeling conventions [4], and I follow this assumption, i.e., Eojeol non-initial is equated with PWd non-initial ([$_{\text{PWd}}$...σ) and Eojeol-initial is assumed to be PWd-initial ([$_{\text{PWd}}\sigma$). While the corpus does not have AP labels, we can still make inferences about the distribution of AP-initial (some of which may also be IP-initial) and AP non-initial syllables in the $\langle$MAXΔ_{F0}, STRENGTH$\rangle$ acoustic space via the PWd labels with the following logic.

Assuming the prosodic constituents are strictly layered, we can assume that PWd non-initial syllables must be AP non-initial. But we do expect the PWd initial syllables to be a mix of AP initial and AP non-initial syllables, since there can be multiple PWds inside an AP. The PWd initial ones that are AP-non-initial should have an overlapping distribution with PWd non-initial syllables (which must be AP non-initial), but the PWd initial syllables that are AP-initial should be separable from the PWd non-initial ones along a cue dimension. If such separability is observed along a dimension, I call that dimension informative. We can also infer that the subset of PWd initial syllables not overlapping with PWd non-initial syllables is AP-initial.

In addition to the exploratory analysis presented above, a clustering analysis was performed on [$_{\text{PWd}}\sigma$ to further explore the systematicity of variation in the distribution [25, 26]. By the logic above, [$_{\text{PWd}}\sigma$ in the corpus can be partitioned into three sub-categories: IP-initial, AP-initial (but not IP-initial), and AP non-initial (see Table 1). To test whether these three sub-categories emerge from the distribution of [$_{\text{PWd}}\sigma$, a K-means analysis was performed by setting the value of K from 2 to an arbitrary number, where K is the number of clusters that the algorithm is supposed to find. Then for each K value, a silhouette analysis was performed, which is a commonly used metric to evaluate clustering analysis. The output of a silhouette analysis ranges from 1 to -1; where 1 indicates perfect separation, 0 indicates a complete overlap, and -1 indicates complete misrepresentation of the data. I hypothesized that the number of clusters that best accounts for the variation in the data to be 3 and expected the found clusters to match the three sub-categories listed above.

4. RESULTS

The marginal plot on the right edge of Fig. 1 shows that some syllables with Lenis onsets have a rising F0 from the previous syllable (positive MAXΔ_{F0}) and some have a falling F0 (negative MAXΔ_{F0}). However, it is clear that MAXΔ_{F0}alone does not distinguish AP-initial syllables from AP non-initial syllables as the [$_{\text{PWd}}\sigma$ (light gray) and [$_{\text{PWd}}$...σ (dark gray) distributions almost completely overlap.

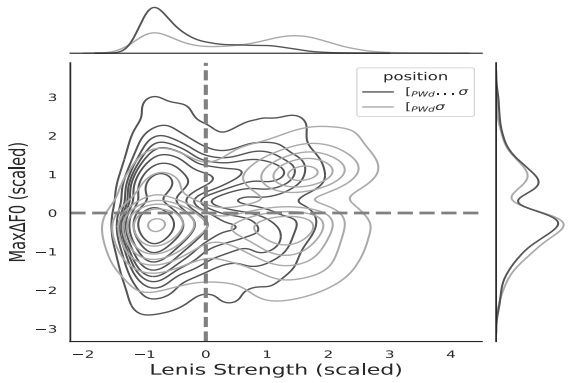


Figure 1: 2D density contour plot with marginal plots showing the distributions of STRENGTH and MAXΔ_{F0} on the top and right

On the contrary, the distributions of [$_{\text{PWd}}\sigma$ and [$_{\text{PWd}}$...σ over STRENGTH only partially overlapped (Fig. 1, top). In particular, while the [$_{\text{PWd}}\sigma$ showed two modes in the distribution, one in the positive region and the other in the negative region, [$_{\text{PWd}}$...σ had a unimodal distribution. It is also noteworthy that [$_{\text{PWd}}\sigma$s are distributed to the right

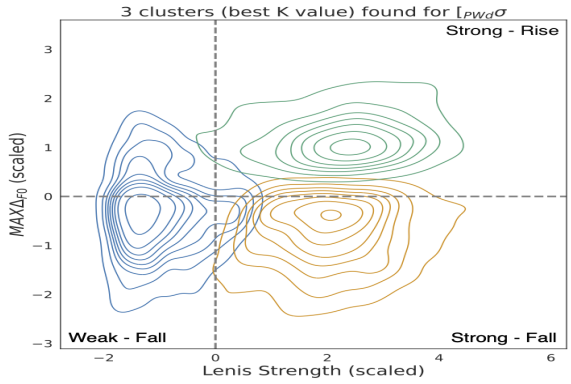


Figure 2: The three found sub-categories of [PWdσ, from the K-means clustering analysis

of [PWd...σs in the positive region of the STRENGTH dimension. This indicates that while [PWd...σ could be realized as voiceless (as the Lenis medial voicing is not obligatory, as previously reported [20]), they were still 'acoustically weaker' than [PWdσ, as reported in [18].

Overall, contrary to expectation, these findings suggest that the tonal cue is not in fact informative (in the sense operationalized in this paper) for separating the AP-initial syllables from the AP non-initial ones in a large and spontaneous dataset—since the distributions of [PWdσ and [PWd...σ almost completely overlapped. Instead, the results support the informativity of the segmental cue, indexing the allophony of the Lenis stop, for AP boundaries.

In addition to the exploratory analysis presented above, a clustering analysis was performed on [PWdσ to further explore the systematicity of variation in the distribution [25, 26]. As hypothesized, it was found that the number of clusters that best explain the variability in the data was 3 (silhouette score: 0.49). The three clusters found roughly occupied the first, third and fourth quadrants in Fig. 2, which would correspond to 'Strong-Rise', 'Weak-Fall' and 'Strong-Fall', respectively. Among the three, it can be inferred that the Weak-Fall cluster corresponds to [PWdσ that are AP non-initial, since it was expected that voiced/lenited Lenis must be in the AP non-initial position.

There is also evidence suggesting that the Strong-Rise cluster corresponds to IP initial syllables and the Strong-Fall cluster to (non IP initial) AP initial syllables. For Strong-Rise syllables, the top ten most frequent preceding syllables included sentence-final particles; for Strong-Fall, they included case markers, which are indicative of the right edge of an AP [4]. The two Strong clusters were separated along the MAXΔ_{F0} dimension but not along the STRENGTH dimension, which suggests that [PWdσ can be first distinguished into AP-initial or AP non-initial syllables along the STRENGTH dimension, but the Strong [PWdσ can be further separated into AP-initial and the higher constituent initial (e.g., IP) syllables, using the tonal cue. More analysis is required to investigate how AP initial and IP initial syllables can be further separated.

5. DISCUSSION AND CONCLUSION

This paper investigated the informativity of previously proposed tonal and segmental cues to AP boundaries in a corpus of Seoul Korean spontaneous speech. Even though the corpus was not prosodically labeled, I showed how the informativity of the acoustic cues for AP boundaries could nevertheless be investigated, via the distributions of [PWdσ and [PWd...σ, and with a clustering analysis. I found that the informativity differs between the tonal and the segmental cues in separating boundary type categories in Seoul Korean spontaneous speech. A clear limitation of this study is that only the syllables with Lenis onset were investigated. It remains to be tested whether other AP-conditioned segmental realizations provide similar informativity as shown in this study (e.g., AP-conditioned denasalization [9]).

Exploring the distributions of the categories in unlabeled data serves as an initial step toward understanding the learnability of the categories, as it resembles the challenge that human learners face, in the sense that they are not given the labels in the data either, cf. the distributional learning literature [27, 28]. Being able to find the expected clusters in the ⟨MAXΔ_{F0}, STRENGTH⟩ acoustic space could indicate that the AP boundary phonological contrast could be learned from the cues investigated.

6. ACKNOWLEDGEMENTS

I thank Kristine Yu and John Kingston for their valuable feedback.

7. REFERENCES

[1] W. Yun, K. Yoon, S. Park, J. Lee, S. Cho, D. Kang, K. Byun, H. Hahn, and J. Kim, "The Korean corpus of spontaneous speech," *Phonetics and Speech Sciences*, vol. 7, no. 2, pp. 103–109, 2015.

[2] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *ICSLP*, vol. 2, 1992, pp. 867–870.

[3] S.-A. Jun, "The accentual phrase in the Korean prosodic hierarchy," *Phonology*, vol. 15, no. 2, pp. 189–226, 1998.
[4] "K-ToBI (Korean ToBI) Labelling Conventions (version 3.1)." [Online]. Available: https://linguistics.ucla.edu/people/jun/ktobi/k-tobi.html
[5] S.-A. Jun, "Intonational phonology of Seoul Korean revisited," *Japanese-Korean Linguistics*, vol. 14, pp. 15–26, 2006.
[6] S. Kim and T. Cho, "The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean," *JASA*, vol. 125, no. 5, pp. 3373–3386, 2009.
[7] S. Kim, T. Cho, and J. M. McQueen, "Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language," *J. Phon.*, vol. 40, no. 3, pp. 443–452, 2012.
[8] A. Tremblay, T. Cho, S. Kim, and S. Shin, "Phonetic and phonological effects of tonal information in the segmentation of Korean speech: An artificial-language segmentation study," *Appl. Psycholinguist.*, vol. 40, no. 5, pp. 1221–1240, 2019.
[9] K. Yoo, "The production and perception of domain-initial strengthening in Seoul, Busan, and Ulsan Korean," Ph.D. dissertation, University of Cambridge, 2020.
[10] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *JASA*, vol. 91, no. 3, pp. 1707–1717, 1992.
[11] S.-A. Jun, S.-H. Lee, K. Kim, and Y.-J. Lee, "Labeler agreement in transcribing korean intonation with K-toBI," in *INTERSPEECH*, 2000, pp. 211–214.
[12] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. J. Venditti, "X-JToBI: an extended j-toBI for spontaneous speech," in *INTERSPEECH*, 2002.
[13] B. V. Tucker and M. Ernestus, "Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," *The Mental Lexicon*, vol. 11, no. 3, pp. 375–400, 2016.
[14] K. Yoon and S. Kim, "A comparative study on the male and female vowel formants of the Korean corpus of spontaneous speech," *Phonetics and Speech Sciences*, vol. 7, no. 2, pp. 131–138, 2015.
[15] S. Kim, "The role of prosodic phrasing in Korean word segmentation," Ph.D. dissertation, University of California, Los Angeles, 2004.
[16] J. Lee and H. Lee, "Korean Intonation Patterns from the Viewpoint of f0 Percentage Change," *Language and Speech Sciences*, vol. 5, no. 1, pp. 123–130, 2013.
[17] T. Cho and S.-A. Jun, "Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean," *UCLA working papers in phonetics*, pp. 57–70, 2000.
[18] T. Cho and P. A. Keating, "Articulatory and acoustic studies on domain-initial strengthening in Korean," *J. Phon.*, vol. 29, no. 2, pp. 155–190, 2001.
[19] D. J. Silva, "The phonetics and phonology of stop lenition in Korean," Ph.D. dissertation, Cornell University, 1992.
[20] S.-A. Jun, "Lenis Stop Voicing Rule," *Theoretical Issues in Korean Linguistics*, p. 101, 1994.
[21] J. Kingston, "Lenition," in *3rd Conference on Laboratory Approaches to Spanish Phonology*. Cascadilla Proceedings Project, 2008, pp. 1–31.
[22] J. Katz and M. Fricke, "Auditory disruption improves word segmentation: A functional basis for lenition phenomena," *Glossa: a journal of general linguistics*, vol. 3, no. 1, p. 38, 2018.
[23] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *J. Phon.*, vol. 54, pp. 35–50, 2016.
[24] C. Dalcher, "Statistical methods for quantitative analysis of multiple lenition components," in *ICPhS*, 2007.
[25] J. Krivokapic, "The planning, production, and perception of prosodic structure," Ph.D. dissertation, University of Southern California, 2007.
[26] E. Chodroff and J. Cole, "Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English," in *Proceedings of INTERSPEECH 2019*. International Speech Communication Association, 2019, pp. 1966–1970.
[27] N. Feldman, T. Griffiths, and J. Morgan, "Learning phonetic categories by learning a lexicon," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 31, no. 31, 2009.
[28] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, pp. B101–B111, 2002.

[1] There is also an intermediate phrase in the most recent version of K-ToBI [5], but we limit our discussion to the AP in this paper.
[2] The initial tone of an AP is 'H' if the initial segment of an AP is an Aspirated or a Fortis obstruent [3]
[3] Different methods of standardization and measuring not the size of the change but the slope of the fall/rise were tried, but all other measurements patterned similarly.

ARTICULATORY MANIFESTATION OF FOCUS-INDUCED PROMINENCE IN SEOUL KOREAN

Jiyoung Jang & Argyro Katsika

University of California, Santa Barbara
jiyoung@ucsb.edu, argyro@ucsb.edu

ABSTRACT

Speech units under phrase-level prominence are known to be produced with longer, larger, and faster constriction gestures. Evidence, however, comes mainly from head-prominence languages. The present electromagnetic articulography study investigates the effect of focus-induced prominence on articulatory manifestation in Seoul Korean, an edge-prominence language that marks phrase-level prominence via Accentual Phrases (APs) and also lacks word-level prosody. Special interest is given in examining the scope of the effect as well as its interaction with boundary marking at the level of the Intonational Phrase (IP). Results from seven native Seoul Korean speakers suggest that constriction gestures are longer, larger, and faster under focus. The scope of the effect extends away from the initial syllable of the focused AP reaching up to its third syllable. An interaction between focus and final IP boundary is detected, indicating an intricate relationship between information structure and prosodic structure, discussed within the framework of Articulatory Phonology.

Keywords: focus-induced prominence, articulation, prosody, Seoul Korean, Articulatory Phonology

1. INTRODUCTION

Prosodic structure serves two important linguistic functions in speech: phrasing, which groups linguistic units into larger cognitive constituents, and prominence, which marks constituents that are important for either rhythmic (e.g., via stress) or conceptual reasons (e.g., via pitch accent). Work on the phonetic realization of these prosodic landmarks has been provided accumulated evidence suggesting that segments are produced with longer, larger, and faster constriction gestures when accented and with longer, less overlapped and slower constriction gestures when phrase-final (e.g., [1, 2]). However, these results come mainly from head-prominence languages, such as English, in which phrase-level prominence is usually marked by placing a pitch accent on the stressed syllable of the prominent word (cf. [3]).

Here, we turn to Seoul Korean with the goal to examine the articulatory correlates of focus-induced prominence in an edge-prominence language. Seoul Korean does not employ word prosody (i.e., lexical stress, lexical tone or lexical pitch accent), and phrase-level prominence is marked by the means of Accentual Phrases (APs). AP serves as the basic intonational unit and is marked by a particular pitch contour [4, 5]. AP's proposed underlying tonal pattern is THLH, where the realization of the initial tone (T) tends to depend on the laryngeal configuration of the AP-initial segment [4 - 7]. Focus-induced prominence in Korean employs this AP level, since the focused word consistently starts an AP (or a higher phrase), and any following AP boundaries up to the end of the Intonational Phrase (IP) often undergo elimination, or possibly attenuation, referred to as dephrasing [6, 8].

Limited work on the correlates of prominence in Korean reports longer, larger, and slightly faster vocalic movements under focus in Korean [9] (see [10] for a review). However, the scope of the effect as well as its interaction with marking higher-level IP boundaries is unclear. The latter is of interest, because in Korean both phrasing and prominence are marked via prosodic boundaries. In particular, we examine the interaction between focus marking (left-edge AP boundary) and IP marking (right-edge IP boundary), with a goal to broaden our understanding on the interplay between information structure and prosodic structure. Another interesting research question involves the phenomenon of dephrasing that arises from the effect of focus-induced prominence; the current study also examines whether the patterns of dephrasing differ with distance of focus from the right edge of the IP.

We use Electromagnetic Articulography (EMA) to assess the effect of contrastive focus on the duration, displacement, and velocity of the consonant gestures of the focused words. We expect gestures under focus to be longer, larger, and faster as compared to that of unfocused ones, in line with previous findings [9]. In addition to these articulatory manifestations, we investigate the relationship between kinematic parameters. Specifically, we test whether duration increases as stiffness decreases, and whether displacement increases proportionally to velocity, as it has been shown to be the case in previous research primarily on head-prominence languages [11, 12].

Based on findings of the extended scope of prominence effect beyond stressed syllable (head) in head-prominence languages [13], we expect the scope of the focus-induced prominence to span more than the

initial syllable of the AP. Nonetheless, to which syllable the effect extends to as well as whether it would interact with the effect spanning from the IP boundary marking from the right-edge needs to be tested.

2. METHODS

2.1. Participants and experimental procedure

Seven native Seoul Korean speakers (5F, 2M; Mean age = 24.5, Age range = 21-28) participated in the study. The speakers were naïve as to the purpose of the study and had no reported speech, hearing, or vision problems. They received financial compensation for their participation.

Prior to the experiment, the participants went through a 15-minute training session in order to become acquainted with the speech materials and their presentation. Kinematic data were collected using an AG501 3D electromagnetic articulograph (Carstens Medizinelektronik). Ten receiver coils were attached to the tongue dorsum (two sensors), tongue tip, upper and lower incisors, upper and lower lips, left and right ears, and nose. Audio recordings were performed simultaneously to the kinematic recordings by the means of a Sennheiser shotgun microphone positioned approximately one foot away from the participant's mouth and set at a sampling rate of 16 kHz. Stimuli were presented on a computer screen placed roughly three feet away from the participant. The experimental session began with a set of simple tasks (e.g., reading lists of short words), the goal of which was to familiarize the participants with the experimental procedure and apparatus. Next, participants were asked to read sets of prompt-target sentences (see Section 2.2). For each set, the prompt sentence was shown in green font and appeared first. The target sentence, shown in blue font, appeared one second after the prompt sentence, and was read aloud. Both prompt and target sentences were presented in regular font, i.e., non-bolded and non-underlined.

2.2. Stimuli

The effect of focus was examined across the test interval /nɛ.maŋ.mi.nam/, a compound word that means *'a handsome guy from Nemang'*. The participants were asked to imagine a situation where they were preparing a play and the compound word was presented as a role in the play. The target word was framed to receive contrastive focus in a set of test stimuli, as shown in Table 1a. Prompt sentences were designed to help appropriate focus placement. For example, for the target sentence (a) in Table 1, the prompt sentence was *'It's not the handsome guy from Nowon (that Uncle Minam is playing)'*. Focused test words in the stimuli were compared to two sets of control stimuli that included the test word in unfocused condition; in one set, focus was placed on the second word of the sentence (Table 1b), and in the other, focus was on the first word of the sentence (Table 1c). We refer to these conditions as Focus-proximate and Focus-distant, respectively. Having these two different unfocused conditions allows examining the degree of dephrasing.

To assess the interaction between focus marking and IP-boundary, test stimuli were further manipulated so that test words either IP-final or IP-medial (compare stimuli (a) and (d) in Table 1).

The combination of Focus Type (Focused, Focus-proximate, Focus-distant) and Within-IP Position (IP-final, IP-medial) yielded six conditions in total, which occurred in an experimental block along with stimuli for other experiments. Each block was differently randomized and repeated eight times. Note that for one speaker, five repetitions were collected due to interruption of the experimental session for technical reasons. In total, 318 tokens were included in the analyses reported here. The acquired data were checked for their prosodic rendition, i.e., focus placement and prosodic boundaries, using K-ToBI [14].

Focus	***Target sentence (IP-final)***
(a) Focused	[pimilpu minami gomopuga **nɛmaŋminam**]? [sʌntækhangʌja]? 'Uncle Minam of the secrecy club is **the handsome guy from Nemang**? Is it decided?'
(b) Focus-proximate	[pimilpu **minami** gomopuga nɛmaŋminam]? [sʌntækhangʌja]? 'Uncle **Minam** of the secrecy club is the handsome guy from Nemang? Is it decided?'
(c) Focus-distant	[**pimilpu** minami gomopuga nɛmaŋminam]? sʌntækhangʌja]? 'Uncle Minam of **the secrecy club** is the handsome guy from Nemang? Is it decided?'
Focus	***Target sentence (IP-medial)***
(d) Focused	[pimilpu minami gomopuga **nɛmaŋ**minam sʌntækhangʌja]? 'Uncle Minam of the secrecy club chose **the handsome guy from Nemang**?'

Table 1: Example stimuli presented by Focus Type (Focused, Focus-proximate, Focus-distant). The IP-final and IP-medial counterparts are given for the Focused condition.

2.3. Data analysis

Consonant (C) gestures in the interval /nɛ.maŋ.mi.nam/ were analyzed, except the coda of the second syllable. Coda /ŋ/ was excluded from the analysis because of its degree of blending with the neighboring vowels. We will refer to the measured consonants as C1, C2, C3, C4, and C5. All of these were onsets, except for C5 which is the coda of the fourth syllable. Kinematic labeling was performed using semi-automatic custom software in Matlab (Mark Tiede, Haskins Laboratories).

For labelling, tongue tip vertical displacement trajectory was used for /n/ and lip aperture trajectory was used for /m/. The labelling procedure detected the following kinematic timepoints in each C gesture on the basis of velocity criteria: onset, time of peak

velocity, target, constriction maximum, release, and offset (Figure 1). The following kinematic measures were calculated based on these timepoints: formation duration (interval between onset and release), displacement (spatial difference between max and onset), and formation's peak velocity (velocity value at point (c) in Figure 1) of each test C gesture.

The retrieved data were analyzed by linear mixed effects analysis using lme4 [15] package in R [16]. To test the effect of focus type and its interaction with Within-IP Position, each C gesture was examined for the dependent variables of formation duration, displacement, and peak velocity. Fixed effects of Focus type (Focused, Focus-proximate, Focus-distant) and Within-IP Position (IP-final, IP-medial) were included. Random effect of speaker was added. To examine the relationship between kinematic parameters [11, 12], two models were built. One model was with formation duration as dependent variable, and stiffness (as calculated by peak velocity over displacement [17]) and Consonant (C1 to C5) as fixed structure. Consonant was added to see whether there was difference in the relationship between the parameters as a function of consonant position in the word or phrase. The other model had displacement as dependent variable, and peak velocity and Consonant as fixed structure. Random effect of speaker was added in the models. In case of significant effects, pair-wise comparisons were assessed by the *emmeans* package [18] with Holm adjustment.

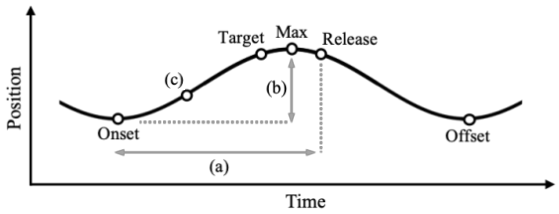


Figure 1: Kinematic measurements of constriction gesture: (a) formation duration (in ms), (b) displacement (in mm), and (c) peak velocity (in cm/sec).

3. RESULTS

3.1. Effect of focus type and its interaction with phrase-final boundary

Table 2 lists the significant main and interaction effects of Focus Type and Within-IP Position, and Figure 2 visualizes the results. Main effect of Focus Type was found in the C gestures of the three first syllables of the test word (C1, C2, C3; Table 2, lines 1-9). In C1, C gestures were longer and larger in the Focused condition than in either of the Unfocused conditions (p<0.001 for all comparisons). Gestures were faster when focused as opposed to when focus fell on the proximate word (p<0.05), but not when it fell on the distant word (p>0.09). No significant difference was found between the two Unfocused conditions in any of the measurements.

Syll	*DV*	*Focus*	*Boundary*
C1	(1) duration	*F*(2)=98.8***	*n.s.*
	(2) displacement	*F*(2)=23.1***	*n.s.*
	(3) peak velocity	*F*(2)=6.6*	*n.s.*
C2	(4) duration	*F*(2)=51.0***	*n.s.*
	(5) displacement	*F*(2)=110.3***	*n.s.*
	(6) peak velocity	*F*(2)=62.7***	*n.s.*
C3	(7) duration	n.s.	*n.s.*
	(8) displacement	*F*(2)=164.5***	*n.s.*
	(9) peak velocity	*F*(2)=114.7***	*n.s.*
C4	(10) duration	*n.s.*	*F*(1)=19.5***
	(11) displacement	*n.s.*	*F*(1)=9.6**
	(12) peak velocity	*n.s.*	*F*(1)=9.3**
C5	(13) duration	*n.s.*	*F*(1)=331.6***
	(14) displacement	*n.s.*	*F*(1)=375.4***
	(15) peak velocity	*n.s.*	*F*(1)=66.6***

Table 2: Summary of main effects ('***' = p<0.001, '**' = p<0.01, '*' = p<0.05, 'n.s.' = p>0.09).

In C2, C gestures were longer, larger and faster when focused as opposed to unfocused (p<0.001 for all comparisons).

In C3, formation duration did not show systematic difference as a function of Focus Type. However, both displacement and peak velocity did, with C gestures being larger and faster when focused as compared to unfocused (p<0.001 for all comparisons). There was also a significant interaction between Focus Type and Within-IP Position on the peak velocity of C3 (F(2)=9.9, p<0.01). Pairwise comparisons detected a significant difference between the two unfocused conditions but only in the IP-medial position. Specifically, IP-medially, C3 gestures were faster when focus fell on the distant word (p<0.05) as opposed to the proximate word (p>0.09).

No further main or interaction effect of Focus Type was found beyond C3. In C4 and C5, there were main effects of Boundary Type in all measurements such that gestures were longer, larger, and faster in IP-final than in IP-medial positions (Table 2, lines 10-15).

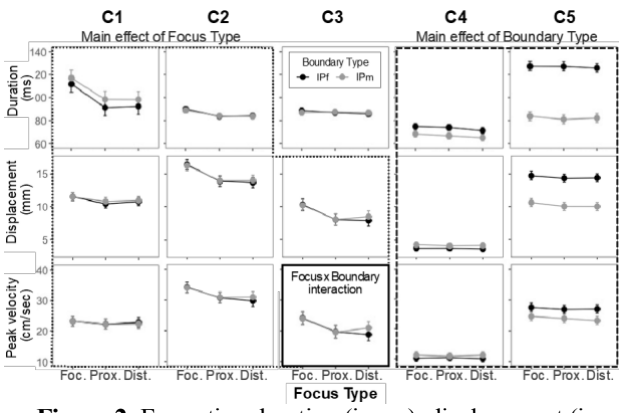


Figure 2: Formation duration (in ms), displacement (in mm), and peak velocity (in cm/sec) of C gestures (C1 to C5) as a function of Focus Type and Within-IP Position. Dotted and dashed boxes indicate main effects of Focus Type and Within-IP position, respectively, and solid boxes indicates interaction between the two factors.

In sum, consonant gestures are longer, larger, and faster under focus. The effect of focus-induced prominence extends rightward to the onset of the third syllable (C1, C2, C3). The effect of IP-boundary marking extends leftward to the onset of the fourth syllable (C4, C5). Interaction between focus and IP-final boundary is detected in the middle, and specifically on peak velocity of C3.

3.2. Relationship between kinematic parameters

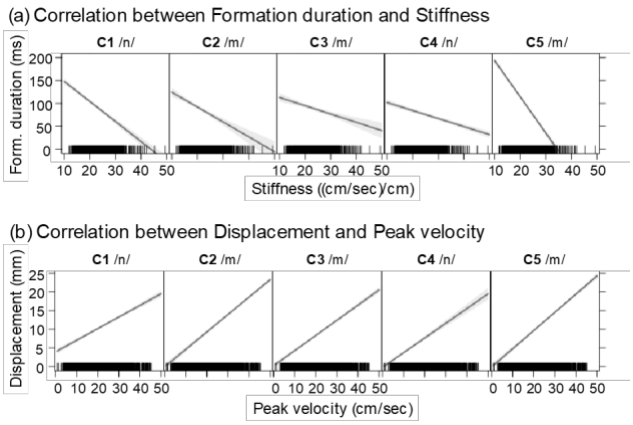


Figure 3: Relationship between (a) formation duration (in ms) and stiffness (in (cm/sec)/cm); (b) displacement (in mm) and peak velocity (in cm/sec) by Consonant.

Figure 3a demonstrates the relationship between formation duration and stiffness by Consonant. Expected covariance between formation duration and stiffness was observed ($F(1)=1100.5$, $p<0.001$) such that duration decreased as stiffness increased (cf. [11, 12]). Consonant interacted with stiffness ($F(4)=452.0$, $p<0.001$). As shown in Figure 3a, the interaction stemmed in part from the fact that the slopes were steeper when the gestures were at the edges of the phrase. Figure 3b shows the relationship between displacement and peak velocity. Expected covariance between the parameters were observed ($F(1)=3568.3$, $p<0.001$), such that displacement increased as peak velocity increased [11, 12]. Consonant interacted with peak velocity ($F(4)=98.1$, $p<0.001$), with slopes being shallower for /n/ than /m/.

4. DISCUSSION

The goal of the present study was to investigate the effect of focus-induced prominence on articulatory modulation in Seoul Korean. We found that constriction gestures under focus are longer, larger, and faster than unfocused counterparts, in line with previous findings on Korean [9]. This kinematic profile of focus is similar to that of head-prominence languages [1], despite the typological difference. One of the main findings of the present study is that the effect of focus-induced prominence goes beyond the edge of the AP initial syllable, extending rightward to the onset of the third syllable of the focused AP. Moreover, an interaction between focus position and Within-IP position was detected, indicating an intricate relationship between phrase-level prominence marking and IP boundary marking. In Articulatory Phonology [19], such prosodic modulations are instantiated by modulation gestures, or μ-gestures [20], that are either spatial or temporal. One specific type is the π-gesture, which is essentially a temporal μ-gesture marking boundaries. μ-gestures have been proposed to alter the spatiotemporal profile of the constriction gestures that overlap with them, with the scope of the effect depending on their coordination [20]. Based on our results, Seoul Korean combines a temporal μ-gesture and a spatial one to instantiate phrase-level prominence, coordinated with the beginning of the focused AP. The temporal μ-gesture may have a shorter activation interval than the spatial one, as reflected in the onset of the third syllable which showed larger and faster movement under focus, but not necessarily longer in duration. The IP-marking π-gesture seems to extend at least two syllables from the right edge of the phrase, overlapping with the μ-gesture in the third syllable. Note that constriction gestures were longer, larger, and faster when preceded by an IP boundary (see also [21]). The faster movement under the activation of π-gesture is interesting, which could possibly be arising from the interaction between π-gesture and μ-gesture.

Another question of the current study was to test whether the distance of focus would exhibit different patterns of dephrasing. Indeed, the dimension of velocity was affected by focus placement: gestures were the slowest when focus fell on the proximate word. This may suggest that the degree of dephrasing might be greater near the focused item and decrease with distance from it.

Finally, we examined the relationship between the kinematic parameters, based on previously reported correlations [11, 12]. Our results showed expected covariance between parameters—duration increases with decreased stiffness and displacement increases with peak velocity, in line with findings of head-prominence languages. However, an interesting pattern arose as a function of consonant position. The relationship between duration and stiffness showed steeper slope, i.e., duration increases more given the same decrease of stiffness, at the edges of the boundaries compared to the medial positions of the phrase. This result may be attributed as a trait of an edge-prominence language, as these positions are expected to be the main locus of focus marking (left-edge) and IP boundary marking (right-edge).

5. ACKNOWLEDGEMENTS

We would like to thank the Korean speakers who participated in the study. This work was supported by NSF Grant #1551428 and funds from the UCSB Academic Senate.

6. REFERENCES

[1] Cho, T. 2006. Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Papers in Laboratory Phonology VIII: Varieties of Phonological Competence (Phonology and Phonetics)* (pp. 519–548). Berlin, Germany: Mouton de Gruyter.

[2] Cho, T., & Keating, P. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, 29(2), 155–190.

[3] Beckman, M. E., & Pierrehumbert, J. B. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255– 309.

[4] Jun, S. A. 1998. The accentual phrase in the Korean prosodic hierarchy. *Phonology*, 15(2), 189-226.

[5] Jun, S. A. 2005. Korean intonational phonology and prosodic transcription. *Prosodic typology: The phonology of intonation and phrasing*, 1, 201.

[6] Jun, S. A. 1993. *The Phonetics and Phonology of Korean*. PhD dissertation, The Ohio State University.

[7] Jun, S. A. 2006. Intonational Phonology of Seoul Korean Revisited. *UCLA Working Papers in Phonetics*, 104, 14-25.

[8] Jeon, H. S., & Nolan, F. 2017. Prosodic marking of narrow focus in Seoul Korean. *Laboratory Phonology*, 8(1), 1-30.

[9] Shin, S., S. Kim, and T. Cho. 2015. What is special about prosodic strengthening in Korean: evidence in lingual movement in V#V and V#CV. *Proc. 17th ICPhS*, Glasgow, U.K.

[10] Cho, T. 2022. The phonetics-prosody interface and prosodic strengthening in Korean. *The Cambridge Handbook of Korean Linguistics Cambridge Handbooks in Language and Linguistics* (pp. 248-293). Cambridge University Press Cambridge.

[11] Edwards, J., Beckman, M. E., & Fletcher, J. 1991. The articulatory kinematics of final lengthening. *J. Acoust. Soc. Am.*, 89, 369–382.

[12] Byrd, D., & Saltzman, E. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2), 149–180.

[13] Katsika, A., & Tsai, K. 2021. The supralaryngeal articulation of stress and accent in Greek. *Journal of Phonetics,* 88, 101085.

[14] Jun, S. -A. 2000. K-ToBI (Korean ToBI) labelling conventions. *Speech Sciences*, 7(1), 143-170.

[15] Bates, D., Maechler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[16] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Ostry, D. J., & Munhall, K. G. 1985. Control of rate and duration of speech movements. *J. Acoust. Soc. Am.*, *77*(2), 640-648.

[18] Lenth, R. V. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.3. https://CRAN.R-project.org/package=emmeans.

[19] Browman, C. P., & Goldstein, L. M. 1992. Articulatory phonology: an overview. *Phonetica*, 49, 155–180.

[20] Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proc. 4th Speech Prosody 2008*, Campinas, Brazil, 175-184.

[21] Kim, J. J., Baek, Y., Cho, T., & Kim, S. 2019. Preboundary lengthening and boundary-related spatial expansion in Korean. *Proc. 19th ICPhS*, Melbourne, Australia, 1019-1023.

TEMPORAL SCOPE OF ARTICULATORY SLOWDOWN UNDER INFORMATIONAL FOCUS: DATA FROM ENGLISH AND MANDARIN

Yuyang Liu[1], Yichen Wang[2], Michael C. Stern[1], Benjamin M. Kramer[1], Jason A. Shaw[1]

[1]Yale University, [2]Michigan State University
yuyang.liu.yl2472@yale.edu, wangy176@msu.edu, michael.stern@yale.edu, ben.kramer@yale.edu, jason.shaw@yale.edu

ABSTRACT

Syllables produced under informational focus are longer than nonfocused syllables. Articulatorily, it is unclear whether this lengthening is the result of longer gestures or of less overlap between them and whether the articulatory basis of this effect varies across languages. We investigated these temporal aspects of focus production in English and Mandarin CV sequences using electromagnetic articulography, measuring four theoretically relevant intervals in nonfocused and focused conditions. Our results indicated that both English and Mandarin showed a temporal slowdown under focus, although the scope and magnitude of this slowdown varied across languages. We propose a generalization of the prosodic gesture model to account for the results.

Keywords: focus, Articulatory Phonology, prosodic gesture, electromagnetic articulography

1. INTRODUCTION

Crosslinguistically, focus has been shown to have a wide range of phonetic effects, impacting pitch, segment duration, and the magnitude of articulatory movements [1–7]. In this paper, we concentrate on how informational focus influences the timing of articulatory gestures in CV sequences. From acoustic studies, we know that focused syllables are longer than their closely controlled nonfocused counterparts in both English and Mandarin [2–5]. However, it is unclear what the articulatory basis of focus-modulated lengthening might be and whether it is uniform across languages. One possibility is that the consonant and vowel gestures of a CV sequence are pushed apart in time, resulting in less overlap and therefore longer syllable duration (cf. *localized hyperarticulation* [8]). Alternatively, it could be that focused syllables are longer because their constituent gestures are lengthened.

The Articulatory Phonology (AP; [9] et seq.) framework has mechanisms available to model each of the possibilities for focus realization listed above. Modeling prosody has received increased attention in AP [10], but there is no consensus on how to model focus. Here, we outline three possibilities. First, focus could make use of the same mechanism deployed to account for articulatory slowdown at phrase boundaries. Boundary-adjacent lengthening and strengthening have been modeled in AP using a prosodic gesture (π-gesture), which locally slows down all concurrently active constriction gestures [11–18]. Generalized to focus, the π-gesture model predicts that both consonants and vowels would show temporal slowdown if they are overlapped with a π-gesture. Another possibility is that the focus-induced lengthening is localized to individual gestures: either the consonant gesture, the vowel gesture, or both. This could be implemented by having focus modulate the gestural parameters, such as activation duration or stiffness [19], directly (e.g., [20]). A third possibility is that focus impacts the relative timing of consonant and vowel gestures, reducing overlap [8], which could be implemented in AP in a number of different ways [21]. This possibility might be less likely for Mandarin than for English, since Mandarin generally has less CV overlap [22, 23] (cf. [24–27] for English).

In order to assess these possibilities, we evaluate the effect of focus on four intervals in CV sequences, selected to capture the theoretically relevant aspects of intra- and intergestural timing referenced above. We evaluate the effect of focus in two languages, English and Mandarin. Both of these languages are known to have longer syllables under focus, but syllable lengthening may have different articulatory bases across languages.

2. METHODS

2.1. Participants

Acoustic and articulatory data were collected from 12 native speakers of (American) English (8 female, 4 male; ages 19–28, $\mu = 20.75$) and 12 native speakers of Mandarin (1 nonbinary, 7 female, 4

Item	Language	Condition	Stimuli	
/ni ma/	English	Nonfocused	Prompt	Is she a knee model client?
			Carrier	She's a knee model **representative**, not a knee model client.
		Focused	Prompt	Is she a knee surgeon?
			Carrier	She's a knee **model**, not a knee surgeon.
	Mandarin	Nonfocused	Prompt	Wǒ yīnggāi mà tā háishì mà nǐ? 'Should I scold him or scold you?'
			Carrier	Nǐ mà **tā** jiù xíng le, bié mà wǒ. 'Just scold him; don't scold me.'
		Focused	Prompt	Wǒ yīnggāi mà tā háishì dǎ tā? 'Should I scold him or hit him?'
			Carrier	Nǐ **mà** tā jiù xíng le, bié dǎ tā. 'Just scold him; don't hit him.'

Table 1: Examples of stimuli.
Note: Focused constituents are bolded, and target syllables are underlined.

male; ages 19–33, $\mu = 24$). No participants reported speech or hearing impairments.

2.2. Stimuli

We elicited productions of eight word-initial CV sequences in each language, in which the initial consonant was a bilabial — either /b/ or /m/ — and the vowel was either /a/ or /i/. Target syllables containing /a/ were immediately preceded by /i/, and those containing /i/ were immediately preceded by /a/. Mandarin target syllables bore a falling tone (T4) and were immediately preceded by a low tone (T3). Each target syllable was produced across two conditions: *focused*, in which the word containing the target syllable was informationally prominent in the carrier sentence, and *nonfocused*, in which the word containing the target syllable was not prominent. In order to encourage natural, communicative speech, a prompt accompanied each carrier sentence. Examples of prompts and carrier sentences are given in Table 1.

2.3. Procedure

Prompts and carrier sentences were presented to participants using E-Prime in a sound-attenuated laboratory. Each prompt was displayed on the screen and accompanied by an audio recording of the prompt. The prompt remained on screen for 5 seconds before the carrier sentence was presented. Participants were instructed to listen to the prompt and to read the carrier sentence that followed.

In total, each participant produced 128 tokens (8 items $\times$ 2 conditions $\times$ 8 repetitions) across 4 blocks of 32 items each. The first two blocks each consisted of 4 repetition cycles of stimuli in the nonfocused condition, and the last two blocks each consisted of 4 repetition cycles of stimuli in the focused condition. Within each repetition cycle, stimuli were presented in a randomized order.

The NDI Wave Speech Research System was used to record movements of nine sensors attached to the articulators and head at a sampling rate of 100 Hz. High-viscosity PeriAcryl was used to attach three sensors to the tongue: tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), placed ~1 cm, ~3 cm, and ~5 cm from the tip of the tongue, respectively. In order to track movements of the jaw, one lower incisor (LI) sensor was attached to the hard tissue of the gum directly below the left incisor. Two sensors were attached at the vermillion border of the upper lip (UL) and lower lip (LL). Reference sensors were attached on the left and right mastoids and on the nasion. Measurements of the occlusal plane and a midsagittal palate trace were also collected. Acoustic data were collected using a Sennheiser shotgun microphone at a sampling rate of 22,050 Hz.

2.4. Analysis

Articulatory data were rotated to the occlusal plane and corrected for head movement computationally.[1] Articulatory gestural landmarks were parsed from the sensor trajectories using MVIEW, a MATLAB-based program for articulatory data visualization.[2] Gestures associated with /b/ and /m/ were extracted from measurements of lip aperture (LA), calculated as the Euclidean distance between the UL and LL sensors. Gestures associated with /a/ were parsed from the trajectory of the TD sensor. Gestures associated with /i/ were parsed from the trajectory of the TB sensor if this sensor was judged to be closest to the palate at the point of maximum /i/ constriction for a participant and from the TD sensor otherwise.

The following gestural landmarks entered into the calculation of the intervals for analysis: the *onset* of controlled movement, the achievement of *target*, the *point of minimal velocity*, the *release* from constriction, and the *offset* of controlled movement. These landmarks were parsed from the continuous trajectories with reference to the velocity signal. The onset and target of a gesture were measured as the timepoints at which the gesture's tangential velocity

Language	Condition	CV lag	C_{CLOS}	C_{OPEN}	V_{OPEN}
English	Nonfocused	36.02 (13.17)	74.39 (8.44)	86.89 (14.06)	146.90 (13.45)
	Focused	41.02 (17.82)	79.73 (10.35)	102.38 (14.21)	174.08 (19.07)
Mandarin	Nonfocused	38.18 (13.25)	81.38 (7.92)	81.58 (6.24)	135.57 (16.09)
	Focused	36.40 (10.91)	83.11 (7.67)	82.93 (10.55)	145.92 (10.96)

Table 2: Means (standard deviations) of key interval durations (ms).

Fixed effect		CV lag	C_{CLOS}	C_{OPEN}	V_{OPEN}
Focus estimate (ms)	English	5.52***	5.56***	16.87***	27.53***
	Mandarin	.51	1.39	2.34	9.25***
Focus + language χ^2		.28	2.78	7.70**	9.46**
Focus × language interaction χ^2		3.77	4.24*	33.22***	32.21***

Table 3: ANOVA results of LMER models.
Note: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

in the movement towards constriction exceeded or sank below, respectively, a 20% threshold of a manually selected local velocity peak. The point of minimal velocity was measured as the timepoint of the velocity minimum. The release and offset of a gesture were measured as the timepoints at which the gesture's tangential velocity in the movement away from constriction exceeded or sank below, respectively, a 20% threshold of a manually selected local velocity peak. Figure 1 provides an example of the gestural landmarks defined above.

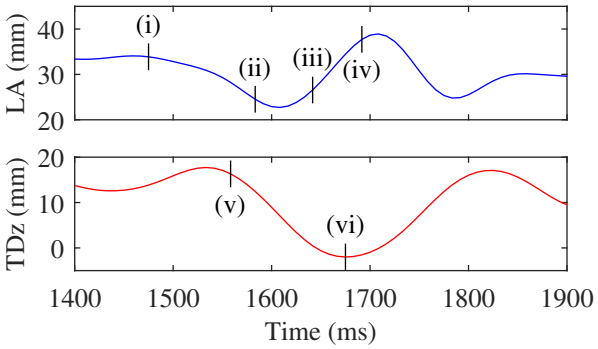


Figure 1: Gestural landmarks.
Note: (i) C onset, (ii) C target, (iii) C release, (iv) C offset, (v) V onset, (vi) V minimal velocity.

Out of the 3,072 tokens elicited, a total of 556 tokens (18.10%) were eliminated from analysis for the following reasons: data storage issues (18); disfluency (5); failure of the participant to produce contrastive focus on the informationally prominent syllable (155); or failure of the gesture parsing tool to extract the consonant or vowel gesture (378).

For all the remaining tokens, we calculated the following four intervals: (i) *CV lag* is defined as the interval from the consonant onset to the vowel onset; (ii) *consonant closing interval* (C_{CLOS}) is defined as the interval from the consonant onset to the consonant target; (iii) *consonant opening interval* (C_{OPEN}) is defined as the interval from the consonant release to the consonant offset; and (iv) *vowel opening interval* (V_{OPEN}) is defined as the interval from the onset to the point of minimal velocity of the vowel gesture. After calculating the intervals, 82 tokens, for which at least one interval was beyond three standard deviations from the mean, were excluded from analysis.

Linear mixed effects regression (LMER) models were fit to each of the intervals defined above [28].[3] We first assessed the effect of focus in each language separately. For each interval in each language, a baseline model included a fixed effect of vowel and random intercepts for subject and item. We then added focus as a fixed effect and compared it to the baseline model. We next fit a model to each interval in the combined data set, including a fixed effect of language. Finally, to statistically assess whether the effect of focus was uniform across languages, we added a fixed effect for the interaction between focus and language. For all these models, statistical significance was determined by a likelihood ratio test against a baseline model that excluded the fixed effect of interest.

3. RESULTS

Means and standard deviations of the durations of each key interval under nonfocused and focused conditions are given in Table 2. In both languages, the difference in the mean duration between the two conditions was the largest for V_{OPEN}. Additionally, for each interval, the effect of focus was larger in English than in Mandarin.

The top rows of Table 3 report the estimate for the fixed effect of focus for each interval and language

along with statistical significance. For English, focus had the biggest effect on V_{OPEN}, followed by C_{OPEN}, followed by C_{CLOS}, followed by CV lag. Mandarin followed the same numerical trend, but the effect was only significant for V_{OPEN}. The effect of language and the interaction between language and focus are reported in the bottom rows. Results show a main effect of language for C_{OPEN} and V_{OPEN} (longer for English) and confirm a significant interaction. The effect of focus is stronger in English for C_{CLOS}, C_{OPEN}, and V_{OPEN}.[4]

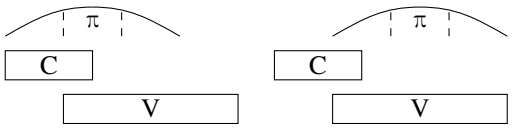
4. DISCUSSION

Our results showed that focus had a lengthening effect in English and Mandarin. In both languages, some intervals were longer in focused syllables than in nonfocused syllables, as expected from past work [2–7]. However, the locus and magnitude of the effect varied across languages. In English but not in Mandarin, focus had a significant effect on the closing and opening phases of the consonant gesture and on the relative timing between the consonant and vowel gestures. Additionally, while focus had a significant effect on the opening phase of the vowel gesture in both languages, it had a stronger effect in English than in Mandarin.

These data allow us to rule out two of the three possible theoretical accounts of the effect of focus on timing outlined in the introduction. It seems that focused syllables are not longer because of less overlap between gestures; there was no effect of focus on CV lag in Mandarin and only a small effect in English, similar in magnitude to the effect on the closing phase of the consonant gesture. We can therefore rule out localized hyperarticulation as an articulatory basis of focus lengthening. It also appears that focus is not directly modulating single gestures; at least in English, the temporal scope of focus encompasses the entire CV sequence.

The third possibility that we raised was the π-gesture model. To date, this has been primarily used for explaining the local lengthening and slowing of articulatory gestures at phrase boundaries [11–18]. The π-gesture is argued to have an extent in time but lack independent articulatory realization; it may only be realized vicariously through its effect on the concurrently active vocal tract constriction gestures [16]. Essentially, the π-gesture triggers a transgestural local slowdown in proportion to its activation level on all constriction gestures with which it overlaps [13, 14, 16]. This approach has promise for accounting for the effects of focus in these data, including the crosslinguistic variation observed.

We propose that the target of the temporal effects of focus is a continuously ramped π-gesture, whose activation level is the highest at the midpoint of the gesture and the lowest at its onset and offset. As a consequence of the focus-induced π-gesture, all concurrently active constriction gestures are slowed down and therefore lengthened, as is observed at phrase boundaries. To account for the language-specific effects that we observed, we posit that the π-gesture may be temporally aligned to different timepoints in the constriction dynamics in different languages: in English, the onset of the π-gesture is aligned to the onset of the syllable, whereas in Mandarin, it is aligned to the onset of the vowel gesture. This hypothesis is illustrated in Figure 2.



(a) English. **(b)** Mandarin.

Figure 2: Proposed gestural alignments.

As in Figure 2(a), in English, the π-gesture has its onset aligned to the onset of the syllable and is maximally activated at the opening phase of the vowel gesture, which partially overlaps with the closing phase of the consonant gesture. As a result, when focus is realized on the syllable, V_{OPEN} and C_{OPEN} get slowed down the most by the π-gesture, and C_{CLOS} less so. In Mandarin, on the other hand, the onset of the π-gesture is aligned to the onset of the vowel gesture, as in Figure 2(b). This way, since the consonant gesture hardly overlaps with the π-gesture, focus has no effect on its duration. In contrast, focus does have an effect on V_{OPEN} but not as strong as the effect found in English, because Mandarin π-gestures are not yet maximally activated at the opening phase of the vowel.

5. CONCLUSION

Past work has shown that informational focus has a lengthening effect in both English and Mandarin. We explored the articulatory basis of this effect, finding that focus increases the gestural duration of vowels (both languages) and onset consonants (English only). Both languages showed focus-driven articulatory lengthening, but its scope and magnitude differed. We argued that this pattern of results is consistent with a π-gesture account of focus. On this account, language variation in focus realization derives from the temporal alignment of the π-gesture.

6. REFERENCES

[1] D. R. Ladd, "Phonological features of intonational peaks," *Language*, vol. 59, no. 4, pp. 721–759, Dec. 1983.

[2] W. Cooper, S. Eady, and P. Mueller, "Acoustical aspects of contrastive stress in question-answer contexts," *J. Acoust. Soc. Am.*, vol. 77, no. 6, pp. 2142–2156, Jun. 1985.

[3] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonol. Yearb.*, vol. 3, pp. 255–309, May 1986.

[4] Y. Xu, "Effects of tone and focus on the formation and alignment of f_0 contours," *J. Phon.*, vol. 27, no. 1, pp. 55–105, Jan. 1999.

[5] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *J. Phon.*, vol. 33, no. 2, pp. 159–197, Apr. 2005.

[6] S. Roessig and D. Mücke, "Modeling dimensions of prosodic prominence," *Front. Commun.*, vol. 4, p. 44, Sep. 2019.

[7] S. Roessig, B. Winter, and D. Mücke, "Tracing the phonetic space of prosodic focus marking," *Front. Artif. Intell.*, vol. 5, May 2022, Art. no. 842546.

[8] K. de Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 491–504, Jan. 1995.

[9] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, May 1992.

[10] D. Byrd and J. Krivokapić, "Cracking prosody in Articulatory Phonology," *Annu. Rev. Linguist.*, vol. 7, no. 1, pp. 31–53, Jan. 2021.

[11] M. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, G. Docherty and D. R. Ladd, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1992, pp. 68–89.

[12] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *J. Phon.*, vol. 26, no. 2, pp. 173–199, Apr. 1998.

[13] D. Byrd, "Articulatory vowel lengthening and coordination at phrasal junctures," *Phonetica*, vol. 57, no. 1, pp. 3–16, Mar. 2000.

[14] D. Byrd, A. Kaun, S. Narayanan, and E. Saltzman, "Phrasal signatures in articulation," in *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, M. Broe and J. Pierrehumbert, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2000, pp. 70–87.

[15] T. Cho, *The Effects of Prosody on Articulation in English*. New York, NY, USA: Routledge, 2002.

[16] D. Byrd and E. Saltzman, "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," *J. Phon.*, vol. 31, no. 2, pp. 149–180, Apr. 2003.

[17] T. Cho, "Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English," in *Laboratory Phonology 8*, L. Goldstein, D. Whalen, and C. Best, Eds. Berlin, Germany: Mouton de Gruyter, 2006, pp. 519–548.

[18] D. Byrd, J. Krivokapić, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1589–1599, Sep. 2006.

[19] C. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *J. Phon.*, vol. 18, no. 3, pp. 299–320, Jul. 1990.

[20] T. Cho, "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ, i/ in English," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3867–3878, Jun. 2005.

[21] E. Saltzman, H. Nam, J. Krivokapić, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," *Speech Prosody*, vol. 4, pp. 175–184, May 2008.

[22] M. Gao, "Mandarin tones: An Articulatory Phonology account," Ph.D. dissertation, Dept. Linguist., Yale Univ., New Haven, CT, 2008.

[23] M. Zhang, C. Geissler, and J. Shaw, "Gestural representations of tone in Mandarin: Evidence from timing alternations," *Int. Congr. Phon. Sci.*, vol. 19, pp. 1803–1807, Aug. 2019.

[24] A. Löfqvist and V. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1864–1876, Mar. 1999.

[25] C. Browman and L. Goldstein, "Competing constraints on intergestural coordination and self-organization of phonological structures," *Bull. Commun. Parlée*, vol. 5, pp. 25–34, Dec. 2000.

[26] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: A coupled oscillator model," in *Approaches to Phonological Complexity*, F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé, Eds. Berlin, Germany: Mouton de Gruyter, 2009, pp. 297–328.

[27] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model," *Mot. Control*, vol. 14, no. 3, pp. 380–407, Jul. 2010.

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using **lme4**," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, Oct. 2015.

[1] The MathWorks, Inc. *MATLAB version: 9.13.0 (R2022b)*. (2022). The MathWorks, Inc.

[2] M. Tiede. *MVIEW: Software for visualization and analysis of concurrently recorded movement data*. (2005). Haskins Laboratories.

[3] R Core Team. *R: A language and environment for statistical computing*. (2022). R Foundation for Statistical Computing.

[4] Data and complete statistical analysis are publicly available at https://osf.io/xkj72/.

A HIERARCHY OF PROMINENCE: THE PRODUCTION AND PERCEPTION OF FOCUS IN AMERICAN ENGLISH

Argyro Katsika[1], Jiyoung Jang[1], Jelena Kriivokapić[2], Louis Goldstein[3] and Elliot Saltzman[4]

[1]University of California, Santa Barbara
[2]University of Michigan
[3]University of Southern California
[4]Boston University

argyro@ucsb.edu, jiyoung@ucsb.edu, jelenak@umich.edu, louisgol@usc.edu, esaltz@bu.edu

ABSTRACT

Pitch accents on stressed syllables mark prominent words. Moreover, articulatory gestures are longer, larger and faster when accented. It is also known that prominence marking interacts with focus structure. However, it is still unclear how many degrees of prominence are encoded phonetically and whether these are perceived. Here, we assess these issues via an electromagnetic articulography study of American English and a corresponding crowd-sourced online perception survey.

Results show that phonetic dimensions (acoustic F0; and kinematic duration, displacement and velocity) differ in the number of degrees of prominence they distinguish. Nonetheless, a hierarchical ordering of the reflected degrees remains consistent across dimensions. Listeners' judgements reflect this: Degrees of prominence are more successfully distinguished by listeners the further away they are from each other hierarchically; neighbouring degrees are distinguished more poorly the lower they are in the hierarchy. An account for this hierarchy is discussed with the framework of Articulatory Phonology.

Keywords: Prominence, focus, prosody, production-perception relationship, Articulatory Phonology.

1. INTRODUCTION

Phrasal pitch accents mark words within phrases as rhythmically or conceptually important, and they are associated to the stressed syllables of the accented words (e.g., [1, 2]). The articulatory movements, called *gestures*, forming the consonant (C) and vowel (V) constrictions in stressed syllables have in turn been found to become longer, larger and faster when accented [3, 4, 5, 6, 7, 8, 9, 10, 11, 12; but see 13]. This spatiotemporal expansion that gestures undergo is often referred to as *prosodic strengthening* (cf. [6]). Later work has presented evidence that it is focus structure and not simply accentuation that causes strengthening ([14, 15, 16] for German; see also [17] for English). This line of research considered different types of focus (i.e., broad focus, narrow focus, contrastive focus, unfocused) and found that phonetic dimensions, such as acoustic F0 and kinematic duration and displacement, increased across focus types. Although there might be typological differences and the set of dimensions used might differ on a language-dependent basis, phrase-level prominence emerges from these findings not only as a multi-dimensional (i.e., employing multiple phonetic dimensions) but also as a multi-level system, which goes beyond the distinction between accented and unaccented. Of course, the presence of a pitch accent automatically denotes an accented vs. unaccented split.

Taken together, the conclusions of previous research suggest three hypotheses: 1) Prominence is organized hierarchically, with levels of the hierarchy possibly representing focus types. However, it is unclear whether all focus types are encoded, and if yes, in what order, since, to date, neither the complete range of degrees of prominence nor the phonetic correlates of these degrees have been established. 2) Whichever set of phonetic parameters a language uses for marking phrase-level prominence, this multi-dimensional system functions as a whole, i.e., with all the phonetic parameters being modified in tandem. Interestingly, [15] presents a dynamical model that captures the observed focus-induced modifications on multiple phonetic dimensions in German by controlling the same, single parameter. 3) Assuming that prominence encodes aspects of information structure, playing thus a significant role in speech comprehension, and given the abundance of phonetic cues used in this encoding, we expect that any degrees of prominence represented phonetically should also be perceived by listeners in their produced order.

Here, we take a first step towards addressing these hypotheses by the means of an electromagnetic

articulography study of American English and a corresponding crowd-sourced online perception survey. By examining both production and perception, a wide range of focus types (i.e., contrastive focus, narrow focus, broad focus, deaccented and unfocused) and multiple phonetic parameters known to be used in prominence marking in American English (acoustic F0; and kinematic duration, displacement and velocity), the current study ultimately aims at examining the hierarchical structure for phrase-level prominence.

2. METHODS

2.1. Production study

The production data from [17] were retrieved and analyzed for this study, and the reader is referred to that study for a full description of the experimental design and procedure. Here, a brief summary of these aspects is offered.

2.1.1. Participants and recording apparatus

There were eight native speakers (2 male, 6 female; mean age: 23) of American English to this study. Kinematic data were collected using the AG501 3D electromagnetic articulograph (EMA; Carstens Medizinelektronik). Receiver coils were attached to the tongue dorsum, tongue body's center, tongue tip, upper and lower lips, upper and lower incisors, left and right ears, and nose. Audio data were collected simultaneously at a sampling rate of 16 kHz.

2.1.2. Stimuli and experimental design

The following nine English words were used as test words: *bee*, *baby*, *melody*, *military*, *design*, *banana*, *humanity*, *matinee*, and *salmonella*. Test words were placed in frame sentences appropriate for their meaning, which were in turn paired with prompt sentences designed to elicit the test words in five focus types: 1) unfocused, following a narrowly focused item (UF), 2) de-accented by virtue of following a contrastively focused item (DA), 3) accented under broad focus (BF), 4) accented under narrow focus (NF), or 5) accented under contrastive focus (CF). To illustrate, a pair of prompt-target sentences for the test word *bee* in the de-accented condition is given in (1).

(1) **Prompt:** *Is it the zoologist who will be testing the bee with the stripes?*

Target: *No, the botanist will be testing the bee with the stripes*

The stimuli were presented on a monitor. The participant read the prompt sentence silently, and the target sentence aloud. Eight blocks of the stimuli were recorded; stimuli in each block were presented in a different random order.

2.1.3. Data analysis

Custom software (Tiede, Haskins Laboratories) was used to semi-automatically detect the C gesture of the stressed syllables of the test words and to label important timepoints in their kinematic progression on the basis of velocity criteria (see Figure 1). Labial and coronal C gestures were labeled on the lip aperture and tongue tip vertical displacement trajectory respectively.

The following timepoints are relevant for the current analyses: gestural onset, peak velocity of constriction's forming movement, and constriction's target (i.e., achievement). Based on these timepoints, the measures of (a) maximum displacement (as the absolute difference between constriction's target and gestural onset) and (b) peak velocity of the formation were calculated. Formation duration (i.e., the interval from onset to release) is not directly reported here, since it is analyzed in [17].

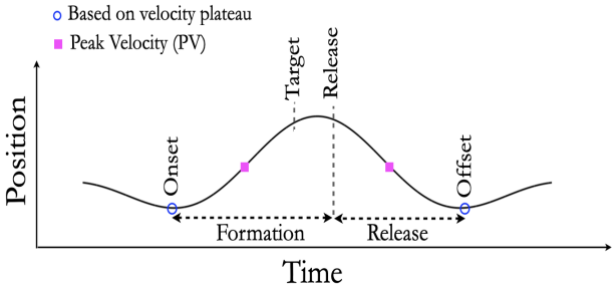


Figure 1: Schematic representation of a constriction gesture and labelled timepoints.

Analysis of acoustic F0 is underway and currently includes the subset of the data that served as the stimuli for the perception study. Stressed vowels were segmented in Praat [18], with their acoustic boundaries being determined at the onset and offset of F2. Maximum F0 at the midpoint of the vowel was automatically extracted using a Praat script.

Statistical analyses were performed by linear mixed effects models using the lme4 package [19] in R [20]. For the dependent variables of displacement, peak velocity, and F0, the fixed effect of Focus Type (UF, DA, BF, NF, CF) and the random effects of Speaker and Word were considered (i.e., *bee* and *melody* for F0; all test words for the kinematic measures). Post-hoc pairwise comparisons with Holm adjustment were assessed using emmeans [21].

2.2. Perception study

In the crowd-sourced online perception survey the fourth repetition of the stimuli for the test words *bee* and *melody* from the production study were assessed by 185 native speakers of American English (age groups: 18-25: 179 speakers; 26-30: 5 speakers; 51-65: 1 speaker). Productions from the middle of the experiment (4th block) were selected as they were considered the least susceptible to possible order effects (e.g., newness of task at the beginning or tiredness at the end). The survey was prepared and presented using Qualtrics [22]. There were eight versions of the survey, each using the data from one speaker from the production study. In each version, there were 23 participants, except for two versions that had 24 participants and one version that had 22.

As a reminder, the production study involved pairs of prompt and target sentences; the prompt was a question and the target was its response. Survey participants listened to the target sentence (produced by the speaker that corresponded to their survey's version) twice at their own pace while viewing on their screen two prompt questions. One question was the original pair in terms of focus type to the target sentence and one that was not. To avoid signaling focus type by other cues, the initial word, which was either 'no' or 'oh', from each target utterance was omitted. The task of the participant was to select the question to which the target sentence they heard was a better response. The pairs of questions the listeners heard represented all possible combinations of the five focus types, yielding 10 combinations. In total, 20 trials were included in each survey (10 focus types combinations x 2 test words x 1 speaker as source). The survey took about 15 minutes to complete. In total, 7400 judgements were acquired.

Survey responses were analyzed in terms of percentage of successful selection between focus types. Additional analyses of kinematic (i.e., duration, maximum displacement, and peak velocity of C gesture formations) and acoustic (i.e., maximum F0 of vowels) dimensions were performed to the subset of data from the production study that was also used in the perception survey. The goal was to examine correlations between selection success rate (in %) and each of these dimensions as well as correlations within each pair of phonetic dimensions. Specifically, for each focus type combination tested in the given version of the survey, we calculated the difference in value (i.e., value in target focus type minus value in competitor focus type) on each phonetic dimension between the corresponding two tokens from the production study. Correlations were analyzed using the GGally [23] package in R [20].

3. RESULTS

3.1. The effect of focus type on phonetic dimensions

The statistical analysis detected main effects of Focus Type on all phonetic dimensions examined (F0: $\chi^2(4)=134.0$, $p<0.001$; Displacement: $\chi^2(4)=78.45$, $p<0.001$; peak velocity: $\chi^2(4)=30.01$, $p<0.001$). The post-hoc pairwise comparisons are listed in Table 1.

		CF	NF	BF	DA
NF	F0	*n.s.*			
	DISP	*n.s.*			
	PV	*n.s.*			
BF	F0	0.008	*n.s.*		
	DISP	0.0001	0.07 (*m.*)		
	PV	0.09	*n.s.*		
DA	F0	<0.0001	<0.0001	<0.0001	
	DISP	<0.0001	0.0001	*n.s.*	
	PV	0.0008	0.006	*n.s*	
UF	F0	<0.0001	<0.0001	<0.0001	*n.s.*
	DISP	<0.0001	<0.0001	0.007	*n.s.*
	PV	0.0001	0.001	*n.s.*	*n.s.*

Table 1: *p* values for pairwise comparisons by Focus Type for F0, displacement (DISP) and peak velocity (PV). Non-significant (*n.s.*) and marginally significant ($0.05 < p < 0.01$) (*m.*) comparisons are indicated.

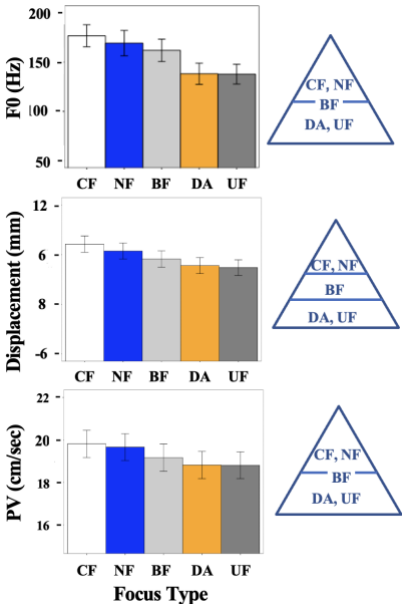


Figure 2: Mean values (with standard error) of stressed vowel's maximum F0 (Hz), and C gesture's formation displacement (mm) and peak velocity (cm/sec) by focus type (DA, UF, BF, NF, CF). A schematic representation of the focus types distinguished per phonetic dimension is juxtaposed.

As per our hypothesis (1), pairwise comparisons indicated that phonetic dimensions differentiated

primarily among focus types and not simply between presence (in all focused conditions) and absence (unaccented and de-accented) of pitch accent. However, phonetic dimensions differed in the number of focus types, and thus degrees of prominence, they distinguished. As illustrated in Figure 2, acoustic F0 and kinematic peak velocity distinguished two degrees: CF, NF > DA, UF with BF not being clearly distinct from either, and kinematic displacement distinguished three degrees: CF, NF > BF > DA, UF. Note that kinematic duration, reported for the same set of data in [17] presented four degrees: CF > NF > BF > DA, UF. Nonetheless, a hierarchical ordering of the reflected degrees remained consistent across dimensions, with CF being on the high extreme and UF on the low.

3.2. Success rate for focus type distinction

Overall, 76% of the total number of judgements successfully selected the target focus type. Figure 3 plots the distribution of successful selections across target-competitor pairs. Listeners' judgments systematically reflected the hierarchical ordering encoded phonetically. In particular, degrees of prominence are more successfully distinguished perceptually the further away they are from each other hierarchically (e.g., 85% in CF vs. DA), whereas neighbouring degrees are distinguished more poorly, especially when lower in the hierarchy (e.g., DA vs. UF is at chance).

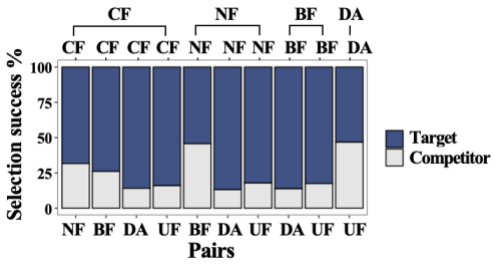


Figure 3: Listeners' selection success rate (%) across target-competitor pairs. Focus types (CF, NF, BF, DA, UF) of each member of the pair are shown on each side of the percentage bar.

3.3. Correlations

Significant correlations were found between selection success rate and all phonetic dimensions as well as within each pair of dimensions (Fig. 4). These patterns, combined with those in 3.1, suggest that phonetic dimensions are modified in tandem across the same hierarchical order of degrees, per hypothesis (2). This seems to facilitate listeners decode focus information, as in hypothesis (3), presumably because of the abundance of phonetic cues. Future work will assess the contribution of individual dimensions to this production-perception relationship.

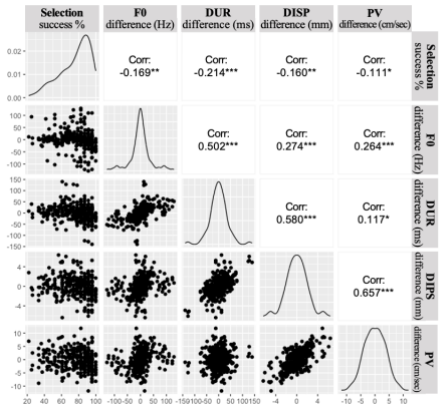


Figure 4: Correlations of selection success rate (in %) and target-competitor difference in F0 (in Hz), duration (DUR in ms), displacement (DISP in mm), and peak velocity (PV in cm/sec).

4. DISCUSSION

Results confirm that prominence encodes focus structure and not solely accentuation [14, 15, 16]. Furthermore, a hierarchical structure of prominence emerges that is reflected in perception. Different dimensions encode different levels of this structure, but this might be an epiphenomenon of different degrees of granularity of the measures used or the perceptual functionality of each phonetic dimension (e.g., velocity does not have a psychoacoustic correspondent). Still, the different dimensions seem to be interconnected. Some of these connections might simply be due to physiological reasons. For instance, it is well established that peak velocity varies with displacement [24, 25]. Others, like the relationship between duration and displacement, are more challenging to capture (see e.g., [6]).

Assuming a single dynamical parameter controlling all of these effects [15] is conceptually compelling, but it is less clear how such an account would capture typological differences. A different account can be offered from within the framework of Articulatory Phonology [e.g., 26], in which prosodic modulations are instantiated by modulation gestures that control the spatial (spatial μ-gestures) and/or temporal (temporal μ-gestures) profile of the C and V constriction gestures [27] and the tone gestures [cf. 28] that overlap with them. The degree of μ'gestures' effect increases at higher hierarchical levels. Based on our findings, English has both a spatial (for effects on displacement and peak velocity) and a temporal μ-gesture (for effects on duration) coordinated with the stressed syllable. There is also a pitch accent gesture coordinated with the same syllable. The higher the prominence level hierarchically, the stronger the strength of the μ-gestures, and thus the stronger their effects on the constriction and pitch accent gestures that overlap with them.

6. ACKNOWLEDGMENTS

The work was supported by collaborative NSF Grants #1551513 / 1551649 / 1551428 / 1551695.

7. REFERENCES

[1] Beckman, M.E., Pierrehumbert, J. 1986. Intonation structure in English and Japanese. *Phonology Yearbook* 3, 255-310.

[2] Silverman, K., Beckman, M.E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard labeling English prosody. *Proc. of the International Conference on Spoken Language Processing 2*, 867–870.

[3] Beckman, M. E., Edwards, J., Fletcher, J. 1992. Prosodic structure and tempo in a sonority model of articulatory dynamics. In: Docherty G. J., Ladd D.R. (eds), *Papers in Laboratory Phonology II: Segment, gesture, prosody*, Cambridge University Press, 68–86.

[4] Beckman M.E., Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In Keating, P.A. (ed), *Phonological Structure and Phonetic Evidence*, Cambridge University Press, 7-33.

[5] Cho, T. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *JASA* 17, 3867-3878.

[6] Cho, T. 2006. Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. In Goldstein, L., Whalen, D.H., Best, C.T. (eds), *Papers in Laboratory Phonology VIII: Varieties of Phonological Competence (Phonology and Phonetics*), Mouton de Gruyter, 519-548.

[7] de Jong, K. 1991. An articulatory study of consonant-induced vowel duration changes in English. *Phonetica,* 48, 1-17, 1991.

[8] de Jong, K. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation, *JASA* 97, 491–504.

[9] de Jong, K., Beckman, M.E., Edwards, J. 1993. The interplay between prosodic structure and coarticulation, *Language and Speech*, 36, 197–212.

[10] Fowler, C.A. 1995. Acoustic and kinematic correlates of contrastive stress accent in spoken English. In: Bell-Berti, F., Raphael, J.J. (eds), *Producing Speech: Contemporary Issues: For Katherine Safford Harris*, Woodbury: American Institute of Physics, 355-373.

[11] Harrington, J., Fletcher, J., Beckman, M.E. 2000. Manner and place conflicts in the articulation of accent in Australian English. In: Broe, M. (ed.), *Papers in Laboratory Phonology 5*, Cambridge University Press, 40–55.

[12] Harrington, J., Fletcher, J., Roberts, C. 1995. Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics* 23, 305-322.

[13] Katsika, A., Tsai, K. 2021. The supralaryngeal articulation of stress and accent in Greek, Journal of Phonetics, 88, 101085.

[14] Hermes, A. Becker, J., Mücke, D., Baumann, S., Grice, M. 2008. Articulatory gestures and focus marking in German. *Proc. of Speech Prosody 2008*, 457-460.

[15] Roessig, S., Mücke, D. 2019. Modeling dimensions of prosodic prominence. *Frontiers in Communication* 4, article 44.

[16] Mücke, D., Grice, M. 2014. The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics* 44, 47-61.

[17] Katsika, A., Jang, J., Krivokapić, J., Goldstein, L., Saltzman, E. 2020. The role of focus in accentual lengthening in American English: Kinematic analyses. *Proceedings of the 10th International Conference on Speech Prosody 2020*, Tokyo, Japan.

[18] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program], Version 6.3.03, retrieved from http://www.praat.org/.

[19] Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[20] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[21] Lenth, R.V. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.3.https://CRAN.R-project.org/package=emmeans.

[22] Qualtrics (2005). Provo, Utah, USA. Version 2021. Available at: https://www.qualtrics.com

[23] Schloerke, B. Cook, D., Larmarange, L., Briatte, F., Marbach, M., Thoen, E., Elberg, A. Crowley, J. 2021. GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally

[24] Munhall, K.G., Ostry, D.J., Parush, A. 1985. Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance* 11, 457–474.

[25] Ostry, D.J., Munhall, K.J. 1985. Control of rate and duration of speech movements. *JASA* 77, 640–648.

[26] Browman, C.P., Goldstein, L.M. 1992. Articulatory phonology: An overview. *Phonetica* 45, 155-180.

[27] Saltzman, E., Nam, H., Krivokapić, J., Goldstein, L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proc. of the Speech Prosody Conference 2008*, 175-184.

[28] Mücke, D., Nam, H., Hermes, A., Goldstein, L. 2012. Coupling of tone and constriction gestures in pitch accents. In: Hoole, P., Bombien, L., Pouplier, M., Mooshammer, C., Kühnert, B. (eds), *Consonant Clusters and Structural Complexity.* Mouton de Gruyter, https://doi.org/10.1515/9781614510772.205.

THE PEDAGOGY OF THE LARYNGEAL ARTICULATOR MODEL

John Esling[1], Allison Benner[1], Silvia Calamai[2], Chiara Celata[3], Lise Crevier Buchman[4], Míša Hejná[5], Scott Moisik[6]

[1]University of Victoria, [2]Università di Siena, [3]Università di Urbino, [4]Université Sorbonne Nouvelle & Université Paris-Saclay, Hôpital Foch, [5]Aarhus University, [6]Nanyang Technological University
[1]esling@uvic.ca, [1]abenner@uvic.ca, [2]silvia.calamai@unisi.it, [3]chiara.celata@uniurb.it, [4]lise.buchman1@gmail.com, [5]misa.hejna@cc.au.dk, [6]scott.moisik@ntu.edu.sg

ABSTRACT

The Laryngeal Articulator Model (LAM) approach to the vocal tract in speech production has emerged over the past decades as an alternative to a strictly glotto-lingual approach. The LAM expands explanatory power by acknowledging: 1. that the larynx is a complex articulator consisting of a network of structures; and 2. that articulations in the lower vocal tract interact with articulations in the oral vocal tract. The LAM approach to speech production is not yet widely implemented in phonetics and phonology teaching materials. The main goal of this paper is to explore the different challenges of laryngeal pedagogy in order to make it easier for scholars in the field to engage with the LAM in pedagogical contexts. To do this we present the authors' specific experiences using the LAM in different pedagogical situations: teaching voice quality, training research/teaching assistants, writing textbooks, training speech clinicians, and modelling speech production.

Keywords: Laryngeal Articulator Model, voice quality, pedagogy, clinical phonetics, modelling

1. INTRODUCTION

This paper explores some of the challenges of laryngeal pedagogy, focusing on the Laryngeal Articulator Model (LAM) [15]. We present the authors' experiences using the LAM in the contexts of teaching voice quality, training research/teaching assistants, writing textbooks, training speech clinicians, and modelling speech production.

2. TRAINING RESEARCH ASSISTANTS

The LAM increases the number of phonation types and states of the larynx for phonetic description and instruction, both in terms of their articulatory production and their auditory correlates. For example, various types of harsh voice, aryepiglottic (AE) trilling, whispery voice and raised larynx voice and the states glottal stop, prephonation and epiglottal stop have revised articulatory definitions. Each type/state in the model represents a stage or level along the continuum in the functioning of the aryepiglottic (AE) sphincter mechanism: from wide open to firmly shut. The most basic auditory distinction dividing this continuum is between "constricted" and "unconstricted" qualities. Research assistants (RAs) coding laryngeal utterances in language or infant-speech databases can be trained to distinguish relatively tight (whispery, creaky, harsh, AE-trilled, pressed) vs. relatively open (breathy, modal, falsetto) postures of the laryngeal mechanism. These postures/qualities represent a larger and more visualizable action as a more 3D concept than if the larynx were only an opening-closing glottis. Here we highlight the application of the LAM in training RAs, using the Infant Speech Acquisition (InSpA) project as an example [4, 13-14].

The InSpA project investigated the incidence and role of laryngeal constriction in infant speech across the first year of life among infants hearing English, French, Bai, or Moroccan Arabic. Initially, we trained RAs to identify the phonation types produced by infants and used this system to code utterances in our database. To check accuracy, each utterance was independently coded by multiple RAs. Where coding differed among the RAs, we listened to the utterances together and viewed wideband spectrograms to reach agreement on which phonation type(s) occurred in these utterances. At the outset, there was high inter-rater variability, particularly in distinguishing between breathy voice and whispery voice and in coding utterances that included compound phonation types. We addressed this problem by teaching RAs to distinguish between constricted and unconstricted utterances. RAs were more likely to agree using this parameter and we began to use it as the primary criterion for coding utterances rather than individual phonation types.

While our goal was to improve inter-rater reliability, we found that this method is more effective at explaining the process of early infant phonetic development. Graphs and tables showing the incidence of individual phonation types would have obscured the most important finding: that infants

from all language groups begin their first year producing primarily constricted sounds and that the incidence of constriction declines linearly over the year as infants begin to produce more unconstricted sounds. Focusing on the parameter of laryngeal constriction also showed that by the end of the first year, the incidence of constriction in infants' babbling correlates with whether or not the language they are hearing employs laryngeal constriction contrastively, e.g., more constriction persists in the babbling of infants in a Bai (tonal-register) context than in an English context [3].

The LAM also helped illustrate how infants gain articulatory control by exploring laryngeal constriction. Starting in months 3 and 4, infants begin to alternate between constricted and unconstricted phonation types within single utterances ("dynamic" utterances in our framework). These alternations in the laryngeal vocal tract help "prime" the exploration of the oral vocal tract. Before producing babbling – the most language-like of prelinguistic utterances – infants alternate between non-syllabic and syllabic sequences within the same utterance ("mixed" utterances in our framework). In these mixed utterances, infants explore coordinated use of the laryngeal and oral vocal tracts, which is essential to the emergence of speech-like utterances. The LAM was not only pedagogically more effective in training RAs, but its unique affordances made patterns in infant speech acquisition more generalizable.

3. EXAMPLE FROM A TEXTBOOK

Next, we report our experience with an introductory manual of speech sciences for Italian university students of languages and linguistics, which is currently under preparation [6]. The book is conceived of as a guide to the linguistic study of speech sounds, from production to acoustics and to hearing and perception and with practical illustrations of common procedures for the recording, archiving / FAIRification, annotation and acoustic analysis of speech. Having a modular structure, with supplementary materials for more advanced students, the book also applies laryngeal articulation ideas to phonological analysis (such as phonological contrast, neutralization, phonotactic constraint, etc.). A final section is devoted to the applications of phonetics in various professional fields (speech impairments, forensic linguistics, L2 pronunciation teaching, etc.).

Taking the LAM as the starting point for describing the dynamics of speech production introduces several novelties in the way phonetics and phonology are taught. One of them has to do with the characterization of vowel qualities. Many aspects could be discussed in this respect but we focus on only two of them for space reasons. First, although vowel quality distinctions are primarily oral, the larynx is conceived of as a generator of "inherent" quality (which most often expands over entire syllables). Second, the quadrilateral space mapping based on the traditional dichotomy between a front-back and a high-low lingual dynamics is re-conceptualized as resulting from three different articulatory dynamics [12]. Two of them, fronting and raising, are oral dimensions, controlled by the movements of the tongue and the jaw; the third one, retraction, is pharyngeal to the extent that it is controlled by the aryepiglottic sphincter. This re-conceptualization is an important departure from traditional approaches in which vowel quality distinctions are described as exclusively dependent on tongue vertical and horizontal movements (with the contribution of lip rounding, which obviously remains essential in the LAM as well).

Terminology reflects this change of perspective. For instance, for an Italian learner, [u] and [o] are raised vowels, whereas [ɔ] (like [ɑ]) is susceptible to retraction. The new conceptualization is more intuitive for learners because it is directly related to the physical dynamics of lingual muscular constriction [17]. Moreover, it nicely explains why the number of height distinctions is larger for front than for "back" vowels in natural languages. In the LAM perspective, the reason is that fronting is governed by the jaw (and the tongue), and both have more freedom to move than the laryngeal articulator, whose action unfolds (or folds) in the lower vocal tract.

The LAM has a direct impact on how the vocal tract itself is conceptually represented. A first departure from traditional approaches is that the rear of the tongue is part of the lower vocal tract. Various phenomena implying larynx constriction nicely demonstrate this. For instance, the so-called [+ATR]/[-ATR] vowel distinction that is typical of (but not limited to) Western African languages is a good example. Instead of focusing on lingual dynamics only, a LAM-based explanation is focused on the co-occurrence of tongue movements with laryngeal constriction and vertical movements (leading to pharyngeal cavity expansion or restriction). The tense/lax vowel distinction typically implemented in Germanic languages such as English and the mid-high/mid-low vowel distinction (as is typically referred to for languages such as Italian) can, consequently, be explained within a larger and much more informative picture. In such a picture, a fundamental part is played by the observation of the existing synergies between articulators. Focusing on synergies rather than on parcellization could be wrongly interpreted as a "complication" from the

pedagogical point of view. Quite the contrary, it is the most realistic way of explaining how speech production works; and realism has the advantage (among various others) of raising the level of intuitiveness in the explanation of many speech phenomena.

While most examples are taken from Italian and other typologically close languages such as English or French, given the audience of the textbook, some sections are developed with data from languages spoken in regions of Asia, Africa, Oceania or the Americas. A relatively large cross-linguistic scope in exemplification is one of the consequences of using the LAM in the teaching of phonetics and phonology. This may be challenging in introductory textbooks but it is certainly highly rewarding in terms of students' awareness of phonetic diversity. For instance, the observed relationship that exists between laryngeal states (constricted vs. unconstricted larynx), phonatory qualities, and pitch variations, can be neatly exemplified by making reference to Bai tonal registers as presented in [11]. The LAM coherently accounts for how larynx constriction, tone and voice quality interact in natural languages; Bai examples are a fully-fledged exemplification of this interaction and can therefore be usefully integrated into a phonetics textbook. The supplementary materials linked with [15], available at: https://drive.google.com/drive/folders/1-Te5DhQ6NApJoDXJOePR5c_HUOBLJsec, further offer a range of video examples for teachers to use in their phonetics and phonology classes.

4. PEDAGOGY OF VOICE AND CLINICAL PHONETICS

In common parlance, voice is a generic term that can have different meanings: i) sometimes used as a synonym for speech, a confusion frequently made by patients consulting for dysphonia or dysarthria or even by doctors; ii) voice can also refer to individual characteristics of vocal behaviour, timbre or prosody; iii) to vocal categories in linguistic systems or; iv) it can refer to phonation, i.e., to the mode of production corresponding to the vibration of the vocal folds, which is a classical but very reductive definition especially in the medical field.

In the realm of pedagogy, the understanding of vocal production systems based on the LAM model is a major contribution in terms of objective description of physiology. The teaching of the complexity of the structures constituting the larynx in the lower vocal tract as an articulator of voice and speech should be part of the learning process in speech therapy and in the medical field, but also for voice pedagogues, singing teachers, voice professionals and all persons interested in vocal quality.

Clinical phonetics is the interface between experimental phonetics and the medical sciences. Synchronized multi-instrument exploration of the oral and laryngeal vocal tracts facilitates the visualization and understanding of the synergy and dynamics of speech production.

The three main objectives of including the LAM in voice and speech pedagogy are: i) to understand typical productions and atypical variations; ii) in pathology to analyze the mechanisms that cause deviant phonatory behaviours or lesions and to anticipate the consequences of surgical treatments; and iii) to optimize the therapeutic choices and rehabilitative possibilities to guide towards an efficient compensation or adaptation [15, ch. 7].

Some examples of the interdependence of the structures of the epilarynx include dysphonia related to an alteration of vocal behaviour due to misuse or overuse involving the glottic plane and having repercussions on the ventricular plane and even involving the aryepiglottic plane [9, 16]. Similarly, after laryngeal surgery [5, 7-8], when the lower part of the epilarynx is removed, the creation of a neovibrator implies the recruitment of the aryepiglottic structures and the base of the tongue, making the oral and laryngeal vocal tract inseparable. Beyond pathology, atypical productions as in certain songs [10] or vocal techniques [28] will also have recourse to the use of the epilarynx by putting in play the aryepiglottic articulator.

The extraordinary human potential for plasticity and agility of the vocal tract leads to the possibility of relearning, adaptation and compensation of the mechanisms of phonatory production. The pedagogy of the LAM contributes to an understanding of the processes of speech production and the mechanisms of its normal and pathological variability and as a guide to the best therapeutic adjustments.

5. HELPING LEARNERS CONCEPTUALLY EXPAND THE LARYNX WITH COMPUTATIONAL MODELLING

Finally, we move to the area of speech modelling. As mentioned above, the vocal folds dominate the attention of researchers and pedagogues, which has in many cases literally motivated the removal of the epilaryngeal structures. This is true both in experimental studies with excised larynges (e.g., [2]) and computational modelling studies (whether 2D lumped element, e.g., [25], or 3D finite element, e.g., [18], formulations are used).

This is not to say that epilaryngeal structures are entirely absent from such research, as they have been the focus of, or have been incorporated into, various experimental (e.g., [1, 19]) and modelling (e.g., [23]) studies. It is also not always practical to include or model every detail of a system under consideration, especially one as complex as the larynx. Removal of the epilaryngeal structures provides a clearer view of the vocal folds in studies using excised larynges, and the omission of these structures in computational modelling studies allows for savings in the cost of designing, implementing, and processing required for conducting simulations. However, from the perspective of the LAM, the absence of the epilaryngeal structures in these studies is regarded as a considerable compromise. In terms of source-filter theory, the epilaryngeal structures contribute a wealth of source generators, including vibratory action of the ventricular folds, aryepiglottic folds, epiglottis, and all of the internal vestibular mucosa spanning these structures and channels, constrictions and baffles that generate turbulent (noise) sources. The tubular spaces and side-pockets characterizing the epilaryngeal spaces provide important resonances for voice production (e.g., the "singing formant" owes its provenance to the epilarynx; see [26]) and underlies the complexity where source-filter theory breaks down (necessitating the non-linear source-filter theory outlined by Titze [27]).

Our own computational modelling work, some of which has a pedagogical orientation, endeavours to draw attention to the whole larynx, integrating the epilarynx into the design of the models as an unelidable component. The 3D-LCM (Laryngeal Constrictor Mechanism) was an early model that provided a "visual synthesis" of the articulatory and vibratory possibilities of the entire larynx [20-21].

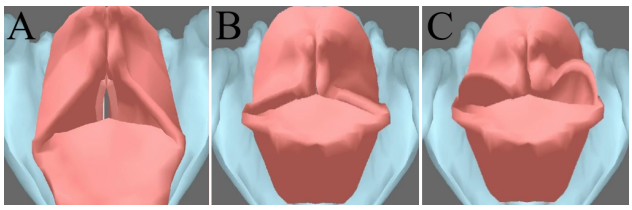


Figure 1: Three frames from the 3D-LCM model showing (A) modal phonation, (B) aryepiglotto-epiglottal stop, and (C) aryepiglottic trilling.

Fig. 1 shows three illustrative frames from this pedagogical tool, which can be manipulated to provide any desired view of the larynx (here we see the larynx as in a laryngoscopic view). Students are able to clearly visualize how the unconstricted laryngeal posture for modal phonation (A) compares to the constricted posture (B) where the aryepiglottic folds have advanced to meet the retracting epiglottis and may engage in aryepiglottic trilling (C). The instructor benefits from the ability to demonstrate the actions of the laryngeal articulator dynamically and three-dimensionally. Computer simulations can be complemented with actual physical sculpture of laryngeal structures to help students understand the anatomy in addition to its actions.

Investigative computational modelling studies of epilaryngeal vibration have been conducted that demonstrate the pulse amplitude-modulating effect that aryepiglottic vibration has on vocal fold vibration [15, ch. 3]. We have also demonstrated that the ventricular folds, through their descent and contact with the vocal folds (a manoeuvre we call vocal-ventricular fold coupling, or VVFC) – evident in ultrasound and MRI data [15] – impacts the vibratory dynamics of the vocal folds, possibly helping to yield creakiness in phonation and aiding in glottal stop production. Finally, we have explored 3D finite-element modelling as a tool to discover the degree of biomechanical quantality (in the sense of Stevens [24]) that exists within the larynx thanks in large part to the presence of the epilaryngeal structures [22]. Apart from video demonstrations, these models can be used in the classroom in an interactive fashion similar to the 3D-LCM, including dynamic simulation if the computational demands of the model setup are manageable enough to be run at a reasonable frame rate.

Because models have a visual appeal and provide a simplified means of understanding the behaviour of complex structures and systems, it is important for both laryngeal research and pedagogy that every effort be made to acknowledge and incorporate the epilaryngeal structures. Doing so will help us to move beyond the conceptualization of the larynx as only a pair of vocal folds.

6. RECAPITULATION

This paper has provided specific examples of some of the pedagogical benefits of the Laryngeal Articulator Model. We hope the readers find these useful in their own teaching practices and that their students will benefit from the supplementary video materials accompanying [15], which are freely available at: https://drive.google.com/drive/folders/1-Te5DhQ6NApJoDXJOePR5c_HUOBLJsec

7. REFERENCES

[1] Alipour, F. and Finnegan, E.M. 2013. On the acoustic effects of the supraglottic structures in excised larynges. *Journal of the Acoustical Society of America* 133, 5, 2984–2992.

[2] Alipour, F., Finnegan, E.M. and Jaiswal, S. 2013. Phonatory characteristics of the excised human larynx in comparison to other species. *Journal of Voice* 27, 4, 441–447.

[3] Benner, A. 2009. *Production and Perception of Laryngeal Constriction in the Early Vocalizations of Bai and English Infants*. PhD dissertation, University of Victoria.

[4] Benner, A., Grenon, I. and Esling, J.H. 2007. Infants' phonetic acquisition of voice quality parameters in the first year of life. *Proc. 16th ICPhS* Saarbrücken, 2073–2076.

[5] Biacabe, B., Crevier-Buchman, L., Laccourreye, O., Hans, S. and Brasnu, D. 2001. Phonatory mechanisms after vertical partial laryngectomy with glottic reconstruction by false vocal fold flap. *Annals of Otology, Rhinology & Laryngology* 110, 935–940.

[6] Calamai, S. and Celata, C. In prep. *Manuale di scienze del parlato*. Carocci.

[7] Crevier-Buchman, L., Laccourreye, O., Weinstein, G., Garcia, D., Jouffre, V. and Brasnu, D. 1995. Evolution of speech and voice following supracricoid partial laryngectomy. *Journal of Laryngology & Otology* 109, 410–413.

[8] Crevier-Buchman, L., Maeda, S., Brasnu, D. and Vaissière, J. 2003. Perceptual and acoustic correlation in consonant identification after partial laryngectomy. *Proc. 15th ICPhS* Barcelona, 2361–2364.

[9] Crevier-Buchman, L., Mattei, A. and Giovanni, A. 2019. Forçage vocal. *EM Consulte Oto-Rhino-Laryngologie*, 20-752-B-10.

[10] Crevier-Buchman, L., Pillot-Loiseau, C., Rialland, A., Vincent Narantuya, C. and Desjacques, A. 2012. Analogy between laryngeal gesture in Mongolian Long Song and supracricoid partial laryngectomy. *Clinical Linguistics and Phonetics* 26, 86–99.

[11] Edmondson, J., Esling, J.H. and Li, S. 2021. Jianchuan Bai. *Journal of the International Phonetic Association* 51, 3, 490–501.

[12] Esling, J.H. 2005. There are no back vowels: The Laryngeal Articulator Model. *The Canadian Journal of Linguistics / La revue canadienne de linguistique,* 50, 13–44.

[13] Esling, J.H. 2009. The control of laryngeal constriction and the emergence of speech in infants in the first year of life. In: Fant, G., Fujisaki, H. and Shen, J. (eds), *Frontiers in Phonetics and Speech Science: Festschrift for Prof. Wu Zongji's 100th Birthday*. The Commercial Press, 191–203.

[14] Esling, J.H. 2012. The articulatory function of the larynx and the origins of speech. *Proc. Annual Meeting Berkeley Linguistics Society* 38, Berkeley (LSA eLanguage).

[15] Esling, J.H., Moisik, S.R., Benner, A. and Crevier-Buchman, L. 2019. *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press.

[16] Esling, J.H., Zeroual, C. and Crevier-Buchman, L. 2007. A study of muscular synergies at the glottal, ventricular and aryepiglottic levels. *Proc. 16th ICPhS* Saarbrücken, 585–588.

[17] Honda, K. 1996. Organization of tongue articulation for vowels. *Journal of Phonetics* 24, 39–52.

[18] Hunter, E.J., Titze, I.R. and Alipour, F. 2004. A three-dimensional model of vocal fold abduction/adduction. *Journal of the Acoustical Society of America* 115, 4, 1747–1759.

[19] Kimura, M., Sakakibara, K.-I., Imagawa, H., Chan, R., Niimi, S. and Tayama, N. 2002. Histological investigation of the supra-glottal structures in human for understanding abnormal phonation. *Journal of the Acoustical Society of America* 112, 2446.

[20] Moisik, S.R. 2008. *A Three-Dimensional Model of the Larynx and the Laryngeal Constrictor Mechanism: Visually Synthesizing Pharyngeal and Epiglottal Articulations Observed in Laryngoscopy*. Masters thesis, University of Victoria.

[21] Moisik, S.R. and Esling, J.H. 2007. 3D auditory-articulatory modeling of the laryngeal constrictor mechanism. *Proc. 16th ICPhS* Saarbrücken, 373–378.

[22] Moisik, S.R. and Gick, B. 2017. The quantal larynx: The stable regions of laryngeal biomechanics and implications for speech production. *Journal of Speech Language and Hearing Research* 60, 3, 540–560.

[23] Sakakibara, K.-I., Imagawa, H., Konishi, T., Kondo, K., Murano, E.Z., Kumada, M. and Niimi, S. 2001. Vocal fold and false vocal fold vibrations in throat singing and synthesis of khöömei. *Proc. International Computer Music Conference* La Habana, 135–138.

[24] Stevens, K.N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17, 3–45.

[25] Story, B.H. and Titze, I.R. 1995. Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America* 97, 2, 1249–1260.

[26] Sundberg, J. 1974. Articulatory interpretation of the 'singing formant'. *Journal of the Acoustical Society of America* 55, 4, 838–844.

[27] Titze, I.R. 2008. Nonlinear source–filter coupling in phonation: Theory. *Journal of the Acoustical Society of America* 123, 5, 2733.

[28] De Torcy, T., Clouet, A., Pillot-Loiseau, C., Vaissiere, J., Brasnu, D. and Crevier-Buchman, L. 2014. A videofiberscopic study of laryngopharyngeal behaviour in the human beatbox. *Logopedics Phoniatrics Vocology* 39, 1, 38–48.

VOICE QUALITY IN NORWEGIAN: EFFECTS OF DIALECT, SPEAKER AGE, SEX, AND SPEAKING MODE

Wim A. van Dommelen

Department of Language and Literature, NTNU, Trondheim, Norway
wim.van.dommelen@ntnu.no

ABSTRACT

This study investigates the effect of Norwegian speakers' dialectal background, age, and sex as well as speaking mode on voice quality. Recordings of read and semi-spontaneous speech were used to analyze short and long /a/ vowels produced by younger (< 40 years) and older (> 40 years) female and male speakers from four main dialectal variants. Acoustic analysis comprised both spectral tilt and harmonicity measures. The analysis revealed that the factor dialect did not have a robust impact on any voice quality measure. Speaker sex affected both types of measures more than any of the other factors, females having steeper spectral tilt and higher harmonicity values than men. Only the spectral measures differed with age, older speakers demonstrating steeper slopes than younger subjects. For spontaneous speech, somewhat steeper slopes and higher harmonicity values were observed than for read speech. Current results conflicting with previous evidence are discussed.

Keywords: Voice quality, Norwegian dialects, speaker age, sex, speaking mode

1. INTRODUCTION

The goal of this study is to explore the relative contribution of Norwegian speakers' dialectal background, age, and sex as well as speaking mode to voice quality. Previous research on dialect-specific voice quality is scarce. Additionally, the presented evidence is not always conclusive. Investigating breathiness in normal female speech, Henton & Bladon [13] measured H1 and H2 in speakers of Received Pronunciation and Modified Northern English. The effect of dialectal accent was not evaluated, but it was small and probably not robust. Szakay [28] investigated phonation in spontaneous speech produced by speakers of Maori English and Pakeha English. H1-H2 values for the former group were found to be generally lower, but no values are specified. Esling & Wong [6] characterize Norwegian voice quality as whispery creaky, but they do not mention any dialectal differences.

The picture emerging from research on the effect of age on phonation is not consistent. While Lee et al. [18] observed H1-H2 and H1-A1 to be smaller for elder than for younger speakers, no significant effect for H1-A1 was found by Gorham-Rowan & Laures-Gore [10]. The latter reported higher noise levels for older speakers (similarly Jin et al. [14], Lortie et al. [19]; see also the meta-analysis by Rojas, Kefalianos & Vogel [24]). In contrast, no age-related changes in noise levels were found by Awan [1] and Goy et al. [11].

Likewise, reported results regarding the impact of speaker sex are confusing. Across publications, all measures apart from H1-H2 (indicating steeper slopes for females) and harmonicity (indicating less noise for females) produced inconclusive evidence. Larger H2-H4 values for females were found by Garellek et al. [9], the opposite by Hejná et al. [12]. The latter measured no difference for H1*-A1*, but lower female CPP compared to higher female CPP values found by Garellek & Keating [8]. Data presented by Mezzedimi et al. [21] and Schaeffer, Knudsen & Small [25] seem to show no impact of speaker sex on HNR / NHR, but no statistical tests were performed.

The two speaking modes investigated in the current study are read speech and semi-spontaneous speech. Studies including the influence of speaking mode on phonation seem to be extremely rare. Participants in Lortie et al. [19] produced a sustained vowel /a/ and retold popular story tales. Acoustic analysis revealed HNR to be lower in connected speech. It is important to note that the latter type of speech contained both vowels and voiced consonants. Starr [26] analyzed Japanese so-called sweet and non-sweet voice performances revealing higher values for sweet voice for H1-H2, H1-A1, and H1-A3. The results for HNR were not consistent across speakers, and no consistent difference was found for CPP.

Due to the scarce and partially contradictory evidence on factors influencing voice quality, the present study has a heuristic character. The two main research questions are:

- What is the effect of speakers' dialectal background, age, sex, and speaking mode (read vs. semi-spontaneous speech) on spectral tilt and harmonicity measures?
- What is the effect size of each of these factors?

2. METHOD

2.1. Speech material

Materials used for this investigation were chosen from the speech database NB Tale [22] provided by the National Library of Norway. The corpus contains annotated recordings of 240 native speakers of Norwegian. This material is divided according to the speakers' dialectal background into 12 subgroups containing 20 speakers each. For the present purposes four homogeneous groups of 20 speakers each were selected, representing the four main dialects: Northern Norwegian (henceforth: North), Central Norwegian (Central), Western Norwegian (West), and Eastern Norwegian (East). In addition to dialect, speaker age (18-40 yrs vs. 41-80 yrs) and sex (female vs. male) are included as systematic factors in the database. Within each subgroup, there are thus five younger and five older speakers of each sex. All recordings were made in a sound-treated environment using 48 kHz sampling frequency and 16-bit quantization. Recordings were in stereo using a Sennheiser 2-1-5 headset and a Shure KSM 44 studio microphone. Since the frequency characteristic of the latter is dependent on the distance from the speaker to the microphone, for the present acoustic analysis the Sennheiser headset channel was selected.

The NB Tale database contains read sentences and semi-spontaneous speech. The read material consists of 20 sentences per speaker. While the first three sentences are identical for all speakers, the remaining 17 have different content across speakers. From the sentence material, stressed /a(:)/ vowels were selected, five from sentences 1-3 that were the same for all speakers and varying numbers of /a(:)/ vowels that were contained in sentences 4-20. The spontaneous speech material contains recordings of participants speaking approximately two minutes about a topic of their own choice. From this material, /a(:)/ vowels and (to increase the number of available tokens) some /æ(:)/ vowels were selected. Our aim was to analyze a total number of 20 spontaneously produced vowel tokens per speaker, but for three out of 80 subjects fewer than 20 tokens were available.

2.2 Acoustic analysis and further processing

Using a Praat [3] script, vowel tokens were extracted from the speech recordings. The freely available tool PraatSauce [15] was used for spectral analysis. Five spectral tilt parameters and five harmonicity measures were calculated: H1*-H2*, H2*-H4, H1*-A1*, H1*-A2*, and H1*-A3*; and CPP (Cepstral Peak Prominence), HNR05, HNR15, HNR25, and HNR35; all in dB. Asterisks denote that PraatSauce outputs for spectral magnitudes of H_i and A_i were corrected for the effect of formant frequencies and bandwidths. (The tool does not correct H4.) To minimize possible effects of phonetic context on measurements, the central 20 ms portion of each vowel was selected for subsequent evaluation. Since an analysis frame rate of 1 ms was used, this corresponds to 20 frames per vowel token for all extracted measures. To eliminate analysis errors, further selection was performed in two steps.

a) All cases with erroneous f0 values of 0 Hz were excluded from the raw data.
b) Subsequently, formant values for all /a(:)/ vowels had to comply with mean values +/- 2 sd as measured for females and males in NB Tale.

Consequently, for a number of vowel tokens there were less than 20 frame values. Remaining frame values were used to calculate mean values for each of the measured parameters. The total number of tokens that went into statistical analysis was 4105 (read speech: 2712; spontaneous speech: 1393).

2.3 Statistical analysis

Statistical analysis of the data was performed with the R program's package lme4 [23] calculating Linear Mixed Effects Models (LMEM). LMEMs included Dialect (North, Central, West, East), Age (< 40, > 40), Sex (female, male) and Mode (read, spontaneous) as fixed factors with by-subject random slopes and intercepts for the factor Mode [2]. Following calculation of each LMEM, the step function developed by Kuznetsova, Brockhoff & Christensen [16] was used to perform backward elimination of nonsignificant factors. Likelihood ratio tests were run in each case comparing a model with n factors with a model having n-1 factors (i.e., without the factor under scrutiny).

Results revealed that the factor Dialect never reached statistical significance, but the other three fixed factors (Age, Sex and Mode) often did. Following statistical analysis, effect size *d* for the levels of the latter three factors was calculated as the absolute difference between two level means divided by the square root of the sum of the variances of the intercept per subject, slope for Mode per subject and the residual variance (Brysbaert & Stevens [4]; Westfall, Kenny & Judd [29]). Calculations of *d* involved both statistically significant and nonsignificant effects.

3. RESULTS

3.1 Main factors

As already stated above, for none of the measures investigated did the factor Dialect reach statistical significance. Between-group differences for the other

factors are presented in Figure 1. As to the effect of Age, for speakers over 40 years significantly higher values than for younger speakers were found for three out of five spectral measures (between-group differences: H1*-H2*: 1.9 dB, H1*-A1*: 1.2 dB, and H1*-A2*: 1.4 dB; $p < 0.001$, $p = 0.025$, and $p = 0.019$, respectively). In contrast, the two age groups did not differ with respect to harmonicity measures. More comprehensive effects were observed for the factor Sex, where except for H2*-H4 female subjects had higher values than males. All spectral measures apart from H1*-A1* as well as all harmonicity measures reached statistical significance (for H2*-H4, $p = 0.007$; for H1*-A2*, $p = 0.031$; else $p < 0.001$). For the spectral measures, significant between-group differences ranged between -0.9 dB for H2*-H4 and 4.9 dB for H1*-A3*; harmonicity measures varied between 3.0 dB for CPP and 10.2 for HNR05. Also speaking mode appeared to affect voice quality in several parameters with values for spontaneous speech generally being higher than for read speech. Three out of five spectral measures were higher for read than spontaneous speech (H2*-H4: 0.3 dB, $p = 0.047$; H1*-A1*: 0.5 dB, $p = 0.003$; and H1*-A3*: 0.7 dB, $p = 0.004$), as well as all four HNR measures (HNR05: 1.1 dB, HNR015: 1.1 dB, HNR25: 1.0 dB, and HNR35: 1.0 dB; in all cases $p < 0.001$).

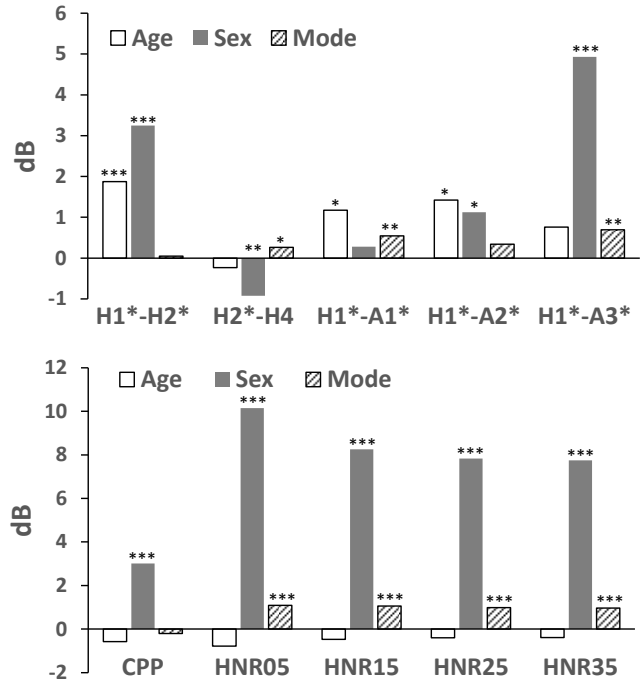


Figure 1: Between-group differences for spectral (top panel) and harmonicity (bottom panel) measures. Age: older (> 40 yrs) - younger (< 40 yrs); Sex: females - males; Mode: spontaneous - read. Probability levels of 0.05, 0.01 and 0.001 are indicated by *, **, and ***, respectively

3.2 Effect sizes

It is evident from Figure 1 that the factor Sex has the largest impact on voice quality, for spectral measures mainly due to H1*-H2* and H1*-A3*. Less important is the factor Age, followed by Mode. To quantify this hierarchy, for each of the acoustic measures effect size *d* was calculated (see Section 2.3). As can be seen from Table 1, across spectral measures, the Sex/Age effect ratio amounted to 0.36/0.21 = 1.8, while the Sex/Mode ratio was even higher (0.36/0.07 = 5.0; rounded values). Among harmonicity measures, the factor Mode had a more consistent impact, and the Sex/Mode *d* ratio was even somewhat lower than the one for Sex/Age (1.25/0.14 = 8.8 vs. 1.25/0.10 = 13.0). Averaged across all acoustic measures, Sex affected voice quality most, followed by Age, with Mode being somewhat less influential (effect ratio for Sex/Age: 0.80/0.15 = 5.3 vs. Sex/Mode: 0.80/0.11 = 7.5).

3.3 Interactions

For reasons of space, this section will only mention some main interaction results. Merely 12 out of 60 two-way interactions turned out to be significant. Five Age x Sex interactions reached significance (for H1*-H2*, and HNR05 – HNR35), all showing larger sex effects for younger vs. older speakers. For four out of five spectral measures, significant Dialect x Sex interactions were observed (H1*-H2*, H1*-A1*, H1*-A2*, and H1*-A3*), in addition to CPP, all five being seemingly incidental. Additionally, one Sex x Mode (H2*-H4), and one Dialect x Mode interaction (HNR05) reached statistical significance.

4. DISCUSSION

The present study investigated the effect of Norwegian speakers' dialectal background, age, and sex as well as speaking mode on voice quality. Acoustic measures characterizing voice quality comprised spectral tilt parameters H1*-H2*, H2*-H4, H1*-A1*, H1*-A2*, H1*-A3* and harmonicity parameters CPP, HNR05, HNR15, HNR25, and HNR35.

4.1 Dialect

Altogether, the picture emerging from the measurements was relatively consistent. For a start, the factor dialect appeared to have no systematic impact on any voice quality measure. This result seems to be in line with the small difference in H1-H2 for two English dialects found by Henton & Bladon [13]. It could be speculated that voice quality

Table 1: Effect size *d* for spectral measures H1*-H2* to H1*-A3* and harmonicity measures CPP to HNR35. Calculations of *d* involved both statistically significant and nonsignificant effects.

Factor	H1*-H2*	H2*-H4	H1*-A1*	H1*-A2*	H1*-A3*	mean	CPP	HNR 05	HNR 15	HNR 25	HNR 35	mean
Age	0.35	0.07	0.25	0.26	0.10	**0.21**	0.15	0.11	0.08	0.07	0.07	**0.10**
Sex	0.60	0.27	0.06	0.21	0.68	**0.36**	0.80	1.42	1.38	1.32	1.31	**1.25**
Mode	0.01	0.08	0.12	0.06	0.10	**0.07**	0.05	0.15	0.18	0.17	0.16	**0.14**

differences may not – or at best rarely – be exploited between dialects, as opposed to phonemically-contrastive phonation types distinguishing registers (DiCanio [5]) or consonants and vowels (e.g., Esposito, Sleeper & Schäfer [7]).

4.2 Speaker age

As to the effect of age on voice quality, the present finding of steeper spectral slopes for older speakers (particularly H1*-H2*) would suggest more breathiness in that age category. This outcome is at odds with the data for Korean female speakers in Lee et al. [18], who found steeper slopes indicated by H1-H2 and H1-A1 for younger women. The authors speculate that the relatively high formant frequencies for Korean /a/ might be partially responsible for their results and that the use of corrected measures could contribute to the interpretation of their results. The current absence of a significant effect of age on harmonicity measures is at variance with the evidence of increased noise in older compared to younger voices in previous investigations ([14], [19], [21], [25]). It seems possible to explain the discrepancies in harmonicity between the present and previous studies by differences in age ranges. The data for SNR presented in Stathopoulos, Huber & Sussman ([27]: Figure 3) show variation but no clear rising or falling tendency for speakers aged 20 – 70 years. For older speakers (70 – 93 years), SNR decreases with age. Whereas in the investigations cited above, age distributions typically range between 20 and 85 years, the current age range was 18 – 71 years, thus excluding ages where decreasing harmonicity could be expected. As we have seen above, our speakers did show an effect of age on spectral slopes. This indicates that spectral slopes and harmonicity measure different acoustic properties of the glottal waveform, breathiness and roughness or hoarseness, respectively (Hejná et al. [12], Latoszek et al. [17], Yumoto, Gould & Baer [30]).

4.3 Speaker sex

With some exceptions, the present effect of speaker sex on acoustic measures parallels previous findings. All previously mentioned studies reported steeper H1*-H2* and H1*-A3* slopes for females. The current higher H2*-H4 values for male speakers are in line with Hejná et al. [12] and Matar et al. [20] but at odds with Garellek et al. [9]. Even less consistent are the results for women's lower H1*-A1* in Gorham-Rowan & Laures-Gore [10] and H1*-H2* and H1*-A1* values in Garellek & Keating [8]. These incongruities may be explained by differences in experimental methodologies and technical conditions. In this connection it is noteworthy that the less robust measures in previous investigations achieved the smallest sex-specific effects in the present study. As to harmonicity (HNR) measures, largely confirming results have been found, women's voices containing less noise than men's.

4.4 Speaking mode

As shown by the current data, also speaking mode had a certain impact on voice quality. In spontaneous speech, three spectral tilt measures (H2*-H4, H1*-A1*, and H1*-A3*) and all HNR measures were larger than in read speech. The effects were relatively small but consistent. The fact that Lortie et al. [19] observed an opposite effect for HNR can be due to their use of portions of spontaneous speech, containing both vowels and consonants. Particularly the presence of obstruents may have resulted in lower HNR values compared to isolated vowels.

5. CONCLUSION

The two research questions formulated in the introduction can be answered as follows:

a) Speaker age and sex as well as speaking mode have been demonstrated to affect both spectral tilt and harmonicity measures. Dialectal background does not exert any systematic effect.

b) In the spectral measures, speaker sex had the greatest impact, followed by speaker age, and mode had by far the weakest effect. As to harmonicity, speaker sex had a more marked effect than the other two factors. While both mode and speaker age were far less important, the latter even failed to reach statistical significance.

6. REFERENCES

[1] Awan, S. N. 2006. The aging female voice: Acoustic and respiratory data. *Clinical Linguistics & Phonetics* 20(2/3), 171-180.

[2] Barr, D.J., Levy, R., Scheepers, C., Tily, H.J. 2013. Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 255-278.

[3] Boersma, P., Weenink, D. 2021. *Praat: Doing phonetics by computer* [Computer program]. Version 6.1.48. http://www.praat.org/.

[4] Brysbaert, M., Stevens, M. 2018. Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition* 1(1): 9, 1-20.

[5] DiCanio, C. T. 2009. The phonetics of register in Takhian Thong Chong. *J. Int. Phon. Ass.* 39 (2), 162-188.

[6] Esling, J. H., Wong, R. F. 1983. Voice quality settings and the teaching of pronunciation. *TESOL Quarterly* 17, 89-95.

[7] Esposito, C. M., Sleeper, M., Schäfer, K. 2021. Examining the relationship between vowel quality and voice quality. *J. Int. Phon. Ass.* 51(3), 361-392.

[8] Garellek, M., Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *J. Int. Phon. Ass.* 41(2), 185-205.

[9] Garellek, M., Samlan, R., Gerratt, B. R., Kreiman, J. 2016. Modeling the voice source in terms of spectral slopes. *J. Acoust. Soc. Am.* 139(3), 1404-1410.

[10] Gorham-Rowan, M. M., Laures-Gore, J. 2006. Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders* 39(3), 171-184.

[11] Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., van Lieshout, P. 2013. Normative voice data for younger and older adults. *Journal of Voice* 27, 545-555.

[12] Hejná, M., Šturm, P., Tylečková, L., Bořil, T. 2021. Normophonic breathiness in Czech and Danish: Are females breathier than males? *Journal of Voice* 35(3), 498.e1-498.e22.

[13] Henton, C. G., Bladon, R. A. W. 1985. Breathiness in normal female speech: Inefficiency versus desirability. *Language & Communication* 5(3), 221-227.

[14] Jin, S. M., Kwon, K. H., Kang, H. G. 1997. Acoustic and stroboscopic characteristics of normal person's voices with advancing age. *Journal of The Korean Society of Laryngology, Phoniatrics and Logopedics* 8, 44-48.

[15] Kirby, J. P. 2021. [Spectral analysis tool] https://github.com/kirbyj/praatsauce.

[16] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1-26.

[17] Latoszek, B. B. v., Maryn, Y., Gerrits, E., De Bodt, M. S. 2018. A Meta-Analysis: Acoustic Measurement of Roughness and Breathiness. *J. Speech Lang. Hear. Res.* 61(2), 298-323.

[18] Lee, S. J., Cho, Y., Song, J. Y., Lee, D., Kim, Y., Kim, H. 2015. Aging Effect on Korean Female Voice: Acoustic and Perceptual Examinations of Breathiness. *Folia phoniatrica et logopaedica* 67(6), 300-307.

[19] Lortie, C. L., Thibeault, M., Guitton, M. J., Tremblay, P. 2015. Effects of age on the amplitude, frequency and perceived quality of voice. *Age* 37(6), 117.

[20] Matar, N., Portes, C., Lancia, L., Legou, T., Baider, F. 2016. Voice quality and gender stereotypes: A study of Lebanese women with Reinke's edema. *Journal of speech, language, and hearing research* 59(6), S1608-S1617.

[21] Mezzedimi, C., Di Francesco, M., Livi, W., Spinosi, M. C., De Felice, C. 2017. Objective evaluation of presbyphonia: Spectroacoustic study on 142 patients with Praat. *Journal of Voice* 31(2), 257.e25-257.e32.

[22] NB Tale – en grunnleggende akustisk fonetisk taledatabase for norsk [NB Speech – a fundamental acoustic-phonetic speech database for Norwegian]. http://www.nb.no/sprakbanken/show?serial=sbr-31&lang=nb

[23] R Core Team. 2021. R: A language and environment for statistical computing. Retrieved from https://ww.R-project.org/.

[24] Rojas, S., Kefalianos, E., Vogel, A. 2020. How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research* 63(2), 533-551.

[25] Schaeffer, N., Knudsen, M., Small, A. 2015. Multidimensional voice data on participants with perceptually normal voices from ages 60 to 80: A preliminary acoustic reference for the elderly population. *Journal of Voice* 29(5), 631-637.

[26] Starr, R. L. 2015. Sweet voice: The role of voice quality in a Japanese feminine style. *Language in Society* 44, 1-34.

[27] Stathopoulos, E. T., Huber, J. E., Sussman, J. E. 2011. Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age. *J. Speech Lang. Hear. Res.* 54(4), 1011-1021.

[28] Szakay, A. 2012. Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. *Journal of Sociolinguistics* 16(3), 382-397.

[29] Westfall, J., Kenny, D. A., Judd, C. M. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *J. Exp. Psych.: General* 143(5), 2020-2045.

[30] Yumoto, E., Gould, W. J., Baer, T. 1982. The harmonics-to-noise ratio as an index of degree of hoarseness. *J. Acoust. Soc. Am.* 71, 1544–1550.

MANDARIN PROSODIC FOCUS BY SPEAKERS WITH AUTISM SPECTRUM DISORDERS

Ho-hsien Pan[1], Shaoren Lyu[1], Chin-po Chen[2], Hai-ti Lin[3], Susan Shur-fen Gau[3]

[1]National Yang Ming Chiao Tung University, [2]National Tsing Hua University, [3]National Taiwan University Hospital
hhpan@nycu.edu.tw, shaorenlyu@gmail.com, stu9116@gapp.nthu.edu.tw, htdaisy125@gmail.com, gaushufe@gmail.com

ABSTRACT

The production of prosodic stress is a potential source of communication difficulties experienced by individuals with autism spectrum disorder (ASD). To investigate the acoustic cues used by individuals with ASD to distinguish given information from narrowly focused new information and to enhance narrow focus during post-focus syllables, read and spontaneous speech corpora were analyzed for both typically developing (TD) and ASD speakers of Taiwan Mandarin. The results revealed that both groups utilized duration lengthening and initial creakiness to mark narrow focus. However, duration shortening, which is a post-focus compression (PFC) cue used by TD speakers, was not used by ASD speakers. Additionally, initial glottalization with tense voice quality, as indicated by low H1-A3c values, was found to be significantly correlated with ASD speakers' memory and executive function scores. These findings suggest that acoustic cues could potentially serve as a noninvasive objective index of ASD.

Keywords: Tense voice quality, duration, executive function, memory, narrow focus.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that often leads to difficulties in communication and interpersonal interactions within a community. The perception and production of prosodic stress on prominent words in utterances are crucial for effective communication. However, individuals with ASD are prone to misperception of stress, particularly when the stress is located early in sentences [1]. Those with ASD who can stress important information are more likely to receive better communication ratings on the Autism Diagnostic Observation Schedule-Generic (ADOS-G) Communication scores [2].

Typically, new and contrastive information that act as the heads of utterances tend to be stressed prosodically. For instance, when asked to produce narrowly focused (NF) new information for words such as "CHOCOLATE" or "ICE CREAM" in contexts like "Did you eat the vanilla ice cream?" "I ate a CHOCOLATE ice cream." or "Did you eat the vanilla cake?" "I ate the vanilla ICE CREAM," individuals with ASD ambiguously distinguish given and NF new information [3, 4, 5].

Previous studies on the production of narrow focus in ASD speech have mainly focused on English. It remains unclear how ASD speakers of native languages with different prosodic typologies, such as Mandarin, express narrow focus. Furthermore, our understanding of the acoustic data used by native Mandarin speakers with ASD to signal NF new information is limited.

Those aims of this study is to investigate the following questions: (1) How do native Mandarin speakers with ASD distinguish NF new information from unfocused given information? (2) How do they enhance NF new information with post-focus compression? (3) Are there significant correlations between acoustic cues for narrow focus and the scores in the Cambridge Neuropsychological Test Automated Battery (CANTAB)?

Works on NF new information in Mandarin were conducted in controlled lab speech produced by typically developing (TD) native Beijing and Taiwan Mandarin speakers. These studies found that duration lengthening and F0 range expansion as the two major acoustic cues for NF syllables [6, 7, 8]. Moreover, Beijing Mandarin speakers lowered and compressed the pitch ranges of post-focus words to enhance the preceding NF new information. However, this post-focus compression (PFC) was not observed among Taiwan Mandarin or Taiwan Min Nan [9].

Although initial vowel glottalization was found to mark prosodic prominence in English [10], the potential use of initial glottalization as a cue for NF in Mandarin never explored. In this study, we aim to investigate this possibility, on top of duration lengthening and F0 range expansion. The two voice quality cues associated with initial glottalization were H1-H2 and H1-A3. Phonation-wise, the lower the H1-H2 values become, the creakier the voice quality is. The lower the H1-A3 values are, the more tense and abrupt the glottal closure becomes.

By comparing the acoustic cues of the same words carrying given versus NF new information or by

comparing the acoustic cues associated with post-given versus post-new information in sequential sentences produced, we explored the focus and post-focus markings by both TD and ASD speakers of Mandarin. Additionally, we reported individual cue weighting in addition to group results to highlight the heterogeneous nature of ASD speakers. Finally, we correlated the effective acoustic measures with cognition scores from the Cambridge Neuropsychological Test Automated Battery (CANTAB).

Based on our hypothesis, we expected that NF new information would be marked by duration lengthening, F0 range expansion, and initial glottalized creaky or tense voice quality. On the other hand, post-focus syllables were predicted to be shorter in duration, narrower in F0 range, and more modal-like in terms of voice quality to enhance preceding NF information.

2. METHODS

2.1. Speakers

Spontaneous speech was collected from 81 participants, and all of whom were of Chinese descent. The participants were divided into eight groups: five TD boys, seven TD adult males, three TD girls, eight TD adult females, seven ASD boys, 44 ASD adult males, two ASD girls and five ASD females. The pre-pubescent ASD and TD speakers were under 13 years of age. The TD participants did not have any history of speech or hearing problems. The ASD participants were clinically diagnosed and confirmed by the Autism Diagnostic Interview-Revised at the National Taiwan University Hospital in Taipei, Taiwan.

2.2. Corpus

Spontaneous speech was collected from both ASD speakers and TD children during the administration of Module 3 of the Mandarin-ADOS-G, a semi-structured clinical instrument designed to assess ASD-related deficits. This module is intended for use with children and adolescents who exhibit fluent spontaneous speech and require the ability to talk about objects or events that are not immediately present. The participants' responses to social-emotional questions on topics such as emotions, friends, loneliness, marriage, social difficulties, and annoyance were recorded using a lapel microphone attached to their collars. During the assessment, the examiners, who were professionally trained therapists or psychiatrists, observed and skipped inappropriate questions that the participants had difficulty answering. They also rated the subjects according to socio-communication codes. The Mandarin-ADOS-G has demonstrated good inter-rater reliability (0.91), test-retest reliability (intraclass correlations 0.55-0.73), and low to high good internal consistency (Cronbach's alpha 0.27-0.86) [11].

The linguistic institute of National Taiwan University recorded the TD adults' spontaneous monologues. Each recording lasted approximately 30 minutes.

2.3. Data transcription

The sound files were transcribed in Praat [12] on multiple tiers, including Chinese orthography, words, syllables, tones, and surface segments. A dual-language forced aligner was utilized to generate segmental boundaries for both Taiwan Mandarin and Min Nan. This aligner was developed through a collaboration with the team that developed Munich Automatic Segmentation System (MAUS) and the Common Language Resources and Technology Infrastructure (CLARIN) [16]. The aligner is available through BAS Web Services [13, 14, 15].

A program developed in the lab was utilized to detect repeated content words in consecutive sentences. The first occurrence of a content word was labelled as new information whereas subsequent repetitions were marked as given information. Similarly, in consecutive sentences, the words following new and given information were labeled as post-new and post-given, respectively.

Next, VoiceSauce [15] was employed to extract acoustic parameters at every 10% vowel intervals during syllables conveying NF new, given, post-new, and post-given information.

2.4. Data analysis

Linear mixed-effect regression models were used to analyze the acoustic measures of syllables carrying NF new information compared to those carrying given information. The models included speakers and syllables as random effects, with given information as the baseline data. Normalized duration, F0 range, and normalized H1-H2c and H1-A3c were compared at initial 10% vowel points. Similarly, the acoustic measures of syllables following NF new information were compared to those following given information using linear mixed-effect regression models, with post-given syllables as the baseline. These models also included speakers and syllables as random effects.

The H1-H2c measures the difference in amplitude between the first and second harmonics, taking into account the resonance effect of the vocal tract. The H1-A3c corrected measures the magnitude difference between the first harmonic and the strongest harmonic of the third formant. The spectral tilt of H1-

H2c reflects the open phase of glottal vibration, with high values indicating breathy voice quality and low values indicating creaky voice quality. H1-A3c reflects the degree of abruptness in glottal closure, with lower values indicating more abrupt glottal closures and tense voices [16, 17].

3. RESULTS

3.1. New vs. Given information

The group results of linear mixed effect regression models comparing given vs. new information showed that TD adult males (β= 0.26***), TD adult females (β= 0.28***), ASD girls (β= 0.46**) and ASD adult males (β= 0.15***) produced NF new information with significantly longer normalized duration than given information.

Among all subject groups, only TD adult females produced a more expanded F0 range for NF new information (β= 0.12***).

Regarding voice quality cues of NF syllables, TD adult males (β= -0.08***), TD adult females (β= -0.06*), and ASD adult males (β= -0.07*) produced significantly lower H1-A3c values, indicating a more tense voice quality during the initial portion of NF syllables. However, the group results for H1-H2c did not show initial creaky voice for NF syllables.

Individual results revealed that five TD adult males, seven TD adult females, eight ASD adult males, and one ASD girl produced narrow focus with significantly longer duration.

Additionally, three TD adult females produced a more expanded F0 range for narrow focus. Furthermore, four ASD adult males produced narrow focus with creakier voice quality.

In summary, the group results revealed that both TD and ASD speakers produced narrow focus with longer duration and a more tense voice quality at the beginning of the syllable. The TD speakers also showed F0 range expansion for narrow focus. However, there were discrepancies between the group and individual results. Individual results showed that narrow focus was marked with longer duration by most speakers, F0 range expansion by a few TD speakers, and initial creaky voice only by some ASD speakers.

3.2 Post-focus vs. post-given syllables

To investigate how PFC enhances NF new information, post-new and post-given syllables were compared. It was hypothesized that post-focus syllables would exhibit significantly shorter duration, narrower F0 range, and less initial glottalization, as indicated by less creaky (high H1-H2c) or less tense (high H1-A3c) voice quality.

Group results showed that both TD adult males (β= -0.10*) and females (β= -0.11**) produced significantly shorter post-new syllables compared to post-given positions. However, individual analysis revealed that only one TD male adult and two TD female adults produced shorter post-focus syllables. No other PFC cues were observed among the eight subject groups.

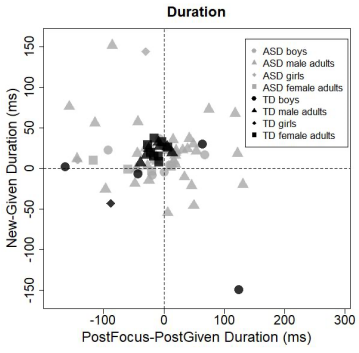
3.3. Individual cue weighting

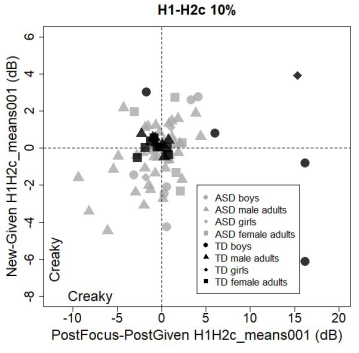
The acoustic differences between syllables carrying given and NF new information were plotted against acoustic differences between post-given and post-new to explore individual cue weightings.

As shown in Figure 1, TD speakers' duration data points were clustered in the second quadrant. In other words, TD speakers produced longer durations during narrow focus and shorter durations during post-focus syllables.

For voice quality, most TD speakers produced more tense (H1-A3c in the third quadrant) but not creakier voice quality (H1-H2c in the first and second quadrants) to mark narrow focus. Post-focus syllables were produced with tense voice qualities to enhance the preceding narrow focus. Post-focus creaky voice was found only among some speakers.

For ASD speakers, the distribution of H1-H2c and H1-A3c data among the four quadrants was heterogeneous. However, it was observed that most ASD speakers produced syllables carrying NF new information with longer duration compared to those carrying given information.





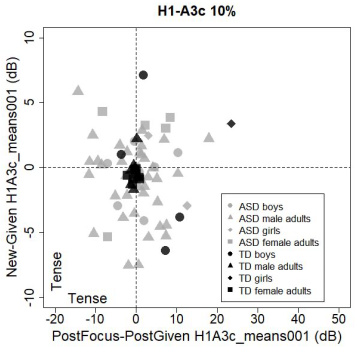


Figure 1: Individual cue weighting of duration, H1-H2c, and H1-A3c differences between new vs. given information and between post-given vs. post-new syllables in ASD and TD speakers.

3.4. Cognitive functions and ASD acoustic cues

In the study of ASD individuals, linear regression models were employed to investigate the correlations between acoustic cues and various measures of memory or executive functions in the cognitive domain, which were assessed using the CANTAB tasks [18].

The correlation analysis revealed that the mean subsequent thinking time (three moves) of the Stocking of Cambridge (SOCst T3) was significantly correlated with the differences in H1-A3c values between syllables carrying new and given information (r=0.66**).

The Stocking of Cambridge (SOC) task is designed to assess visuospatial planning and problem-solving skills as part of the executive functions. During the task, subjects match a set of stockings containing three balls of different colors at the bottom of the screen to a pattern displayed at the top. The SOCstT3 is a measure of the subject's speed of movement, calculated as the difference in time between selecting the first ball and completing the problem in three moves (https://www.cambridgecognition.com/cantab/cognitive-tests/executive-function/stockings-of-cambridge-soc/). Higher SOCstT3 values indicate poorer executive function performance.

The positive correlation between SOCstT3 and H1-A3c during the initial portion of NF syllables suggests that ASD individuals who took longer times to complete the task also produced higher H1-A3c values, indicating a less tense voice quality, during narrow focus. Therefore, it can be concluded that ASD speakers with better executive function tend to produce more tense voice quality during the initial portion of narrow focus. The use of tense voice quality as a NF marking was also observed among the TD speakers.

Apart from the SOC task, the correlation analysis also showed that the H1-A3c values during the initial portion of post-focus syllables were significantly correlated with the total and adjusted total number of presentations required to locate the correct patterns in the Paired Associate Learning task (PALtT & PALtTA: r= -0.45*).

During the Paired Associate Learning (PAL) task, subjects were presented with a sequence of boxes that were opened randomly. The majority of the boxes were empty, except for one or two boxes containing patterns. The subjects were then asked to match the patterns of trials to the correct boxes (https://www.cambridgecognition.com/cantab/cognitive-tests/memory/paired-associates-learning-pal/).

The negative correlations showed that ASD subjects who required fewer trials to complete the PAL task tended to produce higher H1-A3c values, suggesting a less tense voice quality, at the initial portion for post-focus syllables. In other words, subjects who required fewer trials tended to produce a more modal-like voice quality during the initial portion of post-focus syllables to enhance the preceding narrow focus, which was produced with tense voice quality.

In summary, tense voice quality production correlated significantly with the ASD speakers' executive function and visuospatial memory. Those ASD speakers who scored lower in PALtT & PALtTA but higher in executive function tasks (SOCst T3) take less time to memorize the patterns and complete the tasks faster. These ASD individuals also signalled narrow focus with initial tense voice quality and enhanced narrow focus with less tense voice quality during post-focus syllables.

4. DISCUSSION

Both TD and ASD speakers utilized duration lengthening and initial tense voice to mark narrow-focus syllables. However, only TD speakers exhibited post-focus duration shortening following narrow focus, indicating PFC enhancement. ASD speakers showed more variability, however, there was a positive correlation between the executive and visuospatial memory function of ASD speakers and the use of tense voice quality. Overall, these findings reveal the use of duration and tense voice quality mark narrow focus, with some differences in cue weighting between TD and ASD speakers.

5. REFERENCES

[1] Kelley, E. 2011. Language in ASD. In: Fein, D. (ed), *The neuropsychology of autism*, Oxford University Press, 123–137.

[2] Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., Volkmar, F. 2005. Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders. *J. Autism Dev. Disord*. 35(6), 861–869. https://doi.org/10.1007/s10803-005-0031-8

[3] Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A. M., Cohen, D. J, Volkmar, F. R. 2001. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *J. Speech Lang. Hear. Res.* 44(5), 1097–115.

[4] Paul, R., Augustyn, A., Klin, A., Volkmar, F. R. 2005. Perception and production of prosody by speakers with autism spectrum disorders. *J. Autism Dev. Disord*. 35(2), 205–220.

[5] Peppé, S., McCann, J., Gibbon, F., O'Hare, A., Rutherford, M. 2007. Receptive and expressive prosodic ability in children with high-functioning autism. *J. Speech Lang. Hear. Res.* 50(4), 1015–1028.

[6] Jin, S. 1996. *An Acoustic Study of Sentence Stress in Mandarin Chinese.* Ph.D. dissertation. The Ohio State University.

[7] Xu, Y. 1999. Effects of Tone and Focus on the Formation and Alignment of F0 Contours. *Journal of Phonetics*. 27, 55–107.

[8] Li, J. R. 2005. New and contrastive focus in Taiwan Mandarin. *J. Acoust. Soc. Am*. 118 (3). https://doi.org/10.1121/1.4780004

[9] Xu, Y., Chen, S. W., Wang, B. 2012. Prosodic focus with and without post-focus compression: A typological divide within the same language family. *The Linguistic Review* 29(1), 131–147.

[10] Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, 423–444.

[11] Chang JC, Lai MC, Chien YL, Cheng CY, Wu YY, Gau SS. 2023. Psychometric properties of the Mandarin version of the autism diagnostic observation Schedule-Generic. *J Formos Med Assoc.* 23, S0929-6646(23)00008-6. doi: 10.1016/j.jfma.2023.01.008.

[12] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9), 341–345.

[12] Schiel, F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. *Proc. 14th ICPhS* San Francisco, 607–610.

[13] Schiel, F. 2015. A statistical model for predicting pronunciation. *Proc. 18th ICPhS* Glasgow, paper 195.

[14] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

[15] Shue, Y.-L.m Keating, P., Vicenik, C., Yu, K. 2011. VoiceSauce: A program for voice analysis. *Proceedings of the ICPHS XVII,* 1846-1849.

[16] Maddieson, Ian & Peter Ladefoged. 1985. Tense and lax in four minority languages of China. *UCLA Working Papers in Phonetics,* 60, 59–83.

[17] Kuang, J., Keating, J. 2012. Glottal articulations of phonation contrast and their acoustic and perceptual consequences. *UCLA Working Papers in Phonetics,* 111, 123–161.

[18] Goldberg, M.C. 2013. CANTAB. In: Volkmar, F.R. (eds) *Encyclopedia of Autism Spectrum Disorders.* Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1698-3_1915

CREAPY: A PYTHON-BASED TOOL FOR THE DETECTION OF CREAK IN CONVERSATIONAL SPEECH

{Michael Paierl, Thomas Röck}[1] , Saskia Wepner[2], Anneliese Kelterer, Barbara Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology
{paierl, thomas.roeck, saskia.wepner}@student.tugraz.at, anneliese.kelterer@edu.uni-graz.at, b.schuppler@tugraz.at

ABSTRACT

The annotation of creaky voice is relevant for various linguistic topics, from phonological analyses to the investigation of turn-taking, but manual annotation is a time-consuming process. In this paper, we present `creapy`, a Python-based tool to detect creaky intervals in speech signals. `creapy` does not require prior phonetic segmentation and supports the export of Praat TextGrid files, allowing for manual revision of the automatically labelled intervals. `creapy` was developed and tested using Austrian German conversational speech. It was optimised for recall to facilitate a semi-automatic annotation process, and it achieved a better performance for men's (recall: .79) than for women's voices (recall: .60).

Keywords: voice quality, creaky voice, automatic creak detection, conversational speech

1. INTRODUCTION

In this paper, we present `creapy`, a Python-based tool for automatic creak detection. Creak is a phonation type that is characterised by constriction in the glottis [1]. Identifying creaky voice is relevant for a wide range of topics in linguistics [2], such as phonological phonation type contrasts in various languages [3, 4], prosodic phrasing [5, 6], turn-taking strategies [7, 8, 9], hesitation phenomena [10], discourse structure [11], the investigation of emotion and attitude [12, 13, 14] and the sociolinguistic study of social markers [15, 16]. Automatic creak detection facilitates the identification of creak in large amounts of data and, even if manual correction may be required, it may speed up the annotation of data needed for phonetic analyses.

Apart from studying the functions of creak itself, identifying creaky voice is also relevant for an accurate extraction of fundamental frequency (F0) for other types of phonetic studies. In creaky segments, there is a significant increase of errors in the computation of F0 [17]. Either F0 cannot be detected or the detected F0 displays octave jumps due to erroneous identification of periods in the signal. Creak detection can thus support the manual correction of automatically extracted F0 by pointing to portions in the signal in which a faulty detection of F0 is likely.

Given the high relevance of creaky voice in speech science, several tools for its automatic detection have been developed. Ishi et al. [18], for instance, proposed a tool developed for conversational Japanese. Drugman et al. [19, 20] presented a creak detector that was integrated into the MATLAB-based speech processing tool COVAREP. It was developed on read and conversational speech from several languages (i.e., American English, Finnish, Swedish and Japanese). While our approach on detecting creaky voice is fairly similar to theirs, we introduce new features and do not rely fully on those proposed in [19], namely the ones used in [18, 21, 22] (H2-H1, F0, residual peak prominence, power peak parameters, inter-pulse similarity, intra-frame periodicity, energy norm, power standard deviation and zero-crossing rate).

Compared to mentioned tools, `creapy` has the advantage that it is written in the open-source programming language Python [23] and does not need manual speech segmentation. To detect creaky voice, `creapy` reads an audio signal for which it returns a binary creak decision and the underlying probability that yielded that decision for each point in time. `creapy` allows for an uncomplicated adjustment of the decision threshold as well as the parameters that are used to calculate the features. It supports the output of creak intervals in Praat [24] TextGrid files, allowing for an easy integration into a work-flow including a subsequent manual correction step. Finally, as we know that creak tends to occur more frequently in informal settings in some languages [25], we used conversational speech for its development.

2. DEVELOPMENT OF CREAPY

2.1. Materials

`creapy` was trained and tested on four conversations from the Graz corpus of read and spontaneous speech (GRASS) [26]. GRASS' conversational component contains unscripted face-to-face conver-

sations between pairs of Austrian German speakers who knew each other well (e.g., close friends, couples, family members). We found this corpus particularly suitable because we aimed at developing `creapy` on high-quality recordings of spontaneous speech, capturing a broad variety of phenomena that occur in casual conversations.

We used Praat [24] to label creaky and non-creaky intervals in 5 to 10 minutes of 4 conversations (overall 60 minutes) for each of the 8 speakers (4 women and 4 men). Intervals were labelled as being creaky by a combination of auditory analysis and visual inspection of the speech waveform (cf., [27]), spectrogram and F0 contours. For the set of non-creak labels, we chose a variety of vowels and voiced consonants, excluding voiceless fricatives, plosives and noise-like intervals, because creak only occurs in voiced segments. This resulted in 682 creak and 771 non-creak intervals to train and test the tool.

2.2. Methods

`creapy` identifies creaky voice by means of a Random Forest classifier (RFC) [28] trained on data that was labelled for the presence of creak (cf., Section 2.1). For both classification and determining feature importance, RFCs were built with 99 estimators, the default maximum depth, with a minimum samples split of 2, the square root as maximum number of features considered for splitting a node and Gini impurity measure. Given an audio file as input, `creapy` creates a list of intervals classified as creak.

2.2.1. Acoustic Features

We extracted a set of 89 acoustic features: EGEMAPSV02 features from openSMILE [29] and Cepstral Peak Prominence (CPP), which showed to be a good correlate of voice quality in many studies [30]. The features were sorted by relevance with the built-in feature importance of the RFC. The five highest ranked features were the amplitude difference of the first and second harmonic (H1H2), the Harmonics-to-noise-ratio (HNR), the fluctuation of the periodicity in time (Jitter) and amplitude (Shimmer), and the mean of the fundamental frequency (F0-mean). These features were computed for each labelled interval of speech in the training set (cf., Section 2.1). In a minority of cases, features that rely on an accurate F0 detection could not be calculated. Due to the RFC's inability to handle missing values, those had to be imputed, i.e., replaced with an actual numeric value. We chose the median of all valid results of the feature in question. This imputation step is valid for our application because `creapy` pre-eliminates unvoiced intervals prior to the actual creak detection (cf., Section 2.2.3).

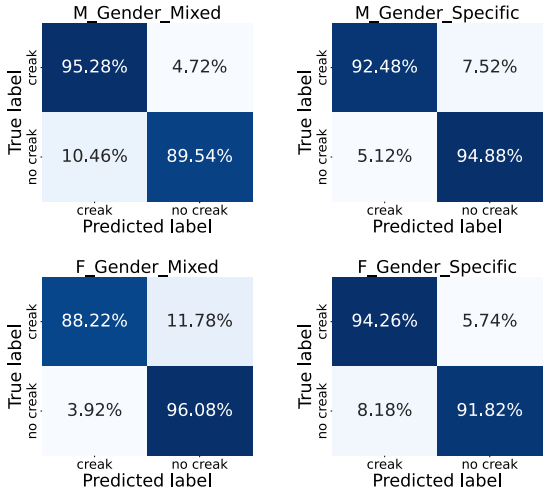


Figure 1: Confusion matrices of creak classification, averaged over speakers of the same gender with the gender-mixed models (left) and the gender-specific models (right).

2.2.2. Evaluation of the Feature Set

75% of the data was used for training and 25% for testing while performing a cross-validation procedure on all speakers. The distribution of labelled creaky and non-creaky intervals was not balanced for each speaker. Speakers were evaluated for two different classification models, one trained on speakers of mixed gender (GenMix) and one trained on speakers of the same gender (GenSpec, gender-specific). Figure 1 shows the confusion matrices of the cross-validation for each model, averaged over the speakers of each gender. Precision, F1-score and recall of approx. 90% for all different models indicates a reliable classification of creak.

2.2.3. Creak Detection

In its main application, i.e., the detection of creak, `creapy` does not need any labelled data to perform creak detection in an audio file. Figure 2 shows a schematic representation of the whole creak detection procedure. For processing the audio file, we found that using windows of 40 ms length and a frame shift of 10 ms showed good results. As creak only occurs in voiced segments, we discard voiceless or silent windows: We first calculate the short-term energy (STE) of each window, and if below 0.005, those windows are discarded as considered to contain silence. For those windows with higher STE, we calculate the zero-crossing rate (ZCR); for ZCR above 0.12, those windows are considered

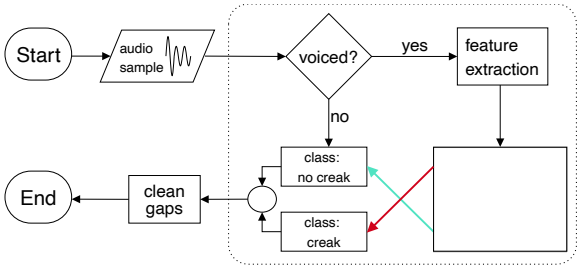


Figure 2: `creapy`'s processing steps.

voiceless and are discarded as well. This approach differs from [19], where ZCR is used as one of the features for the classifier. The five acoustic features used for creak classification (cf., Section 2.2.1) are then only extracted for the remaining intervals. If needed, missing values are imputed using the replacement value taken from the training data (cf., Section 2.2.1). Based on these acoustic features, the probability of the current interval being creaky is determined. If the probability is ≥ 0.75, the interval is classified as creak. We acknowledge that the difference between non-creaky and creaky voice is gradual rather than binary. Nevertheless, we chose this threshold because it maximised the F1-score over several test-runs with different speakers.

To create meaningful creak intervals, a post-processing step ensures creaky intervals with a minimum length of 30 ms and maximum gaps between neighbouring creak intervals of 10 ms (cf., Figure 3). Finally, the resulting creak intervals (bottom row in Figure 3) are exported to a Praat TextGrid file.

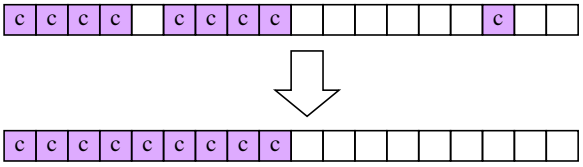


Figure 3: Schematic representation of `creapy`'s post-processing step to join neighbouring or discard very short creak intervals.

3. EVALUATION AND DISCUSSION

In this section, we present how `creapy` performs in long, continuous sound files and we compare cross-validation results from the gender-mixed (GenMix) and gender-specific (GenSpec) models. There are three possible outcomes to evaluate the detection: If the tool detected a creaky interval which overlaps with a manually labelled creak interval, this interval counts as true positive (TP). If there was no manual creak annotation at that point, the interval counts as false positive (FP). If a manually annotated creak interval was not detected by `creapy`, this counts as

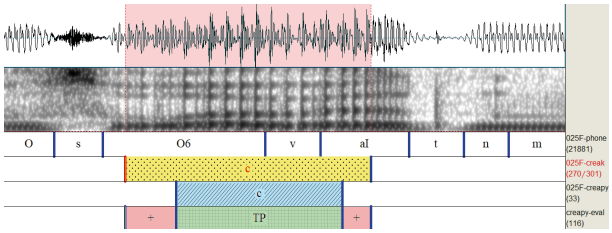


Figure 4: Example for a detected creak interval; second tier (yellow/dotted): manual label, third tier (blue/diagonal): `creapy`'s detection, fourth tier (green and red): evaluation of the detection.

false negative (FN). Figure 4 shows a TextGrid with manually annotated creak (yellow) and creak as detected by `creapy` (blue). The bottom tier shows the duration of a TP in green/checkered and the inaccuracy (i.e., time difference between manually labelled and detected intervals) in red/clean.

Table 1 shows the number of TPs, FPs, FNs, and F1- and recall-scores of the detection for the gender-specific models (GenSpec) and the gender-mixed model (GenMix). Overall, performance decreased in comparison to the evaluation of the feature set from Section 2.2.2. This decrease in performance is not unexpected, as the training material only contained manually annotated creaky and modal intervals. The continuous sound files contain a variety of speech and non-speech sounds, which were more difficult to distinguish from creak. A qualitative analysis showed that FPs often occurred in specific cases: (a) in other non-modal phonation types, e.g., pressed or breathy voice, (b) in speaker noises, such as laughter and throat clearing that may contain other kinds of non-modal phonation, (c) in modal [a], and (d) in overlapping speech when the interlocutor was audible on the speaker's audio channel. Thus, a higher performance can be expected when using `creapy` on data without overlapping speech.

Compared to the evaluation results (cf., Section 2.2.2), Table 1 shows a clearer difference in performance for speech by men vs. women. Men have a higher F1-score with relatively more FPs, and women have an overall lower F1-score and relatively more FNs. Note that this difference is *not* connected to a data imbalance, because the amount of creak labels was gender-balanced. The overall worse performance for women's speech could be related to sex-specific differences in the relationship between H1-H2 and voice quality [31]. The gender-specific model increased the number of FNs in men's data but decreased the amount of FNs in women's data. The fact that F0 octave jumps in women's creak still fall into men's modal F0 range could explain the high percentage of FNs in women's recordings processed with the gender-mixed model. With a semi-

Model	TP	FP	FN	F1	recall
M_GenMix	305	316	80	.606	.792
M_GenSpec	247	129	108	.678	.696
F_GenMix	132	55	193	.482	.406
F_GenSpec	201	103	134	.593	.600

Table 1: Cross-validation results for men and women, trained with data from all speakers (GenMix) or with gender-specific models (GenSpec).

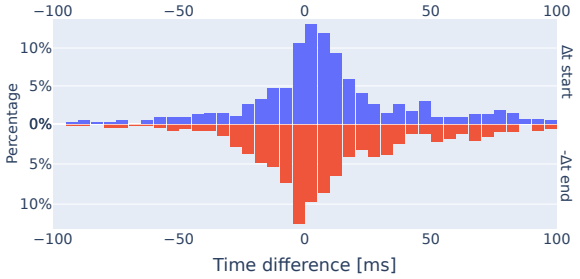


Figure 5: Absolute time difference of manual labels and detection; positive values: interval too short, negative values: interval too long.

automatic creak annotation in mind, a tool that results in fewer FNs is preferable, because it is much more time-consuming to look for creak intervals that were missed by `creapy` than to discard intervals that were incorrectly detected as creak. Therefore, we propose that the gender-specific model is better suited for creak detection for women, but not for men (cf., lower recall, despite the higher F1-score).

Our definition of TP indicates that creak was recognised correctly, but not how precisely the boundaries were detected. We thus compared the difference between start and end time of all overlapping labels (cf., Figure 4, with overlap in green/checkered and deviations in red/clean). Figure 5 shows the time differences between `creapy`'s detection and the manually annotated intervals, separately for start and end time. We inverted the sign of Δt_{end}, so that positive numbers indicate that the automatically detected interval was too short and a negative time difference indicates that the detected interval was too long. The distribution shows a shift towards positive values indicating that detected intervals tended to be too short. Some minor shifts occur inevitably due to the fixed frame shift of `creapy`. About 20 % of the time shifts are within the range of this frame shift (i.e., 10 ms). About 50 % of the intervals had an overlap of 75 % or more with manual annotations. These results indicate that while `creapy` shows a good general performance, manual adjustment of boundaries is still recommended.

`creapy` yielded results comparable to [19]'s COVAREP, a binary decision tree based on Kane et al. [22]; however, our F1-scores in Table 1 are clearly higher than their results based on [18]'s features. Yet, a detailed comparison is not possible, as their evaluation methods are not clearly identifiable.

4. HOW TO USE CREAPY

`creapy`[3] performs automatic creak detection in a continuous audio file. No previous segmentation of speech is necessary. The user can select from the following pre-trained RFC models: 1) a model trained on women only, 2) a model trained on men only, and 3) a mixed model trained on both genders (default). In the configuration file, the user can define the audio files that should be processed. While parameters from acoustic features and thresholding values (e.g., imputation strategy, ZCR and STE threshold, creak threshold) were tuned to generalise well on multiple speakers in the data we used for `creapy`'s development, these parameters can be adjusted manually in a dedicated configuration file. Detected creak-intervals are then written to a Comma-Separated-Values (`.csv`) or Praat (`.TextGrid`) file, the latter containing a new tier with the detected creak intervals. This TextGrid allows for a manual revision of the automatically created labels.

5. CONCLUSION

In this paper, we presented `creapy`, a Python-based toolkit for the detection of creak in speech signals. The best performance with a Random Forest classifier was obtained with only five acoustic features (H1H2, HNR, Jitter, Shimmer, F0-mean). In general, `creapy` achieved a good performance and we observed that a gender-mixed model performed better for men (recall: .79), and a gender-specific model performed better for women (recall: .60). Detected creak intervals had a high temporal match with manual labels. About 20 % of detected intervals overlapped almost entirely with manual labels and half of the detected intervals overlapped by 75 % or more. In the future, `creapy` could be improved with more training data of other non-modal phonation types which were difficult to distinguish from creak.

6. ACKNOWLEDGEMENTS

The work by A. Kelterer and S. Wepner was funded by grant P-32700-NB from FWF (Austrian Science Fund).

7. REFERENCES

[1] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29(4), 383–406.

[2] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 101155.

[3] Garellek, M., Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association* 41(2), 185–205.

[4] Kelterer, A., Schuppler, B. 2020. Phonation type contrasts and tone in Chichimec. *The Journal of the Acoustical Society of America* 147(4), 3043–3059.

[5] Belotel-Grenié, A., Grenié, M. 2004. The Creaky Voice Phonation and the Organisation of Chinese Discourse. *International Symposium on Tonal Aspects of Language: With Emphasis on Tone Languages*, 5–8.

[6] Carlson, R., Hirschberg, J., Swerts, M. 2005. Cues to upcoming Swedish prosodic boundaries: Subjective judgement studies and acoustic correlates. *Speech Communication* 46(3/4), 326–333.

[7] Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31(1), 139–152.

[8] Lennes, M., Anttila, H. 2002. Prosodic features associated with the distribution of turns in Finnish informal dialogues. *The Phonetics Symposium*, 149–158.

[9] Włodarczak, M., Heldner, M. 2022. Contribution of voice quality to prediction of turn-taking events. *Speech Prosody 2022, Lisbon, Portugal*, 485–489.

[10] Carlson, R., Gustafson, K., Strangert, E. 2006. Cues for hesitation in speech synthesis. *INTERSPEECH*. Citeseer.

[11] Lee, S. 2015. Creaky voice as a phonational device marking parenthetical segments in talk. *Journal of Sociolinguistics* 19(3), 275–302.

[12] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40(1-2), 189–212.

[13] Horne, M. 2009. Creaky fillers and speaker attitude: Data from Swedish. In: Barth-Weingarten, D., Dehé, N., Wichmann, A. (eds), *Where Prosody Meets Pragmatics*. Emerald Group Publishing Limited, 277–288.

[14] Grichkovtsova, I., Morel, M., Lacheret, A. 2012. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication* 54(3), 414–429.

[15] Yuasa, I. P. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3), 315–337.

[16] Mendoza-Denton, N. 2011. The Semiotic Hitchhiker's Guide to Creaky Voice: Circulation and Gendered Hardcore in a Chicana/o Gang Persona. *Journal of Linguistic Anthropology* 21(2), 261–280.

[17] Hess, W. 2012. *Pitch determination of speech signals: algorithms and devices* volume 3. Springer Science & Business Media.

[18] Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N. 2008. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 47–56.

[19] Drugman, T., Kane, J., Gobl, C. 2014. Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language* 28(5), 1233–1253.

[20] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S. 2014. Covarep – a collaborative voice analysis repository for speech technologies. *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[21] Drugman, T., Kane, J., Gobl, C. 2012. Resonator-based creaky voice detection. *Thirteenth Annual Conference of the International Speech Communication Association*, 1592–1595.

[22] Kane, J., Drugman, T., Gobl, C. 2013. Improved automatic detection of creak. *Computer Speech & Language* 27(4), 1028–1047.

[23] Van Rossum, G., Drake, F. L. 2009. *Python 3 reference manual*. CreateSpace.

[24] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer. Version 6.1.38, retrieved 2 January 2021.

[25] Kane, J., Pápay, K., Hunyadi, L., Gobl, C. 2011. On the use of creak in Hungarian spontaneous speech. *ICPhS*. Citeseer, 1014–1017.

[26] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[27] Keating, P. A., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of ICPhS*, 2–7.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

[29] Eyben, F., Wöllmer, M., Schuller, B. 2010. Opensmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

[30] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 1–22.

[31] Simpson, A. P. 2012. The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics* 40, 477–490.

[1] equal contribution

[2] corresponding author

[3] https://gitlab.tugraz.at/speech/creapy.git

THE INFLUENCE OF COMMON COLDS AND ITS INTERACTION WITH LEXICAL TONES ON VOICE QUALITY OF MANDARIN SPEAKERS

Yifan Mou[1], Lei Zhu[2]

[1, 2] Institute of Linguistics, Shanghai International Studies University, China
mouyifan@shisu.edu.cn, zhulei@shisu.edu.cn

ABSTRACT

This study examines the influence of common colds and its interaction with lexical tones on the voice quality of Mandarin speakers. We find that speakers with common colds have a significantly lower CPP and a higher Jitter compared with healthy speakers, implying a greater instability of vibrations and more noisy elements in voice production. Such changes are largely due to the inflammation of larynx and hypofunction of vocal folds caused by common colds. Besides, a significant interaction between health condition and tones is found for HNR and Jitter. Mandarin tones with curved pitch contours, including Tone2, Tone3 and Tone4, are more susceptible to colds, presenting a greater degradation of voice quality. This is mainly caused by the tendency for cold speakers to decrease their phonatory efforts and extra vocal efforts required by curved tones.

Keywords: voice quality, Mandarin tones, pathology

1. INTRODUCTION

Acoustic parameters obtained in non-invasive manners are objective indicators for speech-related diseases. The accuracy and robustness of acoustic parameters for disease detection have been confirmed in various studies, including asthma, Covid19, Parkinson's syndrome and Alzheimer's disease [1, 2]. Voice production requires the cooperation of vocal organs, such as pharynx, larynx and oral cavity [3]. However, speech-related diseases would disturb the function of vocal organs, causing significant acoustic differences between pathological speakers and normal ones [4]. These differences could be captured and quantified by various acoustic parameters [5].

Due to the connection with external environment, vocal organs are highly susceptible to viral infections [6]. Upper respiratory tract infection (URTI) is one of the most common diseases, causing three to five million serious infections each year [7]. Upper respiratory tract infection is an acute infection involving the sinuses, pharynx, larynx, trachea, and bronchi [8, 9]. The common cold is the most common symptom of upper respiratory tract infections. Adults are reported to catch two to five colds each year, while children might catch seven to ten colds annually [10, 11].

Early symptoms of common colds are headache, sneezing, chills and sore throat, and the later symptoms involve nasal congestion, cough and hoarseness [9, 12]. Even mild symptoms of colds could affect the function of vocal cords, causing temporary speech disorders including worse voice quality, vocal fatigue and reduced pitch range. Severe inflammation of the larynx caused by colds may lead to a long-term, or even persistent damage to vocal function [13, 14].

Previous phonetic studies on common colds focused on the variation of formants, fundamental frequency, spectral peaks and Mel-frequency cepstral coefficients between speakers with common colds and healthy ones. Related works have proven that frequencies of the first formant (F1) and the second formant (F2) of vowels get lower for the cold condition than for the healthy condition. The second Mel-frequency cepstral coefficient also gets significantly lower in the state of common cold [15]. The spectral peak near 1KHz in nasal consonants shows significant decrease for speakers with common colds as well. As a measurement of nasality, it suggests a higher nasal congestion for cold speakers [16]. Besides, the pitch of vowels is lowered during a cold, and an increase in noise is found for stops and liquids [6].

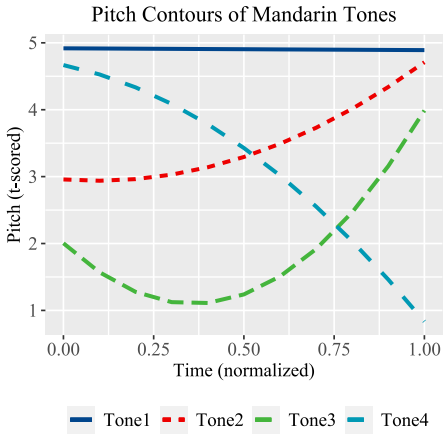
It has been noticed that speakers with common colds could have different voice quality compared with healthy people [17]. However, previous studies paid less attention to the effect of common colds on voice quality, which has been proven to be reliable indicator reflecting the state of the glottis and severity of voice disorder [18]. This study aims to fill this gap by exploring how common colds influence our voice quality, which helps uncover mechanisms of voice production for pathological conditions and improve the diagnosis of speech-related diseases. Based on previous studies, we hypothesized that there exists a significant main effect of health condition on voice quality. Common colds would lead to a worse performance of voice quality. According to the protocol of the European Laryngology Society [19] and American Speech-Language-Hearing Association [18], four parameters of voice quality, including harmonic-to-noise ratio (HNR), cepstral peak prominences (CPPs), Jitter and Shimmer, were included in the study.

We also noticed that recording materials used in previous works on common colds were mainly sustained vowels with flat pitch contours. Mandarin has

four lexical tones classified by their pitch heights and pitch contours. Mandarin tones include Tone 1 (high-flat), Tone 2 (mid-rising), Tone 3 (falling-rising) and Tone 4 (high-falling) [20], as shown in Figure 1.

In previous research on tonal languages, a significant interaction between tones and phonation was found. Different registers of tones show a clear phonation contrast with each other [21], which implies that variations in vocal effort exist for different tones. Considering the vocal fatigue and hypofunction of vocal folds for the cold condition, there is a tendency for speakers with common colds to adjust phonation strategy to reduce their phonatory efforts [16]. Thus, we proposed our second hypothesis that there exists an interaction between health condition and Mandarin tones. Some Mandarin tones might be more susceptible to common colds, presenting a greater degradation of voice quality.

Figure 1: Four types of Mandarin tones, plotted with normalized time and t-scored pitch height



2. METHOD

2.1 Data

Audio files in the study were obtained from the Common Cold Speech Dataset offered by Chinese Linguistic Data Consortium, containing a set of speech and corresponding transcriptions recorded by speakers with common colds and healthy speakers. Recordings were mono-channel speech saved in Wav format, with a sampling rate of 16KHz. Speech data were recorded with professional equipment in a quiet environment. Each phrase was repeated three times by participants to ensure a stable performance.

Recording materials included numbers, places and daily expressions. Almost all possible combinations of Mandarin syllables were covered in the design of recording materials, and there were sufficient recordings for all four Mandarin tones. All participants were native Mandarin speakers without any hearing loss or major voice damage. The dataset contained 92 participants with balanced gender and health condition distributions, including 45 common cold speakers and 47 healthy speakers. Half of participants were males, while another half were females. Each participant recorded approximately 600 words in average.

2.2 Data Processing

We used a pre-trained MFA [22] Mandarin acoustic model to perform forced alignment on the recorded speech and their corresponding transcriptions at the syllable level and the phoneme level. The Mandarin acoustic model has already been trained on various large-scale Mandarin datasets, with quite good performance and robust error rate. After obtaining the alignment, we then manually corrected the inaccurate boundaries of syllables and phonemes in the TextGrids. After all boundaries being marked and correctified, we used Praat's python interface *Parselmouth* [23] to extract parameters of voice quality from the voicing part (the rhyming part with lexical tones) of Mandarin syllables. Extracted data were subjected to outlier detection to ensure the results of linear mixed-effects models not being disturbed by outliers.

As we have mentioned in Section 1, four parameters of voice quality including harmonic-to-noise ratio (HNR), cepstral peak prominences (CPPs), Jitter and Shimmer were included in the study. HNR and CPPs mainly represent the proportion of noise in the sound, while Jitter and Shimmer mainly stand for the stability of voice production during the phonation stage. To improve the accuracy of parameter extraction in pathological condition, we adopted the cross-correlation method for the estimation of fundamental frequency, as suggested by the manual of Praat [24]. The default range for pitch detection was expanded to 50Hz (pitch floor) to 600Hz (pitch ceiling). Straight line was selected as the regression line to fit the overall cepstrum for calculating CPPs. Jitter (local) was measured by the average absolute difference between lengths of consecutive periods, divided by the average length. Shimmer (local) was measured by the average absolute difference between amplitudes of consecutive periods, divided by the average amplitude.

2.3 Statistical Analysis

In order to determine the effect of health condition and its interaction with lexical tones on voice quality of Mandarin speakers, we fitted four linear mixed-effects models for four parameters of voice quality respectively, using *lme4* package [25] in R [26]. Health condition (healthy vs. cold), tones (Tone1 to Tone4) and their interaction were included as fixed effects, while random effects were made of by-participant, by-gender, and by-item random intercepts. Health

condition and tones were treatment-coded, with healthy condition and Tone1 as the reference level.

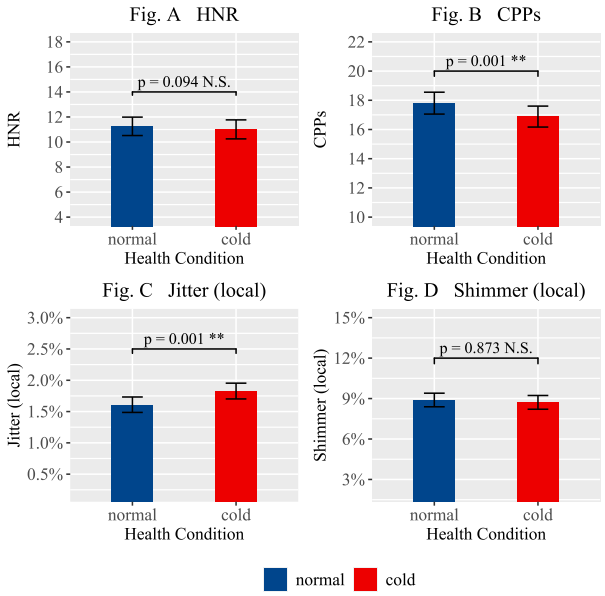
All models were fitted with the backward stepwise method by removing the random effect capturing the smallest variance until convergence [27]. Examination of main effects and interaction effects was based on the Type III analysis of variance, with *anova* function in *lmerTest* package [28]. Degree of freedom was calculated with the Satterthwaite's method. Multiple comparisons were calculated by *emmeans* package [29] with Bonferroni correction. The significance level was set at 0.05, and *p*-values were obtained with *lmerTest*.

3. RESULTS

3.1 Main effect of health condition on voice quality

Linear mixed effects analysis performed on the CPPs and Jitter shows significant differences for speakers under different health conditions, as shown in subplots B and C of Figure 2. Specifically, a significant main effect of health condition is found on CPPs [F (1,90) = 11.441, p = 0.001**] and Jitter [F (1,90) = 11.314, p = 0.001**], indicating that speakers with common colds have different CPPs and Jitter, compared with healthy speakers. Statistics reveals that health condition does not seem to cause significant differences in HNR [F (1,90) = 2.856, p = 0.094] and Shimmer [F (1,90) = 0.0483, p = 0.873], as shown in subplots A and D of Figure 2.

Figure 2: Error bar plots of HNR, CPPs, Jitter and Shimmer for different health conditions



Given the significant difference in CPPs and Jitter of speakers under different health conditions, a Bonferroni corrected post-hoc pairwise comparison was carried out. Results of post-hoc pairwise comparisons (see Table 1) show that speakers with common colds have a significantly lower CPP compared with healthy speakers [β = -0.835, t = -3.383, p = 0.001**]. A lower CPP indicates an increase in noise during the process of voice production, probably caused by hoarseness and hypofunction of vocal folds. Besides, speakers with common colds have a significantly higher Jitter compared with healthy speakers [β = 0.220, t = 3.364, p = 0.001**]. A higher Jitter indicates that more aperiodic vibrations are found in the phonation stage, which is largely due to asymmetrical changes in mass and tension of vocal folds under pathological condition. Coughs and inflammation of larynx, which cause discomfort and disturb vocal function, could lead to an increase in Jitter as well.

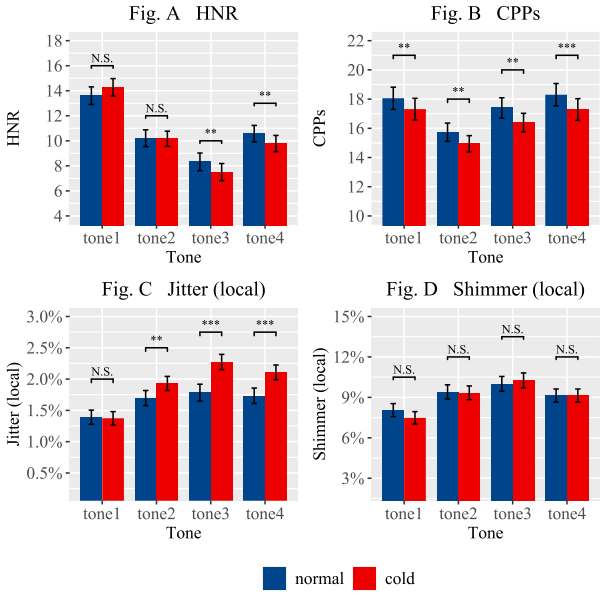
Table 1: Results of post-hoc pairwise comparisons for main effect of health condition. Estimates with significant p-values are put in bold.

parameter	*Contrast*	*β*	*S.E.*	*t.ratio*	*p*
HNR	cold - healthy	-0.764	(0.452)	-1.690	0.094
CPPs	cold - healthy	**-0.835**	**(0.247)**	**-3.383**	**0.001**
Jitter	cold - healthy	**0.220**	**(0.066)**	**3.364**	**0.001**
Shimmer	cold - healthy	-0.075	(0.340)	-0.220	0.873

3.2 Interaction between health condition and tones

Significant interaction effects between health condition and lexical tones are found for HNR [F = 190.85, p < 0.001**] and Jitter [F = 328.12, p < 0.001**], as shown in subplots A and C of Figure 3. No significant interaction effects are found in CPPs [F = 0.694, p = 0.556] and Shimmer [F = 2.398, p = 0.066], as shown in subplot B and D. (No significant interaction effect was found in CPPs, despite the presence of significant simple main effects of health condition in all four lexical tones.) Significant interaction effects in HNR and Jitter indicate that there are different simple main effects of health condition for voice quality in four Mandarin tones.

Figure 3: Error bar plots of HNR, CPPs, Jitter and Shimmer of four tones by different health conditions



Post-hoc pairwise comparisons (Table 2) were carried out to further examine the interaction between health condition and tones. Table 2 shows that difference of HNR between cold speakers and healthy speakers is significant for Tone 3 [β = –1.402, t = –3.075, p = 0.003**] and Tone 4 [β = –1.20, t = –2.651, p = 0.009**], but not for Tone1 [β = 0.08, t =0.177, p = 0.86] and Tone2 [β = –0.533, t = –1.164, p = 0.247]. Results also confirm that the difference of Jitter between speakers with common colds and healthy speakers is significant for Tone2 [β = 0.196, t = 2.881, p = 0.005**], Tone3 [β = 0.418, t = 6.165, p < 0.001***] and Tone4 [β = 0.348, t = 5.297, p < 0.001***], but no significant difference is shown for Tone1 [β = -0.082, t = -1.253, p = 0.213]. Besides, pairwise comparisons of Jitter indicate a larger effect size of health condition in Tone3 [β = 0.418] and Tone4 [β = 0.348] than in Tone2 [β = 0.196].

Table 2: Results of post-hoc pairwise comparisons for the interaction effect. Differences between cold and healthy condition in each tone are listed. Estimates with significant p-values are put in bold.

Parm.	*Lexical Tones (cold - healthy)*				*Statistics*	
	Tone1	*Tone2*	*Tone3*	*Tone4*	*F*	*Sig.*
HNR	0.080	-0.533	**-1.402**	**-1.200**	**190.95**	**< 0.001**
CPPs	**-0.811**	**-0.784**	**-0.841**	**-0.904**	0.694	0.556
Jitter	-0.082	**0.196**	**0.418**	**0.348**	**328.12**	**< 0.001**
Shimmer	-0.535	-0.121	0.331	0.027	2.398	0.066

4. DISCUSSION

Results of linear mixed-effects models suggest that there is a significant main effect of health condition on CPPs and Jitter. Specifically, the CPPs of cold speakers is significantly lower, compared with healthy speakers. Cold speakers also have higher Jitter values, compared with healthy ones. As a measurement of overall noise level in the speech, a lower CPP implies that more aperiodic vibrations and noisy elements are found in the voice production of cold speakers. As for Jitter, it mainly presents the stability of vibration of the vocal folds. A higher Jitter suggests that the vocal folds show more tremors and a greater instability during the phonation stage.

Degradation of voice quality is largely due to the inflammation of larynx caused by common colds, which changes the mass, length and tension of the vocal folds. Such changes could lead to a more deviant vibration. It becomes more difficult for speakers with common colds to maintain a consistent pitch or change their pitch with ease. Besides, asymmetrical changes in the shape and temporary hypofunction of vocal folds make it more difficult to form a complete closure during the contact phase. More airflows could be spilt out of the vocal folds, which accounts for an increase of noise and a greater instability of phonation in cold speakers. Other symptoms of common colds, such as coughing and hoarseness, irritate the vocal folds as well and bring more noisy portions to the sound.

Results of linear mixed-effects models suggest that a significant interaction effect between health condition and lexical tones was found for HNR and Jitter. To be specific, significant difference of HNR between cold speakers and healthy speakers was found for Tone 3 and Tone 4, but not for Tone1 and Tone2. Significant difference of Jitter was found for Tone2, Tone3 and Tone4, but not for Tone1. Significant interaction between health condition and Mandarin tones implies that lexical tones with curved pitch contours, including Tone2, Tone3 and Tone4, are more susceptible to common colds, presenting a more apparent degradation of voice quality.

Modification of fundamental frequency is achieved through changes in mass, length and tension of the vocal folds [30]. Therefore, Mandarin tones with curved pitch contours, including Tone 2 (mid-rising), Tone 3 (falling-rising) and Tone 4 (high-falling), might require extra vocal efforts and muscle controls during the process of voice production. Suggested by previous studies, there is a tendency for cold speakers to adjust their phonation strategy with a decrease of phonatory effort, considering the vocal fatigue caused by common colds [15]. Thus, curved tones which require more phonatory efforts, including Tone2, Tone3 and Tone4, show an apparently poorer voice quality compared to flat Tone1.

Besides, we notice that pairwise comparisons for Jitter show a larger effect size of health condition in Tone3 and Tone4 than in Tone2, despite they are all curved tones. It is largely due to the difference in the magnitude and rate for the change of pitch. As shown in Figure1, Tone3 and Tone4 have greater pitch ranges and steeper slopes of pitch contour, compared with Tone2. More phonatory efforts might be required for the production of Tone 3 and Tone4, to assure a timely and fast-enough change of the height and slope of the pitch. Besides, the production of Tone3 and Tone4 passes the lower part of possible pitch range of speakers, where chaotic oscillations are usually observed during lax phonation [31]. Thus, a greater degradation of voice quality could be found in Tone3 and Tone4, than in Tone2.

The study fills the gap of previous phonetic studies on common colds by exploring the influence of common colds on voice quality and its interaction with lexical tones in Mandarin speakers. The study offers a new interface between speech pathology and tone types, which could be further expanded in relevant studies on other tonal languages. Studies of the voice quality of colds speakers could help uncover mechanisms of voice production of speakers with speech disorders, and improve the performance of speech-related disease detection.

Additional investigation in future studies may be necessary to explore the physiological and articulatory factors underlying the dissimilarities of changing patterns observed in four parameters of voice quality.

5. REFERENCES

[1] V. M. Dixit and Y. Sharma, "Voice Parameter Analysis for the disease detection," *IOSR Journal of Electronics and communication Engineering,* vol. 9, no. 3, pp. 48-55, 2014.

[2] P. Mouawad, T. Dubnov, and S. Dubnov, "Robust detection of COVID-19 in cough sounds," *SN Computer Science,* vol. 2, no. 1, pp. 1-13, 2021.

[3] A. Vahedian-Azimi, A. Keramatfar, M. Asiaee, S. S. Atashi, and M. Nourbakhsh, "Do you have COVID-19? An artificial intelligence-based screening tool for COVID-19 using acoustic parameters," *The Journal of the Acoustical Society of America,* vol. 150, no. 3, pp. 1945-1953, 2021.

[4] J. Kutor, S. Balapangu, J. K. Adofo, A. A. Dellor, C. Nyakpo, and G. A. Brown, "Speech signal analysis as an alternative to spirometry in asthma diagnosis: investigating the linear and polynomial correlation coefficient," *International Journal of Speech Technology,* vol. 22, no. 3, pp. 611-620, 2019.

[5] A. Rybakovas, V. Beiša, K. Strupas, J. Kaukėnas, and G. Tamulevičius, "Inverse Filtering of Speech Signal for Detection of Vocal Fold Paralysis After Thyroidectomy," *Informatica,* vol. 29, no. 1, pp. 91-105, 2018.

[6] J. Wagner, T. Fraga-Silva, Y. Josse, D. Schiller, A. Seiderer, and E. André, "Infected Phonemes: How a Cold Impairs Speech on a Phonetic Level," in *Interspeech*, 2017.

[7] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods,* vol. 151, pp. 41-54, 2018.

[8] R. Eccles, "Understanding the symptoms of the common cold and influenza," *The Lancet infectious diseases,* vol. 5, no. 11, pp. 718-725, 2005.

[9] M. J. Bové, S. Kansal, and C. A. Rosen, "Influenza and the vocal performer: Update on prevention and treatment," *Journal of Voice,* vol. 22, no. 3, pp. 326-332, 2008.

[10] A. S. Monto, "Epidemiology of viral respiratory infections," *The American journal of medicine,* vol. 112, no. 6, pp. 4-12, 2002.

[11] D. A. J. Tyrrell, S. Cohen, and J. E. Schilarb, "Signs and symptoms in common colds," *Epidemiology & Infection,* vol. 111, no. 1, pp. 143-156, 1993.

[12] G. G. Jackson, H. F. Dowling, I. G. Spiesman, and A. V. Boand, "Transmission of the common cold to volunteers under controlled conditions: I. The common cold as a clinical entity," *AMA archives of internal medicine,* vol. 101, no. 2, pp. 267-278, 1958.

[13] R. T. Sataloff, "Common infections and inflammations and other conditions," *Professional Voice: The Science and the Art of Clinical Care,* p. 429, 1997.

[14] P.-T. Lin, J. C. Stern, and W. J. Gould, "Risk factors and management of vocal cord hemorrhages: an experience with 44 cases," *Journal of Voice,* vol. 5, no. 1, pp. 74-77, 1991.

[15] R. G. Tull and J. C. Rutledge, "Analysis of "cold-affected"speech for inclusion in speaker recognition systems," *The Journal of the Acoustical Society of America,* vol. 99, no. 4, pp. 2549-2574, 1996.

[16] R. G. Tull, *Acoustic analysis of cold-speech: Implications for speaker recognition technology and the common cold*. Northwestern University, 1999.

[17] A. K. Suresh, M. SrinivasaRaghavanK., and P. K. Ghosh, "Phoneme State Posteriorgram Features for Speech Based Automatic Classification of Speakers in Cold and Healthy Condition," in *Interspeech*, 2017.

[18] R. R. Patel *et al.*, "Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function," *American journal of speech-language pathology,* vol. 27, no. 3, pp. 887-905, 2018.

[19] P. H. Dejonckere *et al.*, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-rhino-laryngology,* vol. 258, no. 2, pp. 77-82, 2001.

[20] Y. R. Chao, (Mandarin Primer). Cambridge, MA and London, England: Harvard University Press, 1948.

[21] J. Kuang, "Production and perception of the phonation contrast in Yi," University of California, Los Angeles, 2011.

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech*, 2017, vol. 2017, pp. 498-502.

[23] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics,* vol. 71, pp. 1-15, 2018.

[24] P. Boersma, "Praat: doing phonetics by computer," http://www. praat. org/, 2006.

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823,* 2014.

[26] R. C. Team, "R: A language and environment for statistical computing," 2013.

[27] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language,* vol. 68, no. 3, pp. 255-278, 2013.

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," *Journal of statistical software,* vol. 82, pp. 1-26, 2017.

[29] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Emmeans: Estimated marginal means, aka least-squares means," *R package version,* vol. 1, no. 1, p. 3, 2018.

[30] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, "Voice quality evaluation in patients with COVID-19: An acoustic analysis," *Journal of Voice,* 2020.

[31] D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer, "Interpretation of biomechanical simulations of healthy and chaotic vocal fold oscillations with empirical eigenfunctions," *The Journal of the Acoustical Society of America,* vol. 95, no. 6, pp. 3595-3604, 1994.

EXPLORING THE ROLE OF FORMANT FREQUENCIES IN THE CLASSIFICATION OF PHONATION TYPE

Bogdan Ludusan[1], Mattias Heldner[2], Marcin Włodarczak[2]

[1]Bielefeld University, [2]Stockholm University
bogdan.ludusan@uni-bielefeld.de, {heldner,wlodarczak}@ling.su.se

ABSTRACT

Phonation type has been found to interact with other acoustic characteristics of speech, including vowel formants. Here, we investigated the relation between phonation type and formants in three Swedish vowels, produced in three phonation types (neutral, breathy and pressed). Our findings show that the formant values change from neutral to breathy and pressed, respectively, but that this effect is more consistent for pressed than for breathy phonation. Moreover, while some systematic patterns can be identified, there is also substantial speaker variation in the size and in the direction of the change. We then employed the formant values, along with other acoustic features, to automatically classify phonation type. The addition of formant information to the classifier significantly improved the performance of the system.

Keywords: phonation type, formant values, pressed, breathy, classification.

1. INTRODUCTION

Phonation type has been primarily described in terms of the degree of glottal adduction [1], which is, in turn, linked to other aspects of voice production, such as variation in acoustic intensity [2] and fundamental frequency [3]. In addition, given that the degree of adduction is associated with vertical larynx position [4, 5], the glottal configuration also affects formant frequencies by changing the vocal tract length [6].

A review of the effects of phonation type on formant frequencies across a number of languages [1] found that breathy phonation is associated with lower formant frequencies (in particular F1) than creaky phonation (but cf. [7], who found no significant effect of phonation type on F1 in Jalapa Mazatec). In a more controlled experimental setting, with participants producing the syllable /pæː/ in modal and pressed phonations, [8] found that ratings of pressedness by speech-language pathologists were linked to variation in formant frequencies (F1 in females and F2 in males). While all these effects were attributed to the impact of phonatory setting on larynx height, recent developments, such as the Laryngeal Articulator Model [9, 10] have proposed a more comprehensive view, whereby raising the larynx also involves the aryepiglottic constriction and tongue retraction.

Whatever the mechanisms involved, instrumental classification of phonation types has predominantly employed methods explicitly designed to be insensitive to formant effects, whether by using inverse filtering [11], electroglottography [12], miniature accelerometers attached to the neck [13] or acoustic measures in the cepstral domain [14]. However, recent work [15] found that automatic classification of phonation type based on acoustic features derived directly from the speech signal reaches similar performance to those calculated from inverse filtering and neck-surface acceleration.

Given the evidence that supraglottal resonances are affected by phonatory posturing, in this paper we analyse the effect of formant frequencies on the automatic classification of phonation type and their relationship with other acoustic measures of phonation type.

2. MATERIALS

The speech materials consisted of diminuendo sequences of syllables with three different vowels [piː, pæː, poː]; produced in three phonation types (neutral, breathy, pressed); at three self-selected pitch levels (habitual, low, high); by five trained singers (two females: S3 and S5, and three males). Each combination of factors was repeated in at least three diminuendo sequences. The recordings were made in a sound treated room using a head-mounted omnidirectional microphone (Sennheiser HSP2) positioned 6 cm from the mouth and digitized using an Expert Sleepers ES-9 audio interface (48 kHz, 24 bit). The recording included other transducers as well, but only the microphone signals was used in the present paper. This procedure resulted in a total of 2640 syllables (897 pæː, 887

pi: and 856 po:), corresponding to breathy (790), neutral (934), and pressed (916) phonation.

3. METHODS

The syllables were annotated by two of the authors. The position of the burst of the stop consonant and the end of the vowel portions were marked and the middle point between these two landmarks was used in the subsequent analysis.

In a first step, the values of first three formants (F1, F2 and F3) were extracted using the *burg* method of the Praat software [16], employing the default parameters in the case of female speakers and a lower formant ceiling (5 kHz) for male speakers. We then investigated whether differences occur in terms of F1-F3 between the three phonation types, for each speaker. The significance of the differences was determined by means of Wilcoxon rank sum tests, corrected for multiple comparisons with the method in [17]. The statistical analyses were run using the appropriate function of the R software [18]. We also examined the relation between formant values and a number of acoustic features previously employed for the classification of phonation type [15]. These were extracted (using Praat), from the same time instants where the formant values were computed. The considered set includes the following features:

- *alpha* - alpha ratio. It represents the ratio of acoustic energy between the high (1-5 kHz) and the low (0-1 kHz) frequency bands.
- *cpps* - smooth cepstral peak prominence. It is defined as the amplitude of the first cepstrum rahmonic relative to the regression line of the cepstrum of the signal [19].
- *f0* - fundamental frequency. The vibration rate of the vocal folds.
- *h1h2*. It consists of the difference between the level of the first two harmonics.
- *hrf* - harmonic richness factor. Defined as in [20], being the ratio between the summed amplitudes of the 2nd to 10th harmonics and the amplitude of the fundamental.

In a second step, we tested if formant information may be used to improve the automatic detection of phonation type. We used the Random Forest algorithm [21] for the classification experiment (the implementation offered by the *randomForest* R package [22]). Random Forest has previously been employed in phonetic research (e.g., [23]), due to its characteristic of being able to return also the importance of the features involved in the classification task. We considered three cases: neutral-breathy (N-B), neutral-pressed (N-P) and breathy-pressed (B-P). For each case, we trained a Random Forest classifier, consisting of 500 trees, with the set of features employed in [15] and another classifier with the same features plus the values of the first three formants. The out-of-bag (OOB) error was chosen as evaluation measure for the goodness of classification. It is calculated on a subset of the data on which the trees were not fitted and a lower error value represents a better discrimination performance.

We determined the importance of each of the features used in the classification experiments. It was operationalized as the total decrease in node impurities (as given by the Gini index) when performing a split on that feature, averaged across all trees. A more discriminative feature would be represented by a higher importance. We then normalized the importance of each feature by the sum of importance values of all features considered in that condition. After this process, the sum of the importance of all the features in each condition will be equal to 1, allowing for an easier comparison between conditions.

The experiments were run on a per-speaker basis and all vowel instances produced by the speaker in the two phonation types discriminated in the experiment were considered. For each speaker and phonation type pair, a total of 100 runs were performed and the mean OOB error and mean importance of the features over the 100 runs reported. To test the statistical significance of the differences between the discrimination of various conditions, as well as between the two feature sets (without/with formant values), we used bootstrapping (by means of the functions of the R package *boot* [24]). It is a non-parametric method which performs sampling with replacement from the set containing the OOB error rates given by the random forest classifier. We repeated the process for 10,000 times, computing at each iteration the mean error and the 95% confidence intervals. If the obtained confidence intervals do not overlap, it means the differences are statistically significant.

4. RESULTS

4.1. Acoustic analysis

The variation of the formant values with respect to the three investigated phonation types is illustrated in Figure 1. For most speaker/vowels combinations, we can see that changing the phonation type from neutral to breathy or pressed results in a change in the mean formant values of the vowel, and

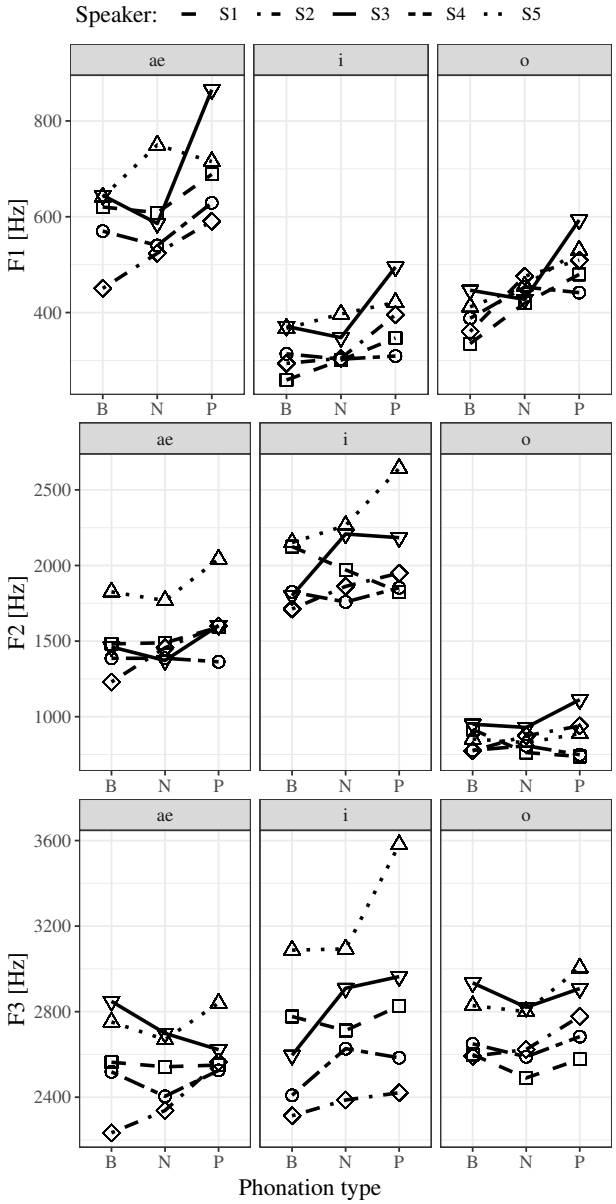


Figure 1: Mean F1, F2 and F3 values across speakers, for each of the investigated vowels in the three phonation type conditions (B - breathy, N - neutral, P - pressed).

formant values generally increased from breathy to pressed. However, one may note a variety of patterns involving the neutral condition, which implies the use of different strategies by the speakers.

We tested the differences between formant values for all pairs of phonation types, within each speaker, by means of Wilcoxon rank sum tests. The results are presented in Table 1. It shows that all speakers modify their formants in one way or another, across phonation types, but also that important individual variation exists. The number of cases (phonation type/formant) where significant changes occurred varied, from the maximal number of cases (9), for

Spkr.	N-B			N-P			B-P		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
S1	×	–	×	×	–	×	×	–	–
S2	×	×	×	×	×	×	×	×	×
S3	–	–	–	×	×	–	×	×	–
S4	–	–	–	×	–	×	×	–	×
S5	–	×	×	×	×	×	×	×	×
All	×	×	–	×	×	×	×	×	×

Table 1: The significance of the change in formant (F1-F3) values between phonation types, for each speaker and for all speakers. $\times/-$ denote statistical significance/lack of it ($\alpha = .05$).

S2, to about half the cases (4), for S3 and S4. While the majority of speakers produced some sort of change in all three investigated phonation type contrasts, speakers S3 and S4 produced no change between neutral and breathy phonation. Overall (last row of Table 1), we see that all but F3 in the N-B case exhibit significant differences (although three of the five speakers changed their F3 between neutral and breathy, the direction differed between the speakers, resulting in a null overall effect).

Finally, we checked the relation between the values of the formants and those of acoustic features previously employed for automatic phonation type classification. The obtained Spearman ρ is illustrated in Table 2. Three of the five features (alpha, h1h2 and hrf) exhibit medium correlations with the first formant value and one of them (alpha) also with the value of the second formant.

4.2. Automatic phonation type detection

The results of the two experiments (without/with the values of the first three formants) are presented in Figure 2. In both experiments, the highest error was obtained in the N-B condition, followed by N-P and the lowest one in the B-P case. A reduction in error rate was seen for all three conditions when the formants were added to the feature set (N-B: 0.9%, N-P: 2.2%, B-P: 1.3%, which corresponds to an additional 25, 57 and 34 correctly classified vowels).

feature	F1	F2	F3
alpha	.49	.55	-.08
cpps	.20	.07	-.14
f0	.31	.12	.29
h1h2	-.45	.19	-.04
hrf	.56	-.23	-.03

Table 2: Spearman rho for the correlation between formant values and acoustic features, across all phonation types and vowels. All but hrf-F3 are significant ($\alpha = .05$).

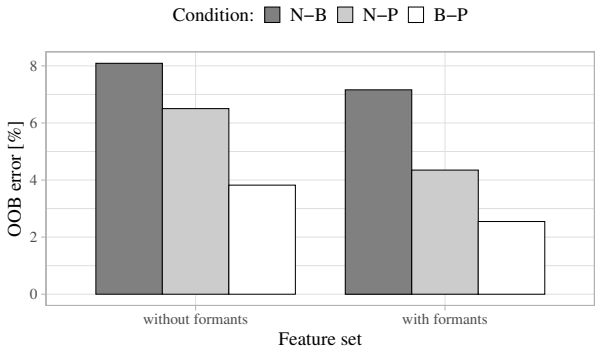


Figure 2: Classification results (mean OOB error) for the two feature sets: without and with formant values, and for the following the phonation type pairs: neutral-breathy (N-B), neutral-pressed (N-P) and breathy-pressed (B-P).

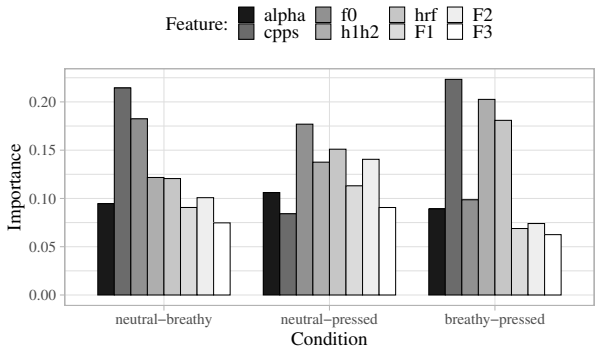


Figure 3: The importance of the features used in the classification experiments considering the formant values, in each of the three conditions.

We tested the differences between conditions within each of the two experiments, as well as between the experiments. There was no overlap between the 95% confidence intervals of any of the pairs compared, which means that all pairwise differences (within and between experiments) are significant. Looking at the per-speaker results in more detail, we observed a decrease in error in 14 of the 15 cases (3 conditions $\times$ 5 speakers): the classification of S2 data did not improve in the N-B condition.

The importance of all the features used in the second experiment (including the formant values) is illustrated in Figure 3. It shows that formant values helped most the N-P classification, all formants being ranked higher than at least one of the other five features (F2 - 3rd, F1 - 5th and F3 - 7th). In the N-B case only F2 was more important than the least important of the other features (alpha), while for the B-P classification the formants had the lowest importance. However, the importance of the formants varied with the speaker, the classification algorithm relying more on formants for some speakers than for others.

5. CONCLUSIONS

We have conducted an investigation of the effects of phonation type on formant values, by systematically analyzing the first three formants of three Swedish vowels, across five speakers. It revealed that all three formants are affected by phonation type, with the changes in formant values for pressed phonation being more systematic than for breathy phonation (compared to neutral phonation production). A higher number of cases showing significant differences in formant values in the comparisons involving pressed speech (as seen from Table 1) suggest that the most marked changes to supraglottal resonances, and implicitly in the position of the formants, occurs in the pressed condition. The observed individual variation might be due to the ability of the speakers to control their larynx height (speakers S1, S3 and S4, who are more experienced singers, show less of a difference between phonation types). We also analysed the relation between formants and other acoustic measures of phonation type. The formant values were mostly correlated to acoustic features that characterize spectral slope (alpha), but also partly the strength of the fundamental (h1h2, hrf). By contrast, cpps, an acoustic measure in the cepstral domain designed to be robust to formant variation, was the least correlated with F1-F3.

We then used the formant information for automatic classification of phonation type, by comparing the performance of one set of features previously used successfully for this task with that of the same set enriched with the formant values. Despite their correlation with some of the other acoustic features and their relatively low mean importance for the classification task (especially in the B-P classification), the use of formant values helped decrease the OOB error rate in all conditions. This shows that formant information is beneficial to automatic detection of phonation type, in a speaker-dependent manner. Future work may explore whether this type of information can be used also in a speaker-independent fashion.

ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council project *Prosodic functions of voice quality dynamics* (VR 2019-02932) to MW. The authors would like to thank Johan Sundberg for the useful feedback.

6. REFERENCES

[1] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[2] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 453–457, 2006.

[3] J.-M. Hombert, J. J. Ohala, and W. G. Ewan, "Phonetic explanations for the development of tones," *Langauge*, vol. 55, no. 1, pp. 37–58, 1979.

[4] J. Sundberg and A. Askenfelt, "Larynx height and voice source. A relationship?" *STL-QPSR*, vol. 22, no. 2–3, pp. 23–36, 1981.

[5] M. Saldias, M. Guzman, G. Miranda, and A.-M. Laukkanen, "A computerized tomography study of vocal tract setting in hyperfunctional dysphonia and in belting," *Journal of Voice*, vol. 33, no. 4, pp. 412–419, 2019.

[6] J. Sundberg and P.-E. Nordström, "Raised and lowered larynx – the effect on vowel formant frequencies," *STL-QPSR*, vol. 17, no. 2–3, pp. 35–39, 1976.

[7] M. Garellek and P. Keating, "The acoustic consequences of phonation and tone interactions in jalapa mazatec," *Journal of the International Phonetic Association*, vol. 41, no. 2, pp. 185–205, 2011.

[8] M. Millgård, T. Fors, and J. Sundberg, "Flow glottogram characteristics and perceived degree of phonatory pressedness," *Journal of Voice*, vol. 30, no. 3, pp. 287–292, 2016.

[9] J. H. Esling, S. R. Moisik, A. Benner, and L. Crevier-Buchman, *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press, 2019.

[10] J. A. Edmondson and J. H. Esling, "The valves of the throat and their functioning in tone, vocal register and stress: Laryngoscopic case studies," *Phonology*, vol. 23, no. 2, pp. 157–191, 2006.

[11] J. Sundberg, "Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental," *Journal of Voice*, vol. 36, no. 1, pp. 4–14, 2020.

[12] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2281–2291, 2017.

[13] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5060–5064.

[14] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech Language and Hearing Research*, vol. 37, no. 4, 1994.

[15] M. Włodarczak, B. Ludusan, J. Sundberg, and M. Heldner, "Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound," *Journal of Voice*, 2022.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Computer program, 2021, http://www.praat.org/.

[17] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[19] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[20] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[21] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[23] B. Ludusan, P. Wagner, and M. Włodarczak, "Cue interaction in the perception of prosodic prominence: The role of voice quality," in *Proc. of Interspeech*, 2021, pp. 1006–1010.

[24] A. Canty, "Resampling methods in R: the boot package," *The Newsletter of the R Project Volume*, vol. 2, no. 3, pp. 2–7, 2002.

VOCAL FOLD VIBRATORY PATTERNS IN BILINGUAL SPEAKERS OF LOW AND HIGH GERMAN

Jörg Peters, Marina Frank, Marina Rohloff

Carl von Ossietzky University Oldenburg
{joerg.peters, marina.frank, marina.rohloff1}@uni-oldenburg.de

ABSTRACT

The aim of this study was to investigate whether changes in stability, periodicity, and rate of vocal fold vibration could serve as indicators of cognitive load in bilingual speakers. We collected speech samples from 95 bilingual speakers of Low German (LG) and High German (HG), aged 15-88 years. Younger speakers had lower levels of jitter and shimmer, along with increased HNR, CPPs, and mean f0 while speaking LG, indicating a greater cognitive load compared to speaking HG. Conversely, older speakers showed minimal or no differences in performance between the two languages. The study found similar results when using the relative dominance of LG over HG instead of age as the predictor variable, which we calculated based on age of acquisition, frequency of use, and self-reported language proficiency. These results highlight the importance of considering vocal fold vibration as a potential marker of cognitive load during language use in bilingual speakers.

Keywords: Voice quality, cognitive load, bilingualism, language dominance, endangered languages.

1. INTRODUCTION

Research on task difficulty has identified several changes in vocal fold vibration that can serve as vocal cues to increased cognitive load [1]-[3]. Studies on speech tasks such as memory and Stroop tasks have found that increased cognitive load is associated with (i) reduced vocal fold perturbation, which results in more stable and synchronized vibratory patterns and is characterized by lower jitter and shimmer [4]-[8]; (ii) reduced additive noise in the voice source and increased periodicity of the speech signal, resulting in a higher Harmonics-to-Noise Ratio (HNR) and Cepstral Peak Prominence (CPP) [4][6][9][10]; and (iii) a higher rate of vocal fold vibration, or fundamental frequency (f0) [4][6]-[8][11][12].

The present study examines whether characteristics of vocal fold vibration can likewise indicate increased cognitive load in non-balanced bilinguals when speaking their weaker language (referred to as L2). Limited research exists on voice quality in bilingual speech, but there is some evidence that L2 speech tends to show a more stable vocal fold vibration and to have a higher mean f0. Decreased jitter and/or shimmer was observed in the L2 of Chinese-English and English-Chinese bilinguals [13][14]. Increased mean f0 was reported for a wide variety of bilingual language pairs including Brazilian Portuguese, Chinese, English, Finnish, French, German, Italian, Korean, Maori, Russian, and Spanish as L1 and/or L2 [14]-[24]. No conclusive findings are available for HNR and CPP in bilinguals [25].

However, many of these findings cannot be interpreted without reservation as evidence of increased cognitive load in the L2. First, it is not clear from many studies how well the L2 was mastered and how its use might have been associated with higher cognitive load. Second, it is not always clear whether the differences found in vocal fold vibration arose from higher cognitive load when speaking the L2 or from language-induced differences in voice quality [26][27]. Numerous studies on bilinguals have even been conducted for the purpose of finding language-dependent differences in voice quality, but only a few studies have controlled for language-specific influences by adopting a design in which the same languages have been acquired by bilinguals once as L1 and once as L2 [13][14][28]-[30].

The association of language use with cognitive load is of particular interest for the study of endangered languages, as increased cognitive load in speaking these languages can be expected to lead to a further decline in their use. Many endangered languages are no longer acquired monolingually but only as a non-dominant language alongside the majority language. This can make it impossible to establish a balanced design which accounts for language-specific influences by including the languages of interest both as L1 and L2. On the other hand, a comparison of voice quality in the use of endangered languages across generations could provide information about the cognitive load associated with speaking these languages without being masked by possible language-specific voice characteristics. For example, a decrease of vocal fold perturbation from older to younger generations in the L2 might indicate a cross-generational increase of cognitive load in speaking the L2, even if this language would generally be associated with less vocal fold perturbation than the L1. In the latter case, increased cognitive load in younger speakers might show up as an additive effect on vocal fold vibration.

A decline in language use from the older to the younger generation is reported for Low German (LG), which historically was spoken monolingually in northern Germany but now is only acquired alongside High German (HG). Recent surveys indicate a sharp decline in the number of speakers of LG and its frequency of use [31].

This study aimed to explore how using LG and HG affects cognitive load across bilingual speakers of different ages in a local speech community. We analyzed measures of vocal fold vibration that are known to be linked to cognitive load in speech production. From [31] we expected that for younger speakers, using LG would be more demanding than using HG, while for older speakers, using LG would be less or equally demanding compared to HG. Accordingly, we hypothesized that younger speakers would show a decrease in jitter and shimmer when speaking LG, as well as an increase in HNR, CPP, and mean f0. In contrast, we expected that older speakers would either exhibit no significant differences or demonstrate an opposite pattern of results for the two languages. Furthermore, we explored whether changes in voice quality could be more directly associated with the relative dominance of LG over HG in bilingual speakers. To accomplish this, we calculated a dominance score for LG and HG and used this score instead of age to predict voice quality features.

2. METHOD

2.1. Participants and data collection

We recruited 95 participants (47 females, 48 males, aged 15 to 88 years) from the municipality of Krummhörn in East Frisia, in the northwest of the federal state of Lower Saxony in Germany. After removing four participants with incomplete data, the final analysis included 91 subjects. All subjects were native speakers of the local variety of Low German, known as East Frisian Low German, and acquired High German simultaneously or at the latest when they entered elementary school.

To explore the relationship between age and language dominance, we administered a survey to all participants which included 55 questions adapted from [31] and [32]. These questions covered topics such as age of acquisition, frequency of language use, and language proficiency. Using the participants' responses, we calculated global language scores for LG and HG for each subject. To predict voice characteristics, we calculated a dominance score by subtracting the global language score for HG from that for LG. We used this score as an alternative predictor variable to the subject's age in our regression models. Positive values indicate dominance of LG over HG, while negative values indicate dominance of HG over LG. The dominance scores of our participants ranged from -119 to +58, within a possible value range of -174 to +174.

Speech samples were recorded with a head-mounted omnidirectional microphone (DPA 4066) and a portable digital recorder (Tascam DR-100 MKIII). The recordings were digitized at 48 kHz sampling rate with 24 bits/sample quantization.

2.2. Procedure, data selection, and acoustic analysis

We asked participants to recite the weekday names from memory in each language, assuming their familiarity with these names in both languages. To avoid order effects, we randomized the language order for each participant.

To minimize potential segmental influences, we selected the /a/ vowel of the last syllable of the LG and HG weekday names for *Monday*, *Tuesday*, *Thursday*, *Friday*, *Saturday*, and *Sunday* for analysis (N = 1092). In LG, these syllables correspond to *dag*, while in HG, they correspond to *tag*. In careful speech, LG *dag* is realized as [da:x] and HG *tag* as [ta:k]. However, in informal speech we also find the variants [da:x] and [dax] in both languages. Inspection of individual tokens revealed that LG onset consonants matched HG consonants in 42% of cases, while LG coda consonants matched HG consonants in 49% of cases. To minimize the impact of neighboring consonants, we extracted the steady-state portion of the target vowel. Furthermore, we excluded vowels that exhibited creak or other unusual vibratory patterns, and we eliminated all instances of the final list member, *Sunday*, from the analysis because they exhibited a high proportion of creak and a lower signal-to-noise ratio. To increase the reliability of measurements, we excluded vowels with a steady-state portion of less than 80 ms, leaving us with a total of 826 vowel tokens for analysis.

We obtained the five voice parameters local jitter, local shimmer, Harmonics-to-Noise-Ratio (HNR), Cepstral Peak Prominence smoothed (CPPs), and mean fundamental frequency (mean f0) from the extracted vowels using default settings in *Praat* [33]. For HNR, CPPs, and mean f0, recordings were downsampled to 16 kHz before analysis. Mean f0 was estimated with the cross-correlation algorithm.

2.3. Statistical analysis

For the dependent variables *Global Language Score* as well as *Jitter*, *Shimmer*, *HNR*, *CPPs*, and *Mean f0* we fitted generalized linear mixed models using the *glmmTMB* package [34] in R [35]. As fixed effects, we included the within-subject effect Language (LG vs. HG) and the between-subject factors Gender (female vs. male) and Age along with interactions between these factors. For all dependent variables, except *Global Language Score*, we fitted additional

models with Dominance Score as a predictor variable in place of Age. As random effects, we had intercepts for both Subject and Item (weekday names), as well as by-subject random slopes for the effect of Language. Order of elicitation (LG or HG first) and consonantal context (*dag*, *tag*, etc.) did not enhance model fit and were therefore not included as random factors. Since the residuals for *Jitter*, *Shimmer*, and *Mean f0* were not normally distributed, we fitted generalized linear mixed models with the Gamma distribution and log-link function for these variables. For the remaining variables, we fitted generalized linear mixed models with the Gaussian distribution. Likelihood Ratio Tests were used to obtain *p*-values for the full models. We will focus on reporting the main effect of Language and interactions involving Language and either Age or Dominance Score. To examine contrasts between slopes for LG and HG, we utilized the *emtrends* function of the R package *emmeans* [36]. The level of significance was set at $p = 0.05$.

3. RESULTS

Figure 1 shows the global language scores for LG and HG as a function of Age. Statistical analysis revealed a significant main effect of Language ($\chi^2 = 55.5619$, $p < .0001$), and a significant interaction between Age and Language ($\chi^2 = 28.7505$, $p < .0001$), but no interaction with Gender. Contrast analysis showed a significant difference between the slopes for LG and HG, suggesting that in younger speakers, HG tended to dominate over LG, with this difference gradually diminishing with age (Estimate = 1.3337, *SE* = 0.2505, *t*-ratio = 5.324, $p < .0001$). Figure 1 indicates a reversal of the dominance ratio for both genders, with LG becoming more dominant than HG around the age of 75. Note, however, that our sample size was limited to only four male speakers aged over 75, while the oldest female speaker was 70 years old.

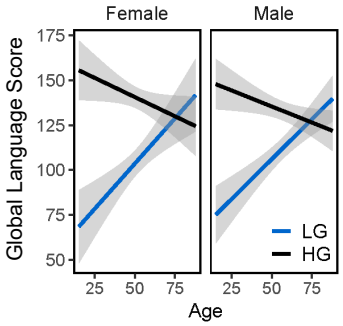


Figure 1: Regression lines for Low German (LG) and High German (HG) global language scores as a function of Age. Shading indicates 95% confidence intervals.

Figure 2 displays regression lines for *Jitter*, *Shimmer*, *HNR*, *CPPs*, and *Mean f0* for LG and HG, with Age (left panels) and Dominance Score (right panels) as predictor variables.

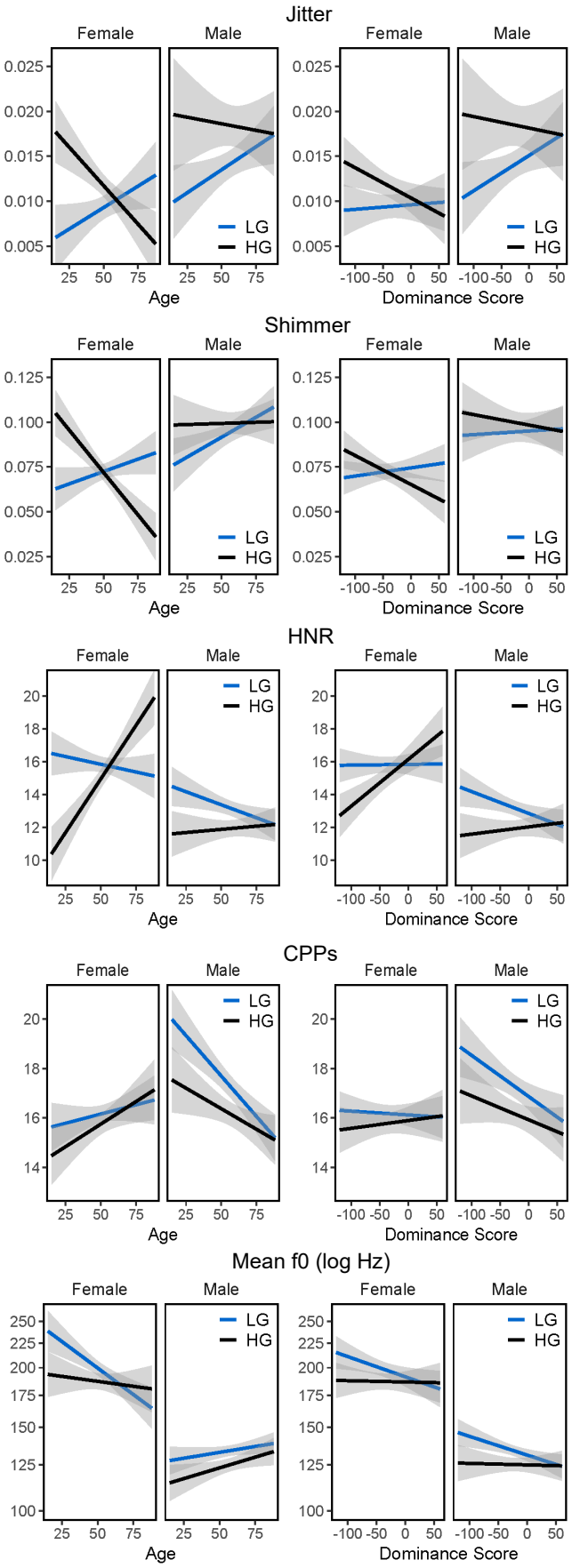


Figure 2: Regression lines for *Jitter*, *Shimmer*, *HNR*, *CPPs*, and *Mean f0* for LG and HG with Age (left panels) and Dominance Score (right panels) as predictor variables. Shading indicates 95% confidence intervals.

In the upper part of Table 1, we report significant interactions between Age and Language for *Jitter*, *Shimmer*, *HNR*, *CPPs*, and *Mean f0*. An additional interaction with Gender was found for *Jitter*, *Shimmer*, and *HNR*. The lower part of Table 1 presents significant interactions between Dominance Score and Language for all dependent variables, except for *CPPs*, where we observed only a main effect of Language.

Contrast analyses in Table 2 reveal in the upper part significant differences between the slopes of LG and HG for all dependent variables as a function of Age, with a greater difference between the slopes of LG and HG for *Jitter*, *Shimmer*, and *HNR* in female speakers compared to male speakers. In the lower part, we report significant differences between the slopes of LG and HG for *Jitter*, *Shimmer*, *HNR*, and *Mean f0* as a function of Dominance Score.

Table 1: Assessing model fit (χ^2) with Age (upper part) and Dominance Score (DS) (lower part) as predictor variables.

	Language	Age:Language	Age:Language:Gender
Jitter	31.5755***	23.5804***	4.0214*
Shimmer	0.2747	17.9844***	4.0175*
HNR	17.9314***	28.5918***	9.5736**
CPPs	15.5287***	4.4534*	0.0699
Mean f0	16.1781***	4.7376*	2.8276
	Language	DS:Language	DS:Language:Gender
Jitter	25.7868***	7.6465**	0.0015
Shimmer	0.2547	5.5552*	1.5447
HNR	16.4462***	12.7671***	0.3652
CPPs	14.9604***	1.2956	0.0031
Mean f0	15.5765***	5.4976*	0.0015

p-levels: * < 0.05, ** <0.01, *** <0.001

Table 2: Contrasts between LG and HG slopes for interactions involving Language and either Age (above) or DS (below).

Variable	Estimate	SE	*t*/*z*-ratio[1]	*p*
Jitter-female	0.0215	0.0047	4.584	<.0001
Jitter-male	0.0095	0.0037	2.566	.0103
Shimmer-female	0.0137	0.0032	4.215	<.0001
Shimmer-male	0.0053	0.0026	2.054	.0400
HNR-female	-0.1466	0.0256	-5.728	<.0001
HNR-male	-0.0462	0.0200	-2.312	.0210
CPPs	-0.0276	0.0139	-1.993	.0466
Mean f0	-0.0025	0.0010	-2.505	.0122
Jitter	0.0033	0.0012	2.767	.0057
Shimmer	0.0019	0.0008	2.304	.0212
HNR	-0.0019	0.0005	-3.557	.0004
Mean f0	-0.0009	0.0004	-2.345	.0190

[1]*t*-ratios: Gaussian distributions; *z*-ratios: Gamma distributions.

4. DISCUSSION

The aim of this study was to explore whether acoustic measures of vocal fold vibration reflect differences in cognitive load during bilingual speech. Specifically, we examined the acoustic parameters of /a/ vowels in weekday names produced by bilingual speakers in their two languages, LG and HG. We observed a decrease in jitter and shimmer and an increase in HNR, CPPs, and mean f0 in LG speech compared to HG speech in most participants. These results suggest that LG speech has higher stability, periodicity, and rate of vocal fold vibration, which has previously been associated with increased cognitive load [1]-[12]. However, we also observed that the effects on vocal fold vibration varied among speakers of different ages. For younger subjects, the acoustic analysis indicates higher cognitive load when speaking LG than when speaking HG. Interestingly, this difference reduced with increasing age. Therefore, our findings suggest that the additional cognitive load required for speaking LG instead of HG decreases as age increases. These findings are consistent with the strong correlation found between age and the dominance scores for LG and HG. Specifically, the younger the speakers were, the more HG dominated over LG. However, the dominance score of our speakers turned out to be a slightly weaker predictor for the dependent variables than age.

Our study demonstrates that acoustic measures can detect the effects of cognitive load on voice quality in bilinguals, irrespective of any language-specific effects. Although there may be differences in LG and HG speech with respect to the acoustic parameters examined, we were able to identify changes in vocal fold vibration associated with differences in cognitive load by comparing speakers differing in age and language dominance. However, our study was limited to reciting weekday names, and future research is necessary to investigate whether similar effects can be observed in other speech tasks.

5. CONCLUSION

Our findings support previous studies suggesting that increased stability, periodicity, and rate of vocal fold vibration result in decreased jitter and shimmer, as well as increased HNR, CPP(s), and mean f0, which can serve as indicators of increased cognitive load. Our results suggest that younger, non-balanced bilingual speakers of Low and High German exhibit these changes when speaking their weaker language, while the differences tend to disappear in older, more balanced bilinguals. These results align with surveys of language use and language proficiency, which indicate an increase of dominance of High German over Low German with decreasing age. By examining cognitive load through voice parameters, our study provides a valuable alternative to self-assessments and language proficiency tests in assessing the mastery of endangered languages.

6. ACKNOWLEDGMENTS

We gratefully acknowledge the support of the German Research Foundation (DFG) through grant no. PE 793/3-1.

7. REFERENCES

[1] Giddens, Ch., Barron, K., Byrd-Craven, J., Clark, K., Winter, A. 2013. Vocal indices of stress. A review. *Journal of Voice* 27, 390.e21-390.e29.

[2] Kirchhübel, Ch., Howard, D., Stedmon, A. 2011. Acoustic correlates of speech when under stress: research, methods and future directions. *Aviat. Space Environ. Med.* 18, 75–98.

[3] Puyvelde, M. van, Neyt, X., McGlone, F., Pattyn, N. 2018: Voice stress analysis: A new framework for voice and effort in human performance. *Front. Psychol.* 9, DOI: 10.3389/fpsyg.2018.01994.

[4] Boyer, St., Paubel, P.-V., Ruiz, R., El Yagoubi, R., Daurat, A. 2018. Human voice as a measure of mental load level. *JSLHR* 61, 2722–2734.

[5] Brenner, M., Shipp, Th., Doherty, E., Morrissey, P. 1985. Voice measures of psychological stress: Laboratory and field data. *Vocal fold physiology, biomechanics, acoustics, and phonatory control*, 239–248.

[6] Kappen, M., et al. 2022. Acoustic speech features in social comparison: how stress impacts the way you sound. *Sci. Rep*., 12. DOI: 10.1038/s41598-022-26375-9

[7] Mendoza, E., Carballo, G. 1998. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice* 12, 263–273.

[8] Rothkrantz, L., Wiggers, P., van Wees, J.-W., van Vark, R. 2004. Voice stress analysis. *International conference on text, speech and dialogue*, 449–456.

[9] MacPherson, M., Abur, D., Stepp, C. 2017. Acoustic measures of voice and physiologic measures of autonomic arousal during speech as function of cognitive load. *Journal of Voice* 31, 504.e1-504.e9.

[10] Yap, T., Epps, J., Ambikairajah, E., Choi, E. 2015. Voice source under cognitive load: Effects and classification. *Speech Communication* 72, 74–95.

[11] Griffin, G., Williams, C. 1987. The effects of different levels of task complexity on three vocal measures. *Aviat. Space Environ. Med.* 58, 1165–1170.

[12] Scherer, K., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, Th. 2002. Acoustic correlates of task load and stress. *7th ICSLP*, Denver, Colorado, USA, 1-4.

[13] Chong, Y. 2012. *Vocal characteristics of English and Mandarin produced by Mandarin-English and English-Mandarin bilingual speakers. A long-term average spectral analysis*. BA Thesis, Univ. of Hong Kong.

[14] Zhu, Sh., Chong, S., Chen, Y., Wang, T., Ng, M. 2022. Effect of language on voice quality - An acoustic study of bilingual speakers of Mandarin Chinese and English. *Folia phoniatrica et logopaedica* 2022, 1-10.

[15] Altenberg, E., Ferrand, C. 2006. Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice* 20, 89–96.

[16] Chan, Y. 2010. *Acoustical differences in vocal characteristics between Cantonese and English produced by Cantonese-English bilingual adult speakers*. BA Thesis, University of Hong Kong.

[17] Mendez, N. 2017. *Do bilingual speakers shift fundamental frequency based on language acquisition or language dominance?* B.A. Thesis, Miami University.

[18] Mennen, I., Schaeffler, F., Docherty, G. 2007. Pitching it differently. A comparison of the pitch ranges of German and English speakers. *16th ICPhS*, 1769–1772.

[19] Ng, M., Chen, Y., Chan, E. 2012. Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers – A long-term average spectral analysis. *Journal of Voice 26*, e171-e176.

[20] Oh, H. 2011. Voice quality processing strategy of Korean learners of Chinese. *17th ICPhS*, 1526–1529.

[21] Peters, J., Frank, M., Rohloff, M. 2020. Pitch range variation in High German (L1) and Low German (L2). *Speech Prosody 2020*, 650–654.

[22] Ruß, A. 2015. *Stimmqualität zwischen Mutter- und Fremdsprache*. Saarbrücken: AV Akademikerverlag.

[23] Schwab, S., Goldman, J.-Ph. 2016. Do speakers show different F0 when they speak in different languages? The case of English, French and German. *Speech Prosody 2016*.

[24] Xue, St., Hagstrom, F., Hao, J. 2002. Speaking fundamental frequency characteristics of young and elderly bilingual Chinese' English speakers. A functional system approach. *Asia Pacific Journal of Speech, Language and Hearing* 7, 55–62.

[25] Engelbert, A. 2014. Cross-Linguistic Effects on Voice Quality: A Study on Brazilians' Production of Portuguese and English. *Proc. of the Int. Symp. on the Acquisition of Second Language Speech*. *Concordia Working Papers in Applied Linguistics*, 5, 157–170.

[26] Andreeva, B., Dimitrova, S. 2022. The influence of L1 prosody on Bulgarian-accented German and English. *Speech Prosody 2022*, 764–768.

[27] Kainada, E., Lengeris, A. 2015. Native language influences on the production of second-language prosody. *J. Int. Phon. Assoc.* 45, 269–287.

[28] Järvinen, K., Laukkanen, A.-M., Aaltonen, O. 2013. Speaking a foreign language and its effect on F0. *Logopedics, Phoniatrics, Vocology* 38, 47–51.

[29] Järvinen, K., Laukkanen, A.-M. 2015. Vocal loading in speaking a foreign language. *Folia Phoniatrica et Logopaedica* 67, 1–7.

[30] Theelen, M. 2016. *Adaptation of fundamental frequency in Dutch and English with second language effects*. B.A. Thesis. Univ. of Amsterdam.

[31] Adler, A., Ehlers, Ch., Goltz, R., Kleene, A., Plewnia, A. 2016. *Status und Gebrauch des Niederdeutschen 2016*. Mannheim: Institut für Deutsche Sprache.

[32] Poarch, G.J., Vanhove, J., Berthele, R. 2019. The effect of bidialectalism on executive function. *International Journal of Bilingualism* 23, 612–628.

[33] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. V. 6.2.16, retrieved 18 Aug. 2022 from http://www.praat.org/

[34] Brooks M., Kristensen K., van Benthem K., Magnusson A., Berg C., Nielsen A., Skaug H., Maechler M., Bolker, B. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 9, 378–400.

[35] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statist. Comp., Vienna, Austria, v. 4.2.1, www.R-project.org

[36] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2019. emmeans: Estimated marginal means, aka least-squares means (R package v. 1.1)

Effects of fundamental frequency and harmonics-to-noise ratio on the perception of Danish laryngealized phonation

Jailyn M. Peña

New York University
jailynpena@nyu.edu

ABSTRACT

This study examines the effect of fundamental frequency (F0) and harmonics-to-noise ratio (HNR) on the perception of laryngealized phonation in Danish. Previous research shows that compared to modal phonation, laryngealized phonation in Danish generally begins with high F0 at the beginning of the syllable rhyme, followed by low F0 and creaky phonation at the end of the syllable. In this study, F0 and HNR were manipulated in naturally produced nonce words to examine how high F0, low F0, and HNR contribute to the percept of laryngealization. Results from an ABX task and a perceptual rhyming task show that listeners were more likely to identify words with both (1) high F0 and (2) a drop in pitch or lower HNR on the coda sonorant as 'laryngealized' compared to other words. Words with only one laryngealization cue did not promote 'laryngealized' responses, indicating that listeners utilize multiple acoustic cues to perceive laryngealization.

Keywords: laryngealization, Danish, phonation, voice quality, speech perception.

1. INTRODUCTION

This study investigates the effects of fundamental frequency (F0) and harmonics-to-noise ratio (HNR) on the perception of the Danish phonation contrast. Danish phonologically contrasts modal phonation and a type of non-modal phonation similar to creaky phonation, e.g., *ven* [vɛn] 'friend' vs. *vend* [vɛnˀ] 'turn!' [1–3]. This non-modal phonation (here, [ˀ]) can only occur in stressed syllables with either a long vowel or a vowel followed by a sonorant coda consonant in the standard dialect and some dialects spoken in Zealand. Prior literature has referred to this phonation type as 'stød,' though here it will be referred to as laryngealization, following descriptions in Grønnum et al. [4]. This is to acknowledge non-modal phonation in Danish is not identical to creak but is on the creakier end of the phonation spectrum put forth by Gordon & Ladefoged [5].

Acoustically, differences in phonation are usually quantified using measures of spectral tilt and noise. Laryngealized segments are often produced with a smaller difference between the amplitudes of the first and second harmonics (lower H1*-H2*, where the asterisks indicate correction for formants) and lower HNR, indicating more noise in the acoustic signal compared to modally phonated segments [6–11]. Additionally, laryngealized segments may also have lower F0 than modal ones [12].

A unique feature of Danish is that phonologically laryngealized syllables are acoustically characterized by high F0 and modal phonation at the beginning of the rhyme. This first phase is then followed by a second phase with a sharp drop in F0 and intensity, irregular amplitude, and oftentimes creaky or compressed phonation (that is, a period of truly non-modal phonation). In terms of spectral tilt and noise, previous work has found a connection between H1-H2, HNR, and laryngealization in Danish [13]. However, previous research has shown that the realization of the second phase is highly variable, and no single acoustic cue is present in all instantiations [2, 14]. Furthermore, studies have yet to investigate how the various acoustic cues that differentiate modal and laryngealized phonation in production affect native Danish listeners' perception of this phonological phonation contrast.

In this study, we focus on the effects of F0 and noise, quantified in terms of HNR, on native Danish listeners' perception of the phonation contrast. The effect of spectral tilt is not examined here, as techniques for manipulating it are not as well developed as those for F0 and noise.

2. METHODS

38 native Danish listeners between 18-40 ($\mu = 27.3$, $\sigma = 3.9$) participated in the study in Copenhagen, Denmark. Due to the highly variable nature of laryngealized phonation in Danish, two tasks were conducted to probe different aspects of perception and ensure the reliability of the results. An ABX task was conducted to facilitate a direct comparison of our test stimuli with both modally phonated and laryngealized words. A rhyming task was also conducted as a test of listeners' metalinguistic awareness of laryngealized phonation and their ability to generalize this knowledge to novel words.

All listeners were from the dialectal region of Zealand, Denmark where the distribution of laryngealized syllables in the lexicon is the same as

in Standard Danish. Listeners performed the tasks in a quiet room on an Acer Spin 5 laptop wearing Sennheiser HD 400S headphones and recorded their answers using a DirectIN high speed button box. All listeners performed the ABX task before the rhyming task. Three listeners who incorrectly performed the rhyming task were excluded from that analysis; all listeners' data were included in the ABX analysis.

To create the stimuli for the two tasks, eight pairs of monosyllabic C(C)VC nonce words produced with both modal and laryngealized phonation as well as four pairs of real monosyllabic words phonologically contrasting in phonation were recorded in a short nonsense phrase consisting of the target word followed by the word *videre* [viðʌ] 'continue.' The recordings were made by a female native Danish speaker (age 27) using a Tascam DR 40x recorder and a head-mounted Shure SM35 condenser microphone at a sampling rate of 44,100 Hz. Examples are given in Table 1. The real words and the laryngealized nonce words were recorded once, and the modally phonated nonce words were recoded twice for a total of 32 recordings. Examples are given in Table 1.

	Nonce Words	**Real Words**
[an]	[gan], [gan^{ʔ}]	*man* [man] 'but' *mand* [man^{ʔ}] 'man'
[al]	[nal], [nal^{ʔ}]	*tal* [t^{h}al] 'number' *bal* [bal^{ʔ}] 'ball'
[ɛn]	[jɛn], [jɛn^{ʔ}]	*ven* [vɛn] 'friend' *vend* [vɛn^{ʔ}] 'turn!'
[ɛl]	[nɛl], [nɛl^{ʔ}]	*vel* [vɛl] 'well' *væld* [vɛl^{ʔ}] 'abundance'

Table 1: Examples of stimuli.

The target words were excised from the following word, and vowel and coda sonorant durations were equalized in Praat [15] to approximately 111 ms each for a total rhyme duration close to 222 ms per recording (1.5 times the original average). This was done because preliminary inspection revealed that the original average rhyme duration sounded unnaturally short in isolation. Minor variations in duration occurred based on the period of the soundwave of the original recordings to prevent acoustic distortions. Onset durations varied from 15.6 – 96.6 ms based on the voicing, manner, and number of consonants. No further manipulations were made to the real words, the laryngealized nonce words, and one set of modally phonated nonce words.

To examine how F0 and HNR affect listeners' perception of the phonation contrast, the second set of modally phonated nonce words (16 recordings) were acoustically manipulated to show one of 6 F0 contours (HH, HL, HM, LL, ML, MM, where 'H' indicates high pitch, 'M' indicates mid pitch, and 'L' indicates low pitch) using a Praat script [16]. The first letter in the sequence indicates the relative F0 of the vowel and the second letter indicates the relative F0 of the coda sonorant. The high F0 value (231 Hz) was based on the speaker's average pitch for real laryngealized words during the first fifth of the vowel. The low F0 value (179 Hz) was selected as the lowest naturalistic F0 value that could be synthesized in Praat. The mid F0 value (205 Hz) was the middle of the two extremes, and only 5 Hz higher than the speaker's own average F0 for modally phonated words during the first fifth of the vowel. These 16 recordings were then duplicated. 30% jitter was then added to the coda sonorant of the resulting files using the *raspiness* function from the Praat Vocal Toolkit [17] to simulate the percept of creakiness or laryngealization, following Huang [18]. This manipulation decreases HNR by introducing more aperiodicity into the acoustic signal. Overall, this process produced 12 types of stimuli for a total of 96 unique stimulus items (6 F0 contours x 2 phonation conditions x 8 unique nonce words). Henceforth these recordings will be referred to as the test stimuli, or simply the stimuli.

Each task consisted of 192 test trials. During the ABX task, for each trial listeners first heard the same nonce word twice, once with naturally produced laryngealization and once with modal phonation, with the order counter-balanced across the experiment. Listeners then heard one of the test stimuli and were instructed to indicate whether they thought the test stimulus sounded more like the modal or laryngealized word by pressing a button on a button box. During the rhyming task, listeners heard a real word with either naturally produced laryngealization or modal phonation before hearing a test stimulus. They were then instructed to indicate whether the two words rhymed by pressing a button on the button box.

The results of the ABX task were coded in binary based on whether a test stimulus was matched to a laryngealized nonce word (1) or not (0). A similar coding scheme was used for trials of the rhyming task in which listeners were asked to rhyme the test stimuli with real, laryngealized words. Trials in which listeners rhymed test stimuli to real, modally phonated words were coded separately based on whether a test stimulus was matched to a real, modally phonated word (1) or not (0). Three logistic mixed effects models were then conducted in R using the *lme4* package [19] with matches to laryngealized/modal words as the dependent variable and the 12 test stimuli types as a sum-coded fixed effect. Listener, trial, and nonce word were included as random intercepts, and an additional binary fixed effect was included in the ABX model to account for a response bias in favour of the second word. In

addition, to determine which types of stimuli promoted more 'laryngealized' responses than chance (50%), binomial tests were conducted on the stimuli types that promoted significantly more 'laryngealized' responses on either task using the *binom.test()* function in R.

3. RESULTS

As shown in Figure 1 and Table 2, for the ABX task, listeners were more likely to match HHʔ ('ʔ' here indicating added raspiness), HLʔ, HM, and HMʔ test stimuli to laryngealized nonce words compared to other stimulus types. This is indicated by the black asterisks in Figure 1 and grey shading in Table 2. LL, LLʔ, and MM stimuli were less likely to be matched to laryngealized nonce words, indicated by the grey asterisks in Figure 1.

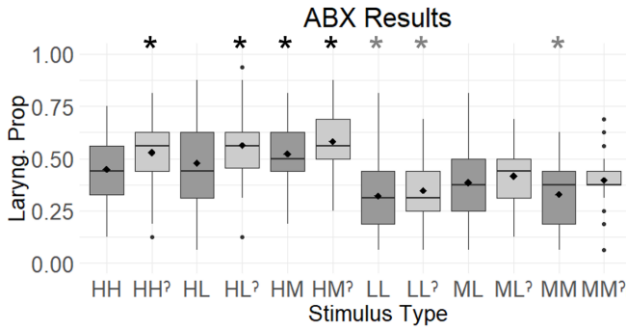


Figure 1: Proportion of 'laryngealized' responses for the ABX task. Box shading indicates whether the stimulus type had added raspiness (light grey) or not (dark grey). Asterisk color indicates whether the stimulus type promoted significantly more 'laryngealized' responses (black) or less (grey).

On the rhyming task, listeners were more likely to rhyme HLʔ, HM, and HMʔ words and less likely to rhyme LL, LLʔ, ML, and MM words to laryngealized words, as shown in Figure 2 and Table 2.

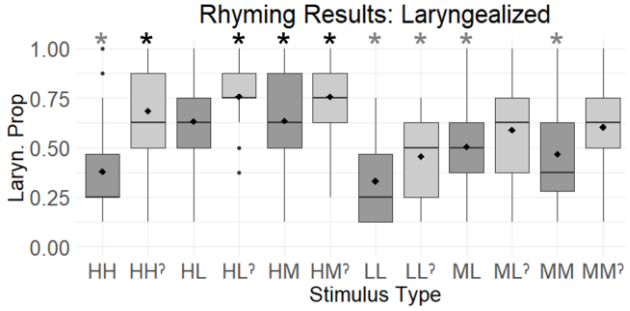


Figure 2: Proportion of 'laryngealized' responses for the rhyming task. Box shading indicates whether the stimulus type had added raspiness (light grey) or not (dark grey). Asterisk color indicates whether the stimulus type promoted significantly more 'laryngealized' responses (black) or less (grey).

Stimuli	Task	Results
HH	ABX	β = 0.05, *z* = 0.226, *p* = *0.821*
	Rhyme	β = -0.95, *z* = -5.293, *p* < *0.001**
HHʔ	ABX	β = 0.43, *z* = 2.056, *p* = *0.040**
	Rhyme	β = 0.55, *z* = 3.046, *p* = *0.002**
HL	ABX	β = 0.22, *z* = 1.036, *p* = *0.300*
	Rhyme	β =0.32, *z* = 1.850, *p* = *0.064*
HLʔ	ABX	β = 0.61, *z* = 2.939, *p*= *0.003**
	Rhyme	β =1.13, *z* = 6.010, *p* < *0.001**
HM	ABX	β = 0.42, *z* = 2.021, *p* = *0.043**
	Rhyme	β = 0.44, *z* = 2.486, *p* = *0.013**
HMʔ	ABX	β = 0.72, *z* = 3.449, *p* < *0.001**
	Rhyme	β = 1.13, *z* = 6.023, *p* < *0.001**
LL	ABX	β = -0.65, *z* = -3.047, *p* = *0.002**
	Rhyme	β = -1.43, *z* = -7.583, p < *0.001**
LLʔ	ABX	β = -0.54, *z* = -2.556, *p* = *0.011**
	Rhyme	β = -0.54, *z* = -3.009, *p* = *0.002**
ML	ABX	β = -0.29, *z* = -1.401, *p* = *0.161*
	Rhyme	β = -0.41, *z* =-2.383, *p* = *0.017**
MLʔ	ABX	β = -0.12, *z* = -.0513, *p* = *0.608*
	Rhyme	β = 0.22, *z* =1.234, *p* = *0.217*
MM	ABX	β = -0.61, *z* = -2.908, *p* = *0.003**
	Rhyme	β = -0.70, *z* = -4.037, *p* < *0.001**
MMʔ	ABX	β = -0.28, *z* = -1.357, *p* = *0.175*
	Rhyme	β = 0.26, *z* = 1.517, *p* = *0.129*

Table 2: Logistic regression results for the ABX task and laryngealized trials of the rhyming task. Shaded rows indicate stimulus types with significantly more 'laryngealized' responses.

The results of the statistical analysis on the modal match data from the rhyming task were similar to that of the other two models. Listeners were significantly less likely to rhyme HL, HLʔ, HMʔ, and MLʔ stimuli to real, modally phonated words than other test stimulus types. Conversely, HH, LL, LLʔ, MM, and MMʔ stimuli were significantly more likely to be rhymed with real, modally phonated words. These results are shown in Figure 3 and Table 3.

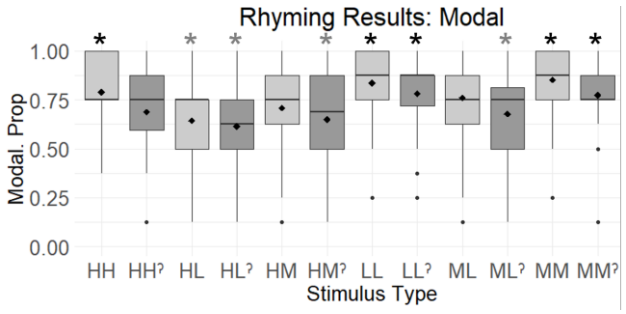


Figure 3: Proportion of 'modal' responses for the rhyming task. Box shading indicates whether the stimulus type had added raspiness (light grey) or not (dark grey). Asterisk color indicates whether the stimulus type promoted significantly more 'modal' responses (black) or less (grey).

Stimuli	Rhyming: Laryngealized Trials
HH	β = 0.47, z = 2.968, p = *0.003**
HHˀ	β = -0.16, z = -1.107, p = *0.268*
HL	β = -0.55, z = -4.195, p < *0.001**
HLˀ	β = -0.69, z = -5.352, p < *0.001**
HM	β = -0.21, z = -1.529, p = *0.126*
HMˀ	β = -0.51, z = -3.903, p < *0.001**
LL	β = 0.82, z = 4.799, p < *0.001**
LLˀ	β = 0.41, z = 2.657, p = *0.008**
ML	β = 0.08, z = 0.587, p = *0.558*
MLˀ	β = -0.27, z = -1.978, p = *0.048**
MM	β = 0.92, z = 5.354, p < *0.001**
MMˀ	β = 0.31, z = 2.103, p = *0.003**

Table 3: Logistic regression results for the modal trials of the rhyming task. Shaded rows indicate stimulus types with significantly more 'modal' responses.

Finally, binomial tests were performed for each test stimulus type which promoted significantly more 'laryngealized' responses than the overall average on either task (HHˀ, HLˀ, HM, HMˀ stimuli). The results show that on the ABX tasks, HLˀ and HMˀ test stimuli promoted significantly more 'laryngealized' responses than chance. All stimulus types tested promoted more 'laryngealized' responses than chance on the rhyming task. These results are summarized in Table 4, along with the percentage of 'laryngealized' responses per stimulus type.

Stimuli	ABX	Rhyming
HHˀ	53.5%, p = 0.109	68.4%, p < 0.001*
HL	47.9%, p = 0.155	64.1%, p < 0.001*
HLˀ	56.3%, p = 0.001*	77.0%, p < 0.001*
HM	52.3%, p = 0.137	65.8%, p < 0.001*
HMˀ	58.2%, p < 0.001*	77.6%, p < 0.001*

Table 4: Binomial test p values and percentage of 'laryngealized' responses per stimulus type by task.

4. DISCUSSION

The results of this study reveal three significant findings. First, listeners only perceived nonce words as 'laryngealized' when they began with a high F0 vowel, regardless of whether there were other acoustic cues to laryngealization in the stimuli (e.g., low F0 or low HNR on the coda sonorant). This suggests that there is a hierarchy between the acoustic cues to laryngealization, with high F0 on the vowel at the top.

Second, listeners only perceived words as 'laryngealized' if there were at least two cues correlating with the unique acoustics of Danish laryngealization. That is, listeners were more likely to perceive words as 'laryngealized,' compared to other test stimulus types, only if the word had both high F0 on the vowel and either a drop in F0 or low HNR (raspiness) on the coda sonorant. This result indicates that listeners utilized multiple, redundant cues to laryngealization in perception.

Furthermore, listeners were able to utilize either low F0 or low HNR as a cue to the second phase of laryngealization in Danish, which is often associated with creaky or compressed phonation. This indicates that listeners' perception of the second phase of laryngealization does not hinge on a single, specific acoustic cue. Rather, listeners may exploit various acoustic cues associated with the non-modal portion of the word to perceive the word as 'laryngealized.' This matches with previous descriptions in the Danish literature, which describe the production of the second phase of laryngealized syllables as extremely variable, acoustically [2, 14].

Finally, the more acoustic cues to laryngealization that were present, the more likely listeners were to identify test stimuli as 'laryngealized,' given the first two generalizations. That is, HLˀ and HMˀ words, which included three acoustic cues to laryngealization (high F0 on the vowel, an F0 drop, and low HNR on the coda sonorant) promoted the most 'laryngealized' responses on both tasks. These results were corroborated by the results of the binomial tests, which found that listeners were above chance at identifying HLˀ and HMˀ words as 'laryngealized' on both tasks.

Also of note is the high degree of agreement between the results of the two tasks. The ABX task was designed to examine which acoustic cue(s) that we tested were associated with the perception of laryngealized phonation by directly pitting modally phonated and laryngealized words against each other. The rhyming task, on the other hand, was designed to test listeners' ability to generalize their knowledge of the Danish phonation contrast to novel forms. Despite the differences between these two tasks, however, the results converged such that on both tasks the same types of test stimuli significantly promoted more 'laryngealized' response compared to other stimulus types. This indicates that Danish listeners are not only sensitive to the fine-grained acoustics of the phonation contrast, as shown by the ABX task, but that they are also able to generalize this knowledge to novel forms.

Overall, these results indicate that the perception of laryngealized phonation in Danish is complex, paralleling results of productions studies which find that the acoustic realization of laryngealization is multifaceted. Future research should investigate the role of other acoustic cues to laryngealization, such as spectral tilt, to the perception of the Danish phonation contrast.

5. REFERENCES

[1] Basbøll, H. 2005. *The phonology of Danish*. Oxford: Oxford University Press.

[2] Fischer-Jørgensen, E. 1989. Phonetic Analysis of the Stød in Standard Danish. *Phonetica*, 46, 1–59.

[3] Grønnum. N. 1998. Danish. *Journal of the International Phonetic Association*, 28, 99–105.

[4] Grønnum, N., Vazquez-Larruscaín, M., Basbøll, H. 2013. Danish Stød: Laryngealization or Tone, *Phonetica*, 70, 66–92.

[5] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29, 383-406.

[6] Garellek, M., Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *The Journal of the Acoustical Society of America*, 41, 185–205.

[7] Esposito, C.M. 2012. An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, 40, 466–476.

[8] DiCanio, C.T. 2012. Coarticulation between tone and glottal consonants in Itunyoso Trique. *J. Phonetics*, 40, 162–176.

[9] Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America*, 142, 1693–1706.

[10] K Yu, K.M., Lam, H.W. 2014. The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, 136,1320–1333.

[11] Jiang, B., Clayards, M., Sonderegger, M. 2020. Individual and dialect differences in perceiving multiple cues: A tonal register contrast in two Chinese Wu dialects. *Laboratory Phonology*, 11.

[12] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of the 18th International Congress of Phonetic Sciences*, 1, 2-7.

[13] Kirkedal, A.S. 2016. Danish Stød and Automatic Speech Recognition. Frederiksberg: Copenhagen Business School (CBS).

[14] Grønnum, N. Basbøll, H. 2001. Consonant length, stød and morae in standard Danish. *Phonetica*, 58, 230–253.

[15] Boersma P. Weenink, D. 2021. Praat: doing phonetics by computer [Computer program].

[16] Chai, Y. 2021. Perception of checked tones in Xiapu Min. *The Journal of the Acoustical Society of America*, 150, A309–A310.

[17] Corretge, R. 2022. Praat Vocal Toolkit. [Online]. Available: http://www.praatvocaltoolkit.com.

[18] Huang, Y. 2020. Different attributes of creaky voice distinctly affect Mandarin tonal perception. *The Journal of the Acoustical Society of America*, 147, 1441–1458.

[19] Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. Fitting Linear Mixed-Effects Models using lme4.

DURATION AS A CUE FOR PHONOLOGICAL VOICING CONTRAST IN WHISPERED CZECH

Michaela Svatošová, Tomáš Bořil

Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czech Republic
michaela.svatosova@atarien.com, tomas.boril@ff.cuni.cz

ABSTRACT

In Czech, the phonological contrast of voicing is primarily realized by the presence or absence of the fundamental frequency. However, this main correlate is missing in whisper, because the vocal folds do not vibrate. The present study explored the acoustical and perceptual side of this phonological contrast. Firstly, it compared the duration of voicing counterparts of Czech plosives and fricatives embedded in modal and whispered pseudowords. The duration of voicing counterparts differed significantly, but the durational ratios were smaller in whisper than in modal phonation. Secondly, a perception experiment was created from the whispered stimuli, assessing the recognisability of whispered obstruents in phonetic context only. Listeners recognised many obstruents especially in the medial position, but substantial variability between individual voicing pairs was found.

Keywords: phonological voicing, whisper, Czech, perception, duration

1. INTRODUCTION

Whisper is used in various communicative situations with many functions [1]. In most cases, speakers want to be understood and the ongoing usage of whisper therefore suggests its (at least partial) intelligibility. Whisper is a type of phonation defined by the absence of vocal folds vibration (the only source signal is noise). However, in many languages, the presence or absence of voicing is exploited phonologically and some minimal pairs could become indistinguishable in whisper. The situational context of a given utterance provides syntactic, semantic and pragmatic cues that usually disqualify one member of the minimal pair. Nevertheless, the question of the relative importance of these cues remains.

Experiments that placed whispered voicing pairs in isolated syllables or short words (limiting the context to phonetic cues) have showed that listeners were able to discriminate voicing counterparts with an above chance success [2, 3, 4]. Other studies have investigated various acoustic cues that could differentiate voicing counterparts, suggesting duration of the target consonants as the most promising parameter [3, 4, 5, 6, 7, 8]. However, the majority of the above mentioned research concerned English in which phonological voicing strongly depends on aspiration (less affected in whisper). A deeper examination of the production and perception of voicing pairs in languages like Czech, which base this contrast on phonetic voicing [9], could be beneficial. Phonological voicing has been rarely studied for whispered Czech, but [10] have found differences in the duration of fricative voicing counterparts and the ability of speakers to distinguish them in minimal phonetic context.

The present study continues in this line of research, but extends the set of voicing pairs to fricatives and plosives. It addresses the perceptual aspect through the evaluation of listeners' ability to identify the voicing counterparts in whisper solely on the basis of phonetic cues (in nonsense pseudowords). From the perspective of production, it compares the duration (a possible correlate of phonological voicing) of voicing counterparts in both phonation types. The main research questions can be summarised as follows:

- Is the phonological contrast of voicing perceptually retained in whisper?
- Do the voicing counterparts differ in duration (in both phonation types)?

2. METHOD

2.1. Material

The study included 7 pairs of Czech obstruents with the same place of articulation, four plosive and three fricative pairs: /p b/, /t d/, /c ɟ/, /k g/, /f v/, /s z/, /ʃ ʒ/. These 14 phonemes were combined with the vowel /a/ into simple CV syllables, subsequently forming trisyllabic pseudowords adhering to Czech phonotactics [9]. Only consonants in the first and second syllable were analysed (the *initial* and *medial* position). Each phoneme appeared in ten different pseudowords for each position, creating a total of

140 unique stimuli (each containing an initial and medial target obstruent). These were complemented by 70 filler pseudowords. All non-analysed syllables featured a wider range of phonemes in order to avoid monotony. Each combination of initial and medial target obstruent repeated maximally three times and the stimuli were controlled for similarity with real Czech words. A sample of target and filler pseudowords is presented in Table 1.

Target stimuli			Fillers
/ta**ʃ**amu/	/**z**a**v**ako/	/**f**a**d**aju/	/jovukɪ/
/**k**a**t**apɪ/	/**p**a**v**alɛ/	/**ʒ**a**ɟ**afo/	/lɛxara/

Table 1: Examples of target and filler pseudowords. Target phonemes are in bold.

2.2. Speakers and recording

All stimuli were randomly divided into 5 blocks (the assignment was done separately for each phonation). The pseudowords were written in normal Czech ortography and they were placed in lines of six (separated by commas) with the first and last word randomly chosen from the set of fillers. The stimuli were recorded in a sound-treated studio by 10 native speakers of Czech (5 men, 18–26 years old) with no speech disorders. They did not receive any reward. Each of them had a unique order of blocks, but modal and whispered blocks always alternated. The speakers were instructed to read with normal effort and at moderate tempo, separating each word with a small pause. Stimuli produced in an incorrect phonation or with hesitation were repeated at the end of the session. The recordings were saved in uncompressed PCM format with a sampling frequency of 44.1 kHz and with 24-bit quantisation.

2.3. Acoustical analysis

All recordings were manually annotated in Praat [11] following the segmentation rules described in [12]. Whenever the nature of whispered speech prevented the use of the typical cues, the placement of boundaries was guided mainly by the changes of amplitude in the oscillogram. The analysis of the temporal data was performed in R using the packages *rPraat*, *tidyverse* and *boot* [13, 14, 15, 16]. In order to compare data from speakers with different articulation rates, the duration of each phone was multiplied by the mean articulation rate of the given speaker and divided by 10. The normalised values thus corresponded to the articulation rate of 10 phones/second. Only phones in the first two syllables were used for the calculation of the mean articulation rate, because the final syllables could have been affected by final lengthening and determining their precise boundaries was problematic. The duration of plosives was not measured in the initial position.

The effect of voicing on normalised duration was statistically evaluated with linear mixed effects models using the package *lme4* [17] in R. The fixed factors were (phonological) voicing, voicing pair, sex and position; random factors were speaker (with random slopes for voicing) and pseudoword. The significance of the factor of voicing and its interactions with other fixed factors was determined by *Likelihood Ratio Tests*.

2.4. Perception experiment

The perception experiment contained stimuli from 4 speakers (2 men). There were 28 whispered pseudowords (representing each target obstruent in both positions) selected from each of them. The stimuli were chosen to have duration values close to the mean values of the given target phoneme and speaker. Each pseudoword occured only once in the whole experiment. The participants (29 native speakers of Czech, 9 men, the interquartile range of age 22–53 years) were presented with a word in which the target phoneme was replaced with an underscore. They could replay the recording once. They chose the missing phoneme from a list that contained 19 consonants that occured in the stimuli. The target stimuli were complemented by 32 fillers (8 from each speaker), asking for non-target consonants. The stimuli followed in a random order. The participants practiced the operation of the experiment with 8 training stimuli. The perception experiment was performed online using Psytoolkit [18, 19] and it took 10–20 minutes. The participants could take two breaks and were instructed to wear headphones and to do the experiment in a silent setting. They did not receive any reward.

Although listeners could choose from a list of 19 consonants, the answers that incorrectly identified features other than voicing amounted to only 2% of all responses and were excluded from further analyses. An identification index was calculated for each listener and target obstruent (e.g. 0.75 for medial /s/, indicating that the given listener recognised 3 of 4 occurences of that obstruent). These indices served for the statistical analysis of the mean identification score of each group with the binomial test using the package *Hmisc* [20] in R. The result was considered significant if the error bar did not include the value 0.5 (chance level).

3. RESULTS

3.1. Duration of target obstruents

Table 2 provides the durational data of the target obstruents in modal phonation. The effect of voicing on normalised duration was statistically significant ($\alpha = 0.05$, $p < 0.001$), as were the interactions of voicing with the fixed factors of voicing pair ($p < 0.001$), position ($p < 0.001$) and sex ($p = 0.0031$).

voicing pair	position	difference (ms)	ratio
/p b/	medial	31.0	1.36
/t d/	medial	42.9	1.63
/c ɟ/	medial	34.9	1.40
/k g/	medial	34.4	1.42
/f v/	initial	53.8	1.55
	medial	54.5	1.72
/s z/	initial	38.4	1.33
	medial	58.3	1.62
/ʃ ʒ/	initial	37.1	1.30
	medial	50.7	1.52

Table 2: Differences and ratios of mean normalised duration of phonologically voiceless to voiced obstruents in modal phonation.

The same measurements were taken for whispered stimuli, as shown in Table 3. The effect of voicing on normalised duration was also statistically significant ($\alpha = 0.05$, $p < 0.001$), as were the interactions with the fixed factors of voicing pair, position and sex ($p < 0.001$ for all interactions). Higher values of voiceless counterparts were present among all voicing pairs, with mean differences in lower tens of milliseconds.

voicing pair	position	difference (ms)	ratio
/p b/	medial	23.6	1.25
/t d/	medial	37.9	1.50
/c ɟ/	medial	25.1	1.27
/k g/	medial	17.8	1.20
/f v/	initial	60.0	1.88
	medial	41.2	1.52
/s z/	initial	36.2	1.36
	medial	34.5	1.35
/ʃ ʒ/	initial	28.9	1.27
	medial	31.1	1.33

Table 3: Differences and ratios of mean normalised duration of phonologically voiceless to voiced obstruents in whispered phonation.

Since longer duration of voiceless obstruents was found in both phonation types, Figure 1 compares the mean voiceless to voiced ratios of normalised duration in both. In medial position, the durational contrast was significantly smaller in whisper for four voicing pairs and the same tendency was present in the other three pairs. On the other hand, modal initial fricatives had similar ratios as their whispered variants. The initial /f v/ showed great variability in duration (also enhanced by the difficulties in delimiting their boundaries), which reduces the reliability of the respective results.

3.2. Perception experiment

Listeners have successfully identified 78% of all target obstruents (success rates ranged between 61–87% for individual listeners). Figure 2 provides the mean identification scores for each target obstruent in both positions. There were apparent differences – while the pairs /t d/ and /f v/ were recognised in both positions, none of the palatal plosives reached statistical significance. The group of initial voiced obstruents tended to have the lowest scores.

4. DISCUSSION

The modal stimuli have mostly confirmed the durational differences reported by [10, 21], although the ratios for fricatives were 20–30 percentage points lower (but in the same order of voicing pairs). The results for whispered fricatives were also in line with [10]. Shorter duration of voiced obstruents in modal phonation is caused by their complex articulation, which includes conflicting aerodynamic conditions [22]. However, the same constraints do not apply for whisper and therefore the observed differences (although smaller than in modal phonation) have to be explained differently.

The perception experiment showed that listeners were quite successful in discriminating the voicing counterparts solely on the basis of phonetic cues. Interestingly, the recognisability of individual obstruents can be linked to the durational ratios of the respective voicing pairs – the pairs /t d/ and /f v/ had the highest ratios (above 1.5), while the bilabial and palatal plosives had the lowest ones (up to 1.3). The importance of duration would also explain another pattern in the perception data, namely the higher recognisability of obstruents in medial position, since it provides more obvious boundaries for the target phones. It is also noteworthy that the listeners did not choose voiceless obstruents more often than voiced ones. On the contrary, they balanced their responses despite the absence of voicing and the bias created by Czech phonological inventory (voiceless obstruents are more frequent than voiced ones, except for the pair /f v/ [23]).

However, the high identification scores of

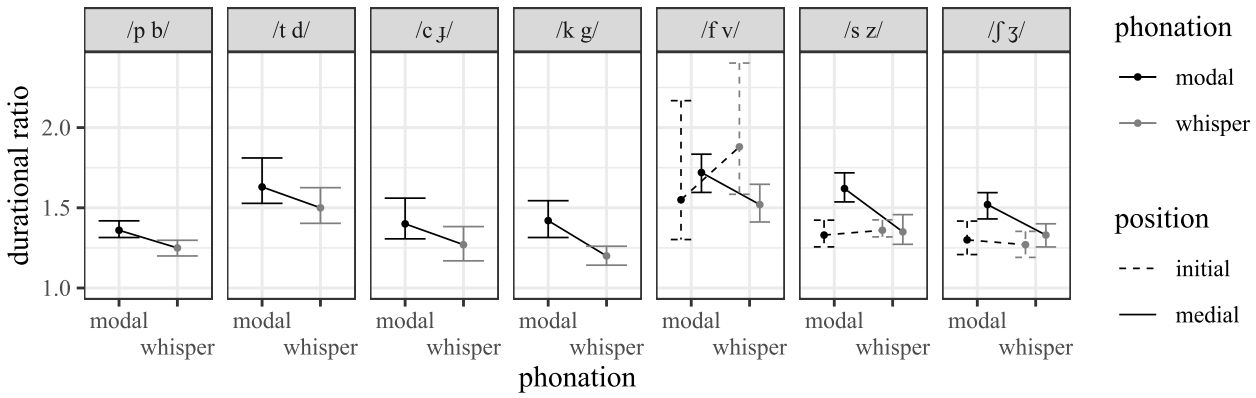


Figure 1: The ratios of the mean normalised durations of all target obstruents. The error bars indicate the 95% confidence intervals (with the Bonferroni correction for $n = 10$) calculated with the bootstrap method.

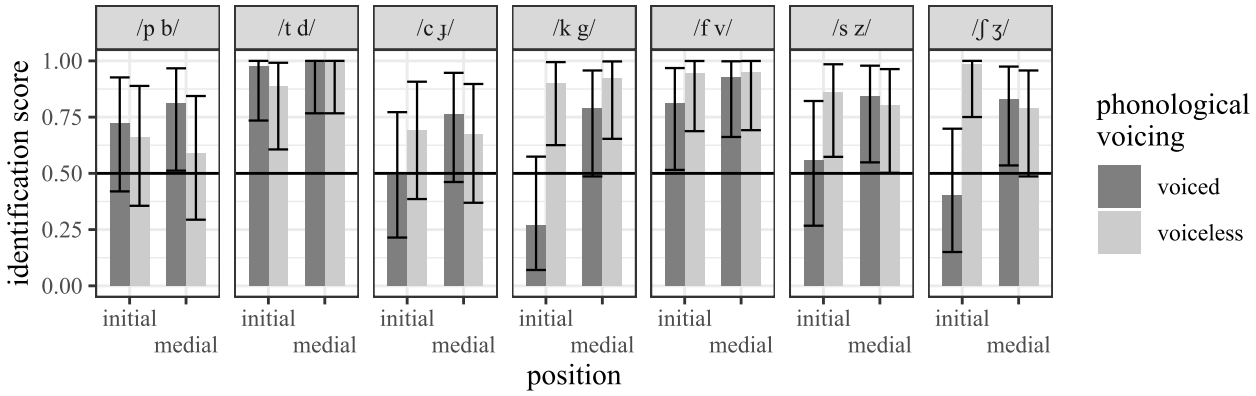


Figure 2: The mean identification scores for each target obstruent in both positions. The error bars indicate the 95% confidence intervals (with the Bonferroni correction for $n = 28$).

phonemes /t d/ and /f v/ are probably caused by multiple factors. Although they phonologically form pairs, their place and/or manner of articulation is considerably different. This is most prominent in the first pair, which consists of a dento-alveolar [t] that contrasts with a (post-)alveolar [d] (often realised as a tap [ɾ]) [24]. Similarly, the Czech /v/ acoustically resembles an approximant rather than a fricative, especially in intervocalic position [25]. Since the features of place and manner are relatively well preserved in whisper (as supported by [3] and the minor amount of other-than-voicing errors in the present perception experiment), they might have raised the identification scores of these two pairs. Moreover, both taps and approximants are associated with shorter durations, which further enhances the durational difference.

It seems that the phonological voicing contrast is partially preserved in whisper, especially in medial position. The voiced obstruents were found to be significantly shorter than their voiceless counterparts. This result supports the *redundant cue hypothesis* as described by van de Velde and van Heuven [8], because the durational ratios in whisper were less pronounced than in modal phonation and speakers were probably not trying to enhance the differences between voicing counterparts in order to compensate for the absent voicing.

Further research is needed to overcome the limitations of the present study, which focused on obstruents in isolated pseudowords with simple syllabic structure. Moreover, the material consisted of read speech and the results cannot be generalised to spontaneous interactions. The production and perception part suggested a connection between durational ratios and recognisability of voicing pairs in whisper, but only indirectly, since the stimuli were not controlled for other parameters. It would be desirable to perform perception experiments with manipulated stimuli differing only in duration.

5. REFERENCES

[1] J. Cirillo, "Communication by unvoiced speech: the role of whispering," *Annals of the Brazilian Academy of Sciences*, vol. 76, no. 2, pp. 413–423, 2004.

[2] G. L. Dannenbring, "Perceptual Discrimination of Whispered Phoneme Pairs," *Perceptual and Motor Skills*, vol. 51, no. 3, pp. 979–985, 1980.

[3] V. C. Tartter, "What's in a whisper?" *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[4] T. Mills, "Cues to voicing contrasts in whispered Scottish obstruents," Master's thesis, University of Edinburgh, 2003.

[5] M. F. Schwartz, "Bilabial Closure Durations for /p/, /b/, and /m/ in Voiced and Whispered Vowel Environments," *The Journal of the Acoustical Society of America*, vol. 51, pp. 2025–2029, 1972.

[6] M. Parnell, J. D. Amerman, and G. B. Wells, "Closure and constriction duration for alveolar consonants during voiced and whispered speaking conditions," *The Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 612–613, 1977.

[7] S. T. Jovičić and Z. Šarić, "Acoustic Analysis of Consonants in Whispered Speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, 2008.

[8] D. J. van de Velde and V. J. J. P. van Heuven, "Compensatory Strategies for Voicing of Initial and Medial Plosives and Fricatives in Whispered Speech in Dutch," in *Proceedings of the International Congress of Phonetic Sciences*, vol. 17, 2011, pp. 2058–2061.

[9] Š. Šimáčková, V. J. Podlipský, and K. Chládková, "Czech spoken in Bohemia and Moravia," *Journal of the International Phonetic Association*, vol. 42, no. 2, pp. 225–232, 2012.

[10] P. Machač and P. Šturm, "The phonological contrast of voicing in whispered Czech and its phonetic correlates – a preliminary study." in *20th Czech-German Workshop – Speech Processing*, R. Vích, Ed., 2010, pp. 34–43.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2020. [Online]. Available: https://www.praat.org/

[12] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Praha: Epocha, 2009.

[13] R Core Team, "R: A Language and Environment for Statistical Computing," 2019. [Online]. Available: https://www.R-project.org/

[14] T. Bořil and R. Skarnitzl, "Tools rPraat and mPraat," in *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer International Publishing, 2016, pp. 367–374.

[15] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the Tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, pp. 1686–1691, 2019.

[16] A. Canty and B. D. Ripley, *boot: Bootstrap R (S-Plus) Functions*, 2021, R package version 1.3-28.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[19] ——, "PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[20] F. E. Harrell, "Hmisc: Harrell Miscellaneous," 2021. [Online]. Available: https://CRAN.R-project.org/package=Hmisc

[21] P. Machač, "Temporální a spektrální struktura českých explozív," Rigorózní práce, Univerzita Karlova, Filozofická fakulta, Ústav obecné lingvistiky, Praha, 2006.

[22] J. J. Ohala, "Aerodynamics of phonology," in *Proc. 4th Seoul International Conference on Linguistics*, 1997, pp. 92–97.

[23] A. Bičan, "Phonological Corpus of Czech," 2016. [Online]. Available: https://ujc.avcr.cz/phword

[24] R. Skarnitzl, "Asymmetry in the Czech Alveolar Stops: An EPG Study," *AUC Philologica 1/2014, Phonetica Pragensia XIII*, vol. 2014, no. 1, pp. 101–112, 2014.

[25] R. Skarnitzl and J. Volín, "Czech Voiced Labiodental Continuant Discrimination from Basic Acoustic Data," in *Proceedings of Interspeech 2005*, 2005, pp. 2921–2924.

SPEECH EMOTION RECOGNITION FROM GLOTTAL FLOW SIGNALS USING 1D CONVOLUTIONAL NEURAL NETWORKS

Itay Ben-Dom, Catherine I. Watson, Clare M. McCann

The University of Auckland, New Zealand
iben350@aucklanduni.ac.nz, c.watson@auckland.ac.nz, c.mccann@auckland.ac.nz

ABSTRACT

Speech emotion recognition can enhance machine-human interaction. While traditional speech emotion recognition systems tend to use prosodic features for analysis, recent studies show glottal features can provide valuable insight into distinguishing different types of emotional expressions. This paper presents a preliminary investigation into the effectiveness of employing the glottal waveform for speech emotion recognition tasks. The JL corpus, a strictly-guided, simulated emotional speech corpus in New Zealand English is used for this study. The modelling and classification tasks are conducted using 1D convolutional neural network. The results show that glottal signal enhances emotion classification accuracy compared to raw audio signals, even with data augmentation. This suggests the glottal source can serve as better alternative input to speech emotion recognition models, particularly when limited training data is available.

Keywords: speech emotion recognition, raw audio, glottal flow, IAIF, convolutional neural networks.

1. INTRODUCTION

The human voice is a powerful conveyor of information. It can provide cues about the speaker's biological, psychological, and emotional state [1]. The design of automatic speech emotion recognition (SER) systems has important applications in human-machine communication [2]. It is desirable for machines to be able to interpret the emotional state of the user and respond to it in an appropriate manner. The development of SER systems is a challenging task, however, due to difficulty in finding proper representations for emotion embedding in speech.

The emotional state of a speaker has direct effect on voice production, leading to physiological changes in respiration, phonation, and articulation [3]. Acoustic descriptors of emotional voice have traditionally been provided in terms of prosodic (e.g. pitch, speech rate) or spectral (e.g. MFCC) characteristics for speech analysis [4]. Over the past decade, however, there has been a growing body of evidence showing glottal source contains emotional content and glottal features can provide valuable insight into distinguishing different types of emotional expressions [5].

The source-filter theory describes human voice production as a linear time-invariant system [6]. During the process of speech production, air flow from the lungs passes through the glottis and generates a quasi-periodic signal called the glottal volume velocity waveform. Separating the source (glottal waveform) and filter (vocal tract) components enables the modeling of their distinct contributions. The glottal waveform can be separated from the speech signal via source-filter deconvolution. One popular approach to achieve this is through iterative adaptive inverse filtering (IAIF) [7]. IAIF has been used and evaluated in various experiments and it has been shown to yield rather robust estimates of the glottal flow.

In recent years, speech emotion recognition has been revolutionised by neural network models, following their demonstrated success in computer vision and speech recognition tasks [8]. Neural networks are a powerful tool and been shown to outperform traditional SER approaches. Convolutional neural networks (CNNs) are commonly used typologies in this area. One-dimensional CNNs that learn acoustic models directly from audio waveforms are becoming a popular method in audio processing due to the ability of these networks to take advantage of the signal's fine time structure [9]. Recent works explored the use of time-series signals as input for audio and speech processing with 1D CNNs [10]. The low-level layer of the model was shown to be able to learn frequency-selective filters from raw audio inputs [11]. It has been shown that features learned directly from the audio waveform can match, or outperform, the performance accuracy of a model trained on hand-crafted features [10, 12] Abdoli *et al* [13] proposed an end-to-end 1D CNN for environmental sound classification from raw audio signals. Different input sizes were

evaluated and the proposed architecture was shown to outperform 2D CNN models.

This study presents presents a preliminary investigation into the effectiveness of incorporating the glottal excitation source for emotion recognition task from speech signal. We utilise the architecture proposed in [13] to compare the model's emotion classification effectiveness when using two different input types; raw audio and glottal flow.To the best of our knowledge, this is the first attempt to consolidate glottal analysis with convolutional neural networks.

2. METHODOLOGY

The glottal flow is estimated from raw audio waveforms using inverse filtering algorithm from the GVV Toolbox package in R [14]. The glottal waveform is computed pitch synchronously using the iterative adaptive inverse filtering algorithm [15]. It is a two-stage iteration process governed by the principles of linear predictive coding, where the glottal waveform is computed by subtracting the contributions of the vocal tract and radiation load from the speech signal for each analysis frame. While the signal obtained through inverse filtering may be only an approximate of the actual glottal waveform and potentially result in experimental bias, the IAIF algorithm was chosen because it is widely used in the literature [16]. Each waveform is downsampled to 16 kHz and split into multiple input frame lengths to be fed into the network architecture; 0.1s (1,600 samples), 0.5s (8,000 samples), 1s (16,000 samples), and 2s (32,000 samples). The frames are extracted with 50% overlap and zero-padded (if required).

The baseline CNN architecture [13] is constructed by stacking two local feature learning block (LFLB), two convolutional layers, and three fully connected layers. The LFLB is designed to extract emotional features from time-series data. Each LFLB consists of one convolutional layer, one batch normalisation (BN) [17] layer, one rectified linear activation unit (ReLU) [18] layer, and one max-pooling layer. The convolution layer learns by sliding filters across the entire spatial dimension of the input. In a CNN, the filters in early layers learn low-level features, while the filters of the deeper layers learn high-level features that resemble concepts. The result of the convolution is a feature map. The filter size, stride, and number of filters are provided in details in [13]. All convolutions and pooling operations are one-dimensional, i.e. only along the axis representing time. The LFLB block is followed by two convolutional layers, each with batch normalisation and ReLU layers. The output of the final convolutional layer is flattened and used as input to two stacked fully connected layers, each with ReLU activation. Due to data sparsity during training, overfitting may be encountered. A popular and easy to implement regularisation technique is dropout [19]. Dropout is used to mitigate the problem of overfitting in neural networks by preventing co-adaptation of features during training. Dropout at a rate of 0.5 is used between each dense layer. The final output layer of this architecture is a softmax classifier with 5 neurons (matching the number of classes). We note that the network depth is proportional to the input size, where the number of convolutional layer is adjusted for the input length, with deeper models used to process longer audio segments.

3. EXPERIMENT

The JL corpus is a New Zealand English simulated emotion speech corpus, developed for emotion classification and synthesis in human-robot interaction research [20]. It contains 5 primary emotions: *happy*, *angry*, *sad*, *neutral* and *excited*. The speech samples were collected from four speakers (two male and two female) who are professionally trained voice actors of New Zealand English. The speech material consists of 15 neutral sentences with equal distribution of English long vowels - /a:/, /o:/, /i:/, /u:/. During the recording session each speaker was asked to repeat the 15 sentences twice for every emotion. Two separate recording sessions were conducted on different days to account for speaker's possible psycho-physiological abnormality at the time of recording. The audio clips were originally sampled at a rate of 44.1 kHz, but for the purposes of this study were downsampled to 16 kHz. Overall, the JL corpus contains 4 (speakers) $\times$ 5 (primary emotions) $\times$ 2 (repetitions) $\times$ 2 (sessions) = 1200 primary emotion sentences.

The 1D CNN network was implemented in TensorFlow Python library v2.9.0. The models were trained on NVIDIA T4 graphical processing unit (GPU) with 16 GB memory. Each model was trained with a batch size of 16 for 1000 epochs with early stopping. An Adadelta optimiser with 0.001 learning rate and 0.95 decay rate was used [21]. Considering speaker independent recognition, leave-one-speaker-out (LOSO) cross-validation was carried out for the discrimination of emotions in the JL corpus. The use of LOSO cross-validation ensures that the models are not over-training to a

Table 1: Test accuracy (%) for different input frame length and input type, with and without data augmentation (DataAug).

Audio Length (s)	*Model Input*			
	Raw Audio		**Glottal Flow**	
		DataAug		*DataAug*
0.1	34.8	35.7	37.7	**39.8**
0.5	37.3	41.4	40.9	**44.7**
1	38.8	43.8	42.2	**45.8**
2	35.0	42.4	36.5	**43.0**

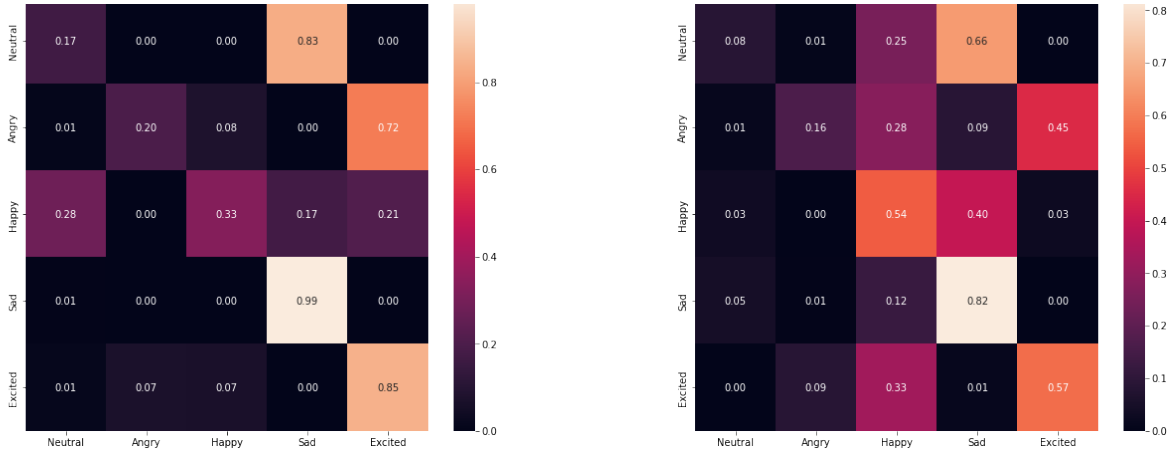
particular speaker. Since there are four speakers in the data set, the training set for each validation consists of three speaker, whereas the validation and test set were taken from an 80/20 split of the 4th speaker samples, respectively. Each validation was preformed 10-fold, with the final test accuracy calculated as the mean accuracy across all folds.

4. DISCUSSION

The models performances for each emotion category can be interpreted from the confusion matrix, shown in Figure 1. Figure 1a depicts the confusion matrix for best accuracy model with raw audio input. Figure 1b depicts the confusion matrix for best accuracy model with glottal waveform input. To study the model's performance for each emotion category, the confusion matrix was calculated by averaging the confusion matrix across ten cross-validation experiments for our best model architecture. For both input types, the high arousal emotions (e.g., angry, excited) are well separated from the low arousal ones (e.g., sad). Some of the most confused pairs are neutral-sad and angry-excited, which are different on the valence level. These erroneous classifications were extensively reported by other SER studies [22, 23]. The results show that classification from raw audio can discriminate between 2 emotions (sad, excited), while glottal waveform input showed discrimination between 3 emotions (happy, sad, and excited). These findings imply that a neural network trained on glottal waveforms is better equipped for discriminating between different emotions, as it has a greater capacity for emotion recognition.

The JL corpus presents speech samples for five primary emotions spoken in New Zealand English: *happy*, *angry*, *sad*, *neutral* and *excited*. Emotion classification performance was tested using both raw audio and glottal waveforms as inputs for an end-to-end 1D CNN models. The performance of the models was compared per input type and input length, and the results are presented in Table 1. The best performance was achieved using glottal input across all model topologies. It is apparent that models trained on glottal waveforms outperform models trained on raw audio. The 16,000-input (1 second) 1D CNN is preferable over other architectures, as it presents the best trade-off between the number of neural network parameters and prediction accuracy. This matches the findings of the original study [13].

The lack of large amounts of training data to train complex models is an ongoing challenge in speech emotion recognition. Insufficient training data can lead to overfitting and resulting in poor generalisation for unseen data. One common strategy to artificially expand the size of the training data set is data augmentation. Creating new samples through transformations of the existing data can help improve classification invariance and generalisation ability for neural network models. Growing the training set through data augmentation transformations have been shown to increased model accuracy in speech recognition tasks across all studies. General audio data augmentation techniques include tempo perturbation, loudness change, additive background noise, time-stretch, and pitch shift [24, 25]. One of the key requirements for data augmentation is to apply transformations to the inputs in a way that preserves their labels (ground truth does not change). In this study, three data augmentation techniques were applied on every utterance, doubling the size of the training set. Over-lapping windows were used to segment the input into various lengths. Since we are interested in continuous frame-based speech emotion recognition, 50% over-lap was applied between successive windows. This was done for both raw audio and glottal waveforms. A common technique for raw audio data augmentation is additive noise. For raw audio waveforms, additional transformation was carried out through distortion of the speech signal by adding noise from Laplacian distribution [26]. We propose a new data augmentation technique for glottal signals in the form of time-shift. We can shift the glottal signal left or right to create additional training data. We used a time-shift degree ranging from -5 to +5 ms. To the best of our knowledge this is the first attempt at applying data augmentation to glottal waveforms. The results show data augmentation boosted the performance accuracy by 3-7% for both raw audio and glottal waveforms, as shown in Table 1. These



(a) **(b)**

Figure 1: Normalised confusion matrix for best accuracy model with raw audio (a) and glottal waveform (b) as input sources.

findings support the current trend in literature and suggest data augmentation is a powerful tool for the training on models on smaller data sets.

Even with the application of data augmentation techniques, CNN models require substantial amount of training data to yield tangible results. The JL speech corpus used in this study has too few samples for the network to adequately capture the variety of emotions found in practice (e.g. *cold anger* vs *hot anger*). This is reflected in the emotion classification error rates observed in Figure 1. If the characteristics of the speech samples used for testing depart too much from those found in the training set, accuracy is expected to be low. When the training set is truly comprehensive, this problem is diminished. The small sample size of speakers (four) offers little variations for the CNN models to learn general characteristics and features. We noticed that the test accuracy for one of the male speakers was consistently lower than that of the other three speakers, which led to lower performance accuracy overall. In addition, since the corpus contains multiple repetitions of the same sentence (2 sessions with 2 repetitions), the corpus can be practically viewed as having only 600 unique samples, where the other 600 samples offer slight variation and therefore can be viewed as a form of data augmentation in itself. Thus, while the networks trained in the context of this work are not ready for practical use, the results they yielded provided a wealth of information that can be explored in future developments. Other investigations in literature often don't address this issue; they don't consider it significant or they use widely-used data sets. It is recommended that all results be interpreted in light of the limitations associated to the respective training data sets.

5. CONCLUSIONS AND FUTURE WORK

This paper demonstrates the effectiveness of using glottal waveforms as an alternative input for speech emotion recognition models. An end-to-end 1D CNN architecture was trained on both raw audio and glottal flow waveforms in order to discriminate between five primary emotions. The results show that neural network models trained on glottal inputs outperform the classification accuracy of those trained on raw audio. This reinforces the prevailing idea in the literature that the emotional content of speech signals is conveyed through the glottal pulse. Additionally, the efficacy of time-shift as a data augmentation approach for glottal signals has been demonstrated. Moreover, the influence of input dimensionality on classification performance underscores the significance of treating input dimensionality as a hyper-parameter in network architecture design. In conclusion, utilising the glottal flow as the input type for 1D CNNs shows potential for replacing raw audio, particularly in studies with limited data constraints. Building upon these findings, we intend to compare the performance accuracy of 1D and 2D CNN models for speech emotion recognition, using raw and spectrogram representations of the glottal flow, respectively. This will form the basis for our future research into speech emotion recognition in disordered speech.

6. REFERENCES

[1] A. Kappas, U. Hess, and K. R. Scherer, "Voice and emotion," in *Fundamentals of nonverbal behavior*, R. S. Feldman and B. Rime, Eds. Cambridge University Press, 1991, ch. 6, pp. 200–238.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] T. Johnstone, "The effect of emotion on voice production and speech acoustics," Ph.D. dissertation, The University of Western Australia, 2017, doi.org/10.31237/osf.io/qd6hz.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[5] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng, "Study of wavelet packet energy entropy for emotion classification in speech and glottal signals," in *Fifth International Conference on Digital Image Processing (ICDIP 2013)*, Y. Wang and X. Yi, Eds., vol. 8878, International Society for Optics and Photonics. SPIE, 2013, p. 887834.

[6] G. Fant, *Acoustic theory of speech production*. Mouton, 1960.

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992, eurospeech '91.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[9] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4624–4628.

[10] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," 2016.

[11] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6964–6968.

[12] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," 2018.

[13] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[14] I. Ben-Dom, "Towards a voice analysis toolbox for the extraction, parametrisation and analysis of the glottal source waveform and its application for senescence voice," Master's thesis, The University of Auckland, 2017, http://hdl.handle.net/2292/35845.

[15] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992, eurospeech '91.

[16] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *INTERSPEECH*, 2016.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[18] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. DudÃk, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[20] J. James, L. Tian, and C. I. Watson, "An open source emotional speech corpus for human robot interaction applications," in *INTERSPEECH*, 2018.

[21] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012.

[22] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] L. Tian and C. Watson, "Emotion recognition using intrasegmental features of continuous speech," in *Proceedings of the 17th Australasian International Conference on Speech Science and Technology (SST2018)*, 2018, pp. 61–64.

[24] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[25] E. Lakomkin, M.-A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 854–860, 2018.

[26] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

Phonetics of glottalized phonations in Yateé Zapotec

Yuan Chai[1], Adrián Fernández, Briseida Mendez

[1]Department of Linguistics, UCLA, Los Angeles CA USA 90095
yuanchai@g.ucla.edu; Dyadee03@gmail.com; briseida.mendezfer@gmail.com

ABSTRACT

This study presents a language, Yateé Zapotec, that contrasts modal, rearticulated (V^{ʔ}V), and checked (V^{ʔ}) phonations independently of tone. Rearticulated and checked phonations both involve glottalization but differ in the phasing of glottalization in vowels. We describe the three phonation types in Yateé Zapotec in terms of their f0, voice quality, amplitude, and duration based on the production of 147 words by a native speaker, and ask what acoustic parameter distinguishes them most effectively. Using multidimensional scaling analysis, we see that rearticulated vowels are best differentiated from modal and checked vowels by amplitude dipping. Checked vowels are best differentiated from modal and rearticulated vowels by their short duration. The results are in accordance with the findings of previous perception studies, such that listeners rely on amplitude dip to perceive vowels interrupted by glottal constriction, whereas they use short duration to identify a checked vowel.

Keywords: checked, rearticulated, phasing of glottalization, amplitude of voicing, duration

1. INTRODUCTION

Yateé Zapotec is a variety of Zapotec spoken in San Francisco Yateé, Oaxaca, Mexico, as well as by diaspora communities around Los Angeles, USA. There were around 480 people in the village of Yateé in 2017, according to a census conducted by the local clinic. Existing documentation of Yateé Zapotec is sparse [1, 2], and both studies focus on the consonants in the language. The current study focuses on the vowel phonation types in Yateé Zapotec and provides a quantitative analysis of their acoustic properties.

There are three vowel phonation types in Yateé Zapotec: modal (/V/), rearticulated (/V^{ʔ}V/), and checked (/V^{ʔ}/). Rearticulated phonation and checked phonation both involve glottalization but differ in the phasing of glottalization. Rearticulated phonation can be realized with either glottalization in the middle of vowels, or with glottalization spanning the whole duration of vowels. Checked phonation, in contrast, has glottalization at the end of vowels. The relative phasing of glottalization and vowel is in Figure 1. Examples of each phasing type in Yateé Zapotec are in Figure 2.

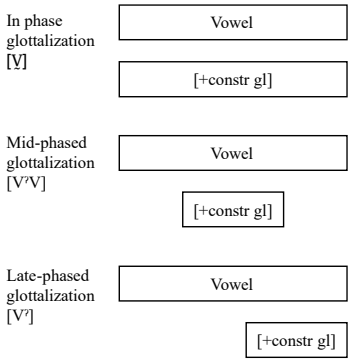


Figure 1: The phasing of in-phase, mid-phased, and late-phased glottalization relative to vowels.

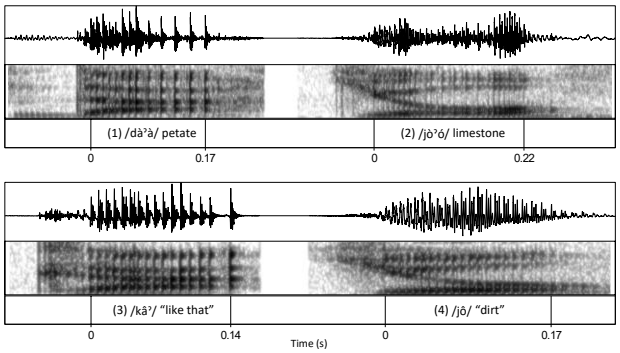


Figure 2: Sample tokens for in phase (1), mid-phased (2), late-phased (3) glottalization, and modal (4) phonation.

Yateé Zapotec has four tones: low, high, rising, and falling. The four tones are fully crossed with the three phonations. Examples of the interaction between phonation and tone are in Table 1.

	Modal	Rearticulated	Checked
Low	/ʂnà/ <xna> "red"	/ʂn:àʔà/ <xna'a> "my mother"	/nàʔ/ <na'> "there"
High	/l:á/ <la> "hot"	/zíʔí/ <zi'i> "heavy"	/ʐáʔ/ <xha'> "clothes"
Rising	/gă/ <ga> "nine"	/wèʔé/ <we'e> "get drunk"	/wĕʔ/ <we'> "wound"
Falling	/jô/ <yo> "dirt"	/jóʔò/ <yo'o> "house"	/kâʔ/ <ka'> "like that"

Table 1: Examples of phonation and tone in Yateé Zapotec. Orthography is in angle brackets.

The research question of the current study is: how do the three phonations differ acoustically? We test the acoustics of the three phonations in five dimensions: f0, spectral tilt, noise, amplitude of voicing, and duration. The first four parameters are related to glottalization: glottalized vowels can have an extra-low f0 or high f0, lower spectral tilt, noisy voicing, and decreased voicing amplitude [3]. Checked vowels have been found to be shorter than modal and rearticulated vowels in several varieties of Zapotec (Yalálag: [4]; Betaza: [5]; Quiaviní: [6]). Further, we ask, among those acoustic parameters, which parameter(s) distinguish(es) the three phonations in Yateé Zapotec most effectively? To answer these questions, we conducted a production experiment and performed quantitative analysis.

2. METHOD

2.1. Participant

The participant is a male speaker of Yateé Zapotec at the age of 37. Yateé Zapotec is his native language acquired at home. He acquired Spanish at the age of 10 and acquired English at the age of 21. He uses Zapotec, Spanish, and English at home and uses English and Spanish at work.

2.2. Stimuli and procedure

The stimuli are 49 Yateé Zapotec words and were elicited in isolation. The distribution of the words in terms of tone and phonation is in Table 2. Table 1 is part of the word list. The full word list and the R script for data analysis are available at https://osf.io/a98sf/?view_only=dc40c474d5064ddaa649892131d41394.

	Modal	Rearticulated	Checked
Low	4	6	4
High	5	5	7
Rising	7	3	2
Falling	5	1	3

Table 2: Distribution of the stimuli by tone and phonation.

During the experiment, each target word was elicited by a picture, its Spanish translation, and its Zapotec orthography. The stimuli were presented on a webpage three times in random order. The participant produced 147 words in total (49 words * 3 repetitions).

2.3. Acoustic measures and data cleaning

The recordings of the target words were processed using VoiceSauce [7]. We drew the f0 track of each token and visually inspected the f0 track. If we saw an f0 track that has an octave jump, we regarded it as an f0 tracking error and excluded it from the analysis of f0 and spectral tilt. We also drew the formant track of each vowel. If a token has a Mahalanobis distance larger than 6 from the mean F1 and F2 of the vowel, we regarded it as a formant tracking error and excluded the token from the analysis of spectral tilt [8]. After manually excluding the f0, F1, and F2 tracking errors, we converted parameter values to z-scores, then treated z-scores larger than 3 as outliers and excluded those tokens from the analysis.

3. RESULTS

In this section, we describe the results of six acoustic parameters for each phonation type: f0, H1*–H2* and residual H1* (spectral tilt), Harmonic-to-noise ratio (noise), Strength of Excitation (amplitude of voicing), and duration. Then we pass all these parameters to a multi-dimensional scaling analysis to illustrate what parameters best distinguish the three phonations.

3.1. F0

We fit f0 contours using a generalized additive model: *gam(f0~phonation*tone+s(time,k=9,by=interaction(phonation,tone)))*. *Phonation*, *tone*, and their interaction are the independent factors. *Time* is the smooth term (degrees of freedom = 9), interacting with the interaction of *phonation* and *tone* to allow an independent time smooth spline for each phonation and tone. The predicted f0 contour of each phonation and tone is in Figure 3. Low tone and falling tone both have a falling contour. Falling tone has a larger falling slope than low tone. High tone and rising tone both have a rising contour. Rising tone has a larger rising slope than high tone. Rearticulated vowels are characterized by a dipping point in the middle of vowels for the four tones, reflecting middle-phased glottalization in rearticulated vowels. High-toned checked and rearticulated vowels have higher F0 than modal vowels, indicating that they are likely to be characterized by tense voice.

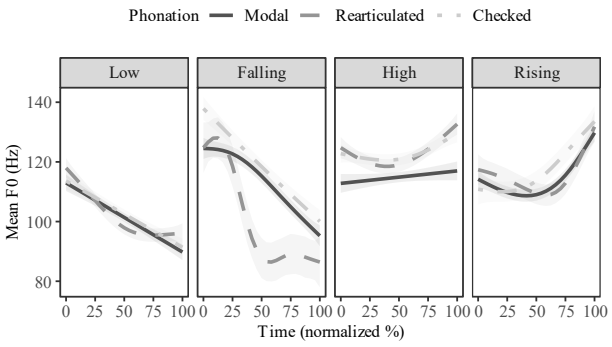


Figure 3: F0 of each tone and phonation.

3.2. Spectral tilt: H1*–H2* and residual H1*

Spectral tilt, such as H1*–H2*, is usually used to represent the degree of glottal constriction. Lower spectral tilt indicates a higher degree of glottal constriction [9, 10]. Residual H1* is a measure proposed in [11], which controls the amplitude of H1 by normalizing the energy of the signal. We fit H1*–H2* contours and residual H1* contours using a generalized additive model with *phonation* as the independent variable and *time* as the smooth term (degrees of freedom = 9), interacting with *phonation*. The same gam model was used for all the following time contours in this paper. Residual H1* (Figure 4b) differentiates the three phonations better than H1*–H2* (Figure 4a). The differences in residual H1* between every two phonations are plotted in Figure 4c. The blue portion of lines indicates that the difference in residual H1* is significant during that period, whereas the red portion means that the difference is not significant. As shown in Figures 4b and 4c, modal vowels have higher residual H1* than the other two phonation types almost all the time, indicating that modal vowels have the least glottal constriction among the three phonations. Rearticulated vowels have the lowest residual H1* in the middle of the vowel, whereas checked vowels have the lowest residual H1* at the end of the vowels. This verifies our prediction that rearticulated vowels have the most glottal constriction in the middle of vowels whereas checked vowels have the most glottal constriction at the end of the vowels.

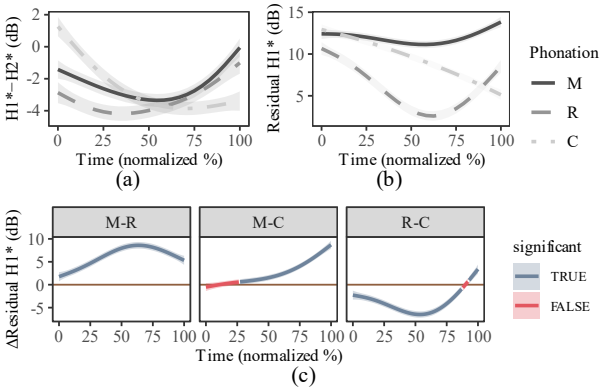


Figure 4: H1*–H2* (a) and residual H1* (b) contours and residual H1* differences (ΔResidual H1*) between phonations (c). M: modal; R: rearticulated; C: checked.

3.3. Noise level: Harmonic-to-noise ratio (HNR)

HNR measures the level of noise in the signal. A lower HNR value represents a higher level of relative spectral noise [12]. As shown in Figure 5, rearticulated vowels have lower HNR values than modal vowels after the first fourth of vowels. Checked vowels have lower HNR than modal vowels in the latter half. This indicates that rearticulated vowels have a noisy quality after the beginning of the vowel. Checked vowels are noisier at the end of the vowel. Modal vowels are the least noisy among the three phonation types.

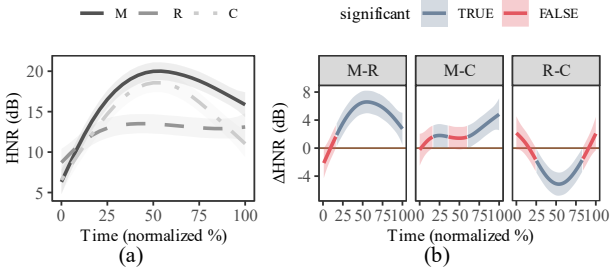


Figure 5: HNR contour (a) and HNR differences (b).

3.4. Amplitude of voicing

Strength of Excitation (SoE) measures the amplitude of voicing [7]. A larger SoE means that the voicing has a larger amplitude. As shown in Figure 8, modal vowels have a near-flat SoE contour. Among the three phonation types, checked vowels have the highest SoE at the beginning of the vowel and the lowest SoE at the end of the vowel. Rearticulated vowels exhibit a dipping SoE contour. The middle point has the lowest SoE. The amplitude of voicing is another parameter that reflects glottalization. Glottalization usually elicits damping in the amplitude of the voicing [3, 13]. The SoE contours in Figure 6 reflect that rearticulated vowels are characterized by amplitude damping in the middle of vowels, whereas checked vowels have amplitude damping at the end of vowels. The amplitude damping corresponds to the middle-phased and late-phased glottalization in rearticulated and checked vowels, respectively.

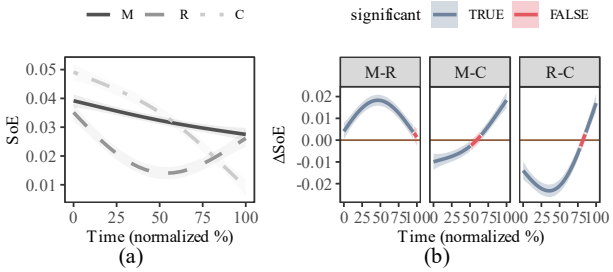


Figure 6: SoE contours (a) and SoE differences (b).

3.5. Duration

Among the three phonation types, checked vowels have the shortest duration (mean = 177 ms), followed by modal vowels (mean = 210 ms) and rearticulated vowels (mean = 216 ms) (Figure 7). A linear regression test shows that checked vowels have a significantly shorter duration than modal and rearticulated vowels ($p < 0.001$). The duration of modal and rearticulated vowels do not differ significantly ($p = 0.5$).

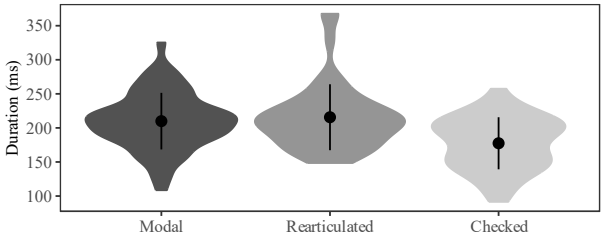


Figure 7: Duration of each phonation type.

3.6. Dimensionality reduction

Among the various parameters described above, which parameter(s) can most effectively distinguish the three phonation types? We performed a non-metric multi-dimensional scaling (MDS) analysis on the data. MDS can reduce the dimensions of the data, then project the data points to the reduced dimensions, and measure the distance among the data points [14]. MDS analysis has been used to analyze the acoustic cues associated with perceptual similarity [15, 16, 17] and the acoustic cues that differentiate phonation types across languages [18].

For each acoustic parameter described in Section 3 (except for duration), we divided the vowel into three equal-timed intervals and measured the mean value of each interval. We measured the Euclidean distances among the data points on a 3-dimensional MDS solution. The stress of the MDS model is 0.115. We plotted the data points on Dimension 1 (MDS1)-MDS2 and MDS1-MDS3 planes (Figure 8). MDS1 separates rearticulated vowels from modal and checked vowels. MDS2 separates checked vowels from modal and rearticulated vowels (Figure 8a). MDS3 separates modal vowels from checked and rearticulated vowels (Figure 8b). The top 3 parameters that have the biggest weight for each dimension are listed in Table 3. MDS1 is represented by the amplitude of voicing (SoE). MDS2 is primarily represented by duration, followed by the amplitude of voicing. MDS3 is represented by the final-third f0 and the last two-thirds of residual H1*.

	Top 1	Top 2	Top 3
MDS1	SoE_3 (5.132)	SoE_2 (4.852)	SoE_1 (2.577)
MDS2	Duration (2.326)	SoE_1 (2.181)	SoE_2 (2.065)
MDS3	F0_3 (2.899)	Residual H1*_2 (1.953)	Residual H1*_3 (1.486)

Table 3: Acoustic parameters with the highest weight on each MDS dimensions. Weight is in parenthesis. Numbers 1, 2, and 3 followed by the parameter name represent the initial, medial, and final third of the vowel, respectively.

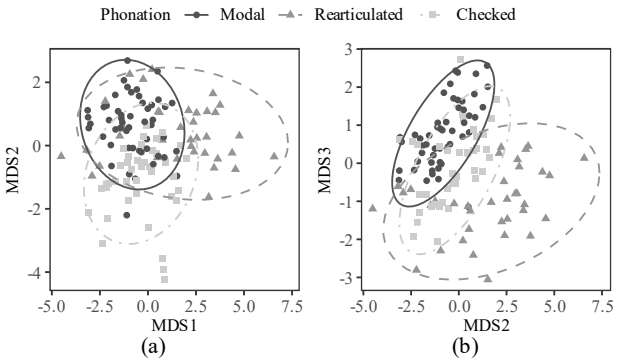


Figure 8: MDS distribution (confidence interval of the ellipses is 95%)

4. DISCUSSION AND CONCLUSION

The descriptive data of the acoustic parameters and the MDS analysis show us that rearticulated vowels are differentiated from modal and checked vowels primarily by their amplitude of voicing. The convex-shaped SoE contour of rearticulated vowels is distinguished from the falling contour of checked vowels and the flat contour of modal vowels. Checked vowels are distinguished from modal vowels primarily by their shorter duration. In addition, modal vowels are further distinguished from checked and rearticulated vowels by having less glottal constriction in the last two-thirds of vowels. Although f0 also emerges as a parameter in MDS3, we do not regard it as a defining feature of phonation because the tonal distribution in the stimuli is unevenly distributed among phonation types. There are more high and rising tones in modal vowels than in rearticulated and checked vowels in the stimuli.

The results agree with some previous findings on the perceptual cues of rearticulated and checked vowels. For rearticulated vowels, listeners of English [19] and Coatzospan Mixtec [20] use amplitude and f0 dips to identify VʔV or V̰. For checked vowels, short duration and falling f0 contour are more important cues than late-phased glottalization (White Hmon [21], Sgaw Karen [22], Xiapu Min [23]). The cues of amplitude and duration that listeners use to perceive rearticulated vowels and checked vowels are also the important cues for distinguishing rearticulated and checked vowels from modal vowels in the production of Yateé Zapotec. F0 not emerging as the primary cue in Yateé Zapotec is probably due to that phonation is fully crossed with all four tones in the language. A constant f0 is required to maintain the tonal contrast. In future studies, we will conduct perception tests with resynthesized stimuli and determine whether listeners of Yateé Zapotec attend to amplitude in voicing, duration, and spectral tilt when distinguishing checked, rearticulated, and modal vowels.

5. REFERENCES

[1] Jaeger, J. J. 1983. The fortis/lenis question: evidence from Zapotec and Jawoñ. *Journal of Phonetics*, 11, 177–189.

[2] Jaeger, J. J., Van Valin, R. D. 1982. Initial Consonant Clusters in Yateé Zapotec. *International Journal of American Linguistics*, 48, 125–138.

[3] Keating P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. in *Proc. 18th ICPhS* Glasgow.

[4] Avelino, H. 2004. Topics in Yalálag Zapotec, with particular reference to its phonetic structures. Dissertation, University of California, Los Angeles, Los Angeles, CA.

[5] Teodocio Olivares, A. 2009. Betaza Zapotec Phonology: Segmental and Suprasegmental Features. Master's thesis, University of Texas at Austin, Austin, TX.

[6] Chávez-Peón, M. E. 2010. "The interaction of metrical structure, tone, and phonation types in Quiaviní Zapotec,"Dissertation, University of British Columbia, Vancouver, Canada.

[7] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. M., 2011. VoiceSauce: A program for voice analysis. in *Proc. 17th ICPhS* Hong Kong, 1846–1849.

[8] Chai, Y., Ye, S. 2022. Checked Syllables, checked tones, and tone sandhi in Xiapu Min. *Languages*, 7, 47.

[9] Klatt, D. H., Klatt, L. C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am*. 87, 820–857.

[10] Zhang, Z. 2016. Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *J. Acoust. Soc. Am*. 139, 1493–1507.

[11] Chai, Y., Garellek, M. 2022. On H1–H2 as an acoustic measure of linguistic phonation type. *J. Acoust. Soc. Am*. 152, 1856–1870.

[12] Garellek, M. 2019. The phonetics of voice. in *Routledge Handbook of Phonetics*, Katz, W., Assmann, P. (eds), Oxford: Routledge, 75–106.

[13] Pan, H.-h. 2017. Glottalization of Taiwan Min checked tones. *Journal of the International Phonetic Association*, 47, 37–63.

[14] Cox, T. F., Cox, M. A. A. 2001. *Multidimensional scaling*, Boca Raton: Chapman & Hall/CRC.

[15] Bradlow, A., Clopper, C., Smiljanic, R., Walter, M. A. 2010. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups," Speech Communication, 52, 930–942.

[16] Fox, R. A., Flege, J. E., Munro, M. J. 1995. The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *J. Acoust. Soc. Am*. 97, 2540–2551.

[17] Gandour, J. T., Harshman, R. A. 1978. Crosslanguage differences in tone perception: a multidimensional scaling investigation. *Language and Speech*, 21, pp. 1–33.

[18] Keating, P., Kuang, J., Garellek, M., Esposito, C., Khan, S. Submitted. A cross-language acoustic space for vocalic phonation distinctions. *Language*.

[19] Hillenbrand, J. M., Houde, R. A. 1996. Role of F0 and amplitude in the perception of intervocalic glottal stops. *Journal of Speech, Language, and Hearing Research*, 39, 1182–1190.

[20] Gerfen, C., Baker, K. 2005. The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, 33, 311–334.

[21] Garellek, M., Keating, P., Esposito, C. M., Kreiman, J. Voice quality and tone identification in White Hmong," *J. Acoust. Soc. Am*., 133, 1078–1089.

[22] Brunelle, M., Finkeldey, J. 2011. Tone perception in Sgaw Karen. in *Proc. 17th ICPhS* Hong Kong.

[23] Chai, Y. 2022. *Phonetics and phonology of checked phonation, syllables, and tones*. PhD diss., University of California San Diego, San Diego, CA.

LABELS FOR VOICES

Jody Kreiman

Departments of Head and Neck Surgery and Linguistics, UCLA
jkreiman@ucla.edu

ABSTRACT

The problem of characterizing voice quality has long caused debate and frustration. The richness of the available descriptive vocabulary is overwhelming, but the density and complexity of the information voices convey leads some to conclude that language can never adequately specify what we hear. Others argue that terminology lacks an empirical basis, so that language-based scales are inadequate *a priori*. Efforts to provide meaningful acoustic characterizations have also had limited success: acoustics may capture sound patterns, but cannot at present explain what characteristics, intentions, or identity listeners attribute to the speaker based on those patterns. However, some terms continually reappear across studies. These terms align with acoustic dimensions accounting for voice variance across speakers and languages, and correlate with size and arousal across species. This suggests that labels for quality rest on a bedrock of biology: We have evolved to perceive voices in terms of size/arousal, and these factors structure both voice acoustics and descriptive language. Such linkages could help integrate studies of signals and their meaning, producing a truly interdisciplinary approach to voice.

Keywords: voice quality, terminology

1. INTRODUCTION

The problem of how to characterize voice quality is an endless source of debate and frustration across disciplines. The richness of the vocabulary available to describe voice is overwhelming, but a shared vocabulary (for example, for clinical or pedagogical use) has not been validated, despite many efforts [e.g., 1, 2]. Although there is some agreement that a few specific terms (especially breathiness and roughness) are important dimensions of quality, listeners do not agree about the extent to which a given voice demonstrates these attributes, or even whether or not they are present at all [3]. Further, the relationship between such terms and the information listeners glean from voice is not clear, so that description does not precisely predict what listeners hear or the way in which voices convey that information. This weak linkage has led some scholars to conclude that language can never adequately specify what we hear [4, 5]. Others have argued that terminology derives from whimsy, analogy, metaphor, and historical tradition and lacks an empirical basis, so that language-based scales are theoretically inappropriate measurement tools *a priori* [6]. However, efforts to provide meaningful acoustic characterizations of the messages conveyed by voice have also had limited success (see [6] for review), and it remains difficult to predict even basic attributes of the speaker (for example, sex, age, emotional state, identity) from instrumental measures. Acoustic parameters may precisely quantify physical sound patterns [7], but beyond that they cannot explain the meaning a listener attributes to a signal, because such meaning derives not only from the sound, but also from the emotional, situational, and cultural context in which the utterance takes place and from the listener's attitudes and background, among many other factors. Listener performance on relatively simple tasks like determining whether two voice samples represent the same or two different speakers (or even the same or two different tokens) cannot be explained by the properties of the signal without reference to higher-level cognitive variables reflecting the listener's contribution to the perceptual process [8, 9].

Thus, a gap emerges between physical sounds and perceived quality in our models of voice, the so-called "timbral abyss" [10]. Acoustic studies quantify signals and link them causally to the bodies that produced them, but do not predict or provide access to any broader meaning conveyed, while studies of the meanings carried by voice are not presently able to explain how such meaning emerges from discrete acoustic signals. How, and to what extent, can we combine these different facets of quality in a single multidisciplinary theoretical framework?

In this paper, I describe a possible partial solution to this long-standing issue in the study of voice quality. I begin by briefly reviewing the literature on descriptive terms for voice from different disciplines. I will argue that a small number of dimensions consistently emerge from these studies (along with a multitude of other terms that vary widely from study to study), and that not only lay individuals, but scholars and clinicians persist in using these terms despite the fact that they cannot be applied reliably to describe or quantify signals. Next, I review work

showing that acoustic variability within and across virtually all speakers can be characterized by a few dimensions that are shared by everyone, regardless of the speaker's gender, language spoken, or the kind of speech sample produced, again accompanied by a large number of other parameters whose relevance depends on the idiosyncrasies of the particular voice in question. Finally, I argue that frequently-emerging descriptive terms align well with these shared acoustic dimensions, which are associated with physical size and arousal across many species and are thus biologically significant. I conclude that the most commonly applied terms for voice remain useful—and continue to be used—because our use of language is partially structured by biology. That is, we perceive voice in terms of these factors, and our terminology reflects this structure without conscious design, because these aspects of the meaning of a voice signal are part of our evolutionary heritage. This provides a link between qualitative and quantitative approaches to measuring voice, forming a common foundation for both kinds of study and a potential basis for truly interdisciplinary approaches to voice studies.

2. THE "DUAL NATURE" OF VOICE QUALITY

Critical theorists (e.g., [11]) propose that the timbral abyss arises from the apparent dual nature of voice quality. From an empirical perspective, quality ultimately depends on speech production (vocal fold movements, vocal tract configurations, and so on (e.g., [12]). This system creates the acoustic voice signal, which can be quantified and serves as input to the voice processing (perceptual) mechanism. From a humanistic perspective, however, voice quality resides in the listener, and not in the speaker. Rather like a tree falling in the woods, which generates meaningless vibrations in the absence of a hearer, a speaker produces something to hear, but what is heard (quality) depends not just on the physical signals, but also on the listener's affect and memory, the conversational setting, cultural structures, and a multitude of other factors [6, 12, 13]. Descriptive terminology remains the most common approach to capturing these aspects of perception, which cannot be assessed by acoustic measures alone because they are not solely functions of the voice signal. Such terminology is relied upon in clinical settings, in humanistic research, and in common discourse.

Thus a gap emerges between acoustic and qualitative approaches to quality. Acoustic studies can provide measures that can reasonably be applied to other voices in other studies, and can precisely quantify quality in the ANSI sense [14] as what makes one voice sound the same or different from another [7]. Qualitative studies can provide very detailed descriptions of specific voices in specific contexts, but results typically do not generalize well to other contexts or other voices [12]. In this sense, qualitative descriptions do not actually measure, or even specify, voice quality in any useful way. This is inevitable given the density and complexity of the kinds of meaning conveyed by voice, but greatly limits the use of qualitative approaches for uncovering general truths about quality or voice perception in general. Acoustic measures also cannot fully characterize quality, because they cannot tap into listeners' contributions to what they hear. Thus, no one kind of analysis appears to be adequate on its own to assess quality, and there is no obvious way to reconcile the different approaches, so that the timbral abyss appears uncrossable.

3. QUALITATIVE STUDIES

Qualitative studies of voice have a long history (see [6] for review), and are supplemented by extensive studies of musical timbre which ask similar questions and use similar methods (see [15] for review). A variety of techniques have been applied to explore the semantic dimensions of quality. Typically authors ask listeners to rate voices (or other sounds) on sets of semantic-differential scales [16] and then apply factor analysis to reveal the structure underlying the ratings [e.g., 16, 17, 18]. For example, Bele [2] asked listeners to rate normal voices on 15 visual analog scales. Four underlying dimensions emerged from factor analyses of these data: sonority, irregularity, noise, and phonatory effort. Similar dimensions (severity of pathology, roughness, breathiness, weakness, and strain) emerged from a study of hoarse voices [17]; and the popular CAPE-V protocol for clinical voice assessment includes scales for severity of pathology, roughness, breathiness, and strain [1]. Across studies, despite differences in the voices, listeners, and scales under examination (see [6] for review), a small set of dimensions consistently emerges across papers. These dimensions include something analogous to brightness/brilliance/sharpness/clarity, which is associated with the distribution of spectral energy in the voice; breathiness and/or roughness, associated with noise or spectral irregularity; and fullness/richness, associated with the location of the spectral centroid [e.g., 18-21]. Interestingly, similar sets of scales have emerged across cultures and languages [22, 23], and from studies of instrumental timbre and animal vocalization.

4. ACOUSTIC VARIATIONS WITHIN AND AMONG SPEAKERS

The question thus arises: Why these terms, and not others? Recent studies of within- and between-speaker acoustic variability suggest a possible answer. Analyses of the speech of large groups of female and male speakers of English, Seoul Korean, Hmong, and Thai [24-26] show remarkably consistent patterns of acoustic variability that are seemingly shared by all speakers, regardless of sex, age, or native language. These factors can be thought of as a low-dimensional "voice space" that represents the ways in which voices differ from each other acoustically. The first dimension of this space always reflects variations in the balance of harmonic and inharmonic energy in the voice source. This combination of parameters is often associated with a quality continuum from "strained" or "pressed" (or "bright") to "breathy" [27, 28], which signals arousal across many species ([27, 29]). The second dimension is associated with formant dispersion, which is sexually dimorphic and varies with the size of the vocal tract. Formant dispersion serves to signal both dominance and reproductive fitness across many species [30]. A second set of dimensions describes variability that is shared by speakers with a common native language, but not by speakers of other languages. These dimensions reflect the phonological structure of the language being spoken [25]: For example, a factor representing H1-H2 (the difference in the amplitudes of the first and second harmonics) appears in the space for speakers of Hmong, which has a phonemic contrast between breathy and modal phonation, but not for English speakers, given that English does not have such a contrast. A final set of dimensions accounts for idiosyncratic variance components that vary from speaker to speaker. Thus, acoustic voice spaces appear to be structured first by biologically driven factors that describe acoustic variability for virtually every voice, secondly by characteristics of the language spoken, and lastly by other miscellaneous attributes of an individual's voice.

5. CONNECTING QUALITATIVE AND ACOUSTIC ASPECTS OF VOICE

The results in Section 3 strongly parallel those in Section 4. The primary dimensions that emerge from qualitative voice research—brightness, breathiness, roughness, and richness, which are associated with the distribution of spectral energy, irregularity/noise, and the location of the spectral centroid—are in essence the same dimensions that emerge from quantitative studies of voice acoustics (variability in the balance of harmonic and inharmonic energy and formant dispersion). These characteristics have empirical associations with biologically important voice characteristics, and are meaningful attributes of vocalization across species. This correspondence between the most common terms for voice and parameters that define the (seemingly universal) human acoustic voice space suggests that the meaning voices carry rests on a bedrock of biology. That is, we have evolved to phonate and to perceive voices in ways that reflect size and state of arousal, and these factors are the basis for both voice acoustics and the language we most commonly use to describe voices.

Wallmark and Kendall [31] have previously suggested a similar association between physical and perceptual measures of quality, pointing out that some commonly-applied descriptors reflect the fact that voices come from bodies, and that this may account in part for the fact that these terms in particular tend to re-appear across studies, cultures, and languages. The arguments in this paper take this account further, by examining not just why some descriptors link bodies to perceived voices, but also why this specific set of descriptors serves this purpose. Of course, finding empirical support for a small set of qualitative descriptors of voice does not mean such descriptors are good tools for quantifying quality. There is ample evidence that ratings on scales like "breathiness" and "roughness" are unreliable and subject to many kinds of measurement error [32]. This small set of acoustic measures is also inadequate to completely specify the sound of a voice in the ANSI sense, although they are an important part of a psychoacoustic model that does specify why specific voice samples sound the same or different [7]. The qualitative and acoustic dimensions discussed in this paper are also only a tiny subset of the measures and labels that can be used to assess or describe voice quality. Recall that both factor analytic and acoustic analyses produced a mass of idiosyncratic detail along with a few widely applicable dimensions. Further, language, as noted above, is a virtually limitless tool for evoking the meaning of a voice. It may be (in fact, it seems likely) that a model that completely maps from one domain to another is both theoretically and empirically impossible. This conclusion is consistent with the observation that many descriptions of voice convey its meaning, but not necessarily the way it sounds. Consider, for example, Raymond Chandler's description of private investigator Philip Marlowe's first encounter with a potential client:

> The voice I heard was an abrupt voice, but thick and clogged, as if it was being strained through a curtain or somebody's long white beard. [33]

Acoustic analysis would be uninformative in this case, which reflects Marlowe's evaluation of what he hears, but not the actual sounds themselves, which are rather hard to imagine.

Nevertheless, the small point of coincidence between qualitative and quantitative analyses of voice suggests that there is at least a foundation of information that links signals and specific aspects of the meaning they convey—in other words, that dependably links speakers to listeners without reference to external variables or context. Although the parameters described here in no way comprise a comprehensive model of voice quality, these results do suggest that there exists a bedrock of meaning—derived from the biological functions subserved by voice—that seemingly underlies both qualitative and acoustic approaches to voice. This foundation provides a way of explaining some aspects of the meaning of voice in terms of specific aspects of production and acoustics, and vice versa, thus spanning, at least in part, the timbral abyss. The task remaining for humanists and empiricists alike is to consider the extent to which a shared foundation can inform and advance their work. Understanding the meanings that inhere in voices, versus those that derive from listener-based factors, could inform humanistic discussions of voice quality; and a focus on those acoustic aspects of voice that are inherently and consistently meaningful could guide and structure the development of better acoustic and biomechanical models of voice. Exploiting the proposed common foundation shared by humanistic and empirical approaches to voice could help integrate studies of physical signals and their meaning, leading eventually to a truly interdisciplinary approach to voice.

6. REFERENCES

[1] Kempster, G.B., Gerratt, B.R., Verdolini Abbott, K., Barkmeier-Kraemer, J., and Hillman, R.E. 2009. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am. J. Speech Lang. Pathol.* 18, 124 – 132.

[2] Bele, I. 2007. Dimensionality in voice quality, *J. Voice* 21, 257 – 272.

[3] Kreiman, J., Gerratt, B.R., and Berke, G.S. 1994. The multidimensional nature of pathologic vocal quality. *J. Acoust. Soc. Am.* 96, 1291 – 1302.

[4] Kendall, R.A., and Carterette, E.C. 1993. Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Percep.* 10, 445 – 468.

[5] Malawey, V. 2020. *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice*. Oxford.

[6] Kreiman, J., Sidtis, D. 2011. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception.* Wiley-Blackwell.

[7] Kreiman, J., Lee, Y., Garellek, M., Samlan, R., Gerratt, B.R. 2021. Validating a psychoacoustic model of voice quality. *J. Acoust. Soc. Am.* 149, 457 – 465.

[8] Kreiman, J., Gerratt, B.R., and Khan, S. D.(2010. Effects of native language on perception of voice quality. *J. Phonetics* 38, 588 – 593.

[9] Lavan, N., Burston, L.F.K., and Garrido, L. 2018. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British J. Psychol.* 110, 576 – 593.

[10] Wallmark, Z. 2022. *Nothing but Noise: Timbre and Musical Meaning at the Edge.* Oxford.

[11] Fales, C. 2002. The paradox of timbre. *Ethnomusicology* 46, 56 – 95.

[12] Heidemann, K. 2016. A system for describing vocal timbre in popular song. *Music Theory Online* 22, 1 – 17.

[13] Hajda, J.M., Kendall, R.A., Carterette, E.C., Harshberger, M.L. 1997. Methodological issues in timbre research. In: Deliege, I., Sloboda, J. (eds), *Perception and Cognition of Music*. Psychology Press, 253 – 306.

[14] ANSI. 1960/1994. *Psychoacoustic Terminology: Timbre.* American National Standards Institute.

[15] Dolan, E.I., Rehding, A. (eds) 2018. *The Oxford Handbook of Timbre* (Oxford).

[16] Osgood, C.E., Succi, G.J., Tannenbaum, P.H. 1957. *The Measurement of Meaning*. University of Illinois Press.

[17] Hirano, M. 1981. *Clinical Examination of Voice.* Springer.

[18] Voiers, W.D. 1964. Perceptual bases of speaker identity. *J. Acoust. Soc. Am*. 36, 1065 – 1073.

[19] Lichte, W.H. 1941. Attributes of complex tones. *J. Exp. Psychol.* 28, 455 – 480.

[20] Pratt, R., Doak, P. 1976. A subjective rating scale for timbre. *Journal of Sound and Vibration* 45, 317 – 328.

[21] von Bismarck, G. 1974. Timbre of steady tones: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica* 30, 146 – 159.

[22] Zacharakis. A., Pastiadis, K., Reiss, J.D. 2014. An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Percep.* 31, 339 – 358.

[23] Alluri, V., Toiviainen, P. 2012. Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. *Music Percep.* 20, 297 – 310.

[24] Lee, Y., Keating, P., Kreiman, J. 2019. Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am* 146, 1568 – 1579.

[25] Lee, Y., Kreiman, J. 2022. Acoustic voice variation in spontaneous speech. *J. Acoust. Soc. Am.* 151, 3462 – 3472.

[26] Lee, Y., and Kreiman, J. 2023. Within- versus between-speaker acoustic variability in Thai. Presented at the 184th Meeting of the Acoustical Society of America.

[27] Anikin, A. 2020. A moan of pleasure should be breathy: The effect of voice quality on the meaning of

human nonverbal vocalizations. *Phonetica* 77, 327 – 349.
[28] Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B.R., Neubauer, J., Alwan, A. 2012. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *J. Acoust. Soc. Am.* 132, 2625 – 2632.
[29] Congdon, J.V., Hahn, A.H., et al. 2019. Hear them roar: A comparison of black-capped Chickadee (Poecile atricapillus) and human (Homo sapiens) perception of arousal in vocalizations across all classes of terrestrial vertebrates. *J. Comp. Psychol.* 133, 520 – 541.
[30] Fitch, W.T. 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques *J. Acoust. Soc. Am.* 102, 1213 – 1222.
[31] Wallmark, Z., Kendall, R.A. 2018. Describing sound: The cognitive linguistics of timbre. In: Dolan, E.I., Rehding, A. (eds), *The Oxford Handbook of Timbre.* Oxford, 578 – 608.
[32] Kreiman, J., Gerratt, B.R. 2000. Sources of listener disagreement in voice quality assessment. *J. Acoust. Soc. Am*. 108, 1867 – 1876.
[33] Chandler, R. 1949. *The Little Sister.* Vintage, p. 41.

What is a 'substantial' voice? An LTAS study on two Cantonese accents

Eugene Wong[1], Roxana Fung[2]

[1]University of Minnesota, [2]The Hong Kong Polytechnic University
wong0703@umn.edu, roxana.fung@polyu.edu.hk

ABSTRACT

Hong Kong Cantonese (HKC) and Guangzhou Cantonese (GZC) are two major accents of Cantonese, but their phonetic differences are understudied. Laymen claim that HKC is 'crispy' whereas GZC is 'substantial'. The current study examined their differences in the domain of voice quality. We analyzed text reading and free speech samples from 35 native talkers using LTAS. Results show that: (i) HKC exhibits more energy than GZC in F1, F2 and higher frequency regions; (ii) an amplitude peak exists at around F4 in GZC but not in HKC; and (iii) the two accents differ in mean spectral energy, spectral tilt, and level difference between f0 and F1 region (L1-L0). We explored the articulatory-acoustic relationships along the dimensions of glottal tension, glottal noise, and resonance between the two accents. A 'substantial' voice may indicate a more resonant speech induced by breathy-lax voice and lowered larynx in GZC.

Keywords: Cantonese, accents, voice quality, LTAS, sociophonetics

INTRODUCTION

1.1. Hong Kong Cantonese and Guangzhou Cantonese

Cantonese is a dialect of Chinese languages spoken in the southern part of China (mainly in Guangdong province) and in Hong Kong. There are two main accents of Cantonese: Hong Kong Cantonese (HKC), which is spoken in Hong Kong, and Guangzhou Cantonese (GZC), spoken in Guangzhou city. HKC and GZC are fully mutually intelligible as they share almost the same phonological system. However, native talkers claim that they have no difficulty in discerning the two accents. Although what phonetic features account for the identification of these two Cantonese variants have perplexed many minds, studies comparing the phonetic differences between the two accents are scare. One study examined the consonants, vowels and lexical tones in university students who spoke HKC and GZC [1]. The acoustic analyses showed that there were minimal differences in segmental features between the two accents.

Unlike professional linguists, many native talkers, however, have been enthusiastically expressing their opinion on the internet. Laymen have pointed out many lexical differences between the two accents. As for pronunciation, their descriptions were very abstract. For instance, they described HKC as 'soft, light, crispy and uplifting'; whereas GZC was "blunt, heavy, low, more cadence, giving people a 'substantial' feeling." Their perceptual descriptions indeed point towards the voice quality domain. Translating into acoustic terms, the distinction between the two accents may be caused by the differences in muscular tension and excitation of glottal noise. The so-called 'substantial' feeling seems to be related to the resonating quality of the voices, induced by larynx height.

1.2. Voice quality and social information

In fact, voice quality settings are considered the quasi-permanent features of someone's voice and are a major component of one's accent (see [2] and [3]). In a broad sense, voice quality features are the habitual settings of the larynx and the vocal tract when someone speaks. These articulatory settings may be constrained by the biological makeup of one's vocal organs, yet they could still be shared by the individuals of a social group. Esling [4] remarks that voice quality characteristics reflect the accent of the entire social group, with individuals' accent differing within the boundaries of the social norms. He further points out that native talkers are very good at detecting the speech patterns that do not originate in their community.

The study by [5] was the first attempt to explore the phonetic differences between HKC and GZC in the voice quality domain. That study involved an acoustic analysis on speech samples produced by 24 talkers of the two accents across three age groups with balanced gender. A total of 864 CV syllables extracted from sentences produced in a reading task were analysed using six acoustic parameters, namely H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3* and CPP. The results showed that there were differences in the phonation of the two accents. The measurements of all the acoustic parameters suggested that HKC was tenser than GZC, except H1*-H2*, which suggested the opposite. The seemingly contradicting results may be due to the small sample size or the nature of the speech samples. Although the vowels and tones were well controlled

in that study, there may be two limitations in the speech samples. First, the naturalness of voice may have been affected by a sentence reading task. Second, the short length of the syllables may not have been able to capture voice quality features that are supposed to be the longest-lasting linguistic components of an accent. Therefore, continuous speech samples are necessary to further examine the voice quality features of HKC and GZC.

1.3. Long-term average spectrum

As for acoustic measurements of continuous speech, [4] suggests that long-term average spectrum (LTAS) is a productive indicator of voice quality because voice quality is indeed an average of the accumulation of traits over time. LTAS is obtained by averaging the spectra over a long duration. It reflects the distribution of sound energy across a range of frequencies. It is assumed that this procedure will eliminate the contribution of segmental features (i.e. linguistic content), and allows the examination of voice quality features that exhibit quasi-permanently in one's voice. For example, one study [6] found that an amplitude peak between 3k to 4k Hz is correlated to a more resonant 'good speaking voice' in actors. There are also some measures that help to further characterize LTAS. Mean spectral energy, i.e., the average amplitude in a spectrum, has been used to quantify glottal tension [7, 8]. Spectral tilt, the ratio of energy between 0-1kHz and 1-5kHz, has been proposed to measure glottal noise [7, 9, 10]. Similarly, spectral slope, i.e., the ratio of energy between 1-5kHz and 5-8kHz, has also been used to quantify glottal noise [11, 12]. Finally, L1-L0 is a measure of the difference between f0 and F1 regions (i.e., 0-300Hz and 3-800 Hz). This measure has been used to examine the distinction between modes of phonation, as it indicates the degree of glottal adduction [11, 13]. L1-L0 has been found to correlate with strained and breathy voices in perception [14].

1.4. The current study

The purpose of this study is to continue the exploration started by [5] on the voice quality features that discern HKC from GZC using LTAS techniques. Sets of longer and more natural speech samples produced by a larger number of talkers are adopted. This study focuses on investigating whether the perceptual descriptions made by laymen can be supported by acoustic measurements. In particular, the following three aspects of voice quality features of the two Cantonese accents will be examined: acoustic resonance, glottal tension, and glottal noise.

METHODOLOGY

2.1. Talkers

We recruited 20 HKC (F= 10) and 15 GZC (F= 8) talkers to provide speech samples. The talkers' age ranged from 23 to 35 years (HKC: mean= 27.1, s.d.= 3.4; GZC: mean= 25.3, s.d.= 2.4). We recruited talkers of this age group because younger talkers of HKC and GZC maybe too similar in their accents due to frequent contact between the cities in recent years.

The HKC talkers and their parents were all born and raised in Hong Kong. All talkers speak HKC in their daily lives and do not use any other Chinese dialects to communicate with their family and friends. They also learnt English as a second language and use English occasionally in school and/or work context.

The GZC talkers were all born and raised in Guangzhou city, China. In addition, they all grew up in the four 'old' districts of Guangzhou (i.e., Dongshan, Yuexiu, Liwan, and Haizhu). The parents of GZC talkers were also born and raised in Guangzhou city. All talkers speak GZC on a daily basis and do not use any other Chinese dialects to communicate with their family and friends. They also learnt Mandarin Chinese as a second language and used Mandarin Chinese occasionally in their education and/or work context.

All talkers reported no history of speech, language nor hearing disorder. They also reported no history of smoking, heaving drinking, laryngeal health issues nor vocal training. All talkers provided consent to participate in the study.

2.2. Production experiment

The production experiment was carried out in a sound-proof booth at the Hong Kong Polytechnic University. Speech samples were recorded through a Telefunken M80 dynamic microphone at the sampling rate of 44.1 kHz. The full production experiment included four tasks: sustained phonation, sentence reading, text reading, and free speech. The current study reported only the data from the text reading and free speech tasks.

In the text reading task, talkers were instructed to read aloud a short passage three times. The passage describes a novel product from an online shopping portal 'Taobao'. It contains 256 Chinese character, which allowed us to extract 30 to 40 seconds of voiced speech in the LTAS analysis detailed below. The speech samples of the second trial were analyzed.

In the free speech task, talkers were asked to elaborate on topics related to their daily lives. The topics included 'Yum Cha' (Cantonese dim-sum food culture), favourite pop culture, experiences on public transport, and about their family and friends. They

were provided with question prompts that assisted them in narration. The speech samples of the first topic 'Yum Cha' were analysed and reported in this study. All talkers were able to talk continuously for 90 seconds, which allowed us to extract at least 40 seconds of voiced speech for subsequent LTAS analysis.

2.4. Data Analysis

Data analyses were conducted through Praat [15]. The voiced portions (including all Cantonese vowels and the approximants /j w l/) were first annotated and extracted. The pitch-corrected LTAS were then obtained at intervals of 100 Hz, up to 8 kHz. The LTAS were then normalized with all amplitudes expressed as negative values relative to the peak.

A total of five parameters were used to measure the LTAS, which include mean spectral energy (MSE), spectral tilt (ST), spectral slope (SS), speaker's formant (SPF), and L1-L0. MSE was measured as the average frequency across the entire LTAS (0-8 kHz). A higher (less negative) MSE is associated with more laryngeal tension. ST was measured as the ratio of energy between 0-1 kHz and between 1-5 kHz. SS was measured as the ratio of energy between 1-5 kHz and between 5-8 kHz. Higher ST and SS are associated with less aspiration noise, hence perceived less breathy. SPF was measured as the energy difference between 0-1 kHz and 3-4 kHz. A lower SPF suggests more resonant speech. L1-L0 is the level differences between the F1 region (300-800Hz) and F0 region (0-300 Hz). A high L1-L0 indicates high glottal adduction, whereas low L1-L0 indicates low glottal adduction.

3. RESULTS

3.1. Overall LTAS

To examine whether HKC and GZC differ in energy distribution in LTAS, we fitted mixed effect regression models to the amplitude values at each frequency bin, controlling for the effects of gender and task. All models included accent, gender, and task as the fixed effects and a random intercept for talker. Before examining the differences between accents, we first report the task effects as the continuous speech samples were collected from two different tasks. Overall, there were task effects only in the 500 and 600 Hz frequency bins across the entire LTAS ($p<0.05$). This suggests that there were minimal differences in voice quality features across the two continuous speech tasks.

As for differences between HKC and GZC, we found significant effects of accent on the amplitude values in the frequency bins of 6-700 Hz, 16-1900 Hz, and 55-6600Hz (all $p<0.05$). For all these frequency bins, HKC had higher amplitudes than GZC. Figure 1 illustrates the mean LTAS for HKC and GZC.

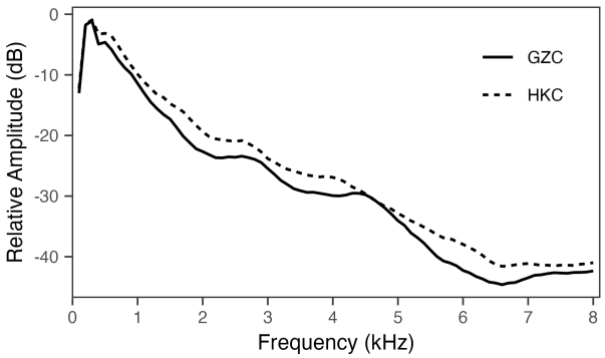


Fig 1. LTAS display of HKC and GZC.

In addition to the energy differences in the various frequency bins reported above, we also observed that GZC had a peak of energy around 4500 Hz, compared to the neighbouring frequencies. In HKC, the 4500 Hz region was rather flat. Although in terms of amplitude values the two accents were quite close, this peak at the F4 region may contribute to the differences in perceptual quality of resonant speech in the two accents of Cantonese. We will return to this in the discussion section below.

3.2. LTAS parameters

We also fitted mixed effect models to the LTAS parameters to determine the effect of accent, after gender and task effects were being controlled for. The models included a random intercept for talker. Due to the exploratory nature of our study, the significant and trending main effects ($p<0.10$) are also reported. Overall, we found a trending main effect of accent on MSE ($p=0.076$), ST ($p=0.086$) and L1-L0 ($p=0.099$). Means and standard deviation for each parameter are provided in Table 1. Overall, HKC has higher MSE, lower ST, and higher L1-L0 than GZC.

Table 1. Means and standard deviations (in parentheses) of the LTAS parameters.

	GZC	HKC
Mean Spectral Energy (MSE, in dB)	-29.3 (3.55)	-27 (3.78)
Spectral Tilt (ST)	0.0663 (0.0139)	0.0587 (0.0155)
Spectral Slope (SS)	0.808 (0.108)	0.783 (0.133)
Speaker's Formant (SPF, in dB)	25.7 (5.21)	22.9 (5.43)
L1-L0 (in dB)	-3.67 (2.81)	-2.09 (2.76)

4. DISCUSSION

In this study, we analyzed the continuous speech samples of HKC and GZC talkers using the LTAS procedure. We will discuss our findings with respect to the three aspects of voice quality: acoustic resonance, glottal tension, and glottal noise.

First, we found higher energy in HKC at the 6-700 Hz (F1 region), 16-1900 Hz (F2 region), and 55-6600Hz regions. In addition, we found that there was a peak at around 4500 Hz (F4 region) in GZC, but not in HKC. This feature, however, was not able to be captured by our statistical analysis. As Wu found no differences in the vowel quality of HKC and GZC, it is assumed that these acoustic features are attributed to articulatory settings in the lower vocal tract, such as larynx height and vocal folds vibratory patterns.

For glottal tension, we found that HKC had higher MSE than GZC. This replicates those in [5] where she found HKC producing tenser phonation than GZC. A higher MSE suggests that HKC has more tension in their voice quality settings than GZC. Such tension may come from adductive tension in the vocal folds, or an overall increase in tightness in the lower vocal tract. However, an acoustic analysis would not allow use to point out which set of muscles give rise to this tension. On the other hand, studies have also reported the interactions between resonance and tension: [7] suggested that a tenser phonation may be related to the use of 'posterior resonance' in bilingual Cantonese-English talkers when speaking HKC. The study by [16] also observed that increased tension is negatively correlated to the perception of resonance. Taken together, HKC talkers may be raising their larynx, which lead to constriction in the lower vocal tract [17]. This constriction increases the overall laryngeal tension and give rise to increased spectral energy in LTAS. Such articulatory settings may cause HKC to be perceived as less resonant than GZC.

As for glottal noise, we found that HKC had higher L1-L0. A lower L1-L0 value suggests more energy in f0 region compared to F1 region. In previous studies [13, 14], low L1-L0 was found to associate with low glottal adduction and the perception of breathy voice. Therefore, the lower L1-L0 suggests that the vocal fold vibratory pattern of GZC is similar to that of lax-breathy voice.

Furthermore, we also found that HKC had lower ST than GZC. As a lower ST is associated with increased aspiration noise and the perception of breathy voice [9, 10], this finding may contradict to the other findings in the current study, which point towards GZC being breathier. Interestingly, this seemingly contradictory result is also attested in [5] and [7]. In the study by [7], LTAS was used to examine the voice quality differences when HKC talkers spoke Cantonese versus English. The findings also showed that HKC talkers exhibited more glottal tension (as indicated by MSE) and more glottal noise (as indicated by ST) when they spoke Cantonese, compared to English. In the study by [5], the various acoustic parameters suggested that HKC was tenser than GZC, except H1*-H2*, which suggest the reverse. In the current study, the increase in glottal tension and noise in HKC maybe a result of the raised larynx posture proposed earlier. According to the Laryngeal Articulator Model [17], raising the larynx would induce constriction in the larynx, narrowing and shortening the epilaryngeal tube, giving rise to turbulent noise. It should be noted that phonation is a complex mechanism, and the interactions of glottal tension and glottal noise are not straightforward. Further studies are required to examine the articulatory mechanism of HKC and GZC. Perceptual effects of these acoustic features are also warranted, particularly for the perception of resonance, as it is traditionally associated with vocal health but not the perception of accent characteristics.

In the last section, we explore how the acoustic findings can relate to the laymen's perceptual descriptions. Overall, our results suggest that HKC exhibits higher glottal tension, and more glottal noise than GZC. Recall that HKC was described as 'light, soft and crispy'. 'Crispy' may refer to the tenser muscular settings, whereas 'light, soft' may be related to increased noise and the perception of less resonant speech, presumably due to raised larynx. Normally, a lowered larynx is considered more relaxed and is the optimal larynx height for speech [18].

On the other hand, GZC was described as "blunt, heavy, low, more cadence, giving people a 'substantial' feeling." The impression of 'heavy and low' may come from a rather lax muscular settings and breathier voice in GZC, compared to HKC. The 'substantial' feeling may be indicating that GZC is more resonant than HKC, presumably due to its lower larynx height.

To conclude, our acoustic analysis provided some evidence to support the voice quality differences between the two accents and the perceptual descriptions by native talkers. Our findings are promising amid the scarcity of studies in the relationships between perception, articulation, and acoustics of voice quality in different accents. They should be followed up by perception experiments to investigate whether native talkers are able to discern the two accents using the acoustic cues of resonance, glottal tension, and glottal noise.

5. ACKNOWLEDEGMENT

This study was supported by an internal competitive research grant from the Hong Kong Polytechnic University offered to the second author.

6. REFERENCES

[1] Wu, W. L. 2006. A comparative analysis of the phonetics of Hong Kong Cantonese and Guangzhou Cantonese. M.Phil. thesis, University of Hong Kong, Hong Kong SAR.

[2] Abercrombie, D. 1967. *Elements of general phonetics*. Edinburgh University Press.

[3] Laver, J. D. M. 1968. Voice quality and indexical information. *British Journal of Disorders of Communication*, *3*(1), 43–54.

[4] Esling, J. H., Edmonson, J. A. 2011. Acoustical analysis of voice quality for sociophonetic purposes. In: di Paolo, M., Yaeger-Dror, M. (eds), *Sociophonetics: A student's guide*. Routledge, 131-148.

[5] Fung, R. S. 2015. Voice quality: A preliminary study on the phonetic distinctions of two Cantonese accents. *Proc. 18th ICPhS* Glasgow.

[6] Leino, T., Laukkanen, A. M, Radolf, V. 2011. Formation of the actor's/speaker's formant: A study applying spectrum analysis and computer modeling. *J Voice* 25(2), 150-158.

[7] Ng, M. L., Chen, Y., Chan, E. Y. 2012. Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers- a long-term average spectral analysis. *J Voice* 26(4), 171-176.

[8] Fuller, B. F., Horii, Y. 1988. Spectral energy distribution in four types of infant vocalizations. *Journal of Communication Disorders*, *21*(3), 251-261.

[9] Mendoza, E., Valencia, N., Muñoz, J. and Trujillo, H., 1996. Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *J Voice* 10(1), 59-66.

[10] Hillenbrand, J., Houde, R. A. 1996. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *J Speech Lang Hear Res* 39(2), 311-321.

[11] Guzman, M., Correa, S., Munoz, D.,Mayerhoff, R., 2013. Influence on spectral energy distribution of emotional expression. *J Voice* 27(1).

[12] Peters, J. 2018. Regional Variation of Voice Quality: A pilot study of High and Low German in the Bersenbrücker Land. *Zeitschrift für Dialektologie und Linguistik*. 1-34.

[13] Kitzing, P. 1986. LTAS criteria pertinent to the measurement of voice quality. *J Phon* 14(3-4), 477-482.

[14] Antonetti, A. E. D. S., Siqueira, L.T.D., Gobbo, M.P.D.A., Brasolotto, A.G., Silverio, K.C.A. 2020. Relationship of cepstral peak prominence-smoothed and long-term average spectrum with auditory–perceptual analysis. *Applied Sciences*, 10(23).

[15] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program]. Version 6.2.14. Accessed: May 2022. [Online]. Available: https://www.praat.org

[16] Master, S., De Biase, N., Chiari, B. M., Laukkanen, A.M. 2008. Acoustic and perceptual analyses of Brazilian male actors' and nonactors' voices: long-term average spectrum and the "actor's formant". *J Voice 22*(2), 146-154.

[17] Esling, J. H., Moisik, S. R., Benner, A., Crevier-Buchman, L. 2019. *Voice quality: The laryngeal articulator model.* Cambridge University Press.

[18] Elliot, N., Sundberg, J., Gramming, P. 1997. Physiological aspects of a vocal exercise. *J Voice* 11(2), 171-177.

VOWEL-INITIAL WORDS AND GLOTTALIZATION: A CORPUS STUDY OF CHICHICASTENANGO K'ICHE'

Elizabeth Wood

University of Texas at Austin
elizabethwood@utexas.edu

ABSTRACT

Although word-initial glottal stops and glottalization are extensively observed in Mayan languages, the exact distribution is debated. This acoustic study measures full glottal closures as well as several measures of glottalization in each third of each word-initial vowel in a corpus of naturalistic speech from the Chichicastenango dialect of K'iche'. Results show that full closures occur only in a minority of cases and are more likely on words with initial stress or in IP-initial position. Words following a pause or a vowel or with initial stress show greater cues to glottalization only in the first part of the vowel or with decreasing effect sizes over the course of the vowel, consistent with coarticulation with a word-initial glottal segment. Words in IP-initial position, in contrast, show consistent glottalization cues across the vowel, suggestive of an IP-initial phonetic effect or boundary marker.

Keywords: Glottalization, corpus phonetics, K'iche' (Mayan)

1. INTRODUCTION

Word-initial glottalization and glottal stops on otherwise vowel-initial words are attested across the Mayan language family, including in K'iche', spoken in the Guatemalan highlands by around 900,000 people [1]. Influenced by the commonly repeated generalization that all roots or words must begin with a consonant in Mayan languages [2, 3], many descriptions of K'iche' take auditory perception of word-initial glottalization as evidence for the presence of glottal stops. However, the exact distribution of these glottal stops is debated, including whether they are phonemic [4] or epenthetic [5, 6] and whether they occur on all otherwise vowel-initial words [4, 7] or only in certain environments, such as on non-prefixed vowel initial roots [5] or on stressed monosyllables or words that are utterance-initial or preceded by a vowel [6].

This paper describes the results of an acoustic study of word-initial glottal stops and glottalization in K'iche' using a corpus of naturalistic speech from the Chichicastenango dialect. Two main questions are addressed: Where do full closures and glottalization occur? And is the time course of these cues consistent with the existence of a word-initial glottal stop segment?

2. METHODS

2.1. The corpus

The data comes from a corpus of 2 hours and 40 minutes of naturalistic, narrative speech recorded in 2018-2019. The two male speakers and 10 female speakers range in age from young adults to elders and all come from the town or vicinity of Chichicastenango, Guatemala, a dialect area that has been historically understudied. Speakers were asked to discuss a particular topic, such as local history and traditions, recipes, or traditional stories, and were allowed to speak freely for as long as they wished. Speakers were recorded using a Zoom Hn4 digital recorder at a sampling rate of 44.1 kHz, using either the internal microphone or connected to a Shure SM10A headset microphone. The recordings were transcribed in the K'iche' orthography and (if the speaker consented) archived in the Archive of the Indigenous Languages of Latin America [8].

2.2. Inclusion criteria and segmentation

Every word in the corpus that begins with a vowel or glottal stop in the transcription was included in this study, with the exception of any uncertain transcriptions, for a total of 2628 tokens. In the segmentation of word-initial vowels, the boundaries between vowels and consonants were marked primarily with reference to the intensity curve: where the amplitude begins to increase adjacent to a fricative, at the sudden jump in amplitude next to a voiced stop or sonorant, after the burst/aspiration for a preceding voiceless stop and at the offset of voicing for a following voiceless stop. Changes in formants and antiformants were used

as a secondary indicator as needed. The boundary between two vowels in hiatus was located at the middle of the formant transition, and secondarily at the location of the sudden change in the intensity curve. When ambiguous, the midpoint of the whole VV interval was used. Word-initial diphthongs, found in the corpus only in the Spanish borrowings /ue.bon/ 'lazy' and /auk.si.liar/ 'assistant', were segmented together as one vowel unit. A full glottal closure was segmented separately from the word-initial vowels when it consisted of either at least 20 ms of silence or a single glottal pulse followed by silence which altogether lasted at least 20 ms (see [9, 10] for similar metrics). Other periods of laryngealized voicing were included in the vowel.

2.3. Measurements

A Praat [11] script was used to measure spectral tilt (H1-H2, H1-A1, H1-A2, H1-A3), periodicity (average HNR, jitter, shimmer), intensity minimum and pitch minimum, as well as the first two formants as these affect spectral tilt measures [12, 13]. The measurements were taken in each third of the vowel, or for very short vowels in three partially overlapping segments that each were the duration of the inverse of the pitch floor used for measurements of that vowel. In order to ensure the most accurate measurements of F0 and other dependent measures, the pitch range was varied as needed for each speaker and vowel. After finding a range mostly suitable for a given speaker, each vowel was visually inspected using those settings and if necessary individual adjustments were made to the pitch floor, pitch ceiling and voicing threshold until Praat accurately located the pitch pulses in that vowel. These specialized settings were then used in the script to take the final measurements.

2.4. Categorization

Each word was classified according to the following factors, including those mentioned in the previous literature as relevant to the occurrence of word-initial glottal stop: syllable count (monosyllabic or polysyllabic), stress position (stress on initial syllable or no), word origin (Spanish loan or no), morpheme type (root or prefix), following a glottalized (ejective or implosive) consonant or glottal stop (yes or no), following a vowel (yes or no), following a pause (yes or no), intonational phrase position based on existence of IP-final boundary tone on the preceding word (initial or non-initial), word initial morpheme, and text (recording code).

2.5. Hypotheses

If a given factor promotes the realization of a glottal stop at the beginning of an otherwise vowel-initial word, then there should be a significantly higher rate of full glottal closures preceding those vowels and/or significantly higher rates of acoustic cues to glottalization in the early part of the vowel (lower spectral tilt, lower HNR, higher jitter and shimmer, lower minimum intensity and pitch). If this glottalization is found throughout the vowel, it must come from a different source. Thus, based on the previous descriptions, higher rates of glottalization cues are expected in the beginning of the vowel for words following a word ending in a vowel or pause [4, 6], for words that are initial in the IP [6], for words that are monosyllabic and have initial stress [6], for roots [5], and for Spanish borrowings [5].

2.6. Statistical analysis

Results were analyzed in R [14] with linear mixed effects models using the package lme4 [15]. A separate model was made for each acoustic measure in each third of the vowel as well as another model for full closures.

In the first phase each model included syllable count, stress position, word origin, morpheme type, preceding glottalized consonant, preceding pause, preceding vowel, IP position, F1 and F2 as fixed effects and initial morpheme and text as random effects. It was found that the variables syllable count, word origin and morpheme type yielded no or almost no significant effects, and were thus eliminated from the model. In the second phase, each model included the random effects, the remaining fixed effects (IP position, stress, preceding pause, preceding glottalized consonant, preceding vowel), and additionally each possible interaction of these fixed effects. The baseline categories for each factor were IP-medial, no initial stress, no preceding glottalized consonant, no preceding pause, and no preceding vowel.

3. RESULTS

3.1. Full closures

Excluding words following a pause, where it is difficult to locate an initial glottal closure, full closures occur in 8.8% of the data. This raises to 11.8% for those following a vowel, 13.63% for those following a glottalized consonant, 15.7% for those IP-initial, and 20.8% for those with initial stress.

The results of the statistical analysis show significant positive effects of IP position and stress (more full closures). The interaction between the two is negative. The two effects do not stack: only unstressed vowels show a greater rate of full closures when IP-initial (stressed vowels do not), and only IP-non-initial vowels shows a greater rate of full closures when stressed (IP-initial vowels do not).

3.2. Acoustic cues to glottalization

The following tables summarize the results of the statistical models for each measure and each of the fixed effects. Only the coefficients of significant effects (with $p < 0.05$) are shown in the tables; blanks indicate no significant effect. Effects consistent with glottalization (negative effect for spectral tilt, HNR, intensity minimum and pitch minimum and positive effect for jitter and shimmer) are shaded in gray. E.g., in Table 1 -8.198 is the coefficient of the effect of preceding glottal(ized) consonant on H1-A1 in the first third of the vowel, and is shaded in gray because it is in the direction of glottalization.

Table 1 shows the results for words that are immediately preceded by word ending in a glottalized consonant or glottal stop. Overall this table shows a negative effect of spectral tilt only in the first third of the vowel.

Measure	1st third	2nd third	3rd third
H1-H2			
H1-A1	-8.198		
H1-A2	-5.107		
H1-A3	-3.516		2.783
HNR			
Jitter			
Shimmer			
Intensity min			
Pitch min			

Table 1: Effects for preceding glott. consonant

Table 2 shows the effects for words immediately preceded by a word ending in a vowel. The negative effect of spectral tilt extends throughout the vowel, with the absolute value of the coefficients decreasing over time. Aperiodicity is greater only in the end of the vowel.

Table 3 shows the significant effects for words that are immediately preceded by a pause. Spectral tilt, periodicity, intensity and pitch all show effects consistent with glottalization only in the beginning of the vowel.

Measure	1st third	2nd third	3rd third
H1-H2	-2.194	-1.402	
H1-A1	-6.310	-3.748	-1.591
H1-A2	-4.381	-2.234	
H1-A3	-4.657	-3.092	
HNR	1.892	-1.081	-1.001
Jitter		0.013	0.013
Shimmer			
Intensity min	2.160		
Pitch min	-14.678	-12.357	

Table 2: Effects of preceding vowel.

Measure	1st third	2nd third	3rd third
H1-H2		2.427	2.117
H1-A1	-2.557	2.546	2.394
H1-A2		2.656	2.072
H1-A3		3.416	2.445
HNR	-2.875	-2.551	
Jitter	0.029		
Shimmer	0.102	0.069	
Intensity min	-5.423		1.353
Pitch min	-9.715		9.378

Table 3: Effects of preceding pause.

Table 4 shows the significant effects for words that have stress on the initial syllable. There is an effect consistent with glottalization for every measure except for the pitch minimum, but these are found only in the first third of the vowel.

Measure	1st third	2nd third	3rd third
H1-H2	-1.622		
H1-A1	-4.045		
H1-A2	-3.409		
H1-A3	-2.783		
HNR	-2.277		2.450
Jitter	0.023		
Shimmer	0.068		
Intensity min	-2.695	1.564	
Pitch min			34.437

Table 4: Effects of initial stress.

Table 5 shows the significant effects for words that are initial in the intonational phrase. There are negative effects to spectral tilt that extend throughout the vowel, with similar coefficients throughout. Additionally, there are positive effects of jitter and shimmer and negative effects to pitch, intensity and HNR, all consistent with greater glottalization, though these are located only in the beginning and middle of the vowel.

Measure	1st third	2nd third	3rd third
H1-H2			-2.024
H1-A1	-2.881	-2.854	-1.996
H1-A2			
H1-A3		-1.625	-1.567
HNR	-2.074	-3.040	
Jitter	0.025	0.031	
Shimmer	0.074	0.042	
Intensity min	-1.572	-1.070	
Pitch min		-9.179	

Table 5: Effects of IP-initial position.

Finally, some interactions were significant, mostly of the measures of periodicity, pitch and intensity. These primarily reflect cases where synergistic forces do not stack (e.g. IP-initial position leads to lower HNR only for words not preceded by a pause, as preceding pause itself already lowers HNR), where the experimental factors result in potentially contrary forces (e.g. stress itself may result in increased pitch/intensity but glottalization in a decrease) or where a particular effect is enhanced when combined with stress (e.g. the increase in jitter in the first third of stressed vowels preceded by a vowel does not also occur for unstressed vowels).

4. DISCUSSION

The results showed two distinct types of effects on the glottalization cues: while the effects of stress, preceding pause, preceding glottalized consonant and preceding vowel are localized (or strongest) at the beginning of the vowel, the effects of IP-initial position are found throughout the vowel. This is exemplified in Figure 1. The stressed vowel in /ˈʊts/ 'good' on the left has visible glottalization in the first part only, while the IP-initial vowel /a/ in the focus marker /a.ˈɾe/ on the right, produced by the same speaker, shows visible glottalization throughout.

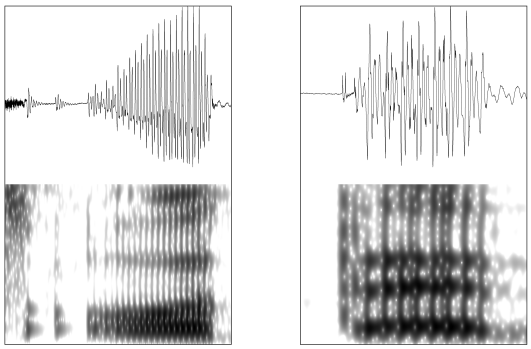


Figure 1: Initial vowels in ˈuts (left), a.ˈɾe (right)

The localized effects of initial stress and preceding pause and the consistently decreasing effects of preceding vowel, as is shown by their similarity to the results for a preceding glottalized/glottal consonant, are consistent with the result of coarticulation with a word-initial glottal segment. Full glottal closures are also significantly more likely on words with initial stress. No support was found for an effect of word origin or morpheme type (c.f. [5]). This provides support for Larsen's [6] description of word-initial glottal stop as occurring only on words that are stressed and monosyllabic (i.e. have initial stress), utterance-initial, or preceded by a word ending in a vowel. However, if the target of this initial glottal gesture is in fact a stop, it is often reduced in naturalistic speech, as full closures only occur in a minority of cases. The prototypical realization of this glottal segment is not a stop at all but rather glottalization of an adjacent vowel.

IP-initial position, in contrast to the other factors, results in glottalization cues that persist strongly throughout the whole vowel. This is not consistent with the effects of an initial glottal segment, but rather must be a phonetic or phonologized effect of this initial prosodic position, which marks the left edge of intonational phrases. There are also higher rates of full closures in IP-initial position; these may represent greater strengthening of this IP-initial glottal constriction gesture. This IP-initial glottalization contrasts with a domain-final spread glottis gesture found in K'iche' and related languages which results in aspiration of stops, devoicing of sonorants and [h] insertion on vowel-final words [16, 17]; combined, these two forces maximize the contrast in voice quality between initial and final position, enhancing the boundary between them.

Each of the significant factors - stress, prosodic position, and preceding context - have been previously identified as relevant to similar alternations of word-initial glottal stops and glottalization in the cross-linguistic literature. However, these previous studies have been mostly focused on heavily studied (Indo-)European languages such as Dutch [18], English [19, 20, 21], German [22, 23, 24], Finnish [25] and Maltese [26]. Thus, this study extends previous research by showing similar patterns in an unrelated and understudied language.

5. ACKNOWLEDGMENTS

I would like to thank all of the K'iche' speakers who are represented in the corpus, without whom this work would have been impossible. I would also like to thank Scott Myers, Brandon Baird, Paige Erin Wheeler and the members of the UT Austin Research in Phonetics and Phonology and Research in Documentary, Descriptive and Historical Linguistics research groups for helpful feedback on this project.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 000392968. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] M. Richards, *Atlas lingüístico de Guatemala*. Editorial Serviprensa Guatemala, 2003.

[2] G. Polian, "Morphology," in *The Mayan Languages*, 1st ed., J. Aissen, N. C. England, and R. Zavala Maldonado, Eds. London: Routledge Ltd., 2017, ch. 8, pp. 201–225.

[3] N. C. England and B. O. Baird, "Phonology and phonetics," in *The Mayan Languages*, 1st ed., J. Aissen, N. C. England, and R. Zavala Maldonado, Eds. London: Routledge Ltd., 2017, ch. 7, pp. 175–200.

[4] T. Kaufman, "Initial glottal stop in Mayan languages," *Unpublished ms*, 2015.

[5] R. Barrett, "The evolutionary phonology of glottal stops in K'ichean," in *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, vol. 33, no. 1, 2007, pp. 19–29.

[6] T. W. Larsen, "Manifestations of ergativity in Quiché grammar," Ph.D. dissertation, University of California, Berkeley, 1988.

[7] C. D. López Ixcoy, *Gramática K'ichee'*. Guatemala: Fundación Cholsamaj, 1997.

[8] E. Wood, "The K'iche' collection of Elizabeth Wood," The Archive of the Indigenous Languages of Latin America, ailla.utexas.org, 2019-, access: public. PID ailla: 271517.

[9] M. Frazier, "The production and perception of pitch and glottalization in Yucatec Maya," Ph.D. dissertation, The University of North Carolina at Chapel Hill, 2009.

[10] B. O. Baird, "Phonetic and phonological realizations of 'broken glottal' vowels in K'ichee'," in *Proceedings of formal approaches to Mayan linguistics: MIT working papers in linguistics*, vol. 63, 2011, pp. 39–49.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2019, [Computer program]. Version 6.0.49, retrieved from http://www.praat.org/.

[12] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 1. IEEE, 2004, pp. I–669–I–672.

[13] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019, [Computer program]. Version 1.4.1103, retrieved from https://www.R-project.org/.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] R. Henderson, "Morphological alternations at the intonational phrase edge: The case of K'ichee'," *Natural Language & Linguistic Theory*, vol. 30, no. 3, pp. 741–787, 2012.

[17] S. AnderBois, "Strong positions and laryngeal features in Yukatek Maya," in *NELS 39*, 2008.

[18] W. Jongenburger and V. J. van Heuven, "The distribution of (word initial) glottal stop in Dutch," *Linguistics in the Netherlands*, vol. 8, no. 1, pp. 101–110, 1991.

[19] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of phonetics*, vol. 24, no. 4, pp. 423–444, 1996.

[20] M. Garellek, "Glottal stops before word-initial vowels in American English: distribution and acoustic characteristics," *UCLA Working Papers in Phonetics*, vol. 110, no. 1, 2012.

[21] N. Umeda, "Occurrence of glottal stops in fluent speech," *The Journal of the Acoustical Society of America*, vol. 64, no. 1, pp. 88–94, 1978.

[22] K. J. Kohler, "Glottal stops and glottalization in German," *Phonetica*, vol. 51, no. 1-3, pp. 38–51, 1994.

[23] B. Alber, "Regional variation and edges: Glottal stop epenthesis and dissimilation in standard and southern varieties of German," *Zeitschrift für Sprachwissenschaft*, vol. 20, no. 1, pp. 3–41, 2001.

[24] B. Pompino-Marschall and M. Żygis, "Glottal marking of vowel-initial words in German," *ZAS Papers in Linguistics*, vol. 52, pp. 1–17, 2010.

[25] M. Lennes, E. Aho, M. Toivola, and L. Wahlberg, "On the use of the glottal stop in Finnish conversational speech," in *The Phonetics Symposium 2006, Publications of the Department of Speech Sciences, University of Helsinki*, vol. 53, 2006, pp. 93–102.

[26] H. Mitterer, S. Kim, and T. Cho, "The glottal stop between segmental and suprasegmental processing: The case of Maltese," *Journal of Memory and Language*, vol. 108, pp. 1–19, 2019.

Laughter patterns in multi-speaker conversation data: comparison between spontaneous laughter and intentional laughter

Kexin Wang[1)], Carlos Ishi[2)], Ryoko Hayashi[3)]

Kobe University[1, 3)], ATR[1,2)], RIKEN[2)], Japan
wkx19970303@hotmail.com, carlos@atr.jp, rhayashi@kobe-u.ac.jp

ABSTRACT

This paper presents the results of an analysis of representative patterns of laughter by 4 males and 4 females, extracted from a corpus of natural conversation speech. In this study, two types of laughter were categorized by a production viewpoint: spontaneous laughter and intentional laughter. The analysis of laughter patterns focused on the number of calls, vowel-quality, and other acoustic characteristics, such as breathy voice and pressed voice. The results showed different trends between the patterns of spontaneous laughter and intentional laughter. A common pattern in spontaneous laughter among all the subjects has been observed, while the patterns of intentional laughter were more individual-dependent and varied, for example, with the usage of nasal sounds and ingressive laughter. However, many of the laughter patterns were found in both spontaneous laughter and intentional laughter, which indicates that other contextual information will also be necessary to understand laughter in dialogue communication.

Keywords: laughter, individuality, spontaneous laughter, intentional laughter, laughter pattern.

1. INTRODUCTION

Laughter is a common behavior throughout life and forms an essential component of nonverbal communication in human social interaction. It is reported that laughter is frequently produced in diverse social situations and with large behavioral variability [1]. For example, [2] classifies laughter into seven types, including humorous laughter, social laughter, ignorance laughter, anxiety laughter, derision laughter, apologetic laughter, and tickling. Moreover, it is reported that there is evident diversity in laughter sex and individual identity [3, 4]. In [3], acoustic analysis was conducted for naturally produced laugh bouts recorded from 97 young adults as they watched funny video clips. The outcomes proved that remarkable variability is found in fundamental frequency characteristics and individuals have their own unique laughter patterns.

On the other hand, the evaluation of synthetic laughter and natural laughter showed that listeners were able to use acoustic characteristics such as laugh duration and F0 to understand the intentions and emotions of the laughing person [5,6]. Furthermore, the contrast between voiced laughter and voiceless laughter has been shown to be an essential difference in laughter variation, with voiced laughter being associated with more positive impressions than voiceless one [7,8].

However, most of the works related to laughter use dialogue data from pairs of speakers, while only a few deal with situations where multiple speakers are present in the environment [9]. We believe that richer phonetic features and laughter patterns exist in multi-speaker conversation data, which is more appropriate for studying the individuality of laughter. Thus, we analyzed laughter in three-way free conversation data in the present work.

2. SPEECH DATA AND ANNOTATION

2.1. Description of the multi-speaker conversation data

A three-party conversation database was used in our analysis. Fig. 1 shows the scenery of the soundproof room where data was collected. The three speakers sat equidistantly around the center table. The database contains several sessions of topic-free conversations of 15 to 20 minutes. Headset microphones and cameras were set up for each subject. Around 13 hours of conversation data by 4 males and 4 females (described as F1-F4 and M1-M4 below) were used in the present laughter analysis. The males are graduate students in their 20s, and the females are research assistants in their 30s to 40s.



Figure 1: The scene of the three-party dialogue data recording.

Depending on the positional relationship between speech and laughter, laughter is classified as "speech laughter" which occurs along with speech utterance, and "stand-alone laughter" which occurs alone without speech [10, 11]. In this work, only stand-alone laughter is targeted.

2.2. Structure of laughter

In [12-15], laughter was analyzed at three primary levels: "segment," "call," and "bout." Fig. 2 shows the hierarchical structure of laughter. At the segmental level, laughter is composed of consonants and vowels. A different segment is recognized when the spectrogram components reflect a clear change in production mode. In actual laughter, it is reported that the first formant is raised and vowels tend to be centralized (schwa) [16]. The call level consists of a combination of consonants and vowels or a single vowel. The call is also called "note" or "syllable." At the bout level, laughter consists of one or more calls, and is typically produced during one exhalation. Although many bouts ended with audible inhalations or exhalations, these sounds were not included in bout-level characterizations unless they were deemed to be critical to the laugh itself [3]. Furthermore, a sequence of multiple bouts is called a laughter episode.

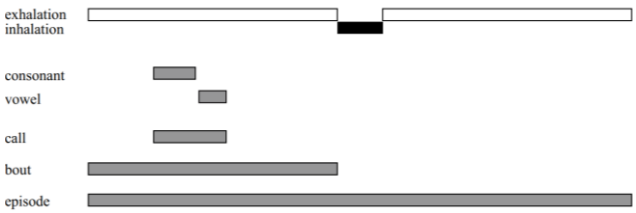


Figure 2: The hierarchical structure of laughter (adapted from [15]).

In this work, we first split the laughter episodes from all speech data, and then split the laughter episodes into bout levels. The following sections 2.3 and 2.4 are both performed at the bout level.

2.3. Spontaneous laughter and intentional laughter

In speech communication, whether the expression is natural or created is a crucial point to be considered [17]. From this production viewpoint, we categorized laughter into two types: spontaneous laughter and intentional laughter.

Spontaneous laughter includes the speaker's involuntary emotion, while intentional laughter is related to the speaker's attitude, intention, and intentionally generated emotion. Unlike spontaneous laughter, when a speaker produces intentional laughter, the listener's reaction might be taken into account for the expression. However, we also recognize the existence of spontaneous laughter with an intentional part, as well as intentional laughter with a spontaneous part. Furthermore, individuality is unavoidably imparted to speech behavior.

The laughter types were annotated by three research assistants, based on audio and video of the laughter events with 5-second pre- and 3-second post-context. Laughter events with two or more matches were used in the analysis. Table 1 shows the number of laughter events per speaker.

Speaker	Spontaneous	Intentional
F1	43	42
F2	16	60
F3	26	26
F4	36	56
M1	22	39
M2	10	9
M3	7	28
M4	21	56

Table 1: The number of laughter events per speaker.

2.4. Annotation of laughter patterns

Taking knowledge from past works on laughter into account, we annotated laughter from the following three viewpoints, as shown in Table 2. Firstly, we divide laughter into four groups by the number of calls: laughter with only 1 call, laughter with 2 or 3 calls, laughter with 4 to 9 calls, and very long laughter with 10 calls or more. Secondly, we describe the vowel-quality of laughter, including Japanese vowels, schwa, and nasal sounds, as well as breathiness. In addition, other prosodic and voice quality characteristics such as pressed voice, falsetto voice, intensity, and some special ways of laughter are also described.

Number of calls		Label 1
1 call		1
2/3 calls		2
4-9 calls		3
10 calls or more		4
Vowel-quality		Label 2
Japanese vowels		a, i, u, e, o
schwa		@
nasal		n
breathiness	pharynx	x
	velar	h
Other voice quality characteristics		Label 3
pressed voice		*
falsetto voice		f
strong (high intensity)		s
weak (low intensity)		w
ingressive laughter		g
whole-voiced (sequence of short calls)		~

Table 2: The annotation of laughter patterns.

The "whole-voiced laugh" item indicates a continuous sequence of calls with short duration, where usually F0 makes small up-down movements between the short calls.

The annotation of laughter patterns is conducted by the first author and two research assistants. The agreement rate between the annotators was not high for laughter patterns (around 50% agreement for two or more annotators). Disagreement mainly occurred in vowel-quality perception (for example, between the vowel /a/ and the schwa vowel). In case of disagreement, the first author's judgment prevailed.

3. ANALYSIS OF THE LAUGHTER PATTERNS

3.1. Variation of laughter call and vowel quality

Table 3 shows the vowel-quality items that appeared in the laughter segments arranged by the laughter type (spontaneous or intentional) and the groups of laughter calls (1 to 4). The phonetic symbols are ordered according to the number of occurrences.

		Vowel-quality
sp (181)	1 (19)	x (14), @ (4), a (2), n (2)
	2 (52)	@ (38), x (23), a (10), n (6), * (4), e (3)
	3 (105)	@ (63), a (45), x (18), * (14), n (3), h (3), f (3)
	4 (5)	@ (4), a (2), * (3)
in (316)	1 (58)	x (47), n (15), @ (10), a (8), e (5)
	2 (147)	@ (67), x (51), n (44), e (14), a (11), h (7), u(5)
	3 (108)	@ (68), x (23), a (18), n (18), e (9)
	4 (3)	@ (3)

Table 3: The distributions of vowel-quality items for different laughter call groups and laughter types. (Number of occurrences are within brackets.)

It can be first observed that laughter call groups 2 and 3 (i.e., the groups of 2/3 calls and 4-9 calls) show the largest occurrences, with group 3 being predominant in spontaneous laughter, and group 2 predominant in intentional laughter.

Regarding vowel-quality, it can be observed that schwa vowels (@) are predominant in all groups except for group 1 (single call), followed by breathiness (x), the vowel /a/, and the nasal sounds /n/. For group 1, the laughter is almost composed of breathy phonation (x). Regarding the call groups 2 and 3, other vowels such as /e/ and /u/, and pressed voice (*) also appear with higher frequency.

We also noticed that vowel-quality could change inside a laughter bout. This change was observed in both spontaneous laughter and intentional laughter, with a tendency for the vowels like /a/, /i/, and /u/ to change into schwa vowels (which can be considered as the default vowel quality during laughter), and for schwa vowels to change into the vowels /a/ (which corresponds to laughter realized by open mouth, mainly in big laughter).

3.2. Relationship between laughter patterns and laughter types

As Table 3 shows, compared to spontaneous laughter, intentional laughter has more vowel-quality variation, especially the vowels /e/ and nasal sounds /n/. On the other hand, spontaneous laughter has more patterns with pressed voice (*).

The typical laughter patterns of each speaker were grouped, and those that occurred twice or more were summarized in Table 4. Fig. 3 shows the waveform, F0 and spectrogram of some representative patterns.

		Typical pattern groups	Other features
F1	sp (43)	1x (5), 2@/2@x (5), 3a (7), 3@/3@x (19)	s (19), * (10)
	in (42)	1x (7), 1@x/1ex/1nx (7), 2@/2@x (4), 3@ (12)	w (6), ~ (9)
F2	sp (16)	2@x (2), 3@/3@x (9)	w (2), g (3)
	in (60)	2@/2@x (14), 2n/2nx (7) 3@/3@x (14), 3e/3ex (6)	s (2), w (5), g (2)
F3	sp (26)	3@ (5), 3a (7)	s (9), g (3), ~ (2)
	in (26)	2a(2), 3@ (4), 3a (3), 3n (3)	s (6), w (2), g (3), ~ (2)
F4	sp (36)	2@ (5), 2a (2), 3@/3@x (11), 3a (3), 3h+g (2)	s (7), w (3), g (3)
	in (56)	1ax (3), 2@/2@x (20), 2n/2nx (12), 3@ (4), 3a (3)	w (9), ~ (3)
M1	sp (22)	3@ (10), 2@ (4), 2@x (3)	f (5), * (7)
	in (39)	1x (5), 1@x (3), 1n/1nx (7) 2@ (10), 3@x(6)	w (5)
M2	sp (10)	2@ (2), 3a+g (4)	s (4), g (4)
	in (9)	2e (3)	
M3	sp (7)	3a+g (2)	g (5)
	in (28)	1x (2), 1@x (2), 2@ (3) 3@+g (8), 3a (2), 4@+g (2)	g (12)
M4	sp (21)	1@x (2), 2@ (3), 2a (2) 2nx (2), 3@+g (3)	s (6), g (6)
	in (56)	1ax/1ex/1nx (6), 2@/2@x (7), 2n/2nx (18), 3@+g (6)	w (10), g (9), ~ (2)

Table 4: Typical pattern groups of spontaneous laughter and intentional laughter. (Number of occurrences are within brackets.)

In spontaneous laughter, call group 3 (4-9 calls) was found to have the largest variety of laughter patterns. 2@, 2@x, 3@, 3@x, 3a are the most typical examples. A common pattern can be observed among the spontaneous laughter of all the subjects, "a ha ha ha", in which the vowel /a/ tends to be centralized and becomes a schwa vowel (such as 3@ example in Fig.

3-3). On the contrary, in intentional laughter, call group 2 (2/3 calls) was found to have the largest variety of laughter patterns. Besides the patterns above, 1x, 2e, 2n, 2nx, and 3n are also commonly found in intentional laughter.

The rightmost column in Table 4 shows the number of occurrences of the prosodic and voice quality characteristics in Label 3, including strong (s), weak (w), ingressive laughter (g), and whole-voiced laugh (~). The strong voice is more predominant in spontaneous laughter, and the weak voice is more predominant in intentional laughter. However, weak spontaneous laughter and strong intentional laughter are also observed. Whole-voiced laugh (Fig. 3-5) is more observed in intentional laughter. Falsetto voice appeared only in M1, but all in spontaneous laughter. Regarding ingressive laughter (Fig. 3-6), no clear trends were observed between spontaneous and intentional laughter.

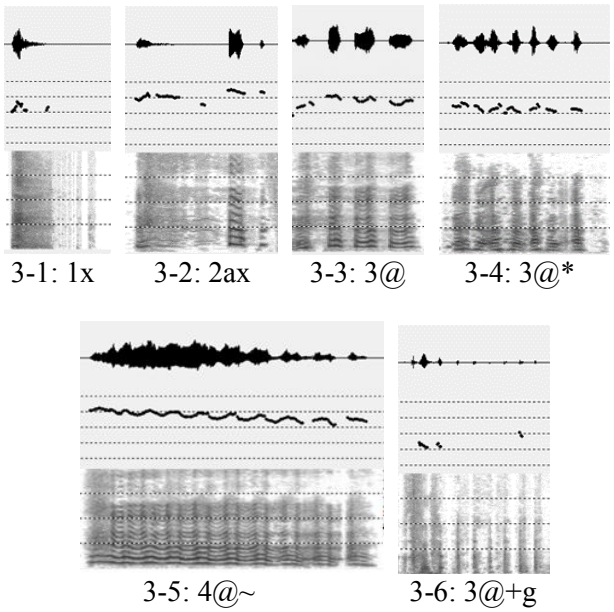


Figure 3: Examples of the representative laughter patterns. Waveform, F0 contour (4 octaves from 55Hz to 880Hz), and spectrogram (8kHz band).

3.3. Individuality of the speakers

The individuality of the speakers regarding laughter patterns can be observed in Table 4. For example, F1 tends to produce pressed voice in spontaneous laughter, and whole-voiced laugh in intentional laughter. M1 tends to use the falsetto voice (f) and pressed voice (*) in spontaneous laughter. F2 and F4 tend to produce patterns with the vowel /e/ and nasal sounds in intentional laughter. In particular, laughter uttered with nasal sounds has a female-like impression and is considered to have a polite effect. Regarding the usage of ingressive laughter (g), F1 and M1 did not use it at all, whereas M3 and M4 made extensive use of a series of ingressive laughter in both spontaneous laughter and intentional laughter. On the other hand, F2, F4, and M2 tend to produce ingressive laughter in spontaneous laughter.

4. CONCLUSION

In the present work, we analyzed the laughter patterns of spontaneous and intentional laughter types in multi-speaker conversation data. The analysis of laughter patterns focused on the number of calls, vowel-quality, and other voice quality characteristics of laughter. The outcomes showed the variation of the laughter depending on the speaker and the social situation.

The laughter with 2 or more calls showed a large number of vowel-quality variations, with the most frequently appearing schwa vowel. In the single-call laughter, breathiness is frequently used. Vowel-quality could change inside one laughter, but usually converges to the schwa or /a/ vowels.

Analysis results also revealed the relationship between laughter types and laughter patterns. Vowel /e/ and nasal sounds /n/ appeared with higher frequency in intentional laughter, whereas pressed voice in spontaneous laughter. A common pattern (“a ha ha ha ha”) was observed in spontaneous laughter among all the subjects, while intentional laughter was more individual-dependent and showed more variations in patterns, such as the usage of nasal sounds, ingressive laughter, and whole-voiced laugh.

Although some trends on the laughter pattern between spontaneous and intentional laughter could be observed, many overlaps in some of the laughter patterns were also observed depending on the speaker. Part of these overlaps are thought to be distinguishable from the visual modality including facial expressions and upper body motions (such visual information was used to classify laughter types, but the analysis of this paper focused on audio-only information). However, others may only be distinguishable from the conversation context. The laughter reason and level of the reason may also be factors to be considered. These are all topics for future investigation. Also, as the present study was limited to qualitative analysis of laughter patterns, our next step is to conduct acoustic analysis to quantitatively show the differences between laughter patterns.

5. ACKNOWLEDGEMENTS

This study was supported by JSPS KAKENHI Grants (No. JP22H04875, 21H04348, and 21K13048). We thank Yuka Nakayama, Naomi Uratani and Yumiko Kondo for contributing to annotations and analyses.

6. REFERENCES

[1] Provine, R. R. 2001. *Laughter: A scientific investigation.* Penguin Books.

[2] Giles, H., Oxford, G. S. 1970. Towards a multidimensional theory of laughter causation and its social implications. *Bulletin of the British Psychological Society*, 23(79), 97–105.

[3] Bachorowski, J.-A., Smoski, M.J., Owren, M.J. 2001. The acoustic features of human laughter. *J. Acoust. Soc. Am.* 111 (3), 1582-1597.

[4] Chafe, W. 2007. *The importance of not being earnest: The feeling behind laughter and humor.* John Benjamins Publishing.

[5] Kipper, S., Todt, D. 2003. The Role of Rhythm and Pitch in the Evaluation of Human Laughter. *Nonverbal Behavior*, 27, 255–272.

[6] Shochi, T., Guerry, M., Suzuki, H., Kanzaki, M. 2019. Vocal aspect of social laughter during virtual interaction. *Working Papers em Linguística*, 19(2), 54-77.

[7] Grammer, K., Eibl-Eibesfeldt, I. 1990. The ritualization of laughter. *Naturlichkeit der Sprache und der Kultur: Acta colloquii*, 192–214.

[8] Owren, M. J., Philipp, M., Vanman, E., Trivedi, N., Schulman, A., Bachorowski, J. A. 2013. Understanding spontaneous human laughter: The role of voicing in inducing positive emotion. *Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man*, 176-190.

[9] Ishi, C., Hatano, H., Hagita, N. 2014. Analysis of laughter events in real science classes by using multiple environment sensor data. *15th Annual Conference of the International Speech Communication Association* Singapore, 1043-1047.

[10] Tian, Y., Mazzocconi, C., Ginzburg, J. 2016. When do we laugh? *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* Los Angeles, 360-369.

[11] Inoue, K., Lala, D., Kawahara, T. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9, 155-160.

[12] Owren, M. J., Seyfarth, R. M., Cheney, D. L. 1997. The acoustic features of vowel-like grunt calls in chacma baboons (Papio cyncephalus ursinus): Implications for production processes and functions. *the Acoustical Society of America*, 101, 2951–2962.

[13] Trouvain, J. 2003. Segmenting phonetic units in laughter. *15th ICPhS* Barcelona, 2793–2796.

[14] Truong, K.P., Trouvain, J., Jansen, M. P. 2019. Towards an Annotation Scheme for Complex Laughter in Speech Corpora. *20th Interspeech* Graz, Austria, 529-533.

[15] Mori, H. Nagata, T. Arimoto, Y. 2019. Conversational and Social Laughter Synthesis with WaveNet. *20th Interspeech* Graz, 520–523.

[16] Szameitat, D. P., Darwin, C. J., Szameitat, A. J., Wildgruber, D., Alter, K. 2011. Formant characteristics of human laughter. *Voice*, 25, 32-37.

[17] Takehara, T. 2019. Emotional communication: facial expressions, voice, body contact, language, cultural differences. In: Uchiyama, I. (eds), Handbook of Psychology of Emotion. Kitaohji Syobo, 147-150.

Cross Cultural perception of Valence and Arousal

Erickson, D.[1], Rilliard, A.[2], Li, Y.[3], Menezes, C.[4], Kawahara, S.[5], Sadanobu, T.[6], Hayashi, R.[7], Shochi, T.[8] Moraes, J.[9], Obert, K.[10]

Haskins, CT. USA[1], Université Paris Saclay, CNRS, LISN, Universidade Federal do Rio de Janeiro, CNPq [2], Institute of Automation, Chinese Ac. Sciences, Beijing, Ch [3], Toledo University, Ohio, USA[4], Keio University, Jn[5], Kyoto University, Jn[6], Kobe Univeristy, Jn[7], Université Bordeaux Montaigne [8], Universidade Federal do Rio de Janeiro, CNPq [9], Voice Study Centre, University of Wales Trinity Saint David, Lampeter, UK [10]
EricksonDonna2000@gmail.com[1], Albert.Rilliard@limsi.fr[2], yongwei.li@nlpr.ia.ac.cn[3], Caroline.Menezes@utoledo.edu[4], kawahara.research@gmail.com[5], sadanobu.toshiyuki.3x@kyoto-u.ac.jp[6], rhayashi@kobe-u.ac.jp[7], takaaki.shochi@labri.fr[8] , jamoraes3@gmail.com[9],kerriebobert@gmail.com[10]

ABSTRACT

This paper reports on the perception of Valence and Arousal by listeners of different cultural backgrounds, trying to reproduce previous results. Listeners (105) from five language groups were presented with vowel sounds produced by one speaker with varying voice qualities; they had to rate each on a "Calm-Excited" (Arousal) and a "Positive Negative" (Valence) five-points scales. The sounds acoustic characteristics were analysed in terms of fundamental frequency, spectral tilt and open quotient. The results show variation across language groups in their use of these acoustic cues to produce their judgements. These differences also apply to results from previous experiments, that differ for the Valence scale, but not for the Arousal one. A multidimensional analysis also showed some tendency of in-group similarity in the way to answer on these two scales.

Keywords: Cross-cultural, Valence, Arousal, Acoustic analysis, Voice quality.

1. INTRODUCTION

The aim of this pilot study is to better understand how the perception of voice quality is affected by a speaker's language and culture. It is well-known that segmental characteristics of speech, i.e., consonants and vowels, are perceived differently depending on the language of the speaker. However, there is a growing body of research reporting that voice quality characteristics of speech may also be perceived differently, depending on the language/culture of the speaker, see e.g., [1]. One approach to examining these language/cultural differences is by using a Valence-Arousal-Dominance 3-Dimensional framework of emotions, e.g., [2]–[4]. In this paper, we look at Valence and Arousal percepts of vowel sounds produced with different voice qualities. We conducted a linear effect mixed model as well as a multifactorial analysis to ascertain which acoustic cues may be contributing to the Valence and Arousal percepts.

According to some of the previous research, voices that are loud, and high-pitched tend to be perceived as aroused/excited (e.g., [1], [4]–[8]). Voices that are breathy [9] or have a steep spectral slope [4] are perceived as having positive valence. Mandarin Chinese and Japanese listeners find high-pitched, less breathy voices pleasant [1] yet German listeners find them unpleasant [8]; Brazilian Portuguese listeners prefer lower pitched, breathy voices [1].

The languages we examine for this study are five: American English (AE), Japanese (JP), Mandarin Chinese (MC), Brazilian Portuguese (BP), and Goa Portuguese (GP), as spoken in Goa, India. The study includes results from [1] re BP listeners, but reports on new perception data collected using headphones from Japanese and Mandarin Chinese listeners, since the previous study collected the perception data in a classroom with loudspeakers.

2. METHODS

2.1. Corpus

Nine vowels (either /i/ or /æ/) were produced during an MRI study by an American English female speaker, with varying voice quality: the controlled changes were (1) the phonation mode (thick, thin, and stiff vocal folds, as described within the Estill model of voice production [10], which roughly correspond to modal, falsetto, and breathy, respectively), (2) vocal tract changes produced by (nasal or not) pharyngeal narrowing (resulting in a twang voice quality (see, e.g., [11]) or a lowered/backed bunching of the tongue, and (3) F0 changes (high or low F0) (see table 1 for details). Since the recordings were collected in an MRI session, which has a limited time requirement, it was not possible to do recordings of a complete set of configurations of voice settings.

2.2. Acoustic analysis

The model does not use inverse filtering, rather it fits the residual signal to the LF model to simultaneously estimate the glottal source waveform and the vocal tract shape parameters using an analysis-by-synthesis strategy (e.g., [12], [13]). The measurements are shown in Table 1.

2.3. Listeners

Five groups of listeners with different cultural backgrounds were recruited: (1) 18 Japanese students from Kobe University; (2) 21 Mandarin Chinese University students from Beijing; (3) 20 Brazilian Portuguese students from Federal University of Rio de Janeiro; (4) 25 American Students from Toledo University (Ohio); (5) 19 listeners from Goa (India), speakers of Goa Portuguese (mean age=66 years). The five language groups were selected due to both their representing diverse languages, and also the location of the authors.

Table 1: *Articulatory and acoustics characteristics of the stimuli: vowel (V), phonation type (Phonation), pharynx narrowed (PhN), presence of nasality (N), fundamental frequency (F0), spectral tilt, and open quotient (OQ).*

	V	Phonation	PhN	N	F0 (Hz)	Tilt	OQ
1	i	stiff (breathy)	-	-	Low 240	-16.9	0.63
2	i	thin (falsetto)	-	-	Low 312	-13.4	0.43
3	i	Thick (modal)	-	-	Low 230	-11.9	0.49
4	i	Stiff (breathy)	-	-	High 480	-11.4	0.43
5	i	Thin (falsetto)	-	-	High 500	-16.1	0.4
6	i	Thick (modal)	-	-	High 520	-10.5	0.4
7	ae	Thin (twang)	+	-	High 520	-13.4	0.34
8	i	Thin (tongue lowered/backed)	+	-	Low 240	-16.1	0.58
9	ae	Thin (twang)	+	+	High 520	-4.8	0.35

2.4. Paradigm

These nine vowels were presented to the five groups of listeners, individually, through headphones, and they had to judge on a scale of 1 to 5 how "Negative" or "Positive" the sound was (if they liked the sounds, it was positive), and how "Calm" or "Excited" the sound were (5 corresponding to "Excited"). There were four randomizations of each of the nine sounds, presented through a LiveCode computer interface.

2.5. Statistics

The perceptual ratings of each of the two scales were fitted using linear mixed-effects models (using R's lme4 library [14], [15]) – one for each language group and each scale. The dependent variable was the answer on each scale (Arousal and Valence), and the fixed effects were the acoustics characteristics of the stimuli, as indicated in Table 1; random intercepts for each listener were included in the model.

Then, the four answers given by each listener to each stimulus, for each of the two scales, were summed up and arranged into a table with one row for each subject, and one column for each stimulus and scale, thus having 105 rows x 18 columns (twice 9 stimuli) table. This table was submitted to a Multiple Factor Analysis (using R's FactoMineR library [16]) so to analyze the main dimensions linked to the way listeners answer to these two scales, according to the acoustic specificities of each stimulus. The spread of the listeners along the 6 first axes of the MFA was used as an input for hierarchical clustering. An inertia gain indicated a 3-cluster solution.

Table 2: *Summary of Linear mixed-effects model for Arousal (calm-excited) for 5 language groups (AE, BP, GP, JP, MC)*

Grp	Factor	Estimate	Std. Error	t value	P
AE	F0	0.275	0.044	6.2	0.000
	tilt	0.063	0.037	1.7	0.089
	OQ	-0.496	0.055	-9.1	0.000
PB	F0	0.094	0.060	1.6	0.121
	tilt	0.018	0.050	0.4	0.719
	OQ	-0.616	0.074	-8.3	0.000
GP	F0	0.317	0.063	5.0	0.000
	tilt	-0.004	0.052	-0.1	0.940
	OQ	-0.236	0.078	-3.0	0.000
JP	F0	0.461	0.048	9.7	0.000
	tilt	0.080	0.039	2.0	0.042
	OQ	-0.281	0.059	-4.8	0.000
MC	F0	0.141	0.051	2.8	0.006
	tilt	0.053	0.043	1.3	0.211
	OQ	-0.441	0.063	-7.0	0.000

Table 3: *Summary of Linear mixed-effects model for Valence (negative-positive) for 5 language groups (AE, BP, GP, JP, MC)*

Grp	Factor	Estimate	Std. Error	t value	P
AE	F0	0.341	0.052	6.5	0.000
	tilt	-0.109	0.044	-2.5	0.012
	OQ	-0.212	0.065	-3.3	0.001

PB	F0	-0.165	0.067	-2.4	0.015
	tilt	-0.305	0.056	-5.5	0.000
	OQ	0.117	0.083	1.4	0.159
GP	F0	-0.082	0.076	-1.1	0.283
	tilt	-0.036	0.063	-0.6	0.573
	OQ	0.007	0.093	0.1	0.940
JP	F0	0.384	0.058	6.6	0.000
	tilt	-0.068	0.048	-1.4	0.154
	OQ	-0.004	0.071	-0.1	0.953
MC	F0	0.101	0.063	1.6	0.109
	tilt	0.028	0.052	0.5	0.588
	OQ	-0.207	0.078	-2.7	0.008

3. RESULTS

3.1. Regression analysis

Tables 2 and 3 present the coefficients of the linear models fitted, for each language group, to the Arousal scale (table 2), and Valence scale (table 3). AE, GP, and MC listeners gave high Arousal ratings to stimuli with high F0 and small OQ, while PB listeners relied on OQ cues only, and JP listeners on high F0, low OQ and higher spectral tilt.

AE listeners tend to rate with a positive valence the stimuli with high pitch and low tilt and OQ; conversely, BP listeners rated as positive stimuli with low pitch and low tilt. GP do not show any systematic relation with acoustic characteristics for their valence judgments, while JP relied on high pitch only, and MC on low OQ only for attributing positive valence judgments.

Table 4: *Main association with the three clusters obtained from the spread of the listeners along the main dimensions of the MFA: language groups more frequent within each cluster; stimuli whose rating by listeners of the cluster are significantly higher (+) or lower (–) than the mean ratings. The numbers in the cells represent the utterance number.*

Cluster	#1	#2	#3
Group	AE, JP	BP	GP, MC
Val +	6, 7, 9, 15	1, 8, 2, 3	3
Val –	1, 8, 3, 2	6, 9, 7, 5	–
Aro +	7, 9	6, 9, 13	1, 8, 3
Aro –	3, 1	1, 4	9, 7, 6

3.2. Multiple Factor Analysis

The hierarchical agglomeration of individual listeners is shown on the dendrogram of Figure 2, from which was derived the three clusters. The characteristics of the stimuli (high positive or negative value given to a specific stimulus on a specific scale) significantly associated with each of these clusters are given in Table 4, with the language groups that are the most representative of each cluster (i.e., groups whose frequency within the cluster is significantly higher than the global frequency).

Figure 1 presents the spread of listeners along the first two main axes of the MFA: one can show that language groups show large overlaps, but that some tendency of listeners from the same cultural background to fit on the same portion of the plane may hint for some amount of culture-specific behavior. Thus, by agglomerating all listeners on the first 6 axes, we can observe the results on a more general level representing about 80% of the variance.

Figure 2 shows how the language groups cluster, based on these first six dimensions of the MFA. Cluster #1 is predominantly composed of AE and JP listeners, cluster #2 with BP listeners, and cluster #3 with GP and MC ones.

In terms of Valence, listeners from clusters #1 and #2 tend to give opposite judgments, with the most typical stimuli for this scale being opposed systematically: (Ut5,6,7,9) were judged positive by #1 and negative by #2; (Ut1,2,3,8) were judged the other way. Listeners in cluster #3 have few associations of stimuli with the valence scale: only Ut3 was judged as positive, as was also seen for cluster #2.

For Arousal, clusters #1 and #3 tend to be opposed in their judgments, with (Ut1,3) receiving high arousal by cluster #3 and low by cluster #1, and (Ut7,9) the other way. Listeners from cluster #2 present a mixed judgment, with some oppositions and some similarities with both of the other clusters.

An interpretation of the data in terms of acoustic cues triggering the perceptions is as follows: Re Valence, AE and JP indicate a positive feeling for high F0 vowel sounds (Ut 6, 7, 9 & 5), and negative feelings for low F0 vowel sounds (Ut 1, 8, 3, & 2). BP, however, show the opposite pattern, in that they prefer low F0 vowel sounds (Ut 1, 8, 2, & 3), not high F0 vowel sounds (Ut 6, 9, 7, and 5). GP and MC show a preference for low F0 vowel sounds produced with a modal somewhat breathy voice (Ut 3).

Re Arousal, all listeners rate low F0 breathy or somewhat breathy vowels as not aroused. And, all groups (AE, JP, BP, MC and GP) rate high F0 not breathy vowels produced with twang as aroused. In addition, BP rate vowels with sustained energy in the upper frequencies are rated as aroused. They rate low breathy vowels (Ut 1) as well as high F0 somewhat breathy vowels (Ut 4) as not aroused. It is interesting that BP rate Ut 3 as aroused, but the other language groups rate it as not aroused.

As for the MC and GP groups, according to the MFA, only moderate ratings are given for arousal by these language speakers. A possible interpretation of this is that for these speakers, these specific acoustic cues are not the salient ones for perceiving Arousal for these vowel sounds.

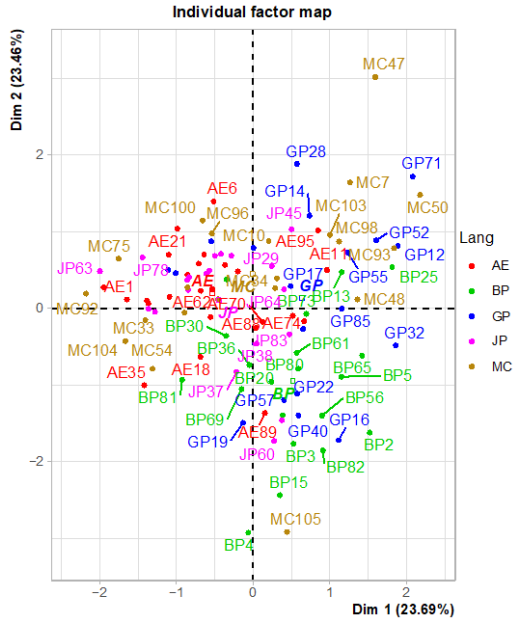


Figure 1: *Spread of the listeners on the first two dimensions of the MFA (about 45% of the variance), coloured by language group*

4. DISCUSSION & CONCLUSIONS

The preliminary results of this study confirm that listeners from different language groups perceive voice quality changes differently. We report an in-group similarity in the way to answer on these two scales, with listeners from each group being consistently clustered together on the basis of their answers.

A comment on OQ and spectral tilt: generally, when OQ increases, spectral tilt becomes steeper. Falsetto voice, often thought of a voice with high F0, generally has a large OQ and steep spectral tilt. However, adding pharyngeal narrowing to a falsetto voice changes the Speed Quotient (rate of increased contact of the vocal folds to decreased contact of the vocal folds per glottal cycle) [17], [18], as occurs when a singer/speaker produces a twang phonation. The abruptness of the vocal fold closure can affect the spectral tilt [19].

Future work is needed to further explore these cross-cultural perceptions of Valence and Arousal characteristics of voice quality in speech; specifically, we would like to include more acoustic measurements of voice quality, as well as more groups of language listeners. The five languages studied here represent a diverse group, and we would like to add more diversity.

Voice quality differences due to language/culture has of yet not been well-investigated; however, we feel this research has valuable applications for a number of professional areas, such as teaching in general, teaching second languages, clinical work, interpersonal relationships, professional speaking, advertisement, etc., basically any situation where we use our voices.

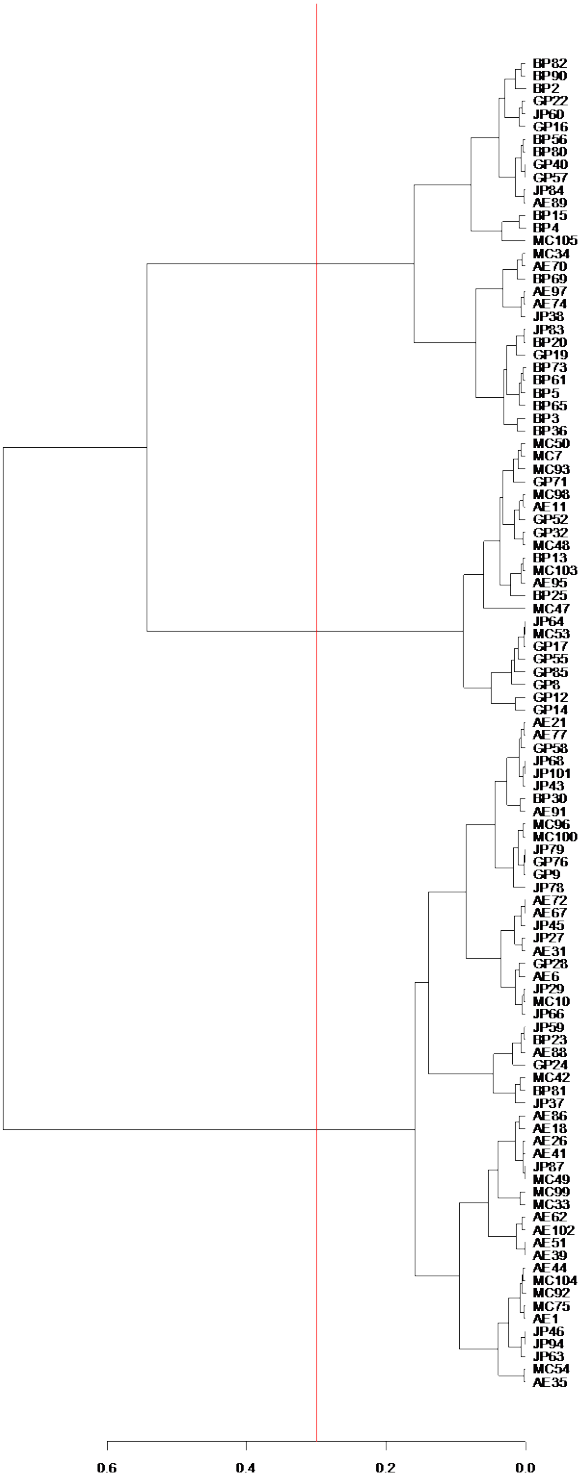


Figure 2: *Dendrogram obtained from the spread of listeners along the first six dimensions of the MFA; the red lines indicate the place where the tree has been cut, according to an inertia gain criterion.*

5. ACKNOWLEDGMENTS

We thank the subjects who participated in the perception tests. This study was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (S), 20H05630

6. REFERENCES

[1] D. Erickson *et al.*, 'Cross cultural differences in arousal and valence perceptions of voice quality', in *Speech Prosody 2020*, ISCA, May 2020, pp. 720–724. doi: 10.21437/SpeechProsody.2020-147.

[2] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975.

[3] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, 'The World of Emotions is not Two-Dimensional', *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007, doi: 10.1111/j.1467-9280.2007.02024.x.

[4] M. Goudbeek and K. Scherer, 'Beyond arousal: Valence and potency/control cues in the vocal expression of emotion', *J. Acoust. Soc. Am.*, vol. 128, no. 3, pp. 1322–1336, 2010, doi: 10.1121/1.3466853.

[5] P. N. Juslin and P. Laukka, 'Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.', *Emotion*, vol. 1, no. 4, pp. 381–412, 2001, doi: 10.1037/1528-3542.1.4.381.

[6] K. Scherer, 'Vocal communication of emotion: A review of research paradigms', *Speech Commun.*, vol. 40, no. 1–2, pp. 227–256, Apr. 2003, doi: 10.1016/S0167-6393(02)00084-5.

[7] T. Bänziger and K. R. Scherer, 'The role of intonation in emotional expressions', *Speech Commun.*, vol. 46, no. 3–4, pp. 252–267, Jul. 2005, doi: 10.1016/j.specom.2005.02.016.

[8] J. Schmidt, E. Janse, and O. Scharenborg, 'Perception of Emotion in Conversational Speech by Younger and Older Listeners', *Front. Psychol.*, vol. 7, May 2016, doi: 10.3389/fpsyg.2016.00781.

[9] A. Anikin, 'A Moan of Pleasure Should Be Breathy: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations', *Phonetica*, vol. 77, no. 5, pp. 327–349, Sep. 2020, doi: 10.1159/000504855.

[10] K. Steinhauer, M. McDonald Klimek, and J. Estill, *The Estill voice model: theory & translation*. Pittsburgh, Pennsylvania: Estill Voice International, 2017.

[11] K. Perta, Y. Bae, and K. Obert, 'A pilot investigation of twang quality using magnetic resonance imaging', *Logoped. Phoniatr. Vocol.*, vol. 46, no. 2, pp. 77–85, Apr. 2021, doi: 10.1080/14015439.2020.1757147.

[12] Y. Li, J. Li, and M. Akagi, 'Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space', *J. Acoust. Soc. Am.*, vol. 144, no. 2, pp. 908–916, Aug. 2018, doi: 10.1121/1.5051323.

[13] Y. Li, J. Tao, D. Erickson, B. Liu, and M. Akagi, 'F_0-Noise-Robust Glottal Source and Vocal Tract Analysis Based on ARX-LF Model', *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3375–3383, 2021, doi: 10.1109/TASLP.2021.3120585.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models Using **lme4**', *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[16] F. Husson, S. Lê, and J. Pagès, *Exploratory multivariate analysis by example using R*, Second edition. Boca Raton: CRC Press, 2017.

[17] D. Erickson, J. Yun, J. Gao, and K. Obert, 'Interaction between phonation mode and pharyngeal narrowing: A pilot EGG study', in *Proceedings of the 12th International Seminar on Speech Production*, M. Tiede, D. H. Whalen, and V. Gracco, Eds., New Haven, CT, USA: Haskins Press, 2020, pp. 190–193.

[18] D. Erickson, J. Yun, K. Obert, M. Reeve, H. Rowson, and K. Møller, 'Voice quality: Interactions among F0, vowel quality, phonation mode and phayngeal narrowing.', in *Nordic Prosody 13*, 2022.

[19] K. Stevens, 'Prosodic influences on glottal waveform: preliminary data', in *international symposium on Prosody, Yokohama, Japan*, 1994, pp. 53–64.

Prosodic word as the domain of breathiness: Evidence from disyllabic words in Kunshan Wu

Wenwei Xu[1], Peggy Mok[2]

[1]Leiden University, [2]The Chinese University of Hong Kong
w.xu@hum.leidenuniv.nl, peggymok@cuhk.edu.hk

ABSTRACT

This study provides new evidence for the claim that the domain of breathiness in Wu Chinese is a prosodic word instead of merely a syllable. Acoustic analysis of disyllabic words in Kunshan Wu shows that breathiness not only can pervade the first syllable but can also continue into the second syllable, provided that F0 is kept low. While previous research on the domain of breathiness in Wu Chinese has mostly focused on Shanghai Wu and only initial syllables, the results from Kunshan Wu, a more conservative dialect, suggest that the duration of breathiness within a prosodic word largely depends on F0 regardless of syllable boundaries. This accords with the fact that the domain of tone sandhi in Wu Chinese is usually a prosodic word, and that breathiness in Wu Chinese is a redundant cue for tone rather than a feature of syllable onset.

Keywords: phonation, voice quality, breathy voice, tone sandhi, Wu Chinese.

1. INTRODUCTION

Cross-linguistic research on the timing of non-modal phonation (e.g., breathy voice, creaky voice) has found both similarities and differences that are of theoretical importance to speech production and perception. In languages where non-modal phonation is typically a result of coarticulation from adjacent segments, its timing is generally restricted to a small portion of the vowel. Conversely, non-modal phonation that is phonologically specified in some languages can often last longer, or even pervade the vowel [1, 2, 3]. For example, breathiness introduced by /h/ or aspirated onsets in English shows only in the first half of the vowel [3], while the phonemic breathiness in Jalapa Mazatec [2, 4] and White Hmong [3, 5] can last over the entire vowel.

In the case of Wu Chinese, the timing of the phonologically specified breathy voice has been long debated. Early impressionistic accounts [6, 7] and experimental studies [8, 9] often attributed the breathiness to syllable onsets, while phonological proposals [10, 11] mostly favoured syllables as the domain of breathiness. Yip [12, 13] and Duanmu [14] extensively discussed this issue and proposed a register contrast for explaining the unique alternation of voicing and voice quality in (Shanghai) Wu, which has been widely accepted.

In Wu dialects, a syllable with a phonologically voiceless onset carries an upper register tone that is presumably modal and higher-pitched, while a syllable with a phonologically voiced onset carries a lower register tone that is usually breathier and lower-pitched. However, in a prosodic word, the phonation difference only surfaces in the initial syllable, while phonetic voicing (VOT < 0) of the voiced onsets only surfaces in non-initial syllables [8, 13, 14].

Such a phonological account has led to a common assumption, although never implying it, that the timing of breathiness should be somewhere within the initial syllable. The consequence is that recent phonetic studies, mostly on Shanghai Wu, all focused on monosyllabic words [15, 16, 17] or the first syllable of disyllabic words [15, 18]. The results showed that breathiness lasted for at least half of the vowel in monosyllables and for the entire first vowel in disyllables in Shanghai Wu [15, 16, 17], while in our previous study on Kunshan Wu (a closely related but more conservative dialect), breathiness did not vanish until the vowel offset in monosyllables carrying certain lexical tones [19].

Nevertheless, auditory impression on disyllabic or polysyllabic words in some Wu dialects also indicates that breathiness may still be perceptible in non-initial syllables, and it is also reasonable to speculate that breathiness can continue from the initial syllable into the following ones. The reasons are two-fold.

On the one hand, phonation is underspecified for non-initial syllables [12, 13], which implies that the phonation is not blocked or 'reset' at the boundary between the initial syllable and the following ones. On the other hand, within a prosodic word, a left-dominant tone sandhi is obligatory in Northern Wu dialects (including Shanghai and Kunshan Wu) [20], and the nature of such kind of tone sandhi is believed to be a process of rightward tonal spreading [21]. If pitch can spread across syllables, voice quality as a suprasegmental feature may also be able to spread.

Therefore, this study will leverage the acoustic data of disyllabic words in Kunshan Wu to show that breathiness can continue across syllable boundaries in certain conditions, and further argue that the domain of breathiness should not be the initial syllable only.

2. METHOD

2.1. Speakers

Sixteen native speakers of the downtown Kunshan Wu dialect were recruited, with age (middle-aged: 43–51; elderly: 67–73) and gender evenly distributed. All the speakers could also speak Mandarin, but none of them learnt it earlier than primary school.

2.2. Materials

The wordlist consisted of 128 disyllabic compounds and proper nouns, which were balanced for tone sandhi patterns (see Section 3.1) and voicing of the intervocalic consonants (voiceless unaspirated vs. voiced). While segments of the first syllable (S1) were not controlled, the second syllable (S2) was always composed of an obstruent onset (/p b t d k ɡ f v s z tɕ dʑ/) and an oral vowel (/a ɛ ɪ ɔ/).

2.3. Procedure and analysis

Each speaker was recorded for three repetitions of the wordlist with the Behringer ECM 8000 Ultra-Linear Measurement Condenser Microphone at their home.

Various acoustic correlates to pitch and voice quality were measured using VoiceSauce [22]. F0 was estimated using the STRAIGHT algorithm [23]. Spectral tilts were then calculated as the differences between the amplitude of selected harmonics (H1-H2, H2-H4, H4-H2K, H2K-H5K, H1-H4, H1-H2K, H1-H5K, H1-A1, H1-A2, H1-A3), which were corrected for formant influence [24] (correction marked by * henceforth). Noise measures included the harmonics-to-noise ratios (HNRs) as well as the Cepstral Peak Prominence (CPP).

For a total of 6136 valid tokens, each acoustic correlates were measured on ten equidistant points in the vowels of both S1 and S2, and standardized into z-scores within speakers. Measures other than F0 were excluded for syllabic fricatives (e.g., /ɣ̩/) in S1.

To investigate the time course of the measures, general additive mixed models (GAMMs) were fitted with the R packages *mgcv* [25, 26] and *itsadug* [27], following [28, 29]. Significance testing was performed using the difference curve method with 99.5% confidence intervals to reduce Type I errors. For visualization of the time courses, fitted values in the GAMMs were plotted in normalized time with adjustment for mean duration of each tone.

3. RESULTS

3.1. Tone sandhi and F0

The disyllabic tone sandhi is, as mentioned earlier, largely left-dominant in Kunshan Wu, which means that the sandhi form of S1 depends solely on the lexical tone of S1, while the sandhi form of S2 depends on the lexical tones of both S1 and S2. The complex mapping between S1 and S2 will not be covered here, but as the focus of this study is on S2, the disyllabic sandhi patterns are grouped by S2 and summarized in Table 1. The F0 contours are plotted accordingly in Figure 1 (top).

For S1, the sandhi register (a = upper and modal, b = lower and breathy) corresponds to that of the underlying lexical tone. The sandhi contour is rising (TS1) or falling (TS2) on an unchecked rime, or short (TS3) on a checked rime. For S2, the sandhi tones do not depend on the underlying tone register, and the sandhi contours can be classified into four groups of similar pitch contours.

3.2. Voice quality

Principal component analysis (PCA) was first performed on a total of sixteen acoustic correlates, (averaged over S1). S1 tones were well separated as two registers along the first principal component (PC1, with 39.2% variance explained), and the useful acoustic correlates ($r > 0.25$ with PC1) were, apart from F0, H1*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, and all the HNRs. Two measures, H1*-H4* and HNR25 (measured over 0–2500Hz), which showed the largest correlations with PC1, were chosen for modelling phonation differences in GAMMs. The interpretation of HNR25 requires some caution because both breathier and creakier phonation can cause more noise (lower periodicity).

3.2.1. S1

The time courses of the acoustic correlates for S1 tones are plotted in Figure 1 (left). Ten words with a nasal onset were excluded for H1*-H4* in case the spectral tilt is biased by nasality.

TS1b (red dashed) showed a significantly larger H1*-H4* than the upper register tones except near the offset of the vowel (two time points). HNR25 was significantly smaller in TS1b than in TS1a except at the vowel onset (two time points), while there is no significant difference between TS1b and TS2a.

As for TS2b (green dashed), H1*-H4* was significantly larger from that of the upper register tones throughout the vowel. HNR25 was significantly smaller in TS2b than in TS1a except at the vowel onset (two time points), while there is again no significant difference between TS2b and TS2a.

Given the extremely short duration of the checked tones (TS3a & TS3b), H1*-H4* and HNR25 were averaged over the entire vowel and plotted as triangular points in Figure 1. Both measures showed a significant difference ($p < 0.001$).

Possible S1	S2	Contour
TS1a/b, TS3a	A	high-falling
TS2a/b, TS3a	B	mid-level
TS2a	C	low-falling
TS2b, TS3b	D	low-dipping

Table 1: Disyllabic sandhi patterns in Kunshan Wu.

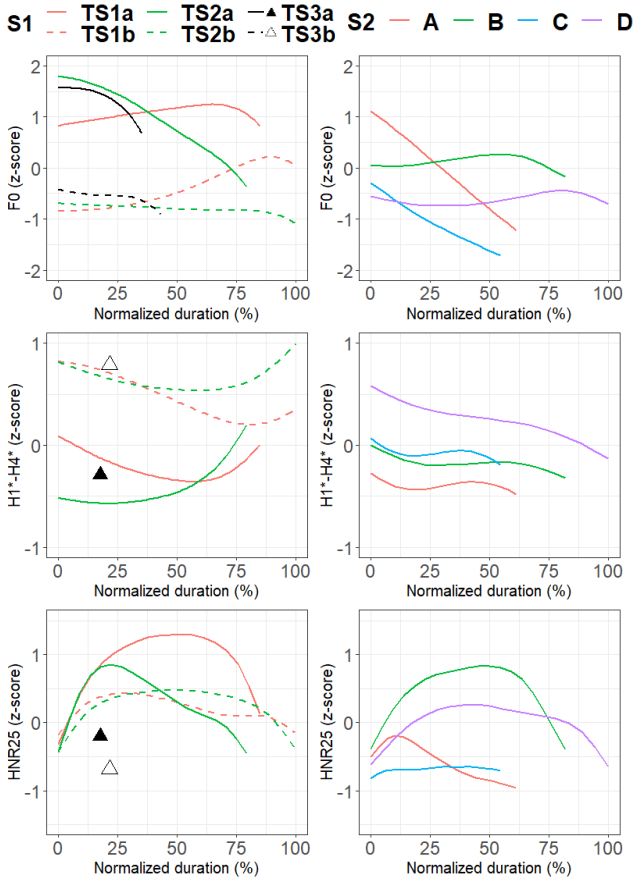


Figure 1: Time courses of F0, H1*-H4* and HNR25 (left = S1; right = S2).

3.2.2. S2

Once tone sandhi is applied, S2 loses its lexical tone identity, but the onset voicing still corresponds to the underlying register. However, no difference was found for H1*-H4* and HNR25 at all between S2 registers, which confirmed that S2 register could not initiate a phonation contrast as S1 register did.

Since the phonation of S2 did not depend on its register, the acoustic correlates were fitted by sandhi tones in S2, and the results are plotted in Figure 1 (right). Tone D (purple) showed a significantly larger H1*-H4* than the other tones for at least half of the vowel from the onset. HNR25 was significantly smaller in Tone D than in Tone B over the middle third of the vowel (discussion for Tone A and C in Section 4.1). Therefore, Tone D was breathier than the other tones.

3.2.3. S1 context for S2

As is shown in Table 1, the disyllabic sandhi is a many-to-many mapping between S1 and S2, which means the same S2 tone can be preceded by different S1 tones.

First, Tone A can be preceded by an upper register tone (TS1a or TS3a) or a lower register tone (TS1b). If breathiness in S1 can continue into S2, breathier phonation would be expected for Tone A when preceded by T1b. However, no acoustic difference was found. This is not surprising because the phonation difference between TS1a and TS1b already vanished before the offset of S1.

Similarly, Tone B can be preceded by an upper register tone (TS2a or TS3a) or a lower register tone (TS2b). Again, no acoustic difference was found between these two contexts, which contradicted our prediction because the phonation difference between TS2a and TS2b did not vanish before the offset of S1.

Finally, different from Tone A and B, Tone C can only be preceded by an upper register tone (TS2a), and Tone D is preceded exclusively by the lower register tones (TS2b or TS3b). Therefore, no further comparison was made.

4. DISCUSSION

4.1. Pitch-dependent voice quality

Voice quality is continuous and multidimensional in both production and perception, which means there is no precise boundary for deciding whether a tone is breathy or not. In this study, a breathy tone is defined in relative terms to a modal tone, for which TS1a is the best reference because as an upper register tone, it also showed the largest HNR25. In comparison with TS1a, the disyllabic sandhi tones are identified with distinct F0 and phonation patterns in Figure 1.

The first sandhi group is {TS1a/b + A} (red lines), which starts with an upper or a lower register tone and ends with a high-falling tone. Breathiness in TS1b is shown by a positive H1*-H4* and an HNR25 smaller than that of TS1a, which are both neutralized before the offset of S1. The phonation in S2 then becomes increasingly creakier as it approaches the pitch floor, which is clearly shown by a relatively low H1*-H4* and a remarkable fall in HNR25.

The second sandhi group is {TS2a/b + B} (green lines). It also starts with an upper or a lower register tone but ends with a mid-level tone. Breathiness in TS2b is shown by a positive H1*-H4* and a HNR25 smaller than that of TS1a, which are maintained at the offset of S1, but there is no phonation difference in S2 between the two S1 contexts. As Tone B generally shows a medial H1*-H4* (around z-score = 0) and

relatively high HNR25 (near z-score = 1), it can be considered modal.

The third group only contains one pattern, {TS2a + C} (green solid + blue), which starts with an upper register tone falling from the pitch ceiling and ends with a low-falling tone towards the pitch floor. The onset of TS2a features a lower H1*-H4* than TS1a, while HNR25 starts high but falls abruptly later. This means TS2a is initiated in regular high pitch with larger glottal constriction, which is typical of tense voice. As pitch drops and becomes more irregular, it turns into (prototypical) creaky voice [30, 31]. In S2, Tone C exhibits a medial H1*-H4* and a negative HNR25, implying considerable noise without too much glottal constriction. This corresponds to the properties of 'unconstricted creaky voice', a form of phrase-final creak found in other languages [30, 31].

The last sandhi group also contains one pattern, {TS2b + D} (green dashed + purple), which starts with a lower register tone (slightly falling) and ends with a low-dipping tone. Breathiness in TS2b, as mentioned above, is maintained at the offset of S1. Compared with TS1a and Tone B, Tone D shows a positive H1*-H4* and a lower HNR, so there is still breathiness in S2.

In summary, voice quality is closely dependent on pitch in Kunshan Wu. Tones are generally realized with tense voice when pitch is extremely high, and with breathy or creaky voice when pitch is low. What is interesting is that the speakers tend to use different strategies to maintain or reach a low pitch. The phonologically relevant low tones are produced with breathy voice, while for the low pitch in falling tones, they switch to creaky voice, similar to where creaky voice is found in Mandarin Chinese [32].

4.2. The domain of breathiness

As discussed above, although the lower register tones are all initiated with breathiness, it vanishes before the offset of TS1b, so only TS2b and TS3b has the potential for carrying over breathiness to S2. Since Tone D is still breathy for at least half of the vowel, such carry-over effect seems to exist. Meanwhile, Tone B is not breathier when preceded by TS2b than TS2a, so there is no carry-over of breathiness from TS2b to Tone B. This raises the question on conditions that permit such kind of continuation.

Comparing the time courses of F0 and H1*-H4* between TS1a and TS1b, and between Tone B and Tone D, it is evident that breathiness tends to vanish as soon as F0 is raised to somewhere around the speakers' mean F0 (z-score = 0). The consequence is that breathiness seems to only continue from TS2b into Tone D as the F0 is kept low, but not into Tone B as the F0 is raised. This pattern is consistent with what was previously found in monosyllabic words in Kunshan Wu [19], where breathiness also vanished earlier in the low-rising tone than in the low-dipping tone. Such covariation is believed to be a result of conflicting perceptual demands for both pitch and voice quality [1, 2, 19].

One may also consider the possibility that a low pitch can automatically cause breathiness in Kunshan Wu, but it is clearly not true because the low pitch in Tone A and C, sequenced after various phonation types, is realized with creaky voice. Therefore, breathiness found in Tone D should only be explained as the continuation of the phonologically motivated breathy voice in the preceding syllable.

In other words, the continuity of breathiness in a disyllabic word is not blocked by syllable boundaries, but constrained by the continuity of a low pitch. This accords with perception studies that have found breathiness to be a redundant cue for low tones in Wu dialects [33, 34]. Given the evidence that breathiness not only pervades the first syllable but also continues into the second syllable in a low-pitched disyllabic word, the domain of breathiness in Kunshan Wu can no longer be just the initial syllable, but should be a larger domain where the alternation of voicing and breathiness is applied. A common notion of such a domain is a prosodic word, where the left-dominant tone sandhi is also obligatory in Kunshan Wu. The domain for variation in pitch and voice quality is thus unified, in which the carry-over of breathiness is not striking, but nicely aligned with the rightward tonal spreading trend in this left-dominant sandhi system. A new question is thus raised whether breathiness also continues across syllable boundaries in Southern Wu dialects, where tone sandhi is mostly right-dominant instead, and further research is called for.

5. CONCLUSION

Acoustic analysis of disyllabic words in Kunshan Wu showed that breathiness not only can pervade the first syllable, but can also continue into the second syllable in some conditions, while the duration of such continuation is believed to be contingent on a continuous low pitch. In addition, the spreading of voice quality across syllable boundaries echoes the spreading of tone across syllables in the left-dominant sandhi, which constitutes evidence for the proposal that the domain of breathiness should be, for example, a prosodic word that is larger than merely a syllable, and unified with the domain of tone sandhi.

Acknowledgements: This study was supported by the RGC General Research Fund 2019/20 Project Reference 14607619 awarded to the second author.

6. REFERENCES

[1] D. Silverman, “Phasing and recoverability,” Ph.D. dissertation, UCLA, CA, 1995.

[2] B. Blankenship, “The timing of nonmodal phonation in vowels,” *J. Phon.*, vol. 30, pp. 163–191, 2002.

[3] M. Garellek, “The timing and sequencing of coarticulated non-modal phonation in English and White Hmong,” *J. Phon.*, vol. 40, no. 1, pp. 152–161, 2012.

[4] M. Garellek and P. Keating, “The acoustic consequences of phonation and tone interactions in Jalapa Mazatec,” *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 185–205, 2011.

[5] C. M. Esposito, “An acoustic and electroglottographic study of White Hmong tone and phonation,” *J. Phon.*, vol. 40, no. 3, pp. 466–476, 2012.

[6] Y. R. Chao, *Studies in the modern Wu-dialects*. Peking: Tsinghua College Research Institute, 1928.

[7] F. Liu, *Étude expérimentale sur les tons du chinois.* Paris: Soicété Les Belles Lettres, 1925.

[8] J. Cao and I. Maddieson, “An exploration of phonation types in Wu dialects of Chinese,” *J. Phon.*, vol. 20, pp. 77–92, 1992.

[9] N. Ren, “An acoustic study of Shanghai stops,” Ph.D. dissertation, Univ. of Connecticut, Storrs, CT, 1992.

[10] M. Sherard. “Shanghai phonology,” Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1972.

[11] P. Rose, “Phonetics and phonology of Yang tone: phonation types in Zhenhai,” *Cahiers de Linguistique-Asie Orientale*, vol. 18, no. 2, pp. 229–245, 1989.

[12] M. Yip, “The tonal phonology of Chinese,” Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[13] M. Yip, "Tonal register in East Asian languages," in *The Phonology of Tone: The Representation of Tonal Register*, H. van der Hulst and K. L. Snider, Eds., Berlin: Mouton de Gruyter, pp. 245–268, 1993.

[14] S. Duanmu, “A formal study of syllable, tone, stress and domain in Chinese languages,” Ph.D. dissertation, MIT, Cambridge, MA, 1990.

[15] J. Gao, “Interdependence between tones, segments and phonation types in Shanghai Chinese,” Ph.D. dissertation, Univ. Sorbonne Nouvelle–Paris 3, 2015.

[16] J. Gao and P. Hallé, “Phonetic and phonological properties of tones in Shanghai Chinese,” *Cahiers De Linguistique-Asie Orientale*, vol. 46, no. 1, pp. 1–31, 2017.

[17] J. Tian and J. Kuang, “The phonetic properties of the non-modal phonation in Shanghainese,” *J. Int. Phon. Assoc.*, pp. 1–27, 2019.

[18] Y. Chen, “How does phonology guide phonetics in segment– f0 interaction?” *J. Phon.*, vol. 39, no. 4, pp. 612–625, 2011.

[19] W. Xu and P. Mok, “The acoustic correlates and time span of the non-modal phonation in Kunshan Wu Chinese,” in *Proc. 12th ISCSLP*, 2021.

[20] E. Selkirk. and T. Shen, “Prosodic domains in Shanghai Chinese,” In *The phonology-syntax connection*, S. Inkelas and D. Zec, Eds., Chicago and London: Univ. Chicago Press, pp. 313–337, 1990.

[21] J. Zhang, “Tonal processes defined as tone sandhi,” in *The Cambridge Handbook of Chinese Linguistics*, C.-R. Huang, Y.-H. Lin, I.-H. Chen, and Y.-Y. Hsu, Eds. Cambridge: Cambridge Univ. Press, 2022, pp. 291–312.

[22] Y. Shue, P. Keating, C. Vicenik and K. Yu, “VoiceSauce: a program for voice analysis,” in *Proc. 17th ICPhS*, 2011, pp. 1846–1849.

[23] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction,” *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[24] M. Iseli, Y. Shue and A. Alwan, "Age, Sex, and Vowel Dependencies of Acoustic Measures Related to the Voice Source," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2283–2295, 2007.

[25] S. N. Wood, “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,” *J. Royal Stat. Soc.: Series B*, vol. 73, no. 1, pp. 3–36, 2011.

[26] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[27] J. van Rij, M. Wieling, R. H. Baayen and H. van Rijn. *itsadug: Interpreting time series and autocorrelated data using GAMMs*. (2022). [Online]. Available: https://cran.r-project.org/web/packages/itsadug/itsadug.pdf

[28] M. Wieling, “Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English,” *J. Phon.*, vol. 70, pp. 86–116, 2018.

[29] M. Sóskuthy, “Evaluating generalised additive mixed modelling strategies for dynamic speech analysis,” *J. Phon.*, vol. 84, pp. 101017, 2021.

[30] P. Keating, M. Garellek and J. Kreiman, “Acoustic properties of different kinds of creaky voice,” in *Proc. 18th ICPhS*, 2015.

[31] M. Garellek, “The phonetics of voice,” in *The Routledge Handbook of Phonetics*, W. F. Katz and P. F. Assmann, Eds. Abingdon, UK, pp. 75–106, 2019.

[32] J. Kuang, “Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice,” *J. Acoust. Soc. Am.*, vol. 142, no. 3, pp. 1693, 2017.

[33] J. Gao, P. Hallé, and C. Draxler, “Breathy voice and low-register: A case of trading relation in Shanghai Chinese tone perception?” *Lang. and Speech*, vol. 63, no. 3, pp. 582–607, 2020

[34] B. Jiang, M. Clayards and M. Sonderegger, “Individual and dialect differences in perceiving multiple cues: A tonal register contrast in two Chinese Wu dialects”, *Lab. Phon.*, vol. 11, no. 1, pp. 11, 2020.

A Perception Study on Voice Quality and Stance in Mandarin Chinese

Qiuyuan Li and Peggy Mok

Chinese University of Hong Kong
qiuyuan.li@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

The social meaning of voice quality has been gaining more attention in sociophonetics in recent years. However, voice quality and its social significance in Mandarin Chinese are still under-researched. This study conducts a perceptual experiment to explore the role of voice quality in stancetaking in Mandarin Chinese. The results show that listeners can make different value judgments about an object based on the different voice qualities used by the speaker to describe it and construct different portraits of the speaker. In Mandarin Chinese, creaky voice leads listeners to place lower values on the object described and to perceive speakers as more gender-neutral and less good-looking. The findings of this study may help investigate the role of voice quality from the production side and the construction of identities and expressing broader social meanings in Mandarin Chinese in the future.

Keywords: voice quality, Mandarin, stance, sociophonetics.

1. INTRODUCTION

Voice quality, by its narrower definition, refers to the status of vocal fold vibration in voiced segments. Voice quality is widely used to contrast meaning in some African languages. However, in several of the most popular world languages, such as English, Spanish, and Mandarin Chinese, voice quality does not function to contrast lexical meaning but rather to express specific social meanings. For example, in American English, young women who use creaky voice are perceived as valley girls, e.g. [1], while they are also perceived as more urban and more educated. In Japanese, [2] shows that Japanese female voice actors use a specific "sweet voice" quality to express the sweetness of the character's personality. However, although some previous studies have examined the social significance of voice quality in Mandarin, e.g. [3], which has found that the evaluation of creaky voice in Mandarin is associated with its prosodic context (IP-final creaky stimuli are rated as more enthusiastic and interested than other positions), it is insufficient in terms of both quantity and depth compared to the research on it in Western languages. Sociolinguistics has developed its third wave of theorizing in variation study ([4]), which shifts the focus of research from macrosocial categories and ethnic groups to individual speakers and subtle identity constructions, yet sociolinguistic research in Chinese is far behind, for example, no Chinese studies have focused on the role of phonetic features in stancetaking for the time being. Though stance has not been given a universal definition in linguistic research, and it may vary from the point of view that speakers want to express ([5]) to the identity they want to construct ([6]) using certain linguistic markers, the stance investigated in this study refers to the perception listeners have of the speakers and the object described by the speaker based on certain linguistic markers. In this study, participants' stance towards the narrator is revealed by their selection of the AI-generated faces, and their stance towards the object described in the narrative passage is indicated by their responses to the 3 statements about the cake.

Hence, this study investigates whether participants' backgrounds affect how they portray the speaker with different voice qualities and whether non-modal voice qualities may be interpreted to express different stances.

2. METHOD

An online experiment was conducted in which participants reported their backgrounds and listened to a short passage spoken in one of the three voice qualities (breathy, modal, and creaky), then rated against the object described by the passage and chose an AI-generated face that they thought would fit the voice the best. Details of the experimental design are described below.

2.1. Experiment design

The online experiment includes four stages, as shown below in Figure 1.

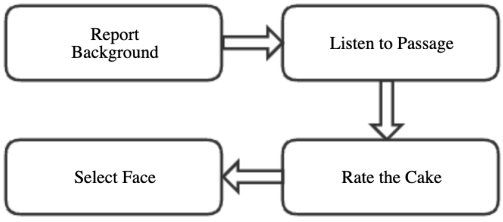


Figure 1: 4 stages in the experiment.

In the first stage, participants reported their backgrounds: gender, age, language spoken at home (Mandarin or not), and Mandarin proficiency.

Participants were then asked to listen to a pre-recorded short passage describing a strawberry cake in the second stage. They were told that a friend bought the cake as a birthday present for them, and the narrator of this passage was exactly that friend. The speaker's gender and the voice quality of this passage were randomized to be female or male, and breathy, modal, or creaky, respectively, while the content remained the same (information on speakers and their voices are to be given in later sections). Participants could listen to the passage as many times as they wanted until they felt comfortable moving on to the rating questions.

The third stage consisted of three statements about the cake described by the passage:

(i). The cake is quite delicious.

(ii). The cake is quite expensive.

(iii). The cake is too much as a gift from this friend and makes you stressed.

The first statement on deliciousness was about the internal evaluation that participants make towards the cake, while the second statement concerned an external and objective value, namely price. The third statement aimed to elicit participants' judgments on the relationship with the narrator.

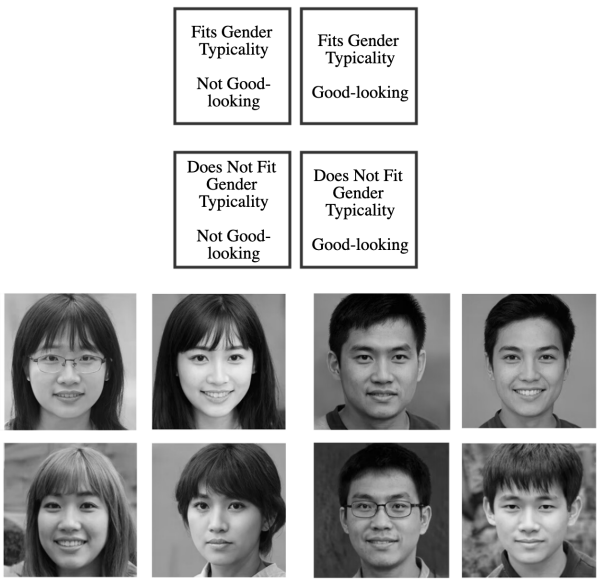


Figure 2: Arrangement of the 4 faces (top) and the actual faces used (female voice bottom-left; male voice bottom-right). Details are explained in section 2.2.

The rating questions were in the format of Likert scale from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree". Participants needed to select from 1 to 5 for each statement. This section was designed to investigate the role of voice quality on the perceived stance of the narrator towards the cake.

After rating each statement, participants were then asked to select 1 out of 4 AI-generated faces (details above) that they thought fit the impression of this "friend" the most.

As shown in Figure 2, the 4 faces were arranged in a 4-quadrant, representing a combination of gender typicality and degree of good-looking. The arrangement was fixed in this order. This section was designed to investigate the role of voice quality on participants' impressions of the speaker.

2.2. Stimuli

This experiment contained audio stimuli and visual stimuli.

There were 6 audio stimuli recorded for this experiment: 3 from a female speaker and 3 from a male speaker.

Both speakers were in their mid-20s and spoke Mandarin natively. The female speaker was a research assistant in the phonetics lab, and the male speaker was an amateur singer. They both have good control over the production of different voice qualities.

A few modifications were made to the audio recordings using a Praat plugin, Praat Vocal Toolkit ([7]). Firstly, the loudness of all sound files was adjusted to 60dB. Secondly, for the female recordings, the fundamental frequency median was adjusted to 190Hz, and that of male recordings was 110Hz. Finally, after noticing that the breathy and modal versions of female recordings have a larger pitch variation and might sound unnaturally dramatic, the pitch variation of these two recordings was reduced to 80% of the original, as it gave the most natural sounding intonation according to 4 naïve listeners.

Gender	Voice Quality	F0 Median	Pitch Variation	Loudness
F	Breathy	190Hz	80%	60dB
F	Modal	190Hz	80%	60dB
F	Creaky	190Hz	100%	60dB
M	Breathy	110Hz	100%	60dB
M	Modal	110Hz	100%	60dB
M	Creaky	110Hz	100%	60dB

Table 1: Parameters of processed audio stimuli.

The length of pauses is also modified to keep all recordings at roughly the same speech rate.

Quality	Gender	H1A1	CPP	HNR05
Breathy	F	14.65	14.98	24.75
Breathy	M	31.56	17.62	18.03
Modal	F	13.65	17.94	33.31
Modal	M	22.08	24.95	25.61
Creaky	F	12.66	19.34	22.07
Creaky	M	25.11	23.29	14.8

Table 2: Voice quality acoustics of stimuli.

We have also measured the voice quality acoustics of the stimuli and ensured that their patterns are typical to specific voice qualities as summarized in [8], as shown in Table 2.

The visual stimuli are human faces generated by [9], a group of face generators based on the Neuro-Network algorithm styleGAN2. The author of this generator provides tens of thousands of free and open-source faces generated by the algorithm, and the stimuli used in this experiment are selected from the dataset.

The stimuli have passed 2 rounds of selection. In the first round, about 50 faces were selected from the raw dataset of [5], making sure they were roughly at the same age and non-fake looking. In the second round, 7 people (4 females and 3 males; all at their mid-20s) were invited to rate the 50 faces on 2 dimensions: whether they fitted gender typicality (feminine for female faces and masculine for male faces), and whether they were good-looking.

After the second round, the 8 faces that received the most votes on each end of the dimensions were selected as the final set for the face-selection stage in the experiment. The 8 faces are demonstrated in Figure 2. Since these faces are all generated by AI from the same set of training data, they appear to be relatively homogeneous, however since the votes of these 7 people showed considerable consistency in both dimensions, the difference between these faces was detectable and could be noticed by the participants. All faces that were finally used as stimuli were evaluated as non-fake and natural.

2.2. Participants

After the experiment was published online, 262 participants completed the experiment. Among them, 148 were females, and 114 were males. Their ages range from 13 to 57 years old, and most are between 19 to 25 years old. 204 of the participants speak varieties of Mandarin at home, while other 58 speak some other languages such as Hmong or Wu Chinese.

As most participants fell between 19 to 25 years old and age was not the primary variable this study aims to investigate, participants younger than 17 and older than 29 years old were excluded to keep the distribution less complicated, giving us 249 final participants with 143 being females and 106 males.

3. RESULTS

The data obtained from the experiment include participant backgrounds, ratings of the cake, and faces they have selected.

3.1. Data Analysis

Participant backgrounds and the faces they have selected are coded into binary form to perform binomial logistic regression. The rating results are not transformed and are analyzed later separately.

Binomial logistic regression is performed to analyze whether the face's gender typicality can be predicted by the participant's gender, language background (Mandarin or not), speaker's gender, and the voice quality used in the passage. All independent variables are transformed into dummy variables. Voice quality is transformed into 2 dummy variables as there are 3 levels within it (breathy, modal, and creaky), and we would like to investigate each level's impact on stances. Modal voice is set to be the default and thus represented by "intercept" in the regression. The results are shown in Table 3.

	Coeffi-cient	SE	z-value	p-value
Intercept	0.75	0.44	1.71	0.0875
PG	-0.11	0.31	-0.35	0.7258
Language	0.33	0.36	0.92	0.3584
SG	1.39	0.33	4.23	**<0.0001**
Creaky	-1.11	0.38	-2.89	**0.0039**
Breathy	-0.53	0.37	-1.44	0.1496

Table 3: Results of binomial logistic regression against face's gender typicality. "PG" represents "participant's gender", and "SG" stands for "speaker's gender". P-values < 0.05 is emphasized in bold. Interactions are non-significant and omitted.

Another binomial logistic regression is performed to see if the good look of the selected face could be predicted by the participant's gender, language background (Mandarin or not), speaker's gender, and the voice quality used in the passage. The results are shown in Table 4.

	Coeffi-cient	SE	z-value	p-value
Intercept	1.25	0.42	2.97	**0.0030**
PG	-0.09	0.28	-0.31	0.7536
Language	0.04	0.33	0.13	0.8937
SG	-0.52	0.28	-1.85	0.0640
Creaky	-0.82	0.34	-2.42	**0.0157**
Breathy	0.05	0.34	0.14	0.8862

Table 4: Results of binomial logistic regression against face's good looks. "PG" represents "participant's gender", and "SG" stands for "speaker's gender". P-values < 0.05 is emphasized in bold. Interactions are non-significant and omitted.

The responses to the 3 statements about the cake are analyzed into frequencies of each rating and displayed in Figure 3.

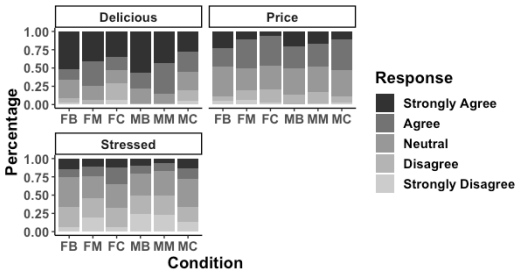


Figure 3: The percentage distribution of participants' responses to 3 statements about the cake after listening to the passage narrated by various gender and voice combinations. For example, "FB" represents "female breathy voice" and "MC" represents "male creaky voice".

As shown in Table 3, participant gender and language background are not related to the face's gender typicality. Keeping other factors the same, the odds of creaky voice being considered as fitting gender typicality is exp(-1.11) = 0.33 times of the odds of a non-creaky voice being considered as fitting gender typicality, which means that creaky voice is recognized as more gender-neutral to both genders. The regression also shows that the speaker's gender affects the gender typicality that participants select. This could be resulted from the pitch of audio stimuli; the female speaker had a deeper voice compared to average females, which could lead her voice being perceived as less feminine ([10]).

As to good looks, it is found that modal and creaky voices have significant impacts on the listener's judgment. While other variables are consistent, participants are exp(1.25) = 3.49 times more likely to select a good-looking face if they heard a modal voice than a non-modal voice. If they heard a creaky voice, on the other hand, the odds of selecting a good-looking face is exp(-0.82) = 0.44 times of hearing a non-creaky voice. It means that modal voice is perceived as more good-looking compared to non-modal voices, and creaky voice is less good-looking compared to modal and breathy voices.

Regarding the ratings of the statements about the cake, the most obvious trend shown in Figure 3 is that breathy voice gains the most positive judgments in deliciousness and price aspects while creaky voice gains the least. The voice quality does not affect how people perceive the cake as too much from such a friend and whether they feel stressed by the gift.

4. DISCUSSION

The fact that participants selected faces that fit gender typicality after hearing modal and breathy voices but not creaky voice is different from our expectation that, at least for male, creakiness indicate a certain degree of masculinity as it often co-occurs with low pitch, which naturally related to larger body size as claimed in [11], as well as previous studies in English, e.g. [12]. The difference between our results and anticipation agrees with the previous theory that gender is a performed sociocultural identity and is not fixed to biological sex [13]. When we view gender in terms of masculinity and femininity but not male and female, we are expressing the sociocultural ideologies about gender, which leads the genderization of linguistic forms [14]. In the current study, voice quality has been shown to be such a gendered linguistic form in Mandarin Chinese.

While gender typicality could be a rather neutral dimension, different stances perceived from these linguistic forms, or voice qualities, are much more subjective as they are reflected by ratings toward the cake and selection on the looks of the speakers. As to the stance toward the cake described in different voice qualities, the results have shown that the more spread the glottis is (which produces breathier voice), the higher the participants would rate the cake in terms of deliciousness and price. Deliciousness and expensiveness are two dimensions that reveal the stances that participants take toward the cake: breathier voice indirectly indexes more positive values in taste and price, and creaky voice indirectly indexes less positive values in taste and price. The looks of speakers, however, are thought to be the best while the voice is modal.

The limitations of this study are that the visual stimuli varied in only two dimensions in which the differences are not extremely obvious, and the experiment is only on the perception side so that it cannot allow us to observe the stances performed and exchanged between speakers like dialogues can do. Future studies may take on the results to design a production study and investigate the stance expressed by voice quality with a greater depth.

5. CONCLUSION

In this perception study, we have investigated the role of voice quality in perceiving stance in Mandarin. It is found that creaky voice will lead the listeners to construct an image of the speaker that is more gender-neutral and less good-looking, while modal voice is associated to more good-looking portrait. It is also found that the evaluation of an object can be affected by the voice quality of the narrator such that the more spread the glottis is, the higher the participants would rate the object in terms of subjective and objective values.

6. REFERENCES

[1] Habasque, Pierre. (2021). Is Creaky Voice a Valley Girl Feature? Stancetaking & Evolution of a Linguistic Stereotype. *Anglophonia.* 32. 10.4000/anglophonia.4104.

[2] Starr, R. L. (2015). Sweet voice: The role of voice quality in a Japanese feminine style. *Language in Society*, *44*(1), 1-34.

[3] Callier, P. R. (2013). *Linguistic context and the social meaning of voice quality variation*. Georgetown University.

[4] Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*(1), 87-100.

[5] Du Bois, John W. (2007). "The stance triangle". In Engebretson, Robert (ed.). *Stancetaking in Discourse: Subjectivity, evaluation, interaction*. Amsterdam: John Benjamins.pp. 139-182. doi:10.1075/pbns.164.07du. ISBN 9789027254085.

[6] Bucholtz, M., & Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. Discourse Studies, 7(4–5), 585–614. https://doi.org/10.1177/1461445605054407.

[7]Corretge, R. (2022). *Praat Vocal Toolkit.* https://www.praatvocaltoolkit.com.

[8] Wright, R., Mansfield, C., & Panfili, L. (2019). Voice quality types and uses in North American English. *Anglophonia. French Journal of English Linguistics*, (27).

[9] a312863063 (2022). *New version of face generators based on StyleGAN2* [electronic resource: python source code]. github.com/a312863063/generators-with-stylegan2.

[10] Krahé, B., Uhlmann, A., & Herzberg, M. (2021). The voice gives it away: Male and female pitch as a cue for gender stereotyping. *Social Psychology, 52*(2), 101–113. https://doi.org/10.1027/1864-9335/a000441

[11] Pisanski, K., Isenstein, S. G., Montano, K. J., O'Connor, J. J., & Feinberg, D. R. (2017). Low is large: spatial location and pitch interact in voice-based body size estimation. *Attention, Perception, & Psychophysics*, *79*(4), 1239-1251.

[12] Lee, K. E. (2016). *The Perception of Creaky Voice: Does Speaker Gender Affect our Judgments?* The University of Kentucky.

[13] Butler, J. (2002). *Gender trouble*. routledge.

[14]Kiesling, S. F. (2019). The 'Gay Voice'and 'Brospeak': Towards a systematic model of stance.

CONVERGENCE OF CREAKY VOICE USE IN AUSTRALIAN ENGLISH

Hannah White, Joshua Penney, Andy Gibson, Anita Szakay, Felicity Cox

Centre for Language Sciences, Department of Linguistics, Macquarie University
hannah.white2@hdr.mq.edu.au, joshua.penney; andy.gibson; anita.szakay; felicity.cox@mq.edu.au

ABSTRACT

Studies of phonetic accommodation have shown that speakers imitate features of their interlocutors' speech such as vowel production, voice onset time and fundamental frequency. While some individual acoustic parameters of voice quality have been examined in accommodation research, a global measure of creaky voice quality accommodation has not yet been investigated. Creak is a voice quality that appears to be increasingly prevalent in many varieties of English including Australian English and carries several social and linguistic meanings. The present study investigated whether 30 pairs of teenagers from Sydney converged in their use of creaky voice across the course of a conversation. Creaky voice was identified using automatic methods of detection. Findings suggest that creaky voice convergence was present among these speakers. Additionally, an effect of speaker sex showed females used creak at more similar rates to each other than males.

Keywords: Phonetic accommodation, creaky voice, voice quality, Australian English.

1. INTRODUCTION

It has been established that speakers imitate phonetic features of their (assumed) interlocutors in both conversational and laboratory settings and with both natural and synthetic speech in a process referred to as phonetic accommodation [2, 3, 5, 12, 24, 25]. Growing evidence suggests that accommodation is an unconscious but not automatic process influenced by many social factors [2, 3, 13, 14, 25]. Communication Accommodation Theory (CAT), as described in [13], proposes that speakers express their attitudes towards one another through strategies of convergence (their speech becomes more similar) and divergence (their speech becomes less similar). Babel [2] found evidence supporting CAT, showing that for target vowels, New Zealand English (NZE) speakers accommodated to Australian English (AusE) speakers in a shadowing task. She also found that the extent of convergence was influenced by the sentiments the NZE-speaker held about Australians. Those with more positive sentiments converged more than those with negative sentiments [2]. Dialect has also been shown to influence degree of accommodation with speakers converging more to accents different from their own [34]. Another shadowing task found that the degree of liking (measured through attractiveness ratings) influenced how much listeners converged/diverged their vowel qualities to/from model speakers in American English (AmE) [3]. Studies using more natural speech contexts have shown similar results [23, 25, 26]. Pardo [23] found that in map tasks, same-sex pairs of AmE-speakers converged in their vowel productions across the course of the conversation and that degree of convergence was influenced by speaker sex and role in the map task: female instruction givers converged more towards female instruction receivers while male instruction receivers converged more towards male instruction givers. Overall, Pardo [23] found that males converged more than females; however, findings regarding speaker sex and convergence are mixed. Studies have found that females converge more than males in certain conditions [5, 22], that females are more sensitive to lexical factors when converging [26], and that females adapt their speech more freely according to social factors than males [5].

Several phonetic features have been examined with regard to accommodation such as vowel quality [2, 3, 25], voice onset time [33], fundamental frequency (f0) [4] and pause and gap length [11]. The few studies that have examined voice quality accommodation have focused on individual acoustic voice quality parameters such as jitter and shimmer or open quotient rather than more global occurrences of different voice qualities such as creaky voice or breathy voice [18, 32]. Levitan and Hirschberg [18] investigated convergence at the conversation and the turn level in speakers of AmE by measuring the prosodic dimensions of mean and maximum intensity, mean and maximum pitch, jitter, shimmer and noise-to-harmonics ratio (NHR). At the conversation level, speakers' rates of mean intensity, shimmer and NHR became significantly more similar across the course of the conversation. At the turn level, only maximum pitch and mean pitch showed significant levels of convergence. Schweitzer et al. [32] measured

voice quality parameters that provide some detail of the glottal cycle, such as open quotient and glottal opening, to investigate turn-level convergence in conversations of female German-speaking pairs. Unlike Levitan and Hirschberg [18], Schweitzer et al. [32] found that most voice quality measures exhibited convergence. In addition, speakers' impressions of their interlocutor's competency and likeability influenced convergence on some parameters.

Creaky voice quality has been documented to varying degrees in many varieties of English, including AmE [27, 29, 37], British English [15] and AusE [7, 19, 36]. It has been labelled as a marker of phrase-finality [15, 27, 29] and is associated with mixed social functions such as sounding educated and casual [37] and less trustworthy and less competent than modal voice [1]. To our knowledge, research is yet to investigate any voice quality accommodation in AusE.

2. AIMS

The present study aims to investigate creaky voice accommodation in male and female same-sex conversations between teenage speakers of AusE. Creaky voice can be realised through a wide range of acoustic parameters and different combinations thereof [16]. While previous studies have investigated acoustic voice quality measures independently of each other, this study will provide insight into creaky voice accommodation as a global voice quality. This is particularly relevant given the suggestion that creak prevalence is continuing to increase in varieties of English [8]. We also investigate whether there are sex differences in creaky voice accommodation.

3. METHODS

3.1. Data

A total of 30 recordings ranging in length from 9.4 – 36.6 minutes were included in the analysis. These recordings were collected as part of a study of Multicultural AusE from various schools throughout Sydney. Speakers were drawn from a variety of areas differing in linguistic diversity and socioeconomic status; however, these factors are not examined in this paper. Recordings consist of conversations between same-sex pairs (16f; 14m) of students aged 15 – 16 years facilitated by a research assistant (RA). Pairs of students attended the same school and were familiar with each other prior to the conversation. All speakers involved in the conversation including the RAs were native speakers of AusE.

Conversations were orthographically transcribed using the Speech to Text API from IBM Watson, which provides automatic time-aligned orthographic transcriptions using an Australian model diarised for each speaker. The orthographic transcriptions of the teenagers were corrected, ensuring that the content of the speech was faithful to the original recordings and also that the speech was assigned to the correct speaker. Speech from the RA was not corrected. Recordings were then processed through WebMAUS to generate textgrids containing segmented and aligned phonemes [31].

Creaky voice was identified in the speech of the teenage speakers using a combination of two automated approaches described in White et al. [36] as the optimised Union method. This method consists of combining the outputs of the AntiMode (AM) method, an f0-based approach to detect creak [7, 9], and a Creak Detector (CD) algorithm, which makes decisions on the presence of creak based on acoustic cues including H2-H1 and residual peak prominence [10]. As different phonetic realisations of creak can be characterised by various combinations of acoustic cues and each tool employs different acoustic parameters, it is likely that the AM method would identify instances of creak that the CD algorithm would miss and vice versa. Steps can be taken to further improve creak detection such as optimising the automatic Creak Detection algorithm, which involves manually annotating a subset of data (also shown in [21] and [35]) and by limiting the analysis to sonorant segments only (including vowels, liquids, nasals and glides) [35, 36]. White et al. [36] showed that the optimised Union method outperformed the use of AM or CD alone. This (mostly) automatic process was carried out on the present data, the output of which is a binary creak decision for every 10 ms of input audio.

3.2. Analysis

To analyse whether speakers converged in their use of creaky voice across a conversation, each conversation was divided evenly into thirds. Dividing the conversations into thirds meant that the amount of speech from each speaker in each bin was more even compared to when the conversations were divided into quarters or fifths (mean ratios of speech from each speaker were close to 0 in all bins). This reduces the likelihood that one speaker dominates in any bin.

Creaky voice prevalence was then calculated per speaker per bin. Creak prevalence was determined as the number of 10 ms intervals spent in creak during sonorant segments divided by the total number of 10 ms sonorant segment intervals used by that speaker.

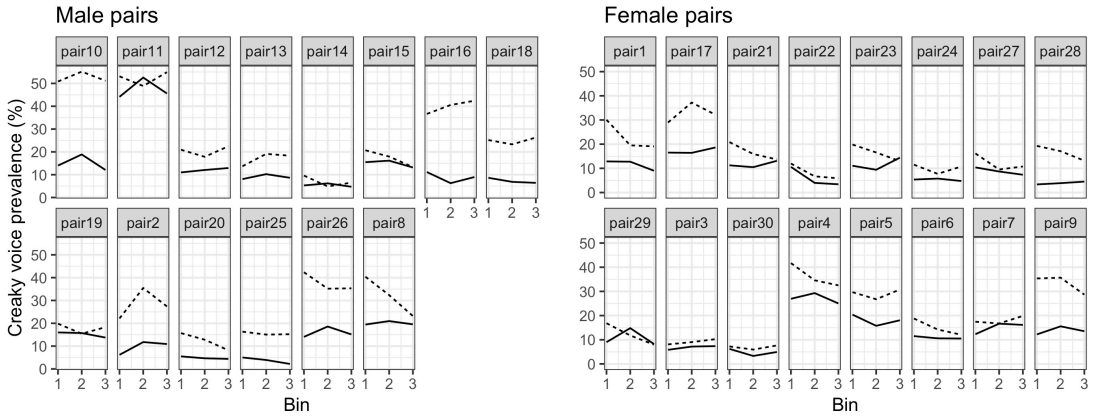


Figure 1: Raw data showing prevalence of creaky voice across thirds of the conversations (i.e., bins). Dotted line = creakier speaker. Male pairs are shown on the left and female pairs are on the right.

For each bin, the difference between each speaker's creak prevalence was calculated. The closer a value was to zero, the more similar the speakers' rates of creaky voice in that third of the conversation.

A linear mixed effects regression (LMER) model was run using the lme4 [6] and lmerTest packages [17] in R [28]. First, a maximal model was built with the difference in creak prevalence as the dependent variable. Independent variables were the third of the conversation (i.e., bin number, treated as a continuous variable henceforth called "bin"), speaker sex (pairs were same-sex), the proportion of the conversation that the RA spoke, the interaction of these three factors, the conversation length and the ratio of speech from each speaker in each bin. Pair was included as a random intercept. All continuous variables (bin, RA proportion, conversation length and speaker ratio) were transformed to z-scores. The data in this study was not collected specifically for the purpose of analysing accommodation. Hence, in each conversation the RA had different levels of engagement with the teenage speakers. Raw means and standard deviations show that, overall, RAs spoke more in the first bin than in the second and third bins and there is considerable variation between conversations (bin 1: mean=21.8% of conversation was RA, sd=9.8; bin 2: mean=16.8%, sd=10.2; bin 3: mean=16.2%, sd=9.4). We proposed that the more an RA talked in a conversation, the greater likelihood that they may have influenced accommodation of the teenagers. We included the proportion of the conversation that the RA spent talking in the model as a way of mitigating their possible influence on creaky voice accommodation. In order to balance model power and Type I error rate, we followed a stepwise elimination process, reducing non-significant interactions followed by non-significant fixed effects, removing terms with highest p-values, and comparing models with ANOVAs to ensure that the removal of each term did not significantly worsen the model [20].

4. RESULTS

Figure 1 shows the raw creaky voice prevalence for each speaker by pair and speaker sex. There appears to be a general trend for speakers to become less creaky over the course of a conversation, as well as evidence of convergence for many of the pairs. This initial impression was tested using LMER modelling.

Our final model included the significant simple fixed effects of bin and gender (with no significant interaction) and the random intercept for pair. The proportion of RA speech, conversation length and the ratio of speech from each speaker in each bin were not significant predictors. The model output is presented in Table 1.

	Est.	Std. error	df	t value	p-value
(Int.)	0.81	0.18	41.9	4.46	<0.001***
bin	-0.10	0.04	60	-2.83	0.006**
sex	0.53	0.25	30	2.15	0.040*

Table 1: Model output showing simple effects of bin and speaker sex on creaky voice convergence. Pair is included as a random intercept. Male is the reference level for sex.

The effect of bin (Figure 2) shows that over the course of a conversation, speakers become more similar to each other in their creaky voice usage. Although the interaction between bin and speaker

sex was not significant, there was a significant simple effect of speaker sex shown in Figure 3. Overall, females were more similar to each other than males.

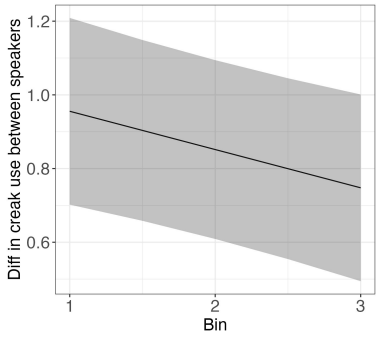


Figure 2: LMER model estimates of the difference in creaky voice prevalence between speakers in a pair across conversations. Values closer to zero indicate convergence.

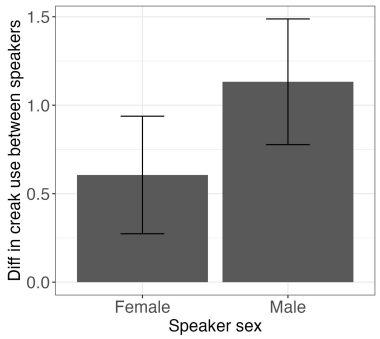


Figure 3: LMER model estimates of the difference in creaky voice prevalence between speakers in a pair for females and males. Values closer to zero indicate more similar creak levels.

5. DISCUSSION

This study explored accommodation of creaky voice in same-sex conversations of teenager speakers of AusE. Results support other studies showing that speakers do tend to phonetically converge towards each other [2, 3, 4, 11, 18, 25, 32, 33]. Unlike other studies of accommodation [5, 22, 23, 26], the present study did not find an interaction of convergence with speaker sex. However, a main effect of sex shows that female pairs were more similar to each other in their levels of creaky voice throughout the conversation than males. In our data, there is greater variance in creaky voice prevalence in males than females (males: mean=19.9%, sd=13.9; females: mean=14.9%, sd=8.45). It is possible that the sex effect is a reflection of this variance; males have more variable creak prevalence, increasing the likelihood that levels of creaky voice use will be more variable between male pairs compared to female pairs. An alternative explanation highlighted by Babel [5] draws on psychology literature indicating that females show more in-group bias than males [30], suggesting that females might be more motivated to accommodate to other females for social reasons.

The present study is limited by the presence of the RA in the conversations. Despite the amount of speech contributed by the RA not significantly influencing creaky voice accommodation, future research should be carried out on data without an RA as it is likely that aspects of the RA's presence other than amount of speech contribute to their influence on accommodation (e.g., RA creak prevalence or attitudes towards the RA). Additionally, it is unknown how participants in this study felt towards their interlocutor. Research has shown that social preferences influence accommodation [2, 3, 5, 25, 32] so future work could investigate whether the same effects exist for creaky voice accommodation and whether this interacts with speaker sex.

6. ACKNOWLEDGEMENTS

This research was supported by a Macquarie University Research Excellence Scholarship to the first author and by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the fifth author.

7. REFERENCES

[1] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatachalam, "Vocal fry may undermine the success of young women in the labor market," *PLoS ONE*, vol. 9, no. 5, p. e97506, 2014.

[2] M. Babel, "Dialect divergence and convergence in New Zealand English," *Language in Soc.*, vol. 39, no. 4, pp. 437–456, 2010.

[3] ——, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phonetics*, vol. 40, no. 1, pp. 177–189, 2012.

[4] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and Speech*, vol. 55, no. 2, pp. 231–248, 2012.

[5] M. Babel, G. McGuire, S. Walters, and A. Nicholls, "Novelty and social preference in phonetic accommodation," *Laboratory Phonology*, vol. 5, no. 1, pp. 123–150, 2014.

[6] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Statistical Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[7] K. Dallaston and G. Docherty, "Estimating the prevalence of creaky voice: A fundamental

frequency-based approach," in *Proc. 19th ICPhS*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. ASSTA Inc, 2019, pp. 581.1–5.
[8] ——, "The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature," *PloS ONE*, vol. 15, no. 3, p. e0229960, 2020.
[9] K. Dorreen, "Fundamental frequency distributions of bilingual speakers in forensic speaker comparison," Master's thesis, University of Canterbury, Christchurch, NZ, 2017.
[10] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Comput. Speech and Language*, vol. 28, no. 5, pp. 1233–1253, 2014.
[11] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proc. INTERSPEECH*, 2009, pp. 2779–2782.
[12] I. Gessinger, E. Raveh, I. Steiner, and B. Möbius, "Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing," *Speech Communication*, vol. 127, pp. 43–63, 2021.
[13] H. Giles and T. Ogay, "Communication accommodation theory," in *Explaining communication: Contemporary theories and exemplars*, B. Whaley and W. Samter, Eds. Lawrence Erlbaum, 2006, pp. 293–310.
[14] H. Giles, D. M. Taylor, and R. Bourhis, "Towards a theory of interpersonal accommodation through language: Some Canadian data," *Language in Soc.*, vol. 2, no. 2, pp. 177–192, 1973.
[15] C. Henton and A. Bladon, "Creak as a sociophonetic marker," in *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, L. M. Hyman and C. N. Li, Eds. Routledge, 1988, pp. 3–29.
[16] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proc. 18th ICPhS*. the University of Glasgow, 2015, pp. 821.1–5.
[17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *J. Statistical Softw.*, vol. 82, no. 13, pp. 1–26, 2017.
[18] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. INTERSPEECH*, 2011, pp. 3081–3084.
[19] D. Loakes and A. Gregory, "Voice quality in Australian English," *JASA Express Letters*, vol. 2, no. 8, p. 085201, 2022.
[20] H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates, "Balancing Type I error and power in linear mixed models," *J. Memory and Language*, vol. 94, pp. 305–315, 2017.
[21] O. Murton, S. Shattuck-Hufnagel, J.-Y. Choi, and D. D. Mehta, "Identifying a creak probability threshold for an irregular pitch period detection algorithm," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. EL379–EL385, 2019.
[22] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender differences in vocal accommodation: The role of perception," *J. Language and Social Psychology*, vol. 21, no. 4, pp. 422–432, 2002.
[23] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2382–2393, 2006.
[24] ——, "Measuring phonetic convergence in speech production," *Frontiers in Pyschology*, vol. 4, p. 559, 2013.
[25] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, "Phonetic convergence in college roommates," *J. Phonetics*, vol. 40, no. 1, pp. 190–197, 2012.
[26] J. S. Pardo, A. Urmanche, S. Wilman, J. Wiener, N. Mason, K. Francis, and M. Ward, "A comparison of phonetic convergence in conversational interaction and speech shadowing," *J. Phonetics*, vol. 69, pp. 1–11, 2018.
[27] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," *Ann. Meet. Berkeley Linguistics Soc.*, vol. 37, no. 1, pp. 427–448, 2013.
[28] R Core Team, "R: A language and environment for statistical computing," 2021. [Online]. Available: https://www.R-project.org/
[29] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *J. Phonetics*, vol. 29, no. 4, pp. 407–429, 2001.
[30] L. A. Rudman and S. A. Goodwin, "Gender differences in automatic in-group bias: Why do women like women more than men like men?" *J. Personality and Social Psychology*, vol. 87, no. 4, pp. 494–509, 2004.
[31] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using MAUS technique," in *New tools and methods for very-large-scale phonetics research workshop*, Philadelphia, PA, USA, 2011.
[32] A. Schweitzer, W. Wokurek, and M. Pützer, "Convergence of harmonic voice quality parameters in spontaneous dialogues," in *Proc. 19th ICPhS*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. ASSTA Inc., 2019, pp. 363–367.
[33] K. Shockley, L. Sabadini, and C. A. Fowler, "Imitation in shadowing words," *Perception and Psychophysics*, vol. 66, pp. 422–429, 2004.
[34] A. Walker and K. Campbell-Kibler, "Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task," *Frontiers in Psychology*, vol. 6, 2015.
[35] H. White, J. Penney, A. Gibson, A. Szakay, and F. Cox, "Optimizing an automatic creaky voice detection method for Australian English-speaking females," in *Proc. INTERSPEECH*, 2021, pp. 1384–1388.
[36] ——, "Evaluating automatic creaky voice detection methods," *J. Acoust. Soc. Am.*, vol. 152, no. 3, pp. 1476–1486, 2022.
[37] I. P. Yuasa, "Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

THE ROLE OF CREAKY VOICE IN CANTONESE TONAL PRODUCTION

Chu-Yan Ho

City University of Hong Kong; The University of Hong Kong
chuyanho2-c@my.cityu.edu.hk

ABSTRACT

Cantonese is a tonal language in which its low tones are reported to be produced with creaky voice inconsistently. The presence of the allophonic creaky voice in Cantonese has driven two hypotheses regarding its source: it is (1) a specific cue for Tone 4 identification or (2) a covariate when producing a low F0 [1]. To test whether creaky voice is tone-specific, native speakers were asked to produce Cantonese lexical tones under normal condition and low-pitch condition. Results revealed a correlation between creaky voice and low F0, and between creaky voice and prosodic positions. When participants were asked to speak in a low-pitch condition, creak occurrence increased significantly for tones other than Tone 4. The results shows that the source of Cantonese creaky voice is likely to be triggered by low F0, and is less likely to be phonemically tied to Tone 4.

Keywords: creaky voice, lexical tones, F0, Cantonese, tone production.

1. INTRODUCTION

Tonal languages use pitch contour and height as the primary acoustic cues for lexical tone identification, while register language refers to languages where phonation is lexically contrastive. However, the distinction has become less clear with the evidence that some tonal languages utilise both pitch and phonation for tone identification, such as Mandarin [2,3], Cantonese [1,4] and Hmong [5]. Notwithstanding the co-occurrence of phonation and pitch for lexical contrasts, tonal languages differ from register languages in that the former mainly relies on pitch for perceptual differences in lexical contrasts and phonation is only adopted as a secondary cue.

Traditionally, Hong Kong Cantonese (HKC) is studied as a tonal language where pitch has been assumed to be the main, if not the only, cue for tonal contrast. It is generally recognized that there are six contrastive tons with three entering tones, which are allotones of the level tones in HKC (See Table 1 for tone descriptions and Figure 2 for tone trajectories). Yu and Lam were the first to identify the relatively high occurrence of creaky voice in the low-falling Tone 4 (24.6%) in HKC and investigate the use of creaky correlates for the identification of T4-T6 pair [1]. They found an enhanced Tone 4 identification when Tone 4 was realized with creaky voice and that creaky voice biased participants toward Tone 4 even with contextual pitch cues. Zhang and Kirby disagreed on the extent of the effect of creaky voice in Tone 4 identification and questioned Yu and Lam's methodology as their manipulated stimuli may belong to double pulsing creak with extra low F0, thus directing the participants to identify Tone 4 [4]. Nonetheless, the two studies acknowledged that creaky voice enhances perception in low pitch range.

Tone no.	Tone category	Tone letter
T1	High-level	55
T2	High-rising	25
T3	Mid-level	33
T4	Low-falling	21
T5	Low-rising	23
T6	Low-level	22
T7 (T1)	High-stopped	5
T8 (T2)	Mid-stopped	3
T9 (T3)	Low-stopped	2

Table 1: The Cantonese tonal system in [6]

Other sociolinguistic studies documented regional, gender and educational differences in voice quality in HKC production in general, but none considered the role of tonal categories [7,8]. Fung verified that HKC is creakier than Guangzhou Cantonese when measuring the H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, and CPP [7] when calculating mean F0 across tonal categories. Notably, her research documented a gender difference between the mean values of the acoustic parameters of phonation – nearly all HKC female participants showed creakier production in all five mean values of the acoustic parameters than male, regardless of age (except for CPP in middle-aged participants). Fung and Lee found a similar gender asymmetry in HKC voice quality [8]. They measured the pitch dynamics, spectral tilt, and periodicity of 60 HKC speakers varying in age, gender, and education level in two passage-reading tasks; their results revealed that female speakers adopted an extremely low pitch floor in general and higher-educated females exhibited creakier voice quality. These sociolinguistic findings are in accordance with creaky voice research in other languages as seen in English [9].

What remains unresolved is the source of creaky voice in Cantonese tonal production. Yu and Lam proposed two hypotheses in [1]; because of the relatively high occurrence of creaky Tone 4 in their study, they hypothesize that creaky voice (1) is a phonemic cue for Tone 4, or (2) a by-product of low F0. The current research intends to address this question by examining the occurrence of creaky voice among Cantonese tonal categories and gauge the effect of low F0 on non-modal phonation of Cantonese tones, especially when the pitch range is manipulated. As limited research has considered the role of tonal categories in Cantonese creaky voice production, this study is the second to document the prevalence of creaky phonation in Cantonese tonal production.

2. EXPERIMENT I: DESIGN

2.1. Participants and stimuli

12 Cantonese native speakers (6M 6F, age: 18-24) with no history of speech or listening disorders were recruited for the current study. As education level and language background have been found to influence Cantonese speakers' voice quality [8], the participants were selected to be local Hong Kong undergraduate students with less than one year of living experience in English speaking countries.

Participants were asked to read a list of Cantonese monosyllables /fu/, /sɛ/, /si/, / soeng/, /fan/, / jau/ and /ma/ in all six Cantonese tones using a SONY A10 Linear PCM Recorder in a sound-attenuated booth. As this experiment intends to answer whether creaky voice is a T4 specific cue, it is crucial that participants accurately produce all six citation forms of Cantonese tones. To avoid contaminating of recorded data by HKC tone mergers [10, 11], the utterances which deviated from the citation form of HKC were discarded. A total of 2016 tokens were collected in this experiment (2 syllables * 7 vowels * 6 tones * 2 repetitions * 12 participants), of which 11 tokens were discarded.

2.2. Segmentation criteria

All speech outputs were segmented and annotated manually using Praat v.6.0.37 [12] to retrieve F0 values. The onset of the vowels was chosen to be the onset point of the tokens, meaning that all initial consonants, regardless of whether it is voiced or voiceless, were discarded in the recordings. All offsets, such as /n/ in [fɐn] and /ng/ in [sœːŋ], were not discarded because the lowering of F0 value that usually accompanies phrase-final creak is of interest to the study.

2.3. Measurements

The presence of creak was defined through both visual and audio inspection by the author, a native speaker of Cantonese. Tokens were coded as creaky if they had the auditory percept of creaky voice, as determined by the authors and if: (1) there were alternating cycles of amplitude and/or frequency or irregular glottal pulses in the wide-band spectrogram and waveform or, (2) there were missing values or discontinuities in the F0 track determined by Praat's autocorrelation algorithm with default settings, or (3) there was the appearance of strong subharmonics or a lack of harmonic structure in the narrow-band spectrogram.

3.EXPERIMENT I: RESULTS AND DISSCUSSION

3.1. Creaky voice occurrence

The results of experiment 1 are shown in Table 2, alongside with the findings given by Yu and Lam's study [1]. It is found that, in addition to 227 Tone 4 tokens (61.0%), 75 Tone 5 tokens (21.7%), 55 Tone 2 tokens (16.9%), and 57 Tone 6 tokens (16.7%) showed creaky phonation. Tone 3 had relatively fewer creaky tokens – 26 (7.5%) while Tone 1 had none. The results suggest that it is unlikely that creaky voice is tied to Tone 4 production to ease perception, but rather, creaky phonation is sensitive to all low pitch tones. If creakiness in Tone 4 is specifically produced to enhance identification as proposed by Yu and Lam [1], then it would be unlikely for Tone 5 and Tone 2 to show 21.7% and 16.9% of creak occurrence in the present study. Consequently, it is unlikely that creakiness attribute in Cantonese tone is specific to any tone for perceptual ease, as Tone 2 and Tone 5 exhibit noticeably similar creak occurrence in this study, and they are argued to be most susceptible to mixed perception as they are both rising tones in Cantonese [10, 11].

Tone	Current Study	Yu and Lam [1]
1	0%	NG
2	16.9%	2.6%
3	7.5%	NG
4	61.0%	24.2%
5	21.7%	NG
6	16.7%	NG

Table 2: Percentage of creak in Cantonese tonal production found in current study and [1]

It should be noted that the distribution of creak occurrence generally matches the pitch heights of the six tones. For example, Tone 4, the lowest tone in

Cantonese tonal categories, generally has the highest creak occurrence. Tone 5 and Tone 2, bearing the tonal letter of 23 and 25 respectively, have the second and third highest creaky tones, which suggest that F0 is the major trigger of creaky voice production.

In addition, creak appears at similar F0 values for the recorded creak occurrences, around 90 Hz for male speakers and 170 Hz for female speakers in Figure 1. The value of pitch threshold is recorded by measuring the lowest F0 value before and/or after the occurrence of creak calculated by Praat. The unity in terms of creak threshold across tonal categories suggests that when a speaker's F0 values fall below a certain threshold, creaky phonation is triggered. This finding is similar to the result presented by Kuang [2], in which she defines 95Hz and 180Hz in her study as the threshold for creak to naturally take place when the pitch values hit a bottom line in Mandarin.

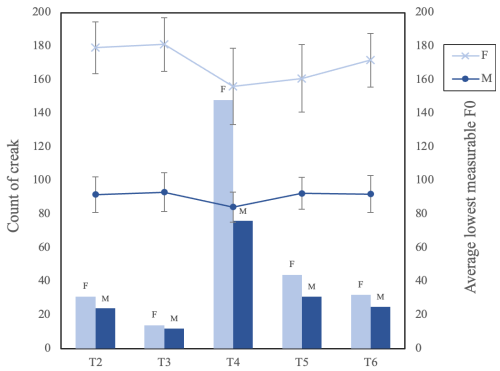


Figure 1: Pitch threshold values – measured by lowest measurable F0 and count of creaks for female and male participants in experiment 1.

Two binomial logistic regression were performed separately for male and female participants to analyse the relationship between creak occurrence and F0 values. Results showed that creak occurrence is correlated with low F0 for both male (β=-0.0852, $p<0.001$) and female (β=0.0311, $p<0.001$) and showed a negative relationship between F0 and creak occurrence for both male and female participants. The higher the F0, the less likelihood for a token to be produced with creaky voice.

3.2 Prosodic positions

The results of experiment 1 also show a significantly higher frequency of creak occurrence on the second syllable ($\chi2$=9.68, $p=0.002$). The second syllable recorded 246 creaky tokens, compared to 191 in the first syllable. A paired t-test was also performed to show the significant difference between the mean F0 values of the first syllable and the second ($t=3.07$, $p<0.001$). These results matched the previous study of Cantonese creaky voice production [1]. Although no previous research has documented the use of creaky voice as a sentence/syllable ending prosodic cue for Cantonese, some studies confirmed the use of creaky to indicate prosodic boundary in Mandarin [13,14]. Future studies should explore the effect of creak on eliciting prosodic cues in Cantonese on different prosodic boundaries.

It should be made clear that because all test words in Experiment 1 were spoken in isolation in a laboratory setting, the reason for the occurrence of creak, especially for the second syllable, could be caused by sentence-final creak accompanied by the lowering of F0. Such limitation could be easily erased if one could replicate the analysis of this experiment with a bigger corpus of daily continuous discourse, especially when tones that are less likely to creak are put at the end of a sentence to gauge the effect of prosodic position on triggering creak.

The difference of creak occurrences based on prosodic positions combined with the results of the logistic regression support that creaky voice is likely to be triggered by producing a low F0 and the percentages of creaky Tone 2 and Tone 5 hint that creaky voice is less likely to be tied to Tone 4. In experiment 2, the claim will be further supported by a manipulated production task asking experiment participants to read the same materials in two conditions: in normal speech and a lowered pitch speech with a lowered pitch range. It is hypothesised that if creak is triggered by low F0 values, when speakers are asked to speak in a low pitch range, more creaky tokens will be observed for tones that are produced in lower pitch ranges other than Tone 4.

4. EXPERIMENT II: EXPERIEMENT DESIGN

The same group of participants was recruited in experiment 2. Following the experiment set up by Kuang [2], Subjects were asked to produce the same set of materials in experiment 1 in two conditions. (1) normal pitch (words are produced in speakers' most comfortable pitch ranges); (2) low pitch (participants were instructed to speak in their lower-than-normal pitch range and were reminded not to whisper). Each token was repeated two times in two conditions monosyllabically.

The segmentation and creak coding were done the same as experiment 1 in which 12-time intervals of each token were extracted for their F0 values. Six seperate paired-sample t-tests were performed to test whether the participants were able to produce a pitch contrast under the two conditions.

5. EXPERIMENTII: RESULTS AND DISCUSSION

The results of experiment 2 are shown in Table 3. Overall, a significantly higher number of creaks is recorded under the low-pitch condition ($\chi2$=74.7, p<0.001). Female participants also produced a significantly higher number of creaks ($\chi2$=33.3, p<0.001). While other tones have recorded a significant increase in creak occurrence, the number of creaky T4 only differs slightly under the low-pitch condition (-4% for female participants and 2% for males. Results of the t-tests confirmed that for Tone 1 (W=10910, p<0.001), Tone 3 (W= 8594, p=0.005), and Tone 4 (W=5464, p=0.049), the average F0 values of the two conditions are significantly different. As for Tone 2 (W=7366, p=0.399), Tone 5 (W=8353, p=0.091), and Tone 6 (W=6993, p=0.426), there is no significant difference in mean F0 values between the two conditions. Figure 2 shows that while the F0 trajectories of female participants are compressed and lowered under the low-pitch condition, male participants' F0 trajectories are only compressed but not lowered.

Tone	Normal		Low pitch		Δ%	
	F	M	F	M	F	M
1	0	0	6	3	-	-
2	30	20	45	30	50%	50%
3	8	9	27	23	238%	156%
4	77	42	74	43	- 4%	2%
5	19	16	42	38	121%	138%
6	21	8	28	32	33%	300%
Total	155	95	222	169	43%	78%

Table 3: Count of creaky phonation in all 6 tones under normal and low-pitch condition by gender

The conflicting results on the mean F0 under two conditions could be caused by two reasons: (1) participants did not fully comprehend the requirements of experiment 2, or (2) some participants were not able to produce an even lower pitch range due to physiological constraints. This could explain why Tone 4 has the lowest percentage change in terms of creak occurrence in the low-pitch condition in Table 3. As Tone 4 generally hits the bottom of the pitch range of the participants, their occurrences in the low-pitch condition are similar to that of the normal condition. For future studies, it is recommended to give clearer guidance to participants in eliciting creaky production. For example, instead of asking the participants to speak in a lower-than-normal pitch range, researchers could ask the participants to speak in their lowest possible pitch range until their voices creak.

The average lowest measurable F0 for creaky tokens was recorded at 165 Hz for female participants and 90 Hz for male participants for experiment 2 for both conditions, as shown in Figure 3. These values are similar to that of experiment 1. As we manipulated speakers' pitch range in experiment 2, having similar F0 values to that of experiment 1 and higher creak occurrences hint that the pitch threshold values we obtained are possible F0 triggers for creaky phonation, regardless of tonal categories.

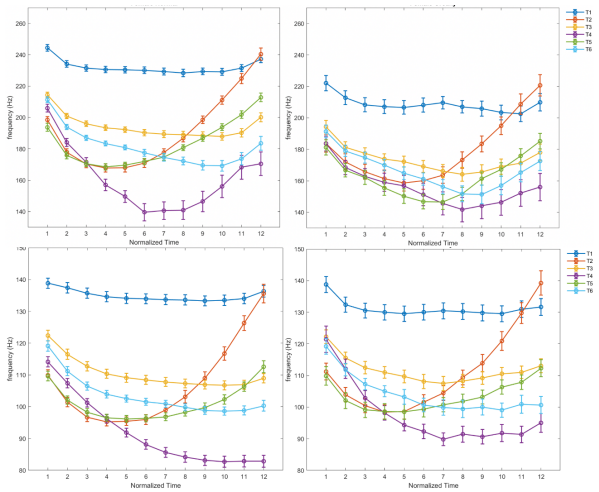


Figure 2: Mean F0 values in two production conditions (left: normal; right: low pitch) for female (upper) and male participants (lower) across time.

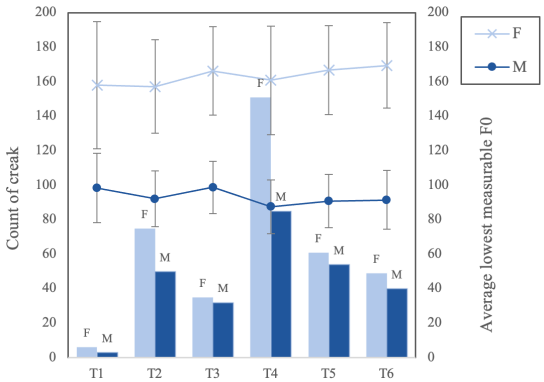


Figure 3: Pitch threshold values – measured by lowest measurable F0 and count of creaks for female and male participants in experiment 2.

6. CONCLUSION

In conclusion, this study demonstrates that creaky voice in HKC is under the influence of covariation of phonation and pitch range, and it is triggered when the speaker's pitch reaches an extremely low pitch, regardless of tonal categories. Some findings could be made from the results of the current study: (1) Tone 4 is not the only tone that creak in tonal production; Tone 2 and Tone 5 also shows considerable a number of creak occurrence and therefore Tone 4 is unlikely a language-specific cue for tonal identification. (2) Creaky voice is more likely to be produced at the second syllable, probably due to F0 declination. (3) A speaker's pitch range manipulation can trigger his/her creaky voice production.

7. ACKNOWLEDGEMENT

The author would like to thank Dr. Vanti Lee for providing access to the sound booth and equipment at CityUHK, as well as her guidance during the revision process. The author also thanks Dr. Jonathan Havenhil's continuous support and encouragement throughout the research process.

8. REFERENCES

[1] Yu, K. M., Lam, H. W. 2014. The role of creaky voice in Cantonese tonal perception. *J. Acoust. Soc. Am*. 136, 1320-1333.

[2] Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *J. Acoust. Soc. Am*. 142, 1693-1706.

[3] Huang, Y. 2019. The role of creaky voice attributes in Mandarin tonal perception. *Proc. 19th ICPhS.* Melbourne.

[4] Zhang, Y., Kirby, J. 2020. The role of F0 and phonation cues in Cantonese low ton perception. *J. Acoust. Soc. Am*. 148, EL40-EL45

[5] Garellek, M. 2012. The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics*, 40, 152-161

[6] Bauer, R. S. 1998. Hong Kong Cantonese tone contours, in S. Matthews (ed.) *Studies in Cantonese Linguistics.* Hong Kong: Linguistic Society of Hong Kong.

[7]Fung, S. Y. R. 2015. Voice Quality: A Preliminary Study on the Phonetic Distinctions of Two Cantonese Accents. *Proc. 18th ICPhS*, Glasgow.

[8] Fung, S.Y.R., Lee, C.K.C. 2019. Voice quality and identity: The case of Hong Kong Cantonese. *The 30th North American Conference on Chinese Linguistics.* The Ohio State University, 16-33.

[9] Yuasa, I. P. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3), 315–337.

[10] Fung, R. S. Y., Lee, C. K. C. 2019. Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *J. Acoust. Soc. Am*. 146, EL424-EL430.

[11] Mok, P., Zuo, D., Wong, P. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, 25(3), 341-370.

[12] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[13] Kuang, J. 2018. The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin. *J. Acoust. Soc. Am* 143, EL509-EL515.

[14] Huang, Y. 2019. The source of creak in Mandarin utterance. *Proc. 19th ICPhS.* Melbourne.

THE PERCEPTUAL EFFECTS OF ALIASING DISTORTION IN GLOTTAL FLOW MODELLING

Zihan Wang, Christer Gobl

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland
zwang7@tcd.ie, cegobl@tcd.ie

ABSTRACT

When modelling the glottal flow signal in a discrete-time system, the aliasing distortion that is produced is typically ignored. The assumption is that the perceptual effects are negligible if the sampling frequency is sufficiently high. In this paper, we implement a recently developed version of the LF model, which eliminates aliasing distortion. By comparing it to the standard application of the LF model, which introduces aliasing distortion, we can explore if, and to what extent, aliasing distortion is perceptible in the modelled voice source signal. The results of a listening test demonstrate that the aliasing distortion is almost always perceptible, even when the sampling frequency is relatively high. The perceptual artefacts are very noticeable when the fundamental frequency is high, particularly in combination with a tense voice quality. In conclusion, therefore, if high-quality modelling of the glottal flow signal is required, aliasing-free source modelling is recommended.

Keywords: glottal flow, voice source generator, aliasing distortion, LF model, voice quality.

1. INTRODUCTION

The glottal flow signal, i.e. the voice source, plays a fundamental role is speech communication. It contributes to the linguistic prosody, such as variations in prominence, accentuation and phrasing e.g., [1, 2, 3]. Furthermore, variation in the voice source carries paralinguistic information, which signals interpersonal information concerning the speaker's state, attitude to the interlocutor and to the discourse [4, 5, 6]. It also coveys extralinguistic information, such as specific characteristics of a speaker's voice, e.g., [7, 8, 9].

Accurate and detailed modelling of the source is therefore important. For the analysis of the glottal flow, a parametric model is often used, where the source parameter data, which define the model waveform, serve as useful descriptors of the source characteristics. Source models can also be used in synthesis, e.g., for the purpose of voice transformation or for the generation of acoustic stimuli to explore the perceptual role of the voice source [10-14].

One aspect of glottal flow modelling generally overlooked is the effect of the aliasing distortion which is produced when the model waveform is sampled. Most source models are defined in the time domain as a set of piecewise elementary functions e.g., [15-21]. Given the discontinuities that will occur in the derivative of these functions, the spectrum is not bandlimited, and thus aliasing will always be produced. Components above the Nyquist frequency (half the sampling frequency) will appear as components at frequencies below the Nyquist frequency, thereby distorting the spectrum.

Since the amplitude level of the source spectrum decreases with increasing frequency, the amount of aliasing is less when the sampling frequency is high. It is therefore often assumed that if the sampling frequency is sufficiently high, the effect will be relatively small and can be disregarded.

However, the perceptual consequences of this distortion in the modelling have not been systematically studied. Hence, in this paper a perception test was carried out to explore the effect of the aliasing distortion and the extent to which it is perceptible.

Two sets of voice source stimuli were used, involving pairs of stimuli which are identical except for the aliasing distortion being present in one and not in the other. Among the stimuli further variables were manipulated to explore other factors that might impact on the perceptibility of aliasing. These tested the following hypotheses: aliasing is consistently audible, even at high sampling frequencies; aliasing is more strongly perceived when the sampling frequency is low and when the f_0 level is high; it is more strongly perceived for a tense voice quality with strong higher harmonics than for a lax voice quality with weak higher harmonics; voice source dynamics might influence the perceived aliasing, the initial expectation being that greater dynamics would results in the aliasing being less strongly perceived, since it is conceivable that there would be greater masking of the aliasing components in these conditions.

2. METHODOLGY

Recently, an alternative discrete-time implementation of the widely used Liljencrants-Fant (LF) model [17]

was developed, which eliminates the aliasing distortion present in the standard application [22]. By comparing the two versions of the model, it is possible to study the perceptual effects of the aliasing distortion in the glottal flow signal.

2.1. The LF model of glottal flow

The LF model is defined by the piecewise function shown in (1).

$$(1)\quad U_g'(t) = \begin{cases} E_0\, e^{\alpha t} \sin \omega_g t & 0 \le t < t_e \\ -\dfrac{E_e}{\varepsilon\, T_a}\left(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon\, T_b} \right) & t_e \le t < t_c \end{cases}$$

It models the derivative of the glottal flow pulse and involves two sub-functions. The first is a segment of an exponentially growing sinusoid, which models the flow derivative during the open phase of the glottal cycle. The duration of the open phase is T_e, and at the end of the open phase the amplitude of the waveform is $-E_e$, which is the negative amplitude of the main glottal excitation (see Fig. 1).

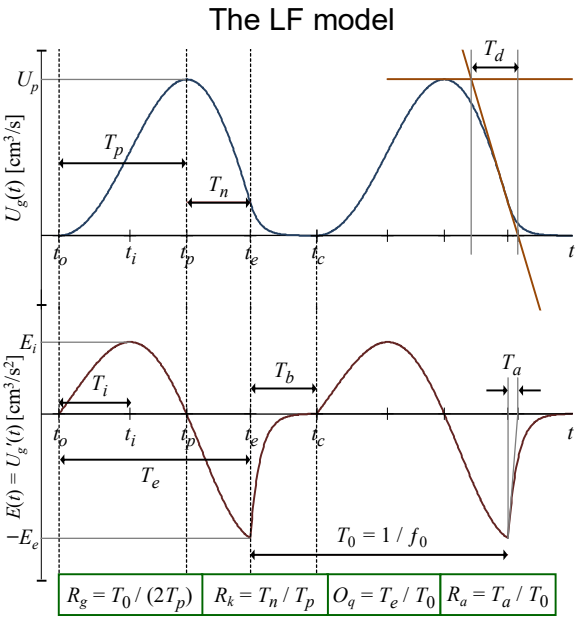


Figure 1: Two LF pulses and parameter definitions. Flow derivative (bottom) and corresponding glottal flow (top).

The second sub-function is a segment of an exponential function, which models the flow derivative during the return phase, i.e. the part of the cycle for which the vocal folds return to full or maximum closure. The duration of this segment is $T_b = t_c - t_e$ (Fig. 1).

The LF model pulse shape is often described by the parameters R_g, R_k and R_a (Fig. 1). Together with E_e, these R-parameters are sometimes called 'the LF-parameters'. However, as has been discussed in [23], the actual parameters of the model, which are required to generate the LF pulse, are E_e, T_e, ω_g, α, ε and T_b.

2.2. The LF model without aliasing

The LF glottal waveform is continuous in time, and when digitised, this is typically done by sampling the time domain functions in (1). The sampling process will introduce aliasing distortion, since the spectrum of the LF model pulse is not bandlimited, and consequently it will always extend beyond the Nyquist frequency regardless of how high the sampling frequency is. Thus, the standard discrete-time version of LF model will have a distorted frequency spectrum.

A method is presented in [22] which avoids aliasing distortion in voice source models and a mathematical framework for generating an aliasing-free version of the LF model is outlined. To avoid aliasing distortion, the model needs to be described in the frequency domain, which can be done by deriving the Laplace transform of the model. This enables the calculation of the true spectrum of the model directly from the model parameters.

The ideal discrete spectrum of the model can then be obtained by sampling the true model spectrum up to the Nyquist frequency. Given this ideal spectrum, which effectively bandlimits the model spectrum to the Nyquist frequency, no aliasing is introduced. The discrete-time LF pulse corresponding to the ideal discrete spectrum is derived by applying the inverse discrete Fourier transform (IDFT). For full details on how the aliasing-free LF model is calculated, see [22].

2.3. The voice source generator

To produce the stimuli to explore the perceptual effects of the aliasing distortion introduced by the sampling process, a system for generating voice source signals was developed. This system, referred to as the voice source generator (VSG) was designed using the MATLAB App Designer [24].

The VSG currently incorporates the two version of the LF model described above: the standard version with aliasing and the aliasing-free version, implemented according to [22].

Furthermore, the VSG offers two different sets of control parameters for the source modelling: one set is the commonly used 'LF parameters', i.e. E_e, R_g, R_k, R_a and f_0. The other set of parameters is R_d, E_e, O_q and f_0. Note that only four parameters need to be specified in this case: R_a is derived from the R_d value using the formula provided in [25].

$R_d = 0.11^{-1} f_0 U_p / E_e = 0.11^{-1} T_d / T_0$ (see Fig. 1) is the global waveshape parameter proposed by Fant [25, 26], which is the central parameter in the alternative parameter control system for the LF model. The R_d

value has been shown to be a good indicator of the tense-lax dimension of voice quality, being low for tense voice and high for lax voice. O_q is the glottal open quotient, here defined as T_e/T_0 (Fig. 1).

To ensure that a possible LF pulse is always produced from the input parameters, parameter constraints as defined in [27] are imposed on the input values to the VSG. Furthermore, amplitude modulated aspiration noise can be added to the voice source pulses, where the modulation is determined by the glottal pulse shape according to [28].

The VSG can also generate LF pulses from the analysis output of the GlórCáil system, a system that can automatically extract LF parameter data from a speech signal [13, 29] or from the output the ISF system, a manual interactive inverse filtering and source modelling system described in [30, 31, 32].

In the future, it is envisaged that additional source models will be included in the VSG, as well as an extended choice of control parameters.

2.4. Acoustic stimuli

The voice source generator described above was used to create 38 acoustic stimuli. These stimuli made up 19 pairs, where the only difference between the stimuli in each pair was the version of the LF model: one producing aliasing and the other aliasing-free.

Out of these 19 pairs, two sets of 8 pairs were generated by specifying input values using the second option of control parameters, i.e. R_d, E_e, O_q and f_0 (aspiration noise was not used). The two sets differed only in terms of the sampling frequency, 10 kHz vs. 20 kHz, to test the hypothesis that aliasing is less perceptible at a higher sampling frequency. These are referred to as stimulus pairs 1 to 8 (see Table 1).

To test if a source signal with higher f_0 and a tenser voice quality produces aliasing that is more perceptible, stimuli 1 to 6 (both sets) differed in terms of f_0 level and voice quality: there were three f_0 levels with either lax or tense settings. Stimuli 1 and 4 had low f_0 (mean f_0 109 Hz), stimuli 2 and 5 medium f_0 (mean f_0 219 Hz), stimuli 3 and 6 high f_0 (mean f_0 437 Hz), stimuli 1 to 3 had high R_d (lax voice, mean R_d was 2.1), stimuli 4 to 6 had low R_d (tense voice, mean R_d was 0.58). These 6 stimuli (both sets) all had relatively little variation in the parameter values (small dynamics). To test the effect of the dynamic variation in the source, stimuli 7 and 8 (both sets) on the other hand involved relatively large dynamics. The only difference between stimuli 7 and 8 was the f_0 level: relatively low for stimulus 7 (mean f_0 166 Hz) and relatively high for stimulus 8 (mean f_0 322 Hz). The R_d mean was 1.2 for both stimuli 7 and 8.

So far, all stimuli are based on stylised synthetically generated source signals. The remaining three pairs of stimuli were generated from source data obtained in a previous study [33] derived from natural speech using the manual interactive ISF system. The three utterances analysed were of the sentence “We were away a year ago” spoken by a male speaker in neutral, angry and sad voice. These stimuli were added to ascertain that the findings on the other stimuli should hold for naturally spoken utterances.

2.5. Listening test

To evaluate the perceptual effect of the aliasing in glottal flow modelling, a listening test was carried out with the 19 pairs of stimuli described above. The test was administered online, and participants were instructed to take the test in a quiet room using high-quality headphones. They were also informed that they could listen to each audio file as many times as they wished. Before the main test, a file containing all the test stimuli was presented. This allowed the participants to familiarise themselves with the range of differences involved in the different pairs of stimuli and to adjust the volume to a comfortable level.

Participants then proceeded to the main test to assess the perceived degree of difference between the pairs of stimuli. To do this, they used a scale from 0 to 5. If no difference between the stimuli was perceived, 0 would be selected. If a difference was perceived, the strength of the perceived difference was rated between 1 and 5, where 1 meant a very small difference and 5 a very large difference (approximately the maximum differences found among the pairs of stimuli). The stimuli were presented in a randomised order. The order of the two stimuli (with or without aliasing) in each pair was also random.

3. RESTULTS AND DISCUSSION

In total, 30 participants completed the listening test. The means and standard deviations of the perceived difference scores for each pair of stimuli are presented in Table 1. One-way analysis of variance was used to test the significance of differences between groups of stimuli. The results show that the aliasing is perceptible in both the stylised stimuli and the stimuli generated using natural spoken data – even at a relatively high sampling frequency. The difference between the difference score and 0 (imperceptible) is statistically significant ($p < 0.001$) for all pairs of stimuli, including the sad stimuli, for which the difference score was the smallest.

Given that the amount of aliasing is less at a high sampling frequency, we would expect the perception scores to reflect this. The overall mean score was indeed lower at the higher sampling rate (Table 1), but the difference was smaller than expected, and was not statistically significant ($p = 0.072$) for these data.

Table 1: Mean and standard deviation values of the difference scores for the pairs of stimuli.

f_s	10 kHz		20 kHz	
Stimuli	*Mean*	*Std*	*Mean*	*Std*
1	1.50	1.18	1.37	1.28
2	2.57	1.43	2.13	1.45
3	2.83	1.46	2.97	1.64
4	1.43	1.28	1.20	1.22
5	2.87	1.36	2.13	1.63
6	3.33	1.58	3.03	1.66
7	1.97	1.45	1.70	1.39
8	2.60	1.52	2.47	1.67
Overall	**2.39**	1.41	**2.13**	1.49
Angry	1.83	1.21		
Neutral	2.63	1.40		
Sad	0.77	0.96		
Overall	**1.74**	1.19		

One would also expect that a high f_0 level would make the aliasing more audible, since when f_0 is high the harmonics close to the Nyquist frequency would be relatively stronger compared with the lower harmonics. Furthermore, the increased frequency gap between the harmonics could potentially lead to less masking of the aliasing components.

This expectation was indeed borne out by the results. Fig. 2 shows that the degree of perceived difference is greater as f_0 increases. The statistical analysis suggests that this effect is significant ($p < 0.01$).

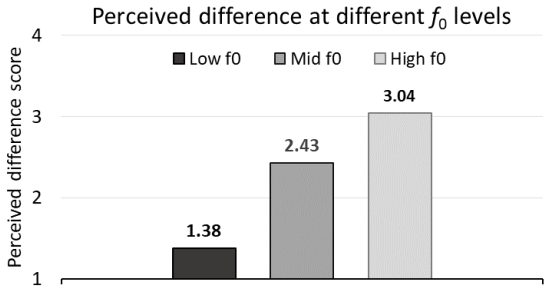


Figure 2: Mean difference scores for stimuli with low, medium and high f_0.

It was also hypothesised that voice quality could affect the perceptibility of aliasing, as a tense voice, with stronger higher harmonics, would be expected to cause more aliasing than a lax voice. Comparing the tense and lax stimuli (stimuli 1-3 vs. 4-6), the overall difference in mean ratings was small (tense 2.33, lax 2.23) and not statistically significant ($p = 0.54$).

However, if one compares only the 10 kHz stimuli for the tense vs. lax voice at low, medium and high f_0 levels (Fig. 3), a more nuanced interpretation is suggested. At low f_0 levels, there is essentially no difference in the scores. At mid f_0 level, a difference does emerge, with tense voice yielding a greater degree of perceived aliasing. At high f_0 level, the difference is greater again. We would tentatively conclude therefore that voice quality plays some role if the sampling frequency is low, with aliasing effects becoming more perceptible with tense voice as f_0 increases.

It was also suggested in the Introduction that a dynamically changing source signal could help mask aliasing components, thus making the aliasing less audible. However, our results do not support this hypothesis. Although the mean rating was slightly lower (2.18 vs. 2.32) for the stimuli with high dynamics, the difference was not significant ($p = 0.48$).

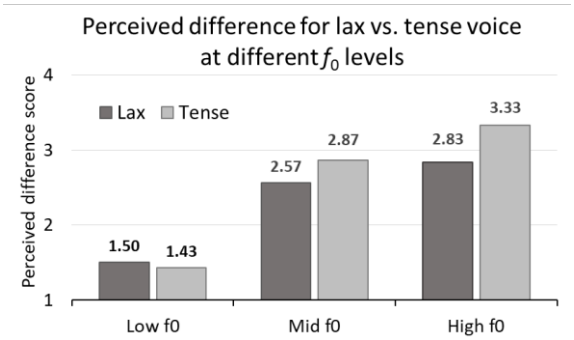


Figure 3: Mean difference scores for stimuli with different voice tension at low, mid and high f_0 ($F_s = 10$ kHz).

Finally, results for the stimuli based on natural spoken utterances showed similar trends to the stylised stimuli. The mean rating may seem somewhat lower, but this is most likely due to the relatively low f_0 values in these stimuli. The mean f_0 across the three stimuli is 106 Hz, comparable to stimuli 1 and 4, which in fact have somewhat lower scores (Table 1). We can therefore be reasonably confident that differences perceived in the stylised voice source stimuli reflect those we would get from real speech data.

4. CONCLUSIONS

The perceptual effect of aliasing distortion is explored by comparing the voice source signals produced by two version of the LF glottal flow model: the standard version which introduces aliasing distortion and a novel implementation based on a frequency domain representation of the model [22]. The perception test shows that the aliasing is almost always perceptible, even when the sampling frequency is high. The distortion is strikingly noticeable when f_0 is high, particularly in combination with tense voice quality. The results therefore suggest that aliasing-free source modelling is beneficial when high-quality modelling of the glottal flow signal is required.

5. ACKNOWLEGEMENTS

This research was carried out in the Róbóglór and ABAIR projects, supported by the Irish Govt. Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, with funding from the National Lottery, supporting An Stráitéis 20 bliain don Ghaeilge 2010-2030.

6. REFERENCES

[1] Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N., van Heuven, V. J. 1995. Supralaryngeal resonance and glottal pulse shape as correlate of stress and accent in English. *Proc. 13th ICPhS*, Stockholm, Sweden, 630–633.

[2] Ní Chasaide, A., Yanushevskaya, I., Gobl, C. 2015. Prosody of voice: declination, sentence mode and interaction with prominence. *Proc. 18th ICPhS*, Glasgow, UK, paper no. 0476, 5 pp.

[3] Yanushevskaya, I., Murphy, A., Gobl, C., Ní Chasaide, A. (2022). Global waveshape parameter R_d in signaling focal prominence: Perceptual salience in the absence of f0 variation. Front. Commun. 7, 1–23.

[4] Sundberg, J., Patel, S., Björkner, E., Scherer, K. R. 2011. Interdependencies among voice source parameters in emotional speech. *IEEE Trans. Affect. Comput*. 2, 162–174.

[5] Patel, S., Scherer, K. R., Björkner, E., Sundberg, J. 2011. Mapping emotions into acoustic space: The role of voice production. *Biol. Psychol.* 87, 93–98.

[6] Yanushevskaya, I., Gobl, C., Ní Chasaide, A. 2018. Cross-language differences in how voice quality and f0 contours map to affect. *J. Acoust. Soc. Am.* 144 (5), 2730-2750.

[7] Hanson, H. M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.* 101, 466-481.

[8] Hanson, H. M., Chuang, E. S. 1999. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* 106, 1064–1077.

[9] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121 (4), 2283–2295.

[10] Cabral, J. P., Renals, S., Yamagishi, J., Richmond, K. 2011. HMM-based speech synthesiser using the LF-model of the glottal source. *Proc. ICASSP*, 4704–4707.

[11] Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., Lin, Q. 1989. Voice source rules for text-to-speech synthesis. *Proc. ICASSP*, Glasgow, UK, vol. 1, 223–226.

[12] del Pozo, A., Young, S. 2008. The linear transformation of LF glottal waveforms for voice conversion. *Proc. INTERSPEECH 2008*, Brisbane, Australia, 1457–1460.

[13] Murphy, A., Yanushevskaya, I., Ní Chasaide, A., Gobl C. 2020. Testing the GlórCáil System in a Speaker and Affect Voice Transformation Task," *Proc. 10th Int. Conf. on Speech Prosody,* Tokyo, Japan, 950–954.

[14] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Comm.* 40, 189–212.

[15] Rosenberg, A. E. 1971. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.* 49 (2B), 583–590.

[16] Ananthapadmanabha, T. V. 1984. Acoustic analysis of voice source dynamics. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 25 (2-3), 1–24.

[17] Fant, G., Liljencrants, J., Lin, Q. 1985. A four-parameter model of glottal flow. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 26 (4), 1–13.

[18] Fujisaki, H., Ljungqvist, M. 1986. Proposal and evaluation of models for the glottal source waveform. *Proc. ICASSP*, Tokyo, Japan, 1605–1608.

[19] Milenkovic, P. H. 1993. Voice source model for continuous control of pitch period. *J. Acoust. Soc. Am.* 93 (2), 1087–1096.

[20] R. Veldhuis, 1998. A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *J. Acoust. Soc. Am.* 103 (1), 566–571.

[21] Shue, Y.-L., Alwan, A. 2010. A new voice source model based on high-speed imaging and its application to voice source estimation," *Proc. ICASSP*, Dallas, Texas, 5134–5137.

[22] Gobl, C. 2021. The LF model in the frequency domain for glottal airflow modelling without aliasing distortion. *Proc. INTERSPEECH 2021*, Brno, Czechia, 2651–2655.

[23] Gobl, C. 2017. Reshaping the transformed LF model: generating the glottal source from the waveshape parameter R_d. *Proc. INTERSPEECH 2017*, Stockholm, Sweden, 3008–3012.

[24] MATLAB 2020. Version 9.8.0.1451342 (R2020a). Natick, Massachusetts: The MathWorks Inc.

[25] Fant, G. 1995. The LF-model revisited: transformations and frequency domain analysis. *STL-QPSR* 2-3, 119–156.

[26] Fant, G. 1997. The voice source in connected speech. *Speech Comm*. 22, 125–139.

[27] Gobl, C. 2003. *The voice source in speech communication – production and perception experiments involving inverse filtering and synthesis*. Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm.

[28] Gobl, C. 2006. Modelling aspiration noise during phonation using the LF voice source model. *Proc. INTERSPEECH 2006*, Pittsburgh, Pennsylvania, 965–968.

[29] Murphy, A. 2020. *Controlling the Voice Quality Dimension of Prosody in Synthetic Speech using an Acoustic Glottal Model*. Unpublished PhD thesis, Trinity College Dublin.

[30] Gobl, C., Ní Chasaide, A. 2010. Voice source variation and its communicative functions. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds), *The Handbook of Phonetic Sciences* (2nd edition). Oxford: Blackwell, 378–423.

[31] Gobl, C., Ní Chasaide, A. 1999. Techniques for analysing the voice source. In: Hardcastle, W. J., Hewlett, N. (eds), *Coarticulation: Theory, Data and Techniques.* Cambridge: Cambridge University Press, 300–320.

[32] Monahan, P. 1996. *Systems for voice source analysis*. Unpublished PhD thesis, Trinity College Dublin.

[33] Wang, Z., Gobl, C. 2022. Contribution of the glottal flow residual in affect related voice transformation. *Proc. INTERSPEECH 2022*, Incheon, Korea, 5288–5292.

CONTRIBUTIONS OF ACOUSTIC MEASURES TO THE CLASSIFICATION OF LARYNGEAL VOICE QUALITY IN CONTINUOUS ENGLISH SPEECH

Chenzi Xu[1], Paul Foulkes[1], Philip Harrison[1], Vincent Hughes[1], Jessica Wormald[1]

[1]Department of Language and Linguistic Science, University of York, UK
{chenzi.xu|paul.foulkes|philip.harrison|vincent.hughes|jessica.wormald}@york.ac.uk

ABSTRACT

Laryngeal voice qualities (e.g. breathy and creaky voice), variable within and across speakers, often pose a challenge in data collection. Their acoustic correlates are still inadequately understood. This study revisits the acoustics of laryngeal voice qualities in high-quality recordings of continuous British English speech produced by experienced phoneticians. Through principal component analysis and multinomial logistic regression with l_1 regularisation, this study identifies contributions of a variety of acoustic measures to the classification of laryngeal voice qualities and provides a multidimensional acoustic profile for breathy, creaky, and modal voice. Classification rates as high as 90% were achieved using the first 5 principal components. The most salient acoustic correlates for creaky voice are, compared to other categories, higher mean H2*, lower mean f0 and HNR below 500 Hz, and for breathy voice, higher mean H1* and spectral tilt measures such as H1*–A1* and H1*–H2*.

Keywords: Voice quality, phonation types, acoustics

1. INTRODUCTION

Variation in laryngeal voice quality, such as breathiness and creakiness, is abundant in speech. Voice quality may be used contrastively as in Jalapa Mazatec [1], enhance perceptual distinctiveness of other phonetic categories such as lexical tones as in White Hmong [2], or function as prosodic configurations such as phrase-final creak in various languages [3]. There is also evidence of sociolinguistic differences in voice quality (e.g. [4]), as well as speaker-specific variation which may be relevant to forensic speech science [5, 6].

Laryngeal voice quality has received increasing attention from the perspectives of articulation, acoustics, and perception in the past two decades. Phoneticians have been dedicated to searching for acoustic features that are capable of distinguishing modal, creaky and breathy voice (e.g. [7]). The two basic articulatory dimensions to laryngeal voice quality are the degree of constriction of vocal folds and aspiration noise [8], which can be captured by spectral tilt and harmonic to noise ratio (HNR) measures respectively as proposed in the psychoacoustic model described in [9]. For spectral tilt there are various proposed measures, of which the most common one is H1–H2 or the corrected H1*–H2*[10], the difference in amplitude between the first and second harmonics, amongst other higher-frequency slopes in source spectrum such as H2–H4, H4–H2kHz, and H2kHz–H5kHz. An alternative set of measures of spectral tilt include H1–A1, H1–A2, H1–A3 varying in their harmonic bandwidth depending on the formant frequencies, or in other words, vowel quality. These may also correlate with the source spectral tilt measures. In addition to spectral tilt and HNR measures, other proposed acoustic correlates of laryngeal voice quality in the literature include corrected amplitude of individual harmonics such as H1, H2, and H4, cepstral peak prominence (CPP) [11], f0, formant frequencies and bandwidths [12].

Although it has been found that both the source and filter characteristics provide important cues for phonation contrasts [9, 13, 14], the influence of the filter on laryngeal voice quality is often ignored. The correlations among the selected acoustic measures are also often neglected, which impacts the interpretability of statistical models.

This paper revisits the acoustics of laryngeal voice quality via an interpretable classification algorithm whereby combinations of acoustic features are used to predict laryngeal voice quality. We examine the contributions of a wide range of the proposed acoustic measures (including features associated with the filter) in distinguishing laryngeal voice quality. While many studies measured voice quality from sustained vowels (e.g. [15]), this study utilises continuous speech, which better represents the dynamic attributes of voice in speech.

2. DATA

This paper reports on a subset of the available materials from a bespoke corpus, collected as part of an on-going project, containing high quality recordings of experienced phoneticians in various vocal conditions.

2.1. Participants

Data from four male adult speakers were included in this study (P1-P4). They are all experienced professional phoneticians.

2.2. Procedure

Each participant read the first two paragraphs of *The Rainbow Passage* in twenty-four vocal conditions, varying laryngeal and supralaryngeal voice quality settings, as well as accent guises and other vocal changes (e.g. holding a pen between their teeth). Each vocal condition was repeated three times non-consecutively within a recording session. All participants took part in three sessions which were at least a week apart. Thus, each vocal condition has 3*3 = 9 repetitions, and the recording of each repetition was saved as a PCM WAV file. The repeated design better captures intra-speaker variation. This paper reports only on the modal, breathy, and creaky voice conditions.

2.2. Recordings

The recording sessions were conducted in an anechoic chamber at the University of York. Each participant wore a DPA omnidirectional headset microphone on the right side of their face. The recording was made on a single channel at a sampling rate of 48 kHz and 24-bit quantization using a Zoom F8n recorder.

3. METHOD

This study employs multinomial logistic regression to predict the phonation types with a wide range of voice source and vocal tract features. The preprocessing, data exploration and analysis were implemented in MATLAB, R, and Python, and the scripts are available as supplemental materials[1].

3.1. Acoustic Measurements

Using VoiceSauce [16] in MATLAB, 29 acoustic features (see Table 1) were extracted every 25ms frame with 10ms frame shift for all audio files. All spectral magnitude measures were corrected using f0 and formants. Both f0 and formants were estimated using the Snack algorithm [17]. The f0 extraction range set from 40 Hz to 300 Hz, and four formants were tracked using the default setting with LPC order being 12 and pre-emphasis set to 0.96.

Only measurements from voiced portions (f0>0) were included. Visual inspection of f0 tracking with spectrogram and waveform revealed that the f0 of each speaker did not exceed 230 Hz. Thus, measures of frames whose f0 was larger than 230Hz were considered tracking artefacts and were excluded.

Each audio file is on average 29.04s in duration (SD = 3.77s) and evenly divided into six breath-group-size [18] intervals of about 4.84s each. The mean duration of voicing portion of the intervals is 2.4s (SD = 0.75s). For the analysis, the mean and standard deviation of measures of each acoustic feature over each interval were used. The division of intervals enriched the dataset, instead of using means and standard deviations for acoustic measures calculated across the entire sample of speech.

There are, hence, 162 sets (9 repetitions of 3 voice qualities with 6 intervals per repetition) of measures for each speaker, and each set contains 58 measures (means and standard deviations of 29 features).

Measure	Explanation
Spectral slope measures (dB)	
H1*, H2*, H4*, H2k*	Amplitude of the first, second, fourth harmonic, and the harmonic nearest 2,000 Hz
H1*–H2*, H2*–H4*	Difference in amplitude between the first and second harmonics, second and fourth harmonics
H4*–H2k*	Difference in amplitude between the fourth harmonic and the harmonic nearest 2,000 Hz
H2k*–H5k*	Difference in amplitude between the harmonic nearest 2,000 Hz and the harmonic nearest 5,000 Hz
A1*, A2*, A3*	Amplitude of the first, second, and third formant
H1*–A1*, H1*–A2*, H1*–A3*	Difference in amplitude between the first harmonic and the harmonic closest to the first, second, and third formant
Energy/Amplitude related measures	
HNR05, HNR15, HNR25, HNR35	Harmonic-to-Noise Ratio for 0-500 Hz, 0-1500 Hz, 0-2500 Hz, and 0-3500 Hz
CPP	Cepstral Peak Prominence
Energy	Root Mean Square (RMS) energy
Vocal tract related measures	
F1, F2, F3, F4	First through fourth formant
B1, B2, B3, B4	First through fourth bandwidth
Glottal measures	
f0	Fundamental frequency (Hz)

Table 1: Acoustic measures extracted from VoiceSauce.

3.2. Multinomial logistic regression

To learn the relationship between phonation types and acoustic features, we fitted multinomial logistic regression models with l_1 regularisation, using the

LogisticRegression() function and SAGA solver [19] in Scikit-Learn Python library [20]. The l_1 regularisation shrinks non-significant coefficients to exactly 0 through penalising the absolute values of magnitude of coefficients during optimization, thereby reducing unnecessary acoustic measures in the final model. The predicted variable was one-hot encoded such that {*breathy* = 0, *creaky* = 1, *modal* = 2}. All acoustic features were z-score normalised, since features measured at different scales might not contribute equally to model fitting and create a bias.

3.3. Cross-validation

Cross-validation is employed to evaluate model performance, given the small-scale dataset. This prevents model overfitting by partitioning a subset of data to validate the model prediction accuracy and utilise all the data both for training and for testing.

In our implementation, two sets of cross-validation were built. In set S, data from one **speaker** was singled out as a test set while the remaining data was used for training, and this process was repeated for every speaker. In set R, data from one **repetition** for all four speakers was held as a test set while the remaining eight repetitions were used for training, and this process was repeated for every repetition.

Two metrics were used to measure the classifier performance:(1) accuracy, the percentage of correctly predicted phonation types in the test data, and (2) F1-score, whether the model predictions are balanced across three predicted categories. An F1-score close to 1 generally means better classification.The average results of all iterations in cross-validation are reported.

3.4. Dimensionality reduction

Regression models with all 58 standardised acoustic features (base) as predictors were initially built. To simulate the chance level predictions, baseline models were also set up by randomly sampling predictions from a discrete uniform distribution of [0,2] and comparing them to true labels. The classification results are presented in Table 2.

	Base models		Random baseline	
Average	Set R	Set S	Set R	Set S
Accuracy	0.98	0.79	0.32	0.34
F1 score	0.98	0.78	0.25	0.27

Table 2: Classification results of full models and random baseline models.

The base models greatly outperformed the random predictions, indicating that there are statistical regularities between phonation types and these acoustic features. Set R has higher accuracy than Set S, suggesting that interspeaker variation is larger than intraspeaker variation in laryngeal voice quality.

The base models, however, are not fully interpretable, as many of the features are highly correlated. Hence, we applied principal component analysis (PCA) to the acoustic features prior to logistic regression models, thereby removing multicollinearity between predictors, as PCA transforms a set of correlated variables into a smaller number of orthogonal variables. We built a pipeline chaining PCA and logistic regression to search for optimal parameters of PCA.

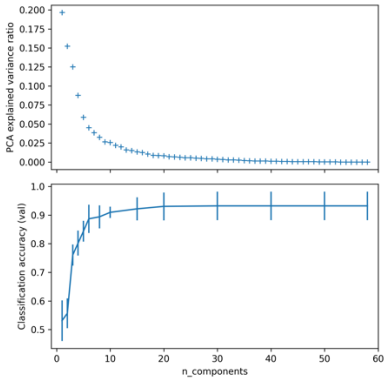


Figure 1: The explained variance and mean classification accuracy across principal components (set R).

We found that with the first five principal components, the classification accuracy consistently reaches over 80%, both for the R and S sets. In Figure 1, the inclusion of the third component greatly increased the accuracy from below 60 % to close to 80%. Hence, we transformed the data with the first five principal components and then used them as predictors in logistic regression (PCA models).

4. RESULTS

4.1. Classification

The PCA models achieved good overall classification results as shown in Table 3, although the first five components captured only about 60% variability in the acoustic data. The confusion matrix[2] from the classification indicates that breathy voice was distinguished from creaky voice with 100% accuracy (and vice versa). Here, we will focus on Set S models tested on unseen speakers.

	PCA models	
	Set R	Set S
Avg. Accuracy	0.9	0.78
Avg. F1 score	0.9	0.78
Avg. Explained variance	0.62	0.63

Table 3: Classification results of PCA models.

The regression model coefficients for each principal component (PC) are listed in Table 4. PCs with larger absolute coefficients indicate their larger contribution to prediction. We can see that all five PCs contribute to the classification of laryngeal voice

quality, with PC1 and PC3 generally the most important variables.

Set S	PC1	PC2	PC3	PC4	PC5
Breathy	-0.96	-0.58	1.43	0	0.90
Creaky	2.52	0	-2.86	0.93	-1.82
Modal	0	0.29	0	-0.13	0

Table 4: Model coefficients for one of the models in Set S. Insignificant coefficients were shrunk to 0 by l_1 regularization.

In order to better understand how principal components predict each phonation type, we visualise the test data in the space of fitted PCs. Figure 2 shows the distribution of test set data (speaker P4; Professor Francis Nolan) in a two-dimensional space of PCs derived from the training data. The inspection of PCA transformed test data in Set S suggests that the first PC is crucial to separate creaky voice (green dots) from the other categories, while PC3 distinguishes breathy voice (orange dots) from the others.



Figure 2: Two-dimensional projection of the test set data using PCs from training set. 0, 1, 2 represents breathy, creaky, and modal voice respectively.

4.2. Acoustic measures contributing to the PCA

The results of section 4.1 can be made interpretable by examining the factor loadings, i.e. the correlation between PCs and acoustic features. Table 5 presents the key factor loadings for the first and third PCs from the data of the first three speakers, used as training data for classification of voice qualities for the test speaker P4 (Figure 2). Full factor loadings can be accessed in the supplemental materials[3].

In Table 5, PC1 negatively correlates with standard deviations of amplitude of harmonics, f0 and HNR mean measures, especially HNR below 500 Hz, and positively correlates with the mean measures of $H2^*$, $H4^*$–$H2k^*$. The first component mainly captures f0 and amplitude or energy related measures, both in inharmonic and harmonic source, distinguishing creaky voice from other categories. PC3 correlates positively with spectral tilt means across all frequency bands $H1^*$–$A3^*$, $H1^*$–$A2^*$, $H1^*$–$A1^*$, $H1^*$–$H2^*$, the mean amplitude of $H1^*$, and the mean and standard deviation measures of first formant bandwidth, and negatively with HNR mean measures.

In Figure 2, creaky voice tends to have higher PC1 values, indicating, as might be expected, lower mean f0 and HNR measures, especially HNR below 500Hz, compared to others. That breathy voice tends to have higher PC3 suggests that breathy voice tends to have higher mean $H1^*$, spectral tilt mean measures such as $H1^*$–$A1^*$ and $H1^*$–$H2^*$, mean/SD of first formant bandwidth, and lower mean HNR measures, especially HNR below 2500 and 3500Hz.

Set S	PC1	PC3
Positive	$H2^*$ (0.14), $H4^*$–$H2K^*$ (0.15) (mean)	$H1^*$, $H1^*$–$H2^*$, $H1^*$–$A1^*$, $H1^*$–$A3^*$, B1 (mean); B1 (sd)
Negative	f0, HNR05 (mean); H1*, H4*, A1*, A3*, H2*, A2*, H2K* (sd)	HNR35, HNR25, HNR15 (mean)

Table 5: Key acoustic measures for PCs (Training set: P1-P3; selected: absolute loadings >0.2 unless specified).

5. DISCUSSION

Laryngeal voice quality is manifested in multiple acoustic features, both source and filter measures. PCA reveals underlying patterns in the wide range of acoustic measures and provides a multidimensional description of the acoustics of phonation types. For instance, PC1 in the sample training set correlates positively with mean $H2^*$ and negatively with mean HNR <500 Hz and f0, assembling features potentially profiling creaky voice, while PC3 correlates positively with mean $H1^*$, spectral tilt measures and B1, and negatively with HNR measures, which related to characterising breathy voice. The orthogonal nature of PCs might also provide insights into person-specific patterns, when test data of different speakers was visualised on the PC space learnt from the same reference set.

6. CONCLUSION

The study revisits the acoustics of laryngeal voice quality via a classification approach using PCA and logistic regression. Laryngeal voice quality was well differentiated using the selected acoustic measures. The average classification accuracy was 78% (Set S) and 90% (Set R); well above chance. Consistent with [13], lower mean HNR measures characterise breathy and creaky voice and our findings further suggests HNR measures at different frequency bands play a role in differentiating breathy voice from creaky voice. Other salient acoustic features include spectral measures $H1^*$, $H2^*$, $H1^*$–$H2^*$, and f0, B1.

7. REFERENCES

[1] M. Garellek and P. Keating, 'The acoustic consequences of phonation and tone interactions in Jalapa Mazatec', *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 185–205, Aug. 2011, doi: 10.1017/S0025100311000193.

[2] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, 'Voice quality and tone identification in White Hmong', *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1078–1089, Feb. 2013, doi: 10.1121/1.4773259.

[3] M. Garellek, 'Perception of glottalization and phrase-final creak', *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 822–831, Feb. 2015, doi: 10.1121/1.4906155.

[4] J. Stuart-Smith, 'Glasgow: Accent and voice quality', in *Foulkes & Docherty 1999*, 1999, pp. 201–222.

[5] E. S. Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, 'The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals', *J. Int. Phon. Assoc.*, vol. 49, no. 3, pp. 353–380, Dec. 2019, doi: 10.1017/S0025100318000130.

[6] N. Francis, 'Forensic Speaker Identification and the Phonetic Description of Voice Quality', in *A figure of speech: A Festschrift for John Laver*, W. J. Hardcastle and J. M. Beck, Eds., New York: Routledge, 2005, pp. 385–411. doi: 10.4324/9781410611888-28.

[7] S. Seyfarth and M. Garellek, 'Plosive voicing acoustics and voice quality in Yerevan Armenian', *J. Phon.*, vol. 71, pp. 425–450, Nov. 2018, doi: 10.1016/j.wocn.2018.09.001.

[8] M. Garellek, 'The phonetics of voice 1', in *The Routledge Handbook of Phonetics*, W. F. Katz and P. F. Assmann, Eds., 1st ed.Abingdon, Oxon ; New York, NY : Routledge, 2019. | Series: Routledge handbooks in linguistics: Routledge, 2019, pp. 75–106. doi: 10.4324/9780429056253-5.

[9] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, 'Toward a unified theory of voice production and perception', *Loquens Span. J. Speech Sci.*, vol. 1, no. 1, p. e009, Jan. 2014, doi: 10.3989/loquens.2014.009.

[10] Y. Chai and M. Garellek, 'On H1–H2 as an acoustic measure of linguistic phonation type', *J. Acoust. Soc. Am.*, vol. 152, no. 3, pp. 1856–1870, Sep. 2022, doi: 10.1121/10.0014175.

[11] J. Hillenbrand and R. A. Houde, 'Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech', *J. Speech Lang. Hear. Res.*, vol. 39, no. 2, pp. 311–321, Apr. 1996, doi: 10.1044/jshr.3902.311.

[12] M. Gordon and P. Ladefoged, 'Phonation types: a cross-linguistic overview', *J. Phon.*, vol. 29, no. 4, pp. 383–406, Oct. 2001, doi: 10.1006/jpho.2001.0147.

[13] M. Garellek, 'Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality', *J. Phon.*, vol. 94, p. 101155, Sep. 2022, doi: 10.1016/j.wocn.2022.101155.

[14] I. R. Titze, 'Nonlinear source–filter coupling in phonation: Theory', *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, May 2008, doi: 10.1121/1.2832337.

[15] B. R. Gerratt, J. Kreiman, and M. Garellek, 'Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech', *J. Speech Lang. Hear. Res.*, vol. 59, no. 5, pp. 994–1001, Oct. 2016, doi: 10.1044/2016_JSLHR-S-15-0307.

[16] Y.-L. Shue, P. Keating, and C. Vicenik, 'VOICESAUCE: A program for voice analysis.', *J. Acoust. Soc. Am.*, vol. 126, no. 4, p. 2221, 2009, doi: 10.1121/1.3248865.

[17] K. Sjölander, 'The Snack Sound Toolkit'. KTH Stockholm, Sweden, 2004. Accessed: Dec. 30, 2022. [Online]. Available: https://www.speech.kth.se/snack/

[18] Y.-T. Wang, J. R. Green, I. S. B. Nip, R. D. Kent, and J. F. Kent, 'Breath Group Analysis for Reading and Spontaneous Speech in Healthy Adults', *Folia Phoniatr. Logop.*, vol. 62, no. 6, pp. 297–302, Oct. 2010, doi: 10.1159/000316976.

[19] A. Defazio, F. Bach, and S. Lacoste-Julien, 'SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives', *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[20] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[1] The supplemental materials are available here: https://github.com/uoy-research/pasr-output/tree/main/icphs_23_voicequality.

[2] Confusion matrices of models are presented in the supplemental materials.

[3] Section 4.2 mainly reported on one of the Set S models with the highest test score to demonstrate how to interpret the principal components. Due to limited space, results of other models can be accessible in the supplemental materials.

ON THE ROLE OF GLOTTAL STOP: FROM BOUNDARY MARKER TO CORRELATE OF FOCUS. A STUDY ON ITALIAN AND FRENCH

Bianca Maria De Paolis[1,2], Bianca Abbà[1], Antonio Romano[1]

[1]Laboratorio di Fonetica Sperimentale "Arturo Genre", Università di Torino (Italy)
[2]SFL - Structures Formelles du Langage, CNRS / Université Paris 8 (France)
biancamaria.depaolis@unito.it; bianca.abba@edu.unito.it; antonio.romano@unito.it

ABSTRACT

Cross-linguistically, word-initial or word-final glottalisation is commonly used to mark prosodic boundaries. In Italian and French glottal stops can occur in the presence of boundaries or pauses, and some authors have also stated that they are common before prosodically prominent words. This latter observation suggests a link between informative prominence (linguistic focus) and glottalisation: the aim of our study is to test this hypothesis, by systematically comparing utterances produced in two situations, focus vs background.

The analysis is conducted on a corpus of task-elicited speech produced by two groups, Italian and French (15+15 speakers and 2069 observations in total). We identified the presence of word-initial and word-final glottalisation, and then counted its occurrence in two different informative contexts, focus vs. background. Our data confirmed that glottalisation is much more likely to happen when the phrase is under narrow focus, both in Italian and French.

Keywords: glottalisation, glottal stops, focus, Italian, French

1. INTRODUCTION

Glottalisation of word-initial or word-final segments has been documented as a common marker of prosodic boundaries in several languages across the world [1]. Even though glottal consonants are not part of the phonological inventory of our languages of interest, several acoustic analyses have been conducted on the presence of glottalisation and glottal stops in Romance languages as well [2,3]. In Italian, glottalisation has been observed to occur predominantly at phrase boundaries, and with target vowels bearing stress. In this sense, the phenomenon can be interpreted as a marker of constituent edges, blocking cohesion processes such as *raddoppiamento fonosintattico* and vowel coalescence [4, 5]. In French, similarly, processes like liaison or *enchaînement* reduce the likelihood of a word-initial vowel being glottalised. Nevertheless, [6] found that vowels are more frequently glottalised on boundaries of higher prosodic constituents. In addition to that, some authors have observed that vowels glottalise to a greater degree if the word is pitch-accented (see [7, 8] for English, [9] for French). Pitch accent being the most commonly cited prosodic cue for linguistic focus [10, 11], these observations seem to suggest a link between the three phenomena: focalisation, pitch accent and glottalisation. To our knowledge, though, no study has tested this hypothesis by systematically comparing the two situations, focus *vs* non-focus, and their respective co-occurrence with word-final or word-initial glottalisation, without necessarily taking the presence of pitch accents into consideration. The aim of the present study is therefore to disentangle these interactions, and answer to the following research question: is there a correlation between the occurrence of glottal stops (here also intended as non-canonical realisations, in the form of laryngealisation) and the expression of focus in Italian and French?

2. METHODOLOGY

The analysis is conducted on a corpus of task-elicited speech, containing a total of 870 utterances, produced by 15 French and 15 Italian speakers. The choice to use semi-spontaneous speech is motivated by the fact that focus is a phenomenon that strictly belongs to communicative interaction, and no real communication between speakers can take place in fully controlled speech. The lack of a precise script certainly implies a less than perfect balance between the occurrences of the target structures; despite this, the number of observations that can be made from our data corpus is largely sufficient to enable precise statistics.

2.1. Participants

Participants were 15 native speakers of Italian and 15 native speakers of French. Both groups are mixed, with a majority of female speakers. Italian speakers were recorded in the city of Turin or adjacent area, where they were raised and were currently working or studying. French speakers were recorded in Paris, and they were raised and were living in the region of Paris and Île-de-France.

Group	L1	M	F	Total	Age
FRL1	French	2	13	15	19-32
ITL1	Italian	2	13	15	22-30

Table 1: Profile of participants.

2.2 Procedures and stimuli

The task is taken from [12] and adapted to the needs of our research. Participants are shown a PPT slide containing a picture story and a caption, serving as baseline. After that, they are shown other slides containing the same picture story, accompanied by different questions, which participants are asked to answer aloud. The questions are worded as to elicit three types of focalisation: broad focus (henceforth BF), narrow identification focus (ID), narrow corrective focus (CF). BF is elicited through a broad *wh*-question of the type "What's going on here?". ID is elicited through a *wh*-question targeting each time a different syntactic constituent (subject, object, verb…), e.g. "Who's buying the newspaper?". Finally, CF is elicited through an assertive statement accompanied by a tag-question, aiming at obtaining a corrective answer, e.g. "Mary is buying a book, isn't she?". Stimuli were created in two versions; Italian speakers performed the task in Italian and French speakers in French. The recordings were conducted in a soundproof booth, with a Focusrite Scarlett interface and Shure SM58 microphones.



Figure 1: Examples of stimuli used for the task.

As already mentioned, the corpus contains non-scripted speech: this implies that the number of phonosyntactic vowel clusters (and consequently possible glottal stops) realised by each participant is not identical. On the other hand, utterances are balanced from the informative point of view: the focus type of participants' utterances is controlled through the question asked.

2.3. Identifying glottal stops

Identification of glottal stops and glottalisation is not a straightforward task; several studies have reported huge variability in their acoustic realisation. Having observed the same variation in the realisations described by [3], in this study we follow the coding of glottalisation described by [13], and take into account every realisation of glottal stop presenting at least two of these acoustic cues:

- aperiodicity;
- diplophonia;
- creak;
- silent gap (full glottal closure).

The total of identified glottal stops (henceforth GS) with this method is 376.

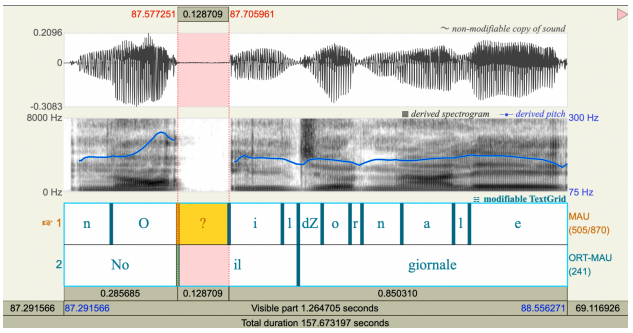


Figure 2: Example of GS between the vowels [ɔ] and [i] in the phrase "No, il giornale" ("*No*, *the newspaper*").

3. PREDICTIONS

As expected from the literature, the presence of a prosodic boundary is a relevant parameter in determining a higher occurrence of glottalisation. For this study, we predict that the informative role of the constituent phrase (focus vs background) is also a relevant parameter: in fact, glottalisation will be more likely to occur on a phrase boundary when the constituent forming the phrase is under narrow (identification or correction) focus. This hypothesised relation could be expressed in the following way:

(1) occurrence of glottal stop:
phrase level < phrase level + focus

4. ANALYSIS

The recorded corpus is composed of 890 utterances, 445 for the French group and 445 for the Italian. We

left out of this analysis 36 ill-formed utterances, namely those presenting main verb elision or internal pauses exceeding 0.2 s. Within this final set of 864 utterances, we identified a total of 2069 vowel clusters situated at possible phrase boundaries, thus potentially leading to the insertion of a GS. Starting from these potential glottalisation points, we have identified the segments where the speakers actually produced a GS. A total of 376 GS were identified and analysed. Vowels located at the beginning of the utterance were excluded from the analysis as well, because their position implied an almost systematic initial glottalisation, regardless of their prosodic position and the informative role they had in the utterance.

We will briefly show the results obtained for each group, Italian and French, and then move to a short comparison between the two languages.

4.1 Italian

The Italian dataset is made up of 434 utterances, within which 1228 vowel clusters at phrase boundaries were observed and 167 GS were detected. Out of these 167, 95 were produced in correspondence of boundaries of focussed constituents. The remaining 72 were produced at the boundaries of all other constituents of the utterance. The raw number and the proportion of GS produced by the speakers in the two contexts (focus vs. non-focus) are shown in the table and graph below.

	Non-focus	Focus
No GS	941	120
GS	72	95

Table 2: Number of occurrences of GS in the two contexts, focus vs non-focus, for the Italian group.

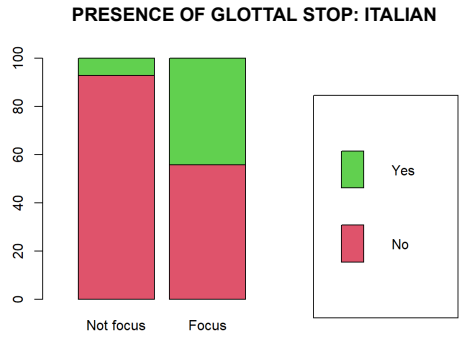


Figure 3: Proportion of occurrences of GS in the two contexts, non-focus vs focus, for the Italian group.

We performed a Pearson's Chi-squared test with Yates' continuity correction to compare the two situations, focus vs non-focus, and the resulting value is $p < 2.2e\text{-}16$. In addition, Bayes' Theorem was applied to calculate prior odds, likelihood ratio and a posteriori probability. The results show a likelihood ratio of 6.28, meaning that we are approximately 6 times more likely to get a GS in a focused environment than a non-focused one. The a posteriori probability is therefore 89%.

4.2 French

The French dataset is made up of 430 utterances, within which 841 vowel clusters at phrase boundaries were observed and 209 GS were detected. Out of these 209, 96 were produced in correspondence of boundaries of focussed constituents. The remaining 113 were produced at the boundaries of all other constituents (background) of the utterance. The raw number and the proportion of GS produced by the speakers in the two contexts (focus vs. non-focus) are shown in the table and graph below.

	Non-focus	Focus
No GS	517	115
GS	113	96

Table 3: Number of occurrences of GS in the two contexts, focus vs non-focus, for the French group.

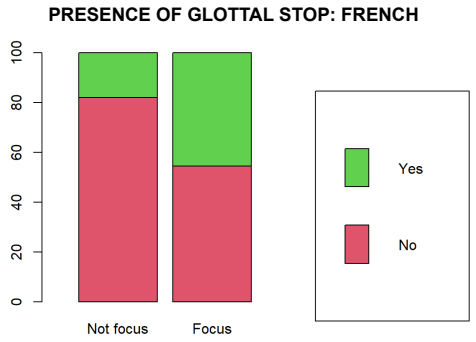


Figure 4: Proportion of occurrences of GS in the two contexts, non-focus vs focus, for the French group.

The Pearson's Chi-squared test with Yates' continuity correction gives a resulting $p = 2.261e\text{-}15$ for the two situations compared, focus vs non-focus. Application of Bayes' Theorem gives in this case a likelihood ratio of 2.68. The a posteriori probability of getting a GS in a focus environment is therefore 69%.

4.3 Interlinguistic comparison

There is a slight difference between the two languages, in which French shows a higher proportion of GS in the non-focus condition than Italian. This is due to the fact that phenomena of vowel coalescence are generally rarer in French than Italian, and vowel clusters are often avoided through various forms of resyllabification like *enchaînement* or *liaison* [14]. Nonetheless, regarding our hypothesis, results are highly significant in both groups.

5. DISCUSSION

Our aim in this study was to test the hypothesis expressed in (1). The results we obtained are highly significant for both groups, so our predictions are confirmed: when a prosodic phrase is under linguistic focus, the probability of observing a GS in correspondence with its boundaries increases considerably. Namely, the increase of probability we could observe is 25% for both groups. This is the first time that the relationship between these two variables has been analysed without it being considered the consequence of other markers of prosodic prominence, e.g. pitch or nuclear accent. The reason why we found interest in the presence of GS in this particular context is that, in our data, it emerges quite strikingly that glottalisation does not necessarily occur in presence of pitch accents: on the contrary, it often surfaces as the only phonetic correlate of *in-situ* focalisation, in absence of major f_0 movements. In our data, GS often happens to be the only distinctive cue between focus and non-focus phrases in an identical environment. As shown in the examples (2) and (3) and their respective spectrograms, the only difference between the non-focus "in edicola" 'at the newsstand' and the contrastive focus "in edicola" is the presence of a GS before the vowel [i] in the second one.

(2) ITL1_sp03:
Che cosa succede qui?
What's going on here?
Maria compra il giornale in edicola.
Mary is buying the newspaper at the newsstand.

(3) ITL1_sp03:
Maria compra il giornale al supermercato, no?
Mary is buying the newspaper at the supermarket, right?
No, Maria compra il giornale [ʔ]in edicola.
No, Mary is buying the newspaper [ʔ]at the newsstand.

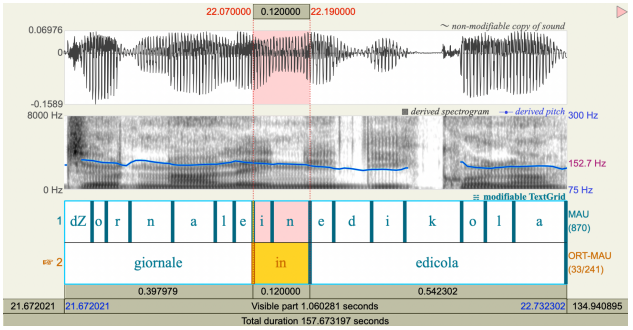


Figure 5: Spectrogram of the phrase "in edicola" ("*at the newsstand*") in a non-focus context as uttered by an Italian speaker (see example 2).

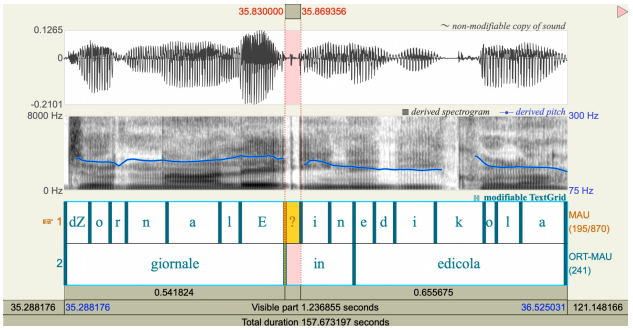


Figure 6: Spectrogram of the phrase "in edicola" ("*at the newsstand*") in a corrective focus context as uttered by an Italian speaker (see example 3).

This goes in the opposite direction than [5, 7, 10, 15]. Another hypothesis on the role of GS has then to be formulated: since it is not the consequence of pitch accent, its presence could be linked to other physiological reasons, see [16]. For example, GS could be functional to the maintenance of subglottal pressure, in order to keep a loud voice at the end of an utterance, or either as a "signal of hard work". Our data support this hypothesis, since the presence of GS is especially visible for focussed phrases in utterance-final position.

In any case, our findings go in the direction of [17] and her definition of "focus as prosodic alignment": in such perspective, the only feature shared by focussed constituents across languages is prosodic alignment to the edge of a prosodic domain, and not prominence intended as nuclear or pitch accent. Italian and French have already been described as "non-plastic" languages from the point of view of focus-induced variation on the intonation level. According to [17], another possible strategy for realising alignment (and then marking focus) is then the insertion or strengthening of a prosodic boundary, which is precisely what Italian and French speakers achieve through the use of GS. It could be the case that both word-initial and word-final GS are used by Italian and French speakers as boundary markers, separating the focused constituent from the rest of the utterance.

7. REFERENCES

[1] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics.* 29, 383–406.

[2] Delattre, P. 1971. Pharyngeal features in the consonants of Arabic, German, Spanish, French and American English. *Phonetica*. 54, 93–108.

[3] Contini, M., Carpitelli, E., Romano, A. 2005. Des occlusives glottales dans l'espace roman. *Estudios Ofrecidos a A. Quilis.* 1, 127–145.

[4] Stevens, M., Hajek, J. 2006. Blocking of word-boundary consonant lengthening in Sienese Italian: some auditory and acoustic evidence. In: Warren, P., Watson, C. (eds), *11th Australasian International Conference on Speech Science and Technology, Proceedings*. Auckland: University of Auckland, 176–181.

[5] Vayra, M. 1994. Phonetic explanations in phonology: laryngealization as the case for glottal stops in Italian word-final stressed syllables. In: Dressler, W. U., Prinzhorn, M., Rennison, J. R. (eds), *7th International Phonology Meeting, Proceedings.* Torino: Rosenberg & Sellier, 275–293.

[6] Fougeron, C. 2001. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*. 29, 109–135.

[7] Pierrehumbert, J., Talkin, D. 1992. Lenition of /h/ and glottal stop. In: Docherty, G., Ladd, D. R. (eds.), *Papers in laboratory phonology, Vol. II: Gesture, segment, prosody*. Cambridge, UK: Cambridge University Press, 90–117.

[8] Cho, T., Keating, P. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics*. 37, 466–485.

[9] Michelas, A., German, J. 2020. Focus Marking and Prosodic Boundary Strength in French. *Phonetica*, 77(4), 244–267.

[10] Avesani, C. 2003. La prosodia del focus contrastivo. Un accento particolare? In: Marotta, G., Nocchi, N. (eds), La Coarticolazione: Atti delle XIII Giornate di Studio del GFS, Pisa, 2002, Pisa: ETS, 157–167.

[11] Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S. A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., Yoo, H. Y. 2015. Intonational phonology of French: Developing a ToBI system for French. In: Frota, S., Prieto, P. (eds), Intonation in Romance. Oxford: Oxford: Oxford University Press, 63–100.

[12] Gabriel, C. 2010. On Focus, Prosody, and Word Order in Argentinian Spanish: A Minimalist Ot Account. *Revista Virtual de Estudos da Linguagem*, 10, 183–222.

[13] Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*. 24, 423–444.

[14] Mertens, P. 2019. *Phonétique, phonologie et prosodie du français*. Leuven: Acco.

[15] van Santen, J., D'Imperio, M. 1999. Positional effects on stressed vowel duration in Standard Italian. *Proc. 14th ICPhS San Francisco,* 241–244.

[16] Lennes, M., Aho, E., Toivola, M., Wahlberg, L. 2006. On the use of glottal stop in Finnish conversational speech. In: Aulanko, R., Wahlberg, L., Vainio, M. (eds), *Fonetiikan päivät - The Phonetics Symposium 2006*. 93–102.

[17] Féry, C. 2013. Focus as prosodic alignment. *Natural Language & Linguistic Theory.* 31 (3), 683–734.

CREAKY VOICE IN THREE SCOTS VARIETIES: USING F0-BASED IDENTIFICATION TO CONSIDER SOCIAL AND LINGUISTIC FACTORS

Joe Pearce

Glasgow University Laboratory of Phonetics, University of Glasgow
j.pearce.1@research.gla.ac.uk

ABSTRACT

While creak in American English is linked to female speakers [1], research on Scots accents using Vocal Profile Analysis (VPA) [2] has found male voices to be creakier than female voices [3, 4]. However, VPA cannot account for within-speaker variation in creak [4]. Here, I consider the potential use of the f0-based method of quantifying creak [5, 6] for examining how social factors (age, gender, area) and linguistic factors affect use of creak in spontaneous speech in Glasgow, Lothian and Insular Scots (n=95).

Creak use was higher among Insular Scots speakers, but showed no difference by age or gender. It was also constrained by linguistic factors (phrase-final position, potential glottal stops, vowel onsets). This work demonstrates how f0-based identification of creak allows consideration of the role of social and linguistic factors; the potential future applications of this method are also discussed.

Keywords: Voice quality, creaky voice, sociophonetics, Scottish English, gender

1. INTRODUCTION

Creaky voice is a mode of phonation characterised by a variety of acoustic characteristics [7], notably low fundamental frequency (f0). This property of some types of creaky voice has been exploited in the f0-based method of automatically identifying creak [5, 6] which codes for creaky voice when f0 drops below a speaker-specific threshold. Here, I consider the potential applications of this method in identifying creaky voice in Scots and Scottish English accents, where Vocal Profile Analysis (VPA) has connected the use of creak with high socio-economic status in male speakers in Edinburgh [8, 9], male speakers in Glasgow [3], and male adolescents in Dumfries, Inverness and Aberdeen [4]. Understanding more about the function of creak requires the ability to consider cases where creak occurs intermittently in more detail than is possible in VPA [4], either due to linguistic factors, or because speakers are using creak to express non-permanent stances, like disengaged affect [10] or toughness[11].

Creak in English is favoured by the presence of glottal stops, which can either be realised with full closure and favour creak in surrounding segments, or be realised without a full closure and manifest as creak in surrounding segments [12, 75] [13]. Glottal stops are common in Scots: In conversational speech, [14] reports a 76% glottalisation rate for /t/ in Glaswegian and [15] reports an 85% glottalisation rate in Edinburgh. Glottals have also been reported in Orkney and Shetland [16, 17]. Creak is also favoured by phrase-final position [18] (see [19] for a review), and phrase-initial vowel onsets [20]. Scottish accents are therefore ideal for exploring the application of this method, allowing us to consider a case where creak is used for both linguistic and social meaning.

Here, I consider creak in data taken from the Scots Syntax Atlas (SCOSYA) [21] in a large sample of younger (approx. 18-25) and older (approx. 65+) male and female speakers from three Scots dialect areas: Glasgow and Lothian, where creak use has been investigated previously, and Insular Scots spoken in Orkney and Shetland, where no previous research has investigated the use of creak. I consider how creak use varies by age, gender and Scots variety, controlling for linguistic factors that favour intermittent use of creak.

2. METHODS

2.1. Corpus

95 speakers from SCOSYA were selected: 19 speakers of Insular Scots (Orkney and Shetland), 29 speakers from Lothian (Edinburgh and surrounding areas), 47 speakers from Glasgow (incl. Greater Glasgow). Speakers were chosen on the basis of an auditory assessment of the level of background noise and recording quality, with noisy and lower quality recordings excluded to improve the results of the forced alignment and subsequent F0 tracking. The varying number of speakers between areas reflect differences in the total available number of each

area, as well as differences in usable recordings. 180 seconds of speech were extracted from each speaker. Time-aligned transcriptions provided by SCOSYA were converted to TextGrids using [22], then force aligned using the Montreal Forced Aligner v.1.1 [23, 24].

2.2. F0-based automatic identification of creak

I followed [6] in using an f0-based method of identifying creak automatically. Files were processed at 16 kHz using MacReaper [6], a drag-and-drop implementation of REAPER [25] for MacOS, to obtain Glottal Closure Instants (GCIs). As [6] notes, REAPER appears to be reliable at detecting GCIs at low and irregular f0, making it well-suited to analysis of creak. Where creak manifested as multiply-pulsed or aperiodic, I found that REAPER tended not to identify all GCIs, resulting in identification of lower f0 and thus classification of creak; however, this is only a passing observation and future work could investigate whether this occurs at scale.

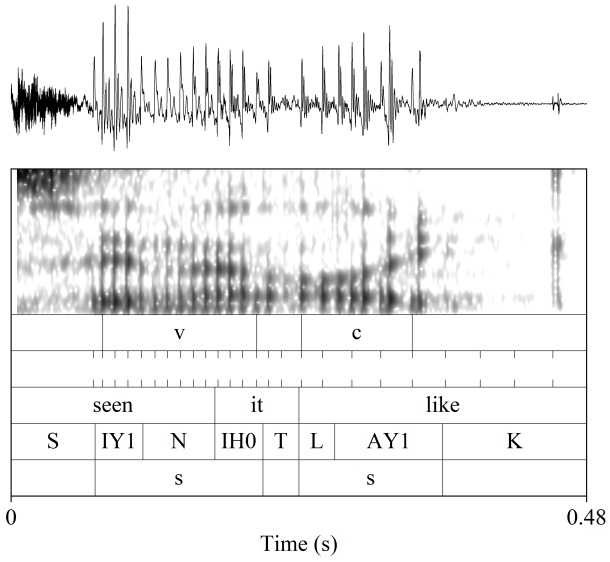


Figure 1: Example of creak identification for Scott, a younger male Glasgow speaker. Tiers show (1) creak identification, (2) GCI detection, (3) words, (4) phones, and (5) initial sonorant stretches

The automated procedure considered only GCIs that occurred within stretches of sonorants. [26] found that this improves the performance of f0-based identification of creak. 52,684 stretches were originally identified, which included 1,188,908 GCIs. Selections were hand-checked for local background noise, constructed dialogue, and major forced alignment errors, which were excluded. After exlusions, 30,792 stretches remained, containing 613,151 GCIs.

Following [6], local f0 was calculated for each GCI by taking the inverse of the time between each GCI within a voiced stretch. Antimodes were then detected using an automated procedure [6] in [27], which `modes`[28](v.0.7.0) to identify the f0 mode of non-creaky speech, the f0 mode of creaky speech, and the antimode between them. Example output is shown in Fig 2. Antimodes were inspected visually for verification. Three speakers were excluded at this stage on the basis that reliable antimodes could not be identified (1 YF Glasgow, 1 OF Glasgow, 1 OF Lothian). 92 speakers remained for analysis.

Speech from the original stretches was then separated into new creaky and non-creaky stretches based on whether local f0 fell below an individual speaker's antimode: In this way, a stretch could either be coded entirely as creaky or non-creaky, or separated into smaller stretches to allow for shorter cycles of creak. An example is shown in Fig. 1: Original stretches are shown on the bottom tier, GCIs on the second tier, and classification as voice or creak on the first tier. I quantified the percentage of creak used by each speaker and group by dividing the total duration of creaky stretches by the total duration of all sonorant stretches. Separation into creaky and non-creaky stretches created 31,698 stretches, of which 26,922 (85%) were not creaky and 4,776 (15%) were creaky. This was equivalent to a total amount of 3,759s (63 mins) of relevant sonorant stretches, of which 340s (9%) was creaky.

2.3. Statistical analysis

The presence of creak was analysed using a binomial generalized linear mixed-effects model in `lme4`[29] (v1.1-28)in R v4.1.2[27]. Effects were added sequentially starting from a minimal model and model fit compared using log-likelihood ratio tests with `anova()`. Fixed effects (reference levels in bold) in the final model were Duration of stretch (log-transformed, scaled), Speech Rate (calculated locally in syllables per second, log-transformed and scaled), potential Glottalisation (**None**, pre-glottalisation context, glottalable context), Phrase-final Position (**Non-final** or Final), Vowel (**Medial vowel**, none, initial vowel, both medial and initial), Contains /l,w,j/ (**No /l,w,j/** or contains /l,w,j/), and Area (**Glasgow**, Lothian, Insular). The final model also included uncorrelated random slopes for Participant by Duration and Speech Rate, and Words by Duration and Speech Rate, where 'Words' was the words that made up the stretch. Fixed-effects that were tested but which did not improve the

fit of the model were as follows: Contains Nasal, Contains Rhotic, Gender (male or female) and Age (older or younger).

Linguistic factors were estimated from the forced aligner output, and so are best understood as proxies for linguistic factors that favour creak. Phrase-Final Position was estimated based on whether the stretch was followed by a pause. Phrase-initial Vowels are coded within the factor Vowel by identifying phrase-initial stretches that began with a vowel. Glottalable contexts are cases adjacent to a /t/, which itself occurs in a context that could be produced as a glottal stop [30, p. 183-184]. Pre-glottalisation contexts were identified as stretches followed by a /p/, /t/, /k/ or /tʃ/ that occurs in a context that could favour pre-glottalisation [30, p. 183-184]. This aimed to account for the fact that /p/ and /k/ can also be glottalised in Glasgow [31].

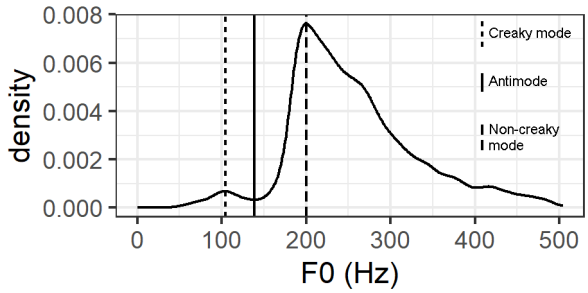


Figure 2: An antimode analysis of the f0 distribution for Alice, a younger female Glasgow speaker

3. RESULTS

3.1. Linguistic factors

Increased voiced stretch Duration and Speech Rate both reduced the log odds of creak (Duration $\beta = -1.121$, $SE(\beta) = 0.054$, $p<0.001$; Speech Rate $\beta = -0.458$, $SE(\beta) = 0.0333$, $p < 0.001$).

By contrast, creak was favoured by Phrase Final position (log(odds) = 0.475, $SE(\beta) = 0.045$, p < 0.001), and a stretch occurring in a Glottalable Context (log(odds) = 0.454, $SE(\beta) = 0.057$, p < 0.001). A stretch being followed by a potential Preglottalisation context did not favour creak. Furthermore, containing an Initial Vowel (log(odds) = 0.660, $SE(\beta) = 0.085$, p < 0.001) and containing no vowel also favoured creak (log(odds) = 1.445, $SE(\beta) = 0.088$, p < 0.001), while containing both a Medial and Initial vowel did not favour creak. Containing /l,w,j/ also favoured creak (log(odds) = 0.177, $SE(\beta) = 0.052$, p < 0.005).

3.2. Social factors

The overall percentage of creak used by each Area, Gender and Age group is shown in Fig. 3. Insular Scots voice quality stands out as particularly creaky: Being an Insular speaker increased the log(odds) of a stretch being creaky by 0.610 ($SE(\beta) = 0.197$, p < 0.01). However, no significant difference was found by Gender or Age.

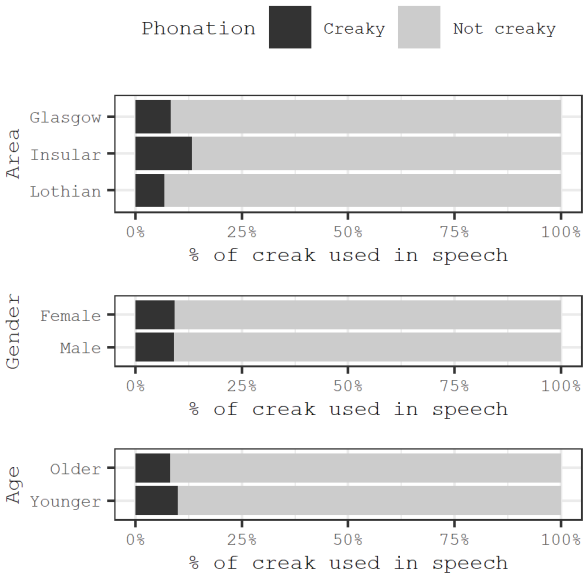


Figure 3: Percentage of creak used across all voiced speech by Area, Gender and Age

4. DISCUSSION

Previous work [5, 6, 32] has shown the potential for f0-based separation of creaky and non-creaky speech on the basis of f0 antimode. Here, this method was used to separate creaky and non-creaky speech, and model how both social and linguistic factors favour the presence of creak in Scots.

4.1. Linguistic variation

Longer stretches were more likely to be identified as non-creaky, showing a limitation of the present analysis: As creak is separated from non-creak, new stretches are created, and these tend to be short if creak is present, as creak is rarely present throughout an entire stretch. This means that it is difficult to make a distinction between speakers who use creak often, but for short periods of time (e.g. for glottals), from those whose creak use may contribute more substantially to their overall voice quality. Future research might consider

using a different unit of analysis: for example, separating speech into 10 ms frames following [26]. Determining what unit of analysis to work with may require considering whether there is a threshold for the presence of creak under which creak does not contribute to the perception of a creaky voice and rather only contributes to the perception of a linguistic phenomenon (e.g. a glottal stop).

There is no surprise here in that vowel onsets, phrase-final position, and potential glottal stops favour creak. However, no previous work appears to have considered the effect of linguistic factors on f0-based estimates of creak. The success of estimating these from the forced aligner rather than hand-coding for each linguistic effect shows the usefulness of f0-based estimation of creak as a coarse-grained method of estimating creak use in large amounts of data, even in spontaneous speech. This makes it possible to distinguish between linguistic uses of creak and potential other reasons, making it useful for sociophonetic research.

4.2. Social factors

Creak was more prevalent in Insular Scots than in Glasgow Scots, while no significant difference was found between Glasgow and Lothian speakers. While [33] find that the use of local Shetland lexical, phonological and morphosyntactic features is lower among younger speakers, they note that the situation is highly complex with different speakers using local forms at different rates. One possible interpretation of the high use of creak in Orkney and Shetland is that creak may be a characteristically local voice quality that younger speakers may be maintaining, while their use of other certain local linguistic forms decreases. Future research could consider differences between Orkney and Shetland, which while grouped together here on the basis of shared historical and phonological characteristics, do form separate varieties, as well and what the function of creak is in Insular Scots.

However, outside of Area, no social factors improved the fit of the model. Unlike [3, 4], I found similar rates of creak between male (9.0%) and female speakers (9.1%). However, as [26] notes, f0-based identification of creak performs better on female speech than male speech; the lack of a higher rate of creak for male speakers could reflect poorer identification of creak in male speakers. I also found similar rates of creak between older (8.1%) and younger (9.9%) speakers. This suggests that in Glasgow and Lothian, creak may serve a more linguistic than social function. However, factors not considered here may be important to consider in future research: Considering the relationship between ethnicity and creak use in other varieties [32, 34] and high creak and high socio-economic status in Edinburgh [8], these social factors may be important in understanding the indexical meanings of creak in Scots accents. Additionally, the present analysis did not consider any of the potential conversational or affective functions of creak, which have been suggested to underlie gender differences in creak found in other varieties (e.g.[10]).

5. CONCLUSION

Work using only a single voice quality measure is often difficult to interpret because measures of Harmonics-to-Noise Ratio and spectral tilt cannot be interpreted in absolute terms in isolation [35], an issue that becomes more apparent when working with corpus data. F0-based categorisation of creak shows promise in this regard: As many forms of creaky voice are characterised by F0 that is low within in a speaker's F0 range, or tracked as such due to F0 irregularity, this method could be used for coarse-grained separation of creaky and non-creaky voice. This could then form the basis of subsequent acoustic analysis of both creaky and non-creaky voice, in turn allowing a more complete view of the types of creak used and where creak fits in a speaker's wider acoustic voice profile.

F0-based estimation of creak has allowed for consideration of how creak use patterns according to both social and linguistic factors, but future research is needed to determine the meaning of creak in Scots beyond relationships with macro-social categories. This present study demonstrates the potential use of the method in sociophonetic research, such as analysing large amounts of spontaneous speech where linguistic context may differ between speakers. It also shows the promise of the method more generally: while this analysis is restricted to macro-level social categories and previously known linguistic constraints, future research could code data for factors such as stance, affect, turn-taking, or topic to gain a closer insight into the indexical, affective and pragmatic meanings of creak in a particular variety or context.

6. REFERENCES

[1] I. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?" *American Speech*, vol. 85, pp. 315–337, 10 2010.

[2] J. Laver, S. Wirz, J. Mackenzie, and S. Hiller, "A perceptual protocol for the analysis of vocal

profiles," *Edinburgh University Department of Linguistics Work in Progress*, vol. 14, pp. 139–155, 1981.
[3] J. Stuart-Smith, "Voice quality in Glaswegian," in *Proceedings of the 14th ICPhS*, 1999, pp. 2553–56.
[4] J. Beck and F. Schaeffler, "Voice quality variation in Scottish adolescents: Gender versus geography," in *Proceedings of the 18th ICPhS*, 2015.
[5] K. Dorreen, "Fundamental frequency distributions of bilingual speakers in forensic speaker comparison," Master's thesis, University of Canterbury, 2017.
[6] K. Dallaston and G. Docherty, "Estimating the prevalence of creaky voice: a fundamental frequency-based approach," in *Proceedings of the 19th ICPhS*, M. T. Sasha Calhoun Paola Escudero and P. Warren, Eds. Canberra, Australia: ASSTA, 2019, pp. 532–536.
[7] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th ICPhS*, 2015.
[8] J. Esling, "Voice quality in edinburgh: A sociolinguistic and phonetic study," Ph.D. dissertation, University of Edinburgh, 1978.
[9] ——, "The identification of features of voice quality in social groups," *JIPA*, vol. 8, pp. 18–23, 1978.
[10] R. Podesva, "The affective roots of gender patterns in the use of creaky voice," in *ETAP 4*, 2018.
[11] N. Mendoza-Denton, *Homegirls: language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell Pub, 2008.
[12] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Blackwell, 1996.
[13] G. Docherty and P. Foulkes, "Sociophonetic variation in 'glottals' in Newcastle English," in *Proceedings of 14th ICPhS, San Francisco*, J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, Eds., 1999, pp. 1037–1040.
[14] J. Stuart-Smith, "Glottals past and present: A study of T-glottalling in Glaswegian," *Leeds Studies in English*, vol. 30, pp. 181–204, 1999.
[15] E. Schleef, "Glottal replacement of /t/ in two british capitals: Effects of word frequency and morphological compositionality," *Language Variation and Change*, vol. 25, no. 2, pp. 201–223, 2013.
[16] P. Sundkvist, ""Standard English" as spoken in Shetland's capital," *World Englishes*, vol. 30, no. 2, pp. 166–181, 2011.
[17] H. Schmitt, "Orkney English phonology: observations from interview data," *Scottish Language*, vol. 34, pp. 58+, 2015.
[18] C. Henton and A. Bladon, "Creak as a sociophonetic marker," in *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin*, L. M. Hyman and C. N. Li, Eds., 1988, pp. 3–29.
[19] L. Davidson, "The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 12, no. 3, pp. 1–18, 2021.
[20] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Prosodic constraints on glottalization of vowel-initial syllables in American English," *JASA*, vol. 95, no. 5, pp. 2978–79, 1994.
[21] J. Smith, D. Adger, B. Aitken, C. Heycock, E. Jamieson, and G. Thoms, "The Scots Syntax Atlas," https://scotssyntaxatlas.ac.uk, 2019.
[22] R. Fromont, "trs2grid.jar," 2022. [Online]. Available: https://sourceforge.net/projects/labbcat/
[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Proceedings of the 18th Conference of the ISCA*, 2017.
[24] ——, "Montreal Forced Aligner [computer program]. v.1.1.0," 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/
[25] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," 2015. [Online]. Available: https://github.com/google/REAPER
[26] H. White, J. Penney, A. Gibson, A. Szakay, and F. Cox, "Evaluating automatic creaky voice detection methods," *JASA*, vol. 152, no. 3, pp. 1476–86, 2022.
[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021.
[28] S. Deevi and D. Strategies, *modes: Find the Modes and Assess the Modality of Complex and Mixture Distributions, Especially with Big Datasets*, 2016, r package v 0.7.0.
[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[30] A. Cruttenden and A. C. Gimson, *Gimson's pronunciation of English*, 8th ed. London: Routledge, 2014.
[31] O. McCarthy and J. Stuart-Smith, "Ejectives in Scottish English: a social perspective," *JIPA*, vol. 43, pp. 273—98, 2013.
[32] A. Szakay and E. Torgersen, "A re-analysis of f0 in ethnic varieties of London English using REAPER," in *Proceedings of the 19th ICPhS*, M. T. Sasha Calhoun Paola Escudero and P. Warren, Eds. Canberra, Australia: ASSTA, 2019, pp. 1675–78.
[33] J. Smith and M. Durham, "A tipping point in dialect obsolescence? Change across the generations in Lerwick, Shetland," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 197–225.
[34] D. Loakes and A. Gregory, "Voice quality in australian english," *JASA Express Letters*, vol. 2, no. 8, p. 085201, 2022.
[35] M. Garellek, "The phonetics of voice," in *The Routledge Handbook of Phonetics*, 2019, pp. 75–106.

WHAT IS GLOTTAL WHISTLE? – EXPLORING EXTREMELY HIGH FUNDAMENTAL FREQUENCIES IN HUMAN VOCAL PRODUCTION

Angela Wingerath[1], Sven Grawunder[1,2]

[1]Martin Luther University Halle-Wittenberg, Halle, Germany; [2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
grawunder@sprechwiss.uni-halle.de

ABSTRACT

Glottal whistle (GW) has been described as a vocal production with very high fundamental frequencies which occur naturally only under extreme conditions. Individual performers are observed to achieve f0-heights of above 1.5kHz and at least up to 6.9kHz, using egressive as well as ingressive airstreams. GW is assumed to be distinct from the whistle (or flageolet) register (M3), whereas its actual production mechanism remains to be unclear. Therefore a vortex mechanism vis-a-vis an extension of the glottal modes (M3 to M4) had been previously proposed. This paper attempts to explore and test these proposals further. Endoscopic data for three subjects reveal different laryngeal configurations on the individual level with in part high degrees of vestibular constriction of a high-positioned larynx. In addition to nonlinear acoustic behaviour in all subjects, for one subject, electroglottographic and acoustic data under air vs. heliox gas conditions seem not to support a vortex-related theory.

Keywords: vocal registers, glottal mechanisms, voice quality

1. INTRODUCTION

1.1. Glottal Whistle

The phenomenon of glottal whistle (GW) had been coined and described by Edgerton et al. [1] as a vocal production type with fundamental frequencies ranging (for both sexes) between 1 and 3kHz. Naturally this type of vocalization occurs under extreme conditions, like child labor or other pain. Aside from this, individual (male and female) vocal performers have demonstrated that the upper limits of the human vocal range can be pushed to enormous extremes: in the example of the singer Demetrio Stratos f0 raised up to 6900Hz (*Cantare la voce*, 1978). In this way, glottal whistles are exerted in the context of vocal performances, either as a peculiar demonstration of vocal range or as one of many *extra normal voices* [2] in the repertoire of so-called multiphonics in avant-garde music. However, these high whistle-like tones are produced on both the inhale and exhale and are usually described by its producers as uncontrollable in pitch. These glottal whistles appear auditorily with an acute (squeaky) timbre, reduced sonority and softer i.e. more narrow loudness ambitus.

1.2. Laryngeal whistling & Whistle register

Unfortunately, the terminology regarding high-pitched voices and whistling-like voice productions used across the literature is often not unambiguous and transparent. Two other similar and partially related phenomena need to be differentiated from glottal whistle, as described here. On the one hand, there is the laryngeal whistling [3, 4, 5, 6] with a pipe-like behavior of a constriction of the larynx entrance, producing whistles ranging between 775Hz and 2500Hz [3]. The lower boundaries and the highly controllable pitch seem here to be indicative [5]. On the other hand, there is the whistle register (also called flageolet, bell or flute register, German *Pfeiffregister*, French *voix de sifflet* or *petit registre*) usually associated with the highest register of the classical female singing voice with frequencies above 1000Hz up to 1397Hz (F6 in Mozart's *Queen of the Night* aria) and higher. The soprano singer Mado Robin used this register up to 2336Hz. Recently other vocal artists (e.g. Georgia Brown, Dimash) have demonstrated to be capable of transitioning into notes beyond such heights (https://youtu.be/7q2oA4sJ3Fs).

1.3. Voice production with high-pitch registers

1.3.1. Vocal fold vibration at high f0

Classifications and systematizations of vocal registers will apart from timbre, frequency range and resonance impressions also consider glottal oscilla-

tion modes. In this respect the whistle register is described as third (M3) glottal mode [7]. Already Garde [8] using stroboscopy had shown that the glottis of the soprano Mado Robin oscillated at least for tones at 1381Hz and at 1550Hz. Echternach [9] was able to provide evidence for such high-pitched glottal activity by means of high-speech videoglottography. Therefore voices produced in the range of glottal whistle proposed by Di Corcia & Fussi [10] as M4, i.e. as an extension of M3 following [7].

1.3.2. Chink Tones

The term *chink tones* is used in analogy to phenomena in wind instruments (*Spalttöne* in German, [11]) as a tone which is produced when air passes a small orifice. The tone results from a periodic vortex shedding as driving force of air particles for inducing a phonation like phenomenon (cf. [11, 12, 13]). This mechanism has been assumed as partially or entirely applicable for some whistling registers in women [4] and children [12]. Tsai et al. [13] suspect a mucosa oscillation induced by vortex shedding in a 4kHz-voice of a Taiwanese singer.

1.3.3. Question & Hypotheses

The main question addressed here concerns if glottal whistles represent the extreme end of the whistle register or some other form of sound production, a distinct vocal register? A proof of glottal activity would be here indicative. This may either be positively confirmed by a corresponding EGG-signal or deviating achieved f0 under differing density conditions of the medium (air vs. heliox). Other aspects concern the homogeneity of the production. As with other extra normal voice phenomena, seemingly similar and associated expressions or performances in individual performers may not have the same voice-mechanical basis.

2. METHODS

2.1. Partcipants & Introspection

For this small scale study in addition to the first author, two other participants were recruited.
M1 – age 31, non-professional singer, experience with heavy-metal singing, experience with glottal whistle for about two months
M2 – age 40, professional singer in the field of experimental improvisation, experience with glottal whistle for several years
F1 – age 48, professional singer in the field of contemporary classical music and experimental improvisation, experience with glottal whistle for several years

According to the singers experiences (mainly here *F1*), when producing a glottal whistle, the singer must find a way to greatly reduce the flow of breath. This is assumed to be caused by a high degree of adduction in the vocal folds and a strong constriction and stiffening of the supra glottal structures. To control the flow of breath, it often helps to alternate between whistling on the inhale and then on the exhale. Another approach to stiffening the tissues is to increase the subglottic pressure. In our case, it became evident that the letter was an approach that *F1* tended to avoid, because of her training in classical singing, and which in turn was more accessible to the other two subjects (*M1, M2*), who were presumably less afraid of the risk of vocal strain.

2.2. Recordings, Endoscopy and Laryngography

For initial recordings (recorded in 2012) a digital recorder (*Roland* R-05) was used for recordings (resolution of 96kHz/24bit, downsampled to 44kHz/16bit).

The descriptive acoustic analysis was carried out with PRAAT [14], wherefore the selected spectrogram range was set to 0-8kHz, window length 0.04s, dynamic range 60dB. The applied pitch algorithm was set to cross-correlation using a range between 1kHz to 4kHz, 0.4 voicing threshold, 0.2 octave cost, 0.47 octave-jump cost and otherwise standard settings.

For video stroboscopy a flexible endoscope from *Xion Medical* (ENDOSTROBE SPECTAR) and EGG (*Xion Medical*) were used. The accompanying EGG would not give reliable results and was therefore left out from analysis.

In a second setting, the audio signals of a phantom-powered head-mounted condenser microphone (AKG C520) and the signals of a two-channel EG-2 electroglottograph (GLOTTAL ENTERPRISE) were acquired by means of a multichannel recorder (788T SOUND DEVICES) at a resolution of 48kHz/16bit.

The commercial helium gas was applied via small conventional rubber balloons. The participants (*M1* and *F1*) were instructed to inhale the balloon content after exhalation and to subsequently attempt a glottal whistle. Participants were then ask to breath normally for several minutes before starting the next attempt (three at maximum for this study). The recordings under heliox condition were made following those under normal (air) condition.

3. RESULTS

3.1. Acoustic Analysis

3.1.1. Fundamental frequencies

Fundamental frequencies appear for all subjects relative fixed and subjectively uncontrollable in pitch height. Nonetheless there is variation in the different settings of this investigation. In the laryngoscopy setting median f0 values for *M1* range between 2.1 and 2.7 kHz and for *M2* between 2.7 and 3.0kHz within 5 sequences of each participant. In the acoustic + EGG setting (air and heliox), *M1* achieves higher f0 namely values between 3.2 and 3.4kHz. (Note that this setting was recorded 6 weeks after laryngoscopy.) The value of both conditions appears as very similar (Air: 3202, 3365, 2960 ; Hel: 3360, 3310, 3290 Hz). *M2* and *F1* demonstrate a long stable GW of up to 15 seconds.

3.1.2. Onset/Offset behavior

All subjects have found a techique to transition (glide) in to the GW.

M1 – achieves short glides in the onset and offset

M2 – downward glides appear especially at the end of GW passages

F1 – glides only appear minimaly in the onset

3.1.3. Nonlinear phenomena

The majority of nonlinear behavior in the recordings are those of sidebands, which are indicated by amplitude modulations (cf. Fig. 1). To a smaller degree there appear short phase of bifurcations into passages with subharmonics, especially with subject *M1*, where GW flip back into a high creak, which he uses as transition for GW. Recordings of *F1* also show phases of biphonation during GW (cf. Fig. 2)

3.2. Laryngoscopic Analysis

Our preliminary endoscopic examination reveals a highly constricted larynx entrance, especially in the mediolateral plane by means of the lateral pharynx walls. Due to the backward movement of the epiglottis during glottal whistle, the tip of the endoscope had to be placed between the epiglottis and the back wall of the pharynx. The laryngeal configurations are described by grid parameters (cf. [15]), following a scale of 0 to 4 (normal to extreme; narrow) in the mediolateral dimension.

For participant *F1* the assessment of distance between VFs, i.e. glottal width <3>, ventricular folds

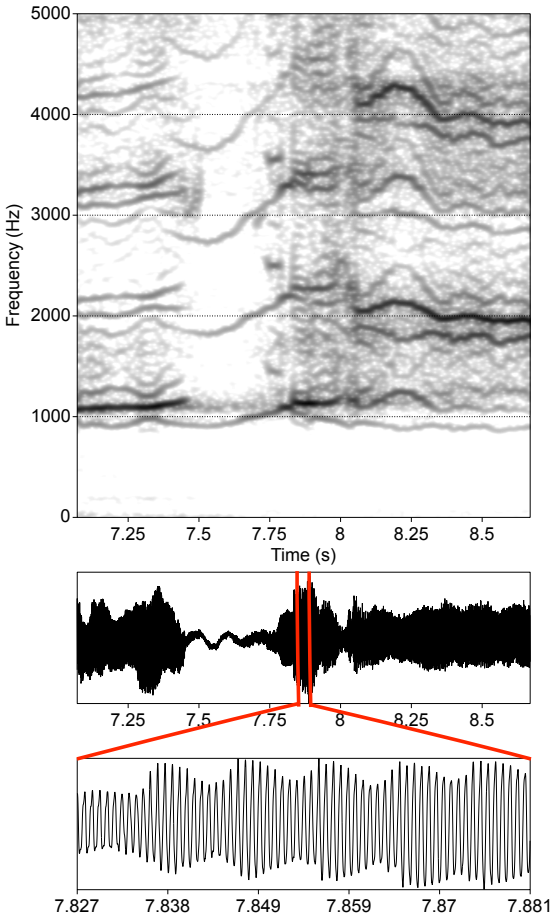


Figure 1: example of egressive GW by *F1*: ambitus hiatus and amplitude modulation (modulation frequency of approximately 110Hz), accompanied by sidebands

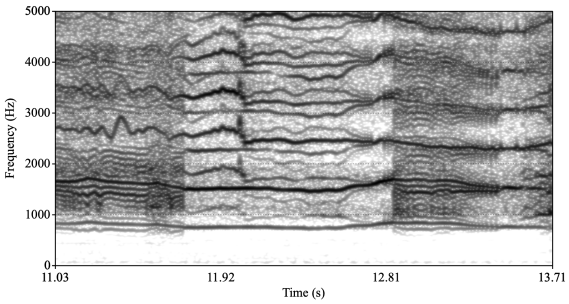


Figure 2: biphonation in passage by *F1*

width <3>, distance between arytenoids apices <3>, epiglottic width as measured the junction with the superior ridge of the aryepiglottic fold on each side <2>; and in the anteroposterior dimension: glottal length <2-3>, epiglottic tip position <2>; cuneiform

fronting <3> (cf. Fig. 3 right column).

Endoscopic footage of participant *M1* shows narrowed pharyngeal walls around the elivated larynx. The narrow ventricular folds width <4> obstructs glottis almost entirely; epiglottic tip position <1>; cuneiform fronting <3> (see Fig. 3 left column).

For participant *M2* occurs a constant obstruction due to a fronting of the arytenoidal apices towards epiglottis (cuneiform fronting <4>) and narrowing of the pharyngeal walls. The so built orifice shows no vibration activity and seems not directly involved in the voice production (cf. Fig. 3 mid column).

In general, all participants show during GW a high positioned larynx with a highly constricted pharyngeal walls, although, observations during practise suggest that this is not essential.

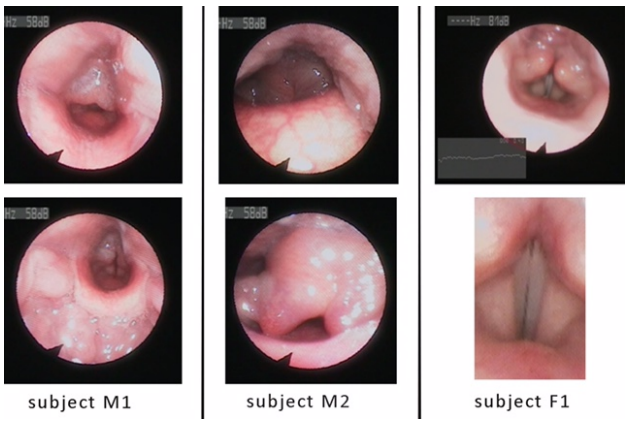


Figure 3: laryngoscopy of GW production by *M1* at 4.5kHz (top) and at 3.3kHz; by *M2* at 3.1kHz ; by *F1* ingressiv at 2.8kHz

3.3. EGG & Heliox

The EGG signal during GW shows in most recordings no detectable periodic behavior, next to clear activity and periodicity in the transition before GW starts (cf. beginning of Fig 4). One remarkable finding is a passage under heliox condition where the same f0 of around 3300Hz can also be detected in the EGG (Fig 4). After initial difficulties achieving a stable GW under heliox *M1* accommodates by higher effort.

4. DISCUSSION

Although the observed laryngeal configurations are reminiscent of a previously inspected 4kHz voice [13] favouring a vortex-like mechanism, our EGG-findings seem to suggest an oscillating behavior at the (close to) glottal level. In addition, the difficulties of *M1* to achieve and maintain the tone

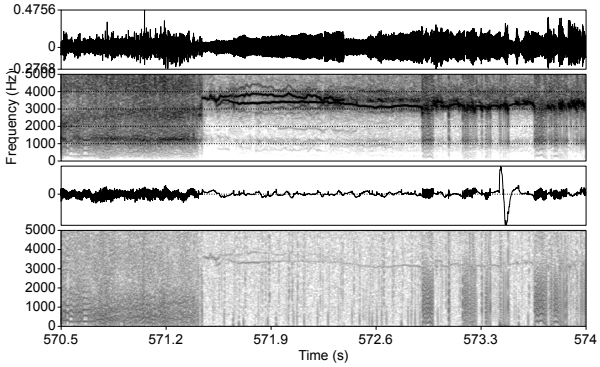


Figure 4: transition from high creak to glottal whistle by *M1* in heliox condition with a clear f0 in EGG spectrogram (lower panel)

under heliox conditions indicate higher transglottal pressure under lower air density conditions. If such difficulties would result from a decoupling (or out of tune) of the higher subglottal formant F1' [16] then this kind of effect would need to be even stronger in their case with M1-M2 transition and longer glottal opening, although it is not. The complex interaction of aerodynamical and biomechanical forces leads additionally to areas of bifurcation and hence a source of biphonation and of subharmonics ([17, 18]).

On the other hand, given the found complete closure in whistle register phonation by means of high-speed video [9], our preliminary results allow to speculate about a possible partial glottis with 'dual appearance', i.e. two separate shorter portions of a medially compressed glottis. Although this would be then also need to be proven undetectable by eletrophysiological (EGG) means, there is no clear indication aside from one instance (see Fig. 4). If this is purely a mechanical effect of coupled neighbouring and due to higher tension / pressure even more narrowed structure as suggested [13], it could perhaps be determined by similar means (ultra sound Doppler image).

5. CONCLUSION

Based on the behavior of the EGG signal, we consider the current findings as indicative though not conclusive for a non-vortex-based (i.e. myoelastic aerodnamic) mechanism as major production mechanism of the glottal-whistle voice. The individual differences hint on the one hand to a more diverse onset behavior and actual expression of the GW tone, i.e. the quality of the voice. The observed approaches of the singers also indicate possible difference between trained and untrained voices.

6. ACKNOWLEDGMENTS

The authors thank prof dr O. König (H Univ., Berlin) for his help in acquiring the laryngoscopic recordings. The authors also thank dr. Christian Herbst (Univ. Vienna) for his helpful advice as well as dr. Matthias Echternach (Univ. Munich) and dr. Coen Elemans (U. of Southern Denmark) for helpful hints and discussions about a prior presentation of the objective of this paper and the anonymous reviewer for her/his comments on the manuscript.

7. REFERENCES

[1] M. E. Edgerton, S. Tan, G. Evans, H. J. Myung, K. K. Bo, F.-Y. Loo, K.-C. Pan, and M. N. Hashim, "Pitch profile of the glottal whistle (m4)," *Malaysian Journal of Science*, vol. 32, no. 1, pp. 78–85, 2013.

[2] M. E. Edgerton, "The extra-normal voice," pp. 728–750, Apr. 2014. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780199660773.013.007

[3] P. Schultz, "Über einen Fall von willkürlichem laryngealen Pfeifen beim Menschen," *Arch. f. Physiol. Suppl*, vol. 523, 1902.

[4] H. Lullies, "Physiologie der Stimme und Sprache," in *Gehör. Stimme. Sprache*, O. F. Rahnke, H. Lullies, and E. . Trendelenburg, F.and Schütz, Eds. Berlin [u.a.]: Springer, 1953, otto F. Ranke.

[5] A. J. Moolenaar-Bijl, "Laryngeal whistle," *Folia Phoniatrica et Logopaedica*, vol. 9, no. 3, pp. 164–168, 1957. [Online]. Available: https://www.karger.com/DOI/10.1159/000262773

[6] J. C. Catford, *Fundamental Problems in Phonetics*. Edinburgh University Press, 1977.

[7] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal vibratory mechanisms: The notion of vocal register revisited," *Journal of Voice*, vol. 23, no. 4, pp. 425–438, Jul. 2009. [Online]. Available: https://doi.org/10.1016/j.jvoice.2007.10.014

[8] E. Garde, "Observation stroboscopique de la vibration des cordes vocales dans le "petit registre" (ou registre "de sifflet") des soprani suraigus," *Folia Phoniatrica et Logopaedica*, vol. 3, no. 4, pp. 248–253, 1951. [Online]. Available: https://doi.org/10.1159/000262518

[9] M. Echternach, M. Döllinger, J. Sundberg, L. Traser, and B. Richter, "Vocal fold vibrations at high soprano fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. EL82–EL87, 2013.

[10] A. Di Corcia and F. Fussi, "Whistle register and m3: A preliminary hsdi investigation by visualization and acoustics in male and female singers," *ePhonoscope*, pp. 267–272, 2016.

[11] F. Krüger and E. Schmidtke, "Theorie der Spalttöne," *Annalen der Physik*, vol. 365, no. 24, pp. 701–714, 1919.

[12] H. Herzel and R. Reuter, "Whistle register and biphonation in a child's voice," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 5, pp. 216–224, 1997. [Online]. Available: https://doi.org/10.1159/000266458

[13] C.-G. Tsai, Y.-W. Shau, H.-M. Liu, and T.-Y. Hsiao, "Laryngeal mechanisms during human 4-khz vocalization studied with ct, videostroboscopy, and color doppler imaging," *Journal of Voice*, vol. 22, no. 3, pp. 275–282, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199706001391

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.2.14, retrieved 1 June 2022 from, 2022. [Online]. Available: https://www.praat.org/

[15] C. Painter, "The laryngeal vestibule, voice quality and paralinguistic markers," *Eur Arch Otorhinolaryngol*, vol. 248, no. 8, pp. 452–8, 1991.

[16] M. Spencer and I. Titze, "An investigation of a modal-falsetto register transition hypothesis using helox gas," *Journal of Voice*, vol. 15, no. 1, pp. 15–24, 2001.

[17] D. A. Berry, I. R. Titze, B. H. Story, and H. Herzel, "Bifurcations in excised larynx experiments," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2930–2930, Nov. 1995. [Online]. Available: https://doi.org/10.1121/1.414146

[18] J. Neubauer, M. Edgerton, and H. Herzel, "Nonlinear phenomena in contemporary vocal music," *Journal of Voice*, vol. 18, no. 1, pp. 1–12, Mar. 2004. [Online]. Available: https://doi.org/10.1016/s0892-1997(03)00073-0

CEPSTRAL PEAK PROMINENCE IN NORMOPHONIC ADULTS: THE EFFECT OF GENDER AND SPEECH TASK SEGMENTAL COMPOSITION

Saoirse O'Regan & Irena Yanushevskaya

Trinity College Dublin
oregans1@tcd.ie, yanushei@tcd.ie

ABSTRACT

Cepstral Peak Prominence (CPP) is currently recommended for clinical use by speech and language therapists as an objective and robust measure of dysphonia. Its widespread use in clinic is stalled, however, by the relative conceptual complexity of CPP and by the lack of baseline normative data. This paper describes CPP values extracted from the audio recordings of normophonic adults 25-55 years old living in Ireland. Audio data was collected remotely, using mobile phones, from 42 participants (21 female). Speech tasks included sustained vowels as well as sentences with different segmental composition (vowels and approximants, nasals, voiced and voiceless obstruents). CPP values were extracted using Praat. We describe the effects of gender, speech task segmental composition and CPP extraction method (with and without voice activity detection) on the obtained CPP values. The results may serve as a normative baseline for clinical voice analysis and assessment in speech and language therapy practice.

Keywords: Cepstral Peak Prominence, normative data, gender, speech task, clinical voice analysis.

1. INTRODUCTION

Clinical voice analysis and evaluation is a multidimensional task and includes both subjective (auditory) and objective (instrumental) methods of assessment [1, 2]; it is expected that at least one objective measurement is included to support auditory assessment [3]. Cepstral Peak Prominence (CPP) is currently recommended by American Speech-Language-Hearing Association (ASHA) for use in clinic as the most robust and reliable measure of dysphonia [3]. Since CPP was first proposed as a measure of breathiness in the voice signal [4, 5], there has been growing interest in the clinical applications of CPP, e.g. [6, 7] and references therein. However, its use by clinicians is still not as widespread as the more traditional measures (jitter, shimmer and HNR). For example, a survey conducted by [8] found that while participant speech and language therapists (SLTs) working with voice consistently used acoustic analysis in voice assessment, none of them reported using CPP. The likely explanation is that CPP is conceptually more complex than commonly used acoustic measures: as stated in [9], 'CPP shares with other cepstral measures the lack of an intuitive interpretation relative to the underlying physiology of vocal fold vibration'. CPP values are influenced by recording conditions (e.g., using mobile phone for remote data collection), segmental composition of speech tasks [6, 10], the choice of software applications [11, 12], analysis settings such as window type and length [13] and the use of additional speech processing, e.g., voice activity detection (VAD) [7]. SPL and speaker f_0 are also contributing factors [13, 14]. Importantly, there is lack of comprehensive baseline normative CPP data by gender and age that would facilitate the interpretation of CPP in clinical voice assessment.

In this paper we describe CPP values extracted using Praat [15] from the audio recordings of various speech tasks (isolated vowels, connected speech) obtained from a sample of normophonic adults 25-55 years old living in Ireland. The data was collected and analysed in an effort to contribute to the growing body of information about normative CPP values which clinicians could use as a baseline when working with disordered voice clients in speech and language therapy clinics and other health care settings.

2. CEPSTRAL PEAK PROMINENCE AND ITS CLINICAL APPLICATIONS

Power cepstrum of a signal equals to the Fourier transform of the logarithm of its power spectrum [9]. Periodic signals will show energy at harmonically related frequencies in their spectrum, and their cepstrum will have a prominent peak at the time ('quefrency') corresponding to the fundamental period of the signal. Cepstral Peak Prominence (CPP) measures, in dB, the magnitude of the cepstral peak relative to a linear regression line fitted to the cepstrum to normalise the prominence of the cepstral peak to the overall amplitude of the signal [4, 5]. CPP is a measure of the periodicity of the spectrum of a

speech signal rather than the periodicity of the signal per se [10, 13].

Over the past years, CPP has been extensively studied, e.g. a comprehensive review in [9] or [6] on clinical applications of CPP. It is well established that CPP is a more robust measure compared to traditional voice quality measures such as jitter, shimmer and HNR: its calculation does not require accurate pitch extraction algorithms and it can be reliably used in the analysis of both sustained vowels and connected speech even if recorded in a non-controlled environment [16, 17]. While CPP is a robust predictor of overall dysphonia severity, it may not perform well discriminating individual voice quality dimensions, such as roughness [10, 18].

Various factors may influence CPP values. Speaker gender and speech task effects have been reported in [19], e.g., significantly higher CPP values were reported in sustained vowels produced by males than by females, but the trend was opposite in sentences. Typically, CPP values in sustained vowels are higher than in connected speech, e.g. summary of studies in [6]. [20] reported higher CPP values for open vowels /ɑ/ and /æ/ than for close vowels /i/ and /u/. The impact of segmental composition of the speech task on CPP was examined by [6]. Their findings suggest that CPP values in CAPE-V sentences varied depending on the amount of voiced segments: the all-voiced sentence 'We were away a year ago' yielded the highest CPP and the sentence with voiceless aspirated plosives 'Peter would keep at the peak' the lowest. The presence of nasals has an effect of lowering CPP [10, 13, 21].

CPP values extracted from the same audio file may differ depending on the software used [12, 22], e.g., Praat [15] and Analysis of Dysphonia in Speech and Voice (ADSV, PENTAX Medical), which have different approaches to voice activity detection [6]. The use of voice activity detection presents a bit of a conundrum: If VAD is not used, the presence of voiceless segments in the speech samples selected for the analysis may artificially lower CPP. If VAD is used and the analysed voice is dysphonic with a lot of aperiodicity, the resulting CPP based on relatively small sections of data will be artificially inflated and not representative of the speaker's voice. The solution proposed in [6] is to use all voiced sentences or to remove voiceless segments prior to the analysis.

Many clinical studies looked at correlations of CPP and auditory-perceptual evaluation of disordered speech and normal controls and at the correlations of CPP with the auditory-perceptual ratings using CAPE-V [23] or GRBAS [24], e.g. reviews in [6, 9]. CPP was found to reliably separate normal and disordered voices and to highly correlate with auditory-perceptual ratings. Several studies aimed to establish CPP cut-off values for both sustained vowels and connected speech, but the findings are varied. Cut-off CPP values in [6] are 11.46 dB (ADSV) and 14.45 dB (Praat) for the sustained /a/ vowels and 6.11 dB (ADSV) and 9.33 dB (Praat) for the Rainbow Passage [25]; they detected the presence of voice disorders with over 90% accuracy.

As mentioned earlier, there are limitations to the use of CPP in clinical settings due to the lack of comprehensive normative data that would assist clinical voice evaluation. One can infer these data from CPP values obtained from healthy controls in the clinical studies; however, there are differences in the reported findings. Studies conducted specifically to collect CPP normative data are still scarce, and the obtained results also vary, e.g., [19, 26].

In this study CPP values were extracted using Praat from the audio recordings of isolated vowels and connected speech obtained from a sample of vocally healthy English speaking adults 25-55 years old living in Ireland. The ultimate goal was to provide normative CPP values which SLTs could use as a baseline when working with disordered voice clients in speech and language therapy clinics. Of particular interest are the effects of speaker gender, speech task and CPP extraction method on CPP values.

3. MATERIAL AND METHOD

Participants The study was approved by Research Ethics Committee; informed consent was obtained from all participants prior to data collection. Forty-two vocally healthy participants (21 female) 25-55 years old living in Ireland, all English speaking, were recruited. Participants had no self-reported history of voice disorders or smoking. Participants' age by gender distribution is shown in Fig. 1.

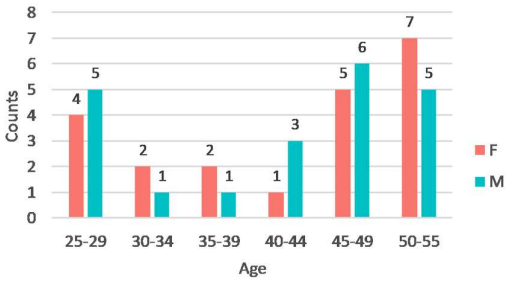


Figure 1: Recorded speakers by age and gender.

Speech task and recording procedure The speech tasks used here were those commonly used in clinical voice assessments and in studies on CPP [6, 12, 22]. The tasks included vowels [a] and [i] sustained for five seconds, CAPE-V [23] sentences with different segmental composition: 'We were away a year ago'

(mainly vowels and vowel-like articulations), 'A man may mean many[1]' (vowels and nasals), 'How hard did he hit him?' (word-initial [h]),'Peter will keep at the peak' (voiceless aspirated stops), and the first two sentences from 'The Rainbow passage' [25]. In the discussion that follows we use shortened names for these tasks.

Participants recorded speech tasks in a quiet environment of their convenience using the default audio recording apps on their mobile phones (iPhone Voice Memos and Android Voice Recorder). They then sent their recordings to the researchers by email. To protect the identity of the participants, the audio files were anonymised. Analysis of data collected remotely, using mobile phones, provides valuable information to facilitate voice assessment via telepractice, whether during the recent pandemic or to accommodate people living in remote areas.

CPP data extraction The obtained M4A files were converted to WAV using Audacity® [27] at the default sampling frequency 48 kHz. The files were segmented in Praat so that each speech task was saved as its own WAV file. An informal auditory analysis was conducted by the authors to ensure that the audio files were of adequate quality and did not contain ambient noise or evidence of atypical phonation. There were 294 sound files in total (42 participants x 7 speech tasks [2 vowels, 4 sentences, 1 passage]).

CPP values were extracted automatically using the Praat plug-in described in [7]. The plug-in allows to extract smoothed CPP using the default Praat settings following [4] which have also been used in [6, 11, 12, 14] and other papers. The plug-in includes the method for extracting CPP with voice activity detection (VAD) performed as the first step. We used both VAD and 'traditional' noVAD extraction methods.

4. RESULTS AND DISCUSSION

The distribution of CPP values by speech task, gender and extraction method (VAD, noVAD) is shown in Fig. 2; to facilitate by-gender and by-method comparison the same data are plotted twice. Table 1 gives median, mean and 95% CI values (due to space constraints, for noVAD method only). As can be inferred from Fig. 2, CPP values get lower for both male and female speakers with the increase of the proportion of voiceless segments, particularly voiceless plosives, in the speech task. The sentence where 39% of segments are voiceless stops ('Peter will keep at the peak') showed the lowest CPP. Generally, it appears that male speakers yield lower mean CPP in most tasks (except sustained vowels) compared to female speakers in both extraction methods.

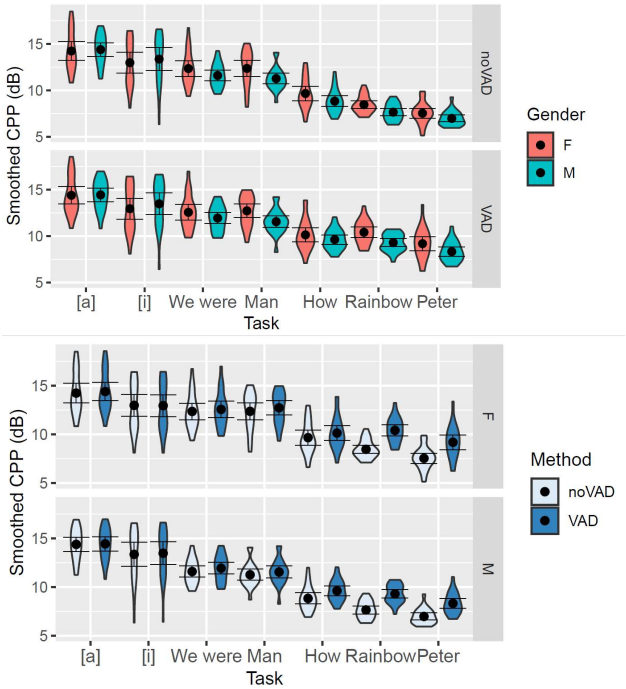


Figure 2: The distribution of CPP values by speech task, gender and extraction method. Black dots and whiskers show mean and 95% CI.

Gender	Task	CPP Median	CPP Mean	95% CI LL–UL
Female (n = 21)	[a]	13.86	14.23	13.24–15.23
	[i]	12.98	12.98	11.85–14.10
	We were	12.09	12.35	11.51–13.19
	Man	13.07	12.36	11.48–13.24
	How	9.18	9.66	8.90–10.43
	Rainbow	8.29	8.47	8.04–8.89
	Peter	7.53	7.54	7.02–8.07
Male (n = 21)	[a]	14.42	14.39	13.64–15.13
	[i]	14.47	13.36	12.13–14.59
	We were	11.46	11.60	11.03–12.17
	Man	11.15	11.27	10.70–11.84
	How	8.66	8.84	8.27–9.42
	Rainbow	7.68	7.66	7.25–8.06
	Peter	6.88	6.99	6.63–7.35

Table 1: CPP median, mean and 95% CI values. Due to space constraints, only noVAD method results are shown.

As it was of interest to establish to what extent speaker gender, the type and segmental composition of speech task and CPP extraction method have an effect on CPP values, we conducted linear mixed-effect model analyses. Our initial model included CPP values as the dependent variable, gender, speech task and extraction method as well as their interactions as the main predictors (fixed effects); random effects included by-subject random intercepts and slopes to account for speaker variability in speech tasks:

[1] Not a CAPE-V sentence.

CPP~Gender*Task*Method+(1+Task|Participant). Analyses were conducted in R environment [28] using *lme4* [29] and *lmerTest* packages [30] for model fitting (using maximum likelihood method) and step-down model simplification by eliminating non-significant effects and interactions. The final model was CPP~Gender+Method+Task+Method:Task+(1+Task |Participant). *sjPlo*t package [31] was used for model visualization. Post-hoc pairwise comparisons were conducted using the *emmeans* package [32] with Tukey correction for multiple comparisons. The results of the mixed model analyses are shown in Table 3; Fig. 4 shows model predicted CPP values.

Predictors	*Est.*	*CI*	*t*	*p*
(Intercept)$_{[F, noVAD, a]}$	14.81	14.17 – 15.44	**45.67**	**<0.001**
Gender$_{[M]}$	-0.98	-1.49 – -0.48	**-3.80**	**<0.001**
Method$_{[VAD]}$	0.10	-0.09 – 0.30	1.04	0.300
Task$_{[How]}$	-5.06	-5.66 – -4.46	**-16.54**	**<0.001**
Task$_{[i]}$	-1.14	-1.68 – -0.60	**-4.14**	**<0.001**
Task$_{[Man]}$	-2.49	-3.08 – -1.91	**-8.41**	**<0.001**
Task$_{[Peter]}$	-7.04	-7.61 – -6.48	**-24.35**	**<0.001**
Task$_{[Rainbow]}$	-6.25	-6.75 – -5.75	**-24.62**	**<0.001**
Task$_{[We_were]}$	-2.34	-2.89 – -1.78	**-8.20**	**<0.001**
Method$_{[VAD]}$× Task$_{[How]}$	0.52	0.25 – 0.80	**3.69**	**<0.001**
Method$_{[VAD]}$× Task$_{[i]}$	-0.06	-0.34 – 0.22	-0.42	0.669
Method$_{[VAD]}$× Task$_{[Man]}$	0.22	-0.06 – 0.50	1.53	0.128
Method$_{[VAD]}$× Task$_{[Peter]}$	1.39	1.11 – 1.67	**9.78**	**<0.001**
Method$_{[VAD]}$× Task$_{[Rainbow]}$	1.69	1.41 – 1.97	**11.91**	**<0.001**
Method$_{[VAD]}$× Task$_{[We_were]}$	0.17	-0.11 – 0.45	1.21	0.225
ICC		0.92		
N $_{Participant}$		42		
Observations		588		
Marginal R^2/Conditional R^2		0.651/0.973		

Table 2: Results of the mixed effect model analyses of the CPP data: coefficients and 95% CI (fixed effects only).

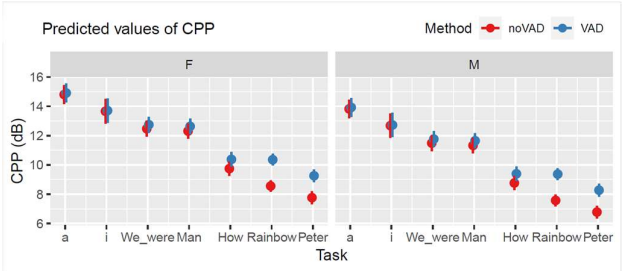


Figure 4: Model predicted CPP values in different speech tasks by gender and extraction method.

The effects of gender, speech task and extraction method (marginal R^2) explain 65% of the data variance; combined fixed and random effects account for about 97% of the variance (conditional R^2). While the general trend is overall similar for male and female speakers, the results suggest significant effect of gender ($t = -3.80$, $p < 0.001$), with CPP values for male speakers being about 1 dB lower than those of female speakers. Unsurprisingly, our results suggest that CPP values are affected by Task. Similar to [20], CPP values of the [i] vowel are significantly lower than those of the [a] vowel for both male and female speakers. CPP values of 'We were' and 'Man' are significantly lower still, and the values get progressively (and significantly) lower as the number of obstruents and voiceless segments in the utterances increases, with the 'Peter' sentence showing the lowest CPP. However, we found significant Method:Task interaction (Fig. 4.). Extraction method has no significant effect on CPP in sustained vowels, but for other speech tasks $CPP_{[VAD]}$ values were significantly higher than $CPP_{[noVAD]}$, the effect gets larger as the number of the voiceless segments in the speech tasks increases.

5. CONCLUSIONS

The paper reported the analysis of audio data (sustained vowels and connected speech) collected with mobile phones from normophonic English speaking adults aged 25-55 living in Ireland. The ultimate goal was to contribute to the information on normative CPP values to facilitate the use and interpretation of CPP by SLTs in clinical voice assessment. The CPP normative values obtained in our study are similar to the ones reported in [6, 33] and the findings generally support earlier studies. Our data suggest significantly higher CPP for female speakers. Provided that the analysis settings are as described in [7], the choice of speech task emerged as another important factor for CPP analysis. CPP extraction method does not seem to matter for the analysis of sustained vowels; in all other speech tasks $CPP_{[VAD]}$ is higher than $CPP_{[noVAD]}$, similar to [6]. This is an important consideration in a busy clinic where pre-processing of the data (i.e. exclusion of unvoiced segments) is not desirable nor feasible unless fully automated.

The findings need to be confirmed on a larger sample and with parallel recording done in optimal conditions to compare mobile phone data with. A potential follow-up study would involve clinicians working with disordered voice clients who would submit mobile phone samples over the course of therapy to be evaluated using CPP. Developing clear guidelines for clinicians, incorporating normophonic data by age, gender, speech task, recording conditions and extraction method in intuitive and easy-to-use tools for clinicians will facilitate a wider use of CPP by speech and language therapists in clinical voice assessment.

6. ACKNOWLEDGEMENTS

The second author acknowledges the support of the Faculty of Arts, Humanities and Social Sciences, Trinity College Dublin (Research Fellows Award).

7. REFERENCES

[1] P. H. Dejonckere *et al.*, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," (in eng), *European Archives of Oto-Rhino-Laryngology,* vol. 258, no. 2, pp. 77-82, 2001.

[2] P. H. Dejonckere, "Assessment of voice and respiratory function," in *Surgery of Larynx and Trachea*, M. Remacle and H. E. Eckel Eds. Berlin Heidelberg: Springer-Verlag, 2010, pp. 11-26.

[3] R. R. Patel *et al.*, "Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function," *American Journal of Speech-Language Pathology,* vol. 27, no. 3, pp. 887-905, 2018.

[4] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research,* vol. 37, no. August, pp. 769-778, 1994.

[5] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech," *Journal of Speech and Hearing Research,* vol. 39, pp. 311-321, 1996.

[6] O. Murton, R. Hillman, and D. Mehta, "Cepstral Peak Prominence values for clinical voice evaluation," *Am. J. Speech Lang. Pathol.,* vol. 29, no. 3, pp. 1596-1607, 2020.

[7] E. S. Heller Murray, A. Chao, and L. Colletti, "A practical guide to calculating Cepstral Peak Prominence in Praat," (in eng), *J. Voice,* 2022.

[8] S. McAlister and I. Yanushevskaya, "Voice assessment practices of speech and language therapists in Ireland," *Clinical Linguistics & Phonetics,* pp. 1-25, 2019.

[9] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control,* vol. 14, pp. 42-54, 2014.

[10] C. A. Ferrer Riesgo, M. S. de Bodt, Y. Maryn, P. Van de Heyning, and M. E. Hernández-Díaz Huici, "Properties of the cepstral peak prominence and its usefulness in vocal quality measurements," in *MAVEBA*, Florence, Italy, 2007, pp. 93-96.

[11] Y. Maryn and D. Weenink, "Objective dysphonia measures in the program Praat: Smoothed Cepstral Peak Prominence and Acoustic Voice Quality Index," *Journal of Voice,* vol. 29, no. 1, pp. 35-43, 2015.

[12] C. R. Watts, S. N. Awan, and Y. Maryn, "A comparison of cepstral peak prominence measures from two acoustic analysis programs," (in eng), *J Voice,* vol. 31, no. 3, pp. 387.e1-387.e10, 2017.

[13] C. A. Ferrer Riesgo and E. Nöth, "What makes the Cepstral Peak Prominence different to other acoustic correlates of vocal quality?," (in eng), *J Voice,* vol. 34, no. 5, pp. 806.e1-806.e6, 2020.

[14] M. Brockmann-Bauser, J. H. Van Stan, M. Carvalho Sampaio, J. E. Bohlender, R. E. Hillman, and D. D. Mehta, "Effects of vocal intensity and fundamental frequency on Cepstral Peak Prominence in patients with voice disorders and vocally healthy controls," *Journal of Voice,* vol. 35, no. 3, pp. 411-417, 2021.

[15] *Praat: doing phonetics by computer [computer program]*. (2023). Accessed: 10/02/2023. [Online]. Available: http://www.praat.org/

[16] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: a meta analysis," *Journal of the Acoustical Society of America,* vol. 126, no. 5, pp. 2619-2634, 2009.

[17] B. Barsties v. Latoszek, Y. Maryn, E. Gerrits, and M. De Bodt, "A Meta-analysis: Acoustic measurement of roughness and breathiness," *Journal of Speech, Language & Hearing Research,* Article vol. 61, no. 2, pp. 298-323, 2018.

[18] S. Y. Lowell, R. H. Colton, R. T. Kelley, and S. A. Mizia, "Predictive value and discriminant capacity of cepstral- and spectral-based measures during continuous speech," *Journal of Voice,* vol. 27, no. 4, pp. 393-400, 2013.

[19] C. Batthyany *et al.*, "A case of specificity: How does the Acoustic Voice Quality Index perform in normophonic subjects?," *Applied Sciences,* vol. 9, no. 12, p. 2527, 2019.

[20] S. N. Awan, A. Giovinco, and J. Owens, "Effects of vocal intensity and vowel type on cepstral analysis of voice," (in eng), *J Voice,* vol. 26, no. 5, pp. 670.e15-20, 2012.

[21] C. Madill, D. D. Nguyen, K. Yick-Ning Cham, D. Novakovic, and P. McCabe, "The impact of nasalance on Cepstral Peak Prominence and Harmonics-to-Noise Ratio," *The Laryngoscope,* vol. 129, no. 8, pp. E299-E304, 2019.

[22] C. Sauder, M. Bretl, and T. Eadie, "Predicting voice disorder status from smoothed measures of Cepstral Peak Prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV)," *Journal of Voice,* vol. 31, no. 5, pp. 557-566, 2017.

[23] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," (in eng), *Am J Speech Lang Pathol,* vol. 18, no. 2, pp. 124-32, 2009.

[24] M. Hirano, *Clinical Examination of Voice*. New York: Springer Verlag, 1981.

[25] G. Fairbanks, *Voice and articulation drillbook*. New York: Harper & Row, 1960.

[26] K. V. Phadke, A.-M. Laukkanen, I. Ilomäki, E. Kankare, A. Geneid, and J. G. Švec, "Cepstral and perceptual investigations in female teachers with functionally healthy voice," *Journal of Voice,* vol. 34, no. 3, pp. 485.e33-485.e43, 2020.

[27] *Audacity® software: free audio editor and recorder [computer application]*. (1999-2021). [Online]. Available: https://audacityteam.org/

[28] *R: A language and environment for statistical computing*. (2021). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.r-project.org/

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1-48, 2015.

[30] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software,* vol. 82, no. 13, pp. 1-26, 2017.

[31] *sjPlot: Data Visualization for Statistics in Social Science*. (2021). Accessed: 27/04/2023. [Online]. Available: https://CRAN.R-project.org/package=sjPlot

[32] *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.5*. (2023). . Accessed: 27/04/2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[33] J. Delgado-Hernández, N. León-Gómez, and A. Jiménez-Álvarez, "Diagnostic accuracy of the Smoothed Cepstral Peak Prominence (CPPS) in the detection of dysphonia in the Spanish language," *Loquens,* vol. 6, no. 1, p. e058, 2019.

VARIATION AND DIALECTAL DIFFERENCES IN THE SEGMENTAL SCOPE OF PHONOLOGICALLY CONTRASTIVE LARYNGEALISATION

Andrea Siem

Lancaster University
a.siem@lancaster.ac.uk

ABSTRACT

Phonological contrasts in voice quality are assumed to be phonetically restricted to the segment(s) on which they are contrastive. However, in the suprasegmental stød (pronounced [støð̰]) in Danish, a contrastive laryngealisation akin to creaky voice, substantial variation occurs in the segmental span of non-modal phonation. In many cases it is continuously realised on the adjacent segment to the contrastive segment and may span up to four segments. This paper presents initial findings of this variation based on acoustic data from 21 native Danish speakers of the Copenhagen (10) and Aarhus dialect (11). It shows that variation in segmental span correlates significantly with the dialect of the speaker and the phonetic realisation of the contrast. Thus, this paper contributes novel findings on the possible quantitative scope of phonetic variation in phonological voice quality contrasts and the role of such variation in dialectal differences as well as individual speaker idiosyncrasy.

Keywords: Phonology, dialectal phonetic variation, creaky voice, Danish stød.

1. BACKGROUND

1.1. Segmental span of non-modal phonation

In languages that utilise non-modal phonation to create phonemic contrasts the change in voice quality is generally restricted to one segment, the vowel. Indeed, the span of non-modal phonation has even been found to be less than the full vowel, e.g. contrastive breathy voice spanning on average 43% of the vowel in Jalapa Mazatec [1]. This lack of full segmental span further persists when comparing languages with phonemic non-modal phonation (Mazatec, Chong, Mpi) and phonetic non-modal phonation (Tagalog) where only a few of the breathy vowels have phonation changes spanning their full duration [2]. Other non-modal phonation types such as contrastive laryngealisation, however, may span the full vowel [2], although percentages were not provided for this type. In a study on Yalálag Zapotec contrastive laryngealisation was realised in a model-nonmodal-modal pattern [3], again spanning only a portion of the vowel as seen with breathy voice in [1], [2]. The limiting factors on non-modal phonation segmental span are suggested to be (i) co-occurrence with other contrasts such as tone, making full vowel non-modal phonation shorter to allow for successive modal vibration that carries tonal information, (ii) possibly greater articulatory effort associated with the production of non-modal phonation and (iii) co-articulatory demands of surrounding segments such as plosives [2]. Whatever the phonological or phonetic limitations may be, however, it appears that a consistent pattern in the segmental span of contrastive non-modal phonation is restriction to one segment and most often less than the full segment. This paper provides data on the contrastive laryngealisation *stød* in Danish that seemingly defies this trend on a proportion of tokens. First some background is provided on the *stød* followed by an account of instances of post-segmental creaky voice and dialectal differences in the occurrence of these instances.

1.2. The Danish *stød*

The *stød* is a laryngealisation that characterises certain syllables which fulfil the phonetic requirement of primary or secondary stress and an adequate stretch of voiced segments called *stød basis* [4]. The *stød basis* is either a long vowel or a short vowel and an adjacent sonorant consonant. The *stød* is contrastive in minimal pairs like ‘ven’ [vɛn] *friend* vs ‘vend’ [vɛnˀ] *turn (imperative)* or can distinguish word classes in minimal pairs like ‘løber’ [ˈløːb̰ɐ] *(a) runner* vs ‘løber’ [ˈløːˀb̰ɐ] *(is) running*. However, not all words with *stød* have a non-*stød* minimal pair contrast and the rules for when a syllable with *stød basis* do not receive *stød* are complex and will not be elaborated here – see e.g. [5] for an overview.

The *stød* is an interesting phenomenon in relation to the segmental scope of phonological voice quality contrasts because it is classified as a syllabic prosody, [5] rather than a segmental contrast, indicating that its scope is the entire syllable, but the timing of the phonetic realisation of the *stød* has its onset roughly in the middle of the syllable [6] and is said to belong to a specific segment in syllables where the *stød basis* consists of more than one segment. Indeed, not just the presence of *stød* is contrastive but there are near-

minimal pairs that contrast based on which segment receives the *stød*. Compare 'lån' [ˈlɔːˀn] *loan* vs 'lund' [ˈlɔnˀ] *grove* where the long vowel in 'lån' receives the *stød* but in 'lund' the nasal receives the *stød*. Further, the phonetic realisation of this contrast varies between the dialects spoken in the two largest cities of Denmark, Copenhagen and Aarhus. Where the Copenhagen dialect realises both instances with laryngealisation akin to creaky voice, the dialect of Aarhus only laryngealises if the *stød* is received by the consonant as a general rule. If the *stød* falls on the vowel, it is instead realised as a long, falling tone, [lɔ̂ːn] instead of [ˈlɔːˀn], leading to what perceptually sounds like a rising-falling bi-tonal pattern [7]. This type of potential *stød* realisation will henceforth be referred to as tonal *stød*, given its phonetic realisation and perceptual properties in the Aarhus dialect, not because I claim that the *stød* is generally a laryngealisation caused by underlying tone as suggested by other scholars – a topic dealt with in [8]. The term tonal *stød* is thus used to characterise the *stød* that occurs in monosyllables with a long vowel and a sonorant consonant coda. Phonetically the tonal *stød* is laryngealised in Copenhagen but tonal in Aarhus.

2. METHODS

The data was not collected to investigate the segmental scope of laryngealisation during the *stød:* however, the presence of creaky voice after the *stød* was something I discovered during the annotation of the data in Praat [9] and decided to mark on a separate tier as it seemed surprisingly prevalent. Thus, this paper is an initial account of the phenomenon, henceforth referred to as post-*stød* creak, and provides a brief quantification of its occurrence. An in-depth phonetic, qualitative acoustic study is outside the scope of this short paper.

The data was collected as part of a PhD research project on phonetic variation in phonological voice quality contrasts and only the parts of the study setup relevant to the present paper are described here. The data consists of minimal pairs of two conditions: (i) no *stød* vs standard *stød* (25 pairs) and (ii) regular *stød* vs tonal *stød* (8 pairs). The pairs were embedded in full sentences read from a laptop screen presented and recorded in SpeechRecorder [10]. Many minimal *stød* pairs are homographs meaning context was needed for participants to know which word was prompted and for this reason carrier phrases were not a feasible option. Different sentences were used for each word instead and all sentences appeared three times in random order. The target word was never the ultimate or penultimate in a sentence to control for potential effects of utterance-final creaky voice to confound the presence of *stød*. Despite this, some tokens had to be discarded from final analysis due the presence of creaky voice on non-*stød* tokens or mispronunciations of the target word. Participants were recruited via social media and consisted of 10 Copenhagen speakers (5 male, 5 female) and 11 Aarhus speakers (6 male, 5 female) aged 21-43 (avg. 30.14). Criteria for inclusion were that the speaker had no vocal pathologies and had grown up in their respective dialect area and not lived outside of it for more than 2 years at the time of recording.

Post-*stød* creak, henceforth PSC, was marked in cases where creak was clearly audible and the waveform showed irregular vibrations, if relevant supported by visible vertical striations in the spectrogram, see Figure 1:

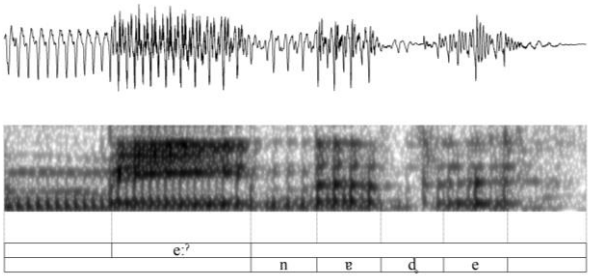


Figure 1: Praat picture of the utterance 'mener det', *(I) mean it,* spoken by a male Copenhagen speaker. The waveform is on top, the spectrogram just below, the stød segment is annotated on Tier 1 and the PSC segments are annotated on Tier 2.

Figure 1 also exemplifies the most extreme case of post-segmental PSC span found in the data. It is possible to produce laryngealisation without any irregular vibration or audible creaky voice [11] but for an initial account of the phenomenon, only segments with audible creak were included. It is possible that the amount of post-stød laryngealisation is larger if acoustic measurements are used to capture additional instances of non-creaky laryngealisation. For the analysis all non-*stød* tokens were excluded as only *stød* tokens are candidates for the presence of PSC. The final number of tokens for analysis of each type per dialect is given in Table 1 below.

3. PREVALENCE OF POST-STØD CREAK AND DIALECTAL DIFFERENCES

Table 1 provides an overview of the prevalence of PSC for each dialect and *stød* type calculated as a percentage of overall tokens of the respective types. As is evident the phenomenon is most prevalent in the Copenhagen dialect, in particular on tonal *stød* tokens

where more than half of the segments receiving *stød* are followed by PSC.

Dialect	PSC tokens regular stød	PSC tokens tonal stød
Copenhagen	194/893 **(21.7%)**	132/225 **(58.7%)**
Aarhus	41/1003 **(4.1%)**	13/249 **(5.2%)**

Table 1: Overview of the percentage of tokens with PSC out of total number of tokens with *stød* stratified by dialect and *stød* type.

There is, however, large inter-speaker variability with some speakers having no PSC and others having PSC on almost all tonal tokens, see Figure 2 and 3:

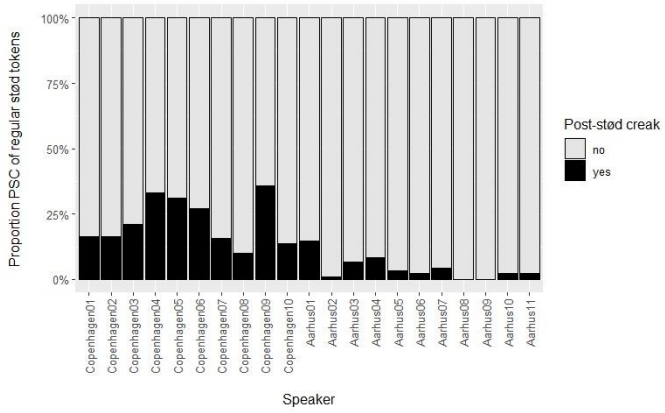


Figure 2: Bar graph showing the proportion of PSC tokens for each individual dialect speaker as a percentage of the total number of regular stød tokens.

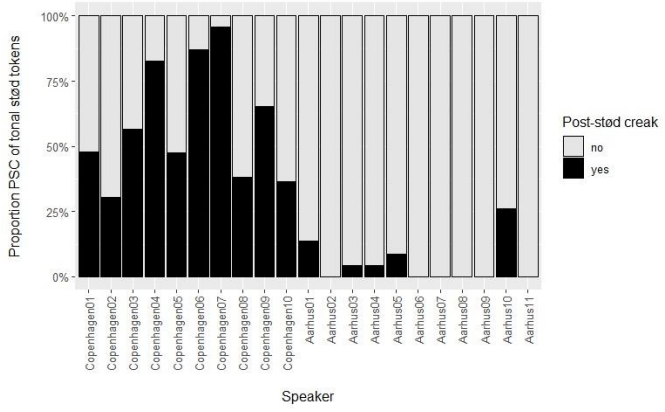


Figure 3: Bar graph showing the proportion of PSC tokens for each individual dialect speaker as a percentage of the total number of tonal stød tokens.

To test the significance of these observed differences a binomial logistic regression model (Generalised Linear Mixed Model, GLMM) was fitted in R [12] for two conditions: (i) the effect of dialect on PSC occurrence and (ii) the effect of *stød* type on PSC occurrence. For each condition a full GLMM was fitted which modelled PSC as a function of stød type + dialect, with speaker and word as random effects. The full model was compared to a reduced model excluding the fixed effect of interest, i.e. dialect for the first condition, *stød* type for the second condition. The effect of speaker dialect on PSC occurrence is significant (χ^2[1]=26.31, p=<0.05) which is to be expected based on the relatively large differences in prevalence observed in Table 1. The overall effect of *stød* type on PSC occurrence is also significant (χ^2[1]=9.52, p=0.002), an effect seemingly driven by the relatively large differences in occurrence between the two *stød* types in the Copenhagen dialect.

4. SEGMENTAL SCOPE

Having quantified the overall prevalence of PSC above the attention will now be turned to variation in the number of segments after the *stød* that are produced with PSC. As evident from the percentages in Table 2 below the overwhelming trend is that PSC occurs on the first segment after the *stød*, a phenomenon observed at least once in all but two speakers.

Dialect	PSC1	PSC2	PSC3	PSC4
Copenhagen	331/1118 **(29.6%)**	33/1118 **(3.0%)**	3/1118 **(0.3%)**	1/1118 **(0.09%)**
Aarhus	54/1252 **(4.3%)**	22/1252 **(1.8%)**	2/1252 **(0.2%)**	0/1252 **(0.0%)**

Table 2: Overview of tokens with PSC on segments 1-4 after the *stød* calculated as a percentage of the total number of tokens with *stød* stratified by dialect.

PSC on the second segment after the *stød* is much less prevalent and PSC on the third and fourth segment is very rare, the latter occurring only on one token in one Copenhagen speaker, see example in Figure 1. Further, the differences in prevalence between dialects decrease the further from the *stød* PSC occurs. It is difficult to make claims about conditioning factors for the production of PSC as the data was not elicited specifically to investigate the phenomenon and thus lack control for many factors that could influence its presence such as sentence

intonation, phonetic environment, stress placement, etc. No single word in the word list consistently has PSC so even though phonetic environment is likely to influence PSC there are clearly other factors to consider in future research. What can be concluded from the initial results here is that both dialect and type of *stød* correlate significantly with the presence of PSC, but with fairly large inter-speaker variability, suggesting that PSC has a degree of idiosyncrasy outside the realm of phonological contrast. The dialectal differences were most prevalent for PSC on the adjacent *stød* segment, much less so for when the segmental scope was larger than one adjacent segment. The fact that non-modal phonation was found to persist beyond the contrastive segment was a surprising finding defying the general trend in languages with this type of contrast and thus worth reporting on.

5. DISCUSSION

In phonetic terms there is nothing remarkable about laryngealisation being followed by creaky voice on the adjacent segment as an effect of coarticulation. However, as demonstrated here, the patterns of PSC suggest that this is more than an effect of co-articulation as it does not occur on all tokens and is significantly correlated with speaker dialect and type of *stød*. Further, it is cross-linguistically unusual for a phonologically contrastive laryngealisation to span more than (half of) one segment, the contrastive status seemingly restricting the segmental span to maximise the perception of the contrast. In regard to the Danish *stød* it is perhaps not unusual to find cases where the vowel receives *stød* and the sonorant consonant coda receives post-*stød* creaky voice considering that the *stød* is generally described as a syllabic prosody and the coda is part of the syllable. But recall that although the *stød* is a syllabic prosody in Danish phonology there are contrastive near-minimal pairs differing in segmental *stød* occurrence rather than syllabic *stød* occurrence, e.g. 'lån' [ˈlɔːˀn] *loan* vs 'lund' [ˈlɔnˀ] *grove*, conditioned by the length of the vowel. Given that these long vowel+sonorant consonant combinations are what produces the tonal *stød* condition and that the prevalence of PSC was statistically significant as a function of *stød* type it could be that the usual restrictive segmental scope of non-modal phonation observed in other languages is less restricted on these tokens because the vowel length already provides salient cues to the contrast, lessening some of the perceptual load from the *stød*. This, however, does not account for why PSC also occurs on regular *stød* tokens where the minimal pairs differ *only* in *stød* with no vowel length difference, e.g. 'låner' [ˈlɔːˌnɐ] *(a) borrower* vs 'låner' [ˈlɔːˀˌnɐ] *borrows*. The PSC here may just be a coarticulatory effect but as it does not occur very consistently on the same words across speakers coarticulation alone cannot sufficiently account for PSC.

A surprising finding was that PSC occurs on tonal *stød* tokens for the Aarhus speakers as these tokens have been found to be phonetically realised with a long, falling tone [7] rather than laryngealisation or creaky voice, meaning no PSC would be expected. It may be a few artifacts of the general effect observed with low tones where the vocal folds start to vibrate in more irregular cycles if the tone is particularly low, which would partly explain why this does not occur on all tonal tokens but only a few – these may have tones ending below a certain threshold where regular vocal fold vibration can no longer be sustained. It could also be an enhancement of the contrast where the tone is perceived as lower if it is produced with creaky voice on the adjacent segment. However, as creaky voice in the Standard Modern Danish dialect *is* the contrast, this seems less plausible. It is entirely possible that the presence of PSC is predicted by sociophonetic factors too, apart of course from the factor of dialect already found here, and given the individual variation evident in Figure 2 and 3, the idiosyncrasy of PSC occurrence could also be of interest to forensic phonetic research on Danish.

6. CONCLUSIONS

The aim of this paper was to provide an initial quantitative account for the phenomenon labelled post-*stød* creak occurring after the Danish laryngealisation *stød* as an example of how phonologically contrastive non-modal voice quality can span over more than one segment, an unexpected observation from research on other languages. Data were compared for two dialects, Copenhagen and Aarhus, and two types of *stød* conditions were included, regular *stød* and tonal *stød*. Post-*stød* creak prevalence was found to be significantly correlated with both dialect and *stød* type, being most prevalent on tonal *stød* tokens in the Copenhagen dialect. The segmental span of PSC was primarily the segment adjacent to the *stød* but could span up to four segments after the contrastive segment in one case. More research is necessary to determine the conditioning factors for PSC but this initial account shows that the Danish *stød* defies the general segmental restriction of contrastive laryngealisation and inspires more research into its conditioning factors, sociophonetic status and individual speaker idiosyncrasy relevant for forensic phonetic research on Danish and beyond.

7. REFERENCES

[1] Silverman, D., Blankenship, B., Kirk, P. & Ladefoged, P. 1995. Phonetic structures in Jalapa Mazatec. *Anthropological Linguistics* 37(1), 70–88.

[2] Blankenship, B. 2002. The timing of nonmodal phonation in vowels. *Journal of Phonetics* 30, 163-191.

[3] Avelino, H. 2010. Acoustic and Electroglottographic Analyses of Nonpathological, Nonmodal Phonation. *Journal of Voice* 24(3), 270-280.

[4] Grønnum, N. & Basbøll, H. 2001. Consonant length, stød and morae in Danish. *Phonetica* 58(4), 230-253.

[5] Basbøll, H. 2003. Prosody, productivity and word structure: the stød pattern of Modern Danish. *Nordic Journal of Linguistics* 26(1), 5-44.

[6] Fischer-Jørgensen, E. 1989. Phonetic Analysis of the Stød in Standard Danish. *Phonetica* 46, 1-59.

[7] Kyst, B. 2007. Toner i århusiansk regiolekt. In Torben Arboe (red.): *Nordisk dialektologi og sociolingvistik*. Fællestrykkeriet, Aarhus Universitet, 238-247.

[8] Grønnum, N., Vazquez-Larruscaín, M. & Basbøll, H. 2013. Danish Stød: Laryngealization or Tone? *Phonetica* 70, 66-92.

[9] Boersma, P. & Weenink, D. 2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.3.03, retrieved 30 December 2022 from http://www.praat.org/

[10] Draxler, C. & Jänsch, K. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. Published in *International Conference on Language Resources and Evaluation*, Lisbon, Portugal. Developed for Bavarian Archive for Speech Signals, Institut für Phonetik, Ludwig-Maximilians-Universität München, Germany, available for download at https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

[11] DiCanio, C. 2009. The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association* 39, 162–188.

[12] R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Interaction between Prosodic Position and Obstruent Voicing: A Case of English Lettered Word in Wenzhou Wu Chinese

Weijun Zhang, Peggy Pik Ki Mok

The Chinese University of Hong Kong
zhangweijun@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

The use of English initialisms (e.g. VIP, henceforth 'English lettered words') in Chinese has increased substantially, with the phonetic form adapted into phonological systems of Chinese languages. Obstruents in Wu Chinese contrast phonologically in voicing, whereas true phonetic voicing is normally found word-medially (similar to English plosives), while in word-initial positions, the voicing contrast of Wu Chinese is realised as a phonation contrast of the following vowel. Given the similarity in prosodic position-dependent realisation of voicing in the two languages, this study focuses on the English lettered words in Wenzhou Wu to investigate whether the phonetic nature varies in prosodic positions as a result of loanword adaptation. The results show that the distribution of true voicing seems to be an effect of the phonetic environment only, regardless of the existence of prosodic boundaries, and the realisation of vowel phonation contrast is dependent on the syllable's position in the prosodic hierarchy.

Keywords: voicing, vowel phonation type, prosodic position, Wu Chinese, loanword adaptation

1. INTRODUCTION

1.1. English lettered words

In modern Chinese languages, the number of words partially or purely consisting of English letters has increased substantially in recent decades [1], and an individual section 'words with letters of Western alphabets' is now included in the authoritative *The Contemporary Chinese Dictionary* [2]. These English lettered words include 1) acronyms, e.g. APEC, Wi-Fi, 2) initialisms, e.g. GDP, ICU, and 3) hybrids of English letters with native Chinese morphemes or numerals, e.g. T *xùshān* ('T-shirt'), PM2.5. This study focused on the latter two types of English lettered words only.

Concerning the phonological form of the English lettered words in Chinese, normally the English pronunciation is used for the letters in initialisms and hybrid lettered words [3]. Though the specific phonetic form of the English letters is likely adapted to the Chinese phonological system – for instance, one of the tones in the inventory would be assigned to each letter – the phonological adaptation is still based on the English pronunciation of the letters for most Chinese languages.

However, Wenzhou Wu is a special case among Chinese languages in that the phonetic form of English letters is not merely the adapted form pronounced in accented English. Instead, our pilot investigation showed that there were a set of localised pronunciations. As in the examples of letters pronounced with the /iː/ vowel in English such as B, D, P, T, they are pronounced with a diphthong /ei/ instead, although /bi/, /di/, /pʰi/ and /tʰi/ are all well-formed syllables in Wenzhou Wu. It may have resulted from historical reasons as a vowel shift of /i/ to /ei/ could be observed comparing the current vowel inventory with the documentation in the first Wenzhou Wu textbook in English, *Introduction to the Wênchow Dialect* published in 1893 [4, 5]. Given that the Romanised script of Wenzhou Wu invented by Christian missionaries was widely used in Wenzhou, it could be inferred that the English letters had been localised since then and were treated equally as native Chinese lexical items in sound change in the past century.

1.2. Prosodic position and voicing realisation

Wu Chinese still preserves the voicing contrast of obstruents in syllable-initial position phonologically, which has been lost in almost all other Chinese languages. However, recent phonetic research on some dialects of Wu Chinese such as Suzhou and Shanghai Wu found that the phonologically defined 'voiced' obstruents are only phonetically voiced in intervocalic positions, while in initial positions the contrast is realised as different phonation types of the following vowels – breathy vowel after phonologically voiced consonants and modal vowel after phonologically voiceless ones, while this vowel phonation contrast does not exist in intervocalic positions [6, 7, 8]. Described as 'intervocalic position' in the literature, only word-medial obstruents were investigated and potential phonetic effects of prosodic hierarchy were usually ignored. A previous study on native Wenzhou words shows a significant age-dependent variation, and the phonetic

realisation of voicing in other Wu dialects mentioned above is applicable for the younger speakers, while the vowel phonation feature may not be consistently observed in the elders' speech [9].

As for English plosives, it is well acknowledged that prevoicing is uncommon in initial position and voicing seems to have little perceptual importance, while the phonetic voicing contrast exists more consistently in medial position [10, 11, 12]. Since there is no vowel phonation contrast like that in Wu Chinese in English, the English 'voiced' plosives followed by a modal vowel in initial position are phonetically highly similar to the Wu syllables beginning with voiceless unaspirated plosives. Given the similarity and difference mentioned above in prosodic position-dependent realisation of voicing in the two languages, it is interesting to investigate how the phonetic realisation of phonological voicing is adapted in language contact. The current study does not limit the 'intervocalic' position to prosodic word level, but includes larger prosodic units to examine the phonetic nature of obstruent voicing in English lettered words in Wenzhou Wu occurring in various prosodic positions.

2. METHOD

2.1. Participants

Since the prosodic position-sensitive realisation of voicing mentioned above seems either unstable or inconsistent among older speakers, only younger speakers were recruited for the current study [9]. Six native speakers (3 male and 3 female) aged 27 to 30 years old participated in the experiment. All participants were Wu-Mandarin simultaneous bilingual speakers born, raised and now live in Wenzhou downtown. The parents of the informants were also all native speakers and Wenzhou Wu is the only language used in the family.

2.2. Prosodic phrasing and experiment stimuli

The current study examined the obstruent voicing in four different prosodic positions in English lettered words as follows, and all positions are exemplified in Table 1.

2.2.1. Absolute initial position

The English lettered words in the absolute initial position (with preceding silence), which could reflect the adapted phonological form of English letters in isolation in Wenzhou Wu, were included and the parameters extracted from the absolute initial (#initial) position could serve as a baseline for comparison with other prosodic positions.

Among the 26 English letters in Wenzhou Wu pronunciation, the eligible candidates for this study with a phonologically voiced obstruent initial include B /bei^{31}/, D /dei^{31}/, G /dʑi^{31}/, J /dʑei^{31}/, V /vei^{31}/ or /ʋei^{44}/ and Z /zei^{31}/. As there is no English lettered word found with an initial J or Z, and the initial consonant of V seems to fluctuate between fricative and approximant, only the three letters, B, D and G, were involved in forming stimuli for the experiment. In other word, this study focused on two phonologically voiced plosives and one affricate.

2.2.2. Prosodic word-medial position

As mentioned above, according to previous research on Wu voicing contrast, the phonological voicing contrast is realised as true phonetic voicing in this position. For Wu Chinese speakers, prosodic word is a prominent unit in their intuition because of its role as the bearing unit of lexical sandhi tones [13]. Some phonologists also proposed that the lexicalisation of short syntactic phrases in Wu Chinese is phonologically marked by the assignment of lexical sandhi tone [14]. The English lettered words, as one single lexical sandhi domain, with B, D, G at medial position were hence included as a part of the stimuli.

2.2.3. Prosodic phrase-medial / word-initial position

Prosodic phrase in Wenzhou Wu is also a prosodic unit with transparent phonological marker proposed as 'phrasal tone sandhi' or 'tone neutralisation' in different phonological analyses [14, 15, 16]. Verb-object constructions are normally typical prosodic phrases with the monosyllabic verb reduced to zero (atonic). This type of experiment stimuli was hence formed by simply combining a monosyllabic verb and the stimuli used for #initial position to form verb-object constructions.

2.2.4. Prosodic phrase-initial position

To form the experiment stimuli for intervocalic prosodic phrase-initial position, the aforementioned stimuli for #initial position were embedded in a carrier sentence 'ŋ^{24} tse^{42} dɤu^{0} i^{0} pi^{42} __ (I read once again __)'. In this case, neither lexical nor phrasal tone sandhi is possible to be applied across the prosodic boundary between 'i^{0} pi^{42} (once)' and the targeted syllable, that is, the targeted syllable cannot be prosodically connected to the preceding word.

2.2.5. Experiment design

In summary, the stimuli used for the current study included three English letters, B /bei^{31}/, D /dei^{31}/ and

G /dʑi³¹/, in four prosodic positions, and there were two tokens for each letter in each prosodic condition, i.e., 24 unique tokens in total. The participants were asked to read the experiment stimuli shown in a random order twice. Simultaneous acoustic and elctroglottography (EGG) signals were recorded using Glottal Enterprises EG2-PCX EGG.

#Initial	**B**RT (bus rapid transit), **D**VD, **G**DP
PW-medial	A**B**C ('basic facts'), 3**D** (film), 4**G**
PP-medial	ma⁰ **D**VD ('to buy a DVD')
PP-initial	ŋ²⁴ tse⁴² dɤu⁰ i⁰ pi⁴² **B**RT ('I read BRT once again')

Table 1: Examples of experiment stimuli ('PW' for prosodic word; 'PP' for prosodic phrase).

2.3. Parameters and data processing

Since the phonological voicing contrast of initial obstruents in native Wenzhou Wu words involves phonetic voicing and vowel phonation type as aforementioned, we would focus on these two phonetic properties in the current study as well. Regarding voicing, the phonetic correlate to be extracted is the voice onset time (VOT).

As for the EGG signal, in this study, the time interval between the peak value of the derived curve of EGG signal (dEGG) and the time point when EGG signal exceeds the 25% threshold was designated as the contact phase [17, 18, 19]. The parameter of contact quotient (CQ), the ratio of the contact phase to the whole cycle, was then calculated. The CQ data were extracted from the vowel, with time normalised as having 10 time points, and the Smoothing Spline ANOVA model was used to test the statistical significance of the difference between CQ curves. The result could be interpreted as statistically significant (at a level of $\alpha = 0.05$) if the different shaded areas do not overlap.

3. RESULTS

3.1. VOT analysis

To figure out the effect of prosodic position on the distribution of true phonetic voicing, the VOT values of the two plosives /b/ in letter 'B' and /d/ in 'D' in different prosodic positions were examined.

As shown in Figure 1, the distribution pattern of negative VOT in English lettered words seems to be an effect of the #initial/intervocalic distinction only, which does not concern the specific prosodic structure. The phonologically voiced plosives were phonetically voiced in all intervocalic positions, no matter whether there was a prosodic boundary at the word or phrase level involved. In the #initial position, the phonologically voiced plosives were in fact voiceless with short positive VOT (12.27 ms for /b/ and 12.62 ms for /d/ on average).

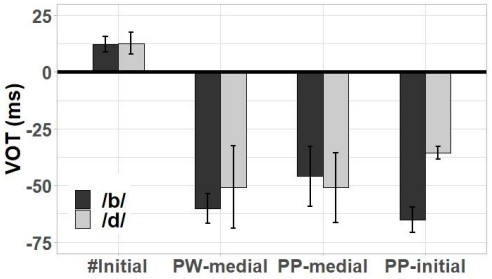


Figure 1: VOT of /b/ and /d/ in different positions.

3.2. Vowel phonation analysis

Different from native Wu syllables, there are no phonologically voiceless unaspirated obstruents contrasting with the phonologically voiced ones in English lettered words. To explore the potential vowel phonation feature existing in these words with obstruents in different prosodic positions, the #initial position (the basic form in isolation) data was used as a baseline in comparison.

3.2.1. Prosodic word-medial position

Consistent with previous findings on native Wu words, the voicing of the obstruents in prosodic word-medial positions in English lettered words is also realised in the following vowels. Figure 2 shows that the vowels following obstruents in #initial and prosodic word-medial positions were significantly differentiated by their phonation type. The CQ of the vowels in prosodic word-medial syllables was higher at least for the first half of the syllable duration, supporting that the vowels in word-medial syllables were less breathy compared to the ones in #initial positions.

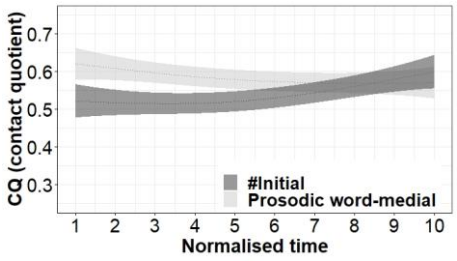


Figure 2: CQ of vowels in #initial and prosodic word-medial syllables.

3.2.2. Prosodic phrase-medial / word-initial position

Though a prosodic word boundary is introduced into the middle of this type of constructions (prosodic phrase), the pattern of CQ values was similar to the

vowels in word-medial syllables (Figure 3). The vowel phonation of this prosodic position is even more differentiated from the vowels following #initial position obstruents. Hence, vowel phonation (breathiness) still served as an articulatory correlates of the phonological voicing of its preceding obstruents in prosodic phrase-medial position.

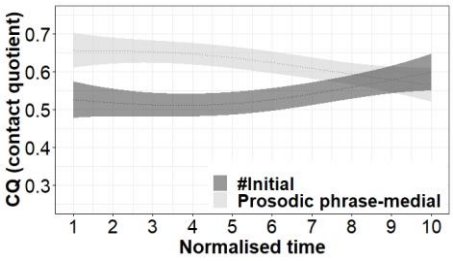


Figure 3: CQ of vowels in #initial and prosodic phrase-medial / word-initial syllables.

3.2.3. Prosodic phrase- initial position

Regarding the voiced obstruents in prosodic phrase-initial syllables, no significant difference on the phonation type of the following vowel could be observed in comparison with the ones in #initial positions as shown in Figure 4.

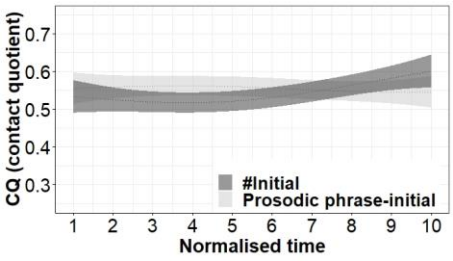


Figure 4: CQ of vowels in #initial and prosodic phrase-initial syllables.

According to the experiment results on vowel phonation in different prosodic positions presented above, it can be concluded that prosodic position does have an effect on the realisation of vowel phonation in English lettered words in Wenzhou Wu. The breathy vowel seems not to appear in word-medial and phrase-medial syllables. In other words, the breathiness feature of the vowels following phonologically voiced obstruents does not exist in medial positions of prosodic domain at the phrase level or lower in the prosodic hierarchy.

4. DISCUSSION

It can be concluded that prosodic position conditions the phonetic realisation of obstruent voicing in English lettered words in Wenzhou Wu Chinese, and both phonetic voicing and vowel phonation are involved. The phonetic adaptation reflected in the data seems strongly relevant with the common position-sensitive feature of the phonetic realisation of obstruent voicing in both Wenzhou Wu and English. The existence/absence of phonetic voicing is shown to be an effect of the #initial/intervocalic distinction only, with no interaction with the prosodic hierarchy – being voiceless in absolute initial position but otherwise phonetically voiced.

However, the phonation contrast of the vowel, which also plays an important role in voicing contrast realisation in native Wu words, depends on the prosodic structure. By virtue of the data which shows that vowels in #initial and prosodic phrase initial syllables are breathier than the ones in prosodic word-medial / phrase-medial syllables, we can infer that the vowel breathiness feature seems to only exist in prosodic phrase-initial positions (including #initial positions). Since the application of tone sandhi (at lexical and phrasal level) is regarded as a transparent prosodic marker of prosodic word and prosodic phrase in Wenzhou Wu, we can also draw the conclusion that the obstruent voicing can be realised as breathiness of the following vowel in the initial position of any sandhi tone bearing domain only. The overall interaction between prosodic position and obstruent voicing in English lettered word in Wenzhou Wu is summarised in Table 2 below. The phonologically constrained vowel phonation feature of Wu voiced obstruents has been applied to the localised pronunciations of English letters. Whether the devoicing only in absolute initial position results from articulatory factors like English [20], or adapts from the realisation of Wu voiced obstruents requires further comparison with non-loanword productions in different prosodic positions.

	#Initial	PW-medial	PP-medial	PP-inital
VOT	−	+	+	+
Phonation	+	−	−	+

Table 2: Interaction between prosodic position and voicing realisation.

It is quite special that prosodic phrasing and segmentation is transparently represented by the tone sandhi patterns in Wu Chinese. Now that the interaction between this prominent prosodic property of Wu Chinese and voicing contrast could be observed in this study, the interaction mechanism between these two segmental and suprasegmental features is also worthy of further research.

5. ACKNOWLEDGEMENTS

This study was supported by the RGC General Research Fund 2019/20 Project Reference 14607619 awarded to the second author.

6. REFERENCES

[1] Xu, M., Tian, C. 2017. Is 'NBA' Chinese or English?: The current situation of English lettered words in Chinese. *English Today* 33(4), 33–39.

[2] Institute of Linguistics of Chinese Academy of Social Sciences. 2012. *The Contemporary Chinese Dictionary*. The Commercial Press. (in Chinese)

[3] Shi, Y. 2020. *Loanwords in the Chinese Language*. Routledge.

[4] Montgomery, P. H. S. 1893. *Introduction to the Wênchow Dialect*. Kelly & Walsh Limited.

[5] Department of Chinese Language and Literature of Peking University. 2003. *A Lexicon of Chinese Dialects*. Language and Culture Press. (in Chinese)

[6] Cao, J., Maddieson, I. 1992. An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics* 20, 77–92.

[7] Song, Y. 2012. *An EGG study on the phonation of obstruents in Wu*. World Publishing Corporation. (in Chinese)

[8] Gao, J. 2015. *Interdependence between Tones, Segments and Phonation Types in Shanghai Chinese*. Ph.D. dissertation, University Sorbonne Nouvelle – Paris 3.

[9] Zhang, W., Xu, W., Mok, P. P. K. 2019. A preliminary study on the age variation of the voicing contrast in Wenzhou Wu Chinese. *Proc. 15th ICPhS* Melbourne, 2411–2415.

[10] Roach, P. 2009. *English Phonetics and Phonology: A Practical Course*. Cambridge University Press.

[11] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3), 384–422.

[12] Davidson, L. 2016. Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics* 54, 35-50.

[13] Duanmu, S. 1998. Wordhood in Chinese. In: *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*. Mouton de Gruyter, 135–196.

[14] Chen, M. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge University Press.

[15] Zheng-Zhang, S. 1964. Tone sandhi of Wenzhou dialect. *Studies of the Chinese Language* 2, 106–152. (in Chinese)

[16] Scholz, F. 1983. *Tone Sandhi, Prosodic Phrasing, and Focus Marking in Wenzhou Chinese*. Ph.D. dissertation, Leiden University.

[17] Childers, D. G., Lee, C. K. 1991. Vocal quality factors: Analysis, syntheses, and perception. *Journal of the Acoustical Society of America* 90.

[18] Kania, R. E., Hans, S., Hartl, D. M., Clement, P., Crevier-Buchman, L., Brasnu, D. F. 2004. Variability of electroglottographic glottal closed quotients: Necessity of standardization to obtain normative values. *Arch Otolaryngol Head Neck Surg* 130(3), 349–352.

[19] Herbst, C., Ternström, S. 2006. A comparison of different methods to measure the EGG contact quotient. *Logopedics Phoniatrics Vocology* 31.

[20] Solé, M. J. 2011. Articulatory adjustments in initial voiced stops in Spanish, French and English. *Proc. 17th ICPhS* Hong Kong, 1878–1881.

CREAKY VOICE IN L2 ENGLISH AND L1 DUTCH

Koen Sebregts, Hielke Vriesendorp, Hugo Quené, Yosiane White

ILS, Utrecht University
k.sebregts@uu.nl, h.a.d.vriesendorp@uu.nl, h.quene@uu.nl, y.z.white@uu.nl

ABSTRACT

While creaky voice is a well-known stylistic-prosodic feature of American English, its use in other languages is under-researched. In Dutch, it has been claimed to be rare and idiosyncratic. Meanwhile, in L2 acquisition studies, creak has been shown to exhibit L1-L2 transfer.

We investigate the prevalence of creaky voice and its development over time in young female speakers of L1 Dutch and L2 English, who had previously been shown to converge on L1/L2 segmental features longitudinally.

Automatic detection of f0 established individual distributions of creaky and modal phonation. Linear mixed-effects models of relative creak prevalence showed that effects of language (L1 vs L2) and style (read vs spontaneous speech) were limited. Rather, speakers showed similarly high levels of creak in their two languages, and stability over time. The results suggest creak is idiosyncratic, rather than gradually acquired or converged on, but also that it is far from rare in Dutch.

Keywords: creaky voice, phonation, L2 acquisition, Dutch, English

1. INTRODUCTION

1.1 Background

Creak and creaky voice are now widely assumed to be umbrella terms for a number of types of non-modal voice qualities [1-3]. Most of these share elements such as low fundamental frequency and some irregularity in glottal pulsing; despite acoustic differences, they appear to be a perceptually coherent category [4].

Cross-linguistically, creak and creaky voice can have anything from phonemic to discourse or affective functions [5]. In English, it is well-known to have prosodic and pragmatic functions, as a boundary marker [6] and turn-taking signal [7], as well as stylistic and sociolinguistic ones. Available as an identity marker of upper-middle class RP speakers to Chicana/o gang members alike [8, 9], it is popularly most associated with young female speech in American English and has received ample attention as a result, both outside of and within linguistic research [10-12].

Considerably less is known about the use of creaky voice in other languages, particularly regarding its sociolinguistic dimensions. Focusing here on languages where creak is not used for phonemic or allophonic purposes, it has been shown to serve as a prosodic marker and turn-taking signal in e.g. Mandarin Chinese, Finnish, Maori, Japanese, and Spanish [13-17]. Mentions of possible stylistic or socio-indexical uses are few and far between (e.g. in Mandarin Chinese, it may signal sarcasm, [18]; in Tzeltal Maya and Lachixío Zapotec, 'commiserating voice' [19]).

Creaky voice has also been studied in second-language contexts, and there is some evidence for L1-L2 transfer. Kim studied US heritage speakers of Mexican Spanish and L1 US English-L2 Spanish learners and found that the female speakers in both groups used creaky voice as a boundary marker in their Spanish in ways similar to those used in their English (and unlike L1 Mexican Spanish speakers) [20]. Kim ties this result explicitly to the socio-indexical meaning of creak in US English. A complicating factor is that L1 Spanish speakers from across the Spanish speaking world have been found to use creak as a boundary marker [17]. However, in that study it was male speakers who had the highest creak levels, suggesting Kim's results may indeed have a social dimension. Conversely, Gibson et al. found that creaky voice is relatively easily acquired through simple exposure to American English speech without sociolinguistic awareness, as "functionally monolingual" Spanish speakers used it more when repeating English non-words than when repeating Spanish non-words [21]. Finally, Pillot-Loiseau et al. found both transfer and acquisition effects in the interactive setting of conversations between L1 French and L1 English speakers in their respective L2s [22]. Not only did the Anglophone speakers use more creak in their French than the native French speakers did (transfer), the Francophone speakers also used more creak in their English than in their L1 (acquisition).

Creaky phonation has also been shown to have idiosyncratic aspects, to the extent that it is a useful feature for individual speaker identification [16, 23]. There appears to be both inter-speaker variation in the amount of creaky phonation (relative to modal

phonation) and, in bilingual speakers, to what extent a speaker's use of creak is similar across their two languages.

1.2 The present study: L1 Dutch – L2 English

The status of creaky voice in Dutch is largely unclear, particularly outside of speech pathology contexts. It has been described as rare and idiosyncratic [24, 25], but also as something that "often" occurs, though highly idiosyncratically [23]. There are occasional mentions in popular and social media of the term "vocal fry" alongside demonstrations of its use by young female speakers of Dutch, reminiscent of the discourse around it in the US [26].

This paper examines the prevalence of creaky voice in a population of L1 speakers of Dutch who are highly proficient L2 speakers of English. They are students at an international undergraduate college in the Netherlands, where they form a relatively tight-knit community who use English as a lingua franca. Speakers in this community have previously been shown to converge over time on particular phonetic features, both segmental (the realisation of /s/ in their L1 and L2, [27]) and prosodic (L2 speech rhythm, [28]), though not on their use of filled pauses [29, 30]. We study the use of creaky voice in both L1 Dutch and L2 English, to address 1) the dearth of studies on creak in Dutch, 2) questions regarding transfer and/or acquisition of creaky voice in second language contexts, 3) the issue of creaky voice as an idiosyncratic vs. a community feature.

1.3 Expected outcomes

Based on the results from Pillot-Loiseau et al., we expect our speakers to use more creak in English than in Dutch. The ELF setting they are in is largely composed of L2 English speakers, but also contains L1 speakers of American English. Our speakers may also have some awareness of the social function of creaky voice in English. We also expect our speakers to increase their use of creak over time, as this function is acquired. Potentially, we may see some level of convergence (speakers displaying more similar levels of creak over time). Here, the results from previous studies point in different directions: while phonetic convergence between speakers in this corpus has previously been shown, creaky voice may be too idiosyncratic a feature for this to happen. A final potential effect may be an influence from the L2 on the L1, with speakers using more creaky voice in their Dutch over time as a result of English having been their dominant language for a number of years.

2. METHODS

2.1. Corpus

The D-LUCEA corpus [31] contains recordings of 285 students (ages: 17-24) at University College Utrecht, an English-language campus college in the Netherlands. Students carried out a number of speech tasks (reading, monologue, dialogue) in L2 English as well as their L1 (Dutch for the majority of speakers, a variety of languages for the others). The corpus is longitudinal: students from three cohorts were recorded a maximum of five times over their three-year stay. The subcorpus used for the current study consists of the first (year 1, semester 1) and last (year 3, semester 2) recordings of all female speakers from the first cohort whose L1 was Dutch (no other home language growing up). The total number of speakers in this subcorpus is 29.

Each recording contains around four minutes of read speech in English, four minutes of spontaneous speech in English, two minutes of spontaneous speech in Dutch, and the final session also contained read speech in Dutch.

2.2. Automatic detection of creak

Given the difficulties associated with manual annotation of creak [32], there has been great interest in methods for the automatic detection of creak over the years [33, 34]. We use a process based on tracking f0 at 5 ms intervals across the signal with REAPER [35], shown to be the optimal creak detection method for female speakers in an evaluative comparison [36]. Previous use of this method has shown that speakers tend to have bimodal f0 distributions, rather than unimodal ones with creak at the lower tail [16, 37]. The local minimum between the two modes, one for creaky voice and one for modal voice, is the so-called antimode (AM). Using this as a cut-off point allows for each f0 measurement to be classified as either creaky (<AM) or modal voice (>AM), which in turn enables the calculation of creak prevalence per speaker (i.e., the percentage of pitch measurements classified as creaky among the total number of pitch measurements).

2.3. Statistical analysis

We analysed the data using a series of linear mixed-effects models using the *lme4* package in R [38, 39] to predict creak percentage based on language (English, Dutch), style (read, spontaneous), and recording (1 - first, 5 - final). Our data are not entirely balanced, as there is no read speech in Dutch in the first recording. This allows us to compare groups in subsets of the data along three dimensions: within the

spontaneous speech data, we examine the effects of language and recording (model 1); within the English language data, the effects of style and recording (model 2); within the data for the final recording, the effects of language and style (model 3). The final models included speaker as a random effect, and by-speaker random slopes for language (models 1 and 3), style (models 2 and 3) and recording (models 1 and 2). Models 1 and 2 did not include interactions, whereas model 3 did; model comparisons showed that these provided the best fit, respectively.

3. RESULTS

All speakers showed bimodal f0 distributions in both their English and Dutch speech, as shown in Fig. 1. Individual speakers' antimodes were highly similar across their two languages, those for Dutch ranging between 114-164 Hz (M=143, SD=15.1) and those for English between 114-162 Hz (M=143, SD=13.7).

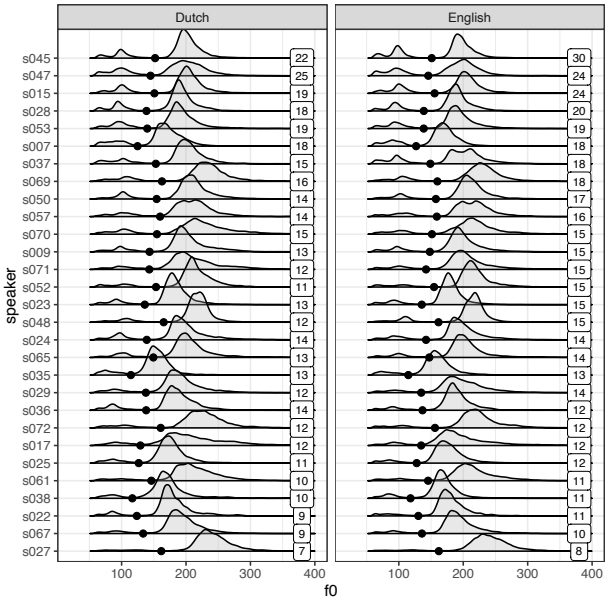


Figure 1: f0 (Hz) distributions, antimodes (●) and creak prevalence (%) by language for all speakers. Speakers sorted by overall creak prevalence.

Mean creak prevalence (over both languages, both tasks, and both recordings) showed a large amount of inter-speaker variability, ranging from 7.67 (SD=1.15) to 26.9 (SD=6.87). In contrast, intra-speaker variation was limited: almost all speakers showed very similar creak prevalence *across* their two languages, with a mean difference of 1.4 percentage points. Inter-speaker variability remained large across recordings, showing no effect of convergence (recording 1: 7-29; recording 5: 9-37).

A linear mixed-effects model predicting creak prevalence in the spontaneous speech data showed an effect of recording, with creak increasing over time (p=.034), but no effect of language. A model predicting creak prevalence in the L2 English data also showed an effect of recording (p=.033), but not of speech style. Finally, a model predicting creak prevalence in the data from the final recording showed an effect of language, with more creak in English than in Dutch (p=.0001), but no effect of style. However, there was a significant interaction between language and style (p=.017): the difference between English and Dutch is found in the read speech style only (with means of 17.5 and 14.6, respectively). The box plot in Fig.2 illustrates the effects of recording, language and style on creak prevalence in the data.

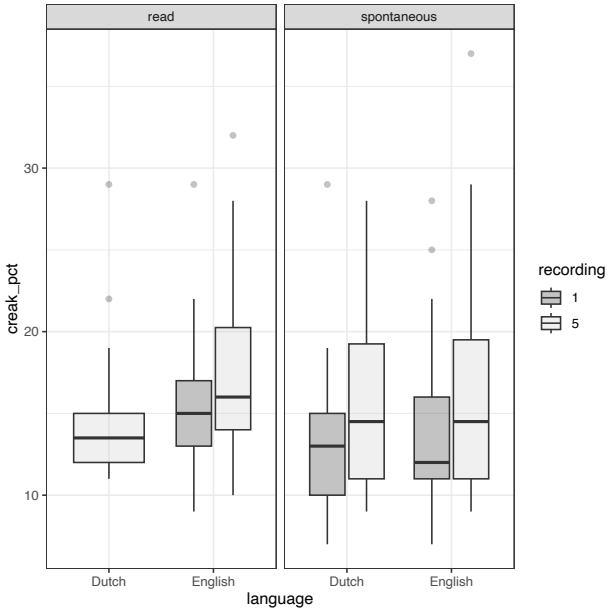


Figure 2: Box plot showing creak prevalence (creak_pct) by language (Dutch, English), speech style (read, spontaneous), and recording (first <1>, final <5>).

4. DISCUSSION

We will start by comparing these results to our expectations. We expected our speakers to use more creak in their L2 English than in their L1 Dutch, because of exposure to creak as a stylistic or socio-indexical feature in English, which – as far as we were aware – is not a function it has in Dutch. We found only very limited effects of language in our data: only in the read speech task during the final recording was there a significant difference between English and Dutch regarding creak prevalence. There are multiple elements to interpreting this result. First, it does not seem that our speakers use particularly low amounts of creak in their English. Although we did not have a clear expectation as to its prevalence, the means and variation displayed by our speakers are comparable to those reported by Dallaston and Docherty [37] for their Australian English speakers, and for those of

American English speakers in some studies [32]. We can, therefore, not conclude that our speakers have failed to acquire English-appropriate creak usage or have limited sociolinguistic awareness. Instead, the lack of a marked difference in creak prevalence between the L2 English and the L1 of our speakers is striking for the opposite reason: the amount of creak used in their Dutch. As noted, very little is known about the creak in Dutch, apart from anecdotal evidence as to its use by young female speakers and claims in older literature that it is rare. The latter notion can easily be put to rest, at least when it comes to the speakers in the D-LUCEA corpus.

The speakers in our corpus are of course not representative of the wider population; they are young, in higher education, and highly proficient in English (a prerequisite for admission to the college). What makes this particular population interesting is that they have previously been shown to display convergence on a number of accent features. The D-LUCEA corpus was collected with the idea of convergence as a "core hypothesis" [31]. This is also partially behind our second expectation, that the use of creak in students' L2 English would increase over time. This appears to be borne out by the data: creak prevalence in English increased between recordings 1 (the students' first semester) and 5 (their last). In fact, creak prevalence also increased in students' spontaneous speech overall, including their Dutch. While it is tempting to relate this to a potential effect of L2 influence on the L1, again the high levels of creak in the participants' Dutch at the outset, and the small difference in creak prevalence between their English and Dutch, casts doubt on this as an explanation. It should also be noted that the effect of recording is relatively small (for both languages, in the order of 3 percentage points).

Finally, actual convergence among speakers, our third expectation, or a potential one, was not observed. The range of creak prevalence among the speakers in the data did not decrease – if anything, it grew slightly wider. This confirms previous studies showing that creak usage is at least partially, and perhaps to a large extent, idiosyncratic, with speakers stable both across languages and over time. While this opens up avenues for forensic research, it puts creaky voice on a par with filled pauses rather than segmental and speech-rhythm features.

This brings us to the three ways this study aimed to contribute to ongoing research. The weight of the evidence on the question of whether creak is mainly an idiosyncratic or community feature seems to firmly land on the former. A second aim of the current study was to engage with the questions of whether creak is a phenomenon easily acquired by L2 learners, or one whose patterns are easily transferred from the L1 to the L2. Given our results, both in terms of the high creak prevalence in Dutch and the stability of speakers across languages and over time, we have to remain agnostic on these questions. The similar levels of creak in individual speakers' L1 and L2 may of course be due to transfer (their use of creaky voice in English simply follows their patterns in Dutch); even more speculatively, these high proficiency, young and female L2 English speakers may have even acquired their use of creak in English and transferred them to their L1. Even if that is the case, however, both its acquisition and its influencing the L1 must have occurred before the recordings in our corpus were made.

Finally, an important contribution of this study is that it is one of very few to look into creak usage in Dutch at all. Our results are to be interpreted with caution given the highly specific nature of our corpus, but they are intriguing even if not representative of wider groups in Dutch society: they suggest that creak may be idiosyncratic, but certainly not rare, among some L1 speakers of Dutch. Creak may simply have flown under the radar of observers, within and outside of linguistics, for a long time. Perhaps its use by Dutch native speakers as a stylistic and/or prosodic feature is not new and not limited to highly proficient English speakers. Going by anecdotal evidence such as mentions on social media platforms, both the term "vocal fry" and its usage (specifically by young female speakers) have increasing visibility. Whether coming from the perspective of forensic linguistics, second language acquisition or sociophonetics, there is ample scope for future study, especially given the tools now at our disposal.

5. REFERENCES

[1] Batliner, A., Burger, S., Johne, B., Kießling, A. 1993. MÜSLI: A classification scheme for laryngealizations. *Proc. ESCA Workshop on Prosody*, University of Lund, 176-179.

[2] Redi, L., Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29(4), 407-429.

[3] Keating, P. A., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th International Congress of Phonetic Sciences*, University of Glasgow, 2-7.

[4] Davidson, L. 2019. Perceptual coherence of creaky voice qualities. *Proc. 19th International Congress of Phonetic Sciences*, Australasian Speech Science and Technology Association Inc., 147-151.

[5] Davidson, L. 2021. The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *Wiley Interdisciplinary Reviews: Cognitive Science* 12(3), e1547.

[6] Kreiman, J. 1982. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics* 10, 163-175.
[7] Laver, J. 1980. *The Phonetic Description of Voice Quality.* Cambridge University Press.
[8] Henton, C., Bladon, A. 1988. Creak as a sociophonetic marker. In: Hyman, L., Li, C. (eds), *Language, speech and mind: Studies in honor of Victoria A. Fromkin*. Routledge, 3-29.
[9] Mendoza-Denton, N. 2011. The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology* 21(2), 261-280.
[10] Yuasa, I. P. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3), 315-337.
[11] Wolk, L., Abdelli-Beruh, N., Slavin, D. 2012. Habitual use of vocal fry in young adult female speakers. *Journal of Voice* 26(3), 111-116.
[12] Davidson, L. 2020. Contributions of modal and creaky voice to the perception of habitual pitch. *Language* 96(1), e22-e37.
[13] Chai, Y. 2019. The source of creak in Mandarin utterances. *Proc. 19th International Congress of Phonetic Sciences*, Australasian Speech Science and Technology Association Inc., 1858-1862.
[14] Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31, 139-152.
[15] Szakay, A., King, J. 2017. Cross-linguistic transfer effects in bilingual English-Māori voice quality and pitch. *ALS 2017: Conference of the Australian Linguistics Society*, Sydney University, 5-7 December.
[16] Dorreen, K. 2017. Fundamental frequency distributions of bilingual speakers in forensic speaker comparison. MA Thesis, The University of Canterbury.
[17] González, C., Weissglass, C., Bates, D. 2022. Creaky Voice and Prosodic Boundaries in Spanish: An Acoustic Study. *Studies in Hispanic and Lusophone Linguistics* 15(1), 33-65.
[18] Li, S., Gu, W., Liu, L., Tang, P. 2020. The role of voice quality in Mandarin sarcastic speech: An acoustic and electroglottographic study. *Journal of Speech, Language, and Hearing Research* 63, 2578-2588.
[19] Sicoli, M. 2015. Voice registers. In: Tannen, D., Hamilton, H., Schiffrin, D. (eds), *Handbook of discourse analysis*. Wiley, 105-126.
[20] Kim, J. Y. 2017. Voice quality transfer in the production of Spanish heritage speakers and English L2 learners of Spanish. In: Perpiñán, S., Heap, D., Moreno-Villamar, I., Soto-Corominas, A. (eds), *Romance languages and linguistic theory II: Selected papers of the 44th Linguistic Symposium on Romance Languages, London, Ontario*. John Benjamins, 191-207.
[21] Gibson, T. A., Summers, C., Walls, S. 2017. Vocal fry use in adult female speakers exposed to two languages. *Journal of Voice* 31(4), 510.e511-510.e515.
[22] Pillot-Loiseau, C., Horgues, C., Scheuer, S., Kamiyama, T. 2019. The evolution of creaky voice use in read speech by native-French and native-English speakers in tandem: A pilot study. *Anglophonia [online]* 27
[23] van Hugte, T., Heeren, W. 2021. An exploratory study into interspeaker variation in creaky voice in Dutch. *Dag van de Fonetiek*, Utrecht, 17 December.
[24] Collins, B., Mees, I. 2003. *The Phonetics of English and Dutch.* E.J. Brill.
[25] Jenner, B. 1987. Articulation and phonation in non-native English: The example of Dutch-English. *Journal of the International Phonetic Association* 17(2), 125-138.
[26] van Oostendorp, M. 2016. Iedereen doet vocal fry! *Neerlandistiek*. https://neerlandistiek.nl/2016/07/iedereen-doet-vocal-fry/.
[27] Quené, H., Orr, R., van Leeuwen, D. 2017. Phonetic similarity of /s/ in native and second language: Individual differences in learning curves. *JASA Express Letters* 142(6), 519-524.
[28] Quené, H., R., O. 2014. Long-term convergence of speech rhythm in L1 and L2 English. *Proc. Speech Prosody 2014*, ISCA, 342-345.
[29] de Boer, M. M., Heeren, W. F. L. 2020. Cross-linguistic filled pause realization: the acoustics of uh and um in native Dutch and non-native English. *Journal Of The Acoustical Society Of America* 148(6), 3612-3622.
[30] de Boer, M. M., Quené, H., Heeren, W. F. L. 2022. Long-term within-speaker consistency of filled pauses in native and non-native speech. *JASA Express Letters* 2(3), 035201.
[31] Orr, R., Quené, H., van Beek, R., Diefenbach, T., van Leeuwen, D. A., Huijbregts, M. 2011. An international English speech corpus for longitudinal study of accent development. *Proc. Interspeech*, ISCA, 1889–1892.
[32] Dallaston, K., Docherty, G. 2020. The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature. *PLoS ONE* 15(3), e0229960.
[33] Martin, P. 2012. Automatic detection of voice creak. *Proc. Speech Prosody 2012*, ISCA, 43-46.
[34] Kane, J., Drugman, T., Gobl, C. 2013. Improved automatic detection of creak. *Computer Speech and Language* 7, 1028-1047.
[35] Talkin, D. 2015. REAPER: Robust Epoch And Pitch EstimatoR. https://github.com/google/REAPER.
[36] White, H., Penney, J., Gibson, A., Szakay, A., Cox, F. 2022. Evaluating automatic creaky voice detection methods. *The Journal of the Acoustical Society of America* 152(3), 1476-1486.
[37] Dallaston, K., Docherty, G. 2019. Estimating the prevalence of creaky voice: A fundamental frequency-based approach. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proceedings of the 19th International Congress of Phonetic Sciences*. 532-536.
[38] R Core Team. 2022. *R: A language and environment for statistical computing.* ISBN 3-900051-07-0.
[39] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1-48.

Principles of voice quality theory: A rtMRI study on labiodental production

Daniel Denian Lee[1,3], Scott Reid Moisik[1], Vimalan Vijayaragavan[2]

[1]Linguistics and Multilingual Studies, Nanyang Technological University, Singapore
[2]Cognitive and Neuroimaging Centre, Nanyang Technological University, Singapore
[3]Phonetics Laboratory, University of Cambridge, UK
ddl26@cam.ac.uk, scott.moisik@ntu.edu.sg, vimalan.vijay@ntu.edu.sg

ABSTRACT

This real-time magnetic resonance imaging (rtMRI) study investigates the theoretic relationship between voice quality settings and phonetic segments, through the lens of two Laverian principles of susceptibility and compatibility. The setting–segment interactions were examined using whole-vocal-tract data of native Singapore English speakers. Qualitative analyses in this paper aim to offer instructive insights into the underlying nature of voice quality, and how its influence on segmental articulation is hierarchically and asymmetrically organised. Specifically, the interaction between voice quality and labiodental articulation will be analysed using the Laryngeal Articulator Model (LAM), which builds upon Laver's seminal descriptions of voice quality as a phenomenon that merits robust phonetic treatment.

Taking stock of the realities of commingling speech patterns amidst 'superdiverse' migratory complexes, voice quality theory will be elucidated through real-world examples from natural speech. Conceptual, terminological, and metatheoretical issues will be discussed, concluding with suggestions for future voice quality research.

Keywords: *voice quality theory*, *supralaryngeal settings*, *segments*, *susceptibility*, *compatibility*

1. INTRODUCTION

VOICE QUALITY is a fundamental property of an individual's speech [1], thus qualifying it as a crucial area of study in phonetic sciences research, and linguistic scholarship more broadly. Voice quality is the phenomenon—and in this paper also a theoretical framework—that consists of visuo-auditory signals present in any given person's speech production. A key concept for elucidating the notion of voice quality is that of a SETTING, which is a shorthand for 'voice quality setting', referring to a specific articulator-wise component of the overall voice quality of a given speaker, language or group. For instance, in examining the voice quality of the 45th President of the United States of America Donald J. Trump, his speech may be described as exhibiting a PROTRUDED LIPS and OPEN JAW setting, and Trump's lips and mandible are the articulatory referents when discussing PROTRUDED LIPS and OPEN JAW setting respectively. These 'sub-phenomenal' strands of Trump's voice quality are manifested, and can be observed even by phonetically untrained laypersons, in the visual (articulation) and aural (audition and acoustics) domain. Therefore, to provide an efficacious examination of voice quality, it is important to investigate both its articulatory and acoustic correlates.

2. NOMENCLATURAL CONCERNS

2.1. Defining voice quality

A well-defined technical definition of voice quality can be traced back to [1]: 'A quasi-permanent quality running through all the sound that issues from his mouth.' This definition was invoked in [2] classic descriptions of voice quality, and later reinforced in the LAM. [1] was astute in in disambiguating between the well-established phonetic sense of 'voice', which refers to the phonatory vibration of the vocal folds, and the 'voice' in 'voice quality'. The latter carries a more generalist sense of 'voice', pointing to the overall notion of a person's speech, thereby including articulations beyond the true vocal folds and other structures in the lower vocal tract. In other words, 'voice quality' should not be treated as an interchangeable synonym for 'voicing' (as in vocal-fold oscillation) or phonatory quality.

For terminological clarity, it is useful to conceive of a 'broad' and 'narrow' sense of voice quality. The narrow sense is associated with phonatory quality (e.g. LOWERED LARYNX VOICE) as expounded in the paragraph above, whereas the broad, Laverian sense of voice quality encompasses articulatory activity across the whole vocal tract, beginning inferiorly from the laryngeal articulator and ending antero-superiorly at the lips.

2.2. Voice quality typology

Voice quality manifests at different levels in naturally occurring speech contexts. At the highest level is the level of the language. That is, each language possesses its own unique aggregate of semi-

permanent gross vocal tract postures, and unsurprisingly languages belonging to the same genealogical family may appear auditorily similar to the naïve listener, such as Indonesian Malay and Tagalog (Austronesian) and Thai and Lao (Austroasiatic). French is a language that is known to feature a preponderance of nasality, which activates the velopharyngeal port (i.e. frequently lowered velic posture), and lip rounding. In LAM terms, therefore, the 'French voice quality' enacts the velopharyngeal and labial settings to give rise to an overall impression of its visuo-auditory characteristics. Another prototypical language that involves salient setting activation is Russian, known for its distinctive LOWER LARYNX configuration.

Voice quality is also analysable at the level of the individual. Speaker-specific voice qualities may be auditorily distinct to hearers owing to pathological bases (e.g. cleft lips, a short lingual frenulum) or simply stylistic factors. For instance, the American former professional boxer Mike Tyson is noted for his signature lisp, likely contributed by the gap in his central incisor space which impedes the canonical production of fricatives. For a non-pathological example of a speaker-specific voice quality, the English historian and television presenter Lucy Worsley features a distinctive CLOSE JAW VOICE that does not appear necessarily linked to pathological factors. Existing intermediately between the levels of the individual and language, is the group (or community). The group may be classified by age, sex, socioeconomic status, and even sexual orientation (e.g. see work done on the 'gay voice/persona').[1]

2.3. Articulatory setting

It is exceedingly crucial to avoid confusing 'voice quality' with another existent term in the literature: ARTICULATORY SETTING (AS). Coined by [3], AS carries a completely disparate meaning, to the extent that it may well qualify as a misnomer based on its prevalent but potentially facile usage among researchers, particularly non-linguists. Put briefly, AS is a term that solely refers to the physiological activity of speech, precluding the acoustic and perceptual correlates of voice quality. Compared against Laver's foundational framework, AS appears to be more like a synonym for the articulatory strand of voice quality (which is a tripartite phenomenon comprising articulation, acoustics and perception) than an analytic paradigm *stricto sensu*. Moreover, the AS nomenclature was founded on the basis of pronunciation pedagogy rather than analytic depth and explanatory power. Thus, the LAM is demonstrably more robust as a framework for describing the mechanisms of voice quality. Additionally, there seems to be a usual pairing of the AS nomenclature with the inter-speech posture (ISP) methodology featured in some works. The ISP technique faces a handful of theoretical and methodological concerns, with the most critical issue being that it discards speech data itself in favour of 'pause states', since only the between-utterance, pre-utterance and absolute rest postures are examined. Finally, the lexical-semantic transparency of 'voice quality' over 'articulatory setting', especially for laypersons (non-linguists), makes it difficult to recommend the continued usage of the AS nomenclature, as it may well propagate the muddying of terminological waters in the literature. With the aforementioned factors taken into consideration, there is a real impetus to promote scholarly concord in the field through the maintenance of the decidedly more appropriate term 'voice quality' in voice quality scholarship.

3. METHODOLOGY

3.1. Subjects

Subjects were recruited through word of mouth and publicity posters placed across several locations at Nanyang Technological University. A total of 53 subjects were recruited for this project, aged between 21 to 41 (inclusive) with 30 males and 23 females. Out of the total of 53, two (one male, one female) possessed atypical mandibular morphology (Class III malocclusion or negative overjet), and one (male) presented with an array of vocal tract abnormalities, including an elongated soft palate that ostensibly impeded his respiration and speech task performance. Hence, the remaining subjects (n = 50) form the usable sample for data analysis. All subjects are native or near-native Singapore English speakers. For the purpose of qualitative analysis in this paper, rtMRI data from five subjects will be presented for visual examination.

The five subjects are male, native speakers of Singapore English, ethnically Chinese (Singapore Mandarin as L2), under the age of 30, and either undergraduates or postgraduate professionals at the time of their participation. Controlling for these

[1] Refer to the author's thesis for a cited sources.

ethnolinguistic factors were aimed at mitigating sociolinguistically-grounded phonetic variation.

3.2. Data acquisition and rtMRI protocol

A 3.0 Tesla Siemens MAGNETOM Prisma scanner at the Cognitive Neuroimaging Centre (CoNiC) of the Lee Kong Chian School of Medicine was utilised for this study's experimental tasks. A real-time 2D and 3D GRE radial sequence (also known as radial FLASH)[2] was used for MR data acquisition, adapted from [4]. The steady-state free precession (SSFP) was attempted during pilot sessions, but a specific 'reconstruction' algorithm was necessary but absent in the system (V. Vimalan, personal communication, October 26, 2021) for the sequence to run successfully. Hence, [5]'s recommendation to utilise the SSFP was unfortunately not applicable for this study, despite the promise of yielding favourable rtMRI data from the 3.0 Tesla system.

Along with the MR data obtained, noise-suppressed audio was concurrently retrieved using an MR-compatible optical microphone (Dual Channel-FOMRI, Optoacoustics, Or Yehuda, Israel). The noise cancellation software of the optical microphone suppressed the operating noise from the scanner, allowing the subjects' speech audio to come through more audibly for auditory and acoustic analyses. A 64-channel head-neck radiofrequency coil was utilised to amplify the MR signal in the vocal tract region. The following acquisition parameters were applied: FOV = 198 × 198 mm; ST = 7.0 mm, with 50 slices/slab; TR/TE = 476.89/2.25 ms; flip angle = 12 °; resolution = 1.5 × 1.5 × 7.0 mm^3 ; TA = 0.43 s. The region of interest (ROI) is approximately bounded infero-superiorly by the base of the 6th cervical vertebra (C6) and nasopharynx roof, and antero-posteriorly by the nose tip and spinal cord. Occasionally, the subcutaneous fat (which displays as bright white in T1 and T2-weighted imaging) posterior to the cervical spinal column can show unwanted visual artefacts in the form of bright streaks, as was the case in pilot sessions conducted for this study. Therefore, it may be recommended to exclude the rear neck subcutaneous fat from the ROI.

3.3. Research questions

To understand the two Laverian principles of SUSCEPTIBILITY and COMPATIBILITY, the following Research Questions (RQ1 and RQ2) are posed: (RQ1) How do voice quality settings impact labiodental articulation?; (RQ2) Are voice quality settings hierarchically ordered, and how do they interact with each other?

To briefly recapitulate Laver's descriptions, susceptibility refers to the openness of segments or settings to external influence, typically a voice quality setting. Compatibility is associated with the articulatory resources necessary and available for the production of speech sounds. It is anticipated that answering RQ1 and RQ2 will illuminate the hitherto less-understood aspects of voice quality.

4. RESULTS

The midsagittal MR images of five select subjects are presented given in the Appendix, with the montage depicting the articulation of the consonantal segment [v] using three voice quality settings (i.e. NEUTRAL, PROTRUDED LIPS, and PROTRUDED JAW). The rtMRI data of five subjects are used for these montages, and to protect their anonymity they are referred to as Sub-A (Subject A, designated with subject ID #20 in the DICOM database), Sub-B (#26), Sub-C (#32), Sub-D (#43), and Sub-E (#51) in this paper. Qualitative analyses are provided based on visual examination of the MRI montage, and categorised by articulatory markers of analysis (AMAs)[3] per subsection. Comparing the inline MR images vertically reveals the cross-setting differences in how voice quality influences segmental articulation, allowing RQ1 and RQ2 to be addressed from an articulatory perspective.

4.1. Labial

The articulation of the voiced labiodental fricative [v] appears to be canonical across all five subjects in the NEUTRAL row of the MRI montage; the upper incisors come into contact with the lower lip. When considering the PROTRUDED LIPS voice quality and its effect on [v] articulation, an interesting pattern emerges unanimously: All subjects realised the labiodental using the inner surface of where the mentalis muscle is located in place of the lower lip. Both upper and lower lips are advanced anteriorly, which prohibits canonical productions of [v] because

[2] Fast low angle shot (FLASH) magnetic resonance imaging.

[3] The four AMAs include LABIAL, LINGUAL, VELOPHARYNGEAL, and LARYNX HEIGHT parameters. AMAs are distinct from settings to distinguish between independent (settings) and dependent (AMAs) variables.

the lower lip is now displaced and unable to meet the upper incisors. Hence, as a natural compensatory action, all subjects made contact between the upper incisors and the inner mentalis muscle wall to achieve an auditorily similar [v] realisation.

Continuing with the analysis of [v] realisation, in the bottom row of the montage, it appears that Sub-A and Sub-E (first and last subjects) made the least use of the dental articulators under the PROTRUDED JAW condition. Corroborating this observation with their speech audio, the non-naturalness of their productions can be detected auditorily. With the physiologic constraint imposed by the PROTRUDED JAW configuration, it appears as though both subjects utilised their upper and lower lips to create a [v]-like sound, in place of the canonical labiodental configuration. Interestingly, the audio data for Sub-A reveal that he managed to achieve a continuous fricative-like sound, but Sub-E clearly struggled to sustain the sound with his lips and only made instantaneous realisations of [v]-like sounds in his utterance. In contrast, the intermediate Subjects B, C, and D appear to adopt a dentolabial strategy to achieve a fricative sound auditorily similar to [v]. This articulation is in theory the more natural outcome as a result of the lower muscular effort required to produce a dentolabial with a negative overjet configuration [6]. This difference in strategy between Subjects A and E on the one hand, and B, C and D on the other is a worthwhile area of further inquiry, using quantitative methods.

4.2. Lingual

All subjects show radically different overall tongue shapes in their production of [v] (top row, NEUTRAL setting), but there is a slight concavity to all subjects' tongues (albeit to varying degrees). This lingual concavity is interesting, because the lingual profile should not act as a significant contributor to shaping the auditory and acoustic properties of the fricative [v]; the most salient articulatory manipulation is to bring together the upper central incisors and the lower lip to create turbulent egressive airflow. It should be reiterated that in the experiment, subjects were tasked to articulate the nonce word <AVA>, and the inter-vocalic [a] context may have a role in biasing this tongue shape. It is interesting to note that Sub-C and Sub-D's NEUTRAL production of [v] appear to involve relatively less tongue retraction (ascertained by the larger linguo-PPW[4] distance) compared to the other three subjects.

4.3. Velopharyngeal

In the top row of the montage, the velopharyngeal port is closed in NEUTRAL productions of [v]. Since the outermost part of the vocal tract (lips, incisors) are the most salient structures for canonical [v] production, it might be natural to conclude that the velum has no relation to the production of labiodentals/dentolabials. This seems to be the case for the PROTRUDED LIPS series data, where the velum remains raised for all subjects (Sub-B may in fact even be showing a tighter velo-PPW constriction, as the appears to be less of a gap compared to its NEUTRAL counterpart). However, in the bottom row of the montage, the PROTRUDED JAW series clearly shows obvious lowering of the velum in Sub-B and Sub-C, compared to the sealed velopharyngeal port seen in Sub-B and Sub-C's NEUTRAL correlates for [v] production.

4.4. Larynx height

A horizontal white line is overlaid across each setting row in the montage, positioned at the base of the 6th cervical spine (C6) belonging to Sub-A, matched against the cricoid lamina. On the whole, it does not seem that the settings analysed in this study have an apparent effect on larynx height in [v] production. There is a possible exception in the case of Sub-A, where his [v] articulation appears to utilise a slight RAISED LARYNX posture in the non-neutral settings. The connection between larynx height and labial/mandibular settings is not explicit, and the current results do not provide enough information to reveal any interesting articulatory patterning in terms of voice quality.

5. CONCLUSION

In answering the research questions: (RQ1) Voice quality settings impact labiodental in asymmetrical and non-linear ways not easily categorisable through preliminary analysis; (RQ2) voice quality settings are hierarchically ordered and take precedence over one another depending on the phone in question. To improve the generalisability of the findings presented, other segmental articulations would need to be examined, and a quantitative analysis using techniques from geometric morphometrics may help yield deeper insights into the hidden structures of voice quality theory.

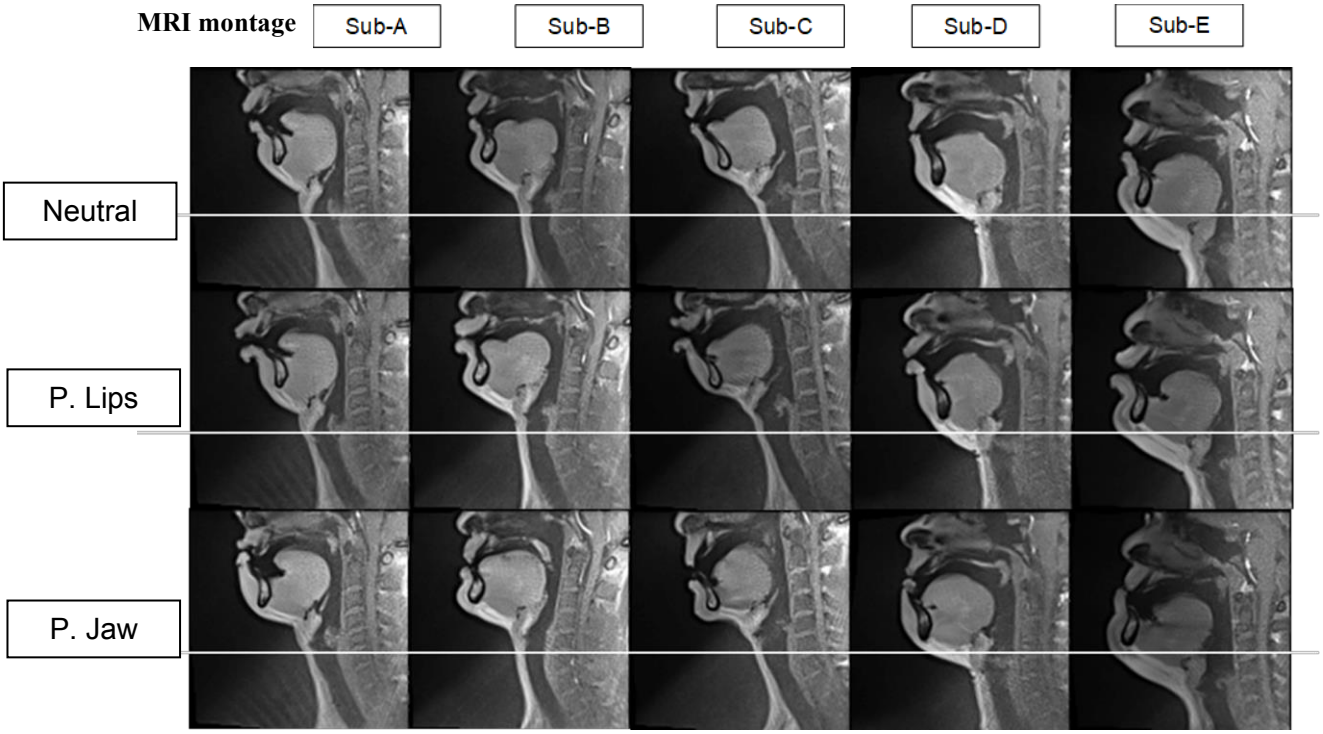
[4] Posterior pharyngeal wall.

6. REFERENCES

[1] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.
[2] J. Laver, *The phonetic description of voice quality*, vol. 31. Cambridge: Cambridge University Press, 1980.
[3] B. Honikman, 'Articulatory settings', in *In honour of Daniel Jones*, D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, and J. L. M. Trim, Eds. London: Longman, 1964, pp. 73–84.
[4] B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, and D. P. Kuehn, 'Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T', *J. Magn. Reson. Imaging*, vol. 32, no. 5, pp. 1228–1237, 2010, doi: 10.1002/jmri.22369.
[5] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, 'Recommendations for real-time speech MRI', *J. Magn. Reson. Imaging*, vol. 43, no. 1, pp. 28–44, 2016.
[6] D. E. Blasi, S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel, 'Human sound systems are shaped by post-Neolithic changes in bite configuration', *Science*, vol. 363, no. 6432, p. eaav3218, 2019.

APPENDIX

MRI montage



I CAN BE BOTH? (PRE-)ASPIRATION AND (PRE-)GLOTTALISATION DO NOT HAVE TO BE MUTUALLY EXCLUSIVE

Míša (Michaela) Hejná

Aarhus University
misa.hejna@cc.au.dk

ABSTRACT

This paper explores the relationship between pre-aspiration and pre-glottalisation, which have traditionally been seen as mutually exclusive when occurring in the same language and/or language family. Acoustic analysis of Welsh English fortis plosives and fricatives produced by 45 speakers reveals that glottalisation is less frequent (17%; n = 545) than pre-aspiration (56%; n =1846). 181 of 3306 tokens show both pre-aspiration and pre-glottalisation (5%). This co-occurrence takes the following forms: 1. the two happen simultaneously, in various forms of whispery/breathy/lax creak; 2. the two happen successively; 3. a combination of 1. and 2. is found. Where the two happen successively, (whispery/breathy/lax) creak typically precedes local breathiness and/or voiceless pre-aspiration. Option 1 is the most frequent in the data, with variable degrees of aspiration and glottalisation present. Whispery/breathy/lax creak emerge(s) as (a) phonatory setting(s) which can be meaningful for our understanding of subsegmental phonatory phenomena, and not only paralinguistic phenomena.

Keywords: pre-aspiration, pre-glottalisation, whispery creak, breathy creak, lax creak, breathiness.

1. INTRODUCTION

This paper focuses on the relationship between pre-aspiration and pre-glottalisation. The former is defined here as a period of primarily glottal friction which occurs in the sequences of sonorants and phonetically voiceless obstruents, prior to the release of the obstruent, as in Welsh English *mat* [ma^{h}t^{s}] and *mass* [ma^{h}s]. The latter is defined here as aperiodic glottal pulsing and/or as a sudden drop in f_0 in the contexts of sonorants and phonetically voiceless obstruents, as in Welsh English *mat* [ma^{ʔ}t^{s}]. The two phenomena are typically investigated separately (e.g., [4, 15, 17-18, 24], but they can enter into a range of relationships.

The research available on the relationship between the two, where these are found in the same language and/or language family, points to a contrastive or an allophonic relationship. In the former, pre-aspirated consonants are found to contrast with pre-glottalised consonants. This has been attested for sonorants, as in Shehret [29] and Huatla Mazatec [26, p. 220), but has not been reported for plosives to the best of my knowledge. [19, p. 340] discuss examples of aspiration and glottalisation being contrastive.

Where pre-aspiration and pre-glottalisation are allophonic, this relationship can be realised in several ways. Firstly, categorical allophony between the two has been found [7; 10: ch. 5; 13]. In their Manchester English data, [13] report (near-)categorical pre-glottalisation in word- and foot-final fortis plosives (*mat* [ma^{ʔ}t^{s}]), but pre-aspiration in word- and foot-medial fortis plosives (*matter* [ma^{h}t^{s}ə]) and in word- and foot-final fortis fricatives (*mass* [ma^{h}s]). This is also reported by [10: ch. 5] for Aberystwyth English and by [11] for various Welsh English accents. Furthermore, [7] report glottalisation in word- and foot-final fortis plosive environment but pre-aspiration in word- and foot-final fortis fricative environment in their Scottish English data. The two can therefore participate in categorically allophonic relationships conditioned by the prosodic environment and/or by the manner of articulation of the consonant involved.

These allophonic relationships, however, do not have to be categorical. [11] reports cases which present us with gradient allophonic constellations. For instance, a word- and foot-final fortis plosive can be associated primarily with pre-glottalisation, but not categorically so, reaching different rates of application with different speakers. Thus, both pre-aspiration and pre-glottalisation can occur in the same prosodic and segmental environment, with one being more preferred than the other.

Finally, historical discussions of pre-aspiration in some languages present us with a situation in which daughter language A shows pre-aspiration where daughter language B shows pre-glottalisation [9, p. 24, 29; 23, p. 154–155]. This diachronic perspective is particularly intriguing as it is not immediately obvious what a precursor to the two outcomes, pre-aspiration and pre-glottalisation, should be. Indeed, aspiration and glottalisation are generally seen as quite different in general discussions of laryngeal contrasts and phonatory settings ([1, p. 102; 8, p. 4, 63; 16, p. 325], but see [14, p. 205; 27, p. 431]).

Although (pre-)aspiration and pre-glottalisation have been seen as mutually exclusive on the whole,

implicitly or explicitly (e.g. [19]), [10, ch. 5; 11] nevertheless makes a cursory comment on cases in which pre-aspiration and pre-glottalisation are not only found in the same prosodic and segmental environments, but also within the same *token*. Similar findings have also been reported by [29] for one speaker of Mehri in the context of consonantal sonorants. Following up on [11], this paper addresses the following questions, using acoustic analyses of a range of Welsh English accents:

- How often do pre-aspiration and pre-glottalisation co-occur in the same tokens?
- What is the phonetic realisation of such cases?

2. METHODOLOGY

2.1. Tokens

The tokens analysed here are taken from the dataset presented in [11]. As Table 1 shows, each participant produced a range of foot- and word-final fortis plosives (/p, t, k/) and one fortis fricative (/s/) in words of a CVC structure. These words were uttered once in isolation, leading to 71-74 words per speaker (due to the occasional presence of a postvocalic /r/, in which case the token was excluded). The dataset comprises of 3306 tokens in total.

Vowel	/p/	/t/	/k/	/s/
/a/	*cap*	*mat*	*back*	*lass*
	map	*cat*	*lack*	*mass*
	sap		*mac*	
/ɛ/	*rep*	*let*	*neck*	*guess*
		pet		*less*
/ɪ/	*lip*	*lit*	*Nick*	*miss*
/ɒ/	*lop*	*lot*	*lock*	*loss*
/ʌ/	*sup*	*but*	*luck*	*bus*
		hut		*fuss*
/ʊ/		*foot*	*cook*	
		put	*look*	
		soot	*took*	
/aː/	*carp*	*art*	*lark*	
	harp	*heart*		
/iː/	*leap*	*beat*	*leak*	*lease*
		meet	*leek*	
/ɜː/	*burp*		*lurk*	*verse*
/oː/	*dorp*	*taught*	*auk*	
	thorp	*wrought*		
/uː/	*loop*	*lute*	*Luke*	*loose*
/aɪ/	*ripe*	*light*	*like*	*lice*
/aʊ/		*lout*		*louse*
/əʊ/	*lope*	*mote*	*oak*	
/eɪ/	*nape*	*late*	*fake*	*face*

Table 1: Words by segmental properties; from [11].

The participants comprise of 45 L1 and L2 speakers of Welsh English. They were born in a range of towns and regions in Wales, UK. The dataset is balanced for gender, but the vast majority of the speakers are L1 (rather than L2) speakers of Welsh. See [11] for more details.

2.2. Data processing

All speakers were recorded reading the same list of words, only part of which is presented and analysed here (as well as in [11]). H4 Zoom Handy recorder was used with a head-mounted AKG C520 microphone. The data was manually annotated and inspected in Praat [2]. Considering the nature of the phenomena of interest (see Section 3), no inferential statistical analyses were conducted. However, see [11] for statistical analyses of pre-aspiration application with respect to the participants' demographics.

2.3. Pre-aspiration

Pre-aspiration was defined as a period of (mostly) glottal friction found in sequences of vowels and phonetically voiceless obstruents. More specifically, voiced and voiceless aspects of this friction were distinguished, following e.g. [10]. The voiceless component is labelled as 'pre' in Fig. 1, while the voiced component, realised as local breathiness, is labelled 'br'. For more details on the relationship between the two, see [10, ch. 3].

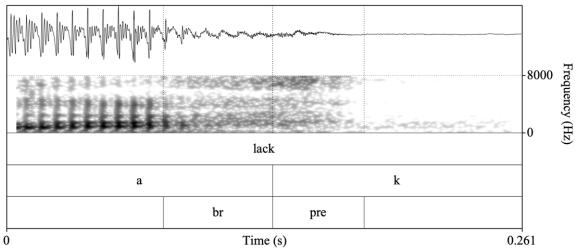


Figure 1: Local breathiness and pre-aspiration. Aberystwyth English speaker.

2.4. Pre-glottalisation

Pre-glottalisation was defined in line with [11, 13] as an aperiodic vocal fold vibration and/or as a sudden drop in f_0. The former is illustrated in Fig. 2. Crucially, glottalisation was only considered to be (sub)segmentally relevant if it occurred towards the end of the vowel, in line with [13]. 545 cases (17%) of the tokens showed vowel-final glottalisation, 514 found in plosives (19% of the plosives) and 31 in fricatives (5% of the fricatives). The 545 cases of vowel-final glottalisation present a conservative

number, as many of the whole-vowel cases affect words with vowels of very short durations, such as *put*.

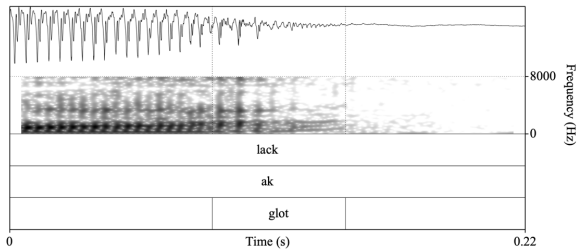


Figure 2: Pre-glottalisation. Dolgellau English speaker.

2.5. Whispery/breathy/lax creak

Whispery creak is a compound phonation [21, p. 161], which combines whisper and creak and manifests itself through noisiness more intense than that of breathiness [3, p. 53, 56, 124; 21, p. 45, 112, 121]. Importantly, "[t]he transition from breathing to breath to whisper can be thought of as an incremental narrowing of the airway from most open to moderately open to constricted" [3, p. 53], which is also reflected in gradient transitions from breathiness to whisperiness in the acoustic signal. The acoustic data analysed here reveals a considerable variation in what could be seen as compound types involving creakiness and breathiness and/or whisperiness [22, 28]. A rich palette of what could be described as whispery, breathy, and lax creak [6] is therefore observed.

3. RESULTS

Pre-aspiration is found in 1846 tokens, i.e. in 56% of the data (62% of the plosives, n = 1643; 32% of the fricatives, n = 203). Every speaker pre-aspirates. In contrast, 10 of the speakers never pre-glottalise, and 15 speakers produce only 1-3 tokens with pre-glottalisation. Pre-glottalisation therefore occurs less frequently than pre-aspiration, with 545 tokens being pre-glottalised in total (17%; 19% of the plosives, n = 514; 5% of the fricatives, n = 31).

181 tokens are both pre-aspirated and pre-glottalised, amounting to 5% of the overall data (Fig. 3). Of the 542 fricatives, 10 show both pre-aspiration and pre-glottalisation (2% of the fricatives). Of the 2655 plosives, 171 display the co-occurrence of the two (6% of the plosives). It needs to be borne in mind that "pre-aspiration" refers to voiceless abduction, i.e. not instances of pre-aspiration realised as voiced, i.e. as local breathiness. This means the numbers of co-occurrence reported here are conservative.

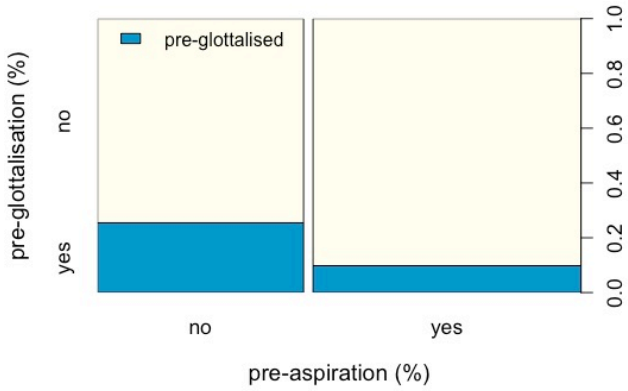


Figure 3: Co-occurrence of (voiceless) pre-aspiration and pre-glottalisation within a token.

Where pre-aspiration and pre-glottalisation occur in the same token, this can take on the following forms. Limiting ourselves to narrowly defined pre-aspiration as voiceless, the vast majority of cases present us with sequences in which we find glottalisation overlaid with aspiration, which is frequently followed by local breathiness and subsequently by voiceless pre-aspiration, or more rarely directly by voiceless pre-aspiration. The number of tokens in which pre-aspiration and pre-glottalisation co-occur increases if instances with local breathiness (voiced pre-aspiration) are considered. Of the 544 cases with pre-glottalisation, the sequencing of the laryngeal phenomena described above applies to 306 cases with local breathiness. Some examples of variably simultaneous glottalisation and aspiration are shown in Fig. 4-7.

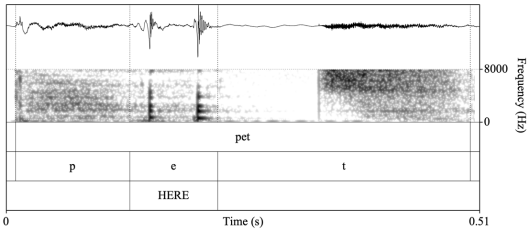


Figure 4: Glottal friction overlapping with glottalisation. A speaker from Yr Wyddgrug / Mold. Token excluded from quantitative analyses (whole vowel affected).

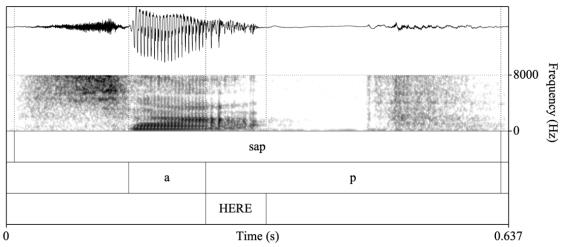


Figure 5: Glottal friction overlapping with glottalisation. A speaker from Dolgellau.

In contrast, only 19 tokens show a much more neatly separated sequence of glottalisation, local breathiness, and pre-aspiration. Few tokens digress from the pattern showing a progression from glottalisation to aspiration: A) 4 tokens show breathiness (some of which is followed by pre-aspiration), which is then followed by glottalisation; B) 23 tokens show breathiness followed by glottalisation overlaid with aspiration (4 of which include pre-aspiration following as the last laryngeal event); C) 3 cases present us with glottalisation overlaid with breathiness which is followed by pre-aspiration with glottal flatulence [25, p. 60]. Finally, 121 tokens display glottalisation with no visible aspiration, and 64 tokens present us with slack glottalisation, or lax creak, mirroring the slack profile of the vowel.

Within themselves, all of these profiles further show a range of gradient possibilities as to the amount and intensity of glottalisation and aspiration present, and these also vary durationally. The data further suggests that the onset of local breathiness, and that of pre-aspiration broadly defined, may be associated with barely noticeable irregularity in the periodicity of the signal. These minor irregularities can become more noticeable and be recognised as more canonical glottalisation (Fig. 6) by the analyst. As Fig. 7 shows, this glottalisation can manifest itself as durationally more prominent. The very short glottalisation durations of the type shown in Fig. 6 could be brought about by relatively sudden abduction, i.e. by a relatively abrupt change of the articulatory settings.

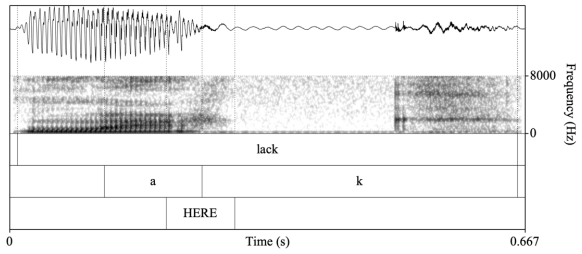


Figure 6: Glottal friction overlapping with glottalisation. A speaker from Clynderwen.

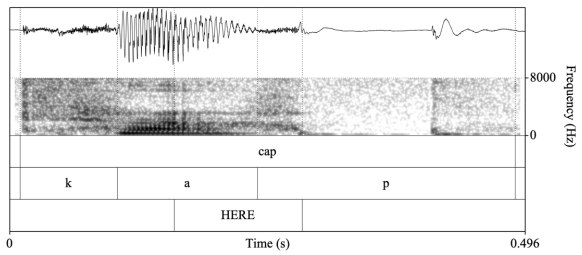


Figure 7: Glottal friction overlapping with glottalisation. A speaker from Caerphilly.

4. DISCUSSION

In this study, whispery creak, breathy creak, and lax creak emerge as phonatory settings which can be meaningful for our understanding of subsegmental phonatory phenomena, and not only paralinguistic phenomena [3, p. 127; 21, p. 122, 140]. This being the case, the co-occurrence of pre-aspiration and pre-glottalisation reported here should be considered in accounts of diachronic patterns whereby a reconstructed pre-aspirated/(pre-)glottalised obstruent gives rise to a pre-aspirated obstruent in daughter language A but a (pre-)glottalised obstruent in daughter language B (e.g. Western Yugur and Tuva; [10, ch. 5]). The two phenomena are necessarily not contrastive in such cases, but rather seem to serve the same articulatory and phonological goal, at least in the obstruent environment. This is also in line with [19, p. 327–350], according to whom [hʔ] and [ʔh] never contrast.

The fact that pre-aspiration and pre-glottalisation can co-occur simultaneously in the same token points out that the glottal continuum model of voice quality [20] is not the most optimal model regarding *phonetic* (as opposed to phonological) typology, as acknowledged in previous work ([3, 5, 12]).

Finally, the results reported here raise the question of what whispery creak and breathy creak are. Although clear definitions of whispery creak are readily available [3, 21], these refer to canonical whispery creak. Canonical whispery creak involves epilaryngeal constriction and should be accompanied by a more turbulent airflow and higher-intensity friction in the acoustic signal. However, research has repeatedly shown that canonical realisations of phenomena are not necessarily what the analyst deals with. A case in point is the glottal stop [5, p. 11]. In the same vein, breathiness can encompass realisations with higher-intensity friction as well as those where friction is practically absent [28]. Pre-aspiration has similarly been debated regarding the intensity and the source of its friction (e.g. [12]). With the acoustic data at hand, the realisations of pre-glottalisation and pre-aspiration often provide a gradient of slack, or lax, creak, and heavily fricated, or whispery, creak. Although [21, p. 113, 133] lists breathy creak as impossible, the data analysed here suggest that this setting is quite possible.

To conclude, the work presented here is used to explicitly state that, just like the glottal stop, creak, and breathiness can take on various shapes, so can whispery creak, breathy creak, and lax creak. The data further suggests that the compound phonatory settings of whispery/breathy/lax creak can be employed for non-paralinguistic purposes, albeit to a rather limited extent.

7. REFERENCES

[1] Bladon, J. 2004. *Evolutionary Phonology*. Cambridge University Press.

[2] Boersma, P., Weenink, D. 2022. "Praat: doing phonetics by computer." Version 6.1.26, 2022. http://www.praat.org/.

[3] Esling, J. H., Moisik, S. R., Benner, A., Crevier-Buchman, L. 2019. *Voice Quality. The Laryngeal Articulator Model*. Cambridge University Press.

[4] Garellek, M. 2013. *Production and Perception of Glottal Stops*. PhD Thesis, University of California.

[5] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: phonetics of voice quality. *Journal of Phonetics* 94, 101155.

[6] Gobl, C., Ní Chasaide, A. 2000. Testing affective correlates of voice quality through analysis and resynthesis. *Proc. ITRW on Speech and Emotion* Newcastle, Northern Ireland, 178–183.

[7] Gordeeva, O., Scobbie, J. 2013. A phonetically versatile contrast: pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality. *Journal of the International Phonetic Association* 43, 249–271.

[8] Gussenhoven, C., Jacobs, H. 2005. *Understanding Phonology*. 2nd edn. Hodder Education.

[9] Hansson, G.Ó. 2001. Remains of a submerged continent. Preaspiration in the languages of Northwest Europe. In: Brinton, L. J. (ed.). *Historical Linguistics 1999: Selected Papers from the 14th International Conference on Historical Linguistics*. John Benjamins, 157–173.

[10] Hejná, M. 2015. *Pre-aspiration in Welsh English: A Case Study of Aberystwyth*, PhD thesis, University of Manchester.

[11] Hejná, M. 2021. Exploration of Welsh English pre-aspiration: how wide-spread is it? *Proc. Interspeech* Brno, 3974–3978.

[12] Hejná, M. Forthcoming. On the rarity of pre-aspirated consonants. In: Anderson, C., Kuznetsova, N., Easterday, S. (eds). *Rarities in Phonetics and Phonology: Evolutionary, Structural, Typological and Social Dimensions.* Language Science Press.

[13] Hejná, M., Scanlon, J. 2015. New laryngeal allophony in Manchester English. *Proc. 18th ICPhS* Glasgow.

[14] Hejná, M., Kimper, W. 2019. Pre-closure laryngeal properties as cues to the fortis-lenis plosive contrast in English. *Yearbook of the Poznań Linguistic Meeting* 4 (1), Mouton de Gruyter, 179–211.

[15] Hejná, M., Kaźmierski, K., Guo, W. 2021. Even Americans pre-aspirate. *English World-Wide. A Journal of Varieties of English* 42(2), 200–226.

[16] Honeybone, P. 2005. Diachronic evidence in segmental phonology: the case of obstruent laryngeal specifications. In: van Oostendorp, M. and van de Weijer, J. (eds). *The Internal Organisation of Phonological Segments*. Mouton de Gruyter, 319–354.

[17] Jatteau, A., Hejná, M. 2016. Dissimilation can be gradient: evidence from Aberystwyth English. *Papers in Historical Phonology* 1, 359–386.

[18] Jones, M. J., Llamas, C. 2003. Fricated pre-aspirated /t/ in Middlesbrough English: an acoustic analysis. *Proc. 15th ICPhS* Barcelona, Spain, 655–658.

[19] Kehrein, W., Golston, C. 2004. A prosodic theory of laryngeal contrasts. *Phonology* 21(3), 325–357.

[20] Ladefoged, P. 1971. *Preliminaries to Linguistic Phonetics*. University of Chicago Press.

[21] Laver, J. 1980[2009]. *The Phonetic Description of Voice Quality*. Cambridge University Press.

[22] Moisik, S., Hejná, M. and Esling, J.H. 2019. Abducted vocal fold states and the epilarynx: new taxonomy for distinguishing breathiness and whisperiness. *Proc. 19th ICPhS* Melbourne, 220–224.

[23] Pentland, D. H. 1977. Proto-Algonquian stop clusters in Cree-Montagnais. *International Journal of American Linguistics* 43(2), 154–156.

[24] Redi L., Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalisation in normal speakers. *Journal of Phonetics* 29, 407–429.

[25] Skarnitzl, R. 2004. Acoustic categories of nonmodal phonation in the context of the Czech conjunction "a". *Phonetica Pragensia* 10, 57–68.

[26] Steriade, D. 1994. Complex onsets as single segments: the Mazateco pattern. In: Kisseberth, C.W. and Cole, J.S. (eds). *Perspectives in Phonology*. CSLI, 203–291.

[27] Stevens, M. and Hajek, J. 2007. Towards a phonetic conspectus of preaspiration: acoustic evidence from Sienese Italian. *Proc. 16th ICPhS* Saarbrücken, 429–432.

[28] Tian, J. and Kuang, J. 2021. The phonetic properties of the non-modal phonation in Shanghainese. *JIPA* 51, 2, 202–228.

[29] Watson, J., Heselwood, B. Forthcoming. Silent sonorant articulations in Mehri and Shehret. In: Anderson, C., Kuznetsova, N., Easterday, S. (eds). *Rarities in Phonetics and Phonology: Evolutionary, Structural, Typological and Social Dimensions.* Language Science Press.

THE DEVIL IS IN THE DETAIL: AN INTERACTIONAL-PHONETIC STUDY OF G-WORD INTERJECTIONS AND SOME METHODOLOGICAL IMPLICATIONS

Marina Noelia Cantarutti

University of York
marina.cantarutti@york.ac.uk

ABSTRACT

This paper illustrates the methodological contributions of Conversation Analysis-Interactional Linguistics to the study of the phonetic and prosodic design of interactional phenomena using conversational corpora. It presents as a case study the analysis of a collection of 74 "(Oh) (my) God" interjections from 3 hours of the CallFriend corpus. We argue that to fully describe conversational practices, an initial exhaustive qualitative approach is required that jointly incorporates interactional and parametric phonetic analyses, so as to better inform how data is to be grouped according to interactionally-relevant criteria (i.e. position, composition, and action) and what phonetic-prosodic features can get eventually measured and compared. This study demonstrates how phonetic-prosodic phenomena that quantitative results may treat as outliers or which get hidden in the sea of de-contextualised aggregate data are in fact speakers' orientations to particular moment-by-moment interactional demands of the local context and thus have an import and organisation of their own.

Keywords: phonetics of talk-in-interaction, conversation analysis, qualitative approaches

1. AN INTERACTIONAL VIEW OF PHONETICS

Phonetics and phonology studies frequently quantify phenomena with (quasi-) experimental corpora mostly collected at the lab. While this is a necessary approach to isolate, control and study particular features, when it comes to determining the role of phonetic and prosodic phenomena in interactional practices -which are context-free and context-sensitive [1]-, the use of conversational data and an association with qualitative approaches becomes paramount.

The phonetics of talk-in-interaction (PoTI; see [2], [3]) is a growing field of phonetic research on naturally-occurring audiovisual data informed by the empirical and inductive methods of Conversation Analysis (CA) and Interactional Linguistics (IL) [1], [4]. Phonetics is considered one of the many resources that co-participants in interaction deploy in real time to carry out social actions (e.g. requesting, repairing, agreeing) and make them recognisable to each other. CA-IL sees the general smoothness of interaction as the result of the successful resolution in real time of "problems" [5] such as securing intersubjective understanding and affiliation needs, the management of turn-taking and of sequences of initiating and responding action, the design of turns, repair, and the calibration of affiliation. By using participants' displayed behaviours ("participant orientation") in next turns as endogenous evidence as well as cumulative evidence from prior research, CA-IL seeks to explain both the single case as well as the aggregate of collections of cases [6].

This paper argues for the importance of a careful qualitative inspection and analysis of conversational data in order to determine interactionally-relevant data groupings and phonetic-prosodic features before doing a quantitative analysis of the aggregate -which may hide some of the local forms of organisation. To demonstrate this, we present an exploratory interactional-phonetic study on the interjectional phrase (Oh) (My) G-word (OMG, henceforth) from a corpus of telephone conversation data. As interjections are context-dependent for their meaning and valence ascription, we show how engaging the methods of CA-IL can shed light on how phonetic-prosodic phenomena are *orderly* deployed to meet particular interactional demands of ongoing courses of action in conversation in what would otherwise be seen as "messy" data.

2. BACKGROUND

Interjections are often taken to be peripheral to language as they are semantically empty, partly phonologically "anomalous" [7] and syntactically unbounded. OMG interjectional phrases [8] in particular are called "expletive" as the G-words (e.g. God, Jesus) have undergone semantic bleaching away from prayer and swearing [9, 10] to be used to manage aspects of subjectivity, textuality, and interactivity [11], [12] in a multifunctional way. Interjections are not necessarily structurally unbounded: they are deployed in specific slots in the interactional structure [13] and belong in ordered sequences of initiating-responding actions to which they may be prosodically integrated [14]. It is in interaction that interjections (as often ambiguous "liminal signs" [15]) acquire social meaning.

Prior CA/IL research has found expletive interjections to happen in response to "good/bad news" and troubles-talk (i.e. complaint and misfortune stories) sequences. OMG in particular has been related to the systematic public display of "surprise" (a newsworthy event that interactants treat as counter to expectations [13]) as well as to the expression of disgust, sympathetic dismay or an engaged empathetic response [16] to someone else's experience. In [17], "Oh God" is treated as a variant of "oh"-prefaced news receipts, where surprise is indexed through the initial glottalization of the vowel in "oh" and wide rising-falling pitch contours. In a study of affect in complaint stories ([18]), OMG and similar alternatives are produced with participants upgrading or matching each other's pitch, loudness or voice quality to manage the fittedness of their mutual (non-)affiliative responses. Because OMG orients to informings and tellings, it may prosodically tie to the design of those initiating actions as well as responses as above. [19] describes how good news tend to be produced with high pitch level and range, abrupt step-ups, louder volume and faster speech rate. Bad news, on the other hand, tend to feature lower pitch level and narrower range, lengthened vowels on which the falling pitch is extended, a slower speech rate and breathy or creaky voice. Our study shows how co-participants orient to both the phonetic-prosodic design of initiating actions as well as to that of each other's responses.

3. DATA AND METHODS

This exploratory analysis of the interactional and phonetic-prosodic properties of OMG interjections is based on 10 telephone conversations (3 hours) from the US English CallFriend corpus [20] rendering 126 G-word tokens, of which 74 featured the words "God/Gosh". The interactional analysis involved the transcription of the extracts using GAT-2 transcription conventions [21] and an inductive description of each, focusing on social action and sequential organisation [22] to register the position and composition [23] of OMG along these lines (adapted from [24]): *position in the sequence (initiating or responsive) and turn (beginning, middle and end); position in the turn relative to other material (pre-positioned, post-positioned, standalone; in overlap or in the clear); social actions OMG implements (stance or affect displays, affiliation, organisation of interaction, other); social actions OMG responds to and their timing (adjacent or delayed); responses to OMG; composition (lexical, syntactic, phonetic-prosodic, other).*

The interactional overview resulted in detailed individual analyses of single cases which were then grouped into collections around patterns of similarity. Of the five subcollections, two feature OMG in initiating slots and three in responsive position. Here we report the analysis of two collections of responsive OMG as displays or surprise or sympathy.

The phonetic analysis was initially done auditorily, incorporating impressionistic observations into the GAT-2 and IPA transcriptions, and parametrically, that is, registering different levels of phonetic detail without ruling out, a priori, the relevance of any feature in OMG tokens and adjacent turns [25], including: *prosodic boundaries, accentuation and f0 contours and shifts; voice quality shifts; duration (at/within the segment and syllable level); tempo and rhythm (noting if any, presence of perceptual isochrony); perceptual levels of loudness; phonetic features surrounding the production of individual forms (e.g. pre-glottalisation, hold and release of plosives, monophthongisation)*. Given dialectal differences across the corpus, vowel qualities for "God" were not measured. Acoustic validation of the registered information was done via manual Praat [26] analyses for those OMG tokens where the audio and voice quality allowed for it, acknowledging the potential effects of telephonic transmission damping of frequencies below 300 Hz. Textgrids at syllable level were created and pitch and boundary accents annotated using a loose ToBI-style system [27]. Fundamental frequency measures were plotted in semitones relative to each speaker's range and baseline. Microprosodic errors in pitch objects were manually corrected. Samples of both phonetic and interactional analyses of the collection were presented for discussion and validation at five different data and lab sessions during 2022.

4. LOOKING AT PHONETICS INTERACTIONALLY: EVIDENCE FROM "OH MY G-WORD"

Our interactional-phonetic analysis confirms the context-sensitivity of OMG while also revealing some interactional normativity. OMG interjections occupy identifiable positions in the sequential structure of interaction supporting the ongoing courses of action, from where they take part of its phonetic-prosodic design. The following account summarises the results of the qualitative interactional phonetic analysis of cases where OMG used by co-participants to affiliate through displays of surprise or sympathy. The examples illustrate how while there are recurrent features, other configurations of OMG are the result of prosodic matching and upgrading [29] of the design of surrounding turns. The table below summarises the constructions identified for the responsive OMG interjections:

INTERACTIONAL STRUCTURE OMG: responsive in second position (parentheses indicate optional components) A: [news informing / telling with + or - valence] B: **[(oh) (my) G-word]** [(assessment)] [(ritualised disbelief)] A: [(yeah/ I know) (elaboration or expansion of informing or telling)]	
I: affiliating with positive stance: surprise displays	**II: affiliating with negative stance: sympathy, dismay**
PHONETIC-PROSODIC FEATURE CLUSTERS *Oh*: often pre-glottalised, longer duration *My*: usually shorter in duration	
God: longer duration [g]: long hold phase, often aspirated/breathy release [d]: audible breathy or vowel offglide release	*God*: longer duration [g]: long hold phase (initially voiced), ejective-like release, aspiration/slow friction noise release [d] normally lenition and audible breathy or vowel offglide release
f0 contours: Sequence of two rise-falls L*+H L-? L*+H L-L% OH (MY) GOD	f0 contours: Sequence of two falls H* L-? H* L-L% OH (MY) GOD Sequence of two narrow falls (fall-to-mid) or narrow fall + level tone starting around estimated midline of speaker's range or higher H* !H-? ^H* H-H% OH (MY) GOD
Wider f0 excursion Pitch above estimated speaker's midline, sometimes falsetto	Narrower f0 excursion Pitch around speaker's midline Breathy and whispery voice qualities
Prosodic matching in subsequent assessments Prosodic matching with features of the prior turn news/telling	Prosodic upgrading in subsequent interjections. Prosodic matching with features of the prior turn news/telling.

Table 1: Summary of interactional and phonetic analyses of two collections of OMG.

In these contexts, OMG is used as a responsive item, pre-positioned to an assessment or a ritualised disbelief marker [13] or as a standalone item. OMG is not intrinsically positively or negatively valenced, and our collection shows that even when no assessment follows that could confirm that valence (compare OMG in Ex.1:line 11 with Ex.1:line 22 "cool"), co-participants' orientations to OMG (Ex.1:line 23 "I know") reveal them to be adequate responses, and this is the kind of endogenous evidence that CA-IL uses to ground their claims, with those recurrent prosodic features in similar contexts then coming to the fore as potential disambiguators.

For announced or retold events shown to bear positive valence, the responsive OMG interjection is designed with a cluster of phonetic-prosodic features that coincide with what prior studies have associated with displays of surprise: perceptually louder and higher in pitch level than in surrounding talk, a wider pitch excursion and series of two rise-fall contours, one completing its domain over "oh my", the other on "God". Example 1 shows a participant making a positive announcement (being able to access extra funding) formatted as recently newsworthy ("just", line 6) and with an incremental building of the unexpectedly positive aspect of the news (a pause after "money" in line 8, an increase in pitch in "eight", line 9).

Example 1: TBCF6239_Script

```
05 DEB: °hhh so `nO::w all <<all> i have to do> is my <<cr>ˇTHEsis,>
06      and <<all> Ive just found ¯OUT =
07      =that even if i just write a> `SCRIPT for it;=
08      =i can GET (.) ?u:hm:: ´M:Oney, (0.4)
09      like ↑`EIGHT thOusand <<cr>↓`DOllars::.> (.)
10      [SO:::  ]
11 SAR: [↑^OH my] ^GO::D.h°
12 DEB: °hh ¯SO::: (0.4) <<cr> uh like to lIve `ON i ´gUEss,=
13 DEB: =whIle im `WRITing i?.> (.) °hh
14      <<all> so `nOw im trying to> FIGure out whether i wAnna ¯DO:
15      (0.6)
16      the `SCRIPT <<all> that im working on> rIght ´NOW, =
17      or do a ˇDIFFerent one, =
18      =tha? (.) i? can Actually use sOme of that `MONey;
19      <<cr> to do RESearch fo:r it.> =
20      =[↓¯SO:]
21 SAR:  [^O::H] my. ^GOD. =
22      =[^THAT would be] sO:: <<cr> ^CO[:OL.]>
23 DEB   [°hhh          ]      [I   ] ^KNO::W.
```

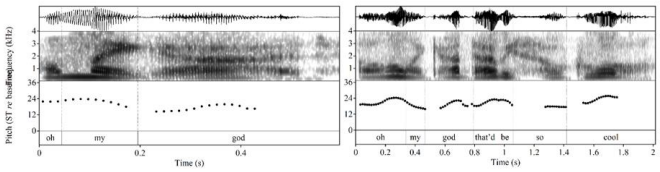


Figure 1: Transcription and acoustic visualisation (waveform, spectrogram, and f0 in ST scaled to speaker's range) of OMG tokens in Ex1 (lines 11 and 21)

At a projectably complete point in the turn where appreciation of the news is a relevant-next, Sarah responds with an OMG token in overlap with Debbie's trail-off "so" (line 10). "Oh" is relatively short (0.04 secs), and there is a reduction of tempo and vowel lengthening in the vowels of "my" and "god", whose final sound is produced with lenition and released audibly with glottal friction. The rise-fall contours do not exhibit great excursion (^1.7 ST ˇ4 ST for "oh"; ^5 ST towards the peak and ˇ2 ST down for "God"). Next, Debbie confirms the positive valence of the news (line 12) and elaborates on the positive implications (lines 14-18). In a timely manner once again at a point of potential completion and in overlap with the trail-off conjunction, Sara orients to the positive news with an upgraded "oh my God" with two wider rising-falling tones, with the pitch excursion spanning "oh my" being 7 ST, while that for "God" is of 5.5 ST. The stance positioning is now verbalised semantically through an explicit positive assessment ("that is so cool") that is prosodically matched to OMG in pitch contour and

loudness, marking coherence between the stance conveyed by both.

The design of OMG can also be locally-relevant with prosodic matching as above but with initiating turns by other speakers. In this case, speaker F2's OMG matches the pitch contour, falsetto voice quality and loudness increase of the announcement and displays alignment with the positive stance:

Example 2: TBCF5000_AnotherCall

```
09 F2: ((talks to someone else while waiting on the line))
10 F1: ´MAYa?
11 F2: (0.2)
12     ^YE::A[:h.]
13 F1:       [you] wOuldnt belIeve whO that ´WAS,
14 F2: `WHO:.
15 F1 <<f> ↑↑^RAchel.>
16     (0.2)
17 F2 ↑oh my <<f> ↑↑^GO:::D.uh°>
18 F1: ´I ^KN[O:W].
```

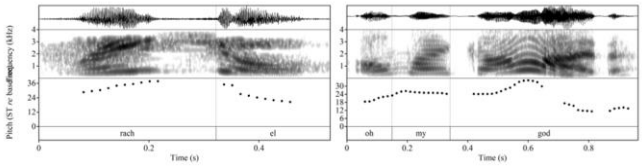


Figure 2: Acoustic visualisation (waveform, spectrogram, and f0 in ST scaled to each speaker's range) of news item + OMG token in Ex 2

OMG tokens also appear as responses to troubles-talk and tellings where experiences or individuals are ascribed negative valence, and are oriented to by displays of dismay or sympathy [16]. These tellings are escalated incrementally [30] until they reach a climax, and thus we may also see escalation and upgrading in co-participant responses. Several cases in this subcollection show lexical and prosodic upgrading from "oh God" to "oh my God", and while the first version is breathy or whispery and perceptually soft (piano) with lenition processes on "God", the second version is produced with modal voice quality a two falling-to-mid tones with minimal pitch excursion. These second versions tend to feature glottalisation on the onset of "oh", a longer vowel in "God", and a long hold phase to the plosive sound, as well as an ejective-like quality or a fricative release for /g/.

In Example 3, Stefanie projects a telling and a list ("so much has happened", line 1; "I mean", line 4), to which Belinda affiliates sympathetically with a first interjectional phrase ("oh no", line 3). After the first misfortune listed, Belinda orients with OMG token with a glottalized onset and a lengthened breathy vowel after a short period of hold of the plosive with a long friction release. "God" is produced with what can be perceived as falling intonation, but due to voice quality f0 cannot be tracked. Stefanie assesses the experience explicitly in a negative way followed by an account (lines 8-9). Belinda replies by grammatically embedding a contribution that displays prior access to Stefanie's experience, a form of affiliation which is confirmed and elaborated on by Stefanie. After a short silence, Belinda aligns with this negative stance with another OMG token (line 13). It also features a glottalized start for "oh", this time with its own falling movement (spanning 3 ST) and lengthened (0.4 secs). The /g/ sound is realised with a long hold phase (69 ms) but the release is ejective-like [g̰ʼʰ] with some friction noise before the onset of the vowel. The production of "my god" here is slightly stylised, with a fall-to-mid that could be initially heard as a long sustained level tone (ᵛ0.5 ST), still produced at a pitch high that is higher than the speaker's estimated baseline.

Example 3: TBCF6750_Car Battery

```
01 STF: ↑`SO much has hAppened to my ´CAR, =
02      since its been `UP here; = [bUt ^THA  ]:T.
03 BEL:                            [`Oh ↑^NO:.]
04 STF: (0.5) °hh i `ME:AN; (0.5)
05      `I:m ?ive °hhh had a ´GEAR go out On me, (0.3)
06 BEL: 2Oh `G(h°)O[:(hh°):D.°hh]
07 STF:           [`A::N       ]D; (.) ↑thAt? thats `BAD though;=
08      =<<all>because id gOne on> `TH(h°)IS c(h°)ircular ´ROAD,
09      and then cOuldn't ´GET anywhere,
10 BEL: oh THATs when it `BURNT;=`YE:AH.
11 STF: `YEAH; =and it wouldn't do anything.
12     (0.3)
13 BEL:`2O:H [my ¯G(k)O:::D]
14 STF:      [↑THIS was    ] jus:: (0.5) i dont `KNO:W. (.)
15      ?uh:m (0.5) <<all>i dont know> ˇWHAT it wAs,
```

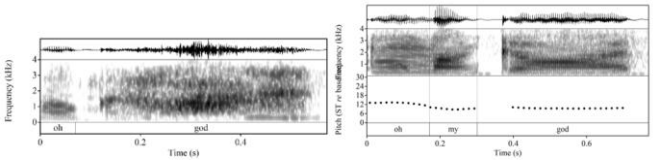


Figure 3: Acoustic visualisation of OMG tokens in Ex. 3

5. FINAL REMARKS

This exploratory interactional-phonetic analysis of (Oh) (My) G-word illustrated recurrent features as well as the variation between groups of responsive OMG interjections in a way that treats phonetics and prosody as interactional resources at the service of the current ongoing courses of action. The detailed CA-IL qualitative analysis of interaction allowed for the identification of endogenous evidence for issues of social action and affiliation, and for proof of the context-free but also context-sensitive role of phonetics-prosody in ways that can be later used to quantitatively investigate and predict possible clustering of features around certain points in the interactional structure. It shows how these features come to be in relation to surrounding talk and the interactional needs that arise moment by moment. If OMG tokens had been extracted from their context for a phonetic analysis, a lot of this interactionally-relevant detail that is intrinsic to our understanding of how phonetics-prosody allow us to "get things done" in interaction may have been blurred or lost.

6. REFERENCES

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] R. Ogden, "The Phonetics of Talk in Interaction," *The Cambridge Handbook of Phonetics*. pp. 657–681, 2021. doi: 10.1017/9781108644198.027.

[3] E. Couper-Kuhlen, and M. Selting, *Prosody in Conversation: Interactional Studies*. Cambridge University Press, 1996.

[4] E. Couper-Kuhlen and M. Selting, *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.

[5] E. A. Schegloff, "Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted," in *Roots of Human Sociality*, N. Enfield and S. Levinson, Eds. 2006, pp. 70–96.

[6] E. A. Schegloff, "Reflections on Quantification in the Study of Conversation," *Research on Language and Social Interaction*, vol. 26, no. 1, pp. 99–128, 1993, doi: 10.1207/s15327973rlsi2601_5.

[7] M. Dingemanse, "Interjections (Oxford Handbook of Word Classes)," Jun. 2021. doi:

[8] F. Ameka, "Interjections: The universal yet neglected part of speech," *J. Pragmat.*, vol. 18, no. 2, pp. 101–118, Sep. 1992, doi: 10.1016/0378-2166(92)90048-G. 10.31234/osf.io/ngcrs.

[9] E. Gehweiler, "11. Interjections and expletives," *Historical pragmatics*, pp. 315–349, 2010,

[10] S. A. Tagliamonte and B. L. Jankowski, "Golly, gosh, and oh my god! What North American dialects can tell us about swear words," *Am. Speech*, vol. 94, no. 2, pp. 195–222, May 2019, doi: 10.1215/00031283-7251241.

[11] M. Ljung, "The functions of expletive interjections in spoken English," in *Corpus Linguistics*, Brill, 2009, pp. 155–171. doi: 10.1163/9789042025981_010.

[12] L. J. Brinton, *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Walter de Gruyter, 2010.

[13] S. Wilkinson and C. Kitzinger, "Surprise As an Interactional Achievement: Reaction Tokens in Conversation," *Soc. Psychol. Q.*, vol. 69, no. 2, pp. 150–182, 2006, doi: 10.1177/019027250606900203.

[14] E. Reber, *Affectivity in Interaction: Sound objects in English*. John Benjamins Publishing Company, 2012.

[15] M. Dingemanse, "Between Sound and Speech: Liminal Signs in Interaction," *Research on Language and Social Interaction*, vol. 53, no. 1, pp. 188–196, Jan. 2020, doi: 10.1080/08351813.2020.1712967.

[16] J. Heritage, "Territories of knowledge, territories of experience: empathic moments in interaction," *The morality of knowledge in conversation*, vol. 29, pp. 159–183, 2011, [Online].

[17] J. Local, "Conversational phonetics: Some aspects of news receipts in everyday talk," *Studies in Interactional Sociolinguistics*, vol. 12, pp. 177–230, 1996,

[18] E. Couper-Kuhlen, "Exploring Affiliation in the Reception of Conversational Complaint Stories," in *Emotion in interaction*, M. Sorjonen and A. Peräkylä, Eds. Oxford: Oxford University Press, 2012.

[19] J. Freese and D. W. Maynard, "Prosodic features of bad news and good news in conversation," *Lang. Soc.*, 1998,

[20] A. Canavan and G. Zipperlen, "CALLFRIEND American English-Non-Southern Dialect," *Linguistic Data Consortium, Philadelphia*, vol. 10, no. 1, 1996.

[21] M. Selting, P. Auer, D. Barth-Weingarten, J. Bergmann, P. Bergmann, K. Birkner, E. Couper-Kuhlen, A. Deppermann, P. Gilles, S. Günthner, M. Hartung, F. Kern, C. Mertzlufft, C. Meyer, M. Morek, F. Oberzaucher, J. Peters, U. Quasthoff, W. Schütte, A. Stukenbrock, S. Uhmann, S., "A system for transcribing talk-in-interaction: GAT 2," *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*, vol. 12, pp. 1–51, 2011, [Online]. Available: http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2011/px-gat2-englisch.pdf

[22] E. A. Schegloff, *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. Cambridge University Press, 2007.

[23] R. Clift, P. Drew, and J. Local, "Why that, now?: Position and Composition in Interaction (Or: don't leave out the position in composition)," *Language, music and interaction*, pp. 211–232, 2013.

[24] R. Ogden and M. Cantarutti, "Coding categories relevant to interaction," in *Dimensions of Linguistic Variation*, M. Yaeger-Dror, C. Cieri, and K. Drager, Eds. Oxford University Press, 2021. In press.

[25] J. Local and G. Walker, "Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech," *Phonetica*, vol. 62, no. 2–4, pp. 120–130, Apr. 2005, doi: 10.1159/000090093.

[26] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program], Version 6.2.23*. Accessed, 2022. [Online]. Available: http://www.praat.org/

[27] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," *The OSU Research Foundation*, vol. 3, no. 30, pp. 255–309, 1997.

[28] B. Szczepek Reed, *Prosodic Orientation in English Conversation*. Springer, 2006.

[29] R. Ogden, "Phonetics and social action in agreements and disagreements," *J. Pragmat.*, vol. 38, no. 10, pp. 1752–1775, Oct. 2006, doi: 10.1016/j.pragma.2005.04.011.

[30] T. Heinemann and V. Traverso, "Complaining in interaction," *J. Pragmat.*, vol. 41, no. 12, pp. 2381–2384, Dec. 2009, doi: 10.1016/j.pragma.2008.10.006.

GAME OF PHONES: A SOCIO-PHONETIC ANALYSIS OF STYLISED MEDIA PERFORMANCE OF YORKSHIRE ENGLISH

Lucy Jackson

Glasgow University Laboratory of Phonetics, University of Glasgow
l.jackson.4@research.gla.ac.uk

ABSTRACT

Stylised linguistic performance in the media can provide an informed snapshot on how social meanings can be ascribed to phonetic variables [1]. This paper explores British English Yorkshire variety (YE) representation in the television series, Game of Thrones. Three actors' distinctive YE variable realisations were acoustically analysed. F1 and F2 measures were extracted from the two vowels' midpoints in the YE STRUT-FOOT merger [2]. The YE monophthongal GOAT and FACE were measured at the 25%, 50% and 75% trajectory points [3]. A lay-listener perception task [4] was conducted to assess YE performances and its socially indexical meaning.
Results found non-native YE actors to unsuccessfully produce the STRUT-FOOT merger and monophthongal variants. Listeners utilised YE variables to detect imitated performances and form social evaluations. Media reliance on phonetic cues to construct fictional social landscapes further entrenches the enregistered variety and its associated ideologies, highlighting a need for further research into media effects on language attitudes.

Keywords: vowels, sociophonetics, media, accents, performance

1. INTRODUCTION

Research into media influence on language attitudes is relatively undocumented [5, 6]. Sociophonetic variables have been found to function as valuable, stylistic resources amongst other semiotic cues (e.g., clothing and appearance) for both performative and interpretive means to navigate the social landscape and identity [7, 8]. This paper examines how broadcast media relies upon sociophonetic variation and its related ideologies as an indexical resource to construct fictional worlds and characters, and how such portrayals can further strengthen these indexical links. Its aim is to encourage further research into media effects on language attitudes and to contribute towards the growing discussion of accent bias in society, its implications [9], and media's role in perpetuating these linguistic ideologies through stereotypical accent representations.

1.1. Yorkshire English

Located in the North of England, the Yorkshire county is the largest county in the UK. There are multiple Yorkshire dialects that can be classified as possessing phonetic and morphosyntactic contrasts [10]. However, this paper utilises the term *Yorkshire English* (YE) to refer to the general representation of the regional variety in the media source. The term is also utilised in reference to the distinctive YE phonetic variables analysed in the paper, which do not appear collectively in any other geographic location than in Yorkshire (see section 2.1.1).
Yorkshire gained prowess in the industrial era of the 18th century onwards due to its textile, steel manufacturing and coal mining trade [11, 12, 13]. This resulted in densely populated industrial cities with a large working-class to meet the demands of the labour workforce. With a population greater than Scotland (over 5.3 million) [14], modern-day Yorkshire faces several social issues: with the Yorkshire city of Bradford ranking highest in Europe per crime index [15] and one third of all Yorkshire children facing poverty [16]. The social history of Yorkshire has arguably accounted for resulting social evaluations of Yorkshire speakers as being of lower status and prestige [17]. Contrastingly, the Yorkshire accent is often rated highly in terms of social attractiveness [18] as one of the most trustworthy British accents. Such ideologised values of the Yorkshire accent as 'honest' have caused YE to gain a market value, with many UK call-centres strategically placed in the region [19].

1.2. Game of Thrones

Adapted from the bestselling works of George R.R Martin, the HBO television series *Game of Thrones* (2011-2019) spanned 8 seasons and became a prominent figure of pop culture. A fictional fantasy world containing medieval warfare, magic and dragons, the screen adaptation drew on sociophonetic variation to aid the audience's sociocultural understanding of the Game of Thrones (GOT) world. The sociolinguistic context is evident in the depiction of the two main feuding families in GOT for overall power of the Westeros Kingdom. The Lannisters are

the most powerful and wealthy family, located in the Southern Kindgom and speak the standard received pronunciation (RP) accent. Although only less than 3% of the real-life population speak RP [20], ideologised values of its speakers as being of high status and intelligence [21] has proven a key tool to create characters of dominance in fictional media [22]. The Starks are the opposing family in the North who are depicted on-screen with the regional YE variety. The Starks' character depiction also parallels with the YE variety perception of being lower in status but, are an extremely loyal family vying for justice. The use of social-indexical variation with existing real-world ideological meanings lays the social groundwork for audiences to access immediate contextual knowledge of the fictional world. This is a growing stylistic tool employed across media platforms, present even in computer games [23].

2. METHODS

The present study employs a two-pronged approach when examining YE representation in the *Game of Thrones* (GOT) media series. The investigation comprises of an acoustic analysis of three actors' YE portrayals in GOT, and a lay listeners' social evaluation (LSE) task of the portrayals.

2.1. Materials

The speech of 3 male actors portraying YE in GOT were analysed: Sean Bean (native YE speaker), Kit Harrington (Standard Southern British English: SSBE speaker) and Richard Madden (Standard Scottish English: SSE speaker). Speech was extracted across 8 seasons of GOT (.mp4 file), video was excluded to minimise perceptual effects in the subsequent evaluation task. Resulting audio files were uploaded into Praat [24] for the data selection process.

2.1.1. Data Selection

In line with previous research into accent imitation in the media [25], this study employed the following selection criteria for speech samples: 1) Speech is in a monologue format (minimum 10 seconds in duration), with no interruptions from external actors/events. 2) No contextual references to the character or plotline. 3) Samples are in full sentences with grammatical structure. This is to ensure listener comprehension and minimise any perceptual effects in the LSE.

The distinctive YE variables functioned as controls for the speech samples, with at least 5 tokens of each YE variable present in each actor's speech compilation (table 1). The compiled recordings of each actor's YE portrayals are 1 minute in duration and utilised as stimuli in the LSE.

YE Variables	Examples	RP	YE
STRUT vowel	*'cut, blood'*	/ʌ/	/ʊ/
FOOT vowel	*'good'*	/ʊ/	/ʊ/
GOAT vowel	*'no, throne'*	/əʊ/	/oː/
FACE vowel	*'name, day'*	/eɪ/	/eː/
Intervocalic /t/	*'battle'*	/t/	/ʔ/

Table 1: YE variables used in speech samples and its minimal realisations in RP and YE.

2.2. Acoustic Analysis

Acoustic results for four YE vowels were reported: STRUT, FOOT, GOAT and FACE vowels [26]. Note that the present study employs the lexical sets [27] for analysing inter-speaker differentiation of YE variable production. An overall distribution of 78 YE tokens were analysed, with an equal distribution of tokens realised by each actor (overall tokens per variable: STRUT = 24, FOOT = 24, GOAT = 15, FACE = 15).

2.3.1. Procedure

All YE tokens were manually labelled in Praat using a TextGrid. To analyse the STRUT and FOOT vowels and their merged nature in YE, F1 and F2 formant measurements were measured by hand from bi-syllabic and monosyllabic lexical items [2]. To examine the monophthongal YE GOAT and FACE variants, the F1 and F2 formant measurements were measured from the 25%, 50% and 75% trajectory points throughout the vowel. Distances between the trajectory points were measured to assess the degree of movement across the vowels [3]. The acoustic data were visualised and statistically analysed using R [28]. Vowels were normalised via the Fabricius et al method [29] and were subsequently visualised utilising the Visible Vowels software [30].

2.3. Listener Social Evaluation (LSE) Task

A total of 18 participants were recruited for the LSE (50:50 native and non-native Yorkshire). The native YE group were aged 22-56 (mean = 30.4) and the non-native YE group were aged 21-26 (mean = 23.5). Conditions were not controlled for gender: native YE group (4F, 5M), non-native YE group (7F, 2M). Listeners were classified as native YE if they received both their primary and secondary education in Yorkshire [31] and had 10 years of uninterrupted residence in the region. All listeners were monolingual speakers of English.

2.3.1. Task Design

Participants were presented the speech compilations of the 3 actors' YE portrayals in GOT as auditory stimuli via Sony SX310 headphones. To eliminate order effects on listener responses, participants were split into three conditions in which 3 actors would be presented in alternating orders. After each stimulus was presented, participants would answer a questionnaire that aimed to elicit both qualitative and quantitative findings. Likert-scale ratings were employed to assess listener ratings on the YE portrayals and the accent's credibility. Participants were able to provide qualitative answers on lay-listener perceptual judgements of the performance, the phonetic variables utilised to form judgements, and the social evaluations/attitudes (status and social attractiveness) of the YE characters from auditory stimuli only. The overall duration of the social evaluation task was no longer than 15 minutes (table 2).

Task	Duration
Stimulus 1	1 minute
Sub-questionnaire	8 questions
Stimulus 2	1 minute
Sub-questionnaire	8 questions
Stimulus 3	1 minute
Sub-questionnaire	10 questions

Table 2: LSE Task Design

3. RESULTS

3.1. Acoustic Analysis

Extensive inter-speaker variability amongst the actors is observed in their STRUT-FOOT realisations (figure 1). Native Yorkshire actor, Sean Bean (SB) produces both STRUT and FOOT in a centralised position with consistent merger, as expected in YE accents. Scotsman Richard Madden (RM) and SSBE speaker Kit Harrington (KH) are unable to produce STRUT-FOOT productions with the positional accuracy needed and are unable to consistently merge the vowels. KH displays instances of hypercorrection with multiple backed FOOT tokens, whilst both actors' productions display acoustic similarity to their own native speech counterparts.

Minimal vowel trajectory was observed by SB in both his GOAT and FACE YE monophthongal variants (Figures 2 and 3). RM and KH both produce the vowels with substantial trajectory movement incompatible with YE monophthongs. Overall, the non-native YE actors were unable to imitate YE speech and produce the distinctive variables with acoustic similarity to native YE speech. RM and KH displayed instances of hypercorrection and inability to alter production from their native accents. YE portrayals in GOT are acoustically inconsistent amongst the actors.

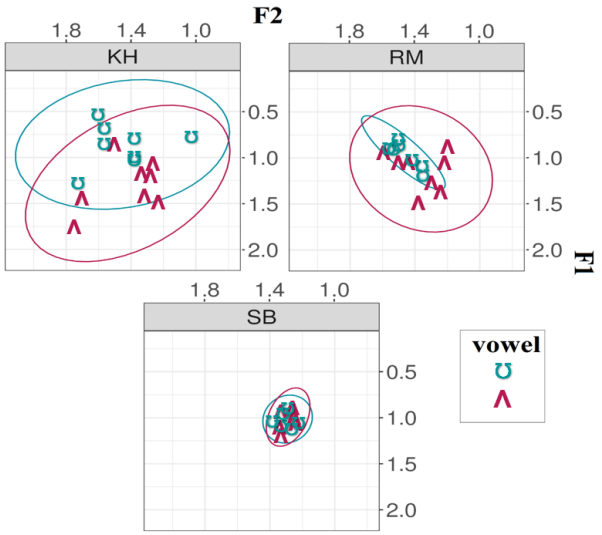


Figure 1: STRUT-FOOT distribution in GOT

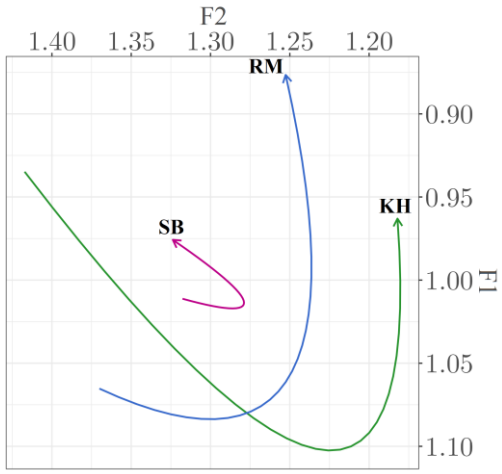


Figure 2: GOAT trajectory in GOT

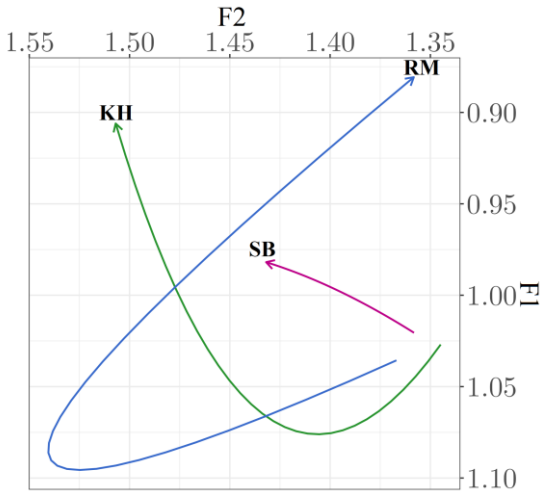


Figure 3: FACE trajectory in GOT

3.2. Listener Perception Task

The majority of listeners across both conditions identified the 3 actors' speech performance to be in a YE accent, with RM and KH receiving mixed results due to their acoustic inconsistencies (e.g., Yorkshire, Manchester, Northern English portrayals). Table 3 shows highest perceptual ratings for SB in relation to a credible YE accent. Non-YE actors received greater variability amongst perceptual ratings, with notable higher ratings for KH by non-YE listeners and conversely, higher ratings for RM by the YE listeners.

Actor	Native YE Listeners		Non-YE Listeners	
	Mean	SD	Mean	SD
SB	4.6	0.5	4.3	0.71
KH	2.7	0.97	3.2	1.39
RM	3.2	1.39	2.4	1.13

Table 3: Credibility Ratings of YE Portrayals

Table 4 demonstrates a high success rate in accurately identifying SB as native YE. Over half of listeners correctly identified KH origins to be London/Southern England. RM performed better in concealing his origins from the listeners. However, all Scottish listeners in the non-YE condition accurately identified his origins, with lay-person qualitative feedback on short vowels, which can be assumed in reference to the Scottish Vowel Length rule [32].

Actors	Overall	Native YE Listeners	Non-YE Listeners
SB	83%	88.8%	77.7%
KH	55.5%	44.4%	66.6%
RM	33.3%	11.1%	22.2%

Table 4: Accuracy rates on Speaker Origin Identification

Listeners showed meta-awareness of phonetic characteristics to form their perceptual judgements: *'The 'ow' in throne sounds too forced… like a posh person trying to sound rough'* [discussing SSBE speaker, KH speech]. *'The way he said the 'O' in home sounded like how I'd say it…really elongated and flat. A bit like the 'a' in chase too'* [Discussing native YE speaker, SB speech].

At the end of the social evaluation task, listeners provided qualitative answers on their expectation of the YE characters' appearances and behaviours. Qualitative data found patterns in YE character descriptions to be *'rough, dirty, strong, warriors'* and lacking any regality or finery. Overall, character behaviour expectations were loyal (*'a man of his word'*), surviving hardship and intent on fighting for justice and family.

4. DISCUSSION

4.1. Stylised YE portrayal in Game of Thrones

The acoustic analysis demonstrates non-native YE actors' unsuccessful attempts of credibly performing the YE variables: lacking in consistency in merging the STRUT-FOOT set or minimal vowel trajectory movement in GOAT and FACE. Lay listeners were found to be able to report on fine-grained phonetic cues to identify YE portrayals, perceptually rate credible YE speech and even detect actors' accent and geographic origins.

The salient, indexical link between a linguistic variety and its social meaning was exemplified through listeners' social evaluations of characters, which paralleled with present societal attitudes of the enregistered YE repertoire [33]. These social evaluations of predicted character appearance and behaviour were similar to actual YE character depictions in GOT, demonstrating how semiotic resources (linguistic and non-linguistic) mutually interact to perform and interpret social identity and personae. Arguably, the most iconic phrases from GOT are *'Winter is coming'* and *'You know nothing, Jon Snow.* Both phrases are frequently produced by 'northern' voiced characters that realise the YE merger set and monophthongal GOAT variables. With GOT becoming a global pop culture phenomenon, YE was commodified via these iconic stylised utterances marketed onto merchandise (e.g., clothing, cups, board games) [34].

Media reliance on sociophonetic variation and its indexical meaning to construct character identities and fictional worlds has reinforced the YE variety and its ideologies as a heavily typified form [33]. This reliance results in the mass-scale reinforcement of the variety and its ideological values in the social landscape and quotidian social interaction [34, 7].

4.2. Wider Implications

This small-scale study has shown how sociophoentic analysis of media accent portrayal can inform on sociophonetic variation and language ideologies in interaction, and the permeance of accent bias and its real-life implications [35]. Accent discrimination has been found to impact job prospects [9], and the likelihood of being convicted of a crime [36].

Research into media performance can inform on sociophonetic theory and our understanding of media influence on accent bias, through its use of the phonetic signal and its social indexicality.

5. REFERENCES

[1] Eckert, P. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language (Baltimore).*

[2] Jackson, L., Gold, E. 2018. An investigation into Bradford speakers' ability to imitate a SSBE accent: Implications for forensic phonetics. *IAFPA 2018,* Huddersfield UK.

[3] Earnshaw, K., & Gold, E. 2018. Variation in the FACE vowel across West Yorkshire: Implications for forensic speaker comparisons. *Interspeech 2018*, Hyderabad.

[4] Shen, C., Watt, D. 2015. Accent categorisation by Lay Listeners: Which Type of 'Native Ear' Works better?

[5] Androutsopoulos, J. 2014.Mediatization and sociolinguistic change. Key concepts, research traditions, open issues. In *Mediatization and Sociolinguistic Change,* 3-48. Berlin: De Gruyter

[6] Stuart-Smith, J., Pryce, G., Timmins, C., Gunter, B. 2013. Television can also be a factor in language change: Evidence from an urban dialect. *J. Ling. Soc. Am. 89,* 501-536.

[7] Eckert, P. 2008. Variation and the indexical field. *Journal of Sociolinguistics, 12*, 453-476.

[8] Moore, E., Podesva, R. 2009. Style, indexicality, and the social meaning of tag questions. *Language in Society, 3,* .447-485.

[9] Sharma, D., Levon, E., Watt, D., Ye, Y., Cardoso, A 2019. Methods for the study of accent bias and access to elite professions. *J. Language and Discrimination*, 150-172.

[10] Haigh, S. 2015. Investigating Regional Speech in Yorkshire: Evidence from the Millennium Memory Bank.

[11] Baxter, J.L. 1977. The origins of the social war in South Yorkshire – a study of capitalist evolution and labour class realization in one industrial region. PhD thesis, University of Sheffield.

[12] Koditschek, T. 1990. *Class formation and urban industrial society: Bradford, 1750-1850.* Cambridge University Press.

[13] Blewett, M. 2006. Yorkshire lasses and their lads: Sexuality, sexual customs and gender antagonisms in anglo-american working-class culture. *J. Social History, 40*(2),317-336.

[14] Yorkshire Network – Facts and Statistics. https://www.yorkshirenetwork.co.uk/yorkshire/yorkshire-facts-and-statistics/.

[15] Numbeo Crime Index. https://www.numebo.com/crime/region_rankings.jsp?title=2022&region=150.

[16] The Yorkshire Post. https://www.yorkshirepost.co.uk/lifestyle/money/one-in-three-yorkshire-children-living-in-relative-poverty-shock-figures-show-3764495.

[17] Coupland, N., Bishop, H. 2007. Ideologised values for British accents. *J. Sociolinguistics, 11*(1), 74-93.

[18] McKenzie, R. M., McNeill, A. 2022. *Implicit and Explicit Language Attitudes: mapping linguistic prejudice and attitude change in England.* Routledge.

[19] Beal, J.C., Carmela, N., Massimo, S. 2009. Perspectives on Prescriptivism. *English Language and Linguistics, 13*(3), 510-516.

[20] Trudgill, P. 2000. *Sociolinguistics of Modern RP.* University College London.

[21] Edwards, J., Jacobsen, M. 1987. Standard and regional standard speech. *Language in Society, 16*(3), 369-379.

[22] Lippi-Green, R. 2012. *English with an Accent: Language, ideology and discrimination in the United States.* London: Routledge.

[23] Kjeldgaard-Christiansen, J., Henjá, M. 2022. The Voices of Game Worlds: A Sociolinguistic Analayis of Disco Elysium. *Games and Culture, 0*(0).

[24] Boersma, P., Weenink, D. 2022. Praat: Doing Phonetics by Computer [Computer program]. Version 6.3.03. Retrieved 17 December 2022 from http://www.praat.org/.

[25] Knooihuizen, R. 2016. Accuracy and acceptability of second-dialect performance on American television. *English Language and Linguistics, 23*(2), 229-252.

[26] Haddican, B., Foulkes, P., Hughes, V., Richards, Hazel. 2013. Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change, 25,* 371-403.

[27] Wells, J. C. 1982. *Accents of English: An Introduction.* Cambridge: Cambridge

[28] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https;//www.R-project.org/.

[29] Fabricius, A. H., Watt, D., Johnson, D. E. 2009. A comparison of three speaker-intrinsic vowel formant frequency normalisation algorithms for sociophonetics. *Language Variation and Change, 21*(3), 413-435.

[30] Heeringa, W., Velde, H.V. 2017. Visible Vowels: A Tool for the Visualisation of Vowel Variation. *Interspeech.*

[31] Gold, E., Ross, S., Earnshaw, K. 2018. The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Interspeech 2018,* Hyderabad.

[32] Trudgill, P., & Hannah, J. 1982. *International English.* London/New York: Arnold.

[33] Agha, A. 2003. The social life of cultural value. *Language & communication, 23*(3), 231-273.

[34] Johnstone, B., Andrus, J., Danielson, A. E. 2006. Mobility, Indexicality, and the Enregisterment of "Pittsburghese". *Journal of English linguistics, 34*(2), 77-104.

[35] Sharma, D., Levon, E., Ye, Y. 2022. 50 years of British accent bias. *English World-Wide. A Journal of Varieties in English.*

[36] Paver, A., Wright, D., Braber, N. 2021. Accent Judgements for social traits and criminal behaviours: ratings and implications.

TEACHING LANGUAGE RHYTHM BY MUSIC TO OLDER ENGLISH LEARNERS OF ITALIAN

Laura Zambianchi[1] and Hae-Sung Jeon[2]

School of Psychology and Humanities, University of Central Lancashire, UK[1, 2]
LDZambianchi1@uclan.ac.uk[1], HJeon1@uclan.ac.uk[2]

ABSTRACT

This paper qualitatively investigates the benefits of explicit instructions on rhythm in foreign language teaching. The study involves teaching Italian to 60+ years old native English speakers. The effects of explicit instructions using musical activities, syllable cards, and body synchronisation on learners' experience were examined. The study was designed around a series of novel educational interventions with regular semi-structured interviews and focus groups with participants, in addition to their regular Italian lessons. Participants reported benefits of the novel approaches in supporting their learning. The present study forms a basis of further empirical investigation on pedagogical development.

Keywords: speech production, syllables, rhythm, pedagogy, older learners

1. INTRODUCTION

In foreign language acquisition, adult learners face particular challenges in learning phonology and phonetics of the target language(s) [1]. In foreign language classroom, however, phonology and phonetics tend not to be taught explicitly for various reasons including that teachers are not sufficiently trained in this area [2, 3].

The present paper reports a qualitative study on teaching and learning rhythm in Italian, using explicit instructions and musical activities. Interpretative phenomenological analysis (IPA) was used (Section 2.2). IPA combines phenomenology (i.e. the study of experience), hermeneutics (i.e. the study of interpretation), and idiography (i.e. a description of the individual). IPA enables a two-stage interpretation process, with the participants trying to make sense of their world, and the researcher trying to make sense of the participants trying to make sense of their world [4].

The teacher involved her students in the research; the study participants were native English speakers who were 60+ years old with French as their dominant second language (L2), learning Italian as their third language (L3). Although incorporating music into language teaching is not a new method [5], the present study shows a case of older learners acquiring so-called 'rhythmically similar' foreign languages which has not been extensively investigated. Despite the seeming rhythmic similarity between French and Italian [6], the learners still found learning Italian rhythm challenging. In addition, although older learners often face challenges in language learning caused by decline in memory capacity [7], they evaluated that embedding music in regular Italian lessons was helpful for improving their pronunciation together with memory retention and recall.

1.1 The syllable in Italian

Italian is often cited as a syllable-timed language showing syllabic isochrony [6, 8, 9]. However, not only the syllable-timing of Italian but also the rhythm typology has been questioned [10, 11, 12, 13].

The impression of syllable-timing seems to be partially caused by the relatively simple syllable structure. Although Italian permits a complex onset (e.g. CCC in spregiare, 'to despise'), [14] reported that 96.40% of the syllables in the LIP (Frequency Lexicon of Spoken Italian) were CV, CVC, V, VC, CVV, or CCV out of 19 possible types.

Despite the simple syllable structure, locating a syllabic boundary is not a trivial matter even for native Italian speakers. The syllable division rules proposed in 1969 by the Ente Nazionale Italiano di Unificazione (an Italian association which performs regulatory activities) state that diphthongs must not be split, and therefore we have, for example, au.to, not a.u.to ('car') and vio.la, not vi.o.la ('purple'). We can, however, split vocalic clusters forming a hiatus, as in a.e.ro.pla.no ('plane') and po.e.ta ('poet'). Sometimes native Italian speakers are uncertain about the intuitive division of syllables. For instance, viola can be considered either a disyllabic word vio.la with a diphthong or a trisyllabic word, vi.o.la with vowel hiatus [15].

1.2 Lexical stress in Italian

Italian words are classified into five categories depending on where the stress is. (In this paper, stressed syllables without orthographic stress marking are capitalised.) The five categories are:

- *tronche*: word-final, *sofà* (‘sofa’), *caffè*
- *piane*: penultimate, *UOmo* (‘man’), *LIbro* (‘book’)
- *sdrucciole*: antepenultimate, *TAvolo* (‘table’), *veRIfica* (‘he/she verifies’)
- *bisdrucciole*: fourth-last, *veRIficano* (‘they verify’), *VENdimelo* (‘sell it to me’)
- *trisdrucciole*: fifth-last syllable, *REcitamelo* (‘recite it to me’)

Stress assignment in Italian is, in general, not predictable in polysyllabic words, and it is predictable from orthography only in a small number of cases. The limited predictability of stress assignment is an issue with which every speaker has to deal [16]. For assigning stress in polysyllabic words, language users tend to rely on the frequency information (i.e. biased towards frequent *piane* pattern) and phonological analogy (e.g. deriving the stress pattern from other words with the same final phonemic sequence) [17].

1.3 Teaching and learning Italian syllables and stress

In Italy, recognising syllables is considered a very important step for learning reading and writing [18]. In primary school, children are explicitly taught the syllabification rules. They are taught that, for instance, geminates are always divided, e.g. *can.ne* (‘reeds’), *far.ro* (‘spelt’), and that, for consonant clusters formed by three or more consonants, the boundary is placed before the second consonant: e.g. *con.trol.lo* (‘control’), *al.tro* (‘other’).

It is probably not surprising that English learners of Italian find it challenging to learn the syllable division and stress assignment (See Sections 1.1 and 1.2). For instance, in [19], English learners of Italian carried out a reading aloud task for non-words with penultimate, antepenultimate, and ambivalent stress neighbourhood sequences in Italian. Learners heavily relied on the dominant stress pattern in Italian, *piane*, assigning stress to the penultimate syllable, although this tendency decreased as their Italian vocabulary size increased (also see [20]). Because most Italian words are *piane* [21], students often make an error producing stress in the penultimate syllable for *sdrucciole*, *bisdrucciole*, and *trisdrucciole*. For instance, *Stefano* is a *sdrucciola* (*STE.fa.no*), but learners tend to mispronounce it as *Ste.FA.no*.

Learners often ask questions such as “why in Italian are accents in writing compulsory only to indicate stress on a word-final vowel, and not on the antepenultimate?” [16]. For writing, words with stress on the last syllable must carry a graphic accent (i.e. *città*, ‘city’). However, the graphic marking of the non-final stress is optional. Therefore, *principi* in writing can be two words differentiated by stress assignment, *PRINcipi* (‘princes’) vs *prinCIpi* (‘principles’), in speech.

Finally, we observed that native English speakers seemed to struggle to perceive and produce geminates, e.g. *capello* (‘hair’) vs *cappello* (‘hat’). Producing a geminate (e.g. *pappa*, ‘mush, baby food’) as a singleton counterpart (e.g. *papa*, ‘pope’) is a rhythmic error changing a word’s meaning.

2. METHODOLOGY

2.1. Participants

Participants were eight 60+ years old native speakers of British English (P1–P8) with varying degrees of proficiency in other languages: all of them studied one or two other European languages (French, German, Spanish, Russian) and Latin, and three of them were language teachers in secondary schools in the North West of England. They were keen language learners. For the majority of participants, French was a dominant L2 (four studied it to degree level, three throughout secondary schooling, one studied it through primary schooling and has been studying it informally in recent years). All participants were learning Italian as L3 by attending the author LZ’s recreational classes (2 hours per week) as a group for a minimum of four years.

2.2. Data collection and analysis

The author LZ, who is a native Italian speaker and the teacher for the participants, carried out semi-structured interviews and led focus groups. Interviews were carried out individually and, on one occasion, with two learners in tandem. Each interview lasted for approximately 40 minutes. The interviews were audio-recorded and manually transcribed.

First, interviews with each participant and one focus group meeting took place; participants discussed their love for the language, their experiences with Italian and with other foreign languages they had acquired and the obstacles they faced. After these meetings, regular conversations and class activities were carried out. The interviews were scheduled in between classes (eight intervention classes) to monitor participants’ progress and reflection about the novel teaching and learning methods. In addition, participants regularly emailed about their experiments with short rhymes (and suggestions for more activities and songs to try out) to the author LZ, showing motivation and desire to explore further. This paper describes the activities which took place over six months.

After the meetings with participants, a list of significant statements was identified for the interpretative phenomenological analysis (IPA) which aims to capture participants' detailed examinations of personal experience without theoretical presumptions [22]. IPA is a rigorous approach employing in-depth qualitative analysis and requiring continuous and active reflexivity of researchers and participants [23]. The participants' experiences of learning the Italian language and their perceptions of their progress, rather than effective approaches to language learning, were focused on in the data collection process. The transcribed notes were subsequently categorised into themes. The final narrative included both the learners' account of their experiences in their own words and interpretative commentary. The study was approved by the ethics committee of the University of Central Lancashire (reference BAHSS0299).

2.3 Teaching materials and activities

In the classes, three methods were broadly used; musical activities, syllable cards, and body synchronisation. For all activities, the teacher emphasised the stressed syllable, particularly for *parole sdrucciole* and words which a number of students identified as having 'too many syllables'. The positive effects of incorporating music into language learning is well-known [24, 25]. For instance, a large body of research shows the benefit of singing in the target language(s) [26] and synchronisation [27].

For musical activities, for instance, the teacher presented the national anthem of Italy upon request of participants who wanted to consolidate the lexicon of kitchen objects. The song starts with *Fra.TEL.li d'I.TA.lia*, with the stress on the 2nd and the 5th syllables. Then participants carried out activities aligning alternative words with the same rhythmic structure producing combinations such as *col.TEL.li* ('knives') and *for.CHET.te* ('forks'), singing and tapping for each syllable with pens.

Another specific example is a full teaching session based on *La donna è Mobile* ('The woman is fickle') from Giuseppe Verdi's *Rigoletto*. The learning activities were designed to emphasise syllable division and stressed syllables. The aria was initially played to students, who were provided with the text, and they immediately started to sing along. The first listening was followed by activities of word stress recognition (e.g. *MObile*). As shown in Fig. 1, there is a correspondence between musical notes and syllables. The third note (D) is aligned to *na* and *è* because of synaloepha [28] which is a contraction of two adjacent vowels into one syllable. *Mo.bi.le* corresponds respectively to the musical notes F, E, and C. The teacher emphasised the importance of the F note: the stressed syllable in *MObile* is aligned to the highest note (F) and this allowed learners to link the stress and melodic prominence.

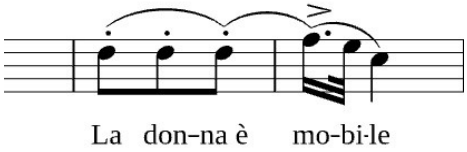


Figure 1: Example of musical score for *La donna è mobile*

For the following activity, learners replaced *MObile* with other trisyllabic adjectives. The students realised that *COmica* would work because it is a *sdrucciola* like *mobile*, whereas *feLIce* would not because it is *piana*, as the *piana* creates a mismatch between stress and the musical prominence (P4: "*La donna è còmica*… It fits..."; P5: "*La donna è felìce*… It doesn't work… Because it's *piana*").

Other examples used for teaching include folk songs (for instance *Bella ciao,* suggested by participants), a poem put to music using the Marseillaise tune, the Italian anthem and arie d'Opera.

Activities with paper cards were used to increase students' awareness of how a polysyllabic word is decomposed into syllables. The teacher prepared paper cards, each card showing a syllable (*ca, vo, lo, etc.*). Participants, in pairs, formed as many words as they could, using the cards (e.g. *cavolo* 'cabbage', *casa* 'house'). At the end of the activity each group shared their words. For words that they found difficult to pronounce and to recall (for instance, students often made a mistake of producing *ca.me.riè.re,* 'waiter' as *ca.mie.re*), they practised pronouncing them with body movement such as clapping and tapping to each syllable (using pens or other objects).

3. RESULTS AND DISCUSSION

The themes identified at the initial interviews included: desire to keep the brain active; high level of competence in other foreign languages; learning Italian later in life and without having to take formal exams; positive perception of the "Italian learning journey"; feeling of having reached a plateau (P5 blames the pandemic and the lack of trips to Italy where they could use the language); their age being partly responsible for memory decline; comparison with being a young learner ("When I was young I

could sit down and learn lists of words without context, but now it doesn't work").

The participants reported their high motivation with a willingness to try anything to improve. All eight participants acknowledged a "change in learning" due to ageing and the fact that they are no longer forced to learn reams of vocabulary as they did for school exams.

Importantly, at early interviews participants identified learning vocabulary and its rhythm as the biggest challenge; they commented that Italian stress is the most difficult part, not knowing how to stress an unknown word (P6: "The pronunciation is just difficult!"; P3: "Words which come from the same root as English are fraught with danger. Thus, A*cido, ..., amBIguo* and *GONdola* are easy because the stress is the same [as in *acid*, *ambiguous*, and *gondola* in English]. However, words like *aCRObata, aNEDdoto, ..., inCOgnito, oRIgano* and *stereOtipo* are trickier: they don't exactly trip off the tongue!"). That is, older learners experience difficulties in L3 vocabulary retention and recall due to a memory loss and a scarcity of strategies to facilitate the learning process. Despite that fact that all participants had experience in learning French which is considered rhythmically similar to Italian [6, 29], the learners reported that their experience in French did not directly benefit their Italian (P1: "I studied French at University level. I never use it with Italian because I know it doesn't help! I usually put the stress on the penultimate syllable because I was told that it works more times than it doesn't, but I am finding out it's not the case!"). The majority of the participants expressed the desire to "experiment" with pronunciation through classical music.

As they were progressing with their learning, they reported the benefits of explicit instructions with music. In particular, Western classical music with clear metrical structure seemed to be useful as educational materials for improving learners' awareness about the syllable division and stress assignment (P3: "The stress on words is my weak point, but the rhyme strategies and music really help"). One participant reported significant improvements in word recall following the session on *La donna è mobile* (P2: "I wrote something using rhyme and rhythm. I did *La donna è mobile* with all the *-abile, -ibile* etc. words; I did something to try help me with the meaning of *pronto* and *presto*, and that worked. It was a rhyme and has worked").

Using the explicit instruction on the syllable division using the cards and movements was also proved to be useful for learning geminates: for instance, when learners break words like *cap.pel.lo* ('hat') into syllables, they were "forced" to produce and perceive the geminates.

The participants' broad evaluation of the musical activities was positive (P7: "without realising it, it's already in your head"). Participants indicated that classical music, due to catchy melodies, a dramatic build and some artists' enunciation, is particularly suited for Italian language learning. Everyone admired the clarity of Luciano Pavarotti's diction, "*as clear as a bell"* [30], as the tenor emphasises stressed syllables in his singing.

The Marseillaise tune, used by Italian songwriter Sergio Endrigo to accompany a Gianni Rodari's children poem called *Napoleone*, also received positive feedback from participants, because they were familiar with the music: P5 said they were having breakfast, clicked on the link which was sent by LZ and stopped eating. They wanted to give full attention to the song because they recognised the music. When P7 was asked "Would it have been the same with an unknown tune?" they responded: "No, not for me".

4. CONCLUSIONS

The present study shows that older learners of Italian benefit from explicit training in syllable division and rhythm. After the explicit training using classical music, syllable cards and body synchronisation, all participants reported to have gained an awareness of syllable structure. Syllable-timing is not a satisfactory label for Italian [31] and Italian has a complex stress system. Explicit instructions on the language-specific syllable division and stress assignment rather than the broad rhythm typology classifying languages into stress-timed or syllable-timed would help learners.

The activities described here show how syllable and stress rhythm can be taught effectively for older learners. They reported that the instructions were particularly helpful for improving their understanding of Italian geminates and stress assignment. Understanding syllables and stress seems to be an important step for learners to develop their strategies to enhance memory retention and recall.

The present study took a qualitative approach. Further study investigating the measurable effects of the training is required. Finally, the present study shows that pedagogical development combining knowledge in phonetics and music can lead to fruitful outcomes.

5. ACKNOWLEDGEMENTS

The authors would like to thank Dr Richard Davies for his invaluable feedback on this project.

6. REFERENCES

[1] Savy, R., Alfano, I. (a cura di) 2016. *La fonetica nell'apprendimento delle lingue.* Milano, Officinaventuno.

[2] Calabrò, L. 2015. Il workshop di fonetica in italiano L2/LS. *Italiano LinguaDue* 7, 40–49.

[3] Grimaldi, M. 2017. L'efficacia dell'istruzione scolastica nell'apprendimento fonetico-fonologico della L2: prospettive d'indagine e applicazioni didattiche. *Studi Italiani di Linguistica Teorica e Applicata (SILTA),* XLVI (1), 109–126.

[4] Smith, J. A., Osborn, M. 2004. Interpretative phenomenological analysis. In: G. M. Breakwell (ed.), *Doing Social Psychology Research,* Blackwell Publishing; British Psychological Society, 229–254.

[5] Pasqui R. 2003. L'utilizzo della canzone in glottodidattica. *Bollettino Itals*, available at: https://www.itals.it/lutilizzo-della-canzone-glottodidattica (accessed Apr. 26, 2023).

[6] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal, *Cognition* 73(3), 265–292.

[7] Murman, D.L. 2015. The impact of age on cognition. *Semin. Hear.* 36(3), 111–121.

[8] Bertinetto, P. M. 1981. *Strutture prosodiche dell'italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici.* Firenze: Accademia della Crusca.

[9] Bertinetto, P. M., Magno Caldognetto, E. 1993. Ritmo e intonazione. In: A.A. Sobrero (a cura di), *Introduzione all'italiano contemporaneo. Le strutture.* Roma-Bari: Laterza.

[10] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalysed. *J. Phon.* 11, 51–62.

[11] Vayra, M., Fowler, C., Avesani, C. 1987. Word-level coarticulation and shortening in Italian and English speech. *Status Report on Speech Research, Haskins Laboratories* 91, 75–89.

[12] Arvaniti, A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66, 46–63.

[13] Nolan, F., Jeon, H. -S. 2014. Speech rhythm: a metaphor? *Phil. Trans. R. Soc. B: Biol. Sci.* 369.

[14] Mancini, F., Voghera, M. 1994. Lunghezza, tipi di sillabe e accento in italiano. In: T. De Mauro (a cura di), *Come parlano gli italiani*, La Nuova Italia, Firenze, 217–245.

[15] Istituto della Enciclopedia Italiana. 2019. Perché trovo il sostantivo "viola" diviso in tre sillabe? La i e la o non formano un dittongo? *Lingua Italiana,* available at: https://www.treccani.it/magazine/lingua_italiana/domande_e_risposte/grammatica/grammatica_1696.html

[16] Indicazione dell'accento in parole dalla grafia ambigua. In: *La Crusca per voi* 19, available at: https://accademiadellacrusca.it/it/consulenza/indicazione-dellaccento-in-parole-dalla-grafia-ambigua/27

[17] Burani C., Paizi D., Sulpizio S. 2014. Stress assignment in reading Italian: friendship outweighs dominance. *Mem. Cogn.* 42(4), 662–75.

[18] Isetta, M.L. 2015. *Leggere per immagini. Giochi e attività fonologici per l'avviamento alla letto-scrittura*. Erickson.

[19] Spinelli, G., Forti, L., Jared, D. 2020. Learning to assign stress in a second language: The role of second-language vocabulary size and transfer from the native language in second-language readers of Italian. *Bilin.: Lang. Cogn.* 24(1), 124–136.

[20] Primativo, S., Rinaldi, P., O'Brien, S., Paizi, D., Arduino, L. S., Burani, C. 2013. Bilingual vocabulary size and lexical reading in Italian. *Acta Psychol.* 144(3), 554–562.

[21] Istituto della Enciclopedia Italiana. 2012. *La grammatica italiana*, available at: https://www.treccani.it/enciclopedia/accentazione-piana_(La-grammatica-italiana)

[22] Eatough, V., Smith, J.A. 2017. Interpretative Phenomenological Analysis. In: Willig, C., Stainton-Rogers, W. (eds.), *The Sage Handbook of Qualitative Research in Psychology* 2nd ed., Sage, London.

[23] Dewey, J. 1933. *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*. Boston, MA: D.C. Heath & Co Publishers.

[24] Cancer A., Stievano G., Pace G., Colombo A., Antonietti A. 2019. Cognitive processed underlying reading improvement during a rhythm-based intervention. A small-scale investigation of Italian children with dyslexia. *Children* 6(8), 91.

[25] Patel, A.D. 2011. Why would musical training benefit the neural encoding of speech? The OPERA Hypothesis. *Front Psychol* 2.

[26] Good, A. J., Russo, F. A., & Sullivan, J. 2015. The efficacy of singing in foreign-language learning. *Psychol. Music* 43(5), 627–640.

[27] Rathcke, T. & Lin, C. & Falk, S., Dalla Bella, S., 2021. Tapping into linguistic rhythm, *Lab. Phon.* 12, 1–32.

[28] Beccaria, G. L. 1994. *Dizionario di linguistica e di filologia, metrica, retorica.* Torino, Einaudi.

[29] Abercrombie, D. 1967. *Elements of General Phonetics.* Edinburgh University Press.

[30] Emmons, S. nd. The Singer's Dilemma: Tone Versus Diction available at: http://www.sci.brooklyn.cuny.edu/~jones/Shirlee/diction.html (accessed Apr. 26, 2023)

[31] Payne, E. 2021. Comparing and deconstructing speech rhythm across Romance languages. In: Gabriel C., Gess R., Meisenburg T. (eds), *Manual of Romance Phonetics and Phonology,* De Gruyter, 264–298.

The influence of tone experience and native tone and intonation categories on nonnative tone learning

Jessica L. L. Chin, Mark Antoniou

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia
jessica.chin@westernsydney.edu.au, m.antoniou@westernsydney.edu.au

ABSTRACT

For many languages, lexical tone is a linguistic feature which assigns meaning to words through variations in pitch. Learning tone as a new feature can be challenging, though training can improve learning outcomes for native tone and nontone listeners alike. Still, there is little consensus on whether native tone experience bolsters nonnative tone learning. This study examined the influence of the listener's native language on the learning of an artificial tone language. After five training sessions, nontonal Australian English, tonal Mandarin, and tonal Vietnamese listeners showed improvements in tone identification and word learning. Tone language listeners demonstrated greater performance in word learning, and Mandarin listeners outperformed the other groups in tone identification. Further investigation of tone identification patterns showed some tones were easier to perceive for all groups, though performance was also influenced by how closely the tones were mapped onto the listeners' native tone and intonation categories.

Keywords: tone training, tone perception, L1 experience

1. INTRODUCTION

Lexical tone is a linguistic feature that occurs in over half of the world's languages [1]. In a tone language, the meaning of words is distinguished by variations in pitch contours. These pitch contours—or tones—can vary in pitch height, direction, and trajectory. Pitch height is commonly numerically notated, where the lowest pitch value is 1 and the highest is 5 [2]. A level (or static) tone remains consistent in pitch height and direction, e.g., high level 55. A contour (or dynamic) tone can change in direction from onset to offset (e.g., rising 35), sometimes in a concave or convex shape (e.g., low dipping 214) [3]. In some languages, shorter "checked" tones occur only in syllables with a stop coda. Further, tone languages vary in the density of tones in their inventory, as well as in tone type. There is much variation between individual tones in a language, and between different languages.

For nontone language listeners, lexical tone is a novel feature to acquire, as these languages would only utilise pitch at the phrase/sentence level in the form of intonation [4]. Various studies show an advantage in nonnative tone perception for native listeners of a tone language over nontone language listeners [5, 6, 7], though some studies show no significant effect of tone experience [8, 9]. Features of individual tones in the target language, as well as tone presentation order, also strongly affect tone perception in native tone and nontone listeners [7]. In general, multi-session tone *training* has led to greater performance for all listeners at the cessation of training, regardless of tone experience. The influence of tone experience on learning performance is less clear [10]. In one study, listeners with a tonal L1 outperformed nontonal L1 listeners when learning nonnative tones [11]. There is also some evidence of a tone density advantage in one study comparing two listener groups with a different tonal L1 [12]. Sometimes, the tonal L1 learning advantage applies only to one task, yet performance is comparable between tonal and nontonal listeners in another (e.g., tone identification) [13]. Further, there are studies in the literature showing no advantages to nonnative tone learning for tonal L1 listeners over their nontonal counterparts. These outcomes were attributed to L1-influenced cue-weightings (towards pitch height and/or direction), and the relationship between L1 tone/intonation categories to the nonnative tones[13, 14]. In sum, a range of studies show variable effects of native tone language experience on nonnative tone learning. While there are advantages to having prior tone experience, this does not always apply across all perceptual tasks, and sometimes there are no advantages at all.

This study aimed to address the uncertain role of tone experience on nonnative tone learning. We examined the following effects: 1) native tone experience, 2) tone inventory density, and 3) similarities between native tone/intonation categories and nonnative tones. Native listeners of nontonal Australian English (AusE), tonal Mandarin, and tonal Vietnamese completed five sessions of tone training in an artificial language based on Hakka Chinese, a regional dialect spoken in Southern China. The tone inventory of Hakka and the tone/intonation

inventories of the listener languages are summarised in Table 1.

We posit that prior tone experience would bolster nonnative tone learning for the native tone language listeners. If tone density predicts tone learning performance, the Vietnamese listeners would outperform the Mandarin. Conversely, if native tone/intonation categories are better predictors of nonnative tone identification, we would assume that the closer the nonnative tone maps onto the native category, the higher the identification accuracy. For example, Hakka T2 (11) maps closely to the low tone contours in Mandarin and Vietnamese. However, despite all listener groups having at least one category with a falling contour, Hakka's two falling tones (41 and 51) could still be too similar for nonnative listeners to distinguish consistently. The two falling tones begin and end at similar pitch heights and are predicted to be more challenging to learn overall.

Language	Tone/Intonation categories
Hakka [16], [17]	T1 (33), T2 (11), T3 (41), T4 (51) Checked: T5 (55), T6 (41)
English [18]	Flat pitch, rising question, falling statement, high falling exclamation
Mandarin [19]	T1 (55), T2 (35), T3 (214), T4 (51)
Vietnamese [20]	A1 (44), A2 (21), B1 (35), B2 (212), C1 (214) Checked: D1 (35), D2 (212)

Table 1: Tone and intonation categories of Hakka, English, Mandarin, and Vietnamese.

2. METHOD

2.1. Participants

Participants were native listeners of Australian English (n = 25; M_{age} = 24.46, SD = 8.2), Mandarin (n = 23; M_{age} = 26.11, SD = 5.4), and Vietnamese (n = 25; M_{age} = 24.53, SD = 6.9). None reported any hearing or neurological impairments, and all participants (except one Vietnamese listener tested remotely) passed an air conduction audiogram at 25 dB HL at 500, 1000, 2000 and 4000 Hz. Five English participants spoke another nontone language at home, and one participant studied Japanese in late adolescence. All Mandarin participants reported Mandarin as their dominant language. None spoke Hakka Chinese, though 14 reported knowledge of another Chinese dialect, and these contain their own lexical tone contrasts (n = 5 Cantonese, n = 3 Shanghainese, and six other dialects where n = 1). All but five Vietnamese participants spoke the Southern variant of the language; three participants spoke the Central variant and two spoke the Northern variant.

2.2. Stimuli

There were three experiment tasks: tone identification, tone word learning, and generalisation. The tokens in the tone identification task consisted of monophthongs /a/, /e/, /i/, /o/, /u/, and VC (checked) syllables /a̰k/, /ḭp/, /ḛt/. The 16 tokens in tone word learning and generalisation consisted of CVC nonwords [fon], [leŋ], [nun], and checked syllable nonwords [wa̰p] and [mḭp].

Auditory stimuli were produced by three female native speakers of Australian English, one for each task. The productions were recorded in a sound-attenuated booth and sampled at 44.1 kHz using a Shure SM10A cardioid microphone connected to a Roland Duo-Capture EX audio interface. All tokens were produced with a level tone, and were later superimposed with Hakka Chinese tones using the pitch-synchronous overlap and add function in Praat [21]. The Hakka tone onset and offset values were drawn from past analyses of three female native Hakka speakers [17].

Visual representations of Hakka's four tones and two checked tones were created for tone identification. In each image, an arrow depicted the height of the tone at onset and offset across five height levels. For tone word learning and generalisation, each of the 16 nonwords corresponded to a monochromatic image of a high-frequency word. These images were taken from a stimuli set used in previous research [21, 22]. Examples of the visual stimuli are provided in Fig. 1.

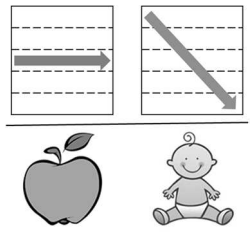


Figure 1: *Top row*: Visual representation of tones 1 (33) and 4 (51) in tone identification. *Bottom row*: Images corresponding to the nonwords [fon51] (apple) and [wa̰p55] (baby) in tone word learning.

2.3. Procedure

Participants completed five sessions of training across separate days. Tasks were presented in a quiet testing space on a laptop PC running E-Prime 3.0. Stimuli were presented at 72 dB SPL over Sennheiser HD280 Pro headphones. In session 1, participants completed a demographics questionnaire, a tone identification pre-test, and tone word learning. In sessions 2 to 4, participants completed tone word

learning only. In session 5, participants completed their final tone word learning session, a generalisation test, and a tone identification post-test. Three weeks after the cessation of training, participants were invited to complete a tone identification retention test. A total of 86% of participants returned for the retention test (68% AusE, 100% Mandarin, and 92% Vietnamese).

2.3.1. Tone identification

The tone identification task consisted of a familiarisation, practice, and test phase. Checked and unchecked tones were presented in separate blocks, and block presentation order was randomised for each participant. The tokens /a/ and /a̰k/ were used in familiarisation and practice, and were not included in the test phase. In familiarisation, participants were exposed to 3 repetitions of /a/ and /a̰k/. During practice, participants were presented with a sound, then instructed to select the tone they think they heard from the two visual response options on the screen (see Fig. 1 for examples). Feedback was provided after each response. In the test phase, participants were presented with the remaining stimuli in a similar format to the practice phase, except without feedback. Stimuli were presented at an interstimulus interval of 3 s; failure to respond within 3 s led to a missed response. A total of 104 trials were presented.

2.3.2. Tone word learning and generalisation

Tone word learning comprised of an exposure and test phase. During exposure, participants were presented with 4 repetitions of a minimal pair, with a unique image accompanying each sound. Participants were then presented with a sound and the two images as response options. Feedback was provided with every response. During the test phase, participants were presented with one sound at a time without feedback. All 16 images were provided as response options. A total of 96 test trials were presented.

The generalisation test was completed during the fifth session, directly after tone word learning. It was identical to the test phase of the tone word learning task, except with stimuli produced by a new talker.

3. RESULTS

Data were analysed using generalised linear mixed-effects models using the *glmer* function from the *lme4* package [24] in R [24, 25]. Post-hoc analyses were conducted with *emmeans* [27].

3.1. Tone identification

Mean identification accuracy is shown in Fig. 2. The model included the fixed factors Language (AusE, Mandarin, Vietnamese; Helmert coded) and Session (pre-test, post-test, retention test; repeated contrast coded), as well as random intercepts by participant and item. The Mandarin listeners outperformed both the AusE (β = 0.82, SE = 0.16, p < .001) and Vietnamese listeners (β = 0.81, SE = 0.19, p < .001). For all groups, performance improved between pre- and post-test (β = 0.22, SE = 0.02, p < .001), as well as between post-test and retention test (β = 0.14, SE = 0.03, p < .001).

We also examined tone identification patterns across listeners at post-test to determine whether identification was influenced by language background, individual tone characteristics, and/or the alternative response option provided in the task (see Table 2). Overall, the level tones were easier to identify than the falling tones, especially when the alternative response option was the other level tone. As predicted, the falling tones were most difficult for all listeners; in fact, the lowest identification scores (<75%) for the highest-performing Mandarin group were for these two tones.

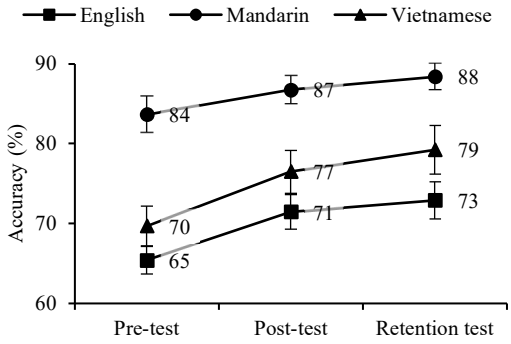


Figure 2: Mean tone identification accuracy by group at pre-, post- and retention test. Error bars depict SEM.

3.2. Tone word learning and generalisation

Fig. 3 shows mean accuracy for both word learning and generalisation. Word learning was analysed with a model that included the fixed factors Language (coded as above) and Session (1–5; repeated contrast coded), as well as random intercepts for participant and item. The model showed that both the Mandarin (β = 0.60, SE = 0.21, p = .012) and Vietnamese groups (β = 0.59, SE = 0.21, p = .013) outperformed their AusE counterparts. There was significant improvement in performance for all groups in subsequent sessions (p < .001). However, there was a

	Stimulus tone identification accuracy per alternative response option																	
	T1 (33)				T2 (11)				T3 (41)				T4 (51)				T5	T6
Group \ Alternative response option	T2 (11)	T3 (41)	T4 (51)	T1 total	T1 (33)	T3 (41)	T4 (51)	T2 total	T1 (33)	T2 (11)	T4 (51)	T3 total	T1 (33)	T2 (11)	T3 (41)	T4 total	T6 (41)	T5 (55)
English	79%	66%	68%	**71%**	83%	75%	78%	**78%**	56%	68%	80%	**68%**	62%	82%	62%	**69%**	80%	76%
Mandarin	92%	97%	97%	**95%**	94%	79%	90%	**88%**	92%	81%	71%	**81%**	93%	94%	74%	**87%**	95%	91%
Vietnamese	89%	83%	79%	**84%**	84%	73%	82%	**80%**	69%	68%	78%	**71%**	72%	83%	72%	**76%**	82%	79%

Table 2: Tone identification (post-test) accuracy of stimulus tones when presented with different tones as alternative response options (e.g., English listeners identified T1 with 79% accuracy when T2 was the alternative response option). Overall tone identification accuracy of the stimulus tone is also provided.

significant interaction between the tonal groups and their performance from session 1 to 2 ($\beta = 0.10$, $SE = 0.05$, $p = .041$), suggesting that the Mandarin group improved more from session 1 to 2 than the Vietnamese group. The model used to analyse generalisation included the fixed factor Language (coded as above), and random intercepts for participant and item. Results showed that the Mandarin group outperformed the Vietnamese ($\beta = 0.53$, $SE = 0.24$, $p = .025$) and AusE groups ($\beta = 1.06$, $SE = 0.24$, $p < .001$).

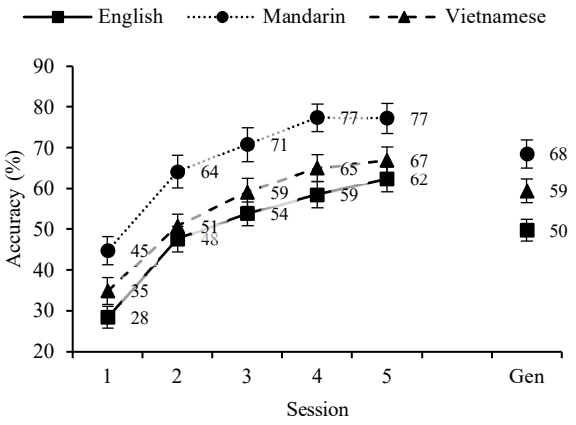


Figure 3: Mean tone word learning accuracy for all groups across all sessions and generalisation.

4. DISCUSSION

In the current study, listeners of varying language backgrounds completed multi-session tone training in an artificial language. All groups, regardless of level of tone experience, improved in tone identification and tone word learning at session 5. Further, all groups were able to retain these learning gains three weeks after the cessation of training. Native tone experience provides advantages to tone word learning, but not necessarily for tone identification: The Mandarin group outperformed both the AusE and Vietnamese groups in this task. Tone density also did not facilitate tone learning. Tone word learning requires listeners to engage in higher-level processes to link nonnative sounds to separate lexical items, much like learning words in a tone language. It would make sense, then, that tone language natives would be able to transfer this knowledge to a nonnative tone language. This finding falls in line with previous training studies demonstrating a tone language advantage [10, 11]. However, it is still uncertain whether native tone experience facilitates tone identification, or other factors bolstered the Mandarin listeners' performance. Studies have shown comparable performance in tone identification between tone and nontone speakers [12, 14], and have attributed this to factors such as cue weighting and L1 tone/intonation categories.

The tone identification patterns revealed a few tone-specific trends. Across listener groups, T2 (11) may be easier to identify due to its lower pitch and distance from the other Hakka tones at onset. Similarly, higher accuracy is observed for the stimulus tone T4 (51) when the alternative tone is T2 (11). Language-specific tone identification patterns were also present. The tone language listeners showed a smaller, but reliable, advantage in identifying T1 (33) than T2 (11), possibly due to T1's closeness to their native level tones relative to T2 (55 for Mandarin, 44 for Vietnamese). Despite the Mandarin group's overall higher identification accuracy, they misidentified the stimulus tone T3 (41) as T4 (51) more frequently than the other groups—a possible reason for this is interference from their own L1 falling tone (51). T3 also does not map closely enough to any categories in English (and neither falling tone maps closely to any Vietnamese category), which may explain the poorer identification accuracy for this tone.

In sum, while tone training can improve overall tone identification and word learning performance, tone- and language-specific identification patterns arise. There appear to be tones which are identified by all listeners with greater or poorer accuracy based on characteristics such as pitch height and direction. Further, the listener's native tone and/or intonation categories can influence nonnative tone identification, either to one's benefit or hindrance.

7. REFERENCES

[1] Yip, M. 2002. *Tone*. Cambridge University Press.

[2] Chao, Y. R. 1930. A system of tone letters. *Maître Phon.* 45, 24–27.

[3] Abramson, A. S. 1978. Static and dynamic acoustic cues in distinctive tones. *Lang. Speech* 21, 319–325.

[4] Ladd, D. R. 2008. *Intonational Phonology*. Cambridge University Press.

[5] Burnham, D. *et al.* 2015. Universality and language-specific experience in the perception of lexical tone and pitch. *Appl. Psycholinguist.* 36, 1459–1491.

[6] Schaefer, V., Darcy, I. 2014. Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Lab. Phonol.* 5, 489–522.

[7] Liu, L. *et al.* 2022. The tone atlas of perceptual discriminability and perceptual distance: Four tone languages and five language groups. *Brain Lang.* 229, 105106.

[8] Bent, T., Bradlow, A. R., Wright, B. A. 2006. The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 97–103.

[9] Tsukada, K., Xu, H. L., Rattanasone, N. X. 2015. The perception of Mandarin lexical tones by listeners from different linguistic backgrounds. *Chin. Second Lang. Res.* 4, 141–161.

[10] Antoniou, M., Chin, J. L. L. 2018. What can lexical tone training studies in adults tell us about tone processing in children? *Front. Psychol.* 9, 1–12.

[11] Wayland, R., Li, B. 2008. Effects of two training procedures in cross-language perception of tones. *J. Phon.* 36, 250–267.

[12] Tong, X., Tang, Y. C. 2016. Modulation of musical experience and prosodic complexity on lexical pitch learning. *Speech Prosody*, 217–221.

[13] Cooper, A., Wang, Y. 2012. The influence of linguistic and musical experience on Cantonese word learning. *J. Acoust. Soc. Am.* 131, 4756–4769.

[14] Wang, X. 2013. Perception of Mandarin tones: The effect of L1 background and training. *Mod. Lang. J.* 97, 144–160.

[15] Francis, A. L., Ciocca, V., Ma, L., Fenn, K. 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294.

[16] Hashimoto, M. J. 2010. *The Hakka dialect: A linguistic study of its phonology, syntax and lexicon*. Cambridge University Press.

[17] Cheung, Y. M. 2011. Vowels and tones in Meixian Hakka: An acoustic and perceptual study. PhD dissertation, City University of Hong Kong, Hong Kong,

[18] Reid, A. *et al.* 2015. Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Atten. Percept. Psychophys.* 77, 571–591.

[19] Chen, C.-C. 2011. *Monographs on linguistics analysis: A synchronic phonology of Mandarin Chinese*. Walter de Gruyter.

[20] Brunelle, M. 2009. Tone perception in Northern and Southern Vietnamese. *J. Phon.* 37, 79–96.

[21] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer.

[22] Antoniou, M., Wong, P. C. M. 2016. Varying irrelevant phonetic features hinders learning of the feature being trained. *J. Acoust. Soc. Am.* 139, 271–278.

[23] Wong, P. C. M., Perrachione, T. K. 2007. Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* 28, 565–585.

[24] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67, 1–48.

[25] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[26] RStudio Team 2020. RStudio: Integrated Development for R. RStudio, PBC., Boston, MA.

[27] Lenth, R. V. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means.

TONE AS A FACTOR INFLUENCING THE DYNAMICS OF DIPHTHONG REALIZATIONS IN STANDARD MANDARIN

Chenyu Li[1], Jalal Al-Tamimi[1], Yaru Wu[2,3,4]

[1]Université Paris Cité, LLF, CNRS; [2]CRISCO, Université de Caen Normandie; [3]LPP, CNRS, Sorbonne Nouvelle; [4]LISN, CNRS, Université Paris-Saclay
chenyu.li@etu.u-paris.fr; jalal.al-tamimi@u-paris.fr; yaru.wu@unicaen.fr

ABSTRACT

This paper investigates the impact of the tonal factor on the realization of diphthongs in Standard Mandarin. We employed a set of Generalized Additive Mixed Models (GAMMs) to respectively test whether and how tones and fundamental frequency ($f0$) influence the realization of the diphthong /ai/ in a journalistic speech corpus of Standard Mandarin. Results show, in general, the realizations of /ai/ to significantly differ with respect to tones, due to different heights and contours of $f0$. A high tone results in a more closed and more front /ai/ and a low tone results in a more open and more back realization; a rising tone strengthens the dynamic features of /ai/ while a falling tone weakens such features and leads to monophthongization. The results suggest that the relation between $f0$ and vowel realization, found in monophthongs, is equally applicable to diphthongs: $f0$ is negatively correlated with F1 and positively with F2.

Keywords: Standard Mandarin, diphthong, tone, $f0$, Generalized Additive Mixed Models

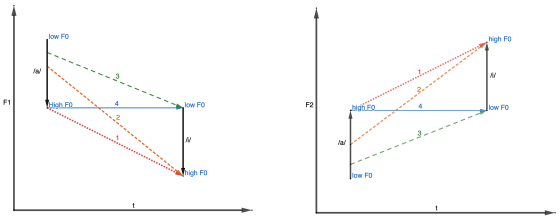
1. INTRODUCTION

Phonological segments and suprasegmental features are two major areas of research in phonology and phonetics, but they are often studied separately. The tone, as a suprasegmental concept, is a major component of Mandarin phonology [1] that was widely studied in monosyllabic sequences [2] or in continuous speech [3]. In Standard Mandarin, there are four lexical tones (with Chao digits): tone 1 (high tone - 55), tone 2 (rising tone - 35), tone 3 (low tone - 21 in natural speech) and tone 4 (falling tone - 51) [1]. In many languages, like Mandarin, there is a dynamic vowel concept, the diphthong, which is considered as a continuously varying single segment [4, 5], or a dynamic vowel sequence which has two targets and a transition between them [6].

The interaction between tones (contours of $f0$) and single vowels has been extensively studied, with accumulating evidence of this interaction in the forms of intrinsic $f0$ and the tonal effect on vocalic realization. A large body of research have confirmed that the intrinsic $f0$ is attested in many languages, including tonal, such as Mandarin [7], and non-tonal languages [8, 9]. Results suggest that a more closed vowel will have a higher intrinsic $f0$. Regarding the effect of $f0$ on vowel realization, different studies have evaluated the impact of different $f0$ levels on vowel realization from an acoustic [7, 10] and an articulatory view point [11, 12, 13]. In general, a higher $f0$ will cause the vowel to be realized as more closed and more front. However, until now, there does not exist any clear evidence for the interaction between the $f0$ and diphthong realization.

Although there is evidence from perception [14] suggesting that this interaction could also apply to the case of diphthongs, there has been no research showing direct evidence of it. As a matter of fact, most of the current literature on Mandarin diphthongs is at the descriptive [6] or the historical/dialectal (e.g., [15]) level and does not take $f0$ into account.

The present study aims to explore how tones influence the realizations of diphthongs in Standard Mandarin and to assess whether the effect of the $f0$ on vowel realization occurs in diphthongs. We hypothesize that for /ai/, the vowel realizations will be affected by $f0$; more concretely, a higher $f0$ will make the vowel produced as more closed and more front, whereas a lower $f0$ will make the vowel produced as more open and more back. Under this hypothesis, we can make various predictions for how /ai/ will be realized with different tones.



(a) F1 **(b)** F2

Figure 1: Predictions following hypothesis

According to Fig. 1, the F1 curve of /ai/ will be

higher with a low falling tone (tone 3) and lower with a high tone (tone 1), flatter with a falling tone (tone 4) and steeper with a rising tone (tone 2). The F2 curve will have an opposite pattern to that of F1.

2. METHODS

2.1. Corpus and data processing

The data came from a journalistic speech corpus from the LDC [16], while only a 2h long data set was provided by the LISN lab to this study for processing and further analyses. The data were automatically segmented, aligned and transcribed with Chinese characters and their phonetic transcriptions with lexical tones at the word and the phoneme levels, using a speech transcription system from LISN in forced alignment mode [17]. Only lexical tones (1 to 4) in Standard Mandarin are included in the lexicon. Four falling diphthongs, i.e., a diphthong whose first component is syllabic [6]: /ai/, /au/, /əi/ and /əu/ were processed [18]. We present the results of /ai/ here due to it being the most common diphthong in the database.

Segmental boundaries between the diphthong /ai/, the previous and following segments were manually verified using *Praat* (version 6.1.54) [19], with the start and end boundaries being chosen using the positions of the *onset* and *offset* of F2 of /ai/, following [20]. We considered the duration of the tone as being that of the *rime*, i.e., from the *onset* of the vowel to *offset* of the syllable [21]: both diphthong and tone had the same duration. This yielded a data set of 215 occurrences: 112 for 8 female speakers (9 for tone 1, 18 for tone 2, 15 for tone 3 and 70 for tone 4) and 103 for 15 male speakers (4 for tone 1, 26 for tone 2, 17 for tone 3 and 56 for tone 4). For each occurrence, we have the following information: the tone; the position of /ai/ in the word, i.e., final or non-final position and monosyllabic; the previous and following segments; speaker's gender and the anonymized identity.

The acoustic data were extracted for each occurrence, with the fundamental frequency ($f0$), and the formant frequencies for F1 and F2. Formant frequencies were automatically obtained using the *Burg* method (window length = 25ms, pre-emphasis from 30 Hz, maximum frequency = 5500 Hz for female and 5000 Hz for male, and a maximum of 5 formants). We then obtained $f0$ measurements using the two-passes method, with auto-correlation (following [22], implemented in [23]); $f0$ estimations are speaker-dependent based on their range that prevents errors in extraction. For each occurrence, we obtained 11 time normalized intervals, at 10% interval for formant and $f0$ frequencies. This yielded a total of 2301 data points (1105 for male speakers, 1196 for female speakers).

2.2. Modelling

Since the diphthong /ai/ is a dynamic vowel sequence, we used Generalized Additive Mixed Models (GAMMs) [24] to capture its dynamic pattern. GAMMs are a statistical technique that models the non-linear relation between predictors and an outcome varying in the time domain. Modelling with GAMMs was done using the packages *mgcv* [24] (version 1.8-33, for modelling) and *itsadug* [25] (version 2.4, for visualization) in R computing language (version 4.2.2) [26]. For model specification, we followed recommendations from [27, 28, 29].

For our specific dataset, we performed two sets of modelling. First, to evaluate the interaction between the diphthong realization using F1 and F2 dimensions and the tonal unit, our predictor was the `tone` (categorical - four levels) and the outcome either F1 or F2 frequencies. We used `time` as a continuous predictor (11 normalized intervals), represented via a *smooth* as a non-linear variable, to track the dynamic pattern during the diphthong realization. The other variables, i.e., the speaker ID, the consonant at syllable onset and the position in the word, were considered as factor smooths modelled as random effects. Different genders were modelled separately to assess differences between the two genders and to reduce computational power.

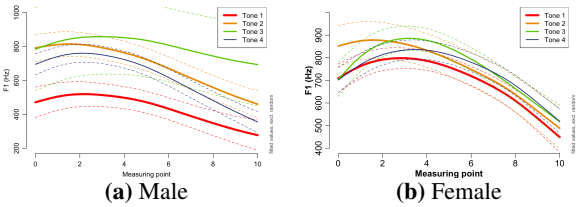
We used an *ordered* categorical predictor for `tone` to reduce the Type I error and increase power in our model [29]. We used a *by* interaction term to adjust for the variable levels of tones for our smooth terms. Then, we verified the auto-correlation levels of our model and obtained an Auto-Regressive GAMMs. After fitting the model, we performed a `gam.check` and the data suffered from a "long tail" distribution; this was corrected using the `"scat"` family [27, 29, 30]. The comparison of the models with `gam.check` showed that the modified ones fitted the data better.

The second model has roughly the same structure, except that the categorical predictor `tone` is replaced by two continuous ones: `f0` and `duration`, to assess the correlation between tone and $f0$/duration, because tone, as a phonological unit, is represented phonetically by the $f0$ contour over time [1].

3. RESULTS

3.1. First model

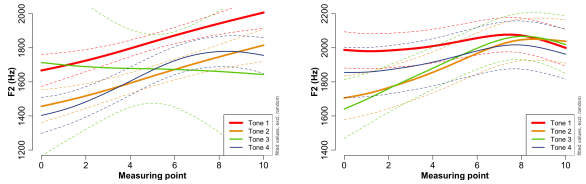
The purpose of the first model is to evaluate any possible influences of tone on the diphthong realization. Figs. 2 and 3 illustrate the predictions obtained from the model for F1 and F2 curves, differentiated by the four tones.



(a) Male **(b)** Female

Figure 2: F1 curves with 4 tones

All F1 contours are falling from interval 2 to 8. In tone 1 (high level), the curve is generally lower than in other tones, signalling a more closed realization of the diphthong /ai/. The curve with tone 3 (low falling), is higher than in other tones, which indicates that /ai/ would be realized as the most open. The curves with tone 1 and 3 have a relatively flat pattern, which indicates partial monophthongization. The curve with tone 2 (rising) starts relatively high (800 Hz) and stops relatively low (600 Hz), which indicates minimal monophthongization, compared with the other tones. The curve with tone 4 (falling) of the female speakers (Fig. 2b) starts relatively low (700 Hz) and ends relatively high (700 Hz), showing a flat pattern which is related to the monophthongization. Using the plot_diff function in the *itsadug* package, we report on the regions of statistical differences (95% of CI). We found that the curves of male speakers are significantly different in the 1-2, 1-3, 1-4, 2-4 (first and final parts), 3-4 (final part) tonal pairs; 1-2 (first part), 1-3, 1-4, 2-3 (starting part), 2-4 (first part) pairs for the female ones are significantly different.



(a) Male **(b)** Female

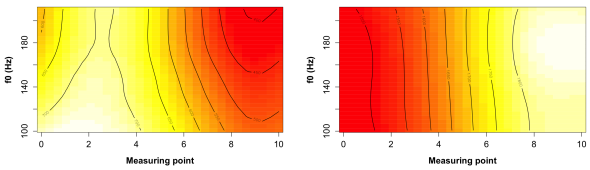
Figure 3: F2 curves with 4 tones

As for the F2 (Fig. 3), the results point to rising contours over time, except for tone 3 in male speakers (Fig. 3a). The curve with tone 1 is generally higher than in other tones, i.e., /ai/ with tone 1 would be realized as most front. The curve with tone 3 of female speakers is generally lower than those with the other tones. For the male speakers, although the curve with tone 3 does not follow the rising trend, it is lower than those with the other tones in the second part (intervals 6 - 10). This indicates that /ai/ would be realized as most back with tone 3. We observed a pattern with a steeper slop in the curve with tone 2, indicating a minimal monophthongization. Based on the predictions of our hypothesis (Fig. 1), tone 4 (falling), is predicted to monophthongize the diphthong. Fig. 3 reveals a flatter curve with tone 4, on the full contour for female speakers and between intervals 5-10 for male speakers. The significance test (95% of CI) show that the curves of pairs 1-2, 1-4 for male speakers, and 1-2, 1-3, 1-4 (first part), 2-4 (starting part), 3-4 (starting part) pairs for female speakers, are significantly different from each other.

The predictions of the first model generally confirm the hypothesis: the tonal effect on vowel realization is observed on the diphthong /ai/ in this study. More concretely, with a high tone (tone 1), /ai/ tends to be realized as more closed and more front; with a low tone (tone 3), it is more open and more back; with a rising tone (ton 2), it tends to have a typical diphthongized realization; and with a falling tone (tone 4), the diphthong tends to be monophthongized (but see *Summary and discussion* for specific issues related to tone 4).

3.2. Second model

In the second model, we consider the combination of $f0$ and duration, to more accurately model the impact of tones on the realization of the diphthong /ai/. To visualize the result of the second model, we selected three intervals at the quartiles, i.e., 25%, 50% and 75% of the duration. We present the summed effects 3D plots using the function fvisgam. In the figures, the dependent variable, i.e., F1 or F2 for female or male speakers, is determined by two different continuous predictors on two axes, i.e., the normalized time on the x axis and the $f0$ on the y axis, with F1 or F2 changes on the z axis, whereby the lighter the colour, the higher the frequency values of F1 or F2; and vice versa.

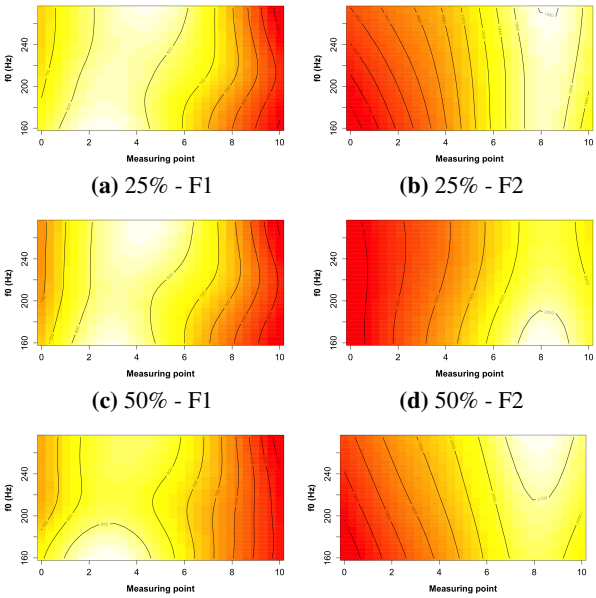


(a) F1 **(b)** F2

Figure 4: F1/F2 with different $f0$ (Male, 50% of duration)

The Fig. 4 demonstrates the model's predictions for male speakers at 50% of the duration due to minimal differences across all three portions (at 25-50-75%); this specific portion shows that the effect of $f0$ on F1 or F2 is independent from the effect of duration on F1 or F2. Overall, F1 falls during the realization of the diphthong, whereas F2 rises. Meanwhile, the contour lines (which show the same value of F1 or F2) and colors show that

when $f0$ rises, F1 falls and F2 rises (especially in the transition part).



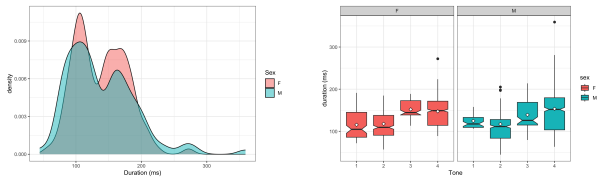
(a) 25% - F1 **(b)** 25% - F2

(c) 50% - F1 **(d)** 50% - F2

(e) 75% - F1 **(f)** 75% - F2

Figure 5: F1/F2 with different $f0$ (Female)

The patterns of the female speakers are more complex. According to Fig. 5, we found that the predictions at 50% of the duration (Figs. 5c and 5d) suggest that F1 has a positive correlation with $f0$ and F2 has a negative correlation (from interval 4); results that are contra to the predictions of our hypothesis, which is likely due to the fact that the reliability of model predictions in some specific cases is influenced by the amount of data. Fig. 6a shows that the distribution of duration of female speakers is *bi-modal* than that of male speakers, which makes the predictions obtained at 50% of duration less reliable and a more dynamic account is needed.



(a) Duration density of genders **(b)** Duration of each tone

Figure 6: Duration of /ai/

4. SUMMARY AND DISCUSSION

In this paper, we were interested in modelling how the diphthong /ai/ in standard Mandarin is realized and assessed any possible relationships with tone, on the one hand, and with the combination of $f0$ and duration as multi-dimensional predictors, on the other. The first model predicts well how the /ai/ is realized in different tones. Overall, if it is associated with a high tone (tone 1), /ai/ is produced as more closed and more front. If it is associated with a low tone (tone 3), the diphthong is produced as more open and more back. /ai/ would be realized as more diphthongized when it is associated with a rising tone (tone 2). As for a falling tone (tone 4), the expected monophthongized pattern has only been confirmed in the result of F1 and F2 curve of female speakers. The second model allows for more refined modelling to capture the effects of $f0$ and of the duration on F1 or F2 and its change over time. In general, $f0$ is negatively correlated with F1, and positively with F2. The models and their predictions in this paper show a good fit to the data and allowed to evaluate how diphthongs are realized in Standard Mandarin. They also allowed to verify the hypotheses with respect to effects of tones and $f0$, independently of each other.

The curves with tone 4, according to the hypothesis, should be more flat, but from the prediction obtained from the model, they show an obvious dynamic pattern. In Figs. 2a and 3a, the F1 and F2 curves with tone 4 for male speakers show a "diphthongized" dynamic feature similar to that with tone 2, which contradicts our hypothesis. From Fig. 6b, the duration of /ai/ varies with different tones: in tones 4 for female and male speakers and in tone 3 for female speakers, the duration is the longest. This allows for more time to complete the vowel realization from the *onset* to the *offset*. In the visualizations (Figs. 2 and 3), the "compression" effect of normalized time will cause curves with longer absolute duration to appear more dynamic. In this case, the duration effect is possibly more important than the $f0$ effect. The results of the model 2 and of the significance test in model 1 also highlighted the impact of gender on F1 or F2. The effects of male speakers are more reflected on F1 (aperture) level, while those for female speakers are more reflected on F2 (backness) level.

The present study has revealed that in addition to $f0$, duration and gender can also affect diphthong realizations. Further research may focus on two aspects. Firstly, through comparative studies of different dialects in Mandarin, we aim to investigate the universality of the tonal influence on diphthongs and provide explanations for various phenomena of monophthongization in different dialects. Secondly, we will attempt to incorporate the current study into Articulatory Phonology framework and explain its relationship with the intrinsic $f0$ of vowels.

5. ACKNOWLEDGEMENTS

This work was supported by the Laboratoire de Linguistique Formelle, the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083), contributing to the IdEx Université Paris Cité - ANR-18-IDEX-0001. The data set was supported by LISN, as part of the IdEx Emergence project *Son-Discours*. Thanks to the *Chinese Student Council* for funding. Special thanks to Prof. Ioana Chitoran.

6. REFERENCES

[1] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.

[2] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[3] Y. Xu, F. Liu *et al.*, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 125, 2006.

[4] B. Malmberg, *Structural linguistics and human communication: An introduction into the mechanism of language and the methodology of linguistics*. Springer Science & Business Media, 2012, vol. 2.

[5] P. Delattre, "Comparing the vocalic features of English, German, Spanish and French," 1964.

[6] H. Ren, "On the acoustic structure of diphthongal syllables," Ph.D. dissertation, University of California, Los Angeles, 1986.

[7] P. Wang, *A Statistical Study on the Tones and Vowels of Beijing Dialect*. PhD Thesis, Nankai University, 2007.

[8] D. H. Whalen and A. G. Levitt, "The universality of intrinsic f0 of vowels," *Journal of phonetics*, vol. 23, no. 3, pp. 349–366, 1995.

[9] J.-M. Hombert, "Consonant types, vowel height and tone in Yoruba," *UCLA Working Papers in Phonetics*, vol. 33, pp. 40–54, 1976.

[10] D. Erickson, R. Iwata, M. Endo, and A. Fujino, "Effect of tone height on jaw and tongue articulation in Mandarin Chinese," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[11] F. Hu, "Tonal effect on vowel articulation in a tone language," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[12] P. Hoole and F. Hu, "Tone-vowel interaction in standard Chinese," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[13] J. A. Shaw, W.-r. Chen, M. I. Proctor, and D. Derrick, "Influences of tone on vowel articulation in Mandarin Chinese," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. S1566–S1574, 2016.

[14] O. Niebuhr, "Intrinsic pitch in opening and closing diphthongs of German," in *Speech Prosody 2004, International Conference*, 2004.

[15] W. Zhang, *Evolution and competition: changes in the phonological structure of the Guanzhong dialect*. Peaple's Press of Shaanxi Province, 2002.

[16] A. Morris, B. Antonishek, X. Li, and S. Strassel, *HAVIC MED Progress Test–Videos, Metadata and Annotation*. Linguistic Data Consortium, University of Pennsylvania, 2019.

[17] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[18] C. Li, *The role of factors influencing the realization of diphthongs in Standard Mandarin: dynamic modeling with GAMMs*. Master Thesis, niversité Paris Cité, 2022.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[20] Ö. Ünal-Logacev, S. Fuchs, and L. Lancia, "A multimodal approach to the voicing contrast in Turkish: Evidence from simultaneous measures of acoustics, intraoral pressure and tongue palatal contacts," *Journal of Phonetics*, vol. 71, pp. 395–409, 2018.

[21] J. M. Howie and J. M. Howie, *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press, 1976, vol. 18.

[22] J. Al-Tamimi and G. Khattab, "Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops," *Journal of Phonetics*, vol. 71, pp. 306–325, 2018.

[23] J. Al-Tamimi. (2022) Praat *f*0 Accurate Estimation. https://github.com/JalalAl-Tamimi/Praat-f0-Accurate-Estimation.

[24] S. N. Wood, *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006.

[25] J. Van Rij, M. Wieling, R. H. Baayen, and D. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2015.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[27] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[28] M. Sóskuthy, "Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction," *arXiv preprint arXiv:1703.05339*, 2017.

[29] M. Soskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, p. 101017, 2021.

[30] Y.-Y. Chuang, J. Fon, I. Papakyritsis, and H. Baayen, "Analyzing phonetic data with generalized additive mixed models," in *Manual of clinical phonetics*. Routledge, 2021, pp. 108–138.

THE ROLE OF PRIOR KNOWLEDGE IN SECOND-LANGAUGE LEARNERS' OVERNIGHT CONSOLIDATION OF CANTONESE TONES

Zhen Qin[1], Rui Jin[2], Ruofan Wu[3]

[13]The Hong Kong University of Science and Technology, [2] The University of Florida
hmzqin@ust.hk, jinr@ufl.edu, hmruofanwu@ust.hk

ABSTRACT

This study examines the role of prior linguistic knowledge in the sleep-mediated memory consolidation of lexical ones. Memory consolidation is beneficial in learning new sounds in a second language. And new information consistent with the existing knowledge is consolidated better than that which is not during sleep. What remains unclear from the research is whether prior knowledge from a native language influences the overnight consolidation of lexical tones. The study adopts an overnight design, using two contrasts of Cantonese contour and level tones, for two perceptual learning experiments conducted separately on Mandarin and English-speaking learners. The results indicate that Mandarin-speaking participants successfully consolidated novel words contrasting in contour tones (but not in level tones) thanks to their prior knowledge of contour tones. Without prior knowledge of tones, English-speaking L2 learners did not show a difference in consolidating tones. The findings suggest a beneficial role of prior knowledge in overnight consolidation.

Keywords: Lexical tones, memory consolidation, prior knowledge, perceptual learning, Cantonese.

1. INTRODUCTION

While adult learners find it challenging to learn novel sounds from second languages (L2), sleep is well known for facilitating L2 learning through *memory consolidation* [1]. By focusing on *lexical tones* (pitch variation that signals word identity), this study investigated sleep-mediated memory consolidation (i.e., overnight consolidation). It examined the role of *prior knowledge* in the consolidation of novel Cantonese tonal contrasts by Mandarin-speaking participants (and English-speaking controls).

Accumulating evidence has demonstrated that overnight sleep, especially right after training, offers the advantage of transforming newly learned linguistic knowledge into stabilized long-term memory [2]–[4]. What remains less clear is why some linguistic information is often prioritized for consolidation [5], [6]. Previous studies proposed an effect of prior knowledge (i.e., schema) on memory consolidation, which suggests that new information consistent with the existing knowledge is encoded more strongly and then consolidated more rapidly than that is not [7]–[9]. The beneficial effect of prior *linguistic* knowledge, that is, knowledge of familiar sounds/words in the native language (L1), on memory consolidation was found for L2 phonological and lexical learning [10], [11].

Different from schema studies in which prior knowledge was manipulated through an encoding or training phase [7], this paper aims to examine the effect of prior knowledge from L1 on the overnight consolidation of lexical tones. Lexical tones are used to contrast segmentally identical words (/si 55/ 'silk' (Tone 1 (T1)), /si 25/ 'history' (T2), /si 33/ 'to try' (T3), /si 21/ 'time' (T4), /si 23/ 'city' (T5), and /si 22/ 'matter' (T6)), for instance, in Cantonese [12]. Recent studies investigated sleep-mediated consolidation in Cantonese tone learning by Mandarin-speaking participants. The findings suggested that overnight sleep after evening training facilitated the stabilization of newly learned tonal representations through memory consolidation [13], [14].

It is well established that an individual's prior knowledge of lexical tones impacts how they learn lexical tones. Depending on L1 experience (e.g., tone or non-tone language), L2 learners may learn certain tones well but learn others poorly [15]–[17]. For instance, prior work on tone perception and learning found that Mandarin-speaking participants showed a different pattern in learning Cantonese *contour and level tones* [18], [19]. Mandarin listeners encode pitch lexically in their contour-tone system (/pa 55/ 'eight' T1, /pa 35/ 'to pull out' T2, /pa 214/ 'to hold' T3, /pa 51/ 'father' T4). They found it easier learning to differentiate contour-level tones (a low-rising tone, /23/ vs. a low-level tone, /33/) than level-level tones (a mid-level tone, /33/ vs. a low-level tone, /22/) as Mandarin does not contrast in level tones. English-speaking participants did not consistently use pitch to encode lexical meanings in their L1 and did not show such a difference in their learning of tonal contrasts. However, it remains unclear whether learners' prior linguistic knowledge, for instance, Mandarin listeners' familiarity with contour tones, will benefit its consolidation over level tones after an overnight interval. Cantonese contour and level tones are a good case for testing the effect of prior knowledge on tonal consolidation.

2. EXPERIMENT 1

2.1. Participants

Forty-six Mandarin-speaking participants (age range: 19-33 years; 11 males) were recruited from a university in Hong Kong. The participants were pre-screened to ensure: 1) they were native Mandarin speakers who do not speak Southern Chinese dialects which may contrast in level tones; 2) they had minimal exposure to Cantonese with their length of residence in Hong Kong less than 1 year; (3) they had no more than five years of professional music lessons. None of the participants reported a history of hearing impairment or neurological disorders. Participants maintained a regular sleep pattern with at least 6 hours of sleep time in the past month. The participants provided written informed consent and received monetary compensation for their participation.

2.2. Materials

The auditory stimuli are two sets of Cantonese monosyllabic words carrying different tone types: 1) **contour-level tonal contrast** (low-rising contour tone T5 /23/ vs. low-level tone T6 /22/; henceforth, contour tones), carried by two base syllables [ji] and [fan]; 2) **level-level tonal contrast** (mid-level tone T3 /33/ vs. low-level tone T6 /22/; henceforth, level tones), carried by two base syllables [jau] and [fu], as illustrated in Fig. 1. The eight tone-words had a similar syllable frequency and pronounceability based on a norming study [15]. A male native speaker of Hong Kong Cantonese who can clearly distinguish the six Cantonese tones recorded the stimuli. As seen in Fig. 1, the two tones in the contour tone set are similar in pitch height (M_{T5} = 103Hz, M_{T6} = 105Hz), but they differ only in pitch contour (T5-a rising contour and T6-a level contour). In contrast, the two tones in the level tone set differ in pitch height (M_{T3} = 114Hz, M_{T6} = 103Hz), but they both have stabilized pitch as level tones. The stimuli were normalized in duration (500 ms) and intensity (70 dB).

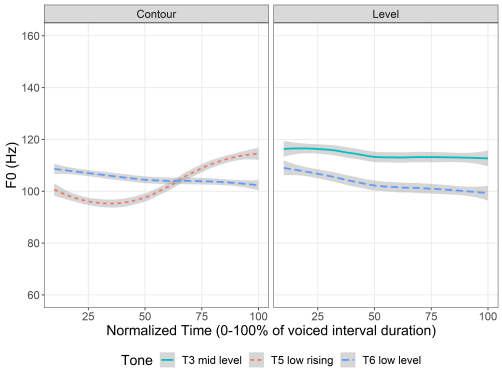


Figure 1: F0 track of a Cantonese contour-level tonal contrast (left) and a level-level tonal contrast (right) produced by a male native speaker.

There are 40 tokens (2 tones x 2 syllables x 2 sets x 5 exemplars). The eight novel words contrasting in tones, with a similar syllable frequency and pronounceability based on a norming study, were used in a novel word learning task. To minimize the influence of the Mandarin lexicon, each auditory (tone-)word will be paired with a novel object from the Novel Object and Unusual Name (NOUN) database [20]. The same word-object pairings will be used for all participants through training and tests in each experiment.

2.3. Procedure

As illustrated in Fig. 2, participants were asked to complete two consecutive experiment sessions (Day 1 and 2) with a 12-hour overnight interval. Participants' alertness levels were assessed at the beginning of each session via Stanford Sleepiness Scale (SSS) [21] to control for the time-of-day effect. On the **evening** of Day 1 (i.e., 8-10 PM), the participants went through a training in which they were instructed to learn the association between an auditory word and novel objects through feedback ("Correct" or "Incorrect"). On each trial, participants first heard an auditory word, and then they saw a pair of two novel objects on each side of the screen. Their task was to press the left or right key to choose the correct object corresponding to the word. The word pair contrasted in the level tones (T3-6) or the contour tones (T5-6). The training adopted a blocked design that each block only contained one syllable, and two syllables alternated across blocks. In total, there were six blocks (2 syllables x 3 blocks) and 40 trials (2 tones x 5 tokens x 4 repetitions) within each block.

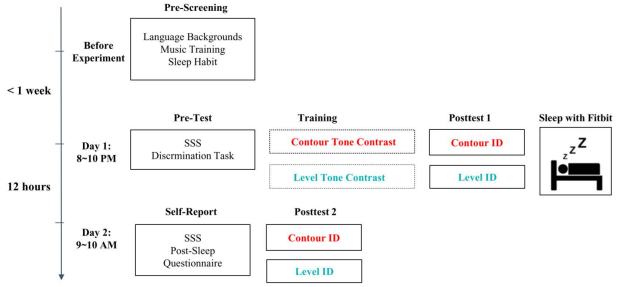


Figure 2: Experiment tasks and procedure.

Immediately following the training, participants completed the first set of word identification (ID) tests in the Posttest 1. There was no feedback in the test, and participants needed to respond when they saw two objects on the screen (1000 ms as time-out time). Participants were given 20 practice trials before the test. The ID test contains 80 trials (2 syllables x 2 tones x 5 tokens x 4 repetitions).

After the Day 1 session, participants went home to sleep wearing actigraphy (i.e., Fitbit). On the **morning** of Day 2 (i.e., after 12 hours), participants

returned for another set of word identification tests in the Posttest 2, which was identical to Posttest 1. They were also asked to complete a post-sleep questionnaire, reporting their subjective estimation of sleep last night. The experiment was conducted via the Paradigm software, Perception Research Systems, Inc. The experiments lasted for around 1 hour (i.e., a 30-min training and a 20-min ID posttest) on Day 1 and around 20 minutes on Day 2. The order of the two tonal contrasts was counterbalanced in separate sessions across the participants throughout the experiments.

2.4. Statistical analysis

To examine the effect of overnight consolidation (and training) on the tonal contrasts, mixed-effects logistic regression models were run on response accuracy (binary, 1 for correct and 0 for incorrect). The models were fitted in R using the lme4 package [22]. All models followed the backward-stepping procedure to determine a maximal random effects structure best justified by the data [23]. All categorical predictors were deviation coded (−0.5, 0.5) to test for the main effects. For training data, the first and last two blocks were coded as initial and outcome blocks to track participants' performance improvement. Block (initial vs. outcome; −0.5, 0.5) and tone (contour vs. level; −0.5, 0.5) were entered as fixed effects. For data of ID tests, session (posttest 1 vs. posttest2; −0.5, 0.5) and tone (contour vs. level; −0.5, 0.5) were entered as fixed effects.

2.5. Results

To assess whether training helped Mandarin-speaking participants learn tonal contrasts, a mixed-effects logistic model was performed on participants' response accuracy. Block and tone were entered as fixed effects, with by-participant intercept and slope, and by-trial intercept as random effects. Results of the model revealed a main effect of the block (β = 0.28, SE = 0.11, z = 2.55, p = .011) and tone (β = − 0.52, SE = 0.14, z = −3.57, p < .001), but no interaction between block and tone (β = 0.21, SE = 0.16, z = 1.33, p = .184). The findings suggests that Mandarin-speaking participants' accuracy improved in both tonal contrasts throughout the training, with higher accuracy in contour tones than level tones.

To examine the overnight sleep consolidation on tonal contrasts, session and tone were entered as fixed effects, with by-participant intercept and slope as random effects, in a mixed-effects model. Results showed a main effect of session (β = 0.20, SE = 0.05, z = 4.00, p < .001) and tone (β = −0.69, SE = 0.14, z = −5.26, p < .001). Crucially, they yielded a significant interaction between session and tone (β = −0.23, SE = 0.10, z = −2.30, p = .022). As illustrated in Fig. 3, post-hoc analyses showed that the effect of the session was significant in contour tones (β = 0.32, SE = 0.07, z = 4.04, p < .001). In contrast, the effect of the session was not significant in level tones (β = 0.09, SE = 0.06, z = 1.38, p = .169). The result of the tone effect indicates that contour tones were better identified than level tones after training (i.e., two posttests). The difficulty of learning level tones may have made them more vulnerable to interference than contour tones, so some participants exhibited declining performance for level tones over time [14].

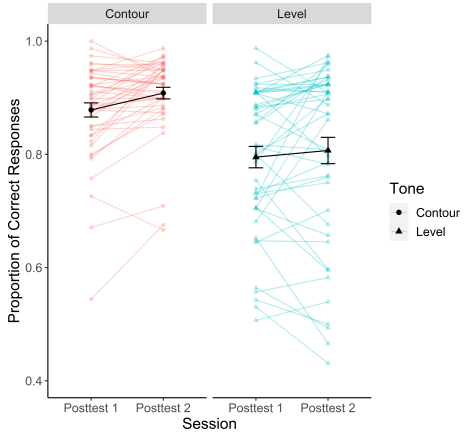


Figure 3: Mandarin-speaking participants' overnight change (Posttest 1 vs. 2) of the identification accuracy across the contour tones (left) and level tones (right) with by-participant data. Error bars indicate standard error.

The results suggest that Mandarin-speaking participants improved their accuracy in identifying novel words contrasting in contour tones, but not those contrasting in level tones, through overnight consolidation. The findings support the beneficial effect of prior knowledge on overnight consolidation. However, they do not rule out the possibility that contour tones are intrinsically better consolidated (and learned) than level tones, regardless of learners' prior knowledge. Thus, Experiment 2 examines whether Cantonese contour tones and level tones are consolidated to different degrees after an overnight interval for a control group of participants, that is, English speakers. English speakers do not have prior knowledge of lexical tones in their L1 and thus are NOT predicted to show the differential consolidation of contour and level tones.

3. EXPERIMENT 2

3.1. Participants and Methods

Twenty English-speaking participants (age range: 19-39 years; 10 males) were recruited in Hong Kong for this experiment. The participants were all native speakers of English (10 from the USA, 8 from the UK, and 2 from other English-speaking areas).

Importantly, they did not know any tone languages (e.g., Mandarin). The screening criteria (e.g., minimal exposure to Cantonese and regular sleep pattern, etc.) were identical to Mandarin-speaking participants. Experiment 2 was a replication of Experiment 1, so the procedure and statistical analyses were the same.

3.2. Results

For the model of training blocks among the English-speaking participants, block and tone were entered as fixed effects, with by-participant intercept and slope, and by-trial intercept as random effects. Results revealed a main effect of block ($\beta = 0.29$, SE = 0.12, $z = 2.38$, $p = .017$), but neither the main effect of tone ($\beta = -0.30$, SE = 0.17, $z = -1.73$, $p = .09$) nor an interaction between tone and block ($\beta = -0.17$, SE = 0.19, $z = -0.89$, $p = .37$). The findings suggest that English-speaking participants' accuracy improved in both tonal contrasts throughout the training. However, there was no difference between the two tonal contrasts.

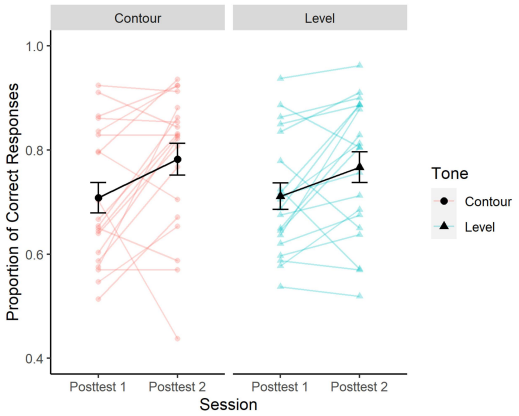


Figure 4: English-speaking participants' overnight change (Posttest 1 vs. 2) of the identification accuracy across the contour tones (left) and level tones (right) with by-participant data. Error bars indicate standard error.

For the model of ID tests among the English-speaking participants, session and tone were entered as fixed effects, with by-participant intercept and slope as random effects. The results, as illustrated in Fig. 4, showed the main effect of the session ($\beta = 0.41$, SE = 0.13, $z = 3.27$, $p = .001$) but not the main effect of tone ($\beta = -0.04$, SE = 0.06, $z = -0.61$, $p = .54$). Unlike the results of Experiment 1, they did not yield a significant interaction between session and tone ($\beta = -0.09$, SE = 0.12, $z = -0.77$, $p = .44$). The findings suggest that, unlike Mandarin-speaking participants, their English counterparts learned both tonal contrasts similarly (e.g., a similar accuracy in two posttests). Crucially, they improved their accuracy in identifying novel words contrasting contour and level tones, without a difference, through overnight consolidation.

4. DISCUSSION

This study investigated the effect of prior knowledge on L2 learners' overnight consolidation of Cantonese contour and level tones [14]. The results showed that Mandarin-speaking participants consolidated contour tones to a larger degree than level tones after an overnight interval because of their familiarity with contour tones in their L1. In contrast, a control group of participants who did not have prior knowledge of lexical tones, that is, English-speaking participants, did not show such a pattern with both tonal contrasts consolidated in a similar degree.

Consistent with prior work on tone perception and learning [18], [19], the findings about tonal contrasts indicate that contour tones were better learned than level tones among Mandarin-speaking participants who used pitch contour to distinguish native tones, but not among English-speaking participants. Importantly, the findings about overnight change suggest a beneficial role of prior knowledge with schema-related memories (i.e., contour tones) recalled better than non-schema memories (i.e., level tones) after sleep among Mandarin-speaking participants alone. Supporting the schema effect on memory consolidation [7], [10], the findings align with the model proposed by Lewis and Durrant [24], which accounts for the beneficial role of prior knowledge. The (selective) memory consolidation model states that new and old memories are replayed simultaneously during sleep, with information shared across new and old memories more strongly activated than that from only a single source.

The present findings imply that which types of tones are consolidated to a larger degree depends on learners' prior knowledge (e.g., familiarity with lexical tones in L1). They thus advanced our understanding of the mechanisms of how sleep supports successful L2 tone learning through memory consolidation. They may also facilitate tailored pedagogical applications in various educational settings. For instance, the findings enable researchers to tailor the training paradigm through the technique of re-exposure during sleep (i.e., Targeted Memory Reactivation-TMR) [25], according to L2 learners' language backgrounds. In closing, it is necessary to acknowledge the limitation of this study. A limited number of English-speaking participants, who do not have prior knowledge of lexical tones, may not be the best control group. Future studies should include a control group with a comparable number of participants who have prior knowledge of level tones (but not contour tones) in their L1 to investigate further the effect of prior knowledge on overnight consolidation of lexical tones in a different scenario.

5. ACKNOWLEDGEMENTS

This work was supported by a Start-up Fund, the Anti-epidemic Fund and the IEMS Research Grant at the Hong Kong University of Science and Technology. The authors would like to thank Dr. Caicai Zhang for her helpful comments on the early version of the research, Jeremy Chui and Yuxin Zhu for their help in data collection.

6. REFERENCES

[1] Rasch, B., 2017. Sleep and language learning. *Brain Lang.* 167, 1–2.

[2] Batterink, L. J.Oudiette, D.Reber, P. J.Paller, K. A., 2014. Sleep facilitates learning a new linguistic rule. *Neuropsychologia* 65, 169–179.

[3] Davis, M. H.Di Betta, A. M.Macdonald, M. J. E.Gaskell, M. G., 2009. Learning and consolidation of novel spoken words. *J. Cogn. Neurosci.* 21, 803–820.

[4] Fenn, K. M.Nusbaum, H. C.Margoliash, D., 2003. Consolidation during sleep of perceptual learning of spoken language. *Nature* 425, 614–6.

[5] Payne, J. D. *et al.*, 2015. Napping and the selective consolidation of negative aspects of scenes. *Emotion* 15, 176–186.

[6] Studte, S.Bridger, E.Mecklinger, A., 2017. Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain Lang.* 167, 28–35.

[7] Hennies, N.Ralph, M. A. L.Kempkes, M.Cousins, J. N.Lewis, P. A., 2016. Sleep spindle density predicts the effect of prior knowledge on memory consolidation. *J. Neurosci.* 36, 3799–3810.

[8] Tse, D. *et al.*, 2007. Schemas and memory consolidation. *Science* 316, 76–82.

[9] Tse, D. *et al.*, 2011. Schema-dependent gene activation and memory encoding in neocortex. *Science* 333, 891–895.

[10] Zion, D. BenNevat, M.Prior, A.Bitan, T., 2019. Prior knowledge predicts early consolidation in second language learning. *Front. Psychol.* 10, 1–15.

[11] Havas, V.Taylor, J. S. H.Vaquero, L.de Diego-Balaguer, R.Rodríguez-Fornells, A.Davis, M. H., 2018. Semantic and phonological schema influence spoken word learning and overnight consolidation. *Q. J. Exp. Psychol.* 71, 1469–1481.

[12] Yip, M. J. W., 2002. *Tone*. Cambridge : Cambridge , 2002.

[13] Chapelle, A. de la *et al.*, 2022. Sleep affects higher-level categorization of speech sounds, but not frequency encoding. *Cortex* 154, 27–45.

[14] Qin, Z.Zhang, C., 2019. The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLoS One* 14, e0221498.

[15] Francis, A. L.Ciocca, V.Ma, L.Fenn, K., 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294.

[16] Hao, Y. C., 2012. Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *J. Phon.* 40, 269–279.

[17] Gandour, J. T., 1983. Tone perception in far eastern-languages. *J. Phon.* 11, 149–175.

[18] Qin, Z.Jongman, A., 2016. Does Second Language Experience Modulate Perception of Tones in a Third Language?. *Lang. Speech* 59, 318–338.

[19] Laméris, T. J.Post, B., 2022. The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers. *Second Lang. Res.*.

[20] Horst, J. S.Hout, M. C., 2016. The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behav. Res. Methods* 48, 1393–1409.

[21] Hoddes, E.Zarcone, V.Smythe, H.Phillips, R.Dement, W. C., 1973. Quantification of Sleepiness: A New Approach. *Psychophysiology* 10, 431–436.

[22] Bates, D. M.Maechler, M.Bolker, B.Walker, S., 2015. lme4: linear mixed-effects models using S4 classes. *J. Stat. Softw.* 67, 1–48.

[23] Matuschek, H.Kliegl, R.Vasishth, S.Baayen, H.Bates, D., 2017. Balancing Type I error and power in linear mixed models. *J. Mem. Lang.* 94, 305–315.

[24] Lewis, P. A.Durrant, S. J., 2011. Overlapping memory replay during sleep builds cognitive schemata., *Trends in Cognitive Sciences*, 15,. 343–351, 2011.

[25] Schreiner, T.Rasch, B., 2017. The beneficial role of memory reactivation for language learning during sleep: A review. *Brain Lang.* 167, 94–105.

VOICING AND REGISTER IN MNONG RÂLÂM

Marc Brunelle[1], Đinh Lư Giang[2], Tạ Thành Tấn[3]

[1,3]University of Ottawa, [2]University of Social Sciences and Humanities, Hồ Chí Minh City
[1]mbrunell@uottawa.ca, [2]lugiangdinh@gmail.com, [3]tta061@uottawa.ca

ABSTRACT

In many Southeast Asian languages, an original voicing contrast in onset stops has transphonologized into a binary *register* contrast realized on following vowels through modulations of vowel quality, phonation and pitch. Register languages with optional residual voicing do exist, but no such language has been studied instrumentally so far. For this reason, the early stages of register formation, in which voicing conditions the phonetic perturbations making up register, are still ill-understood.

Mnong Râlâm, an Austroasiatic language previously described as preserving a voicing contrast in onset stops, has been anecdotally reported to have redundant register. We investigated it to determine if it can shed light on the relation between voicing and register. Results suggest that register is the primary contrastive element in Mnong Râlâm and that onset voicing is a common but optional cue. We found no trade-off between voicing and register.

Keywords: Register; Voicing; Austroasiatic; Vowel quality; Phonation

1. INTRODUCTION

In many Southeast Asian languages of the Austroasiatic and Austronesian phyla, onset voicing was *transphonologized* into register: the contrastive role of voicing in onset obstruents was taken over by a binary register contrast on following rhymes [1-4]. The term *register* refers to a type of phonological contrast realized as a combination of various possible phonetic properties: the *low register* that follows former voiced stops typically has a lower pitch, a laxer/breathier phonation and higher vowels than the *high register* following former voiceless stops. These registral properties have been instrumentally investigated in languages reported to have fully neutralized onset voicing [5-15] as well as in languages in which there are optional traces of voicing in the onset system [16, 17]. However, there has not yet been any systematic study of a language at a conservative stage of register development in which onset voicing and register are still redundant.

As a result of this empirical gap, the mechanisms proposed to explain how register is phonologized and replaces onset voicing are based on diachronic reconstructions and inferences about purported coarticulatory processes. It has been proposed that the loss of voicing results in an increased aspiration that perturbates the spectral properties of the following vowel [2, 3, 18]. Another view is that articulations meant to increase the size of the supraglottal cavity in order to override the aerodynamic voicing constraint have effects on these spectral properties [4, 19, 20].

Mnong is a term used to designate a continuum of language varieties spoken in Eastern Cambodia and the Highlands of Central Vietnam. It belongs to the South Bahnaric branch of Austroasiatic. While Central Mnong (also Bunong/Phnong) is uncontroversially registral [21-23], Eastern Mnong is claimed to preserve a voicing contrast in obstruents and has never been reported to be registral [24-26]. We chose to study the Râlâm variety of Eastern Mnong (Yang Tao commune, Lăk district, Đăk Lăk province, Vietnam) because of anecdotal evidence that it may have redundant register.

According to Blood [24], Mnong Râlâm has a voicing contrast in plain onset stops (bolded in Table 1), alongside implosives and aspirated series. If it also has redundant register, vowels following plain voiced stops should bear low register properties.

p	**t**	**c**	**k**	ʔ
pʰ	tʰ	cʰ	kʰ	
b	**d**	**ɟ**	**g**	
ɓ	ɗ	ʃ		
	s			h
m	n	ɲ	ŋ	
w	l, r	j		

Table 1: Mnong Râlâm onsets, adapted from [26]

In this paper, we investigate the phonetic realization of stop voicing and register in Mnong Râlâm. Our research questions are the following:

RQ1: Does Mnong Râlâm preserve a voicing contrast in plain onset stops?

RQ2: Does Mnong Râlâm have a register contrast on vowels following plain stops?

RQ3: Is there a trade-off between voicing and register in Mnong Râlâm? More specifically, is there evidence that syllables headed by onset stops with weaker voicing have clearer registral properties, as we may expect if the language is transitioning from voicing to register?

2. METHODS

Twenty-three native speakers (12 women) of Mnong Râlâm born between 1944 and 1992 (mean: 1975, sd: 15) were recorded in the hamlet of Buôn Dơng. They were all originally from the area, except for one man who was raised in a nearby city. They all spoke Vietnamese and 21 of them also spoke Êđê (also Rhade), a Chamic language that is the lingua franca used between minority groups in the province of Đăk Lăk.

2.1. Experiment

Sixty real words comprising target syllables made up of combinations of all dental and velar onsets and the five vowels /iː, ɛː, aː, ɔː, uː/ were selected. Open syllables were preferred; when they were not available, syllables closed by sonorants were chosen. Target words were either monosyllables or disyllables ending in target syllables. In this short paper, we report full results for 19 target syllables starting with the plain stops /t, k, d, g/ (16 of which being open monosyllabic words) and an additional 9 target syllables with the aspirates and implosives /ɗ, tʰ, kʰ/ in §3.1. These words were read four times in a fixed frame sentence, randomized in SpeechRecorder [27] along with the rest of the wordlist.

Two signals were acquired during the experiment: an audio signal recorded with a Shure Beta 53 and an EGG signal with a Glottal Enterprises EG2-PCX.

2.2. Data analysis

In order to assess the robustness of the voicing contrast in onset stops, we measured the VOT and the onset and offset of closure voicing from the EGG signal. The presence of register was determined by measuring F0 (pitch), F1 and F2 (vowel quality), and H1*-H2* (phonation) from the audio signal at each millisecond of the target vowels. As 25 ms windows were used for spectral measures, we disregarded the initial and final 12 sampling points to avoid extracting results from windows spanning adjacent segments or pauses. Measurements were excluded 1) if their derivatives were 0.95 sd larger or smaller than the mean derivative or 2) if they were more than 3 sd away from the mean of all tokens produced by the same speaker with the same vowel and onset voicing. All remaining measures were z-normalized by speaker. To ensure readability, z-scores were converted back to familiar scales based on means and standard deviations obtained from the entire data set.

Linear mixed models were fitted on VOT and on all relevant vocalic properties averaged over the first 10 ms of vowels (which, as we will see, is the area of maximal contrast between the high and low registers). Fixed factors included *Register/Voicing*, *Vowel* and *Place of Onset* and all their two-way interactions. Random intercepts for *Word* and *Speaker* were also included. Reference levels for all models presented below are *Register: High*; *Vowel: ɔː*; *Place: Coronal*.

3. RESULTS

3.1. Onsets

VOT values for stops are reported in Fig. 1. As expected, aspirated stops have a strong lag VOT and most implosives have a strong lead VOT. The few implosives with a slight positive VOT are for the most part prenasalized. Voiceless plain stops (in orange) all have a moderate lag VOT, but voiced stops have a bimodal distribution, some of them having a strong lead VOT while others have a moderate lag VOT comparable to that of voiceless stops (Voicing: $\beta = 1.2$, $t = 0.67$, $p = 0.54$, no sign. interactions).

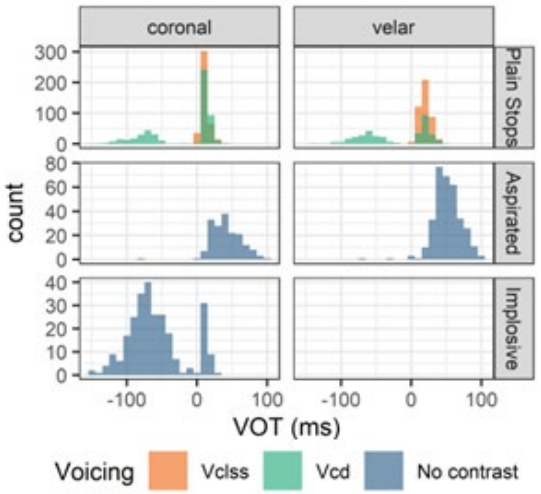


Figure 1: VOT in onset stops

This unexpected distribution of VOT in voiced stops led us to look into more detail at the realization of voicing during their closure. We divided voiced stops into three groups: those that are fully voiced, those that are devoiced (including tokens with carryover voicing from a preceding sonorant that lasts for less than 30% of the closure) and tokens that have a voiced closure with a voiceless release, i.e., tokens that have voicing over more than 30% of their closure, but that have no voicing around their release. An example of the latter is given in Fig. 2.

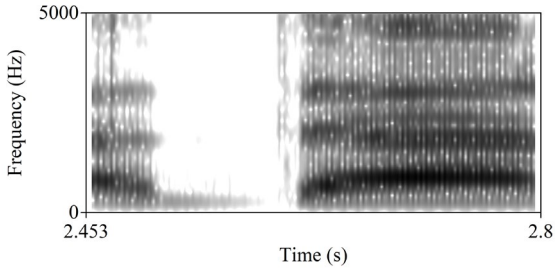


Figure 2: Voiced stop with a voiceless release in the word /ada/ ‘kind of duck’

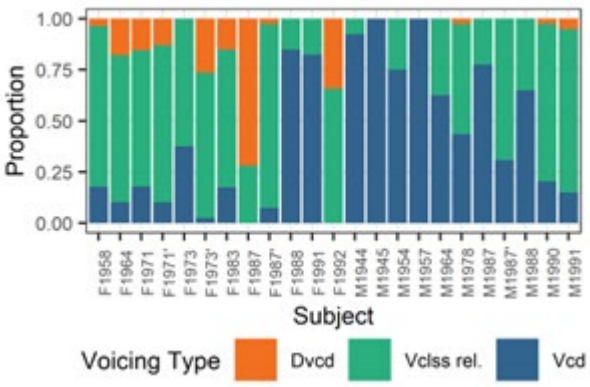


Figure 3: Distribution of types of voicing in phonemically voiced stops, by speaker

The distribution of these three types of voiced stops by speaker is given in Fig. 3. Women, on the left, have a greater tendency to devoicing, with a significant proportion of devoiced tokens and a clear majority of tokens with voiceless releases. This is an indication that phonetic voicing may not be the primary cue of the "voicing" contrast in Mnong stops.

It should also be noted that there is no significant difference between the VOT of fully devoiced stops and voiced stops with voiceless releases.

3.2. Vowels

Let us turn to vocalic cues to determine if there is evidence that register has developed out of the original voicing contrast. In Fig. 4, we see that mean f0 is higher in the high than the low register at vowel onset. However, this effect is weak: it is not significant after coronals (RegisterLow: $\beta = -9.6$, $t = -0.91$, $p = 0.42$) but is after velars (RegisterLow: PlaceVelar: $\beta = -54$, $t = -7.3$, $p < 0.01$).

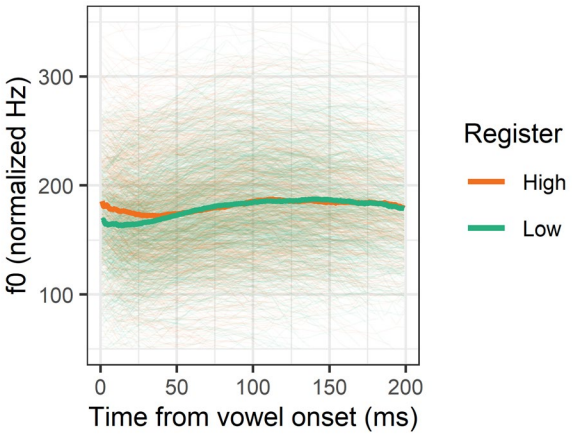


Figure 4: f0, by register (thick lines are means, thin lines are individual tokens)

Registers differ much more in mean H1*-H2* than in mean f0, as can be seen in Fig. 5. The low register has a higher H1*-H2* value than the high one over the first 100 ms, suggesting that it is breathier or laxer. However, this seems to be a weak, vowel-specific effect: it is not significant when the vowels /ɔː, aː/ are used as references in our mixed model (e.g., with a: RegisterLow: $\beta = 2.45$, $t = 1.5$, $p = 0.213$) but is significant with vowels /iː, ɛː, uː/ (e.g., with u: RegisterLow: $\beta = 5.95$, $t = 3.6$, $p = 0.023$).

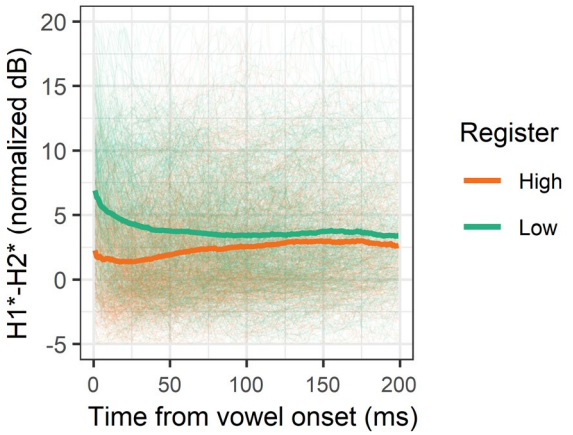


Figure 5: H1*-H2*, by register (thick lines are means, thin lines are individual tokens)

Turning to vowel quality, we can see in Fig. 6 that open and mid vowels have a significantly lower onset F1 (RegisterLow: $\beta = -223$, $t = -3.0$, $p = 0.039$; no significant interactions including register) and more dramatically falling onglides in the low register than in the high one. Close vowels also start on a lower F1 in the low register, but they have slightly rising onglides in the high register. F2 patterns are more homogeneous, with low register vowels having a systematically higher onset F2, but this difference is not significant (RegisterLow: $\beta = 227$, $t = 2.1$, $p = 0.106$; no significant interactions including register).

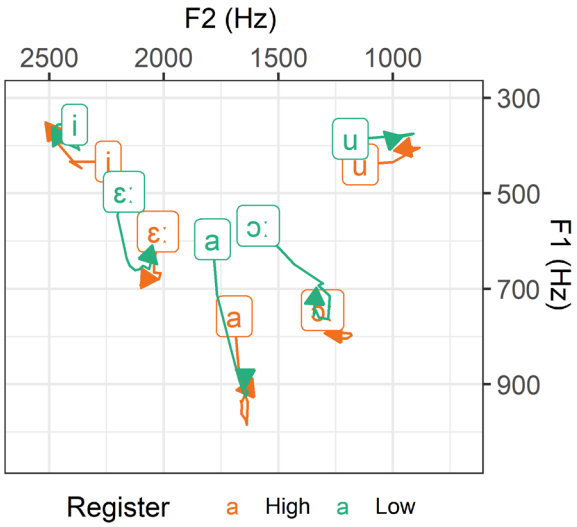


Figure 6: Vowel trajectories, from onset (boxes) to 300th ms (arrowhead), by register

3.3. Register cues by type of voicing

Is there a relation between the type of onset voicing and the salience of register properties on the following vowel? To answer that question, we plotted the four acoustic properties explored in §3.2 at vowel

onset following phonologically voiced stops in Fig. 7. The x-axis largely corresponds to the three-way classification in §3.1: fully voiced stops have a proportion of 1, devoiced stops are located in the leftmost area of the chart (from 0 to 0.3), and voiced stops with a voiceless release fall in the middle.

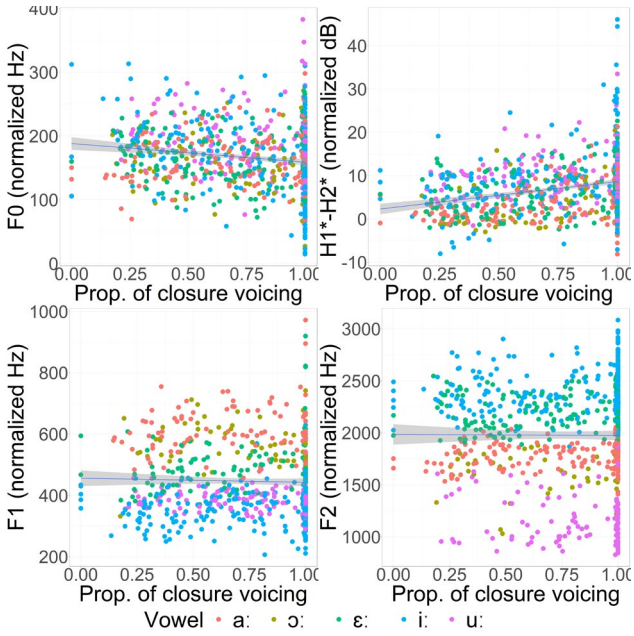


Figure 7: Acoustic properties of the low register at vowel onset, by proportion of closure voicing in "voiced" stops.

Fig. 7 shows weak significant correlations between the proportion of closure voicing and f0 ($r = -0.16$, $p < 0.001$) and H1*-H2* ($r = 0.27$, $p < 0.01$). Rather than trading-off, voicing and register seem to reinforce each other: more robust voicing is associated with more pronounced low register cues, like a lower f0 and a greater H1*-H2* (laxer/breathier phonation).

4. DISCUSSION

As shown in §3.1 and in Fig. 7, there is no longer a systematic voicing contrast in Mnong Râlâm stops (RQ 1). A significant proportion of the stops that were described as voiced in other sources are either fully devoiced or only exhibit limited carryover voicing stemming from a preceding sonorant. Nonetheless, the majority of "voiced stops" preserve significant closure voicing, even if it does not persevere until the release in about half of them. As shown in Fig. 3, women have a higher proportion of devoiced voiced stops and voiced stops with voiceless releases, while men have more fully voiced closures, an asymmetry that was also found in other register languages [15-17]. The fact that this asymmetry is not language-specific could be an indication that the trend to more phonetic devoicing in women is in part anatomically driven rather than fully attributable to sociolinguistic factors.

Contrary to our expectations and to previous descriptions, Mnong Râlâm already seems to have a well-developed register system largely based on vowel quality and phonation (RQ2). As shown in Fig. 6, open and mid vowels /ɛː, aː, ɔː/ tend to start with dramatic falling onglides in the low register, while close vowels /iː, uː/ have more limited rising onglides in the high register. This pattern of register-conditioned diphthongization corresponds to that observed in other register languages [4, 28, 29].

Register-conditioned phonation differences, here measured with H1*-H2*, appear relatively salient if we aggregate all vowels, but this seems largely attributable to high vowels. Other phonation measures (both spectral and EGG-derived) will be reported in longer papers, along with preliminary perceptual results confirming that F1 and phonation are the main register cues and that they are perceptually more salient than closure voicing.

We found a weak correlation between two acoustic properties of register, f0 and H1*-H2*, and the proportion of closure voicing in preceding low-register onset stops (RQ3). Closure voicing is negatively correlated with f0 and positively correlated with H1*-H2*. This is the opposite of the expected trade-off between voicing and register and is reminiscent of patterns found in register languages with less prevalent closure voicing [16, 17]. Overall, this suggests that register is already well-established in Mnong Râlâm and that closure voicing is either a way of making the low register more salient or a secondary consequence of clearly articulating it.

We conclude that Mnong Râlâm has a stable register system that is reinforced by optional closure voicing (especially in men). It is likely that previous researchers described it as a language with a voicing contrast in onset stops because of this frequent secondary closure voicing. Since its register contrast and secondary voicing appear to be in a stable relation and do not trade-off, Mnong Râlâm cannot inform us on the early stages of register formation.

5. ACKNOWLEDGEMENTS

We would like to thank H Hê Bkrông and H Wăn Du for their help with data collection and recruitment, and all our participants for their patience and generosity with their time. We are also grateful for the People's Committee of the Lăk district (Đăk Lăk province) for their help and efficiency in granting us permission to work in the commune of Yang Tao. We also acknowledge the contributions of Sabrina McCullough, Sue-Anne Richer and Jeanne Brown, who annotated and processed the acoustic recordings.

6. REFERENCES

[1] Henderson E. 1952. The main features of Cambodian pronunciation *Bulletin of the School of Oriental and African Studies,* 14, 453-476.

[2] Haudricourt A. 1965. Les mutations consonantiques et les occlusives initiales en mon-khmer *Bulletin de la Société Linguistique de Paris,* 60, 160-172.

[3] Huffman F. E. 1976. The register problem in fifteen Mon-Khmer languages *Oceanic Linguistics special publication,* vol. Austroasiatic Studies, part 1, 575-589.

[4] Ferlus M. 1979. Formation des Registres et Mutations Consonantiques dans les Langues Mon-Khmer *Mon Khmer Studies,* VIII, 1-76.

[5] L.Thongkum T. 1987. Another Look at the Register Distinction in Mon *UCLA Working Papers in Phonetics,* 67, 132-265.

[6] L.Thongkum T. 1988. Phonation Types in Mon-Khmer Languages. In: Fujimura O. (ed), *Vocal Fold Physiology: Vocal Physiology, Voice Production, Mechanisms and Functions*, New York: Raven Press, 1988, 319-334.

[7] L.Thongkum T. 1989. An acoustic study of the register complex in Kui (Suai) *Mon-Khmer Studies,* 15, 1-19.

[8] L.Thongkum T. 1990. The interaction between pitch and phonation type in Mon: Phonetic implications for a theory of tonogenesis *Mon-Khmer Studies,* 16-17, 11-24.

[9] Abramson A. S., Luangthongkum T., Nye P. W. 2004. Voice register in Suai (Kuai): An analysis of perceptual and acoustic data *Phonetica,* 61, 147-171.

[10] Abramson A. S., Nye P. W., Luangthongkum T. 2007. Voice register in Khmu': Experiments in production and perception *Phonetica,* 64, 80-104.

[11] Abramson A. S., Tiede M. K., Luangthongkum T. 2015. Voice Register in Mon: Acoustics and Electroglottography *Phonetica,* 72, 237-256.

[12] Watkins J. 2002. *The Phonetics of Wa: Experimental Phonetics, Phonology, Orthography and Sociolinguistics*. Canberra: Australian National University.

[13] Brunelle M. 2006. A phonetic study of Eastern Cham register, In: Sidwell P., Grant A. (eds.), *Chamic and Beyond*, Sidney: Pacific Linguistics, 2006, pp. 1-36.

[14] Brunelle M. 2009. Contact-induced change? Register in three Cham dialects *Journal of Southeast Asian Linguistics,* 2, 1-22.

[15] Brunelle M., Brown J., Hà P. T. T. 2022. Northern Raglai voicing and its relation to Southern Raglai register: evidence for early stages of registrogenesis *Phonetica*, 79 (2), 151-188

[16] Brunelle M., Tạ T. T., Kirby J., Đinh L. G. 2020. Transphonologization of voicing in Chru: Studies in production and perception *Laboratory Phonology: Journal of the Association for Laboratory Phonology,* 11 (1), 15.

[17] Tạ T. T., Brunelle M., Nguyễn T. Q. 2022. Voicing and register in Ngãi Giao Chrau: Production and perception studies *Journal of Phonetics,* 90, 101115.

[18] Wayland R., Jongman A. 2002. Registrogenesis in Khmer: A phonetic account *Mon-Khmer Studies*, 32, 101-115.

[19] Gregerson K. 1976. Tongue-root and Register in Mon-Khmer in *Austroasiatic Studies*. vol. 1, Jenner P. N., Thompson L., Starosta S. (eds.), Honolulu: University Press of Hawaii, 1976, 323-369.

[20] Brunelle M. 2010. The role of larynx height in the Javanese tense ~ lax stop contrast In: Mercado R., Potsdam E., Travis L. (eds.), *Austronesian Contributions to Linguistic Theory: Selected Proceedings of AFLA*, Amsterdam/Philadelphia: John Benjamins, 2010, 7-24.

[21] Butler B. 2010. *Registrogenesis in Bunong: The Interaction of Phonetics and Phonology*, PhD thesis, Cornell University.

[22] Butler B. 2015. Bunong, In: Jenny M., Sidwell P. (eds.), *The Handbook of Austroasiatic Languages,* 1, Brill, 719-745.

[23] Vogel S., Filippi J.-M. 2006. *Elements de langue Phnong,* Phom Penh: Editions Funan.

[24] Blood H. 1976. The Phonemes of Uon Njun Mnong Rơlơm *Mon-Khmer Studies,* 4-23.

[25] Đinh L. T. 1994. Tiếng Mnông: sự phân chia và thông nhất giữa phương ngữ *Tập sản khoa học Á ĐHTH TP.HCM - Annals of HCM University,* 1, 71-79.

[26] Đinh L. T. 2005. *Từ Điển Việt-Mnông Rlâm*: Nhà xuất bản Đại học Quốc gia TP. Hồ Chí Minh.

[27] Draxler C., Jänsch K., "SpeechRecorder-a Universal Platform Independent Multi-Channel Audio Recording Software," in *LREC*, 2004.

[28] Huffman F. E. 1985. Vowel permutations in Austroasiatic languages: Papers presented to Paul K. Benedict for his 71st birthday In: Thurgood G., Matisoff J., Bradley D. (eds.), *Linguistics of the Sino-Tibetan Area: The State of the Art*, Canberra: Pacific Linguistics Series C 87, Australian National University, 1985, 141-145.

[29] Jenner P. N. 1974. The Development of Registers in Standard Khmer In: Liêm N. Đ. (ed.), *Southeast Asian Linguistic studies*. vol. 1, Canberra: Autralian National University, 1974, 47-60.

REALIZATION OF LOW TONE SEQUENCES IN DISYLLABIC WORDS IN A LARGE MANDARIN SPEECH CORPUS

Yaru Wu[1,2,3], Yiya Chen[4,5], Lori Lamel[3]
[1]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[3]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[3]LISN, Univ. Paris-Saclay, 91405 Orsay cedex, France
[4]Leiden University Centre for Linguistics (LUCL)
[5]Leiden Institute for Brain and Cognition (LIBC)
yaru.wu@unicaen.fr, yiya.chen@hum.leidenuniv.nl, lori.lamel@lisn.fr

ABSTRACT

This study investigated the acoustic realization of disyllabic Low tone (T3T3) words in Mandarin. A large corpus of 850 hours of journalistic speech was used. The segmentation of the continuous Mandarin speech was carried out using the LISN (former LIMSI) speech transcription system in forced alignment mode. The full corpus was aligned twice using two different pronunciation dictionaries, one with and one without systematic tone variants. We analyzed the tonal representations of the underlying T3T3 sequences that the LISN system produced with the two alignment strategies. We further investigated the effect of different factors, namely prosodic position, word frequency, tonal contexts, and parts of speech, on conditioning the surface realization of the underlying T3T3 sequences. The LISN outputs suggest that the first T3 of disyllabic T3T3 words is not always realized as T2 in continuous speech and a range of factors conditions its acoustic realizations.

Keywords: Mandarin, tone realization, large corpora, variation factors

1. INTRODUCTION

Over the last decades, there has been great progress in both the construction and availability of large speech corpora (e.g., LDC) and automatic speech recognition techniques that are capable of learning powerful speech representations in acoustic audio signals (e.g., [1]). In this paper, we capitalize on this progress and investigate the acoustic realizations of Low tone sequences (T3T3) in disyllabic words in a large Mandarin speech corpus.

It is well-known that Low tone sequences in Mandarin can change through sandhi, where a Low tone is pronounced as a rising pitch before another Low tone [2] (known as Low tone sandhi or T3 sandhi). Many factors can influence the application of T3 sandhi (e.g., [3]). While there is debate over how to characterize the factors that condition T3 sandhi and whether the rising pitch of the Low tone sandhi variant is completely neutralized with the rising lexical tone (T2), it has been generally accepted, based on both impressionistic observations and lab speech, that T3 undergoes the sandhi process in disyllabic words. Taking a corpus approach, Yuan & Chen [4] investigated the f0 realization of T3 in disyllabic words (T3T3) using the LDC corpus [5] (i.e. LDC2005S15 & LDC98S73) and compared them to disyllabic words with T2T3 sequences. They found that the T3 sandhi variant differs from T2 in terms of the magnitude and time span of the f0 rise.

In this study, we used the LISN (former LIMSI) speech transcription system to test the categorization of different T3 realizations in disyllabic words with T3T3 sequences. The output of the LISN system enables self-supervised learning which has been shown to be effective in situations with limited labeled data. It can be used to discover structural patterns in a significantly larger amount of unlabeled data by exploiting multi-dimensional contextual information in the corpus. The goals of the study was to (1) gain a better understanding of the realization of Low tone in a sandhi context from naturally occurring corpus speech, (2) learn how it varies according to different factors, including prosodic position, word frequency, tonal contexts, and parts of speech, and (3) determine the impact of these variation factors on the acoustic realizations of disyllabic T3T3 sequences in standard Mandarin.

2. METHOD

2.1. Corpus and alignments

A Mandarin journalistic corpus, distributed by LDC [6, 7], was used. It contains about 850 hours of continuous Mandarin speech (e.g., [8, 9]), with 9M word-tokens. The segmentation at the

word and phone levels was carried out using the speech transcription system from LISN in forced alignment mode [10, 11, 12]. The pronunciation lexicon provides phone level representations for each word (tones included). The best matching pronunciations among potential realizations were selected automatically during the forced alignment. Pauses, hesitations, and breath were also detected automatically by the system. Neutral tone was not included in the pronunciation lexicon. The minimal phone segment duration is 30 ms given the acoustic modeling technical constraints of a 3-state model and a 10 ms (frame) step [13, 14].

The full corpus was aligned twice using two different pronunciation dictionaries. The first set of alignments used a base pronunciation lexicon, which includes almost no variants (alignments V0). Words with pronunciation variants (less than 4% of all word-tokens) in the first set of alignments were not concerned by our analyses. We then used an expanded pronunciation dictionary for the second set of alignments (V1), in which tone variants were introduced for the first and last syllable of each word. This means words' first and last syllables can be aligned with any of the four tones. In this study, we decided to focus on T3T3 disyllabic words. Therefore, both syllables of the words can be aligned with any of the four tones the system sees fit. Using this method, we investigated the differences between V0 (reference tone) and V1 (surface realizations of tones) based on the LISN alignments. For instance, 打理(/da3li3/, "manage") can only be aligned with T3T3 in V0; in V1, however, it can be aligned as any of the 16 possibilities ($T_{1-4}T_{1-4}$). These two sets of alignments, i.e., without (V0) and with tone variants (V1), made it possible to quantify the realization of tones with respect to different variation factors.

2.2. Investigated factors

We analyzed the realization of the first T3 in T3T3 disyllabic words according to the following factors: prosodic position, left tonal context, relative word frequency, and part of speech. To better understand the realization of the first T3 in T3T3 disyllabic words, we decided to exclude T3T3 words immediately following another T3 syllable.

The two prosodic positions concerning the first T3 of disyllabic words are phrase-initial and word-initial. Phrase-initial tone is defined as the tone of the first syllable of a disyllabic word immediately following a pause (including silence, hesitation and breath) $\geq$ 100 ms. This threshold was empirically determined given that pause durations of 50, 100 and 200 ms gave similar results with respect to prosodic position. Word-initial tone refers to tones in the first syllable of disyllabic words that do not appear at a phrase-initial position. The left tonal context concerns the tone of the syllable that immediately precedes the T3T3 word. We used the relative word frequency values provided in SUBTLEX-CH for our analyses on the word frequency factor [15]. The SUBTLEX-CH provides word and character frequency measures based on a corpus of film subtitles, covering 33.5 million words or 46.8 million characters. The relative frequency ranges from 0 to 100. The part of speech (POS) of each word-token of the corpus was automatically annotated using Stanza [16]. This supplementary annotation allows us to examine the link between tone realization and POS. Words without matching POS were excluded from our analyses. Given that few word tokens were concerned for "conjunction" (CCONJ) and for "particle" (PART), we did not include these two categories in our analyses. The grammatical words, i.e., words that are adposition (ADP), auxiliary (AUX), determiner (DET), numeral (NUM), pronoun (PRON), were categorized into the same group named GRAM. Details on the concerned level of each variation factor can be found in Tab. 1.

Factors	Levels	Examples
Left tonal context	Tone1#_	T1#打(/da3/)理
	Tone2#_	T2#打(/da3/)理
	Tone3#_	T3#打(/da3/)理
	Tone4#_	T4#打(/da3/)理
Prosodic position	Phrase-initial	##打(/da3/)理(*manage*)
	Word-initial	#打(/da3/)理
Relative frequency	Scaled to the range of 0 to 100	
Parts of speech	ADJ (adjective), ADP (adposition), ADV (adverb), AUX (auxiliary), CCONJ (coordinating conjunction), DET (determiner), NOUN (noun), NUM(numeral), PART (particle), PRON (pronoun), PROPN (proper noun), VERB (verb)	

Table 1: Factors investigated for the first T3 of T3T3 disyllabic words. # stands for word boundary; ## for phrase boundary.

2.3. Statistical analyses and random forest model

We carried out the statistical analyses using MCMCglmm package in R [17], allowing us to fit Generalized Linear Mixed Models using Markov chain Monte Carlo techniques. The results of the statistical analyses are reported with a 95% highest posterior density (HPD) interval for each coefficient.

We also carried out a random forest classifier on the prediction of the realization of the first T3 in T3T3 words. The factors mentioned in Table 1 were included as independent variables: relative

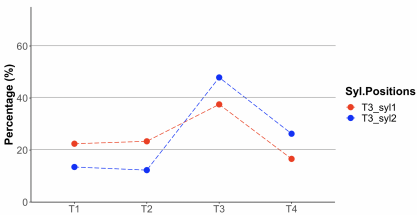


Figure 1: Realization of the first and the second T3 of T3T3 disyllabic words. The four types of realizations are shown on the x-axis. The first T3 is shown in red and second in blue.

word frequency, left tonal context, part of speech, and prosodic position. In addition to this list, we also included the part of speech of the preceding and the following word as independent variables. The contribution of each factor to the model was then evaluated. The data were randomly selected and divided into two parts: (1) 70% of the data was used for training and (2) the remaining 30% of the data was reserved as test data. The 10-fold cross-validation method was applied to extract the best combination of hyperparameters.

3. RESULTS

Fig. 1 presents the percentage of the first (in red) and the second (in blue) T3 of T3T3 disyllabic words realized as T1, T2, T3 or T4. For both T3s, the percentages of the four realizations add up to 100%. The figures show that the first T3 of T3T3 words is realized less as T3 than the second T3 of T3T3 words.

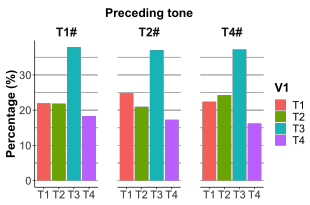


Figure 2: Realization of the first T3 in T3T3 words according to the preceding tone.

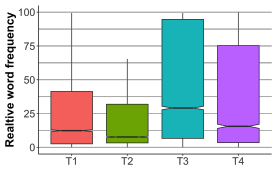


Figure 3: Relative word frequency range for each realization of the first T3 of T3T3 words.

Fig. 2 shows the realization of the first T3 in T3T3 words as a function of the left tonal context (post-lexical tonal context) of the word in question.

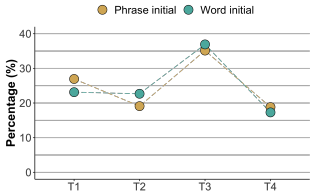


Figure 4: Realization of the first T3 in T3T3 words according to the prosodic position of the syllable carrying the tone.

The first T3 of T3T3 words is more likely to be realised as T1 when it is preceded by T2 than when it is preceded by T4 ($T1 \times PrecededByT2$ = 0.19372, 95%HPD [0.06056, 0.28655], p< 0.01). It is less likely to be realised as T2 when it is preceded by T1 or T2 than when it is preceded by T4 ($T2 \times PrecededByT1$ = -0.10165, 95%HPD [-0.21333, -0.01836], p< 0.05; $T2 \times PrecededByT2$ = -0.13801, 95%HPD [-0.21717, -0.05662], p< 0.01). Interestingly, when the T3 is preceded by T2, it is more likely to be realized as T1 than as T2, probably due to the delayed high f0 peak realization from the preceding T2 [18]: T2(3-**5**)#T1(**5**-5)T3.

Fig. 3 shows the relationship between relative word frequency and the realization of the first T3 in T3T3 words. The relative word frequency ranges from 0 to 100 (see more details in [15]). Results based on the multinominal logistic confirmed that the first T3 of T3T3 words is less likely to be realized as T1, T2 or T4 in high frequency words ($T1 \times frequency$ = -0.00364, 95% HPD $[-0.00400, -0.00326]$, p < 0.01; $T2 \times frequency$ = -0.00198, 95% HPD $[-0.00232, -0.00161]$, p < 0.01; $T4 \times frequency$ = -0.00202, 95% HPD $[-0.00250, -0.00151]$, p < 0.01).

Fig. 4 presents the realization of the first T3 in T3T3 words according to the prosodic position of the word (i.e. IP initial *vs* word-initial positions). Similar patterns are found for both positions as far as the realization of the first T3 is concerned. The T3 in question is observed to be realized as T1 slightly more often than as T2 at the IP-initial position.

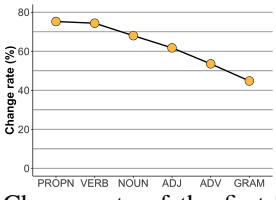


Figure 5: Change rate of the first T3 in T3T3 words according to different parts of speech.

Fig. 5 shows the change rate of the first T3 in T3T3 disyllabic words as a function of the parts

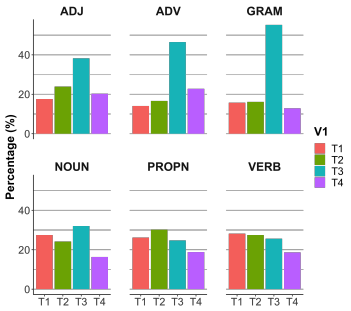


Figure 6: Realization of the first T3 in T3T3 words for the concerned parts of speech.

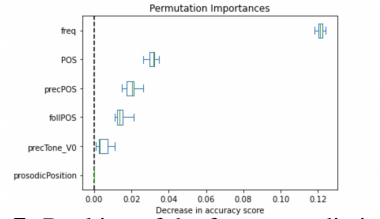


Figure 7: Ranking of the factors predicting of the realization of the first T3 in T3T3 words.

Comparisons	95% HPD	p value
T1 × ADJ = 0.60152	[0.38861, 0.80995]	p < 0.01
T1 × ADV = 0.51933	[0.39077, 0.65743]	p < 0.01
T1 × NOUN = 1.33442	[1.19102, 1.43624]	p < 0.01
T1 × PROPN = 1.07729	[0.66248, 1.45905]	p < 0.01
T1 × VERB = 1.73975	[1.57341, 1.83614]	p < 0.01
T2 × ADJ = 0.79490	[0.56019, 1.04017]	p < 0.01
T2 × ADV = 0.48989	[0.38881, 0.59259]	p < 0.01
T2 × NOUN = 1.16182	[1.07119, 1.27842]	p < 0.01
T2 × PROPN = 1.54294	[1.23952, 1.79054]	p < 0.01
T2 × VERB = 1.66008	[1.54706, 1.74707]	p < 0.01
T4 × ADJ = 1.12799	[0.94772, 1.29447]	p < 0.01
T4 × ADV = 1.07407	[0.94964, 1.19133]	p < 0.01
T4 × NOUN = 1.01376	[0.83692, 1.16253]	p < 0.01
T4 × PROPN = 1.52619	[1.26715, 1.80068]	p < 0.01
T4 × VERB = 1.44156	[1.31836, 1.55031]	p < 0.01

Table 2: Results of the multinominal logistic model concerning parts of speech.

of speech of the words. The first T3 of the T3T3 words tends to be realized as other tones the least in grammatical words (GRAM) and the most in proper nouns (PROPN). Fig. 6 illustrates the realization of the first T3 as a function of the parts of speech. The results indicate that the realization patterns of T3 in T3T3 words differ among different parts of speech categories. The first T3 of a T3T3 word is more likely to be realized as T1 when the word in question is an adjective (ADJ), an adverb (ADV), a noun (NOUN) a proper noun (PROPN), or a verb (VERB) than when it is a grammatical word/GRAM ($p < 0.01$ for all related comparisons; see details in Table 2). Similar trends are found for the first T3 of T3T3 words realized as T2 or T4 (cf. Tab. 2).

The ranking of the factors is presented in Fig. 7 according to the random forest model. Relative word frequency is observed to be the factor that contributed the most to the prediction of the realization of the first T3 in T3T3 words, followed by part of speech of the word containing the T3 in question (POS), part of speech of the preceding word (precPOS), part of speech of the (follPOS) and the lexical tone of the preceding syllable. Prosodic position, however, does not seem to contribute much to the prediction concerning the realization of the first T3 in T3T3 words, which is likely due to the limited levels (i.e. phrase initial vs. word initial) we were able to code and to the relative low occurrence in speech of phrase-initial T3s in T3T3 words.

4. DISCUSSION AND CONCLUSIONS

This study aimed to gain a better understanding of the acoustic realization of Low tone in a tone sandhi context (T3T3) in continuous Standard Chinese speech. To do so, the speech data was segmented twice in forced alignment mode using the LISN speech transcription system: once without and once with systematic tone variants. The output of the first set of alignments can be interpreted as the reference pronunciation of words. The output of the second set of alignments is related to the surface realization of words. This offered us an opportunity to compare directly between the underlying T3T3 sequences and their acoustic realizations, classified as one of the four lexical tones, according to the LISN transcription system. The system learns about the categorization of the four tones without supervision. In this paper, we focused on the tonal representations of the first T3 in the underlying T3T3 sequences of disyllabic words. The results showed that the realization of the first T3 varied as a function of a set of factors which included prosodic position, word frequency, tonal contexts, and part of speech. The results of the random forest algorithm suggest that the impact of these factors on T3 realization is unequal, as indexed by the ranking of the factors based on their contribution to the prediction of the T3 realizations. Among the factors investigated, the relative word frequency tends to contribute the most to predicting the realization of the first T3 in T3T3 words. It would be interesting to explore including neutral tone as a category, in addition to the four lexical tones. The application of T3 sandhi in bi-syllabic words is known to be non-optional and consistent in the literature. The varied realizations of T3 in the LISN output beg the question of how acoustic modeling systems such as that of LISN learn and classify the tones. For future research, it is important to compare native listeners' classification to LISN's output to gain further insights.

5. REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] M. Y. Chen, *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press, 2000, vol. 92.

[3] C. Shih, "Mandarin third tone sandhi and prosodic structure," *Linguistic Models*, vol. 20, pp. 81–124, 1997.

[4] J. Yuan and Y. Chen, "3rd tone sandhi in standard chinese: A corpus approach," *Journal of Chinese Linguistics*, vol. 42, no. 1, pp. 218–237, 2014.

[5] P. Fung, S. Huang, and D. Graff, "Hkust mandarin telephone speech/transcripts, part 1 ldc2005s15/t32," 2006.

[6] S. M. Strassel, C. Cieri, A. Cole, D. DiPersio, M. Liberman, X. Ma, M. Maamouri, and K. Maeda, "Integrated linguistic resources for language exploitation technologies." in *LREC*, 2006, pp. 185–190.

[7] A. Morris, B. Antonishek, X. Li, and S. Strassel, *HAVIC MED Progress Test–Videos, Metadata and Annotation*. Linguistic Data Consortium, University of Pennsylvania, 2019.

[8] K. Walker, C. Caruso, K. Maeda, D. DiPersio, and S. Strassel, "Gale phase 3 chinese broadcast news speech ldc2015s13," *Web Download. Philadelphia: Linguistic Data Consortium*, 2015.

[9] M. Glenn, H. Lee, S. Strassel, and K. Maeda, "Gale phase 3 chinese broadcast news transcripts ldc2015t25," *Web Download. Philadelphia: Linguistic Data Consortium*, 2015.

[10] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[11] M. Adda-Decker and L. Lamel, "The use of lexica in automatic speech recognition," in *Lexicon Development for Speech and Language Processing*. Springer, 2000, pp. 235–266.

[12] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin, "Mandarin broadcast news speech (hub4-ne)," *Linguistic Data Consortium*, 1998.

[13] L. Lamel, J.-L. Gauvain, V. B. Le, I. Oparin, and S. Meng, "Improved models for mandarin speech-to-text transcription," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4660–4663.

[14] L. Chen, L. Lamel, G. Adda, and J.-L. Gauvain, "Broadcast news transcription in mandarin," in *Sixth International Conference on Spoken Language Processing*, 2000.

[15] Q. Cai and M. Brysbaert, "Subtlex-ch: Chinese word and character frequencies based on film subtitles," *PloS one*, vol. 5, no. 6, p. e10729, 2010.

[16] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[17] J. Hadfield, "Mcmcglmm: Markov chain monte carlo methods for generalised linear mixed models," *Tutorial for MCMCglmm package in R*, vol. 125, 2010.

[18] Y. Xu, "Fundamental frequency peak delay in mandarin," *Phonetica*, vol. 58, no. 1-2, pp. 26–52, 2001.

L2 PHONOLOGICAL CONTRASTS ARE NOT NECESSARILY LOST: EVIDENCE FROM FRENCH LISTENERS DETECTING MANDARIN TONES

Volker Dellwo, Sandra Schwab, Rushen Shi

University of Zurich, University of Bern, University of Quebec at Montreal
volker.dellwo@uzh.ch, sandra.schwab@unibe.ch, shi.rushen@uqam.ca

ABSTRACT

Adult listeners are typically poor in perceiving non-native phonemic contrasts that are absent in their native language. French listeners, for example, whose native language lacks contrastive word stress, have a documented poor contrastive stress distinction ability. Here, we tested the performance of French listeners to detect tone contrasts, another phonological contrast that is not present in French. Forty-nine French listeners detected Mandarin level- and contour-tone contrasts in an odd-one-out task in either mono- or bisyllabic words. Performance was high and significantly above chance for all contrasts but contrasts in monosyllabic words were significantly better detected than in bisyllabic words. We conclude that the ability to detect non-native phonological contrasts depends highly on the nature of the contrast (here: stress-contrast vs. tone contrasts), a finding that might be important in the light of language evolution.

Keywords: L2 contrast perception, tone perception, stress perception, language evolution

1. INTRODUCTION

Early in development, infants possess an acute ability to perceive both native and non-native contrasts. The sensitivity to non-native contrasts, however, gradually declines towards the end of the first year of life while the perception of native contrasts continues ([1], [2]) or further strengthens ([3]). In French, for example, there is extensive evidence that adult listeners have very poor ability to distinguish contrastive stess in languages like Spanish or English, a phenomenon commonly known as *stressdeafness* ([4], [5]). The most likely reason given is that French phonology does not have a contrastive stress system, hence sensitive of listeners to such contrast is lost at some point during language acquisition.

Here, we tested how good French listeners are at distinguishing lexical tones, another phonological contrast that does not exist in French. In infants, contrastive tone detection performance in listeners from non-tonal languages varies. English- and French-learning infants discriminate rising versus falling contrast and the rising versus low contrast in Thai at 4 months of age, but the ability declines by 9 months ([6], [7]). Similarly, English-learning infants discriminate Cantonese high-rising and mid-level tones at 4 months of age but declined in this ability by 9 months ([8]). However, decline was not always observed. For example, French-learning infants aged 4 to 11 months showed no decline in discriminating the high-level versus falling contrast in Mandarin, while they showed a tendency of decline for another acoustically more similar tonal contrast, rising versus low-dipping ([9]). Similarly, Dutch-learning infants showed no decline in their discrimination of the Mandarin high-level versus falling tones during the first year of life, although a decline was observed when the acoustical difference of this contrast was artificially reduced ([10]). On the other hand, [11] showed that English-learning infants improved in their discrimination of a tonal contrast (high-level versus low-dipping) in Mandarin from 6 to 12 months of age, suggesting the possible contribution of growing general auditory processing capacities independent of the exposure to a pertinent phonemic system.

These results suggest that the perception of lexical tones by non-tonal language learners is less following the general decline of non-native contrasts replicated in typical non-native consonant and vowel perception tasks. Thus, it is unclear how the lack of lexical tone contrasts in a listener's native language affects its recognition ability. [12] tested 11-to-12-month-old French-learning infants on two types of tonal contrasts in Mandarin, level tone contrasts (high versus low, i.e., tone 1 vs tone 3; henceforth: T1, T3) and contour tone contrasts (rise versus fall, tone 2 vs tone 4; T2, T4). When tone contrasting syllables were the first syllable in bisyllabic pseudowords (second syllable was always T1), infants showed strong discrimination performance for both level and contour contrasts. However, when only the first syllable was presented to infants, performance dropped for contour tones and was at chance for level tones. The superior discrimination of both types of tonal contrasts when the bisyllabic forms were presented suggests that tonal differences between the first and the second syllable possibly offered salient acoustic cues. Importantly, the lack of lexical tones in infants' native language did not prevent them from perceiving tonal contrasts, at least under some conditions.

Here, we tested French native adults on the same stimuli that were used for the infants in [12] to understand whether the continued lack of experience with tone contrasts ultimately leads to a poor ability to perceive either level-tone (T1, T3) or contour-tone contrasts (T2, T4). If non-native language phonological contrasts decay with age, we expect French speakers to show poor discrimination performance for Mandarin tonal contrasts, for example similar to performances found for stress contrasts [5]. Additionally, we wanted to know (a) whether French adults – if able to detect Mandarin tones – rely on the second syllable or whether the tone in the first syllable is sufficient and (b) whether discrimination ability varies between contour and level contrasts. Given the larger degree of acoustic dynamic changes in contour contrasts, performance should be better compared to level contrasts that are mainly distinguished by tonal height differences.

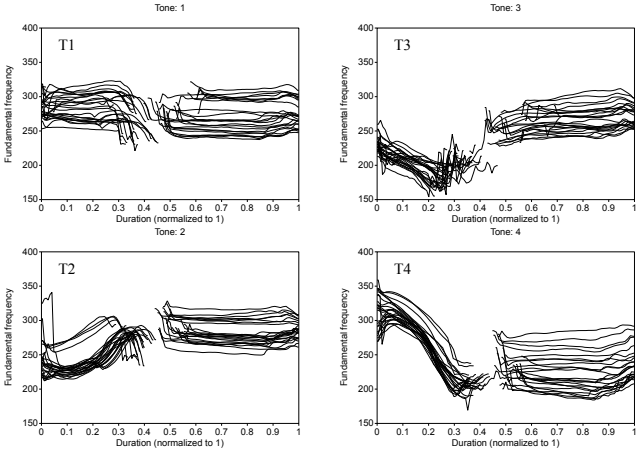


Figure 1: Pitch contours of all stimuli (with repetitions; N=96) separated by tones with level-tone examples (T1, T3; top) and contour tone examples (T2, T4; bottom). Duration has been normalized between 0 (onset of tone in first syllable) and 1 (offset of tone in last syllable).

2. METHOD

2.1. Participants

Forty-nine native listener of French from Switzerland and France (age range: 18 to 40, mean age = 27.27 years). No participant had knowledge of a tone language.

2.2. Material

Stimuli: Six Mandarin Chinese bisyllabic pseudowords (*fadi, quigu, pengbu, wangpi, xianda, yuanta*) were recorded from a female native speaker of Mandarin Chinese. Each word was produced with T1 (high tone), T2 (rising tone), T3 (low tone) and T4 (falling tone) on the first syllable and with T1 on the second syllable which was always /di:/ (see example in Fig. 1). Four repetitions of each of the 24 pseudowords were produced by the native speaker resulting in a total of 96 stimuli (6 words * 4 tones * 4 repetitions).

Trials: Trials of three stimuli were created in which one stimulus (odd) differed from the other two stimuli (even) regarding its tone. While the tone on the even stimuli was from the same category, the two stimuli were two randomly chosen realisations of the four repetitions. Combinations were either created with level- or contour tone contrasts, resulting in four combinations (Table 1). Two variants from different repetitions were created for the four tone combinations for each pseudoword resulting in 48 trials (4 tone combinations * 6 pseudowords * 2 variants). The three stimuli in a trial were separated by 500 ms of silence.

Trial sets: A second trial set was created that was identical to the first set with the exception that the second syllable of the bisyllabic pseudoword stimuli was deleted (from the onset of the stop gap of the second syllable /di:/). In total, the experimental material thus consisted of two trial sets, one containing the bisyllabic words (henceforth: *All*) and the other containing only the first syllable (henceforth: *1st*).

Contrasts	Odd	Even
Level (T1-T3)	1	3
Level (T1-T3)	3	1
Contour (T2-T4)	2	4
Contour (T2-T4)	4	2

Table 1: Combinations of tones in experimental trials with two level tone combinations (rows 1, 2) and two contour tone combinations (rows 3, 4).

2.3. Procedure

Participants were randomly split in two groups, one group (N=30) listening to the *All* trial set the other (N=19) to *1st* trial set. In an odd-one-out task, participants were presented each of the 48 trials of the respective trial set via high-quality headphones connected to a computer sound card in a quiet classroom at the University of Geneva. After each trial presentation, listeners task was to detect the position of the odd stimulus (the stimulus varyin in tone from the other two) in each trial and indicate the sequential position in the trial (1st, 2nd or 3rd) on corresponding buttons on the screen. Combinations of stimuli within each trial and the sequence of trials in each trial set were fully randomised for each individual listener. Experimentas were run in Praat

using scripts (www.praat.org) available from the first author.

2.4. Data analysis

Statistical analyses were carried out using R (version 3.4.0, [13]; R Development Core Team, 2016; lmerTest R package, [14]). A mixed-effects logistic regression model was run on the correct (odd detected)/incorrect (odd not deteceted) responses ([15]). The fixed part of the model was comprised of trial set (*All*, *1st*), tone contrast (level, contour), and their interaction. Both level contrasts and contour contrasts from Table 1 were respectively grouped together so that there were only two types of contrasts: level tone and contour tone contrasts. The random part of the model included random intercept for participants, and random slopes allowing for the effect of 'tone contrast' to differ across participants. The significance of the main effects and interaction was assessed with likelihood ratio tests that compared the model with the main effect or interaction to a model without it. For descriptive reasons, Fig. 2 presents percent correct, however, the analysis was performed on correct/incorrect responses.

In addition, participants' responses to each tonal contrast, i.e., T1 vs T3 (Level) and T2 vs T4 (Contour), in the *All* and *1st* contexts were compared with chance in exact binomial tests, allowing us to assess if the two types of tonal contrasts were perceived in these contexts by adults.

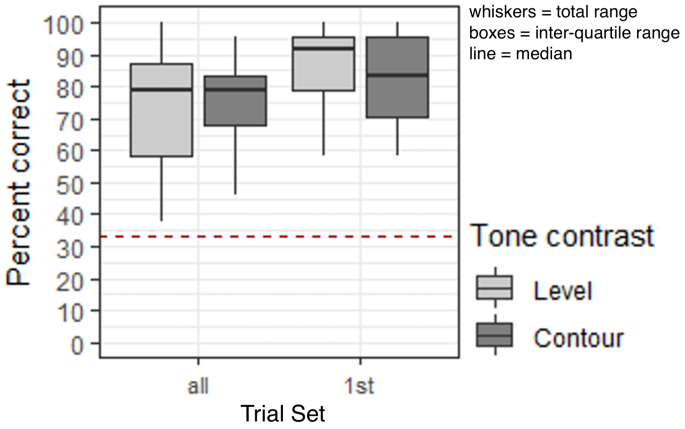


Figure 2: Percent correct as a function of *Trial Set* (All, first syllable) and 'tone contrast' (level, contour). The dashed line represents chance level (33%).

3. RESULTS

Fig. 2 presents the percent correct as a function of stimulus set and tone contrast. Performance for each tonal contrast in the two trial sets was overall high and each performance was significantly above the 33.3% chance level (All Level = 74%, exact binomial p [one-tailed] < 0.001; All Contour = 77%, exact binomial p [one-tailed] < 0.001; Level = 86%, exact binomial p [one-tailed] < 0.001; Contour = 82%, exact binomial p [one-tailed] < 0.001).

The interaction between 'trial set' and 'tone contrast' was close to significance (Tab. 2). For this reason we carried out simple effect post-hoc analyses (with Tukey correction) rather then interpretation of main effects. Simple effect analysis revealed that the difference between the two types of tonal contrasts (i.e., Level, Contour) was not significant in *All* (Level = 74%; Contour = 77%; $\beta = 0.172$, SE = 0.127, z = 1.351, p = .177) nor in *1st* (Level = 86%; Contour = 82%; $\beta = 0.327$, SE = 0.189, z = -1.728 , p = .084). Simple effect analyses also showed that the performance for contour tones did not significantly differ across the two stimulus sets (All-1st: $\beta = -0.433$, SE = 0.292, z = -1.486, p = .137). On the contrary, for level tones, the performance was better with the extracted first syllable (86%) than with the entire bisyllabic pseudoform (74%) ($\beta = 0.933$, SE = 0.297, z = 3.138, p = .002)

Predictors	*Log-Odds*	*CI*	*p*	*Likelihood ratio test*
(Intercept)	1.243	0.840 – 1.646	**<0.001**	
wordPart (ref = all)				$\chi^2(1) = 7.288$, p = .007
1st	0.952	0.280 – 1.623	**0.005**	
toneContrast (ref = level)				$\chi^2(1) = 0.176$, p = .675
Contour	0.070	-0.251 – 0.391	0.668	
wordPart x toneContrast				$\chi^2(1) = 3.411$, p = .065
1st:Contour	-0.531	-1.076 – 0.015	0.057	

Table 2: Summary of mixed-effects logistic regression model with likelihood ratio tests. Reference is 'incorrect' for the dependent variable, 'All' for wordPart and 'Level' for toneContrast.

3. DISCUSSION

We obtained a high detection rate of Mandarin tone-contrasts by French listeners in bisyllabic pseudowords (~75%) and significantly stronger performance when only the tone bearing syllable was presented (~84%). This suggests that the lack of exposure to tonal contrasts in their native language did not considerably impede French listener's detection performance of Mandarin Chinese tone contrasts. Thus non-native tone contrasts do not necessarily decay with age as they do with stress-contrasts in French ([4], [5]).

How much better are French listeners in detecting tone contrasts compared to stress contrasts? In a contrastive word-stress detection task using an odd-one-out design, [5] tested French listeners – amongst others – on their ability to detect stress differences in single talker and double talker condition. [5] found that using one talker and one intonation – identical to the present experiment – yielded ~63% correct for

stress-contrast detection. This result is drastically lower than any of the conditions tested in the present study and is strong evidence that for French listeners the ability to detect tone is generally better maintained compared to the ability to detect stress-contrasts.

These findings are interesting in the light of the evolution of phonological contrasts in language systems, since they suggest that some contrasts are acoustically more prone to be stable throughout the lifetime of a listener while others seem to be more subtle and more likely to be lost. Stress-contrasts, for example, rely on a detection of a combination of cues like duration, pitch and loudness, while fundamental frequence is dominant in tone detection (at least for the stimuli in our experiment; not necessarily for tone languages in general). It seems possible that contrasts which rely on an auditorily less complex interplay of acoustic parameters might be more stable throughout the lifetime. This might be a reason for the vast success of tone systems around the world, in particular for the adoption of tones across typologically non-related languages during tono-genisis in Southeast Asia [16], [17].

How robust are the findings of the present study? In [5] French listeners were tested under additional conditions: (a) there was a single and double speaker design and (b) intonation variability was controlled in having either one statement intonation or varying between a statement and a question intonation contour. Interestingly, [5] found that performance dropped significantly when a second talker was introduced and was at chance level with two intonation contours. We thus expect that performance in tone perception might significantly drop when the task gains in difficulty, e.g. by increasing the number of speakers, styles, phrase positions where the tone contrasts occur, etc. In fact, it seems obvious from our results that the use of a single speaker might have led to a disproportional advantage for adults. As can be seen in Fig. 1, f0 is systematically higher than 250 Hz in T1 and lower then 250 Hz in T3 for this particular speaker. This means that in an odd-one-out task, a tone higher than 250 Hz is easy to contrast against two others that are lower than 250 Hz. It would be interesting to test French listener performance when speakers in the stimulus trials vary (e.g. male and female with drastically different average f0). A native control group would be necessary for such a task as it is unclear whether native Mandarin listeners would overcome such strong difference in level tones in single syllable presentation.

What does the finding reveal about the tone discrimination ability in adults versus infants? The test procedures in [12] were very different from the odd-one-out design in the present experiment which must necessarily be the case. The infants in [12] were tested with a preferential head turning procedure, a method by which the infant is first habituated for 30 seconds with one target tone (in various bisyllables or monosyllables). Subsequently, by presenting a novel contrast versus new exemplars of the familiarized tone in separate trials, it can be tested whether the infant recognizes the earlier familiarized tone as distinct from the novel tone by turning their head to the respective sound source and yielding systematic differences in looking durations. Here, adult listeners were tested with an odd-one-out design in which they identified the odd stimulus that carried a different tone compared to two even stimuli in a three stimuli trial. This might also explain why adults were better with single syllable task compared to the double while for infants this was the other way round. Given the lack of comparative information in the single syllable condition, it seems somehow counter intuitive that adults revealed a higher performance. Again, this might very likely be related to the single speaker design in which adults might have gained a disproportional advantage in the odd-one-out task compared to the head-turning task in infants. For a preferential head-turning task, the nature of a simple tone height difference (i.e., citation tones in monosyllables) might not be interesting enough for the infant to change their attention levels. In contrast, the fact that infants responded strongly to contour tone differneces may reveal that a dynamic difference between the direction of F0 change might increase attention sufficiently. For adults, this might mean that not their ability to detect abstract Mandarin tone categories led to their high performance in the single syllable condition but the fact that the tone differences in the trials were auditorily well distinguishable. We take this as further evidence that general, non-language or speech specific factors contributed to the performance because adding more speech detail led to confusion and a drop in performance. This means that the auditory clarity between the tone contours in a single syllable outweighed additional language specific information. In native listeners, we would expect this effect to be reversed.

In summary, further experiments will be necessary by using multiple speakers, stimulus variability, listener groups with varying L2 background and control groups. This will be interesting to understand the maintenance of non-L1 phonolgical contrasts. It will have high impact on the understanding of the development of contrast perception throughout the human lifetime and the evolution of phonological systems of languages. It will also be interesting to compare tone perception mechnisms across speech and non-speech sounds [18].

5. REFERENCES

[1] J. F. Werker and R. C. Tees, “Cross-language speech perception: Evidence for perceptual reorganization during the first year of life,” *Infant Behav Dev*, vol. 7, no. 1, pp. 49–63, 1984, doi: 10.1016/S0163-6383(84)80022-3.

[2] L. Polka and J. F. Werker, “Developmental Changes in Perception of Nonnative Vowel Contrasts,” *J Exp Psychol Hum Percept Perform*, vol. 20, no. 2, pp. 421–435, 1994, doi: 10.1037/0096-1523.20.2.421.

[3] P. K. Kuhl, E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson, “Infants show a facilitation effect for native language phonetic perception between 6 and 12 months,” *Dev Sci*, vol. 9, no. 2, pp. F13–F21, Mar. 2006, doi: 10.1111/J.1467-7687.2006.00468.X.

[4] E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler, “A Destressing ‘“Deafness”’ in French?,” 1997.

[5] S. Schwab and V. Dellwo, “Intonation and talker variability in the discrimination of Spanish lexical stress contrasts by Spanish, German and French listeners,” *J Acoust Soc Am*, vol. 142, no. 4, pp. 2419–2429, 2017, doi: 10.1121/1.5008849.

[6] K. Mattock and D. Burnham, “Chinese and English infants’ tone perception: Evidence for perceptual reorganization,” *Infancy*, vol. 10, no. 3, pp. 241–265, 2006, doi: 10.1207/S15327078IN1003_3.

[7] K. Mattock, M. Molnar, L. Polka, and D. Burnham, “The developmental course of lexical tone perception in the first year of life,” *Cognition*, vol. 106, no. 3, pp. 1367–1381, Mar. 2008, doi: 10.1016/J.COGNITION.2007.07.002.

[8] H. H. Yeung, K. H. Chen, and J. F. Werker, “When does native language input affect phonetic perception? The precocious case of lexical tone,” *J Mem Lang*, vol. 68, no. 2, pp. 123–139, Feb. 2013, doi: 10.1016/J.JML.2012.09.004.

[9] R. Shi, E. Santos, J. Gao, and A. Li, “Perception of Similar and Dissimilar Lexical Tones by Non-Tone-Learning Infants,” *Infancy*, vol. 22, no. 6, pp. 790–800, Nov. 2017, doi: 10.1111/INFA.12191.

[10] L. Liu and R. Kager, “Perception of tones by infants learning a non-tone language,” *Cognition*, vol. 133, no. 2, pp. 385–394, Nov. 2014, doi: 10.1016/J.COGNITION.2014.06.004.

[11] F. M. Tsao, “Perceptual improvement of lexical tones in infants: Effects of tone language experience,” *Front Psychol*, vol. 8, no. APR, p. 558, Apr. 2017, doi: 10.3389/FPSYG.2017.00558/BIBTEX.

[12] E. Santos and R. Shi, “Perception of level tones and contour tones by French-learning infants. ,” in *14th International Congress for the Study of Child Language (IASCL)*, Lyon, 2017.

[13] “R: a language and environment for statistical computing.” https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing (accessed Dec. 17, 2022).

[14] “Tests in Linear Mixed Effects Models [R package lmerTest version 3.1-3],” Oct. 2020, Accessed: Dec. 18, 2022. [Online]. Available: https://CRAN.R-project.org/package=lmerTest

[15] R. H. Baayen, D. J. Davidson, and D. M. Bates, “Mixed-effects modeling with crossed random effects for subjects and items,” *J Mem Lang*, vol. 59, no. 4, pp. 390–412, Nov. 2008, doi: 10.1016/J.JML.2007.12.005.

[16] J. A. Matisoff, “Tonogenesis in Southeast Asia,” *Consonant Types and Tone. Southern California Occasional Papers in Linguistics*, vol. 1. pp. 73–95, 1973.

[17] A. S. Abramson, “The plausibility of phonetic explanations of tonogenesis,” in *From traditional phonology to modern speech processing: Festschrift for Professor Wu Zongji’s 95th birthday*, G. F. Fant Hiroya; Cao, Jianfen; Xu, Yi, Ed., Beijing: Foreign Language Teaching and Research Press, 2004, pp. 17–29.

[18] T. Bent, A. R. Bradlow, and B. A. Wright, “The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds.,” *J Exp Psychol Hum Percept Perform*, vol. 32, no. 1, pp. 97–103, 2006, doi: 10.1037/0096-1523.32.1.97.

TONE IS NOT PREDOMINANT: TONE IS NOT PRIMORDIAL

Ian Maddieson

University of New Mexico and University of California, Berkeley
ianm@berkeley.edu

ABSTRACT

Human language was originally tonal according to some theories, especially those suggesting a common origin of language and music. Support is sought in assertions that most modern spoken languages are tonal, and that loss of tone is more frequent than its acquisition. A detailed effort to compile data from all language families containing 20 or more "living" languages in *Ethnologue*, plus a sample of isolated and unclassified languages, was carried out to determine how many are tonal. Results compiled from (near-) exhaustive surveys, individual language reports and sample-based estimates show tonal languages are not in the majority today. Moreover, diachronic studies show that evidence for innovation of tone is far more frequent than for original tone, and loss of tone is quite rare, casting doubt on a 'primordial tone' hypothesis. Pitch was undoubtedly important in human proto-language, but seems more likely to have had phrasal and/or discourse functions.

Keywords: Tone, language origins, language and music, tonogenesis, tonoexodus.

1. INTRODUCTION

The idea that human language at its origin was tonal, specifically having lexical contrasts involving tone, has been frequently proposed, particularly by those who envisage language and music as having a shared origin. It was espoused (in pre-evolutionary terms) by French Enlightenment philosophers Condillac and Rousseau. Jespersen [1] proposed that "primitive language [was] sung rather than spoken." More recently, Brown [1, 2, 3], Mithen [4], Scherer [5] and others have renewed the suggestion of an original 'musilanguage' from which language(s) with tone contrasts emerge(s). Brown finds support in the claim that (1) "*the majority of spoken languages in the world today are lexical-tonal*" and asserts that (2) *"lexical tone is the ancestral state of spoken language, and the loss of tone is a derived feature of non-tonal languages, rather than the reverse progression.*" Both these statements are contestable. (1) appears to be based on the suggestion by Yip [6] that "*as many as 60-70 per cent of the world's languages may be tonal.*". Yip called this "a very rough estimate" but it has been widely repeated in popular and scientific literature without this qualification. A more precise estimate of how many modern languages are tonal is provided in the present paper. In addition, an analysis of loss or acquisition of tone across all major language families is reported to see if (2) is plausible. If (1) and (2) are false, it doesn't mean music and language couldn't have a common origin but shows that these particular arguments cannot be used to support the idea.

2. NUMBER OF TONAL LANGUAGES

The *Ethnologue* [7] estimates there are 7151 known currently spoken languages, though the total of languages listed in real named families is 7281; this number includes well-documented recently spoken languages, but excludes non-families such as creoles, pidgins and 'mixed languages' as well as ancient languages. This latter number provides a pragmatic basis from which to start an estimate of what proportion of languages are tonal, despite the familiar problems with how to define a language. Surveys were carried out of all language families with 20 or more members listed in *Ethnologue*, a total of 6520 languages, 89.5% of the listed individual languages.

Table 1. Select survey volumes consulted (not listed in References but applicable to relevant families).

Region/ *Family*	**Publisher**	**Editor(s)/ Author(s)**	**Year**
Africa	DeGruyter	Güldemann	2018
Amazonia	Cambridge	Dixon, Aikhenwald	1999
Andes	Cambridge	Adelaar, Muysken	2004
Australia	Cambridge	Dixon	1970
Austro-Asiatic	Brill	Jenny, Sidwell	2015
Dravidian	Cambridge	Krishnamurti	2003
Indo-European	Routledge	Ramat, Ramat	1998
Papuan	DeGruyter	Palmer	2018
Sino-Tibetan	Routledge	Thurgood, LaPolla	2003
South America	DeGruyter	Campbell, Grondona	2012
Turkic	Routledge	Johanson, Csato	1998
Uralic	Routledge	Abondalo	1998

Information was aggregated from survey volumes, on-line databases, book chapters, journal articles, and in a few cases personal communications. Space limits preclude an exhaustive bibliography, but some of the major survey volumes consulted are listed in Table 1.

For some language families, explicit discussion of the distribution of tones is available, as for Eyak-Athabaskan. Krauss [8] listed and mapped the Athabaskan languages known at that time to be tonal. Additional languages recognized in the Ethnologue and the non-Athabakan languages of the extended family can then be added based on language-specific studies; Eyak is non-tonal and Tlingit (standard) is tonal. Out of 44 languages of this family, it appears that 28 should be considered tonal.

Table 2. Number of tonal languages by family

Family	#lgs	# tonal	basis
Niger-Congo	1551	1458	E
Austronesian	1257	126	E
Trans-New Guinea	481	167	E
Sino-Tibetan	457	367	E
Indo-European	446	51	E
Afro-Asiatic	384	252	E
Australian	380	0	S
Nilo-Saharan	207	180	E
Oto-Manguean	178	178	S
Austro-Asiatic	167	39	E
Kra-Dai	91	91	S
Dravidian	86	0	S
Tupian	76	5	S
Uto-Aztecan	61	2	S
Toricelli	57	4	S
Arawakan	55	9	E
Sepik	54	0	S
Quechuan	44	0	S
Eyak-Athabaskan	44	28	S
Algic	41	5	S
Turkic	41	0	S
Hmong-Mien	39	39	S
Uralic	38	2	S
Ramu-Lower Sepik	32	2	E
Cariban	31	0	S
Mayan	31	2	S
Nakh-Dagestanian	31	2	S
Panoan	26	3	S
Salish	26	0	S
Tor-Kwerba	23	0	E
West Papuan	23	4	E
S-Central Papuan (Yam)	22	0	E
Chibchan	20	8	S
Lakes Plains	20	20	S

For other families, an estimate of the percentage of tonal languages is made based on a sample, which is then applied to the number of languages in *Ethnologue*. For example, SAPhon [9] lists 45 Arawakan languages, 7 of which are reported as tonal, that is, 15.56%. This percentage, applied to the 55 languages of the corresponding family (called 'Maipurean') in *Ethnologue*, gives an estimate of 8.5 tonal languages, rounded up to 9. Throughout, an overestimation bias in the number of tonal languages is maintained. For example, the tonal status of 18 of the 20 Chibchan languages is known, with 7 being tonal (Teribe only vestigially). So the percent of tone languages is calculated as 7/18 = 3.89%. For a family with 20 languages this predicts 7.78 (rounded to 8) tone languages, though the two undocumented varieties of Tunebo are almost certainly non-tonal. Indo-European languages such as Livonian and Slovenian are classified as tonal even though alternative accentual analyses are possible; any language identified as having a "melodic accent" is classified as tonal.

The results of this exercise are shown in Table 2, which reports the number of languages by family, and the estimated number that are tonal. The final column indicates whether the number of tonal languages is estimated from a sample ('E') or is based on an explicit statement or comprehensive survey ('S'). For the larger families a sampling strategy usually had to be used as many languages lack good descriptions, and the same was true for most of the Papuan groups.

Of the 6520 languages in the families listed in Table 2, the indications are that 3044, or 46.7 %, are tonal. This is well below the estimate of 60-70% often cited and suggests that tone languages form a minority of modern languages. Moreover, given the way that the sampling was conducted this figure is more likely to be an over-estimate than and under-estimate of the number of tone languages. Although they make up a substantial proportion of current and recently spoken languages, tone languages are not so predominant that tonality should be taken as the default condition of any original human language, however that is envisaged.

3. LOSS OR GAIN OF TONE

Brown proposes that "*the loss of tone is a derived feature of non-tonal languages*". We don't know if early human languages were tonal, but we can perhaps better understand the likelihood of Brown's claim being correct by examining what we know about the evolution of tone in the reconstructible past. For most of the language families on which serious comparative work has been carried out, there is no reason to reconstruct tone as an original feature of the

phonology of the proto-language. For 18 of the 34 language families listed in Table 2, there is no indication that (in descending order of size) Austronesian, Australian, Dravidian, Tupian, Uto-Aztec, Arawakan, Algic, Turkic, Quechuan, Uralic, Cariban, Mayan, Nakh-Dagestanian, Panoan, Salish, Tor-Kwerba, West Papuan or South-Central Papuan require consideration of the possibility of reconstructing tonal contrasts, since all or the vast majority of the languages they contain have no tone and they often have obvious phenomena of accentuation as the main dimension of lexical prosody. Substantial historical-comparative work has been carried out in all these groups except for the last three, where basic documentation is still the priority.

In Austronesian, the largest predominantly non-tonal language family, there is debate as to whether the accent position in Proto-Austronesian was predictable or had a limited contrastive role, but absolute agreement on the absence of original tonality. This does not mean that no Austronesian language is tonal. A spectacular exception is Utsat (also known as Hainanese Cham) which has five tonal distinctions [10], stemming from earlier properties of consonantal onsets and codas. Simpler tone systems are found in several other Austronesian languages, but these are clearly local and independent developments.

None of the modern Australian or Dravidian languages are tonal and there is no reason to suggest that ancestral forms ever had tones. The vast majority of Uto-Aztec languages are non-tonal [11], although some Hopi dialects have tones recently developed from voicing contrast in final consonants [12], and Northern Tepehuan has reinterpreted a stress accent as a high tone. The Algic family includes some languages with a melodic accent, e.g. Blackfoot, but no tone is reconstructed for earlier stages [13, 14]. In South America, the Arawak, Tupi [15] and Carib families each have only a handful of tonal languages and Quechuan has none [16, 17]. The Turkic and Uralic families of Eurasia also include few or no tonal languages. For these last six language families, no reconstruction has proposed original tonal contrasts in the proto-language.

Indo-European presents a more complex case. Modern languages are mostly non-tonal and have either a contrastive accent, as in Russian, or no role for the lexical accent, as in French. However, from contemporary sources and later commentaries, there are indications of contrasting pitch height patterns in Ancient Greek [18] and Sanskrit [19]. A distinction between high and descending "accents" as well as a low tone on unstressed syllables is shared. Modern Baltic and South Slavic languages also have contrasts in pitch that some Indo-Europeanists consider to be related to these ancient patterns, leading to the assumption that the proto-language possessed contrasting 'melodic accents'. However, others [20] argue that quasi-tonal contrasts in the Baltic languages are of more recent origin, not inherited. Others remain undecided about the tone of Proto-Indo-European [21], or suggest that the accentual system was simpler (perhaps with a high tone as a normal realization of stress). That stress in Proto-Indo-European had a very important role is evident from the patterns of ablaut in which certain vowels are suppressed in unstressed syllables. This is not incompatible with tonality, as demonstrated by Thai and other 'sesquisyllabic' languagesin Asia and Mesoamerica, but it suggests that Indo-European was not originally a canonical tonal language.

Certain other families in Table 2 have a substantial proportion of tonal languages, including (in descending order of size) Sino-Tibetan, Afro-Asiatic, Nilo-Saharan, Oto-Manguean, Austro-Asiatic, Kra-Dai, Eyak-Athabaskan, and Hmong-Mien. In the majority of these families, it is clear that tone is not a characteristic of the reconstructible proto-language.

The Sino-Tibetan family is of particular interest. All Sinitic languages are tonal and Mandarin Chinese is often considered the prototype of a tonal language, for example by Rousseau and Brown [3]. However, it is in fact atypical both in the number and complexity of its tonal contrasts and in the fact that the tones of adjacent syllables are largely independent of each other. But tone is not original in this branch [22, 23]. As Baxter & Sagart [24] write, "There were no tones in Old Chinese." The four categories of traditional Middle Chinese tones: shǎngshēng (rising tone), qùshēng (starting tone), rùshēng (input tone) and píngshēng (even tone) derive respectively from forms with a final -*ʔ, a final -*s, a final voiceless oral stop, or a final sonorous consonant or vowel. In the much larger Tibeto-Burman branch, tones appear in some languages, but their origins are clearly independent of each other.

The second largest family, Afro-Asiatic, has a majority of languages with contrasting tones. Ehret [25] suggests that this family was originally tonal but has lost this characteristic in some branches. His argument seems to be based primarily on a principle of "majority rule." He writes: "*Phonemic tone is a widespread feature of Afroasiatic, appearing regularly in the languages of the Omotic, Chadic, and Southern and Eastern Cushitic divisions of the family. Only the Boreafrasian subgroup (Semitic, Berber, and Egyptian ...) has entirely deleted tone*." This family, much like Indo-European, has several relatively well-defined branches, of which 3 of the 6, as Ehret notes, show no signs of tone. In the Chadic branch, there is strong evidence that consonant types

influence tonal development [26], as is also the case for some Omotic languages [27]. There are no believable attempts to make reconstructions at the proto-Afro-Asiatic level. Ehret's reconstruction is not convincing because it relies on only a few languages and the proto-tone seems to be assumed rather than demonstrated. The independent development of tones from consonantal environments in different branches of Afro-Asiatic cannot be ruled out.

Even less is known about early forms of proto-Nilo-Saharan. Ehret [28] suggests that the proto-language had 3 tones, but again the presentation of the data is not convincing, partly because it seems to project the patterns of a single individual language, Uduk, onto the proto-language without reconstruction of intermediate groups.

In the Oto-Manguean family, Rensch [29] finds that the tone must be reconstructed as original; All daughter languages are tonal. As for Austro-Asiatic, Kra-Dai (= Tai-Kadai), Eyak-Athabaskan (= Na-Dené) and Hmong-Mien, it is clear that tonal contrasts, where they exist, developed from the properties of the consonantal environment. There are striking similarities with the evolution of tone in Chinese in the way tones originate from Kra-Dai, Hmong-Mien and some Austro-Asiatic languages [30, 31, 32]. In Eyak-Athabaskan, tonal contrasts arise from consonantal laryngeal constriction, and take two distinct forms in different languages [8].

Of the seven families from the Papuan region shown in Table 2, the Toricelli and Sepik languages are not reported as tonal. A considerable proportion of the languages of the trans-New Guinean family are tonal [34] but geography plays a striking role in their distribution. Donohoe [35] notes that languages on or adjacent to the Central Cordillera of the island of New Guinea are much more likely to be tonal than those more distant. There is a clear regional pattern here, and linguistic contact is probably a factor, but if it is a process of spreading tone in the highlands or a loss of tone in the lowlands, this cannot be established at the moment and may never be. Only the Lakes Plains family, which is 100% tonal, argues for tone as original, but the possibility of eventual wider comparison may change this impression.

This leaves the Niger-Congo family, by most accounts, the language family with the largest number of members. There is currently no serious attempt to recontruct Proto-Niger-Congo, and there are many challenges to the proposition that it is in fact a coherent language family. Several subgroups have been the subject of comparative work, in particular the Bantu subgroup, as well as smaller divisions such as Edoid [36] and Jukunoid [37]. The presence of a binary tone contrast in proto-Bantu is well established [38], and other reconstructions of subgroups also suggest that the tone is original within each subgroup. However, correlations between different groups are not established. Nevertheless, Niger-Congo has a high probability of being tonal originally. Non-tonal languages in this family, such as Swahili, Wolof, Ndut, Bisa and Koromfe, would therefore exemplify cases of loss of tone, some clearly due to contact.

The only language family in which tone is reconstructed as original but which now contains a majority of non-tonal languages is Chibchan [39]. Brown's claim that loss of tone is the common pattern is not supported by what we know of the loss or acquisition of tone. We know of far more cases in which languages acquire tone than cases where tone is lost.

4. SUMMARY

We find that less than half of modern languages are tonal and that of the 34 families examined, only four have a high probability of having original tone (Niger-Congo, Oto-Manguean, Chibchan and Lakes Plains), although Afro-Asiatic and Nilo-Saharan might have also been tonal at their origin. Proto-language, or rather pre-language, is generally thought of as consisting of holistic statements, rather than analyzable structures. Rousseau and Brown agree with this view. The question is what traits would have been retained from this stage when a segmentable lexicon and syntactic organization began to emerge. Human language in its earliest form would undoubtedly have used the variation in pitch of the voice. However, it seems more likely that this would have taken a form closer to the patterns described as intonation in modern languages, such as the frequent use of a global or final rise to signal a question, a descending sequence to signal an affirmation, or the stepped tones that are common in the call of a name. In modern languages, such patterns coexist with lexical tone, even in strongly tonal languages such as Mandarin Chinese and Yoruba where they overlap with tonal melody [40]. Intonation is a truly universal feature of modern spoken languages. It is much more likely that the legacy of a hypothetical "musilanguage" stage of protocommunication is intonation, not lexical tone. Tone probably arrived late in the diversity of human languages.

5. REFERENCES

[1] Brown, S. 2000. The 'musilanguage' model of music evolution. In N. L. Wallin, B. Merker, S. Brown, eds, *The Origins of Music*. MIT Press, 271-300.

[2] Brown, S. 2001. Are music and language counterparts? In *The Biological Foundations of*

Music, eds R. Zatorre, I. Peretz. New York Academy of Sciences, 372–374.

[3] Brown, S. 2017. A joint prosodic origin of language and music. *Frontiers in Psychology*. 8:1894. doi: 10.3389/fpsyg.2017.01894

[4] Mithen, S. 2005. *The Singing Neandertals.* Harvard University Press.

[5] Scherer, K. R. 1991. Emotion expression in speech and music. In *Music, Language, Speech, and Brain,* ed. J. Sundberg, L. Nord, & R. Carlson: 146-56. Macmillan

[6] Yip, M. 2002. *Tone*. Cambridge University Press.

[7] Eberhard, D. M et al. 2022. *Ethnologue: Languages of the World, 25th edition*. SIL International.

[8] Krauss, M. 2005 [1979]. Athabaskan tone. In S. Hargus & K. Rice, eds, *Athabaskan Prosody.* Amsterdam**:** John Benjamins, 55-136.

[9] Michael, L. et al. 2015. *SAPhon: South American Phonological Inventory Database v2.1.0.* Survey of California and Other Indian Languages. U. of California, Berkeley.

[10] Maddieson, I.; K-F. Pang. 1993. Tone in Utsat. In *Tonality in Austronesian Languages*, ed K. Gregerson & J. Edmondson: U of Hawaii: 75-89.

[11] Heath, J. 1977. Uto-Aztecan Morphophonemics. *IJAL* 43: 27–36

[12] Manaster-Ramer, A. 1986. Genesis of Hopi tones. *IJAL* 52: 154–160

[13] Proulx, p. 1984. Proto-Algic I: Phonological sketch. *IJAL* 50: 165-207.

[14] Goddard, I. 1979. Algonquian Comparative. In L. Campbell, M. Mithun, eds, *The Languages of Native America.* U. of Texas Press, 70–132

[15] Souza Mello, A. A. 2000. *Estudo Histórico da Família Lingüística Tupi-Guarani: Aspectos Fonológicos e Lexicais.* Ph. D., U. Federal de Santa Catarina.

[16] Parker, G. J. 1989. Comparative Quechua phonology and grammar II: Proto-Quechua phonology and morphology. *Working Papers in Linguistics, U. of Hawai'i* 2: 123-147.

[17] Landerman, P. N. 1991. Quechua Dialects and their Classification. Ph.D. UCLA

[18] Allen, W. S. 1953. *Phonetics in Ancient India*. Oxford U.P.

[19] Allen, W. S. 1987. *Vox Graeca*, 3rd edition. Cambridge U. P.

[20] Kortlandt, F. 2010. Glottalization and tonogenesis in Athabaskan, Balto-Slavic and Germanic. *Suvremena Lingvistika* 36: 37-44.

[21] Beekes, R. S. P. 1995. Comparative Indo-European Linguistics: An Introduction. John Benjamins.

[22] Haudricourt, A-G. 1954b. How to reconstruct archaic Chinese. *Word* 10: 351–64.

[23] Baxter, W. H. 1992. *A Handbook of Old Chinese Phonology*. Mouton de Gruyter

[24] Baxter, W. H., L. Sagart. 2014. *Old Chinese: A New Reconstruction*. Oxford U.P.

[25] Ehret, C. 1995. Reconstructing Proto-Afroasiatic (Proto-Afrasian): Vowels, Tone, Consonants, and Vocabulary. U. of California Press.

[26] Wolff, E. 1987. Consonant-tone interference in Chadic and its implications for a theory of tonogenesis in Afroasiatic. In D. Barreteau, ed, *Langues et cultures dans le bassin du Lac Tchad*. ORSTOM, 193-216

[27] Tesfaye, A, Wedekind, K. 1994. Aspects of Omotic tonogenesis: Shinasha. *J. of Ethiopian Studies* 27: 1-19.

[28] Ehret, C. 2001. A historical-comparative reconstruction of Nilo-Saharan. R. Köppe Verlag.

.[29] Rensch, C. A. 1976. *Comparative Otomanguean Phonology*. Indiana U.

[30] Norquest, P. K. 2007. *A Phonological Reconstruction of Proto-Hlai.* Ph. D., U. of Arizona.

[31] Ratliff, M. 2010. *Hmong-Mien Language History*. Pacific Linguistics, ANU.

[32] Haudricourt, A-G. 1954a. From the origin of the tones of Vietnamese. *Asian Journal* 242: 69-82

[34] Pawley, A., 2012. How reconstructable is proto Trans New Guinea? Problems, progress, prospects. In H. Hammarström, W. van den Heuvel, eds., *History, Contact and Classification of Papuan Languages*. Linguistic Society of Papua New Guinea: 88–164

[35] Donohue, M. 1997. Tone systems in New Guinea. *Linguistic Typology* 1:347–386.

[36] Elugbe, B. O. 1989. *Comparative Edoid: Phonology and Lexicon.* U. of Port Harcourt Press.

[37] Shimizu, K. 1980. *Comparative Jukunoid.* Vienna: Afro-Pub.

[38] Meeusssen, A. E. 1967. Bantu grammatical reconstructions. *Africana Linguistica* 3: 79-121

[39] Costenla Umaña, A. 1981. Comparative Chibchan Phonology. Ph. D. thesis, U. of Pennsylvania.

[40] Xu, Y. 1997. Contextual tonal variations in Mandarin. *J. of Phonetics* 25: 61-83.

Cue weighting in Mandarin tone perception: A comparison between native speakers and learners of Mandarin

Song Jiang

University of Toronto
soong.jiang@mail.utoronto.ca

ABSTRACT

Some studies have shown that advanced learners of Mandarin perceive Mandarin tones more categorically compared to non-native speakers without prior experience. However, these studies have mainly focused on one single cue.

In the current study, the author investigated the interaction between two acoustic-phonetic cues (turning point position (TP) and f0 at the end (End f0)) during perception between Tone 2 and Tone 3 and how English listeners of Mandarin adapt their perceptual strategies to distinguish categories that do not exist in their L1. The results show that native Mandarin listeners tend to use a TP-dominant strategy, whereas non-native listeners rely less on TP. No significant difference was found in cue weighting between the learners and naïve English listeners. We suggest that native Mandarin listeners are able to exploit TP to filter out the uncritical portion, whereas learners of Mandarin focus more on the canonical textbook forms of Mandarin tones.

Keywords: Mandarin tones, contour shape, turning point position, cue weighting, L2 acquisition.

1. INTRODUCTION

In natural speech, phonetic categories are usually defined with multiple acoustic cues [1]. During speech perception, listeners need to integrate acoustic information across multiple dimensions for successful speech categorization. For example, Lisker [2] identified 16 acoustic properties that may play a role in the perception of English voicing. Nevertheless, these acoustic cues do not contribute to speech categorization equivalently. Instead, listeners rely on some cues more than others. This phenomenon has been referred to as cue weighting [3].

There has been controversy over the ability of L2 learners to adapt their perceptual cue weighting in response to different target languages [4, 5, 6]. These studies mainly focused on the acoustic properties that exist in the L1 of the learners (*e.g.*, VOT, formant information, duration). Mandarin tones perception is a good case for us to investigate if learners can use some acoustic cues in Mandarin tone perception that are non-contrastive in non-tonal languages (*e.g.*, f0 curve shape [7]). For example, distinguishing between Tone 2 (a rising tone) and Tone 3 (a low dipping tone) has been considered a more difficult task for both native and non-native speakers due to their similar concave contours [8, 9, 10, 11]. Multiple cues have been found to serve as important cues to the native perception of Tone 2 and Tone 3, such as the temporal position of the f0 minimum (a.k.a. turning point position or TP) [10, 11, 12] and the f0 at the vowel offset (End f0) [13]. A later TP and a lower End f0 serve as a cue for Tone 3. It has also been reported that advanced learners of Mandarin significantly outperformed non-native speakers without prior experience in the identification and discrimination of Tone 2-Tone 3 [8, 9]. What remains unclear is whether advanced learners of Mandarin achieve native-like categorical perception by modulating their cue-weighting strategies.

In the current study, we aim to investigate the relative weights of different cues in Mandarin tone perception. A comparison is made between native Mandarin speakers, English learners of Mandarin, and naïve English speakers. This study focuses specifically on two acoustic-phonetic cues, TP and End f0. Here arise our research questions:

1. What are the cue weights of TP and End f0 for the three groups?
2. How different is the cue weighting between native and non-native speakers?
3. Will L2 learning experience influence the cue weighting in Mandarin tone perception?

TP, which implies the pitch contour shape, is not an acoustic feature that listeners of non-tonal languages can use productively. Hence, I hypothesize that native Mandarin speakers may rely more on TP than non-native speakers. Learners of Mandarin may have acquired the cue weighting that native Mandarin speakers use in tone perception, or learners may use other cue weightings to achieve native-like categorical perception.

2. METHODS

2.1. Participants

One male native Mandarin speaker from North China (Liaoning Province) was recruited for stimuli

recording. Seventy-two participants were recruited for the perception experiment through the Linguistics Participant Pool at the University of Toronto. It consisted of three groups:

- MAN: native Mandarin speakers (n = 26)
- L2M: English learners of Mandarin (n = 26)
- ENG: native English speakers with no previous exposure to Chinese (n = 20)

All participants were first asked to complete a questionnaire on language background before the task. The L2M participants were asked to report their years of learning. Their years of learning averaged 7.5 years (range: 0.125 years – 20 years). None of the ENG participants had learned any tonal language as their second language. No speech or hearing impairments were self-reported.

2.2. Stimuli

Ten repetitions of the syllables /pa, pu, pi, fa, fu, a, u, i/ with Tone 2 and Tone 3 were recorded in the University of Toronto Phonetics Lab. The recording was done in a sound-attenuated booth using an AudioTechnica AT831b microphone and a SoundDevices 722 digital audio recorder, with a 16-bit depth and a sampling rate of 48 kHz.

Figure 1 illustrates the contours along the TP and End f0 continua and the distribution of the stimuli. The stimuli were designed to vary in TP along a continuum from the second to the eighth deciles. The duration was constant at 375 msec, the average of all recording tokens. The End f0 also varies in 7 Hz steps in the range of 144 Hz ~ 200 Hz, which is similar to the range of the End f0 of the recording.

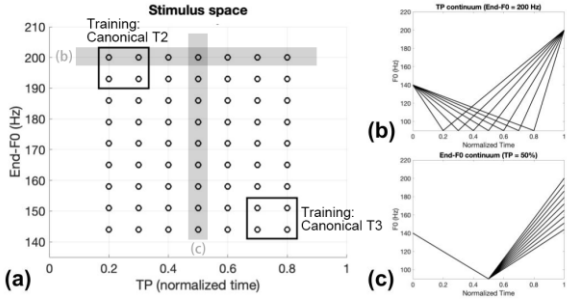


Figure 1: The diagrams of (a) the TP × End f0 stimulus space; (b) a TP continuum (End f0 = 200 Hz), and (c) an End f0 continuum (TP = 50%).

Acoustic measurements for the recordings were performed to determine the f0 range and vowel duration. Tone 2 and Tone 3 have close f0 values at the onset, around 140 Hz. The Onset f0 was set at 140 Hz. According to [12], 50 Hz of the difference between the onset and the minimum (Δf0) results in the most ambiguous perception along the TP continuum. The Δf0 was therefore set at 50 Hz – that is, the minimum is 90 Hz. No onset effect on f0 was found in our data. The onsetless /i/ was chosen as the target syllable because it was also used in previous studies [8, 11]. The target syllable /i/ was synthesized with different f0 contours covarying in TP and End f0, resulting in a set of stimuli spanning a two-dimensional perception space. All stimuli were created by using f0 synchronous overlap and add (PSOLA) method in PRAAT [14]. In total, 63 stimuli were created.

2.3. Procedure

A 2AFC identification task was implemented online using Gorilla [15]. The identification task included two phases: a training phase and a testing phase. In the training phase, two repetitions of "canonical" Tone 2 and Tone 3 stimuli were presented randomly in a single block. As illustrated in Figure 1(a), the four stimuli in the upper left box were the canonical forms of Tone 2, and the four stimuli in the lower right box were the canonical forms of Tone 3. The stimuli in isolation were presented to participants. Then, they were asked to identify the tone and click on the corresponding button on the screen. Depending on their native language, the question was provided in either English ("*Please identify the tone of the syllable you just heard. Did it have Tone 2 or Tone 3?*") or Mandarin ("*请问您刚刚听到的字是二声? 还是三声?*"). Apart from the names of the tones, the diacritics of tones in *Pinyin* were also provided as visual aids to facilitate responses. Feedback was provided during this phase.

Following training, the testing phase began. Ten repetitions of the entire stimulus space in Figure 1(a) were presented to participants. There were ten blocks in this phase. Each repetition (63 tokens) was presented and randomized in a separate block. A short break was provided between blocks. Participants were asked to do the same identification task as in the training phase, but no feedback was provided.

In total, there were 16 trials (2 repetitions × 8 stimuli) in the training phase and 630 trials (10 repetitions × 63 stimuli) in the testing phase. Depending on their first language, participants were provided written instructions in English or Mandarin.

2.4. Statistical analysis

To predict the model of identification results, we fitted generalized logistic models (GLM) to each individual [16, 17] and extended the one-dimensional model to a multi-dimensional stimulus space.[1] This allows us to estimate the relationship between the identification score (p, the probability of Tone 2 responses) and the two predictors, that is, TP and End f0. The MATLAB function *fitglm*() was used to predict the logistic model, using the formula in (1).

(1) $$\ln\left(\frac{1-p}{p}\right) = a_{TP}x_{TP} + a_{F0}x_{F0} + b$$

The categorical boundary can be derived from (1) corresponding to the 50% identification score, as (2). It can be simplified as (3), where the slope m represents the cue-weight ratio between TP and End f0. The intercept on the TP axis (1-c)/m represents the TP step before which all End f0 steps have Tone 2 identification.

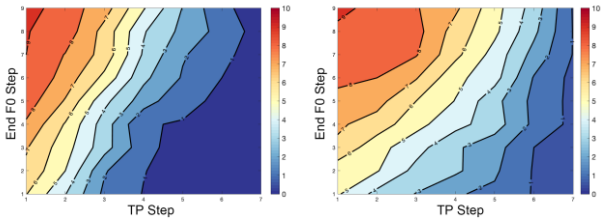
(2) $$a_{TP}x_{TP} + a_{F0}x_{F0} + b = \ln\left(\frac{1-0.5}{0.5}\right) = 0$$
$$\Rightarrow \quad x_{F0} = -\frac{a_{TP}}{a_{F0}}x_{TP} - \frac{b}{a_{F0}}$$

(3) $$x_{F0} = mx_{TP} + c$$

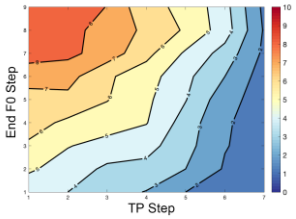
The boundary width of each participant's responses was also attained by calculating the distance between the 20% and the 80% categorical boundaries. It helps us quantitatively evaluate how categorical the perception of each group is.

3. RESULTS

The contour maps in Figure 2 show the average responses for each group. The MAN group exhibits a more vertical boundary than the other two groups. It implicates a higher cue-weight ratio between TP and End f0, which means that MAN listeners relied more on TP. The L2M and ENG groups show boundaries with similar slopes, which implies that the L2M and ENG groups have similar cue-weight ratios.



(a) MAN (n = 22) (b) L2M (n = 26)



(c) ENG (n = 20)

Figure 2: Contour maps of the average responses of each group. (Dark red area: Tone 2-dominant perception; dark blue area: Tone 3-dominant perception)

If we visually inspect the area between two and eight responses, which represents the area of ambiguous identification, the MAN group exhibits a narrower boundary than the other two groups. The ENG group has a wider ambiguous area than the L2M group. That means, in the current space, L2M listeners perceived tones in a more categorical way than ENG listeners since the identification pattern of L2M is more compact. The outperformance of L2M in categorical perception is consistent with [8].

Figure 3 shows the box plot of the boundary width of the ambiguous area. The wider the boundary width, the larger the stimulus space that is perceived ambiguously by the listener, which indicates a worse performance of categorical perception. The ANOVA results showed a significant effect of *Group* on *Boundary width* (F (2, 57) = 9.112, p < 0.001***). The results of Tukey's HSD *post hoc* analyses revealed that the ENG group had a significantly wider boundary width than the other two groups (ENG > MAN, p < 0.001***; ENG > L2M, p = 0.036*). No significant difference between MAN and L2M was found. The results confirmed that L2M identified tones in a native-like categorical way.

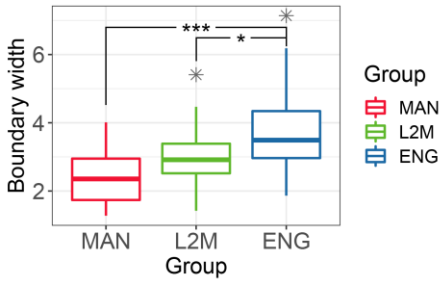


Figure 3: Box plot of the boundary width by *Group*.

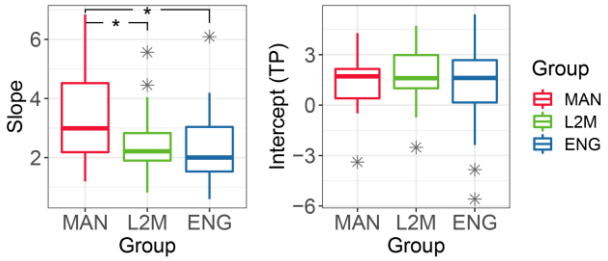


Figure 4: Box plots of the slope m (left) and the intercept (1-c)/m (right) by *Group*.

The slope m and the intercept on TP-axis (1-c)/m were predicted from the generalized logistic model fitted to each individual. If m is greater than 1, the speaker relied more on TP than End f0. The greater the slope, the higher the reliance on TP. According to the box plots in Figure 4, most participants show m values greater than 1. Only one L2M participant and two ENG participants have a slope of less than 1. The distributions of *Intercept* (*TP*) for the three groups overlapped. The ANOVA results revealed a significant effect of *Group* on *Slope* (F (2, 57) = 4.278, p = 0.018*). The results of Tukey's HSD *post hoc* analyses revealed that the MAN group had a significantly steeper boundary slope than the other two groups (MAN > L2M, p = 0.041*; MAN > ENG, p = 0.038*). No significant difference between L2M

and ENG was found. As for *Intercept (TP)*, no effect of *Group* was found (F (2, 57) = 0.736, p = 0.483).

Figure 5 shows the classification of listeners' reliance types based on the predicted slopes (χ^2 (6) = 10.78, p = 0.096·). The thresholds used here are the same as those used in [5]. The MAN group has 36.4% of speakers who relied exclusively on TP, and the percentage is larger than the other two groups. Moreover, only the L2M and ENG groups have speakers of the "more on f0" type; the ENG group has 14.3% more speakers of this type than ENG. The L2M and ENG groups have similar percentages of "Exclusively on TP" and "more on TP" speakers; however, the L2M group has 12.5% more speakers whose reliance type is "Mainly on TP" than ENG.

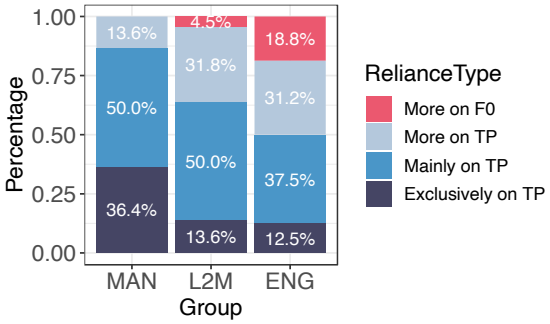


Figure 5: Percentage of each reliance type by *Group*. (More on F0: 0.5<*m*<1; more on TP: 1<*m*<2; mainly on TP: 2<*m*<4; exclusively on TP: *m*>4)

For the L2M group, Pearson's r was used to test the correlation between *Slope* and self-reported learning time of Mandarin. As Figure 6, no correlation between *Slope* and learning years was found (r = 0.0358, p = 0.8774).

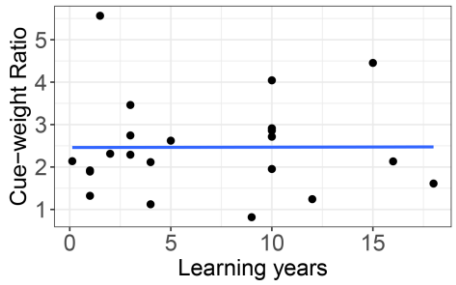


Figure 6: Scatterplot showing the distribution of Cue-weight Ratio (slope) by Learning year(s). Each L2M participant is represented by a dot.

4. DISCUSSION

The present findings demonstrated that the perceptual difference in cue weighting of TP and End f0 is significant between native (MAN) and non-native listeners (learners, L2M, and naïve English speakers, ENG). The results revealed that Mandarin listeners relied more on TP than the other two groups. For native Mandarin speakers, TP is associated with the proportion and duration of the critical portion of Tone 2 and Tone 3. An earlier TP renders a longer duration of the rising part, and a later TP renders a longer duration of the falling part. It has been proposed that the rising part and the falling part are the critical components of Tone 2 and Tone 3, respectively [18]. Moreover, the full dipping Tone 3 (MLM) usually becomes the half falling Tone 3 (ML) when it is not in the phrase-final position. In some Mandarin dialects (*e.g.*, Taiwanese Mandarin), the half Tone 3 is the only form of Tone 3, regardless of the context. Therefore, if the duration of the low falling part is long enough to be perceived, the tone will be identified as Tone 3 by Mandarin speakers. However, non-native listeners, who do not have the same experience with this tone patterning, may focus on the entire pitch contour, since for them, the full Tone 3 is the only prototypical form of Tone 3.

Recall that two predictions were on hold for the question of whether learners have acquired the native-like weighting to achieve native-like categorical perception or learners use other cue weightings, with which learners can achieve the same goal. Our present data did not show that the perception of learners is closer in terms of cue-weighting to that of native Mandarin listeners. Despite similar ability to categorize Mandarin tones, learners still have difficulty modulating their perceptual mechanisms to match native perception during learning. A possible explanation could be that the modulation failure is due to insufficient input; however, we did not find any correlation between the slope and proficiency. Learning experience appears not to take apparent effect on the cue-weighting of perceptual cues to Mandarin tone perception. Similar modulation failure has also been found in Spanish [5] and Japanese [6].

From the "top-down" perspective, adjusting cue weightings is never a specific learning or teaching target in L2 Mandarin courses, and therefore, it is typically not included in Mandarin course materials. Once learners can categorically perceive Mandarin tones, even though their cue weighting is not native-like, they can keep using the cue weight ratio that helps them steadily perform the identification. Moreover, the rising part is overtly emphasized in teaching and learning. Most learners of Mandarin are taught that the pitch contours are consistent with the diacritics in *Pinyin* (Tone 2: *á*, Tone 3: *ǎ*). The "v" shape of the Tone 3 diacritic may emphasize the impression that Tone 3 has a low rising at the end. The survey of Mandarin teaching materials in [19] showed that 14 out of 15 Mandarin textbooks for learners from Mainland China and Taiwan only represented Tone 3 as a low dipping tone (falling-then-rising). This knowledge-driven modulation could also result in attenuating the cue weight of TP, which implies the proportion of the falling part.

5. REFERENCES

[1] Keyser, S. J., & Stevens, K. N. 2006. Enhancement and Overlap in the Speech Chain. *Language* 82(1), 33–63.

[2] Lisker, L. 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1), 3-11.

[3] Schertz, J., & Clare, E. J. 2020. Phonetic cue weighting in perception and production. *WIREs Cognitive Science* 11(2), e1521.

[4] Schertz, J., Cho, T., Lotto, A., & Warner, N. 2016. Individual differences in perceptual adaptability of foreign sound categories. Attention, Perception, and Psychophysics, 78(1), 355-367.

[5] Escudero, P., & Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. Studies in Second Language Acquisition, 26(4), 551-585.

[6] Yamada, R. A., & Tohkura, Y. I. 1992. The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception and Psychophysics* 52(4), 376-392.

[7] Tupper, P., Leung, K., Wang, Y., Jongman, A., & Sereno, J. A. 2020. Characterizing the distinctive acoustic cues of Mandarin tones. *The Journal of the Acoustical Society of America* 147(4), 2570-2580.

[8] Shen, G., & Froud, K. 2016. Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America* 140(6), 4396-4403.

[9] Han, J. I., & Tsukada, K. 2020. Lexical representation of Mandarin tones by non-tonal second-language learners. *The Journal of the Acoustical Society of America* 148(1), EL46-EL50.

[10] Shen, X. S., & Lin, M. 1991. A perceptual study of Mandarin tones 2 and 3. *Language and Speech* 34(2), 145-156.

[11] Chow, R., Liu, Y., & Ning, J. 2019. The Categorical Perception of Mandarin Tone 2 and Tone 3 by Tonal and Non-tonal Listeners. In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren (eds.) *Proceedings of the 19th ICPhS*, Melbourne, Australia 2019, 3877-3881.

[12] Moore, C. B., & Jongman, A. 1997. Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America* 102(3), 1864-1877.

[13] Shen, J., Deutsch, D., & Rayner, K. 2013. On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America* 133(5), 3016-3029.

[14] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[15] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1), 388–407.

[16] Xu, Y., Gandour, J. T., & Francis, A. L. 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* 120(2), 1063-1074.

[17] Morrison, G. S., & Kondaurova, M. V. 2009. Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America* 126(5), 2159-2162.

[18] Liu, S., & Samuel, A. G. 2004. Perception of Mandarin lexical tones when f0 information is neutralized. *Language and Speech* 47(2), 109-138.

[19] Linge, O., 2011. Teaching the third tone in Standard Chinese: tone representation in textbooks and its consequences for students. *LUP Student Papers.* Online reference: http://lup.lub.lu.se/student-papers/record/1979823.

[20] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25(1), 61-83.

[21] Hallé, P. A., Chang, Y. C., & Best, C. T. 2004. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics* 32(3), 395-421.

[1] Participants who did not identify Tone 2 and Tone 3 in a correct way were excluded from the logistic regression because no meaningful boundary could be predicted from their responses.

Discrimination of Cantonese lexical tones by listeners with and without learning experience of tone languages

Yee-ping WONG

College of Professional and Continuing Education, The Hong Kong Polytechnic University
kookihee@yahoo.co.jp

ABSTRACT

This study investigated the perception of the six Cantonese lexical tones by (1) native Mandarin listeners, (2) novice Japanese-speaking listeners, (3) Japanese-speaking learners of Mandarin, and (4) Japanese-speaking learners of Cantonese. Analyses of A-prime scores in an AXB discrimination task revealed that all four groups showed high difficulty in discriminating acoustically similar pairs within the level-level pairs (T1-T3, T1-T6 and T3-T6) as well as the rise-level pair (T2-T5). Although native Mandarin listeners and learners of Cantonese generally performed better, the performance varies with tone pairs. In contrast, experience learning Mandarin tones was not significantly advantageous for tonal discrimination. Moreover, the overall performance was worse when speaker's gender was female. This result could be explained by F0 overlapping of the two female speakers. Native Mandarin listeners and learners of Cantonese are more resilient to intertalker variations as they managed to outperform the novice listeners in the more challenging voice condition.

Keywords: Chinese lexical tones, L2 speech perception, Cantonese, Japanese, Mandarin

1. INTRODUCTION

Both Cantonese and Mandarin use fundamental frequency (F0) as the primary perceptual cue to distinguish lexical meanings. There are six lexical tones in Cantonese—T1 (high level, 55), T2 (high rising, 25/35), T3 (mid-level, 33), T4 (low falling, 21), T5 (low rising, 23), and T6 (low level, 22) [2]. On the other hand, there are four tones that contrast in meaning in Mandarin—T1 (high level, 55), T2 (rising, 35), T3 (dipping, 214), and T4 (falling, 51) [10].

Although results of many previous studies suggest that speakers of tone languages generally outperformed those of non-tone languages in perception of lexical tones (e.g., [9, 24]), it is also known that in lexical tone perception, listeners with different first language (L1) and linguistic experience use different perceptual cues. For example, while both native speakers of Cantonese and Mandarin attend to both F0 height and direction in distinguishing lexical tones, Cantonese speakers do not appear to make use of duration as a useful cue [5, 6, 8, 11]. Moreover, the weighting on F0 direction and height may be different according to speakers' L1. It is reported that Taiwan Mandarin speakers were more confused than their English counterparts in tone perception. This result may be attributed to the difference between the tonal systems of Cantonese and Mandarin. Native Mandarin speakers assigned more weight to F0 direction than F0 height, whereas native Cantonese speakers were sensitive to both cues [5]. Mandarin speakers also showed difficulty in perceiving the three Cantonese level tones, especially between acoustically difficult pairs such as T3 and T6 [14]. In contrast to Mandarin speakers, native English speakers tend to rely on F0 height and average F0 [3, 4, 5, 6, 7]. [2] points out that listeners of non-tone languages perceived lexical tones mainly on a psychoacoustic mode. Although speakers of non-tone language perceive tones acoustically as non-linguistic units, they could somehow show greater sensitivity to subtle F0 differences within a tonal category [14]. It is found that native speakers of English and French could categorize, albeit in a different pattern [19, 20]. Furthermore, native Japanese speakers were able to categorize the Mandarin tones [16] and Cantonese tones [24] into their native pitch-accent categories.

Apart from the influence of L1 and psychoacoustic factors, learning the target tonal language as a second language (L2) also plays an important role in tone perception (e.g., [7, 23]). For instance, late advanced native English learners of Mandarin were approaching the perception pattern of native Mandarin speakers compared to the control group by attending to both F0 contour and average F0 in [7]. Meanwhile, although the influence of L1 is well-documented, less is known about the role of L2 in the perception of lexical tones in a third language (L3). Recently, L2 learning experience of Mandarin has been shown to modulate the perception of L3 Cantonese tones for English-speaking learners [14]. In addition, [25] reported that native Japanese learners of Mandarin outperformed novice Japanese speakers in perception of L3 Cantonese as well.

The use of F0 as perceptual cue is different in Cantonese, Mandarin, and Japanese. Japanese is a non-tonal language and pitch-accent pattern is used to contrast meaning. Combinations of high (H) pitch, low (L) pitch, and an accent (*) are used to indicate

the meaning contrast in lexical words [27]. Contrastive words like *hashi* (H*L) 'chopsticks', *hashi* (LH) 'edge', and *hashi* (LH*) 'bridge' can form minimal pairs. Japanese pitch accent is primarily realized by F0 contour [21]. Thus, Japanese speakers are expected to be skilled at processing F0 variations. However, pitch accents in Japanese is said to be phonetically different from lexical tones as they are not realized within a single syllable [18, 22].

To obtain a more complete picture of the acquisition of lexical tones and the influence of L2, the present study further expanded the discrimination of Cantonese tone contrasts to native Japanese L2 learners of Cantonese and native Mandarin speakers. To the best of my knowledge, there is no previous studies that examined the tone perception of Japanese-speaking learners of Cantonese. This study fills this gap by comparing Mandarin and Cantonese learners' performance in an AXB discrimination task and provides additional evidence as to the role of L1 and/or L2 in non-native tone perception.

2. METHOD

2.1. Participants

The study included four groups of listeners. Novice native Mandarin listeners (NM), who had no or minimal exposure to tonal languages other than their first language, consisted of 10 listeners (mean age: 25.5 years). Their performance provided baseline data for benchmark comparison. Novice native Japanese listeners (NJ) who had no or minimum exposure to any tone languages, consisted of 21 listeners (mean age: 30.0 years). To minimize the influence of accent pattern in dialects, only those who are from the Greater Tokyo Area were recruited. L2 Mandarin group (NJL2M) consisted of 12 learners (mean age: 25.0 years) who have learnt Mandarin as foreign or second language for an average of 2.1 years (SD=1.2). L2 Cantonese group (NJL2C) consisted of 10 learners (mean age: 47.2) who have learnt Cantonese as L2 for an average of 8.9 years (SD=8.1). NJL2M self-reported their Mandarin proficiency as beginner–upper intermediate and they had no prior experience with non-Mandarin tonal languages. NJL2C also reported the same for Cantonese. All participants reported no hearing problems or language disorders, and neither had professional musical training.

2.2. Stimuli

Two female and two male native speakers of Hong Kong Cantonese produced all stimuli. All combinations of the six tone pairs were included to allow a more thorough comparison. The two test stimuli of [jɐu] and [se] (/jau/ and /se/ in Jyutping used in Hong Kong), which were used in previous studies (e.g., [14]), were selected for the present study. The six tones with these two syllables all form real Cantonese words. The four speakers each read two sets of words (one set for /jau/ and one set for /se/) in the carrier sentence at a normal speech rate. They were asked to read several times and the tokens with duration which were closest to the average of the six tones were selected and segmented for the discrimination task. The intensity of all the stimuli was normalized at 70dB. The F0 values of the stimuli of the two male speakers were 124.6Hz and 128.2Hz, while that of the two female speakers were 184.5Hz and 132.3Hz on average. Figure 1 shows an example using F0 of [se]. Three native speakers identified the stimuli and the accuracy was approximately 96%.

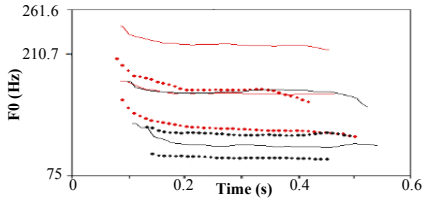


Figure 1: T1 and T6 of [se] showing F0 range of the 4 native speakers (Black=Male, Red=Female).

2.3. Procedure

The AXB discrimination task was carried out using Praat. There were 15 A–B pairs for the 6 tones and all four possible combinations (AAB, ABB, BAA and BBA) were included. Participants received 480 trials (2 syllables x 15 pairs x 4 combinations x 2 voices x 2 repetitions) presented in a mixed-talker design, in random order, and blocked by 60 trials. Within each trial, the first and third sound were uttered by one speaker and the second sound was uttered by another speaker of the same gender. 10 extra trials were included for practice. The interstimulus interval was set to 1s and the intertrial interval was 3s. For each trial, participants chose either the label '1' (X=A) or the number '3' (X=B) on the keyboard. The process was self-paced and lasted for around 1 hour, with short breaks scheduled between blocks.

2.4. Data analysis

Responses of each tone pairs were used to calculate A-prime (A') scores [16], an index of discrimination accuracy. A' was calculated based on the proportion of 'hits'(H) and 'false alarms'(F) for each pair. If H was equal to or exceeded F, A' was calculated as (1). Otherwise, A' was calculated as (2).

(1) $A' = 0.5 + ((H-F)*(1+H-F)) / (4H*(1-F))$
(2) $A' = 0.5 - ((F-H)*(1+F-H)) / (4F*(1-H))$

Mixed-effects model ANOVA was then performed on the A' scores data with Tone Pair and Speaker Gender as within-subjects factors, and Group as between-subjects factor. In the model, participant and syllable were crossed random variables. Post-hoc pairwise comparisons (with Bonferroni adjustment) were conducted to further explore the effects and interactions using the emmeans package of R.

3. RESULTS

The results of mixed-effects model ANOVA using Satterthwaite approximation revealed that the main effects of all three factors were significant: tone pair ($F(14,1421) = 161.6$, $p < .001$); gender ($F(1, 1421) = 632.1$, $p < .001$); group ($F(3, 49) = 3.6$, $p = .02$). The two-way interactions of the three factors (pair x group: $F(42, 1421) = 1.5$, $p = .02$; gender x group: $F(3, 1421) = 52.4$, $p < .001$; pair x gender: $F(14, 1421) = 52.4$, $p < .001$) are also all significant. Three-way interaction was non-significant. Paired t-tests were then carried out for multiple comparisons.

First, the overall scores across the 15 tone pairs were significantly lower for the difficult tone pairs which have similar F0 direction. The A' score was the lowest for T3-T6, followed by T1-T3, T1-T6, and T2-T5, and then T3-T4 and T4-T6. The scores of other pairs were significantly higher. Second, the A' score was higher when the stimuli were presented in female voice than in male voice. Third, the averaged A' scores across the 15 tone pairs of NJL2C group was significantly higher than the NJ group (0.894 vs. 0.835). The NM group, with mean score 0.889, also outperformed NJ group with marginal significance ($t(49) = 2.5$, $p = .09$).

The concerning interaction term of the current study—group x tone pair, is illustrated in Figure 2. Results of post-hoc analyses showed that NJL2C outperformed NJ for T1-T3 ($t(259) = 3.8$, $p = .001$) and T1-T6 ($t(259) = 3.0$, $p = .02$). NJL2C also outperformed NM for T2-T5 ($t(259) = 2.7$, $p = .05$). The A' score of NJL2C was also marginally significantly higher than NJL2M for T3-T6 ($t(259) = 2.5$, $p = .08$). On the other hand, NM outperformed NJ for T1-T6 ($t(259) = 2.6$, $p = .06$), T3-T4 ($t(259) = 2.4$, $p = .09$), T3-T5 ($t(259) = 2.5$, $p = .09$), T4-T6 ($t(259) = 2.6$, $p = .05$), and T5-T6 ($t(259) = 2.4$, $p = .09$). Further, the score of NM was also significantly higher than NJL2M for T3-T6 ($t(259) = 3.1$, $p = .01$).

Table 1 displays the mean scores of the four groups in voice of different gender. The performance was always better for all four groups when the speaker's voice was male (all $p < .001$). Moreover, only when the stimuli were presented in a female voice, significant between-group difference was observed (NJL2C > NJ, $t(59) = 3.6$, $p = .004$; NM > NJ, $t(59) = 4.0$, $p = .001$).

The mean discrimination scores by tone pair between the two types of speaker gender are given in Table 2. The score differences between the speaker gender were significant for all but level-rising tone pairs T1-T2, T2-T3, T2-T6, T3-T5, and T5-T6. Additionally, when the speaker is female, the scores of pairs T1-T6, T1-T3, and T3-T6 were substantially lower, followed by T2-T5, then T4-T6 and T3-T4. By contrast, when the speaker is male, the scores were significantly lower for T3-T6, followed by T2-T5, then T1-T3, and next T1-T6, T5-T6, and T4-T6.

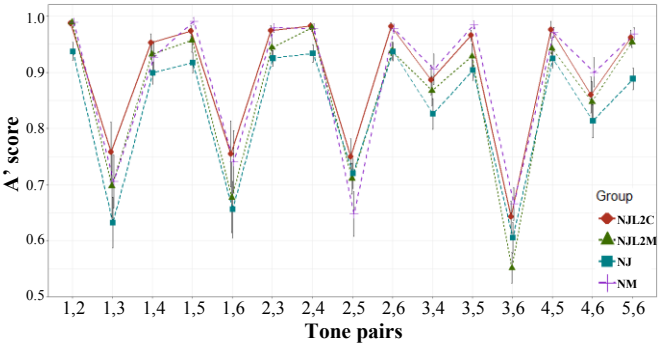


Figure 2: Average A' scores for each tone pair by the four groups (error bars = ±1 standard error).

Group	Female	Male
NJ (n=21)	0.754 (0.236)	0.915 (0.123)
NM (n=10)	0.843 (0.197)	0.935 (0.123)
NJL2M (n=12)	0.795 (0.223)	0.927 (0.126)
NJL2C (n=10)	0.835 (0.194)	0.953 (0.088)

Table 1: Mean A' scores of the four groups by speaker's gender. Standard deviations are in parentheses.

Tone pairs	Female	Male
T1-T2	0.953 (0.093)	0.984 (0.036)
T1-T3	0.477 (0.208)	0.892 (0.116)
T1-T4	0.864 (0.126)	0.980 (0.041)
T1-T5	0.921 (0.110)	0.981 (0.071)
T1-T6	0.444 (0.210)	0.946 (0.068)
T2-T3	0.936 (0.095)	0.962 (0.061)
T2-T4	0.939 (0.083)	0.985 (0.047)
T2-T5	0.637 (0.157)	0.783 (0.127)
T2-T6	0.946 (0.082)	0.961 (0.081)
T3-T4	0.769 (0.151)	0.954 (0.065)
T3-T5	0.918 (0.117)	0.955 (0.072)
T3-T6	0.559 (0.157)	0.664 (0.136)
T4-T5	0.911 (0.101)	0.984 (0.034)
T4-T6	0.740 (0.164)	0.952 (0.089)
T5-T6	0.916 (0.097)	0.949 (0.090)

Table 2: Mean A' scores for the 15 Cantonese tones by gender of the speaker. Standard deviations are in parentheses.

4. DISCUSSION

In general, tonal discrimination was poorest between the level-level tone pairs (i.e. T3-T6, T1-T3, T1-T6) and the rising-rising tone pair T2-T5 across the four groups. The performance was also poor for the lower register tone pairs, T3-T4 and T4-T6. These pairs were reported to be more difficult contrast pairs (e.g., [12, 15]). Previous studies have pointed out that the acoustic properties of the tonal stimuli robustly affect listeners' perception of tones regardless of their language backgrounds. Tone pairs having the same F0 direction within each pair are deemed to be acoustically hard to distinguish. Moreover, although substantial differences were found due to the voice of speaker, the discrimination of level-rising tone pairs were minimally affected. This suggests that phonetic similarities between tone pairs might be the most influential factor in the perception of non-native tones in the present study. Some L2 speech perception models (e.g., [1]) propose that phonetic similarities determine the assimilation or categorization of L2 sounds into L1 sounds. Tones that are more dissimilar acoustically will be easier to discern whereas tones with more similar features are more likely to cause confusion and learning difficulties.

Overall, the findings of this study show that the three native Japanese speaker groups exhibit different perception patterns in relation to their linguistic experience of tone languages. NJL2C group and NM group performed considerably better than the other two groups. Furthermore, NJL2C outperformed NJ group for the more difficult pairs, T1-T3, T1-T6, T2-T5, and T3-T6. This suggests the positive learning effect of Cantonese. In contrast, while NM group also performed better for six relatively difficult tone pairs (i.e. T1-T6, T3-T4, T3-T5, T3-T6, T4-T6, and T5-T6), it is surprising that NJL2M did not have advantage over the NJ group. This result is in contrast with findings of many previous studies [e.g., 14, 26]. Specifically, in [26], the data of 21 novice Japanese listeners and 10 Japanese-speaking learners of Mandarin were also included in the present study. The learner group showed certain advantages over the novice NJ group. Additionally, [14] reported that native English learners of Mandarin benefited from both their L1 and L2 (Mandarin) in perceiving L3 (Cantonese) tones and they outperformed the native English controls only when the F0 direction of the tone pairs were different (T1-T2 and T2-T6). By contrast, the results of this study revealed that NJL2M group did not perform better with pairs that differ in F0 direction nor pairs that differ in F0 height.

There are no middle level nor real level low-low pitch accent pattern in Japanese. A sequence of high tones is also not allowed in word-initial position in Tokyo Japanese. Additionally, there are no rising pitch accent patterns which differ in the magnitude of F0 change. It is assumed that NJ's selections depended on phonetic similarities. Meanwhile, NM are sensitive to F0 direction. It seems that their L1 (Mandarin) profoundly affects the discrimination of pairs with different F0 direction, especially when F0 are crowded in the lower register region, i.e. T3-T4, T3-T5, T4-T6 and T5-T6. NM outperformed the NJ group in distinguishing tone pairs with close F0. It is possible that the complex Cantonese tone system may have posed difficulties for Mandarin learners since their phonetic system have been reorganizes in response to Mandarin. The interlanguage phonology of NJL2M is in-between NJ and NM. Apart from psychoacoustic cues, NJL2M learned to use F0 direction (and duration) as cues in tone perception. Yet, the Mandarin learning experience of NJL2M group might not be sufficient to facilitate distinction of difficult Cantonese tone contrasts. Comparing to [26], given the fact that the two newly added Mandarin learners were still at the beginner stage (one year learning experience, they could easily feel confused when they perceive unfamiliar tone contrasts. Although the present findings do not support the claim that exposure to L2 Mandarin tones could positively modulate L3 (Cantonese) tonal perception, still, it is believed that tonal experience in Mandarin has strengthened the sensitivity of tone contour of NJL2M. In fact, the result of a Pearson correlation test showed that there is weak positive correlation ($r(358) = .1$, $p = .01$) between the length of learning and A' scores of NJL2M group. With more learning experience, Mandarin learners could potentially show additional advantages over novice listeners. To test this claim in future studies, data from more native Japanese learners of Mandarin with different proficiency levels are needed.

Similar to [26], the performance of all four groups were significantly higher when the speaker's voice was male. One justification may be that the two female speakers in this experiment spoke in very different pitches (as illustrated in Figure 1). The F0 of the lower register tones of one female speaker actually overlap with the F0 of the higher tones of the other female speaker. Although it has been noted in [13] that Cantonese tone identification can be heavily influenced by intertalker variations, the performance of NJL2C and NM were more robust even though the discrimination was conducted using female voices. This result could be accounted for by the linguistic experience of L2 Cantonese learning and the higher sensitivity to lexical tones of native Mandarin learners respectively. In other words, these two groups were less adversely affected by the speaker variations than the other two groups.

5. ACKNOWLEDGEMENTS

The author would like to thank Keiichi Tajima, Albert Lee and the reviewer for their helpful advice and constructive comments, as well as all the participants for their time. Any errors are, of course, my own.

6. REFERENCES

[1] Best, C. T. 1995. A direct realistic view of cross-language speech perception. In: Strange, W. (ed), *Speech perception and linguistic experience: Issues in cross-language research*. Baltimore: York Press, 171-204.

[2] Burnham, D., Mattock, K. 2007. The perception of tones and phones. In: Bohn, O. S., Munro, M. J. (eds), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*. Amsterdam: John Benjamins, 259-280.

[3] Francis, A. L., Ciocca, V., Ma, L., Fenn, K. 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics* 36(2), 268–294.

[4] Gandour, J. 1981. Perceptual dimensions of tone: evidence from Cantonese. *Journal Of Chinese Linguistics* 9(1), 20–36.

[5] Gandour, J. T. 1983. Tone perception in far Eastern languages. *Journal of Phonetics* 11, 149-175.

[6] Gandour, J. T. 1984. Tone dissimilarity judgments by Chinese listeners. *Journal of Chinese Linguistics* 12, 235-261.

[7] Guion, S. G., Pederson, E. 2007. Investigating the role of attention in phonetic learning. In: Bohn, O. S., Munro, M. J. (eds), *Language experience in second language speech learning: In honor of James Emil Flege*. Amsterdam: John Benjamins, 57-77.

[8] Khouw, E., Ciocca, V. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35(1), 104–117.

[9] Lee, Y. S., Vakoch, D. A., Wurm, L. H. 1996. Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research* 25(5), 527-542.

[10] Li, C. N., Thompson, S. A. 1989. *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.

[11] Liu, S., Samuel, A. G. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech* 47(2), 109–138.

[12] Mok, P., Lee, C. W. T., Yu, A. C. L. 2018. Perception and production of Cantonese tones by South Asians in Hong Kong. *Proc. Speech Prosody 2018* Poznán, 458-462.

[13] Peng, G., Zhang, C., Zheng, H. Y., Minett, J. W., Wang, W. S. Y. 2012. The effect of intertalker variations on acoustic–perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research* 55, 579-595.

[14] Qin, Z., Jongman, A. 2016. Does second language experience modulate perception of tones in a third language?. *Language and Speech*, *59*(3), 318-338.

[15] Qin, Z., Mok, P. 2013. Discrimination of Cantonese tones by speakers of tone and non-tone languages. *Kansas Working Papers in Linguistics* 34, 1-25.

[16] Snodgrass, J. G., Levy-Berger, G., Haydon, M. 1985. *Human experimental psychology*. New York: Oxford University Press.

[17] So, C. K. 2010. Categorizing Mandarin tones into Japanese pitch-accent categories: The role of phonetic properties. *Proc. Second Language Studies: Acquisition, Learning, Education and Technology* Tokyo.

[18] So, C. K., Best, C. T. 2010. Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech* 53(2), 273-293.

[19] So, C. K., Best, C. T. 2011. Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Studies in Contemporary Linguistics* 47(1), 133-145.

[20] So, C. K., Best, C. T. 2014. Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition* 36, 95–221.

[21] Sugiyama, Y. 2017. Perception of Japanese Pitch Accent without F0. *Phonetica* 74(2), 107–123.

[22] Sunaoka, K., Sun, Q., Hiki, S. 2009. Corpus-based Chinese language education: Development and assessment of Chinese tone discrimination CAI self-learning system based on learner corpus. *Journal of Liberal Arts (The School of Political Science and Economics, Waseda University)* 126, 79-98.

[23] Tsukada, K., Kondo, M., Sunaoka, K. 2016. The perception of Mandarin lexical tones by native Japanese adult listeners with and without Mandarin learning experience. *Journal of Second Language Pronunciation* 2(2), 225-252.

[24] Wayland, R. P., Guion, S. G. 2004. Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning* 54, 681-712.

[25] Wong, Y. 2019a. Categorization of Cantonese Lexical Tones by Japanese-speaking Novice Listeners and Learners of Chinese Mandarin. *Proc. International Conference on Phonetics and Phonology* Tokyo, 77.

[26] Wong, Y. 2019b. Perception of Cantonese lexical tones by Japanese non-learners and learners of Mandarin. *Proc. 19th ICPhS* Melbourne, 3867-3871.

[27] Wu, X., Tu, J., Wang, Y. 2012. Native and nonnative processing of Japanese pitch accent. *Applied Psycholinguistics* 33(3), 623-641.

Grammatical tone in Kanise Khumi: exclusive/inclusive pronominal agreement

Elissa Ikeda & Ryan Gehrmann
Payap University, Chiang Mai, Thailand
elissa_i@payap.ac.th, ryan_g@payap.ac.th

ABSTRACT

Kuki-Chin languages are known for using tonal variation to mark grammatical relationships [1]–[3]. This paper provides evidence that tone is used to mark an inclusive-exclusive distinction in pronominal subject agreement markers on Kanise Khumi verbs. An inclusive-exclusive distinction is common for Kuki-Chin independent pronouns, but not for pronominal agreement marking on the verbal complex [4], [5]. Using elicited sentences, this paper compares exclusive and inclusive agreement prefixes with respect to vowel quality (F1) and pitch (F0).

When the object is 3rd person, segmental information (i.e. vowel quality) is one feature that distinguishes exclusive from inclusive in the person-marking prefix. Pitch is another. However, in reflexive/reciprocal interpretations, pitch is the distinguishing feature between exclusive and inclusive agreement prefixes. Thus, exclusive-inclusive pronominal agreement provides an example of morphosyntactic tone in Kanise Khumi.

Keywords: grammatical tone, pronominal agreement, Tibeto-Burman, Kuki-Chin, Khumi.

1. INTRODUCTION

As Hyman [6] pointed out, tone can do everything that segments can do and more. In Kuki-Chin languages, tone can mark a variety of grammatical relationships such as genitive, benefactive, causative, number, negation, and nominalization to name a few. Peterson [7] argues that to his knowledge, "no other language of the area is reported to exhibit as widespread use of tone for marking morphosyntactic information as [Bangladesh] Khumi does." This paper provides evidence that tone is used to mark an inclusive-exclusive distinction in pronominal subject agreement markers on Kanise Khumi verbs. Kanise is a Khumi variety spoken mainly in Paletwa Township, Chin State, Myanmar.

Kanise Khumi distinguishes between exclusive and inclusive using both independent pronouns and agreement prefixes on the verb. For first-person reference, exclusive indicates that the speaker is included but the addressee is excluded. In contrast, an inclusive pronominal includes both the speaker and addressee. Table 1 displays the clusivity distinction in independent pronominals in Kanise Khumi.

	Sing.	Dual	Plural
Exclusive	kai˥	kai˥ n̥ḭ˧	kai˥ tsi̤˨˩
Inclusive		ʔai˧ n̥ḭ˧	ʔai˧ tsi̤˨˩

Table 1: First-person independent pronominals in Kanise Khumi.

Kuki-Chin languages may have bound pronominal agreement prefixes on verbs that mark subject and object [4]. Among Khomic (Southwestern) Kuki-Chin languages, Kanise Khumi appears to have the most complex pronominal agreement marking system (cf. [8]–[11]). Prefixes agree with subject person (See Table 2). Subject number is marked through suffixes: dual suffix /hɔe˥/, plural suffix /u˧/, and singular -ø.

Object	Subject			
	1 excl	1 incl	2	3
1			n̩˥	aŋ˥
2	kaŋ˥	kaŋ˥		aŋ˥
3 or Intransitive	kə˥	ka˧	n̩˥	ə˥
Reflex./Recipr.	ka˥	ka˧	ka˧	a˥

Table 2: Person agreement prefixes in Kanise Khumi.

Three lexical tones have been observed on noun and verb roots in isolation. Tone 1 has high level pitch and modal voice quality. Tone 2 is a mid checked tone (short duration & glottalized voice quality). Tone 3 is low-falling with breathy voice quality [12], [13]. A consistent association between one tone on the inclusive prefix and a different tone on the exclusive prefix would provide evidence that tone is used to distinguish this grammatical relation [14]. However, inconsistent variation of pitch on any one of the first-person subject prefixes might indicate that 1) the prefixes are simply not specified for tone, 2) the tone of the verb root might affect the tone of the agreement prefix, or 3) that a full syllable prefix might vary freely with a reduced syllable [15], [16]. In the final case, we might expect other features of syllable reduction to correlate (i.e., neutral or mid-central vowel quality, shorter duration, gradient realizations of pitch [17]–[21]).

2. METHODOLOGY

2.1 Data collection

We recorded pronominal agreement marking in sentence frames. Ezra, a 50+ year-old male speaker, provided the data from Yangon, Myanmar. We presented sentence frame prompts using PowerPoint during a Zoom meeting. Each slide contained one of five sentence frames written in Kanise orthography with a blank space for the verb. The sentence frames varied only by the independent pronoun provided as the subject of the sentence: 1st singular, 1st dual inclusive, 1st dual exclusive, 2nd singular, or 2nd dual. Each slide also included the Burmese gloss for one of 48 verbs to supply the blank space. For example, *Jandyh kai hni* ____ (ကသည်): "Yesterday we [dual exclusive pronoun] ____ (dance)." The 48 verbs represented three tone groups based on the tone melodies found on verb roots in citation forms. For transitive verbs, we did not specify an object pronoun so as to keep the sentence frames constant.

During the call, the speaker responded to each of the 240 PowerPoint slides by producing three repetitions of the sentence and then whistling the tone melody of the entire sentence. He recorded the sentences using the on-board microphone of a Samsung 11 phone in .wav format at 44.1 kHz using the ASR app [22]. He transferred recordings through Signal [23] with the "sent media quality" setting at the "standard" level.

2.2 Data analysis

We selected the two sentence frames with a first-person dual subject, either exclusive or inclusive. These frames maintained a constant grammatical structure and tone environment surrounding the subject agreement prefix and verb root. In examples 1-2, the grammatical structure appears in the first line and the tone environment in the second.

(1) [Yesterday] [**1.excl pronoun**] [**1.excl**-verb-dual]
[M H] [H M] [**1.excl**-verb-H]

(2) [Yesterday] [**1.incl pronoun**] [**1.incl**-verb-dual]
[M H] [M M] [**1.incl**-verb-H]

For each group, we eliminated sentences where 1) an extra suffix was added; 2) a phrase intervened between the independent subject pronominal and the verbal complex; 3) a valency-changing prefix intervened between the person-marking prefix and the verb root; or 4) person agreement involved a reflexive/reciprocal interpretation. Thus, we were left with 28 sentences with an inclusive subject and 27 sentences with an exclusive subject using the same set of verbs minus one. Reciprocal interpretations with six verbs were analyzed separately (See Section 3.2).

We used Praat Text Grid to annotate the vowel portion of each subject agreement prefix, each verb root, and each dual suffix in the second and third sentence repetitions.

Acoustic measures were extracted at 5 ms intervals in Praat using the PraatSauce script [24]. The data was analyzed and the plots drawn in python using the plotnine module [25]. F0 measurements are displayed in semitones relative to the mean F0 value for all of the tokens annotated for this project. All F0 and F1 plots are smoothed via loess regression with confidence intervals displayed.

3. RESULTS

3.1 Exclusive-inclusive distinction with 3rd person objects

3.1.1 Vowel quality

With third-person objects, there is a segmental distinction between 1st exclusive and 1st inclusive in terms of the vowel quality of the prefix: exclusive [kə] vs. inclusive [ka]. Figure 1 compares the F1 values of the two prefixes where values associated with the exclusive are plotted with solid lines and inclusive with dashed lines; means are shown in bold.

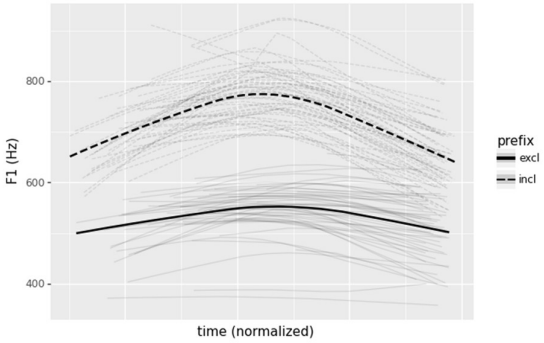


Figure 1: F1 comparison of the prefix.

At the midpoint of the vowel of the exclusive prefix, the mean F1 is around 575 Hz contrasted with around 775 Hz for the inclusive prefix. Across all observations, there is no overlap between exclusive and inclusive at the center of the vowel. This suggests that there is a distinct vowel quality target associated with each prefix, and that the vowel quality associated with the inclusive prefix is consistently more open than the exclusive prefix.

3.1.2 Pitch

There is also consistent mapping between fundamental frequency and clusivity. Figure 2 displays F0 in semitones relative to the mean where time is normalized. There is a difference of 2-3 semitones between the mean F0 curve for the exclusive prefix and the mean F0 curve for the inclusive prefix. When all tokens are considered, there is no overlap between the pitch curves of the two prefixes. This suggests that there is a distinct pitch target for each prefix, with the exclusive prefix produced with a consistently higher F0 than the inclusive.

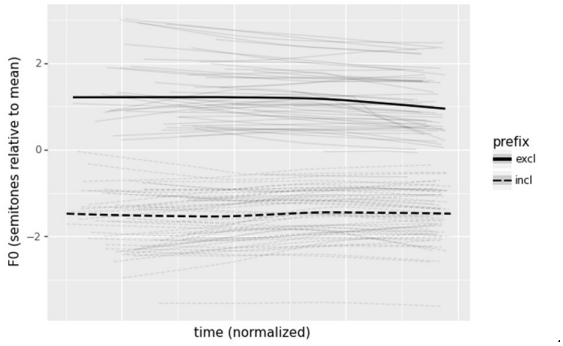


Figure 2: F0 comparison of the prefix.

The verb complex investigated in this study is comprised of three syllables: 1) person prefix, 2) verb root, 3) number suffix. Mean pitch curves spanning the verb complex are given in Figure 3. Verb complexes with an exclusive subject are plotted on the left, inclusive subject on the right. On the x-axis, time is normalized by syllable: The person prefix is represented between 0-1; the verb root between 1-2; and the number suffix between 2-3.

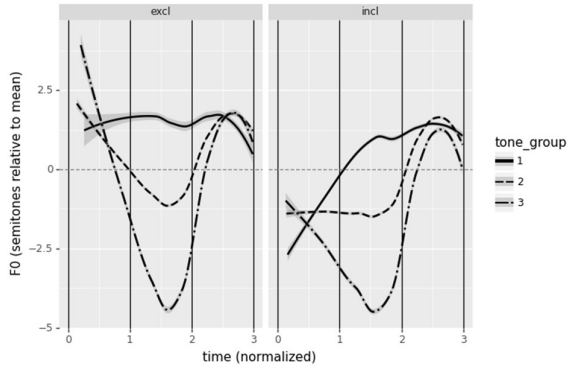


Figure 3: F0 comparison of the verb complex.

As follows from the above, prefixes associated with an exclusive subject are set higher than prefixes with an inclusive subject. The tone of the verb root does not appear to affect this pattern. The pitch target of the verb root is generally realized by the midpoint of the root (x-axis = 1.5). As expected, the suffix is consistently associated with a high pitch target.

The left boundary of each prefix was a word boundary, so the pitch of the prefix at onset gives a good sense of the pitch target of the prefix, while the right edge of the prefix indicates movement toward the pitch target of the verb root. The mean onset of the exclusive prefix is at or higher than the mean pitch target for Tone 1 verb roots (left graph, solid line at x = 1.5) Assuming for declination across the phonological word, this suggests that the pitch of the exclusive prefix could be associated with the pitch level of Tone 1. In contrast, the onset pitch of the inclusive prefix is near or below -1 semitones, close to the level of the mean pitch target for verbs associated with Tone 2 (right graph, dashed line at x = 1.5). There is an inverse relationship where the mean onset pitch of the exclusive prefix is highest before a low-tone root, while the mean onset pitch of the inclusive prefix is lowest before a high-tone root.

3.2 Reflexive and reciprocal interpretations

We designed the experiment with a third-person object in mind; however, no object was specified in the sentence prompts. Without prompting, Ezra produced sentences with a reflexive or reciprocal interpretation for six verbs. These verbs were /təl/ ‘fight’, /gõl/ ‘meet’, /jo̱ɬ/ ‘sell’, /ʔu̱ɬ/ ‘bury’, /ʔo̱ɬ/ ‘cut’, and /ŋa̱i̱ɬ/ ‘love’. Many Kuki-Chin languages have preverbal middle voice markers that distinguish person [4]. In Kanise Khumi, a reflexive/reciprocal interpretation affects the person-marking prefix as shown previously in Table 2. In these cases, vowel quality is not relevant in the exclusive-inclusive distinction. Pitch, however, is a reliable cue. Examples 3-6 demonstrate this point. Examples 3 and 4 involve a third-person object. The exclusive and inclusive agreement markers differ in both vowel quality and pitch. Examples 5 and 6 involve a reciprocal interpration where the exclusive and inclusive agreement markers differ only in pitch.

(3) (kaiˈl-n̥i̱ɬ) (ʔəlni̱ɬ-teŋəl) **kə**l-təl-hɔel
(1.excl-dual) (3.sg-obj) **1.excl**-fight-dual

(4) (ʔaiˈl-n̥i̱ɬ) (ʔəlni̱ɬ-teŋəl) **ka**ɬ-təl-hɔel
(1.incl-dual) (3.sg-obj) **1.incl**-fight-dual

(5) (kaiˈl-moˈl-n̥i̱ɬ-teŋəl) **ka**l-təl-hɔel
(1.excl-recip-dual-obj) **1.excl.recip**-fight-dual

(6) (ʔai˥-mo˥-n̥i̥˧-teŋə˥) **ka˧**-tə˥-hɔe˥
(1.incl-recip-dual-obj) **1.incl.recip**-fight-dual

3.2.1 Reciprocal interpretation and pitch

Figure 4 illustrates the pitch contrast between the exclusive and inclusive prefixes that correspond to a reciprocal interpretation.

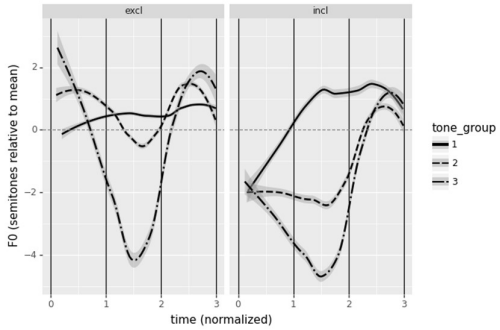


Figure 4: F0 comparison of the verb complex in reciprocal interpretations.

The prefix pitch is displayed between timepoints 0 to 1. The pitch of the prefix is above 0 semitones for the exclusive and close to -2 for the inclusive. Like Figure 3, this F0 comparison of the prefixes shows that pitch distinguishes exclusive and inclusive.

3.2.2 Reciprocal interpretation and vowel quality

Vowel quality does not distinguish the exclusive and inclusive prefixes in reciprocal interpretations (Figure 5).

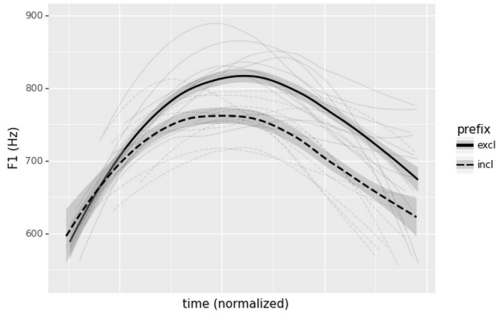


Figure 5: F1 comparison of the prefix in reciprocal interpretations.

In Figure 5, the F1 of the prefix ranges between 700-800 Hz regardless of whether it is exclusive or inclusive, signifying an open vowel for both. This contrasts with Figure 1, which showed a difference in F1 when the object is 3rd person. In those cases, the exclusive prefix had a mean F1 of 550 Hz, signifying a mid vowel height.

3.2.3 Alignment with other subjects

Reflexive and reciprocal constructions with other subjects confirm a clusivity distinction signalled through pitch. The person agreement prefix [ka˥] is used with first-person singular and first-person exclusive (dual/plural) subjects. On the other hand, [ka˧] is used with first-person inclusive (dual/plural) and second-person subjects. In other words, [ka˥] includes the speaker but excludes the addressee(s) in a reflexive/reciprocal construction. In contrast, [ka˧] includes the addressee(s) as sole or part-actor. Table 3 displays this alignment pattern.

	High prefix ka˥	Mid prefix ka˧
1st singular	X	
1st dual exclusive	X	
1st dual inclusive		X
2nd singular		X
2nd dual		X

Table 3: Tone alignment in pronominal subject agreement marking with reciprocal/reflexive interpretations.

4. DISCUSSION

In Kanise Khumi, exclusive and inclusive are distinguished in independent pronouns and preverbal agreement. Vowel quality (F1) and pitch (F0) signal this distinction in first-person subject prefixes when acting on a third person object. However, sentences with a reflexive or reciprocal interpretation reveal that pitch is the most consistent feature distinguishing clusivity on agreement prefixes. While there may be redundant cues in the noun phrase, this constitutes a form of *morphosyntactic tone* where "a form carrying a specific meaning must be marked by a specific tone, which is interpreted as being directly linked to this meaning" [14, p. 13].

This paper has not addressed other suprasegmental features that typically bundle together with pitch in Southeast Asian tone systems such as voice quality and duration [26], [27]. To accurately interpret duration and voice quality in a preverbal position, we must first investigate the reliability of duration and voice quality as tone cues on verb roots that precede a suffix. In addition, the results of this study should be corroborated by testing listener discrimination and the production of other speakers.

6. REFERENCES

[1] N. T. Par, "Agreement and verb stem alternation in Senthang Chin," Unpublished MA thesis, Payap University, Chiang Mai, Thailand, 2016.

[2] L. M. Hyman, "Kuki-Thaadow: An African tone system in Southeast Asia," *UC Berkeley Phonol. Lab Annu. Rep.*, vol. 3, 2007.

[3] E. J. Henderson, "Grammar and tone in South-East Asian languages," *Wiss. Z. Karl-Marx-Univ.-Leipz.*, vol. 16, pp. 171–78, 1967.

[4] K. Van Bik, "Typological profile of Kuki-Chin languages," in *The Languages and Linguistics of Mainland Southeast Asia*, De Gruyter Mouton, 2021, pp. 369–402.

[5] S. DeLancey, "Argument indexation (verb agreement) in Kuki-Chin," presented at the Workshop on Issues in Kuki-Chin Linguistics 46th International Conference on Sino-Tibetan Languages and Linguistics, Dartmouth College, 2013.

[6] L. M. Hyman, "Tone: Is it different?," *Handb. Phonol. Theory*, vol. 2, 2011.

[7] D. A. Peterson, "Bangladesh Khumi," in *The Mainland Southeast Asia Linguistic Area*, A. Vittrant and J. Watkins, Eds. Berlin, Boston: De Gruyter Mouton, 2019, pp. 12–55.

[8] H. So-Hartmann, "Valence-changing Prefixes in Kuki-Chin," presented at the ICSTLL 46, Dartmouth College, Aug. 2013.

[9] D. A. Peterson, "On Khumi verbal pronominal morphology," *Proc. Twenty-Eighth Annu. Meet. Berkeley Linguist. Soc. Spec. Sess. Tibeto-Burman Southeast Asian Linguist.*, pp. 99–110, 2002.

[10] D. A. Peterson, "The stem alternation in Rengmitca," *Lang. Peoples East. Himal. Reg.*, vol. 1, no. 1, pp. 80–94, 2020, doi: 10.5070/H91150997.

[11] H. So-Hartmann, "Properties of the Lemi Verb Complex."

[12] E. Ikeda, "Tone patterns in Kanise Khumi monosyllables," in *Proceedings of the 11th Payap University and UNRN Research Symposium 2021*, Chiang Mai, Thailand, Feb. 2021, pp. 1223–1234.

[13] E. Ikeda, "Syllable structure and morphemicity in tone patterns on verbs in Kanise Khumi," in *JSEALS Special Publication No. 8, Papers from the 30th Meeting of the Southeast Asian Linguistics Society (2021)*, 2022, pp. 145–162. [Online]. Available: http://hdl.handle.net/10524/52498

[14] M. Konoshenko, "Tone in grammar: What we already know and what we still don't," *Vopr. Jazyk.*, no. 4, pp. 101–114, 2017.

[15] C. S. Hornéy, "A Phonological Analysis of Mro Khimi," Master's thesis, University of North Dakota, Grand Forks, ND, 2012.

[16] J. Watkins, "A first account of tone in Myebon Sumtu Chin," *Linguist. Tibeto-Burman Area*, vol. 36.2, pp. 97–127, 2013.

[17] L. West, "Pitch specification of minor syllables in Sgaw Karen," *J. Acoust. Soc. Am.*, vol. 139, no. 4, pp. 2218–2218, Apr. 2016, doi: 10.1121/1.4950643.

[18] B. A. Butler, "Deconstructing the Southeast Asian sesquisyllable: a gestural account," Doctoral Dissertation, Cornell University, Ithaca, New York, 2014.

[19] P. Pittayaporn, "Typologizing sesquisyllabicity: The role of structural analysis in the study of linguistic diversity in Mainland Southeast Asia," in *The languages of Mainland Southeast Asia: The state of the art*, N. J. Enfield and B. Comrie, Eds. Berlin, München, Boston: De Gruyter Mouton, 2015, pp. 500–528.

[20] W. H. Tan, "Descriptive Analysis of Minor Syllables in Kanise Khumi," Independent Study, Payap University, Chiang Mai, Thailand, 2022. [Online]. Available: https://drive.google.com/file/d/1eAT4ZAVNrCt4fBb1v8EUHuyM2pa92ZYQ/view?usp=sharing

[21] K. E. Herr, "The phonological interpretation of minor syllables, applied to Lemi Chin," Master's thesis, Payap University, Chiang Mai, Thailand, 2011.

[22] NLL Apps, "Another Sound Recorder (ASR)." Nov. 01, 2022. [Online]. Available: https://nllapps.com/apps/asr/

[23] "Signal." Signal, Nov. 01, 2022. [Online]. Available: https://signal.org/en/

[24] J. Kirby, "Praatsauce: Praat-based tools for spectral analysis." 2018. [Online]. Available: https://github.com/kirbyj/praatsauce

[25] H. Kibirige, "plotnine: A grammar of graphics for Python. v0.10.1." 2018. [Online]. Available: https://github.com/has2k1/plotnine

[26] E. J. Henderson, "The topography of certain phonetic and morphological characteristics of Southeast Asian languages," *Lingua*, vol. 15, pp. 400–434, 1965.

[27] J. Kirby and M. Brunelle, "Southeast Asian tone in areal perspective," *Camb. Handb. Areal Linguist.*, pp. 703–731, 2017.

CONTACT-INDUCED TONOGENESIS IN HONG KONG ENGLISH

Ping Hei Yeung

Georgetown University
py101@georgetown.edu

ABSTRACT

Tonogenesis has been commonly described as a language-internal phonetic process. Yet, recent literature on contact languages in Africa suggests that it may also occur when languages with different word-prosodic systems come into contact. To test whether language contact is indeed a possible trigger of tonogenesis, this study compares the pitch production of monosyllabic content and function words by 39 speakers of Hong Kong English (HKE), a contact variety with tonal Cantonese substrate, with 30 speakers of American English (AE). Dynamic F0 was measured using Praat and statistically evaluated with Generalized Additive Mixed Modelling (GAMM). HKE speakers were found to produce content words with a significantly higher F0 than function words, while AE speakers did not show any significant difference in F0. The results indicate that HKE is a tone language with two contrastive lexical tones. Thus, apart from language-internal phonetic factors, language contact may also motivate tonogenesis.

Keywords: Tone, language contact, tonogenesis, sound change, English

1. INTRODUCTION

Tonogenesis, the phonologization of lexical tones in a previously non-tonal language, is often described as a language-internal process motivated by laryngeal articulation of consonants, e.g. Kammu [1]–[3], vowel height, e.g. U [2], [4], and F0 correlates of stress, e.g. Swedish [5], [6]. Recently, studies on contact languages in Sub-Saharan Africa and the Caribbean suggest that tonogenesis may also occur as a result of contact between languages with different word-prosodic systems, especially when the substrates are tonal Niger-Congo languages and the superstrates are non-tonal Indo-European languages. Such examples include Central African French [7], Equatorial Guinean Spanish [8], Ghanaian English [9], Nigerian English [9], Pichi [10], and Saramaccan [11]. Even though their superstrates are non-tonal Indo-European languages, they have all become tonal with a {H, L} two-tone system, which can be illustrated by the minimal pair *wood* (H tone) vs. *would* (L tone) in Nigerian English [8], [9].

The reinterpretation of F0 correlates of non-tonal L2 stress systems by L1 tone language speakers has been proposed to be the source of tonogenesis in these languages. Since stressed syllables in many non-tonal languages like English receive a H* pitch accent, tone language speakers would treat the pitch accent as lexical and assign the stressed syllables a H tone. Conversely, since unstressed syllables do not carry any pitch accent, they would tend to have a lower pitch than stressed syllables, and therefore tone language speakers would assign unstressed syllables a L tone [7], [8].

Thus, tonal minimal pairs in these contact languages often originate from distinctions between content words and function words. In Pichi, an English-based creole, the minimal pair *bay* 'by' (L tone) vs. *báy* 'buy' (H tone) differ not only in their tone but also their syntactic category: the function word *bay* receives a L tone while the content word *báy* receives a H tone [10]. Function words in English often have two variants: a weak form and a strong form. The unstressed weak form does not receive any pitch accents while the strong form does. Since the unstressed variants of function words have a much higher frequency than the stressed variants [12], the former would be more salient to L2 speakers, favoring them to analyze those words as having a L tone.

A similar observation has been made for Hong Kong English (HKE), a contact variety of English with tonal Cantonese substrate. Due to the presence of lexical pitch contours in polysyllabic content words, Yiu [13] and Wee [14] classified HKE as a tone language with a {H, Ø} two-tone system. However, it remains unclear whether the tones in HKE are contrastive as previous studies did not provide any examples of minimal pairs or compare pitch production with non-tonal varieties of English. Thus, the objective of this study is to examine whether tonogenesis has occurred in HKE by comparing pitch production of monosyllabic content and function words between speakers of HKE and American English (AE). Given the evidence from other contact languages, it is hypothesized that speakers of HKE would make a tonal distinction between content and function words by using a higher pitch for the former and a lower pitch for the latter, while speakers of AE would not.

2. METHODOLOGY

2.1. Participants

39 speakers of HKE aged between 18-58 (22 females, 17 males) and 30 speakers of AE aged between 18-30 (15 females, 15 males) were recruited. Participants in the HKE group were all born and raised in Hong Kong, and they were fluent bilinguals in both English and Cantonese. Participants in the AE group on the other hand were all born and raised in the United States, and they spoke AE as their native language. Participants reported no history of speech or hearing impairments.

2.2. Materials and procedure

The HKE data were drawn from a larger sociolinguistic study that includes sociolinguistic interviews, a word list reading task, and a minimal pair reading task. The AE group performed the same two reading tasks but they did not take part in the sociolinguistic interviews. For the purpose of this study, only the speech data from the word list reading task were analyzed. In this task, both groups were instructed to read aloud a list of 155 words, among which 9 pairs of segmentally identical content words and function words were the targets, which are shown in Table 1. By using content word/function word homophones, the segmental effects on F0 were controlled. These words were chosen also because they were reported to be tonal minimal pairs in tonal contact varieties of English [9].

Table 1: Syntactic category of the 18 target words. Words on the same row are segmentally identical in both varieties.

Function word	Content word homophone
be	bee
but	butt
or	ore
an	Ann
their	there
by	buy
for	four
to	two
would	wood

Data collection took place in quiet public spaces or the participants' home in Hong Kong for the HKE group and a sound-attenuated booth at Georgetown University for the AE group. Words were presented one at a time in a pseudo-random order on a laptop computer and they were embedded in the carrier phrase "say ___ again" to keep the phonological environment constant. The target words were placed at the middle of the carrier phrase to avoid the effect of boundary tones. The target words were primed to be the focus of the phrase such that they receive a H* pitch accent in AE regardless of their syntactic category [15], [16]. Participants were instructed to read aloud each word along with the carrier phrase three times, resulting in a total of 54 target tokens per speaker. Audio was recorded at a 44.1 kHz sampling rate and 16-bit sample depth using condenser microphones.

2.3. Data analysis

The audio recordings were segmented using Montreal Forced Aligner [17], which were then manually reviewed and corrected. Dynamic F0 was measured using Praat [18] at every 10% interval between 10% and 90% of the sonorant portion of the target words. The tokens were then coded for their syntactic category, adjacent segments, duration, number of repetitions, and English variety. Duration was defined as the duration of the sonorant portion of the word, while adjacent segment was defined as the onset and coda consonants of the word. After that, the F0 values of each token were converted into semitones and z-score normalized by speaker to account for human pitch perception and individual differences in voice pitch.

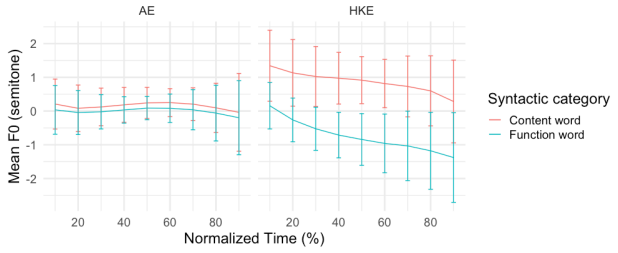
The normalized F0 was analyzed using Generalized Additive Mixed Modelling (GAMM) [19]. GAMM was fit using the *mgcv::bam()* function [19], [20] and evaluated using functions from the *itsadug* and *gratia* packages [21], [22]. The dependent variable was the normalized F0. The independent variable was the interaction of English variety and syntactic category. The reference level of this interaction term was content word by AE speakers. The interaction was included as a parametric effect and as a smooth in the model. The smooth included time, which is the normalized time point of the measurements. A smoothing term models the non-linear F0 values over time. The parametric effects of duration and repetition, as well as their fixed interactions with time were included to model their influence on F0. A difference smooth accounted for the non-linear effect of adjacent segments. Speaker was included as a random intercept and slope interacting with syntactic category to model the variable effect of syntactic category on individual speakers. Word was included as a random intercept and slope interacting with English variety to model variety-specific differences in the articulation of individual words. To account for the relationship between measurements taken at consecutive time points, an autoregressive error term was included. An AR1 model was incorporated to account for autocorrelation of residuals. The model was

calculated using a scaled-t distribution to correct for non-normality of the model residuals as indicated by *mgcv::gam.check()*.

3. RESULTS

The F0 values of content and function words in the two varieties relative to the mean F0 before z-score normalization are given in Figure 1. At the 10%, 50%, and 90% intervals, AE speakers had a mean F0 of 0.208 (SD±0.74), 0.242 (SD±0.461), and -0.038 (SD±1.154) semitones for content words and 0.036 (SD±0.723), 0.087 (SD±0.346), and -0.198 (SD±1.099) semitones for function words. HKE speakers on the other hand had a mean F0 of 1.344 (SD±1.053), 0.917 (SD±0.697), and 0.283 (SD±1.227) semitones for content words and 0.157 (SD±0.687), -0.842 (SD±0.766), and -1.378 (SD±1.329) semitones for function words. Content words had a higher F0 than function words by 0.172, 0.155, and 0.16 semitones in AE and 1.187, 1.759, and 1.661 semitones in HKE.

Figure 1: Mean F0 (semitone) of content and function words by English variety.



In regard to the results of GAMMs, Figure 2 provides the predicted F0 trajectory by each syntactic category for speakers of AE and HKE, while Figure 3 demonstrates the significance of the F0 difference between the two syntactic categories for each variety. HKE speakers produced the content words with a significantly higher F0 throughout the entire sonorant duration than the function words. As shown in Figure 2, the 95% confidence interval of the content words did not overlap with that of the function words. Likewise, Figure 3 indicates a significant difference in F0 throughout the whole sonorant duration. Then, as shown in Figure 3, the degree of F0 difference between content and function words changed throughout the sonorant duration of the target words. The difference increased from the beginning to around the 60% point of sonorant duration, and then the difference decreased from that point onwards. The shape of the F0 trajectory also differed between the two syntactic categories. While a falling F0 contour was observed across both categories, function words were characterized by a sharper drop in F0 than content words.

Figure 2: Fitted GAMM smooths with 95% confidence interval for F0 (z-score) of content and function words by English variety.

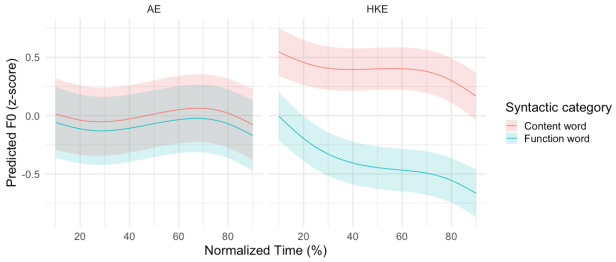
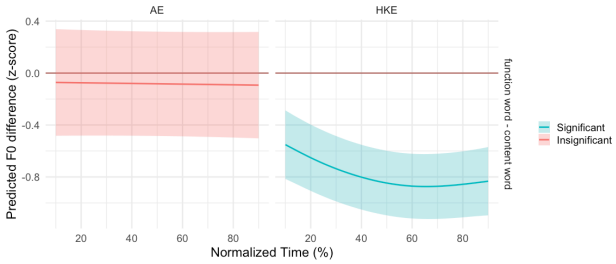


Figure 3: Difference plot for function and content words by English variety. Blue line indicates significant difference while red line indicates insignificant difference.



Conversely, the F0 of the AE group stayed relatively constant throughout the sonorant duration regardless of syntactic category. As shown in Figure 3, although the function words had a slightly lower F0 than the content words, the difference was not significant. It is also shown in Figure 2 by the complete overlap of the 95% confidence intervals of the predicted F0 contours of content and function words. Therefore, there was no significant effect of syntactic category on the dynamic F0 of AE speakers. Furthermore, the confidence intervals for each syntactic category in the AE group were wider than those of the HKE group, which indicates that AE speakers produced the target words with greater variability in F0. Thus, speakers of HKE and AE differed in their F0 production of monosyllabic content and function words. Throughout the sonorant duration, HKE speakers produced the content words with a significantly higher F0 than function words, while no effect of syntactic category was found for AE speakers.

4. DISCUSSION

The results of the current study support the hypothesis that HKE is a tone language with two lexical tones. Unlike speakers of AE who maintain the same pitch contour regardless of syntactic category, speakers of HKE produced the content words with a significantly higher pitch than the function word homophones. This difference is apparent in the production of the two phrases "*Say four again*" and "*Say for again*", which are shown in Figures 4 and 5. The HKE speaker in Figure 4 produced two distinct pitch

contours: a higher pitch for "*four*" and a lower pitch for "*for*". The target words in the two phrases are specified for their lexical tone, with the content word *four* being assigned a higher tone and the function word *for* a lower tone. On the other hand, the AE speaker in Figure 5 produced the two sentences with the same pitch contour, corroborating existing literature on English prosody [15], [16]. Since the target words were the focus of the carrier phrase, they became prosodically prominent regardless of their syntactic category. The function word *for* appeared in its stressed form, making it both segmentally and suprasegmentally indistinguishable from its content word homophone *four*.

Figure 4: F0 contour of "*Say four again*" (upper) and "*Say for again*" (lower) by a speaker of HKE.

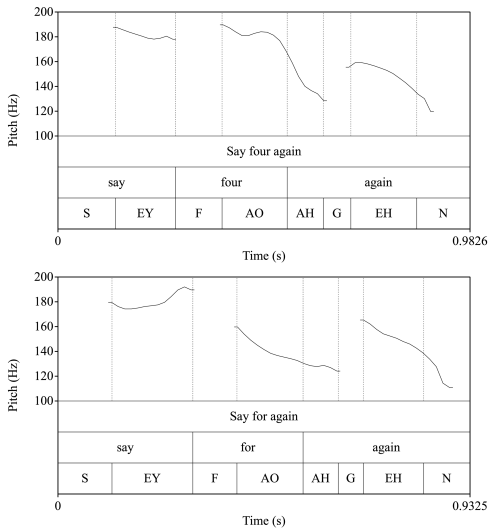
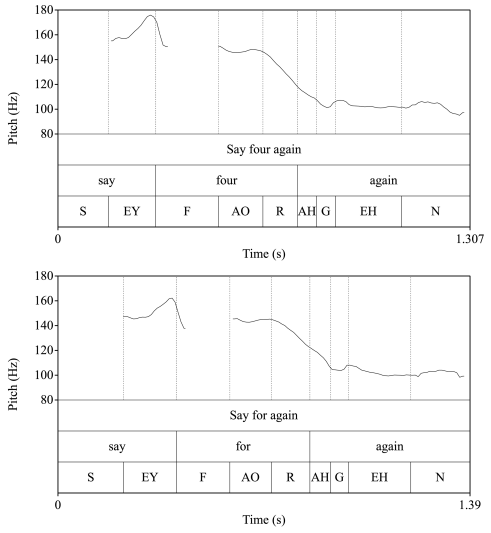


Figure 5: F0 contour of "*Say four again*" (upper) and "*Say for again*" (lower) by a speaker of AE.



However, the HKE tones observed in the current study differ in their trajectory from those reported in Yiu [13] and Wee [14]. Instead of level tones with a stable pitch throughout the duration, the ones found in monosyllabic content words and function words involve a falling pitch contour, which is shown in both the predicted F0 trajectory by GAMM in Figure 2 and the raw F0 trajectory in Figures 1 and 4. Such a difference might not be phonological but instead a result of tonal coarticulation in the carrier phrase. Following Wee [14], the preceding word in the carrier phrase *say* would have a H surface tone and the following word *again* would have a M-H surface tone. When the tone in the target word differs from those of the adjacent tone-bearing units, there would be transitions in pitch target that create pitch contours. More in-depth analysis is required to determine whether the tones assigned to the monosyllabic content and function words are phonologically the same as the ones found in previous studies.

HKE and contact languages in Sub-Saharan Africa and the Caribbean share the same tonal inventory and the same strategy in tone assignment: they all have two lexical tones differing in their pitch levels, and they all assign a higher tone to content words and a lower tone to function words [9]. This could be caused by similarities in both the substrates' tonal inventories and the superstrate's intonation system. The various Niger-Congo languages spoken in Sub-Saharan Africa and Cantonese spoken in Hong Kong all have level tones distinguished by their pitch level relative to each other [23], [24]. On top of that, the superstrates like English, French, and Spanish all assign pitch accents to specific syllables of a word, and their function words are all generally unstressed by default [7], [15], [16], [25], [26]. Both factors might have facilitated the formation of the {H, L} two-tone systems found in these tonal contact languages.

5. CONCLUSION

With the production of tonal minimal pairs in HKE as evidence, this study demonstrates that language contact between tonal substrates and non-tonal superstrates is indeed a possible trigger of tonogenesis. Although some North Germanic languages like Swedish also underwent tonogenesis that stemmed from the reinterpretation of F0 correlates of stress [5], [6], the case of HKE is unique as it is a contact-induced change rather than a language-internal change. Given that tonogenesis has occurred repeatedly in various contact languages from HKE to Central African French, the notion of lexical tone being a marked phonological feature that readily disappears during language contact [27], [28] is challenged.

6. REFERENCES

[1] S. Premsrirat, 'Khmu dialects: a case of register complex and tonogenesis', in *Cross-linguistic studies of tonal phenomena: historical*

development, phonetics of tone, and descriptive studies, Tokyo, 2003.
[2] J. O. Svantesson, 'Tonogenetic mechanisms in northern Mon-Khmer', *Phonetica*, vol. 46, no. 1–3, pp. 60–79, 1989, doi: 10.1159/000261829.
[3] J. O. Svantesson and D. House, 'Tone production, tone perception and Kammu tonogenesis', *Phonology*, vol. 23, no. 2, pp. 309–333, Aug. 2006, doi: 10.1017/S0952675706000923.
[4] J. Kingston, 'Tonogenesis', in *The Blackwell Companion to Phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds., John Wiley & Sons, Ltd, 2011, pp. 1–30. doi: 10.1002/9781444335262.wbctp0097.
[5] E. Gårding, *The Scandinavian word accents*. Lund, Sweden: Gleerup, 1977.
[6] T. Riad, 'The Origin of Scandinavian Tone Accents', *Diachronica*, vol. 15, no. 1, pp. 63–98, Jan. 1998, doi: 10.1075/dia.15.1.04ria.
[7] B. Guri, 'Prosodie et contact de langues: le cas du système tonal du français centrafricain', PhD Thesis, Université de Nanterre - Paris X, 2012. [Online]. Available: https://tel.archives-ouvertes.fr/tel-00789349
[8] G. B. Steien and K. Yakpo, 'Romancing with tone: On the outcomes of prosodic contact', *Language*, vol. 96, no. 1, pp. 1–41, 2020, doi: 10.1353/lan.2020.0000.
[9] C. Gussenhoven, 'On the intonation of tonal varieties of English', in *The Oxford Handbook of World Englishes*, M. Filppula, J. Klemola, and D. Sharma, Eds., Oxford University Press, 2014.
[10] K. Yakpo, *A grammar of Pichi*. Language Science Press, 2018. [Online]. Available: https://langsci-press.org/catalog/view/85/91/1370-1
[11] J. Good, 'Tone and accent in Saramaccan: charting a deep split in the phonology of a language', *Lingua*, vol. 114, no. 5, pp. 575–619, May 2004, doi: 10.1016/S0024-3841(03)00062-7.
[12] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, 'Effects of disfluencies, predictability, and utterance position on word form variation in English conversation', *J. Acoust. Soc. Am.*, vol. 113, no. 2, p. 1001, Jan. 2003, doi: 10.1121/1.1534836.
[13] S. Yiu, 'Aspects of tone in Cantonese English', The University of Hong Kong, Hong Kong, 2014. doi: 10.5353/th_b5481868.
[14] L.-H. Wee, 'Tone assignment in Hong Kong English', *Language*, vol. 92, no. 2, pp. e67–e87, 2016, doi: 10.1353/lan.2016.0039.
[15] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.
[16] E. Selkirk, 'The prosodic structure of function words', in *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1996, pp. 187–213.
[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, 'Montreal Forced Aligner'. 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/
[18] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer [Computer program]'. 2022. [Online]. Available: http://www.praat.org/
[19] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. New York: Chapman and Hall, 2017. doi: 10.1201/9781315370279.
[20] S. N. Wood, 'mgcv: Mixed GAM computation vehicle with automatic smoothness estimation'. 2022. [Online]. Available: https://CRAN.R-project.org/package=mgcv
[21] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, 'itsadug: Interpreting time series and autocorrelated data using GAMMs'. 2022. [Online]. Available: https://CRAN.Rproject.org/package=itsadug
[22] G. L. Simpson, 'gratia: Graceful ggplot-based graphics and other functions for GAMs fitted using mgcv'. 2022. [Online]. Available: https://gavinsimpson. github.io/gratia/
[23] S. Matthews and V. Yip, *Cantonese: A Comprehensive Grammar*. Routledge, 2013.
[24] D. Odden, 'Tone: African languages', in *The handbook of phonological theory*, Cambridge: Blackwell Publishers Ltd., 1995, pp. 444–475.
[25] J. I. Hualde and P. Prieto, 'Intonational variation in Spanish: European and American varieties', in *Intonation in Romance*, S. Frota and P. Prieto, Eds., Oxford University Press, 2015, pp. 350–391. doi: 10.1093/acprof:oso/9780199685332.003.0010.
[26] I. Roca, 'Saturation of Parameter Settings in Spanish Stress', *Phonology*, vol. 22, no. 3, pp. 345–394, 2005.
[27] P. Trudgill, 'Contact and Sociolinguistic Typology', in *The Handbook of Language Contact*, R. Hickey, Ed., John Wiley & Sons, Ltd, 2010, pp. 299–319.
[28] J. H. McWhorter, 'Identifying the Creole Prototype: Vindicating a Typological Class', *Language*, vol. 74, no. 4, pp. 788–818, 1998, doi: 10.2307/417003.

PHONETIC REALIZATIONS OF NEUTRAL TONES IN WUHAN CHINESE DISYLLABIC WORDS

Zhou Duanchi

Hitotsubashi University
duanchizhou@gmail.com

ABSTRACT

The purpose of this study is to clarify the realisation of the neutral tone (NT) phenomenon in Wuhan Chinese, one of the Southwestern Mandarin varieties, based on quantitative analysis using acoustic-phonetic methods. Specifically, we examine 1) how many types of NT exist in Wuhan Chinese, and 2) what factors determine the pitch value of NT syllables. According to previous studies based on impressionistic observations of the Wuhan tone system, there are two types of NT, realized with a pitch value of 5 (highest pitch) or 3 (mid pitch level). Contrary to previous studies, however, our quantitative analysis revealed that there are four types of NT in Wuhan Chinese and that the pitch value of NTs is determined by the underlying tones of both the first and second syllables.

Keywords: neutral tone, Wuhan Chinese, cluster analysis.

1. INTRODUCTION

In Wuhan Chinese, one of the sub-varieties of Southwestern Mandarin Chinese, there are four basic lexical tones in the citation form. These are referred to as T1, T2, T3, and T4, respectively. Figure 1 schematically shows the pitch contours of the four tones. Pitch values are expressed using the 5-level tone letters of Chao [2], where 1 indicates the low range, 2 the mid-low range, 3 the middle range, 4 the mid-high range, and 5 the high range.

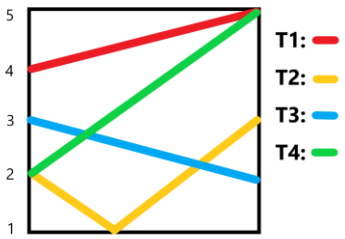


Figure 1: Schematic pitch contours for the four basic tones in Wuhan Chinese based on five-level tone letters.

In words with two or more syllables, the phenomenon of neutral tone (NT) may occur. This study defines NT as a phenomenon where the duration of a non-initial syllable is shortened and the number of tonal contrasts in that syllable is neutralised. In other Chinese dialects, there are rule-governed and non-rule-governed NT words [5,7]), and this holds true in Wuhan Chinese, too. In the former, NT is predictable grammatically, whereas it is lexically determined in the latter.

According to previous studies, there are two types of NT in Wuhan Chinese [3,6,9]: NT with pitch value 5 and with pitch value 3. Opinions differ regarding the factors that determine which of the two types of NT appears. Hou [3] and Zhu [10] argued that this factor is the tone of the first syllable. According to these studies, NT has a pitch value of 5 if the first syllable is {T1, T4}, whereas it has a pitch value of 3 if the tone of the first syllable is {T2, T3}. (Hereafter, "{A, B}" stands for "A or B"). In contrast, Li [6] states that the underlying tone of the second syllable itself determines the realisation of NT, and that if it is {T2, T3, T4}, then NT has a pitch value of 5. The abovementioned studies are based on impressionistic observations of Wuhan tones and provide no quantitative data, which may account for the absence of a consensus on the factors determining the pitch value of NTs. This study aims to clarify 1) how many types of NT exist in Wuhan Chinese, and 2) what factors determine the pitch value of NTs based on a quantitative analysis using acoustic-phonetic methods.

2. METHODS

2.1. Test words and speakers

A total of 300 disyllabic words, 150 that are expected to have an NT in the second syllable and 150 that do not, were used as test words. The criteria for the selection of each of the 150 words included the combinations of underlying tones of the first and second syllables. According to a widely accepted theory, Wuhan Chinese has four underlying tones; therefore, there are 4×4=16 possible combinations of underlying tones in disyllable words. However, Zhou [8] suggests that in Wuhan Chinese, the T2s derived from *Yangping* tones in Middle Chinese behave differently from those from *Rusheng* tones in the NT phenomenon. Therefore, in this study, T2 was divided into those derived from *Yangping* tone (henceforth,

T2) and those derived from *Rusheng* tone (henceforth, T2′). This gives 5×5=25 combinations of underlying tones. Providing six words for each combination yields 25×6=150 words of each category.

The speakers were two native speakers of Wuhan Chinese, F (born in 1961, female) and M (born in 1964, male), with no history of living abroad. Test words were presented to the speakers in a randomized order and the speakers were asked to read each word aloud in isolation. The produced speech data were measured using the speech analysis software Praat [1].

2.2. Measurements and statistical analysis

First, the test words were acoustically analysed. Based on a visual inspection of the waveforms and spectrograms, we identified the beginning and end of the rhyme of each syllable and measured its duration. After that, we divided the rhyme section of the second syllable into 20 sections and extracted the fundamental frequency (f_0) of that section.

Since there is a speaker-specific difference in the realisation of NT words (that is, some words are realised as NT words by one speaker but non-NT words by the other), the test words were classified as NT and non-NT words by means of a non-hierarchical cluster analysis that was performed for each speaker separately based on the rhyme duration of the first and second syllables of all test words. The cluster analysis classified the test words into two groups, and we considered the group with the shorter rhyme length to be the group with NT words.

Subsequently, we categorised the pitch values of NT words by performing a hierarchical clustering analysis for each speaker based on the f_0 of the rhyme section of the second syllable. As a method for measuring the distance between test words, we adopted the Dynamic Time Warping method (DTW), which is suitable for measuring time-series data, such as f_0 contours. Ward's method was used to measure the distance between the clusters. The mean value of the f_0 contour was then calculated for each obtained cluster.

Finally, we examined the factors that determine the pitch value of NT syllables by examining the correspondence between the underlying tone of the first and second syllables and the pitch value of NTs.

3. RESULTS

3.1. Classification of NT and non-NT words

Figure 2 shows the results of the non-hierarchical clustering analysis based on the duration of rhymes in the first and second syllables of all test words. The horizontal and vertical axes are the durations of the rhymes of the first and second syllables, respectively.

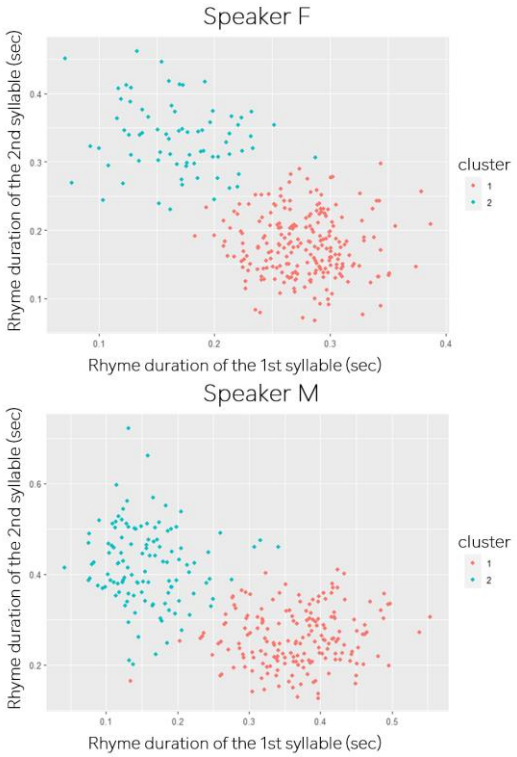


Figure 2: Results of non-hierarchical clustering analysis based on the duration of the rhymes of the 1st and 2nd syllables.

All words of both speakers were classified into Cluster 1 (red), with a longer first syllable and a shorter second syllable, and Cluster 2 (green), with a shorter first syllable and a longer second syllable. Words classified into Cluster 1 (red) were defined as NT words. The results of the non-hierarchical clustering analyses are almost consistent (with a match rate of more than 90%) with the classification based on the auditory impression of the author, who is a native speaker.

3.2. Classification of pitch values of NTs

Four interpretable clusters were extracted by the hierarchical clustering analysis based on the f_0 of the second-syllable rhyme of the NT words for each speaker. They are labelled A, B, C, and D. The max and min on the vertical axis are roughly the upper and lower limits of the pitch range of each speaker so that the gender differences are minimized.

To examine the pitch values of NTs, the mean f_0 of the rhymes in the second syllable was obtained for each cluster. Figure 3 illustrates the results for speaker F. The solid line indicates the mean f_0 for all words in each cluster, and the dotted lines indicate the mean f_0 obtained for each underlying tone of the second syllable in that cluster. In cluster A, the mean f_0 is located in the lowest region in the pitch range. In cluster B, the mean f_0 is located in the highest region.

In cluster C, the mean f_0 is located in the middle region. In Cluster D, the mean f_0 rises from the low to high regions. We refer to these pitch values of NTs respectively as L-NT, H-NT, M-NT, and R-NT.

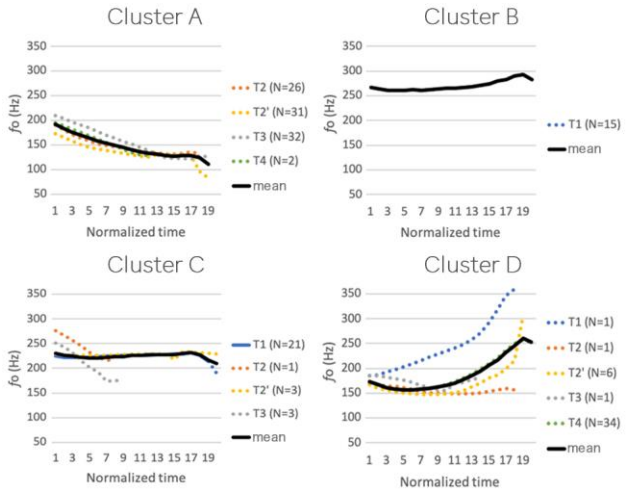


Figure 3: Mean f_0 of the rhyme in the second syllable (Speaker F).

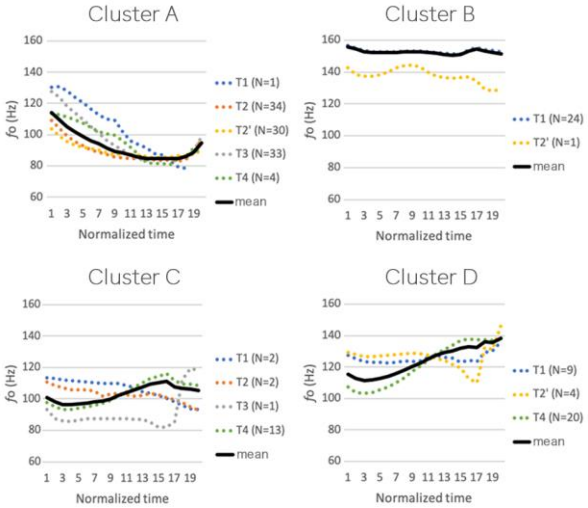


Figure 4: Mean f_0 of the rhyme in the second syllable (Speaker M).

Figure 4 shows the results for speaker M. Similar to speaker F, L-NT appears in cluster A, and H-NT appears in cluster B. However, as far as the overall average value (solid line) is examined, both clusters C and D exhibit R-NT. The difference between the two clusters is in their pitch range; that is, the mean f_0 is lower in cluster C than in cluster B. In contrast, examination of the mean f_0 for each underlying tone of the second syllable (dotted line) reveals that M-NT and R-NT coexist in clusters C and D. Importantly, only words with an underlying T4 on the second syllable exhibit R-NT. Thus, as far as speaker M is concerned, f_0-based cluster analysis could not separate M-NT from R-NT. This was likely due to clustering based on differences in the overall pitch range rather than differences in f_0 contour shape (rising or flat). The unstable phonation at the end of many of speaker M's words might have contributed to this failure.

3.3. Correspondence between underlying tones and NTs

To determine the factors conditioning the realisation of NTs, we examined the correspondence between each cluster and the underlying tone of its words.

Table 1 shows the distribution of the underlying tone of the second syllable (NT syllable) in each cluster for Speaker F. Almost all words with underlying {T2, T3} in the second syllable belong to cluster A (L-NT). Almost all words with an underlying T4 belonged to cluster D (R-NT). Words with an underlying T1 on the second syllable belong to clusters B (H-NT) and C (M-NT).

Since the realisation of NTs cannot be predicted from the underlying tone of the second syllable, we examined the underlying tone of the first syllable, finding first that all but 2 of the 15 words belonging to cluster B (H-NT) have a first syllable with an underlying T1 (5 words) or T4 (8 words), and second that all but 3 of the 21 words belonging to cluster C (M-NT) have a first syllable with an underlying T2 (7 words), T2′ (6 words), or T3 (5 words).

Since most of the words with underlying T2′ on the second syllable (31 words out of 40) belong to cluster A (L-NT), it seems reasonable to assume that they behave identically to words with T2, although the fact that a few T2′ words belong to clusters C (M-NT) and D (R-NT) deserves attention.

	T1	T2	T2'	T3	T4	Total
Cluster A (L-NT)	0	**26**	**31**	**32**	2	91
Cluster B (H-NT)	**15**	0	0	0	0	15
Cluster C (M-NT)	**21**	1	3	3	0	28
Cluster D (R-NT)	1	1	6	1	**34**	43
Total	37	28	40	36	36	177

Table 1: Breakdown of the underlying tone of the second syllable for each cluster (Speaker F).

Table 2 shows the results for Speaker M. Similar to Speaker F, almost all words with underlying {T2, T3} belong to Cluster A (L-NT). Most of the words with an underlying T1 (24 out of 36 words) belonged to cluster B (H-NT), whereas 11 words belonged to cluster C or D. As shown in Section 3.2, although M-NT and R-NT coexist in clusters C and D, words with underlying T1 on the second syllable are always realised as M-NT. Unlike speaker F, however, it is not possible to predict whether H-NT or M-NT will appear from the underlying tone of the first syllable.

As for those words with an underlying tone of T4, most belong to either clusters C or D, in which M-NT and R-NT coexist. However, words with an underlying T4 on the second syllable are only realised as R-NT (Figure 3). As in the case of T1, we cannot predict whether words with an underlying T4 belong to Cluster C (lower pitch range) or Cluster D (higher range) from the underlying tone of the first syllable. Thus, unlike the case for speaker F, we cannot confirm the involvement of the first syllable in the realisation of the NT.

Most of the words with underlying T2′ on the second syllable (30 out of 35 words) belonged to cluster A (L-NT). This leads to the assumption that they exhibit the same behaviour as words with T2, although the fact that some of them exhibit H-NT or M-NT may deserve attention.

	T1	T2	T2'	T3	T4	Total
Cluster A (L-NT)	1	**34**	**30**	**33**	4	102
Cluster B (H-NT)	**24**	0	1	0	0	25
Cluster C (M-NT, R-NT)	2	2	0	1	**13**	18
Cluster D (M-NT, R -NT)	9	0	4	0	**20**	33
Total	36	36	35	34	37	178

Table 2: Breakdown of the underlying tone of the second syllable for each cluster (Speaker M).

4. DISCUSSION

The results showed that there are four types of NTs in Wuhan Chinese, namely L-NT, H-NT, R-NT, and M-NT, and revealed that the underlying tones of both the first and second syllables determine which of the four types of NT is realized. The results can be summarised as (1).

(1) a. {T1, T2, T3, T4} + {T2, T3} → L-NT
b. {T1, T2, T3, T4} + T4 → R-NT
c. {T1, T4} + T1 → H-NT
d. {T2, T3} + T1 → M-NT

If the underlying tone of the second syllable is {T2, T3, T4} (1ab), then the underlying tone of the first syllable does not contribute to the realisation of NTs. On the other hand, if the underlying tone of the second syllable is T1 (1 cd), then the underlying tone of the first syllable contributes to NT realisation, although conditioning by the first syllable was not confirmed by the results of speaker M.

The NT phenomena make it possible to propose that in Wuhan Chinese, T2 and T3 constitute one natural class, whereas T1 and T4 form the other class. Each class can be considered to share its own distinctive tonal feature because the former has a low pitch that does not exceed the middle range in a citation form, whereas the latter has a high pitch above the middle range.

5. CONCLUSION

Wuhan Chinese has hitherto been described as having two types of NT, but the factors that determine the realisation of NTs have remained controversial [3,5,7]. The quantitative analyses of the present study based on acoustic-phonetic methods revealed that, in contrast to previous studies, there are four types of NT, and the underlying tones of both the first syllable and second syllables contribute to the realisation of NTs. Because the contribution of the first syllable's underlying tones was not detected for one of the two speakers, further studies with more participants should be conducted in the future. Our ongoing analysis of other speakers is yielding preferable results. At any rate, the previous description that there are only two types of NTs in Wuhan Chinese is not supported by the quantitative analysis of this study.

ACKNOWLEDGEMENTS

This study was supported by the NINJAL collaborative project "Empirical Study on the Intonational Diversity in Japanese and Ryukyuan Dialects" (project leader: Yosuke Igarashi).

6. REFERENCES

[1] Boersma, P., Weenink, D. 2011. Praat: Doing phonetics by computer [Computer program]. Version 5.2.16, from http://www.praat.org/.
[2] Chao, Y. R. 1930. A system of tone letters. *Le Maître Phonétique* 45, 24–27.
[3] Hou, J. Y. 1997. *Wuhanhua Yindang*. Shanghai: Shanghai Jiaoyu Chuban.
[4] Institute of Linguistics CASS. 2003. *Hanyu Fangyan Ciyu Diaocha Tiaomubiao. Fangyan*, 2003-01.
[5] Jing, S. 2022. *Xiandai Hanyu Qingsheng Dongtai Yanjiu*. Beijing: Minzu Chubanshe.
[6] Li, M. X. 2004. Neutral tones in disyllabic sequences across Chinese dialects: An OT account. Master's thesis, Tianjin Normal University.
[7] Lu, Y. Z. 2001. *Qingsheng he Erhua*, Beijing: Shangwu Yinshuguan.
[8] Zhou, D. C. 2021. Wuhan Fangyan Shuangyinjie Qingsheng De Shengdiao Tixi Yu Zhonggu Yinyun, *Proceedings of the 12th International Conference on Middle Chinese Studies*, 568–573.
[9] Zhu, J. S. 1992. *Wuhan Fangyan Yanjiu*. Wuhan: Wuhan Chubanshe.
[10] Zhu, J. S. 2017. *Wuhan Fangyan Cidian*. Wuhan: Chongwen Shuju.

TONE CHANGE IN MANDARIN: COMPARING ACOUSTIC ANALYSES FROM THE EARLY 20th CENTURY TO TODAY

Cathryn Yang
Payap University & SIL International
cathryn_yang@sil.org

ABSTRACT

This real-time study of Standard Mandarin tones presents a new acoustic analysis of Yuen Ren Chao's 1922 archival recordings and compares the results to studies from the early 21st century. Three positional variants (citation, final, and non-final) are compared across time, and Tone 2 and Tone 3 are shown to have undergone subtle changes to their contours. Two phonetic mechanisms induce contextual variants that, over time, spread to more environments and result in contour change. For both tones, carryover effects of a high preceding offset leads to raising of f0 onset and rightward sliding of f0 turning point. For Tone 3, the final rise is truncated due to the time pressure on f0 movement, which leads to contour reduction. Overall, these findings support Pittayaporn's phonetically-grounded model of tone change, which connects synchronic tone variation with diachronic tone change.

Keywords: Tone change, real-time study, Mandarin, archival recordings, phonetically-grounded sound change

1. INTRODUCTION

Phonetic tone change (i.e., change over time of a tone's f0 height and/or contour) has not received much scholarly attention and may be one of the least understood types of sound change [1]. Lack of empirical studies makes it difficult to answer basic questions, e.g., is there directionality in tone change? Real-time studies could provide valuable evidence for inferring the direction of tone change [2], but few such studies exist, e.g., [3] on Bangkok Thai. This study investigates tone change in Standard (Beijing) Mandarin in real time, by conducting a new acoustic analysis of archival recordings from 1922 [4] and comparing the results to acoustic studies from 2006 [5] and 2018 [6].

Early 20th century acoustic studies of Mandarin tone such as [7] and [8] focused exclusively on the citation form (i.e., the surface form in monosyllables uttered in isolation), without consideration of positional variants. However, tone shapes are known to vary according to the segmental and prosodic environment, including the position within an utterance [9].

In this study, a new analysis of phonograph recordings of Yuen Ren Chao (赵元任) from 1922 [4] enables comparison of three positional variants: citation form, the first syllable in a disyllabic word (non-final form), and the second syllable in a disyllabic word (final form). The 1922 data afford examination of the synchronic effects on tone contour of the preceding and following tone. Comparing contextual variants from 1922 to those seen in the early 21st century allows us to track which variants have changed in their distribution, that is, spread beyond their original conditioning environment. This study thus presents an opportunity to see links between synchronic variation and diachronic change, following the approach proposed in [3].

Mandarin has four lexical tones (Figure 1) and a neutral tone in unstressed syllables (not included in the present study) [10]. In Chao pitch numbers [11], with "1" representing low and "5" high, Tone 1 (T1) is usually transcribed as 55; Tone 2 (T2) as 35; Tone 3 (T3) as 214; and Tone 4 (T4) as 51 [10]. The above transcriptions correspond well with the 1922 tone system in Figure 1 (left panel), but less well with the system in 2006 [5] (Figure 1, right panel). T1 and T4 citation forms have not changed substantially, but the contours of T2 and T3 have changed noticeably. This study focuses on describing their changes.

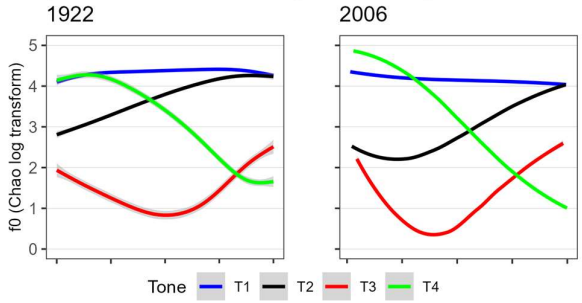


Figure 1: Left panel: Mandarin citation tones in 1922, n=187; Right panel: Mandarin citation tones in 2006, n= 2080 (adapted from [12] with permission).

2. DATA AND METHOD

Four acoustic studies, two from the early 20th century and two from the early 21st century, are compared with the new analysis from 1922. [7] from 1915 and [8] from 1934 are some of the earliest acoustic studies on Standard Mandarin citation tone. The early 21st century acoustic studies selected for comparison are

[5] from 2006 and [6] from 2018 on Beijing Mandarin. Standard Mandarin tone values are based on Beijing Mandarin pronunciation [13].

In the 2006 study [5], 10 monosyllabic words were recorded for each of the four tones. The sample included 52 speakers (balanced for gender) across a wide age range. To enhance comparability, f0 data from the 30–40-year age group (7 speakers) were selected for comparison with Chao, who was 32 at the time of the 1922 recordings. In the 2018 study [6], 50 speakers (balanced for gender; mean age = 20.9 years) were recorded reading 16 disyllabic words (1 combination of each tone pair); thus, the dataset was balanced for preceding and following tone.

In both 21st century studies, f0 was measured at 9 equidistant time points throughout the rime. Inter-speaker variation of f0 range was normalized by logarithmic transform according to the equation in (1), where *x* represents the observed f0 value, and *max* and *min* represent speaker-specific maximum and minimum f0 values [5]:

$$T = 5 * \frac{(lg\ x - lg\ min)}{(lg\ max - lg\ min)} \quad (1)$$

The subject of the present study's new analysis is Yuen Ren Chao, well-known linguist and composer. Chao was born in 1892 and spent his early childhood in northern China; as a child, he learned several Chinese lects including Mandarin [14]. A gifted scholar, Chao taught an early form of Standard Mandarin at Harvard in 1922 [13]. This early version of the national standard, replaced by the 1930's, included a fifth tone (reflex of the *ru* tone) from Nanjing Mandarin, but the other four tones followed Beijing Mandarin pronunciation [13].

The 1922 phonograph recordings, originally produced to accompany the textbook [4], were downloaded from [15]. 200 monosyllabic words (citation form) and 234 disyllabic words were selected and annotated in Praat [16]. Tokens with pitch measurement errors or with the fifth or neutral tone were excluded, leaving a total of 549 tokens. f0 of the rhyme was extracted every 10 msec. Annotated recordings and the R [17] code used to produce the relevant figures are available at [18]. f0 was normalized according the function in (1) to facilitate comparison with the 21st century studies.

The 1922 disyllabic data was not balanced for preceding and following tone: there were more tokens of the T2-T2 sequence compared to other sequences. To enhance comparability with the 2018 data, which was balanced for prosodic context, 28 T2 tokens were excluded, resulting in 9-15 tokens for each sequence. The one exception was the T2-T3 sequence, which had only 4 example words. For T3, there were 7-14 tokens for each preceding and following environment (except T2-T3).

3. RESULTS

3.1. Tone 2

T2 citation form has subtly but visibly changed over the past 100 years. In 1915 [7] and in 1934 [8] (Figure 2), T2 appears as a mid rising tone that starts its rise within the first 10% of the rhyme; there is no dip. Results from the 1922 data (Figure 3, left panel) show a similar rising contour. The similarity between the studies' results implies that the no-dip variant was standard in the early 20th century, not simply an idiosyncrasy of Chao.

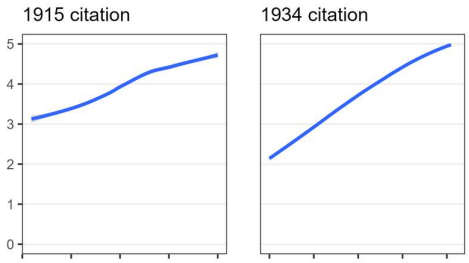


Figure 2: Left: T2 citation form in 1915, redrawn from [7] and 1934, redrawn from [8] (© The Physical Society of Japan).

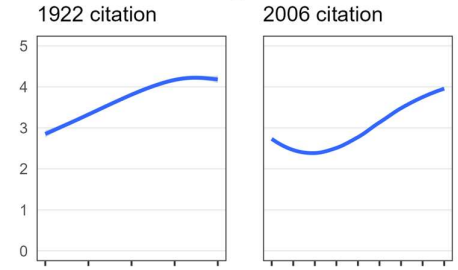


Figure 3: T2 citation form in 1922 (left, n=40); and in 2006 (right, n=70, redrawn from [5]).

In contrast, there is a noticeable difference between the early 20th century and the early 21st century (Figure 3 right panel). In the 2006 study [5], the first third of the f0 trajectory entails a small fall before the rise begins. A similar contour is seen in [19] from 1997 (not pictured here), which has a slight dip and a f0 turning point at around 25% into the syllable.

Besides citation form, Chao's 1922 recordings of disyllables afford the opportunity to compare contours in other positions. Figure 4 compares non-final and final contours in 1922 with those of the 2018 study [6]. In 2018, the f0 onset is higher and the rise begins later (at approximately 40% relative time point) compared to the early 20th century—similar to the change observed for citation tone.

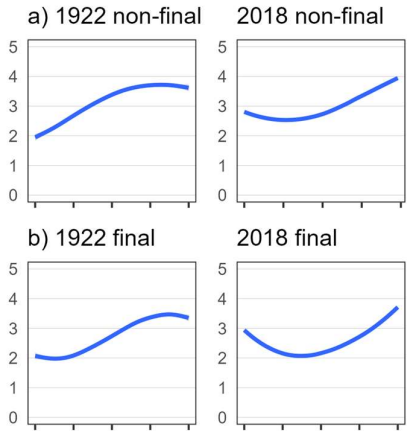


Figure 4: a) T2 non-final in 1922 (n=42) and in 2018 (n=400); b) T2 final in 1922 (n=50) and in 2018 (n=400). 2018 plots adapted from [12] with permission.

Where did the innovative dipping variant of originate from? Figure 5 compares T2 final forms in different prosodic contexts: after a tone with a high f0 offset (that is, T1, T2, and T3 sandhi form) vs. after T4, which has a low f0 offset. In the high context, T2 f0 onset is raised slightly compared to the low context; this raising is an assimilatory carryover effect of the same type reported in [19]. Also, there appears a short recovery period that results in a later f0 turning point. It is this rightward-shifted variant that has spread outside of its original conditioning environment and now appears in the 21st century citation and non-final forms.

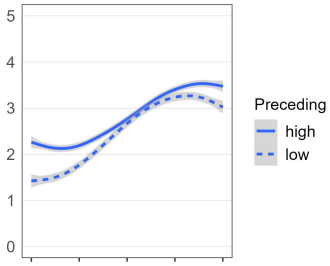


Figure 5: T2 in final position in 1922 (n=42); line type represents pitch height of the preceding tone's offset.

3.2. Tone 3

In the majority of tone descriptions from the mid-19th century to 1915, T3 is described as low rising, but between 1915 and the 1930's, descriptions vary, some reporting low rising and others low dipping [20]. Since the 1940's, most descriptions are low dipping [20]. Based on the above trends, Endo [21] suggests that T3 citation form changed from low rising to low falling-rising sometime in the early 20th century. The comparison of Figure 6 with Figure 7 supports Endo's [21] proposal. In the 1915 and 1934 studies, T3 has a "short level run" [7] and then rises—there is no dip. However, in 1922 (Figure 7 left panel), f0 onset is slightly higher (at Chao pitch number 2 rather than 1), and this raising results in an early dip before the rise begins (similar to the change in T2).

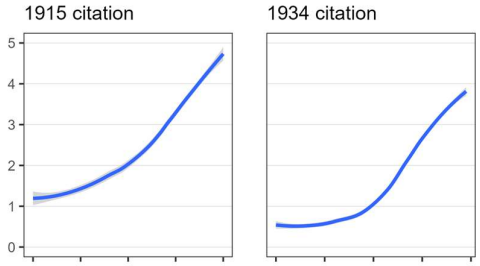


Figure 6: Left: T3 citation form in 1915, redrawn from [7] and 1934, redrawn from [8] (© The Physical Society of Japan).

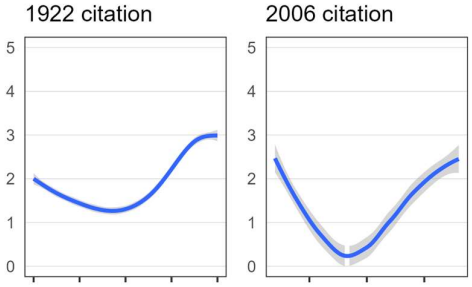


Figure 7: T3 citation form in 1922 (left, n=70) and in 2006 (right, n=70, redrawn from [5]).

The citation form in 1922 and 2006 (Figure 7 right panel) both have a dipping contour, though T3's f0 minimum in 2006 appears considerably lower. The current data is insufficient to confirm whether there was a true lowering of f0 minimum over time, but since the same normalization function was used across the compared studies, this potential change warrants further investigation, along with the potential connection between lower f0 minimum and the development of creaky voice as a phonetic correlate of T3 [22].

Differences are also apparent in non-final and final forms (Figure 8). Chao [4] describes the low falling non-final form as the "half T3" variant, because only the first half of the contour is present. Contour simplification in non-final position is often due to shorter duration [23]. In 1922, mean duration of non-final form is 78% of that of the final form. Shorter duration leads to increased time pressure on f0 [24]. The 1922 non-final form shows both target undershoot (raised f0 minimum) and truncation (cut-off of the rise) compared to citation form. The half T3 variant appears only in non-final form in 1922, but has spread to final form in 2018. Xu's [19] study of disyllabic sequences in 1997 report a similar spreading of the non-final allotone to final position.

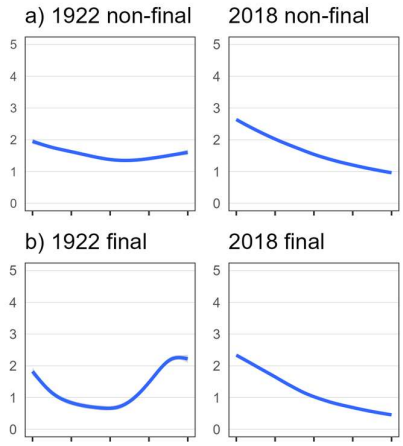


Figure 8: a) T3 in non-final position in 1922 (n=37) and in 2018 (n=350); b) in final position in 1922 (n=34) and in 2018 (n=400, adapted from [12] with permission).

In addition, T3 f0 onset has raised over time: below Chao level 2 in 1922 but consistently higher than level 2 in 2018. In 1922, T3 raised onset is observed after a preceding high offset, as seen in Figure 9. This variation is similar to the results for T2, implying that the same conditioning factors, i.e., preceding tone context, affected both T2 and T3.

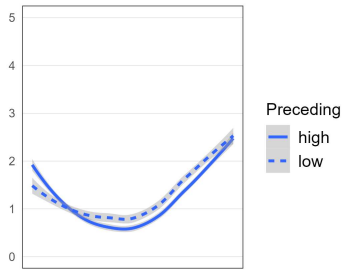


Figure 9: T3 in final position in 1922 (n=34).

4. DISCUSSION

The observed changes in Mandarin show similarity to the changes reported for Bangkok Thai [3], seen in Figure 10. Over the past 60 years, Thai Tone 4 (High) has gradually developed a slight dip or level period before its rise, similar to Mandarin T2. Thai Tone 5 (Rising) is gradually losing its final rise, similar to Mandarin T3. These similarities are not shared innovations or caused by language contact; instead, it is likely that the same phonetic biases are at play in both languages.

From a typological perspective, the observed changes follow several crosslinguistic trends in tone change: 1) raising of f0 onsets that are below mid, 2) rightward sliding of f0 turning points, and 3) reduction of low rising contour [25]. Raising of low f0 onsets is a well-attested carryover effect of high preceding offset in both Thai [26] and Mandarin [19]. Carryover effects also result in a short recovery period before the initial tone target is reached, causing a later temporal location of the f0 turning point. In the case of T3, shorter duration in the non-final position results in truncation of the rise. This study shows that the variants with raised f0 onset, later alignment of f0 turning points and/or reduced contour have spread outside their original conditioning environment. The change over time is a result of these variants' change in distribution.

What phonetic biases underlie these changes? Pittayaporn [3] points to time pressure on f0 contour as a generator of synchronic variants that may then spread to other environments. Xu [27] identifies two time constraints on f0: 1) minimum duration for f0 movement and 2) synchronization of the tone with the segmental string. In other words, it takes a certain amount of time to make f0 transitions, but there is also a limit on how much time is available, depending on the tonal domain. These articulatory constraints are responsible for time pressure on f0 movement, and their diachronic consequences can be seen in crosslinguistic trends in tone change.

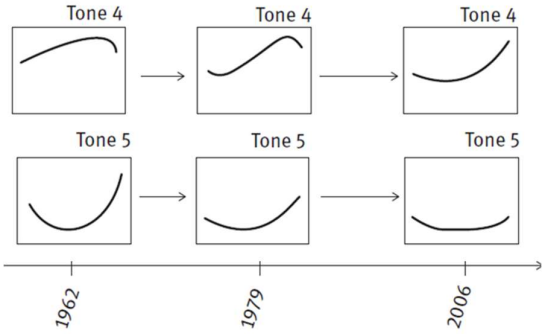


Figure 10: Bangkok Thai citation tone over the past 60 years. Adapted from [3] with permission.

5. CONCLUSION

This study conducts tonal acoustic analysis on Yuen Ren Chao's recordings from 1922 and compares the analysis to acoustic studies from 1915 [8], 1934 [9], 2006 [6] and 2018 [7]. The differences between the 20th century studies and the 21st century studies for T2 and T3 are interpreted as reflecting tone change over time; of course, more data from the early 20th century is needed to confirm this. In the 1922 data, certain prosodic environments condition a raised f0 onset and later temporal location of the f0 turning point and/or a truncated contour. The raised-onset, rightward-shifted and/or truncated variants spread beyond the original conditioning environment by the early 21st century, resulting in contour change. This study thus reveals a link between synchronic variation and diachronic change, evidence that supports Pittayaporn's [3] phonetically-grounded tone change model.

6. ACKNOWLEDGEMENTS

Sincere thanks to Elissa Low for segmentation and labelling. This project was supported by SIL International.

7. REFERENCES

[1] E. W. Campbell, “Why is tone change still poorly understood, and how might documentation of less-studied tone languages help?,” in *Historical linguistics and endangered languages: Exploring diversity in language change*, P. Epps, D. Law, and N. Pat-El, Eds., Routledge, 2021, pp. 15–40.

[2] J. Tillery and G. Bailey, “Approaches to real time in dialectology and sociolinguistics,” *World Englishes*, vol. 22, no. 4, pp. 351–365, 2003, doi: 10.1111/j.1467-971X.2003.00305.x.

[3] P. Pittayaporn, “Phonetic and systemic biases in tonal contour changes in Bangkok Thai,” in *Tonal change and neutralization*, H. Kubozono and M. Giriko, Eds., in Phonology and Phonetics, vol. 27. Berlin: Mouton de Gruyter, 2018, pp. 249–278.

[4] Y. R. Chao, *Guoyu liushengpian keben (國語留聲片課本) [A textbook of the national language with gramophone records]*. Shanghai: Commercial Press, 1922.

[5] F. Shi and P. Wang, “Beijing hua danziyin shengdiao de fenzu tongji fenxi (北京话单字音声调的分组统计分析) [A statistical analysis of tone groups in Beijing Mandarin],” *Dangdai Yuyanxue*, vol. 8, no. 4, pp. 324–333, 2006.

[6] Y. Zhang, “Beijing Putonghua de shengdiao shiyan fenxi (北京普通话的声 调实验分析) [An experimental analysis of the tones of Putonghua in Beijing],” PhD dissertation, Beijing Language and Culture University, Beijing, 2018.

[7] C. B. Bradley, “The tone-accents of two Chinese dialects,” *J. Am. Orient. Soc.*, vol. 35, pp. 199–206, 1915, doi: 10.2307/592645.

[8] S. Obata and T. Toshima, “Shinago no butsuri oto koe manabu teki kenkyū: (sore ichi) shi koe no seishitsu (支那語の物理音聲學的研究 (其一) 四聲の性質) [Acoustic study of Chinese (Part 1: Tone)],” *J. Math. Phys. Soc. Jpn. 日本数学物理学会誌*, vol. 8, no. 1, pp. 1–10, 1934, doi: 10.11429/subutsukaishi1927.8.1.

[9] C. DiCanio, J. Benn, and R. Castillo García, “Disentangling the effects of position and utterance-level declination on the production of complex tones in Yoloxóchitl Mixtec,” *Lang. Speech*, vol. 64, no. 3, pp. 515–557, 2021, doi: 10.1177/0023830920939132.

[10] Y. R. Chao, *A grammar of spoken Chinese*. Berkeley: University of California Press, 1968.

[11] Y. R. Chao, “ə sistim əv ‘toun-letəz,’” *Maître Phon.*, vol. 8 (45), no. 30, pp. 24–27, 1930.

[12] F. Shi, “Shilun Putonghua shengdiao de bendiao: Jiantan wuduzhi jidiaofa de xingzhi (试论普通话声调的本调: 兼谈五度值记调法的性质) [On the essential tones in Standard Chinese: Concurrent discussion on the nature of five-degree method],” *Int. J. Chin. Linguist.*, vol. 7, no. 1, pp. 140–157, Jun. 2020, doi: 10.1075/ijchl.00008.shi.

[13] Y. R. Chao, “What is correct Chinese?,” *J. Am. Orient. Soc.*, vol. 81, no. 3, pp. 171–177, 1961, doi: 10.2307/595651.

[14] R. Levenson, “Chinese linguist, phonologist, composer and author, Yuen Ren Chao,” 1974. Accessed: Oct. 29, 2022. [Online]. Available: https://oac.cdlib.org/view?docId=hb8779p27v&query=&brand=calisphere

[15] “Lao guoyin xuexi (老国音学习) [Study of the Old Standard Mandarin pronunciation],” 2022. https://zh.wikiversity.org/wiki/老國音學習 (accessed Oct. 18, 2022).

[16] P. Boersma and D. Weenick, “Praat: doing phonetics by computer.” 2022. [Online]. Available: https://www.praat.org

[17] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org

[18] Yang, Cathryn, “CAYang-blip/Tone-change: v1.0.” Zenodo, Jan. 2023. doi: 10.5281/zenodo.7509125.

[19] Y. Xu, “Contextual tonal variations in Mandarin,” *J. Phon.*, vol. 25, no. 1, pp. 61–83, Jan. 1997, doi: 10.1006/jpho.1996.0034.

[20] X. Chen, “Qing zhonghouqi zhi Minguo chuqi Beijing hua de shengdiao bianhua (清中后期至民国初期北京话的声调变化) (Tone change in Beijing dialect from mid-late Qing to early Republic),” *Zhongguo Yuwen*, vol. 2018, no. 04, pp. 435–445, 2018.

[21] M. Endo, “Bidirectional change in tone: Evidence from Chinese,” *Econ. Rev.*, vol. 12, pp. 1–29, 2020.

[22] J. Kuang, “Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice,” *J. Acoust. Soc. Am.*, vol. 142, pp. 1693–1706, Sep. 2017, doi: 10.1121/1.5003649.

[23] J. Zhang, *The effects of duration and sonority on contour tone distribution — A typological survey and formal analysis*. London: Routledge, 2002.

[24] J. Hanssen, J. Peters, and C. Gussenhoven, “Responses to time pressure on phrase-final melodies in varieties of Dutch and West Frisian,” *J. Phon.*, vol. 93, p. 101150, Jul. 2022, doi: 10.1016/j.wocn.2022.101150.

[25] C. Yang and Y. Xu, “Crosslinguistic trends in tone change: A review of tone change studies in East and Southeast Asia,” *Diachronica*, vol. 36, no. 3, pp. 417–459, 2019.

[26] J. Gandour, S. Potisuk, S. Dechongkit, and S. Ponglorpisit, “Tonal coarticulation in Thai,” *J. Phon.*, vol. 22, pp. 477–492, 1994.

[27] Y. Xu, “Timing and coordination in tone and intonation—An articulatory-functional perspective,” *Lingua*, vol. 119, no. 6, pp. 906–927, Jun. 2009, doi: 10.1016/j.lingua.2007.09.015.

INTERMINGLING TONE SYSTEMS: THE RELATIONSHIP OF NANNING MANDARIN TO NANNING CANTONESE AND STANDARD MANDARIN

Xinzi Hou[1], Liang Zhao[1], Eleanor Chodroff[2]

[1)]Department of Language and Linguistic Science, University of York, York, United Kingdom
[2)]Department of Computational Linguistics, University of Zürich, Zürich, Switzerland
xinzi.hou@york.ac.uk, liang.zhao@york.ac.uk, eleanor.chodroff@uzh.ch

ABSTRACT

Nanning Mandarin, the lingua franca of a multilingual city in China, presents a special case of lexical tone development arising from language contact between Nanning Cantonese and Standard Mandarin. Previous research has suggested that the realisation of Nanning Mandarin tones derives from the transfer of Nanning Cantonese to Standard Mandarin. However, this research was limited to auditory methods, differed in reports of the tone system, and lacked a direct acoustic comparison of all three language varieties. In the present study, we conducted an acoustic-phonetic analysis of Nanning Mandarin tones from 30 talkers, and directly compared the f0 contours of Nanning Mandarin with those in Nanning Cantonese (48 talkers) and Standard Mandarin (21 talkers). Though the four-tone Nanning Mandarin tone inventory shows a resemblance to Standard Mandarin, the phonetic realisation of Nanning Mandarin tones exhibits considerable similarity to that of the primary source language, Nanning Cantonese in f0 height and contours.

Keywords: lexical tone, dialects, intermingling communities, Mandarin, Cantonese

1. INTRODUCTION

Language contact creates a unique environment for tone systems to develop. As a multilingual city, Nanning, China presents one such environment for investigating the relationships between lexical tone systems. Nanning is the capital of the Guangxi Zhuang Autonomous Region in southern China. Traditionally, the dominant language of the city has been Nanning Cantonese, a Sinitic language, but the city has also had a relatively high number of Pinghua (Sinitic) and Zhuang languages (Tai) speakers. Since the 1950s, Standard Mandarin (or Putonghua, also a Sinitic language) has been officially promoted as the lingua franca throughout China and is the required language in education, news media and government affairs. Moreover, Nanning has received an influx of immigrants with different language backgrounds from within and beyond Guangxi, providing further reason to use the lingua franca. In the 1990s, the strong nationwide promotion resulted in the widespread prevalence of a distinct Mandarin dialect, namely Nanning Mandarin, which has become the current lingua franca in the region.

Nanning Mandarin (NM) is the local version of Standard Mandarin (SM) produced by native speakers of the traditional regional languages (mainly Nanning Cantonese, NC) speakers. It is usually categorised as a type of Local Mandarin (or Local Putonghua) [4] or an interlanguage produced by L1 Nanning locals in approximating Standard Mandarin as an L2 language [1]. While Nanning Cantonese was the primary L1 for older generations, Nanning Mandarin is now frequently acquired as the native language. Though Nanning Mandarin retained the four tone categories from its target language (Standard Mandarin), the phonetic realisation of those tones has been noted to be unique.

Few studies have directly examined the lexical tone systems of Nanning Mandarin, and they mostly rely on auditory transcription. One study reported the tone inventory of the interlanguage between Nanning Cantonese and Standard Mandarin, which was presumably Nanning Mandarin; however, the methodology was not made explicit [12]. In addition, Fan [21] conducted an auditory analysis of the tone inventory in Nanning Mandarin from one older and one younger speaker, and found that the differences between generations were minor. Both studies identified a level tone for NM Tone 1, a rising tone for NM Tone 2, and a low falling tone for NM Tone 3; however, they differed in the description of NM Tone 4, as well as the exact Chao tone numbers assigned to many of these tones. Two additional studies have indirectly examined Nanning Mandarin through an auditory comparison of Nanning Cantonese and Standard Mandarin [9, 12].

Overall, previous studies have found that Nanning Mandarin has a narrower pitch range than Standard Mandarin and the absence of a dipping tone [9, 10, 12, 18]. A similar observation was reported in an intonational study of Nanning Mandarin [10]. These studies conjectured that the identified patterns can be explained as a partial transfer of the phonetic tone realisation from Nanning Cantonese to Mandarin. All four studies indicated intermediate and unstable

phonetic realisations between Nanning Cantonese and Standard Mandarin [23]. Nevertheless, discrepancies still remain regarding the precise characterisations of the NM tone system.

In the present study, we extend previous research on the NM lexical tone system by directly examining the relationship of the NM tone system to the NC and SM tone systems. In contrast to previous studies, we assess these relationships using acoustic-phonetic analysis from a large number of NM, NC, and SM speakers. The presented research aims to demonstrate the acoustic representations of tone in Nanning Mandarin and conduct a comparison of these with the historically dominant source language, Nanning Cantonese, and the target language, Standard Mandarin. This study sheds new light on the research on tone in language contact environments by expanding the documentation of tonal languages with systematic descriptions of two understudied Sinitic varieties, Nanning Mandarin and Nanning Cantonese. The findings further our understanding of tone in complex language contact, as well as source-language effects on interlanguages.

2. METHODS

2.1. Materials

For Nanning Mandarin and Nanning Cantonese, a wordlist containing 72 monosyllabic words and 72 disyllabic words was designed based on the diachronic correspondence between Mandarin and Cantonese tone categories. Nanning Mandarin and Standard Mandarin have four citation tones, while Nanning Cantonese has six citation tones that have historical correspondences with Mandarin tones [11, 19]. The present study focuses on the monosyllabic words alone, so only these will be described in detail. The 72 monosyllabic words cover all four tone categories in Nanning Mandarin with 12 to 24 words for each NM tone and all six tone categories in Nanning Cantonese with 12 words for each NC tone.

All words are shared between Nanning Mandarin and Nanning Cantonese and recognised in spoken language, though a few sound old-fashioned. The words in both languages have similar phonetic contexts with a V, CV or CVN structure. Examples in Nanning Mandarin and Nanning Cantonese include 姑 '*aunt*' /ku^{55}/–/ku^{55}/; 有 '*have*' /jɐu^{24}/–/jou^{214}/. For clearer segmentation and to minimise reading errors, syllables with multiple readings were disambiguated with a disyllabic context and instructions for the participant to only read the target word.

For Standard Mandarin, ten unique syllables were used to create a wordlist of 40 monosyllabic words, with each syllable realised in one of the four SM lexical tone categories (see [24] for more detail).

2.2. Participants

For Nanning Mandarin and Nanning Cantonese, 52 native speakers were recruited; 48 were bilingual speakers and four were Nanning Mandarin monolingual speakers. With a few exceptions, speakers were born and raised in Nanning; started speaking Nanning Mandarin or Nanning Cantonese before the age of 12; had not stayed outside of Nanning for more than one year within the past ten years; did not have any explicit training or test in Standard Mandarin; and did not have close family members or friends who spoke Northern Mandarin dialects. Participants who did not meet the above criteria went through additional validation of their NM accent for further exclusion (see Section 2.4).

Participants were roughly balanced for gender (27 females and 25 males) and age (12 each for age 24–39, 40–49 and 50–59, 16 for age 60–69). Speakers aged 30 to 60 formed the majority because they were less likely to have settled elsewhere in recent years and have had less exposure to Standard Mandarin at school. Therefore, they were less affected by other languages or dialects than the younger speakers.

For Standard Mandarin, 21 fluent speakers of Standard Mandarin (13 females and eight males aged 23–41) were recruited from six Mandarin-dominant cities of China [24]. Though these cities have their own regional dialects, the participants reported an average score of 4.14 on a 1–5 scale (5 as native-like fluency) on their fluency in Standard Mandarin. Each audio used for analysis was also manually checked for its representativeness of Standard Mandarin.

2.3. Procedure

For Nanning Mandarin and Nanning Cantonese, recordings were conducted in person in a quiet room or booth in a location convenient for the participant in Nanning. The instructions and wordlist were presented in Simplified Chinese on a Macbook Pro via PsychoPy v.2021.2.3 [20]. The production was recorded using Awesome Voice Recorder X v2.0.4 on an iPhone 7 smartphone. The audio was recorded in WAV format with a sampling rate of 22,050 Hz and a bit rate of 320 kbps. Before the study, participants completed a demographic questionnaire, and received the instructions and overview of the wordlist. This was then followed by five practice words.

The recording involved two repetitions of four blocks: monosyllabic and disyllabic words in each of Nanning Mandarin and Nanning Cantonese. Language order was counterbalanced across bilingual participants, and word order was randomised within

each block. Target words were presented one by one on the screen; the participant controlled the progression by pressing any button on the keyboard. Participants were allowed to repeat a token if needed for correction. In total, 7486 monosyllabic NM and 6905 monosyllabic NC words were collected.

For Standard Mandarin, recordings were collected remotely via participant-controlled smartphones. Participants were instructed to install a specific recording app and record themselves while reading the words presented on the Gorilla Experiment Builder [19]. Words were presented trial by trial and fully randomised (see [24] for more detail).

2.4. Data processing

Recordings were automatically segmented and aligned using Praat [8] and the Montreal Forced Aligner (MFA) [18] at the syllable and phone levels. Given the formal register of reading, it was possible that some speakers produced tokens with a strong SM accent. Validation of the naturalness of the NM-accent was performed for 26 speakers based on their demographic background and an initial auditory evaluation was performed by the first author. Six native NM auditors listened to three to five random tokens from each of these speakers and rated their accents on a scale from 1 to 5. NM speakers rated as 4 or above were retained in the acoustic analysis and 20 NM speakers were thus excluded.

For all three language varieties, f0 values were automatically extracted from 11 equidistant points in the sonorant portion of each syllable. The first point was excluded as its value was highly affected by the preceding consonant and was frequently undefined. After averaging by speaker and tone, the f0 values were standardised by speaker-specific f0 ranges for each language, and then transformed into the tone values in Chao tone numbers ranging from 0 to 5, using the Log T-value formula [2]:

$$LT = \frac{\log_{10} mean - \log_{10} min}{\log_{10} max - \log_{10} min} \times 5 \quad (1)$$

Reference to Chao tone numbers in the paper thus reflects the LT value. Tokens with at least six usable f0 values and speakers with at least 72 (50%) usable tokens were retained for acoustic analysis (two speakers were excluded). After a combination of validation tasks, there were 6365 tokens from 48 NC speakers (24 female, 24 male) and 3732 tokens from 30 NM speakers (14 female, 16 male).

3. RESULTS

Figure 1 displays the f0 contours in Chao tone numbers for each tone and language variety; Table 1 shows the Chao tone numbers for each tone and language variety. Both varieties of Mandarin have four phonological tones, whereas Cantonese has six phonological tones, with a likely tone merger between Tone 2 and Tone 6.

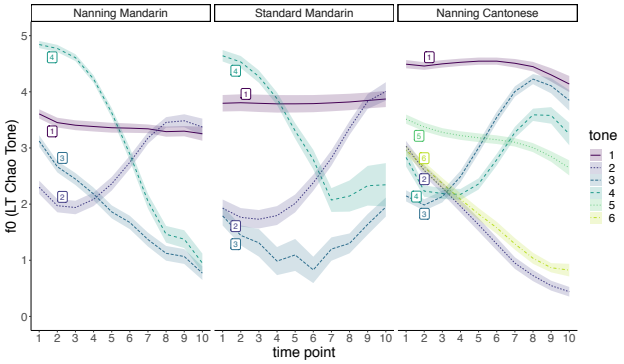


Figure 1: f0 contours in Chao tone numbers of Nanning Mandarin, Standard Mandarin, and Nanning Cantonese for each phonological tone. Each ribbon represents 1 standard error from the mean.

Tone	Nanning Mandarin	Standard Mandarin	Tone	Nanning Cantonese
1	44	44	**1**	55
2	34	24	**2**	31
3	41	212	**3**	35
			4	34
4	52	53	**5**	43
			6	31

Table 1: Chao tone numbers for each phonological tone in Nanning Mandarin, Standard Mandarin, and Nanning Cantonese. Tones are aligned according to their historical correspondence.

At a glance, the NM tone–contour mappings reflect those of Standard Mandarin, but with a noticeable lack of dipping in NM Tone 3 and a considerably higher onset in NM Tone 4. When we zoom out to the use of the f0 space, however, Nanning Mandarin appears to contrast with Standard Mandarin and shares greater resemblance to Nanning Cantonese. Unlike Standard Mandarin where a High–Low contrast is utilised, Nanning Mandarin displays a High–Mid–Low contrast, most saliently at the onset, resembling Nanning Cantonese. Specifically, NM Tone 4 begins in the high range, NM Tone 1 and Tone 3 begin in the middle range, and NM Tone 2 begins in the low range.

To investigate whether individual NM tones share greater similarity with Standard Mandarin or Nanning Cantonese, we examined the overall similarity of each NM f0 contour to similarly-shaped NC and SM f0 contours using the root-mean-squared

deviation (RMSD) between the averaged f0 contours from all ten timepoints in Chao tone numbers. Lower RMSD values correspond to a higher similarity between contours. As shown in Table 2, all four NM f0 contours were most similar to NC f0 contours, although some were closely followed by SM tones.

To further investigate differences in the exact contour independent of f0 height and any potentially remaining individual variation, the RMSD values were also calculated after *z*-scoring the raw f0 in hertz by speaker and tone. The results still indicated that all four NM tones shared greatest similarity to NC tones, though the falling SM Tone 4 became substantially more similar to NM Tone 3 and Tone 4.

NM Tone	Tone Comparisons			
NM1 (Level)	**NC1**	**NC5**	**SM1**	
	1.12	0.26	0.50	
	0.56	0.27	0.70	
NM2 (Rising)	**NC3**	**NC4**	**SM2**	
	0.59	0.20	0.38	
	0.23	0.19	0.39	
NM3 (Mid-fall)	**NC2**	**NC6**	**SM3**	**SM4**
	0.30	0.26	1.11	1.40
	0.12	0.11	1.05	0.32
NM4 (High-fall)	**NC2**	**NC6**	**SM3**	**SM4**
	1.77	1.72	2.43	4.00
	0.25	0.23	1.13	0.38

Table 2: RMSD values on the Chao tone numbers listed first in each cell, and on the *z*-scored values, second. Each row presents a comparison of an NM tone to tones with a similar contour in Nanning Cantonese and Standard Mandarin.

4. DISCUSSION

Overall, the lexical tone system of Nanning Mandarin resembled Nanning Cantonese in implementing a three-level tone system with High–Mid–Low contrasts in the phonetic space, in line with transcriptions in previous studies [9, 10, 12, 18, 20]. Moreover, individual contours were most similar to those of Nanning Cantonese in both Chao tone numbers and *z*-scored hertz, but closely followed by some SM tones. This suggests a clear influence of Nanning Cantonese on Nanning Mandarin.

Nevertheless, individual NM tones still approximated those of Standard Mandarin in f0 contours. Though NM Tone 3 did not dip like SM Tone 3, it did exhibit a similar initial fall. In Standard Mandarin, the realisation of Tone 3 is highly contextual, and critically only dips in phrase-final position [8, 13]. In Nanning Mandarin, it may be that only the core portion of Tone 3 was realised without the boundary rise. The absence of the Tone 3 prepausal rise resulted in two falling tones (NM Tone 3 and Tone 4). The dispersed f0 onsets and different slopes of the two falling tones could reflect a need for tonal contrast. The dispersed f0 onsets also correspond to the mid and high levels of the three-level tone system that is common to Nanning Cantonese, but distinct from Standard Mandarin.

With respect to Nanning Cantonese, we also observed a merger of Tone 6 (traditionally low-level [3–6, 11, 13, 15]) with Tone 2 (mid-falling), consistent with findings in previous auditory [3, 5, 12] and acoustic studies [15, 17]. This tone merger resulted in a lack of a low-level tone in the inventory, potentially accounting for why NM Tone 3 was not realised even as the non-phrase-final SM Tone 3 (21 contour). The falling NC Tone 6 could have potentially served as the initial template for NM Tone 3, giving rise to a mid- instead of low-falling tone (41 contour).

The findings of both Nanning Mandarin and Nanning Cantonese show similarity to the data in previous literature, but with some minor differences in tone values and contours. Results of the three language varieties show that Nanning Mandarin, as a variety of Standard Mandarin, has presented a hybrid tone system with a clear and large influence from Nanning Cantonese in its approximation to Standard Mandarin. Whether these similarities and differences are significant and affect pitch perception, however, remains to be tested.

5. CONCLUSION

The present study investigated the lexical tone system of Nanning Mandarin in comparison to Nanning Cantonese and Standard Mandarin. The results showed that the NM tone–contour mappings were clearly inherited from Standard Mandarin, but displayed considerable similarity to and influence from Nanning Cantonese. Overall, this study provided acoustic data of two understudied Sinitic language varieties in Nanning. The findings set the basis for further investigation of the relation between tone and intonation in Nanning Mandarin, as well as variation in tone realisation arising from complex language contact, from Nanning to a broader range of world languages.

6. ACKNOWLEDGEMENTS

The authors would like to thank Chenzi Xu, Dom Watt, and the anonymous reviewer for helpful feedback and suggestions. We also thank the NM auditors for their help with data annotation. The final author was supported by SNSF grant 208460.

7. REFERENCES

[1] L. Selinker, 'Interlanguage', *International Review of Applied Linguistics in Language Teaching*, vol. 10, no. 1–4, pp. 209–232, Jan. 1972.

[2] F. Shi, '天津方言双字组声调分析 [Analysis of bi-syllabic tones in Tianjin dialect]', *语言研究 [Studies in Language and Linguistics]*, no. 01, pp. 77–90, 1986.

[3] J. Hong, '南宁白话的语音特点 [Phonetic Features of Nanning Cantonese]', *广西师范大学学报(哲学社会科学版) [Journal of Guangxi Normal University (Philosophy and Social Science Edition)]*, no. 01, pp. 47–54, 1989.

[4] Y. Yao, '应该开展对"地方普通话"的研究 [Research Should be Conducted on "Local Standard Mandarin"]', *语文建设 [Language Planning]*, no. 03, pp. 20–23, 1989.

[5] J. Xie, '南宁白话同音字汇 [Nanning Cantonese homophone glossary]', *方言 [Fangyan]*, no. 04, pp. 286–303, 1994.

[6] H. Yang, *南宁话音档 [Sound archive of Nanning languages]*. Shanghai: Shanghai Educational Publishing House, 1997.

[7] 广西壮族自治区地方志编纂委员会 [Guangxi Zhuang Autonomous Region Local Documentation Compilation Committee], Ed., *广西通志·汉语方言志 [Comprehensive documentation of Guangxi: Sinitic dialects]*. Nanning: Guangxi People's Press, 1998.

[8] P. Boersma, 'Praat, a system for doing phonetics by computer', *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[9] M. Yip, *Tone*. Cambridge University Press, 2002.

[10] H. Yang, '南宁普通话的语调、语气词及其声学表现 [Intonation and Modal Particles Used in Nanning Putonghua and Their Manifestations in Acoustics]', *语言研究 [Study in Language and Linguistics]*, vol. 23, no. 04, pp. 104–109, 2003.

[11] Y. Liang, '"南宁普通话"韵母研究 [A Study of the Finals in Nanning People's Oral Mandarin]', Master, Guangxi University, 2005.

[12] J. Yu, Ed., *广西推广普通话的理论与实践研究 [A Study of the theory and practice of promoting Putonghua in Guangxi]*. Shanghai: Shanghai Educational Publishing House, 2008.

[13] Y. Lin and F. Qin, *广西南宁白话研究 [A study of Cantonese in Nanning, Guangxi]*. Guilin: Guangxi Normal University Press, 2008.

[14] F. Shi and Q. Ran, '普通话上声的本质是低平调——对《汉语平调的声调感知研究》的再分析 [A Comment on "Mandarin Tone Perception: A Report on a Low Level Tone"]', *中国语文 [Studies of the Chinese Language]*, no. 6, pp. 550-555+576, 2011.

[15] T. Li, '浅谈南宁白话对南宁方言普通话的声调影响 [A brief discussion of the impact of Nanning Cantonese on the tones of Nanning dialect Mandarin]', *剑南文学(经典阅读) [Jiannan Literature (Classics Reading)]*, no. 2, pp. 97–98, 2012.

[16] T. Luo, '南宁白话单字调实验语音研究 [An Experimental Study on the Mono-Syllable of Nanning Cantonese]', Master, Guangxi University, 2015.

[17] K. Wang, '南宁白话(新派)单字调声学实验研究 [An acoustic experimental study on the monosyllabic tones in Nanning Cantonese from the younger generation]', *语文学刊 [Journal of Language and Literature Studies]*, no. 04, pp. 37–39, 2015.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, 'Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi', in *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017, pp. 498–502.

[19] A. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. Evershed, 'Gorillas in our Midst: Gorilla.sc, a new web-based Experiment Builder'. bioRxiv, p. 438242, 22-Oct-2018.

[20] J. Peirce *et al.*, 'PsychoPy2: Experiments in behavior made easy', *Behav Res*, vol. 51, no. 1, pp. 195–203, Feb. 2019.

[21] F. Fan, '南宁普通话语音研究 [Nanning Mandarin Phonetic Research]', Master, Guangxi University, 2020.

[22] X. Li and M. Xiang, *汉语方言学基础教程 [Basic tutorials in Chinese dialectology]*, 2nd ed. Beijing: Peking University Press, 2020.

[23] J. Lyu and J. Yu, '多语接触与通语变异——广西"南普"形成及变异研究之考察 [Multi-language Contact and Standard Language Variation: A Study on the Form and Variation of "Nanpu" in Guangxi]', *西南交通大学学报(社会科学版) [Journal of Southwest Jiaotong University (Social Science)]*, vol. 22, no. 01, pp. 47–55, 2021.

[24] L. Zhao and E. Chodroff, 'The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1985–1990.

DISYLLABIC TONES IN MANDARIN PRESCHOOL CHILDREN AND CHILD-DIRECTED SPEECH

Zhiwei Wang[1], Jun Gao[2], Aijun Li[2]

[1]Beijing International Studies University (BISU), China
[2]Institute of Linguistics, Chinese Academy of Social Sciences, China
2549066586@qq.com gao-jun@cass.org.cn liaj@cass.org.cn

ABSTRACT

This study investigates the developmental performance of disyllabic tones in Mandarin-speaking preschoolers (aged 3-6) and child-directed speech (CDS). Comparing the acoustic parameters of tones in disyllabic words produced by preschoolers in four age groups and CDS, we found that: 1) the preschoolers have already acquired the tone sandhi rules phonologically; 2) their phonetic accuracy is over 95% after 4 years; the error rate of the first syllable is higher than those on the second, declining significantly with age; more errors are found in Tone 2 and Tone 3 than Tone 1 and Tone 4; 3) preschoolers' off-standard production is mainly characterized by the 'undershoot' of tonal target regarding the slope and curvature of the pitch contour; 4) pitch register, pitch range, and duration of disyllabic tones show significant difference between child-Mom dyads, but exhibit consistent developmental patterns between preschoolers and CDS, which reflects a synchronous entrainment.

Keywords: Lexical tone acquisition, Disyllabic tones, Mandarin-speaking preschoolers, child-directed speech

1. INTRODUCTION

Mandarin Chinese has four lexical tones: Tone 1 (T1, high level, HH), Tone 2 (T2, rising, LH), Tone 3 (T3, low rising, LLH), and Tone 4 (T4, falling, HL), transcribed as 55, 24, 213, and 51 in a 5-point scale [1]. The T3T3 disyllabic tone in Mandarin undergoes a phonological tone sandhi rule whereby T3T3 is realized as T2T3. When positioned in the first syllable, T3 and T4 are frequently realized as half-T3 (21) before T1\T2\T4 and half-T4 (53) before T4 due to tone coarticulation. The neutral tone in Mandarin (T5), is typically short in duration and found in the final unstressed syllable. The pitch contour of the neutral tone may vary with the preceding tones [2]. So, for normal stressed words the phonological and phonetic variations of the tone on the first syllable in disyllables are more complex than the tone on the second. Although many studies have examined Mandarin tones and tone sandhi, children's tone production and the relation between children's production speech (CPS) and child-directed speech (CDS) have been insufficiently investigated.

Previous auditory-based research suggested that Mandarin-speaking children acquire tones earlier than other linguistic elements by the age of 1.5 [3, 4]. The analysis of a children's cross-sectional corpus (CASS_Child_Word) revealed that at 18 months (1.5 yrs), Mandarin-speaking children exhibited a high tone production error rate (around 21%) in each syllable context, which decreased with age and dropped to around 2% at 42 months (3.5 yrs) [5].

However, research using the method of acoustic speech analysis yields different results [6-8]. Three-year-old children are found to be capable of perceiving T1, T2, T4 accurately, but their accuracy of tone production is much lower than adults [6]. There are also clear distinctions between children's and adults' tone production of monosyllabic words [7]. The acquisition of monosyllabic tones in Mandarin is suggested to be incomplete even at the age of 4-5 [8]. Adult-like tone (T3 and T4) and sandhi tone production is mastered after the age of 5 [9, 10].

When talking to young children, people employ a genre of child-directed speech (CDS) that is markedly different from adult-directed speech (ADS). Two hypotheses were proposed to explain the features of CDS against ADS: the hyperarticulation hypothesis [11] and the pragmatic hypothesis [12]. The former suggests that mothers enhance phonemic contrasts to facilitate speech and language acquisition in children. The latter hypothesis proposes that the acoustic differences between CDS and ADS result from mothers' affective emotions, to draw children's attention [13, 14].

Given that children's development of tone and tone sandhi rule has not been sufficiently explored, this study investigates the development of the disyllabic tone in Mandarin preschool children, including the distribution of production errors, the acoustic patterns, and the interaction between CPS and CDS. We predict that the error rate of the tone on the first syllable will be higher than that on the second due to the high variability of the first tone mentioned above [4]. Also, due to the changes in the CDS mentioned above [14], we predict that there will be a

synchronous entrainment effect [15] between CPS and CDS which is closely related to the development of children's speech.

2. MATERIALS AND METHODS

2.1. Participants and speech data

A total of 39 children and 11 mothers participated in this study. Mandarin is the language the mothers used exclusively with their children. The children were roughly divided into four age groups as shown in Table 1. None of the participants have a history of speech/hearing disorders.

Group	Age of the children	Children#	Moms#
G3	3 yrs (3;0-3;10, Mean: 3;4)	10 (2 males)	3
G4	4 yrs (4;0-4;11, Mean: 4;8)	11 (5males)	4
G5	5 yrs (5;0-5;10, Mean: 5;2)	12 (7 males)	2
G6	6 yrs (6;0-6;5, Mean: 6;3)	6 (3 males)	2

Table 1: Demographic information of participants.

The data were extracted from CASS_Child_Word [16], collected through a picture-naming task. If the child failed to produce the target word, the mother would say it and ask the child to imitate and repeat. Hence, some trials including both the child's and the mother's speech were selected as the recording samples in this study. There were a total of 1580 usable tokens, including 1360 produced by the children and 220 in CDS.

Two native speakers of Mandarin Chinese classified CPS as accurate, off-standard, and erroneous based on their auditory judgment. The recordings were annotated by a professional annotator in Praat. F0 was then extracted at 10 points of equal interval for each tone, then transformed into semitones with reference frequency of 54 Hz.

Ten acoustic parameters of each tone were calculated according to [17], namely Initial Pitch(pitch at time point 1), Mid Pitch (pitch at time point 5), Final Pitch (pitch at time point 10), Min Pitch 50% (minimum pitch in the second half of the tone), Max Pitch 50% (maximum pitch in the second half of the tone), Mean Pitch 50% (mean pitch in the second half of the tone), Pitch Range 50% (the absolute value of the difference between Max Pitch 50% and Min Pitch 50%), Pitch Range 100% (the absolute value of the difference between the maximum and minimum pitch in the whole tone contour), Slope 50% (directional excursion in the second half of the tone divided by the duration between Max Pitch 50% and Min Pitch 50%) and Tone Duration.

3. RESULTS

3.1. Accuracy distribution and confusion matrix in all age groups of children

Figure 1 shows the accuracy of the 16 disyllabic tone pairs produced by children aged 3;0-6;5. The overall accuracy is high, with an average of 95.36% (*Sd* = 4.99%) among the 16 tone combinations. Tone sandhi rules can cause confusion for children. Hence, the tone pairs containing T2 or T3 have lower accuracy, including T2T3 (92.31%), T4T2 (92.68%), T4T3 (93.10%), T2T2 (93.33%) (errors involving the first and the second syllables). It is also noteworthy that the accuracy of T4T4 is just 80%.

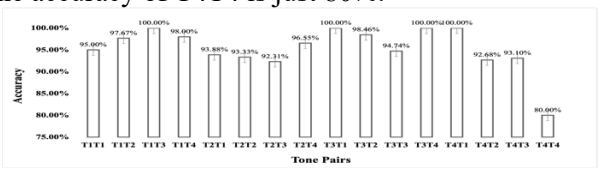


Figure 1: Accuracy of disyllabic tone pairs produced by children (regardless of age).

The confusion matrix of the 16 disyllabic tone pairs is presented in Table 2. The first tone tends to be mispronounced when followed by T1. Similarly, the tone pairs with T4 as the second tone tend to substitute whatever first tone with T1 or T3. Concerning the tone pairs with T2 or T3 as the second tone, the error patterns are more diverse. Among these tone pairs, the second tone is more frequently mispronounced. T2 tends to be replaced by T3 or T1 on the first syllable, and T3 by T1 or T4. On the second syllable, T2 tends to be replaced by T3. There is no case where both the first and the second tone are wrong at the same time.

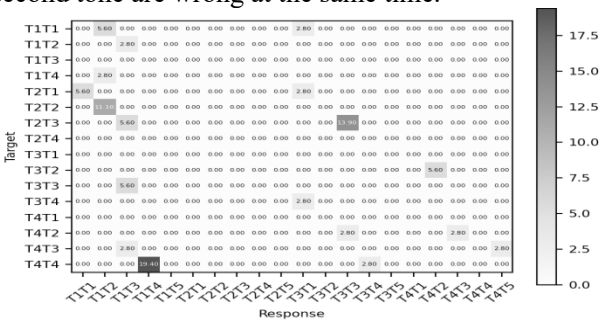


Table 2: The confusion matrix of disyllabic tones produced by children (regardless of age).

Off-standard production is defined as the case where the phonological category of tones is correct, but there are still problems with the phonetic realization. Specifically, for T1 (HH), the children's off-standard production shows a F0 contour similar to the rising or falling tone. For T2 (LH), T3 (LLH), the children's off-standard production is of the insufficient slope or curvature, showing a rather flat F0 contour. Similarly, for T4 (HL), the pitch slope is smaller with a lower start and insufficient falling.

3.2. Accuracy and error rate by age groups

Figure 2 shows the accuracy of disyllabic tones produced by children in each age group, in which S1 and S2 respectively stands for the first and second syllable. Statistical results indicate that the accuracy of CPS has significant positive correlations with age ($r1$=0.85, $p1$=0.001<0.05; $r2$=0.79, $p2$=0.001<0.05).

Figure 3 shows the error and off-standard rates of disyllabic tones produced in each age group. There is a clear downward tendency in the error rate of S1 from G3 to G5, and it tends to stabilize between G5 and G6. The off-standard rate of S1 declines from G3 to G6. The error rate of S2 is higher in G3, which indicates an anomaly in G4. However, in G5 and G6, the error rate stabilizes. The off-standard rate of S2 decreases from G3 to G6. Statistical results indicate that the error and off-standard rates of CPS have significant negative correlations with age (error rate: $r1$= -0.86, $p1$=0.047<0.05; $r2$=-0.99, $p2$=0.000<0.05; off-standard rate: $r1$=-0.65, $p1$=0.04<0.05; $r2$=-0.92, $p2$=0.000<0.05). Both the error and off-standard rates are higher for S1 than S2. Before the age of 4, the error rate of S1 was higher than S2, and the two rates of G4 were both less than 5%.

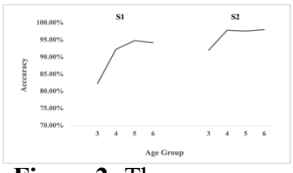


Figure 2: The accuracy of disyllabic tones in CPS.

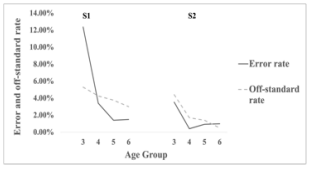


Figure 3: The error and off-standard rate of disyllabic tones in CPS.

Figure 4 shows the error and off-standard rates of CPS. In terms of S1, T2 has the highest error rate, followed by T4 and T3. As for S2, T3 is correct and those of the other tones are below 2% on average. The off-standard rate of T3 on S1 is the highest, followed by T2, and T4 and T1; and that of T3 on S2 is the highest, followed by T1, T4 and T2. Except T1, the error rate and off-standard rate of T2, T3, T4 on S2 are lower than on S1.

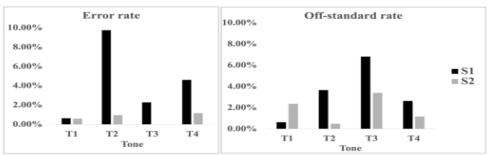


Figure 4: The error and off-standard rates of the tones on S1 and S2 in CPS.

3.3. Developmental patterns of the acoustic properties of disyllable tones in CPS and CDS

3.3.1 Pitch

This section analyses the patterns of disyllabic tones in children's correct production and in CDS. Figure 5 shows the mean pitch contours of the four tones on S1 and S2 in both CPS (in different age groups) and in CDS.

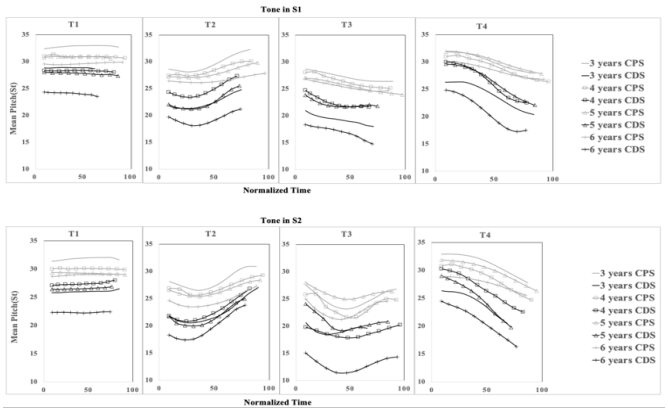


Figure 5: Pitch contours of disyllabic tones in CPS and CDS.

Comparing CPS and CDS, the pitch contours of T2 and T4 on the S1 position in CPS are flatter than those in CDS with lower absolute values of the rising and falling slope. The realization of the low target of T3 in CPS is poorer than that of CDS on both S1 and S2. Among different age groups in CPS, the overall pitch registers are downward with age. The contour of falling part of T3 and T4 on S1 grows with age, gradually approaching CDS's performance. As children age grows, S2 demonstrates three developments in tone production: the falling of T4, the forward movement of the pitch feature points of T2, and an increased curvature of contour tones. Finally, comparing the CDS for each age group of children, except for G3, the pitch of CDS for the other groups declines with age, following the same pattern as CPS.

Figure 6 depicts the developmental pattern of pitch range and pitch register in CPS and CDS. The mean pitch range of S2 is larger than that of S1 in both CPS and CDS, and that in CDS is larger than that in CPS for both syllables. Other than the CDS for 3-year-olds, the mean pitch register of S1 and S2 exhibits a decreasing trajectory in both CPS and CDS as the children grow older.

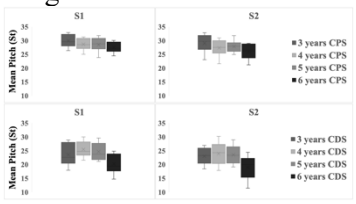


Figure 6: Pitch register and pitch range of S1 and S2 in CPS and CDS.

3.3.2 Duration

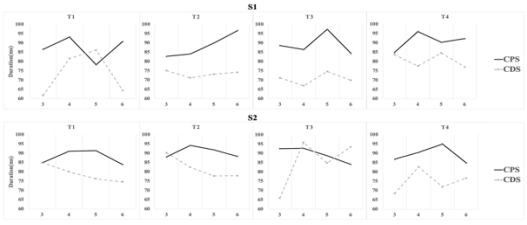


Figure 7: The mean duration of CPS and CDS.

Figure 7 shows the mean duration of disyllabic tones in CPS and CDS. The duration of each tone produced in CPS is longer than that in CDS. This phenomenon is more prominent in S1 than in S2.

3.4. Statistical analysis of the acoustic properties of disyllabic tones in CPS and CDS

For each of the acoustic parameters listed in Section 2.1, a multivariate analysis of variance (MANOVA) was performed using genre (CPS and CDS) as between-subjects factor, syllable context, and tone as within-subject factors and the acoustic parameter as the dependent variable.

As expected, significant main effects of tone were found in all comparisons (all $p<0.05$). There were significant main effects of genre in Pitch Range 50% (St) (CDS>CPS, $p<0.001$), Pitch Range 100% (St) (CDS>CPS, $p<0.001$), and Tone Duration (ms) (CPS>CDS, $p<0.001$), which means the pitch range of CDS is significantly wider than that of CPS, and the duration of CPS is significantly longer than that of CDS. Significant interactions between tone and genre were found for three acoustic parameters: the significance of CPS in Pitch Range 100% (St) (CPS: T4>T2=T3>T1, $p<0.05$; CDS: T4>T2>T3>T1, P<0.001), and Slope 50% (St/ms) (CPS: T2>T3>T1>T4, $p<0.05$; CDS: T2>T3>T4>T1, $p<0.001$) is lower than that of CDS.

4. DISCUSSION AND CONCLUSION

While children generally produce disyllabic tones well and their accuracy improves with age, the tone pairs containing T2 and T3 were found to be relatively difficult to acquire, and the T4T4 pairs had the lowest accuracy. The errors of replacing the T4 on the first syllable with T1 were mostly made by the lower age groups (3-4 years). These error patterns may be attributed to the proportion of tone combinations we selected from CASS_Child_Word. Furthermore, in the T4T4 tone pair, the physical limitations of children's pronunciation prevent F0 from dropping to an appropriate level. As a result, children may realize T4 at the position of S1 as T1, influenced by S2. The positive correlation between the accuracy and age indicates developmental improvements. The higher error and off-standard rates in S1 compared to S2 can be attributed to the "right-dominant-like" tone sandhi in Mandarin (tone sandhi of neutral tone word is a left-dominant-like one, which is not concerned in this paper), which echoes the results of previous studies on similar tone sandhi patterns [18].

The significance interaction between tone and genre shows that although children produced appropriate tone contours for all tones, children aged 3 to 6 do not produce adult-like disyllabic tones in terms of pitch range, slope, and length, consistent with previous study [9].

Comparing CPS and CDS, we found that their pitch registers share a same downward trend with children's age. This similarity in developmental patterns can be explained by the entrainment effect where the parent makes effort to communicate with children effectively and convey pragmatic information.

To conclude, this study investigates the production of disyllabic tones by Mandarin-speaking preschoolers and analyses the development of acoustic patterns in both CPS and CDS. We found that preschoolers have a phonological mastery of the tone sandhi rules. After the age of 4, the phonetic accuracy of children's disyllabic tone production has exceeded 95%, but more errors are made in Tone 2 and Tone 3 than in Tone 1. Since Mandarin has a kind of "right-dominant-like" sandhi pattern for normal stressed word, the first (left) syllable is more likely to be pronounced erroneously than the second (right). The error rate decreases dramatically with age. Moreover, we analyzed the phonetically off-standard pattern of CPS, which is characterized primarily with the "undershoot" of tonal goal with respect to the slope and curvature of pitch contours. T1 has a steeper slope, while T2, T3 and T4 have insufficient slope and curvature. Significant differences were also identified in the pitch register, pitch range, and duration of disyllabic tones across the four age groups of children and CDS. CPS and CDS follow similar developmental patterns with age, reflecting a synchronous entrainment. However, current evidence is insufficient to verify or contradict both the hypotheses. The results of this study therefore support our prediction that the error rate of the first syllable would be higher than that of the second, and that there is a synchronous entrainment between CPS and CDS, which is closely related to the developmental trajectory of child speech.

The present study would be helpful in the following aspects. Firstly, the confusion matrix in our study will be beneficial for the analysis of the speech of atypical children. Secondly, the understanding of the developmental pattern of the suprasegmental characteristics of Mandarin contributes to the studies and practices on Mandarin teaching to child learners with special needs and in ethic areas, as well as to teach Chinese as a second language for children.

For future work, we will verify the results with more age-grouped test data from the corpus. Tone pairs containing neutral tones would also be considered. We are conducting a perceptual experiment for Mandarin preschoolers, to explore the relationship between tone perception and production.

5. ACKNOWLEDGEMENTS

This work was supported by the Project of Cultural Experts and "Four Batches" of Talents directed by Aijun Li and CASS Innovation Project.

6. REFERENCES

[1] Chao, Y.-R. 2002. *Linguistic Essays*. The Commercial Press.

[2] Cao, J. 1992. On neutral-tone syllables in Mandarin Chinese. *Canadian Acoust.* 20, 49–50.

[3] Li, C., Thompson, S. 1977. The acquisition of tone in Mandarin-speaking children. *Journal of Child Language* 4, 185–199.

[4] Hua, Z., Dodd, B. 2000. The phonological acquisition of Putonghua (Modern Standard Chinese). *Journal of Child Language* 27, 3–42.

[5] Gao, J., Li, A., Zhang, Y. 2017. The production of tone in monosyllabic and disyllabic words by Mandarin-speaking children. *Studies on Speech Production–11th International Seminar*, 2017, Tianjin, China, October 16–19, 2017, Revised Selected Papers 89–98.

[6] Wong, P. 2012. Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *Journal of Phonetics* 40, 141–151.

[7] Wong, P., Schwartz, R. G., Jenkins, J. J. 2005. Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. *Journal of Speech, Language & Hearing Research* 48, 1065–1079.

[8] Wong, P. 2013. Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan. *The Journal of the Acoustical Society of America* 133, 434–443.

[9] Wong, P., Strange, W. 2017. Phonetic complexity affects children's Mandarin tone production accuracy in disyllabic words: A perceptual study. *PloS one.* 12.

[10] Xu Rattanasone, N., Tang, P., Yuen, I., Gao, L., Demuth, K. 2018. Five-Year-olds' acoustic realization of Mandarin tone sandhi and lexical tones in context are not yet fully adult-like. *Frontiers in Psychology* 9, 817.

[11] Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., Lacerda, F. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686.

[12] Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., Cristia, A. 2015. Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science* 26, 341–347.

[13] Tang, P., Xu R. N., Yuen, I., Demuth, K. 2017. Phonetic enhancement of Mandarin vowels and tones: Infant-directed speech and Lombard speech. *The Journal of the Acoustical Society of America.* 142. 493-503.

[14] WANG, L., Kalashnikova, M., Kager, R., LAI, R., WONG, P. 2021. Lexical and Prosodic Pitch Modifications in Cantonese Infant-directed Speech. *Journal of Child Language*. 48. 1-27.

[15] Edlund, J., Heldner, M., Hirschberg, J. 2009. Pause and gap length in face-to-face interaction. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* 2779-2782.

[16] Gao, J., Li, A., Xiong, Z., Shen, J., Pan, Y. Normative database of word production of Putonghua-speaking children — Beijing articulahon norms project: CASS_CHILD_WORD. *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Gurgaon, India, 2013, pp. 1-6.

[17] Wong, P. 2018. Mothers do not enhance tonal contrasts in child-directed speech: Perceptual and acoustic evidence from child-directed Mandarin lexical tones. *The Journal of the Acoustical Society of America* 143, 3169.

[18] Li, X., Mok, P. P. K. 2020. The acquisition of tone sandhi of the Xiamen dialect. *Speech Prosody*, 479–483.

[19] Liu, Y., Li, A., Dang, J., Zhou, D. 2021. Semantic and acoustic-prosodic entrainment of dialogues in service scenarios. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 71–74.

[20] Pickering, M. J., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 169–190.

AN ACOUSTIC STUDY OF TONE SANDHI IN LISHUI WU

Wupeng Lan[1], Xiaocong Chen[1], Caicai Zhang[1]

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China
21046493g@connect.polyu.hk, xiaocong.chen@polyu.edu.hk, caicai.zhang@polyu.edu.hk

ABSTRACT

This study investigated the tone sandhi in the Lishui Wu dialect, an understudied variety of Southern Wu dialects in Zhejiang Province, China. Previous studies showed that the tone sandhi in the Lishui dialect is mostly right-dominant, but there seem to be no uniform patterns across items and speakers. We recruited 4 elderly and 4 young native speakers of Lishui Wu and conducted acoustic analyses of the f0 of the first syllable in both disyllabic real and pseudo words. Overall, our results did not reveal consistent sandhi patterns for both real and pseudo words, but the first syllable tended to be mostly produced as falling tones (with different slopes), with the pitch onset height conditioned by the yin and yang tones. Further analyses also found that young speakers showed a larger similarity between the sandhi and citation tones than elderly speakers, indicating some generational changes in Lishui tone sandhi.

Keywords: Tone sandhi, Lishui Wu, age, lexicality

1. INTRODUCTION

It has been shown that Wu Chinese exhibits complex tone sandhi patterns, but previous acoustic studies mostly focused on the major varieties of Wu dialects such as Shanghai [1][2] and Wuxi dialect [3]. In this study, we investigated the tone sandhi pattern of an understudied variety of Southern Wu, the Lishui (丽水) Wu dialect. The Lishui dialect belongs to the Chuqu (处衢) subgroup of Wu dialects and is spoken mostly in the Liandu district, the main urban area of Lishui city in Zhejiang province, China [4].

The f0 patterns of the citation tones in Lishui Wu are displayed in Fig 1. Due to the historical evolution, the T4 (yangshang) category has merged with other tonal categories, with those derived from the voiced obstruent initials in Middle Chinese (T4a) merging with T2 (yangping), and those derived from the voiced sonorant initials (T4b) merging with T3 (yinshang) [5]. Thus, there are only seven distinctive citation tones [5]-[9], which consist of six smooth tones (T1: yinping 34; T2: yangping 32; T3: yinshang 43; T5: yinqu 51/41; T6: yangqu 341), as well as two checked tones (T7: yinru 45; T8: yangru 23).

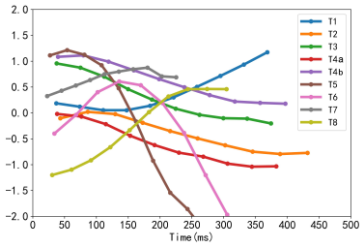


Figure 1. Pitch contour of citation tones in Lishui Wu (y-axis stands for normalized z scores of log-transformed f0)

Previous studies showed that the disyllabic tone sandhi pattern in Lishui dialect is mostly right-dominant, whereby the final syllable in a disyllabic word preserves its citation tone in most cases, while the first syllable tended to change to another tone [5]-[9]. However, previous impressionist phonetic transcriptions and acoustic studies did not consistently reveal well-formed patterns for the sandhi tones of the first syllable [5]-[9], though it was shown that most of the sandhi tones were level tones or falling tones [5][8]. Moreover, as shown in [8], the tone sandhi patterns seemed to vary among three speakers of different ages, indicating that the tone sandhi may undergo some changes across generations. It remains unclear whether old and young speakers show any difference in their tone sandhi patterns. Besides, it also remains a myth whether native speakers have any implicit knowledge of tone sandhi, as no consistent tone sandhi patterns were revealed in previous studies, and less is known about whether tone sandhi in the Lishui dialect can be productive in novel words.

Thus, this study aims to investigate how age and lexicality influence the tone sandhi patterns of the first syllable in disyllabic words of the Lishui dialect. We focused on addressing two questions: (a) What are the tone sandhi patterns of the first syllable in disyllabic real words and pseudowords in the Lishui dialect? (b) Are there any differences in the tone sandhi patterns between young and elderly speakers?

2. METHODS

2.1. Subjects

We recruited 8 native speakers of the Lishui dialect, including four elderly speakers (2 males and 2 females, aged between 72 and 78) and four young

speakers (2 males and 2 females, aged between 23 and 35), all of whom were born and raised in the main urban area of Lishui and had not lived outside Lishui for a long time before 18. They could speak the Lishui dialect fluently and used it frequently in their daily life. All the participants' parents were also born and raised in the main urban area of Lishui and spoke the Lishui dialect. They reported no language or hearing deficits and could read Chinese characters using the Lishui dialect.

2.2. Materials and procedure

Given the practical consideration for the experimental duration, we only examined the tone sandhi patterns of the smooth tones as the initial syllables. A real-word and a pseudo-word list were created. The real words were selected from common disyllabic words in the Lishui dialect, and the pseudowords were constructed by combining two existing morphemes to form nonsense items albeit with some syntactic structures (see below). All the real words and pseudowords were from 48 tonal combinations (6 tonal categories for the first syllable including T1, T2, T3, T5, T6 and the merged T4; 8 tonal categories for the second syllable including T1, T2, T3, T5, T6, T7, T8 and the merged T4), with four items for each of the tonal combinations in both real words and pseudowords respectively. For the four items of each tonal combination for both real and pseudowords, two had the Modifier+Noun (MN) structure and the other two had the Verb+Object (VO) structure (also see [10]). There were 384 items (48×4×2) in total. Four additional Lishui native speakers confirmed the selected real word words as frequent words and the constructed pseudowords as non-existing items. A single-character list was also created for the measurement of citation tones.

The recording took place in a quiet room. The stimuli were presented one by one by DMDX on a laptop computer screen and participants' pronunciations were recorded by a condenser microphone SAMSON C03U sampled at 22050 Hz with 16-bit quantification. Participants were asked to pronounce the single-character list first, and then completed the real-word and the pseudoword list, with the order of the two lists counterbalanced across participants. Before completing the pseudoword list, the experimenter also told participants to pronounce two characters together as if they formed a real disyllabic word and provided four additional pseudowords for practice. The stimuli within each list were presented randomly. Participants were asked to pronounce each item twice. The experimenter listened to each participant's pronunciation and make sure that they did not make a mistake.

2.3. Data preprocessing and coding

All stimulus recordings were imported into Praat for manual annotation and segmentation. 10 equidistant f0 points in the annotated rime portion were automatically extracted using the STRAIGHT algorithm in VoiceSauce [11]. We removed those tokens with inaccurate f0 tracking due to creaky voicing and obtained 2536 tokens for real words and 2528 tokens for pseudowords. We computed the log-z (LZ) values by log-transforming f0 values and performing the z-score normalization to remove any speaker/gender differences [12]. We also averaged the LZ values of the tokens for each item for each participant and transcribed the tones on a 5-level scale. We did not analyse the items when the T4 was the first syllable due to its merging with other tonal categories. We thus only focused on the sandhi tonal realizations for T1, T2, T3, T5 and T6 as the first syllable in the analysis.

3. RESULTS

3.1. Tone sandhi patterns in real and pseudo words

In general, we found that for each tonal combination, the sandhi tones of the first syllables from the same tonal category were realized as more than one type of pitch contour in most cases, and the sandhi tone realizations also varied from speaker to speaker. Thus, no consistent patterns were clearly detected. However, we found that, the pitch onset height of the sandhi tones was mostly conditioned by the yin and yang tones, a phenomenon similarly observed in other Wu dialects [5]. Specifically, the sandhi tones of the yin tones (T1, T3, T5) started at a higher f0 onset (above LZ score 0) whereas those of the yang tones (T2, T6) started at a lower f0 onset (below LZ score 0) (see Fig 2).

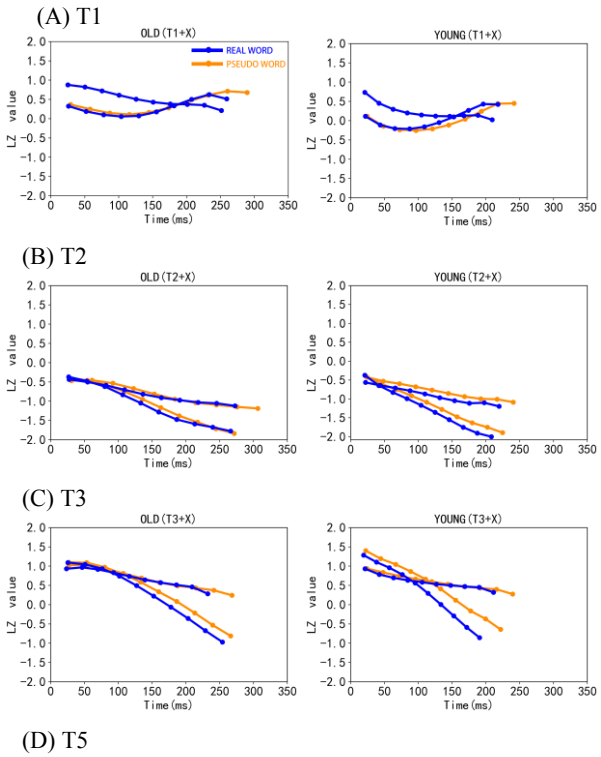
Given the various realizations of the sandhi tones, we further categorized the sandhi tones into the following types: Falling, Level, Rising, and Convex tones, based on the pitch contour of the sandhi tones.[1] Moreover, for falling tone, yin tones (T1, T3 and T5) mainly produced to a high falling tone, with its f0 onset over LZ score 0, whereas yang tones (T2, T6) mainly produced to a low falling tone, with its f0 onset below LZ score 0. Additionally, different slopes of falling tones were also detected. We subdivided the falling tones into four subcategories: high-onset steeper-slope falling tone (FSP+), high-onset shallower-slope falling tone (FSL+), low-onset steeper-slope falling tone (FSP-), and low-onset shallower-slope falling tone (FSL-). We then calculated the proportion of each tonal type for each tonal category and displayed those dominant types of pitch contour of the sandhi tones from each tonal

category in both real and pseudo words for the elderly and young age groups separately in Fig 2. In general, we found that the sandhi tones of the same tonal category exhibit different realizations in both real and pseudo words. Moreover, we also found that the realizations of the sandhi tones were not clearly conditioned by the following tone.

As for T1, in real words, all participants tended to produce it as a high-onset shallower-slope falling tone /43/ (old:29% vs. young:43%), or a rising tone /34/ (old:45% vs. young:41%) that was very similar to the citation tone of T1 /34/. In contrast, for pseudowords, all participants tended to produce it as the rising tone /34/ (old:71% vs. young:83%), indicating that they may choose to preserve the citation tone of T1 in novel word contexts.

For T2, in real words, all participants tended to produce it as low-onset falling tones with a shallower slope /32/ (old:31% vs. young:37%) or a steeper slope /31/ (old:25% vs. young:29%). The same pattern was also observed in pseudowords (/32/: 34% for old people and 39% for young people; /31/: 20% for old people and 23% for young people). Note that these two falling tones are also similar to the citation tone of T2 /32/.

For T3, in real words, all participants tended to produce it as high-onset falling tones with a shallower slope (/43/ (old:50%; young:51%) or a steeper-slope falling /42/ (old:48%; young:30%). A similar pattern was also detected in pseudowords (/43/: 48% for old people and 59% for young people; /42/: 50% for old people and 29% for young people). Again, the sandhi tones are very similar to the citation tone of T3 /43/.



(D) T5

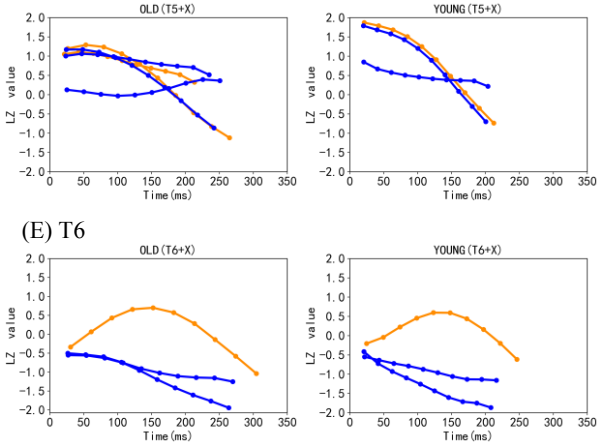


Figure 2. The pitch contour of the dominant types of sandhi tones from T1, T2, T3, T5 and T6 as the first syllable in real and pseudo words for the elderly and young age groups (—real word, —pseudoword).

For T5, in real words, old speakers tended to produce it as a high-onset steeper-slope falling tone /42/ (43%), a high-onset shallower-slope falling tone /43/ (28%) or a rising tone /34/ (20%). In contrast, young speakers tended to produce it as a high-onset steeper-slope falling tone /52/ (44%) or a high-onset shallower-slope falling tone /43/ (24%). In pseudowords, old speakers tended to produce it as a high-onset steeper-slope falling tone /42/ (71%) and a high-onset shallower-slope falling tone (43) (25%) while young speakers tended to dominantly produce it as a high-onset steeper-slope falling tone /52/ (73%), which seemed to more resemble the citation tone of T5 /51/.

For T6, in real words, all participants tended to produce it as low-onset falling tones with a shallower slope /32/ (old: 27% vs. young: 31%) or a steeper slope /31/ (old:21% vs. young:24%). In contrast, for pseudowords, all participants tended to produce it as the convex tone /342/ (old:44% vs. young:42%), which was very similar to the citation tone of T6 /341/. This seemed to indicate that they tended to preserve the citation tone of T6 contexts.

In sum, although we did not detect consistent patterns for each tonal category, we found that the dominant types of sandhi tones were mostly falling tones (with different slopes), and the pitch onset height was conditioned by the yin and yang tones. Moreover, it seemed that the tone sandhi patterns of old and young speakers mostly converged except for T5. In addition, it seemed that people tended to preserve the citation tones for some tonal categories in pseudowords.

3.2 Analysis of the similarity between citation tones and sandhi tones

As shown in the previous section, it seemed that some of the dominant types of the sandhi tones were similar

to the citation tones of the same tonal category, indicating that participants may choose to preserve the citation tone in some cases. Thus, to quantify the degree of tone sandhi, we followed [13] and measured the pitch similarity between the sandhi tones and citation tones by computing the Euclidean distance between the 10 f0 points of the derived tones from each token of the initial syllable and the 10 averaged f0 points of the citation tone from the same corresponding tonal category for each participant respectively. The Euclidean distance was calculated by the following equation:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_{10}-y_{10})^2}$$

To further investigate how age and lexicality influenced the degree of tone sandhi, we conducted the linear mixed-effect model analysis on the Euclidean distance using JASP [14], with age and lexicality included as the fixed-effect predictors, and the structure as the control covariate into the model. Due to the model convergence problem, we only included the by-participant and by-item intercept. The model results were summarized in Table 1.

Table 1. Fixed effect estimates for Linear Mixed Models results of Euclidean distance.

Effect	df	F	p
Structure	1, 320.50	0.207	.650
Age	1, 6.00	5.970	.050
Lexicality	1, 320.5	0.109	.741
Structure*Age	1, 4736.40	0.063	.801
Age * Lexicality	1, 4736.42	11.885	< .001
Structure * Lexicality	1, 320.50	<.001	.996
Structure*Age*Lexicality	1, 4736.42	0.066	.797

As shown in Table 1, we found that there was a significant main effect of age and a significant interaction effect between age and lexicality on the Euclidean distance. Further analyses revealed that the mean Euclidean distance was smaller for young speakers than old speakers in both real words (β=88.793, *SE*=34.09, z=2.605, p=.028) and pseudo words (β=77.586, *SE*=34.09, z=2.276, p=.046). This showed that the sandhi and citation tones were more similar for young than old speakers, indicating that young speakers tended to preserve the citation tones. Moreover, there was also a significant difference in the mean Euclidean distance between the real and pseudo words among old speakers (β=-7.588, *SE*=2.506, z=-3.028, p=.010), as old speakers showed more similarity and tended to preserve the citation tones more in pseudowords than real words. In contrast, no lexicality effect was found among young speakers (β=3.618, *SE*=2.501, z=1.447, p=.148), possibly because young speakers tended to preserve the citation tones in both real and pseudowords.

4. DISCUSSION AND CONCLUSION

In this study, we found that in Lishui disyllabic words, the dominant sandhi tone types of first syllables were mostly falling tones (with different slopes). This is slightly different from previous results [5]-[9], which found that the sandhi tones were predominantly level tones and falling tones. This may be due to the slight difference between the perceptual and subjective judgment for the classification of level tone. Some of the tones classified as shallower-slope falling tones in this study may be perceptually classified as level tones in previous studies. Moreover, we also found that the pitch onset height of the sandhi tones was conditioned by the yin and yang tones, consistent with similar previous observations in Wu dialects [5]. However, this experiment revealed that the tone sandhi patterns of the Lishui disyllabic words were extremely complex and did not exhibit clear patterns for each tonal combination (also see similar observations in other Wu dialects [5]). The reasons for this perplexing complexity are still unknown, which need further examination.

Our further analyses revealed that young speakers tended to show a larger similarity between the sandhi and citation tones in real words. This seems to indicate that the sandhi tones in the Lishui dialect have a trend of preserving the original citation tones from the corresponding tonal category instead of changing to another different tone in younger generations. Moreover, for pseudowords, old speakers also showed a larger distance between the sandhi and citation tones than young speakers, which indicated that old speakers showed a higher degree of tone sandhi and could apply some knowledge of tone sandhi even in pseudowords. However, we also found that old speakers had a greater degree of tone sandhi in real words than pseudowords. That is to say, the realizations of tone sandhi in real words did not fully apply to pseudowords. This seemed to indicate that the sandhi tones were probably stored with the specific lexical items in native speakers' mental lexicon [15]. On the whole, the degree of tone sandhi in old speakers is greater than young speakers. This shows that Lishui dialect has a trend of becoming less tone sandhi as time changes. However, note that our study did not manipulate the semantic transparency [3], which could further affect the tone sandhi productivity. Thus, further research is required.

Moreover, the Euclidean distance used in this study only took the mean LZ value into consideration and may not fully reflect the similarity in terms of the pitch contour. Other similarity measures taking the tonal shape into account need to be explored in future studies.

5. ACKOWLEDGEMENTS

We thank Wanxiang Subdistrict Offices, Liandu District, Lishui City. We thank Tu Huimin, Chong Shufen, Zhao Lijun and Tong Chang, who gave us selfless help in the pre-participant recruitment and the recording place. We thank the eight participants for taking the time to participate in our experiment. Additionally, we would like to thank Gao Ziqi, for his help in writing the code.

6. REFERENCES

[1] Zhu, X. 2006. *A Grammar of Shanghai Wu*. Lincom Europa.
[2] Zhang, J., Meng, Y. 2016. Structure-dependent tone sandhi in real and nonce disyllables in Shanghai Wu. *Journal of Phonetics*, *54*, 169-201.
[3] Yan, H. 2017. *The nature of variation in tone sandhi patterns of Shanghai and Wuxi Wu*. Springer.
[4] Xiong, Z.H., Zhang, Z.X., Huang, X., Dao, B., & Zou, J.Y. 2013. *Language Atlas of China*. The Commercial Press.
[5] You, R.J., & Yang, J.Q. 2001. *Wuyu shengdiao de shiyan yanjiu*[An Experimental Study of Wu tone sandhi]. Fudan University Press.
[6] Ji, X.L. 1988. *Zhongguo Dabaike quanshu:Yuyan Wenzi*[Encyclopedia of China: language and character]. Encyclopedia of China Publishing House.
[7] Wang, W.S. 2002. Zhejiang Lishui fangyan de yuyin tedian[Phonetic features of Lishui dialect of Zhejiang]. *Lishui shifan zhuanke xuexiao xuebao*, *24*(1), 46-49.
[8] Steed, W.J. 2010.*Lishui Wu tone and tone sandhi: an acoustically-based description*. Dissertation, ANU.
[9] Lei, Y.P. 2019. *Zhejiang fangyan ziyuan diancang: Lishui*[Zhejiang dialect resource collection: Lishui]. Zhejiang University Press.
[10] Yan, H. 2013. *The productivity of tone sandhi patterns in Wuxi Chinese.* Master thesis, University of Kansas
[11] Shue, Y.-L., P. Keating , C. Vicenik, K. Yu. 2011. VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.
[12] Zhu, X. 2010. *Yuyin xue*[Phonetics]. The Commercial Press.
[13] Rhee, Nari., Kuang, J.J., & Chen, Aoju. 2022. The effect of musicality on the development of Mandarin prosody. *Speech Prosody*, 704-707.
[14] JASP Team. 2022. JASP (Version 0.16.3)[Computer software].
[15] Hsieh, H. I. 1976. On the unreality of some phonological rules. *Lingua*, *38*(1), 1-19.

[1] Concave tones and rising tones were classified into the same category as the number of concave tones was small and concave tones did not form phonological contrast with rising tones in the Lishui dialect.

SYNTACTIC AND PROSODIC EFFECTS ON RECONSTRUCTION OF TONE SANDHI DOMAINS: EVIDENCE FROM XIAMEN DIALECT

Yiying Hu[1], Hui Feng[1]*, Aijun Li[2]

[1]School of Foreign Languages, Tianjin University, [2]Institute of linguistics, Chinese Academy of Social Sciences
huyiying@tju.edu.cn, fenghui@tju.edu.cn, liaj@cass.org.cn

ABSTRACT

Recent studies have addressed the syntactic and prosodic conditions on tone group reconstruction, indicating the syntax-prosody competition, but previous results show less consensus. This study investigates syntactic and prosodic effects on reconstruction of the tone sandhi domain in Xiamen Dialect. Findings are as follows: (1) The tone sandhi domain in Xiamen Dialect is reconstructed into a smaller one by producing the break of phonological phrases after the head verb and the adjunct XP. (2) The syntactic adjuncthood affects positions of the break of phonological phrases. Such break occurs more frequently after AdvPs than after head verbs and CLPs. (3) Prosody exerts restrictions on the number and positions of the break of phonological phrases, which limits syntactic effects on reconstruction of the tone sandhi domain.

Keywords: tone sandhi domain, prosody, syntax, Xiamen Dialect

1. INTRODUCTION

1.1. Effects of syntax and prosody on tone sandhi domain reconstruction

Previous studies addressed syntactic and prosodic conditions on tone group reconstruction, indicating an interface between syntax and prosody [1].

Two factors affecting reconstruction of the tone sandhi domain in Taiwanese have been discovered: syntactically the adjuncthood and prosodically the number of syllables in the domain. In terms of syntactic effects, the break of phonological phrases produced by young speakers may occur after a head verb and an adjunct XP, which appears inconsistent with prior research on general Taiwanese that breaks of phonological phrases only occur at the right edge of non-adjunct XPs [2, 3, 4, 5]. When it comes to prosodic effects, trisyllabic and tetrasyllabic fragments are preferred lengths for tone groups produced by young speakers [4, 6]. Yet, prior work shows less consensus as some studies on Taiwanese produced by the young generation found that a phonological phrase is restricted within three syllables [2, 3].

1.2. Tone sandhi conditioning in Xiamen Dialect

The tone inventory of Xiamen Dialect comprises seven tones, including five smooth tones and two checked tones [7]. All these seven tones have tonal alternations between base tones and sandhi tones. Table 1 lists the tone inventory of Xiamen Dialect.

	T1	T2	T3	T4	T5	T6	T7
Base tones	H	LM	HL	L	M	M	H
Sandhi tones	M	M	H	HL	L	H/HL	L/L

Table 1: Tone Inventory [8, 9].

Base tones are regarded as underlying forms in the derivational process of tonal alternations while sandhi tones are regarded as surface forms deriving from corresponding base tones [10, 11, 12]. The tone sandhi rule is shown in (1).

(1) Tone Sandhi Rule [13]
base tone → sandhi tone / __ X #

The tone sandhi domain conditioning tone sandhi in Xiamen Dialect is related to syntactic structures [4, 13, 14, 15, 16]. Within a tone sandhi domain, base tones are only produced at right edges of sandhi domains while sandhi tones are produced in other positions. Chen [13] suggested that the tone group in Xiamen Dialect is marked at the right edge of a non-adjunct XP but not of an adjunct XP. Chen [5] defined that XP is an adjunct of Y if XP appears in $[...XP...]_{YP}$ and is not a strictly subcategorized argument of Y [17]. Examples are given in (2) and (3), with edges of tone groups marked with "#", base tones marked as B and sandhi tones marked as S. The NP in (2) is a non-adjunct XP, which is the internal argument of the verb 'take'. The base tone is retained at the right edge of this NP. By contrast, the AdvP in (3) is an adjunct XP modifying the verb 'rise-up' and carries the sandhi tone.

(2) XP as a non-adjunct [13]
t'eq tsih$]_{NP}$ # *k'i puaq-kiao*
S B S S B
take money to gamble
'take the money to gamble'

(3) XP as an adjunct [13]
tsa$]_{AdvP}$ *k'i-lai*
S S B
early rise-up
'get up early'

The tone group in Xiamen Dialect is suggested as the tone sandhi domain, also referred to as the phonological phrase (φ) [4, 18]. The break of the phonological phrase (φ-break) is marked at the right or left edge of a syntactic XP and X^{head} [19]. So the φ-break coincides with the right edge of the non-adjunct XP [6, 18, 20, 21]. And the phonological phrase is considered as the tone sandhi domain in previous studies [6, 13, 18, 20, 21].

So far, tone group reconstruction in Xiamen Dialect has been rarely studied. In addition, few attempts have been made to explore the discrepancies between effects of different adjunct XPs on tone group reconstruction. And prosodic restrictions on tone group reconstruction have not been convincingly presented as prior studies showed less consensus. Given great phonological similarities between Taiwanese and Xiamen Dialect but significant discrepancies in their contact with Mandarin Chinese, this study addresses the following research question. That is, whether syntax and prosody affect reconstruction of the tone sandhi domain in Xiamen Dialect in the same way as they do in Taiwanese.

2. METHOD

2.1. Participants

Nineteen participants were invited who were born and brought up in Xiamen, and they speak both Xiamen Dialect in their daily life and Mandarin Chinese when at work. They were divided into three groups, shown in Table 2.

	Participants	Mean Age
Group A	3 female, 4 male	20.6 (SD=0.5)
Group B	2 female, 4 male	48.0 (SD=2.1)
Group C	4 female, 2 male	75.2 (SD=2.5)

Table 2: Information of participants.

2.2. Materials

The materials include 21 carefully designed test sentences, with various combinations of the number of syllables and syntactic structures, constructing adjunct XPs in different lengths. For the adjunct XP, the study focuses on the adverbial phrase (AdvP) and the classifier phrase (CLP) as adjuncts. For the X^{head}, this study focuses on the head verb. The example of the AdvP as an adjunct, the CLP as an adjunct and the head verb are respectively provided in (6), (7) and (8) in the discussion part. All sentences were read three times by all participants, producing 1,197 tokens (21 sentences× 3 repetition× 19 participants).

2.3. Procedure

The recording was carried out in a quiet room. Participants were instructed to read materials presented on an iPad screen at normal speed as if talking to a close friend. The recording equipment was a Lenovo YOGA C740 computer and a Plantronics Blackwire 3220 Series headset microphone. All sounds were recorded with PRAAT [22] at 44.1 kHz 16 bits and a mono soundtrack.

2.4. Data analysis

Tone bearing units in test sentences were manually annotated based on the spectrogram. Voice Sauce was used to extract F0 values [23]. Two well-trained native speakers of Xiamen Dialect worked as judges to decide whether or not tone sandhi occurred. Listening to all recorded sentences, judges annotated base tones and sandhi tones produced for each syllable in sentences, with the help of F0 for each recording. All their annotations reached a consensus. For each test sentence produced by each age group, patterns of the φ-break were calculated and presented in the percentage.

3. RESULTS AND DISCUSSION

3.1. The discovery of an intermediate φ-break

The study discovers an intermediate state of the φ-break (noted as φ'-break), which is only found after head verbs produced by GA, shown in (4b) and (5b). The tone bearing this φ'-break is a high level tone H, and such a tone value is included in neither the base tone L nor the sandhi tone HL of the word *sau*.

(4)		[[*pʰaʔ* *sau*]$_V$		[[*lak* *e*]$_{CLP}$		[*paŋ* *kūī*]$_N$]$_{NP}$]$_{VP}$		Frequency of φ-break patterns (%)		
		clean		six	CL	room				
		'Clean six rooms.'								
Base tones		M	L	H	LM	LM	H	GA	GB	GC
φ-break Patterns	a.	(HL	HL	L	M	M	H)φ	28.57	100	100
	b.	(HL	H)φ'	(L	M	M	H)φ	57.14	0	0
	c.	(HL	L)φ	(L	M	M	H)φ	14.29	0	0

(5)		[[*kūā* *kɪŋ*]$_{AdvP}$		[[*pʰaʔ* *sau*]$_V$		[[*lak* *e*]$_{CLP}$		[*paŋ* *kūī*]$_N$]$_{NP}$]$_{V'}$]$_{VP}$		Frequency of φ-break patterns (%)		
		quickly		clean		six	CL	room				
		'Clean six rooms quickly.'										
Base tones		HL	HL	M	L	H	LM	LM	H	GA	GB	GC
φ-break Patterns	a.	(H	H	HL	HL	L	M	M	H)φ	14.29	100	100
	b.	(H	H	HL	H)φ'	(L	M	M	H)φ	28.57	0	0
	c.	(H	H	HL	L)φ	(L	M	M	H)φ	28.57	0	0
	d.	(H	HL)φ	(HL	HL	L	M	M	H)φ	28.57	0	0

In (4) and (5), either φ-break or φ'-break occurs after the head verb, which is the way for a high percentage of GA to reconstruct shorter sandhi domains. We suppose that φ'-break is an intermediate state of phonological phrasing break, as it provides a substitute for the base tone to distinguish the tone bearing such a break from the sandhi tone.

3.2. Syntactic effects on reconstruction of sandhi domains

3.2.1. Patterns of sandhi domain reconstruction

Results illustrate that the tone sandhi domain in Xiamen Dialect is reconstructed into a shorter domain. As presented in (6), (7) and (8) where VP and NP are non-adjunct XPs, Group C, whose members are older than 70, only preserve the base tone at the rightmost edge of the non-adjunct XP, while the younger group GA and a part of middle-aged group GB reconstruct shorter tone sandhi domains by producing base tones at the right edges of the adjunct XP and the head verb. Present results show at least three patterns for young Xiamen Dialect speakers to reconstruct shorter tone sandhi domains.

(6)		[[*kũã* *kin*]$_{AdvP}$ [*pʰaʔ* *sau*]$_{V}$]$_{VP}$				
		quickly		clean		Frequency of
		'Clean quickly .'				φ-break patterns (%)
Base tones		HL	HL	M	L	GA GB GC
φ-break	a.	(H	H	HL	L)φ	28.57 66.67 100
Patterns	b.	(H	HL)φ	(HL	L)φ	71.43 33.33 0

(7)		[[*tsap* *li* *e*]$_{CLP}$ [*paŋ* *kãĩ*]$_{N}$]$_{NP}$					
		twelve		CL	room		Frequency of
		'Twelve rooms.'					φ-break patterns (%)
Base tones		H	M	LM	LM	H	GA GB GC
φ-break	a.	(L	L	M	M	H)φ	85.71 100 100
Patterns	b.	(L	L	LM)φ	(M	H)φ	14.29 0 0

(8)		[[*pʰaʔ* *sau*]$_{V}$		[[*lak* *e*]$_{CLP}$	[*paŋ* *kãĩ*]$_{N}$]$_{NP}$]$_{VP}$			
		clean		six CL		room		Frequency of
		'Clean six rooms.'						φ-break patterns (%)
Base tones		M	L	H	LM	LM	H	GA GB GC
φ-break	a.	(HL	HL	L	M	M	H)φ	28.57 100 100
Patterns	b.	(HL	H)φ'	(L	M	M	H)φ	57.14 0 0
	c.	(HL	L)φ	(L	M	M	H)φ	14.29 0 0

First, the tone sandhi domain produced by the younger generation can be blocked by the adverbial phrase (AdvP) which is an adjunct, presented in (6b). The AdvP in (6) is an adjunct modifying the verb and *kin* is at the right edge of this adjunct XP. All GC produce sandhi tones for *kin*, while 33.33% of GB and 71.43% of GA retain base tones.

Second, the tone sandhi domain produced by 14.29% of GA is blocked by the classifier phrase (CLP) which is an adjunct, shown in (7b). The CLP in (7) is an adjunct modifying the noun 'room' and the syllable *e* is at the right edge of this adjunct XP.

Third, the φ-break also occurs after the head verb produced by GA, shown in (8c). *Sau* is at the right edge of the head verb 'clean'. All GB and GC produce sandhi tones for *sau*, while a high percentage of GA produce base tones for it..

Generally, young speakers of Xiamen Dialect reconstruct shorter tone sandhi domains by producing φ-breaks after head verbs and adjunct XPs, including AdvPs and CLPs.

3.2.2. Frequency of the three patterns

Between three tone-group reconstruction patterns in Xiamen Dialect, discrepancies are found in the use frequency in tone group reconstruction and in the order of appearance across the three age groups.

First, φ-breaks occur more frequently after the AdvP than the CLP when both of them are adjunct XPs, shown in (9) and (10). Results show 71.43% of GA and 33.33% of GB retain base tones at the right edge of AdvP in (9), while only 14.29% of GA preserve base tones at the right edge of CLP in (10). Such a trend is inconsistent with previous research in [4], which shows that φ-breaks after both AdvPs and CLPs are used by 99.6% of young Taiwanese speakers when reconstructing tone groups.

From the diachronic perspective, results suppose that the φ-break after the CLP appears later than that after the AdvP in the ongoing change. When the AdvP and CLP are both adjunct XPs, 33.33% of GB produce the φ-break after the AdvP and none of GB produces that after the CLP.

(9)		[[*kũã* *kin*]$_{AdvP}$ [*pʰaʔ* *sau*]$_{V}$]$_{VP}$				
		quickly		clean		Frequency of
		'Clean quickly .'				φ-break patterns (%)
Base tones		HL	HL	M	L	GA GB GC
φ-break	a.	(H	H	HL	L)φ	28.57 66.67 100
Patterns	b.	(H	HL)φ	(HL	L)φ	71.43 33.33 0

(10)		[[*tsap* *li* *e*]$_{CLP}$ [*paŋ* *kãĩ*]$_{N}$]$_{NP}$					
		twelve		CL	room		Frequency of
		'Twelve rooms.'					φ-break patterns (%)
Base tones		H	M	LM	LM	H	GA GB GC
φ-break	a.	(L	L	M	M	H)φ	85.71 100 100
Patterns	b.	(L	L	LM)φ	(M	H)φ	14.29 0 0

Second, the φ-break occurs more frequently after the AdvP than after the head verb. As (11) shows, 85.71% of GA and 33.33% of GB produce φ-breaks after AdvPs, while only 28.57% of GA produce φ-breaks after head verbs.

From the diachronic perspective, results imply that the φ-break after the head verb appears later than that after the AdvP in the ongoing change, as 33.33% of GB produce the φ-break after AdvPs but none of GB produces that after the head verb.

(11)		[[*ko*	*kʰa*	*kũã*	*kin*]$_{AdvP}$	[[*pʰaʔ*	*sau*]$_{V}$	[*paŋ*	*kãĩ*]$_{N}$]$_{V}$]$_{VP}$	
		still	more	quickly		clean		room		Frequency of
		'Clean the room much more quickly.'								φ-break patterns (%)
Base tones		HL	M	HL	HL	M	L	LM	H	GA GB GC
φ-break	a.	(H	H	H	H	HL	HL	M	H)φ	14.29 66.67 100
Patterns	b.	(H	H	H	HL)φ	(HL	HL	M	H)φ	57.14 33.33 0
	c.	(H	H	H	HL)φ	(HL	H)φ'	(M	H)φ	28.57 0 0

Third, the φ-break (also the φ'-break) occurs after the head verb but not after the CLP when the head verb and the CLP appear in the same sentence, presented in (12).

(12)		[[pʰaʔ sau]V clean	[[lak e]CLP six CL	[paŋ kuĩ]N]NP]VP room	Frequency of φ-break patterns (%)		
		'Clean six rooms.'					
Base tones		M L	H LM	LM H	GA	GB	GC
φ-break	a.	(HL HL	L M	M H)φ	28.57	100	100
Patterns	b.	(HL H)φ'	(L M	M H)φ	57.14	0	0
	c.	(HL L)φ	(L M	M H)φ	14.29	0	0

Taken together, in these three syntactic patterns of the tone group reconstruction, the φ-break after AdvPs occurs more frequently than that after head verbs, which occurs more frequently than that after CLPs. Their order of appearance in the ongoing change is that φ-breaks after AdvPs and head verbs appear earlier than those after CLPs. The Kruskal-Wallis test reports a significant difference in the use frequency of φ-breaks after AdvPs (p=0.002<0.01) and φ-breaks after head verbs (p=0.000<0.01) between age groups, while there is no significant difference in the use frequency of φ-breaks after CLPs between groups (p=0.368>0.05).

3.3. Prosodic restrictions on reconstruction of sandhi domains

Syntactic effects on tone group reconstruction are limited by prosodic restrictions. Although the φ-break after AdvPs occurs more frequently than that after head verbs, younger speakers tend to use the latter to balance the lengths of tone sandhi domains, shown in the comparison between (13) and (14).

(13)		[[kũã kin]AdvP quickly	[[pʰaʔ sau]V clean	[paŋ kuĩ]NP]V']VP room	Frequency of φ-break patterns (%)		
		'Clean the room quickly.'					
Base tones		HL HL	M L	LM H	GA	GB	GC
φ-break	a.	(H H	HL HL	M H)φ	14.29	100	100
Patterns	b.	(H H	HL H)φ'	(M H)φ	28.57	0	0
	c.	(H HL)φ	(HL HL	M H)φ	57.14	0	0

(14)		[[kũã kin]AdvP quickly	[[pʰaʔ sau]V clean	[[lak e]CLP six CL	[paŋ kuĩ]N]NP]V']VP room	Frequency of φ-break patterns (%)		
		'Clean six rooms quickly.'						
Base tones		HL HL	M L	H LM	LM H	GA	GB	GC
	a.	(H H	HL HL	L M	M H)φ	14.29	100	100
φ-break	b.	(H H	HL H)φ'	(L M	M H)φ	28.57	0	0
Patterns	c.	(H H	HL L)φ	(L M	M H)φ	28.57	0	0
	d.	(H HL)φ	(HL HL	L M	M H)φ	28.57	0	0

In (13), 57.14% of GA produce the φ-break after the AdvP. As (14) adds a CLP to increase syllables in NP, the longer length of the tone group forces young speakers to break the string into shorter fragments. Due to the lower use frequency of the φ-break after the CLP, younger speakers tend to shift the φ-break from after the AdvP to after the head verb, resulting in that the percentage of GA producing φ-breaks after the AdvP drops to 28.75%.

Moreover, prosodic restrictions on tone group reconstruction exert impacts on the number and positions of φ-breaks, shown in (15), (16) and (17).

(15)		[[ko kʰa	kũã kin]AdvP	[[pʰaʔ sau]V	[paŋ kuĩ]NP]V']VP	Frequency of φ-break patterns (%)		
		still more	quickly	clean	room			
		'Clean the room much more quickly.'						
Base tones		HL M	HL HL	M L	LM H	GA	GB	GC
φ-break	a.	(H H	H H	HL HL	M H)φ	14.29	66.67	100
Patterns	b.	(H H	H HL)φ	(HL HL	M H)φ	57.14	33.33	0
	c.	(H H	H HL)φ	(HL H)φ'	(M H)φ	28.57	0	0

(16)		[[ko kʰa	kũã kin]AdvP	[[pʰaʔ sau]V	[[lak e]CLP	[paŋ kuĩ]N]NP]V']VP	Frequency of φ-break patterns (%)		
		still more	quickly	clean	six CL	room			
		'Clean six rooms much more quickly.'							
Base tones		HL M	HL HL	M L	H LM	LM H	GA	GB	GC
φ-break	a.	(H H	H H	HL HL	L M	M H)φ	14.29	66.67	100
Patterns	b.	(H H	H HL)φ	(HL HL	L M	M H)φ	42.86	33.33	0
	c.	(H H	H HL)φ	(HL H)φ'	(L M	M H)φ	42.86	0	0

(17)		[[ko kʰa	kũã kin]AdvP	[[pʰaʔ sau]V	[[ɓap li	e]CLP	[paŋ kuĩ]N]NP]V']VP	Frequency of φ-break patterns (%)		
		still more	quickly	clean	twelve	CL	room			
		'Clean twelve rooms much more quickly.'								
Base tones		HL M	HL HL	M L	H M	LM LM	H	GA	GB	GC
φ-break	a.	(H H	H H	HL HL	L L	M M	H)φ	14.29	66.67	100
Patterns	b.	(H H	H HL)φ	(HL HL	L L	M M	H)φ	28.57	33.33	0
	c.	(H H	H HL)φ	(HL H)φ'	(L L	M M	H)φ	57.14	0	0

From (15) to (17), with the increasing syllables in the CLP, the percentage of sentences with two φ-breaks (or φ'-breaks) rises from 28.57% to 57.14% in GA. Results also demonstrate that pressures brought by prosodic restrictions may transfer from CLPs to AdvPs and head verbs.

To elucidate prosodic restrictions and make a comparison with previous studies, preferred lengths of tone group are calculated in this study. GA tend to break the string into disyllabic and tetrasyllabic fragments, while GB prefer tetrasyllabic ones and GC prefer tetrasyllabic and hexasyllabic ones.

4. CONCLUSION

The purpose of this study is to explore syntactic and prosodic effects on reconstruction of the tone sandhi domain in Xiamen Dialect.

The study finds that the tone sandhi domain in Xiamen Dialect is reconstructed into a shorter domain. Previous studies have argued that sandhi domains are delimited only by the right edge of non-adjunct XPs [13, 24]. Our study has discovered that with reconstruction in the younger generation, sandhi domains can also be confined by the right edge of adjunct XPs and head verbs. In terms of syntactic effects, φ-breaks occur more frequently after AdvPs than after head verbs and CLPs. And φ-breaks after AdvPs and head verbs appear earlier than those after CLPs in the ongoing change of tone sandhi domains. In terms of prosodic effects, lengths of tone groups exert restrictions on the number and positions of φ-breaks. Younger speakers tend to break the string into disyllabic or tetrasyllabic fragments. Syntactic effects on reconstruction of the tone sandhi domain are limited by prosodic restrictions. And an intermediate state of the φ-break is discovered in this study.

5. ACKNOWLEDGEMENTS

This work was supported by the Project of Cultural Experts and "Four Batches" of Talents directed by Aijun Li and CASS Innovation Project.

6. REFERENCES

[1] J. Guo, C. Zhou, and X. Tang, "Review of Zhang (2017): Syntax Phonology Interface: Argumentation from Tone Sandhi in Chinese Dialects," *Asian Languages and Linguistics*., vol. 2, no. 2, pp. 318–326, 2021.

[2] Y. E. Hsiao, "Tone group reconstruction in iGeneration Taiwanese," *Proceedings of the 51st International Conference on Sino-Tibetan Languages and Linguistics.*, Kyoto, 2018.

[3] Y.E. Hsiao, "Tone sandhi of young Taiwanese speakers," *Proceedings of the Linguistic Society of America*, New York, pp. 13-16, 2018.

[4] Y. E. Hsiao, "The syntax-prosody competition: Evidence from adjunct prosodic parsing in iGeneration Taiwanese," *Lingua*, vol. 237, p. 102805, Apr. 2020.

[5] Y. Cao, "Revisiting Tone Sandhi Domain in Xiamen Chinese," *Proceedings of the Annual Meeting on Phonology*, Los Angeles, 2022.

[6] Y. E. Hsiao, "The competition between syntax and rhythm in iGeneration Taiwanese," *Proceedings of the Linguistic Society of America*, vol. 5, no. 1, p. 94, Mar. 2020.

[7] C.P. Luo, "xiamen yinxi [The Phonology of Xiamen Dialect]," *The Collected Linguistic Works of Luo Changpei*, Shandong Education Press, Jinan, 1956.

[8] S.J. Cai, "Southern Min Tone Sandhi and The Allomorph Selection Hypothesis," *Contemporary Linguistics,* Chinese Academy of Social Sciences, Beijing, pp. 176–200+238, 2002.

[9] S. Duanmu, *The phonology of standard Chinese*, 2nd ed. Oxford ; New York: Oxford University Press, 2007.

[10] F.F. Hsieh, "Tonal Chain-Shifts as Anti-Neutralization-Induced Tone Sandhi," *University of Pennsylvania Working Papers in Linguistics*, Pennsylvania, pp. 99–112, 2005.

[11] J. Zhang and Y. Lai, "Phonological Knowledge beyond the Lexicon in Taiwanese Double Reduplication," *Interfaces in Chinese Phonology*, pp. 183–222, 2008.

[12] Q. Lin, "The Diachrony of Tone Sandhi: Evidence from Southern Min Chinese," Hong Kong University of Science and Technology, Hong Kong, 2017.

[13] M. Y. Chen, "The syntax of Xiamen tone sandhi," *Phonology Yearbook*., vol. 4, no. 1, pp. 109–149, May 1987.

[14] R. L. Cheng, "Tone Sandhi in Taiwanese," *Linguistics*, vol. 6, no. 41, 1968.

[15] H.-C. Hsu, "Constraint-based phonology and morphology: a survey of languages in China," Doctoral dissertation, University of California, San Diego, 1994.

[16] J.-W. Lin, "Lexical government and tone group formation in Xiamen Chinese," *Phonology*, no. 11, pp. 237–275, 1994.

[17] D. Dowty, "Grammatical Relations and Montague Grammar," In: Jacobson, P., Pullum, G.K. (eds) *The Nature of Syntactic Representation*, Springer, Dordrecht, 1982.

[18] Y. E. Hsiao, "Syntax, rhythm and tone: a triangular relationship," Doctoral dissertation, University of California, San Diego, 1991.

[19] E. Selkirk, "On derived domains in sentence phonology," . *Phonology Yearbook*., vol. 3, pp. 371–405, May 1986.

[20] Y. E. Hsiao, *Southern Min tone sandhi and theories of prosodic phonology*. Taipei, 1995.

[21] M.Y. Chen, *Tone sandhi: Patterns across Chinese dialects*, Cambridge University Press, Cambridge, 2000.

[22] P. Boersma, "Praat, A System for Doing Phonetics by Computer." *Glot Internationl*, pp. 341-345, 2001.

[23] Y.L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," *Proceedings of the ICPhS XVII*, Los Angeles, pp. 1846-1849, 2011.

[24] C. Shih, "The prosodic domain of tone sandhi in Chinese," Doctoral dissertation, University of California, San Diego, 1986.

NON-NATIVE TONE PERCEPTION – WHEN MUSIC OUTWEIGHS LANGUAGE EXPERIENCE

Antonia Götz [1,2] & Liquan Liu [1,3,4,5]

1 The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia,
2 Department of Linguistics, University of Potsdam, Potsdam, Germany,
3 School of Psychology, Western Sydney University, Sydney, Australia,
4 Center of Multilingualism across the Lifespan, University of Oslo, Oslo, Norway,
5 Centre of Excellence for the Dynamics of Language, Australian Research Council, Canberra, Australia.

ABSTRACT

This study examined how language (e.g., bilingualism, L2) and music (e.g., years of practising) experiences improve lexical tone perception. A substantial number of 532 participants from L1 Mandarin, L1 non-tone, bilingual L1 non-tone & L2 non-tone, and bilingual L1 non-tone & L2 tone backgrounds were tested on their discrimination of Mandarin tones. Results revealed that neither bilingual nor second (tone or non-tone) language experience affects novel tone perception. However, listeners' years of music training predicted perception outcomes regardless of listeners' language backgrounds. These results indicate 1) learning a tone language as L2 does not guarantee perceptual advantage of non-native tones, even after years of learning; 2) a myth of "bilingual advantage" in tone perception, challenging the bilingual enhanced acoustic sensitivity hypothesis in the perceptual domain; and 3) learning a musical instrument helps with tone perception across language groups, exhibiting a cross-domain effect in the processing of linguistic and musical pitch.

Keywords: tone perception, bilingualism, second language perception, years of musical training, cross-domain effect.

1. INTRODUCTION

In tone languages, tones distinguish word meanings. Around 60-70% of the world's languages are tonal [1], and more than half of the world population speak a tone language [2]. Most tone perception studies have been conducted on individuals from tonal languages, such as Mandarin or Cantonese, and there is a limited understanding of how individuals from non-tonal languages perceive tone and how second language learning assist in perceiving lexical tones. This project extended from existing tone perception studies and investigated whether tone experience learned from one's second language, sequential bilingual experience and music experience play a role in tone perception. Specifically, the study can inform the development of more effective language learning strategies and interventions for individuals with tone perception difficulties, as well as shed light on the cognitive benefits of bilingualism and musical training.

Some tones or tone contrasts are naturally more discriminable than others. In Mandarin tones, T1-T3 (level-dipping) appears more salient than T2-T3 (rising-dipping) for tone and non-tone language speakers [3], and T2-T4 (rising-falling) exceeds other contrasts in discriminability [4]. In Cantonese tones,

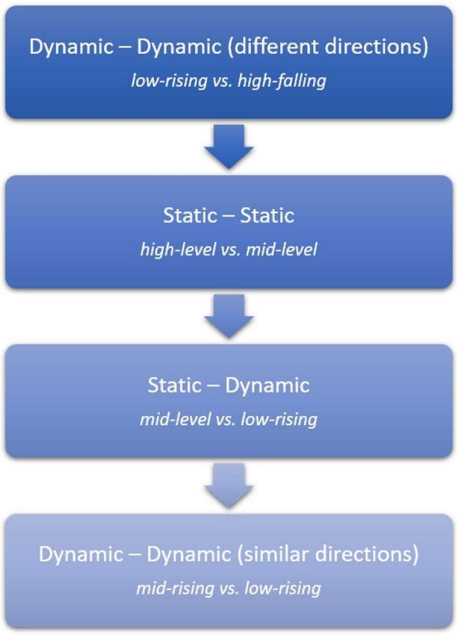


Figure 1: In [6], participants from five language backgrounds (Australian English, Cantonese, Chinese Mandarin, Singaporean Mandarin, Thai) unanimously show the same pattern when perceiving static and dynamic tones across four tone languages (Cantonese, Chinese Mandarin, Singaporean Mandarin, Thai). A salience hierarchy is hypothesised marking which type of tone contrast is easier (higher) to discriminate than others (lower) regardless of listeners' language backgrounds.

T2-T6 (high rising-low rising) leads to lower discrimination accuracy than other contrasts across listeners from tone and non-tone, monolingual and bilingual backgrounds [5]. A recent collaborative study involving participants from five tone and non-tone language backgrounds reports a salience hierarchy regarding the type of tone contrast (Figure 1). Prior research has established that tone language speakers perceive tones more categorically than their non-tone language peers [7]. What remains less clear is whether tone language speakers perceive non-native tones as well as native ones, and to what extent listeners' L2 tonal experience would lead to successful tone discrimination. One recent study testing Cantonese, Chinese, Singaporean and Thai listeners' perception of tones from their own versus other languages reveals that tone language speakers do not perceive non-native tones as good as their native ones [6]. When testing Mandarin perception by L2 Mandarin learners from L1 Cantonese or English backgrounds, the two groups do not differ in their overall accuracy on Mandarin tones, and both groups show difficulty with the acoustically most difficult T2-T3 tone pair in Mandarin [8]. Tone perception appears to be difficult even for advanced L2 learners [9], and listeners' L1 knowledge and the acoustic features of L2 tones both interfere with tone perception.

With respect to bilingualism, simultaneous bilingual experience has been shown to strengthen tone perception in infancy [10], [11] and adulthood [12] even when tone is not part of the language repertoires of the bilingual speakers. Such perceptual advantage is often attributed to the bilingual environment typically more complex than a monolingual one. The need of establishing two phonological systems may well enhance listeners' sensitivity to a third. Whether such advantage extends to the cognitive domain is under heated debate [13]. As speakers' L2 experience can also be considered as a sequential bilingual experience, it is interesting to see whether (non-tone) L2 experience may play a role, as well as whether the length of L2 learning may be a relevant factor in this case.

In terms of music experience, music training has been shown to improve non-tone language speakers' tone perception [14]–[16]. Listeners with no prior tone language experience discriminate tones more accurately when they are more musically trained [17], [18]. Some studies suggest that non-tone language speakers perceive (linguistic) tones in the same fashion of musical tunes, as their performances of pitch in language and music correlate [19]. These findings point to a domain-general effect showing that music experience can enhance non-tone language speakers' tone perception. Bilinguals and tone language speakers do not show such correlation [12], suggesting the interference of bilingual and L1 experience on tone perception. The research questions of the current study are: 1. Do different tone contrasts lead to different perceptual outcomes? 2. Does (L1 or L2) tone language or sequential bilingualism experience strengthen novel tone perception? Do listeners' years of L2 experience play a role? 3 Does music experience (years of music training) modulate tone perception? We predicted that based on the salience hierarchy [6], Mandarin T2-T4 (dynamic different) would be the most salient contrast, followed by T1-T2/T3/T4 (static-dynamic), and then T3-T2/T4 (dynamic similar). Moreover, listeners' L2 (tone, non-tone) and music experiences facilitate tone perception, and the magnitude of facilitation is directly relevant to their years of experience.

2. METHODS

2.1. Participants

A total number of 814 subjects participated in the study. Of these participants, 191 did not complete the online experiment and another 91 participants did not provide valid consent sharing data for analysis. The final sample included 532 participants from multiple language backgrounds who were divided into three language groups (Table 1).

Group	N	M_{age} (range)	Years of L2 (mean (SD))	Years of music (mean (SD)
L1 Mandarin	25	18-25	12.65 (4.55)	2.93 (2.91)
Mono (non-tone L1)	44	18-25	0.48 (0.09)	2.83 (2.93)
Bi (non-tone L1 + non-tone L2)	429	18-25	8.97 (5.14)	3.26 (3.53)
Bi (non-tone L1 + tone L2)	34	18-25	9.85 (4.94)	3.09 (3.05)

Table 1: Participants were categorised into three groups. Here, Mono refers to monolinguals and Bi refers to sequential bilinguals.L1 Mandarin speakers had L2 experience with non-tone languages but not in another tone langauge. Years of music refers to the mean and standard deviation of years of musical training that was self-indicated by the participants.

2.2. Stimuli

The stimuli consisted of 12 monosyllabic Mandarin non-words (/tou/, /bou/, /ɕye/, /pye/, /pian/, /fian/, /jy/, /ty/, /bi/, /gi/, /gua/, /lua/) with legal phonotactic structures. Each syllable was produced with the four Mandarin tones (T1, T2, T3 and T4). The length of each syllable was 250 ms. The final stimulus set consisted of 72 stimuli: 12 syllables x 6 tone contrasts (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4). A Mandarin native speaker produced six tokens of each syllable and we included two of the tokens into the experiment. By including different tokens, we prevent that the participants make their decision based on acoustic information alone. The same token was never repeated within one trial of the task. For example, in AAB trials, if token 1 was used as the A sound, token 2 would appear as the sound for X (the second A. In addition to the variability of the different tokens, the same word was never repeated within one Tone Contrast. All stimuli were normalized in intensity (70 dB).

2.3. Procedure

The experiment ran online by using the online platform Labvanced [20] at a quiet place and participants were asked to wear headphones. Each experiment started with two practice trials to familiarise the participants with the AXB discrimination task. Participants were asked to press a key as accurate and quick as possible if the second syllable was more similar to the first one (AAB, via key 1) or the third (ABB, via key 3) syllable. All contrast were also presented in the reverse order to counterbalance participants' responses. In each trial, the A and B sounds consisted of a different tone category. The interstimulus interval was 1000 ms and the intertrial interval was 3000 ms. The time-out of the response time was set to 2500 ms, measured at the end of the third syllable. Trials with responses after the time-out were not repeated. The next trial started immediately following the previous trial and trials were randomised across participants. A break was included after 25%, 50% and 75% of the experimental trials and its length was participant controlled; participants continued the experiment by pressing a key. No feedback was provided to the participants.

3. RESULTS

All statistical analyses were performed by using R [21] and the lme4 package [22]. Plots were generated by using ggplot2 [23]. General Linear Mixed Effects regression models were constructed with the maximal random and fixed factor structure with accuracy (binomial answer as 1 or 0) as dependent variable. Language background was coded as the comparison between L1 native Mandarin (coded as 0.5) and monolingual L1 (coded as -0.5), monolingual L1 (coded as 0.5) and bilingual L2 Non-tone speakers (coded as -0.5) and bilingual L2 Non-tone speakers (coded as 0.5) versus L2 Tone speakers (coded as -0.5).

Within the Tone Contrast discrimination, we predicted the following hierarchy: T2-T4 > T1-T2/T3/T4 > T3-T2/T4. Following this hierarchy, we applied the subsequent contrast comparisons: T2-T4 was compared to the mean of T1-T2, T1-T3 and T1T4, T3-T4 was compared to the mean of T1-T2, T1-T3 and T1T4, T3-T4, T3-T4 was compared to T2-T3, T1-T3 was compared to T1-T4 and T1-T2 was compared to T1-T3. Listener's discrimination of the Tone Contrast showed the expected patterns: T2-T4 yielded higher accuracy than T1-TN contrasts (β(SE) = 0.461 (0.019), z = 23.289, p < .001), which yielded higher accuracies than T3-T4 (β(SE) = 0.165(0.014), z = 11.957, p < .001), which was further higher than T2-T3 ($\beta(SE)$ = 0.254(0.024), z = 10.390, p < .001).

Our results (see also Figure 2) revealed that L1 Mandarin speakers are not better in discriminating Tone Contrasts than L1 Non-tone speakers ($\beta(SE)$ = -0.031(0.065), z = 0.478, p = 0.633). L2 tone language speakers are not better in discriminating Tone Contrasts than L2 Non-tone language speakers ($\beta(SE)$ = -0.043(0.041), z = -1.054, p = 0.291) and the listener's years of experience did not interact ($\beta(SE)$ = 0.035(0.070), z = 0.498, p = 0.619). Similarly, L2 listeners of a non-tone language were not better in discriminating the Tone Contrasts than L1 Non-tone language listeners ($\beta(SE)$ = -0.020(0.034), z = -0.594, p = 0.553), and duration of the second language did not interact ($\beta(SE)$ = -0.034(0.070), z = 0.483, p = 0.629).

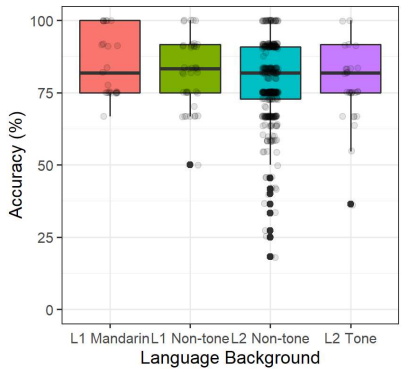


Figure 2: Accuracy results from the speech discrimination task separated by the four groups of L1 and L2 tone language experience (Mandarin, Monolingual L1 Non-tone, Bilingual L2 Non-tone, Bilingual L2 Tone)

However, music experience modulated Tone Contrast discrimination. The more music experience listeners have, the better their tone discrimination ($\beta(SE)$ = 0.011(0.005), z = 2.077, p = 0.0383).

4. DISCUSSION

The current study investigated the extent to which L2 (tone or non-tone) language and music experiences modulate tone perception. With respect to tone contrasts, a salience hierarchy as predicted was reported. Contrasting our predictions, listeners' L1 and L2 language experience does not alter tone perception, regardless of whether the language is tonal or how many years they have learned the language. However, music experience plays an important role for speakers across language backgrounds, with positive relationship between years of music training and successful tone perception outcomes.

In terms of Mandarin tone contrasts, dynamic tones with different pitch directions (T2-T4, rising-falling) were the easiest to discriminate, followed by static-dynamic tone contrasts (T1-T2/T3/T4, flat-contour tones) and then dynamic tones with similar pitch patterns (T3-T2/T4, dipping-rising/falling). Our findings replicate the salience hierarchy reported in Liu and colleagues [6], suggesting that tone acoustics is critical for listeners' tone perception across language backgrounds.

Learning a tone language as L2 does not lead to successful tone perception. Previous studies have reported similar findings for L1 speakers of a non-tone and even a tone language learning another tone language as L2 [8]. Even advanced L2 learners may find it difficult to (re-)establish a tone category [9], conforming to our results that years of L2 experience does not affect tone perception outcomes. The finding is in line with another study on segmental contrast perception. Catalan has an /e/–/ɛ/ contrast while Spanish has only one /e/ closer to and more open than the Catalan /e/. Pallier and colleagues [24] examined two types of Catalan-Spanish bilinguals' perception of the Catalan contrast. The Catalan-dominant bilinguals are exposed to Catalan since birth, whereas the Spanish-dominant bilinguals are exposed to Spanish first and get in touch with Catalan after 6 years of age when they start in kindergarten or primary school. Only Catalan-dominant but not Spanish-dominant bilinguals discriminate the Catalan contrast. Results of the vowel perception among Spanish-dominant bilinguals are in accordance with the findings of tone perception in the current study. The overall findings also demonstrate the importance of early exposure on phonological category establishment and later perception, mirroring studies reporting perceptual advantages [25], [26] and neural traces [27] of phonology of the birth language among adoptees who are adopted to a new language environment as early as 6 months after birth.

Although a simultaneous bilingual experience has been shown to facilitate tone perception [10], a sequential one does not appear to reach a similar effect. While Dutch sequential bilinguals show reduced tone discrimination abilities than their Mandarin peers, Dutch simultaneous bilinguals stand in the middle between these two groups in their performances [12]. The overall findings add to the limited data of how sequential bilingual may impact speech perception, suggesting the importance of early linguistic diversity on tone perception. The lack of significant results between the language groups may be attributed to the differences in group sizes and the resulting larger variability in Bilingual L2 Non-tone group.

Regarding music experience, the current results conform to and extend on prior studies [14]–[16]. That is, listeners do not have to be musicians to show an advantage in tone perception. Their years of music training is directly relevant to tone perception irrespective of their language backgrounds. The finding that overall music experience can facilitate speech perception provides new insights into potential cross-domain strategies in L2 language learning, especially when the target language is tonal.

5. CONCLUSION

When perceiving non-native tones, listeners' music experience plays an important role. Its effect may be more evident than L2 experiences, as neither L2 tone language nor sequential bilingual experience appears to affect perception of foreign tones. The overall findings provide implications for early, consistent diversity in and exposure to language and music.

6. REFERENCES

[1] M. Yip, *Tone*. Cambridge University Press, 2002.

[2] V. A. Fromkin, *Tone: A linguistic survey*. Academic Press, 2014.

[3] Z. Zeng, L. Liu, A. Tuninetti, V. Peter, F.-M. Tsao, and K. Mattock, "English and Mandarin native speakers' cue-weighting of lexical stress: Results from MMN and LDN," *Brain and Language*, vol. 232, p. 105151, 2022.

[4] T. Huang and K. Johnson, "Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners," *Phonetica*, vol. 67, no. 4, pp. 243–267, 2010.

[5] X. Tong, S. M. K. Lee, M. M. L. Lee, and D. Burnham, "A tale of two features: Perception of Cantonese lexical tone and English lexical stress in Cantonese-English bilinguals," *PloS one*, vol. 10, no. 11, p. e0142896, 2015.

[6] L. Liu *et al.*, "The tone atlas of perceptual discriminability and perceptual distance: Four tone languages and five language groups," *Brain and Language*, vol. 229, p. 105106, 2022.

[7] P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of Phonetics*, vol. 32, no. 3, pp. 395–421, 2004.

[8] Y.-C. Hao, "Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers," *Journal of Phonetics*, vol. 40, no. 2, pp. 269–279, Mar. 2012, doi: 10.1016/j.wocn.2011.11.001.

[9] E. Pelzl, E. F. Lau, T. Guo, and R. DeKeyser, "Advanced second language learners' perception of lexical tone contrasts," *Studies in Second Language Acquisition*, vol. 41, no. 1, pp. 59–86, 2019.

[10] L. A. Petitto, M. S. Berens, I. Kovelman, M. H. Dubins, K. Jasinska, and M. Shalinsky, "The 'Perceptual Wedge' hypothesis as the basis for bilingual babies' phonetic processing advantage: New insights from fNIRS brain imaging," *Brain Lang*, vol. 121, no. 2, pp. 130–143, May 2012, doi: 10.1016/j.bandl.2011.05.003.

[11] L. Liu and R. Kager, "Perception of tones by bilingual infants learning non-tone languages," *Bilingualism: Language and Cognition*, vol. 20, no. 3, pp. 561–575, 2017.

[12] L. Liu, A. Chen, and R. Kager, "Simultaneous bilinguals who do not speak a tone language show enhancement in pitch sensitivity but not in executive function," *Linguistic Approaches to Bilingualism*, vol. 12, no. 3, pp. 310–346, 2022.

[13] M. Lehtonen, A. Soveri, A. Laine, J. Järvenpää, A. De Bruin, and J. Antfolk, "Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review.," *Psychological bulletin*, vol. 144, no. 4, p. 394, 2018.

[14] D. Burnham, R. Brooker, and A. Reid, "The effects of absolute pitch ability and musical training on lexical tone perception," *Psychology of Music*, vol. 43, no. 6, pp. 881–897, Nov. 2015, doi: 10.1177/0305735614546359.

[15] A. Cooper and Y. Wang, "The influence of linguistic and musical experience on Cantonese word learning," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4756–4769, 2012.

[16] P. C. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Applied Psycholinguistics*, vol. 28, no. 04, pp. 565–585, 2007.

[17] J. A. Alexander, P. C. M. Wong, and A. Bradlow, "Lexical tone perception in musicians and non-musicians," presented at the European Conference on Speech Communication and Technology, Lissabon, Portugal, 2005.

[18] B. Chandrasekaran, A. Krishnan, and J. T. Gandour, "Relative influence of musical and linguistic experience on early cortical processing of pitch contours," *Brain Lang*, vol. 108, no. 1, pp. 1–9, Jan. 2009, doi: 10.1016/j.bandl.2008.02.001.

[19] A. Chen, L. Liu, and R. Kager, "Cross-domain correlation in pitch perception, the influence of native language," *Language, Cognition and Neuroscience*, vol. 31, no. 6, pp. 751–760, 2016.

[20] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "LabVanced: a unified JavaScript framework for online studies," in *International conference on computational social science (cologne)*, 2017.

[21] R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing." Vienna, Austria., 2021. [Online]. Available: https://www.R-project.org/

[22] D. M. Bates, Maechler, M., Bolker, B., and Walker, S., "Fitting Linear Mixed-Effects Models using lme4," *Journal of Statistical Software*, 2015.

[23] H. Wickham, "Programming with ggplot2," in *ggplot2: Elegant Graphics for Data Analysis*, H. Wickham, Ed. Cham: Springer International Publishing, 2016, pp. 241–253. doi: 10.1007/978-3-319-24277-4_12.

[24] C. Pallier, L. Bosch, and N. Sebastián-Gallés, "A limit on behavioral plasticity in speech perception," *Cognition*, vol. 64, no. 3, pp. B9–B17, 1997.

[25] W. Choi, X. Tong, F. Gu, X. Tong, and L. Wong, "On the early neural perceptual integrality of tones and vowels," *Journal of Neurolinguistics*, vol. 41, pp. 11–23, 2017.

[26] W. Choi, X. Tong, and S. H. Deacon, "Double dissociations in reading comprehension difficulties among Chinese–English bilinguals and their association with tone awareness," *Journal of Research in Reading*, vol. 40, no. 2, pp. 184–198, 2017.

[27] L. J. Pierce, J.-K. Chen, A. Delcenserie, F. Genesee, and D. Klein, "Past experience shapes ongoing neural patterns for language," *Nature Communications*, vol. 6, no. 1, pp. 1–11, 2015.

DIALECTAL DIFFERENCES IN THE REALISATION OF TONE 1 IN TAIWAN MANDARIN

Janice Fon[1] & Yu-Ying Chuang[2]

[1]National Taiwan University, Taiwan; [2]Eberhard-Karls University of Tübingen, Germany

jfon@ntu.edu.tw; yychuang2012@gmail.com

ABSTRACT

The central dialect of Taiwan Mandarin is said to have an utterance-final rise. Previous literature indicated that the low-fall Tone 3 and the high-fall Tone 4 in the standard northern dialect are realised respectively as low-dipping and high-level utterance-finally in the central variety. This study thus explored whether such a tendency also extended to the high-level Tone 1. Twenty-five native speakers (13 northern and 12 central) were recruited to read 24 bisyllabic stimuli containing Tone 1. Results showed rising was indeed the predominant realisation for Tone 1, but it was not exclusive to the central dialect, nor was it to the utterance-final position. Reference points for the rising portion also did not show a more drastic rise for the central variety. Rather, central speakers consistently realised their Tone 1 with a steeper initial fall before the rise, likely creating a percept of a more prominent rise than their northern counterparts.

Keywords: Dialectal difference, Tone 1, Level tone, Tonal realisation, Taiwan Mandarin.

1. INTRODUCTION

Mandarin is the official language of Taiwan, a small island country in Southeast Asia. It is a tone language with four lexical tones, a high-level Tone 1, a mid-rising Tone 2, a low-falling Tone 3, and a high-falling Tone 4 [1, 2]. Although the language is vibrant with phonetic variations due to frequent contact with Min, a substrate language and a former *lingua franca* [3], most are in fact nontonal [4, 5]. When it comes to tonal variations, however, many would unanimously point to the central dialect [6].

Two traits are often mentioned by the general public [6]. First, the central dialect favours a lower pitch register and a narrower pitch range than the northern standard dialect, and the effect is especially prominent on high targets [6, 7]. Min proficiency might be an underlying cause for pitch register lowering [7], and pitch range narrowing likely stemmed from a negative transfer from the local Min variety, in which the phonologically high-level tone is realised as phonetically mid-level instead [6].

Secondly, the central dialect tends to have an utterance-final rise [6]. Although this trait is often mentioned in the mass media, systematic research is rare. To the best of our knowledge, there are only two relevant studies. Fu [8] found that junior high schoolers from central Taiwan tend to realise their utterance-final Tone 3 as dipping, instead of the default low fall. Similarly, Wu [9] found that adult central speakers are inclined to realise their utterance-final Tone 4 as high level, rather than the prescribed high fall. Both studies adopted subjective judgement as a means for tonal shape identification, and acoustic analyses are yet to be done.

To evaluate whether the preference for a final rise is specific to only Tone 3 and Tone 4 or common across all four tones, this study extended the inquiry to Tone 1, which would potentially be an ideal candidate since it is realised with lower pitch in the central dialect [6, 7], and any preference for ending with a rise could thus be readily detected.

This study has two aims. First, we would like to use acoustic measures to see if Tone 1 indeed tends to be realised with a rise in the central variety. Previous studies [8, 9] used only subjective judgments, which are efficient for determining phonological categories, but difficult to detail phonetic differences. This study would thus like to complement subjective judgments with acoustic measurements so as to provide a fuller view of the dialect. Secondly, we intend to directly compare the central variety with the northern dialect. Although studies regarding both varieties are existent (cf. northern: [1, 4, 5, 10, 11]; central: [8, 9]), few have included both for direct comparisons, except for Fon [2], which focused on tonal contour, and Huang & Fon [7] and Khoo [6], which focused on tonal range and register. None has examined the utterance-final rise. Therefore, we included both dialects to directly gauge the degree of divergence between the two.

2. METHOD

2.1. Subjects

Twenty-five native Mandarin speakers aged between 18 and 25 were recruited in this study ($\overline{X}$= 20.72, *SD* = 1.97). About half were from northern Taiwan (*N* =

13), and half from central. The two groups were comparable in age [$t(23) = -0.07$, ns]. All subjects were born and raised in their respective areas and had not lived outside the region for more than six months before 18 years old. Each group was about equally divided in gender.

2.2. Stimuli

Three Tone 1 syllables, *ban*[1], *dan*[1], and *gan*[1], were chosen as target stimuli and were paired with another syllable to form disyllabic words. The target stimuli were placed in either the first or the second position, and the pairing syllables were in one of the four tones (e.g., ***dan***[1]*xin*[1] 'to worry' *vs.* *ming*[2]***dan***[1] 'roster'). In total, there were 3 (syllables) × 2 (positions) × 4 (tonal environments) = 24 stimuli. An additional set of 72 bisyllabic fillers that did not contain any of the target syllables was also included.

2.3. Equipment

Recording was done with a sampling rate of 48 kHz using a SONY PCM-M1 Digital Audio Recorder and a SHURE SM10A head-mounted microphone, and was later downsampled to 22050 kHz using Praat 6.1 [12].

2.4. Procedure

Subjects read the stimuli in a clear and natural fashion in a quiet room. They were randomly assigned to one of the three semi-randomization orders. Words containing the same target syllable were hand-adjusted to avoid juxtaposition.

2.5. Analyses

The voiced portions of the stimuli were hand-labelled using Praat [12], and pitch contours extracted by a Praat script were interpolated and smoothed after hand correction for pitch-doubling and -halving. Tonal shapes of the stimuli were hand-labelled based on independent perceptual judgments on the pitch excursions by three native speaker judges, two of whom were the authors.

To facilitate comparisons across different tokens, two sets of pitch information were extracted for further analyses. The first included reference points from each token (cf. [1, 6, 7, 10]). For the rising contour, it is the initial minimum and final maximum of the rising portion. The second set extracted ten pitch points at equal time intervals from each token, so that analyses could be performed on time-normalised tonal contours.

3. RESULTS

In total, 24 (stimuli) × 25 (subjects) = 600 tokens of stimuli were collected, among which 439 were perceived as rising by at least two judges, accounting for 73% of the data (Table 1). In other words, rising is the most prevalent for both dialects, and all speakers adopted it to various extents. Fig. 1 shows the pitch excursion for the perceived level and rising contours for both genders. The two tonal excursions occupied similar pitch registers, and rising was realised as slightly dipping acoustically, which is common in Taiwan Mandarin [1, 2, 10].

Table 1: Perceived tonal contours of the target stimuli by at least two of the three judges. Numbers before the slashes are counts for the first syllable and those after are counts for the second.

	Level	Rising	Others	Total
Northern	48/12	106/142	2/2	156/156
Central	60/18	71/120	13/6	144/144
Total	108/30	177/262	15/8	600

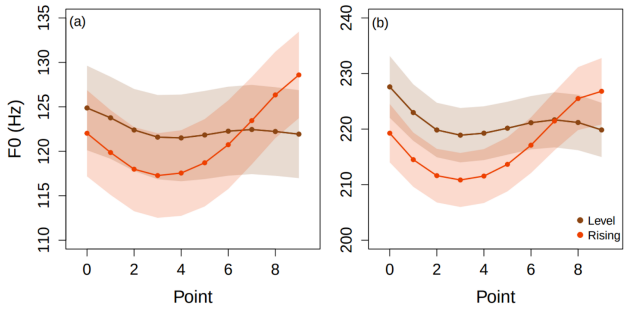


Figure 1: Time-normalised mean pitch excursions of perceived level and rising contours in (a) male and (b) female speakers. Shaded areas represent standard error.

To study how the distribution of rising contours was affected by various factors, a GENDER × DIALECT × POSITION × TONE four-way mixed ANOVA was performed on the rising percentages of each subject. Results showed that all main effects except for DIALECT were significant [GENDER: $F(1, 21) = 4.86$, $p < .05$; POSITION: $F(1, 21) = 28.37$, $p < .0001$; TONE: $F(3, 63) = 5.92$, $p < .01$]. The interaction between POSITION and TONE was also significant [$F(3, 63) = 2.82$, $p < .05$]. Post hoc analyses using Bonferroni's adjustments showed that females had more rising realisations than males (Fig. 2), and final syllables had more rising than prefinals (Fig. 3). With regard to the interaction, there was an effect of TONE for the prefinal position. Tone 1 pairing syllables elicited fewer rising realisations than Tone 2 ($p = .06$) and Tone 3 ($p < .05$), as shown in Fig. 3. No effect involving TONE was found for the final position.

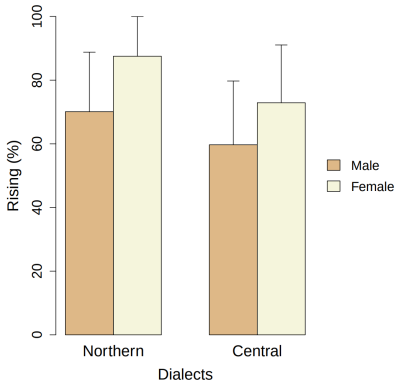


Figure 2: Rising percentages regarding GENDER and DIALECT. Error bars represent standard error.

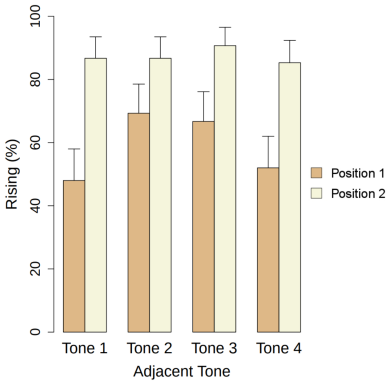


Figure 3: Rising percentages regarding POSITION and TONE. Error bars represent standard error.

Fig. 4 shows the mean pitch of the initial and final points of the rise. Two separate linear mixed effects analyses were performed on males and females using the `lme4` package [13] in R [14], as shown in (1). Reference pitch values were included as the dependent variable, and fixed effects of DIALECT, POSITION, and POINT, along with their interaction terms, were entered into the models. Random effects included by-subject and by-item intercepts, as well as the by-subject random slopes for POINT. Significance was calculated using normal approximation [15].

(1) F0 ~ DIALECT * POSITION * POINT
+ (1+POINT|subject) + (1|item)

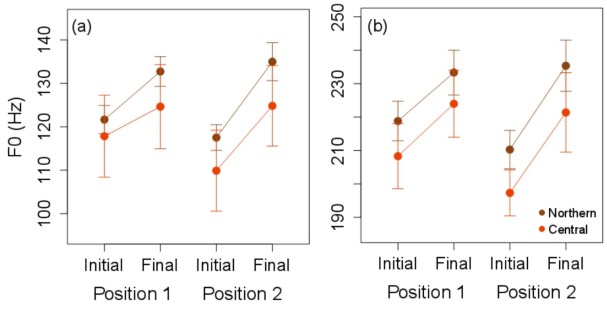


Figure 4: Mean pitch of the initial and final points of the rise across the two dialects for (a) males and (b) females.

Table 2 shows the fixed effects for males and females. For both genders, the main effects of POSITION and POINT were significant. This was expected, since reference points were extracted from the rising portion of a tone, and there was a slight declination effect for the initial reference point. A two-way interaction of POSITION × POINT was also significant. The second syllable consistently showed a larger rise than the first.

Table 2: Fixed effects of the linear mixed models for males and females using the initial point of the first position in the northern dialect as the reference. '**' $p < .01$, '***' $p < .001$.

	Est.	SE	*t* value	
		Male		
Intercept	109.86	6.74	16.29	***
Dialect2	-0.68	9.56	-0.07	
Position2	-9.56	3.30	-2.90	**
Point2	10.86	2.15	5.06	***
Dialect2:Position2	-5.68	4.54	-1.25	
Dialect2:Point2	-4.28	3.19	-1.34	
Position2:Point2	6.67	1.90	3.52	***
D2:Pn2:Pt2	2.36	2.86	0.83	
		Female		
Intercept	203.86	6.04	33.76	***
Dialect2	-8.22	8.92	-0.92	
Position2	-19.00	4.42	-4.30	***
Point2	14.49	3.06	4.73	***
Dialect2:Position2	-4.48	6.21	-0.72	
Dialect2:Point2	-1.08	4.72	-0.23	
Position2:Point2	10.84	2.49	4.36	***
D2:Pn2:Pt2	0.41	3.92	0.11	

Fig. 5 shows the 10-point pitch extractions of the rising realisations from both dialects. To compare the tonal contours, the generalised additive mixed model (GAMM), a powerful statistical method that could mathematically separate tonal undulations into pitch height and pitch excursions [16], was adopted as a means of analyses to accommodate the fact that the contours were not linear. Separate GAMM models were built on males and females using the first position of the northern dialect as the reference to fit the pitch extraction values using the `bam` function of the `mgcv` package [17] in R [14], as shown in (2).

(2) F0 ~ POSITION
+ DIALECT2.POSITION1
+ DIALECT2.POSITION2
+ s(POINT, by = POSITION)
+ s(POINT, by = DIALECT2.POSITION1)
+ s(POINT, by = DIALECT2.POSITION2)
+ s(subj, bs = "re") + s(syll, bs = "re")

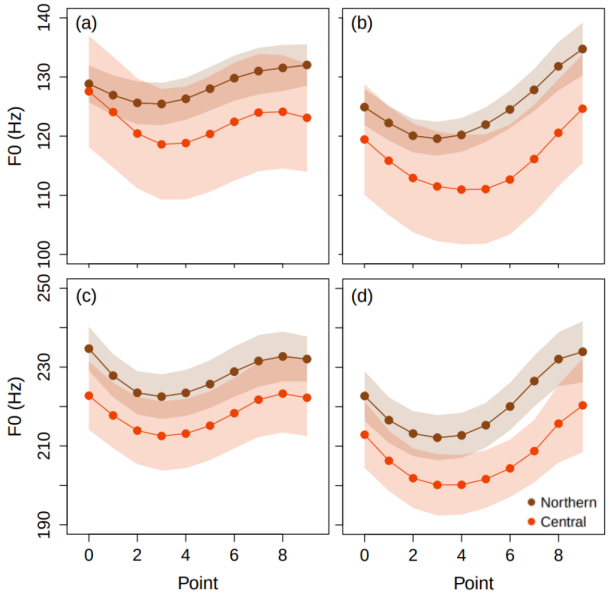


Figure 5: Ten-point pitch extractions in the two dialects for males (a-b) and females (c-d). The first column is the prefinal position and the second is the final. Shaded areas represent standard error.

Table 3 shows the smooth terms of the GAMM results of the difference smooths, which directly compared the tonal excursions of the two dialects while disregarding differences in pitch register. Results showed that except for the prefinal position in females, central speakers differed consistently from their northern counterparts in terms of their tonal contours. Fig. 6 shows the difference smooths predicted by GAMM, which generally started high in the beginning, and gradually declined toward the end. Since the reference level of DIALECT was set at the northern variety, this suggests that compared to their northern counterparts, central speakers tended to have a steeper fall in the beginning, resulting in a large positive difference that gradually diminished, followed by a shallower rise, resulting in a negative difference, the magnitude of which gradually increased.

4. CONCLUSION

Results in this study were interesting yet surprising, as some of the previously assumed patterns were blatantly disconfirmed (cf. [6]). Although rising was indeed found to be fairly common for Tone 1 in the central dialect, as was predicted from studies on other tones [8, 9], it was not exclusive to the variety, nor was it to the final position. This implies the cues that formed the impression of the general public are definitely not as straightforward as one had previously assumed. Based on the results of this study, both dialects showed a preference for a rising Tone 1, and had a comparable degree of rising realisation and a similar extent of the rise. However, the central dialect consistently showed a steeper fall before the rise, which potentially created a stronger contrast to the following rise and thus a percept of a more prominent rise. While the initial falling portion is usually not considered phonologically relevant in a rising contour, perception studies have shown that it could still effectively contribute to the percept of a rising tone [10]. Therefore, we suspect that the steepness of the initial fall might play a key role in forming the impression of the general public for the central dialect. This would merit further studies.

Table 3: Smooth terms of GAMM fitted to male and female F0 contour shape data using the first position from the northern dialect as the reference level.'*' $p < .05$, '**' $p < .01$, '***' $p < .001$.

Smooth terms	edf	Ref.df	*F* value	
		Male		
Position1	4.29	5.29	15.38	***
Position2	4.67	5.75	35.23	***
Dialect2:Position1	1.30	1.53	7.19	**
Dialect2:Position2	2.13	2.65	5.30	**
Subject	9.99	12.00	945.71	
Syllable	1.26	2.00	23.39	
		Female		
Position1	5.10	6.22	24.31	***
Position2	5.37	6.54	76.66	***
Dialect2:Position1	1.00	1.00	0.02	
Dialect2:Position2	1.00	1.00	21.98	***
Subject	10.98	13.00	547.39	
Syllable	1.50	2.00	14.49	

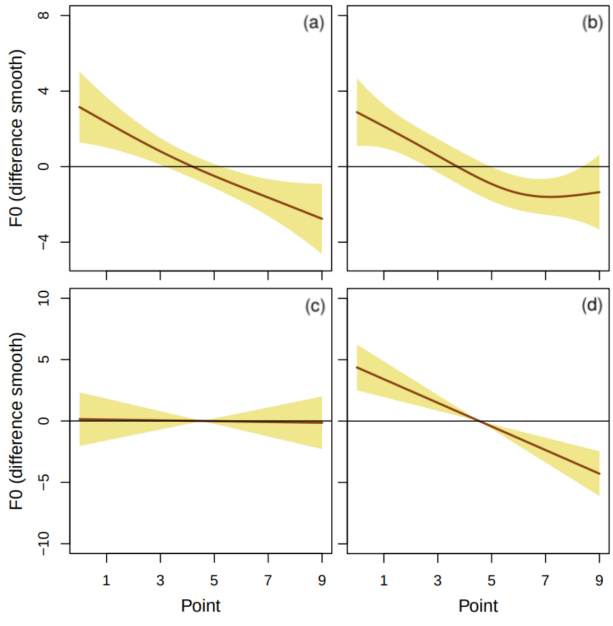


Figure 6: Difference smooths predicted by GAMM for males (a-b) and females (c-d). The first column is the prefinal position and the second is the final. Shaded areas represent two standard errors.

7. REFERENCES

[1] Fon, J., Chiang, W.-Y. 1999. What does Chao have to say about tones? --A case study of Taiwan Mandarin. *Journal of Chinese Linguistics*, 27, 15–37.

[2] Fon, J. 2020. The phonetic realizations of the Mandarin phoneme inventory: The canonical and the variants. In: Liu, H.-M., Tsao, F.-M., Li, P. (eds), *Speech Perception, Production and Acquisition - Multidisciplinary approaches in Chinese languages*. S Springer, 11–36.

[3] Huang, S. 1993. *Language, society, and ethnic identity: A study on language sociology in Taiwan.* Crane.

[4] Kubler, C. C. 1985a. *The Development of Mandarin in Taiwan: A case study of language contact*. Student Book.

[5] Kubler, C. C. 1985b. The influence of Southern Min on the Mandarin of Taiwan. *Anthropological Linguistics*, 27, 156–176.

[6] Khoo, H. 2020. A preliminary study of the tonal features of central Taiwan Mandarin. *Taiwan Journal of Linguistics*, 18, 115–157.

[7] Huang, Y.-H., Fon, J. 2011. Investigating the effect of Min on dialectal variations of Mandarin tonal realization. *Proc. 17th ICPhS Hong Kong*, 918–921.

[8] Fu, J.-W. 1999. Chinese tonal variation and social network-a case study in Tantzu Junior High School Taichung, Taiwan. MA thesis. Providence University.

[9] Wu, S.-J. 2003. A sociolinguistic study of Chinese tonal variation in Puli, Nantou, Taiwan. MA Thesis. Providence University.

[10] Fon, J., Chiang, W.-Y., Cheung, H. 2004. Production and perception of two dipping tones (T2 and T3) in Taiwan Mandarin. *Journal of Chinese Linguistics*, 32, 249–280.

[11] Tsao, F.-F. 2000. Taiwanized Japanese and Taiwan Mandarin--Two Case Studies of Language Contact during the Past Hundred Years in Taiwan. *Hanxue Yanjiu (Chinese Studies)*, 18, 273–297.

[12] Boersma, P., Weenink, D. 2009. Praat: Doing phonetics by computer (Version 5.1) [Computer software]. http://www.praat.org/

[13] Bates, D., Maechler, M., Bolker, B., Walker, S., & Others. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, 1, 1–23.

[14] R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[15] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.

[16] Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC Press.

[17] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73, 3–36.

DISYLLABIC TONE SANDHI PATTERNS IN SHAOXING WU CHINESE

Xinyi Wen, Yiya Chen, Lisa Lai-Shen Cheng

Leiden University Centre for Linguistics, Leiden University, the Netherlands
x.wen@hum.leidenuniv.nl, yiya.chen@hum.leidenuniv.nl, l.l.cheng@hum.leidenuniv.nl

ABSTRACT

We report acoustic data on disyllabic tone sandhi patterns in Shaoxing Wu Chinese. With quantitative analyses, we confirmed the impressionistic observations that 1) both tonal register and contour condition disyllabic tone sandhi forms in Shaoxing Wu Chinese; and 2) not only the initial but also the non-initial tones play a role. These findings present novel experimental evidence on the independence of register and contour for lexical tone representations and sandhi processes. Furthermore, they provide a more nuanced picture of the left- vs. right-dominant tone sandhi patterns in Northern Wu dialects; both can be present in the same dialect. No durational correlates of the left- and right-dominant sandhi patterns, however, were found, raising questions about the phonetic substance of metrical prominence structure based on tone sandhi dominance patterns.

Keywords: Tone Sandhi, Northern Wu Chinese, Tonal Register, Contour

1. INTRODUCTION

Lexical tones can contrast in both pitch register and contour [1]. The structural relationship between the two has been a long-debated issue [2], with different models proposed on tonal representations (e.g., [3], [4]). These models differ in their predictions of whether tonal pitch register and contour are independent or interdependent. While some data suggest that they can spread separately and therefore are independent (e.g., [5], [6]), these observations have been drawn mainly upon introspective impressionistic descriptions. Experimental results are therefore needed to confirm the tonal patterns and investigate further how exactly tonal register and contour interact. One goal of this paper is to report data from the *f0* realizations of disyllabic tone sandhi patterns in Shaoxing Wu Chinese, aiming to shed light on this issue and provide insights on the way register and contour give rise to the surface *f0* contours of contextual tonal realization.

Shaoxing Wu Chinese is a Northern Wu dialect, known to contrast in both register and contour. There are eight lexical tones, with four different contours in both high and low pitch registers (e.g., [7], [8]), as illustrated in Figure 1. Impressionistic studies have shown that tone sandhi patterns of disyllabic words in Shaoxing Wu are conditioned by both tonal register and contour of initial syllables. In addition, the contours of final syllables are preserved when following an initial falling tone, though neutralized following a rising tone [9]. Such complex tone sandhi patterns present an interesting test case on how tonal register and contour condition contextual tonal realization. Yet, their effects on *f0* trajectories have not been thoroughly examined with controlled experiments.

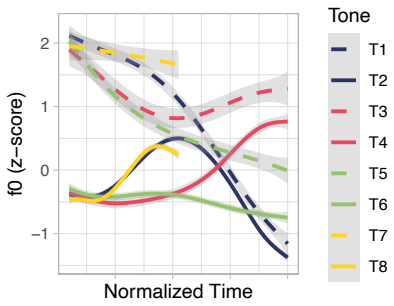


Figure 1: Lexical tones in Shaoxing Wu.

The impressionistically observed disyllabic tone sandhi patterns, if borne out by experimental data, will shed light on the issue of left- vs. right-dominant sandhi patterns in Wu Chinese. Words in Northern and Southern Wu are known to differ in their tone sandhi patterns (e.g., [6], [10]). The former tends to undergo left-dominant sandhi, with the initial tone determining the sandhi form and the non-initial tone often neutralized. The latter undergoes right-dominant sandhi, with the non-initial tone preserving the citation form and the initial one neutralized. However, there are also Northern Wu dialects with the non-initial tone conditioning tone sandhi (e.g., [11], [12]), and the effects of register and contour vary across dialects [13]. Given the potential role tones in the final position play in Shaoxing Wu, the second goal of this paper is to report on data which enable us to have a detailed look at how tones in the final position condition sandhi patterns, especially the effects of register and contour separately. This would further our understanding of the right-dominant patterns of tone sandhi in Northern Wu dialects, in which the prevalent sandhi pattern is left-dominant.

Related to the left vs. right-dominant sandhi is the proposal that such a tone sandhi difference is due to a typological difference in metrical prominence across

Chinese dialects. In the existing literature, it has been argued that only prominent syllables can preserve underlying tones (e.g., [6], [14]). Yet, few of them provided evidence other than tonal patterns, leading to a circular argument for the relationship between sandhi-dominant patterns and metrical structures. While evidence from cross-dialect comparisons suggests the need to delink metrical prominence from durational correlates [15], it is nevertheless worth investigating whether there is a durational difference between words undergoing left- vs. right-dominant tone sandhi within the same dialect. As mentioned earlier, in Shaoxing Wu Chinese, words starting with falling tones resemble right-dominant tone sandhi patterns, in contrast to words starting with rising tones that show left-dominant tone sandhi patterns. Corresponding differences in syllable duration for words with falling vs. rising tones in the initial syllable can elucidate the potential link between tone sandhi-derived metrical prominence structure and its possible durational correlates.

In sum, this study aimed to address the following research questions: (1) How do tonal registers and contours of initial (σ1) and final syllables (σ2) condition the *f0* trajectories of disyllabic words in Shaoxing Wu? (2) Do words starting with falling and rising tones differ in their duration patterns? In doing so, we will better understand the interplay of register and contour in contextual tonal realizations as well as the potential link between sandhi dominance and metrical prominence.

2. METHOD

2.1. Participants

16 middle-aged native speakers of Shaoxing Wu were recruited as participants of the current study (10 female and 6 male, mean = 50.7, SD = 3.6). They were all born and raised in the urban area of Shaoxing. They use Shaoxing Wu frequently for daily communication. None of them reported difficulties in speaking or hearing.

2.2. Stimuli

A list of disyllabic words was constructed as the stimuli. To examine whether and how registers and contours of σ1 and σ2 condition sandhi contours, the tone of both syllables was controlled as follows: for *registers*, both High and Low register tones were included; for *contours*, Rising and Falling tones were included. Two different items were included for each tonal combination. In total, 512 tokens were collected (2 σ1 tonal registers × 2 σ1 tonal contours × 2 σ2 tonal registers × 2 σ2 tonal contours × 2 items × 16 speakers).

2.3. Procedures

Recordings were conducted in a quiet room in Shaoxing, using a laptop equipped with an external sound card (Cakewalk UA-G1) and a headset (Sennheiser PC150). Stimuli were displayed on the screen with a randomized order for each participant using FieldPhon [16]. Participants read the stimuli aloud at a normal speech rate, which were recorded with 44 kHz sampling rate and 16-bit rate.

2.4. Data Preparation

Data from one male speaker were excluded from analysis due to creakiness in most of the tokens produced. Another 52 tokens were excluded due to disfluency or mistakes in production, as judged by a native speaker of Shaoxing Wu. As a result, 428 tokens from 15 speakers were included for further analysis. The sound files were manually segmented using Praat. *f0* values in Hz were extracted by taking ten equidistant points in the rime of each syllable and normalized to within-speaker z-scores. Rime duration of each syllable was also extracted.

2.5. Data Analysis

To analyse the *f0* trajectories, generalized additive mixed models (GAMM; [17]) were applied with the *mgcv* package in R [18]. The response of the model is the normalized *f0* value of disyllabic words. To understand whether and how the underlying registers and contours of both syllables condition disyllabic tone sandhi patterns, it is important to formally assess their effects on the height or shape of *f0* trajectories separately. We first created a variable as the predictor to present the interaction between underlying tonal registers and contours of σ1 and σ2, and then re-specified the models with ordered-factor difference smooths (following [19]). Factor smooths modelling non-linear differences over time for each speaker and each item were included as random effects.

For the analysis of duration, linear mixed-effects models were built with the *lme4* package in R [18]. Tonal context (start with falling vs. rising tone), syllable position (σ1 vs. σ2), and their interaction were entered into the models as fixed factors in a stepwise fashion. By-speaker and by-item intercepts and slopes for the effect of syllable position were included as random effects.

3. RESULTS

3.1. *f0* trajectories

We first investigated the effect of σ1 register and contour on the *f0* trajectories of disyllabic words.

Fitted *f0* trajectories of words ending with High Falling, Low Falling, High Rising, and Low Rising tones (T1-T4) were plotted in Figure 2A-D. As shown in each panel of Figure 2, the contrasts of register and contour are preserved on σ1 (before the vertical dashed lines). Underlyingly high register tones are realized higher than low register tones, and underlyingly falling and rising tones differ in their surface contours: falling tones surface with high falling and low level pitch contours, while rising tones surface with high level and low rising contours. Zooming in on the *f0* realizations of σ2 (after the vertical dashed lines), they are also conditioned by the underlying contour of σ1. For words ending with rising tones (Figure 2C and 2D), their σ2 surface with falling tones after an initial rising tone, and surface with level-ish contours after a falling tone.

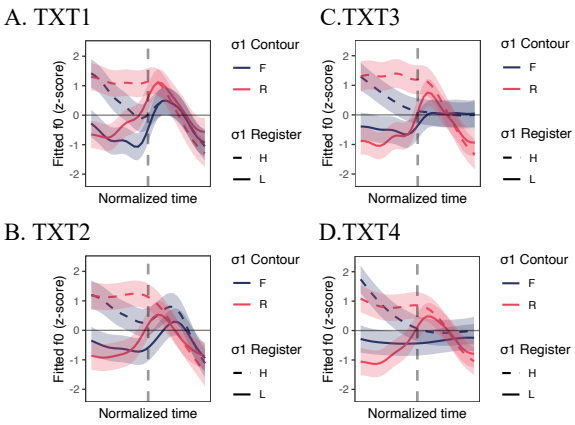


Figure 2: Fitted *f0* trajectories of disyllabic words ending with the same tone. Line types represent the register, colours the contour of σ1, and vertical dashed line syllable boundary.

To illustrate the effect of σ2 register and contour, fitted *f0* trajectories were re-plotted by presenting words starting with the same tone in the same panel of Figure 3.

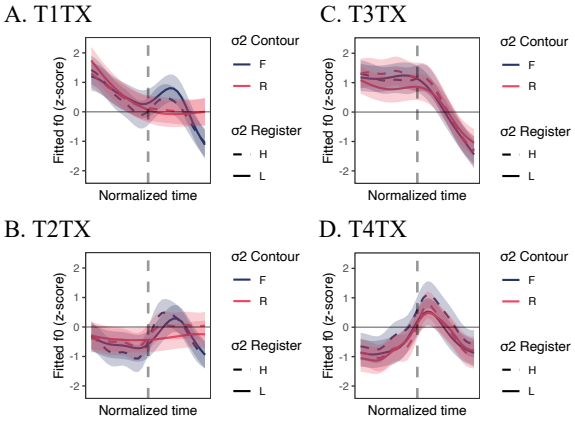


Figure 3: Fitted *f0* trajectories of disyllabic words starting with the same tone. Line types and colours represent the register and contour of σ2 respectively.

Figure 3 illustrates that the *f0* realizations of σ1 are neither conditioned by the underlying register nor the contour of σ2: T1-T4 surface with high falling, low level, high level, and low rising pitch contours respectively, regardless of the tonal identities of their following syllables. As for the *f0* realizations of σ2, the underlying registers of σ2 seem to be neutralized: in each panel, underlyingly high register and low register tones surface alike. Regarding the effect of σ2 contour, however, there is a sharp contrast between words starting with underlyingly falling (T1 and T2, Figure 3A-B) and rising tones (T3 and T4, Figure 3C-D). Following a falling tone, the contours of σ2 are preserved to some extent. After a rising tone, σ2 are neutralized to a falling contour.

The statistical results in Table 1-4 further confirmed the above observations. Briefly speaking, both the register and contour of σ1 showed a significant effect, while only the contour of σ2 had a significant effect in words starting with falling tones.

σ1 × σ2		*β*	SE	*t*	*p*
F × H.F	H − L	0.678	0.254	2.670	0.015 *
R × H.F	H − L	0.597	0.248	2.404	0.032 *
F × L.F	H − L	0.705	0.236	2.978	0.006 **
R × L.F	H − L	0.913	0.256	3.567	< 0.001 ***
F × H.R	H − L	0.630	0.247	2.548	0.022 *
R × H.R	H − L	1.065	0.264	4.022	< 0.001 ***
F × L.R	H − L	0.773	0.250	3.087	0.004 **
R × L.R	H − L	0.902	0.233	3.871	< 0.001 ***

Table 1: Parametric terms modelling the constant difference between disyllabic words starting with High and Low register tones.

σ1 × σ2		edf	*F*	*p*
H × H.F	F − R	4.039	6.942	< 0.001 ***
L × H.F	F − R	3.137	2.439	0.068
H × L.F	F − R	5.279	10.118	< 0.001 ***
L × L.F	F − R	5.255	8.431	< 0.001 ***
H × H.R	F − R	1.000	58.777	< 0.001 ***
L × H.R	F − R	1.000	34.672	< 0.001 ***
H × L.R	F − R	1.000	49.57	< 0.001 ***
L × L.R	F − R	3.637	7.734	< 0.001 ***

Table 2: Smooth terms modelling the non-linear difference between disyllabic words starting with Falling and Rising tones.

σ1 × σ2		*β*	SE	*t*	*p*
H.F × F	H − L	-0.179	0.287	-0.622	1
H.R × F	H − L	0.193	0.272	0.711	0.955
L.F × F	H − L	0.190	0.277	0.685	0.987
L.R × F	H − L	0.451	0.273	1.649	0.198
H.F × R	H − L	0.082	0.273	0.301	1
H.R × R	H − L	0.044	0.288	0.153	1
L.F × R	H − L	0.430	0.273	1.576	0.230
L.R × R	H − L	0.095	0.273	0.349	1

Table 3: Parametric terms modelling the constant difference between disyllabic words ending with High and Low register tones.

σ1 × σ2		edf	*F*	*p*
H.F × H	F − R	4.927	9.765	< 0.001 ***
H.R × H	F − R	1.000	0.197	1
L.F × H	F − R	4.796	9.131	< 0.001 ***
L.R × H	F − R	1.623	0.267	1
H.F × L	F − R	5.520	12.697	< 0.001 ***
H.R × L	F − R	1.000	3.488	0.124
L.F × L	F − R	5.444	10.445	< 0.001 ***
L.R × L	F − R	1.002	0.498	0.956

Table 4: Smooth terms modelling the non-linear difference between disyllabic words ending with Falling and Rising tones.

3.2. Duration

Figure 4 shows that there is a difference in the duration of σ1 and σ2 in words starting with falling and rising tones, respectively. The effect of tonal context was significant, with the mean duration of words starting with falling tones being longer than those starting with rising tones ($\chi2(1) = 18.62$, $p < 0.001$). Additionally, the mean duration of σ2 was significantly longer than that of σ1 ($\chi2(1) = 15.64$, $p < 0.001$). The interaction between tonal context and syllable position was significant ($\chi2(1) = 11.86$, $p < 0.001$): the mean duration of σ2 was significantly longer than that of σ1 in both words starting with falling ($\chi2(1) = 19.42$, $p < 0.001$) and rising tones ($\chi2(1) = 5.66$, $p = 0.017$), and the interaction is mainly due to different magnitudes of lengthening in σ2.

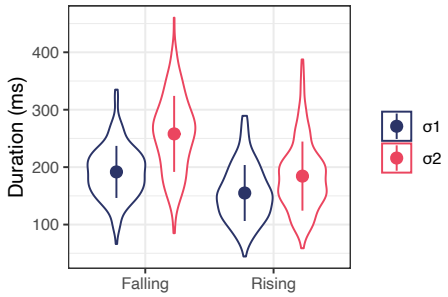


Figure 4: Duration of σ1 (blue) and σ2 (red) in words starting with falling and rising tones. The error bars represent the standard deviation, and the outlines of the violin plot indicate the kernel probability density.

4. DISCUSSION AND CONCLUSION

Our results demonstrated that in Shaoxing Wu, both the tonal register and contour of σ1 have significant effects on the *f0* trajectories of a disyllabic word. Specifically, the underlying register of σ1 is preserved; the contour contrast is also preserved on σ1, via conditioning the surface forms which are not identical to the underlying contours. The tone of σ2 also plays a role: after an initial falling tone, a falling tone preserves its contour, while a rising tone surfaces with a level-ish contour. Taken together, the results confirm impressionistic observations on Shaoxing Wu (e.g., [9]) with acoustic data.

These findings confirm that in Shaoxing Wu, registers can be preserved when contours undergo changes (e.g., σ1 in Figure 2A), and contours can be maintained despite register neutralization (e.g., σ2 in Figure 3A). These patterns contradict the idea that tonal register dominates tonal contour [3], and provide experimental evidence for the independence of tonal register and contour in sandhi processes of Wu dialects. They lend support to models of tonal representation proposing independent relationship between register and contour (e.g., [4]), which draw evidence only from impressionistically observed register or contour spreading.

Our findings also offer a more nuanced picture of the left- vs. right-dominant sandhi typology in Wu dialects. We have shown that σ2 can maintain their contours despite the overall left-dominant tendency, indicating that both patterns are present in the same dialect. This supports previous research on the complex tone sandhi patterns in Northern Wu dialects (e.g., [11], [12]). In addition, the results demonstrate that the effects of register and contour can differ between σ1 and σ2. [13] proposed that in Northern Wu, σ1 tends to preserve registers, and σ2 preserves contours. Our data align with their findings, showing that σ2 does preserve contours to some extent, and σ1 preserves both register and contour.

The findings on syllable duration provide further insights into the relationship between sandhi dominance and metrical prominence. We found that σ2 is longer than σ1, be it in words starting with falling or rising tones, although the syllable-position-based durational contrast is greater when σ1 bears a falling tone (Figure 4). As confirmed by the results on *f0*, words starting with falling (Figure 3A-B) and rising tones (Figure 3C-D) undergo right- and left-dominant sandhi respectively. If this distinction is due to different metrical structures [14], we would expect to see different durational patterns between the syllables in each type of metrical structures; σ2 should be longer than σ1 in a right-dominant structure, and vice versa in a left-dominant structure. However, no such durational differences were observed, raising questions about cues of metrical prominence motivated by tone sandhi dominance patterns.

To conclude, with controlled experimental data, this study provides evidence that tonal register and contour behave independently in sandhi processes, and offers a more nuanced understanding of the complex tone sandhi patterns in Northern Wu dialects. In addition, our findings suggest the lack of a straightforward relationship between tone-sandhi based metrical prominence and syllable durational contrasts.

5. ACKNOWLEDGEMENTS

This research is supported by CSC scholarship to Xinyi Wen and NWO VICI grant (No. vi.c.181.040) to Yiya Chen.

6. REFERENCES

[1] K. Pike, *Tone languages*. Ann Arbor: University of Michigan Press, 1948.
[2] M. Yip, *Tone*. Cambridge University Press, 2002.
[3] M. Yip, "Contour Tones," *Phonology*, vol. 6, no. 1, pp. 149–174, 1989.
[4] Z. Bao, *On the Nature of Tone*. Oxford University Press, 1999.
[5] Z. Bao, "Tonal Contour and Register Harmony in Chaozhou," *Linguist. Inq.*, vol. 30, no. 3, pp. 485–493, 1999.
[6] M. Y. Chen, *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge University Press, 2000.
[7] F. Wang, "Shaoxinghua Jiyin (Sounds of Shaoxing)," in *Yuyanxue Luncong (Issues of Linguistics)*, vol. 3, Shanghai: Shanghai Educational Publishing House, 1959.
[8] J. Zhang, *The phonology of Shaoxing Chinese*. Utrecht: LOT, 2006.
[9] F. Wang, *Shaoxing Fangyan Yanjiu (Studies on Shaoxing Dialect)*. Beijing: Language and Culture Press, 2015.
[10] J. Zhang, "A directional asymmetry in Chinese tone sandhi systems," *J. East Asian Linguist.*, vol. 16, no. 4, pp. 259–302, Nov. 2007, doi: 10.1007/s10831-007-9016-2.
[11] M. Y. Chen and H. Zhang, "Lexical and postlexical tone sandhi in Chongming," in *Studies in Chinese Phonology*, J. Wang and N. Smith, Eds. Berlin, Boston: De Gruyter, 1997. doi: 10.1515/9783110822014.13.
[12] M. K. M. Chan, "Contour-tone spreading and tone sandhi in Danyang Chinese," *Phonology*, vol. 8, no. 2, pp. 237–259, Aug. 1991, doi: 10.1017/S0952675700001391.
[13] J. Kuang, J. Tian, and Y. Zhou, "The common word prosody in Northern Wu," in *Tonal Aspects of Languages 2018*, Jun. 2018, pp. 7–11. doi: 10.21437/TAL.2018-2.
[14] S. Duanmu, "Metrical and Tonal Phonology of Compounds in Two Chinese Dialects," *Language*, vol. 71, no. 2, p. 225, Jun. 1995, doi: 10.2307/416163.
[15] C. Guo and F. Chen, "Phonetic Realizations of Metrical Structure in Tone Languages: Evidence From Chinese Dialects," *Front. Psychol.*, vol. 13, pp. 1–17, Jul. 2022.
[16] W. Pan, L. Li, and X. Han, "Field Phon (Version F2.0.2)." 2015.
[17] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. New York: Chapman and Hall/CRC, 2017. doi: 10.1201/9781315370279.
[18] R Core Team, "R: A language and environment for statistical computing." 2021.
[19] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *J. Phon.*, vol. 70, pp. 86–116, Sep. 2018, doi: 10.1016/j.wocn.2018.03.002.

PRODUCTION OF CANTONESE TONES BY MANDARIN-SPEAKING IMMIGRANTS: ACOUSTIC AND PERCEPTUAL MEASUREMENTS

Yike Yang[1], Dong Han[1], Sze Man Wong[1], Chak Sum Chan[1], Chun Yeung Leung[1], Xiaocong Chen[2]

[1]Hong Kong Shue Yan University, [2]The Hong Kong Polytechnic University
yyang@hksyu.edu, dhan@hksyu.edu, 201042@hksyu.edu.hk,
221043@hksyu.edu.hk, 221039@hksyu.edu.hk, xiaocong.chen@polyu.edu.hk

ABSTRACT

Although Cantonese has a complex tonal system, there is a lack of research on adult learners' acquisition of second language (L2) Cantonese tones, particularly studies of learners with a tone language background. The present study attempted to explore whether Mandarin-speaking immigrants could acquire the Cantonese tonal system and whether there would be category assimilation or dissimilation of lexical tones in their L2 Cantonese. A tone production experiment involving 41 participants was conducted, and both acoustic and perceptual measurements were employed to analyse the speech samples. The immigrants showed a smaller tonal space in comparison with the native speakers; they also had very low accuracy rates in their tone production, indicating that they had not fully acquired the Cantonese tonal system. Explanations for the confusion patterns are provided, and the effects of the first language on L2 tone acquisition are discussed.

Keywords: speech production, speech prosody, lexical tone, second language, Cantonese

1. INTRODUCTION

1.1. Lexical tones in Cantonese and Mandarin

In Cantonese and Mandarin, lexical tones contribute to differences in word meanings. As shown in Table 1, Cantonese has six lexical tones, with three level tones (T1, T3 and T6), two rising tones (T2 and T5) and one falling tone (T4) [1]. Unlike Cantonese, the four lexical tones in Mandarin are distinguished by pitch contours [2]. The differences between the two tonal systems may bring about interactions of bilingual speakers' Cantonese and Mandarin.

Table 1: The tonal systems of Cantonese and Mandarin.

	Cantonese tonal system		Mandarin tonal system	
Name	Category	Letter	Category	Letter
Tone 1 (T1)	High Level	55	High Level	55
Tone 2 (T2)	High Rising	25	Rising	35
Tone 3 (T3)	Mid Level	33	Dip-rising	214
Tone 4 (T4)	Mid-low Falling	21	Falling	51
Tone 5 (T5)	Mid-low Rising	23	N/A	N/A
Tone 6 (T6)	Mid-low Level	22	N/A	N/A

According to recent studies, several Cantonese tone pairs (such as T3 and T6) are merging in either the production or the perception [3], which is probably part of a sound change in contemporary Cantonese [4]. If immigrants in Hong Kong are exposed to different types of Cantonese input (some merged and some not), it is interesting to explore whether they can maintain the contrasts amongst the six tones in their second language (L2) Cantonese.

1.2. Modelling L2 speech production

To account for the differences in the learnability of L2 speech, the Speech Learning Model (SLM) [5] and its revised version, the Revised Speech Learning Model (SLM-r) [6], have proposed that the processes and mechanisms that guide first language (L1) speech acquisition remain intact and accessible for L2 speech learning across the lifespan. According to the SLM and the SLM-r, there is a common phonetic space in a bilingual speaker's mind that stores the phonetic categories of both the L1 and the L2, and in which the L1 and L2 categories exert mutual influence.

One hypothesis in the SLM is the category assimilation hypothesis (CAH), which posits that an L2 sound that is perceived as being similar to an L1 sound does not form a new category in the common space and is understood as a variant of the L1 sound at an allophonic level. In this case, only one single phonetic category is used to process the two linked diaphones, and this mapping will eventually give rise to a new merged category, a phenomenon that has been documented in several studies [7]–[9]. Another claim in the SLM is the category dissimilation hypothesis (CDH). A new category will be established if an L2 sound is absent in the L1 system, which will make the combined phonetic space more crowded, resulting in the phonemes tending to disperse to ensure that the phonetic contrast is maintained. Recent support for the CDH comes from [10], which showed that Spanish-Catalan bilinguals had developed two categories to accommodate the mid-back vowels in the two languages.

1.3. The present study

Studies of Cantonese tone acquisition have mainly focused on younger populations, such as monolingual children [11], bilingual children [12] and bilingual

youths [13], but adult L2 learners' acquisition of Cantonese tones has not been investigated. Adults are fundamentally different from younger populations because adult learners have fully acquired their L1 when they begin to acquire their L2. In addition, the acquisition of L2 tones by learners with a tonal language background has received little attention. Thus, the present study attempts to examine whether adult learners with a tonal language background (Mandarin-speaking immigrants) can acquire the complex Cantonese tonal system, and whether there is category assimilation or dissimilation of lexical tones in their L2 Cantonese due to the influences of their L1 Mandarin, with the aim of determining whether the hypotheses of the SLM and the SLM-r hold for lexical tones. The purpose of this study is to answer the following research questions:

1) Are there acoustic and perceptual differences in the production of Cantonese tones by native speakers and immigrants?

2) Is there any category assimilation or dissimilation in the Cantonese tone production of the immigrants?

2. METHODS

2.1. Participants

Two groups of participants attended a tone production experiment in a soundproofed booth at a local university. The participants were 32 Mandarin-speaking immigrants who were born and raised in Northern China, and who had spoken Mandarin as their only Chinese dialect prior to their arrival in Hong Kong after puberty. To assess their language profile, the immigrants completed a language background questionnaire [14] prior to the recording session. The results of the questionnaire indicated that the immigrants were fluent speakers of Cantonese, although they were more dominant in Mandarin. Nine native Cantonese speakers were included as the control group. The Cantonese speakers was born and raised in Hong Kong, where Cantonese is the dominant language. None of the participants reported any history of speech, language or hearing disorders.

2.2. Materials and procedures

The target stimuli were 12 monosyllabic words contrasting the six lexical tones in two base syllables (/si/ and /fu/), as listed in Table 2. Another ten monosyllabic words with different consonant and vowel combinations were included as the filler trials. Both the target and filler syllables are of high frequency in Cantonese. The syllables were presented in two contexts; that is, in isolation and in a carrier phrase '我讀__呢個字 *ngo5 duk 6_ nei1 go3 zi6* (I read the character _)'. Each stimulus appeared twice. In total, there were 1,968 target trials (2 target syllables * 6 tones * 2 context * 2 times * 41 speakers).

Table 2: The target syllables.

Tone	Syllable
T1	詩 /si1/ 'poem', 夫 /fu1/ 'husband'
T2	史 /si2/ 'history', 苦 /fu2/ 'bitter'
T3	試 /si3/ 'to try', 富 /fu3/ 'rich'
T4	時 /si4/ 'time', 扶 /fu4/ 'to hold'
T5	市 /si5/ 'market', 婦 /fu5/ 'married woman'
T6	事 /si6/ 'thing', 父 /fu6/ 'father'

The participants were briefed about the task requirements and were allowed to read the stimulus list prior to the experiment. During the experiment, the stimuli were presented randomly in E-Prime 2.0 [15] on a computer screen, and the production sessions were recorded in Audacity [16] at a sampling rate of 44,100 Hz.

2.3. Data processing and analysis

The vowel portions of the target syllables, which bear lexical tones, were segmented manually in Praat [17]. Following the segmentation, 20 time-normalised F0 values were extracted for each vowel using a Praat script. To eliminate individual differences in the F0 range and make direct cross-group/speaker comparisons possible, the F0 values, which were originally measured in Hz, were converted to a five-point scale from 1 to 5 with Equation 1:

$$c_m = ((5-1) * \frac{o_m - min_n}{max_n - min_n}) + 1 \qquad (1)$$

where c_m and o_m represent the converted and original F0 values of the mth point, respectively, and max_n and min_n stand for the maximal and minimal values of all the original F0 values of the nth speaker.

The F0 values were analysed using generalised additive mixed models (GAMMs) via the 'mgcv' package [18] in R [19], [20], in which *the converted F0 values* was the dependent variable, and *time*, *group*, *tone*, *context* and *syllable* were included as the predictors. The GAMM was adopted because it does not assume a linear relationship between the dependent variable and the predictors, which makes it appropriate for modelling time-dependent datasets such as the data in this study. The figures were plotted using the 'ggplot2' package in R [21].

Perceptual evaluations were also included in this study. Two native speakers of Hong Kong Cantonese who did not exhibit tone merging were invited to listen to all the target trials. The trials were presented randomly to the listeners; thus, the listeners did not know whether the speaker was a native speaker or an

immigrant before they began the judgement task. The listeners were instructed to judge which character (corresponding to a tone) each trial represented and were allowed to listen to the trials several times if they deemed it necessary. The listeners completed the task independently, and the agreement amongst them was 77.69%. We adopted the more stringent criterion from [22]: Only the trials that both listeners considered to be the intended tones were counted as having been pronounced correctly.

3. RESULTS

3.1. F0 contours of the tone production

Figure 1 presents the F0 contours of the six tones produced by Cantonese speakers and immigrants. According to Figure 1(a), the Cantonese speakers clearly distinguished the six tones in their production. For the immigrants, as shown in Figure 1(b), T1 was extremely high and far from the remaining five tones, and the five tones were crowded in a narrower space in comparison to the native speakers. In addition, the native speakers showed a greater F0 range than did the immigrants.

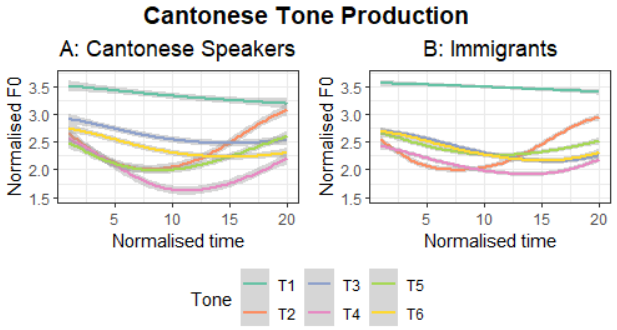


Figure 1: Production of Cantonese tones by Cantonese speakers and immigrants.

3.2. Acoustic analyses

A GAMM was first fitted with the entire dataset. There were main effects of *time*, *group*, *tone*, *context* and *syllable* (*p*s < .05). The predictors *context* (in isolation and in sentences) and *syllable* (/si/ and /fu/) are beyond the scope of this study. In the following of this section, the predictors *group* and *tone* will be investigated in more detail, with *time* always being included as a one-dimensional smooth.

A GAMM was constructed for native speakers' F0 data, with *tone* as the predictor and T1 as the reference. The results suggested that T1 was statistically different from the remaining tones, and it always had higher F0 values (*p*s < .001). More GAMMs were fitted to test whether the proposed merging-in-progress tone pairs were distinguishable. The models indicated that none of the tone pairs (T2 and T5, T3 and T6, T4 and T6, and T4 and T5) had merged in our data (*p*s < .001).

The same procedures were applied to the immigrants' data. The immigrants also showed very high F0 values for T1 compared to the remaining five tones (*p*s < .001). Although the immigrants exhibited a narrower tonal space for the remaining five tones, many of the tone pairs were separable by GAMMs: T2 and T5 ($p < .001$), T3 and T4 ($p < .001$), T3 and T5 ($p = .027$), and T4 and T6 ($p < .001$). However, T3 and T6 produced by the immigrants were merged as one category ($p = .467$).

Next, separate GAMMs were fitted to compare each tone produced by the native speakers and the immigrants, with data from the native speakers as the reference. Table 3 lists the statistics from the GAMMs, which suggests divergence between the two speaker groups for all the tones except for Tone 6, which was also marginally significant. Specifically, while the immigrants' T2, T3 and T6 were not as high as the same tonal categories that were produced by the native speakers, their T1, T4 and T5 were higher than were those of the native speakers. Given that the F0 values had been normalised to the same scale, it can be inferred that the immigrants failed to pronounce the Cantonese tones in a native-like way.

Table 3: Statistics from the GAMMs (with native speakers as the reference).

Tone	Estimate ± SE	T value	P value
T1	0.152 ± 0.028	5.429	< .001
T2	-0.053 ± 0.022	-2.402	< .001
T3	-0.252 ± 0.028	-11.09	< .001
T4	0.144 ± 0.023	6.644	< .001
T5	0.174 ± 0.023	7.696	< .001
T6	-0.042 ± 0.023	-1.835	0.067

3.3. Perceptual evaluations

The overall accuracy rates of the two groups were calculated first. As introduced in Section 2.3, only the tokens that were judged as being the intended ones by both listeners were counted as accurate. The tones produced by the native Cantonese speakers were generally perceived as being correct, with the accuracy rate of each tone ranging from 66.67% to 91.67% (average rate: 79.40%). However, the immigrants made many tone production errors, with an average accuracy rate of 44.99% (ranging from 20.83% to 68.75%); in other words, the two native listeners considered more than half of the trials to have been pronounced incorrectly.

Figure 2 displays the production accuracy rates for each tone. As can be seen in Figure 2(a), the native speakers' T2 and T4 were always perceived as the intended tones (100% accuracy rate), and T1 was also

almost always perfect (97.22% accuracy rate). The accuracy rate for T3 was 72.22%, and the majority of the incorrectly identified T3 syllables were perceived as being T6. For T5 and T6, the accuracy rates were slightly above 50%. The T5 syllables tended to be perceived as T2, and the T6 syllables were likely to be perceived as T3.

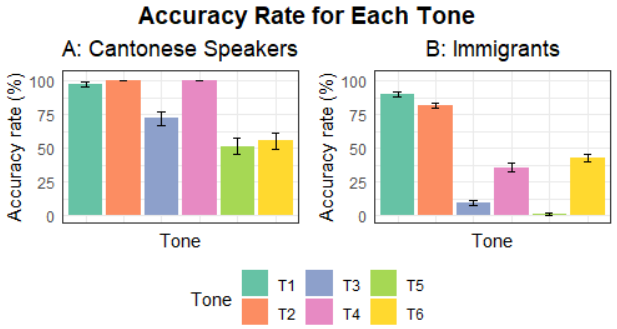


Figure 2: Accuracy rate of Cantonese tone production by Cantonese speakers and immigrants.

The accuracy rates for each tone produced by the immigrants are presented in Figure 2(b), according to which T1 and T2 had the highest accuracy rates, although they were not as accurate as were the native speakers' T1 and T2. The accuracy rates for the remaining four tones were very low, and the tones that the four tones were most likely to be perceived as are presented in Table 4. Both T3 and T4 were very likely to be perceived as T6, and T6 tended to be perceived as T4 or T3, thus suggesting that the immigrants had difficulty in distinguishing the T3-T6 and T4-T6 pairs. Another observation was that T5 was perceived more frequently as T6 than it was as T2.

Table 4: Listeners' judgements of immigrants' tones.

Tone	Perceived as
T3	T6 (68.18%), T4 (24.12%)
T4	T6 (54.48%), T2 (29.60%)
T5	T6 (45.40%), T2 (31.43%)
T6	T4 (48.94%), T3 (31.34%)

4. DISCUSSION AND CONCLUSION

This study investigated Mandarin-speaking immigrants' production of Cantonese tones using acoustic analyses and perceptual evaluations. The acoustic results suggested that the native speakers had a larger tonal space than did the immigrants, which is in line with previous findings that Cantonese speakers exhibit larger F0 range than Mandarin speakers [23]. Consequently, the native speakers clearly distinguished the six tones, and the immigrants' T2 to T6 were extremely crowded and even revealed the phenomenon of tone merging [13]. The perceptual evaluations confirmed the acoustic data. Although the native listeners generally considered the immigrants' T1 and T2 to have been pronounced correctly, the overall accuracy rates for the immigrants' tone production were very low, which suggested that they had not acquired the Cantonese tonal system completely.

A closer examination of the data showed common confusion patterns in the immigrants' tone production: T3-T6, T4-T6 and T5-T6. The first two pairs have been reported as being tones that are undergoing the process of merging [4], which could partially account for the immigrants' pattern from their linguistic input. However, more explanations are needed as native speakers did not always exhibit T3-T6 or T4-T6 confusion patterns. It is possible that T3, T4 and T6 are challenging for Mandarin speakers because these tonal categories are missing in their L1 system, and the acoustic and perceptual differences between these tones and their L1 tones are huge. It is therefore difficult for Mandarin speakers to establish the new categories when they are learning Cantonese, given the similarities between T3 and T6 (in pitch direction) and between T4 and T6 (in pitch height) within the tonal system of Cantonese.

The confusion between T5 and T6 was unexpected because the two tones differ in pitch direction, and native speakers do not confuse these two tones. A possible source of the confusion might be the learners' difficulty in correctly establishing the tonal categories because T5 and T6 share similar pitch height (23 vs 22). This indicates that the acoustic or perceptual similarity within a tonal system might surpass linguistic input and cause confusion in L2 category formation. Further studies should be conducted to compare the weight of linguistic input and phonetic similarity in L2 speech acquisition.

The data provide support for CAH and CDH from the SLM and the SLM-r. The identification accuracy rate for T5 was below 2%, suggesting that the T5 category was missing in the immigrants' Cantonese. As there is only one rising tone in Mandarin, it is likely that the learners had merged T5 with T2 to form one rising category, which is an example of category assimilation. Note that the immigrants also merged T5 with T6, which may be explained by the more similar pitch height between T5 and T6. Moreover, the immigrants attempted to raise their T1 to make it farther away from the remaining tones, and their T1 category was apparently different from that of the native speakers based on the acoustic analysis, which lends support to the CDH, although whether the immigrants' Cantonese T1 and Mandarin T1 are separate still needs to be confirmed when their L1 and L2 T1 categories are compared directly. Such investigation will further our understanding of L1 and L2 speech interactions (e.g. [24]).

5. ACKNOWLEDGEMENTS

The authors would like to thank all the informants for their participation and the ICPhS committee and reviewers for their constructive suggestions. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No.: UGC/FDS15/H15/22) and an ASA International Student Grant from the Acoustical Society of America to the first author.

6. REFERENCES

[1] R. S. Bauer and P. K. Benedict, *Modern Cantonese Phonology*. Berlin: Walter de Gruyter, 1997.

[2] S. Duanmu, *The Phonology of Standard Chinese*. Oxford University Press, 2007.

[3] R. S. Y. Fung and C. K. C. Lee, "Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. EL424–EL430, Nov. 2019, doi: 10.1121/1.5133661.

[4] P. P. K. Mok, D. Zuo, and P. W. Y. Wong, "Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese," *Lang. Var. Change*, vol. 25, no. 3, pp. 341–370, Oct. 2013, doi: 10.1017/S0954394513000161.

[5] J. E. Flege, "Interactions between the native and second-language phonetic systems," in *An integrated view of language development: Papers in honor of Henning Wode*, T. Piske, A. Rohde, and P. Burmeister, Eds. Trier: Wissenschaftlicher Verlag, 2002, pp. 217–244.

[6] J. E. Flege and O.-S. Bohn, "The Revised Speech Learning Model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*, R. Wayland, Ed. Cambridge: Cambridge University Press, 2021, pp. 3–83.

[7] R. C. Major, "Losing English as a First Language," *Mod. Lang. J.*, vol. 76, no. 2, pp. 190–208, 1992, doi: 10.2307/329772.

[8] C. B. Chang, "Rapid and multifaceted effects of second-language learning on first-language speech production," *J. Phon.*, vol. 40, no. 2, pp. 249–268, 2012, doi: 10.1016/j.wocn.2011.10.007.

[9] M. L. Sancier and C. a. Fowler, "Gestural drift in a bilingual speaker of Brazilian Portuguese and English," *J. Phon.*, vol. 25, no. 4, pp. 421–436, 1997, doi: 10.1006/jpho.1997.0051.

[10] M. Simonet, "Production of a catalan-specific vowel contrast by early Spanish-Catalan bilinguals," *Phonetica*, vol. 68, no. 1–2, pp. 88–110, 2011, doi: 10.1159/000328847.

[11] P. P. K. Mok, H. S. H. Fung, and V. G. Li, "Assessing the Link Between Perception and Production in Cantonese Tone Acquisition," *J. Speech, Lang. Hear. Res.*, vol. 62, no. 5, pp. 1243–1257, May 2019, doi: 10.1044/2018_JSLHR-S-17-0430.

[12] Y. Yao *et al.*, "Cantonese tone production in pre-school Urdu–Cantonese bilingual minority children," *Int. J. Biling.*, vol. 24, no. 4, pp. 767–782, Aug. 2020, doi: 10.1177/1367006919884659.

[13] A. C. L. Yu, C. W. T. Lee, C. Lan, and P. P. K. Mok, "A New System of Cantonese Tones? Tone Perception and Production in Hong Kong South Asian Cantonese," *Lang. Speech*, vol. 65, no. 3, pp. 625–649, Sep. 2022, doi: 10.1177/00238309211046030.

[14] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism," *COERLL, University of Texas at Austin*, 2012. https://sites.la.utexas.edu/bilingual/.

[15] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime User's Guide*. Pittsburgh: Psychological Software Tools Inc, 2012.

[16] Audacity Team, "Audacity(R): Free Audio Editor and Recorder." 2019, [Online]. Available: https://audacityteam.org/.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2015, [Online]. Available: http://www.praat.org/.

[18] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2017.

[19] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2018, [Online]. Available: https://www.r-project.org.

[20] RStudio Team, "RStudio: Integrated Development for R." RStudio, Inc., Boston, MA, 2016, [Online]. Available: http://www.rstudio.com/.

[21] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer, 2016.

[22] P. P. K. Mok and A. Lee, "The acquisition of lexical tones by Cantonese–English bilingual children," *J. Child Lang.*, vol. 45, no. 6, pp. 1357–1376, Nov. 2018, doi: 10.1017/S0305000918000260.

[23] Y. Yang, S. Chen, and X. Chen, "F0 Patterns in Mandarin Statements of Mandarin and Cantonese Speakers," in *Proc. Interspeech 2020*, Oct. 2020, pp. 4163–4167, doi: 10.21437/Interspeech.2020-2549.

[24] Y. Yang, "First Language Attrition and Second Language Attainment of Mandarin-speaking Immigrants in Hong Kong: Evidence from Prosodic Focus," Ph.D dissertation, The Hong Kong Polytechnic Univ., Hong Kong, 2022.

LEXICAL TONES ARE TIMED TO ARTICULATORY GESTURES

Francesco Burroni

Institute of Phonetics and Speech Processing, Spoken Language Processing Group, LMU, Munich
francesco.burroni@phonetik.uni-muenchen.de

ABSTRACT

We investigate whether speakers more stably time f0 movements, associated with lexical tones, to articulatory gestures or to the acoustic outputs of articulatory gestures. Acoustic and articulatory data collected from eight Bangkok Thai speakers show that f0 movements are more stably timed to the articulatory onsets of vowels. F-tests show that the lag between f0 movements and vocalic gestures onsets has the lowest variance compared to the lag of f0 movements onsets and any other acoustic/articulatory landmark. Additionally, Gaussian Process Regression models trained with articulatory features outperform models trained with acoustic features in providing a generative model of f0 movements onset timing. Taken together these two lines of evidence suggest that speakers time f0 movements onsets to articulatory gestures rather than their acoustic consequences, like the acoustic onsets of syllables/vowels. Issues regarding surface f0 timing, articulation, and perception are also discussed to shed light on the uncovered patterns.

Keywords: lexical tone, timing, articulation, anchoring, articulatory phonology

1. INTRODUCTION

An open question in the phonetic literature is how speakers control the timing of f0 movements, associated with lexical tones and pitch accents, relative to the movements of supralaryngeal articulatory gestures and their acoustic consequences.

Research on tone languages holds that f0 movements are synchronized with syllables e.g., [1]–[4]. Yet, it is not clear whether "syllable" should be interpreted in terms of articulatory gestures affiliated with an articulatory syllable or their acoustic consequences associated with an acoustic syllable. Research on the timing of f0 peaks and valleys of pitch accents has uncovered regularities in their timing relative to anchoring sites in the acoustic segmental string e.g., [5]–[9]. These findings were interpreted as evidence that speakers time relevant f0 movements to acoustic landmarks, such as consonantal closures' or stressed vowels' acoustic onsets. This hypothesis was termed Segmental Anchoring Hypothesis (SAH). Further work on testing the *SAH*, however, showed that the timing of f0 movements to acoustic landmarks can be influenced by linguistic and paralinguistic factors, such as syllable structure, speech rate, and diatopic variation e.g., [8], [10]–[12]. These findings led several researchers to hypothesize that f0 movements may be anchored to articulatory rather than acoustic events e.g., [13]–[15]. Non-linearities in the articulatory to acoustic mapping could then be responsible for variability in f0 movements timing to acoustic landmarks. Investigations testing whether f0 movements are more stably aligned to acoustic or articulatory events, however, failed to provide conclusive evidence e.g., [13], [16].

Meanwhile, researchers investigating both acoustic and articulatory signals, especially from the perspective of gestural phonology [17], assumed that f0 movements associated with lexical tones and pitch accents can be treated on par with articulatory gestures and that these are timed to the supralaryngeal articulatory gestures associated with the production of consonants and vowels e.g., [18]–[24]. Crucially, the timing of f0 movements to supralaryngeal articulatory gestures was posited without assessing whether acoustic or articulatory landmarks provided more stable "anchoring" sites.

The main research question we investigate is the following: do speakers time f0 movements associated with lexical tones to the articulatory gestures underlying the production of consonants and vowels or to the acoustic consequences of said gestures, such as the acoustic onset of consonants/vowels/syllables? We can entertain two hypotheses. ***H1*****:** we expect speakers to time relevant f0 movements to articulatory gestures. ***H2***: we expect speakers to time relevant f0 movements to the acoustic consequences of articulatory gestures, e.g., the (acoustic) onsets of syllables or vowels. The predictions associated with ***H1*** are that, if speakers time f0 movements to articulatory gestures, the variance in lags between f0 movements onsets, associated with lexical tones, and articulatory gestures onsets, associated with consonants and vowels, should be the lowest; indicating that the two are produced with stable relative initiations. Additionally, knowledge of the time of initiation of the articulatory gestures should help predict the timing of the tonal f0 movements initiations. Conversely, the predictions associated with ***H2*** are that, if speakers time f0 movements to the acoustic consequences of articulatory gestures,

the variance in lags between f0 movements onsets, associated with lexical tones, and acoustic onsets of consonants/syllables and vowels should be the lowest. Additionally, knowledge of acoustic onsets of consonants/syllables and vowels should help predict the timing of the tonal f0 movement initiation. A production study with acoustic and articulatory data collected from eight Bangkok Thai (BKKT) speakers was conducted to test ***H1***/***H2*** and their predictions.

2. METHODOLOGY

Eight L1 speakers of Thai participated in the experiment. They did not disclose any speech or hearing impairment. All speakers were screened for nativeness in BKKT by a native speaker trained in phonetics. Participants produced BKKT Falling and Rising tones (F/R), w2, followed by all five tones of the language (Mid, Low, Falling, High, Rising), w3. The disyllabic targets were embedded in a carrier sentence with a fixed number of words and syllables; all carrier words are Mid-toned Table 1.

w1	w2	w3	w4	w5
dū:	mî:	mā:	bōn	dā:w
	mǐ:	mà:		dīn
		mâ:		lāŋ
		má:		
		mǎ:		

Table 1: Experimental materials

F/R in w2 were chosen because their f0 trajectory resembles an articulatory trajectory and can easily be landmarked. All five BKKT tones were used in w3 to introduce variation due to coarticulatory effects. The transition between [u] and [i] in the carrier was chosen to maximize tongue movement in the longitudinal direction and facilitate locating vowels onsets of the target word [mi:]. Participants were asked to produce the stimuli at one of five rates: very slow, slow, normal, fast, or very fast to introduce variation. W5 was varied at every trial to function as a distractor. In total participants produced 2 (F/R) × 5 (M/L/F/H/R) × 3 (w5) = 30 unique stimuli × 5 (rates) × 3 (repetitions) = 450, distributed in 9 blocks. Not all participants completed the entire experiment due to time constraints. Not all tokens for all participants could be analyzed due to errors, equipment malfunctions in articulatory data tracking, and a coding oversight. After screening all the data, a total of 3157 tokens were retained for analysis.

For the experiment, participants sat in a front of a computer monitor. A custom MATLAB GUI was used to present the stimuli and collect synchronized acoustic and articulatory data. Audio was collected with a sampling frequency of 44.1 kHz and 16 bits per sample using a shotgun microphone positioned around 1.25 m away from the participant. Articulatory data were collected at a sampling frequency of 400 Hz using an NDI Wave electromagnetic articulometer (EMA). Articulatory data sensors were adhered midsagittally on the lower and upper lip (LL, UL) vermilion border; on the lower right incisor to capture jaw movement (JAW). Two sensors were placed on the tongue, one posterior to the tongue apex, about 1 cm, to measure tongue tip (TT) movement, and one posterior to the TT sensor of approximately 6–7 cm to measure tongue body (TB) movement. Reference sensors were positioned on the nasion and left and right mastoid processes.

Speaker-specific monophone Hidden Markov Models were trained in Kaldi [25] and used to perform forced alignment by speaker, followed by manual checking and correction. F0 was extracted using the Sum of Residual Harmonic algorithm [26] in MATLAB with a 52 ms window and a 10 ms overlap. For men, we used a range [60 200] Hz and, for women, a [100 400] Hz range. The raw f0 trajectories were cleaned of f0 "jumps" greater than 20 Hz, interpolated with spline interpolation, and smoothed using a moving median followed by a moving average filter.

Articulatory trajectories' missing values were obtained with linear interpolation. The position of articulatory sensors was corrected for head movement. The trajectories were smoothed using a 3rd order low pass Butterworth filter with a 10 Hz cutoff. In this paper, we focus on the trajectories involved in the production of [m] and [i:] in w2. The [m] gesture closure and release were identified using a lip aperture (LA) time series. LA is defined as the Euclidean distance between the vertical and horizontal components of the LL and UL movements. The formation of [i:] was studied from the horizontal component of the tongue body (TB) movement. We focus on the horizontal component because the carrier sentence has a transition from [u:] to [i:] that is maximally differentiated on the horizontal plane. Tones were landmarked based on the f0 trajectory. We first located an inflection/midpoint, then proceeded to locate velocity extrema that precede and follow the inflection/midpoint. We then located positional extrema that precede and follow the first and second velocity extrema, respectively. Within a region spanning the first positional extremum and the first velocity extremum, we located the onset of the movement, defined as the first time point at which the movement velocity surpasses 20% of the maximum velocity. We then located the gestural target as the last time point after the first velocity extremum and before the inflection point where movement velocity falls below a 20% threshold of maximum velocity. The operation was repeated between the inflection

point and second velocity extremum to locate the release and between the second velocity extremum and second positional extremum to locate the movement offset. An example of landmarked articulatory and (acoustic) f0 trajectories is presented in Figure 1.

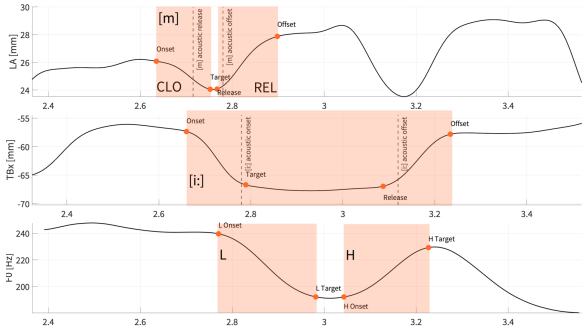


Figure 1: Example of LA, TBx, and F0 landmarking

All acoustic landmarks were obtained from forced alignment boundaries. All landmarking was manually checked and corrected. From the landmarks, second-order landmarks were derived. The C-T center was calculated as the midpoint between the closure (CLO) and tonal (T) onsets. Two versions of the C center (1 and 2) were also calculated. C center 1 is the midpoint of CLO onset and offset, while C center 2 is the midpoint of closure and release. Pairwise lags among the onset of different acoustic and articulatory gestures were calculated as illustrated in Figure 2.

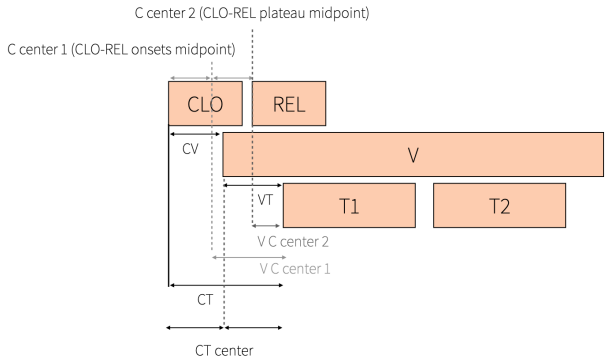


Figure 2: Lags and second-order landmarks

We examined the means and variances of lags between the tonal onset and all other articulatory and acoustic landmarks. F-tests for variance were used to assess differences in variance, indicating different stability of f0 movements onset to the other landmark forming the lag. Gaussian Process Regression (GPR) models were trained to predict the onset of f0 movements based on three sets of features:

1. Control Features: (1-5): (1) Subject, (2) Block, (3) Utterance Duration, (4) Tone, (5) Following Tone
2. Acoustic Features: (6-9) (6) C onset, (7) C Offset, (8) V Onset, (9) V Offset
3. Articulatory Features: (10-14): (10) LA movement onset, (11) LA mov.t target, (12) TB (TB) mov. onset, (13) TB mov. target, (14) C center 1

The control GPR model was used to obtain a root mean squared error (RMSE) baseline based on speakers' nonstationary behavior [27]. Models trained with acoustic and/or articulatory features were used to compare how acoustic and articulatory information could improve the prediction of f0 movements onsets. The accuracy reported in the paper is from tenfold cross-validation.

3. RESULTS

3.1 Variability analysis

The main finding is that the landmark with the lowest variability compared to the f0 movement onset is the articulatory onset of vowels, with a lag standard deviation estimated at 55.8 ms. The second lowest variance is with the acoustic onset of the consonant/syllable, with a standard deviation estimated at 62 ms. An F-test for equal variance ($F = .81$, df_1, $df_2=3156$, $p < .0001$) reveals that the lag between the vowel articulatory onset and the tonal f0 movement has lower variance than the lag with the consonant/syllable acoustic onset.

Variance patterns do not align with the mean patterns. The onsets of tonal f0 movements are more stably timed to the onsets of the vocalic gestures, however, the vocalic gesture initiation *precedes* the f0 movement initiation by 48.4 ms on average. On the other hand, the acoustic onset of the consonant syllable precedes the f0 movement by 2.4 ms on average. To sum up, the onset of f0 movements associated with lexical tones is more stably timed to the articulatory onset of vowels, but it is closest in time to the consonant/syllable acoustic onset. These results are summarized in Figure 3.

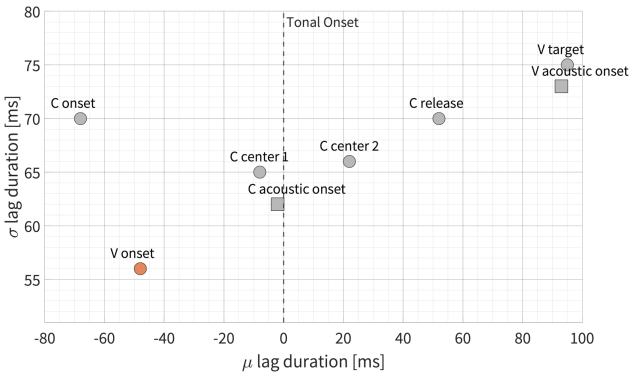


Figure 3: Plot of lag std.dev. (σ) *vs*. lag mean (μ) duration

3.2 Gaussian Process Regression analysis

The main findings are that adding acoustic and/or articulatory information reduces the RMSE in predicting the f0 movement onset by around 5.5

times. Second, if we compare the addition of acoustic *vs*. articulatory information to the control feature set, we observe that articulatory information yields better predictions, in line with the variability analysis. Third, the combination of both acoustic and articulatory information results in a ~1 ms improvement performance of the GPR model. This last finding suggests that both types of information play a role in determining the timing of f0, yet, when articulatory information is available, the addition of acoustic information does not lead to a huge improvement and vice versa. The results of the Gaussian Process are summarized in Figure 4.

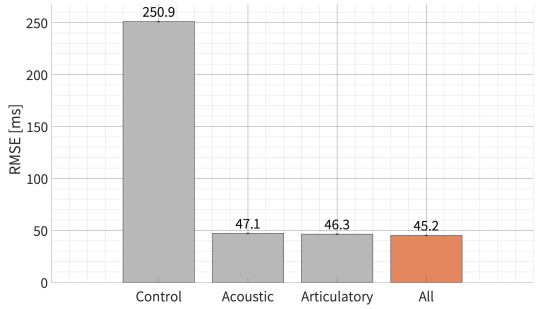


Figure 4: GPRs RMSE for different training features

4. DISCUSSION

Both the variability analysis and the GPR models suggest that f0 movements seem to be timed to articulatory events rather than their acoustic consequences. This is in line with the lower variability observed between f0 movements onsets and the articulatory onset of vocalic gestures rather than any other acoustic/articulatory landmark. The GPR models show that predictions of f0 movements onsets achieve better results based on the articulatory, rather than acoustic, landmarks with a lower RMSE of around ~1 ms. Additionally, the combination of both acoustic and articulatory information only leads to a reduction in RMSE of around ~1 ms, despite a larger number of parameters. We can, thus, conclude that articulatory timing of f0 movements onsets is more in line with the data.

F0 movements are, however, closer in time, with a mean lag of ~2 ms, to the acoustic onset of syllables. The fact that, on average, f0 movements onsets are closest to the acoustic onset of syllables could be the reason why previous work has considered acoustic boundaries as possible anchoring sites for f0 movements. However, the near-synchronization between f0 movements onsets and acoustic landmarks goes hand in hand with a higher variability compared to articulatory landmarks. This is an additional problem for a hypothesized timing to acoustic landmarks since lower means in speech timing are expected to correlate with *lower*, not higher, standard deviations e.g., [28]. We can hypothesize that f0 movements are timed, but not synchronized, to articulatory gestures because f0 movements onsets are acoustic, however, these acoustic changes may be delayed compared to the laryngeal articulatory adjustments necessary to produce them. For instance, consonantal articulatory onsets occur ~68 ms before their acoustic onsets in the data presented. If this reasoning is correct, the articulatory onset of the laryngeal adjustments resulting in f0 movements could be closer in time or synchronized, to the onset of the vocalic gesture.

A second more speculative hypothesis worth considering is that the muscles controlling laryngeal adjustments may have higher latencies compared to the muscles controlling oral movements of e.g., jaw, lips, and tongue. The physiological underpinning of these longer latencies may be the greater length of the recurrent laryngeal nerve, controlling laryngeal muscles, compared to the hypoglossal and trigeminal nerves, controlling the jaw, tongue, and lips. For instance, bilateral adduction of vocal folds, time-locked to a longer-latency response (R2) of recurrent laryngeal nerve stimulation, occurs around 60 ms after external stimulation via electromyography [29]. If this is correct, speakers may co-plan movements underlying f0 changes and oral articulatory gestures, but these are not synchronous due to intrinsic differences in their innervation response latencies.

A third hypothesis worth considering is tied to perceptual effects. A serendipitous consequence of the latency between vocalic gestures articulatory onsets and acoustic f0 movements onsets is that tones are produced synchronized to acoustic syllables, thus, largely, over perceptually salient rhymes where f0 is unperturbed. Tones could then be another case where "gestural" timing is constrained by perceptual recoverability [30]. Whatever the exact mechanisms, several factors may be responsible for the uncovered pattern whereby f0 movements onsets are stably timed to articulatory gestures, but, at the surface, they look near-synchronous to acoustic syllable boundaries.

A final note of caution is in place. Articulatory gestures and their acoustic consequences have a causal relationship that makes identifying the control structure of speech timing difficult: the standard deviation differences and model accuracy differences we reported are admittedly modest in size. However, they do suggest that speakers time f0 movements to the movements of articulators producing consonants and vowels, not to their acoustic consequences. More work on tonal languages beyond Thai and on pitch accents languages will help assess the generality and merits of the findings presented in this paper, as well as the possible cross-linguistic instantiations of f0 movements timing.

5. REFERENCES

[1] Y. Xu and S. Prom-on, "Degrees of freedom in prosody modeling," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Springer, 2015, pp. 19–34.

[2] Y. Xu, F. Liu, and others, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 125, 2006.

[3] Y. Xu, "Syllable is a synchronization mechanism that makes human speech possible." PsyArXiv, Mar. 2020. doi: 10.31234/osf.io/9v4hr.

[4] Y. Xu and A. Lee, "Tonal Processes Defined as Articulatory-based Contextual Tonal Variation," in *The Cambridge Handbook of Chinese Linguistics*, C.-R. Huang, Y.-H. Lin, I.-H. Chen, and Y.-Y. Hsu, Eds. Cambridge University Press, 2022, pp. 275–290. doi: 10.1017/9781108329019.016.

[5] A. Arvaniti, D. R. Ladd, and I. Mennen, "Stability of tonal alignment: the case of Greek prenuclear accents," *Journal of Phonetics*, vol. 26, no. 1, pp. 3–25, 1998.

[6] D. R. Ladd, D. Faulkner, H. Faulkner, and A. Schepman, "Constant 'segmental anchoring' of F0 movements under changes in speech rate," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1543–1554, 1999.

[7] D. R. Ladd, I. Mennen, and A. Schepman, "Phonological conditioning of peak alignment in rising pitch accents in Dutch," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2685–2696, 2000.

[8] M. Atterer and D. R. Ladd, "On the phonetics and phonology of 'segmental anchoring' of F0: evidence from German," *Journal of Phonetics*, vol. 32, no. 2, pp. 177–197, 2004.

[9] T. Ishihara, "Tonal alignment in Tokyo Japanese," The University of Edinburgh, Edinburgh, 2006.

[10] C. Petrone and R. Ladd, "Sentence-domain effects on tonal alignmenr in Italian," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, 2007, pp. 1253–1256.

[11] P. Welby and H. Løevenbruck, "Anchored down in Anchorage: Syllable structure and segmental anchoring in French," *Italian Journal of Linguistics/Rivista di linguistica*, vol. 18, pp. 74–124, 2006.

[12] P. Prieto and F. Torreira, "The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish," *Journal of Phonetics*, vol. 35, no. 4, pp. 473–500, 2007.

[13] M. D'Imperio, R. Espesser, H. Loevenbruck, C. Menezes, N. Nguyen, and P. Welby, "Are tones aligned to articulatory events? Evidence from Italian and French," *Laboratory Phonology*, vol. 9, Jan. 2007.

[14] D. R. Ladd, "Segmental anchoring of pitch movements: Autosegmental association or gestural coordination?," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 19, 2006.

[15] M. D'Imperio, N. Nguyen, and K. G. Munhall, "An Articulatory Hypothesis for the Alignment of Tonal Targets in Italian," in *Proceedings of the 15th international congress of phonetic sciences*, 2003, pp. 253–256.

[16] D. Mücke, M. Grice, J. Becker, and A. Hermes, "Sources of variation in tonal alignment: Evidence from acoustic and kinematic data," *Journal of Phonetics*, vol. 37, no. 3, pp. 321–338, 2009, doi: https://doi.org/10.1016/j.wocn.2009.03.005.

[17] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, 1992.

[18] M. Gao, "Mandarin tones: An articulatory phonology account," Ph.D. Thesis, Yale University, 2008.

[19] D. Mücke, H. Nam, A. Hermes, and L. Goldstein, "Coupling of tone and constriction gestures in pitch accents," *Consonant clusters and structural complexity*, vol. 26, p. 205, 2012.

[20] D. Mücke, A. Hermes, and S. Tilsen, "Strength and Structure: Coupling Tones with Oral Constriction Gestures.," in *INTERSPEECH*, 2019, pp. 914–918.

[21] R. P. Karlin and S. Tilsen, "The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai," in *Proceedings of Meetings on Acoustics 168ASA*, 2014, vol. 22, no. 1, p. 060006.

[22] H. Yi, "Lexical tone gestures," PhD Thesis, Cornell University, 2017.

[23] H. Niemann, D. Mücke, H. Nam, L. Goldstein, and M. Grice, "Tones as Gestures: The Case of Italian and German.," in *ICPhS*, 2011, pp. 1486–1489.

[24] R. P. Karlin, "Towards an articulatory model of tone: a cross-linguistic investigation," Ph.D. Thesis, Cornell University, 2018.

[25] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011, no. CONF.

[26] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of the Annual Conference of the International Speech Communication Association*, Florence, 2011, pp. 1973–1976.

[27] S. Tilsen, "Structured nonstationarity in articulatory timing.," in *ICPhS*, 2015.

[28] J. A. Shaw, A. I. Gafos, P. Hoole, and C. Zeroual, "Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters," *Phonology*, vol. 28, no. 3, pp. 455–490, 2011.

[29] T. Yamashita, E. A. Nash, Y. Tanaka, and C. L. Ludlow, "Effects of stimulus intensity on laryngeal long latency responses in awake humans," *Otolaryngology—Head and Neck Surgery*, vol. 117, no. 5, pp. 521–529, 1997.

[30] C. P. Browman and L. Goldstein, "Competing constraints on intergestural coordination and self-organization of phonological structures," *Les Cahiers de l'ICP. Bulletin de la communication parlée*, no. 5, pp. 25–34, 2000.

A PRELIMINARY EXPLORATION OF TONES IN PAITE

Rosie Nempalching Hangshing, Shakuntala Mahanta

Phonetics and Phonology Lab, Department of HSS, Indian Institute of Technology Guwahati
h.rosie@iitg.ac.in , smahanta@iitg.ac.in

ABSTRACT

In this paper, we report the results of a series of experiments carried out on tones in Paite (ISO 639-3: pck) and provide an acoustic analysis of its tones. Paite is an endangered and under-documented Tibeto-Burman language spoken primarily in Manipur, India. This study confirms the presence of three contrastive tones in Paite. We built a linear mixed-effects models of f0 and duration to examine the tonal contrast using the likelihood ratio test which shows statistically significant contrast for f0 and confirms the presence of three contrastive tones in Paite.
Further, an examination of consonant-tone interaction suggests the presence of a fourth tone which occurs in a restricted environment. We conclude that Paite has three contrastive tones: rising, level and falling, and a fourth tone, arising from the interaction of a contour tone with final stops. We refer to the fourth tone as a low checked tone.

Keywords: Paite, tone, consonant-tone interaction.

1. INTRODUCTION

Paite (ISO 639-3: pck) belongs to the Northern Kuki-Chin sub-group of the Tibeto-Burman family [1], [2], [3], [4]. It is primarily spoken in the state of Manipur in India by about 79,500 speakers [5]. As Manipur is home to about 33 Tibeto-Burman languages, there is pressure from larger neighbouring languages and other sociolinguistic influences such as language contact between multiple similar languages which is causing decline in intergenerational transmission. Consequently, this study is a preliminary attempt devoted to investigating tones in Paite, an under-documented language.
In the previous literature, conflicting claims have been made regarding the number of tones found in Paite [6], [7], [8], [9]. It is also observed that none of the accounts were backed by instrumental evidence. Moreover, the previous literature also does not provide lexical items which are contrastively specified for tones.
In order to establish the tonal contrasts in Paite, we prepared a list of homophonous words exhibiting tonal contrast and this was primarily done by the first author who is a native speaker of Paite. After consulting with many speakers and analysing the lexicon in these dictionaries [10], [11], we were convinced that there are three linguistically-contrastive tones in the language. This study also looks into the nature of the consonant-tone interaction found in Paite which triggers the occurrence of a fourth tone.

2. ACOUSTICAL ANALYSIS OF PAITE TONES: METHODOLOGY

A production experiment was conducted with 7 sets of minimal tone triplets (7x3=21 unique words). The word was embedded in a medial position in a carrier sentence of the type “I ‘x’ said”, x being the target word. 10 native speakers (7 male and 3 female, aged between 25 to 40) of Lamzang variety from Churachandpur district of Manipur participated in the experiment. All the participants spoke Paite as their first language and could also speak English. There were 3 iterations each resulting in 630 tokens. The target words in the dataset were of CV, CVV, and CVC syllable types.

Word	Gloss	Tone
/bêl/	pot	falling
/bēl/	shelter	level
/bĕl/	early	rising
/pôl/	straw	falling
/pōl/	mingle	level
/pŏl/	colourful	rising

Table 1: Examples of tone minimal triplets considered for the experiment.

Sentence	Gloss	Tone
/kēn bêl tsī/	I ‘pot’ said	falling
/kēn bēl tsī/	I ‘shelter’ said	level
/kēn bĕl tsī/	I ‘early’ said	rising
/kēn pôl tsī/	I ‘straw’ said	falling
/kēn pōl tsī/	I ‘mingle’ said	level
/kēn pŏl tsī/	I ‘colourful’ said	rising

Table 2: Examples of target tone minimal triplets embedded in carrier phrase.

The recordings were carried out at the Phonetics and Phonology Lab, IIT Guwahati in a sound proof recording booth using a Tascam DR-100 MKII recorder. Four tier TextGrid files were created in Praat [12]. The tiers were TBU, individual phonemes, words, and sentences. The pitch of the onset (the point of initiation) and offset (the point of termination) of the TBU was observed in Praat. TBUs in Paite are the syllable nucleus or the entire rhyme if the coda is a sonorant. A Praat script was run to extract the pitch values at every 10% of the total duration of the TBU. Thus, there were 11 consecutive points at which the pitch was calculated, starting from the onset at 0% till the offset at 100%. Each point represented 10% of the total length of the pitch track. We did two measurements i) Averaging of percentage-wise pitch values for all iterations and ii) Z-score normalization of all raw pitch values using the formula Z= (f0i – x)/SD, where f0i is the sampling point, x is the average f0 of all the sampling points, and SD representing the standard deviation of all the sampling points. The percentage-wise pitch values in Hz were normalized and averaged across all the tokens of all the speakers for all the 11 points and plotted on a graph to observe the distinct pitch contours.

2.1. Acoustical analysis of f0

The minimal tone triplets were examined to understand the possible tonal contrasts in Paite. The normalised pitch values of all iterations of all speakers at 11 consecutive points were used to draw line diagrams for visual examination. Figure 1 & 2 display the contrastive tones.

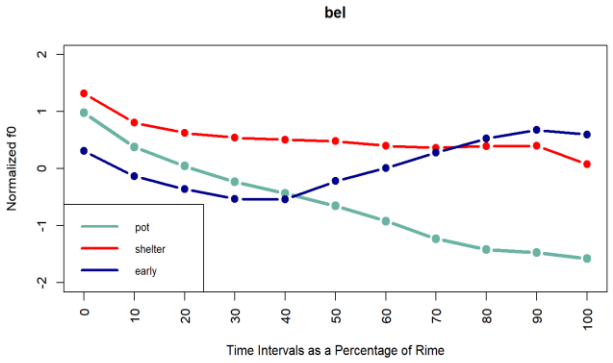


Figure 1: Normalized f0 interval contrast for the tonal triplet bêl ‘pot’, bēl ‘shelter’ and bĕl ‘early’.

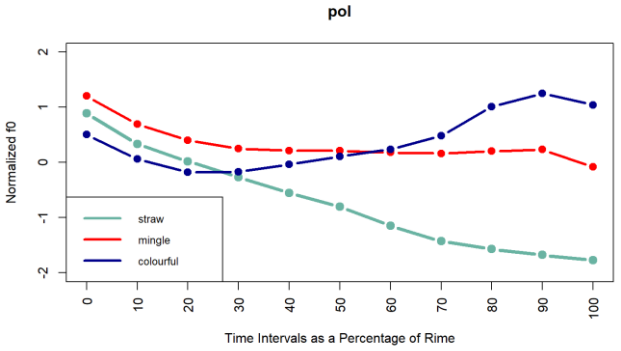


Figure 2: Normalized f0 interval contrast for the tonal triplet pôl ‘straw’, pōl ‘mingle’ and pŏl ‘colourful’.

The y-axis in figures 1 & 2 represent the normalized f0 values and the x-axis represents the time interval at 11 consecutive points of the rime. The graphs show a clear contrast between the f0 of falling, level and rising tones. The f0 contour of the three tones are significantly different from each other.

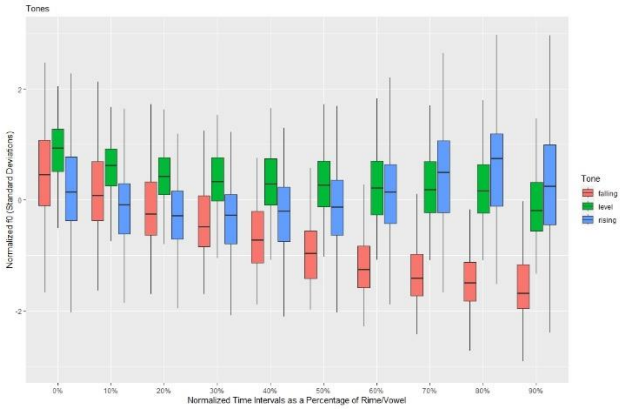


Figure 3: Boxplot of normalized f0 intervals showing falling, level and rising tones.

The deviation from the mean of the three tones is visualized in the boxplot in figure 3 which shows the f0 contrast at every 10% of the total duration. The f0 for the three tones is plotted at each interval point clearly showing the movement of the pitch. The results of the acoustic experiment indicate the presence of three contrastive tones in Paite. All the three tones show a dynamic f0 contour.

2.2. Statistical Analysis

For statistical analysis, a linear mixed-effects model was built using lme4 in R [13]. Tone.model= lme4::lmer(F0~Tone+(1+Tone|Subject)+(1+Tone|Word),data=tone) was the full model created. Normalized f0 and tone were the fixed effects. Speakers and words were random factors.

	Estimate	Std. error.	t value
Intercept	-0.3312	0.2204	-1.502
Level	0.7535	0.2789	2.701
Rising	0.3069	0.3570	0.860

Table 3: Fixed effects for normalized f0 contrast between falling, level and rising tones.

The fixed-effects results showed a distinction between the level tone, rising tone and f0 of the intercept which represents the falling tone. The estimate of the intercept is about -0.3312. A comparison of the full and null model using the likelihood ratio test [14] showed that tone affects pitch ($\chi 2$ (1) = 6.8331, df = 1, p = 0.008948), which reveals that the contrast is statistically significant.

	Estimate	Std. error.	t value
Intercept	259.93	16.887	15.393
Level	7.81	19.210	0.407
Rising	16.66	22.122	0.753

Table 4: Fixed effects for duration contrast between falling, level and rising tones.

The raw values of duration of each target word were measured to check if duration is a possible cue for tonal contrast. Duration.model= lme4::lmer(Duration ~ Tone + (1+Tone|Speaker) + (1+Tone|Word), data = dur) was the model created with duration and tone as the fixed effects. Random effects were gender and iteration. No significant difference for duration between the tones was found ($\chi 2$ (1) = 4.8654, df = 2, p = 0.0878). The level and rising tones are lengthened by a meagre 7.81 ms and 16.66 ms respectively when compared to the intercept which represents the falling tone.

3. CONSONANT-TONE INTERACTION IN PAITE LEAD TO A FOURTH TONE

Paite has a two-verb stem system where there is tonal and segmental alternation between the two verb forms. This phenomenon is observed in other Kuki-Chin languages such as Thadou [15], Vaiphei [16], Mizo [17] and Hakha Lai [18] as well. There is not enough linguistic work on Paite. There is hardly any literature that discusses its verbal morphology and we aim to explore this in future works. Verb-stem alternations may arise from various morphological and syntactic functions which are beyond the scope of this paper. This paper is limited to the interaction between segments and tones involved in these alternations and how these processes suggest the presence of a fourth tone in Paite.

Our research shows how this verb-stem alternation necessitates segmental as well as tonal change. The difference between the two alternates can be observed tonally as the two forms carry different tones. We believe that stem II forms are derived from stem I forms as the distribution of stem II tones is considerably limited. Usually, the tone of stem II undergoes a process of lowering, along with a change in segment word finally. Stem II typically carries a low tone.

3.1. Data sets

We present below in tables 5, 6 & 7 data sets to show how a low tone emerges as a result of the alternation. We followed the same process of data collection for the low tone as for the other tones. The previous 10 subjects recorded these words. The data set contained 6 words with 3 iterations each. A total of 180 tokens were considered. The falling tone is described as a sequence of high-low (HL) and the rising tone as low-high (LH).

HL tone	L tone	Gloss
/pân/	/pàt/	‘start’
/dîn/	/dìt/	‘return’
/kêm/	/kèp/	‘keep’

Table 5: HL in verb stem I changing to L in verb stem II, along with nasals changing to final stops.

HL tone	L tone	Gloss
/bâa/	/bàt/	‘borrow’
/nâa/	/nàt/	‘pain’
/tûu/	/tùt/	‘sit’

Table 6: HL in verb stem I changing to L in verb stem II, along with the insertion of a final stop.

LH tone	L tone	Gloss
/tʰǒu/	/tʰòʔ/	‘get up’
/lǒu/	/lòʔ/	‘pluck’
/zǒu/	/zòʔ/	‘finish’

Table 7: LH in stem I changing to L in stem II. Additionally, there is glottal stop insertion along with a vowel change.

We observe that the final consonant is always either an oral or a glottal stop. In table 5, HL in verb stem I changes to L in verb stem II, along with nasals changing to final stops. In table 6, HL in verb stem I changes to L in verb stem II, along with the insertion of a final stop. In table 7, LH in stem I changes to L in stem II along with a glottal stop insertion. We observe that a fourth tone emerges as a result of the interaction between a final stop consonant and a contour tone (rising LH or falling HL) in the verb-stem II form. The resulting stem II tones are consistently low.

3.2. Low checked tone in Paite.

The f0 contour of the checked tone when compared to the other three tones has a lower pitch than the falling and level tones. It also exhibits a low-falling contour. An examination of figure 4 helps us visualize how the pitch of the tone that results from the consonant-tone interaction is invariably low.

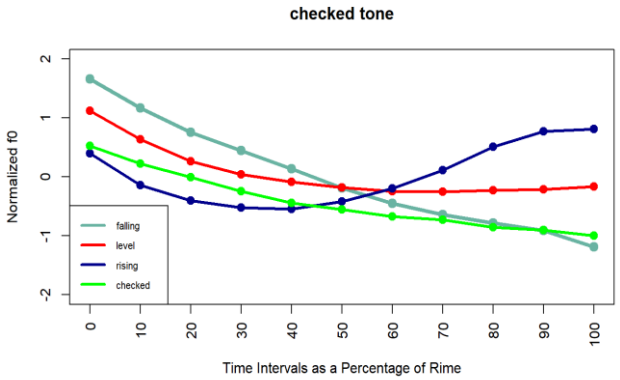


Figure 4: Normalized f0 of checked tone in contrast with falling, level and rising tones.

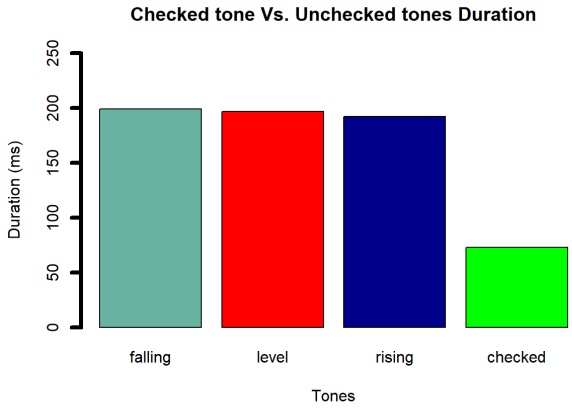


Figure 5: Duration of low checked tone compared to falling, level and rising tones.

This tone is observed to behave in a similar manner as the checked tones found in other tonal languages such as Taiwanese [19] and Xiapu Min [20] where the checked tones show a clearly distinct f0 contour. The low checked tone in Paite is also found to be significantly shorter in duration than the other three tones as shown in figure 5. In Burmese [21], Mandarin [22] and Taiwan Min [23] as well, checked tones are distinguished phonetically by a notably shorter duration.

4. CONCLUSION

Acoustic analysis performed on the recorded speech of 10 native speakers confirms the presence of three contrastive tones in Paite. This is further substantiated by likelihood ratio test results revealing that the f0 contrast between the three tones is statistically significant. On the other hand, it is found that duration does not play a role in providing a cue for tonal contrast. Further, a fourth tone is detected in Paite which occurs only with short vowels followed by either a final oral or glottal stop. It is also observed that the pitch of the fourth tone is consistently low. This tone is identified as a low checked tone. This tone is also significantly shorter in duration when compared to the other three tones. So, the results obtained in this study establish that there are three contrastive tones in Paite: falling, level and rising, and a fourth low tone that occurs in a restricted environment.

5. REFERENCES

[1] Grierson, G. A. 1904. *Linguistic Survey of India, Vol III, Part III.* Office of the superintendent of government printing, India.

[2] Kamkhenthang, H. 1988. *The Paite: A Transborder tribe of Burma and India.* Mittal Publications.

[3] Vanbik, K. 2009. *Proto-Kuki-Chin: A Reconstructed Ancestor of the Kuki-Chin languages*. PhD Thesis, University of California, Berkeley.

[4] Button, C. 2011. *Proto northern chin.* University of California.

[5] Census of India. 2011. Office of the Registrar General, Govt. of India.

[6] Elangbam, S. 1998. *A Grammatical Comparison of Kom, Manipuri, Paite and Tangkhul*. PhD Dissertation, Manipur University, Imphal, India.

[7] Muana, T.K. 1998. *Nominal Phrase in Paite and English: A Contrastive Study.* PhD Dissertation, Manipur University, Imphal, India.

[8] Singh, N. S. 2006. *A Grammar of Paite*. Mittal Publications.

[9] Yumnam, S. D. 2010. *A Comparative Phonology of Hmar, Paite and Vaiphei*. PhD Dissertation, Manipur University, Imphal, India

[10] Kamkhenthang, H. 2022. *Paite-English Dictionary*. Published by Babmang Hauzel.
[11] Tawmbing, C. 2014. *English-Paite Dictionary*. Lamka, Manipur: Hornbill Publication.
[12] Boersma, P., Weenink, D. 2021. Praat: *doing phonetics by computer* (version 6.1.38). Available: http://www.praat.org.
[13] R Core Team. 2021. R: a language and environment for statistical computing (version 4.0.4*). R Foundation for Statistical Computing*, Vienna, Austria. Available: https://www.R-project.org.
[14] Ben-Akiva, Moshe & Joffre Swait. 1986. The Akaike likelihood ratio index. *Transportation Science* 20, 133–136.
[15] Sreenivasan, A. 2015. *Phonetics and Phonology of Tones in Thadou: Implications for Typology.* PhD Dissertation, EFLU Hyderabad, India.
[16] Prichard, J. A. 2018. *Verbal Stem Alternation in Vaiphei*. Doctoral dissertation, Portland State University.
[17] Chhangte, L. 1986. *A Preliminary Grammar of the Mizo Language (Tibeto-Burman)*. Doctoral dissertation, The University of Texas at Arlington.
[18] Hyman, L.M. and VanBik, K. 2002. Tone and stem2 formation in Hakha Lai. *Linguistics of the Tibeto-Burman Area*, *25*(1), pp.113-121.
[19] Kuo, G. 2010. Tone recognition of the two checked tones in Taiwanese. *The Journal of the Acoustical Society of America* 128(4), 2476-2476.
[20] Chai, Y., & Ye, S. 2022. Checked Syllables, Checked Tones, and Tone Sandhi in Xiapu Min. *Languages* 7(1), 47.
[21] Gruber, J.F. 2011. *An articulatory, acoustic, and auditory study of Burmese tone*. Georgetown University.
[22] Bo, W.U. 2018. An acoustic analysis of vowels in checked syllables in Chinese. *Chinese Journal of Acoustics*, *4*.
[23] Pan, H.H., 2017. Glottalization of Taiwan Min checked tones. *Journal of the International Phonetic Association*, *47*(1), pp.37-63.

PHONETIC IMPLEMENTATION OF PHONOLOGICALLY DIFFERENT HIGH TONE SPANS IN LUGANDA

Cerys Hughes, Seung Suk Lee, Alessa Farinella & Kristine M. Yu

University of Massachusetts, Amherst, USA
ceryshughes@umass.edu; seungsuklee@umass.edu; afarinella@umass.edu; krisyu@linguist.umass.edu

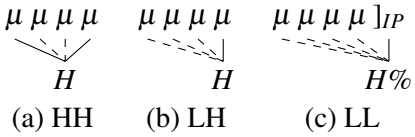
ABSTRACT

Substantial phonetic work exists both on long-distance phonological processes and on incomplete neutralization for segments, but not for tone. To help fill these gaps, we re-analyze Myers et al. 2018's Luganda production experiment on high tone spans derived by distinct phonological processes. They found differences in peak delay between two kinds of surface-identical high tone spans.We build on their analysis by analyzing the shape of the whole high span f0 trajectory using functional data analysis (FDA) and generalized additive mixed models (GAMMs). A previous reanalysis suggested the tone spans also differ in degree of sag in their plateau. We found this wasn't supported with different f0 normalization and incorporating richer random effects in trajectory shapes with GAMMs, although GAMMs supported different rise shapes to the plateau. In addition to addressing an empirical gap, we thus provide a cautionary tale on studying incomplete neutralization in dynamic trajectories.

Keywords: Tone, phonetics-phonology interface, incomplete neutralization, FDA, GAMMs

1. INTRODUCTION

A string of like-tones in the surface representation, e.g., a high tone span, can arise from multiple phonological tonal spreading processes. For example, [1] describes high tone spans in Luganda derived from three different spreading processes, as schematized in Fig. 1. Tones in Luganda are either high (H) or low (L), with both lexical and boundary tones. HH spans are high tone plateaus, which occur between two lexical high tones that spread to all of the syllables between them (subject to syntactic conditioning). LH spans occur when there is a low tone followed by a lexical high tone within a phrase and the high tone spreads leftward. LL high tone spans occur when there are no lexical high tones but instead an intonational H% that spreads leftward. Our current investigation does not focus on an additional "short" span discussed in [1], HL, but we include it in our replication of [2].



(a) HH (b) LH (c) LL

Figure 1: Relevant high tone span types in Luganda from [1], each derived by different tone spreading processes.

There is some work on the phonetic implementation of sequences of like-tones in languages with lexical and/or grammatical tone where each tone bearing unit (TBU) in the tone span is already specified in the underlying representation (e.g., in Yoruba and Mandarin [3, 4]). But [1] is, to our knowledge, the only instrumental study of like-tone spans arising from different spreading processes. It offers a case study for exploring whether underlying phonological contrasts are incompletely neutralized in spans with the same surface tones, in line with findings of incomplete neutralization of segmental contrasts.

There is a proliferation of studies of segmental incomplete neutralization, as evidenced by the feasibility of [5]'s meta-analysis of fourteen studies of German final devoicing alone. In contrast, there are just a handful of studies of incomplete neutralization of tonal processes, and these seem to be limited to a few Chinese varieties. Unlike the point-based acoustic measures of interest in segmental studies of incomplete neutralization such as in devoicing or flapping, the acoustic object of interest in tonal studies is the f0 trajectory shape. Statistical methods for analyzing shape components of trajectories, e.g., smoothing spline ANOVAs, functional data analysis (FDA) [6], and generalized additive mixed effects models (GAMMs) [7], have enabled prosodic analyses to address f0 trajectory shape in addition to f0-based measures at inflection points [8, 9, 10, 11], a.o., including in studies of incomplete neutralization in tonal sandhi [12, 13].

[1] focused on f0 turning points and pointwise measures in their analysis of high tone spans in

Luganda, which characterized the timing and f0 excursion size of the rise into and fall out of the high f0 plateau. They found that the lexical spans (e.g. HH and LH) have similar normalized f0 levels, HH and LH spans have a long plateau between f0 rise and fall, and that intonational span (LL) f0 contours are below those of the lexical spans (see Fig. 4). [1] also found that LH spans have more peak delay (with respect to the onset of the syllable initiating the H span) relative to all other classes, and took this as evidence that underlying phonological contrasts are not fully neutralized between LH and HH spans, which differ only in whether the initial syllable of the H span is underlyingly H-toned or not.

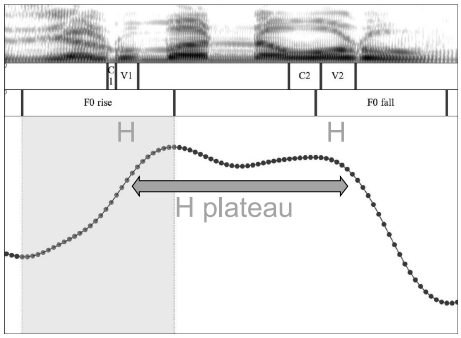


Figure 2: An HH token with noticeable sag in its f0 plateau, figure adapted from [1], Fig. 4

While [1] did not analyze the f0 trajectory over the high f0 plateau, a re-analysis of the same data from [1] in [2, 14] noticed that exemplar f0 contours of HH and LH spans in [1] both showed "sag" (a dip) in the middle of the plateau, with a greater degree of sag for the HH tokens (Fig. 2). This "sag" is reminiscent of the description of a "sagging" transition between two high tone targets in American English intonation in [15] that is a function of the distance between them, with greater distance resulting in a greater degree of sag, see also [16, 17]. [2, 14] used FDA to characterize shape components of the high tone plateau f0 trajectories and provided additional evidence for incomplete neutralization between HH and LH spans, due to more sag in the plateau of HH spans than LH spans.

This paper shows that: (i) [2, 14]'s finding of a difference in the degree of sag of the high plateaus of HH and LH span is not supported with the addition of richer random effects structures, (ii) but that GAMMs support a difference in the shape of the rise trajectory into the plateau which could have a perceptual effect [18] consistent with [1]'s reported acoustic LH-HH peak delay difference. We also make a methodological contribution in providing a case study of the effect of researcher degrees of freedom in assessing incomplete neutralization, an issue raised in [5]'s meta-analysis of German final devoicing. Specifically, the findings about sag and incomplete neutralization are greatly impacted by the choices of including speaker or item as random effects and of normalizing f0 by speaker or token.

2. METHODS

We re-analyzed the recordings of 10 native speakers of Luganda from [1], also analyzed in [2, 14]. Data consisted of approximately 20 sentences per span type for each of the four span types (the 3 shown in Fig. 1 in addition to the short span HL) for a total of 80 items. All items were elicited from all 10 speakers, and [1] included 798 tokens for analysis in total. We additionally discarded 4 HL tokens with span durations greater than 3SD higher than the mean HL span duration, and 1 LL token with heavy laryngealization. We followed the data pre-processing procedure in [2, 14]. F0 measurements were extracted using speaker-specific f0 floors and ceilings with STRAIGHT implemented in VoiceSauce [19]. A handful of clear gross f0 tracking errors were manually corrected. For both FDA and GAMMs, the analysis window was the span duration—the interval from the rise onset through the fall offset marked in Praat TextGrids [20] from [1] (Fig. 2).

FDA studies of f0 contours have varied in how they process f0, e.g., raw f0 values in Hz [9], conversion to semitones and centering about the mean within each token [12, 13], and by-speaker z-score normalization [10, 11]. While [2, 14] followed [12, 13]'s within-token mean-centering, [1] used by-speaker z-score normalization. For comparison to previous analyses of [1]'s data, we therefore ran parallel analyses with both types of f0 transformation for FDA and GAMMs. Span type contrasts were defined as treatment contrasts with HH-span as the reference level, since we were interested in comparing the LH-span (and other span types) to the HH-span. Statistical significance was assessed at an α-level of 0.05.

Data analysis was done in R [21]. Like in [2, 14], FDA was implemented with the *fda* package [6] on the full data set, including HL spans, B-spline basis functions were used for smoothing F0 measurements and parameterizing the principal components, and mixed effects regressions were conducted using *lme4* [22] and *lmerTest* [23]. For the long span subset (we obtained the same pattern of results with the full data set) we built GAMMs with by-speaker z-scored f0 values or

by-token centered f0 values (30 evenly spaced samples, with normalized time defined as proportion of span duration) as the dependent variable, span and duration as parametric variables, an interaction between proportion of span duration and span type, and random smooths of proportion of span duration for participants and items with *mgcv* [7]. We corrected for autocorrelation in the residuals (due to the non-independence of time series data) and heavy tails in the distribution of measurements observed in GAMMs diagnostics by including an autoregressive (AR1) model and specifying a scaled-*t* distribution [24]. We coded the interaction between proportion of span duration and span type as nominal factor difference smooths for *itsadug* [25] visualization of f0 trajectory shape differences, and as ordered factor difference smooths for inferential tests on the difference between spans relative to the HH reference level [26]. Further details for data processing and modeling (including for other span type comparisons besides HH vs. LH), can be found in the OSF repository at https://osf.io/p4gj5/?view_only=95e88cf6724447fc93eb8684723762b6.

3. RESULTS

3.1. FDA

We first replicated [2, 14]'s FDA procedure using within-token f0 centering and data from all four tone spans. FDA with by-speaker z-scored f0 values instead resulted in a fourth functional principal component (PC4) (5.8% of variability in data, see Fig. 3) that looked nearly identical to the PC3 indexing sag in [2, 14]. A higher coefficient for these PCs indexes a decrease in degree of sag.

A mixed effects logistic regression model in [14, 2] with span type as the dependent variable, the PCs as non-interacted fixed effects, and a by-speaker random intercept showed that more sag (as indexed by a PC like in Fig. 3) significantly increased the probability of HH relative to an LH span. However, once we also included by-item random intercepts, none of the PCs reached significance for classifying span type. Similarly, for the by-speaker f0 normalized data, the PC4 sag component was significant when only by-speaker random intercepts were included, but no PCs were significant when by-item intercepts were included (the model didn't converge with both types of random intercepts).

Since incomplete neutralization studies typically estimate potential differences in acoustic measures between different categories, we also ran linear regressions with each of the sag components (PC3 for by-token centering; PC4 for by-speaker

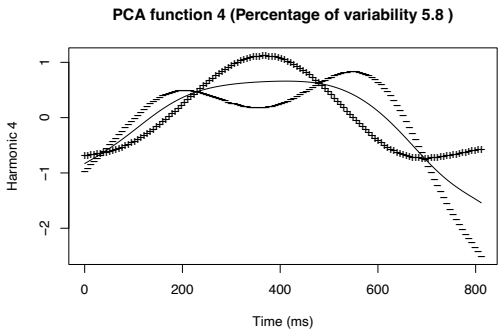


Figure 3: PC4 from by-speaker f0 normalization indexes a sag shape component nearly identical to one found in [2, 14] with within-token f0 centering. The minus sign curve indicates the effect on the resulting curve when the PC's coefficient is decreased, and the plus sign curve shows the effect when it is increased.

normalization) as the dependent variable, span type interacted with span duration as the fixed effects, and random intercepts by-speaker and by-item (random slopes models did not converge). For the by-token centering model, all span types, duration, and span type-duration interactions were significant. In particular, for the LH vs. reference level HH contrast, the sag component coefficient was higher for LH (less sag) than HH spans, particularly for shorter span durations. However, for the by-speaker normalization model, the LH vs. HH span contrast was marginal (β=11.2, (SE 3.5), t = 1.8, p = 0.07) and its interaction with duration was insignificant. Thus, while both f0 normalization choices showed the same direction of effects for the sag component, they differed in whether LH vs. HH span significantly affected the degree of sag. Regression diagnostics indicated non-normality—a heavy-tailed distribution for both linear models, which we addressed with GAMMs.

3.2. GAMMs

With richer random effects structures in GAMMs (by-speaker and by-item adjustments for trajectory shape), as well as corrections for heavy tails and the non-independence of f0 values within the same time course, differences in the degree of sag between the LH and HH spans are not supported with either f0 normalization choice. Estimated GAMMs partial effects for the two kinds of normalization are in Fig. 4.

The LL span trajectory is clearly lower than other span trajectories, as found in [1]. The overlapping confidence intervals between the LH

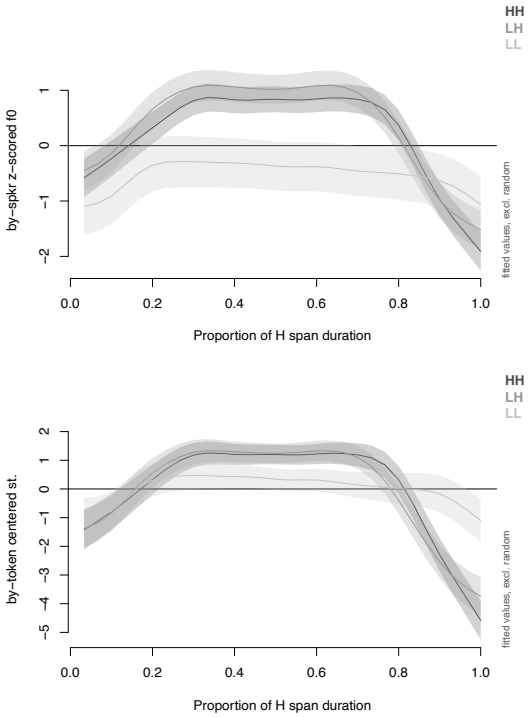


Figure 4: Estimated GAMMs partial effects of span type among long spans. (top) by-speaker z-scores (top); (bottom) by-token centered st.

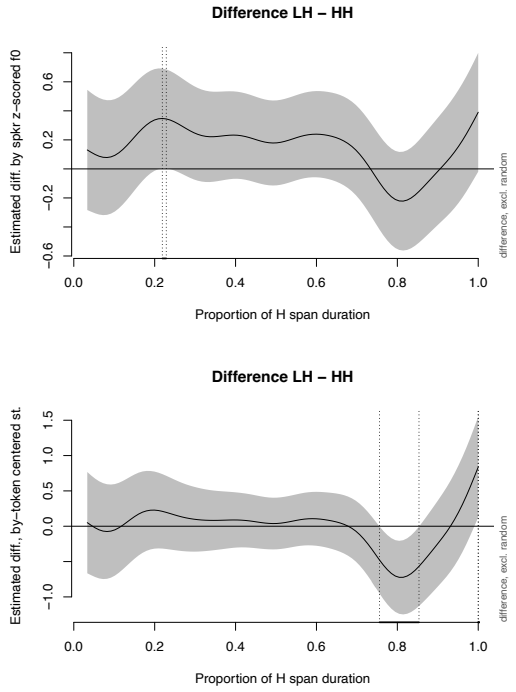


Figure 5: GAMMs estimated differences for LH span trajectories relative to HH ones; pointwise 95% CIs shown as shaded bands; by-speaker z-scored f0 (top); by-token centered st. (bottom)

and HH span plateaus indicate a lack of support for differences in trajectory shape in the plateau (formally assessed by non-significance of the LH-HH ordered factor difference smooths), but suggest potential differences in the rise/fall. We can better visualize the regions of differences with the difference smooths between LH and HH shown in Fig. 5. The difference smooth for by-speaker z-scores shows that significant differences between the LH and HH spans occur only in the rise to the plateau—the dashed region at 22-24% of the H span duration—but the difference smooth for by-token centered st. shows that significant differences occur only in the fall out of the plateau, at about 80% of the H span duration. While the robustness of these small differences at the contour edges is unclear, they demonstrate the impact of normalization choice.

By combining Figs. 4, 5, we can see that the tiny shape difference between the by-speaker z-scored LH and HH spans is that the LH span rise into the plateau bulges out more (a "domey" shape) than the straighter HH span rise (perceptually consistent with [1]'s peak delay difference finding [18]). But for by-token centered st., the difference is that the LH span fall bulges out downward more than the straighter HH span fall. (Differences at the end are likely an artifact of edge effects in f0 estimation.)

4. DISCUSSION AND CONCLUSION

GAMMs do support incomplete neutralization between LH and HH spans through small but consistent f0 differences. However, the effect of f0 standardization choice on whether the difference was in the rise or fall demonstrates the impact of analysis choices in incomplete neutralization studies. It's not obvious that one choice is more theoretically-motivated than the other. Together, with our failure to replicate the difference in degree of sag between the two spans reported in [2, 14] with richer random effects structures, these results show the need for caution in interpreting implications for phonological representations based on small phonetic implementation differences. As we have shown, even where phonetic differences appear to be statistically significant, different methodological choices can result in non-significant effects. A broader concern is that even statistically significant differences in phonetic implementation do not necessarily support the existence of underlying phonological contrasts [27]. Irrespective of incomplete neutralization effects, our trajectory-shape analyses enrich our understanding of the very limited data available on like-tone spans arising from different spreading processes.

5. ACKNOWLEDGEMENTS

We thank Scott Myers for helpful correspondence as well as sharing the Luganda tone span data he collected and analyzed.

6. REFERENCES

[1] S. Myers, E. Selkirk, and Y. Fainleib, "Phonetic implementation of high-tone spans in Luganda," *Laboratory Phonology*, vol. 9, no. 1, p. 19, Dec. 2018.

[2] C. Hughes, A. Farinella, S. S. Lee, and K. M. Yu, "Phonetic implementation of phonologically different high tone plateaus in Luganda," LabPhon 18, Virtual, Jun. 2022.

[3] Y. O. Laniran, "Intonation in tone languages: The phonetic implementation of tones in Yoruba," Ph.D. dissertation, Cornell University, 1992.

[4] C. Shih, "A declination model of Mandarin Chinese," in *Intonation: Analysis, Modeling and Technology*, A. Botinis, Ed. Kluwer Academic Publishers, 2000, pp. 243–268.

[5] B. Nicenboim, T. B. Roettger, and S. Vasishth, "Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German," *J. Phon.*, vol. 70, pp. 39–55, Sep. 2018.

[6] J. Ramsay and B. Silverman, *Functional Data Analysis*, 2nd ed. New York, N.Y.: Springer Verlag, 2005.

[7] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[8] J. E. Andruski and J. Costello, "Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong," *JIPA*, vol. 34, no. 02, pp. 125–140, 2004.

[9] J. A. D. Aston, J.-M. Chiou, and J. P. Evans, "Linguistic pitch analysis using functional principal component mixed effect models," *Journal of the Royal Statistical Society: Series C*, vol. 59, no. 2, pp. 297–317, 2010.

[10] G. Lohfink, A. Katsika, and A. Arvaniti, "Variability and category overlap in the realization of intonation," in *Proc. ICPhS 19 Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 701–705.

[11] S. Gryllia, K. Marcoux, K. Jepson, and A. Arvaniti, "The many shapes of H*," in *Speech Prosody 2022*, May 2022, pp. 754–758.

[12] C. Cheng, J.-Y. Chen, and M. Gubian, "Are Mandarin Sandhi Tone 3 and Tone 2 the Same or Different? The Results of Functional Data Analysis," in *Proceedings of PACLIC27*, 2013, pp. 296–301.

[13] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *J. Phon.*, vol. 49, pp. 16–40, Mar. 2015.

[14] S. S. Lee, A. Farinella, C. Hughes, and K. M. Yu, "Phonetic implementation of phonologically different high tone plateaus in Luganda," 1st Intl. Conf. on Tone and Intonation, University of Southern Denmark, Dec. 2021.

[15] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[16] ——, "Synthesizing intonation," *JASA*, vol. 70, no. 4, pp. 985–995, 1981.

[17] D. R. Ladd and A. Schepman, ""Sagging transitions" between high pitch accents in English: Experimental evidence," *J. Phon.*, vol. 31, pp. 81–112, 2003.

[18] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology," *Laboratory Phonology*, vol. 3, no. 2, Jan. 2012.

[19] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," *Proceedings of ICPhS XVI*, 2011.

[20] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (Version 6.3.03) [Computer program]." 2022.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[24] J. van Rij, P. Hendriks, H. van Rijn, R. H. Baayen, and S. N. Wood, "Analyzing the Time Course of Pupillometric Data," *Trends in Hearing*, vol. 23, p. 233121651983248, Jan. 2019.

[25] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using GAMMs," 2022.

[26] G. Simpson, "Comparing smooths in factor-smooth interactions II," https://www.fromthebottomoftheheap.net/2017/12/14/difference-splines-ii/, Dec. 2017.

[27] J. A. Shaw and S. Kawahara, "Assessing surface phonological specification through simulation and classification of phonetic trajectories," *Phonology*, vol. 35, no. 3, pp. 481–522, 2018.

Duration and Sonority in Blackfoot Syllable Structure

Donald Derrick

University of Canterbury, New Zealand Institute of Language, Brain & Behaviour
donald.derrick@canterbury.ac.nz

ABSTRACT

Blackfoot, an Algonquian language spoken in northern Montana and Southern Alberta, reveals mechanisms of syllabification not well-described in Linguistic literature. The Blackfoot alveolar fricative /s/ can occupy any syllable position based on its duration [1]. Blackfoot's back fricative /x/ has a center of gravity based on the preceding vowel and always closes a syllable [2, 3]. Singleton plosives, affricates, nasals, and glides form syllable onsets – glottal plosives can also be codas. Geminates also form part of the nucleus of preceding syllables [2, 3]. Nasals, glides, and vowels all have the same amplitude and duration, but only the vowels form as tautosyllabic nuclei. The production-based mechanism that distinguishes vowels from nasals and glides must therefore include other information, which may be any of acoustical salience, vocal tract constriction, or visible mouth/jaw opening. Taken together, Blackfoot syllables appear to be constructed through an interaction of duration and continuant frication with vocal tract openness.

Keywords: Syllable Structure, Sonority, Duration, Multi-modal speech, Blackfoot

1. INTRODUCTION

One goal of phonology is to describe and explain cross-linguistic syllable structure. Two major theories of syllable organization are known as the 'Sonority Sequencing Principle (SSP)' [4] and the 'Syllable Contact Law (SCL)' [4-6]. These theories use a sonority scale [4, 7], which categorizes speech sounds in terms of how easily people can hear them and is acoustically correlated with acoustic energy [8-10], relative resonance [4], and segmental duration [11]. In Clements' version of the sonority scale, sound classes are ranked from least to most sonorous in the order: *Obstruents (e.g. /p/) < Nasals (e.g. /n/) < Liquids (e.g. /l/,/r/) < Glides (e.g. /w/,/j/) < Vowels (e.g. /a/, /i/)* [4].

According to the SSP, segments are syllabified in such a way that sonority, or perceptible distinctiveness, increases from the syllable margin to the peak. According to the SCL, in a sequence of syllables, the coda on the first syllable is more sonorous than the onset on the second syllable [4]. For example, the word "grandpa" [gɹæm.pɑ] has two syllables. According to the SSP, [g] is less sonorous than [r], and [r] is less sonorous than [æ], and according to the SCL, the coda [m] of the first syllable is more sonorous than onset [p] of second syllable.

While SSP and SCL are described as universal principles, only one sound class division within SSP and SCL conforms to syllable organization cross-linguistically: *Non-sonorants* (sounds produced with interrupted or turbulent air-flow) versus *sonorants* (sounds produced with continuous and non-turbulent air-flow) [12]. Other sound class divisions in the SSP and SCL are violated by many languages [12]. These violations have frustrated Linguists so much that many resorted to descriptions based on sequences that do and do not appear in the relevant language – distributional evidence [13-15]. Distributional evidence describes the facts but explains none of them.

As a good example, Blackfoot, an Algonquian language spoken in northern Montana and Southern Alberta, partially violates even the supposedly universal non-sonorant vs. sonorant SSP, sometimes treating voiceless fricatives like sonorants (e.g. [ʔiː.kóm.ʔsː.pi.kaʔ.ps.si], *iikómsspika'pssi*, 'he is hard to take care of', where it does so twice).

Here I examine segmental intensity, duration, formant position, as well as discussing unmeasured visual and tactile speech, to identify and describe the phonetic basis of syllabification in Blackfoot as an example of how to address the limitations of the analysis techniques used to describe and generate sonority theory (SSP and SCL).

1.1. Blackfoot Phonetics and Syllable Structure

Blackfoot has 21 contrastive consonants, including plosives /p, pː, t, tː, k, kː, ʔ/, fricatives /s, sː, x/, affricates /ts, tːs, ks/, nasals /m, mː, n, nː/, glides /w, j/, and the marginal pre-assibilants /ˢt, ˢtː/. Blackfoot also has 5 long /iː, ɛː, aː, ɔː, oː/ and 3 short /i, a, o/ vowels, along with several predictable variants /ɪ, ɐ, ʊ/ along with rising /jV/ and falling /Vj/ diphthongs.

Any consonant may form a Blackfoot onset except for the dorsal fricative /x/, which obligatorily closes a syllable.

Vowel nuclei may be a short vowel, long vowel, or diphthong. Nuclei may also be voiceless fricatives. The consonants /s/, along with /x/ (realized

as [ç], [x], or [x^{w}]) can act as vowels and are the only consonants which can occur between consonants. The /s/ can be short or long, whereas the /x/ is always short, coalescing with a preceding short vowel [16], forming the whole nucleus, or the second mora in a long syllable nucleus (eg [ʔiç.po.kón.ʔsː.kaː], *iihpokónsskaa*, 'he has gotten the ball').

A coda consonant may be /s/, /x/, a glottal plosive /ʔ/, the first half of a geminate consonant, or a nasal before glottal stop.

1.1.1. Syllabification of Blackfoot /s/

Blackfoot /s/ can occupy any position and length in a syllable, including: 1) short interconsonantal [CsC] ([ʔá.ps.pi.niːks], *ápspiniiksi,* 'geese'); 2) short word-initial [#sC] ([s.pé.t͡sɪ.ko], *spátsiko,* 'sand'); 3) short intervocalic [(V)sV] ([niː.síp.po], *niisíppo*, 'forty'); 4) long postconsonantal [Cs:V] ([pi.sé.ts.sɛs.ki], *pisátssaisski*, 'flower'); 5) long preconsonantal [VsC] ([ʔɪs.póm.ʔç.taːt], *isspómmihtaat,* 'help out!'); 6) long interconsonantal [Cs:C] ([mo.wé.psː.pɪn], *mowápsspin*, 'eye'); 7) short word-final [Vs#] ([mox^{w}.kín.ʔsː.t͡sɪs], *moohkínsstsis*, 'elbow'); 8) long intervocalic [Vs:V] ([nɪs.síːs], *nissíss*, 'my yng'r sibling' (of fem.)); 9) superlong postconsonantal [Cs::V] ([ní.tsː.soʔ.toː.kạ], *nítssso'tooka*, 'he felt me (up)'). As well as being part of an 10) affricate ([ʔiː.t͡síʔ.ts.ks.o.ji, *iitsí'tsksoyi*, 'picnic') or 11) pre-assibilant ([miː.^{s}t͡s.ís], *miistsís*, 'tree') [17]

1.2. Hypothesis

Given these distributional patterns of syllabification in Blackfoot, the hypothesis is that Blackfoot syllabification requires not only intensity, but also duration and vocal tract openness information to be describable through speech production outputs. The predictions tested here are: 1) Blackfoot /s/ syllable position is dependent on duration and segment adjacency, but not amplitude differences. 2) Blackfoot /x/ combines with acoustical information from the underlying preceding vowel identifiable through center-of-gravity. This information, along with its obligatory distribution before a stop or fricative, ensures it always closes a syllable. 3) Blackfoot vowels have the same amplitude and duration as glides and nasals, and so if production-based information explains their distribution, it must come from another source such as visual speech or acoustic information on vocal tract constriction as identified through air flow stoppage (for nasals) or 4) formant position (for glides).

2. METHODS

2.1. Participant

One native speaker of Blackfoot language consultant (XXX) provided all of the data. They spoke the "new Blackfoot" Káinai dialect and was recorded when they were in their 60s. The consultant was literate in the Blackfoot orthography written by Franz and Russel and originally published in 1978 [18]. This data is currently part of Blackfoot Illustration under review [19].

2.2. Materials

Recordings were completed using a Marantz 660 solid-state recorder with a countryman (phantom power) wired lapel microphone.

2.3. Procedure

The participant was seated in a sound-attenuated booth, and the wired label microphone was clipped on to their shirt collar. They were then verbally given English glosses and asked to translate the gloss and repeat it 3 times. If the consultant translated differently from what the source material suggested, a second recording of 3 repetitions was completed after they read the translation provided in the Blackfoot dictionary [18].

2.4. Analysis

Measurements were transcribed using PRAAT [20], and duration, amplitude, and center of gravity were extracted using automatic settings with a PRAAT script. Analysis was completed using R [21]. All statistical analyses were completed with simple linear models using R's built-in linear model predictor tools.

3. RESULTS

The results of the measurements are presented in order of the predictions to be tested.

3.1. Blackfoot /s/ syllable position

Comparison of the durations of Blackfoot /s/ show that /s/ duration falls into five groups (Figure 1), which are statistically significant (Table 1): A) pre-assibilation and affrication are of similar duration; B) all the short /s/ except for word-final /s/ form a bigger group, including the two tautosyllabic short syllable nuclei /s/; C) all the long /s/ except long intervocalic, including the tautosyllabic long nuclei /s/ (long interconsonental), and the short nucleus + following onset (long postconsonantal); D) the short word final

and long intervocalic /s/, which syllabifies as the other geminates do; E) the superlong postconsonantal /s/, which forms a long syllable nucleus and a following syllable onset.

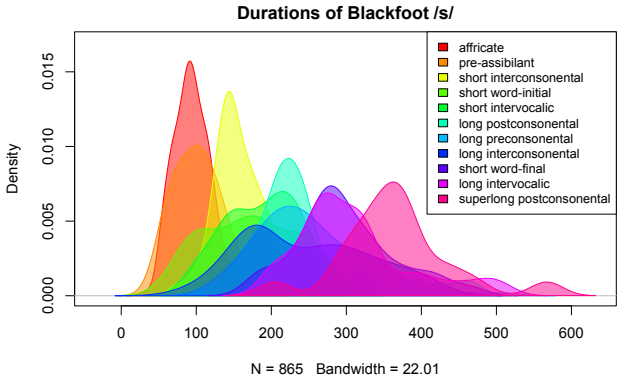


Figure 1: Duration of Blackfoot /s/ by syllable position and segment adjacency

Position	Dur (ms) Med (SD)	t-value	Group
Affricates	93 (13.4)	N/A	A
Pre-assibilant	103 (40.9)	1.5	A
Short interconsonantal	150 (32.8)	9.0	B
Short word-initial	173 (59.8)	6.4	B
Short intervocalic	191 (50.0)	12.0	B
Long postconsonantal	225 (7.3)	19.6	C
Long preconsonantal	241 (70.7)	21.0	C
Long interconsonantal	245 (84.4)	23.5	C
Short word-final	287 (66.3)	18.8	D
Long intervocalic	292 (76.7)	18.8	D
Superlong postconsonantal	368 (74.7)	20.3	E

Table 1: Duration groupings for /s/: Dur = duration, ms = milliseconds, Med = median, SD = standard deviation.

In contrast, the intensity of Blackfoot /s/ is much more difficult to group or interpret. Blackfoot /s/ intensities shows extreme overlap that do not relate well to syllable structure or duration, as seen in Figure 2.

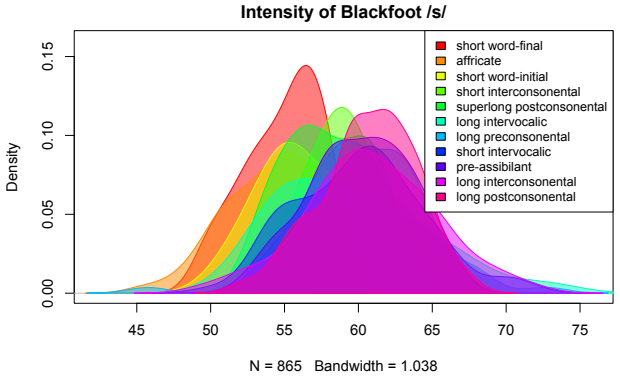


Figure 2: Intensity of Blackfoot /s/ by syllable position and segment adjacency

3.2. Blackfoot /x/ center of gravity

Blackfoot /x/ center of gravity is much higher for the fricative [ç] that follows /i/ than for the [x] that follows [a] and [o] and the [x^{w}] that follows [o] and [ɔ], as seen in Figure 3.

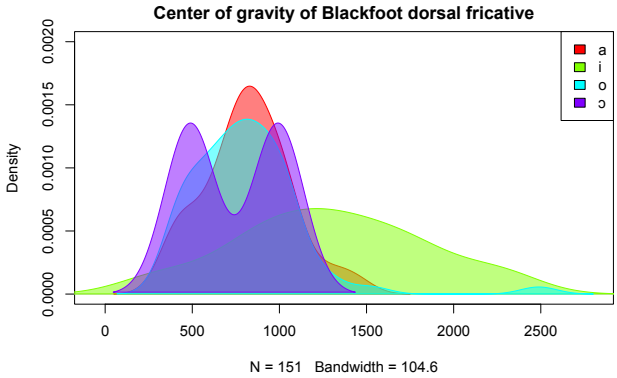


Figure 3: Center of gravity for Blackfoot /x/ variant

The differences between [ç] and the rest are significant, as seen in table 2.

Position	COG (ms) Med (SD)	t-value	Group
[x] 'ax'	825 (247)	N/A	A
[ç] 'ix'	1269 (538)	5.79	B
[x/x^{w}] 'ox'	787 (320)	-0.13	A
[x^{w}] 'ɔx'	741 (357)	-0.28	A

Table 2: Center of gravity groupings for /x/. COG = center of gravity.

3.3. Blackfoot segment intensity

Short segments are shown because there are no short glides in Blackfoot, but results are similar for long segments. Blackfoot vowels, glides, and nasals have similar amplitudes, with fricatives having lower amplitudes, and stops having the lowest amplitude, as seen in Figure 4.

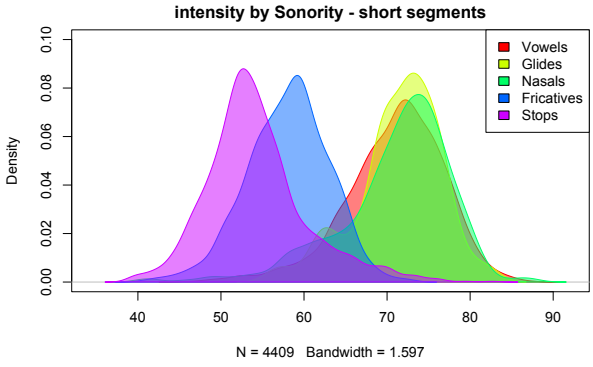


Figure 4: Blackfoot segment intensity by type

The differences between the three observable groups in Figure 4 are statistically significant, forming the three groups seen in Table 3.

Position	Int (dB) Med (SD)	t-value	Group
Vowels	71.6 (5.73)	NA	A
Glides	72.3 (5.13)	0.7	A
Nasals	71.8 (6.54)	-0.9	A
Fricatives	58.3 (4.96)	-62.1	B
Stops	53.2 (6.06)	-86.1	C

Table 3: Center-of-gravity groupings for /x/. Int = intensity, dB = decibels.

Duration distribution can be seen in Figure 5.

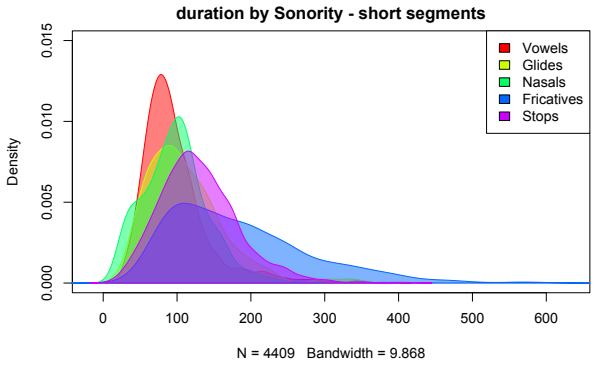


Figure 5: Blackfoot segment duration by type

Vowels and nasals are the shortest (group A), followed by glides (group B), stops (group C), and fricatives (group D), as seen in Table 4.

Position	Int (dB) Med (SD)	t-value	Group
Vowels	88 (47.1)	NA	A
Nasals	98 (43.2)	0.10	A
Glides	100 (52.3)	2.08	B
Stops	123 (53.5)	13.0	C
Fricatives	163 (94.1)	33.4	D

Table 4: Center-of-gravity groupings for /x/. Int = intensity, dB = decibels.

3.4. Blackfoot glide formants

Blackfoot glides have different formants than their related vowels (the vowel [i] for [j], and the vowel [o] for [w]), as seen in Figure 6.

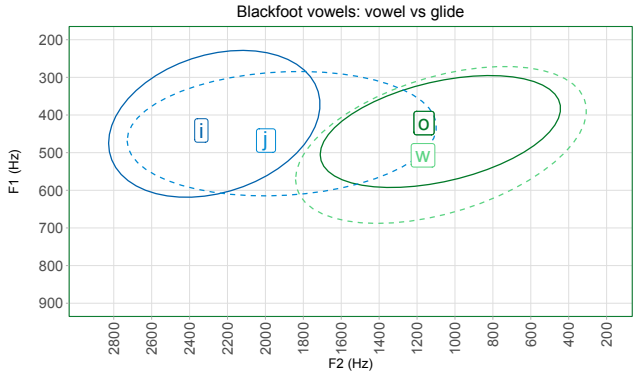


Figure 6: Formants Blackfoot glides vs vowels

4. DISCUSSION

The results show support for prediction 1, that /s/ duration, when combined with information about adjacent segments (vowels, consonants, or a word boundaries) to provide sufficient information to identify the syllable position of the given /s/ segment. In contrast, /s/ intensity does not provide enough information to aid in identifying syllable position.

The results partially support prediction 2 in that the center of gravity for [ç] are distinguishable from those for [x] and [x^{w}]. However, center of gravity alone does not provide sufficient information to distinguish all of the underlying vowels from each other. Nevertheless, when listening to this dataset, it is relatively easy to hear the effects of the specific merged vowel, so information from speech production is sufficient. This acoustic information, combined with the distributional evidence, identifies the dorsal fricative /x/ as a segment that always closes syllables.

The results support prediction 3 in that intensity of vowels, glides, and nasals are similar in Blackfoot, and while the duration of all three are not quite identical, it is hard to imagine that intensity differences provide enough to explain why nasals and glides cannot form syllable nuclei when vowels can. Instead, nasals stop airflow from leaving the mouth – they are stops – which provides its own production-based distinction between nasals and vowels. Similarly, in support of prediction 4, glides have differing formant values and are distributed before vowels.

Blackfoot also sometimes ends words in nearly inaudible voiceless vowels; speakers who notice mispronunciations may ask learners to watch them speak [22]. This behaviour suggests native Blackfoot speakers learn their syllable structure rules in part through visual, rather than acoustic, cues. This additional acoustic and visual information provides the information needed to improve our understanding of the connection between multi-modal speech production and syllable structure.

Taken together, the results show that Blackfoot syllables are constructed through an interaction of segment adjacency, duration for /s/, and the degree of oral vocal tract openness for voiced continuants.

5. ACKNOWLEDGEMENTS

Thanks to the Jacobs Research Funds and to SSHRC funds to Martina Wiltschko for providing the resources to record and document the Blackfoot data. Thanks also to Natalie Weber for teaching me so much about Blackfoot and for her support. This

proceedings paper is dedicated to the memories of Tootsinam (Beatrice Bullshields), our Blackfoot consultant who provided the recordings used in this Illustration, to Abigail Scott who worked on Blackfoot research with us, and to Donald G. Frantz who authored so much excellent work on Blackfoot. You will all be missed.

6. REFERENCES

[1] Derrick, D. 2007. Syllabification and Blackfoot /s/. In *Proceedings of the Northwest Linguistics Conference (NWLC) 22*, 62–76. Vancouver.

[2] Elfner, Emily. 2006b. Contrastive syllabification in Blackfoot. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, Baumer, Donald and Montero, David and Scanlon, Michael (eds.), 141–149. Somerville, MA: Cascadilla Proceedings Project.

[3] Weber, Natalie. 2020. *Syntax, prosody, and metrical structure in Blackfoot*. PhD dissertation, University of British Columbia.

[4] G. N. Clements, "The role of sonority cycle in core syllabification," *Papers in laboratory phonology,* vol. 1, pp. 283-33, 1990.

[5] R. W. Murray and T. Vennemann, "Sound change and syllable structure in Germanic phonology," *Language,* vol. 59, pp. 514-528, 1983.

[6] J. B. Hooper, *An introduction to natural generative phonology.* . Academic Press, New York, 1976.

[7] E. Sievers, *Grundziige der Phonetik, Leipzig*. Breitkopf and Hartel., 1881.

[8] R.-M. S. Heffner, *General Phonetics*. University of Wisconsin Press, 1950.

[9] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. Wadsworth, 2014.

[10] S. Parker, "Sound level protrusions as physical correlates of sonority," *Journal of Phonetics,* vol. 36, no. 1, pp. 55-90, 2008.

[11] M. Komatsu, S. Tokuma, W. Tokuma, and T. Arai, "Multi-dimensional analysis of sonority: Perception, acoustics, and phonology," in *Seventh International Conference on Spoken Language Processing*, 2002.

[12] C. Jany, M. Gordon, C. M. Nash, and N. Takara, "How universal is the sonority hierarchy?: A cross-linguistics acoustic study," presented at the ICPhS XVI, 2007.

[13] J. J. Ohala and H. Kawasaki-Fukomori, *Alternatives to the sonority hierarchy for explaining segmental sequential constraints*. De Gruyter Moutton, 2011.

[14] R. Wright, "A review of perceptual cues and cue robustness. ," in *Phonetically based phonology*, vol. 34, B. Hayes, R. Kirchner, and D. Steriade Eds., 2004, p. 57.

[15] E. Henke, E. M. Kaisse, and R. Wright, "Is the sonority sequencing principle an epiphenomenon," in *The Sonority Controversy*, vol. 18, S. Parker Ed.: Walter de Gruyter, 2012, pp. 65-100.

[16] Miyashita, Mizuki. 2011. Five Blackfoot lullabies. *Proceedings of the American Philosophical Society* 155(3): 276–293.

[17] Weber, Natalie. 2020. *Syntax, prosody, and metrical structure in Blackfoot*. PhD dissertation, University of British Columbia.

[18] Frantz, Donald G. & Norma Jean Russell. 2017. *Blackfoot Dictionary of Stems, Roots and Affixes*. Third edition. Toronto: University of Toronto Press.

[19] XXX (citation withheld pending blind review).

[20] Boersma, P., & Weenink, D. 2019. PRAAT: Doing phonetics by computer [Computer program].

[21] R Development Core Team. 2007. R: A language and environment for statistical computing, R Foundation for Statistical Computing.

[22] Gick, B., Bliss, H., Michelson, K., & Radanov, B. 2011. Articulation without acoustics: “Soundless” vowels in Oneida and Blackfoot, Journal of Phonetics. 40: 46-53.

VOICELESS STOPS IN SYLLABLE CODA IN TUSCAN ITALIAN

Giovanna Marotta[1], Piero Cossu[1], Giuditta Avano[1]

[1]Università di Pisa
giovanna.marotta@unipi.it, piero.cossu@fileli.unipi.it, giuditta.avano@phd.unipi.it

ABSTRACT

In Tuscan varieties of Italian, postvocalic stop consonants are affected by a spirantization process (*gorgia toscana*). In current literature, stops have been analysed in syllable onset only, while no studies examined the coda context. Our study examines the phonetic outputs of /k t p/ in coda in order to verify whether the postvocalic position is a sufficient constraint for *gorgia* to apply. Although Italian words derived from Latin do not allow a stop in coda position, items like *cactus* and *taxi* are now well integrated in the Italian lexicon, being frequently produced by speakers. We recorded 24 Florentine speakers and built a corpus of 474 postvocalic voiceless stops occurring in two different contexts: 168 consonants in coda position (e.g. *cactus* "cactus") and 288 occur in onset position (e.g. *oca* "goose"). The results of the acoustic analysis (including duration and intensity of the target segments) show a stop output in coda position for all speakers, providing further evidence for the syllable constraint.

Keywords: Italian, lenition, Tuscan, syllable, acoustics

1. INTRODUCTION

In Tuscan varieties of Italian, stop consonants are affected by a regular weakening process when occurring in postvocalic position, i.e in onset position, given that in the native lexicon stops cannot occupy the coda position (cf. [1; 2; 3; 4; 5; 6; 7]).

The process gives rise to fricative segments and occurs not only in word intervocalic context (e.g. [ˈa:ɸe] *ape* "bee", [ˈdi:θo] *dito* "finger", [ˈɔ:xa] *oca* "goose", [ˈaʊ̯θo] *auto* "car", [ˈbaɪ̯θa] *baita* "cabin") but also when the stop is followed by a liquid or a glide (e.g. [ˈka:ɸra] *capra* "goat", [ˈa:xre] *acre* "acrid", [riˈxja:mo] *richiamo* "recall", [ˈa:xwila] *aquila* "eagle"). In literature this spirantization process is traditionally named *gorgia toscana* (henceforth GT). Voiceless plosives are the preferred target for GT. Amongst Italian speakers, the lenition of /k/ is considered the prototypical feature of Tuscan, since such weakening process mostly affects /k/ with respect to /t/ and /p/ in all the region of Tuscany.

In Italian as well as in Tuscan lexicon derived from Latin, a stop may only be followed by a vowel, a liquid, or a glide. Since all these segments do not block GT, the process appears to be constrained only by the left side of the string. Therefore, the triggering context of GT may simply be defined as postvocalic (cf. [6]; [7]). The lenition process is thus blocked if the stop is either geminated (e.g. [ˈgat:o] *gatto* "cat", [ˈtap:o] *tappo* "cap",) or preceded by another consonant (e.g. [ˈarko] *arco* "arch", [ˈkampo] *campo* "field"). Stops remain stable even in absolute initial position of an utterance, i.e. after a pause.

In some loanwords a voiceless stop may be followed by another stop (e.g. *cactus* "cactus" [ˈkaktus], *captato* "intercept" [kapˈta:θo]), a fricative (e.g. *taxi* "taxi" [ˈtaksi]) or a nasal segment (e.g. *ritmo* "rhythm" [ˈritmo], *acne* "acne" [akne]). In these cases, the stop can be associated with a syllable coda. However, from a general point of view it is possible that the clusters stop + *s* or *n, m* could be interpreted as branching onsets, since they respect the Sequencing Sonority Principle [8]. It has to be underlined that nowadays lexemes like the ones aforementioned are well integrated in the Italian lexicon, since they are frequently produced by speakers in every register of the language, especially for young people.

In current literature on GT, stops have been analysed in syllable onset only, that is with reference to words belonging to the native lexicon derived from Latin. Up to now there are no acoustic studies that have examined the context where the target consonant is followed by an obstruent or a nasal. On the assumption that GT applies when the stop associated with a postvocalic onset position, we may expect that the target consonant will be not affected by GT if it is associated with syllable coda. As we shall see, our acoustic results will show that stops in this specific context do not undergo GT, thus confirming the role played by syllable structure in this lenition process, as already proposed by [6] and [7]: stops are a target for GT only in postvocalic onset position.

2. METHOD

Our study includes 24 Florentine participants (12 females and 12 males) who were born in Florence and have never moved to other cities, and whose parents were born in town as well. Age ranged from 29 to 64 years, whereas education varied from lower-level secondary diploma to master's degree. Speech

material was recorded at 48 kHz 24 bit mono, using a Shure headset microphone, which guaranteed a constant distance of the microphone from the speaker, connected to a Zoom H2next recorder. The participants have been recorded in a quiet room. They were asked to read randomised sentences with target words including voiceless stops (the main target of GT [1; 2],), both in stressed and unstressed syllables.

The global corpus is composed of 474 postvocalic voiceless stops: 168 /k/, 168 /t/ and 138 /p/. 168 consonants occur in coda position: 72 /k/, 72 /t/ and /42 /p/ (e.g. *cactus* "cactus", /ˈkak.tus/, *ritmo* "rhythm" /ˈrit.mo/), whereas 288 occur in onset position (96 for each place of articulation); e.g. *buca* "hole", /ˈbu.ka/; *foto* "photo", /ˈfɔ.to/). The acoustic analysis was performed using PRAAT [9] with the help of webMAUS [10] for supervised alignment. Following [11] and [12] we classified the acoustic outputs of each target segment as follows:

a) **stop** (i.e. [k t p]): silence followed by spike and release phase; see Figure 1;
b) **fricative** (i.e. [x/h θ ɸ]): frication noise through all phone duration without voicing; see Figure 2;
c) **approximant** (i.e. [ɣ/ɦ ð β]): frication noise with voicing (and visible formant structure); see Figure 3;
d) **deletion** of the segment: no acoustic quality typical of the consonant between the preceding vowel and the following segment; see Figure 4;
e) **assimilation** (e.g. /kt/ > [t:]): production of the following consonant as long, i.e. geminated; see Figure 5.

We also found segments in which the frication noise was preceded by a brief hold phase (around 30 ms; e.g. see Figure 6), known in GT literature as semifricatives [4; 7]; these outputs have been included into the class of the fricatives, as they showed clear frication noise.

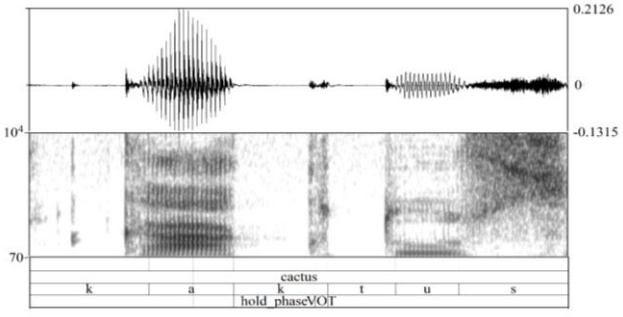


Figure 1: Soundwave and spectrogram of *cactus* [ˈkak.tus], female Florentine speaker, 58 years old.

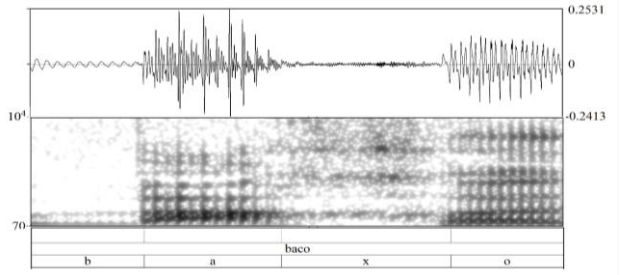


Figure 2: Soundwave and spectrogram of *baco* [ˈba:.xo] "worm"; female Florentine speaker, 49 years old.

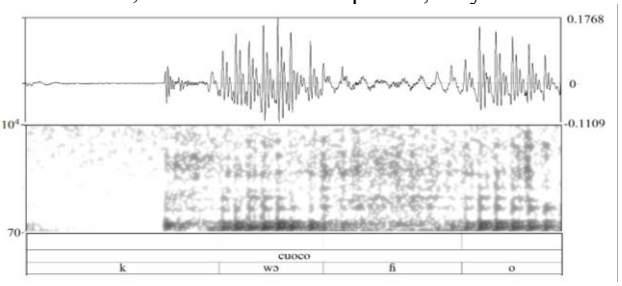


Figure 3: Soundwave and spectrogram of *cuoco* [ˈkwɔ:.ɦo] "cook"; male Florentine speaker, 63 years old.

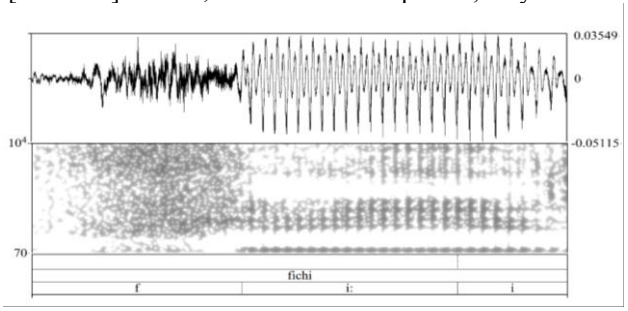


Figure 4: Soundwave and spectrogram of *fichi* [ˈfi:.i] "figs"; female Florentine speaker, 64 years old.

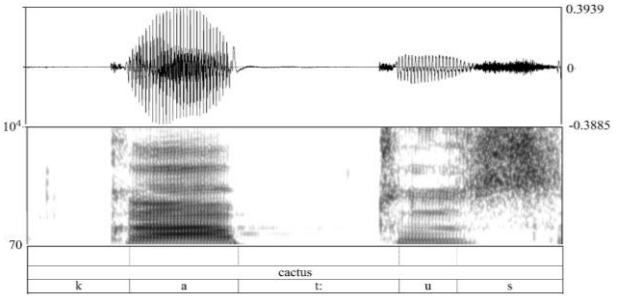


Figure 5: Soundwave and spectrogram of *cactus* "cactus" [ˈkat.tus]; female Florentine speaker, 48 years old.

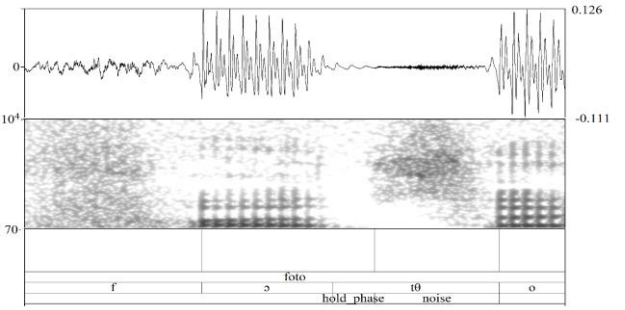


Figure 6: Soundwave and spectrogram of *foto* "photo" [ˈfɔ:.tθo]; female Florentine speaker, 48 years old.

For each token, we measured phone duration (ms) and intensity (dB), two parameters that have been previously found correlating with stop lenition due to GT [4, 5].

3. RESULTS

The different surface outputs with their percentage values results are presented in Table 1. A clear-cut difference in the phonetic outputs of underlying voiceless plosives /p t k/ emerges with crucial reference to the syllable position of the target consonant.

	Stop	Fric	Approx	Delet	Assim
Onset	**20 (7%)**	**200 (69%)**	**61 (21%)**	**7 (2%)**	**0**
/k/	1 (1%)	41 (43%)	47 (49%)	7 (7%)	0
/t/	7 (7%)	86 (90%)	3 (3%)	0	0
/p/	12 (13%)	73 (76%)	11 (11%)	0	0
Coda	**167 (90%)**	**13 (7%)**	**0**	**0**	**6 (3%)**
/k/	65 (90%)	3 (4%)	0	0	4 (6%)
/t/	73 (88%)	9 (12,50%)	0	0	0
/p/	39 (93%)	1 (2%)	0	0	2 (5%)

Table 1: Phonetic outputs of /k t p/ with reference to syllable; Stop = stop, Fric = fricatives, Approx = approximants, Delet = deletion, Assim = assimilation.

Post-vocalic onsets are mostly lenited and surface as continuous segments, i.e. as fricatives and approximants (69% fricatives; 21% approximants; 90% in total). As previously observed in literature ([1] [2] [3]), the most affected segment by GT is the velar plosive /k/, which is even deleted in part of the tokens, with concomitant lengthening of the preceding vowel, like in [ˈfiːi] *fichi* "figs". In general, full closure articulations for post-vocalic onsets amount only to 7% of total tokens. On the contrary, segments in post-vocalic coda show the opposite trend, since 90% tokens show complete closure, whereas continuous segments are only 7% of total phonetic outputs. No approximants or segment deletions are found among the outputs occurring in coda position. Moreover, we observed that the few fricative outputs occur before nasals (/n m/; 8%), and sibilant (/s/; 13%; see Table 2), whereas before a following stop full regressive assimilation may occur, although quite rare (3%); typically /kt/ > [tː].

Acoustic measures relative to phone duration and intensity are in line with previous studies ([4; 5]; see Table 3). In onset position, going from stops and fricatives to approximants, a decrease in duration and an increase in intensity is observed, regardless of the place of articulation.

	Stop	Fric	Approx	Delet	Assim	Total
Coda	**167 (90%)**	**13 (7%)**	**0**	**0**	**6 (3%)**	**186 (100%)**
/C.N/	99 (92%)	9 (8%)	0	0	0	108 (100%)
./C.t/	42 88%	0	0	0	6 12%	48 (100%)
/C.s/	26 (87%)	4 (13%)	0	0	0,00%	30 (100%)

Table 2: Phonetic outputs of /k t p/ (C) in coda position with reference to the following segment; Stop = stop, Fric = fricatives, Approx. = approximants, Delet = deletion, Assim = assimilation

In coda position, a progressive decrease in duration is found going from stops to fricatives: the mean duration of the stops is 119 ms, whereas that of fricatives is 93 ms. The mean value of intensity in coda is similar for the two phonetic outputs: 46 dB for stops; 45 dB for fricatives; however, since only 13 tokens (7% of the outputs) are fricatives (see Table 1), a proper comparison is not possible. The comparison of the results of the acoustic analysis shows a longer duration of the segments in coda position compared to those in onset position. In particular, coda stops are 40 ms longer than onset stops and coda fricatives are 20 ms longer than onset fricatives. The intensity shows similar values for stops and fricatives in both contexts (from 45 to 48 dB), whilst the approximant outputs are louder (58 dB on average).

		Stop Mean (sd)	Fric Mean (sd)	Appr Mean (sd)
Onset				
	Duration (ms)	**82 (11)**	**77 (17)**	**51 (18)**
	Intensity (dB)	**47 (9)**	**48 (6)**	**58 (6)**
/k/	Duration (ms)	84	77 (18)	50 (19)
	Intensity (db)	45	48 (6)	60 (5)
/t/	Duration (ms)	77 (11)	76 (18)	62 (8)
	Intensity (dB)	46 (5)	48 (6)	51 (4)
/p/	Duration (ms)	85 (10)	77 (14)	54 (9)
	Intensity (db)	47 (11)	49 (7)	53 (6)
Coda				
	Duration (ms)	**118 (35)**	**93 (23)**	
	Intensity (dB)	**46 (8)**	**45 (5)**	
/k/	Duration (ms)	119 (45)	78 (18)	
	Intensity (dB)	45 (6)	43 (4)	
/t/	Duration (ms)	125 (30)	102 (19)	
	Intensity (dB)	45 (8)	45 (4)	
/p/	Duration (ms)	106 (20)	56	
	Intensity (dB)	48 (9)	53	

Table 3: Duration and intensity (mean and standard deviation) of phonetic outputs with reference to syllable position and place of articulation; Stop = stop, Fric = fricatives, Approx = approximants.

We fitted a linear mixed model created with the R (version 4.3.0) [13] function *lmer()* of the package *lme4* [14] to test if the syllable position interacts with the manner of articulation, interpreted as continuous (i.e. fricatives, approximants and deletion) and non-continuous (i.e. stops). Speaker, gender, word, and place of articulation were set as random effects. The results show that the effect of the syllable constituent (onset vs. coda) is statistically significant ($\chi^2 = 63,7$, $p < 0.001$). Therefore, our study provides further evidence in favour of the syllable constraint on GT.

4. DISCUSSION

GT is traditionally described as a phenomenon affecting intervocalic stops, especially if voiceless ([1; 2; 3]). However, since GT applies also before liquid consonants and glides (cf. § 1. Introduction), the context of application of GT is best described as postvocalic ([4; 7]). With reference to syllable structure (see [15; 16]), GT can be interpreted as a lenition process affecting any postvocalic stop segment associated with the onset constituent. If this interpretation is correct, it follows that a stop in coda position does not undergo GT. In order to verify such prediction, we investigated the acoustic output of voiceless stops in coda position produced by Florentine speakers. The results collected show that stops in this context are not a target for GT, thus confirming the hypothesis that GT affects stop segments only if associated to the onset position.

However, we also found some fricative outputs when the stop is followed by a nasal consonant or /s/. These data seem to be compatible with a possible tautosyllabic interpretation of the clusters stop + *s* or *n, m*, i.e. as complex onsets. Indeed, such clusters obey the *Sequencing Sonority Principle* ([8]), being the second segment weaker than the preceding one; in parallel, the degree of sonority increases from the plosive consonant to the nasal as well as to the sibilant. Moreover, this view agrees with the data concerning the selection of the definite article in Italian [17]: before words beginning with the clusters stop + *s* or *n, m* Tuscan speakers optionally select the allomorph *il*; e.g. *il psicologo* "the psychologist", *il pneumatico* "the tire", unlike standard Italian where the allomorph *lo* is preferred. The allomorph *il* usually occurs before onsets, both simple and complex (i.e. branching), both in Standard Italian and Tuscan varieties; e.g. *il ponte* "the bridge", *il prato* "the meadow", *il toro* "the bull", *il treno* "the train".

In conclusion, our study is the first one investigating the phonetic output of stops in the context of syllable coda in Tuscan Italian. The results collected show that stops in coda do not lenite, except in a few cases. Although the data are consistent and statistically significant, we believe that further studies are needed to verify the consistency of our empirical data. We planned to extend the analysis to a larger sample of speakers as well as to other areas of Tuscany.

The data presented highlight the relevance of the syllable structure in the process of Tuscan lenition. Three conditions must be satisfied in order for GT to apply. Firstly, the target segment has to be a stop. Secondly, the stop must be preceded by a vocalic segment (vowel or glide), since after a consonant GT does not apply. Thirdly, the stop has to be associated with the syllabic onset position; more precisely, the segment needs to be the head of an onset, since we know that the first segment of a *muta cum liquida* is also affected by lenition, mirroring simple onset contexts (e.g. [ˈlaːxrima] *lacrima* "tear", [ˈliːθro] *litro* "liter", [ˈkaːɸra] *capra* "goat"). The occurrence of a lenition process in syllabic onsets is not surprising; as a matter of fact, many languages show similar phenomena in the same context. In particular, lenition is a well-documented process in Romance languages (see for instance [18]).

On the other hand, coda position is traditionally considered a weak context although segments in syllable coda are affected by different weakening processes as compared to onsets ([19]). We might expect plosives to become spirants in coda position too, via GT. As this is not the case, it is possible that Tuscan speakers preserve stops in coda position because of the constraints of their native lexicon that do not allow lenition in this context. GT is a phonological process acting on the native lexicon which may not apply to new items belonging to peripheral or more recent areas of the lexicon. However, even though recent loanwords were not analysed in this study, our corpus also includes some English loans recently entered in the lexical competence of Italian speakers including a postvocalic onset (e.g. *meeting, Microsoft*).

According to a preliminary analysis, it appears that GT regularly applies also to these items, if the conditions are met (e.g. [ˈmaɨ̯xrosoft] *Microsoft*, [ˈmiːθiŋg] *meeting*). It is also possible that the particular kind of data elicitation we adopted (i.e. reading) influenced the participants, encouraging a more standard production, with stricter correspondence to the orthographic input. Future investigations will also be aimed at analysing spontaneous speech, in order to verify the hypotheses advanced here with the support of more exhaustive data.

5. REFERENCES

[1] Giannelli, L., Savoia, L.M. 1978. L'indebolimento consonantico in Toscana, I. *Rivista Italiana di Dialettologia* 2, 25-58.

[2] Giannelli, L., Savoia, L.M. 1979-80. L'indebolimento consonantico in Toscana, II. *Rivista Italiana di Dialettologia* 3–4, 39–101.

[3] Giannelli, L. 1997. Tuscany. In: Maiden, M., Parry M. (eds), *The Dialects of Italy*, Routledge, 297-302.

[4] Marotta, G. 2001. Non solo spiranti: La Gorgia Toscana nel parlato di Pisa. *L'Italia Dialettale* 62, 27–60.

[5] Sorianello, P. 2001. Un'analisi acustica della Gorgia fiorentina. *L'Italia Dialettale* 62, 61-94.

[6] Bafile, L. 1997. La spirantizzazione toscana nell'ambito della teoria degli elementi. In: Catagnoti A. (eds), *Studi linguistici offerti a G. Giacomelli dagli amici e dagli allievi*, Unipress, 27-38.

[7] Marotta, G. 2008. Lenition in Tuscan Italian (*Gorgia Toscana*). In: Brandao de Carvalho, P., Scheer, T., Ségéral, Ph. (eds), *Lenition and fortition*, Mouton-DeGruyter, 235–272.

[8] Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M.E. (eds), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*. Cambridge University Press. 283-333.

[9] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program] Version 6.2.17, retrieved 23 August 2022 from https://www.praat.org.

[10] Schiel, F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech, *Proc. 14th ICPhS 1999*. San Francisco, 607-610.

[11] Stevens, K. N. 1999. *Acoustic phonetics*. MIT Press.

[12] Sorianello, P. 2003. Spectral characteristics of voiceless fricative consonants in Florentine Italian. *Proc. 15th ICPhS, Barcelona,* 3081–3084.

[13] RStudio Team. 2020. *RStudio: Integrated Development for R. RStudio*, PBC, Boston, MA URL http://www.rstudio.com/

[14] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. *Fitting linear mixedeffects models using LME4*, *Journal of Statistical Software*, 67, 1–48.

[15] Goldsmith, J. 1990. *Autosegmental and metrical phonology*. Blackwell.

[16] Kenstowicz, M. 1994. *Phonology in Generative Grammar*. Blackwell.

[17] Marotta, G. 1993. Selezione dell'articolo e sillaba in italiano: un'interazione totale? *Studi di Grammatica Italiana* XV, 255-296.

[18] Brandao de Carvalho, P., Scheer, T., Ségéral, Ph. 2008. *Lenition and fortition*, Mouton-DeGruyter.

[19] Ségéral, Ph., Scheer, T. 2001. La Coda-Miroir. *Bulletin de la Société de Linguistique de Paris* 96, 107-152.

INTRUSIVE VOCOIDS AND SYLLABLE STRUCTURE IN GEORGIAN

Caroline Crouch[1], Ioana Chitoran[2], Louis Goldstein[3], Argyro Katsika[4]

[1]Rice University, [2]Université Paris Cité, [3]University of Southern California, [4]University of California, Santa Barbara
cc185@rice.edu; ioana.chitoran@u-paris.fr; louisgol@usc.edu; argyro@ucsb.edu

ABSTRACT

Complex onsets in Georgian are characterized by long lags between consonant constrictions, which differ depending on the sonority shape of the onset cluster. These lags result in relatively open vocal tract configurations where vocoids sometimes intrude. In this study, we examine the phonetic properties of such vocoids and present evidence that, although they emerge from the cluster's articulatory configuration, they may create conditions for resyllabification. Specifically, results show that, although the distribution of vocoids depends on sonority shape (more appear in rises, fewer in falls), their formant structure and duration remain stable and unaffected by the cluster's sonority and the nucleus vowel's quality. They are shorter in duration than Georgian lexical vowels. Yet, preliminary analyses show that vocoids affect the syllable-scale structure of the amplitude envelope in ways that could lead to resyllabification in sonority falls. Implications for Georgian syllabic structure are discussed.

Keywords: vocoids, sonority, syllabic structure, resyllabification, Georgian

1. INTRODUCTION

Georgian syllable structure is typologically unusual: onsets can be up to seven consonants in length, with minimal restrictions on consonant combinations. Complex syllable onsets in Georgian often include transitional, schwa-like elements in between consonants. Word onset clusters are also characterized by long lags between the constrictions of the consonants [1][2]. Interestingly, the duration of these lags differs according to sonority shape, as does the distribution of vocoids [1]. Sonority rises (e.g., /br/) have the longest lags and the highest percentage of vocoid occurrence, while plateaus (e.g., /bg/) have shorter lags, and fewer vocoids, and falls (e.g., /rb/) have the shortest lags and fewest reported vocoids.

The primary goal of this study is to determine whether these vocoids are epenthetic vowels or rather are artifacts of the vocal tract configuration during the open transition between consonants, i.e., during the lag between consonant constrictions. For this purpose, we compare the vocoids to phonological vowels acoustically, in terms of formant structure and acoustic duration. In parallel, we show kinematic data that illustrate the position of major articulators during the open transitions between consonants.

Our second goal is to explore syllable-level motivations for the sonority shape-based differences in the distribution of vocoids, and consequently in the duration of the lag between consonants as well. Syllable-level motivations originate in previous research [1][3], which has shown that constriction lag occurs in all clusters in Georgian, regardless of sonority shape or place of articulation [1][3] (see also [4] on place of articulation). These patterns thus strongly suggest that segment-level perceptual concerns do not motivate differences in degree of lag, as previously proposed [4]. Instead a different proposal is put forward in [1][3], according to which the modulation of consonant-consonant timing in Georgian is due to syllable-level planning: Speakers plan for syllables to have a single, vocalic nucleus, and in onsets with initial approximants (i.e., sonority falls) the sequence of approximant+vocoid could result in an additional large peak in the amplitude envelope that could be reinterpreted as a syllable nucleus, and could lead to syllable misparsing. Here we take a first step towards assessing this hypothesis more directly by examining the structure of the syllable-scale amplitude envelope across syllables with and without audible vocoids, and of all sonority shapes. We use the amplitude envelope, because it has been shown to relate to speech comprehension [5], and thus differences in the amplitude envelope in vocoid-containing syllables of different sonority shapes could reflect different possible syllabification parsings by the listener.

2. METHODS

2.1. Participants and procedure

Three native speakers of Georgian in their twenties (S1, S2, and S3) participated in the experiment. They were recruited through an announcement circulated by a local Georgian cultural association.

The data analysed here were collected as part of a larger series of Electromagnetic Articulography (EMA) experiments that included the simultaneous collection of audio data. EMA sensors were attached at three points along the midsagittal line of the

tongue: one on the tongue tip (TT), one on the tongue dorsum (TD), and one on the midpoint between TT and TD, referred to as TB. Two sensors were attached on the lips: one on the upper lip and one on the lower. The Euclidean distance between these sensors at each timepoint was calculated and the resulting variable—Lip Aperture (LA)—was used for labial segments. Sensors were also attached on the upper and lower incisor for reference and jaw movement respectively, and on the bridge of the nose and behind each ear for head correction. Audio was recorded with a Shure SCM262 microphone mixer at a 16 kHz sampling rate, with a Sennheiser shotgun microphone positioned a foot away from the participant's mouth. Kinematic data was automatically synchronized with the external audio data.

2.2. Stimuli

Stimuli were presented in Georgian orthography on a computer screen approximately four to five feet from the EMA. Test words appeared first in isolation and then in a frame sentence, without the word in isolation disappearing from the screen. Two frame sentences were used. The sentence was changed following S1's participation in order to facilitate kinematic labelling.

(1) _____. kal-ma _____ mo-m-ts'er-a. [S2 & S3]
_____. woman.ERG ____ PRVB-1SG.OBJ-write-3SG.SUBJ.AOR
_____. The woman wrote _____ to me.

(2) _____. k'idev _____ v-tkv-i. [S1]
_____. again ____ 1SG.SUBJ-say-AOR.
_____. I said _____ again.

Tables 1 and 2 show the test words analysed in the current study. As with the frame sentence, the test words were refined and expanded following S1's participation. Test words manipulate both sonority shape and order of place of articulation of the consonants in the onset. Order of place of articulation was included as a factor because it has been demonstrated to affect gestural overlap or lag in Georgian [4].

	Front to back	Back to front
Sonority rise	**br**elo 'chaff' **p'l**edi 'rug'	**tm**aze 'hair.in' **dm**anisi 'Dmanisi (town in Georgia)'
Sonority plateau	**mn**axe 'see me' **bk'**ich'i 'raisin' **bg**era 'sound'	**t'b**aze 'lake.in' **k'b**ili 'tooth' **k'b**ena 'sting'
Sonority fall	**mt**aze 'mountain.in' **md**are 'worthless'	**rb**ena 'running (n)' **lp'**eba 'decaying (n)'

Table 1: Test words for S1

	Front to back	Back to front
Sonority rise	**br**egi 'mound' **bn**eli 'darkness' **bn**eda 'epilepsy' **bl**epi 'bluff' **ml**ode 'waiting'	**tm**aze 'hair.in' **dm**anisi 'Dmanisi (town in Georgia)' **gr**evi 'gift' **gl**exi 'peasant'
Sonority plateau	**ʃx**ama 'poison' **bg**era 'sound' **pt**ila 'hair lock' **mn**axe 'see me'	**xʃ**avs 'shut off' **gd**eba 'lie about' **tb**eba 'warm up'
Sonority fall	**mt**aze 'mountain.in' **md**are 'worthless' **rg**eba 'benefitting'	**rb**ena 'running (n)' **lb**eba 'softening (n)' **lm**oba 'feeling sadness'

Table 2: Test words for S2 and S3

Stimuli were randomized and included test words for other experiments alongside those reported here. S1 and S2 recorded each word eight times, while S3 recorded 7 due to time constraints. The underlined words are those that were analysed for the current study. They were selected because a) at least one production had an intrusive vocoid and b) because the vowel following the cluster of interest (bolded in Tables 1 and 2) is either /e/ or /a/. The other phonological vowels of Georgian (/o/, /i/, /u/) are far less frequent in the test words, so in the interest of having a balanced data set, words with those vowels were excluded. Only productions with vocoids were considered in this study. There were a total of 240 vocoid-vowel tautosyllabic pairs across all words, speakers, and phrasal positions (isolation vs. quotative). 58 vocoids precede /a/ and 82 precede /e/.

2.3. Analysis

2.3.1. Acoustic analysis

All vocoids and vowels were segmented in Praat [6]. We define a post-release vocalism as a vocoid, if it has both a voicing bar and a distinct formant structure (see Fig. 3). Onset and offset of both vocoids and vowels were determined by the presence of f2. Formant values at the midpoint of the vowel or vocoid, and acoustic duration, were automatically extracted using a Praat script [7]. Values were checked by hand and corrected where necessary.

2.3.2 Articulatory analysis

Articulatory analysis of consonant (C) constriction lag was done semi-automatically using Mview (Mark Tiede, Haskins Laboratories). EMA sensor trajectories were used to label the C constrictions for each onset cluster. Lip aperture was used for labial consonants, tongue tip vertical displacement for coronal, and tongue dorsum displacement for dorsal. Relevant timepoints for the constrictions were labelled using velocity criteria as shown in Figure 1.

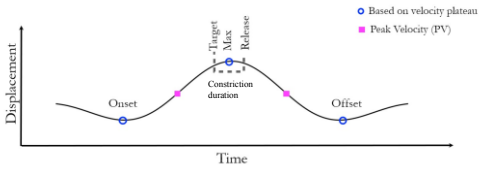


Figure 1: Timepoints labelled for each gesture

The lag between C constrictions is calculated as the time between the first consonant's release and the second consonant's target.

2.3.3 Amplitude envelope analysis

The amplitude envelope was extracted following Tilsen and Arvaniti [8] using a Matlab script [9].

2.3.4 Statistical analysis

All statistical analysis was done in R [10] with linear mixed effects models [11]. Minimal adequate fixed effects structures were determined using drop1 [10] and minimal adequate random effects structures using rePCA [12]. Post-hoc pairwise comparisons with a Holm correction were done using emmeans [13]. Graphics were made with ggplot2 [14].

3. RESULTS

3.1 Distribution

The distribution of vocoids in the data are not random. Of the 240 vocoids measured, 149 appear in sonority rises, 62 in sonority plateaus, and 29 in falls.

3.2 Duration

Vocoids are significantly shorter than vowels ($p<.001$). Mean vocoid duration overall is 36.6ms, while mean vowel duration overall is 118.3ms Phrasal position is also highly significant, with both vocoids and vowels produced in the isolation condition being longer than those in the quotative condition ($p<.001$). Neither sonority shape nor order of place of articulation have an effect on vocoid or vowel duration.

Sonority shape does, however, significantly affect the duration of the lag between consonant constrictions [2]. Sonority falls have shorter lags than rises ($p<.001$). Across all sonority conditions, lags are on average 100ms, nearly equivalent to the duration of a full vowel. As with vocoid duration, phrasal position significantly affects lag: lags are longer in the isolation condition ($p<.001$). However, constriction lag duration and vocoid duration are not correlated.

Figure 3 shows the spectrogram and EMA sensor trajectories for /bneda/ with an intrusive vocoid. The colored boxes mark the gestures for /b/, /n/, /e/. The filled boxes represent the constriction plateaus. The horizontal arrow shows the lag between consonant constrictions, while the dotted vertical lines delimit the vocoid. We can see from this example that not only does the vocoid occupy only a fraction of the lag, but also that the tongue dorsum (TD) is not moving towards an articulatory target during the vocoid, instead it is in the process of moving along with the tongue tip (TT) to the following /n/ target.

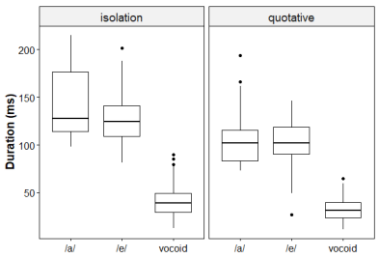


Figure 2: Durations (ms) for vocoids and vowels in the isolation (left) and quotative (right) conditions.

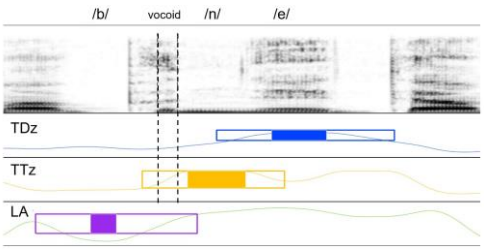


Figure 3: Spectrogram and sensor trajectories for /bneda/, with vocoid (dotted lines) and constriction lag (arrow) marked. TDz stands for tongue dorsum vertical displacement; TTz for tongue tip vertical displacement; LA for lip aperture. Boxes on each trajectory mark the target (filled box) and movement into and out of the target (empty boxes).

3.3 Formants

Vocoids are distinct from both /a/ and /e/ in both first and second formants ($p<.001$). Moreover, the quality of the following vowel does not affect the formant values of the vocoid. Figure 4 shows the f1 by f2 plot. Three clouds are visible: one for the /a/ vowel, one for

the /e/ vowels and one for the vocoids, regardless of the quality of the following vowel.

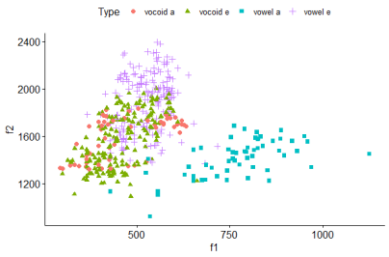


Figure 4: f1 and f2 values for /a/, /e/, and vocoids preceding either /a/ or /e/.

3.4 Amplitude envelope

Figure 5 shows the waveform and corresponding amplitude envelope per sonority shape.

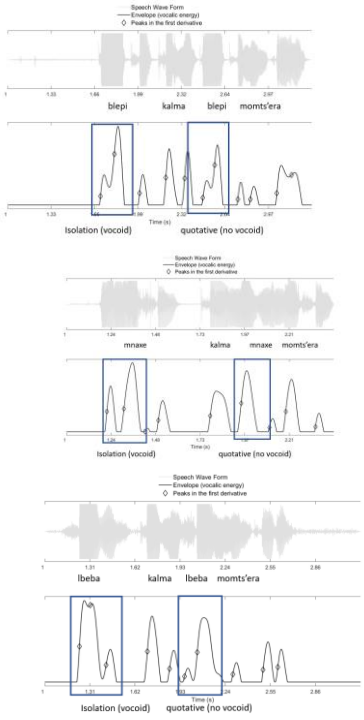


Figure 5: Waveform and amplitude envelope for sonority rise /blepi/ (5a), sonority plateau /mnaxe/ (5b), and sonority fall /lbeba/ (5c) with and without a vocoid, as produced by S2. Regions of interest are outlined; the box on the left of each pair marks the production with a vocoid, and the box on the right marks the one without.

In these examples, isolation productions have a vocoid, while the quotative ones do not. In sonority rises (Fig. 5a), the vocoid and non-vocoid productions have the same amplitude envelope shape, with the same number of amplitude peaks identified. The vocoid peak is smaller in magnitude and peak velocity than the vowel peak. In a sonority plateau like /mnaxe/, we see different shapes in the envelope for the vocoid and non-vocoid productions (Fig. 5b). The production with a vocoid has two distinct peaks, while the non-vocoid production has one. We see something similar for the sonority fall /lbeba/ in Figure 5c. The production with the vocoid has three peaks in the first derivative identified, while the non-vocoid production has only two. The first peak in the vocoid production, which coincides with the vocoid itself, is greater in magnitude than many lexical vowels in the utterance.

4. DISCUSSION

These results indicate that intrusive vocoids in Georgian are articulatory artifacts that emerge due to the long lag between consonant constrictions. According to [3], this lag is created because C constriction gestures in Georgian onsets are in anti-phase coordination with each other, with the prenuclear C gesture also anti-phase with the nucleus V gesture. This means that each successive gesture is initiated only after the previous gesture has reached its target. As articulators move from one C to the other, a relatively neutral configuration is created in the vocal tract. When the second consonant is voiced, voicing may occur during some part of this transition resulting in the vocoid, as illustrated in Fig. 3. Since the overlap between lag and voicing is brief and the oral tract is target-less during this interval—hence the non-effect of sonority on vocoid quality—the resulting vocoid has a schwa-like quality (see Fig. 4; F1 ~ 500 Hz and F2 ~ 1500 Hz for the vocoids). The fact that the quality of the vocoid is unaffected by the following vowel is consistent with the account of the nucleus V gesture being anti-phase with the onset. If the vowel gesture were in-phase with the onset, the lag between consonant gestures would briefly 'unmask' the V gesture, and the formant structure of the resultant vocoid would be expected to be affected by the quality of the phonological vowel.

Sonority shape does affect the duration of the lag and the distribution of vocoids. It has been hypothesized that this may avoid resyllabification [3]. Preliminary examination of the syllable-scale amplitude envelope supports this hypothesis, since it shows that the presence of a vocoid in a sonority plateau or fall (Figures 5b-c) results in a distinct additional peak, which could be interpreted as an additional syllable. Conversely, in a sonority rise, the vocoid is less distinct from the nuclear amplitude peak (Figure 5a). The distributional asymmetry of intrusive vocoids across onsets of different sonority shapes thus mirrors the potential disruptiveness of the vocoid. In those onsets where a vocoid may result in resyllabification, we find fewer vocoids and shorter lags between consonants, as speakers plan for a syllable with a single nucleus.

5. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (#1551428) to Argyro Katsika (PI) and National Institutes of Health (NIDCD-DC-002717) to Douglas Whalen (PI).

6. REFERENCES

[1] C. Crouch, A. Katsika, and I. Chitoran, "Sonority sequencing and its relationship to articulatory timing in Georgian," Journal of the International Phonetic Association, pp. 1–24, 2023.

[2] M. Pouplier et al., "Language and cluster-specific effects in the timing of onset consonant sequences in seven languages," Journal of Phonetics, vol. 93, p. 101153, Jul. 2022, doi: 10.1016/j.wocn.2022.101153.

[5] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," Science Advances, vol. 5, no. 11, p. eaay6279, Nov. 2019, doi: 10.1126/sciadv.aay6279.

[3] C. Crouch, "Postcards from the syllable edge: sonority and articulatory timing in complex onsets in Georgian," UC Santa Barbara, 2022. Accessed: Jan. 04, 2023. [Online]. Available: https://escholarship.org/uc/item/5w18167d

[4] I. Chitoran, L. Goldstein, and D. Byrd, "Gestural overlap and recoverability: Articulatory evidence from Georgian," Laboratory phonology, vol. 7, no. 4–1, pp. 419–447, 2002.

[6] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.3.03, http://www.praat.org/

[7] Lennes, Mietta, "SpeCT - Speech Corpus Toolkit for Praat (v1.0.0). First release on GitHub". Zenodo, Mar. 08, 2017. doi: 10.5281/zenodo.375923.

[8] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," The Journal of the Acoustical Society of America, vol. 134, no. 1, pp. 628–639, 2013.

[9] A. MacIntyre, "AcousticLandmarks" Accessed: 4 January 2023 [Matlab script] https://github.com/alexisdmacintyre/AcousticLandmarks

[10] R Core Team, "R: A language and environment for statistical computing." R foundation for statistical computing Vienna, Austria, 2018.

[11] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," Journal of Statistical Software, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[13]R. Lenth, J. Love, and M. Herve, "emmeans: Estimated Marginal Means, aka Least-Squares Means." Jan. 10, 2018. Accessed: Jan. 04, 2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[14] H. Wickham, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

TEMPORAL ORGANIZATION OF WORD-INITIAL CONSONANT SEQUENCES IN NAJDI ARABIC

Omar A. Alkhonini, Harim Kwon

Majmaah University, Seoul National University
om.alkhonini@mu.edu.sa, harimkwon@snu.ac.kr

ABSTRACT

This study investigated the syllabic affiliations of Najdi Arabic word-initial consonant sequences (#CC), by examining the temporal relation among segments using acoustic measures. Using minimal pairs: #C_1C_2VX vs. #C_2VX produced by Eight Najdi Arabic speakers, two intervals were compared: C-CENTER (from the mean of the midpoint(s) of C(s) to V offset) and RIGHT-EDGE (from the release of immediately prevocalic C to V offset). Results demonstrated RIGHT-EDGE stability. While C-CENTER was significantly longer in #CCVX than in #CVX, RIGHT-EDGE was not. Additionally, the durational variance was smaller in RIGHT-EDGE than in C-CENTER. These suggest #CC is heterosyllbically parsed in Najdi Arabic, corroborating the claim that Najdi Arabic is restricted regarding word-initial tautosyllabic clusters. Najdi Arabic may not allow word-initial tautosyllabic consonant clusters.

Keywords: Najdi Arabic, Syllable, Word-initial Consonant Sequence, Temporal Organization.

1. INTRODUCTION

1.1. Background

Syllabic parsing of word-initial consonant sequences can be informed by the component consonants' temporal patterns (e.g., [1, 2]). Syllable onset consonants are temporally coordinated with the rest of the syllable such as the rime. That is, when a word-initial sequence of multiple consonants forms a tautosyllabic onset cluster (i.e., #CCVX), all component consonants are temporally coordinated relative to the rime. On the other hand, when a sequence of consonants is heterosyllabically parsed (i.e., #C.CVX), the consonant(s) not forming the syllabic onset would not have a stable temporal relation with the rime.

The relation between the syllabic organization and the temporal patterns has been examined in different languages. For example, English has been claimed to have complex onsets, or onset clusters, because the intervals between the c-center (the mean of the midpoints of each consonant in consonant sequences) to the anchor (the end of the syllabic nucleus) is more stable than those between the right-edge (the release of the immediately prevocalic consonant) to the same anchor (e.g., [1]). The c-center to anchor interval remains stable because each consonant in the consonant sequence is coordinated with the rest of the syllable structure. On the other hand, if the right-edge to anchor interval is more stable than the c-center to anchor interval, it suggests that only the rightmost consonant in the sequence forms the syllabic onset with the preceding consonant(s) being parsed heterosyllabically (i.e., #C.CVX). This pattern has been reported in Moroccan Arabic [3, 4], Tashlhiyt Berber [5], and Jazani Arabic [6].

While most of these previous studies [1, 2, 3, 4, 5] rely on articulatory measures to investigate these temporal patterns, acoustic measures have also been shown to exhibit consistent temporal patterns in [6], opening the possibility of using acoustic measures to examine syllabic affiliations of consonant sequences.

1.2. Current Study

This study investigates the syllabic affiliations of Najdi Arabic word-initial consonant sequences (#CC) using acoustic measures. Najdi, an understudied variety of Arabic spoken mainly in the center of Saudi Arabia, has been described as having onset consonant clusters (e.g., [7]). However, the temporal patterns of the component consonants have not yet been examined.

It has recently been claimed that Najdi may be very limited in terms of onset consonant clusters, based on Najdi speakers' production patterns [8]. Najdi speakers in [8] produce prothetic vowels before word-initial C_1C_2 when C_1 is more sonorous than, or as sonorous as, C_2. That is, #CCVX with a falling or plateau sonority profile are realized as [#vC.CVX] while those with a rising sonority profile would surface as [#CCVX]. These findings, according to [8], provide evidence that Najdi allows only the onset clusters with a rising sonority profile but prohibits those with a less preferred sonority profile, consistent with the predictions of the Sonority Sequencing Principle (e.g., [9, 10, 11]). However, it remains unclear if Najdi #CC sequences surfacing without a prothetic vowel are tautosyllabic onset clusters.

The present study aims to determine the syllabic affiliations of the component consonants of Najdi #CC sequences with rising sonority profiles. To that end, we investigate the temporal patterns of consonant sequences that surface with no simplifications (i.e., a vowel prothesis) to determine whether their syllabification is tautosyllabic [#CCVX] or heterosyllabic [#C.CVX]. The main research question of this study is the following: When word-initial consonant sequences surface without prothesis in Najdi Arabic, are they parsed as tautosyllabic CC onset clusters?

2. METHODS

2.1. Participants

Eight native speakers (3F, 5M, mean age = 26) of Najdi Arabic, from the central part of Saudi Arabia (Riyadh and surrounding towns), participated in the study in Az-Zulfi, Saudi Arabia. They all reported speaking the Modern Standard Arabic, but no other dialects of Arabic. The participants also reported speaking English as the second language.

2.2. Stimuli

The stimuli, shown in Table 1, consisted of seven minimal pairs differing in the number of initial consonants: #C_1C_2VX vs. #C_2VX.

#C_1C_2VX	Gloss	#C_2VX	Gloss
ʃbuk	fences	buk	wallet
bzur	children	zur	fake
ʃmus	very hot	mus	razor
snun	teeth	nun	letter “N”
dru:s	lessons	ru:s	heads
flan	someone	lan	become soft
ħraseh	protecting	raseh	his head

Table 1: 14 target words in the stimuli

All #C_1C_2 sequences were of rising sonority profile with varying sonority distances between C_1 and C_2. The current study adopts the scale proposed by Gouskova [12] for its thorough display of the categories.

2.3. Tasks and Procedures

The participants produced the target words embedded in two different carrier phrases. In one of the carrier phrases, the target word was preceded by a vowel /i/ (e.g., [haði ___ mar:ah] “This is ___ once.”), and in the other one, by a consonant /k/ (e.g., [ðik ___ mar:ah] “That is ___ once.”). These carrier phrases were used to examine if the preceding segment would influence the syllabification. A preceding vowel can possibly promote a heterosyllabic parsing of #CCVX (e.g., [ði#C.CVX]) compared to a preceding consonant (e.g., [ðik.#CCVX]).

Each speaker repeated the target words 7 times in each carrier phrase, and their speech was acoustically recorded.

2.4. Measurements and Analysis

To determine the syllabification of consonant sequences, we used acoustic measures that were similar to the ones used by previous studies (e.g., [3, 4, 6]). Each target word #(C_1)C_2VX yielded three acoustic landmarks: c-center (the mean of acoustically measured midpoints of C_1 and C_2), right-edge (the acoustic release of C_2) and the anchor (the acoustic onset of postvocalic X). From these landmarks, we obtained two interval measures: c-center to anchor (henceforth C-CENTER) and right-edge to anchor (henceforth RIGHT-EDGE).

The two interval measures, C-CENTER and RIGHT-EDGE, were then examined to determine which was more stable. The stability was determined based on two different analyses. First, duration analysis compared the interval measures between #CCVX and #CVX words. If an interval measure remains constant when a consonant is added, it is considered stable. For example, if /dru:s/ and /ru:s/ are more similar in their C-CENTER measures than in the RIGHT-EDGE measures, C-CENTER is considered to be more stable than RIGHT-EDGE. Second, variance analysis compared the durational variances of the two interval measures, using the relative standard deviation (RSD). The interval measure with smaller durational variances is considered to be more stable. The RSD is more conservative measure than the more commonly used standard deviation in our case, as it is less inclined toward the RIGHT-EDGE stability (see [3, 4] for more information on RSD calculation).

For annotation, the boundaries for words and segments were initially determined using the Montreal forced aligner [13], and then hand-corrected in Praat [14].

2.5. Predictions

The most stable interval would indicate the syllabification of the word-initial CC sequences in Najdi Arabic. If #CC sequences are tautosyllabic and form true onset clusters (e.g., [7]), it is predicted that C-CENTER would be more stable than RIGHT-EDGE. On the other hand, if Najdi #CC sequences are heterosyllabically parsed (#C.CVX), as suggested in [8], RIGHT-EDGE is predicted to be more stable than C-CENTER.

3. RESULTS

3.1. Duration analysis

Figure 1 shows the difference between C-CENTER and RIGHT-EDGE measures across #CCVX and #CVX words. The mean C-CENTER interval was 229.2 ms for #CCVX and 185.2 ms for #CVX, while the mean RIGHT-EDGE interval was 140.5 ms for #CCVX and 142.2 ms for #CVX.

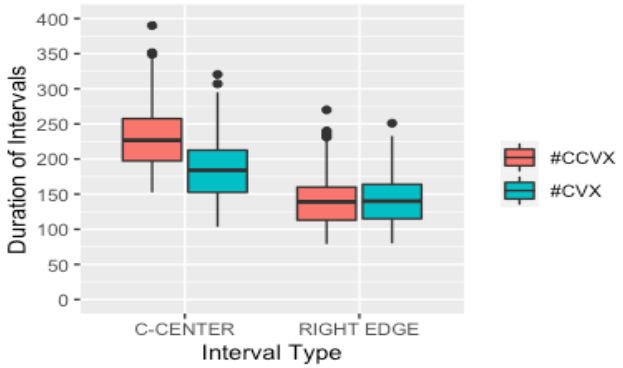


Figure 1: C-CENTER vs. RIGHT-EDGE by word size

To examine if the observed differences in duration are statistically significant, two logit mixed-effects model analyses [15] were performed in R [16] with the lme4 package [17], one for C-CENTER and the other for RIGHT-EDGE. The dependent variable for these models was the respective interval (C-CENTER vs. RIGHT-EDGE). Each model included word size (#CCVX vs. #CVX) and environment ("target preceded by a consonant" vs. "target preceded by a vowel"), and their interactions as fixed effects. The factors were contrast coded. Participants and items were included as random effects with by-items random slopes for environment. Including more random slopes led to convergence errors. P-values were obtained by comparing the full models with the models without the effects in question.

Model comparisons for C-CENTER revealed a significant effect of word size [$\chi^2(1) = 14.48$, $p < .001$]: #CCVX were longer than #CVX. However, environment [$\chi^2(1) = .002$, $p = .97$] and the interaction between environment and word size [$\chi^2(1) = 1.02$, $p = .31$] were not significant. These outcomes suggest the C-CENTER duration was influenced by word size, not the environment or the interaction between word size and environment. The duration of C-CENTER was significantly longer in #CCVX than #CVX, across the environments (i.e., carrier phrases).

Model comparisons for the RIGHT-EDGE models revealed that none of the fixed effects or their interaction contributed significantly to the model fit. The effects of word size [$\chi^2(1) = .12$, $p = .73$], environment [$\chi^2(1) = .18$, $p = .67$], and their interaction [$\chi^2(1) = 0$, $p = .99$] were not significant. These outcomes suggest that the RIGHT-EDGE duration was not influenced by the word size or the environment. The #CVX and #CCVX words, regardless of the environment, had similar RIGHT-EDGE interval.

This pattern held across speakers. Figures 2 and 3 show each speaker's C-CENTER and RIGHT-EDGE measures. For all speakers, C-CENTER, but not RIGHT-EDGE, was longer in #CCVX than #CVX. Despite individual differences in their overall interval measures, all speakers showed more stability in RIGHT-EDGE than C-CENTER intervals.

The RIGHT-EDGE stability was sustained across all word pairs, regardless of their sonority distance. Figure 4 shows the RIGHT-EDGE interval duration of #CCVX and #CVX in each pair.

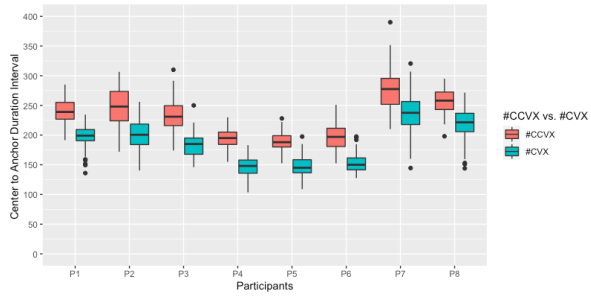


Figure 2: C-CENTER in #CCVX vs. #CVX by speaker

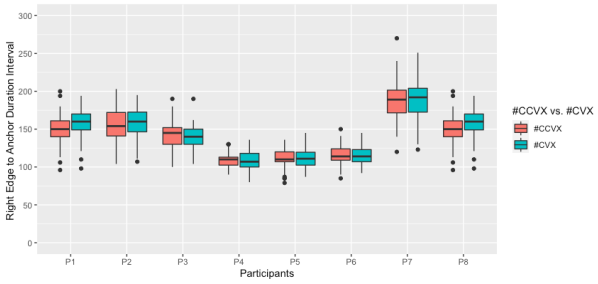


Figure 3: RIGHT-EDGE in #CCVX vs. #CVX by speaker

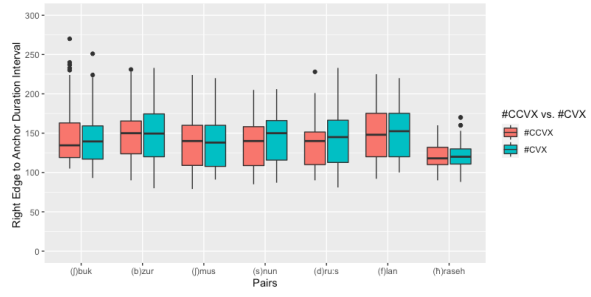


Figure 4: RIGHT-EDGE in #CCVX vs. #CVX by item

3.2. Variance analysis

The relative standard deviation (RSD) was calculated across the two intervals, and then compared using logit mixed-effects model analysis in R [16] with the lme4 package [17]. The dependent variable was the RSD. The predictor was the interval type (C-

CENTER vs. RIGHT-EDGE). Participants and items were included as random effects in the model.

Model comparison revealed a significant effect of the interval type [$\chi^2(1) = -3.92$, $p < .001$]. The RSDs of RIGHT-EDGE were significantly smaller than those of C-CENTER (Figure 5).

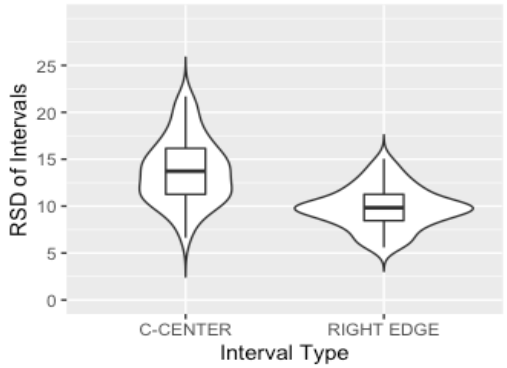


Figure 5: RSD of C-CENTER vs. RIGHT-EDGE

3.3. Summary of findings

The results of the duration analysis and the variance analysis, taken together, suggest that RIGHT-EDGE was more stable than those of C-CENTER. Words of a different size (#CCVX vs. #CVX) had different C-CENTER durations but the RIGHT-EDGE durations remained constant. This pattern was found across all speakers and items examined. In addition, the RIGHT-EDGE intervals had smaller RSD than the C-CENTER intervals.

4. DISCUSSION

Najdi #CC sequences with a rising sonority profile, unlike those with a falling or plateau sonority profile, were reported to surface without vowel insertion [8]. This raised a question if the #CC sequences, not modified by a prothetic vowel, would form a true onset cluster. This study examined the temporal patterns of Najdi word-initial CC sequences with a rising sonority profile to answer this question.

The overall results confirmed that C-CENTER and RIGHT-EDGE were substantially different. RIGHT-EDGE was shown to be not different between #CVX and #CCVX (see Figures 1 and 3), remaining constant regardless of the number of consonants in a sequence. Although individual speakers and word pairs had their own range of RIGHT-EDGE measures (Figures 3 and 4), RIGHT-EDGE stability was found in all speakers and word pairs. On the other hand, C-CENTER was notably less stable and increased as the number of consonants in the sequences increases. This was indicated by that #CCVX had significantly longer C-CENTER than #CVX (Figure 2). Moreover, the results of the variance analysis using RSD established that C-CENTER interval showed a greater variance than RIGHT-EDGE interval (Figure 5). All these outcomes indicated a greater stability in RIGHT-EDGE than in C-CENTER.

Interestingly, these patterns were held constant regardless of whether the preceding segment was a vowel /i/ or a consonant /k/. The preceding segment did not encourage or impede heterosyllabic or tautosyllabic parsing of the #CC sequences.

A greater stability in RIGHT-EDGE than in C-CENTER suggests that Najdi #CC sequences show the temporal patterns more consistent with hetero-syllabic parsing [#C.C] than tautosyllabic parsing [#CC]. Similar to Moroccan Arabic [3, 4] and Jazani Arabic [6], Najdi Arabic appears to have simplex onsets, with #CC sequences being heterosyllabically parsed as [#C.C].

This study relied on acoustically measured landmarks to obtain C-CENTER and RIGHT-EDGE intervals, similar to [6]. The acoustic measures used in the current study did not capture exactly the same information as the articulatory measures used in most previous studies on the temporal organization of consonant sequences (e.g., [1, 2, 3]). Still, the acoustic measures revealed the consistent patterns in both duration and variance analyses, supporting the interpretation that one interval (RIGHT-EDGE) was more stable than the other (C-CENTER). This provides a support to the claim in [6] that acoustic measures can be used to examine syllabic organizations of consonant sequences.

5. CONCLUSION

The current study provides evidence, based on the temporal measures of the #C_2VX and #C_1C_2VX sequences, that #CC in Najdi may not be tautosyllabic consonant clusters even when they surface without a prothetic vowel. Instead, the #CC sequences are likely syllabified as [#C_1.C_2]. As suggested in [8], Najdi Arabic may have an underlying vowel before the consonant sequences /#VC_1.C_2/, which surfaces only when C_1 is more sonorous than, or as sonorous as, C_2. When C_1 is less sonorous than C_2 (i.e., sonority rise), the underlying vowel is deleted in the surface form but the temporal organization of the consonants still show the heterosyllabic parsing.

Our findings suggest that Najdi is the third Arabic dialect that shows the RIGHT-EDGE stability, following Moroccan Arabic [3, 4] and Jazani Arabic [6]. This raises the question of whether there would be other Arabic varieties showing the same pattern, calling for more studies that examine the temporal patterns of word-onset consonant sequences in other Arabic varieties.

6. ACKNOWLEDGMENTS

We would like to give special thanks to Dr. Steven Weinberger, Dr. Ghada Khattab, and Dr. Douglas Wulf for their valuable comments and feedback. Our appreciation extends also to all faculty members and PhD students at the English department—linguistics major—in George Mason University.

7. REFERENCES

[1] Browman, C. P., Goldstein, L. 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45, 140–155.

[2] Byrd, D. 1995. C-centers revisited. *Phonetica* 52, 285–306.

[3] Shaw, J. A., Gafos, A. I., Hoole, P., Zeroual, C. 2009. Syllabification in Moroccan Arabic: Evidence from patterns of temporal stability in articulation. *Phonology* 26, 187–215.

[4] Shaw, J. A., Gafos, A. I., Hoole, P., Zeroual, C. 2011. Dynamic invariance in the phonetic expression of syllable structure: A case study of Moroccan Arabic consonant clusters. *Phonology* 28, 455–490.

[5] Hermes, A., Mücke, D., & Auris, B. (2017). The variability of syllable patterns in Tashlhiyt Berber and Polish. *Journal of Phonetics* 64, 127–144.

[6] Durvasula, K., Ruthan, M., Lin, Y., Heidenreich, S. 2021. Probing syllable structure through acoustic measurements: Case-studies on American English and Jazani Arabic. *Phonology* 38(2), 173–202.

[7] Abboud, P. 1979. The verb in northern Najdi Arabic. *Bulletin of the School of Oriental and African Studies* 42, 467–499.

[8] Alkhonini, O. (2021). *Underlying Representation and Acoustic Analyses of Two-Consonant Initial Clusters and Prothesized Vowels in Najdi Arabic* [Unpublished PhD Dissertation]. English Department, George Mason University.

[9] Hooper, J. B. (1976). *An introduction to natural generative phonology*. Academic Press.

[10] Kiparsky, P. (1979). Metrical structure assignment is cyclic. *Linguistic Inquiry*, *10*(3), 421–441.

[11] Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff, R. Oehrle, & M. Halle (Eds.), *Language sound structure: Studies in phonology, presented to Morris Halle by his teacher and students* (pp. 107–136). MIT Press.

[12] Gouskova, M. (2004). Relational hierarchies in optimality theory: The case of syllable contact. *Phonology, 21*(2), 201–250.

[13] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal forced aligner (Version 1.0.0) [Computer software]. Retrieved 5 May 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[14] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer (Version 6.3.03) [computer software]. Retrieved 17 December 2022 from http://www.praat.org/

[15] Jaeger, T. F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59, 434–446.

[16] R Core Team. 2021. R: A language and environment for statistical computing [computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48.

A SIMPLE ACOUSTIC MEASURE OF ONSET COMPLEXITY

Karthik Durvasula

Michigan State University
durvasul@msu.edu

ABSTRACT

The kinematic consequences of onset complexity have been extensively probed since the 1980s — complex and simple onsets generally show different temporal stability patterns with respect to the following vowel. More recently, it has been argued that acoustic methods can also be used to establish the same dichotomy. In this paper, I show that a rather simple acoustic measure of pre-vocalic consonant duration in word-initial CV and CCV sequences can be used to tease apart complex and simple onset organisations across languages. I show this through an analysis of controlled experimental data on American English and Jazani Arabic, and through an analysis of corpus data from three different American English corpora of sociolinguistic interviews or oral histories. The results in turn suggest that a relatively coarse measure (and a relatively simple learning algorithm) may be sufficient for a learner to acquire some aspects of syllable-structure from the acoustic input.

Keywords: Onset complexity, Production experiments, Corpora, English, Jazani Arabic

1. INTRODUCTION

The kinematic consequences of different onset organisations have been extensively probed since the 1980s [1, 2, 3, amongst others]. In languages with complex onsets, it's generally observed that word-initial consonants along with the following vowel show a 'global timing stability' pattern. More specifically, the mean of the mid-points of the consonantal gestures of a word-initial consonant sequence (called the c-centre) is temporally aligned to the end of the following vowel (called the anchor), *i.e.*, the c-centre is at a stable distance away from the end of the following vowel no matter how many consonants are in the sequence (see Figure 1, left). This pattern of c-centre-to-anchor interval stability has since been replicated for American English [4], and has also been observed in a variety of languages with complex onsets: Georgian [5], Italian [2], Polish [6], Romanian [7], and Spanish [8]. In contrast, word-initial consonant sequences in languages which allow at most a single consonant (simple onsets languages) have been observed to show a local timing stability' pattern. More specifically, word-initial consonant sequences have a right-edge-to-anchor interval stability, wherein the last consonantal gesture of a word-initial consonant sequence is in a stable temporal relationship with the following vowel, and the presence of more consonants word-initially before the prevocalic consonant does not change the timing between the prevocalic consonant and the anchor (see Figure 1, right). This pattern of right-edge-to-anchor interval stability has been observed in Tashlhiyt Berber [5, 6], Moroccan Arabic [9, 10] and Jazani Arabic [11],

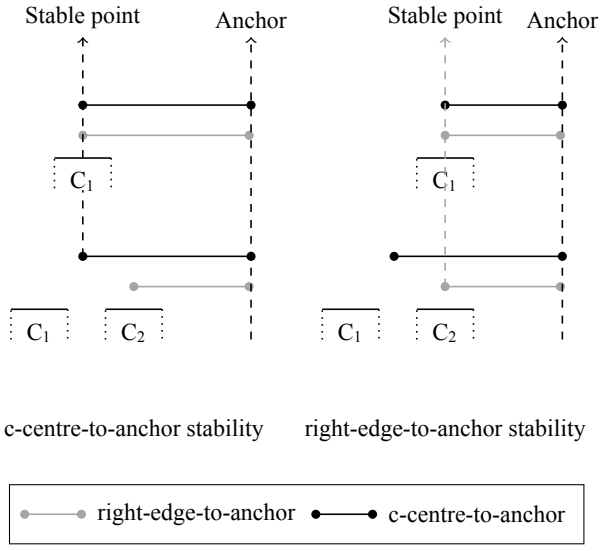


Figure 1: Schematic representations of c-centre-to-anchor stability patterns (left) and right-edge-to-anchor stability patterns (right). The x-axis in the figure represents time. The anchor marks the end of the following vowel, and C_1-C_2 represent word-initial consonants.

There are however a few results that contradict the above linking hypothesis between temporal stability patterns and onset complexity in a language. There are arguments in the phonological literature that suggest that word-initial consonant sequences in Hebrew [12], French [13] and German [14] form complex onsets. However, the three languages have been observed to show a right-edge alignment, at

least for some consonant sequences [15, 16, 17]. Relatedly, the c-centre-to-anchor stability pattern was observed for only some of the consonant clusters studied in Romanian [18] and Polish [6], though all the clusters studied were previously argued to be complex onsets.

While the observations might be at first blush seem problematic for the linking hypothesis discussed above, Mücke, Hermes, and Tilsen [19] recently argued that the patterns observed in the languages are consistent with a complex onset organisation, and previous research likely misinterpreted the relevant articulatory data. More specifically, they suggest that explicitly modelling the speaker-specific coupling strength between gestures and speaker-specific biomechanical interactions between articulators allows us to still understand the patterns in such languages as stemming from a c-centre organisation. They show this is possible at least for German. A second possibility to account for the discordant results is discussed by Durvasula, Ruthan, Heidenreich, and Lin [11], who suggest based on their results that it is possible that the relevant articulatory data is in fact indirect evidence of the the stability patterns present in the acoustics, and that there might be more stability for the c-centre-to-anchor than for the right-edge-to-anchor for the three languages, when the intervals are extracted from acoustic measurements. Both of the above suggestions raise the possibility that the articulatory stability patterns observed in the three languages may not be counterexamples to the linking hypothesis after all. A third possibility raised by Sotiropoulou, Gibson, and Gafos [8] is that the temporal stability pattern is only a part of the gamut of correlates that distinguish global timing stability in complex onset languages from local timing stability in simple onset languages, and really the distinction between the two is simultaneously expressed over a set of different phonetic parameters rather than just through a single measure such as c-centre or right-edge stability. In this paper, I take inspiration from the last two possibilities.

Most relevant for current purposes are two recent findings: (a) Sotiropoulou, Gibson, and Gafos [8] point out an important correlate of global timing stability in complex onset languages — a segment in the C_2 position of a $\#C_1C_2V$ sequence may be shorter than when the same segment is in the C_1 position of $\#C_1V$ sequence. A corollary they do not explore, but is necessary to complete their argument, is the following: in simple onset languages, a segment in the C_2 position of a $\#C_1C_2V$ sequence should be the same duration as when the same segment is in the C_1 position of $\#C_1V$ sequence. (b) Durvasula, Ruthan, Heidenreich, and Lin [11] and Durvasula and Selkirk [20] show that correlations between c-centre/right-edge stability patterns and onset organisation are also observable through an analysis of *acoustic* measurements. I combine the above two lines of research and ask if the simple durational measurement of pre-vocalic consonant duration observable in acoustic recordings correlates with onset complexity both in a lab-based production experiment and in three different corpora of American English.

2. EXPERIMENT 1: PRODUCTION EXPERIMENT

I used the data published by [21] in an Open Science Foundation repository; it includes TextGrids from production experiments on 10 American English speakers (a complex onset language) and 7 Jazani Arabic speakers (a simple onset language). The crucial test items consisted of C_1VC and C_1C_2VC words in both languages, and the crucial word-initial consonant sequences consisted of fricatives and nasals (e.g., English: <nap, snap>; Jazani Arabic: [məʕ] 'with', [səmʕ] 'listen'). There were 16 words for the American English experiment (8 test), and 78 words for the Jazani Arabic experiment (all test). The actual counts of the consonantal positions in each experiment are shown in Table 1.

Corpus	C_1	C_2
English	389	388
Jazani Arabic	1850	1469

Table 1: Counts of the test consonants in relevant positions (underlined) in $\#\underline{C_1}V\ldots$ and $\#C_2\underline{C_2}V\ldots$ words in each of the three corpora.

Praat [**?**] scripts were used to extract the relevant consonant duration measurements. Visual inspection suggested that a consonant in the second position of a word-initial consonant sequence is in fact shorter than the same consonant in a simple onset for American English but not for Jazani Arabic (see Figure 2).

The English and Jazani Arabic consonant durations were modelled using the programming language R [22]. Separate linear mixed effects models were fitted using the R packaes `lme4` and `lmerTest` [23, 24] for each language with a random effects structure that included by-subject, and by-segment random intercepts and by-subject, by-segment and by-corpus random slopes for the position of the consonant. The independent variable

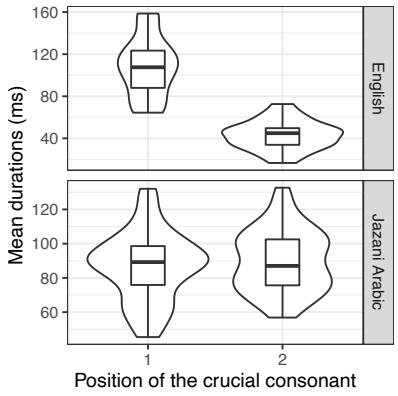


Figure 2: Durations of pre-vocalic consonants in initial and second positions of word-initial consonant sequences in the experimental data; English (top) vs. Jazani Arabic (bottom)

was Consonant Position (initial vs. second). As with the initial visual inspection, there was a statistically clear shortening of a consonant in the second position of word-initial consonant sequences compared to the same consonants in initial position for English ($\hat{\beta}$ = -64.8 ms; 95% CIs = -83.7 – -45.5 ms; see Table 2), but not for Jazani Arabic ($\hat{\beta}$ = 3.3 ms; 95% CIs = -4.0 – 10.6 ms; see Table 3). [Note: throughout, the phrase "statistically clear", and variants, are used instead of "statistically significant" on the recommendation of [25].]

	Est.	df	t	Pr(>|t|)
Interval	107.86	9.20	11.19	< 0.0001
Position: 2	-64.78	4.65	-6.64	0.002

Table 2: Linear mixed-effects model for the English consonant duration (random effects structure: by-subject, by-segment random intercepts and random slopes of Consonant Position)

	Est.	df	t	Pr(>|t|)
Interval	86.66	9.63	12.22	< 0.0001
Position: 2	3.32	12.60	0.92	0.37

Table 3: Linear mixed-effects model for the Jazani Arabic consonant duration (random effects structure: by-subject, by-segment random intercepts and random slopes of Consonant Position)

3. EXPERIMENT 2: CORPUS DATA

Subsequent to the analysis presented above, I probed if the consonant shortening was observable in naturalistic productions *without* highly controlled stimuli that are usually the case in laboratory experiments. In short, I aimed to probe if such shortening was observable simply by comparing prevocalic consonantal durations across-the-board *without* controlling for rhymes or any other factors (word frequency, morphological complexity, …).

For this purpose, I used high-quality recordings of conversational American English speech from three different corpora: (a) **The Buckeye corpus**: conversational speech from 40 speakers [26]. The speakers were all natives of Central Ohio, and were recorded in Columbus, Ohio before Spring 2000. The sample design was stratified for age (under thirty and over forty) and sex. The recordings were hand-transcribed, and both the recordings and the transcriptions are available at https://buckeyecorpus.osu.edu. (b) **The Influence of Higher Education on Local Phonology (I-Help) corpus**: 60 sociolinguistic interviews conducted in 2014 [27]. The speakers were all natives of the Greater Lansing area of Michigan. In this article, I report the the results of an analysis of 44 speakers (29 self-reported female, and 15 self-reported male) that were annotated and I was given access to by the authors. The interviews were transcribed and time-aligned by the original authors in ELAN [28]. (c) **The Auto Town Corpus**: 21 oral histories collected in the mid-1990s and 2000s [27]. The speakers (9 self-reported women, 12 self-reported male) who varied in ages (date of birth range: 1907-1971) were all former auto plant workers of the Local 602, Fisher Body and Diamond REO assembly plants in Lansing, Michigan. The interviews were again transcribed and time-aligned by the original authors in ELAN.

The focus was on consonants that can appear as the second consonant in word-initial consonant sequences in American English [m, n, l, r, w, p, t, k]. All the words where the relevant consonants were the first consonant in #CV… sequences and the second consonant in #CCV… were identified. No other restrictions were imposed to narrow down the word-lists. Praat scripts were used to extract the relevant measurements from a total of 149,044 words from the three corpora (Buckeye = 84,046, I-Help = 56,643, Autotown = 8,355). The actual counts of the relevant consonantal positions in each corpus are shown in Table 4.

Corpus	C_1	C_2
Autotown	7111	1244
Buckeye	72488	11558
I-Help	48165	8478

Table 4: Counts of the test consonants in relevant positions (underlined) in #$\underline{C_1}$V… and #$C_2\underline{C_2}$V… words in each of the three corpora.

As with the American English results in

Experiment 1, the results suggest that a consonant in the second position of a word-initial consonant sequence is in fact shorter than when it is a simple onset (see Figure 3). The durations were modelled using a linear mixed effects model with a random effects structure that included by-subject, by-segment and by-corpus random intercepts and by-subject, by-segment and by-corpus random slopes for the position of the consonant. In line with the visual inspection, a statistically clear shortening ($\hat{\beta}$ = -17.5 ms; 95% CIs = -28.2 – -6.5 ms) was observed for the second consonant in complex onsets compared to the same consonants in initial position (see Table 5).

	Est.	df	t	Pr(>|t|)
Interval	71.89	8.07	10.57	< 0.0001
Position: 2	-17.53	7.99	-3.35	0.01

Table 5: Linear mixed-effects model for the consonant duration (random effects structure: by-subject, by-segment and by-corpus random intercepts and random slopes of Consonant Position)

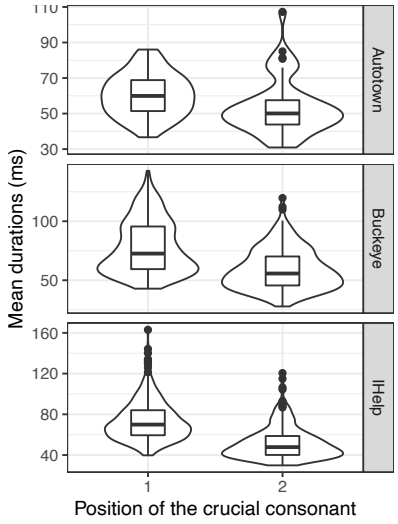


Figure 3: Durations of the consonants in relevant positions (underlined) in #$\underline{C_1}$V… and #$C_2\underline{C_2}$V… words in the three corpora.

4. DISCUSSION

A clear effect of onset complexity on acoustic consonant durations was found. There was no statistically clear difference in the durations of pre-vocalic consonants in word-initial CV and CCV words in Jazani Arabic, but there was a statistically clear difference in the durations of similar words for American English, both in controlled production data collected in a laboratory and in much more naturalistic sociolinguistic interviews or oral histories. Furthermore, the 95% confidence intervals for the controlled production experiment were quite far away from zero (-83.7 – -45.5 ms), while the same for the controlled Jazani Arabic production experiment included a very small range close to zero (-4.0 – 10.6 ms). This suggests that the differences observed in the American English cases are quite meaningful, and substantially different in character from Jazani Arabic.

The results therefore suggest that the acoustic shortening of the pre-vocalic consonant duration could indeed be a reliable cue to differentiate consonant sequences that form complex onsets from those that don't. We echo the important point made in Durvasula, Ruthan, Heidenreich, and Lin [11] that the presence of such syllable-structure signatures in the acoustic recordings makes it easier to collect phonetic data relevant to syllable complexity both in the lab and during fieldwork than extant methods.

The results also provide further support for claims in the phonological literature that word-edge (consonant) sequences do not constitute direct evidence of syllable structure in a language [29], and therefore highlight the need for clear linking hypotheses connecting abstract structure to phonological patterns or phonetic data (such as the one explored in the current paper) in order to identify the abstract structure.

Finally, this effect was observed in naturalistic corpora of acoustic measurements quite robustly. Given the uncontrolled nature of the stimuli, it is of course to be expected that the unstandardised effect size in the corpus analysis was much smaller than the same in the controlled production experiment for American English. Furthermore, the corpus results are particularly exciting because they suggest, contrary to standard views, that the acoustic input not only has information about some aspects of syllable-structure, but that relatively coarse data is sufficient to observe this pattern, which thereby suggests that an extremely simple learning algorithm that keeps track of pre-vocalic consonant durations is potentially sufficient for a learner to acquire some aspects of syllable-structure directly from the acoustic input they receive.

5. REFERENCES

[1] C. P. Browman and L. H. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140–155, 1988.

[2] A. Hermes, D. Mücke, and M. Grice, "Gestural coordination of Italian word-initial clusters: The case of 'impure s'," *Phonology*, vol. 30, no. 1, pp. 1–25, 2013.

[3] D. Byrd, "C-centers revisited," *Phonetica*, vol. 52, pp. 263–282, 1995.
[4] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in american english: Testing the predictions of a gestural coupling model," *Motor Control*, vol. 14, no. 3, pp. 380–407, 2010.
[5] L. H. Goldstein, I. Chitoran, and E. Selkirk, "Syllable structure as coupled oscilator models: Evidence from georgian vs. tashlhiyt berber," *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 241–244, 2007.
[6] A. Hermes, D. Mücke, and B. Auris, "The variability of syllable patterns in tashlhiyt berber and polish," *Journal of Phonetics*, vol. 64, pp. 127–144, 2017.
[7] S. Marin and M. Pouplier, "Articulatory synergies in the temporal organization of liquid clusters in romanian," *Journal of Phonetics*, vol. 42, pp. 24–36, 2014.
[8] S. Sotiropoulou, M. Gibson, and A. Gafos, "Global organization in spanish onsets," *Journal of Phonetics*, vol. 82, p. 100995, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0095447020300863
[9] J. A. Shaw, A. I. Gafos, P. Hoole, and C. Zeroual, "Syllabification in moroccan arabic: evidence from patterns of temporal stability in articulation," *Phonology*, vol. 26, no. 1, pp. 187–215, 2009.
[10] ——, "Dynamic invariance in the phonetic expression of syllable structure: a case study of moroccan arabic consonant clusters," *Phonology*, vol. 28, no. 3, pp. 455–490, 2011.
[11] K. Durvasula, M. Q. Ruthan, S. Heidenreich, and Y.-H. Lin, "Probing syllable structure through acoustic measurements: case studies on american english and jazani arabic," *Phonology*, vol. 38, no. 2, p. 173–202, 2021.
[12] S. Bolozky, "Israeli hebrew phonology," in *Phonologies of Asia and Africa (including the Caucasus). Vol. 1*, A. S. Kaye, Ed. Eisenbrauns, 1997, pp. 287–311.
[13] F. Dell, "Consonant clusters and phonological syllables in french," *Lingua*, vol. 95, no. 1, pp. 5 – 26, 1995, french Phonology. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0024384195900993
[14] R. Wiese, *The Phonology of German*. Clarendon, 1996.
[15] M. Pouplier, "The gestural approach to syllable structure: universal, language- and cluster-specific aspects," pp. 63–96, 2012.
[16] J. Brunner, C. Geng, S. Sotiropoulou, and A. Gafos, "Timing of german onset and word boundary clusters," *Laboratory Phonology*, vol. 5, pp. 403–454, 2014.
[17] S. Tilsen, D. Zec, C. Bjorndahl, B. Butler, M.-J. L'Esperance, A. Fisher, L. Heimisdottir, M. Renwick, and C. Sanker, "A cross-linguistic investigation of articulatory coordination in word-initial consonant clusters," *Cornell Working Papers in Phonetics and Phonology 2012*, pp. 51–81, 2012.
[18] S. Marin, "The temporal organization of complex onsets and codas in romanian: A gestural approach," *Journal of Phonetics*, vol. 41, no. 3, pp. 211–227, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447013000168
[19] D. Mücke, A. Hermes, and S. Tilsen, "Incongruencies between phonological theory and phonetic measurement," *Phonology*, vol. 37, no. 1, pp. 133–170, 2020.
[20] E. Selkirk and K. Durvasula, "Acoustic correlates of consonant gesture timing in english," 2013, paper presented at the 166th Meeting of the Acoustical Society of America, San Francisco, USA.
[21] K. Durvasula, M. Ruthan, S. Heidenreich, and Y.-H. Lin, "Osf repository for *Probing Syllable Structure Through Acoustic Measurements: Case-studies on American English and Jazani Arabic*," 2021.
[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org
[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using `lme4`," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[24] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.
[25] J. Dushoff, M. P. Kain, and B. M. Bolker, "I can see clearly now: Reinterpreting statistical significance," *Methods in Ecology and Evolution*, vol. 10, no. 6, pp. 756–759, 2019. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13159
[26] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.
[27] S. E. Wagner, A. Mason, M. Nesbitt, E. Pevan, and M. Savage, "Reversal and re-organization of the northern cities shift in michigan," *University of Pennsylvania Working Papers in Linguistics*, vol. 22, no. 2, pp. 171–179, 2016. [Online]. Available: https://repository.upenn.edu/pwpl/vol22/iss2/19/
[28] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *5th international conference on language resources and evaluation (LREC 2006)*, 2006, pp. 1556–1559.
[29] T. Borowsky, "Topics in the lexical phonology of English," Ph.D. Dissertation, University of Massachusetts, Amherst, MA, USA, 1986. [Online]. Available: https://scholarworks.umass.edu/dissertations/AAI8701140/

SYLLABIC REDUCTION IN ITALIAN CONNECTED SPEECH: TOWARDS THE INTEGRATION OF LINGUISTIC AND COMPUTATIONAL APPROACHES

Loredana Schettino, Vincenzo Norman Vitale, Francesco Cutugno

University of Naples Federico II
{loredana.schettino, vincenzonorman.vitale, cutugno}@unina.it

ABSTRACT

In spontaneous speech, words are normally uttered in continuous chains whereby boundaries are blurred and segments can undergo changes, e.g., vowel centralization, consonant lenition, or deletion, due to structural linguistic factors or contingent sociolinguistic factors. However, such "reduction" processes are often studied by comparing connected speech realizations with isolated words in their expected, phonological forms. This study aims at investigating systematic patterns of phonetic variation in a dataset of spontaneous Italian discourse by considering the syllabifications of the speech chain at the phonological and phonetic levels. Moreover, since phonetic segmentation, based on human perception, implies a discretization of continuous productions, we employ unsupervised computational techniques, namely clustering algorithms, which allow for the observation of patterns emerging from the data without external (human) interference. The linguistic analysis shows systematic reduction processes related to the syllabic structure and lexical stress. The computational representation highlights a convergence of syllables to the non-marked structure in Italian (CV).

Keywords: reduction, syllabic structures, connected speech, clustering, unsupervised learning

1. INTRODUCTION

In spontaneous speech, the phonetic realization of words may vary with respect to their "canonical form" and speakers' utterances can be represented as continuous chains of words which, for the economy of production efforts, are commonly under-specified in the speech signal. This normally results in acoustic "reduction" which can manifest itself as spectral reduction, temporal reduction, and absence or change of segments [1] and ranges from subtle weakening up to segmental changes, i.e. from vowel centralization or consonant lenition to deletion of segments or even of multiple syllables [2, 3]. Speech reduction phenomena have been described in many different languages (a.o. German [4], Dutch [5], American English [6], Italian [7], French [2]) as representing, in fact, the rule rather than the exception in spoken language communication [8].

The type or extent of this variation may depend on sociolinguistic factors related to the communicative situation and the speakers [9] and linguistic factors, like prosodic features, lexical category, the discursive function [4, 10, 5]. In particular, a study on Italian [11] showed that vowel centralization represents a structural feature independent of diaphasic and diatopic variation. As for the listeners' processing of pronunciation variants, it was highlighted that in speech understanding, the frequency of occurrence of the variants with relation to specific contexts plays a more important role than the degree of reduction [3, 12]. Therefore, in-depth investigations of the patterns of phonetic variation in speech are crucial to deepen our understanding of speech production and comprehension mechanisms.

Reduction processes have often been studied by comparing, segment by segment, phonetic realizations with the phonological forms of isolated words, even when they appeared in connected speech. At the same time, evidence has been provided that the syllable can represent a relevant basic unit of speech production and perception rather than phonetic segments [13, 8]. In particular, [14] shows that the observable variation in connected speech is more systematic at the syllabic level than at the segmental one. Consequently, comparing the syllable sequence of the speech chain as expected in the phonological stream to that observed phonetically seems to be most suitable to describe reduction phenomena in connected speech.

In the last couple of decades, technological tools have been introduced to support the analysis of acoustic reduction by providing automatic and internally coherent phonetic annotations [2, 15, 1]. Further computational tools could be employed in the analysis of systematic phenomena in phonetic

realizations. In particular, clustering algorithms are unsupervised tools that could be exploited to observe how segments identified as phonetic syllables will be grouped on the basis of the features extracted from the corresponding signal. In a recent work on the explainability of Deep Learning based Automatic Speech Recognition (ASR) systems [16], authors employed the K-Means algorithm to evaluate the similarity of information encoded in Deep Neural Networks (DNN) layers. However, the considered approach employs unknown features which are automatically extracted and encoded by DNN. Conversely, we aim to conduct an informed analysis by extracting specific features generally used in speech analyses [17, 18] through unsupervised computational tools in order to provide further support to the analysis of the patterns of phonetic variation in speech.

The main purpose of this study is to build upon the research proposed in [11] by investigating the phonetic variation patterns that may be observed in the speech chain, and their relation to specific linguistic structures, namely, syllabic structures and stress with the support of computational methods.

2. METHOD

2.1. Data and annotation

The investigation concerns the Italian dataset of the Nocando corpus [19] which consists of spoken narrative texts by 11 subjects (university students).

To obtain reliable annotations, the audio files and transcriptions were processed using the WebMAUS Basic services [20] and the provided phonological and phonetic transcriptions were manually edited in Praat [21] and syllabified according to the principles of sonority sequencing and onset maximization [22]. These principles were applied to both the phonological and phonetic representation of the speech chain, which means that

- resyllabification is expected between words when the second one begins with a vowel (e.g. *un'orecchio* [u.no.ˈre.kkjo];
- long consonants and geminates are considered as *onset* of the same syllable;
- the initial [s] of a consonantal nexus is assigned to the *coda* of the preceding syllable (e.g. *questo* [ˈkwes.to] in Italian). Exceptionally, it is considered *onset* only in the rare cases when it occurs at the absolute beginning of utterance.

So "phonological syllables" are those expected after resyllabification but before reduction phenomena, whereas "phonetic syllables" are those that are actually realised (see Fig. 1).

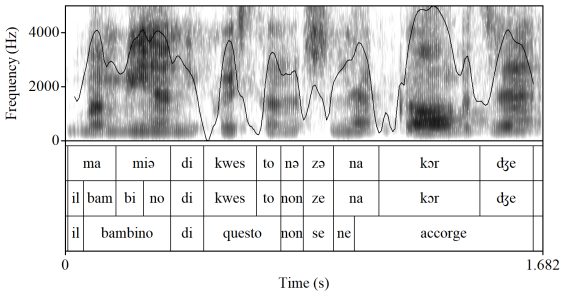


Figure 1: Example of annotation. I: phonetic level; II: phonological level; III: ortographic level.

2.2. Linguistic analysis

The alignment between the sequence of phonological syllables and one of the phonetic syllables was evaluated using SCLITE, which is a tool included in the Speech Recognition Scoring Toolkit (SCTK) provided by the American National Institute of Standards and Technology (NIST). The evaluation output reports the following cases:

- "Deletion", a phonetic syllable is found in place of two phonological ones;
- "Substitutions", a phonetic syllable is different from the corresponding phonological one;
- "Correct", the phonetic and the phonological syllables are alike.

The phonological syllables were also annotated for their phonetic structure (CV, CCV, CCVC, CVC, VC, V) and contextual saliency (lexical stress). For syllables subjected to substitution, the position of the syllabic structure in which it occurs is indicated, namely, Onset (O), Nucleous (N) and/or Coda (C).

To evaluate the role of lexical stress and syllabic structure on phonetic variation and control for individual variability, a statistical analysis was conducted on R [23] and Generalized Linear Mixed Models were fitted [24]. First, the evaluation levels were considered as binomial dependent variables; Syllabic Structure and Lexical Stress as interacting independent variables, and Speaker as a random effect. Then, on the subset of syllables subjected to substitution, the positions of the changed segments were considered as binomial dependent variables, Lexical Stress as an independent variable.

2.3. Computational Analysis

Considering the phonetic syllable as the base unit, we also assume the syllable is a data point represented with a discrete set of features. Based on the above-mentioned discrete representation, we employ a class of unsupervised learning approaches

named *Clustering*. Approaches of this kind are meant to group similar items based on similarity criteria, which in our case means comparing syllables on the basis of a discrete set of features rather than the whole signal. The following clustering techniques, implemented in the Scikit-Learn toolkit [25], were considered:

- **K-Means** [26] is a vector-quantization method that divides n objects in k clusters based on their mean distance;
- **Hierarchical Agglomerative Clustering (HAC)** [27] is a greedy technique that aims at grouping (or splitting) clusters based on a similarity measure. The final output is a clusters hierarchy which could be divided based on the number of desired clusters.

These techniques require, in some way, an apriori expert intervention to determine the number of desired clusters. In our case, for both **HAC** and **K-Means**, the desired *number of clusters (k)* was fixed to the five most frequent classes of syllable structures among those found in the dataset described in Section 2.1, namely those with more than ten samples. Then, through the OpenSmile toolkit [17], we extract, for each considered sample, the GeMAPS set [18], composed of 62 features.

3. RESULTS

3.1. Linguistic Analysis

The dataset considered for the analysis consists of 940 syllables at the phonological level. Most of these are phonetically realized as expected (C=67%), a quarter of them have been subjected to variation (24%) and the rest have been deleted (D=9%). As expected, most syllables are unstressed (71%) and the stressed ones are fewer (29%). As for the syllabic structure, the following patterns are observed: CV (65%), CVC (16%), CCV (9%), VC (3%), CCVC (3%), V (2%).

The presence of lexical stress and the syllabic structure are both found to be significant predictors of phonetic variation (see Fig. 2 and 3). In particular, lexical stress predicts the occurrence of deletions, with unstressed syllables being most likely to be deleted (*Est*=-1.56, *SE*=0.40, *z*=-3.88, *p*=0.0001), but not the occurrence of substitutions (*Est*=-0.33, *SE*=0.18, *z*=-1.91, *p*=0.055). As for syllable structures, CV represents the most stable structure, which is significantly less subjected to variation (*Est*=-0.63, *SE*=0.28, *z*=-2.26, *p*=0.023), whereas V and VC structures are significantly more subjected to deletion (V: *Est*=2.43, *SE*=0.92, *z*=2.64, *p*=0.008; VC: *Est*=1.68, *SE*=0.81, *z*=2.06, *p*=0.039).

As for the cases of substitution, most changes concern either the syllabic onset (46%) or the nucleus (45%). Moreover, lexical stress (in Fig. 4) is found to predict which part of the syllable is affected by variation. Namely, the nuclei in unstressed syllables (*Est*=-1.13, *SE*=0.40, *z*=-2.82, *p*=0.005), whereas the syllable onsets in stressed ones (*Est*=0.81, *SE*=0.32, *z*=2.45, *p*=0.012).

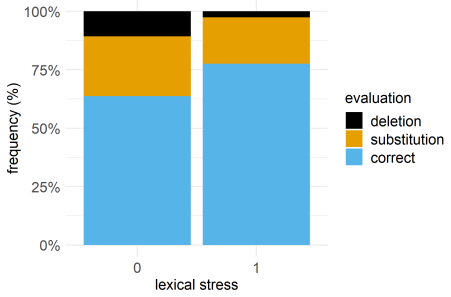


Figure 2: Frequency (%) of the evaluation output cases per stressed (1) and unstressed (0) syllables.

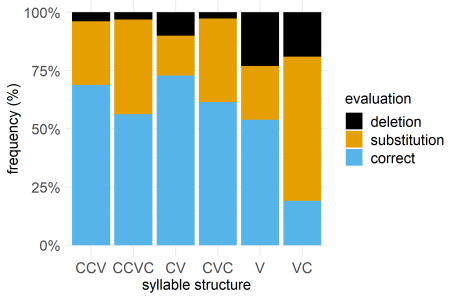


Figure 3: Frequency (%) of the evaluation output cases per syllabic structure.

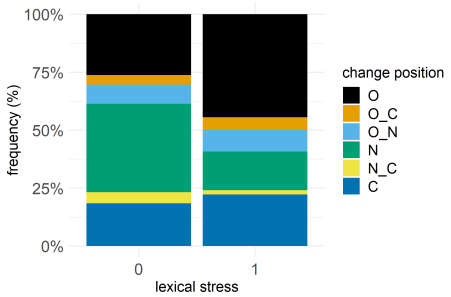


Figure 4: Frequency (%) of change position in the syllabic structure per lexical stress condition.

3.2. Computational Analysis

In table 1, we report results for each considered clustering technique. Each pattern label is associated with a cluster based on frequency; for example, the CV pattern label is assigned to the cluster with the most significant number of samples

belonging to the CV pattern. Results highlight a particularly high similarity between the CV pattern, which is also the most frequent in the considered dataset, and the CCV and CCVC patterns (the K-Means assigns the CV label to most items per each class of syllabic structure, see the CV column in Table 1). This means that, as for the considered statistical features, the CV pattern is acknowledged to be very similar to the other two ones when considering algorithms requiring some prior knowledge about data, like the number of clusters for K-Means and HAC. Please note that behind all patterns a lot of acoustic information is encapsulated in both the onset *C* and *V*, which are representative symbols for different, respectively, consonant phones and vowel phones.

Algorithm	**Class**	CCV	CCVC	CV	CVC	VC
K-Means	CCV	**12**	6	21	17	9
K-Means	CCVC	0	**4**	9	6	0
K-Means	CV	85	128	**227**	55	81
K-Means	CVC	0	15	68	**35**	3
K-Means	VC	1	0	5	2	**1**
HAC	CCV	**21**	17	6	17	4
HAC	CCVC	0	**5**	4	7	3
HAC	CV	166	140	**174**	84	12
HAC	CVC	3	44	21	**49**	4
HAC	VC	2	2	3	2	**0**

Table 1: Clustering results comparison.

4. DISCUSSION

The comparison between the syllabic annotation of the speech chain at the phonological and phonetic levels provided insight into the reduction processes, or rather the systematic phonetic variation, occurring in connected speech net of the expected syllabic restructuring phenomena across word boundaries. The analysis of the systematic relation between the variation phenomena and the syllabic structural characteristics, that is, in this case, the presence of lexical stress and structural configuration in terms of vowel and consonant(s) patterns allowed for the description of reduction processes related to linguistic structures which are, to a certain extent, independent from sociolinguistic factors. As expected [11], unstressed syllables are most likely to undergo reduction phenomena. However, distinguishing between deletions and substitutions, we found that unstressed syllables can be deleted rather than stressed ones, but both, stressed and unstressed syllables, can be almost equally subjected to substitution. Then, further investigation on substitution phenomena highlighted that systematic differences concern the position of the change, in that the presence of lexical stress prevents vowel deletion or change (such as centralization, which concerns unstressed syllables) but allows for onset changes (such as lenition or assimilation). As for the syllabic structure, CV is confirmed as the most frequent structure [28] and the most stable and resistant to variation. Instead, V and VC structures are most prone to deletion and to restructuring processes in the speech chain. This seems also in line with Greenberg's observation that syllable onsets are generally preserved while coda or nuclear constituents are more frequently subjected to underspecification [14].

The computational representation supported this finding by highlighting that most syllables, even when belonging to different structural classes according to the manual annotation, are associated with the CV label, which means that more complex structures (CCV, CCVC, CVC patterns) show, in a way, a considerable degree of similarity with CV structures. In particular, the stability of the syllable onset seems to ensure the recognition of a consonantal onset. As for syllables with CCV and CCVC structures, probably, the system recognizes consonantal acoustic features but, since it doesn't account dynamically for internal variations, it assigns a generic C to complex onsets. As for CVC structure, in Italian, closed syllables are most likely to contain a sonorant coda, so the cluster could assimilate the whole sonorant group to a nucleus and **CVC** is associated with **CV**. This suggests that the acoustic realization of syllables tends to converge to more simple structures. Nonetheless, caution is needed when interpreting this result given the risk of oversimplification by a computational representation that relies on features depicting each syllable as a whole and not allowing for investigation of the internal structure. Follow-up studies will concern the in-depth analyses of different clustering processes. However, this study supports the importance of an informed investigation that relies on domain knowledge rather than purely statistical computations.

The presented findings result from the constructive integration between more traditional linguistic and computational approaches to the description of systematic phonetic variation in speech, which can be relevant not only to unveiling the functioning of speech production and comprehension mechanisms but also to improving the performances of speech technologies such as ASR systems. [1]

[1]This work results from the collaboration between the authors: Loredana Schettino, §§ 1, 2.1, 2.2, 3.1, 4; Vincenzo Norman Vitale, §§ 1, 2.3, 3.2, 4; Francesco Cutugno, §§ 1, 4.

5. REFERENCES

[1] B. Schuppler, "Automatic analysis of acoustic reduction in spontaneous speech," Ph.D. dissertation, Radboud University Nijmegen, The Netherlands, 2011.

[2] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous french," *Speech communication*, vol. 46, no. 2, pp. 119–139, 2005.

[3] M. Ernestus and N. Warner, "An introduction to reduced pronunciation variants," *Journal of Phonetics*, vol. 39, no. 3, pp. 253–260, 2011.

[4] K. J. Kohler, "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations," in *Speech production and speech modelling*. Dordrecht: Kluwer, 1990, pp. 69–92.

[5] M. T. C. Ernestus, *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: LOT, 2000.

[6] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous speech: Data and analysis. Proc. 10th international symposium*, 2004, pp. 29–54.

[7] R. Savy, "Riduzioni foniche nel parlato. il ruolo della morfologia flessiva nellâinterpretazione del messaggio e nella comunicazione," Ph.D. dissertation, University of Roma Tre, Italy, 1999.

[8] F. Cangemi and O. Niebuhr, "Rethinking reduction and canonical forms," in *Rethinking reduction*, F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers, Eds. Berlin: De Gruyter, 2018, pp. 277–302.

[9] M. Ernestus, I. Hanique, and E. Verboom, "The effect of speech situation on the occurrence of reduced word pronunciation variants," *Journal of Phonetics*, vol. 48, pp. 60–75, 2015.

[10] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–1024, 2003.

[11] R. Savy and F. Cutugno, "Hypospeech, vowel reduction, centralization: how do they interact in diaphasic variations," in *Proc. XVIth International Congress of Linguists*, Paris, 1998.

[12] S. Brand and M. Ernestus, "Understanding reduced words. The relevance of reduction degree and frequency of occurrence," in *Proc. 19th International Congress of Phonetic Sciences, Melbourne*, 2019.

[13] F. Albano Leoni, "The boundaries of the syllable," in *The Notion of Syllable across History, Theories and Analysis*, D. Russo, Ed. Cambridge: Cambridge Scholars Publishing, 2016.

[14] S. Greenberg, "Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, 1999.

[15] M. Adda-Decker and N. D. Snoeren, "Quantifying temporal speech reduction in French using forced speech alignment," *Journal of Phonetics*, vol. 39, no. 3, pp. 261–270, 2011.

[16] A. Prasad and P. Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3739–3753.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[19] L. Brunetti, S. Bott, J. Costa, and E. Vallduví, "A multilingual annotated corpus for the study of information structure," in *Grammar & corpora 2009*, 2011.

[20] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." *Online: http://www.praat.org*, 1999-2022.

[22] M. Nespor, *Fonologia*. Bologna: Il Mulino, 1993.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2022. [Online]. Available: https://www.R-project.org/

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[27] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The computer journal*, vol. 26, no. 4, pp. 354–359, 1983.

[28] S. La Torre, "Le sillabe del lessico italiano: tipi, pattern e disposizioni. saggio di interrogazione del dbsli (data base delle sillabe del lessico italiano)," Ph.D. dissertation, University of Chieti-Pescara "G. D'Annunzio", Italy, 2005.

Emergent syllable types in Brazilian Portuguese

Thaïs Cristófaro Silva, Maria Mendes Cantoni

Federal University of Minas Gerais
thaiscristofaro@gmail.com, mmcantoni@gmail

ABSTRACT

This paper considers the emergence of syllable types in Brazilian Portuguese (BP). Studies on syllable inventory show that open syllables are the most frequent, corresponding to 78% of the syllables in BP words, whereas closed syllables occur in 22% of cases [23]. These syllable patterns are frequently assumed to reflect the phonotactic patterns of the language. We challenge this view and intend to show that closed syllables are becoming a common pattern in BP.

Our claim is supported by the analysis of different phonological processes which were empirically examined. We show that unstressed high vowel weakening and loss triggers the emergence of consonant-final syllables. Such a phenomenon is implemented phonetically and lexically in a gradual fashion. The emergent syllable types in BP may be similar to those from European Portuguese (EP). However, as the vowel which is weakened and lost in each language differs, the outcome for each language varies considerably. We argue that phonetic and phonological properties must operate together in order to promote the emergence of new syllable types in the language.

Keywords: vowel reduction, syllable types, emergent patterns, emergence, Portuguese.

1. INTRODUCTION

This paper considers the emergence of syllable types in Brazilian Portuguese (BP) based on works which examined unstressed high vowel weakening and loss in the language. Studies on syllable inventory show that open syllables are the most frequent, corresponding to 78% of the syllables in BP words, whereas closed syllables occur in 22% of cases [1]. These syllable patterns are frequently assumed to reflect the phonotactic patterns of the language. We challenge this view and show that closed syllables are becoming the most common pattern in BP. In this paper we argue that independent phonological processes may affect syllable structure yielding to the emergence of new syllable types.

In order to understand our proposal, it is important to consider consonantal clusters in BP. BP has tautosyllabic and heterosyllabic clusters [2]. Tautosyllabic clusters are formed by stops or labiodental fricatives followed either by a tap or a lateral (1a-d). Heterosyllabic clusters must have either a sibilant, a rhotic, a lateral (which may be vocalized as [w]), and a nasal element /N/ (which is not phonetically manifested), as illustrated in (1e-h).

(1) Tautosyllabic clusters

a.	[ˈfrakʊ]	*fraco*	‘weak’
b.	[ˈkɔbrə]	*cobra*	‘snake’
c.	[ˈbluzə]	*blusa*	‘blouse’
d.	[ˈãplʊ]	*amplo*	‘ample’

Heterosyllabic clusters

e.	[ˈfɛstə]	*festa*	‘party’
f.	[ˈkaɾtə]	*carta*	‘letter’
g.	[ˈsɛlvə]	*selva*	‘jungle’
h.	[ˈsãtʊ]	*santo*	‘saint’

Segmental restrictions define that tautosyllabic clusters may occupy any position within a word and be stressed or unstressed. The only restriction is that /vl/ is restricted to loans and /vr/ does not occur word-initially. /dl/ does not occur and /tl/ is restricted to few lexical items.

Heterosyllabic three-consonant clusters may exceptionally have two consonants in coda position, in which case a sibilant is always the final consonant of the syllable: *per**s**picaz* ‘insightful’, *sol**s**tício* ‘solstice’ and *tran**s**porte* ‘transport’. There is a considerable amount of variability in coda position depending on regional varieties: sibilants may be alveolar or alveopalatal and agree in voicing with the following consonant; rhotics can be either a back fricative or a tap as a consequence of neutralization (phonological contrast is observed intervocalically and lost in coda position), laterals may be vocalized in coda position. The nasal element /N/ is usually not pronounced but triggers the nasalization of the preceding vowel [3]. These constraints apply to BP as well as EP.

Another relevant constraint to the present discussion is the set of consonants that may traditionally occur word-finally. Consider (2).

(2) Word-final consonants

a.	[ˈlus]	*luz*	‘light’
b.	[ˈmaɾ]	*mar*	‘sea’
c.	[ˈsaw]	*sal*	‘salt’

Only sibilants, rhotics and laterals may occur word-finally in BP where similar segmental constraints apply as to heterosyllabic clusters. An epenthetic vowel has been posited to adjust syllable structure in cases a word-final consonant occurs, as in: [ĩtɛrˈnɛtʃi] 'internet' or [iuˈtubi] 'YouTube' [4]. The epenthetic vowel is a short high front vowel that mainly occurs in unstressed position but may also occur in stressed ones [5].

Finally, a few consonantal clusters occur word-finally being that the final consonant is always a sibilant.

(3) Word-final consonantal clusters
a. [ˈtɔɾaks] *tórax* 'chest'
b. [ˈbisɛps] *bíceps* 'biceps'

The segmental patterns observed word-finally for BP also apply to European Portuguese (EP). However, a wider range of consonants may occur word-finally in EP as apocope has suppressed word-final unstressed vowels. Vowel loss is also observed in unstressed pretonic position which triggers consonantal clusters to occur: [dɨlɨˈgadʊ] ~ [dɨlˈgadʊ] *delegado* 'delegated' [6, 7, 8]. Vowel reduction and loss in EP affected mostly unstressed /e/ which is reduced to [ɨ]. It is a well-advanced phenomenon that may be assumed to be under completion.

In recent years it has been observed that vowel weakening and loss is becoming recurrent in BP, being favoured to occur when a sibilant is adjacent to an unstressed high front vowel. In the next section we examine those studies to have an overall view of vowel weakening and loss in BP. We suggest that phonological processes may affect syllable structure yielding to the emergence of new syllable types.

2. VOWEL LOSS ACROSS WORD POSITIONS

Data presented in this section come from several works which examined unstressed high vowel weakening and loss in BP.

2.1. Word final consonants

Word-final consonants have been reported in several varieties of BP. Word-final vowel loss triggers any BP consonant to appear word-finally [9, 10, 11].

(4) Word-final consonants
a. [ˈlãʃ] *lanche* 'snack'
b. [ˈʃav] *chave* 'key'
c. [ˈʃɛk] *cheque* 'cheque'

Assis [10] claimed that the vowel weakening and loss is gradient and leaves fine phonetic detail traces in the acoustic signal. According to her, the loss of the word-final vowel triggers the shortening of the primarily stressed vowel. Thus, examples as /pas/ *paz* 'peace' and /ˈpasi/ *passe* 'entry ticket' which apparently sound the same, in fact present different pronunciations. She showed that when a word-final vowel occurs, as in [ˈpa:si] *passe* 'entry ticket' the primarily stressed vowel is longer than when apocope takes place: [ˈpas] *passe* 'entry ticket'. This result is in accordance with the general tendency for vowels to shorten in closed syllables [12]. She also showed that when a word-final sibilant occurs, as in [ˈpas] *paz* 'peace', the primarily stressed vowel presents the shortest durational values when compared to [ˈpas] or [ˈpasi] *passe* 'entry ticket'.

The main finding of this study is that several consonants may occur word-finally in BP and that vowel weakening and loss is gradient. She experimentally only examined sibilants occurring word-finally and found results compatible with previous studies [13, 14].

2.2. Word final consonantal clusters

Soares [5] examined regular plural forms in BP where the singular form ends in an unstressed word-final front vowel and is followed by [s] which is the suffix for regular plural in BP.

(5) Word-final consonantal clusters
a. [ˈʃavis] ~ [ˈʃavs] *chaves* 'keys'
b. [ˈpahtis] ~ [ˈpahts] *partes* 'parts'
c. [ˈklubis] ~ [ˈklubs] *clubes* 'clubs'

He found that the high front vowel did not occur in 67% of the cases. He examined cases of word-final consonantal clusters, as in [ˈtɔɾaks] tórax 'chest' in contrast with word-final clusters in plural forms, as in [ˈʃɛks] *cheques* 'cheques'. No vowel was observed between the word-final cluster in the former group whereas in the latter group zero-vowel alternation occurred. He found a very high rate of [ks] (96%) which seems to indicate the sound change is towards completion at least for [ks] sequences. Other works had also reported high rates of consonant-sibilant sequences word-finally in BP [16,17].

He expected to find evidence for fine phonetic detail in cases where vowel weakening and loss took place. However, no evidence to it was found which may be attributed to the fact that the phenomenon is well advanced. This result may suggest that phonetic stability may be achieved when a sound change is fully implemented.

2.3. Word-internal consonantal clusters

Leite [17] and Ivo [18] examined cases where an unstressed high vowel was weakened and lost in pretonic position. In both studies a sibilant was adjacent to the high vowel. In Leite's [17] work the sibilant followed the high front vowel as in (6a-b) and in Ivo [18] the sibilant preceded the high vowel, as in (6c-d).

(6) Word-internal consonant clusters
a. [kõdˈsõɪs] 'conditions'
b. [batˈzadʊ] 'baptized'
c. [kapasˈtadʊ] 'capable'
d. [motosˈklɛtə] 'motorbike'

In (6a-b) a consonant-sibilant cluster occurs whereas in (6c-d) a sibilant-consonant cluster occurs following unstressed high vowel loss. Whereas the absence of a vowel was observed in 62% of cases where a consonant-sibilant cluster occurred (6a-b), a rate of 33% lack of a vowel was observed when a sibilant-consonant cluster occurred (6c-d). Although there might be methodological differences, it appears that consonant-sibilant clusters are favoured when compared to sibilant-consonant clusters.

Ivo [18] experimentally analysed sibilant-consonant clusters, i.e., [sC] and heterosyllabic clusters where a sibilant was the first consonant. She considered the duration of the sibilant-consonant sequences. She found that consonant-sibilant clusters which emerge from high vowel deletion, as in [ãtesiˈpadə] ~ [ãtesˈpadə] *antecipada* 'antecipated', have a greater duration than sibilant-consonant clusters which are traditional in the language, as in [ispaˈsadə] *espaçada* 'spreaded out'. Her results show that although in both cases a sibilant-consonant occurs, they are in fact different with regards to fine phonetic detail as sibilants are longer in emergent sibilant-consonant clusters than in similar clusters that do not result from vowel loss.

2.4. Word-initial consonantal clusters

Cristófaro Silva and Freitas [19] analysed word-initial sC consonantal clusters. Consider (7).

(7) Word-initial consonants
a. [isˈpasʊ] ~ [ˈspasʊ] *espaço* 'space'
b. [isˈtradə] ~ [ˈstradə] *estrada* 'road'
c. [isˈkɔlə] ~ [ˈskɔlə] *escola* 'school'
d. [isˈpɾeɪ] ~ [ˈspɾeɪ] *spray* 'spray'
e. [isˈtɛs] ~ [ˈstɛs] *stress* 'stress'
f. [isˈkeɪtʃi] *skate* 'skate'

They observed that a vowel may or may not occur in both cases. The lack of a vowel presented higher rates (75.8%) in cases of loans as in (7a-c) and the presence of a vowel was favoured (39%) in cases as in (7d-f).

Experimental data showed that when a vowel is pronounced it is shorter in loans (7.1%) than in native words (9.7%). Their results also indicated that sibilants have a longer duration in loan words when a vowel does not occur (16.8%), than when a vowel was produced (14.2%). A similar result was observed in native words where sibilants displayed a greater duration when a vowel was not produced (16.9%), than when a vowel was produced (13.9%).

In sum, their results showed that sibilants in sC-clusters are longer when the vowel is not produced (16.8% and 16.9%), than when the initial vowel is produced (14.2% and 13.9%), and this result was statistically significant. Furthermore, the different durational properties of vowels in loan and native words, as well as the lengthening of the sibilant when the vowel does not occur are evidence for the relevance of phonetic detail in the implementation of sound changes.

3. RESULTS

The works reported in section 2 considered cases where an unstressed high vowel may or may not occur. When the vowel is suppressed an emergent syllable type occurs.

These emergent syllable types are not considered as part of the typological inventory of PB syllables. They are understood as a consequence of phonological processes and thus not part of phonological representations. We challenged this view and intent to show that closed syllables are becoming a common pattern in BP. Based on Exemplar Models principles [20, 21, 22] we suggest that fine phonetic detail and specific paths of sound changes lead BP towards emergent syllable types.

The studies we considered had a sibilant adjacent to the unstressed high vowel, except in the case of word-final consonants which could have any BP consonant adjacent to the high front vowel. Although sibilants may offer the appropriate articulatory and acoustic conditions to accommodate vowel loss [9] and conform to the segmental limitations on complex codas, we claim that unstressed high vowel loss reflects a pathway of change that affects BP phonology in general. What triggers the loss of the vowel is its intrinsic short duration when compared to other vowels in general [23]. Our argument is strengthened as the high back vowel [u], which also has an intrinsic short duration, may be weakened and lost in BP as in [busˈkãdʊ] ~ [ˈbskã.dʊ] *buscando* 'searching' and [hobusˈtes] ~

[hobsˈtes] *robustez* ‘robustness’ [24]. We suggest that the combination of the short duration of high vowels together with the lack of stress offer the conditions for vowel weakening and loss to take place in BP. And the emergent consonantal clusters converge with the phonotactic typology of the language, since Portuguese is complex-onset and complex-coda language, although the latter with strict segmental limitations.

Furthermore, we claim that the accountability of fine phonetic detail in the implementation of those sound changes promote the emergence of new syllable types – as for example several word-final consonants, expansion of Cs syllables (cf. (3)), word-medial Cs and s.C and sC word-initially. It might be the case that in the final stages of the implementation of a sound change - for example the 96% rate for word-final Cs sequences we discussed by Soares [15] – fine phonetic detail is accommodated within the system and less variability is attested. This would account for why once a sound change is completed previous generations do not have access to it. Once a sound change is completed lexical representations are permanently affected since sound changes cannot be undone [20: p. 59]. This view is in accordance with Exemplar Models which claims that fine phonetic detail is part of phonological representations [20]. It also accommodates the claim that phonological development follows from pathways of sound changes. All case studies we presented showed that the pathways of sound change are phonetically and lexically gradual [20, 21, 22].

A final word has to be said about vowel loss in EP and BP. In the beginning of this paper we pointed out that EP has an advanced process of vowel weakening and loss which triggers several consonantal clusters to occur [6, 7, 8]. Interestingly, although EP and BP are closely related language varieties, the emergent clusters may be similar in terms of their segmental content or syllable types, however the lexical outcome may be different. This is because the vowel which is weakened and lost in EP is /e/ which is phonetically manifested as [ɨ]. In BP the vowel which is weakened and lost is /i/ (and /e/ which undergoes vowel raising to [i]) [3].

Consider, for example, the word *tessitura* ‘tessiture’. In EP the leftmost vowel may be deleted yielding [tsiˈturə] ~ [tɨsiˈturə]. In BP, the vowel that may be deleted is /i/ so that [tesˈturə] ~ [tesiˈturə] occurs. However, word-finally the same outcome happens in both languages. A word as *chave* ‘key’ will be pronounced as [ˈʃav] in EP and BP. Alternative pronunciations with a word-final vowel are: EP [ˈʃavɨ] and BP [ˈʃavi].

The facts discussed in this section provide us with evidence that a combination of phonetic and phonological factors trigger the emergence of new syllable types in BP. Unstressed short vowels are subject to weakening and loss in EP and BP. Additionally, an adjacent sibilant appears to contribute to vowel loss in BP, which can be due to articulatory, acoustic and phonotactic reasons. We have seen that vowel loss is implemented gradually, where fine phonetic detail plays a role in the implementation of the sound change. From a phonological perspective, the unstressed vowels /e/ in EP and unstressed /i/ in BP have exemplars in competition and the most common exemplar becomes more robust so that emerging syllable types appear as a consequence of vowel loss.

4. CONCLUSIONS

This paper considered the emergence of syllable types in BP. Open syllables represent 78% of syllable types whereas closed syllables number 22% of cases [1]. We challenged this result and showed that complex syllables are becoming the most common pattern in the language. The syllable patterns which come from high front vowel deletion all have a sibilant. We showed that independent phonological processes which trigger unstressed high front vowel weakening and loss in BP may affect syllable structure yielding to the emergence of new syllable types.

Our claim is supported by the analysis of different phonological processes which were empirically examined. We showed that the emergence of new syllable types is implemented phonetically and lexically in a gradual fashion.

We also showed that emergent syllable types in BP may be similar to those from EP. However, as the vowel which is weakened and lost in each language differs, the outcome for each language may vary, although pronunciations would merge in some instances. We argue that phonetic and phonological properties must operate together in order to promote the emergence of new syllable types.

5. ACKNOWLEDGEMENTS

Thaïs Cristófaro Silva acknowledges support from CNPq 311934/2020-0 and Maria Mendes Cantoni acknowledges the support from FAPEMIG, grant number APQ-02997-18.

6. REFERENCES

[1] Viaro, M. E; Z. de O. G. Filho. (2007). Análise quantitativa da frequência dos fonemas e estruturas silábicas portuguesas. Estudos Linguísticos 36, 27-36.

[2] Câmara Jr., M. (1970). Estrutura da Língua Portuguesa [The Portuguese language structure]. Petrópolis: Vozes.

[3] Cristófaro Silva, T. (2017). *Fonética e fonologia do português: Roteiros de estudos e guias de exercícios* [Portuguese phonetics and phonology: Study guides and guidelines for exercises]. São Paulo: Editora Contexto.

[4] Collischonn, G. (2000). A epêntese vocálica no português do sul do Brasil: Análise variacionista e tratamento pela Teoria da Otimalidade [Vowel epenthesis in Southern Brazilian Portuguese: Variationist and Optimality Theory analyses]. *Letras de Hoje*, 35(1), 285–318.

[5] Cantoni, M. (2015). A epêntese no português brasileiro em uma perspectiva multirrepresentacional [A multirepresentational approach to epenthesis in Brazilian Portuguese]. *Gragoatá*, 1(38), 231–246.

[6] Mateus, M. H. M.; E. d'Andrade. (2000). The Phonology of Portuguese. Oxford University Press.

[7] Miguel, M. A. (1993). Os padrões das alternâncias vocálicas e da vogal zero na fonologia portuguesa. Unpublished PhD Thesis. Universidade dos Açores, Ponta Delgada.

[8] Miguel, M. A. (2003). As estruturas silábicas e a redução vocálica no Português Europeu. Revista de Estudos da Linguagem, 11, 95-118.

[9] Albano, E.; F. Meneses. (2015). From Reduction to Apocope: Final Poststressed Vowel Devoicing in Brazilian Portuguese. Phonetica, 72, 121-137.

[10] Assis, A. (2017). A emergência de consoantes finais no Português Brasileiro na microrregião de Araguaína/Tocantins Unpublished PhD thesis, UFMG. Belo Horizonte.

[11] Dias, E.; I. Seara. (2013). Redução e apagamento de vogais átonas finais na fala de crianças e adultos de Florianópolis: uma análise acústica. Letrônica, 6, 71-93.

[12] Maddieson, I. (1985). Phonetic cues to syllabification," In V. Fromkin (ed.) Phonetic Linguistics. Orlando: Academic Press. 1985, 203-21.

[13] Cristófaro Silva, T. (in press). Word-final onsets: A Brazilian Portuguese case study. In F. Breit, Y. Yoshida, & C. Youngberg (Eds.), Elements, Government and Licensings: Developments in phonology. UCL Press.

[14] Vieira, M. J. B.; Cristófaro Silva, T. (2015). Redução Vocálica em Postônica Final. Revista da ABRALIN, 14, 379.

[15] Soares, V. (2016). Encontros consonantais em final de palavra no português brasileiro. Unpublished master's dissertation. Belo Horizonte: Universidade Federal de Minas Gerais.

[16] Cristófaro-Silva, T.; Mendes-Jr., W. (2022). Plural formation in English: a Brazilian Portuguese case study. In: *Second Language Pronunciation: Different Approaches to Teaching and Training*. De Gruyter Mouton.

[17] Leite, C. (2006). Seqüências de (oclusiva alveolar + sibilante alveolar) como um padrão inovador no português de Belo Horizonte. MA dissertation. Universidade Federal de Minas Gerais, Belo Horizonte.

[18] Ivo, A. (2021), O apagamento da vogal alta anterior pretônica em sequências [si].C. MA dissertation. Universidade Federal de Minas Gerais, Belo Horizonte.

[19] Cristófaro Silva, T.; M. Freitas. (2020). sC-clusters in Brazilian Portuguese. Journal of Portuguese Linguistics, 19, 1-17.

[20] Bybee, J. (2001). Frequency effects on French liaison. *Typological studies in language*, *45*, 337-360.

[21] Bybee, J. (2010). Language, Usage and Cognition. Cambridge: Cambridge University Press.

[22] Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.) Frequency effects and the emergence of lexical structure. John Benjamins, Amsterdam. 137-157.

[23] Keating, P. (1985). Universal phonetics and the organization of grammars", *Phonetic Linguistics*, ed. V. Fromkin, Academic Press, 115-132.

[24] Souza, R. (2014). The emergence of syllable structure? Data from gradient vowel reduction in Brazilian Portuguese. In Proceedings of the 10th High Desert Linguistics Society (103–118). Albuquerque: High Desert Linguistics Society.

TRACKING CREAK FROM EARLY TO LATE ADULTHOOD: A PANEL STUDY FROM THE NORTH EAST OF ENGLAND

James Grama, Mirjam E. Eiswirth, & Isabelle Buchstaller

Sociolinguistics Lab, University of Duisburg-Essen
james.grama@uni-due.de, mirjam.eiswirth@gmail.com, isabelle.buchstaller@uni-due.de

ABSTRACT

This study investigates how creak and F0 change across the life-span in a panel of six English speakers from the North East of England, recorded between two and three times. Across timepoint, we observe a general decrease in F0 for all but one speaker. However, creak follows complex patterns as speakers age. In early adulthood, all speakers exhibit creak, and show evidence that creak can be recruited to indicate turn-finality. Towards the end of their working lives, speakers reduce their proportion of creak dramatically, and show little evidence of its previous turn-yielding function. In older age, creak reappears, and a subset of speakers show renewed use of its turn-yielding function. These findings suggest that both global and interactional uses of creak change over the life-span, and suggest that—for some communities—creak may be age-graded.

Keywords: lifespan change, panel study, creaky voice, F0, North East of England.

1. INTRODUCTION

This paper explores the use of creaky voice across a small panel sample of speakers. As a highly multiplex phonation type, creak is contingent on both physiological and social-interactional factors. Here, we explore its use by six English speakers from the North East of England who were interviewed once in early adulthood, and either one or two times in older age. We ask: (1) do we observe age-related changes in fundamental frequency (F0) and creak; and (2) do speakers vary in their use of creak as an interactional strategy to signal turn-transition?

Our analysis suggests that in early post-adolescent life, speakers make frequent use of creak, especially at the end of turns; however, the rate of creak and its turn-final use decrease across the middle life-span. Following retirement, speakers show a general resurgence in creak, both as a global feature and especially as a turn-final indicator.

1.1. Life-span change and creak

While linguistic theory tends to rely on the observation that speech patterns are relatively fixed by adulthood [1], panel studies (which follow the same individual over time) have revealed post-adolescent malleability (see [2]). Voice quality would appear to lend itself particularly well to panel studies, given that it has been characterized as a semi-permanent feature of individuals [3], despite the age-related changes that affect laryngeal structures (e.g., vocal fold atrophy [4], histological changes [5], and changes in glottal competence [6]). In elderly populations, these physiological changes partly motivate increased irregularity in F0, jitter, and shimmer (e.g., [7]). Particular focus has been placed on changes in F0, which has been found to decrease as speakers age, followed by an increase in elderly men, and a decrease in elderly women [8-11]. Given the relationship between low F0 and creak [12], we might expect creak to parallel changes in older age; however, we are not aware of any panel research that explores changes in creak across the adult life-span.

1.2. Social and stylistic functions of creak

Creak has shown to be highly multifunctional in conversation, indexing both social information about the speaker, as well as doing interactional work [13]. Creak's frequent occurrence in turn-final position in English and Finnish suggests a turn-completion signalling function [14-15]. As a global feature, creak is both used in greater proportion by [16-20] and heavily associated with young, urban, college-educated women [16, 20-21]. From the perspective of language change, there are two possible explanations for this pattern: (1) we could be seeing a change in progress led by this group (as suggested by [16] for English); or (2) creak could be an age-graded feature. The latter hypothesis finds some support in the distribution of creak in early studies [22-23], which similarly find that creak is used more by young women. This suggests the possibility of a repeated pattern at the individual level, while the community as a whole remains stable.

What is thus needed is a panel study that explores changes in both the extent of overall creak, as well as how speakers operationalise creak in interaction across the life-span. We begin to address these questions here, focusing on malleability in F0 and in

creak, both globally and as an interactional resource across the adult life-course of six individuals.

2. METHODS

2.1. Panel data-set

The data come from sociolinguistic interviews conducted with the same participants in 1971, 2013, and 2019. These speakers (see Table 1) offer a rich picture of the linguistic choices made by individuals as they transition from young adulthood to the end of their professional careers and into older age; we argue that these trajectories carry more weight than treating age as purely chronological [24-25].

Sex	Speaker	YOB	T1	T2	T3
F	Edith	1939	32	74	--
	Nelly	1942	29	71	78
	Anne	1948	23	65	71
M	Aidan	1946	25	67	--
	Rob	1948	23	65	--
	Fred	1950	21	63	69

Table 1: Panel participants and ages by time point.

2.2. Preparing the data and delineating a 'turn'

Interviews were transcribed and time-aligned using ELAN [26]; speech intervals were divided into sections delineated by pauses, roughly falling into 'breath groups'. Transcriptions were force aligned using LaBB-CAT [27], a process that created smaller .wav files based on aligned intervals (n=10,012).

We make use of the conversation analytical concept of a "turn" [28-29], which we define (following established practice) as a spate of talk produced by one participant when they are holding the floor. Because turns are delimited following interactional parameters (i.e., floor transition), their length is not predetermined, though they tend to contain no more than a few turn-constructional units [30]. We restrict our analysis to turns where floor transition is unproblematic (i.e., non-competitive [31]; n=2,156 turns).

2.3. F0 extraction and creak assessment

Though creak is notoriously multiplex [32], Keating et al. [12] argue that one of its prototypical features is low or irregular F0. Without intending to wash away inherent complexities, we operationalize low F0 here as a correlate of creak. This approach has been reliably implemented in [33], who use REAPER (the Robust Epoch and Pitch EstimatoR; [34]) to identify likely creak. REAPER produces estimations of Glottal Closure Instants (GCIs), from which F0 can be estimated by taking the reciprocal of the duration between adjacent GCIs. REAPER's default settings were used, but to prioritize detection of low F0, minimum expected F0 was lowered from 40 to 20 Hz. We excluded overlapped segments, resulting in 2,006,501 F0 measurements across 15 recordings.

Creak was estimated using binary categorization of F0 measurements (see [33]). The mode of F0 values was calculated for each recording, as well as the most prominent mode below this mode; the antimode between these two modes was calculated, and all values below this antimode were coded as creaky. Given reported age-related changes to F0, this categorization was done for each recording to produce individualized pictures of a speaker's F0 and creak at each timepoint. Of the total F0 measurements, 95.2% (n=1,910,850) were classified as modal, and 4.8% as creaky (n=95,651).

We present three analyses: (i) of changes to F0, (ii) of global creak, and (iii) of creak over the duration of a turn. While claims can certainly be drawn from group-aggregated data, we echo [35] in arguing that such analyses can hide important speaker-specific changes which bear on life-span changes. Therefore, our analysis of F0 includes linear mixed-effects models fit to each speaker. The dependent variable was modal F0 (excluding creaky tokens), with timepoint as a predictor, and turn and word as random intercepts (Section 3.1). Changes to creak across the life-span and the turn (Sections 3.2-3.3) were modelled using logistic mixed-effects models fit to each timepoint separately, due to uneven representation of creak across recording. These models included an interaction between speaker and relative position in the turn, with a random slope of turn by relative position in the turn. Alpha for all models was 0.05.

3. RESULTS

3.1. F0 over the life-span

We turn first to an analysis of how modal F0 has changed over the life-span. First, we note a decrease in F0 across timepoint for all speakers (Table 2), with greater changes observed between T1 and T2 (where there is a larger time-gap) than between T2 and T3. Of the speakers with a third timepoint, only Anne shows significant F0 lowering between T2 to T3. In contrast to findings from [9-11], men in our sample show no evidence of an F0 increase in later life. Also, women tend to show numerically larger life-span changes in F0 than men; in fact, Rob shows no significant change across timepoint at all. Thus, while F0 decreases with age, the extent of these changes is highly speaker- and life-stage specific.

If we plot all F0 measurements (Figure 1), we observe a bimodal distribution for F0: women show a primary peak between 150-220 Hz, while men exhibit one between 80-130 Hz. We also note a second, low peak that represents the F0 of identified creaky tokens (between 80 and 100 Hz). This secondary peak shows considerable intra-speaker variance, being clear for some, and absent for others. This is early evidence that speakers show different extents of creak across their life-spans.

	Estimate (T1-T2)	Estimate (T1-T3)	Estimate (T2-T3)
Edith	-13.40 (***)	--	--
Anne	-27.47 (***)	-35.46 (***)	-7.99 (***)
Nelly	-48.79 (***)	-50.25 (***)	-1.46
Aidan	-8.39 (*)	--	--
Rob	-2.49	--	--
Fred	-9.93 (***)	-11.99 (***)	-2.06

Table 2: Estimated F0 differences (by speaker) across timepoint: (***)=p<0.001, (*)=p<0.05.

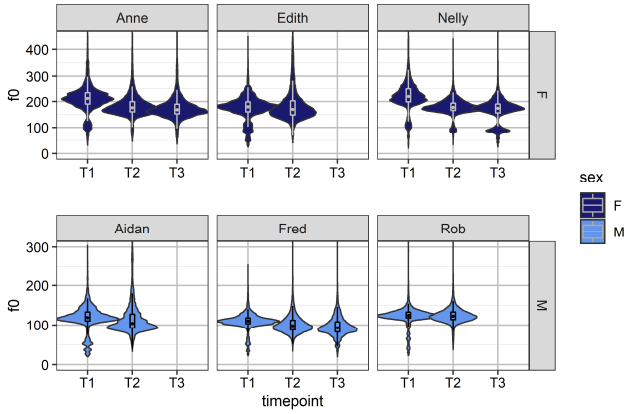


Figure 1: Distribution of F0 by speaker and timepoint.

3.2. Change in creak over time

Figure 2 shows the proportion of creak across time point and speaker. It demonstrates two things. First, while creak rates are generally low, all cohort members produce creak to some extent at T1, followed by a decrease in the amount of creak at T2. Three speakers (Edith, Fred and Aidan) produce no creak at all at T2. Second, all speakers with a third timepoint register an increase in their proportion of creak from T2 to T3. As with changes to F0, speakers vary in the extent to which their rates of creak increase; Anne and Fred show modest increases, while Nelly's rate rises to about 18% of all GCIs.

In the context of variability across the life-span, these findings suggest that speakers tend to exhibit more creak when they are young and relatively new members of the workforce (at T1), followed by a decrease in creak at or near the age of retirement (at T2). At T3, when speakers are well-and-truly retired from the workforce, they show an increase relative to T2. This pattern maps closely on to the U-shaped curve typical of age-graded variability, in which a middle-aged trough is followed by a tail (see [36]).

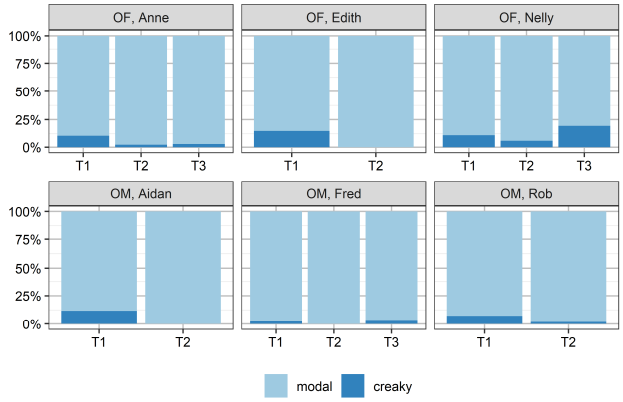


Figure 2: Proportion of GCIs coded as creaky (dark) vs. modal (light) across speaker and timepoint.

3.3. Creak across the turn

We now turn to an investigation of *where* creak is deployed in the turn, and whether this, too, changes over the life-span. Figure 3 shows the marginal effects from the models fit to timepoint, organized by speaker; normalized relative turn position is on the x-axis, with predicted creak on the y-axis.

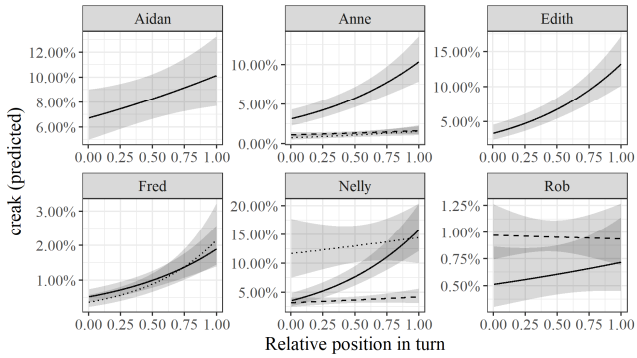


Figure 3: Predicted creak across time-normalized position in the turn for T1 (solid), T2 (dashed), and T3 (dotted); y-axis scaled to show individual behavior.

First, these findings corroborate that speakers exhibit vastly different overall rates of creak across timepoint. Second, creak use is mediated by turn position and timepoint. At T1, all speakers show a positive relationship between creak likelihood and turn position—that is, creak is more likely to occur at the end of a turn. However, the strength of this effect varies by speaker. All speakers show a significant effect of turn position (Table 3), except for Rob, whose use of creak is also relatively low. At T2, we observe a drastic change that corresponds with the global decrease of creak; no speakers exhibit

a meaningful relationship between turn position and the likelihood of creak, even those with turn-mediated creak at T1. T3 reveals more heterogeneous patterns. Fred shows a resurgence of creak, both overall and in turn-final position. Anne, who showed creak at T2 with no effect of turn-position, exhibits the effect anew at T3, although less strongly than at T1. And Nelly, who mirrored Anne at T1 and T2, shows a significant global increase in creak at T3, but no relationship between creak and turn position.

	T1	T2	T3
Edith	1.493(***)	--	--
Anne	1.273(***)	0.380	0.740(*)
Nelly	1.632(***)	0.283	0.246
Aidan	0.450(*)	--	--
Rob	0.341	-0.036	--
Fred	1.311(***)	--	1.806(***)

Table 3: Estimated log-odds of turn position on creak; Šidák corrected *p*-values: (***)=$p<0.001$, (**)=$p<0.01$, (*)=$p<0.05$.

4. DISCUSSION

We can now start to paint a picture of how F0 and creak vary over the life-span in this panel of older speakers. First, we provide corroborating panel evidence that F0 does indeed lower in older age (cf. [8-11]), though without the rise shown by elderly men. While this could be explained by individual differences across studies, it is more likely that our men are simply not old enough to show this effect (all are at least 8 years younger than the mean age of elderly men in [9]). Second, while [3] argues that voice quality is a quasi-permanent feature of a person's speech, the current panel speakers exhibit differentiated and variable creak over their lifespans, behavior which appears to follow the classic U-shaped curve associated with age-grading in sociolinguistic theorizing [36]. High rates of creak among the women in the sample, especially at T1, supports research that links creak with young women's speech (despite studies also placing a disproportionate focus on this very group; see [20]). The decrease in the proportion of creak we observe at T2, thus, could be driven by increased pressures to avoid some of the indexicalities of this feature while the uptick at T3 would represent a relaxation of these pressures (cf. the lessening of marketplace pressures in old age; [35]).

Finally, creak's behavior throughout the turn provides additional insight into its function. At T1, when speakers are relatively young and just beginning their career trajectories, all show an increased likelihood for creak to occur at the end of turns. Presumably and as argued in [14-15], this is evidence of creak's turn-yielding function. At T2, however, this use all but disappears, paralleling the overall decrease in global creak. Again, it is possible that this change is due to the indexical association of creak as "young" and as a feature of women's speech (see [16]). Two of the three speakers who creak at all at T2 are women (Anne, Nelly), and all other speakers (except Rob, whose rates are low) avoid creak categorically. At T3, speakers exhibit variable patterns, with Anne and Nelly showing a resurgence in their use of creak, but in different ways. Anne demonstrates little change in overall proportion of creak, but makes renewed use of its turn-yielding function, while Nelly drastically increases her rate of global creak, which numerically but not statistically rises across the turn. Fred, by contrast, reproduces his behavior at T1, seemingly deploying creak as he once did after abandoning it at T2. These complex patterns indicate that creak is recruited for social as well as interactional uses which change over the life-span and suggests that changes in creak cannot be solely explained by changes in motor-control.

It is, of course, important to note the limitations of our study. REAPER has been shown to be a robust sociophonetic tool, but it is also a coarse-grained one [33]. Because our study relies heavily on the unsupervised identification of creak, it is likely that some differences exist between human coded creak and creak as coded by REAPER. In addition, the methodology applied here does not account for creak that is *not* characterized by low F0 (e.g., aperiodic creak; [12]). Verification of the observed patterns is therefore necessary to gain higher-resolution insight into how creak is operationalised across the life-span. In spite of these caveats, we believe our results point to the possibility that creak is partly an age-graded phenomenon.

5. ACKNOWLEDGMENTS

We acknowledge: funding from the DFG (#BU-2902/7-1); our transcribers and interviewers; the Newcastle DECTE team for providing access to the original data; two anonymous reviewers for their helpful feedback; and Katharine Dallaston for her assistance with REAPER. We are greatly indebted to our panel speakers for sharing their time and speech.

6. REFERENCES

[1] Lenneberg, E.H. 1967. The biological foundations of language. *Hospital Practice* 2(12), 59-67.

[2] Sankoff, G., & Blondeau, H. 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83(2), 560-588.
[3] Abercrombie, D. 1967. *Elements of general phonetics*. Edinburgh University Press.
[4] Honjo, I. & Isshiki, N. 1980. Laryngoscopic and voice characteristics of aged persons. *Arch Otolaryngol* 106(3), 149-150.
[5] Kahane, J.C. 1987. Anatomic and phyiologic changes in the aging peripheral speech mechanism. In *Aging: Communication Processes and Disorders*. Grune & Stratton.
[6] Vaca, M., Cobeta, I., Mora, E., Reyes, P. 2017. Clinical assessment of glottal insufficiency in age-related dysphonia. *J. Voice* 31(1).
[7] Decoster, W. & Debruyne, F. 1997. Changes in spectral measures and voice onset time with age: A across-sectional and a longitudinal study. *Folia Phoniatr Logop*, 49, 269-280.
[8] Harrington, J., Palethorpe, S., & Watson, C.I. 2000. Does the Queen speak the Queen's English? *Nature* 408(6815), 927-928.
[9] Dehqan, A., Scherer, R.C., Dashti, G., Ansari-Moghaddam, A., Fanaie, S.. 2012. The effects of aging on acoustic parameters of voice. *Folia Phoniatr Logop*, 64, 265-270.
[10] Reubold, U., Harrington, J., & Kleber F. 2010. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication* 52(7-9), 638-651.
[11] Reubold, U., & Harrington, J. 2017. The influence of age on estimating sound change acoustically from longitudinal data. In *Panel Studies of Variation and Change*. Routledge.
[12] Keating, P., Garellek, M., & Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. ICPhS*, 2-7.
[13] Mendoza-Denton, N. 2011. The semiotic hitchhikers guide to creaky voice: Circulation and gendered hardcore in a Chicano/a gang persona. *J. Linguistic Anthropology* 21(2), 261-280.
[14] Monsen, R.B., & Engebretson, A.M. 1977. Study of variations in the male and female glottal wave. *J. Acoust. Soc. Ame.* 62(4), 981-993.
[15] Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the IPA* 31(1), 139-152.
[16] Yuasa, I.P. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3), 315-337.
[17] Wolk, L., Abdelli-Beruh, N.B., & Slavin, D. 2012. Habitual use of vocal fry in young adult female speakers. *Journal of Voice Foundation* 26(3), 111-116.
[18] Podesva, R. 2013. Gender and the social meaning of non-modal phonation types. *Proc. 37th Annual Meeting of the BLS*, 427–448.
[20] Dallaston, K. & Docherty, G. 2020. The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature. *PLoS One*, 15(3).
[21] Callier, P. & Podesva, R. 2015. Voice quality and identity. *Ann Rev Applied Ling* 35, 173-194.
[22] Linke, C.E. 1953. A study of pitch characteristics of female voices and their relationship to vocal effectiveness. PhD dissertation. The University of Iowa.
[23] Austin, W.M. 1965. Some social aspects of paralanguage. *Canadian J. of Linguistics* 11(1), 31-39.
[24] Eckert, P. 1996. Age as a sociolinguistic variable. *The Handbook of Sociolinguistics*. OUP.
[25] Pichler, H., Wagner, S.E., Hesson, A. Old-age language variation and change: Confronting variationist ageism. *Language and Linguistics Compass* 12(6), e12281.
[26] Lausberg, H., & Sloetjes, H. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behav Res Methods* 41(3), 841-849.
[27] Fromont, R., & Hay, J. 2012. LaBB-CAT: An annotation store. In *Proc. Australasian Language Technology Association Workshop*, 113-117.
[28] Sacks, H., Schegloff, E.A., & Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696-735.
[29] Schegloff, E.A. 1995. Discourse as an interactional achievement III. The omnirelevance of action. *Res Lang Soc Interact* 28(3), 185-211.
[30] Ford, C.E., 1993. *Grammar in interaction.* CUP.
[31] Local, J., Kelly, J., & Wells W.H.G. 1986. Towards a phonology of conversation: Turn-taking in Tyneside English. *J. Linguistics* 22(2), 411.
[32] Garellek, M. 2019. The phonetics of voice. In *The Routledge Handbook of Phonetics*, 75-106. OUP.
[33] Dallaston, K., & Docherty, G. 2019. Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proc. ICPhS*, 532-536.
[34] Talkin, D. 2019. *REAPER*.
[35] Mechler, J., & Buchstaller, I. Instability in the use of a stable variable. *Linguistics Vanguard Special Issue on Aging.*
[36] Downes, W. 1998. *Language and society, 2nd ed.* Cambridge.

LONGITUDINAL INVESTIGATION OF THE AGING VOICE

Benjamin V. Tucker[1,2], Stephanie Hedges[2], Mark Berardi[3], Eric J. Hunter[4],

[1]Northern Arizona University, [2]University of Alberta, [3]University Hospital Bonn, [4]Michigan State University

benjamin.tucker@nau.edu, shedges@ualberta.ca, Mark.Berardi@ukbonn.de, ejhunter@msu.edu

ABSTRACT

Acoustic and perceptual research has shown that an individual's vocal characteristics change with age. In this paper, we report a later-life longitudinal investigation of three male speakers using publicly available archives of speeches given to large audiences on a semi-regular basis (generally with a couple of years between each instance). The collection of speeches was given during the last 40-50 years of each speaker's life. From each speech, 5-minute samples (recordings and transcripts) were force-aligned to identify word and phoneme boundaries. We investigate non-linear changes in vowel duration, word duration, speech rate, fundamental frequency, and vowel space using Generalized Additive Mixed Models. We describe the vocal changes over the lifespan in the speakers' later years. The current results do not support expected outcomes from previous research and illustrate idiosyncratic and dynamic effects for individuals in later years.

Keywords: speech production, aging, duration, fundamental frequency, vowel space

1. INTRODUCTION

With advancing age, our bodies undergo a variety of physiological and cognitive changes that impact speech [1]. These changes have been shown to impact speech production and speech acoustics metrics. In order to understand how age influences speech, we use longitudinal data from three separate male speakers covering the last 40-50 years of life. This paper investigates the longitudinal effects of age on vowel and word duration, speech rate, vowel space, and fundamental frequency.

Research on speech and aging has found variable effects [2, for a review]. Generally, segment and word duration increases with advancing age [3, 4, 5]. Jacewisz et al. [6] found that reading and speaking rate increased from early adulthood to middle adulthood following which it slowly decreased throughout late adulthood. However, Quene et al., [7] found that the speech rate of one Dutch speaker (Queen Beatrix) between the ages of 42 and 74 years decreased from 42 to 62 years of age and then increased from 62 to 74 years. They argue that it is possible that the increased speech rate in Dutch culture and decades of experience reading the *Troonsrede* may have impacted her speech rate.

Many studies have found changes in fundamental frequency with age [8, 9, 10, 11, 12, 13, 14]. Researchers performing long-term longitudinal investigations have primarily focused on individual speakers' fundamental frequency changes. For example, Harrington found a significant decrease of an average of 21 percent in fundamental frequency for four speakers from an earlier recording to a recording made at a later age [10]. In a more in-depth analysis of the speech of both Queen Elizabeth II and Alistair Cooke, Reubold et al. [11] found that the fundamental frequency of Queen Elizabeth decreased as a function of age in a linear pattern while for Alistair Cooke fundamental frequency decreased until the age of 87, but then increased after 87. Hunter et al. [15] used one of the same subjects as the current study (S1) and found that F0 decreased until the age of 68 where it then increased until the age of 98 years. They attribute this change to general physiological changes. In cross-sectional studies, researchers have found differences in the patterns of male and female participants with a u-shaped result for female participants [13, 14, 3].

The vowel space of aging speakers has also been investigated. This research is also variable with some research indicating an increase in phonetic height (decreased F1) and backness (decreased F2) with age [10, 11, 1, 16]. In a longitudinal analysis of the speech of participants from the UP corpus, Gahl & Baayen [17] examined how the vowel space shifts with aging between early to late adulthood. They found that vowel space shifts more toward the periphery with an increase of age, even when controlling for an increase in vowel duration.

The purpose of this study was to build on previous reports of later-life speech changes. From the perspective of using available recordings and automatic analysis techniques, later-life speech changes could be investigated over a larger period

of time. This approach allows for a larger number of longitudinal samples compared to what could be accomplished in a laboratory setting.

2. METHOD

The present study analyzes regular speeches by three religious leaders to university students that cover a time period ranging from 40 to 50 years. This source includes longitudinal data covering middle adulthood to late adulthood. The speeches were given in the same style keeping the style characteristics consistent across the years recorded.

2.1. Participants

This study analyzes the speech of three male speakers: S1 (born: 1910), S2 (1899), and S3 (1922). Speakers S1 and S3 were born in Utah, and S2 was born in Idaho. The majority of the speakers' lives were spent in the western United States, especially Utah and Idaho. However, both S2 and S3 spent time in the eastern United States living and working in Washington DC, Massachusetts, and New York. Additionally, per lifestyle choice, the speakers reportedly did not consume alcohol or coffee, and they did not smoke; these substances all have been shown to affect voice production and vocal health [18, 19].

2.2. Forced Alignment

Five-minute segments were extracted from each recorded speech and were then annotated orthographically using the linguistic annotator ELAN software [20]. Phonetic alignment of the speech was obtained using the Penn Forced Aligner [21] with no manual correction. We extracted acoustic measures using Praat [22] for 73,305 stressed and unstressed vowels from 51,842 words.

3. RESULTS

We performed a generalized additive mixed models (GAMM) [23, 24] on the extracted data to account for non-linear trends found within the acoustic trajectories over time with speaker and word included as random effects. Four separate GAMMs were run on the data to investigate trends for vowel duration, word duration, speech rate (syllables per second for each recording), and fundamental frequency with advancing `age` as a smoothed predictor and word and speaker as random effects. Convex hulls for each speaker's vowel space were calculated in order to investigate vowel space shifts

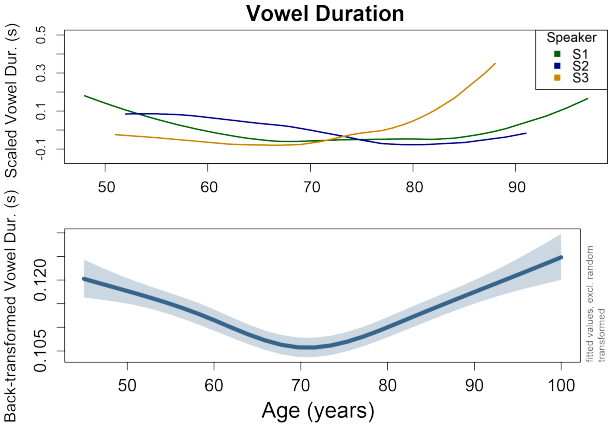


Figure 1: Mean vowel duration overtime for S1, S2, and S3

with age [25, 26].

Figure 1 shows the mean vowel duration of the stressed vowels ($n = 41,359$) for each speaker (top) and a visualization of the GAMM results (bottom). Each speaker shows a unique pattern of mean vowel duration over the later-decades of the lifespan. Vowel duration for S1 decreases to about the age of 68 and then vowel duration starts to increase at the age of 85. While vowel duration for S2 generally decreases to about the age of 80 with a slight increase after that. S3, interestingly, shows a largely flat, with maybe a slight decrease to about the age of 70 followed by a sharp increase in duration, producing the longest vowel durations of all the speakers. The GAMM with duration measures for all three speakers, illustrated in lower part Figure 1, shows that the modeled vowel duration trajectory follows a U-shaped pattern, with vowel duration decreasing until approximately the age of 71 years, where it then increases. The model shows a significant effect of `Age` ($edf = 3.8144, F = 66.4155, p < 0.0001$).

The average word duration data by speaker is plotted at the top of Figure 2 and the effect of `Age` in the GAMM model is found at the bottom of Figure 2. The top part of the figure illustrates the individual variation present for each speaker. While the word duration begins at about the same point in their early 50s, S1 decreases slightly, then increases slightly where word duration then levels off for the last decade and a half. For S2, however, there is a slight increase after which there is a U-shaped pattern of a decrease followed by an increase occurs. The trajectory for S3 shows a relatively stable duration until his mid-70s, where it then sharply increases. The bottom part of Figure 2 summarizes the result of a GAMM with Age as the primary predictor of word duration for all speakers. The model indicates that as speakers age word duration follows a U-shaped

pattern, similar to that which was found for vowel duration. From the ages of approximately 48 to 70 years, word duration is flat or decreases. Following the age of 70 years, words are produced with longer durations as speakers age ($edf = 3.9258, F = 81.3261, p < 0.0001$).

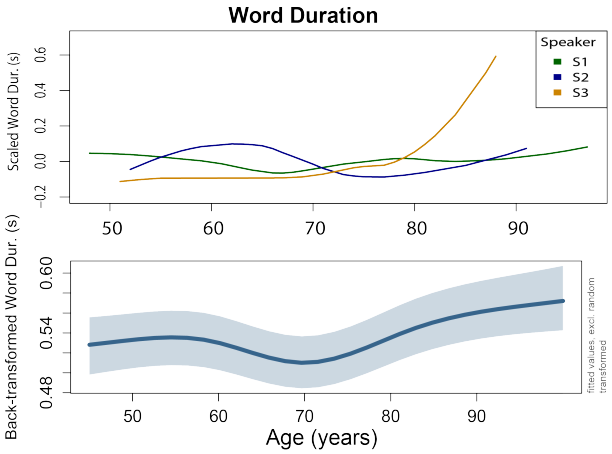


Figure 2: Word duration over time for S1, S2, and S3 and the results of a GAMM analysis of word duration (s) over time.

The average speech rate for each speaker (the number of vowels divided by the total duration of the recording) is plotted in the top half of Figure 3. As was the case with both vowel and word duration, the average speech rate pattern is unique to each individual. With S1, speech rate increased until around the age of 68 years where it then seemed to level off until the age of approximately 81 years, where it then decreased. S2 followed a somewhat similar trajectory with an increase of speech rate until the age of around 80 years, followed by a decrease. The speech rate of S3 increased until the age of 68 years, when rate decreases sharply. The pattern for all three speakers is generally consistent; however, the age at which the decrease in speech rate begins varies across the three speakers. We generated a GAMM for speech rate using data from all three speakers again focusing on the effect of Age. In the model, the trajectory of speech rate was found to increase until the age of 70 years, where it then decreases ($edf = 2.983, F = 2.84, p < 0.05$). This aligns nicely with the results reported for both vowel and word duration.

The top part of Figure 4 plots the average fundamental frequency of the stressed vowels for each speaker as they age. While a general U-shaped pattern can be seen for each speaker, there are differences between the speakers. Fundamental frequency for S1 decreases slightly until the late 70's, then gradually increases. Fundamental

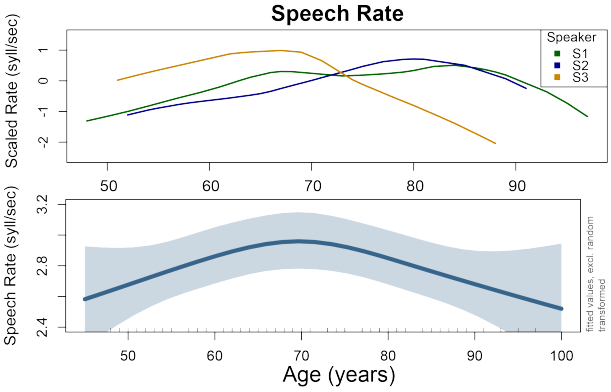


Figure 3: Changes in speech rate (syllables/second) for S1, S2, and S3 with advancing age (top) and Results of a GAMM analysis of speech rate over time (bottom)

frequency for S2 actually increases from 52 years to 59 years, then decreases until the mid 70's where, like S1, it begins to increase. S3 decreases very slightly until the early 70's and is followed by the sharpest increase out of all the speakers. As can be seen in Figure 4, modeled fundamental frequency follows a significant U-shaped pattern, decreasing from the age of around 50 years to 70 years, where it then follows an increasing trend with advancing age ($edf = 3.997, F = 298.528, p < 0.0001$).

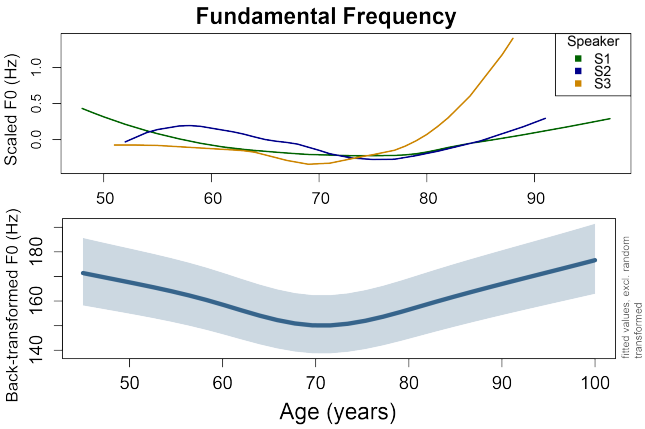


Figure 4: Changes in fundamental frequency for S1, S2, and S3 with age and GAMM of fundamental frequency for all speakers over time.

In Figure 5 (top left) we see S1's vowels space and how it changes over time. In this case, the vowel space shifts to lower F1 and F2 values. Changes in the vowel space of S2 can be seen in Figure 5 (top right). S2 also follows a similar trend as S1 although the shift is much less extreme. We note however that for S2 in his 50s there seem to be some surprisingly high F2 values. At the bottom of Figure 5, S3's vowel space is plotted, indicating that there was no shift to a decreased F1 but there may be a slight lowering of F2 over each decade. The vowel space

area for speakers remains relatively constant with age (e.g., 1255 to 1366 for S1).

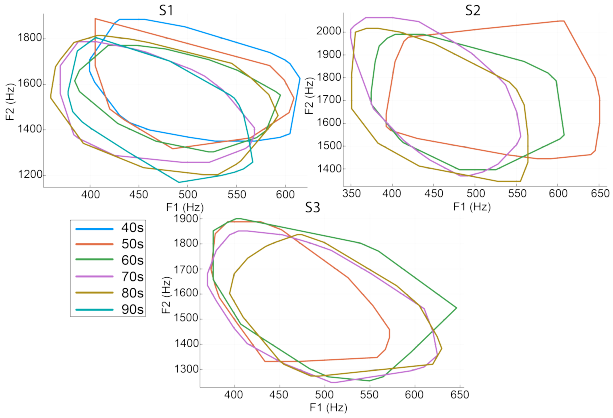


Figure 5: Overlaid vowel plots reflecting production for all three speakers over time. Each decade of the speaker's life is represented by a convex hull [27].

4. DISCUSSION AND CONCLUSION

This paper investigates the effects of age in three male speakers on segment and word duration, speech rate, fundamental frequency, and vowel space. Our data is longitudinal, controlling for potential confounding variables that arise with cross-sectional studies, such as individual differences. Our findings show that the acoustic changes that occur with age are dynamic and idiosyncratic. They are dynamic in that the acoustic changes do not follow a linear trend, but are more U-shaped (or upside-down U-shaped) with age. The results also indicate that each speaker presents individual patterns in their acoustic data.

Our results of vowel and word duration build on the previous literature. These results align with the literature in that segment and word duration is longer in late adulthood than in young adulthood. However, it adds to previous literature by suggesting that the trajectory follows a U-shaped pattern and not a strictly linear pattern.

There are interesting similarities between the trajectories of vowel duration and word duration. Both trajectories follow a U-shaped pattern where the segments and words are produced faster until around the age of 70 years when both start to be produced slower with advancing age. The relation of the two patterns makes sense as vowel duration ultimately affects whole word duration.

It is also important to note the similarities found between the two durational GAMMs and the speech rate GAMM. The two measures are related as segment and word duration influence speech rate. As vowel and word duration shorten, speech rate increases, and vice versa, as vowel and word duration lengthen, speech rate decreases as less syllables are able to be produced per second. In this study, the speech rate refers to specifically read speech given in a talk. Different results are expected in the cases of spontaneous speech or speech produced from a different elicitation tasks.

Speech rate follows an inverted U-shaped trajectory expected from results found by [6] who also found an inverted U-shaped pattern. However, [6] found the peak to occur at the age of 45 years, whereas our data peaks at around 71 years. This is likely due to the fact that our data only samples later years and doesn't include anything below 45 years old. Also, our results are different for speech rate than that found by [7] who found that speech rate increases in late adulthood. However, this strengthens Quene and colleagues' arguments that sociolinguistic factors contribute to their result.

Our findings that show a U-shaped trajectory for fundamental frequency support the findings of other researchers for males and females [14] and for Alistair Cooke [11]. However, a key difference between our results and the previous analyses is the age at which the U-shaped pattern occurs, as with speech rate. Our results suggest that on average for our three speakers there is a change in their early 70s. However, Reubold et al. [11] found that an increase in fundamental frequency occurred at the age of 87 years for Alistair Cooke, and Stathopoulos et al. [14] found it to be around 50 and 60 years for males and females respectively. This suggests that the age at which the change in the direction of fundamental frequency occurs is likely an indicator of individual aging processes and will be unique for every individual.

For vowel space, we find that in general for all of our male speakers, there is a decrease in F1 and F2 with aging. This corresponds with earlier findings [10, 11, 1, 16] but does not agree with findings by Gahl & Baayen who find a more peripheral vowel space with age. Again this may be due to differences is the sampled speakers' ages.

While the effects of age on vowel and word duration, speech rate, fundamental frequency, and vowel space from our data follow general trends that have been found in previous research, our results also indicate that there are dynamic and idiosyncratic patterns for individual speakers. Furthermore, the pattern of change is not linear. As a result, a statistical model that allows for non-linear trends is necessary for research on the acoustic changes related to aging.

5. REFERENCES

[1] S. A. Xue and G. J. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 3, pp. 689–701, 2003.

[2] B. V. Tucker, C. Ford, and S. Hedges, "Speech aging: Production and perception," *WIREs Cognitive Science*, vol. 12, no. 5, pp. 1–23, 2021.

[3] J. D. Harnsberger, R. Shrivastav, W. S. Brown, H. Rothman, and H. Hollien, "Speaking rate and fundamental frequency as speech cues to perceived age," *Journal of Voice*, vol. 22, no. 1, pp. 58–69, Jan. 2008.

[4] E. B. Slawinski, "Acoustic correlates of [b] and [w] produced by normal young to elderly adults," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2221–2230, Apr. 1994.

[5] M. Bruckl, "Women's vocal aging: A longitudinal approach," in *INTERSPEECH 2007*, Antwerp, Belgium, Aug. 2007, pp. 1170–1173.

[6] E. Jacewicz, R. A. Fox, and L. Wei, "Between-speaker and within-speaker variation in speech tempo of American English," *The Journal of the Acoustical Society of America*, vol. 128, no. 2, pp. 839–850, Aug. 2010.

[7] H. Quené, "Longitudinal trends in speech tempo: The case of Queen Beatrix," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. EL452–EL457, May 2013.

[8] S. N. Awan, "The aging female voice: Acoustic and respiratory data," *Clinical Linguistics & Phonetics*, vol. 20, no. 2-3, pp. 171–180, Jan. 2006.

[9] E. de Leeuw, "Native speech plasticity in the German-English late bilingual Stefanie Graf: A longitudinal study over four decades," *Journal of Phonetics*, vol. 73, pp. 24–39, Mar. 2019.

[10] J. Harrington, S. Palethorpe, and C. Watson, "Age-related changes in fundamental frequency and formants: A longitudinal study of four speakers," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, Jan. 2007, p. 2756.

[11] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers," *Speech Communication*, vol. 52, no. 7-8, pp. 638–651, Jul. 2010.

[12] A. Russell, L. Penny, and C. Pemberton, "Speaking fundamental frequency changes over time in women: A longitudinal study," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 101–109, Feb. 1995.

[13] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice*, vol. 32, no. 5, pp. 644.e1–644.e9, Sep. 2018.

[14] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.

[15] E. J. Hunter, M. Kapsner-Smith, P. Pead, M. Z. Engar, and W. R. Brown, "Age and speech production: A 50-year longitudinal study," *Journal of the American Geriatrics Society*, vol. 60, no. 6, pp. 1175–1177, 2012.

[16] E. Jacewicz, R. A. Fox, and J. Salmons, "Cross-generational vowel change in American English," *Language Variation and Change*, vol. 23, no. 1, pp. 45–86, 2011.

[17] S. Gahl and R. H. Baayen, "Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood," *Journal of Phonetics*, vol. 74, pp. 42–54, May 2019.

[18] L. C. Cantor-Cutiva, P. Bottalico, and E. Hunter, "Factors associated with vocal fry among college students," *Logopedics Phoniatrics Vocology*, vol. 43, no. 2, pp. 73–79, Apr. 2018.

[19] L. A. R. Rossi-Barbosa, M. R. Barbosa, R. M. Morais, K. F. de Sousa, M. F. Silveira, A. C. C. Gama, and A. P. Caldeira, "Self-Reported Acute and Chronic Voice Disorders in Teachers," *Journal of Voice*, vol. 30, no. 6, pp. 755.e25–755.e33, Nov. 2016.

[20] H. Brugman and A. Russel, "Annotating multimedia / multi-modal resources with ELAN," *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Mar. 2009.

[21] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics '08*, pp. 5687–5690, 2008.

[22] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2019. [Online]. Available: www.praat.org

[23] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. CRC Press, Jun. 1990.

[24] S. N. Wood, *Generalized Additive Models*. New York: Chapman & Hall/CRC, 2006.

[25] M. C. Kelley and D. Aalto, "Measuring the dispersion of density in head and neck cancer patients' vowel spaces: The vowel dispersion index," *Canadian Acoustics*, vol. 47, no. 3, pp. 114–115, Oct. 2019, number: 3.

[26] B. H. Story and K. Bunton, "Vowel space density as an indicator of speech performance," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. EL458–EL464, 2017.

[27] M. C. Kelley, "Phonetics.jl," Nov. 2021, original-date: 2020-05-27T04:51:46Z. [Online]. Available: https://github.com/maetshju/Phonetics.jl

VARIABILITY IN SPACE, COMPENSATION IN TIME: EFFECTS OF AGING ON TONGUE BODY KINEMATICS IN PROSODIC FOCUS MARKING

Doris Mücke[1], Tabea Thies[1,2], Philipp Buech[3], Noemi Furlani[1], Janine Schreen[1,2], Anne Hermes[3]

[1]IfL-Phonetics, University of Cologne; [2]Department of Neurology, University Hospital Cologne; [3]Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle, Paris
doris.muecke@uni-koeln.de

ABSTRACT

The study investigated how the movements of the tongue vary across different ages for prosodic focus marking. We collected articulatory recordings from 44 German speakers, aged 19 to 79 years, and analyzed the phonetic characteristics of vowel production, such as position, velocity, and duration of the vocalic gesture. Across all ages we found more extreme vocal tract configurations to signal prominence, but the variability of the target positions in the vowels was considerably higher with advancing age. The spatial modifications were accompanied by longer vocalic movements of the tongue body under prominence induced by longer deceleration phases for all speakers. While younger speakers showed a fine-grained temporal marking for all degrees of prominence under investigation (background, broad and contrastive focus), older speakers produced durational differences only to differentiate between unaccented and accented conditions (background vs. broad/contrastive focus). Interestingly, the highlighting strategies were less fine-grained but considerably strengthened with advancing age.

Keywords: prosodic prominence, aging, speech kinematics, hyper-articulation, vowel production

1. INTRODUCTION

In West-Germanic languages, the main prominence of the sentence is usually placed on the focussed parts [1], while other parts within the sentence are considered out-of-focus, i.e., in background position. It has been shown that focus structure determines the position of the nuclear pitch accent and that pitch is a dominant parameter to distinguish between unaccented (background) and accented (broad/narrow/contrastive focus) parts of the utterance on the level of speech production and perception [1, 2]. Other constant phonetic parameters of prosodic modulation involve more subtle, but systematic changes in articulation. Prosodic modulations of vowels and consonants lead to longer, larger and faster articulatory movements to signal prominence in terms of hyper-articulation [3, 4, 5].

To adapt to communicative demands, speakers modulate multiple phonetic cues of the rather flexible prosodic system to differentiate between different degrees of prosodic prominence. A recent study [6] compared different phonetic variables for focus marking. The authors report that the relative importance of the phonetic variables differ for comparisons across accentuation (background vs. broad focus) and within accentuation (broad vs. narrow vs. contrastive focus). Spatio-temporal characteristics of tongue body kinematics play a moderate role to encode the opposition between accented and unaccented syllables (e.g., background vs. broad/contrastive focus), while modifications across different focus types with nuclear accents on the same syllable are more subtle, e.g., broad vs. contrastive focus. The authors argue that tongue body positions are already peripheral under accentuation even in broad focus so there might not be enough space for further adjustments [6].

Prosodic prominence requires a high amount of physical control. Previous studies on prominence marking involving localized hyper-articulation strategies are restricted to younger or middle-aged adults. Thus, little is known about how prosodic strategies evolve with advancing age. For gross motor control, such as limb movement, aging involves the loss of flexibility and muscular strength resulting in smaller and slowed down movements [7]. Further, it can affect the execution of a movement leading to asymmetrical movement patterns in terms of prolonged deceleration phases and greater motor variability decreasing the degree of accuracy of goal-directed movements in older adults [8]. Studies on speech motor control report on a slowed-down speech system when measuring articulation rates in the acoustic domain [9, 10],

while spatial modifications in terms of a more centralized vowel space are rather unclear [11]. Interestingly, there is some evidence from an acoustic speech production study, that variability in older speakers may increase, at least in the temporal domain [12]. Kinematic studies of speech production found slower vocalic tongue body movements in older speakers accompanied by asymmetrical velocity profiles in terms of longer deceleration phases [13]. However, tongue body velocities are either reported to decrease, or to be unaffected by aging [13, 14]. For prosodic focus marking, there is one small articulatory study investigating effects of prominence on aging including 4 older and 4 younger speakers [14, 15]. The results indicate that older speakers might prolong vocalic tongue body movements and the respective deceleration phase to signal prominence in a stronger way than younger speakers.

A clear picture of articulatory focus marking under advancing age is still missing. Hence, the present study investigates focus-related modulations of the tongue in 44 German speakers of different ages by using electromagnetic articulography.

2. METHODOLOGY

Speakers and Recordings: Speech data of 44 German speakers (19 females, 25 males) aged between 19 and 79 years were recorded with an Electromagnetic Articulograph (AG 501). Due to the world pandemic situation[1], there are no speakers between the ages of 31 and 44. However, based on the literature on effects of aging on speech motor control, it can be assumed that this should be a rather uncritical time span for our research question. For all speakers, sensors were placed on the lower and upper lip, tongue tip and tongue body. The tongue sensors were placed approximately $1cm$ and $4cm$ behind the tongue tip. Reference sensors on the cartilage behind the ears were used for head correction and rotation on the midsagittal plane.

Speech material: Ten target words were elicited in an interactive animated game scenario to keep sentence prosody as natural as possible. The target words were disyllabic girl names ('$C_1V_1.C_2V_2$) with stress on the first syllable. All target words were embedded in a carrier sentence, such as "Der Opa hat der Mali/Mila/Mula/Lani/Lina/Luna gewunken" (English: "*The grandpa waved to Mali/Mila/Mula/Lani/Lina/Luna*") and appeared in three different focus conditions, i.e., background (target word is unaccented), broad and contrastive focus (target word bears the nuclear pitch accent). V_1 contains one of the five peripheral vowels: /i:, e:, a:, o:, u:/. In total 1320 target words went into the analysis (10 target words $\times 3$ focus conditions $\times 44$ speakers).

Kinematic tongue body measures of V_1 production: For temporal measures we calculated (i) movement duration (in ms), i.e., the gestural activation interval for V_1 from start to maximum target, (ii) maximum velocity (in mm/s), (iii) duration of the acceleration phase (in ms), i.e., the interval from start to peak velocity and (iv) duration of the deceleration phase (in ms), i.e., the interval from peak velocity to target (mm/s). In a mass-spring model, the acceleration phase can be related to the stiffness of the movement [4]. For positional measures, the (v) vertical and (vi) horizontal tongue body positions according to the window method were obtained for V_1 [16]. Therefore, all position values of the articulatory tongue body trajectory during the first half of the acoustic vowel were averaged to detect the articulatory position for V_1.

Statistics: The statistical analysis was done in R v4.2.2 [17]. Generalized additive mixed models (GAMMs) were performed using the packages `mgcv` v1.8-40 [18] and `itsadug` v2.4.1 [19]. The GAMMs were applied to the duration, velocity, acceleration and deceleration of the vocalic tongue body movement. The model structure was the same for all measurements: focus condition (background, broad, contrastive) was entered as a fixed parameter and as a smooth term by age. Random intercepts for the preceding consonant and vowel were added. The formula for the model structure was the same for all measurements: `parameter ~ focus condition + s(age,by=focus condition) + s(vowel,bs='re') + s(consonant,bs='re')`.

3. RESULTS

3.1. Spatial modifications

For the spatial analysis, speakers have been divided into five groups to provide a better overview of the descriptive data: 19-30 years (n=17), 45-59 years (n=11), 60-69 years (n=8), and 70-79 years (n=8). Fig. 1 provides vowel charts for the kinematic tongue body positions, separately for focus conditions (background, broad focus, contrastive focus), vowel type (/i:, e:, a:, o:, u:/) and the five age groups.

The figure demonstrates that all age groups show more extreme vocal tract configurations from background to contrastive focus to enhance vowel place features in terms of hyper-articulation (Fig. 1; from left to right). This is especially the case for

the low vowel /a/, which is directly related to the opening of the oral cavity to expand sonority. The main difference between the age groups lies in the variability of the target positions. While in the younger group, inter-speaker variability is low with a clear distinction between neighboring vowels, the vowel dispersion considerably increases with advancing age resulting in greater overlapping vowel positions, e.g., for /i - e/ or /u - o/.

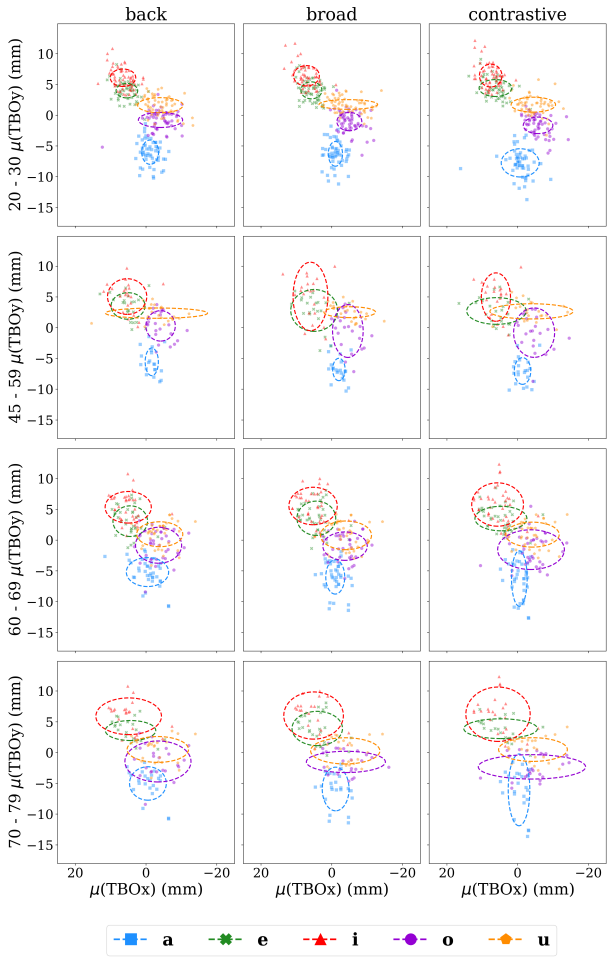


Figure 1: Articulatory vowel space based on vertical and horizontal tongue body positions during the target vowels. Centers of the ellipses are based on the means, width and height are based on the variances. Each row represents a different age group.

3.2. Temporal modifications

Fig. 2 visualizes the estimates of focus structure on the vocalic tongue body durations. The horizontal axis shows the age of the speakers. The upper plot shows the shaded confidence bands (95% confidence interval) for the movement durations separately in the background (red), broad focus (green) and contrastive focus (blue). The following plots (b-d) show differences between the focus conditions. Note, that the differences between 'background-broad' and 'background-contrastive' reflect the comparison across accentuation, while 'broad-contrastive' implies a comparison within accentuation. The red-shaded vertical areas in the difference plots indicate that the confidence intervals of the compared conditions are significantly different from zero. Regarding the fixed effect, the GAMM for duration revealed a significant effect of focus condition ($p < 0.001$). No data was available for speakers between 31-44 years.

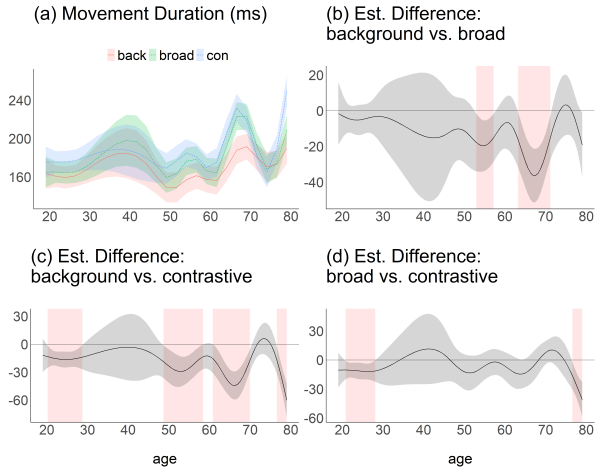


Figure 2: (a)–(d) Non-linear smooths and difference plots for durations of the vocalic tongue body movement.

The younger speakers below 30 years increase the vocalic tongue body movement on average by $16ms$ across ('background-contrastive') and $11ms$ within accentuation ('broad-contrastive'). The middle-aged and older speakers, however, differentiate only across accentuation at an age older than 48 years. Interestingly, the articulatory strengthening of accented syllables in terms of kinematic vowel prolongations are much stronger for speakers beyond 60 years (about $28ms$ from background to contrastive). When taking the background condition as the baseline in the upper plots (2(a), red-shaded bands), we can see that overall durations of vocalic movements increase with age (about $24ms$), but the prominence marking appears to have an 'extra boost' on the prominent syllable (about $33ms$).

The maximum velocity of the tongue body movement during vowel production is another parameter we investigated. However, the GAMM (see Fig. 3(a)) revealed no effect of focus condition ($p > 0.05$). Moreover, no differences between the focus conditions across age were observed on the maximum speed. There is a trend of

decreased maximum speed in speakers older than 60 years when compared to younger and middle-aged speakers (about $13mm/s$ slower for older speakers), though the pattern is not clear.

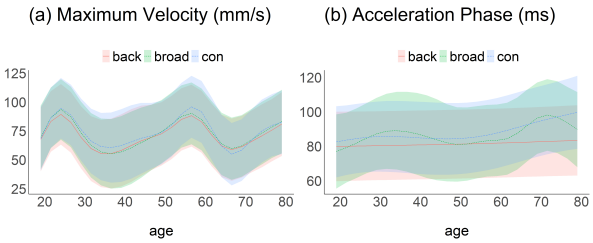


Figure 3: Non-linear smooths for (a) maximum velocity and (b) acceleration phase of the tongue body movement.

Fig. 3(b) displays the effect of focus conditions on the acceleration phase of the vocalic tongue body movement. Interestingly, most speakers do not modulate the acceleration phase under prominence. The GAMM for testing differences in acceleration phases revealed a significant effect of focus condition ($p < 0.001$). Significant differences were found between 'background-broad' in the age of 65-77 years (about $12ms$ longer in broad focus) and between 'background-contrastive' from 59-79 years (about $12ms$ longer in contrastive).

The deceleration phase appears to be a strong prominence parameter across all speakers (see Fig. 4). Especially younger speakers systematically prolong the deceleration phase across and within accentuation (e.g., 11ms for maximum diverging focus structure), while most speakers older than 52 years show modulations only between accented and unaccented parts of the utterance (about $8ms$ longer from background to broad and $14ms$ longer from background to contrastive). There is a clear trend for longer deceleration phases related to advanced age, as can be taken from the red-shaded confidence intervals in background conditions: When being compared to speakers younger than 30 years, we find $11ms$ longer deceleration phases for speakers older than 60 years and $18ms$ longer for speakers beyond 70 years. The GAMM revealed only a significant effect of the focus condition ($p < 0.001$) on the deceleration phase.

4. DISCUSSION & CONCLUSION

Prosodic focus marking implies modifications of the supralaryngeal system. In line with the literature, more extreme tongue configurations are produced to distinguish between accented and

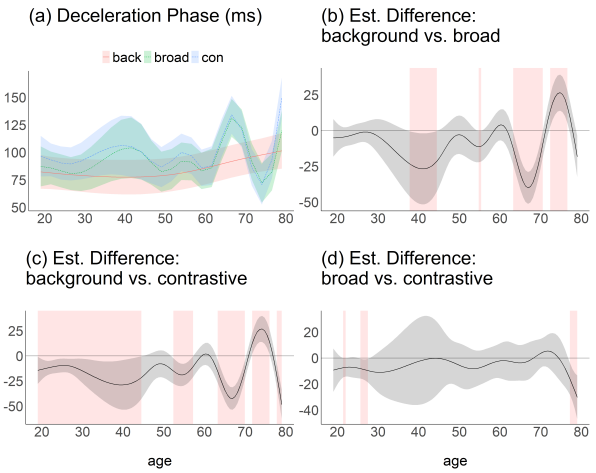


Figure 4: (a)–(d) Non-linear smooths and difference plots for deceleration phase of the vocalic tongue body movement.

unaccented syllables as well as within accentuation. It cannot be confirmed that vowel spaces are more reduced in older speakers, but inter-speaker variability and therefore accuracy of the spatial targets in the vertical and horizontal dimensions appear to be rather high. Even though all speakers produce more peripheral vowels under prominence, the variability in the spatial tongue positions considerably increases with advanced age. Older speakers show more overlap between neighboring vowels than younger speakers, and this process already starts for middle-aged speakers.

The spatial modifications are accompanied by longer vocalic movement durations of the tongue body. Vocalic prolongations are especially induced by longer deceleration phases of the gestural activation interval, while stiffness-related acceleration phases remain stable. The temporal modulations are more fine-grained for younger speakers, i.e., below 30 years, as they differentiate between all degrees of prominence. In contrast, the strength of prosodic modulations are enhanced especially in the temporal domain for speakers older than 60 years. This aging-induced process already starts in middle-aged speakers. We interpret the temporal adjustments as compensation strategies for the increasing variability in the spatial dimension within older speakers. Durational adjustments of vowels in accented target syllables are known to be effective since they lead to an increase in sonority on the perceptual level [20]. We conclude that the results on prosodic focus marking strategies with increasing age reveal a plasticity of the prosodic system that can be adapted through life [21].

5. ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) in the collaborative research center SFB1252 Prominence in Language (Project-ID 281511265) as well as by the French ANR (ANR-10-LABX-0083, LABEX-EFL).

6. REFERENCES

[1] D. R. Ladd, *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press, 1980.

[2] S. Baumann and B. Winter, "What makes a word prominent? Predicting untrained German listeners' perceptual judgments," *Journal of Phonetics*, vol. 70, pp. 20–38, Sep. 2018.

[3] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," in *Speech Production and Speech Modelling*, ser. NATO ASI Series, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Springer Netherlands, 1990, pp. 403–439.

[4] T. Cho, "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3867–3878, Jun. 2005, publisher: Acoustical Society of America.

[5] J. Harrington, J. Fletcher, and M. E. Beckman, "Manner and place conflicts in the articulation of accent in Australian English," in *Papers in Laboratory Phonology: Volume 5*, ser. Papers in Laboratory Phonology, M. B. Broe and J. B. Pierrehumbert, Eds. Cambridge: Cambridge University Press, 2000, vol. 5, pp. 40–55.

[6] S. Roessig, B. Winter, and D. Mücke, "Tracing the Phonetic Space of Prosodic Focus Marking," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[7] C. J. Ketcham and G. E. Stelmach, *Movement Control in the Older Adult*. National Academies Press (US), 2004, publication Title: Technology for Adaptive Aging.

[8] E. A. Christou, "Aging and Variability of Voluntary Contractions," *Exerc Sport Sci Rev*, vol. 39, no. 2, pp. 77–84, Apr. 2011.

[9] J. D. Amerman and M. M. Parnell, "Speech timing strategies in elderly adults," *Journal of Phonetics*, vol. 20, no. 1, pp. 65–76, Jan. 1992.

[10] P. H. Ptacek, E. K. Sander, W. H. Maloney, and C. C. R. Jackson, "Phonatory and Related Changes with Advanced Age," *Journal of Speech and Hearing Research*, vol. 9, no. 3, pp. 353–360, Sep. 1966, publisher: American Speech-Language-Hearing Association.

[11] D. D'Alessandro and C. Fougeron, "Changes in Anticipatory VtoV Coarticulation in French during Adulthood," *Languages*, vol. 6, no. 4, p. 181, Dec. 2021, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[12] P. Tremblay, I. Deschamps, P. Bédard, M.-H. Tessier, M. Carrier, and M. Thibeault, "Aging of speech production, from articulatory accuracy to motor timing," *Psychol Aging*, vol. 33, no. 7, pp. 1022–1034, Nov. 2018.

[13] A. Hermes, J. Mertens, and D. Mücke, "Age-related Effects on Sensorimotor Control of Speech Production," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1526–1530.

[14] D. Mücke, T. Thies, J. Mertens, and A. Hermes, "Age-related effects of prosodic prominence in vowel articulation," in *12th International Seminar on Speech Production*, New Haven, United States, Dec. 2020, pp. 126–129.

[15] T. Thies, A. Hermes, and D. Mücke, "Compensation in Time and Space: Prominence Marking in Aging and Disease," *Languages*, vol. 7, no. 1, p. 21, Mar. 2022, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[16] S. Roessig, *Categoriality and continuity in prosodic prominence*. Language Science Press, Jan. 2021, publication Title: Language Science Press.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[18] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society (B)*, vol. 73, no. 1, pp. 3–36, 2011.

[19] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2022, r package version 2.4.1.

[20] M. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," in *Gesture, Segment, Prosody*, ser. Papers in Laboratory Phonology, D. R. Ladd and G. J. Docherty, Eds. Cambridge: Cambridge University Press, 1992, pp. 68–89.

[21] E. De Leeuw and C. Celata, "Plasticity of native phonetic and phonological domains in the context of bilingualism," May 2019, accepted: 2019-10-16T14:01:23Z Publisher: Elsevier.

CHANGE ACROSS THE LIFESPAN IN GOAT: EVIDENCE FROM A PANEL STUDY OF TYNESIDE ENGLISH

Lea Bauernfeind, Carina Ahrens & James Grama

Sociolinguistics Lab, University of Duisburg-Essen
lea.bauernfeind@uni-due.de, carina.ahrens@uni-due.de, james.grama@uni-due.de

ABSTRACT

This study investigates changes in GOAT over time in two sets of panel speakers (old and young) from Tyneside, North-East England. Data were collected at two or three timepoints between 1971 and 2019. We observe a general preference for the incoming Northern monophthong and standard diphthong, while localized variants lose ground over time. Speakers also show evidence of lifespan change. In the old panel, middle-class speakers show a preference for the standard closing diphthong that follows a U-shaped curve, while working-class speakers exhibit a similar preference for the back monophthong. The young panel, by contrast, shows no clear class-based patterns. Acoustic analysis of a single young panel speaker reveals further changes in the phonetic realization of GOAT variants over time. Together, these results suggest that GOAT is mediated across the lifespan by class and linguistic marketplace pressures, both of which are highly dependent on speaker age.

Keywords: panel study, lifespan change, Tyneside English, auditory and acoustic methods, GOAT

1. INTRODUCTION

A notable quality of many Northern varieties of British Isles English is the monophthongal realization of the mid-vowels FACE and GOAT [1-3]. Work on the English spoken on Tyneside, an urban conurbation in the North-East of England, has demonstrated that these vowels have undergone a shift in apparent time towards pan-northern monophthongs, mainly at the expense of localized inglides [4-6]. Real-time longitudinal work by Buchstaller et al. [7] found that speaker participation in changes to FACE was highly contingent on mobility, sex, and marketplace pressures. However, no studies have as of yet investigated whether (and if so, to what extent) GOAT shows similar lability over the lifespan.

In Tyneside English (TE), GOAT can be realized as one of four variants (see Table 1). Each variant (except [ɵː]) has a corresponding variant in FACE with similar socio-indexical meanings (see [4-6] for the bulk of the work in this area). The pan-northern monophthong has been described as the default form in TE, resulting from a leveling process towards a general, supra-local northern variety [6, 8]. The closing diphthong, then, serves as the prescribed standard variant. Evidence from style shifting suggests that TE speakers associate this form with 'correctness', and its use is largely restricted to young middle-class speakers and women [5]. The inglide and central monophthong are both favored by working-class men, and are connected to local orientation and regional pride. And while the inglide is slowly abating, the central monophthong remains a prominent male variant across social class and age.

Type	Phonetic form
Pan-northern monophthong	[oː]
Localized central monophthong	[ɵː]
Localized ingliding diphthong	[ʊə]
Standard closing diphthong	[oʊ]

Table 1: GOAT variants in Tyneside English.

Thus, while GOAT is well-described in TE in apparent time, and FACE has been investigated using real-time panel data, no work to our knowledge has examined how individual speakers pattern across their lifespan with respect to GOAT. The current study explores this by using two samples of panel speakers. Variability and change in GOAT realizations over the lifespan provide insight into how niched marketplace pressures and age bear on a variable undergoing change in progress.

2. DATA AND METHODS

Data come from interviews conducted with an old and young panel. The old panel includes the same six speakers investigated in [7], who were first recorded in 1971 in their early twenties, and again in 2013, when they were near- or immediately post-retirement. Half of the old panel were interviewed a third time in 2019, when they were well into retirement. Six younger speakers form the young panel, first recorded between 2007 and 2010 when they were students at university, then again in 2014 when they had recently entered the workforce. These panel samples combined provide real-time evidence of how age-stratified cohorts behave with a variable undergoing change in the broader speech community. The panel

speakers are listed in Table 2, along with their social trajectories. All speakers were broadly classified into three social types: stably working class, upwardly mobile (moving from working to middle class), or stably middle class. Class designations were based on occupation (or parent's occupation for young speakers at T1 who were still at university, following recommendations in [9]).

Panel	Speaker	YOB	Trajectory
Old	Rob	1948	Stable WC
	Edith	1939	Stable WC
	Anne	1948	Stable WC
	Aidan	1946	Upward
	Fred	1950	Upward
	Nelly	1942	Stable MC
Young	Jane	1990	Stable WC
	Paul	1990	Stable WC
	Charlotte	1990	Upward
	Lynn	1990	Upward
	Jake	1985	Upward
	Amelia	1989	Stable MC

Table 2: Speaker sample with class trajectories.

Data were transcribed in ELAN [10], and two methodological pipelines were used: one for auditory and another for acoustic analyses. The auditory analysis pipeline aimed to capture proportional changes in GOAT variants over time (see Section 3.1-3.2). The first and second authors auditorily coded GOAT tokens from all 12 speakers; variants fell into one of four possible categories, outlined in Table 1, and only five tokens were coded for any one lexeme. This resulted in 2,484 tokens for analysis. Second, an acoustic analysis pipeline was used to investigate phonetic changes in a single young woman, Charlotte (see Section 3.3). This analysis draws from all content words across timepoint, and includes a third timepoint. To facilitate acoustic analysis, these tokens (n=540) were also coded auditorily by the second author. Acoustic measurements were extracted from force-aligned segments in LaBB-CAT [11], boundaries were hand-checked in Praat [12], and F1 and F2 were extracted from 20-80% of the vowel's duration at 10% intervals. Overlapped or noisy tokens were excluded from all analyses. Statistical analysis was conducted in R [13], and linear mixed-effects models were built using `lme4` [14].

3. RESULTS

3.1. Old panel

Figure 1 shows the proportional distribution of GOAT variants for the old panel speakers across timepoint. In general, working-class speakers show higher proportions of local forms, while upwardly mobile and stably middle-class speakers produce more standard forms. The pan-northern monophthong is the most frequent variant overall, particularly for working-class speakers (compare [5, 6]).

Crucially, we observe individual instability over the lifespan that is mediated by social class. Rob, an engraver, shows the highest rates of localized variants of any speaker. However, at T2, he shifts away from the central monophthong in favor of the inglide. Alongside this, he exhibits an increase in the pan-northern monophthong and a reduction in the standard closing diphthong. By contrast, Edith, a local shop assistant and home help, shows a much less differentiated system. She increases her use of the pan-northern monophthong across timepoint at the expense of the closing variant, and produces no localized variants. These two speakers nicely mirror findings from apparent time (compare [5, 6]) in that they both register an increase in the pan-northern monophthong. Anne, for whom we have a third timepoint, shows the most complex patterns. Her proportion of the pan-northern monophthong increases from T1 to T2 at the expense of both the closing and the inglide variant. However, at T3, she reverts to a distribution that mirrors her system at T1, a pattern that is reminiscent of the U-shaped curve postulated in sociolinguistic theorizing [15].

Changes in middle-class speakers are strikingly different. Nelly and Fred, both educators, show clear evidence of a U-shaped curve. From T1 to T2, they exhibit retrenchment towards the closing form, with only marginal examples of localized variants. This time span represents polar ends of when Nelly and Fred worked in professional settings that uphold pervasive standard language ideologies (see recent critical school research in the UK [16, 17]). While the effect of such 'correction' towards the standard varies across the two speakers, both show renewed uptake of the pan-northern monophthong at T3, well into retirement. This relaxation is particularly stark for Nelly, who shows a nearly categorical shift away from the closing variant. Aidan, an upwardly mobile lecturer, exhibits somewhat more complex patterning from T1 to T2. While he, too, displays an increase in the standard closing variant during his teaching career, he exhibits commensurate upticks in local forms, both at the expense of the pan-northern monophthong. This seemingly contradictory pattern can partly be explained by Aidan's social profile. He emphasizes that he is a proud Geordie with a working-class heritage. Moreover, as a lecturer at a welding college, he reports feeling connected with blue-collar workers. Thus, his use of local forms can be viewed as an expression of working-class solidarity, despite his position as an educator. While

we lack a third recording for Aidan, we hypothesize that he would show similar behavior to Fred and Nelly, and move towards more vernacular forms.

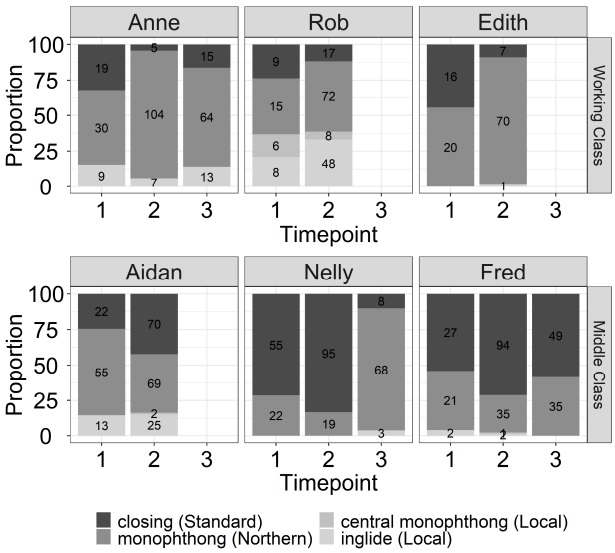


Figure 1: Proportional distributions in the old panel.

3.2. Young panel

Figure 2 shows the distribution of GOAT variants for the young panel. In contrast to the old panel, variants in the young cohort are functionally restricted to just the pan-northern monophthong and the standard closing diphthong (the lone inglide is produced by Jane). Furthermore, changes across timepoint are not clearly delineated by social class.

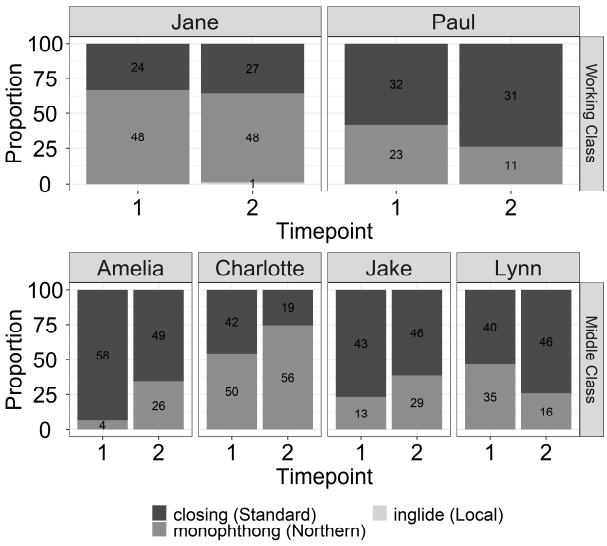


Figure 2: Proportional distributions in the young panel.

Three distinct patterns emerge when comparing individual trajectories across timepoint. First, a single speaker, Jane, is stable. Second, we observe a shift towards the pan-northern monophthong for one middle-class and two upwardly mobile speakers, Amelia, Charlotte, and Jake. Finally, the opposite pattern characterizes Lynn and Paul's behavior, who register an increase in the standard closing variant at the expense of the pan-northern monophthong. These latter two patterns are difficult to square with results from apparent time, which find that the closing variant is on the rise in middle-class speakers [5, 6]. If this is indeed the case, then movement away from standard forms would constitute retrograde change among the middle class. Working-class speakers, too, differ; in [5, 6], young working-class women strongly favor the pan-northern monophthong, and men produce more localized forms. In the present analysis, localized forms are all but absent, and lifespan uptake of supra-local and standard forms is variable.

3.3. Probing acoustic patterns: A case study

While auditory coding can capture changes in the variable use of sociolinguistically relevant forms, acoustic data enables us to investigate changes in the phonetic implementation of those variants. We focus here on Charlotte, whose changing career path (university student at T1, PhD student at T2, and lecturer at a university in the North-East at T3) allows us to explicitly test the impact of marketplace pressures on GOAT. We treat age as categorical, rather than chronological, to capture non-linear behavior as the speaker ages. When considering a broader range of content words (see discussion of methods in Section 2), Charlotte's use of the pan-northern monophthong increases substantially. This form is near-categorical at all timepoints, but Charlotte nevertheless shows an increase from T1 (159/171, 93%) to T2 (179/186, 96%), while the added T3 (167/183, 91%) shows rates parallel to T1.

Figure 3 plots the trajectories of the variants Charlotte produces: the pan-northern monophthong and the standard diphthong. Linear mixed-effects models (with T2 as the baseline) fit to the Euclidean distance of both variants reveal that monophthongal GOAT has a shorter trajectory at T2 than T1 (β=77.90, t=2.24, p<0.05) or T3 (β=112.52, t=3.33, p<0.01). We observe similar changes in the position of monophthongal GOAT in F1/F2 space. Specifically, linear mixed-effects models reveal a higher midpoint at T2 than at T1 (β=0.22, t=2.08, p<0.05) or T3 (β=0.24, t=2.33, p<0.05). A preceding coronal also motivates significantly fronter monophthongal GOAT midpoint values over non-coronals at T3 (β=0.43, t=2.80, p<0.01), but no significant differences at either T1 (β=-0.05, t=-0.44, p=0.66) or T2 (β=0.14, t=-1.18, p=0.24). No effects were found to bear significantly on closing GOAT in either trajectory length or position in F1/F2 space. This null effect might be expected with low token numbers, but it

may be that infrequent variants are less susceptible to lifespan change.

Together, these results paint a picture of lifespan change in the acoustic realization of Charlotte's most frequent GOAT variant. At T2, Charlotte exhibits not only higher proportions of the pan-northern monophthong, but also realizes these variants as higher, and with a shorter trajectory length than at any other timepoint. In addition, there is some evidence that Charlotte innovates a phonological effect in real-time; monophthongal GOAT is fronter at T3 following coronals. Whether this is a broader change in the speech community is a question for future research.

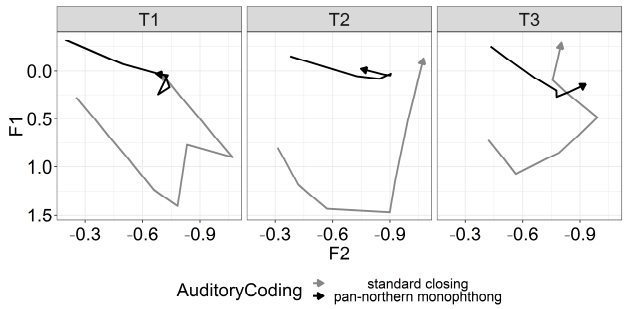


Figure 3: Charlotte's trajectory for closing and monophthongal variants across timepoint.

4. DISCUSSION

These results suggest that the realization of GOAT in both the old and young panels are socially- and occupationally niched. In the old panel, speaker behavior is tied to socio-economic trajectory, with working-class speakers showing a high incidence of localized forms and a preference for the pan-northern monophthong that increases (Rob and Edith) or follows a U-shaped curve (Anne). Middle-class speakers, by contrast, show a much clearer preference for standard closing GOAT, with both Fred and Nelly following U-shaped curves, as they register an increase in the standard variant during active teaching, followed by an increase in vernacularity in post-retirement. One way to interpret these patterns is via niched marketplace pressures [18]. Specifically, middle-class and upwardly mobile old panel speakers serve roles in education, occupations where overt valuation of standard language is particularly high [15, 16]. For old middle-class speakers, then, evidence points to some movement towards standard GOAT variants during times when these speakers would have had maximum daily exposure to such beliefs. At T3, Nelly and Fred show evidence of relaxation of these pressures, consistent with the oft-hypothesized 'tail' towards vernacular forms [15].

We can compare these findings with those of Buchstaller et al. [7] for FACE, noting that their analysis includes more tokens (n=1,922) but lacks a T3 recording for any speaker. Comparing their findings with ours, the two vowels do not follow the same trajectories over the lifespan. Edith and Anne's FACE show minor shifts away from monophthongal variants from T1 to T2, the precise opposite of what we observe here for GOAT. Nelly produces largely stable proportions of FACE across timepoint, with a clear preference for the pan-northern monophthong, but GOAT outside of T3 is dominated by the standard variant. Aidan exhibits an increase in his preference for the localized inglide for FACE, but GOAT shows the clearest increase in the standard closing form. Only Rob and Fred show parallel behavior in the two vowels. Thus, while FACE and GOAT are structurally and socio-indexically linked (see [4-6, 19]), this differential patterning may be evidence that they index slightly different meanings.

The young panel, by contrast, shows more individuated patterns, dominated by changing preferences for the pan-northern monophthong (Amelia, Charlotte, Jake) or the standard closing variant (Paul, Lynn), with a single case of stability (Jane). This makes it difficult to generalize across social trajectory or occupation; educators (Amelia, Charlotte, Lynn) exhibit both patterns. It is possible that these changes reflect more individualized relationships with marketplace pressures, or that these pressures act differently on younger speakers. What can be stated with some confidence is that variability in GOAT of younger TE speakers centers around the pan-northern monophthong and the standard closing form, both of which enjoy greater representation than any localized forms (cf. [20]).

Finally, we present evidence from a single speaker, Charlotte, of acoustic instability over the lifespan. Specifically, an increase in proportion appears to be correlated with more peripheral realizations of the pan-northern monophthong. This suggests a connection between the proportional use of variants and their phonetic implementation. Evidence also points to an emerging phonological contrast at T3, further accentuating speaker lability over the lifespan. These insights are afforded by the marriage of auditory and acoustic methods in the analysis of a complex phonological variable (cf. [19-21]).

Overall, our findings show that individual responses to changes in progress are heavily mediated by social trajectory and its associated marketplace pressures, as well as age. Moreover, we exemplify how the non-linear treatment of age as life-stage dependent (rather than chronological) can reveal changes in acoustic implementation. That GOAT's patterning differs from that of FACE suggests that change in variants of these vowels may not manifest as symmetrical over the lifespan of individual speakers as it does in the broader speech community.

5. ACKNOWLEDGEMENTS

We acknowledge funding provided by the Deutsche Forschungsgemeinschaft (#BU-2902/7-1). We are also grateful to: the transcribers in the Sociolinguistics Lab (Anabel Büscher, Paul Enninger, Anne-Marie Mölders, Marie Phillips, Hannah Sawall, and Deborah Veiter); Mirjam E. Eiswirth for project management and data curation; the on-site interviewers; and the DECTE team at Newcastle University for providing access to the original data. Additional thanks go to Isabelle Buchstaller and the anonymous reviewers for their helpful comments on an earlier draft. We are greatly indebted to our panel speakers for sharing their time and speech. All remaining errors are our own.

6. REFERENCES

[1] Wells, J.C. 1982. Accents of English. Cambridge, UK: Cambridge University Press.

[2] Beal, J., Burbano-Elizondo, L., & Llamas, C. 2012. *Urban North-Eastern English: Tyneside to Teesside*. Edinburgh, UK: Edinburgh University Press.

[3] Haddican, B., Foulkes, P., Hughes, V., & Richards, H. 2013. Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change* 25, 371-403.

[4] Watt, D. 1999. Phonetic variation in two Tyneside vowels: Levelling in lockstep. *Proc. ICPhS*, 1621-1624.

[5] Watt, D. 2000. Phonetic parallels between the close-mid vowels of Tyneside English: Are they internally or externally motivated? *Language Variation and Change* 12, 69-101.

[6] Watt, D. 2002. 'I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced levelling in the Tyneside vowel system. *Journal of Sociolinguistics* 6(1), 44-63.

[7] Buchstaller, I., Krause, A., Auer, A., & Otte, S. 2017. Levelling across the life-span?: Tracing the FACE vowel in panel data from the North East of England. *Journal of Sociolinguistics* 21(1), 3-33.

[8] Watt, D. & Milroy, L. 1999. Patterns of variation and change in three Newcastle vowels: Is this dialect levelling? In P. Foulkes & G. Docherty (eds.) *Urban Voices*. London: Arnold, 25-46.

[9] Baranowski, M. & Turton, D. 2018. Locating speakers in the socioeconomic hierarchy: Towards the optimal indicators of social class. Paper presented at *NWAV 47*. New York University.

[10] Lausberg, H, & Sloetjes, H. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41(3), 841-849.

[11] Fromont, R., & Hay, J. 2012. LaBB-CAT: An annotation store. In *Proc. Australasian Language Technology Association Workshop*, 113-117.

[12] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

[13] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.Rproject.org/.

[14] Bates, D., Mächler, M. Bolker, B., & Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1-48.

[15] Downes, W. 1998. *Language and society, 2nd ed.* Cambridge.

[16] Cushing, I. 2020. The policy and policing of language in schools. *Language in Society* 49(3): 425-450.

[17] Cushing, I., Snell, J. 2022. The (white) ears of Ofsted: a raciolinguistic perspective on the listening practices of the schools inspectorate.

[18] Bordieu, P., & Boltanski, L. 1975. Le fétichisme de la langue. Actes de la recherche en sciences sociales 4, 2-32.

[19] Amand, M. 2019. Analyse sociophonétique de l'anglais de Tyneside dans le corpus DECTE: le cas des voyelles FACE, GOAT, PRICE et MOUTH. PhD thesis, Université de Paris.

[20] Warburton, J. 2020. The merging of the GOAT and THOUGHT vowels in Tyneside English: Evidence from production and perception. PhD thesis, Newcastle University.

[21] Maguire, W.N. 2007. What is a merger, and can it be reversed? The origin, status and reversal of the 'NURSE-NORTH merger' in Tyneside English. PhD thesis, Newcastle University.

AGE-RELATED VOWEL VARIATION IN FRENCH

Anne Hermes, Nicolas Audibert, Angélina Bourbon

Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle, Paris
anne.hermes@sorbonne-nouvelle.fr, nicolas.audibert@sorbonne-nouvelle.fr, angelina.bourbon@sorbonne-nouvelle.fr

ABSTRACT

This paper presents a study on the temporal and spectral characteristics of oral and nasal vowels in French read speech, produced by speakers from 25 to 90 years. The study utilizes a set of eight vowel categories and combines classical vowel space-related metrics with MFCC parametrization to accurately include nasal vowels. In total 15,375 vowels were analyzed from data from 37 French speakers (20 females, 17 males). Our data showed (a) a slowing down (increase in vowel duration) with age and (b) no clear change in the overall centralization/dispersion of vowels with increasing age. However, for older male speakers, there was a decrease in the variability within vowel categories and a moderate increase in the separation between categories. Additionally, both male and female speakers demonstrated a greater acoustic distinction between oral and nasal vowels with aging.

Keywords: Aging, speech production, acoustics, French, MFCC, vowel space

1. INTRODUCTION

1.1. Age-affected speech parameters

Several studies investigated age-related effects on the supraglottal and glottal systems. Aging effects have been found in voice quality parameters (pitch: a decrease in F0 in women [1]; less clear for men [2, 3, 4]), in Voice Onset Time (shorter VOT in voiceless stops; [5, 6, 7], and in some instances, in formant frequencies (centralization, [8]). Overall, the most robust and documented age-related differences in speech production concern speech rate in that a slowing down is observed for older speakers [9, 10, 11, 12, 13, 14]. A recent study on French, [15] showed that older speakers produce longer pauses, longer sentence durations and lower articulation rates than younger speakers.

An important effect of aging over the lifespan (from young over middle-aged to older) has been reported by [16] for German and [17] for French in that effect, such as slowing down, and decrease in coarticulation is reported to be non-linear.

The relationships between aging and the realization of vowels have been addressed in various studies, focussing in most cases on the frequencies of the first two formants. While some authors report a decrease in formant frequencies, only the decrease in F1 for women appears to be rather robust across studies [18], but no clear tendency is found regarding vowel space area.

It is rather difficult to determine at what exact age these changes occur, as different studies have involved participants of different ages/age groups with different methods of analysis. Potential explanations for these changes in formant frequencies are inter alia the lengthening of the vocal tract [19], reduced and/or slowed down articulatory movement [10, 20], and other adjustments in articulation. Further research is needed to understand the relative contributions of these factors and to determine whether these changes are seen in all older adults or only in certain subgroups, such as those with health conditions. Normative data on vowel formant frequencies in older adults is necessary for interpreting acoustic changes that may occur in age-related conditions and for developing methods for automatically detecting speaker age.

To our best knowledge, such studies investigated individual formant frequencies and in some cases vowel space area, but did not examine the interplay between vowel categories and their evolution with aging. This study focuses on aging effects on temporal (vowel duration) and spectral shape in oral and nasal vowels in French read speech produced by speakers across different ages. This study aims to contribute to our knowledge of aging effects on speech production and better account for age-related vowel variation by using a rather large set of vowel categories (8) and combining classical vowel space-related metrics with MFCC parametrization to reliably include nasal vowels.

2. METHOD

2.1. Subjects

Acoustics recordings from 41 French native speakers were done in a quiet room with a head-mounted microphone (SHURE SM35-TQG), using Pyaudio (48 kHz sampling rate). The subjects' chronological age ranges from 23-90 years old. For the study, all of the subjects had normal vision or had corrected their vision to be normal, and none of them had any

reported issues with speech, voice, language, psychological health, neurological functioning, or neurodegenerative disorders at the time of the study. Participants were screened for cognitive alterations using a standardised clinical tool, the Mini-Mental State Examination (MMSE, [21]). All participants were above the threshold of 24 (out of 30).

2.2. Speech material

The speech material is taken from Bourbon & Hermes (2021) [15] and consisted of French sentences (based on [22]), which differed in length and syntactic complexity (e.g., short-simple "Mélanie Dupont a réservé ses vacances" (Engl. *Mélanie Dupont booked her holidays*); long-complex condition "Mélanie Dupont a dit à sa mère, qui l'a appelée dans la matinée, qu'elle attend avec impatience ses vacances." (Engl. *Melanie Dupont told her mother, who called her in the morning, she's looking forward to her vacation*). Each sentence (short-simple, short-complex, long-simple, long-complex) was repeated three times, i.e., in total each speaker produced 36 sentences.

2.3. Annotations/Labelling

All sentences were segmented into phones using WebMaus [23]. After a sanity check, 2.2% of recordings with erroneous segmentation were discarded. To keep the number of utterances equal across speakers, two repetitions of each utterance were retained. We ended up with 37 speakers (20 females: $\bar{x}$=62.5 years, sd=23.3, 17 males: $\bar{x}$=56.7 years, sd=20.4). Fig. 1 displays density plots for the age distribution for female and male speakers.

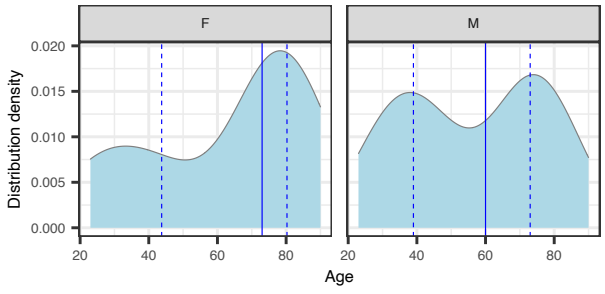


Figure 1: Density plots of age distribution in female (left) and male (right) speakers. The vertical lines represent the median (solid lines) and the quartiles (dashed lines).

The final corpus included exemplars of the oral vowels /i, y, e, ɛ, a, o/ and the nasal vowels /ɔ̃, ɑ̃/. After discarding the final vowel of each utterance, lengthened due to the accentuation pattern in standard French, for each vowel, we had a minimum of 14 exemplars (see Table 1).

Vowel	meanCount	minCount	maxCount
i	65	53	74
e	75	58	83
ɛ	51	40	56
a	118	96	122
o	35	26	36
y	16	14	18
ɔ̃	31	24	32
ɑ̃	25	18	27

Table 1: Exemplars for all oral and nasal vowels.

Note that we discarded /ɔ/ which was sparser and more appeared only in specific contexts. For the final dataset, we had a total of 15,375 vowels which were included in the analysis.

2.4. Measurements

For the temporal analysis, we extracted the durations of all target vowels. This temporal parameter was used as an indicator of speech rate and is expected to slow down with age as it has been reported in many studies before.

Automatic formants measurement is known to be error-prone, particularly in the case of nasalized segments [24]. Although thresholding may be applied to discard aberrant formant values [25], such a procedure may lead to discarding a large number of vowels (see e.g. [26]) and can hardly be applied to nasal vowels. In order to avoid including erroneous measures in analyses, we use MFCC measures to reliably capture the spectral characteristics of target vowels, following [26]. A custom Praat [26] script is applied to extract on each target vowel 12 MFCCs (not including coefficient 0 related to the overall level of energy) on a 15 ms frame centred on the middle of the vowel, using a filter bank spaced by 100 Mel.

Separately for each speaker, different metrics *DistCentroid*, *VDispersion* and *ContrastLoss* [28] are computed from 12-dimensional MFCC values, resulting in one value for each metric for each vowel exemplar. *DistCentroid* is defined as the Euclidean distance from a vowel exemplar to the vowel space centroid (computed as the centroid of vowel-specific centroids); *VDispersion* as the Euclidean distance from an exemplar to the centroid of the corresponding vowel category; *ContrastLoss* as the posterior probability for the exemplar to belong to the reference vowel category, computed by a linear discriminant analysis (LDA). In addition to ContrastLoss, confusion rates between vowel categories were also extracted from the LDA results. The R code for the computation of all metrics applied is provided as a Shiny interactive application and for download under https://shiny.laboratoirephonetiquephonologie.fr/vowel_space_metrics_computation/. Since we assume

that there are differences due to age, we display in addition to the results for all speakers, the results for the older speakers separately (≥ 60 years; solid lines).

3. RESULTS

3.1. Age effects on vowel duration

For the analysis of this dataset, we analyzed male and female speakers separately, since it is known that age-related effects are reported to be different.

Fig. 2 displays the MFCC for all vowel durations across male and female speakers over age. In line with what has been reported for age-related effects on temporal parameters, our data also revealed a slowing down with increasing age (females: r=0.554; males: r=0.789). Comparable trends are observed for the older speakers, with a larger between-speakers variability (females: r=0.204; males: r=0.385).

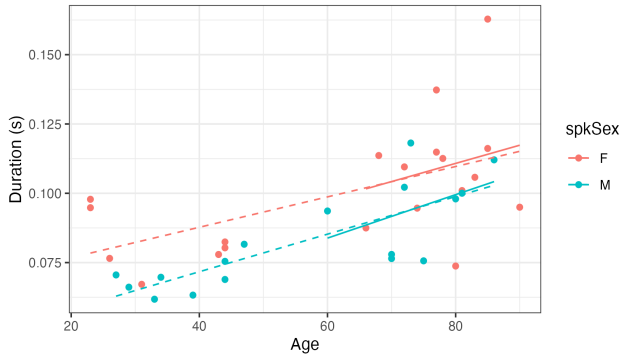


Figure 2: Scatterplot of mean duration per speaker against age, for female and male speakers. The regression is illustrated by dashed lines for all speakers, and by solid lines for older speakers.

3.2. Age effects on formants

As said before, for the formant analysis, we used several metrics to look at age-related effects. For the metric *DistCentroid* (vowel exemplar to the vowel space centroid), no overall pattern of change with age is found either for female (r=.106) or male speakers (r=.083), as shown by Fig. 3.

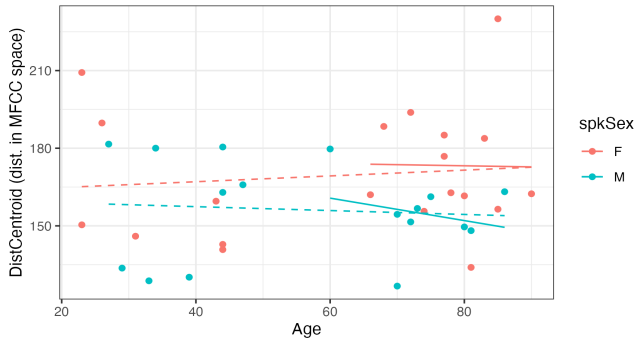


Figure 3: Scatterplot of mean DistCentroid values per speaker against age, for female and male speakers. Higher values of DistCentroid correspond to a more dispersed vowel space. The regression is illustrated by dashed lines for all speakers, and by solid lines for the older speakers.

A moderate tendency toward centralization is found for older male speakers (r=-.233) but not for older female speakers (r=-.013).

For the metric *VDispersion* (distance from exemplar to the centroid of the corresponding vowel category), Fig. 4 illustrates diverging patterns found for female (r=.211) and male speakers (r=-.387) across the lifespan. For older speakers, category-wise variability decreases with age for male speakers (r=-.569) while this tendency is much weaker for female speakers (r=-.120).

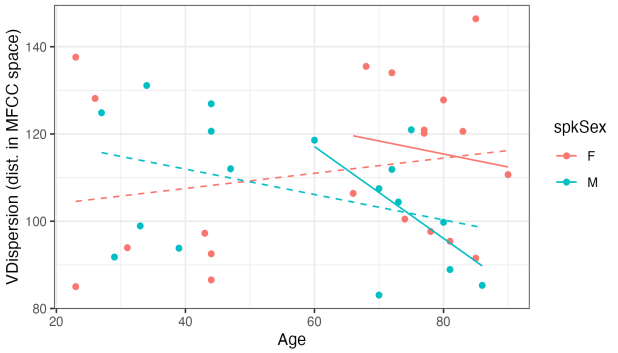


Figure 4: Scatterplot of mean VDispersion values per speaker against age, for female and male speakers. Higher values of VDispersion correspond to a larger variability within vowel categories. The regression is illustrated by dashed lines for all speakers, and by solid lines for the older speakers.

For the metric *ContrastLoss* (posterior probability of vowel category membership), Fig. 5 illustrates that there is no effect of age on the metric for female speakers (r=.035), neither across ages in the older group (r=.085). For male speakers (r=.343), it can be observed that overlap decreases with age, across all ages and in the older group specifically (r=.300). Note that *ContrastLoss* analysis considers all vowel categories together and may hide category-specific patterns of change with aging.

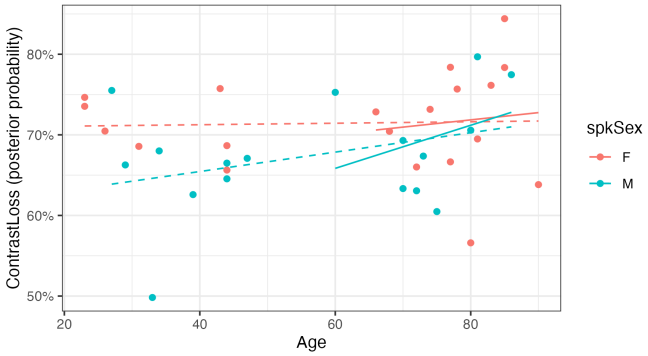


Figure 5: Scatterplot of mean ContrastLoss values per speaker against age, for female and male speakers. Higher values of ContrastLoss correspond to a lesser overlap between vowel categories. The regression is illustrated by dashed lines for all speakers, and by solid lines for the older speakers.

In order to get more insights into how the organization of the speaker's vocalic system is affected by aging, the evolution with age of confusion attributed by the LDA between broader vowel classes was further explored. Although diverging patterns are observed for most confusion types considered, a moderate but consistent tendency is found for confusions between oral and nasal vowels, which undergo less acoustic overlap with aging, as illustrated in Fig. 6 (females: r=-.533; males: r=-.235). However, the lesser overlap between oral and nasal vowels is reached earlier by female speakers, which may explain the diverging patterns observed between female and male speakers in the older group (females: r=.244; males: r=-.47).

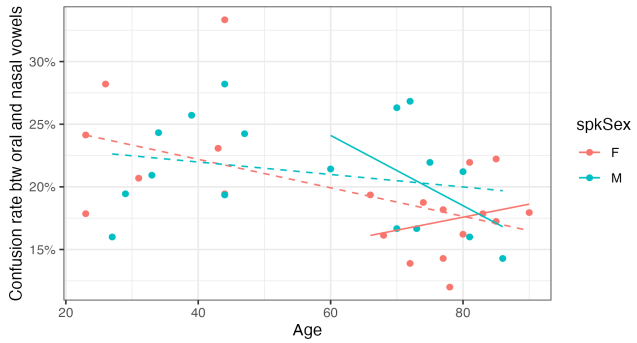


Figure 6: Oral-nasal confusions based on LDA classification per speaker against age, for female and male speakers. The regression is illustrated by dashed lines for all speakers, and by solid lines for the older speakers.

4. DISCUSSION AND CONCLUSION

In this study, the effects of age on vowel duration and formants were investigated using a dataset of French read speech produced by male and female speakers of different ages.

In line with what has been reported in the literature on aging and speech, the results also showed that speech, here measured as vowel duration, is slowed down with increasing age for both male and female speakers. For formants, no overall pattern of change with age was found for either male or female speakers. However, a moderate tendency toward centralization was found for older male speakers, and there were diverging patterns of change with age for female and male speakers in terms of vowel category variability and overlap between vowel categories. Specifically, category-wise variability decreased with age for male speakers, while this tendency was much weaker for female speakers. There was also a moderate but consistent tendency for acoustic overlap between oral and nasal vowels to decrease with age, with this reduction in overlap being reached earlier by female speakers. These findings suggest that aging may differentially affect the production of vowel sounds in men and women. They also suggest that in languages such as French in which the contrast between oral and nasal vowels is phonological, the acoustic separation between those categories may increase with aging. Further research is needed to understand the underlying mechanisms and potential consequences of these changes.

5. ACKNOWLEDGEMENTS

This work was supported by the French ANR ANR-10-LABX-0083 (Labex EFL) and ANR-21-CE28-0015-01 (PASDCODE) and by the Swiss NSF Sinergia Program (ChaSpeePro, CRSII5_202228).

6. REFERENCES

[1] L. A. Ramig and R. L. Ringel, "Effects of Physiological Aging on Selected Acoustic Characteristics of Voice," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 1, pp. 22–30, Mar. 1983, doi: 10.1044/jshr.2601.22.

[2] S. E. Linville, "The sound of senescence," *Journal of Voice*, vol. 10, no. 2, pp. 190–200, Jan. 1996, doi: 10.1016/s0892-1997(96)80046-4.

[3] R. J. Baken, "The Aged Voice: A New Hypothesis," *Journal of Voice*, vol. 19, no. 3, pp. 317–325, Sep. 2005, doi: 10.1016/j.jvoice.2004.07.005.

[4] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers," *Speech Communication*, vol. 52, no. 7–8, pp. 638–651, Jul. 2010, doi: 10.1016/j.specom.2010.02.012.

[5] B. J. Benjamin, "Phonological performance in gerontological speech," *Journal of Psycholinguistic Research*, vol. 11, no. 2, pp. 159–167, Mar. 1982, doi: 10.1007/bf01068218.

[6] R. J. Morris and W. S. Brown, "Age-related differences in speech variability among women," *Journal of Communication Disorders*, vol. 27, no. 1, pp. 49–64, May 1994, doi: 10.1016/0021-9924(94)90010-8.

[7] J. Ryalls, A. Cliche, J. Fortier-blanc, I. Coulombe, and A. Prud'hommeaux, "Voice-onset time in younger and older French-speaking Canadians," *Clinical Linguistics & Phonetics*, vol. 11, no. 3, pp. 205–212, Jan. 1997, doi: 10.3109/02699209708985191.

[8] M. P. Rastatter, R. A. McGuire, J. Kalinowski, and A. Stuart, "Formant Frequency Characteristics of Elderly Speakers in Contextual Speech," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 1, pp. 1–8, 1997, doi: 10.1159/000266431.

[9] M. Bilodeau-Mercure, V. Kirouac, N. Langlois, C. Ouellet, I. Gasse, and P. Tremblay, "Movement sequencing in normal aging: speech, oro-facial, and finger movements," *AGE*, vol. 37, no. 4, Jul. 2015, doi: 10.1007/s11357-015-9813-x.

[10] A. Hermes, J. Mertens and D. Mücke, "Age-related effec ts on sensorimotor control of speech production", *Interspeech*, 2018, pp. 1526-1530.

[11] B. L. Smith, J. Wasowicz, and J. Preston, “Temporal Characteristics of the Speech of Normal Elderly Adults,” *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 4, pp. 522–529, Dec. 1987, doi: 10.1044/jshr.3004.522.
[12] E. D. Mysak and T. D. Hanley, “Aging Processes in Speech: Pitch and Duration Characteristics,” *Journal of Gerontology*, vol. 13, no. 3, pp. 309–313, Jul. 1958, doi: 10.1093/geronj/13.3.309.
[13] P. H. Ptacek, E. K. Sander, W. H. Maloney, and C. C. R. Jackson, “Phonatory and Related Changes with Advanced Age,” *Journal of Speech and Hearing Research*, vol. 9, no. 3, pp. 353–360, Sep. 1966, doi: 10.1044/jshr.0903.353.
[14] J. D. Amerman and M. M. Parnell, “Speech timing strategies in elderly adults,” *Journal of Phonetics*, vol. 20, no. 1, pp. 65–76, Jan. 1992, doi: 10.1016/s0095-4470(19)30254-2.
[15] A. Bourbon and A. Hermes, “Have a break: Aging effects on sentence production and structuring in French”, *International Seminar on Speech Production,* 2021, pp.102-105.
[16] T. Thies, A. Hermes, and D. Mücke, “Compensation in Time and Space: Prominence Marking in Aging and Disease,” *Languages*, vol. 7, no. 1, p. 21, Jan. 2022, doi: 10.3390/languages7010021.
[17] D. D’Alessandro and C. Fougeron, “Changes in Anticipatory VtoV Coarticulation in French during Adulthood,” *Languages*, vol. 6, no. 4, p. 181, Oct. 2021, doi: 10.3390/languages6040181.
[18] R. D. Kent and H. K. Vorperian, “Static measurements of vowel formant frequencies and bandwidths: A review,” *Journal of Communication Disorders*, vol. 74, pp. 74–97, Jul. 2018, doi: 10.1016/j.jcomdis.2018.05.004.
[19] S. E. Linville and J. Rens, “Vocal Tract Resonance Analysis of Aging Voice Using Long-Term Average Spectra,” *Journal of Voice*, vol. 15, no. 3, pp. 323–330, Sep. 2001, doi: 10.1016/s0892-1997(01)00034-0.
[20] P.J. Watson and B. Munson. "Parkinson’s disease and the effect of lexical factors on vowel articulation", *The Journal of the Acoustical Society of America* 124, no. 5, 2008, EL291-EL295.
[21] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “‘Mini-mental state,’” *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, Nov. 1975, doi: 10.1016/0022-3956(75)90026-6.
[22] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, “Acoustic and respiratory evidence for utterance planning in German,” *Journal of Phonetics*, vol. 41, no. 1, pp. 29–47, Jan. 2013, doi: 10.1016/j.wocn.2012.08.007.
[23] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services,” *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017, doi: 10.1016/j.csl.2017.01.005.
[24] F. de Wet, K. Weber, L. Boves, B. Cranen, S. Bengio, and H. Bourlard, “Evaluation of formant-like features on an automatic vowel classification task,” *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1781–1792, Sep. 2004, doi: 10.1121/1.1781620.
[25] C. Gendrot and M. Adda-Decker, “Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German,” *Interspeech*, Nov. 2005, pp. 2453-2456.
[26] E. Ferragne and F. Pellegrino, “Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics,” *Journal of Phonetics*, vol. 38, no. 4, pp. 526–539, Oct. 2010, doi: 10.1016/j.wocn.2010.07.002.
[27] P. Boersma and D. Wennink, “Praat: doing phonetics by computer [Computer program]”, Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/.
[28] N. Audibert, C. Fougeron, C. Gendrot and M. Adda-Decker, "Duration-vs. style-dependent vowel variation: A multiparametric investigation", 1*8th International Congress of Phonetic Sciences*, 2015, pp. 5-10.

Phonation and Aging in White Hmong

Christina M. Esposito[a] , Kevin Schäfer[b] & Sameer ud Dowla Khan[c]
[a]Macalester College, USA [b]UCSB, USA [c]Reed College, USA
esposito@macalester.edu, kevinjschaef@gmail.com, skhan@reed.edu

ABSTRACT

Age-related changes in physiology such as the ossification of the ribcage cartilages [19, 25] and the thickening of the vocal folds [15] may change speakers' voices, such that older voices have greater variability [21, 22, 23] and phonatory instability [13]. Unlike previous linguistic work on aging, the current study examines the effects of aging in a language with contrastive phonation: White Hmong. Forty-five native speakers ranging from 23–71 years of age were recorded producing words with contrastive breathy, modal, and creaky phonation. H1*−H2* and cepstral peak prominence (CPP) were measured based on previous studies [6, 7, 11]. Results showed that phonation measures were surprisingly stable across age groups, while variability (dispersion) substantially increased with age. This demonstrates that speakers of a language with contrastive phonation can maintain production distributed around stable targets, even as control of articulation is reduced.

Keywords: phonation; White Hmong; aging

1. INTRODUCTION

The aging process produces a number of changes to the muscular system, the larynx, the vocal folds, and the respiratory system, all of which can affect speech. These include physical changes such as the ossification and calcification of the ribcage cartilages [19, 25], and the thickening of the vocal folds [15], as well as neurological changes such as decreased synaptic activity [27].

Studies have examined the effects of aging on speech, finding age-based differences in 1) f0 [4, 5, 16], 2) intonational patterns [1], 3) spectral mean for /s/ [28] and 4) overall variability [3, 19, 21, 22, 23, 27]. Most relevant for the current study, voice quality differences have also been observed: older voices are often perceived as hoarse or breathy [13, 26]. This might be due to an increase in vocal fold mass, especially for post-menopausal women [15, 16] and/or incomplete glottal closure in both older women and men [3, 16]. (Younger women also possess incomplete glottal closure, but in a different location than older women [24]). These age-related changes to the vocal folds and to the degree of glottal closure can result in an increase in noise and/or breathiness. [8] reported that elderly speakers' voices are characterized by lower harmonic-to-noise ratio, an indicator of increased noise, when compared to middle-aged and younger voices. In addition, [12] found that older speakers have lower cepstral peak prominence (CPP) values; as CPP is a measure of noise and/or aperiodicity, this result suggests more aperiodic/non-modal phonation. Furthermore, elderly speakers display increased phonatory instability, changing glottal configuration more frequently than young speakers [13].

All the aforementioned studies have focused on languages that do not contrast phonation type, raising the question of how aging might affect phonation when phonemic contrast is at play. In such a language, can we expect changes similar to those reported for languages like English? Or will the production of phonation types as indicated in the acoustic signal remain consistent across age? To answer these questions, we examined three phonemic phonation types (breathy, modal, creaky) produced by speakers of White Hmong ranging from 23–71 years of age, resulting in the first study of its kind to our knowledge.

2. THE LANGUAGE

White Hmong (*Hmoob Dawb*) is a Hmong-Mien language spoken in Laos and Thailand and by large diasporic communities in the US, Australia, and France. The phonetic properties of phonation (and by extension tone, as the two features are linked) in Hmong have been the subject of a number of studies [6, 7, 9, 10, 11, 17, 20]. White Hmong is described as contrasting seven tones: rising (45), mid (33), low (22), mid-rising (24), high-falling (52), high-falling breathy (42), and low-falling creaky (21). However, recent research on connected speech [11] proposed revisions to this description, finding that the high-falling breathy tone is more accurately 54 or 53.

Several acoustic measures of spectral tilt and noise have successfully distinguished phonation types in White Hmong, including: the amplitude of the first harmonic (H1*) [6], and the amplitude of the first harmonic minus the amplitude of the second harmonic (H1*−H2*) [6, 7, 20], where the asterisk indicates that the harmonic amplitudes were corrected for the effects of adjacent formants using [18], an

extension of [14]. [11] showed H1*−H2* did not adequately distinguish modal vs. non-modal phonation in connected speech and found CPP to be a better measure.

3. METHODS

3.1. Participants

Forty-five native speakers (20 female, 25 male) of White Hmong were recorded, with ages ranging from 23–71. They were born in Laos, Thailand, or the US, but all resided in the Minneapolis/St. Paul area of Minnesota, US, at the time of recording. The number of participants per age by decade is provided in Table 1.

Number of Participants	Age (by decade)
8	20s
12	30s
9	40s
3	60s
3	70s

Table 1: Number of participants per age by decade.

Twenty speakers over 60 were originally recorded, but they struggled with the task, and their data was unusable. All spoke English in addition to White Hmong; the reported age of English onset ranged from 5–57 years of age. All reported daily use of and literacy in White Hmong.

All speakers indicated their exercise level using guidelines from the US Centers for Disease Control and Prevention; all reported some degree of regular exercise. Speakers reported no history of smoking and no history of vocal fold disorders or strokes.

3.2. Recording

Speakers were recorded producing the list of 70 monosyllabic words used in [6]. These included all six vowels [i, e, ɨ, a, u, ɔ] and all seven tones of the language: 38 words had modal phonation, 17 had creaky phonation, and 15 had breathy phonation. Onset consonants varied, but were restricted to oral obstruents. Aspirated consonants and nasalized vowels were avoided. Words were read in the frame *rov hais* ____ *dua* [ʈɔ24 hai22 ____dua33] 'Say ____ again'.

3.3. Measurements

The measures used in the current study were CPP and H1*–H2*, as defined in section 2; these are the measures that successfully distinguished phonation in previous studies of the language [6, 7, 10, 11, 20].

3.4. Analysis

The data were modeled in terms of phonation and dispersion to determine if there were correlations between age and phonation, or between age and dispersion of phonation measures, calculated as the standard deviation. (Age was treated as a continuous variable.) Models were fit using one of the measures of voice quality (H1*−H2* and CPP) as the dependent variable; CPP was log-transformed for regularity. Based on exploratory statistics, the data from three of the 45 participants were removed, given insufficient evidence that they significantly distinguished more than two phonation types at any time point using either measure. The three unusable participants were all female and aged 23, 30, and 43.

The resulting linear mixed-effects regression models were fitted for each of the dependent phonemic phonation (breathy, modal, or creaky), variables (H1*−H2* or CPP) with fixed effects of time point of the measurement (beginning, middle, end), and age group, with random intercepts for individuals and age of English exposure.

4. RESULTS

4.1. Mixed effects models of phonation measures

Linear mixed effects regression models were fit to H1*−H2* and log CPP. A maximal model was reduced by eliminating interaction effects when doing so did not result in a significantly different model. To prevent overfitting, the random effects were subject to a Principal Components Analysis following the method outlined by [2]. This process eliminated the random effects of gender and exercise level, which at any rate were unbalanced, with reported exercise covariant with gender and age. Models without these random effects outperformed those that included them. Furthermore, adding gender as a random effect was deleterious to the predictive power of the models, which suggests that no effect of menopause was observable in the data. Age of English onset had little effect on the data. Thus, gender, exercise level, menopausal status, and age of English onset will not be discussed here.

The model for H1*−H2* (Fig. 1) was moderately successful in characterizing variance (conditional r^2=0.369). There were significant main effects of time point (χ^2=111.9, df=2, $p<0.01$) and phonation (χ^2=912.0, df=2, $p<0.01$). The main effect of age was not significant (χ^2=0.0223, df=1, $p<1.0$). All two-way

interactions were also significant: age and time point (χ^2=10.7, df=2, p<0.01); age and phonation (χ^2=6.29, df=2, p<0.05); and time point and phonation (χ^2=138, df=4, p<0.001). The effect sizes (regression coefficients) are, however, miniscule. The greatest change is found in creaky vowels, but mean H1*–H2* does not exceed an increase of 4dB over the span of five decades.

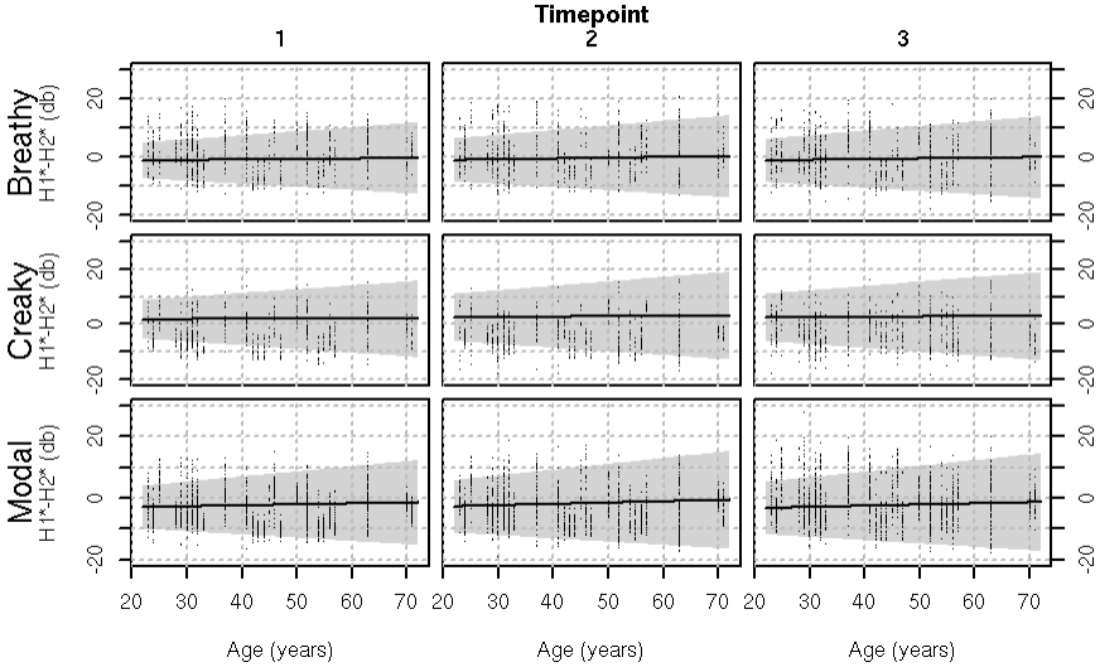


Figure 1: Token H1*−H2* (dB) values by participant age at each time point (1 = beginning, 2 = middle, 3= end of vowel) for each phonation type with a regression line and 95% confidence intervals (light gray).

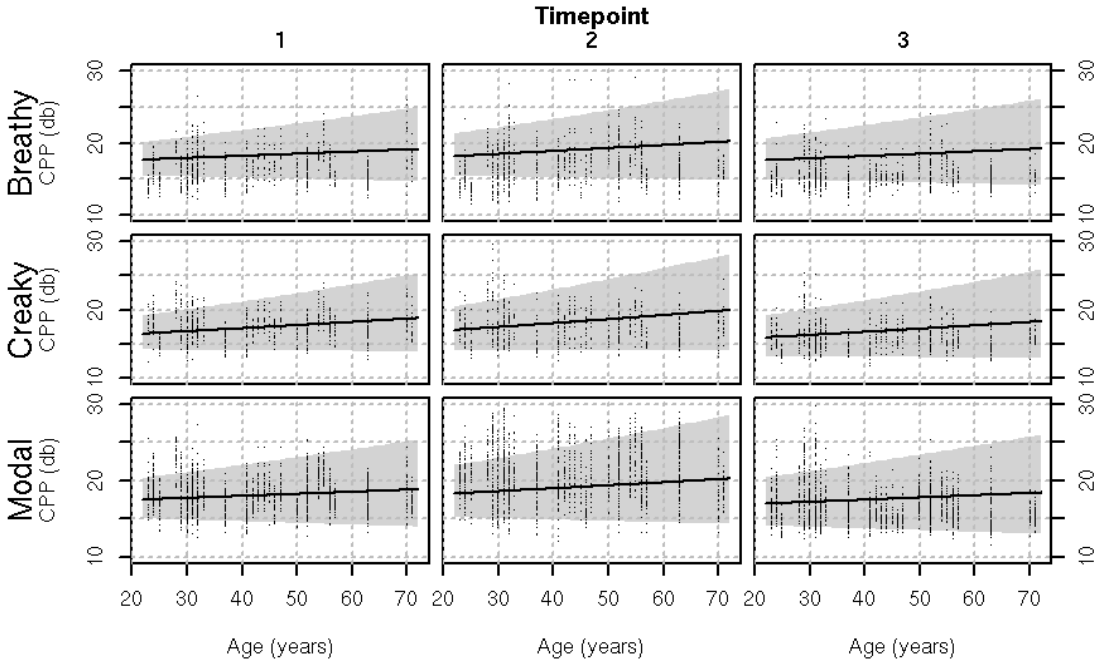


Figure 2: Token CPP (dB) values by participant age at each time point (1 = beginning, 2 = middle, 3= end) for each phonation type with a regression line and 95% confidence intervals (light gray).

The model for CPP (Fig. 2) characterized comparatively more of the variance (conditional r^2=0.475). As with the model of H1*−H2*, age was not a significant main effect (χ^2=0.0989, df=1, p<1.0), but there were significant main effects of time point (χ^2=2410, df=2, p<0.001) and phonation (χ^2=1308, df=2, p<0.001). All two-way interactions were also significant: age and time point (χ^2=36.21, df=2, p<0.001); age and phonation (χ^2=70.70, df=2, p<0.001); and time point and phonation (χ^2=551.8,

df=4, p<0.001). The effects were similarly small, but minute increases in CPP (less than 3dB over five decades) were seen across phonation types and time points with age, suggesting slightly *less* noise in older speakers' voices. Generally, CPP was highest across phonation types at the midpoint of the vowel. The interaction with the greatest effect size involving age resulted only in a ~2.88dB increase in CPP at the midpoint of breathy vowels (i.e. slightly less breathy at that time point).

4.2. Mixed effects models of dispersion

Finally, dispersion was quantified as the standard deviation for *each participant's* (untransformed) CPP measures at each timepoint within each phonation category. Another pair of linear mixed effects models was fitted to dispersion measures as dependent variables, with the same fixed effects of phonation type, time point, and age, with random intercepts for individuals and age of English onset. The model with H1*−H2* as dependent variable did not have any significant main effect of age or interactions with age and is not reported here.

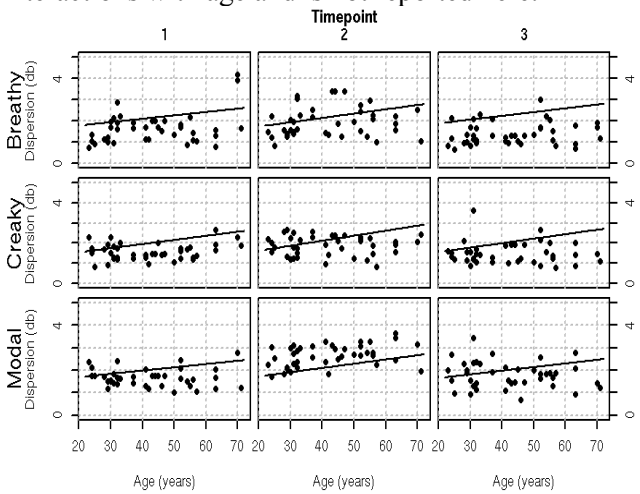


Figure 3: Dispersion of individual participants' CPP for each time point (1=beginning, 2=middle, 3= end); each phonation type is plotted with a regression line.

The model with CPP as the dependent variable (Fig. 3) was moderately successful at characterizing the variance (conditional r^2=0.402). All of the main effects were significant: age (χ^2=5.80, df=1, p<0.05), time point (χ^2=102, df=2, p<0.001), and phonation (χ^2=42.0, df=2, p<0.001). There was a significant interaction of time point and phonation (χ^2=23.9, df=4, p<0.001) and an interaction of age and time point (χ^2=5.77, df=2, p<0.1) approaches significance.

In this model, breathy vowels have the lowest dispersion for younger speakers at the midpoint, and other phonation types at the end of the vowel (and secondarily the midpoint). There is a trend for all phonation types throughout the vowel toward greater dispersion with greater age. This effect is most limited for modal vowels and strongest for creaky vowels. Dispersion at the midpoint is most affected, such that the model predicts that dispersion should be greater at the midpoints for participants in their 50s and older.

5. DISCUSSION AND CONCLUSION

When compared to research on languages without phonemic phonation contrasts, the results of the current study are mixed. Counter to expectation, this study finds no evidence that the voices of older White Hmong speakers are substantially more non-modal than those of younger speakers by either measure of phonation (H1*−H2* or CPP). In fact, the H1*−H2* models strongly suggest stable means across age groups. If anything, CPP measures suggest (negligible) *increase* with age, i.e. marginally *less* aperiodic noise in older voices.

However, in line with previous studies, older speakers' voices were found to be more variable than those of younger speakers. That is, the primary effect of age appears to involve increased variability (dispersion) rather than an overall change to or a failure to approximate a target level of phonation. This variability is strongest for non-modal voice qualities, in particular creaky tone.

It is possible that creak and breathiness in White Hmong are more affected by variability because they are already distinguished from modal phonation by tonal correlates. Indeed, previous research on tone identification in White Hmong has shown that listeners do not rely on creaky voice to perceive the low-falling creaky *-m* (21) tone, but instead, rely on f0 and duration [10]; this might explain why the variability effect in the current study is strongest for creaky vowels. As speakers of White Hmong age, they may allocate their limited capacity to coordinate and modulate articulation between different correlates of tone. Next steps in this research would be looking at variability in the tone, to determine if tonal patterns remain consistent as speakers age.

Replicating these methods across a diverse range of languages, particularly those that have different types of phonemic phonation contrasts, would shed light on the generalizability of these results. In future work, the inclusion of a non-tonal language with phonemic phonation contrasts, such as Gujarati, would provide an indication of the role of functional load, if any, in aging-related changes.

6. REFERENCES

[1] Barnes, D. R. 2013. Age-related changes in the production of linguistic prosody [Master's thesis, Purdue University].

[2] Bates, D., Kliegl, R., Vasishth, S., & Baayen H. 2015. Parsimonious Mixed Models. arXiv:1506.04967 [stat], arXiv: 1506.04967.

[3] Biever, D. M., & Bless, D. M. 1989. Vibratory characteristics of the vocal folds in young adult and geriatric women. *Journal of Voice* 3, 120–131.

[4] Brown, W. S., Jr., Morris, R. J., Hollien, H., & Howell, E. 1991. Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of Voice* 5, 310–315.

[5] Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. 2018. Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice 32*, 644.e1-644.e9.

[6] Esposito, C. M. 2012. An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics* 40, 466–476.

[7] Esposito, C.M., & Khan, S.D. 2012. Contrastive breathiness across consonants and vowels: A comparative study of Gujarati and White Hmong. *Journal of the International Phonetic Association* 42, 123-143.

[8] Ferrand, C. T. 2002. Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice* 16, 480–487.

[9] Garellek, M. 2012. The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics* 40, 152–161.

[10] Garellek, M., Keating, P., Esposito, C., & Kreiman, J. 2013. Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America* 133, 1078–1089.

[11] Garellek, M., & Esposito, C. 2021. Phonetics of White Hmong vowel and tonal contrasts. *Journal of the International Phonetic Association,* 1-20.

[12] Garrett, R.K. 2013. *Cepstral-and spectral-based acoustic measures of normal voices*. Doctoral dissertation, The University of Wisconsin-Milwaukee.

[13] Gorham-Rowan, M. M., & Laures-Gore, J. 2006. Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders* 39, 171-184.

[14] Hanson, H.M. 1995. Glottal characteristics of female speakers. Ph.D. dissertation, Harvard.

[15] Hirano, M., Kurita, S., & Nakashima, T. 1983. Growth, development, and aging of the human vocal folds. In D. B. J. Abbs (Ed.), Vocal fold physiology: Contemporary research and clinical issues (pp. 22–43). San Diego, CA: College-Hill Press.

[16] Honjo, I., & Isshiki, N. 1980. Laryngoscopic and voice characteristics of aged persons. *Archives of Otolaryngology* 106, 149–150.

[17] Huffman, M. 1987. Measures of phonation type in Hmong. *The Journal of the Acoustical Society of America* 81, 495–504.

[18] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America* 121, 2283-2295.

[19] Kahane, J. 1980. Age related histological changes in the human male and female laryngeal cartilages: Biological and functional implications. In V. Lawrence (Ed.), Transcripts of the ninth symposium care of the professional voice (pp. 11-20). New York: The Voice Foundation.

[20] Keating, P., Kuang J., Garellek, M., Esposito, C.M., & Khan, S.D. to appear. A cross-language acoustic space for vocalic phonation distinctions. *Language.*

[21] Linville, S. E., & Fisher, H. B. 1985. Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *The Journal of the Acoustical Society of America* 78, 40–48.

[22] Linville, S. E. 1988. Intraspeaker variability in fundamental frequency stability: An age-related phenomenon? *The Journal of the Acoustical Society of America* 83, 741–745.

[23] Linville, S. E., Skarin, B. D., & Fornatto, E. 1989. The interrelationship of measures related to vocal function, speech rate, and laryngeal appearance in elderly women. *Journal of Speech and Hearing Research* 32, 323–330.

[24] Linville, S. E. 2002. Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice* 16, 472–479.

[25] Malinowski, A. 1967. The shape, dimensions, and process of calcification on the cartilaginous framework of the larynx in relation to age and sex in the Polish population. *Folia Morphologica* 26, 118-128.

[26] Ptacek, P. H., & Sander, E. K. 1966. Age recognition from voice. *Journal of Speech & Hearing Research 9*(2), 273–277.

[27] Ramig, L. A., & Ringel, R. L. 1983. Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech and Hearing Research* 26, 22–30.

[28] Taylor S., Dromey C., Nissen S.L., Tanner K., Eggett D., Corbin-Lewis K. 2020. Age-Related Changes in Speech and Voice: Spectral and Cepstral Measures.

[29] *Journal of Speech Language Hearing Research* 63(3), 647-660.

[30] Shue, Y.-L., Keating, P., Vicenik, C. 2009. VoiceSauce: A program for voice analysis. *The Journal of the Acoustical Society of America* 124(4), 2221.

PHONETIC AND PRAGMATIC FEATURES OF THE FRENCH DISCOURSE MARKER *ET PUIS*: A LONGITUDINAL STUDY IN 10 FEMALE SPEAKERS

Annette Gerstenberg, Susanne Fuchs, Laura L. Koenig

Universität Potsdam, Potsdam (Germany), Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin (Germany), Adelphi University, New York (USA)
gerstenberg@uni-potsdam.de

ABSTRACT

We contribute to the growing field of phonetic studies on aging. Our data is not limited to laboratory settings, but comes from an extensive corpus of spontaneous speech data recorded over 10 years. We take the example of the frequent French discourse marker *et puis* 'and then'. Its occurrences are compared in 10 French female participants recorded at two timepoints 10 years apart. Following earlier work showing interaction between acoustic features and prosody, we assume a similar interaction with regard to discourse functions. We labelled all occurrences of *et puis*, accounting for differences in duration, f0, formants (F1, F2) and central discourse functions between the two timepoints. Our results show a significant and consistent effect of age for F2 and in some positions for F1. Duration marked selected discourse functions, no interaction was found for f0. We propose to consider this as a stabilizing effect of the frequent discourse marker.

Keywords: aging, discourse markers, vowels, duration, formants

1. INTRODUCTION

The inclusion of older speakers in age-related phonetic research has become more and more frequent both in cross-generational and longitudinal designs. Such studies explore a wide range of questions: age-related effects on speech acoustics; sociolinguistically meaningful variables and their age-specific use; cross-generational [1] or longitudinal data focusing on individual trajectories from a sociolinguistic perspective [2]. However, it remains a challenge to trace changes with age and at the same time account for the particular properties of spontaneous language.

Since many aspects of speech acoustics, including vowel formants, vary systematically with gender [3, 4] and increase statistical complexity, we base our study on a sample of 10 female speakers from a longitudinal corpus of biographical interviews recorded in France [5]. A previous longitudinal study on the French filler particle *euh* [6] established the interaction of aging mainly with prosodic positions, and suggested that aging may not be visible throughout the speech signal. An effect of age for f0 was only found in the final position of a segment, and also for F2, with both measures showing lower values at older ages. Using a similar approach, we will investigate another frequent element of spoken French, further exploring pragmatic functions, the discourse marker *et puis*. Discourse markers are known to be linked to generation-specific patterns [7]; e.g., French *enfin* meaning 'that is' is used to introduce a substitution marker in the context of potential age-related disfluencies [8]. The two parts of *et puis* /e pɥi/ are considered to operate pragmatically as a unitary discourse marker [9]. It is frequently used, and its pragmatic functions are well understood [10]. This allows for the composition of a highly comparable sample, in which the discourse marker is affected by age-related prosodic changes. Following earlier work, we expected to see an effect of aging between the two time points (t1 and t2) for F2 and interaction with pragmatic factors. While [11] found a significant age-related effect only for f0 and F1 (but not F2), other studies have found that F2 is also subject to age-related changes [1, 12]. The cited studies differ greatly in methods, in the composition of the underlying samples, and in syllable position and phonetic context; a comprehensive review found no evidence for an invariant pattern of age-related changes in vowel formants [13]. In cross-generational settings, age differences can be due to individual and social conditions. The advantage of the repeated recordings of the same speakers as used for the present study is the resulting high degree of speaker stability on many levels, such as education, lifestyle factors such as singing and smoking, communication styles and

basic health factors (although the latter may also differ between t1 and t2 in longitudinal designs). Finding interactions between phonetic measures of age and pragmatic functions could provide further evidence that acoustic features of aging do not appear across the board, but can be overridden by discourse functions. Prosody is known to be a resource in interaction, used to convey speaking styles, power relations or managing conflicts [14]. Analysis of entrainment, looking for tendencies of participants in a conversation to develop a similar behaviour as their counterparts, revealed that it can be found on the level of discourse markers, but only partly; some features did not participate in entrainment [15]. While some features are speaker-specific, other prosodic cues are characteristics of specific discourse markers and play a central role in fulfilling their pragmatic functions as the marking of discourse relations. Discourse markers such as French *et* 'and' as well as *alors* 'then' create discourse relations such as temporality, addition, consequence, concession, specification and topic shift. It has been shown that the actualisation of these discourse relations is supported by prosodic means: less frequently used discourse relations are marked by a stronger "prosodic segmentation" with long silent pauses, prosodic boundaries, lenghthening of the discourse marker and pitch reset, while cues of "prosodic integration" (no pause, less frequent prosodic boundary, no lengthening) are used in the case of unmarked, predictable discourse relations [16]. The setting used for the present study allows for the combination of the two approaches which, on the one hand, investigate potential interactions of age and the acoustic measures of a frequent discourse marker, and on the other hand observe interactions in the use of prosodic and phonetic cues for different discourse relations, again, with a potential interaction with age.

2. METHODOLOGY

The sample is comprised of two series of recordings with 10 female speakers showing no signs of cognitive impairment. Mean age was 75.7 for t1 (median=78) and 86 for t2 (median=88), covering three levels of education: four speakers had finished high school (Baccalauréat), 5 attended school until age 15/16 (Brevet d'études primaires) and 1 until age 12/13 (Certificat d'études). The interviews of the first (t1, 2005) and the second (t2, 2015/2016) series comprise 120 635 words. The selection of *et puis* was motivated by its corpus-specific frequency: after the conjunction *parce que* 'because' [17] with a frequency of 638 occurrences, *et puis* is the second most frequent connector, with 598 occurrences. We included instances of *et puis* in the analysis (n=431) where /e/ was voiced and labeled the onset of the first vowel, the burst and the offset of the second vowel in PRAAT [18].

2.1. Acoustic analyses

In PRAAT, formant and f0 measures were obtained automatically for the vowel /e/. Preliminary analyses showed that the vowel glide in *puis* was frequently quite short and its pronunciation placed on a continuum between the proper vowel glide and /i/. This, combined with carry-over effects of the plosive burst, meant that automatic formant and f0 measures frequently yielded unrealistic values or else were undefined. PRAAT analysis settings used the burg method, a window size of 25 ms, a maximum frequency range of 5500 Hz, and pre-emphasis. The default setting was to estimate 5 formants; that value was changed to 4 or 6 for individual tokens when the automatic output yielded unreasonable values based on visual and auditory assessment. The f0 analysis used the PRAAT default values of 75 Hz and 600 Hz as the minimum and maximum allowable values. Although data loss for *et* was less extreme than for *puis*, a number of tokens still yielded undefined or unrealistic values. The final number of measured tokens was 393 for f0 and 315 for formants in /e:/. We did not analyze *puis*, because of the extreme data loss.

2.2. Annotating discourse functions

For the pragmatic functions of discourse markers, different hierarchical models combining the levels of semantics, cognition and discourse segmentation have been proposed. In a recent model that yielded a high level of replicability [10], the authors propose a two-level annotation. On the first level, four different "domains" of discourse markers are identified. Upon application to our data two of them turned out to be relevant, the "ideational domain", "linked to states of affairs in the world, semantic relations between external events" and the "sequential domain", "linked to the structuring of local and global discourse segments such as topics and turns" (p. 11-12). On the second level, the model presents 15 different functions which can be flexibly combined with the domains on the first level.

Accordingly, *et puis*, corresponding to *and then* in the English translation from French, is labeled as "ideational" on the level of domain

and "temporal" on the level of functions in the following context: "I continued my studies and then uh at the second internship I didn't uh I didn't carry it out" [10] (p. 33). In order to avoid small numbers of occurrences and missing values for certain levels, we ultimately grouped them into four functional categories: the original categories of Addition (ADD) for discourse-new information; Monitoring (MNT) used to control the discourse flow, in our approach joined with hesitations; semantical categories (SEM) grouping the less frequent categories of, e.g., temporal and causal relations; Topic (TOP) for start / change of or return to a topic. As for the embedding of *et puis*, we also reduced the number of categories and annotated whether a pause preceded the discourse marker or not (Category Prae: Pause vs. Other) and did the same for the part following *et puis* (Category Post: Pause vs. Other).

2.3. Statistical treatment

Statistic analyses were computed in R [19] with Posit [20], using the packages *lme4* [21], *lmerTest* [22] and *ggplot2* [23]. Specifics of the models are given in the results section below. All values with alpha$\leq$0.05 were considered significant. For each dependent acoustic measure, linear mixed models were run, with speaker-specific random slopes. In every subsection, we focus on results with significant outcomes.

3. RESULTS

3.1. Does fundamental frequency in *et puis* change with aging and discourse function or prosody?

We obtained an overall f0 from /e:/ in *et puis* and investigated whether it changes significantly with the speakers' age as well as with the use of this word in discourse. We first ran a model with speaker-specific random slopes for the interaction between the two predictors. Since this showed a singular fit, we reduced the model to speaker-specific random slopes of age only, resulting in lmer(overall_f0)$\sim$ pragfunct*timepoint+(1+timepoint|speaker). Our findings revealed no main effect of age, no effect of pragmatic function, and no interaction between the two factors, as graphically displayed in Fig. 1. In a next step, we ran a model with age and the occurrence of pauses before or after *et puis*. All of them were similarly constructed as described above, only including Prae (Pause vs. Other) or Post (Pause or Other) as predictors. None of the obtained results reached significance.

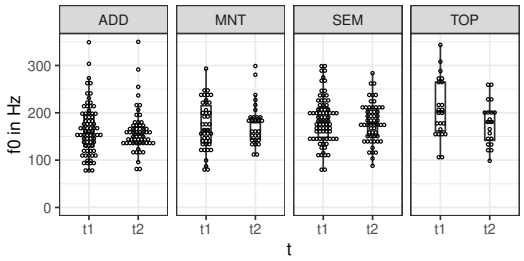


Figure 1: Boxplots with dot values showing overall f0 in Hz. Different subplots correspond to the four pragmatic functions. Addition (ADD), Monitoring (MNT), Semantic relation (SEM), Topic change (TOP).

3.2. Do formants change with aging and discourse function or prosody?

Models similar to the ones for f0 (pragmatic function * age; Prae * age; Post * age) were run for F1 and F2. While F1 did not change with age or pragmatic function, F2 showed a main effect of age only (Fig. 2). At the later time point (t2), speakers had on average a lower F2 (t=-2.39, p=0.0276*) which means they produced the /e:/ from a more retracted place of articulation.

For pauses, we found that there was no effect when a pause did or did not follow *et puis*, but we found a main effect (t=2.817, p=0.00517**) for F1 for the Prae condition (Pause or no pause was preceding). Moreover, an interaction with age was found as well (t=-2.163, p=0.0314*). The interaction reveals that with younger age (t1) F1 is lower than with older age (t2), but only in cases where no pause was present before *et puis*. If a pause was present before the target utterance, speakers started with a similar F1 value independent of their age.

When we ran the model with either pauses preceding (Prae) or following (Post) the target utterance and age, we found that F2 was always affected by age (Prae: t=-3.342, p=0.003201**, Post: t=-3.696, p= 0.00465**). Thus, the second formant is a very robust marker of age and can be found in all conditions, independently of whether they have certain pragmatic functions or occur in prosodically marked positions. Some additional effects were found: when a pause preceded the utterance, F2 was also higher than without a pause (t=3.687, p= 0.000268***). Finally, a marginal interaction was found (Post*age), when a pause followed the utterance, speakers produced no F2 difference at different time points, while they did when no pause was present t=2.017, p= 0.04458*).

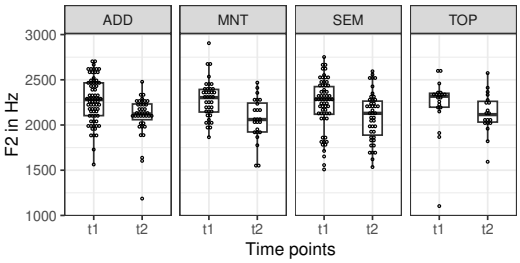


Figure 2: Boxplots with dot values showing F2 in Hz. Different subplots correspond to the four pragmatic functions. Addition (ADD), Monitoring (MNT), Semantic relation (SEM), Topic change (TOP).

3.3. Does duration change with aging and discourse function or prosody?

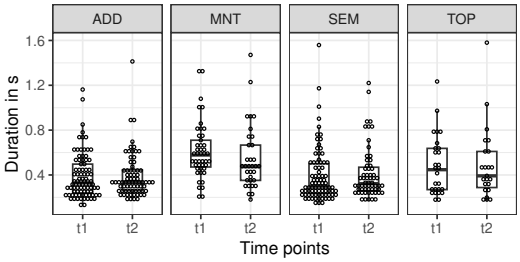


Figure 3: Boxplots with dot values showing the duration of *et puis* in seconds. Different subplots correspond to the four pragmatic functions. Addition (ADD), Monitoring (MNT), Semantic relation (SEM), Topic change (TOP).

While pragmatic functions had no imprint on the production of either f0, F1 or F2, i.e. spectral characteristics of the utterance, they did have an influence on the duration of said utterance. *et puis* was produced for a significantly longer duration in Monitoring (MNT: +0.443s) than in Adding a new topic (ADD: t=4.761, p=0.000269***). No other effects were found. Not surprisingly, the prosodic parameters (Prae and Post) also affected the duration of the utterance, but in all cases without an interaction with the speakers' age. When a pause preceded the utterance, the duration of *et puis* was slightly longer than when no pause was present (t=2.469, p= 0.014*). The same happened when a pause followed the utterance. Here we found again, a longer duration of *et puis* (t=3.634 , p=0.0104*). Thus, pauses surrounding the utterance caused general lengthening.

4. DISCUSSION

Changes in f0 did not turn out to differ significantly between t1 and t2. This could reflect the specific 10 year time-span for our older speakers which might be not be indicative of age-related f0 changes in female speakers: age-related changes develop on a continuum and are gradual in nature [1, p. 14]. Previous studies have often focused on younger vs. older speakers (see review in [6]). Our findings show that fine-grained analyses of changes within the limited timespan of 10 years refine our understanding of age-related changes in aspects of voice. Further, focusing on specific segments and/or words with different discourse functions may show different patterns from what has been attested in past work. Alternatively, we could assume a general stability of the basic measure of f0 linked to the frequency of this specific discourse marker. An intriguing finding is represented by the interaction of a lower F1 at t1 with specific discourse functions, i.e., exclusively after a speech sequence, not after a pause. This raises the question of the potentially changing mechanisms of coarticulation with age. This is also the case for F2 as a marker of age which was found to be very consistent. The finding that *et puis* was longer in the less frequent Monitoring (MNT) function than in the more frequent function of Adding (ADD) is in line with [16].

5. CONCLUSION

As in [6], F2 turned out to be a robust marker of age, found in all prosodic contexts and pragmatic functions. This was less true for F1, which turned out to be age-sensitive only in very specific contexts (after a preceding speech segment vs. pause). Related challenges for future research concern the involvement of F1 and F2 in coarticulation, and with regard to articulation rate and age. The duration of *et puis* was influenced by its pragmatic function, but not age. We interpret this finding in two directions: First, temporal phonetic properties mark pragmatic functions and second, pragmatic functions may be stable across the age span we investigated. Our results provide further evidence that certain prosodic positions show age effects while others do not. In turn, f0 of /e/ in the highly frequent discourse marker *et puis* did not show age-related change. In a next step, comparisons could be drawn to other portions of less routinized speech, where age-related changes may be more visibly reflected. Could it be the case, that the discourse marker *et puis* shows stable prosodic patterns as has been shown for pieces of formulaic language [24]?

6. REFERENCES

[1] C. Fougeron, F. Guitard-Ivent, and V. Delvaux, "Multi-dimensional variation in adult speech as a function of age," *Languages*, vol. 6, no. 4, p. 176, 2021.

[2] D. Bowie, "Individual and group trajectories across adulthood in a sample of Utah English speakers," in *Language variation and language change across the lifespan*, ser. Routledge studies in language change, K. V. Beaman and I. Buchstaller, Eds. New York, NY: Routledge, 2021, pp. 101–118.

[3] F. Karlsson and L. Hartelius, "On the primary influences of age on articulation and phonation in maximum performance tasks," *Languages*, vol. 6, no. 4, p. 174, 2021.

[4] J. Kreiman and D. Sidtis, *Foundations of voice studies. An interdisciplinary approach to voice production and perception*. Chichester and West Sussex and Malden (MA): Wiley-Blackwell, 2011.

[5] E. El Sherbiny Ismail, A. Gerstenberg, M. Lupica Spagnolo, F. Schulz, and A. Vandenbroucke, "L'âge avancé en perspective longitudinale et ses outils : Langage, un corpus au pluriel," *SHS Web Conf. (SHS Web of Conferences) - 8e Congrès Mondial de Linguistique Française*, vol. 138, pp. 10 003 [1–14], 2022.

[6] S. Fuchs, L. L. Koenig, and A. Gerstenberg, "A longitudinal study of speech acoustics in older french females: Analysis of the filler particle 'euh' across utterance positions," *Languages*, vol. 6, no. 4, p. 211, 2021.

[7] B. Murphy, *Corpus and sociolinguistics. Investigating age and gender in female talk*, ser. Studies in corpus linguistics. Amsterdam: Benjamins, 2010, vol. 38.

[8] C. T. Bolly, G. Christodoulides, and A. C. Simon, "Disfluences et vieillissement langagier. De la base de données Valibel aux corpus outillés en français parlé," *Corpus*, no. 15, pp. 1–13, 2016.

[9] L. Crible and L. Degand, "Co-occurrence and ordering of discourse markers in sequences: a multifactorial study in spoken French," *Journal of Pragmatics*, vol. 177, pp. 18–28, 2021.

[10] ——, "Domains and functions: A two-dimensional account of discourse markers," *Discours*, no. 24, 2019.

[11] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers," *Speech Communication*, vol. 52, no. 7–8, pp. 638–651, 2010.

[12] C. Oliveira, A. R. Valente, L. Albuquerque, F. Barros, P. Martins, S. Silva, and A. Teixeira, "The Vox Senes project: a study of segmental changes and rhythm variations on European Portuguese aging voice. IberSPPECH 24-25 March 2021," *IberSPEECH*, pp. 135–138, 2021.

[13] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice: Official journal of the Voice Foundation*, vol. 32, no. 5, pp. 644.e1–644.e9, 2018.

[14] A. Wichmann, "7. Prosody and pragmatic effects," in *Pragmatics of Society*, G. Andersen and K. Aijmer, Eds. de Gruyter, 2011, pp. 181–214.

[15] Š. Beňuš, "Conversational entrainment in the use of discourse markers," in *Recent Advances of Neural Network Models and Applications*, ser. Smart Innovation, Systems and Technologies, S. Bassis, A. Esposito, and F. C. Morabito, Eds. Springer International Publishing, 2014, vol. 26, pp. 345–352.

[16] I. Didirkova, G. Christodoulides, and A. C. Simon, "The prosody of discourse markers alors and et in French. A speech production study," in *Speech Prosody 2018*. ISCA: ISCA, 2018, pp. 503–507.

[17] V. Hekkel, *Eine soziolinguistische Betrachtung von 'parce que'-Strukturen in Synchronie und Diachronie*. Universität Potsdam: publishUP, 2021.

[18] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 9/10, no. 341–347, 2001.

[19] R Development Core Team, "R: A language and environment for statistical computing (r 4.2.2): R," MPI Göttingen, 2022. [Online]. Available: http://www.r-project.org

[20] Posit Software, *Posit (formerly RStudio) 22.12.0*. Boston (MA): PBC, 2022. [Online]. Available: https://posit.co/

[21] D. Bates, M. Mechler, B. Bolker, and S. Walker, *Package 'lme4'*. Repository CRAN: r-project.org, 2018.

[22] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, *Package 'lmerTest'*. CRAN: r-project.org, 2020-10-23.

[23] H. Wickham, *ggplot2: Elegant graphics for data analysis*, 2nd ed. New York (NY): Springer, 2016.

[24] A. E. Hallin and D. van Lancker Sidtis, "A closer look at formulaic language: Prosodic characteristics of Swedish proverbs," *Applied linguistics*, vol. 38, no. 1, pp. 68–89, 2017.

SPRAAKLAB: a mobile laboratory for collecting speech production data

Martijn Wieling, Teja Rebernik and Jidde Jacobi
University of Groningen, The Netherlands
{m.b.wieling,t.rebernik,j.jacobi}@rug.nl

ABSTRACT

In this paper, we discuss the specifications of a mobile laboratory, dubbed SPRAAKLAB, and how we use it for acquiring research-grade acoustic and articulatory data in the field, thereby providing access to participant populations which are otherwise hard to study. In addition, we illustrate how the mobile laboratory supports public engagement activities in combination with research data acquisition, allowing us to entertain and inform an interested audience about speech research, while simultaneously collecting speech production data from dozens of participants in a matter of days.

Keywords: Mobile speech laboratory, large-scale data collection, public engagement.

1. INTRODUCTION

In two of the research lines in our lab, we focus on (1) investigating regional language variation and (2) investigating disordered speech (articulation). Both lines require collecting high-quality acoustic and articulatory research data in populations that are more difficult to reach than undergraduate university students. Unfortunately, university-based facilities for collecting acoustic and articulatory data are less convenient when the goal is to study populations (such as dialect speakers, elderly speakers, or individuals with a speech disorder) which are relatively dispersed and do not always live close to the university, nor have a particular reason to regularly visit one.

As collecting acoustic and articulatory data is very hard to do remotely, one option (see [5]) is to use portable acoustic and articulatory recording equipment to visit participants at their homes or a nearby location (e.g., [6]). However, there are two disadvantages to this. The first is that setting up and packing the equipment when leaving is rather involved, especially when using costly equipment with many components such as a portable electromagnetic articulograph (EMA; e.g., [3]). The second disadvantage is that the locations where the equipment may be set up are variable, which may affect the quality of the recordings. For example, usually there is no sound-dampened room, which often results in background noise being present in acoustic recordings. Consequently, data cannot be obtained in a controlled manner, thereby negatively affecting the quality of the acquired data.

In this paper, we illustrate how using a custom-made mobile laboratory alleviates both issues. In addition, we explain how it lowers the threshold for potential participants to participate and how it enables us to inform a broader audience about speech research.

2. MOBILE LABORATORY SPECIFICATIONS

After a successful funding request in the beginning of 2020, our university board provided us with a grant of €150,000 to purchase a custom-made mobile laboratory. We spent an additional €35,000 for all additional hardware and equipment inside (excluding our EMA and ultrasound systems). Our faculty board additionally committed to paying the variable costs of operating the mobile laboratory (i.e., the diesel fuel, insurance, road tax, roadside assistance, the alarm system subscription, and yearly maintenance costs) which are estimated at about €15,000 per year. We found a coach work company (Carrosserie Akkermans B.V. in Oud Gastel, the Netherlands; the chassis was made by Coxx Mobile Systems B.V. in Rijen, the Netherlands) that constructed the laboratory according to our specifications. The whole process from the initiation of the design until delivery of the vehicle lasted about nine months (April 2020 to January 2021).

2.1. Outside: dimensions, safety, power and design

The mobile laboratory (SPRAAKLAB: SPeech Recorded Acoustically And Kinematically LABoratory) was constructed on the basis of a custom-made chassis with air suspension attached to a Fiat Ducato 3.5T Heavy Chassis Cabin. The air suspension allows the chassis to be lowered to facilitate the entrance of people with reduced mobility. This is additionally aided by an extendable step and a security handrail. Stabilizer legs (manually extendable) allow the van to be stabilized when parked. The total dimensions of the mobile laboratory are 7 meters in length, 2.20 meters in width (2.75 meters including the mirrors of the cabin), and 2.95 meters in height. The length of the laboratory area itself is about 4.50 meters. The inside floor-to-ceiling height is about 2.10 meters, which is comfortable. The weight of the van is just below 3,500 kilograms,

which means that it can be driven with a regular EU class B driving license. A 360-degree bird-eye camera system, a digital rear-view mirror, a rear cross-traffic collision warning system (for detecting objects when driving in reverse), as well as a loud beeping noise clearly audible at the outside of the van while driving in reverse, were installed to help with safely navigating. The mobile laboratory is furthermore secured with a Class-5 alarm system with a vehicle tracking system.

The van obtains its electricity by plugging in its power cable in an available external power socket (rated 230V, 16A). A 50-meter extension cable is present to allow for flexibility in placing the mobile laboratory. For emergency power, a Honda EU22i generator is stored in the van. The (locked) storage area for the generator and extension cable is only accessible from the outside of the van and also contains two power sockets. These sockets can be used to plug in two cable extension reels (also stored in the storage area), each of which has multiple power sockets. When electricity is needed outside, a cable extension reel can be positioned outside where needed (usually close to the entrance of the mobile laboratory), while the connecting cable of the cable box can be guided below the van, through a closable hatch, into one of the power sockets in the storage area.

The outside design of the van (designed by the last author; see Figures 1 and 5) includes photos showing the different techniques we use for our research, which allow us to explain those techniques to a general audience. In the summer, a large awning above the door provides shade during public engagement activities in front of the van (see also Section 3.2).



Figure 1: The outside of SPRAAKLAB. See Figure 5 for a photo of the other side.



Figure 2: The experimenter room inside SPRAAKLAB (fixed desk on the left, foldable desk on the right).

2.2. Inside: laboratory rooms

SPRAAKLAB contains two separate rooms, separated by a (soundproof) door. In addition to a control room (of 2 x 2.50 meters) for at most two researchers, the mobile laboratory contains a separate sound-dampened (approx. -40 dB) room of about 1.90 x 1.90 meters. Since we use EMA for our data collection, bitumen (as opposed to metal) was used for sound-dampening the room. A (soundproof) window allows the researcher to look inside the sound-dampened room. While the setup of the mobile laboratory has been chosen to optimally support speech (articulation) research, it is flexible enough to also be suitable for running reaction time, eye-tracking and EEG experiments for up to two participants simultaneously.

2.3.1. Experimenter room

The experimenter room contains two desks (with multiple power plugs and UTP ports close by) to allow two researcher(s) to be present simultaneously (see Figure 2). One of the desks is fixed, and the other is foldable (but can be secured while driving). The desk chairs can be securely attached to the two desks. Below the fixed desk, a cupboard contains two small high-performance PCs from which all audio, video and articulatory recording equipment can be controlled. Sound recording equipment (sound cards and sound processors) is on top of the fixed desk and connected to the PCs below. A network attached storage (NAS) system with several terabytes of hard drive storage, connected to a 16-port switch and an uninterruptible power supply, provides a robust storage system for all acquired research data.

In addition, the experimenter room contains a small kitchen with a sink and a cold-water tap (see Figure 3), a small refrigerator (as the glue we use for EMA studies needs to be stored at a low temperature), a heater, an air-conditioning system, and a small

portable toilet for emergencies stored in a cupboard. In addition, an intercom system is available to communicate with the participant inside the sound-dampened room. Finally, there is a large number of cupboards available to store equipment and supplies. The lights in the room are dimmable.



Figure 3: The kitchen of the experimenter room inside SPRAAKLAB.

2.2.2. Sound-dampened room

The sound-dampened room (see Figure 4) contains two foldable desks (to allow for dyadic experiments) and two chairs, all of which can be secured while driving. A 32" computer monitor is attached to a flexible arm to optimally position it in front of participants of different heights (including children). There are three pan-tilt-zoom IP cameras, placed above and below the monitor, as well as in the outer top corner of the room. These cameras allow the researcher to observe (and optionally record) the participant during an experiment. The camera below the monitor is useful for checking the stability of an ultrasound probe, for example, whereas the camera above the monitor is useful for checking whether EMA sensors remain attached to the tongue throughout the duration of the experiment. The camera in the top corner provides a clear overview of the entire room and is useful for checking in on the participant(s) during the experiment.

In addition, the sound-dampened room contains two high-quality microphones (Sennheiser SE66) attached to flexible arms that can reach all places where participants can sit. Other microphones (such as headset microphones) can be attached to the flexible arms if necessary. One of the walls of the room has a shelf to which our NDI Vox-EMA field generator can be attached. The room furthermore contains ample power sockets and UTP sockets, a stereo speaker system, and two headphone connections, as well as cupboard space for the EMA and ultrasound tongue imaging (UTI) equipment. All equipment is connected to and can be controlled from the fixed desk in the experimenter room (using audio, video, and UTP cables guided through sound-isolated ports). The lights in the room are dimmable (controlled separately from the dimmable lights in the experimenter room).



Figure 4: The sound-dampened room inside SPRAAKLAB.

3. SPRAAKLAB USE

3.1. Bringing the laboratory to the participant

When investigating the speech articulation of (for example) people with Parkinson's disease, it is often more convenient for them if data can be collected at their homes. However, recording participants in their own homes may negatively affect the quality of the data as there is a great deal of variability in the environment where the data is recorded. Using SPRAAKLAB, we are now not only able to travel to the participants' homes but also able to collect high-quality research data due to the constant, optimal environment. Rather than bringing our recording equipment into the field [5], we now bring an entire laboratory *setting*. In doing so, we do not only improve the quality and consistency of the collected data, but the data collection itself is also more efficient, as the time needed to set up and store equipment is reduced to a minimum (about 15-20 minutes for even the most complicated experiments). In addition, the use of SPRAAKLAB significantly lowers the participation threshold for our participants, as they do not have to travel to us, nor do they have to host us at their home. A disadvantage of using the mobile laboratory is that researchers first need to get used to driving a vehicle of this size (to prevent damaging it while driving), and need sufficient training to be able to use all equipment inside before data collection can start.

3.2. Public engagement

Besides using SPRAAKLAB to collect research data, we also employ it to participate in public engagement



Figure 5: SPRAAKLAB during a public engagement activity in 2021.

events. For these events, usually organized during the summer, we place a few tables and chairs in front of the mobile laboratory, underneath the awning (see Figure 5). At the tables, we then set up various activities for children and adults to learn about their speech in general, and our speech research in particular. For example, we usually show people the movement of their tongue using UTI (and also give them a printout of the ultrasound image as a keepsake) while explaining the basics of speech motor control. We promote regional language appreciation and use by demonstrating an app for learning the regional language, and an app which can predict where someone is from on the basis of the dialectal variants they use (similar to [2]). We additionally provide a short speech quiz (on a tablet), and some puzzles. Finally, we explain the workings of the speech production system using a physical model of the vocal tract, and by generating vowel sounds using an electrolarynx in combination with a sliding vocal tract model [1].

3.3. Combining public engagement and data collection

During most events we visit with our mobile laboratory, we also collect experimental data in the sound-dampened room. Visitors to the events are usually very interested in participating in a relatively short experiment (lasting up to 20 minutes; while people are enthusiastic to participate, being recruited on the spot also means they usually have limited time available), despite us not providing a monetary reimbursement for their time. This approach allows us to collect high-quality data from a substantial number of people in a relatively short time span.

For example, in three days we collected formant perturbation data for 30 Dutch children in a very short (five-minute) experiment during *Zwarte Cross* [4], which is the largest music festival in the Netherlands. At the cultural festival *Noorderzon* in Groningen (the Netherlands), we collected UTI data (likewise in a five-minute experiment) on the production of liquids in a sample of 70 adult participants during only six days. Finally, at the *Festival della Scienza* in Genoa (Italy) in October 2022, we collected formant perturbation data in a 15-minute experiment for over 90 Italian native speakers in the span of a week. For all experiments, ethical approval is always obtained beforehand, and people (in the case of children: their parents) give informed consent before participating.

While collecting data at an outreach event does result in less data per participant than in a regular experimental setup, it has the benefit of not only attracting many participants but also ensuring that we are more likely to get a diverse group of participants regarding age and educational background.

4. CONCLUSION

In this paper, we have illustrated how we have literally "take[n] the laboratory into the field" [5]. Our approach using SPRAAKLAB has the benefit of being able to easily access hard-to-reach populations while maintaining a strictly controlled laboratory environment in which sophisticated experimental speech production techniques can be employed. A further advantage of having access to SPRAAKLAB is that it can be used for combining public engagement activities and data collection. In this way, we can collect sufficient research-grade data with minimal effort in a span of only a few days.

5. ACKNOWLEDGEMENTS

We thank the board of the University of Groningen for financing the purchase of the mobile laboratory, and the board of the Faculty of Arts of the University of Groningen for providing the funding for the operational costs of SPRAAKLAB. We thank all Speech Lab Groningen members and others who have brainstormed with us while designing the van, and thank the photographer Miguel Santín as well as everyone who was willing to have their photo taken for the outside of the van. We thank both Coxx Mobile Systems B.V. and Carrosserie Akkermans B.V. for their excellent support during the design and construction of the mobile laboratory. Finally, we thank all organizers of festivals and public engagement events who have facilitated our participation with SPRAAKLAB.

6. REFERENCES

[1] Arai, T. 2016. Vocal-tract models and their applications in education for intuitive understanding of speech production. *Acoustic Science and Technology* 37, 148-156.

[2] Hilton, N. 2021 Stimmen: A citizen science approach to minority sociolinguistics. *Linguistics Vanguard 7*.

[3] Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., Wieling, M. 2021. A review of data collection practices using articulography. *Laboratory Phonology* 12, 1-42.

[4] Rebernik, T., Tienkamp, T.B., Polsterer, K.M., Hukker, V., Medvedeva, M., van der Ploeg, M., Schepers, I., Sekeres, H., de Vries, W., Abur, D., Jonkers, R., Noiray, A., Wieling, M. 2023. 5-minute formant adaptation task in Dutch children. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023).*

[5] Whalen, D., McDonough, J. 2015. Taking the laboratory into the field. *Annual Review of Linguistics* 1, 395-415.

[6] Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S.N., Baayen, R.H. 2016 Investigating dialectal differences using articulography. *Journal of Phonetics* 59, 122-143.

/R/ LENITION IN QUEBEC FRENCH : EVIDENCE FROM THE DISTRIBUTION OF 9 ALLOPHONES IN LARGE CORPORA

Lancien Mélanie[1,2], Hutin Mathilde[3,5], Stuart-Smith Jane[4], Adda-Decker Martine[2,5], Vasilescu Ioana[5]

[1] FNSRS, Switzerland, [2] LPP, France, [3] ILC-FNRS, Belgium, [4] GULP, UK, [5] LISN-CNRS, France
melanie.lancien@unil.ch, mathilde.hutin@uclouvain.be, jane.stuart-smith@glasgow.ac.uk,
martine.adda-decker@sorbonne-nouvelle.fr, ioana.vasilescu@lisn.upsaclay.fr

ABSTRACT

Lenition is a process whereby a segment shifts to a "weaker" variant (i.e., closer to deletion in the history of languages). Lenition is also a positional phenomenon, typically affecting intervocalic or coda consonants before post-coda or post-pausal onsets. While the lenition of stops is well studied in Romance languages, investigations about other segments are rare. We propose to fill this gap by focusing on /R/ in Québec French (QF), a variety documented to exhibit up to 9 allophones. We examine 50K+ read words from the PFC-Québec Corpus [1] that we manually annotated (for voicing, manner and place of articulation – based on perception and spectrograms). The analysis of the distribution of /R/ in different syllabic positions shows that lenited (approximantized, vocalized and non-realized) variants indeed appear in leniting positions (coda and intervocalic), thus showing that /R/ realizations in QF are not in free variation but indeed an instance of lenition.

Keywords: Lenition, French, rhotics, big data phonetics, laboratory phonology

1. INTRODUCTION

Lenition is a well-known process in historical phonetics whereby a segment shifts to a "weaker" variant. A variant is considered as "weaker" than its former realization if it can be found on the path from the former realization to complete deletion in the history of languages. Lenition and its opposite, fortition, are positional phenomena, i.e., some syllabic positions are, regardless of phonetic context, weakening positions, favoring lenition, while other positions are strengthening positions, favoring fortition, or at least resistance to lenition [2]. Typically, the intervocalic context (henceforth, V_V as /R/ in Fr. "bureau" /byRo/, *office*) and the coda position, either word-internal (V_C as in Fr. "verdit" /vɛRdi/, *turns green*) or word-final (V_# as in Fr. "dur" /dyR/, *hard*) are leniting contexts, as in (1), while the onset position, either word-initial (#_V as in Fr. "rouge" /Ruʒ/, *red*) or word-internal (C_V as in the family name "Malraux" /malRo/ *Malraux*[1]), are fortifying ones as in (2).

(1) Lat. /t/ → Fr. /∅/:
V_V = Lat. pa**t**er → pa**d**rem → pe**ð**re → Fr. père

(2) Lat. /t/ → Fr. /t/:
C_V = Lat. por**t**a → Fr. por**t**e

Studies on lenition in Romance languages however rarely focus on rhotics (to the exception of some studies on Catalan [3] or Brazilian Portuguese [4, 5, 6, 7]). The reason behind the scarcity of these studies probably lies in the fact that /R/[2] in Standard French shows little variation besides the well-known voicing assimilation (e.g., "quatre" [ka**tχ**], *four* but "cadre" [ka**dʁ**], *frame*), arguably an instance of lenition itself [8]. However, this relative uniformity is not shared by all varieties of French, and Canadian French shows an impressive amount of allophones in the realization of the rhotic ([9, 10] for instance).

In Canada, most of the native speakers of French are found in the province of Québec (inhabitants of Québec representing 23.2% of Canada's population). The last survey published by the "Institut de la statistique du Québec" about Francophonie (2016), stated that 79% of the population of Québec were identifiable as native speakers of French, i.e., approximately 6.4M out of 8.1M inhabitants. In Québec, the population is mainly distributed along the shores of the Saint-Laurent river, which leads to the terminology Eastern/Western Québec French in sociolinguistic studies (see [10] for instance). Another group of speakers is to be mentioned here: Acadian speakers, who are gathered in the North-East part of Québec and considered belonging to a different regiolect (see [11] among others). However, according to a demographic study conducted by Laval University [12], in the last decade, 97% of the population was concentrated in 20% of the territory, out of which 80% in the cities of Sept-Îles and Montreal.

Previous works [13, 14, 15, 16] show that there are both apical and dorsal realizations of

/R/ co-occurring as allophones in Québec French (henceforth QF). The use of these allophones varies across speakers depending on their region (apical variants being a marker of Western QF and uvular fricatives of Eastern QF) but also their age (younger speakers displaying close to no apical allophones), thus making /R/ a sociolinguistic marker [16] evolving over time (tending towards the disappearance of the apical variant). The articulation mode also varies (fricatives, trills, approximants...) resulting in up to 9 allophones of /R/ in Québec French [17].

In the present paper, we propose to fill the gap regarding the phonetics and phonology of /R/ in French, and in particular in Québec French. We aim to analyse 50k+ items pronounced by native QF speakers to identify their phonetic realizations and possible patterns of variation. In particular, we focus on positional factors to establish whether this variation in the realization of rhotics in QF can be considered an instance of lenition.

In the following, we present our data and methodology in Section 2, our results in Section 3 and our conclusions and discussion in Section 4.

2. METHOD

2.1. Corpus

For this study, we use a subset of the PFC-Québec corpus [1, 18], in which speakers had to read 2 word lists (max. 304 items)[3] specifically designed to bring out commonly discussed variation patterns in French phonology.

The corpus covers 32 locations, though only 29 were selected for this paper (we ruled out the data from incomplete list readings or from illiterate speakers). Among them, 13 are Western Laurentian, 12 Eastern Laurentian, and 4 Acadian, for a total of 396 speakers (206 females, 190 males) born between 1921 and 1999.

2.2. Coding

For this study, we manually coded the 58,083 exemplars of /R/ occurring in the recorded lexical words of the corpus. The purpose of this coding was to differentiate the 9 allophones of /R/ displayed by QF speakers. A trained phonetician listened to the recordings and looked at the spectrograms and oscillograms generated by Praat [19] in order to segment and code the phone. The choice of allophone was made according to both the perception of the annotator and a set of features:

- apical tap [ɾ] or trill [r][4]: Perceived as apical + 1 to 3 flappings in the spectrogram,
- uvular trill [ʀ]: Perceived as uvular + 2 to 3 flappings in the spectrogram,
- voiceless fricative [χ]: Friction noise and no F0 or voicing bar in the spectrogram,
- voiced fricative [ʁ]: Friction noise in the spectrogram and F0 detected + voicing bar,
- retroflex [ɻ]: Perceived as a retroflex + formants in the spectrogram,
- approximant variant [ɹ]: Formants in the spectrogram, glide-like,
- vocalized variant [ɚ]: Very stable formants in the spectrogram, schwa-like,
- deleted: No visible trace of a phone in the spectrogram - very often in clusters.

For the purpose of our analysis, the phonological contexts in which each phone appears is also automatically coded. One of the 6 contexts listed below (by order of strength) was attributed to each token depending on its position in the word:

- word-initial onset: #_V (e.g., rouge /Ruʒ/, *red*),
- part of an onset TR-cluster: C_V (e.g., bras /bRɒ/, *arm*),
- intervocalic onset: V_V (e.g., bureau /byRo/, *office*),
- word-internal coda: V_C (e.g., verdit /vɛRdi/, *turns green*),
- word-final coda: V_# (e.g., dur /dyR/, *hard*),
- part of a coda TR-cluster: C_# (e.g., quatre /katR/, *four*).

2.3. Data

For the present study, we excluded retroflex variants (only 21 tokens in a reduced number of word-forms) as well as word-final clusters (almost exclusively /tR/ clusters thus biasing the results), which results in a subset of 55,672 tokens.

The remaining variants are grouped into 4 categories according to strength of /R/[20]: Apical (apical trills [r] and taps [ɾ]), Uvular (uvular trills [ʀ]), Fricatives ([ʁ] and [χ]), and Lenified ([ɹ], [ɚ] and non-realized). Counts are given in Table 1.

3. RESULTS

In this section, we present our results regarding the distribution of our 4 /R/ categories (Apical, Uvular, Fricatives and Lenited) as a function of their syllabic position in the word. However, it appears that apical and uvular variants are not generalized among all speakers: Most speakers (n=153) use only the voiced and voiceless uvular fricatives and no

	Apical	Uvular	Fricatives	Lenited	Total
#_V	9.9% 16.6% 786	6% 33.1% 475	49.1% 16.5% 3917	35% 10.8% 2793	100% 14.13% 7971
C_V	10.6% 44.8% 2116	3.4% 47.8% 686	50.7% 42.8% 10159	35.3% 27.4% 7065	100% 36% 20026
V_V	10.7% 25.5% 1207	1.7% 13.1% 188	28.2% 13.3% 3169	59.4% 25.9% 6681	100% 20.2% 11245
V_C	6.8% 5.9% 281	0.5% 1.5% 21	46.9% 8.1% 1926	45.7% 7.3% 1876	100% 7.4% 4104
V_#	2.7% 7.1% 338	0.5 % 4.5 % 65	37.1% 19.2 % 4569	59.7% 28.5% 7354	100% 22.1% 12326
Total	8.5% 100% 4728	2.6% 100 % 1435	42.6% 100 % 23740	46.3% 100% 25769	100% 100% 55672

Table 1: Distribution of variants for QF /R/ according to predicted context of lenition (top to bottom). The rates in blue are the rates of each variant in the given position, the rates in green are the rates of each variant across positions. The numbers in black are the counts of each variant in each position.

apical nor uvular variants, some speakers (n=47) use almost exclusively the apical variant[5], a few fricatives and never the uvular, while other speakers (n=49) use the apical variant only marginally and prefer the uvular. We will therefore present the results for the "Fricative speakers" (21,523 tokens) in subsection 3.1, then for the "Apical speakers" (6,111 tokens) in subsection 3.2 and for the "Uvular speakers" (6,852 tokens) in subsection 3.3.

Our statistical analyses consist in multinomial logistic mixed models (mclogit package [21] under R 4.0). We started with a full model with speaker and word variables as random effects as in (3), and used model comparison through log likelihood comparison to find the best-fit model. In this paper we present only the results for linguistic context.

(3) Formula : $Coding \sim Context * Birthyear * Sex$

3.1. Fricative Speakers

Regarding speakers using exclusively fricative and lenited variants, as can be seen in Fig.1, lenited /R/ appears more in weakening positions, i.e., intervocalic onset and both word-internal and word-final codas, than in strengthening positions, i.e., word-initial onsets (Δ=28.2%, Δ=7.4% and Δ=17.0% respectively).

The multinomial mixed effect model shows that there is indeed a main effect for Context (χ^2(4, n=21,523)= 24.576, p<0.001).

Posthoc tests then show that fricatives are more likely to be used in every context except V_V and V_# where they are realized as often as lenited variants.

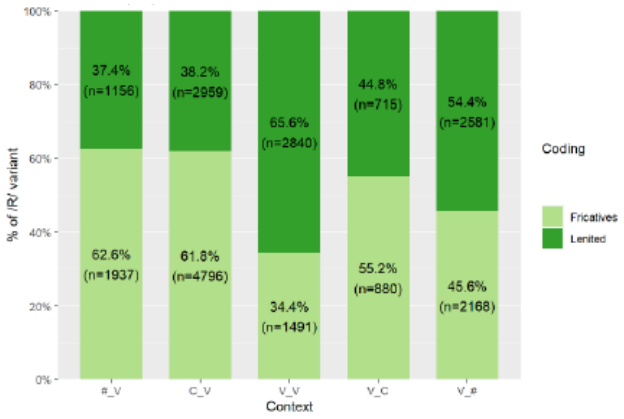


Figure 1: Rates of each variant of /R/ in QF according to predicted context for lenition (from left to right) for the subset of fricative /R/ users (n=153 speakers, n=21523 /R/).

It thus seems that /R/ variation patterns qualify as lenition, although it has not yet fully expanded from word-final to word-internal coda position, which is not surprising since the latter usually undergoes change later than the former [2, 22].

3.2. Apical Speakers

Regarding speakers with at least 5% apical /R/ in their productions, Fig.2 shows that they do not display any uvular trills, and few uvular fricatives. Moreover, it is clear that word-internal and word-final codas, i.e., weakening positions, display more lenited variants than word-initial onsets, i.e., strengthening positions (Δ=21.7% and Δ=58.0% respectively).

The multinomial mixed effect model shows that there is indeed a main effect for Context (χ^2(4, n=6,111)=17.3072, p<0.01).

Posthoc tests then show that all the tested pairs differ in each context (p<0.001). Thus, V_# is the context that triggers the most lenited variants (p<0.001), followed by V_C. Very few lenified variants are used in the other contexts, including the intervocalic position, where apicals are largely preferred. It therefore seems that Apical speakers also display a variation pattern that qualifies as lenition.

3.3. Uvular Speakers

Regarding speakers with at least 5% uvular trilled /R/ in their productions, as can be seen in Fig.3, they display very few apical variants and many uvular

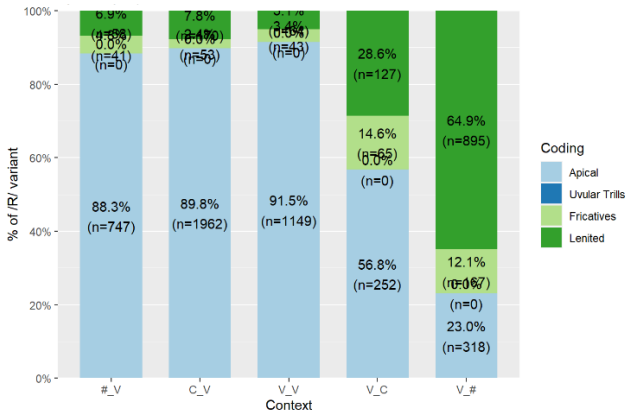


Figure 2: Rates of each variant of /R/ in QF according to predicted context for lenition (from left to right) for the subset of apical /R/ users (n=47 speakers, n=6,111 /R/).

fricatives. The rates of lenited variants are higher in all positions than those in Fig.2, suggesting that these speakers have a stronger tendency to lenition than Apical speakers. However, there are still more lenited variants in weakening positions, i.e., word-internal coda, word-final coda and even intervocalic positions, than in strengthening positions, i.e., word-initial onsets (Δ=22.9%, Δ=57.2% and Δ=28.9% respectively). One additional interesting observation is that rates of uvular fricatives seem rather constant across all positions while the variant losing ground *vis-à-vis* lenited variants is the uvular trill, which can therefore be identified as a properly "strong" variant rather than as a "default" variant.

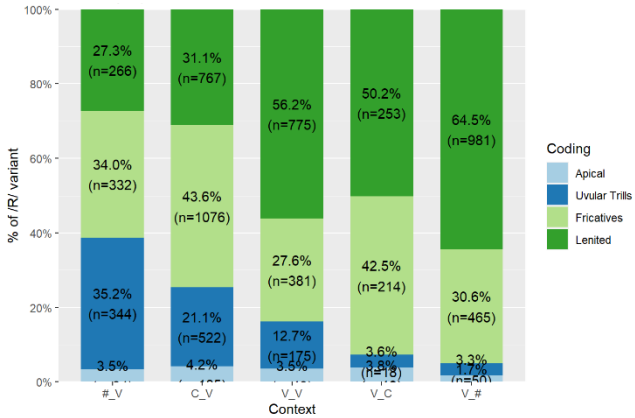


Figure 3: Rates of each variant of /R/ in QF according to predicted context for lenition (from left to right) for the subset of uvular trill /R/ users (n=49 speakers, n=6,852 /R/).

The multinomial mixed effect model shows that there is indeed a main effect for Context (χ^2(4, n=6,852)=39.105, p<0.001).

Posthoc tests then show that all rates are significantly different, except for Uvular Fricatives *vs* Uvular Trills in word-initial onset #_V positions, Uvular Fricatives *vs* Lenited variants in intervocalic V_V and word-internal coda V_C positions, and finally Apical *vs* Uvular Trills in word-internal V_C and word-final V_# coda positions. In sum, uvular fricatives are equally distributed with uvular trills in strong onset position but with lenited variants in weak intervocalic and internal coda positions, while apical and uvular trills are equally distributed in both word-internal and word-final coda position.

For these speakers too, then, the distribution of /R/ realizations qualifies as lenition, where the lenited variants are in quasi complementary distribution with a strong uvular variant.

4. CONCLUSION AND DISCUSSION

In the present paper, we analyze the distribution of 55,672 realizations of /R/ in Québec French with regards to their syllabic positions. We find that "weaker", i.e., approximant-like, vowel-like or even non-realized variants of /R/ indeed occur in syllabic positions identified by historical linguists as weakening positions, thus advocating in favor of an ongoing lenition of /R/ in this variety of French. Our work corroborates the idea of intervocalic and coda positions being triggering positions for lenition overall in Romance languages [2, 22]. It also sheds light on the fact that QF speakers can be grouped according to their preferred "strong" allophone of /R/ (i.e., apical, uvular, or fricative) and that these groups all behave similarly regarding lenition.

In future work, even though single transcribers have been shown to be internally consistent [23], additional corroboration from random sample coding will be undertaken. Future studies will also allow us to explore the geo- and socio-linguistic factors at play in the lenition of /R/ and even compare these patterns with those found in other varieties such as Metropolitan French. We also intend to provide observations regarding each /R/ variant separately, and to include in our models frequency and various other phonetic characteristics (duration, formants, speech-rate, etc.), which have been shown to correlate with lenition [24, 25].

5. ACKNOWLEDGEMENTS

This research was supported by the Swiss National Science Foundation, the Belgian Scientific Research Fonds (CR 40011029 (PPaDisM)), the French National Research Agency (ANR CLD20-25 and ANR-21-CE38-0019 (DIPVAR)) and the SPADE

consortium.

6. REFERENCES

[1] M.-H. Côté, "Le projet PFC et la géophonologie du français laurentien," J. Durand, G. Kristoffersen, and B. Laks, Eds. Nanterre : Presses Universitaires de Paris Ouest, 2014, pp. 173–198.

[2] P. Ségéral and T. Scheer, "Positional factors in lenition and fortition," in *Lenition & fortition*, J. B. de Carvalho, T. Scheer, and P. Ségéral, Eds. Mouton de Gruyter, 2008, pp. 131–172.

[3] D. Recasens and A. Espinosa, "Phonetic typology and positional allophones for alveolar rhotics in catalan," *Phonetica*, vol. 64, pp. 1 – 28, 2007.

[4] A. H. P. Silva and E. C. Albano, "Brazilian Portuguese rhotics and the phonetics/phonology boundary," in *ICPhS-14*, 1999, pp. 2211–2214.

[5] A. H. P. Silva, "Towards a dynamical representation for gradient allophony of Brazilian Portuguese rhotics," in *ICPhS-15*, 2003, pp. 1863–1866.

[6] I. Rennicke, "Variation and change in the rhotics of brazilian portuguese," Ph.D. dissertation, 2015.

[7] P. Howson and A. Kochetov, "Gestural Lenition of Rhotics Captures Variation in Brazilian Portuguese," in *Proceedings of Interspeech 2018*, 2018, pp. 182–186.

[8] E. R. Webb, "Minimalism and French /R/: Phonological representations in phonetically based phonology," *Journal of French Language Studies*, vol. 19, no. 1, pp. 87–115, 2009.

[9] G. Dulong and G. Bergeron, "Le parler populaire du Québec et de ses régions voisines," *Atlas linguistique de l'Est du Canada. Collection Études et dossiers Québec*.

[10] C. Verreault and T. Lavoie, "«La langue de nos gens» du Père Laurent Tremblay: Une première synthèse sur la variation géolinguistique du français parlé au Québec au début des années 1940," *Langues et linguistique*, no. 25, pp. 145–212, 1999.

[11] P. Balcom, L. Beaulieu, G. R. Butler, W. Cichocki, and R. King, "The Linguistic Study of Acadian French," *Canadian Journal of Linguistics/Revue canadienne de linguistique*, vol. 53, no. 1, pp. 1–5, 2008.

[12] "La situation démolinguistique du Québec," https://www.axl.cefan.ulaval.ca/amnord/Quebec-2demo.htm.

[13] J. Clermont and H. Cedergren, "Les "R" de ma mère sont perdus dans l'air," *Le français parlé. Études socio-linguistiques*, pp. 13–28, 1979.

[14] L. Santerre, "Les (r) montréalais en régression rapide," in *Les français régionaux du Québec, Numéro spécial de Protée VII (2)*, L. Lavoie, Ed., 1979, pp. 117–132.

[15] ——, "Des [r] montréalais imprévisibles et inouïs," *Revue québecoise de linguistique*, vol. 12, pp. 77–96, 1982.

[16] G. Sankoff and H. Blondeau, "Language change across the lifespan: /r/ in Montreal French," *Language*, pp. 560–588, 2007.

[17] M. Lancien, "More on the transition from apical to dorsal /R/ in Quebec," in *R-atics'7, International Congress on Rhotic Sounds*, 2021.

[18] J. Durand, B. Laks, and C. Lyche, "La Phonologie du français contemporain (PFC): usages, variétés et structure," C. Pusch and W. Raible, Eds. Günther Narr Verlag Tübingen, Germany, 2002, pp. 93–106.

[19] P. Boersma, "Praat: Doing phonetics by computer version 5.2. 03," *http://www. praat. org/*, 2015.

[20] E. R. Webb, "From r to ɾ via z and l: Rethinking the history of french/r," *High Desert Linguistics Societyâ¦ Albuquerque, New Mexico*, p. 55, 2002.

[21] M. Elff, M. M. Elff, M. Suggests, and L. Yes, "Package 'mclogit'," 2022.

[22] D. Steriade, *Phonetics in Phonology: The Case of Laryngeal Neutralization*. UCLA, 1997.

[23] J. Stuart-Smith, "A sociophonetic investigation of postvocalic/r/in glaswegian adolescents," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 1307–10.

[24] M. Pukli, "L'effet de la fréquence lexicale sur les réalisations des rhotiques en Ecosse," in *XXXIIe Journées d'Études sur la Parole*, Aix en Provence, France, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02530429

[25] Y. Wu, M. Hutin, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Extracting linguistic knowledge from speech: A study of stop realization in 5 Romance languages," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3257–3263. [Online]. Available: https://aclanthology.org/2022.lrec-1.348

[1] In French, /R/ is almost never an internal onset since sequences of an obstruent followed by /R/ usually result in a so-called *muta cum liquida*, or CR-cluster.

[2] Throughout the paper, we use a capital R to note the rhotic of French without posing an underlying representation. This notation is not the same as the one for the uvular trill ʀ, and is merely a way to refer to some sort of rhotic archiphoneme.

[3] Full lists can be found on the project's webpage https://public.projet-pfc.net

[4] Apical taps and trills are usually confused and merged as a single "apical" macro-category in classical studies on Quebec French, thus we confuse them in this paper but will address the possibility of the tap being a lenited variant of the trill in further analyses.

[5] A threshold was set to 5% apical realizations (for Apical speakers) and 5% uvular realizations (for Uvular speakers) to ensure that the use of these /R/ variants was not an articulatory "accident" but an actual use of the variant. The groups of Fricative, Apical and Uvular speakers are mutually exclusive.

PRODUCTION PLANNING HYPOTHESIS: EVIDENCE FROM PALATALIZATION IN AMERICAN ENGLISH

Małgorzata Kul and Kamil Kaźmierski

Faculty of English, Adam Mickiewicz University
kgosia@amu.edu.pl; kamil.kazmierski@amu.edu.pl

ABSTRACT

The study examines palatalization in American English (the Buckeye corpus) to test the predictions of the Production Planning Hypothesis: its prediction is that the rate of palatalization is positively correlated with smoothed conditional probability of the palatalizing context. We controlled for target segment, height of the vowel, grammar, speech rate,gender and age by including them as covariates in the mixed effects binomial logistic regression model.

The statistically significant effect of probability is in line with our prediction. This substantiates the claim that the size of the planning window increases the likelihood of reduction processes. Our findings provide further evidence for Production Planning Hypothesis from a non-reductory process.

Keywords: palatalization, yod coalescence, Buckeye, Production Planning Hypothesis, Smoothed Conditional Probability

1. INTRODUCTION

A number of phonetic and phonological processes operate across words. They give rise to variability in pronunciation: a speaker has a choice between the palatalized variant (*did you* /dɪdʒjə/) and the unpalatalized form (*did you* /dɪd jə/). This variability is restrained by the phonetic context of the cross word processes. The claim that speakers plan a pronunciation variant in advance, using cues from the previous and following words, is central to the Production Planning Hypothesis (PPH) [1, 2, 3, 4]. Th hypothesis does not assume that processes occur by default or at random. It presumes that processes and ensuing variants result from an interaction between planning-related predictors (e.g. the rate of speech, predictability of surrounding words, pauses) with the effects of the phonological context from surrounding words. PPH proposes that the predictors linked to speech planning are capable of modulating the size of the planning window and as such, can account for the variability encountered in pronunciation.

Previous studies have shown that in reduction processes at word junctures such as /t, d/ deletion [5] and flapping [4], the predictors of speech planning modulate the effects of the upcoming words on the preceding ones. As a consequence, this bleeds the random impact exerted by the following sounds. [5] found probabilistic effects of reduction and proved that planning the trigger word prior to planning the target word exhibits itself by higher deletion rate. The study by [4] reports that high probability of the following words is directly tied to flapping rate whereas no such link exists between predictability and glottaling process.

Our study adds the process of palatalization to redress the imbalance between widely studied reduction processes and non-reductory ones as palatalization is a special case of assimilation. We consider palatalization a good testing ground for Production Planning Hypothesis given that the process involves a significant amount of planning, i.e. more than one initial sound of the following word. In this connection, we extend the probabilistic implications to variants such as palatalized *dis yourself*. The study treats palatalization (also known as yod coalescence ([6]: 50) as an umbrella term covering a variety of assimilatory processes (see [7], [1] for a more extensive discussion on palatalization). In phonology, the change from the alveolar place of articulation in the lexical form into the post-alveolar one in the surface form is dubbed palatalization [8]. Both acoustic and articulatory oriented research takes palatalization to be a merger of two sounds, an alveolar and palatal when an increased gestural overlap occurs [9].

By palatalization we mean in this study the byproducts in the form of postalveolar fricatives /ʃ, ʒ/ and affricates /tʃ, dʒ/ if the palatal /j/ directly follows one of the alveolar consonants /t, d, s, z/. In terms of distribution, the palatalization process occurs word internally (*spiritual*) or across the boundaries of two words (*would usually*). Our study considers the latter pattern. The process results in variable outcomes and by variation, we understand the choice between two forms, palatalized (*as usual*

/æʒjuʒəl/) and unpalatalized (*as usual* /æzjuʒəl/). Following the probabilistic assumptions of PPH, the research question this paper pursues is whether speakers lay palatalization out in advance and if so, what the extent of the planning strategy is. We hypothesize that both smoothed conditional probability of $word_1$ given $word_2$ and smoothed conditional probability of $word_2$ given $word_1$ will be positively associated with palatalization rate.

2. METHOD

2.1. Data

The data come from the Buckeye corpus [10] of American English. It contains approximately 40 hs of speech, nearly 307,000 word types and 9,600 word tokens. We searched for all bigrams in which $word_1$ ended with /s, z, t, d/ and $word_2$ started with /j/. Corpus querying was performed with LaBB-CAT [11], manual annotation was performed by the first author by categorizing each token (N = 2,316) as palatalized (n = 1,136) or unpalatalized (n = 1,180). Categorization followed augmented analysis (both auditory and acoustic). Unclear cases were omitted if the spectrogram revealed no cues characteristic for fricatives before /j/. Data in the .csv format and scripts are available at https://osf.io/w47d9/.

2.2. Modelling

We fit a mixed effects binary logistic regression model of whether or not palatalization occurred, using the *lme4* package [12] in R [13].

2.2.1. Test variables

Previous studies found a positive correlation between contextual probability, operationalized as transitional probability, and palatalization (e.g. [14, 15]). Our study furnishes a more precise measure of contextual predictability, i.e. the smoothed conditional probability, which better captures the characteristics of infrequent lexical items. We applied Modified Kneser-Ney smoothing [16], using the *cmscu* package [17] and trained the smoothing function on the SUBTLEX-US corpus [18]. We included the probability of $word_1$ given $word_2$ and $word_2$ given $word_1$.

2.2.2. Control variables

- Target segment: /s, z, t, d/ [19]
- Height of vowel following /j/: high /u, ʊ, ɪ/ and non-high /ɝ, ɛ, ʌ, æ, ɑ, ɔ/ [20]
- Lexical vs. function words [21]
- Mean speech rate and speech rate deviation: [5]
- Gender, age and their interaction:gender expressed as male and female [22], age as two groups, young (up to 30) and old (above 40) [23]

We also included by-speaker and by-bigram random intercepts [24].

2.3. Results

Fixed effects	Estimate	p value
$(Intercept)$	-4.03	0.01
scp_giv_prev	6.61	< 0.001
scp_giv_fol	8.07	0.32
$targ\_segs$	2.11	< 0.001
$targ\_segt$	1.31	< 0.001
$targ\_segz$	3.08	< 0.001
$vowel\_heightnon-high$	-1.44	< 0.001
$gram\_wlex\_w$	0.48	0.131
$rate\_dev$	0.17	0.027
$mean\_rate$	0.55	0.132
$genderM$	-0.37	0.552
$ageY$	-0.81	0.271
$genderM : ageY$	0.73	0.275

Table 1: Model summary: p values calculated with likelihood ratio tests

We hypothesized that smoothed conditional probability given previous word (scp_giv_prev) raises likelihood of palatalization across word-boundaries. As can be seen in Fig. 1, scp_giv_prev is positively associated to probability of palatalization ($b = 6.6$, $p < 0.05$).

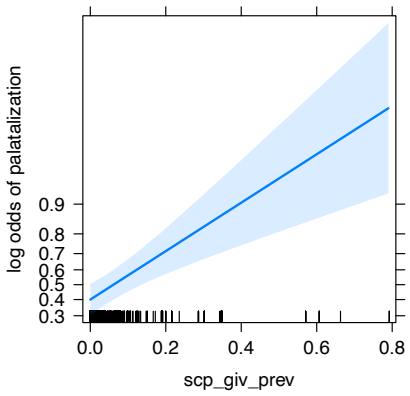


Figure 1: Partial effect plot of smoothed conditional probability of $word_2$ given $word_1$

We were also interested in the smoothed conditional probability given the following word. It turned out (Fig. 2) that the influence of the following word was not significant ($p = 0.32$).

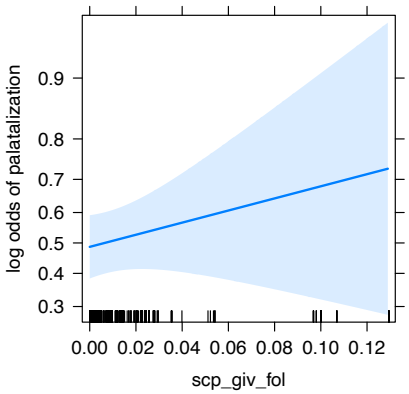


Figure 2: Partial effect plot of smoothed conditional probability of $word_1$ given $word_2$

Out of the control variables, included in our model, target segment and vowel height turned out to be statistically significant. The remaining control variables, i.e. grammar, rate deviation and mean, gender, age and the interaction between gender and did not come out as significant. For a full list of the estimates and p values, see Table 1.

Two of the control variables turned out to be significant, target segment and vowel height, and may therefore warrant further scrutiny. As for target segment, /d/ has the the lowest predicted rate of palatalization among the segments which are targets for the process. For all four target segments, see Fig. 3. Since the effects of target segment and vowel height do not directly pertain to our hypothesis, we do not discuss them further here.

Regarding vowel height, high vowels have a higher predicted rate of palatalization than non-high vowels do (see Fig. 4).

3. DISCUSSION

Production Planning Hypothesis assumes that the speaker plans articulation in advance as the application of processes does not happen by default [2], [4]. As a consequence, the speaker may choose the variant using cues from e.g. the preceding and following context, especially in processes which span word boundaries.

In our study, we aimed to extend the ramifications of the PPH on palatalization. So far, PPH has studied reduction processes such as liaison

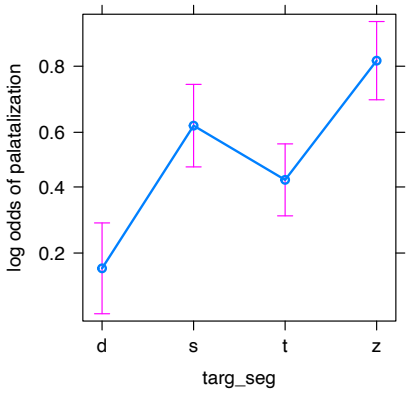


Figure 3: Partial effect plot of target segment identity

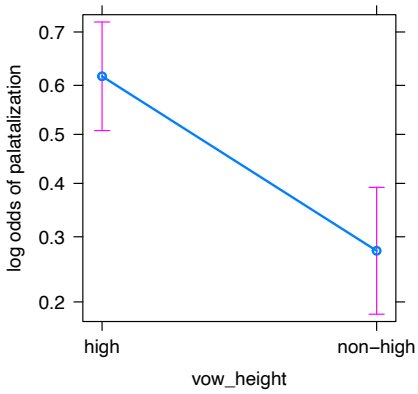


Figure 4: Partial effect plot of the vowel height

in French, flapping and /t,d/ deletion in English. Our contribution to the PPH is to test the planning window on a non-reductory process, i.e. palatalization.

We have positively verified our hypothesis, deriving from PPH, that smoothed conditional probability has a significant effect on palatalization, cf. Fig. 1. This validates the claim that the speakers from the Buckeye corpus indeed selected to apply palatalization in the target word in the situation where $word_2$ is predictable from $word_1$.

The other operationalization of contextual probability, that is the predictability of the $word_1$ given $word_2$ did not turn out to be significant. The latter was found for reduction of word duration [25] and perhaps the fact that ours is a non-reductory process, explains the discrepancy between ours and [25]'s result. Further support for this observation comes from [4] who found such effect for flapping

where sounds do not assimilate and substantiates our suggestion that reduction processes pattern differently from coalescence processes.

We have also observed a statistically significant effect of vowel height on palatalization [20], in that the high vowel group increased the probability of a palatalized variant, unlike the non-high vowels. This is a very promising outcome for PPH as we demonstrate that palatalization is sensitive to the height of the following vowel and as such, takes the following vowel to be a part of the planning window.

This result offers an interesting testing ground for PPH if we consider in a future study an interaction of smoothed conditional probability and vowel height. If the effect of the vowel is larger for higher scp values and smaller for lower scp values, it would constitute another piece of evidence that the size of the planning window is adaptable and truly context-sensitive.

4. CONCLUSIONS

Firstly, we found that in their choice of a pronunciation variant, speakers are informed by the predictability of the earlier phonetic material, relative to the target word. This provides evidence for Production Planning Hypothesis by establishing the size of the planning window: it spans the predictability of the preceding but not the following word. Secondly, vowel height predicts whether a variant is palatalized or not which seems to imply that speakers plan in advance to palatalize given the category of the upcoming vowel following /j/. We suggest that other processes may be tested within the frameworks of Production Planning Hypothesis to shed more light on speech planning. Another implication is to find a possible interaction between the vowel height and smoothed conditional probability given the previous word to produce a more comprehensive and explanatory model which captures palatalization in the probabilistic paradigm.

5. REFERENCES

[1] N. Bateman, "On the typology of palatalization," *Language and Linguistics Compass*, vol. 5, no. 8, pp. 588–602, 2011.

[2] M. Wagner, "Locality in phonology and production planning," *McGill Working Papers in Linguistics*, vol. 22, no. 1, pp. 1–18, 2012.

[3] A. Bürki, "Variation in the speech signal as a window into the cognitive architecture of language production," *Psychonomic Bulletin & Review*, vol. 25, no. 6, pp. 1973–2004, 2018.

[4] O. Kilbourn-Ceron, M. Clayards, and M. Wagner, "Predictability modulates pronunciation variants through speech planning effects: A case study on coronal stop realizations," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, p. 5, 2020.

[5] J. Tanner, M. Sonderegger, and M. Wagner, "Production planning and coronal stop deletion in spontaneous speech," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 15, pp. 1–39, 2017.

[6] J. Wells, *Longman Pronunciation Dictionary. Second edition*. Harlow: Pearson Education, 2000.

[7] D. N. S. Bhat, "A general study of palatalization," in *Universals of human language. Volume 2: Phonology*, J. H. Greenberg, Ed. Stanford, CA: Stanford University Press, 1978, pp. 47–92.

[8] M. Halle and K. Monahan, "Segmental phonology of modern english," *Linguistic Inquiry*, vol. 16, pp. 57–116, 1985.

[9] E. C. Zsiga, "Phonetic alignment constraints: consonant overlap and palatalization in English and Russian," *Journal of Phonetics*, vol. 28, pp. 69–102, 2000.

[10] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[11] R. Fromont and J. Hay, "LaBB-CAT: An annotation store," in *Proceedings of the Australasian Language Technology Association Workshop 2012*, Dunedin, New Zealand, 2012, pp. 113–117.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[13] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[14] J. Bybee, "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change," *Language Variation and Change*, vol. 14, pp. 261–290, 2002.

[15] M. E. Renwick and C. N. Cassidy, "Detecting palatalization in spontaneous spoken english," *Proceedings of Meetings on Acoustics*, vol. 23, pp. 1–10, 2015.

[16] S. Chen and J. Goodman, "An empirical study of smoothingtechniques for language modeling," Harvard University, Tech. Rep. TR-10-98, 1998.

[17] J. K. Davis and D. W. Vinson, *cmscu: A Count-Min-Sketch with Conservative Update implementation for R*, 2016, r package version 1.0. [Online]. Available: https://github.com/jasonkdavis/r-cmscu

[18] M. Brysbaert and B. New, "Moving beyond KuÄera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior Research Methods*,

vol. 41, no. 4, pp. 977–990, 2009. [Online]. Available: https://doi.org/10.3758/brm.41.4.977

[19] O. Glain, "The phonological fuzziness of palatalisation in contemporary english: A case of near-phonemes?" in *The Corpus Phonology of English: Multifocal Analyses of Variation*, A. Przewozny-Desriaux, S. Navarro, and C. Viollain, Eds. Edinburgh: Edinburgh University Press, 2020, pp. 50–73.

[20] L. M. Hyman, *Phonology: theory and analysis*. Holt, Rinehart and Winston, 1975.

[21] R. Shi, B. Gick, D. Kanwischer, and I. Wilson, "Frequency and category factors in the reduction and assimilation of function words: Epg and acoustic measures," *Journal of Psycholinguistic Research*, vol. 34, pp. 341–364, 2005.

[22] D. Byrd, "Relations of sex and dialect to reduction," *Speech Communication*, vol. 15, no. 1-2, pp. 39–54, 1994.

[23] W. D. Raymond, R. Dautricourt, and E. Hume, "Word-medial /t, d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors," *Language Variation and Change*, vol. 18, no. 1, pp. 55–97, 2006.

[24] K. Kaźmierski, "Prevocalic t-glottaling across word boundaries in Midland American English," *Laboratory Phonology*, vol. 11, no. 1, pp. 1–23, 2020.

[25] S. Seyfarth, "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," *Cognition*, vol. 133, no. 1, pp. 140–155, 2014.

LAXING HARMONY IN LAURENTIAN FRENCH: COARTICULATION OR PHONOLOGICAL PROCESS?

Phillip Burness[1], Beth MacLeod[2], Talia Tahtadjian[1], and Suzy Ahn[1]

University of Ottawa[1] and Carleton University[2]

ABSTRACT

In Laurentian French, high vowels are laxed in certain closed syllables, and several sources report that this laxing can optionally spread to preceding high vowels in open syllables. A past articulatory study suggested that this laxing harmony may be (at least partially) a coarticulatory phenomenon, although the study only analyzed words where harmony could apply just one step. We collected ultrasound data for chains of potential harmony targets and found that some speakers maintain an intermediate or complete degree of laxing throughout such chains, suggesting that such iterative harmony is phonological. We additionally observed a speaker who applies harmony in a somewhat non-local manner, harmonizing penultimate vowels only if they can also harmonize a word-initial antepenultimate vowel in the same word.

1. BACKGROUND

Laurentian French (henceforth LF) encompasses the varieties of French spoken in Ontario and Quebec, excluding other Canadian varieties such as Acadian French, which is spoken in the Maritime provinces. The LF dialect is well-known for a process of allophonic laxing that affects the high vowel phonemes /i/, /y/, and /u/ when they appear in certain closed syllables, and this process serves as the base for a reported case of vowel harmony [1, 2, 3, 4, 5] which is the focus of this paper. Examples of each process are given in (1) and (2). We start by summarizing some work on the distribution and acoustics of closed-syllable laxing before discussing the harmony pattern.

(1) Closed syllable laxing in a final syllable
- /pəti/ → [pə.t^{s}i] 'small (masc.)'
- /pətit/ → [pə.t^{s}ɪt] 'small (fem.)'

(2) Harmonic laxing of a non-final vowel
- /minyt/ → [mɪ.nʏt] 'minute'

All sources agree that laxing occurs before codas other than voiced fricatives (/v/, /z/, /ʒ/) and rhotics (/ʁ/), which are sometimes grouped together under the label of *lengthening* consonants in the LF literature. Some sources additionally claim that voiced fricatives and/or rhotics either trigger laxing or diphthongization. All the varying claims regarding lengthening codas are summarized in Table 1. Furthermore, the process is reported to be obligatory for closed final syllables but optional for non-final closed syllables [5].

/ʁ/ coda	/v/ coda	/z/ coda	/ʒ/ coda	Sources
L	L	L	L	[6, 7]
N	L	L	L	[8]
L	N	N	N	[9, 10]
N	N	N	N	[11, 4, 12]
D	D	D	D	[3, 13, 14, 15] [16, 17, 18]

Table 1. Claims regarding lengthening codas. L = laxing, D = diphthongization, N neither

As for the process's acoustics, every acoustic study of LF laxing finds that laxed vowels have a significantly higher first formant (F1) than their regular tense counterparts, and most acoustic studies find that laxed front vowels have significantly lower second formant (F2) and that laxed back vowels have significantly higher F2 [19, 20, 11, 6, 21, 7]. A minority of studies, however, find no F2 differences [22, 17, 18]. Based on these formant differences, we expect a lower tongue position (since F1 is negatively correlated with vowel height) and a more central tongue position (since F2 is negatively correlated with backness), although the single existing articulatory study by [22] saw only tongue lowering.

Normally, we expect laxing only on high vowels in closed syllables, although several sources report that high vowels in open syllables can take on a lax quality if followed by a laxed vowel under certain circumstances [1, 2, 3, 4, 5]. The most in-depth investigation of this harmony process was done by [5], who classified speakers into categories based on their answers to an acceptability judgment task. Given a word with a chain of 4 high vowels like /similityd/ 'similarity', there are 8 logical possibilites for laxing placement, but only 4 of these are

accepted by the LF speakers that [5] examined: absence of harmony (3a), local non-iterative harmony (3b), local iterative harmony (3c), and non-local harmony (3d). The location of secondary non-final laxing is underlined for clarity.

(3) Accepted types of harmony [5]
   a. Absent [si.mi.li.tˢʏd]
   b. Local, non-iterative [si.mi.l̲ɪ.tˢʏd]
   c. Local, iterative [s̲ɪ.m̲ɪ.l̲ɪ.tˢʏd]
   d. Non-local harmony [s̲ɪ.mi.li.tˢʏd]

(4) Unaccepted types of harmony [5]
   a. [si.m̲ɪ.li.tˢʏd]
   b. [si.m̲ɪ.l̲ɪ.tˢʏd]
   c. [s̲ɪ.mi.l̲ɪ.tˢʏd]
   d. [s̲ɪ.m̲ɪ.li.tˢʏd]

The other experimental investigation of LF harmony was done by [2]. Collecting acoustic and ultrasound data for disyllabic words, he found that penultimate high vowels which have potentially undergone harmony (as determined impressionistically) are articulated midway between fully lax and fully tense. He interpreted his results to mean that the spreading of LF laxing is less than harmony but more than co-articulation, representing an intermediate stage in the synchronic development of a true vowel harmony system.

While the findings for disyllabic words in [2] can certainly be interpreted to mean that LF harmony is not a fully phonological process, they do not necessarily preclude the possibility that harmony is fully phonological but causes an intermediate result. If laxing harmony is (partially) a phonetic process of co-articulation, we expect to observe a diminishing degree of laxing as we get further from the final syllable. Our study builds on that of [2] by including trisyllabic words with two goals:

- Ascertain whether laxing due to harmony is weaker in antepenultimate syllables
- Gain acoustic and articulatory evidence for the 4 types of speakers reported by [5]

2. METHODS

We analyze data collected from 7 native speakers of LF French (6 female) who read a word list aloud in a soundproof booth using a Micro ultrasound system with a Telemed MC4-2R20S-3 convex probe that was stabilized with a helmet. Ultrasound and audio data were collected simultaneously using Articulate Assistant Advance software [23], with the ultrasound probe recording at a framerate of approximately 82 Hz. Following the experiment, participants completed a basic demographic questionnaire. Data were collected from 4 additional people, but they were excluded from analysis because they spoke a different dialect of French (2 female), because their tongue imaged poorly (1 male), or because the recording session ended early (1 male).

The word list consisted of 58 words in total and was constructed for use in several experiments including the present one. The list was repeated 6 times, with a different randomization for each repetition. This randomization was the same across all participants. Of the 58 words, 12 were analyzed in the present experiment: /maʒi/ ‘magic’, /beni/ ‘blessed’, /avi/ ‘opinion’, /pɛʁmi/ ‘permit’, /bazilik/ ‘basil’, /ɑ̃piʁik/ ‘empirical’, /imyn/ ‘immune’, /sɔlityd/ ‘solitude’, /ʁidikyl/ ‘ridiculous’, /disimyl/ ‘conceal’, /pʁimitif/ ‘primitive’, and /miʁifik/ ‘fabulous’. This gave us 4 instances of each of the categories in (5). Note that we distinguish penultimate instances that begin a chain of high vowels (5c) from those that are in the middle of such a chain (5d). All vowels analyzed were instances of /i/ in order to make comparing tongue shapes across conditions more straightforward.

(5) Types of /i/ analyzed:
   a. Open final
      /maʒi̲/
   b. Closed final
      /bazili̲k/
   c. Chain-initial open penult
      /bazi̲lik/
   d. Chain-medial open penult
      /miʁi̲fik/
   e. Chain-initial open antepenult
      /mi̲ʁifik/

Using Praat [24], we marked the beginning and end of each vowel by hand, using these points to calculate the midpoint of each vowel as well as its duration. We then normalized duration values on a speaker-by-speaker basis using z-scores. Finally, tongue tracing was done by hand with Articulate Assistant Advance software [23], using the frame closest to the calculated vowel midpoint.

3. RESULTS

We began our analysis by inspecting tongue shape across syllable types. For each participant, a Smoothing Spline Analysis of Variance plot (SSANOVA: [25]) was created using the extracted

tongue splines. These plots used a polar coordinate system, as this makes comparisons roughly perpendicular to the tongue surface and thus easier to see and interpret [26]. For all but one participant (F06), final closed syllables were produced with a lower tongue body/blade and a more retracted tongue root than in final open syllables, as we would expect from the formant differences observed in the existing literature on LF laxing.

Speaker F06's tongue shape did not differ across any of the conditions. Visual inspection of her formant values showed higher F1 and lower F2 in final laxed syllables only, which would be indicative of a lower tongue body/blade and a more retracted tongue root. We thus tentatively classify this speaker as a non-harmonizer, although with the mismatch between her acoustic and audio data, it is not fully clear whether she truly laxes high vowels in final closed syllables.

The majority of participants (5 of 7) exhibit what we interpret to be local iterative behaviour. Speakers M01, F02, F03, and F07 displayed tongue splines that were intermediate between a fully open and fully laxed shape for both chain-initial and chain-medial penultimate open /i/, as was observed by [2]. Furthermore, speakers M01, F02, F03, and F06 produced antepenultimate open /i/ with the same intermediate shape as they did penultimate open /i/. Speaker F04 displayed a nearly laxed tongue shape for both types of penultimate open /i/. Speaker F04 also extended her penultimate behaviour to antepenultimate syllables, producing them with a fully lax tongue shape. To illustrate, Figure 1 provides the SSANOVA plots for speaker M01 (left) and speaker F04 (right).

The remaining speaker, speaker F05, displayed what we interpret to be a type of non-local behaviour. Her antepenultimate instances of open /i/ had a fully laxed tongue shape, as were her chain-medial penultimate instances. Crucially, her chain-initial penultimate /i/ displayed only a partially laxed tongue shape, suggesting that her harmony system will spread through a chain of high vowels specifically when it can reach the word-initial syllable. The intermediate tongue shape in chain-initial penults would, in this scenario, be coarticulatory in nature. It could also be driven by the word /imyn/ 'immune', where the chain-initial /i/ also happens to be in the first syllable; in all other words the chain-initial /i/ was in the second syllable.

To analyze the normalized duration values, we ran a mixed effects model in R [27] using the lme4 package [28]. Estimated p-values were obtained with the lmerTest package [29]. Each model included a fixed effect of syllable type, with final open syllables as the reference level. For random effects, we included by-participant intercepts and by-participant slopes for syllable type. To include by-item intercepts, each syllable of each word was treated as a different item (so, for example, the three vowels in /miʁifik/ 'fabulous' are considered to be three different items).

The duration model revealed that final closed syllables were shorter than final open syllables ($\beta = -0.98$, $t = -2.78$, $p = 0.012$). Penultimate syllables were even shorter, both for chain-initial instances ($\beta = -1.28$, $t = -2.82$, $p = 0.012$) and chain-medial instances ($\beta = -1.84$, $t = -5.14$, $p = 0.0000587$). Antepenultimate syllables were shorter than final closed syllables and longer than penultimate syllables ($\beta = -1.26$, $t = -3.06$, $p = 0.0064$). Figure 3 illustrates these differences. In light of the small number of participants, a reviewer suggested we do a post-hoc power analysis, although we chose not to do so since [30] advise that such analyses do not change the interpretation of the p-values.

4. DISCUSSION AND CONCLUSION

In his ultrasound study, [2] argued that the intermediate degree of laxing on penultimate open instances of /i/ meant that the process of LF laxing harmony was at least partially a phonetic process of coarticulation. If this were true, we would expect the degree of laxing to diminish when it spreads to antepenultimate syllables, although this is not what we found in our data. For three of the speakers that we categorized as local iterative harmonizers (following the terminology in [5]), the intermediate degree was maintained throughout the chain of high vowels, and for one speaker the entire chain exhibited complete or near-complete laxing.

One possibility here is that laxing harmony is indeed phonological, although it may create intermediate results for some speakers. Another possibility, though, is that the vowel space may be reduced in non-final syllables due to their shorter duration, making non-final /i/ inherently less high and front. In this latter scenario, what has been reported as harmony could in fact be a perceptual phenomenon whereby non-final /i/ is inherently lower and more central due to vowel space reduction and gets perceived as laxed when followed by an actually laxed vowel. The fact that non-final laxing was always complete for one of the iterative speakers (F04) suggests that her harmony at least is not due solely to a reduced vowel space, but to mediate conclusively between the possible explanations, a future study will compare non-final instances of open /i/

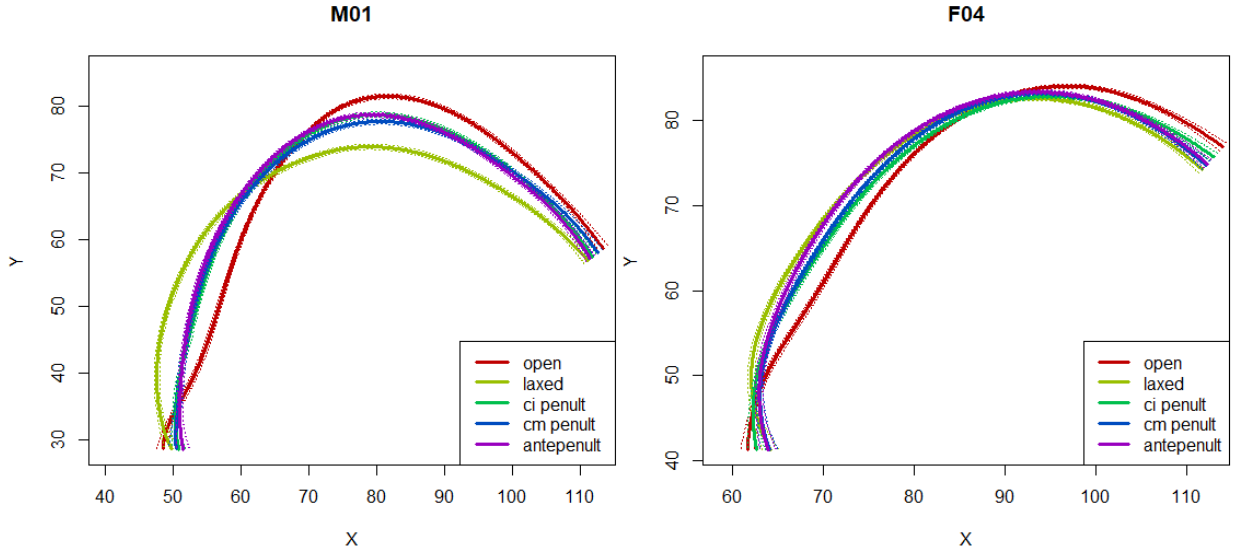


Figure 1. SSANOVA plots for speakers M01 (left) and F04 (right), ci = chain-initial and cm = chain-medial

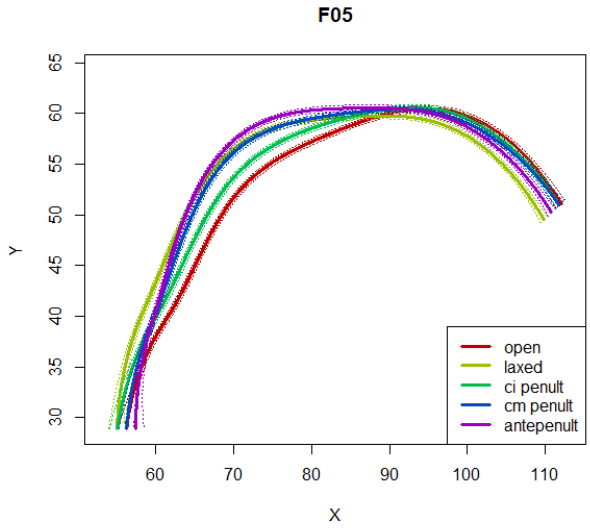


Figure 2. SSANOVA plot for speaker F05, ci = chain-initial and cm = chain-medial

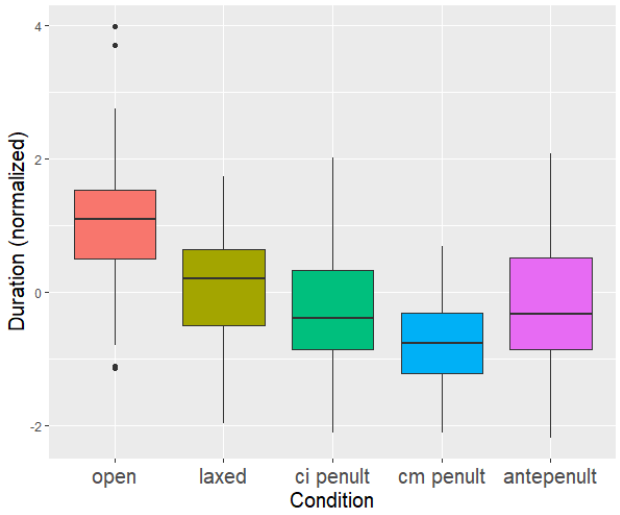


Figure 3. Duration by syllable type, ci = chain-initial and cm = chain-medial

that are followed by a laxed high vowel (e.g., /i̱myn/ 'immune') to those which are not (e.g., /si̱zo/ 'scissors').

That being said, one of our speakers exhibited non-local behaviour that would be difficult to explain as entirely due to a reduced vowel space in non-final syllables. This participant laxed penultimate instances of /i/ specifically when they could also lax word-initial antepenultimate /i/. This is not exactly as we would expect from past descriptions of non-local LF harmony [3, 5], though it is reminiscent of a pattern from Tutrugbu, where non-high prefix vowels undergo leftward [+ATR] harmony only when the leftmost prefix is also non-high [31].

To conclude, the main contributions of our study are that it provides:

- articulatory evidence that LF harmony is a phonological phenomenon which may produce an intermediate result for some speakers
- articulatory evidence for local iterative and non-local harmonizers described in [5]

We plan to collect more data from LF speakers in order to confirm and expand our findings.

5. REFERENCES

[1] A. Cuerrier and C. Reiss, "Geminates and vowel laxing in Quebec French," in *Romance Languages and Linguistic Theory 16: Selected Papers from the 47th Linguistic Symposium on Romance Languages*, I. Vogel, Ed., 2020, pp. 66–76.

[2] W. Dalton, "Deconstructing Canadian French vowel harmony," 2012, presented at the 36th Penn Linguistics Colloquium.

[3] D. Dumas, "Durée vocalique et diphtongaison en français québécois," *Cahier de linguistique*, vol. 4, pp. 13–55, 1974.

[4] D. C. Hall, "Contrast and phonological activity in the vowel system of Laurentian French," *Toronto Working Papers in Linguistics*, vol. 37, 2016.

[5] G. C. Poliquin, "Canadian French vowel harmony," Ph.D. dissertation, Harvard University, 2006.

[6] L. MacKenzie and G. Sankoff, "A quantitative analysis of diphthongization in Montreal French," *University of Pennsylvania Working Papers in Linguistics*, vol. 15, no. 2, p. article 11, 2010.

[7] C. Paradis, "An acoustic study of variation and change in the vowel system of Chicoutimi and Jonquère (Quebec)," Ph.D. dissertation, University of Pennsylvania, 1985.

[8] A. Boulanger, "Les parlers en [Iz]," *Revue québécoise de linguistique théorique et appliquée*, vol. 5, no. 4, pp. 129–142, 1986.

[9] M.-H. Côté, "Laurentian French (Quebec): extra vowels, missing schwas and surprising liaison consonants," in *Phonological Variation in French: Illustrations from Three Continents*, R. Gess, C. Lyche, and T. Meisenburg, Eds. Amsterdam: John Benjamins Publishing Company, 2012, pp. 235–274.

[10] J. Lamontagne, "Contrasting community and individual grammars in Laurentian French laxing," in *AMP 2021 Proceedings*, 2021.

[11] J.-D. Gendron, *Tendances phonétiques du français parlé au Canada*. Laval University Press, 1966.

[12] D. Walker, *The Pronunciation of Canadian French*. Ottawa: University of Ottawa Press, 1984.

[13] A. MacLaughlin, "Une (autre) analyse de la distribution des variantes des voyelles hautes en français montréalais," *Revue québécoise*, vol. 5, no. 4, pp. 21–60, 1986.

[14] M.-H. Côté, "La longueur vocalique devant consonne allongeante en contexte final et dérivé en français laurentian," in *Vues sur les français d'ici*, C. Lebland, F. Martineau, and Y. Frenette, Eds. Quebec: Presses de l'Université Laval, 2010, pp. 49–75.

[15] G. Leblanc, "Une étude acoustique des voyelles orales susceptibles d'être diphtonguées en français québécois," Master's thesis, Université Laval, 2012.

[16] L. Ostiguy and C. Tousignant, *Le français québécois: normes et usage*. Montréal: Guérin Universitaire, 1993.

[17] C. Sigouin and V. Arnaud, "Analyse acoustique des voyelles fermées tendues, relâchées et allongées en français québécois: la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique," in *Actes des XXXe journées d'étude sur la parole*, 2014.

[18] ——, "Quebec French close vowels in lengthening contexts: tense, lax or diphthogized?" in *Proceedings of ICPhS 2015*, 2015.

[19] V. Arnaud, C. Sigouin, and J.-P. Roy, "Acoustic description of Quebec French high vowels: First results," in *Proceedings of the 17th ICPhS*, 2011, pp. 244–247.

[20] J. Dolbec, C. Ouellon, and M. Ouellet, "Considérations préliminaires sure le trapèze vocalique de locuteurs québécois," *Dialangue*, vol. 4, pp. 44–48, 1993.

[21] P. Martin, "Le système vocalique du français du québec: De l'acoustique à la phonologie," *La Linguistique*, vol. 38, no. 2, pp. 71–88, 2002.

[22] W. Dalton, "An ultrasound imaging study of the tense-lax distinction in Canadian French vowels," *Canadian Acoustics*, vol. 39, no. 3, pp. 168–169, 2011, proceedings of the Acoustics Week in Canada.

[23] Articulate Instruments Ltd, *Articulate Assistant Advanced version 220.2.0*, Edinburgh, UK, 2022. [Online]. Available: http://www.articulateinstruments.com/

[24] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, 2022. [Online]. Available: https://www.fon.hum.uva.nl/praat/

[25] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[26] J. Mielke, "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons," *Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2858–2869, 2015.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[30] L. Kumle, M. L.-H. Võ, and D. Draschkow, "Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r," *Behaviour Research Methods*, vol. 53, pp. 2528–2543, 2021.

[31] A. G. McCollum, E. Baković, A. Mai, and E. Meinhardt, "Unbounded circumambient patterns in segmental phonology," *Phonology*, vol. 37, pp. 215–255, 2020.

Auditory and written perception of Japanese sequential voicing*

Timothy J. Vance[1], Manami Hirayama[2] and Hinako Masuda[3]

[1] National Institute for Japanese Language and Linguistics, [2,3] Seikei University
[1] tjvance825@gmail.com, [2] hirayama@fh.seikei.ac.jp, [3] h-masuda@st.seikei.ac.jp

ABSTRACT

Japanese *rendaku* affects second elements (E2s) of compounds, replacing a voiceless obstruent with a voiced obstruent (/**t**ama/ 'ball', /me+**d**ama/ 'eye-ball'). Lyman's Law is ordinarily stated as a constraint preventing *rendaku* when E2 already contains a voiced obstruent. An alternative proposal is that the constraint is orthographic, preventing a second diacritic in the *kana* representation of an E2 that already contains one. The crucial cases involve E2s with a medial /p/: a voiceless obstruent written with a diacritic. We report two experiments involving compounds with made-up E2s. Speakers chose between forms with and without *rendaku*. In an audio experiment, E2s with medial /p/ patterned with E2s with medial voiceless obstruents, but in a written experiment, E2s with medial /p/ patterned between those with medial voiceless obstruents and those with medial voiced obstruents. These results lend credence to the orthographic version Lyman's Law, but only when orthographic information is visually present.

Keywords: sequential voicing, voicing, Japanese, perception, orthography

1. INTRODUCTION

Japanese (i.e., modern Tokyo "standard" Japanese) has a well-known set of morphophonemic alternations known collectively as *rendaku* (sometimes translated as 'sequential voicing') [8]. Many morphemes have one allomorph that begins with a voiceless obstruent and another allomorph that begins with a voiced obstruent, as in (1).

(1) a. /tara/ 'cod'
b. /tara+ko/ 'cod roe'
c. /boH+dara/ 'dried cod fillet'

The initial voiced obstruent in the /dara/ allomorph in (1c) is an instance of *rendaku*, and the prototypical environment for *rendaku* is the beginning of the second element of a two-element compound. Only the non-*rendaku* allomorph of an alternating morpheme can occur word-initially, as in (1a) and (1b), with /tara/ rather than /dara/.

Although each *rendaku* alternation pairs a voiceless obstruent with a voiced obstruent, the phonetic difference between the two phonemes in each pair is in many cases more than just the absence vs. presence of voicing [17]. All the Japanese obstruent phonemes except /p/ are involved in these alternations. Each voiceless obstruent phoneme other than /p/ (i.e., /f h t c s č š k/) appears word-initially in at least one pairing, and each voiced obstruent (i.e., /b d z ǰ g/) appears as a *rendaku* partner in at least one pairing.

Lyman's Law is a constraint on *rendaku*. The now standard formulation is that Lyman's Law prohibits *rendaku* in the second element (E2) of a compound if that element already contains a medial voiced obstruent (e.g., */umi+ga**z**e/ 'sea breeze'; cf. /kaze/ 'wind'). Only a few exceptions to Lyman's Law are found in the existing vocabulary [6, 9, 14].

The modern Japanese writing system incorporates two parallel subsystems, *hiragana* and *katakana*, each of which makes it possible to represent the pronunciation (i.e., the phonemic form) of any Japanese word accurately. Together, these two subsystems are known as *kana*, and they are basically moraic: for the most part, each symbol represents a short syllable or half of a long syllable, although some moras are written with digraphs. Orthographic norms require a few minor mismatches between kana "spelling" and pronunciation [16], but these norms can be violated when the goal is to convey pronunciation. Both *hiragana* and *katakana* use the diacritic ⟨ ゛ ⟩, known as *dakuten*, to represent moras that begin with a voiced obstruent (as in だ for /da/ vs. た for /ta/). As noted, /p/ is the only modern Tokyo phoneme realized as an obstruent that is not involved in a *rendaku* alternation. The phoneme /b/ alternates not with /p/ but with /f/ or /h/. The term *han-dakuon* 'half voiced sound' denotes a mora that begins with /p/, and such a mora is written in *kana* with the diacritic ⟨ ゜ ⟩, known as *han-dakuten*, added to a symbol for a mora that begins with /h/ or /f/ (as in ぱ for /pa/ vs. は for /ha/ and ば for /ba/).

There is one conspicuous difference between Lyman's own statement of Lyman's Law [7] and the now standard version: Lyman's list of E2-medial consonants that inhibit *rendaku* includes /p/ in addition to all the voiced-obstruent phonemes of modern Japanese. The existing vocabulary, however, provides little reason to believe that E2-medial /p/ inhibits *rendaku* [18].

Most present-day linguists simply ignore /p/ when discussing *rendaku* and Lyman's Law, but it should

be noted that the modern Japanese vocabulary contains very few examples of *rendaku* in an E2 with a medial /p/. The almost total absence of such examples lends plausibility to the notion that an E2-medial /p/ inhibits *rendaku*. It is not far-fetched to imagine that an E2-medial /p/ might have a psychologically real inhibiting effect on *rendaku* for present-day native speakers of Japanese.

Recent work by [5] shows that native speakers seem to treat /p/ and voiced obstruents alike in triggering the optional devoicing of underlyingly voiced "geminates," that is, sequences of the moraic obstruent /Q/ followed by a voiced obstruent within the same element. Such sequences are confined to recent loans [15], and the examples below in (2) illustrate the putative pattern.

(2) a. /beQdo/ ベッド 'bed' [bed:o]~[bet:o]
b. /paQdo/ パッド 'pad' [pad:o]~[pat:o]
c. /heQdo/ ヘッド 'head' [hed:o] (*[het:o])

Given the optional devoicing pattern in (2), [4] suggests that it might be better to view *rendaku* as an orthographic phenomenon rather than as a phonological phenomenon. Their evidence is that /p/ seems to depress the acceptability of forms with more than one "marked" segment just as a voiced obstruent does, but their experimental data is not from forms with/without *rendaku* but from forms with/without an initial voiced obstruent (i.e., the idea that /gaQta/ is better than /gaQda/ but also better than /gaQpa/). In other words, (supposedly) /p/ patterns like a voiced obstruent in such cases.

Our study provides a straightforward test of the effect of /p/ and voiced obstruents in inhibiting *rendaku*. If the *rendaku* rate is lower when the test item's E2 contains a medial voiced obstruent than when it contains a medial voiceless obstruent other than /p/, the results corroborate the usual version of Lyman's Law. If test items with E2s that contain a medial /p/ pattern like test items with E2s that contain a medial voiced obstruent, then the results support the orthographic interpretation of Lyman's Law: E2 cannot contain more than one diacritic (゛ or ゜) in *kana* representation. If the test items with E2s that contain medial /p/ pattern like test items with E2s that contain a medial voiceless obstruent other than /p/, then the results support the usual version of Lyman's Law: E2 cannot contain more than one voiced obstruent.

2. METHODOLOGY

In order to test whether Lyman's Law is an orthographic effect rather than a phonological effect, in particular the possible interaction of Lyman's Law and /p/, we conducted two perception experiments, one with aural stimuli (without presenting *kana* orthography) and the other with written stimuli (including *kana* orthography).

2.1. Stimuli

Table 1 shows examples of the 38 test words we prepared. These were all non-existing Sino-Japanese words, so their familiarity scores were controlled for.[1] We chose to use Sino-Japanese words, although they often do not undergo *rendaku*, to obviate possible interaction with lexical-stratum factors involving /p/ and *rendaku*. The test words were compounds with the first element (E1) *nise-* (偽"fake") (e.g., /nise+soNpei/). If E2 undergoes *rendaku*, the first segment of E2 becomes voiced (e.g., /nise+**z**oNpei/).

E2 medial	Word	Orthography
/p/	/soN**p**ei/	村兵（そんぺい）
voiced obstr.	/soN**g**uu/	村宮（そんぐう）
voiceless obstr.	/soN**k**uu/	村空（そんくう）

Table 1: Examples of test words

Participants were asked to choose whether the form with *rendaku* or the form without *rendaku* (e.g., /nise+**z**oNpei/ or /nise+**s**oNpei/) was more natural. Half of the participants in each experiment were presented with the form undergoing *rendaku* first, and the other half were presented with the form not undergoing *rendaku* first, thereby counterbalancing the presentation order.

2.2. Participants

Fifty-three native Japanese speakers participated in the audio experiment, and 46 native Japanese speakers participated in the written experiment. All participants were from the Tokyo area (Tokyo, Chiba, Kanagawa and Saitama prefectures). Participants received a small monetary compensation upon completion of the experiment.

The participants for the two experiments were recruited via the online crowdsourcing platform Crowdworks. In both the audio and the written experiments, the participants were first presented with a consent form and with instructions, and they completed two practice trials before proceeding to the main section of the experiment.

2.3. Audio and written experiments

The audio experiment was a binary forced-choice test coded with jsPsych and conducted via a web-based test implemented using Cognition.run. The participants were asked to wear either headphones or earphones. The participants were visually presented with the non-existing test word in *kanji* and its "meaning"

in Japanese on the computer screen and auditorily presented with the forms undergoing and not undergoing *rendaku* (the choice between /nise+**z**oNpei/ and /nise+**s**oNpei/, etc.) and asked to choose the more natural form. The *kanji* orthography uses Chinese characters and thus, unlike *kana*, does not directly encode the voicing contrast; participants were given the voicing information aurally but not visually. The participants were able to listen to the audio stimuli repeatedly, as many times as they wanted.

The written experiment was a binary forced-choice test conducted via a web-based test implemented using PsyToolKit [12, 13]. The presentation for each test word consisted of two pages: a first page that explained the "meaning" of the non-existing word (in Japanese), and a second page asking the preference for the form undergoing *rendaku* or the form not undergoing *rendaku* (e.g., the choice between /nise+**z**oNpei/ and /nise+**s**oNpei/, presented in Japanese orthography). The participants were given the information about voicing and about /p/ visually, i.e., with and without the diacritics ⟨゛⟩ and ⟨゜⟩, as in にせきんぼん /nise+**k**iN**b**oN/ vs. にせぎんぼん /nise+**g**iN**b**oN/ and にせとんぴん /nise+**t**oN**p**iN/ vs. にせどんぴん /nise+**d**oN**p**iN/.

3. RESULTS

Figure 1 gives the mean rates of the preference for *rendaku* in the two experiments. The numbers indicate the percentage of participants in each experiment that chose the forms with *rendaku* as more natural.[2] The y-axis is the response rate, and the x-axis is the category of E2-medial consonants. Overall, the rates are not very high even for test words with voiceless E2-medial obstruents other than /p/ (0.21 in the audio experiment and 0.26 in the written experiment). This is the condition most likely to elicit *rendaku*, but the low rates can be explained by the general fact that Sino-Japanese E2s often do not undergo *rendaku*.

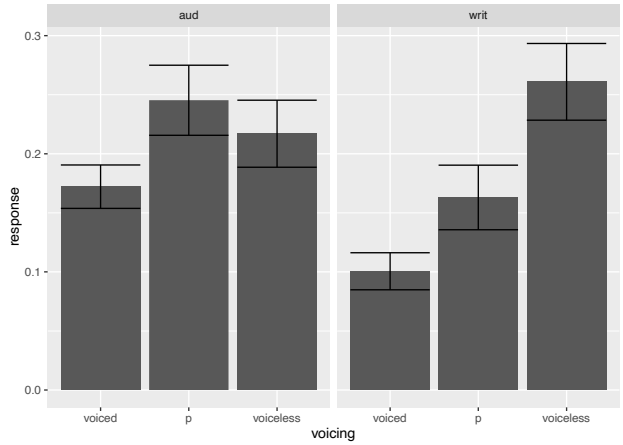


Figure 1: Mean *rendaku* preference rates; Left panel: audio data, Right panel: written data

To examine the effects of the E2 consonants, a mixed-effects logistic regression analysis was carried out with *rendaku* preference response as the dependent variable. The predictors included the voicing (three levels: /p/, voiceless consonants other than /p/, voiced consonants), the mode (two levels: audio and written) and the interaction of these two. Speaker and item were included in the model as random effects (a random intercept and a random slope for speaker and a random intercept for item). The analysis used the software R [11] with R package *lme4* [1].

Type II Wald chi-square tests, with the *Anova* function of the *car* package [2], determined that the main effect of voicing is overall statistically significant (Table 2).

	χ^2	*df*	*p*-value
voicing	11.9457	2	< 0.01 **
mode	1.7454	1	0.186
voicing:mode	9.5157	2	< 0.01 **

Table 2: Wald chi-square test of predictors in the mixed-effects logistic regression analysis

The results of the mixed-effects logistic regression analysis (Table 3) reveal that in the audio data, E2 /p/ is significantly different from E2 voiced consonants but not different from other voiceless consonants.

	Estimate	S.E.	z-value	Pr(>|z|)
Intercept	-1.502	0.288	-5.208	< 0.001 ***
Voiced	-0.636	0.287	-2.215	< 0.05 *
Voiceless	-0.322	0.319	-1.009	0.312
Modewritten	-0.637	0.403	-1.579	0.114
Voiced:writ	-0.108	0.378	-0.286	0.774
Voiceless:writ	0.968	0.404	2.397	< 0.05 *

Table 3: Mixed-effects logistic regression analysis

Follow-up pairwise comparisons were made for the three levels of voicing within each mode. This was done using the *testInteractions* function of the *phia* package with no adjustment. The results reveal that (a) in the audio data, in addition to the above-reported significance (/p/ vs. other voiceless obstr., /p/ vs. voiced obstr.), voiceless obstruents other than /p/ and voiced obstruents are not significantly different (p=0.29) and (b) in the written data, all three pairwise comparisons were significant (/p/ vs. voiced: p=0.039, voiced vs. voiceless except /p/: p<0.001), although the difference between /p/ and other voiceless obstruents was only marginally significant (p=0.068).

To summarize, /p/ patterns differently in the audio and written data. In the audio data, /p/ patterns with other voiceless obstruents and not with voiced obstruents: the *rendaku* rate is higher in E2s with medial /p/ or another voiceless obstruent than in E2s with a medial voiced obstruent. In the written data, /p/ patterns neither with other voiceless obstruents nor with voiced obstruents: the *rendaku* rate increases in the

order voiced obstruents < /p/ < voiceless obstruents excluding /p/, positioning /p/ between voiced and voiceless excluding /p/.

4. DISCUSSION

The results of the audio experiment are consistent with the standard version of Lyman's Law, according to which an E2-medial voiced obstruent inhibits *rendaku*. E2-medial /p/ patterned with other voiceless obstruents and not with voiced obstruents.

The results of the written experiment are less clear-cut. An E2-medial /p/ inhibited *rendaku* more than an E2-medial voiceless obstruent but less than an E2-medial voiced obstruent. If /p/ had patterned with the other voiceless obstruents, the results would be consistent with the standard version of Lyman's Law. If /p/ had patterned with the voiced obstruents, the results would be consistent with the orthographic version of Lyman's Law, according to which a medial diacritic (either *dakuten* or *han-dakuten*) in the *kana* representation of E2 inhibits *rendaku*.

The statistically significant interaction (see Table 2) between mode of presentation (audio vs. written) and voicing (/p/ vs. other voiceless obstruents vs. voiced obstruents in E2-medial position) indicates that mode of presentation affects participant behavior. The "intermediate" results of the written experiment are consistent with the idea that more than one diacritic in the *kana* representation of an E2 is disfavored. On the other hand, the statistically significant difference between /p/ and voiced obstruents contradicts the idea that any combination of two diacritics has the same effect. Recall that the diacritics used for moras beginning with a /p/ ⟨ ゜ ⟩ and for those beginning with a voiced obstruent ⟨ ゛ ⟩ differ in Japanese orthography (§1). Thus, one possible explanation for the difference between /p/ and voiced obstruents in *rendaku* inhibition is that although the orthography (in particular diacritic usage) plays a role, its effect is stronger when there are two identical diacritics (i.e., two ⟨ ゛ ⟩) than when there are two different diacritics (i.e., one ⟨ ゛ ⟩ and one ⟨ ゜ ⟩). This difference could be understood as a type of OCP effect.

The involvement of orthography in phonological representations has been argued for [4] (§1), and the results of this study support the idea. However, how do we reconcile the difference between the results with aural and written stimuli? The aural results for /p/ are arguably unexpected, since in Japanese morphophonology /p/ often does not pattern with the other phonemes realized as voiceless obstruents. For example, the *rendaku* alternation parallel to /t/~/d/, /k/~/g/, etc., is /h, f/~/b/, not /p/~/b/ (§1); /p/ alternates with /h/ and /b/ only in certain environments (e.g., immediately following moraic /N/). It is known that there are some phonological patterns that are not found elsewhere in the phonology of a language that may be explained by perception or perceptual knowledge [10], e.g., in the rhyming patterns in Japanese rap music [3]. Our results with the aural stimuli may simply reflect the perceptual similarity between /p/ and the other voiceless obstruents in Japanese. These phonemes are all realized phonetically as voiceless consonants, and this perceptual knowledge may have played a role in the participants' responses in the online processing. It would be interesting to test the hypothesis with a production experiment. If this perception account is on the right track, the results would be different there. Another possibility for further investigation would be a third presentation mode, providing both audio stimuli and *kana* representation and thus differing in just one respect from each of the other two modes: audio and *kana* vs. audio only vs. *kana* only.[3]

5. CONCLUSION

This paper has reported the results of two experiments designed to test the suggestion that Lyman's Law can be construed as an orthographic constraint rather than as a phonological constraint.

The phonological construal (standard Lyman's Law) is that an E2-medial voiced obstruent prevents *rendaku*. Since *rendaku* produces an E2-initial voiced obstruent, the phonological construal blocks the addition of a second voiced obstruent to an E2 that already contains one. The orthographic construal of Lyman's Law is that an E2-medial diacritic in *kana* orthography prevents *rendaku*. Since *rendaku* yields a *dakuten* (voicing) diacritic on the initial symbol of the *kana* representation of an E2, the orthographic construal blocks the addition of a second diacritic to the *kana* representation of an E2 that already contains a diacritic (either *dakuten* or *han-dakuten*).

The experimental results show that native speakers of Japanese respond differently depending on whether the test items are presented aurally or visually (i.e., in *kana*). The results of the audio experiment indicate that there is no diacritic effect when response choices are presented aurally. The results of the written experiment indicate that a second diacritic does inhibit *rendaku* when the response choices are presented visually but that the effect is stronger when E2 contains two *dakuten* than when it contains one *dakuten* and one *han-dakuten*, possibly showing an OCP effect involving orthographic representation. The different patterning of /p/ in the aural and written results, in particular the patterning of /p/ in the aural data, might be explained in terms of the perception of surface phonetic forms: [p] is just like other voiceless obstruents.

6. REFERENCES

[1] Bates, D., Mächler M., Bolker B., Walker S. 2015. Fitting linear mixed-effects models usinglme4. *J Stat Softw* 67(1), 1–48.

[2] Fox, J., Weisberg, S. 2019. An R Companion to Applied Regression, 3rd edition. Sage.

[3] Kawahara, S. 2007. Half rhymes in Japanese rap lyrics and knowledge of similarity. *Journal of East Asian Linguistics* 16, 113–144.

[4] Kawahara, S. 2018. Phonology and orthography: The orthographic characterization of rendaku and Lyman's Law. *Glossa* 3(1), 10, 1–24.

[5] Kawahara, S., Sano, S. 2016. /p/-driven geminate devoicing in Japanese: Corpus and experimental evidence. *Journal of Japanese Linguistics* 32, 57–77.

[6] Kindaichi, H. 1976. Rendaku-no kai (Explaining rendaku). *Sophia Linguistica* 2, 1–22.

[7] Lyman, B. S. 1894. The change from surd to sonant in Japanese compounds. *Oriental Studies of the Oriental Club of Philadelphia*, 160–176.

[8] Martin, S. E. 1952. *Morphophonemics of Standard Colloquial Japanese.* (*Language* Dissertation No. 47) Linguistic Society of America.

[9] Martin, S. E. 1987. *The Japanese Language through Time*. Yale University Press.

[10] Martin, A., Peperkamp, S. 2011. Speech perception in phonology. In: van Oostendorp, M., Ewen, C. J., Hume, E., Rice, K. (eds), *The Blackwell Companion to Phonology*. Wiley-Blackwell, 2334–2356.

[11] R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

[12] Stoet, G. 2010. PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods* 42(4), 1096–1104.

[13] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44(1), 24–31.

[14] Suzuki, Y. 2005. Raiman-no hoosoku-no reigai-ni tsuite (On exceptions of Lyman's Law). *Journal of Bunkyo Gakuin University, Department of Foreign Languages and Bunkyo Gakuin College* 4, 249–265.

[15] Vance, T. J. 2008. *The Sounds of Japanese.* Cambridge University Press.

[16] Vance, T. J. 2014. What students of Japanese can learn from Akkadian cuneiform. *Japanese Language and Literature* 48, 341–379.

[17] Vance, T. J. 2016. Introduction. In: Vance, T. J., Irwin, M. (eds), *Sequential Voicing in Japanese: Papers from the NINJAL Rendaku Project*. John Benjamins, 1–12.

[18] Vance, T. J. 2022. *Irregular Phonological Marking of Japanese Compounds: Benjamin Smith Lyman's Pioneering Research on Rendaku* (Mouton-NINJAL Library of Linguistics, 4) Walter de Gruyter.

* We thank an anonymous reviewer for their comments. We also appreciate input from Ami Eda and advice on statistics from Yoonjung Kang. This work was supported by JSPS KAKENHI 19K00613 (Manami Hirayama) and 20K00754 (Hinako Masuda).

[1] The list of words contained three types: (1) those with an E2-medial obstruent preceded by a moraic nasal /N/ (20 words), (2) those with an E2-medial obstruent preceded by a moraic obstruent /Q/ (12 words), and (3) those with E2-medial /h/ (6 words). This paper discusses only cases with moraic /N/.

[2] Words with initial /č/ and medial /š/ and /ǰ/ and those with initial /h/ and medial /p/ and /b/ were excluded from the analysis because they showed unexpectedly lower rates, possibly due to OCP effects.

[3] We thank the anonymous reviewer for this suggestion.

THE WINDOW OF OPPORTUNITY: ANTICIPATORY NASAL COARTICULATION IN THREE LANGUAGES

Marianne Pouplier[1], Francesco Rodriquez[1], Roy Alderton[1], Justin J.H. Lo[2], Eva Reinisch[3], Bronwen Evans[2], Christopher Carignan[2]

[1]Institute of Phonetics and Speech Processing (IPS), LMU Munich; [2]Dept. of Speech, Hearing and Phonetic Sciences, University College London; [3]Acoustics Research Institute Austrian Academy of Sciences, Vienna
{pouplier|f.rodriquez}@phonetik.uni-muenchen.de, roy.alderton@city.ac.uk,
{justin.lo|bronwen.evans|c.carignan}@ucl.ac.uk, eva.reinisch@oeaw.ac.at

ABSTRACT

This paper compares anticipatory nasal coarticulation in American English, French, and German. These languages differ in whether nasality is contrastive (French), phonologized but not contrastive (American English), or neither (German). We measure nasal intensity during a specific temporal interval preceding a nasal or oral control consonant. In English, coarticulation has the greatest temporal extent whereas in French, anticipatory nasalization is more constrained. German differs from English, but not French. While results confirm some of the expected language-specific effects, they underscore that the temporal extent of anticipatory nasal coarticulation can be greater than often reported if this is allowed for by the stimulus material. For all languages, the onset of coarticulation may considerably precede the prenasal vowel in VN sequences, especially so for English. Overall, our data further add to our understanding of the non-local temporal scope of anticipatory coarticulation and its language specific expressions.

Keywords: nasality, anticipatory coarticulation, phonological contrast, long-distance coarticulation

1. INTRODUCTION

The goal of the current experiment is to contribute to our understanding of the temporal extent of anticipatory nasal coarticulation and its cross-linguistic variation. While it is generally assumed that coarticulation arises from the influence of adjacent segments on each other, it is well-known that anticipatory coarticulation may go well beyond the adjacent segment and transgress prosodic boundaries (e.g. [12, 26, 29]). This non-local scope of anticipatory coarticulation has been especially evident in the context of V-to-V and labial coarticulation [20, 23]. In this paper, we expand this research by comparing the maximal extent of nasal coarticulation between languages.

Moll and Daniloff [19] were among the first to show that in English, nasal coarticulation may occur across multiple segments and cross word boundaries. More recently, Basset et al. [1] investigated nasal airflow in French spontaneous and read speech and reported cases in which the onset of nasal airflow starts before the segment preceding the nasal, especially so for $C_{voiced}\tilde{V}$. Yet there are comparatively few studies that explicitly investigate the maximum possible scope of nasal coarticulation, and how this may differ between languages. While current theories of speech production model coarticulation on the basis of overlap of adjacent segments, long-distance coarticulation is usually not part of these accounts ([11], [26]). Overall, our knowledge of long-distance effects is comparatively small, hampering a more comprehensive theoretical treatment of how coarticulation is planned.

Another important factor in understanding the control of coarticulation has been language-specific differences, and a number of studies have sought to link these to phonological contrast [17, 21]. Evidence for such a link has been equivocal [3, 18]: in the context of nasality, a contrast perspective predicts that contextual nasalisation of a vowel due to a following nasal consonant should be quite restricted in its temporal extent if a language has contrastive vowel nasality, but free to vary if no contrast is at stake. While the presence of contrast may indeed constrain coarticulation (e.g., French), or the absence of contrast may enable extensive coarticulation (e.g., English), for other languages, coarticulation has been found to be either unexpectedly limited in the absence of a contrast (e.g., Spanish, Italian, Japanese [10, 25, 28]), or unexpectedly extensive despite the presence of a contrast (e.g., Lakota [24]). Also dialectal variation within Spanish and French renders a more complex picture ([5], and discussion in [24]). Yet numerous differences in study design and methodologies hamper meta-comparisons across languages.

In sum, there is currently still relatively little predictive knowledge about the factors conditioning one language-specific pattern versus another. Few studies have examined the anticipatory scope of nasal coarticulation cross-linguistically using comparable material that allows for nasality to spread beyond the adjacent segment. The goal of our

current study is therefore to explore the maximal range of anticipatory nasal coarticulation for three languages: French, American English (henceforth English), and German. These languages differ in whether nasality is contrastive (French), phonologized but not contrastive (English, see [25]), or neither (German). We measure nasal intensity during a relatively large temporal interval preceding a nasal or oral control consonant. A difference metric is used to predict, based on changes to nasal intensity, an upcoming nasal vs. oral consonant.

The French vowel inventory comprises both nasal and oral vowels. In keeping with the literature, we expect French to show a limited scope of anticipatory nasalization (but see [8] on carryover coarticulation in French). For English, nasal coda consonants cause extensive nasalization on the preceding vowel and this nasalization co-varies with the duration of the preceding vowel. This has been taken to mean that English speakers target a pre-nasal tautosyllabic vowel as nasal, not oral [25, 30]. Recent publications on the physiology of velum behavior in German [6, 14] suggest a moderate degree of contextual vowel nasalization, despite the absence of contrast. How exactly German compares to French or English has to our knowledge not been investigated directly. For German, nasality is not known to be coupled to any phonological factors and is therefore in principle free to vary.

2. METHODS

2.1. Participants

Our dataset currently includes 77 participants (30 German, 30 English, 17 French). More speakers for French are being recorded (target: 30). All speakers were recorded in Germany or the UK and self-identified as native, dominant-language, standard speakers of the respective language.

2.2. Stimuli and recording

For each language, all stimuli were real words containing VN sequences in non-initial position, where N stands for any of the nasal consonants of a given language. Control oral minimal pairs with VC where C=/p, t, k/ were also recorded (Table 1). Any segments preceding the nasal consonant did not control velum position (i.e., consonants were /r, l, j/), allowing for nasality to spread across several segments. There were 10 nasal-oral minimal pairs per language (8 for German). Each word was recorded three times in randomized blocks, giving a targeted total of 10 items x 2 conditions (nasal, oral) x 3 repetitions = 60 per speaker. Data loss occurred due to speech errors and technical problems. All tokens with phrasal breaks immediately preceding the target word were removed from analysis. The token total for the current analyses is 1414 for English, 979 for French, and 1390 for German. The data were recorded as part of a larger dataset investigating coarticulation across different articulators.

English	French	German
rhymer, riper [raɪmɚ, raɪpɚ]	*l'anis, lapis* [lanis, lapis]	*Leine, leite* [laɪnə, laɪtə]

Table 1: Examples of nasal-oral word pairs in each language.

Target words were embedded in a carrier phrase constructed to be similar across the three languages (Engl.: *He'll tell Cleo X soon*. French: *Je dis à Cléo X samedi*. Ger.: *Er las Kleo X zweimal vor.*).

Speakers were recorded using a nasalance device which captures oral and nasal intensity based on two microphones separated by an acoustic baffle.

2.3. Data processing

All recordings were segmented automatically [13], with manual correction of all relevant boundaries. A region of interest (ROI) was defined over the interval from the end of the liquid in *Kleo/Cléo/Cleo* up to the onset of the nasal/oral target consonant. Recordings were filtered [4] with a passband of 80-10,000 Hz. Nasal intensity was mean-corrected for overall (nasal+oral) intensity on a token-by-token-basis and extracted for the ROI.

The onset of coarticulation was defined as the *divergence point* between oral and nasal condition curves and calculated as follows (cf. [16] for a similar procedure in a different context): For each nasal-oral minimal pair, the average nasal intensity curve was calculated across the repetitions of the oral tokens. A difference curve was obtained by subtracting each nasal token's nasal intensity curve from this oral condition average curve. For example, for the *Lohn-Lot* pair, the average nasality of all *Lot* tokens was subtracted from the nasality curve of each *Lohn* token on a by-speaker basis. Curves were trimmed to whichever curve was shorter.

To determine the onset of coarticulation, a sigmoid was fitted [27] to the time- and magnitude-normalized difference curves. The point at which a line tangential to the midpoint of the sigmoid rise intersects with the x-axis was defined as the *divergence point* (Fig. 1). Tokens were fit twice, once using a single and once an inverted double sigmoid (cf. [27]), with the one with the lower AIC value being chosen as the best fit. Normalized divergence points were then transformed back into

absolute time. A 5% RMS threshold was used per language to exclude tokens for which no good fit could be obtained, leaving a total of 1764 data points for further analysis.

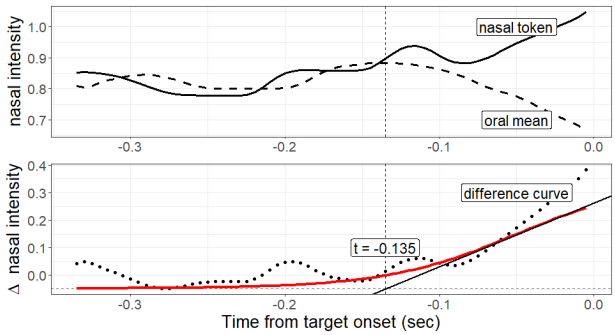


Figure 1, top: Nasal intensity curves of the mean of the oral condition tokens (dashed) and a single token from the nasal condition (solid). **Bottom**: Dotted line: difference curve obtained by subtracting the curves in the top graph from each other. Solid (red) line: fitted sigmoid. The point of divergence *t* is where the solid black line intersects the abscissa and defines the onset of coarticulation.

2.4. Statistical analysis

Mixed models were run in R [2, 22] with random intercepts for speaker and word pair. Repetition did not account for any variance and was excluded from statistical modelling. Significance was evaluated by model comparison; posthoc tests were conducted using [15]. Significance was assumed at $p<.05$. Divergence points and durations were log transformed for statistical analyses.

3. RESULTS

Fig. 2 gives the distribution of the divergence points by language. Recall that the divergence point quantifies the onset of coarticulation as the time point at which the nasal intensity curves of the oral-nasal minimal word pairs diverge. English has the smallest median, followed by German and then French. This confirms the expectation that French would show the latest onset of coarticulation and English the earliest. A statistical model with divergence point as dependent variable and fixed factor LANGUAGE is a significantly better fit over a model with random effects only ($\chi^2(2)=17.9$, $p<.001$). Posthoc pairwise comparisons indicate that English differs significantly from both French ($p<.001$) and German ($p=.028$), but the latter do not differ from each other ($p=.188$). Notably for English, there are instances of divergence points as early as 400ms before the onset of the nasal consonant.

We now ask whether durational differences in the ROI between languages may lie behind the seemingly language-specific effects, either due to carrier phrase or speech rate differences (if anticipatory nasality were to systematically vary with rate across languages).

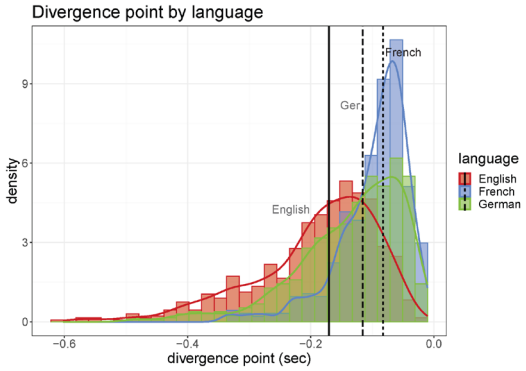


Figure 2. Divergence point distribution relative to consonant onset (zero) by language. Smaller values mean more extensive coarticulation. Vertical lines: median divergence point by language.

Table 2 gives the mean and standard deviation for the ROI duration per language. French has the shortest average ROI duration and hence by inference the fastest speech rate; English and German are very similar. A mixed model with LANGUAGE as fixed factor is a significantly better fit than a model with random factors only ($\chi^2(2)=10.9$, $p=.004$) with posthoc comparisons confirming a significant difference between German and French ($p=.01$) as well as English and French ($p=.01$).

To follow up on the role of duration, we compare our first statistical model on divergence points with LANGUAGE as a fixed factor with a model that also includes DURATION as a predictor. Model comparison is significant ($\chi^2(1)=36.6$, $p<.01$), yet importantly, there is no significant interaction between LANGUAGE and DURATION ($\chi^2(2)<1$).

	English	French	German
Mean	*0.45*	*0.38*	*0.46*
SD	*0.08*	*0.08*	*0.08*

Table 2. Mean and standard deviation (SD) of ROI duration in seconds.

Previous work on nasality argued for the crucial role of preceding vowel duration for anticipatory nasality in English [25]. It may thus be the case that speakers time the onset of coarticulation to the onset of the prenasal vowel. Fig. 3 presents the distribution of the divergence points calculated relative to the vowel onset; negative numbers indicate that the divergence points precede prenasal vowel onset. For French and German, the median is 20 and 30ms post vowel onset, respectively, while for English, the median precedes the prenasal vowel onset by -33ms. In VN sequences, the upcoming nasal consonant can

thus be predicted *before* the acoustic vowel onset for more than 50% of the data in English. Fig. 3 underscores that this also is the case for the other two languages for a sizable portion of the data. When comparing divergence points relative to vowel onset statistically, a model with fixed factor LANGUAGE is not a significantly better fit than a model with random factors only ($\chi^2(2)=3.9$, $p=.14$)

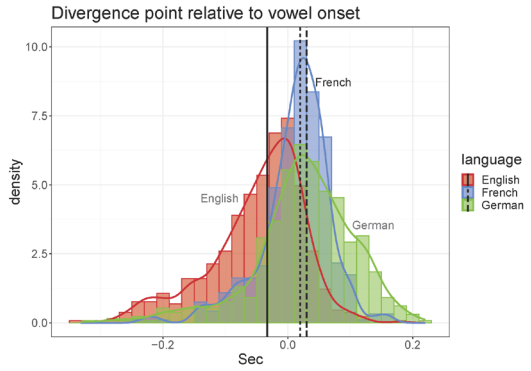


Figure 3. Divergence point distribution relative to vowel onset (zero) by language. Positive numbers: divergence point falls within the prenasal vowel, negative numbers mean it precedes the prenasal vowel. Vertical lines: median divergence point by language.

4. DISCUSSION

The goal of our paper was to investigate the maximal temporal extent of nasal coarticulation in three languages, given a relatively large window of opportunity. Our stimulus composition thus explicitly allowed for nasal coarticulation to spread beyond the prenasal vowel. We further asked whether the different phonological status of nasality in the chosen languages would play a role in terms of contrast constraining coarticulation. Based on a nasal-oral difference metric, we traced the time-point at which an upcoming nasal or oral consonant can be predicted based on nasal signal intensity.

Our results confirm the expectation that French is most limited in its coarticulatory scope, whereas English shows the greatest temporal extent of anticipatory nasalization. German, however, aligns with French, even though nasality is not contrastive in the former. For English, the divergence point precedes the onset of the prenasal vowel in more than 50% of the data. While language systemic factors may play a role in determining certain coarticulatory patterns, our results resonate with other studies which have cast doubt on phonological contrast being a good predictor of the temporal extent of coarticulation (among others, [3, 24]).

We were further able to show that nasal coarticulation may be quite extensive in scope, ranging up to around 300ms before the onset of the nasal consonant (Fig. 2). For all languages it may extend considerably before the prenasal vowel onset (Fig. 3). This aligns nasal coarticulation with studies on other articulators which have reported that anticipatory coarticulation can spread beyond the preceding vowel [12, 23], and challenges accounts of coarticulation as local co-production. Relatedly, Tilsen [26] recently has argued that anticipatory information is traceable in the signal much earlier than commonly assumed, a point raised in a different context also by [16]. Future work will have to examine to what extent predicted points of divergence as determined here can be exploited by listeners. Moreover, while the mere temporal extent of coarticulation may not vary as a function of phonological contrast, the overall signal dynamics may very well do so ([7], [25]); this is another point we aim to follow up in future analyses.

Solé [25] famously proposed that in English, velum opening for a coda consonant is phonologically timed to the preceding vowel, whereas in Spanish, velum opening is timed to the nasal consonant itself. Dow [9] recently presented a similar argument for French anticipatory nasal coarticulation being like Spanish, arguing that the timing of velar opening is tied to a short, physiologically determined time window preceding the nasal consonant. In our current work, however, we clearly see that the divergence point precedes the prenasal vowel for a considerable portion of the data for all three languages, most extremely so for English (Fig. 3). This latter observation for English may not be so surprising if one follows Solé in assuming that English speakers target an independently nasal vowel before nasal consonants. In that case, one could expect anticipatory nasality to spread from this nasal vowel. Yet the lack of a significant difference between languages in divergence points relative to vowel onset is unexpected from that perspective.

In sum, while our data confirm language-specific patterns of anticipatory nasal coarticulation, they also align with findings that phonological contrast per se is not a strong predictor of the temporal extent of coarticulation. We show that the onset of coarticulation, quantified as nasal-oral condition divergence, may be observed even earlier than the onset of the prenasal vowel in all languages, providing further evidence for coarticulation stretching over longer time windows than foreseen in models of coarticulation as local coproduction.

Acknowledgments. Work supported by DFG grant PO 1269/5-1 to M. Pouplier and C. Carignan. Katharina Neubert was of invaluable help during data collection and segmentation.

7. REFERENCES

[1] Basset, P., Amelot, A., Vaissière, J., Roubeau, B., 2001, Nasal airflow in French spontaneous speech. *J. Int. Phon. Assoc.,* 31, 1, 87-99.

[2] Bates, D.M., Maechler, M., Bolker, B., Walker, S., 2015, Fitting linear mixed-effects models using lme4. *J. Statistical Software,* 667, 1, 1-48.

[3] Beddor, P.S., Harnsberger, J.D., Lindemann, S., 2002, Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *J. Phon.,* 30, 591-627.

[4] Boersma, P., Weenink, D., 2022, Praat: doing phonetics by computer [Computer program]. http://www.praat.org.

[5] Bongiovanni, S., 2021, Acoustic investigation of anticipatory vowel nasalization in a Caribbean and a non-Caribbean dialect of Spanish. *Linguistics Vanguard,* 7, 1.

[6] Carignan, C. *et al.*, 2021, Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German. *Language*.

[7] Cohn, A., 1990, Phonetic and phonological rules of nasalization. *UCLA Working Papers in Phonetics,* 76.

[8] Delvaux, V., Demolin, D., Harmegnies, B., Soquet, A., 2008, The aerodynamics of nasalization in French. *J. Phon.,* 36, 4, 578-606.

[9] Dow, M., 2020, A phonetic-phonological study of vowel height and nasal coarticulation in French. *J. French Lang. Studies,* 30, 3, 239-274.

[10] Farnetani, E., 1986, A pilot study of the articulation of /n/ in Italian using electro-palatography and airflow measurements. *15e Journée d'Études sur la Parole,* 23-6.

[11] Farnetani, E., Recasens, D., 2010, Coarticulation and connected speech processes, in *The Handbook of Phonetic Sciences (2nd. ed.)*, W. J. Hardcastle, J. Laver, and F. E. Gibbon Eds.: Wiley-Blackwell, pp. 316-352.

[12] Grosvald, M., 2009, Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *J. Phon.,* 37, 2, 173-188.

[13] Kisler, T. *et al.*, 2016, BAS speech science web services. *Proc. of LREC 2016,* paper id 668.

[14] Kunay, E. *et al.*, 2022, Vowel height and velum position in German: Insights from a real-time magnetic resonance imaging study. *J. Acoust. Soc. Am.,* 152, 6, 3483-3501.

[15] Lenth, R.V., 2022, emmeans: Estimated Marginal Means, aka Least-Squares Means.

[16] Liu, Z., Xu, Y., Hsieh, F.-f., 2022, Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *J. Phon.,* 90, 101116.

[17] Manuel, S., 1990, The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America,* 88, 3, 1286-1298.

[18] Mok, P., 2012, Does vowel inventory density affect vowel-to-vowel coarticulation? *Language and Speech,* 56, 2, 191-209.

[19] Moll, K.L., Daniloff, R.G., 1971, Investigation of the Timing of Velar Movements during Speech. *J. Acoust. Soc. Am.,* 50, 2B, 678-684.

[20] Noiray, A., Cathiard, M.-A., Ménard, L., Abry, C., 2011, Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *J. Acoust. Soc. Am.,* 129, 1, 340-349.

[21] Öhman, S.E., 1966, Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America,* 39, 1, 151-168.

[22] *R: A Language and Environment for Statistical Computing*. (2021). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org

[23] Redford, M., Kallay, J., Bogdanov, S., Vatikiotis-Bateson, E., 2018, Leveraging audiovisual speech perception to measure anticipatory coarticulation. *J. Acoust. Soc. Am.,* 144, 4, 2447-2461.

[24] Scarborough, R., Zellou, G., Mirzayan, A., Rood, D.S., 2015, Phonetic and phonological patterns of nasality in Lakota vowels. *J. Int. Phon. Assoc.,* 45, 3, 289-309.

[25] Solé, M.-J., 1995, Spatio-temporal patterns of velopharyngeal action in phonetic and phonological nasalization. *Language and Speech,* 38, 1-23.

[26] Tilsen, S., 2020, Detecting anticipatory information in speech with signal chopping. *J. Phon.,* 82, 100996.

[27] Umut Caglar, M., Teufel, A.I., Wilke, C.O., 2018, Sicegar: R package for sigmoidal and double-sigmoidal curve fitting. *PeerJ,* 6, e4251.

[28] Ushijima, T., Hirose, H., 1974, Electromyographic study of the velum during speech. *J. Phon.,* 2, 315-26.

[29] West, P., 1999, Perception of distributed coarticulatory properties of English /l/ and /r. *J. Phon.,* 27, 4, 405-426.

[30] Zellou, G., Dahan, D., Embick, D., 2017, Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience,* 32, 6, 776-791.

MODELLING THE PARAMETERS OF FIXED STRESS IN SLAVIC WITH DYNAMIC COMPUTATIONAL NETWORKS

Toby Hudson

University of Oxford
toby.hudson@ling-phil.ox.ac.uk

ABSTRACT

This paper examines the acoustic correlates of word stress in Czech and Polish, Western Slavic languages with fixed accent position. The objective of the study is to model the data using a pair of variables, calculated by an optimisation algorithm. These variables represent lateral excitation or inhibition influencing prominence through the word in each direction.

A repository of recordings of stand-alone trisyllabic words in both languages is built using publicly-available internet resources. The duration, average intensity and average fundamental frequency for each syllable are measured. The optimisation algorithm is applied to the normalised measurements using a computational network that operates recursively until equilibrium and a pair of weights is obtained.

The network performs well where data are not diffuse, e.g. fitting the intensity data with a mean squared error of 0.04. The trendline for this parameter differs minimally between the two languages due to peak delay in the Czech data.

Keywords: Computational networks, modelling, word stress, Czech, Polish

1. INTRODUCTION

The possibility of modelling word stress patterns by using dynamic computational networks was introduced by Larson [1]. Here, the relative prominence of syllables is considered to be the result of lateral inhibition and excitation in the neighbouring environment, i.e. each syllable exerts a force on the others, akin to edge phenomena in the visual domain. This approach was heralded for its power and descriptive adequacy since it can accurately model phonological structures such as syllabic sonority and word stress with fewer variables than the number of units under scrutiny.

Larson [1] successfully uses this approach to model all attested general prominence patterns of word stress. Words with lexical/morphological stress or syllabic weight-to-stress are given additional weightings (M, H/L). The core settings are a weighting for initial syllable (I), optionally a weighting for the final syllable (F), a leftwards propagating 'wave' (α), and a rightwards propagating 'wave' (β). Word-prosodic shapes are generated by altering the relative values of these variables within the equation

$$(1) \quad d_i^{t+1} = u_i + \alpha^* d_{i+i}^t + \beta^* d_{i-1}^t$$

where d is the derived prominence of the syllable in whichever parameter we are investigating, subscript references its position within the word, superscript means time index, u indicates inherent sonority (weighting I above), α is the coefficient of leftward inhibition, and β is the coefficient of rightward inhibition [1: 21]. When the network converges we are left with values for α and β which characterise the shape of the trajectory.

The question remains whether or not such an iterative network will fit attested phonetic data with sufficiently low error. The present study is a preliminary attempt to apply equation (1) to data from a pair of languages which are understood to have static stress (i.e. no M or H/L): Czech (CZ) and Polish (PL). Since words in these languages are relatively short, the goal is to model the trajectory of potentially salient parameters with fewer variables than the number of syllables: here, this will be to characterise the durational, intensity and fundamental frequency (f0) contours of trisyllabic words in CZ and PL using only α and β. This is study is therefore both an exploration of the realisation of stress in these languages and a proof of concept for a modelling technique not commonly used for this purpose which may subsequently be used to model the prosody of longer words and more complex stress scenarios.

Since this study begins by obtaining acoustic readings, no assumption is made about whether and how stress is manifested in the languages in question beyond the selection of CZ and PL as likely candidates for manifesting left-edge and penultimate weighting. Duration, intensity and f0 are considered to be the prime acoustic candidates to represent prominence but this investigation could be expanded to include spectral prominence and finer-grained distinctions such as consonant lengthening.

Western Slavic is understood to have lost the earlier dynamic lexical accent – this was one of its common innovations sometime after the second century AD – and regularised a demarcative stress on the first syllable. Polish later regularised the stress on the penultimate syllable, and recent research has shown Polish to exhibit clear acoustic signalling of word stress in duration and f0 [2], and intensity [3]. However, present-day Czech does not systematically manifest clear acoustic patterns associated with stress. Recent studies (e.g. [4]) have demonstrated a pitch peak delayed to the post-stressed syllable, and [5] reports a shorter duration for the canonically stressed syllable than the second syllable, suggesting that differences between the prosody of CZ and PL may be subtle. Since stress is non-contrastive here, the acoustics of word-level prominence have become weak ([5]) though, as [3] demonstrates, this need not be the case *de facto* for non-contrastive stress.

2. METHOD

A dataset was built by obtaining publicly available audio files from the internet (primarily from pronunciation sites *forvo.com* and *dict.cc*). For this study only trisyllabic words were searched for; a trial-and-error process consisted of looking up suitable candidates to see which would bring up clear, noise-free audio recordings. Almost all recordings consisted of single words; a few were short phrases.

Czech	**Polish**
nadšený	nieufny
kajuta	nadymać
kdekoli	mojemu
nabidnout	kuśtykać
jizdenka	księgosusz
hniloba	kryjówka
smažený	kozodój
obézní	koniuszyc
dobrodruh	klarowny
hoření	kablówka
odjinud	jedynie
koleno	jajowód
kravata	instrument
obrovský	gliwicki
domácí	geneza
německý	esencja
vidlička	dopłacić
žvýkačka	cesarski
pracovník	bajeczny
pomeranč	abstrakcja

Table 1: List of words represented in the set of audio recordings

Only those deemed to be male voices were included here. The set of 20 Polish and 20 Czech tokens, covering a variety of phonotactic possibilities, is listed in Table 1.

Acoustic readings were obtained in *Praat* ([6]) with the use of a script to extract the timing of syllables as well as mean f0 and mean intensity values for each syllable. These were tabulated with very few missing values.

Duration measurements were normalised by expressing the value for each syllable of a word as a percentage proportion of the whole word. Intensity measurements were normalised by expressing the value for each syllable as a percentage proportion of the mean intensity for the word in question. F0 measurements were normalised by conversion of the deviation from the mean value of each word to semitones (multiplication by 12*log2). A very few obvious outliers were removed.

A Python 3 script was written which expresses equation (1) as a 3*3 matrix with only α and β as variables. Each normalised list of values (duration, intensity, f0) was read into the script as a comma-separated datafile. The script called the *scipy.optimize* package running the *minimize* routine with the *Nelder-Mead* simplex algorithm. This routine looked for a minimum optimal pair of values for α and β in a function that fitted each data set as accurately as possible.

The output for each of the six datasets is given as the optimal values for α and β, and mean squared error (MSE) along with filled contour plots to represent the degree of optimality of each pair of variables to the data at hand (the darkest coloured contour represents the solution space within which the network performs well). The dashed trajectory overlaid on the input trajectories shows in each case how neatly the function matches the input.

There arises the question of the value to which the invariant 'positional activation' I should be set. Larson [1: 94] shows that this must be greater than zero to produce canonical patterns of initial stress. The model was run with I set to values from 0.1 to 1.0 by increments of 0.1; the best fit from these (declared for each output below) was selected for each dataset.

3. RESULTS

Figures 1 and 2 represent the optimization for the CZ and PL *duration* measurements respectively. The data for durational measurements were not tightly clustered, such that no obvious outliers could be removed to improve the optimization which suffered from these rather messy sets of data.

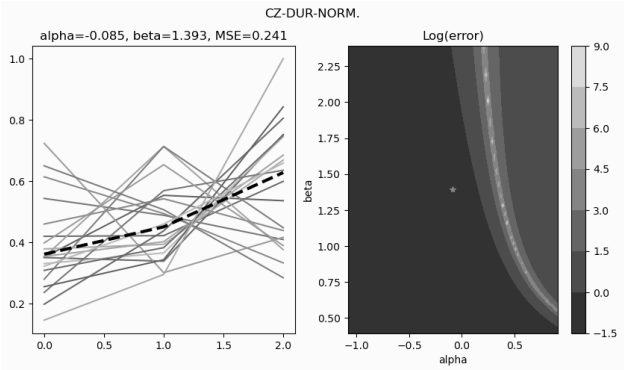


Figure 1: Optimization for syllable duration in Czech; *I*=0.4

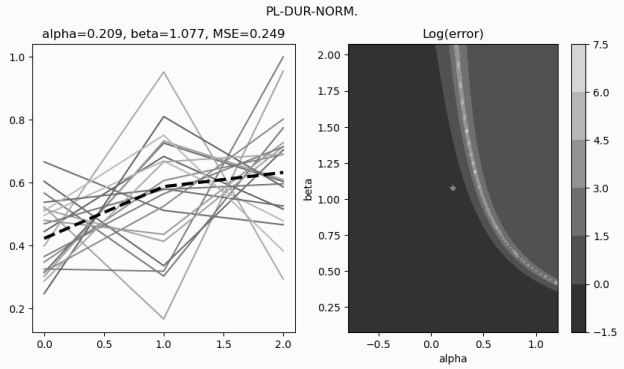


Figure 2: Optimization for syllable duration in Polish; *I*=0.3

A second syllable peak is more apparent for the *intensity* readings, and with an MSE of approximately 0.04 for both CZ and PL the model fits the data much better for this parameter (Figures 3 and 4).

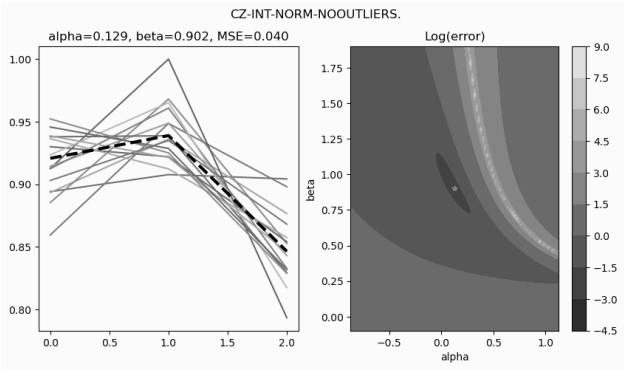


Figure 3: Optimization for syllable intensity in Czech; *I*=0.8

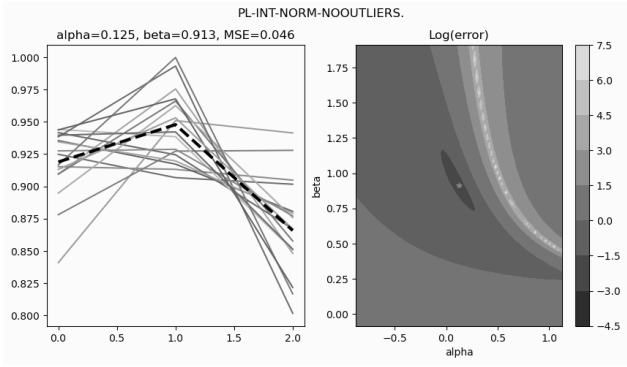


Figure 4: Optimization for syllable intensity in Polish; *I*=0.8

Optimization was relatively successful for *f0* with MSE at approximately 0.1 for both languages with a different setting for *I* for CZ and PL (Figures 5 and 6). The reported pitch delay may be in evidence for CZ but there is apparently a split in the data between trajectories with an f0 rise from syllable 1 to syllable 2 and those with a fall in f0 in this position. This same split is evident in the PL data. Optimization would be greatly improved without this split.

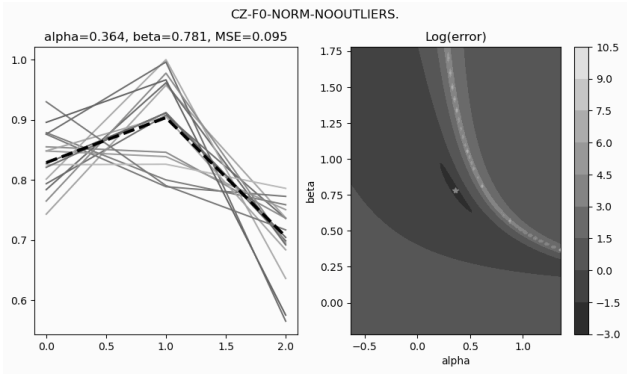


Figure 5: Optimization for syllable f0 in Czech; *I*=0.5

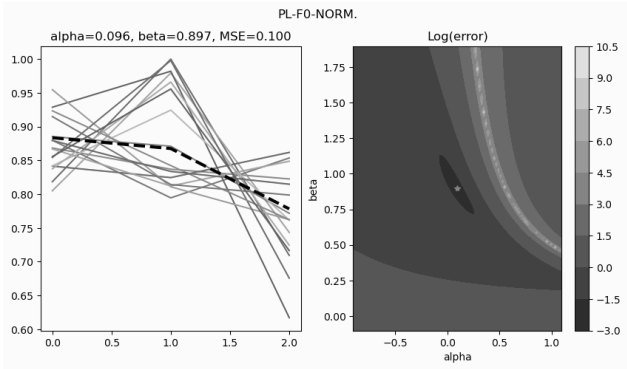


Figure 6: Optimization for syllable f0 in Polish; *I*=0.8

Fig. 7 plots the goodness of *I* at the intervals tested. If it were necessary to set this to a single value for all parameters then it would be possible to adopt a value around the centre of the range with the least mean error for all data sets being modelled (from left to right the red arrows indicate mean, median and mode activation weight).

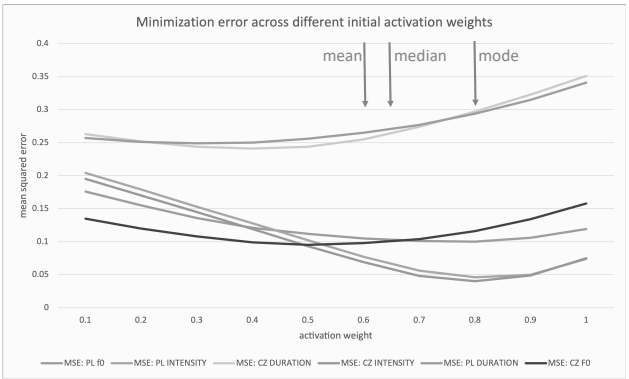


Figure 7: Minimization error (mean squared error) as a result of different activation weighting (*I*)

Larson [1:78] shows how different combinations of A and B generate the stress patterns of the world's language. Negative α in combination with negative β, for example, produces a rhythmic stress pattern, but positive α with positive β results in cumulative stress. When α is a higher positive value and β is a lower positive value, left edge cumulative stress is generated. With the exception of the more disparate data set for Czech duration (for which β had a negative value), all other data sets here do indeed conform to this pattern, as can be seen in Fig. 8 which plots the convergence values for the six optimizations.

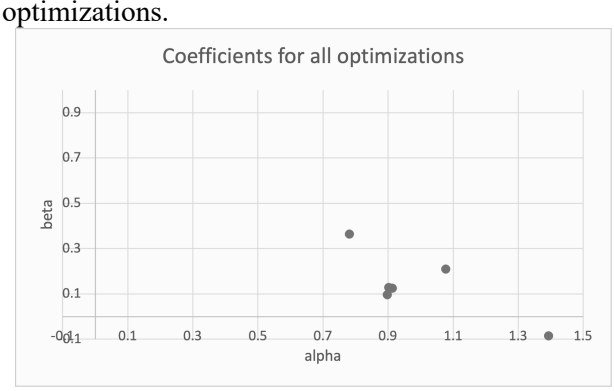


Figure 8: Convergence values for all six optimizations

4. DISCUSSION AND CONCLUSIONS

On the basis of these findings, average syllable *intensity* is most neatly modelled, and the data themselves are most consistent in this parameter. With optimization at the same initial activation value and a similar goodness of fit for CZ and PL it may be concluded that CZ and PL are modelled equally well for intensity, and that the model is successful in these two data sets. Additionally it would seem that there is little to distinguish the two languages in this parameter, i.e. they both exhibit a second syllable peak – the canonical position for PL, but representing peak delay in CZ.

With an MSE greater than 0.2 for both outputs we cannot deem the model to be powerful for *duration*. It is possible that an improvement would be seen with longer words if the data are split by those examples with a word-final lengthening effect and those without. Visual inspection of the trajectories does not lend strong confirmation to the notion that stress is cued durationally in either language, though the trajectories of several tokens have a durational peak on the second syllable which suggests an inconsistent durational prominence in the expected part of the word for Polish and in delayed position for Czech.

Undeniably the words with long vowels (é, á, í, ý) introduce a durational confound here as might be expected in a language with phonemic vowel length and quantity-insensitive stress. It is also possible that the variability of onset and rime consonant clusters has confounded patterns which would be apparent in a much larger investigation or where only words of matching syllabic complexity are compared.

Similarly to the intensity readings, the *f0* data are split into trajectories which have an f0 peak on the second syllable and others with an f0 trough in the same position. It is possible that intonation has been a confound here, for example if recordings were made with a 'listing' intonation. Alternatively it may be that f0 plays an inconsistent role in cueing stress in both languages.

The relative strength of intensity for CZ and PL here as cue to word stress tallies with the perspective that word stress is allied to *loudness*, a percept which relates to vocal effort in addition to e.g. jaw aperture and vowel timbre ([7]). The similarity in trend for intensity in CZ to that of PL is suggestive of (typologically rare) peninitial stress.

Where the data (once normalised and bereft of outliers) follow the same general trend the network has successfully modelled the trend with a fixed activation value and two variables. The practical implication is that the activation value in combination with α and β as cited for each output may be employed in synthesis to generate prosodic trajectories that accord with attested data. A theoretical implication of the above may be that the success of the model in representing intensity contours for Czech and Polish words that fit the data well vindicates its reality, i.e. it may indeed represent oscillatory ('wave') processes of excitation and inhibition in speech production and perception in which successive units in a string are quantifiably related to one another ([8]). However, the implications of these promising preliminary findings would be strengthened by extending the application of the model to longer words, which would also serve to test the applicability of the method to modelling secondary stresses.

5. REFERENCES

[1] Larson, G. 1992. Dynamic computational networks and the representation of phonological information. Doctoral thesis.

[2] Oliver, D., Grice, M. 2003. Phonetics and Phonology of lexical stress in Polish verbs. *Proc. 15th ICPhS* Barcelona, 2027–2030.

[3] Malisz, Z., Żygis, M. 2018. Lexical stress in Polish: evidence from focus and phrase-position differentiated production data. *9th International Conference on Speech Prosody* Poznań, Poland, 1008–1012.

[4] Palková, Z., Volín, J. 2003. The role of F0 contours in determining foot boundaries in Czech. *Proc. 15th ICPhS* Barcelona, 1783–1786.

[5] Skarnitzl, R, Eriksson, A. 2017. The acoustics of word stress in Czech as a function of speaking style. *Proc. Interspeech*, 3221–5.

[6] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.2.21, retrieved 1 October 2022 from http://www.praat.org/.

[7] Kochanski, G., Grabe, E., Coleman, J., Rosner, B. 2005. Loudness predicts prominence: fundamental frequency lends little. *J. Acoust. Soc. Am*. 118, 1038–1054.

[8] Goldsmith, J. 1994. A Dynamic Computational Theory of Accent Systems. In: Cole, J., Kisseberth, C. (eds), *Perspectives in Phonology*. Stanford: Center for the Study of Language and Information, 1–28.

DYNAMICS OF THE TONGUE CONTOUR IN THE PRODUCTION OF GUTTURAL CONSONANTS IN LEVANTINE ARABIC

Jalal Al-Tamimi[1] and Pertti Palo[2]

[1]Université Paris Cité, LLF, CNRS, France;[2]Indiana University, Speech Production Laboratory, USA
[1]jalal.al-tamimi@u-paris.fr; [2]pertti.palo@taurlin.org

ABSTRACT

This paper presents a dynamic account of the tongue contour imaged with Ultrasound Tongue Imaging (UTI) and quantified via Generalised Additive Mixed Models, during the articulation of guttural consonants (uvular, pharyngealised, pharyngeal) in Levantine Arabic. Gutturals are claimed to form a natural class and we aim to quantify the degree of (dis)similarity between the members of this class. UTI data were obtained from 8 participants (4 females), producing a variety of consonants (21) in the medial position of a disyllabic frame, with symmetric vowels /iː aː uː/. When compared to plain coronal consonants, and through a dynamic analysis of tongue contours throughout the VCV sequence, the three members of the guttural class show similarities in overall tongue changes towards the front, dorsum, back and root across the three vowel contexts, providing articulatory evidence for the legitimacy of gutturals as a natural class.

Keywords: Gutturals, Levantine Arabic, natural class, GAMMs, Ultrasound Tongue Imaging

1. INTRODUCTION

Guttural consonants are assumed to form a natural class due to specific phonological patterning, and/or to the use of a *common* oro-sensory zone in the pharynx [1, 2]. Its members were traditionally identified as pharyngeals /ħ ʕ/ (or epilaryngeal /ʜ ʢ ʡ/) and uvulars (/χ ʁ q/) [1]. Pharyngealised (/tˤ dˤ ðˤ sˤ/) [2], and/or Laryngeals (/ʔ h/) [3] are also considered to be part of this class, due to pharyngealised sharing a similar place of articulation to that of pharyngeals, albeit with a different degree of constriction [4, 5], or due to the increase in the frequency of the first formant in the vowels surrounding gutturals when compared with plain coronals [3]. However, phonetic research failed to find a common acoustic and/or articulatory correlate that unites all members of such class [6]. Indeed, gutturals in Arabic are variable in tongue configuration, degree of "retraction" and larynx rising/lowering [7, 8, 9, 4]. They favour regressive, rather than progressive feature spreading, which is strongest in pharyngealised consonants [10].

Following the "Laryngeal Articulator Model" (LAM) [11] "epilaryngeal" consonants are predicted to induce a maximal "retraction" of the tongue, with a back and down gesture and a raised larynx, "pharyngeal"/"pharyngealised" consonants to show partial "retraction" and "uvular" consonants to show nil "retraction". When considering the "pharyngeal"/"pharyngealised" pair, it is important to note that the former is assumed to have an upper-to-mid pharyngeal location (depending on the vowel, whereas the latter has a variable constriction location going from upper-to-low pharyngeal, depending on the language [11, 12]. Relying on the predictions from LAM, we propose that members of the guttural class (uvular, pharyngealised and pharyngeal) will share similarities in how they are realised, within consonants and across surrounding vowels. These similarities can be related to the constriction location, changes in specific sections of the tongue, e.g., tongue body, dorsum, root, or in the degree of tongue "retraction".

The aim of this study is then to assess whether there are any (dis)similarities between the members of the guttural natural class. We examine the full tongue contour (from tip to root) quantified via Ultrasound Tongue Imaging (UTI) and modelled via Generalised Additive Mixed-effects Models (GAMMs) on tongue shapes throughout the production of a VCV sequence. We quantify any changes not only at the tongue root, but also throughout the tongue contour, by using a 3D representation of the full tongue contour to guide our investigation.

2. METHOD

2.1. Data recordings

Ten Levantine Arabic Urban speakers (5 males, 5 females), aged 25-45 were recorded using synchronised UTI, ElectroGlottoGraphy (EGG), and audio recordings through a multichannel breakout [13]. The UTI data used a Mindray DP-6600, NTSC video output at 30fps, with a scan depth of 7.55cm, sampling Frequency of 5MHz, with

an endocavity microconvex probe (10mm radius; 120°Field Of View) with a metallic stabilisation headset (developed by Articulate Instruments). The EGG data used a 2-channel ElectroGlottoGraph, with larynx contact and height. Finally, the acoustic signal was recorded using a Roland Pro Microphone connected to a Roland Quad-Capture, sampled at 44.1 kHz, 16 Bit quantisation in mono channel; the microphone was placed at ≈ 15 cm from the speaker's mouth. Due to using the stabilisation headset, the EGG electrodes and the UTI probe, the angle of view across all participants was identical.

2.2. Material

Speakers were asked to produce a list of real and nonce-words in the following frame /ˈʔVːˈCVː/ (V ː = symmetric /iː aː uː/; C = all possible consonants in Levantine and other Arabic varieties = /b t d m n r f θ ð s z ʃ ʒ l w j k g x ɣ q tˤ dˤ ðˤ sˤ zˤ lˤ ħ ʕ ʔ h/), with three symmetric repetitions (with a maximum theoretical number of items of 2790 across all participants; 31 C * 3 V * 3 repetitions * 10 speakers). We then chose 21 consonants that were divided into 6 contexts, totalling 1940 items (21 C * 3 V * 3 repetitions * 10 speakers; with additional repetitions): **Plain** ⇒ /t d ð s z l/; **Velar** ⇒ /k g x ɣ/; **Uvular** ⇒ /q/; **Pharyngealised** ⇒ /tˤ dˤ ðˤ sˤ zˤ lˤ/; **Pharyngeal** ⇒ /ħ ʕ/; **Glottal** ⇒ /h ʔ/.

2.3. Data segmentation

Acoustic signals were first transliterated using the newly developed romanisation system for Arabic (ATR conventions) [14], then force-aligned using MAUS-universal language [15], with PraatAlign [16], adapted to Arabic. The boundaries obtained from the automatic forced alignement system were then manually corrected to prevent any errors, using criteria adapted from [5, 17]

2.4. UTI tracing

UTI data from 8 participants (4 males; 4 females) were analysed, using Articulate Assistant Advanced (AAA, version 2.18.04) [18]. First, the UTI video recordings were de-interlaced to 59.977 fps, to increase number of frames within each recording (frame length = 16 ms). The acoustic boundaries from the segmented speech were used in our subsequent analysis to guide landmarks selection. Because we wanted to dynamically track changes in tongue contours throughout the VCV sequence, we specified nine-time intervals (timeFrame) within a VCV sequence with a 25% interval shift, starting with the temporal location at 50% of the preceding vowel (V1) and ending at 50% of the following vowel (V2); the other seven intervals were equally spread across the remaining portion at 75% of V1; at 0%, 25%, 50%, 75%, and 100% of the medial consonant (C2), and at 25% of V2.

At each time interval, we traced the full tongue contour, in an unrotated view, from the visible portions on the UTI videos, using a 42 fanline shaped window, which was limited between the hyoid and the mandible bones. The tongue contours were first automatically traced using batch processing in AAA and were then all hand corrected. This resulted in a total of 13698 tongue splines. We extracted the 42 fanline coordinates in Polar coordinates with the Rho values representing tongue height in mm (r) and each fanline representing an angle value in Radians (φ). The first 4 and last 4 fanlines that were hidden by the hyoid and the mandible bones were excluded from subsequent analyses.

2.5. Statistical design

UTI data from 34 fanlines obtained throughout the VCV sequence from the nine intervals were considered as time series and were modeled via an Auto-Regressive Generalised Additive Mixed Model (AR-GAMM) using the package `mgcv` [19]. The total number of measurements submitted to our AR-GAMMs was 465732 datapoints (13698 splines * 34 fanlines). We used an ordered predictor for each of our predictors; Context (plain, velar, uvular, pharyngealised, pharyngeal, and glottal), Vowel (/iː aː uː/) and gender; order predictors reduce Type I error and increase power [20]. To account for inter-dependencies in the data (multiple speakers, items, contexts and two time series) and to allow for normalisation within and between speaker and gender, we used a maximal specification model, similar to [21], with the following specification:

- Outcome ⇒ Radius value in mm
- Fixed ⇒ Context by Vowel interaction by gender
- Smooths ⇒ Angle (34) and timeFrame (9) by the Context by Vowel interaction and by gender
- Tensor product interaction (ti) ⇒ between Angle and timeFrame by the Context by Vowel interaction and by gender
- Factor smooths interactions ⇒ Angle and timeFrame by speaker adjusted by Context by Vowel interaction
- Factor smooths interactions ⇒ Angle and timeFrame by word adjusted by gender

Our model improved the fit when compared to a simpler one ($\chi^2_{(4)}$ = 8101.9, p<0.0001); it accounted for 88.7% of the deviance explained in the data. We evaluated the structure of the model using the function `gam.check`. To quantify the overall tongue contour throughout the VCV sequence, we

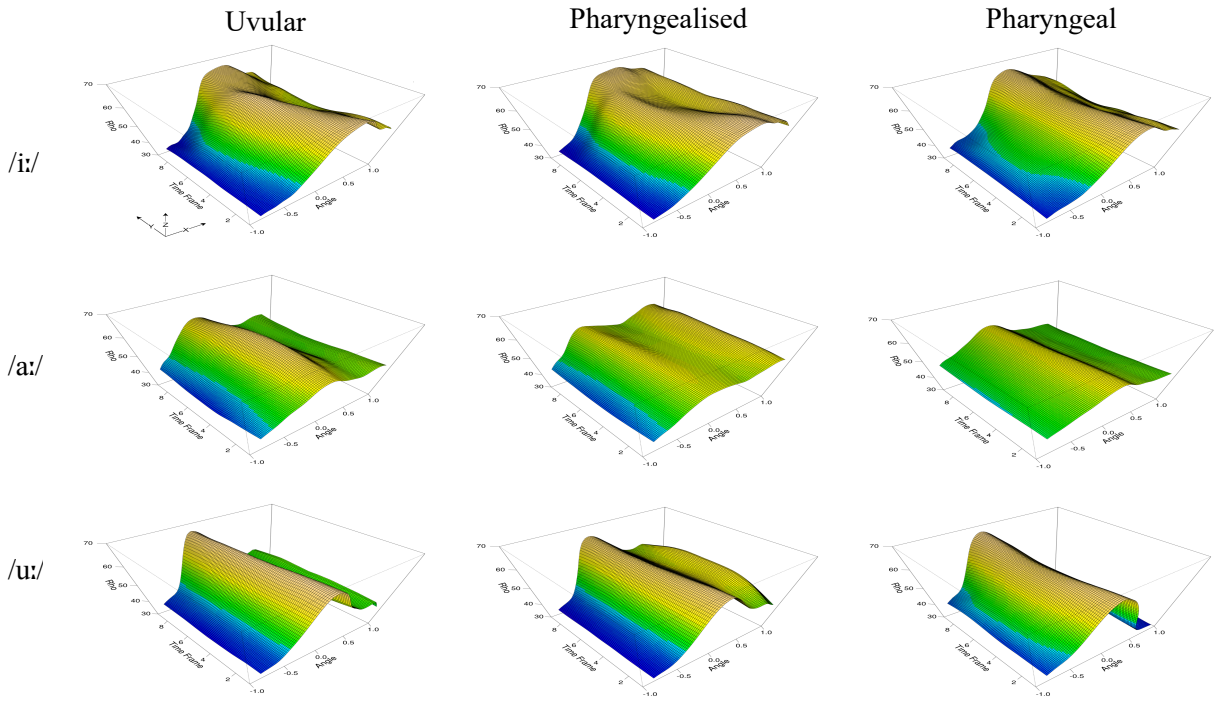


Figure 1: 3D surface contours for the three classes (Uvular Pharyngealised and Pharyngeal) in the vowels /i: a: u:/ (top, middle and bottom rows), according to the Angle (x-axis; oriented from bottom right to top right, tongue root to tip), timeFrame (y-axis; from bottom left to top left, V1 50% to V2 50%) and Rho (z-axis; bottom to top). Tongue height is indicated by both height within the image and colour continuum blue-green-orange.

used two types of plots as 3D visualisations, with the angle values on the x-axis, the timeFrame on the y-axis and the tongue height (Rho) on the z-axis. The first is a 3D surface plot for a specific context by vowel interaction using `vis.gam` from `mgcv`. The second is a differences plot between two tongue contours in a specific vowel environment (using `plot_diff2`) from `itsadug` [22]. We used predictions of our model for angle value ranging between -1 and 1, to estimate the constriction location in a similar fashion to that used in [23].

3. RESULTS

3.1. 3D surface plots

The results presented in Fig. 1 show the tongue surfaces for the interaction between specific contexts and vowels (uvular, pharyngealised and pharyngeal in columns; /i: a: u:/ in rows). One interesting observation across the three contexts is the similarities in tongue surfaces with respect to the specific portion of the tongue that is impacted upon. For instance, starting with the three contexts in an /i:/ context (top row), we can easily observe tongue front and dorsum depression (angle = 0 to +0.84) and retraction towards the tongue back/dorsum area (angle = 0 to -0.84) and root. Uvular and Pharyngealised show similarities in tongue surfaces, with small differences related to dorsum root retraction and tongue height. In the three classes, most of the similarities are within the consonant itself spreading into the V2. Within /a:/ (middle row), similar patterns are observed between the three contexts, with most of the changes observed towards the front, mid, back/dorsum and root of the tongue. Tongue body is lower than in /i:/, but shows minimal depression specifically in the pharyngealised context. Here again, most of the similarities are within the consonant spreading into V2. Finally, within /u:/ (bottom row), most of the changes are observed within the uvular and the pharyngeal context, with less differences within the pharyngealised context. The three contexts show similarities in type of tongue changes from front to back, with differences in tongue height. This first comparison showed how uvular, pharyngealised and pharyngeal contexts share similarities in how they impact on the tongue contour throughout the VCV sequence and within vowels. These changes are compatible with the “double bunching” for a “pharyngeal” place suggested by [11, 12], specifically for /i: u:/ contexts, with changes within /a:/ context located at the root of the tongue.

3.2. 3D difference plots

Given the similarities observed between the three contexts highlighted in Fig. 1, we wanted to formally

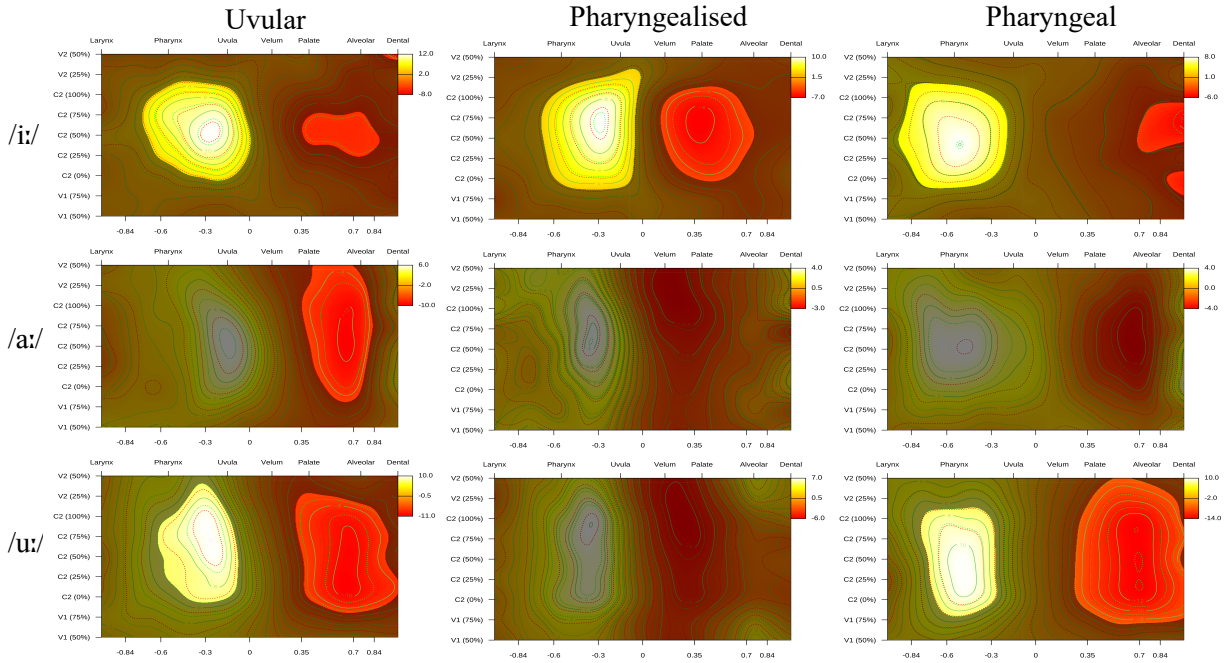


Figure 2: 3D difference plots, with statistically significant regions 95% CI highlighted. The plots show differences between each of the three classes (Uvular, Pharyngealised and Pharyngeal) and the plain class, in the vowels /iː aː uː/ (top, middle and bottom rows), according to the Angle (x-axis; tongue root to tip), timeFrame (y-axis; V1 50% to V2 50%), Rho (z-axis; tongue height difference indicated by lighter and darker colours, with lighter = increase in tongue height; darker = decrease), with estimated constriction location (secondary x-axis, top).

assess whether there are any differences between each of the contexts and the plain coronal context. Results presented in Fig. 2 show the 3D difference plots, which compares the overall tongue contour differences across two contexts. It is interesting to note that across the three vowels, the results are similar and show overall that most of the differences between the three contexts and the plain coronal context are located at the front and the back tongue that goes into the root, specifically within pharyngeals. Most of the observed differences are located throughout the C2 that goes into V2 (up to the middle). Within V1, the only differences observed are around and after 75% of the vowel. The results suggest overall that the constriction location of "gutturals" is similar and is located within an upper to low pharyngeal that shows a gradient constriction across the contexts. Uvulars show the highest location, followed by pharyngealised and then pharyngeals.

4. DISCUSSION

This paper attempted to identify whether there are any (dis)similarities between uvular, pharyngealised and pharyngeal contexts. The results presented offer an empirical evidence compatible with legitimacy of the guttural natural class. Gutturals in Arabic show similarities in overall tongue shape and impacts on the front and back cavities. The uvular context shows partial "retraction" with a "raised" and "back" configuration following LAM [11]. The pharyngealised context has an intermediate tongue "retraction", with a back and mid-down gesture in /iː aː/, and a back and mid-up gesture in /uː/. The uvular and pharyngealised contexts are different in their constriction location with differences mostly in the degree of tongue rising. The pharyngeal context shows a near maximal "retraction" with a lowered tongue dorsum and tongue root changes, but with a fronted tongue position. The tongue dynamics via 3D whole Tongue contours and differences in comparison with a plain context, show *articulatory* differences throughout the C2 and in V2 (up to mid), with minimal differences within V1. This suggests that coarticulatory patterns are progressive in gutturals. The results of the estimated constriction location on the secondary x-axis suggest tongue root changes and retraction that are possibly indicative of larynx raising. Our results suggest gutturals to show possible double bunching in an /iː uː/ contexts, with tongue fronting towards the palatal region and tongue retraction towards the pharynx [11, 12].

5. ACKNOWLEDGMENT

This work is supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083), contributing to the IdEx Université Paris Cité - ANR-18-IDEX-0001; A

British Academy/Leverhulme small research grant, UK (SG160181; 2017-2019); A Leverhulme International Academic Fellowship, UK (IAF-2018-016), for the first author; and a personal post-doctoral grant from Eemil Aaltonen's foundation via The Finnish Foundations' Post Doc Pool, for the second author. We acknowledge usage of the High Power Computing (HPC) facilities at Newcastle University, UK and at the CNRS/TGIR HUMA-NUM, IN2P3 and GENCI-IDRIS, France (Grant 2022-AD010613733).

6. REFERENCES

[1] J. McCarthy, "The phonetics and phonology of Semitic pharyngeals," in *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, P. Keating, Ed. Cambridge University Press, 1994, pp. 191–233.

[2] J. Sylak-Glassman, "An Emergent Approach to the Guttural Natural Class," *Proceedings of the Annual Meetings on Phonology*, pp. 1–12, 2014.

[3] B. Zawaydeh, "The phonetics and phonology of gutturals in Arabic," Ph.D. dissertation, Bloomington, IN: Indiana University, 1999.

[4] A. Laufer and T. Baer, "The emphatic and pharyngeal sounds in Hebrew and in Arabic," *Language and Speech*, vol. 31, pp. 181–205, 1988.

[5] J. Al-Tamimi, "Revisiting acoustic cues of pharyngealization in jordanian and moroccan arabic: Implications for feature specification," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.

[6] M. Bin-Muqbil, "Phonetic and phonological aspects of Arabic emphatics and gutturals," Ph.D. dissertation, University of Wisconsin-Madison, 2006.

[7] C. Zeroual and G. Clements, "The feature [pharyngeal]," in *Features in Phonology and Phonetics: posthumous Writings by George N. Clements and Coauthors*, A. Rialland, R. Ridouane, and H. van der Hulst, Eds. Mouton de Gruyter: Berlin, 2015, pp. 247–276.

[8] F. Al-Tamimi and B. Heselwood, "Nasoendoscopic, videofluoroscopic and acoustic study of plain and emphatic coronals in Jordanian Arabic," in *Instrumental studies in Arabic phonetics*, B. Heselwood and Z. Hassan, Eds. John Benjamins, 2011, pp. 165–191.

[9] B. Heselwood and F. Al-Tamimi, "A study of the laryngeal and pharyngeal consonants in Jordanian Arabic using nasoendoscopy, videofluoroscopy and spectrography," in *Instrumental studies in Arabic phonetics*, B. Heselwood and Z. Hassan, Eds. John Benjamins, 2011, pp. 101–128.

[10] S. Hellmuth, *The Oxford handbook of Arabic linguistics*. Oxford University Press, 2013, ch. Phonology, pp. 45–70.

[11] J. Esling, S. Moisik, A. Benner, and L. Crevier-Buchman, *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press, 2019.

[12] S. Moisik, E. Czaykowska-Higgins, and J. Esling, "Phonological potentials and the lower vocal tract," *Journal of the International Phonetic Association*, pp. 1–35, apr 2019.

[13] A. Wrench and J. Scobbie, "High-speed cineloop ultrasound vs. video ultrasound tongue imaging: Comparison of front and back lingual gesture location and relative timing." in *Proceedings of the Eighth International Seminar on Speech Production (ISSP)*, 2008, pp. 57–60.

[14] J. Al-Tamimi, F. Schiel, G. Khattab, N. Sokhey, D. Amazouz, A. Dallak, and H. Moussa, "A Romanization System and WebMAUS Aligner for Arabic Varieties," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), © European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*, Marseille, 20-25 June 2022, 2022, pp. 7269–7276.

[15] F. Schiel, "A statistical model for predicting pronunciation," in *Proc. of the International Conference on Phonetic Sciences, Glasgow, United Kingdom, Paper*, vol. 195, 2015.

[16] M. Lubbers and F. Torreira, "Praatalign: an interactive praat plug-in for performing phonetic forced alignment," https://github.com/dopefishh/praatalign, 2013-2016, version 2.0a.

[17] J. Al-Tamimi and G. Khattab, "Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops," *Journal of Phonetics*, vol. 71, pp. 306–325, nov 2018.

[18] A. Wrench, "Articulate assistant advanced user guide (version 2.17)," *Edinburgh: Articulate Instruments Ltd*, 2018.

[19] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[20] M. Sóskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, 2021.

[21] M. Heyne, D. Derrick, and J. Al-Tamimi, "Native language influence on brass instrument performance: An application of generalized additive mixed models (GAMMs) to midsagittal ultrasound images of the tongue," *Frontiers in Psychology*, vol. 10, nov 2019.

[22] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2017, r package version 2.3.

[23] C. Carignan, P. Hoole, E. Kunay, M. Pouplier, A. Joseph, D. Voit, J. Frahm, and J. Harrington, "Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, 2020.

Variable Word Onset Vowel Deletion in Najdi Arabic

Rana Alsabhan[1], Jane Setter[2]

[1]Princess Nourah Bint Abdulrahman University, [1,2]University of Reading
r.alsabhan@pgr.reading.ac.uk, j.e.setter@reading.ac.uk

ABSTRACT

The literature on Arabic dialects agrees that short vowels are deleted word-initially, resulting in surface consonant clusters. This study supplements the existing literature by acoustically analysing word onset vowel deletion in Najdi Arabic (NA), a dialect spoken in central Saudi Arabia. It aims to examine the phonetic deletion of the vowel within an obstruent context. Given the relatively lower frequency of obstruents cross-linguistically [1], the abundance of such sounds in the Arabic language presents a compelling context for investigating vowel deletion and its conditioning factors.

Forty-five NA native speakers (31 females, 14 males) were recruited to do three production tasks: diapix, reading and shadowing tasks. The results showed that 54% of the data were produced with vowel deletion, the favouring factor being the place of articulation of the first obstruent, where the deletion was more likely to occur following a dorsal obstruent than following coronal, labial and pharyngeal obstruents. The gender of the speaker also influenced vowel deletion, where male listeners were more likely to delete more vowels than female speakers.

Keywords: vowel deletion; Najdi Arabic; obstruents; phonetics; consonant clusters

1. INTRODUCTION

The process of variable short vowel deletion word-initially is attested in several Arabic dialects, such as Algerian Arabic, Jordanian Arabic and Omani Arabic [2, 3]. E.g., the word *kitab* 'book' can be realised with a deleted vowel as /ktab/, or with a retained vowel as /kitab/. The same word can be produced with variable phonetic realisations. Similar processes have been also observed in other languages, such as the variable deletion of the French schwa [4] and the American schwa [5]. This process and its possible causes have received attention in the linguistic literature.

The conditioning factors of short vowel deletion are mostly attributed to speech rate and style [5, 6]. In Arabic dialects, the constraints that account for this phenomenon are attributed to stress and syllable structure, where vowel deletion occurs in unstressed open syllables and does not result in a sequence of more than two consonants [7, 8]. Sociolinguistic research on Arabic attributed vowel deletion more to male speakers than female speakers given that females tend to adhere more closely to Standard Arabic since it has higher status [9].

The literature on phonological contexts that trigger vowel deletion cross-linguistically lists several factors. For example, in Japanese, vowels are deleted/devoiced between voiceless consonants [10], which suggests an effect of the laryngeal feature of the surrounding consonants. In Kozani Greek, back vowels tend to be deleted more frequently than front vowels [11], indicating an effect of the quality of the target vowel. In Montreal French, vowels are deleted when preceded by fricatives [12], implying that the manner of articulation of the surrounding consonants affect the deletion of the vowel.

The phonetic realisation of vowel deletion and the factors that might influence the deletion process have not received much attention in Arabic dialects. We aim to address this issue in our empirical study, which has two objectives: 1) to document the phonetic variability by means of an acoustic study examined over 21,000 occurrences of words; and 2) to examine the effect speaker's gender and phonological contexts on the deletion process.

2. METHODOLOGY

This study employed different speech styles to examine vowel deletion thoroughly. Three production tasks were conducted: a diapix task to elicit spontaneous speech; a reading task to elicit controlled speech based on visual (written) stimuli; and a shadowing task to elicit controlled speech based on audio stimuli. The reading task is a prerequisite for the shadowing task, because this allows participants to access the orthographic forms of stimuli first to help them perceptually recognise the words.

2.1. Participants

45 native NA speakers (31 females, 14 males) were recruited. Their ages ranged from 18 to 50 years (mean= 27, range = 25–34, *SD* = 7.8). They had all given informed consent for their data to be used and did not report any speech or hearing difficulty. A total of 41 participants completed the diapix task and 39 participants completed the shadowing task.

2.2 Recording Procedure

Participants' speech was recorded using a Tascam DR05X stereo handheld digital audio recorder with built-in, omni-directional condenser microphone. Due to the Covid-19 pandamic, an external microphone was not used in order to ensure the safety of the participants. The recording was made at a sampling rate of 44.1 KHz and 16-bits resolution and were saved in waveform audio file (WAV) formats. Data collection rook place in quiet locations within participants' homes and participants were provided with instructions on how to optimise the recording environment. Prior to the recording, participants recived training on how to operate the recorder.

2.3 Diapix Task

A diapix task is an interactive task that elicits spontaneous speech in a controlled setting [13]. It employs two participants, each with a version of the same picture that is different in some ways, and they are asked to spot the differences by conversing with each other without looking at each other's picture.

2.3.1. Stimuli

Six picture pairs were adapted from Baker and Hazan [14]. There were 12 differences between the two versions of each pair. The pictures were printed out in colour in A5 size.

2.3.2. Procedure

Participants were given a folder that included six numbered envelopes (1–6), envelope (1) being for practice. They were instructed to sit at a table facing each other and place the recorder in the middle between them (approximately 30 cm from each participant). Each participant was instructed to take one version and converse with the other participant to find the 12 differences.

2.4. Reading Task

A sentence-reading task was utilised to examine participants' productions at a systematic and controlled level.

2.4.1. Stimuli and Design

We extracted 66 disyllabic and trisyllabic high-frequency NA nouns, in which the first and second consonants were obstruents, from Arabic Web Corpus usinf Sketch Engine. The target words comprised 17 words with a fricative–fricative sequence (FF), 16 words with a fricative–plosive sequence (FP), 19 words with a plosive–fricative sequence (PF) and 14 words with a plosive–plosive sequence (PP). A list of 40 filler words was also included to divert participants' attention.

This task involved two types of stimuli, namely visual and audio, following the protocol used in Hall [15]. The visual stimuli consisted of orthographic forms of the target words, while the audio stimuli were audio recordings of example sentences for each word, produced by a NA native speaker. The target word in each example sentence was excised and replaced with white noise for a duration of 400 ms.

A self-paced PowerPoint presentation was created, where each slide displayed one target word presented within two carrier sentences, accompanied by the matching audio recording of an example sentence which played automatically 300 ms after viewing the slide. Each word was repeated twice in the presentation, which totalled 344 slides.

2.4.2. Procedure

Participants were emailed a link to the task and were instructed to place the recorder on the right side of their computer. They were then instructed to read the sentences after listening to the audio stimulus and press the right arrow on their keyboard to advance to the next slide. Participants were allowed to take as many breaks as needed.

2.5. Shadowing Task

The aim of this task was to elicit the pronunciation of target words with a specific variant, namely, the vowel-absence variant /#CC/.

2.5.1. Stimuli and Design

The target and filler words used in this task were identical to those used in the reading task, except for one PP word. A NA speaker was audio-recorded producing the words with the onset vowel deleted /#CC/. A self-paced PowerPoint presentation was created, where each slide had an audio recording of one word played automatically twice with a 400 ms interstimulus. Each word was repeated twice in the presentation, resulting in a total of 210 slides.

2.5.2. Procedure

Participants were instructed to wear headphones (Beats Solo3 wireless on-ear), place the recorder to the right side of their computer, and sit about 30 cm from the computer. They were directed to listen to the audio stimuli and repeat the word after the second repetition as quickly and clearly as possible. Then, they were asked to press the right arrow on

their keyboard to advance to the next slide.

2.6. Acoustic Analysis

Recordings were manually segmented and annotated in Praat [16]. One annotation tier was created to mark word and segment boundaries, following the segmentation guidelines created by Skarnitzl and Machač [17]. Another tier was created to mark the presence or absence of vowels in experimental items.

Vowel presence was identified by visual inspection of the spectrogram and waveform, and was coded as 'CVC' when high amplitude peaks of energy were visible in spectrograms as horizontal dark bands. For example, the onset of the vowel following a fricative was marked at the point where periodicity began. The offset of the vowel preceeding a fricative was marked at the point of onset of turbulent noise, as shown in Figure 1.

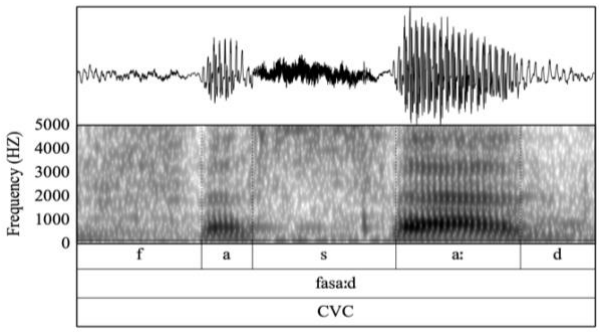


Figure 1: An example of word onset vowel presence for the FF sequence /fasa:d/.

Acoustic evidence for word onset vowel absence was marked when there was a complete absence of voicing and formant structure, and was coded as 'CC'. For example, in an FF sequence, both waveforms and spectrograms showed two distinct aperiodic frequencies for each fricative, with no evidence of formant structure or voicing, as illustrated in Figure 2.

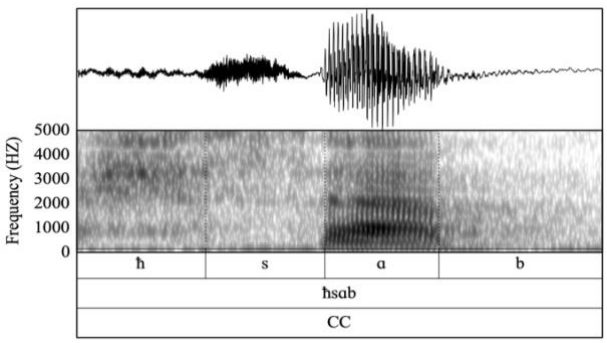


Figure 2: An example of word onset vowel absence for the FF sequence /ħsɑb/.

2.7. Statistical Analysis

The data were analysed by means of mixed-effects logistic regression model in R [18] with the the glmer() function using 'glmmTMB' package [19]. The dependent variable was was dichotomous (0 = vowel was deleted, 1= vowel was retained, 0 was the success level in the analysis). The fixed part of the model consisted of the following independent variables:

- Gender: male vs female speakers. Since past literature shows that gender differences affect deletion rates, this was the only sociolinguistic variable addressed in this study.
- The manner of articulation of the flanking obstruents with four levels: FF, FP, PF, PP.
- The place of articulation of the first obstruent, with four levels: labial–initial, coronal–initial, dorsal–initial and pharyngeal–initial.
- The laryngeal feature of the flanking obstruents (matching laryngeal: both voiced /zb/ or both voiceless /kt/ or different laryngeal features e.g. voiced–voiceless /bs/).
- The order of the place of the articulation of the sequence: front-to-back, e.g. /sk/, and back-to-front e.g. /kt/.
- The front or backness of the target vowel.
- The height of the target vowel.

The random effects considered in this study were participants and items. The significance of each varable was tested against the null hypothesis. If the p-value for a variable was > 0.05, we concluded that there was no effect on the outcome, and that the result was due to chance and therefore not statistically significant. Alternatively, if the p-value was ≤ 0.05, we could reject the null hypothesis and infer that the variable had an effect on the outcome.

For reasons of space, in the next paragraphs we report only the most interesting results in relation to our research aims.

3. RESULTS

The results of the three production tasks were combined, yielding a dataset of 21,205 tokens, with 58% of them exhibiting with vowel deletion. Vowel deletion was attested in all four obstruent sequences: FF, FP, PF, and PP. Word onset vowels were deleted across participants (M = 58%, SD = 4) and across items (M = 54%, SD = 30). Two factors were shown to be significant predictors of vowel deletion, namely gender and place of articulation of the initial obstruents (PoA). The result for each variable is discussed in detail below.

3.1 Gender

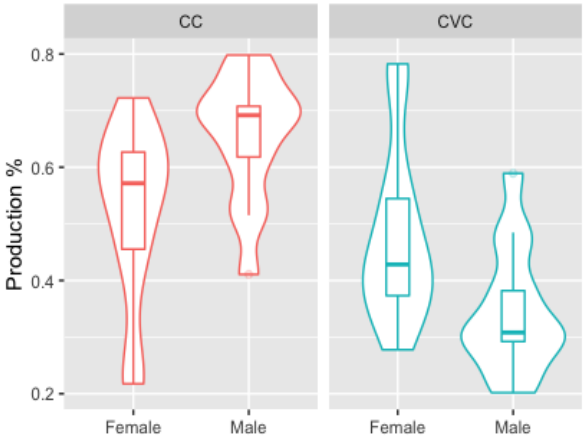


Figure 3: Vowel deletion (CC) and vowel retention (CVC) by gender.

Male speakers ($M = 65$, $SD = 11$) tended to delete more vowels than female speakers ($M = 52$, $SD = 15$). Overall, the rate of word onset vowel deletion among male speakers (66%) was higher than among female speakers (54%). The result of the mixed-effect logistic regression model showed a significant rise in vowel deletion among male speakers ($p < 0.01$), with a positive coefficient estimate of $b = 0.99$, indicating a significant rise in vowel deletion in males compared to females. The likelihood of vowel deletion by male speakers is 2.7 times higher than by female speakers.

3.2 PoA

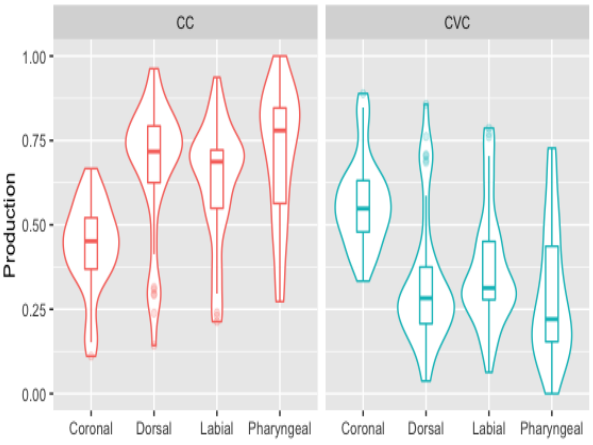


Figure 4: Vowel deletion (CC) and vowel retention (CVC) by PoA.

Participants produced pharyngeal-initial and dorsal initial sequences with the highest percentage of vowel deletion (71% and 69%, respectively), while coronal-initial sequences which were produced with the lowest percentage of vowel deletion (43%).

The findings of the mixed-effects logistic regression model showed a statistically significant effect of PoA for dorsal–initial sequences ($p = 0.03$), with a positive coefficient estimate of $b = 1.28$, indicating that dorsal–initial sequences were more likely to undergo vowel deletion compared to the the baseline reference (coronal-initial sequences).

The probability of vowel deletion in dorsal-initial sequences is 3.63 times higher than in coronal-initial sequences. For the labial-initial sequences, the likelihood of vowel deletion is 2.6 times higher than in coronal-initial sequences. Pharyngeal-initial sequences had the highest likelihood of vowel deletion, being 89% higher than coronal-initial sequences. Overall, coronal-initial sequences exhibited the least percentage of vowel deletion.

4. DISCUSSION AND CONCLUSIONS

In this paper, we have offered an acoustic analysis of vowel deletion in the dialect of NA. We show that vowel deletion is not solely a phenomenon of fast or casual speech, as it occurs at slow rates and with careful pronunciations, which is consistent with findings from previous research on NA [20, 21]. The data also shows that vowel deletion was observed in all four obstruent sequences: FF, FP, PF, and PP

In the present study, we also investigated the factors that contribute to vowel deletion. The findings indicated that two factors affect the rate of the vowel deletion. The first factor is the gender of the speaker, with male speakers tended to delete vowels more frequently than their female counterparts. Vowel-deleted variants are associated with the dialectal form, whereas vowel-present variants are linked with Standard Arabic. This suggests that female speakers tend to adhere to the prestigious variant of the language, as observed in other Arabic dialects [9].

The findings also revealed a statistically significant effect of the place of articulation of the first obstruents, particularly for dorsal-initial sequences. This finding is noteworthy for two reasons: firstly, it is the first time that PoA has been identified to play a role in vowel deletion in NA. Secondly, it contradicts with the findings from other languages, such as Japanese [22], where vowels are more likely to be deleted following coronals than labials or dorsals.

The data, while still preliminary, are enlightening as to the degree of vowel deletion in NA. This study highlights two additional factors that influence vowel deletion, beyond the conventional factors of stress and syllable structure that are widely acknowledged by Arabic linguists.

Acknowledgements: *This work was kindly supported by Princess Nourah Bint Abdulrahman University. We also thank the two anonymous reviewers.*

5. REFERENCES

[1] L. Davidson and K. Roon, "Durational correlates for differentiating consonant sequences in Russian," *Journal of the International Phonetic Association,* vol. 38, no. 2, pp. 137-165, 2008, doi: 10.1017/S0025100308003447.

[2] R. Al-Balushi, "Omani Arabic: More than a dialect," *Macrolinguistics,* vol. 4, no. 4, pp. 80-125, 2016, doi: 10.26478/ja2016.4.4.5.

[3] K. H. AbuAbbas, "Topics in the phonology of Jordanian Arabic: An optimality theory perspective," Dissertation/Thesis Doctoral dissertation, University of Kansas ProQuest Dissertations Publishing, 2003.

[4] A. Bürki, C. Fougeron, C. Gendrot, and U. H. Frauenfelder, "Phonetic reduction versus phonological deletion of French schwa: Some methodological issues," *Journal of Phonetics,* vol. 39, no. 3, pp. 279-288, 2011, doi: https://doi.org/10.1016/j.wocn.2010.07.003.

[5] L. Davidson, "Schwa elision in fast speech: Segmental deletion or gestural overlap?," (in English), *Phonetica,* vol. 63, no. 2-3, pp. 79-112, 2006, doi: 10.1159/000095304.

[6] F. Torreira and M. Ernestus, "Vowel elision in casual French: The case of vowel /e/ in the word c'était," (in English), *Journal of phonetics,* vol. 39, no. 1, pp. 50-58, 2011, doi: 10.1016/j.wocn.2010.11.003.

[7] R. Bani-Yasin and J. Owens, "The phonology of a Northern Jordanian Arabic dialect," *Zeitschrift der Deutschen Morgenländischen Gesellschaft,* vol. 137, no. 2, pp. 297-331, 1987. [Online]. Available: http://www.jstor.org/stable/43379559.

[8] M. Alqahtani, "Syllable structure and related processes in Optimality Theory: An examination of Najdi Arabic," Doctoral dissertation, Newcastle University, Newcastle University, 2014. [Online]. Available: http://hdl.handle.net/10443/2757

[9] A. M. A. Al-Harahsheh, "Language and gender differences in Jordanian spoken Arabic: A sociolinguistics perspective," (in English), *Theory and Practice in Language Studies,* vol. 4, no. 5, p. 872, 2014, doi: 10.4304/tpls.4.5.872-882.

[10] N. Ogasawara, "Lexical representation of Japanese vowel devoicing," *Language and speech,* vol. 56, no. 1, pp. 5-22, 2013, doi: https://doi.org/10.1177/0023830911434118.

[11] N. Topintzi and M. Baltazani, "The acoustics of high-vowel loss in a Northern Greek dialect and typological implications," in *Consonant clusters and structural complexity*, vol. 26, P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, and B. Kühnert Eds. Berlin, Boston: DE GRUYTER, 2012, pp. 373–402.

[12] E. Farnetani and M. G. Busà, "Quantifying the range of vowel reduction in Italian," in *14th International Congress of Phonetic Sciences (ICPhS-14)*, San Francisco, CA, USA, 1999, pp. 491-494., doi: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14_0491.html.

[13] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The Wildcat corpus of native-and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles," *Language and Speech,* vol. 53, no. 4, pp. 510-540, 2010.

[14] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods,* vol. 43, no. 3, pp. 761-770, 2011, doi: https://doi.org/10.3758/s13428-011-0075-y.

[15] N. Hall, "Acoustic differences between lexical and epenthetic vowels in Lebanese Arabic," *Journal of Phonetics,* vol. 41, no. 2, pp. 133-143, 2013, doi: https://doi.org/10.1016/j.wocn.2012.12.001.

[16] *Praat: doing phonetics by computer* (2018). [Online]. Available: http://www.praat.org/

[17] R. Skarnitzl and P. Machač, "Principles of phonetic segmentation," (in English), *Phonetica,* vol. 68, no. 3, pp. 198-199, 2011, doi: 10.1159/000331902.

[18] *R: A language and environment for statistical computing*. (2021). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[19] M. E. Brooks *et al.*, "GlmmTMB balances speed and flexibility among packages for Zero-inflated: Generalized Linear Mixed Modeling," *The R Journal,* vol. 9, no. 2, pp. 378-400, 2017, doi: 10.32614/RJ-2017-066.

[20] P. F. Abboud, "The verb in northern Najdi Arabic," *Bulletin of the School of Oriental and African Studies,* vol. 42, no. 3, pp. 467-499, 1979.

[21] B. Ingham, *Najdi Arabic: Central Arabian*. Amsterdam; Philadelphia: John Benjamins Publication Company, 1994.

[22] J. A. Shaw and S. Kawahara, "More on the articulation of devoiced [u] in Tokyo Japanese: Effects of surrounding consonants," (in English), *Phonetica,* vol. 78, no. 5, pp. 467-513, 2021, doi: 10.1515/phon-2021-2011.

LANGUAGE SPECIFICITY VS SPEAKER VARIABILITY OF ANTICIPATORY LABIAL COARTICULATION IN GERMAN AND ENGLISH

Justin J. H. Lo[1], Christopher Carignan[1], Marianne Pouplier[2], Roy Alderton[2], Francesco Rodriquez[2], Bronwen G. Evans[1], Eva Reinisch[3]

[1]University College London, UK, [2]Ludwig-Maximilians-Universität München, Germany, [3]Austrian Academy of Sciences, Austria

{justin.lo|c.carignan|bronwen.evans}@ucl.ac.uk, {pouplier|f.rodriquez}@phonetik.uni-muenchen.de, roy.alderton@city.ac.uk, eva.reinisch@oeaw.ac.at

ABSTRACT

This paper presents a study of the sources of cross-linguistic variation in anticipatory labial coarticulation. We collected audiovisual data from speakers of German and American English, two languages distinguished by the presence of phonological rounding contrast for vowels, to examine (i) how patterns of coarticulation differ in these languages and (ii) to what extent such patterns are speaker-specific.

Our results indicate that anticipatory labial coarticulation is temporally extensive in both German and English. While our data demonstrate that in both languages there is considerable variability between speakers in their individual degree of coarticulation, there is nonetheless an overarching influence of language on the patterns of coarticulation. In particular, anticipatory coarticulation is less extensive and more constrained in German than in English. These findings contribute to our understanding of the roles of phonological contrast and speaker variability in the temporal organization of anticipatory coarticulation.

Keywords: Lip rounding; coarticulation; individual variation; language specificity; speech production.

1. INTRODUCTION

In speech, the production of individual sound units is influenced by, and in turn influences, their surrounding contexts, as gestures required to articulate successive sounds overlap and interact with each other. The anticipatory initiation of articulatory movement, such as lip protrusion before rounded vowels, provides important articulatory and acoustic cues that are useful for perception in advance of the target sound itself [16]. Anticipatory lip movement in relation to rounding, the focus of this study, is particularly known to be temporally far-reaching, easily extending up to hundreds of milliseconds [10].

The temporal dynamics of anticipatory coarticulation are known to exhibit much cross-linguistic variation [2, 4, 20], but the source of such variation remains debated. Language-specific sound systems and contrasts have been argued to be influential factors governing the degree of coarticulation [12], in that the maintenance of phonological contrasts is expected to restrict the freedom employed by coarticulatory gestures [5, 17]. Yet, evidence for such a claim from various types of coarticulatory phenomena has thus far been mixed (e.g. [13, 18]). In particular, speakers of Swedish, in which rounding is phonologically contrastive, have been found to initiate anticipatory lip protrusion earlier, not later, than speakers of American English, which has no rounding contrast [11]. Even within individual languages, there is substantial speaker variability in the extent of coarticulation (e.g. [7, 19]), casting further doubt on the role of phonological contrast in the temporal organization of coarticulation.

In a cross-linguistic comparison of American English and Canadian French, Noiray et al. [14] examined the initiation of rounding movement in progressively longer [i]-C_n-[u] sequences and showed that the onset of lip constriction extended in proportion to the possible window of coarticulation at a speaker-specific rate. With no evidence of systematic patterns by language, they concluded that the implementation of anticipatory lip rounding is chiefly speaker-dependent.

In this study, we pursue the issue of the sources of cross-linguistic variation in coarticulation. Specifically, we investigate anticipatory labial coarticulation in German, which has phonological rounding contrast, and North American English, which does not, by tracking the dynamic movement of the lips. We examine how the temporal extent

of coarticulation differs between German and English, and address the question of whether patterns of coarticulation are predominantly driven by language-level variation or speaker variability.

2. METHODOLOGY

2.1. Materials

Stimuli consisted of a list of (near-)minimal pairs, wherein the target vowel differed by rounding (e.g. German *Sehne–Söhne* /zenə/-/zønə/). Each target word was embedded in a carrier sentence (German: *Aber Elsa legt gern __ beiseite*; English: *But Tessa had said __ pleasantly*), designed to be maximally permissible for anticipatory labial coarticulation. This environment, demarcated by the bilabial anchor /b/ (in *Aber* and *But*) and the onset of the target vowel, formed our region of interest (ROI). Table 1 lists the vowel pairs tested, but the German pairs involving /ɛ/ were excluded from the present analysis so that pairs were matched by phonological vowel height across languages. Three repetitions of the list were elicited in randomized blocks, yielding 144 utterances in German and 174 utterances in English per speaker.

Language	Vowel pairs
German	i–u, i–y, e–o, e–ø ɪ–ʊ, ɪ–ʏ, (ɛ–ɔ)*, (ɛ–œ)*
English	i–u, e–o, ɪ–ʊ

Table 1: Vowel pairs in German and English. Pairs marked * were excluded from analysis.

2.2. Data collection

We collected recordings from 30 native speakers each of German and North American English, using an adapted version of the "blue lip" technique developed by Lallouache [9] to track lip movement. Two video cameras were placed at right angles to capture frontal and side views of the speaker's blue-painted lips at 50 frames per second. Speakers were shown the stimuli on a computer monitor and prompted for production one at a time using *SpeechRecorder* [6], which simultaneously collected audio recordings. Utterances where prosodic breaks were identified immediately before the target word, or where the target word involved mispronunciation or false starts, were excluded from further analysis.

2.3. Measuring lip rounding & coarticulation

In line with previous research [14], we used the area of constriction between the lips (aperture) as a reliable indication of lip rounding. Within each utterance, we processed each frame from the front camera in Julia [1], first applying thresholding in HSV (hue, saturation, value) to isolate the blue lip and reference components, then fitting an ellipse to the area bounded by the inner edges of the upper and lower lips. We took the area of the ellipse to be the aperture and scaled it from pixels to mm^2. At this stage, data from one German and two English speakers were excluded due to technical difficulties with lip tracking. We smoothed the time-varying aperture signals by means of Hampel and Savitzky-Golay filters, and removed any trials with remaining erroneous measurements. A total of 4,074 trials in German and 4,726 trials in English remained.

To determine the onset of coarticulation, we analysed lip aperture in the rounded condition in relation to that in the unrounded condition and considered their point of divergence in time. We calculated the time-varying distance between the aperture of each utterance in the rounded condition and the mean aperture of all utterances in the unrounded condition by the same speaker in the corresponding word pair, within a fixed window of up to 500 ms before the acoustic onset of the target vowel. We then fitted sigmoid curves to each of the resulting difference curves (using the *sicegar* package [3] in R [15]), taking the point where the tangent line with the maximum slope intersected with the minimum asymptote as the onset of coarticulation. The example item in Fig. 1 illustrates this derivation. We excluded poor fits whose RMS error was amongst the highest 5% of each language, leaving 4,139 items in total. For the purposes of reporting and statistical analysis, we used the absolute value of the timing (and henceforth refer to onset of coarticulation as such), such that a higher positive value corresponds to an earlier onset of coarticulation.

2.4. Statistical analysis

To evaluate the effect of language on coarticulation, we fitted a linear mixed-effects model to the onsets of coarticulation obtained above (using the *lmerTest* package [8]). The main predictor of interest, language, was included as a sum-coded fixed effect. As controls, we also included phonological vowel height (sum-coded), ROI duration (mean-centred, as proxy for speech rate) and their interactions with language as fixed effects. To account for

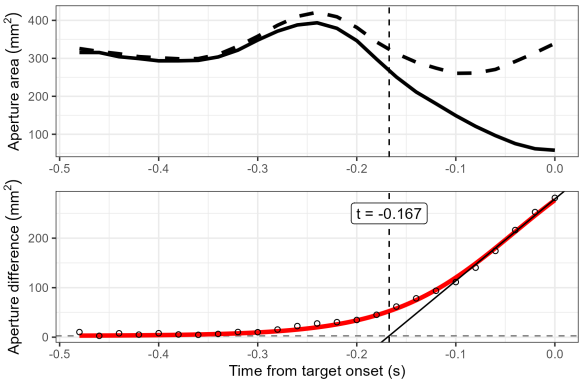


Figure 1: (Top) Aperture preceding /ø/ (solid) vs mean aperture preceding /e/ (dashed) in German *Löhne–Lehne*. (Bottom) Difference curve and its sigmoidal fit, with derived onset of coarticulation (dashed vertical line).

variability across items and speakers, we included random intercepts by word pair and by speaker, as well as random slopes for ROI duration by speaker. In addition, to assess speaker-level effects, we calculated the mean and standard deviation (SD) of onset of coarticulation for each speaker to examine speaker variability in closer detail.

3. RESULTS

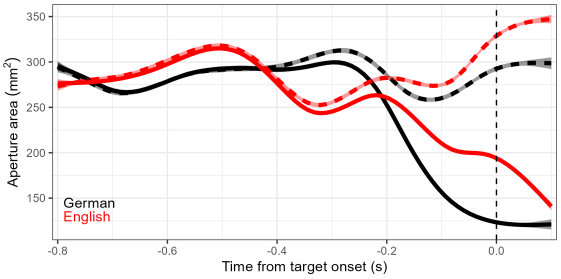


Figure 2: Overall aperture (mm^2) from 800 ms before to 100 ms after target onset (rounded: solid, unrounded: dashed; German: black, English: red; ribbons show ±2SD).

The overall trajectories of aperture in German and English are shown in Fig. 2. In German, aperture in both rounded and unrounded conditions progressed in tandem until shortly after 400 ms before the onset of the target vowel, when the aperture in the rounded condition grew increasingly smaller than that in the unrounded condition. Aperture in English followed trajectories that displayed broadly similar trends, with divergence between the two conditions emerging after 400 ms before target vowel onset. The distance between the conditions continued to rise, as aperture preceding unrounded vowels increased while that preceding rounded vowels fell.

Notwithstanding the similarities in the overall trajectories for German and English, our measure of onset of coarticulation, calculated on the level of individual utterances, revealed a significant effect of language ($p = .0002$). Onset of coarticulation was 53 ms earlier in English than in German, although, as illustrated in Fig. 3, divergence between rounded and unrounded conditions consistently commenced over 200 ms before target vowel onset in both languages. Vowel height was found to be significant as a main effect, but not in interaction with language: coarticulation was more extensive before /i/–/u/ or /y/ than before /ɪ/–/ʊ/ or /ʏ/ ($p = .0081$), and in turn more extensive than before /e/–/o/ or /ø/ ($p = .0024$). There was a significant interaction between language and ROI duration ($p = .0183$), such that earlier onset of coarticulation was found for longer ROIs, to a greater extent in German than in English (Fig. 4).

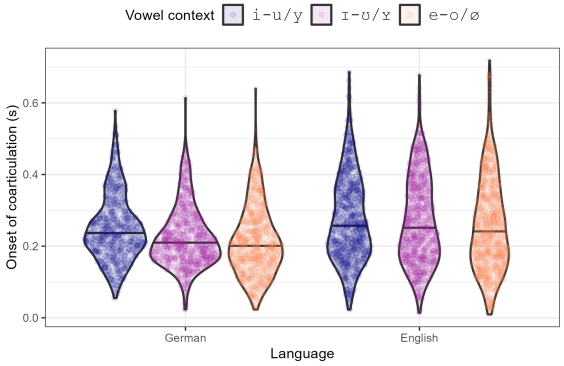


Figure 3: Distributions of onset of coarticulation (in s), measured relative to target vowel onset, in German and English by vowel context.

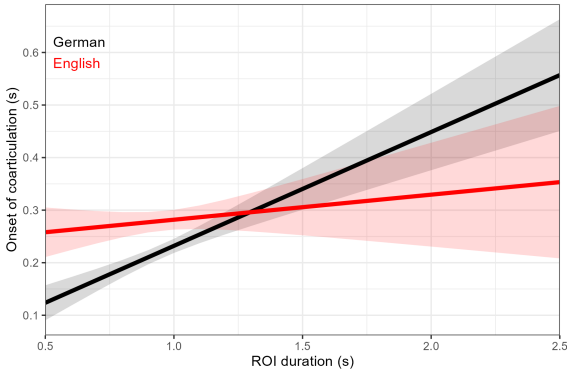


Figure 4: Model-predicted effect of ROI duration on onset of coarticulation (in s) in German (black) and English (red). Ribbons show ±2SD.

In addition to systematic cross-linguistic differences, we found substantial between-speaker variability in both languages, as is evident in Fig.

5. Mean onset of coarticulation for individual speakers, which shows considerable overlap between the two languages, spanned a similarly wide range in German (152–331 ms) and English (182–352 ms). It is also clear from Fig. 5 that onset of coarticulation among English speakers generally had a higher SD than among German speakers, suggesting that English, whose speakers reported higher means on the whole, also exhibited greater within-speaker variability.

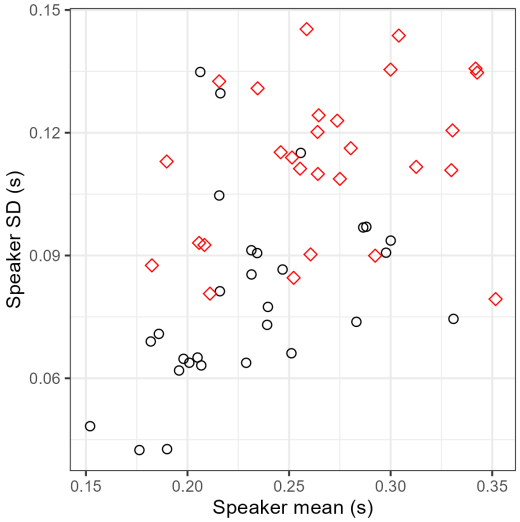


Figure 5: Scatterplot showing mean and SD of onset of coarticulation per speaker (German: black circles; English: red diamonds).

4. DISCUSSION

In this study, we compared the temporal extent of anticipatory labial coarticulation in German and English. Using audiovisual data to capture the dynamics of lip aperture in speech production, we have shown that anticipatory labial coarticulation is extensive in both languages. Coarticulatory information, in the form of differential lip aperture before rounded and unrounded vowels, regularly emerged over 200 ms before the onset of the target vowel in both languages, closely matching the temporal extent of anticipatory lip rounding reported in the literature [16].

Our findings here shed light on two opposing views on the driving force behind cross-linguistic variation in coarticulation. The evidence here lends support to the argument that language-specific phonological contrasts (in this case, rounding contrast in German) can act as a constraining factor on the degree of coarticulation (cf. [12]). Specifically, contrary to previous findings for Swedish [11], we showed that speakers of German, which similarly has phonological rounding contrast, initiated anticipatory coarticulation later, not earlier, than English speakers across all vowel contexts examined. Moreover, on an individual level, English speakers were comparatively less restricted in how much their extent of coarticulation varied from utterance to utterance.

The results here are also consistent with the position that implementation of labial coarticulation is highly speaker-specific [14]. Onset of coarticulation showed an impressive range of inter-speaker variation in both languages, ranging from an average of under 200 ms preceding the onset of the rounded vowel for speakers with the latest onset to over 330 ms for speakers with the earliest onset. Unlike in English and French [14], however, in the present comparison of English and German, speaker idiosyncrasy alone does not override the role of language. The significant influence of language is evidenced not only in the overall degree of coarticulation, but also in its differential response to speech rate variation (cf. [17]). It may be the case that a relatively small difference between languages is more effectively captured in our much larger set of speakers. Another possibility is that, while German and French both contrast rounding, the systemic influence on coarticulation is not a uniform one. We plan to address these issues with a parallel corpus of French speakers.

In the current study, we have focused on the temporal dynamics of lip aperture, yet anticipatory labial coarticulation may well vary by language in other dimensions. In future analyses, we plan to incorporate measures such as lip protrusion and spread to further investigate the question of how coarticulatory information is conveyed. Speakers of different languages may also follow distinct trajectories of coarticulation. While not our present focus, it is worth noting the dynamic differences within the target vowel itself in Fig. 2: in both rounded and unrounded vowels, lip aperture in German was maintained after reaching target onset, whereas in English sharp divergence between the two conditions continued well into the target vowel. More in-depth examination is thus warranted to better understand the time course of lip movement, as well as to fully explore the scope of coarticulatory variability across individuals and languages.

ACKNOWLEDGEMENTS

This work is jointly funded by the British Arts and Humanities Research Council and the German Research Foundation (AH/V002082/1).

5. REFERENCES

[1] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM review*, vol. 59, no. 1, pp. 65–98, 2017.

[2] S. E. Boyce, "Coarticulatory organization for lip rounding in Turkish and English," *J. Acoust. Soc. Am.*, vol. 88, no. 6, pp. 2584–2595, 1990.

[3] M. U. Caglar, A. I. Teufel, and C. O. Wilke, "Sicegar: R package for sigmoidal and double-sigmoidal curve fitting," *PeerJ*, vol. 6, p. e4251, 2018.

[4] H. Clumeck, "Patterns of soft palate movements in six languages," *J. Phon.*, vol. 4, no. 4, pp. 337–351, 1976.

[5] V. Delvaux, D. Demolin, B. Harmegnies, and A. Soquet, "The aerodynamics of nasalization in *French*," *J. Phon.*, vol. 36, no. 4, pp. 578–606, 2008.

[6] C. Draxler and K. Jänsch, "SpeechRecorder - a universal platform independent multi-channel audio recording software," in *Proc. LREC '04*, Lisbon, 2004, pp. 559–562.

[7] M. Grosvald, "Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation," *J. Phon.*, vol. 37, no. 2, pp. 173–188, 2009.

[8] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017.

[9] M. T. Lallouache, "Un poste visage-parole couleur: Acquisition et traitement automatique des contours des lèvres," Ph.D. dissertation, Institut National Polytechnique de Grenoble, 1991.

[10] J. Lubker, "Temporal aspects of speech production: Anticipatory labial coarticulation," *Phonetica*, vol. 38, no. 1–3, pp. 51–65, 1981.

[11] J. Lubker and T. Gay, "Anticipatory labial coarticulation: Experimental, biological, and linguistic variables," *J. Acoust. Soc. Am.*, vol. 71, no. 2, pp. 437–448, 1982.

[12] S. Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1286–1298, 1990.

[13] P. P. K. Mok, "Does vowel inventory density affect vowel-to-vowel coarticulation?" *Lang. Speech*, vol. 56, no. 2, pp. 191–209, 2013.

[14] A. Noiray, M.-A. Cathiard, L. Ménard, and C. Abry, "Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in french and english speakers," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 340–349, 2011.

[15] R Core Team, *R: A language and environment for statistical computing*, 2019. [Online]. Available: https://www.R-project.org/

[16] M. Redford, J. Kallay, and J. Potratz, "Temporal extent of anticipatory lip rounding in child and adult speech," *J. Acoust. Soc. Am.*, vol. 146, no. 4, p. 2925, 2019.

[17] M.-J. Solé, "Spatio-temporal patterns of velopharyngeal action in phonetic and phonological nasalization," *Lang. Speech*, vol. 38, no. 1, pp. 1–23, 1995.

[18] T. Ushijima and H. Hirose, "Electromyographic study of the velum during speech," *J. Phon.*, vol. 2, no. 4, pp. 315–326, 1974.

[19] G. Zellou, "Individual differences in the production of nasal coarticulation and perceptual compensation," *J. Phon.*, vol. 61, pp. 13–29, 2017.

[20] S. E. G. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, vol. 39, no. 1, pp. 151–168, 1966.

INVESTIGATING OVERLAPPING GESTURES IN ACOUSTIC SIGNALS: THE CASE OF FRICATIVES AND SONORANTS IN CLUSTERS.

Jérémy Genette

Universiteit Antwerpen
jeremy.genette@uantwerpen.be

ABSTRACT

It is investigated whether the difference between sonorants and fricatives in terms of overlapping articulatory gestures can be tracked in the acoustic signal. To this aim, a new method using spectral comparisons is introduced to make possible the consistent analysis of speech sounds characterised by markedly different acoustic features. The efficiency of this method is tested by evaluating overlapping articulatory gestures in French initial CCV syllables with either a fricative or a sonorant as C2 (N=1904). The stimuli were produced by 20 native speakers of Belgian French who participated in a reading task. The main findings confirm the results of previous articulatory studies, indicating that sonorants exhibit longer articulatory overlap with the following vowels than fricatives. By obtaining consistent results, this paper validates the suggested method for evaluating the dynamics of overlapping articulatory gestures.

Keywords: methodology, coarticulation, acoustics, epenthetic vowels, sound change.

1. INTRODUCTION

Phonetic explanations for recurrent sound change patterns have triggered much interest throughout the history of linguistics [1, 2]. According to Ohala [3, 4, 5], sound changes are the result of phonetic synchronic variations that are phonologized over time. In this framework, physical constraints on speech production and perception may favour the misinterpretation of the acoustic signal by listeners who do not compensate for coarticulation and consequently do not retrieve the speaker's intended phonological target. Therefore, such confusions can trigger the emergence of a sound change by incorporating the fortuitous – even if somehow constrained – phonetic variation into the phonology of the language. Articulatory gestures which overlap to different extents are typical examples of physical constraints that potentially lead listeners to reinterpret the signal due to coarticulation.

Research on such overlapping articulatory gestures has been implemented by analysing directly articulatory data [6] or by observing how overlapping gestures can affect perception [7]. If differences in overlap timing can be observed in speech production and if they are consequently perceived by listeners, it is assumed that cues for such overlapping gestures can be found in the acoustic signal. Such coarticulatory effects have been acoustically studied via dynamic analyses of formants for vowels (e.g. [8]) and sonorants (e.g. [9]), of the centre of gravity for fricatives (e.g. [10]), etc. However, to the best of our knowledge, there is no technique that can be consistently applied to speech sounds characterised by markedly different acoustic features, such as sonorants and fricatives.

The main objective of this paper is to introduce a general acoustic method for assessing the amount of coarticulatory influence between flanking phones, by using a Spectral Similarity Index. The efficiency of the method will be tested by observing whether it grasps the expected differences in overlapping articulatory gestures in CCV syllables whose C2 is either a sonorant or a fricative.

2. OVERLAPPING GESTURES IN CCV

One of the most recurrent sound change patterns is the simplification of complex syllabic structures into CV syllables. As for CCV syllables, the simplification can consist of either the deletion of one of the two consonantal segments (i.e., CCV > CV) or the insertion of a vowel between the consonants (i.e., CCV > CVCV).

In *Articulatory Phonology* (cfr. [11, 12]), the insertion of epenthetic vowels can be attributed to a mistiming of the different articulatory gestures. More specifically, an epenthetic vowel can result from a reduced articulatory overlap between C1 and C2 or from the articulatory gesture of the V being anticipated up to the C1-C2 transition [13]. The former account suggests the presence of a short gap between the articulatory implementation of C1 and C2, leaving the vocal tract open. The latter account suggests that the articulatory gesture for the vowel might be anticipated as early as the C1-C2 transition. Both gestural timings are thought to

favour the reinterpretation of CCV as CVCV by the listener. The gestural scores of both accounts are shown in Fig. 1.

It has long been noted, however, that some phonetic contexts favour vowel epenthesis more than others. For example, it is commonly assumed that the insertion of epenthetic vowels between consonants is favoured by plosive + sonorant rather than by plosive + fricative clusters [7]. This can reasonably be attributed to the precise articulatory configuration in terms of constriction width and pressure required for fricatives [6], while sonorants are characterized by a reduced build-up of oral pressure [14]. In Ohala's [3, 4, 5] terms, the specific articulatory configuration needed for the production of fricatives, as opposed to that of sonorants, could be the physical constraint that make the stop + fricative clusters less subject to reinterpretation than stop + sonorants clusters, hence less subject to sound change.

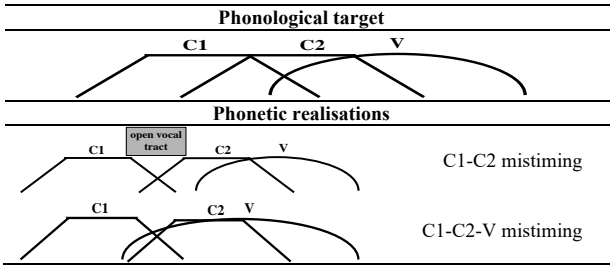


Figure 1: Gestural scores of the phonological target of CCV syllables and potential phonetic realisations (adapted from [13]).

In other words, the longer gestural overlap between a sonorant and the following V causes the listener to reinterpret a CCV syllable more easily as CVCV. Both articulatory and perception studies previously confirmed the propensity of plosive + sonorant clusters to favour the insertion of epenthetic vowels [6, 7]. On the one hand, the current evidence indicates that the vocalic articulatory overlap is longer in plosive + sonorant than in plosive + fricative clusters [6]. On the other hand, perceptual experiments confirm that CCV syllables are more often reinterpreted as CVCV if C2 is a sonorant [7]. Given the consistency of articulatory and perceptual evidence, it can reasonably be assumed that such differences in articulatory timing are reflected in the acoustic signal. The acoustic lens sheds light both on what is produced and perceived and enables a better integration of speech production and perception in the study of sound change.

In the following, a method is presented to study overlapping gestures acoustically. It is then observed whether it detects the difference in articulatory timing between sonorants and fricatives.

3. EXPERIMENT

To assess the extent of articulatory overlap between a speech sound (the target phone) and its flanking phones, its spectrum is compared to a theoretically non-coarticulated spectrum. The non-coarticulated spectrum is computed by instructing speakers to produce several repetitions of words containing the target phone with the preceding and following phones chosen to maximise the range of articulatory dimensions that might modify its spectral composition through coarticulation. From those stimuli, the spectrally stable portion of the target-phone is manually extracted, and its Long-Term Average Spectrum (LTAS) is computed. The gathered LTAS of all repetitions are averaged per-target phone and per-speaker to obtain a per-phone and per-speaker reference LTAS.

Then, the speech sounds to be analysed are divided into time-normalized frames, and the LTAS of each frame is computed. The spectral comparison is then performed between the LTAS of the extracted frames and the per-phone and per-speaker reference LTAS via a Spectral Similarity Index (SSI). By doing so, the amount of similarity between both spectra can be evaluated throughout the duration of the sound.

In this study, speech samples were collected from native French speakers via a self-paced reading task to investigate the dynamics of articulatory overlap between sonorants and fricatives in CCV syllables.

3.1. Participants and procedure

Data were collected from 20 native speakers of Standard Belgian French (11 women and 9 men, age: $M = 47.6$ years, $SD = 18.9$). They participated in a randomized self-paced reading task of French (pseudo-)words. All stimuli were presented visually using a *PsychoPy* experiment [15] on a computer screen placed at a comfortable reading level for each participant. The productions were recorded on a TONOR TC 30 in a quiet room. Before the start of the task, subjects were invited to read practice stimuli. The stimuli comprised both (pseudo-)words with initial CCV syllables containing the test stimuli and the reference stimuli used to elicit the reference LTAS.

3.2. Materials

3.2.1. Test stimuli

The materials for this study consisted of (pseudo-) words with a CCV syllable in initial position. The C1 of the CCV syllable was one of the plosives /p,

b, t, d, k, g/. The C2 was either one of the sonorants /ʀ, l, m/ or one of the fricatives /s, z/. The V was one of the three cardinal vowel /i, a, u/. In total, the stimuli set amounts to 90 pseudo-words (6 C1 * 5 C2 * 3 V) and 40 existing French words. The selected existing words were the most frequent lexical items with a given C1-C2-V combination, according to the *Lexique 3.83* database [16]. Some stimuli were excluded due to technical issues and omissions in the responses of the participants, leaving 1904 tokens.

3.2.2. Reference stimuli

The selected sonorants and fricatives were produced in 4 repetitions of 15 words in which they were followed by one of the 3 cardinal vowels /i, a, u/. For instance, /z/ was produced in 4 repetitions of the words *hasard* /azaʀ/ "chance", *visite* /vizit/ "tour" and *bouzouki* /buzuki/ "bouzouki". The reference stimuli were randomly inserted within the test stimuli during the reading task.

3.3. Acoustic analysis

3.3.1. Test stimuli

The C2 of the CCV syllables were segmented manually after the release burst of the C1 and before the beginning of the vowel. Then, each C2 was divided into 10 time-normalized frames. For each frame, the LTAS was computed using a frequency band of 100 Hz across the sampling frequency (48 kHz) via PRAAT [17] and the *PraatR* package [18].

3.3.2. Reference stimuli

The spectrally stable portion of the sonorants and fricatives – i.e., the portion of the sound that is least affected by coarticulation – in the reference stimuli were segmented manually through visual inspection. The LTAS was then computed for each segment, using a frequency band of 100 Hz throughout the sampling frequency (48 kHz) via PRAAT [17]. The LTAS of each C2 were averaged to obtain a per-speaker and per-C2 reference LTAS.

3.4. Spectral Similarity Index

By considering each LTAS as a vector of n dimensions, the 10 frames of the target-C2 were compared with the reference LTAS of the corresponding C2 via the Spectral Similarity Index (SSI) as in Eq. 1, where x stands for the reference spectrum, y is the test spectrum, i is the frequency bin and n is the number of frequency bins.

The resulting SSI values indicate the similarity between the reference C2 and each frame of the test stimuli. A high SSI value indicates a high similarity between both LTAS.

$$(1)\ SSI = \frac{\sum_{i=1}^{n}(x_i-\bar{x}).(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

3.5. Statistical analysis

Multi-level modeling was used as the statistical technique. Models of increasing complexity were built step by step by including random and fixed effects one after the other [19, 20]. The statistical analysis was carried out in R [21] using the *lme4* package [19] and the *lmerTest* package [22] to obtain *p*-values. The final model included time (i.e., the frame number), a quadratic effect of time, C2 type (sonorant vs. fricative), lexicality (word vs. pseudo-word), and their interactions, as well as subjects and items nested within subjects as random effects.

4. RESULTS

Table 1 shows the results of the final model, and a graphical representation of the SSI dynamics throughout C2 is presented in Fig. 2. In Table 1, only the main fixed effects are presented, as well as the interaction effects that are significant. The results show that SSI increases with time, as indicated by a significant effect of time (p<.001). The effect of time is not linear, as shown by the significant quadratic effect of time (p<.001). The main effect of C2 type (sonorant vs fricative) is also significant (p<.001), but the nonlinear effect of time is different depending on the type of C2, as shown by the significant interaction between both factors (p<.001) and by the significant three-way interaction with time (p<.001). The main effect of lexicality is, however, not significant (p=.68).

	Estimate	SE	*t*-value	*p*-value
Intercept	809.30	0.005	149.94	<.001***
Time	0.18	0.03	6.02	<.001***
Quadratic time	0.24	0.06	3.89	<.001***
C2 type [son.]	0.06	0.006	10.28	<.001***
Lexicality [word]	0.004	0.009	0.42	0.68
C2 type [son.] * Quadratic time	-3.69	0.09	-4.24	<.001***
Time * Quadratic time	-0.42	0.03	-11.18	<.001***
C2 type [son.] * Time * Quadratic time	0.45	0.005	8.55	<.001***

Table 1: Fixed main effects and significant interaction effects on SSI; C2 type [fri.] & Lexicality [pseudo-word] = ref. category.

As can be seen in Fig. 2, the SSI dynamics of fricatives exhibit a bell-shaped trajectory, while sonorants show a shallow SSI increase at the beginning followed by a relatively flat SSI trajectory. The sharp and short decreases in SSI at the end of fricatives indicate that the coarticulary influence of the vowel begins late and does not reach the first half of the fricative. The steep

increase at the beginning can be attributed to the coarticulatory influence of C1. In contrast, in sonorants, the maximum level of dissimilarity due to the vowel is quite stable over the last two-thirds of C2 and gently decreases towards its beginning. Yet, after visual inspection of the spectrograms, the low SSI at the beginning of sonorants, must be attributed to a period of relative silence between the release of the C1 occlusion and the occlusion needed to produce /m/ (see Fig. 3.a.), rather than to coarticulatory transition patterns as in /ʀ/ and /l/ (see Fig. 3.b.). Nevertheless, the more stable SSI curve in sonorants indicates that the effect of the vowel, which is expected to be maximal at the C2-V boundary, remains constant for a longer period of time in sonorants than fricatives.

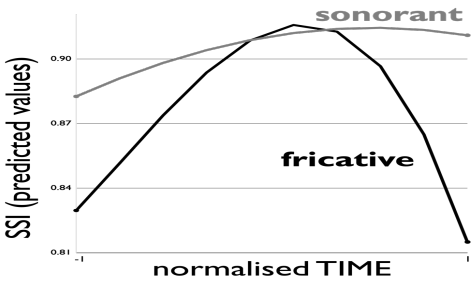
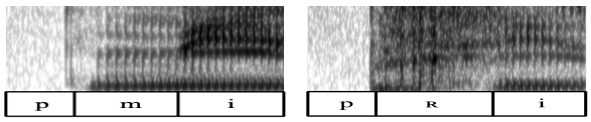


Figure 2: SSI dynamics (predicted values) as a function of C2 type (fricative vs sonorant).



(a) (b)

Figure 3: Example spectrograms of CCV syllables with /m/ (a) and /ʀ/ (b) as C2.

5. DISCUSSION

This study aimed at observing whether differences in the timing of coarticulatory gestures in CCV syllables can be observed in the acoustic signal. For this purpose, the SSI metrics was designed to assess the extent to which a given frame of C2 is spectrally different from a theoretically non-coarticulated spectrum. The C2 of 1904 clusters were analysed with the suggested method. The main finding is that sonorants tend to overlap more with the successive vowel than fricatives, which is consistent with previous articulatory and perceptual studies. In fact, the results show a clear difference in behaviour between fricatives and sonorants.

The evolution of the SSI in fricatives shows a bell-shaped curve, indicating that the coarticulatory influence of the vowel covers the second part of the C2, but not earlier. This suggests that the articulatory gestures of the vowel do not extend to the beginning of the C2. Therefore, it is assumed that the phonetic realization of the CCV syllable is relatively close to the phonological target, as shown in Fig. 1. This makes it unlikely that the signal of the CCV syllable is misinterpreted by the listener as CVCV. The large decrease in gestural overlap could be attributed to the precise articulatory configuration in terms of airflow and constriction width necessary to produce frication noise [6]. In contrast, the SSI trajectory of the sonorants is much more stable and follows a rising trend. It potentially indicates that the coarticulatory influence of V covers a larger portion of C2 and merges with the influence of C1 at the beginning of C2. The observed major articulatory overlap between sonorants and vowels, attributed to the less constrained articulatory specifications of sonorants, might lead listeners to reinterpret the signal as CVCV, rather than as a CCV syllable.

It must be noted tough that the absolute SSI values do not provide information about the amount of the coarticulatory influence *per se* because the SSI heavily depends on the spectral variability between reference stimuli, hence the average high SSI values observed with sonorants. However, the shape of the trajectories provides insights into how the preceding and following speech sounds affect the SSI for each C2 type. An analysis considering the different types of sonorants as well as the effect of potential allophonic realisations of French /ʀ/ would provide more detailed insights, but this is beyond the scope of the present paper.

In short, this technique could provide an alternative to more expensive and invasive articulatory methods, albeit with limitations. By complementing articulatory and perceptual studies, acoustic analyses such as this one could also lead to a deeper understanding of sound change processes.

6. CONCLUSIONS

The present study aimed at observing whether differences in the timing of coarticulatory gestures in CCV syllables can be tracked in the acoustic signal. A new method is suggested by computing a per-speaker and per-C2 reference LTAS with which to compare the LTAS of the test C2.

The results show that the coarticulatory influence of V on C2 was longer for sonorants than fricatives, confirming the results of previous studies. The results indicate the feasibility of tracking overlapping articulatory gestures in the acoustic signal. The suggested technique cannot identify specific articulatory gestures, but it permits the assessment of coarticulatory dynamics in speech sounds differing in acoustic features, such as sonorants and fricatives.

7. ACKNOWLEDGEMENTS

Thanks are due to the participants of this study. The research reported in this paper was partly supported by a research grant from the Research Foundation – Flanders (FWO) [grant G004321N].

8. REFERENCES

[1] Pinget, A.-F. 2015. *The actuation of sound change.* LOT.

[2] Garrett, A., Johnson, K. 2013. Phonetic bias in sound change. In: Yu, A. C. L. (ed.), *Origins of sound change: Approaches to phonologization*. Oxford University Press, 51–97.

[3] Ohala, J. J. 1989. Discussion of Lindblom's 'Phonetic invariance and the adaptive nature of speech'. In: Elsendoorn, B. A. G., Bouma, H. (eds.), *Working models of human perception*. Academic Press, 175–183.

[4] Ohala, J. J. 1989. Sound change is drawn from a pool of synchronic variation. In: Breivik, L. E., Jahr, E. H., (eds.), *Language change: Contributions to the study of its causes*. Mouton de Gruyter, 173–198.

[5] Ohala, J. J. 1993. Sound change as nature's speech perception experiment. *Speech Communication* 13, 155–161.

[6] Recasens, D. 2018. *The production of consonant clusters: Implications for phonology and sound change*. De Gruyter Mouton.

[7] Fleischhacker, H. A. 2005. *Similarity in phonology: Evidence from reduplication and loan adaptation*. University of California.

[8] Krull, D. 1989. Consonant-vowel coarticulation in spontaneous speech and in reference words. *Speech Transmission Laboratory Quarterly Progress and Status Report* 1(5), 101–105.

[9] Themistocleous, C., Fyndanis, V., Tsapkini, K. 2022. Sonorant spectra and coarticulation distinguish speakers with different dialects. *Speech Communication* 142, 1–14.

[10] Lulaci, T., Tronnier, M., Söderström, P., Roll, M. 2022. The time course of onset CV coarticulation. *Fonetik 2022 - the XXXIIIrd Swedish Phonetics Conference*.

[11] Browman, C. P., Goldstein, L. M. 1986. Towards an articulatory phonology. *Phonology* 3, 219–252.

[12] Browman, C. P., Goldstein, L. M. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

[13] Buchwald, A. B., Rapp, B., Stone, M. 2007. Insertion of discrete phonological units: An articulatory and acoustic investigation of aphasic speech. *Language and Cognitive Processes* 22(6), 910–948.

[14] Ladefoged, P. 1997. Linguistic phonetic descriptions. In: Hardcastle, W. J., Laver, J. (eds.), *The handbook of phonetic sciences*. Wiley, 589–618.

[15] Peirce, J.W. 2007. PsychoPy–psychophysics software in Python. *Journal of neuroscience methods* 162(1-2), 8–13.

[16] New, B., Pallier, C. 2020. Lexique 3.83. http://www.lexique.org

[17] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot. International* 5(9), 341–345.

[18] Albin, A. 2014. PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language. *Journal of the Acoustical Society of America* 135(4), 2198.

[19] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1).

[20] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390–412.

[21] R Development Core Team 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[22] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2015. Package 'lmertest'. *R Package Version* 2(0), 734.

VARIABLE RHOTICITY IN GLASGOW ENGLISH: SOCIOLINGUISTIC FACTORS AND ABSTRACTIONS VS. EXEMPLARS

Kamil Kaźmierski

Faculty of English, Adam Mickiewicz University, Poznań, Poland
kamil.kazmierski@amu.edu.pl

ABSTRACT

This paper uses a corpus of Glasgow English unscripted speech to investigate the variability in the realization of pre-consonantal /r/. The relationship between the gender of the speaker, as well as their familiarity with the interlocutor, and the presence of pre-consonantal /r/ is investigated using mixed-effects binomial regression modeling. Additionally, the dataset is probed for evidence of abstract representations and phonetically-rich exemplars in the storage of /r/.

A statistically significant influence of the interaction between familiarity with interlocutor and speaker gender on the likelihood of non-rhoticity provides some evidence that non-rhoticity is a prestige variant for the speech community that the corpus is representative of. The statistical significance of by-word random intercepts, attesting to word-specific effects, provides evidence for exemplars. The lack of a statistically significant effect of a clearly definable phonological context (the backness of the vowel preceding /r/) on the likelihood of the presence of pre-consonantal /r/ means that the current study does not provide support for abstractionism or hybrid models.

Keywords: variable rhoticity, Glasgow English, sound change, corpus phonology

1. INTRODUCTION

Scottish English accents are generally thought of as rhotic, with /r/ being realized as a consonant (an approximant [ɹ], a tap [ɾ], or a trill [r]) regardless of its position in a syllable [1, 2].

However, there is actually considerable variation with regard to the presence of pre-consonantal /r/ in Scottish English accents [3, 4], including Glaswegian [5, 6]. Early reports of weakening rhoticity in Glasgow English identified the loss of pre-consonantal /r/ in working-class speech [7, 2], suggesting it is a local development, specifically not driven by the non-rhoticity of the London-based English accent.

1.1. Change from below or from above?

The incipient non-rhoticity in working-class speakers in Glasgow has been interpreted as a local development, given low volume of face-to-face interactions with speakers of non-rhotic English accents. Findings such as [3], on the other hand, suggest that non-rhoticity might be a new overt prestige variant after all.

In cases of change in progress, style-shifting, with higher rates of the innovative variant over the conservative variant when more attention is paid to speech, has been shown to be indicative of 'change from above', i.e. conscious adoption of a variant associated with a language variety regarded as higher in prestige [8]. Additionally, women have often been shown to be the leaders of change, regardless of whether the change is from above (driven by the overt prestige of a language variety it is imported from) [9] or from below [10] the level of consciousness (imbued with the 'covert' prestige of the local variety). The present study considers the influence of subtle changes in speech style due to familiarity with the interlocutor or lack thereof. Style-shifting, in particular among women, in the direction of non-rhoticity will be seen as indicative of non-rhoticity being a prestige variant. Style-shifting in the direction of rhoticity, on the other hand, will be seen as indicative of speakers orienting themselves to the rhotic variant as prestigious, with non-rhoticity carying 'covert' prestige.

1.2. Phonological storage: Abstractions vs. exemplars

The abstractionist approach to phonological storage and processing is exemplified by most generative frameworks, and rests upon the model of speech production proposed by Levelt et al. [11], with abstract lexical representations and feed-forward architecture of speech production involving discrete modules. The phonetic shape of a lexical item is based on the abstract phonological building blocks that it is composed of. The phonetic shape is further allowed to be influenced by frequency effects

- with higher frequency items showing a greater degree of reduction. Any word-specific effects beyond frequency effects, however, are ruled out by this approach. Finding an effect of an abstract phonological feature (here: vowel backness) on the rate of non-rhoticity will be taken as supporting the abstractionist approach.

The rich storage approach, on the other hand, is associated with the exemplar theory [12, 13, 14] and posits phonetically rich representations as playing a pivotal role in speech production. Phonetically rich representations of lexical items are assumed to be stored in the mental lexicon, and lexical items, which if analyzed into abstract phonemes would be identical, might in fact have different phonological representations. Multiple exemplars of lexical items are drawn upon in speech production. Hence, word-specific effects which go beyond frequency effects are expected to be commonplace. Finding an effect of by-word random intercepts will be taken as supporting the rich storage approach.

Reconciling the two approaches by recognizing the need for both kinds of representations in the mental lexicon, so-called hybrid models of phonology have been proposed [15, 16, 17]. Finding an effect of both vowel backness and by-word random intercepts will be taken as supporting hybrid models.

2. METHOD

2.1. Data

The data come from the freely available HCRC Map Task corpus [18]. It contains recordings of unscripted speech of speakers coming from three areas: (1) Glasgow, (2) elsewhere in Scotland, and (3) 'rest of the world' (as per speaker metadata). Each speaker interacted with a person they knew - a familiar interlocutor, and a person they did not know - an unfamiliar interlocutor. For the present study, only speakers from Glasgow were included. After the exclusion of non-Glasgow speakers, some Glasgow speakers were left speaking only to either a familiar, or only to an unfamiliar interlocutor; these speakers were also excluded (final number of speakers: $N = 29$). Speaker age ranged from 18 to 26, with most speakers closer to the lower end of the spectrum (*median* = 19). Pairs of speakers performed an information-gap speaking task, where speaker A, using a map they were given, had to give instructions to speaker B, who was also provided with a map, so that speaker B would trace a path from start to finish. Due to mismatches in the placement of landmarks on the maps provided to each member of the pair, the need for clarification would often ensue. The HCRC Map Task corpus is distributed with word-level time-aligned orthographic transcriptions in XML format. They were converted to TextGrids and imported into LaBB-CAT [19] for additional automatic annotation and for querying. Using a phonemic transcription layer created in LaBB-CAT, 5,330 word tokens with pre-consonantal /r/ were retrieved. The tokens were then manually annotated by the author based on audition, supported by inspection of waveforms and spectrograms, as belonging to one of the four categories: non-rhotic (1), containing an approximant (2), tap (3), or trill (4). Examples of each of the four categories are shown in in Figs. 1-4.

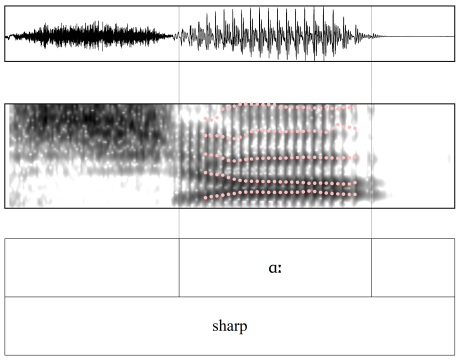


Figure 1: A token of *sharp* tagged as **non-rhotic**, produced by a 17-year-old male speaker

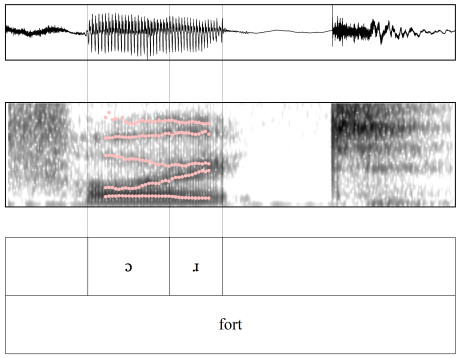


Figure 2: A token of *fort* tagged as an **approximant**, produced by a 19-year-old female speaker

2.2. Empirical results

2,039 tokens (37.7%) were non-rhotic, 1,969 tokens (36.4%) were realized as approximants, 1,373 tokens (25.8%) were realized as taps, and only 7 tokens (0.1%) were realized as trills.

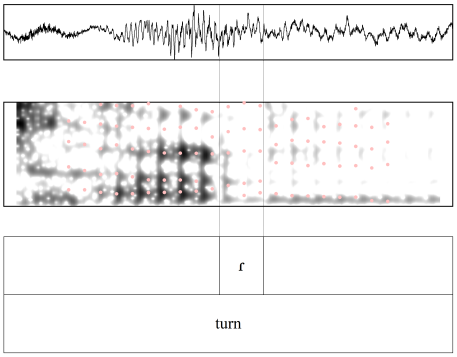


Figure 3: A token of *turn* tagged as a **tap**, produced by an 18-year-old male speaker

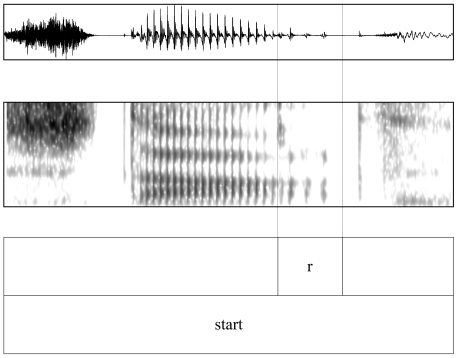


Figure 4: A token of *start* tagged as a **trill**, produced by an 18-year-old male speaker

For further analysis, trills were removed due to their extreme rarity, and approximants and taps were grouped together as 'rhotic' realizations.

There is a good deal of individual variation: some speakers are close to being categorically rhotic, some are close to being categorically non-rhotic, with most positioned somewhere along these two end points. A Bayesian mixed-effects binary regression intercept-only model with by-item and by-speaker varying intercepts was fitted to estimate the population-level and speaker-level probabilities of non-rhoticity. The population-level estimated mean probability is 0.3 (95% Credible Interval = [0.21, 0.4]). Speaker-level probabilities are shown in Fig. 5.

The results, broken down by speaker gender and by familiarity with the interlocutor are presented in Fig. 6. For men, there is no appreciable difference in non-rhoticity rate contingent on the familiarity with their interlocutor: it is very close to the population-level mean (37%) regardless of whether they are speaking with a familiar (36%) or an unfamiliar interlocutor (39%). For women, however, the non-rhoticity rate goes above the

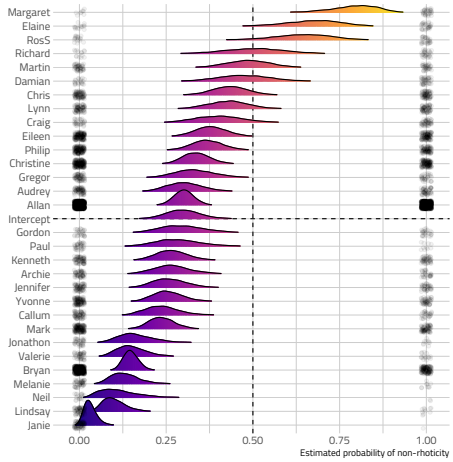


Figure 5: Estimated probabilities of non-rhoticity for individual speakers. 'Intercept' is the population-level estimate, jittered and semi-transparent points are individual tokens (rhotic at 0; non-rhotic at 1)

population-level mean when they are speaking with an unfamiliar interlocutor (42%), and sinks well below the population-level mean when they are speaking with a familiar interlocutor (31%).

Breaking the results down by vowel backness, 49% of tokens with back vowels were non-rhotic (618/1,253), 34% of tokens with central vowels (1,155/3,401), and 32% of tokens with front vowels (219/676). Hence, back vowels seem to favor non-rhoticity.

The statistical significance of the effects is assessed in the next section.

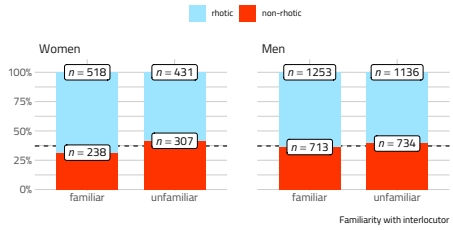


Figure 6: Rate of non-rhoticity by gender by familiarity with interlocutor

2.3. Modeling

A mixed-effects binary logistic regression model of non-rhoticity, i.e. the absence of pre-consonantal /r/ was fitted using the *lme4* package [20] in R [21]. Calculation of all *p*-values for the purposes of significance testing was conducted by means of likelihood ratio tests.

2.3.1. Test variables

Speaker gender (binary categorical predictor, reference = *female*, and familiarity with the interlocutor (binary categorical predictor, reference = *yes*) were coded based on corpus metadata. Additionally, their interaction was included to allow for the fact that familiarity/unfamiliarity with the interlocutor might have a larger influence on women than on men.

In the present data set, /r/ occurred before front vowels in the NEAR /i/ and SQUARE /e/ lexical sets, before central vowels in the NURSE /ʌ/, START /a/, and CURE /ʉ/ lexical sets, and before back vowels in the FORCE /ɔ/ and NORTH /o/ lexical sets

Back vowels /ɔ/ and /o/ have been previously shown to favor non-rhoticity [22], which is in line with the descriptive statistics of the present data set. Vowel class was coded as a binary categorical predictor with the levels *non-back* and *back* (reference: *non-back*). This enabled significance testing relevant to the question of an abstract phonological feature (here: vowel backness) influencing non-rhoticity.

To capture word-specific effects, a by-word random intercept was included in the model.

2.3.2. Control variables

Since non-rhoticity can be argued to be influenced by frequency effects, word forms were tagged with *Log*10 frequency from *SUBTLEX-UK* [23] and entered as a numerical predictor.

To allow for speaker-specific effects (which is always necessary when multiple data points from a single speakers appear in the data set, but compare also Fig. 5), a by-speaker random intercept was included.

2.4. Results

Table 1 presents the estimates of the model and their associated *p*-values. Note that all estimates are in log-odds. Women are more likely (b = 0.52) to produce non-rhotic tokens when speaking to unfamiliar interlocutors than when speaking to familiar interlocutors. As attested by the familiarity by gender interaction term, the degree to which familiarity with the interlocutor affects each gender is different. When the estimate for the interaction (b = -0.41) is added to the estimate for familiarity speakers (b = 0.52) the result is very close to zero (0.11): there is hardly any difference in the predicted rate of non-rhoticity for men, regardless of whether they are speaking to a familiar or to an unfamiliar interlocutor.

The higher rate of non-rhoticity for back vowels has not turned out to be statistically significant (p = 0.196).

The by-word random intercept (not shown in the table), on the other hand, has turned out to be statistically significant ($p < 0.001$), which was assessed, like for fixed effects, with a likelihood-ratio test.

The frequency predictor has a positive estimate (b = 0.32) and is statistically significant (p = 0.018), showing that increased frequency is associated with a higher probability of non-rhoticity.

Fixed effect	Estimate	p-value
$(Intercept)$	-2.5	< 0.001
$familiar = no$	0.52	< 0.001
$gender = m$	0.38	0.02
$log10frequency$	0.32	0.018
$vowel = back$	0.39	0.196
$familiar = no : gender = m$	-0.41	0.005

Table 1: Fixed effects of the mixed-effects binary logistic regression model of **non-rhoticity**: p-values calculated with likelihood ratio tests

3. DISCUSSION

The results with regard to gender and familiarity are consistent with the hypothesis that non-rhoticity is an innovative prestige variant in the speech community that the corpus is representative of. Given that reports of non-rhoticity in Scottish English accents are relatively recent, we are more likely seeing a case of language change in progress than stable stylistic variation. Non-rhoticity is more likely when speaking to an unfamiliar interlocutor, when, arguably, more attention is being paid to speech than when speaking to a familiar interlocutor. Thus, the results attests to style shifting, and non-rhoticity appears to be a prestige variant. The effect is observable in women, who often display more style-shifting than men. The finding that non-rhoticity carries overt prestige concurs with [3], and seems to contrast with earlier reports of non-rhoticity carrying covert prestige in working-class Glasgow speech [7, 2]. Conceivably, though, non-rhoticity has different socio-indexical meaning for Glasgow English speakers depending on their socio-economic status (for the corpus used here, unfortunately, no data pertaining to speakers' *SES* is available).

As for the question of abstractionist versus phonetically rich versus hybrid models of

phonological storage, the word-specific effect on the probability of non-rhoticity which goes beyond frequency effects provides support for the rich storage approach. Even though non-rhoticity was more likely in the context of back vowels than in the context of non-back vowels, this effect was not statistically significant. Thus, no support for abstractionism, and, by extension, for hybrid models, is provided.

4. ACKNOWLEDGMENTS

This research was supported by the National Science Centre (Poland) grant No. UMO-2017/26/D/HS2/00027.

5. REFERENCES

[1] J. C. Wells, *Accents of English*. Cambridge: Cambridge University Press, 1982, vol. The British Isles.

[2] J. Stuart-Smith, *Urban voices: Accent studies in the British Isles*. London: Routledge, 1999, ch. Glasgow: accent and voice quality, pp. 203–222.

[3] O. Schützler, "Variable Scottish English consonants: The cases of /ʍ/ and non-prevocalic /r/," *Research in Language*, vol. 8, pp. 5–21, oct 2010.

[4] P. Meer, R. Fuchs, A. Gerfer, U. Gut, and Z. Li, "Rhotics in standard scottish english," *English World-Wide*, vol. 42, no. 2, pp. 121–144, may 2021.

[5] J. Stuart-Smith, M. Sonderegger, T. Rathcke, and R. Macdonald, "The private life of stops: VOT in a real-time corpus of spontaneous glaswegian," *Laboratory Phonology*, vol. 6, no. 3-4, jan 2015.

[6] M. Sóskuthy and J. Stuart-Smith, "Voice quality and coda /r/ in Glasgow English in the early 20th century," *Language Variation and Change*, vol. 32, no. 2, pp. 133–157, jul 2020.

[7] C. Macafee, *Glasgow*, ser. Varieties of English around the world. Amsterdam: John Benjamins, 1983.

[8] W. Labov, *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 1966.

[9] ——, *Principles of linguistic change. Vol. 2: Social factors*. Malden, MA: Wiley-Blackwell, 2001. [Online]. Available: https://www.ebook.de/de/product/4266481/labov_principles_of_linguistic_change_v2.html

[10] P. Eckert, "Adolescent social structure and the spread of linguistic change," *Language in Society*, vol. 17, no. 2, pp. 183–207, 1988.

[11] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and Brain Sciences*, vol. 22, pp. 1–75, feb 1999. [Online]. Available: https://doi.org/10.1017/s0140525x99001776

[12] J. Bybee, *Phonology and language use*. Cambridge: Cambridge University Press, 2001.

[13] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker variability in speech processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 145–165.

[14] J. B. Pierrehumbert, "Word-specific phonetics," in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 101–139.

[15] ——, "The next toolkit," *Journal of Phonetics*, vol. 34, pp. 516–530, 2006.

[16] S. D. Goldinger, "A complementary-systems approach to abstract and episodic speech perception," in *Proceedings of the 16th International Congress of Phonetic Sciences*, J. Trouvain and W. J. Barry, Eds., SaarbrÃ¼cken, 2007, pp. 49–54.

[17] M. Ernestus, "Acoustic reduction and the roles of abstractions and exemplars in speech processing," *Lingua*, vol. 142, pp. 27–41, 2014.

[18] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, oct 1991.

[19] R. Fromont and J. Hay, "LaBB-CAT: An Annotation store," in *Proceedings of the Australasian Language Technology Association Workshop*, Dunedin, New Zealand, 2012, pp. 113–117.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[22] V. Dickson and L. Hall-Lew, "Class, gender, and rhoticity: the social stratification of non-prevocalic /r/ in Edinburgh speech," *Journal of English Linguistics*, vol. 45, no. 3, pp. 229–259, jul 2017.

[23] W. J. B. V. Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-UK: a new and improved word frequency database for British English," *Quarterly Journal of Experimental Psychology*, no. 67, pp. 1176–1190, 2014.

An Electropalatographic study of the Singleton-Geminate contrast in Arabic

Amel Issa

University of Gharyan
amel_issa@hotmail.com

ABSTRACT

This study investigates the articulatory correlates of the singleton and geminate consonants in Tripolitanian Libyan Arabic (TLA) using Electropalatographic (EPG) data. The aim of this study is to ascertain whether underlying differences are reflected in articulatory dissimilarity. The location, size and shape of linguopalatal contact were examined together with the flatness of the tongue degree. The articulatory results show systematic articulatory differences between the singleton and geminate consonants. The data provide evidence of the use of different regions of articulators for gemination. The data provide evidence for apical contact for singletons as opposed to laminal contact for geminates and a flatter shape of the tongue in geminate articulation. The results are suggestive of a strengthening effect on the articulatory level. The results of the current study confirm the robustness of the effect of the phonological context on the articulatory gestural plans of the singleton and geminate consonants.

Keywords: Electropalatography (EPG), gemination, articulatory data, Libyan Arabic, sonorant sounds.

1. INTRODUCTION

Traditionally, durational cues to gemination have formed the main emphasis in investigating the singleton-geminate contrast and, generally, these studies in various languages have shown that duration is the most robust correlate of gemination (see e.g. [1], [2] and [3], among others). A number of non-durational correlates of geminates have also been investigated by researchers for many languages (see e.g. [4], and [5], among others). Some of these cues can be investigated acoustically, while others can also be investigated through articulatory analysis. For instance, [5] and [6] showed that, in Italian, geminate laterals are more palatalized compared with their singleton counterparts both acoustically and articulatorily. She also found articulatory evidence, using Electropalatography, for the existence of apical contact for singletons as opposed to laminal contact for geminates and a flatter shape of the tongue in geminate articulation [6].

Moreover, some articulatory studies showed that, in languages that allow word-initial voiceless geminate stops, even though the geminate contrast is neutralized perceptually and acoustically (in cases where no significant secondary cues to gemination exist), articulatorily, the distinction still holds in languages such as Tashlhiyt Berber [7] and Swiss German [8, 9]. It is shown in these articulatory studies that word-initial geminates were systematically longer in their articulation than their singleton counterparts. Another finding of these studies was that stops were longer in phrase-initial position than in phrase-medial position, which was interpreted as prosodic lengthening and strengthening in cases whereby it involved more linguopalatal contact. [6] also reported that gemination in Italian involved longer consonant duration as well as greater linguopalatal contact and the use of different regions of articulators. Based on these findings, [6] analysed gemination in Italian as a fortitional (i.e., strengthening) process.

Although these geminate articulatory correlates have been investigated for some languages, they have not been investigated for Arabic. All consonant phonemes in TLA have geminate counterparts which can occur both word-medially and word-finally with the exception of the glottal stop /ʔ/ ‘hamza’ which can only be geminated word-medially. Geminate consonants in TLA are contrastive, very frequent, and play an important role in the grammar of the language. However, very little is known about the phonetic realisation of gemination in this dialect (see [3]), with the articulatory cues to the singleton-geminate contrast have not been investigated yet. The acoustic results of [10] show that the singletons and geminates sonorants in TLA have similar formant structure and intensity values, which indicates that gemination in TLA does not show strengthening effects. However, it will be interesting to investigate whether these results are mirrored at the articulatory level. That is, to find out whether the sonorant geminates in TLA show longer and/or greater linguopalatal contact despite the fact that they do not show strengthening by the acoustic parameters.

This study contributes to the literature on gemination and the literature on Arabic language (and TLA) by examining the articulatory correlates of the singleton-geminate contrast in TLA using Electropalatography (EPG). The drive for

investigating the articulatory gestures of the singleton-geminate contrast is to ascertain whether underlying differences are reflected in phonetic dissimilarity.

2. METHODOLOGY

2.1. Language variety and subject

TLA is a variety of Arabic spoken in the northwest region of Libya (a country in the Maghreb region of North Africa) known as Tripolitania (Trablus) province. Due to the difficulty in finding subjects willing to take part in this kind of research and, also, due to the limited financial resources, only one male native TLA speaker participated in this experiment. He was naive as to the purpose of the study. At the time of the experiment, he was 34 years old, and had no obvious speaking or hearing defects. He was born and lived in Tripoli. He had lived and been educated there until he got his first degree. He speaks a typical TLA dialect. He was monolingual during childhood and his parents do not speak languages other than TLA. He was a postgraduate student who speaks English as a second language, and he lived in West Yorkshire during the time of the recording. He had been in the UK for about four years.

2.2. EPG data collection and recording

The system used was WinEPG (Articulate Instruments Ltd.). The model of the custom-made electro-palate was that of the Articulate-style. In this system there are eight rows of electrodes in the electro-palate, six electrodes on the first row (the one behind the front teeth) and eight on each of the other seven rows, forming a total of 62 electrodes. These sensors are spread between the point behind the upper front teeth and the back of the hard palate. The audio signal was sampled at 22.05 KHz and the EPG sampling rate was 100 frames per second.

The speaker was asked to read a list composed of trisyllabic minimal or near minimal word pairs containing the sonorant sounds /l, n, r/ in medial intervocalic position at normal speech rate. The test tokens were randomized and filler words were inserted after each 2-3 utterances. Equal number of singleton and geminate contrasts of these sounds were included. The carrier sentence was “ma tguli:ʃ _____ ta:ni” ‘Don't say (fm)_____ again’. Table 1 shows examples of the test tokens.

Singleton	Gloss	Geminate	Gloss
/lili:na/	‘for Lina’	/fil:ina/	‘cork’
/lina:ʒi/	‘for Naji’	/bin:a:ʒi/	‘Surname’
/liri:ma/	‘for Rima’	/bir:i:ma/	‘valve’

Table 1: Examples of test tokens.

2.3. Data analysis and measurements

The analysis of articulatory data was done with the Articulate assistant software (version 1.18). The EPG data was annotated using the analysis task window of the software, in which the waveform display, spectrogram display and EPG palate display are presented simultaneously. The constriction area for the alveolar region was defined as the first three rows from the front of the palate (R1-3). The research presented in this paper is part of a bigger project examining gemination in TLA. Many articulatory parameters were measured for the singleton and geminate sonorants, however only two subsets of the results are presented here.

3. RESULTS

3.1. The location, size, and shape of contact

The location, size, and shape of contact are presented graphically using representative palatograms for each sound combination separately. The contact patterns can provide information about the location of contact on the passive articulator and can be suggestive of the involvement of the active articulator. The depth of the constriction was taken into account as well in order to extrapolate tongue configurations (see [6]).

Figure 1 presents the palatograms for the alveolar lateral /l/. It is clear that the geminate laterals show more linguopalatal contact than singletons at rows 2 and 3. They also show somewhat deeper contact (i.e., go further back in the mouth – in this case, extends to row 4) than singletons, something that could be

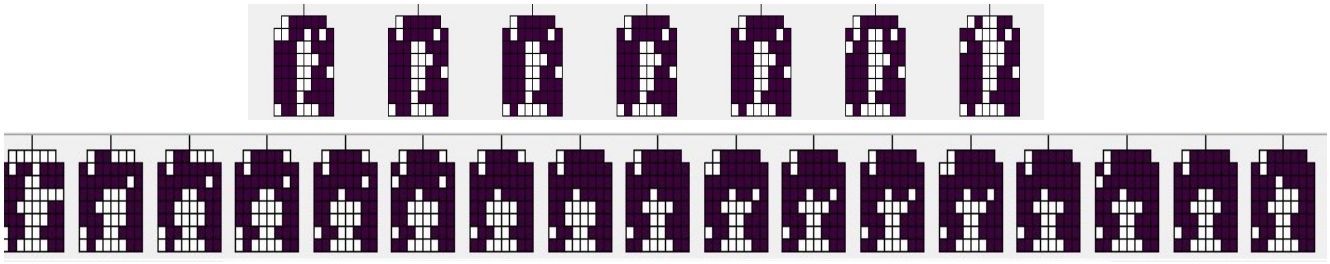


Figure 1: EPG palatograms of the singleton (top) and geminates (bottom) /l/.

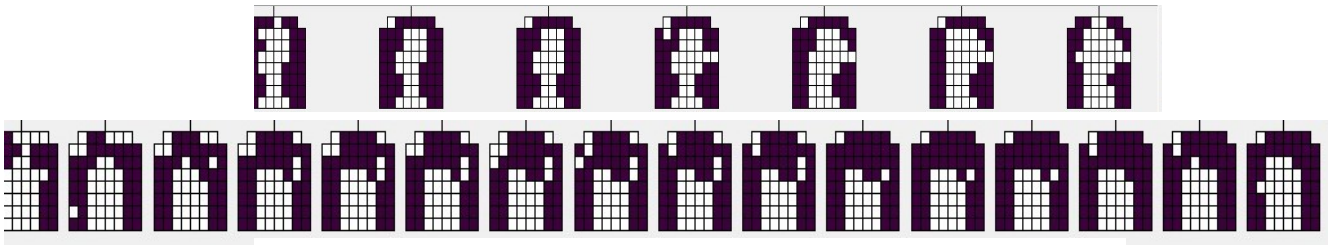


Figure 2: EPG palatograms of the singleton (top) and geminates (bottom) /n/.

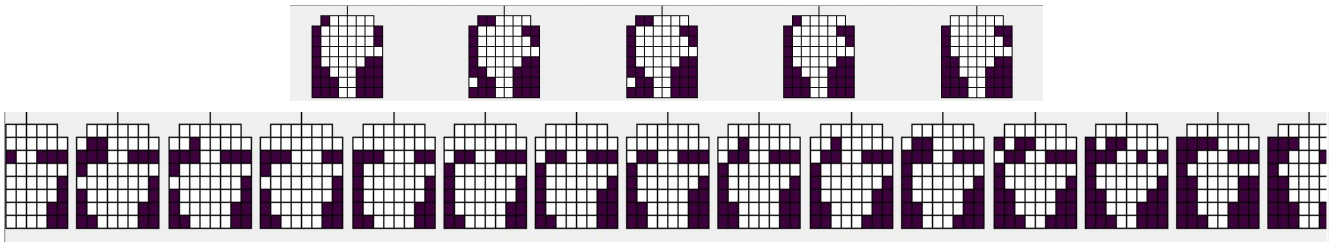


Figure 3: EPG palatograms of the singleton (top) and geminates (bottom) /r/.

interpreted as a more laminal contribution to the articulation of geminates and an apical production of singleton.

The differences between the singleton and geminate palatograms in the case of the alveolar nasal /n/ are presented in Figure 2. Geminates show more linguopalatal contact than singletons. Regarding the location of the contact, for singletons, the main contact lay along the first row indicating an apical articulation; in the case of geminates, the occlusion occurred in the first three rows with more linguopalatal contact at rows 2-3, something that could be taken to indicate an apico-laminal articulation. The firm contact at the two lateral columns for both singleton and geminate nasals could correspond to a 'cupped' (concave) tongue configuration. This shape of contact is compatible with the apical and apico-laminal interpretation of the contact patterns.

Figure 3 presents the palatograms for the alveolar rhotic /r/. Geminates show more linguopalatal contact than singletons. The main contact lies in the first row for the singletons and in third row for the geminate. This indicates that the /r/ is produced with an apical tongue configuration in all cases. However, the tongue is fronted in the case of singletons and retracted in the case of geminates. That is, as an apical anterior alveolar for singletons vs. apical posterior alveolar for geminates.

3.2. Flatness index

Flatness of the tongue, which is the mean contact divided by the maximum possible contact of the relevant region (see [11]), was measured for the singleton and the geminate consonants. If the contact profile is relatively flat, then the contact remains near maximum for more of the duration of the consonant and, therefore, this index will be closer to 1. The assumption is that gestures that maintain a constriction will be flatter than those that form a closure that is quickly released. The index of flatness was measured to investigate whether geminates are produced with flatter tongue configurations than singletons.

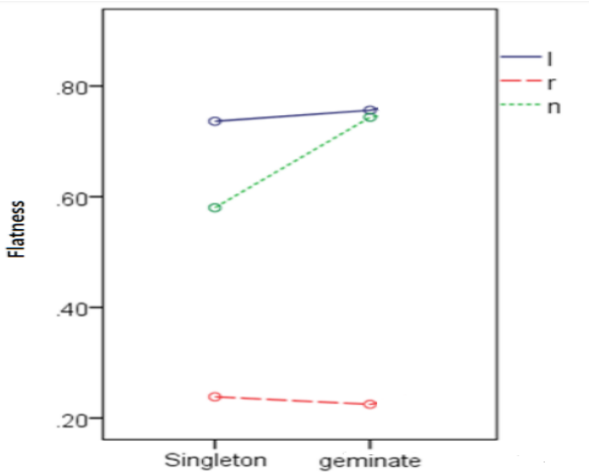


Figure 4: The flatness degree for the singleton and geminate sounds.

Figure 4 shows the flatness of the tongue degree for the singleton and geminate sounds. It is clear from Figure 4 that there is difference in the degree of flatness between the three sounds tested in this study with the alveolar rhotic /r/ showing the least degree of flatness compared to /n/ and /l/. The alveolar lateral /l/ show the highest degree of flatness only as a singleton. The alveolar lateral /l/ and the alveolar nasal /n/ show similar degree of flatness as geminates and both show flatter tongue configurations as geminates than singletons. On the contrary, the alveolar rhotic /r/ does not show flatter tongue

configurations as a geminate. The results of the flatness of the tongue degree supports the visual observations of the frames presented above.

4. DISCUSSION

The visual inspection of the palatograms was used in this study to provide information about the location of contact on the passive articulator which can be suggestive of the involvement of the active articulator.

The location, size and shape of contact presented in the current study are found to be robust articulatory cues for the distinction between singleton and geminate consonants in TLA. This study provides evidence that the articulation of geminate consonants involves deeper contact than singletons, something that could be interpreted as a more laminal (in the case of /l/) or apico-laminal (in the case of /n/) contribution to the articulation of geminates and an apical production of singletons. This observation of the existence of apical contact for singletons as opposed to laminal contact for geminates is generally comparable to what has been found for Italian [6].

Although the alveolar rhotic showed apical tongue configurations both as a singleton and a geminate, it was produced as an apical anterior alveolar for singletons but apical posterior alveolar for geminates. This difference in the articulatory configurations contributes to the singleton-geminate distinction of this sound.

The palatograms presented in Figures 1, 2, and 3 show longer articulatory duration for geminates compared to singletons. The articulatory gestures of geminate consonants show more contact frames than those of the singletons reflecting an increased articulatory duration. Geminates also show more linguopalatal contact than singletons. The number of the activated electrodes over the course of the production of the target sounds is higher in the case of geminates. This observation is generally comparable with what has been found for Italian nasals, laterals and stops by Payne [6] with the geminates involving more linguopalatal contact than singletons.

As regarding the flatness of the tongue index, the results of the current study provide clear evidence that the tongue is flatter during the production of geminates. The degree of flatness among the sound categories investigated here is different, however. The /l/ shows the highest degree of tongue flatness followed by /n/ and then /r/. The results of the flatness of the tongue among the three sound categories matches the extrapolated observations of the location, size and shape of contact.

As mentioned in in the introduction section above, the results of longer articulatory duration together with the greater amount of linguopalatal contact for geminates have been interpreted as lengthening and strengthening for some languages. The general assumption is that if geminates involve a stronger articulation, then the amount of contact with which it is produced is expected to show positive correlation. In the current study, and as mentioned above, geminates are found to involve more linguopalatal contact than singletons. However, whether to consider this as a strengthening effect for TLA geminates may not be straightforward since the behaviour of sonorant consonants may be difficult to understand with respect to consonant strengthening. That is, they are inherently quite vowel-like and thus might weaken or strengthen by becoming less like a vowel. As suggested by [12], weak sonorants will resemble vowels more closely in terms of formant structure and intensity whereas strong sonorants will have less formant structure and less intensity. In the current study, geminates are found to have more linguopalatal contact than singletons. However, the results of [10] for TLA sonorants show that the singleton and geminate consonants have similar formant structure and intensity values. Based on the acoustic evidence alone, there seem to be no indication of strengthening effects for gemination in TLA. However, and by considering the current results, it could be that the strengthening effects for TLA sonorants are evident only on the articulatory level.

In conclusion, the articulatory results of the current study show systematic articulatory differences between the singleton and geminate consonants in TLA. These results show that underlying differences are reflected in articulatory dissimilarity. The data provide clear evidence of the use of different regions of articulators for gemination. The data provide evidence for apical contact for singletons as opposed to laminal contact for geminates. The data also show a flatter shape of the tongue in geminate articulation. Furthermore, there is evidence of longer articulatory duration and greater linguopalatal contact for geminates. These results are suggestive of a strengthening effect on the articulatory level. The location, size and shape of linguopalatal contact together with the flatness of the tongue degree provide clear evidence that singletons are dynamically distinct from geminates in TLA. These results confirm the robustness of the effect of the phonological context on the articulatory gestural and temporal plans of the singleton and geminate consonants.

5. REFERENCES

[1] Khattab, G. and Al-Tamimi, J. 2008. "Durational cues for gemination in Lebanese Arabic". *Language and Linguistics*, 22, 39-55.

[2] Arvaniti, A. 1999. "Effects of speaking rate on the timing of single and geminate sonorants". In *Proceedings of the XIVth International Congress of Phonetic Sciences*. San Francisco, CA. pp. 599-602.

[3] Issa, A. 2015. "On the phonetic variation of intervocalic geminates in Libyan Arabic," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK.

[4] Local, J. and Simpson, A. 1988. "The domain of gemination in Malayalam," In: D. Bradley., E. J. A. Henderson., and M. Mazaudon, (Eds). *Prosodic Analysis and Asian Linguistics*: To Honour R. k. Sprigg. Pacific Linguistics, C-104. pp. 33-42,

[5] Payne, E. M. 2005. "Phonetic variation in Italian consonant gemination," *Journal of the International Phonetic Association*, 35(2), pp153-189.

[6] Payne, E. M. 2006. "Non-durational indices in Italian geminate consonants", *Journal of the International Phonetic Association*, 36(1), pp. 83-95.

[7] Ridouane, R. 2007. "Gemination in Tashlhiyt Berber: an acoustic and articulatory study," *Journal of the International Phonetic Association*, 37(2), pp119-142.

[8] Kraehenmann, A. and Lahiri, A. 2007. "Non-neutralizing quantity in word-initial consonants: artriculatory evidence", In J. Trouvain and W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (Saarbrücken, 6–10 August 2007), pp. 465–468.

[9] Kraehenmann, A. and Lahiri, A. 2008. "Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops", *Journal of the Acoustical Society of America,* 123(6), pp. 4446–4455.

[10] Issa, A. 2017. "Acoustic cues to the singleton-geminate contrast: the case of Libyan Arabic sonorants". *Proceedings of the 18th conference of the international speech communication – Interspeech* 2017, Stockholm, Sweden, pp. 2988-2992.

[11] Byrd, D., Flemming, E., Mueller C. A., and Tan, C. C. 1995. "Using regions and indices in EPG data reduction", *Journal of Speech and Hearing Research*, 38, pp. 821-821.

[12] Lavoie, L. 2001. *Consonant strength: phonological patterns and phonetic manifestations*. Garland Publishing, Inc,

INTERGESTURAL CV TIMING OF HOMOPHONOUS WORDS WITH DIFFERENT MORPHOLOGICAL STRUCTURES: A CASE STUDY OF LIQUID /L/ IN KOREAN

Jiyoung Lee[1], Sahyang Kim[2], and Taehong Cho[1]

[1]*Hanyang Institute for Phonetics & Cognitive Sciences of Language (HIPCS), Hanyang University,* [2]*Hongik University*
Ljy1004@hanyang.ac.kr, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This EMA study investigated whether and how underlying versus resyllabified onset /l/s in Korean may be differentiated, by examining homophonous sequences varying in morphological structures ($C_1V_1\underline{C_2V_2}$ vs. $C_1V_1\underline{C_2}$+$\underline{V_2}$ where C_2 is /l/). In Korean, /l/ is produced as [l] in the coda, but as [ɾ] in the onset. But in $C_1V_1\underline{C_2}$+$\underline{V_2}$, /l/ (in C_2) is resyllabified as an onset where it is produced as [ɾ], predicting no surface difference between the underlying and resyllabified onset conditions. Results showed that while /l/ was produced as [ɾ] in both conditions, the structural differences were reflected in temporal dimensions. [ɾ] was longer in the resyllabified versus underlying onset condition. And intergestural CV timing showed less overlap in the resyllabified versus underlying onset condition, indicating greater gestural cohesiveness of CV in the latter case. These differences were further augmented under prosodic prominence, reflecting speakers' deliberate effort to encode and maintain the morphological structure difference.

Keywords: intergestural timing, morphological structure, prosodic structure, Korean liquids, Articulatory Phonology

1. INTRODUCTION

The Korean liquid phoneme 'ㄹ', /l/, is renowned for its wide allophonic variations [1, 2]. The occurrence of an allophone of Korean /l/ is often determined by its position in a syllable. It is produced as a lateral (light /l/) in the coda position, and as a flap in the onset or intervocalic position [19, 20, 28]. Although Korean listeners perceive them as a single phoneme, the allophones are known to have distinct acoustic and articulatory features. Acoustically, a lateral variant is longer in duration than a flap variant [2]. In the articulatory dimension, the realization of /l/ is closely related to the adjacent vocalic gestures [2, 14, 19, 20]; a lateral is realized as a raising of the tongue tip during the preceding vowel while a flap is featured as a ballistic and short tongue tip movement between vowels [17].

The current study investigates whether and how the same allophone of /l/, occurring in different morpho-phonological environments, would be distinguished in its articulatory manifestation. This question is examined by comparing homophonous sequences of the resyllabified intervocalic /l/, which is underlyingly a coda as is produced in isolation (e.g., /pal + i/, where '+' is a morpheme boundary), and the underlying onset intervocalic /l/ (e.g., /pali/). In the former, the underlying lateral coda would be resyllabified in the intervocalic environment created by a morphological concatenation and thus produced as a flap [paɾi] (= a derived flap condition); and in the latter, with /l/ in an underlying onset intervocalic environment, the word would also be produced as a flap as in [paɾi] (= an underlying flap condition). At the surface phonetic level, therefore, in both cases, /l/ is produced as a flap, despite having different syllable structures due to underlying morphophonological compositions as in /pal + i/ and /pali/.

Previous studies allow us to make competing predictions regarding the production of the derived versus underlying flap. On one hand, studies on speech planning have suggested that the spelled-out segments go through the same process in creating the word's syllabified form, and thus the speech output may have little chance to reflect its internal morphophonological or syllable structure [22, 23]. If this is the case, no articulatory differences would be observed between the two homophonous sequences. On the other hand, the theory of Articulatory Phonology [4, 5, 6, 7, 15] assumes that the gestural coordination, specified as phase relations within a syllable, is stored in the lexicon. According to their assumption, the sequence of CV shows an in-phase relationship (i.e., the C and V gestures more or less start synchronously). The underlying flap in our study would show an in-phase relationship specified in the lexicon whereas the resyllabified flap may show a different pattern as their relationship is not specified in the lexicon (as C and V in our study do not belong to the same word). In addition, previous articulatory studies have provided some evidence that the gestures are coordinated differently depending on their underlying structures [10, 18]. Furthermore, some acoustic studies have also demonstrated differences in homophonous sequences [24]. Therefore, these studies allow us to predict some differences between the surface flap in the two different underlying morphological (and syllable) structures, particularly in terms of their gestural coordination.

This study also examines how the articulatory realization of the resyllabified and the underlying flap are further modulated by prosodic structure reflected in prosodic boundary strength and prominence. Prosodic strengthening via prominence or prosodic position is well known to maximize lexical contrast or to enhance the gestural bonding relationship [9, 11, 12, 13]. The prosodic modulation on the articulation of the allophone, therefore, will further shed light on whether and how speakers encode the underlying structural differences of the surface identical allophones.

2. METHOD

2.1. Participants

Twelve native speakers of Seoul Korean (6F, 6M in their 20s) participated in the experiment and were paid for their participation. In this paper, a subset of the collected articulatory data was analyzed first (4F, 3M) and the remaining data is still under analysis.

2.2. Speech materials and procedures

Two monomorphemic CVC-CVCV word pairs were used as target words, in which /l/ is placed in the coda or in the intervocalic onset position as in /pal/ 'foot' - /pali/ 'bowl' and /mal/ 'horse' - /mali/ 'Mari, *a proper name*'. They were followed by a grammatical particle /(i)lago/, of which the function is to directly cite the preceding word(s). Note that /i/ in the particle is inserted only when it follows a closed syllable. The combination of a target word and the particle yielded two sets of identical segmental strings with different underlying syllable structures: /pal+ilago/-/pali+lago/ and /mal+ilago/-/mali+lago/ where '+' refers to a morpheme boundary, and the underlines (i.e., $\underline{C_1V_1C_2+V_2}$ and $\underline{C_1V_1.C_2V_2}$) indicate the target sequences.

As shown in Table 1, target words were inserted in carrier sentences where two prosodic factors, Boundary and Focus were manipulated. The Intonational Phrase initial (IP-initial) condition and the IP-internal word initial (Wd-initial) condition were distinguished by the presence or absence of a prosodic phrasal juncture (including a pause) between the target and the preceding word. For the focus condition, target words were contrasted in a carrier sentence (e.g., "Did you write **/pali/** or **/pal/**?"). For the unfocused condition, non-target words were contrasted (e.g., "Did **you** write /pal/ or did **that person** write /pal/?).

Participants were asked to read out each written sentence presented as a written text on a computer screen. To induce each intended prosodic context, typographical cues were employed. A comma and a space were used at the IP boundary. The Wd boundary condition employed the pronoun /wuli/, which establishes a possessive relationship with the following noun. The targets were presented immediately after the pronoun without any space in between (e.g., /wulipal/ 'ourFoot'). The focused words were highlighted in red and bold.

During the recording sessions, two trained Korean ToBI transcribers checked whether each token was produced as the intended prosodic rendition or not. 240 sentences were produced per participant (4-target * 2-boundary * 2-focus * 15-repetition). In total, 1680 tokens were collected from 7 speakers, and by excluding 45 tokens with unintended prosodic renditions, 1635 tokens (97.3%) were analyzed.

Articulatory data were collected using EMA (AG501, Carstens Electronics), and acoustic data were recorded using a Tascam US 4*4 audio interface and a SHURE VP88 microphone simultaneously with the articulatory data. Five sensors were attached to the primary articulators: tongue tip (TT), tongue body (TB), upper lip (UL), lower lip (LL), and the middle of the lower gumline (LG). In this paper, kinematic data from TT (for the consonantal /l/ gesture) and TB (for the vocalic gesture) movements were analyzed.

Table 1: Examples of the test sentences. Targets are underlined, and contrasted words are marked in bold. '#' and '+' refer to phrase boundary and morphological boundary, respectively.

Conditions		Test sentences
#=IP	Foc	/ʧikimjʌki, # **pali**+lago s*ʌnni, # **pal**+ilago s*ʌnni/ Right here, did you write a **bowl** or a **foot**?
	Unf	/ʧikimjʌki, # pal+ilago **nika** s*ʌnni, # pal+ilago **ʧjeka** s* ʌnni/ Right here, did **you** write a foot or did **that person** write a foot?
#=Wd	Foc	/ʧikimjʌki, wuli#**pali**+lago s*ʌnni, wuli#**pal**+ilago s*ʌnni/ Right here, did you write our **bowl** or our **foot**?
	Unf	/ʧikimjʌki, wuli#pal+ilago **nika** s*ʌnni, wuli#pal+ilago **ʧjeka** s* ʌnni/ Right here, did **you** write our foot or did **that person** write our foot?

2.3. Measurements and statistical analyses

A matlab-based software, MVIEW [30] was used to define kinematic landmarks such as the onset and target of each movement. The articulatory landmarks were determined on the basis of changes in the velocity profiles of the articulatory movements by using a 20% threshold criterion. [25, 27].

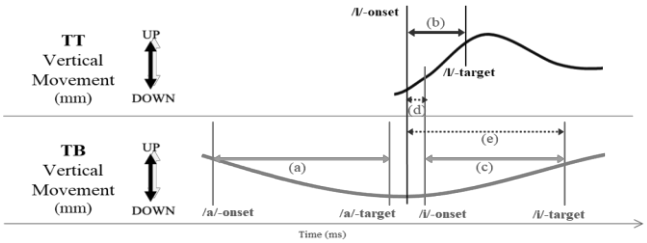


Figure 1: Schematized illustration of the tongue tip and tongue body vertical movements and the kinematic measurements: (a) V_1 duration, (b) C_2 duration, (c) V_2 duration, (d) time interval between C_2 and V_2 onsets, and (e) C_2V_2 duration

The duration of V_1, C_2, and V_2 was measured from the onset to the target of the gestural movement (Fig.1a~c). C_2V_2 duration (Fig. 1e) and the intergestural interval between the onset of C_2 /l/ and the onset of /i/ (Fig.1d) were also measured. For the intergestural time interval, a positive value indicates that the C_2 starts earlier than V_2 (hereafter "Earlier C_2") while a negative value shows that the C_2 starts later than V_2 (hereafter "Later C_2"). Since the actual degree of proximity between the onsets of C_2 and V_2 may reflect how much the two gestures occur synchronously, the absolute time intervals, as well as the Earlier C_2 and Later C_2 values, were analyzed.

A series of linear mixed-effects models were run separately for each measure with the lme4 package [13] in R. Fixed effects were Syllable Structure (CVC+V *vs.* CVCV), Focus (Focused *vs.* Unfocused), Boundary (IP-initial *vs.* Wd-initial), and all their 2-way and 3-way

interactions. All the factors were contrast-coded and the reference level of each factor is underlined above. The random structure included by-subject intercepts and slopes for all the fixed factors and their interactions. Nearly maximal models were fitted for each measurement as long as the model converged. When there was a case of non-convergence, a slope with the smallest variance was removed. When there were interactions, post-hoc pairwise comparisons were conducted with the emmeans package [21].

3. RESULTS

In what follows, only the results relevant to the research questions are reported (i.e., the main effects of Syllable Structure and its interaction with Focus and Boundary).

3.1. Durational analyses on individual gestures

There was a significant main effect of Syllable Structure on V_1 duration (β=6.28, t=2.06, p=.04; Fig.2a), and some marginal effect on C_2 duration (β=2.27, t=2.32, p=.059). Both V_1 and C_2 were longer in the underlying coda condition ($C_1V_1C_2$+V_2) than in the underlying onset condition ($C_1V_1.C_2V_2$). In other words, the vowel /a/ and the consonant /l/ gesture in the underlying $C_1\underline{V_1C_2}$ words were longer than the same phonemes in the $C_1\underline{V_1}.\underline{C_2}V_2$ words. For V_2 that follows /l/, there was no main effect of Syllable Structure.

There was a significant interaction between Syllable Structure and Focus on both C_2 (β=3.08, t=4, p<.001) and V_2 duration (β=2.67, t=2.47, p=.014). As can be seen in Fig.2b and Fig.2c, the interactions stem from the fact that the Syllable Structure effect (i.e., longer C_2 and V_2 in $C_1V_1\underline{C_2}$+$\underline{V_2}$ than in $C_1V_1.\underline{C_2V_2}$) was significant only under focus (C_2: β=3.81, t=3.629, p=.028; V_2: β=2.06, t=2.58, p=.019). The emergence of the difference is mainly due to the fact that the consonantal and vocalic gestures temporally expanded more for $C_1V_1\underline{C_2}$+$\underline{V_2}$ in the Focus condition as illustrated in Fig.2b and 2c. No other significant interactions were observed.

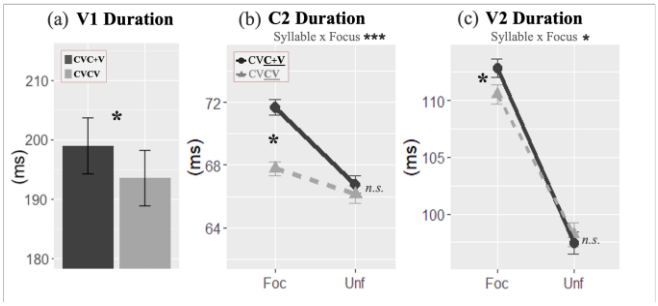


Figure 2: (a) Main effect of Syllable Structure on V_1 duration; (b) Syllable x Focus interaction on C_2 duration, (c) Syllable x Focus interaction on V_2 duration. Error bars represent standard errors. ('*' refers to p<.05, and '***' to p<.001)

3.2. Intergestural analyses

As for the intergestural time intervals between C_2 and V_2 onsets, there was a significant main effect of Syllable Structure on the absolute values and on the later C_2 values (Absolute, β=1.98, t=2.17, p=.031; Later C_2, β=-15.1, t=-2.59, p=.014). The intergestural intervals were longer for the resyllabified flap ($C_1V_1\underline{C_2}$+$\underline{V_2}$) than for the underlying flap ($C_1V_1.\underline{C_2V_2}$). That is, the consonantal and the following vocalic gestures were farther apart in the heteromorphemic than in the monomorphemic conditions.

There was no significant Syllable Structure by Focus interaction, but a marginal interaction was observed in the absolute time interval values (Absolute, β=3.43, t=1.87, p=.062). As shown in Fig 3a, this is presumably due to the fact that the temporal interval difference in terms of syllable structure was significant only in the focused condition, with longer interval in $C_1V_1\underline{C_2}$+$\underline{V_2}$ than in $C_1V_1.\underline{C_2V_2}$ (β=3.7, t=2.88, p=.009).

There were significant interactions between Syllable Structure and Boundary in Absolute and Later C_2 (Absolute, β=-4.22, t=-2.3, p=.022; Later C_2, β=29.48, t=2.87, p=.005). As can be seen in Fig.3b, the interactions seem to have come from the fact that the significant difference between the two syllable structures were observed only in Wd condition: Intergestural interval between C_2 and V_2 gestures were longer in $C_1V_1\underline{C_2}$+$\underline{V_2}$ than in $C_1V_1.\underline{C_2V_2}$ (Absolute, β=-4.09, t=3.1, p=.02; Later C_2, β=-29.84, t=-2.72, p=.04).

Finally, as for C_2V_2 duration, there was a significant Syllable Structure (longer in $C_1V_1\underline{C_2}$+$\underline{V_2}$ than in $C_1V_1.\underline{C_2V_2}$; β=1.96, t=2.21, p=.05), and this effect further interacted with Focus (β=5.8, t=3.94, p<.001). As shown in Fig 3c, the interaction stems from the fact that Syllable Structure difference was significant only under focus (β=18.21, t=4.18, p=.021).

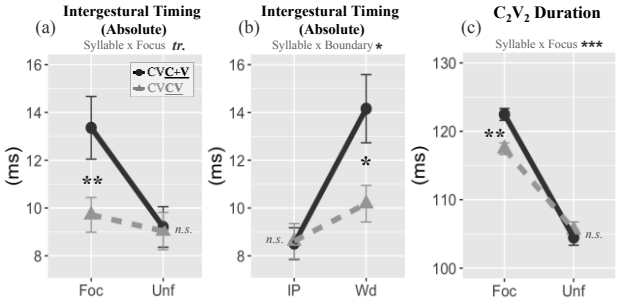


Figure 3: (a) Syllable x Boundary interaction on the intergestural timing for the Absolute condition; (b) Syllable x Boundary interaction on the intergestural timing for the Later C_2 condition; (c) Syllable x Focus interaction on C_2V_2 duration. Error bars represent standard errors ('*' refers to p<.05, '**' to p<.01, and '***' to p<.0001).

4. DISCUSSION

The present study examined kinematic characteristics related to the Korean /l/ in order to understand whether and how its allophone [ɾ], occurring in two different morphophonological environments, is articulatorily

manifested and further modulated by higher order prosodic structure. Comparisons were made between the resyllabified flap derived from the underlying coda from a heteromorphemic environment as in /pal + i/ and the underlying intervocalic flap from a monomorphemic environment in /pali/, both of which are produced as [pari] on the surface. The articulatory durations of the consonantal (i.e., /l/) and the surrounding vocalic gestures (i.e., /a/ and /i/) were measured and the intergestural time interval between the flap and the following vowel /i/ was calculated.

One of the basic findings was that the articulatory duration of the tongue body movement for /a/ was longer in the underlying heteromorphemic $C_1\underline{\mathbf{V_1}}C_2$+$V_2$ structure than in the underlying $C_1\underline{\mathbf{V_1}}C_2V_2$ monomorphemic environment, and this difference was not further modulated by prominence or neighbouring boundary strength. The result suggests that the intrinsic temporal difference between the monosyllabic and disyllabic words, which must be encoded in the lexicon, arises on the surface in the articulatory dimension.

Moreover, the tongue tip gesture for the target liquid showed a subtle but significant temporal difference as a function of morphophonological structure, but only under focus. That is, the derived vs the underlying flap did not differ from each other at least in the temporal dimension when they were produced in the unfocused environment, yet the morphophonological difference emerged articulatorily when the two structures were contrasted under focus. Note that the same kind of focus-induced modulation was observed in the tongue body gesture for the following /i/ vowel, and in the intergestural duration measured from the onset of the tongue tip gesture to the offset of the tongue body gesture. All these temporal adjustments were mainly manifested by more lengthening in the heteromorphemic sequence under focus. The results therefore seem to suggest that speakers engage in a deliberate effort to make a distinction between the two homophonous sequences, referring to the underlying structures, when the contrast is called for by the information structure. The results therefore support the views, in line with previous studies, that the lexical contrast is maximized in a fine-phonetic detail under prominence [cf. 9, 11, 12, 13], and that the gestures are coordinated differently depending on their underlying structures [10, 18, 29].

Another important finding of the present study is that the morphophonological structural difference between the two homophonous sequences is clearly reflected in the intergestural timing between the consonantal and the following vocalic gesture. The fact that the absolute distance between the two gestures were shorter for the monomorphemic $C_1V_1.C_2V_2$ than for the heteromorphemic $C_1V_1C_2$+V_2 suggests that the gestural cohesiveness is stronger when the two gestures belong to the same syllable/morpheme underlyingly. Note also that such a structure-sensitive temporal difference in the absolute term was found significant only under focus. This can be accounted for within the Articulatory Phonology, which views that the gestural coordination is specified within the lexicon [4, 5, 6, 7]. C_2 and V_2 in the $C_1V_1.C_2V_2$ belong to the same word and form an underlying CV syllable. The smaller temporal interval between the onsets of C_2 and V_2 can be deemed as revealing a stronger in-phase relationship specified in the lexicon. In contrast, C_2 and V_2 in the $C_1V_1.C_2$+V_2 do not belong to the same word, and hence their timing relationship is not intrinsically specified in the lexicon. The farther distance between the two gestures under focus therefore appears to reflect a relatively loose relationship in their gestural timing as the two gestures do not belong to the same word. Their yet enhanced distance under focus seems to provide another piece of evidence for the speakers' effort to maximize the structural difference when on demand.

The present study also found that the intergestural timing difference due to the syllable structure (stemming from morphological structure) was further modulated by the strength of prosodic boundary at which the target-bearing word occurs. The aforementioned intergestural timing difference between the resyllabified and the underlying CV structure (i.e., longer in the former than in the latter) was found significant only in the phrase-medial (Word) condition, not in the IP-initial condition. That is, at the IP-initial position, despite the fact that C_2V_2 were not strictly at the edge of the prosodic phrase (forming the second syllable of the IP-initial word), they showed smaller intergestural timing. This indicates a strong bonding relationship between C_2 and V_2 in the IP-initial position, regardless of whether the flap is derived or underlying. On the other hand, the gestural bonding appears to be loose in the medial condition. This result is in line with the previous studies showing a strong gestural bonding at a larger boundary, possibly reflecting domain-initial strengthening [8, 12].

In conclusion, the current study showed that the flap allophone of Korean /l/ occurring in homophonous surface sequences from two different underlying structures (i.e., a resyllabified flap and an underlying flap), can be articulatorily distinguished and their difference particularly shows up under focus. The results take the support away from the studies arguing that the underlying representations (e.g., morphological differences) are dimmed through the resyllabification process (i.e., during the phonological encoding stage) by making the adjacent words spelled-out segments "on the fly" (i.e., phonological syllable) [22, 23]. Rather, the present study provides evidence that speakers fine-tune the articulatory realization of gestures and their coordination depending on the underlying morphological structure and a higher-order prosodic structural demand.

5. ACKNOWLEDGMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

6. REFERENCES

[1] Ahn, S. W. (2002). The electropalatographic evidence of the Korean flap: An intervocalic Korean liquid sound. Speech Sciences, 9(3), 155-168.

[2] Ahn, S. W. (2006). An Electropalatographic Study of English 1, r, Flap and the Korean Liquid ㄹ. Korean Journal of English Language and Linguistics, 6(1), 113-133.

[3] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

[4] Browman, C.P., & Goldstein, L. (1988). Some notes on syllable structure in Articulatory Phonology. Phonetica, 45, 140–155.

[5] Browman, C. P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. Papers in laboratory phonology I: Between the grammar and physics of speech, 341-376.

[6] Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica, 49(3-4), 155-180.

[7] Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. Producing speech: Contemporary issues, 19-33.

[8] Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. Journal of phonetics, 29(2), 155-190..

[9] Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. Journal of Phonetics, 39(3), 344-361.

[10] Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. Phonetica, 58(3), 129-162.

[11] Cho, T. (2005) Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. Journal of the Acoustical Society of America

[12] Cho, T., Yoon, Y., & Kim, S. (2014). Effects of prosodic boundary and syllable structure on the temporal realization of CV gestures. Journal of Phonetics, 44, 96-109.

[13] de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. The journal of the acoustical society of America, 97(1), 491-504.

[14] Hwang, Y., Lulich, S. M., & De Jong, K. J. (2019). Articulatory and acoustic characteristics of the Korean and English word-final laterals produced by Korean female learners of American English. The Journal of the Acoustical Society of America, 146(5), EL444-EL450.

[15] Iskarous, K., & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. Journal of Phonetics, 95, 101195.

[16] Krakow, R. A. (1999). Physiological organization of syllables: a review. Journal of Phonetics, 27(1), 23-54.

[17] Ladefoged, P. & Maddieson, I. (1996). The sounds of the world's languages. Malden, MA (USA): Blackwell Publishing.

[18] Lee, J., Kim, S., & Cho, T. (2019). Effects of morphological structure on intergestural timing in different prosodic-structural contexts in Korean. In Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia.

[19] Lee, S. & Kang, S. (2003). Acoustic and phonological properties of Korean liquids. Journal of Language Sciences, 10(2), 77-92.

[20] Lee, Y. J., Goldstein, L., & Narayanan, S. S. (2015). Systematic variation in the articulation of the Korean liquid across prosodic positions. In ICPhS.

[21] Lenth, R. V. (2022). Emmeans: Estimated Marginal Means, aka Least-Squares Means, Version 1.8. 1-1.

[22] Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and brain sciences, 22(1), 1-38.

[23] Levelt, W. J. (1993). Speaking: From intention to articulation. MIT press.

[24] Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English1. Journal of Linguistics, 53(1), 181-216.

[25] Pouplier, M., & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. Language and Cognitive Processes, 25(5), 616-649.

[26] Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. Journal of psycholinguistic research, 25(2), 193-247.

[27] Shaw, J., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. Phonology, 26(1), 187-215.

[28] Shin, J. (2015). The sounds of Korean. In Brown, L., & Yeon, J. (Eds.). The handbook of Korean linguistics (Chapter 4). John Wiley & Sons.

[29] Son, M. (2011). The Projection of Syllable Structure: A Case Study of Intervocalic/k/ in Korean. The Korean Journal of Linguistics, 36(2), 395-414.

[30] Tiede, M. (2010). MVIEW: Multi-channel visualization application for displaying dynamic sensor movements. development.

PERCEPTION OF TASHLHIYT CONSONANT QUANTITY CONTRASTS BY NATIVE VS. NONNATIVE LISTENERS FROM THREE LANGUAGES

Pierre Hallé[a,b], Philipp Buech[a], Yueh-chin Chang[c], Feng-fan Hsieh[c], Jiayin Gao[a], and Rachid Ridouane[a]

[a] LPP (CNRS/Sorbonne Nouvelle, Paris), [b] LMC2lab (Paris Cité), [c] Tsing Hua University (Hsinchu, Taiwan)
pierre.halle@sorbonne-nouvelle.fr, philipp.buech@sorbonne-nouvelle.fr, ycchang@mx.nthu.edu.tw,
ffhsieh@mx.nthu.edu.tw, jiayin.gao@cnrs.fr, rachid.ridouane@sorbonne-nouvelle.fr

ABSTRACT

Remarkably, Tashlhiyt has gemination contrasts for all consonants in all word positions. Perception of these contrasts by natives vs. nonnatives is little documented, with the exception of [Hallé, Ridouane, and Best, 2016], which found that natives performed near ceiling on all contrasts under scrutiny, whereas nonnative French listeners were much better on word-final than -initial contrasts. The authors tentatively proposed that the word-initial and word-final positions engage different perceptual processes. We re-examine these findings using a more comprehensive design including initial, medial, and final positions, using Tashlhiyt native listeners and nonnative listeners from three languages differing in terms of quantitative and qualitative incidence of gemination: Chinese, French, and Japanese.

Overall, easy perception of non-initial Tashlhiyt geminate-singleton contrasts by nonnative listeners is confirmed, supporting the earlier proposition that they engage universal perceptual processes. In contrast, word-initial contrasts would require language-specific learning.

Keywords: geminate consonants, within-word position, Tashlhiyt, cross-linguistic speech perception

1. INTRODUCTION

The perception of geminate vs. single consonants by listeners whose language has no geminates has been mainly studied in the framework of second language acquisition, most often using stimuli from a gemination language and nonnative learners vs. control native listeners. Japanese and Italian are the most popular gemination languages used in these studies. Worth to note, in these two languages, lexical geminates are restricted to the word-medial position. One common finding across studies is that no-gemination language listeners learning a gemination language have substantial difficulty, especially beginners, at perceiving word-medial geminates in identification or discrimination tasks (English learners: [1-6]; Korean learners: [7-9]; Chinese learners: [10-13]). [6] tested American listeners of four different learning levels of Japanese (including naïve listeners) on a same-different discrimination test and found error rates in the 40-30% range. Identification tasks seem to show better performances: in [8] (also see [9]), Korean learners of Japanese identify correctly Japanese singletons at the 64% level but geminates near ceiling level. Chinese listeners tend to show a reverse asymmetry [12]. More generally, perception of the Japanese geminate-singleton contrast differs according to the phonological and phonetic systems of the listener's native language.

A first cause of variation is the different extents to which languages allow gemination or not. Some languages, such as Lingala, have no gemination at all. Concatenated gemination occurs in Chinese dialects with checked syllables such as Cantonese (e.g., 拍佢, *pak koy* 'tap him'); it is restricted to /-n/#/n-/ in Mandarin. It is far more common in English or French, which allow many consonant codas. Korean "tense" stops are somewhat akin to geminates due to their long closure duration. Tashlhiyt is an extreme case of gemination incidence since it allows gemination in all utterance positions and for all consonants.

A second cause of variation is related to the possible secondary cues to gemination and how listeners exploit them. The primary cue –consonant duration– may be accompanied by prosodic cues such as flanking vowels' durations and/or pitch. [4] showed that consonant and preceding vowel durations interact in the perception of gemination in language-specific ways: a perceptual bias toward gemination is induced by preceding vowel *shortening* for Italian vs. preceding vowel *lengthening* for Japanese listeners, reflecting differences in how secondary cues to gemination are implemented in Italian versus Japanese.

In this study, we focused on the first cause of variation: the extent of gemination incidence in the listener's native language. Tashlhiyt fits this purpose quite well since its geminates are produced with primary durational cues, with little or no secondary cues to gemination, whether prosodic or phonetic-acoustic, as shown in [14-15]. That is, the differences that may be found across listeners will not likely be attributable to differences in the language-specific ways they exploit secondary cues to gemination in perception. [16] examined the discrimination of some Tashlhiyt geminate-singleton contrasts by French vs. Tashlhiyt listeners. The contrasts were

/b/-/bb/, /d/-/dd/, /g/-/gg/, /f/-/ff/, and /s/-/ss/ in word-initial position, and /t/-/tt/ in word-final position. Whereas Tashlhiyt listeners performed near ceiling level on all contrasts, French listeners did rather poorly on word-initial contrasts (in the 65-75% correct discrimination range) but did quite well on word-final /t/-/tt/ contrasts (~90% correct). From these rather scarce data, the authors nevertheless proposed that the good French performance on word-final /t/-/tt/ reflects a universal sensitivity to the "beat" given by successive salient phonetic events (here, between the vowel preceding /t, tt/ and the release of /t, tt/). In contrast, the perception of word-initial contrasts requires sensitivity to "sound substance duration," which would need language-specific learning. Anecdotal evidence supporting the authors' claim was drawn from a single word-medial contrast used in the training phase (*jutid-juttid*). To begin with, we believed that more data and a more complete design were needed before drawing reliable conclusions about the perceptual mechanisms at work when listening to geminate compared to single consonants. In particular, the design should cover all the three initial, medial, and final positions, different types of consonants, and use a more varied sample of languages with regards to gemination incidence. Moreover, since [16] was restricted to word-initial and -final positions, no comparison could be made with the data reported in the literature, which indeed massively deal with the medial position.

In the present study, we used a rather exhaustive design including the three within-word positions and four different target consonants /d, t, s, n/ covering major manners of articulation and voice settings. We recruited participant listeners from four different languages: Tashlhiyt as the control language, Mandarin (limited incidence of false gemination), French (higher incidence of false gemination arising from word contact or vowel deletion), and Japanese (gemination restricted to word-medial obstruents and /n/).

Several remarks are in order. First, word-initial /t/-/tt/ is very difficult to perceive in the absence of acoustic cues to duration, even for Tashlhiyt listeners [17]. We nevertheless included this contrast for sake of completeness. Second, regarding Japanese, the native "yamato" lexical stratum bans geminate voiced obstruents. However, established loanwords such as /eggu/ 'egg' or /beddo/ 'bed' are quite common [18]. Also, Japanese /nn/ may be special for gemination as it most often occurs as a concatenated geminate (in Sino-Japanese compounds or /-n/#/n-/ contacts). In Mandarin, /nn/ can only be a concatenated geminate, as in 山鳥 *shān niǎo* 'mountain bird'. Note that we recruited Mandarin-speaking subjects in Taiwan: they all had Mandarin as their main language but some of them also had a good command of the Hakka and/or Minnan dialects spoken in Taiwan, in which /p, t, k/ false geminates do occur, as in Cantonese. Since [12] reported an advantage of mainland China Cantonese- over Mandarin-speaking listeners in the perception of Japanese geminates, we collected detailed language background data on Taiwanese participants for a possible effect of their familiarity with Hakka or Minnan.

Keeping these points in mind, several predictions can be made. First, given the incidence and phonological status of gemination in the four languages under scrutiny, we expect the best performance for the Tashlhiyt control group, followed by the Japanese, French and Chinese groups in that order, with possible differences within the Chinese group. Weak differences across groups for the non-initial positions plus poor nonnative performance in initial position would support [16]. In more detail, Japanese listeners might have extra trouble with /d/-/dd/ and possibly /n/-/nn/, whereas Mandarin-dominant Taiwan listeners might find the latter contrast easier.

2. DISCRIMINATION EXPERIMENT

We tested the nonnative language groups (Chinese, French, and Japanese) and the control Tashlhiyt group on their discrimination of Tashlhiyt geminate-singleton contrasts, using the AXB paradigm.

2.1. Methods

2.1.1. Participants

Fifty-four Chinese-speaking students at Tsing Hua University in Hsinchu, Taiwan (aged 18-47 y, mean 23.1, SD 5.0; 21 males), 20 Japanese-speaking students at Sophia University in Tokyo (20-28 y, mean 22.4, SD 2.4; 5 males), 21 French-speaking students at Université Sorbonne Nouvelle in Paris (20-39 y, mean 25.1, SD 5.5; 5 males), and 26 Tashlhiyt-speaking students at Ibnou Zohr University in Agadir (18-53 y, mean 33.2, SD 12.4; 16 males) participated in the experiment voluntarily. Nonnative participants never had exposure to Tashlhiyt. French participants were naïve to any true-gemination language. Based on the collected metadata, Taiwan participants were either clearly Mandarin-dominant speakers (n=22), or with good (n=14) or moderate (n=16) practice of Minnan and/or Hakka. Japanese participants had not been exposed to other gemination language than Japanese. Participants reported no hearing deficit or any kind of language impairment.

2.1.2. Stimuli and design

Twelve geminate-singleton contrasts (4 consonants x 3 positions) of Tashlhiyt were used (/d/: *diɤ-ddiɤ,*

tidi-tiddi, fad-fadd; /t/: *tili-ttili, juti-jutti, jufat-jufatt*; /s/: *sir-ssir, tisi-tissi, ifis-ifiss*; /n/: *niʁ-nniʁ, inas-innas, imun-imunn*). This made a total of 24 words. Four repetitions of each item produced in isolation by a native speaker were retained as experimental stimuli. Tashlhiyt /d/ is entirely voiced (i.e., [d]) and differs from /dd/ by a longer voiced closure duration (geminate-to-singleton duration ratio ~2.5 in the recorded materials). The mean geminate-to-singleton duration ratios were 2.01, 2.25, and 2.53 for initial, medial, and final position, respectively. Closure could not be measured acoustically for word-initial /t/ and /tt/ but electropalatographic articulatory data show much longer closures for geminates than single consonants in absolute initial position [14].

As can be seen, the duration ratios increased from the initial to the final position. They were the highest for /n/ and /d/ and the lowest for /s/. Apart from consonant duration, gemination was not accompanied by *f0* or intensity cues in our stimuli (except the smaller intensity of prevoicing with /dd/ than /d/, as found in [16]). Yet, gemination was *accompanied* by consistent, however weak secondary duration cues in flanking vowels. The vowel preceding word-final consonants, was shorter before geminates than singletons (158 < 214 ms, p<.01); this trend was much weaker in word-medial position (94 < 108 ms, n.s.). A weak trend in the opposite direction was found in the following vowel (194 vs. 188 ms, n.s.).

Each contrast was presented 16 times (four times for each of the AAB, ABB, BBA, and BAA patterns), hence 192 trials divided in two blocks of 96 trials with the same distribution of the 12 contrasts. Each stimulus appeared equiprobably in each within-AXB position. The test phase was preceded by a training phase of eight trials on six contrasts –not used in the test trials– bearing on manner (*dar-daʁ),* voicing (*dar-tar, giji-kiji,* each used twice), or gemination (*fit-ffit, kiji-kkiji*, and *tafa-ttafa*).

2.1.3. Procedure

Nonnative participants were tested individually in a quiet room and received the speech stimuli through professional quality headphones. On each AXB trial, participants were presented with the stimuli and had to press one of the keys labeled “1” or “3” to indicate whether X was a better category match to the first (A) or third (B) stimulus. The inter-stimulus (offset to onset), inter-trial, and inter-block intervals were set to 1, 4, and 9 s, respectively. Response times (RTs) were measured from the onset of the B stimulus, with a 3 s time-out. The experiment was run by a custom-made Python program [19], using the Pygame extension for timing control and RT measurement. For the Tashlhiyt-native subjects, the experiment was run using an on-line PsyToolkit [20-21] script, which replicated the settings of the off-line Python program. All participants received feedback on accuracy, time-out, and RT in the training phase, and only on possible time-out in the test phase, in order to foster speed and attention.

2.2. Results

The data of two Taiwanese and two Tashlhiyt participants were discarded due to high miss-rates, leaving 52 Taiwanese, and 24 Tashlhiyt subjects for accuracy and RT data analyses. We first looked whether Taiwanese subjects should be partitioned into subgroups, based on their metadata: 22 subjects were clearly Mandarin-dominant (native mandarin-speaking parents, little or no practice of Minnan or Hakka), 14 had substantial practice of Minnan or Hakka (both parents native speakers of one of these dialects, high self-rating in hearing/speaking fluency), and 16 were in-between. The 14 Minnan- or Hakka-fluent subjects did not perform better than the 22 Mandarin-dominant ones, as [12] would suggest. On the contrary, their average proportion of correct discrimination was numerically lower (.75 vs. .77), and this held for most position × consonant conditions. We therefore treated the Chinese-speaking group as a single, unified group.

We ran analyses of variance on the unfiltered d-prime and correct-response RT data with the following fixed factors: Group (four language groups) as a between-subject factor, Position (three positions), Consonant (d, t, s, n), and Block (two blocks) as within-subject factors. All the contrasts were quite well perceived except word-initial /t/-/tt/, which was perceived at chance level by all four groups. We therefore present the data separately for word-initial /t/ and /d, s, n/ in Figs. 1 and 2. Block was marginal for d-prime, p=.055 (numerical trend for increased performance in block 2) but n.s. for RTs. The other factors were all significant, ps<.0001. More detail is given in the next sections. We begin with d-prime.

The four groups did not differ for word-initial /t/, F<1 (~55% correct); for word-initial /d, s, n/ only, the pattern was Tashlhiyt > French > Japanese ≈ Chinese, p<.0001, p<.0001, and F<1, respectively; for word-medial all four /d, t, s, n/, it was Tashlhiyt ≈ French > Japanese > Chinese, F<1, p<.05, p<.01, respectively; word-finally, it was Tashlhiyt ≈ French >> Japanese ≈ Chinese, F<1, p<.0001, F<1, respectively. For /d, s, n/ only, Tashlhiyt subjects performed near ceiling regardless of Position, which was not significant, p>.05. For all four /d, t, s, n/, the three nonnative groups’ performance differed with Position, following the pattern initial << medial ≥ final, p<.0001, p=.16, respectively. The same pattern

held for Tashlhiyt subjects on /t/ alone. Overall, conflating the medial and final positions, all four groups performed very well, with Tashlhiyt and French subjects near ceiling (~96%), somewhat better than Japanese and Chinese subjects (~90%). In contrast, the three nonnative groups performed quite poorly on word-initial contrasts, as shown in Fig. 1.

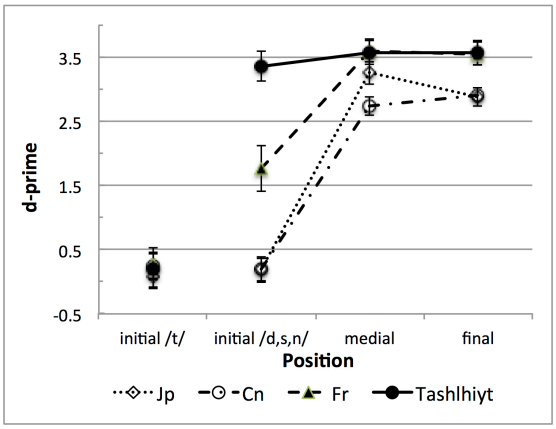


Figure 1: d-primes according to Position and Group, separately for /t/ and /d, s, n/ word-initially.

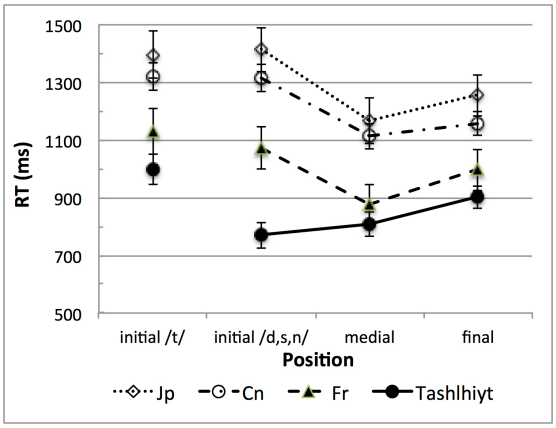


Figure 2: RTs according to Position and Group, separately for /t/ and /d, s, n/ word-initially.

The RT data paralleled well this overall picture. Tashlhiyt subjects were the fastest, $p<.0001$. In non-initial position, the order in RTs was Japanese ≥ Chinese >> French ≥ Tashlhiyt, $F<1$, $p<.0001$, $p=.23$, respectively. Word-initially, it was Japanese ≥ Chinese >> French > Tashlhiyt, $p=.23$, $p<.0001$, $p<.001$, respectively. Nonnative groups responded faster for non-initial than initial position (all four consonants) by ~180 ms, $p<.0001$, with the fastest RTs in medial position, $F(1,90)=99.3$, $p<.0001$. In contrast, Tashlhiyt subjects were faster on non-final (799 ms), than final (918 ms) positions, $p<.0001$ for /d, s, n/, and the slowest for word-initial /t/ (1000 ms vs. 809-902 ms non-initially), $p<.0005$.

In Japanese, /dd/ is nonnative and /nn/ is a weaker case of gemination, hence the prediction of a lesser performance on /d/-/dd/ or /n/-/nn/. This prediction is borne out only word-initially but clearly not in non-initial position (see Table 1). As for Taiwanese, they performed the worst on /n/-/nn/ although /nn/ is the sole concatenated geminate in Mandarin. Unpredictably, /s/-/ss/ was generally the best-perceived contrast word-initially but the worst perceived in non-initial positions.

consonant	initial	medial	final	means
d	**-0.24**	**3.84**	**3.40**	**2.33**
t	0.07	3.04	3.30	2.14
s	**1.11**	**2.82**	**1.89**	**1.94**
n	**-0.31**	**3.32**	**2.95**	**1.99**
means	0.16	3.26	2.88	2.10

Table 1: Japanese d' data: Consonant x Position.

3. DISCUSSION

Both the accuracy and RT data pointed to a consistent pattern: Nonnative listeners performed well on Tashlhiyt word-medial or -final gemination contrasts, but much less well on word-initial contrasts. (The case of word-initial /t/-/tt/ is special: it was perceived at chance level in all four groups of listeners and induced the longest RTs.) Therefore it seems that, in non-initial position, singleton-geminate contrasts are much easier than can be expected from the incidence of gemination in the listeners' languages: Japanese listeners did worse than French and not better than Mandarin-speaking Taiwanese. French listeners did as well as Tashlhiyt listeners. The data thus fit well with [16]: sensitivity to the "beat" given by successive salient events (e.g., the vowels surrounding a consonant) is likely universal; in contrast, sensitivity to word-initial consonant duration, where no beat can help, likely requires language-specific learning.

An alternative account may be provided by the secondary durational cues that accompany gemination in Tashlhiyt: vowels are shorter before geminate than singleton consonants, which should help perceiving gemination contrasts in non-initial position. Yet, this shortening is more marked for word-final than -medial geminates (26 vs. 12%), which runs counter to the perceptual advantage for medial over final gemination we observed. The "beat" account provides a simple explanation for this differential: the beat given by two vowels flanking a medial consonant must be more salient than that given by a preceding vowel and a following consonant-release because the latter phonetic event is conceivably perceptually weaker than a vowel is.

At any rate, further investigation of the potential contribution of vowel duration as a secondary cue to gemination is warranted as a follow-up study.

4. ACKNOWLEDGEMENTS

This study was supported by the French ANR (ANR- 10-LABX-0083, Labex-EFL).

5. REFERENCES

[1] Han, M. S. 1992. The timing control of geminate and single stop consonants in Japanese: A challenge for non-native speakers. *Phonetica* 49: 102–27.

[2] Harada, T. 2006. The acquisition of single and geminate stops by English-speaking children in a Japanese immersion program. *Studies in Second Language Acquisition* 28: 601–32.

[3] Toda, T. 2007. Issues regarding geminate consonants in Japanese language education. *Journal of the Phonetic Society of Japan* 11: 35–46.

[4] Kingston, J., Kawahara, S., Chambless, D., Mash, D., Brenner-Alsop, E. 2009. Contextual effects on the perception of duration. *J. Phon.* 37, 297–320.

[5] Hardison, D. M., Saigo, M. M. 2010. Development of perception of second language Japanese geminates: Role of duration, sonority, and segmentation strategy. *Applied Psycholinguistics*, 31, 81-99.

[6] Hayes, R. 2001. The perception of novel phoneme contrasts in a second language: A developmental study of native speakers of English learning Japanese singleton and geminate consonant contrasts. *Coyote Papers* 12, 28–41.

[7] Sonu, M., Kato, H., Tajima, K., Akahane-Yamada, R., Sagisaka, Y. 2013. Non-native perception and learning of the phonemic length contrast in spoken Japanese: Training Korean listeners using words with geminate and singleton phonemes. *Journal of East Asian Linguistics* 22: 373–98.

[8] Tsukada, K., Cox, F., Hajek, J., Hirata, Y. 2018. Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Language Research* 34, 179–200.

[9] Tsukada, K., Rong, Y., Kim, J. Y., Han, J. I., Hajek, J. 2021. Cross-Linguistic Perception of the Japanese Singleton/Geminate Contrast: Korean, Mandarin and Mongolian Compared. *Proc. Interspeech*, 3910-3914.

[10] Hung, H. Y. 2012. Perception of Japanese geminate stops among Taiwanese learners of Japanese. *Journal of the Phonetic Society of Japan* 16. 15–27.

[11] Lu, S., Chen, L., Yamada, J. 2016. Taiwanese learners' durational patterns for Japanese obstruent geminates/singletons and adjacent vowels. *Perceptual and Motor Skills* 122. 67–75.

[12] Ren, H. 2022. Perception of Japanese singleton and geminate contrasts: A case of Chinese learners with different dialectal backgrounds. *Front. Psychol.* 13, 1070107.

[13] Uchida, T. 1993. Characteristics of auditory cognition of long vowels and double consonants for Chinese students in learning Japanese language. *Japanese Journal of Educational Psychology* 4. 414–23.

[14] Ridouane, R. 2007. Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *J. Int. Phon. Assoc*. 37, 119–142.

[15] Ridouane, R. 2010. Geminate at the junction of phonetics and phonology. In: Fougeron, C., Kühnert, B., D'Imperio, M., Vallée, N. (eds), *Papers in Laboratory Phonology X*. Mouton de Gruyter, 61–90.

[16] Hallé, P., Ridouane, R., Best, C. 2016. Differential difficulties in perception of Tashlhiyt Berber consonant quantity contrasts by native Tashlhiyt listeners vs. Berber-naïve French listeners. *Front. Psychol*. 7, 209.

[17] Ridouane, R., Hallé, P. 2017. Word-initial geminates: from production to perception. In Haruo Kubozono (ed.), *Phonetics and Phonology of Geminate Consonants,* 66-84. Oxford Studies in Phonology and Phonetics - Oxford University Press.

[18] Kawahara, S. 2015. The phonetics of sokuon, or geminate obstruents. In: Kubozono, H. (ed.) *Handbook of Japanese phonetics and phonology*. De Gruyter Mouton, 43–78.

[19] Yu, S. 2018. Psypsy program. Github repository: https://github.com/shi4yu2/psypsy.

[20] Stoet, G. 2010. PsyToolkit: a software package for programming psychological experiments using Linux. *Behavior Research Methods* 42, 1096–1104.

[21] Stoet, G. 2017. PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44, 24–31.

INTRA- AND INTERSEGMENTAL DURATIONAL COMPENSATION OF KOREAN PLOSIVES

Yeongeun Choi

Department of Computational Linguistics, University of Zurich
yeongeun.choi@uzh.ch

ABSTRACT

This study investigates the durational characteristics of three types of word-initial Korean plosives (fortis, lenis, aspirated), with a focus on the correlations between *Closure duration* (CD), *Voice onset time* (VOT), and the following vowel. Read speech of 15 native Korean speakers was recorded. The results show that (1) the VOT of aspirated plosives is as short as that of lenis plosives (even shorter in the case of velars), (2) CD plays a more important role in distinguishing Korean plosives than VOT, and (3) the durations of CD and VOT are dependent on one another in a complementary relationship. (4) In general, plosive duration influences vowel length (the longer the plosive, the shorter the vowel); particularly, (5) VOT is closely correlated with vowel length across place and manner of articulation. In sum, CD and VOT of Korean plosives influence each other and partially also the duration of the following vowel.

Keywords: Korean plosives, Closure duration, Voice onset time, Vowel duration, Durational compensation

1. INTRODUCTION

Plosives (henceforth also called 'stops') are characterized by two articulatory phases, with closure and release occurring consecutively before the onset of the following vowel. The time intervals of stop closure, known as *Closure duration* (CD), and of stop release, known as *Voice onset time* (VOT), vary depending on plosive type and place of articulation. The variation in the durations of CD and VOT arises in a complementary relationship, i.e. a long CD induces a short VOT and *vice versa*, and results in a limitation of the total duration of vocal fold opening [19, 4]. The durational adjustment occurs not only within a plosive, but also with an adjacent vowel. In particular, previous studies on pre-consonantal vowels in different languages have provided evidence that vowel length is determined by the relative length of neighboring consonants, i.e. longer vowels shorten the VOT of the following stops, e.g. in varieties of English [25], and also in the Central Bavarian dialect of German [9]. Conversely, durational compensation between vowels and preceding consonants seems to be less common.

The three types of homorganic plosives in Korean are all voiceless in word-initial position and are differentiated by the amount of [aspiration] as follows: fortis /p*, t*, k*/[1] (unaspirated), lenis /p, t, k/ (slightly aspirated), and aspirated /p^h, t^h, k^h/ (strongly aspirated). These properties lead to the differences in VOT, fortis with short-lag, lenis with rather longer, namely 'intermediate'-lag, and aspirated with long-lag [18, 6]. However, more recent studies (e.g. [23, 11, 3]) have found an overlap, commonly known as 'merger' of VOT between lenis and aspirated plosives, especially in the Seoul dialect, which is generally considered as Standard Korean. Regarding CD duration in Korean word-initial plosives, fortis plosives have the longest CD with a large duration overlap with aspirated plosives, which have the second longest CD, while lenis plosives have the shortest duration [16, 5].

Korean has been posited to lose its distinctive vowel quantity. Although vowel length distinctions are still preserved in a few homonyms, native Korean speakers barely distinguish between long and short vowels [20, 21] (see also [12, 24]). Despite the loss of phonemic vowel length contrasts, vowels before or after plosives appear to vary according to allophonic distribution. An electromyographic study [10] found that a preceding plosive affects the duration of the neighboring vowel more than a following plosive does, in terms of the voicing gesture: the larger the glottal opening area of stops, the larger the degree of devoicing of the following vowel. For instance, vowel devoicing was rarely observed after word-initial fortis stops associated with a very small glottal opening area (see also [7]). The voicing gesture may trigger a reduction in the actual time interval of the post-consonantal vowel, but this aspect was not examined in the current study.

Based on these insights, we hypothesize that the two basic acoustic cues of Korean word-initial plosives (i.e. CD and VOT) adjust their durations in a compensatory behavior, and are also related to the duration of post-consonantal vowels. In order to verify these hypotheses and to gain a better understanding of the relative durational behavior in Korean stops, a carefully designed corpus of read speech was collected from 15 native speakers of Korean (cf. § 2).

2. METHOD

2.1. Participants

15 native Korean speakers were recruited (13 female, 2 male; aged 20-29 years, mean 22). All subjects speak the Seoul dialect and had lived abroad for less than one year.

2.2. Speech materials

A reading task was conducted with a list of Korean interrogative sentences, which all begin with the interrogative adverb *eonje* 'when'. A word pair item 'A and B' was embedded in the carrier sentences, where the target word containing a word-initial plosive always appeared in the A position (e.g. /b/ in *eonje* ***baji****-wa chima-reul ibeoyo*? 'When do you wear **pants** and skirts?'). A total of 72 test sentences were created (3 places of articulation x 3 plosive types x 2 vowels x 4 word pair types). Additionally, 10 warm-up sentences with other initial consonants were presented to the participants at the beginning of the experiment. The warm-up sentences only contained actual Korean words.

The pairs of the target words were mixed with real and pseudo-words. Only disyllabic words (CVCV, CVCVC, CVCCV, and CVCCVC, with C indicating a consonant and V a vowel) were used to elicit a minimal pair-like speaking style with a maximally careful articulation. Only the vowels /a/ and /o/ followed the initial plosives of the target words. Overall, a total of 1,080 tokens (72 words x 15 subjects) were collected for analysis. One recording was excluded due to an excessive pause before the target word.

2.3. Experimental procedure

The experiment began with the warm-up sentences, followed by the sentences containing the target words, which were randomly displayed on a computer screen, following the procedure used by [15]. Normally, speakers read the given sentences aloud only once, but they could repeat them if they mispronounced a part of the sentence.

The production task was recorded using SpeechRecorder software [8] in soundproof recording facilities at the University of Zurich, Goethe-University Frankfurt, and the Acoustics Research Institute in Vienna. The following microphones were used for the recordings: a Røde NT2-A microphone in Zurich; a Røde Lavalier microphone attached to a Zoom H4n audio recorder in Frankfurt; and a AKG C451EB microphone in Vienna. The recordings were all sampled at 44.1 kHz and 16 bits.

2.4. Acoustic analysis

The three main acoustic cues, i.e. the CD and VOT of the initial plosive and the duration of its following vowel, were measured using Praat [2]. The following time intervals were all manually segmented: for CD, the interval from the voicing offset of the preceding vowel to the stop release; for VOT, the interval from the stop release burst to the voicing onset of the following vowel; and for the vowel, the interval from the onset to the offset of vocal cord vibration.

The durations of all target intervals (CD, VOT, vowel) were normalized by calculating z-scores to reduce the impact of the phonotactic context and to control for speech rate in the collected recordings.

2.5. Statistical analysis

Linear mixed-effects models were implemented in R [22] to estimate the complementary durational relationship between the main acoustic cues depending on their segmental context using the R packages *lme4* [1] and *lmerTest* [14]. The analysis of variance models included as dependent variables the z-scored CD duration (CD_{norm}), the z-scored VOT duration (VOT_{norm}), the z-scored plosive (CD plus VOT) duration ($Plosive_{norm}$), and the z-scored duration of the following vowel ($Vowel_{norm}$), respectively. All models used the fixed effects *plosive type* (lenis, fortis, aspirated), *place of articulation* (bilabial, alveolar, velar), *first syllable* (CV, CVC), and *vowel* (/a/, /o/), as well as the random effects *speaker* and *target word*, including both random intercepts and random slopes.

We report the results of type III F-tests and the corresponding p-values computed with the Satterthwaite's method, and corrected with the Tukey method, using the R package *emmeans* [17]. Correlations were calculated with Pearson's product-moment correlation.

3. RESULTS

3.1. Closure duration and Voice onset time

The results for CD_{norm} showed that both *plosive type* ($F[2, 71.4] = 227.3$, $p < .001$) and *place of articulation* (henceforth *PoA*, $F[2, 71.4] = 66.9$, $p < .001$) had highly significant effects on CD duration. The interaction of *plosive type***PoA* ($F[4, 71.4] = 4.3$, $p < .01$) also influenced the CD values. As shown in Figure 1, fortis plosives had the longest CD duration, followed by aspirated plosives, with lenis plosives having the shortest CD duration. Regarding *PoA*, CD values consistently decreased from bilabial to velar (a pattern which is well-known for VOT as well; cf. [9]).

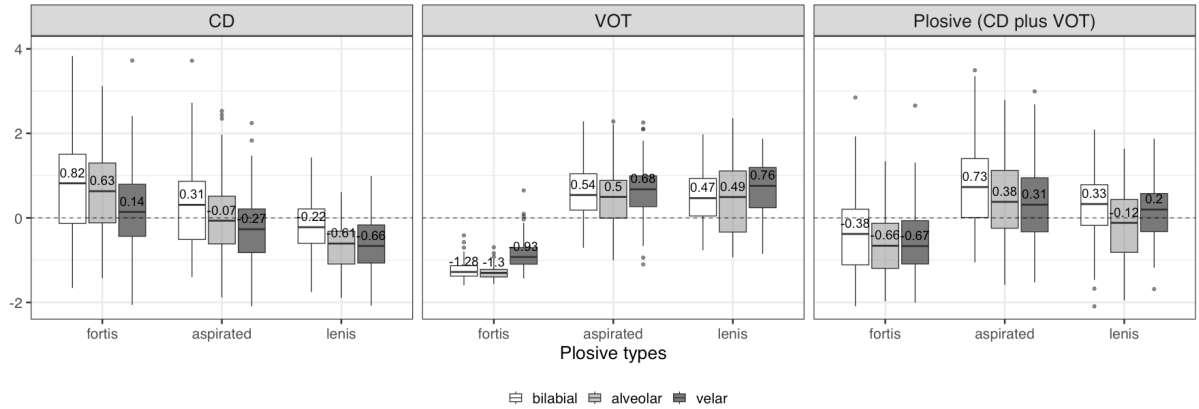


Figure 1: CD_{norm}, VOT_{norm}, and $Plosive_{norm}$, for fortis, aspirated and lenis plosives according to *PoA* (bilabial, alveolar, velar).

As concerns the three types of plosives, the VOT values showed exactly the opposite results. Fortis stops had very short absolute VOT values and lenis stops had relatively longer VOT values (see Table 1), consistently with previous findings (e.g. [13, 6, 23]). Aspirated stops, however, exhibited much shorter VOT values, which is inconsistent with previous reports (cf. ibid.).

	Absolute VOT (ms)	**Z-scored** VOT
fortis	18 ± 8	−1.13 ± 0.31
aspirated	61 ± 17	0.58 ± 0.67
lenis	61 ± 19	0.55 ± 0.73

Table 1: Mean and standard deviation for plosive types calculated from absolute VOT values (in milliseconds) and z-scored VOT values.

The results of pairwise comparisons in the model of VOT_{norm} supported that the 'merger' of VOT between lenis and aspirated plosives was clearly present in all *PoA*s (see Table 2 and light gray boxplot in Figure 1). Additionally, lenis plosives had a relatively greater standard deviation than aspirated plosives, suggesting that some lenis obstruents were even longer than aspirated ones. This occurred primarily in the case of alveolar and velar plosives.

	estimate	SE	df	t.ratio	p-value
bilabial	0.09	0.13	72.0	0.70	0.76
alveolar	−0.03	0.13	73.4	−0.24	0.97
velar	−0.07	0.13	71.9	−0.57	0.84

Table 2: Pairwise comparison evaluation of VOT_{norm} between aspirated and lenis plosives according to *PoA* (bilabial, alveolar, velar). Results are averaged for the levels *first syllable* and *vowel*.

Regarding *PoA*, VOT durations were longest for velar plosives compared to the other two more anterior places (bilabial-velar: estimate = −0.22, Standard error, henceforth SE = 0.08, df = 72.4, t = –2.89, $p < 0.05$; alveolar-velar: estimate = −0.29, SE = 0.07, df = 73.0, t = –3.91, $p < .001$) for all plosive types. The difference between bilabial and alveolar plosives was not significant (estimate = −0.07, SE = 0.07, df = 72.5, t = −0.96, p = 0.60; see also Figure 1).

3.2. Intrasegmental interaction within a plosive

The duration of the plosive phases changed steadily in an inverse relationship: a longer CD was followed by a shorter VOT and *vice versa* (see Figure 1). The result of the Pearson's correlation test supported the negative correlation between CD_{norm} and VOT_{norm} ($r = −0.32$, $p < .001$). This duration correlation was observed across *plosive type* and *PoA*, except for fortis and aspirated plosives at the alveolar place of articulation, where the shorter CD duration was not accompanied by longer VOT duration. Velar lenis plosives also displayed a marginally reduced CD duration compared to the alveolar lenis plosives, but they consistently maintained the 'short CD plus long VOT' pattern.

3.3. Intersegmental interaction between the target plosive and its following vowel

The target words were classified into two groups based on the first syllable: the CV group, embedding an initial consonant and a vowel, and the CVC group, containing an additional consonant in coda position. In both groups, the results of the regression analysis showed that *plosive types* had a significant effect on $Plosive_{norm}$ ($F[2, 71.9] = 92.8$, $p < .001$) and $Vowel_{norm}$ ($F[2, 70.6] = 132.3$, $p < .001$). Vowel duration was reduced in the order of fortis > lenis > aspirated, whereas plosive duration was marginally negatively correlated ($r = −0.14$, $p < .001$; see also Figure 2).

Figure 2 shows that the plosive and vowel durations were both shorter in the CVC syllables than in the CV syllables. In particular, the vowel duration was drastically reduced.

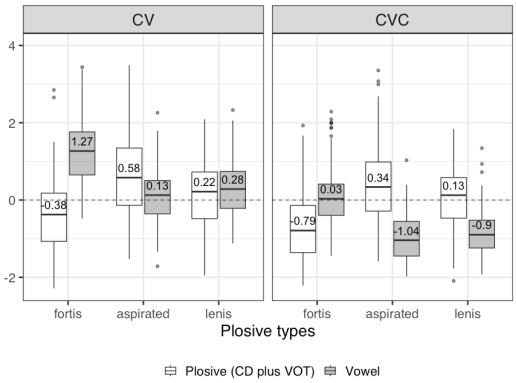


Figure 2: *Plosive*$_{norm}$ and *Vowel*$_{norm}$ for fortis, aspirated, and lenis plosives in the two different syllable types.

In contrast to the plosive type, the effect of *PoA* was less relevant to the interplay between *Plosive*$_{norm}$ and *Vowel*$_{norm}$. For aspirated and lenis plosives, the decrease in vowel duration values was proportionate to the reduced consonant length in both CV and CVC syllables (see Figure 3). Conversely, for fortis plosives, the variation in the duration of the two segments was irregular for both types of syllables on the basis of *PoA*.

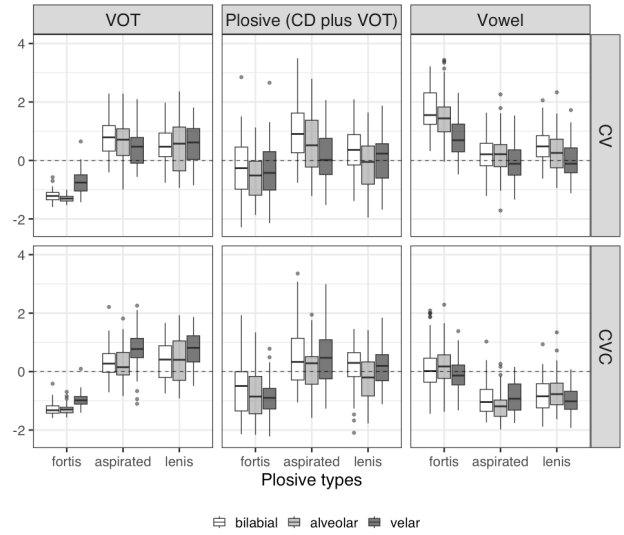


Figure 3: Durational variation of *VOT*$_{norm}$, *Plosive*$_{norm}$, and *Vowel*$_{norm}$, for fortis, aspirated, and lenis plosives according to *PoA* and syllable type.

Nevertheless, vowels gradually decreased in duration from the bilabial to the velar place of articulation of the preceding plosive (*PoA* effect on *Vowel*$_{norm}$: $F[2, 70.7] = 11.8$, $p < .001$), with the exception of alveolar fortis, velar aspirated, and alveolar lenis plosives in the CVC syllables. Similarly to the complementary relationship between CD and VOT (see § 3.2), this pattern was closely related to VOT alone in terms of durational compensation, as indicated above.

4. DISCUSSION

The present study investigated durational correlations within a plosive (intrasegmental) and between a plosive and a vowel (intersegmental) in word-initial plosives in Standard Korean. The findings on the acoustic cues of plosives and their internal and external interactions with adjacent segments can help shed more light on the acoustic properties of Korean plosives.

The results showed that (young) Korean speakers tend to produce aspirated obstruents with shorter VOT than reported in previous studies (e.g. [23, 11, 3]), especially for velar place of articulation. This points towards a merger between lenis and aspirated plosives and suggests that VOT alone does not seem to distinguish the different amount of aspiration any longer; rather the presence/absence of aspiration does, particularly when plosives are articulated in word-initial position (for word-medial plosives, see [10]). Instead, the lack of distinctiveness of VOT seems to be compensated by CD as an additional acoustic cue. This coherent pattern of variation in closure duration indicates that CD may play an essential role in discriminating lenis and aspirated plosives, regardless of place of articulation. As hypothesized, the lengthening of CD may serve as a means of adjusting the durations of CD and VOT in terms of a complementary relationship (see also [5, 6]).

Moreover, total plosive duration and vowel length were interdependent in that they were inversely related across plosive types. In particular, VOT considerably affected the duration of the following vowel: vowels were longer after unaspirated stops than after (slightly/strongly) aspirated stops. This variation in vowel duration is in line with a universal pattern of vowel length in CV or VC sequences [19, 4]. Furthermore, both directions of the transition from C to V or from V to C did not cause an unexpected pattern of vowel length variation. In other words, the adjacency between plosives and vowels itself plays a key role in adjusting the duration of target segments in a complementary relation.

5. CONCLUSION

In the present study, the three types of Korean plosives were clearly distinguishable in CD and VOT. These acoustic cues were closely interdependent in terms of a complementary adjustment of their durations, and this durational compensation showed a rather regular pattern across places of articulation. Moreover, vowel length distinction was highly interrelated with the presence/absence of aspiration.

6. ACKNOWLEDGMENTS

My sincere appreciation goes to my supervisors Stephan Schmid and Eva Maria Luef, as well as the two anonymous reviewers, for their valuable feedback. I am also grateful to Carolina Lins Machado and Franka Zebe for their assistance with statistical analyses. Any remaining errors are solely my own.

7. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[2] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program], version 6.1.52. http://www.praat.org/.

[3] Byun, H.-G. 2016. F0 as a primary cue for signaling word-initial stops of Seoul Korean. *Phonetics and Speech Sciences* 8(1), 25-36.

[4] Cho, T., Ladefoged, P. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27(2), 207-229.

[5] Cho, T., Keating, P. A. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics* 29(2), 155-190.

[6] Cho, T., Jun, S.-A., Ladefoged, P. 2002. Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics* 30(2), 193-228.

[7] Choi, H. 2011. Vowel Duration as a Perceptual Cue for Preceding Stop Laryngeal Contrast in Korean. *Proceedings of the 17th International Congress of Phonetic Sciences* Hong Kong, 17-21.

[8] Draxler, C., Jänsch, K. 2004. SpeechRecorder: A universal platform independent multi-channel audio recording software. In: Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (eds), *Proceedings of the IV international conference on language resources and evaluation* Lisbon, 559–562.

[9] Jochim, M., Kleber, F. 2022. Fast-Speech-Induced Hypoarticulation Does Not Considerably Affect the Diachronic Reversal of Complementary Length in Central Bavarian. *Language and Speech*, 1-35.

[10] Jun, S.-A., Beckman, M., Niimi, S., Tiede, M. 1998. Electromyographic Evidence for a Gestural-Overlap Analysis of Vowel Devoicing in Korean. *UCLA Working Papers in Phonetics*, 1-42.

[11] Kang, Y. 2014. Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics* 45, 76-90.

[12] Kang, Y., Yoon, T.-J., Han, S. 2015. Frequency effects on the vowel length contrast merger in Seoul Korean. *Laboratory Phonology* 6(3-4), 469–503.

[13] Kim, M. R. C. 1994. *Acoustic characteristics of Korean stops and perception of English stop consonants* (Ph.D. dissertation). Madison: University of Wisconsin-Madison.

[14] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2018. lmerTest: Tests in linear mixed effects models. R Package, version 3.1.3.

[15] Ladd, D. R., Schmid, S. 2018. Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops. *Journal of Phonetics* 71, 229-248.

[16] Lee, S. H. 1998. On the Effects of Places of Articulation of Stops on Their Closure Duration in Korean. *The Journal of the Acoustical Society of Korea* 17(6), 8-13.

[17] Lenth, R. V., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., Singmann, H. 2022. emmeans: Estimated marginal means, aka least-squares means. R package, version 1.8.3.

[18] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3), 384-422.

[19] Maddieson, I. 1997. Phonetic Universals. In: Laver, J., Hardcastle, W.J. (eds), *The handbook of phonetic sciences*. Oxford: Blackwell, 619-639.

[20] Magen, H. S., Blumstein, S. E. 1993. Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics* 21, 387-409.

[21] Park, J. W. 1994. Variation of vowel length in Korean. In: Y. K. Kim-Renaud (ed), *Theoretical issues in Korean linguistics*. Stanford, CA: Center for the Study of Language and Information, 175-187.

[22] R Core Team. 2022. R: A language and environment for statistical computing, version 4.2.2. http://www.R-project.org/.

[23] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287-308.

[24] Shin, J., Kiaer, J., Cha, J. 2012. *The sounds of Korean*. Cambridge University Press.

[25] Tanner, J., Sonderegger, M., Stuart-Smith, J., Fruehwald, J. 2020. Toward "English" Phonetics: Variability in the Pre-consonantal Voicing Effect across English Dialects and Speakers. *Frontiers in Artificial Intelligence* 3, 38. https://doi.org/10.3389/frai.2020.00038

[1] The [+tense] feature of Korean fortis consonants is often marked with the diacritic /*/ (e.g. in [6]).

ARTICULATORY CORRELATES OF STOP PHONATION IN L1 KOREAN AND L2 ENGLISH

Gwanhi Yun[1], Jae-Hyun Sung[2]

[1]Daegu University, [2]Kongju National University
ghyun@daegu.ac.kr, jsung@kongju.ac.kr

ABSTRACT

A growing body of research has identified and documented various acoustic cues for laryngeal contrasts, but it is still elusive what articulatory correlates various laryngeal contrasts are mapped onto, and whether such articulatory correlates emerge in a speaker's L1 and L2. Using ultrasound imaging of Korean-English late bilinguals, this study examines articulatory patterns of L1 Korean and L2 English alveolar stops. Results from eight speakers show that Korean fortis and aspirated stops and English voiced stops manifest as advanced tongue root when compared to Korean lenis stops and English voiceless stops respectively, along with some individualized patterns. Oral and nasal stops in both languages, albeit less robustly, also yielded some degree of articulatory distinction. These articulatory patterns offer new and exciting insights into abstract laryngeal categories and supralaryngeal articulatory gestures. Tongue root advancement serves as an articulatory correlate of Korean stop contrast, and also a cue to enhance voicing contrast in L2 English.

Keywords: stop phonation, tongue root advancement, ultrasound imaging, Korean, English

1. INTRODUCTION

Laryngeal stop contrast yields two major types: 'true-voice' and 'aspirating' types [1, 2]. Languages such as French, Spanish and Russian exhibit true-voice pattern where stops are classified as prevoiced and voiceless unaspirated stops. On the other hand, 'aspirating' languages such as English and German, word-initial stops appear as either plain, voiceless unaspirated and voiceless aspirated stops. It is widely acknowledged that various types of consonants are differentiated by a myriad of acoustic and perceptual cues, such as VOT for unaspirated vs. aspirated stops and f0 of the following vowel for fortis vs. non-fortis stops [3, 4]. Despite the numerous studies on acoustic and perceptual cues for laryngeal contrasts, it still calls for an investigation what articulatory correlates are mapped onto such cues. Focusing on tongue body/root movement, this study examines articulatory gestures that are associated with three laryngeal stops in Korean, and voiceless and voiced stops in English produced by Korean L2 learners of English.

As one of the aspirating languages, Korean exhibits a 3-way contrast of lax (/p, t, k/), fortis (/p', t', k'/), and aspirated (/p^h, t^h, k^h/) stops. Of our special interest are the fortis stops, which have traditionally been identified as 'tense'. Previous studies on tense consonants show that tense consonants are produced with more muscular "tension" in the pharyngeal region of the tongue than lax consonants [5], and result in longer duration than lax counterparts [6]. English, another aspirating language, shows voiceless (/p, t, k/) and voiced (/b, d, g/) stop distinction, which are also associated with various aerodynamic and acoustic correlates. Voiced stops are made with the increased supralaryngeal volume than voiceless stops, along with variation among different places and phonological contexts [7]. Recent ultrasound studies showed that tongue root is moved further front and tongue body is moved more downward during English voiced stop closure [8], and tongue root displacement is made for different stops and fricatives in Korean [9], albeit based on a small number of speakers.

While aerodynamic, acoustic and articulatory correlates provide complementary documentation of laryngeal contrast in two languages, an additional articulatory investigation with more speakers would further our understanding of how laryngeal contrast is made with articulatory gestures. Furthermore, this study also looks into articulatory correlates of stops produced by L2 speakers, represented by Korean leaners of English. In the present study, we address three research questions to examine the articulatory correlates of L1 Korean three-way stop contrast and L2 English voicing contrast, along with the gestural differences between oral and nasal stops in two languages. First, we examine if there is an intraoral articulatory correlate of L1 Korean stop

distinction. Second, we investigate whether the enlargement of oral cavity is associated with tongue root advancement for L2 English stop voicing. Lastly, we look for tongue body configuration to differentiate oral and nasal stops in L1 Korean and L2 English.

2. METHODS

2.1. Participants

Eight native speakers of Korean (4 females) participated in the production experiment. All participants were undergraduate students majoring in English, recruited from a university in South Korea, with the age range of 20 to 26 (mean age = 21). They learned English through the formal education system in South Korea for an average of 11 years. Their averaged English proficiency test score and self-rated proficiency score are 560 out of 990 on TOEIC® (Test Of English for International Communication) and 5.8 out of 10 point-scale, respectively, representing a group of upper-intermediate learners of English.

2.2. Materials

Various di- and trisyllabic words, where all target segments were word-initial, were employed in the production experiment. Given the potential V-to-C coarticulatory effects, all target segments were produced in three vowel contexts: /i, a, u/.

	lax /t/	fortis /t'/	aspirated /t^h/	nasal /n/
/i/	tita	t'ita	t^hita	nita
/a/	tata	t'ata	t^hata	nata
/u/	tuta	t'uta	t^huta	nuta

Table 1: Examples of Korean stimuli.

	voiceless /t/	voiced /d/	nasal /n/
/i/	teem	deep	need
/a/	tah-dah	darling	narcissism
/u/	too	doodle	noon

Table 2: Examples of English stimuli.

2.3. Procedures

Ultrasound imaging was employed to identify, extract and analyze the midsagittal tongue surface images. The production experiment with ultrasound imaging was proceeded in a quiet room at a university in South Korea. Articulatory and acoustic data were simultaneously collected using Articulate Assistant Advanced (AAA) software, connected to the EchoB ultrasound machine and 5-10 MHz convex-curved transducer. Participants were asked to wear a head-stabilizing headset, sit comfortably in front of a computer monitor and read out a list of stimuli at a normal speed. Following the data collection, midsagittal tongue images of interest were identified and extracted manually.

Extracted tongue contours were statistically analyzed using SSANOVA. As shown in Figure 1, when the averaged tongue curves and confidence intervals in different conditions do not overlap, tongue curves in different colors are considered statistically different, equivalent to $p<.05$.

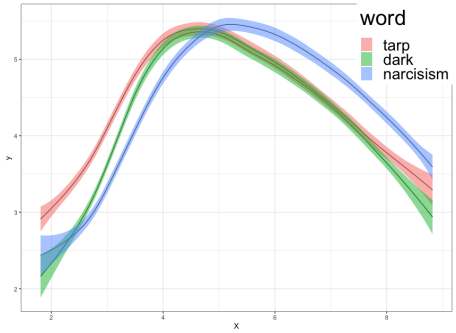


Figure 1: SSANOVA plot.

3. RESULTS

3.1. L1 Korean laryngeal contrast

Table 3 presents the percentages of speakers who showed 3-way and 2-way articulatory contrast for Korean lax, fortis and aspirated stops. Figure 2 illustrates 3-way comparisons of lax (/t/), fortis (/t'/), and aspirated (/t^h/) stops in Korean produced by Speakers #1 and #6, and Figure 3 shows 2-way comparisons of oral (/t/) and nasal (/n/) stops produced by the same speakers.

The results from Table 3 indicate that a vast majority of speakers make articulatory distinction for different laryngeal stops. Gestural patterns represented by two speakers in Figure 2 show a clear sign of tongue root advancement for the fortis stop (/t'/) when compared to non-fortis stops, and also for aspirated stop (/t^h) when compared to the lax counterpart. Tongue configuration is mediated by different vowel conditions, in which the /a/ environment yield more articulatory contrast

among three stops. As illustrated in Figure 3, the articulatory distinction between oral and nasal stops is not as robust as that among three laryngeal oral stops.

	lax vs. fortis vs. aspirated /t/ vs. /t'/ vs. /t^h/		oral vs. nasal /t/ vs. /n/
/i/	0%(3-way)	25%(2-way)	25%
/a/	50%(3-way)	37.5%(2-way)	37.5%
/u/	12.5%(3-way)	37.5%(2-way)	50%

Table 3: Summary of the Korean results.

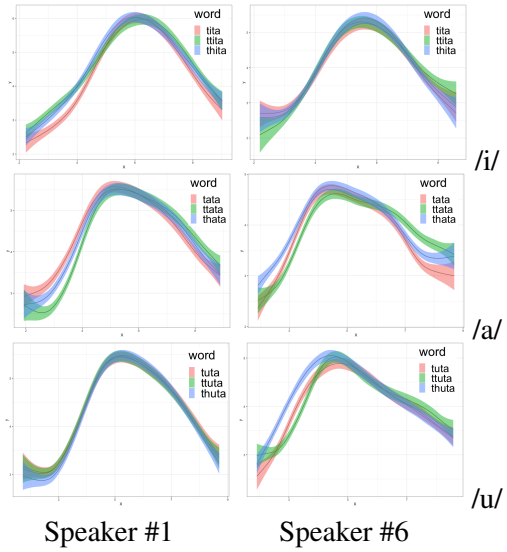


Figure 2: Korean lax (pink), fortis (green), and aspirated (blue) stops in three vowel conditions, produced by Speakers #1 and #6.

3.2. L2 English laryngeal contrast

Table 4 presents the percentages of speakers who showed 3-way articulatory contrast for English voiceless, voiced and nasal stops. Figure 4 shows 3-way comparisons of voiceless (/t/), voiced (/d/), and nasal (/n/) stops in English produced by Speakers #1 and #6. All 2-way articulatory contrasts in Table 4 were made for voiceless vs. voiced stops, with an exception of one instance by Speaker #5 who differentiated oral and nasal stops only in the /a/ context.

Figure 4 indicates that Korean learners of English do produce the articulatory distinction between voiceless and voiced stops, manifested as tongue root advancement for the voiced stop, in line with the findings from L1 English speakers [7, 8]. As for oral and nasal stops, only one speaker (Speaker #5) makes articulatory distinction between nasal

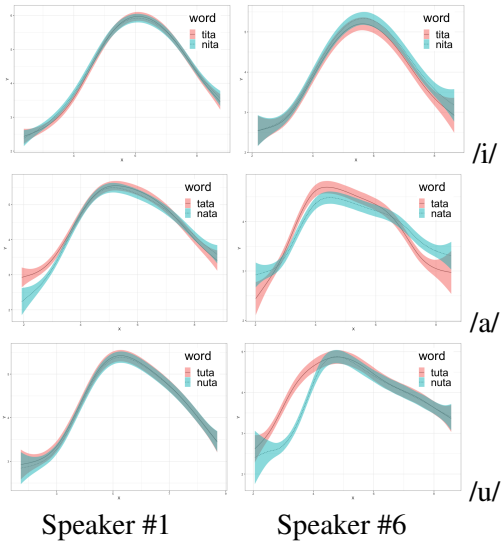


Figure 3: Korean oral (pink) and nasal (blue) stops in three vowel conditions, produced by Speakers #1 and #6.

	voiceless vs. voiced vs. nasal /t/ vs. /d/ vs. /n/	
/i/	0%(3-way)	25%(2-way)
/a/	37.5%(3-way)	50%(2-way)
/u/	25%(3-way)	12.5%(2-way)

Table 4: Summary of the English results.

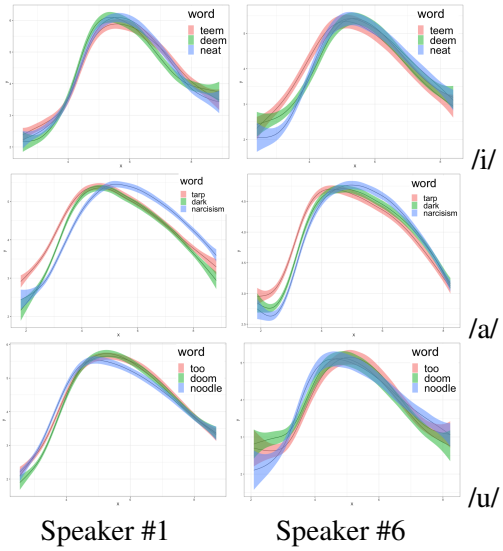


Figure 4: English voiceless (pink), voiced (green), and nasal (blue) stops in three vowel conditions, produced by Speakers #1 and #6.

and non-nasal stops. The gestural patterns of English voiceless and voiced stops produced by L2

learners suggest that Korean learners of English make articulatory efforts to produce voiced stops, even when voiceless and voiced stops are markedly different in the acoustic dimension, as illustrated in Figure 5. Voiceless and voiced stops produced by Speakers #1 and #3 show that they are clearly different in VOT and closure voicing. Tongue configuration for voiceless and voiced stops in this study suggests that tongue root advancement acts as L2 learners' cue that enhances the voicing contrast in L2.

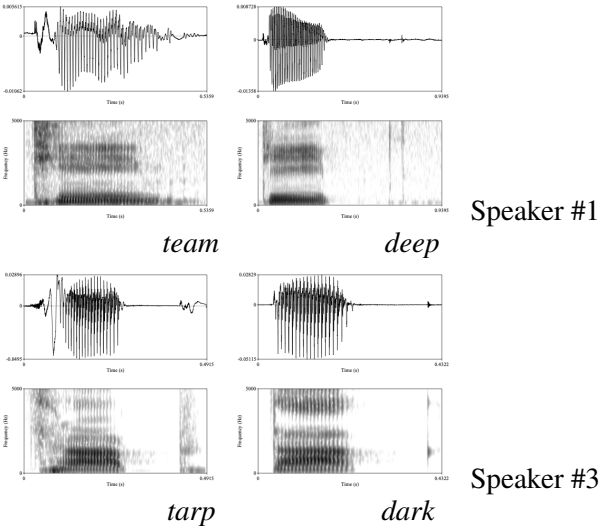


Figure 5: Waveforms and spectrograms of *team*, *deep*, *tarp*, and *dark* produced by Speakers #1 and #3.

In sum, our findings from Korean-English late bilinguals provide empirical evidence for the articulatory cue [ATR] at play both native and non-native stop contrast.

4. DISCUSSION AND CONCLUSIONS

The gestural patterns from our study indicate that the tongue root is advanced further front for Korean fortis stops than lenis or aspirated ones, and for English voiced stops than voiceless ones. In addition, compared to the contrast among oral stops, oral and nasal stops in both L1 and L2 do not differ markedly in the magnitude of tongue root displacement. The findings from this study offer new insight into tongue configuration for Korean and English stops. Tongue root advancement, or [ATR] serves as an articulatory cue to laryngeal contrast in L1 Korean and L2 English, accompanied or enhanced by the adjustment of supralaryngeal oral tract volume in conjunction with tongue root displacement. Less marked distinction between oral and nasal stops can be explained as the lowering of larynx, which obviates the need to advance the tongue root for nasal stops and expand the supralaryngeal volume.

Our findings can be further supported by future investigation with a wider variety of phonological and extralinguistic factors. Of our interest is whether different places of articulation and prosodic positions, e.g., utterance-medial vs. -final, result in different articulatory configuration. Furthermore, examining L2 speakers with various proficiency levels will lead to a more comprehensive understanding of tongue configurations for L1 and L2 stops.

5. REFERENCES

[1] J. Beckman, M. Jessen, and C. Ringen, "Empirical evidence for laryngeal features: Aspirating vs. true voice languages," *Journal of Linguistics*, vol. 49, pp. 259–284, 2013.

[2] S. Fuchs, "Articulatory correlates of the voicing contrast in alveolar obstruent production in german," *ZASPiL*, vol. 41, pp. 1–238, 2005.

[3] K. H. Kang and S. G. Guion, "Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3909–3917, 2008.

[4] E. J. Kong, M. E. Beckman, and J. Edwards, "Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese," *Journal of Phonetics*, vol. 40, pp. 725–744, 2012.

[5] J. S. Perkell, *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. MIT Press, 1969.

[6] M. Jessen, *Phonetics and phonology of tense and lax obstruents in German*. John Benjamins, 1998.

[7] J. Westbury, "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," *The Journal of the Acoustical Society of America*, vol. 73, pp. 1322–1336, 1983.

[8] S. Ahn, "The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese," *Journal of Phonetics*, vol. 71, pp. 451–467, 2018.

[9] H. Kwon and S. Ahn, "Tongue positioning during Seoul Korean lax, tense, and aspirated obstruents," Paper presented at the UltraFest IX, Indiana University, 2020.

/S/-LENITION AND RESYLLABIFICATION IN SOUTHERN CONE SPANISHES

Gilly Marchini, Michael Ramsammy

University of Edinburgh
G.E.M.Marchini@sms.ed.ac.uk, m.ramsammy@ed.ac.uk

ABSTRACT

This paper presents experimental results on the interaction between word-final /s/-lenition and vowel reduction in Spanish. Data from 6 speakers of Southern Chilean Spanish (SCS) and 5 speakers of Buenos Aires Spanish (BAS) were analysed. Results show that final debuccalisation to [h] (i.e. /s/-*aspiration*) both shortens preceding vowels and the fricative itself in both dialects. Our data further highlight that, with both [s] and [h]-allophones of /s/, vowel and /s/-durations in the /Vs#V/ context are not significantly different from /Vs#C/ in either dialect. However, realisations of /Vs#V/ do show significant durational reduction in comparison to /V#sV/ in both SCS and BAS.

Despite established theoretical claims about the operation of phrasal resyllabification in Spanish, our data therefore suggest that /s/ in the /Vs#V/ environments behaves as though it remains in canonical coda position word-finally. Consequently, our findings contribute to growing scholarship disputing the categorical nature of resyllabification in Spanish.

Keywords: Vowel reduction, /s/-debuccalisation, Southern Cone Spanish

1. INTRODUCTION

This study analyses word-final /s/-lenition in Southern Chilean Spanish (SCS) and Buenos Aires Spanish (BAS) on the basis of a corpus of spontaneous speech recordings. Background on /s/-lenition is provided in §2, and the methods employed in the study are outlined in §3. We analyse the effect of aspiration and elision on preceding vowel duration (§4.1) prior to examining the interaction of /s/-lenition and vowel shortening with syllabic affiliation (§4.2). We compare vowel and /s/-duration in canonical coda position (i.e., /Vs#C/) with realisations in which /s/ remains in canonical onset position (i.e., /V#sV/) and those in which it is said to resyllabify across a word boundary as a *derived onset* (i.e., /Vs#V/). The discussion in §5 focuses on evidence for dialect-specific phrasal resyllabification operations in Spanish.

2. BACKGROUND

2.1. /s/-lenition

Southern Cone Spanishes are a group of dialects spoken in the southern tip of South America sharing several features, including /s/-lenition [1]. This is a phonological process that causes /s/ to display some degree of phonetic weakening in syllable-final contexts. For example, in *conos* ('cones'), /s/ may be retained as [s] in more formal speech (i.e., ['ko.nos]). Elsewhere, it may aspirate to [h] (['ko.no**h**]) or completely elide (['ko.no**Ø**]). Lenition patterns vary cross-dialectally: advanced lenition varieties, such as Caribbean and Peninsular Spanish dialects, routinely elide /s/ [2, 3]. By contrast, semi-conservative varieties, such as Southern Cone Spanishes, may favour aspiration [1, 4].[1] This occurs both with inflectional /s/, as in *conos*, and /s/ in monomorphemes, like *lunes* 'Monday'.

Particularly pertinent to the present study are the effects of /s/-lenition on preceding vowels. In Eastern Andalusian Spanish, mid and low-vowels (/e, a, o/) in contact with elided /s/ open and lax [5, 6] with regressive harmony effects noted for bisyllabic words containing two mid vowels (e.g., *conos*, ['kɔ.nɔØ]) [5]. In contrast, shortening and centralisation effects are observed in Caribbean Colombian Spanish: e.g., ['ko.nə̆Ø] [6]. In Cuban Spanish, word-final /s/-elision has no observable effect on surrounding vowels, whereas word-medial /s/-elision triggers lengthening of preceding vowels [7, 8]: e.g., *bosque* ['boːØ.ke], 'forest'.

Whilst similar dialect-specific patterns are noted in other advanced varieties (e.g., Western Andalusian Spanish [9], Puerto Rican Spanish [10, 2] and Dominican Spanish [3]), research on Southern Cone varieties is scarce. Regarding SCS, Bolyanatz [11, 12] showed that qualitative and/or quantitative vowel alternations are dependent on key social factors (e.g., gender, age). For BAS, impressionistic analysis from Honsa [13]

indicates that when plural /s/ elides, mid and low vowels lengthen and lax to disambiguate otherwise homophonous singular items: e.g., *cono* [ˈko.no] vs *conos* [ˈko.nɔːØ]. However, the occurrence of this pattern has not been verified through acoustic analysis.

Thus, existing research has focused overwhelmingly on the effects of /s/-elision in advanced varieties such that similar effects in Southern Cone varieties have typically only been discussed in impressionistic terms. Furthermore, the effects of aspiration on vowel and /s/-duration as an intermediary step towards elision have not yet been explored. With the exception of Bolyanatz [11, 12], analyses have focused on lenition in controlled speech, such that the question of whether similar effects occur in spontaneous speech—where lenition is said to be most commonly observable [4]—remains to be addressed. It is also unclear how lenition effects may interact with other phonological processes targetting syllable-final consonants, particularly phrasal resyllabification.

2.2. Resyllabification

Phrasal resyllabification is the process in which word-final consonants in Spanish are said to relocate to occupy the onset position of a following vowel-initial word: e.g., /las**#a**las/ → [la.ˈ**sa.**lah] 'the wings'. It has been argued that phrasal resyllabification operates pan-dialectally in Spanish [1, 14]. Thus, resulting derived onsets are predicted to be phonetically indistinguishable from canonical onsets: cf. /V#sV/, as in [la.ˈ**sa**.la] 'the room', versus /Vs#V/, as in [la.ˈ**sa**.lah] 'the wings'.

However, there is increasing scholarship that disputes these claims. Of particular relevance is Strycharczuk and Kohlberger [15], who showed that in Peninsular Spanish, /s/ in derived onset (/Vs#V/) environments is durationally distinct from /s/ in both canonical coda (/Vs#C/) and onset position (/V#sV/). This suggests that derived onsets may form their own unique phonetic category (see §5). Although there has been some phonetic investigation on /s/-allophony in Southern Cone varieties [4, 11, 12], no research to date has closely examined the putative outcomes of resyllabification on the basis of spontaneous speech data.

2.3. Research questions

Given the paucity of research into /s/-lenition and resyllabification in Southern Cone Spanishes, this paper focuses on the following questions:

RQ1. What are the durational consequences of /s/-lenition on word-final /s/-realisations and preceding vowels in SCS and BAS?

RQ2. How do /s/-realisations and preceding vowels vary durationally in SCS and BAS according to the putative syllabic affiliation of /s/ (i.e., across /Vs#C, Vs#V, V#sV, Vs#sV/ contexts)?

RQ3. To what extent can analysis of spontaneous speech provide wider insights into the dialect-specific operation of /s/-aspiration and phrasal resyllabification in Spanish?

Regarding RQ3, we specifically aim to test whether /s/ in resyllabification contexts (/Vs#V/) patterns phonetically with canonical onsets (/V#sV/), as is predicted by phrasal resyllabification. By contrast, should vowels and /s/-realisations (either [s] or [h]) in these environments show different durational properties—e.g. greater similarities to /s/ in canonical coda position (/Vs#C/)—this may provide further evidence against the operation of phrasal resyllabification (cf. [15]). Another possibility is that /s/ in resyllabification sites is phonetically distinct from realisations in other phonological environments, thus supporting the existence of a homogeneous third category.

3. METHODS

In accordance with previous work, we restricted our analysis to the realisation of word-final /s/ in the context of preceding mid and low vowels, i.e., /e, a, o/. The high vowels (/i, u/) were excluded from the analysis due to their infrequency in appearing in word-final syllables in Spanish. We analysed realisations of /s/ in the following environments: word-final pre-vocalic /Vs#V/ (*cosas o* 'things or'); word-initial pre-vocalic /V#sV/ (*cosa suya* 'her/his thing'); pre-consonantal /Vs#C/ (*cosas que* 'things that'); pre-pausal /Vs##/ (*cosas ##*); false geminate /Vs#s/ (*cosas son* 'things are'). Tokens with diphthongs or long vowels derived by /s/-elision (e.g., *cosas y* → [ˈko.saj] 'things and'; *cosas así* → [ˈko.saː.ˈsi] 'things like that') were also excluded.

The data come from sociolinguistic interviews conducted with 11 speakers of Spanish (6 SCS, 5 BAS) aged between 21 and 37 years old. Participants self-identified as cis-gender male or female, were university-educated and had been born and raised in either Santiago de Chile or Buenos Aires. None reported hearing or speech pathologies. Approximately 35 minutes worth of speech was analysed per speaker totalling 2,765 /s/ tokens. Due to the COVID-19 pandemic, speech was recorded

remotely: we opted to use the video calling platform Zoom due to its accuracy in capturing relevant acoustic information [16, 17]. All participants wore headphones with an integrated microphone to ensure consistent distance from the recording device.

Speech was segmented using the Montreal Forced Aligner [18] and manually corrected. Praat scripts [19] extracted vowel, /s/ and combined rhyme duration (ms). Tokens were coded for dialect (BAS, SCS), vowel (a, e, o), syllabic affiliation of /s/ (canonical coda, canonical onset, derived onset, false geminate), stress context (stressed, unstressed), speaker gender (female, male), following context (pre-pausal, pre-consonantal, pre-vocalic) and /s/-realisation (sibilant, aspirated, elided).

Protocol for coding /s/ was adapted from Rogers [4]: /s/ was considered retained ([s]) when high-intensity frication was observed in the upper frequency regions. [h]-realisations were identified through reduced intensity and more evenly distributed turbulence throughout the lower frequency ranges. /s/ was considered elided where no fricative energy was visible on the waveform or in any spectral regions.

Statistical analysis was carried out in R [20] with plots made using the *ggplot2* package [21]. Linear mixed-effects regression (*lme4* package, [22]) was used to model the duration data. Post-hoc pairwise comparisons were conducted using estimated marginal means (*emmeans* package) with bonferroni adjustments.

4. RESULTS

4.1. Durational correlates of /s/-lenition

First and foremost, we note that vowels adjacent to [h] are significantly shorter than in contexts where /s/ is realised as [s] or [Ø] (p <.001 for all comparisons, dialectal differences n.s.). Thus, durational variation between vowels in [s] and [Ø] contexts varied non-significantly (t = 1.733, p >.1).

In order to compare the effect of aspiration on /s/-duration, the data were subset to exclude syllables in which /s/ was elided. In this data subset (N = 1340), vowels before [h] in both dialects are significantly shorter than in unlenited contexts (t = 5.705, p <.001). Fricative durations in [h]-realisations are also significantly shorter than in sibilant [s] (t = 13.204, p <.001). This double shortening effect is therefore visible in overall rhyme duration (cf. Figure 1): [h] rhymes are significantly shorter than [s] rhymes (t = 11.969, p <.001). Although segment duration was generally greater in SCS than BAS, differences between the two dialects fail to reach significance for all measures (vowel, /s/, and rhyme duration; all comparisons p >.1).

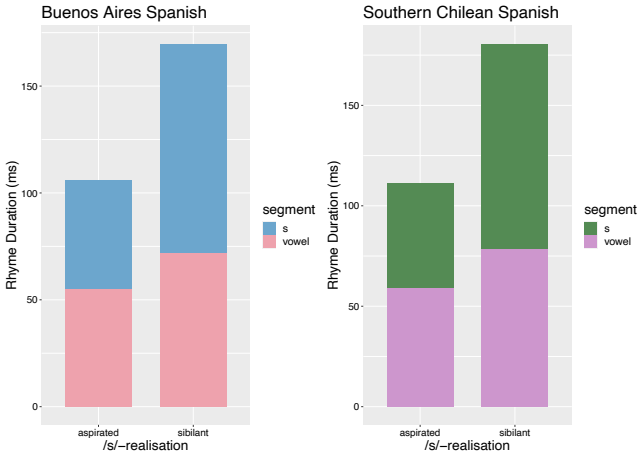


Figure 1: Mean vowel and /s/-durations (values in ms) by /s/-realisation and dialect.

4.2. Syllabic affiliation of /s/

In order to address RQ2, data were subset to include /Vs#C, Vs#V/ and /Vs#sV/ contexts: pre-pausal tokens were removed from the analysis due to obvious pre-boundary lengthening effects in /Vs##/.

Regarding /s/-duration, [s]-realisations are significantly longer (p <.01) in /Vs#s/ than in all other environments except /V#sV/ (t = 1.769, p >.1). [s] in /V#sV/ is longer than in both /Vs#C/ and /Vs#V/ (t = -3.253 and t = 6.095 respectively, both p >.1), whilst differences between /Vs#C/ and /Vs#V/ did not reach significance (t = 1.509, p >.1).

Regarding [h]-realisations, only comparisons between /Vs#C/ and /Vs#V/ were possible due to the infrequency of aspiration outwith these contexts. However, results again reveal non-significant variation between these environments (t = 1.509, p >.1). Comparisons between aspirated and sibilant tokens in /Vs#C/ and /Vs#V/ reveal that, in both contexts, [h]-realisations are consistently shorter than [s]-realisations (both p <.001). Comparisons between the two dialects were non-significant across all conditions (all p >.1).

Similar trends are observed for vowel duration. Vowels preceding [s] in /Vs#s/ and /V#sV/ are significantly longer than those in /Vs#C/ and /Vs#V/ (all p <.01). Thus, vowel duration varies non-significantly between /Vs#s/∼/V#sV/ and /Vs#C/∼/Vs#V/ pairs (both p >.1). Comparisons between aspirated realisations of /Vs#C/ and /Vs#V/ failed to reach significance (t = -1.697, p

>.1). Again, across all conditions, vowels preceding [h] are significantly shorter than vowels preceding [s] (all p <.001) whilst comparisons between dialects are non-significant (p >.1).

Our results therefore confirm that vowels and /s/-realisations (whether lenited or not) pattern similarly in /Vs#V/ and /Vs#C/. By comparison, significantly longer vowel and /s/-realisations are observed in /V#sV/.

dialect	s-type	syllable context	vowel dur.	/s/ dur.	rhyme dur.
BAS	[s]	/Vs#C/	84.45	106.57	191.02
		/Vs#s/	60.77	87	147.77
		/V#sV/	59.2	93.89	153.1
		/Vs#V/	68.72	85.15	153.88
	[h]	/Vs#C/	56.82	53.27	110.09
		/Vs#V/	55.7	57.23	112.93
SCS	[s]	/Vs#C/	79.89	107.51	187.39
		/Vs#s/	62.42	106.85	169.27
		/V#sV/	59.86	95.9	155.77
		/Vs#V/	71.44	81.94	153.38
	[h]	/Vs#C/	56.83	51.87	108.7
		/Vs#V/	61.8	50.32	112.12

Table 1: Mean vowel, /s/ and rhyme duration (values in ms) for BAS and SCS according to /s/-realisation and syllabic affiliation.

5. DISCUSSION

In answer to RQ1, the data show that aspiration significantly shortens word-final /s/ and preceding vowels in both SCS and BAS: in other words, /s/-lenition compresses the entire rhymal unit. By contrast, /s/-elision has no significant effect on preceding vowel duration in either dialect. Although SCS and BAS behave similarly, the observed patterns are distinct from those reported for other /s/-leniting dialects. Our findings therefore concur with the general observation that the phonetic outcomes of /s/-lenition are dialect-specific.

Regarding RQ2, word-final /s/ and vowel durations in the derived onset environment (/Vs#V/) did not vary significantly from measurements extracted from canonical coda /s/ (/Vs#C/). However, the durations of /s/ and preceding vowels in canonical onset position are significantly longer than both /Vs#V/ and /Vs#C/. With reference to RQ3, these findings suggest that so-called derived onsets are not durationally identical to canonical onsets, nor do they form their own unique category in these varieties. Rather, in spite of theoretical claims about the alleged exceptionless operation of phrasal resyllabification in Spanish, our findings are consistent with the assumption that /s/ remains in the coda in /Vs#V/. Accordingly, the data provide no positive evidence that phrasal resyllabification operates in these dialects.

These findings echo previous studies that highlight the non-categorical, dialect-specific nature of resyllabification in Spanish. For example, in advanced varieties of Caribbean Spanish, word-final, pre-vocalic /s/-aspiration may occur despite resyllabification. Kaisse [23] therefore suggests that word-final /s/ continues to behave (i.e., aspirate) as if it were in canonical coda position and thus that resyllabification may simply not apply in these environments. In Highland Ecuadorian Spanish, resyllabification varies according to the phonetic realisation of the sound: Robinson [24] showed that word-final, pre-vocalic [n] and [s] were always judged as onsets by native speakers, whilst [ŋ] and [z]-allophones in the same environments were judged as codas. As mentioned, Strycharczuk & Kohlberger [15] showed that in Peninsular Spanish, derived onset /s/ formed its own unique category from both canonical coda and onset /s/, suggesting that this variety may indeed evidence incomplete resyllabification across word boundaries.

In light of these facts, our findings present additional evidence that phrasal resyllabification varies cross-dialectally. Specifically, durational data extracted from unmonitored spontaneous speech do not support the claim that coda /s/ resyllabifies in word-final pre-vocalic contexts.

6. CONCLUSION

This study shows that /s/-aspiration—but not elision—shortens vowels in word-final syllables in both SCS and BAS. Shortening affects not only the fricative but the entire rhymal unit: as a whole, [Vh]-rhymes are consistently shorter than [Vs]-rhymes in both dialects under consideration. Furthermore, final /-Vs/-syllables in putative resyllabification contexts behave as rhymes: i.e., /s/ patterns like a canonical coda. This is to say that final /-Vs/-syllables are significantly shorter than /V#s/ sequences containing a canonical onset. Despite potential methodological limitations of remote data collection using Zoom and the necessity to exclude certain sequences of interest from the analysis (derived diphthongs and long vowels), we conclude that the predictions of phrasal resyllabification are not borne out by our spontaneous-speech data from SCS and BAS.

7. REFERENCES

[1] J. I. Hualde and S. Colina, *Los sonidos del español*. Cambridge: Cambridge University Press, 2014.

[2] S. Poplack, "The notion of the plural in Puerto Rican Spanish: Competing constraints on (s) deletion." in *Locating Languages in Time and Space*, W. Labov, Ed., 1980, pp. 55–67.

[3] T. Terrell, "La desaparición de /s/ posnuclear a nivel léxico en el habla dominicana," in *Estudios sobre el español del Caribe*, R. Nuñez Cadeño, I. Paez Urdaneta, and J. Guitart, Eds. Caracas: Casa de Bello, 1986, pp. 117–134.

[4] B. M. Rogers, "The state of Spanish /s/ variation in Concepción, Chile: Linguistic and social trends," *Open Linguistics*, vol. 6, no. 1, pp. 132–153, 1 2020.

[5] N. Henriksen, "Patterns of vowel laxing and harmony in Iberian Spanish: Data from production and perception," *Journal of Phonetics*, vol. 63, pp. 106–126, 7 2017.

[6] M. A. Rincón Perez, "An acoustic and perceptual analysis of vowels preceding final /-s/ deletion in the speech of Granada, Spain and Cartagena, Colombia." Ph.D. dissertation, Purdue University, 2015.

[7] M. C. Resnick and R. M. Hammond, "The status of quality and length of Spanish vowels," *Linguistics: An Interdisciplinary Journal of Language Sciences*, vol. 13, no. 156, pp. 79–88, 1975.

[8] R. M. Hammond, "An experimental verification of the phonemic status of open and closed vowels in Spanish," Ph.D. dissertation, Florida Atlantic University, 1978.

[9] K. M. Carlson, "An Acoustic and Perceptual Analysis of Compensatory Processes in Vowels Preceding Deleted Post-Nuclear /s/ in Andalusian Spanish," *Concentric: Studies in Linguistics*, vol. 38, no. 1, pp. 39–67, 2012.

[10] N. Figueroa, "An acoustic and perceptual study of vowels preceding deleted post-nuclear /s/ in Puerto Rican Spanish," in *Hispanic linguistics at the turn of the millennium: papers from the 3rd Hispanic linguistics symposium*, H. Campos, E. Herburger, A. Moreales-Front, and T. J. Walsh, Eds. Somerville, MA: Cascadilla Press, 2000, pp. 66–79.

[11] M. Bolyanatz, "Plural production in Santiago Spanish," Ph.D. dissertation, University of California, Los Angeles, 2017.

[12] M. A. Bolyanatz, "Evidence of Incomplete Neutralization in Chilean Spanish," *Phonetica*, vol. 77, pp. 107–130, 2018.

[13] V. Honsa, "The phonemic systems of Argentinian Spanish," *Hispania*, pp. 275–283, 1965.

[14] R. Bermúdez-Otero, "Cyclicity," in *The Blackwell companion to phonology*, M. van Oostendrop, C. J. Ewen, E. Hume, and K. Rice, Eds. Malden, MA: Wiley-Blackwell, 2010, pp. 2019–2048.

[15] P. Strycharczuk and M. Kohlberger, "Resyllabification Reconsidered: On the Durational Properties of Word-Final /s/ in Spanish," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 7, no. 1, 2016.

[16] V. Freeman and P. De Decker, "Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1211–1223, 2 2021. [Online]. Available: https://doi.org/10.1121/10.0003529

[17] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 3910–3916, 6 2021. [Online]. Available: https://doi.org/10.1121/10.0005132

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," 2017.

[19] P. Boersma and D. Weenik, "Praat: doing phonetics by computer," 2021.

[20] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2022.

[21] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] E. M. Kaisse, "Resyllabification precedes all segmental rules. Evidence from Argentinian Spanish," in *Formal Perspectives on Romance Linguistics: Selected papers from the 28th Linguistic Symposium on Romance Languages (LSRL XXVIII)*, J.-M. Authier, B. Bullock, and L. A. Reed, Eds., 1999, pp. 197–210.

[24] K. Robinson, "The dialectology of syllabification: A review of variation in the Ecuadorian Highlands," *Romance Philology*, vol. 66, pp. 132–153, 2012.

[1] When describing Southern Cone varieties, we opt for the term 'semi-conservative' over conservative since, although not quite as advanced as those from Southern Spain or the Caribbean, research shows increasingly advanced patterns, particularly for metropolitan, Chilean varieties [4].

INTERACTING SOUND SYMBOLIC PATTERNS: THE CASE OF VOWEL SOUND SYMBOLISM IN KOREAN GIVEN NAMES

Lisa Sullivan[1] & Yoonjung Kang[2,3]

[1]University of Manitoba, [2]University of Toronto, [3]University of Toronto Scarborough
lisa.sullivan@umanitoba.ca, yoonjung.kang@utoronto.ca

ABSTRACT

The light-dark vowel pattern in Korean sound symbolic vocabulary runs contrary to the proposed cross-linguistic vowel height symbolism and, therefore, presents an opportunity to explore how cross-linguistic and language specific patterns interact when they operate on the same sounds. Analysis of a corpus of Korean given names suggests that the language-specific pattern overrides the cross-linguistic one and light vowels (which are lower) are more frequently found in female names than dark vowels (which are higher). This study examines the productivity of this pattern by testing if native speakers of Korean extend it to novel names in a nonce-name gendering experiment. The results show that Korean speakers judge the names with /o/ (light but lower) to be more female than those with /u/ (dark but higher), mirroring the findings of the corpus study and the native sound symbolism, but English speakers do not.

Keywords: Korean, sound symbolism, names

1. INTRODUCTION

Sound symbolic patterns occur when a sound or sound property is associated with a particular meaning. Most research in this area uses nonce name experiments (e.g. [1, 2]). However, some recent work has shown that these patterns are attested in areas of the lexicon such as animal names (e.g. [3]), character names (e.g. [4, 5, 6]), given names (e.g. [7, 8, 9, 10, 11, 12, 13]) and brand names (e.g. [14, 15, 16]).

The sound symbolic pattern we focus on in this paper is vowel height [1]. High vowels (e.g. /i u/) are associated with smallness and low vowels (e.g. /æ ɑ/) with largeness. Some suggest an iconic, and potentially universal, connection between the size of the oral cavity and the meaning of larger and smaller size (e.g. [17, 2]). This pattern appears to be present in the phonology of given names, with female names having more high vowels than male names in languages such as English [18, 13], Armenian [8] and Cantonese [12]. Smallness may be associated more with females than males, leading to more high vowels in female than male names.

1.1. Korean Dark-Light Sound Symbolism in Given Names

Despite being a proposed universal, there are counterexamples to the vowel height pattern. One is the dark-light pattern found in Korean sound symbolic vocabulary [19]. Korean has a sound symbolic vocabulary in which certain vowels and consonants convey specific meanings [20, 21, 22, 19]. We focus on the sound symbolism of vowels, which is the main topic of our study. Dark vowels (/e y ʌ u/) are associated with largeness while light vowels (/ɛ ø a o/) are associated with smallness, amongst other meanings [23]. For example, /tultul/ and /toltol/ refer to a large and a small object rolling into a round shape, respectively. This pattern runs opposite the height pattern as light vowels are lower than dark vowels (see Table 1), so its presence suggests that the vowel height pattern may not be present in Korean, or that it is suppressed by the language specific dark-light pattern.

	Front		Central	Back
High	i	y	ɨ	u
Mid	e	ø	ʌ	o
Low	ɛ		a	

Legend
Neutral
Dark
Light

Table 1: Korean Vowels

Previous studies that examined the Korean speakers' association of vowel height and size found mixed results. Taylor and Taylor [24] found that speakers rate nonce words with light vowels as smaller than those with dark vowels. Kwon [23] found that unlike English speakers, Korean speakers chose dark vowel words (with /u/) as more augmentative in meaning more often than light vowel words (with /o/), but this did not hold in /i-ɛ/, /ɨ-a/, and /i-a/ pairs, where the universal height pattern overruled the native pattern. Shinohara and Kawahara [25] found a weaker effect of height

for Korean speakers than Chinese, English, and Japanese speakers, presumably due to the native sound symbolism.

One question this raises is whether the sound symbolic vocabulary pattern extends to other areas of the lexicon. This appears to be the case for names. A corpus analysis of the most popular two syllable Korean newborn names between 2008 and 2020 (F: 974, M: 970) [26] shows that male names tend to have more dark vowels than female names, which is consistent with the light-dark sound symbolic pattern, rather than the vowel height pattern [27]. However, the height pattern still seems to be at play, as /i/ and /ɨ/, neutral vowels in the sound symbolic vocabulary, are more prevalent in female names than male names, suggesting that both patterns are present, but that the language-specific sound symbolism may override the more universal pattern.

1.2. Lexicalized Pattern or Productive Phonological Process?

Even though the height and dark-light patterns are found in the English and Korean name corpora, respectively, this does not reveal if they are productive in speakers' minds. Previous studies [8, 10, 28, 11, 29] indicate that cross-linguistically recurring sound-meaning associations, such as the vowel height pattern, are part of speaker's productive grammar and that they are able to use them to assign gender to nonce names, at least in their native language.

Korean presents an opportunity to explore how cross-linguistic and language specific patterns interact as the height and dark-light patterns operate on the same vowels. If the Korean dark-light pattern is a fossilized historical accident and is not reflected the speaker's productive grammar we should see English and Korean speakers assigning gender to names in the same way. If, on the other hand, the Korean pattern is active in speaker's minds, we would expect Korean, but not English, speakers to use it to assign gender to names, leading to names with dark vowels being considered male more often than those with light vowels. If this is the case, it suggests that speakers are able to learn sound symbolic patterns specific to their language, and apply them to novel situations.

To answer this question, we conducted a "nonce name gendering" experiment using Korean nonce names with the vowels /u/ and /o/ in the second syllable. In the Korean name corpus, /u/, (dark and high), is more common in male names than /o/ (light and mid) is, while the opposite is true for English (Figure 1) [30]. We expect that, if the pattern present in the lexicon is active in speakers' minds, Korean speakers should rate names with /u/ as more male than names with /o/. Conversely, English speakers should rate names with /o/ as more male due to the vowel height pattern.

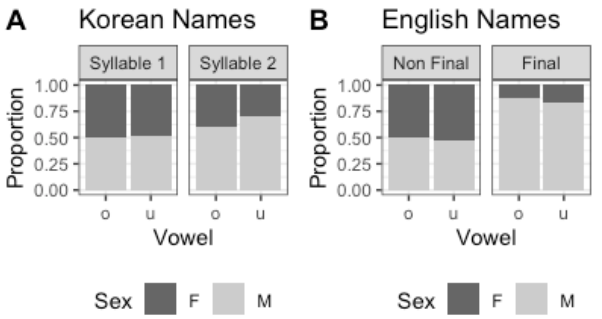


Figure 1: Ratio of male to female names containing /u/ and /o/ in Korean and English

We investigate the alternation in the second syllable as the light-dark pattern is significant on this syllable, but not on the initial one, in the corpus. We focus on /u/ and /o/ because the light-dark status of the front vowel pairs (/e/-/ɛ/ and /y/-/ø/) is less robust than that of the back vowel pairs due to sound changes which have resulted in mergers amongst these vowels [31, 32] and it is not clear how English speakers can distinguish the /a/-/ʌ/ pair.

2. METHOD

2.1. Participants

40 native Korean speakers (28 Female, 12 Male) and 40 native North American English speakers (27 Female, 13 Male) over the age of 18 completed the study. Most participants were under the age of 40. They were recruited through prolific.co (79) and word of mouth (1) and paid $5.00CAD for their time.

2.2. Stimuli

The stimuli consisted of 78 bisyllabic nonce names including 10 target names, 60 fillers and 8 training items. The target names were 5 bisyllabic CVCV nonce name minimal pairs which varied based on their final vowel. One member of the pair contained /u/ (dark) and the other, /o/ (light). The nonce names were not words or names in Korean or English. The stimuli were recorded by a middle-aged male native speaker of the Seoul dialect of Korean. Korean does not have contrastive lexical stress. The stimuli were produced with no obvious stress prominence on either syllable.

2.3. Procedure

The procedure was implemented online using jsPsych [33]. Participants were instructed to imagine that a Korean family with a boy and a girl had moved in next door and they were trying to figure out what each kid's name was. They would hear the name and then be presented with a 6-point Likert-type scale ranging from 'Definitely Female' to 'Definitely Male' which they used to rate how male or female they thought the name was. The order of presentation was randomized by participant.

2.4. Analysis

A full interaction mixed effects linear regression model was constructed using the *lmer*() function from the lme4 package [34] and the lmerTest package [35] in R [36]. Participant ratings were converted to a numerical scale ranging from 1 (definitely female) to 6 (definitely male). Rating was the response variable and vowel type (dark or light) and participant language (English or Korean) were the predictor variables. Random intercepts for participant, and minimal pair were included. Random slopes were not included due to singular fit issues. Language and vowel type were contrast coded such that /o/ and English were coded as -0.5 while /u/ and Korean were coded as 0.5. For a post-hoc analysis of the interaction between language and vowel type, separate models were constructed for each language group with the same structure as the full model except that language was removed as a predictor variable.

3. RESULTS

Figure 2 shows the Korean and English participants' mean ratings for the nonce names across all five minimal pairs. Korean participants' mean rating of names with the dark vowel /u/ was more male than their mean rating of names with the light vowel /o/. English participants' mean ratings for names with /u/ and /o/ are very similar, though names with /o/ have a slightly higher mean rating. The linear regression showed a significant effect of language (Est. = 0.215, *SE* = 0.0937, df = 78, t = 2.295, p = 0.0244) and a significant interaction between language and vowel type (Est. = 0.350, *SE* = 0.648, df = 714, t = 2.989, p = 0.0029). The main effect suggests that Korean participants, on average, gave names more male ratings than English participants.

To understand the interaction between language and vowel type, separate post-hoc models were constructed for each language. The Korean model

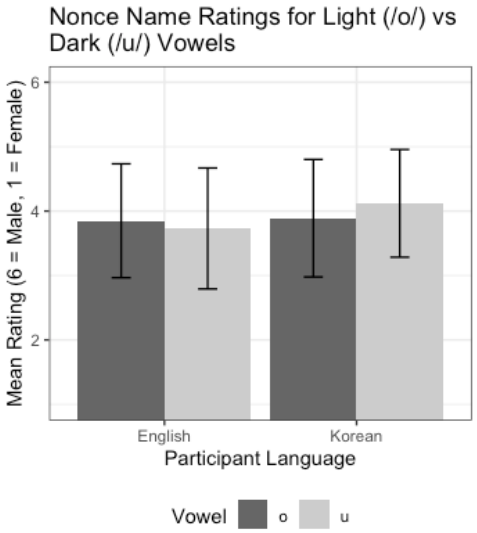


Figure 2: Nonce-name ratings by vowel and participant language. Error bars represent 1 standard deviation above or below the mean.

showed a significant effect of vowel type whereby names with the dark vowel /u/ were rated as more male than those with the light vowel /o/ (Est. = 0.230, *SE* = 0.077, df = 355, t = 2.983, p = 0.003). The English showed a non-significant difference in rating between /o/ and /u/ in the direction of universal height pattern (Est. = -0.120, *SE* = 0.084, df = 355, t = -1.422, p = 0.156). The lack of vowel height effect may be due to the fact that Korean /o/ is phonetically high and may be difficult to distinguish from /u/ for English listeners.

These results indicate that Korean participants were using the dark-light vowel pattern to assign gender to names, but English speakers do not. The main effect of language is due to the fact that English and Korean speakers gave similar ratings to names containing /o/, but Korean speakers gave higher (more male) ratings to those containing /u/, as can be seen in Figure 2.

4. DISCUSSION

The purpose of the experiment was to investigate how native (Korean) vs. non-native (English) language speakers use language-specific and cross-linguistic patterns to assign gender to nonce names using Korean dark-light vowel sound symbolism and cross-linguistic size symbolism as a case study. The results of the name gendering experiment suggest that only native speakers make use of language-specific patterns.

Korean participants rated nonce names with /u/ (dark) as more male than nonce names with /

o/ (light), indicating that Korean speakers use the language specific dark-light pattern when it conflicts with the universal height one. This is consistent with what is found in the corpus of real Korean names [27], suggesting that, like in other languages [11], the sound symbolic patterns found in real names are also actively used by speakers to determine name gender. Furthermore, note that while the universal height pattern is iconically grounded – the smaller oral cavity size of high vowels signals smallness – and can operate without the mediation of the phonological grammar, the Korean light-dark pattern has no natural phonetic motivation and the fact that Koreans are following the light-dark pattern in name gender preference suggests that sound symbolism isn't fully phonetic and hinges on the speaker's language-specific phonological knowledge.

Unlike Korean participants, English participants did not display a difference in their ratings between nonce words with /u/ and those with /o/. This might be unexpected if English speakers are using the height pattern to assign gender to names, since /u/ is higher than /o/, and would be expected to be rated more female. However, as both vowels are phonetically high in Korean, it is possible that there is not sufficient difference in their heights for English speakers to consider them different in terms of size symbolism. This is also the finding of Kwon [23] for English speaker's judgement of size symbolism in Korean sound symbolic words. Regardless, the results indicate that English speakers do not make use of this Korean language-specific pattern.

While the results of this study show that Korean, but not English, speakers use the light-dark pattern, they do not show if Korean participants use the universal height pattern in cases where light-dark symbolism is not at play. In the current study, we were primarily interested in whether or not participants used the light-dark pattern at all, and how this compared between our two language groups. Future work should investigate the potential interaction between the two sound symbolic patterns further by examining alternations throughout the Korean vowel space. In this study we focused on /u/ and /o/ as they are free from neutralizing sound changes and are expected to be reliably distinguished by English listeners. Future studies can examine the rest of the vowel space, where the traditional dark-light pattern is obscured by mergers (/e/ vs. /ɛ/) or neutral vowels (/i/ vs. /ɛ/, /ɨ/ vs. /a/). The findings from the Korean name corpus [27] show that universal height pattern emerges with neutral vowels and we predict the same to hold true in a name gendering experiment.

5. CONCLUSION

This study used a name gendering experiment to investigate sound symbolism in Korean given names and explore how language-specific and cross-linguistic sound symbolic patterns interact. Korean light-dark sound symbolism presented an opportunity to investigate this as it runs somewhat contrary to the cross-linguistic height pattern. Korean participants rated names with /u/ (dark) as more male than names with /o/ (light) whereas English speakers did not display a difference in their ratings between these two vowels, suggesting that native speakers of a language with a language-specific sound symbolic pattern can make use of that pattern to assign meaning to nonce names. Our study shows that sound-meaning associations that are grounded in a particular group of native vocabulary are productively extended to the domain of personal names, which mostly consists of Sino-Korean morphemes. This finding has implications beyond the sound symbolism of names in demonstrating the fluidity of phonological knowledge that "leaks" across different lexical strata.

6. REFERENCES

[1] S. S. Newman, "Further experiments in phonetic symbolism," *The American Journal of Psychology*, vol. 45, no. 1, pp. 53–75, 1933.

[2] E. Sapir, "A study in phonetic symbolism." *Journal of experimental psychology*, vol. 12, no. 3, p. 225, 1929.

[3] B. Berlin, "Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature," *Sound symbolism*, pp. 76–93, 1994.

[4] S.-H. Kim, S. Mun, and E.-S. Ko, "Sound symbolism in the gender of personal names: With focus on the game characters in animal crossing [이름의 성별에 따라 나타나는 음성상징성: 동물의 숲 주민 캐릭터 이름을 중심으로]," *Studies in Phonetics, Phonology, and Morphology* [음성음운형태론연구], vol. 28, no. 1, pp. 33–55, 2022.

[5] S. Shih, J. Ackerman, N. Hermalin, S. Inkelas, H. Jang, J. Johnson, D. Kavitskaya, S. Kawahara, M. Oha, R. L. Starre, and A. Yu, "Cross-linguistic and language-specific sound symbolism: Pokémonastics," 2019, under review.

[6] L. Sutton, "Aliens are just like us: Personal names in the legion of super-heroes," *Names*, vol. 64, no. 2, pp. 109–119, 2016.

[7] T. Ackermann and C. Zimmer, "The sound of gender-correlations of name phonology and gender across languages," *Linguistics*, vol. 59, no. 4, pp.

1143–1177, 2021.
[8] G. Ananthathurai, L. Bradford, A. Derohan, S. Galeazzi, K. Jagani, and Y. Kang, "Sound symbolism of gender in personal names: Western Armenian and Kutchi," in *Summer Phonology Forum 2019*. University of Toronto, 2019.
[9] H. Cho, "A maximum-entropy model of phonotactics for korean male and female names," *Studies in Phonetics, Phonology, and Morphology* [음성음운형태론연구], vol. 27, no. 1, pp. 99–129, 2021.
[10] N. Mohsin, L. Sullivan, and Y. Kang, "Sound symbolism in Urdu first names," in *UTSC Undergraduate Research Poster Forum*. University of Toronto Scarborough, 2019.
[11] L. Sullivan and Y. Kang, "Phonology of gender in English and French given names," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 2124–2128.
[12] K. W. Y. Wong and Y. Kang, "Phonology of gender in Cantonese first names," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberrra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 2129–2133.
[13] S. K. Wright, J. Hay, and T. Bent, "Ladies first? phonology, frequency, and the naming conspiracy," *Linguistics*, vol. 43, no. 3, pp. 531–561, 2005.
[14] G. A. Abel and L. H. Glinert, "Chemotherapy as language: Sound symbolism in cancer medication names," *Social Science & Medicine*, vol. 66, no. 8, pp. 1863–1869, 2008.
[15] R. R. Klink and L. Wu, "The role of position, type, and combination of sound symbolism imbeds in brand names," *Marketing Letters*, vol. 25, no. 1, pp. 13–24, 2014.
[16] L. J. Shrum, T. M. Lowrey, D. Luna, D. B. Lerman, and M. Liu, "Sound symbolism effects across languages: Implications for global brand names," *International Journal of Research in Marketing*, vol. 29, no. 3, pp. 275–279, 2012.
[17] S. Kawahara, "Sound symbolism and theoretical phonology," *Language and Linguistics Compass*, vol. 14, no. 8, 2020.
[18] L. Sullivan, "Phonology of gender in French and English given names," Master's thesis, University of Toronto, 2018.
[19] H.-M. Sohn, *The Korean Language*. Cambridge University Press, 1999.
[20] Y.-m. Y. Cho, "Sound symbolism in Korean," in *Korean Language in Culture and Society*, H.-M. Sohn, Ed. University of Hawai'i Press, 2006, ch. 6, pp. 64–73.
[21] K.-O. Kim, "Sound symbolism in Korean," *Journal of Linguistics*, pp. 67–75, 1977.
[22] J.-S. Lee, *Phonology and sound symbolism of Korean ideophones*. Indiana University, 1992.
[23] N. Kwon, "The natural motivation of sound symbolism," Ph.D. dissertation, University of Queensland, Brisbane, 2015.
[24] I. K. Taylor and M. M. Taylor, "Phonetic symbolism in four unrelated languages." *Canadian Journal of Psychology/ Revue canadienne de psychologie*, vol. 16, no. 4, p. 344, 1962.
[25] K. Shinohara and S. Kawahara, "A cross-linguistic study of sound symbolism: The images of size," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 36, no. 1, 2010, pp. 396–410.
[26] Kimkkikki. (2020) 한국인의 이름 통계 [Korean name statistics]. 출생 이름 랭킹 [birth name ranking]. [Online]. Available: https://koreanname.me/
[27] L. Sullivan and Y. Kang, "Sound symbolism of vowels in Korean given names," 2022, submitted.
[28] D. M. Sidhu, P. M. Pexman, and J. Saint-Aubin, "From the Bob/Kirk effect to the Benoit/éric effect: Testing the mechanism of name sound symbolism in two languages," *Acta Psychologica*, vol. 169, pp. 88–99, 2016.
[29] K. W. Y. Wong and Y. Kang, "The emergence of the frequency code in sound symbolism of gender in Cantonese names," in *LabPhon 17*. University of British Columbia, 2020.
[30] L. Sullivan, "Universal and cross-linguistic gender-based sound symbolism in Korean given names," 2020, unpublished Manuscript.
[31] S.-H. Kim and S. Nam, "Revisiting vowel harmony in Korean sound-symbolic words: A corpus-based quantitative approach [이름의 성별에 따라 나타나는 음성상징성: 동물의 숲 주민 캐릭터 이름을 중심으로]," *The Journal of Studies in Language* [언어연구], vol. 35, no. 3, pp. 309–325, 2019.
[32] N. Kwon, "Iconicity correlated with vowel harmony in Korean ideophones," *Laboratory Phonology*, vol. 9, no. 1, 2018.
[33] J. R. de Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, 2015.
[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[35] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.
[36] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://R-project.org/

Acknowledgements: This work draws on research supported by the Social Sciences and Humanities Research Council of Canada.

PREBOUNDARY LENGTHENING AND ITS KINEMATIC CHARACTERISTICS IN MANDARIN CHINESE IN INTERACTION WITH FOCUS AND LEXICAL TONES

Hongmei Li[1,3], Sahyang Kim[2], Taehong Cho[3]

[1]Department of English, Yanbian University; [2]Hongik University;
[3]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS), Hanyang University
angela.hongmeili@gmail.com, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study investigates preboundary lengthening (PBL) in Mandarin Chinese by exploring kinematic variation of Lip Aperture conditioned by boundary, focus, and lexical tones. Results showed that the lip opening and the following hold duration of CV words were longer phrase-finally, but in interaction with prominence. Under focus, PBL came with slower velocity with no spatial expansion, but under no focus, PBL came with spatial expansion but no slowing-down. This suggests that in the absence of focus-induced hyperarticulation, PBL comes with both temporal and spatial expansion, possibly counteracting a slowing-down effect. PBL also interacted with lexical tones. Compared to simplex Tone1, PBL was much more robust for Tone3 (low-dipping) and Tone4 (falling) with further augmented PBL for Tone3 under focus, presumably to make sufficient room for realizing their tonal complexity. These results imply that PBL is modulated by the phonetically-driven phonological needs for maximizing tonal contrast when licensed by prosodic structure.

Keywords: preboundary lengthening, kinematics, Mandarin Chinese, focus, lexical tones

1. INTRODUCTION

Preboundary lengthening (henceforth PBL) refers to a temporal expansion of phonological units at the right edge of a prosodic domain [4,12,19]. As a near-universal phonetic phenomenon, PBL is unequivocally realized across many languages as temporal lengthening in the acoustic domain, but its detailed articulatory manifestation and its scope are known to be language-specifically modulated by other higher-order linguistic factors [17,18,22,25].

The phonetic realization of PBL is known to be gradient in nature, with its effect strongest when closest to the boundary and more robust before a larger boundary than a smaller one [12,16,17,18,22]. Within the framework of Articulatory Phonology [2,14], these characteristics of PBL have been understood with the π-gesture model [3,4,7]. The π-gesture is a non-tract variable prosodic gesture that does not have a specified constriction degree and location of its own but overlaps with segmental constriction gestures. The π-gesture is assumed to modulate the rate of clock and control the realization of articulatory movements. As a consequence, articulatory gestures are realized with a longer, slower, and in some languages, spatially larger movement under a stronger influence of the π-gesture (i.e., at a stronger boundary) than when not [3,4,15]. In addition, its effect is the strongest at the juncture and gradually wanes across gestures even within a syllable in proportion to its distance from the boundary. The present study investigates the kinematic characteristics and the scope of PBL of monosyllabic CV words in Mandarin Chinese by examining the lip closing and opening gestures at the phrase-final and phrase-medial positions.

The investigation of PBL in Mandarin is particularly interesting, given the accumulating crosslinguistic evidence that the detailed manifestation of PBL is modulated by the prominence system of language [16,18,22,24]. In English, for example, PBL was not only realized on the final syllable regardless of stress, but also on a non-final stressed syllable [24]. PBL was further affected by phrase-level prominence in English, being modulated by the degree of prominence [18]. The interaction between PBL and the language-specific prominence system was also found with Japanese, a lexical pitch-accent language, and with Korean, an edge-prominence language without lexical level prominence, although no further interaction between PBL and phrase-level prominence was found in either language [16,22]. In order to further our understanding of the fine-grained kinematic realization of PBL and its interaction with language-specific prominence system, the current study focuses on how lexical tone system and phrase-level prominence in Mandarin interact to affect articulatory manifestation of PBL.

Mandarin has four lexical tones which are specified with different tonal targets [5,11]: a High-level tone (T1), a Low-High rising tone (T2), a Low-dipping tone (T3), and a High-Low falling tone (T4). There is a close correlation between contour-bearing ability and intrinsic duration of tone [26]: a longer duration is required when a tone has more pitch targets, a greater excursion, or a rising pitch excursion. Considering that each tone has distinctive tonal targets and intrinsic temporal structure, a question arises as to how PBL may be modulated by lexical tones in Mandarin. On the one hand, complex tones with more tonal targets would require more time to manifest phonological contrasts compared to simplex

tones. Thus, one might expect a greater PBL effect for complex tones than for simplex tones. On the other hand, since lexical contrasts should be maintained through distinctive lexical tones both phrase-finally and phrase-medially, both complex and simplex tones may still be produced with a similar degree of PBL when phrase-final relative to when phrase-medial.

The present study also examines the effects of focus-induced prominence on PBL and lexical tones to explore whether and how the assumed interaction between lexical tones and PBL may be further conditioned by prominence. In particular, given that a boundary-related strengthening effect on tonal realization could be masked by the presence of tonal hyperarticulation under focus, testing the focus factor may further inform us whether and how the assumed contrast maximization of tones associated with PBL becomes more or less evident in the presence or absence of focus [7,8,10].

2. METHOD

2.1. Participants and procedures

Twelve native Mandarin speakers (6 females, 6 males, M_{age} = 25.2) participated in this study. They were born and raised in Northern China and had resided in Korea for fewer than 3 years at the time of recording. Two CV sequences (/pa/, /ma/) were recorded across four lexical tones. Each target word was embedded in a carrier sentence that was an answer to a question in a mini dialogue where Boundary (IP-medial vs. IP-final) and Focus (UnFoc vs. Foc) conditions varied, as in Table 1.

Table 1: Examples of target words in carrier sentences. Target words are underlined and italicized, and focused words are in bold. "#" represents an IP boundary. A number after each word indicates the lexical tone of the word (e.g., 1 = Tone1).

Boundary	Focus	Example sentences
IP-medial	UnFoc	A: [**maʊ1** mi1 pa1 pi4 ʂʏŋ4 ma?] Does **Cat** EIGHT win? B: [pu4] # [**ma1** mi1 ***pa1*** pi4 ʂʏŋ4.] No. **Mommy** ***EIGHT*** wins.
	Foc	A: [ma1 mi1 **ta1** pi4 ʂʏŋ4 ma?] Does Mommy **BUILD** win? B: [pu4] # [ma1 mi1 ***pa1*** pi4 ʂʏŋ4.] No. Mommy ***EIGHT*** wins.
IP-final	UnFoc	A: [ni3 ʈʂʰu1 **maʊ1** mi1 pa1 ma?] Do you play **Cat** EIGHT? B: [pu4] # [ʊɔ3 ʈʂʰu1 **ma1** mi1 ***pa1***] # [pi4 ʂʏŋ4 pa?] No. I play **Mommy** ***EIGHT***. Must win, right?
	Foc	A: [ni3 ʈʂʰu1 ma1 mi1 **ta1** ma?] Do you play Mommy **BUILD**? B: [pu4] # [ʊɔ3 ʈʂʰu1 ma1 mi1 ***pa1***] # [pi4 ʂʏŋ4 pa?] No. I play Mommy ***EIGHT***. Must win, right?

During the recording, the speakers were presented with a mini dialogue written in Chinese on the top of a computer screen, along with two picture cards (Fig.1). Next to the picture on each card, a monosyllabic word was displayed in Chinese. The target word (/ma/ or /pa/) was always on a card with a picture of a mommy, marked by a red seal. The combination of a picture and the target word yielded critical sequences such as /maʊ1mi1 pa1/ ('cat eight'), /ma1mi1 pa1/ ('mommy eight'), and /ma1mi1 ta1/ ('mommy build'). A target word (e.g., pa1) was followed by either an IP boundary (IP-final condition) or an IP-medial word boundary (IP-medial condition). In the focused condition, the target word on a mommy card was contrasted with a word on another mommy card, given in the prompt question (e.g., 'mommy build /ta1/' vs. 'mommy eight' /pa1/, Fig 1b). In the unfocused condition, the picture words were contrasted in the question-answer pair (e.g., 'cat eight' /pa1/ vs. 'mommy eight' /pa1/, Fig.1a), such that the target word was unfocused.

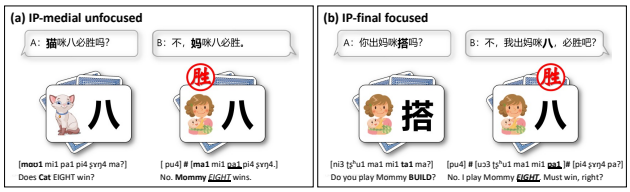


Figure 1: An example of the visual prompts used in the designed card game. Displayed in the right mommy card is a target word (*pa1*, 'eight').

The experiment was conducted at Hanyang Institute for Phonetics and Cognitive Sciences in Seoul. The articulatory data were obtained by using EMA (Electromagnetic Articulograph, Carstens AG501) with a sampling rate of 250 Hz. Acoustic data were recorded simultaneously with the articulatory data, with a Tascam US-4x4 audio interface and a SHURE KSM44A microphone at a sampling rate of 44 kHz. In total, 1920 tokens were collected: 2 CV sequences x 4 lexical tones x 2 boundaries x 2 focus conditions x 5 repetitions x 12 speakers. Thirty-five tokens with unintended prosodic renditions were excluded, leaving 1885 tokens for further data analysis.

2.2. Measurements and Statistical analysis

Articulatory data of lip movements were analyzed using a MATLAB-based software MVIEW [23]. Lip aperture was calculated using the Euclidean distance between the two sensors attached to the upper and lower lips. The gestural landmarks were identified based on the velocity profile, and seven kinematic measures for lip closing and opening gestures of CV were taken as follows (schematized in Fig. 2).

(1) **movement duration:** duration from onset to target (ms)
(2) **hold duration:** duration from target to release (ms)
(3) **formation duration**: duration from onset to release (ms)
(4) **time-to-pkvel**: duration from onset to peak velocity (ms)
(5) **movement displacement**: displacement from onset to target (mm)
(6) **maximum displacement**: displacement from onset to max constriction (mm)
(7) **peak velocity**: the highest velocity during lip movements (cm/sec)

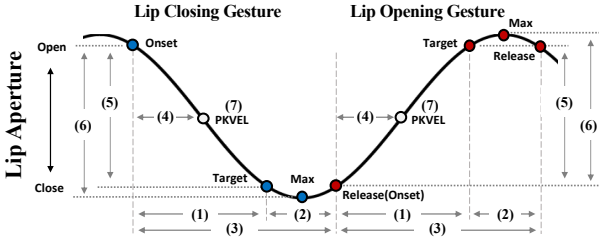


Figure 2: Schematized trajectories of lip closing and opening movements with kinematic landmarks and temporal-spatial measures.

A series of linear mixed-effects models were conducted by using lme4 [1] and lmerTest [20] packages. The seven kinematic measures were taken as dependent variables. Boundary, Focus, Tone, and their interactions were fixed factors. Boundary (IP-medial, IP-final) and Focus (unfocused, focused) were contrast-coded, while Tone (T1, T2, T3, T4) was treatment-coded. (The reference levels are underlined). The random structure included a random intercept by Speaker with random slopes by Boundary, Tone, and Focus and a random intercept by Item. The maximal random effect structure was employed if the model converged. When there were interactions between factors, pairwise comparisons of the posthoc tests were conducted. For the purpose of the present study, only the main effect of Boundary and its interactions with Focus and Tone are reported.

3. RESULTS

3.1. Articulatory characteristics of Mandarin PBL

The PBL effects were significant in almost all measures of the lip opening gesture. The lip opening gesture was longer in duration (in all temporal measures but *time-to-pkvel*), larger in displacement, and slower in peak velocity in the IP-final than in the IP-medial positions.

In contrast, none of the temporal measures showed any significant PBL effects for the lip closing gesture (although there was a trend effect in *total duration*). The lip aperture was significantly smaller IP-finally than IP-medially for both *movement* and *maximum displacement*. *Peak velocity*, however, was slower IP-finally than IP-medially for the lip closing gesture, as has been found with the lip opening gesture.

3.2. Interaction between PBL and prominence

As for the lip opening gesture, significant Boundary x Focus interactions were found in *movement duration* (Fig.3a) and *formation duration* measures. These interactions were due to the fact that the PBL effect was greater in the unfocused condition (β=-14.93, p<.001) than in the focused condition (β=-9.60, p<.001). No interaction was found in *time-to-pkvel* (Fig.3b).

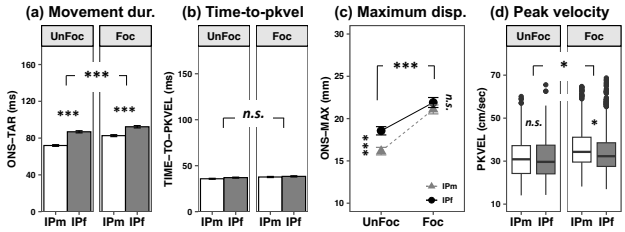


Figure 3: PBL x Prominence interactions for the lip opening gesture. Error bars show standard errors. (*** p<.001; * p<.05)

In the spatial dimension, a significant Boundary x Focus interaction was found in *maximum displacement* (Fig.3c) and in *movement displacement*, which was attributable to the fact that the larger displacement due to Boundary was only observed in the unfocused condition (β=-2.45, p<.001).

Peak velocity also showed a robust Boundary x Focus interaction, which may arise from the fact that the significantly slower movement in the IP-final than IP-medial position was only observed under focus (β=1.68, p<.05; Fig.3d).

As for the lip closing gesture, a trend interaction effect was found in *formation duration* only, which may be due to the fact that the significant PBL effect was only detected in the unfocused condition (β=-4.58, p<.05). No other significant interaction was found.

3.3. Interaction between PBL and lexical tone

Each of the three (potentially) complex tones, that is, T2 (a Low-High rising tone), T3 (a Low-dipping tone), and T4 (a High-Low falling tone) was compared with a simplex, High-level tone (T1) to examine the interaction between PBL and the lexical tone.

As for the lip opening gesture, there was no significant interaction between Boundary and Tone for T2, either in the temporal or in the spatial dimension. It means that T1 and T2 did not differ from each other in terms of their Boundary effects.

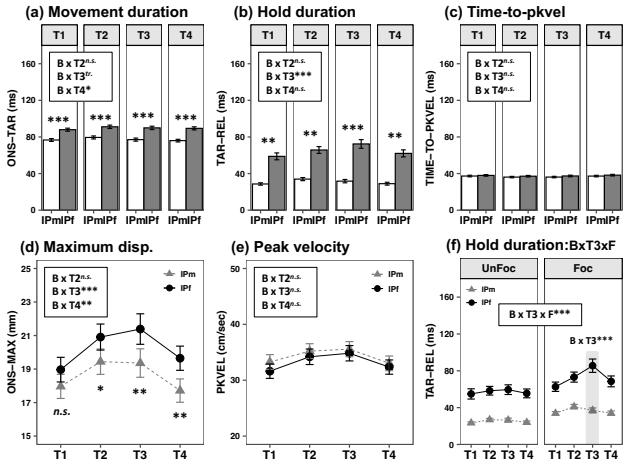


Figure 4: PBL x Tone interactions for the lip opening gesture. Error bars show standard errors. (*** p<.001; ** p<.01; * p<.05; *tr.* .05<p<.08)

T3 showed a significant Boundary x Tone interaction in *hold duration* (Fig. 4b), which was mainly due to the fact that the PBL effect (i.e., longer *hold duration* in the IP-final than in the IP-medial position) was significantly greater for T3 (β=-40.90, p<.001) than for T1 (β=-30.16, p<.01). This effect further interacted with Focus (β=19.27, p<.001; Fig.4f), showing that the Boundary x Tone interaction found with T3 was significant only in the focused condition. Presumably due to the extreme PBL effect of T3 in *hold duration*, similar Boundary x Tone interaction was also found in *formation duration*. However, no interaction was detected in *time-to-pkvel* as shown in Fig.4c.

T4 showed a significant interaction between Boundary and Tone in *movement duration* (Fig. 4a). The posthoc tests revealed that the interaction may arise from

the fact that the PBL effect was greater in T4 (β=-13.32, p<.001) than in T1 (β=-11.15, p<.001). Significant two-way interaction was also found in *formation duration* but not in *hold duration* and *time-to-pkvel*.

In the spatial dimension for the lip opening gesture, there were significant Boundary x Tone interactions for T3 and T4 in both *movement displacement* and *maximum displacement*. As shown in Fig.4d, the interactions stemmed from the fact that lip opening was larger in the IP-final position than in the IP-medial position for T3 and T4, but that this boundary effect was not observed for T1.

As for *peak velocity* (Fig.4e), no interaction between Boundary and Tone was found in any tone comparisons.

For the lip closing gesture, significant Boundary x Tone interactions were found in *hold duration* and *formation duration* for T2, which may be contributable to the fact that the PBL effect was more reduced in T2 than in T1. There was a Boundary x Tone interaction found in *time-to-pkvel* for T4. No other interaction was found.

4. SUMMARY AND DISCUSSION

A basic finding of our study is that preboundary lengthening (PBL) in Mandarin Chinese was found for the (C-to-V) lip opening that was proximal to the boundary but not for the lip closing gesture for the onset C which was distal. This proximity effect could be accounted for the π-gesture model [3,4,7,15], in that its influence was stronger when the gesture was closer to the boundary. The lip opening gesture showed a *longer*, *larger*, and *slower* movement in the IP-final than in the IP-medial position. The slower movement could also be attributed to a clock-slowing effect of the π-gesture. Interestingly, however, *time-to-pkvel* was not necessarily longer associated with PBL, although it is often assumed to be influenced by the π-gesture. Instead, *hold duration* (the duration of the plateau-like portion after the target attainment of the lip opening gesture) contributed significantly to the PBL. As for the lip closing gesture preceding the opening gesture, there was no sign of PBL, though its movement itself was still *slower* in velocity but *smaller* in displacement which could counteract the lengthening effect (the smaller the displacement, the shorter the duration).

The PBL effect in Mandarin further interacted with focus-induced prominence. The boundary-induced temporal enhancement was greater in the unfocused than in the focused condition, and the spatial enhancement was only found when the target word was not focused. Results are in line with the findings in English [18], which showed the most robust PBL in the least prominent condition. These findings seem to suggest that when gestures are locally hyperarticulated by focus [7,10,21], and hence already spatio-temporally expanded enough, there is no sufficient room left to expand the gesture further to mark the end of a phrase boundary (a kind of ceiling effect).

What is also interesting is that the phrase-final slowing down was observed only in the focused condition, which yielded longer and slower (but not larger) movement under focus, unlike longer and larger phrase-final movement found in the absence of focus. This may indicate that speakers employ multiple strategies to achieve the prosodic goal of boundary marking, instead of utilizing one single kinematic parameter to encode linguistic structures [3,6,24].

Most crucially for the purpose of the present study, results showed that the PBL-related articulatory manifestation was differentially modulated by the lexical tone of the phrase-final word. The High-level tone (T1) and the rising tone (T2) were comparable in their PBL effects in both the temporal and spatial dimensions. In contrast, the Low-dipping tone (T3) and the falling tone (T4) showed Boundary x Tone interactions when compared with T1. In the temporal dimension, T3 and T4 showed greater PBL effects than T1. Recall that both T3 (low-dipping tone) and T4 (falling tone) have a low tonal target, and it is physiologically demanding to fully attain the underlying low pitch. Due to restricted time in the phrase-medial position, T3 and T4 must have been temporally suppressed and thus implemented with a short duration on the surface, which resulted in no durational difference between T1 and T3/T4 in the phrase-medial position. In the phrase-final position, however, both T3 and T4 are more robustly lengthened compared to T1, presumably to realize their tonal targets fully in the position. This pattern is notable particularly for T3. As it is physiologically challenging to continuously maintain a low target with vocal fold vibration, the pitch goes up to some extent after reaching the low target, such that vocal folds may reach a more comfortable equilibrium state. As a result, a hidden rising component of the dipping tone is manifested when more time is allowed in the phrase-final position, which makes T3 manifest like a complex contour tone. It is also notable that the lengthening of *hold duration* for T3 was observed under focus. This seems to indicate that the underlying tonal features of T3 became visible and fully implemented phrase-finally, at least partially due to local hyperarticulation to achieve tonal contrast maximization [8,9,10,21].

In the spatial dimension (in the tonal space), T3 and T4 showed larger displacement in the phrase-final than in the phrase-medial position, which was not observed with T1. This is also consistent with the full realization of L tone target in the phrase-final position. That is, with sufficient time, the jaw is likely to be lowered more, which further aided the larynx to reach the Low target in both T3 and T4. Taken together, the tonal targets for T3 and T4 were fully realized with longer duration and accompanied by larger spatial magnitude phrase-finally than phrase-medially.

In conclusion, our study shows that although PBL in Mandarin follows the cross-linguistically applicable patterns, it is modulated by higher-level prominence and

lexical tones. From the point of view of the Articulatory Phonology, tone can also be deemed as laryngeal gesture [13]. In this context, our results suggest that segmental constriction gestures are coordinated with tone gestures and prosodic gestures (e.g., the π-gesture), and they are all activated to regulate the spatio-temporal kinematics of speech production. Results further support the view that low-level fine phonetic details can be fine-tuned by higher linguistic structures of a given language in a language-specific way.

5. ACKNOWLEDGEMENT

We thank the Mandarin Chinese speakers for their participation and the reviewers for the helpful comments. This work was supported by the Ministry of Education of the Republic of Korean and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884 awarded to TC (PI) and SK (Co-PI)).

6. REFERENCES

[1] Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint.arXiv:1506.04967*.

[2] Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. *Mind as motion: Explorations in the dynamics of cognition*, *175*, 194.

[3] Byrd, D., & Krivokapić, J. (2021). Cracking prosody in articulatory phonology. *Annual Review of Linguistics*, *7*(1), 31-53.

[4] Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149-180.

[5] Chao, Y. R. (1965). A grammar of spoken Chinese.

[6] Cho, T. (2006). Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. *Laboratory phonology*, *8*, 519-548.

[7] Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, *10*(3), 120-141.

[8] Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121-157.

[9] Choi, J., Kim, S., & Cho, T. (2016). Phonetic encoding of coda voicing contrast under different focus conditions in L1 vs. L2 English. *Frontiers in psychology*, *7*, 624.

[10] De Jong, K. (2004). Stress, lexical focus, and. segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, *32*(4), 493-516.

[11] Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.

[12] Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *the Journal of the Acoustical Society of America*, *89*(1), 369-382.

[13] Gao, M. (2008). Tonal alignment in Mandarin Chinese: An articulatory phonology account. *Unpublished Doctoral Dissertation (Linguistics), Yale University, CT.*

[14] Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. *Action to language via the mirror neuron system*, 215-249.

[15] Iskarous, K., & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. *Journal of Phonetics*, *95*, 101195.

[16] Jang, J., & Katsika, A. (2020). The amount and scope of phrase-final lengthening in Seoul Korean. In *Proceedings of the 10th International Conference on Speech Prosody, Tokyo, Japan* (pp. 270-274).

[17] Kim, J. J., Baek, Y., Cho, T., & Kim, S. (2019). Preboundary lengthening and boundary-related spatial expansion in Korean. In *Proceedings of the 19th International Congress of Phonetic Sciences*.

[18] Kim, S., Jang, J., & Cho, T. (2017). Articulatory. characteristics of preboundary lengthening in interaction with prominence on tri-syllabic words in American English. *The Journal of the Acoustical Society of America*, *142*(4), EL362-EL368.

[19] Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208-1221.

[20] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

[21] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403-439). Springer, Dordrecht.

[22] Seo, J., Kim, S., Kubozono, H., & Cho, T. (2019). Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter?. *The Journal of the Acoustical Society of America*, *146*(3), 1817-1823.

[23] Tiede, M. (2010). MVIEW: Multi-channel. visualization application for displaying dynamic sensor movements. *development*.

[24] Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445-472.

[25] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of phonetics*, *27*(1), 55-105.

[26] Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. *The handbook of Chinese linguistics*, 443-464.

OBSTRUENT VOICING AND LARYNGEAL FEATURES IN ARABIC

Abdulrahman Dallak[1,2], Ghada Khattab[2], Jalal Al-Tamimi[3]

[1]Jazan University, KSA, [2]Newcastle University, UK, [3]Université Paris Cité, LLF, CNRS, France
adallak@jazanu.edu.sa, ghada.khattab@ncl.ac.uk, jalal.al-tamimi@u-paris.fr

ABSTRACT

The paper investigates the laryngeal specification of obstruents in Jazani Arabic and whether fundamental frequency ($f0$) interacts with consonant voicing. It presents an acoustic analysis of voice onset time (VOT) and $f0$ perturbations. It is hypothesized that $f0$ perturbations should be predictable from the patterning of VOT in that voicing lead should induce a lowering effect on $f0$ and long lag should induce an increasing effect on $f0$. Twenty Jazani adult speakers produced target words with voiced and voiceless stops in controlled carrier sentences. VOT and $f0$ estimates were measured, the latter dynamically from the midpoint of the closure to the midpoint of the vowel at 10 ms intervals. The results show that VOT patterns in Arabic signal two distinct laryngeal specifications and that $f0$ perturbations appear to be predictable from the patterning of VOT. The results are then discussed in light of Laryngeal Realism and how they inform phonological representation.

Keywords: Jazani Arabic, voicing contrast, Laryngeal Realism, $f0$ perturbation

1. INTRODUCTION

It has long been observed that voiced stops /b,d,g/ in voicing languages such as Russian and Dutch are characterised by robust voicing lead whilst voiceless stops /p,t,k/ in aspirating languages such as English and German are characterised by robust long lag [1, 2]. The former implicates the specification of [voice] on the voiced series whilst the voiceless series in the same (voicing) language are unspecified. For the latter case, this implicates the specification of [spread glottis] on the voiceless series whilst the voiced series are unspecified. This observation is made formally explicitly in the so-called Laryngeal Realism (LR) [3, 4]; an experimental approach mapping VOT realisations into phonological features.

1.1. Laryngeal feature beyond VOT

While VOT patterning has taken centre stage when it comes to its role in laryngeal representation, $f0$ perturbation has been argued to also participate in the specification of laryngeal contrasts. In fact, recent studies [5] have shown that some languages appear to replace the VOT distinction with other emerging cues such as $f0$ and closure duration to maintain the laryngeal contrast. According to [6], the stops in Swiss German are realised as lenis stops [b̥,d̥], fortis unaspirated stops [p,t], and aspirated stops [p^h,t^h]. Following an investigation of the voicing contrast using VOT, stop closure and $f0$, the authors found that the (voicing) contrast in Swiss German stops is primarily signalled by closure duration for all the three categories, but not VOT. They reported that the unaspirated categories are accompanied by $f0$ raising, resembling that of prototypical voiceless stops. This shows that while VOT did not signal the voicing contrast in Swiss German stops, closure duration and $f0$ exhibited a robust pattern, reliably differentiating the three categories.

[5] examined the role of $f0$ in French and Italian (voicing languages) where the voiced stops are specified for [voice] and hence expected to induce a lowering effect on $f0$ while the voiceless stops are unspecified and expected to exhibit an inert pattern similar to the nasal baseline. The results depicted that $f0$ during the closure phase of the voiced stops was statistically more lowered than the nasal baseline in both languages, and following voiceless stops $f0$ was significantly raised in both languages. But what was unexpected is that $f0$ at the vowel onset and even midpoint was higher following the voiceless series given that French and Italian are both voicing languages, and based on LR, their voiceless series are laryngeally unspecified. Even more surprisingly, the authors reported that $f0$ at the vowel onset following voiced stops was statistically indistinguishable from the nasal baseline. Kirby et al. attribute the raising pattern of $f0$ after the unspecified voiceless stops to aerodynamic factors which inhibit phonation and which are independent from VOT and segment phonological voicing

In relation to Arabic, while the dialects surveyed in the literature such as Lebanese [7], Qatari [8], Najdi [9] etc. vary on what is encoded for the

voiceless series, they are in agreement when it comes to the voiced series. That is, voiceless stops in Arabic have been observed to exhibit positive VOT with a varying degree of length within the aspiration range whilst voiced stops have been observed to exhibit voicing lead [7, 8, 9]. According to LR, Arabic stops appear to be over-specified. Both voicing categories of stops are specified simultaneously; [voice] for the voiced series and [spread glottis/tense] for the voiceless series [7, 8, 9, 10]. But whether both VOT and $f0$ exhibit a similar (or dichotomous) pattern and can be predictable from each other in Arabic is the focus of the present study.

1.2. [tense] vs. [spread glottis] characteristics

The feature [tense] was proposed in [11] to account for the aspiration contrast with an eye on the participation of the other cues in signaling the voicing/aspiration contrast [12]. [tense] is not necessarily accompanied by long VOT but explicitly presupposes that the voicing contrast is still maintained by other acoustic cues such as burst duration, intensity, voicing%, etc. In contrast, [spread glottis] was first proposed to signal the aspiration contrast per se [13]. The scope of each feature is still debatable among phoneticians and phonologists [2, 14, 15, 16], but it is crucial to highlight that [tense] does not necessarily implicate (robust) aspiration.

2. THE PRESENT STUDY

Given (i) the dichotomous pattern of VOT and $f0$ in Swiss German, French, and Italian, (ii) the unexpected raising pattern of $f0$ following lenis and unaspirated stops in Swiss German, and (iii) the unexpected raising pattern of $f0$ following the unspecified voiceless stops in voicing languages (i.e., French and Italian), the present study aims to investigate the mapping between these two acoustic correlates in the Jazani Arabic dialect to assess whether both VOT and $f0$ exhibit a similar pattern and can be predictable from each other. In order to examine the nuances of $f0$ pattern, it is crucial to utilise a nasal baseline as was done in [6, 5, 17].

2.1. Methods

2.1.1. Participants

Twenty male subjects from Jazan (Saudi Arabia), aged 22 to 30 and with no reported speech or hearing disorders, were recruited for the study. The subjects' speech was recorded using Tascam Dr-05 Linear PCM in a sound-proof booth. Subjects were asked to read the sentences out loud twice while being recorded. The recordings were digitised at 44.1 kHz sampling rate, with 16-bit quantization and in a mono channel.

2.1.2. Materials

Disyllabic words containing the target sounds /b,d,g,t,k/ and /m/ were inserted in varying natural sentences. /m/ was used as nasal baseline for examining consonant voicing effect on $f0$ [6]. The study only looked at VOT and $f0$ patterns of post-vocalic word-initial stops in three vowel contexts /i:,a:,u:/. This experiment generated 600 tokens for the analysis (20 subj$\times$ 3 vowel $\times$ 2 rep $\times$ 5 cons).

2.1.3. Acoustic analysis

Data were manually transcribed using a newly developed transliteration system for Arabic (ATR), which was then forced aligned using the WebMAUS Arabic via the Munich Automatic Segmentation (MAUS) [18] as in Fig 1. Two acoustic correlates were measured in this experiment: (i) VOT, and (ii) $f0$. VOT was manually measured for post-vocalic word-initial stops from the onset of the release burst to onset of the vowel using visual inspection of oscillogram and broadband spectrograms [19]. The onset of the release burst was identified by an abrupt vertical "spike" in the spectrogram following stop closure. The onset of the vowel was identified by the first vertical striations on the spectrogram. For voiced stops, VOT represents closure plus release duration. $f0$ estimates were measured dynamically from the midpoint of the closure (even for the nasal) to the midpoint of the vowel at 10 ms intervals using `Praat`'s version of `VoiceSauce` [20] For voiceless stops, $f0$ estimates were measured from vowel onset to midpoint.

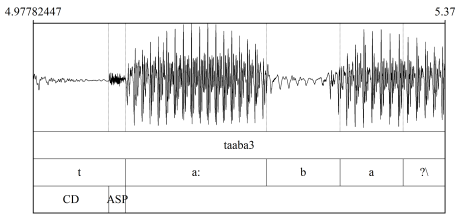


Figure 1: Schematic TextGrid showing how VOT was measured for /ta:/baʔ/ 'followed'

2.1.4. Statistical analysis

Statistical analysis was carried out in `R` [21]. We modelled the difference between voiced vs. voiceless stops using estimation statistics [22] based on median difference, 95% CI, and effect size.

The CIs coupled with effect size are derived from 5000 bootstrap resampling, bias-corrected and accelerated. The effect size used Cohen's *d*. For modelling $f0$ contour, we used a generalised additive mixed model (GAMM) using `mgcv` [23]. It was fitted to the (Hz) value of $f0$ with voicing by phoneme interaction as parametric term, and following the recommendation of [24] for random effect. The results reported here are for the context /aː/. The model used 'scat' family [25], and accounted for autocorrelated residuals. GAMM graphs were produced using `tidyverse` [26], and `tidymv` [27].

3. RESULTS

3.1. VOT MODEL

Fig. 2 presents an estimation plot using two different y-axes. The y-axis on the left side represents the VOT value in ms whilst that on the right side represents effect size of the median difference Δ between the two voicing categories. The grey curve indicates the resampled distribution of Δ, median, and 95% CI, given the observed data. The results

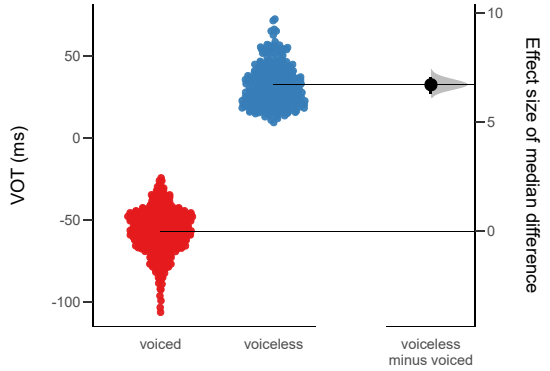


Figure 2: Estimation plot showing the distribution of each voicing category coupled with effect size.

show that voiced stops exhibited an averaged VOT value of -75 ms ([b]= -57.9, [d]= -56.9, [g]= -56.5). In contrast, the voiceless series exhibited an averaged VOT value of 32.5 ms ([t]=31.0, [k]=34.1). The estimated difference between both categories is 90.5 ms [95% CI 87.6, 93.6], exhibiting a robust effect size d = 6.7 [95% CI 6.3, 7.06]. More specifically, for the difference between [t]-[d], the results exhibited a difference value of 87.9 ms [95% CI 84.6, 91.6]. For [k]-[g], the estimated difference was 90.5 ms [95% CI 87.6, 93.6].

3.2. $f0$ CONTOUR MODEL

Table 1 presents the GAM model summary of $f0$. It shows that voiced stops induced a significantly reduced pattern of -19.45 Hz averaged across all the intervals measured relative to the intercept which represents the nasal baseline /m/. For the voiceless stops, the model shows that voiceless stops induced a statistically non-significant pattern of -8.29 Hz averaged across all the intervals measured relative to the nasal baseline.

Predictors	Estimates
Intercept (nasal)	152.57^{***}
	(8.65)
voiced.stop	-19.45^{**}
	(6.21)
voiceless.stop	-8.29
	(6.41)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Summary of the GAM model fitted to $f0$

Fig.3 presents the estimates of $f0$ contour for voiced stops and the nasal baseline along with the difference curve. It shows that voiced stops induced a robust reduced pattern throughout the measured intervals from closure midpoint to vowel midpoint.

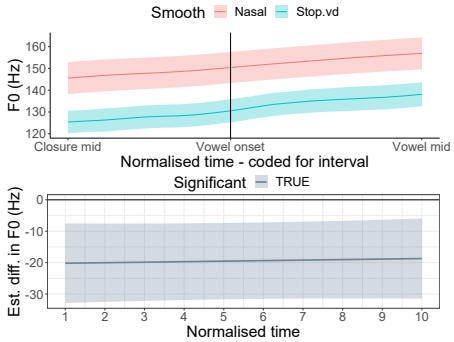


Figure 3: Estimated $f0$ splines for nasal baseline and vd stops coupled with a difference curve

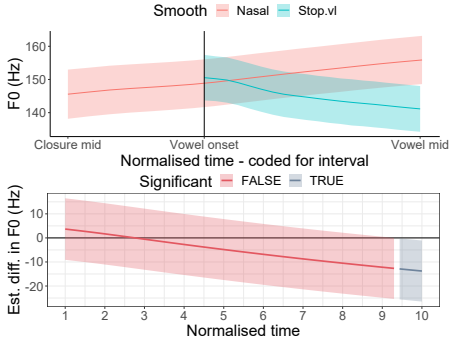


Figure 4: Estimated $f0$ splines for nasal baseline and vl stops coupled with a difference curve

Fig. 4 presents the estimates of $f0$ contour for voiceless stops and the nasal baseline along with the difference curve from vowel onset to midpoint. It shows that while the voiceless stops induced a raising pattern at vowel onset, this raising pattern is weak and negligible. Crucially, the difference curve

presents no evidence for significant differences between the two trajectories except toward the vowel midpoint.

4. DISCUSSION

This study investigated two acoustic measures, VOT and $f0$ contour, to assess the laryngeal specification of stops in the Jazani Arabic dialect along with exploring whether VOT and $f0$ exhibit a predictable mapping pattern, given the cases addressed briefly in Introduction.

4.1. VOT and LR

According to the predictions of LR, robust voicing lead implicates the specification of [voice] on the voiced stops whilst long lag implicates the specification of [spread glottis] on the voiceless stops. The results show that voiced stops exhibited an average VOT value of -75 ms, congruent with the available literature on Arabic [7, 8, 9]. For the voiceless stops, the present results show that voiceless stops exhibited an average VOT value of 32.5 ms. Compared to the pattern of voiceless stops in other Arabic dialects such as Najdi and Qatari [9, 8], respectively, this observed value constitutes a short VOT value for the voiceless series. More specifically, in Najdi, voiceless stops showed a mean VOT value of 47.6 ms. Likewise, voiceless stops in Qatari Arabic showed a value of 55 ms. This long positive VOT values for Najdi and Qatari Arabic are similar to that of the aspirating languages such as German 67.33 ms [2] and English 56.05 ms [28].

The observed value of the voiceless series in the present dialect is instead relatively similar to the voiceless series in the Lebanese dialect, which is 28.62 ms [7], and that of the unspecified (lenis) stops in aspirating languages such as English (25 ms) [28].

Taken together (i) the absence of robust long lag on the voiceless series and (ii) the null pattern exhibited by the voiceless series on $f0$ contour, the present results demonstrate that voiceless stops in the current dialect are not specified with [spread glottis]. While this suggests that voiceless stops are unspecified, this paper is an excerpt from a larger project [29] where robust evidence is presented showing active participation of other cues in signalling the voicing contrast in the present dialect such as voicing%, harmonics, intensity, burst duration, etc. Thus, we propose that voiceless stops in the present dialect are specified with [tense], which in the current context indicates that the voicing contrast is upheld by other acoustic cues beyond VOT (and $f0$). The fact that researchers working within the LR tradition use [spread glottis] to signal robust aspiration while others [2] use [tense] to signal the same manifestation reveals that the scope of each feature is still controversial. In this study, we deviate from [spread glottis] due to the absence of robust aspiration and adopt [tense] due to the active participation of the other cues in signalling the voicing contrast (cf.[14, 12]).

4.2. $f0$ contour and LR

Given the specification of [voice] on the voiced series, it follows naturally that voiced stops should induce a robust lowering pattern compared to the nasal baseline [6, 5]. This robust pattern is observed in $f0$ contour results where the voiced stops exhibit a significant lowering pattern throughout all the measured intervals. For the voiceless stops, the pattern observed rules out the specification of [spread glottis] on the voiceless series; otherwise [spread glottis] should induce significantly higher $f0$ contour in relation to the baseline [1, 9]. Similar to the phonological mapping of VOT, the pattern exhibited by the voiceless stops on the $f0$ contour largely indicates a null effect and specification. This indicates that VOT and $f0$ (contour) exhibit a matching pattern and appear to be predictable from each other at least for the present dialect. This suggests that they provide evidence for a laryngeal matching specification. This is in line with the predictions of LR, which appear to be language-specific.

5. CONCLUSION

This study examined the acoustic realisation of voicing distinction in stops in the Jazani Arabic dialect. It showed that voiced stops are specified as [voiced], and voiceless stops are unspecified or at best specified with a lower numerical value for [tense] [1] building on an ongoing project [29]. It also showed that VOT and $f0$ contour exhibit a matching pattern given the appropriate feature specified for the voiceless series.

6. ACKNOWLEDGEMENTS

For AD, we acknowledge Jazan University for PhD scholarship. For JAT, this work is partially supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083), contributing to the IdEx Université Paris Cité - ANR-18-IDEX-0001.

7. REFERENCES

[1] J. Beckman, M. Jessen, and C. Ringen, "Empirical evidence for laryngeal features: Aspirating vs. true

voice languages," *Journal of linguistics*, vol. 49, no. 2, pp. 259–284, 2013.
[2] M. Jessen, "Phonetics and phonology of tense and lax obstruents in german," *Phonetics and Phonology of Tense and Lax Obstruents in German*, pp. 1–414, 1998.
[3] G. K. Iverson and J. C. Salmons, "Aspiration and laryngeal representation in germanic," *Phonology*, vol. 12, no. 3, pp. 369–396, 1995.
[4] P. Honeybone, "Diachronic evidence in segmental phonology: the case of obstruent laryngeal specifications," *The internal organization of phonological segments*, vol. 319, p. 54, 2005.
[5] J. P. Kirby and D. R. Ladd, "Effects of obstruent voicing on vowel f0: Evidence from 'true voicing' languages," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2400–2411, 2016.
[6] D. R. Ladd and S. Schmid, "Obstruent voicing effects on f0, but without voicing: Phonetic correlates of swiss german lenis, fortis, and aspirated stops," *Journal of Phonetics*, vol. 71, pp. 229–248, 2018.
[7] J. Al-Tamimi and G. Khattab, "Acoustic correlates of the voicing contrast in lebanese arabic singleton and geminate stops," *Journal of Phonetics*, vol. 71, pp. 306–325, 2018.
[8] V. Kulikov, "Laryngeal contrast in qatari arabic: Effect of speaking rate on voice onset time," *Phonetica*, vol. 77, no. 3, pp. 163–185, 2020.
[9] N. Al-Gamdi, "Voicing contrast in najdi arabic stops: Implications for laryngeal realism," Ph.D. dissertation, Newcastle University, 2022.
[10] N. Al-Ghamdi, J. Al-Tamimi, and G. Khattab, "The acoustic properties of laryngeal contrast in najdi arabic initial stops," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, S. Calhoun, P. Escudero, M. Tabain, and . P. Warren, Eds., Melbourne, Australia; Canberra, Australia, 2019, pp. 2051–2055.
[11] R. Jakobson, C. G. Fant, and M. Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press, 1952.
[12] H. Kim, G. N. Clements, A. Rialland, R. Ridouane, and H. van der Hulst, "The feature [tense]," *Features in phonetics and phonology: Unpublished work from George N. Clements and his colleagues*, pp. 159–178, 2015.
[13] C.-W. Kim, "A theory of aspiration," *Phonetica*, vol. 21, no. 2, pp. 107–116, 1970.
[14] R. Ridouane, G. N. Clements, and R. Khatiwada, "Language-independent bases of distinctive features," in *Tones and Features*, J. A. Goldsmith, E. Hume, and L. Wetzels, Eds. Berlin, Boston: DE GRUYTER, Oct. 2011, pp. 264–291.
[15] M. Halle and K. N. Stevens, "A note on laryngeal features," *Quarterly Progress Report, Research Laboratory of Electronics*, pp. 198–213, 1971.
[16] N. Chomsky and M. Halle, *The sound pattern of English.* ERIC, 1968.
[17] J. J. Ohala *et al.*, "Phonetic explanations for nasal sound patterns," in *Nasálfest: Papers from a symposium on nasals and nasalization*. Stanford University Language Universals Project Palo Alto, CA, 1975, pp. 289–316.
[18] J. Al-Tamimi, F. Schiel, G. Khattab, N. Sokhey, D. Amazouz, A. Dallak, and H. Moussa, "A romanization system and webmaus aligner for arabic varieties," in *13th Conference on Language Resources and Evaluation (LREC 2022)*, 2022.
[19] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2020.
[20] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," *Energy*, vol. 1, no. H2, pp. H1–A1, 2010.
[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/
[22] J. Ho, T. Tumkaya, S. Aryal, H. Choi, and A. Claridge-Chang, "Moving beyond p values: data analysis with estimation graphics," *Nature methods*, vol. 16, no. 7, pp. 565–566, 2019.
[23] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 1, pp. 3–36, 2011.
[24] M. Soskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, p. 101017, 2021.
[25] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.
[26] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. FranÃ§ois, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. MÃ¼ller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019.
[27] S. Coretta, *tidymv: Tidy Model Visualisation for Generalised Additive Models*, 2022, r package version 3.3.2. [Online]. Available: https://CRAN.R-project.org/package=tidymv
[28] G. J. Docherty, "The timing of voicing in british english obstruents," in *The Timing of Voicing in British English Obstruents*. De Gruyter Mouton, 2011.
[29] A. Dallak, "Arabic Obstruents: Laryngeal Contrast and Representation," Ph.D. dissertation, Newcastle University, forthcoming.

PRE-ASPIRATION IN SOUTHERN SWEDISH AND ITS RELATIONSHIP TO PROSODY

Mechtild Tronnier

Lund University, Lund, Sweden
Mechtild.Tronnier@ling.lu.se

ABSTRACT

Pre-aspiration has been observed in various languages as a transition between a vowel and a voiceless obstruent. It occurs in northern Swedish dialects, but also as a non-normative feature in Southern Swedish (Scanian). The distribution and magnitude of pre-aspiration in the Scanian variant of Swedish has been investigated in this study, primarily highlighting its relationship to either of the two Swedish word accents. This study is motivated by the fact that laryngeal properties – i.e. stød – also occur in Danish, a language, geographically and historically close to Southern Swedish. Furthermore, the role of vowel length was considered.

The results show that the magnitude, of pre-aspiration is more substantial in Accent 1–words in both: the measured duration of pre-aspiration and in its proportion in the VHC-sequence. Furthermore, largest magnitude of pre-aspiration was found in Accent 1-words containing short vowels.

Keywords: Pre-aspiration, Swedish word accents, non-normative pre-aspiration, Southern Swedish.

1. INTRODUCTION

Pre-aspiration was observed, when carrying out data-base-work – i.e. segmentation – for Southern Swedish dialects [1], which in contrast to more Northern Swedish varieties [2] are not recognized for that trait. The same observation was made in a comparative study on quantity: a section of pre-aspiration was revealed between long/short vowels and long/short voiceless obstruents in one of the languages under investigation – Swedish – but not in the other one. This made a comparison of durational aspects in the coda between the two languages difficult [3].

Previous work on pre-aspiration showed that neither gender nor age was a factor for preferred production of pre-aspiration in Southern Swedish dialects [1]. In addition, neither vocalic type nor vowel quantity was a reason for increased occurrence of pre-aspiration in monosyllabic words [4].

Some studies on the interaction of pre-aspiration and quantity features for various dialects in Sweden have been carried out ([1, 4, 5]). A possible connection between tonal features in Swedish and the occurrence and magnitude of pre-aspiration, however, has not been investigated so far. This contribution, thus, aims to shed some light on the relationship between prosodic features – i.e. contrasting tonal word accents and quantity characteristics occurring in stressed syllables – and pre-aspiration in Southern Swedish, primarily the dialectal variant spoken in Scania.

1.1 Pre-aspiration

A very appropriate definition of pre-aspiration is given by Page [6]:
"Pre-aspiration results from the anticipation of the glottal opening gesture associated with a voiceless obstruent. The glottal opening precedes the oral closing gesture and devoices the latter part of the vowel."

Pre-aspiration is shown in a spectrogram of a relevant speech sample in Figure 1.

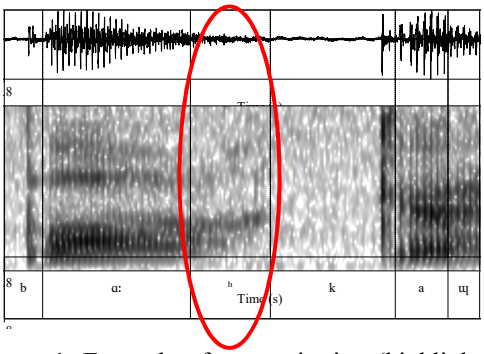


Figure 1: Example of pre-aspiration (highlighted) in the Swedish word *bakar* ("bake", /ˈbaːkar/, pronounced as [ˈbɑːʰkaɯɻ]).

1.2. Pre-aspiration in Languages in Northern Europe

Hejná has published a very comprehensive bibliography of studies on pre-aspiration [7]. This shows that pre-aspiration occurs in languages all over the world, but has predominantly been studied for the Northern European languages.

In Icelandic, pre-aspiration occurs between a short vowel and a voiceless long stop and contrasts with the lack of pre-aspiration in exactly the same context [2]. Pre-aspiration is therefore obligatory and normative for specific lexical items.

In Swedish [2] and Norwegian [8], however, pre-aspiration can be either normative or non-normative, depending on the dialect.

Pre-aspiration also occurs in different variants of Sami, spoken in the north of both, Norway and Sweden [9].

In Western Europe, Irish and Scottish Gaelic ([10, 11]) presents us with this phenomenon. In some English dialects, spoken in or close to the areas in the UK where Gaelic is rooted, pre-aspiration can also be found, e.g. in Scottish English [12] and Welsh English [13], among others.

1.3. Relationship between pre-aspiration and prosodic features

In Icelandic, pre-aspiration occurs after short vowels and is inherited from Old Norse geminates. This does not mean that pre-aspiration is obligatory in all cases in a sequence of short vowels and a voiceless obstruent. Pre-aspiration is a contrastive feature here, and does not occur in words, not inherited from Old Norse [6].

Studies on quantity in various dialects of Swedish have taken the phenomenon of pre-aspiration into consideration [5]. When comparing two dialects from the North of Sweden, both with non-normative pre-aspiration, pre-aspiration showed its relationship to quantity contrast differently [5]. The phenomenon was considerably longer between a long vowel and a voiceless stop (V:HC) than after a short vowel in one of the dialects (Arjeplog), whereas the reverse was the case in the other dialect (Vemdalen): pre-aspiration was considerably longer between a short vowel and a voiceless stop (VhC:), and almost lacking after a long vowel (V:C).

According to studies on present-day Italian [14, 15], pre-aspiration mainly occurs before geminate stops. In addition, if optional pre-aspiration occurs, closure duration is shorter.

So far, the possible interaction between pre-aspiration and tonal features is not widely studied. As there are connections between Swedish tonal word accents and Danish stød, this matter is presented in more detail in the following section.

1.4. Swedish word accents and Danish stød

Stød in Danish is produced with laryngeal constrictions and occurs mainly in stressed syllables containing a long vowel or a short vowel followed by a sonorant in the coda. Historically, stød occurred in monosyllabic words and is therefore related to the word Accent 1 in Swedish and Norwegian [6].

The distribution of both, Danish stød and the two Swedish word accents is very complex and related to the morphological structure of a word A word carrying stød/Accent 1 – when inflected with a particular morpheme – may change to non-stød/Accent 2 – when inflected with another morpheme.

Pre-aspiration may have emerged as an assimilatory process in Danish – and in other languages – when a sonorant became devoiced due to its placement before a voiceless stop [6]. In the case of Danish, this process may have given room for pre-aspiration in a sequence, which in the non-assimilated form had been the place for stød.

In a recent study [16] stronger signs of laryngeal constriction in Accent 1-vowels in Southern Swedish was discovered, which is coherent with voice measures for Danish stød.

It is for the reasons above, that the current study is interesting: non-normative pre-aspiration was investigated, taking the prosodic characteristics of quantity and tonal word accents in Southern Swedish into consideration.

2. THE STUDY

2.1. Material

The gathered material consisted of read speech, with target words imbedded in a carrier phrase. The target words were in focus. They were disyllabic, with the stress on the first syllable and a ˈCV(:)CV(C)-structure. The stressed vowel was either short or long and varied between three different phonemes: /i a o/. The medial consonant consisted of a voiceless stop at one of three different places of articulation: /p t k/. The consonants also varied slightly in length and in their role in the sequence of syllables: in the case of a long stressed vowel, the following consonant introduced the unstressed syllable, whereas in the case of a short stressed vowel, the following consonant was slightly longer and ambisyllabic, serving as a coda and an onset. All these combinations were present in both: words with Accent 1 and Accent 2. The balance was maintained by the choice of underlyingly monosyllabic words, where the disyllabic definite form carries Accent 1 and the plural form Accent 2.

V	Accent 1 definite	Accent 2 plural	Engl.
/o/	kocken	kockar	‘chef’
/o:/	kåken	kåkar	‘building’
/i/	tippen	tippar	‘dump’
/i:/	pipen	pipar	‘pipe’

Table 1: Examples of some target words

There was one repetition for each target word, so that the sum of recorded target words was 72 per speaker.

2.2. Participants

Ten students identifying themselves as speakers of the Scanian dialect were recorded. They varied in age between 18 - 32 years, seven of them were female and three were male speakers. The speakers were not aware of the aim of this study and did not have specific knowledge in phonetics, i.e. beyond pronunciation training in a foreign language.

2.3. Recordings

The recordings were made in a quiet room, equipped with wall panels for acoustic dampening, on a portable digital recorder of the type "Roland R-05" in *.wav*-format (16bit/44.1 sampling rate). A lavalier microphone was used, attached to the upper body of the speaker. The sentences were presented to each speaker one-by-one on the screen of a laptop computer.

2.4. Data analysis

The recorded material was further analysed in PRAAT: at first, an automatic segmentation with the help of the PRAAT VOCAL TOOLKIT was carried out, and then manual correction followed. Hereby, the vowel in the stressed syllable (V) in each recorded item of all target words, the section of pre-aspiration (H) and the following voiceless stop (C) was singled out.

To begin with, the frequency of occurrences was extracted from the data, as pre-aspiration is not normative in the Southern Swedish dialect. This gives a good overview over how much the phenomenon is generally present. In addition, the distribution of its occurrence in relationship to word accents and vowel quantity was considered.

Moreover, the magnitude of occurring sections of pre-aspiration was calculated from the segmentations in the PRAAT-TextGrids. Hereby, the absolute duration of pre-aspiration was calculated for each obtained target item and its proportion in relationship to the VHC-sequence it occurred in in percentage (%). The two calculations were chosen for the following reasons: absolute duration measures are interesting for the inspection of the general perceptual relevance of the phenomenon and the measure of proportion in the VHC-sequence accounts for normalization over speech tempo.

Magnitude measures were than compared with independent T-tests for the variables word accent, vowel quantity and across word accent and vowel quantity.

3. RESULTS

All ten speakers did produce pre-aspiration, and everyone produced items without it. Due to pronunciation errors, only 718 items could be included in the analysis.

3.1. Frequency of occurrence

Under all conditions, pre-aspiration occurred in 70-85% of the utterances – as can be seen in Table 2. There is however a slight advantage for items with Accent 2 over items with Accent 1. The same is the case for quantity, where items with short vowels present us with slightly higher occurrencies then items with long vowels.

	With pre-aspiration	Without pre-aspiration
All ∑718	569 *(79%)*	149 *(21%)*
Accent 1	275 *(77%)*	84 *(23%)*
Accent 2	294 *(82%)*	65 *(18%)*
Long V:	271 *(75%)*	88 *(25%)*
Short V	298 *(83%)*	61 *(17%)*

Table 2: Frequency of occurrence of pre-aspiration under different conditions.

3.2. Magnitude of pre-aspiration

Table 3 shows the results for the differences of the magnitude of pre-aspiration for both, duration and proportion under the various conditions. Generally, Accent 1-words present us with longer durations and proportions of pre-aspiration than Accent 2-words.

Conditions	Probabilities for durations	Probabilities for proportions
Acc1/Acc2	0.00205** *Acc1 > Acc2*	0.00034*** *Acc1 > Acc2*
Long/Short V	0.0258* *L < S*	8.5E-07*** *L < S*
Acc1 Long/Short V	0.02739* *L < S*	1.4E-05*** *L < S*
Acc2 Long/Short V	0.44568 *L = S*	0.01341* *L < S*
Long V: Acc1/Acc2	0.2116 *Acc1 = Acc2*	0.19986 *Acc1 = Acc2*
Short V Acc1/Acc2	0.00547** *Acc1 > Acc2*	0.00161** *Acc1 > Acc2*

Table 3: Probabilities and significance levels for durational and proportional differences of pre-aspiration for various conditions.

Furthermore, words with short vowels occur with a larger magnitude of pre-aspiration than those with long vowels. A more detailed analysis shows that in

Accent 1-words, pre-aspiration is larger in the case of short vowels. In addition, for words with short vowels, pre-aspiration is larger for Accent 1-words. In words with Accent 2 or a long vowel pre-aspiration does not appear to be equally substantial.

3.2.1. Duration of pre-aspiration

The length of pre-aspiration in this data is on average around 40ms (Figure 2a)), which makes the phenomenon perceptually noticeable.

Figure 2a) also shows, that pre-aspiration is longest in Accent 1-words with short vowels ("1S") and it is shortest in Accent 2-words with long vowels ("2L"). Accent 1-words with long vowels ("1L") have slightly longer pre-aspiration than Accent 2-words with short vowels ("2S"). The last two composites are thus found between the previous two extremes.

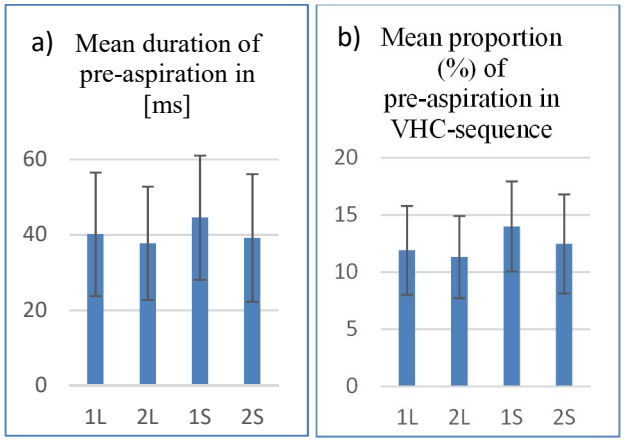


Figure 2: The mean magnitude with standard variation of pre-aspiration (a) duration and b) proportion) under accent-conditions ("1" and "2") and vowel quantity-conditions ("**L**ong V" and "**S**hort V").

3.2.2. Proportion of pre-aspiration in VHC-sequences

The range of the proportions of pre-aspiration spreads around 12% on average in the VHC-sequence under all conditions (Figure 2b)).

The distribution of the mean proportion of pre-aspiration is similar for the utmost conditions as for the duration measurements: pre-aspiration has the largest proportion in Accent 1-words with short vowels ("1S") and the smallest proportions in Accent 2-words with long vowels ("2L"). In the middle, Accent 1-words with long vowels (1L") have more or less the same proportion of pre-aspiration as Accent 2-words with short vowels ("2S").

4. DISCUSSION AND CONCLUSIONS

Pre-aspiration does occur in about 80% of the collected data of the Scanian variant of Swedish and it was produced irrespectively of which accent a word carries and whether the stressed syllables contains a long or a short vowel. There was, however, a slight preference for words with Accent 2 and those with short vowels. The results from quantity measures are not coherent with the results in [4], where monosyllabic words were analysed and vowel quantity did not have an influence on number of occurrence.

In what concerns the magnitude of the produced pre-aspiration on the other hand, a clear trend was observed. Hereby, a larger magnitude of pre-aspiration occurs for Accent 1-words and those with short vowels in comparison to Accent 2-words and words with long vowels. With regard to quantity features, the current results are similar to what had been found in [5] for the dialect of Vemdalen, in that pre-aspiration was longer in the case of a preceding short vowel. The measures of the magnitude however varied: in [5], pre-aspiration after a short vowel took about 28% of the VHC-sequence, whereas in the current data it was only around 12%.

When it comes to the word accents, larger magnitude was favoured for Accent 1-words. Furthermore, Accent 1-words with short vowels show the largest magnitude of the phenomenon.

At this point, the relationship between the current findings and stød and pre-aspiration in Danish comes into the picture. Stød occurs in stressed syllables with either long vowels or short vowels followed by a sonorant. Historically, words with stød are related to words with Accent 1 in Swedish. If a sonorant in a historical stage lost its voicing due to tonal assimilation to the following voiceless stop, that sequence could have ended up as a pre-aspiration and without stød, i.e. stød got lost [6].

Without considerations about the historical connection of the target words with Danish words, which underwent such a process, it is still striking, that words in Southern Swedish with similar structures expose the largest degree of pre-aspiration.

Strong signs of laryngeal constriction were also found in the vowel of Accent 1-words in Southern Swedish in [16], similarly to Danish stød. The investigated material in [16] did not only contain words where a voiceless obstruent followed the vowel. Therefore, the occurrence of pre-aspiration was not accounted for. From that study [16] and from the current one, it can, however, be acknowledged, that some laryngeal action is taking place for words with Accent 1 – in addition to the production of a tonal pattern, which usually characterises the difference between the two word accents – in Southern Swedish. Similar laryngeal actions, however, could not be confirmed for words with Accent 2.

5. ACKNOWLEDGEMENT

The author gratefully acknowledges Lund University Humanities Lab for making available their recording facilities.

I addition, the author wants to thank Svenja Willenborg, who assisted with data collection and analysis during her internship at Lund University.

6. REFERENCES

[1] Tronnier, M. 2002. Preaspiration in Southern Swedish Dialects. *Proceedings of Fonetik 2002*, *Speech, Music and Hearing Quarterly Progress and Status Report, Vol.44,* KTH, Stockholm, 33-36.

[2] Helgason, P. 2002. *Preaspiration in the Nordic Languages. Synchronic and Diachronic Aspects*. PhD thesis, Stockholm University.

[3] Tronnier, M. 2018. Observations on the transitions from vowels to voiceless obstruents: a comparative study. *Proceedings of Fonetik 2018.* Department of Philosophy, Linguistics and Theory of Science, Göteborg University, 73-74.

[4] Tronnier, M. 2002. Preaspirated Stops in Southern Swedish. *Proceedings of ICSLP 2002*, Denver, Col., 1117-11

[5] Wretling, P., Strangert, E., Schaeffler, F. 2003. Preaspiration as a Quantity Feature. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 03-Barcelona)*, 2701-2704.

[6] Page, B. R. 1997. On the origin of preaspiration in Scandinavian. *American Journal of Germanic Linguistics & Literatures* 9(2), 167-190.

[7] Hejna, M. Pre-aspiration biblio. https://misprdlina.wordpress.com/pre-aspiration-biblio/

[8] van Dommelen, W. A. 1999. Preaspiration in intervocalic /k/ vs. /g/ in Norwegian. *14th ICPhS, San Francisco*, 2037-2040.

[9] Engstrand, O. 1987. Preaspiration and voicing contrast in Lule Sámi. *Phonetica* 44, 103-116.

[10] Ní Chasaide, A. 1985. *Preaspiration in Phonological Stop Contrasts*. PhD thesis, University College of North Wales.

[11] Ó Murchú, M. 1985. Devoicing and pre-aspiration in varieties of Scots Gaelic. *Ériu* 36, 195-198.

[12] Gordeeva, O., Scobbie, J. 2007. Non-normative preaspirated voiceless fricatives in Scottish English: phonetic and phonological characteristics. *QMU Speech Science Research Centre Working Papers*.

[13] Hejná, M. 2015. *Pre-aspiration in Welsh English: A case study of Aberystwyth*. Ph.D. dissertation, University of Manchester.

[14] Stevens, M. 2011. Consonant length in Italian: gemination, degemination and preaspiration. *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*, 21-32.

[15] Stevens, M., Reubold, U. 2014. Pre-aspiration, quantity, and sound change. *Laboratory Phonology* 5(4): Ed. J. Cole, 455-488.

[16] Hjortdal, A. 2022. Creaky voice in South Swedish accent 1. *Phonetic Experimental Research at the Institute of Linguistics (PERILUS), Proceedings of Fonetik 2022*: XXXIIIrd Swedish Phonetics Conference. Royal Institute of Technology, Stockholm.

How measures of gestural overlap relate to dynamics:
Evidence from German and English word-initial stop-lateral clusters

Shihao Du[1], Stephan Kuberski[1], Adamantios I. Gafos[1,2]

[1]Department of Linguistics and Cognitive sciences, University of Potsdam, Germany
[2]Haskins Laboratories, New Haven, Connecticut, USA

ABSTRACT

In data from English and German clusters (C1C2), we examine if and how the stiffness of C1 opening and C2 closing movements (the two relevant movements in the C1-to-C2 transition) modulate overlap, using four overlap measures. Results show a variegated picture where different overlap measures do or do not depend on the stiffness parameters. We seek explanations for this patterning that lead to a better understanding of the relation between the plethora of overlap measures used in the literature and the dynamics of the gestures whose overlap is at issue.

Keywords: stiffness, overlap, consonant clusters

1. INTRODUCTION

Studies on inter-segmental coordination have amassed substantial evidence for the cross-linguistic patterning of consonant clusters (e.g., among others, [1]-[23], [25]-[27]). By and large, the approach has been to document patterns of overlap between segments and study the dependence of various overlap measures on the segmental composition of the cluster (stiffness of the component gestures, place of articulation and so on). There has been little to no consensus on what measures to use in characterizing patterns of overlap and in cases where different measures are used within the same study why some measures show dependence on the conditioning factors studied and other overlap measures do not. Here, we seek to address this latter issue by studying the dependence of four different measures of overlap on the conditioning factors of the stiffness of the two relevant movements in the C1-to-C2 transition, C1 opening stiffness and C2 closing stiffness. Whereas the role of the latter parameter in modulating overlap has been studied by Roon et al. [23] and Du and Gafos [10], that of the C1 opening movement remains so far unexplored. Yet, it is intuitively clear that this parameter should play some role in modulating overlap in C1C2 clusters: C1 opening stiffness controls temporal properties of the C1 opening movement which is co-extensive with the transition between C1 and C2, and thus any measure concerning their overlap. Moreover, in terms of overlap measures, while both Roon et al. [23] and Du and Gafos [10] used different quantifications of overlap, their measures were by no means extensive, especially when concerning the transition period between the C1 and C2 gestures. In particular, the interval between the plateaus of C1 and C2, often referred to as the inter-plateau interval (IPI), was absent in the aforementioned two works even though this interval is often used in the literature as an index of overlap (e.g., among others, [1], [12], [21], [26]). Hence, the current study extends this line of research by examining the role of also C1 opening stiffness in modulating overlap and includes IPI as one of the indices of overlap. It will be shown, first, that, similarly to C2 closing stiffness, C1 opening stiffness significantly contributes to modulating overlap, at least for some of the measures of overlap. The dependence of the overlap measures on the conditioning parameters (stiffness of C1 opening and stiffness of C2 closing movement) in our results shows a variegated picture: two overlap measures are modulated by only one stiffness parameter, one measure is modulated by both stiffness parameters, and another measure is not modulated by either of the two stiffness parameters. This patterning sets up the stage for the question we take up here in the Discussion section, namely, why some overlap measures are and other are not modulated by the dynamical parameter of stiffness of the two gestures whose overlap is at issue.

2. METHODS AND DATA

Electromagnetic articulography (EMA) data from three adult German native speakers and three adult American English native speakers were analyzed. The data were collected using the Carstens AG501 at the authors' institution. No participants reported any hearing or speech problems. The experimental procedures were approved by the Ethics Committee at the authors' institution.

The corpus was comprised of German and English word-initial stop-lateral clusters in which C1 was /p, k/ and C2 was /l/. Stimulus words were the lexical items *Plage, Klage* for German and *plight, played, pledge, plead, climb, clip, clean* for English.

German (/English) participants produced ten (/eight) repetitions of each stimulus in a carrier phrase. Overall, a total of 151 German and 156 English tokens went into data analysis.

Gestural landmarks in the C1C2 clusters were identified using the Matlab-based software Mview developed at Haskins Laboratories by Mark Tiede. Stiffness was computed by dividing the peak velocity of the C1 opening or C2 closing movement by the amplitude of that movement, following Roon et al. [23] and Du and Gafos [10].

Overlap in the C1C2 stop-lateral clusters was quantified by four different temporal intervals, each delineated by one landmark from C1 (release, offset) and one from C2 (onset, target). The resulting four intervals are listed in Table 1 and schematized by braces in Figure 1. All four overlap measures were computed by subtracting the C1 landmark timestamp from the C2 landmark timestamp.

Overlap measure	Landmarks delineating the interval corresponding to each overlap measure
1	C1 release to C2 onset
2	C1 release to C2 target (IPI)
3	C1 offset to C2 onset
4	C1 offset to C2 target

Table 1: The four overlap measures used in the present work with their delineating landmarks.

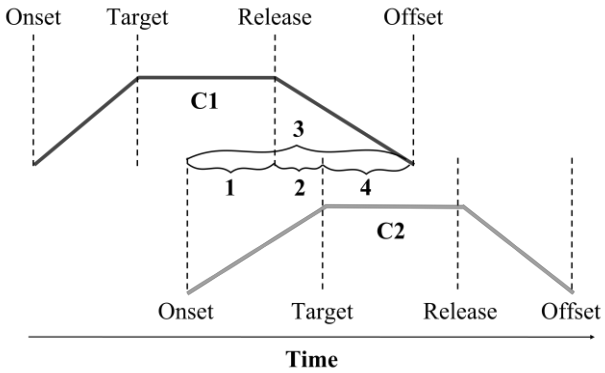


Figure 1: Schematic representation of C1 and C2 gestures with temporal landmarks indicated by vertical dashed lines. The four overlap measures (1, 2, 3, 4) listed in Table 2 are labelled by the horizontal curly brackets.

Linear mixed effects models were fitted to the data to assess relations between the two stiffness parameters (C1 opening stiffness and C2 closing stiffness) and the four overlap measures using the lme4 package in R [24]. C1 opening stiffness (continuous), C2 closing stiffness (continuous), cluster (categorical, levels: /pl/, /kl/) and language (categorical, levels: German, English) were modelled as independent variables with interactions among all possible combinations (C1 opening stiffness × C2 closing stiffness × cluster × language). Each of the four overlap measures was modelled as the dependent variable, thus yielding four models. Random intercepts by subject were also included. All models were assessed using the ANOVA function in the R stats package.

3. RESULTS

We begin with descriptive characterizations of the relations between the four overlap measures (y-axis) and C1 opening stiffness (x-axis) in Figure 2. Figure 3 shows the same for C2 closing stiffness (x-axis). A linear regression line was fitted for each cluster within each language across all speakers.

C1 opening stiffness seems positively related to overlap 3 and 4 (Fig. 2C, 2D) across languages and clusters, while its relation with the other two overlap measures (Fig. 2A, 2B) seems ambivalent, if at all existent. C2 closing stiffness appears to be positively related to overlap 1 and 3 (Fig. 3A, 3C), but less so with the other two overlap measures (Fig. 3B, 3D).

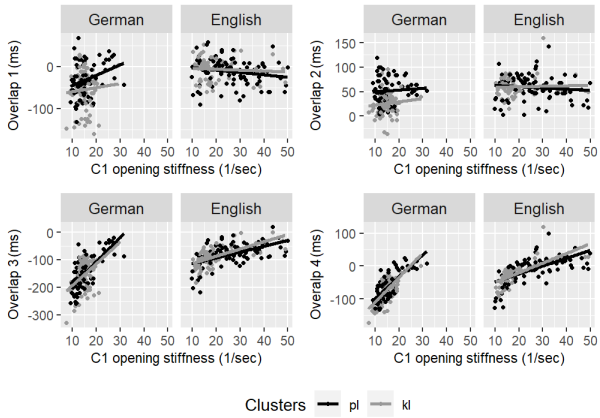


Figure 2: Relations between C1 opening stiffness and the four overlap measures (A: overlap 1, B: overlap 2, C: overlap 3, D: overlap 4). Linear regression lines were fitted across speakers for each cluster in each language.

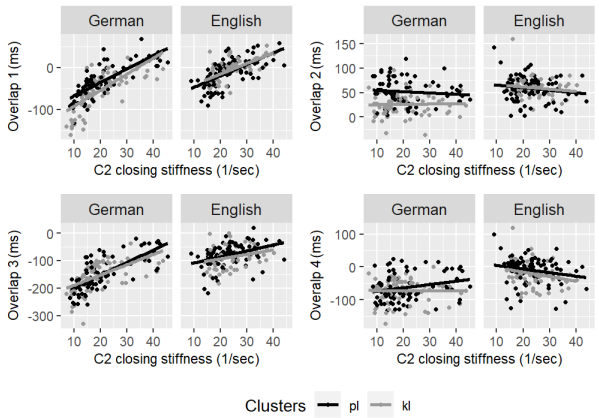


Figure 3: Relations between C2 closing stiffness and the four overlap measures (A: overlap 1, B: overlap 2, C: overlap 3, D: overlap 4). Linear regression lines were fitted across speakers for each cluster in each language.

Statistical results confirm the above observations: for overlap 1 (Fig. 2A and Fig. 3A), only the main effect of C2 closing stiffness was significant ($p < 0.0001$, F-value = 39.35); the effect of C1 opening stiffness was not significant ($p = 0.81$, F-value = 0.06). When overlap was indexed by IPI (overlap 2, Fig. 2B and Fig. 3B), neither the main effect of C1 opening nor that of C2 closing stiffness was significant (C1: $p = 0.36$, F-value = 0.84; C2: $p = 0.44$, t-value = 0.61); for overlap 3 (Fig. 2C and Fig. 3C), both main effects were significant (C1: $p < 0.001$, F-value = 39.12; C2: $p < 0.001$, F-value = 38.12); and finally, for overlap 4 (Fig. 2D and Fig. 3D), only the main effect of C1 opening stiffness showed significance ($p < 0.001$, F-value = 53.42) but not that of C2 closing stiffness ($p = 0.12$, F-value = 2.42). Overall, overlap 1 and 4) are modulated by C2 closing and C1 opening stiffness respectively, overlap 3 is modulated by both stiffnesses, and overlap 2 (IPI) is not modulated by either C1 opening or C2 closing stiffness.

4. DISCUSSION

An examination of the dependence between our four overlap measures on our conditioning parameters (stiffness of the C1 opening and stiffness of the C2 closing movement) gives rise to a variegated picture: two overlap measures are modulated by only one stiffness parameter, one measure is modulated by both stiffness parameters, and another measure is not modulated by either of the two stiffness parameters. This patterning of how overlap measures may or may not be modulated by the dynamics of the two most relevant movements in the transition between C1 and C2 invites explanation.

We begin by recalling the basic reason why stiffness relates to interval duration and hence why overlap measures, which delineate intervals of different durations, should in principle depend on stiffness: stiffness controls the durational properties of a movement – stiffness is the reciprocal of time; recall that stiffness is computed as Velocity v divided by Amplitude A, in which v bears the unit of mm/s and A the unit of mm; hence stiffness has the unit of 1/s (or the reciprocal of time). Thus, the more coextensive (contemporaneous) the movement controlled by the stiffness parameter with the overlap measure in question, the stronger the relation between the two. Note now that, in Figure 1, the different overlap measures are coextensive to different degrees with the two relevant movements (C1 opening, C2 closing) in the C1-to-C2 transition. For instance, overlap 1 is in its entirety coextensive with C2 closing, but not with C1 opening (the latter is so because overlap 1 ends at C1 release whereas C1 opening movement starts at that landmark). More precisely, note that overlap 1 is exclusively coextensive with C2 closing, because during the lifetime of that interval, C1 opening movement is not unfolding). We submit that it is this notion of exclusive coextensiveness that contributes to understanding of how the relations between the four overlap measures and the two stiffnesses play out.

To give a quantitative expression of this thesis, we calculated the percentage of exclusive coextensiveness (PEC) for each overlap measure with the two relevant movements. A PEC value is a property of a pair of an overlap measure (overlap 1, 2, 3, 4) and a movement (C1 opening, C2 closing) and is computed as the proportion of that movement that coextends (is contemporaneous with) the interval delineated by the overlap measure, under the exclusiveness condition that the other movement must not be unfolding at the same time during that interval. For example, in Figure 1, to calculate the PEC for overlap measure 1 with the C2 closing movement, we divide the duration of the interval delineated by overlap 1 by the duration of the C2 closing movement (no C1 opening movement is unfolding during the lifetime of overlap measure 1). Similarly, for the PEC of overlap 3 with the C1 opening movement, we divide the duration of the interval delineated by overlap 4 by the C1 opening movement duration. The results are shown in Figure 4. Overlap 1 and 4 show exclusive coextensiveness predominantly with the C2 closing and C1 opening movement respectively, and thus they are modulated by either C2 closing stiffness or C1 opening stiffness individually but not both at the same time. Only overlap 3 shows considerable coextensive portions with both the C2 closing and the C1 opening movement. Therefore, it is predicted to be the only measure modulated by both stiffnesses; this prediction is borne out in our statistical results in §3.

Moreover, language differences can also be seen in Figure 4, as PEC in each of the eight overlap-movement pair for German is larger than for English; that is, for example, PEC for overlap 1 with C2 closing in German /kl/ is higher than in English /kl/. These PEC differences mesh well with the model fits in §3. In each of the four models, whenever there are significant effects of stiffness, the interaction between these effects and language is statistically significant as well ($p < 0.001$). For instance, even though C1 offset to C2 onset (overlap 3) is modulated by both stiffness terms in German and English, the relations in German are significantly stronger than those in English. This can also be read off from the differences in the regression slopes between German and English in Figure 2C and 3C; the German clusters always have

a steeper slope than their English counterparts. The same observation can be made for overlap 1 (Figure 2A and 3A) and 4 (Figure 2D and 3D): here too, the German clusters have steeper regression line slopes than those in English, likewise confirmed by the model fits, wherein interactions between language and the main effects of stiffnesses are significant.

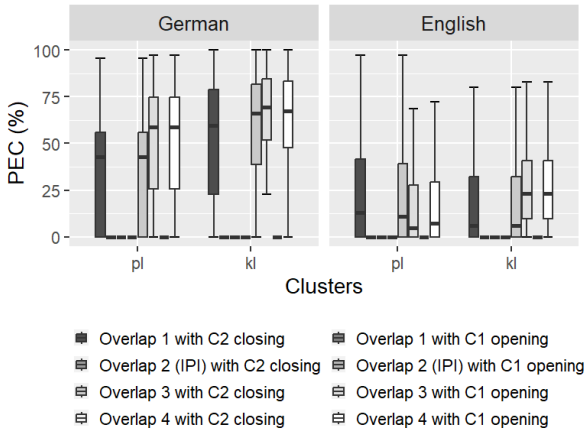


Figure 4: Percentage of exclusive coextensiveness (PEC) for each pair of overlap measure and C1 opening / C2 closing movement across languages and clusters.

Finally, what remains to be explained is why overlap 2 (IPI) is not modulated by either of the two stiffness parameters. As shown in Figure 4, within each cluster type from each language, overlap 2 (IPI) always shows the least PEC with both C1 opening and C2 closing movement. To illustrate more concretely, let us use the schema in Figure 1 to compare overlap 2 (IPI) to overlap 1: it can be seen that IPI exhibits the least amount of exclusive coextensiveness with the C1 opening and C2 closing movements. In fact, during the interval delineated by IPI, the two movements unfold at the same time. Therefore, assessing the relation between (the duration of) IPI and any one of the two stiffness parameters is contaminated by any effects the other stiffness parameter may have on the duration of the same interval (because movement controlled by the other stiffness unfolds at the same time).

Another (non-exclusive to the previous) way to make sense of IPI's lack of relation with the stiffness parameters in our data can be sought from the perspective of variability. In absolute duration terms, IPI is a relatively short interval as depicted in Figure 5, which shows distributions of the overlap measures. Perhaps, then, because IPI is a short interval, its variability, relative to other overlap measures, is limited to such an extent that assessing the relation of its duration and the value of any of the two stiffness parameters becomes infeasible (assessing the nature of a relation between two parameters is possible only if both parameters vary).

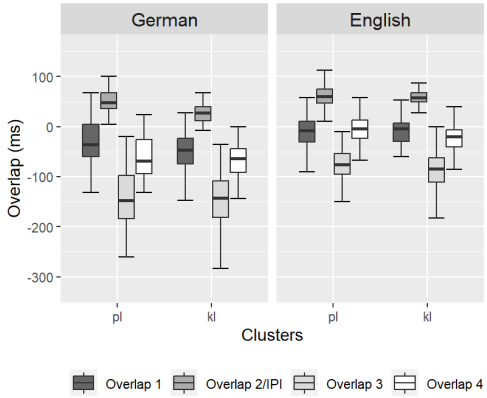


Figure 5: Distributions of the four overlap measures by cluster across German and English. Note that IPI is the interval: 'C1 release to C2 target'.

To compare variability across overlap measures, we used the standard deviation for each overlap measure (because some of our measures take negative values, it is not appropriate to use their means in comparing coefficients of variance, as opposed to standard deviations, of these measures). The results are displayed in Table 2. Both across and within each language, IPI is the least variable interval. Hence, it seems that an appeal to extent of variability may also contribute to the lack of a relation between IPI and the two stiffnesses.

Overlap measure	SD (across languages)	SD (German)	SD (English)
1	41.211	44.616	29.353
2 (IPI)	27.096	27.463	22.586
3	59.471	57.900	39.657
4	47.283	39.808	38.643

Table 2: SD(s) of the four overlap measures across languages and within German and English.

5. CONCLUSION

We investigated the effects of C1 opening and C2 closing stiffness on temporal overlap in English and German word-initial stop-lateral clusters (C1C2), using four overlap measures. Results show that the various overlap measures were not uniformly modulated by the stiffness parameters. Analyses indicate that the extent to which any given overlap measure was modulated by the stiffness parameters is related to the variability of the interval and the degree to which that overlap measure is coextensive with the two movements whose stiffness parameters are under scrutiny. Our results constitute a first step towards a better understanding of the relation between the plethora of overlap measures used in the literature and the dynamics of the gestures whose overlap is at issue.

6. REFERENCES

[1] Bombien, L., & Hoole, P. 2013. Articulatory overlap as a function of voicing in French and German consonant clusters. *The Journal of the Acoustical Society of America*, *134*(1), 539–550. doi: 10.1121/1.4807510

[2] Brunner, J., Geng, C., Sotiropoulou, S., & Gafos, A. 2014. Timing of German onset and word boundary clusters. *Laboratory Phonology*, *5*(4). doi: 10.1515/lp-2014-0014

[3] Byrd, D. 1994. *Articulatory timing in English consonant sequences* [PhD dissertation]. UCLA.

[4] Byrd, D. 1996. Influences on articulatory timing in consonant sequences. *Journal of Phonetics 24*: 209–244. doi: 10.1006/jpho.1996.0012

[5] Byrd, D., & Saltzman, E.L. 1998. Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics* 26, 173–199. doi:10.1006/jpho.1998.0071.

[6] Byrd, D., & Choi, S. 2010. At the juncture of prosody, phonology, and phonetics – the interaction of phrasal and syllable structure in shaping the timing of consonant gestures. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Phonology and Phonetics [PP]*. De Gruyter Mouton. doi: 10.1515/9783110224917.1.31

[7] Chitoran, I., Goldstein, L., & Byrd, D. 2002. Gestural overlap and recoverability: Articulatory evidence from Georgian. In C. Gussenhoven, T. Rietveld, & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 419–447). Mouton de Gruyter.

[8] Cho, T., Lee, Y., & Kim, S. 2014. Prosodic strengthening on the /s/-stop cluster and the phonetic implementation of an allophonic rule in English. *Journal of Phonetics*, *46*, 128–146. doi: 10.1016/j.wocn.2014.06.003

[9] Cho, T. 2016. Prosodic boundary strengthening in the phonetics-prosody interface: Prosodic boundary strengthening. *Language and Linguistics Compass*, *10*(3), 120–141. doi: 10.1111/lnc3.12178

[10] Du, S., & Gafos, A. I. 2023. Articulatory overlap as a function of stiffness in German, English and Spanish word-initial stop-lateral clusters. *Laboratory Phonology*.

[11] Gafos, A. I., Hoole, P., Roon, K., & Zeroual, C. 2010. Variation in overlap and phonological grammar in Moroccan Arabic clusters. In C. Fougeron, B. Kühnert, M. D'Imperio, & Nathalie Vallée (Eds.), *Laboratory Phonology* 10 (pp. 657–698). De Gruyter Mouton. doi: 10.1515/9783110224917.5.657

[12] Gibson, M., Sotiropoulou, S., Tobin, S., & Gafos, A. I. 2019. Temporal aspects of word initial single consonants and consonants in clusters in Spanish. *Phonetica*, *76*(6), 448–478. doi: 10.1159/000501508

[13] Hardcastle, W. 1985. Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication 4*: 247–263.

[14] Hermes, A., Mücke, D., & Auris, B. 2017. The variability of syllable patterns in Tashlhiyt Berber and Polish. *Journal of Phonetics*, *64*, 127–144. doi: 10.1016/j.wocn.2017.05.004

[15] Hoole, P., Bombien, L., Kühnert, B., & Mooshammer, C. 2009. Intrinsic and prosodic effects on articulatory coordination in initial consonant clusters. In G. Fant, H. Fujisaki, & J. Shen (Eds.), *Frontiers in phonetics and speech science* (pp. 275–286). The Commercial Press.

[16] Lialiou, M., Sotiropoulou, S., & Gafos, A. I. 2021. Spatiotemporal coordination in word-medial stop-lateral and s-stop clusters of American English. *Phonetica, 78*(5-6), 385-433. doi: 10.1515/phon-2021-2010

[17] Marin, S., & Pouplier, M. 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control*, *14*(3), 380–407. doi: 10.1123/mcj.14.3.380

[18] Marin, S. 2013. The temporal organization of complex onsets and codas in Romanian: A gestural approach. *Journal of Phonetics, 41*(3–4), 211–227. doi: 10.1016/j.wocn.2013.02.001

[19] Mücke, D., Hermes, A., & Tilsen, S. 2020. Incongruencies between phonological theory and phonetic measurement. *Phonology*, *37*(1), 133–170. doi: 10.1017/S0952675720000068

[20] Pastätter, M., & Pouplier, M. 2013. Temporal coordination of sibilants in Polish onset clusters. *The Journal of the Acoustical Society of America, 134*(5), 4201–4201. doi: 10.1121/1.4831417

[21] Pouplier, M. 2020. Articulatory phonology. In M. Pouplier, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.745

[22] Pouplier, M., Pastätter, M., Hoole, P., Marin, S., Chitoran, I., Lentz, T. O., & Kochetov, A. 2022. Language and cluster-specific effects in the timing of onset consonant sequences in seven languages. *Journal of Phonetics, 93*, 101153. doi: 10.1016/j.wocn.2022.101153

[23] Roon, K. D., Hoole, P., Zeroual, C., Du, S., & Gafos, A. I. 2021. Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. *Laboratory Phonology*, *12*(1), 8. doi: 10.5334/labphon.272

[24] Bates, D. et al. (2022). lme4: Linear Mixed-Effects Models using 'Eigen' and S4. R package version 1.1-31. url: https://CRAN.R-project.org/package=lme4.

[25] Shaw, J., Gafos, A. I., Hoole, P., & Zeroual, C. 2009. Syllabification in Moroccan Arabic: Evidence from patterns of temporal stability in articulation. *Phonology*, *26*(1), 187–215. doi: 10.1017/S0952675709001754

[26] Sotiropoulou, S., Gibson, M., & Gafos, A. 2020. Global organization in Spanish onsets. *Journal of Phonetics*, *82*, 100995. doi: 10.1016/j.wocn.2020.100995

[27] Yanagawa, M. (2006). *Articulatory timing in first and second language: a cross-linguistic study* [PhD Dissertation]. Yale Univerisity.

Glottalisation of voiceless stops in Multicultural London English

Adam J. Chong[1], Marc Garellek[2]

[1] Queen Mary University of London; [2] University of California, San Diego
a.chong@qmul.ac.uk; mgarellek@ucsd.edu

ABSTRACT

Glottalisation of voiceless stops is attested in many English varieties. Most commonly, this occurs with /t/ especially in coda position. The occurrence of glottalisation has been shown to be influenced by a range of phonological, lexical and social factors. Some recent work has also sought to formulate an explanatory theory for why glottalisation occurs more often in certain phonetic and phonological contexts. Many studies, including those examining glottalisation in the UK English varieties, however, have focussed primarily on the glottalisation of /t/, even though glottalisation in coda position is also attested with voiceless stops at other places of articulation. This exploratory study examines glottalisation of voiceless stops in a small corpus of Multicultural London English. We examine the relative frequency of glottalisation of voiceless stops at different places of articulation in different positions. We further find that previous coarticulatory theories of glottalisation are not supported in this set of data.

Keywords: glottalisation, voiceless stops, Multicultural London English, phonetic theory.

1. INTRODUCTION

In many varieties of English, voiceless stops, in particular /t/, have glottal or (pre)glottalised variants (e.g. [ʔ], [ʔ͡t]) in various positions [5,12,13,18,28]. Glottalisation refers to a phonological phenomenon in which a particular segment is associated with vocal fold constriction [11]. This is often phonetically realised as creaky voice, either co-occurring with or without a full glottal stop. For the current study, we do not distinguish what is often referred to as glottal replacement (full replacement by glottal stop, e.g. /t/ -> [ʔ]) and glottal reinforcement (glottal closure along with oral stop). We refer to both phonetic realisations as 'glottalisation' [11]. Glottalised variants of /t/ can occur in several phonological contexts. In a number of British varieties of English, glottalised variants can occur in word-medial intervocalic position, when /t/ is either post-tonic (e.g. *pre**tt**y*) or in between unstressed vowels (e.g. *iden**t**ity*) [26]. More typically, glottalisation occurs with coda /t/, and also with /p/ and /k/, although the rates at which this occurs differs across varieties [13,20]. Listeners seem to use glottalisation to more quickly perceive an upcoming voiceless (vs. voiced) stop [4,21]. Additionally, glottalisation has been shown to occur with voiced stops in some varieties like African American English [11] and with [t͡ʃ] in some British varieties [22]. Previous work, especially examining British English varieties, has shown that glottalisation (typically full replacement of /t/) varies across varieties and is also subject to a range of social (e.g., gender, age, social class and style) and lexical variables, with glottal variants of /t/ often perceived as a stigmatised form [5,7,8,17,24,26]. Further, variable coda glottalisation is also subject to phonological constraints. Coda glottalisation has been shown, for example, to occur more often before sonorants in mainstream North American English [13,25].

Most work on coda glottalisation cited above describes the relation between the occurrence of glottalised variants and other phonological events (e.g. adjacent phonological context) and/or social factors (e.g. speaker age, gender). Explanatory theories, particularly phonetic ones, for *why* coda glottalisation occurs more frequently in these environments, are more limited [10]. Here we consider whether coda glottalisation in Multicultural London English (MLE, [2,3]) can be explained by the theory elaborated in [25]. The authors claim that, at least for mainstream American English of Ohio, the occurrence of coda glottalisation can be partly explained by distinct phonological and phonetic factors, including but not limited to near-categorical allophony (/t/ → [ʔ] before sonorants), the fact that [t] allophones already include glottalisation (to inhibit voicing) [29], and that the glottalisation already associated with [t] can be rendered more perceptible due to gradient reduction of the coronal gesture in certain environments (e.g. phrase-finally, and before non-coronal and aspirated sounds), leading to higher rates of *observed* glottalisation where [t]'s coronal gesture is reduced.

In this exploratory study, we seek to present an initial examination of the quantitative patterns of glottalisation of voiceless stops at all places of articulation in MLE. The questions we address are as follows: (1) Do all voiceless stops /p, t, k/ in coda position show perceptible glottalisation in MLE (c.f. [27]) compared to word-medial intervocalic position? (2) If so, does the rate of perceptible glottalisation

differ by place of articulation? (3) Which phonological factors affect the variation in occurrence of perceptible coda glottalisation? (4) Can [25]'s explanatory theory of glottalisation account for the patterning of glottalisation observed in MLE? Given the small sample examined in this study, we are unable to address the question of how glottalisation of voiceless stops in MLE varies as a function of social factors, such as gender or social class, and how these influence rates of glottalisation across different phonological environments. We leave these questions open for future examination.

2. METHODS

2.1. Corpus information

The data examined in this exploratory study comes from the *Multicultural London English* project corpus [15]. This corpus consists of recordings of 120 individuals in range of age groups: 4-5, 8-9, 12-13, 16-19, 20s-30s and roughly 40 years-old. Recordings were made in a multi-ethnic and multilingual north London borough in the late 2000s and involved sociolinguistic interviews recorded in small groups or pairs, and sometimes individually. No information is available regarding the recording apparatus.

For the current study, we examined four (M: 1; F: 3) of the eight speakers (M: 5; F: 3) in the 20s-30s year-old age group. These were chosen since the transcripts contained the fewest utterance alignment errors as well as fewest overlapping utterances in the recordings. Information about the sociolinguistic background of each speaker is presented in Table 1.

2.2. Annotation procedure

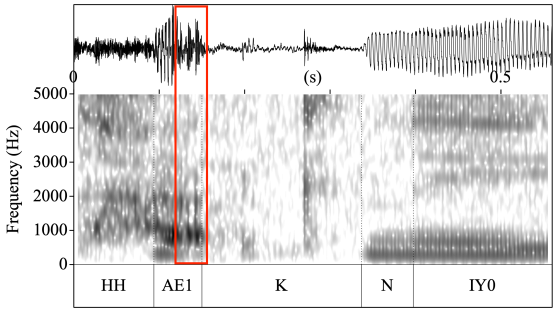


Figure 1: Spectrogram and waveform of the word *Hackney*. The irregular pulses in the final portion of the first vowel indicate glottalisation (red box).

To facilitate manual annotation and coding, the interview recordings from the four speakers were first segmented into utterances in Praat [1] textgrids based on time-points in the ELAN [8] transcripts. These were then forced-aligned using the Montreal Forced Aligner (MFA) [16].

The current study focuses on singleton voiceless stops in word-medial and word-final coda positions. For comparison, we also examined singleton voiceless stops in word-medial intervocalic position (e.g. in *better*) when the following syllable was not stressed. This is a context where glottalisation of /t/, but not /p/ and /k/, is common in many British varieties of English. Coding for perceptible glottalisation was initially done by two phonetically trained annotators, and all the tokens were then reviewed by a third. Annotators labelled a target stop as glottalised based on both auditory analysis, common in sociolinguistic studies [26], and visual inspection of the waveform and spectrograms. Following previous work on glottalisation [6,13,19], words were labelled as glottalised if there was visual evidence for irregularity in F0 and amplitude, based on irregularity pitch periods visible in the waveform and spectrogram, especially in the spacing of vertical striations in the spectrogram (Figure 1). Of course, irregular voicing can also arise due to phrase-final prosody (phrase-final creak) [11], however, in this study we follow [19] in not attempting to distinguish between glottalisation and creak. We leave open for future work the effects of phrase position on glottalisation [19,25].

The following factors were coded for each token (following [25]):

- **Position** in word: whether the target stop was word-medial or word-final.
- **Word**: word containing the target stop.
- **Preceding and following segment**: the (underlying) segment preceding and following the target according to the CMU pronouncing dictionary (dictionary used by the MFA). Pauses following the target stop were also annotated.
- **Stress and following stress:** stress on the syllable containing the target and the following syllable were marked as stressed if it had primary/secondary stressed marked by the MFA. Following [25], monosyllabic function words were always coded as unstressed and following pauses were marked as unstressed.

Tokens which involved speaker overlap, laughter or which were too quiet to recover the variant produced were discarded. The final data set contained 4503 tokens. A breakdown by speaker is shown in Table 1. The data included 3,658 word-final singleton voiceless coda stops (188 /p/, 2608 /t/ and 862 /k/). with 95 word-medial codas (7 /p/, 39 /t/ and 49 /k/). There were additionally 750 word-medial non-foot-initial intervocalic tokens (254 /p/, 269 /t/ and 227 /k/). Following [25], the following consonant was based on the dictionary pronunciation rather than

surface forms.

Speaker	Self-defined ethnicity	Age	N tokens
F1	White British	21	2005
F2	Jamaican	31	671
F3	White British	33	857
M1	Bangladeshi	23	970
		Total	4503

Table 1: Background information for speakers in the current sample and token counts by speaker.

3. RESULTS

This is an exploratory study, for which we have a small number of speakers and small number of tokens for some subset comparisons. Therefore, the analysis focuses on providing the descriptive statistics (means, standard deviations) and overall patterns in the data.

3.1. Glottalisation by place of articulation and position

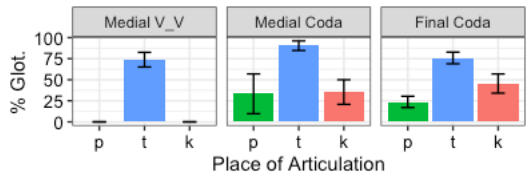


Figure 2: Rates of glottalisation by place of articulation and position (med. intervocalic, medial coda, final coda). Error bars indicate +/- 1 SE.

POA	Medial V_V	Med. Coda	Final Coda
/p/	0 (0)	33.3 (47.1)	23.7 (13.5)
/t/	**73.8 (17.4)**	**90.4 (11.3)**	**75.9 (13.8)**
/k/	0 (0)	35.4 (29.2)	45.5 (22.8)

Table 2: Mean and standard deviation (in parentheses) of glottalisation rate (%) by place of articulation and position.

Figure 2 shows the rate of glottalisation of voiceless stops in medial intervocalic position, word-medial coda position and word-final coda position. Means and standard deviations are shown in Table 2. First, we confirm that glottalisation (or glottaling) in word-medial intervocalic position (e.g. in *better*) only occurs with /t/. We did not observe glottalisation of /p/ or /k/ in this position, as expected. On the other hand, in coda position, /t/ is most frequently glottalised, but we also observed frequent glottalisation of /k/ and less frequent glottalisation of /p/, with slightly lower rates of glottalisation of /p/ in final vs. medial position.

3.2. Factors affecting coda glottalisation

Next, we examine what factors affect the rate of coda glottalisation. Due to the overall low number of tokens in word-medial position, and of /p/ in general, we focus here on examining word-final coda /t/ and /k/. We explore below the effect of the following phonological environment on glottalisation, focussing on the role of the following manner and place of articulation, as well as speaker and lexical effects. First, we find that the rate of glottalisation varies due to speaker and lexical effects (Table 3), with the female speakers glottalising /t/ more often than the one male speaker. Further, we observe that although the overall rates of glottalisation differ across speakers, the relative rate of glottalisation between the different voiceless stops is the same: /t/ > /k/ > /p/, suggesting that the overall trend in glottalisation rates by place of articulation is generalizable across speakers.

Speaker/ POA	/p/	/t/	/k/ (all)	/k/ (without *'like'*)	/k/ (only *'like'*)
F1	36.8	86.7	71.5	32.5	83.7
F2	20	77.1	28.3	16.9	42.6
F3	31.6	83.7	57.5	50	65.0
M1	6.5	56.0	24.5	9.0	40.3
Mean	**23.7**	**75.9**	**45.5**	**27.1**	**57.9**
St. Dev.	**13.5**	**13.8**	**22.8**	**18.1**	**20.5**

Table 3: Rate (%) of coda glottalisation by speaker and place of articulation.

Further, the overall higher rate of glottalisation of /k/ compared to /p/ seems to primarily be driven by a relatively high rate of glottalisation with the word *"like"*. The last two columns of Table 3 show the rates of /k/ glottalisation in a subset of data without tokens of *'like'* (n = 311) and of just *"like"* tokens (n = 551). More than half of the total /k/ tokens come from *"like"*, and glottalisation rate for /k/ is much lower (closer to /p/) when *"like"* is excluded. Given this, we excluded *"like"* in our examination below of phonological effects on glottalization.

Figure 3 and Table 4 shows rates of glottalisation for coda /t/ and /k/ based on following manner and place. Here we grouped individual segments into broad manner and place classes due to the low token count for some segments. Overall, /t/ is glottalised at a very high rate regardless for the following phonological environment. In almost all cases, /t/ was glottalised more than 70% of the time, suggesting that glottal stops are likely a near-categorical allophone of /t/ in coda position.

Next, we examine coda /k/ where glottalisation is more variable and less frequent compared to /t/. [25] proposed that the (voicing inhibiting) laryngeal gesture associated with voiceless stops is more likely to be perceptible when the oral gesture is more liable

to be obscured/weakened. For /k/, this would occur mainly before non-velars (which can obscure the velar gesture if contemporaneous) or aspirated sounds (whose glottal spreading could weaken the glottalisation). However, we observed the opposite: compared to before pauses, /k/ was glottalised as often before coronals, but less often before labials and palatals. Unfortunately, there were too few tokens before velars to make any observations. Further, we observed less frequent glottalisation when the following segment was a vowel or voiceless obstruent. We discuss these results in the context of [25]'s coarticulatory account below.

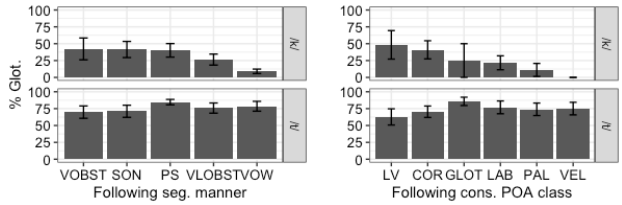


Figure 3: Rates of glottalisation by following (L) segment manner and (R) consonant POA. Top = /k/ without *"like"*, bottom = /t/. LV = Labial-velar (/w/), PS = Pause. Error bars indicate +/- 1 SE.

	/k/ (without *"like"*)		/t/	
	Mean (SD)	*n*	Mean (SD)	*n*
Manner				
SON	41.4 (23.9)	47	71.1 (17.9)	569
VOBST	42.3 (32.4)	40	69.9 (18.3)	417
VLOBST	26.4 (16.2)	83	75.9 (15.1)	431
VOW	9.1 (6.6)	105	78.5 (14.6)	827
PAUSE	40.1 (20)	36	84.7 (8)	364
POA				
LAB	21.7 (20.7)	34	76.9 (19)	283
LAB-VEL	48.3 (42.3)	15	62.7 (24.1)	212
COR	40.9 (26.9)	96	70.4 (16.9)	620
PAL	11.1 (19.2)	6	74 (18.5)	101
VEL	0 (0)	4	75.1 (18.7)	123
GLOT	25 (50)	15	85.8 (12.3)	78

Table 4: Means and standard deviations (with token counts) of glottalisation rate for /k/ and /t/ by following segment (top) manner and (bottom) POA.

4. DISCUSSION AND CONCLUSION

In this exploratory study, we examined the extent to which glottalisation occurred with MLE voiceless stops at in different prosodic positions. We found that all three voiceless stops are glottalised in coda position [27], where /t/ is glottalised at much higher rates than both /p/ and /k/. Unsurprisingly, only /t/ is glottalised in medial intervocalic position.

A second goal of this study was to examine what factors influence coda glottalisation, and to determine if a proposed explanatory theory of coda glottalisation [25] could apply to MLE. Here, we found that coda glottalisation patterns in MLE can be explained by a combination of phonetic and phonological factors. Firstly, the very high rate of /t/ glottalisation across all environments suggests that [ʔ] might be a (near)-categorical allophone of /t/ in this position. As for /k/, we did not find clear support for [25]'s coarticulatory account of glottalisation. We expected that glottalisation would be more observable before non-velars, where the velar gesture is more likely to be weakened. While we found relatively higher rates of glottalisation before coronals, glottalisation was less frequent before labials and palatals, contrary to [25]. The lower rate of glottalisation before palatals (/j/) might be explained by resyllabification across word boundaries of /k.j/ to [.kj]. The lower glottalisation rate before vowels could, similarly, be due to resyllabification processes across a word boundary [14]. The lower rate before labials, however, is problematic under [25]'s account. Glottalisation also occurred less often before voiceless obstruents, which here would have included initial aspirated stops and voiceless fricatives, which both involve glottal spreading. We further found that /k/ was glottalised more frequently before sonorants, a pattern similar to that found for /t/ in North American English [13,25]

Glottalisation of /k/ is also influenced partly by speaker and lexical effects. Particularly, /k/'s overall glottalisation pattern was heavily influenced by high rates of glottalisation with *"like"*, the discourse marker, which was the most frequent /k/ word. Other words showed far lower rates of /k/ glottalisation. Whether this is due to *"like"* occurring more often before phrase boundaries (where glottalisation is more likely to occur) or to word-specific effects is still unclear. Certainly, this allows for future work to examine the trajectory of change for /k/ glottalisation in present-day MLE.

More generally, future work would further test [25]'s proposal with a larger sample of MLE. Our data also raises the question of how listeners perceive glottalisation in MLE, and whether they associate it primarily with voicelessness or whether they show place asymmetries in favour of recovering /t/ given the much higher rates of glottalisation [cf. 4]. Finally, most previous work has typically focused on /t/ glottalisation. Since coda glottalisation also occurs with /p, k/, it would be interesting to examine if these are similarly conditioned by social variables shown to constrain /t/-glottalisation in many English varieties, and whether these differ by position [cf. 24], especially given that glottalisation of /p, k/ does not occur medially but does in codas, unlike /t/.

5. ACKNOWLEDGEMENTS

We would like to thank Paul Kerswill, Jenny Cheshire, Sue Fox and Eivind Torgersen for sharing part of the MLE corpus for the current study. This work would not have been possible without the help of Daniel Price and Francesca Ziegler who were funded by a QMUL Humanities and Social Sciences Undergraduate Research Bursary Award (AJC).

6. REFERENCES

[1] Boersma, P., Weenink, D. 2019. *Praat: Doing phonetics by computer* (version 6.0.50).

[2] Cheshire, J., Fox, S., Kerswill, P., Torgersen, E., 2008. Ethnicity, friendship network and social practices as the motor of dialect change: linguistic innovation in London. *Sociolinguistica* 22, 1–23.

[3] Cheshire, J., Kerswill, P., Fox, S., Torgersen, E., 2011. Contact, the feature pool and the speech community: the emergence of Multicultural London English. *Journal of Sociolinguistics* 15, 151–196.

[4] Chong, A. J., Garellek, M. 2018. Online perception of glottalized coda stops in American English. *Lab. Phon.* 9, 4.

[5] Docherty, G., Foulkes, P. 1999. Derby and Newcastle: Instrumental phonetics and variationist studies. In: Foulkes, P., Docherty G. J. (eds.), *Urban voices: Accent studies in the British Isles.* Routledge. 105-142.

[6] Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M. 1996. Glottalization of wordinitial vowels as a function of prosodic structure. *J. Phonetics* 24, 423–444.

[7] Eddington, D., Channer, C. 2010. American English has got a lot of glottal stops: Social diffusion and linguistic motivation. *American Speech* 85, 338–351.

[8] ELAN (Version 6.4) [Computer software]. 2022. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

[9] Fabricius, A. 2002. Ongoing change in modern RP: Evidence for the disappearing stigma of t-glottalling. *English World-Wide* 23, 115–136.

[10] Farrington, C. 2018. Incomplete neutralization in African American English: The case of final consonant voicing. *Language Variation and Change* 30, 361–383.

[11] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *J. Phonetics* 94, 101155.

[12] Gordeeva, O. B., Scobbie, J. M. 2013. A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality. *J. Int. Phon. Assoc.* 43, 249–271.

[13] Huffman, M. 2005. Segmental and prosodic effects on coda glottalization. *J. Phonetics* 33, *335–362*

[14] Kaźmierski, K., Wojtkowiak, E., Baumann, A. 2016. Coalescent assimilation across word boundaries in American English and in Polish English. *Research in Language* 14, 235–262.

[15] Kerswill, P., Cheshire, J., Fox, S., Torgersen, E. 2007-2010. *Multicultural London English: The Emergence, Acquisition and Diffusion of a New Variety.* ESRC Research Project, RES-062-23-0814.

[16] McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

[17] Milroy, J., Milroy, L., Hartley, S., Walshaw, D. 1994. Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English. *Language Variation and Change* 6, 327–357.

[18] Pierrehumbert, J. B. 1995. Prosodic effects on glottal allophones. In: Fujimura, O., Hirano, M. (eds.), *Vocal Fold Physiology: Voice Quality Control.* Singular. 39–60.

[19] Penney, J., Cox, F., Szakay, A. 2020. Glottalisation, coda voicing and phrase position in Australian English. *J. Acoust. Soc. Am,* 148, 3232-3245.

[20] Penney, J., Cox, F., Szakay, A. 2019. Glottalisation of word-final stops in Australian English unstressed syllables. *J. Int. Phon. Assoc.*

[21] Reinisch, E., Penney, J. 2019. The role of vowel length and glottalization in German learners' perception of the English coda stop voicing contrast. *Lab. Phon.* 10, 18.

[22] Roach, P. J. 1973. Glottalization of English /p/, /t/, /k/ and /ʧ/—A re-examination. *J. Int. Phon. Assoc.* 3, 10–21.

[23] Turton, D. 2019. t-glottalling, flapping and pre-glottalisation in British Englishes: Patterns in phonological and social variability. Talk given at *Cambridge University Linguistics Forum.* Cambridge, UK. 5 December.

[24] Schleef, E. 2013. Glottal replacement of /t/ in two British capitals: Effects of word frequency and morphological compositionality. *Language Variation and Change* 25, 201-223.

[25] Seyfarth, S. and Garellek, M. 2020. Physical and phonological causes of coda /t/ glottalization in the mainstream American English of central Ohio. *Lab. Phon.* 11, 1–33.

[26] Smith, J., Holmes-Elliott, S. 2018. The unstoppable glottal: Tracking rapid change in an iconic British variable. *English Language and Linguistics* 22, 323–355.

[27] Tollfree, L. 1999. South East London English: Discrete versus continuous modelling of consonantal *reduction.* In: Foulkes, P., Docherty G. J. (eds.), *Urban voices: Accent studies in the British Isles.* Routledge. 163–184

[28] Wells, J. C. 1982. *Accents of English*. Cambridge University Press.

[29] Westbury, J. R., and Keating, P. A. 1986. On the naturalness of stop consonant voicing. *J. Ling.* 22, 145–166.

FORMANT CHARACTERISTICS OF UNSTRESSED UNREDUCED VOWELS IN AMERICAN ENGLISH: ONLY EXPLAINED BY DURATION?

Mariko Sugahara

Doshisha University, Japan
msugahar@mail.doshisha.ac.jp

ABSTRACT

It has been considered that there are four levels in English vowels: primarily stressed (P), secondarily stressed (S), so-called "unstressed unreduced" (U) and unstressed reduced (R). According to Fear et al., U vowels were shorter than P and S but longer than R, and they were more reduced than P and S though not as reduced as R in terms of their vowel qualities. The question asked in this study is whether U's partial reduction is only a by-product of its shorter duration. A production experiment was conducted with 25 American English speakers, focusing on /æ/ and /ɔː~ɑː/. We performed statistical analyses using LMM and the results indicate for both of the vowels that U's partial phonetic reduction is not simply a by-product of its shorter duration.

Keywords: unstressed unreduced vowels, American English, phonetic reduction, duration, formant characteristics

1. INTRODUCTION

It has been considered that there are four levels in English vowels: primary stress (P), secondary stress (S), so-called "unstressed unreduced" vowels (U), and unstressed reduced vowels known as schwa ([1], [2]). According to the production study conducted by [1], the acoustic characteristics of U were in between those of the stressed vowels (P and S) and R. U vowels were shorter than P and S but longer than R, and they were more reduced than P and S though not as reduced as R in terms of their vowel qualities which are indicated by *F2-F1* differences. Nonetheless, the perception experiment conducted by [1], in which the four types of vowels were cross-spliced and listeners rated the acceptability of the resulting words, the listeners more steadily grouped U with the stressed vowels (P and S) rather than R. This suggests that the U vowels are perceptually categorized as full vowels just as P and S are despite their partial spectral reduction.

According to [3], [4] and [5], the major source of vowels' phonetic reduction in their spectral characteristics is their shorter duration: The shorter the duration of a vowel is, the more difficult for its articulators to reach their targets, resulting in reduction in their spectral characteristics. This is called target undershoot. The strongest hypothesis under this undershoot model is that English U vowels are primarily associated with shorter duration than P and S, and the reduction of their spectral characteristics is nothing but a by-product of their shorter duration.

However, it is still possible to imagine that the articulatory target of U vowels is set at a slightly slacker position than that of P and S vowels independently of U's shorter duration, and the spectral characteristics of U may differ from those of P and S even when the effect of duration is removed.

A production experiment was conducted and statistical analyses using LMM (Linear Mixed Model) were performed to test this alternative view.

2. PRODUCTION EXPERIMENT

2.1. Participants

The participants were 25 paid native speakers of American English, comprising 15 females and 10 males, who were undergraduate or graduate students at universities in the United States, from 18 to 29 years old at the time of the experiment. Eight of them were from the New England variety area, six from the West, five from the South, three from the Midland, one from the North Central area, one from the Inland North area. One person was raised in multiple areas.

2.2. Word sets

Among all the word sets used in our production experiment, 16 words were considered in the subsequent analyses, which are shown in Sets 1-4. They all started with non-high monophthongs. The initial vowels of the four words in each set were divided into the four types: P, S, U, and R. The words in Sets 3 & 4 except for *authenticity* and *Athena* were adopted from [1].

/æ/ as the initial full vowel (P, S, U)

Set 1	(P) ˈ**a**ctive	(U) **a**cˈtivity
	(S) ˌ**a**ctiˈvation	(R) **a**/ə/cˈcept
Set 2	(P) ˈ**a**sphalt	(U) **a**sˈphaltic
	(S) ˌ**a**spiˈration	(R) **a**/ə/sˈparagus

/ɔː~ɑː/ as the initial full vowel (P, S U)

Set 3	(P) ˈ**au**diences	(U) **au**ˈdition
	(S) ˌ**au**diˈtoria	(R) **a**/ə/ˈddition
Set 4	(P) ˈ**au**thorize	(U) **au**ˈthentic
	(S) ˌ**au**thenˈticity	(R) **A**/ə/ˈthena

The initial vowels of P, S and U in Sets 1 &2 were /æ/, and those in Sets 3 & 4 were /ɔː~ɑː/. Whether they are pronounced as /ɔː/ or /ɑː/ depends on speakers' varieties ([6], [7], [8]). (The spectral variation of /ɔː~ɑː/, however, seems to be even smaller than that of /æ/, as shown in Figures 2 and 3).

The initial vowels of all the four words in each set were followed by the same consonant so that the effect of the following consonant is kept the same across the members in the same set. We included the reduced vowel /ə/ as the control case in order to verify if the U vowels keep their unreduced vowel qualities.

Additional word sets were also included in the production experiment, which were not considered in this study. It is because the initial vowels in the additional sets did not satisfy the criteria of being a non-high monophthong with no onset consonant.

2.3. Recordings

The recordings took place in a sound-proof studio. Cardioid condenser microphones are used: Countryman Headset Microphone for 17 participants and RØDE NT2-A for eight participants. Sound was recorded as WAV files in monaural using Marantz Solid State Recorder at 16 bits with the sampling rate of 44.1 kHz.

The target words were embedded in two types of carrier sentences: (i) *What did you say? I said* "___" *this time*, and (ii) *I didn't say* "___." *X said* "___." For each word, the carrier sentences (i) and (ii) were produced once respectively. About 300 filler sentences were also included in each speaker's recording session and the presentation order of sentences was randomized. This study, however, only focused on the target word tokens produced in the carrier sentences of type (i), where the word was interpreted as new and its primary-stress syllable coincided with a nuclear pitch accent.

2.4. Acoustic measurements

Acoustic data were obtained from the initial vowels of the 16 target words in Sets 1 to 4. Although the total number of the initial vowels should be 400 (4 words*4 sets*25 speakers), four tokens were missing (two tokens of *accept* and two tokens of *authorize*) and the number of recorded tokens considered in this study was 396 (16 from 21 participants and 15 from four participants).

The WAV file of each token was segmented using PRAAT ([9]), and we obtained (i) the duration values of the initial vowel intervals in ms, and (ii) their formant values (*F1, F2*) in Hz at their midpoint.

The starting point of each vowel was defined as where its formants explicitly appeared. The point sometimes coincided with the beginning of glottalized phonation. Occasionally, there were cases in which an epenthetic schwa vowel followed the coda consonant /d/ of the preceding word *said* and no clear break in the spectrogram was available between the epenthetic vowel and the beginning of the target initial vowel. In such cases, the starting point was identified as where the amplitude of the waveform narrowed down or where the vowel quality of the epenthetic vowel changed to that of the target initial vowel. The ending point of each vowel was defined as where consonant frication or closure started.

The *F1* and *F2* values in Hz were converted into the Bark scale, and the durational values were converted into the logarithmic scale (Ln). We regard the distance between the Bark-scaled *F1* and the Bark-scaled *F2*, i.e., *F2-F1* henceforth, as the indicator of vowel reduction following [1] that greater *F2-F1* indicates more reduction. We adopted this measure for empirical reasons. Our speakers' *F1* of /æ/ decreased while that of /ɔː~ɑː/ was almost unchanged, and their *F2* of /ɔː~ɑː/ increased while that of /æ/ was almost intact, as the stress level shifted from P to U. In order to analyse the spectral characteristics of the two vowels in a parallel manner, the measure of *F2-F1* was needed.

The outliers of the *F2-F1* values were identified separately for each stress level of each vowel using the 1.5*IQR rule. There were ten *F2-F1* outliers: one instance of *activation*, *active* and *authorize* each, two instances of *audition* and *authentic* each, three instances of *authenticity*, which were removed when creating Fig. 1 and Fig. 3, and when conducting statistical analyses. After the removal of the ten outliers, the number of the *F2-F1* data was 386: P=96, S=100, U=98, R=92.

3. PRELIMINARY OBSERVATIONS

Before performing the main statistical analyses, we made some preliminary observations.

3.1. Duration and *F2-F1*

Given that a greater *F2-F1* value is an indicator of phonetic reduction, and if shorter duration incurs more phonetic reduction, we expect *F2-F1* to increase as vowel's duration decreases.

To see the relationship of duration and *F2-F1*, durational values were normalized to remove speaker differences. First, each vowel token's duration was

converted into the logarithmic (Ln) scale. Secondly, for each speaker, the mean of the log durations was obtained separately from the initial vowels of the eight words in Sets 1 & 2 (i.e., the /æ/ sets) and from those in Sets 3 & 4 (i.e., the /ɔ:~ɑ:/ sets). For example, for each speaker, the mean duration of the initial vowels of P, S, U, R in Sets 1 & 2 and the mean duration of those in Sets 3 & 4 were obtained separately. Thirdly, we subtracted the mean duration from the duration of the corresponding vowel token, which is the normalized duration value.

The box plots in Fig 1. summarize the normalized duration of the four vowel levels, and the scatter plots in Fig. 2 show the normalized duration values plotted on the x-axis and *F2-F1* plotted on the y-axis. It is evident from Fig. 1 that as the stress level decreased from P to R, duration also decreased, which replicates the observation by [1].

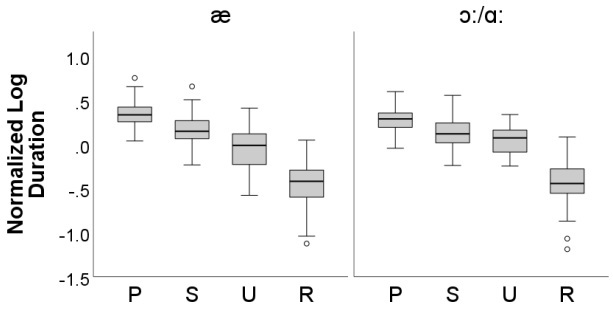


Figure 1: Normalized duration of the four levels

Fig. 2 shows a negative correlation between the *F2-F1* values and the duration values across the four different vowel levels. There are, however, differences between U and the other vowel types: U's slopes of fitted lines are steeper than the other groups' slopes. We test if the differences are significant in the following statistical analyses.

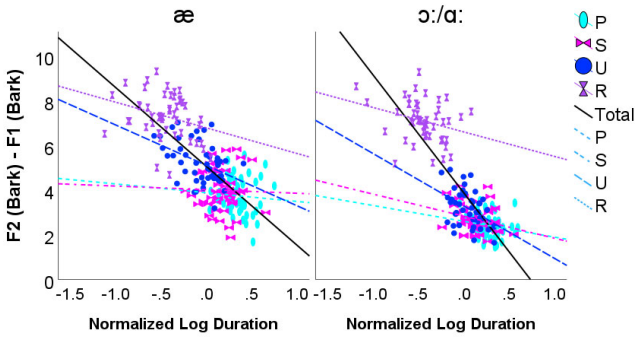


Figure 2: *F2-F1* plotted against normalized duration

Another observation from Fig. 2 is that the overlap between U's and R's data points for /æ/ is greater than those for /ɔ:~ɑ:/ on both the x-coordinate (duration) and the y-coordinate (*F2-F1*). This implies that the degree of U's phonetic reduction is different between /æ/ and /ɔ:~ɑ:/. This point will be discussed in Sec. 5.

3.3. Sex and *F2-F1*

Since *F1* and *F2* values are known to vary according to vocal tract length ([10] among others), sex may affect the values of *F2-F1*. Fig. 3 presents the box plots of *F2-F1* from females and males. Although the ranges of the boxes and the whiskers slightly differed between F and M, their medians were approximately the same for all of the vowel types regardless of the vowel differences. It could be argued that normalisation between the two sex groups was achieved to some extent by taking the difference between *F2* and *F1*.

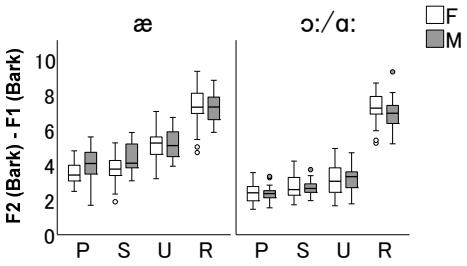


Figure 3: *F2-F1* box plots of females and males

4. STATISTICAL ANALYSES AND RESULTS

4.1. Analyses

Statistical analyses were performed on SPSS using LMM (Linear Mixed Model) to investigate whether (i) the vowel types and duration both contribute to predicting *F2-F1*, (ii) *F2-F1* differs between U and the other vowel types even when given the same duration, and (iii) the relationship between *F2-F1* and duration is the same across the four vowel types.

For each of the vowels, three models were tested: M1, M2 and M3. The target variable was *F2-F1* in all of the models. One fixed factor (*Type*, i.e., P, S, U, R) was included in M1, two factors (*Type* and *Duration*) in M2, and three factors (*Type, Duration, Type*Duration*) in M3. We consider collinearity between *Type* and *Duration* not to be a problem here because their VIF was 2.94 for /æ/ and 2.69 for /ɔ:~ɑ:/, less than 5. (The VIF scores were obtained by converting *Type* into a scalar variable.) Random intercepts of *Speaker* and those of *WordSet* were also included in the three models.

4.2. Results

First, the three models were compared. For both of the vowels, M3's AIC was the smallest (for /æ/, 501.76 for M1, 489.81 for M2, 481.46 for M3; for /ɔ:~ɑ:/, 404.79 for M1, 381.61 for M2, 374.79 for M3). From this, M3 seems to be the most appropriate model among the three. However, not all of the three fixed factors in M3 were significant. Although M3 itself was significant: [F (7, 188) =

102.9, $p < .001$ for /æ/; F (7, 182) = 308.3, $p < .001$ for /ɔː~ɑː/], and *Type* and *Duration* in M3 were also significant respectively: [*Type*: F (3, 188) = 43.65, $p < .001$ for /æ/ and F (3, 182) = 112.06, $p < .001$ for /ɔː~ɑː/; *Duration*: F (1, 188) = 6.77, $p < .025$ for /æ/ and F (1, 182) = 22.48, $p < .001$ for /ɔː~ɑː/], the factor of *Type*Duration* was not significant for both of the vowels: [F (3, 188) = 1.56, $p = .2$ for /æ/; F (3, 182) = 1.12, $p = .34$ for /ɔː~ɑː/]. (Since there are separate analyses for the two, we adopted Bonferroni correction and the α level was set at .025.) Tables 1 and 2 summarize the results of the coefficients of the fixed factors in M3.

Table 1: Results of the fixed factors for /æ/ (M3)

	β	SE	t	95%CI	
Intercept (U)	5.1	.20	25.9***	4.7	5.5
diff (P-U)	-1.2	.29	-4.2***	-1.8	-.66
(S-U)	-1.1	.19	-6.0***	-1.5	-.77
(R-U)	1.7	.25	6.7***	1.2	2.2
Slope (U*D)	-1.6	.51	-3.2**	-2.6	-.60
diff (P*D-U*D)	1.2	.90	1.3	-.59	3.0
(S*D-U*D)	1.7	.85	2.0	.05	3.4
(R*D-U*D)	.49	.69	.71	-.87	1.8

'***' <.001, '**' <.01, '*' <.025

Table 2: Results of the fixed factors for /ɔː~ɑː/ (M3)

	β	SE	t	95%CI	
Intercept (U)	3.3	.19	17.5***	2.9	3.7
diff (P-U)	-.57	.23	-2.5*	-1.0	-.12
(S-U)	-.51	.14	-3.5***	-.79	-.22
(R-U)	3.2	.21	15.8***	2.9	3.7
Slope (U*D)	-2.2	.60	-3.6***	-3.4	-.98
diff (P*D-U*D)	.99	.91	1.1	-.81	2.8
(S*D-U*D)	1.5	.84	1.8	-.16	3.2
(R*D-U*D)	1.0	.73	1.4	-.42	2.5

The tables above show that for both of the vowel phonemes, U's estimated slope and intercept were significant. The term 'diff (*TypeX*-U)' represents the estimated difference between U's intercept and the intercepts of the other vowel types' slopes. The negative values of 'diff (P-U)' and 'diff (S-U)' mean that U's intercept is significantly greater than P's and S's whereas the positive value of 'diff (R-U)' means that U's intercept is significantly smaller than R's.

The estimated differences in slopes between U and the other vowel types represented as 'diff (*TypeX**D-U*D)' were not significant, which is already expected as mentioned above. (Because U's slope was negative, their positive values mean less steep slopes.)

5. DISCUSSION AND CONCLUSION

The results above show that *Vowel Type* contributes to predicting *F2-F1* together with *Duration*. U's intercept being greater than P/S's and it being smaller than R's means that when given zero on the x-coordinate, i.e., the coordinate of *Duration*, U's estimated *F2-F1* value is expected to be greater than P/S's and smaller than R's. The value of zero on the x-coordinate corresponds to the mean of the normalized durations in our study. As already mentioned in Sec. 3.1, we normalized durations by taking deviations from the mean of the initial vowels' durations for each vowel phoneme group. Because they are deviations, their mean is zero. Given this, U's estimated *F2-F1* is greater than P/S's and smaller than R's when given the durational mean of all the tokens. This suggests that U's spectral quality is more reduced than P's and S's, and less reduced than R's independently of duration, i.e., U's partial reduction is not simply a by-product of its shorter duration. This further implies that U's articulatory target is set at a slightly slacker position than P's and S's.

Although U's slopes of fitted lines were steeper than the other vowel types' slopes (see Fig. 2), the difference was not statistically significant. That is, there is little persuasive evidence to support the claim that the U's relation with duration is different from the other vowels' relation to it.

One thing to note is that the magnitude of U's phonetic reduction is more pronounced for /æ/ than for /ɔː~ɑː/ as mentioned in Sec. 3.2. A key for explaining this may be that the former is lax while the latter is tense. Lax vowels have been claimed to be more prone to reduction than tense vowels in metrically weak position ([11]), with which our outcome seems to comply. Because the only lax and tense vowels tested here are /æ/ and /ɔː~ɑː/ respectively, it is too early to draw any firm conclusion regarding this, which needs to be investigated further.

Related to this point, a perception study may be necessary, too. In their perception study, [1] found that U was grouped with P and S, i.e., the full vowels, rather than R by native speakers. They, however, only used tense vowels. Given a possibility that the lax vowel /æ/ of the U type is more reduced than the tense vowel /ɔː~ɑː/ of the same type, it is still unresolved how the former is perceptually grouped.

The target vowels considered in this work were limited to non-high monophthongs preceded by no onset consonant. We need to look at other vowels, too. The presence or absence of onset consonants may also affect the outcome. As mentioned in Sec. 2.2, additional word sets with other vowels with and without onset consonants were also recorded in our production experiment. Their spectral patterns will be considered in our future study. We have also recorded the target words in a post-focus environment as mentioned in Sec. 3.3., and it is of interest to us how U vowels behave in that environment.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP17K02828 and JP22K00627. We are thankful to reviewers for their valuable comments, which helped us improve the quality of the paper. All errors are my own.

REFERENCES

[1] Fear, B. D., Cutler, A., Butterfield, S. 1995. The strong/weak syllable distinction in English. *J. Acoust. Soc. Am*. 97, 1893-1904.
[2] Ladefoged, P., Johnson, K. 2015. *A Course in Phonetics*, 7th ed. Cengage Learning.
[3] Lindblom, B. 1963. Spectrographic Study of Vowel Reduction. *J. Acoust. Soc. Am*. 35, 1773-1781.
[4] Moon S-J., Lindblom, S. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am*. 96, 40-55.
[5] Barns, J. 2006. *Strength and Weakness at the Interface.* Mouton de Gruyter.
[6] Jones, D., Roach, P., Setter, J., Esling, J. 2011. *Cambridge English Pronouncing Dictionary*, 18th ed. Cambridge University Press.
[7] Upton, C., & Kretzschmar, Jr. W. A. 2017. *The Routledge Dictionary of Pronunciation for Current English*, 2nd ed. Routledge.
[8] Wells, J. C. 2008. *Longman Pronunciation Dictionary*, 3rd ed. Longman.
[9] Boersma, P., Weenink, D. 2022. *Praat: Doing Phonetics by Computer* [Computer program]. Version 6.1.51.
[10] Johnson, K. 2020. The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology*. 11(1), 1-16. DOI:https://doi.org/10.5334/labphon.196
[11] Chomsky, N., Halle, M. 1968. *The Sound Pattern of English*. MIT Press.

DISAPPEARING SUPRALARYNGEAL ARTICULATORY DISTINCTION OF THE THREE-WAY LARYNGEAL CONTRAST OF KOREAN VELAR STOPS

Jiyeon Song[1], Sahyang Kim[2], Taehong Cho[1]

[1]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS), Hanyang University
[2]Hongik University
jiyeoni00@hanyang.ac.kr, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This EMA study investigated articulatory characteristics of the three-way contrast of Korean velar stops in CV contexts. Building on a previous small-scaled study that indicated a three-way kinematic distinction of bilabials, the current study examined velars with 22 speakers to further explore the kinematic evidence of the three-way stop contrast, by considering the recent VOT merger between the lenis and aspirated stops. The results showed that both males and females maintained the aspirated-lenis distinction only minimally in different kinematic dimensions. Males showed a three-way contrast in peak velocity of C-to-V opening movement (/k^h/</k/</k*/) with a lenis-aspirated distinction. Females never showed a three-way contrast in any kinematic measure, but a two-way lenis-aspirated distinction only in C-to-V deceleration duration (/k*/=/k^h/</k/). It appears that the lenis-aspirated kinematic distinction is being attenuated, led by female speakers, in line with the loss of lenis-aspirated distinction in VOT, which would otherwise contribute to the supralaryngeal articulation.

Keywords: Korean three-way stop contrast, VOT merger, sound change, kinematic characteristics

1. INTRODUCTION

In many languages, stops are mostly classified based on voicing and aspiration. However, Korean shows an unusual three-way contrast of stops that are all voiceless word-initially. The three-way contrastive stops are often called the lenis, the fortis, and the aspirated stops and occur at three places of articulation (bilabial, denti-alveolar, and velar). In general, these stops are assumed to be contrastive in laryngeal terms: the lenis stops are characterized as breathy, and slightly aspirated; the fortis stops as tensed, laryngealized, and unaspirated, and the aspirated stops as strongly aspirated (e.g., [6]). It was also previously suggested that the three-way contrast can be explained with VOT alone (see [6] for related discussion) which could be defined as the timing between the supralaryngeal release gesture and the laryngeal voicing gesture (cf. [12]). (Note, however, that the classification of stops is traditionally defined along two phonetic dimensions: voicing and aspiration (e.g., [13]). Traditionally, the three-way contrast was considered to be reflected in the mean VOTs, which would be longest for the aspirated, shortest for the fortis, and intermediate for the lenis stops, although VOT ranges may overlap quite substantially between the lenis and the aspirated stops (cf. [4, 6]). Subsequent studies, however, showed that the distinction in VOT between lenis and aspirated stops has been neutralized, showing a VOT merger between the two stop categories (e.g., [1, 8, 11, 17]). Further, it has been suggested that females are more likely to show the VOT merger than males in Seoul Korean (e.g., [9]).

Aerodynamic and physiological factors involved in determining VOT differences are related to articulatory movement velocity and a temporal adjustment between stop closure duration and VOT ([5]). According to [14], there is a trade-off between the closure duration and the following VOT within the constant duration of abduction and adduction cycle of the vocal cords. In other words, the longer the closure duration, the shorter the VOT. Therefore, fortis consonants with the shortest VOT have the longest closure duration. This was partially supported in the previous research in kinematic terms with the fortis having the longest closure duration by a binary distinction (/p/</p^h p*/) in lip constriction duration in the articulatory phonology research about Korean labial stops (e.g., [7]) and the three-way contrast in the tongue dorsum closure duration (/k/</k^h/</k*/) (e.g., [3]).

Brunner et al. [3] found the three-way contrast in kinematic terms in peak velocity of the tongue dorsum closing movement (/k/</k^h/</k*/). In a similar vein, a kinematic study of labials by [7] found a three-way contrast in lip constriction degree (/p/</p^h/</p*) at the constriction offset in both IP-initial and IP-medial positions, suggesting that the three-way contrast becomes greater as the constriction progresses. Brunner et al. [3] also showed the three-way contrast in deceleration duration of the tongue dorsum closure movement (/k*/</k^h/</k/). This is in line with the perspective of the virtual target hypothesis in which different target positions could be defined virtually even beyond the

physical limit of the palate, leading to different closure durations.

These kinematic studies taken together indicate that the three-way stop contrast in Korean, although they may be defined primarily in laryngeal terms, can still be distinctive at the supralaryngeal level as reflected in kinematic measures. But these studies were based on a small number of speakers (three speakers for [3], five speakers for [7]), and therefore it would be difficult to generalize the findings. Moreover, recent studies, as introduced above, indicate that there has been a VOT merger between the lenis and the aspirated stop whose distinction is now made by F0 at least among younger speakers of Seoul Korean. Given that those kinematic studies were carried out several years ago and the sound change of the VOT merger appears to have been in progress, an important question arises as to whether and how the kinematic properties of the three-way stop contrast may have also been modified, reflecting possible effects of the sound change that might have impacted supralaryngeal articulation. In particular, if we define VOT articulatorily (as in [5]), it is possible that the tongue posture associated with VOT may vary depending on how long the hold (closure) duration continues, which tends to be inversely related to VOT (cf. [14]). However, it is possible that due to the VOT merger, the lenis-aspirated distinction would no longer be associated with the differential hold duration-related impacts on articulation. If so, the supralaryngeal differences between the lenis and the aspirated stops that previously existed may now have disappeared along with the VOT merger.

With these possibilities in mind, the present study examines production of monosyllable words of 22 young Seoul Korean speakers to explore how the three-way laryngeal contrast of velars is realized kinematically at the supralaryngeal level, and how the results may be interpreted in connection with the recent VOT merger between the lenis and aspirated stops. Given that the on-going sound change is assumed to be led by female speakers, it will also be considered whether any loss of supralaryngeal distinction especially between the lenis and the aspirated may be more evident in the female than in the male speech.

2. METHODS

2.1. Participants and Procedure

Twenty-two Seoul Gyeonggi speakers participated in the experiment as part of a large-scale acoustic-articulatory database building project in Korean at Hanyang Institute for Phonetics and Cognitive Science of Language [18]. Each participant was asked to produce a sequence of eighteen syllables of the form /Ca/ of which the three-way contrastive velar stops (/ka/, /k^{h}a/, /k*a/) were analyzed for the current study. A total of 132 tokens (3 velars x 22 speakers x 2 repetition) are reported in this paper.

The articulatory data were collected using the 3D electromagnetic articulography (EMA, AG501, Carstens Electronics). Eight movement sensors were used to track the movement of articulators. Six sensors were attached to the articulators: tongue dorsum (TD), tongue body, tongue tip, the lower lip and upper lips at the vermilion borders, and the middle of the lower gumline. Two fixed reference sensors were attached to the nasion and upper incisor. The data of TD sensor has been analyzed to track the closing and opening movement of TD.

2.2. Measurements and Statistical analysis

Mview (a set of Matlab scripts developed by Mark Tiede at Haskins Labs) was used to analyze the obtained kinematic data. Figure 1 shows a schematic of the kinematic measures and relevant landmarks for the closing and opening movement trajectory of the tongue dorsum, and the tangential velocity of the tongue dorsum.

Kinematic measurements of the tongue dorsum in accordance with the labels of Figure 1 include the following: (a) Closing Movement Duration , (a-1) Acceleration Duration 1: duration from Onset 1 to Peak Velocity 1, (a-2) Deceleration Duration 1: from Peak Velocity 1 to Target 1, (a-3) Peak Velocity 1, (b) Opening Movement Duration, (b-1) Acceleration Duration 2: from Onset 2 to Peak Velocity 2, (b-2) Deceleration Duration 2: from Peak Velocity 2 to Target 2, (b-3) Peak Velocity 2, (c) Closure Duration: from Target 1 to Onset 2. Forward and vertical movements duration from Onset 1 to Target 1 were also measured.

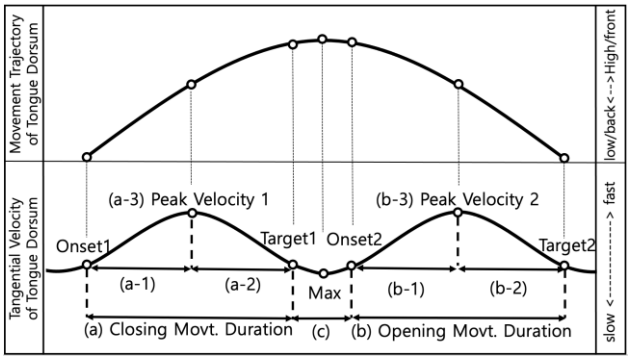


Figure 1. Schematized tongue dorsum movement trajectory measures and tangential velocity

Linear mixed-effects models (LMMs)[1] were created with the *lme4* package ([2]), followed by the post-hoc Tukey's pairwise comparisons via the

emmeans package ([16]) in R ([15]). *Gender* (male and female) and *Consonant* (lenis, aspirated and fortis) were employed as fixed factors and *Speaker* as a random effect. Given that the gender-related difference is a crucial aspect of our research question, we performed post-hoc Tukey tests with Tukey correction to examine the *Consonant* effect separately for male and female groups. We carried out these tests even in the absence of significant interactions found in the initial linear-mixed effects models to ensure a complete analysis.

3. RESULTS

Results of LMMs show that there were no significant effects either of *Consonant* or of *Gender* on the acceleration and deceleration duration and peak velocity in the closing movement, but significant effects were obtained on the peak velocity and deceleration duration in the opening movement, the opening movement duration, and closure duration. Their mean values and the standard deviations are given in Table 1.

Table 1. Mean values and the standard deviations in parentheses of Closure Duration, Peak Velocity 2, Deceleration Duration 2, and Opening Movement Duration

	Closure Duration (ms)			Peak Velocity 2 (cm/sec)		
	k	k^h	k^*	k	k^h	k^*
Female	63.7 (43.9)	68.3 (35.1)	108 (78.3)	13.1 (2.67)	12.0 (1.93)	14.6 (3.30)
Male	61.3 (29.3)	71.6 (34.3)	94.6 (46.0)	17.7 (4.97)	15.7 (3.35)	20.7 (4.91)
	Deceleration Duration 2 (ms)			**Opening Movt. Duration (ms)**		
	k	k^h	k^*	k	k^h	k^*
Female	94.6 (30.4)	74.3 (23.4)	73.4 (29.3)	160 (39.6)	142 (27.2)	138 (34.5)
Male	85.3 (21.7)	76.2 (22.9)	54.8 (22.6)	151 (37.3)	150 (42.8)	114 (22.4)

Closure Duration showed a significant effect of *Consonant* (β=39.57, SE=11.06, p<0.001), indicating that the fortis stop had a longer closure duration than the aspirated stop. There was no significant effect of *Gender*, and no interaction between *Consonant* and *Gender*. However, separate analyses of *Consonant* effects across genders revealed some gender-related differences. Post-hoc comparisons, with Tukey correction, of the three velar stops for Closure Duration suggested that, for females, the fortis stop had a significantly longer closure duration than the lenis stop (β=44.1, SE=10.9, p<0.001) and the aspirated stop (β=-39.6, SE=15.8, p<0.05). However, there was no significant difference between the aspirated and lenis stops (/k/, /k^h/</k*/). For males, the fortis stop displayed a significantly longer duration than the lenis stop (β=-33.3, SE=10.9, p<0.01), but there was no significant difference between the lenis and aspirated stops, nor between the aspirated and fortis stops (/k/</k*/).

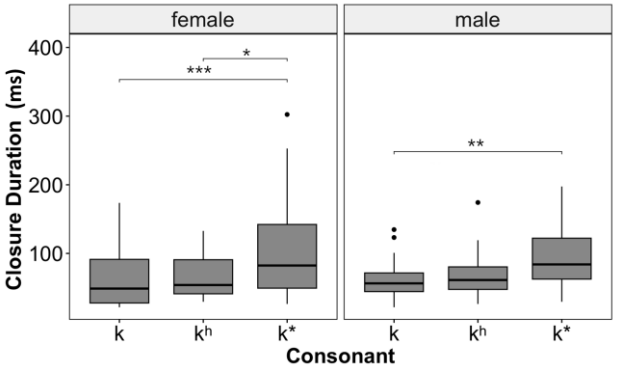


Figure 2. Closure Duration for different stop categories. Error bars indicate standard errors.

Peak Velocity 2 of the opening movement shows main effects of *Consonant* (β=2.67 SE= 0.76, p<0.001) and *Gender* (β= 3.73, SE= 1.40, p<0.05). There was also an interaction between *Consonant* and *Gender* (β=2.39, SE=1.08, p<0.05). Follow-up post-hoc comparisons (with Tukey correction) of the three velar stops for Peak Velocity 2 of the opening movement showed that for *females*, the fortis stop had significantly faster peak velocity than the aspirated stop (β =-2.67, SE=0.87, p<0.01), but there was no significant difference between the lenis and aspirated stops, nor between the lenis and fortis stops (/k^h/</k*/). For *males*, the three velar stops were presented with significantly different peak velocities of the opening movement (lenis and aspirated: β =-2.01, SE=0.73, p<0.05; aspirated and fortis: β =-5.06, SE=0.87, p<0.001; lenis and fortis: β =3.05, SE=0.95, p<0.01) in the following order: /k^h/</k/</k*/.

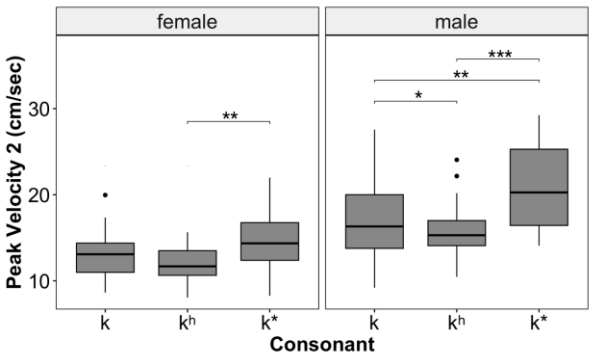


Figure 3. Peak velocity 2 for different stop categories. Error bars indicate standard errors.

Deceleration Duration 2 of the opening movement shows main effects of *Consonant* (β =20.28, SE=6.36, p<0.01), but there was no significant effect of *Gender*. There was an interaction between *Consonant* and

Gender (β=-20, SE=8.99, p<0.05). Post-hoc comparisons (with Tukey correction) of the three velars for the deceleration duration of the opening movement indicated that for *females*, the lenis stop showed significantly longer deceleration duration than the aspirated stop (β=-20.3, SE=6.95, p<0.01) and the fortis stop (β=-21.2, SE=6.2, p<0.01), but there was no significant difference between the aspirated and fortis showing the following order: /k^h/, /k*/</k/. For *males*, the fortis stop was presented with the shorter deceleration duration of the opening movement than the lenis (β=-30.5, SE=6.2, p<0.001) and aspirated stops (β=21.42, SE= 6.35, p<0.01) (/k/, /k^h/</k*/).

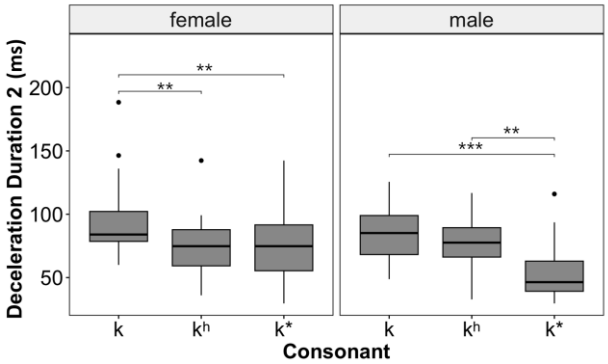


Figure 4. Deceleration Duration 2 of the opening movement for different stop categories. Error bars indicate standard errors.

Opening Movement Duration shows no main effects of *Consonant* and *Gender*. Nonetheless, a trend effect of *Consonant* was observed (β=17.66, SE=9.12, p=0.056). The interaction effect of *Consonant* and *Gender* was significant (β=-31.48, SE= 12.90, p< 0.05).

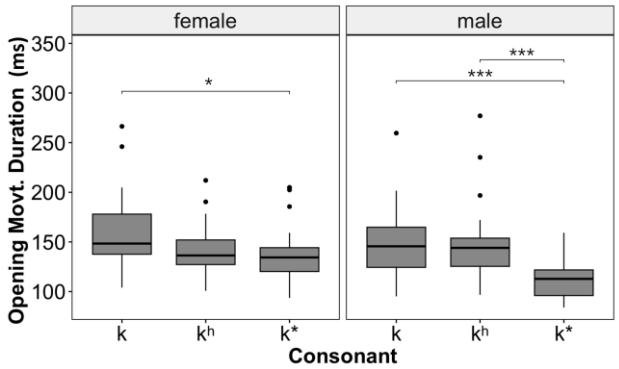


Figure 5. Opening Movement Duration for different stop categories. Error bars indicate standard errors.

Post-hoc comparisons (with Tukey correction) of the three velar stops for the movement duration of the opening movement indicated that for *females*, the fortis had a significantly shorter opening movement duration than the lenis stop (β=-21.6, SE=8.69, p< 0.05) (/k^*/</k/). For *males*, the fortis stop had a significantly shorter opening movement duration than the lenis stop (β=-36.7, SE=8.69, p< 0.001) and the aspirated stop (β=35.39, SE=8.68, p< 0.001) in the following order: /k*/</k/, /k^h/.

Forward movements of TD did not show significant differences between the consonants but there was a significant difference between genders showing more forward movement in *males* than in *females* (β: 3.10, SE= 0.53, p<0.001). There was no significant result in vertical movements, either, showing no effect of *Consonant* or *Gender*.

4. SUMMARY AND CONCLUSION

The results from Closure Duration showed no three-way contrast for either females or males: females showed a binary contrast (/k/=/k^h/</k*/) while males showed a difference only between the lenis and fortis stops (/k/</k*/). For Peak Velocity 2, males showed the three-way contrast (/k^h/</k/</k*/) while females show the difference only between the aspirated and fortis (/k^h/</k*/). As for Deceleration Duration 2 of the opening movement, females showed the binary contrast with no evidence of the aspirated and fortis distinction (/k^h/=/k*/</k/) while males showed a binary distinction with no contrast between the lenis and aspirated (/k/=/k^h/</k*/). The results from Opening Movement Duration indicated that females showed a distinction only between the lenis and fortis stops (/k^*/</k/) whereas males displayed the binary distinction without a contrast of the lenis and aspirated (/k*/</k/=/k^h/). However, there was no significant result between the three consonants in the vertical and forward movements.

Turning to the results of closure duration, the results appear to illustrate a trade-off between closure duration and the following VOT, such that fortis stops with the shortest VOT are presented with the longest closure duration. However, there was no distinction for the closure duration between the lenis and aspirated stops. This may imply that neutralized closure duration is caused by the loss of lenis-aspirated distinction in VOT, and thereby there were no hold duration-related differential impacts on supralaryngeal articulation between the lenis and aspirated stops. Another interesting finding is that a three-way contrast still remained for Peak Velocity 2 of the opening movement in males in the reverse order of the traditional three-way VOT contrast (/k^h/</k/</k*/), but females did not show the same three-way distinction. This may suggest that males are more conservative than females when it comes to the on-going sound change of laryngeal contrast even in supralaryngeal articulation.

It is worth noting that the results of the present study did not show any differences in kinematic measures associated with the consonantal closing movement in CV consonants, despite previous findings by [3] that reported consonantal effects on the closing movement kinematics in VCV contexts. This discrepancy may be due to the time gap between the two studies or the difference in syllable structure, or a combination of both factors.

With these findings taken together suggest the current study offers some evidence of disappearing supralaryngeal articulatory distinction of the three-way stop contrast in velars along with the on-going sound change related to the VOT merger. Further, the results also support the view that the lenis-aspirated kinematic distinction is disappearing and led by female speakers. However, our results must be taken with caution and cannot be fully generalized because the current study was based on stops in monosyllable words produced in isolation. Therefore, it remains to be seen how our findings can be generalized as more data become available.

5. ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884).

6. REFERENCES

[1] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., & Yoon, T. J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120-144.

[2] Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

[3] Brunner, J., Fuchs, S., & Perrier, P. 2011. Supralaryngeal control in Korean velar stops. *Journal of Phonetics*, *39*(2), 178-195.

[4] Cho, T. 1996. Vowel correlates to consonant phonation: an acoustic-perceptual study of Korean obstruents. MA thesis, University of Texas at Arlington.

[5] Cho, T., & Ladefoged, P. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics*, *27*(2), 207-229.

[6] Cho, T., Jun, S. A., & Ladefoged, P. 2002. Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of phonetics*, 30(2), 193-228.

[7] Cho, T., Son, M., & Kim, S. 2016. Articulatory reflexes of the three-way contrast in labial stops and kinematic evidence for domain-initial strengthening in Korean. *Journal of the International Phonetic Association*, *46*(2), 129-155.

[8] Choi, J., Kim, S., & Cho, T. 2020. An apparent-time study of an ongoing sound changes in Seoul Korean: A prosodic account. *Plos one*, *15*(10).

[9] Kang, Y. 2014. Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. Journal of Phonetics, 45, 76-90.

[10] Kim, H., Maeda, S., & Honda, K. 2010. Invariant articulatory bases of the features [tense] and [spread glottis] in Korean plosives: New stroboscopic cine-MRI data. *Journal of Phonetics*, *38*(1), 90-108.

[11] Kim, M. R. 2013. Interspeaker variation on VOT merger and shortening in Seoul Korean. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 060212). Acoustical Society of America.

[12] Ladefoged, P., & Cho, T. 2001. Linking linguistic contrasts to reality: The case of VOT. *TravauxDu CercleLinguistiqueDe Copenhague*, *31*.

[13] Lisker, L., & Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word, 20(3), 384-422.

[14] Maddieson, I. 1997. Phonetic Universals. In 1he handbook of phonetic sciences (J. Laver & W. J. Hardcastle, editors), pp. 619-639. Oxford: Blackwells.

[15] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[16] Russell V. Lenth, 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.5.

[17] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(2), 287-308.

[18] The Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS). 2022, Dynamics of Speech Production through Articulatory DB Construction.

[1] Full models

(i) ClosureDuration~gender*consonant+(1|speaker)

(ii) PeakVelocity2~gender*consonant+(1+gender | speaker)

(iii) DecelerationDuration2~gender*consonant+(1+gender |speaker)

(iv) OpeningMovementDuration~gender*consonant+(1+gender | speaker)

‘O LO/R/D, OPEN THOU OU/R/ LIPS’: RHOTICITY IN CHORAL SINGING FROM GLASGOW AND CAMBRIDGE

Edward Marshall

Glasgow University Laboratory of Phonetics, University of Glasgow
Edward_Marshall1@outlook.com

ABSTRACT

Front vowel phonology and realization in British choral singing is thought to be based on the non-regional standard accent SSBE. Is choral singing affected by regional variation in speech? Scottish Standard English is rhotic; SSBE is non-rhotic. Do Scottish choirs produce postvocalic /r/ in choral singing? This study compares the realization of postvocalic /r/ in choirs from Glasgow and Cambridge (1925-2017). 5016 tokens were auditory coded for rhoticity (presence/absence of /r/) and analyzed with Bayesian binomial mixed models. The predicted probability of producing /r/ was 0.68 for Glasgow and 0.10 for Cambridge. Furthermore, Glasgow produced linking /r/ categorically whereas Cambridge was variable ($\sim$ 60%). The difference between the choirs decreased over time perhaps reflecting a change in the spoken accent of the Glasgow choir director. British choral singing may be partly based on SSBE, but may also be influenced by regional features.

Keywords: rhotics, phonology, singing, sociophonetic, standard accent, SSBE

1. INTRODUCTION

Previous sociolinguistic work has found that the front vowels of choral singing have lowered over time [1] following a well-evidenced pattern of change observed in Received Pronunciation [2]. This supports the notion that choral singing is based on a standard form of the language as suggested by musicologists [3–5]. In [1], for example, choirs from different regions showed shared front vowel phonology, suggesting a generalized non-regional singing style. The question remains, are there any phonological differences between these choirs that *are* affected by dialect area? In this paper, I investigate whether choirs produce postvocalic /r/ in choral singing in a non-rhotic dialect area, Cambridge, and a rhotic dialect area, Glasgow (c.f. [2]).

For postvocalic /r/ in English (e.g. *car*, *card*), both the type of realization and whether /r/ is produced at all can vary between speakers and speaker groups making it the ideal variable for exploring variation by dialect area. In speech there is a strong effect of phonetic context on whether postvocalic /r/ is produced or not. Articulated /r/ is most likely pre-vocalically (e.g. *car and*), then pre-consonantally (e.g. *car could*), and least likely pre-pausally (e.g. *car#*) [6].

Scottish Standard English (SSE) is considered a rhotic variety [2, 7, 8] meaning that postvocalic /r/ is usually articulated in Scottish varieties of English [6], though there is some evidence that the frequency and strength of the variants produced is reducing over time [9]. In SSE, /r/ in words like *car* would be articulated e.g. /kar/, and could be phonetically realized as a post-alveolar approximant [ɹ], retroflex approximant [ɻ] (tip-up), or bunched [10], as a tap [ɾ], or as a trill [r] [11]. In SSE, historically we would expect a higher frequency of trills and taps compared to SSBE where we would expect more approximants [2].

In contrast, Standard Southern British English (SSBE) is a non-rhotic variety of English meaning that postvocalic /r/ is usually *not* produced apart from pre-vocalically [2].In other words, in SSBE the /r/ in the word *car* would not be produced at all e.g. /kɑ:/, apart from when followed by a vowel (e.g. *car and*, which could be realized as [kɑɹand]), known as linking /r/.

Here I present an empirical phonological analysis: is choral singing rhotic or non-rhotic? If the choral accents in the Glasgow and Cambridge singer groups relate to the spoken accent in each region, we would predict articulated postvocalic /r/ in Scotland but not in Southern England. If the choral accents are not related to the spoken accent of each region, we would expect both groups to produce postvocalic /r/ in linking /r/ contexts (e.g. *car and*)

1.1. Previous linguistic studies of /r/ in singing

Previous sociolinguistic analyses of singing have focused on auditory coding of consonant realizations in solo singing [12–15] using a variationist Labovian approach [16]. In the first

sociolinguistic study of singing, investigating multiple variables including postvocalic /r/ in recordings of British rock/pop bands [12] found that 'words such as *girl, more* tend to be pronounced with an /r/ even by those English English speakers (the majority) who do not have non-prevocalic /r/ in their speech' (pg. 142). Trudgill suggested the insertion of /r/ in words where they would not be found in speech was a form of hypercorrection, arguing that this was a form of accommodation towards an imagined American audience.

[17] moved away from thinking about variation in singing in a purely linguistic or conversational model, towards incorporating the notion of performance. Morrissey argued that certain features had become enregistered as appropriate for the performance of a certain genre, and that deviating from those features would essentially be considered unstylistic by listeners. These overarching styles [17] calls 'reference styles'.

More recently, in the first study of singing in Scotland, [14] investigated postvocalic /r/ in indie artists from Glasgow. They find that 'overall, there is a high rate of the variants at the weakly, rather than the strongly rhotic end of our continuum. This is despite postvocalic /r/ being a classic stereotype of Scots.' [14] (pg. 228). The authors attributed this surprising finding to the reduction of postvocalic /r/ in working class speech in the Central Belt [9].

In all studies of postvocalic /r/ in singing to date, the style of music investigated has principally been popular forms such as pop, rock, heavy metal, and indie. What we know about realization of postvocalic /r/ in classical singing is limited to classical singing pronunciation guides e.g. [18]. Previous studies have all investigated genres of singing which fall under the same popular 'reference style'. What are the constraints on the realization of postvocalic /r/ within the classical reference style?

1.2. Research questions

Evidence from front vowels suggests that British choral singing is based on a non-regional standard accent like SSBE [1]. However, *is choral singing affected by regional variation in speech?* In order to answer this broader question, this study asks; *do Scottish choirs produce postvocalic /r/ in choral singing? Is there evidence of change over time in their production of postvocalic /r/?*

2. METHODOLOGY

In order to answer these research questions two electronic time-aligned corpora were constructed in LaBB-CAT [19]. The Glasgow corpus consists of commercially released recordings of the Glasgow Orpheus (1906-1951) and Glasgow Phoenix (1951-present) choirs with audio recordings from 1925 to present day. The King's corpus consists of commercially released recordings and public broadcasts of the choir of King's College, Cambridge, with audio recordings from 1945 to 2019. Audio recordings and texts were aligned in LaBB-CAT [20] using Praat [21] and all tokens of /r/ in the corpus (8407 tokens) were auditorily coded for realization. This yielded 5016 tokens of postvocalic /r/. As the realizations were vastly dominated by approximants ($\sim$ 95% of the realized tokens of postvocalic /r/), I collapsed the levels to a binary factor (Realized vs No realization). The data were analysed with Bayesian binomial mixed models using `brms` [22] [version 2.18.0] in R [23] [version 4.1.2]. Tables were produced with `xtable` [24] [version 1.8-4] and `BayesPostEst` [25] [version 0.3.2].

2.1. Linguistic context

There are four main contexts for postvocalic /r/ as shown in Table 1.

Table 1: postvocalic /r/ contexts

coding	example	name
'r.V'	*car and*	pre-vocalic (linking /r/)
'r'	*car#*	pre-pausal
'r.C'	*car could*	pre-consonantal word boundary
'rC'	*card*	pre-consonantal within word

The four contexts are listed in the order of most likelihood of rhoticity to least likelihood [9]. i.e. /r/ is most likely to be produced in *car and* (linking /r/ context) in all varieties of English, including SSE and SSBE, whereas /r/ in *card* is least likely to be produced. In Table 2, rhoticity is broken down by Corpus and Context.

Table 2: Rhoticity by Corpus and Context

Corpus	Context	/r/	No /r/	/r/(%)
Glasgow	r	102	89	53%
-	r.C	329	282	54%
-	r.V	176	13	93%
-	rC	247	459	35%
Cambridge	r	6	273	2%
-	r.C	126	1046	11%
-	r.V	219	158	58%
-	rC	130	1361	9%

2.2. Modelling, priors, and convergence criteria

The data were modelled using Bayesian binomial mixed models. The model structures were:

```
  Combined_rhoticity ~ Corpus + Context
+ Corpus:Context + (1|Album) + (1|Song)
+ (1|Corpus:Word)

  Glasgow_rhoticity ~ Time/Director
+ Context + Time/Director:Context +
(1|Album) + (1|Song) + (1|Director:Word)
```

The dependent variable of both models is Rhoticity (Realization/Non Realization of postvocalic /r/). The predictor variables for the Combined model were the factors Corpus (Cambridge, Glasgow) and Context (r.V, r, r.C, rC). For the Glasgow only model, Context is the same, but instead of Corpus we have Time/Director which is a three-level factor: 'HSR' = Hugh S. Roberton (1925-1951), 'PM' = Peter Mooney (1959-1975), 'MJS' = Marilyn J. Smith (1987-2016).

Following recommendations by [26], I used weakly-informative regularising priors using Cauchy distributions centred on 0 with a scale factor of 2.5 for all fixed and varying effects and a cauchy distribution centred on 0 with a scale factor of 10 for the intercept. Cauchy distributions were used as opposed to normal distributions as they have more weight in the tails and therefore allocate more probability space to values further from zero while still giving values closer to zero greater probability [27]. Model chains were visually inspected for convergence, Rhat was 1 for all coefficients and the minimum effective sample size for all coefficients was greater than 100 times the number of chains. I was satisfied that the models converged successfully and that the posterior summaries are amenable to interpretation. All factors are sum coded for ease of interpretation.

3. COMBINED (SYNCHRONIC) RESULTS

As seen in Table 3, the combined rhoticity model revealed a main effect of Corpus; the predicted probability of producing postvocalic /r/ was 0.10 for Cambridge and 0.68 for Glasgow (logit difference: 1.47, 95% CI [1.22, 1.73]). The model supports an interaction of Corpus by Context. As articulated /r/ is least likely pre-pausally, there is a positive adjustment for pre-pausal tokens in the Glasgow corpus (logit difference 0.62, CI [0.25; 1.03]). As the likelihood of articulated /r/ is so high for Glasgow overall, there is a negative adjustment for pre-consonantal contexts (logit difference −0.42, CI [−0.67; −0.18]). The predicted probability for articulating postvocalic /r/ in pre-pausal context *car#* (context_r) was 0.025 for Cambridge and 0.47 for Glasgow. The predicted probability of articulating postvocalic /r/ in pre-consonantal context across word boundary *car could* (context_r.C) was 0.09 for Cambridge and 0.59 for Glasgow (interaction visualized in Fig. 1).

Table 3: Combined rhoticity model summary

	Estimate	l-95% CI	u-95% CI
Intercept	**−0.70**	**−0.95**	**−0.45**
CorpusGlasgow	**1.47**	**1.22**	**1.73**
Context_r	**−1.51**	**−1.92**	**−1.14**
Context_r.C	−0.20	−0.41	0.01
Context_rC	**−0.99**	**−1.24**	**−0.74**
CorpusGlasgow:Context_r	**0.62**	**0.25**	**1.03**
CorpusGlasgow:Context_r.C	−0.19	−0.40	0.02
CorpusGlasgow:Context_rC	**−0.42**	**−0.67**	**−0.18**

Point estimate displayed: median
Results are given on the log odds ratio (not the response) scale.
Bold font indicates 0 outside 95% credible interval.

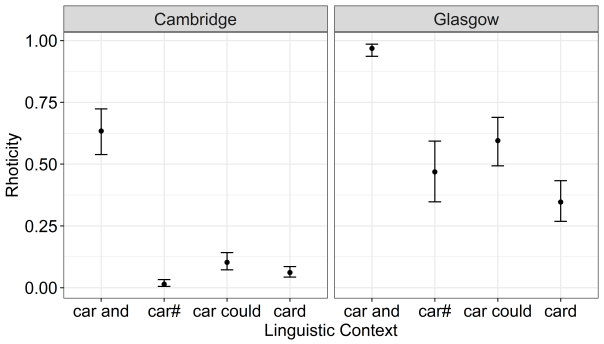


Figure 1: Combined rhoticity model: Corpus by Context interaction. y-axis: estimated proportion of rhoticity

4. GLASGOW (DIACHRONIC) RESULTS

Both King's and Glasgow data were modelled separately to investigate change over time. King's showed stable variation in linking /r/ *car and* contexts only; and only minimal /r/ produced in the other linguistic contexts; there was no evidence of change over time; model omitted for brevity.

In a model of the Glasgow data (as found in Table 4) there is a main effect of Time/Director. For Time/Director Peter Mooney (1959-1975), the predicted probability of producing postvocalic /r/ is 0.7. The predicted probability for postvocalic /r/ being produced for Marilyn J. Smith (1987-2016) is 0.46. The credible interval for Marilyn J. Smith

Table 4: Glasgow rhoticity model summary

	Estimate	l-95% CI	u-95% CI
Intercept	**0.73**	**0.24**	**1.24**
Director_PM (1959-1975)	0.10	−0.46	0.66
Director_MJS (1987-2016)	**−0.89**	**−1.65**	**−0.18**
Context_r	**−1.48**	**−2.18**	**−0.89**
Context_r.C	−0.12	−0.54	0.30
Context_rC	**−1.44**	**−1.89**	**−1.01**
Director_PM:Context_r	0.64	−0.02	1.38
Director_MJS:Context_r	**−1.73**	**−2.95**	**−0.75**
Director_PM:Context_r.C	−0.38	−0.85	0.09
Director_MJS:Context_r.C	0.32	−0.34	0.98
Director_PM:Context_rC	−0.15	−0.65	0.35
Director_MJS:Context_rC	0.36	−0.32	1.05

Point estimate displayed: median

Results are given on the log odds ratio (not the response) scale.

Bold font indicates 0 outside 95% credible interval.

does not include zero (logit difference −0.89, 95% CI −1.65; −0.18) meaning that it is different to the grand mean for Time/Director. As Peter Mooney (1959-1975) is not different to the grand mean but Marilyn J. Smith (1987-2016) is, there is evidence of change over time. That is, postvocalic /r/ is least likely to be produced in the later time period.

The model also supports an interaction of Time/Director by Context. The interaction is driven by Marilyn J. Smith (1987-2016) and the pre-pausal Context *car#* as shown in Fig. 2. The late time period Marilyn J. Smith (1987-2016) is much less likely to produce /r/ in pre-pausal Context than previous Time/Director pairs. The predicted probability of producing /r/ in pre-pausal context for Marilyn J. Smith was 0.03.

5. DISCUSSION

Glasgow choirs produce more rhoticity than King's in all linguistic contexts as we might expect for speakers of Standard Scottish English. It appears that the Glasgow choral singers are using their underlying SSE phonology which contains postvocalic /r/ in both linking /r/ and non-linking /r/ contexts. King's underlying phonology based on SSBE does not have postvocalic /r/, apart from in the linking /r/ context. Linking /r/ at King's is articulated approximately 60% of the time leaving a very high proportion of hiatus compared to rates in speech [28], perhaps evidence of corrected RP [2]. While evidence from vowels may indicate a homogeneous standard of British choral singing based on a non-regional accent like SSBE [1] (perhaps evidence of a classical reference style [17]), this work demonstrates that a choir's sung phonology may also be impacted by the phonology of the spoken accent of the singers and/or their choir directors.

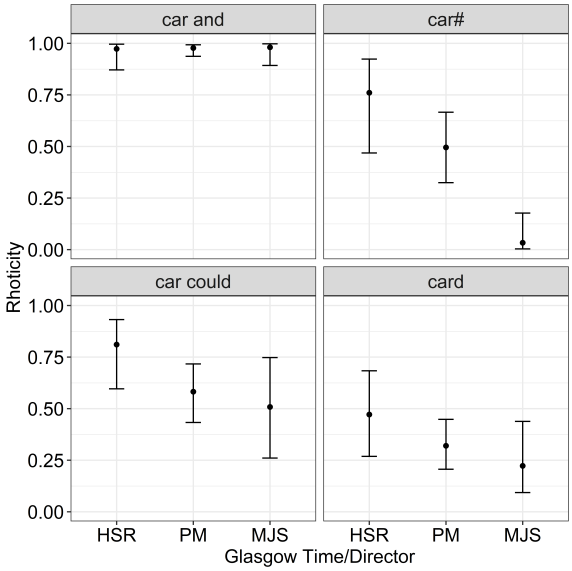


Figure 2: Glasgow rhoticity model interaction of Time/Director by Context. y-axis: estimated proportion of rhoticity. 'HSR' = Hugh S. Roberton 1925-1951; 'PM' = Peter Mooney 1959-1975; 'MJS' = Marilyn J. Smith 1987-2016

Change over time in the Glasgow data, with a reduction in production of postvocalic /r/ in pre-pausal contexts may provide evidence of levelling. One explanation is that choir's director in the later time period speaks with an SSBE accent. Whether the singers imitate the director, or their pronunciation is "corrected" in this case is unknown. The reduction of postvocalic /r/ in pre-pausal context could also be evidence of a wider pattern of levelling to a non-rhotic norm in British choral singing. The reduction of postvocalic /r/ also reflects the pattern of derhoticization in Scottish English [9] as found by [14] in indie bands in Glasgow. However, it is less likely that an ongoing change in a working-class variety has influenced the sung phonology within the classical reference style compared with indie, as channeling 'authenticity' has been recognised as integral to indie as a genre [29], perhaps suggesting it would more readily incorporate phonetic features connected with working-class identities. Future research is needed to establish whether the reduction rhoticity in pre-pausal contexts is restricted to these choirs and directors, or if it is part of a wider pattern of change in choral singing in Scotland.

6. REFERENCES

[1] E. Marshall, J. Stuart-Smith, J. Butt, and T. Dean, "Variation and change over time in Standard British Choral Accent (1925-2019)," 2023, in review.

[2] J. C. Wells, *Accents of English*. Cambridge University Press, 1982, vol. 1.
[3] J. Potter, *Vocal Authority: Singing Style and Ideology*. Cambridge University Press, 1998.
[4] J. Sagrans, "Early music and the choir of King's College, Cambridge, 1958 to 2015," Ph.D. dissertation, Schulich School of Music, McGill University, Montreal, 2016.
[5] T. Day, *I Saw Eternity the Other Night - King's College, Cambridge, and an English Singing Style*. Allen Lane, 2018.
[6] J. Stuart-Smith, "The phonology of modern urban Scots," in *The Edinburgh companion to Scots*, J. Corbett, J. D. McClure, and J. Stuart-Smith, Eds. Edinburgh University Press, 2003, pp. 110–137.
[7] D. Abercrombie, "The accents of Standard English in Scotland," in *Languages of Scotland*, A. J. Aitken and T. MacArthur, Eds. Chambers of Edinburgh, 1979.
[8] T. Jauriberry, "Variation and change of middle-class /r/ in Standard Scottish English," *Lingua*, vol. 256, p. 103059, 2021.
[9] J. Stuart-Smith and E. Lawson, "Scotland: Glasgow and the Central Belt," in *Listening to the Past: Audio Records of Accents of English*, ser. Studies in English Language. Cambridge University Press, 2017, pp. 171–198.
[10] E. Lawson, J. Stuart-Smith, and J. M. Scobbie, "The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1646–1657, 2018. [Online]. Available: https://doi.org/10.1121/1.5027833
[11] D. Watt, C. Llamas, and D. E. Johnson, "Sociolinguistic variation on the Scottish-English border," in *Sociolinguistics in Scotland*, R. Lawson, Ed. London: Palgrave Macmillan UK, 2014, pp. 79–102. [Online]. Available: https://doi.org/10.1057/9781137034717_5
[12] P. Trudgill, *On Dialect: Social and Geographical Perspectives*. Blackwell, 1983.
[13] P. Simpson, "Language, culture and identity: With (another) look at accents in pop and rock singing," *Multilingua - Journal of Cross Cultural and Interlanguage Communication*, vol. 18, pp. 343–367, 1999.
[14] M. Krause and J. Smith, "'I stole it from a letter, off your tongue it rolled.' the performance of dialect in Glasgow's indie music scene," in *Language and a Sense of Place: Studies in Language and Region*, C. Montgomery and E. Moore, Eds. Cambridge University Press, 2017.
[15] C. Caillol and E. Ferragne, "The sociophonetics of British heavy metal music: T voicing and the FOOT-STRUT split," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australiasian Speech Science and Technology Association, Inc., 2019.
[16] W. Labov, *Sociolinguistic Patterns*. University of Pennsylvania Press, 1972.
[17] F. A. Morrissey, "Liverpool to Louisiana in one lyrical line: Style choice in British rock, pop and folk singing," in *Standards and Norms in the English Language*, M. A. Locher and J. Strässler, Eds. De Gruyter Mouton, 2008.
[18] D. Adams, *A Handbook of Diction For Singers: Italian, German, French*, 2nd ed. Oxford University Press, 2008.
[19] R. Fromont and J. Hay, "LaBB-CAT: an annotation store," in *Proceedings of Australasian Language Technology Association Workshop*. Australasian Language Technology Association, 2012, pp. 113–117.
[20] R. Fromont, "Forced alignment of different language varieties using LaBB-CAT," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australiasian Speech Science and Technology Association, Inc., 2019.
[21] P. Boersma and D. Weeninck, *Praat: doing phonetics by computer [computer program]. Version 6.0.43*, 2018. [Online]. Available: http://www.praat.org/
[22] P. C. Bürkener, "Advanced Bayesian multilevel modeling with the R package brms." *The R Journal*, vol. 10, pp. 395–411, 2018.
[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org/
[24] D. B. Dahl, D. Scott, C. Roosen, A. Magnusson, and J. Swinton, *xtable: Export Tables to LaTeX or HTML*, 2019, r package version 1.8-4. [Online]. Available: https://CRAN.R-project.org/package=xtable
[25] J. Karreth, S. Scogin, R. Williams, and A. Beger, *BayesPostEst: Generate Postestimation Quantities for Bayesian MCMC Estimation*, 2021, r package version 0.3.1. [Online]. Available: https://CRAN.R-project.org/package=BayesPostEst
[26] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, vol. 2, pp. 1360–1383, 2008.
[27] A. E. Kimball, K. Shantz, C. Eager, and J. Roy, "Confronting quasi-separation in logistic mixed effects for linguistic data: A Bayesian approach," *Journal of Quantitative Linguistics*, vol. 26, no. 3, pp. 231–255, 2019.
[28] J. A. Mompeán and A. F. Gómez, "Hiatus resolution strategies in non-rhotic English: The case of/r/-liaison." in *ICPhS*, 2011, pp. 1414–1417.
[29] J. C. Beal, "'You're not from New York City, you're from Rotherham' dialect and identity in British indie music," *Journal of English Linguistics*, vol. 37, pp. 233–240, 2009.

STRESS CONDITIONED PHONOLOGICAL PROCESS: A CASE STUDY OF ITALIAN PALATALIZATION

[1,2] Bowei Shao, [2] Philipp Buech, [2] Anne Hermes, [1] Maria Giavazzi

[1] Département d'études cognitives, École Normale Supérieure, Université PSL, Paris, France
[2] Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/Université Sorbonne Nouvelle, Paris, France
{bowei.shao; maria.giavazzi}@ens.psl.eu, {philipp.buech; anne.hermes}@sorbonne-nouvelle.fr

ABSTRACT

This study examines the interaction between lexical prominence and palatal/velar consonants in Italian. The specific phenomenon of interest is the palatalization of /k/ in, e.g., 'cardiaci' [kar.ˈdi.a.**tʃ**i] but not in 'ubriachi' [u.bri.ˈa.**k**i], which is argued to be conditioned by the position of /k/ relative to lexical prominence. The aim of the study is to uncover potential mechanisms that block palatalization in post-stress position, as in [u.bri.ˈa.**k**i]. The results revealed that the stressed vowel /a/ is more peripheral, i.e., hyper-articulated compared to unstressed /a/, which is more closed and fronted, especially towards its end. Further, the immediately post-stress consonants show longer closure durations compared to far-from-stress. We argue that the blocking of velar palatalization could be related to the longer closure duration in post-stress context, which is an articulatory by-product of the hyper-articulated stressed vowel. These findings are discussed within the predictions of the Task Dynamic model for the *μ-gesture.*

Keywords: Lexical prominence, stress, Italian palatalization, *μ*-gesture, acoustics.

1. INTRODUCTION

Prosodically prominent positions within the word are known (1) to be privileged, manifesting the positional maintenance of contrasts otherwise neutralized and the resistance to processes applying elsewhere. For instance, vocalic contrasts are preferentially realized in stressed position, but they may be reduced in unstressed position (e.g., in English, Brazilian Portuguese, Western Catalan [1, 2, 3]). Similar patterns are observed for consonantal contrasts (e.g., in Copala Trique, Italian, and Finnish [4, 5, 6]). Prominent positions are however also (2) the preferred target for a small class of frequent processes: for instance, consonants are often lengthened in pre-tonic and post-tonic positions (as in English and Somali [7, 8]), consonants in these two positions are described as having louder (e.g., Farsi [9, 10]) or affricate-like bursts (e.g., Maori [11]).

The two seemingly contradictory linguistic behaviours, (1) and (2), may have a common phonetic basis: they may be the articulatory by-product of lexical prominence. The nucleus bearing lexical prominence is known to be longer in duration, higher in f0, more peripheral and more intense [12]. While f0, duration and intensity are less related to oral cavity activities, the hyper-articulated vowel qualities related to lexical prominence shall manifest on the lingual articulation of the nucleus and have effects on adjacent segments [13, 14]. This interaction between lexical prominence and segmental processes can be illustrated by velar palatalization in Italian masculine plural nouns and adjectives. This process is triggered by /i/, and conditioned by lexical prominence [15, 16, 4]. For example:

(3) FAR: cardiaci /kar.ˈdi.a.k-i/→[kar.ˈdi.a.tʃi] 'cardiac'
POST: ubriachi /u.bri.ˈa.k-i/→[u.bri.ˈa.ki] 'drunk'

In (3) FAR, stress is located on the second syllable of the word and considered here as far from the target of palatalization /k/ (hereafter FAR), the /k/ is palatalized to /tʃ/. However, in (3) POST, palatalization is blocked in the immediate post-stress position (hereafter POST). Overall, the examples in (3) illustrate the triggering and the blocking of palatalization in Italian. The triggering of palatalization by the following /i/ is argued to be both acoustically and auditorily motivated [17, 18, 19]: the velar stop /k/ and the palato-alveolar affricate /tʃ/ before a palatal vowel /i/ share acoustic similarities, which have been analyzed as the cause of perceptual reanalysis by speakers/listeners. The reason for the blocking of palatalization, however, is still unclear.

In this study, we aim to investigate how the blocking of palatalization could be explained in relation to lexical prominence by analyzing the acoustic characteristics of stressed vowels and the following segments in two stress conditions (POST and FAR). The hypothesis is that a tempo-spatial modulation gesture related to lexical prominence, i.e., *μ*-gesture [14], may have played an important role in modulating the articulatory behaviour of both the stressed vowel and the consonant in POST condition.

This prominence related gesture modulation, in turn, has an impact on the perceptual contrast between /k, g/ *vs.* /tʃ, dʒ/.

2. METHOD

2.1. Material and speakers

Target words were trisyllabic nonce words, structured /$C_1V_1.C_2V_2.C_3V_3$/, differing solely by the position of stress on the first or the second syllable (e.g., /ˈpi.ta.ki/, /pi.ˈta.ki/). The nonce words were designed to compare how the target consonants /k, g, tʃ, dʒ/ (in C_3 position) were produced in both FAR and POST contexts. The V_3 position is occupied by /i/. $C_1V_1.C_2V_2$ sequences were /pita/, /fesa/, /pufa/ /tipa/ and /suta/. For example, in the nonce words /ˈpi.ta.ki/ and /pi.ˈta.ki/, C_3 /k/ was in FAR and POST contexts respectively.

The nonce words were written in their orthographic forms with stress location marked by an acute lexical stress accent (e.g., *pítachi, pitáchi*). They were embedded in a carrier phrase *Dimmi* ___ *di nuovo* ('Say ___ again'). The phrases were randomized and presented on a computer screen.

The acoustic recordings were conducted in a double-walled sound booth on a TASCAM DR-100MKIII Linear PCM Recorder. Sound files were recorded at a sampling rate of 44.1 kHz. The phrases were repeated three times by 18 speakers, yielding 3328 tokens (two speakers were excluded due to misplacement of stress and dialectal background).

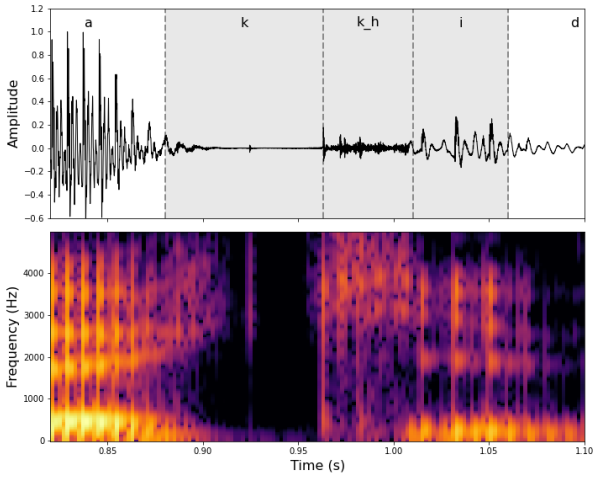


Figure 1: Segmentation of the target sequence '*áchi*' in the utterance '*Dimmi pitáchi di nuovo'*.

Utterances were pre-segmented using the Label Sounds function of Audacity v3.1.3 [20]. They were then extracted and automatically annotated using the Montreal Forced Aligner v2.0.6 [21]. Stop bursts were automatically identified by custom script written in Python v.3.10.6 [22] utilising a modified version of the algorithm described in [23]. The alignments (see Fig. 1) were manually inspected and corrected.

Acoustic analyses were conducted on V_2, C_3 and V_3 using Praat (6.1.16) [24]. We measured the duration and the formants for V_2 and V_3. Formants were measured by taking ten equal-distanced time points across the entire duration of the vowels (including onset and offset, with limits set to 5000 Hz for male and 5500 Hz for female speakers), in order to account for the dynamic evolution during the vowels. For C_3, we measured the mean Centre of Gravity (CoG) using windows of 5 ms shifted by 1 ms across the release phase (i.e., the aspiration phase of /k, g/ and the fricative phase of /tʃ, dʒ/). The closure duration and total duration of the plosives and affricates was also measured for C_3. We calculated the closure ratio by taking the ratio between closure duration and total duration of C_3.

2.2. Statistical analyses

The formant trajectories and the duration of V_2 were modelled with generalized additive mixed models (GAMM) [25] using the *mgcv* package (1.8-40) [26] in *Rstudio* [27]. The model was constructed as shown in the footnote[1]. With this model, we were able to analyze how formant trajectories vary as a function of stress, and how the time-varying formant trajectories were affected by the duration of the vowels.

To compare the differences in CoG and in closure duration across C_3, we conducted linear mixed-effects analyses, using *Rstudio* [27] and the *lme4* (1.1-30) [28] package. Mean CoG/closure ratio were taken as dependent variables, and the interaction between C_3 and stress as fixed effects. As random effects, we had by-speaker random slopes for the effect of C_3. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

3. RESULTS

3.1. Duration and formant structure of V_2

The stressed vowel is known to be hyper-articulated, entailing *inter alia* longer duration [12]. The durational difference between stressed and unstressed V_2 is evident, with the stressed V_2 being more than twice as long ($\bar{x} = 185$ ms, $\sigma = 45$) as the unstressed V_2 ($\bar{x} = 82$ ms, $\sigma = 29$).

The vocalic spaces of F1 and F2 (in Hz) of V_2 are shown in Fig. 2. The stressed vowel is more open, the unstressed vowel is less open, fronted and clearly more variable.

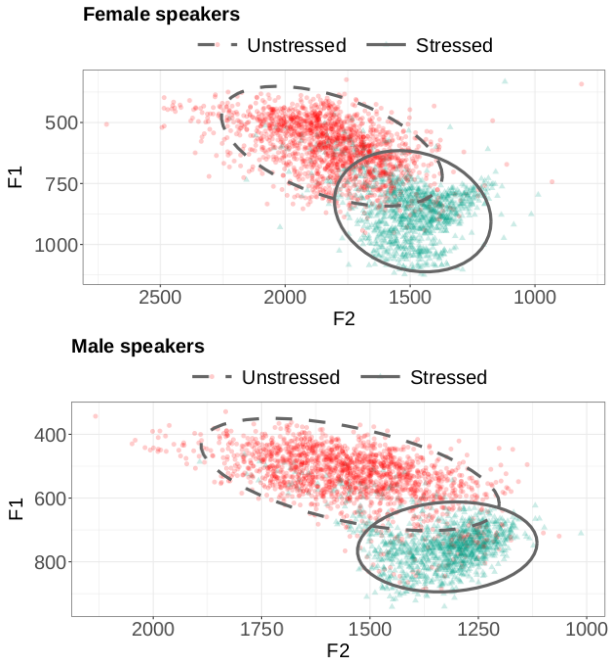


Figure 2: Formants (F1 and F2 in HZ) of /a/ according to stress conditions (red/dashed lines for unstressed; green/solid for stressed), with 95% confidence ellipses; female speakers on top, male speakers below.

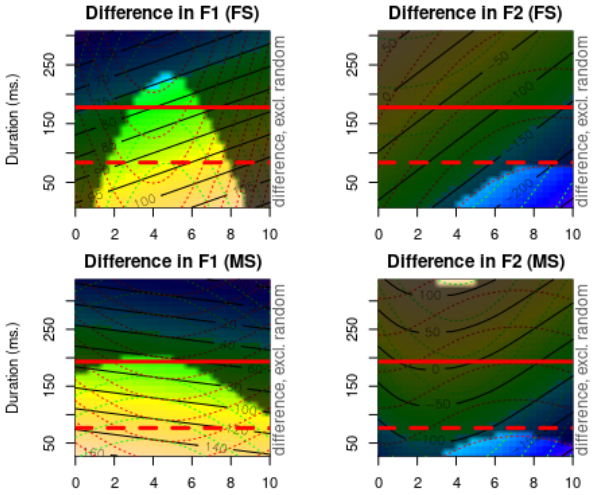


Figure 3: Difference between stressed /a/ and unstressed /a/ over time (x-axis) modulated by duration (y-axis). The shaded region indicates the area where the difference is non-significant. The red solid and dashed lines indicate the mean duration of stressed and unstressed /a/ respectively. FS and MS are for female and male speakers respectively.

The duration and formant structure of V_2 is further explored with GAMMs. More precisely, the models investigate how the formant trajectories of V_2 evolve in relation to the difference in duration. That is, does a longer, stressed V_2 differ from a shorter, unstressed V_2 in terms of formant trajectory? This comparison is shown in Fig. 3, which presents the estimated differences between stressed and unstressed V_2. The general observation would be that the F1 is overall higher in stressed /a/, except at the beginning and the end, probably due to coarticulation with surrounding segments. The models estimate that the difference is larger when the duration is shorter. This could be understood as follows: the shorter the unstressed vowel /a/, the more closed it is. The difference in F2 is less pronounced compared to F1. However, there is a consistent pattern towards the end of the vowels: the models predict that the shorter the vowel, the higher its F2.

In general, the differences between stressed and unstressed V_2 show a duration-related pattern. When V_2 is short, it is less open over its entire duration, and particularly fronted towards its end.

3.2. Closure duration and mean centre of gravity in C_3

The mean CoG of a consonant indicates its place of articulation in the oral cavity [29]. The CoG in C_3 is analysed with linear mixed effect models, grouped by their voicing categories. The results are presented in Table 1.

Centre of gravity	Estimate	S.E.	*p*
/k/ POST (Intercept)	3896.1	126.2	***
/tʃ/ POST	-88.1	120.1	0.476
/k/ FAR	-6.7	33.2	0.841
/tʃ/ FAR	-29.7	47.1	0.528
/g/ POST (Intercept)	1442.7	111.0	***
/dʒ/ POST	1350.1	137.0	***
/g/ FAR	-124.7	85.1	0.144
/dʒ/ FAR	103.2	121.4	0.396
Closure ratio			
/k/ POST (Intercept)	0.574	0.017	***
/g/ POST	0.080	0.035	*
/k/ FAR	-0.044	0.012	***
/g/ FAR	0.055	0.017	**
/tʃ/ POST (Intercept)	0.371	0.015	***
/dʒ/ POST	0.117	0.013	***
/tʃ/ FAR	-0.028	0.009	**
/dʒ/ FAR	0.030	0.013	*

Table 1: Fixed effects of linear mixed-effects analyses conducted on CoG and closure ratio of C_3, grouped by voicing and manner of articulation, respectively. *=<0.5, **<0.01, ***<0.001.

We note that CoG of plosives and affricates are overall similar, especially for the voiceless ones. The release phase of /k/, whether in POST or in FAR, has similar CoG compared to /tʃ/ in both contexts. The /g/ in POST has lower CoG compared to /dʒ/ in POST, but in FAR both segments are not significantly different from each other and from /g/ in POST. These results show that CoG is not related to the relative position of lexical stress; it is most probably related

to the coarticulation with the following vowel /i/, as shown in [17].

The results of two mixed models on closure ratio, grouped by manners of articulation, are presented in Table 1. The important general pattern is that, for all consonants, the closure phase is more important in POST than in FAR. This difference is more pronounced in the plosives compared to the affricates.

3.3. Formant structure and duration of V_3

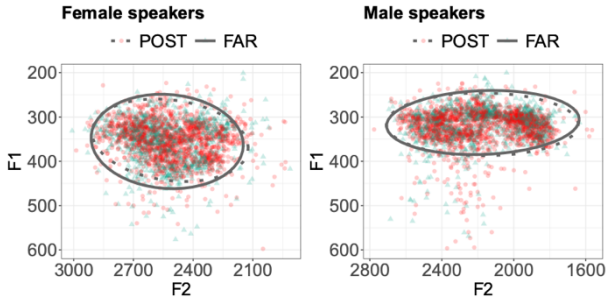


Figure 4: Formant (F1 and F2 in Hz) of /i/ (V_3) according to position (red/dotted line for POST, green/solid line for FAR), with 95% confidence ellipses.

V_3 /i/ is analysed in relation to its relative position to lexical stress. The formant structures of /i/ show no difference, as can be seen in Fig. 4. The F1/F2 spaces for /i/ in POST and FAR contexts are virtually identical. The duration of V_3 showed no difference either (POST: $\bar{x} = 73$ ms, $\sigma = 42$; FAR: $\bar{x} = 86$ ms, $\sigma = 40$). This suggests that V_3 is not directly influenced by the position of lexical stress. Whether in POST or FAR, the acoustic characteristics of /i/ are virtually the same. This further suggests that /i/ has the potential of triggering palatalization in both conditions, however palatalization is not realised in POST.

4. DISCUSSION AND CONCLUSION

This study investigated the influence of lexical prominence on the following segments in a /$C_1V_1.C_2V_2.C_3V_3$/ sequence in Italian, aiming to uncover potential mechanisms that block velar palatalization in post-stress position. By putting stress on V_1 or on V_2, the syllable C_3V_3 was either FAR from stress or POST stress. The analysis on stressed vowels showed that the μ-conditioned stressed vowels were longer in duration and more peripheral in acoustic space. The analyses of C_3V_3 in both conditions showed that V_3 (i.e., /i/ and trigger of palatalization) could not be responsible for the blocking of velar palatalization as it had similar acoustic characteristics in both FAR and POST conditions. However, C_3 had longer closure duration in POST compared to in FAR. This difference, combined with the acoustic shape of μ-conditioned V_2, may shed light on the stress-conditioning of palatalization.

We interpret the stress-conditioned velar palatalization as resulting from a perceptual effect rooted in articulation. The durational increase of the stressed vowel follows the predictions of the μ-gesture model [14]. As illustrated in Fig. 5, given that gestures are temporally coordinated and potentially overlapping with other gestures, μ-gestures also have an extent in time and overlap with vocal tract constriction gestures, namely here the C_3 closure gesture. The intergestural coarticulatory overlap between two adjacent movements, i.e., an opening gesture of V_2 and a closure gesture of C_3, are realized in different degrees according to the presence or absence of the μ-gesture (i.e., POST and FAR conditions) [30].

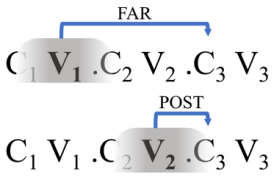


Figure 5: Illustration of μ-gesture associated with different stressed positions. Shading represents the scope of the μ-gesture, with strongest effect at the midpoint of the stressed vowel.

In POST, the μ-conditioned V_2 shows more coarticulatory resistance [31] with the following C_3. The large opening gesture of V_2 leads to a later target achievement of C_3, reflected acoustically in the longer closure duration. In FAR, the μ-gesture is not having an effect. The V_2 gesture shows less coarticulatory resistance, resulting in more gestural blending and a shorter closure in C_3.

The articulatory-conditioned C_3 closure duration discussed above is likely to have played an important perceptual role in recovering the plosive *vs.* affricate category. That is, a longer closure duration may facilitate the recovering of a plosive consonant, leading to the blocking of palatalization.

In sum, the difference in C_3 closure duration is interpreted as an articulatory by-product directly related to the μ-conditioned V_2. To confirm our interpretation, an articulatory study is planned to fully understand the gestural coordination between μ-conditioned vowels and consonants.

5. ACKNOWLEDGEMENTS

We'd like to thank Sam Mitchell for his contribution in the study's earlier stage. This study was funded by ANR-JCJC DIA-SYN-PHON funding attributed to Maria Giavazzi (ANR-21-CE28-0008). This study was partially supported by EUR Frontiers in Cognition and the Idex PSL (ANR-17-EURE-0017), and Labex EFL (ANR-10-LABX-0083).

5. REFERENCES

[1] Fourakis, M. 1991. Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical society of America*, 90(4), 1816-1827.

[2] Marcet, A., Fernández-López, M., Baciero, A., Sesé, A., & Perea, M. 2022. What are the letters e and é in a language with vowel reduction? The case of Catalan. *Applied Psycholinguistics*, 43(1), 193-210.

[3] Nobre, M. A., & Ingemann, F. 1987. Oral vowel reduction in Brazilian Portuguese. *In Honour of Ilse Lehiste*, 195-206.

[4] Giavazzi, M. 2010. *The Phonetics of Metrical Prominence and its Consequences on Segmental Phonology*. Doctoral dissertation, MIT.

[5] Hollenbach, B. 1977. Phonetic vs. phonemic correspondence in two Trique dialects. In WilliamMerrifield (ed.), Studies in Otomanguean Phonolog. Dallas: SIL and University of Texas atArlington. 3567.

[6] Suomi, K., Ylitalo, R. 2004. On Durational Correlates of Word Stress in Finnish. Journal ofPhonetic, 32:35-63.

[7] Armstrong, L. E. 1964. *The phonetic structure of Somali*. Gregg International Publishers Limited.

[8] Turk, A. E., & Sawusch, J. R. 1997. The domain of accentual lengthening in American English. *Journal of Phonetics*, *25*(1), 25-41.

[9] Gonzalèz, C. 2003. *The effect of stress and foot structure on consonantal processes*. Doctoral dissertation, USC

[10] Samareh, Y. 1977. *Arrangement of segmental phonemes in Farsi*. University of Tehran, Tehran University Publications.

[11] Bauer, W. 1993. *Maori*. London, New York: Routledge.

[12] Gordon, M., & Roettger, T. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, 3(1), 1-11.

[13] Byrd, D., & Saltzman, E. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2), 149-180.

[14] Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. 2008. A Task-dynamic Toolkit for Modeling the Effects of Prosodic Structure on Articulation. In *Proceedings of the 4th International Conference on Speech Prosody (speech prosody 2008)*, Campinas, Brazil. 175-184.

[15] Dressler, W. U. 1985. *Morphophonology: The Dynamics of Derivation*. Ann Arbor, Mich.: Karoma.

[16] Albano Leoni, F., Cutugno, F., Savy, R., 1995. The vowel system of Italian connected speech, in Elenius K., Branderud P., ed., *Proceedings of the XIIIth ICPhS*, Stockholm, 4, 396-399

[17] Guion, S. G. 1998. The role of perception in the sound change of velar palatalization. *Phonetica*, 55(1-2), 18-52.

[18] Ohala, J. J., & Jones, C. 1993. The phonetics of sound change. *Historical linguistics: Problems and perspectives*, 237-278.

[19] Wilson, C. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5), 945-982.

[20] Audacity Team. 2022. *Audacity(R): Free Audio Editor and Recorder* [Computer application]. Version 3.1.3 retrieved December 25th 2021 from https://audacityteam.org/.

[21] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498-502.

[22] Van Rossum, G., & Drake, F. L. 2009. *Python/C Api Manual-Python 3*. CreateSpace.

[23] Ananthapadmanabha, T. V., Prathosh, A. P., & Ramakrishnan, A. G. 2014. Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index. *The Journal of the Acoustical Society of America*, 135(1), 460-471.

[24] Boersma, P. & Weenink, D. 2022. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.16, retrieved 21 November 2020 from http://www.praat.org/.

[25] Wood, S. 2006. *Generalized additive models: an introduction with R*. CRC Press.

[26] Wood S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3-36.

[27] R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[28] Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.

[29] Stevens, K. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

[30] Mücke, D., & Grice, M. 2014. The effect of focus marking on supralaryngeal articulation–Is it mediated by accentuation?. *Journal of Phonetics*, 44, 47-61.

[31] Browman, C. P., & Goldstein, L. 1990. Tiers in articulatory phonology, with some implications for casual speech. Papers in laboratory phonology I: Between the grammar and physics of speech, 341-376.

[1] GAMM models

```
bam(valuef1 ~ stressPattern*duration_ms +
   s(pt_f1, by = stressPattern, k = 10) +
   s(pt_f1, speaker, by = stressPattern,
      bs = "fs", m = 1, k = 10) +
   s(pt_f1, cons, by = stressPattern,
      bs = "fs", k = 10) +
   s(pt_f1, cons_aft, by = stressPattern,
      bs = "fs", k = 10),
   data = fmt_melt_a_m,
   scat(theta = NULL, link = "identity",
       min.df = 3),
   nthreads = 30,
   discrete = T,
   AR.start = fmt_melt_a_m$start.event,
rho = fmt_autocorr)
```

Compensatory breathiness in Canary Islands Spanish – an acoustic study of vowels in /s/ weakening contexts

Karolina Broś

University of Warsaw
k.bros@uw.edu.pl

ABSTRACT

This study investigates the compensatory effects of /s/ aspiration and deletion on the preceding vowel in Canary Islands Spanish. No such effects have been reported to date for this dialect. A phonetic analysis focused on the amount of breathiness in the target vowel, as marked by cepstral peak prominence, shows that the second part of the vowel is significantly breathier than the first when followed by an aspirated coda /s/ or complete coda /s/ elision compared to vowels followed by other segments. Moreover, the shorter the following aspirated /s/, the greater the effect, thus breathiness is modulated by the duration of the following sound. This compensatory effect is possibly phonological and not purely phonetic given that breathiness is also caused by deleted segments and that vowels preceded by an (onset) aspirated /s/ are not affected. Differential effects of /s/ weakening depending on syllable affiliation suggest morphological conditioning.

Keywords: Spanish, /s/ weakening, incomplete neutralisation, phonetics-phonology interface.

1. INTRODUCTION

The weakening of /s/ is perhaps the most well-known phenomenon pertaining to Spanish phonetics and phonology. In its most common form, it consists in the production of a glottal frication [h] in coda position, also referred to as /s/ aspiration [1], although different surface realisations have been reported, including velar fricatives, [ɦ] and [ʔ], among others [2–4]. Furthermore, as an instance of lenition, the process can lead to complete elision in some contexts. Both the degree to which /s/ weakening ensues and the contexts in which it is observed differ from one variety of Spanish to another. Starting with word-internal coda position, word-final prepausal, preconsonantal and prevocalic /s/ can also be affected [5, 6]. Against this background, the dialect of interest in this paper, i.e. Canary Islands Spanish, shows quite extended /s/ weakening with and without voicing, and with optional deletion. Thus, vowels followed by /s/ can surface as [Vs], [Vh], [Vɦ] or [V] in many contexts, both inside words and across word boundaries, before consonants, pauses and vowels standing in a subsequent word. Examples are provided below.

(1) /s/ weakening in Canary Islands Spanish[i]

estás 'you are'
[eh.ˈtas] [eh.ˈtah] [eh.ˈtaɦ] [eh.ˈta]
estás guapa 'you look nice'
[eh.ˈtah.ˈgwa.pa] [eh.ˈtaɦ.ˈgwa.pa] [eh.ˈta.ˈgwa.pa]
estás aquí 'you are here'
[eh.ˈta.sa.ˈki][eh.ˈta.ha.ˈki][eh.ˈta.ɦa.ˈki][eh.ˈta:.ˈki]

Quite importantly, there is a tendency in the literature to focus on categorical distinctions between the above variants [7], although there have also been some works centred around the continuous dimension of coda /s/ weakening [8–10]. In this context, it must be noted that whenever the /s/ is weakened but not deleted completely, its strength and duration may differ from one phonetic environment or speaker to another. The question is whether such differences are systematic and whether they affect the flanking segments. Since coda /s/ is an important grammatical marker in Spanish, corresponding to the plural in nouns, adjectives and pronouns, and to the second person singular in verbs, its weakening may bring important consequences for meaning disambiguation. The complete elision of /s/ can be interpreted as contrast neutralisation. We can think of numerous minimal pairs of words and phrases that sound the same after /s/ deletion, e.g. *estas chicas* 'these girls' vs *esta chica* 'this girl'. However, there have been reports of incomplete neutralisation in some dialects of Spanish. For instance, in Eastern Andalusian and Puerto Rican Spanish vowels followed by weakened /s/ are produced as more open [11–13]. There have also been some reports of vowel duration changes as a result of /s/ weakening [14, 15], which involve both lengthening and shortening. Additionally, it has been claimed that an aspirated vowel is produced in weakening contexts, which may be interpreted as breathy voice, some effects of which have been described recently for Chilean Spanish [16–17]. Nevertheless, research on such incomplete neutralization effects is scarce and based on a variety of dialects, each of which has its own specificity. Many of the existing studies have brought inconclusive or contradictory results. As for the

dialect in question, no compensatory effects have been reported in the literature. This study is therefore the first to investigate the influence of /s/ weakening on the preceding vowels in Canary Islands Spanish.

1.1. Rationale for the present study

The immediate inspiration for the present study was an impressionistic exploration of the corpus. The inspection of spectrograms containing /s/ productions in a variety of contexts suggested that preceding vowels may be affected by the turbulence of the following fricative sound in its weakened form. This was followed by personal communication with native speakers who participated in an unpublished perception study consisting in the identification of plural vs singular word forms with variable durations of /s/ aspiration ranging from 115 ms to 0 ms [18]. Some of them claimed that there is a perceptual difference between plural and singular nouns and cutting out the aspiration following the final vowel does not cancel this effect. Thus, given our own observations and speaker intuitions, we set out to investigate the effect of /s/ weakening on the preceding vowels in Canary Islands Spanish. Since glottal fricatives have a tendency to be produced as breathy voice [19], and many of the underlying /s/ are produced as the undoubtedly breathy [ɦ] in the dialect, our expectation was that the preceding vowels are realised as (partially) breathy when followed by a glottal frication.

1.2. Goals and hypotheses

Given the above, the general prediction concerning the dialect is that there is a compensatory effect of breathiness on vowels followed by /s/. Additionally, since the duration of the /s/ is variable, the shorter the following aspirated /s/ ([h/ɦ]), the breathier the preceding vowel. By extension, this implies that full segment deletion should lead to even more breathiness on the vowel because the noise from the consonant is realised partially, if not fully, on the preceding vocoid. Furthermore, there should be a significant difference between the supposedly modal vowels not followed by /s/ and those followed by /s/, and the second half of the vowel should be the one affected given the direction of the exerted influence. The metric used to measure breathiness in this paper is cepstral peak prominence – a measurement considered to be especially robust as a marker of breathiness compared to other metrics [20, 21].[ii] The specific hypotheses are provided below.

H1. CPP(S) is lower (more breathiness) in vowels followed by /s/ compared to vowels followed by other segments (lowest when there is no acoustic cue to the oral gesture of /s/, i.e. /s/ is deleted).
If Hypothesis 1 is supported, there is a compensatory effect of breathiness that comes from the deleted or aspirated segment.
H2. CPP(S) is lower (more breathiness) in the second half of the vowel compared to the first.
If Hypothesis 2 is supported, this confirms the subphonemic and compensatory nature of the change.
H3. CPP(S) is lower (more breathiness) with the decrease in duration of the following aspirated /s/.
If there is a linear trend in the data as per Hypothesis 3, there is evidence for gradient change in the phonation of the vowel preceding the /s/.

In all cases, we do not predict differences in the breathiness effect depending on vowel quality, except that high vowels should be breathier in general (/i u/) [22]. We also look at other parameters associated with sound breathiness, such as intensity [23–25], F0 [26], and the presence of noisy or aperiodic energy in the signal [21, 27], as measured by the harmonics-to-noise ratio (HNR). All of them are expected to be lower in breathy voice.

2. METHODOLOGY

2.1. Corpus and speakers

The corpus gathered for the study consists of digital recordings of semi-structured interviews with 33 native speakers from the north of Gran Canaria (15 females, 18 males) aged 16–79. The recordings were made in February 2016 with the use of a Zoom H4N digital recorder and a Shure SM10a headworn microphone at 44,100 Hz sampling frequency. The corpus contains a total of 105 files containing 3,650 sentences and 96,076 sounds. The sound files were automatically aligned with orthographic transcriptions and pretranscribed phonetically using EasyAlign [28], followed by manual correction. We then extracted all the vowels (N=43,757) together with their acoustic measurements and contextual information. To avoid confusion with intervocalic instances of /s/ realised as [h ɦ], we only included vowels followed by coda /s/ and other consonants or vowels, which gave us a total of 28,134 observations.

2.2. Variables

The extracted variables of interest include target vowel and following sound duration in seconds, as well as smoothed cepstral peak prominence (CPPS) measured at 11 time points in the vowel, following previous literature [29]. Additional measurements include: vowel intensity and standard deviation of

intensity, mean harmonics-to-noise ratio, mean pitch, and pitch measured at 11 time points.

2.3. Statistical analysis

Statistical analyses were conducted in R [30] using the packages *lme4* [31] for building models, and *emmeans* [32] for the calculation of simple effects. Two main mixed effects models were built: one on the full database and one on a database excluding segmental deletions. Both models used CPP(S) as a dependent variable. Model 1 used vowel part (part 1, part 2) and following sound type (consonant, vowel, deleted /s/, aspirated /s/) as fixed factors. Model 2 used following sound type and following sound duration in seconds as fixed factors. Both random intercepts and random slopes were included whenever possible in terms of model convergence.

3. RESULTS

3.1 Model 1

The model checking the effects of /s/ weakening on vowels compared to the presence of other following segments was run on a total of 208,968 observations given that the initial number was multiplied by converting the data into long format to include multiple values of CPP(S) as measured at different vowel points. We performed F-testing of fixed effects, obtaining denominator degrees of freedom via Satterthwaite approximation.

The results show a main effect of following sound type ($F=2129.92$, $p<0.001$) and a significant interaction between vowel part and following sound type ($F=513.05$, $p<0.001$). We then estimated marginal means from the model in order to assess simple effects and take a closer look at the interaction term. As for the pairwise comparisons between the different levels of following sound, all of them resulted significant. Most importantly, following vowels are correlated with a higher value of CPP(S) compared to all other sounds ($t=25.17$, $p<0.001$ compared to consonants other than /s/, $t=47.78$, $p<0.001$ compared to deleted /s/ and $t=70.16$, $p<0.001$ compared to aspirated /s/). Similarly, other consonants are correlated with a higher value of CPP(S), i.e. less breathiness, compared to deleted /s/ ($t=37.5$, $p<0.001$) and aspirated /s/ ($t=63.75$, $p<0.001$), and deleted /s/ is correlated with a higher value of CPP(S) compared to aspirated /s/ ($t=20.47$, $p<0.001$). Fig. 1 illustrates these dependencies. It also shows differences in CPP(S) between the two vowel parts depending on the following sound, with other consonants and vowels raising the CPP(S) of the target vowel in the second part, and all types of /s/, deleted or not, decreasing it, and hence showing a breathiness effect. Additionally, although differences in CPP(S) in the first part of the vowel are significant between all four levels of following sound, we can see that the mean value is especially low in the case of vowels followed by aspirated /s/ (in this case the level of CPP(S) in the first part of the vowel is the same as the level for vowels followed by deleted /s/ in the second part of the vowel, $p=1$).

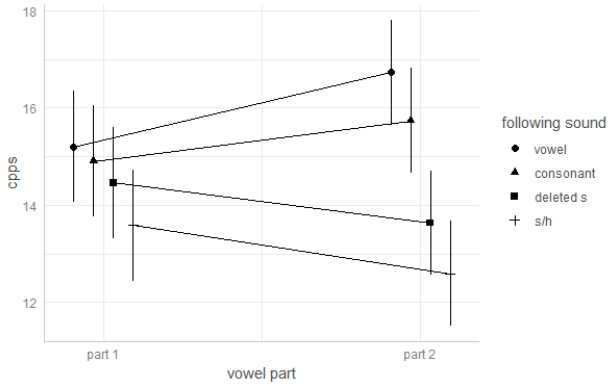


Figure 1: Effects plot for Model 1 showing predicted values of CPP(S) as per vowel part and following sound.

3.2 Model 2

Model 2 (without deleted /s/, i.e. 194,048 observations) shows significant main effects of both following sound duration ($F=6.79$, $p<0.05$) and following sound type ($F=129.94$, $p<0.001$), and a significant interaction ($F=45.97$, $p<0.001$).

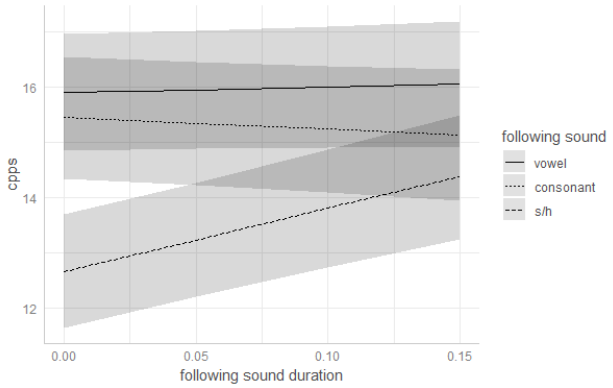


Figure 2: Effects plot for Model 2. Predicted values of CPP(S) as per following sound type and sound duration.

Pairwise comparisons show significant differences in CPP(S) between vowels standing before all types of following sounds when their mean duration is taken into account (see Fig. 2). However, we can see a positive correlation between following sound duration and the CPP(S) of the target vowel only in the case of following aspirated /s/ and no such tendencies when the vowel is followed by other segments, which confirms the compensatory breathiness effect of /s/ weakening (Fig. 2).

3.3 Additional models

The results of additional models with intensity, mean pitch and HNR as dependent variables are shown in a summary table below, where we can see that the expected pattern corresponding to greater breathiness in vowels followed by /s/ obtains in the case of intensity and HNR, while F0 shows the opposite.

model	following sound	*t* statistic
intensity	other vs deleted /s/	59.143***
	other vs aspirated /s/	40.86***
	deleted vs aspirated	-12.38***
HNR	other vs deleted /s/	6.38***
	other vs aspirated /s/	55.51***
	deleted vs aspirated	36.75***
F0	other vs deleted /s/	-23.66***
	other vs aspirated /s/	-14.97***
	deleted vs aspirated	5.95***

Table 1. Pairwise comparisons: additional models.

4. DISCUSSION

The results of the study support Hypotheses 2 and 3. In /s/ weakening contexts, the second part of the vowel is breathier than the first, and the amount of breathiness is modulated by the amount of aspiration. Hypothesis 1 was partially supported in that vowels are breathier when followed by a weakened /s/ compared to other segments but the amount of breathiness is greater before aspirated than deleted /s/, contrary to our predictions. Furthermore, most of the other acoustic parameters associated both with consonant weakening and with breathy voice are in line with the compensatory breathiness interpretation. Thus, our data show that there is a compensatory effect that differentiates vowels followed by coda /s/ from other vocalic segments in Canary Islands Spanish. Since we only looked at following segments, and vowels in the studied dialect of Spanish are equally likely to be found after weakened /s/ as before it, or to flank a weakened /s/ on both sides, it is possible that the observed effect is coarticulatory rather than regressively caused by /s/ weakening. In other words, it may be that neighbouring frication is what affects the vowel and this may take place bidirectionally. To exclude this possibility, we compared vowels followed by /s/ with vowels preceded by /s/, and used following and preceding sonorant consonants as a baseline for comparison. Fig. 3 shows the results of this enquiry. Note that the first part of the vowel is not affected in any case, which is what we would expect given the hypothetical influence of preceding aspiration, and only following aspirated /s/ results in a decrease in CPP(S).

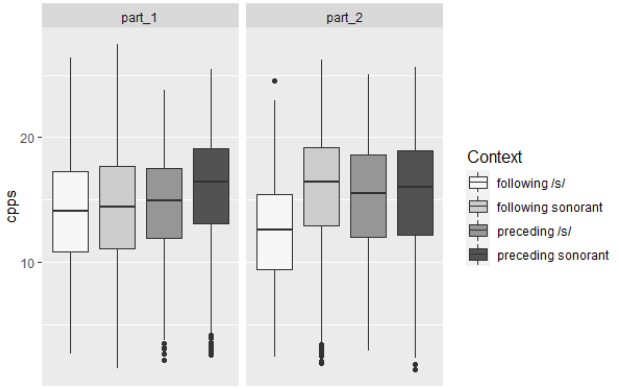


Figure 3: Mean CPP(S) depending on vowel part and /s/ or sonorant position with respect to the target vowel.

We might also expect that vowel duration and not only type of phonation is affected by /s/ weakening [cf. 15-16]. An additional mixed effects model shows that the following sound does affect the duration of the target vowel ($F=628.91$, $p<0.001$). Curiously, while /s/ deletion causes vowel lengthening, aspiration is associated with the shortest vowel duration, which suggests a compensatory effect in the former case and coarticulation in the latter.

All in all, the results of the study are interesting from the point of view of the phonetics-phonology interface. On the one hand, we see gradient effects in the form of increasing breathiness of the second half of the vowel with the decreasing duration of following aspiration. Additionally, aspiration leads to lower values of CPP(S) throughout the vowel and not only its second part, as opposed to deletion. On the other hand, /s/ deletion causes both breathiness and lengthening of the target vowel, which is clearly a compensatory change. Both types of weakening are also correlated with lower vowel intensity and HNR, which confirms that such vowels differ from those followed by other sounds, be it consonants or vowels. Furthermore, the effect is only observable when coda /s/ is involved. All of those factors put together suggest that phonology is involved. Vowel breathiness in /s/ contexts has a functional effect: it helps distinguish morphological contrasts, assists in perception when disambiguating meanings and possibly plays a role in learning. It is therefore possible that we are dealing with an ongoing phonologization of contrast, i.e. a transition from gradient breathiness with /s/ aspiration to a categorical change from [Vs] to a breathy vowel [V^{h}]. Future studies should focus on morphological and prosodic effects, such as disambiguating between coda /s/, onset /s/, onset /x/ realised as [h] and coda /s/ resyllabified as onset.

7. ACKNOWLEDGEMENTS

The study was supported from the funds of project no. UMO-2017/26/D/HS2/00574 granted to the author by the National Science Centre, Poland.

8. REFERENCES

[1] Terrell, T. 1981. Diachronic reconstruction by dialect comparison of variable constraints: *s*-aspiration and deletion in Spanish. *Variation omnibus,* 115–124.

[2] Marrero, V. 1990. Estudio acústico de la aspiración en español. *Revista de Filología Española* 70, 345–397.

[3] Widdison, K.A. 1995. On the value of an experimental paradigm in linguistics and its application to issues in Spanish phonology. *Neophilologus* 79(4), 587–598. doi: 10.1007/BF01126890.

[4] Lipski, J. 2011. Socio-phonological variation in Latin American Spanish. In: Díaz-Campos, M. (ed.), *Handbook of Hispanic Sociolinguistics*. Blackwell. https://doi.org/10.1002/9781444393446.ch4.

[5] Lipski, J. 1994. *Latin American Spanish.* Longman.

[6] Hualde, J.I. 2005. *The sounds of Spanish*. Cambridge University Press.

[7] Erker, D. 2012. Of Categories and Continua: Relating Discrete and Gradient Properties of Sociophonetic Variation. *University of Pennsylvania Working Papers in Linguistics* 18(2), Article 3.

[8] Erker, D. 2010. A subsegmental approach to coda /s/ weakening in Dominican Spanish. *International Journal of the Sociology of Language* 203, 9–26.

[9] Campos-Astorkiza, R. 2014. Sibilant voicing assimilation in Peninsular Spanish as gestural blending. In: Côte, M.-H., Mathieu, E., Poplack, S. (eds), *Variation within and Across Romance Languages.* John Benjamins, 17–38.

[10] Gerfen, C. 2002. Andalusian codas. *Probus* 14, 247–277.

[11] Alonso, D., Zamora Vicente, A., Canellada de Zamora, M.J. 1950. Vocales andaluzas: Contribución al estudio de la fonología peninsular. *Nueva revista de filología hispánica* 4(3), 209–230.

[12] Rincón Pérez, M.A. 2015. *An acoustic and perceptual analysis of vowels preceding final /-s/ deletion in the speech of Granada, Spain and Cartagena, Colombia*. Unpublished doctoral dissertation, Purdue University.

[13] Navarro Tomás, T. 1948. *El español en Puerto Rico: Contribución a la geografía lingüística hispanoamericana*. Universidad de Puerto Rico.

[14] Oroz, R. 1966. *La lengua castellana en Chile*. Facultad de Filosofía y Educación: Universidad de Chile.

[15] Cepeda, G. 1990. La alofonía de /s/ en Valdivia (Chile). *Estudios Filológicos* 25, 5–16.

[16] Lenz, R. 1940. Estudios chilenos (Fonética del castellano de Chile). In: Alonso, A. (ed.), *El español en Chile*. Universidad de Buenos Aires, 87–258.

[17] Bolyanatz, M.A. 2020. Evidence for Incomplete Neutralization in Chilean Spanish. *Phonetica* 77(2), 107–130. https://doi.org/10.1159/000493393.

[18] Wołłejko, J., Broś, K. 2021. Perception of word-final aspirated /s/ – a cross-dialectal and cross-linguistic study. *Phonetics and Phonology in Europe Conference,* Barcelona.

[19] Garellek, M., Chai, Y., Huang, Y., Van Doren, M. 2021. Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association,* 1–28. doi:10.1017/S0025100321000116.

[20] Khan, S.D. 2012. The phonetics of contrastive phonation in Gujarati. *Journal of Phonetics* 40(6), 780–795.

[21] Samlan, R.A., Story, B.H., Bunton, K. 2013. Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling. *Journal of Speech, Language, and Hearing Research* 56, 1209–1223.

[22] Lotto, A.J., Holt, L.L., Kluender, K.R. 1997. Effect of voice quality on perceived height of English vowels. *Phonetica* 54(2), 76–93.

[23] Fischer-Jorgensen, E. 1968. Phonetic Analysis of Breathy (Murmured) Vowels in Gujarati. *Annual Report of the Institute of Phonetics University of Copenhagen*.

[24] Thongkum, T. 1988. Phonation types in Mon-Khmer languages. In: Fujimura, O. (ed.), *Vocal fold physiology: voice production, mechanisms and functions*. Raven Press, 319–334.

[25] Traill, A., Jackson, M. 1988. Speaker variation and phonation type in Tsonga nasals. *Journal of Phonetics* 16, 385–400.

[26] Hombert, J., Ohala, J.J., Ewan, W.G. 1979. Phonetic Explanations for the Development of Tones. *Language 55*, 37.

[27] Gordon, M. 2002. *Linguistic aspects of voice quality with special reference to Athabaskan*.

[28] Goldman, J-P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. *Proceedings of Interspeech 2011*, 3233–3236.

[29] Esposito, C., Khan, S. 2012. Contrastive breathiness across consonants and vowels: A comparative study of Gujarati and White Hmong. *Journal of the International Phonetic Association* 42(2), 123–143.

[30] R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org.

[31] Bates, D., Maechler, M., Bolker, B., Walker, S. 2018. *Lme4: linear mixed-effects models using 'Eigen' and S4. R package (version 1.1-17).* cran.r-project.org/web/packages/lme4.

[32] Lenth, R. 2019. *Emmeans: Estimated Marginal Means, aka Least-Squares Means.* https://CRAN.R-project.org/package=emmeans.

[i] All major pronunciations occurring in the dialect are listed, although some are more frequent than others. We ignore other than glottal productions reported e.g. by [3]. Variation in stop weakening typical in this dialect is irrelevant for the present study and is therefore not included.

[ii] Another popular measure, not explored here, i.e. H1 – H2, has proven weak in differentiating the degree of breathiness and can be influenced by nasality [see 21]. It must be noted, however, that at least some speakers use it more reliably than CPP in perception (e.g. in Gujarati, [20]).

CONSONANT-VOWEL COARTICULATION PATTERNS IN SWEDISH AND MANDARIN

Man Gao[1], Malin Svensson Lundmark[2]

[1]Dalarna University, [2]Lund University, [2]University of Southern Denmark
[1]mao@du.se, [2]malin.svensson_lundmark@ling.lu.se

ABSTRACT

This paper reports a cross linguistic study that compares the coarticulation patterns between consonant and vowel (CV) in Mandarin Chinese and Southern Swedish. Kinematic data were collected using the Electromagnetic Articulography (EMA) for both languages and were subjected to three types of CV time lag measurement, based on more or less equivalent landmarks on lips and tongue, and partially adopted in previous studies [1, 2, 3]. We found rather consistent CV coordination patterns in these two typologically different languages with both the velocity-based and the acceleration-based measurements on the lips and the tongue body. The most striking result to emerge from the data is the same effect of gender on the variation of CV coarticulation in both languages, which has not been reported previously. In addition, only when gender was added as a factor, did we find the language differences on the CV time lags.

Keywords: CV coarticulation, EMA, Mandarin Chinese, Southern Swedish, gender difference

1. INTRODUCTION

In the past half century, new methods and equipment have been developed to obtain a greater understanding of the temporal and spatial coordination patterns of various speech articulators. For example, some studies [1, 2, 4] have used the Electromagnetic Midsagittal Articulometer to collect 2D kinematic data for analyzing the interarticulator coordination in sequences of segment or in words that carry lexical tones. Other studies [3, 5] have used the ElectroMagnetic Articulography (EMA) to collect 3D kinematic data for analyzing the consonant and vowel coordination in pitch accents. It has been observed in these studies [2, 3, 5] that the word initial consonant and following vowel (CV) exhibit a strong articulatory overlap to a considerable extent.

Löfqvist and Gracco [1] recruited American English speakers to investigate the issue of coarticulation in the sequence of V1CV2, in which C is a labial stop. They found that the tongue articulator moves for the second vowel synchronously or even before the start of the lips close for the consonant. In a study which set out to examine the segment-to-segment and tone-to-segment temporal coordination patterns in Mandarin words, Gao [2] reported that the word-initial consonant and its following vowel are aligned in a constant manner, about 45 ms apart with the consonant articulator activated first. Most recently, detailed examination of the word initial C-V coordination in Swedish words by Svensson Lundmark et al. [3] reported both synchronous coarticulation as in [1] and coarticulation with a temporal lag as in [2], depending on the methodological choices such as landmarks measured and the participants.

While the articulatory overlap patterns may be a language-specific feature, the varied patterns observed in previous studies could be a result of lack of standardized measures. In the aforementioned studies, different landmarks were selected for computing the coarticulation. For example, Gao [2] measured the onset of consonant and vowel on the basis of velocity curve of the corresponding articulators, but the measurement in Löfqvist and Gracco [1] was based on the tangential velocity of the tongue body movement for vowel and the peak acceleration of the lips for consonant. Recent work on peak acceleration of the lip articulator shows that it is systematically timed with the acoustic segments across speakers [6]. While the nature of peak acceleration of the tongue body onset movement is yet to be investigated, it is not entirely clear whether the consonantal and the vocalic gestural onsets should be measured using equivalent landmarks on velocity or acceleration, or because of the different linguistic functions of the articulators, use non-equivalent landmarks.

This paper reports an initial attempt to investigate the consonant-vowel coarticulation patterns in Mandarin Chinese and Southern Swedish. Specifically, two issues will be addressed: first is to better understand the interarticulator coordination pattern in two typologically different languages. And the other issue is methodological: to evaluate three types of measurements with more or less equivalent landmarks, some of which have been used previously for computing the CV coarticulation on both sets of language data.

2. METHOD

The speech material is EMA data from two Mandarin (one female) and four Southern Swedish speakers (two female). It contains 474 syllables on the CV sequence /ma/: 46 tokens from Mandarin, and 428 from Swedish (which is part of a larger corpus of EMA data on 21 Swedish speakers, see e.g. [3]).

In Swedish, the target /ma/ receives the primary stress and is the first syllable in a disyllabic word. It's an open syllable which contains a bilabial consonant [m] and a vocalic nucleus [ɑː], followed by either a [l] or an [n] in the next syllable. In Accent 1 (a tonal fall) these are in turn followed by the suffix +en, while in Accent 2 (a tonal rise) by +ar. The target words are placed in statements preceded by leading questions, to ensure a non-focused elicitation.

In Mandarin the target syllable /ma/ can carry four lexical tones, which correspond to four different words. The target word is embedded in the statements with conflicting tonal environment, which are the same as the speech material used in [2]. The test syllable /ma/ of the two languages, occurs in disyllabic and monosyllabic words respectively, due to the different tone carrying units in Mandarin and Swedish. As they are both embedded in sentence-medial positions, and receive primary stress, this difference should not affect the comparison of CV coarticulation.

2.1. Procedure

All speakers were recorded at 250 Hz with EMA; a Carstens AG501 at the Lund University Humanities Laboratory. Audio was recorded simultaneously using an external condenser microphone (a t.bone EM 9600) at a sampling rate of 48 kHz. The Swedish speakers read leading and target sentences from a prompter in a random order, each set appearing eight times. The Mandarin speakers read a set of target sentences six times in random order.

The authors separately segmented the acoustic data of the two languages in Praat [7]. The textgrid files were later used in R [8] as reference time windows for collection of the articulatory data. An inter-annotation agreement (IAA) between the two authors was performed on 60 tokens randomly collected from the large Swedish corpus, showing a good agreement for segment boundaries between the two authors with a mean deviation of 2.4 ms, and 93.4% of tokens (213 of 228) within 10 ms [9].

2.1.1. Articulatory data

Articulatory data were collected from six sensors: two placed on the upper and the lower lip at the vermilion border, one on the lower incisor, and three sensors placed on the midline of the tongue. The first tongue sensor was placed on the tongue body where the participant made a bite mark after having stretched out his or her tongue as far as possible. The second tongue sensor was placed between the sensor at the back and a third tongue sensor, which was placed approximately 1 cm from the tongue tip. To correct for head movements three additional sensors were used: one behind each ear and one on the nose ridge. Post-processing was done in the Carstens software. Only the sensor positions of the lips and the sensor furthest back on the tongue body (TB) were further analyzed in this study.

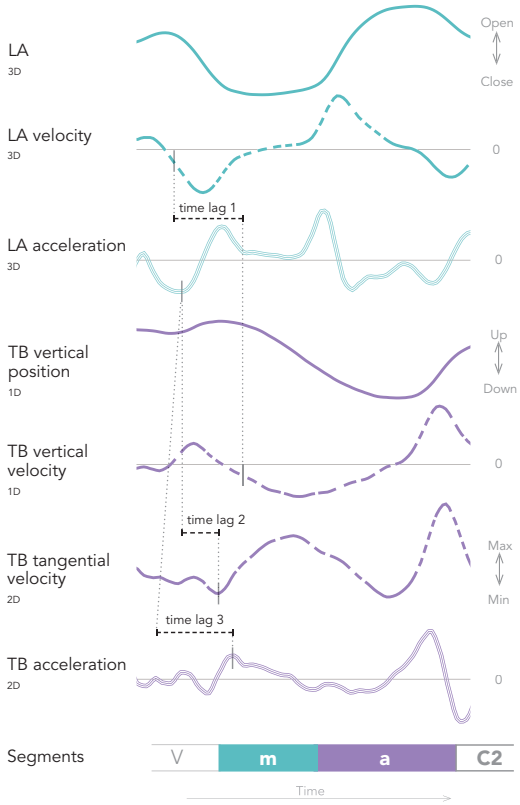


Figure 1: Lip aperture (LA) and tongue body (TB) landmarks used to calculate the CV time lags (Swedish word example). Positive time lags signify that TB onset follows LA onset, while negative time lags mean TB precedes LA.

2.1.2. Articulatory measurements

The data were then processed in R [8]. Lip aperture (LA) was calculated using the three-dimensional (3D) Euclidian distance between the sensors on the upper and the lower lips. The articulatory data was smoothed using locally weighted regression by the R function *loess* (low span 0.1). Articulatory landmarks were automatically collected using the textgrid files as reference time windows. CV time lags was calculated using the following velocity and acceleration landmarks at the onsets of the bilabial closing and the tongue movement (Fig. 1):

1) 20% threshold from 0-crossing to peak velocity of LA (3D), and of TB lowering (one-dimensional, 1D) (based on [2]);
2) peak acceleration of LA (3D) and the minimum tangential velocity of TB (two-dimensional, 2D) (based on [1]);
3) peak acceleration of LA (3D) and of TB (2D).

2.2. Statistical analysis

All statistical tests were run in R [8]. Generalized linear mixed models (GLMM) were run with speaker and tone/accent as random effects (random intercept). Language, and subsequently also gender, were set as fixed effects. Likelihood ratio tests were performed to evaluate added complexity (following [10]). The models were run using the lme4-package [11] and the lmer Test-package [12]. In addition, one-way ANOVA tests were also used to determine whether there were any statistically significant differences between the tones/accents within each language.

3. RESULTS

3.1. CV time lag measures

Results from the CV time lag 1 measure are similar to previously reported results [2, 3]: we find time lags of about 60 ms for Swedish speakers, and about 42 ms for Mandarin speakers (Table 1). Even though Figure 2a suggests otherwise, there is no statistically significant difference between the languages (t =-1.14, p =.297). The one-way ANOVA tests reveal no statistically significant differences among the four Mandarin tones or between the two Swedish word accents.

CV time lag 2, the time lag of the combined acceleration/velocity measurement, displays more or less synchronous timing for Swedish, and negative time lags of about -28 ms for Mandarin (Fig. 2b). The language difference is not statistically significant (t =-1.8, p =.112) (Table 1). Figure 2b indicates less varied time lags for some of the tones of the Mandarin speakers. The one-way ANOVA reveals no statistically significant difference between the four Mandarin tones. However, a significant difference is found between Swedish A1 and A2 (F (1,125) = 8.255, p <.01), which is similar to previous reported results [3].

Similarly, the third time lag measure (CV time lag 3, based on acceleration peaks) does not display any difference between the two languages (t =-0.8, p =.452) (Table 1). Figure 2c suggests less varied time lags for the Mandarin speakers only. The one-way ANOVA indicates statistically significant difference among the Mandarin tones (F (3,42) = 4.243, p <.05), and the post hoc Tukey test reveals a statistically significant difference between T1 and T3 (p <.01). Between the Swedish word accents the difference is marginally significant (F (1,116) = 3.352, p =.07).

		Estimate	SE	df	t-value	p-value
CV time lag 1	(Intercept) Sw	60.58	9.66	5.95	6.27	.001
	Language Ma	-19.19	16.80	6.06	-1.14	.297
CV time lag 2	(Intercept) Sw	2.48	9.53	7.56	0.26	.801
	Language Ma	-27.65	15.38	7.49	-1.80	.112
CV time lag 3	(Intercept) Sw	34.86	7.05	4.87	4.95	.005
	Language Ma	-9.17	11.44	6.34	-0.80	.452

Table 1: GLMM models on the CV time lags, with language as fixed effect.

3.2. Gender differences

Because of the large time lag variation observed for the Swedish speakers, ad-hoc one-way ANOVA tests were performed on gender variations within each language. Figure 3 displays the results on the combined gender and language approach: taking LA onset as the reference point, TB onset for male speakers precedes TB onset for female speakers. The ANOVA tests show that the temporal lags between the male and female speakers differ significantly at the level of p <.01 for all three measures. However, as the ANOVAs do not account for speaker or tone variability, ad-hoc GLMM models adding gender are also performed: complexity was only warranted for CV time lag 1, showing significant differences on gender and language (Table 2). Thus, language differences are found, but only for the velocity-based measure, and only when gender is added as a factor.

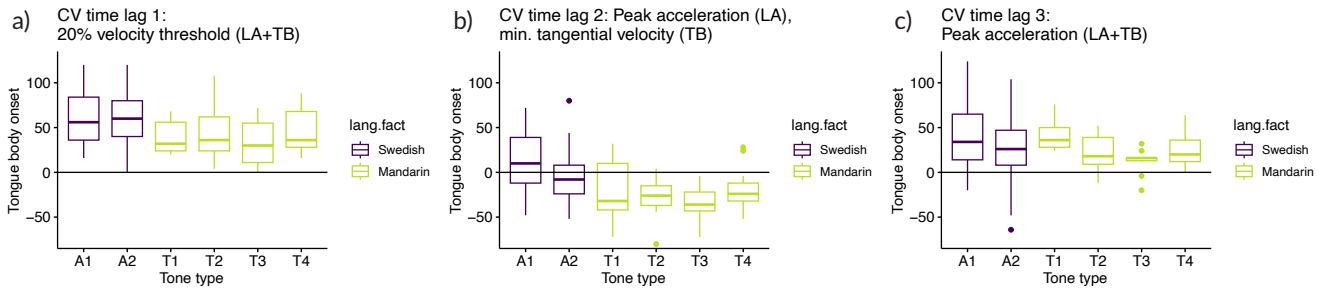


Figure 2: CV time lags (in ms): a) time lag 1, based on velocity; b) time lag 2, based on acceleration + velocity; c) time lag 3, based on acceleration. 0 marks LA onset, boxes marks TB onset. Swedish word accents: A1 and A2; Mandarin lexical tones: T1, T2, T3 and T4.

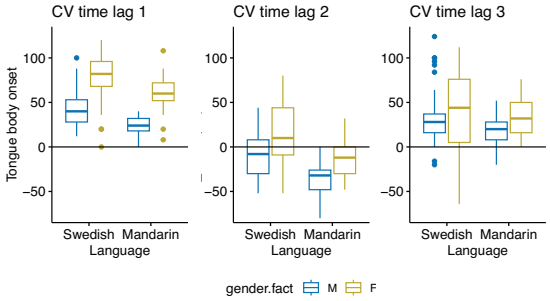


Figure 3: CV time lags 1-3 (in ms) divided according to gender and language. 0 marks LA onset, boxes marks TB onset.

		Estimate	SE	df	*t*	*p*
	(Intercept) Sw M	40.19	5.00	9.30	8.04	.000
CV time	Gender F	40.43	3.06	6.06	13.24	.000
lag 1	Language Ma	-17.50	7.64	12.60	-2.29	.039
	Gend F:Lang Ma	-3.63	5.94	167.08	-0.61	.541
CV time	(Intercept) M	-19.78	9.14	6.20	-2.16	.072
lag 2	Gender F	24.71	12.15	4.67	2.03	.102
CV time	(Intercept) M	26.57	7.16	6.39	3.71	.009
lag 3	Gender F	9.58	9.37	5.69	1.02	.348

Table 2: GLMM models on language and gender as fixed effects (complexity added only for CV time lag 1).

4. DISCUSSION AND CONCLUSION

The velocity-based (CV time lag 1) and the combined acceleration/velocity-based measurements (CV time lag 2), which were adopted repeatedly in previous studies [1, 2, 3, 13], yield similar results that are in line with previous work [2, 3], and further suggest no significant difference between the two languages. Since pitch variations are used in both languages to contrast lexical meanings, the possible effect of pitch on the CV coarticulation has also been examined. Our results have been, once again remarkably close to what were reported in previous studies, no effect of lexical tones is observed for Mandarin Chinese (time lag 1 and 2), and only time lag 2 reports an effect of the word accent for Swedish. CV time lag 2 includes the TB front-back dimensions, possibly yielding the observed difference between word accents, as previously reported in [3]. The use of CV time lag 3 was motivated by the missing standard measure of articulatory data. Although the acceleration-based measures did not show any significant difference between Swedish and Mandarin, only CV time lag 3 displayed differences between both the word accents and the lexical tones.

One surprising factor that was significantly associated with the CV timing is gender. For both languages, and all measurement types, speakers' gender was proven to affect the temporal lag in the same direction. However, when the speaker variation was taken into account, the gender differences were not as obvious. Only on the velocity-based measure (CV time lag 1), we found different patterns of co-articulation between Mandarin and Swedish. This suggests that CV coordination between the two languages only differ when comparing male and female speakers separately. The rather unexpected findings on gender could be due to the anatomical and physiological features of the speakers. There are obvious biological differences between sex, where generally speaking, male speakers have bigger head size than female speakers, thus they should also have bigger vocal tracts. Because the larger vocal tract is closely correlated with the velocities of tongue in a positive manner [14], subsequently, the shorter CV temporal lags observed with male speakers may be associated with the greater velocity of tongue body (responsible for vowel). Moreover, compensatory articulations as a consequence of anatomically differences [15], linked to biological differences between sex, or socially constructed roles of gender, could also play a part in gender-specific articulation (see e.g. [16]). Clearly this needs more investigation to establish the role that speakers' gender is playing when comparing CV coarticulation patterns.

CV time lag 1 is based on equivalent landmarks on velocity of LA and TB, which yield not only the largest gender difference, but also the longest time lags of the three measures. We presume the longer time lags are due to the movement characteristics of the lips and tongue (e.g., the lips being smaller and faster), notably, a difference not as evident for the acceleration landmarks in CV time lag 3. Peak acceleration (equated with adding force to the movement) of both lips and tongue (time lag 3) does not only display shorter time lags, but also, as already mentioned, differences between tones/accents in both languages. Although non-equivalent landmarks (time lag 2) can be motivated by different linguistic functions, and movement characteristics, of the lips and the tongue, both measures on equivalent landmarks seem to reveal significant differences between groups: the velocity-measured (time lag 1) differs between gender and language; the acceleration-based (time lag 3) between tones and word accents. However, such a hypothesis, linking systematic tendencies in CV timing to movement characteristics, needs to be confirmed with bigger sample size, specifically of the Mandarin data.

In summary, our results provide first evidence that the word-initial consonant and vowel may be coarticulated in a similar manner in typologically different languages. For the first time, the present data also show that speakers' gender is most likely to affect the CV coarticulation. However, this study has some limitations: besides the small sample size and un-balanced data sets, it does not consider the potential influence of the prosodic environment, which is due to the limited understanding of this variable in studies of articulatory coordination.

5. ACKNOWLEDGMENTS

This work was supported by an International Postdoc grant from the Swedish Research Council (Grant No. 2021-00334), and by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, 2018–2024; Grant No. 2017-00626). The authors gratefully acknowledge the Lund University Humanities Lab.

6. REFERENCES

[1] Löfqvist, A., Gracco, V. L. 1999. Interarticulator programming in VCV sequences: Lip and tongue movements. *J. Acoust. Soc. Am.* 105(3), 1864–1876.

[2] Gao, Man. 2008. *Tonal alignment in Mandarin Chinese: An articulatory phonology account*. New Haven: Yale University Doctoral dissertation.

[3] Svensson Lundmark, M., Frid, J., Ambrazaitis, G., Schötz, S., 2021. Word-initial consonant–vowel coordination in a lexical pitch-accent language. *Phonetica* 78(5-6), 515–569.

[4] Mücke, D., Nam, H., Hermes, A., Goldstein, L. M. 2012. Coupling of tone and constriction gestures in pitch accents. In Hoole, P., Bombien, L., Pouplier, M., Mooshammer, C., Kühnert B. (eds.), *Consonant clusters and structural complexity.* Mouton de Gruyter, 205–230.

[5] Niemann, H., Grice, M., Mücke, D. 2014. Segmental and positional effects in tonal alignment: An articulatory approach. *Proc. 10th ISSP* Cologne, 285–288.

[6] Svensson Lundmark, M. 2023. Rapid movements at segment boundaries. *J. Acoust. Soc. Am.* 153 (3).

[7] Boersma, P., Weenink, D. 2018. Praat: Doing phonetics by computer [Computer software]. Version 6.0.37. Retrieved 3 February 2018 from http://www.praat.org/.

[8] R Core Team. 2015. R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

[9] Machač, P., Skarnitzl, R. 2009. *Principles of phonetic segmentation*. Epocha.

[10] Wieling, M., Tiede, M. 2017. Quantitative identification of dialect-specific articulatory settings. *J. Acoust. Soc. Am.* 142(1), 389–394.

[11] Bates, D., Maechler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1). 1–48.

[12] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82(13), 1–26.

[13] Zhang, M., Geissler, C., Shaw, J. 2019. Gestural representations of tone in Mandarin: Evidence from timing alternations. *Proc. 15th ICPhS* Melbourne, 1803–1807.

[14] Kuehn, D. P., Moll, K. L. 1976. A cineradiographic study of VC and CV articulatory velocities. *J. Phon.* 4(4), 303–320.

[15] Engwall, O., Delvaux, V., Metens, T. 2006. Interspeaker variation in the articulation of nasal vowels. *Proc. 7th ISSP* Ubatuba, 3–10.

[16] Foulkes, P., Docherty, G. 2006. The social life of phonetics and phonology. *J. Phon.* 34(4), 409–438.

A SIMULATED ANALYSIS OF REPEATED MEASURES IN VOT: WE NEED MORE TOKENS!

Kyle Parrish

Rutgers University
kyle.parrish@rutgers.edu

ABSTRACT

In speech research, studies often make use of repeated measures, in which multiple productions of a given sound are taken from the same participant and incorporate them into multilevel models. For instance, in studies of voice onset time (VOT), researchers often elicit productions of stop consonants from multiple words produced by the same participant. It is unclear, however, how many repetitions of the same segment are necessary and how researchers go about choosing this number. The present study used reported data from previous literature to simulate an underlying distribution of 1000 points for each stop consonant in American English (/p/, /t/, /k/, /b/, /d/, /g/) in order to determine how many of these tokens were necessary so that the random sample would be practically equivalent sample to the full simulated distribution. The results suggest that at least 60 tokens are necessary for all 6 stop consonants to achieve a practically equivalent sample (equivalence bounds: Cohen's d = +/- .4). All materials used to create this paper, the application and the R code used to run the simulation can be found at https://osf.io/k8534/.

Keywords: Repeated Measures, VOT, simulation.

1. INTRODUCTION

Linguistic and psychological research of human subjects is often concerned with characterizing the behavior of a participant in distinct experimental conditions. It is generally agreed that, in most cases, it is necessary for experimenters to collect a sample of participants from a hypothetical underlying distribution, since the resources involved make it impossible to collect data from every possible subject meeting some given group criteria. For example, if one wanted to determine how the segment /a/ is produced in American English, it would likely not be feasible to collect tokens of /a/ from every American English speaker and only a sample of these speakers could be collected due to resource constraints. In general, the number of participants necessary for a faithful sample are determined by power analysis, where a researcher specifies an effect size, power level, significant threshold and finds the minimum number of participants needed to reliably find that effect.

In addition to knowing the appropriate number of participants for a study, the number of repetitions of a particular sound by each of those participants is another factor which has received much less attention. Following the same logic as sampling from a distribution of participants, it is also the case that researchers cannot collect all productions of a given sound from a subject, and must collect a sample. While most studies in speech research utilize repeated measures, it is unclear how the number of productions of a given segment are justified.

The quantity of repeated measures has varied in the literature. For instance, in their tutorial for multilevel models, Baayen and colleagues [2] simulated three tokens per condition in their multilevel model. Although their data set was simulated to demonstrate the nested structure of data and that data points from the same individual are not independent, the decision of three repetitions per token in this paper appears to be arbitrary and was not specifically justified.

In speech research, the quantity of repeated measures has also varied. Given their abundance, VOT studies provide a good example of the variation in repeated measures in speech research, and are the focus of the present paper. To investigate the use of repeated measures in recent VOT studies, a brief analysis of relevant studies in VOT was carried out focusing on the Journal of Phonetics and Google Scholar. Using both websites, a search for the term "VOT production" sorted by relevance revealed that the six most relevant articles range from 3 to 50 tokens per segment and condition [6], [1], [18], [11], [9], [4]. To be clear, these studies ranged in how they divided conditions. In some cases, the all stop consonants were presented in a single vocalic context [18], while others elicited stops with many vowel sounds [4]. Here, the focus of repeated measures is the quan-

tity of tokens per segment per language (in a bilingual context), thus the same stop in the same language, but in a different vocalic context were counted as belonging to the same condition. For example, Chodroff and Wilson [4] analyzed all 6 stops in American English and included 5 repetitions of each stop in 10 vocalic contexts, for a total of 50 productions of each stop per participant. It is important to note, however, that there is evidence that repetitions of the same word by the same participant cease to be independent of one another [22]. As a result, the underlying distribution may not be a normal distribution, which is a key assumption of the current work. Ideally, repetition of the same word should be avoided.

Table 1 shows a breakdown of the top studies returned by the search and their tokens per condition.

Table 1: Most relevant studies by filter in the Journal of Phonetics

Study	Repetitions
Olson (2013)	3
Antoniou (2011)	4
Hussain 2018	5
Fish (2017)	9 – 12
Gorba & Cebrian (2022)	10
Chodroff & Wilson (2017)	45 – 50

In addition to the studies from the Journal of Phonetics, the top 10 studies from the Google Scholar search for "VOT production" were also analyzed [23], [13], [8], [10], [7],[12], [21], [17], [15], [20]. The mean number of productions of a given sounds in this subset was 23.9 (sd = 28.6). This high standard deviation was due to the Nielsen study [17], which elicited 100 productions of a given sound. Omitting this study, the mean tokens were 15.4 (sd = 10.8; range 3-35).

The source of the variation in the number of repeated measures in the literature is unclear, but could be made more consistent by carrying out a power analysis, similarly to how sample size is justified. The power of a sample refers to the probability of detecting an effect when it exists [5]. A power analysis consists of four parts, an significance threshold (typically .05), a power level (typically .8), a sample size and an effect size. Given three of these four, a power analysis calculates the missing number. For example, with the significance threshold at .05 and the power level of .8, a power analysis reveals that about 80 participants are needed per group to reliably detect a small effect (Cohen's D = .4). Although speech research is often concerned with finding differences in

Table 2: Most relevant studies of "VOT production" on Google Scholar

Study	Repetitions
Zlatin & Koenigsknecht (1976)	35
Kupske (2017)	18
Gandour & Dardarananda (1984)	30
Gabriel et al. (2018)	3
Harada (2003)	9
Khattab (2002)	17
Ryalls et al. (1997)	9
Nielsen (2011)	100
Llama et al. (2016)	6
Riney et al. (2007)	12

groups or conditions, it is also possible and useful to conduct a power analysis to determine the probability of detecting practical equivalence within a particular effect size. In this case, a threshold for practical equivalence can be specified, along with a number of samples and significance threshold. As a result, the power analyses of the present study refer to the probability of detecting practical equivalence, rather than a particular effect size, when it exists.

The present study implements this approach using the reported means and standard deviations in Chodroff and Wilson (2017) [4]. The present study carries out several power analyses on large simulated data sets of a 6 stop consonants in English to determine how many tokens are necessary to consistently generate a sample that reliably falls within a small effect size of the simulated underlying distribution. That is, if we consider the speaker to be producing VOT from an underlying distribution of possible values, we can analyze how many samples from that distribution are needed to achieve a particular level of precision when we characterize an individual's stop production. In particular, the present study is guided by the following research question:

RQ1: For each of the 6 stop consonants in English, how many tokens are necessary to produce a sample that is practically equivalent (falls within Cohen's d = +/- .4 on the test of equivalence when p > .05) to the underlying sampling distribution (a simulated data set of 1000 points).

It was predicted that, just like power analyses of needed participants, that more samples would be associated with higher power.

2. METHODS

All data used for the present study were simulated using the rnorm function in R. First, for each stop

(/p/,/t/,/k/,/b/,/d/,/g/), an underlying distribution of 1000 points each was generated. The means and standard deviations of each distribution came from the from the literature (see Table 3) [4] for all 6 top consonants. This underlying distribution was generated to serve as a the representation of all possible realizations of a given consonant which were each equally probable to be produced by a given individual, although the true mean and standard deviation of this distribution will never be known.

Table 3: Means and Standard Deviations for each Stop Consonant in American English from Chodroff & Wilson (2017)

Segment	Mean (SD)
/p/	89(27)
/t/	98(28)
/k/	99(24)
/b/	13(5)
/d/	21(7)
/g/	28(10)

Then, using the generated underlying distribution a loop randomly selected a given number of tokens from the total of 1000. The sample sizes analyzed were 10, 20, 30, 40, 50, 60, 70, and 80 which were looped 100 times each. For example, if the sample size was 10, then the function chose 10 random rows of the total of 1000 per iteration. Following each selection of the random tokens, a test of equivalence [14], and a Welch's t-test were carried out between the random sample (e.g., the 10 random tokens of that iteration) and underlying distribution (all 1000 simulated data points). This process, which generated 48000 total data points (100 simulations x 8 sample sizes x 6 segments) was repeated 10 times. Then, the percentage of equivalent samples per segment per sample size per repetition was calculated. Essentially, this process generated 10 power analyses in each possible case of segment and sample size.

The test of equivalence used equivalence bounds of +/- d = .4, or a less than a small effect in language research based on a recent study [19]. In other words, the present study considered a random sample from a participant's greater distribution to be practically equivalent when the difference between the sample and all their productions was less that a small effect. The data were coded for whether or not the test of equivalence or t-test were was significant for a given iteration (1 for yes/0 for no).

3. RESULTS

Figure 1 shows the percentage of practically equivalent results in each sample size for each segment. Each box in the plot is made up of 10 total data points, where each of them is the total number of practically equivalent samples out of 100. The figure suggests that at least 60 samples of a given segment are needed to achieve a sample that is within d = +/- .4 at least 80 percent of the time.

10 tokens per segment was the least successful at producing a practically equivalent sample, as all segments were practically equivalent 0 percent of the time. Table 4 reports the percentage of significant tests of equivalence for each segment and each sample size. Again, for each segment, at least 60 tokens are necessary to produce a practically equivalent sample at least 80 percent of the time.

Figure 1: Quantity of practically equivalent instances per sample size

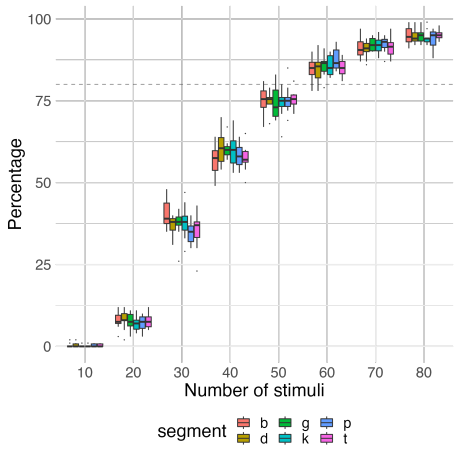


Table 4: Percentage of significant Tests of Equivalence per sample size and segment

n	p	t	k	b	d	g
10	0	0	0	0	0	0
20	0.08	0.08	0.07	0.07	0.07	0.08
30	0.4	0.37	0.37	0.38	0.35	0.35
40	0.57	0.61	0.6	0.6	0.58	0.58
50	0.75	0.74	0.75	0.74	0.75	0.76
60	**0.84**	**0.85**	**0.86**	**0.86**	**0.88**	**0.85**
70	**0.91**	**0.91**	**0.92**	**0.92**	**0.92**	**0.91**
80	**0.95**	**0.95**	**0.95**	**0.94**	**0.94**	**0.95**

3.1. False Positive results

In frequentist analysis, the significance threshold (alpha: typically .05 in linguistic research), refers to the false positive rate. Given that our significance threshold was .05, it is expected that around 5 percent of the total t-tests would be (falsely) positive. In total, 1903 of 48000 t-tests were significant (3.9 percent). Each of these iterations also had a significant test of equivalence, and there were 1900 instances of a significant t-test and an insignificant test of equivalence and 3 cases of a significant t-test and a significant test of equivalence. Figure 2 visualizes the pooled quantity of false positive tests in the whole data set by sample size. The figure suggests that the false positive rate does not vary much by segment or sample size, and stays around or below .05. Table 5shows the mean false positive rate for each segment and also shows that the false positive rate ranged from .04-.06, regardless of sample size or segment.

Figure 2: False Positive rate per sample size

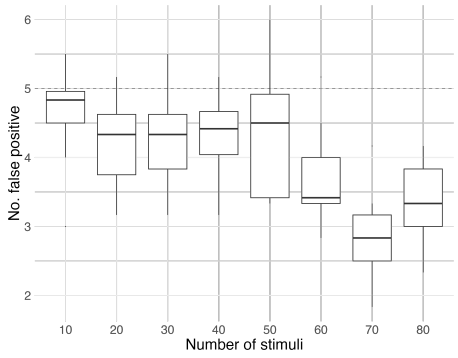


Table 5: Percentage of significant Welch's T-tests per sample size and segment

n	p	t	k	b	d	g
10	0.04	0.05	0.05	0.04	0.06	0.04
20	0.04	0.04	0.05	0.04	0.04	0.05
30	0.05	0.04	0.04	0.05	0.03	0.05
40	0.04	0.04	0.05	0.03	0.05	0.05
50	0.05	0.04	0.05	0.05	0.03	0.04
60	0.04	0.04	0.04	0.04	0.03	0.04
70	0.03	0.03	0.02	0.02	0.03	0.03
80	0.03	0.04	0.03	0.04	0.03	0.02

4. DISCUSSION

The results suggest that, at large, more tokens per segment should be used in speech research, at least for VOT studies in English. These results do not directly take into account various additional factors which have been found to impact VOT, and it is possible that fewer tokens might be needed in the event that there is reason to believe that less variation exists in the underlying distribution from which an experiment is sampling. The means and standard deviation from the present study [4] included each stop 50 times in 10 vocalic contexts, which therefore likely includes a wider range of possible values since the following vowel has been shown to impact VOT [16].

In general, this study serves as an example for speech researchers to consider how repeated measures are justified in their research, and should be considered in tandem with the smallest effect size of interest. For the purpose of this paper, the assumption was made that any sample that fell within a small effect size of the underlying distribution was practically equivalent, but this may not be a consensus for other measures, or even in VOT for different purposes. For example, very small effects might be taken as evidence for subtle changes in speech such as cross-linguistic influence, studies in phonetic drift or phonetic accommodation. In these cases, the quantity of repeated measures may need to be increased even more for better precision.

These results provide additional context in the issues surrounding low statistical power in not linguistic research with and emphasis on speech research. A recent review of linguistic research at large found that studies are typically under powered in regard to number of participants [3]. That is, the low quantity of participants per group or condition increase the probability of a false negative finding and virtually guarantee an over-estimated effect size when an effect is found. The finding, together with the present study, suggest that researchers analyze how many participants are necessary and how many tokens to elicit from each participant per condition prior to carrying out a study. Power analysis, again, provides researchers with a necessary tool for examining these questions.

The present study also includes open materials, including the code used to run all analyses and to produce this manuscript. Additionally, a shiny application was created to easily reproduce this analysis for a single segment given a mean and standard deviation. That is, by using the app and plugging in the mean and standard deviation, an underlying distribution of 1000 points is created, and samples of 10, 20,

30, 40, 50, 60, 70 and 80 are taken, randomly, 100 times each. A plot is then produced which shows the number of times that each of these samples is practically equivalent in a test of equivalence (bounds d = +/- .4). The equivalence bounds and sample sizes can also be adjusted.

5. CONCLUSION

In conclusion, the present study has argued that more attention should be paid to the justification of the quantity of repeated measures in speech research. In specific, it is recommended that 60 tokens per segment be given to a speaker in a specific condition. For example, if one wants to study VOT of a bilingual in two languages, at least 60 tokens per segment per language would be ideal, based on the current data. A companion shiny application was also created together with this paper, and can be used to quickly replicate this analysis during planning of research projects: https://kparrish92.shinyapps.io/repeated_measures_app/.

6. REFERENCES

[1] Antoniou, M., Best, C. T., Tyler, M. D., Kroos, C. 2011. Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of phonetics* 39(4), 558–570.

[2] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language* 59(4), 390–412.

[3] Brysbaert, M. 2021. Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition* 24(5), 813–818.

[4] Chodroff, E., Wilson, C. 2017. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in american english. *Journal of Phonetics* 61, 30–47.

[5] Cohen, J. 1992. Statistical power analysis. *Current directions in psychological science* 1(3), 98–101.

[6] Fish, M. S., García-Sierra, A., Ramírez-Esparza, N., Kuhl, P. K. 2017. Infant-directed speech in english and spanish: Assessments of monolingual and bilingual caregiver VOT. *Journal of Phonetics* 63, 19–34.

[7] Gabriel, C., Krause, M., Dittmers, T. 2018. VOT production in multilingual learners of french as a foreign language: Cross-linguistic influence from the heritage languages russian and turkish. *Revue française de linguistique appliquée* 23(1), 59–72.

[8] Gandour, J., Dardarananda, R. 1984. Voice onset time in aphasia: Thai ii. production. *Brain and Language* 23(2), 177–205.

[9] Gorba, C., Cebrian, J. 2022. The acquisition of l2 voiced stops by english learners of spanish and spanish learners of english. *Speech Communication*.

[10] Harada, T. 2003. L2 influence on L1 speech in the production of VOT. *Proceedings of the 15th international congress of phonetic sciences*. Causal Productions Barcelona 1085–1088.

[11] Hussain, Q. 2018. A typological study of voice onset time (VOT) in indo-iranian languages. *Journal of Phonetics* 71, 284–305.

[12] Khattab, G. 2002. VOT production in english and arabic bilingual and monolingual children. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* 1–38.

[13] Kupske, F. F. 2017. A complex approach on integrated late bilinguals' english VOT production: a study on south Brazilian immigrants in London. *Ilha do Desterro* 70, 81–94.

[14] Lakens, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8(4), 355–362.

[15] Llama, R., López-Morelos, L. P. 2016. VOT production by spanish heritage speakers in a trilingual context. *International Journal of Multilingualism* 13(4), 444–458.

[16] Nearey, T. M., Rochet, B. L. 1994. Effects of place of articulation and vowel context on VOT production and perception for french and english stops. *Journal of the International Phonetic Association* 24(1), 1–18.

[17] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39(2), 132–142.

[18] Olson, D. J. 2013. Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics* 41(6), 407–420.

[19] Plonsky, L., Oswald, F. L. 2014. How big is "big"? interpreting effect sizes in l2 research. *Language learning* 64(4), 878–912.

[20] Riney, T. J., Takagi, N., Ota, K., Uchida, Y. 2007. The intermediate degree of VOT in japanese initial voiceless stops. *Journal of Phonetics* 35(3), 439–443.

[21] Ryalls, J., Zipprer, A., Baldauff, P. 1997. A preliminary investigation of the effects of gender and race on voice onset time. *Journal of Speech, Language, and Hearing Research* 40(3), 642–645.

[22] Winter, B. 2015. The other n: The role of repetitions and items in the design of phonetic experiments. *ICPhS*.

[23] Zlatin, M. A., Koenigsknecht, R. A. 1976. Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research* 19(1), 93–111.

THE TIMING OF SECONDARY DORSAL ARTICULATIONS ACROSS SYLLABLE POSITIONS IN IRISH

Jaye Padgett[a], Ryan Bennett[a], Grant McGuire[a], Máire Ní Chiosáin[b], Jennifer Bellik[a]

University of California, Santa Cruz[a] and University College Dublin[b]
padgett@ucsc.edu, rbennett@ucsc.edu, gmcguir1@ucsc.edu, maire.nichiosain@ucd.ie, jbellik@ucsc.edu

ABSTRACT

All consonants in Irish are contrastively velarized /Cˠ/ or palatalized /Cʲ/. This study investigates the timing of /Cˠ Cʲ/ gestures across syllable positions in Irish. In Russian, which has a similar contrast, the dorsal gesture for /Cʲ/ seems to peak at the CV transition in prevocalic onset consonants. In postvocalic codas, /Cʲ/ gestures peak near the VC transition, may be held until C release, and may have more variable timing. We report comparable gestural timing in Irish, and propose that these patterns are grounded in perception. Both CV/VC transitions and C release carry important cues to /Cˠ Cʲ/ contrasts. In onsets, aligning /Cˠ Cʲ/ gestures at CV transition maximizes the salience of those cues. In codas, /Cˠ Cʲ/ gestures must be aligned to both VC transitions and C release to achieve the same effect. Cross-linguistically, /Cˠ Cʲ/ contrasts are often limited in coda position, plausibly reflecting these articulatory asymmetries.

Keywords: palatalization, velarization, Irish, articulatory coordination, syllable structure

1. THE TYPOLOGY OF SECONDARY PALATALIZATION CONTRASTS ACROSS SYLLABLE POSITIONS

Typologically, secondary dorsal contrasts involving palatalization are more likely to occur in syllable onsets (≈ prevocalically) than in syllable codas (≈ non-prevocalic) [1–4]. In Bulgarian, for example, palatalization contrasts only occur before (non-front) vowels: word-finally and before a consonant, there are no contrasts between plain /C/ and palatalized /Cʲ/ [5,6]. While there are some marginal exceptions to this typological generalization [7,8], the presence of a /C(ˠ) Cʲ/ contrast in coda position normally entails a corresponding contrast in onset position (but not vice-versa).

It has been argued that this asymmetry reflects perceptual factors: secondary dorsal contrasts like /C(ˠ) Cʲ/ may be easier to perceive in onsets when compared to codas [2,3,9–12]. It has also been noted that the articulatory coordination of secondary dorsal articulations may vary across syllable positions [2, 4, 13–15]. In onset /Cʲ/, the palatalization gesture tends to continue fronting during the course of consonant constriction, peaking at the release of the consonant at the CV transition. Though coda palatalization contrasts are less commonly studied, in Russian the secondary dorsal gesture for /Cʲ/ seems to peak earlier in coda position, around the VC transition, and is sometimes maintained throughout the consonant constriction [2, 4]. Coda /Cʲ/ may also show more variability in the timing of the secondary dorsal constriction than onset /Cʲ/. These articulatory asymmetries plausibly contribute to differences in the relative perceptibility of /C(ˠ) Cʲ/ contrasts across different syllabic contexts.

In this paper we examine the articulatory timing of secondary dorsal constrictions for palatalized /Cʲ/ and velarized /Cˠ/ stops in Irish, in word-initial onset /#CV/ and word-final coda /VC#/ contexts. Since Irish, like Russian, maintains a /Cˠ Cʲ/ contrast in both onsets and codas alike, it provides another language in which we can test the hypothesis that the typological dispreference for coda /C(ˠ) Cʲ/ contrasts is at least partially grounded in articulatory asymmetries between onset and coda positions.

2. BACKGROUND ON IRISH

Irish (or 'Gaelic') is spoken daily by ∼74,000 people in Ireland [16]. These speakers are concentrated in Irish-speaking communities on the western coast, though significant speaker populations are also present in urban areas. Much of the Irish population reports at least some ability in the language (∼1.75 million people). While Irish has some limited state support in the Republic of Ireland, the language is at risk even in traditional Irish-speaking communities, due to the centuries-long hegemony of English.

2.1. Secondary dorsal articulations in Irish

All consonants in Irish have contrastively palatalized /Cʲ/ and velarized /Cˠ/ variants [10, 11, 17]. The /Cˠ Cʲ/ distinction can be used to indicate grammatical

as well as lexical contrasts (1). Palatalization and velarization are contrastive in pre-vocalic (≈ onset) position as well as in word-final (≈ coda) position.

(1) a. *bó* /bˠoː/ 'cow'
beo /bʲoː/ 'alive'
b. *súil* /sˠuːlʲ/ 'eye'
siúl /sʲuːlˠ/ 'walk'
c. *casúr* /kˠasˠuːrˠ/ 'hammer'
casúir /kˠasˠuːrʲ/ 'hammers'

In Irish, palatalized /Cʲ/ is consistently associated with tongue body raising and fronting (= an [i]-like gesture), while velarized /Cˠ/ is consistently associated with tongue body backing, and some degree of raising (= an [u]- or [o]-like gesture) [13]. This is true across places and manners of articulation, as well as different vowel contexts (at least before [iː uː]). Coronal and velar consonants shift their primary place of articulation depending on their secondary articulations: /Tˠ/ tends to be dental or alveolar, and /Tʲ/ tends to be alveo-palatal or postalveolar; while /Kˠ/ tends to be velar or uvular, and /Kʲ/ palatal [18, 19].

Acoustically, palatalized consonants have raised second formants (F2) in neighboring vowels relative to their velarized counterparts [10, 18]. Velarized consonants are frequently produced with lip rounding, which likely enhances F2 differences associated with the /Cˠ Cʲ/ contrast [20]. Palatalized stops have louder and longer bursts, with higher spectral center of gravity in their release noise, especially when comparing coronal to labial stops. The spectral shape of fricative noise very audibly distinguishes the coronal and velar fricatives /sˠ sʲ xˠ xʲ/. [10, 11] report that speakers of Irish perceive /Cˠ Cʲ/ stop contrasts more reliably in onset position than in coda position.

2.2. Timing of /Cˠ Cʲ/ in Irish onsets

[13] examine tongue body position in /CVː/ sequences in Connacht Irish on the basis of ultrasound tracings of stops and fricatives, analyzed using principal component analysis. They find that palatalized /Cʲ/ tends to have more extreme tongue body fronting at C release (= CV transition) than at the beginning of C constriction, or at C midpoint. In contrast, velarized /Cˠ/ does not show any clear, consistent asymmetries across these timepoints: the magnitude of dorsal backing tends to be comparable between the beginning of C constriction, C midpoint, and C release. [13] only examine consonants in word-initial, prevocalic (onset) position, and so do not address the timing of secondary dorsal constrictions in word-final (coda) consonants.

3. METHODS

We recorded two native speakers of Ulster Irish (24, M; 40, M), three of Connacht Irish (42, F; 50, F; 43, F), and two of Munster Irish (34, M; 56, M). The materials were monosyllabic words with a target stop consonant controlled for secondary articulation (/Cʲ/ or /Cˠ/), place of articulation (labial, coronal, dorsal), position (word-initial or -final) and adjacent vowel /iː uː ɔː/. (We do not address the vocalic environment here for lack of space.) Speakers repeated the 36 experimental items five times each in the frame sentence [ˈdˠuːrʲtʲ ˈiːfˠə __ əˈnˠʊrˠə] 'Aoife said ___ last year'. Midsagittal images of the tongue body were obtained with a portable Terason T3000 ultrasound system and model 8MC3 probe (≈ 60 fps, or 1 frame every 17 ms). The probe was held in place with an Articulate Instruments ultrasound stabilization headset [21]. Using a time-synchronized audio recording, we extracted video frames closest to the onset and offset of closure of the target consonant. These ultrasound images were then traced using EdgeTrak [22], and trace coordinates were range-normalized by speaker. The analyses reported here are based on the position of the highest point of the tongue body in each frame. (If this was a plateau, the plateau's center was used).

4. HYPOTHESES

Considering prior research on Russian and Irish [2, 4, 10, 13, 14], we expect that the secondary dorsal articulation for /Cʲ/ in onset position will reach a maximum at the C release, which coincides with the CV transition. Stop releases are known to be important landmarks for articulatory alignment [23]; further, aligning secondary /Cʲ/ articulations to the consonant release should maximize the acoustic cues to the /Cˠ Cʲ/ contrast at the perceptually-important CV transition [4, 12, 24–26]. We expect particularly consistent release alignment for labial consonants: secondary lingual articulations like /Cˠ Cʲ/ are for the most part biomechanically independent from labial gestures, and so there is greater freedom for labial consonants to implement a perceptually-optimal /Cˠ Cʲ/ alignment pattern [2].

Our expectations for coda position, and for velarization, are less firm. In coda position, the VC transition does not align with stop release. Both the VC transition and stop release carry potentially important acoustic cues to /Cˠ Cʲ/ contrast. Hence, we may find more variable patterns of alignment in coda position, depending on whether speakers

prioritize the acoustic expression of /Cˠ Cʲ/ contrasts during the VC transition, or during stop release instead. (Non-stop consonants also have steady-state cues to /Cˠ Cʲ/ contrasts, which may lead to different expectations about their internal gestural timing.) Speakers may also align the peak of /Cˠ Cʲ/ gestures at the VC transition, and maintain that gesture until C release. As for velarization, past work has not found any consistent tendency for /Cˠ/ gestures to align with consonant release in onset position, as observed for /Cʲ/. Regardless of syllabic position, velarization seems fairly static over the duration of the consonant. However, relevant data are limited.

It has been reported for English that more open constrictions (e.g. secondary velarization in /l/) tend to precede more narrow constrictions (e.g. tongue tip closure for /l/) for articulatorily complex consonants in coda position [14, 27, 28]. This tendency may reflect anti-phase coordination of gestures in the coda [29], and could lead to a greater bias toward VC alignment of /Cˠ Cʲ/ gestures in coda position. However, [2, 4, 14] cast doubt on this notion for Russian, finding that peak constriction for secondary /Cˠ Cʲ/ articulations often aligns with consonant closure in codas, and may be maintained through the entire consonant.

5. RESULTS

Fig. 1 plots the difference in tongue body backness (measured at the highest point of the tongue body) between consonant end (= release) and consonant start (= beginning of closure). A value of zero indicates that the peak of the dorsum has the same backness at C start and C end. Distributions significantly different from zero are marked with a boxed REL or VC, indicating their apparent temporal alignment with C release or VC transition, respectively. Statistical significance was assessed with Bonferroni-corrected one-way *t*-tests, with the significance threshold $\alpha = 0.05/12 = 0.0042$.

For palatalized /Cʲ/, a *positively* shifted distribution reflects a more extreme *fronting* gesture at C release than at the beginning of closure. This is the expected distribution in syllable onsets, where the C release coincides with the CV transition: the gestural maximum for the palatalization gesture should occur at C release. As expected, the distributions for onset /Cʲ/ are all significantly positively shifted in onset position.

For velarized /Cˠ/, a *negatively* shifted distribution reflects a more extreme *backing* gesture at C release than at the beginning of closure. This is a plausible expectation for onset position:

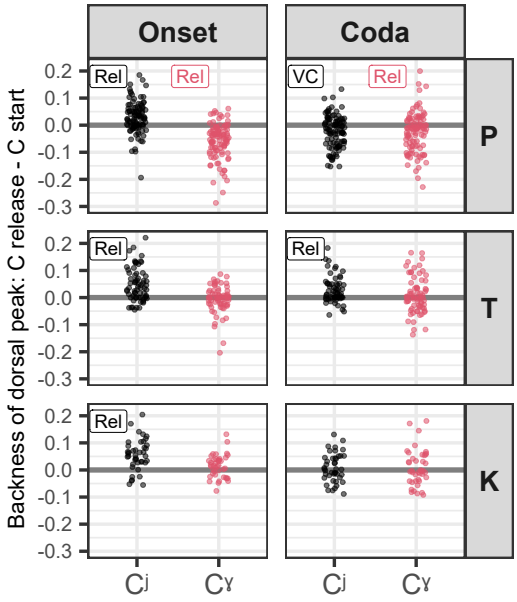


Figure 1: Backness of dorsal peak at C release (C end) − beginning of closure (C start). Units are in speaker-normalized [0,1] range. Distributions significantly different from zero (= no difference) are marked with 'Rel' for alignment with C release, or 'VC' for alignment with VC transition.

dorsal backing for /Cˠ/ should peak at C release, as fronting for /Cʲ/ does. However, prior work has not clearly shown asymmetries across consonant landmarks for /Cˠ/ [13]. In fact, we find a negative shift in onset position only for /Pˠ/; onset /Tˠ Kˠ/ show no significant shift at all.

We now turn to coda position. Coda /Pʲ/ has a negative shift, suggesting that the greatest fronting of the palatalization gesture for /Pʲ/ tends to occur at the VC transition rather than at C release. In contrast, coda /Tʲ/ has a positive shift, suggesting that palatalization peaks at C release instead. Coda /Kʲ/ shows no significant shift in either direction.

Coda /Pˠ/ has a negative shift, which indicates that the backing gesture for coda /Pˠ/ tends also to peak at C release. Otherwise, there are no significant shifts for coda /Tˠ Kˠ/.

In summary, onset /Cʲ/ shows alignment to the C release = CV transition, across places of articulation. The same is true of onset /Pˠ/, but not onset /Tˠ Kˠ/. In coda position, alignment patterns are less consistent. Coda /Pʲ/ shows VC alignment, while coda /Tʲ/ shows release alignment. Coda /Pˠ/ shows release alignment, but coda /Tˠ Kˠ/ do not show any particular alignment pattern.

However, care must be taken in interpreting non-significant differences between C start and C end (apart from the general risks of interpreting null results). A null result in Fig. 1 could reflect

at least three different scenarios. First, it could be that the gestural alignment of /C$^{\gamma}$ C^{j}/ varies rather categorically between C start and C end. In this scenario, we might expect a wide, bimodal distribution with both positive and negative clusters, corresponding to a mix of VC vs. C release alignment in coda position. No such distribution is evident in Fig. 1.

Second, it could be that the peak dorsal gesture is inconsistently timed, landing anywhere between the closure and release of the primary constriction, depending on the utterance. In this scenario we would expect a loose and unimodal distribution of points around zero in Fig. 1.

Third, it is possible that some /C$^{\gamma}$ C^{j}/ gestures are held mostly consistent across the consonant closure. In coda position this pattern would reflect VC alignment, but with the dorsal gesture maintained until release. Given that important perceptual cues to coda /C$^{\gamma}$ C^{j}/ contrasts occur at *both* C start (= VC transition) and C end (= C release), we might in fact expect exactly this alignment pattern for coda consonants. In this scenario we would expect a tight spread of points around zero in Fig. 1.

To tentatively explore our null results, we computed Shapiro-Wilk tests to assess whether any of the distributions in Fig. 1 might be more spread out than a normal distribution (with Bonferroni-corrected $\alpha = 0.05/12 = 0.0042$). Onset /T$^{\gamma}$/ significantly departs from normality according to this test, along with onset /P^{j} P$^{\gamma}$ T^{j}/ and coda /T^{j}/ (which are also shifted away from zero). We also computed Hartigan's dip test to investigate whether any of the distributions in Fig. 1 are multimodal: this test did not reject the hypothesis of unimodality for any of those distributions, even with uncorrected $\alpha = 0.05$. All of the distributions in Fig. 1 appear to be unimodal, and most distributions centered at zero appear to be normal in shape. We thus cautiously conclude that coda /C$^{\gamma}$ C^{j}/ may show either VC alignment of the dorsal gesture, maintained until C release, or inconsistent alignment of the dorsal gesture across different landmarks (scenarios two and three above). However, we reiterate that we are drawing these conclusions mostly from null results, and so they should be interpreted cautiously.

6. DISCUSSION

Our main finding is a difference in gestural timing in syllable onset vs. coda position (corresponding to word-initial vs. -final position in our experiment) for palatalized consonants in Irish. While the peak of tongue body frontness in /C^{j}/ coincides more with the release of a consonant's primary constriction (= CV transition) in syllable onsets, timing in the coda is more ambiguous, not clearly favoring alignment with either C closure or C release.

As noted earlier, /C$^{\gamma}$ C^{j}/ contrasts are more stable in syllable onset (or word-initial) position compared to syllable-final (or word-final) position, and the contrast is also harder to perceive in coda (word-final) position. It is possible that our different findings for syllable onset vs. coda gestural timing bear on these facts. In syllable onset position, all cues to stop palatalization vs. velarization coincide at the CV transition point; in such a case, it seems adaptive to time the peak dorsal gesture for palatalization with the CV transition. In coda position, on the other hand, formant transition cues and consonant release cues are necessarily *disjoint* in time, separated by the consonantal closure. If gestural timing can serve to optimize cues to contrast, then uniform alignment to consonant release will not accomplish this goal in the case of coda consonants. Articulatorily, anti-phase gestural coordination patterns in codas could also lead to greater variability in timing [29]. In future work we plan to further explore the nature of the more ambiguous alignment facts in coda position.

As a separate question, why is it that labial stops stand out in our results as showing more consistent alignment with a consonantal landmark? This may be due to articulatory constraints: in the case of /T^{j} T$^{\gamma}$/ or /K^{j} K$^{\gamma}$/, there is tongue coupling or direct articulatory competition between the primary articulator and a secondary dorsal one, while this is not true of /P^{j} P$^{\gamma}$/. This might allow other factors, such as constraints on perceptibility, to more freely determine timing in the latter case. In addition, perception of the secondary contrast may depend more on formant transition cues for /P^{j} P$^{\gamma}$/ than for /T^{j} T$^{\gamma}$/ or /K^{j} K$^{\gamma}$/, because the former contrast has less robust support from cues associated with a release burst [10]. It is interesting in this context to note that palatalization contrasts are typologically less stable in labials than in coronals or dorsals [1,2]. How far these ideas go in illuminating our results for labials is a matter for future research.

Finally, why has prior work (discussed above) found more ambiguous alignment of velarization gestures compared to palatalization ones, regardless of position, something that may be true also of our current results? This is also a question for future research, but we note that lip rounding is known to play a role in velarization, at least in Irish [20]. A fuller understanding of the timing facts may require attention to the lips as well as the tongue.

7. REFERENCES

[1] Y. Takatori, *A study of constraint interaction in Slavic phonology*. New Haven, CT: Yale University, 1997.

[2] A. Kochetov, *Production, perception and emergent phonotactic patterns: A case of contrastive palatalization*. Routledge, 2002.

[3] ——, "Perception of place and secondary articulation contrasts in different syllable positions: language-particular and language-independent asymmetries," *Language and Speech*, vol. 47, no. 4, pp. 351–382, 2004.

[4] ——, "Syllable position effects and gestural organization: Evidence from Russian," in *Papers in Laboratory Phonology VIII*, L. Goldstein, D. Whalen, and C. Best, Eds. Berlin: Mouton de Gruyter, 2006, pp. 565–588.

[5] E. Scatton, "Bulgarian," in *The Slavonic languages*, G. Corbett and B. Comrie, Eds. London: Routledge, 1993, pp. 188–248.

[6] K. Iskarous and D. Kavitskaya, "Sound change and the structure of synchronic variability: Phonetic and phonological factors in Slavic palatalization," *Language*, vol. 94, no. 1, pp. 43–83, 2018.

[7] I. Chitoran, *The Phonology of Romanian*. Berlin: De Gruyter, 2013.

[8] A. Malmi, P. Lippus, and E. Meister, "Spectral and temporal properties of Estonian palatalization," *JIPA*, vol. TBD, no. TBD, p. TBD, to appear.

[9] A. Kochetov, "Testing licensing by cue: A case of Russian palatalized coronals," *Phonetica*, vol. 63, no. 2-3, pp. 113–148, 2006.

[10] M. Ní Chiosáin and J. Padgett, "An acoustic and perceptual study of Connemara Irish palatalization," *JIPA*, vol. 42, no. 2, pp. 171–191, 2012.

[11] J. Padgett and M. Ní Chiosáin, "The perception of a secondary palatalization contrast: a preliminary comparison of Russian and Irish," in *A Festschrift for Junko Itô and Armin Mester*, R. Bennett, A. Brasoveanu, D. Buckley, N. Kalivoda, S. Kawahara, G. McGuire, and J. Padgett, Eds. Santa Cruz, CA: Department of Linguistics, University of California, Santa Cruz, 2018. [Online]. Available: https://itomestercelebration.sites.ucsc.edu/

[12] R. Wright, "A review of perceptual cues and cue robustness," in *Phonetically based phonology*, B. Hayes, R. Kirchner, and D. Steriade, Eds. Cambridge, UK: Cambridge University Press, 2004, pp. 34–57.

[13] R. Bennett, M. Ní Chiosáin, J. Padgett, and G. McGuire, "An ultrasound study of Connemara Irish palatalization and velarization," *JIPA*, vol. 48, no. 3, pp. 261–304, 2018.

[14] K. Iskarous and D. Kavitskaya, "The interaction between contrast, prosody, and coarticulation in structuring phonetic variability," *Journal of phonetics*, vol. 38, no. 4, pp. 625–639, 2010.

[15] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Malden, MA: Blackwell, 1996.

[16] R. o. I. Government of Ireland, "20-year strategy for the Irish language 2010–2030," Dublin, Ireland. [Online]. Available: http://www.chg.gov.ie/gaeltacht/20-year-strategy-for-the-irish-language-2010-2030/

[17] M. Ní Chiosáin, "Topics in the phonology of Irish," Ph.D. dissertation, University of Massachusetts, Amherst, 1991.

[18] A. Ní Chasaide and G. Fealy, "A preliminary investigation of palatal/palatalised consonants of Irish," 1991, unpublished ms.

[19] A. Ní Chasaide, "Irish," *JIPA*, vol. 25, no. 1, pp. 34–39, 1995.

[20] R. Bennett, J. Padgett, M. Ní Chiosáin, G. McGuire, and J. Bellik, "Contrast enhancement and cue trading in Irish secondary consonant articulations," in *Proceedings of ICPhS 19, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 3584–3552. [Online]. Available: https://assta.org/proceedings/ICPhS2019/papers/ICPhS_3597.pdf

[21] A. Wrench, *Ultrasound Stabilisation Headset users manual*, Articulate Instruments Ltd.

[22] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 545–554, 2005.

[23] J. Kingston, "Articulatory binding," in *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman, Eds. Cambridge, UK: Cambridge University Press, 1990, pp. 406–434.

[24] O. Fujimura, M. Macchi, and L. Streeter, "Perception of stop consonants with conflicting transitional cues," *Language and Speech*, vol. 21, pp. 337–343, 1978.

[25] J. Ohala, "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman, Eds. Cambridge, UK: Cambridge University Press, 1990, pp. 258–275.

[26] D. Steriade, "Directional asymmetries in place assimilation: a perceptual account," in *The role of speech perception in phonology*, K. Johnson and E. Hume, Eds. New York: Academic Press, 2001, pp. 219–250.

[27] R. A. Krakow, "Physiological organization of syllables: a review," *Journal of Phonetics*, vol. 27, no. 1, pp. 23–54, 1999.

[28] R. Sproat and O. Fujimura, "Allophonic variation in English /l/ and its implications for phonetic implementation," *Journal of phonetics*, vol. 21, pp. 291–311, 1993.

[29] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: A coupled oscillator model," in *Approaches to phonological complexity*, F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé, Eds. Berlin: Mouton de Gruyter, 2009, pp. 297–328.

COARTICULATION AND CONTRAST IN A VOWEL HARMONY SYSTEM: COARTICULATORY PROPENSITY IN KHALKHA MONGOLIAN V-C-V SEQUENCES

Auromita Mitra[1], Meghavarshini Krishnaswamy[2], Indranil Dutta[3]

New York University[1], University of Arizona[2], Jadavpur University[3]
auromita.mitra@nyu.edu, mkrishnaswamy@arizona.edu, indranildutta.lnl@jadavpuruniversity.in

ABSTRACT

Vowel harmony has been understood to emerge when listeners fail to perceptually compensate for acoustic variation due to coarticulation. Assuming such an account, what explains the maintenance of non-harmonic domains in the grammar? Towards understanding this, we examine coarticulation within a synchronic system with well-established patterns of harmony and non-harmony. In Khalkha Mongolian, vowels in non-compound words share the features [ATR] and [round], harmony operating in the carryover (left-to-right) direction. The high-front vowel /i/ does not participate in harmony, giving "non-harmonic" VCV sequences. We quantify coarticulatory variation by comparing dependencies in first- and second-formant frequencies (F1&F2) of vowels in harmonic vs non-harmonic VCV sequences. Unlike the former, the latter show greater coarticulation in the anticipatory (right-to-left) direction—opposite to that of vowel harmony. /i/, which is transparent to harmony, demonstrates high coarticulatory resistance [1]. We argue that in systems where vowel harmony is well-established, synchronic patterns of coarticulatory propensity serve to limit feature-sharing in non-harmonic domains.

Keywords: Khalkha Mongolian, vowel harmony, coarticulation, vowels, formants

1. COARTICULATION AND VOWEL HARMONY

Vowel harmony is a grammatical feature of languages whereby sequences of vowels (contiguous or non-contiguous) within a certain domain, e.g. the word, obligatorily have the same value for certain phonological features. This is a productive rule, and leads to predictable variation in vowels depending on phonological context. This could be left-to-right: *carryover*, where features of a preceding vowel affect those of the following vowels, or right-to-left: *anticipatory* where the converse happens. Independent of this grammatical process, an essential feature of speech is that sounds are not produced in discrete sequences. Rather, articulatory gestures overlap during connected speech, and this kind of *coarticulation* results in acoustic variation. Coarticulation could be carryover when the gestures of an earlier target perturb the latter in a mechanico-inertial fashion, or anticipatory, where articulatory planning for a subsequent vowel affects the production of an earlier vowel. To extract stable phonological categories, from such a variable signal, listeners must perceptually compensate for the acoustic variation resulting from coarticulation [2, 3]. Given the parallels between this kind of physiological-psychoacoustic process and grammatical harmony processes, vowel harmony has been understood to diachronically emerge when listeners fail to perceptually compensate for acoustic variation due to coarticulation, thereby perceiving the coarticulated variation as part of the phonological feature of the category [4, 3]. Under this hypothesis, the greater persistence of carryover coarticulation would favor carryover harmony, leading to left-to-right harmony systems, while the persistence of anticipatory coarticulation would result in to anticipatory, right-to-left harmony systems. Of note here is that coarticulation is usually understood to be a non-voluntary by-product of articulation, and therefore arguably continues to exist even after harmony has developed. Assuming the account of causation, once a psycho-acoustic process such as the perceptual persistence of e.g. carryover coarticulation has been set in motion, what prevents the system from coarticulating further, assimilating more features, and eventually resulting in a radical system of vowel copying in the grammar? While such systems exist (e.g. Telugu; [5]), other harmony systems appear to be in a synchronic state of equilibrium, where harmony operates over certain features and segments, but not others. Specifically, most harmony systems also

contain some form of *non-harmony*– sequences or segments that do not participate in harmony, either by completely ignoring it (transparent segments), or blocking it (opaque segments). How do such sequences persist in synchronic systems, given the continued existence of coarticulation in the direction of harmony, and the system's proclivity for its perceptual non-compensation? That is, what checks does the language system put in place to prevent the spillover of the very processes that initiated harmony, into its grammatically non-harmonic domains? Towards understanding this, this project examines how coarticulation functions within a system with well-established and robust vowel harmony. We measure the acoustic consequences of coarticulation in harmonic and non-harmonic sequences, and compare the patterns of coarticulatory propensity in these sequences.

2. VOWEL SYSTEM OF KHALKHA MONGOLIAN

Khalkha Mongolian has seven monophthongal vowel categories, classified as non-pharyngeal (+ATR) and pharyngeal (-ATR) [6]:

	[+ATR]	[-ATR]	neutral
high	u	ʊ	i
non-high	e, o	a, ɔ	

Table 1: Monopthongs in Khalkha Mongolian, classified by harmony class

The inventory has four diphthongs that are contrastive: ai, ɔi, ui, ʊi. For the purposes of vowel harmony, these pattern identically to their corresponding monophthongal counterparts (a, ɔ, u, ʊ). There is also a vowel length distinction that does not play a role in vowel harmony, and is not discussed it here.

Vowel harmony

Vowels in non-compound words must have identical value for the [ATR] feature. A subset of vowels (non-high: e, o, a, ɔ) also show rounding harmony, whereby any instance of e, o, a, ɔ in a non-initial position must match the [round] feature of the initial vowel. We focus on ATR harmony. Vowel harmony proceeds left-to-right (carryover harmony): ATR is contrastive only in word-initial position, and vowels in non-initial positions must match the value of the initial vowel. This means that the language has two types of words: [+ATR] and [-ATR]. /i/ does not show phonological alternation; it can occur in both [+ATR] and [-ATR] environments. Thus, although the phonetic realization has been shown to vary slightly across contexts, causing researchers to analyze it as an allophone [6], the category is phonologically unaffected by harmony. Moreover, when /i/ intervenes between the initial and subsequent vowels, harmony proceeds across it – it is 'transparent' to harmony. In our study, we focus exclusively on disyllablic words. Thus, words in which the second vowel is /i/ do not demonstrate harmony, resulting in 'non-harmonic' sequences.

2.1. Questions and Hypotheses

[ATR] (advanced tongue root) is a phonological feature, whose acoustic correlates include the first formant frequency (F1) [7, 8]. Our aim is to examine whether patterns of coarticulation differ between words that involve grammatical ATR harmony (harmonic sequences), and those that don't (non-harmonic sequences). Since we are using acoustic data, we want to identify acoustic consequences of coarticulation in the acoustic dimensions of interest (here, formant frequencies), and quantify those. One approach is to think about coarticulation as constraining the random variability of individual articulatory gestures– if the articulation of sound B is physiologically affected by that of sound A, then we expect that each time a speaker produces the sequence AB, some of the variability in the articulation of B will be explained by the presence of A (compared to, e.g., CB). Assuming (simplistically) that articulatory variability has acoustic consequences, we operationalizie V-to-V coarticulation as constraints on variability in vowel acoustics. We assume that if the articulation of vowel 1 affects that of vowel 2, then some of the variability in the acoustics of 2 will be explained by the identity of 1. We will treat the degree of dependency as a proxy for the extent of coarticulatory effect. Within a V1CV2 sequence, a greater dependency of V1 on V2 suggests greater anticipatory coarticulation, while the opposite suggests greater carryover coarticulation. This gives us the *directionality* of coarticulation within a sequence. The aim of the current study is to examine how coarticulation behaves in synchronic data in a system where vowel harmony (and its domains) have already been grammaticalized. Therefore, we ask whether the extent and directionality of coarticulation differ between harmonic and non-harmonic VCV sequences.

Since vowel harmony in Khalkha Mongolian is in the left-to-right carryover direction, we expect an overall privileging of carryover coarticulation in harmonic domains of the language. This would

follow from previous accounts of the diachronic development of harmony. However, this does not necessarily translate to an assumption about a greater extent of coarticulation per se. The accounts in, e.g. Ohala (1994) [4] are perceptual. Carryover coarticulation could be privileged by means of lesser perceptual *compensation*, even if the extent of coarticulation were equal in both directions.

Our main interest in this study is in the non-harmonic sequences. We expect that if the system actively preserves non-harmony in certain domains, then such domains will differ in coarticulatory patterns, compared to domains in which harmony has developed (diachronically) as a result of coarticulation. Exactly what devices a language might employ to preserve grammatical domains is not something we have a priori assumptions about. However, we hypothesize that carryover coarticulation in non-harmonic sequences would be restricted in some way. This could manifest as either: (i) extent: coarticulation is primarily in the carryover direction, but the magnitude is smaller than in harmonic sequences; (ii) directionality: coarticulation has a different pattern across harmonic and non-harmonic words.

3. METHODOLOGY

3.1. Participants and materials

Participants were 14 first-language speakers of Khalkha Mongolian, recruited at the EFL University campus, Hyderabad, where they were enrolled in a short-term English language course. All the participants were from Ulaanbataar, between 25 and 40 years of age. We recorded speakers individually in a sound-attenuated booth, reading a set of words in frame sentences, and a passage, all presented on a computer screen in the Cyrillic script that is used commonly to read and write Khalkha Mongolian. Only the former was analyzed for this study. Each participant read 4 blocked repetitions of 59 target words, in the frame [*pi X gesen*] "I said X" (critical items: 59×14×4 = 3304). Targets were disyllabic non-compound words of the form (C)V1CV2(C), taken from a book on the description of the sound system of Mongolian [6]. We included words from examples throughout the text, creating a set with all allowed sequences of vowels in initial (v1) and non-initial (V2) positions. Words in which /i/ occupied the V2 position were classified as "non-harmonic", and all others as "harmonic". We removed one item from an initial set of 60 targets, because it was found that it is pronounced as monosyllabic, in spite of having two vowel characters in the orthography.

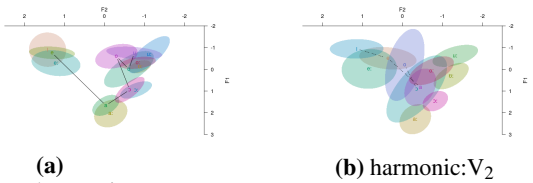
3.2. Acoustic analysis

A subset of recordings (including reading passages and frame sentences) was used to train an acoustic model with the Montreal Forced Aligner [9] to segment and annotate the data. F1 and F2 data were extracted with a customized Praat script [10]. We analyze Lobanov-normalized F1 and F2 at vowel midpoints. To quantify coarticulatory propensity, we use linear mixed effects models to examine how well the identity of V2 explains variance in the acoustics of V1 (anticipatory) and vice-versa (carryover).

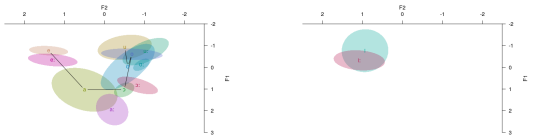
4. RESULTS

4.1. Vowel space diffusion

Plotting F1 and F2 values of vowels in the initial (V1) and non-initial (V2) position (figure 1) shows the following patterns: for harmonic sequences, the vowel space for V2 is more diffused than V1, suggesting greater variability due to coarticulation. Since the non-harmonic sequences contain only a subset of the vowel space (which are in complementary distribution to V1), the V2 vowel space is not informative. However, the vowel space for V1 shows that the low front vowel /a/ and the high back vowel /u/ are more diffused in the initial position of a non-harmonic sequence, compared to a harmonic sequence. This suggests differences in patterns of variability between harmonic and non-harmonic subsets. This is probed in the statistical analyses.



(a) harmonic:V_1 **(b)** harmonic:V_2



(c) non-harmonic:V_1 **(d)** non-harmonic:V_2

Figure 1: Steady-state vowel spaces for harmonic and non-harmonic vowel sequences in Khalkha Mongolian

4.2. Statistical analyses

For any vowel, we expect F1 and F2 values to be predicted by the phonological identity of the vowel. To quantify coarticulation, we used mixed effects models [11] in R (version 4.1.1) [12] to probe how well the formant frequency of a segment (dependent variable) is predicted by the identity of the *other* vowel in the word (independent variable), giving a measure of coarticulatory variability. Moreover, we examine whether this tendency differs between harmonic and non-harmonic words by comparing effect sizes of the independent variable in each model using the *effectsize* [13] package in R. For each model, we started with the maximal random effects structure motivated by the study design: by-item and by-speaker random slopes. This was iteratively simplified until the model converged [14, 15]. For all mixed models, the alpha criterion was set at $|t| > 2$. Table-1 summarizes the final model and output for F1. Here t5 refers to the formant measure at the steady state of the vowel. The results indicate robust coarticulation in both directions, with larger effects in the carryover (left-to-right) direction.

5. DISCUSSION AND CONCLUSION

The statistical analyses demonstrate an asymmetry in coarticulatory propensity between harmonic and non-harmonic sequences in Khalkha Mongolian. Specifically, while harmonic sequences show greater coarticulation in the carryover direction, parallel to that of harmony, non-harmonic sequences show the opposite pattern: coarticulation is greater in the anticipatory direction, which is opposite to that of harmony. We suggest that greater anticipatory coarticulation in these sequences counteracts the overall tendency of the system to privilege carryover coarticulation perceptually, and thus serves to check the gradual development of harmony in grammatically non-harmonic domains. Note that in Khalkha Mongolian, the only vowel that occurs in the non-initial position in non-harmonic sequences if /i/. The lack of significant carryover coarticulation in these sequences suggests that the segment has high coarticulatory resistance compared to the other vowels. This phonetic 'special-ness' tallies with research that has ascribed a special status to the segment in the phonological inventory of Mongolian, where it functions as a 'default' vowel in diphthongs and epenthetic processes [6]. These facts are compatible with two possible threads of interpretation: (i) the language system maintains non-harmony by developing higher coarticulatory resistance in the segments that occur in non-harmonic domains; (ii) non-harmonic domains develop diachronically around segments that have higher coarticulatory resistance for independent reasons. More cross-linguistic investigations in a variety of harmony systems, particularly those with non-harmonic domains involving multiple segments, will allow us to tease these apart. The findings of the present study suggest that in vowel harmony systems, articulatory and psychoacoustic processes might themselves function as checks to maintain synchronic equilibrium and limit harmony to certain domains.

Harmony type	Direction	Model fixed effects	ChiSq	Df	p	effect size (η^2) [1]
harmonic	anticipatory	F1V1t5 ~ V1+**V2**	17.174	9	0.04606 *	0.322
	carryover	F1V2t5 ~ V2+**V1**	34.131	11	0.003443 ***	0.536
non-harmonic	anticipatory	F1V1t5 ~ V1+**V2**	100.87	1	< 2.2e-16 ***	0.133
	carryover	F1V2t5 ~ V2+**V1**	133.41	10	< 2.2e-16 ***	0.174

Table 2: Model outputs for coarticulation in F1, compared to a null model (bold)

Harmony type	Direction	Model fixed effects	ChiSq	Df	p	effect size (η^2)
harmonic	anticipatory	F2V1t5 ~ V1+**V2**	9.3863	9	0.4024	0.191
	carryover	F2V2t5 ~ V2+**V1**	22.79	11	0.01892 *	0.404
non-harmonic	anticipatory	F2V1t5 ~ V1+**V2**	110.57	1	< 2.2e-16 ***	0.146
	carryover	F2V2t5 ~ V2+**V1**	74.809	10	5.182e-12 ***	0.101

Table 3: Model outputs for coarticulation in F2, compared to a null model (bold)

6. REFERENCES

[1] D. Recasens, J. Fontdevila, and M. D. Pallarès, "Velarization degree and coarticulatory resistance for /i/ in Catalan and German," *Journal of Phonetics*, vol. 23, no. 1–2, pp. 37 – 52, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S009544709580031X

[2] P. S. Beddor and R. A. Krakow, "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," *The Journal of the Acoustical Society of America*, vol.

106, no. 5, pp. 2868–2887, 11 1999. [Online]. Available: https://doi.org/10.1121/1.428111

[3] P. S. Beddor, J. D. Harnsberger, and S. Lindemann, "Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates," *Journal of Phonetics*, vol. 30, no. 4, pp. 591 – 627, 2002.

[4] J. J. Ohala, "Towards a universal, phonetically-based, theory of vowel harmony," in *Third International Conference on Spoken Language Processing*, 1994.

[5] I. Dutta, S. Irfan, and K. Harsha, "Using ANNs for vowel identification from V-to-V coarticulation in non-harmonic VCV sequences," in *15th Conference on Laboratory Phonology, Ithaca, NY*, 2016. [Online]. Available: http://labphon.org/labphon15/long_abstracts/LabPhon15_Revised_abstract_279.pdf

[6] J.-O. Svantesson, A. Tsendina, A. Karlsson, and V. Franzén, *The phonology of Mongolian*. OUP Oxford, 2005.

[7] C. G. A. Starwalt, "The acoustic correlates of atr harmony in seven-and nine-vowel african languages: A phonetic inquiry into phonological structure," Ph.D. dissertation, The University of Texas at Arlington, 2008.

[8] P. Olejarczuk, M. A. Otero, and M. M. Baese-Berk, "Acoustic correlates of anticipatory and progressive [atr] harmony processes in ethiopian komo," *Journal of Phonetics*, vol. 74, pp. 18–41, 2019.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2009. [Online]. Available: http://www.praat.org

[11] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[13] M. S. Ben-Shachar, D. Lüdecke, and D. Makowski, "effectsize: Estimation of effect size indices and standardized parameters," *Journal of Open Source Software*, vol. 5, no. 56, p. 2815, 2020.

[14] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

[15] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," *arXiv preprint arXiv:1506.04967*, 2015.

EARLY ACTIVITIES IN THE FIELD OF X-RAY IMAGING IN EXPERIMENTAL PHONETICS AND SPEECH TRAINING

Rüdiger Hoffmann

TU Dresden, Institut für Akustik und Sprachkommunikation, 01062 Dresden, Germany
ruediger.hoffmann@tu-dresden.de

ABSTRACT

This year is the 100th anniversary of the death of WILHELM C. RÖNTGEN. With his discovery of the X-rays in 1895, phonetics received a powerful tool for observing the articulation process. The history of the resulting development was described by several authors. Nevertheless, there are some aspects of the early years of this technology and their application in the deaf-mute education, which were never discussed before. This paper has the aim to present such details using the archival material from the former Hamburg Phonetics Laboratory, which is now in the historic acoustic-phonetic collection (HAPS) of TU Dresden, and hitherto unpublished material from several German archives and libraries.

Keywords: X-rays in phonetics; sagittal sections; sound charts; Hamburg Phonetics Laboratory.

1. INTRODUCTION

The description of the positions of the articulators for the speech sounds forms an essential task in phonetics. A sagittal cross section of the head proves to be a very useful representation which was firstly applied by LEONARDO DA VINCI (1452–1519) (Fig. 1 a). Later, FRANCISCUS MERCURIUS VAN HELMONT (1614–1699) used this type of imaging when he tried to prove his idea, that the position of the articulators for the different sounds is coded in the shape of the Hebrew letters (Fig. 1 b). WOLFGANG VON KEMPELEN (1734–1804) used it to justify the construction of his talking machine (Fig. 1 c).

With growing importance of the "deaf-mute education", supporting material for lip-reading and speech training was required. It included views on the lower part of the face, which were at first carried out as drawings [5] [4], later as photographs (ca. 1885, cf. [22, p. 82]). FRANZ HERRMANN CZECH, professor at the deaf-mute institute in Vienna, added sagittal sections as one of the first (Fig. 1 d), and the physiologist ERNST BRÜCKE (1819–1892) used them in his famous book [3] (Fig. 1 e).

The development of experimental phonetics at the end of the 19th century enabled improved representations. OTTO BREMER (1862–1936), the founder of the speech science education at Halle-Wittenberg University, published a set of 15 precise sagittal sections, accompanied by palatograms, which must be counted as a milestone (Fig. 1 f).

Later, such illustrations were manufactured in large size for usage in the classroom, called sound charts. A set of 16 sound charts was designed for the Munich deaf-mute institute. They combined a photograph of a speaker with a sagittal section and a palatogram, performed by the well-known painter and scientific illustrator GUSTAV KELLER (1860–1911) following the design (mentioned in 1906 [10, p. 17]) of the deaf-mute teacher MAX KOLLER. They were published in small size as appendix of a book of another teacher of this institute, FRANZ XAVER RÖTZER, in 1908 [24].

The acquisition of the sagittal sections by means of mechanical tools (including palatography) was complicated to perform. Additionally, the measurements influenced the speakers physically and mentally. Therefore it was an important event when the detection of the so-called X-rays was published by WILHELM CONRAD RÖNTGEN (1845–1923) at the end of 1895. This promised to map the articulators in action in a contactless and easy way.

MICHAEL ASHBY found out that "the earliest surviving X-ray image of any part of the vocal tract was made by the Glasgow laryngologist JOHN MACINTYRE as early as April 1896" [1, p. 103]. The corresponding development in Germany, which will be partially pursued in this paper, started five months later with a paper of the Berlin laryngologist MAX SCHEIER (1864–1921) at the *68th Meeting of German Natural Scientists and Physicians* [25]. His work will be considered in Section 2. In the following, it will be shown how the X-ray technology influenced the development of sound charts for educational purposes (Section 3) and the research activities at the Hamburg Phonetics Laboratory, that was an important place for the development of experimental phonetics (Section 4). Section 5 describes the relics of this historic development, which are preserved in the HAPS of the TU Dresden.

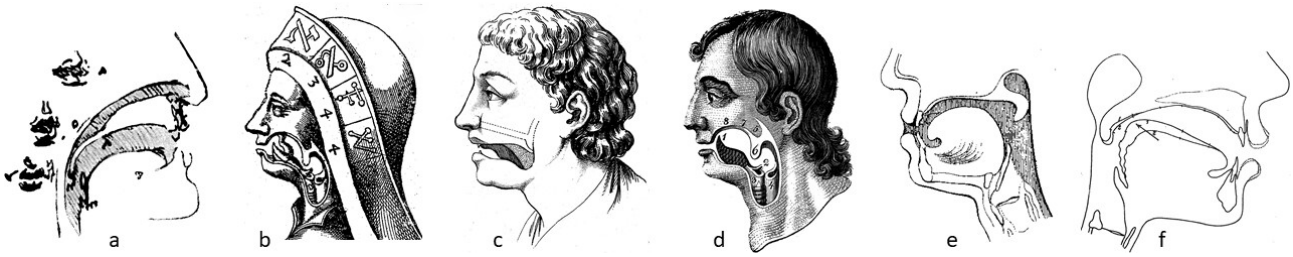


Figure 1: Historic sagittal sections in phonetics and deaf-mute education before X-ray technology was invented. (a) LEONARDO ca. 1500 [21, p. 148]; (b) HELMONT 1667 [6]; (c) VON KEMPELEN 1791 [9]; (d) CZECH 1836 [4]; (e) BRÜCKE 1856 [3]; (f) BREMER 1893 [2].

2. LIFE AND WORK OF MAX SCHEIER

Although the work of MAX SCHEIER is widely cited, little is known about his biography. The following data are based mainly on the habilitation file [38] and the Berlin address books [31]. He was born in Jastrow (West Prussia) on December 4, 1864, and studied medicine at the universities of Greifswald, Berlin, Würzburg, and again Berlin, where he received his doctorate in 1889. He specialized in rhinolaryngology in Berlin and Frankfurt.

Since 1890, he had a medical practice in Berlin which was specialized for diseases of neck and nose in 1893. Additionally, he founded an outpatient department for diseases of neck and nose at the Berlin Johanneum in 1898.

SCHEIER was active in the Berlin Laryngological Society and published a lot. His own list of publications from the beginning until 1902 includes 37 entries. One of his main topics was the ossification of the larynx. He gave courses in rhinolaryngology for practitioners since 1899. He was elected as a member of the French society of otorhinolaryngology. Due to this successful carrier, he applied for his habilitation at Berlin University in 1902. However, the assessments of the reviewers were negative. It seems that this setback caused SCHEIER to change his focus. Apart from a number of summarizing papers in 1908/09, few publications can be found since then. His medical specialization was extended by otology in 1902. His office moved from Friedrichstraße to the noble area of Tauentzienstraße (1910) and Kurfürstendamm (1915). The (hitherto unpublished) date of his death is July 3, 1921 [34]; he was buried at the Jewish cemetery in Berlin-Weißensee.

Regarding the contents of his papers since 1896, a considerable part deals with the application of X-rays to rhinolaryngology, but only few are immediately directed to the requirements of speaking and singing. They include 7 papers in the years 1897/98 and 3 papers in 1909 and are discussed in more detail in several reviews (e.g. [11], [29]).

SCHEIER published his observation that X-rays can be useful to show the positions of the articulators as a rather short paper at the *15th Congress of Internal Medicine* in June 1897 [26]. A more elaborated version was published in a phonetic journal [27]. The results based on the observation of the fluorescent X-ray screen and were explained verbally. X-ray photographs still needed an exposure time of several minutes which was unsuited for phonetic experiments. When SCHEIER delivered an updated version ten years later [28], he also added photographs because the exposure time was then reduced to several seconds due to the technical progress.

3. F. RAUSCH AND HIS SOUND CHARTS

The improved knowledge about the articulation of sounds raised the quality of the teaching material for deaf-mute institutes and similar applications. The sound charts of the deaf-mute teacher FRIEDRICH RAUSCH (1871–1956), the first edition of which appeared in 1903, must be mentioned because a reference is made in the accompanying booklet, that the sagittal sections are based “on the X-ray photographic observations of Dr. SCHEIER on the articulators” [23].

The biography of RAUSCH was investigated for this paper, but due to the limited space, it will be published in more detail at another place. The question arises in the present context, where he acquired his special phonetic knowledge. It is known from a short curriculum [36], that he studied at Halle-Wittenberg University “in his leisure time”, while he worked as a teacher at the deaf-mute institute in Halle 1894–1896 [37]. Indeed, he was inscribed as a guest student for the study of philosophy at April 20, 1894 [39]. There is no further information, but a connection to the aforementioned OTTO BREMER, who acted as an outside lecturer at the German Seminar in Halle since 1888 [7], could be assumed.

Howsoever the scientific network of RAUSCH

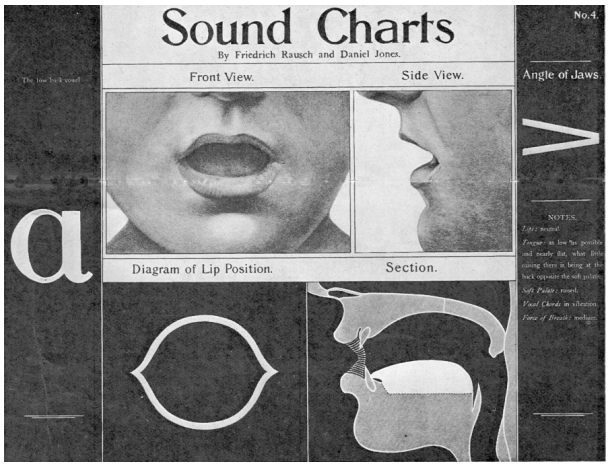


Figure 2: Example for the sound charts (No. 4 of the English version); original size 18 × 23 inches. Reproduced from a flyer of the publisher Dent & Sons in the Dresden HAPS collection.



Figure 3: G. PANCONCELLI-CALZIA with his co-workers doing X-ray inspections of the articulation organs. Photograph no. 1501 from the Hamburg negative archive, now in the HAPS.

might have been, when he came as a teacher to Nordhausen (Thuringia, Germany) in 1900, he started activities in deaf-mute education and in teaching aids manufacturing. The first edition of his 20 big-sized sound charts was elaborated in cooperation with the renowned photographer CARL SCHIEWEK (1871–1934) and appeared in Nordhausen with a print run of 2.000 copies in 1903 [23]. It seems that there is only one complete copy left in public institutions, namely in the library of Halle-Wittenberg University. An English version of nine charts was prepared in cooperation with the British phonetician DANIEL JONES (1881–1967), who was with the University College London since 1907, and appeared in London as a volume of "Dent's Modern Language Series" in 1911 (Fig. 2). A specimen copy is listed in the catalogue of British Library.

In 1911, the publisher was changed to N. G. Elwert in Marburg [35]. The charts were redesigned in portrait size and extended by six charts for French and English sounds. Elwert published 18,000 of the big-sized charts in two editions (1911 and 1926), which were still available after WW II. A complete set of the 2nd edition is preserved in the HAPS. Additionally, a "handy" version in postcard size was available in several editions.

4. EARLY X-RAY STUDIES IN HAMBURG

4.1. The Hamburg Phonetics Laboratory

After some advances in X-ray technology, some phoneticians felt encouraged to apply it for own investigations, thus starting a process which intensified over the years. As a prominent example, ERNST A. MEYER (1873–1953) should be mentioned who tried rather early to verify the alleged efficiency of the new method [13] [14]. All researchers in this phase had in common that they made their investigations in X-ray laboratories of medical facilities. There is only one exception: the Hamburg Phonetics Laboratory.

After the Hamburg Colonial Institute was founded in 1908, the Africanist CARL MEINHOF (1857–1944) demanded a Phonetics Laboratory, which was installed in 1910 and directed by GUIDO PANCONCELLI-CALZIA (1878–1966), a postgraduate from Abbé PIERRE-JEAN ROUSSELOT (1846–1924). He developed this laboratory to an excellent place for research and teaching in experimental phonetics, from 1919 as a part of Hamburg University. Already in the third year of existence, the following departments were established [15]: the phonographic center, the cinematographic department, and the X-ray department with own equipment (Fig. 3).

4.2. X-ray investigations and publications

Except from the study of different "colonial languages" [15], an early application case for the brand-new X-ray equipment in Hamburg came from outside. The Hamburg "pedagogue for hygienic speaking and singing" [32] CLARA HOFFMANN cooperated with the Lübeck "special doctor for diseases of the lung, the neck, and the nose" [33] ALFONS OTT with the aim to control the progress in the education process in speaking and singing by X-ray images since ca. 1910. Both came in contact with PANCONCELLI-CALZIA who contributed his phonetic expertise and the laboratory. They jointly published a paper at the *First International Congress of Music Pedagogy* in 1913 [8], which was hitherto ne-



Figure 4: X-ray photograph of the articulators of CLARA HOFFMANN, singing the sound /a/ with pitch a^1. Photograph no. 328 from the Hamburg negative archive, now in the HAPS.

glected by historic reviews [11] [29]. They investigated the singing and the speaking voice of as much as 186 people.

For the *First International Congress of Experimental Phonetics*, held in Hamburg in April 1914, X-ray methods played no prominent role. An overview lecture of MAX SCHEIER was announced, but not presented. In the following WW I, the focus of PANCONCELLI-CALZIA shifted to phoniatric tasks in a military hospital where he also published a note about X-ray techniques [17]. Nevertheless, he still published a detailed paper on the X-ray analysis of the production of certain sounds in African languages [18], and another on X-ray polygrams [19].

4.3. The "secret book" of Panconcelli-Calzia

Anyone who studies the work of PANCONCELLI-CALZIA will sometime discover the citation given here as [16], which prompts him to a monograph on the application of X-rays in phonetics as volume 33 of a series. If he orders this 33rd volume in his library, he receives a monograph on the X-ray investigation of the hand, foot, and elbow. How to explain this?

Indeed, PANCONCELLI-CALZIA wrote the book in 1914. It consists of chapters on phonetic examination techniques (including a description of the Hamburg X-ray laboratory), on the elements of articulation, and on the application in education (by CLARA HOFFMANN). The appendix includes 29 X-ray photographs which were previously unpublished.

Habent sua fata libelli. As in the case of the proceedings of the congress of 1914, WW I prevented the publication of this important book. The publisher used the number of the volume for another title in 1918. Luckily for posterity, a proof with many notes by the author has survived. It is preserved in the State and University Library in Hamburg. The author valued his own work so highly that he included it (although unpublished) 1941 in his survey on the history of phonetics as "the first work of this kind" [20, p. 70].

5. MEMORABILIA IN THE HAPS

When the Hamburg Phonetics Institute was closed in the Winter Term 2006/07, the collection of historic devices was transferred to the historic acoustic-phonetic collection (HAPS) of the TU Dresden along with many documents and audiovisual material, including a collection of approx. 2.000 photographic glass negatives which were digitized in 2018. Among them are ca. 80 early X-ray photographs, of which Fig. 4 shows an example.

Although PANCONCELLI-CALZIA was very interested in the production of educational films, there is only one cinematographic X-ray film left in the HAPS. Is is undated and shows the articulation of several sound sequences. It was digitized to accompany the presentation of this paper.

In the time following WW II, X-ray investigation experienced a revival in Hamburg in the time of PANCONCELLI-CALZIA's successor OTTO VON ESSEN (1898–1983). The phonetician HANS-HEINRICH WÄNGLER (1921–2001) worked there 1954–1964 and published the first edition of his *Atlas of German speech sounds* in 1958, including 29 sound charts with X-ray photographs [30]. Besides other similar material, the HAPS preserves a set of original X-ray photographs from this project, which were produced by the X-ray department of the University ENT clinics in Hamburg.

Even the tradition of the big-sized sound charts was continued in Hamburg in conjunction with the phonetic teaching material of CARL MARTENS (employee of the Phonetics Laboratory since 1913) and his son PETER MARTENS (1919–2012; professor for German phonetics at Hamburg University 1977–1985) [12]. They developed a convolute of 31 scrolled tables (sized 83 cm × 58 cm) which was edited by the publisher Hanex in Hamburg. Apart from this complete set, the HAPS owns 62 tables from cardboard, too, which use the same pictorial material but are manufactured individually.

It can be assumed that the tables by MARTENS were the last classical sound charts which were regularly published. With the development of electronic media, new types of phonetic teaching material were introduced.

6. REFERENCES

[1] Ashby, M. G. 2016. *Experimental phonetics in Britain, 1890–1940*. Oxford: PhD thesis.

[2] Bremer, O. 1893. *Deutsche Phonetik*. Leipzig: Breitkopf & Härtel (Sammlung kurzer Grammatiken deutscher Mundarten; 1).

[3] Brücke, E. 1956. *Grundzüge der Physiologie und Systematik der Sprachlaute*. Wien: Gerold.

[4] Czech, F. H. 1836. *Versinnlichte Denk- und Sprachlehre mit Anwendung auf die Religions- und Sittenlehre und auf das Leben*. Wien: Mechitaristen.

[5] Graser, J. B. 1829. *Der durch Gesicht- und Tonsprache der Menschheit wiedergegebene Taubstumme*. Bayreuth: Grau.

[6] Helmont, F. M. v. 1667. *Kurtzer Entwurff des eigentlichen Natur-Alphabets der heiligen Sprache*. Sultzbach: Lichtenthaler. – Reprint ed. by Wilhelm Viëtor as addendum to *VOX* vols. 26 & 27.

[7] Hirschfeld, U., Solms, H.-J., Stock, E. (eds.) 2016. *Otto Bremer – Wegbereiter der sprechwissenschaftlichen Phonetik an der Universität Halle*. Berlin: Frank & Timme.

[8] Hoffmann, C., Ott, A., Panconcelli-Calzia, G. 1913. Wissenschaft und Gesangspädagogik, Röntgenstrahlen und praktische Stimmbildung. *Vorträge u. Referate, 1. Internat. Musikpädagog. Kongress, Berlin 1913*, 54–66.

[9] Kempelen, W. v. 1791. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. Wien: Degen. – Commented Transliteration & Translation into English by F. Brackhane, R. Sproat & J. Trouvain. 2 Parts, Dresden: TUDpress 2017 (Studientexte zur Sprachkommunikation; 87/88).

[10] Königliches Zentral-Taubstummen-Institut. 1906. *Jahresbericht 1905/06*. München.

[11] Macmillan, A. S., Kelemen, G. 1952. Radiography of the supraglottic speech organs – A survey. *AMA Archives of Otolaryngology* 55, no. 6, 671–688.

[12] Martens, C. und P. 1961. *Phonetik der deutschen Sprache*. München: Hueber.

[13] Meyer, E. A. 1907. Röntgenographische Lautbilder. *Medizinisch-pädagogische Monatsschr. für die ges. Sprachheilkde.* 17, no. 8/9, 225–243.

[14] Meyer, E. A. 1910. Untersuchungen über Lautbildung. *Festschrift Wilhelm Viëtor zum 25. Dezember 1910*. Marburg: Elwert, 166–248.

[15] Panconcelli-Calzia, G. 1913. Mitteilung über das dritte Arbeitsjahr des Phonetischen Laboratoriums des Seminars für Kolonialsprachen zu Hamburg. *VOX* 23, no. 6, 313–318.

[16] Panconcelli-Calzia, G. 1914. *Die phonischen Bewegungen des Menschen im Röntgenbild.* Hamburg: Gräfe & Sillem (Fortschritte auf dem Gebiet der Röntgenstrahlen, Ergänzungsbände = Archiv und Atlas der normalen und pathologischen Anatomie in typischen Röntgenbildern, 33).

[17] Panconcelli-Calzia, G. 1916/17. Experimentelle Versuche zur Erweiterung des Müllerschen Verfahrens zur Fremdkörperlokalisation. *Fortschritte auf dem Gebiete der Röntgenstrahlen* 24, no. 2, 123–125.

[18] Panconcelli-Calzia, G. 1917/18. Untersuchungen über ɬ ɺ ɓ ɠ ǃ ǂ ǁ mit Röntgenstrahlen. *Zeitschrift für Kolonialsprachen* 9, no. 1, 20–25.

[19] Panconcelli-Calzia, G. 1919. Phonetische Röntgenpolygramme. *VOX* 29, no. 5, 170–179.

[20] Panconcelli-Calzia, G. 1941. *Geschichtszahlen der Phonetik. 3000 Jahre Phonetik*. Hamburg: Hansischer Gildenverlag.

[21] Panconcelli-Calzia, G. 1943. *Leonardo als Phonetiker*. Hamburg: Hansischer Gildenverlag.

[22] Panconcelli-Calzia, G. 1961. *3000 Jahre Stimmforschung*. Marburg: Elwert.

[23] Rausch, F. 1903. *Laut-Tafeln zum Gebrauch beim deutschen und fremdsprachlichen Sprech-Schreib-Leseunterricht nach phonetischen Grundsätzen*. Nordhausen: Schiewek.

[24] Rötzer, F. X. 1908. *Übungsbuch für Schwerhörige und Ertaubte*. München/Berlin: Oldenbourg.

[25] Scheier, M. 1896. Über die Photographie der Nase und des Kehlkopfes mittelst Röntgenstrahlen. *Verhandl. d. Ges. deutscher Naturforscher u. Ärzte, 68. Vers., Frankfurt 1896*, 2. Theil, 2. Hälfte, 416–420.

[26] Scheier, M. 1897. Die Anwendung der X-Strahlen für die Physiologie der Stimme und Sprache. *Verhandl. d. Congr. f. Innere Medizin, Berlin 1897*, Wiesbaden: Bergmann, 416–420.

[27] Scheier, M. 1897/98. Über die Bedeutung der Röntgenstrahlen für die Physiologie der Sprache und Stimme. *Phonetische Studien* 11 = *Die neueren Sprachen* 5 (Supplement), 40–56.

[28] Scheier, M. 1909. Die Berdeutung des Röntgenverfahrens für die Physiologie der Sprache und Stimme. *Archiv für Laryngologie und Rhinologie* 22 no. 2, 175–208 plus 9 tables.

[29] Simon, P. 1967. *Les Consonnes Françaises. Mouvements et positions articulatoires a la lumière de la radiocinématographie*. Paris: Klincksieck.

[30] Wängler, H.-H. 1964. *Atlas deutscher Sprachlaute*. Berlin: Akademie-Verlag, 3rd ed.

[31] Adressbuch Berlin, URL (17.10.2022): digital.zlb.de/viewer/berliner-adressbuecher/

[32] Adressbuch Hamburg, URL (31.03.2022): agora.sub.uni-hamburg.de/subhh-adress/digbib/start

[33] Adressbuch Lübeck, URL (25.10.1922): digital-stadtbibliothek.luebeck.de/viewer/toc/L13031/

[34] LA Berlin, Bestand: P Rep. 551 (Charlottenburg I), Nr. 342, SU 565/1921.

[35] LA Hessen, StA Marburg, Bestand: 334/2, Nr. 286 (Herstell. u. Vertrieb der "Laut-Tafeln für Sprech-, Schreib- und Leseunterricht" von F. Rausch).

[36] LA Thüringen, HStA Weimar, Bestand: Archivberatungsstelle der Provinz Sachsen, Nr. 7.

[37] LA Thüringen, StA Gotha, Bestand: Regierung Erfurt, Nr. 29627, Personalakten F. Rausch.

[38] UA HU Berlin, Med. Fak. 01, Nr. 1348 (Habilitationen 1901–1902), Bl. 196.

[39] UA MLU Halle-Wittenberg, Rep. 4 (Rektorat der Universität Halle-Wittenberg von 1817 bis 1948), Nr. 2149 (Gasthörer Bd. 3: 1890–1904).

BENT VS. STRAIGHT MODELS OF THE HUMAN VOCAL TRACT BASED ON MEASUREMENTS BY CHIBA AND KAJIYAMA

Takayuki Arai

Sophia University
arai@sophia.ac.jp

ABSTRACT

Chiba and Kajiyama measured three-dimensional configurations of the human vocal tract during the production of five Japanese vowels and published the resultant data in "The Vowel: Its Nature and Structure." The data were estimated by palatograms as well as X-ray images and were naturally bent. They also derived cross-sectional area functions for the areas perpendicular to the midline along the length of the vocal tract. The overall area functions were simplified by interpolating them in a piecewise-linear fashion.

In this study, we first compared replicated physical models of the original bent configurations with approximated straight models based on the same area functions. We then created bent models with further simplifications and exhibited them at a museum. These models have pedagogical usefulness because they are simply shaped but also bent, which makes it easy to imagine where they are located in the head.

Keywords: Chiba and Kajiyama, The Vowel, physical models, vocal tract, speech production

1. INTRODUCTION

Chiba and Kajiyama's 1941–42 book "The Vowel: Its Nature and Structure" [1] introduced several fundamental topics on vowel production and perception that paved the way to modern speech science [2–6]. One key issue pointed out in the book is how vocal-tract configurations determine vowel qualities, and to examine it, the authors measured 3D configurations of the vocal tract during the production of five Japanese vowels. These measurements were performed using the state-of-the-art X-ray apparatus at that time (Shimazu Co.), palatography (in which a set of shapes is molded out of plaster), and laryngoscopy (for observing the pharyngeal cavity). By combining these three techniques, they were able to obtain detailed 3D configurations of the vocal tract.

Chiba and Kajiyama derived cross-sectional area functions based on the 3D measurements by 1) drawing the midline of each vocal-tract configuration of each vowel and 2) estimating the cross-sectional areas of points on the midline where the cross-sections are perpendicular to it. They further simplified the area functions and created physical models in round-bottle shapes. After feeding in a sound source, they found that the physical models based on the area functions were close enough to their naturally produced counterparts. Arai [7] replicated their physical models not only for historical but also pedagogical purposes.

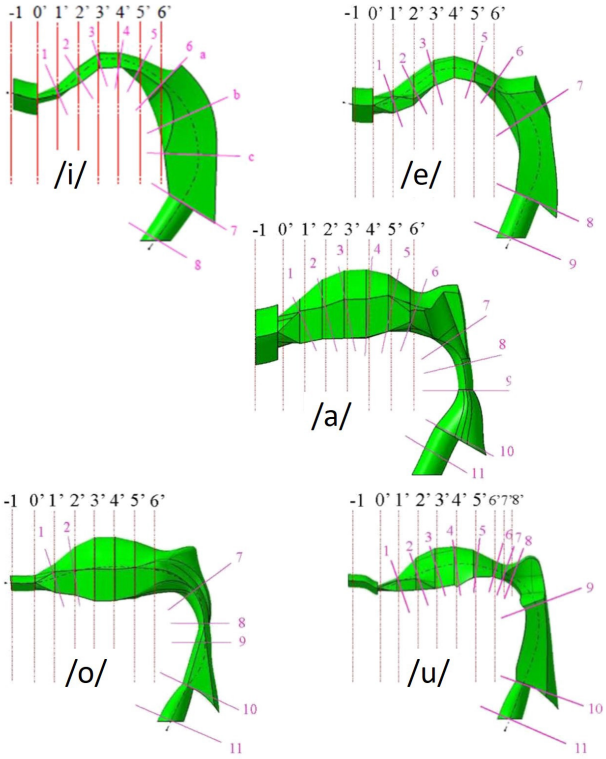


Figure 1: Midsagittal view of five Japanese vowels based on measurements in [1].

One of the options they might have selected was to model vocal tracts by keeping the bent shapes. Therefore, in this study, we compared bent vs. straight models with the same area functions. We also further simplified the configurations for bent models and displayed them at a museum exhibition.

2. PHYSICAL MODELS

Figure 1 shows the midsagittal view of the vocal-tract configurations of five Japanese vowels based on the measurements in [1]. The vertical lines in each plot show the sections (numbered with the prime symbol)

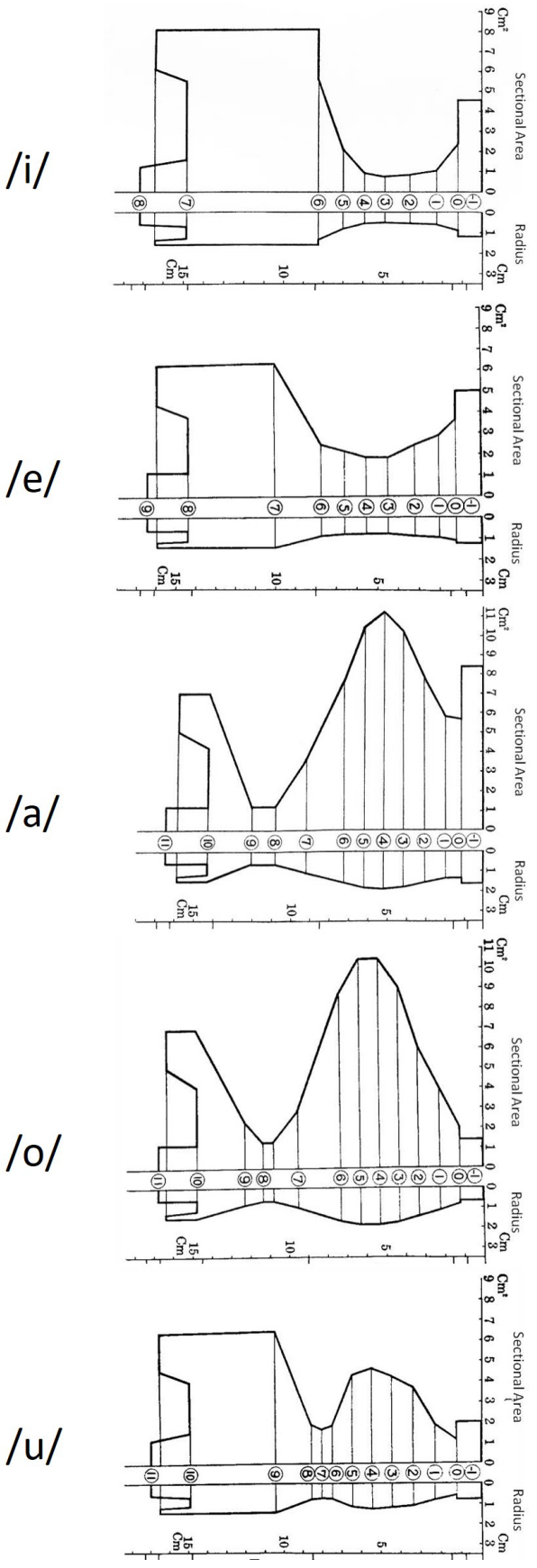


Figure 2: Cross-sectional area functions of five Japanese vowels adapted from [1].



Figure 3: Replicated physical models of five Japanese vowels (from left: /i/, /e/, /a/, /o/, and /u/) [7].

that were originally measured by palatograms. At the same point of the midline in each plot, additional lines show the sections (numbered with no prime symbol) that are perpendicular to the midline. In Fig. 2, the area above the horizontal axis in each plot shows the cross-sectional area function based on the perpendicular lines in Fig. 1, where the numbers in both figures correspond to each other. These area functions were drawn by piecewise linear approximation. The area below the horizontal axis in each plot in Fig. 2 shows the radius function when the cross-sections were modeled as circles.

Using the radius function as a basis, Chiba and Kajiyama created physical models in the "Artificial Vowels" section in [1], and after feeding in a sound source, they showed that artificially produced vowels were close enough to the naturally produced counterparts. However, the original models no longer exist (presumably due to wartime disruption), so the replication of the physical models created by Arai [7] is shown in Fig. 3. Arai [8] tried to model vocal tracts of vowels /i/ and /a/ by keeping the bent configurations based on Chiba and Kajiyama's original measurements. In [8], there were two major design constraints for the bent models: 1) the midsagittal cross-section must be the same as the one in Fig. 1, and 2) the cross-sectional area function must be the same as the one in Fig. 2. Although Chiba and Kajiyama drew outlines of the uvula, epiglottis, and piriform fossa (as shown in Fig. 1), they were simplified in Fig. 2, so the designs in [8] omitted them. The designs for vowels /i/ and /a/ were successful, but there were no descriptions for the rest of the vowels (i.e., /e/, /o/, and /u/), and no acoustic analysis was ever reported.

This is why, in this study, we expand the design of these bent models so that all five vowels can be implemented. We followed the same design constraints as above and then 3D-printed the bent models, as shown in Fig. 4.

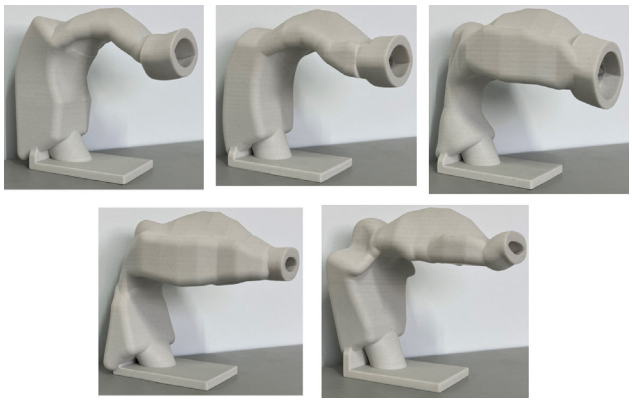


Figure 4: 3D-printed bent models of five Japanese vowels based on measurements in [1].

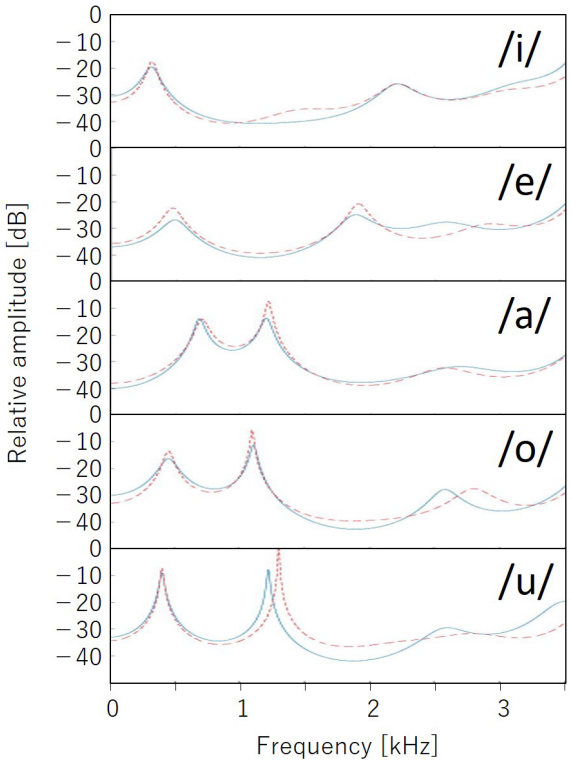


Figure 5: LPC-based spectral envelopes of bent vs. straight models for five Japanese vowels (LPC order: 10).

3. ACOUSTIC ANALYSIS

In this section, we compare the frequency characteristics of the bent vs. straight models.

3.1. Recordings

We wanted to subjectively judge the vowel quality of each output sound, so the recordings were done by feeding in a simple impulse train from a deriver unit of a horn speaker. We digitally recorded in a quiet laboratory at Sophia University and set the original sampling frequency to 44.1 kHz and the quantization to 16 bits. A microphone (Rode, NT6) were placed approximately 150 mm away from the mouth opening.

3.2. Analysis method

We utilized Matlab to analyze each spectrum during vowel production by the linear predictive coding (LPC) method. First, the speech signals were downsampled to 8 kHz, and then the LPC-based spectral envelope was calculated for each utterance.

3.3. Results

Figure 5 shows the comparison between the bent (solid lines) and straight (dashed lines) models for each vowel. As we can see, the first two formant frequencies (F1 and F2) were almost identical between the bent and straight models for each vowel, although there was a slight difference between the two models in high-frequency regions above 2 kHz.

3.4. Discussion

Our findings suggest that Chiba and Kajiyama presumably already knew that the bent and straight models have almost the same frequency characteristics as long as they have the cross-sectional area function in common, where the cross-dimensions are taken perpendicularly against the midline along the vocal-tract length. This is especially evident in lower-frequency regions below 2 kHz. The results in Fig. 5 prove that Chiba and Kajiyama were correct in their original methods. Fant [9] and other researchers have followed the same methodology [2]. The reason for the difference in frequency characteristics might not be caused by the issue of straight vs. bent tubes. When Chiba and Kajiyama converted a bent tube to a straight tube, they also changed the shape of each cross-section to that of a circle. As a result, this might have yielded such differences.

4. PEDAGOGICAL APPLICATIONS

Various previous studies [e.g., 7, 10–12] have demonstrated the usefulness of physical models of the human vocal tract. Here, we introduce yet another set of bent models with extra simplifications.

4.1. Simplification

As discussed in Section 2, there are two design constraints for the bent models in Fig. 5. However, in in Chiba and Kajiyama's work [1], coronal sections were only available in the oral cavities, and not in the pharyngeal cavities. Therefore, some parts of the configurations in Fig. 4 are not realistic and might lead to misunderstandings if the models are applied as educational tools. Therefore, we further simplified the configurations so that the inner cavities (except the hypopharynx and the lips) are approximated by simple box shapes, instead.

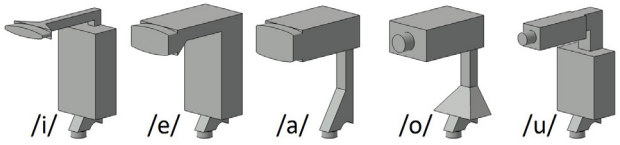


Figure 6: Simplified inner cavities of five Japanese vowels.

The resultant vocal-tract configurations are shown in Fig. 6, where only the inner cavities (i.e., not the outer shapes) for all five vowels are depicted. As we can see, for vowels /i/ and /e/, the pharyngeal cavities are wide open, the oral cavities are narrow, and the lips are spread. For vowels /a/ and /o/, on the other hand, the oral cavities are wide open and the pharyngeal cavities are narrow. For vowel /u/, the major parts of the oral and pharyngeal cavities are relatively wide, and the corner part (the middle part of the vocal tract) is narrow. For vowels /o/ and /u/, the lips are rounded, so each model has a narrow cylinder shape at the mouth end.

4.2. Hypopharynx

When designing the hypopharynx in Fig. 6, there were two options we could have taken. The first was to design the so-called "larynx tube" [13], as shown in Fig. 7(a), which is closer to Chiba and Kajiyama's original measurement. The second option was to design it as a narrow and straight tube, as shown in Fig. 7(b).

We conducted an acoustic analysis between sounds for the two cases in Fig. 7. The results (shown in Fig. 8) indicate a difference around the 3-kHz region in the spectra as well as around the F1 and F2 frequencies. The F1 and F2 frequencies are both within vowel /a/ in Japanese, whereas the frequency range of 2.5–3.5 kHz matches the one for the "singer's formant" in [13]. The case in Fig. 7(a) is similar to the situation where the larynx is lowered [13], which is what caused the higher peak level in this frequency range; the difference was approximately 6.5 dB (Fig. 8). The bandwidths of F1 and F2 for the model in Fig. 7(b) were wider, which might suggest a breathier voice quality.

In the case of vowel /o/, the hypopharynx is much wider than vowel /a/. In addition, the pharyngeal cavity above that section should be sufficiently narrow. These points are reflected in the design of /o/ in Fig. 6, and result in a better sounding quality for /o/.

4.3. Museum exhibition

We 3D-printed the models (Fig. 9) discussed in this work and temporarily exhibited them at the National Museum of Ethnology, Japan in 2022 [14]. Previous studies [10–12] have featured mainly straight tubes or head-shaped models. These newly designed models (VTM-B50) have an additional pedagogical usefulness because they are simply shaped but also bent, which makes it easy to imagine where they are located in the head.

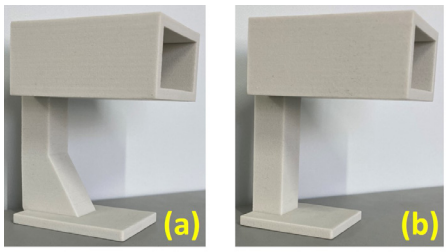


Figure 7: Two versions of simplified bent models of vowel /a/.

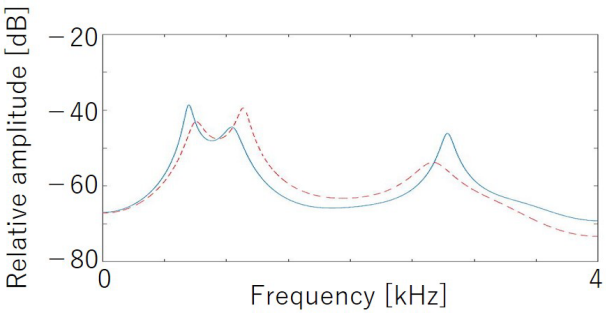


Figure 8: LPC-based spectral envelopes of the two models in Fig. 7 (solid: model in Fig. 7(a); dashed: model in Fig. 7(b)). F1 and F2 frequencies were approximately 693 and 1050 Hz for the Fig. 7(a) model and 761 and 1135 Hz for the Fig. 7(b) model.

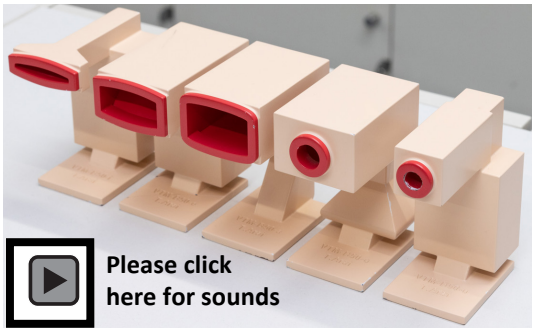


Figure 9: 3D-printed and painted version of VTM-B50 models. A video clip is available here [15].

5. SUMMARY

In this study, we examined bent and straight models of the human vocal tract for five Japanese vowels based on the measurements by Chiba and Kajiyama [1]. The lower frequency characteristics between the two were similar, thus proving that Chiba and Kajiyama conducted the analyses correctly the first time. We further simplified the design process for pedagogical purposes and exhibited the models at a museum.

6. ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Numbers 21K02889 and Sophia University Special Grant for Academic Research (Research in Priority Areas).

7. REFERENCES

[1] Chiba, T. and Kajiyama, M. 1941–42. *The Vowel: Its Nature and Structure*. Tokyo-Kaiseikan.
[2] Maekawa, K. and Honda, K. 2001. On the Vowel, Its Nature and Structure and related works by Chiba and Kajiyama. *J. Phonetic Soc. Jpn.* 5(2), 15–30.
[3] Arai, T. 2004. History of Chiba and Kajiyama and their influence in modern speech science. *From Sound to Sense*. MIT.
[4] Maekawa, K. 2002. From articulatory phonetics to the physics of speech: Contribution of Chiba and Kajiyama. *Acoust. Sci. Tech.* 23(4), 185–188.
[5] Stevens, K. N. 2001. The Chiba and Kajiyama book as a precursor to the acoustic theory of speech production. *J. Phonetic Soc. Jpn.* 5(2), 6–7.
[6] Fant, G. 2001. T. Chiba and M. Kajiyama, pioneers in speech acoustics. *J. Phonetic Soc. Jpn.* 5(2), 4–5.
[7] Arai, T. 2001. The replication of Chiba and Kajiyama's mechanical models of the human vocal cavity. *J. Phonetic Soc. Jpn.* 5(2), 31–38.
[8] Arai, T. 2017. Chiba and Kajiyama measured 3D configurations of the human vocal tract and made physical models based on the measurements. *Proc. Second International Workshop on the History of Speech Communication Research (HSCR 2017)*. 55–60.
[9] Fant, G. 1960. *Acoustic Theory of Speech Production*. Mouton.
[10] Arai, T. 2007. Education system in acoustics of speech production using physical models of the human vocal tract. *Acoust. Sci. Tech.* 28(3), 190–201.
[11] Arai, T. 2012. Education in acoustics and speech science using vocal-tract models. *J. Acoust. Soc. Am.* 131(3), Pt. 2, 2444–2454.
[12] Arai, T. 2016. Vocal-tract models and their applications in education for intuitive understanding of speech production. *Acoust. Sci. Tech.* 37(4), 148–156.
[13] Sundberg, J. 1974. Articulatory interpretation of the "singing formant," *J. Acoust. Soc. Am.* 55(4), 838–844.
[14] https://loquens.site/
[15] https://youtu.be/aLyqaafEM38

THE FIRST X-RAY SOUND FILM (1935): ITS ORIGINS, IMPACT AND RESTORATION

Michael Ashby[1], Angelika Braun[2]

[1]University College London (UCL), UK, [2]University of Trier, Germany
m.ashby@ucl.ac.uk, *brauna@uni-trier.de*

ABSTRACT

Cineradiography was developed in the 1920s as a medical diagnostic tool for examining movement inside the body. The earliest applications to speech produced only silent films with low frame rates. The recording of synchronised sound required improvement of the filming speed to 24 fps, a problem solved by Bonn radiologist Robert Janker, who cooperated with Paul Menzerath, then head of the phonetics laboratory at the University of Bonn, to produce a sound X-ray film first screened at the second ICPhS in London in 1935. The film created a sensation and established a paradigm for research in speech production for the next 30 years. Digital restoration of two surviving incomplete prints of the film preserved in London permits reconstruction of the original 1935 version, including the first ever X-ray slow-motion sequence of speech movements.

Keywords cineradiography, speech production, coarticulation, Robert Janker, Paul Menzerath

1. INTRODUCTION

Röntgen announced the discovery of X-rays in 1895 and the potential application to speech research was quickly pointed out [1]. Still images could be used to investigate the articulation of sustained sounds such as steady vowels but although a moving X-ray (of a joint in a frog's leg) was demonstrated by Macintyre as early as 1897 [2], the capturing of speech movements required filming from the dim and indistinct fluoroscope screen. By the late 1920s improvements in cameras, lenses, film emulsions, X-ray tubes and fluorescent screens permitted the making of short silent films within radiation dose limits then thought safe. In Bonn in 1930 the younger Hermann Gutzmann (1892–1972) presented a silent X-ray film of speech [3], but identified as shortcomings the low frame rate (16 fps), and the potentially harmful exposure levels. A sound film would be ideal, he said.[1]

The 'ideal' was realised at the second ICPhS in London in 1935 when Paul Menzerath (1883–1954) presented the first X-ray sound film to an astonished audience.

As far as we can determine and reconstruct, what the audience saw was a film a little over 3 minutes in total duration. Of this, the first 1 minute 13 seconds is occupied by lateral X-rays of a succession of speakers seen from the left side. They are seen and heard to speak successive sentences of a rather prosaic 50-word passage which explains what a sound X-ray achieves and points out that it is particularly applicable to the study of speech production but also to the recording of heart sounds and movements. The synchronously recorded sound is nicely demonstrated when one of the speakers stumbles over and has to repeat the word *Zungengrundes* within the phrase *des Zungenrückens, und Zungengrundes* ('of the back of the tongue and of the root of the tongue'). At approximately 01:13 a caption explains that without sound the filming rate can be raised to 48 fps. The remainder of the film is in slow motion, with further captions identifying the spoken material as various isolated vowels and consonants and the numbers one to ten. The film was shown in conjunction with Menzerath's paper to the Congress [4] (though the film is not itself the central focus of the paper) and additionally screened in demonstrations attached to the Congress and in a larger theatre nearby.

2. THE MAKERS OF THE FILM. MENZERATH AND JANKER

At the time of the film's production, Paul Menzerath was the head of the laboratory in the phonetics institute at the University of Bonn, considered one of the best-equipped laboratories world-wide. Born in Düren, Germany, in 1883. he was a psychologist by training but developed an interest in experimental phonetics and was put in charge of the phonetics laboratory at the University of Bonn when it was established in 1921. Menzerath gained wide notice for a ground-breaking monograph written with his Portuguese colleague Armando de Lacerda [5] which is still cited as the source of the concept 'coarticulation', and he also gives his name to Menzerath's Law, which states that the more elements in a chain, the shorter the individual elements are (for instance, the more syllables in a word, the shorter the individual syllables). It happened that one of Menzerath's colleagues at the medical school of the same university was a world-renowned specialist on cineradiography: Robert Janker (1894–1964).

Janker had to abandon a career as a surgeon in Munich after developing a wash eczema, and turned instead to radiology. He went to Bonn in 1928 and received his *Habilitation* from that university in 1930, the topic being X-ray cinematography, a field in which he continued to work, making prolific contributions and gaining an international reputation. Originally, he experimented with X-ray films of the lungs and the heart, adding speech to the range of subjects at a later stage.

3. IMPACT

Brief as it was, the film was widely seen as one of the highlights of ICPhS. In a report on the Congress for the journal *Language* [6], the American linguist Hans Kurath (1891–1992) wrote 'In the field of experimental phonetics the outstanding contribution was the X-ray sound film by P. Menzerath …Movements of the larynx, the soft palate, the tongue, the jaw and the lips are all visible in this remarkable film'.

In a similar review for the journal *English Studies* [7], the British linguist J. R. Firth (1890–1960) singled out 'Professor Menzerath's X-ray film' as 'of outstanding interest'.[2]

Perhaps the most enthusiastic account is that given by Louise Kaiser (1891–1973) in a report for the Bulletin of the International Society of Experimental Phonetics [8: 134]. She wrote 'The impression is as overpowering as that of looking through a microscope for the first time. The world that is revealed is quite a new one. Speech is seen to consist physiologically of a continuous complexity of almost countless movements. There are no "speech positions", no "on-glides" and "off-glides" or any of the other familiar clichés of non-experimental phonetics. The entire voice physiology of the present day is swept away at the first glance'.

She foresaw that cineradiography could provide a new paradigm for speech research: 'The individual pictures—"stills"—afford opportunities for study and measurement that can furnish a basis for a new voice physiology that will meet the demands of modern science'. This was prescient to the extent that cineradiographic speech research was pursued in many centres around the world until around 1970.[3] The fourth ICPhS in Helsinki in 1961 [9] saw five major contributions based on cineradiography, and the quantitative approach foreseen by Louise Kaiser reached a peak in the work of Joseph Perkell at MIT [10]. She was over-optimistic, however, in the hope that still frames from the 1935 film might be used this way. Although *movements* of the larynx, soft palate and tongue are clearly seen when the film is projected, the structures themselves are barely discernible in individual frames. The resolution falls far short of what was achieved in still X-ray images of the day (see Figure 1).

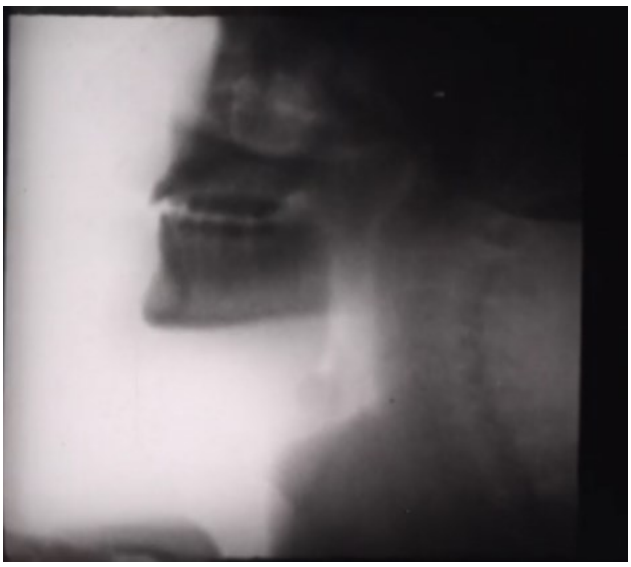


Figure 1: A typical single frame from the 1935 film, illustrating the relatively low resolution achieved.

4. MENZERATH'S VIEW OF THE FILM

Menzerath showed the film on at least two further occasions: at the 4th International Congress of Linguists in Copenhagen in 1936 [11] and at the XIe Congrès International de Psychologie in Paris in 1937 [12]. Neither in 1935 nor in these later publications did Menzerath attempt a systematic—let alone quantitative—analysis of the film. It seems that he was satisfied with using the film in a general way to support his views on coarticulation and did not make any effort to develop an experimental routine for applying cineradiography in phonetic research. His own estimation of the value of the film is certainly more reserved than that of Louise Kaiser. Remarks in [11] show that he realizes that the limitations in frame rate and the poor definition on individual frames can warrant only very guarded conclusions about the 'continuity' of speech movements, and that the relevant comparison is of average segment production rate with the achieved frame rate.

In all of these publications, Menzerath is named as the sole author. Menzerath concludes his 1935 paper with an acknowledgment to radiologist Robert Janker and [11] does make somewhat more explicit the cooperation with Janker, but the precise nature of their respective contributions is nowhere specified. Menzerath [4] refers to Janker as his *Mitarbeiter* (= someone working under his guidance), although all the technical innovations which made the film possible were certainly Janker's.

It is hard to discern what Menzerath's contribution may have been. He held old-fashioned and prescriptive views on German pronunciation [13], and the speakers heard in the 1935 film (who sound like Bonn locals) can hardly have been selected by him. Similarly, the speech materials recorded show no obvious sign of being designed by a phonetician: the use in the slow-motion section of five vowels *a e i o u* in alphabetical order, the consonants *l m n r s t f w b p h* in unspecified contexts and the spoken numbers 1 to 10 rather suggests the opposite.

5. JANKER'S ACCOUNTS OF THE FILM

We have two documents authored by Robert Janker which shed light on the production of the film. The first [14] is a typeset document containing explanatory notes to the film C 150, *Röntgentonfilm der Sprache*. C 150 was the number assigned by the *Reichsstelle für den Unterrichtsfilm* (RfdU), founded in 1934, which released and distributed the film in 1937, and the explanatory notes by Janker were originally published along with the film. The date of publication of this document has led to the film itself being dated to 1937, although we take C 150 to be the film as screened in 1935. It is uncertain whether the number C 150 was already assigned to the film in 1935.

The notes [14] explain that while cineradiography as such was not new, the combination of picture and sound presented a major problem. The method employed for the film was developed as a response to a contest which was advertised by the Zeitler foundation in Berlin in 1932 for the development of an X-ray film that would capture picture and sound of processes inside the body. Janker won the contest with a film using disks to record the sound. The 1935 film evidently represents the third generation of his steadily-improved sound-film attempts. The document includes the text which is heard spoken in the film, and an account of the 48 fps slow-motion section.

6. SOUND ON DISK

It is important to realize that by 1935 disks were used to *capture* the sound clips but they were dubbed to a conventional optical film soundtrack for projection. The advantage of using disks arose from the peculiar circumstances of working in the X-ray laboratory: radiation exposure limits meant each speaker appeared in a separate take lasting less than 10 seconds, and it would have been cumbersome and wasteful to attempt direct optical recording to film.

The fact that the individual takes are very short meant that synchronisation between the separately recorded sound and picture within each short section would not be a major problem, provided only that in the editing process the picture and sound could be correctly aligned at the start. No doubt some sort of clapper-board system was used, but all evidence of the method used has been edited out of the resulting film. Assembling the edited film must have been a somewhat laborious process.

As indicated in [15], the direct-to-disk recording system used by Janker was one manufactured by the Dralowid Company in Berlin (which produced and marketed all manner of electrical and optical components). The blank disks, sold under the name Draloston, were coated with a resin which could be hardened and made durable after recording by baking in a special oven called the Draloston Duotherm.

7. THE MAKING OF THE FILM

A second document by Janker [15], dated 30 April 1963, is a typewritten transcript of the soundtrack for a 34-minute film *Entwicklung und Stand der Roentgenkinotechnik* (R 598) [16] made by Janker. That film presents a systematic and detailed history of the development of cineradiography, and of Janker's role in that development. It seems possible that Janker himself may be the speaker of the voice-over; certainly no other speaker is credited.



Figure 2: The recording setup used for the 1935 film. The light rectangular panel is the fluoroscope screen and the camera (not shown) is located to the right. Notice the large microphone mounted at the top right.

X-ray sound films of speech are dealt with at two places in the film R 598. Around 06:44 a photograph shows the 1930s filming setup (Figure 2) and at 08:25 begins a section just over 41 seconds in duration reproducing examples of the results.

Around 23:14 is an illustration of the more advanced and more comfortable recording setup used in the postwar period (Figure 3). By now Janker was using a 13 cm image intensifier manufactured by Philips. An example recorded in 1956 by a female subject begins at about 24:06 (Figure 4). The reduced radiation exposure resulting from use of the image intensifier permits the same speaker to remain in shot for the whole take of 1 minute 20 seconds.

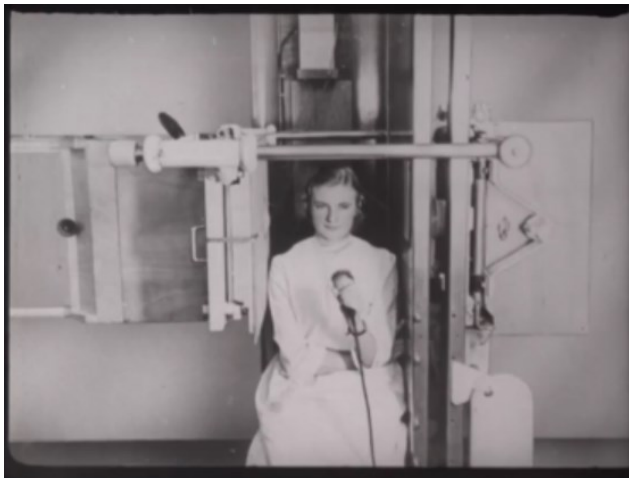


Figure 3: Recording setup of the postwar period, using a handheld microphone. The image intensifier and camera are to the right.

The script is only slightly modified from that used in 1935, and the film is again titled *Röntgentonfilm der Sprache*. This has led to confusion in library and archive catalogues between films produced more than 20 years apart: a DVD of the later film [17] wrongly gives its production date as 1937.[4] In fact, X-ray films from the postwar period are easily recognised at a glance by the circular image field resulting from the image intensifier (see Figure 4).

Janker also provided input to a 17-minute film called *Röntgenstrahlen* [18], directed by Martin Rikli and subsidized by the *Reichsstelle für den Unterrichtsfilm* which won the prize for best scientific film at the Biennale in Venice in 1937. A sequence on speech extends from 13:42 to 14:46 and includes material which does not form part of the 1935 film and is not known elsewhere, consisting of some multi-vowel words, the days of the week and the months from January through May.

8. RESTORATION

It is uncertain whether any complete copy of the 1935 film survives in Germany, and the search for one is made difficult by the unhelpful confusion which has arisen with the superficially-similar film of 1956.

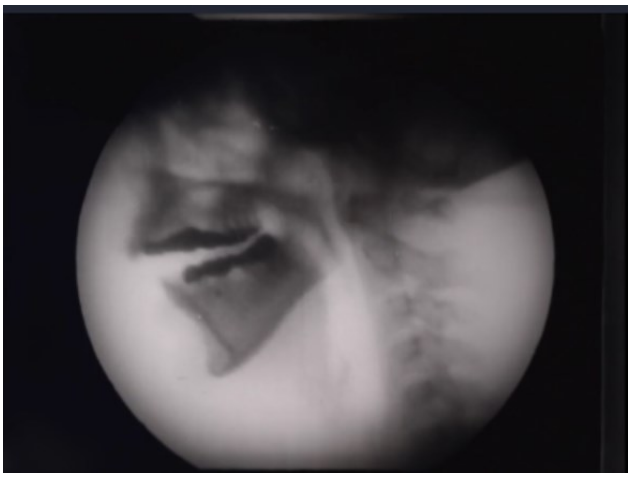


Figure 4: A typical single frame from the 1956 film.

Two 16 mm prints do however survive in London and came to light with other legacy film in 2008. One is a sound print of C 150, but is unfortunately severely damaged at the perforations along its entire length. It may have been left behind as a gift after the London congress, or obtained from RfdU soon afterwards. The other film surviving in London is a (double-perforation) silent print which is intact. There are reasons to think that it is a duplicate made locally from the sound copy before that was damaged.

The silent copy was digitised and screened by the first author in 2010. It runs for 3 minutes 20 seconds, and appears to contain the entire script as set out in [14] including the slow-motion (48 fps) section. It was possible to add synchronised sound to the first 40 seconds, taken from a fragment available online.[5]

But scanning and digitisation techniques have advanced over the last decade, and it is now possible to scan damaged film without access to the perforations. Our current restoration efforts are therefore directed at extracting as much information as possible from the damaged sound copy, and piecing-together from all the sources available a reconstruction of the film which contemporary judges found so 'remarkable', 'outstanding' and 'overpowering'.

Acknowledgements: Our thanks to Elfriede Tillian, Peter Adamson, Anthony Saffrey, Jo Griffin and the team at R3store Studios, London, Marcel Michels (Deutsches Röntgen-Museum, Remscheid) and two anonymous reviewers.

9. REFERENCES

[1] Scheier, M. 1897. Über die Bedeutung der Röntgenstrahlen für die Physiologie der Sprache und Stimme. *Phonetische Studien* 11. 40–56.

[2] https://scotlandonscreen.org.uk/browse-films/007-000-000-095-c

[3] Gutzmann, H. 1930. Röntgenfilmaufnahmen des Sprechens. *Bericht über die 1 Tagung der Internationalen Gesellschaft für experimentelle Phonetik in Bonn vom 10. bis 14. Juni 1930*, Bonn, 56–57.

[4] Menzerath, P. 1935. Neue Untersuchungen zur Steuerung und Koartikulation. *2nd ICPhS*, London, 1935, 220–225.

[5] Menzerath, P. and de Lacerda, A. 1933. *Koartikulation, Steuerung und Lautabgrenzung: eine experimentelle Untersuchung*. Berlin: F. Dümmler.

[6] Kurath, H. 1936. Second International Congress of Phonetic Sciences, July 22-26, 1935. *Language* 12(1), 59–62. doi: 10.2307/409021

[7] Firth, J. R. 1935. The Second International Congress of Phonetic Sciences. *English Studies* 17, 179–182.

[8] Kaiser, L. 1936. Bulletin of the International Society of Experimental Phonetics III. *Archives Néerlandaises de Phonétique Expérimentale*, 12, 111–147.

[9] *International Congress of Phonetic Sciences, 1962. Proceedings of the Fourth International Congress of Phonetic Sciences: held at the University of Helsinki, 4-9 September 1961*, edited by Antti Sovijärvi and Pentti Aalto. The Hague.

[10] Perkell, J. S. 1969. *Physiology of speech production : results and implications of a quantitative cineradiographic study*. Cambridge (Mass). M.I.T. Press.

[11] Menzerath, P. 1938. Neue Untersuchungen zur Wortartikulation (mit Vorführung eines Röntgentonfilms von Janker und Menzerath). Actes du quatrième Congrès International de Linguistes tenu à Copenhague du 27 Août au 1er Septembre 1936. Copenhagen: Einar Munksgaard, 67-75.Paris.

[12] Menzerath, P. 1937. Die Struktur der Artikulation (mit Lichtbildern und Vorführung eines neuen Lichtton-Abhörgerätes). XIe Congrès International de Psychologie. Paris, 25.-31 Juillet 1937. Rapports et Communications. Paris. Agen: Imprimerie Moderne, 1-2.

[13] Braun, A. & B. Möbius. 2023. Armando de Lacerda and his contemporaries: Paul Menzerath. In Q. Lopes, A. Braun & M. Ashby (eds.), *HSCR 2022. Lacerda 120. Proceedings of the Fifth International Workshop on the History of Speech Communication Research, Porto, May 28–29, 2022* (Studientexte Zur Sprachkommunikation 104), 17–28. Dresden: TUDpress.

[14] Janker, R. 1937. Röntgentonfilm der Sprache. Veröffentlichung der Reichsstelle für den Unterrichtsfilm zu dem Hochschulfilm Nr. C 150. Berlin.

[15] Janker, R. 1963. Entwicklung und Stand der Röntgenkinotechnik. Institut für den Wissenschaftlichen Film. Sonderarchiv-Film W 598. [ms.] 26p.

[16] https://av.tib.eu/media/28765?hl=keyword:"Pfeife"

[17] *Röntgentonfilm der Sprache*. 2007. DVD. Göttingen: IWF Wissen und Medien gGmbH..

[18] https://youtu.be/EkMlw6-aywI.

[19] Plug, L. 2008. J. R. Firth: A new biography. *Transactions of the Philological Society* 106(3), 337–374. doi: 10.1111/j.1467-968X.2008.00203.x.

[1] *Ideal wäre Tonröntgenfilm, für dessen Schaffung Verhandlungen angebahnt sind*. 'A sound X-ray film would be ideal, for the production of which negotiations are in progress'. As far as we know Gutzmann's film is not preserved. It was made in Berlin with radiologist Viktor Gottheiner (1899–1974). The rivalry between Janker and Gottheiner in the development of cineradiography will be covered in detail elsewhere.

[2] Firth had taught English pronunciation and intonation at a summer school at the University of Bonn in 1929 [19] and it seems probable that he was thus already acquainted with Menzerath.

[3] Its impact upon phonetic research is perhaps comparable to the development of the sound spectrograph in the acoustic domain.

[4] The confusion may have been made worse by the (presumably accidental) fact that the *Bestell-Nr.* (order number) of the DVD was D 150, resembling the C 150 number assigned to the earlier film.

[5] https://vlp.mpiwg-berlin.mpg.de/library/data/lit38421? The fragment is identical with the relevant section of film R 958 [15] and stops short at the same point in the text. It may thus have been taken from the 1963 film rather than from an original 1935 film.

RESTORATION OF A PORTABLE KYMOGRAPH

Angélique Amelot[1], Alain Ghio[2], Didier Demolin[1]

[1]LPP (UMR7018, CNRS/Sorbonne Nouvelle), Paris, France
[2]Aix Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France

ABSTRACT

The presence of an old portable kymograph in the Phonetics and Phonology Laboratory (LPP) of La Sorbonne Nouvelle gave the idea for a restoration project. Alexis Rygaloff purchased a portable Gauthier kymograph for field investigations in China around the middle of the 20th century. The kymograph (about 10 kg) consists of a recording cylinder moved by a clockwork mechanism associated with a Foucault regulator. The clockwork mechanism is connected to a drive bar by means of a belt placed on two wheels, one connected to the mechanism and the other to a drive bar. The registering devices (Marey's drums) built by Verdin are attached to a trolley fixed on a foot connected to the bar. The regulator allows to choose different movements (slow, medium, fast) depending on the nature of the experiments that are planned.

Keywords: Kymograph, Clockwork mechanism, Foucault regulator.

1. INTRODUCTION

At the beginning of the 17th century, Galileo was concerned with representing sounds graphically. Nevertheless, acoustic analyzes of the sounds' descriptions began before their graphic representations. We owe the first acoustic analyzes to Lissajous in 1857. The graphic representation of uniform sounds of short duration is linked to the invention of the *chronograph*, and Young, as early as 1807, showed his students a graphic recording of the sound vibrations of a tuning fork.

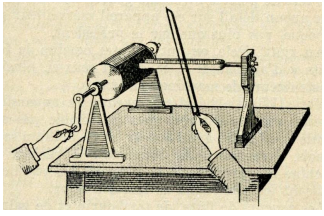


Figure 1: A picture of Thomas Young's from [1] p.4.

Another chronograph which constitutes an improvement on Young's model is the *vibrophone* invented by Duhamel in 1840. The major contribution of this device is the addition of a tuning fork-chronograph, which brings a precise temporal reference. This contribution allowed Marey to represent other phenomena graphically and no longer in a uniform way. This biologist who claimed as early as 1868 that all the sciences could be represented graphically, launched the starting point for the first recordings of speech for the purpose of sound description [2].

The kymograph belongs to the family of speech recorders in the same way as the *phonotaugraph* and the *phonograph*. The former allows the visualization of speech sounds, the latter is used to archive them. The speech recorder is inseparable from the speech recorders which are better known as *Marey's drums.* The recorder is the plane that records and traces the movements in the form of oscillations induced by the inscribers.

In this paper, we will describe the mechanism of a portable kymograph used by Rygaloff for field studies in China. We will also show the importance of the adjustment of the speech inscribers (Marey's drums) and the speed of rotation (Foucault regulator) which are fundamental elements to obtain interpretable and reproducible tracings.

2. HISTORY

2.1. Recording speech movements

The first trace of a device recording a graphic representation related to the measurement of time is linked to the use of the *kymographion* described by [3] to reproduce the activity of blood vessels in living animals. The originality of this device is linked to the fact that the brush color is changed with each turn of the cylinder, the lines are superimposed and can be fairly easily distinguished. The limit of this device is related to its use, which requires vivisection and therefore cannot be used in fields other than animal physiology [4]. In order to be able to represent the exact form of a pulsation and to represent it with a unit of time, Marey undertook the construction of a system based on the *kymographion* and the *sphygmograph* of Vierordt. He ended up with Chauveau in the development of a system called cardiograph which shows the importance of the inscribers which he will call levers.

At the same time, it was at the time when the abbot Rousselot wondered about devices that could help him to characterize phonetic phenomena by taking into account the speed parameter. He wondered about the interest of the cylinder and the speed of its drive [5: 76]. We will return to these notions in section 3.
The device selected by Rousselot which seemed to be the most suitable for recording the movements of speech was nothing other than the recording cylinder with clockwork and regulator mechanism built by Verdin, presented on the first page of his catalog of precision instruments [6]. This is, in our opinion, the first version of the apparatus which will later be called the kymograph.

2.2. Movement inscribers

To represent a movement faithfully, whether in terms of length, force or duration, implies that the marker used is reliable and precise [4] This started from the mechanism of Scott de Martinville's phonautograph (1857), a thin membrane which vibrates differently according to the influence of the sounds. The vibrations of this membrane are transmitted to a stylus which rubs on the smoked surface of a rotating cylinder. Marey [2], developed several inscribers of different movements like the transmission myograph, which in our opinion is closer to the needs of the more complex mechanisms of speech is the transmission myograph (p. 196). This is at the origin of the lever drum or Marey's drum used with the kymograph described in this paper.

3. PORTABLE KYMOGRAPH

3.1 Gauthier Kymograph

The Gauthier's portable kymograph (*Fig. 3*) consists of a stainless-steel drum (a) on which the paper is fixed, a Foucault regulator (b and *Fig. 4*), and a drive bar (c) on which the recording drums are positioned. This apparatus can be transported in a small wooden box (*Fig. 2*). Small boxes with the accessories necessary for the proper functioning of the kymograph can also be placed in the wooden box. The accessories are: a metal mouth mask, olives of different sizes, glass tubes for measuring nasal pressure, and laryngograph for laryngeal vibrations (*Fig. 10*). Recording drums of different diameters are associated with this set. These accessories can be connected to the recording drums by rubber tubes. Hollow straws, pens for scratching the paper, ink pens, glue, wooden pinches for handling the paper and adjustable tuning forks are also part of the apparatus.



Figure 2: Gauthier portable Kymograph and its travel box.

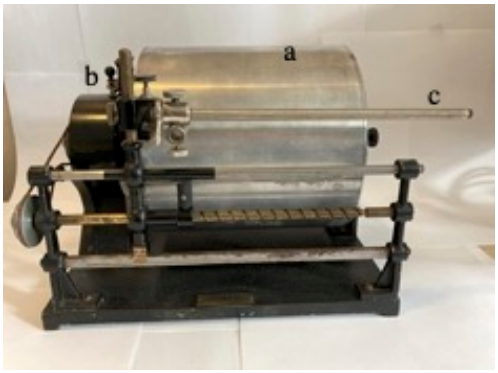


Figure 3: Gauthier Kymograph parts: a) drum, b) Foucault regulator, c) drive bar.

3.2 Foucault Regulator

At the end of his life, Foucault was especially known for the construction of regulators, initially intended to stabilize the drive of lighthouses, and telescopes. A huge quantity of regulator models can be found in the "Recueil des Travaux Scientifiques de Léon Foucault" [7].
The regulator is made up of a spring-loaded clockwork mechanism that winds up with a key. It is connected to a drive bar through a belt placed on two wheels, one connected to the clockwork mechanism and the other to a drive bar. The recording devices are attached to a trolley fixed on a foot connected to the bar. The controller can allow to choose different speed movements (slow, medium, fast) depending on the nature and the precision of the experiments to be made. The clockwork gives the cylinder 3 different speeds (*Fig. 4*), 1 and 8 revolutions per second and 1 revolution per minute. A brake placed on the regulator frame allows it to be started and stopped at will. The regulator itself consists of a disk with 4 adjustment masses which regulate the rotation of the drum. To adjust the regulator, it is possible count the number of revolutions that the cylinder performs in a

given time. Once the right speed has been chosen, it is noticed that the first revolutions are irregular, but that after the rotation of the drum becomes more regular. The last rotation is also irregular because of the braking imposed by the regulator.

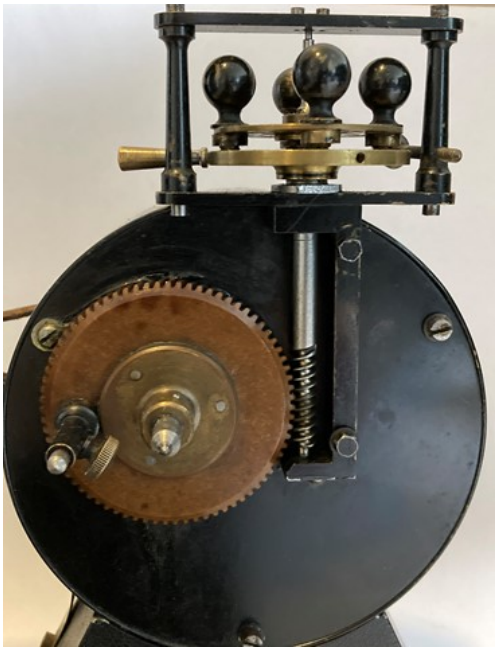


Figure 4: Drive wheel at the output of the clockwork mechanism and Foucault regulator.

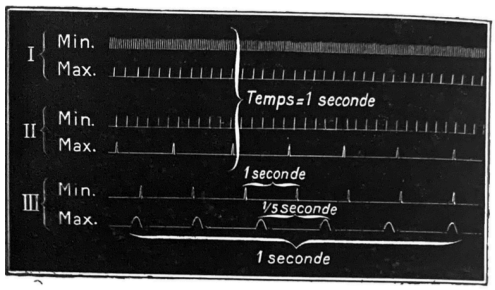


Figure 5: Illustration to explain the 3 possible rotation speeds of the Foucault regulator from [8]

The recording apparatus which we have just described does not, therefore, have a uniform speed of rotation. It is, therefore, necessary to associate it with a tool allowing to have a reliable time base. The means adopted for a portable kymograph is the use of a tuning fork. A light feather is attached to the end of one of its branches, which allows direct measurement of its vibrations at the same time as the traces are obtained on the paper placed on the drum. An adjustable tuning fork provides 25 or 50 Hz vibrations that are suitable for slow movements. A tuning fork tuned for vibrations between 200 and 1000 Hz is associated with rapid movements.

Blackened paper is placed on the drum. The blackening of the paper is obtained either by a small torch or by the smoke coming from a candle with a thick wick which gives a thick black smoke. This candle called *cellar rat* (Rat de cave in French) is placed under the drum which is rotated on itself to blacken the entire sheet of paper. There were devices simplifying the smoking system where the smoke of cellar rat was distributed under several cylinders that were turned manually (*Fig. 6*).

Another method employed to inscribe the paper was to leave the paper white and obtain black lines with ink placed in a quill at the end of the straws placed on the inscribing drums.

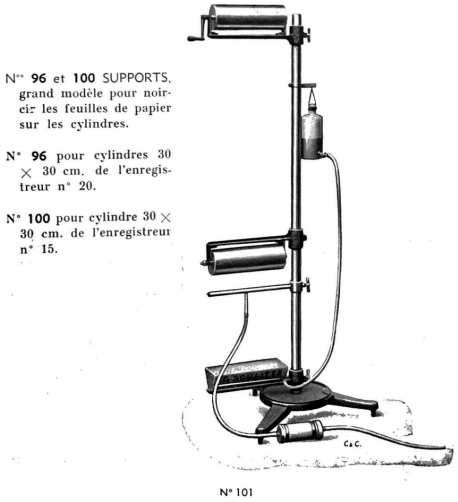


Figure 6: Smoke holder from [8]

Lever drums consist of a metal tank closed at the top by a thin rubber membrane with some tension (*Fig. 7* & F on *Fig. 8*). They carry a metal tube that opens inside and fits a rubber tube (C in *Fig. 8*). The air that passes through the tube and enters the tank and lifts the rubber membrane. On the membrane an aluminum disc is glued and connected with a lever which is articulated by one of its ends to a small straw placed in the extension of the axis (E in *Fig. 8*). This joint allows the lever to perform vertical movements. The straw is equipped with a pen that traces on the smoky paper the movements induced by the membrane [4].

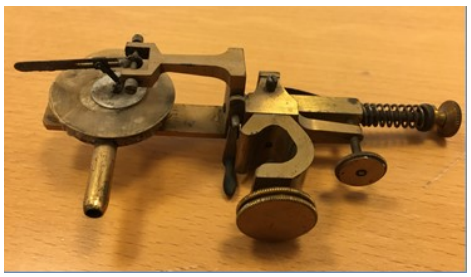


Figure 7: Marey's drum built by Verdin used with our kymograph.

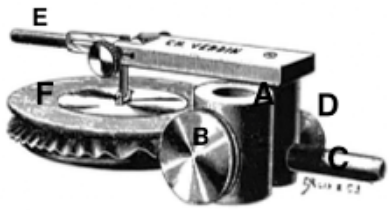


Figure 8: Marey's drum with rubber cover (F), hole to fix the drum on drive bar (A), Adjustment knob (B & D), bar to put the straw which is used to mark the cylinder (E), tube through which the air passes (C) [8] p.26.

Once the assembly of the drum is done, it is necessary to choose the point of the cylinder where the inscription will take place, which amounts to adjusting the position of the drum and that of the carriage. The feather placed at the end of the straw should touch the surface of the blackened paper.

4. RESTORATION STEPS

The first restoration step of this kymograph was to make the clockwork mechanism work again and to test the various parts involved to make it work. We also had to test the tension of the drive belt to have a continuous and regular rotation.

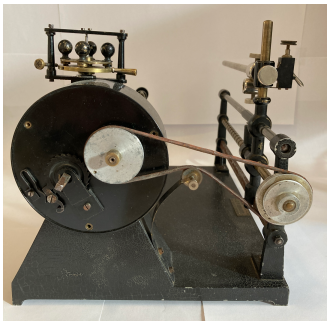


Figure 9: Belt, rolling disks, and drive bar.

Once the regulator of Foucault was put back in operation, we tested the various instruments allowing to capture the differences in air movements that we had at our disposal: the nasal olives for the air passing by the nose, the mouthpiece for the air passing by the mouth and the laryngograph with a double bowl of Zund-Burguet; two drums whose membrane is raised by springs, fixed together and opened into a Y-shaped pipe, described in the catalog [9].

a b c

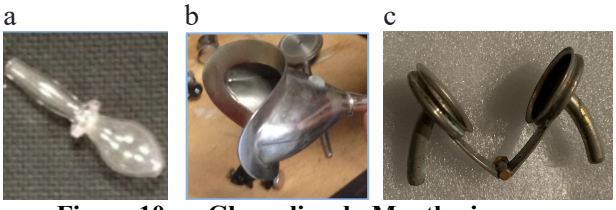


Figure 10: a. Glass olive, b. Mouth piece, c. Laryngograph

The second restoration step was to attach the tabs to the bar, to space them far enough apart so that the curves would not overlap, and to find an effective angle for the straws to scrape the blackened paper. The different instruments (*Fig. 10*) allowed us to test different stiffnesses of the membrane attached to the drum, different bending angles of the straws, and different speeds of the controller, see *Fig. 11* for the complete setup used and *Fig. 12* to see the obtained plots.

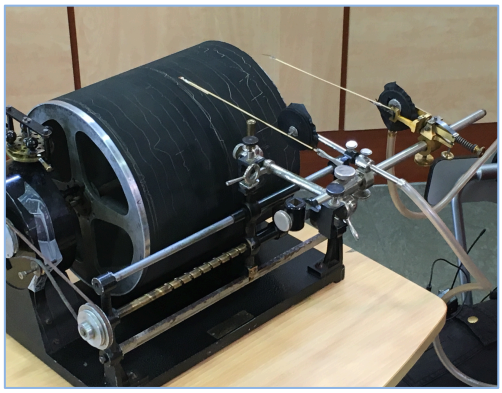


Figure 11: Complete Kymograph with 2 Marey's drums, straws and pen on blackened paper.

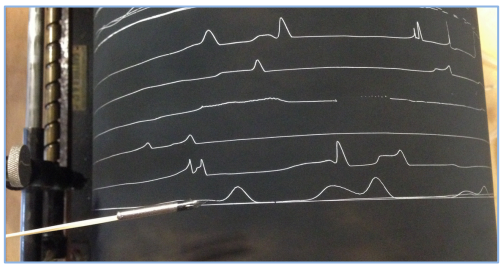


Figure 12: Straw with pen scratching blackened paper.

The next step in this restoration process is to test the instruments (the same ones used by Rousselot and recently restored) with the kymograph in an attempt to reproduce the experiments made by Rousselot at the end of the 19th century.

5. ACKNOWLEDGEMENTS

We wish to dedicate this paper to the memory of Bernard Teston who took an essential part in the first steps of this restoration. We also wish to thank Annie Rialland who made the Kymograph available to us. Finally, the Labex-EFL who supported parts of the costs. The Labex EFL, Aix Marseille University and the LPL for supporting parts of the restoration, Jean Shoentgen and the University of Brussels for providing us the Verdin Marey's drum.

6. REFERENCES

[1] Hémardinquer, P., 1930, *Le phonograph et ses merveilleux progrès*, Masson et Cig. [Online]. Available: http://www.hervedavid.fr/francais/phono/HEMARDINQUER - Phonographe.htm

[2] Marey, É.-J., 1868, *Du mouvement dans les fonctions de la vie : leçons faites au Collège de France*, Paris: Germer Baillière.

[3] Ludwig, C. F. W., "Beiträge zur Kenntniss des Einflusses der Respirationsbewegungen auf den Blutlauf im Aortensysteme," in *Archiv für Anatomie, Physiologie und wissenschaftliche Medizin*, pp. 242–302.

[4] Marey, E. J., 1878, *La méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine*, Paris.

[5] Rousselot, P. J., 1897, *Principes de phonétique expérimentale*, Paris.

[6] Verdin, C., 1895, "Catalogue des instruments de précision construits par Charles Verdin," vol. 1895, no. c, p. 164.

[7] Gariel, C.-M., 1878, *Recueil des travaux scientifiques de Léon Foucault*, Paris: Gauthier-Villars.

[8] Boulitte, G., 1928, *Cylindres enregistreurs, Inscripteurs à air et électromagnétiques. Appareils pour la mesure et l'enregistrement du temps. Support pour inscripteurs*, Paris: Boulitte, G.

[9] Boulitte, G., 1924, *Catalogue G : phonétique*, Paris: Boulitte, G.

ACOUSTIC CORRELATES AND CONTRAST MAINTENANCE IN BURMESE VOICELESS SONORANTS

Chiara Repetti-Ludlow

New York University
ChiaraRepettiLudlow@nyu.edu

ABSTRACT

This study examines acoustic correlates to the Burmese sonorant voicing contrast. In Burmese, voiceless sonorants are identified largely by preaspiration, and previous literature has found other correlates are important to the contrast as well. However, no study has yet examined how these acoustic correlates vary in different phrasal positions. To address this topic, a production study was carried out with participants reading minimal pairs in three phrasal positions. Results show that preaspiration and F0 on the sonorant are important correlates in all three phrasal positions. In initial and medial position, SoE on the following vowel also significantly differentiated voiced and voiceless sonorants. However, in final position, vowel devoicing causes SoE information to be lost, which results in HNR becoming a more important correlate. Thus, there seems to be a trading relationship.

Keywords: acoustic correlates, Burmese, voiceless sonorants, vowel devoicing, prosody

1. INTRODUCTION

Voiceless sonorant consonants are typologically rare segments, occurring in only 2% of the world's languages [1]. Burmese is one such language, with 6 phonemic voiceless sonorant consonants: [m̥, n̥, ɲ̥, ŋ̥, l̥, w̥] [2, 3], with previous literature typically describing these segments as voiced sonorants with a preceding voiceless period, similar to preaspiration [4, 5, 6]. Previous work on these segments has focused on correlates such as fundamental frequency (F0) [7], duration [4, 6], electroglottography [8], and oral and nasal airflow [6, 8]. Although this research has furthered our understanding of Burmese voiceless sonorants, the focus has been on aerodynamic and temporal aspects of the sonorants, rather than correlates related to voicing and aspiration. Given the salience of aspiration as an acoustic correlate and the effects it can have on adjacent segments due to coarticulation [9], a more thorough examination of the acoustic correlates to the Burmese sonorant voicing contrast is warranted, with a focus on preaspiration.

Another question addressed in this study is how phrasal position and acoustic correlates interact. Acoustic correlates often re-weight due to prosodic constraints, such that a given correlate is fully or partially neutralized in a certain prosodic environment. For example, in languages like German [10], cues to the voicing contrast, such as voice onset time (VOT), are lost in word-final position. To compensate for this loss of acoustic information, other cues receive a higher relative weight, thereby allowing listeners to recover the contrast. In the case of German, when VOT information is lost, the duration of the preceding vowel receives a higher weight and allows for the contrast to be maintained. Phrase-final loss of acoustic information has been observed for other languages and types of contrasts as well. For example, in Mandarin, F0 is a primary cue to tonal contrasts, but in phrase-final position, the F0 cue weakens due to devoicing [11]. This results in correlates like HNR becoming more important to compensate for the neutralization. Similarly, in Newcastle English, /t/ has preaspiration in phrase-final position, which reinforces the segment that would otherwise be partially neutralized due to the absence of release burst information [12]. Put generally, in cases where a correlate of a contrast becomes unavailable in some prosodic environments, other correlates may be sufficient to compensate for lost acoustic information.

In order to construct a full picture of the Burmese sonorant voicing contrast, it is necessary to consider the relative importance of acoustic correlates and the ways in which correlates re-weight based on prosodic constraints. This work therefore sets out to answer three primary research questions:

1. What are the main acoustic correlates to the Burmese sonorant voicing contrast?
2. Are there any prosodic neutralization processes and do they result in re-weighting?
3. Does the presence of preaspiration change the central correlates to the voicing contrast?

2. METHODS

To examine the acoustic realization of voiced and voiceless sonorants in Burmese, a production study was conducted in which participants read target words in one of three phrasal positions. Target words were monosyllabic minimal pairs consisting of every combination of 8 onset consonants ([w, w̥, l, l̥, ŋ, ŋ̥, m, m̥]), 2 tones (high and low), and the nucleus vowel, [a]. Each target word appeared in 9 different carrier phrases, 3 for each phrasal position (phrase-initial, phrase-medial, and phrase-final). Each phrase targeting a given prosodic position had the same number of syllables, the same preceding consonant, and no adjacent aspirated stop. All stimuli were checked by a native speaker of Burmese and target phrases were sorted into three randomized lists for a total of three participants per list. Altogether, these combinations resulted in 144 target phrases, such as [ŋa **la** hãʊ̃] 'my old mule.' Further sonorants, tones, and vowels were not included and no filler sentences were used in order to keep the experiment under an hour. Phrases were presented to participants in Burmese orthography one at a time in a slideshow and participants were given a break after every 10 phrases.

Nine native speakers of Burmese were recorded: six women and three men, ranging in age from 19-42. All participants were living in New York at the time of the study and had been living in the United States between 1-20 years. Participants were contacted primarily through personal channels and Facebook groups and were reimbursed $25 for participating in the hour-long study. No participants were excluded. Participants were recorded in a sound-attenuated booth using a Zoom H4n Pro recorder and a head-mounted Shure WH20 microphone. The recordings were done with a bit depth of 16-bit for a sampling rate of 44.1 kHz. After collecting the data, target words were labeled in Praat [13], with the preceding vowel, preaspiration (when relevant), sonorant, and following vowel individually labeled in a TextGrid, as shown in Figure 1. Aspiration was segmented according to the beginning and end of clear frication in the spectrogram and waveform, and the sonorant and vowels were segmented based on the beginning and end of the second formant in order to reliably attain voice quality measures.

The measures considered in the analysis were duration, H1-H2, strength of excitation (SoE), harmonics-to-noise ratio 0-3500 Hz (HNR), and F0, in line with recent work on phonation and sonorants [14, 15]. A summary of these measures can be found in Table 1. Average measures taken over the sonorant and adjacent vowels were obtained using VoiceSauce [16]. A linear discriminant analysis (LDA) was run in R using the MASS package lda() function to determine the acoustic correlates that were best able to predict the voicing category of the sonorant [17]. Linear mixed-effects regressions were then carried out using the lme4 package lmer() function to verify that the values of these acoustic correlates were significantly different between voiced and voiceless sonorants [18].

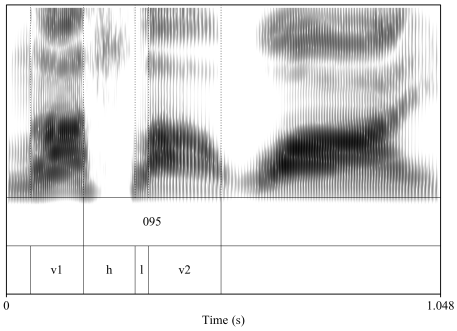


Figure 1: Labeling of [ŋa l̥a wãĩ] 'my round sword' with preaspiration.

Acoustic measure	Correlate
Duration (ms)	Segment length
H1-H2 (dB)	Breathiness/Creak
SoE	Degree of voicing
HNR (dB)	Asp./Non-modal phonation
F0 (Hz)	Aspiration/Voicing

Table 1: Acoustic measures and their correlates.

3. RESULTS

In order to determine the most important correlates to the contrast, three LDAs were carried out, one for each phrasal position. Duration, H1-H2, SoE, HNR, and F0 measures taken over the sonorant and adjacent vowels were used in the analysis, as was the duration of preaspiration (0 if absent). The phrase-initial LDA predicted voicing category with 84% accuracy, and the phrase-medial and phrase-final LDAs predicted voicing category with 87% accuracy. In all phrasal positions, the most important correlate of the contrast was the duration of preaspiration (LDs: initial = 0.99, medial = 1.07, final = 1.17). Preaspiration was present in 79% of voiceless sonorants, with no significant differences based on phrasal position. The second most important correlate in all phrasal positions was F0 of the sonorant (LDs: initial = 0.87, medial =

0.38, final = 0.50). In initial and medial position, SoE of the following vowel was the third most important correlate (LDs: initial = -0.42, medial = -0.23), but in final position HNR of the sonorant was the third most important (LD: final = -0.42). Linear mixed-effects regressions were carried out on these correlates and the results, shown in Table 2, confirm that the top three LDA correlates for each phrasal position are significantly different between voiced and voiceless sonorants. Of particular interest is the third most important cue, which is different in phrase-final position compared to the other positions, as illustrated in Figure 2.

Initial	β	z	p value
Aspiration duration	58.65	6.73	**<0.01***
F0 of sonorant	36.06	10.99	**<0.01***
HNR of sonorant	-0.28	-0.29	0.77
SoE of vowel	-0.0017	-6.54	**<0.01***
Medial	β	z	p value
Aspiration duration	84.96	11.27	**<0.01***
F0 of sonorant	21.18	6.30	**<0.01***
HNR of sonorant	-2.31	-2.10	0.06
SoE of vowel	-0.0009	-2.53	**0.02***
Final	β	z	p value
Aspiration duration	78.97	13.70	**<0.01***
F0 of sonorant	21.61	4.26	**<0.01***
HNR of sonorant	-3.67	-2.58	**0.02***
SoE of vowel	-0.0004	-0.75	0.46

Table 2: Results of linear mixed-effects models for the relevant LDA correlates in all phrasal positions.

One interpretation of the finding that SoE does not differentiate voiced vs. voiceless sonorants in phrase-final position is that Burmese undergoes phrase-final vowel devoicing, as indicated by a lower SoE. A linear mixed-effects model was run with the vowel's SoE as the dependent variable, and the interaction between sonorant voicing and phrasal position as a fixed effect. The results of this model, shown in Table 3, indicate that vowel devoicing occurs phrase-finally regardless of whether the preceding sonorant is voiced or voiceless, as shown in Figure 3. This suggests that there is a neutralizing process in phrase-final position that impacts one of the correlates that is used for the sonorant voicing contrast: SoE. This results in sonorant HNR receiving a higher weight.

Finally, given that there are two possible realizations of voiceless sonorants: one with a period of preaspiration and one without it, an analysis was carried out to determine whether the

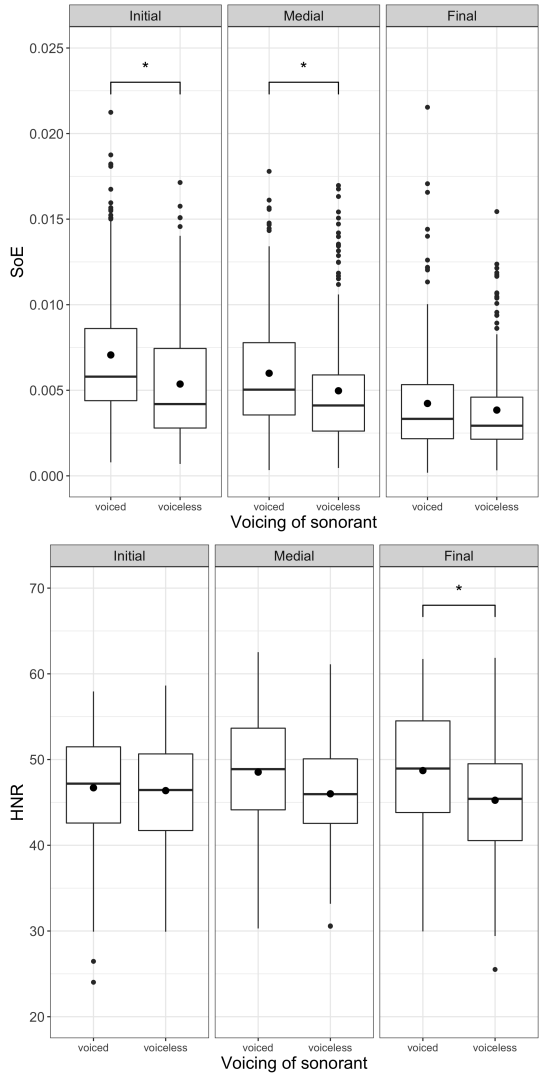


Figure 2: SoE on the vowel (top) and HNR on the sonorant (bottom) faceted by phrasal position. Stars represent significant differences and dots in the box plots represent mean values.

Difference	Son.	β	z	p value
Initial-Med.	voiced	0.0004	2.20	0.06
Initial-Final	voiced	0.003	13.74	**<0.01***
Med.-Final	voiced	0.002	8.48	**<0.01***
Initial-Med.	vless	0.0003	1.51	0.39
Initial-Final	vless	0.002	7.36	**<0.01***
Med.-Final	vless	0.001	5.83	**<0.01***

Table 3: Results of the linear mixed-effects model examining vowel SoE, phrasal position, and voicing of the preceding sonorant.

same correlates significantly differentiated voiced vs. voiceless preaspirated and voiced vs. voiceless unaspirated sonorants. Linear mixed-effects models

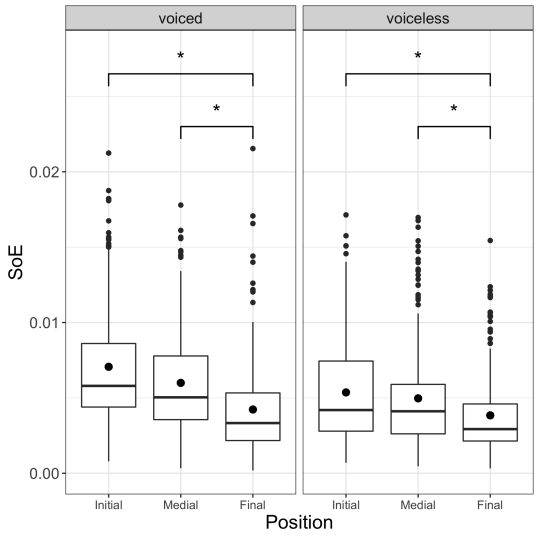


Figure 3: SoE on the vowel faceted by sonorant voicing. Stars represent significant differences and dots in the box plots represent mean values.

were carried out for the three relevant correlates (F0 of sonorant, HNR of sonorant, SoE of vowel) in each phrasal position, and results are shown in Table 4. F0 on the sonorant significantly differentiated sonorant voicing in all three phrasal positions regardless of aspiration. HNR on the sonorant was significantly different for voiced vs. voiceless preaspirated sonorants only in medial and final positions. SoE on the following vowel significantly differentiated voiced and voiceless sonorants in initial and medial position regardless of aspiration, but not in final position.

4. DISCUSSION

First, the results of this study suggest that preaspiration is the primary correlate of the Burmese sonorant voicing contrast in all three phrasal positions, but it is not the only correlate that is important to the contrast. Correlates such as F0 and HNR on the sonorant, and SoE on the following vowel all provide important acoustic information, demonstrating that both correlates related to complex phonation (preaspiration, HNR) and correlates traditionally associated with voicing contrasts (F0, SoE) are used. Second, this study found that Burmese undergoes phrase-final vowel devoicing, similar to what has been observed in languages like French [19] and Bulgarian [20]. This is reflected by a lower SoE on final vowels compared to vowels in other phrasal positions. Given that SoE on the vowel following the sonorant is an important correlate to the voicing contrast in initial and

Initial	Asp.	β	z	p value
F0 son	preasp	39.39	12.99	**<0.01***
HNR son	preasp	-1.05	-1.13	0.28
SoE vowel	preasp	-0.0003	-6.30	**<0.01***
F0 son	unasp	29.13	8.30	**<0.01***
HNR son	unasp	1.31	1.30	0.21
SoE vowel	unasp	-0.0013	-3.42	**<0.01***
Medial	Asp.	β	z	p value
F0 son	preasp	24.12	7.68	**<0.01***
HNR son	preasp	-2.81	-2.62	**0.02***
SoE vowel	preasp	-0.0008	-2.33	**0.03***
F0 son	unasp	11.08	2.82	**<0.01***
HNR son	unasp	-0.62	-0.52	0.61
SoE vowel	unasp	-0.0011	-2.54	**0.02***
Final	Asp.	β	z	p value
F0 son	preasp	25.19	5.22	**<0.01***
HNR son	preasp	-3.90	-2.74	**0.02***
SoE vowel	preasp	-0.0005	-0.93	0.37
F0 son	unasp	10.33	1.88	**0.05***
HNR son	unasp	-2.93	-1.94	0.07
SoE vowel	unasp	-0.0001	-0.12	0.91

Table 4: Results of linear mixed-effects models for voiced vs. voiceless preaspirated sonorants and voiced vs. voiceless unaspirated sonorants.

medial positions, this means that important acoustic information is lost in final position. However, HNR measured over the sonorant becomes more important in this phrasal position, potentially to compensate for the lost acoustic information. Since the LDA's predictive power is equally strong in medial and final position, this means that the model isn't performing worse despite lost SoE information. Thus, the voicing contrast is not neutralized in phrase-final position, but rather, correlates re-weight to preserve the contrast. Third, the presence or absence of preaspiration did not seem to impact F0 on the sonorant or SoE on the following vowel, but it did impact HNR on the sonorant. While HNR was never a significant correlate phrase-initially, in medial and final position, it was only significantly lower for voiceless sonorants when preaspiration was present. This suggests that HNR is tied to the period of preaspiration.

These findings demonstrate that in Burmese, the neutralization of acoustic correlates can lead to re-weighting in certain prosodic positions, similar to what has been observed in German [10]. However, this is contingent on other factors, such as the presence of preaspiration, as some correlates are interdependent.

5. REFERENCES

[1] S. Moran and D. McCloy, "PHOIBLE online," 2011. [Online]. Available: phoible.org

[2] P. Ladefoged and I. Maddieson, "The sounds of the world's languages," *Language*, vol. 74, no. 2, pp. 374–376, 1998, publisher: Linguistic Society of America.

[3] J. W. Watkins, "Burmese," *Journal of the International Phonetic Association*, vol. 31, no. 2, pp. 291–295, Dec. 2001.

[4] M. Dantsuji, "A study on voiceless nasals in Burmese," *Studia phonologica*, vol. 18, pp. 1–14, 1984.

[5] P. Ladefoged, *Preliminaries to linguistic phonetics*. University of Chicago press, 1971.

[6] P. Bhaskararao and P. Ladefoged, "Two types of voiceless nasals," *Journal of the International Phonetic Association*, vol. 21, no. 2, pp. 80–88, Dec. 1991.

[7] I. Maddieson, "The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis," *Journal of Phonetics*, vol. 12, no. 1, pp. 9–15, Jan. 1984.

[8] K. Chirkova, P. Basset, and A. Amelot, "Voiceless nasal sounds in three Tibeto-Burman languages," *Journal of the International Phonetic Association*, vol. 49, no. 1, pp. 1–32, Apr. 2019.

[9] J. Blevins, "Evolutionary phonology and the life cycle of voiceless sonorants," *Typological hierarchies in synchrony and diachrony*, pp. 29–60, 2018, publisher: John Benjamins Amsterdam & Philadelphia.

[10] R. Port, F. Mitleb, and M. O'Dell, "Neutralization of obstruent voicing in German is incomplete," *The Journal of the Acoustical Society of America*, vol. 70, no. S1, pp. S13–S13, 1981, publisher: Acoustical Society of America.

[11] N. Rhee and J. Kuang, "The different enhancement roles of covarying cues in Thai and Mandarin tones," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2407–2411.

[12] P. Foulkes, G. Docherty, and D. Watt, "Tracking the emergence of sociophonetic variation," in *Proceedings of the 14th international congress of phonetic sciences*, 1999, pp. 1625–1628.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer Program]," 2015.

[14] K. H. Berkson, "Acoustic correlates of breathy sonorants in Marathi," *Journal of Phonetics*, vol. 73, pp. 70–90, Mar. 2019.

[15] M. Garellek, "Acoustic discriminability of the complex phonation system in !Xóõ!" *Phonetica*, vol. 77, no. 2, pp. 131–160, Mar. 2020.

[16] Y. Shue, "The voice source in speech production: Data, analysis and models," 2010.

[17] W. N. Venables and B. D. Ripley, "Random and mixed effects," in *Modern applied statistics with S*. Springer, 2002, pp. 271–300.

[18] R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, Nov. 2008.

[19] A. Dalola and K. Bridwell, "The shape of [u]: Towards a typology of final vowel devoicing in continental french," *ICPhS19, Melbourne, Australia, August*, pp. 5–9, 2019.

[20] B. Andreeva and J. Koreman, "The status of vowel devoicing in Bulgarian: Phonetic or phonological," in *Formal description of slavic languages: The fifth conference (FDSL5), leipzig*, 2003, pp. 81–91.

ANALYZING VARIABILITY IN CLOSURE VOICING AND CO-INTRINSIC F0 IN CENTRAL STANDARD SWEDISH

James Kirby[1, a)], Maryann Tan[2, b)]

[1]Institute for Phonetics and Speech Processing (IPS), LMU Munich
[2]Centre for Research in Bilingualism, Stockholm University
a) jkirby@phonetik.uni-muenchen.de, b) maryann.tan@biling.su.se

ABSTRACT

The voicing contrast of Central Standard Swedish has been argued to be typologically unusual, on the grounds that it contrasts phonetically prevoiced with voiceless aspirated plosives. However, reports vary regarding how often voiced plosives are actually realized with closure voicing. In this study, we measure how often phonologically voiced plosives /b d/ are devoiced, along with F0 trajectories and aspiration durations following release of closure. Results reveal considerable variation in voicing rates: some speakers consistently prevoice /b d/, while others consistently devoice them. Phonologically voiced plosives are frequently realised with short-lag VOT, regardless of whether or not they are also realized with closure voicing. Finally, F0 trajectories following nasals and phonologically voiced plosives are indistinguishable, regardless of whether plosives are prevoiced or devoiced, but differ significantly from those following phonologically voiceless plosives. F0 may thus be a more reliable phonetic indicator of laryngeal status in Swedish than the presence of closure voicing.

Keywords: Swedish, microprosody, voicing, VOT

1. INTRODUCTION

Central Standard (Stockholm) Swedish (CSS) has been argued to have a typologically unusual, phonetically overspecified two-way laryngeal contrast between prevoiced lenis /b d g/ and postaspirated fortis /p t k/ plosives [1, 2]. However, reports vary regarding the extent to which /b d g/ are actually prevoiced, at least in utterance-initial position. While some researchers [1, 2] have reported primarily strong prevoicing for both males and females (as do [3] for Umeå Swedish), others [4, 5] maintain that they are typical voiceless unaspirated, especially in the speech of female talkers, presumably due to a smaller supralaryngeal vocal tract disfavoring the aerodynamic conditions required for voicing [6].

This reported variability in the realization of prevoicing makes CSS an interesting language to consider in terms of the microprosodic or *co-intrinsic* pitch perturbations (hereafter CF0) conditioned by onsets. First observed in so-called "aspirating" languages such as German [7] and English [8], CF0 effects have also been documented in "true voicing" languages such as French [9, 10], Italian [10], and Spanish [11]. The basic observation is that a CF0 dichotomy exists between phonologically voiced and voiceless (or lenis and fortis) obstruents, F0 being higher following the voiceless/fortis member compared with the voiced/lenis member, regardless of other aspects of their phonetic realization [12, 13].

If CF0 effects are primarily the result of a gesture or gestures designed to support or enhance voicing [12, 14], CF0 might be expected to differ following devoiced and prevoiced tokens of phonologically voiced stops, particularly if the effect is contingent on the realization of glottal pulsing during the closure phase. If the dichotomy is primarily driven by gestures implemented to suppress voicing during the closure phase of voiceless plosives [13, 15, 16], on the other hand, F0 trajectories following phonologically voiced plosives would not be expected to differ regardless of whether or not plosives are realized with closure voicing.

Because CF0 effects have not previously been reported for CSS, and because reports differ regarding the frequency with which CSS voiced plosives are realized with closure voicing, in this study we revisit the acoustic realization of this laryngeal contrast in a sample of CSS speakers. We focus on the questions of how often phonologically voiced plosives in CSS are realized with/out glottal pulsing during the closure phase (Q1), as well as how prevoiced and devoiced plosives differ in terms of their post-release voicing lag (Q2) and F0 trajectories (Q3).

2. METHODS AND MATERIALS

2.1. Participants

Forty native speakers of Standard Central (Stockholm) Swedish (24 female, ages 20-44, median 28;

16 male, ages 18-43, median 34) were recruited for this study. They were paid 100SEK in the form of cash or gift vouchers for their effort. No participants reported a history of speech or hearing disorders. Data from one female speaker was removed as she was later revealed not to speak the Stockholm variety, leaving data from 39 speakers for analysis.

2.2. Speech materials

Forty-eight words with /b p d t m n/ onsets were selected for recording (see Appendix) along with 22 real and non-words of the form /hVd/ or /hVr/ for a separate vowel analysis not reported here. We aimed to include as many monosyllabic voicing pairs as possible, but did not control for word frequency or other lexical characteristics. Nasal onsets were included to provide a baseline for assessing differences in the F0 trajectories [13]. There were 19 items with long vowels /ɑː oː iː/ and 29 with short vowels /a u i ʉ/. All disyllables bore the grave accent (accent 2).

2.3. Procedure

Speakers were recorded in a sound-attenuated room in the Multilingualism Lab of the Centre for Research in Bilingualism at Stockholm University, using an audio-Technica AT 3505 microphone at a sampling rate of 44.1 kHz. Words were distributed across four blocks such that word triplets did not appear in the same block. Each word within each block was repeated five times, with the order of words within each block pseudo-randomized so that repetitions of a word did not appear consecutively. Block presentation across participants followed a latin-square order. Items were displayed on a monitor and the researcher controlled the presentation.

2.4. Annotation

Recordings were manually annotated and stored as an EMU speech database [17]. Items were annotated to indicate oral (c)losure and (o)pen intervals and points indicating the onset (v) and possible cessation (cv) of voicing during the closure, along with a point indicating the onset or resumption (r) of voicing following the release burst (see Figs. 1-2).

2.5. Analysis

In what follows, we use different terms to distinguish the *phonological* voicing specification of an onset (*voiced* = /b d/, *voiceless* = /p t/) from its *phonetic* realization (*prevoiced* = [b d], *devoiced* = [b̥ d̥], *aspirated* = [pʰ tʰ]). While in theory /p t/ could be

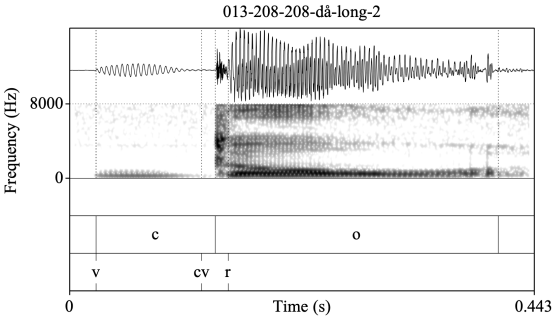


Figure 1: Example of prevoiced *då* 'then', speaker F13, repetition 2, illustrating both (negative) voice lead and (positive) voice lag. See Sec. 2.4 for explanation of labels.

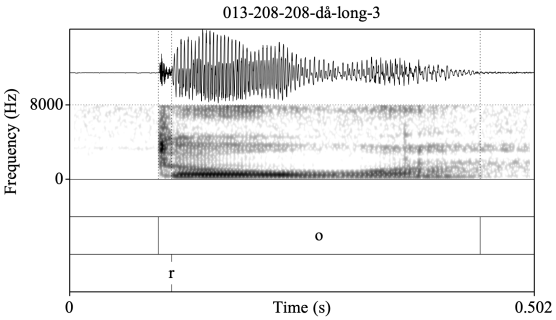


Figure 2: Example of devoiced *då* 'then', speaker F13, repetition 3, illustrating post-release voice lag but no voice lead.

realized as [b d], no tokens of /p t/ showed evidence of voicing during the closure in our data. Duration of voicing preceding the release burst was measured for each token containing a phonetically prevoiced plosive [b d] as either the time of the interval c, or the time of the difference between the points cv and v, whichever was smaller. In addition, we calculated the After Closure Time (ACT) [18] as the time from closure release (start time of o interval) to the onset of regular glottal pulsing of the following vowel (time of point r; see Figs. 1-2). For voiceless aspirated plosives, this corresponds to VOT, but while VOT for a prevoiced plosive may be negative, its ACT may be positive (as in Fig. 1). F0 was measured at 5 msec intervals throughout the open phase using the `ksvF0` estimator in the *wrassp* package [19].

3. RESULTS

3.1. Proportion of prevoicing

Out of 3282 tokens of phonologically voiced plosives (/b d/), slightly over half (1842) had at least

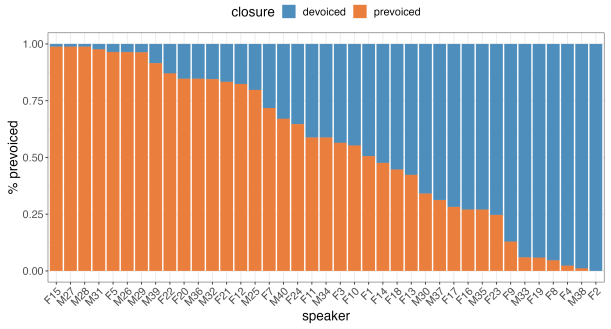


Figure 3: Proportion of /b d/ tokens realized with at least some prevoicing by speaker.

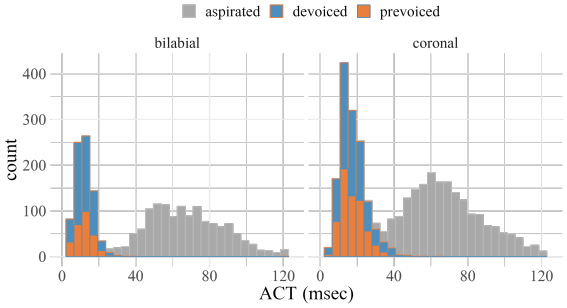


Figure 4: Distributions of After Closure Time (in msec) by phonetic realization of obstruent.

some closure voicing; we refer to these tokens as *prevoiced*. As shown in Fig. 3, however, the distribution of prevoicing across speakers is not uniform: some speakers prevoiced nearly every instance of every /b d/ token, while others consistently devoiced them. Visual examination did not reveal any obvious effects of lexical item, although impressionistically, tokens produced with closure voicing by primarily "devoicing" speakers were often coronals with high vowels (e.g. *dill, dimm, dipp*).

To explore the factors influencing devoicing rates, we fit a generalized linear mixed model (GLMM) with a logistic link function using the *lme4* package [20] to predict the probability of devoicing from log word *frequency*, vowel *height* (high or low), *place* of articulation of the onset (bilabial or coronal), and speaker *age* and *sex*. Interactions were explored but dropped after model criticism. Word frequencies (occurrence per million) were taken from the Swedish PAROLE corpus [21]. Random intercepts were included for items and speakers, with (uncorrelated) by-speaker random slopes for *place*, *height*, and *frequency*. Onset *place*, speaker *sex*, and vowel *height* were Helmert-coded. Only *sex* and *age* were significant predictors ($p < 0.05$) and both significantly improved model fit (*sex*: $\chi^2 = 5.55, Pr(> \chi^2) = 0.018$; *age*: $\chi^2 = 4.88, Pr(> \chi^2) = 0.027$). Conditional and marginal R^2 values computed with the `MuMIn` package [22] indicate that much of the variance is explained by the random effects ($R^2_m = 0.13, R^2_c = 0.66$), primarily due to the random term for *speaker* (without which $R^2_m = 0.14, R^2_c = 0.15$).

3.2. After Closure Time

Fig. 4 shows the distribution of prevoiced, devoiced, and voiceless plosives by place of articulation. To assess magnitude differences, we fit a generalized linear mixed-effects regression model predicting ACT from onset *place* (bilabial, coronal), *phonetic realization* (devoiced, prevoiced, aspirated), and their interaction, along with syllable *nucleus*, speaker *sex*, and the interaction of *place*, *sex*, and *phonetic realization*. Post-hoc pairwise comparisons of the estimated marginal means [23] showed a significant difference of 55 msec in ACTs between devoiced and aspirated plosives for both males ($SE = 2.1, z = 25.2$) and females ($SE = 2.1, z = 26.9$, both $p < 0.001$), but ACTs of devoiced and prevoiced plosives differed only for females and then only by 3 msec ($SE = 0.9, z = 3.6, p = 0.001$). Differences (averaged over sex) between levels of *place* were small but significant for devoiced (9 msec, $SE = 2.9, z = -2.9, p = 0.004$) and prevoiced (7 msec, $SE = 3, z = -2.5, p = 0.01$), but not voiceless (aspirated) plosives.

3.3. F0 trajectories

A generalized additive mixed model (GAMM: [24]) was fit to the speaker-centered F0 values, with *phonetic realization* (prevoiced, devoiced, aspirated, or nasal) of the onset as a predictor variable and smooth terms for F0 trajectory over the open phase for each of the four onset types. Item- and speaker-level differences were captured using factor smooths, analogous to random slopes and intercepts in a linear mixed model. A scaled-t family was selected on account of the heavy-tailed distribution of residuals.

Deviance explained by the GAMM model was 49.6%. Fig. 5 shows the predicted trajectories over the open phase (measured from closure release rather than vowel onset). The difference smooths (Fig. 6) show that while F0 following devoiced [b̥ d̥] and aspirated [pʰ tʰ] are significantly different for most of their trajectories, the same is not true of devoiced [b̥ d̥] and prevoiced [b d]: while there is a small difference immediately following release, the trajectories rapidly converge and remain similar throughout the remainder of the vowel. A similar pattern is observed for the difference smooth between devoiced [b̥ d̥] and the baseline nasals [m n].

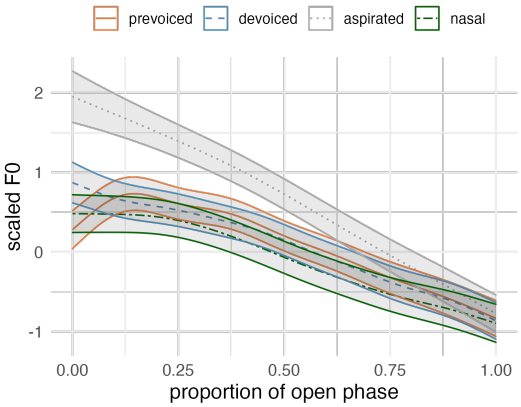


Figure 5: GAMM predicted F0 trajectories by phonetic realization of onset.

4. DISCUSSION

Our results indicate that in CSS citation forms, realization of /b d/ as [b̥ d̥] is common (Q1), particularly (but by no means exclusively) in the speech of female talkers, a trend previously reported elsewhere [4, 5], cf. [1, 2, 3]. However, devoicing of /b d/ did not substantially affect either the realization of ACTs (Q2) nor of the following F0 trajectories (Q3). Barring a small perturbation immediately following closure release (also documented for English by [25]), F0 trajectories following devoiced [b̥ d̥] are comparable to those following prevoiced [b d] and nasal [m n], consistent with previous reports of the (lack of) effect of devoicing on co-intrinsic F0 (e.g. [26].) These findings are what we would expect if the CF0 dichotomy is primarily driven by gestures implemented to suppress voicing during the closure phase of voiceless plosives [13, 15, 16].

Over 50 years ago (at the 7th ICPhS), Jan Lindqvist presented fiberoptic data showing that while Swedish lenis plosives were often acoustically devoiced utterance-initially, the glottis was nonetheless invariably in a voicing position [4]. The acoustic devoicing we observe here is thus likely the result of a failure to overcome the Aerodynamic Voicing Constraint [6, 27] rather than a categorical difference in articulatory posture. Taken together with Lindqvist's findings, the present results suggest that the Swedish plosive contrast may indeed be phonetically overspecified [1, 2], but they also demonstrate that presence or absence of glottal pulsing during the closure is not a reliable acoustic-phonetic correlate of phonological voicing in this language. Rather, the co-intrinsic F0 dichotomy may be a more reliable indicator, since it persists regardless of whether or not closure voicing is achieved. This may also be true of other so-called "true voicing" languages

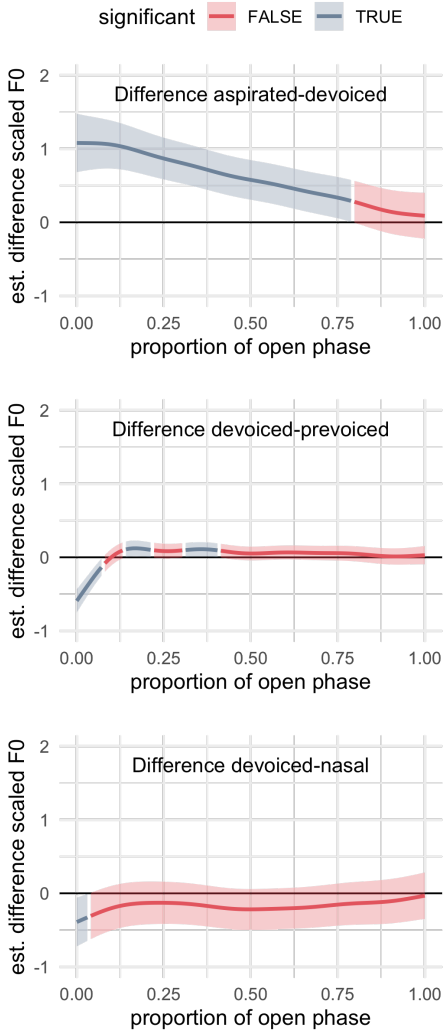


Figure 6: Difference smooths for F0 between different onset types.

in which spontaneous devoicing is common such as Tokyo Japanese [26], Dutch [28, 29], and European Portuguese [30].

5. ACKNOWLEDGMENTS

This work was funded in part by ERC Starting Grant 758605 to James Kirby and a Helge Ax:son Johnson Foundation grant to Maryann Tan. Thanks to Nicole Benker for assistance with data annotation.

6. APPENDIX: WORDLIST

bål, ball, bar, bil, biff, borr, buss, då, damm, dal, dill, dina, ding, dipp, ditt, dopp, dugga, mål, mall, mat, mil, mitt, morr, must, nå, namn, nav, nick, Nina, norr, nunna, Pål, pall, par, pil, piff, porr, puss, tå, tand, tal, till, ting, tipp, titt, topp, tugga, Tina.

7. REFERENCES

[1] Beckman, J., Helgason, P., McMurray, B., Ringen, C. 2011. Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics* 39(1), 39–49.

[2] Helgason, P., Ringen, C. 2008. Voicing and aspiration in Swedish stops. *Journal of Phonetics* 36(4), 607–628.

[3] Karlsson, F., Zetterholm, E., Sullivan, K. P. H. 2004. Development of a gender difference in Voice Onset Time. *Proceedings of the 10th Australian International Conference on Speech Science Technology* 316–321.

[4] Lindqvist, J. 1972. Laryngeal articulation in Swedish. *Proceedings of the seventh International Congress of Phonetic Sciences* 361–365.

[5] Keating, P. A., Linker, W., Huffman, M. 1983. Patterns in allophone distribution for voiced and voiceless stops. *Journal of Phonetics* 11, 277–290.

[6] Ohala, J. J. 1983. The origin of sound patterns in vocal tract constraints. In: MacNeilage, P. (ed), *The Production of Speech*. New York: Springer Verlag, 189–216.

[7] Meyer, E. A. 1897. Zur Tonbewegung des Vokals im gesprochenen und gesungenen Einzelwort. *Phonetische Studien (Beiblatt zu der Zeitschrift die neuren Sprachen)* 10, 1–21.

[8] House, A. S., Fairbanks, G. 1953. The influence of consonant environment upon the secondary acoustic characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1), 105–113.

[9] Di Cristo, A., Hirst, D. J. 1986. Modelling French micromelody: Analysis and synthesis. *Phonetica* 43(1-3), 11–30.

[10] Kirby, J., Ladd, D. R. 2016. Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages. *Journal of the Acoustic Society of America* 140(4), 2400–2411.

[11] Dmitrieva, O., Llanos, F., Shultz, A. A., Francis, A. L. 2015. Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics* 49, 77–95.

[12] Kingston, J., Diehl, R. L. 1994. Phonetic knowledge. *Language* 70(3), 419–454.

[13] Hanson, H. M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America* 125(1), 425–441.

[14] Honda, K., Hirai, H., Masaki, S., Shimada, Y. 1999. Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech* 42(4), 401–411.

[15] Halle, M., Stevens, K. N. 1971. A note on laryngeal features. *MIT Research Laboratory of Electronics Quarterly Research Report* 101, 198–213.

[16] Löfqvist, A., Baer, T., McGarr, N. S., Story, R. S. 1989. The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America* 85, 1314–1321.

[17] Winkelmann, R., Harrington, J., Jänsch, K. Sep 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech Language* 45, 392–410.

[18] Mikuteit, S., Reetz, H. 2007. Caught in the ACT: The timing of aspiration and voicing in East Bengali. *Language and Speech* 50(2), 247–277.

[19] Bombien, L., Winkelmann, R., Scheffers, M. 2022. wrassp: an R wrapper to the ASSP Library. R package version 1.0.2.

[20] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[21] Borin, L. 2014. The Swedish Parole corpus. http://hdl.handle.net/11372/LRT-421.

[22] Bartoń, K. 2022. MuMIn: Multi-model inference. R package version 1.47.1.

[23] Lenth, R. V. 2022. emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.2.

[24] Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. New York: Chapman and Hall/CRC 2 edition.

[25] Xu, Y., Xu, A. 2021. Consonantal F0 perturbation in American English involves multiple mechanisms. *The Journal of the Acoustical Society of America* 149(4), 2877–2895.

[26] Gao, J., Arai, T. 2019. Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics* 77, 100932.

[27] Ohala, J. J. 2011. Accommodation to the aerodynamic voicing constraint and its phonological relevance. *Proceedings of The 16th International Congress of Phonetic Sciences* Hong Kong 64–67.

[28] van Alphen, P. M., Smits, R. 2004. Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. *Journal of Phonetics* 32(4), 455–491.

[29] Pinget, A.-F., Kager, R., Van de Velde, H. 2020. Linking variation in perception and production in sound change: evidence from Dutch obstruent devoicing. *Language and Speech* 63(3), 660–685.

[30] Pape, D., Jesus, L. M. 2015. Stop and fricative devoicing in European Portuguese, Italian and German. *Language and Speech* 58(2), 224–246.

MORPHO-PHONOLOGICAL EFFECTS ON THE PHONETIC CHARACTERISTICS OF TENSE CONSONANTS IN KOREAN COMPOUNDS

Chloe Dokyung Kwon, Sam Tilsen, John Whitman

Cornell University
dk837@cornell.edu, tilsen@cornell.edu, jbw2@cornell.edu

ABSTRACT

This study examines the phonetic characteristics of Korean word-medial stops in two morphological contexts - compound and simple nouns - and probes the realization of compound tensification, a process whereby plain medial stops surface as tense consonants. Using utterances extracted from a large-scale Korean conversational corpus to see what speakers actually do, phone duration, F0, and spectral tilt were measured. These were used to investigate the differences in the phonetic patterns of plain and tense stops, described to be categorically distinct regardless of morphological context. However, the results show that morphological context does play a role. Specifically, the plain medial stops in compound nouns exhibit phonetic characteristics in the direction expected for tense stops, while the compound tense medial stops exhibit the characteristics in the direction expected for plain stops. The results call into question the assumption that compound tensification results in categorically distinctive tense consonants.

Keywords: phonetics, morphophonology, Korean, corpus study, stop consonants

1. INTRODUCTION

Korean obstruents show a typologically unusual three-way laryngeal contrast not including a voicing distinction: plain (lenis), tense (fortis), and aspirated [1]. Plain consonants are characterized by slight aspiration word-initially, tense consonants are characterized by tenseness in the vocal folds, and aspirated consonants are characterized by strong aspiration [2]. Acoustic correlates of these characteristics are well documented, including duration [3], [4], [5], [6], F0 [2], [4], [7], and spectral tilt [2], [8]. It is still uncertain how different morphological contexts affect the distribution of the three-way contrast and its implications for the morphological and phonological representations of words.

Korean compound tensification or *sai-sios* exemplifies the ambiguity in the relationship between the morpho-phonological status of words and their phonetic realizations. One of the common types of Korean nominal compounds is a subordinate compound [1]. It consists of two component words, of which the meaning of the first (W_1) modifies the meaning of the second (W_2), the head. Compound tensification refers to the process whereby a plain obstruent, onset of W_2 in a subordinate compound, becomes tense.

	W_1		W_2	
(1)	/pom	+	**p**i/	
	[pom		**p***i]	‘spring rain’
	spring		rain	
(2)	/pom	+	**p**at^h/	‘spring field’
	[pom		**b**at]	
	spring		field	

The historical account is that compound tensification is triggered by insertion of the reflex of a historical genitive marker, attested as a genitive particle *=s* in middle Korean [1]. This triggers post-obstruent tensification in the following plain obstruent [9], [10], causing the plain stop to surface as tense.

However, the occurrence of tensification is not fully predictable. For example, both (1) and (2) are subordinate compounds with the same component word preceding the bilabial stop in the target position of tensification. Korean dictionaries [11], [12] describe the target consonant in (1) as a tense stop (transcribed with an asterisk) and the one in (2) as a plain stop, which likely undergoes intersonorant voicing given the phonological context [7]. The way these forms are produced does not always agree with what is in the dictionary, as the occurrence of tensification tends to vary across speakers. For instance, [13] surveyed 35 native speakers of Yanbian Korean who were asked to indicate whether they would use a tense or plain pronunciation for 1,171 compound nouns. She found that the responses were variable by speaker, such that all 35 speakers agreed on only four of the words with respect to tensification.

To account for the variation, previous studies considered several possible factors. One of them is the frequency of the whole compound, which was found to be positively correlated with the occurrence of tensification [13], [14], [15]. Another factor is the segment type of the coda in the first component word. The likelihood of tensification increased in the order of vowel, liquid, nasal, and obstruent [13], [14]. The place of articulation of the target consonant also

showed an effect, with coronal consonants experiencing tensification more frequently [14].

Although the above studies provide insight into the complex nature of compound tensification, there are two main unresolved issues that the present study aims to address. First, previous studies rely on linguistic judgments to infer phonological representations rather than directly investigating the realizations. In this study, this is done by systematic phonetic analyses of the expected acoustic correlates of the three-way contrast. Second, previous studies assume compound tensification is a categorical phenomenon in which the plain and derived tense consonants can be clearly distinguished, and the derived tense consonants are the same as the underlyingly contrastive tense consonants. To investigate whether this is true in terms of phonetic characteristics, this study examines word-medial plain and tense stops in Korean simplex and compound nouns. Acoustic measures are hypothesized to vary not only by the types of consonants (i.e., plain or tense) but also by their morphological contexts (i.e., compound or simplex).

Compound	C	V	{N, L, V}	#	__	V	(C)
Simplex	C	V	{N, L, V}	.	__	V	(C)

Table 1: Schematization of the morphological contexts. The shaded region shows the position of the target consonant.

In Table 1, the shaded region marks the position of the target consonant, which can be a plain or tense bilabial, alveolar, or velar stop. The preceding sound can be a nasal (N), liquid (L), or a non-high vowel (V), whereas the vowel following the target consonant were controlled to be non-high vowels. Based on prior acoustic studies, we expect tense consonants to be longer than plain consonants [3], [4], [5], [6]. F0 and spectral tilt measures on this vowel were predicted to be higher and lower, respectively, for tense consonants [2], [4], [7], [8]. The acoustic measures are hypothesized to vary depending on whether the sounds occur in a compound noun due to compound tensification, in which the plain consonants are expected to behave more similarly to tense consonants.

2. METHOD

2.1. Data

The speech data was extracted from the Korean Conversational Speech Corpus [16], which consists of conversations between two speakers on a common topic. Each conversation lasts for approximately 15 minutes, totaling about 500 hours of data. The corpus records data from 2,739 native speakers who speak different regional varieties and belong to various age groups. The data comes in a total of 870,682 audio files and 2,232 text transcription files. The transcriptions include speaker metadata such as age and the region of principal residence.

2.2. Data processing

The search for simplex and compound nouns was carried out using the Standard Korean Language Dictionary [11]. It provides information about the orthography, pronunciation after the phonological rules have been applied, and the position of the syllable and morpheme boundaries. From this dictionary, the nouns were extracted if (1) they are two to three syllables long, (2) they are native words, (3) they occur with a plain or a tense stop in the onset of the second or third syllable for simplex nouns or the onset of the second component word for compound nouns, (4) these stop sounds are followed by a non-high vowel (/a, ʌ, e/), and (5) the stops are not preceded by high vowels or obstruents. After compiling a list of the nouns, the phrases were parsed into words using the Kiwi parser [17] and kept only the words that appear at least once in the corpus. Regarding compound types, subordinate compounds were separated from coordinate compounds, which is another common type described not to exhibit compound tensification but were set aside for future analysis. The process resulted in 60 subordinate compounds (626 tokens) and 10 simplex nouns (1935 tokens).

The speech data was force-aligned with Montreal Forced Aligner [18]. After segmentation, the following measures were extracted and processed: the duration of the target consonant, fundamental frequency (F0) of the following vowel, and spectral tilt of the following vowel. The duration was measured using the segmentation information, and 219 (8.5%) outliers were removed. For F0 analysis, the values were extracted over the vowel following the target consonant using the autocorrelation method in Praat [19]. These were then processed to discard outliers and smoothed using a smoothing spline method with the spline parameter set to 9.9×10^{-4}. During this process, 247 tokens (9.6%) were discarded. The onset of F0 was used for the analysis.

For spectral tilt, the differences in the amplitudes of the first harmonics and second harmonics (H1-H2) were measured using Voicebox in MATLAB [20]. The processing involved down-sampling the audio signal to 11kHz for LPC analysis [21] and creating 20-ms frames with the step size of 1-ms. After

applying the hamming window and estimating LPC filter coefficients, the inverse-filtered frames were used to estimate the narrowband spectrum. Fifty of the spectra were randomly selected and plotted individually to check whether the first and second harmonics fall within the reasonable range. 327 tokens (12.8%) were removed after this process. The analysis used the average of H1-H2 lying within the first 30% of the vowel interval.

2.3. Analysis

Linear mixed-effects modeling (LME) was implemented with *fitlme()* in MATLAB and was utilized to examine the phonetic realizations of plain and tense consonants in different morphological contexts. The dependent measures are the duration of the stop consonants, F0 of the following vowel, and the spectral tilt of the following vowel. The main predictors are the types of stop consonants (Plain or Tense) and the morphological contexts (Simplex or Compound). The interaction effect of these two predictors were included in the model to examine whether the effects of morphological contexts vary by the types of the stop consonants. Gender (Female or Male) was also considered for the F0 analysis. The model included individual subject and word items as random effects.

3. RESULT

3.1. Target consonant duration

Target consonant duration patterned consistently with the hypothesis that it would vary by the consonant types and morphological contexts. This is illustrated in Figure 1 which shows the differences in the distributions of the duration. The boxes are grouped by the types of consonants, with the blue box on the left representing those in simplex nouns and the orange box on the right representing those in compound nouns. The duration is longer for the tense consonants than the plain consonants, but the duration is shorter for tense consonants in compound nouns. This is opposite from the pattern of the duration of the plain consonants that become longer in compound nouns.

The effect of the consonant type is confirmed by the LME result (Plain: -45.8 $\pm$ 19.2, t(2338)=-2.4, $p<0.05$). There was no significant effect of the morphological context alone but an interaction effect was found between the types of consonants and the morphological contexts (Plain and Compound: 42.7$\pm$20.2, t(2338)=2.1, $p<0.05$). The duration of the plain consonants was longer in compound nouns while the duration of the tense consonants was shorter in compound nouns.

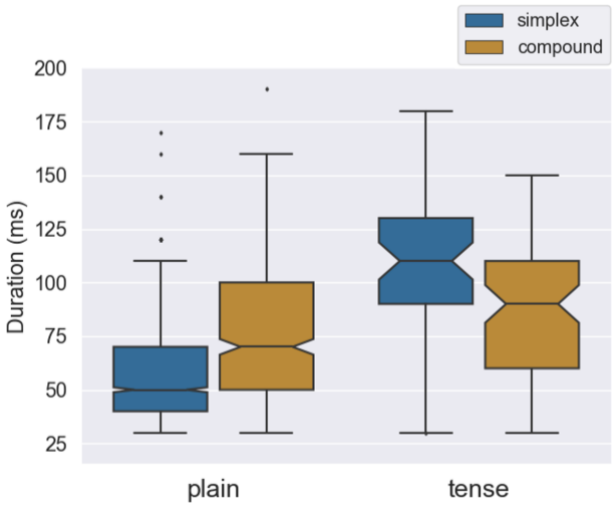


Figure 1: Distribution of phone duration of the target consonants

3.2. F0 of following vowel

Previous studies report that F0 values on the vowel following tense consonants are higher than those of the vowel following plain consonants [2], [4], [7]. Most studies focus on the consonants in the word-initial position, but the results from this study suggest that the pattern also arises in word-medial consonants in simplex nouns.

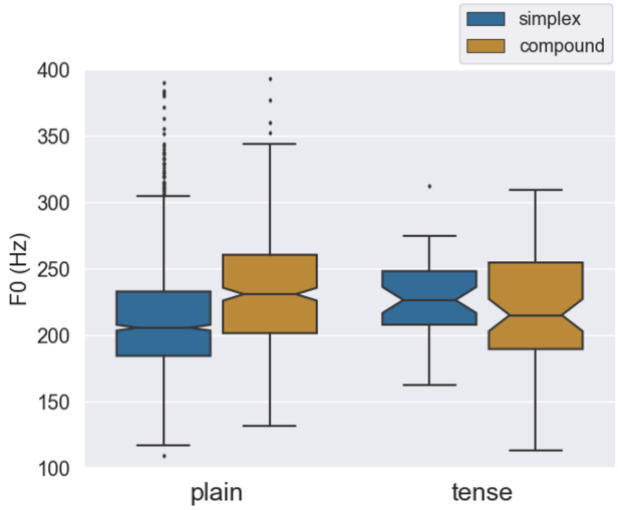


Figure 2: Distribution of F0 from female speakers

However, the plain consonants showed higher F0 than the tense consonants in compound nouns, as shown in Figure 2. Only the effect of gender was confirmed by the LME analysis.

3.3. H1-H2 of following vowel

H1-H2 was measured by comparing the difference between the amplitudes of the first and the second harmonics (H1-H2). It is measured to distinguish between breathy and creaky voicing followed by

plain and tense consonants, respectively [2], [8]. The vowels following plain consonants are likely to show higher H1-H2 since the articulation of plain consonants tend to result in breathy vowels. On the other hand, the vowels following tense consonants are likely to be creaky, which would result in lower H1-H2.

Figure 3 shows that the plain and tense consonants exhibit similar H1-H2 values. What is striking is the difference within the tense consonants across the morphological conditions. Tense consonants in compound nouns show higher H1-H2 than the level of H1-H2 for plain consonants. The LME result shows that the differences between the types of consonants and the morphological contexts are significant (Plain: 1.6 ± 0.7, $t(2298)=2.3$, $p<0.05$; Compound: 2.1 ± 0.8, $t(2298)=2.6$, $p<0.01$), as well as their interaction effect (Plain and Compound: -1.9 ± 0.8, $t(2298)=-2.3$, $p<0.05$).

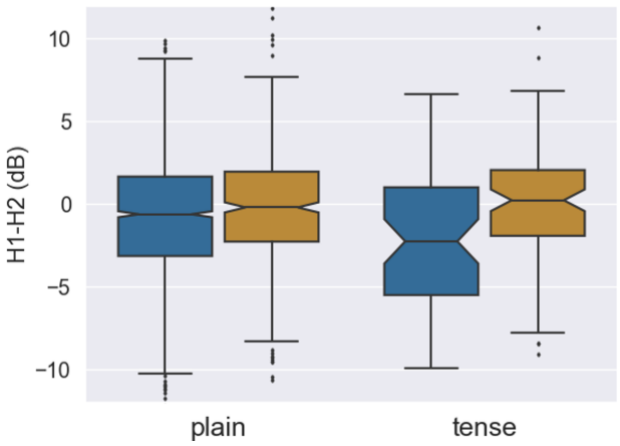


Figure 3: Distribution of averaged H1-H2 over the 30% of the vowel interval following the target consonants

4. DISCUSSION

The results show that phonetic characteristics of the consonants vary not only by consonant type but also by morphological context. Specifically, the duration of the plain consonants in a compound noun showed longer duration than those in a simplex noun. Also, although the differences were not significant, the direction of the effect on F0 was consistent with the hypothesis: it was higher for the plain consonants in a compound noun. It suggests that the plain consonants in compound nouns may be exhibiting characteristics of tense consonants as a result of compound tensification. However, it needs to be further investigated for more conclusive evidence.

All three phonetic measures showed that the tense consonants exhibit different phonetic characteristics depending on the morphological context. Interestingly, the measures patterned in the direction expected for the plain consonants. For example, the duration was longer, F0 was lower, and H1-H2 was higher for tense consonants in compound nouns compared to those in simplex nouns, which are characteristics expected for a plain obstruent in Korean. It seems that the tense consonants are almost neutralized to a plain consonant in a compound noun. It also indicates gradience in the phonetic details of a consonant type. The source of this gradience is a topic for further investigation.

This study offers an important insight into the nature of phonological processes through the examination of naturalistic data. While many studies explore phonetic implementation of phonological features or processes, not many have investigated the exact relation between the phonetic and phonological representations together with the role of the morphological factors. This is especially true in the context of compound tensification in Korean where the orthographic or metalinguistic representation have been taken for granted. The results support the importance of morphological contexts in predicting the phonetic patterns of consonants, which warrants further investigation in a more controlled production study. Another important dimension is to look at how these representations are processed and perceived by the listeners, which will add further insight into how the phonological knowledge is stored in the mind of the speakers/listeners.

While the corpus data set a solid ground for a preliminary analysis, it also came with variation from speakers and other contexts that were difficult to be controlled. For one, factors such as the category of the preceding sounds had to be excluded from the model since it correlated with one of the main predictors, the morphological context. Obtaining a more balanced set of observations from an experiment will allow systematic investigation of inter- and intra-speaker variation. It will also enable examining the difference between the frequency effects and lexicalization. A random-slope analysis will be performed on the more controlled dataset for reliable statistical results.

5. ACKNOWLEDGEMENTS

We thank Abigail C. Cohn for her detailed feedback in developing the ideas and improving this paper. We also thank all the members of Cornell Phonetics Lab for their time and input. Finally, we thank two anonymous reviewers for their thoughtful and helpful comments.

6. REFERENCES

[1] S. Cho and J. Whitman, *Korean: A linguistic introduction*. Cambridge University Press, 2020.

[2] T. Cho, S.-A. Jun, and P. Ladefoged, “Acoustic and aerodynamic correlates of Korean stops and fricatives,” *J. Phon.,* vol. 30, pp. 193-228, 2002.
[3] J.-I. Han, “The phonetics and phonology of “tense” and “plain” consonants in Korean,” Ph.D dissertation, Cornell Univ., Ithaca, NY, 1996.
[4] M.-R. C. Kim, “Acoustic characteristics of Korean stops and perception of English stop consonants,” Ph.D dissertation, Univ. of Wisconsin, Madison, 1994.
[5] D. J. Silva, “The phonetics and phonology of stop lenition in Korean,” Ph.D dissertation, Cornell Univ., Ithaca, NY, 1992.
[6] T. Cho and P. Keating, “Articulatory and acoustic studies of domain-initial strengthening in Korean,” *J. Phon.*, vol. 29, no. 2, pp. 155-190, 2001.
[7] S.-A. Jun, “The phonetics and phonology of Korean prosody,” Ph.D dissertation, Ohio State Univ., Columbus, 1993.
[8] M. R. Kim, P. S. Beddor, and J. Horrocks, “The contribution of consonantal and vocalic information to the perception of Korean initial stops,” *J. Phon.*, vol. 30, pp. 77-100, 2002, doi: 10.006/jpho.2001.0152.
[9] S.-C. Ahn, “The interplay of phonology and morphology in Korean,” Ph.D dissertation, Univ. of Illinois, Urbana, 1985.
[10] K. Chung, “Neutralization in Korean: a functional view,” PhD dissertation, Univ. of Texas at Austin, 1980.
[11] N. Park, “Stochastic learning and well-formedness judgement in Korean phonotactics,” Ph.D dissertation, Seoul National Univ., Korea, 2020.
[12] “Naver Korean Dictionary,” https://ko.dict.naver.com/#/main.
[13] C. Ito, “Compound tensification and laryngeal co-occurrence restictions in Yanbian Korean”, *Phonology*, vol. 31, no. 3, pp. 349-398, 2014, doi: 10.1017/S0952675714000190.
[14] S, Kim, “Phonologtical trends in Seoul Korean compound tensification, M.A. thesis, Seoul National Univ., Korea, 2016.
[15] K. Zurrow, “Predicting Korean sai-siot: phonological and non-phonological factors,” handout, *21st Japanese Korean Linguistics Conference,* Seoul National Univ., Korea, 2011.
[16] The National Institute of Korean Language, “Ilsang daehwa malmungchi [Converstional speech corpus],” https://corpus.korean.go.kr, 2020.
[17] M. Lee, “Kiwi, Korean Intelligent Word Identifier,” Version 0.14.0, 2022, doi: 10.5281/zenodo.7041425.
[18] M. McAuliffe, M. Socolof, E. Stengel-Eskin, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner”. Version 1.0. 2017.
[19] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot Intl.*, vol. 5, pp. 341-345.
[20] “Voicebox: Speech processing toolbox for MATLAB,” http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.
[21] J. D. Markel and A. H. Gray, *Linear prediction of speech.* New York: Springer-Verlag, 1976.

POSITIONAL ASYMMETRY IN LENGTH AND VOICING CUES IN THE ALVEOLAR PLOSIVES OF IKEMA MIYAKO RYUKYUAN

Shigeko Shinohara[1], Angélique Amelot[1], Qandeel Hussain[2]

[1] Laboratoire de Phonétique et Phonologie, CNRS/Sorbonne Nouvelle, France, [2] Universität Bamberg, Germany
shigeko.shinohara; angelique.amelot@sorbonne-nouvelle.fr, qandeel.hussain@students.mq.edu.au

ABSTRACT

In Ikema (a Miyako Ryukyuan dialect), obstruents and nasals contrast in length and/or voicing in word-initial and medial positions. Yet, none of them has the four-way distinction in both positions. Our study focuses on asymmetric alveolar plosive contrasts: a triplet of /t tt d/ in the word-initial position and a quadruplet /t tt d dd/ in the medial position. The study investigates relevant temporal and non-temporal acoustic properties of these consonants. The classificatory importance of each acoustic property was also evaluated. The results suggest that many of the acoustic values of the initial triplet are more dispersed in the acoustic space than those of the medial quadruplet. The orders of importance of acoustic parameters showed their distinct roles depending on the positions. Duration-related cues are dominant in both positions; however, voice quality, intensity and f0 have a more important role in the initial position.

Keywords: Initial geminate, laryngeal contrast, Ryukyuan, Ikema, alveolar plosive.

1. INTRODUCTION

Ikema is a dialect of Miyako Ryukyuan (a Japonic language) spoken by around 1300 elderly people in Ikema island, the Sarahama area of Irabu and the Nishihara area of Miyako islands of Japan [11, 21, 34]. The obstruent system of Ikema has a rich set of length and voicing contrasts in word-initial and medial positions [29] but with segmental gaps. Based on her fieldwork in the Nishihara area, [11] identifies the following phonemic consonants: /p b t d k g f v s z h ts m n̪ n̥ n r w j/ (symbols are adapted). Among them, the obstruents that appear as word-initial geminates include /tt tts kk ff ss vv zz/ (words starting with /kk/ are very rare) and geminate obstruents that appear in the medial position include /pp kk dd/ on top of the initial geminate set (occurrence of /pp/ or kk/ is very low). Singleton /v/ does not occur anywhere; occurrence of singleton /p/ depends on idiolect. We focus on the asymmetric contrast of alveolar plosives, which remains the most complete set among the obstruents. A four-way distinction /t tt d dd/ is present in the word-medial position ([bata] 'stomach', [tatta] 'more', [hada] 'period' [badda] 'trench') while the initial position limits the contrast to three types /t tt d/ ([tiː] 'hand', [ttiː] 'pipe' [diː] 'cane') lacking /dd/.

In previous acoustic and articulatory studies, full voicing (with prevoicing) was observed in the voiced obstruents in both positions [33, 8]. Concerning the alveolar plosives, longer Voice Onset Time (VOT) is characteristic of initial voiceless singleton /t/ [33, 17]. Thus, in terms of VOT, the initial position hosts three types: a negative valued /d/, a short lagged /tt/ and a long lagged /t/. [17] reported more acoustic results (formants and spectral tilts of the following vowel) on initial obstruents /b t tt d k g ts ([tɕ]) z ([dʑ]) f ff vv s ss z zz/ followed by /a/ in Ikema. Overall, consonant duration (VOT for stops and affricates; total duration for fricatives) is the most reliable acoustic correlate of voicing and length contrasts. As to the non-temporal properties, geminates and voiceless obstruents generally induced higher fundamental frequency (f0) at the onset of the following vowel, with exception of /t/ with a long aspiration, which induced slightly higher f0 than /tt/. In addition, geminates /tt ff ss zz/ were differentiated from singletons by the higher intensity and lower F1/H1-H2. Non-temporal aspects of medial obstruents of Ikema remained unreported.

In this paper, we examine the phonetic properties of alveolar plosives in initial and medial positions in Ikema to explain its asymmetric distribution in the synchronic state. We explore selected acoustic measures proven to be most efficient in the previous study of initial obstruents in Ikema [17]: duration of the consonants and f0, F1, intensity, H1-H2c at the onset of the following vowel (see 3.2 for the detail).

2. THEORETICAL ASSUMPTIONS

Cross-linguistically, consonant and closure durations are known to be the most salient cues for length [9, 25, 10, 18] and voicing [27, 16] distinctions. However, there may be many other cues. Voicing is well-known to impact the microprosody of the following vowel [13, 27] to the extent of triggering tonal changes in some languages [12]. An increasing number of studies are confirming the impact of geminates on f0 onsets [1, 10]. Other acoustic parameters are also

influenced by gemination: lower F1 [20, 2], higher intensity [1, 10] and arguably smaller H1-A1 [19]. In contrast, the VOT difference between geminates and singletons may be variable among languages [28, 3]. VOT differences, in turn, may influence f0 [22]. Some of the non-durational properties may be used as enhancing cues and their role may not be the same when the phonetic and phonological contexts differ [32].

Initial obstruent length is hard to perceive [31, 23] because there is no preceding sound that conveys various acoustic cues [20, 16]. Especially, plosives are the most handicapped consonant type in the initial position for length distinction because the length difference is mainly conveyed by the closure duration in intervocalic position. Without this duration, consonant length may be mostly contrasted on the laryngeal dimension. Cross-linguistic studies suggest that there are typically only three VOT contrasts [6, 14]. We examined the release duration of the plosives and four non-temporal values at the onset of the following vowel in the initial position to determine if they are used differently than in the medial position.

A functional assumption, the Dispersion theory presupposes that maximally distant acoustic dispersion assures the perceptibility of the contrastive elements [7]. In this regard, given the asymmetric numbers of alveolar plosives, their acoustic realisations may be different between the positions, despite the same distinctive feature combinations for three of the consonants: /t tt d/.

We shall measure, thus, the selected acoustic values in the two positions and the closure duration of the medial plosives and consider the positional differences. After reporting the acoustic measurements, we shall evaluate the classification power of each of the acoustic parameters.

3. METHODS

3.1. Recording

Ten Ikema speakers (three males born 1952-1959, avg. age 61.7; seven females born 1932-1952, avg. age 75.6) were recorded in 2018 at a participant's house or a local community centre.

A word list of singleton and geminate obstruents was created. This study uses only seven words involving alveolar plosives out of the whole corpus: [ta:] 'rice field', [tta] 'tongue', [dariru] 'being tired', [bata] 'stomach', [batta] 'grasshopper', [bada] 'trench' and [badda] 'side.' A Sony Linear PCM-M10 digital voice recorder with an integrated microphone was set on the table to record the audio sounds (44.1 kHz, encoded in 16 bit). Participants were introduced to the word list and instructed to produce each word in citation form. The Ikema syntax allows for *pro*-drop and bare forms in the phrase-initial position, creating a natural language usage situation in citation forms. During the session, an equivalent of each target word was presented orally in standard Japanese, and participants produced each target word three times. Some participants provided additional repetitions that were included in the final analysis, while other items had to be discarded due to poor quality, such as uneven tempo or creaky voice.

3.2. Analyses

A total of 227 tokens were included in the analyses (word-initial: 104; word-medial: 123). The data were segmented and analysed in Praat [4]. The release interval of the plosives was segmented from the beginning of the release burst to the first glottal pulse of the following vowels. (Fig. 1). The closure duration of word-medial consonants was defined as the temporal interval from the offset of the preceding vowel to the release of the closure. f0, F1, intensity (cubic interpolation) and H1-H2c (corrected) at the onset of the following vowels [a] or [a:] were measured. The formants were extracted using LPC analysis over a sliding 25 ms window.

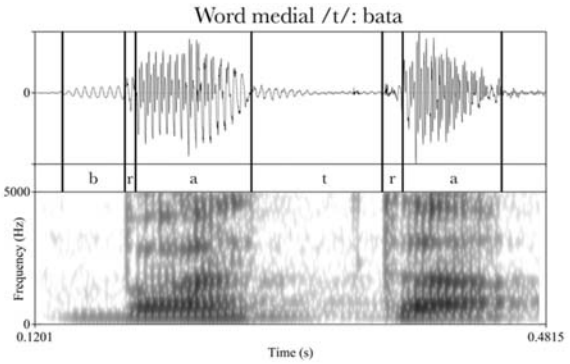


Figure 1: Spectrogram of /bata/ 'stomach.' r refers to the release interval, t is the measured closure duration in this example.

A series of linear mixed effects (LME) models was conducted in R [30] *lmerTest* packages [24]. The *emmeans* package [26] was used to perform pairwise comparisons. In all the LMER models, we performed pairwise comparisons, which included all the words (α value: p = 0.05) but run separately for each position. For all the LME models, Speaker was included as a random factor, Voicing (voiceless and voiced) and Length (singleton and geminate) as fixed factors. The following measurements were included as dependent variables: release duration of the plosives (and closure duration of the medial ones) (ms), f0 (Hz), F1(Hz), intensity (dB), H1-H2c (dB). All the non-temporal acoustic measures were z-score

normalised. Classification of the target segments was performed using the *rpart* package [35] and predict() function in R [30]. Variable importance was calculated to further investigate the ranking and role of each acoustic measure in the classification of voicing and length contrasts [15].

4. RESULTS

The results of acoustic measurements are illustrated in Figs. 2-7 and the main results of the pairwise comparisons among the consonants are summarised for each measurement. The symbol '>' indicates where α value appeared in the difference. ',' indicates where no such difference appeared between the consonants.

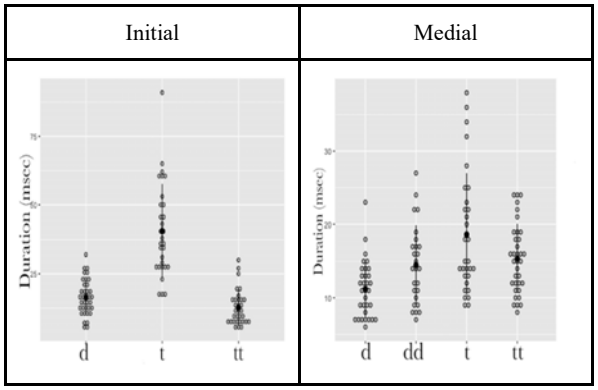


Figure 2: Release duration of the consonants.

The release duration of /t/ was longer than the others in both positions. That of /d/ in the medial position was shorter than the rest. (Initial: t > tt, d; Medial: t > tt, dd > d).

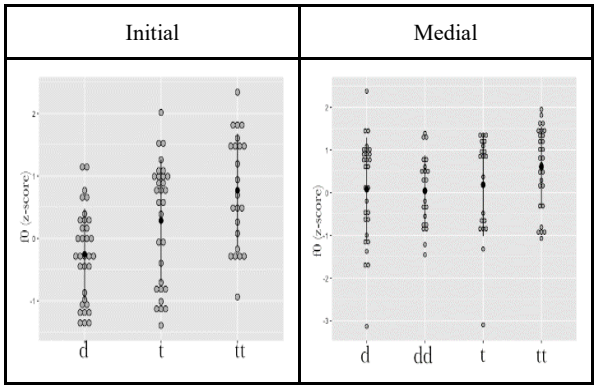


Figure 3: f0 (z-score) measured at the onset of the vowel.

The geminate had higher f0 than the singletons and the voiceless consonants induced higher f0 than the voiced counterpart in the initial position (tt > t > d) while no difference was observed among any pairs in the medial position.

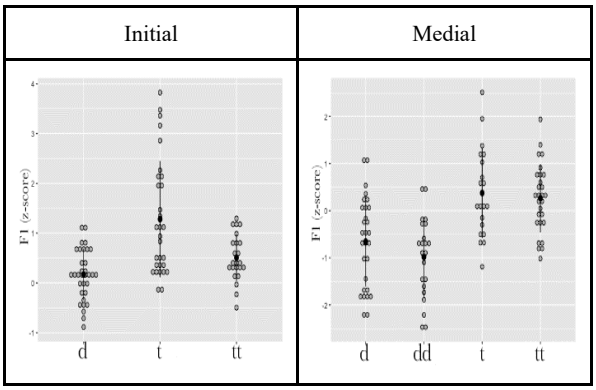


Figure 4: F1 (z-score) measured at the onset of the vowel.

/t/ had higher F1 than its geminate and voiced counterparts in the initial position (t > tt, d). No difference was found between the members of the voiceless pair in the medial position (t, tt > d > dd).

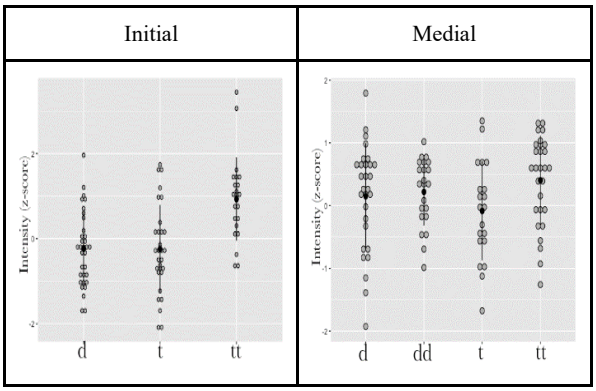


Figure 5: Intensity (z-score) measured at the vowel onset.

Voiceless geminate had higher intensity than its singleton counterpart in both positions (tt > t). The voiced pair in the medial position did not show any difference. (Initial: tt > d, t; Medial: tt, dd, d > t).

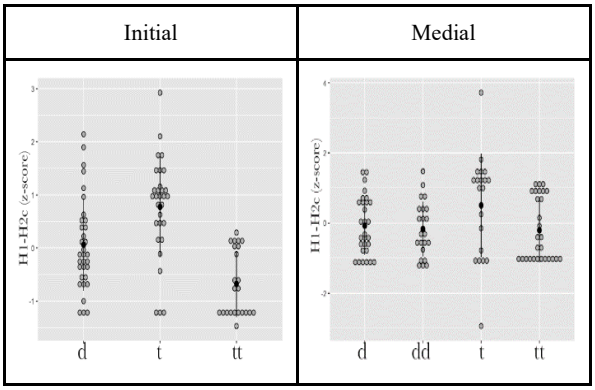


Figure 6: H1-H2c (corrected / z-score) measured at the vowel onset.

The initial triplet was separated by H1-H2c (t > d > tt) but it was not the case in the medial position.

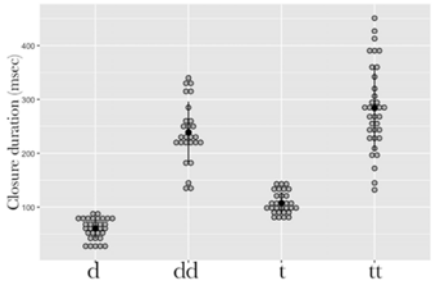


Figure 7: Closure duration of medial consonants.

The closure duration of the medial consonants was longer in geminates than in singletons and voiceless ones than in voiced ones. (tt > dd > t > d).

The different orders in the acoustic measurements imply different cue weights in each position. To assess the importance of each acoustic cue in each position, we conducted an additional test as outlined in section 3.2. Figs. 8-9 display the results using a unitless scale.

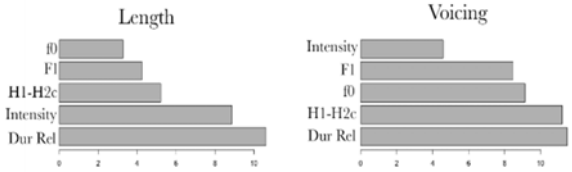


Figure 8: Importance of acoustic parameters for the initial plosives in Ikema.

In terms of length distinction among the initial triplet, intensity is the second most important cue after release duration (Dur Rel). Regarding voicing, voice quality (H1-H2c) along with release duration are the most significant cues.

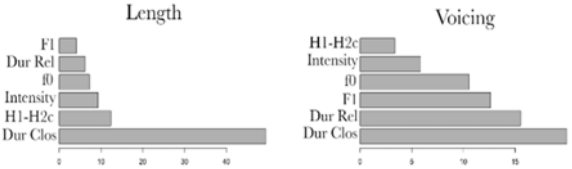


Figure 9: Importance of acoustic parameters for the medial plosives in Ikema.

The analyses for the medial quadruplet included closure duration (Dur Clos). For length distinction, closure duration is by far the most important cue. In terms of voicing, closure duration is also more significant than any other attributes of the consonant.

5. DISCUSSION

We investigated acoustic properties of the length and voicing contrasts in the asymmetrically distributed alveolar plosives of Ikema (/t tt d/ initially vs. /t tt d dd/ medially). Dispersion theory predicts that the acoustic space would be used to optimise the perception of the contrasts. Figs. 2-6 illustrate that acoustic values of each consonant category are mostly more dispersed in the initial position and more clustered in the medial position. A larger difference between the minimum and maximum values in certain measures (f0, Intensity, H1-H2c) between the two positions may be partly a by-product of the domain initial strengthening [5, 10], however, different variable importance between the positions cannot be explained by this effect alone.

The variable importance analysis revealed the predominance of the closure duration in the medial plosives, while release duration is the most important variable for the initial plosives. Non-temporal properties are more influential in the initial length contrast and are likely also more important in the initial voicing contrast than in the medial ones.

None of the obstruents has a four-way contrast in the initial position in Ikema. Since release duration has been confirmed as the more powerful variable than non-temporal properties in this position, our results are in line with [6] claiming up to three VOT contrasts. For instance, [6] noted that an apparent four-way laryngeal contrast in Pakistani languages (Jangli and Urdu) relies on other phonetic parameters rather than VOT. It is worth noting that in the triplet, /d/ had a mean voicing lead of 75.1 ms (*sd*: 52.3 ms), which could also be useful for its identification. However, it was excluded from the measurements to enable comparison with the medial voiced plosives where closure duration often overlaps with voicing lead in this language. This implies that VOT may play a more decisive role than release duration in the initial triplet. We hypothesise that without context preceding it (*i.e.*, word-initial position), no more than three contrasts can be hosted without introducing contrast in other laryngeal dimensions such as distinctive voice quality or tone on the following vowel. In the case of Ikema, non-temporal properties can be at best characterised as enhancing cues at the synchronic state.

Our study confirmed that the closure duration, the privileged attribute of the medial consonants, clearly separates the four categories (Fig. 7). Adjacent vowel durations are also known to give cues to consonant length and voicing [19, 16]. A future study invites us to study this aspect in Ikema.

6. ACKNOWLEDGEMENTS

We thank our participants and the staff of the NPO Kyuunu Hukarasha for allowing us to use their facilities. The study was supported by Labex EFL.

7. REFERENCES

[1] Abramson, A. S. 1998. The complex acoustic output of a single articulatory gesture: PattaniMalay word-initial consonant length. In *Fourth Annual Meeting of the Southeast Asian Linguistics Society 1994*, Tempe, Arizona, 1–20.

[2] Al-Tamimi, J., Khattab, G. 2015. Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *J. Acoust. Soc. Am*. 138(1), 344–360.

[3] Arvaniti, A., Tserdanelis, G. (2000). On the phonetics of geminates: Evidence from Cypriot Greek, *Sixth international conference on spoken language processing*, 559-562. Beijing: Chinese Friendship Publishers.

[4] Boersma, P. Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.10, retrieved 17 March 2022 from *http://www.praat.org/*

[5] Cho, T., Keating, P. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean, *J. Phon.* 29, 155-190.

[6] Cho, T., Whalen, D. H., Docherty, G. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *J. Phon.*, 72, 52–65.

[7] Flemming, E. S. 2002. *Auditory representations in phonology*. New York: Routledge.

[8] Fujimoto, M., Shinohara, S., Mochihashi, D. 2021. Articulation of geminate obstruents in the Ikema dialect of Miyako Ryukyuan: A real-time MRI analysis. *J. Int. Phon. Assoc.*, 1–25.

[9] Ham, W. 2001. *Phonetic and phonological aspects of geminate timing*. New York: Routledge.

[10] Hamzah, M. H., Hajek, J., Fletcher, J. 2020. Non-durational acoustic correlates of wordinitial consonant gemination in Kelantan Malay: The potential roles of amplitude and f0. *J. Int. Phon. Assoc.* 50(1), 23–60.

[11] Hayashi, Y. 2010. Ikema (Miyako Ryukyuan). In: Shimoji, M., Pellard, T, (eds.), *An introduction to Ryukyuan languages*, 167–188. Tokyo: RILCAA.

[12] Hombert, J.-M., Ohala, J. J., Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, 55(1), 37–58.

[13] House, A. S., Fairbanks, G. 1953. The influence of consonant environment on secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.*, 25, 105–113.

[14] Hussain, Q. 2018. A typological study of Voice Onset Time (VOT) in Indo-Iranian languages. *J.Phon., 71*, 284–305.

[15] Hussain, Q., Kochetov, A. 2022. Acoustic classification of coronal stops of Eastern Punjabi, *Phonetica*, 79(1), 77–110.

[16] Hussain, Q., Shinohara, S. 2019. Partial devoicing of voiced geminate stops in Tokyo Japanese. *J. Acoust. Soc. Am.,* 145(1), 149–163.

[17] Hussain, Q., Shinohara, S. 2020. Acoustic correlates of word-initial singletons and geminates of endangered Ikema, *LabPhon. 17*, UBC.

[18] Idemaru, K., Guion, S. 2008. Acoustic covariants of length contrast in Japanese stops, *J. Int. Phon. Assoc.* 38, 167–186.

[19] Idemaru, K., Guion, S. 2010. Relational timing in the production and perception of Japanese singleton and geminate stops, Phonetica 67(1-2), 25-46.

[20] Kawahara, S. 2006. A faithfulness ranking projected from a perceptibility scale: the case of [+Voice] in Japanese.Language *82:* 536–574

[21] Kibe, N. 2019. *General study for research and conservation of endangered dialects in Japan: Research report on Miyako Ryukyuan*. NINJAL *Collaborative Research Project Reports 12–02*.

[22] Kirby, J. P 2018. Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages, *J.Phon., 71*, 326-354.

[23] Kraehenmann, A. 2001. Swiss German stops: Geminates all over the word. *Phonology* 18, 109–145.

[24] Kuznetsova, A., Per Brockhoff, B., Rune H. B. C. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. R package version 3.1-1. 10.18637/jss.v082.i13.

[25] Lahiri, A., Hankamer, J. 1988. The timing of geminate consonants. *J. Phon,* 16(3), 327–338.

[26] Lenth, R. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.5. https://CRAN.R-project.org/package=emmeans

[27] Lisker, L. 1986. "voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. Language and Speech, 29(1), 3–11.

[28] Muller, J. S. 2003. The production and perception of word initial geminates in Cypriot Greek. *Proc. of the 15th ICPhS,* 1867–1870.

[29] Nakama, H., Takubo, Y., Iwasaki, S., Igarashi, Y., Nakagawa, R. 2022. *Minami Ryukyu Miyako Ikema Hogen Jiten*, NINJAL.

[30] R Core Team. 2022. R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/

[31] Ridouane, R., Hallé, P. 2017. Word-initial geminates from production to perception. In: Kubozono, H. (ed.), *Phonetics and phonology of geminate consonants.* Oxford University Press, 66-84.

[32] Seyfarth, S., Garellek, M. 2018. Plosive voicing acoustics and voice quality in Yerevan Armenian, *J.Phon.,* 71, 425-450

[33] Shinohara, S., Fujimoto, M. 2018. Acoustic characteristics of the obstruent and nasal geminates in the Ikema dialect of Miyako Ryukyuan, In: Babatsouli, E. (ed.), *Cross-linguistic research in monolingual and bilingual speech*, 253-270.

[34] Shinohara, S., Hussain, Q. 2021. Ikema Miyako (Japan) – *Language Snapshot. Language Documentation and Description* 19, 240-248.

[35] Therneau, T., Atkinson, B. 2019. rpart: Recursive Partitioning and Regression Trees [Computer software manual]. (R package version 4.1-15. https://CRAN.R-project.org/package=rpart)

FORTIS STOPS IN POLISH – EVIDENCE FROM ACOUSTIC MEASURES OF VOICE QUALITY

Geoffrey Schwartz, Maral Asiaee, Ewelina Wojtkowiak

Adam Mickiewicz University, Poznań, Poland
geoff@amu.edu.pl, marasi@amu.edu.pl, ewelina.wojtkowiak@amu.edu.pl

ABSTRACT

An acoustic study of the effects of the Polish laryngeal contrast on the phonation of vowels following voiced and voiceless obstruents is presented. Data from 15 native speakers of Polish reveal voicing-based phonation differences in H1*-H2*, H2*-H4*, H4*–H2kHz*, and H2kHz*–H5kHz, providing evidence that phonation properties may help distinguish between Polish laryngeal categories in onset position. Implications for the phonological encoding of the laryngeal contrast are discussed. It is suggested that Polish voiceless stops may be categorized as [fortis], postulated to be an abstract category with a wide range of possible phonetic effects.

Keywords: laryngeal contrasts, phonetics-phonology interface, voice quality, Polish.

1. INTRODUCTION

It is well known that the phonetic realization of two-series laryngeal contrasts is phonetically much more complex than one might infer from phonological descriptions based on features such as [voice] or [spread glottis]. Phonological voicing has been shown to be cued by a large number of acoustic properties, including voice onset time (VOT; [1]), the duration and height of neighboring vowels [2], [3], f_0 in the vicinity of the consonant [4], F1 transition duration [5], noise burst duration or frication duration, as well as the amplitude of stop release bursts, frication, or aspiration. Of these parameters, the phonological literature has focused largely on VOT [6]. On the basis of VOT, it has often been claimed that two-series systems may be divided into two types, commonly referred to as aspiration and true-voice languages. Further, in the approach referred to as laryngeal realism (LR; e.g. [7], [8]) it is suggested that these two types differ with regard to the phonological feature specification that appears, [spread glottis] in aspiration systems as opposed to [voice] in true-voice systems.

If voicing and aspiration systems were truly phonologically distinct, we would expect phonetic distinctions in additional parameters aside from VOT. However, phonetic studies of parameters beyond VOT often reveal parallel behavior between voicing and aspiration systems. The voicing effect on f_0, by which pitch is higher in the vicinity of voiceless obstruents, is an example. In American English, [9] found that voiceless onsets raise f_0, while voiced obstruents have no effect, relative to the baseline produced by nasal onsets. Similar findings have been obtained for voicing languages [10], [11] including French, Italian, and Polish. F1 transition duration, in which clipped transitions after fortis consonants induce higher F1 measurements in non-high vowels, has been shown to behave in a parallel fashion in English [5] and Polish [11]. In short, the current state of research into the phonetics of two-series laryngeal systems has not identified a single guiding phonetic feature upon which phonological categories may be constructed.

In this connection, recent research has begun to identify additional phonetic correlates, based on phonation, which may play a role in maintaining contrast in two-series laryngeal systems (e.g. [12], [13]). Most of this research has come from varieties of English and arose out the observation that voiceless obstruents, most commonly the plosive /t/, frequently undergo glottal reinforcement or even glottal replacement. This glottalization is typically associated with voiceless obstruents in English, while it is for the most part absent in the case of voiced obstruents (but see [14]). Acoustically, glottalization is often described in terms of acoustic measures of voice quality, including a number of measures quantifying spectral tilt. Since glottalization weakens the first harmonic of the vocal wave, glottalized consonants typically show more gradual dropoffs in spectral tilt and lower values for acoustic measures comparing the amplitude differences between lower and higher harmonics. Notably, [15] found perceptual differences between glottalization induced by voiceless consonants and creaky voice that often appears phrase-finally.

The research connecting phonation type with laryngeal contrast discussed above focuses on post-vocalic or coda position in syllables. Thus, for English, experimental findings have linked aspects of phonation with the perception of vowel duration [13] which in this language has been found to be a primary cue to the contrast [16]. At the same time, given the

overall phonetic complexity of the laryngeal contrasts, we suggest that it is worth investigating the effects of phonological voicing in pre-vocalic or onset positions as well.

Polish is a language in which the laryngeal contrast in the coda position is said to be neutralized, completely or incompletely. In the initial position, the familiar VOT contrast between pre-voiced and short-lag stops is attested. At the same time, the non-VOT cues such as f_0 and F1 transitions, housed on vowels following the consonants, have been shown to play an important role in maintaining laryngeal contrasts [11], [17]. In light of the discussion above, we raise the question of whether acoustic measures of voice quality may also constitute part of the phonetic anatomy of the contrast in Polish between voiced and voiceless consonants. A working hypothesis, on the basis of the findings [11], is that voiceless stops in Polish may be described as [fortis], which may be reflected in smaller values for spectral tilt measures gleaned from neighboring vowels, even if the effect is not dramatic enough to be described as glottalization. In what follows, we present a preliminary acoustic study investigating this question.

2. METHOD

2.1. Participants and recordings

15 monolingual Polish speakers took part in a production study. They were all females, aged 17-38 (median age: 25). While they had some history of learning foreign languages in school, they claimed not to be fluent in any of them.

The materials consisted of a word list of forty-eight monosyllabic and disyllabic words in Polish, each of them beginning with a plosive (24 voiced, 24 voiceless). The dataset was counterbalanced for the place of articulation of the stop (16 labial, 16 coronal, 16 dorsal), and the following vowel was always non-high. The participants were recorded in a quiet room in a language school where they were enrolled in an elementary English course. They were recorded directly onto a laptop, using a head-mounted microphone and a USB interface. The items were shown to the participants on PowerPoint slides.

2.2 Segmentation and extraction of acoustic parameters

The acoustic annotation was performed manually using Praat [18] on three separate tiers. Tier 1 contained the consonantal onsets (VOT: positive for voiceless stops, negative for voiced stops, measured according to standard criteria). Tier 2 categorised the onsets into voiceless or voiced. Tier 3 contained the measurements of the vowel only.

Two Praat scripts were used to extract the VOT values from tier 1 and F1 (Bark difference: F1-f_0), and f_0 (in Hz) averages from the first 20% of the vowel from tier 3. VoiceSauce [19] was used to extract voice quality measures from the vocalic part in each word in tier 3. After obtaining all data frames at each 1ms intervals, impossible or incorrect 0 values were omitted.

Four spectral tilt measures, H1*-H2*, H2*-H4*, H4*–H2kHz*, and H2kHz*–H5kHz), were chosen for analysis. These measures allow for the quantification of spectral tilt over a wide range fo the acoustic spectrum [20]. Since in this study we are only looking at phonologically conditioned variation within the confines of modal voice, measures of periodicity such as Cepstral Peak Prominence or Harmonics to Noise Ratio, which quantify the presence of noise in the vocal wave and are more effective in distinguishing between modal and non-modal voices [20], are not analysed here.

2.3. Statistical analysis

All statistical analyses were performed using R version 4.0.4 [21], while visualisations of the data were made using JASP [22]. Since the data were not normally distributed ($p<.05$), the Mann-Whitney U test was performed to find out whether spectral tilt values in vowels differ after voiced/voiceless obstruents.

3. RESULTS FOR TRADITIONAL LARYNGEAL PARAMETERS

We start by reporting results for acoustic parameters traditionally associated with the laryngeal contrast.

As far as VOT is concerned, the results are summarised in Table 1.

Table 1: mean VOT measures of Polish monolinguals

Series	VOT (ms)	Std Dev
voiceless	41.70ms	16
voiced	-92.05ms	26

Table 2 presents the average values of F1 (in Bark) and f_0 (in Hz) after voiceless- and voiced-initial onsets.

Table 2: mean F1 and f_0 measures of Polish monolinguals

Parameter	Onset	Average
F1-f_0	voiceless	4.9 Bark
	voiced	4.4 Bark
f_0	voiceless	195 Hz
	voiced	186.6 Hz

The results from the traditional acoustic parameters associated with the laryngeal contrast may be summed up as follows. In line with previous studies, VOT in Polish voiceless stops is in a range approaching what [23] would call 'slightly aspirated', while prevoicing in voiced stops is very common. With regard to correlates deriving from the vowel following the stops, the effects of underlying voicing were detectable in both the f_0 (difference between voiceless and voiced onsets of 8.4Hz; p=.019) and F1 (difference between voiceless and voiced onsets of 0.5 Bark; p=.003) parameters.

4. RESULTS OF VOICE QUALITY MEASURES

Figures 1-4 show raincloud plots of the four spectral tilt measures analysed here. From the plots it is clear that the voicing induced differences are relatively subtle and small in magnitude, as might be expected for within-modal-voice variation, as opposed to modal-non-modal contrasts.

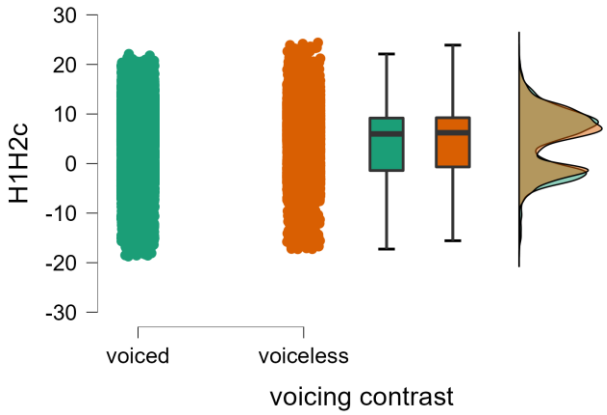


Figure 1: The difference between voiced and voiceless onsets in the H1*-H2* measurement.

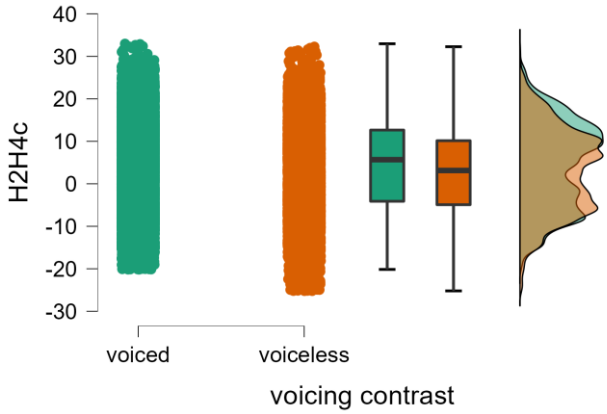


Figure 2: The difference between voiced and voiceless onsets in the H2*-H4* measurement.

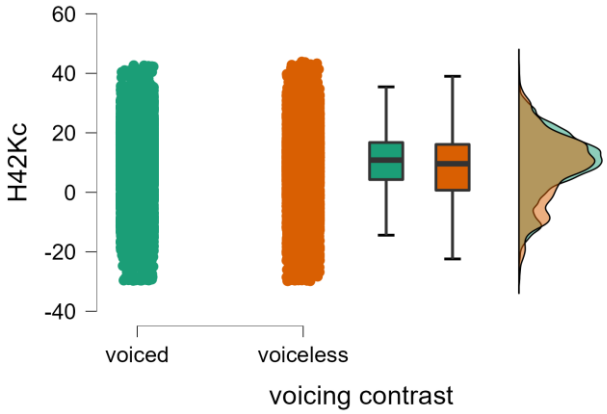


Figure 3: The difference between voiced and voiceless onsets in the H4*-H2kHz* measurement.

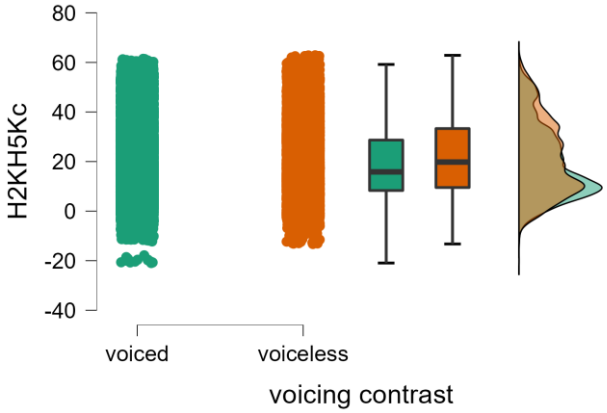


Figure 4: The difference between voiced and voiceless onsets in the H2kHz*-H5kHz* measurement.

The median and Interquartile range (IQR) of voice quality parameters are presented in Table 3.

Table 3: Descriptive data exhibiting median and (IQR) of spectral tilt measures

Spectral tilt measures	Voiceless	Voiced
H1*-H2*	6.22 (9.93)	5.98 (10.6)
H2*-H4*	3.14 (15.1)	5.68 (16.8)
H4*–H2kHz*	9.60 (15.4)	10.8 (12.5)
H2kHz*–H5kHz	19.8 (23.8)	15.8 (20.4)

Results from the Mann-Whitney U test revealed that the median of all parameters depicting the harmonic source spectral shape were statistically different (p<.001) between the vowels following the voiced and voiceless obstruents.

5. DISCUSSION

The present study investigated the effects of obstruent voicing on voice quality in Polish. Our working hypothesis was that voiceless consonants in Polish should be characterized as 'fortis', and as such should induce a slightly stiffer voice quality on the following vowel than voiced consonants. We presented results

for four spectral tilt measures, H1*-H2*, H2*-H4*, H4*–H2kHz*, and H2kHz*–H5kHz. In all four measures, we observes significant differences resulting from underlying voicing. Thus, overall, the results provide evidence for the claim that voice quality is indeed an acoustic correlate of voicing in Polish.

With regard to our working hypothesis, results from two of the measures were compatible with the 'voiceless as fortis' hypothesis, while results for the other two measures showed a difference in the opposite direction. For the H1*-H2* measure, which compares the difference in amplitude between the first and second harmonics of the vowel wave, voiced consonants induced smaller measures, indicating a stronger H2 relative to H1. Likewise, the H2kHz*–H5kHz measure was slightly lower after voiced consonants, indicating a stronger high-frequency element in the spectrum that is expected in stiffer voice qualities. By contrast, the H2*-H4* and H4*–H2kHz* measures both showed signs of stiffer phonation after voiceless obstruents. The H2*-H4* and H4*-H2kHz results provide additional support for the 'voiceless as fortis' hypothesis, which originally arose out of the f_0 and F1 results described by [11].

Overall, the voice quality results suggest a situation in which the bottom and the top of the spectrum are indicative of slacker phonation, while the middle of the spectrum is indicative of stiffer phonation. In this situation, a comparison of the relative merits of the different acoustic measures of voice quality is warranted The H1*-H2* measure describes amplitude of harmonics at the very bottom of the acoustic spectrum, typically in the area below 300Hz. Perceptual sensitivity to amplitude differences in this frequency range is relatively limited, as is visible in equal loudness curves [24]. H2*-H4*, by contrast, compares harmonics in a higher frequency range that shows slightly greater auditory sensitivity. Meanwhile H2kHz, the harmonic nearest 2000 Hz, is in the most sensitive frequency range when it comes to amplitude differences. For this reason, it may be hypothesized that the differences in the H2*-H4* and H4*–H2kHz* measures observed here may be more perceptually robust than H1*-H2* in characterizing the effects of voiceless consonants on voice quality. It is also possible that since Polish voiceless stops have been described as 'slightly aspirated', as shown above, they contributed to a more breathy result in some measures but not others.

We also note that the H4*–H2kHz* measure shows an acoustic resemblance to measures of spectral balance that have been shown to distinguish stressed from unstressed syllables [25], [26]. Spectral balance measures typically divide signal up into two bands, 0-1 kHz and 1-5 kHz. More prominent (or stressed) syllables typically have higher relative amplitude in the higher band where perceptual sensitivity is greater. It is also worth noting that the comparison inherent within the H2kHz*–H5kHz measure, between harmonics closest to 2000 Hz and 5000 Hz, falls entirely within the upper band of spectral balance measures. With regard to the split results observed here, in future studies it is worth investigating the hypothesis that some measures of phonation are manipulated for phonological purposes, while others are more crucial for speaker identification.

Overall, our results provide preliminary evidence of a connection between 'fortisness' in Polish stop production and perceptual prominence of the voice quality that appears on the vowel following the stop. In other words, there may be a phonetic connection between the phonatory effects elicited by stress and those elicited by voiceless consonants. In this connection, we note the presence of onset-sensitive stress systems in the world's languages. In a few such systems, a pattern is attested in which voiceless onsets attract stress, but voiced onsets do not. The opposite pattern is not found [27]. We suggest that in certain instances, voicing-induced differences in phonation of the type observed here may evolve into phonological prominence.

Returning to the phonological issue of laryngeal contrasts, we suggest that their phonetic complexity points to an approach to the phonetics-phonology relationship in which phonological specifications are abstract in nature and do not highlight any one phonetic property. Instead of a specification such as [-voice], which makes direct reference to the presence or absence of periodicity in the signal, we would opt in favour of an abstract feature such as [fortis], cued by a wide range of phonetic properties, but not tied to any one property in particular. Notably, this feature has not only consonant-internal acoustic correlates such as VOT and burst amplitude, but also encodes phonetic properties observed on neighbouring vowels, including raised f_0, clipped F1 transitions, and, as we have shown stiffer voice quality as reflected in higher amplitude harmonics in perceptually prominent areas of the acoustic spectrum.

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the Polish National Science Centre (Narodowe Centrum Nauki), project number 2021/41/B/HS2/00239.

7. REFERENCES

[1] Lisker, L., Abramson, A.S. 1964. A cross-language study of voicing in initial stops: acoustical measurements, *Word* 20, 384–422.

[2] Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22(3), 129-159.

[3] Esposito, A. 2002. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica* 59, 197−231.

[4] Ohde, R. N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing, *J. Acoust. Soc. Am.* 75, 224–230.

[5] Stevens, K., Klatt, D. 1974. Role of formant transitions in the voiced-voiceless distinction for stops, The *J. Acoust. Soc. Am.* 55, 653–659.

[6] Beckman, Jill, Petur Helgason, Bob McMurray & Catherine Ringen. 2011. Rate effects on Swedish VOT: Evidence for phonological over specification. *Journal of Linguistics* 39, 39-49.

[7] Honeybone, P. 2005. Diachronic evidence in segmental phonology: the case of obstruent laryngeal specifications. In: van Oostendorp, M., van der Weijer, J. (eds), *The internal organization of phonological systems.*Mouton de Gruyter, 319–354.

[8] Beckman, Jill, Michael Jessen & Catherine Ringen. 2013. Empirical evidence for laryngealfeatures: Aspirating vs. true voice languages. *Journal of Linguistics* 49, 259–284.

[9] Hanson, H. M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English, *J. Acoust. Soc. Am.* 125, 425-441.

[10] Kirby, J., Ladd D. R. 2016. Effects of obstruent voicing on vowel f0: evidence from 'true voicing' languages", *J. Acoust. Soc. Am.* 140, 2400–2411.

[11] Schwartz, G., Wojtkowiak, E., Brzoza, B. 2019. Beyond VOT in the Polish laryngeal contrast, *Proc. 19th ICPhS*, Melbourne, 2383-2387.

[12] Penney, J,, Cox, F., Miles, K., Palethorpe, S. 2018. Glottalisation as a cue to consonant voicing in Australian English. *J Phon* 66, 161-184.

[13] Sanker, C. 2019. Influence on coda stop features in perceived vowel duration. *J Phon* 75, 43-56.

[14] Koops, C., & Niedzielski, N. 2009. /t, d/-glottalization and vowel variation in Houston AAE. Paper presented at New Ways of Analyzing Variation (NWAV) 38, University of Ottawa, Ottawa, Canada.

[15] Garellek, M. 2015. Perception of glottalization and phrase-final creak. *J. Acoust. Soc. Am.*, 137, 822-831.

[16] Port, R. F., Dalby, J. 1982. Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics* 32, 141–152

[17] Schwartz, G., Arndt, D. 2018. Laryngeal realism vs. modulation theory – evidence from VOT discrimination in Polish, *Language Sciences* 69, 98-112.

[18] Boersma, P. 2001. Praat, a system for doing phonetics by computer, *Glot International* 5, 341-345.

[19] Shue, Y.-L., Keating, P., Vicenik, C. 2009. "VOICESAUCE: A program for voice analysis.," *J. Acoust. Soc. Am.*, 126, 2221- 2009.

[20] Garellek, M. 2019. The phonetics of voice. *Routledge Handbook of Phonetics*, Routledge, 75-106.

[21] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[22] JASP Team. 2023. JASP (Version 0.17.1) [computer software].

[23] Cho, T. and P. Ladefoged. "Variation and universals in VOT: evidence from 18 languages". *JPhon* 27, 207-229, 1999.

[24] Robinson, D. W., Dadson, R. S. 1957. Threshold of Hearing and Equal-Loudness Relations for Pure Tones, and the Loudness Function, . *J. Acoust. Soc. Am.* 29, 1284.

[25] Plag, G. K., Schramm, M. 2011. Acoustic correlates of primary and secondary stress in North American English. *J Phon* 39, 362-374.

[26] Sluijter, A., van Heuven, V.J. 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proc. ICSLP96*, 630-633.

[27] Gordon, M. 2005. A perceptually driven account of onset-sensitive stress, *Natural Language and Linguistic Theory* 23, 595-653.

PLANART: A MODULAR PLATFORM FOR COMPUTATIONAL MODELING OF ARTICULATORY PLANNING

Alice Turk[1], Benjamin Elie[1], Juraj Šimko[2]

[1]The University of Edinburgh, [2]University of Helsinki
a.turk@ed.ac.uk, benjamin.elie@ed.ac.uk, juraj.simko@helsinki.fi

ABSTRACT

We introduce the initial developments of PlanArt, a modular platform for testing different computational modeling approaches for speech articulation planning. The impetus for building this software comes from theoretical debates in the fields of Phonology and Phonetics which have led to vastly different modelling architectures, proposed mechanisms, as well fundamentally different dynamic models of speech articulatory trajectories. One way to address these debates is to test different models using a common platform for model testing in different contexts. In this paper, we present the platform architecture, designed to initially compare two Optimal Control Theory approaches: (1) Šimko & Cummins' Embodied Task Dynamics based on Articulatory Phonology, and (2) Turk & Shattuck-Hufnagel's XT/3C approach based on symbolic phonological planning. In addition, we discuss some of the challenges of this type of modelling task.

Keywords: speech articulation, articulatory modeling, speech production, planning, phonetics and phonology

1. INTRODUCTION

The creation of PlanArt is inspired by debates in the theoretical literature about the nature of representations and processes in planning speech articulations. Our hope is to contribute to this debate by providing a single modular platform for the comparison of models used for predicting articulatory patterns in different segmental and prosodic contexts. These comparisons will allow us to better understand different computational approaches to speech production planning as well as identify the interdependencies inherent within each approach. We believe that the comparative platform will ultimately lead to a fair evaluation of different theoretical approaches to speech motor control.

Here we first introduce the theoretical issues that motivate the development of PlanArt, followed by the goals of our system and the general architecture of the modular platform. We will illustrate the platform by the current partial implementation designed to compare the architectures and components required for two existing approaches to speech articulation modelling:

(1) **ETD**: A development of Articulatory Phonology called Embodied Task Dynamics [1, 2], and

(2) **XT/3C**: Turk & Shattuck-Hufnagel's Phonology-extrinsic Timing/3-Component approach [3, 4].

These two approaches both subscribe to the principles of Optimality Control Theory (OCT) which assume that speech articulation patterns arise as an optimal solution to both production and perceptual constraints (see Section 3).

Some of the differences between the two approaches that will be addressed by the PlanArt platform include (1) a symbolic planning component in XT/3C that is not used in ETD; (2) speech production goals defined in terms of acoustic cues in XT/3C *vs.* articulatory constriction goals in ETD, and (3) different dynamic models for trajectory planning: second order dynamics for ETD and Lee's General Tau Theory control [5] for XT/3C.

2. THEORETICAL BACKGROUND

This section provides a general discussion of different theoretical issues and their implications for computational models of speech articulation, to be addressed by a complete version of PlanArt.

2.1. Systematic phonetic variability

Words that speakers consider 'the same' (*phonologically equivalent*) are nevertheless pronounced systematically differently in different contexts (often referred to as *systematic phonetic variability*). Although there is strong agreement that all words that sound the same must share sub-lexical phonological representations, there is lack of agreement about the nature of these representations (spatiotemporal vs. symbolic). These different types of proposed representations lead to different system architectures that relate these representations to produced speech that

includes appropriate variability in surface articulatory patterns. Models assuming symbolic phonological representations traditionally separate phonological and phonetic planning in the speech production process and require mapping the Phonological (symbolic) representations to the Phonetic (articulatory) space [3].

An alternative approach assuming a spatiotemporal (gestural) nature of phonological units [6, 1] provides a way of unifying the planning process into a single Phonological planning component.

While there is growing appreciation for the influence of multiple phonological and stylistic factors on the phonetic form of particular utterances, there are different ways of modelling this, e.g. Optimal Control Theory (e.g. [7] and Dynamic Field theory [8]).

2.2. The underlying dynamics of speech movements and the nature of speech goals

Most speech movements exhibit the hallmarks of practiced, purposeful movement, that is, single-peaked, and (often but not always) symmetrical velocity profiles. Modifications of damped oscillator-based systems have been previously used to generate speech movement trajectories in several articulatory models [9, 10, 11]. Lee's General Tau theory [5] (see also below) addresses the issue of targeted motor action more generally. It controls the time-course of movement (movement acceleration and deceleration over time) via a single parameter in a relatively simple equation. Most importantly, this equation presents a better fit to articulatory measurements than oscillator-based systems [12].

A closely related issue is the nature of the goals of speech movements. Models with context-independent articulatory goals (such as most damped oscillator-based ones) require accounts of contextual differences in spatial positions at movement endpoints, e.g. via undershoot from shorter activation intervals, gestural blending, or sensitivity to feedback about target achievement. Models with context-dependent goals–such as those using General Tau theory that require full spatial and temporal specification of the realized movement targets–require ways of determining these goals.

While models with articulatory goals (where sound is the result of motor action and does not need to be explicitly specified) are arguably simpler, acoustic goals may be required to account for the equivalence of different motor actions that result in similar sounds [13]. Approaches relying on acoustic speech targets need a mechanism for solving the acoustic-to-articulation inversion problem [14].

2.3. The nature of coordination and speech timing patterns

Speech involves the skilled, efficient coordination of sets of articulators. However, how coordination works is still unclear. Possibilities include coordination based on the time of movement onset, the time of target achievement, and based on feedback about particular spatial or spectral states (e.g. associated with target achievement). Different dynamic models of speech movement trajectories have implications for modelling coordination. Because targets are never reached in damped oscillator-based systems, these appear to require either coordination based on movement onsets [15], or coordination based on feedback about target achievement in articulatory or acoustic space [14, 16, 17]. In contrast, Lee's Tau Theory equation provides a mechanism for coordination based on the *time* of target achievement.

Debates relate to whether speech timing patterns emerge from adjustments of default activation intervals of spatiotemporal phonological representations [6], from sensitivity to feedback about when speech targets are achieved [16, 17], or if they are explicitly specified, as part of a Phonetic Planning process [3], via optimization procedures [2], [1].

3. MODELING OBJECTIVES

Our overall goal is to develop a flexible and modular modelling platform allowing for parallel implementation, testing and comparison of sets of competing theoretical comparisons, such as those introduced above.

In the initial stages of development, our focus is on creating a platform architecture that allows a comparison of two theories–ETD and XT/3C–that share a common Optimal Control Theory approach in which trade-offs between competing requirements of production efficiency and perceptual efficacy are satisfied [18, 3].

The modularity of the platform assumes that the transformation of the input to the model–expressed in a form of a gestural score (for ETD) or a sequence of symbolic phonemic units with associated acoustic cues, for XT/3C–to the output (articulatory trajectories) proceeds in identifiable discrete steps, and that each step can be computationally implemented as an independent module with its own inputs and outputs. The platform will be flexible in the sense that individual modules can be replaced by other modules with essentially the same functionality albeit using different theoretical underpinnings.

Briefly, the competing requirements are concep-

tualized in the form of a multi-objective cost function, combining an articulatory *effort* E, linked to the forces applied to embodied speech articulators during speech, requirements of perceptual distinctiveness expressed as a *parsing* cost, denoted P, and the cost of *time* it takes to produced the entire utterance, denoted D. The composite cost function can then be expressed as a weighted average of these partial cost components

$$C = \alpha_E E + \alpha_P P + \alpha_D D, \quad (1)$$

where α_E, α_P, α_D are weights assigned to individual components within the trade-off.

The weights of these components and, potentially, other parameters of the cost function (*optimization parameters*) quantify the relative priorities of task requirements, including context-dependent constraints, as well as costs of movement (e.g. effort). The task requirements are assumed to be provided to the optimization procedure (conceptualized as part of *dynamic planning*) through prior phonological processing. For a given intended utterance, the optimal timing and kinematic characteristics of speech movements that minimize the value of the combined cost function reflect the requirements encoded by the optimization parameters. Increasing the weight α_D, for example, imposes a greater premium (cost) on time, and the optimal trajectories can be expected to correspond to articulation at a faster speaking rate; this type of adjustment can be imposed globally or locally, eliciting global or local articulation rate changes.

4. PLATFORM ARCHITECTURE

The modular PlanArt architecture is illustrated in Fig. 1. The Symbolic Planning Component, required only for XT/3C (corresponding to its Phonological Planning Component) provides the parameters for the task requirements (context-dependent constraints) to be used in the objective function.

The Dynamic Planning module is the component implementing the optimization procedure discussed above. This processing step leads to generation of a detailed articulatory plan in the form of a vector of static and dynamic articulatory parameters that can be executed during a speech production event by the Motor-Sensory Implementation module. During the optimization process itself, the Articulatory trajectory module (potentially a copy of the Motor-Sensory Implementation one) computes the articulatory trajectories (given a static and a dynamic model) and returns a cost C to the dynamic plan following Eq. (1). The optimization procedure consists of finding the articulatory plan for which the

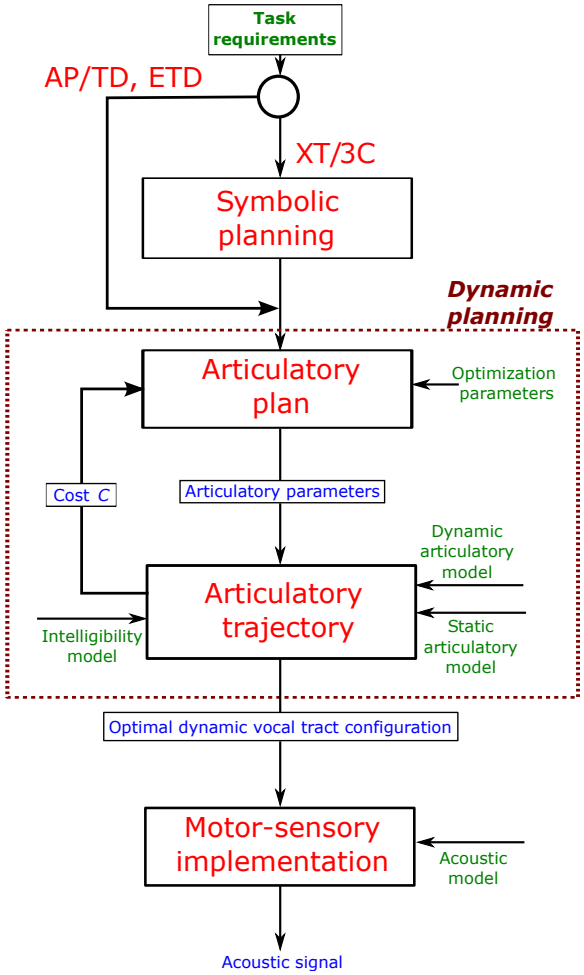


Figure 1: Architecture of the PlanArt software for comparing ETD and XT/3C. Elements in green and elements in blue are module inputs and outputs, respectively. Modules names are in red fonts.

articulatory trajectory component returns a minimal cost. Within ETD this thus corresponds to a gestural score formed during phonological planning (cf. Articulatory Phonology), where the gestural score is a phonological representation of an utterance. The dynamical planning stage corresponds to the Phonetic Planning Component within XT/3C theory. The optimal articulatory trajectories are then fed to the motor-sensory implementation which computes the acoustic signal given a user-defined acoustic model.

The modular architecture can also capture other key differences between the two approaches, such as different dynamic models for trajectory planning.

As a development of the Task Dynamics implementation of AP [19, 20], ETD uses a (mutually coupled system of) second order critically damped dynamics to model the time-course of articulatory movement and task realization. While this choice does not yield the best fitting articulatory movements, it reflects the theoretical commitment of the

approach, namely seeing the task oriented articulatory action (a gesture) as a phonological primitive. As in Articulatory Phonology, systematic phonetic variability results from the properties of the dynamics. For example, the *phonological* target conceptualized as a parameter of dynamical description is never reached, and variability in ETD arises partly from a degree of undershooting the target. The degree of undershoot in turn depends on other dynamical parameters, such as stiffness and activation interval, that are within ETD computed as a part of the phonological planning step as optimal solutions of the objective function.

On the other hand, XT/3C proposes the time between acoustic landmarks, as well as General Tau Theory and its equations of movement as underlying sources of timing and articulatory dynamics which are used in its Phonetic Planning Component. Unlike damped mass-spring dynamics, the Tau Guide equation assumes a pre-specified movement duration and a *reached* target of the movement.

Another difference between the XT/3C and ETD approaches is related to speech production goals defined in terms of acoustic cues for the former *vs.* articulatory constriction goals for the latter. Within the Optimal Control driven PlanArt architecture, the appropriate degree of proximity to the target is evaluated within the parsing cost part of the objective function. Consequently, this modeling decision (acoustic or articulatory targets) can be to a large degree implemented as part of the mathematical definition of the cost component and does not necessarily influence the remaining parts of the implementation.

In this way (and leaving aside some theoretical assumptions), the modular architecture of PlanArt platform will thus allow testing various setups by, for example, swapping the dynamic models used for trajectory planning or the nature of targets between two approaches. This possibility will allow us to test the practical relevance and importance of various modelling commitments (although see some potential challenges to the modular approach below).

5. MODELING CHALLENGES

The optimization approach facilitates the modular architecture of the PlanArt system by allowing inclusion of alternatives in terms of dynamic models of articulation (tau-guided *vs.* damped attractor dynamics), or in terms of the nature of targets. The details of the implementation might, however, differ quite dramatically depending on the choice, in particular in terms of the overall number of optimized parameters.

The AP approach (including ETD) assumes context-independent phonological targets (in the task space) implemented as target parameters of the dynamical description; each possible gesture within the given language repertoire has one such target. The contextual variation arises through the context-dependent nature of realization of the movement towards this target, determined by other dynamical parameters (such as stiffness) and by the activation patterns. For a given utterance, the parameters to be optimized are thus limited to this relatively small "context dependent" subset.

On the other hand, an assumption of context dependent targets (articulatory or acoustic) implies a need to optimize much higher number of parameters. For a Tau-guided movement, for example, every articulatory movement has to be fully specified in terms of its (achieved) end-point and duration of the movement. That means that in addition to the dynamical parameters of the context-dependent movement dynamics (a Tau-coupling constant in case of the Tau-guided movement), the optimization process needs to find appropriate context-dependent targets for each individual articulator for each allophonic token within the utterance.

As argued above, thanks to the optimization approach this is not a problem in principle, but may have practical (and, possibly, theoretical) implications regarding the potential efficacy of the system.

Another modelling challenge will be the appropriate implementation of acoustic (rather that articulatory) targets of speech production that will essentially require an explicit inverse mapping from acoustic to articulatory space. The team developing the PlanArt software are currently exploring probabilistic models specially designed for this task.

6. DISCUSSION

Model comparison via a common platform is a promising way of assessing the benefits and computational challenges of theoretical alternatives. Our hope is that PlanArt can be used by the research community to compare different modeling approaches reflecting different theoretical assumptions (beyond the presently discussed ETD and XT/3C). The software is intended to be an open-source software. For the sake of anonymity, the link to a public repository, where the source codes of the software are freely available, will be provided only in the final version of the paper. PlanArt is in its beginning stages; we anticipate many insights will come of this project.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the European Research Council (ERC) under the European Unionâs Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A. Turk; Grant agreement No. 101019847).

8. REFERENCES

[1] J. Simko and F. Cummins, “Embodied task dynamics,” *Psychological review*, vol. 117, no. 4, p. 1229, 2010.

[2] ——, “Sequencing and optimization within an embodied task dynamic model,” *Cognitive Science*, vol. 35, no. 3, pp. 527–562, 2011.

[3] A. Turk and S. Shattuck-Hufnagel, *Speech timing: Implications for theories of phonology, speech production, and speech motor control*. Oxford University Press, USA, 2020.

[4] ——, “Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production,” *Frontiers in Psychology*, p. 2952, 2020.

[5] D. N. Lee, “Guiding movement by coupling taus,” *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[6] C. P. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[7] E. Todorov and M. I. Jordan, “Optimal feedback control as a theory of motor coordination,” *Nature neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[8] K. D. Roon and A. I. Gafos, “A dynamical model of the speech perception-production link,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, no. 35, 2013.

[9] D. Byrd and E. Saltzman, “The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening,” *Journal of Phonetics*, vol. 31, no. 2, pp. 149–180, 2003.

[10] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, “Model-based reproduction of articulatory trajectories for consonant–vowel sequences,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[11] T. Sorensen and A. Gafos, “The gesture as an autonomous nonlinear dynamical system,” *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[12] B. Elie, D. Lee, and A. Turk, “Modeling trajectories of human speech articulators using general tau theory,” *Speech Communication*, to appear.

[13] J. S. Perkell, “Movement goals and feedback and feedforward control mechanisms in speech production,” *Journal of neurolinguistics*, vol. 25, no. 5, pp. 382–407, 2012.

[14] F. H. Guenther, “A neural network model of speech acquisition and motor equivalent speech production,” *Biological cybernetics*, vol. 72, no. 1, pp. 43–53, 1994.

[15] H. Nam, L. Goldstein, and E. Saltzman, “Self-organization of syllable structure: A coupled oscillator model,” *Approaches to phonological complexity*, vol. 16, pp. 299–328, 2009.

[16] S. Tilsen, “A dynamical model of hierarchical selection and coordination in speech planning,” *PloS one*, vol. 8, no. 4, p. e62800, 2013.

[17] ——, “Three mechanisms for modeling articulation: selection, coordination, and intention,” *Working Papers of the Cornell Phonetics Laboratory*, 2018.

[18] J. Simko, “The embodied modelling of gestural sequencing in speech,” University College Dublin. School of Computer Science and Informatics, Tech. Rep., 2009.

[19] E. Saltzman, “Task dynamic coordination of the speech articulators: A preliminary model,” *Experimental brain research series*, vol. 15, pp. 129–144, 1986.

[20] E. L. Saltzman and K. G. Munhall, “A dynamical approach to gestural patterning in speech production,” *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989.

ACOUSTIC PROPERTIES OF PALATALIZED CONSONANTS IN JAPANESE

Yu Tanaka

Doshisha University
yutanak@mail.doshisha.ac.jp

ABSTRACT

Palatalization affects the acoustics of segments in various ways. The present study investigates the spectral and durational properties of Japanese palatalized stops and fricatives in the data of 90 female and 111 male speakers taken from a large speech corpus. The results show, as expected, that palatalized consonants tend to have higher F2 and higher CoG (except /sʲ/) of their release or frication noise and mostly (all but labial /pʲ/) have longer VOT or total duration than their plain counterparts. The findings suggest that phonemically palatalized consonants in Japanese, though variously realized in phonetic terms, exhibit cross-linguistically common acoustic characteristics of palatalization.

Keywords: Palatalization, F2, CoG, VOT, Japanese

1. INTRODUCTION

The phonology of Japanese systematically contrasts plain and palatalized consonants [1, 2]; many consonants have a plain form /C/ and a palatalized form /Cʲ/. As is common cross-linguistically [3, 4, 5], the actual phonetic realizations of the phonemically palatalized consonants vary. Some are produced with secondary palatalization, as in /pʲ/ [pʲ] and /kʲ/ [kʲ], while others are true palatals or alveolo-palatals, as in /hʲ/ [ç] and /sʲ/ [ɕ].

There is a large body of literature describing the phonetic correlates of palatalization in other languages (see e.g. [6, 7, 8, 9, 10] and references therein). Common acoustic effects of palatalization on stops and fricatives are as follows. When palatalized, they tend to have release or frication noise that is louder, longer, and spectrally distinct, with higher second formant frequency (F2) and higher center of gravity (CoG) values (except for sibilants) than their plain counterparts.

Relatively little research has examined the acoustic properties of Japanese palatalized stops and fricatives. One experimental study [11] analyzed /bʲa, gʲa, mʲa, nʲa, ɾʲa/ and /ba, ga, ma, na, ɾa/, showing that the syllable duration of the former group is longer; however, it did not examine the duration of voiceless stops or fricatives. Other studies [12, 13] conducted acoustic analyses of word-medial /t, k, tʲ, kʲ/ in recordings taken from a speech corpus, reporting that /kʲ/ tends to have shorter closure and longer release noise than /k/, but no such compensatory effect is found between /t/ and /tʲ/. Note, however, that /tʲ/ [tɕ] involves affrication, and direct comparison with plain /t/ [t] or with palatalized dorsal /kʲ/ [kʲ] is not easy. The study also excluded the labial stop /pʲ/ and fricatives such as /sʲ/ and /hʲ/ from consideration.

The current study provides further analyses of palatalized consonants in Japanese. Specifically, I measure the spectral and durational features of palatalized stops and fricatives /pʲ, kʲ, sʲ, hʲ/, and compare them to those of plain /p, k, s, h/.

2. METHOD

Following [12, 13], I used the *Corpus of Spontaneous Japanese* (CSJ) [14], which contains speech data of 90 female and 111 male speakers from various age groups with detailed annotations. Fig. 1 is an image of annotated speech containing *ky* /kʲ/ in the word *keNkyuH* /keɴkʲuː/ 'research'. The palatalized stop is segmented into closure (<cl>) and release (labeled *ky*) in the segment ("seg") tier.

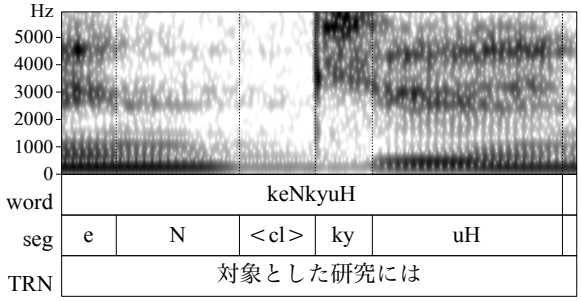


Figure 1: Annotated speech in CSJ

To examine the acoustic properties of palatalized stops, I measured the duration and midrange F2 and CoG of the release of /pʲ/ and /kʲ/, and those of /p/ and /k/ for comparison. (The midrange was defined

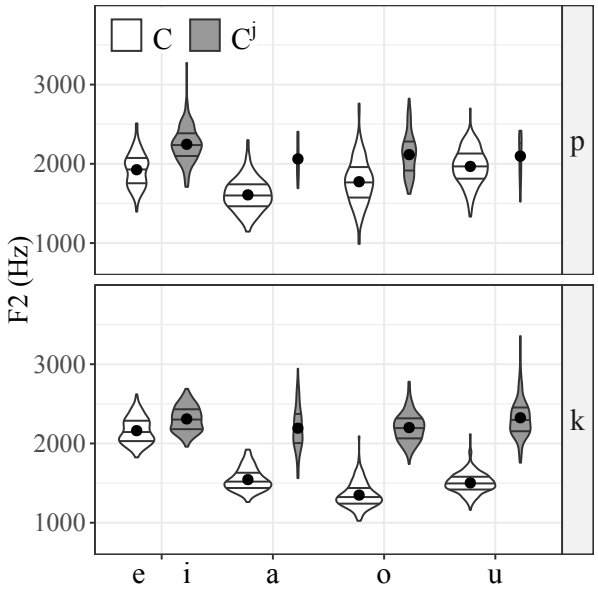


Figure 2: F2 of stop release noise

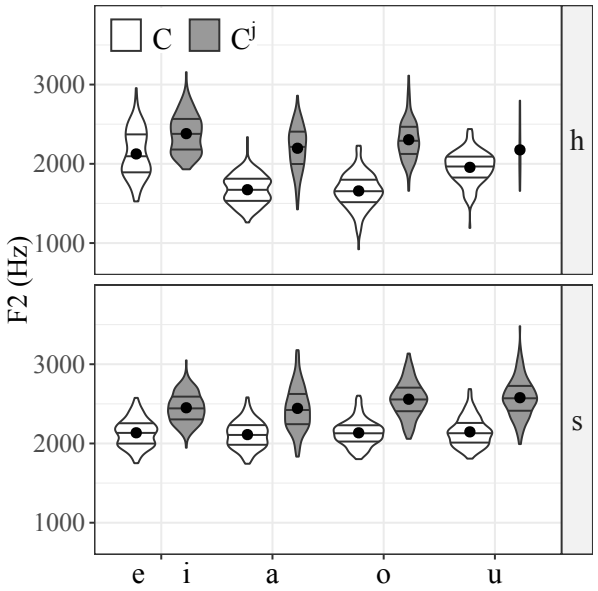


Figure 3: F2 of frication noise

as a 20ms window, or a 5ms window if the release was shorter than 20ms, centered at the midpoint.) For fricatives, I measured the duration and midrange F2 and CoG of /sʲ/ and /hʲ/, and those of /s/ and /h/. (I excluded cases where the following vowel was devoiced.) The measurement procedure was automated using a modified version of the *formant-logger* script [15] on *Praat* [16]. In total, 157,823 tokens of the target segments were analyzed.

3. RESULTS

3.1. F2 and CoG of release and frication noise

I first present the results of F2 of stop release noise. Fig. 2 shows a violin plot of the speakers' mean F2 of plain /C/ and palatalized /Cʲ/ by consonant type and vowel context. Lines (—) and dots (•) represent quartiles and grand means, respectively. There are differences in sample size, as shown by width, since some speakers' data do not contain any instances of the relevant sequences. The /e/ and /i/ contexts are given for reference; in principle, /e/ only occurs with plain /C/ (i.e. [pe], [ke]), and /i/ only with palatalized /Cʲ/ (i.e. [pʲi], [kʲi]) on the surface level.

It can be seen that /Cʲ/ has relatively high F2 in general. A mixed-effects model was constructed on *R* [17, 18, 19] to predict F2 values with the presence of palatalization (/C/, /Cʲ/), the consonant type (/p/, /k/), and their interaction as its predictors. The random structure included by-speaker random intercepts and slopes, as well as random intercepts for the vowel context (/a/, /o/, /u/). The data in the /e/ and /i/ contexts with no real contrast in palatalization were excluded, and the remaining 117,120 tokens were used as fitting data.

The model shows that /Cʲ/, or palatalization, significantly raises F2 in /pʲ/ ($\beta = 525.77$, $t = 20.47$, $p < 0.001$) with the baseline intercept being /p/; the same is true for /kʲ/ ($\beta = 830.26$, $t = 77.12$, $p < 0.001$) if the baseline is set to /k/. The main effect of the consonant type is also significant; /p/ has higher F2 than /k/ ($\beta = 219.44$, $t = 17.93$, $p < 0.001$). (However, see §4 for discussion of possible measurement errors.) In addition, the interaction term of the two predictors is significant, indicating that the F2-raising effect of /Cʲ/ is greater on /kʲ/ than /pʲ/ ($\beta = 304.49$, $t = 11.43$, $p < 0.001$).

Fig. 3 plots the F2 values of fricatives, contrasting /C/ and /Cʲ/. Here too, /Cʲ/ shows consistently higher F2 in every vowel context. Another mixed-effects model was applied in the same manner as above, except that the consonant type was fricatives (/h/, /s/). It predicts that palatalization raises F2 in both /sʲ/ ($\beta = 403.73$, $t = 28.19$, $p < 0.001$) and /hʲ/ ($\beta = 587.48$, $t = 38.70$, $p < 0.001$). /h/ generally has lower F2 than /s/ ($\beta = -402.33$, $t = -41.05$, $p < 0.001$), but palatalization has a greater effect on /hʲ/ than /sʲ/, as is indicated by a significant interaction ($\beta = 183.75$, $t = 8.55$, $p < 0.001$).

Fig. 4 gives the results for the CoG of stop release noise. Generally speaking, /Cʲ/ shows higher CoG than /C/. A mixed-effects model indicates that the effect of /Cʲ/ is significant in both /kʲ/ ($\beta = 1819.92$, $t = 43.96$, $p < 0.001$) and /pʲ/ ($\beta = 625.76$, $t = 9.69$, $p < 0.001$). Also, CoG is higher in dorsal /k/ than

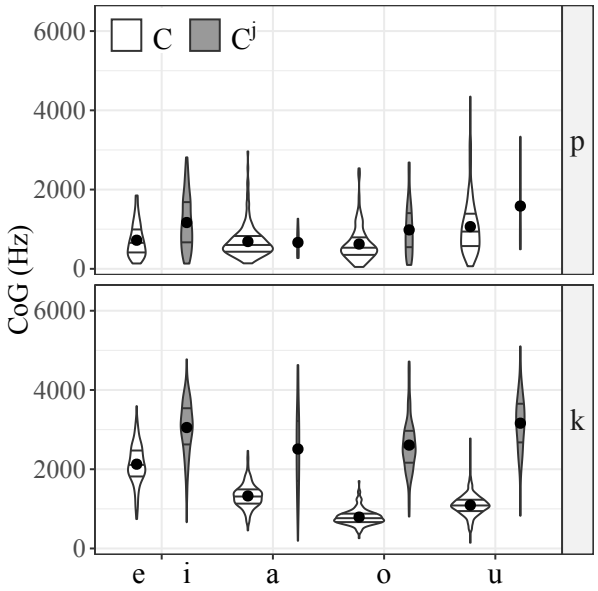


Figure 4: CoG of stop release noise

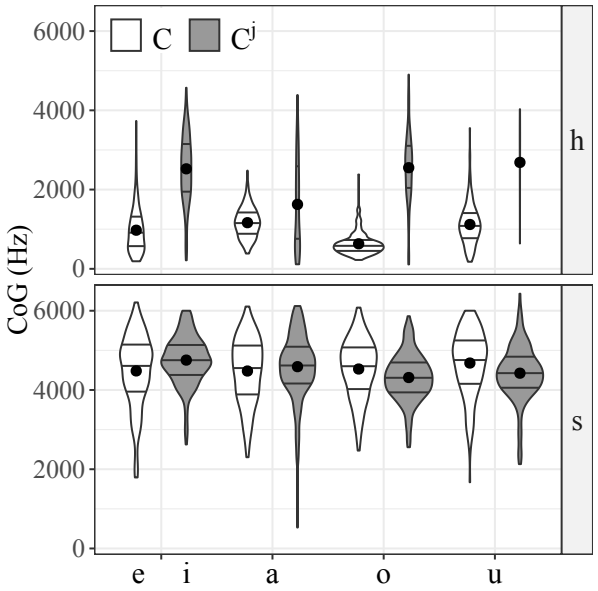


Figure 5: CoG of frication noise

labial /p/ ($\beta = 446.19$, $t = 16.78$, $p < 0.001$), and a significant interaction further shows that the effect of palatalization is greater on the dorsal ($\beta = 1194.16$, $t = 16.08$, $p < 0.001$).

Fig. 5 shows the results for fricatives. As can be seen, /s/ has higher CoG than /h/ ($\beta = 3602.75$, $t = 69.18$, $p < 0.001$). The effect of palatalization is significant, with CoG higher in /h^j/ than in /h/ ($\beta = 1221.57$, $t = 17.63$, $p < 0.001$). However, a significant interaction between palatalization and the consonant type suggests that no such relationship holds for /s^j/ and /s/ ($\beta = -1355.83$, $t = -15.87$, $p < 0.001$); in fact, when directly compared, the CoG of /s^j/ is lower than that of /s/ ($\beta = -134.265$, $t = -2.90$, $p = 0.004$). That is, palatalization rather lowers CoG in the case of sibilants.

In sum, palatalization is associated with high F2 and CoG of stop release and frication noise, except that sibilant /s^j/ has relatively low CoG.

3.2. VOT and total duration

Turning to the durational properties of palatalized consonants, I present in Fig. 6 the duration of voiceless stop release noise, namely voice onset time (VOT). Notice that in the /k/ conditions, /C^j/ tends to have longer VOT than /C/. As for the /p/ conditions, in contrast, the differences between /C/ and /C^j/ seem at best marginal.

A mixed-effects analysis reveals that the effect of /C^j/ is significant in /k^j/ ($\beta = 9.95$, $t = 16.84$, $p < 0.001$), but its interaction with the consonant type of /p/ is negatively significant ($\beta = -8.96$, $t = -7.92$, $p < 0.001$); /C^j/ actually does not lengthen VOT in the case of /p^j/ ($\beta = 0.99$, $t = 1.03$, $p = 0.305$). Additionally, the main effect of the consonant type is significant; /k/ has longer VOT than /p/ ($\beta = 9.11$, $t = 22.29$, $p < 0.001$).

Fig. 7 shows the duration of fricatives. For the most part, /C^j/ is longer than /C/. Statistically, palatalization lengthens both /s^j/ ($\beta = 6.07$, $t = 8.13$, $p < 0.001$) and /h^j/ ($\beta = 9.49$, $t = 8.24$, $p < 0.001$), although its significant interaction with the consonant type indicates that the effect is stronger in the latter ($\beta = 3.42$, $t = 2.53$, $p = 0.013$). Also, /s/ is longer than /h/ ($\beta = 23.88$, $t = 37.39$, $p < 0.001$).

The results as a whole thus show that palatalization lengthens the VOT of stops and duration of fricatives, except in the case of labial /p^j/ where no such lengthening effect is found.

4. DISCUSSION

Acoustic analyses of plain and palatalized stops and fricatives in Japanese have shown that palatalization lengthens their release and frication noise and raises their F2 and CoG, although these effects are not observed with certain segment types. These results, including the null effects, in fact align well with cross-linguistically common patterns, and also hold some theoretical implications.

As stated above, palatalized consonants tend to be longer. [9] further reports that in Connemara Irish, palatalization lengthens stop release duration, but the effect is weaker in labials than in coronals. This is parallel to the null result regarding the VOT of

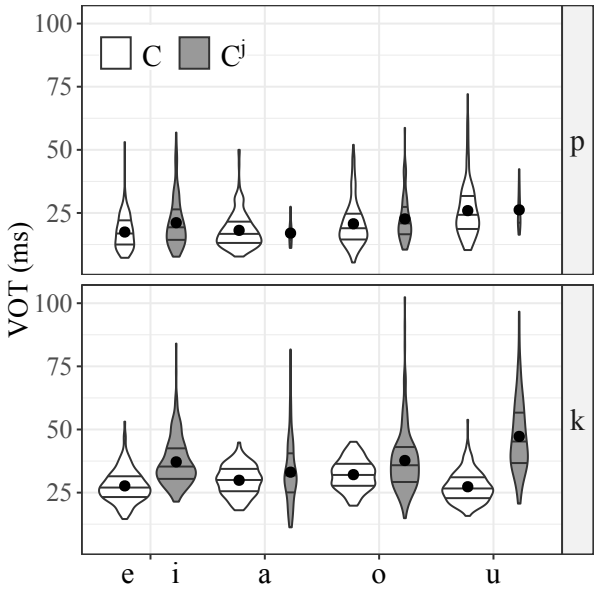


Figure 6: VOT of stops

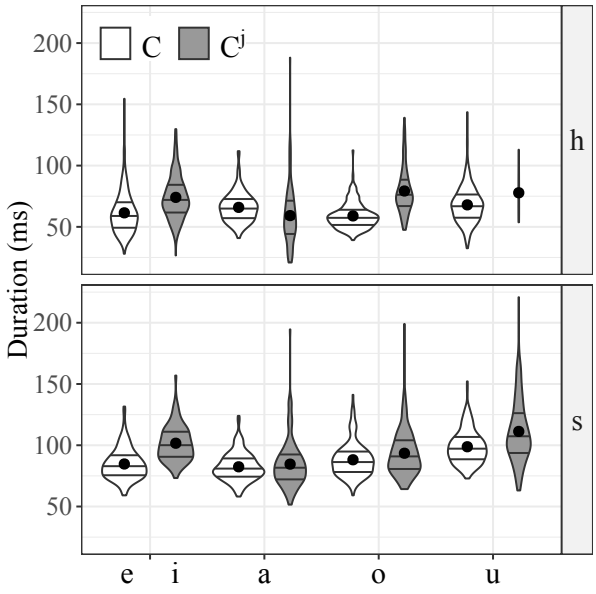


Figure 7: Duration of fricatives

/pʲ/ in Japanese (Fig. 6). Labial stops are known for already having relatively short VOT [20, 21]. The facts above suggest that they even resist the lengthening effect of palatalization.

The results of CoG are also compatible with cross-linguistic data. Secondary palatalization itself tends to raise CoG, but alveolo-palatal and post-alveolar sibilants such as [ɕ] and [ʃ] inherently show lower CoG than alveolar [s] (see [8, 9, 22, 23]). As we have seen, Japanese palatalized consonants also have higher CoG than their plain counterparts, except that /sʲ/, phonetically realized as alveolo-palatal [ɕ], has lower CoG than /s/ [s] (Fig. 5).

The results for fricative duration (Fig. 7) can be relevant to debates on representations in traditional Japanese phonology. For alveolo-palatal sibilants such as [ɕ] and [ʑ], some scholars posit underlying /Cʲ/, as in /sʲ/ and /zʲ/, based on the fact that they pattern with other phonemically palatalized consonants [1, 2]. Others propose representations that are closer to the surface forms, as in /ɕ/ and /ʑ/, partly because they are phonetically single sounds [11, 24]. Although arguments about phonological representations should be based on phonological evidence, the results here suggest that [ɕ] (as well as [ç]), which is transcribed as a single sound, shows common characteristics of "palatalized" consonants; it is longer, as if involving two articulatory gestures, as implied by the notation /Cʲ/.

The results for the F2 of stop release noise (Fig. 2.) also require some attention. Recall that plain /p/ has higher F2 than /k/. This is somewhat unexpected given that labials generally lower F2. It is possible that the automated measurement had errors when F2 was too low, especially in the cases of /po/ and /pu/. That said, the low F2 values of /k/ in back vowel contexts seem genuine. Impressionistically, some speakers' production of /k/ before /a, o, u/ is quite back, sometimes sounding like uvular [q]. This may be interpreted as a kind of enhancement effect, as seen in Russian [25]; not only are palatalized consonants fronted, but non-palatalized ones are backer.

Another caveat is that the phonological contexts and the frequencies of the target segments are not uniform. In principle, /p(ʲ)/ occurs word-medially as a geminate or after a moraic nasal [26] (e.g. /happa/ 'leaf', /kaɴpai/ 'toast'). The occurrence of /h(ʲ)/ is also limited to word-initial position in the native vocabulary. Even though the analyzed corpus data include recent loans where these restrictions do not hold (e.g. /paɴ/ 'bread'), the results above may still have been affected by the distributional and frequency differences. Conducting a production experiment using nonce words with more control on these factors would solve the issue.

There are other possible acoustic features of palatalization which this study has not explored. For example, previous studies show that the intensity of stop release and F2 of following vowels could also differentiate palatalized and plain consonants. Intensity was especially difficult to address using speech data of different speakers recorded in various environments. Future experimental work should examine these other acoustic properties with respect to palatalized consonants in Japanese.

5. ACKNOWLEDGMENTS

I thank two anonymous reviewers for their comments. This study was partially supported by JSPS Kakenhi Grants 20K13019 and 22K13106.

6. REFERENCES

[1] J. Ito and A. Mester, *Japanese Morphophonemics*. Cambridge, MA: MIT Press, 2003.

[2] L. Labrune, *The Phonology of Japanese*, ser. The Phonology of the World's Languages. Oxford: Oxford University Press, 2012.

[3] D. N. S. Bhat, "A general study of palatalization," in *Universals of Human Language*, J. Greenberg, Ed. Stanford: Stanford University Press, 1978, vol. 2: Phonology, pp. 47–92.

[4] N. Bateman, "On the typology of palatalization," *Language and Linguistics Compass*, vol. 5, no. 8, pp. 588–602, 2011.

[5] A. Kochetov, "Palatalization," in *The Blackwell Companion to Phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds. Malden: Wiley-Blackwell, 2011, vol. 3: Phonological Processes, pp. 1666–1690.

[6] ——, *Production, Perception, and Emergent Phonotactic Patterns: A Case of Contrastive Palatalization*. New York: Routledge, 2002.

[7] ——, "Testing licensing by cue: A case of Russian palatalized coronals," *Phonetica*, vol. 63, no. 2-3, pp. 113–148, 2006.

[8] ——, "Acoustics of Russian voiceless sibilant fricatives," *Journal of the International Phonetic Association*, vol. 47, no. 3, pp. 321–348, 2017.

[9] M. Ní Chiosáin and J. Padgett, "An acoustic and perceptual study of Connemara Irish palatalization," *Journal of the International Phonetic Association*, vol. 42, no. 2, pp. 171–191, 2012.

[10] L. Spinu, I. Vogel, and H. T. Bunnel, "Palatalization in Romanian – Acoustic properties and perception," *Journal of Phonetics*, vol. 40, no. 1, pp. 54–66, 2012.

[11] A. Nogita, "Arguments that Japanese [Cj]s are complex onsets: Durations of Japanese [Cj]s and Russian [C^{j}]s and blocking of Japanese vowel devoicing," *Working Papers of the Linguistics Circle of the University of Victoria*, vol. 26, no. 1, pp. 73–99, 2016.

[12] M. F. Matsui, "A case study of the temporal features of Japanese palatalized consonants [Nihongo yōon no jikan tokusei ni kansuru jirei kenkyū]," *Theoretical and Applied Linguistics at Kobe Shoin*, vol. 24, pp. 434–443, 2021.

[13] ——, "The temporal features of Japanese palatalized consonants [Onsei jizoku jikan kara mita Nihongo yōon no seishitsu]," *Proceedings of the 38th Meeting of the Japanese Cognitive Science Society*, vol. 38, pp. 434–443, 2021.

[14] National Institute for Japanese Language and Linguistics, "The Corpus of Spontaneous Japanese," 2012, https://ccd.ninjal.ac.jp/csj/index.html.

[15] K. Crosswhite, "formant-logger," 2009, Praat script.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2022, Computer program. Version 6.3.03. http://www.praat.org/.

[17] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using 'Eigen' and S4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48., 2015.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26., 2017.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2022.

[20] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[21] T. Cho and P. Ladefoged, "Variations and universals in VOT: Evidence from 18 languages," *Journal of Phonetics*, vol. 27, pp. 207–229, 1999.

[22] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *The Journal of the Acoustical Society of America*, vol. 108, pp. 1252–1263, 2000.

[23] M. Gordon, P. Barthmaier, and K. Sands, "A cross-linguistic acoustic study of voiceless fricatives," *Journal of the International Phonetic Association*, vol. 32, no. 2, pp. 141–174, 2002.

[24] T. J. Vance, *The Sounds of Japanese*. Cambridge: Cambridge University Press, 2008.

[25] J. Padgett, "Contrast dispersion and Russian palatalization," in *The Role of Speech Perception in Phonology*, E. Hume and K. Johnson, Eds. New York: Academic Press, 2001, pp. 187–218.

[26] J. Ito and A. Mester, "Japanese phonology," in *The Handbook of Phonological Theory*, J. Goldsmith, Ed. Cambridge, MA: Blackwell Publishers, 1995, pp. 817–838.

Aspiration length and intensity as potential acoustic cues to [voice] in final position: evidence from Shughni

Yury Makarov

HSE University / Institute of Linguistics of the Russian Academy of Sciences
yurmak@iling-ran.ru

ABSTRACT

Final Shughni plosives tend to be released and aspirated, at least in citation forms. Data from 14 Shughni speakers showed that underlyingly voiced plosives have a shorter period of aspiration, and its intensity is weaker compared to that of voiceless ones. Although sometimes neutralized with regard to closure voicing, the contrast may be maintained by means of final aspiration length and intensity. Nevertheless, typological accounts do not usually treat these two characteristics as potential acoustic cues to underlying voicing in final position, and this study gives evidence that final aspiration length and intensity should be added to the list (cf. such cues as the duration of closure and preceding vowel, closure voicing, and F_1 frequency). Moreover, possible implications for languages with final aspiration (like Kashmiri) are discussed.

Keywords: neutralization, final devoicing, aspiration, acoustic cues, Iranian languages.

1. INTRODUCTION

1.1. Shughni language and speakers

Shughni is an Iranian language spoken by ca. 100,000 in the Pamir Mountains in Tajikistan and Afghanistan. It has several varieties, constituting the so-called Shughni-Rushani group. From now on, the variety spoken in Khorugh, the capital of the Gorno-Badakhshan Autonomous Region in Tajikistan, is described. Shughni is used primarily as a spoken language, though there are a number of poetry books and even novels. It is usually the case that Shughni speakers are bi- or even multilingual, having some knowledge of Tajik (official language) and Russian (lingua franca) [1].

1.2. Final laryngeal neutralization

A typical example of final laryngeal neutralization is final devoicing. It occurs when consonants of the voiced and voiceless series are realized without vocal fold vibration in final position, e.g., Russian /kod/ ‘code’ and /kot/ ‘cat’ pronounced as [kot]. Although neutralization may be complete in respect of closure voicing, acoustic cues to underlying voicing are likely to be present (thus making the neutralization process incomplete). For example, utterance-final postvocalic stops in English may have closure voicing, but there are other means of signalling the [voice] feature [2]. In fact, for American English, vowel duration is claimed to be a perceptually sufficient cue to underlying voicing in word-final stops, fricatives and clusters [3], [4]. On the other hand, in German, it was shown to be significant only in production [5].

Well-established acoustic cues to voicing in English are summarized in Table 1.

Cue	Voiced	Voiceless
preceding vowel	longer	shorter
closure	voiced shorter	voiceless longer
F_1 at the edges of adjacent vowels	lower	higher

Table 1: Acoustic cues to underlying voicing in utterance-final postvocalic stops in English (after [2]).

To be considered an acoustic cue, an acoustic characteristic must aid in the recognition of some feature of the segment [6], [7]. Until this perceptual significance is shown, this acoustic characteristic may be regarded as a potential cue, provided that it correlates with the feature in question. As for [voice], potential cues are the presence/absence of release (voiced stops are less often released) [8], [9], the intensity and spectral properties of release [10], and burst (expected differences are not specified) and aspiration (longer in underlyingly voiceless) duration [5], [7], [11] There is also a claim that ‘any noise attendant upon consonant production will have greater intensity in the voiceless sound’ [12], though few works are focused on considering the linguistic implications of this acoustic fact, regarding final stops in particular.

Final devoicing is not the only possible scenario of neutralization [13]. For instance, in Kashmiri (Indo-Aryan) the unaspirated and aspirated series are neutralized in favour of the latter, i.e., /wat/ ‘way’ and /kat^h/ ‘story’ are realized as [wat^h] and [kat^h] [14]. An interesting question is then what kind of cues (or at least acoustic correlates) are employed to maintain the underlying contrast when final aspiration occurs?

1.3. Final obstruents in Shughni

As recent studies have shown, plosives in Shughni are normally released in final position, and the preceding vowel is significantly longer before the underlyingly voiced consonant [15]. Final obstruents may be devoiced word-finally (F_0 present for less than half of the closure); it is less probable in careful speech. Interestingly, even if the final closure is fully voiced, the burst can be followed by a period of voiceless aspiration, see Fig. 1.

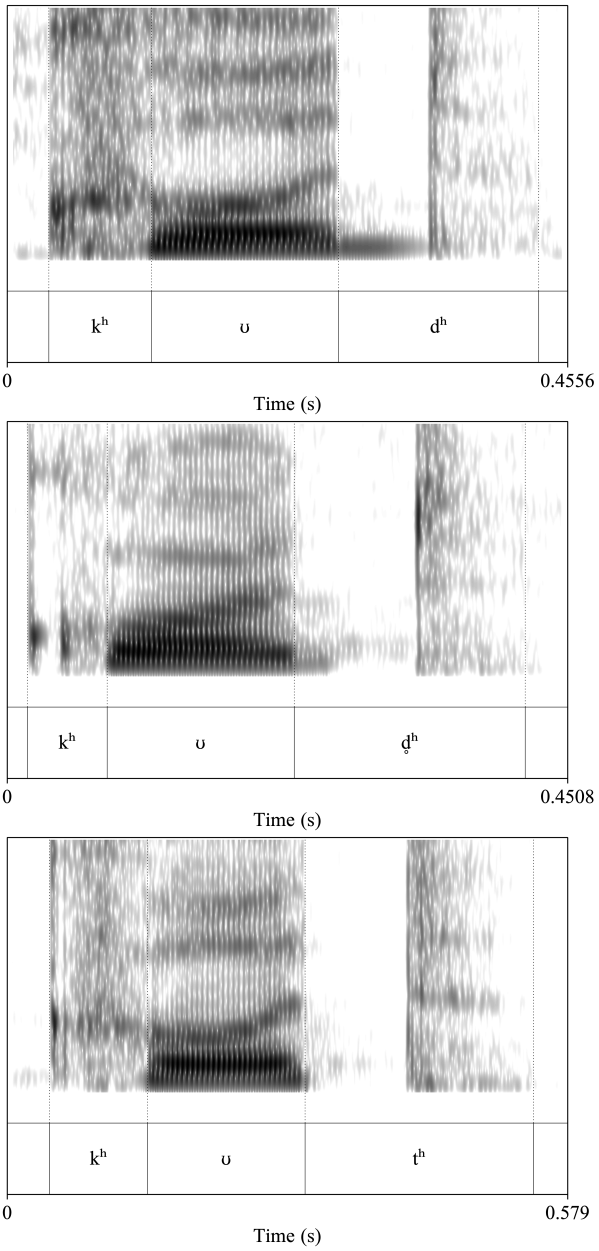


Figure 1: Dynamic spectrograms for /kʊd/ 'dog' (produced with a fully voiced and a devoiced final plosive) and /kʊt/ 'short.' Note voiceless aspiration in all cases. The view range is 0 to 5500 Hz.

The difference between /kʊd/ 'dog' and /kʊt/ 'short', though, impressionistically remains quite perceivable: the aspiration of final /t/ is 'stronger.'

1.4. Research purpose

In this paper, it is proposed that aspiration duration and intensity are the acoustic correlates of the underlying voicing contrast in word-final plosives in Shughni. The hypothesis is that underlyingly voiceless plosives have a longer and more intensive period of aspiration than voiced ones. The purpose of the next two sections is to test this hypothesis statistically. After that, implications for voiced/voiceless distinction and the typology of final laryngeal neutralization in phonetic theory are discussed.

2. DATA AND THEIR ANALYSIS

Data for this research were collected during fieldwork in Khorugh in 2022.

2.1. Speakers

14 native speakers of Shughni participated in the recording (mean age 28.9, SD 18.8; 12 females, 2 males). In accordance with Section 1.1, they all were bi- or multilingual. The speakers were not paid for participation, though received souvenirs.

2.2. Recording procedure and stimuli

The speaker and researcher sat in a quiet room. The researcher asked the speaker to translate a Russian word into Shughni. The stimuli were given along with fillers (they were Shughni words of the CVC structure ending in other obstruents and sonorants); the wordlist was obtained using *pamiri.online* [16]. If the speaker faced some trouble, an additional Russian synonym was given. Once the speaker suggested the intended translation, they were asked to repeat it three times with pauses between each of the utterances.

Shughni stimuli along with their English translations are given in Table 2.

Shughni	English
/bad/	bad
/bat/	embrace
/jed/	bridge
/jet/	open
/vud/	(it) was
/but/	footwear
/ba:d/	then
/ba:t/	a kind of national dish; bat
/kʊd/	dog
/kʊt/	short

Table 2: IPA transcriptions of the Shughni stimuli with their English translations.

A dynamic headset microphone (Shure WH20 XLR), connected to a digital recorder (Tascam DR-40x), was used to record utterances as 44.1 kHz 16-bit mono WAV files.

2.3. Data annotation

Only correctly pronounced words were annotated in Praat [17]. An utterance was considered mispronounced if there was no pause at its end or if the word was mumbled.

Every final plosive was segmented into closure, burst and aspiration; vowel duration was also annotated so as to be used as a reference point in relative measurements.

2.4. Data analysis

All of the annotated parts—namely vowel, final plosive's closure, burst and aspiration—were extracted from the recordings and represented as separate Sound objects in Praat. After that, duration and intensity were extracted using a Praat script. The `Get intensity (dB)` function was used in Praat for measuring intensity [18].

On average four pronunciations for a speaker were obtained: one when suggesting a Shughni translation and three when repeating in isolation. One of the speakers occasionally repeated the translation four times (resulting in 5 pronunciations in total); I decided to include all of them. The resulting spreadsheet with the data used in this study is available in [19].

In addition to the absolute values of the duration and intensity of aspiration, relative values were obtained using (1). Unlike absolute values, relative ones are expected to remain valid even when comparing two pronunciations with different intensity (due to dissimilar recording circumstances and/or manners of speaking): one with overall greater intensity and another with overall lesser one. While absolute measurements will likely show a significant difference, relative ones will not.

(1) $Relative = \frac{Absolute}{Vowel}$, where *Vowel* is for the absolute duration or intensity of the preceding vowel.

I used linear mixed models from the *lme4* package [20] in R [21] to test the hypothesis from Section 1.4 statistically. The final plosive (/d/ or /t/) was considered the fixed effect, and the intercepts for word meaning, speaker and utterance number were included as random effects. The formula is shown in (2).

(2) $DurationORIntensity \sim FinalPlosive + (1 \mid WordMeaning) + (1 \mid Speaker) + (1 \mid UtteranceNumber)$

2.5. Limitations

The dataset has a few shortcomings. First, only one place of articulation (alveolar plosives) was taken into consideration. Second, careful speech was elicited, and, according to some sources [22], 'hyperarticulated speech may transform unreleased stops into released stops.' Third, there is a gender imbalance in favour of female speakers; as for the age, half of the speakers are 13 to 16; the other half is above 30 so at least two generations are represented.

Nevertheless, all these limitations, at least impressionistically, do not seem to be critical. Other places of articulation appear to have similar tendencies, and citation forms, while confined to specific contexts, are eligible subject matter.

3. RESULTS

Overall, the assumed presence of the [voice] feature in the final plosive significantly affected duration and intensity either in absolute or relative values. Table 3 contains the details.

All p-values (i.e., Pr(>|t|) values) are far below 0.01, though since they simply indicate that the estimates of the intercepts and differences between /t/ and /d/ are significantly different from 0, these data are of little interest for the current research purpose.

Fig. 2 depicts error bars for the reported measurements.

Response	Estimate	Standard error
Absolute Duration	0.044 s	0.004
Relative Duration	0.385	0.031
Absolute Intensity	2.877 dB	0.665
Relative Intensity	0.047	0.008

Table 3: Estimates of the difference between the expected values for /t/ and those for /d/ along with standard errors for each response obtained from the linear mixed models in Section 2.4.

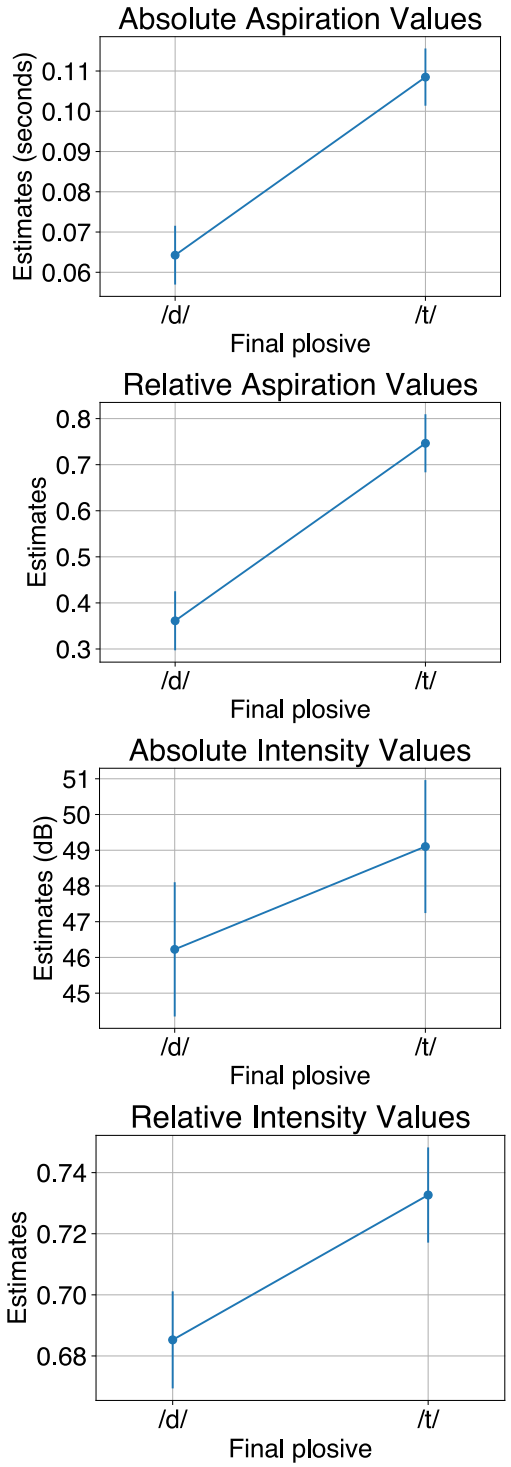


Figure 2: Error bars for each type of response.

4. DISCUSSION

The purpose of this study was to find out whether the duration and intensity of aspiration of final plosives in Shughni serve as acoustic correlates of underlying voicing and, hence, are potential acoustic cues to it. The hypothesis, based on both fieldwork observations and typological expectations, was that voiceless plosives would exhibit higher values either for duration or intensity compared to voiced ones.

The analysis of the data reveals that, regardless of the means of measurement (expressed in absolute or relative values), in final position, voiceless plosive /t/ has a longer period of aspiration with greater intensity. The difference between the predicted values of the absolute duration of aspiration in /t/ and /d/ is 44 ± 4 ms. It seems to be perceptible since such a duration is comparable to half of the vowel in /kʊt/ 'short.' Moreover, the just noticeable difference (JND) in discrimination of duration is estimated as a square root of the full duration of the stimulus (if this duration is lower than 100 ms) [23], [24] so a difference of 44 ms is quite discernible. The same is true for the difference of 2.877 ± 0.665 dB, with minimal values of the corresponding JND ranging from 0.3 to 1.0 dB [ibid.]. As for the relative values, their estimated differences are all significantly different from 0.

From a linguistic perspective, these results have three sides: one language-specific and two relevant to phonological typology.

The language-specific outcome is that Shughni final plosives (alveolar ones were tested, other places are assumed to have similar trends) may not be completely neutralized in final position. Although final devoicing sometimes occurs, underlyingly voiceless plosives have a longer period of aspiration of greater intensity in comparison to underlyingly voiced ones. This is a potential acoustic cue (among other candidates, see Section 1.3) and an additional study of Shughni plosives is needed to assess its role in perception. While the existing descriptions of Shughni do not agree on whether Shughni voiceless plosives are aspirated [25], [26], final aspiration supports adding [spread glottis] to their phonological specification.

As for the typological relevance, aspiration duration and intensity as acoustic correlates of underlying voicing have not been thoroughly studied (cf., for instance, [7], [22] and also Section 1.2 above). The Shughni data give evidence for adding them to the inventory of potential acoustic cues to [voice] in final position.[1]

Lastly, the current study provides support to the expectation that in a language with final aspiration (like Kashmiri above, see Section 1.2) the neutralization process may be incomplete due to the difference in aspiration duration and intensity. The similarity between the voiced unaspirated series (like

[1] This is also supported by the studies on final devoicing in German [11], [27], [28]; all of these, though, lack the instrumental measurement of intensity, and the difference in aspiration duration was generally less significant than found by my study. I thank Timofey Mets for drawing my attention to that.

that of Shughni) and the voiceless unaspirated one (like that of Kashmiri) is that neither of them is specified with the feature [spread glottis] (taken as the privative one, cf. 'laryngeal realism' [13], [29]). The Shughni pattern suggests that mere phonetic ('automatic') aspiration not reinforced with the underlying [spread glottis] feature is by default less prominent, especially in time and intensity domains.

5. ACKNOWLEDGEMENTS

I am indebted to Dr. George Moroz for his help with data analysis as well as to the Shughni speakers and the University of Central Asia, without whom this research would not have been possible. I am also grateful to the two anonymous reviewers, whose informative comments helped a lot, and Dr. Pavel Iosad, who pointed out the role of intensity in this case of incomplete neutralization. The publication was prepared within the framework of the Academic Fund Program at HSE University (grant № 22-00-034 'Field and Corpus Research on the Languages of the Shughni-Rushani Group').

6. REFERENCES

[1] D. (Joy) I. Edelman and L. R. Dodykhudoeva, 'Shughni', in *The Iranian languages*, G. Windfuhr, Ed., London and New York: Routledge, 2009, pp. 787–824.
[2] J. Kingston and R. L. Diehl, 'Phonetic Knowledge', *Language*, vol. 70, no. 3, pp. 419–454, 1994, doi: 10.1353/lan.1994.0023.
[3] L. J. Raphael, 'Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English', *J. Acoust. Soc. Am.*, vol. 51, no. 4B, pp. 1296–1303, Apr. 1972, doi: 10.1121/1.1912974.
[4] R. F. Port and J. Dalby, 'Consonant/vowel ratio as a cue for voicing in English', *Percept. Psychophys.*, vol. 32, no. 2, pp. 141–152, Mar. 1982, doi: 10.3758/BF03204273.
[5] T. B. Roettger, B. Winter, S. Grawunder, J. Kirby, and M. Grice, 'Assessing incomplete neutralization of final devoicing in German', *J. Phon.*, vol. 43, pp. 11–25, Mar. 2014, doi: 10.1016/j.wocn.2014.01.002.
[6] R. L. Trask, *A dictionary of phonetics and phonology*, Reprinted. London; New York: Routledge, 2006.
[7] D. Steriade, 'Phonetics in Phonology: The Case of Laryngeal Neutralization', 1997.
[8] W. S.-Y. Wang, 'Transition and Release as Perceptual Cues for Final Plosives', *J. Speech Hear. Res.*, vol. 2, no. 1, pp. 66–73, Mar. 1959, doi: 10.1044/jshr.0201.66.
[9] J. E. Flege and J. Hillenbrand, 'A differential effect of release bursts on the stop voicing judgments of native French and English listeners', *J. Phon.*, vol. 15, no. 2, pp. 203–208, Apr. 1987, doi: 10.1016/S0095-4470(19)30548-0.
[10] M. Halle, G. W. Hughes, and J. -P. A. Radley, 'Acoustic Properties of Stop Consonants', *J. Acoust. Soc. Am.*, vol. 29, no. 1, pp. 107–116, Jan. 1957, doi: 10.1121/1.1908634.
[11] R. F. Port and M. L. O'Dell, 'Neutralization of syllable-final voicing in German', *J. Phon.*, vol. 13, no. 4, pp. 455–471, Oct. 1985, doi: 10.1016/S0095-4470(19)30797-1.
[12] D. B. Fry, *The physics of speech*. in Cambridge textbooks in linguistics. Cambridge: Cambridge University Press, 1996.
[13] G. K. Iverson and J. C. Salmons, 'Final Devoicing and Final Laryngeal Neutralization', in *The Blackwell Companion to Phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds., Oxford, UK: John Wiley & Sons, Ltd, 2011, pp. 1–22. doi: 10.1002/9781444335262.wbctp0069.
[14] S. M. Syeed, 'The Himalayan way of breathing the last: On the neutralizing status of the word-final aspiration in Kashmiri', *East. Anthropol.*, no. 31, pp. 531–541, 1978.
[15] Y. Makarov, 'Plosives in Shughni', HSE University, Moscow, 2022. [Online]. Available: https://www.researchgate.net/publication/361584042_Realizacia_vzryvnyh_soglasnyh_v_sugnanskom_azyke_Plosives_in_Shughni
[16] Y. Makarov, M. Melenchenko, and D. Novokshanov, 'Digital Resources for the Shughni Language', in *Proceedings of The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 61–64. [Online]. Available: https://aclanthology.org/2022.eurali-1.9
[17] P. Boersma and D. Weenik, 'Praat: doing phonetics by computer [Computer program]'. Apr. 15, 2021. [Online]. Available: http://www.praat.org/
[18] ppgb, 'Sound: Get intensity (dB)', *Praat Manual*, Jan. 29, 2007. https://www.fon.hum.uva.nl/praat/manual/Sound__Get_intensity__dB_.html
[19] Y. Makarov, 'Raw data (measurements)'. Unpublished, 2023. doi: 10.13140/RG.2.2.23847.62888.
[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models Using lme4', *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.
[21] R Core Team, 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/
[22] J. Blevins, *Evolutionary Phonology: The emergence of sound patterns*. New York: Cambridge University Press, 2004.
[23] K. N. Stevens, *Acoustic phonetics*. in Current studies in linguistics series, no. 30. Cambridge, Mass.: MIT Press, 2000.
[24] E. Koffi, 'A Comprehenisve Review of the Acoustic Correlate of Duration and Its Linguistic Implications', *Linguist. Portf.*, vol. 10, 2021, [Online]. Available: https://repository.stcloudstate.edu/stcloud_ling/vol10/iss1/2
[25] K. Olson, *Shughni Phonology Statement*. SIL International, 2017.
[26] D. Karamshoev, *Badzhuvskij dialekt shugnanskogo jazyka [Bajuwi dialect of Shughni]*. Dushanbe: Izdatel'stvo AN Tadzhykskoj SSR, 1963.
[27] M. Fourakis and G. K. Iverson, 'On the "Incomplete Neutralization" of German Final Obstruents', *Phonetica*, vol. 41, no. 3, pp. 140–149, May 1984, doi: 10.1159/000261720.
[28] M. O'Dell and R. Port, 'Discrimination of word-final voicing in German', *J. Acoust. Soc. Am.*, vol. 73, no. S1, pp. S31–S31, May 1983, doi: 10.1121/1.2020331.
[29] P. Honeybone, 'Germanic Obstruent Lenition; some mutual implications of theoretical and historical phonology', PhD Thesis, University of Newcastle upon Tyne, Newcastle upon Tyne, 2002.

GEMINATE TYPES AND THEIR ACOUSTIC EFFECTS ON ADJACENT VOWELS IN HUNGARIAN

Tilda Neuberger

Hungarian Research Centre for Linguistics
neuberger.tilda@nytud.hu

ABSTRACT

The present study examines how phonologically different geminates (i.e., lexical, assimilated, and concatenated types) are implemented phonetically in Hungarian, with particular attention to the temporal and spectral parameters of the adjacent vowels. By investigating geminates in two positions (word-internally and at word boundary), the influence of boundary on geminate production is also observed. Results show that differences in the autosegmental representation of geminate types are mirrored in the durational and amplitude features of the consonants themselves and their neighbouring vowels. Moreover, the temporal properties of geminates seem to be affected by boundary type. However, formant frequencies of adjacent vowels did not show consistent change depending on geminate types. Findings shed more light on the relationship between abstract phonological categories and continuous phonetic details.

Keywords: consonant length, geminate types, vowel acoustics, autosegmental representation, Hungarian

1. INTRODUCTION

Many genetically unrelated languages use consonant length contrastively (e.g., Bengali [12], Hungarian [5, 15], Italian [7, 17], Japanese [8], Tashlhiyt Berber [7, 19]). Previous studies were consistent in reporting significant durational differences between singletons and geminates ([19] and the references listed there). With respect to other potential attributes of length, such as preceding vowel duration, VOT, f0, or amplitude, findings are not as consistent across languages [8, 12, 16, 19]. Therefore, it is claimed that the primary acoustic correlate of geminates is the increased duration of the closure or constriction. This supports the representation of the autosegmental framework (e.g., [14]), i.e., differences between singletons and geminates are represented as differences on the timing tier: one vs. two timing slots, respectively. However, the 'geminate' category is not homogeneous in the sense that it groups together long consonants derived from different sources phonologically. The autosegmental approach allows distinguishing three geminate types with different underlying representations, as illustrated on Fig. 1. Lexical geminates are part of the phonemic inventory of a language. Assimilated geminates derive from assimilatory processes, e.g., voicing assimilation. Concatenated ("fake") geminates are merged sequences of identical consonants.

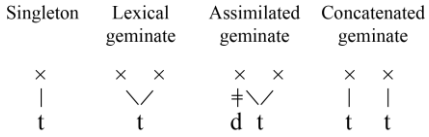


Figure 1: The autosegmental representation of singletons and three geminate types.

Studies dealing with the durational features of geminate types have yielded inconclusive results across languages, potentially reflecting language-specific characteristics, e.g., [9, 12, 17, 19].

No significant differences were found between the three above-mentioned geminate types in preceding vowel duration, closure duration and VOT in Bengali stops [12]. Geminate types did not differ in consonant duration; however, they differed in the duration of the preceding vowel in Libyan Arabic [9] and in Tashlhiyt Berber [19]. In contrast, durational differences in the geminates themselves were found in Italian [17] with concatenated geminates being produced longer than lexical ones.

The differences found between geminate types may be related to the position of the geminates analysed. In languages that do not use consonant length as a distinctive feature but fake geminates occur across morpheme boundaries, via affixation, in compounds or across word boundaries (like in English or German), it was observed that word-boundary geminates were frequently pulled apart in careful speech via f0 changes and pause insertion [2, 10, 16]. It is assumed that differences between geminate types may be due to the boundary-dependent lengthening process in case of fake geminates [16]. Assuming that the boundary strength is not a binary (as in SPE phonology: strong/'word', or weak/'morpheme' boundary [4]) but a gradual property, it was found that the more segmentable the prefix the longer the consonant duration [2]. The question arises whether boundary types have an

influence on the secondary acoustic correlates of gemination (e.g., adjacent vowel duration) as well.

Hungarian is promising ground for investigating this question since it is an agglutinative language with various geminate types occurring at different boundaries in a variety of vowel contexts. Although the durations of geminate types have been examined under uncontrolled conditions (spontaneous speech) in Hungarian [15] and results indicated differences in the consonants themselves (showing short lag VOT as irrelevant in length contrast), secondary acoustic correlates have not been investigated so far. The present study examines the phonetic correlates of geminate types with particular attention to the surrounding vowel context in controlled material. It is hypothesised that (i) different geminate types show durational and/or spectral dissimilarity; (ii) the position/boundary type has an influence on geminate production, resulting in the same geminate type being articulated differently at morpheme boundary within a word than across word boundaries (e.g., indicating the word boundary by incomplete neutralization in voicing assimilation or by pulling the segments apart in concatenation).

2. METHODS

The research material consisted of singletons (S), lexical geminates (Glex), assimilated geminates (Gass) and concatenated geminates (Gcon). The latter two types appeared in two positions: word-internal (WI) and at word-boundary (WB). WI geminates arose from consonant sequences that span a morpheme boundary within a word (root+suffix, weak boundary), whilst WB geminates appeared at the connection of two lexical free morphemes composing a phrase (word#word, strong boundary).

For each geminate type we chose meaningful words/phrases where the target sounds occurred in the same vowel context. For frequency and distributional reasons, the consonant type under study was the (short/long) voiceless alveolar stop [t], and the preceding and following vowels were [ɒ] or [ɛ]: identical V1 and V2 in the VCV sequences. The target words were embedded in 12 sentences as the first words of the sentences, either in preverbal focal position or as verbs of the sentence (Table 1). Word-level stress is highly predictable in Hungarian; always assigned to the initial syllable of a prosodic word [21]. The segments we studied occurred in unstressed syllables except for word-boundary geminates, which concern the first syllable, thus, prosodic effects cannot be ruled out. However, stress tends to affect intensity rather than duration, whereas lengthening may serve as vowel and consonant quantity distinction is this language.

Types	Sentences	Glosses
S	Köve**t**elem [køvɛtɛlɛm] a pénzem.	I demand my money.
Glex	Helye**tt**esít [hɛjɛt:ɛʃi:t] a matektanár.	The math teacher substitutes.
Gass WI	Enge**dt**elek [ɛŋgɛt:ɛlɛk] téged is vezetni.	I let you drive too.
Gass WB	Neke**d t**eszek [nɛkɛt:ɛsɛk] szívességet.	It's you I'm doing a favour for.
Gcon WI	Éle**tt**elen [e:lɛt:ɛlɛn] volt a bolygó.	The planet was lifeless.
Gcon WB	Néme**t t**ervek [ne:mɛt:ɛrvɛk] voltak a táskában.	German blueprints were in the bag.

Table 1: Example sentences ([ɛ] context).

The linguistic material was read aloud in three repetitions by twenty monolingual, Hungarian-speaking adults (10 males and 10 females, aged between 21 and 54 years, mean: 36.7±5 years). None of them reported any speech or hearing disability. Sentences were displayed to participants in random order on a screen using SpeechRecorder [6]. Speakers were instructed to read the sentences in their normal (not careful) speaking style [11]. The experiment was run in a sound-proof booth. The acoustic data were recorded at a sample rate of 44.1 kHz and 16-bit resolution. A total of 720 tokens were collected (12 sentences × 3 repetitions × 20 speakers). The following temporal and spectral parameters generally associated with gemination [7, 8, 19] were analysed using Praat [3]:

- Closure duration (CD, ms): The closure phase was measured between the termination of the preceding vowel and the stop burst.
- Duration of the preceding (V1) and following (V2) vowel (ms): The segmentation of the vowels was based on their second formants supported by visual analysis display of the spectrograms and oscillograms.
- CD/V1 and CD/V2: Closure duration related to preceding and following vowel duration.
- Relative RMS amplitude: A proportional variable measured as the root-mean-square amplitude of the stop release compare to the RMS amplitude of the following vowel.
- First two formants of adjacent vowels (Hz): F1 and F2 were extracted at the temporal midpoint of the vowel using the built-in Burg algorithm in Praat.

The raw data were standardised within speakers using *z*-transformation (for the statistical analysis). Linear mixed models were computed in R [18], using the lmer [1] and lmerTest [11] packages. The dependent variables were the above-mentioned

acoustic parameters; the fixed factor was *geminate type* (including the three geminate types and the two positions), while the random factor was *speaker* (N = 20). *Vowel quality* was incorporated as fixed factor in the models on amplitude and formant frequencies. The effect of *gender* contributed no improvement to the models and was thus excluded during model selection. Pairwise comparisons with Tukey method were performed with emmeans [13]. F-values and corresponding *p*-values were computed using the Satterthwaite method. Plots were made with the ggplot2 package [23], also containing the singleton category as reference.

3. RESULTS

Besides the typical single-articulated, lengthened geminate realisation, rearticulated geminates occurred in 2.3% of the cases. These were produced solely at word boundary (5 assimilated and 9 concatenated geminates by five participants). (These occurrences were excluded from the CD measurements due to two closure phases.)

3.1. Absolute duration

As expected, the average **closure duration** of each geminate type was longer than that of singletons (Fig. 2). The statistical analysis revealed that closure duration was significantly affected by geminate type: $F(4, 586) = 28.968$; $p < 0.001$. For both derived geminate types, position influenced the durations in such a way that word-internal geminates were realised shorter and word-boundary ones longer, on average; however, the difference was not significant ($p > 0.05$).

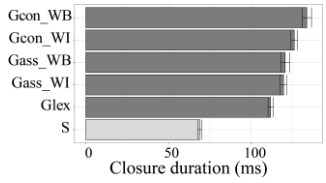


Figure 2: Closure duration of singletons and various geminate types (mean and standard error).

Concerning the **duration of the preceding vowel** (Fig. 3), a significant main effect of geminate types was shown: $F(4, 600) = 26.189$; $p < 0.001$. The V1 before lexical geminates was significantly shorter than V1 before any other geminate types except for Gass_WB ($p < 0.001$ in the three cases). The assimilated geminates did not differ between the two positions, but concatenated ones did ($p < 0.001$).

Geminate type had a significant effect on the **duration of the following vowel** as well (Fig. 3): $F(4, 600) = 337.56$; $p < 0.001$. V2 was realised with significantly longer duration after word-boundary geminates compared to word-internal ones ($p < 0.001$). Consistent differences by boundary types provided evidence for the manifestation of prosodic hierarchy by fine-grained phonetic details. The remarkable difference between the two WB geminates suggests that word-initial lengthening has a stronger effect on vowel duration after sequences of identical consonants than after assimilated geminates (which presumably pull the words together).

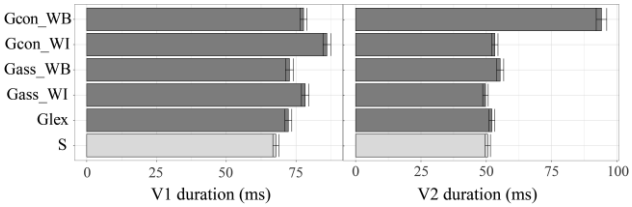


Figure 3: Preceding and following vowel duration as a function of geminate type (mean and standard error).

3.2. Relative duration

The **CD/V1** was higher for all geminate types than for singletons (in which the mean value was 1, indicating that the consonant and the vowel had the same duration, on average). The mean CD/V1 values were the same for the three word-internal geminates (1.5 in each case), but higher for the word-boundary geminates (1.6±0.3 for Gass and 1.7±0.5 for Gcon). Statistical analysis indicated that only Gcon_WB was significantly different from all the word-internal geminates ($p \leq 0.005$ in each contrast). The mean **CD/V2** was higher (2.2-2.5) for most geminate types than for singletons (1.4±0.4), except for the Gcon_WB (1.4±0.3). Statistical analysis showed that geminate types differed in this parameter: $F(4, 586) = 89.751$; $p < 0.001$.

3.3. Spectral parameters

Statistical analysis showed a significant effect of geminate type ($F(4, 586) = 8.878$; $p < 0.001$) and vowel quantity ($F(4, 586) = 18.427$; $p < 0.001$) on **relative RMS amplitude**. The [ɒ] vowel showed higher mean values than [ɛ], but the pattern by geminate type was the same for the two vocalic contexts. Pairwise comparisons confirmed that lexical geminates were realised with significantly higher amplitude than word-internal assimilated ones ($p = 0.004$) and word-boundary concatenated ones ($p < 0.001$). Furthermore, there was a significant difference between the two types of derived geminates at word boundary ($p = 0.024$). Same types of geminates did not differ in amplitude depending on the boundary type.

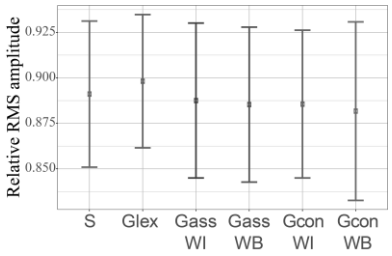


Figure 4: Relative RMS amplitude (mean and standard error).

Concerning the **formant frequencies of the preceding vowel**, a significant interaction effect of geminate type and vowel quality was found on F1 ($F(4, 600) = 34.252$; $p < 0.001$) and on F2 ($F(4, 600) = 6.088$; $p < 0.001$). This indicates that the change in F1 and F2 by geminate types was different depending on the vowel: Larger differences between geminate types were observed in the case of [ɛ] than in [ɒ]. Moreover, F1 values varied more widely than F2 values as a function of geminate types (Fig. 5). F2 values remained mostly unaltered regardless of geminate type.

Turning to the **following vowel** (Fig. 5), a similar picture was shown as in V1. Formant frequencies varied strongly with type of the geminate in [ɛ], but not in [ɒ] – in accordance with the significant interaction of geminate type and vowel quality in F1 ($F(4, 600) = 33.967$; $p < 0.001$) and F2 ($F(4, 600) = 21.624$; $p < 0.001$). In sum, [ɛ] appeared with higher F1 values in the lexical and assimilated geminate context than in the concatenation geminate context.

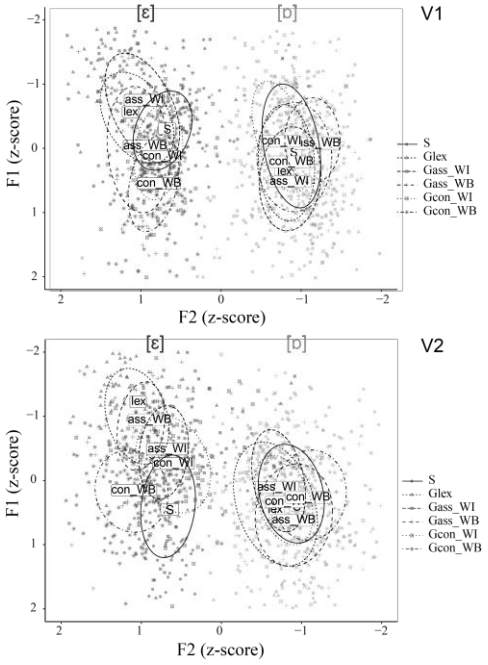


Figure 5: Formant frequencies of the preceding and following vowels as a function of geminate type.

4. DISCUSSION AND CONCLUSION

This study examined how phonological representations of geminate types are reflected in phonetic data. Besides the three commonly studied geminate types [9, 12, 19], we distinguished between types based on position to explore possible boundary-dependent and/or prosodically conditioned realisations [2, 16, 20].

Our first hypothesis was confirmed by the data, showing that the type of geminate significantly affects closure duration, adjacent vowel duration, relative duration and amplitude. Durational results are consistent with those found in Italian [17] (a syllable-timed language, like Hungarian), indicating that lexical geminates are realised with shorter duration than post-lexical ones. V1 and RMS results are in line with those found in Tashlhiyt Berber [19], showing that lexical geminates are produced with higher burst energy and with shorter V1 compared to concatenated ones. The frequencies of the first two formants of adjacent vowels did not show consistent change depending on geminate types. F1 variation by geminate types implies that their difference might be associated with the vertical movement of the tongue during adjacent vowel production. However, it is not uniform across vowel qualities and may be affected by prosodic factors or by the place of articulation of the other adjacent consonants. There may also be a non-local gemination effect, but more research is needed to understand this phenomenon in Hungarian, as previous studies suggests language-specific articulatory mechanisms in this respect [22].

As for the second hypothesis, results confirmed that the temporal properties of geminates are affected by boundary type. Even geminates of the same type were realised differently depending on whether they occurred at a weak/morpheme boundary (WI) or a strong/word boundary – in agreement with the results for English fake geminates [16]. Word-boundary concatenated geminates differed consistently from other geminates and exhibited longer CD, longer V2, higher CD/V1 ratio, lower CD/V2 ratio, and lower RMS amplitude. In some cases, word boundary blocked full assimilation in derived geminates by pulling the segments apart. These findings indicate a stronger coherence of articulation within words than across word boundaries, similarly to what was observed in case of two-term consonant clusters [20]. These findings may contribute to a better understanding of the relationship between distinct phonological categories and continuous phonetic variables, and shed more light on the role of morphological structure and prosodic factors in gemination.

5. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48.

[2] Ben Hedia, S., Plag, I. 2017. Gemination and degemination in English prefixation: Phonetic evidence for morphological organization. *Journal of Phonetics*, 62, 34–49.

[3] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.43, http://www.praat.org/

[4] Chomsky, N., Halle, M. 1968. *The sound pattern of English*. Cambridge, MA: MIT Press.

[5] Deme, A., Bartók, M., Grászi, T. E., Csapó, T. G., Markó, A. 2019. Articulatory organization of geminates in Hungarian. In *Proc. of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, 1739–1743.

[6] Draxler, C., Jänsch, K. 2004. SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. In *Proc. of the IV. International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 559–562.

[7] Hermes, A., Tilsen, S., Ridouane, R. 2021. Cross-linguistic timing contrast in geminates: A rate-independent perspective. In *Proc. of the 12th International Seminar on Speech Production (ISSP2020)*, 52–55.

[8] Hirata, Y., Whiton, J. 2005. Effects of speaking rate on the single/geminate stop distinction in Japanese. *The Journal of the Acoustical Society of America*, 118(3), 1647–1660.

[9] Issa, A. 2015. On the phonetic variation of intervocalic geminates in Libyan Arabic. In *ICPhS*. Paper. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proc. of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 0564.1–5.

[10] Kotzor, S., Molineaux, B. J., Banks, E., Lahiri, A. 2016. "Fake" gemination in suffixed words and compounds in English and German. *The Journal of the Acoustical Society of America*, 140(1), 356–367.

[11] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26. http://doi.org/10.18637/jss.v082.i13.

[12] Lahiri, A., Hankamer, J. 1988. The timing of geminate consonants. *Journal of Phonetics*, 16(3), 327–338.

[13] Lenth, R., Lenth, M. R. 2018. Package 'lsmeans'. *The American Statistician*, 34(4), 216–221.

[14] McCarthy, J. J. 1986. OCP effects: Gemination and antigemination. *Linguistic inquiry*, *17*(2), 207–263.

[15] Neuberger, T. 2015. Durational correlates of singleton-geminate contrast in Hungarian voiceless stops. In *ICPhS*. Paper. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proc. of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 0422.1–5.

[16] Oh, G. E., Redford, M. A. 2012. The production and phonetic representation of fake geminates in English. *Journal of phonetics*, 40(1), 82–91.

[17] Payne, E. M. 2005. Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association*, 35(2), 153–181.

[18] R Core Team. R: A language and environment for statistical computing. 2015 Retrieved from https://www.r-project.org/

[19] Ridouane, R. 2010. Geminates at the junction of phonetics and phonology. *Papers in laboratory phonology*, 10, 61–90.

[20] Slis, I. H. 1986. Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of speaker and listener. *Journal of Phonetics*, 14(2), 311–326.

[21] Szalontai, Á., Wagner, P., Mády, K., Windmann, A. 2016. Teasing apart lexical stress and sentence accent in Hungarian and German. In *12. Tagung Phonetik und Phonologie im deutschsprachigen Raum.* LMU Munich. 215–218.

[22] Turco, G., Braun, B. 2016. An acoustic study on non-local anticipatory effects of Italian length contrast. *The Journal of the Acoustical Society of America,* 140(4), 2247–2256.

[23] Wickham, H. 2016. *ggplot2 – Elegant Graphics for Data Analysis (2nd Edition).* Springer-Verlag, New York.

REVISITING CV TIMING WITH A NEW TECHNIQUE TO IDENTIFY INTER-GESTURAL PROPORTIONAL TIMING

Karthik Durvasula and Yichen Wang

Michigan State University
durvasul@msu.edu, wangy176@msu.edu

ABSTRACT

In this article, we probe the timing relationship in CV sequences. To do so, we develop a new technique that is sensitive to the distinction between proportional lag relationships and constant lag relationships, which is a problem for previous explorations of CV lag. Using the new technique, we show that our data is consistent with the vowel onset being synchronised with the preceding consonant offset (what we call "offset-onset alignment"). We also observed more stability with target onset alignment than with gesture onset alignment. Our results suggest that speakers might be planning target achievement more so than alignment of gesture onsets. Furthermore, they suggest a need for a new explanation for the C-centre effect.

Keywords: CV timing, in-phase, anti-phase, offset-onset alignment, X-Ray microbeam data.

1. INTRODUCTION

In Articulatory Phonology [1, 2, 3, amongst others], inter-gestural timing between two gestures (G_1 and G_2) boils down to one of three phasal relationships, where an internal clock cycle of a gesture is represented by 0°-360°: (a) *in-phase*: when gestures start simultaneously (0°phase difference); (b) *anti-phase*: when one gestural onset is perfectly asynchronous with another gestural onset (180° phase difference); (c) *eccentric phase*: gestures have a phasal relation that is different from 0° and 180°.

With the above timing relationships as backdrop, one can layout three different views of CV timing that have been directly or indirectly argued for in prior research. The first hypothesis is what might be called the standard position [1, 2, 4] — the consonant gesture (G_1) and the vocalic gesture (G_2) are in-phase (Figure 1a). When an additional claim that consonants in CCV sequences are in an anti-phase relationship is added, then this first viewpoint of CV timing is able to predict that C-CENTRE-TO-ANCHOR interval stability pattern wherein the consonantal gestures in a complex onset are in a global timing relationship with the subsequent vowel [5, 6, 7, 8, amongst others].

A second hypothesis about CV timing depends on a generalisation of the *split-gesture hypothesis*, whereby a stop gesture is decomposed into two gestures related to the closure and release portions, respectively [9]. This hypothesis was used by Nam [9] to account for the fact that impressionistic observations of prior work suggested that the *centre* of the observed stop gesture, and not the onset, was in a stable temporal relationship with the following vowel. A similar view of a separately manipulable neutral attractor gesture for all segment gestures was discussed in earlier research work [10, 11, 12]), and, in fact, a similar observation about the timing of CV sequences that, impressionistically, the V gesture starts roughly mid-way through the observed consonant gesture can be seen in some of the earliest work on gestural timing [5] not just for stops, but also liquids and fricatives. Therefore, one could perhaps generalise Nam's split-gesture hypothesis to all segments, and claim that vowel gestures start mid-way through any preceding consonantal gesture. However, the actual simulations that Nam presented do not bear this prediction out, even for stops. His simulations predict that the vowel gesture has a 60° phase difference with a preceding consonant (stop) closure gesture and a phase difference of -60° with a preceding (stop) release gesture, when an equal coupling strength is assumed for all timing relationships. Consequently, it is straightforward to show that, if each gesture needs about 240°-360° of its internal clock duration to attain the target (see [2] for the lower bound), the vowel itself is expected to start about 12.5%-16.7% of the entire observed consonant (or specifically, stop) gesture duration, and not mid-way through the consonant gesture (derivation not presented in the interest of space).

A third view about CV timing can be seen as a generalisation of the work by Shaw, Durvasula, Kochetov and Oh [13, 14] who argue that, in consonant sequences, G_2 is aligned to the end of G_1 — we call this an *offset-onset alignment*

relationship. If we were to generalise this relationship to all segment sequences, we would predict that in CV sequences, the vowel gesture is in an *offset-onset alignment* relationship with the preceding consonant gesture, *i.e.*, the proportional lag is 100% between the two gestures (Figure 1b).

Most of the evidence furnished in support of in-phase timing in CV sequences is either based on observed lag (see next section) or based on general work on motor control [15, 10]. For example, in a detailed quantitative/experimental evaluation of this question done quite recently, [16, 4] develop a new experimental paradigm that uses triplets of stimuli (*e.g.*, [lolju] vs. [loju] vs. [lolu], which allowed them identify the actual gestural onset of the consonant and vowel gestures through a comparison of minimal pairs. They found that the labial gesture of an [u] starts in-phase with the preceding consonant [l] gesture, but the tongue tip gesture of the same vowel starts towards the end of the consonant gesture. They interpret this result as consistent with in-phase CV timing, with the caveat that a vocalic gesture (*e.g.*, tongue tip) that conflicts with the preceding consonant gesture is delayed and starts during the offset of the consonantal gesture. The comparative technique they employ has long been argued to allow one to separate out the effect of context [17, 18, 19]; however, there are three issues with their technique/interpretation. First, based on their data, one could have also argued for the opposite conclusion that the vowel gesture starts during the offset of the consonantal gesture, but the rounding of the vowel gesture starts earlier, *i.e.*, the in-phaseness is specific to rounding. In fact, Benguerel and Cowan [17] observe that rounding in French starts up to 6 segments before the vowel (across syllables and even words), even when one uses a comparative technique to identify the onset of the gesture. So, it is possible that the rounding pattern is simply a language-specific fact or assimilation in English and French. Second, an issue with the comparative technique is that it requires *phonetic* minimal pairs, and not phonological minimal pairs. For example, if there is fronting of [u] after [l] (see [20]), then [lu] doesn't form a good minimal pair with [ju]. Third, the technique doesn't allow us to separate absolute lag from proportional lag (see next section).

2. A NEW WAY TO IDENTIFY PROPORTION LAG BETWEEN TWO GESTURES

The main technical issue that we address in this paper is that any observed lag (*i.e.*, the actual duration of time) between gestures can be due to two different sources: (a) a proportional (relative) temporal lag between the two gestures, wherein G_2 starts at a certain proportional point of G_1. Note, this is the temporal relationship represented by the phasal relationships discussed above; (b) an absolute (or constant) temporal lag between two gestures, wherein G_2 starts at a constant time in relation to G_1. Absolute lag may be a result of planning, or biomechanical factors [21, 8], or due to differences in measurement errors related to different gestures.

Given that there are two sources that can contribute to the observed lag between gestures, one can't simply look at the observed average lag between two gestures to establish the phasal relationship between them. For example, G_2 onset could be timed in-phase with G_1 but with a positive absolute lag (Figure 1c), or G_2 onset could be timed to the end of G_1 but with a negative absolute lag (Figure 1d). Both the alternatives would superficially look the same. If we calculated the average observed lag between the onsets of the two gestures and then used that duration to calculate the proportional lag (compared to G_1 duration), we would get the same proportional lag value in both cases. This is quite problematic as the cases have quite different underlying temporal organization of the gestures, while one has in-phase alignment the other has offset-onset alignment.

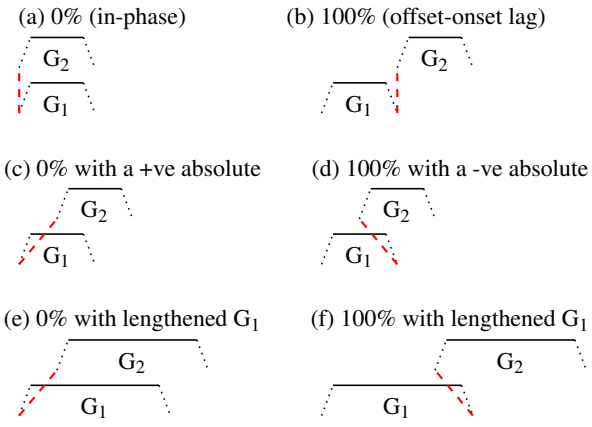


Figure 1: Different timing possibilities between two gestures and their consequences for a lengthened G_1 (proportional lag point = red dashed line)

One way to solve this measurement issue is to recognise that the two possibilities in Figure 1c-d make different predictions for how the simple lag duration *co-varies* with a change in G_1 duration. In the first case (represented in Figure 1e), the observed lag from G_1 onset to G_2 onset remains constant. In the latter case (represented in Figure 1f), the

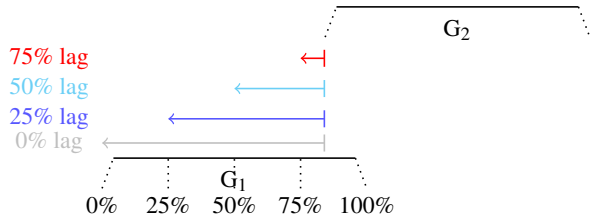


Figure 2: Measuring lag between G_2 and G_1 from different points along G_1

observed lag from G_1 onset to G_2 onset increases, but that from G_1 *offset* to G_2 onset remains constant. That is, in both cases, the observed lag between G_2 onset and a point along the duration of G_1 has the *least variance* when the observed lag is measured from the point along G_1 that G_2 is aligned to.

So, to identify the proportional lag between G_1 and G_2, we can calculate the observed lag between G_2 onset and different proportional points along G_1's duration. This is shown schematically in Figure 2. We can then calculate the variances of the measurements for each of these measured lags to identify the proportional point along G_1 with the lowest variance for the lag measure to G_2 onset.

To test this technique, we created 6 different simulated datasets with random Gaussian noise added to the gesture durations, which differed in the proportional and absolute lags between G_1 and G_2. Specifically, we simulated 3 datasets with an absolute lag of 0 ms that varied in the proportional lag (0% vs. 60% vs. 100% of G_1 duration), and 3 more datasets with an absolute lag of 50 ms and the same three proportional lags. Our approach accurately identified the proportional lags between the two gestures (see Figure 3).

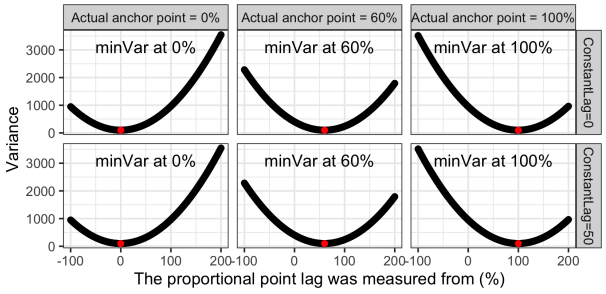


Figure 3: Variances corresponding to different proportional lags on simulated data (red point = proportional lag point with least variance.)

Note, our technique doesn't suffer from the problem of part-whole correlation that plagues the calculation of proportional lag using a ratio of observed lag and G_1 duration [22]. In fact, it is guaranteed to identify the proportional timing relationship between the two gestures as the basic premise of the technique is the statistical theorem: *Variance(X + constant) = Variance(X)* [23]. Essentially, calculating the variances of the observed lags between the two gestures from different points along G_1 allows us to remove the effect of the absolute lag. Note, based on another statistical theorem, it is straightforward to show that the technique works even if the absolute lag has its own variance, as long as the absolute lag and the proportional lag are independent of each other.

3. EXPERIMENTAL DATA

3.1. Methods

In this section, we use the above technique to identify the actual proportional lag in articulatory data. Data were collected from the Wisconsin X-Ray Microbeam Speech Production Database [24], which consists of production data from 57 speakers in three different tasks: wordlists, sentences, and paragraphs. We focussed on words where the initial consonant was a labial, and had multiple repetitions for each speaker with the same preceding context. We further looked for a variety of vowels following the initial consonant. Five words satisfied our requirements: <back, fiber, make, much, people>, where the relevant vowels were [æ, ɑɪ, eɪ, ʌ, i], respectively. Two other words <five, before> came close but were eliminated due to inconsistent contexts. We collected measurements from a total of 789 word productions (back=143,fiber=127, make=171,much=141,people=207).

Lip aperture, defined as the euclidean distance between sensors on the upper and lower lips, was used to track the bilabial gesture; a lower lip ('LL') sensor was used to track the labio-dental gesture; and a tongue blade (either 'T2' or 'T1') sensor was used to track the vocalic gesture. The gestures from each token were parsed using the `findgest()` algorithm in **mview**, a Matlab-based program for data visualization and analysis [25].

For each gesture, we identified the following: (a) gesture onset/offset, identified using a threshold of 10% of the peak maximum, and used, (b) target onset/offset, identified using a threshold of 30% of the peak maximum.

After identifying the onsets and offsets of the gestures/targets, we first calculated the lag duration from 10% steps (range: -50% – 150%) of G_1 duration to the onset of G_2 for the gestural and target measurements, separately. Secondly, for the gestural and target measurements, we calculated the variances for each combination of speaker, word, and step. Finally, we identified the step with the

lowest variance for each combination of speaker and word, also separately for the gestural and target measurements.

3.2. Results

We fitted simple regression models for each of the words, where the dependent variable was the step with the minimum lag for a speaker. Since our interest is simply in identifying the mean step value, each model had only an intercept (therefore, this is equivalent to a one-sample t-test). We also calculated 95% confidence intervals for each of the estimated intercepts. These results are presented in Figure 4. Note, in the interest of space, we don't present tabled summaries, as the crucial information needed for inference is present in the figure.

We focus our discussion on the interpretation of the 95% confidence intervals (CI) for each model [26, 27, 28], instead of just null hypothesis testing of an underlying proportional lag of 0% — this is because our interest here is not only in seeing a consistency (or lack thereof) with the null hypothesis of a proportional lag of 0%, but actually on the range of proportional lag values that are reasonable estimates for the underlying proportional lag. Note, if the CIs don't include a certain proposed lag, that is evidence that the underlying true proportional lag is inconsistent with the proposed lag value.

As can be seen in Figure 4 (left), the onset of the G_2 gesture is not consistent across words/vowels in our data. The 95% CIs suggest the gestural onset of [i] in <people> is consistent with 0% lag (an in-phase alignment with G_1) or with 12.5%-16.7% proportional lag between the onsets of the two gestures (as per the split-gesture hypothesis). However, the 95% CIs of the other four words/vowels are inconsistent with the two hypotheses. Furthermore, two of the three 95% CIs [æ, eɪ] are consistent with the *offset-onset alignment* hypothesis.

In contrast, the 95% CIs in Figure 4 (right) suggest that the proportional timing of onset of the G_2 *target* is consistent only with the offset-onset alignment hypothesis in all five cases.

In fact, when pooled together in a mixed-effects model [29] with a random intercept for speaker and word/vowel, the overall estimated proportional lag for gestural targets was 94% (95% CI: 80% - 107%), which is almost exactly what one would expect as per the offset-onset alignment hypothesis. However, given the by-word/vowel variability for the proportional lags of gestural onsets, it is not surprising that the overall estimated proportional lag for the gestural onsets was 63% with a much wider 95% CI range (33% - 93%).

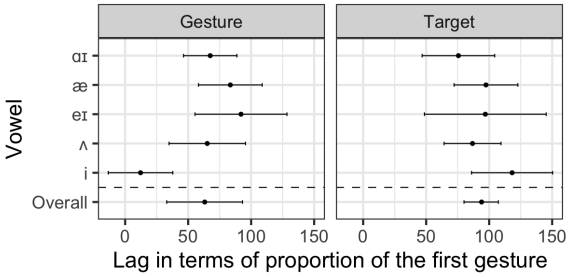


Figure 4: Gestural lag measurements. (points = proportional lag estimates; error bars = 95% CIs)

4. CONCLUSION

In this article, we presented a new technique based on the variance of the observed lag to establish the proportional inter-gestural timing relationship between any two proximate gestures. Using the technique, we observed that vowel gestural onsets were not consistently aligned with the preceding consonant gesture, but the vowel target onsets consistently had offset-onset alignment.

It is important to point out that this is a first attempt using a new technique to establish proportional lag. While the technique is accurate in identifying proportional lags, we caution the reader from over-interpreting our experimental results—while the number of cases here are (at minimum) comparable to previous works looking at the issue either impressionistically or through a careful quantitative study, it is still quite small and is therefore in need of replication.

If replicable and generalisable to other cases, our results have two important implications. First, speakers might actually be planning to coordinate gestural *targets* in a consistent way and not gestural *onsets*. The differences in gestural onsets could be a result of trying to ensure the consistent alignment of gestural targets in a sequential manner. Second, standardly, the C-centre effect noted earlier is seen as a result of competitive coupling wherein consonant gestures in a complex onset are all in an in-phase relationship with the following vowel gesture while the consonant gestures themselves are in an anti-phase relationship. But, if CV timing is an offset-onset relationship then we can't derive the C-centre effect in the standard way. It is, however, possible to derive it from a situation wherein all gestural pairs (CC and CV) in complex onsets have offset-onset timing.

5. REFERENCES

[1] C. P. Browman and L. H. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

[2] ——, *Tiers in articulatory phonology, with some implications for casual speech*, ser. Papers in Laboratory Phonology. Cambridge University Press, 1990, vol. 1, pp. 341–376.

[3] L. Goldstein, "Back to the past tense in english," *Representing language: Essays in honor of Judith Aissen*, pp. 69–88, 2011.

[4] Z. Liu, Y. Xu, and F.-f. Hsieh, "Coarticulation as synchronised cv co-onset–parallel evidence from articulation and acoustics," *Journal of Phonetics*, vol. 90, p. 101116, 2022.

[5] C. P. Browman and L. H. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140–155, 1988.

[6] J. A. Shaw, A. I. Gafos, P. Hoole, and C. Zeroual, "Syllabification in moroccan arabic: evidence from patterns of temporal stability in articulation," *Phonology*, vol. 26, no. 1, pp. 187–215, 2009.

[7] A. Hermes, D. Mücke, and M. Grice, "Gestural coordination of Italian word-initial clusters: The case of 'impure s'," *Phonology*, vol. 30, no. 1, pp. 1–25, 2013.

[8] D. Mücke, A. Hermes, and S. Tilsen, "Incongruencies between phonological theory and phonetic measurement," *Phonology*, vol. 37, no. 1, pp. 133–170, 2020.

[9] H. Nam, "Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure," *Laboratory Phonology*, vol. 9, pp. 483–506, 2007.

[10] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989. [Online]. Available: https://doi.org/10.1207/s15326969eco0104_2

[11] M. E. Beckman and J. Edwards, "Intonational categories and the articulatory," *Speech perception, production and linguistic structure*, p. 359, 1992.

[12] J. Edwards, M. E. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 369–382, 1991. [Online]. Available: https://doi.org/10.1121/1.400674

[13] J. A. Shaw, K. Durvasula, and A. Kochetov, "The temporal basis of complex segments," in *Proceedings of the 19th International Congress of Phonetic, Sciences (ICPhS 2019)*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australasian Speech Science and Technology Association Inc., 2019, pp. 676–680. [Online]. Available: https://assta.org/proceedings/ICPhS2019/papers/ICPhS_725.pdf

[14] J. A. Shaw, S. Oh, K. Durvasula, and A. Kochetov, "Articulatory coordination distinguishes complex segments from segment sequences," *Phonology*, vol. 38, no. 3, pp. 437–477, 2021.

[15] J. Kelso, E. Saltzman, and B. Tuller, "The dynamical perspective on speech production: data and theory," *Journal of Phonetics*, vol. 14, no. 1, pp. 29 – 59, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0095447019306084

[16] Z. Liu, Y. Xu, and F. Hsieh, "Coarticulation as synchronised sequential target approximation: An ema study," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020. International Speech Communication Association (ISCA), 2020, pp. 1381–1385.

[17] A.-P. Benguerel and H. A. Cowan, "Coarticulation of upper lip protrusion in french," *Phonetica*, vol. 30, no. 1, pp. 41–55, 1974.

[18] S. E. Boyce, "Coarticulatory organization for lip rounding in Turkish and English," *The Journal of the Acoustical Society of America*, vol. 88, no. 6, pp. 2584–2595, 1990.

[19] S. E. Boyce, R. A. Krakow, and F. Bell-Berti, "Phonological underspecification and speech motor organisation," *Phonology*, vol. 8, no. 2, pp. 219–236, 1991. [Online]. Available: http://www.jstor.org/stable/4420035

[20] S. Hawkins and S. B. Jones, "The perceptual magnet effect reflects phonetic context," *Journal of the Acoustical Society of America*, vol. 115, no. 5, 2630, 2004.

[21] M.-J. Solé, "Phonetic and phonological processes: The case of nasalization," *Language and Speech*, vol. 35, no. 1-2, pp. 29–43, 1992. [Online]. Available: https://doi.org/10.1177/002383099203500204

[22] W. J. Barry, "Some problems of interarticulator phasing as an index of temporal regularity in speech," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 5, 1983.

[23] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer, 2004, vol. 26.

[24] J. R. Westbury, "X-ray microbeam speech production database user's handbook," 1994.

[25] M. Tiede, "Mview: software for visualization and analysis of concurrently recorded movement data," 2005.

[26] J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 236, no. 767, pp. 333–380, 1937.

[27] J. Cohen, "The earth is round (p<. 05)," *American Psychologist,*, vol. 49, no. 12, pp. 997–1003, 1994.

[28] G. Cumming, *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013.

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using `lme4`," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

MEASURING VARIATION IN CENTRAL PAME VOWELS

Borja Herce, Catalina Torres
University of Zurich

ABSTRACT

Central Pame (cent2145), also known by its endonym Xi'iui, is a threatened Otomanguean language spoken in central Mexico. An earlier impressionistic account described Central Pame as having an asymmetric 5-vowel system with one more front than back vowels. This study examines a set of monophthongal, oral vowels in the language. We explore the acoustic phonetic characteristics in different speakers who are bilingual (with Spanish). Results confirm that five vowel qualities can be determined acoustically. Results show three front /i, e, ɛ/, one central vowel /ɐ/ and one back vowel /o/. We find that the realisations of the front vowel /e/ differ between speakers, with young female speakers showing a more fronted vowel. The evidence suggests a vowel chain shift is taking place in front vowels.

Keywords: Otomanguean, under-documented languages, language contact

1. INTRODUCTION

Central Pame (Oto-Pamean, Otomanguean, cent2145), also known by its endonym Xi'iui, is a language spoken by around 3000 people in and around Santa María Acapulco, in the southern part of the Mexican state of San Luis Potosí. The language was first documented via the work of SIL linguists/missionaries in the 1950s [1], [2], [3], [4]. The description of the language's phonology by Gibson [5] remains the most thorough source available to date. For all its virtues, this description is of necessity dated in its methods, as it described the state of the language 70 years ago. In addition, notable recent changes have been reported in neighbouring related languages [6]. Impressionistic field work observations suggest that front Central Pame vowels may be changing as well.

1.1 Central Pame Vowels

The vowel system of Central Pame was described as an asymmetric 5-phoneme one consisting of four front /i, e, ɛ, a/, and one back vowel /o/ [5]. The instantiations and allophonic ranges of the phonemes were impressionistically described as /i/∼/ɪ/, /i̞/∼/ɛ̞/, /ɛ/∼/æ̞/, /a/, and /ʊ/∼/u/∼/o/∼/ɒ/ respectively. No phonemic distinction between different back vowels was reported despite the wide allophonic domain. Additionally, inter-speaker variation was reported regarding the quality of vowels /i/, /e/, and /ɛ/ in certain words. Additionally, /i/ was described as alternating quite freely with a diphthong /ei/. All vowels were reported to occur phonemically as either oral or nasal, as well as stressed and unstressed. According to [5], no phonemic vowel length exists in the language.

Neighbouring related languages have been described to be relatively similar to Central Pame. Some descriptions of the closest-related one, Northern Pame [7], report a 6-vowel inventory (essentially the one described for Central Pame by [5] plus an additional innovative phoneme /ə/), while others [8] find an additional phonemic distinction between /u/ and /o/. For more distantly related Chichimec, [6] also finds 7 vowels (and a contrast between /o/ and /u/) in the speech of the oldest speakers, but increasing evidence for mergers (of /ɛ/ with /e/, and /o/ with /u/) in the speech of younger speakers more heavily influenced by Spanish. Some of these disagreements and/or findings in related languages highlight uncertainties and points of controversy that also exist in Central Pame. More recent (but much less thorough) accounts of the vowel system of the language, thus, differ from the one in [5]. Accounts in [9] and [10] for example, consider that some words have an invariably /o/ or /u/-like pronunciation of the back vowel, which might point towards a (marginal) phonemic distinction between two different back vowels.

Considering observations on neighbouring languages and the lack of phonetic acoustic

investigations in Central Pame, this study sets out to investigate its vowel system acoustically. Additionally, we examine whether, as reports for Chichimec [6], there is evidence for a potential merger of /e/ and /ɛ/, or other changes to front vowels.

2. METHODS

2.1 Data collection

Fieldwork was conducted in the communities of Santa María Acapulco and the neighbouring La Parada in 2022. Recordings were performed through an H1n Zoom sound recorder at a 48kHz sampling rate. They were performed indoors and in a quiet location of the respective communities. This study is based on data from 6 speakers (4f, 2m) between the ages of 18 and 31.

2.2 Materials

In consultation with native speakers of Central Pame we elaborated a list of 80 target words, which were designed to represent all possible combinations of stress-tone and vowel qualities. An elicitation paradigm was used to prompt participants to utter the selected words in carrier sentences in a naturalistic setting (a market where different objects, animals, etc. are sold). Speakers where exposed to visual stimuli consisting of images, and Spanish translations for clarity.

2.3 Data processing

Sound files were manually transcribed, and force aligned in WebMAUS, using a parameter model based on SAMPA [11]. Phonetic alignment was manually corrected in Praat [12]. A hierarchical database was build using the EMU Speech Database Management system [13]. Utterances containing disfluencies or failing to elicit the intended target word were discarded. Acoustic measures for the first and second formants were taken at vowel mid-point using the emuR [14] package in R [15]. This study examines vowel tokens in stressed syllables, from a subset of 21 target words. These were selected to avoid coarticulation effects on the vowels due to adjacent liquid, nasal, or approximant consonants. Outliers that could be linked to tracking errors were removed. The final dataset contains 339 vowel tokens.

2.4 Statistical analysis

Measurements for F1 and F2 were Lobanov normalised and these values were used in the vowel plots made with phonR [16]. We use the Pillai-Bartlett Trace (Pillai score) to measure a potential merger between vowel pairs. The Pillai score is well suited to compare between speaker differences, capture the degree of overlap or distance between vowel categories, and analyse a small number of tokens [17]. Pillai score values range from 0 to 1 whereby values closer to 0 indicate more overlap and 1 means complete separation in the F1/F2 space. To further investigate F1 and F2 we used lmerTest [18] and computed two linear mixed effects models with vowel and sex as fixed factors and participant and word as random intercepts. To examine speaker-specific differences, we carried out additional models with participant as fixed factor and word as random intercept. We also carried out pairwise comparisons using emmeans [19].

3. RESULTS

Figure 1 shows ellipses of the five Central Pame vowels for female and male speakers. The plots show some degree of overlap between /i/, /e/, and /ɛ/ while the /ɐ/ and /o/ vowels are more clearly separated in the vowel space. Table 1 summarises the Pillai scores and statistical significance obtained. The results show that for all speakers the vowels /e/ and /ɛ/ are separate, although the separation is less strong in male speakers. These results are statistically significant, thus opposing the possibility of a vowel merger between /e/ and /ɛ/. Additionally, we find that /ɛ/ and /ɐ/ are also clearly separate for all speakers, which is also confirmed by the statistical significance obtained. Although the scores of three speakers (Fe3, Fe4, M2) suggest a (near) merger between /i/ and /e/, we find that the results for speakers Fe1 and M1 show a clear (i.e. statistically significant) separation between /e/ and /i/. Results of the linear mixed effects models confirm that most vowels differ on the F1 and F2 plan. We will report on relevant pairwise comparisons to focus on the vowels of interest. Figure 2 shows the normalized F1 values for five

Central Pame vowels. Plots indicate that for the vowel /i/ speakers vary on the F1 plan, and male speakers display lower values than females.

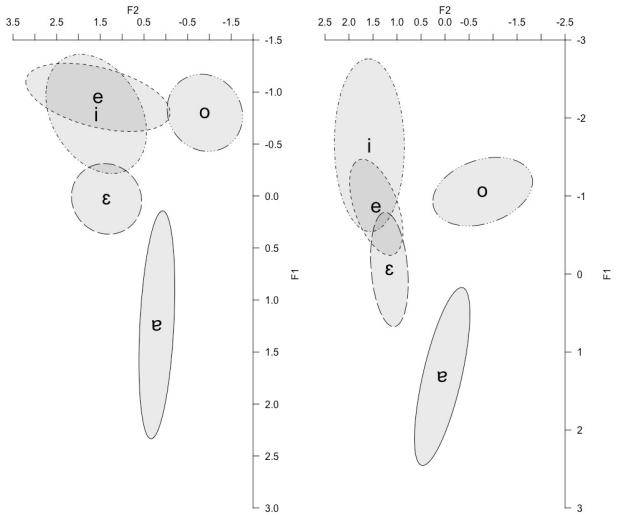


Figure 1: Lobanov normalised F1 and F2 values at midpoints of female (left) vs male (right) speakers.

Results of a model investigating F1 confirms that vowels differ. The pairwise comparisons reveal that /e/ and /ɛ/ ($\beta = -106$, $p < .009$) as well as /ɛ/ and /i/ ($\beta = -120$, $p < .003$) are separate. However, it appears that this cannot be confirmed for /i/ and /e/ ($\beta = -14$, $p < .9$ (n.s)). As Figure 1 shows, it appears that /i/, /e/, and /o/ are realised at a comparable height. Comparison between these vowels do not confirm a statistically significant difference.

	/e/ vs /ɛ/	/ɛ/ vs /ɐ/	/i/ vs /e/
Fe1	0.958 (p<.0001)	0.932 (p<.0001)	0.756 (p<.0008)
Fe2	0.968 (p<.0001)	0.743 (p<.0001)	0.315 (n.s)
Fe3	0.781 (p<.0001)	0.870 (p<.0001)	0.062 (n.s)
Fe4	0.778 (p<.0001)	0.865 (p<.0001)	0.152 (n.s)
M1	0.552 (p<.0001)	0.868 (p<.0001)	0.430 (p<.0001)
M2	0.387 (p<.0001)	0.934 (p<.0001)	0.078 (n.s)

Table 1: Pillai scores for speakers of different sex, statistical significance is given in parentheses.

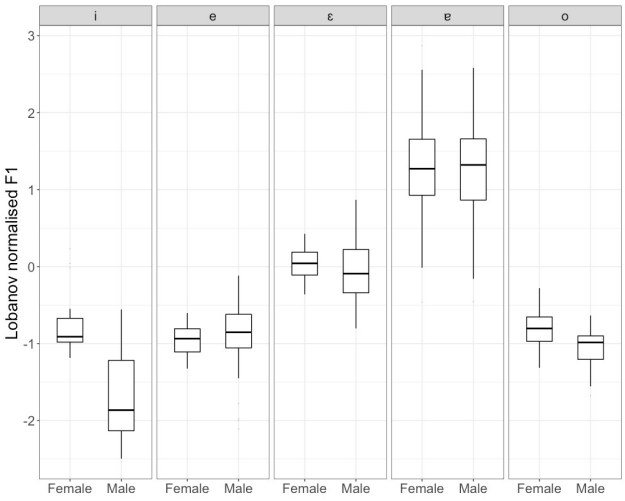


Figure 2: Boxplots of Lobanov normalised F1 values for five vowels in female (left) and male speakers (right).

Results of the model investigating F2 show a contrast between a front and a central vowel (ɛ vs a: $\beta = 493$, $p < .002$). Figure 3 shows the vowel /e/ displays a high degree of variability, especially for female speakers. Similarly, values for /o/ show high variance for female and male speakers. Our statistical investigation of F2 reveals that the front vowels /e/, /i/ and /ɛ/ do not differ significantly (i vs e: $\beta = 54$, $p < .9$ (n.s)).

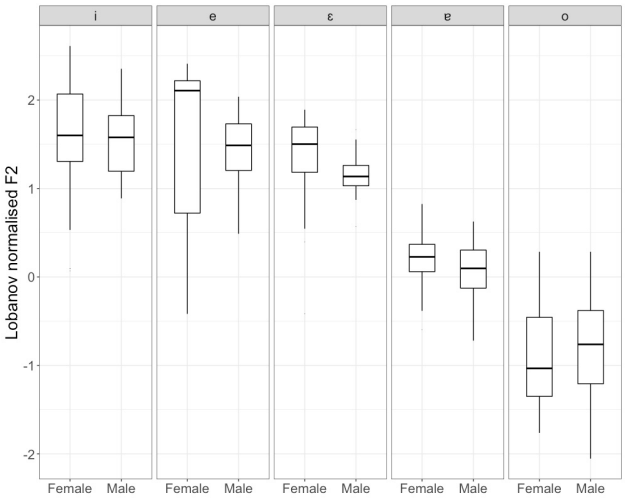


Figure 3: Boxplots of Lobanov normalised F2 values for five vowels in female (left) and male speakers (right).

Figure 4 shows the ellipses of /e/ vowel tokens for female speakers. Speakers Fe2, Fe3, and Fe4 show a high degree of fronting. Instead, speaker Fe1 realises the /e/ vowel further to the back in a less frontal position. A model that investigates speaker specific differences for F2 finds that speaker Fe1 significantly differs from all other female speakers (Fe1 vs Fe2: $\beta = -1019$, $p <$

.0001). We find that the two male speakers do not differ significantly. We performed the same model to investigate F1 and find a significant difference between the male speakers (M1 vs M2: $\beta = -73$, $p < .002$), although the difference is of smaller magnitude. The pairwise comparisons did not yield a speaker-specific significant difference for F1 in females.

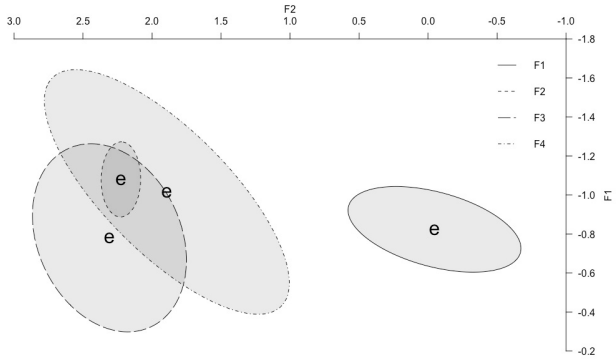


Figure 4: Lobanov normalised F1 and F2 values at midpoints of female /e/ vowels.

4. DISCUSSION

Our investigation of Central Pame confirms a five vowel inventory that can be determined acoustically. Our description of monophthongal oral vowels finds evidence for three front vowels /i, e, ɛ/, one central vowel that may be best characterised as /ɐ/, and one back vowel /o/. We find evidence that speakers vary in the realisation of front vowels, especially of the vowel /e/. In contrast to the related language Chichimec, we do not find any evidence for a vowel merger between /e/ and /ɛ/. Although male speakers seem to produce /e/ and /ɛ/ with a stronger degree of overlap, whereas female speakers clearly separate the two. Interestingly, statistical models investigating F1 and F2 find no statistically significant difference between /e/ and /i/. For the vowel /e/ we find speaker-specific differences in F2 in the female speakers and for F1 for the male speakers, showing that this vowel may be realised variably. Thus, we find that the older speakers (Fe1 31 years; M1 29 years) are different from younger speakers (18-23 years). We find increased F2 values confirming that young female speakers show a stronger degree of fronting of /e/. We speculate that this realisation could be related to some (to us unknown) sociolinguistic factors or possibly to age, since older speakers appear to display different patterns from younger speakers. The data suggests a chain shift might be taking place in the language's front vowels. Similar to the Great Vowel Shift in 15th c. in English [20], front vowels appear to have become more closed, and may diphthongise when they cannot become more closed (i.e. /ɛ/>/e/, /e/>/ɪ/, /i/>/ei/) although this remains to be explored in future research with older speakers. Regarding F1 the vowel /o/ appears to be at a similar height as /e/ and /i/. However, some degree of variability is visible in the F2 values which could be indicative of a more central /ʊ/, possibly /ɵ/ or back realisation /o/. The extent to which the /o/ vowel is influenced by lexical or sociolinguistic factors goes beyond the scope of this study and remains to be explored in future work.

5. CONCLUSION

This study presents the first acoustic phonetic analysis of the Central Pame vowel system. We find evidence for a five vowel system /e, i, ɛ, ɐ, o/. Our data suggest a sound change in progress by which the front vowel /e/ is being fronted and raised. Possibly, this shift could be affecting the phonetic specification of other vowels in the inventory (/i/ and /ɛ/) which will be examined in greater detail in future work. With respect to (the) back vowel(s), although the absence of native minimal pairs suggests that the language has a single back vowel phoneme, the values for F2 are quite variable. The integration of Spanish loan words into the Pame's phonemic and phonetic system may be exerting some influence. If the original Spanish vowel distinction (i.e. /u/ vs /o/) are preserved in an increasing number of loan words (e.g. tambòl 'drum', Sp. *tambor*, krús 'cross', Sp. *cruz*), this could put additional pressure into a phonemic split in the vocabulary.

Future research will include more data from more (especially older) speakers and further examine the role of age, sex, language use, and sociolinguistic factors to account for variability in the vowel space.

7. REFERENCES

[1] Gibson, Lorna F. 1950a. Three Chichimeca texts. Unpublished manuscript. Online at https://www.sil.org/resources/archives/57422.

[2] Gibson, Lorna F. 1950b. A pedagogical grammar of Central Pame. Unpublished manuscript. Online at https://www.sil.org/resources/archives/57479.

[3] Gibson, Lorna F. 1950c. Verb paradigms in Pame. Unpublished manuscript. Online at https://www.sil.org/resources/archives/53023.

[4] Olson, Donald. 1955. Mode-Aspect-Person Inflection in Pame. (Unpublished manuscript, online at http://www.language-archives.org/item/oai:sil.org:57423)

[5] Gibson, Lorna F. 1956. Pame (Otomi) phonemics and morphophonemics. *International Journal of American Linguistics* 22, 4: 242-265.

[6] Avelino, Heriberto. 2019. The ecology of variation and change in the context of language attrition: ʔuzãʔ (Chichimec) vowels. Language Ecology 3 (1), 28–57.

[7] Berthiaume, Scott Charles. 2013. *A phonological grammar of Northern Pame*. PhD Thesis, The University of Texas at Arlington.

[8] Avelino, Heriberto. 1997. Fonología y morfofonología del Pame Norte. BA Thesis. México DF: Escuela Nacional de Antropología e Historia.

[9] Avelino, Heriberto. 1996. El sistema consonántico del Pame Central. Tercer Encuentro de Lingüística en el Noroeste. Vol I. Universidad de Sonora.

[10] Hurch, Bernhard. 2022. Pame (central) de Santa María Acapulco, Santa Catarina, San Luis Potosí. In Yolanda Lastra (Ed.) Archivo de Lenguas Indígenas de México. México DF: El Colegio de México.

[11] T. Kisler, U. Reichel, and F. Schiel. 2017. "Multilingual processing of speech via web services," Computer Speech & Language, vol. 45, pp. 326–347.

[12] P. Boersma and D. Weenink. 2021. "Praat: doing phonetics by computer (version 6.1.52)[computer program]".

[13] R. Winkelmann, J. Harrington, and K. Jänsch. 2017. "EMU-SDMS: Advanced speech database management and analysis in R," Computer Speech & Language, vol. 45, pp. 392–410.

[14] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington. 2017. emuR: Main Package of the EMU Speech Database Management System, R package version 0.2.3.

[15] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[16] McCloy, Daniel R. 2016. "phonR: Tools for phoneticians and phonologists." R package version 1.0-7. http://drammock.github.io/phonR/

[17] Adank, P., Smits, R. and van Hout, R. 2004. "A comparison of vowel normalization procedures for language variation research", *The Journal of the Acoustical Society of America*, 116(5): 3099–3107.

[18] Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software*, 82(13), 1-26. https://doi.org/10.18637/jss.v082.i13.

[19] Lenth, R .2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.4-1, https://CRAN.R-project.org/package=emmeans.

[20] Wolfe, Patricia M. 1972. *Linguistic change and the great vowel shift in English*. Univ of California Press.

Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production

Nicholas Aoki and Georgia Zellou

University of California, Davis
nbaoki@ucdavis.edu, gzellou@ucdavis.edu

ABSTRACT

Extensive work in speech perception indicates that given the same speech signal, listeners behave differently when viewing East Asian faces compared to Caucasian faces. An untested question is whether visual guise also affects *speech production*. If speakers assume that an Asian face depicts a non-native English speaker, we predict that speech towards an Asian face should be hyper-articulated compared to speech towards a Caucasian face and should acoustically resemble speech towards an imagined non-native listener. Moreover, individual differences in ratings of how likely the faces depict non-native speakers should predict variation. Results reveal that: (1) speech towards an East Asian face is hyper-articulated compared to a Caucasian face through a slower speaking rate; (2) non-native-directed speech is even more hyper-articulated than speech towards an East Asian face; (3) ratings do not predict differences between visual conditions. This study has implications for the relationship between speech perception, production, and social expectations.

Keywords: Speaking style, Visual guise, Non-native-directed speech, Perceived ethnicity, Sociolinguistics

1. INTRODUCTION

Extensive research has shown that the perceived ethnicity of visual guises can impact speech perception by triggering sociolinguistic expectations about speakers [1, 2, 3, 4]. For example, Rubin [5] presented listeners with the same recording of a lecture produced by a native speaker of American English and found that listeners who see a picture of an East Asian woman show lower comprehension (i.e., lower cloze test accuracy) of the lecture material and rate the voice as more accented compared to listeners who see a picture of a Caucasian woman. This phenomenon has been explained as a type of linguistic stereotyping – visual cues to a speaker's group membership can generate expectations about their speech as being more accented due to listener biases (e.g., that East Asians are non-native English speakers) [5, 6, 7, 8]. Other work has found that an East Asian guise can enhance comprehension. McGowan [9] showed that for Mandarin-accented speech, transcription accuracy in a speech-perception-in-noise task is higher when a picture of an East Asian woman is presented to listeners compared to a Caucasian woman or a silhouette. The greater congruency of the East Asian face with accented speech thus increases comprehension compared to the less compatible Caucasian face [9, 10]. In general, these prior studies in speech perception are consistent with the assumption that seeing an East Asian face may trigger a "forever foreigner" stereotype [11] and that the sociolinguistic expectations of East Asian and Caucasian faces can induce distinct perceptual responses.

An untested question is whether the perceived ethnicity of visual guises also impacts *speech production*. Prior work has shown that speakers tend to produce more effortful speech (e.g., through a slower speaking rate) when communicating with listeners who may have difficulty understanding them, such as non-native speakers or individuals who are hard-of-hearing [12]. These acoustic-phonetic modifications, also known as "clear speech", benefit listeners by increasing intelligibility compared to more "casual speech" [13]. Clear and casual speech are often elicited through explicit instructions to speakers (e.g., "Speak clearly to someone who may have trouble understanding you"; "Say the sentences in a natural, casual manner" [14]). However, it is not yet known whether the presentation of different visual guises (without written instructions) will affect speech production.

The current study tests three hypotheses. If seeing an Asian face triggers expectations that the listener is a non-native speaker with difficulty understanding English, then: (1) speakers should hyper-articulate (i.e., produce greater intensity, higher pitch, slower speaking rate, and greater pitch range) when looking at an Asian face compared to a Caucasian face; (2) speech patterns in the Asian and Caucasian face conditions should be similar to non-native-directed speech and casual speech, respectively (where the latter two styles are elicited via explicit instructions); (3) individuals who rate the Asian face as being more likely to be a non-native speaker relative to the Caucasian face should hyper-articulate more when looking at the Asian face.

2. METHODS

2.1. Participants

48 native English speakers (35 female, 13 male, 0 non-binary; mean age = 19.17; sd = 1.71) were recruited from the University of California, Davis (UC Davis) Psychology subjects pool and received course credit. One participant was removed for lack of attentiveness during the task.

2.2. Visual Guises

Figure 1 shows the images of the Asian and Caucasian faces. The pictures, which were selected from the Chicago Face Database, were rated as being females in their mid-20s and as conforming to their self-reported ethnicity in a norming study [15].



Figure 1: The Asian face (left) and the Caucasian face (right) used in this study.

2.3. Procedure

Participants completed two recording sessions on separate days between one and three days apart. In one session, there were two blocks – one presented the Caucasian face, and another presented the Asian face. The order of the blocks was counterbalanced. On each trial, speakers were shown a face and a sentence and asked to "produce the sentence to the listener in the image". After both blocks, participants rated the faces on how much they looked like non-native English speakers from 1 ("Strongly Disagree") to 7 ("Strongly Agree").

In another session, there were three blocks: (1) non-native-directed speech; (2) casual speech; (3) hard-of-hearing-directed speech (not analyzed here). To elicit non-native-directed speech, speakers were asked to imagine "talking to a listener who is a native speaker of Mandarin and is learning English". For casual speech, participants were asked to talk "casually" as if "to a listener who is a native speaker of English". The non-native-directed speech block always preceded the casual speech block. Neither block presented a visual image.

For all blocks, speakers produced the same 80 semantically unpredictable sentences (e.g., "Tom discussed the hay") from the Speech Perception in Noise Test [16]. Session order was counterbalanced. Productions were recorded in a sound-attenuated booth using a Shure WH20 XLR head-mounted microphone and digitally sampled at a 44-kHz rate.

2.4. Statistical Analysis

Intensity, pitch, speaking rate, and pitch range were measured over each sentence using Praat [17]. Speech rate was calculated by dividing the number of syllables by the sentence duration (in seconds). Each variable was modelled via separate linear mixed-effects regression models with the *lme4* R package [18]. All models had a treatment-coded fixed effect of Block (4 levels: Asian face, Caucasian face, Non-native-directed, Casual [reference level]) with by-listener and by-sentence random intercepts and by-listener random slopes for Block. The marginal and conditional coefficients of determination were found with the *r.squaredGLMM* function [19].

To test whether the ratings of the faces affected individual differences between the visual guise conditions, a correlational analysis was conducted. For each participant, the difference in ratings between the Asian and Caucasian faces was calculated, where a more positive difference indicates a greater perception that the Asian face corresponds to a non-native English speaker. For each participant, the difference between the mean value in the visual conditions was calculated for each acoustic variable. Pearson product-moment correlations were computed between the difference in ratings and the difference in the acoustic variables (e.g., testing if those who rated the Asian face as being more likely to be a non-native speaker also spoke with greater average intensity in the Asian face condition).

3. RESULTS

3.1. Intensity

Neither non-native-directed speech (*Coef* = 0.59, *SE* = 0.37, z = 1.62, p = 0.11), nor the Asian face condition (*Coef* = 0.11, *SE* = 0.57, z = 0.20, p = 0.84), nor the Caucasian face condition (*Coef* = 0.40, *SE* = 0.54, z = 0.74, p = 0.47) are significantly different in mean intensity from casual speech. The marginal and conditional coefficients of determination were 0.001 and 0.88, respectively.

3.2. Pitch

Neither non-native-directed speech (*Coef* = 1.86, *SE* = 1.10, z = 1.70, p = 0.10), nor the Asian face condition (*Coef* = -1.72, *SE* = 0.98, z = -1.77, p = 0.08), nor the Caucasian face condition (*Coef* = -1.44,

SE = 2.97, *z* = -1.14, *p* = 0.26) are significantly different in mean pitch from casual speech. The marginal and conditional coefficients of determination were 0.004 and 0.81, respectively.

3.3. Pitch Range

Mean pitch range across conditions is provided in Figure 2. Non-native directed speech contains a significantly larger pitch range than casual speech (*Coef* = 8.12, *SE* = 1.32, *z* = 6.14, *p* < 0.001). There is no difference in pitch range between the Asian (*Coef* = 2.43, *SE* = 1.70, *z* = 1.43, *p* = 0.16) and Caucasian face conditions (*Coef* = 0.65, *SE* = 1.76, *z* = 0.37, *p* = 0.71) from casual speech. The marginal and conditional coefficients of determination were 0.008 and 0.65, respectively.

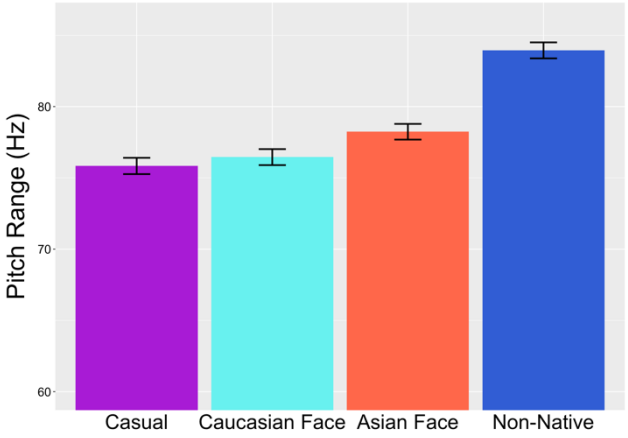


Figure 2: Mean pitch range (Hz) by condition. Error bars represent standard errors.

3.4. Speaking Rate

Figure 3 shows mean speaking rate by condition. Non-native-directed speech (*Coef* = -1.48, *SE* = 0.1, *z* = -14.75, *p* < 0.001), the Asian (*Coef* = -0.45, *SE* = 0.09, *z* = -4.81, *p* < 0.001), and Caucasian face conditions (*Coef* = -0.26, *SE* = 0.06, *z* = -4.14, *p* < 0.001) are all slower than casual speech. The marginal and conditional coefficients of determination were 0.30 and 0.85, respectively. A model with the same random effects structure that excluded casual speech and coded non-native-directed speech as the baseline level indicated that non-native directed speech is even slower than the Asian (*Coef* = 1.03, *SE* = 0.10, *z* = 10.54, *p* < 0.001) and Caucasian face conditions (*Coef* = 1.23, *SE* = 0.11, *z* = 11.35, *p* < 0.001). A final post-hoc model that only included the two visual guise conditions revealed that speaking rate in the Asian face condition is slower than in the Caucasian face condition (*Coef* = -0.20, *SE* = 0.08, *z* = -2.55, *p* = 0.01). The speaking rate results can be summarized as follows, from fastest to slowest: Casual > Caucasian Face > Asian Face > Non-Native-Directed.

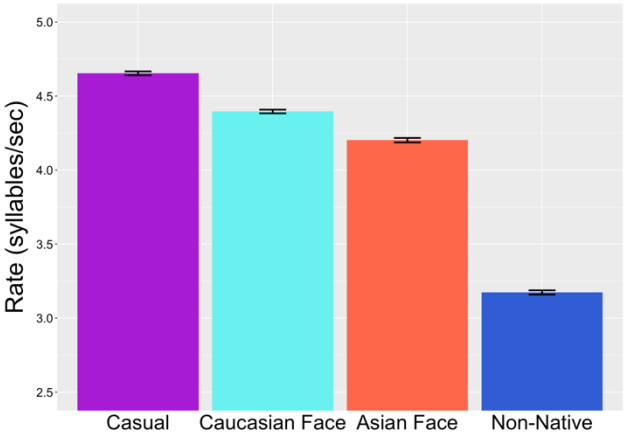


Figure 3: Mean speaking rate (syllables per second) by condition. Error bars represent standard errors.

3.5. Non-Native Speaker Ratings

The difference in ratings between the Asian and Caucasian faces did not correlate significantly with differences between the visual conditions for any acoustic variable (Intensity: *r* = -0.01, *p* = 0.97; Pitch: *r* = -0.22, *p* = 0.14; Pitch Range: *r* = -0.01, *p* = 0.95; Rate: *r* = -0.28, *p* = 0.053).

4. DISCUSSION

The current study supports the hypothesis that speakers hyper-articulate more when looking at a picture of an East Asian face compared to a picture of a Caucasian face. More specifically, speakers talk more slowly towards an East Asian face than a Caucasian face. However, speech towards an imagined non-native listener, which is hyper-articulated compared to casual speech through greater pitch range and slower speaking rate, is even slower than the Asian face condition. We also find that explicit ratings of how much the faces correspond to non-native English speakers do not correlate with differences in production between the visual conditions.

Overall, the slower speaking rate in the East Asian face condition relative to the Caucasian face condition is in line with a considerable body of work in speech perception on effects of visual guise [5, 6, 7, 8, 9, 10]. Visual cues, such as apparent ethnicity, can trigger expectations about group membership based on linguistic stereotypes. The assumption that East Asians are non-native English speakers leads to downstream effects in both speech perception and production. Similar to how listeners demonstrate better understanding of Mandarin-accented speech with a congruent East Asian face [9], speakers talk

more slowly towards an East Asian face to aid listener comprehension. The current results also align with prior work in syntax and pragmatics indicating that listeners accommodate non-native speakers to facilitate communication (e.g., through greater leniency for syntactic errors [20] and for pragmatically odd or under-informative utterances [21, 22]).

Although speakers talk more effortfully towards an East Asian face than a Caucasian face, even greater hyper-articulation is observed in non-native-directed speech on pitch range and speaking rate. One way to account for the larger effect size in the non-native-directed speech condition is by appealing to Hypo- and Hyper-articulation (H&H) Theory [23] and the concept of *informativity* from Bayesian frameworks of speech perception [24, 25]. According to H&H Theory, articulatory patterns reflect competing goals between conserving effort and maximizing clarity. Whether speakers choose to produce more hyper-articulated speech, at the expense of articulatory effort, depends on how likely they think the listener will understand them. This likelihood is partially determined based on the informativity of socio-indexical cues. For example, in the non-native-directed speech condition in the current study, speakers are given an explicit description of the imagined listener as a "native speaker of Mandarin" who is "learning English". This written guise provides high certainty about the language background of the listener and is thus an informative cue. Given prior experience interacting with non-native listeners, speakers may talk with a highly effortful style to maximize clarity and reduce potential comprehension difficulties. In contrast to a written guise, visual cues to apparent ethnicity are less informative. Although ethnicity and language background can be correlated [26], they are not in a one-to-one relationship [27]. In fact, the majority of Asian Americans in the United States are actually proficient in English [28]. Visual guises thus offer less information about the language background of the listener, leading to greater uncertainty about whether the listener is a native or a non-native speaker, and therefore, whether hyperarticulation is necessary to facilitate communication. While linguistic stereotypes and biases may lead to a slower speaking rate towards East Asian faces than Caucasian faces, uncertainty about listener identity means that speakers are less willing to expend articulatory effort, thereby resulting in a smaller effect size compared to the non-native-directed condition.

Although speech towards an East Asian face is hyper-articulated compared to a Caucasian face, both visual conditions are significantly slower compared to casual speech towards an imagined listener who is a native English speaker. One possible explanation is that, relative to speech directed towards an imagined listener, visual guises simulate a more ecologically valid interaction with a real listener, thus resulting in a distinct speaking style. This is consistent with prior work showing acoustic differences between speech towards a real listener and speech towards an imagined listener [29, 30, 31]. Future work should more explicitly test how real-listener-directed speech compares acoustically to guise-directed speech.

Another direction for future research is to examine the role of implicit biases on speech towards visual guises. In the current study, speakers were asked to provide explicit ratings for the East Asian and Caucasian faces on how likely they corresponded to a native or non-native speaker, and no correlations were found between the ratings and the acoustic results. Prior studies, however, have shown that explicit judgments and implicit beliefs do not necessarily overlap [32]. More work is needed to see whether more implicit measures, such as the Implicit Association Test, can better account for individual differences in production towards visual guises.

Speaking style could also differ between other pairs of guises, such as a picture of a digital device and a human. Recent studies have proposed that, similar to East Asian individuals, people may be biased towards thinking that voice-activated artificially intelligent assistants have difficulty understanding them [33, 34, 35]. Consistent with these accounts, Aoki, Cohn, and Zellou [36] showed that presenting a picture of a cylindrical device decreases transcription accuracy in a speech-perception-in-noise task compared to a picture of a human for both a naturally produced voice and a text-to-speech voice. Bias against digital devices may elicit a more hyper-articulated speaking style towards a picture of a device compared to a human.

5. CONCLUSION

Although much past work has examined how the perceived ethnicity of visual guises affects speech perception, little work has explored effects on speech production. The current study begins to fill this gap, showing that speakers hyper-articulate towards East Asian faces compared to Caucasian faces through a slower speaking rate. These results lead to further questions about the nexus between speech perception, speech production, and social expectations.

6. REFERENCES

[1] Zheng, Y., Samuel, A. G. 2017. Does seeing an Asian face make speech sound more accented? *Atten. Percept. Psychophys.* 79, 1841-1859.

[2] Hanulíková, A. 2018. The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguist. Vanguard* 4, 1-9.
[3] Hanulíková, A. 2021. Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PloS One*, 16, e0259230.
[4] Melguy, Y. V., Johnson, K. 2021. General adaptation to accented English: Speech intelligibility unaffected by perceived source of non-native accent. *J. Acoust. Soc. Am.* 149, 2602-2614.
[5] Rubin, D. L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Res. High. Educ.* 33, 511-531.
[6] Kang, O., Rubin, D. L. 2009. Reverse Linguistic Stereotyping: Measuring the Effect of Listener Expectations on Speech Evaluation. *J. Lang. Soc. Psychol.* 28, 441-456.
[7] Yi, H.-G., Phelps, J. E. B., Smiljanić, R., Chandrasekaran, B. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *J. Acoust. Soc. Am.* 134, EL387-EL393.
[8] Hu, G., Su, J. 2015. The effect of native/non-native information on non-native listeners' comprehension. *Lang. Aware.* 24, 273-281.
[9] McGowan, K. B. 2015. Social Expectation Improves Speech Perception in Noise. *Lang. Speech.* 58, 502-521.
[10] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137, 2823-2833.
[11] Tuan, M. 1998. *Forever foreigners or honorary whites?: the Asian ethnic experience today.* Rutgers University Press.
[12] Uchanski, R. M. 2005. Clear speech. In: Pisoni, D. B., Remez, R. E. (eds), *The Handbook of Speech Perception.* Blackwell, 207-235.
[13] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Lang. Linguist. Compass.* 3, 236-264.
[14] Cohn, M., Pycha, A., Zellou, G. 2021. Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition* 210, 104570.
[15] Ma, D. S., Correll, J., Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* 47, 1122-1135.
[16] Kalikow, D. N., Stevens, K. N., Elliott, L. L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337-1351.
[17] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341-345.
[18] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1-48.
[19] Nakagawa, S., Johnson, P. C. D., Schielzeth, H. 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* 14, 20170213.
[20] Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., Weber, A. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *J. Cogn. Neurosci.* 24, 878-887.
[21] Gibson, E., Tan, C., Futrell, R., Mahowald, R., Konieczny, L., Hemforth, B., Fedorenko, E. 2017. Don't underestimate the benefits of being misunderstood. *Psychol. Sci.* 28, 703-712.
[22] Fairchild, S., Papafragou, A. 2018. Sins of omission are more likely to be forgiven in non-native speakers. *Cognition* 181, 80-92.
[23] Lindblom, B. 1990. Explaning phonetic variation: A sketch of the H&H theory. In: Hardcastle, W. J., Marchal, A. (eds), *Speech production and speech modelling.* Springer, 403-439.
[24] Kleinschmidt, D. F., Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148-203.
[25] Kleinschmidt, D. F. 2019. Structure in talker variability: How much is there and how much can it help? *Lang. Cogn. Neurosci.* 34, 43-68
[26] Wolfram., W., Schilling, N. 2015. *American English: Dialects and Variation.* John Wiley & Sons.
[27] King, S. 2020. From African American vernacular English to African American language: Rethinking the study of race and language in African American's Speech. *Annu. Rev. Linguist.* 6, 285-300.
[28] Budiman, A., Ruiz, N. G. 2021. Key facts about Asian Americans, a diverse and growing population. Pew Research Center. https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-americans/ (accessed Apr. 24, 2022).
[29] Scarborough, R., Bernier, J., Zhao, Y., Hall-Lew, L., Dmitrieva, O. 2007. An Acoustic Study of Real and Imagined Foreigner-Directed Speech. *Proc. 16th ICPhS* Saarbrücken, 2165-2168.
[30] Scarborough, R., Zellou, G. 2013. Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J. Acoust. Soc. Am.* 134, 3793-3807.
[31] Scarborough, R., Zellou, G. 2022. Out of sight, out of mind: The influence of communicative load and phonological neighborhood density on phonetic variation in real listener-directed speech. *J. Acoust. Soc. Am.* 151, 577-586.
[32] Devos, T., Banaji, M. R. 2005. American = white? *J. Pers. Soc. Psychol.* 88, 447-466.
[33] Cohn, M., Liang, K. H., Sarian, M., Zellou, G., Zhou, Y. 2021. Speech Rate Adjustments in Conversations With an Amazon Alexa Socialbot. *Front. Commun.* 6, 671429.
[34] Cohn, M., Zellou, G. 2021. Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. *Front. Commun.* 6, 675704.
[35] Cohn, M., Segedin, B. F., Zellou, G. 2022. Acoustic-phonetic properties of Siri- and human-directed speech. *J. Phon.* 90, 101123.
[36] Aoki, N. B., Cohn, M., Zellou, G. 2022. The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters* 2, 045204.

ACOUSTIC MEASURES OF NON-NATIVE ADDRESSEE REGISTER FOR MID TO HIGH PROFICIENT ENGLISH LEARNERS OF GERMAN

Bianca Sell, Megumi Terada, Malte Belz, Christine Mooshammer

Humboldt-Universität zu Berlin, Germany
bianca.maria.sell@hu-berlin.de

ABSTRACT

Up to now, most studies investigating speech adaption towards non-native addressees focused on L2 speakers with a low proficiency of English. Here, we examine task-based conversations when speaking to mid to high proficient L2 speakers of German. We compare articulation rate, vowel duration, and vowel space area of 20 German L1 speakers when conversing with L1 German speakers and with L2 German speakers of L1 English. As expected, we find no effect of addressee on vowel duration, while articulation rate is slower for non-native addressees, regardless of proficiency, compared to native addressees. However, vowel space area does not change for different addressees, contrary to enlargement reported for English non-native addressee register. The results will be discussed with respect to L2 proficiency and phonetic convergence.

Keywords: Foreigner-directed speech, German, hyperarticulation, register

1. INTRODUCTION

Speakers adapt their linguistic behaviour to the specific demands of the communicative situation – they employ different registers [1]. Besides situational variables such as setting or topic, the addressee's characteristics are another factor influencing register choices. Previous studies [2, 3, 4] show that speakers accommodate their linguistic behaviour to the (presumed) traits of the addressee. One of these traits is the addressee's linguistic background: Speakers shift into another register when communicating with a non-native interlocutor, which is called *non-native addressee register* (NNAR) or *foreigner directed speech* (FDS) [5]. On the phonetic level, NNAR features lower articulation rates, louder speech, and hyperarticulations [6, 7, 5]. The latter has often been accessed by the vowel space area (VSA), i.e. the area spanned by the first and second format of the corner vowels /iː, aː, uː/, and was mainly found for English NNAR, probably as a means to extend the perceptual distance of vowel categories in order to avoid confusion between them [5]. Vowel space extension was not consistently found in all languages investigated (e.g., not for Arabic NNAR, see [8]) and might be dependent on the size of the vowel inventory [5]. Results on VSA extension remain inconclusive and mostly focussed on the English language. As has been found by [9, 10], hyperarticulation in vowels is not related to vowel length. A much more consistent adaptation towards non-native speakers is a lower articulation rate, cf. [9, 7, 11]. While [9] report that adaptations in speech rate are mostly a result of longer pauses instead of a slower articulation rate, [11] find a longer mean word duration for English NNAR.

For German, research on NNAR has been focused mostly on qualitative parameters adapted towards low proficiency addressees [12, 6]. It is, however, difficult to differentiate in these studies between the addressee's low-level proficiency, the lower prestige of the addressee's L1, and possible power imbalances that might create a negative bias influencing linguistic behaviour [12]. It remains largely unknown what the features of German NNAR are when elicited by mid to high proficiency addressees whose L1 is (ideally) not subject to a negative bias. In the present study, we address this issue: We investigate whether German L1 speakers adapt their vowel space area (VSA), vowel duration (VD), and articulation rate (AR) when talking to non-native addressees with mid to high proficiency levels in German, with English as their L1 (considered as presitigous in Germany [13]), when compared to the same setting, but with German L1 interlocutors. We additionally assess the influence of the addressee's proficiency (self-reported mid vs. high proficiency).

For this purpose, we conduct a corpus-based study comparing participants' VSA, VD, and AR when talking to German native addressees or non-native addresses with mid to high proficiency. We expect to find an increased VSA when speakers are conversing with non-native interlocutors with a stronger effect for interlocutors with a lower proficiency (Hyp. 1). Likewise, it is predicted that the participants decrease their AR when speaking with non-native interlocutors, with a stronger effect for lower L2 proficiency (Hyp. 2), as shown in previous

studies [5]. We do not expect any adaptions concerning vowel duration (Hyp. 3), as indicated by [9, 10]. The present study thus allows us to gain a better understanding of the phonetic level of NNAR in German for mid-to-high proficiency interlocutors.

2. METHOD

2.1. Corpus and annotation

In the Corpus of Non-Native Addressee Register (CoNNAR) we compare the data of 20 German L1 speakers (10 female; age 20–38 years, mean age = 26 years, sd = 4.5 years) who passed through our experiment twice with instructed interlocutors (confederates): Once with another German L1 speaker (L1 confederate, n = 4) and once with an English L1 speaker (L2 confederate, n = 4) with mid to high proficiency in German (self reported proficiency levels of B1 to C1 CEFR [14]). Two of the L2 confederates originate from the US, two from Southern England. Both varieties of English can be considered similarly familiar and prestigious in German [15]. L1 and L2 confederates were matched in age and gender, creating confederate pairs differing in the non-natives' German proficiency. Each confederate conversed with five experimental participants. The order of L1 and L2 confederate experiments was counterbalanced.

Each experimental session (n = 40) consisted of a word list with the sentence *Sage X bitte* 'Say X please' with X being a bisyllabic word containing all monophthongs of German in stressed position. The word list was read in a break-out room with only the participant and the experimenter. Next, participants conversed for 8 minutes about a topic of their choice and then solved two Diapix tasks, 8 minutes each (a spot-the-difference task where both interlocutors are given slightly different versions of the same picture with the goal of finding all differences, [16, 17]). The original Diapix picture materials by Baker & Hazan [17] were translated into German and adapted to elicit as many corner vowels without being overly obvious [18]. In the second Diapix confederates were instructed to ask for a clearer pronunciation after 3 to 4 found differences. Free conversation and the Diapix tasks were held in break-out rooms without the experimenter. Lastly, participants read the word list again in a break-out room and filled out a questionnaire after the second session.

Due to the COVID-19 pandemic the interlocutors were placed in different rooms and connected via Zoom. The data was recorded using directional microphones (Sennheiser) via Audacity [19] in a stereo file. The recordings lasted about one hour per session, participants were paid 11€/h. The collected data was transliterated using Praat [20], the segmentation of sounds was carried out automatically using the WebMAUS service [21]. We measured the frequencies of the first and second formant and vowel durations for the target vowels [aː], [iː], and [uː] in stressed and accented position within content words. In the Diapix task, we focused on target words in the pictures' written parts. If participants produced less than 2 target vowels per Diapix from reading those written parts, additional target vowels in stressed and accented position were annotated manually, avoiding diphthongisation due to r-vocalisation or following vowels. In case there were not enough target words following these conditions syllables without word stress were also accepted (7 cases). All target vowels were annotated manually. The beginning of each vowel was marked at the rising zero crossing of the first complete period, its end at the zero crossing of F2 fading.

2.2. Data preparation and measurements

The annotated word lists and Diapix tasks were converted into an EMU database [22] with the R package *emuR* version 2.3.0 [23]. The Praat formant tracker was used to calculate formants [24], using gender-specific formant ceilings (5000 Hz for male, 5500 Hz for female speakers). All formant trajectories were checked and manually corrected for obvious tracking errors in the EMU database.

For measuring the three corner vowels /iː, aː, uː/, five measure points in the center of the vowel were extracted for F1 and F2 and normalised using Lobanov's normalisation method [25]. Based on the normalised F1 and F2, we calculated the VSA using the following formula for triangle area in a Euclidean plane: $VSA = |\frac{1}{2} * ((F1_i * (F2_a - F2_u) + F1_a * (F2_u - F2_i) + F1_u * (F2_i - F2_a))|$

For articulation rate the syllable number is automatically detected based on the orthographic transcription by using the R-package *sylly.de* version 0.1.2 [26]. The total number of syllables per speaker in each Diapix task was summed and divided by total length of articulation (in sec.) excluding silent pauses, turn transitions, and extra- and paralinguistic events such as laughing, clicks and background noises as well as pseudonymised tokens.

3. RESULTS

According to Hypothesis 1, we expected an increase in VSA for non-native addressees compared to L1 addressees with a stronger effect for mid proficient learners (B1-B2) than high proficient learners (C1).

As can be seen in Figure 1, our data does not confirm this hypothesis. Even though participants show a larger VSA when speaking to non-natives with B2 proficiency than to the L1 confederate (cf. L1B2 in Figure 1), the results did not show any significant effects of addressee or their L2 proficiency on the VSA, as evaluated by a linear mixed-effects model with normalised VSA as dependent variable and confederate pair (L1B1, L1B2, L1C1_1, L1C1_2) as well as language (L1 vs. L2) as independent variable, including random intercepts for participants.

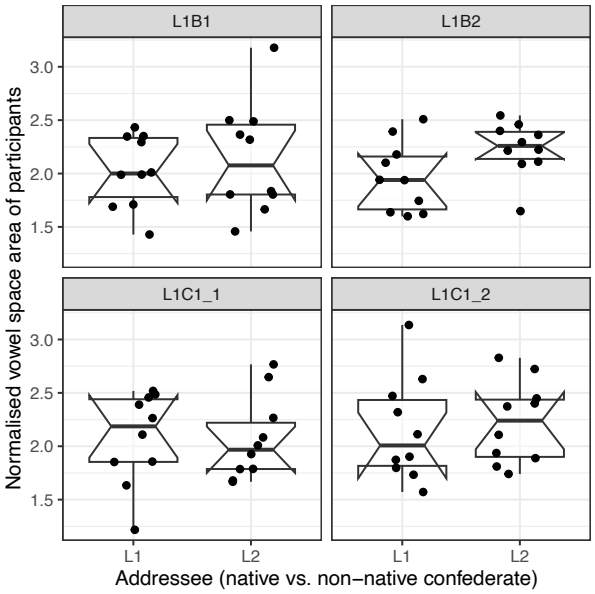


Figure 1: Normalised vowel space area of the L1 participants per confederate pair (L1B1, L1B2, L1C1_1, L1C1_2) and addressee (L1 vs. L2) for each of the two Diapix tasks.

In Hyp. 2 we expected to find a slower AR for non-native addressees compared to L1 addressees with a stronger effect for mid than high proficient learners. Figure 2 shows the AR per speaker sorted by confederate pairs. It can be seen that participants tend to decrease their AR with non-native interlocutors with mid level L2 proficiency (B1, B2) compared to L1 interlocutors. These speakers tend to behave similarly though individual speakers deviate from the described pattern. In contrast, participants maintain or increase their AR when conversing with advanced L2 speakers (C1_1, C1_2). We calculated a linear mixed-effects model with AR as dependent variable and confederate pair (L1B1, L1B2, L1C1_1, L1C1_2) as well as language (L1 vs. L2) as independent variable, including random intercepts for participants. Participants speak slower when talking to an L2 confederate than to an L1 confederate ($\beta = .1, se = .05, p < .05$). There was no significant effect for confederate pair.

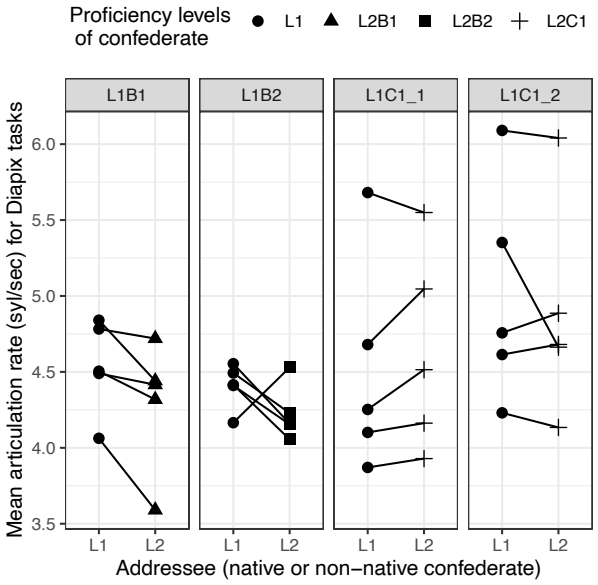


Figure 2: Mean articulation rate of L1 participants per confederate pair (L1B1, L1B2, L1C1_1, L1C1_2) and addressee (L1 vs. L2) for all Diapix tasks

Figure 3 shows a scatterplot with regression lines of articulation rates and VSA per participant sorted by the interlocutors' proficiency (L2B1-2 vs. L2C1 vs. L1). The correlations of normalised VSA and AR for every proficiency level did not reach significance, as calculated by a Pearson's correlation test.

Figure 4 shows boxplots for vowel duration per target vowel and proficiency. Speakers do not change vowel duration depending on the addressee. For none of the target vowels /iː aː uː/ the differences in duration based on the confederates' L1 or L2 proficiency reached significance, based on a linear mixed-effects model with duration as dependent variable and vowels as well as confederate pairs as independent variables, with participants as random intercepts. The null result of vowel duration confirm Hypothesis 3.

4. DISCUSSION

We investigated whether speakers change their articulation rate, vowel space area, and vowel duration when conversing with a non-native addressee with mid to high German L2 proficiency compared to German L1 addressees. In summary, VSA and vowel duration show no effect of the addressee while articulation rate is slower for non-native addressees compared to L1 addressees independent of

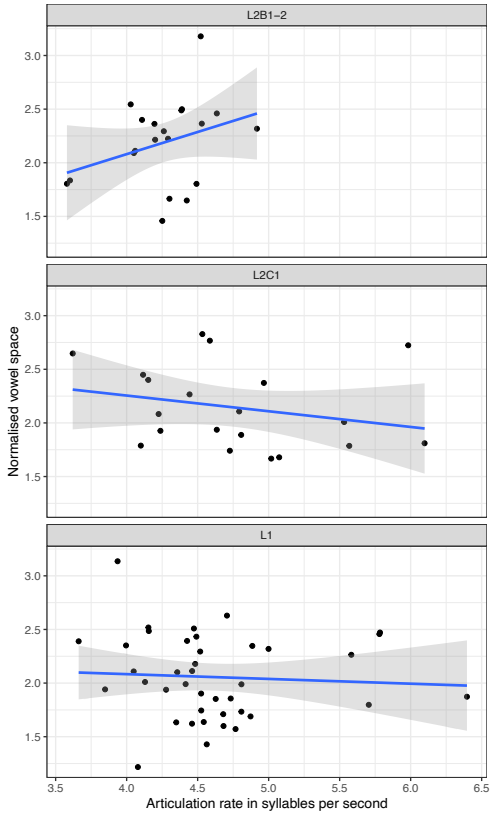


Figure 3: Vowel space area and articulation rate in both Diapix tasks per participant speaking to L1 or L2 confederates with varying proficiency levels.

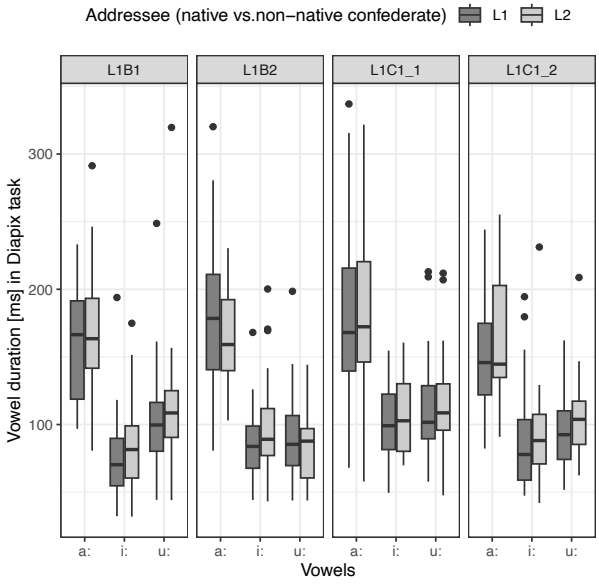


Figure 4: Duration of target vowels in Diapix tasks per confederate pair and the non-native's L2 proficiency.

the non-native addressees' proficiency. [27] report that hyperarticulation, measured as VSA extension, might be language-specific: A larger vowel inventory might lead to more hyperarticulations in order to minimise overlap between categories. German and English have a similarly large vowel inventories [28] (with 15 full vowels in German [29]). However, while hyperarticulations are regularly reported for English [5, 7], we find none for German. The absence of VSA extension in our data questions hyperarticulation as a feature dependent on the vowel inventory – at least in NNAR for mid to high proficiency addressees. Concerning vowel duration, our results confirm previous studies [9, 10] that find no influence of the addressee on vowel duration. As indicated by [11], articulation rate tended to be slower for non-native addressees though we do not find an effect of the non-natives' proficiency. This could be due to the addressees' higher proficiency or the fact that we examine *articulation rate* instead of *speech rate* (where slower speech rate was a consequence of longer pauses rather than vowel lengthening). It is possible that pauses are longer depending on the non-natives' proficiency in German NNAR (similarly to English NNAR, cf. [9]). The differences between our results and previous studies could stem from varying situational factors such as the absence of an actual interlocutor in favour of an imaginary one in [9] which [30] found to result in a stronger effect than a natural interlocutor, non-native addressees with different L1 or proficiency levels that might be subject to a bias [11], or different conversational tasks (e.g., Diapix task compared to free conversations [10]).

For a more in depth analysis, more data on German NNAR are necessary to statistically confirm the tendencies found in the present study. Additionally, variation in AR could be due to phonetic convergence: [31] report that Dutch L1 speakers converge to L2 speakers (inter alia in speech rate) though the effect was stronger for non-native addressees with lower accent ratings. This indicates that our participants' AR measurements might be influenced by the confederates AR. Future studies should take alignment into account and include the confederates' data as well. Therefore, we are planning to include effects of phonetic convergence in our analyses.

FUNDING

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

5. REFERENCES

[1] Lüdeling, A., Alexiadou, A., Adli, A., Donhauser, K., Dreyer, M., Egg, M., Feulner, A. H., Gagarina, N., Hock, W., Jannedy, S., Kammerzell, F., Knoeferle, P., Krause, T., Krifka, M., Kutscher, S., Lütke, B., McFadden, T., Meyer, R., Mooshammer, C., Müller, S., Maquate, K., Norde, M., Sauerland, U., Solt, S., Szucsich, L., Verhoeven, E., Waltereit, R., Wolfsgruber, A., Zeige, L. E. 2022. Register: Language users' knowledge of situational-functional variation. *Register Aspects of Language in Situation (REALIS)* 1(1), 1–58.

[2] Bell, A. 1984. Language style as audience design. *Language in Society* 13(2), 145–204.

[3] Campbell-Kibler, K. 2010. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change* 22(3), 423–441.

[4] Hay, J., Jannedy, S., Mendoza-Denton, N. 1999. Oprah and/ay: Lexical frequency, referee design and style. *Proc. 14th ICPhS*, 1389–1392.

[5] Piazza, G., Martin, C. D., Kalashnikova, M. 2022. The acoustic features and didactic function of foreigner-directed speech: A scoping review. *Journal of Speech, Language, and Hearing Research* 65(8), 2896–2918.

[6] Roche, J. 1998. Variation in xenolects (foreigner talk). *Sociolinguistica* 12(1).

[7] Bradlow, A. R., Bent, T. 2002. The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America* 112(1), 13.

[8] Kendi, A. A., Khattab, G. 2019. Acoustic properties of foreigner directed speech. *Proc. 19th ICPhS*, 810–814.

[9] Biersack, S., Kempe, V., Knapton, L. 2005. Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. *Ninth European Conference on Speech Communication and Technology*, 2401–2004.

[10] Uther, M., Knoll, M. A., Burnham, D. 2007. Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication* 49(1), 2–7.

[11] Hazan, V., Uther, M., Grunlund, S. 2015. How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles? *Proc. 18th ICPhS*.

[12] Hinnenkamp, V. 1982. *Foreigner Talk und Tarzanisch: eine vergleichende Studie über die Sprechweise gegenüber Ausländern am Beispiel des Deutschen und des Türkischen*. Hamburg: Buske Verlag.

[13] Schroedler, T., Purkarthofer, J., Cantone, K. F. 2022. The prestige and perceived value of home languages. insights from an exploratory study on multilingual speakers' own perceptions and experiences of linguistic discrimination. *Journal of Multilingual and Multicultural Development*, 1–18.

[14] Council of Europe, 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

[15] Meer, P., Hartmann, J., Rumlich, D. 09 2021. Attitudes of German high school students toward different varieties of English. *Applied Linguistics* 43(3), 538–562.

[16] Van Engen, K., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., Bradlow, A. R. 2010. The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53(4), 510–540.

[17] Baker, R., Hazan, V. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43, 761–770.

[18] Bullock Oliveira, M., Sell, B. PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels. Type: dataset.

[19] Audacity team, 2020. Audacity (version 2.4.2). https://www.audacityteam.org/.

[20] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

[21] Schiel, F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. *Proc. 14th ICPhS*, 607–610.

[22] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, 392–410.

[23] Winkelmann, R., Jaensch, K., Cassidy, S., Harrington, J. 2020. emur: Main package of the emu speech database management system.

[24] Winkelmann, R. 2015. Praattoformants2asspdataobj.r.

[25] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49(2B), 606–608.

[26] Michalke, M. 2017. *sylly.de: Language Support for 'sylly' Package: German*. (Version 0.1-2).

[27] Al-Tamimi, J.-E., Ferragne, E. 2005. Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. *9th European Conference on Speech, Communication and Technology*. Newcastle University.

[28] Maddieson, I. 2013. Vowel quality inventories. *The world atlas of language structures online*.

[29] Kohler, K. J. 1999. German. In: *Handbook of the International Phonetic Association*. Cambridge: Cambridge Univ. Press, 86–89.

[30] Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., Brenier, J. 2007. An acoustic study of real and imagined foreigner-directed speech. *Journal of the Acoustical Society of America* 121(5), 3044–3044.

[31] Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., Lemhöfer, K. 2021. Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *Journal of Phonetics* 88, 101076.

A STUDY ON CAREGIVERS SPEECH IN RETIREMENT HOMES

Jean-Luc Rouas[1], Yaru Wu[2,3,4], Takaaki Shochi[1,5]

[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[2]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[3]Laboratoire de Phonétique et Phonologie UMR7018, CNRS-Sorbonne Nouvelle, France
[4]LISN UMR 9015, CNRS, Univ. Paris-Saclay, France
[5]CLLE-ERSSàB CNRS UMR 5263, Bordeaux
jean-luc.rouas@labri.fr, yaru.wu@unicaen.fr, takaaki.shochi@labri.fr

ABSTRACT

This paper describes our efforts at characterising the caregivers voice in retirement homes. Towards this aim, we recorded 20 professional caregivers in two retirement homes. These caregivers were recorded using a headset microphone connected to a smartphone, allowing a total liberty of movement while keeping a very high sound quality. The caregivers were recorded while performing three different tasks: text reading, informal interview and professional roleplay with a fictive patient. All the recordings were then processed by an automatic speech recognition system that gives words and phones sequences and their timestamps. In our analysis, we focus particularly on the differences between spontaneous speech conditions and look for contrasts in terms of duration, speech rate, fundamental frequency and intensity. Our final aim is to capture the specificity of the professional caregivers voice in order to build automatic training tools.

Keywords: Affective speech, caregivers voice, speaking styles, spontaneous speech

1. INTRODUCTION

Human social interaction is an exchange of social information conveyed by voice, eye contact, gestures and facial expressions [1, 2]. Among these modalities, prosodic features of affective expressiveness are well known as an important modality which conveys a speaker's various affective meanings [3, 4].

Verbal communication is an important part of interaction, especially with dependent elderly people. However, in the hospital setting, most caregivers are focused on their work, and their verbal interactions with patients are brief. One report describes that verbal communication lasted only 2 minutes each day for a bedridden patient with dementia in a long-term care facility [5]. Caregivers are indeed often discouraged by laconic patients who do not respond or give irrelevant answers.

However, it has been shown that an intentional positive expressiveness demonstrated by professional caregiver can have a strong impact in the contexts with elderly dementia patients [6]. For instance, [7] investigated the influence of vocally expressed emotions and moods by professional caregivers to / or with persons with severe dementia. The results showed the positive effect of vocally expressed caregiver singing for these patients. They reported that the singing enhanced patients' positive emotions and diminished their aggressive behavior.

To systematize positive interactions, the "Humanitude" method aims at maintaining good communication skills that must be developed based on face-to-face interactions, verbal communication and touch interaction. Several studies have shown that using this method results in a significant reduction in patients' aggressive behavior and less dependence on neuroleptics [8, 9, 10].

Concerning the vocal communication skills, the "Humanitude" method relies on phonetic and lexicological elements as well as on a technique called *auto-feedback* where caregivers speak without interruption even when care recipients respond inadequately. Thus, two categories of parameters must be studied. On the one hand, prosodic parameters (intensity, speech rate, fundamental frequency) should be in line with the recommended soft, calm and melodious voice, on the other hand, lexical elements should aim to convey positive emotions.

As a first step, this current work investigates common and different points of three speech style text reading, spontaneous talk, professional talk). Then, we aim to identify some prosodic patterns of these speaking styles.

This paper briefly summarizes the recording protocol of the "tender care" corpus produced by French professional caregivers with the description of the specific equipment used to allow freedom of

movement, the tasks carried out, and the recording settings (Section 2). Then, it describes how we pre-processed the files to obtain the phonetic transcription which is then used to extract acoustic features on Inter-Pausal Units (Section 3), followed by the analysis of the features (Section 4) and the discussion of the results (Section 5).

2. RECORDING PROTOCOL

2.1. Equipment

To ensure total mobility of the recorded participants, we designed a completely autonomous recording setup. We equipped our participants with a headset directional microphone DPA 4288 CORE linked to a iRIG PRO preamplifier connected to a Samsung Galaxy A51 smartphone. Both the preamplifier and the smartphone are fitted in a waist bag ensuring complete freedom of movement while the headset microphone guarantees a high recording quality.

2.2. Tasks

2.2.1. Text Reading

The first task is text reading. The text chosen is *The North Wind and the Sun*. This text has been used for more than a century by the International Phonetic Association to illustrate many of the world's dialects and languages. The text is presented to our French speakers in its French version "La bise et le soleil". The goal here is to record a first control voice in a very structured context. The text reading generally lasts less than a minute.

2.2.2. Informal interview

The interview task consists of a questionnaire with open-ended questions about their work as a caregiver and their general daily routine. The goal is to get the person to talk about themselves as much as possible. This is why impersonal questions about work were chosen so as not to embarrass the interviewee. This exercise also allows the subject to gain confidence.

2.2.3. Professional care task

The experiment consists in recording speech during the staging of a care task with a fictitious unresponsive patient. The chosen care task is dressing up, which includes buttoning a shirt and the techniques of waking up the body. At the end of the dressing, the caregiver performs an uprighting of the fictitious patient and makes them walk. The fictitious patient is completely mute – which allows to evaluate the technique of *auto-feedback* on the part of the caregiver – and lets themselves be cared for by the caregiver. For the care task, the context was set up to create a familiar environment for the caregiver. The bed used is similar to that of the regular patients and partitions are put in place to create intimacy with the fictitious patient.

2.3. Collected data

The acquisition of the audio recordings was done in two retirement homes for dependent elderly people: *Les Balcons du Lot* in Prayssac and *Les résidences du Quercy Blanc* in Castelnau-Montratier in the southwestern part of France. Three recording sessions were carried out: two at the Prayssac establishment, on September 24, 2021 and March 25, 2022 and one at Castelnau-Montratier on November 24, 2021.

We recorded a total of 25 participants during the three sessions : 21 female participants and 4 male. Given that we only had 4 male participants, we decided to focus on female participants in this study. One recording of a female participant was excluded due a poor recording quality. The total duration of remaining 20 participants' recordings is 2 hours and 30 minutes. The total duration recorded for each task and the mean duration per speaker per task is given in Table 1.

Task	mean dur.	total dur.
Text reading	45.0 s.	15 m. 03 s.
Interview	134.8 s.	44 m. 56 s.
Professional care	188.9 s.	62 m. 58 s.

Table 1: Mean duration per speaker per task and total recorded duration per task. All 20 recorded subjects participated in each task.

3. FEATURES

3.1. Automatic phonetic transcription

Phonetic transcriptions were obtained by using an automatic system based on the Kaldi framework [11] and trained using the ESTER database [12]. The model used for the transcriptions is a *Time Delay Neural Network* coupled with a hidden Markov model. The neural network is based on a time-delayed subsampled neural network (TDNN) with 7 TDNN layers and 1024 units in each. The input for the acoustic model is a 40-dimensional high-resolution MFCC vector concatenated with a

100-dimensional I-vector [13].

3.2. Parameter extraction

Automatic segmentation in Inter-Pausal Units (IPUs) is carried out using the output of the automatic transcription system with a threshold of 250 ms. Some shorter pause segments were still present after the automatic segmentation in IPUs. The IPUs were therefore further divided into smaller IPUs between pauses. Hereafter, the IPU refers to the smaller IPUs between pauses.

Fundamental frequency (F0) and intensity were extracted every 10 ms using the snack routines [14]. Mean and standard deviation of the F0 and intensity values are computed on each IPU. The fundamental frequency was converted to semitones (ST) relative to a reference value (here, 50 Hz).

The speech rate was computed on each IPU using the phones detected by the phonetic transcription system and is therefore given in phones per second.

4. RESULTS

Linear mixed models (LMM) were used to analyze the acoustic parameters using the *lme4* package in R [15]. A model is conducted for each acoustic parameter. Speech style was included as a fixed factor for all models. For the models on IPU duration and speech rate, intercepts were included for subject as random effect. Concerning the models on fundamental frequency in semitone, the standard deviation of fundamental frequency across IPUs and the intensity, intercepts were included for subject and item as random effects.

Figure 1 presents the duration of the IPUs (in seconds) as a function of the three speech styles. The results show that IPUs on average are shorter for the interview [β = -2.210; t = -2.87; SE = 0.771] and reading tasks [β = -3.334; t = -4.32; SE = 0.772] than for professional voice.

This observation is in line with what is expected by the use of the *auto-feedback* technique in which the caregivers are asked to keep a steady vocal flow through the length of the care.

This lengthening of IPUs in the professional task is linked to a lower speech rate, as seen on Figure 2. Our speakers tend to speak faster when reading [β = 14.48; t = 6.99; SE = 2.07] or being interviewed [β = 11.17; t = 5.62; SE = 1.99] than when they are caring. This phenomenon is in line with our expectation, since when trying a keep the steady vocal flow when caring, our caregivers naturally tend to slow their speech rate. A lower speech rate is also a indication of a calmer voice.

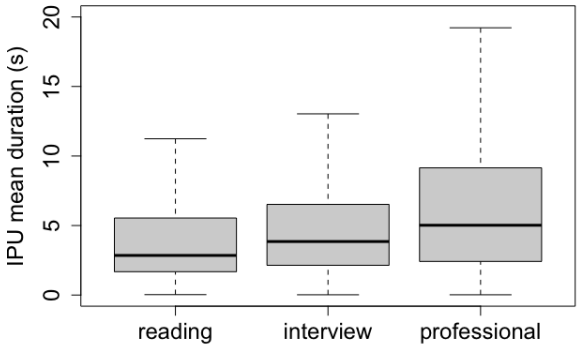


Figure 1: Boxplot of IPU length in seconds for the three speaking styles

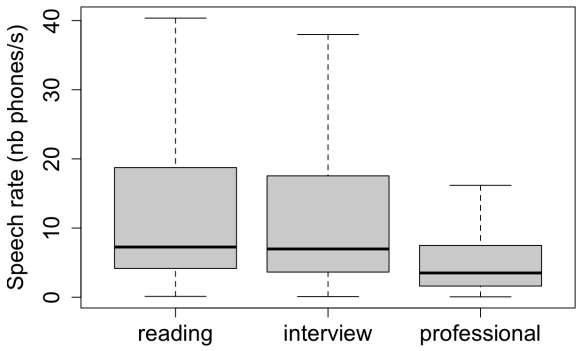


Figure 2: Boxplot of the Speech Rate in phones per second for the three speaking styles

Figure 3 illustrates the fundamental frequency (in semitones) as a function of the three speech styles. We observe that fundamental frequency is on average lower for the informal interview [β = -2.797; t = -2.71; SE = 1.034] than for the professional voice. No significant difference is found between reading and the professional voice. This finding implies that the caregiver try to speak with a significantly higher voice following the instruction of the "Humanitude" care method during a professional setting.

The melodic variations in terms of standard deviation of the fundamental frequency across IPUs are displayed on Figure 4. This figure shows that caregivers produce less variation in fundamental frequency when interviewed than when they are doing professional care [β = -0.210; t = -1.25;

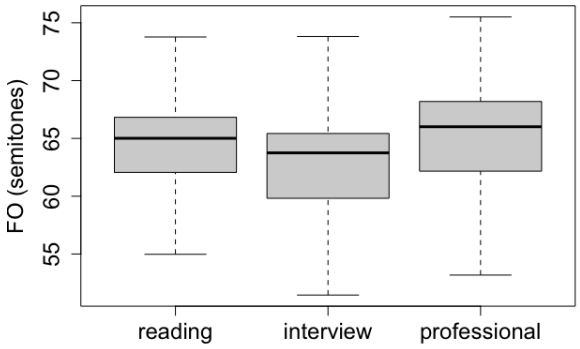


Figure 3: Boxplot of Fundamental Frequency values average over IPUs for the three speaking styles

SE = 0.168]. As far as the standard deviation of the fundamental frequency is concerned, no significant difference is found between reading and the professional voice, based on the relevant LMM model. Having greater variations in their use of fundamental frequency is corresponding to the expectations of producing a melodic voice as recommended by the Humanitude care method.

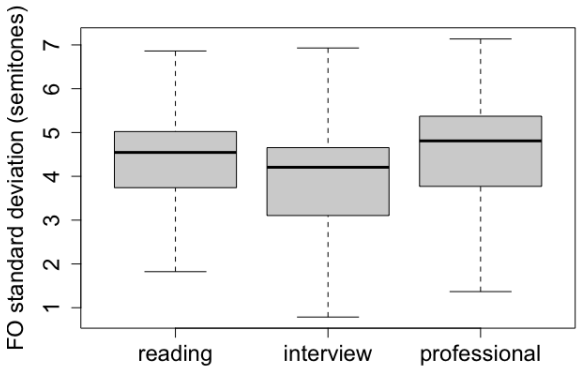


Figure 4: Fundamental frequency standard deviation across IPUs in semitones for the three speaking styles

Figure 5 shows the measured intensity (in dB) as a function of the three speech styles in question. The intensity is observed to be higher for the interview [β = 2.98; t = 2.48; SE = 1.20] and reading [β = 4.78; t = 3.98; SE = 1.20] conditions than for the professional voice. This finding also corresponds to our presumptions since it relates to the fact that caregivers are expected to speak in a calm voice.

The differences observed on intensity for the three different styles may also explain the fact that our speakers tend to produce higher pitch speech when caring because it compensates for the loss in intelligibility induced by the lower intensity.

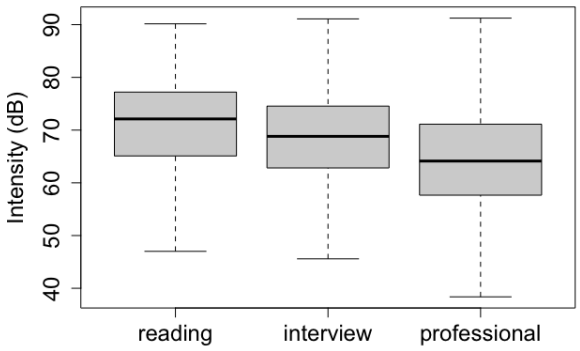


Figure 5: Intensity in decibels for the three speaking styles

5. DISCUSSION

This paper aims to identify the characteristics of caregivers professional speech in contrast with their production while reading or performing an informal interview. As stated in the introduction, caregivers are recommended to produce a soft, calm and melodious voice.

The three prosodic parameters (i.e. fundamental frequency, intensity, duration) we studied here are all in line with this recommendation. For instance, we observed longer IPUs, lower speech rate, greater F0 variation, higher F0 and lower intensity during professional care than during interview or/and reading. The "soft voice" characteristics may be related with this lower intensity and higher F0 voice instructed by Humanitude method. This Â« softness Â» of the voice may also be related to other acoustic parameters like voice quality in interaction with fundamental frequency and intensity. Further studies are needed to confirm or refute this hypothesis.

The next step of our research will focus on the lexical content of the caregivers' speech. It is indeed expected that their professional interactions convey positive emotions. We plan to investigate this assumption by using the transcriptions generated by a high-performance end-to-end automatic speech recognition system coupled with lexical and sentiment analysis tools.

6. ACKNOWLEDGMENTS

This work has been partially funded by the French RNMSH "Humavox" project.

7. REFERENCES

[1] A. Wichmann, "The attitunal effects of prosody, and how they relate to emotion," in *Proc. of ISCA Workshop on Speech and Emotion*, Newcastle, 2000, pp. 143–148.

[2] N. Campbell, "Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language," *Language Resources and Evaluation*, vol. 39, no. 1, pp. 109–118, Feb. 2005.

[3] K. R. Scherer and T. Bänziger, "Emotional expression in prosody: A review and an agenda for future research," in *Speech Prosody*, Nara, Japan, 2004.

[4] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson, and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and speech*, vol. 52, no. 2-3, pp. 223–243, 2009.

[5] Y. Gineste, R. Marescotti, and J. Pellissier, "L'humanitude dans les soins," *Recherche en soins infirmiers*, vol. 94, no. 3, pp. 42–55, 2008.

[6] Y. Gineste and J. Pellissier, *Humanitude*, nouvelle édition: armand colin ed., 2007.

[7] E. Götell, S. Brown, and S.-L. Ekman, "The influence of caregiver singing and background music on vocally expressed emotions and moods in dementia care: A qualitative analysis," *International Journal of Nursing Studies*, vol. 46, no. 4, pp. 422–430, Apr. 2009.

[8] M. Honda, M. Mori, S. Hayashi, K. Moriya, R. Marescotti, and Y. Gineste, "The effectiveness of French origin dementia care method; Humanitude to acute care hospitals in Japan," *European Geriatric Medicine*, vol. 4, p. S207, Sep. 2013.

[9] M. Honda, M. Ito, S. Ishikawa, Y. Takebayashi, and L. Tierney, "Reduction of Behavioral Psychological Symptoms of Dementia by Multimodal Comprehensive Care for Vulnerable Geriatric Patients in an Acute Care Hospital: A Case Series," *Case Reports in Medicine*, vol. 2016, p. 4813196, 2016.

[10] M. Ito and M. Honda, "An examination of the influence of Humanitude caregiving on the behavior of older adults with dementia in Japan," in *Proceedings of the 8th International Association of Gerontology and Geriatrics European Region Congress*, vol. 2018, 2015.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[12] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *In In: Proceddings of Interspeech, Brighton (United Kingdom*, 2009.

[13] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2014.

[14] K. Sjölander. (2004) The snack sound toolkit.

[15] R Development Core Team, *R: A La nguage and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

Acoustic characteristics of dog-directed speech: The role of individual empathy level and personality traits

Jasmin Devi Nuscheler, Nathalie Elsässer, Angelika Braun, Katharina Zahner-Ritter

University of Trier, Department of Phonetics, Germany
jd.nuscheler@web.de, {s2naelsa | brauna | k.zahner-ritter}@uni-trier.de

ABSTRACT

We typically adjust our speech when addressing children or pets. So far, we know very little about the (speaker-specific) factors that explain the degree of such speech style modifications. Here, we compare the acoustic characteristics of speech to a dog (a Mini Maltese, dog-directed speech, DDS) vs. speech to an adult (adult-directed speech, ADS) in German. We test whether the degree of modification between the two speech styles (DDS vs. ADS) relates to a speaker's empathy capacity and personality traits. Results corroborate previous findings on a higher f0 in DDS. Additionally, vowels were fronted and lowered (hyperarticulation), and the postalveolar fricative /ʃ/ tended to be retracted in DDS. The degree of modification between DDS and ADS differed considerably between speakers, but speakers' personality profiles were only weakly correlated with the extent of individual speech style modification – except for spectral cues and conscientiousness, and a tendency for mean f0 and empathy.

Keywords: dog-directed speech, acoustic features, individual variation, empathy, personality

1. INTRODUCTION

A number of studies have shown that speakers phonetically tailor their speech to their interlocutors [1], e.g., when talking to children [2], foreigners [3], or dogs [4]. We know, however, far too little about the speaker-specific factors that determine the degree to which speakers modify their speech. The present study contributes to this new line of research (cf. [4, 5]), focusing on dog-directed speech by non-dog owners in German.

Dog-directed speech (DDS) differs acoustically from adult-directed speech (ADS) [2, 6]: Several studies have reported a higher mean f0 in DDS compared to ADS [6]. Findings on f0 range are inconclusive, with some studies reporting a wider f0 range in DDS than ADS [7], while others do not [4]. Beyond f0 adjustments, [7] and [8] have also demonstrated hyperarticulated vowels in DDS as compared to ADS, i.e., an extension of the vowel space measured via the first (F1) and second formant frequencies (F2), but see [6]. Using higher mean f0 as well as hyperarticulated vowels has been associated with the expression of emotions [8], attracting attention [9], or with the idea of appearing smaller and thus less threatening [2, 10]. Indeed, a dog's size seems to affect speakers' f0, with small dogs being addressed in a higher-pitched voice than big ones [11]. These acoustic differences between ADS and DDS have so far only been studied in dog-owners.

While the acoustic characteristics of different speech styles are quite well understood, we still know very little about the factors that account for inter-speaker variability in the degree of speech style modification [4, 5]. Speakers' personality traits might affect the degree to which they modify their speech: In particular, personality traits have been shown to be relevant with respect to the degree of phonetic accommodation, with openness and neuroticism correlating with how far speakers converge to their interlocutor [12]. For DDS, [4] recently showed that the personality dimension *openness* (positively) predicted the f0 range of female (but not male) dog-owners in DDS. Similarly, a speaker's empathy level might influence acoustic adjustments in DDS. Humans empathise with animals [13], and individual empathy levels affect the way we talk and interpret language [14, 15], for instance with respect to the use of prosodic cues to question intonation or ambiguity resolution [15].

The present study sets out to systematically test whether acoustic modifications between German ADS and DDS also pertain to non-dog owners (**RQ1**). If non-dog owners modify their speech in a similar way to dog owners, we expect higher and more variable mean f0, and more peripheral vowel quality in DDS as compared to ADS.

The second research question (**RQ2**) is more explorative in nature and investigates whether speaker-specific personality traits and a speaker's individual empathy level relate to larger acoustic differences between speech to an adult vs. a dog (in the sense that higher levels of empathy are associated with larger speech style modifications between ADS and DDS etc.).

2. METHODS

Non-dog owners interacted with an adult (ADS) and a 3,5-year-old mini-Maltese, called *Jenna* (DDS, Fig. 1). We used the same dog to control for effects of size [11]. She has a calm and reserved temper and is generally judged as cute. The study was approved by the Ethics Committee of the University of Trier (EK Nr. 35/2022); all participants gave written consent.

2.1. Participants

Twelve German non-dog owners (six male, six female; mean age: 24.3 years, SD = 2.1) participated for a small remuneration. To avoid effects of familiarity, they did not have children, nor did they know the experimenter or the dog. They all showed high affection to dogs, as revealed by a questionnaire (mean = 0.82 on a scale between 0-1, SD = 0.12).

2.2. Materials

Three German nouns were selected to elicit the three cardinal vowels /i:/, /a:/, /u:/ in the two scenarios (see [2, 6, 7]): *Biber* [ˈbi:.bɐ] 'beaver', *Schal* [ʃa:l] 'scarf', and *Schuh* 'shoe' [ʃu:]. Targets were selected such that they were picturable and easy to use as toys. For the picture description task (ADS), 15 (licence-free) pictures were selected (5 pictures for each target). The objects were shown in different contexts (e.g., beaver swimming in a pond or eating some fruit) to increase variation. For the play situation (DDS), we used real objects as toys, i.e., two cuddly toy beavers, two scarfs, and a pair of black sport shoes, see Fig. 1.

2.3. Procedure

Speech recordings. Participants were recoded when interacting with an adult (female, 23 years) and a dog (3,5-year-old mini-Maltese, fixed order) in a quiet room at the University of Trier using a DPA-headset microphone (4088 DCA) connected to a portable recorder (Zoom H4n Pro; 44.1 kHz, 16 bit), which they carried in a little waist bag to be able to move around freely. ADS was elicited via a picture-description task [5] in which participants described 15 pictures to the experimenter (pseudo-randomized such that the same targets did not occur right after each other). In DDS, participants used toys to engage the dog in a shared play, see Fig. 1. Recordings (both phases) lasted about 10 minutes.

Questionnaires. The Big Five Personality Test (B5T, 5-point Likert Scales) [16] was used to assess the five personality dimensions: extraversion (8 items), neuroticism (8 items), conscientiousness (9 items), openness (10 items), and agreeableness (10 items), totalling 45 items.

Participants' empathy capacity was assessed via the *Saarbrücker Persönlichkeitsfragebogen zur Messung von Empathie* (SPF, 5-point Likert Scales) [17], which has been used successfully to predict empathy levels in different contexts [18]. The questionnaire includes four subgroups (perspective taking, fantasy, empathic concern and personal distress), but only the first three (16 items in total) are summed up to an individual empathy score [19]. Questionnaires were filled in via *SoSci Survey* [20] after completion of the recordings (15-25 minutes).

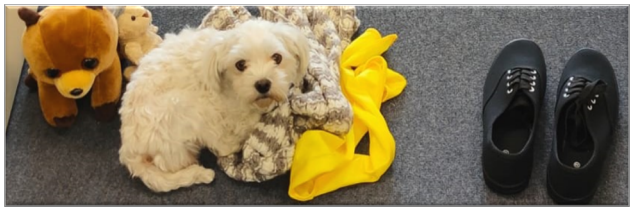


Figure 1: Study-setup in DDS speech style.

2.4. Data processing and statistical analysis

Recordings were manually annotated in Praat [21]: On Tier 1, intonational phrases were segmented, and on Tier 2, the stressed vowels in the three targets were labelled. We additionally labelled the post-alveolar fricative /ʃ/ in *Schuh* and *Schal*, based on an auditory impression of a retracted fricative in DDS.

We excluded phrases in which participants talked in ADS while the dog was present (N = 8), as well as whispered targets (N = 1 in ADS; N = 20 in DDS), see Table 1 for an overview of the dataset.

	ADS	DDS
No. of utterances	1614	1452
No. of target *Biber*	121	79
No. of target *Schal*	103	41
No. of target *Schuh*	221	75

Table 1: Number of utterances and targets per speech style.

Acoustic information was extracted via a Praat script: mean f0 and SD f0 on the phrase level, and the first two formants in the middle of the stressed vowel, as well as the centre of gravity (COG) and its dispersion (SD COG) in the middle of /ʃ/. For formant extraction, we used the linear predictive coding algorithm ([5], t-step = 0.01s, nr of formants = 5, wind. length = 0.025 s, pre-emphasis from = 50 Hz). For male speakers, the maximum formant value was set to 5000 Hz; for females, to 5500 Hz (except for /u:/, where the number of formants was set to 2 and the max frequencies to 1200/1500 Hz). Formant values were manually corrected if necessary. For fricatives, we used a stop Hann band filter to remove

f0 before extracting a spectral slice from the middle of the fricative ([22], wind. length: 0.005 s; max freq: 12000 Hz; t-step: 2 ms; freq. step: 20 Hz, wind. shape: Hamming raised sine-squared).

Statistical analyses were run in R(Studio) [23]. To assess the acoustic differences between ADS and DDS (**RQ1**), we used linear mixed regression models. *Speech style* (ADS vs. DDS) was entered as fixed factor in all models. The model for formants additionally contained the factor *vowel*. Random slopes (subject, items) were retained if they improved the fit of the model [24, 25].

To test for a relationship between the personality traits and the degree of modification between the speech styles (**RQ2**), correlation analyses were conducted. To this end, so called difference scores were calculated for each subject for each individual acoustic parameter (e.g., mean f0 in DDS minus mean f0 in ADS). We then correlated these difference scores with the personality trait scores (the empathy level and the five personality dimensions).

3. RESULTS

3.1. Acoustic differences between ADS and DDS

F0. There was an effect of speech style on speakers' mean f0 ($\chi^2(1) = 56.2$, $p < 0.0001$), with DDS being higher than ADS for both female and male speakers (interaction between gender*speech style: $p = 0.14$; males had a numerically larger difference between speech styles). All speakers produced DDS with higher f0, even though the difference between speech style showed considerable individual variation, see Fig. 2. F0 was significantly more variable in DDS than in ADS for male speakers (ß = 1.39, SE = 0.38, $t = 3.63$, $p < 0.01$), but not for females ($p = 0.1$); the extent of modification again varied across speakers.

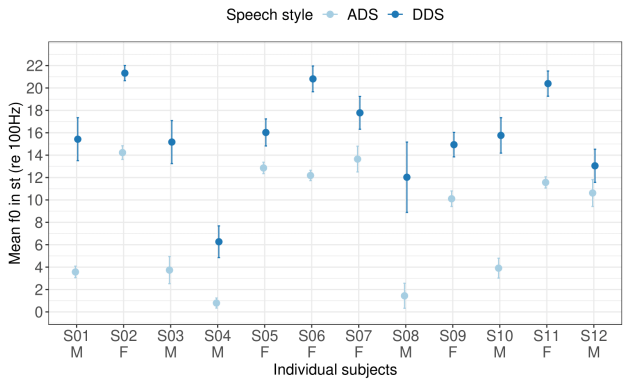


Figure 2: Mean f0 in semitones per speech style.

Vowel quality. Figure 3 shows the averaged vowel space spanned by F1 (vertical dimension, higher values indicate lowered vowel) and F2 (horizontal dimension, higher values indicate fronted vowel). There was no interaction between speech style and gender ($p = 0.23$) or vowel and speech style ($p = 0.13$). However, there was an effect of speech style on the first formant ($\chi^2(1) = 4.6$, $p < 0.05$), which was carried by /u:/ ($p = 0.03$) and /i:/ ($p = 0.06$), and absent in /a:/ ($p = 0.69$). For the second formant, there was an interaction between speech style and vowel ($\chi^2(2) = 22,5$, $p < 0.0001$) and no interaction between speech style and gender ($p = 0.11$): All vowels showed a higher second formant, but the difference was smallest for /u:/. However, it must be taken into account that F2-values for /u:/ are relatively small anyway. The distinction for both F1 and F2 across speech styles varied considerably across speakers.

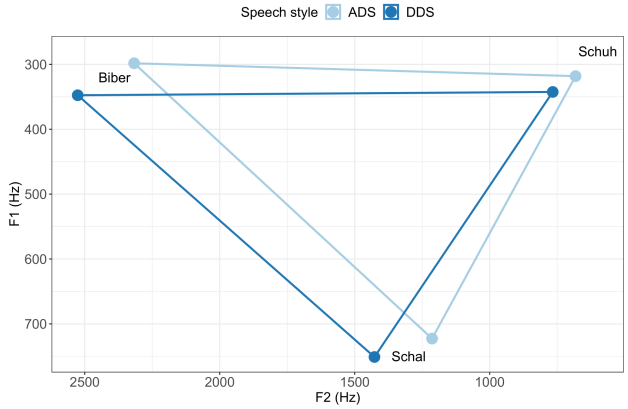


Figure 3: Vowel spaces in both speech styles.

Fricative. An exploratory analysis of the postalveolar fricative /ʃ/ revealed a tendency for the dispersion (SD COG) to be larger in DDS than ADS (ß = 141.24, SE = 69.3, $t = 2.0$, $p = 0.07$), possibly suggesting a retraction of the fricative. Gender and speech style did not interact ($p = 0.26$); again, participants varied in the extent of modification.

3.2. Relation between personality traits and acoustic modification between DDS and ADS

Figure 4 shows the individual personality profiles (B5T and Empathy) in proportions (0 = lowest, 1 = highest). Reliability of the scales was tested via Cronbach's α, with the empathy subscales ranging between 0.57 and 0.79, and the five dimensions of the B5T between 0.77 and 0.92. Speakers not only differed in the extent of speech style modification (3.1), but also showed considerable variation in their personality traits (see Fig. 4). We now turn to the question of whether the extent of individual differences between DDS and ADS relates to a speaker's personality traits (RQ2).

The correlations between the acoustic changes according to speech style and the personality profiles generally fall short of significance. There is one

strong correlation between the dispersion (SD COG) in the postalveolar fricative and the B5T personality dimension *conscientiousness* ($r = 0.57$, $p = 0.05$): The more conscientious a participant, the bigger the difference in dispersion between the fricative in DDS and ADS, i.e., the modification of speech style.

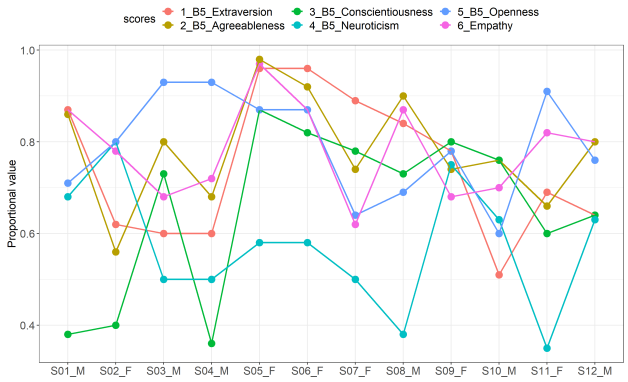


Figure 4: Personality profiles per subject.

On a descriptive level, however, we see some tendencies for relations between personality traits and the extent of speech style modification, e.g., for empathy and the difference in mean f0 (see Fig. 5). Figure 5 shows that for female speakers, higher empathy scores go along with a larger difference between DDS and ADS in mean f0 (red dots; except for one speaker with a high empathy score and a small extent of modification). Post-hoc analyses revealed that larger differences between DDS and ADS in females were due to an increase in f0 in DDS (not ADS) with increasing empathy levels.

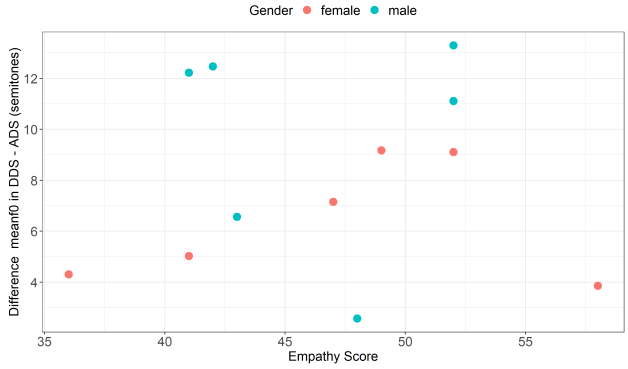


Figure 5: Relation between difference in mean f0 (DDS-ADS) and empathy score.

4. GENERAL DISCUSSION

This study investigated acoustic differences between German DDS and ADS in non-dog owners (**RQ1**), considering the potential influence of personality traits on speech style modification (**RQ2**). Regarding **RQ1**, our findings corroborate previous findings on dog-owners with respect to a higher mean f0 [6]. The participants' general affection to dogs as well as the small size of the dog [11] might have led to this robust difference in all speakers. Results on the variability of f0 were less consistent, showing a larger SD in f0 for male speakers only (see also [4, 7]). Previous studies (on infant-directed speech, IDS) have suggested that an increase in vowel space could be due to increased smiling and hence an overall increase in F2 [26, 27]. Our study, however, revealed differences for both F1 and F2 (lowering and fronting) and is hence most compatible with the interpretation of speakers trying to sound small and non-threatening (shortened vocal tract through raised larynx, cf. [2]), signalled by higher f0 and higher formants [5]. The finding on the spectral distribution in the postalveolar fricative /ʃ/ (tendency towards a larger dispersion of the COG in DDS than ADS), which suggests a backward movement in the place of articulation (i.e., towards the palatal fricative /ç/) [22]), might also be interpreted along these lines: The retraction in place of articulation might reduce the hissing sound in sibilants and hence be an attempt to sound less threatening. This is different from IDS, where sibilants have been shown to be enhanced compared to ADS [28], possibly because IDS serves additional purposes compared to DDS (cf. sound category acquisition [29]). Such interpretations need to be made with care, and future work needs to corroborate this tendency on consonantal differences between DDS and ASD, particularly with regard to the movement of the lips during the interaction with dogs (e.g., smiling). Taken together, our results clearly indicate that non-dog owners (similar to dog-owners) adapt their speech when interacting with dogs (**RQ1**).

The second aim of our study was to test whether speakers' personality traits (B5T dimensions and empathy), factors that have recently been shown to affect phonetic convergence [30] or prosody interpretation [14, 15], account for interspeaker variability in the extent to which speakers adapt their speech in DDS as compared to ADS (**RQ2**). However, speakers' personality traits correlated only weakly with the acoustic modifications, except for conscientiousness and SD COG of /ʃ/. Contrary to [4] we find no relation between openness and f0 range in DDS. Our findings reveal a tendency for higher empathy levels to go hand in hand with larger modifications between DDS and ADS in this group (Fig. 5). Given these tendencies, it is conceivable that relations between personality traits and acoustic cues would be more evident in larger datasets.

To conclude, the results of our study corroborate acoustic differences between speech registers, and at the same time underline the complexity of the relationship between speaker-specific modulations and personality that yet needs to be resolved.

5. ACKNOWLEDGMENTS

The study was financially supported by the *Student Fund* of the Department of Phonetics at the University of Trier.

6. REFERENCES

[1] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Springer, 1990, pp. 403–439.

[2] M. Kalashnikova, C. Carignan, and D. Burnham, "The origins of babytalk: smiling, teaching or social convergence?," *Royal Society Open Science*, vol. 4, no. 8, p. 170306, 2017.

[3] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech," *Speech Communication*, vol. 49, no. 1, pp. 2–7, 2007.

[4] R. Lesch, K. Kotrschal, I. Schöberl, A. Beetz, J. Solomon, and W. T. Fitch, "Talking to dogs: Companion animal-directed speech in a stress test," *Animals,* vol. 9, no. 7, 2019.

[5] M. Weirich and A. Simpson, "Effects of gender, parental role, and time on infant- and adult-directed read and spontaneous speech," *JSLHR*, vol. 62, no. 11, pp. 4001–4014, 2019.

[6] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? On talking to babies and animals," *Science*, vol. 296, no. 5572, p. 1435, 2002.

[7] D. Burnham *et al., Are you my little pussy-cat? Acoustic, phonetic and affective qualities of infant- and pet-directed speech*, 1998.

[8] H. J. Kim, M. Diehl, R. Panneton, and C. Moon, *Hyperarticulation in mothers' speech to babies and puppies*, 2006.

[9] A. Fernald and P. K. Kuhl, "Acoustic determinants of infant preference for motherese speech," *Infant Behavior and Development*, vol. 10, no. 3, pp. 279–293, 1987.

[10] J. J. Ohala, "An ethological perspective on common cross-language utilization of F0 of voice," *Phonetica*, vol. 41, no. 1, pp. 1–16, 1984.

[11] Y.-F. Li and P. P. K. Mok, "Does size matter? An preliminary investigation on the effects of physical size on pitch level in pet-directed speech," in *Proceedings of the International Conference on Speech Prosody 2016*, 2016, pp. 1196–1200.

[12] N. Lewandowski and M. Jilka, "Phonetic convergence, language talent, personality and attention," *Front. Commun.*, vol. 4, 2019.

[13] P. Steen, "Selektive Empathie mit Tieren," in *Sprache und Empathie*, K. Jacob, K.-P. Konerding, and W.-A. Liebert, De Gruyter, 2020, pp. 249–284.

[14] N. Esteve-Gibert, A. J. Schafer, B. Hemforth, C. Portes, C. Pozniak, and M. D'Imperio, "Empathy influences how listeners interpret intonation and meaning when words are ambiguous," *Memory & Cognition*, vol. 48, no. 4, pp. 566–580, 2020.

[15] R. Orrico and M. D'Imperio, "Individual empathy levels affect gradual intonation-meaning mapping: The case of biased questions in Salerno Italian," *Laboratory Phonology*, vol. 11, no. 1, 2020.

[16] L. Satow, *SEA - Skala zur Erfassung von Testverfälschung durch positive Selbstdarstellung und sozial erwünschte Antworttendenzen*: ZPID (Leibniz Institute for Psychology Information) – Testarchiv, 2012.

[17] C. Paulus, *Der Saarbrücker Persönlichkeitsfragebogen SPF(IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index*. https://doi.org/10.23668/psycharchives.9249.

[18] S. K. Poppinga, "Empathievermögen und Belastungsempfinden bei Müttern mit physischen Misshandlungserfahrungen in der Kindheit: Selbstauskunft und hirnmorphometrische Korrelate," Charité - Universitätsmedizin Berlin, 2021.

[19] C. Paulus, "Ist die Bildung eines Empathiescores in der deutschen Fassung des IRI sinnvoll?," 2012. http://dx.doi.org/10.22028/D291-23347.

[20] D. J. Leiner, *SoSci Survey. Professionelle Onlinebefragung*. https://www.soscisurvey.de

[21] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. Computer program.

[22] S. Jannedy and M. Weirich, "Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives," *JASA*, vol. 142, no. 1, p. 395, 2017.

[23] R Core Team, *R: A language and environment for statistical computing.*

[24] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[25] H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates, "Balancing Type I error and power in linear mixed models," *Journal of Memory and Language*, vol. 94, pp. 305–315, 2017.

[26] T. Benders, "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent," *Infant Behav & Development*, vol. 36, no. 4, pp. 847–862, 2013.

[27] E. B. Burnham, E. A. Wieland, M. V. Kondaurova, J. D. McAuley, T. R. Bergeson, and L. C. Dilley, "Phonetic modification of vowel space in storybook speech to infants up to 2 years of age," *JSLHR*, vol. 58, no. 2, pp. 241–253, 2015.

[28] A. Cristià, "Phonetic enhancement of sibilants in infant-directed speech," *JASA*, vol. 128, no. 1, pp. 424–434, 2010.

[29] P. K. Kuhl, "Cracking the speech code: How infants learn language," *Acoust. Sci. & Tech.*, vol. 28, no. 2, pp. 71–83, 2007.

[30] I. Gessinger, E. Raveh, I. Steiner, and B. Möbius, "Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing," *Speech Communica*, vol. 127, pp. 43–63, 2021.

ASSESSING REGISTER VARIATION IN LOCAL SPEECH RATE

Daniel Duran[1], Melanie Weirich[2] & Stefanie Jannedy[1]

[1]Leibniz-Centre General Linguistics (ZAS), Germany [2]Friedrich Schiller University Jena, Germany
duran@leibniz-zas.de, melanie.weirich@uni-jena.de, jannedy@leibniz-zas.de

ABSTRACT

We present results on phonetic register variation (i.e. conventionalized and socially recurring linguistic patterns of intra-individual speech behavior) in contrasting situated interactions. In this paper, we analyze intra-individual differences in local speech rate. Data was elicited with a novel method where participants talk to a video-taped interlocutor in a simulated tele-conference. Each participant mastered a "formal" situation (e.g. requesting a deadline extension from a superior) and an "informal" one (e.g. describing their favorite recipe to an acquaintance). The silent interlocutor persona appeared in two guises differing in hair style, clothing and make-up. This set-up allows us to systematically and consistently manipulate the experimental condition while eliciting laboratory induced differences in fine phonetic detail. We hypothesize that formal situational requirements slow down speech production due to processing and memory load constraints. Statistical analyses for 45 German participants show a slower speech rate in the formal condition.

Keywords: Speech register, intra-speaker variation, formality, speech rate, sociophonetics.

1. INTRODUCTION

This work is embedded in elucidating the various levels of linguistic awareness that speakers have when making contextual linguistic choices. Such selections are being made because speakers do not act uniformly, independent of context, situation or interlocutor. Rather, speakers seem to adjust their vocabulary and wording, forms of addressing and most certainly also the fine phonetic detail of their speech.

In fact, a social awareness of variation in speech style depending, for example, on the addressee [1] or referee [2] has been shown previously. Specific studies on cross-situational variation of single speakers showed that differences in fine phonetic detail like the use of phonation type [3], creak [4] or palatalization of fricatives [5] may serve to construct differences in social meaning, group affiliation and personas. Differences in speech style due to situational (i.e. level of formality or in Labov's sense "attention paid to speech" [6]) and functional (i.e. request, narration) variation and perceived social rank of interlocutor we here refer to as differences in register [7, 8]. Previous studies found a slower speech rate in polite and formal speech in contrast to informal speech in German or Dutch [9, 10]. A slower speech rate has also been associated with a higher cognitive load [11, 12], though findings of effects in the opposite direction, i.e. faster speech rates under higher cognitive load, have also been reported [13].

In order to study if and how fine phonetic detail is being altered based on the speech situation, the level of formality and the interlocutor, we have devised an experimental paradigm that tightly controls for these three factors. The paradigm is novel in the sense that it simulates a set-up that people have become quite used to during the pandemic. Pre-recording the addressee assured that all facettes of the set-up were held constant: the movements, gaze and posture of the interlocutor as well as their clothing, makeup and hair style and the surrounding prompts (coffee cup, smart phone). Thus, in our work, we set out to elicit laboratory induced intra-individual stylistic differences in fine phonetic detail, specifically, in speech rate, with the interlocutor held constant. Our hypothesis is that formal situational requirements slow down speech production due to processing and memory load constraints while casual and informal conversations facilitate faster speaking rates.

2. METHOD

2.1. Stimuli

In a pre-study, the level of formality of the interlocutor was rated by 26 participants on 15 personality attributes. Two stimulus videos were produced based on these previous ratings of perceived personality traits given different hair styles, clothing and make-up of the depicted interlocutor. One video was produced to be perceived as most formal and one to be perceived as least formal. The silent interlocutor persona thus appeared in two guises differing in hair style, clothing and make-up in the formal and the informal stimulus videos (cf. Figure 1). A third neutral video was produced to serve as a familiarization and test-trial during the experiments. One female actor was recorded sitting at a desk and changing slightly in head, arm and upper body

position in a choreographed manner in these different guises while apparently listening to her interlocutor but not speaking.

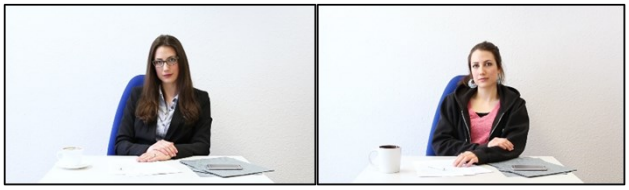


Figure 1: The two personas chosen for portraying the greatest (left) and the least (right) level of formality.

2.2. Task

The experimental setup was designed to simulate a video-call. Participants were seated at a table in front of a large screen which showed the stimulus video. The duration of each stimulus video was roughly 2 minutes. Participants were informed of the duration of each recording and asked to fill the entire recording time if possible. To facilitate conversation, ideas for things to mention were presented to each participant. After the familiarization scenario, participants were recorded in two conditions: a formal scenario and an informal scenario. Three cover stories were prepared for each of the two settings and each participant was assigned one story for the formal and one story for the informal setting.

The participants' task was to talk to the pre-recorded interlocutor in two situations: (a) in the formal, "at-stake" situation, for example, they had to request a deadline extension for a term paper or negotiate a pay raise with a superior, and (b) in the informal, "nothing-at-stake" situation, they had to describe their favorite recipe or their favorite things to do around town to a new neighbor or an acquaintance. These situations were designed to induce the largest possible contrast between the social contexts in terms of formality. Each participant interacted with the formal and the informal video in randomized order.

2.3. Participants

A total of 45 participants were recorded: 15 speakers were recruited in Bremen (9 female, 6 male) and 30 speakers in Berlin (16 female, 14 male, mean age: 24, range: 19 to 35 years). All participants received written information about the procedure and data handling and gave their written consent prior to each recording session. A compensation in accordance with the German minimal-wage requirement was paid for each participation. None of the participants included in this analysis reported any known speech impairments. They were all German speaking residents of Bremen or Berlin (and surroundings).

Data was recorded in two different environments: the group of speakers from Bremen were recorded in a home environment (due to contact restrictions during the early phases of the pandemic, "home space") and speakers from Berlin have been recorded in a sound-attenuated laboratory setting ("lab space").

2.4. Speech Material

All speech data was segmented and annotated with WebMAUS [14] with the "German (DE)" language model. All annotations were manually corrected in Praat [15]. Syllables have been generated automatically with the "emuR" R package [16]. Canonical transcriptions have been adopted from MAUS with manual corrections according to [17], where necessary.

We evaluated approx. 165 minutes of speech data from the 45 speakers. Excluding pauses, approximately, 117 minutes of total vocalization time were analyzed. Figure 2 visualizes the amount of speech data by speaker gender (female vs. male), experimental condition (formal vs. informal) and space (Berlin lab vs. Bremen home). On average, male participants tend to speak less than female participants. Also, both participant groups, tend to speak more in the informal condition in comparison to the formal condition.

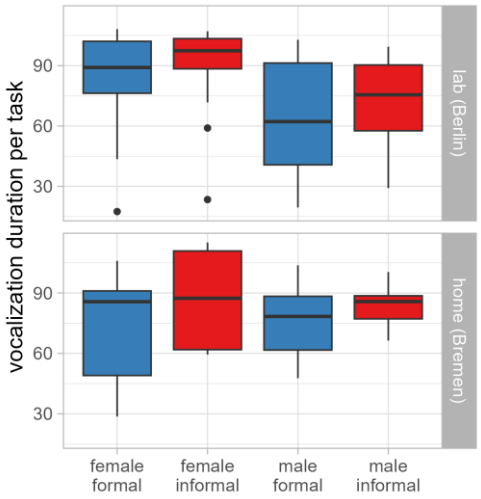


Figure 2: Speech material (only words excluding pauses, i.e. total vocalization time) by speaker gender, condition and recording environment.

The local speech rate was assessed for a total of 7,531 inter-pausal intervals as the number of syllables per second (SPS), calculated as the number of syllables divided by the total duration of the syllables measured in seconds. In addition, we compute "reduction" as the percentage of realized phonetic segments from the total number of canonical segments (based on the manually corrected canonical transcriptions of each word segment). In other words,

we compute reduction as the proportion of "realized" sounds to "intended" phonological segments [cf. 18].

3. ANALYSIS AND RESULTS

We fit linear mixed-effects models (LMER) in R using the "lme4" package [19] for the dependent variables *total vocalization duration (TVD)*, *syllable rate* (expressed as syllables per second, *SPS*) and *reduction*. Speaker *gender* ("female" vs. "male"), *space* ("Bremen home" vs. "Berlin lab") and *condition* ("informal" vs. "formal") are defined as categorical fixed effects. We also include the *session duration* for the modelling of total vocalization time as a control variable. For easier interpretation, we center the values for session duration, which sets the reference level at the intercept of the model to the mean duration of all sessions. We include *speaker* as random effect. We apply forward selection in model fitting by step-wise adding variables and all two- and three-way interaction terms and comparing model fits based on ANOVAs. We compute post hoc pairwise comparisons using the "emmeans" package [20].

4.1 Total vocalization duration (TVD)

We first model total vocalization duration (*TVD*) in order to identify differences between sub-groups within our data set. We start with an intercept-only null model including only random intercepts by speaker. Adding *condition* improves the model fit ($\chi^2(1) = 6.474$, $p = 0.011$). The variables *gender* or *space* do not further improve the model fit. However, adding the interaction term *condition*space* improves the model ($\chi^2(2) = 6.523$, $p = 0.038$). We also find that adding the centered session duration (*CSD*) improves the model ($\chi^2(1) = 53.763$, $p < 0.001$). The best fitting model for the total vocalization time is (in R notation):

$$TVD \sim CSD + condition*space + (1|speaker)$$

The intercept estimate is at 73.98 (SE = 3.54). This corresponds to the base levels *condition* = "formal", *space* = "Berlin lab", and *CSD* = 0, i.e. the mean session duration. The term for session duration is significant (est. 0.793, SE = 0.087, $p < 0.001$), indicating as expected more vocalization in longer session durations. Pairwise comparisons between "formal" and "informal" by *space* show no significant difference for the Berlin-lab group ($p = 0.569$) or the Bremen-home group ($p = 0.881$). Note though that the estimate is larger for the lab speakers (est. 1.385, SE = 2.59) than for the home speakers (est. 0.555, SE = 3.70). Comparisons between the "lab" and "home" *space* settings by formality *condition* show a significant difference between the Berlin-lab and the Bremen-home participants in formal ($p = 0.018$) and informal ($p = 0.023$) conditions, indicating longer vocalization durations in the home space independent of condition. Figure 3 shows the data by *condition* and *space*, and the effect plot of the interaction between these variables.

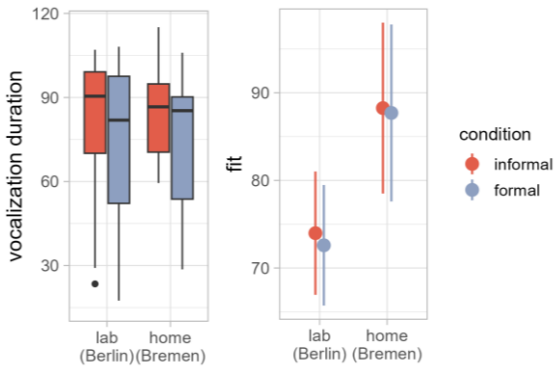


Figure 3: Vocalization duration in seconds by space and condition (left) and the predicted model fit (right panel).

4.2 Local speech rate in SPS

We again start with an intercept-only model for the local speech rate in SPS. Adding *condition* improves the model fit ($\chi^2(1) = 6.149$, $p = 0.013$). Adding *gender*, *space* or any interaction term does not further improve the model. The best fitting SPS model is thus:

$$SPS \sim condition + (1|speaker)$$

The intercept for *condition* = "informal" is at 4.78 (SE = 0.09). The *condition* term is statistically significant (est. -0.137, SE = 0.055, $p = 0.013$), i.e. the model predicts fewer SPS in the formal condition in comparison to the informal condition. Figure 4 shows the local speech rate in syllables per second (SPS) by experimental condition and the model fit.

4.3. Reduction

We start building the reduction model with an intercept-only model including only random intercepts by speaker, as with the previous two models. Adding *gender* ($\chi^2(1) = 4.824$, p = 0.028), *space* ($\chi^2(1) = 20.035$, $p < 0.001$) and their interaction *gender*space* ($\chi^2(1) = 5.169$, $p = 0.023$) improves the model fit. Adding *condition*, however, does not improve the model fit ($\chi^2(1) = 3.522$, $p = 0.061$), nor does any interaction with condition. The best reduction model is thus:

$$reduction \sim gender*space + (1|speaker)$$

The intercept of the reduction model is at 11.83 (SE = 0.81), i.e. for *gender* = "female" and *space* = "Berlin lab". Pairwise comparisons between "male" and "female" by *space* show a significant gender difference in reduction for the "lab" participants ($p =$

0.002), but not for the "home" participants (p = 0.684). A pairwise comparison of *space* by *gender* shows significant differences between home and lab in phonetic reduction for both female (p = 0.021) and male speakers (p < 0.001), with overall more reduction in the home space than in the lab space but a larger difference for male (est. −7.99, SE = 1.60) than for female speakers (est. −3.29, SE = 1.37). Figure 5 shows the reduction by gender and space, and the model fit.

4. CONCLUSIONS

We found that the vocalization duration not only (trivially) depends on the duration of a session. There was also a significant difference between the vocalization duration in the lab and home space in the informal and the formal experimental condition. The participants in the home group (recorded in Bremen) speak more than the participants in the lab group (recorded in Berlin).

The local speech rate, measured in syllables per second (SPS), differs significantly between informal and formal speech with a slower speech rate in the formal condition, independent of speaker gender or space. This is in line with findings of a slower speech rate in polite and formal speech in other studies [9, 10]. In comparison, fastest SPS-rates obtained from a professional speaker were up to 10 SPS [21]. It also supports the hypothesis that formal situational requirements slow down speech production due to a higher cognitive load while casual and informal conversations facilitate faster speaking rates [11, 12].

For reduction, a significant interaction between speaker gender and space was found, revealing more reduction in the home than in the lab space for both genders but with a larger difference for males. This leads to a gender difference only in the lab condition with females showing more reduction than males. Several acoustic studies have found that male speakers show more vowel reduction and elision, shorter sentence and sound durations, and thus speak faster than females [e.g. 22, 23, 24, 25 and 26, 27, 28]. In our study, females show more reduction than males in the lab setting. One reasoning might be that females are better in adjusting to the situation and fulfilling the communicative tasks, elaborating on a given topic, with only little time to prepare, reflecting this in a more reduction.

All in all, our results point at the fact that situational and functional phonetic variation is not only dependent on the speech task, the speaker (gender) and the addressee, but also on the physical location at which conversations take place. It therefore appears that the role of the specific location and therefore mind-set that the speaker was made believe to be speaking in is a factor worth exploring further. Anecdotal evidence already suggests that there is an effect of place of conversation, as speaking more tamed and quietly in a library or in an elevator is culturally agreed upon in most western societies. In further work we will therefore expand our exploration of the effect of place on phonetic realization.

The reasons for the slower speaking rate in formal situations remain to be analyzed in more detail but potentially comprise processing and memory load constraints connected with formal situational requirements but also factors associated with speaker perception and attributed personality styles such as seriousness or integrity.

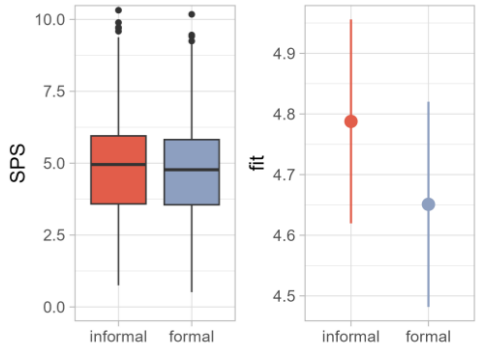


Figure 4: Syllables per second by experimental condition (left panel, omitting 13/3889 extreme data points, i.e. outliers, with SPS > 10 from the plot for readability), and the predicted model fit (right panel).

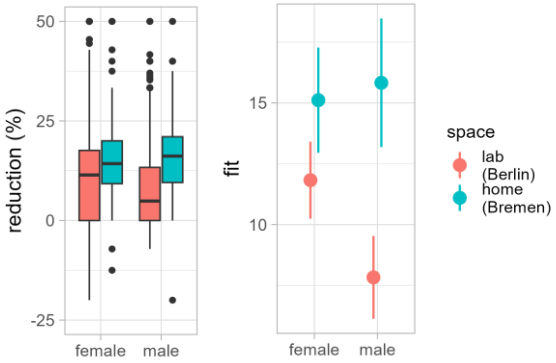


Figure 5: Reduction by speaker gender and space (left, omitting 11/3889 extreme data points from the plot with reduction > 50, and 2 with reduction < −25 for readability), and the predicted model fit (right).

4. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Council) — SFB 1412, grant 416591334, project C02. Ethics approval was obtained from the Ethics Committee of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS, German Society of Linguistics) under the number #2019-07-190625.

5. REFERENCES

[1] A. Bell, "Language style as audience design," *Language in Society*, vol. 13, no. 2, Art. no. 2, 1984.

[2] J. Hay, S. Jannedy, and N. Mendoza-Denton, "Oprah and /ay/: Lexical frequency, referee design, and style," in *The Routledge Sociolinguistic Reader*, M. Meyerhoff and E. Schleef, Eds. Routledge, 2010, pp. 53–58.

[3] R. J. Podesva, "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," *Journal of Sociolinguistics*, vol. 11, no. 4, Art. no. 4, 2007, doi: 10.1111/j.1467-9841.2007.00334.x.

[4] N. Mendoza-Denton, "The Semiotic Hitchhiker's Guide to Creaky Voice: Circulation and Gendered Hardcore in a Chicana/o Gang Persona," *Journal of Linguistic Anthropology*, vol. 21, no. 2, pp. 261–280, Dec. 2011, doi: 10.1111/j.1548-1395.2011.01110.x.

[5] S. Jannedy and M. Weirich, "Sound change in an urban setting: Category instability of the palatal fricative in Berlin," *Laboratory Phonology*, vol. 5, no. 1, Jan. 2014, doi: 10.1515/lp-2014-0005.

[6] W. Labov, The Social Stratification of English in New York City, 2nd ed. Cambridge University Press, 1966/2006. doi: 10.1017/CBO9780511618208.

[7] D. Biber and S. Conrad, *Register, Genre, and Style*. Cambridge, UK: Cambridge, 2009. doi: 10.1017/cbo9780511814358.

[8] V. N. Pescuma et al., "Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods," *Frontiers in Psychology*, vol. 13, 2023, doi: 10.3389/fpsyg.2022.964658.

[9] S. Grawunder, M. Oertel, and C. Schwarze, "Politeness, culture, and speaking task - paralinguistic prosodic behavior of speakers from Austria and Germany," in *Speech Prosody 2014*, May 2014, pp. 159–163. doi: 10.21437/SpeechProsody.2014-20.

[10] K. Koppen, M. Ernestus, and M. van Mulken, "The influence of social distance on speech behavior: Formality variation in casual speech," *Corpus Linguistics and Linguistic Theory*, vol. 15, no. 1, pp. 139–165, May 2019, doi: 10.1515/cllt-2016-0056.

[11HUTT] K. H. Huttunen, H. I. Keränen, R. J. Pääkkönen, R. Päivikki Eskelinen-Rönkä, and T. K. Leino, "Effect of cognitive load on articulation rate and formant frequencies during simulator flights," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1580–1593, Mar. 2011, doi: 10.1121/1.3543948.

[12MP] M. K. MacPherson, "Cognitive Load Affects Speech Motor Performance Differently in Older and Younger Adults," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 5, pp. 1258–1277, May 2019, doi: 10.1044/2018_JSLHR-S-17-0222.

[13LIV] S. E. Lively, D. B. Pisoni, W. Van Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2962–2973, May 1993,

[14] T. Kisler, U. Reichel, and F. Schiel, 'Multilingual processing of speech via web services', *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017, doi: 10.1016/j.csl.2017.01.005.

[15] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2022. [Software]. Available: http://www.praat.org/

[16] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*. 2021.

[17] S. Kleiner, R. Knöbl, Dudenredaktion (Bibliographisches Institut), and Institut für Deutsche Sprache, Eds., *Duden, das Aussprachewörterbuch, 7. Auflage*. Berlin : Mannheim: Dudenverlag ; Institut für Deutsche Sprache, 2015.

[18] J. Koreman, "Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 582–596, Jan. 2006, doi: 10.1121/1.2133436.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[20] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*. 2023. [Software]. Available: https://CRAN.R-project.org/package=emmeans

[21] S. Jannedy, S. Fuchs, and M. Weirich, 'Articulation beyond the usual: Evaluating the fastest German speaker under laboratory conditions', in *Between the Regular and the Particular in Speech and Language*, S. Fuchs, P. Hoole, C. Mooshammer, and M. Żygis, Eds., Peter Lang, 2010, pp. 205–234.

[22] D. Byrd, 'Preliminary results on speaker-dependent variation in the TIMIT database', *The Journal of the Acoustical Society of America*, vol. 92, no. 1, pp. 593–596, Jul. 1992, doi: 10.1121/1.404271.

[23] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, 'Acoustic characteristics of American English vowels', *The Journal of the Acoustical Society of America*, vol. 97, no. 5, p. 3099, 1995, doi: 10.1121/1.411872.

[24] C. Ericsdotter and A. M. Ericsson, "Gender differences in vowel duration in read Swedish: Preliminary results," *Proceedings of Fonetik 2001, XIVth Swedish phonetics conference. Working papers / Lund University, Department of Linguistics and Phonetics*, vol. 49, pp. 34–37, 2001.

[25] Adrian P. Simpson and Christine Ericsdotter, "Sex-Specific Durational Differences in English and Swedish," in *15th International Congress of Phonetic Sciences*, 2003, pp. 1113–1116.

[26] S. P. Whiteside, "Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker sex differences," *Journal of the International Phonetic Association*, vol. 26, no. 1, pp. 23–40, Jun. 1996, doi: 10.1017/S0025100300005302.

[27] A. P. Simpson, 'Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung', *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung (AIPUK)*, 33, 1998.

[28] K. Johnson and J. Martin, "Acoustic Vowel Reduction in Creek: Effects of Distinctive Length and Position in the Word," *Phonetica*, vol. 58, no. 1–2, pp. 81–102, Jun. 2001, doi: 10.1159/000028489.

How do children mark gender phonetically? An analysis of voice quality

Eugene Wong, Benjamin Munson

University of Minnesota
wong0703@umn.edu, munso005@umn.edu

ABSTRACT

Children assigned female at birth (AFAB) have identical vocal tracts and larynxes to children assigned male at birth (AMAB) before puberty. Nonetheless, previous studies have found that adult listeners perceive the speech of AFAB children as more girl-like than that of AMAB children, suggesting that some gendered speech characteristics are learned before the development of vocal tract sex dimorphism. AFAB and AMAB children differ in many segmental acoustic measures. This study examined differences in voice quality between the two groups of children, given the well documented voice quality differences between adult men and women. We examined various acoustic indices of voice quality of 110 English-speaking children at three and at five years-of-age. We found only minimal differences in the voice quality between the groups at both time-points. We found that these measures can predict gender ratings of children's speech, suggesting that people apply stereotypes about adult speech when rating children.

Keywords: sociophonetics, voice quality, gender, selective learning, language acquisition

1. INTRODUCTION

It is well documented that adult men and women speak differently. Some of the phonetic differences found between cisgender men and women can be attributed to the sexual dimorphism in their speech-production mechanisms. As a group, cisgender women tend to speak with higher fundamental frequencies and resonant frequencies than cisgender men. One reasonable hypothesis is that these differences are a passive consequence of sex dimorphism in the vocal tract. However, there are numerous phonetic differences between men and women that are not attributable to sex dimorphism. Studies of children's speech lend support to the conjecture that gendered speech represents culturally and linguistically specific learned ways of speaking.

Prior to puberty, there are no consistent group-level differences in vocal-tract sizes between children assigned female at birth (AFAB) and children assigned male at birth (AMAB) [1, 2]. Throughout this paper, we use AFAB and AMAB instead of 'girl' and 'boy', as we reserve the latter set of terms for studies in which children were asked directly for their gender identity, and in which options more than just 'boy' and 'girl' were given. The speech of AFAB and AMAB children is not identical: the speech of children as young as 2.5 years of age can be robustly distinguished based on sex assigned at birth (SAB) by adult listeners [3]–[5]. In these studies, the speech of AFAB children were rated as more girl-like and AMAB children as more boy-like. These findings suggest that children have learned to express gender phonetically by age 2.5. However, it is unclear what phonetic features distinguish the speech of AFAB and AMAB children, and what acoustic cues allow adults to distinguish the two groups of children.

Previous studies examining acoustic features of gender in children's speech have found that some differences in vowel formant frequency values between AFAB and AMAB children emerge at around 4 years of age, but are not markedly and consistently different until 8 years of age. Moreover, there is no gender difference in habitual f0s until 8 years of age [3, 6, 7]. One study [8] found differences in the overall scaling of vowels' formant frequencies differed between AFAB and AMAB children at age 5, but not at age 3. The measure in [8] used the average distance between adjacent formants to estimate the length of the vocal tract during vowel production, which was adapted from [9]. Another analyses of the same group of children's /s/ productions found gender differences in the spectral mean frequency at both 3 and 5 years of age, consistent with differences between adult men and women [10]. Taken together, these studies show that children mark features related to gender early in life, and that their expression of gender change across early development.

One phonetic domain that has rarely been discussed in previous literature on gender differences in children's speech is voice quality. It is generally thought that there are two dimensions of voice quality. One of these reflects the habitual voice settings of an individual. In adults, part of this habitual voice quality is tied to sex dimorphism of the laryngeal structures between adult men and women, given that individuals can still manipulate their voices to a certain degree. This dimension of voice quality variation also reflects upper respiratory health, and explains why voice quality is often discussed in the

context of speech pathologies. The other dimension is intentional variation in voice quality to signal lexical contrasts, indexical meanings, or both. For instance, creaky voice is a variable commonly used by young Americans for social indexical purposes [11]. As the laryngeal structure in prepubertal children are the same [1], any acoustic differences found in the voice quality between AMAB and AFAB children should be indicative of their learning and expression of gender. In previous studies, [12] found that five-years-old AMAB and AFAB children exhibit differences in one index of voice quality, contact quotients in the English /a/ vowel, yet the gendered pattern did not mirror those found in adults. The authors argued that such difference might arise from the larger variability in contact quotients in AMAB children, and concluded there were only minimal differences in laryngeal features between the two groups of children. Another study [13] found that voices of nine-years-old German-speaking AMAB children were perceived as less breathy than AFAB children. The findings from [12, 13] suggest that children might learn ways of producing voice quality that reflect their gender. Nevertheless, recent studies from [4, 5] found that gendered speech emerges much earlier than the age range in [12, 13]. To date, there is no literature that examines gender differences in voice quality in children of 3 to 5 years of age with a large group of participants. To address this gap, we examine the acoustic voice quality differences between AMAB and AFAB children. We also examine whether voice quality measurements predict adult listeners' gender ratings of children's speech, using data reported previously in [5].

2. METHODS

2.1. Participants

Speech samples were collected from 55 AMAB and 55 AFAB children who are monolingual speakers of American English. The children participated in a longitudinal study of vocabulary development, where speech samples were collected at the three time-points at one-year intervals. The first time-point (FTP) was at 2.5 to 3.5 years of age, the second time-point was at 3.5 to 4.5 years-of-age, and the last time-point (LTP) being 4.5-5.5 years of age. In this study, we only examined the speech samples collected in FTP and LTP. The children were recruited in Minneapolis, MN or Madison, WI. Among the 110 participants, seven children (5 AFAB) speak African American English (AAE), and the remaining 103 children speak the local white variety of American English (wAE, 50 AFAB). All children reported no history of speech, language nor hearing disorders. All caregivers provided consent for their children to participate in the study.

2.2. Productions

The productions were a subset of the approximately 100 productions elicited in the larger study. That word set was designed to elicit a variety of word-initial consonants in a variety of vowel contexts. The words were chosen because they were likely to be familiar to children of the age being tested, and hence were not identical across time-points. In this study, we examined a subset of productions consisting of the low vowels /æ/ and /a/ in a primary-stressed syllable for acoustic measurement of children's voice quality.

2.3. Data collection

The children completed a real-word-repetition task at each time-points. In this task, an experimenter played the recorded target words one at a time and asked the child to repeat the words, while showing a picture of the target words, at a randomized order. Some words were repeated twice to elicit sufficient tokens for analysis. The prompts used with the wAE speaking children was a speaker of the local wAE variety, while those used with AAE speaking children was a native speaker of AAE.

In addition to the production experiment, gender ratings of these children were collected from 80 naïve adult listeners. The listeners ratings were the same as those obtained in [5]. In short, the adult listeners were presented the children's productions one-by-one, and were asked to rate on a visual-analogue scale from "definitely a boy" to "definitely a girl". A mean gender rating was then obtained by averaging the listeners' numerical responses for each child. The numerical values ranged from 0 "definitely a boy" to 1 "definitely a girl".

2.4. Data Analysis

Textgrids were created using Praat [14] and aligned with the Montreal Forced Aligner [15]. The alignments were manually inspected and corrected when necessary. Acoustic measurements were performed on the vowels using VoiceSauce [16]. Voice quality of the vowels was gauged using two spectral tilt parameters (H1*-H2* and H2*-H4*), five spectral slopes parameters (H1*-A1*, H1*-A2*, H1*-A3*, H4*-H2k* and H2k*-H5k*), Cepstral Peak Prominence (CPP) and Harmonic-to-noise ratios at four frequency ranges 0-500Hz (HNR05), 0-1500Hz (HNR15), 0-2500Hz (HNR25), 0-3500Hz (HNR35). Asterisks denote the correction of bandwidth and formants. Results of the acoustic measurements were then time-normalized into three

segments of equal duration. Only the middle segment was used for subsequent analysis. In addition, vowel tokens with total duration shorter than 50ms were removed. As a result, our final data consisted of 1,661 vowels for FTP, and 2,071 vowels for LTP.

A principal component analysis (PCA) was then performed. Since the initial data showed a bimodal distribution, which was related neither to vowel differences nor to the productions of particular child or subset of children, the PCA analysis would potentially reduce the bimodality of the data. Moreover, as the acoustic parameters also showed moderate correlation with each other, the PCA analysis was expected to yield a smaller set of acoustic variables. All parameters were centred and scaled before entering the computation of PCA.

3. RESULTS

3.1. Principal components analysis

There were four principal components (PCs) that had eigenvalues greater than 1. In total, they accounted for 84.8% of the variance in the data. Acoustic parameters with loadings over 0.32 in absolute values were considered to form a component. Table 1 shows the acoustic parameters that form the four PCs, together with the percentage of variance explained.

Table 1: Components of PC1 to PC4

PC	Acoustic parameters	% of variance explained
PC1	HNR05, HNR15, HNR25, HNR35	35%
PC2	H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H2k*-H5k*	30%
PC3	H1*-H2*, H1*-A1*, H4*-H2k*, H2k*-H5k*	12%
PC4	H1*-H2*, H2*-H4*, H4*-H2k*	9%

For PC1, the four HNRs shared a positive loading to this component. For PC2, all parameters except H2k*-H5k* shared a positive loading towards the component. This suggest that for PC2, a higher H2*-H4*, and the three higher H1*-An* parameters were associated with a lower H2k*-H5k* value. For PC3, only H4*-H2k* had a positive loading, whereas H1*-H2*, H1*-A1* and H2k-H5k* shared a negative loading. In other words, a higher H4*-H2k* was associated with lower values in the other three parameters. Finally, for PC4, H1*-H2* and H4*-H2k* shared a positive loading, whereas H2*-H4* had a negative loading. This indicates that higher H1*-H2* and H4*-H2k* were associated with a lower H2*-H4* value.

3.2. Voice quality differences between AMAB and AFAB children

After we reduced the acoustic parameters to four PCs, we examined whether the AFAB and AMAB children differ in these measures at each time-point. Linear mixed effect models were constructed using the fitted values obtained from each of the four PCs as response variables. The models included a random intercept for participant. Coefficient-level results are presented in Table 2. AFAB children are used as reference level.

Table 2: Coefficient-level results of the LMER models on voice quality predicted by SAB.

PC	Predictor	β (SE)	t	p
First time-point (FTP)				
PC1	(Intercept)	-0.37 (0.17)	-2.23	0.03
	SAB	0.32 (0.23)	1.35	0.18
PC2	(Intercept)	-0.09 (-0.01)	-0.83	0.41
	SAB	-0.011 (0.15)	-0.07	0.94
PC3	(Intercept)	-0.13(0.06)	-2.21	0.03
	SAB	0.18 (0.08)	2.18	0.03
PC4	(Intercept)	0.06 (0.05)	1.33	0.19
	SAB	-0.06 (0.07)	-0.83	0.41
Last time-point (LTP)				
PC1	(Intercept)	0.20 (0.18)	1.12	0.27
	SAB	-0.086 (0.25)	-0.35	0.73
PC2	(Intercept)	-0.012 (0.11)	-0.12	0.91
	SAB	0.15 (0.15)	0.99	0.32
PC3	(Intercept)	0.028 (0.06)	0.50	0.62
	SAB	0.025 (0.08)	0.32	0.75
PC4	(Intercept)	-0.06 (0.04)	-1.45	0.15
	SAB	0.077 (0.06)	1.22	0.23

As revealed in Table 2, at FTP, only PC3 was found to have a main effect from SAB. Recall that PC3 consisted of a negative loading of H1*-H2*, H1*-A1*, H2k*-H5k* and a positive loading of H4*H2k*. The coefficient intercept of -0.13 indicates that AFAB children had a higher H1*-H2*, H1*-A1* and H2k*-H5k*, and a lower H4*-H2k* than the AMAB counterpart. In other words, AFAB children had laxer voice quality than AMAB children at FTP. At LTP, none of the PCs were found to have a main effect from SAB. Overall, the statistical results suggest that there was at best only a minimal difference in voice quality between the two groups of children.

3.3. Predicting gender ratings

We then examined whether perceived gender ratings collected form adult listeners were predicted by the

measurements of children's voice quality. To facilitate this analysis, for each child we computed average values for the four PCs using the fitted values from the mixed models. We then constructed simple regression models using the four average PC values as predictors, and the gender ratings as the response variable. The regression models were constructed separately for each group of children at each time-points. The initial model included all four PCs, since together they represented the voice quality of the children. The models then went through a backward stepwise procedure to select the best model, based on AIC values. The results showed that only one PC remained in each of the models. This suggests that voice quality parameter might be an acoustic cue to children's SAB, and adult listeners rely on one PC of voice acoustic cue. Table 3 and 4 summarize the coefficient-level results of the best fitted model for each SAB and for each time-point.

Table 3: Coefficient-level results of linear models of gender ratings predicted by voice quality at FTP.

SAB/ Predictor	β (SE)	t	p
AFAB			
(Intercept)	0.586 (0.015)	39.76	<0.001
PC3	0.051 (0.035)	1.49	0.143
AMAB			
(Intercept)	0.498 (0.015)	33.4	<0.001
PC2	-0.041 (0.019)	-2.1	0.041

As shown in Table 3, PC2 had a main effect on gender ratings of AMAB children. The coefficient estimate of -0.041 suggests that a lower PC2 (i.e. a lower H2*-H4*, H1*-An* and higher H2k*-H5k*) was associated with a lower gender ratings, i.e. more towards "definitely a boy". PC3, on the other hand, was retained in the model for predicting gender ratings of AFAB children, suggesting that PC3 might contribute to gender ratings. However, the p-value for the coefficient in the model itself was not statistically significant. Together, these findings indicate that the effect of PC3 on ratings was, at best, small.

Table 4: Coefficient-level results of linear models of gender ratings predicted by voice quality at LTP.

SAB/ Predictor	β (SE)	t	p
AFAB			
(Intercept)	0.584 (0.016)	36.89	<0.001
PC1	0.026 (0.012)	2.24	0.029
AMAB			
(Intercept)	0.45 (0.015)	30.31	<0.001
PC1	0.019 (0.012)	1.52	0.134

In LTP, both the models for predicting gender ratings of AFAB children and of AMAB children consist of PC1. The positive coefficient estimates from both models (0.019 and 0.026) indicate that higher HNR values (the components of PC1) were associated with an increase in gender ratings (towards "definitely a girl").

4. DISCUSSION

We found that the two groups of children differed in one PC (PC3) at 3 years of age, but not at 5 years of age. Specifically, the two groups of children showed speech differences in their spectral shape. A lower H1*-H2*, H1*-A1* and H2k*-H5k*, but higher H4*-H2k*, were associated with the group of AMAB children. To our knowledge, this finding is just the second speech parameter, next to analyses of /s/ production in [10], that distinguished children's SAB at 3 years of age. However, the overall variance explained by PC3 was relatively low (the r^2 is only 12%) in this set of data. This may suggest that only a small subset of children was utilizing voice quality cues in their expression of gender. However, as we found no SAB difference in the voice quality of children at 5 years of age, we argue that there were overall minimal differences in voice quality between the groups of children.

Our results showed that adult listeners seemed to use some acoustic cues of voice quality to rate children's gender, although there were minimal differences in voice quality between the two groups of children. This reflects that adult listeners were generalizing acoustic cues for gender differences in adult talkers when rating children's gender. This may be because f0 cues, which is typically used to distinguish adult voice gender, are absent in young children. In LTP, PC1 (which consists of HNR parameters) predicted children's gender ratings. The use of HNR was also found in previous study in which older children use breathiness to express gender [13]. In FTP, PC2 and PC3 (consists of spectral shape parameters) predicted adults' gender ratings, but not PC1. These pieces of evidence seem to suggest that adults, when they are asked to identify speaker gender in the absence of f0 differences, turn to noise cue (PC1) before spectral cues (PC2 and PC3). The noise cue may be a more salient cue for gender expression in older children and adults. When noise cues are also not salient, adults turn to spectral shape cues. However, both PC1 and PC2 were not the acoustic cues that children use in their gendered speech. The mismatch in production and perception cues suggest that voice quality is not a robust cue for gendered speech in children. Our results also confirm that gender can be expressed and perceived with multiple variables and idiosyncrasy exhibit across individuals. Future studies may explore the set of acoustic cues for perception of gendered speech.

5. REFERENCES

[1] Fitch, W. T., Giedd, J. 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *J. Acoust. Soc. Am.* 106(3), 1511–1522.

[2] Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., Yandell, B. S. 2005. Development of vocal tract length during early childhood: a magnetic resonance imaging study. *J. Acoust. Soc. Am.* 117(1), 338–350.

[3] Perry, T. L., Ohde, R. N., Ashmead, D. H. 2001. The acoustic bases for gender identification from children's voices. *J. Acoust. Soc. Am.* 109(6), 2988–2998.

[4] Fung, P., Schertz, J., Johnson, E.K. 2021. The development of gendered speech in children: Insights from adult L1 and L2 perceptions. *JASA Express Lett.* 1, p. 014407.

[5] Munson B., Lackas, N., Koeppe, K. 2022. Individual Differences in the Development of Gendered Speech in Preschool Children: Evidence From a Longitudinal Study. *J. Speech Lang. Hear. Res.*, 65(4), 1311–1330.

[6] Lee, S., Potamianos, A., Narayanan, S. (1999) Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am.* 105(3), 1455–1468.

[7] Vorperian, H.K., Kent, R.D. 2007. Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. *J. Speech Lang. Hear. Res.* 50 (6), 1510–1545.

[8] Wong, E., Munson, B. 2021. A longitudinal study of gender-specific characteristics of children's vowels. J. *Acoust. Soc. Am.* 150(4), A152.

[9] Lammert, A. C., Narayanan, S. S. 2015. On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLOS ONE* 10(7).

[10] Koeppe, K. 2021. The emergence of gendered phonetic variation in preschool children: Findings from a longitudinal study. M.A. thesis, University of Minnesota, Minneapolis. Available: https://hdl.handle.net/11299/224487.

[11] Podesva, R. J., Callier, P. 2015. Voice quality and identity. *Annual Review of Applied Linguistics* 35, 173–194.

[12] Robb, M. P., Simmons, J. O. 1990. Gender comparisons of children's vocal fold contact behavior. *J. Acoust. Soc. Am.* 88(3), 1318–1322.

[13] Simpson, A. P., Funk, R., Palmer, F. 2017. Perceptual and Acoustic Correlates of Gender in the Prepubertal Voice. *INTERSPEECH 2017*, 914–918.

[14] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program]. Version 6.2.14. Accessed: May 2022. [Online]. Available: https://www.praat.org

[15] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *INTERSPEECH 2017*, 498–502.

[16] Shue, Y.-L., Keating, P. A., Vicenik, C., Yu, K. 2011. VoiceSauce: A program for voice analysis. *Proc. 17th ICPhS* Hong Kong,1846–1849.

Mandarin tone production can be learned under perceptual guidance — A machine learning simulation

Hanyi Meng, Chien-Tzu Chen, Yue Chen, Zirui Liu, Yi Xu
University College London
hanyi.meng.20@ucl.ac.uk; chien-tzu.chen.21@alumni.ucl.ac.uk; yue.chen.1@ucl.ac.uk; zirui.liu.17@ucl.ac.uk; yi.xu@ucl.ac.uk

ABSTRACT

Intuitively, speech production can be learned by imitating proficient speakers in language acquisition. But a recent computational simulation has shown that learning to produce English words can be achieved under the guidance of speech perception, without direct mimicry. In this study, we tested whether similar perception-guided learning also applies to Mandarin tone acquisition. We used PENTAtrainer, a pitch modelling tool to simulate learners' tone articulation, and a trained tone recognizer to simulate tone perception. Three learning methods with different optimization objectives were tested: 1) closeness of fit of f_0 contours, 2) tone recognition by an automatic tone recognizer, and 3) tone recognition plus minimization of mean f_0 difference at the initial learning phase. The results show that method 3 achieved the best learning outcome as evaluated by the tone recognizer and human listeners. Perception-guided tone learning is therefore shown to be effective if learners' exploration range can be reduced first.

Keywords: speech perception, vocal tone learning, computational modelling, speech synthesis

1. INTRODUCTION

It is still unclear how children acquire language spontaneously without explicit adult instructions. A popular idea is that they do it through imitation [16, 23, 24]. However, it has been difficult to computationally simulate such imitative learning [11, 20, 21, 25]. And a major source of difficulty has been the speaker normalization [13] or correspondence [5] problem. For example, because children's vocal tracts are much shorter than adults' [10, 27], their formants are much higher and more dispersed, making it difficult for children to directly imitate adult speech. However, recently it has been demonstrated computationally that this correspondence problem can be largely solved by using speech perception as a guide in production learning [15, 26, 34]. The production of simple English words with high intelligibility can be learned this way without directly imitating any specific utterances.

The effectiveness of simulating perception-guided vocal learning raises the question of whether tone learning can be simulated in a similar way. Theoretically, this is conceivable, and in fact should be easier since tones mainly involve a single acoustic dimension, i.e., fundamental frequency (f_0), and tones have been successfully modelled with PENTAtrainer, a Praat-based prosody modelling tool [31]. However, perception-guided learning with only a single acoustic dimension could also present problems due to the lack of cross-reference to other parameters.

This study is a preliminary test of perception-guided tone learning. But a strategy slightly different from [15, 26, 24] is applied. First, the acoustic imitation was simulated by an algorithm that optimized the matching of f_0 contours across a whole corpus consisting of many utterances produced by multiple speakers (including both males and females). Second, data from the same corpus were used to assess the learning outcome of both imitative and perception-guided learning, thus minimizing confounding in comparison.

2. METHOD

2.1. Corpus

The Xu1999 corpus used in this project was originally recorded for an experimental study of tone and focus in Mandarin [28], which consisted of utterances produced by four female and four male speakers. The utterances were five-syllable sentences composed of three words (two disyllabic and one monosyllabic), as shown in Table 1. As can be seen, the second, third and fourth syllables have varying tones, while the tone of the first and last syllable is always H. The speech was fluent, with a speech rate of roughly five syllables/s.

Word 1		Word 2		Word 3	
HH māomī	'kitty'	H mō	'touches'	HH māomī	'kitty'
HR māomí	'cat-fan'	R ná	'takes'	LH mǎdāo	'sabre'
HL māomǐ	'cat-rice'	F mài	'sells'		
HF māomì	'cat-honey'				

Table 1: Tone patterns and corresponding sentences used as recording material. H, R, L, and F represent high, rising, low, and falling tones, respectively [28].

The sentences in the full corpus also varied in focus conditions: initial, medial, final and no focus [28]. The present study only used the neutral focus sentences, however, to simulate tone learning only.

The corpus was divided into a training set consisting of 6 of the 8 speakers, 3 males and 3 females, and a testing set consisting of 2 speakers (1 male, 1 female).

2.2. Modelling tool

The computational tool was a special-purpose version of PENTAtrainer [31] — an interactive Praat [3] script for modelling speech prosody. PENTAtrainer models tone and intonation by combining built-in articulatory dynamics (target approximation) [22, 33], parallel encoding [29], and global stochastic learning (simulated annealing [14]) [31]. The original version has been shown to generate intelligible and natural-sounding tone and intonation by optimizing f_0 contour fitting [31]. The f_0 fitting can be viewed as a form of learning by imitation, as it tried to maximize the similarity between the learned and the original f_0 contours.

The special-purpose version of PENTAtrainer used in this study included two additional learning methods, *learning by optimizing tone recognition*, and *learning by optimizing tone recognition and minimizing mean f_0 difference* (deltaf_0). The computationally intensive learning task was run by a Python executable called by the Praat script. The tone recognizer, also called by the Praat script, was a support vector machine (SVM) trained by the scikit-learn package [6] in Python with syllable-sized f_0 contours as input data. The trained model was able to recognize both tone and focus with high accuracy [7, 8]. In learning method 3, deltaf_0 was the utterance-wide mean f_0 difference between each original and synthetic contour, and it added a fraction of weight to the tone recognition error in the coarse-tuning phase of the learning:

$$e = 0.9\, e_r + 0.1\, d \quad (1)$$

where $e_r = 10$ (1 – recognition rate [0,1]), and $d = f_{0\text{orig}} - f_{0\text{orig}}$, where $f_{0\text{orig}}$ and $f_{0\text{orig}}$ were utterance-wide mean f_0 of original and synthetic tones, respectively.

The coarse-tuning phase, consisting of the first 450 of the total 750 training iterations, optimized all tonal targets at once in each iteration, while the fine-tuning phase optimized each parameter (height, slope and strength [31]) of each tonal target at a time.

2.3. Procedure

The experiment proceeded in 5 steps:

Step 1. Training the tone recognizer on all the neutral-focus utterances in the Xu1999 corpus. The overall post-training recognition rate was 94.7%.

Step 2. Running PENTAtrainer in the training set with three learning methods, each repeating five times. The main simulated annealing parameters used were: *iteration* = 750, *learning rate* = 0.1, *starting temperature* = 700, and *reduction factor* = 0.98.

Step 3. Averaging the pitch targets of each tone learned from the five runs of each learning method to obtain three sets of tone targets.

Step 4. Running PENTAtrainer in the testing set with the mean tone targets to a) generate f_0 contours that were input to the automatic tone recognizer for tone recognition, and b) resynthesize all the utterances in the testing set with the model-generated f_0 contours.

Step 5. Playing the resynthesized utterances to listeners for perceptual tone identification and judgment of naturalness.

For step 5, the stimuli contained 320 recordings from the testing set, which were divided into four conditions: a) original recordings, b) recordings resynthesized with parameters learned from f_0 fitting, c) recordings resynthesized with parameters learned from recognition only, and d) recordings resynthesized with parameters learned from recognition+deltaf_0. The tones to be identified belonged to the second syllable, which was always /mi/, c.f. Table 1. Syllables in other positions carried fewer tones, varied in segmental compositions, and were not included in the perception test. For the naturalness rating, listeners were asked to judge whether they have heard a human utterance or a computer-generated sound.

The listening subjects were 20 native Beijing Mandarin speakers, who performed the perception tasks on Gorilla, an online experiment platform. They had no history of neurological or communication disorders, and passed a hearing screening at 20 dB HL bilaterally at 125, 250, 500, 750, 1000, 2000, 3000, and 4000 Hz.

3. RESULTS

3.1. Numerical evaluations

Table 2: Numerical assessments of tone learning separated by learning methods.

Learning method	RMSE	Correlation	Recog. rate
F_0 fitting	1.64	0.81	85%
Recog. only	4.12	0.18	66%
Recog.+deltaf_0	2.09	0.72	95%

Table 2 shows the root mean square error (RMSE), Pearson's correlation coefficient (*r*) and tone recognition rate for the three learning methods. As expected, learning by *f_0 fitting* worked well, achieving low RMSE and high correlation, consistent with previous findings on the same corpus [31]. Interestingly, the tone recognition rate, at 0.85, was also fairly high, which was consistent with [7]. *Recognition only* showed poor results, with high RMSE and low correlation, whereas

recognition+deltaf$_0$ had the highest recognition rate at 95%, although its RMSE was higher and correlation was lower than those of f_0 fitting.

Figure 1 shows learning progression in terms of mean RMSE and recognition rate. Both indicators were recorded during learning with all methods, regardless of whether the method itself used them as optimization objectives. As can be seen, in all cases, RMSE was reduced over the iterations while recognition was improved. In all cases, the sudden improvement from iteration 450 was due to the shift from coarse- to fine-tuning phase of learning, as explained in 2.1.

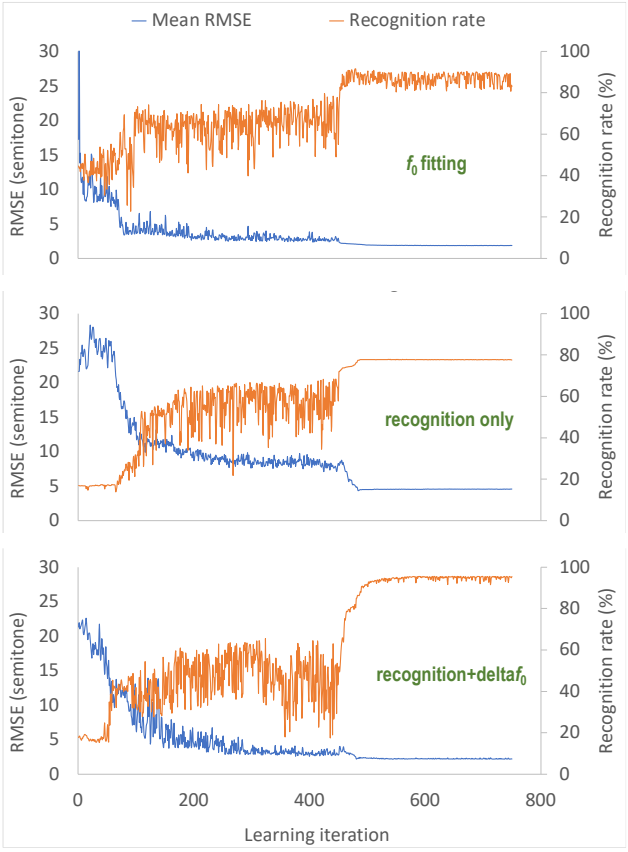


Figure 1: Examples of learning progression per iteration in terms of mean RMSE and recognition rate.

As can be seen, for f_0 fitting, RMSE was reduced to a very low level in the fine-tuning phase, but the increase of tone recognition hovered around 90%. For recognition only, both RMSE and recognition failed to improve much further in the fine-tuning phase. For recognition+deltaf_0, RMSE stopped to reduce below two semitones, but recognition went quickly above 95% after the onset of fine-tuning.

3.2. Human perceptual evaluation

Figure 2 shows perceptual tone identification rates for the four types of stimuli: a) original utterances, b) audios resynthesized with parameters learned from f_0 fitting, c) audios resynthesized with parameters learned from recognition only, and d) audios resynthesized with parameters learned from recognition+deltaf_0.

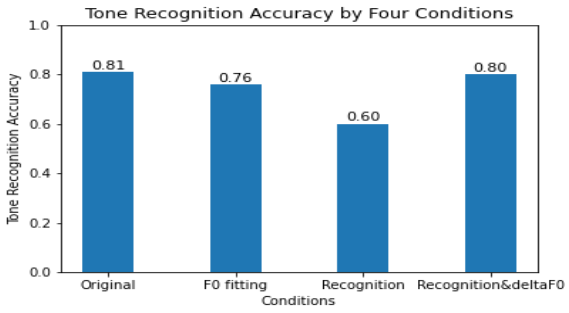


Figure 2: Tone recognition rate in four conditions.

The original utterances achieved the best tone recognition rate at 81%, while recognition+deltaf_0 was the second best at 80%. A two-tailed t-test found the difference between the two conditions non-significant. However, neither of these conditions performed nearly as well as the overall automatic tone recognition rate of 94.7% mentioned in section 2.3. One likely reason is that the recognizer performance was for tones of all syllables in each sentence. Syllables 1 and 5 both always had T1, the high-level tone, whose recognition rate was very high, partially due to over-training. For the tone of the second syllable alone, the recognizer achieved only 90% for the testing set, although this is still much higher than the 81% of the listener recognition of the original tones in Figure 2. Therefore, the superior performance of the recognizer is more likely because it has been trained on the same corpus, whereas the listeners relied on their real-life listening experience, which would include many more speakers with diverse tone articulations. The recognition rate for f_0 fitting was 76%, which was significantly lower than both the original ($t(16) = 6.57$, $p < 0.001$) and recognition+deltaf_0 ($t(16) = 4.89$, $p < 0.001$) conditions.

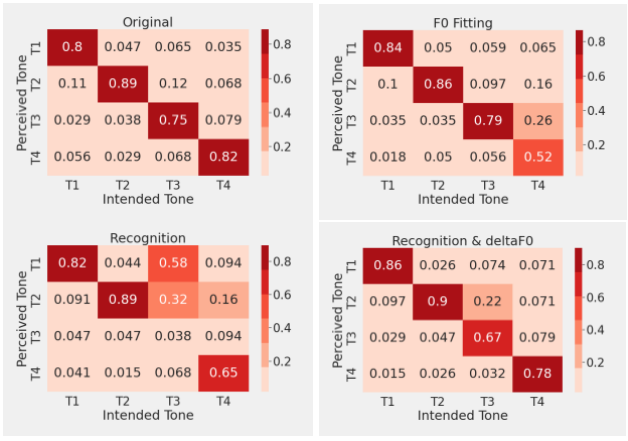


Figure 3: Heat map of confusion matrices for the perception of tones in four conditions.

Confusion matrices of tone perception are shown in Figure 3. These confusions can be examined together with the mean f_0 contours in Figure 4, which are separated by tone of the second syllable. First, for both T1 and T2, the recognition-only condition learned obviously wrong targets, high-fall for T1 and high-level for T2. Curiously, at 82% and 89%, the perception of these two tones did not seem to be severely affected. Second, For T2, the recognition+deltaf_0 condition generated a contour with an extra low minimum f_0 and a sharp terminal rise. This allowed the tone to be perceived (90%) as well as in the original condition (89%). Third, for T4, f_0 fitting generated a contour with a lower f_0 peak than both the original and recognition+deltaf_0 conditions. This is probably why it had a 26% confusion with T3. Finally, in the recognition+deltaf_0 condition, T3 was heard as T2 around 22% of the time. Although this is similar to the original condition where confusion with T2 was 12%, the f_0 contour in the bottom left plot of Figure 4 shows that the greater confusion was likely due to an earlier rise than the original T3.

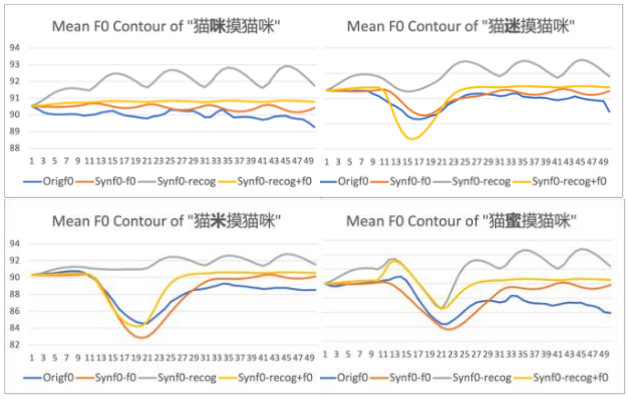


Figure 4: Mean F_0 contours, separated by the tone of the second syllable, clockwise from the top left: T1, T2, T3 and T4. The horizontal axis is normalized time (10 points/syllable). The vertical axis is f_0 in semitones.

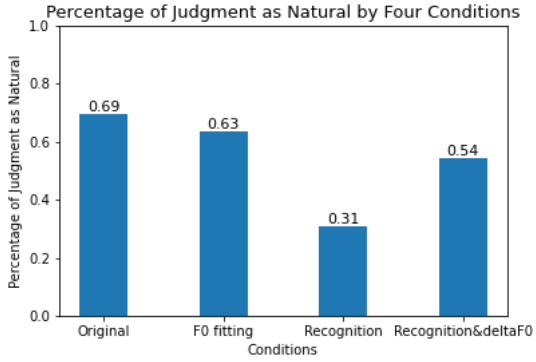


Figure 5: Percentage of utterances judged as natural speech (rather than synthetic) in the four conditions.

Figure 5 shows the results of naturalness judgment by listeners. As can be seen, even the original utterances were judged only 69% as human articulation. Interestingly, utterances from the f_0 fitting condition were judged as more likely to be humanly articulated than those from the recognition only ($t(16) = 7.55$, $p < 0.001$) and recognition+deltaf_0 ($t(16) = 4.97$, $p < 0.001$) conditions.

4. DISCUSSION AND CONCLUSION

This preliminary simulation study has demonstrated that perception-guided vocal learning [15, 26, 34] may also work for tone acquisition, provided that the target exploration range is constrained in the early learning phase, as done in the recognition+deltaf_0 condition. The quality of the learned tones with that method was better than those learned with f_0 fitting as assessed by both automatic tone recognition and human tone perception, except in terms of naturalness.

It is important to note, however, that the f_0 fitting method in this study was not strictly simulating direct mimicry, because it optimized f_0 contour match in all instances of each tone across all the repetitions by all speakers in the training set of the corpus (120 in total). But the optimization in recognition+deltaf_0 was also performed across all the utterances in the training set. In other words, the only difference between these two methods was the learning objective: to maximize the similarity of f_0 contours, or to maximize the tone recognition accuracy. On the face of it, the differences may be hard to comprehend. Why wouldn't achieving maximum acoustic similarity to multiple speakers lead to the best learning outcome? But the results clearly show an advantage for recognition-guided learning. This suggests that the difference in learning objectives is not trivial, as it may reflect *the core nature of speech as a communication system*. Given this nature, the proper objective of vocal learning should be to gain the ability to produce maximally intelligible speech rather than to just sound like other speakers. And the same may also be true of adult speech. That is, what makes a contrastive phonetic unit equivalent across different speakers is that it has been learned in such a way that it is most likely to be perceived as that unit. While this may sound circular as a factual definition, the circularity would disappear once it is treated as an operational definition, as shown in this study.

It is unclear, however, why recognition guidance alone did not work well in this study. Is it indeed due to a lack of cross-reference to other parameters as speculated in the Introduction? If yes, is it possible to introduce some minor adjustments to the current learning algorithm? This will need to be addressed in future studies.

5. REFERENCES

[1] Asada, M. (2016). Modeling early vocal development through Infant – Caregiver interaction: A review. *IEEE Transactions on Cognitive and Developmental Systems, 8* (2), 128-138.

[2] Beecher, M. D., & Brenowitz, E. A. (2005). Functional aspects of song learning in songbirds. *Trends in Ecology & Evolution, 20* (3), 143e149.

[3] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**: 341-345.

[4] Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. *Nature, 417* (6886), 351-358.

[5] Brass, M. and Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in cognitive sciences* **9**(10): 489-495.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[7] Chen, Y., Gao, Y. and Xu, Y. (2022). Computational modelling of tone perception based on direct processing of f0 contours. *Brain Sciences 12* (3), 337.

[8] Chen, Y. and Xu, Y. (2021). Parallel Recognition of Mandarin Tones and Focus from continuous F0. *1st International Conference on Tone and Intonation (TAI)*. Sonderborg, Denmark.

[9] Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science, 298* (5600), 2013-2015.

[10] Fitch, W. T. & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106* (31), 1511–1522.

[11] Howard, I. S. and Huckvale, M. A. (2005). Training a vocal tract synthesizer to imitate speech using distal supervised learning. In *Proceedings of International Conference on Speech and Computer (SPECOM)*, Patras, Greece**:** 159-162.

[12] Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Neuroreport, 17* (10), 957-962.

[13] Johnson, K. (2005). Speaker normalization in speech perception. In *The handbook of speech perception*. D. B. Pisoni and R. E. Remez pp**.** 145-176.

[14] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**(4598): 671-680.

[15] Krug, P., Birkholz, P., Gerazov, B., Niekerk, D. R. v., Xu, A. and Xu, Y. (2023). Artificial vocal learning guided by phoneme recognition and visual information. IEEE Transactions on Audio, Speech and Language Processing.

[16] Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences* **97**(22): 11850-11857.

[17] Kuhl, P. K. (2003). Human speech and birdsong: communication and the social brain. *Proceedings of the National Academy of Sciences, 100* (17), 9645-9646.

[18] Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21* (1), 1–36.

[19] Lin, J., Li, W., Gao, Y., Xie, Y., Chen, N. F., Siniscalchi, S. M., … Lee, C.-H. (2018). Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks. *Journal of Signal Processing Systems, 90* (7), 1077–1087.

[20] Pagliarini, S., Leblois, A. and Hinaut, X. (2021). Vocal Imitation in Sensorimotor Learning Models: A Comparative Review. *IEEE Transactions on Cognitive and Developmental Systems* **13**(2): 326-342.

[21] Philippsen, A. K., Reinhart, R. F. and Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *Proceedings of 4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE**:** 195-200.

[22] Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* **125**: 405-424.

[23] Skinner, B.P. (1957). *Verbal Behavior*. New York: Appleton/Century Crofts.

[24] Speidel, G. E., & Nelson, K. E. (1989). A fresh look at imitation in language learning. *In The many faces of imitation in language learning* (pp. 1-21). Springer, New York, NY.

[25] Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes* **26**(7): 952-981.

[26] Van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., Halliday, L., Prom-on, S. and Xu, Y. (2023). Simulating vocal learning of spoken language: Beyond imitation. Speech Communication 147: 51-62.

[27] Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data.

[28] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics 27: 55-105.*

[29] Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* **46**: 220-251.

[30] Xu, Y. (2019). Prosody, Tone and Intonation. *In The Routledge Handbook of Phonetics*. W. F. Katz and P. F. Assmann: Routledge, New York. pp. 314-356.

[31] Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57, 181-208.

[32] Xu, Y. and Sun X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America, 111* (3): 1399-1413.

[33] Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* **33**: 319-337.

[34] Xu, Y., Xu, A., Niekerk, D. R. v., Gerazov, B., Birkholz, P., Krug, P. K., Prom-on, S. and Halliday, L. F. (2022). Evoc-Learn — High quality simulation of early vocal learning. In *Proceedings of Interspeech 2022.*

Vowel duration before a geminate in children and adults: is lexicality important?

Lucia Colombo* and Joanne Arciuli°

*Università degli Studi di Padova, Italy
°Flinders University, Adelaide, Australia

ABSTRACT

The phonetic realization of Italian vowels includes shorter vowel durations before geminate consonants compared to singleton consonants but only in stressed syllables (/p'ɛttine/, comb). Previous research using nonword repetition showed that Italian children reach adult-like mastery of the vowels production in such contexts after 4 years of age [7]. We investigated whether this developmental trajectory holds when children produce familiar words. Italian children (3-6 years old) and adults performed a picture naming task, eliciting productions of familiar words. Stimuli were two and three-syllable words labelling familiar objects or people, with medial singleton or geminate consonants, and with stress on the initial or penultimate syllable. Acoustic analyses showed overall longer durations of vowels preceding the medial consonant (V1) for young children compared to adults. In contrast to the previous study of nonword production, even the youngest children were able to produce the appropriate adult-like variation in V1 duration in singleton/geminate contexts.

Keywords: Vowel production, Geminates, Lexical stress, Lexicality, Language acquisition.

1. INTRODUCTION

Geminate consonants occur frequently in some languages (e.g., Arabic, Finnish, Italian). Phonetically, they are realized as long consonants. Acoustically, a geminate consonant has a longer duration by comparison with the singleton consonant (e.g., the minimal pair in Italian *note*, notes, vs *notte*, night). In Italian, and some other languages, the realisation of vowels that surround medial consonants differs depending on whether the medial consonant is singleton or geminate [9;12]. In particular, in Italian, the duration of the vowel preceding a geminate is shorter compared to when it precedes a singleton. This difference in vowel duration, that is dependent on the length of the following consonant, is useful for the listener, to better discriminate the geminate/ singleton contrast [9; 11]. However, this shortening of the preceding vowel only occurs for tonic vowels, that is, where the vowel preceding the medial geminate/singleton consonant carries stress [4; 7]). When this vowel does not carry stress, its duration is not shortened. In the present study we investigated the impact of these contextual factors of type of consonant and stress on vowel duration in the production of real words by children and adults.

1.1. Stress in Italian

In Italian vowel duration depends on different contextual factors, one of which is stress. Most Italian words are stressed on the penultimate syllable (~75% of words) (e.g., /bam'bino/, child; /'kane/, dog), while a smaller number of words have stress on the antepenultimate syllable, (~18% of words (e.g., /'tavolo/, table) [6; 16]. Stress on the final syllable is much less frequent (about 2-3%). The main acoustic correlate of stress in Italian is the increased duration of the vowel, compared to a non-stressed syllable, in particular in open non-final syllables [3; 8]. However, the phonetic realization of the vowel preceding the medial consonant is shorter when it precedes a geminate compared to a singleton (e.g., V1 is shorter in CV1CCV2 than in CV1CV2, where the first syllable is stressed: compare/'notte/, night vs /'note/, notes), This poses no problem for the adult native speaker, but for young children there is a developmental trajectory towards adult-like mastery of this aspect of speech production. That is, children must acquire the distinction between geminate and singleton consonants, as well as variable vowel duration related to the nature of the following consonant, as well as stressed vs. unstressed conditions. Italian children show adult-like mastery in terms of variable vowel duration in stressed and unstressed syllables, even at quite a young age [1; 13]. Recent evidence, however, shows that this is not so in productions that involve the medial geminate/ singleton distinction. Note that we distinguish between intelligibility and adult-like mastery. Thus, children's productions may be intelligible but not yet adult-like as revealed by fine-grained acoustic analyses.

1.2. Children's acquisition of stress and geminates

Three- and four-years-old children may not show adult-like mastery of the geminate/singleton contrast [7], which is associated to the duration of both the

consonant, and of the preceding vowel. This has been interpreted as meaning that in children aged 3-4 years the contrast between geminate and singletons, and the ability to phonetically realize this distinction, has not been fully acquired (i.e., is not adult-like).

These results have been shown in a task that involved measuring vowel duration produced via repetition of nonwords which exhibited the geminate/singleton contrast in medial position [7]). In the present study we report the results of an experiment where children were required to name pictures of real objects (i.e., real word production). The aim was to investigate if adult-like mastery of the production of vowels preceding the singleton/ geminate contrast depends on the lexical status of the stimuli (i.e., when familiar words are produced as opposed to nonwords). Indeed, the phonological representation of known words may be stronger and more stable [17] such that the phonetic realization of vowels preceding medial geminate/singleton consonants might be easier to acquire for young children by comparison with their productions of vowels in nonwords.

To test this hypothesis, we selected a number of pictures whose names were: three-syllabic words with penultimate syllable stress and medial singleton, (/ma'tita/, pencil) or medial geminate consonants (/tak'kino/, turkey); disyllabic and three-syllabic words with initial syllable stress and medial singleton, (/'note/, notes; /'sɛdano/, celery) or medial geminate consonants (/'gatti/, cats; /'pattino/, skate shoe). We expected the duration of the vowel preceding a medial singleton consonant to be longer than the vowel preceding a medial geminate consonant in words where this vowel carries stress (i.e., in words like /ˈnote/, notes, and /'sɛdano/, celery, compared to words like /'gatti/, cats, and /'pattino/, iceskate). This prediction should hold for adults and older children, as it did for the repetition of nonwords in Colombo et al.'s [7] study. The crucial question was whether young children (3 and 4 years of age) would be able to realize vowels differently depending on these contextual factors as adults do: when repeating nonwords they did not show adult-like mastery but we might see a different pattern of results when children are naming familiar words.

2. METHOD

2.1. Participants

Overall, 77 children from a pre-school in a small town in north-eastern Italy were recruited for the purposes of the research. Due to technical issues, only the recording from 74 participants were analyzed (42 female and 32 male). The age of the participants ranged from 3 to 6 years: 17 children aged 3 years (mean = 3.62, SD = 0.22), 24 children aged 4 years (mean = 4.54, SD = 0.23), 21 children aged 5 years (mean = 5.58, SD = 0.28) and 12 children aged 6 years (mean = 6.22, SD = 0.18). The children were assigned both the word stimuli of the present study and a task of nonword repetition [7]. The recordings were undertaken individually and in a quiet room of the pre-school building. Only children whose parents signed an agreement form participated to the research. Additionally, 28 adults from north-east Italy were recruited, ranging from 19 to 62 years of age (mean = 33.89, SD = 13.94, 16 females and 12 males). Like the children, the adults were interviewed and recorded in an isolated and quiet room.

2.2. Materials

Several constraints were considered for the word-picture pairs selection. First, the objects corresponding to the words had to be easy to name by young children as well as adults. The tonic syllable of the words must be the first or second syllable. The CV structure of the words had to be as simple and common as possible, to avoid confounds due to structure complexity (in Italian the most common structure is CVCVCV).

Given that the majority of Italian words have penultimate stress, it was quite difficult to find three-syllabic words with initial stress associated with easily visualised familiar objects. Forty-three 2- and 3-syllable words with penultimate or initial stress and medial singleton or medial geminate consonants were initially selected. The CV structure was mostly CVCV- CVCCV for disyllables and CVCVCV for 3 syllables. Words corresponded to common objects, familiar to the younger children. They were photographs of real objects/people, found on the web. Out of the 43 words, 20 were three-syllabic with penultimate stress, 13 of which with a geminate consonant, and 7 with a medial singleton consonant. Five three-syllabic words had initial syllable stress, 3 of which had a medial geminate, and 2 a singleton. Finally, there were 18 disyllabic words, 17 of which with initial stress and a medial geminate. The pictures corresponding to these words were selected and presented in a pilot test to 11 adult participants for picture agreement verification. Pictures which did not elicit the expected name for the picture by at least 50% of the adult participants were discarded.

On the basis of the pilot, some words were excluded from the statistical analysis (mostly because they contained a diphthong after the medial consonant or because of low naming agreement.) After the exclusion of words that did not correspond to the

above-mentioned criteria, the following stimuli remained for the statistical analyses: 13 words with the stressed vowel before the medial consonant (V1) (6 with singleton, 7 with geminate consonant) and 13 words with an unstressed vowel (10 with geminate, 3 with singleton consonants). The corresponding pictures were presented in the naming task in order to elicit speech which could be analysed acoustically.

2.3. Procedure

The naming task was undertaken on a personal computer which displayed each picture embedded as a slide in a PowerPoint presentation. Each test slide was followed by an empty slide. The slide sequence was controlled by the experimenter. Four practice pictures, not part of the experimental materials, preceded the test stimuli. Two different random sequences of the stimuli were assigned to two groups of participants. The participants' responses were recorded with a microphone headset (Reloop RUF 1NH/HS) connected to a computer and with the software Audacity. The original (.aup) format was converted to another format (.wav 32 bit) for the acoustic analyses.

The software PRAAT, version 6.0.28, [5] was used to undertake acoustic measurements of vowels. Waveforms and wide-band spectrograms with a 300-Hz bandwidth were generated for each sound file. Vowel segmentation was made considering concurrent information from amplitude traces, intensity curves and F0 contour. Both the vowel preceding (V1) and the vowel following the medial consonant (V2) were analyzed. The onset of V1 was determined by an increase in amplitude and appearance of the formant structure. The offset of V1 was determined by a drop in amplitude and change in formant structure. The onset of V2 was determined by the release burst of the preceding stop consonant and appearance of formant structure, and its offset by an intensity variation in the waveform and spectrogram. All analyzed vowels were intermediate vowels that preceded a medial consonant (either singleton or geminate consonant).

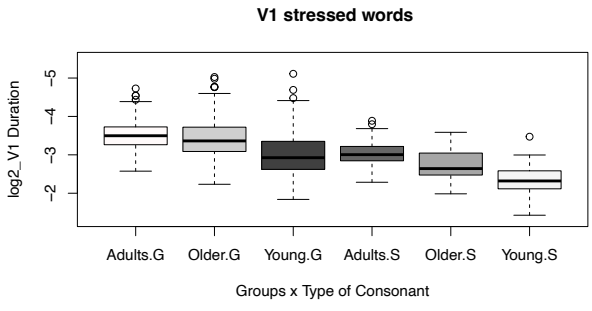


Figure 1: Log2 V1 duration in the stressed vowel V1 in the three groups for words stressed on V1.

3. RESULTS

Perceptual evaluations were made by the first author and the experimenter (both native speakers of Italian) in order to determine whether each production was correct and analyzable. Productions were considered correct if all segments were pronounced correctly, and with correct stress and distinctive single/geminate pronunciations, and if the utterance corresponded to the expected name of the picture. For two of the picture stimuli two names were possible and both were analyzed, when appropriate, one of which was a diminutive (/'gatti/, /ga'ttini/; cat, small cats; /'tattsa/, /tatts'ina/; cup, small cup). Both words of each pair had a medial geminate consonant, with V1 unstressed in the diminutive words.

Acoustic analyses were only undertaken on correct productions. For children, 1518 out of 1920 productions (79.06%) and for adults 668 out of 669 (99.85%) were available for acoustic and statistical analyses. The data from the children aged 3-6 were sorted into two age groups: 3-4 and 5-6, so as to have a simpler statistical design.

The analyses of the dependent variable of vowel duration were conducted using linear mixed-effects modelling in R version 3.5.2 GUI [14], treating participants and items as random effects. Models with random slopes in addition to random intercepts by participants showed convergence problems, so only random effects for intercepts were included. On the basis of the reviewed literature, we expected an effect of the type of consonant only on stressed V1 (not unstressed V1), thus the data for the two types of stimuli were analyzed separately. Fixed factors were group with three levels (3-4, 5-6, and adults, between-subjects and within-items), and type of consonant (singleton/ geminate, within-subjects and between-items). The models were fit by maximum likelihood with the Laplace approximation technique. The lme4 package, version 1.1-21 [2] was used to run the linear mixed-effects models. As no effect was found in the analyses of V2 durations, and none was predicted, only the results of the analyses on V1 durations will be reported. The analyses were carried out on log2 transformed data, in order to ensure a normal distribution.

3.1. V1 in stressed syllables

We used a statistical model with a two-way interaction: lmer(log2_V1_Duration ~ Group * Type of consonant + (1|Participant) + (1|Word)), and it explained more variance compared to the model with only the additive effects, χ^2 = 9.93, p = .001. The ANOVA showed significant effects of group, χ^2 = 228.01, p = 0, type of consonant, χ^2 = 17.27, p = 0,

and of the interaction, $\chi^2 = 9.98$, p = .001. The group effect was due to decreasing V1 durations in young children by comparison with adults (Figure 1). The effect relating to type of consonant was due to longer durations of V1 preceding a singleton than V1 preceding a geminate. Separate contrasts between the means for vowel duration in singleton/geminate contexts for each group showed that the difference in vowel duration for singleton vs. geminate contrasts was significant for each group, but was numerically larger for younger children due to the longer overall V1 duration: t(22.8) = 5.26, p < .0001; t(19.3) = 5.43, p < .0001; t(118.3) = 3.90, p = .001, respectively for young children, older children and adults. Analyses conducted on just three-syllabic words gave a similar pattern of results, with longer durations of V1 preceding a singleton than preceding a geminate for all the groups.

3.2. V1 in unstressed syllables

The means for the unstressed vowel are reported in Figure 2. In contrast to the former analysis, the statistical model including the interaction did not significantly differ from the model with only the two fixed factors. The ANOVA was carried out on the model with two fixed factors, group and type of consonant, and the random factors, participants and words. Only the group effect was significant, with decreasing V1 durations from younger children to adults: $\chi^2 = 289.39$, p = 0. Comparisons of the vowel durations of all the groups were significantly different: (t (874) = 16.37; t(629) = 8.88; and t(282) = 5.24, all p's < .0001, respectively for the young and the older children compared to adults, and for the two children's groups.

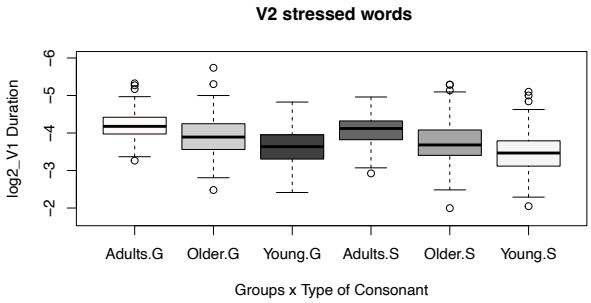


Figure 2: V1 duration in msec in the unstressed vowel V1 in the three groups for words with stress on V2.

4. DISCUSSION

The results of the experiment using speech production elicited via picture naming confirmed that, in Italian, vowel duration in stressed vowels before a geminate is shorter than before a singleton in adults [1; 7]. In the groups of children we tested, the same pattern was found, even though speech rate is slower in children, and therefore vowel duration was overall longer than in adults. In contrast to previous research that elicited speech via nonword repetition, even young children are able to realize vowels differently depending on the contextual factors of a following singleton/ geminate, and on whether the vowel occurs in a stressed syllable or not. Note that the participants in the present study were the same as in the previously published study with nonwords [7]. As has been discussed in the latter paper, this contrast between production of words and nonwords might depend on a variety of factors. One is the familiarity associated with real words. For example, children aged 26 months are able to repeat familiar words better than nonwords even when the words include unfamiliar consonants [10; 15]. This in turn might depend on different factors. First, children may learn words, and thus produce them, as units [17]. Moreover, learned words are retrieved from memory, where their phonological representation is also connected to semantics, which may act as support to the phonological representation. Indeed, the inter-connections between the word phonology and its meaning may reinforce the phonological representation held in the memory buffer.

In contrast, nonwords are sequences of segments never heard before. In order to be repeated, they must be held in a memory buffer, while the production processes, required to articulate them aloud, are activated. Further, nonwords are presumably represented as sequences of phonemes or groups of phonemes in a specific order. The motor program to articulate them is based on this memorized sequence from short -term- memory.

In conclusion, in Italian, realization of vowels duration is affected by a number of contextual factors including length of the following consonant and whether the vowel occurs in a stressed syllable or not. The present study confirms that lexicality is another crucial factor in the acquisition of this contrast. However, the present results do not allow us to determine at what level of representation or implementation the failure to produce the singleton/geminate contrast in V1 duration is located (whether it is perceptual, at the level of phonological or phonetic representation, or even at the motor production level).

In further studies it will be important to investigate whether this word/nonword difference only depends on familiarity of the word to pronounce, or also from other factors, like its meaning.

5. REFERENCES

[1] Arciuli, J., Colombo, L. 2016. An acoustic investigation of the developmental trajectory of lexical stress contrastivity in Italian. *Speech communication*, *80*, 22- 33.

[2] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

[3] Bertinetto, P.M. 1980. The perception of stress by Italian speakers. *Journal of Phonetics*, *8*(4), 385-395.

[4] Bertinetto, P.M. & E. Vivalda. 1978. Recherches sur la perception des oppositions de quantité en italien. *Journal of Italian Linguistics 3*, 97-116.

[5] Boersma, P., & Weenink, D. (2017). Praat v. 6.0.28 [Computer software]. *Amsterdam: Institute of Phonetic Sciences*.

[6] Colombo, L. 1992. Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology, 18,* 987-1003.

[7] Colombo, L., Infanti, M., Arciuli, J. 2022. How is vowel production in Italian affected by geminate consonants and stress patterns? *J Child Lang.* 2022 Nov 3:1-19. doi: 10.1017/S0305000922000526.

[8] D'Imperio, M.R., & Rosenthal, S.S. 1999. Phonetics and phonology of main stress in Italian. *Phonology*, *16*, 1-28.

[9] Esposito, A., & Di Benedetto, M. G. (1999). Acoustical and perceptual study of gemination in Italian stops. *Journal of the Acoustical Society of America, 106*(4, Pt 1), 2051–2062. https://doi.org/10.1121/1.428056

[10] Keren-Portnoy, T., Vihman, M. M., DePaolis, R. A., Whitaker, C. J., Williams, N. M. 2010. The role of vocal practice in constructing phonological working memory. *J Speech Lang Hear Res*. *53*(5), 1280–1293. https://doi.org/10.1044/1092-4388(2009/09-0003)

[11] Krull D., Traunmüller, H. & Bertinetto, P.M. (2006). Local speaking rate and perceived quantity: an experiment with Italian listeners. *Working Papers 52* (Lund University, Dept. of Linguistics), 81-84.

[12] Maddieson, I. 1985. Phonetic cues to syllabification. In: *Phonetic linguistics: Essays in honor of Peter Ladefoged*, Academic Press, 203-221.

[13] Olivucci, F., Pasqualetto, F., Vayra, M., & Zmarich, C. (2016). Lo sviluppo dell'accento lessicale nel bambino in età prescolare: una prospettiva fonetico- acustica. In R. Savy & I. Alfano (cur.), *La fonetica nell'apprendimento delle lingue* (pp. 209-228). Milano: Officinaventuno.

[14] R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org

[15] Roy P, Chiat S. A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: evidence from typically developing children. J Speech Lang Hear Res. 2004 Feb;47(1):223-34. doi: 10.1044/1092-4388(2004/019). PMID: 15072541.

[16] Spinelli, G., Sulpizio, S., & Burani, C. (2017). Q2Stress: A database for multiple cues to stress assignment in Italian. *Behavior Research Methods, 49*(6), 2113–2126. https://doi.org/10.3758/s13428-016-0845-7

[17] Vihman, M., & Keren-Portnoy, T. (Eds.). (2013). *The Emergence of Phonology: Whole-word Approaches and Cross-linguistic Evidence*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511980503

INFANTS' LEARNING OF NOVEL SEGMENTS IS MODULATED BY PROSODY

Kateřina Chládková[a,b], Václav Jonáš Podlipský[c], Natalia Nudga[a,b], Nikola Paillereau[a,b], Kateřina Kynčlová[a,b], Šárka Šimáčková[c]

[a]Faculty of Arts, Charles University, [b]Institute of Psychology, Czech Academy of Sciences, [c]Faculty of Arts, Palacký University Olomouc

katerina.chladkova@ff.cuni.cz

ABSTRACT

Young infants recognize atypical realisations of native-language speech. Later they learn words better from native-accented talkers. However, 6-month-olds preferentially listen to unfamiliar speech. We tested whether the learning of new vowels matches 6-month-olds' listening preferences, being more effective from nonnative-accented speech. We exposed Czech six-month-olds to delexicalised utterances with consonants replaced by [f] and vowels by 405 tokens sampled from a bimodal [ɛ]-[æ] distribution, a contrast absent from Czech, and with either native or atypical rhythm. Discrimination of [ɛ]-[æ] was then tested in an alternating/non-alternating paradigm. Longer first-look duration to non-alternating than to alternating trials – indicating a learning effect – was found in infants familiarised with the novel contrast in atypical rhythm; such effect was not detected after familiarisation with native rhythm. Six-month-olds thus more effectively exploit distributional information about novel vowels from non-native rhythm, which matches their previously reported preferences for listening to novel over familiar accents.

Keywords: distributional learning, selective learning, non-native rhythm, vowels, infancy

1. INTRODUCTION

Infants recognize an atypical accent when their native language is being spoken and have accent-specific listening (and social) preferences. For instance, 5- to 6-month-olds prefer watching a talker who speaks the infants' native language with a native accent over a talker with a foreign accent [1], 10-month-olds prefer engaging with native-accented talkers [1], 17-month-olds learn new words better from native- and familiar-accented speakers [2, 3], and pre-school children preferentially learn the function of new objects form native-accented talkers over foreign-accented ones [4].

To recognize accented speech, infants strongly rely on prosodic cues [5–8], as well as on phonotactic patterning [5] and even on vowel properties [9]. Recently, it has been shown that infants as young as 3.5 to 4.5 months discriminate atypical accents in their native-language speech even when those deviate from the native norm only in minimal changes to durational relations between strong and weak syllables and all other cues are equal [10]. Not only could the infants in that study discriminate native and accented speech solely on the basis of the durational patterning, but they also manifested listening preferences with stimuli differing only in durational rhythm cues: 4-month-olds preferentially listened to the native accent and 6-month-olds preferred the non-native accent.

Interestingly, this behaviour was observed at the developmental stage when infants acquire the categories for their native segments. For vowels, perceptual narrowing occurs between the 5th and 6th month of age [11]. This leads us to ask: how does infants' early sensitivity to, and listening preferences for, durational rhythmic patterns interact with the acquisition of native vowel categories?

A wide-spread view is that prosody bootstraps further language development. Previous studies have demonstrated that prosodic cues facilitate word recognition [12], the acquisition of the lexicon and grammar [13, 14], and that rhythmic priming facilitates language comprehension in typically-developing as well as language-delayed children [15]. Here we ask whether prosody can bootstrap the formation of segmental categories. The specific question is whether infants exploit distributional information about a novel vowel contrast more effectively from speech with native versus atypical rhythm, implemented as atypical patterning of durational relations between strong and weak syllables.

The literature suggests that infants might learn about segments more effectively from the type of input for which they also have listening

preferences. Infants preferentially listen to infant-directed speech (IDS) over adult-directed speech (ADS) [16] and it has been proposed that they may also learn new segments selectively better from IDS than ADS [17]. One of the functions of IDS, differing from ADS in suprasegmental as well as segmental cues [18], may be to facilitate language acquisition. The question is whether selective learning would align with preferential listening if that means not only learning from more perceptually attractive and often also more salient input (as in the case of IDS) but even learning from *less native* input. This is what we put to test here by investigating whether 6-month-old infants, who have been shown to preferentially listen to rhythmically atypical speech, also learn novel segments more readily from such rhythmically atypical speech.

We measure infants' ability to exploit distributional information about a novel vowel contrast. Infants have been reported to be able to employ a distributional learning mechanism to uncover category-like structures in their speech sound environment [19, 20], although the nature of the (to-be-)learnt speech categories has been questioned [21]. The present experiment tests whether infants discover the input's distributional structure of segmental categories with different success depending on whether it is embedded in native versus nonnative rhythm, implemented, respectively, as native-like versus atypical patterning of durational relations between strong and weak syllables.

2. METHOD

2.1 Participants

Forty-one 5.5- to 6.5-month-old infants acquiring Czech participated (5 other infants were excluded for fussiness). All were born full-term, had normal hearing and no familial risk of dyslexia. They were randomly assigned to the native (n = 21) or the non-native rhythm (n = 20) condition.

2.2 Stimuli

The training stimuli were adapted from our prior study on rhythm preferences [10]. Three women, native speakers of Czech, recorded utterances in Czech with native rhythm and with atypical rhythm, in which the durational ratio between word-initial and non-initial syllables was increased (to achieve consistent rendition, the speakers imitated a resynthesized model recording, see [10]).

In the present "[fɛfæfɛ] design," the utterances were delexicalized by replacing all Cs with the same token of [f] and all Vs with one of 405 unique tokens sampled from a continuous bimodal distribution [22] with peaks at [ɛ] and [æ], whose midpoint F1 and F2 values were 730 Hz and 2125 Hz, and 1015 Hz and 1885 Hz, respectively. Original segment durations were always preserved. Different [fɛfæfɛ] utterances from the 3 speakers were combined to create 8- to 9-sentence trials with native or non-native rhythm. The post-exposure test stimuli were isolated [ɛ] and [æ] with F1 and F2 of the training-distribution peaks and 125 ms long.

2.3 Procedure

In the training phase, the infants were exposed to the bimodal [ɛ]-[æ] distribution embedded in delexicalized utterances either in native or non-native rhythm. There were 5 training trials, with the average duration of 24.5 s, amounting to ~2 minutes of training. The subsequent test phase comprised of two blocks, each containing an alternating [ɛ]-[æ] and a non-alternating [ɛ]-[ɛ] trial (order counterbalanced). Each test trial contained 16 vowels (8 pairs) and lasted 8.3 s.

The infants were tested in a central fixation paradigm implemented in PyHab [23] while seated on the lap of their parent, in a sound-treated booth. A camera mounted on top of the central screen recorded the infants' looking behaviour. Stimulus presentation was not contingent on infant looking. The data were coded offline by a coder blind to the stimuli and trial types.

2.4 Looking time measures

We analysed first-look duration, following [24, 10]. In a paradigm such as ours, with stimulus presentation *not* contingent on infant looking – i.e. a paradigm in which trials do not end as soon as the infant stops looking – first look duration may better capture infants' true discrimination abilities or listening preferences than does total looking time. Arguably, the latency of the first looking away reflects the infant's genuine attention to the presented auditory stimulus, while any repeated looks towards the stimulus throughout the trial that goes on irrespective of the infant's attention, may as well be just random. Besides first look duration we also analysed total looking time because it is a more commonly used measure in the literature.

Both measures were analysed separately for the Training and for the Test. The models for the Training phase assessed preferential listening, and those for the Test phase assessed distributional learning.

2.5 Statistical analyses

First look duration and total looking time were log-transformed and each submitted to a linear mixed-effects model, *lme4, lmerTest* in R [25–27].

The analyses of Training modelled the fixed effects of Familiarization type (sum-coded: -native, +non-native) and Trial number (mean-centred), and random per-participant intercepts and slopes for Trial number.

In the analyses of the Test phase, the fixed factors were Familiarization (-native, +non-native), Test-type (-alternating, +nonalt.), Test-block (-first, +second), and their interactions. The random effects were per-participant intercepts and slopes for Test-type and Test-block.

3. RESULTS

3.1 Training phase

The model for first look duration in the Training phase yielded a significant intercept and an effect of Trial number, indicating that the infants' attention decreased throughout the familiarisation phase (estimated slope [log scale] = -0.192, *SE* = 0.094, *df* = 40.134, *t* = -2.034, *p* = 0.049). The effect of Familiarization type was not significant: the mean first look duration to nonnative rhythm was numerically (but not significantly) longer than the mean first look duration to native rhythm; see left panel in Figure 1.

The model for total looking time detected a significant intercept and an effect of Trial number, again confirming that the infants' looking times decreased throughout familiarisation (estimated slope = -0.084, *SE* = 0.031, *df* = 39.881, *t* = -2.750, *p* = 0.009). The effect of Familiarization type with the *t* value of 1.961 (estimated at 0.062, *SE* = 0.032, *df* = 38.502, *p* = 0.057) indicated that infants in the non-native rhythm condition tended to look longer during familiarisation than infants in the native rhythm condition (native: mean = 17.6 s, 95% CI = 16.0–19.3 s; non-native: mean = 19.9 s, 95% CI = 18.1–21.9 s); see left panel in Figure 2.

3.2 Test phase

The model summary for first look duration in the Test phase is shown in Table 1.

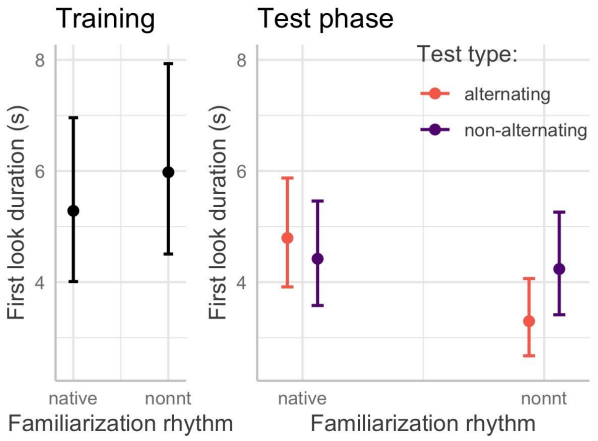


Figure 1. *First look duration* in Training and Test. Estimated means and 95% confidence intervals.

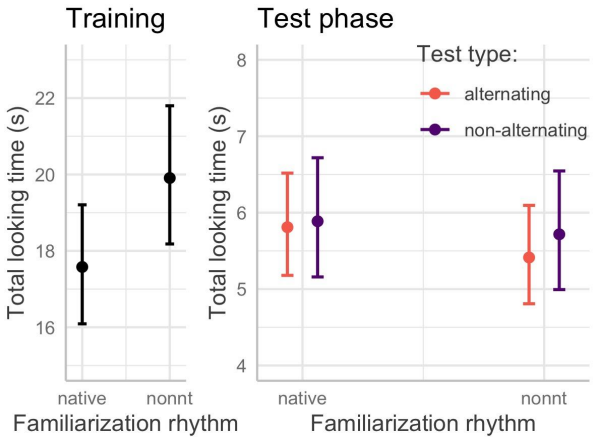


Figure 2. *Total looking time* in Training and Test. Estimated means and 95% confidence intervals.

There was a significant intercept, meaning that the infants' first looks were real, i.e. they were reliably longer than 0. For two other interaction parameters, the absolute *t*-value approached 2, just missing out the commonly used alpha threshold of 0.05. Unpacking the interaction of Familiarization and Test-type (see right panel in Figure 1) revealed that infants trained with non-native rhythm looked longer to non-alternating than to alternating trials at test, by nearly 1 second (mean difference = 0.93 s, *t* = -2.065, *p* = 0.046); no such effect was detected for native-rhythm exposure (mean difference = -0.37, *t* = 0.690, *p* = 0.494).

Pairwise comparisons for the interaction of Familiarization and Test-block showed that infants exposed to native rhythm during familiarisation looked longer to the first block of test trials than to the second block, by about 1.1 second (mean difference = 1.13 s, *t* = 2.350, *p* = 0.024); no such effect was detected for infants exposed to non-native familiarisation (mean difference = -0.2 s, *t* = -0.507, *p* = 0.615).

The model for total looking time in the Test phase yielded a significant intercept, confirming that, overall, total looking time during test trials was greater than 0. No other effects were significant, only one had a potentially meaningful *t*-value, namely, the effect of Test-block (estimated mean effect = -0.040, *SE* = 0.023, *df* = 35.907, *t* = -1.752, *p* = 0.088), indicating that the total looking time tended to be longer in the first than in the second block.

Parameter	Estim.	*SE*	*df*	*t*	*p*
Intercept	1.422	0.064	39.100	22.251	<.001
Familiarization (-nat +nonnt)	-0.110	0.064	39.100	-1.722	.093
Test type (-alt+nonalt)	0.042	0.042	38.844	1.002	.323
Test block (-1st+2nd)	-0.047	0.037	37.786	-1.272	.211
Familiarisation * Test type	0.083	0.042	38.844	1.963	.057
Familiarisation * Test block	0.074	0.037	37.786	2.000	.053
Test type * Test block	-0.022	0.032	39.173	-0.680	.501
Familiarisation *T.type*T.block	-0.007	0.032	39.173	-0.229	.820

Table 1: Fixed-effects model summary for first look duration (log transformed) in the Test phase.

4. DISCUSSION

This experiment tested whether infants' learning of novel segments is modulated by the rhythm of the speech in which the to-be-learnt segments are embedded. Given that 6-month-old Czech-exposed infants have demonstrated listening preferences for non-native over native rhythm [10], we tested whether this would be reflected in selective learning from non-native rhythm.

We exposed 6-month-old monolingual Czech infants to a bimodal distribution representing a novel vowel contrast, an English-like [ɛ]-[æ] distinction. Infants passively listened to 405 vowel tokens sampled from the [ɛ]-[æ] distribution embedded in delexicalized [fɛfæfɛ] utterances with either native Czech rhythm or with atypical, non-native rhythm. Immediately after the ~2-minute training phase, the infants were tested on their discrimination of [ɛ]-[æ] in an alternating/ non-alternating, central fixation, paradigm.

The present results for the training data corroborate previous findings [10] that Czech 6-month-olds preferentially listen to non-native rather than native rhythm. Note however that this effect was found here for total looking time (and not for first look duration) while in our prior study, with the same population and similar stimuli, we found this effect for first look duration. This discrepancy in outcomes for the different measures might be due to the nativeness being a within-subjects factor in the prior work (and thus more robust) and only a between-subjects factor here. It could also be attributed to stimulus properties: the previous study used low-frequency words that had rich segmental structure while the present study used completely delexicalized stimuli with reduced segmental complexity.

The present results for the test phase indicate that infants learn novel segments better from input to which they preferentially attend, even if it is less native-like. Based on the distributional training literature [19], learning effects should be reflected in longer looking to non-alternating than to alternating trials. This is because the preference for non-alternating test trials translates as infants' preference for *novelty at test* as opposed to the – at test already *familiar* – *bimodal alternation* between the two peaks, [ɛ] and [æ], that they were exposed to during training. Our analyses indicate longer first look duration to non-alternating than alternating test trials in infants trained with non-native rhythm. This suggests that infants were able to discover the underlying bimodal distribution better when it was presented in non-native than in native rhythm.

In sum, the present experiment suggests that infants' distributional learning of novel segmental contrasts is modulated by prosody, specifically, the rhythm of the training speech input. When both measures (i.e. first look duration and total looking time) are considered, the selectivity of the learning mechanism matches the infants' preferential listening behaviour. Future research is needed to further test the trends observed here, as well as to tackle the question of which perceptual and learning mechanisms are tapped at by first look versus total looking times, and compare these across paradigms contingent versus non-contingent on infant looking behaviour.

5. ACKNOWLEDGMENT

This work was funded by a Czech Science Foundation grant 21-09797S.

6. REFERENCES

[1] Kinzler, K.D., Dupoux, E. Spelke, E.S., 2007. The native language of social cognition. *Proceedings of the National Academy of Sciences* 104(30), 12577-12580.

[2] Mattock, K., Polka, L., Rvachew, S. Krehm, M., 2010. The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants. *Developmental Science* 13, 229-243.

[3] Byers-Heinlein, K. Fennell, C.T., 2014. Perceptual narrowing in the context of increased variation: Insights from bilingual infants. *Developmental Psychobiology* 56, 274-291.

[4] Kinzler, K.D., Corriveau, K.H. Harris, P.L., 2011. Children's selective trust in native-accented speakers. *Developmental Science* 14, 106-111.

[5] Ramus, F., 2002. Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition* 2, 85-115.

[6] Gasparini, L., Langus, A., Tsuji, S. Boll-Avetisyan, N., 2021. Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition* 213, 104757.

[7] White, L., Floccia, C., Goslin, J. Butler, J., 2014. Utterance-Final Lengthening Is Predictive of Infants' Discrimination of English Accents. *Language Learning* 64, 27-44.

[8] Martinez-Alvarez, A., Benavides-Varela, S., Lapillonne, A. Gervain, J., 2022. Newborns discriminate utterance-level prosodic contours. *Developmental Science*, e13304.

[9] Zacharaki, K. Sebastian-Galles, N., 2021. The ontogeny of early language discrimination: Beyond rhythm. *Cognition* 213, 104628.

[10] Paillereau, N., Podlipský, V.J., Smolík, F., Šimáčková, Š., & Chládková, K. 2021. The development of infants' sensitivity to native versus non-native rhythm. *Infancy*, 26, 423-441.

[11] Tsuji, S. Cristia, A., 2014. Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology* 56, 179-191.

[12] Johnson, E.K. Jusczyk, P.W., 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* 44, 548-567.

[13] Mandel, D.R., Jusczyk, P.W. Nelson, D.G.K., 1994. Does sentential prosody help infants organize and remember speech information? *Cognition* 53, 155-180.

[14] Gleitman, L., Wanner, E. 1982. The state of the state of the art. In Wanner E., Gleitman, L. (Eds.), *Language acquisition: the state of the art* (pp. 3-48). Cambridge, UK: CUP.

[15] Ladányi, E., Lukács, Á. Gervain, J., 2021. Does rhythmic priming improve grammatical processing in Hungarian-speaking children with and without developmental language disorder?. *Developmental Science* 24, e13112.

[16] ManyBabies Consortium, 2020. Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science* 3, 24-52.

[17] Adriaans, F. Swingley, D., 2017. Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The Journal of the Acoustical Society of America* 141, 3070-3078.

[18] Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., Fusaroli, R., 2022. A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, 1-20.

[19] Maye, J., Werker, J.F., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101-B111.

[20] Wanrooij, K., Boersma, P., Van Zuijen, T.L., 2014. Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in Psychology* 5, 77.

[21] Schatz, T., Feldman, N.H., Goldwater, S., Cao, X.N. Dupoux, E. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences* 118, e2001844118.

[22] Wanrooij, K., Boersma, P., 2013. Distributional training of speech sounds can be done with continuous distributions. *The Journal of the Acoustical Society of America* 133, EL398-EL404.

[23] Kominsky, J.F. 2019. PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behavior and Development* 54, 114–119.

[24] Lewis, M., Kagan, J., Kalafat, J. 1966. Patterns of fixation in the young infant. *Child Development* 37, 331– 341.

[25] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1–48.

[26] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 1–26.

[27] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

VOWEL LENGTH IN INFANT-DIRECTED SPEECH: THE REALISATION OF SHORT-LONG CONTRASTS IN CZECH IDS

Michaela Svoboda[a,b], Kateřina Chládková[a,b], Tanja Kocjančič Antolík[b,c,d], Nikola Paillereau[b,c], Petra Slížková[b]

[a]Institute of Czech Language and Theory of Communication, Faculty of Arts, Charles University; [b]Institute of Psychology, Czech Academy of Sciences; [c]Institute of Phonetics, Charles University; [d]Faculty of Education, University of Ljubljana

michaela.svoboda@post.cz

ABSTRACT

When interacting with young children, talkers across many languages use a speech style that reflects positive affect, draws infants' attention, and supposedly facilitates language acquisition. As for the latter, a well-documented feature of infant-directed speech is an exaggeration of spectrally-cued vowel contrasts. Here we tested whether talkers exaggerate also durationally cued contrasts.

Sixty-three mothers, native speakers of Czech, were recorded while playing with their infant (4- to 10-month-olds, IDS) and while speaking to an adult (ADS). The durations of the five Czech phonemically short vowels were compared to their long counterparts. Vowel duration (normalised for word duration) was longer in IDS than in ADS more for phonemically long vowels at the younger infant ages, indicating a developmentally specific early exaggeration of length contrasts in Czech infant-directed speech. The present finding suggests that in a language with phonemic length, caregivers' realisation of speech sounds may go beyond merely being longer and slower overall.

Keywords: infant-directed speech, vowel length, development of early input, Czech

1. INTRODUCTION

This paper investigates the realisation of phonemic short-long contrasts in Czech-speaking mother's infant-directed speech (IDS). Across many cultures and languages, adults use a distinct speech style when communicating with infants and young children. Recently conducted a meta-analysis reflecting results of 87 unique IDS studies on more than 20 different languages or language varieties [1] demonstrated that IDS often differs from adult-directed speech (ADS) in several aspects, including acoustic parameters such as mean F0 (which tends to be higher in IDS than in ADS), F0 range or standard deviation (larger in IDS than in ADS), speech tempo (slower in IDS), the acoustic duration of vowels (longer in IDS), and the size of the vowel space (larger in IDS). Some of these features occur already in the speech addressed to unborn infants [2]. Although, in general, vowel spaces tend to be exaggerated in IDS compared to ADS, the literature also reports different tendencies for some languages. For instance, [2] did not find enlarged vowel spaces in Dutch mothers' IDS but reported vowel fronting which she attributed to the mothers' positive affect, i.e. smiling, while speaking to their child.

Researchers have argued that the exaggeration of the vowel space facilitates young children's acquisition of native phoneme contrasts (*the hyperarticulation hypothesis*, see [4, 5]). Whereas the literature provides much data on vowel spectral properties and enlargement (or absence or even shrinkage [3, 7]) of vowel quality contrasts [1], less is known about potential exaggeration of length contrasts. If acoustic enhancement occurs to aid children in learning native contrasts, one would expect to find the exaggeration also for vowel length distinctions in languages whose phonology contains phonemic length. However, data from languages with phonemic length are mixed. Some studies report exaggerated duration differences between short and long vowels in IDS (e.g. [6] for Swedish) while others only report longer duration overall indexing the slower speech tempo typical of IDS [e.g. 7 for Norwegian]. In Japanese IDS, an exaggerated length difference was found to be context-specific, reportedly occurring only in word-final positions [8]. Interestingly, Japanese infants seem to acquire context-independent native length contrasts relatively late (at about 9.5 months of age [9]), which one could potentially attribute to insufficient exposure to the long member of length contrasts in the input [10] or even to the non-exaggerated nature of short-long distinctions (outside the word-final position).

Here we assess the realisation of short-long vowel contrasts in Czech IDS. Infants acquiring Czech are perceptually sensitive to vowel duration differences from an early age on: already at birth

their brains respond to a native vowel length contrast as strongly as to a native spectral contrast [11] and throughout the infants' first year of life, phonological length differences remain behaviourally discriminable at least as robustly as spectral contrasts [12]. Based on these developmental patterns, we hypothesize that the Czech infants' input facilitates the learning of vowel length, likely through the exaggeration of short-long differences in IDS. On the other hand, infants acquiring other languages with phonemic length master the native length contrasts even though studies have not detected exaggeration of length contrasts in their IDS [13]. An alternative hypothesis therefore is that Czech IDS will be slower overall, with similarly prolonged durations of both short and long vowels, and without an exaggeration of the duration-cued differences.

Regarding the developmental trajectory of infants' early input, one of the findings of [1] most relevant to our research question is that vowel duration in IDS appears to change over time: the older the infant gets, the shorter the vowels tend to be. It is worth pursuing this phenomenon further, considering the developmental changes associated with infants' perceptual acquisition of native vowel contrasts. As noted above, Czech infants' sensitivity to length contrasts seems to develop rather early, at or even before the 4th month of age. One could hypothesise that if durational contrasts are exaggerated in developmentally sensitive ways, larger exaggeration would be found for younger than for older infants. To address this, we investigate IDS spoken to infants across 4 to 10 months of age.

The phonological system of Czech contains ten monophthongal vowels, five short and five long phonemes. While duration is the primary cue for most Czech short-long contrasts, the high-front vowel contrast /iː/-/ɪ/ is cued also by spectral properties [14–16]. In some dialects, the spectral cue can even outweigh duration for /iː/-/ɪ/, and partially cues also the /uː/-/u/ contrast [15, 17]. We therefore test whether the durational contrast in Czech IDS is realised differently in the high front vowels than in the other vowel pairs.

To sum up, we investigate whether (1) the durational difference cueing vowel length contrasts is exaggerated in Czech IDS, whether (2) the exaggeration depends on infant age, and whether it is (3) specific to vowel pairs for which vowel duration is the primary cue in ADS.

2. METHOD

2.1. Participants

Sixty-three mother-infant dyads participated in the experiment. The women were native speakers of Czech who had lived in the metropolitan Prague area for at least 5 years prior to the experiment. They were between 24 and 39 years old and had no speech or hearing disorders. Their children participating in the experiment were aged 4 (n = 18), 6 (n = 13), 8 (n = 19), or 10 (n = 13) months (±2 weeks).

2.2. Materials and procedure

Each of the 10 Czech monophthongs /ɪ iː ɛ ɛː a aː o oː u uː/ was elicited in two words, altogether represented by a set of 20 objects which participants spontaneously commented on (following [16]). The target vowels always occurred in a word-initial syllable (which is, formally, stressed); the flanking consonants were bilabials and alveolars. Participants were instructed to name each object at least twice while talking about it. Participants talked about the set of objects twice, once to their infant (only the parent and the infant present in the recording booth) and once to an adult experimenter (with only the parent and the experimenter present), with order of the conditions counterbalanced. The recordings were done in a sound-treated booth using a condenser microphone AKG C520 and an Edirol UA 25 sound card, with Audacity run on a PC (44.1-kHz sampling frequency and 16-bit quantisation).

2.3. Acoustic analysis

Duration was measured over segmented word and vowel tokens. Only disyllabic words, and their corresponding first-syllable vowels, were included in the analysis. All tokens were manually segmented using Praat [18], based on visual inspection of the waveform and spectrogram. Word onset and offset corresponded to the onset of the first and the offset of the last segment. The target vowel interval had to include visible formants, particularly F2. All boundaries were aligned to zero crossing in the waveform. The final data set included 10249 target vowel tokens, 5782 of them in the IDS condition and 4467 in ADS roughly equally distributed across vowel categories.

2.4. Statistical analysis

Vowel durations were normalised for word duration, by dividing the duration of each vowel in milliseconds by the duration of the word in which it was embedded. The normalised vowel durations were submitted to a linear mixed-effects model using the lme4 and lmerTest packages in R [19–21]. The modelled fixed effects were Style (ADS vs. IDS, coded as -1 vs. +1), Length (short vs long, coded as -1 vs +1), Vowel pair (with 4 sum-to-zero contrasts comparing the /ɪ/-/iː/ pair to each of the remaining 4 pairs), and Age (in months, mean-centered). The models contained a full random-effects structure with main and interaction effects of all the three within-subjects factors.

3. RESULTS

Table 1 lists the outcomes for the interaction effects involving Style and Length, since those can answer our research questions, as well as the simple main effects involved in those interactions.

Fixed effects	estimate	SE	t
Intercept	**0.193**	**0.0018**	**109.22**
Style (-ADS +IDS)	**0.009**	**0.0007**	**12.43**
Length (-lo + sh)	**0.054**	**0.0007**	**76.04**
Age	**-0.004**	**0.0024**	**-2.25**
Style*Length	**0.004**	**0.0070**	**5.38**
Style*Length*Age	**-0.002**	**0.0010**	**-2.04**
Sty*Len*V.pair(-a +i)	0.002	0.0014	1.35
Sty*Len*V.pair(-e +i)	-0.0005	0.0015	-0.31
Sty*Len*V.pair(-o+i)	-0.002	0.0014	-1.41
Sty*Len*V.pair(-u+i)	-0.0006	0.0015	-0.41

Table 1: Model outcome of selected fixed effects, namely, of interactions involving the factors Style and Length, and the related simple effects. DFs for Intercept and Age = 63, other DFs ~ 10200. Effects yielding $p < 0.05$ are in bold.

The parameters relevant to our research question about realisation of the Czech length contrast in IDS are those that involve the interaction of Style and Length. The significant two way-interaction of Style and Length suggests that the length contrast is realised differently in IDS than in ADS: pairwise comparisons show that long vowels differ between IDS and ADS more than short vowels (see Fig. 1, left panel). This two-way interaction is further licenced by the significant three-way interaction additionally involving Age: as shown in Figure 1 (right panel), the greater lengthening of long vowels in IDS as compared to short vowels decreases with infant age.

None of the interactions of Style and Length with Vowel pair turned out significant. We thus failed to find evidence that the differential cue-weighting of duration for the high-front vowel pair documented for Czech adult speech would affect length contrasts in Czech IDS. Given this null result for the interaction of Style, Length, and Vowel pair we will not make any conclusions about vowel-quality-specific realisation of length in Czech IDS.

4. DISCUSSION

In this paper, we investigated whether in Czech, a quantity language, phonemic length contrasts are exaggerated in infant-directed speech. To this end, we measured the durations of phonemic short and long vowels in the speech of Czech mothers addressed to their 4- to 10-month-old infant and to an adult. The analyses revealed that not only are vowels longer in Czech IDS than in ADS in general (main effect of Style) but that the lengthening is larger for long vowel phonemes than for short phonemes (interaction of Style and Length) and increases with decreasing age of the infant who is being spoken to (three-way interaction of Style, Length, and Age). To our knowledge, such age-specific enhancement of length contrasts in IDS has not been widely reported for quantity languages, very probably because length contrasts as such have been investigated only in a few IDS studies to date [6, 8].

The present findings for Czech suggesting the exaggeration of length contrast in speech to very young infants (and less so to older ones) are partly in line with the results for exaggeration of vowel length distinctions reported earlier for Swedish IDS [6]. The developmentally conditioned early exaggeration of length distinctions in Czech, could have a facilitative function, aiding infants to acquire the contrast between short and long vowel phonemes. After all, infants acquiring Czech seem to be perceptually sensitive to native length contrasts at birth as well as at 4 months [11, 12], unlike for instance infants acquiring Japanese, who distinguish native vowel length contrasts

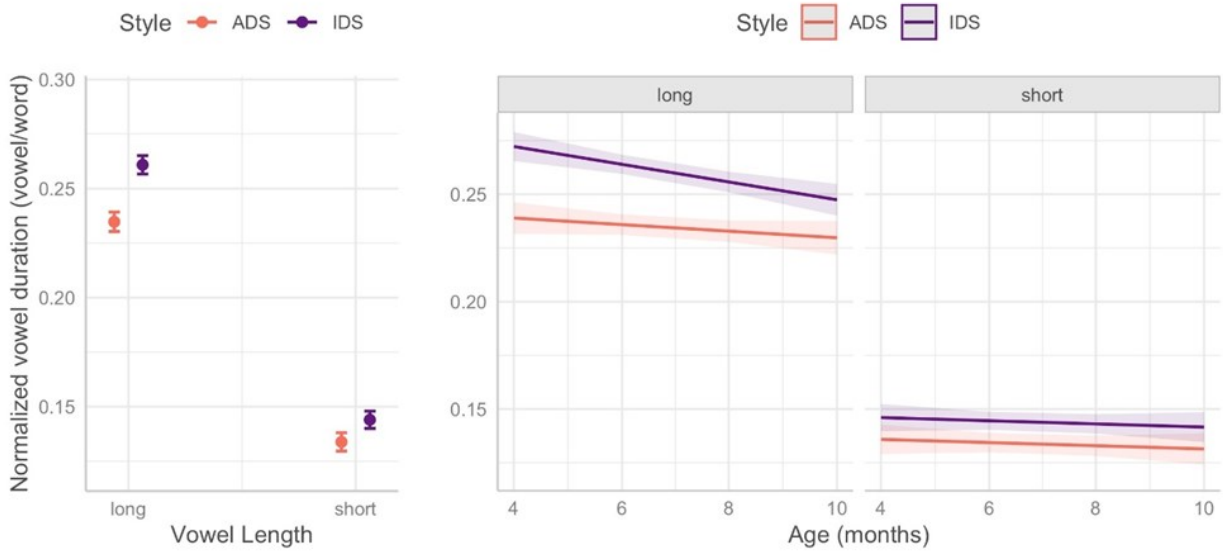


Figure 1: The estimated mean normalised vowel duration and its 95% confidence intervals for the two way-interaction of Style and Length (left) and for the three-way interaction of Style, Length, and Age (right).

much later, at 9.5 months [9].

A plausible explanation for this between-language difference in infants' perceptual abilities could lie in the input, where contrast exaggeration promotes early vowel length acquisition in Czech and low frequency of occurrence of long vowels postpones vowel length acquisition in Japanese [10]. It would be worth to further compare the realisation of length contrasts in IDS across various languages in which vowel length is phonemic, such as Slovak, Hungarian, Finnish, Arabic or Estonian, and monitor its relationship to the perceptual development of infants acquiring those languages.

Contrary to our predictions, according to which we expected the length contrast to be realised differently in IDS for the high-front vowel pair than for other pairs, we did not detect any such effects. This could mean that Czech speakers realise the length contrast in high front vowels similarly across ADS and IDS. But note that we did not analyse the participant´s native dialects that might affect the realisation of this length contrast. In the present study, the mothers came from various parts of Czechia, the only criterion for inclusion in the study being that they had lived in the Prague area at least 5 years prior to the recording. At least for some vowel pairs, the weighting of vowel duration as a cue to vowel length contrasts varies between the western and eastern Czech dialects [14]. One could thus speculate that the realisation of durationally-cued contrasts may differ also in IDS across those dialects. In future work on IDS, dialectal variation should be taken into account.

The literature reports on cross-linguistically observed changes in IDS due to infant age, such as the overall shortening of vowel durations with increasing infant age, thus contributing to progressively faster speech tempo [1]. It remains to be investigated whether developmental changes in IDS also occur for spectral properties of vowels such as diphthongisation, for vowel space size, or for consonants, and to what extent they depend on the language, dialect, and infant age at hand.

5. CONCLUSION

When talking to their infants, Czech-speaking mothers not only produce longer vowels overall but also exaggerate the durational differences between phonologically short and phonologically long Czech vowels, as compared to when speaking to an adult. This exaggeration of length contrasts seems to be larger for speech addressed to younger infants than for speech addressed to older infants. Future research is needed to better understand how infant age or mother's native language variety modulate IDS.

6. ACKNOWLEDGMENT

This work was funded by Charles University grant Primus/17/HUM/19 and Czech Science Foundation grant 21-09797S. We are grateful to the participating families and also to Kristýna Hrdličková, Veronika Ungrová, and Zuzana Oceláková for help with data collection and annotation.

7. REFERENCES

[1] Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., Fusaroli, R. 2022. A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, 1-20.

[2] Chládková, K., Černá, M., Paillereau, N., Skarnitzl, R. and Oceláková, Z. 2019. Prenatal infant-directed speech: vowels and voice quality. *Proceedings of ICPhS 2019*, pp.1525-1529.

[3] Benders, T. 2013. Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, 36(4), 847-862.

[4] De Boer, B., Kuhl, P. K. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4, 129

[5] Cristia, A. 2013. Input to Language: The Phonetics and Perception of Infant-Directed Speech. *Language and Linguistics Compass* 7(3), 157–170

[6] Sundberg, U. 1998. Mother tongue - Phonetic Aspects of Infant-Directed Speech (PhD dissertation, Department of Linguistics, University of Stockholm).

[7] Englund, K., Behne, D. 2006. Changes in infant directed speech in the first six months. *Infant and Child Development: An International Journal of Research and Practice,* 15(2), 139-160.

[8] Tajima, K., Tanaka, K., Martin, A., Mazuka, R. 2013. Is the vowel length contrast in Japanese exaggerated in infant-directed speech? *Proceedings of the Annual Conference of the International Speech Communication Association,* INTERSPEECH, 3211-321.

[9] Mugitani, R., Pons, F., Werker, J.F., Amano, S. 2009. Perception of vowel length by Japanese and English-learning infants. *Developmental Psychology,* 45, 236– 247.

[10] Bion, R. A., Miyazawa, K., Kikuchi, H., Mazuka, R. 2013. Learning Phonemic Vowel Length from Naturalistic Recordings of Japanese Infant-Directed Speech. *PLOS ONE* 8(2): e51594.

[11] Chládková, K., Urbanec, J., Skálová, S., Kremláček, J. 2021. Newborns' neural processing of native vowels reveals directional asymmetries. *Developmental Cognitive Neuroscience,* 52, 101023.

[12] Paillereau, N., Podlipský, V. J., Šimáčková, Š., Smolík, F., Oceláková, Z., Chládková, K. 2021. Perceptual sensitivity to vowel quality and vowel length in the first year of life. *JASA Express Letters*, 1(2), 025202.

[13] Bernstein Ratner N, Luberoff A. 1984. Cues to post-vocalic voicing in mother–child speech. *Journal of Phonetics,* 12, 285–289.

[14] Šimáčková, Š, Podlipský, V., Chládková, K. 2012. Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association,* 42(2), 225-232.

[15] Podlipský, V. J., Chládková, K., Šimáčková, Š. 2019. Spectrum as a perceptual cue to vowel length in Czech, a quantity language. *The Journal of the Acoustical Society of America,* 146(4), EL352.

[16] Paillereau, N. & Chládková, K. (2019). Spectral and temporal characteristics of Czech vowels in spontaneous speech. AUC PHILOLOGICA. 2019. 10.14712/24646830.2019.19.

[17] Bořil T., Veroňková J. 2020. Perceived Length of Czech High Vowels in Relation to Formant Frequencies Evaluated by Automatic Speech Recognition. In: Sojka P., Kopeček I., Pala K., Horák A. (eds), *Text, Speech, and Dialogue. Lecture Notes in Computer Science,* 12284. Springer, Cham, 409-417

[18] Boersma, P., Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.0.25). http://www.praat.org/

[19] Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[20] Kuznetsova A, Brockhoff PB, Christensen RHB. 2017. lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software,* 82(13), 1–26.

[21] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Influence of bilingualism or caregiver input? Variation in VOT in simultaneous bilingual preschoolers in Singapore

Jasper H. Sim

National Institute of Education, Nanyang Technological University, Singapore
jasper.sim@nie.edu.sg

ABSTRACT

This exploratory study examines variation in voice onset time (VOT) in 12 English-Mandarin and English-Malay simultaneous bilingual preschoolers born and raised in Singapore. These children were not only different in their second L1, but they and their caregivers were also different in their language dominance, as are typical of speakers in such sociolinguistic contexts. This study seeks to explore effects of bilingualism and caregiver input on VOT development. The findings revealed differences in English VOT between Chinese and Malay children, and between Malay children—all of whom highly English dominant—that reflect the same variation previously found in the child-directed speech of Singaporean caregivers. The results also revealed category assimilation of Malay voiceless unaspirated and English fortis stops in some Malay children that could be attributed to the phonetic overlap in their input. Some unexplained individual variation and differences further highlight the complexity of bilingual phonological acquisition in such contexts.

Keywords: quality of input, language contact, multi-dialectal, English-Malay, English-Mandarin.

1. INTRODUCTION

In many bilingual contexts, the phonetic input that children receive is not homogeneous. Caregivers who are late-L2 speakers may provide phonetic input that is inconsistent, and these caregivers, as well as those experiencing L1 attrition, may also exhibit phonetic details that differ from their monolingual peers in their child-directed speech (CDS) [1]–[3].

Not all differential production is due to acquiring an L2 late. In multicultural and multilingual contexts, between-speaker variation can arise from, *inter alia*, caregivers speaking a different other L1 or heritage language; even those speaking the same languages may differ in their language dominance (e.g. in terms of amount of use of their 2L1s), thereby exhibiting accent differences [4], [5]. Moreover, in communities that have experienced long-term language contact and/or shift, features that had once emerged from effects of individual bilingualism may be transmitted to and retained by later generations of L1 speakers [5]–[7]. In other words, what appear as effects of bilingualism may in fact be learnt through the input.

Children are sensitive to sub-phonemic details in the input, and its fine-grained variation can moderate language outcomes [8], [9]. Stoehr and colleagues [2] examined the production of VOT by Dutch-German preschoolers who acquired the heritage language German from their mothers who spoke L1 German, and the majority language, Dutch, from L1-speaking fathers and L2-speaking mothers. They found that individual variation in the children's VOT was associated with the variation in both their mother's non-native L2 and attrited L1. Sim and Post [9] also found that Singaporean mothers who released English coda stops to a lesser degree also had children who tended to not release their stops, and the same was true for mothers who released their stops to a higher degree. Importantly, these studies show that variable language outcomes in bilinguals may be attributed to input properties and not (solely) due to effects of bilingualism such as cross-linguistic influence (CLI) that involves assimilation of or interactions between the two phonological systems [10]–[13].

This present study explores the variation in the implementation of stop voicing contrast in English-Mandarin and English-Malay simultaneous bilingual preschoolers. Singaporean Mandarin employs a two-way distinction between unaspirated and aspirated /p, t, k/ stops, and Malay employs a two-way distinction between truly voiced /b, d, g/ stops and unaspirated voiceless /p, t, k/ stops. In a recent work [14], it was revealed that while Singaporean Malay and Chinese early bilingual mothers were similar in their use of VOT for stop voicing contrast in their English CDS towards preschoolers (i.e. longer positive VOT for fortis, and shorter VOT/lead VOT for lenis stops), they were dissimilar in where in the VOT continuum the contrasts were made: Malay mothers employed shorter positive VOT but longer lead VOT, whereas Chinese mothers produced longer positive VOT but shorter lead VOT. Mothers who were less English dominant produced even shorter VOT in their fortis stops, thereby exhibiting smaller phonetic contrasts. The findings suggest that Malay children, especially those raised by Malay-dominant caregivers, are exposed to English stops that phonetically overlap with the same stops in Malay to a large degree.

This study builds on this work by exploring the effects of individual bilingualism and caregiver input on VOT development. It aims to examine (i) whether and how English VOT patterns between Singaporean Malay and Chinese children who are equally highly English dominant are different, and (ii) whether and how English and Malay VOT patterns are different between Malay-dominant and English-dominant Malay children.

2. METHODOLOGY

2.1. Participants

Participants (10 boys, 2 girls) were nine Malay and three Chinese Singaporean preschoolers between the ages of 44 and 76 months (M_{age} = 60.8) who were raised by early bilingual caregivers. The children were exposed to local varieties of English and Malay or Mandarin (for the Chinese) by the age of three. The language dominance of their mothers was measured using the Bilingual Language Profile (BLP) [15]; possible scores ranged from –218 (Malay or Mandarin dominant) to +218 (English dominant). The amount of use of English by the children, the proxy for language dominance for this study, was calculated from an accumulated measurement of the estimated amount and proportion of time for which their 2L1s were used with significant people in their immediate ecosystem, in self-interaction and in their exposure to media, as reported by their caregivers.

The three Chinese children (**CHI**) were all highly English dominant ($M_{\%Eng\ use}$ = 88, *SD* = 5.8), and were raised by highly English-dominant mothers (M_{BLP} = 119.7, *SD* = 50.1). The Malay children were divided into three groups: six Malay children were highly English dominant ($M_{\%Eng\ use}$ = 80.3, *SD* = 7.5), three of whom (**MLY-L.ENG**) were raised by mothers who were *less* English dominant (M_{BLP} = 47.1, *SD* = 2.0), while three others (**MLY-M.ENG**) were raised by mothers who were *more* English dominant (M_{BLP} = 76.0, *SD* = 22.3). The final three Malay children (**MLY-DOM**) were more Malay dominant ($M_{\%Eng\ use}$ = 54.7, *SD* = 7.1) and were raised by Malay-dominant mothers (M_{BLP} = -9.25, *SD* = 21.1).

2.2. Materials and recording procedure

Target words with close/high vowels (Table 1) were elicited through a picture card naming task primarily conducted by one of the caregivers (typically the mother). Each word was elicited twice. The Malay target words were only elicited from the Malay children, after a short interaction in Malay between caregiver and child.

The recording took place in a quiet room in the respective homes of the participants. The child's responses were recorded using a NAGRA ARES-MII recorder, through a lapel microphone that was pinned to the collar of the child. Back-up recording was also performed by the author using a Zoom H5 recorder. Both recorded at a sampling rate of 44.1 kHz at 16 bit.

Stop	English	Malay
p	pea (swimming) pool	pipi *cheek* putih *white*
t	tea, 't' two	tiga *three* tujuh *seven*
k	key cooking	kitab *book* kuda *horse*
b	bee (peek-a-)boo	bibir *lips* buka *open*
d	'd' doing	diri *stand* duduk *sit*
g	geese good	gigi *teeth* gula-gula *candies*

Table 1: Stimuli

2.3. Acoustic analysis

VOT boundaries for all word-initial stops were placed manually based on acoustic cues in the waveforms spectrograms on Praat [16]. VOT was defined as the time interval between the release burst as signalled by a sharp peak in waveform energy and the onset of voicing, taken to be the nearest zero-crossing of the onset of periodicity in the waveform. If voicing occurred during the closure, VOT was measured from the onset of glottal pulses in the waveform and/or visual cues in the spectrogram that indicated voicing onset up to the first release burst. 44 tokens could not be reliably measured due to noise and overlapping speech and were excluded. Faster speaking rate is associated with shorter VOT; vowel duration was used in this study as a proxy for local speech rate [17].

2.4. Statistical analysis

Stops with positive VOT (+VOT) were analysed separately from prevoiced stops (–VOT) because of the bimodal distribution. Linear regression analyses were conducted using R. Model selection was based on parsimony. In all models, the random effect structure included random intercepts for subject and word only if they significantly improved model fit. To evaluate the contribution of each predictor, pairwise model comparisons between a full model and a more restricted model that excluded the predictor under consideration were performed using likelihood ratio tests. Continuous predictors were z-standardised. Categorical predictors were weighted effect coded. Outliers in the raw measurements were detected using the interquartile range method. After checking the normality of residuals, influential outliers were

removed using Cook's distance, before the final, optimised models were obtained and reported below.

3. RESULTS

The VOT means of English fortis and lenis stops (with +VOT and –VOT) and Malay voiceless (VL) and voiced (V) stops are presented in Table 2.

Stops	Chi	Mly-dom	Mly-L.Eng	Mly-M.Eng
Eng Fortis *SD (n)*	**120.6** 63.3 (41)	**62.1** 34.2 (44)	**48.8** 50.0 (41)	**46.8** 36.5 (46)
Eng Lenis *(+) SD (n)*	**27.1** 24.1 (19)	**35.7** 15.2 (19)	**27.4** 17.0 (21)	**30.4** 15.6 (10)
Eng Lenis *(–) SD (n)*	**–193.5** 97.4 (16)	**–133.5** 41.9 (16)	**–79.2** 64.0 (11)	**–104.5** 41.5 (23)
Mly VL *SD (n)*	—	**38.6** 28.2 (38)	**31.2** 18.8 (29)	**36.3** 24.3 (37)
Mly V *SD (n)*	—	**–102.9** 22.6 (11)	**–71.1** 45.8 (9)	**–80.5** 37.2 (22)

Table 2: VOT means (ms) of Eng. and Mly stops.

3.1. Stops with +VOT

Figure 1 shows the distribution of +VOT of English fortis/lenis and Malay voiceless stops for each child.

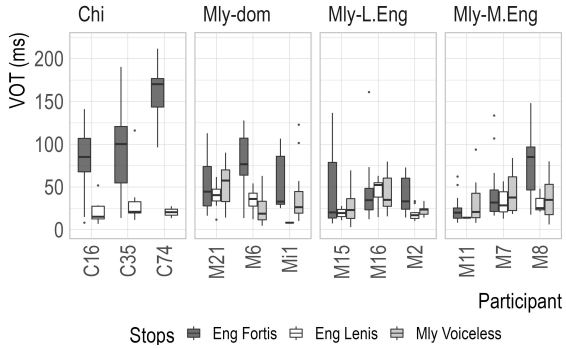


Figure 1: +VOT (ms) of English fortis/lenis and Malay voiceless stops of individual child by group.

3.1.1 Phonemic contrast of English stops with +VOT

The +VOT of English fortis and lenis stops produced by all children were first analysed. The full model included as fixed effects (first in list = reference level for all models henceforth) vowel duration, place of articulation (POA; velar, alveolar, bilabial), group (MLY-M.ENG, CHI, MLY-DOM, MLY-L.ENG), and type (fortis, lenis), and the two-way interaction between group and type. In the reduced model (*obs.* = 223, influential outliers removed = 18, marginal R^2 = .62, conditional R^2 = .75), all were significant predictors. Only for fortis stops, Chinese children produced longer VOT than Malay-dominant Malays, b = 69.64, t(9.05) = 6.32, p < .001, and even more so than English-dominant Malays (MLY-L.ENG: b = 87.32, t(9.33) = 7.86, p < .001; MLY-M.ENG: b = 84.69, t(8.99) = 7.70, p < .001). Marginal differences between the Malay groups were not significant.

In addition, Chinese children exhibited phonemic contrast between fortis and lenis stops by producing significantly longer +VOT for fortis stops than lenis, b = 97.6, t(43.4) = 11.81, p < .001. The contrast was smaller in MLY-DOM Malays, b = 27.7, t(37.8) = 3.48, p = .001, marginally in MLY-M.ENG Malays, b = 18.7, t(64.1) = 2.04, p = .04, and was not significant in MLY-L.ENG Malays, b = 11.7, t(36.6) = 1.47, p = .15.

In sum, after controlling for vowel duration and POA, Chinese children produced much longer +VOT for English /p, t, k/ than all Malays, thereby exhibiting greatest phonetic contrast between English fortis and lenis stops. Compared to English-dominant Malays, MLY-DOM Malays produced greater fortis-lenis contrasts; the fortis-lenis contrasts produced by MLY-L.ENG Malays were marginal and not significant.

3.1.2 English and Malay VL stops with +VOT

Malay children's English fortis and lenis stops with +VOT and their Malay /p, t, k/ stops were compared. Malay voiceless stops are normally unaspirated, as are English lenis stops. The full model included vowel duration, POA, group, type (Eng. fortis, Malay voiceless, Eng. lenis) and the two-way interaction between group and type. In the reduced model (*obs.* = 268, influential outliers removed = 17, marginal R^2 = .41, conditional R^2 = .54), the main effects of vowel duration, POA and type were significant predictors. Compared to the weighted grand mean, only English lenis stops had significantly shorter +VOT (β = -0.50, b = -13.26, t = -3.47, p < .001). The +VOT of English fortis stops was not significantly longer than Malay voiceless unaspirated stops (b = 3.48, t = 0.70, p = .77).

In sum, after controlling for vowel duration and POA, their Malay voiceless unaspirated stops were more similar to English fortis stops in +VOT than to their English lenis stops. That there was no significant interaction between type and group suggests that there were no significant differences between groups.

3.2. Stops with –VOT

3.2.1 Likelihood of prevoicing

English lenis and Malay voiced stops were more likely short-lag than prevoiced. The Chinese children prevoiced 16/35 (46%) English lenis stops, 16/35 (46%) by MLY-DOM, 11/32 (34%) by MLY-L.ENG, but 23/33 (70%) by MLY-M.ENG. For Malay voiced stops, MLY-DOM prevoiced 11/31 (35%) stops, 9/31 (29%) by MLY-L.ENG, and 22/33 (67%) by MLY-M.ENG. As

can be seen in Figure 2 below, some children did not (consistently) produce prevoiced stops.

3.2.2 Length of –VOT

Figure 2 shows the distribution of –VOT of English lenis and Malay voiced stops for each child.

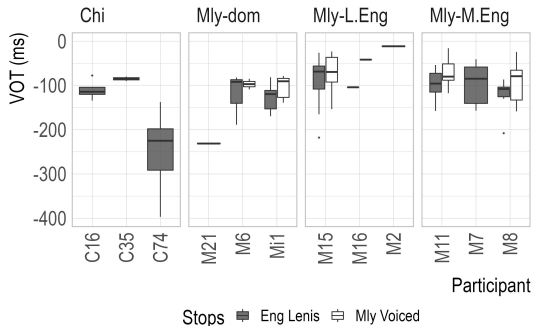


Figure 2: –VOT (ms) of English lenis and Malay voiced stops of individual child by group.

The regression analysis for –VOT length included only data of Malay children who produced at least five prevoiced stops: M6, Mi1, M15, M11 and M8. In the full model that included vowel duration and the two-way interaction between language and group as predictors (*obs.* = 81, R^2 = .18), only vowel duration significantly improved model fit. In other words, there was no significant difference between English lenis and Malay voiced stops in their –VOT for the five children producing prevoiced stops.

4. DISCUSSION

Variation in VOT in 12 English-Mandarin and English-Malay simultaneous bilingual preschoolers raised in Singapore was examined. They were not only different in their second L1, but they and their caregivers were also different in their language dominance, as is typical of speakers in such sociolinguistic contexts. This is an exploratory study; the findings should be interpreted with caution due to the small sample size and number of words per child that were examined.

The analysis on the +VOT of English stops of all children revealed that, for fortis stops only, Chinese children produced longer +VOT than all the Malays, thereby also exhibiting the largest phonetic contrast between English fortis and lenis stops with +VOT. Amongst the Malay children, the fortis-lenis contrast was greater for MLY-DOM children than for English-dominant MLY-M.ENG Malays. The contrast was marginal and not significant for MLY-L.ENG children.

Effects of bilingualism such as CLI and language dominance bear less explanatory power for these findings, as we should expect small differences between CHI, MLY-L.ENG, and MLY-H.ENG children, who all used English highly frequently, on an average of 83% of the time [13]. These differences instead suggest an influence of caregiver input, as they reflect the same variation found in the VOT of Singaporean caregivers [14]: as mentioned, in their CDS, Chinese mothers produced fortis stops with longer +VOT than all Malay mothers (as did CHI children). Compared to their English-dominant peers, Malay-dominant mothers produced shorter +VOT in their fortis stops, thereby exhibiting smaller contrasts (as did MLY-L.ENG children). The results, along with those of previous work, e.g. [2], [9], [18], indicate that an explanatory model of phonological acquisition needs to consider qualitative properties of caregiver input.

That MLY-DOM children produced greater fortis-lenis phonetic contrasts than their English-dominant peers was unexpected, since previous studies showed that Malay-dominant educated Singaporean adults had much shorter +VOT for English /p, t, k/ than their English-dominant peers, due to potential CLI or input of caregiver who were L2 English speakers [5], [14]. Interestingly, the mothers of MLY-DOM children were all Malay language teachers. Whether the greater contrast between English fortis and lenis stops in these children was due to the influence of caregiver input (i.e. their mothers also exhibiting greater English contrasts) or due to richer Malay input (and therefore enhanced phonemic category dissimilation between their 2L1s [19]) remains to be studied.

In the analysis that compared Malay children's Malay voiceless unaspirated stops with their English stops produced with +VOT, it was revealed that, across groups, the VOT of Malay voiceless unaspirated stops was not significantly different from the VOT of English fortis stops. Instead of being a result of CLI, since this was observed across groups, it could be attributed to the considerable phonetic overlap in the input between Malay voiceless unaspirated stops and English fortis stops in the short-lag region that could have led to category assimilation [11], [19]. Note, however, that based on visual inspection of the raw VOT in Figure 1, some children did distinguish these stops by producing longer English fortis stops, although such individual variation was averaged within groups.

Finally, it remains unclear if the lack of consistent prevoicing of Malay stops, if produced at all, is due to developmental reasons [13] or a result of mothers' reduced prevoicing [7]. One question that cannot be answered in this study is why MLY-M.ENG children produced more prevoiced English and Malay stops than others overall, when it is the MLY-DOM children who should presumably have been exposed to more prevoiced stops in their input by their Malay-dominant mothers.

5. REFERENCES

[1] M. S. Fish, A. García-Sierra, N. Ramírez-Esparza, and P. K. Kuhl, "Infant-directed speech in English and Spanish: Assessments of monolingual and bilingual caregiver VOT," *Journal of Phonetics*, vol. 63, pp. 19–34, Jul. 2017, doi: 10.1016/j.wocn.2017.04.003.

[2] A. Stoehr, T. Benders, J. G. van Hell, and P. Fikkert, "Bilingual preschoolers' speech is associated with non-native maternal language input," *Language Learning and Development*, vol. 15, no. 1, pp. 75–100, 2019, doi: 10.1080/15475441.2018.1533473.

[3] G. Khattab, "Acquisition of Lebanese Arabic and Yorkshire English /l/ by bilingual and monolingual children: A comparative spectrographic study," in *Current Issues in Linguistic Theory*, vol. 319, Z. M. Hassan and B. Heselwood, Eds. Amsterdam: John Benjamins Publishing Company, 2011, pp. 325–354. doi: 10.1075/cilt.319.15kha.

[4] M. Amengual, "Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals," *International Journal of Bilingualism*, vol. 23, no. 5, pp. 954–970, Oct. 2019, doi: 10.1177/1367006917741364.

[5] J. H. Sim, "'But you don't sound Malay!': Language dominance and variation in the accents of English-Malay bilinguals in Singapore," *English World-Wide*, vol. 40, no. 1, pp. 79–108, 2019, doi: 10.1075/eww.00023.sim.

[6] S. Kirkham, "Ethnicity and phonetic variation in Sheffield English liquids," *Journal of the International Phonetic Association*, vol. 47, no. 1, pp. 17–35, Apr. 2017, doi: 10.1017/S0025100316000268.

[7] R. Mayr and A. Siddika, "Inter-generational transmission in a minority language setting: Stop consonant production by Bangladeshi heritage children and adults," *International Journal of Bilingualism*, vol. 22, no. 3, pp. 255–284, Jun. 2018, doi: 10.1177/1367006916672590.

[8] A. Cristià, "Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category," *J Acoust Soc Am*, vol. 129, no. 5, pp. 3271–3280, May 2011, doi: 10.1121/1.3562562.

[9] J. H. Sim and B. Post, "Variation in quality of maternal input and development of coda stops in English-speaking children in Singapore," *Journal of Child Language*, pp. 1–26, Aug. 2021, doi: 10.1017/S0305000921000593.

[10] J. E. Flege, C. Schirru, and I. R. A. MacKay, "Interaction between the native and second language phonetic subsystems," *Speech Communication*, vol. 40, no. 4, pp. 467–491, Jun. 2003, doi: 10.1016/S0167-6393(02)00128-0.

[11] J. A. Barlow, P. E. Branson, and I. S. B. Nip, "Phonetic equivalence in the acquisition of /l/ by Spanish–English bilingual children," *Bilingualism: Language and Cognition*, vol. 16, no. 1, pp. 68–85, Jan. 2013, doi: 10.1017/S1366728912000235.

[12] S. Kirkham and K. M. McCarthy, "Acquiring allophonic structure and phonetic detail in a bilingual community: The production of laterals by Sylheti-English bilingual children," *International Journal of Bilingualism*, vol. 25, no. 3, pp. 531–547, Jun. 2021, doi: 10.1177/1367006920947180.

[13] M. Kehoe and K. Kannathasan, "Development of voice onset time in monolingual and bilingual French-speaking children," *Lingua*, vol. 252, p. 102937, Mar. 2021, doi: 10.1016/j.lingua.2020.102937.

[14] J. H. Sim, "Variation in VOT in English child-directed speech of English-Mandarin and English-Malay early bilinguals in Singapore," Under review.

[15] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual Language Profile: An easy-to-use instrument to assess bilingualism." COERLL, University of Texas at Austin, 2012. [Online]. Available: https://sites.la.utexas.edu/bilingual/

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. [Online]. Available: http://www.praat.org/

[17] R. H. Kessinger and S. E. Blumstein, "Effects of speaking rate on voice-onset time in Thai, French, and English," *Journal of Phonetics*, vol. 25, no. 2, pp. 143–168, Apr. 1997, doi: 10.1006/jpho.1996.0039.

[18] M. Ramon-Casas, S. Cortés, A. Benet, C. Lleó, and L. Bosch, "Connecting perception and production in early Catalan–Spanish bilingual children: language dominance and quality of input effects," *J. Child Lang.*, pp. 1–22, Dec. 2021, doi: 10.1017/S0305000921000787.

[19] J. Flege, "Language contact in bilingualism: Phonetic system interactions," *Laboratory Phonology*, vol. 9, Jan. 2007.

The effects of vision on speech production in children

Lucie MÉNARD[1,2] and Christine TURGEON[3,4]

[1]Laboratoire de phonétique, Center for research on Brain, Language and Music
[2]Université du Québec à Montréal, Département de linguistique, Montréal, CANADA
[3]Centre interdisciplinaire de recherche en réadaptation et intégration sociale (Cirris)
[4]Université Laval, Département de réadaptation, Québec, CANADA

ABSTRACT

When producing French vowels, adult speakers with congenital visual deprivation produce smaller displacements of the lips (visible) but larger displacement of the tongue (invisible) than their sighted peers. To further investigate the impact of visual experience on the implementation of phonological targets, a speech production study was conducted with blind and sighted school-aged children. Eight congenitally blind children (mean age: 7 years old; range: 5 to 11 years) and eight sighted children (mean age: 7 years old; range: 5 to 11 years) were recorded while producing repetitions of the French cardinal vowels. Tongue and lip positions were tracked using a synchronous ultrasound and audio-visual recording system. Results show that blind children have reduced magnitude of tongue and lip displacement compared to their sighted peers, unlike adults. Overall, blind children display phonetic patterns typical of earlier stages of speech development in sighted children.

Keywords: speech production, speech development, multimodality, articulatory phonetics, vision.

1. INTRODUCTION

In face-to-face conversation, speech is produced and perceived through various modalities. Movements of the lips, jaw, and tongue, for instance, are heard and seen by the perceiver. Visually salient articulatory movements (of the lips and jaw) also contribute to speech identification in acoustically degraded conditions ([1, 2]) and in non-degraded conditions ([3]). The seminal McGurk effect ([4]) shows how high-level speech perception integrates auditory and visual features.

The above-mentioned sensory modalities are also involved in speech production. Indeed, many studies have emphasized the close relationship between speech production and speech perception systems. These studies led to the proposal of sensorimotor theories of speech ([5, 6] for instance). One of them, the *Perception for action control theory* (PACT) described in [7], posits that perceptual processes and procedural knowledge of speech actions are both involved in the planning and recovering of speech units. According to PACT, speech goals correspond to multisensory perceptuo-motor units. In the course of speech development, perception and action are tightly linked, and speech perception necessarily involves procedural knowledge of speech production mechanisms. Furthermore, perceptual mechanisms provide gestures with auditory, visual, and somatosensory templates that guide and maintain their development. In this paper, we explore the role of vision in speech production through a study of congenitally blind children, who never had access to perceptual visual templates.

2. VISUAL DEPRIVATION AND SPEECH PRODUCTION

The role of visual input has mainly been described in the speech perception domain. Yet, in the last decade, several studies have provided evidence that phonetic implementation of phonological goals is guided in part by visual constraints.

2.1. Speech production in blind adults

The fact that congenitally blind speakers learn to produce correct speech sounds suggests that visual cues are not mandatory in the control of speech movements. Nevertheless, differences exist between speech sounds produced by congenitally blind adults and sighted adults. For instance, in Canadian French-speaking adults, congenital blindness was found to reduce the size of the acoustic vowel space ([8]). At the articulatory level, the magnitude of lip contrasts (visible articulator) produced between phonologically rounded and unrounded vowels is also reduced in blind adults. Conversely, tongue contrasts are greater in blind speakers than in their sighted peers. In conditions where speech intelligibility was enhanced, such as in contrastive focus or in clear speaking conditions, blind speakers made less use of their visible articulators than sighted speakers. Perceptually, both speaker groups reach comparable intelligibility scores in clear auditory conditions ([9]). Interestingly, a different pattern of results was found in Dutch ([10]): a larger acoustic vowel space in blind than in sighted adult speakers was obtained, thus

suggesting that strategies to cope with visual deprivation are language-dependent. Altogether, those results suggest that the phonetic correlates of phonological contrasts are guided by sensory templates involving vision and acquired during speech development.

2.2. A developmental approach

At the language acquisition stage, babies establish relationships between auditory parameters and visual events ([11]). Although numerous studies have focused on the impact of auditory deprivation on speech production, not much is known about congenital visual impairment. This kind of deprivation could have consequences for the strategies used to develop language ([12, 13, 14]) and more specifically to produce phonological targets. Lewis (1975) ([15]) reported that, at the pre-babbling stage, blind babies imitated lip gestures less than sighted babies. Blind babies also show longer babbling phases, as well as delays in the production of their first words ([16, 17]). Elstner (1983) ([18]) and Mills (1987) ([19]) presented various studies showing phonological delays and phonetic-phonological disorders in older children. In a study of syllables produced by a congenitally blind 2-year-old German child, a higher number of phonological confusions between groups of visually dissimilar consonants (labial /b/ vs. velar /k/) was reported for the blind child compared to two English-speaking sighted children ([20]). Comparable results were found for three blind speakers ([20]). Not much is known, however, about the effects of visual deprivation on the development of the phonetic implementation of phonological targets.

2.3. Objective

The objective of this study is to describe the phonetic strategies (at the articulatory and acoustic levels) used to produce cardinal vowels in French-speaking congenitally blind children and sighted children.

3. METHOD

3.1. Participants

Eight congenitally blind children (mean age: 7 years old; range: 5 to 11 years) and eight sighted age-matched and gender-matched sighted children (mean age: 7 years old; range: 5 to 11 years) were recruited in the Montreal area. The blind children had congenital visual deprivation and had no other medical issues or language delays. All children were native speakers of Canadian French and were screened for auditory thresholds.

3.2. Corpus

Children had to produce multiple repetitions of the target vowels /i/, /u/ and /a/ embedded in /bVb/ syllables. Each syllable was produced ten times in each of the following conditions: “I am /bVb/ the musician” (“Je suis /bVb/ le musicien”, neutral condition); “No, I am /bVb/ the musician” (“Non, je suis /bVb/ le musicien”, focus condition), as a response to a question by the experimenter in which an error was introduced in the target vowel “Are you /bVb/ the musician?” The focus condition was elicited to generate vowel production under local hyperarticulation ([21]). It is known that, in contexts where perceptual saliency is enhanced, speakers hyperarticulate the gestures that are weighted more heavily in the phonetic representation. Furthermore, the ability to alter speech production along the hypoarticulation-hyperarticulation scale is an index of mature speech control. In young children (at 4 years old), in line with the sequential acquisition of lip and tongue motor control, the effects of contrastive focus are first observed in labial gestures, later followed by lingual gestures at 7-8 years old ([21]).

3.3. Experimental procedure

Articulatory and acoustic data were recorded using a combined ultrasound (Sonosite 180Plus) and audio-visual recording system (frontal and lateral views of the face). The synchronized acoustic signal was also recorded using a unidirectional SHURE microphone, with a sampling frequency of 22050 Hz.

3.4. Data analysis

All articulatory and acoustic measures were extracted at vowel midpoint. Lip positions were extracted using homemade Matlab programs, as described in [21]. Upper lip protrusion corresponded to the distance (in mm) between the reference position of the head and the vermilion border of the upper lip, on the lateral view. Ultrasound images corresponding to each vowel’s midpoint were selected. Tongue contours were semi-automatically tracked using the GetContour program (described in [22]) and corrected for head movements using an adapted version of the HOCUS procedure ([22, 26]). Measures of tongue height and maximum curvature index (hereafter referred to as MCI), as described in [23], were also extracted. Acoustic data were analysed using Praat ([24]). The first (F1) and second (F2) formant frequencies were also extracted at the vowel midpoint, using the Linear predictive coding algorithm embedded in Praat. To normalize for

between speaker variability in vocal tract size, all data were transformed in z-scores.

Linear mixed-effects models ([25]) were built using R. The dependent variables were articulatory and acoustic measures (one model for each variable), and the independent variables corresponded to the vowel (/i/, /a/, and /u/), the prosodic condition (neutral and focus) and the participant's group (sighted and blind). The participant was included in the models as a random effect (slope).

4. RESULTS

4.1. Production results

Figure 1 shows the dispersion ellipses of the three cardinal vowels /i u a/ for sighted and blind children, in the neutral and focus conditions. As this graph shows, the two speaker groups differ in terms of both vowel space organization and the effects of prosodic focus. First, a significant effect of the interaction between speaker group and prosodic condition is observed on both F1 and F2 values ($\chi^2(3)=16.74$; $p<0.001$ and $\chi^2(3)=9.08$; $p<0.05$): F1 values are significantly higher in the focus condition than in the neutral condition for the sighted children only. Regarding F2, only the sighted speakers produced the vowel /i/ with higher F2 in the focus condition than in the neutral condition ($p<.001$).

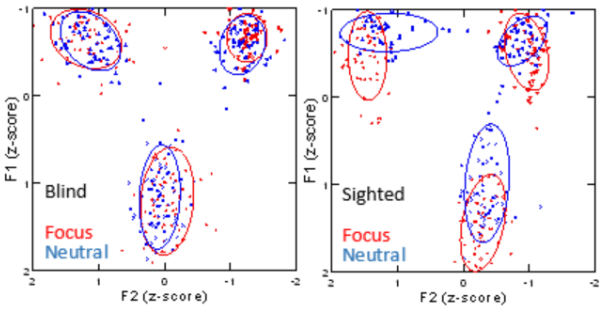


Figure 1: Mean values of F1 and F2 for sighted children (right panel) and blind children (left panel) in the neutral (blue line) and focus (red line) conditions.

The articulatory strategies the children used to implement the three phonological targets are displayed in Figures 2 and 3. Figure 2 shows the upper lip protrusion values. Data are averaged across speakers, for each vowel. In this articulatory dimension, sighted speakers produced significantly larger differences between the rounded vowel /u/ and the unrounded vowel /i/ ($\chi^2(5)=25.42$; $p<0.001$). Furthermore, a significant effect of the interaction between speaker group and prosodic condition is found ($\chi^2(3)=14.82$; $p<0.01$), with overall lip protrusion values being enhanced in the focus condition compared to the neutral condition for the sighted group only.

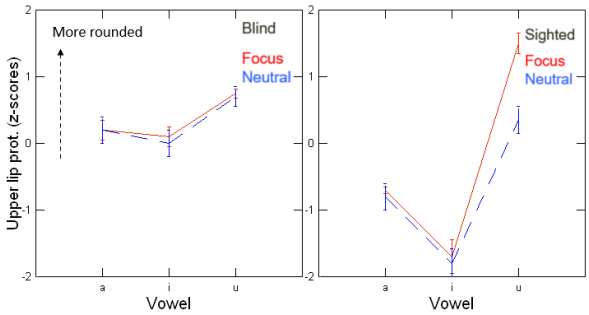


Figure 2: Mean values of upper lip position for sighted children (right panel) and blind children (left panel) in the neutral (blue line) and focus (red line) conditions. Error bars are standard errors.

Figure 3 depicts the mean values of tongue front-back position, for each vowel and each prosodic condition. As was the case for lip protrusion, lip height values are averaged across speakers. A significant effect of the interaction between speaker group, prosodic condition and vowel is also found for this articulatory dimension ($\chi^2(7)=21.36$; $p<0.001$). Indeed, the difference between tongue position in /u/ relative to /i/ is significantly larger for sighted children than for blind children ($p<0.01$). No significant effect of prosodic condition was found for blind children, whereas sighted children produced less fronted /u/ and /a/ in the focus condition compared to their neutral counterparts ($p<0.001$).

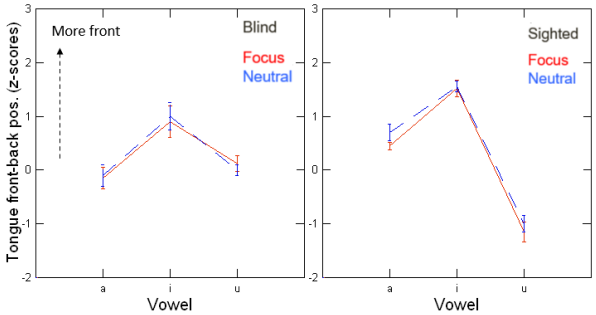


Figure 3: Mean values of tongue front-back position for sighted children (right panel) and blind children (left panel) in the neutral (blue line) and focus (red line) conditions. Error bars are standard errors.

Finally, MCI values (tongue curvature) are shown in Figure 4. As was the case for tongue position, a significant effect of speaker group, prosodic condition and vowel is revealed by the statistical analysis ($\chi^2(7)=27.85$; $p<0.001$). Indeed, the difference between tongue curvature values in /u/ and /i/ is significantly larger in sighted children than in their blind peers ($p<0.001$). Furthermore, for /i/ and /u/, tongue curvature was reduced (indicating a more

bunched shape) in the focus condition compared to the neutral condition in sighted children only ($p<0.001$). No such effect was found in the blind group.

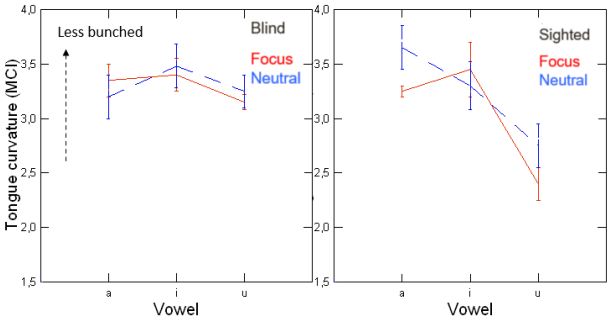


Figure 4: Mean values of tongue curvature for sighted children (right panel) and blind children (left panel) in the neutral (blue line) and focus (red line) conditions. Error bars are standard errors

5. DISCUSSION

The results presented so far suggest that congenital visual deprivation impacts the phonetic development of vowels in children. Indeed, in our study, congenitally blind children did not differentiate between tongue and lip positions to the same extent as their sighted age-matched controls: for both these articulators, the use of labial and lingual contrasts was reduced when access to visual input was not available. Furthermore, no effect of contrastive focus was observed on the vowels produced by congenitally blind children. This pattern of results points to delayed development, corresponding to an earlier stage of motor control maturity than in sighted children (see [21]).

It should be noted, however, that those differences do not indicate specific speech disorders. Indeed, as reported in other studies ([8]), by adulthood, blind speakers have acquired mature motor control of the lips and tongue such that those gestures are recruited in complementary ways to implement phonological targets in sighted and blind groups. Developmental trajectories are therefore likely linked to the richness of sensory input surrounding the child. A longitudinal analysis is currently under way to better characterize speech development in blind children.

Despite the significance of the results, some limitations must be acknowledged. First, our analyses include only eight children from each group. In an effort to limit the effects of associated motor or cognitive deficits, we included only children with isolated congenital deprivation and with a similar linguistic background (native Canadian French speakers, in our case), which greatly reduces the number of participants. Second, no measure of speaker's auditory perception skills was taken (apart from auditory thresholds). Although we cannot rule out the possibility that part of the group differences reported here at the production level could arise from perceptual acuity differences, it is highly unlikely that the production results are entirely related to such perceptual differences. Indeed, in our previous studies on speech production by blind and sighted adults, perception did not explain production patterns ([8, 21]). Furthermore, only three vowels were targeted in our study. Further studies should undoubtedly be conducted to extend our results to other phonemes, and possibly in other languages. Recent studies have indeed shown that, although blindness impacts both acoustic and articulatory correlates of vowel production, its effect varies across languages ([10]).

Acknowledgements

This work was supported by an SSHRC grant and an NSERC grant. Thanks to Zofia Laubitz for copy editing the paper.

6. REFERENCES

[1] Sumby, W. H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26(2): 212–215.

[2] Summerfield, Q. 1979. Use of visual information for phonetic perception. *Phonetica* 36: 314–331.

[3] Arnold, P. and Hill, F. 2001. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology* 92: 339–355.

[4] McGurk, H. and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264: 746–748.

[5] Hickok G, Houde J, Rong F. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*. 201, 169(3):407-22.

[6] Parrell, B., Ramanarayanan, V., Nagarajan, S., and Houde, J. 2018. FACTS: A hierarchical task-based control model of speech incorporating sensory feedback, *Proceedings of Interspeech* 2018, 1497–1501.

[7] Schwartz, J.-L., Basirat, A., Ménard, L. and Sato, M. 2012. The Perception for Action Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics* 25: 336–354.

[8] Ménard, L., Toupin, C., Baum, S. R., Drouin, S. Aubin, J. and Tiede, M. K. 2013. Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *Journal of the Acoustical Society of America* 134(4):2975–2987.

[9] Ménard, L., Trudeau-Fisette, P. and Tiede, M. 2022. Intelligibility of speech produced by sighted and blind adults, *PlosOne*.

[10] Veenstra, P., Everhardt, M. K., & Wieling, M. (2018). *Vision deprived language acquisition: Vowel production and ASR efficacy.* Poster presented at the

16th Conference on Laboratory phonology (LabPhon), Lisbon, Portugal.
[11] Kuhl, P. K. and Meltzoff, A. N. 1982. The bimodal perception of speech in infancy. *Science* 218(4577): 1138–1141.
[12] Reynell, J. 1978. Development patterns of visually handicapped children. *Child: Care, Health and Development* 4: 291–303.
[13] McConachie, H. 1990. Early language development and severe visual impairment. *Child: Care, Health, and Development* 16(1): 55–61.
[14] Cappagli, G., Cuturi, L.F., Signorini, S. *et al.* 2022. Early visual deprivation disrupts the mental representation of numbers in visually impaired children. *Scientific Reports,* 12, 22538.
[15] Lewis, M. M. 1975. *Infant Speech: A Study of the Beginnings of Language*. New York: Arno Press.
[16] Burlingham, D. 1961. Some notes on the development of the blind. *Psychoanalytic Study of the Child* 16: 121–145.
[17] Warren, D. H. 1977. *Blindness and Early Childhood Development*. New York: American Foundation for the Blind.
[18] Elstner, W. 1983. Abnormalities in the verbal communication of visually impaired children. In A.E. Mills (ed.), *Language Acquisition in the Blind Child*, 18–41. London: Croom Helm.
[19] Mills, A. E. 1983. *Language Acquisition in the Blind Child: Normal and Deficient*. San Diego, CA: College-Hill Press.
[20] Mills, A. E. 1987. The development of phonology in the blind child. In B. Dodd and R. Campbell (eds.), *Hearing by Eye: The Psychology of Lip-Reading*, 145–163. London: Lawrence Erlbaum Associates.
[21] Ménard, L., Prémont, A., Trudeau-Fisette, P., Turgeon, C. and Tiede, M. 2020. Probing the Development of Phonemic Goals in French through Prosodic Focus, *Journal of Speech, Language, and Hearing Research*.
[22] Noiray, A., Ries, J., Tiede, M., Rubertus, E., Laporte, C., & Ménard, L. 2020. Recording and analyzing kinematic data in children and adults with SOLLAR: Sonographic & Optical Linguo-Labial Articulation Recording system. *Laboratory Phonology*, http://doi.org/10.5334/labphon.241
[23] Dawson, K. M., Tiede, M. K. and Whalen, D. H. 2016. Methods for quantifying tongue shape and complexity using ultrasound imaging, *Clinical Linguistics & Phonetics*, 30:3-5, 328-344.
[24] Boersma, P. and Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/.
[25] Bates, D., Maechler, M., Bolker, B. and Walker, S. 2012. Lme4: *Linear mixed-effects models using Eigen & S4*. R package version 1.0-5.
[26] Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., & Vatikiotis-Bateson, E. 2005. The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research, 48,* 543–553.

The Origins of Phonology and Lexicon in Infancy (OPAL): Phonological Abstraction Before Perceptual Attunement?

Denis Burnham[1], Catherine Best[1], Antonia Götz[1], Marina Kalashnikova[2,1], Elizabeth Johnson[3], Eylem Altuntas[1], Anne Cutler[1,4,*]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University; [2]Basque Center on Cognition, Brain & Language [3]Child Language & Speech Studies Lab, University of Toronto; [4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
denis.burnham@westernsydney.edu.au; c.best@westernsydney.edu.au; a.goetz@westernsydney.edu.au; m.kalashnikova@bcbl.eu; elizabeth.johnson@utoronto.ca;

Inspired by, and dedicated to the memory of, Distinguished Professor Anne Cutler

ABSTRACT

Recently it was found that Dutch-speaking adults adopted from Korea at less than 6 months learned to perceive a Korean consonant contrast that is non-native in Dutch better than native Dutch speakers lacking Korean experience. They also better generalised that perceptual learning to other articulation places and to speech production. The inference that phonological abstraction occurs before 6 months, prior to language-specific attunement (6-10 months), begs direct evidence from infants. Our project, 'Origins of Phonology and Lexicon' (OPAL), is designed to address that lacuna. In this paper we (i) derive and describe tasks for evaluating three types of phonological abstraction: Protoword Structure, Phonological Inventories, and Phoneme Features; (ii) describe our method for testing these in infancy; and (iii) validate these tasks with adults. While adults showed differences among tasks, they were able to perform each abstraction with just one feedback trial, providing a firm basis for our ongoing infant studies.

Keywords: phonological abstraction, infant speech perception, perceptual attunement, adult validation

1. INTRODUCTION

1.1. Background

In recent studies by Choi, Cutler and colleagues, a group of Dutch-speaking adults who had been adopted from Korea as infants between 3 and 6 months, with no further experience of Korean, were trained to identify the Korean fortis/lenis/aspirated alveolar stops /t*/-/t/-/t^h/, which are not phonemic in Dutch [1, 2]. The adoptees learned the distinction more quickly than a control group of native Dutch-speaking adults lacking any Korean language experience. While this might suggest that the adoptees had stored a complete compilation of the Korean phoneme repertoire, it was also found that the adoptees surpassed the controls in (i) generalising from the learned alveolar contrast to the same fortis/lenis/aspirated contrast in bilabial and velar places, (ii) producing these Korean consonants, and (iii) showing higher speech perception-production correlations. Together these results suggest that what had been gleaned before the age of 6 months was not mere memory for the Korean consonant inventory, but rather phonologically abstract representations, in this case the contrasting fortis/lenis/aspirated features of stop consonants. Consequently, phonological knowledge appears to be gained before the emergence of perceptual attunement, the process by which infants tune in to the phonemes and contrasts of their ambient language between about 6 to 10 months [4].

These adult results beg the questions of how and when abstract phonological representations are laid down in infancy and this is what our 'Origins of Phonology and Lexicon' (OPAL) project has been funded to investigate [3]. This paper concerns the development (i) of tasks for investigating the timing, nature and extent of phonological abstraction in infancy, (ii) of methods for testing performance on the tasks; and (iii) validating the tasks and measures with adult participants.

1.2. Development of Phonological Abstraction Tasks

A recent EEG study has provided a proof of concept for a method to investigate phonological abstraction by 5-month-old infants [5]. Infants were familiarised with two classes of spoken non-words differing only in syllable location, viz, ABA (e.g., *baluba. gotigo, etc.*), paired with an arbitrary visual cartoon, e.g., a fish, and AAB (e.g., *babalu, gogoti, etc.*) paired with another visual cartoon, e.g., a lion. In the subsequent test trials with new non-words of each syllable structure, infants showed neural mismatches to incongruent word-label pairings (i.e., reverse pairings

to those in familiarisation), but not to congruent pairings (same as familiarisation), suggesting that infants had abstracted the opposing multi-syllabic structures, ABA vs AAB. We have adapted this method to test three types of phonological abstraction in infants under 6 months. Derivation of these three phonological abstraction tasks, and determination of familiarisation and test stimuli, are set out below.

1.2.1. Protoword Structure (PRO)

Newborns discriminate differences in the structure of syllable sequences, e.g., AAB (*babamu, gegeba*) vs ABC (*bamuge, gebamu*) [6,7]. In addition to discriminating such structural differences, young infants can also detect changes in position of a syllable within a series (A*B*A→AA*B*) as in [5] and in statistical learning studies with artificial grammars [8]. Our protoword structure task (Table 1) goes beyond syllable sequences in *words* to test sensitivity to phonotactics, i.e., phoneme position in *syllables*.

Phase & Phonemes	**Category A CCV**	**Category B CVC**
Learning 3-syll sets from:	/blV, glV, spV, stV, ʃmV/	/bVl, gVl, sVp, sVt, ʃVm/
example:	*blee-glar-spoo*	*beel-garl-soop*
Test	As for Learning but new nonwords	
example:	*blar-gloo-spee*	*barl-gool-seep*

Table 1: Proto-Word Structures (PRO) Task (V = vowel).

1.2.2. Phonological Inventories (INV)

Infant sensitivity to language differences in prosodic rhythms is supported by French newborns' successful discrimination of stress-timed English from mora-timed Japanese but not from stress-timed Dutch [9]. It has been proposed that rhythm differences between languages reflect differences in proportions of vowels and consonants [10]. Moreover, differences in consonant/vowel distributions have consequences for the phonological abstraction of permissible word structures [11]. Our phonological inventories task (Table 2) will provide the first test of infants' ability to distinguish among consonant/vowel distributions.

Phase & Phonemes	**Category A 2V-8C**	**Category B 8V-2C**
Learning 3-CV words from:	C: /b, k, dʒ, s, l, w, m, n/ V: /i:, æɔ/	C: /m, k/ V: /i:, ɐ:, o:, ʉ:, oɪ, æɪ, æɔ, ɑe/
example:	*bee-kou-jee-sou*	*mee-koy-mar-kay*
Test	As for Learning but new nonwords	
example:	*bou-kee-jou-see*	*moy-kar-may-zor*

Table 2: Phonological Inventories (INV) Task (V = vowel, C = consonant)

1.2.3. Phoneme Features (FEA)

Between 14 and 19 months infants develop phonological categories that: are durable across changes in English accents, mature over age, and are correlated with infants' vocabulary size [12,13,14].

We will investigate a possible precursor of phonological category formation in infancy – assessing whether infants can categorise consonants by a phonological feature contrast akin to that tested by Choi, Cutler and colleagues [1, 2], in this case voiced vs voiceless consonants (see Table 3).

Phase & Phonemes	**Category A Voiced**	**Category B Voiceless**
Learning 3-syll words from:	/bV, vV, dʒV/	/pV, fV, tʃV/
example:	*bee-var-jor*	*par-for-choo*
Test*	/dV, zV, gV/	/tʃV, sV, kV/
example	*dor-zoo-gae*	*too-sae-kee*

Table 3: Phoneme Features (FEA) Task (V = vowel). *Test phase uses new consonants, to test feature generalisation.

1.2.4. Summary

As can be seen the three tasks are highly complex, requiring us to calculate the size of the expected effects in order to determine sample sizes and analysis models for the infant experiments. Therefore, we first needed to validate the three tasks with adults, which is what we report in this paper.

2. METHOD[i]

2.1 Design

A within-subject design was employed: 3 tasks (Protoword Structure / Phonological Inventories / Phonemic Features) x 2 phases (Learning / Test).

2.2 Participants

65 adults participated but 11 had incomplete data. 54 data sets were included for trials to criterion (TTC), and 53 for reaction times (RT) (one had missing RT data in one cell). Language background varied within the constraint that each participant must be proficient in English (18 L1-English [8 monolingual]; 36 L2-English). Mean age was 38.54 years (median = 34.5, range = 17-77, standard deviation = 14.21).

2.3. Experimental Environment

The software for the three tasks was developed in PsychoPy, a Python-based package for behavioural studies [15]. The study was conducted online (due to the COVID pandemic) using Pavlovia, an extension of PsychoPy [16]. All instructions were in English.

2.4. Stimulus Materials

For each task, two contrasting categories (A, B) of multi-syllabic non-words were created, composed of CCV vs CVC syllables (C=consonant, V=vowel) for *Protoword Structure* (PRO); 2V-8C vs 8V-2C inventories for *Phonological Inventories* (INV); and voiced (Vd) vs voiceless (Vl) initial consonants for *Phonemic Features* (FEA) (see Tables 1-3). For each task, 480 words were created, 240 each for Category A and B. Words were recorded by a different female Australian English speaker for each task.

2.5. Procedure

In each task within a two-alternative forced-choice implicit learning task, participants heard non-words from categories (A or B) and were required to identify the category of each word by pressing the A computer key for Category 'A', and the 'L' key for Category B. There were two phases, Learning and Test. Each drew stimuli from 120 Category A and 120 Category B words selected randomly without replacement from the relevant category. No more than three words from either category appeared in succession.

At the start of the Learning phase participants heard, 'This sound belongs to A', followed by a word from List A. This was the *only* trial with feedback; *all* remaining trials in Learning and Test lacked feedback. The sequence in ensuing trials was: (i) 'Ready' on the screen for 0.5secs; (ii) auditory word presented; (iii) a required key press ('A' or 'L'); (iv) next trial after response or after a time-out period of 1.5secs. No-response trials were not repeated. Key press responses and reaction times were recorded.

Participants proceeded from Learning to Test phase, and from Test phase to task-end if they (i) made 7 correct responses in a moving window of 8 trials (binomial probability = .03516, p<.05), *or* (ii) if they did not reach this criterion within 50 trials. Between Learning and Test phases, and between the Test phase of one task and the Learning phase of the next, there was a break and participants advanced via pressing the space bar. All participants completed all three tasks, and task order was counterbalanced.

3. RESULTS

The dependent measures were Response Criterion, and Reaction Times for correct responses in both Learning and Test phase in each task.

3.1. Response Criterion Measurements

3.1.1 Participants Reaching Criterion

The number of participants reaching criterion (7 correct responses in a moving window of 8 trials within the maximum of 50 trials) in each phase of each task are shown in the upper row of Table 4. A 2 x 3 chi-square analysis found no differences in number of participants reaching criterion across tasks or phases, χ^2 (2) = 0.22, p > .05. Further chi-square tests for each phase in each task were also not significant.

See the lower row of Table 4 for the binomial probability of reaching criterion against chance (0.5). In both phases of the PRO and INV tasks, the number of participants (/54) reaching criterion was greater than chance. But in neither FEA phase did the number reaching criterion exceed chance.[ii]

3.1.1 Trials to Criterion (TTC)

Figure 1 (checkered plots) and Table 4 upper row) show, for each task, mean trials to criterion for participants reaching criterion within the maximum of 50 trials per phase. Those who did not reach criterion were given a score of 57 (maximum number of trials (50) + the minimum number to reach criterion thereafter (7) see solid bars in Figure 1). This latter 'All Participants' measure affords repeated measures analysis as each of the 54 participants has a score for each of the six task x phase cells.

	PRO		INV		FEA	
	Learn	Test	Learn	Test	Learn	Test
n	35	33	34	34	26	29
p =	.014*	.049*	.027*	.027*	.608	.292

Table 4: n = number (of N= 54) per task and phase reaching criterion within 50 trials. p = binomial probability the proportion is above chance (.5). *p <.05.

A 3-task x 2-phase analysis of variance (ANOVA) of All Participants revealed no significant main effect for phase, F(1,53) = 0.05, p > .05, or its interactions. As FEA was the only task that incorporated (i) categories distinguished by a language-specific phonological feature distinction (voicing) and (ii) a

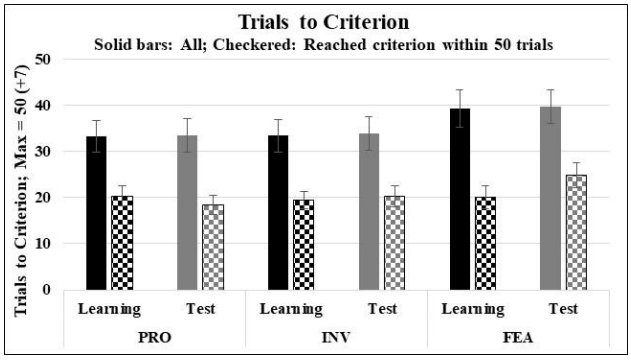


Figure 1: Trials to Criterion (TTC) for all participants in 50 (+7) trials and for only those reaching criterion within 50 trials. Error bars are standard errors of the mean. (s.e.m.)

generalisation from Learning to Test Phases, we ran planned comparisons to compare (i) FEA to the other two tasks combined (FEA vs PRO+INV), and (ii) the other two tasks (PRO vs INV), both of which had neither a feature distinction nor generalisation from Learning to Test. There were significantly more TTCs for FEA than for the mean of PRO and INV, $F(1,53) = 5.39$, $p < .05$, but no significant difference between PRO and INV, $F(1,53) = 0.01$, $p > .05$.[iii]

3.3.2. Mean Reaction Times (RTs)

Reaction Times (RTs) cannot be less than zero, so they are typically skewed. To correct for this, $\log^{10}$ values were calculated. These were used in the analyses and are plotted in Figure 2.

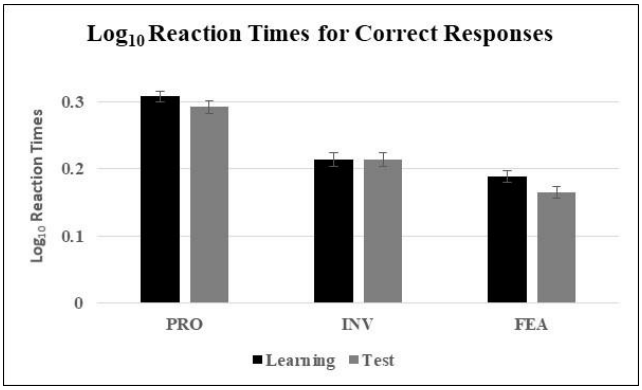


Figure 2: Mean $\log^{10}$ Reaction Times for Learning and Test phases in each of the three tasks. Error bars are standard errors of the mean.

A 3-task x 2-phase ANOVA revealed that RTs decreased significantly from Learning to Test across tasks, $F(1, 52) = 11.34$, $p < .01$. There was a main effect of task, indicating that RTs were significantly shorter for FEA than for PRO and INV combined, $F(1, 52) = 98.81$, $p < .001$. In turn, RTs were significantly shorter for INV than PRO, $F(1, 52) = 255.89$, $p < .001$.[iv]

4. DISCUSSION

Three phonological abstraction tasks – *Protoword Structure* (PRO), *Phonological Inventories* (INV), *Phonetic Features* (FEA) – were developed and validated with adults with a view to using these in tests of phonological abstraction in infancy.

Adults engaged in all three tasks, but the FEA task appears to be more difficult than either the PRO or the INV tasks because: (i) the number of participants reaching criterion within 50 trials was not significantly above chance in either Learning or Test for FEA (in PRO and INV it was significant in both phases) and (ii) there were more TTCs in FEA than in PRO and INV. Inspection of TTC data for only those reaching criterion (Figure 1) suggests an increase in TTC for FEA from Learning to the (Generalisation) Test. However, this observation is weakened by the fact that (i) this comparison was not significant in the All Participants TTC ANOVA, (ii) in neither FEA Learning nor Test was the number of participants reaching criterion greater than chance, and (iii) the faster RTs in FEA than for PRO or INV occurred irrespective of phase.

In contrast to the TTC analyses, RTs were surprisingly significantly shorter for FEA than PRO and INV, which may seem to imply, somewhat paradoxically, that FEA was an easier task. However, the TTC analyses imply that FEA was more difficult (were significantly greater TTCs than PRO ad INV), so it appears that participants may have responded more quickly in the FEA task simply because they had given up trying to figure out the basis of the two categories.

It is of passing interest that there was no effect of language background (L1-English vs L2-English) on either TTC or RT (see endnotes ii, iii, iv). There is a possible effect of language background on numbers reaching criterion in the FEA task (due to the use of an English voicing contrast), but the only firm conclusion that can be drawn is that, in general, the results here are robust over L1-English and L2-English adults.

Despite differences in FEA versus PRO and INV, the adults were able to complete all three tasks under difficult conditions, i.e., there was feedback on only one preliminary trial and that feedback was for only one of the two categories. In addition, they were given no information regarding the basis of the categories, which is, itself, the same constraint that the infants will face. So, together the results augur well for using these tasks (in EEG adaptations) with infants.

5. SUMMARY AND CONCLUSIONS

Three tests of phonological abstraction were developed and were validated with adults. One task, FEA, was more difficult that the other two, PRO and INV, but adults were able to engage with all tasks under very difficult conditions that would also be the case for infants. The results provide a solid foundation for the infant studies in this project.

6. APPENDIX – FURTHER DATA

In addition to the three tasks here, the OPAL project also includes a fourth task – a double test of phonological abstraction. In this the two categories to be learned are based on a consonantal place of articulation contrast, *plus* an additional learning-to-test change of modality (auditory-only to visual-only or visual-only to auditory-only). Preliminary infant data for this task is reported in ICPhS 2023 paper 42.

7. REFERENCES

[1] Choi, J., Broersma, M. and Cutler, A., 2017. Early phonology revealed by international adoptees' birth language retention. *Proceedings of the National Academy of Sciences 114*, pp.7307-7312.

[2] Choi, J., Cutler, A. and Broersma, M., 2017. Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory. *Royal Society Open Science 4*, 60660.

[3] Burnham, D., Best, C. (Cutler, A., Johnson, E., Kalashnikova, M.) Origins of Phonology and Lexicon: Abstract representations before 6 months. Australian Research Council Discovery Grant, DP00025567.

[4] Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology 50*, pp.509-535.

[5] Kabdebon, C. and Dehaene-Lambertz, G., 2019. Symbolic labeling in 5-month-old human infants. *Proceedings of the National Academy of Sciences*, *116*(12), pp. 5805-5810.

[6] Gervain, J., Macagno, F., Cogoi, S., Peña, M. and Mehler, J., 2008. The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, *105*(37), pp.14222-14227.

[7] Gervain, J., Berent, I. and Werker, J.F., 2012. Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, *24*(3), pp.564-574.

[8] Saffran, J.R., Aslin, R.N. and Newport, E.L., 1996. Statistical learning by 8-month-old infants. *Science*, *274*(5294), pp.1926-1928.

[9] Nazzi, T., Bertoncini, J., & Mehler, J. 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), pp. 756

[10] Ramus, F., Nespor, M. and Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*,*73*, pp.265-292.

[11] Nazzi, T., & Cutler, A. 2019. How consonants and vowels shape spoken-language recognition. *Annual Review of Linguistics*, *5*, pp. 25-47.

[12] Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. 2009. Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychological Science*, *20*(5), pp. 539-542.

[13] Mulak, K.E., Best, C.T., Tyler, M.D., Kitamura, C. and Bundgaard-Nielsen, R.L., 2010. Vocabulary size predicts the development of phonological constancy: An eyetracking study of word identification in a non-native dialect by 15-and 19-month-olds. In *Proceedings of the 20th International Congress on Acoustics*.

[14] Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. 2013. Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development*, *84*(6), pp. 2064-2078.

[15] Peirce, J.W., 2007. PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), pp.8-13.

[16] Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J.K., 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), pp.195-203.

[i] We acknowledge assistance with *Programming:* Dr Leidy Castro-Meneses, Johnson Chen, Dr Benjawan Kasisopa, Prof Ronan Reilly, Dr Chris Wang, Mary Youkhanna; *Design*: Dr. Pelle Söderström; *Stimulus construction*: Dr Benjawan Kasisopa, Dr Chris Wang; *Stimulus voices*: Dr Ruth Brookman, Caitlin Hooper, Kate Jones; *Collation, stimuli & data*: Michael Barrett, Nur Hazwani Abd Hadi, Supreet Kaur; *Recruiting*: Jessica Simpson.

[ii] See the following table for the percentage of L1-English and L2-English participants reaching criterion. Low and unequal numbers preclude firm conclusions, but we note the following differences:

- PRO: similar percentages
- INV: more L2-English than L1-English participants reached criterion (maybe due to the more diverse language experience of L2 speakers)
- FEA: more L1-English than L2-English reached criterion (maybe due to the voicing contrast being based on an English phonetic voicing distinction).

English status	N	PRO		INV		FEA	
		Lrn	Tst	Lrn	Tst	Lrn	Tst
L1-Eng	**18**	61	61	50	39	50	67
L2-Eng	**36**	64	61	69	75	44	47

[iii] Further analyses of TTC for L1-English vs L2-English revealed no significant interactions of either Phase or Task with English status (L1, L2), so the results of the main analysis – more TTC for FEA than for the mean of PRO and INV – hold across both language background groups.

[iv] Further analyses of log RTs for L1-English vs L2-English revealed no significant interactions of Phase or Task with English status (L1, L2), so the results of the main analysis – increased RTs from Learning to Test, FEA with shorter RTs than PRO and INV, and INV with shorter RTs than PRO – hold across both language background groups.

Input distribution of Korean affricates in the context of a diachronic sound change: comparison between IDS and ADS

Minji Nam[a], Youngon Choi[a], Hyun Kyung Hwang[b], Mazuka Reiko[c,d,e]

[a]Chung-Ang University, Korea, [b]University of Tsukuba, Japan, [c]RIKEN Center for Brain Science, Japan, [d]Waseda University, Japan, [e]Duke University, USA
minji.namm@gmail.com; yochoi@cau.ac.kr; hwang.kyung.gu@u.tsukuba.ac.jp; reiko.mazuka@riken.jp

ABSTRACT

Compared to stops, little is known about the distributions of affricates along the acoustic spectrums provided by caregivers for their infants. Korean affricates have three-way contrasts (fortis, lenis, and aspirated), all of which used to be distinguishable by frication and aspiration duration. However, due to a diachronic change, lenis no longer differs from aspirated in the durational cue but differs in fundamental frequency (*f*0) from aspirated.

We examined the distributional pattern of maternal input along these acoustic parameters in their speech to infants and adults. Like stop categories, mothers' productions reflected the diachronic sound change regardless of speech register, showing the difference in *f*0 not in duration for lenis-aspirated contrast. However, the cue differences were selectively enhanced for relevant contrasts in infant-directed speech: larger *f*0 differences for the lenis-aspirated contrast and greater durational differences between fortis and other categories. This might provide additional phonetic support for young Korean learners.

Keywords: affricates, Korean, phonetic input, diachronic change, speech register

1. INTRODUCTION

An affricate is a type of consonant that is produced by combining the articulation of a plosive with the fricative [1]. Therefore, both the duration of frication noise and voice onset time (VOT) play crucial roles in distinguishing affricate categories. Affricates have posed challenges in specifying their underlying phonetic structure [1], and have received relatively little attention in the field of speech development due to their unique combination of articulatory features.

Korean presents an even more challenging case for learners because it has typologically unusual three-way affricate contrasts, which are entirely voiceless. The three affricate categories in Korean are known as fortis (or tense), lenis (lax or plain), and aspirated (/c*, c, c^h/, see examples in Table 1). Among the three categories, fortis has the shortest total noise duration (combining durations of frication and aspiration, [1]). The aspirated category used to be produced with the longest noise duration, followed by that of lenis (intermediate duration between fortis and aspirated, [2]).

However, like three-way stop categories in Korean, two of the three affricate categories—lenis and aspirated—have undergone a diachronic phonetic change in the past several decades [3]–[6]. As a result, noise duration has shortened for aspirated, largely overlapping with the range of noise duration for lenis. By contrast, the fundamental frequency (*f*0) difference was enhanced between the lenis and aspirated categories [1], [2].

Adult native Korean speakers relied on durational cues in categorizing and discriminating fortis from other categories [1], [5], whereas *f*0 was the main cue to distinguish lenis from others [1]. These demonstrated that native Korean adults utilized the relevant acoustic cues to distinguish the target affricate categories, showing that their perception pattern reflects the shift in the use of phonetic cues.

These results raise a possibility that the Korean caregiver's phonetic input of affricates to young learners has characteristics reflecting the phonetic changes. However, it is also possible that caregiver input may contain exaggerated acoustic parameters to support early learners [7]–[10] such that the durational and frequency cues are presented with greater differences than the typical input observed among speech directed to adults. Early maternal phonetic quality is associated with infants' consonant discrimination abilities [10]–[12], the child's later language outcomes [13], and children's production [14].

In the context of a diachronic change, it is also possible that individual caregivers may differ in terms of reflecting the change in their input to young infants, as reported by Choi et al. [7]. This may result in mixed distributions along the acoustic dimensions.

To our best knowledge, no studies have examined the input distribution patterns of the Korean affricates along the acoustic parameters of noise duration and *f*0 available to young Korean-learning infants. Therefore, the present study investigated the phonetic input pattern of three affricate categories in the

speech of mothers whose infants are between 4 and 10 months of age.

Korean	Meaning	IPA	Category
짜다	salty	t'sada	Fortis
자다	sleep	tsada	Lenis
차다	kick	tsʰada	Aspirated

Table 1: Examples of words used to elicit word-initial three-way affricate categories in Korean.

2. DATA & METHODS

2.1. Participants and Materials

Thirty-six native Korean-speaking mothers participated in the recording session conducted in our lab. The age of the mothers ranged from 23 to 39 years (mean = 31.9 years) and their infants were between 4 and 10 months of age (mean = 6.8 months). All the mothers reported that they grew up and currently reside in Seoul.

A list of three sets of minimal triplet words (짜다/자다/차다, 찌다/지다/치다, 짱구/장구/창구) was provided for the mothers and they were asked to utter each word in the carrier sentence as below in two speech registers: infant-directed speech (IDS) and adult-directed speech (ADS). In the present paper, we specifically report the analyses of the three words listed in Table 1 from the full list.

- ADS: 이것은 ____라고 합니다.
  (ikʌsɨn ____ lako hapnita,
  'This is ____'.)
- IDS: 이것은 ____래.
  (ikʌsɨn ____ le, 'This is ____'.)

In the IDS register, mothers were provided with a set of picture cards depicting target words. They were asked to show the infant the picture cards while recording. In the ADS register, mothers were asked to read the target words from the list as if casually speaking to another adult. The target words were recorded five times for each speech register.

Some mothers missed reading some of the target words. Some repeated the word less than five times. We included partial sets of obtained tokens in the analyses. However, we excluded tokens with noise. As a result, a total of 883 tokens were included in the final analyses, 463 for ADS and 420 for IDS, respectively.

2.2. Procedure

The mothers were taken into a small, sound-treated room with their infants. The utterances were recorded using a portable Marantz digital recorder (PMD 661MKII) and a head-mounted microphone (Shure SM10A).

2.3. Measurements

The noise duration, including frication duration, and aspiration duration, and *f*0 in the following vowel were measured by implementing a Praat script [15]. Frication duration (Fric.) was determined by the presence of high-frequency, high-amplitude noise in the spectrogram. Aspiration duration (Asp.) was measured from the offset of high-frequency noise to the onset of the following vowel and was identified by low amplitude and low-frequency noise. F0 was measured by converting the average wavelength of the first 25 milliseconds of the vowels. Vowel onset was indicated by the first zero-crossing of periodicity after the aperiodic noise.

3. RESULTS

3.1. Acoustic characteristics of the maternal input

The distribution patterns along the noise duration and *f*0 were very similar across the two speech registers. As shown in Figure 1 and Table 2, the noise duration of fortis affricates was much shorter than the other two categories. However, the range of the noise duration for lenis affricates overlapped with those of aspirated affricates in both ADS and IDS.

We estimated the measured durations of Fric., Asp., noise, and *f*0 properties using linear mixed-effects regression models (*lmer* in lme4 package in R version 4.1.0, [16]). Model comparison was done using the likelihood ratio test. The full model included the fixed effects of category (fortis, lenis, & aspirated), speech register (IDS, ADS), and their interactions. Speaker was included as random effects. Model estimates with the absolute value of the *t*-statistic greater than 2 were considered significant at a 5% significance level [17].

For all the acoustic parameters, significant main effects were found for both the category and register. The interaction effects between the two were also significant. The estimated *f*0 was significantly lower for lenis ($\beta = -95.93$, $t = -20.21$) than aspirated ($\beta = 279.88$, reference category) and this difference was enhanced ($\beta = -42.05$, $t = -6.05$) in IDS. The estimated *f*0 was also significantly lower for fortis ($\beta = -41.18$, $t = -8.80$) than aspirated but the difference

Comparison pair	Acoustic Properties	Speech Register	Mean (difference)	SD
Aspirated – Fortis	Fric. duration difference (ms)	ADS	9.43	7.88
		IDS	14.17	13.44
	Asp. Duration difference (ms)	ADS	46.89	14.63
		IDS	55.75	21.70
	Noise duration difference (fric. + asp., ms)	ADS	53.11	15.72
		IDS	69.50	21.62
	*f*0 differnece (Hz)	ADS	39.53	27.17
		IDS	54.98	38.45
Lenis – Fortis	Fric. duration difference (ms)	ADS	9.45	7.92
		IDS	13.41	11.49
	Asp. Duration difference (ms)	ADS	56.36	18.41
		IDS	66.29	35.78
	Noise duration difference (fric. + asp., ms)	ADS	62.69	17.06
		IDS	78.51	31.72
	*f*0 differnece (Hz)	ADS	66.84	43.14
		IDS	114.93	80.73
Aspirated – Lenis	Fric. duration difference (ms)	ADS	5.77	4.09
		IDS	4.57	4.81
	Asp. Duration difference (ms)	ADS	13.33	10.25
		IDS	19.30	16.36
	Noise duration difference (fric. + asp., ms)	ADS	12.49	11.33
		IDS	18.99	16.00
	*f*0 differnece (Hz)	ADS	97.25	30.08
		IDS	139.00	54.12

Table 2. Mean differences in acoustic properties among the three sound pairs in ADS and IDS register.

was not greater in IDS (β = −1.12, t = −0.87) Removing the interaction effect significantly reduced model fit ($\chi^2(2) = 45.70$, $p < .001$), suggesting that *f*0 cue was enhanced selectively for lenis-aspirated contrast.

The estimated noise duration (Fric. + Asp.) was shorter for fortis (β = −53.42, t = −28.04) and longer for lenis (β = 11.84, t = 6.13) than aspirated. The durational difference was greater in IDS (β = −15.47, t = −5.65) for fortis-aspirated but no change was observed for lenis-aspirated contrast (β = 1.45, t = 0.51). Removing the interaction effect significantly reduced model fit ($\chi^2(2) = 44.92$, $p < .001$), supporting that the combined durational cue was enhanced selectively for fortis-aspirated contrast in IDS. Similar patterns were observed with both Fric. and Asp. duration measure.

3.2. Individual differences in how mothers speak the three-way affricates: modern way vs. the traditional way

Lastly, we examined the variability in the production of durational cues and *f*0 cue among these mothers.

According to Lee [2], the estimated noise duration differences between the aspirated and lenis categories were about 10ms and 75 Hz (cf. Fig. 1 in Lee [2]) for the estimated *f*0 differences among speakers in the younger age group. Based on the estimated average differences in the two cues produced by each mother, we categorized them into two groups: (1) those who produce the cue "the modern way" and (2) those that

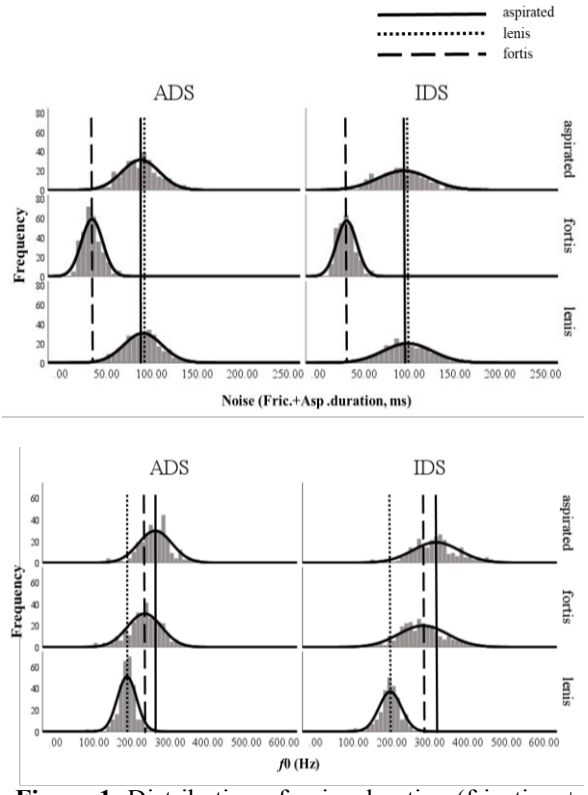


Figure 1: Distribution of noise duration (frication + aspiration) and *f*0 cues in three affricate categories (top panel: aspirated, middle panel: fortis, & bottom panel: lenis) produced by Korean mothers in ADS (left panels in each) and IDS (right panels in each). Vertical lines indicate mean values.

produce the cue "the traditional way." The criterion for the noise duration for this division was 10ms: if the mothers on average produced a noise duration difference shorter than 10ms, they were considered as producing the cue "modern way" but as "traditional way" if the difference was greater than 10ms. A similar approach was applied to the use of the *f*0 cue, so using the 75Hz as the criterion, those mothers that produced the *f*0 difference between aspirated and lenis greater than 75Hz were considered to produce the cue "modern way" and vice versa for other mothers.

Among 35 mothers in ADS speech (1 mother did not record ADS), 20 mothers (67%) produced VOT in a modern way, whereas 15 of them still produced noise duration difference in the traditional way. For *f*0, 28 mothers (80%) produced the difference in line with the recent change, but 7 of them (20%) still produced an *f*0 difference of less than 75Hz, following the traditional pattern. Mothers who produced both cues, reflecting the complete diachronic change were 46% (see Table 3). The mean and standard deviation of mothers' cues split by this grouping are summarized in Table 3.

In IDS, among 31 mothers (1 mother did not record IDS, 4 mothers did not record either lenis or

	ADS AspLenis Noise duration (ms)		ADS AspLenis *f*0 (Hz)		IDS AspLenis Noise duration (ms)		IDS AspLenis *f*0 (Hz)	
	modern	traditional	modern	traditional	modern	traditional	modern	traditional
Mean	4.55	23.23	107.15	55.33	5.85	25.77	149.88	54.72
SD	3.77	8.68	22.01	15.8	3.32	15.24	44.92	19.34

Table 3. Mean and SD in noise duration and *f*0 of mothers who produced the acoustic cues in modern versus traditional way in ADS and IDS.

aspirated words), 10 mothers (32%) produced noise duration in the modern way, whereas 21 of them (68%) still produced noise duration difference in the traditional way (Table 4). For *f*0, 27 mothers (87%) produced the difference in line with the recent change, but 4 of them (13%) still produced an *f*0 difference of less than 75Hz, following the traditional pattern. Mothers who produced both noise duration and *f*0 reflecting the complete diachronic change were about 29% (n = 9). Taken together, we observed that more mothers produced longer noise duration (considered traditional way) in IDS than ADS (21, 15, respectively, fisher's exact test, p = .051). However, the group variation on the *f*0 cue did not differ much between the two registers (fisher's exact test, p = .659).

Overall, the cue production patterns revealed that not all mothers today are faithful in reflecting the shifted cue pattern, especially for noise duration in IDS.

Diachronic Change		Speech Register ADS	IDS	Total
Traditional	Count	3	3	6
	% of total	5%	5%	9%
Modern	Count	16	9	25
	% of total	24%	14%	38%
Mixed	Count	16	19	35
	% of total	24%	29%	53%
Total	Count	35	31	66
	% of total	53%	47%	100%

Table 4. The distribution of mothers in reflecting the diachronic change in two acoustic cues (noise duration & f0): traditional (no reflection of cue change in both acoustic cues), modern (reflecting change in both cues), & mixed (one cue reflecting the change, the other preserving traditional cue pattern).

4. DISCUSSION

The results of our acoustic analyses showed that native Korean-speaking mothers today provide their young infants with affricate phonetic input that largely reflects the phonological changes that have occurred in the past several decades. The input distribution patterns were overall very similar between mothers' speech to infants as well as to other adults.

These suggest that, unlike in the past, durational cues may no longer play key roles in discriminating between lenis and aspirated as young Korean learners derive these categories, like Korean-speaking adults [1].

However, we observed that some mothers still tried to supply the durational cue difference for the contrast as suggested in our examination of individual differences among mothers (see a similar report on Korean stop input patterns, [7]). This raises an interesting possibility that their infant learners might diverge in terms of relying on different cues, given the strong relationship between children's production and their mothers' quality of production was observed in a previous study [14]. The infants of mothers who provided salient durational and *f*0 cues might also develop sensitivity to this cue, whereas those infants whose mothers consistently relied on a single cue might develop sensitivity to that cue alone. Future studies can illuminate this by testing Korean-learning infants' developing sensitivities to these cues and how they weigh them as they derive these phoneme categories.

Importantly, the specifically relevant cue parameter for category distinction was selectively enhanced in mothers' speech to their infants. In IDS, mothers produced larger differences in *f*0 between lenis and aspirated. By comparison, the durational difference was enlarged for fortis and other categories but not for lenis and aspirated contrast in IDS. These were consistent with many previous reports on the enhanced cue distinctions in IDS than ADS (e.g., [8]). These results suggest that early maternal phonetic input may scaffold the early development of native phonology.

5. ACKNOWLEDGMENTS

We would like to thank Yoonsung Kim, Jisoo Oh, and Gayoung Kim for their tremendous help in this project. This work has been supported in part by a Korean National Research Foundation grant to Y. Choi (NRF2014R1A1A2054072), by Japanese Society for Promotion of Science Grant-in-Aid for Scientific Research (17K13458) to H. Hwang, S(16H06319), Specially Promoted Research (20H05617) and MEXT Grant-in-aid for Innovative Areas #4903, 17H06382 S(16H06319) to R. Mazuka.

6. REFERENCES

[1] J. Schertz and Y. Kang, "Phonetic cue competition within multiple phonological contrasts: Perception of Seoul Korean sibilants," *Korean Linguist.*, vol. 18, no. 1, pp. 1–17, 2022.

[2] H. Lee, "Consistent sound change between stops and affricates in Seoul Korean within and across individuals: A diachronic investigation," *J. Acoust. Soc. Am.*, vol. 140, no. 6, pp. EL491–EL496, Dec. 2016, doi: 10.1121/1.4971203.

[3] H. Bang, M. Sonderegger, Y. Kang, M. Clayards, and T. Yoon, "The effect of word frequency on the timecourse of tonogenesis in Seoul Korean," in *Proceedings of the 18th International Congress of Phonetic Science (ICPhS2015)*, 2015.

[4] Y. Kang, "Voice onset time merger and development of tonal contrast in seoul korean stops: A corpus study," *J. Phon.*, vol. 45, no. 1, pp. 76–90, 2014, doi: 10.1016/j.wocn.2014.03.005.

[5] M. Kim, "Correlation between VOT and F0 in the Perception of Korean Stops and Affricates," in *INTERSPEECH 2004 - ICSLP (8th International Conference on Spoken Language Processing), Jeju*, 2004, pp. 1–6.

[6] D. J. Silva, "Acoustic evidence for the emergence of tonal contrast in contemporary Korean Acoustic evidence for the emergence of tonal contrast in contemporary Korean *," *Phonology*, vol. 23, pp. 287–308, 2006, doi: 10.1017/S0952675706000911.

[7] Y. Choi, M. Nam, R. Mazuka, H.-K. Hwang, and N. Yamane, "Korean mothers' production of laryngeal stops to their infants as compared with adults in the context of tonogenesis," in *the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019, pp. 3260–3264.

[8] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao, "Mothers' speech clarity and infants' discrimination An association between mothers' speech clarity and infants' speech discrimination skills," *Dev. Sci.*, vol. 6, no. 3, pp. F1–F10, 2003.

[9] C. Narayan and T.-J. Yoon, "VOT and f0 in Korean infant-directed speech," *Can. Acoust.*, vol. 39, no. 3, pp. 152–153, 2011.

[10] E. Ko, "Evaluating the clear speech hypothesis of child-directed speech: The case of Korean stop consonants," presented at the The Korean Society of Speech Sciences (KSSS 2016), Seoul, Korea, 2016.

[11] A. Cristià, "Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3271–3280, May 2011, doi: 10.1121/1.3562562.

[12] M. Nam, Y. Choi, and R. Mazuka, "Caregiver phonetic input supports infants' learning of sound categories: The role of voice onset time distribution in natural phonological acquisition," presented at the in the 46th Boston University Conference on Language Development (BUCLD 2021), Boston, US, 2021.

[13] K. M. Hartman, N. B. Ratner, and R. S. Newman, "Infant-directed speech (IDS) vowel clarity and child language outcomes," *J. Child Lang.*, vol. 44, no. 5, pp. 1140–1162, Sep. 2017, doi: 10.1017/S0305000916000520.

[14] J. H. Sim and B. Post, "Variation in quality of maternal input and development of coda stops in English-speaking children in Singapore," *J. Child Lang.*, vol. 49, no. 6, pp. 1147–1172, Nov. 2022, doi: 10.1017/S0305000921000593.

[15] Xu, Y. 2013. ProsodyPro — *A Tool for Large-scale Systematic Prosody Analysis*. In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France. 7-10.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4." arXiv, Jun. 23, 2014. Accessed: Jan. 07, 2023. [Online]. Available: http://arxiv.org/abs/1406.5823

[17] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *J. Mem. Lang.*, vol. 59, no. 4, pp. 390–412, Nov. 2008, doi: 10.1016/j.jml.2007.12.005.

A LONGITUDINAL STUDY OF PAUSES, INTERPAUSAL UNITS AND CLAUSES IN INFANT-DIRECTED SPEECH

Anna Kohári and Katalin Mády

Hungarian Research Centre
kohari.anna@nytud.hu, mady.katalin@nytud.hu

ABSTRACT

Infants are able to perceive prosodic features, including pauses, and to detect phrase and clause boundaries from fluent speech. Yet, studies investigating the pauses in infant-directed speech (IDS) with respect to syntactic or prosodic structure are relatively sparse and mostly limited to a given age of the child. Here we analyze 14 mothers' semi-spontaneous narratives in Hungarian IDS compared to adult-directed speech (ADS). The longitudinal analysis showed that the adults tended to talk slower to their 6 and 18 month-old children than to other adults, and that pauses were more frequent in IDS as compared to ADS. Furthermore, the ratio between the frequency of pauses at clause boundaries and within clauses was found to be higher in IDS than in ADS. We also found that in speech directed to 18-month-old infants the interpausal units and clauses tended to consist of less syllables than in ADS.

Keywords: infant-directed speech, pausing, longitudinal study

1. INTRODUCTION

Infant-directed speech (IDS) is characterized by lexical and syntactic simplification, repetition, higher fundamental frequency, and slower speech rate [1] compared to adult-directed speech (ADS). Besides these widely known characteristics, there has been little discussion on pauses in IDS and their relation to syntactic or prosodic structure, although perception tests revealed that infants preferred listening to speech where pauses occur between clauses instead of within clauses [2, 3]. Since the occurrence of pauses could provide crucial cues for infants to perceive clause and phrase boundaries, facilitating the acquisition of linguistic structures [3], in the present work we investigate the complex interplay between the acoustic features (articulation rate, pauses, and length of interpausal units (IPUs)) related to pausing and the clause boundaries.

Although numerous studies have demonstrated that adults talk slower to infants [1, 4, 5, 6, 7], others did not find speech or articulation rate differences between the two registers [8, 9]. A plausible explanation behind these seemingly contradictory results may be that the children's age influences the tempo of IDS. This assumption is supported by the findings of Narayan and McDermott [6], who observed that caregivers gradually increased their speech rate as their infant addressees developed, and by the second year of the children's age their IDS rate approximated that of their ADS.

Several studies have established that pauses tend to last longer in spontaneous IDS than in ADS [1, 5, 10, 11]. However, in an experiment where the participants were asked to describe given objects in a semi-spontaneous manner to adults and to infants, the pause durations in the two registers did not differ significantly [12]. Pause durations in the context of syntactic or prosodic structure received less attention, yet, it has been demonstrated for both registers that the pauses at utterance, clause, or phrase boundaries are typically longer than those within speech units [13, 14, 15]. The two registers were reported to differ in terms of pause durations between independent clauses, which were found to last longer in IDS than in ADS [14]. Furthermore, the aforementioned difference between pause durations at phrase boundaries and at non-boundaries was more pronounced in IDS than in ADS [15].

The available data on the age-related differences of pause durations is also quite sparse. Stern [10] concluded that pauses tended to last longer in American English speech directed towards newborns than to 4-, 12-, and 18-month-old infants, however, no such trend was found for later ages of the children. Kondaurova [14] also found that pauses between clauses tend to be longer when addressing 6-month-olds than 12-month-olds.

Not only the duration, but also the frequency of pauses may be different between the two registers. Several studies concluded that pauses occurred more often in IDS than in ADS [7, 12]. It has also been reported that above 90% of the pauses within IDS were detected between utterances, and only a small fraction in them. This ratio is in huge contrast with the result from ADS, where only 50–68% of the pauses occurred in inter-utterance positions [1, 4, 16].

In our longitudinal study, we address the pausing- and timing-related peculiarities of IDS at different ages of the infants and analyze the extent to which the occurrence of pauses coincides with clause boundaries.

2. MATERIALS AND METHODS

2.1. Participants

Recordings from 14 primiparous mothers (i.e., mothers giving birth for the first time), native speakers of Hungarian were used for our analysis. The ages of the participants ranged from 24 to 41 years (M = 31.6, SD = 5.7), and all of them earned a high school diploma or college degree. The properties of IDS change with the child's age [6], therefore besides investigating the mothers' speech to their newborns, we also repeated the experiments at 6 months of the babies' age when they reached the preverbal stage, and later when the children were 18 month old and could communicate actively.

2.2. Materials

The speakers were asked to tell a fairytale about pixies based on a story book with a sequence of pictures, using their own words, first to an (adult) experimenter (i.e., producing ADS) and then to their own child (IDS). During the ADS recording sessions another experimenter was playing with the children in the same room. Besides the pictures, the pages also contained certain scripted sentences, which the mothers had to incorporate to their otherwise spontaneous storytelling to ensure that the content of the speech was approximately identical in both registers. For the analyses of the present study only the spontaneous speech parts and their pauses were investigated. The first recording session with each mother was conducted in a silent separated room at the Birth Centre of the Military Hospital in Budapest a few days after the childbirth. The experiments were later repeated in the baby lab of the Research Centre for Natural Sciences.

2.2. Analysis

Recordings were annotated [17], and the boundaries of interpausal units and clauses were labelled manually using Praat 6.1.08 [18]. For our analysis a clause was defined as a unit containing a subject and a related predicate with their modifiers, while an incomplete clause lacking a predicate was merged with the nearest and semantically connected complete clause, and the two were then considered as one unit. Articulation rates were acquired by measuring the duration of IPUs and clauses (in seconds) and dividing their syllable counts by the obtained durations. The minimum duration above which a pause was considered for the analysis was set to 30 ms, the typical duration of certain speech sounds in spontaneous speech [19]. The duration of the closure phases of voiceless plosives and affricates occurring before pauses was taken to be 50 ms [19]. Since pauses shorter than 300 ms are traditionally ignored in the IDS-related literature [1, 5, 10], we also repeated the analysis of the investigated parameters, setting the lower limit for the duration of the considered pauses to 300 ms. However, this modification did not affect our findings, except for a single aspect: when only the pauses longer than 300 ms were considered, pause durations within the clauses did not exhibit detectable dependence on the infant's age (whereas applying the 30 ms pause limit they did, see Results). Beside the durations of the pauses their frequencies were also calculated. On the one hand, we determined the number of pauses (normalized to 100 syllables) in the speech of each mother in both registers as a function of the infant's age. On the other hand, the frequencies of pauses at clause boundaries and within clauses were evaluated separately (for each mother, infant age, and register). Since disfluencies may also affect pause durations and articulation rates, and as disfluencies are more common in ADS than in IDS [20, 21, 22], our analysis was restricted to clauses – and the pauses therein – lacking disfluencies (false starts, hesitations). The ratio of pauses between clauses was found to be unaffected by whether we included the disfluencies in the analysis, or omitted them. 4227 IPUs and 3301 clauses were investigated.

Linear mixed effect models were fitted for the dependent variables (i.e., articulation rate; the logarithm of the pause durations) using the R 3.6.2 software [23] with the *lmerTest* package [24]. For the syllable counts, a negative binomial regression mixed model was applied. Register, age, and their interactions were modeled as fixed factors, and subjects as random factors (random slopes and intercepts) in each model. For articulation rate measurements, the number of syllables was also treated as a fixed factor of the model. The maximal models were optimized using stepwise backward elimination [25]. We performed Tukey post-hoc tests utilizing the *emmeans* package [26]. Marginal (R^2m) and conditional R^2 (R^2c) estimates were determined using the *MuMIn* package [27]. For the speaker- and register-wise pause frequencies, after evaluating whether the test assumptions were met, repeated measures ANOVA was performed [28].

3. RESULTS

3.1. Articulation rate

Figure 1 shows the results of articulation rate of IPUs. The analysis revealed that the register (ADS or IDS) in interaction with the baby's age affected the articulation rate ($p < 0.05$, R^2m = 0.14, R^2c = 0.28).

The Tukey post-hoc test showed that IDS tended to be slower than ADS at 6 and 18 months of the infant's age. (6m: β = 0.21, SE = 0.07, z = 2.97, p = 0.003, 18m: β = 0.22, SE = 0.07, z = 2.97, p = 0.003). However, immediately after childbirth the two registers did not exhibit articulation rate differences. With increasing syllable number of the IPU, articulation rate also increased (β = 0.09, SE = 0.03, t = 27.3, p < 0.001), but the model did not show an interaction between the syllable count and the babies' age or the register. Thus, the articulation rate of IPUs was lower when talking to 6 and 18 month-old infants compared to ADS, even after adjusting to the differences between the syllable counts. Repeating the same analysis on clauses instead of IPUs yielded the same results.

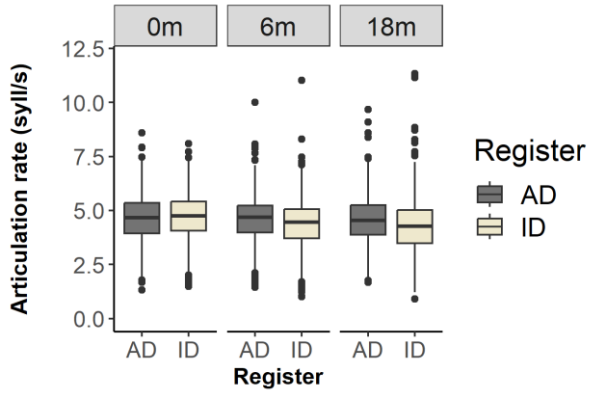


Figure 1: The articulation rate of IPUs for registers and the babies' age.

3.2. Number of syllables

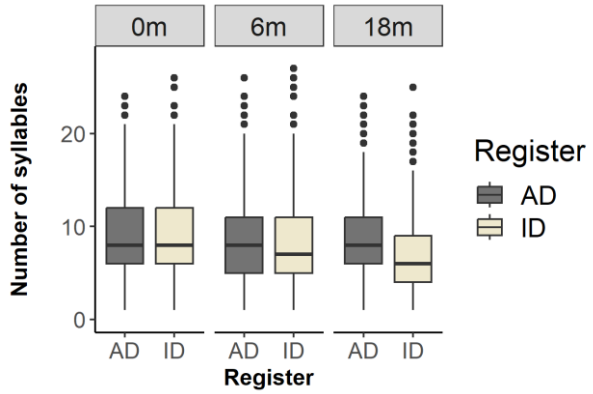


Figure 2: The number of syllables per clauses for registers and the babies' age.

The number of syllables in IPUs and clauses showed a peculiar relationship with the register and the baby's age. The Tukey post-hoc test indicated that ADS exhibited more syllables than IDS both for IPUs and clauses in the experiments conducted 18 months after childbirth (IPU: β = 0.08, SE = 0.04, z = 2.20, p = 0.028; clause: β = 0.17, SE = 0.03, z = 5.96, p < 0.001) but no such effect was found in the other two investigated ages of the babies. (IPU: β = 0.08, SE = 0.04, z = 2.20, p = 0.028; clause: β = 0.17, SE = 0.03, z = 5.96, p < 0.001). Although for the syllable counts in IPUs, the model's explanatory power was weak (R^2m = 0.01, R^2c = 0.07), it was found to be moderate for the case of clauses (R^2m = 0.17, R^2c = 0.19), which may well be the consequence of the fact that here the presence or absence of pauses could also be incorporated in the model as a further independent variable.

3.3. Pause duration

Besides register and the babies' age, the position of pauses – whether they occur within or between clauses – constituted a further fixed factor for modeling the effects that influence their durations. The model revealed an interaction between the babies' age and the pause position (p < 0.05, R^2m = 0.22, R^2c = 0.24). The duration of pauses was significantly longer between clauses than within clauses for all studied ages of the babies in both registers. The analysis also showed that pauses within the clauses were longer in the 0-month experiments than when the babies were 6 and 18 months old (0m-6m: β = 0.27, SE = 0.07, z = 4.19, p < 0.001; 0m-18m: β = 0.36, SE = 0.07, z = 5.14, p < 0.001). We found no effect of the register (not even in interaction with the other factors) on the durations of pauses.

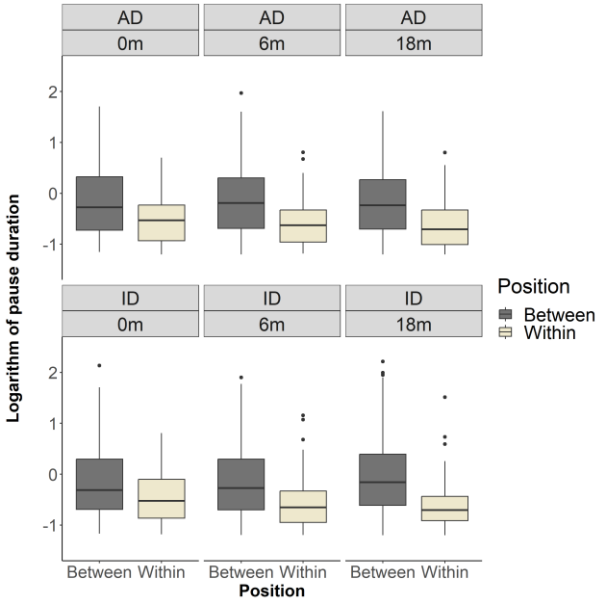


Figure 3: Pause durations for registers and the babies' age between and within clauses.

3.4. Frequency of pauses

The occurrence of pauses (pauses per 100 syllables) was evaluated for each speaker concerning the register and the babies' age. The results indicate a

significant interaction between the register and the babies' age (RM ANOVA, $F(1, 24) = 5.51$, $p = 0.01$, $\eta^2 = 0.32$). Tukey post-hoc tests revealed that pauses occurred more often in IDS than in ADS, considering the experiments at 6 and 18 months of the babies' age ($p < 0.05$). However, no such difference between the registers appeared in the 0-month recordings. Post-hoc tests also showed that the pause frequencies in ADS remained unchanged throughout the three recording sessions. To compare pause frequencies between and within the clauses, we calculated the ratio (percentage) of pauses falling between clauses compared to the total number of pauses for each speaker, register, and age. We found that this ratio was higher in IDS than in ADS ($F(1, 12) = 14.99$, $p = 0.002$, $\eta^2 = 0.56$). However, there was no effect of babies' age or interaction with babies' age.

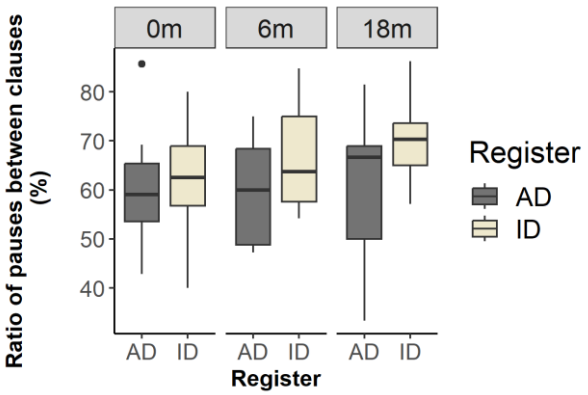


Figure 4: The ratio (percentage) of pauses falling between clauses compared to the total number of pauses for registers and the babies' age.

4. DISCUSSION AND CONCLUSION

In this study, we analyzed the pausing and timing characteristics of infant-directed speech (IDS) and explored the connections between pausing and clause boundaries. Evaluating our records of (semi-) spontaneous narrative IDS and ADS, we found – in agreement with previous research (e.g. [1, 5, 6, 7]) – that clauses and inter-pausal units (IPUs) were both slower in IDS than in ADS, but only in the recording sessions at 6 and 18 months of the babies' age. The observation that this difference did not appear in the 0-month experiments – conducted only a few days after childbirth – could possibly be explained by the special environmental (hospital) conditions and by the lack of actual mother-child interaction in the cases where the infant was asleep during the session [9].

The analyses of the syllable counts in clauses and IPUs revealed that we could observe a difference from ADS only in the sessions with 18-month-old infant addressees in terms of the number of syllables per speech unit. Interestingly, for Mandarin Chinese IDS, it has been reported that speakers tended to use IPUs consisting of less syllables already at two months of the babies' age [11]. The difference may be related to the experimental conditions: their recording sessions were conducted in a way that the babies and their parents were playing freely, whereas in our case, the speech of the subjects was controlled by the arrangement of the experiment (storytelling based on pictures). As for the latter speech situation, it appears that in the earlier ages of the babies, speakers seldom used pauses to isolate and emphasize words and phrases [cf. 29]. Future research involving pragmatic analysis – beyond the scope of the present study – is needed to explore the properties of the shorter IPUs and clauses in the IDS addressing 18-month-olds, who are already capable of producing words from their rapidly increasing vocabulary.

The durations of the pauses did not exhibit differences between the two registers, seemingly contradictory results from earlier research [1, 5, 10]. However, another analysis in which the content was controlled by the experiment setting also did not find any difference between German IDS and ADS in terms of pause durations [12]. Thus, those results and the ones presented in the present paper imply that the speech situation and certain resulting speech planning procedures may have an effect on the characteristics of IDS, including the durations of pauses. However, the duration of pauses was found to be longer between clauses than within clauses for all studied ages of the babies in both registers. Thus, though the durations of pauses were not exaggerated at the clause boundaries in IDS, their lengths still served as boundary markers, and could facilitate the infants' segmentation from fluent speech.

Pauses were more frequent in IDS as compared to ADS at 6 and 18 months of the infants' ages, in agreement with earlier results [7, 12]. Furthermore, the ratio between the frequency of pauses at clause boundaries and within clauses was found to be higher in IDS than in ADS (in agreement with previous research on pauses following utterances in other languages [1, 16]). Therefore, the occurrence of pauses coincided with clause boundaries more often in IDS than in ADS in all three investigated ages of the babies, since pauses may provide crucial cues for the infants to perceive clause boundaries facilitating the acquisition of linguistic structures [3]. It is to be emphasized, however, that although here we have focused merely on a single acoustic cue – pauses – detecting phrase-final lengthening and f0 movements may also help the babies parse the speech. Furthermore, extending the analysis to intonational phrases and their boundaries may provide an even more comprehensive picture of the relations between syntax and prosody in IDS.

5. ACKNOWLEDGEMENTS

We would like to thank Katalin Pirsel and Veronika Harmati-Pap for the manual transcriptions and annotations. This work was funded by the National Research, Development and Innovation Fund (Grants no. NKFI-115385, NKFI-135038 and NKFI-134775).

5. REFERENCES

[1] Fernald, A. Simon, T. 1984. Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 20(1), 104.

[2] Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., Kennedy, L. 1987. Clauses are perceptual units for young infants. *Cognition* 26(3), 269–286.

[3] Johnson, E. K. Seidl, A. 2008. Clause segmentation by 6-month-old infants: A crosslinguistic perspective. *Infancy* 13(5), 440–455.

[4] Broen, P. A. 1972. *The Verbal Environment of the Language-Learning Child*. American Speech and Hearing Association.

[5] Fernald, A., Taeschner, T., Dunn, J., Papousek, Mechthild, de Boysson-Bardies, B., Fukui, I. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* 16(3), 477–501.

[6] Narayan, C. R. McDermott, L. C. 2016. Speech rate and pitch characteristics of infant-directed speech: Longitudinal and cross-linguistic observations. The *Journal of the Acoustical Society of America* 139(3), 1272–1281.

[7] Martin, A., Igarashi, Y., Jincho, N., Mazuka, R. 2016. Utterances in infant-directed speech are shorter, not slower. *Cognition* 156, 52–59.

[8] Han, M., de Jong, N.H., Kager, R.. 2018. Infant-directed speech is not always slower: Cross-linguistic evidence from Dutch and Mandarin Chinese. *Proc. of the 42nd annual Boston University Conference on Language Development*, Somerville. 331–344.

[9] Mády, K., Reichel, U., Szalontai Á., Kohári, A., Deme, A. 2018. Prosodic characteristics of infant-directed speech as a function of maternal parity. *Proc. of the 9th International Conference on Speech Prosody*, Poznań, 294–298.

[10] Stern, D. N., Spieker, S., Barnett, R. K., MacKain, K. 1983. The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language* 10(1), 1–15.

[11] Grieser, D. L. Kuhl, P. K. 1988. Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. Developmental Psychology 24(1), 14–20.

[12] Menn, K. H., Michel, C., Meyer, L., Hoehl, S., Männel, C. 2022. Natural infant-directed speech facilitates neural tracking of prosody. *NeuroImage* 251, 118991.

[13] Soderstrom, M., Blossom, M., Foygel, R., Morgan, J. L. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. Journal of Child *Language* 35(4), 869–902.

[14] Kondaurova, M. V., Bergeson, T. R. 2011. The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech. *Journal of Speech, Language, and Hearing Research* 54, 740–754.

[15] Ludusan, B., Cristia, A., Martin, A., Mazuka, R., Dupoux, E. 2016. Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America* 140(2), 1239–1250.

[16] Fisher, C. Tokura, H. 1996. Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development* 67(6), 3192–3218.

[17] Harmati-Pap, V., Vadász, N., Kas, B., Tóth, I. 2021. Anyai dajkanyelvi narratívák lexikai és szintaktikai jellemzőinek longitudinális vizsgálata. *Beszédtudomány - Speech Science* 2, 207–242.

[18] Boersma, P. Weenink, D. 2019. Praat: doing phonetics by computer. Computer program.

[19] Trouvain, J., Werner, R. 2022. A phonetic view on annotating speech pauses and pause-internal phonetic particles. In: Schwarze, C. Grawunder, S. (eds.) *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion: Konzepte, Probleme, Lösungen*., 55–73.

[20] Björkenstam, K. N., Wirén, M., Eklund, R.. 2013. Disfluency in child-directed speech. *Proc. of Fonetik 2013, the XXVIth Swedish Phonetics Conference*, Linköping, 57–60.

[21] Soderstrom, M., Morgan, J. L. 2005. Disfluency in speech input to infants? the interaction of mother and child to create error-free speech input for language acquisition. *Proc. of the DiSS-2005*, Aix -en-Provence, 157–162.

[22] Newport, E. L., Gleitman, H., Gleitman, L. R. 2020. Mother, I'd rather do it myself. Lidz, L. (ed.), *Sentence First, Arguments Afterward: Essays in Language and Learning*. Oxford University Press, 141–177.

[23] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

[24] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82, 1–26.

[25] Winter, B. 2019. *Statistics for linguists: An introduction using R*. Routledge.

[26] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2022. Package 'emmeans'. Estimated marginal means, aka least-squares means. https://cran.r-project.org/web/packages/emmeans/.

[27] Bartoń, K. 2022. MuMIn: Multi-Model Inference. https://cran.r-project.org/package=MuMIn.

[28] Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M., S. 2022. Package 'afex'. https://cran.r-project.org/web/packages/afex/

[29] Laing, C. E., Vihman, M., Keren-Portnoy, T. 2017. How salient are onomatopoeia in the early input? A prosodic analysis of infant-directed speech. *Journal of Child Language* 44(5), 1117–1139.

EVIDENCE FOR PRESERVED PHONETIC SENSITIVITY TO LEXICAL TONES IN CHINESE INTERNATIONAL ADOPTEES

Andrea C.-L. Chang[1], Tyler K. Perrachione[1], & Sung-Joo Lim[1,2]

[1]Department of Speech, Language, and Hearing Sciences, Boston University, USA
[2]Department of Psychology, Binghamton University, USA
sungjoo@binghamton.edu

ABSTRACT

During infancy, children gain perceptual sensitivity to phonetic contrasts in their native language while losing sensitivity to foreign contrasts. It remains unclear whether traces of language exposure in infancy remain after sudden and permanent interruption of heritage language acquisition, as in international adoption. We examined sensitivity to lexical tones in adults (age 18-23 years) with a history of international adoption (age 3-18 months) whose heritage language was Chinese. We assessed phonetic, phonological, and lexical sensitivity to Chinese tones using tone contour identification and discrimination, auditory Stroop, and Chinese sentence identification tasks, and the auditory brainstem frequency-following response (FFR). International adoptees had native-like FFR, and their tone discrimination and identification were intermediate between English and Mandarin control groups. However, higher-level processing of Chinese did not differ between adoptees and the English control group. Exposure to heritage languages in infancy may create traces of early phonetic environments that persist into adulthood.

Keywords: tonal language perception, speech perception, critical period, frequency-following response

1. INTRODUCTION

Infants become sensitive to the speech sounds of their native language within the first year of life. They exhibit enhanced sensitivities to native speech contrasts [4] while losing sensitivity to foreign-language speech contrasts [11]. Such "neural commitment" [3] to one's native-language speech contrasts suggests that early linguistic experience may have a significant impact on lifelong tuning of speech processing.

However, it remains unclear whether traces of language exposure from early infancy remain even after sudden and radical changes in the linguistic environment, as in international adoption. The existing evidence is equivocal as to whether early language experiences leave traces that persist across development [7,10]. Nevertheless, two recent EEG studies demonstrated that children adopted from China retain neural sensitivity to Mandarin lexical tones [6,9]. Based on this evidence from cortical recordings, it is possible that the impact of early language exposure to Mandarin also leaves persistent traces of neural sensitivity to the phonetic representations of lexical tones at the subcortical level (i.e., auditory brainstem) [2].

To this end, the present study examined phonetic, phonological, and lexical sensitivity to Mandarin Chinese lexical tones using a battery of behavioral tests in normal hearing adults who were adopted from China compared to native English and native Chinese controls. We also examined differences in the sensitivity of the subcortical auditory system in these groups via FFRs elicited by Mandarin lexical tones.

2. METHODS

2.1. Participants

We recruited three participant groups with different language backgrounds: English-speaking adults who had been adopted from China to the U.S. as children (n=9; mean age = 20 years), native Mandarin speakers (n=21; mean age = 22.2), and native American English speakers with no tone language experience (n=21; mean age = 21.2). On average, the adoptees arrived in the U.S. at 9.6 months of age (Table 1). Adoptees' heritage language cannot be known for certain and may have been Mandarin, Cantonese, or another dialect; six of the adoptees had some post-adoption exposure to Mandarin (usually in college). We quantified their Mandarin knowledge using the *Hanyu Shuiping Kaoshi* (HSK) Level 1 test [1].

All participants had normal hearing and provided written informed consent, approved by the Institutional Review Board at Boston University.

Subj.	Age (yr.)	Adoption age (mo.)	Mandarin Experience (yr.)	HSK score
P1	19	6	0	N/A
P2	21	3	0	N/A
P3	18	10	1	28/40
P4	21	6	6	32/40
P5	19	12	3	23/40
P6	20	12	11	39/40
P7	20	18	2	39/40
P8	23	8	2	14/40
P9	21	11	0	N/A

Table 1: International adoptees' demographics.

2.2. Behavioral tasks

Participants performed all behavioral tasks in a sound attenuated booth. All auditory stimuli were RMS normalized to 70 dB SPL.

2.2.1. Pitch-contour perception test (PCPT)

Participants performed a non-lexical pitch pattern identification task, which is designed to assess listeners' aptitude for learning lexical tones [8,12]. Five vowels (/a/, /i/, /o/, /e/, and /u/) recorded by four native American English speakers (2 females) were resynthesized to have three Mandarin tonal contours (level, rising, and falling) using the pitch synchronous overlap-and-add algorithm (PSOLA) [5]. On each trial, listeners heard one vowel and matched its synthesized pitch contour to a corresponding arrow icon using assigned keys on a response box. Participants completed 120 trials (4 speakers × 5 vowels × 2 repetitions). No feedback was provided.

2.2.2. Tone discrimination task

Participants performed an AX lexical tone discrimination task on pairs of Mandarin monosyllables and synthesized sinewave tones. On each trial, participants heard a pair of stimuli that differed in consonant, vowel, and speaker, and they judged whether the pitch contours of the pair were the same (e.g., mǎi/yě) or different (e.g., tīng/xià). The speech stimuli were recorded by one female and one male native Mandarin speaker. The sinewave tones were synthesized from the natural speech sounds using PSOLA.

Participants completed 192 trials (half with same and half with different tone contours) for both speech and non-speech stimuli. No feedback was provided.

2.2.3. Auditory Stroop task

To assess international adoptees' implicit semantic processing of their heritage language (whether as a result of latent knowledge or due to re-exposure in adulthood), we conducted an auditory Stroop task. On each trial, listeners heard one of the two words meaning "high" or "low" spoken in English, Mandarin, and Cantonese; each word was spoken in either an exaggeratedly high- or low-pitched voice by one male native speaker of each language. Participants judged whether the voice pitch was high or low on each trial, and we hypothesized that participants' familiarity with a language would be reflected in slower response times (semantic interference) when a word's meaning and its voice pitch were incongruent vs. congruent.

Participants performed 40 trials (20 congruent) per language condition. Corrective feedback was provided on each trial.

2.2.4. Sentence identification task

We conducted a sentence identification task to assess whether the adoptees retained enhanced familiarity for recognizing their heritage language (or had gained such familiarity from post-adoption exposure). In this task, participants heard sentences spoken in Mandarin, Cantonese, Japanese, Korean, or Arabic, and were asked to identify whether the sentence was "Chinese" or not. The sentences were taken from "The North Wind and Sun" passage recorded by four native speakers of each language (2 females, 2 males). Listeners performed a total of 80 trials (4 sentences × 4 speakers × 5 languages) presented in a random order. No feedback was provided.

2.3. Frequency-following response (FFR)

Four minimally contrastive Mandarin monosyllables (/yi1/ "clothing", /yi2/ "aunt", /yi3/ "chair", /yi4/ "easy"), recorded by one native Mandarin speaker, were used to elicit the FFR. In an electrically shielded, sound attenuated booth, participants passively listened to randomly presented syllables while watching a silent movie. The stimuli were delivered via Etymotic ER-1 insert earphones at 70 dB SPL.

A total of 6400 trials were collected from each participant (1600 trials per stimulus). All stimuli were presented at alternating polarities.

2.3.1. Data acquisition, preprocessing, and analyses

We measured FFR using EEG, recorded from a standard 10/20 layout of 32 active scalp electrodes (Biosemi ActiveTwo) at a sampling rate of 4096 Hz. Two reference electrodes were placed on the earlobes. Electrode offsets for all channels were maintained below 30 μV. EEG data were high-pass filtered at 70 Hz cutoff to isolate subcortical responses from cortical activity. Trials were removed if the activity range exceeded 35 μV. The remaining trials for each stimulus per participant (M ± SD = 1475 ± 133 trials) were averaged for further analyses.

For each participant, we extracted f_0 contours for the stimuli and FFR via autocorrelation using a 40-ms sliding window (3-ms steps). We quantified the pitch tracking accuracy of the FFR by computing stimulus-to-response correlation; this measure indexed how well the FFR tracked the f_0 of the stimulus.

3. RESULTS

3.1. Pitch-contour perception test

We used a generalized linear mixed-effects model to examine whether participants' language background had an impact on identifying pitch contours of vowels

(Fig. 1). Mandarin controls performed significantly better on the PCPT than did international adoptees (z = 2.85, p = 0.004); and the adoptees had significantly higher accuracy than the native English listeners (z = 2.00, p = 0.046).

Next, we examined whether the pitch identification accuracy of the adoptees depended on their age of adoption or Mandarin proficiency (HSK score). These correlations did not reveal any significant relationships (both $p > 0.57$).

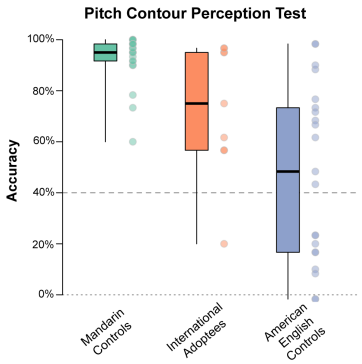


Figure 1: Pitch contour identification accuracy by participants in each group.

3.2. Tone discrimination performance

We used linear mixed-effects models to examine whether the participants in each group differed in their ability to discriminating spoken Mandarin lexical tones and their sinewave analogs (Fig. 2).

This analysis revealed significant main effects of group ($\chi^2(2) = 89.35$, $p < 0.001$) and stimulus type ($\chi^2(2) = 18.25$, $p < 0.001$), as well as a significant interaction between group and stimulus type ($\chi^2(2) = 84.99$, $p < 0.001$). The linear model with pairwise contrast coding revealed that Mandarin controls performed significantly better than the international adoptees overall ($z = 6.24$, $p < 0.001$). Mandarin controls also had a significant advantage for discriminating speech vs. sinewave tones compared to the international adoptees ($z = 5.77$, $p < 0.001$). Furthermore, the adoptees performed significantly better than native English group overall ($z = 4.05$, $p < 0.001$), and likewise had a relative advantage for discriminating speech vs. sinewave tones ($z = 2.06$, $p = 0.040$).

We also examined whether the adoptees' tone discrimination was related to their age of adoption or HSK score. While HSK scores were positively correlated with speech tone discrimination ($r = 0.78$, $p = 0.01$), this relationship was not found for sinewave tones ($p = 0.72$). Additionally, age of adoption was marginally correlated with speech tone discrimination ($r = 0.63$, $p = 0.068$), but was not related to discrimination of sinewave tones ($p = 0.42$).

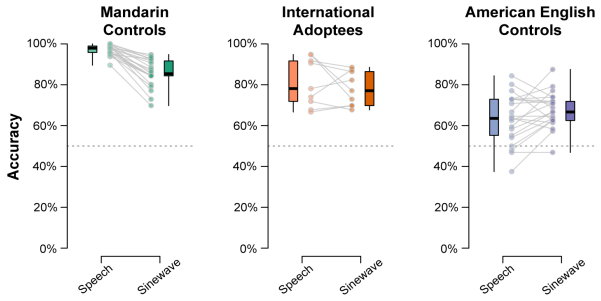


Figure 2: Accuracy discriminating Mandarin speech and sinewave tones, by group.

3.3. Auditory Stroop task performance

Table 2 shows the amount of interference by group and language condition when the word meaning and vocal pitch were incongruent (i.e., "high" said with a low-pitched voice) compared to when they were congruent (i.e., "high" said with a high-pitched voice).

	Mandarin Controls	International Adoptees	Native English
Mandarin	55 ±110	31 ± 72	16 ± 73
Cantonese	67 ± 103	–7 ± 46	–13 ± 70
English	79 ± 176	54 ± 73	102 ± 79

Table 2: Response time (RT) differences (mean ± s.d. in ms) in the incongruent vs. congruent trials in each language condition and participant group.

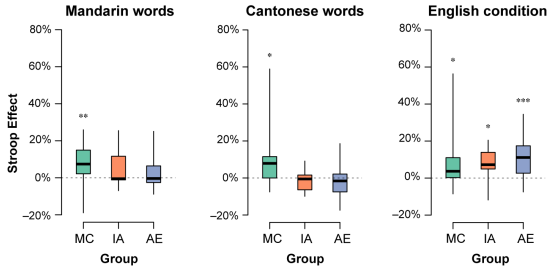


Figure 3: Semantic interference in each language condition by group. Auditory Stroop effect: (mean incongruent RT – mean congruent RT) ÷ mean congruent RT.

We performed a separate linear mixed-effects model on RT for incongruent vs. congruent trials for each language condition and each participant group (Fig. 3). When identifying the voice pitch of Mandarin words, Mandarin controls were significantly slower to respond to incongruent vs. congruent trials ($t_{772} = 3.36$, $p < 0.001$), but no significant interference (Stroop effect) was found in the other two groups (both $p \geq 0.23$). Likewise for the Cantonese words, only Mandarin controls exhibited a significant Stroop effect ($t_{771} = 3.30$, $p = 0.001$). In the English condition, all groups exhibited significant Stroop effects (Mandarin controls: $t_{748} = 2.42$, $p = 0.02$; Adoptees: $t_{331} = 2.56$, $p = 0.01$, native English: $t_{777} = 5.13$, $p < 0.0001$).

The magnitude of international adoptees' Stroop effect for Mandarin words was not correlated with their age of adoption or HSK score (both $p > 0.17$).

3.4. Sentence identification performances

As expected, Mandarin controls were highly accurate at identifying Chinese sentences (i.e., Mandarin and Cantonese) vs. other languages (Fig. 4). The linear mixed-model analysis revealed that Mandarin controls exhibited significantly higher rates of identifying Mandarin and Cantonese sentences as Chinese compared to both the adoptee ($z = 2.07$, $p = 0.038$) and native English groups ($z = 3.67$, $p < 0.001$). In contrast, the international adoptees and native English groups misidentified Korean and Japanese sentences as Chinese significantly more often than did Mandarin controls (both $p < 0.001$).

We found that correct identification of Chinese sentences in the adoptee group exhibited a strong relationship to the HSK score ($r = 0.83$, $p = 0.0052$), but not age of adoption ($p = 0.25$).

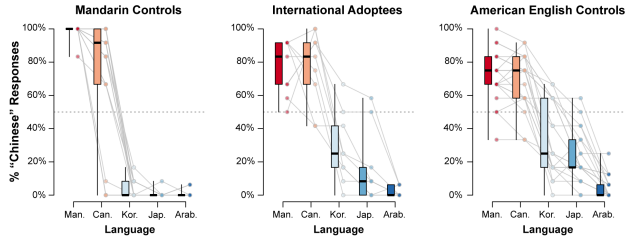


Figure 4: Percentage of sentences identified as Chinese when hearing sentences from five different languages.

3.5. FFR pitch tracking accuracy

Table 3 lists average FFR pitch tracking accuracy by contour and group (Fig. 5). We conducted independent samples t-tests to determine pairwise group differences in the pitch tracking accuracies. We found that Mandarin controls and international adoptees exhibited similar pitch tracking accuracies ($p > 0.06$). Compared to the native English group, significantly higher FFR pitch tracking accuracy for T1 was found in the Mandarin control ($t_{39.8} = 2.08$, $p = 0.047$). Interestingly, we found that compared to native English listeners, pitch tracking accuracy for T4 was significantly higher in the adoptees ($t_{18.3} = 2.30$, $p = 0.034$).

	Mandarin Controls	International Adoptees	English Controls
T1: Level	0.20 ± 0.28	0.25 ± 0.34	0.04 ± 0.17
T2: Rise	0.61 ± 0.47	0.92 ± 0.15	0.67 ± 0.49
T3: Dip	0.77 ± 0.31	0.75 ± 0.33	0.70 ± 0.35
T4: Fall	0.84 ± 0.27	0.96 ± 0.06	0.82 ± 0.25

Table 3 Mean FFR pitch tracking accuracy (Pearson's *r*) in the participant groups for each Mandarin tone contour.

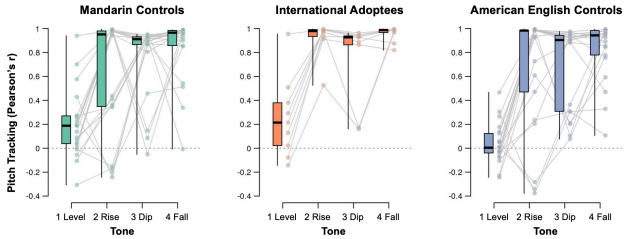


Figure 5: Participants' FFR pitch tracking accuracy by group for each Mandarin tone.

4. DISCUSSIONS

This study examined whether international adoptees maintained traces of their heritage language even after a prolonged period of limited or no exposure to it. Behaviorally, we found that the international adoptees partially maintained phonological sensitivity to pitch contours; although the international adoptees were significantly poorer at perceiving pitch compared to Mandarin controls, adoptees were significantly more accurate at pitch perception than native English controls. These results suggest that adoptees do not retain fully native-like phonological representations. However, the adoptees' superior performance to the native English group suggests that they may still have some representation of early acquired linguistic features when re-exposed to Mandarin [7,10].

We also found that higher-level linguistic knowledge (lexical/semantic) was not retained by the international adoptees as demonstrated by their poorer performances in the Chinese Stroop tasks and Chinese sentence identification compared to Mandarin controls. Given the fact that the heritage language's words are lost within months of adoption and into adulthood [7,10], it is expected that adoptees would have weaker semantic connections between the word and the associated semantic meaning.

Analysis of the FFR data yielded modest results potentially due to noise and sample size; nevertheless, the adoptee group showed generally better FFR pitch tracking than their native English-speaking peers.

While we cannot rule out influences of post-adoption Mandarin exposure on the enhanced phonological and pitch sensitivities in the adoptees vs. native English speakers, we doubt that adult language learning alone explains their better tone sensitivity, as these low-level behavioral measures were not reliably correlated to their Mandarin competence (i.e., HSK).

In sum, these results suggest that the adoptees may have latent phonological representations of their heritage language, which may support their ability to learn Mandarin [13]. These data are consistent with neural commitment theory, in which early language environments have long-lasting effects on auditory and linguistic representations.

5. REFERENCES

[1] Chinese Proficiency Test 汉语水平考试 Hanyu Shuiping Kaoshi (HSK): www.chinesetest.cn

[2] Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research* 25, 161-168.

[3] Kuhl, P.K., Tsao, F.M., Liu, H.M., Zhang, Y., & De Boer, B. 2001. Language/culture/mind/brain: Progress at the margins between disciplines. *Annals of the New York Academy of Sciences* 935, 136-174.

[4] Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606-608.

[5] Moulines, E. & Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453–467.

[6] Norman, G., Bylund, E., & Thierry, G. 2022. Irreversible specialization for speech perception in early international adoptees. *Cerebral Cortex* 32, 3777-3785.

[7] Pallier, C., Dehaene, S., Poline, J.B., LeBihan, D., Argenti, A.M., Dupoux, E., & Mehler, J. 2003. Brain imaging of language plasticity in adopted adults: Can a second language replace the first?. *Cerebral Cortex* 13, 155-161.

[8] Perrachione, T.K., Lee, J., Ha, L.Y.Y., & Wong, P.C.M. 2011. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J. Acoust. Soc. Am.* 130, 461-472.

[9] Pierce, L.J., Klein, D., Chen, J.K., Delcenserie, A., & Genesee, F. 2014. Mapping the unconscious maintenance of a lost first language. *Proc. Nat. Acad. Sci. USA.* 111, 17314-17319.

[10] Ventureyra, V.A., Pallier, C., & Yoo, H.Y. 2004. The loss of first language phonetic perception in adopted Koreans. *Journal of Neurolinguistics*, 17, 79-91.

[11] Werker, J.F. & Tees, R.C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7, 49-63.

[12] Wong, P.C.M. & Perrachione, T.K. 2007. Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics* 28, 565-585.

[13] Zhou, W., Broersma, M., & Cutler, A., 2021. Asymmetric memory for birth language perception versus production in young international adoptees. *Cognition*, 213, 104788.

USING SPECTRAL COHERENCE BETWEEN TEMPORAL ENVELOPE AND MOUTH MOVEMENTS TO REVEAL RHYTHMIC REGULARITIES IN SPEECH

Lei He
[a] Department of Computational Linguistics, University of Zurich, Switzerland
[b] Department of Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich, Switzerland
lei.he@uzh.ch

ABSTRACT

This paper features a method to characterize the rhythmicity in speech from the temporal envelope and mouth opening-closing movements using spectral coherence [He (2022) *J. Acoust. Soc. Am*. **152**, 567–579]. Such a coherence reveals the joint regularities in both domains of signals. For illustrations, recordings from both L1 and L2 English utterances are used for comparisons.

The recordings utilized for this Prosody Visualization Challenge (PVC-2023) have been chosen from the EMA-MAE corpus [1],[1] which is openly accessible. As such, there are no ethical concerns associated with their current scientific and not-for-profit usage.

Keywords: speech rhythm, spectral coherence, temporal envelope, mouth opening-closing movements

1. INTRODUCTION

Rhythm is a ubiquitous phenomenon in nature. The rhythmicity in speech has been a central interest to phoneticians, speech scientists, psychologists, and neuroscientists. Different approaches to speech rhythm have been proposed over the decades (see [2] for a general overview). Here, I follow the "frame/content" perspective of speech rhythm put forward by MacNeilage [3].

The crux of this theory is under the backdrop of speech evolution: Speech production is evolved from pre-existing cyclical jaw movements in ancestral primates in the form of lip-smacking. Such cyclical lip movements were important visual-facial gestures in non-human primate communications [4]. In the course of human evolution, the coupling between mouth opening-closing cycles gradually emerged. As a result, the sonority of speech typically waxes and wanes following the mouth movement cycles. Such opening-closing alternations are organized into syllable-sized unites corresponding to the amplitude modulations, which constitute the rhythmic frames [3]. It is thus evident that the mouth opening-closing movements are pivotal to speech rhythm. In the previous Prosody Visualization Challenge (ICPhS 2019 in Melbourne), Erickson et al. [4] has presented how jaw movement patterns are indicative of prosodic differences across languages.

Here, I combine both mouth opening-closing movements and the temporal envelope to reveal the rhythmic frame using spectral coherence. Spectral coherence is a Fourier transform-based signal processing technique that captures the connectivity between both domains of signals. By nature, the coherence is nothing but a power spectrum, whose magnitude shows the correlations of both signals in the frequency domain.

2. SIGNAL PROCESSING STEPS

2.1. Extracting the temporal envelope

The audio signal (typically in the format of WAV) is bandpass filtered between 700 and 1,300 Hz to keep the vocalic energy. Then the filtered signal is full-wave rectified and further lowpass filtered at 20 Hz (This can also be achieved by re-sampling the rectified signal to 40 Hz. Thus, high frequencies not responsible for rhythmicity are naturally excluded by the Nyquist theorem).

Here, I use **ENV** to notate the temporal envelope time series.

2.2. Extracting the mouth opening-closing time series

Depending on the equipment used, mouth opening-closing functions can be extracted, for example, using a particular piece of hardware [5], or counting the number of pixels within the lip contour from a video clip [6].

Here, the mouth opening-closing time series is extracted from the electromagnetic articulography data that include the *x-y-z* coordinates from the sensors attached to the upper and lower lips in the midsagittal plane and the lip corner. The area of the triangle formulated by these three sensors is used to approximate the mouth opening magnitude at each sampled instant.[2] The whole area time series is re-sampled to 40 Hz, ad modum § 2.1 above.

I use **Om** to notate the mouth opening-closing time series.

2.3. Calculating the spectral coherence from ENV and Om

(1) The DC-biases in both **ENV** and **Om** are removed, and the Tukey window (α = .1) is used to taper the edges [7].

(2) Both signals are zero-padded to the nearest $2^{\wedge N}$ ($N \in \mathbb{Z}^{+}$), and fast Fourier transformed to yield their spectra, $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$.

(3) The cross-spectrum of both $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$ is calculated (i.e., taking the product of the Fourier coefficients of $\mathbf{FFT_{ENV}}$ and the complex conjugate of the Fourier coefficients of $\mathbf{FFT_{Om}}$). The cross-spectrum is then normalized to the product of the amplitudes of both $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$ to yield the spectral coherence of both **ENV** and **Om**.

The coherence illustrates the magnitudes of common periodicities in both **ENV** and **Om**. It is intrinsically a spectrum, and can thus be converted to the time domain via the inverse Fourier transform if desired.

3. VISUALIZATION

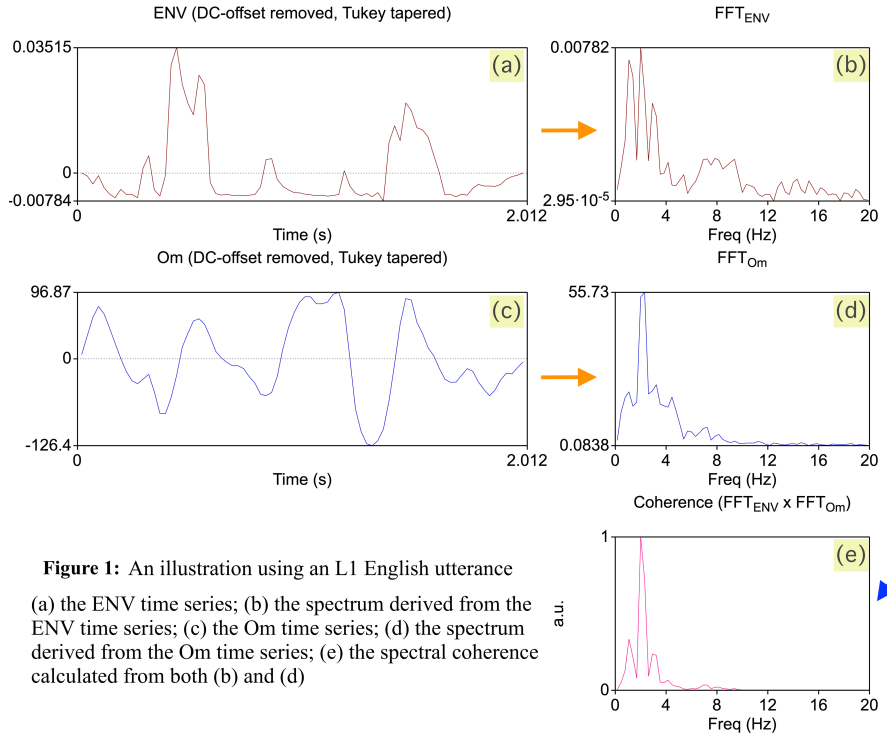


Figure 1: An illustration using an L1 English utterance

(a) the ENV time series; (b) the spectrum derived from the ENV time series; (c) the Om time series; (d) the spectrum derived from the Om time series; (e) the spectral coherence calculated from both (b) and (d)

The materials for the visualization are accessible online in a repository: https://osf.io/vq2jn/ The materials include a readme, the original waveforms, extracted **ENV** and **Om** time series, plots, and the main Praat scripts to do major signal processing steps. This repository may be updated with more materials in the future. All updates will be documented in the readme.

Figure 1 illustrates the steps used to calculate the coherence from an L1 English utterance; figure 2 illustrates the steps used to calculate the coherence from an L2 English utterance, and is only accessible in the repository due to the page limit. An initial inspection of the two coherences in both plots suggests that the joint regularities occur at a frequency of ~2 Hz for the L1 utterance (Fig.1); whereas for the L2 utterance, multiple regularities coexist (Fig. 2).

4. ACKNOWLEDGEMENTS

Swiss National Science Foundation (Grant № PZ00P1_193328) / University of Zurich Forschungskredit (Grant № FK–20–078).

5. REFERENCES

[1] A. Ji, J. J. Berry, and M. T. Johnson, 2014. The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In: Proc. ICASSP, Florence (May 4–9, 2014), pp. 7769–7773.

[2] L. He, 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am.* 152, 567–579.

[3] P. F. MacNeilage, 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499–511.

[4] D. Erickson, T. Huang, C. Menezes, and S. Kawahara, 2019. Using jaw movement patterns to visualize prosody. Poster presented for the Prosody Visualization Challenge (PVC-2019) at the 19th ICPhS, Melbourne (August 7–8, 2019).

[5] D. Erickson, O. Niebuhr, W. Gu, T. Huang, and P. Geng, 2020. The MARRYS cap: A new method for analyzing and teaching the importance of jaw movements in speech production. In: Proc. 12th International Seminar on Speech Production (Haskins Press, New Haven), pp. 48–51.

[6] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, e1000436.

[7] L. He, 2022. tukey_window.praat (A Praat script to apply the Tukey window to a signal). URL: https://osf.io/6r839

[1] The EMA-MAE dataset is based on work supported by the National Science Foundation of the United States under Grant #IIS-1142826 to Marquette University, which support does not constitute an endorsement.

[2] The formula below is used to calculate the area of this triangle. In the formula, *x, y,* and *z* refer to the coordinates from sensors attached to the upper, lower lips and the lip corner (vis-à-vis the subscripts UL, LL and LC):

$$\frac{1}{2}\sqrt{\begin{vmatrix} y_{UL} & z_{UL} & 1 \\ y_{LL} & z_{LL} & 1 \\ y_{LC} & z_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} z_{UL} & x_{UL} & 1 \\ z_{LL} & x_{LL} & 1 \\ z_{LC} & x_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} x_{UL} & y_{UL} & 1 \\ x_{LL} & y_{LL} & 1 \\ x_{LC} & y_{LC} & 1 \end{vmatrix}^2}.$$

Disentangling intonation in tone languages

Katrina Kechun Li, Francis Nolan, Brechtje Post

University of Cambridge
kl502@cam.ac.uk, fjn1@cam.ac.uk, bmbp2@cam.ac.uk

ABSTRACT

Describing intonation in tone languages can be difficult, as tone and intonation both heavily rely on f0. In this paper, we propose a method to disentangle intonation from tonal contours by plotting the f0 contour differences of the same sentence uttered under different focus conditions. We demonstrate this method through a case study of two Chinese languages, Chengdu and Changsha. The results reveal that while intonation realisation is not entirely independent of lexical tones, it is possible to identify a uniform picture of intonational realisation across various tonal conditions. This method lessens the reliance on prior knowledge of lexical tones in interpreting intonation, and it contributes to the development of an abstract representational system for intonation in tone languages.

Keywords: difference curve, tone languages, f0 contour, prosodic focus, Chinese languages

1. INTRODUCTION

Intonation and lexical tones are intertwined in tone languages, making it challenging to disentangle them. Consequently, studies of intonation in tone languages often require f0 contours to be interpreted separately for each tonal condition, which is less straightforward than in non-tonal languages. This limitation has hindered the application of an abstract intonational model in tone languages, even if ToBI-style models have been proposed [1]. This study examines a single intonational function, focus, in multiple tonal conditions, but explores a method that subtracts lexical tones from the intonation contour. Despite its simplicity, the visualisation yields a uniform picture of intonation patterning, which could support an abstract intonation model in tone languages.

2. METHODOLOGY

We present data from two Chinese languages with distinct tonal systems, Chengdu and Changsha. In each dataset, participants uttered a sentence with a comparable SVO structure as a response to a precursor wh-question eliciting focus on either the subject (Name) or the object (Object). Within the sentence, the Name, Verb, and Object all had the same tone. Four sentences were constructed with four different tones (T1-T4) used at the three target positions. The example below shows one of the four tonal conditions in each language (citation tones represented in Chao number).

Name Prefix	Name	Verb	Aspect	Classifier	Object	Object Suffix
			Chengdu			
ɕiaʊ53	**jɪn53**	**mai53**	lə0	pa53	**ji53**	tsɹ0
Xiao-yin		bought		a	chair	
			Changsha			
ɕiaʊ42	**jɪŋ45**	**mɑi45**	tɑ21	tsɑ34	**iou45**	tsɹ33
Xiao-yin		sold		a	grapefruit	

The audio files were processed using *Praat* [2] and *Parselmouth* in Python [3], which generated a comparable data format that can be plotted using *ggplot2* in R [4]. For each syllable, f0 was extracted from 20 equidistant points of the rhyme and converted to semitones based on the participant's average f0.

For each dataset, we report the visualisations of four tonal conditions based on three repetitions of one male participant. The pattern revealed will become more obvious with a full dataset, but the current paper aims to demonstrate that even a small dataset can convey useful information through this visualisation. Information about the full dataset can be found in the congress paper 'Variations of focus prominence in three tone languages', and visualisations and sound samples are available on the website [5].

3. RESULTS AND DISCUSSIONS

The left panels of Figure 1 display the raw f0 contours for each syllable in Chengdu and Changsha respectively. Each row corresponds to a tonal condition, and each column corresponds to a syllable. It is notable that the f0 deviations happen mostly on the 'Name' and the 'Object' where the focus is placed (bold lines). However, the specific changes in f0 are heavily influenced by the underlying lexical tones, making it difficult to discern a clear pattern.

The difference curve introduced in the right panels of Figure 1, which equals the subject focus contour minus the object focus contour, presents a more consistent pattern across different tonal conditions. Instead of comparing two curves, one can compare a single curve with the reference line y=0 (red lines). Positive values indicate that the f0 of subject focus is higher than that of object focus, and vice versa.

The difference curves across tonal conditions seem to form groups. At the 'Name' in Chengdu, for example, T1 and T4 show a dipped difference curve below the reference line, whereas T2 and T3 have a curve above the reference line. This difference reflects the influence of underlying tonal target: the low target in T1 and T4 in the middle of the syllable gets lower under focus, and the high target in T2 and T3 gets higher, at the beginning and end of the syllable respectively. In Changsha, the difference curves cross the reference line in T1 and T2 at the 'Name', suggesting a bitonal target underlyingly, where the high target is scaled higher while the low target is scaled lower. This observation is in line with the proposal that focus is realised through hyper- or hypo-articulation of tonal targets [6].

In addition to the paradigmatic view of comparing to a reference line, the difference curve also promotes a syntagmatic view of comparison within the sentence. For example, the difference curve at the 'Object' can be close or above 0 (Chengdu T2/T4, Changsha T2/T3), which is not the expected focus effect. However, the relative difference between 'Name' and 'Object' still reflects the effect of focus. Although this relational view of intonation is not uncommon in non-tonal languages, it is often missing in the previous discussions in tone languages, due to the complexities in traditional visualisations.

In conclusion, the difference curve proposed in this study offers a simple yet effective way to disentangle intonation in tone languages. By visualising the f0 difference between subject and object focus, the method reveals a uniform picture on the effect of focus while taking into account the underlying lexical tones. The approach also renders easier both the paradigmatic and syntagmatic interpretation. The study shows that the proposed method can be used to generalise intonation patterns in different tonal languages. Future studies can explore the robustness of this method by examining its application on different sentence structures and tonal combinations.

5. REFERENCES

[1] Peng S., Chan M. K. M., Tseng C., Huang T., Lee O. J. and Beckman M. E. 2005. Towards a Pan-Mandarin System for Prosodic Transcription. In *Prosodic Typology*, edited by Jun S-A. Oxford University Press.

[2] Boersma P. and Weenink D. 2019. Praat: doing phonetics by computer. Accessed: May 26, 2019.

[3] Jadoul Y., Thompson B. and de Boer B. 2018. Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics* 71, 1–15.

[4] Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[5] Li K. K. 2023. InTone [Website]. Available: https://kc-li.github.io/InTone_Visualisation_Site

[6] Chen, Y. and Gussenhoven C. 2010. Post-Focus F0 Compression—Now You See It, Now You Don't. *Journal of Phonetics* 38 (4), 517–25.

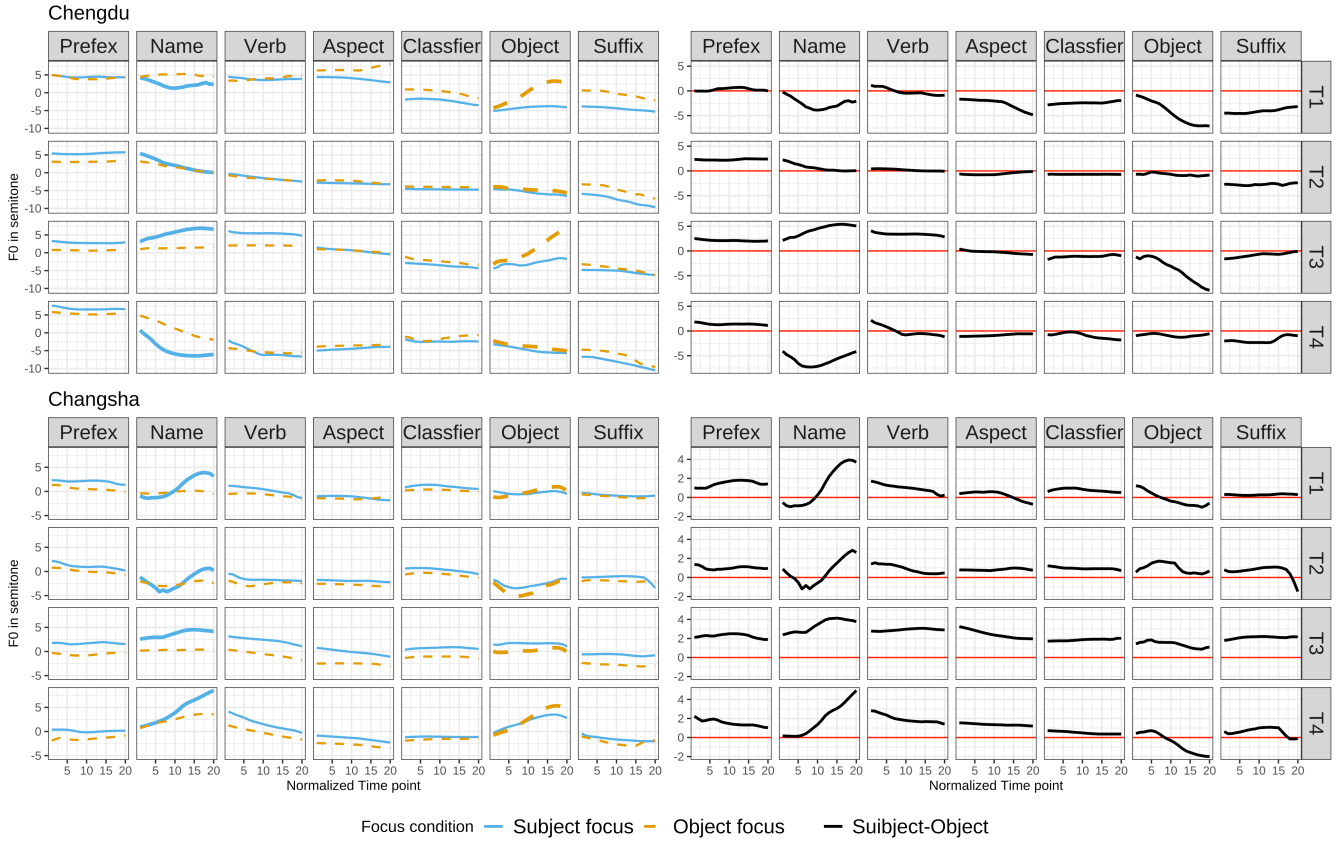


Figure 1: Raw f0 contour (left panel) vs. Difference curve (right panel), in Chengdu (top) and Changsha (bottom).

RELATION BETWEEN MUSICAL APTITUDE AND L2 STRESS PERCEPTION IN FRENCH- AND KOREAN-SPEAKING LISTENERS

Maria Teresa Martinez Garcia[1], Sandra Schwab[2]

[1]University of Utah Asia Campus, [2]University of Bern
[1] maria.martinezgarcia@utah.edu, [2] sandra.schwab@unibe.ch

ABSTRACT

Several factors could explain individual differences in the perception of non-native stress contrasts. This study investigates the effect of the native language (Spanish, French, Korean) and the impact of music aptitude on French and Korean listeners' performance at stress-detection in Spanish. The results showed an effect of L1, with Korean and French listeners showing poorer discrimination than native speakers. Moreover, only French (but not Korean) listeners showed a positive correlation between music aptitude and stress discrimination. These findings suggest the existence of a link between music and L2 stress discrimination in French, but not in Korean listeners. Individual differences and the high music aptitude of Korean listeners are discussed as potential explanations for the different results reported for the two groups of non-native listeners.

Keywords: music aptitude; L2 prosody; stress 'deafness'; Korean, French

1. INTRODUCTION

1.1. Stress 'deafness' hypothesis

Depending on the properties of their native language, some speakers can easily detect second language (L2) stress, while others cannot, known as the stress deafness hypothesis (e.g., [1][2][3]). According to this hypothesis, the degree of stress deafness is related to the stress properties of the L1, specifically to the nature of their lexical stress (free or fixed). The hypothesis claims that listeners of fixed-stress languages have more difficulties in perceiving stress contrasts in the L2 than listeners of free-stress languages, as they cannot encode stress information in their mental lexicon (e.g., [1][2][3]). Among other factors, the ability to discriminate L2 stress contrasts also depend on the cognitively complexity of the perceptual task that listeners have to perform. For example, listeners tend to perform better in AX than in ABX discrimination tasks ([2]). The present study investigates L2 stress perception with two cognitively different tasks (i.e., identification and Odd-One-Out tasks) in listeners with fixed-stress (Korean and French) while listening to a foreign/L2 language with free stress (i.e., Spanish), and explores the role of musical aptitude in this perception.

1.2. Spanish, Korean and French stress properties

The languages included in this study –Spanish, Korean, and French– differ in their accentual properties. Spanish is a free-stress language where lexical stress has a distinctive function ([4]), distinguishing segmentally identical words such as in *número* (/'numero/, engl. (the) *number*) and *numero* (/nu'mero/, engl. I *number*). French is a fixed-stress language in which primary stress is realized at the phrasal level, precisely on the last syllable of the Accentual Phrase (AP) ([5]). Consequently, stress has a demarcative function (i.e., it marks the right end of the accentual phrase; [5]). Moreover, both the realization of primary stress and the intonation contours are anchored on the final syllable of the accentual phrase, which leads to a "fusion" between accentual and intonational structures (i.e., the so-called "syncretism" between accentuation and intonation; [6]). Similarly, Korean is a fixed-stress language (i.e., [7]), which uses phrase-level prominence, also at the end of Accentual Phrase, cued primarily by f0 in marking phrasal boundaries. The Accentual Phrase in standard (Seoul) Korean is demarcated by a phrase final High tone, while the initial boundary of the prosodic domain could be either a High or a Low tone, depending on the initial segment of the utterance (e.g., [8][9]). While certain dialects of Korean (e.g., Chonnam or North Kyungsang) may have a vowel length distinction in the first syllable of certain words (e.g., *sa:.kwa 'apology'* vs. *sa.kwa 'apple'*) ([10][11][12]), Seoul Korean (the dialect studied) has lost this feature.

1.3. L2 stress perception and music aptitude

In most studies dealing with musical expertise and pitch detection (e.g., [13][14][15][16]), pitch variations did not convey a linguistic/semantic meaning (i.e., no distinctive function). In this study, we examine whether musical skills enhance the perception of f0 variations that convey linguistic meanings. In languages such as Spanish, f0 variations are involved in the realization of lexical stress. Then, a change in f0, coupled with changes in duration and

amplitude, implies a change in the position of the stressed syllable (e.g., [4]) and, hence, a change in meaning.

To our knowledge, only a few studies have examined the influence of musical expertise on the perception of lexical stress in L2 French listeners (e.g., [17][18][19][20]). Using different experimental tasks (sequence repetition, discrimination of stress contrasts, or identification of stressed syllables), they all showed that musical expertise was significantly related to French listeners' performance at perceiving stress contrasts in an L2 with unpredictable stress (i.e., Dutch, English, or Spanish). On the other hand, while musical notation has been shown to help Korean learners understand stress placement in English ([21]), no other study has investigated the possible link between musical aptitude and L2 stress recognition among Korean speakers.

Our main aim is to examine to what extent French and Korean listeners differ in their L2 stress detection, in comparison with native listeners. We hypothesize that, given the similarities the two languages, both groups would show similar patterns, performing lower than native listeners. Like [22], we used a stress identification task (not cognitively very demanding). Participants heard Spanish words and had to indicate whether they perceived stress on the first, second or third syllable. The second task was an Odd-One-Out task which was more memory demanding. Listeners heard trials of three Spanish words and had to indicate the word with a different stress pattern. Secondly, we aim to determine how music aptitude and L2 stress detection are related in French and Korean listeners. Because of the similar nature of stress in French and Korean, we predicted a similar relationship between music aptitude and L2 stress detection for both groups of listeners.

2. METHOD

2.1. Participants

Three groups completed the study. The first group included 20 Swiss French listeners (hereafter 'FR') recruited at the University of Neuchâtel, Switzerland (*mean age*: 22.1 years, *sd*: 2.15, *age range*: 21-26 years). They had no knowledge of Spanish or other Romance language. The second group included 20 Korean listeners (hereafter 'KR') recruited at the University of Utah *Asia Campus*, Korea (*mean age*: 21.9 years, *sd*: 2.7, *age range*: 19-29 years). At the time of the experiment, they had been enrolled in a basic course of Spanish for 3 weeks (4 hours per week). The third group included 41 native Spanish/Catalan listeners (hereafter 'ES') recruited at the University of Pompeu Fabra, Spain (*mean age*: 21.56 years, *sd*: 4.46, *age range*: 19-44 years).

All the FR and KR performed the identification task, the Odd-One-Out task, and the music aptitude tests, whereas 24 ES performed the identification task and the other 17 ES performed the Odd-One-Out task. ES did not perform the music aptitude test.

2.2. Identification

Participants completed a stress identification task where they heard 120 trisyllabic Spanish words (e.g., *numero*) and had to indicate which syllable was stressed (i.e., see [22]). The experiment was run using Praat software scripts for all participants [23]. Different voices and both declarative and interrogative intonations were included.

2.3. Odd-One-Out

Participants performed an Odd-One-Out task in which they heard 216 trials of three segmentally identical Spanish words (e.g., *numero*). Among them, two words presented the same stress pattern (e.g., stress on the penultimate syllable) and one (i.e., the odd) presented a different stress pattern (e.g., stress on the final syllable). Participants' task was to indicate which of the three words was the deviant word (i.e., 'odd-one-out'; see [22] for details). The experiment was run using Praat software scripts for FR and KR participants [23] and using DMDX for ES participants [24]. Different voices and type of intonation (declarative and interrogative) were included in the stimuli.

2.4. Music aptitude test

We adapted the Advanced Measures of Music Audiation test (i.e., AMMA) developed by [25]. Participants performed a discrimination task in which they heard 16 sequences of two melodies. They had to indicate whether the two melodies were the same or different. As in the original test, the differences between the two melodies were either tonal or rhythmic. The experiment was run using Praat software scripts for all participants [23].

2.5. Data analysis

For the music aptitude test, the percent correct for each participant was calculated. For the identification and Odd-One-Our tasks, the accuracy (in percentage) was calculated for each participant.

Statistical analyses were conducted using R (lme4 R package; [26]). Two analyses were conducted. First, two mixed-effects logistic regression models were run on the correct/incorrect responses ([27]) to

test the effect of L1 in the identification and Odd-One-Out tasks. In each model, the fixed part was 'L1' (ES, FR, KR) which was recoded into [0, 1] dummy variables. The variable 'trial number', centered on the mean (for R operational reasons the variable 'trial number' was first rescaled between 0 and 1), was included as a control variable. The random part of the model included random intercepts for participants and items. In each model, the significance of the main effect was assessed with likelihood ratio tests that compared the model with the main effect to a model without it. The estimates (β) are expressed in logit and were computed taking 'incorrect response' as the reference level for the dependent variable. Post-hoc analyses with Tukey correction for multiple comparisons were performed to obtain 2 by 2 comparisons. The figures in the results section show percentage of correct responses, although the models were run on raw data (correct/incorrect responses).

Second, correlation analyses were performed, separately for FR and KR, between the participants' music aptitude and their average accuracy (%) at the identification and Odd-One-Out tasks respectively. We did not run regression analyses with L1, task, music aptitude, and their interaction as predictors because of the limited number of data points and hence, low statistical power.

3. RESULTS

3.1. Effect of L1

Figure 1 shows the results for the ES, FR, and KR groups in the identification (left) and Odd-One-Out task (right). For the identification task, the effect of L1 was significant ($\chi2(2) = 109.02$, $p < .001$). As expected, post-hoc analyses revealed that ES presented a better performance than FR and KR, ($\beta = -2.83$, SE = 0.20, $z = -14.23$, $p < .001$; $\beta = -2.86$, SE = 0.20, $z = -14.03$, $p < .001$). As predicted, we observe no significant difference between FR and KR ($\beta = 0.03$, SE = 0.20, $z = -0.17$, $p = .98$).

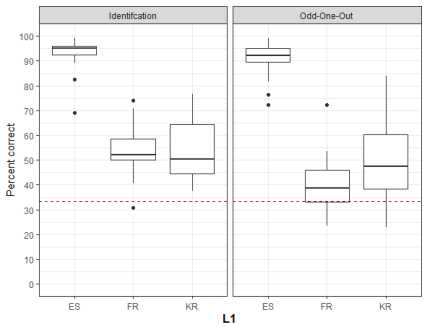


Figure 1: Percent correct by L1 (ES, FR, KR) in the identification (left) and the Odd-One-task (right).

Similar results were found for the Odd-One-Out task. The effect of L1 was significant ($\chi2(2) = 77.41$, $p < .001$). ES performed significantly better than FR and KR ($\beta = -3.30$, SE = 0.208 $z = -11.98$, $p < .001$; $\beta = -2.80$, SE = 0.28, $z = -9.94$, $p < .001$) whose performance did not significantly differ from each other ($\beta = 0.50$, SE = 0.26, $z = 1.92$, $p = .13$).

3.2. Relationship between music aptitude and stress detection accuracy

Figure 2 presents the relationship between music aptitude and accuracy for FR (top panels) and KR (bottom panels) in the identification task (left) and Odd-Out-Out tasks (right). For FR participants, results showed a marginally significant correlation for the Odd-One-Out task ($r = .44$, $p = .051$), but no signification correlation for the identification task ($r = .36$, $p = .12$). The better the French-speaking listeners' music aptitude, the better their Odd-One-Out performance. Although there seems to be an outlier in the Odd-One-Out task for the FR group (i.e., a participant who scored extremely low in the music test), the marginal correlation does not disappear when excluding this participant.

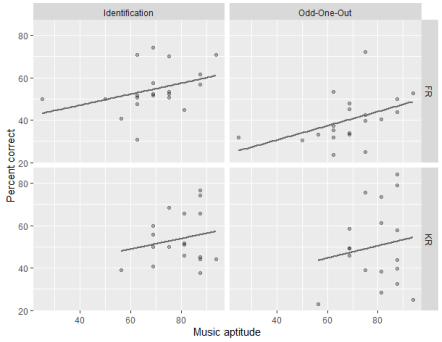


Figure 2: Percent correct by music aptitude for FR (top) and KR (bottom) in the identification (left) and Odd-Out-Out task (right).

For KR participants, the relationship between music aptitude and accuracy was not significant in neither the identification task ($r = .20$, $p = .42$) nor the Odd-One-Out task ($r = .15$, $p = .55$).

4. DISCUSSION

4.1. L2 stress perception in non-native listeners

The first goal was to examine the non-native stress detection performance of two groups of native listeners of a predictable stress language (i.e., French, Korean), who supposedly have difficulties in detecting stress contrasts in L2 Spanish. Their performance was compared with the performance of native listeners of Spanish. Results showed that, as expected, both French and Korean listeners showed

lower performance than native listeners in both tasks. According to our predictions, Korean listeners performed as poorly as French-speaking participants. The former, however, presented a higher degree of variability in their results. Two aspects should be considered to understand this variability difference. First, their knowledge of Spanish was different. French participants had no knowledge of Spanish (i.e., all were *ab initio* learners). However, Korean participants had been enrolled in a Spanish course for three weeks, which could generate interindividual perceptual differences. Second, the groups also differed in their Proficiency in English. While Korean's English (subjective) proficiency ranged from beginner to native-like, most French participants (i.e., 15/20) estimated their English proficiency as advanced. Spearman correlation analyses showed a significant relationship between the listeners' proficiency and their performance in the tasks for Korean participants ($rs(16) = .47$, $p = .05$; $rs(16) = .48$, $p = .04$), but not for French-speaking listeners ($rs(18) = .29$, $p = .22$; $rs(18) = .04$, $p = .86$). In sum, although Korean participants performed as poorly as French-speaking listeners, their stress detection performance in Spanish could have been related to their proficiency in English. Therefore, English (or other L2) proficiency constitutes a factor that needs to be either controlled or entered as a possible predictor of L2 stress detection performance, while studying the acquisition of L2 stress contrasts.

4.2. L2 stress perception and music aptitude

The second goal was to determine whether there was a relationship between the non-native listeners' music aptitude and their performance at detecting stress in a foreign language. Regarding the French group, we found that listeners with high musical aptitude tended to show better performance at discriminating stress contrasts in Spanish, but not at identifying the stressed syllable (i.e., the relationship was marginally significant in the Odd-One-Out, but not significant in the identification task). This finding, in line with previous research ([20]), suggests that similar perceptual mechanisms might be called in the discrimination of musical melodies (i.e., music test) and in the discrimination of non-native stress contrasts (i.e., in the Odd-One-Out task).

As for Korean listeners, no significant relationship between musical aptitude and L2 stress detection performance was found for either task, in disagreement with our expectations. One explanation is the fact that most of the Korean participants performed well in the music test (i.e., only one participant had a performance lower than 68%). Consequently, music aptitude, being high and relatively constant among Korean listeners, is not related to L2 stress detection. Korean students have mandatory music training during their education until the age of 18, which makes them be well trained in music. Their theoretical and practical music training could have had an input on how they perceived melodies in the music test performed in this study. Moreover, there is anecdotal evidence that the majority of Koreans have Karaoke as one of their hobbies and spend a couple of hours a week in Karaoke places (94% of 20-year-old Koreans spend at least 1.5 hours a week in a Karaoke, [28]).

Thus, possible explanations for the divergence between French and Korean results regarding musical aptitude could be either related to the lack of variability in the Korean data or task-dependent. On the one hand, all the Korean listeners performed well in the musical aptitude test, which could hinder any potential correlation between musical aptitude and stress perception. To confirm this hypothesis, it would be necessary to test Korean listeners without musical training or to find a more fine-tuned task, such as the PROMS tests proposed [29] to evaluate musical aptitude among individuals who are already trained in music.

On the other hand, a (marginal) correlation was found for French listeners in the Odd-Out-One task, not in the identification task. To rule out task-dependent results, other tests should be used to evaluate musical aptitude (e.g., PROMS tests, [29]). If the link between music aptitude and non-native stress perception is not task-dependent, a correlation should be found with different types of tasks.

.

5. CONCLUSION

Our findings support the hypothesis that the stress properties of the native language influence the perception of stress contrasts in a foreign language, shown by how the accentual similarities between French and Korean made them pattern similarly to each other. Two factors were found to explain the interindividual variability observed in non-native stress detection ability: musical aptitude and proficiency in a foreign language (i.e., English in this case). While these two factors correlated with the non-native stress perception performance, further studies should deeper explore their contribution in the perception of non-native contrasts by either examining other language combinations or using a wider range of experimental tasks.

6. REFERENCES

[1] Dupoux, E. Sebastian-Gallés, N. Navarette, E., Peperkamp, S. 2008. Persistent stress 'deafness': The

case of French learners of Spanish. *Cognition, 106*, 682-706.
[2] Dupoux, E., Pallier, C., Sebastian-Gallés, N., Mehler, J. 1997. A destressing 'deafness0 in French ? *Journal of Memory and Language, 36*, 406-421.
[3] Dupoux, E., Peperkamp, S., Sebastian-Gallés, N. 2001. A robust method to study stress 'deafness '. *Journal of the acoustical Society of America, 110,* 1606-1618.
[4] Quilis, A. 1981. *Fonética acústica de la lengua española,* Madrid, Gredos.
[5] Lacheret-Dujour, A., & Beaugendre, F. (1999). *La prosodie du français* (French prosody) (CNRS éditions, Paris), pp. 1-354.
[6] Rossi, M. (1979). Le français, langue sans accent (French, language without stress), *Studia Phonetica, 15*, 13-52.
[7] Lee, G. (2015). Production and Perception of Korean and English Prominence by Korean Speakers. Unpublished doctoral dissertation. Kansas University.
[8] Jun, S.-A., 1993. The Phonetics and Phonology of Korean Prosody. Ph.D. dissertation. Ohio State University.
[9] Jun, S.-A., 1998. The accentual phrase in the Korean prosodic hierarchy. *Phonology, 15*(2), 189–226.
[10] Heo, W., 1965. Kwuke Umwoonhak [Korean Phonology]. Cengumsa, Seoul.
[11] Kenstowicz, M., Park, C., 2006. Laryngeal features and tone in Kyungsang Korean: a phonetic study. *Stud. Phon. Phonol. Morphol. 12* (2), 247–264.
[12] Ko, E.-S., 2013. A metrical theory of Korean word prosody. *Ling. Rev. 30*(1), 79–115.
[13] Magne, C., Schön, D., Besson, M. 2006. Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neuroscience, 18*, 199–211.
[14] Marques, C., Moreno, S., Castro, S. L. Besson, M. 2007. Musicians detect pitch violation in a foreign language better than non-musicians: behavioural and electrophysiological evidence. *Journal of Cognitive Neuroscience, 19*, 1453-1463.
[15] Micheyl, C., Delhommeau, K., Perrot, X. et al. 2006. Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*, 36-47.
[16] Schön, D., Magne, C., Besson, M. 2004. The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*, 341–349.
[17] Degrave, P. 2017.Can music help learners and teachers in word stress perception? *Travaux du Cercle Belge de Linguistique, 11,* 1-20.
[18] Kolinsky, R., Cuvelier, H., Goetry, V., Peretz, I., Morais, J. 2009. Music training facilitates lexical stress processing. *Music Perception, 26*(3), 235-246.
[19] Schwab, S., Calpini. N. 2018. Expertise musicale et perception de variations de f0 en L1 et en L2. *Revue Française de Linguistique appliquée, 23*, 15-30
[20] Schwab, S. & Dellwo, V. (2019). *Music and L2 prosody: the role of musical aptitude on the discrimination of stress contrasts.* In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019 (pp. 1927-1931). Canberra, Australia: Australasian Speech Science and Technology Association Inc.
[21] Kim, K. 2019. The efficacy of lexical stress diacritics in the English comprehensibility and accentedness of Korean speakers. Ph.D. dissertation. University of Victoria.
[22] Schwab, S. & Dellwo, V. (2021) Explicit versus non-explicit prosodic training in the learning of Spanish L2 stress contrasts by French listeners. *Journal of Second Language Studies*.
[23] Boersma, P., Weenink, D. 2011. Praat: Doing phonetics by computer. [Computer Software]. Consulté sur: http://www.praat.org.
[24] Forster, J. C., DMDX updates page. Online: http://www.u.arizona.edu/~jforster/dmdx.htm
[25] Gordon, E. 1989. *Advanced measures of music audition.* Chicago, GIA Publications.
[26] Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.
[27] Baayen, R. H., Davidson, D. J., & Bates, D. M. 2008. Mixed effects modeling with crossed random effects for subjects and items, *J. Memory Lang., 59*, 390-412.
[28] KB Bank (2019). 노래방 현황 및 시장여건 분석.Analysis of karaoke status and market conditions (KB Self-Employment Analysis Report) Retrived from: https://eiec.kdi.re.kr/policy/domesticView.do?ac=0000148861
[29] Law, L., & Zentner, M. (2012). *Assessing Musical Abilities Objectively: Construction and Validation of the Profile of Music Perception Skills.* PLoS ONE 7(12): e52508. doi:10.1371/journal.pone.0052508

PERCEPTION OF WORD STRESS AMONGST FRENCH LEARNERS OF ENGLISH: NUCLEAR TONE & SUFFIX

Takeki Kamiyama[1], Maelle Amand[2]

[1]TransCrit, Université Paris 8 Vincennes - Saint-Denis
[2]CeReS, Limoges University
takeki.kamiyama@univ-paris8.fr, maelle.amand@unilim.fr

ABSTRACT

French-speaking learners of second languages are said to be partially *stress deaf* (Dupoux et al. 2008). However, little is known about error tendencies in auditory identification of lexical stress in English. 55 advanced learners identified the primary stress of 46 utterance-final words pronounced by an American English speaker with an aggregated 36% of errors rates. Words ranged from 2 to 6 syllables (mode = 4). The study measures the impact of 6 stress-imposing endings (plus the addition of *-ing* for 3 of them) with 3 nuclear tones (fall, rise, fall-rise) on word-stress identification. Results show a significant difference in the error rate between the rising tone and the other tones ($p < 0.006$). Amongst all participants, stress-imposing endings also had a significant impact on error rates ($p < 0.001$) and even more so when combined with the neutral suffix *-ing*, with stresses being more frequently identified on the syllable preceding *-ing* with an unreduced vowel, e.g. ˈ*celebrate* vs. *celeˈbrat***ing**.

Keywords: L2 word-stress perception, nuclear tone, French learners of English, stress-imposing endings

1. INTRODUCTION

French-speaking learners of second languages are said to be somewhat *stress-deaf* ([1] [2] for Spanish; [3] for French Canadian learners of English). They are perceptually less sensitive to lexical stress in second languages and thus have difficulty in learning to perceive it. This study aims to investigate how advanced learners of English perceive lexical stress in words of two syllables or more, the majority of which contain a stress-imposing ending (*-ate*,*-ify*). It also tests whether varying F0 patterns have an impact on the identification of stress by these learners.

Native listeners of English are known to identify lexical stress through various acoustic parameters: fundamental frequency (f0), duration, intensity and vowel quality [4]; [5]; [6]. The use of such a combination of parameters is challenging for French-speaking learners of English since French has no lexically distinctive stress and the parameters used to cue initial and final phrase accents in French are mainly duration (lengthening of the onset of the initial syllable and of the rhyme of the final syllable) and f0 (more or less high f0 depending on the prosodic hierarchy of the boundary, except in utterance-final position) ([7]; [8] *inter alii*). Are French-speaking learners simply *stress-deaf* and insensitive to any prosodic parameters, or are they sensitive to specific cues when identifying lexical stress in English? Do they also resort to segmental parameters such as the lack of reduction in vowels, be they lexically stressed or unstressed? The correlation between f0 movement and perceived primary stress has often been examined in both native [9] and non-native speakers [10]. Tremblay [3] highlighted the difficulty of French-speaking Canadian learners of English in hearing stress in nonce words in isolation (AXB test). [11] also examined the perception of stress in disyllabic word-pairs using synthetic speech.

However, the stimuli used in the above-mentioned studies were restricted to disyllabics, and to our knowledge, the use of f0 by French learners as a cue to perceiving lexical stress in English on suffixed words has not yet been examined.

Although many frequent words in English are stressed on the first syllable [12], many words used in Academic English do no follow this pattern. Words derived from Latin/French [13] often contain stress-imposing endings which modify the stress-pattern of the words they derive from, e.g. ˈ*stable* => *staˈbility*. Some of these word-endings, such as *-ate*, *-ify* and *-ize* can be combined with a neutral suffix, which do not affect the stress-pattern (e.g. *-ing*): ˈ*celebrating*, *iˈdentifying*, ˈ*harmonising*. Secondary stress is also found on an earlier syllable if primary stress falls on the third syllable or later (ˌ*visualizˈation*). Such combinations and constraints lead to an increase in the learner's difficulties and strategies to identify primary stress. [14] listed at least 6 factors influencing stress placement

amongst Brazilian learners of English (reading of target words within sentences). Some of them include the tenseness of the vowel in the final syllable of verbs, the presence of a tertiary stress (i.e. syllables with unreduced vowel not receiving primary or secondary stress), the stress pattern of derivatives along with the predominant stress-patterns of English.

The present analysis examines multiple effects influencing the perception of lexical stress across two groups of advanced learners, one specialising in the English language, literature and culture (henceforth *Eng*), and the other, in applied uses of English and another modern language, namely, German or Spanish (henceforth *AppLang*). We expect *Eng* to perceive correct primary stresses more often than *AppLang* since only the former received formal training on pronunciation and explicit knowledge on stress placement in relation to common stress-imposing endings. We also hypothesise that French learners are likely to identify primary stress on the initial syllable or on the last syllable based on acoustic patterns observed in French. Finally we predict that a rising nuclear tone may lead to the perception of a stress on the syllable receiving a rise.

2. METHODOLOGY

The 55 listeners were French-speaking learners of English at university level (year 2). (1) 26 *Eng* (440 hours of courses related to English in year 1 + semester 1 of year 2): in year 1, they had received 24 hours of formal teaching in the pronunciation of English (introduction to phonemes, word-endings and word stress, etc.). In year 2, the pronunciation module did not include word stress. 2 students were removed from the study because they were native speakers of English living in France. (2) 29 *AppLang*, who all receive courses in English and in either Spanish or German (264 hours of courses related to English). They spend 30% less time in class studying English and had not yet received any training or explicit knowledge on word stress. However, the other language they major in also has lexical stress, which may raise their awareness on lexical stress indirectly.

The 46 test words shown at the end of the paper mostly contain a stress-imposing ending (*-ate*, *-i/yse*, *-ify*, *-igible*, *-ion*, *-ity*), with or without a neutral suffix (*-ing*, *-ment*) and their length ranges from 2 to 6 syllables (mode=4). They were placed in sentences eliciting 3 different nuclear tones, namely, fall, rise, fall-rise in the British School of intonation analysis ([15] *inter alii*): e.g. *No, I didn't use the word ...* for fall; *Have you ever said ...* for rise; *I didn't like the word ...* for fall-rise. These utterances were presented in a randomized order for each nuclear tone and read once by a male speaker of General American from New England in his twenties. Fig. 1 shows f0 values of some of the utterances thus obtained.

The utterances were then arranged in a semi-random order so that the same stress-imposing ending would not be presented in two succeeding trials, and that the same nuclear tone succeeds maximally twice. The list of 138 (= 46 test words x 3 nuclear tones) utterances were then presented twice in two different orders to the participants on a screen, who were asked to select the primary stress of the test words presented in orthography and broken down into syllables. A clearly indicated training phase that preceded the test phase helped the participants familiarise with the task. The experiment was compiled with Praat [16] and lasted around 30 minutes. A mixed effect model was run with correct/incorrect placement as dependant variable and student group, tone, suffix and syllable number as independent ones. Participants were treated as random effects.

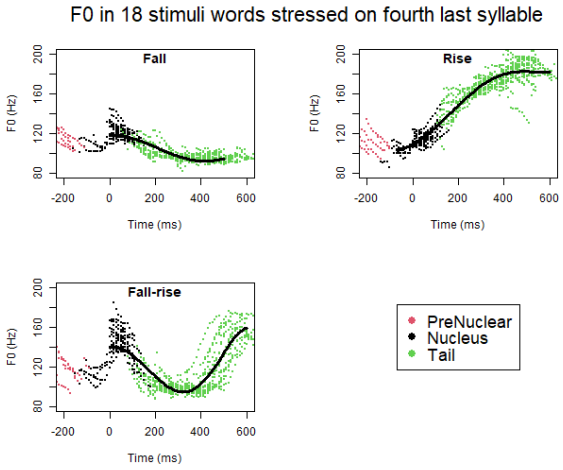


Figure 1: F0 measured (automatically then corrected manually on Praat [16] Pitch object) every 10ms for the 18 target words with stressed syllable followed by 3 unstressed ones (e.g. ˈnotifying, inˈterrogating) pronounced with fall, rise and fall-rise. 5th-degree polynomial fitting curve. Time 0: beginning of rhyme of stressed syllable.

3. RESULTS

Eng and *AppLang* exhibited significantly different scores of correct/incorrect stress identification (henceforth *cor/inc*) (χ^2 = 261.81, df = 1, p<2.2e-16), with *Eng* outperforming *AppLang* (Table 1).

Learner group	% Correct	% Incorrect
Eng	69.0%	31.0%
$n = 26$	(5025)	(2200)
AppLang	56.0%	43.1%
$n = 29$	(4609)	(3461)

Table 1: % of cor. vs. incor. primary stress placement by learner group. Counts are between parentheses (n=15295).

Overall, *Eng* tend to favour the correct syllable over any other incorrect syllable, unlike *AppLang* students fro specific words (e.g. **anaLYsing*, **celeBRAting*, **notiFYing*; **eliGIble*, **speciaLIsation*). A GLME model confirmed a group difference with *Eng* being 20.5% more likely of correct hits (p = .03, Std.E. = .28).

3.1. Nuclear tone and *cor/inc* stress identification

The model also confirmed that a rising tone yields more incorrect answers (p = .006, Std.E. = .046). Stimuli with a rise are 41.7% more likely to lead to incorrect hits compared to a fall. Differences in scores of incorrect stress identification between fall and fall-rise were not significant.

3.2. Direction of misidentification: beginning or end?

Across incorrect answers, *AppLang* favoured a right-shifting of stress 30% of the time (vs. 20% for *Eng*). However, only 17.4% of the incorrectly chosen syllables were word-initial and 10.4% were word-final. Incorrect scores on initial syllables are higher than for final partly due to the presence of a secondary or tertiary stress in some of the words (**VIsualisation*,**ORganisation*). Figures drop to 2.5% when the final syllable contain a monophthong. Namely, the proportion of word-final syllables chosen as primary stress in both groups was considerably higher when they ended in a stress-imposing suffix containing a final diphthong (18%-46% of incorrect syllables).

Table 2 displays the aggregated percentages of stress identification on last syllable by suffix(es) and group. Although average scores by group are reasonably low, when the suffix contains a word-final diphthong, scores are at least twice as high, which indicates that vowel quality in final unstressed syllables may enhance L1 transfer of identifying prominence on the final syllable. In the two words without suffix (*interpret* and *develop*), *AppLang* chose the final syllable 4 times as much as *Eng*, who opted for initial stress (89.2% of misidentification on these two words). However, it is rather on the penultimate syllable that most common stress misidentification occurs in words of 3 syllables, with a final unstressed diphthong, and with a correct stress on the first syllable (e.g. ˌ*symbolise*, ˌ*celebrate*, ˌ*notify*); *Eng*: 78.9% of incorrect responses vs *AppLang*: 65%). And for the same word-type with *-ing*, stress on the penultimate remains high (55.7%, both cohorts), albeit with a more even share of stresses identified on the ante-penultimate syllables (43.9%, both cohorts).

Eng	7.5%	AppLang	11.7%
'ATE	24.5%	'ATE	25.5%
AT'ING	0.4%	AT'ING	0.4%
I'FY	21%	I'FY	46.50%
IFY'ING	0%	IFY'ING	0%
'IZE	42.2%	'IZE	42.3%
IZ'ING	2%	IZ'ING	2%
I'BLE	4.8%	I'BLE	2.7%
I'TY	0.6%	I'TY	2.2%
'TION	5.2%	'TION	3%
'MENT	4%	'MENT	3.4%
none	10.7%	none	42%

Table 2: % **incorrect** primary stress identification on last syllable by learner group and suffix(es). Top line: aggregate % incorrect primary stress on **last syllable only**. *Eng*: n=26; *AppLang*: n=29; 6 *-ate* and *-ating* words respectively, 4 *-ify*, *-ifying*, *-i/yse* and *-i/ysing* words respectively x 3 tones x 2 repetitions. None = *interpret* and *develop*.

3.3. Misidentification by suffix & combination of suffixes

In both groups of learners, primary stress in words with *-tion* (e.g. *simplification*) and *-ize+-ing* (e.g. *symbolising*) exhibited higher scores of misidentification (*Eng*: 46.9% *-tion*, 52.1% *-i/ysing*; *AppLang*: 57.9% *-tion*, 78.8%). In addition, a systematic increase of incorrect responses is observed when a strong suffix containing a diphthong is combined with *-ing* (Fig. 2), with a majority of stress misidentification on the diphthong contained in the stress-imposing suffix: 54% amongst all other possibe answers for both cohorts. Only the word *interrogating* never followed this pattern, while *realising* followed it 93% of the time (**reaLISing*).

4. DISCUSSION

The results presented above show that *Eng* with some formal training and explicit knowledge on English lexical stress (only during the previous semester) perceived the correct primary stress better

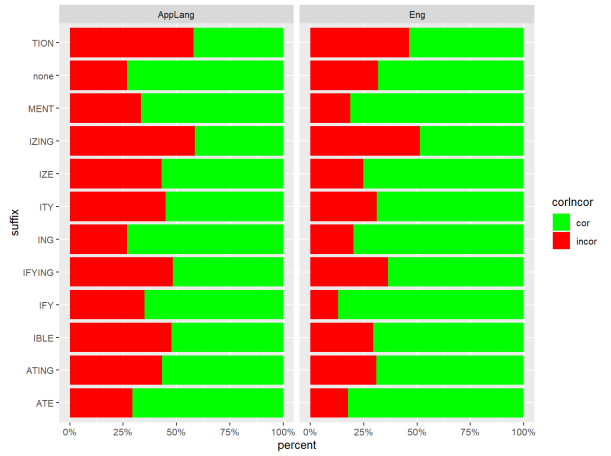


Figure 2: % cor/inc response by suffix and group. *Eng*: n=26; *AppLang*: n=29; 5 *-tion* words, 2 words without suffix, the word *development*, 4 *-ify*, *-ifying*, *-i/yse* and *-i/ysing* words repectively, 16 *-ing* words, 4 *-igible* words and 6 *-ate* and *-ating* words repectively x 3 tones x 2 repetitions.

than *AppLang* (69% vs. 56%). While this difference suggests the impact of formal training and explicit knowledge, it is also possible that the two cohorts initially had a different proficiency level of English upon entering university.

Tone seems to play a certain role in the perception of primary stress, which led to more misidentification on words carrying a rising tone, suggesting a possible impact of the prosodic pattern of the phrase accent in their L1. This should be examined further in a production experiment.

Contrarily to the expectations based on simple contrastive analysis on prosody, primary stress was rarely identified on initial or final syllables overall. However, stress-imposing endings containing a diphthong (/aɪ/ in *-ise* & *-ify*, /eɪ/ in *-ate*) were more often incorrectly identified as stressed on the last syllable. [14] also remarks that Brazilian learners of English at university level have similar strategies in production but that this may not be confined to Portuguese-speaking learners. [17] made a specific remark on the suffix *-ise* which is often stressed by learners of English as a foreign language due to disyllabics like *demise*, *arise* being stressed on the final syllable. Although this should be investigated further in a production experiment, the difficulties raised by [17] and [14] may account for the higher proportions of stress identification on these suffixes carrying a diphthong, and more specifically, on the considerably higher incorrect scores observed on the suffix *-ise* regardless of the learners' L1. A more systematic study should be extended to learners of L1s other than Romance languages.

Unexpected results concerned the above-mentioned suffixes when followed by *-ing*, which led to considerably higher error rates than when they are not. The same pattern was also found in a preliminary survey on paper carried out on the same words prior to the perception test with a larger number of students from the same university programmes and year. It is possible that students over-generalise the Latin stress rule in English by re-organising primary stress on the heavy penultimate syllable (with a diphthong, at least) and potentially over-generalising the *-ion* rule for which the stressed vowel is free, *i.e.* either diphthong or long (except when the vowel letter is the monograph <i>, [13]). This over-generalisation can be accounted for by the fact that *-ion* endings are the most frequent suffixes in Academic English [18].

5. CONCLUSION

The present study indicates that French-speaking learners of English are far from identifying lexical stress at random out of *stress-deafness* and that they show interlanguage perceptual patterns beyond expected transfers from their L1 (initial / final prominence) and mobilise multiple cues such as F0, the presence of a suffix or the lack of reduction of unstressed vowels to detect primary stress in English. Over-generalisations of stress rules (e.g. *-ion* stress rule) are an indicator of learners' strategies to inhibit transfers from their L1. Awareness raising on such strategies may help the learner to go further towards a more accurate perception and acquisition of primary stress in English.

6. REFERENCES

[1] E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler, "A destressing 'deafness' in French?" *Journal of Memory and Language*, vol. 36, pp. 406–421, 1997.

[2] S. Schwab, N. Giroud, M. Meyer, and V. Dellwo, "Working memory and not acoustic sensitivity is related to stress processing ability in a foreign language: An ERP study," *Journal of Neurolinguistics*, vol. 55, 2020, article number 100897.

[3] A. Tremblay, "Phonetic variability and the variable perception of L2 word stress by French Canadian listeners," *International Journal of Bilingualism*, vol. 13, pp. 35–62, 2009.

[4] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[5] ——, "Experiments in the perception of stress," *Language and Speech*, vol. 1, no. 2, pp. 126–152,

1958.
[6] I. Lehiste, "Influence of fundamental frequency pattern on the perception of duration," *Journal of Phonetics*, vol. 4, pp. 113–117, 1976.
[7] S.-A. Jun and C. Fougeron, "A phonological model of French intonation," in *Intonation: Analysis, modelling, and technology*, A. Botinis, Ed. Dordrecht: Kluwer, 2000, pp. 209–242.
[8] J. Vaissière, "Cross-linguistic prosodic transcription: French vs. English," in *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, N. B. Volskaya, N. D. Svetozarova, and P. A. Skrelin, Eds., 2002, pp. 147–164.
[9] W. Choi, "Towards a native OPERA hypothesis: Musicianship and English stress perception," *Language and Speech*, vol. 65, no. 3, pp. 697–712, 2022.
[10] Q. Wang, "Perception of English stress by Mandarin Chinese learners of english: An acoustic study," Ph.D. dissertation, University of Victoria, 2008.
[11] D. Frost, "The perception of word stress in English and French: Which cues for native English and French speakers?" in *EPIP1 (English Pronunciation: Issues and Practices)*, Chambéry, France, 2009, pp. 57–73.
[12] A. Cutler and D. M. Carter, "The predominance of strong initial syllables in the English vocabulary," *Computer Speech & Language*, vol. 2, no. 3, 1989.
[13] L. Guierre, *Drills in English stress-patterns*. London: Longman, 1970.
[14] B. Baptista, "Strategies for the prediction of English word stress," *IRAL*, vol. 27, no. 1, pp. 744–747, 1989.
[15] J. C. Wells, *English Intonation: An Introduction*. Cambridge University Press, 2006.
[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.1.29, retrieved 30 octomber 2020 from http://www.praat.org/," 2020.
[17] R. Kingdom, *The groundwork of English stress*. Longman, 1958.
[18] S. Suhandoko and D. R. Ningrum, "A corpus-based list of academic english derivational suffixes," *Indonesian Journal of Applied Linguistics*, vol. 10, no. 2, pp. 481–490, 1976.

Test words

-ATE: ˈcelebrate; ˈcelebrating; ˈdecorate; ˈdecorating; doˈmesticate; doˈmesticating; faˈcilitate; faˈcilitating; inˈcorporate; inˈcorporating; inˈterrogate; inˈterrogating

-I/YSE: ˈanalyse; ˈanalysing; ˈauthorise; ˈauthorising; ˈrealise; ˈrealising; ˈsymbolise; ˈsymbolising

-IFY: ˈnotify; ˈnotifying; ˈpacify; ˈpacifying; ˈpurify; ˈpurifying; ˈspecify; ˈspecifying

-IGIBLE: ˈeligible; inˈcorrigible; inˈtelligible; ˈnegligible

-TION: ˌamplifiˈcation; ˌorganiˈsation; ˌsimplifiˈcation; ˌspeciaˈlisation; ˌvisuaˈlisation

-ITY: ˌexpresˈsivity; ˌproducˈtivity; ˌreguˈlarity; ˌsimiˈlarity

(no stress-imposing ending): deˈvelop; deˈveloping; deˈvelopment; inˈterpret; inˈterpreting

Production of Mandarin Disyllabic Tones by Vietnamese Speakers

Jung-Yueh Tu

National Chengchi University
jytu@nccu.edu.tw

ABSTRACT

This study investigates the tone production of Mandarin disyllabic words by Vietnamese speakers. In the study, there were 30 Vietnamese learners of Mandarin, who were requested to produce 80 (4 tones * 4 tones * 5 words of each tonal combination) disyllabic words in Mandarin. The overall results showed that Tone 4 was the most difficult. In the first syllable, most errors were found for Tone 3 when followed by another Tone 3 (the first Tone 3 should be pronounced as a rising tone but mispronounced as Tone 3), which indicated that Vietnamese speakers tend to underapply *Mandarin third tone sandhi*. In the second syllable, most errors were found for Tone 4 when preceded by Tone 4 (mispronounced as Tone 1). The findings will be discussed in terms of relevant L2 production models as well as effects of phonetic/phonological nature of Mandarin lexical tones and the interference from L1 phonology.

Keywords: Mandarin tones, L2 tonal production, disyllabic words, Vietnamese CSL learners

1. INTRODUCTION

The goal of this study to enrich the empirical evidence for the second language (L2) research. Models accounting for the production of L2 speech, particularly at the suprasegmental level, have been proposed. For example, *Prosodic Transfer Hypothesis* [1, 2] assumes that the types of prosodic representations that can be built in the L2 are restricted by the L1 prosodic constraints. Also, it hypothesizes that L2 learners fail to produce native-like speech because they are not able to construct prosodic representations required for the L2 speech production if those representations are disallowed in the L1. In addition, *Asymmetry Hypothesis* [3] proposes that the differences of prosodic domains and rule applications between L1 and L2 will affect L2 phonological acquisition. When the prosodic domain in L1 is smaller than L2, the learners will have positive evidence for the occurrence of alternation in a larger prosodic domain, in which the L2 acquisition will eventually take place. If the prosodic domain in the L1, however, is larger than L2, the learners will need negative evidence (e.g., forms and correction) for learning the proper domain setting of L2 alternation, in which L2 acquisition will not take place.

In terms of L2 speech, learning a tonal contrast involves both perception and production. Since most of previous studies on the learning of Mandarin tones concern the processing of monosyllabic tones, this study investigates the production of disyllabic tones. In modern Mandarin, it is the disyllabic words with a high percentage (69.8% of the total words) [4] that dominate the vocabulary [4, 5, 6]. Thus, this study aims to address the problem of tonal production in a larger linguistic unit with contextual effects, focusing on disyllabic words, which can reflect the collaborative patterns of tones in speech, such as tone sandhi and tonal combination.

In Mandarin Chinese, tonal acquisition is found to be difficult for L2 learners. How to master Mandarin tones is generally quite challenging for L2 learners. Building on the previous study regarding error patterns of Mandarin disyllabic words by Japanese and Korean learners [7, 8], the current study intends to expand our understanding of L2 tonal production by investigating the Vietnamese CSL learners with experience of another tonal language profile.

1.1. Mandarin tone system

As a tonal language, Mandarin has four lexical tones, including Tone 1, a high level tone (55), Tone 2, a mid-rising tone (35), Tone 3, a low-dipping tone (214), Tone 4, a high-falling tone (51), and one neutral tone. The tonal system can be depicted either by a numerical system, with 5 indicating the highest point and 1 the lowest [9] or by a system of tonal features, H, M, and L, as presented in Table 1.

Table 1: Tonal system in Mandarin Chinese.

Tone categories	Tonal feature	Pitch value	Height
1	high-level (HH)	55	High
2	mid-rising (MH)	35	
3	low-dipping (MLH)	214	Mid
4	High-falling (HL)	51	Low

In connected speech, it is found that Tone 3 always appears as a *half third tone*, with only the low-falling contour shape [10]. A low-dipping Tone3 is converted to a rising Tone (similar to Tone 2) when followed by another Tone3, known as *third tone sandhi* [11].

1.2. Vietnamese tone system

Like Mandarin, Vietnamese is a tonal language. As for the number of tones in Vietnamese, most linguists ([11, 12, 13]) take there to be six, including thanh ngang 'flat,' a mid-level tone (33), thanh huyền 'deep,' a low-falling tone (21), thanh sắc 'sharp,'a mid-rising tone (35), thanh nặng 'heavy,' a mid-falling tone (3ˀ2ʔ), thanh hỏi 'asking,' a mid-falling(-rising) tone (313), and thanh ngã 'tumbling,' a mid-rising tone (3ˀ5). In Vietnamese, there are variations among speakers concerning how tone is realized phonetically. Also, there are differences among varieties of Vietnamese spoken in the major geographic areas (northern, central, southern) and among individuals [14]. The tone system in northern varieties is given in Table 2.

Table 2: Tonal system in (Northern) Vietnamese.

Tone ID	Tone name	Tonal feature	Pitch value
A1	ngang 'flat'	mid-level	33
A2	huyền 'deep'	low-falling (breathy)	21
B1	sắc 'sharp'	mid-rising, tense	35
B2	nặng 'heavy'	mid-falling, glottalized, heavy	3ˀ2ˀ
C1	hỏi 'asking'	mid falling(-rising), emphasis	313
C2	ngã 'tumbling'	mid rising, glottalized	3ˀ5

Revised from [11, 14]

Unlike Mandarin, Vietnamese tones do not rely only on pitch contour. Instead, Vietnamese often uses a register complex (which is a combination of phonation type, pitch, length, vowel quality and so on). The tonal feature of each tone is shown in Table 2.

1.3. Research questions

The current study aims to understand how native language experience with tones affects L2 speech production at the suprasegmental level by examining how Vietnamese speakers produce Mandarin tone in disyllabic words. The current study attempts to address three research questions: 1) What are the general error patterns of Mandarin disyllabic tones produced by Vietnamese CSL learners? 2) How do Vietnamese tone system and phonology influence tonal production of Mandarin? 3) How can the production of Mandarin disyllabic tones by Vietnamese speakers be accounted for by relevant L2 production models?

2. Method

2.1. Participants

There were 30 Vietnamese learners participating in the study (23 females, 7males; age range: 20-37 years old; mean age: 24.97 years old). All the participants have passed the self-made exam revised from TOCFL Band B test [15], which was corresponding to CEFR B1-B2 before the experiment. None had difficulty in hearing and speaking.

2.2. Stimuli

There were eighty disyllabic Mandarin words used in the stimuli list, which consisted of four tones in the first syllable, four tones in the second syllable, and five disyllabic words in each tonal combination (4 tones * 4 tones * 5 disyllabic words = 80). Those words included all Mandarin tonal combinations, excluding the neutral tone. The four tones were arranged into disyllabic words, and then 16 tonal combinations can be retrieved (4 tones * 4 tones = 16 pairs). The numbers 1, 2, 3, 4 represent Tone 1, Tone 2, Tone 3, and Tone 4, respectively. Note that the tonal combination 3-3 should be pronounced as similar to 2-3 due to the *third tone sandhi*. Consider that tonal production depends on participants' knowledge of the words with their tonal specifications, the disyllabic words were adopted from the stimuli used in [7], which were familiar to the participants. The presentation order of the words was randomized to avoid participants' expectation of a pattern. Every word was presented with Mandarin phonetic symbols (Hanyu Pinyin) and traditional Chinese characters.

2.3. Procedure

The experiment was conducted in a quiet room. The participants' utterances of the stimuli were recorded by a stand-alone microphone with a sampling frequency of 16 kHz and a resolution of 16 bit on a desktop. The recording was attained by the software PRAAT. First, the participants were familiarized with the practice section. Later, they were requested to produce and record the stimuli. They saw a disyllabic word on the screen and then produce the word when they were ready. The participants were asked to read out those disyllabic words with the correct lexical tones at a normal rate.

2.4. Analysis

The participants' recordings of the stimuli were analyzed by three phonetically-trained native speakers of Mandarin. The recordings were judged by the 3 native speakers, who labelled the tone (with a choice among the four lexical tones) of each syllable of the disyllabic words. If the production was too ambiguous to be categorized as any lexical tone, it would be labelled as 'other.' When there was any disagreement on the tone labelling, the decision made by the majority was selected. Then the acoustic analysis was carried out using the software PRAAT with visual pitch contour to decide the label of the tone.

3. Results

3.1. Overall accuracy

One-way repeated-measures ANOVAs were conducted for the accuracy rate of tonal production. The result showed that the main effect of tone was significant in the study [MSE=.020, F=3.201, p=.026(<.05)]. The accuracy of Tone 4 (M=.825, SD=.187) was significantly lower than Tone 2 (M=.931, SD=.121) [SE=.036, p=.043(<.05)], and also lower than Tone 3 (M=.910, SD=.092) and Tone1 (M=.877, SD=.149). (see Figure 1)

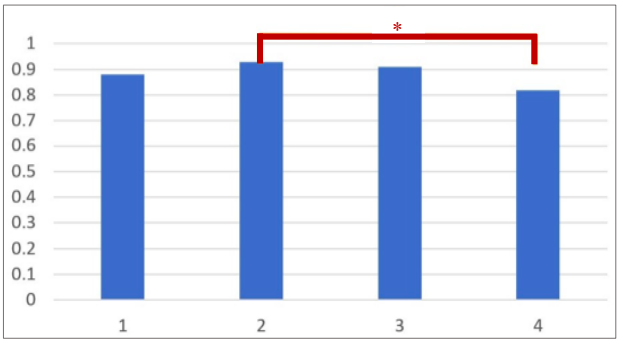


Figure 1. Accuracy of production for individual tone (*means p<.05).

3.2. Syllable effect

Paired Samples *t*-tests were conducted for the syllable effect on the accuracy of individual tone. The result showed that in Tone 1, the accuracy of the second syllable (M=.920, SD=.144) was significantly higher than the first syllable (M=.833, SD=.195) [t(29)=-2.802, p=.009(<.01)]. For the other tones, the syllable effect on accuracy was not significant (all p>.20). In Tone2, the accuracy of the second syllable (M=.945, SD=.138) was a little higher than the first syllable (M=.920, SD=.129). In Tone 3, the accuracy of the second syllable (M=.912, SD=.171) was also slightly higher than the first syllable (M=.908, SD=.126). In

Tone 4, the accuracy of the second syllable (M=.845, SD=.193) was also a little higher than the first syllable (M=.804, SD=.249). (see Figure 2)

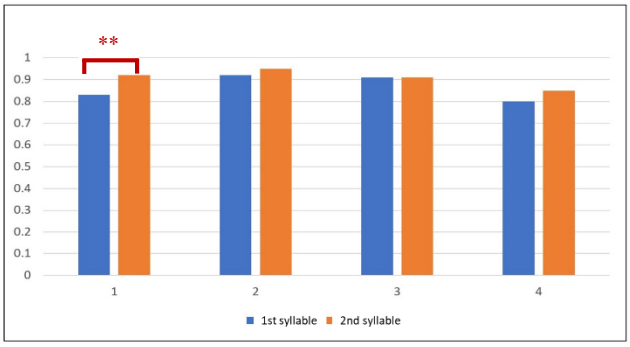


Figure 2. Accuracy for the four tones in the first and second syllables (**means p<.01)

3.3. First syllable

One-way repeated-measures ANOVAs were conducted for the accuracy of tonal production in the first syllable. The result (see Figure 3) showed that when Tone 1 in the first syllable was followed by different tones in the second syllable, its accuracy was not significantly different [MSE=.087, F=2.102, p=.104(>.05)]. When Tone 2 in the first syllable was followed by different tones in the second syllable, its accuracy was not significantly different [MSE=.034, F=.886, p=.451(>.05)]. When Tone 3 in the first syllable was followed by different tones in the second syllable, its accuracy was significantly different [MSE=.050, F=11.712, p<.001]. When Tone 4 in the first syllable was followed by different tones in the second syllable, its accuracy was not significantly different [MSE=.107, F=2.305, p=.081(>.05)].

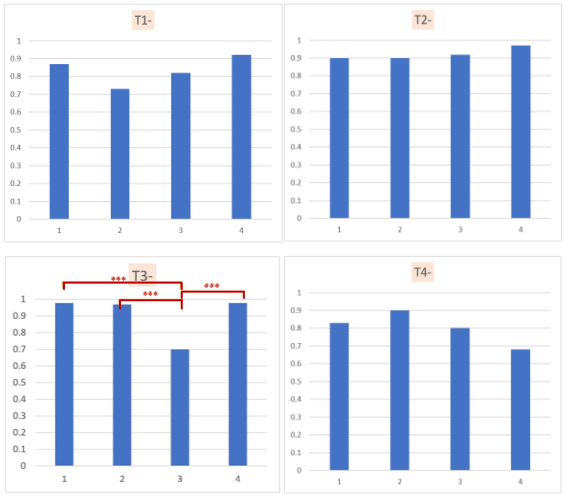


Figure 3. Accuracy of production for individual tone in the first syllable (***means p<.001)

As shown in Figure 3, the accuracy of production for Tone 3 followed by Tone 3 (M=.700, SD=.406) was significantly lower than Tone 3 followed by Tone 1 (M=.983, SD=.091), Tone 3 followed by Tone 2 (M=.966, SD=.126) and Tone 3 followed by Tone 4 (M=.983, SD=.091). Due to the *third tone sandhi* rule, when Tone 3 was followed by Tone 3, the first syllable should be pronounced as a rising tone, which is similar to Tone 2.

3.4 Second syllable

One-way repeated-measures ANOVAs were conducted for the accuracy of tonal production in the second syllable. The result showed that when Tone 1 in the second syllable was preceded by different tones in the first syllable, its accuracy was not significantly different [MSE=.049, F=2.442, p=.068(>.05)]. When Tone 2 in the second syllable was preceded by different tones in the first syllable, its accuracy was also not significantly different [MSE=.036 , F=2.392, p=.072(>.05)]. When Tone 3 in the second syllable was preceded by different tones in the first syllable, its accuracy was not significantly different as well [MSE=.53, F=0.978, p=.406(>.05)]. When Tone 4 in the second syllable was preceded by different tones in the first syllable, its accuracy was significantly different [MSE=.067, F=5.302, p=.002(<.01)].

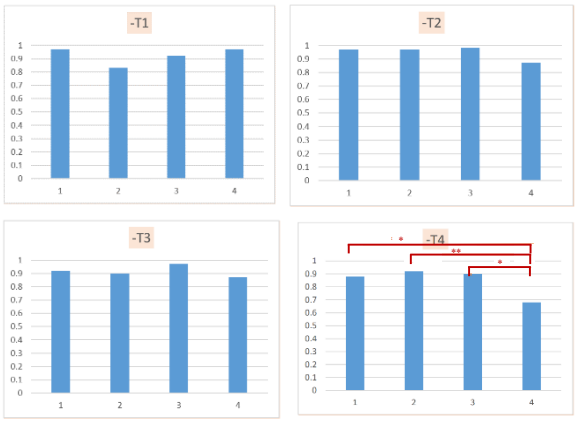


Figure 4. Accuracy of production for individual tone in the second syllable (**means p< .01, ***means p<.001)

As shown in Figure 4, the accuracy of production for Tone 4 preceded by Tone 4 (M=.683, SD=.352) was significantly lower than Tone 4 preceded by Tone 1 (M=.883, SD=.252), Tone 4 preceded by Tone 2 (M=.917, SD=.190) and Tone 4 preceded by Tone 3 (M=.9, SD=.359).

3.5 Error pattern in each tonal combination

The tonal error matrix in each combination for the first syllable and second syllable was shown in Table 3 and 4, respectively. Note that according to the *third tone sandhi* rule, a T3 followed by another T3 should be pronounced as a rising tone, similar to T2. Thus, in the tonal combination 3-3, the realization of T3 as T2 was presented as correct response. The correct response was highlighted in blue. If the accuracy rate of response was lower than 70%, the wrong response would be highlighted in pink.

Table 3. Tonal error matrix-1st syllable

		1	2	3	4	others
T1	11	0.88	0.00	0.00	0.07	0.05
	12	0.72	0.00	0.00	0.20	0.08
	13	0.82	0.02	0.00	0.15	0.02
	14	0.92	0.00	0.02	0.07	0.00
		1	2	3	4	others
T2	21	0.00	0.90	0.05	0.00	0.05
	22	0.00	0.90	0.07	0.00	0.03
	23	0.00	0.90	0.07	0.00	0.03
	24	0.00	0.97	0.02	0.00	0.02
		1	2	3	4	others
T3	31	0.00	0.02	0.98	0.00	0.00
	32	0.00	0.00	0.97	0.00	0.03
	33	0.00	0.70	0.22	0.00	0.08
	34	0.00	0.02	0.98	0.00	0.00
		1	2	3	4	others
T4	41	0.13	0.00	0.00	0.83	0.03
	42	0.08	0.00	0.00	0.90	0.02
	43	0.18	0.00	0.00	0.80	0.02
	44	0.27	0.00	0.00	0.68	0.05

Table 4. Tonal error matrix-2nd syllable

		1	2	3	4	others
T1	11	0.97	0.00	0.00	0.03	0.00
	21	0.83	0.00	0.00	0.12	0.05
	31	0.92	0.00	0.02	0.03	0.03
	41	0.97	0.00	0.00	0.03	0.00
		1	2	3	4	others
T2	12	0.00	0.97	0.03	0.00	0.00
	22	0.00	0.97	0.02	0.00	0.02
	32	0.00	0.98	0.02	0.00	0.00
	43	0.00	0.87	0.08	0.00	0.05
		1	2	3	4	others
T3	13	0.00	0.07	0.92	0.00	0.02
	23	0.00	0.05	0.90	0.00	0.05
	33	0.00	0.00	0.97	0.00	0.03
	43	0.00	0.08	0.87	0.00	0.05
		1	2	3	4	others
T4	14	0.10	0.00	0.00	0.88	0.02
	24	0.08	0.00	0.00	0.92	0.00
	34	0.05	0.00	0.00	0.90	0.05
	44	0.28	0.00	0.00	0.68	0.03

4. DISCUSSION

The overall accuracy rate for individual tone showed that the percent correct of T4 was significantly lower than T2. The result indicated a level of difficulty: T4> T1 & T3 > T2. That is, T4 is the most difficult and T2 is the least difficult. In addition, the syllable effect was found for Tone 1. That is, the accuracy of T1 in the second syllable was significantly higher than that in the first syllable, as indicated that the participants were better at producing Tone 1 correctly in the second syllable. Moreover, in the first syllable, the accuracy of production for Tone 3 followed by Tone 3 was significantly lower than that followed by the other tones. This result implied that the participants tend to *underapply* Mandarin *third tone sandhi*, which means they did not always pronounce the Tone 3 followed by another T3 as a rising tone (similar to T2) in the disyllabic words. In the second syllable, the accuracy of production for Tone 4 preceded by Tone 4 was significantly lower than that preceded by the other tones.

With regard to the error matrix in each tonal combination, most errors in the first syllable were found in the tone pairs 3-3 and 4-4. In the tone pair 3-3, Tone 3 in the first syllable was mostly mispronounced as Tone 3. In the tone pair 4-4, Tone 4 in the first syllable was mostly mispronounced as Tone 1. In addition, most errors in the second syllable were found in the tone pair 4-4, in which Tone 4 in the second syllable was mostly mispronounced as Tone 1. As previously mentioned, Tone 3 in the first syllable was mostly mispronounced as Tone 3 indicated that the participants tend to underapply the *third tone sandhi* rule. Some researchers document that in Vietnamese, sandhi occurs only in reduplicated constructions and that sandhi occurs only variably and does not apply to register, which must remain unchanging [16]. Thus, it is probably due to language transfer from Vietnamese phonology. Furthermore, the effects of Vietnamese phonology in the production of Mandarin tones were revealed in the tonal confusion between Tone 1 and Tone 4. As reviewed in 1.2., Vietnamese has six lexical tones. In Mandarin, Tone 1 (55) and Tone 4 (51) both start with a high pitch (5) while in Vietnamese, there is a mid-level tone (33) but no high-falling tone and the highest pitch of tones is (3). Therefore, the Vietnamese speakers tend to mispronounce Mandarin Tone 4 as Tone 1 for both the first syllable and second syllable in the disyllabic words.

The findings in this study suggest that the Mandarin tonal error patterns produced by Vietnamese learners were consistent with the assumptions in *Prosodic Transfer Hypothesis* and *Asymmetry Hypothesis*. The prosodic representations and rule applications in Mandarin bring about difficulties in L2 tonal acquisition for Vietnamese learners.

5. CONCLUSION

The current study investigated the tonal production of Mandarin disyllabic words by Vietnamese learners. The results showed that most errors were found in the tone pairs 3-3 and 4-4. The Vietnamese learners tend to underapply the *third tone sandhi* rule to Mandarin disyllabic words and mispronounce Tone 4 as Tone 1.

6.ACKNOWLEDGEMENTS

This study is funded by National Science and Technology Council, Taiwan (Grant No. MOST 110-2410-H-004-046). I thank Mei-chi Chao, Wei-ching Wang at National Chengchi University and Chih-yu Liao at University of Taipei for their help with partial data collection and analysis.

7.REFERENCES

[1] Goad, H., White, L. & Steele, J. 2003. Missing inflection in L2 acquisition: Defective syntax or L1-constrained prosodic representations? *Canadian Journal of linguistics, 48*, 243-263.

[2] Goad, H. & White, L. 2006. Prosodic transfer: L1 effetcs on the production L2 determiners. In: Bamman, D., Magnitskaia T., Zaller C. (eds), *Proceedings of the 30th Annual Boston University Conference on Language Development*, 231-224.

[3] Young-Scholten, M. 1994. On positive evidence and ultimate attainment in L2 phonology. *Second language research, 10*, 193-214.

[4] Chang, K.J., Chen, L.M, Lee, N.C. 2006.Tonal errors of Japanese students learning Chinese: A study of disyllabic words. *Computational Linguistics and Chinese Language Processing, 11*(3), 281-296.

[5] Duanmu, S.1999. Stress and the development of disyllabic words in Chinese. *Diachronica XVI, 1*, 1-35.

[6] He, K., Li, D.1987. *Xiandai Hanyu san qian changyong ci biao [Three thousand most commonly used words in modern Chinese]*. Beijing: Beijing Shifan Daxue Chubanshe.

[7] Tu, J., Hsiung, Y., Wu, M., Sung, Y. 2014. Error patterns of Mandarin disyllabic tones by Japanese learners. *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech-2014)*, 2558-2562.

[8] Tu, J., Hsiung, Y., Cha, J., Wu, M. & Sung, Y. 2016. Tone production of Mandarin disyllabic words by Korean learners. *Proceedings of Speech Prosody 8*, 375-379.

[9] Chao, Y. R.1930. A system of tone letters. *La Maître phonétique 45*, 24-27.

[10] Chao, Y. R.1955. Mandarin Primer, Harvard University Press.

[11] Doan, T.T. 2003. *A Concise Vietnamese Grammar*. Ha Noi: NXB The Gioi.

[12] Han, M. S. 1969. Vietnamese tones. *Studies in the Phonology of Asian Languages 8.* Acoustic Phonetics Research Laboratory, University of Southern California.

[13] Han, M.S., Kim, K.O. 1974. Phonetic variation of Vietnamese tones in disyllabic utterances. *Journal of Phonetics 2*, 223-232.

[14] Vietnamese phonology https://en.wikipedia.org/wiki/Vietnamese_phonology, retrieved on December, 30, 2022.

[15] Test of Chinese as Foreign Language (TOCFL): https://tocfl.edu.tw/index.php/exam/test/page/1, retrieved on April, 16, 2023.

[16] Nguyen, A.-T. T., Ingram, J. C. L. 2007. Stress and tone sandhi in Vietnamese tone reduplications. *Mon Khmer Studies 37*, 15-39.

PERCEPTION OF JAPANESE CONSONANT LENGTH BY ADVANCED LEARNERS FROM MANDARIN- AND MONGOLIAN-SPEAKING BACKGROUNDS

Kimiko Tsukada[1,2], Yurong[3]

[1]Macquarie University, [2]The University of Melbourne, [3]Inner Mongolia University
kimiko.tsukada@gmail.com, umyurong@yahoo.co.jp

ABSTRACT

Japanese uses durational variation contrastively for both vowels and consonants. It is widely acknowledged that length contrast is difficult for non-native speakers from diverse L1 (first language) backgrounds. We examined the perception of Japanese consonant length (i.e., short/singleton vs long/geminate) by advanced learners from Mandarin- and Mongolian-speaking backgrounds.

Unlike Japanese, consonant length is not contrastive in Mandarin or Mongolian. However, vowel length is contrastive in Mongolian. Furthermore, unlike Japanese and Mandarin which predominantly use open syllables, Mongolian frequently uses closed syllables and permits a wide range of consonants in coda position.

The participants responded to 200 trials via an AXB discrimination task. The overall mean discrimination accuracy was 86%, 96% and 99% for the Mandarin, Mongolian and Japanese groups, respectively. The Mongolian group's advantage over the Mandarin group suggests that the above-mentioned L1-related factors may continue to affect the processing of Japanese consonant length even for highly advanced learners.

Keywords: Japanese consonant length, short/singleton vs long/geminate, advanced learners, Mandarin, Mongolian

1. INTRODUCTION

Contrastive use of consonant length is one of the most prominent characteristics of Japanese phonetics and phonology [1-4] and is known to pose difficulties to non-native learners from diverse first language (L1) backgrounds including Mandarin and Mongolian, i.e., target languages in the present study [5-9].

Consonant length is not contrastive in Mandarin or Mongolian at the level of words [10], but vowel length is contrastive in Mongolian [11]. Furthermore, unlike Japanese and Mandarin which predominantly use open syllables [12, 13], Mongolian frequently uses closed syllables and a wide range of consonants occur in coda position [14].

Mandarin places even greater restrictions than Japanese on the distribution of coda consonants and only two nasal stops (/n/ and /ŋ/) are allowed in that syllable position [10, 15, 16]. Given that geminates are one of the exceptional consonant categories that may close a syllable in Japanese [13], the difference in the preferred L1 syllable structure may also play a significant role in cross-language speech processing.

Furthermore, Mongolian learners of Japanese who participated in the present study were from the Inner Mongolia Autonomous Region in China and Mongolian-Mandarin sequential bilinguals. The above-mentioned cross-linguistic phonetic characteristics, together with Mongolian speakers' exposure to both Mongolian and Mandarin since childhood, were expected to facilitate and further boost their perception of Japanese singleton/geminate in comparison with Mandarin speakers. Indeed, the Mongolian speakers were shown to discriminate Japanese consonant length contrasts more accurately than did the Mandarin speakers regardless of Japanese experience [17, 18].

L1 effects in cross-language speech processing may be most obvious for naïve listeners or inexperienced learners [19]. However, L1 tone language advantage, specifically Vietnamese over English, was observed in Mandarin tone learning even for advanced learners [20]. Research involving highly advanced second/foreign language (L2/FL) learners in languages other than English is still limited. We hope to fill this gap by examining if and how advanced learners of Japanese from Mandarin- and Mongolian-speaking backgrounds differ from each other on the one hand and from native Japanese speakers on the other hand in their perception of Japanese singleton/geminate contrasts. Cross-language comparison involving multiple groups of learners differing in L1 and proficiency levels would advance our current understanding of universal vs language-specific characteristics of how difficult L2/FL sounds are processed and learned. The findings would have implications for extensive L2/FL experience and ultimate attainment in cross-language speech learning as well as pedagogy of Japanese pronunciation.

2. METHOD

2.1. Stimuli preparation

2.1.1. Speakers

The experimental stimuli and procedure were identical to those used in our previous research ([21, 22]). Six (3 males, 3 females) native speakers of Japanese participated in the recording sessions, which lasted between 45 and 60 minutes. The speakers' age ranged from late twenties to early forties. According to self-report, which was confirmed by the first author (native speaker of Japanese originally from Tokyo), all speakers spoke standard Japanese, having been born or having spent most of their life in the Kanto region surrounding the Greater Tokyo Area. The speakers were recorded in the recording studio at the National Institute of Japanese Language and Linguistics, Tokyo.

2.1.2. Speech materials

	Singleton		Geminate	
	word	gloss	word	gloss
Alvolar	he**t**a	unskilled	he**tt**a	decreased
	ka**t**o	transient	ka**tt**o	cut
	ma**t**e	wait	ma**tt**e	waiting
	o**t**o	sound	o**tt**o	husband
	sa**t**e	well, then	sa**tt**e	leaving
	wa**t**a	cotton	wa**tt**a	broke
Velar	a**k**e	open	a**kk**e	appalled
	ha**k**a	grave	ha**kk**a	mint
	i**k**a	below	i**kk**a	lesson one
	ka**k**o	past	ka**kk**o	parenthesis
	sa**k**a	slope	sa**kk**a	author
	shi**k**e	rough sea	shi**kk**e	humidity

Table 1: Twelve pairs of Japanese words used with target sounds underlined and bolded.

Table 1 shows 12 Japanese word pairs used in this study. The /(C)V**C(C)**V/ tokens contained singleton (n = 96) or geminate (n = 96) consonants intervocalically (underlined and bolded). Only tokens with stops were considered in this study. As voiced geminates are limited in Japanese [2, 23, 24], only voiceless stops (/t, k/) were used. On average, the closure durations were 96 ms and 262 ms for singletons and geminates, respectively. The geminate-to-singleton ratios were 2.7 for alveolars (/t/-/t:/) and 2.8 for velars (/k/-/k:/), respectively. These durational values are in good agreement with what has been reported in previous research [e.g., 25] (see, however, [24] for alveolars).

2.2. Participants

Three groups of young adults participated in an AXB discrimination task. The first group consisted of 10 (5 males, 5 females) native speakers of Mandarin (*mean age* = 24.9 years, *sd* = 1.9) who were on short-term exchange and living in Tokyo, Japan for less than 1 year. They were born and raised in various provinces of China. The second group consisted of 10 (5 males, 5 females) native speakers of Mongolian (*mean age* = 23.3 years, *sd* = 2.0). Two of the male Mongolians lived in Japan for more than 3 years while the other three lived in Japan for less than 1 year. One of the female Mongolians was staying in Tokyo for less than 2 months while the rest participated in the study in China. All non-native learners had passed JLPT (Japanese Language Proficiency Test, according to which, the easiest level is N5 and the most difficult level is N1) at N1 and were considered highly advanced learners. The last and a control group consisted of 10 (2 males, 8 females) native speakers of Japanese (*mean age* = 21.0 years, *sd* = 0.8) who were students at University of Oregon in Eugene, OR, USA. All Japanese speakers were born and spent the majority of their life in Japan. Their mean length of stay in the US was 0.4 years (*sd* = 0.22) at the time of participation. None of the Japanese speakers participated in the recording sessions. According to self-report, all participants had normal hearing.

All participants were tested individually in a session lasting approximately 30 to 40 minutes in a sound-attenuated laboratory or a quiet room at their own university. The experimental session was self-paced. The participants heard the stimuli at a self-selected, comfortable amplitude level over the high-quality headphones on a computer.

2.3. Procedure

The participants completed a two-alternative forced-choice AXB discrimination task, in which they were asked to listen to trials arranged in a triad (A-X-B). The presentation of the stimuli and the collection of perception data were controlled by the PRAAT program [26]. In the AXB task, the first (A) and third (B) tokens always came from different length categories, and the participants had to decide whether the second token (X) belonged to the same category as A (e.g., '$kato_2$'-'$kato_1$'-'$katto_3$') or B (e.g., '$kako_3$'-'$kakko_1$'-'$kakko_2$'; where the subscripts indicate different speakers).

The participants listened to a total of 200 trials. The first eight trials were for practice and were not analyzed. The three tokens in all trials were spoken by three different speakers. Thus, X was never acoustically identical to either A or B. This was to ensure that the participants focused on relevant

phonetic characteristics that group two tokens as members of the same length category without being distracted by audible but phonetically irrelevant within-category variation (e.g., in voice quality). This was considered a reasonable measure of participants' perceptual capabilities in real world situations [27]. All possible AB combinations (i.e., AAB, ABB, BAA, and BBA, 48 trials each) were tested.

The participants were given two ('A', 'B') response choices on the computer screen. They were asked to select the option 'A' if they thought that the first two tokens in the AXB sequence were the same and to select the option 'B' if they thought that the last two tokens were the same. No feedback was provided during the experimental sessions. The participants could take a break after every 50 trials if they wished. The participants were required to respond to each trial, and they were told to guess if uncertain. A trial could be replayed as many times as the participants wished to reduce their anxiety, but responses could not be changed once given. The interstimulus interval in all trials was 0.5 s.

3. RESULTS

We used R version 3.6.0 for statistical analyses and data visualization reported below [28]. The packages used include ez [29] and tidyverse [30].

3.1. Overall results

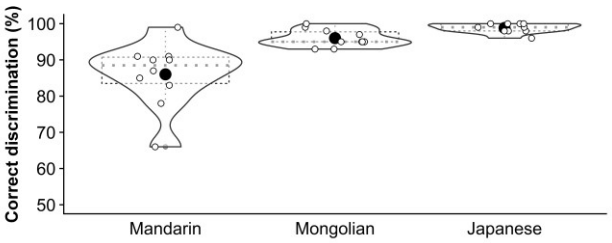


Figure 1: Accuracy (%) of length discrimination by three groups of participants. The horizontal line and the black circle in each box indicate the median and mean, respectively.

Figure 1 shows the distributions of percentages of correct discrimination by the three groups of participants. The overall mean discrimination accuracy was 86%, 96% and 99% for the Mandarin, Mongolian and Japanese groups, respectively. The Japanese group was at near ceiling with little individual variation. Of the two learner groups, the Mandarin group was much more variable than the Mongolian group as clearly seen in Figure 1.

One-way analysis of variance (ANOVA) with group (Mandarin, Mongolian, Japanese) reached significance [$F(2, 27) = 15.4$, $p < .001$, $\eta_G^2 = .53$]. According to the post-hoc *t*-tests, all between-group differences were statistically significant [$t(9.4 – 13.9) = -3.2 – -4.5$, $p < .05$]. The Japanese group was significantly more accurate than the Mongolian group which, in turn, was more accurate than the Mandarin group.

3.2. Comparison of the direction of category change (Geminate (G) > Singleton (S) or Singleton > Geminate)

Figure 2 shows the distributions of percentages of correct discrimination for trials differing in the direction (from G to S, from S to G) of length category change.

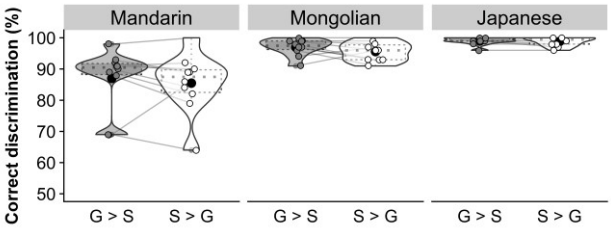


Figure 2: Accuracy (%) of length discrimination for trials differing in the direction of category change (G > S, S > G). The light lines connect individual participants' scores.

Two-way ANOVA with group (Mandarin, Mongolian, Japanese) and direction of category change (G > S, S > G) reached significance only for the main effect of group [$F(2, 27) = 15.3$, $p < .001$, $\eta_G^2 = .49$]. Lack of significant two-way interaction suggests that the overall group effect was mirrored whether the length category changed from geminate to singleton or from singleton to geminate within a trial.

3.3. Comparison of the length category (Geminate vs Singleton) of the target token (X in AXB)

Figure 3 shows the distributions of percentages of correct discrimination for trials differing in the length category (Geminate, Singleton) of the target token.

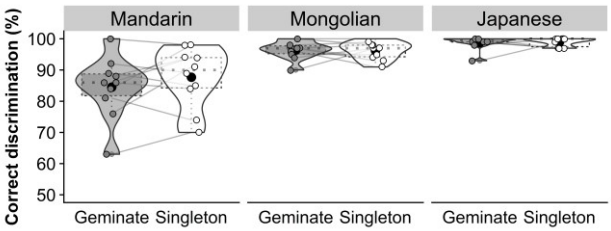


Figure 3: Accuracy (%) of length discrimination for trials differing in the length category (Geminate, Singleton) of the target token.

Two-way ANOVA with group (Mandarin, Mongolian, Japanese) and length (Geminate,

Singleton) reached significance only for the main effect of group [$F(2, 27) = 15.9$, $p < .001$, $\eta_G^2 = .49$]. Again, lack of significant two-way interaction suggests that the overall group effect was mirrored whether X in the AXB sequence was singleton or geminate.

3.4. Comparison of the place of articulation (Alveolar vs Velar) of the target token (X in AXB)

Figure 4 shows the distributions of percentages of correct discrimination for trials differing in the place of articulation (Alveolar, Velar) of the target token.

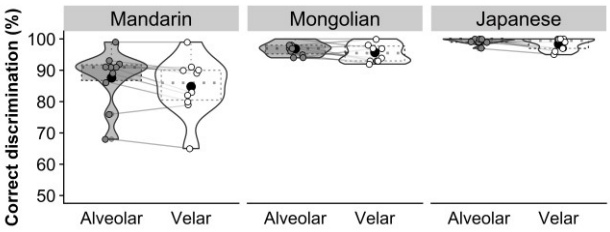


Figure 4: Accuracy (%) of length discrimination for trials differing in the place of articulation (Alveolar, Velar) of the target token.

Two-way ANOVA with group (Mandarin, Mongolian, Japanese) and place (Alveolar, Velar) reached significance for the main effects of group [$F(2, 27) = 15.4$, $p < .001$, $\eta_G^2 = .52$] and place [$F(1, 27) = 11.5$, $p < .01$, $\eta_G^2 = .02$], but not the two-way interaction. This suggests that the overall group effect was mirrored for both places of articulation and the participants were slightly more accurate when the target token was alveolar (95%) than when it was velar (93%).

3.5. Comparison of length discrimination at alveolar (/t/-/t:/) and velar (/k/-/k:/) places of articulation

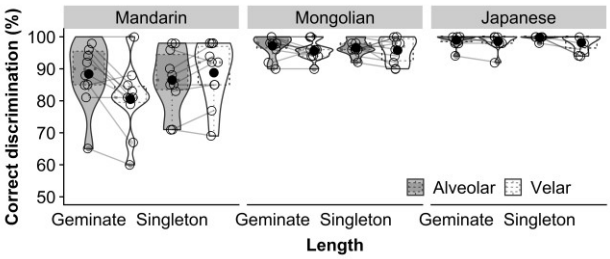


Figure 5: Accuracy (%) of length discrimination for trials differing in the place of articulation (Alveolar, Velar) and the length category (Geminate, Singleton) of the target token.

Figure 5 shows the distributions of percentages of correct discrimination for trials differing in the place of articulation (Alveolar, Velar) and the length category (Geminate, Singleton) of the target token.

Three-way ANOVA with group (Mandarin, Mongolian, Japanese), length (geminate, singleton) and place (Alveolar, Velar) reached significance for the main effects of group [$F(2, 27) = 15.9$, $p < .001$, $\eta_G^2 = .46$] and place [$F(1, 27) = 12.1$, $p < .01$, $\eta_G^2 = .02$]. Some of the interactions involving the place factor were significant [length x place: $F(1, 27) = 8.3$, $p < .01$, $\eta_G^2 = .02$, group x length x place: $F(2, 27) = 9.6$, $p < .001$, $\eta_G^2 = .04$]. As seen in Figure 5, the Mandarin group was differentially affected by the place of articulation when the target token was geminate (88% for Alveolar vs 81% for Velar) as opposed to when it was singleton (87% for Alveolar vs 89% for Velar). This pattern was absent in the Japanese and Mongolian groups.

4. DISCUSSION

This study examined how advanced learners from Mandarin- and Mongolian-speaking backgrounds perceive Japanese consonant length contrasts known to be difficult for non-native speakers. Vowel length is contrastive in Mongolian, but not in Mandarin. Further, unlike Japanese and Mandarin, Mongolian frequently uses closed syllables. We were interested in determining if the Mongolian learners of Japanese continue to outperform the Mandarin counterparts at an advanced level of proficiency.

The Mongolian group's advantage over the Mandarin group observed in this study may be due to the former group's early bilingualism (in Mongolian and Mandarin) and further examination with monolingual Mongolian learners of Japanese is needed. Our results are consistent with the view on Mandarin tone learning by L1 Vietnamese learners that "L1 experience exerts specific and lasting influences on L2 tone perception, well into advanced levels of proficiency" [20].

5. CONCLUSIONS

We investigated the perception of Japanese singleton/geminate by highly advanced learners from Mandarin- and Mongolian-speaking backgrounds to assess the lasting influence of L1. We observed that the Mongolian group resembled the control Japanese group to a greater extent than did the Mandarin group. This suggests that L1-related factors may continue to affect the processing of Japanese consonant length even for highly advanced learners.

6. ACKNOWLEDGEMENTS

This research is supported by the Sumitomo Foundation Fiscal 2020 Grant for Japan-related Research Projects (Grant Number: 208012). We thank reviewers for their time and input for this work.

7. REFERENCES

[1] Fujisaki, H., Nakamura, K., Imoto, T. 1973. Auditory perception of duration of speech and nonspeech stimuli. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics* 7, 45–64.

[2] Kawahara, S. 2015. The phonetics of sokuon, or geminate obstruents. In: Kubozono, H. (ed.), *Handbook of Japanese Phonetics and Phonology.* Walter de Gruyter, 43–78.

[3] Kubozono, H. 2013. Introduction to the special issue on Japanese geminate consonants. *Journal of East Asian Linguistics* 22, 303–06.

[4] Vance T. J. 2008. *The Sounds of Japanese.* Cambridge University Press.

[5] Han, M. S. 1992. The timing control of geminate and single stop consonants in Japanese: A challenge for non-native speakers. *Phonetica* 49, 102–127.

[6] Hung, H. Y. 2012. Perception of Japanese geminate stops among Taiwanese learners of Japanese. *Journal of the Phonetic Society of Japan* 16, 15–27.

[7] Liu, Y. L. 2014. The process of acquiring geminates by native speakers of Mongolian (in Japanese). *Nihongo Kenkyu (Japanese Language Studies)* 34, 1–13. (http://nihongo.hum.tmu.ac.jp/tmu_j/archive31-40.html)

[8] Sonu, M., Kato, H., Tajima, K., Akahane-Yamada, R., Sagisaka, Y. 2013. Non-native perception and learning of the phonemic length contrast in spoken Japanese: Training Korean listeners using words with geminate and singleton phonemes. *Journal of East Asian Linguistics* 22, 373–398.

[9] Tsukada, K, Cox, F., Hajek, J., Hirata, Y. 2018. Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Language Research* 34(2), 179–200.

[10] Feng, Q., Busà, M. G. 2022. Mandarin Chinese-speaking learners' acquisition of Italian consonant length contrast. *System* 111, Article 102938.

[11] Iivonen, A., Harnud, H. 2005. Acoustical comparison of the monophthong systems in Finnish, Mongolian and Udmurt. *JIPA* 35(1), 59–71.

[12] Kubozono, H. 1988. *Phonetics/Phonology.* Kuroshio.

[13] Shibatani, M. 1990. *The Languages of Japan.* Cambridge: Cambridge University Press.

[14] Kuribayashi, H. 1988. The development of reduced vowels and the change of syllabic structure in Mongolian. *The Study of Sounds* 22, 209–223.

[15] Lee, W.-S., Zee, E. 2003. Standard Chinese (Beijing). *JIPA* 33(1), 109–112.

[16] Meng, Y., Kotzor, S., Xu, C., Wynne, H. S. Z., Lahiri, A. 2021. Mismatch negativity (MMN) as an index of asymmetric processing of consonant duration in fake Mandarin geminates. *Neuropsychologia* 163, Article 108063.

[17] Tsukada, K., Yurong, Kim, J.-Y., Han, J.-I., Hajek, J. 2021. Cross-linguistic perception of the Japanese singleton/geminate contrast: Korean, Mandarin and Mongolian compared. *Proc.* 22nd *Interspeech* Brno, 3910–3914.

[18] Tsukada, K., Yurong. 2022. Non-native perception of the Japanese singleton/geminate contrast: Comparison of Mandarin and Mongolian speakers differing in Japanese experience. 23rd *Interspeech* Incheon, 3068–3072.

[19] Best, C. T., Tyler, M. D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, O.-S., Munro, M. J. (eds), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege.* John Banjamins Walter de Gruyter, 13–34.

[20] Pelzl, E., Liu, J., Qi, C. 2022. Native language experience with tones influences both phonetic and lexical processes when acquiring a second tonal language. *Journal of Phonetics* 95, Article 101197.

[21] Tsukada, K., Idemaru, K., Hajek, J. 2018. The effects of foreign language learning on the perception of Japanese consonant length contrasts. *Proc.* 17th *SST* Sydney, 37–40.

[22] Tsukada, K., Hajek, J. 2022. Adaptation to L3 phonology? Perception of the Japanese consonant length contrast by learners of Italian. *Proc.* 18th *SST* Canberra, 171–175.

[23] Hussein, Q., Shinohara, S. 2019. Partial devoicing of voiced geminate stops in Tokyo Japanese. *J. Acoust. Soc. Am.* 145, 149–163.

[24] Sano, S. 2019. The distribution of singleton/geminate consonants in spoken Japanese and its relation to preceding/following vowels. *Proc. 19th ICPhS* Melbourne, 1833–1837.

[25] Hayes-Harb, R. 2005. Optimal L2 speech perception: Native speakers of English and Japanese consonant length contrasts. *Journal of Language and Linguistics* 4, 1–29.

[26] Boersma, P., Weenink, D. Last viewed June 13, 2016. *Praat: Doing Phonetics by Computer* [version 6.0.19], retrieved from http://www.praat.org

[27] Strange, W., Shafer, V. L. 2008. Speech perception in second language learners: The re-education of selective perception. In: Hansen Edwards, J. G., Zampini, M. L. (eds), *Phonology and Second Language Acquisition.* John Benjamins, 153–191.

[28] R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[29] Lawrence, M. A. 2016. *ez: Easy Analysis and Visualization of Factorial Experiments.* R package version 4.4-0. https://CRAN.Rproject.org/package=ez

[30] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R. et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43), 1686.

THE CONTRIBUTION OF TEMPORAL CUES TO PERCEIVED NATIVENESS IN THE NATIVE SPEECH OF ENGLISH MIGRANTS TO AUSTRIA

Ineke Mennen[1], Ulrich Reubold[1], and Robert Mayr[2]

[1]University of Graz, [2]Cardiff Metropolitan University
ineke.mennen@uni-graz.at, johann.reubold@uni-graz.at, rmayr@cardiffmet.ac.uk

ABSTRACT

Learning a second language can affect pronunciation in the first language, and may result in the impression of non-nativeness. The role of segments and prosody in perceived non-nativeness is unknown. We investigated the role of temporal features to perceived non-nativeness in the L1 speech of English migrants to Austria, using a design where segmental and temporal cues are swapped. Results revealed that the relative importance of temporal features in listeners' impression of non-nativeness depends on whether the segmental string comes from monolingual (i.e. non-attrited) or bilingual (i.e. attrited) speakers. In the former situation, listeners attend to the transferred temporal information and judge the manipulated stimuli as less native. However, when segments originate from bilingual speakers, listeners ignore the transferred temporal information. This suggests that there needs to be some correspondence between segmental and prosodic information for listeners to attend to temporal information in their judgements of nativeness[i].

Keywords: L1 phonetic attrition, temporal cues, durationional cues, prosodic manipulation.

1. INTRODUCTION

The languages used by bilingual individuals are in constant interaction with each other (e.g. [1]). At the phonetic level, such interaction typically leads to foreign accented speech, where a speaker's second language (L2) accent retains perceivable traces of an individual's native language (L1) accent (see [2] for an overview). Far less is known about the opposite effect, i.e. that the L2 can exert a long-term influence on a bilingual's L1 pronunciation with traces of the L2 system occuring in a speaker's L1 accent. This phenomenon is usually referred to as L1 phonetic attrition (e.g. [3]). Listeners are very sensitive to such L2-induced changes to L1 pronunciation: accent rating studies show that listeners often detect a non-native accent when judging the L1 pronunciation of late-sequential bilinguals who have been long-term exposed to an L2 [4, 5, 6].

There is ample evidence for L2-induced changes to L1 pronunciation in a wide range of segmental and prosodic areas of speech production. Changes at the segmental level have been observed for among others voice onset time in plosives, formants of vowels and laterals, rhotics, and sibilants (see [7], for a recent overview). At the prosodic level, changes have been observed in the realisation of prosodic prominence, pitch range, the choice and frequency of use of intonation patterns, and how pitch accents are timed in relation to segments (see [8] for an overview).

Only few studies have tried to determine what cues listeners use when judging individuals as non-native in their L1. [9] asked monolingual Spanish listeners to rate short Spanish speech samples produced by Spanish English bilingual speakers living in the UK and indicate which features they associated with non-nativeness. Listeners predominantly listed segmental features; there were fewer comments on prosody. [10] examined the salience of different speech cues in the L1 of English migrants to Austria using a similar methodology. Contrary to [9], their results showed that listeners associated prosodic and segmental features approximately equally often with non-native speech. Given these contradictory findings, and the fact they were based on comments by listeners who may not have been able to verbalize what they based their judgments on, it remains unclear to what extent prosodic cues play a role in judgements of nativeness.

This study investigates the role of temporal features to listeners' impression of non-nativeness in the L1 speech of English migrants

to Austria. We use the prosody transplantation or morphing paradigm [11], where we swap segmental and temporal cues, so that the two levels are disentangled and their role in perceived non-nativeness can be examined. In our study, we created stimuli where the segments of monolingual native speakers of Standard Southern British English (SSBE) who live in England are transplanted ('morphed') onto the temporal (durational) features of late-sequential English-Austrian German bilinguals (i.e. English migrants to Austria), and vice versa. The resulting stimuli were then presented to monolingual English listeners and judged in an accent rating task. So far, no studies have used this method for assessing L1 attrited speech. However, transferring native temporal cues to L2-accented speech has been found to improve foreign accented ratings [12, 13]. If one assumes that listeners use similar cues in rating L1 naiveness and in rating L2 accentedness, then the prediction is that the accentedness ratings will improve when monolingual temporal cues are transferred to BIL speech but worsen in the reverse scenario.

2. METHODS

2.1. Speakers

Two groups of speakers participated in this study: (1) English-Austrian German bilingual speakers (BIL, N = 7, 3 females, 4 males), who were raised as monolingual speakers of SSBE and moved to Austria in adulthood where they acquired Austrian German as an L2; (2) monolingual speakers of SSBE residing in England (MON, N = 7, 3 females, 4 males) who have never lived outside England and reported no more than high-school level knowledge of other languages.

Six of the bilingual participants were rated for their degree of perceived nativeness in an earlier global accent rating experiment [10] and were perceived as moderately accented in their L1, with an average score of 3 on a 6-point accentedness scale ("1"="certainly native" – "6"="certainly non-native") and a range of segmental and prosodic features were identified by listeners as deviant from the L1 norm. The remaining bilingual participant was recruited at a later stage, but was also perceived (by a native SSBE listener) as moderately accented with both segmental and prosodic features contributing to this perception.

2.2. Sentence material

Twelve sentences were used to create the stimuli for this experiment. The choice of these sentences was based on the fact that the sentences consisted of a variety of statements and questions, and provided plenty of scope for the speakers to produce a variety of segmental and prosodic cues. The audio recordings of the sentences were automatically segmented using text-to-phoneme conversion and forced-alignment algorithms [14] after which they were checked and manually corrected where necessary.

2.3. Stimuli

In order to create the stimuli, we paired the MON and BIL speakers in such a way that their versions of each sentence used for morphing contained the same number of syllables and that the speakers were similar in voice quality, median pitch, and speech tempo (in terms of the total durations of the paired utterances, cf. Fig.1). This was done to counteract a possible negative influence on the quality of the morphed stimuli. This resulted in 7 MON-BIL pairs.

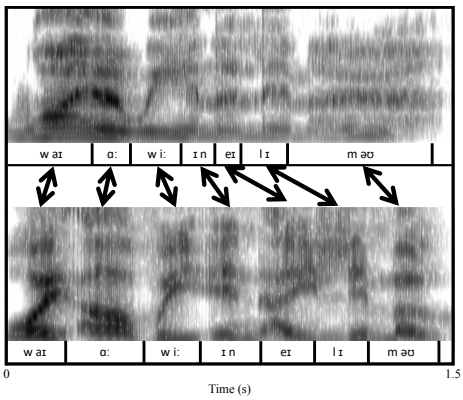


Figure 1: Duration morphing on syllable level of sentence "Why are we in a limo?" produced by a MON (top) and a BIL speaker (bottom).

We then created two sets of stimuli from the (sound level normalized) readings of the 12 sentences by the speakers: (1) unmorphed speech, which was, however, resynthesized to make sure that the listeners heard manipulated speech in both sets; (2) temporally morphed speech in which durations of syllables are

adjusted by means of the PSOLA algorithm [15][ii]. For the latter, we segmented the speech signals into syllables, measured their durations, calculated for each MON/BIL syllable pair the proportions dur_{MONO}/dur_{BIL} and vice versa, and used these as factors in duration tiers in the *To Manipulation* procedure in *Praat* [16]. This resulted in 336 stimuli, i.e. 168 stimuli per condition.

2.4. Listeners and procedure

We recruited 40 monolingual listeners (20 per condition) living in England through Prolific (https://www.prolific.co), an online research platform providing help with the recruitment of participants for online experiments (16 females, 4 males for the unmorphed condition; 18 females, 4 males for the duration morphed condition). Mean age for the unmorphed speech condition was 42 years (range 20-69), mean age for the duration morphed condition was 31 years (range 18-57). The experiment was presented to the listeners on Qualtrics [17]. Listeners were asked to use their desktop or laptop computers and a headphone for the experiment and given instructions to ensure adequate volume settings. They were informed that they would hear samples from fluent English speakers and would see an orthographic transcript of the sentence that was spoken. Samples were played in random order, and listeners were asked after hearing each sample whether it was spoken by a native or non-native speaker of English. They then had to indicate how confident they were of their choice (uncertain, semi-certain, certain). These two ratings were combined into a 6-point accentedness scale ranging from "1" = "certainly native" to "6" = "certainly non-native" – a method which is commonly used in studies on L1 attrition of speech [4, 5, 9, 10].

The experiment lasted approximately 35 minutes and listeners were paid a small fee for their participation. They could listen to each stimulus three times. Listeners were told that the samples they would hear may sound a bit artificial, that there was no right or wrong answer, and that they should respond intuitively. In order to familiarize listeners with the manipulated samples and experimental set-up, we presented three practice stimuli, randomly selected from the experimental stimuli.

2.5. Data analysis

Two measures were taken to examine whether the BIL and MON showed production differences in the temporal domain of the unmanipulated data: speech rate and VarcoSyll [18, 19], as speech rate and rhythm were expected to show cross-language differences [20, 21]. Speech rate was measured as the number of uttered syllables per second. VarcoSyll was computed as the standard deviation of a syllabic interval duration divided by the mean syllabic interval duration, multiplied by 100. It was computed for non-phrase final syllables only [22] to avoid pre-phrase boundary lengthening effects. We used paired t-tests to statistically examine the production differences in the temporal domain, unless otherwise mentioned.

To test whether potential production differences between the groups in speech rate or VarcoSyll could predict perceived nativeness in the unmanipulated stimuli, we applied linear models to the speaker-specific means of each measure and the ratings, both separately for MON and BIL, as for all speakers combined.

The accent ratings were tested for interrater reliability by means of Fleiss' Kappa [23], using the *R* package *irr* [24]. The 6-point accentedness scale was converted to an ordered factor, which then was the dependent variable in a Cumulative-Link Mixed Model using the function *clmm()* from the *R* package *ordinal* [25]. *Speaker Base* (*MON* vs. *BIL*) and *Duration* (*MON* vs. *BIL*) were treated as fixed factors, *Listeners* and *Utterance* as random factors (with random intercepts and slopes for each combination of random:fixed effects). Post-hoc comparisons were made with *emmeans* [26].

3. RESULTS

3.1. Production of temporal measures

As shown in Fig. 2a, BIL have a significantly lower speech rate in the unmorphed stimuli than MON ($t[6] = 3.0$, $p < 0.05$). No main effect was found for VarcoSyll ($t[6] = 0.3$, n.s.). However, as Fig. 2b suggests, variability of this measure is considerably higher in BIL as compared to the rather consistent MON. This was confirmed by an F-Test ($F[6,6] = 0.08$, $p < 0.01$).

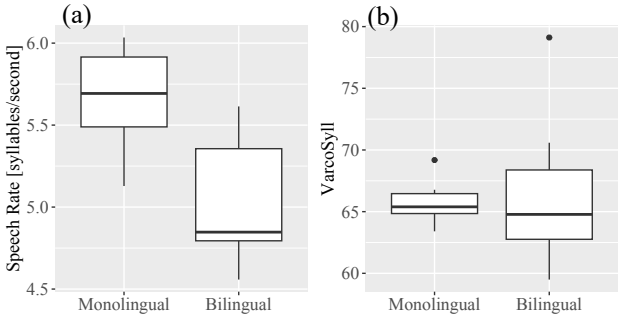


Figure 2: (a) Boxplots of speech rate and (b) VarcoSyll against Group.

3.2. Perceived non-nativeness in L1

We first tested the interrater reliability of the listeners' binary responses of (non-)nativeness. Fleiss' Kappa for the listener ratings of unmanipulated stimuli was 0.48 ($z = 85.7$, $p < 0.001$); for manipulated stimuli it was 0.43 ($z = 773$, $p < 0.001$). This constitutes 'moderate agreement' in both cases [27]. For the 6-point scale responses, Fleiss' Kappas were expectedly lower (unmanipulated: 0.26, i.e. 'fair agreement'; $z = 82.7$, $p < 0.001$); manipulated: 0.19, i.e. 'slight agreement'; $z = 64.2$, $p < 0.001$)).

Linear models showed that nativeness ratings (6-point scale) of unmanipulated stimuli could not be predicted from VarcoSyll, neither for the groups separately, nor combined. However, speech rate (Fig. 3b) significantly correlated with nativeness ratings for BIL ($F[1,5] = 12.89$, $p < 0.05$; *Adj.* $R^2 = 0.66$), MON ($F[1,5] = 8.4$, $p < 0.05$; *Adj.* $R^2 = 0.55$), and all speakers combined ($F[1,12] = 45.7$, $p < 0.001$; *Adj.* $R^2 = 0.77$).

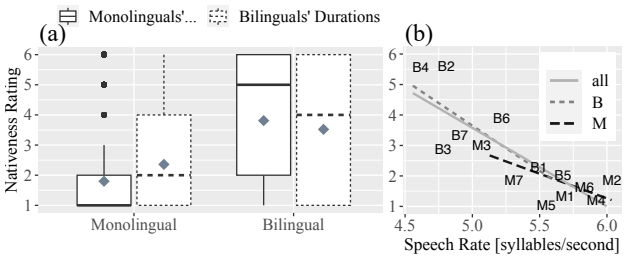


Figure 3: (a) Nativeness ratings and effect of duration on MON vs BIL speech. Diamonds show mean ratings; (b) nativeness ratings as a function of speech rate; for MON, BIL, and combined (all), with numbers indicating pairs.

Results of a Cumulative-Link Mixed Model (Fig. 3a) showed an effect of *Speaker Base*, with *BIL* rated as less native than the *MON* ($\chi^2[1] = 36.1$, $p < 0.001$). Also *Duration* did elicit significant changes in nativeness ratings ($\chi^2[1] = 6.0$, $p < 0.05$). There was a significant interaction between *Duration* and *Group* ($\chi^2[1] = 10.5$, $p < 0.01$). Post-hoc comparison showed a significant effect of *Duration* for the *MON Speaker Base* only (i.e., the manipulation led to more *non-native* responses: $p < 0.001$), but not for the *BIL Speaker Base* (n.s.). Under both conditions of *Duration*, MON and BIL *Speakerbases* differed significantly ($p < 0.001$), with the latter always being perceived as less native than the former.

4. DISCUSSION AND CONCLUSION

Results of the production data showed that BIL spoke the L1 more slowly and had a tendency of having a less stable rhythm compared to MON L1 speakers. This suggests an L2-induced influence, as Austrian German is reported to have a relatively slow speech rate [21] and research suggests that its rhythm may be situated more toward the syllable-timed end of the rhythm continuum than SSBE [20]. Moreover, the differences in speech rate were found to correlate with listeners' judgments of nativeness of the unmanipulated stimuli: slower speech sounds less native to listeners.

The influence of transferring durational cues, however, only partially confirmed our predictions. While transferring BIL durational cues to MON speech makes individuals sound less native, transferring MON duration cues to BIL speech did not have the expected positive effect on nativeness ratings. This suggests that the influence of durational cues depends on the degree to which other cues, such as segments and intonation, are native. When they conform to the native norm, as in the MON participants, durational cues influence perceptions of nativeness. However, when other cues deviate from L1 norms, as is the case in the BIL participants, the improved durational cues are simply ignored by listeners. This shows that segmental cues are more important than temporal ones, with the latter only considered when no other cues speak against them. This is in line with previous work on L1 perception of L2 speech showing that "there needs to be some degree of correspondence between segments and prosody for small deviances in prosody to be perceived" [28, p. 522]. Our results show that much the same holds true for nativeness perception in L1 attrited speech.

5. REFERENCES

[1] Green, D. W. 1998. Mental control of the bilingual lexico-semantic system. *Bil Lang Cogn* 1, 67–81.

[2] Edwards, H., Zampini, M. L. 2008. Phonology and Second Language Acquisition. Amsterdam: John Benjamins.

[3] Major, R.C. 2010. First language attrition in foreign accent perception. *Int J Bil* 14, 163–83.

[4] Bergmann, C, Nota, A., Sprenger, S.A., Schmid, M.S. 2016. L2 immersion causes non-native-like L1 pronunciation in German attriters. *J Phon* 58: 71–86.

[5] de Leeuw, E., Schmid, M.S., Mennen, I. 2010. The effects of contact on native language pronunciation in an L2 migrant context. *Bil Lang Cogn* 13: 33–40.

[6] Hopp, H. Schmid, M.S. 2013. Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Appl Psycholinguist* 34, 361–94.

[7] Reubold, U., Ditewig, S., Endes, K., Mayr, R., Mennen, I. 2021. The effect of dual language activation on L2-Induced changes in L1 speech within a code-switched paradigm. *Languages* 6: 114.

[8] Mennen, I., Reubold, U., Endes, K. & Mayr, R. 2022. Plasticity of native intonation in the L1 of English migrants to Austria. *Languages*, 7(3), 241.

[9] Mayr, R., Sánchez, D., Mennen, I.. 2020. Does teaching your native language abroad increase L1 attrition of speech? The case of Spaniards in the United Kingdom. *Languages* 5: 41.

[10] Ditewig, S., Reubold, U., Mayr, R., Mennen, I. Under review. The relation between perceived non-native features and deviations from the L1 norm in adult sequential bilinguals.

[11] Boula de Mareüil, P., Vieru-Dimulescu, B. 2006. The contribution of prosody to the perception of foreign accent. *Phonetica 63*, 247–267.

[12] Polyanskaya, L., Ordin, M., Busa, M.G., 2016. Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Lang Speech* 60 (3), 333–355.

[13] Van Maastricht, L., Zee, T., Krahmer, E., Swerts, M. 2021. The interplay of prosodic cues in the L2: How intonation, rhythm, and speech rate in speech by Spanish learners of Dutch contribute to L1 Dutch perceptions of accentedness and comprehensibility. *Speech Comm* 133, 81090.

[14] Kisler T., Reichel, U.D., Schiel, F. 2017. Multilingual processing of speech via web services. *Comput Speech Lang* 45, 326–47.

[15] Moulines, E. & Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun* 9, 453–467.

[16] Boersma, P., Weenink, D. 2006. Praat (Version 6.2.06). https://www.praat.org.

[17] Qualtrics. 2022. Qualtrics XM PlatformTM. Computer Program. Available online: http://www.qualtrics.com.

[18] Lai, C., Evanini, K., Zechner, K. 2013. Applying rhythm metrics to non-native spontaneous speech. *Proc. of SLaTE*, 159–163.

[19] Leemann, A. Kolly, M-J., Nolan, F., Li, Y. 2018. The role of segments and prosody in the identification of a speaker's dialect. *J Phon*. 68, 69–84.

[20] Kelly, N. Variability of rhythm across dialects: Austrian German. Unpublished Abstract. https://www.academia.edu/31492968/Variability_of_rhythm_across_dialects_Austrian_German.

[21] Kleber, F., Jochim, M., Klinger, N., Pucher, M., Schmid, S., Zihlmann, U. 2021. Sprechgeschwindigkeitsunterschiede zwischen den natonalen hochsprachlichen Varietäten Deutschlands, Österreichs und der Schweiz. Talk presented at the Österreichische Linguistik Tagung.

[22] Gibbon, D., Gut, U. 2001. Measuring speech rhythm. *Proc. of Eurospeech*, 95-98.

[23] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters, *Psychol. Bull.*, 76 (5), 378–382.

[24] Gamer, M., Lemon, J., Fellows, I., Singh, P. 2019. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1.

[25] Christensen, R. H. B. (2019). Ordinal: Regression Models for Ordinal Data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

[26] Lenth, R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.1-1, <https://CRAN.R-project.org/package=emmeans>.

[27] Landis, J. R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data, *Biometrics*. 33 (1) 159–174.

[28] Ulbrich, C., Mennen, I. 2016. When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accent. *Int J Bil* 20: 522–49.

[i] We would like to thank the Austrian Science Fund (FWF) for their financial support under grant number P33007-G.

[ii] While the durational morphing transferred the proportional duration of syllables from MON to BIL speakers and vice versa, there is some overlap with speech rate (measured as the number of syllables per second), as we selected the MON/BIL pairs to have a similar speech tempo (in terms of the total durations of the paired utterances).

Effects of L1 Phonotactic Constraints on Learning Novel Words

Tamami Katayama

Kumamoto University
katayamat@kumamoto-u.ac.jp

ABSTRACT

In this study, it was examined whether native Japanese speakers learn novel phonotactics that violates L1 phonotactic constraints. In Experiment 1, it was investigated whether native Japanese speakers generalize syllable position regulations of restricted consonants that occur word medially (e.g., in ba**F**.**P**ev, ti**D**.**Z**ek) to novel words in which the consonants occur at word edges (e.g., **Z**ɪb.tə**F**). The Japanese participants learned regulation of phonotactics at the syllable level rather than at the word level. In Experiment 2, it was examined whether the participants track the co-occurrence of consonants across syllables within a word. They discerned test items with the same co-occurrence of consonants between syllables (e.g., FP as in bɛ**F**.**P**ək) as in the training items (e.g., ba**F**.**P**ev) from those with different co-occurrence of consonants (e.g., FZ as in bɪ**F**.**Z**ət). The Japanese speakers also learned the co-occurrence restrictions of consonants across syllables within a word.

Keywords: L1 phonotactic constraints, syllable structure, L2 word learning

1. INTRODUCTION

Syllable structures and phonotactic constraints differ depending on the language. Infants are able to identify patterns of phonotactics of given words by being exposed themselves to speech [1][2][3]. This pattern finding skill is considered to be an essential ability for infants acquire language. It was reported that 4-year-old children used the combination of talker variability and multiple word types to generalize phonotactics [4]. Learning phonotactic regulations facilitates word learning since there is no apparent acoustic cues for word boundaries [5][6][7][8]. It is not clear whether phonotactic learning is caused by top-down processing that is mediated by the lexicon [9][10] or by experience of exposure to speech without storage of abstractions [11][12].

Dupoux, Kakehi, Hirose, Pallier, and Mehler [13] reported that phonotactics of the listener's first language affected speech perception. Native Japanese speakers perceived a vowel inside consonant clusters (VCCV) and recognized /ebzo/ as /ebuzo/, and the proficiency of the foreign language did not affect the effect of epenthesis. Dupoux et al. concluded that not only are L2 sounds assimilated into the inventories of listeners' first language but that they are distorted to conform to their L1 phonotactic regulations.

On the other hand, Trapman and Kager [14] also reported that L2 learners are able to access the phonotactics of the target language they learn. They tested L2 learners who learned L2 syllable structures that are not allowed in their first language using three different language groups: Russian, Dutch, and Spanish. Even Spanish speakers, whose legal syllable structures were fewer than those of the other two languages, were able to learn some L2 phonotactic structures that are illegal in their L1. As L2 proficiency increases, they discriminated legal codas and illegal codas in L2 similarly to the native speakers of the target language. The results can be accounted for by the episodic theory that proposes that perceptual and contextual details of each incident are stored in a listener's mental lexicon [15].

Bernard [16] reported that native English speakers learned the regularities of syllable structures when they were given non-words in an experiment and argued that phonotactic constraints can be represented as a unit of syllable. Subsequently, Bernard [17] investigated whether listeners solely depend on phonotactics at the syllable level when they learn novel words. Bernard examined whether native English speakers use information on co-occurrence of consonants across syllables when syllable positions of the consonants are invariant. In the training phase, the participants listened two times to bisyllabic non-words that included four restricted consonants (F, P, D and Z) in onset and coda positions (e.g., ba**F**.**P**ev, ti**D**.**Z**ek). In the test phase, novel words with the same co-occurrence of consonants across syllables as in the training items (e.g., bɛ**F**.**P**ək) and those with different co-occurrence (e.g., bɪ**F**.**Z**ət) were provided along with repeated training items and fillers. The participants judged whether they heard the words before and the false recognition rates were compared. Bernard found that the listeners generalized the syllable position of consonants to a novel word. Since Bernard tested the learning of artificial words using native English speakers, the question as to whether native Japanese speakers can learn novel phonotactics as native English speakers

do when it violates L1 phonotactic constraints remains unanswered.

In order to examine the effect of phonotactics on novel word learning by native Japanese speakers, the following research questions were raised:

1. Do native Japanese speakers generalize regularities of syllable positions when they learn novel words whose phonotactics is illegal in their L1?
2. Do native Japanese speakers track co-occurrence of consonants across syllables?

If they learn words at the syllable level, they will generalize the syllable positions regardless of the word positions of consonants.

2. EXPERIMENT 1

2.1. Methodology

In the present study, the methodology used by Bernard [17] was replicated, and experiments were conducted employing native Japanese speakers. In Experiment 1, it was investigated whether native Japanese speakers generalize regulations of syllable positions of restricted consonants that occur in medial positions in non-words and apply them to novel words with consonants that occur at the positions of word edges.

2.1.1. Participants

Sixteen native Japanese speakers, who are undergraduate students in Japan, participated in the experiment. Seven of them were male and nine were female and their mean age was 20.6 years. Two of them had visited Canada for one months, but the others had no experience of being in English-speaking countries. None of the participants reported a hearing impairment.

2.1.2. Design

In the training phase, following Bernard [17], artificial words composed of two consonant-vowel-consonant (CVC) syllables with four restricted consonants (/d/, /f/, /p/, and /z/) in the onset and the coda position to occur word medially (e.g., ba**F**.**P**ev, ti**D**.**Z**ek) were repeated twice (24 items x 2 conditions x 2 cycles), so that the participants were allowed to track syllable-position restrictions (e.g., P and Z as onsets with D and F as codas) and co-occurrence restrictions (e.g., FP clusters and DZ clusters: oPZ/cDF).

In the subsequent test phase, novel items with the restricted consonants in word-edge positions were presented under two conditions: items with the same syllable positions as those in the training items (e.g., **Z**ɪb.tə**F**) and items with different syllable positions from those in the training items (e.g., **F**ɪb.tə**Z**). The novel items were intermixed with two additional repetitions of the training items and presented to the participants. The participants were asked whether the items were repeated or novel. The rates of false recognition for the novel items with the same syllable positions (Same Test items) and those with different syllable positions (Different Test items) were compared. It was predicted that the participants would falsely recognize the test items with the same syllable positions more often than those with different syllable positions if they track the restricted consonants at the syllable level. This experiment made it possible to determine whether native Japanese speakers identify consonants at the syllable level rather than linear order occurrence.

2.1.3. Materials

Following Bernard's study [17], four restricted consonants (/d/, /f/, /p/ and /z/) and five vowels were used in the present study. The first syllable consisted of a lax vowel (/ɪ/, /ʌ/, /ɛ/ and /æ/), while the second syllable always had a schwa (/ə/). Two types of syllable-position restrictions and two types of co-occurrence restrictions (2 x 2) were used. One of the types of syllable-position restriction had /p/ and /z/ in the onset position (oPZ) and /d/ and /f/ in the coda position (cDF), and the other had /d/ and /f/ in the onset position (oDF) and /p/ and /z/ in the coda position (PZ). One of the types of co-occurrence restrictions was /f/ that occurs with /p/ and /d/ with /z/ (FP/DZ) (e.g., bɛ**F**.**P**əv) and the other had /f/ with /z/ and /d/ with /p/ (FZ/DP). Eight consonant-to-role-assignments were created with each restricted consonant that occurred with an equal frequency to onset and coda positions, and co-occurrence of restricted consonants was thus equally frequent. Four of them had the restricted consonants in word medial positions (e.g., ba**F**.**P**ev), while the other half had the restricted consonants in word edge positions (e.g., **P**ɪb.tə**F**). Each consonant-to-role assignment had four conditions and each condition consisted of 24 items, and thus 768 non-words were created (8 consonant-to-role assignments x 4 conditions x 24 items). Each restricted consonant occurred with an equal frequency to onset and coda positions. A native English speaker from the United States pronounced the 768 bisyllabic non-words with the stress on the first syllables in a recording studio. The recordings were stored in a computer and a phonetician who is also a native English speaker checked whether the sound stimuli were pronounced as intended. The program of the experiment was written using E-prime 3.0 [18].

2.1.4. Procedure

After filling in a questionnaire regarding their language background, each participant sat in front of the computer and followed an instruction written on the screen. The volume of the sound given through the headphone was adjusted to a comfortable level by individual participants. In the experiment, a sound file of non-words was given and the question "Have you heard it before?" with an answer of "1 for Yes" and "0 for No" appeared on the screen in Japanese after each non-word. The participants were asked to respond as to whether the word was presented before or not in the experiment using *Chronos*, a psychology software tool [19]. As soon as the response was made, the next trial was presented automatically. After the participants practiced 9 trials using filler items, they were given an opportunity to clarify what they should do in the experiment. The main session consisted of a training phase and a test phase, though transition between them was not noticeable for the participants.

In the training phase, the participants listened to 2 cycles of the training items (48 items x 2) and filler items (24 items x 2), and thus half of correct responses were "Yes" and the others were "No." In the test phase, Same Test items (24 items) and Different Test items (24 items) were intermixed with two additional repetitions of training items and fillers and they were randomly presented. Since the test items were novel, the correct responses for them were "No," while those for training items and filler items were "Yes."

2.2. Results

Table 1 shows the mean ratio of "Yes" responses for each type of items: fillers, Training items, Same Test items, and Different Test items. Since the training items and the fillers were repeatedly presented, the correct answers for them were "Yes", while the correct answers for the test items were "No". Therefore, responses for the test items indicate that the participants falsely recognized the novel test items.

Item	Syll-pos	Co-occ	Yes-recog
Filler			0.85 (0.10)
Training			0.90 (0.09)
Test	Same	Different	0.50 (0.22)
	Different	Different	0.42 (0.18)

Table 1: Ratio of "Yes" responses for each type of items

The participants differentiated the repeated items (0.89) from the test items (0.46). The "Yes" recognition ratios were 0.90 for the training items, 0.85 for the fillers, 0.50 for Same Test items, and 0.42 for Different Test items. Thus, the participants were able to differentiate Same Test items from Different Test items. Analysis using the linear mixed model was performed for the responses to the test items in order to test the effect for test type (i.e., Same Test items or Different Test items). The model included Test type (Same or Different), Training syllable position restrictions (oPZ/cDF or oDF/cPZ), and Training co-occurrence restrictions (FP/DZ or FZ/DP). The model's fit was significantly improved by the inclusion of the fixed effect of Test-type (F (1, 15) = 6.4, $p < 0.05$). The participants tracked the syllable positions in the word medial positions and differentiated the test items with the same syllable positions from those with different syllable positions, suggesting that they generalized syllable positions to the novel words even when the restricted consonants appeared at word edges.

3. EXPERIMENT 2

3.1. Methodology

In Experiment 2, the methodology used by Bernard [17] was replicated to examine whether native Japanese speakers can track co-occurrence restrictions when syllable-position restrictions were maintained.

3.1.1. Participants

Sixteen native Japanese speakers at college in Japan who did not participate in Experiment 1 were recruited as participants. Fourteen of them were female and two were male. Their mean age was 19.4 years. None of them had experience staying in English-speaking countries for more than two weeks. None of the participants reported a hearing impairment.

3.1.2. Design

Non-words that had restricted consonants in word medial positions were used in Experiment 2. One type of co-occurrence restrictions was /f/ followed by /p/ and /d/ followed by /z/ (FP/DZ) (e.g., bɛ**F**.**P**əv) and the other had /f/ followed by /z/ and /d/ followed by /p/ (FZ/DP) (e.g., bæ**F**.**Z**ət). Each restricted consonant occurred with an equal frequency to onset and coda positions and the co-occurrence of restricted consonants was thus equally frequent. In the training phase, the participants heard two repetitions of non-words with either type of co-occurrence restriction above (24 items x 2 conditions x 2 cycles) (e.g., ba**F**.**P**ev, ti**D**.**Z**ek) and were given chances to track syllable-position restrictions (e.g., oPZ and cDF) and co-occurrence restrictions (e.g., FP/DZ). In the test

phase, novel items with the same co-occurrences as those in the training items and novel items with different co-occurrences from those in the training items were intermixed with two additional repetitions of the training items and presented to the participants. It was investigated whether the participants track co-occurrence restrictions when the syllable-position restrictions were maintained.

3.1.3. Materials

The same sound stimuli as those in Experiment 1 were used, with restricted consonants occurring in the word medial positons.

3.1.4. Procedure

The same procedure as that in Experiment 1 was used.

3.2. Results

Table 2 shows the mean ratio of "Yes" responses for each type of items: fillers, Training items, Test items with the same co-occurrence (Same Test items), and test items with different co-occurrence (Different Test items). As in Experiment 1, the participants differentiated the repeated items (0.75) from the test items (0.58). False recognition ratios were 0.77 for the training items, 0.70 for the fillers, 0.64 for the test items with the same co-occurrences as the training items, and 0.53 for the test items with the different co-occurrences. The responses for the test items indicate that the participants falsely recognized the novel test items and were able to differentiate Same test items from Different test items.

Item	Syll-pos	Co-occ	Yes-recog
Filler			0.70 (0.21)
Training			0.77 (0.09)
Test	Same	Same	0.64 (0.27)
	Same	Different	0.53 (0.24)

Table 2: Ratio of "Yes" responses for each type of items

Analysis using a linear mixed model was performed for the responses to the test items in order to examine the effect for test type as in Experiment 1. The model's fit was significantly improved by the inclusion of the fixed effect of Test-type (F (1, 15) = 8.1, $p < 0.05$). The participants tracked the co-occurrence in the word medial positions and differentiated Same Test items from Different Test items, suggesting that they learned co-occurrence restrictions across syllables to occur word-medially.

4. DISCUSSION

The aim of the present study was to determine whether native Japanese speakers learn novel phonotactics even when it violates L1 phonotactic constraints. In both experiments in the present study, large proportions of the Japanese participants identified the training items and distinguished the same test items from the different test items with respect to syllable position and co-occurrence. As an answer to the first research question ("Do native Japanese speakers generalize regularities of syllable positions when they learn novel words whose phonotactics is illegal in their L1?"), the native Japanese speakers learned restrictions of syllable positions of artificial words with illegal phonotactics in their L1 even when the target consonants appeared in the different word positions (i.e., from the word medial positions to the word edges). The Japanese participants learned phonotactics at the syllable level rather than at the word level. Thus, they decomposed bisyllabic non-words into syllables and generalized syllable position of the consonants. As for the second research question ("Do native Japanese speakers track co-occurrence of consonants across syllables?"), the Japanese speakers tracked co-occurrence of the consonants across syllables as well. It is assumed that they learned novel words by relying on multiple factors available.

The results were in line with those of Bernard [17] in which native English speakers were employed. The results showing that the participants learned syllable positions and co-occurrence of consonants across syllables, despite L1 phonotactic constraints, through exposure to artificial words during training phase suggest that experience is a determinant factor in learning novel words rather than exertion of the L1 lexicon. That is, the Japanese participants did not distort their perception of illegal phonotactics in the non-words in the present study so that they could align them with their L1 inventory, which contrasts with the study by Dupoux et al. [13]. The results can be explained by the exemplar model [15] that advocates that a listener's memory of experiences for speech is stored as episodes in the listener's mental lexicon, since they were not affected by L1 phonotactic constraints. However, there is a possibility that the participants depended solely on onset without considering codas. Further study is required to examine whether native Japanese speakers track codas when onset is invariant. Furthermore, what unit native Japanese speakers track, onset followed by a vowel or by a rhyme, needs to be investigated.

7. REFERENCES

[1] Saffran, J. R., Aslin, R. N., Newport, E. L. 1996. Statistical learning by 8-month-old infants. *Science* 274, 1926-1928.

[2] Mattys, S. L., Juscyzk, P. W. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91-121.

[3] Tomasello, M. 2003. Origins of language. In: Tomasello, M. (ed), *Constructing a Language: A Usage-based Theory of Language Acquisition*. Havard University Press, 8-42.

[4] Richtsmeier, P.T., Gerken, L., Ohala, D. K. 2011. Contributions of phonetic token variability and word-type frequency to children's phonological representation. *Journal of Child Language*, 38, 951-978.

[5] Mehler, J., Dommergues, J. Y., Faruenfelder, U., Segui, J. 1981. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* 20, 298-305.

[6] Cutler, A., Mehler, J., Norris, D., Segui, J. 1992. The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology* 24, 381-410.

[7] McQueen, J. M. 1988. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language* 39, 21-46.

[8] Norris, D., McQueen, J. M., Cutler, A. Butterfield, S. 1997. The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology* 34(3), 191-243.

[9] Gow, D. W., Schoenhaut, A., Avcu, E., Ahlfors, S. P. 2021. Behavioral and neurodynamic effects of word learning on phonotactic repair. *Forntiers in Pychology*, 12, 1-15. doi: 10.338g/fpsyg.2021.590155.

[10] Dupoux, E., Pallier, C., Kakehi, K., Mehler, J. 2001. New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*, 16 (5/6), 491-505.

[11] Dell, G. S., Reed, K. D., Adams, D. R., Meyer, A. S. 2000. Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26 (6), 1355-1367. doi: 10.1037//0278-7393.26.61355.

[12] Samuel, G. A. 2020. Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of memory and language, 111.* *h*ttps://doi.org/10.1016/j.jml.2019.104070

[13] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Mehler, J. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Peformance,* 25(6), 1568-1578.

[14] Trapman, M. Kager, R. 2009. The acquisition of subset and superset phonotactic knowledge in a second language. *Language Acquisition* 16, 178-221. doi: 10.1080/10489220903011636.

[15] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review,* 105(2), 251-279.

[16] Bernard, A. 2015. An onset is an onset: Evidence from abstraction of newly-learned phonotactic constraints. *Journal of Memory and Language*, 78, 18-32. http://dx.doi.org/10.1016/j.jml.2014.09.001

[17] Bernard, A. 2017. Novel phonotactic learning: Tracking syllable-position and co-occurrence constraints. *Journal of Memory and Language,* 96, 138-154.

[18] Psychology Software Tools, Inc. [E-prime 3.0.]. 2016. Retrieved from https://support.pstnet.com/.

[19] Psychology Software Tools, Inc. [Chronos]. 2015.

PREDICTING NONNATIVE CONSONANT IDENTIFICATION: THE CASE OF ENGLISH CONSONANTS AS PERCEIVED BY NATIVE SPEAKERS OF KALAALLISUT (WEST GREENLANDIC)

Ocke-Schwen Bohn[1] and Kulunnguaq Korneliussen[2]

[1]Aarhus University, Denmark; [2]National Advisory Center, Nuuk, Greenland
[1]ocke.bohn@cc.au.dk; [2]kkor@aqqut.gl

ABSTRACT

We report two experiments which examine the perception of English initial consonants by native speakers of Kalaallisut (West Greenlandic). English allows 23 consonants in initial position, with voicing contrasts for all obstruents, whereas Kalaallisut has 12 initial consonants with no voicing contrasts. We generated predictions for the identification of English consonants from a comparison of the consonant inventories (initial position) of the two languages, and from Experiment 1, in which 12 native speakers of Kalaallisut perceptually assimilated English consonants, presented in [Ca] syllables, to their native inventory. The predictions were tested in Experiment 2, in which the participants of Experiment 1 identified the same consonants using English labels. Neither the comparison of the consonant inventories nor the results of the perceptual assimilation experiment account satisfactorily for the patterns of (mis-) identification observed in Experiment 2. We conclude our presentation with a discussion of factors which contribute to the observed patterns.

Keywords: Cross-language consonant perception, West Greenlandic, perceptual assimilation.

1. INTRODUCTION

A fairly large number of studies have been inspired by Best's Perceptual Assimilation Model (PAM) [1] to predict the level of discrimination of nonnative contrasts from cross-language perceptual assimilation patterns (e.g., [5, 8]). However, even though PAM was originally developed to predict cross-language *discrimination* by naïve listeners, this model has very rarely been used and adapted to predict the *identification* of nonnative speech sounds, with [10] being a recent exception.

The present study applies the predictions for cross-language identification as generated in [10] (and inspired by PAM predictions for discrimination) to the identification of English initial consonants by native (L1) speakers of Kalaallisut (West Greenlandic). These languages differ in that English has a fairly large consonant inventory (23 in initial position) in which all obstruents differ in voicing, whereas Kalaallisut has a smaller initial consonant inventory (12) with no voicing contrasts.

Table 1: Kalaallisut initial consonants.

	Plosive	Nasal	Fricative	Approx	Lateral
Labial	p	m		ʋ	
Alveolar	t	n			l ~ ɬ
Alv.-pal			ɕ		
Palatal					
Velar	k		ɣ	j	
Uvular	q		ʁ		

Table 2: English initial consonants.

	Plosive	Affr	Nasal	Fricat	Appr	Lateral
Labial	pʰ p		m	f v	w	
Dental				θ ð		
Alveolar	tʰ t		n	s z	ɹ ~ ɻ	l ~ ɬ
Alv.-pal		tʃ dʒ		ʃ ʒ		
Palatal					j	
Velar	kʰ k					
Glottal				h		

Specifically, Kalaallisut plosives are short-lag unaspirated, whereas English has aspiration contrasts. Only English has affricates, and it has a rich fricative inventory. The inventory of initial nasals does not differ between the languages, and the approximant inventories differ only with respect to the phonetic realization of /r/, and in that English has the velarized bilabial [w], whereas Kalaallisut has a labiodental [ʋ].

This leads one to expect that L1 Kalaallisut speakers would encounter problems especially in the identification of English stops (with respect to the voicing contrasts) and fricatives and affricates with respect to both voicing and place of articulation. The perception of English sonorants, however, could be expected to be comparatively unproblematic because both languages share [m, n, j, l ~ ɬ], but differ with respect to the phonetic realization of /r/ and with respect to the place of articulation of the approximants [w] (English) and [ʋ] (Kalaallisut).

As discussed and demonstrated by, e, g, [4], comparisons of sound inventories are, at best, a first step in attempts to predict cross-language perception. In an attempt to come up with more specific and testable predictions regarding the identification of English consonants by L1 Kalaallisut listeners, we conducted Experiment 1 in which L1 Kalaallisut

listeners perceptually assimilated English consonants to Kalaallisut categories. The results from Experiment 1 were then used to generate predictions, based on the general hypotheses presented in [10], for the identification of English consonants by L1 Kalaallisut listeners (Experiment 2).

3. EXPERIMENT 1

Experiment 1 examined the perceptual assimilation of English initial consonants to Kalaallisut. We used the results from Experiment 1 to generate predictions for L1 Kalaallisut listeners' identification accuracy of English initial consonants. Our specific predictions are derived from the general predictions for the identification of nonnative speech sounds from assimilation patterns as developed in [10] and presented below. These are related, respectively, to PAM's assimilation types a) Two-Category, b) Single Category, c) Category Goodness, and d) Uncategorized:

a) L2 phones which have a unique match in an L1 category irrespective of the goodness of fit will be accurately identified by all L2 learners, irrespective of L2 experience.
b) Assimilation of two or more L2 phones to one L1 category with comparable goodness ratings will result in inaccurate identification of these phones, irrespective of L2 experience.
c) Assimilation of two or more L2 phones to one L1 category with different goodness ratings will result in accurate identification of the better match(es) irrespective of L2 experience, and in identification accuracy depending on L2 experience for the poorer match(es).
d) The identification accuracy of L2 phones which are not consistently assimilated to any L1 category (assimilation frequency of < 50%, see [6]) will depend on L2 experience.

The predictions a), b), and the first part of c) can be tested with any nonnative listeners. Testing the second part of prediction c) and prediction d) requires listener groups differing in L2 experience. Participant availability for the present study did not allow for testing the effect of English language experience.

3.2. Methods

Twelve L1 speakers of Kalaallisut (6 f, 6 m; mean age: 22.7 years) participated. All participants were fluent speakers of their second language, Danish, and none had spent any extended period of time in an English-speaking country.

The stimuli were drawn from the Shannon et al. (1999) corpus [11] as used by the Alvin3 software [9]. Twenty-one tokens each of [Ca] syllables with [C] = [p^h, p, w, f, v, θ, ð, t^h, t, s, z, ʃ, ʒ, k^h, k, ɻ, ɬ, tʃ, dʒ, h, j, h], produced by two L1 English speakers (1f, 1m) were presented in two randomizations for a total of 84 trials with praat's *ExperimentMFC* [3].

Participants were presented with each stimulus over high-quality headphones and they selected, via mouse click, one of 11 orthographically presented response alternatives corresponding to graphemes used for Kalaallisut, i.e., *pa, ta, ka, qa, fa, sa, ga, ra, va, la, ja*. (Selection of these alternatives was based on piloting in which participants used *ka* and *ga*, and *fa* and *va*, even though the graphemes in each pair do not correspond to different Kalaallisut categories.) Immediately after selecting the Kalaallisut category corresponding to the English consonant, participants indicated the goodness of their match on a 5-point Likert scale, whose endpoints were labeled (in Kalaallisut) as *Ajorpoq* ("bad") and *Ajunngilaq* ("good"). Instructions were provided in Kalaallisut by the second author, a native Kalaallisut speaker.

3.3. Results

Tables 3a-e present the results of the perceptual assimilation (PA) task in terms of fit indices, which are derived by multiplying the proportions of identifications with the goodness rating for this match (see [8]). For example, the proportion of English [p] matched with Kalaallisut <p> was 0.583, with a mean goodness rating of 4.8, resulting in a fit index of (0.583 x 4.8 =) 2.8. In the tables below, cells of interest are highlighted with bold frames. Cells with significantly different fit indices are separated by a bold line, and cells with Single Category assimilations are not separated by a line.

Table 3a: Fit indices for English initial plosives to Kalaallisut categories.

EN Stm	Kalaallisut response				
	p	v	t	K	g
p^h	2.7				
p	2.8	1.5			
t^h			2.8		
t			4.6		
k^h				3.1	0.3)
k				3.0	0.5

Table 3a suggests problems for the identification of English [p^h, p] and [k^h, k], which are each assimilated to just one Kalaallisut category with comparable fit indices, and it predicts that [t] will be identified more correctly than [t^h] because of a significant difference in fit indices for [t] (high) vs [t^h] (low).

Table 3b: Fit indices for English initial approximants to Kalaallisut categories.

EN Stm	Kalaallisut response			
	v	l	j	r
v	4.1			
w	3.3			
ɬ	0.4	3.8		
ɻ				1.7
j			4.4	

Table 3b suggests identification problems for [v, w], which are assimilated to Kalaallisut /v/ without a significant difference in goodness of fit. [j, ɻ, ɬ] are predicted to be identified accurately because they are each uniquely assimilated to Kalaallisut categories.

Table 3c: Fit indices for English initial (labio-) dental fricatives to Kalaallisut categories.

EN Stm	Kalaallisut response		
	f	v	t
f		4.3	
v		4.1	
θ	3.1		0.3
ð		0.8	1.5
t^h			2.8
t			4.5
ʧ			1.3
ʤ			1.1

Table 3c suggests low identification accuracy for English [f, v], and higher identification accuracy for [θ] than [ð].

Table 3d: Fit indices for English initial sibilants and affricates to Kalaallisut categories.

EN Stm	Kalaallisut response			
	t	s	k	g
s		4.3		
z		2.6		
ʃ		2.4		
ʒ		1.2	0.3	0.4
ʧ	1.3	0.5	0.3	0.4
ʤ	1.1	1.2		0.3

Table 3d suggests higher identification accuracy for [s, ʃ] than their voiced counterparts [z, ʒ], respectively. The affricates [tʃ, dʒ] will not be identified correctly.

Table 3e: Fit indices for English [h] to Kalaallisut categories.

EN Stm	Kalaallisut response			
	p	v	j	g
h	0.2	0.1	0.4	0.6

Table 3e suggest that [h] is uncategorized for Kalaallisut listeners. Given their extensive experience with L2 Danish, which has a /h/ category, we predict that [h] will be identified well.

4. EXPERIMENT 2

Experiment 2 examined the identification of English initial consonants by the same L1 Kalaallisut listeners as in Experiment 1, using the same stimuli as in Experiment 1. Each stimulus was presented, in random order, three times each, resulting in (2 talkers x 23 consonants x 3 randomizations =) 138 trials. Stimuli were presented via high-quality headphones using the Alvin3 software [9]

4.1. Results and discussion

The results of Experiment 2 and their relation to the predictions are presented separately for each class of consonants in Tables 4 a-e. The tables list the stimuli on the vertical axis and the responses on the horizontal axis. Only responses of > 10% are listed.

Table 4a: Identification of English stop consonants by L1 Kalaallisut listeners (rounded % correct).

	p^h	p	t^h	t	k^h	k	v
p^h	94	6					
p	14	50					26
t^h			67				
t				81			
k^h					90	6	
k					39	61	

The prediction from the comparison of the inventories was that identification accuracy would be compromised because Kalaallisut obstruents do not differ in voicing. The PA experiment suggested that [p^h, p] and [k^h, k] would be less accurately identified than [t]. Contrary to predictions, the aspirated [p^h, k^h] are identified highly and more accurately than [t].

Table 4b: Identification of English approximants by L1 Kalaallisut listeners.

	w	v	ɬ	ɻ	j
w	71	29			
ɬ			91		
ɻ				97	
j					97

As predicted from the PA experiment and, in part, from the comparison of the inventories, the identification accuracy for [ɬ, ɻ, j] was high. These approximants were each assimilated uniquely to an L1 category. However, [w] was identified less accurately, which was predicted because [w] and [v] were both assimilated with comparable goodness ratings to Kalaallisut <v>.

Table 4c: Identification of English (labio-) dental fricatives by L1 Kalaallisut listeners.

	f	v	θ	ð	t	w
f	96					
v		70				30
θ	39		36	18		
ð			33	61		

The comparison of the inventories resulted only in the general prediction that identification of English fricatives and approximants would be difficult because of the voicing contrasts and comparably large number of places of articulation. - Contrary to predictions from the Experiment 1, [f] was identified highly, and [v] somewhat accurately even though both were assimilated with comparable goodness ratings to L1 <v>. Identification rates for both [θ] and [ð] were low. This was expected for [ð] but not for [θ], which was assimilated uniquely to <f>.

Table 4d: Identification of English sibilants and affricates by L1 Kalaallisut listeners.

	s	z	ʃ	ʒ	ʧ	ʤ
s	73		24			
z	38	33	29			
ʃ	12		79	6		
ʒ			62	10	5	
ʧ			20		39	15
ʤ					35	24

As expected, identification rates for the voiceless sibilants [s, ʃ] were higher than for their voiced counterparts [z, ʒ] because [s, ʃ] were assimilated with significantly higher fit indices to one native category than their voiced counterparts. As expected, the identification rates for the affricates were low because both [ʧ] and [ʤ] were assimilated to one native category with comparable fit indices. – Finally, [h] was identified highly accurately at 98%, consistent with its uncategorized assimilation pattern.

5. CONCLUSION

The present study examined the perception of the relatively large initial consonant inventory of English by listeners with an L1 (Kalaallisut) which has a comparably small consonant inventory. Predictions for the accuracy of English consonant identification by L1 Kalaallisut listeners were derived from a comparison of the inventories of the two languages and from a cross-language mapping experiment, in which L1 Kalaallisut listeners perceptually assimilated English initial consonants to native categories.

The comparison of the inventories was a first step in an attempt to predict identification accuracy by the nonnative listeners. The predictions were, in part, necessarily vague (for fricatives) and mostly unsuccessful in identifying identification problems (except for some of the approximants).

We had expected that perceptual assimilation patterns would provide valid predictions for cross-language identification accuracy because current speech learning models [2, 7] assume that nonnative speech perception is largely guided by how the sounds of the L2 are perceptually mapped on to the L1.

This expectation was only met, to some extent, for the identification of English approximants, sibilants, and the glottal fricative by L1 Kalaallisut listeners. The predictions for the identification of plosives and (labio-)dental fricatives were not borne out. We are currently unable to account for these unexpected results. The most puzzling result was that the aspirated [p^h, k^h], but not [t^h], were more accurately identified than their unaspirated counterparts, even though Kalaallisut has only unaspirated stops. Familiarity with the L2 Danish of our participants cannot help to explain this result, but it may be the reason for the surprisingly accurate identification of [f, v] since Danish has the very similar [f - ʊ] contrast. Several inaccurate predictions were derived from perceptual assimilations which differed in category goodness, and we suggest that the ability of nonnatives to detect a difference in goodness of fit to an L1 category is a necessary, but not sufficient prerequisite for accurate identification. Finally, the low identification accuracy for the dental fricatives could be due to confusable response labels for [θ, ð], i.e., *thigh* and *thy*.

We hope that future studies will be able to further explore the relation between the perceptual assimilation of nonnative sounds to native categories on the one hand and identification accuracy of these nonnative sounds on the other.

6. REFERENCES

[1] Best, C. T. 1995. A direct realist view of cross-language speech perception. In: Strange, W. (ed.) *Speech Perception and Linguistic Experience*. York Press, 171-206.

[2] Best, C. T., Tyler, M. D. 2007. Nonnative and second-language speech. In Bohn, O.-S., Munro, M. J. (eds.) *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, 17, 13.

[3] Boersma, P., Weenink, D. 2016. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.14. Online: http://www. praat. org/retrieved.

[4] Bohn, O.-S. 2002. On phonetic similarity. In Burmeister, P. Piske, T., Rohde, A. (eds.), *An Integrated View of Language Development: Papers in Honor of Henning Wode.* Wissenschaftlicher Verlag Trier, 191-216

[5] Bohn, O.-S, Best, C. T., Avesani, C., Vayra, M. 2011. Perceiving through the lens of native phonetics: Italian and Danish listener's perception of English consonant contrasts. *Proc. 17th ICPhS* Hong Kong, 336-339.

[6] Bundgaard-Nielsen, R. L., Best, C. T., Tyler, M. D. 2011. Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Appl. Psycholing.* 32, 51-67.

[7] Flege, J. E., Bohn, O.-S. 2021. The revised Speech Learning Model (SLM-r). In: Wayland, R. (ed.) *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge, UK: Cambridge University Press, 3-83.

[8] Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J. C. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am*. 107, 2711-2724.

[9] Hillenbrand, J., Gayvert, R. T. 2005. Open source software for experiment design and control. *J. Speech, Language, and Hearing Res.* 48, 45-60.

[10] Horslund, C. S., Bohn, O. S. (2022). Assimilation patterns predict L2 identification accuracy of English initial consonants. *J. Second Language Pronunciation*, 8, 218-247.

[11] Shannon, Robert V., Jensvold, Angela, Padilla, Monica, Robert, Mark E., Wang, Xiaosong. 1999. Consonant recodings for speech testing. *J. Acoust. Soc. Am*. 106, L71-L74.

ORAL RECAST AND TASK-BASED INTERACTION ON THE PROSODY OF L2 CHINESE LEARNERS

Chun-Mei Chen

National Chung Hsing University
chench@dragon.nchu.edu.tw

ABSTRACT

This study investigates the effect of oral recast on the acquisition of Mandarin prosody of L2 learners from diverse L1 linguistic backgrounds in Task-Based Language Teaching (TBLT) classroom-based context. The participants were thirty-six L2 learners of Chinese, and tokens of prosodic tasks were collected from the pre-test and post-test, and task-based interaction sessions in beginner L2 Mandarin classrooms. The results indicated that the accuracy rate of tones and *Third Tone Sandhi* production with oral recast corrective feedback was significantly higher than those without oral recast provision in the same condition. The findings verified the effectiveness of oral recast in tonal pedagogy and shed light on the tonal acquisition of L2 Chinese learners. Additionally, tonal production with oral recast feedback treatment facilitated the application of *Third Tone Sandhi* rules and prosodic fluency in L2 Chinese speech.

Keywords: oral recast, TBLT, L2 learners, prosody, Mandarin Chinese

1. INTRODUCTION

This study investigates how oral recast, one of the most frequently occurred corrective feedback (CF) in beginner classrooms of Chinese as a second language [1][2] affects the production of prosody of the second language (L2) learners in Task-Based Language Teaching (TBLT) classroom-based context. Teacher-student interactions have been believed to be beneficial to language development [3][4][5]. Recent studies carried out further investigation on the levels of L2 proficiency and paired task performance [6][7][8]. The phonetic pedagogy on prosodic tasks for learners to notice and correct errors still needs to be explored in L2 Chinese classrooms. Additionally, corrective feedback aimed at pronunciation and tonal errors in second language classrooms has been relatively understudied. The present study fills in the gap by investigating the effect of interactional corrective feedback, specifically oral recast, in face-to-face (FTF) and synchronous computer-mediated communication (SCMC) task-based interaction on the prosodic gains of L2 Chinese learners. The degree to which the impact of oral correct feedback on the prosodic tasks in beginner L2 Chinese classrooms was investigated.

An oral CF move in L2 instruction ranges from a reformulation, an indication of an error, prompting a target phrase or structure, to a metalinguistic explanation. Oral implicit CF includes recasts, clarification requests, elicitation, and repetition [9][10], whereas oral explicit CF includes metalinguistic feedback and other explicit corrections. While the effects of recasts have been confirmed [11], prompts have been found to be more effective than recasts [12], and form-focused instruction, combined with prompts, contributes to L2 grammar development [13]. Specifically, empirical studies suggest that metalinguistic feedback outperformed recasts [14][15] in grammar instruction, albeit such issues as processing prosodic and phonetic knowledge in task-based interaction are relatively unexplored. This study investigates oral CF effects and contextual factors, exploring the role of oral CF provision in FTF or SCMC task-based interaction with potential benefits in facilitating L2 Chinese prosodic acquisition. The study positions oral recast treatments as mediated by noticing and meaning-making processes and embedded in phonetic pedagogy with relevance to L2 acquisition. Through an empirical examination of the integration of oral CF in task-based L2 Chinese phonetic instruction, this study adds to the discussion on using linguistic resources in the classroom context and contributes to understanding L2 Chinese prosodic gains in the TBLT framework.

2. CORRECTIVE FEEDBACK AND L2 PHONETIC ACQUISITION

Oral CF in this study refers to verbal responses to a learner's errors [11][14][16]. Recasts, defined as the reformulation of a learner's utterance, have been argued to be the most effective type of CF in facilitating learning [11]. Additional or alternative explicit CF practices include a metalinguistic explanation with the provision of rules and examples. Studies have shown that recasts are the most frequently occurring oral CF type in language classrooms [15][17][18]. Although teaching pronunciation in beginner-level L2 Chinese

classrooms has become essential in teaching practices, the majority of studies on Chinese phonology and phonetics were theoretically-based or conducted in laboratory contexts. It should be noted that such issues as teaching L2 learners from diverse linguistic backgrounds [19], and raising L2 learners' attention in classroom tasks [20][21][22] can be crucial in promoting phonetic instruction and L2 phonetic acquisition. The practices and application of oral CF in L2 Chinese phonetic instruction and their effects on prosodic production, specially, in the FTF or SCMC modality, inform practitioners of learners' behaviour and shed light on the tonal acquisition of L2 Chinese learners.

In addition, L2 Chinese learners must use prosodic elements such as tonal changes to communicate effectively. One of the most commonly taught tonal change patterns is the *Third Tone Sandhi* rule. The Tone 3 (T3) Sandhi in Mandarin Chinese is a phonological pattern in which a T3 syllable becomes a Tone 2 (T2) or a T2-like syllable when it is followed by another T3 syllable in connected speech [23]. In phonetic drills, when two T3 syllables occur in succession, the first syllable is instructed to change to T2, while the second syllable keeps the T3. In strings of more than two T3 syllables, the tone changes depending on syntactic bracketing structure or the context in teaching practices. Accordingly, the "context" needs more elaboration and concrete illustration in task-based phonetic instruction. To read the seven-syllable T3 string, as in '*lǐ xiǎojiě xiǎngmaǐ shuǐjiǎo*' ('Ms. Li would like to buy dumplings'), form-focused instruction and prosodic grouping could be crucial in developing L2 learners' L2 phonetic acquisition and avoiding breakdowns.

Focusing on the form within communicative tasks is a central component of task-based frameworks [24][25]. As classroom tasks "provide opportunities for noticing the gap between a participant's production and input provided and for metalinguistic reflection in the form of output" [26], meaning messages between the instructor and the L2 learner are conveyed. Although attention or awareness may not be a significant concern in TBLT, learning can be predicated on learners' attending to the target structures when engaged in meaningful task-based interaction [25]. TBLT research also demonstrates the integration of traditional acoustic analysis of L2 pronunciation and L2 pronunciation accuracy within the TBLT context.

This study seeks to provide an empirical understanding of how phonetic and phonological knowledge can be transformed into successfully uptake in the context of second language learning and L2 Chinese tonal instruction. Given that implicit knowledge is considered to represent a learner's true competence and knowledge without awareness [27], explicit knowledge is often described as knowledge with awareness [28]. Concrete form-focused instruction activities have been proposed in L2 learners' interlanguage development [16]. In terms of phonological deviant tokens, implicit feedback has been shown to be salient to learners [29]. Although learning contexts in L2 Chinese classrooms are different from laboratory settings with controlled conditions, it is evident that tasks and learning contexts influence the production accuracy and processing of tonal knowledge.

This study sets out to investigate the effects of oral recast and TBLT interaction on the prosodic gains of beginner L2 Chinese learners' production. The research questions are formulated as follows:

RQ1 What are the effects of oral recast on learners' production of Mandarin tones and Third Tone Sandhi?
RQ2 What are the effects of task-based interaction on beginner L2 Chinese learners' prosody?

3. METHOD

3.1. Instructional setting and participants

The study was conducted in an intensive L2 Chinese program. The classes were delivered for five hours per week. Three teachers conducted and their 36 students participated in the investigation, and they were divided into three groups, Class A and Class B constituted the experimental group with oral recast, in contrast to Class C in the control group without oral recast in phonetic instruction. In the experimental group, Class A utilized the oral recast and task-based interaction in the phonetic instruction in the FTF modality, whereas the instruction intervention of Class B was conducted in the SCMC modality. The participating teachers were highly experienced, with a mean experience of 10.8 years. The class size was the same, with 12 students per class, and the students were beginner L2 learners of Chinese from linguistically diverse backgrounds. None were heritage learners, and they began learning Mandarin in adulthood. The mean age for the L2 Chinese learners was 26.5 years (ranged 18-36), and they reported no history of hearing impairments or speech disorders at the time of the investigation.

3.2. Materials

A pre-test and equivalent post-test was designed to assess the prosodic gains of the participants in both groups from three classes. The participants were asked to complete a total of five task-based activities in Chinese consisting of sentence completion and information exchange. Strings in consecutive T3

syllables were presented in pinyin with diacritics marking tones without grouping or syntactic bracketing. There were 30 items on the pre-test and post-test representing potential application of the T3 Sandhi rules (in T2 or a T2-like syllable) and the other 120 tones. The participants were instructed in Week 1 that if two syllables in succession use the third tone, the first syllable changes to the second tone, and the second syllable keeps the third tone, such as *shuǐ jiǎo* becomes *shuí hǎo* 'dumpling' and *shuǐ guǒ* becomes *shuí guǒ* 'fruit.' The production task was conducted individually, and picture-based interaction was administered to assess L2 learners' explicit and declarative knowledge, in addition to the prosodic gains, along with elicitation inquiry that guided the correct application to the T3 Sandhi rule in the task-based interaction. Participants were required to describe the assigned pictures in Chinese, using the provided items to produce oral tokens with tone changes. The teacher in each class initiated the questions, assigned the task in the activities, and gave instructions to elicit the production of the participants. The production task took participants in the groups approximately 15-20 minutes to complete. All the tests were recorded and conducted in the classrooms as part of the course assessments. In addition to the oral recast treatment sessions, classroom discourse and voice tokens collected from task-based interaction recordings were compiled from Week 4 to Week 12, along with an optional follow-up task recording in Week 24. The elicitation materials and the familiarity or unfamiliarity with the production tasks were controlled.

3.3. Procedure

All participants in the experimental group, in FTF and SCMC modality, received oral recast when making errors on tone changes in phonetic instruction treatments. The oral recast treatment sessions in the experimental group lasted 30 minutes every week, in both FTF and SCMC modes, from Week 4 to Week 8. The control group did not receive any oral recast feedback in the phonetic instruction. All the learner participants were instructed in task-based interaction to know the difference among a full T3, a half T3, and T2. In the present study, all the participating teachers provided tonal instruction by asking their students to change the marked parts to T2 in the strings longer than three syllables, such as [*Lǐ **xiǎo**jiě*] *mǎi* [***shuǐ**jiǎo*] 'Ms. Li buys dumplings' and [*Lǐ **xiǎo**jiě*] ***xiǎng**mǎi **shuǐ**guǒ*] 'Ms. Li would like to buy the fruit' with bracketing and grouping. L2 Chinese learner participants spent 30-40 minutes interacting with the instructor when practicing task-based activities. The procedure of the pre-test, treatments, and the post-test is illustrated in Fig. 1.

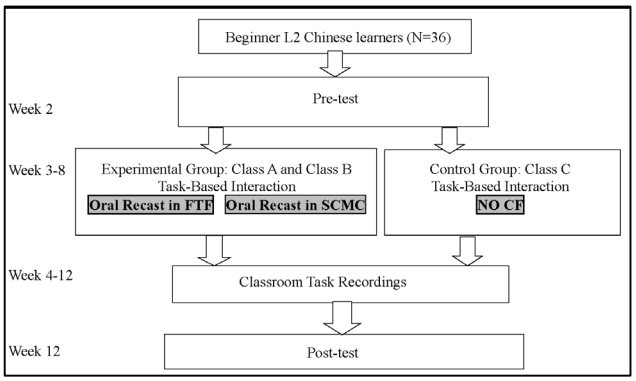


Figure 1: An overview of the procedure

As for the production scoring, the present study adopted a combination of an acoustic analysis and human rating method. The pre-test and post-test were scored for accuracy and fluency. Changes in crucial acoustic properties will either immediately or eventually enhance the overall intelligibility of L2 speech production. Five native speakers of Mandarin instructors who did not teach the participants were recruited to rate the accuracy of L2 learners' production tasks in speech strings. Acoustic analyses are conducted on the dataset using Praat with a focus on the significant acoustic variables, correlates of prosodic accuracy and fluency when inconsistency of the rating emerged. The recruited raters were presented with production tokens in a randomized order and asked to rate them. The inter-class correlation was calculated for tonal accuracy ($r = .85$) and fluency ($r = .81$). The values were statistically significant at $p < .001$, indicating adequate reliability among the human rating.

4. RESULTS

The data were checked for normality and revealed two potential outliers in the post-test scores of the experimental and control groups. Outliers have been retained for the analysis as no significant changes were found in testing performed with and without outliers. A mixed ANOVA tested for a significant interaction effect ($\alpha = .05$) between test time and groups with and without oral recast feedback group. Significance testing was calculated. The first research question investigated the effects of oral recast on learners' production of Mandarin tones and T3 Sandhi. Results indicated that the experimental and control groups with task-based interaction significantly improved their tone production scores from Week 2 to week 12. All scores in the post-test were higher than those in the pre-test, indicating gains

in prosodic accuracy and fluency of the testing. The main effect for time was statistically significant (F=115.96, df =1, p = .00), which addressed the second research question by confirming the effect of task-based interaction on the prosodic production of L2 Chinese learners. The comparison of prosodic tasks of L2 Chinese learners in the post-test is shown in Fig. 2.

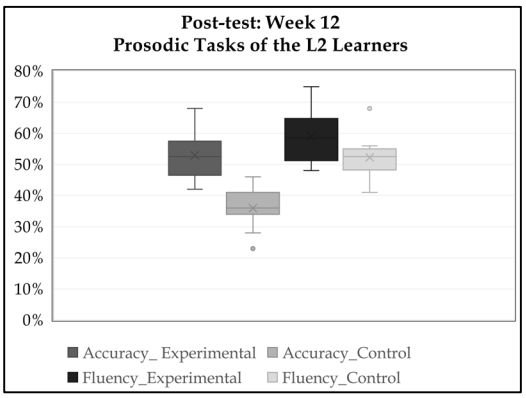


Figure 2: Comparison of prosodic tasks of L2 Chinese learners in the post-test

In terms of the experimental group in Class A (in FTF mode) and Class B (in SCMC), the FTF group showed a mean gain of 33.2% from the pre-test to post-test, while the SCMC group saw a mean gain of 28.6%. The interaction between time and modalities was not statistically significant (F = .983, df =1, p =.356). Therefore, the model was simplified to a 2 (Group, Experimental and Control Group) × 2 (Time) mixed design ANOVA, which showed the significant Time by Group interaction (F (1, 34) = 3.85, p = .027, η^2_p = .08), as well as a significant main effect of Time (F (1, 34) = 7.28, p = .0125, η^2_p = .09). The main effect of Group on the prosodic accuracy was significant (F (1, 34) = 5.33, p = .023, η^2_p = .07), indicating L2 Chinese tone and T3 Sandhi production with oral recast feedback was significantly higher than those without oral recast provision in the same task-based interaction condition. However, the main effect of Group on the fluency was not significant (F (1, 34) = 2.78, p = .278, η^2_p = .02). The accuracy of L2 production of T3 Sandhi in classroom task recordings is illustrated in Fig. 3. As can be seen, both experimental and control groups improved their accuracy of T3 Sandhi production in task-based interaction classrooms, though the gains in the experimental group were more salient and retained in Week 24. The result suggests that the retrieval access to target form in oral recast treatment can contribute to the long-term prosodic gains. It is worth noting that the study focused on the oral recast feedback in phonetic instruction, and individual learner factors warrant a long-term follow-up to assess the gains.

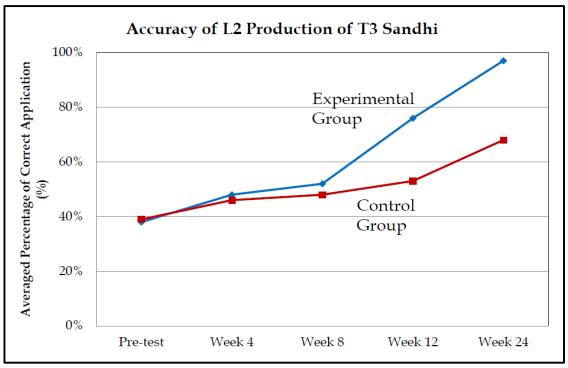


Figure 3: Accuracy of T3 Sandhi in the groups

5. DISCUSSION AND CONCLUSION

This study aimed to investigate the effect of oral recast and task-based interaction on the prosodic production of beginner L2 Chinese learners. By examining the impact of oral recasts on beginner L2 Chinese learners' prosodic development, whether recasts provided in FTF and SCMC classrooms benefited prosodic development can be further examined in different modalities. It is evident that with task-based interaction, beginner L2 learners can apply the T3 Sandhi rule in their speech. The accuracy rate of tones and T3 Sandhi production with oral recast corrective feedback was significantly higher than those without oral recast provision in the same condition. Additionally, tonal production with oral recast feedback treatment facilitated the application of T3 Sandhi rule and prosodic fluency in L2 Chinese speech.

Gains in the prosodic production throughout the investigation also confirmed the impact of oral recast feedback on beginners' tonal production. Oral feedback with task interaction assets of tonal contrasts in communicative mode improved the production of a tonal system. Findings suggest that oral recast feedback provision in FTF and SCMC classrooms facilitates L2 Chinese prosodic acquisition. Pronunciation tasks in task-based dialogue could be effective in improving learners' speech fluency, and focus-on-forms tasks improved L2 learners' communicative strategies.

Prosodic issues in L2 learners' connected speech are seldom explored in earlier studies. The study confirmed that FTF and SCMC oral recast feedback enhanced learners' tonal production, prosodic familiarity, and appropriate application of tonal changes in their pronunciation tasks, regardless of their L1 linguistic backgrounds. Task-based interaction could be relevant to the prosodic achievements of L2 learners in beginner classrooms. The findings in this study verified the effectiveness of oral recast in tonal pedagogy and shed light on the tonal acquisition of L2 Chinese learners.

6. REFERENCES

[1] Lyster, R., Ranta, L. 1997. Corrective feedback and learner uptake. *Studies in Second Language Acquisition* 19, 37–66.

[2] Chen, C.-M. 2020. Teachers' practices of corrective feedback in beginner L2 Chinese classrooms. *Journal of Chinese Language Teaching* 17(4), 1–40.

[3] Philp, J., Adams, R., Iwashita, N. 2014. *Peer Interaction and Second Language Learning*. Routledge.

[4] Sato, M., Ballinger, S. (Eds.). 2016. *Peer interaction and second language learning: Pedagogical potential and research agenda*. Amsterdam, the Netherlands: Benjamins.

[5] Zabihi, R., Ghahramanzadeh, B. 2022. Proficiency pairing and engagement with language in L2 peer interaction. *System* 105, 102725.

[6] Basterrechea, M., Gallardo-del-Puerto, F. 2020. Language-related episodes and pair dynamics in primary school CLIL learners: A comparison between proficiency-matched and student-selected pairs. *Studies in Second Language Learning and Teaching* 10(3), 423–447.

[7] Niu, R., Jiang, L., & Deng, Y. 2018. Effect of proficiency pairing on L2 learners' language learning and scaffolding in collaborative writing. *The Asia-Pacific Education Researcher* 27(3), 187–195.

[8] Storch, N., Aldosari, A. 2013. Pairing learners in pair work activity. *Language Teaching Research* 17(1), 31–48.

[9] Carroll, S., Swain, M. 1993. Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition* 15, 357–386.

[10] Li, S. 2010. The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning* 60, 309–365.

[11] Long, M. H. 2007. Recasts in SLA: The story so far. In: Long, M. H. (ed.), *Problems in SLA*. Mahwah, NJ: Erlbaum, 75–116

[12] Ammar, A., Spada, N. 2006. One size fits all? Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition* 28, 543–574.

[13] Lyster, R. 2004. Different effects of prompts and effects in form-focused instruction. *Studies in Second Language Acquisition* 26, 399–432.

[14] Ellis, R., Yuan, Y. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26, 59–84.

[15] Sheen, Y. 2004. Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research* 8, 263–300.

[16] Lyster, R. 2007. *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam, the Netherlands: Benjamins.

[17] Ellis, R., Basturkmen, H., Loewen, S. 2001. Learner uptake in communicative ESL lessons. *Language Learning* 51(2), 281–318.

[18] Panova, I., Lyster, R. 2002. Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly* 36(4), 573–595.

[19] Derwing, T. M., Thomson, R. I., Foote, J. A., Munro, M. J. 2012. A longitudinal study of listening perception in adult learners of English: Implications for teachers. *Canadian Modern Language Review* 68(3), 247–66.

[20] Saito 2012, Saito, K. 2012. Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly* 46, 842–854.

[21] Saito, K. 2013. Re-examining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition* 35, 1–29.

[22] Saito and Lyster 2012 Saito, K., Lyster, R. 2012. Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning* 62, 595–633.

[23] Chen, M. Y. 2000. *Tone sandhi: Patterns across Chinese dialects.* Cambridge University Press.

[24] Robinson, P. (Ed.). 2011. *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. Amsterdam, the Netherlands: Benjamins.

[25] Robinson and 2007 Robinson, P., Gilabert, R. 2007. Task complexity, the Cognition Hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching* 45, 161–176.

[26] Long, M. H. 2015. *Second language acquisition and task-based language teaching*. Malden: Wiley Blackwell.

[27] Spada, N. 2015. SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching* 48(1), 69–81.

[28] Ellis, R. 2003. *Task-based language learning and teaching*. Oxford: Oxford University Press.

[29] Mackey, A., Gass, S., McDonough, K. 2000. How do learners perceive interactional feedback? *Studies in Second Language Acquisition* 22, 471–497.

FILLER PARTICLES IN ENGLISH AND SPANISH L1 AND L2 SPEECH

Beeke Muhlack

Language Science and Technology, Saarland University
muhlack@lst.uni-saarland.de

ABSTRACT

The filler particle (FP) distribution and vowel quality of 10 female English native speakers and 10 female native Spanish speakers were investigated both in their native language (L1) and their second language (L2). As expected, speakers produce more FPs in their L2 speech and speaker groups prefer different types of FPs. Native English speakers produce more vocalic-nasal FPs (*um*) than Spanish speakers while the latter produce more vocalic FPs (*uh*). The vowel quality in the FPs differs between the speakers' L1 (open central vowel for English and close-mid front vowel for Spanish) but vowel areas merge in the L2, which can be explained with the speakers' attempt to produce native-like FPs of the target language.

Keywords: filler particles, disfluencies, second language learning, English, Spanish

1. INTRODUCTION

The aim of this paper is to compare the production of filler particles (FPs), i.e., the frequency distribution and vowel quality, in English and Spanish as a native (L1) and second language (L2). Work on FPs and their phonetic characteristics has increased over the last decades. However, languages other than English (e.g., Spanish [1]) are still under-researched in this area. Relevant studies for Spanish FPs show the following: The preferred type of FP in Spanish seems to be the vocalic FP with approx. 76% of all FPs being realised as *uh*, 13% as nasal FP *hm* and only 11% with a vocalic-nasal FP *um* [2]. The investigation of FP usage and quality of 24 L1 Spanish speakers living in the US shows that the vowel quality of FPs in their L1 changes from [e] to a lower, more central variant ([ə] or [a]) with increasing use of English (proficiency and length of stay) [1].

Similarly, [2] showed that Spanish monolinguals use an FP vowel with a lower F1 and a higher F2 than Spanish-Afrikaans bilinguals. It seems that Spanish speakers produce a fronted and more close vowel in their FPs also compared to other languages [3]. In comparison with vowels in lexical material, [4] show that the FP vowel is closest to the realisations of /e/ but claim it is produced with a higher F1 and a lower F2 than the lexical vowels. Unfortunately, in [3, 4] the degree of overlap between the realisations of /e/ and the vowel in FPs was not reported.

From the numerous studies on FPs in English, for instance [5] reports the vocalic-nasal FP *um* as the preferred type for British English speakers (81% *um*, 18% *uh*). The vowel quality of English FPs is often reported to vary between the central vowels [ə] and [ʌ][1] and the back vowel [ɑ] [6]. The vowels in the FPs usually spread over a large part of the central vowel space [7]. Results in [8] suggest that the vowel in FPs in English is closest (in terms of smallest Euclidean distance) to the open-mid central vowel [ʌ], although the English speaker group is quite heterogeneous in terms of their English accents (which may introduce further variation).

Work on L2 learning has shown that the use of disfluencies, and thus also FPs, is typically higher in an L2 [9, 10, 11, 12] but decreases for pauses with rising L2 proficiency [13]. An L2 effect for pauses but not for FPs was also found by [14] for advanced English learners of German. The preference of the vocalic FP in L2 English (intermediate and advanced learners) by native Spanish speakers is reported by [15]. This seems to suggest that learners transfer their native FPs to their L2. The results in [16] show that advanced learners of an L2 (English and German) are able to adapt the vowel quality of the target language in their FPs but intermediate learners are not. The latter group seems to transfer their L1 FPs to the L2.

Furthermore, [17] found that German-French bilinguals produce distinct vowel qualities in the two languages but that the weaker language would show a shift towards the vowel quality of the dominant language. For Afrikaans-Spanish bilinguals from Patagonia, Argentina, [2] found that the speakers do not share the vowel quality of Afrikaans and Spanish monolinguals, but they still produce separate vowel qualities for their FPs in both languages.

Based on the reviewed literature, it is assumed that English native speakers prefer the vocalic-nasal FP *um* while Spanish native speakers prefer the vocalic FP *uh*. It is expected that language

learners also employ their preferred FPs in their L2 [15]. Distinct vowel qualities for each language are expected to be observed and, furthermore, a shift towards the L2 vowel quality is expected in the second language context.

2. MATERIAL

For this study, a subset of the Diapix-FL corpus [18] is used, which consists of 20 female speakers (10 native English, 10 native Spanish/Basque) completing a dialogue task in both their L1 and L2 (Spanish and English, respectively). Native speakers of English were recorded at the University of Edinburgh, native Spanish/Basque speakers were recorded at the University of the Basque Country in Vitoria. The speakers are quite proficient in their L2 (B2-C1 on CEFR[2]) as both groups study their L2 in their second year at university. The native Spanish participants all passed a B2 proficiency test for English as part of their study programme to reach their second year, native English speakers self-reported their proficiency level for Spanish as their L2. Four male speakers (2 English, 2 Spanish) from the corpus were excluded in order to keep the speaker group as homogeneous as possible.

The speakers were grouped in same-L1 pairs and given a spot-the-difference task in both L1 and L2. Each speaker was given a picture, different versions of a scene, for which they had to cooperate to find 12 differences without being able to see each other's pictures. Dialogue partners solved three spot-the-difference tasks per language. Each speaker was recorded on a separate channel. For more details on the recording method, see [19, 20].[3]

Orthographic annotations are included in the corpus along with annotations of silent pauses (including breath noises), elongations, and FPs. All FPs are marked with one symbol, so for the re-annotation distinct labels for vocalic (*uh*), nasal (*hm*) and vocalic-nasal (*um*) FPs were used. Additionally, ten tokens (annotated in stressed positions) of each corner vowel (/i u a/ for Spanish and /i u ɑ/ for English) for all L1 speech were annotated along with ten tokens of one additional vowel that is frequently reported to occur in the FPs of the respective language (/e/ for Spanish, /ʌ/ for English).

3. RESULTS

A total of 2,737 FPs was found in the subset of the corpus: 245 nasal FPs (*hm*), 1,118 vocalic FPs (*uh*) and 1,374 vocalic-nasal FPs (*um*).

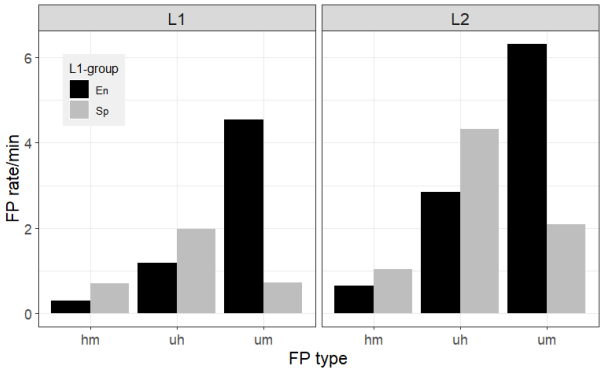


Figure 1: FP rate per minute in L1 and L2 speech by both speaker groups. Native English speakers in black, native Spanish speakers in grey.

3.1. Frequency distribution

FPs are more frequent in each L2 than in the respective L1 speech. L1 English shows an overall FP rate of 1.9 per minute while L1 Spanish shows a lower rate with 1.0 FP per minute. Native English speakers increase their FP rate only slightly in their L2 (2.2 FPs/min for L2 Spanish), whereas native Spanish speakers show a rate of 2.5 FPs per minute in their L2, which is considerably higher compared to their low L1 rate. The frequency distribution of the four FP types in both speaker groups (English and Spanish native speakers) and both conditions (L1, L2) is shown in Figure 1. In their respective native languages (left panel), L1 Spanish speakers use more vocalic FPs than the other FP types, while L1 English speakers use the vocalic-nasal FP type *um* more frequently (4.8 vs. 0.7 *um*/min). The nasal FP *hm* is used rarely but more often by Spanish speakers. The same pattern is also visible in L2 speech (right panel), with numbers increasing overall.

During the classification of the FPs, an observation was made: speakers, when producing a vocalic-nasal FP, do not always use the bilabial nasal /m/ but also an alveolar /n/ in some instances. An auditory inspection by one annotator (the author) revealed that 3.4% (i.e., 38/1,108) of all vocalic-nasal FPs of the English native speakers (in both L1 and L2) were produced with a nasal that was not bilabial and 18.1% (i.e., 48/266) of those by the Spanish native speakers were produced with a nasal that was not bilabial. The majority of non-bilabial nasals were alveolar, while a few were also categorised as labiodental, velar or a sequence of alveolar and bilabial nasals. To shed more light on the phenomenon, the annotation of several annotators should be taken into account.

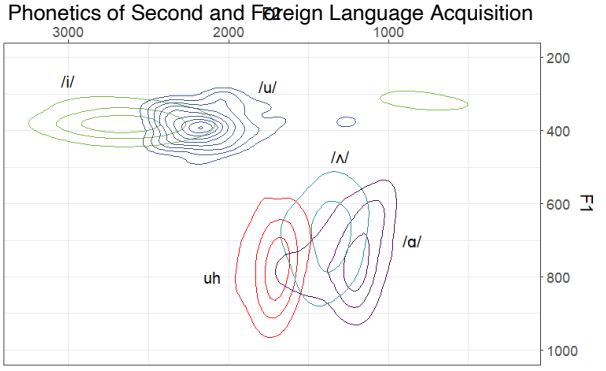


Figure 2: Formant values of realisations of *uh* (red) in English L1 compared to those of the lexical vowels /ɑ i u ʌ/.

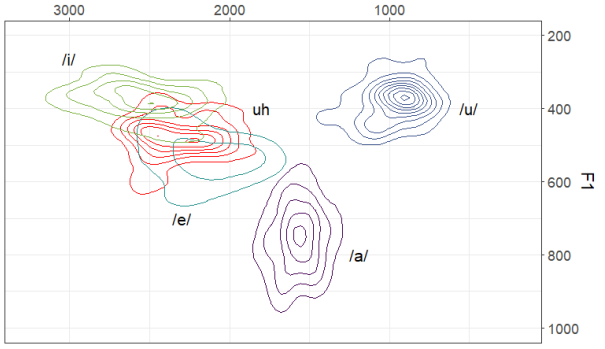


Figure 3: Formant values of realisations of *uh* (red) in Spanish L1 compared to those of the lexical vowels /a i e u/.

3.2. Vowel quality

The first and second formant of all purely vocalic FPs and the corner vowels of L1 English and L1 Spanish were measured in Praat [21] at the midpoint of the vowel using the Burg method.[4] To visualise the vowel quality produced in the FP *uh*, two-dimensional (F1, F2) kernel density distributions of the lexical vowels and the FP vowel were plotted in R [22] using ggplot2 [23].

Figure 2 shows that the vowel in the native English FPs partly overlaps with the realisations of the lexical vowels (also L1 English) /ɑ/ and /ʌ/. Based on this observation, the overlap in the two-dimensional vowel spaces between these two lexical vowels and the FP vowel was calculated using the Pillai-score [24], that ranges from 1 (no overlap) to 0 (complete overlap). The vowel quality of the FP *uh* overlaps more with the open-mid vowel /ʌ/ (Pillai: 0.27) than the open vowel /ɑ/ (Pillai: 0.46). The /u/-fronting has been described before for the same data set [19].

Figure 3 shows that the vowel in the native Spanish FPs overlaps with the two front vowels /i/ and /e/ of native Spanish speech to a considerable degree. The Pillai scores show more overlap with the close-mid vowel /e/ (Pillai: 0.22) than the close vowel /i/ (Pillai: 0.34).

In order to answer the question whether speakers transfer the vowel quality of their L1 FP to their L2, Figure 4 compares the FP *uh* by both speaker groups in their L1 and L2 speech. The left panel shows that FPs in English and Spanish L1 are produced with distinct vowel qualities (Pillai: 0.67). The L2 FPs (right panel) show vowel qualities that spread over a larger area of the central vowel space for both speaker groups. However, they also show a bimodal distribution for both groups. Some FPs in L2 Spanish by the English speakers are produced with a lower F1 and a higher F2 while the majority of FPs is produced with the same vowel quality as

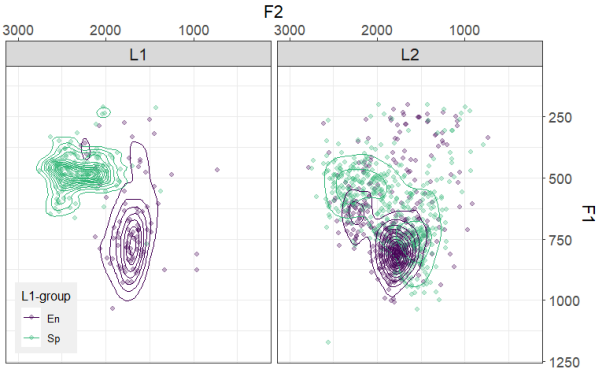


Figure 4: Vowel quality of *uh* in L1 and L2 speech by native English speakers (purple) and native Spanish speakers (green).

observed in their L1. Conversely, for L2 English by the Spanish speakers some of the FPs are produced with a native-like vowel quality (similar to the close-mid /e/) but a large portion is produced with a higher F1 and a lower F2, approximating the native English vowel quality.

4. DISCUSSION

The results of the current study confirm that L1 English prefers the vocalic-nasal FP (*um*) and L1 Spanish prefers the vocalic FP (*uh*) [2, 5]. Nasal FPs are not very frequent in either speaker group but for L1 Spanish *hm* is as infrequent as *um*. The frequency of FPs in L2 speech increases as reported before [9, 10, 11, 12], and moreover, the preferred type of FP remains constant [15] with native Spanish speakers preferring the vocalic FP *uh* and the native English speakers preferring the vocalic-nasal type *um* in both their L1 and L2. A reason for the preference of the specific FP type may be the syllable structures in both languages. In English, closed syllables seem to be more common than open syllables as observed from the data reported in [25], while in Spanish closed syllables do occur, but most often word-internally [26]. Another reason for the

preference of *uh* over *um* may be the restriction for Spanish word-final closed syllables which only allows the consonants /l, n, r, s, d/ in the coda[5] [26]. This restriction would also explain the occurrence of alveolar nasals in the vocalic-nasal FPs of the native Spanish speakers, yet not those produced by the native English speakers.

The results of the comparison of vowel qualities between lexical vowels and the FP *uh* in English show that the realisations of the open-mid central vowel /ʌ/ best represent the vowel of the FP, although vowel spaces do not entirely overlap. The FP vowel in English is produced more centrally than the vowel /ʌ/. The Spanish FP vowel, on the other hand, is produced as a close-mid front vowel, similar to the realisations of the Spanish lexical vowel /e/. In Spanish, [e] and [ɛ] are allophones of the phoneme /e/ [26], so the annotated corner vowels in Spanish include tokens of both allophones. The FP vowel overlaps with (the upper) half of the vowel space covered by the vowel phoneme /e/, suggesting that the vowel quality in *uh* is represented by [e] rather than [ɛ]. This is also supported by previous literature [3, 2]. However, the claim by [4] that the Spanish FP is produced with a lower F2 is not supported by the data presented here. Although a higher F1 can be observed, this may be due to the inclusion of the allophone [ɛ] in the set of corner vowels. Note that native Spanish speakers were bilinguals in Spanish and Basque, which may have influenced the vowel quality of the FPs. However, as the vowel systems of Spanish and Basque are the same [27, 26] this influence is assumed to be rather small.

The comparison between L1 and L2 speech shows that while the FP vowels in L1 English and Spanish are very distinct, this cannot be observed in the speakers' L2 speech. The distributions of FP vowels of Spanish and English L1 speakers, producing FPs in their L2 show two peaks, which suggests that some hesitations approximate the realisations of Spanish /e/ and some approximate the realisations of English /ʌ/. It seems that advanced learners are able to adapt the FP's vowel quality, but this is not the case for all the data. The results suggest that L1 Spanish speakers may be better at approximating the vowel quality of English FPs and that L1 English speakers are more likely to keep their native FP in their L2. It is likely that individual speakers adapt the vowel quality of the target language while others keep their native vowel quality. Factors that may influence the switch from L1 to L2 FP vowel quality may be L2 proficiency, the exposure to the target language by native speakers, or the speakers' stance towards the foreign language. Another possibility is that speakers do not switch from one vowel quality to the other entirely, but they employ a mixture of native-like and foreign FPs. Further investigations into speaker-specific patterns are planned for future work.

5. CONCLUSION

To conclude, this study has shown that filler particles (FPs) in English and Spanish native speech are quite distinct in (i) the type that is preferred (*um* vs. *uh*) and (ii) the vowel quality that is used approximating the realisations of the lexical vowels /ʌ/ in English and /e/ in Spanish. Furthermore, it is shown that learners of an L2 are able to produce FPs with a native-like vowel quality, even though a full adaptation was not observed. It is assumed that the degree of target language FP adaptation is speaker-specific, but this hypothesis is in need of further investigation.

The results discussed here offer insights into the realisations of FPs in two languages, suggesting that different languages employ specific FP paradigms. The tendency that native-like qualities of FPs (their vowel quality and preferred type) transfer to the speaker's L2 even for advanced learners may be highly relevant in fields where the speaker's background is unknown (forensic phonetic casework, language analysis for determination of origin). The results also support the view that disfluencies should be discussed in the L2 classroom to raise the learners' awareness of the foreign FP realisations. Whether the discussion of FPs in the L2 context helps the learners better approximate the L2 FPs is still an open question in L2 research.

[1] In phonetic transcription for English the symbol ʌ is commonly used to describe an open-mid central vowel [28] as opposed to a back vowel as described by the IPA.
[2] The Common European Framework of Reference for Languages (CEFR) [29] provides guidelines for the categorisation of L2-fluency and proficiency.
[3] The DIAPIX-FL corpus is freely available at http://datashare.is.ed.ac.uk/handle/10283/346
[4] Maximum formant: 5.5 kHz, max. number of formants: 5, window length: 0.025 s, dynamic range: 50 Hz
[5] Other consonants may occur in loanwords.

6. REFERENCES

[1] D. Erker and J. Bruso, "Uh, bueno, em...: Filled pauses as a site of contact-induced change in Boston Spanish," *Language Variation and Change*, vol. 29, no. 2, pp. 205–244, 2017.
[2] L. Garciá-Amaya and S. Lang, "Filled pauses are susceptible to cross-language phonetic influence,"

Studies in Second Language Acquisition, pp. 1–29, 2020.

[3] M. Candea, I. Vasilescu, and M. Adda-Decker, "Inter- and intra-language acoustic analysis of autonomous fillers," in *Workshop on Disfluency in Spontaneous Speech Workshop (DiSS 2008), Aix-en-Provence*, 2008, pp. 47–52.

[4] I. Vasilescu, R. Nemoto, and M. Adda-Decker, "Vocalic hesitations vs vocalic systems : a cross-language comparison," *Proceedings of the 16th International Congress of Phonetic Sciences*, no. August, pp. 1101–1104, 2007.

[5] E. de Leeuw, "Hesitation markers in English, German, and Dutch," *Journal of Germanic Linguistics*, vol. 19, no. 2, pp. 85–114, 2007.

[6] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, 1994. [Online]. Available: ftp://130.107.33.205/pub/papers/shriberg-thesis.pdf

[7] V. Hughes, S. Wood, and P. Foulkes, "Strength of forensic voice comparison evidence from the acoustics of filled pauses," *International Journal of Speech, Language and the Law*, vol. 23, no. 1, pp. 99–132, 2016.

[8] B. Muhlack, "The vowel quality of non-lexical hesitation particles in German and English L1 and L2 speech [Poster]," in *Phonetik und Phonologie im deutschsprachigen Raum (P&P16), Trier*, 2020.

[9] C. Brand and S. Götz, "Fluency versus accuracy in advanced spoken learner language," *International Journal of Corpus Linguistics*, vol. 16, no. 2, pp. 255–275, 2011.

[10] G. Gilquin, "Hesitation markers among EFL learners: Pragmatic deficiency or difference?" in *Pragmatics and Corpus Linguistics: A Mutualistic Entente*, J. Romero-Trillo, Ed. Berlin, Heidelberg, New York: Mouton de Gruyter, 2008, pp. 119–149.

[11] L. Temple, "Second language learner speech production," *Studia Linguistica*, vol. 54, no. 2, pp. 288–297, 2000.

[12] R. Wiese, "Language production in foreign and native languages: Same or different?" in *Second Language Productions*, H. W. Dechert, D. Möhle, and M. Raupach, Eds. Tübingen: Gunter Narr Verlag, 1984, pp. 11–25.

[13] A. Riazantseva, "Second lanuage proficiency and pausing: A study of Russian speakers of English," *Studies in Second Language Acquisition*, vol. 23, no. 4, pp. 497–526, 2001.

[14] M. Belz, S. Sauer, A. Lüdeling, and C. Mooshammer, "Fluently disfluent?: Pauses and repairs of advanced learners and native speakers of German," *International Journal of Learner Corpus Research*, vol. 3, no. 2, pp. 118–148, 2017.

[15] J. Cenoz, "Pauses and hesitation phenomena in second language production," *ITL - International Journal of Applied Linguistics*, vol. 127-128, no. January 2000, pp. 53–69, 2000.

[16] B. Muhlack, "L1 and L2 production of non-lexical hesitation particles of German and English native speakers," in *Workshop on Laughter and Other Non-Verbal Vocalisations, Bielefeld*, 2020, pp. 44–47.

[17] J. J. Lo, "Between äh(m) and euh(m): The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals," *Language and Speech*, vol. 63, no. 4, pp. 746–768, 2020.

[18] M. Cooke, M. L. Garcia Lecumberri, and M. Wester, *DiapixFL*. LISTA Consortium: (i) Language and Speech Lab, Universidad del Pais Vasco, Spain and Ikerbasque, Spain; (ii) CSTR, University of Edinburgh, UK; (iii) KTH Royal Institute of Technology, Sweden; (iv) Institute of Computer Science, FORT, 2013.

[19] M. L. Garcia Lecumberri, M. Cooke, and M. Wester, "A bi-directional task-based corpus of learners' conversational speech," *International Journal of Learner Corpus Research*, vol. 3, no. 2, pp. 175–195, 2017.

[20] M. Wester, M. L. García Lecumberri, and M. Cooke, "DIAPIX-FL: A symmetric corpus of problem-solving dialogues in first and second languages," in *Interspeech, Singapore*, 2014, pp. 509–513.

[21] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org

[22] R Core Team, "R: A Language and Environment for Statistical Computing," in *R Foundation for Statistical Computing*. Vienna, Austria: R version 4.1.3, 2022. [Online]. Available: https://www.r-project.org/

[23] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[24] M. C. Kelley and B. V. Tucker, "A comparison of four vowel overlap measures," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 137–145, 2020.

[25] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 101–112, 1990.

[26] C. Gabriel, "Phonetik und Phonologie des Spanischen," in *Linguistik im Sprachvergleich: Germanistik, Romanistik, Anglistik*, R. Klabunde, W. Mihatsch, and S. Dipper, Eds. Berlin: J.B. Metzler, 2022, ch. 2, pp. 27–48.

[27] J. I. Hualde and J. Ortiz de Urbina, *A Grammar of Basque*. Berlin, New York: Mouton de Gruyter, 2003.

[28] P. J. Roach, *English Phonetics and Phonology: A practical course*, 4th ed. Cambridge, New York, Melbourne: Cambridge University Press, 2009.

[29] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Strasbourg: Cambridge University Press, 2011. [Online]. Available: https://rm.coe.int/1680459f97

L2 acquisition of Polish sibilant fricatives by L1 Czech speakers

Phil J. Howson[1] & Irfana Madathodiyil[2]

[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft; [2]All India Institute of Speech and Hearing
howson@leibniz-zas.de; irfana@aiishmysore.in

ABSTRACT

Polish language has an uncommon three-way sibilant contrast /s, ʂ, ɕ/. For learners of Polish who have an L1 system that has two-way sibilant contrast (e.g., /s, ʃ/), this means that they must learn two segments (i.e., /ʂ, ɕ/) that are acoustically and perceptually similar to one L1 segment.

In this study, we examined 5 A-level and 5 C-level L2 leaners of Polish who were L1 Czech speakers. Participants produced sentences in both Czech and Polish containing their L1 and L2 sibilants, while we recorded ultrasound data. Tongue contours were traced and extracted. GAMMs were used to compare contours.

The results revealed that the acquisition process was long and extended at least past the A-level. But that typically, only /ʂ/ was acquired, while /ɕ/ remained articulatorily linked to /ʃ/. The results implicate goodness-of-fit as an important predictor for acquisition and that L2 segments are integrated into the L1 phonetic system.

Keywords: L2 acquisition, Polish, Sibilants, phonetics, Czech

1. INTRODUCTION

Second language acquisition requires that learners develop new articulatory routines and acquire the phonetic and phonological representation for novel segments. Two influential theories for second language acquisition are the Perceptual Assimilation Model L-2 [2] and the Speech Learning Model [3, 4].

2.1. Perceptual Assimilation Model-L2

The perceptual assimilation model-L2 [2], henceforth PAM-L2, is an extension of the perceptual assimilation model [1] to second language learners. In this model, novel L2 segments can either be perceptually assimilated to an L1 segment if they are similar in some phonetic dimension or they can remain uncategorized if they are sufficiently different from L1 segments. In the context of Czech learners of Polish, single category assimilation is expected for /ʂ, ɕ/ because they are perceptually similar to /ʃ/ for non-native listeners [7]. Potential differences in goodness-of-fit to the L1 category, or rather how perceptually similar /ʂ, ɕ/ are to /ʃ/, can vary between the two segments and may potentially inhibit or facilitate learning of one or both segments. In either case, the model predicts that if there are enough perceptual differences between the assimilated L2 segments and the L1 segments, learning can occur.

2.2. Speech Learning Model

The speech learning model [3, 4], henceforth the SLM, posits that L2 segments are mapped onto perceptually similar L1 segments during learning. The SLM additionally posits that the mechanisms involved in L1 acquisition are intact for L2 acquisition and that the learning process, similar to L1 acquisition, takes a significant amount of time. However, for learning to take place, disruption of L2-to-L1 perceptual links is necessary.

2.3. Hypothesis

While both the PAM-L2 and SLM were developed for second language acquisition contexts (SLA, i.e., contexts where learners are immersed in a second language), we believe it has applications in foreign language contexts as well (FLA, i.e., contexts where learners are not immersed and learn in a classroom). We hypothesize that the novel Polish sibilants, /ʂ, ɕ/, will initially develop as both perceptually linked to /ʃ/. However, due to larger differences in formant transitions and spectral cues, /ʂ/ will be easier to acquire as a new category, while /ɕ/ will be perceptually linked to /ʃ/ much longer, delaying, or even blocking acquisition. To test this hypothesis, we performed an ultrasound study of L1 Czech speakers' acquisition of L2 Polish. We recorded ultrasound data for L1 and L2 sibilant inventories.

2. METHODS

2.1. Participants

Ten L1 speakers of Czech learning Polish participated in this experiment (5 female, 5 male; 5 A-, 5 C-level). Participants were all college-educated, aged between 20-24 years, and had no self-reported history of speech or hearing disorders.

2.2 Stimuli & Procedure

Participants read sentences in Polish and Czech. Each language was blocked so that only Czech or only Polish words were produced, and the order of the blocks were counterbalanced. Target words were embedded in carrier phrases, for Polish "powiedz *target* proszę" and for Czech "řekněte *target* ještě

jednou." Each of the target segments were placed in three vocalic environments, /e, a, o/. Each target word was produced 8 times for a total of 1,200 token (8 repetitions x 5 segments x 3 environments x 5 learners x 2 groups). Tables 1 and 2 present the target words for Polish and Czech, respectively.

Segment	Target	English
/s̪/	de**s**ek	planks
	ta**s**ak	cleaver
	o**s**o	wasp
/ʂ/	rze**sz**e	multitudes
	pa**sz**a	feed
	o**sz**o	fallen
/ɕ/	le**si**e	forest
	na**si**ał	has clouded
	o**si**oł	donkey

Table 1. Polish stimuli with target segments in bold.

Segment	Target	English
/s/	ve**s**el	oars
	já**s**at	to rejoice
	po**s**olit	to salt
/ʃ/	ve**š**el	went in
	Dá**š**a	Dasha (name)
	po**š**oupat	to push/move

Table 2. Czech stimuli with target segments in bold.

Data were recorded in a quiet room in Prague using the AAA suite and a Micro ultrasound with a 20mm probe that had 92-degree field of view. The average frame rate was 90 frames per second.

2.2 Analysis

We annotated the data for each of the relevant segments based on the spectrogram. The onset was marked as the end of the formant structure of the preceding vowel and the onset of aperiodic noise, while the offset was marked as the end of the aperiodic noise and the onset of formant structure for the following vowel. DeepCutLab for Speech [13] was then used to trace tongue contours at the midpoint of each segment. DeepCutLab [5] is a machine learning algorithm for automatic feature extraction and has been trained on ultrasound data to detect edges for use with AAA. x- and y-coordinates were then extracted, normalized, and compared with generalized additive mixed models (GAMMs) [9] using the mgcv package [12] in R [6]. Posthoc comparisons were performed with tidymv [8]. We generated models for the Czech sibilant system and the Polish sibilant system for a total of 4 models (2 groups x 2 models). We included smoothing terms for the x-axis and x-axis by Segment. A factor smooth for x-axis and the interaction between Segment, Environment, and Participant was also included. Data were visualized using ggplot2 [13].

3. RESULTS

3.1 A-Level Learners

The statistical analysis for A-level learners of Polish revealed significant effects for the smoothing term for the interaction between the x-axis and /s̪/ [$F(4.5, 4.7) = 98$; $p < 0.01$], /ʂ/ [$F(1.6, 1.7) = 8$; $p = 0.01$], and /ɕ/ [$F(3.3, 3.5) = 30$; $p < 0.01$]. The R^2 for the model was 0.78. Figure 1 presents the GAMM contours for /s̪, ʂ, ɕ/. Figure 2 presents the estimated differences between /ʂ, ɕ/.

The GAMM contours revealed A-level learners distinguished /s̪/ from /ʂ, ɕ/, and that there was a slightly lower body for /ʂ/ than /ɕ/ (Figure 2). The results further indicated that /s̪/ had a more retracted tongue dorsum, compared to /ʂ, ɕ/ and a lower tongue body/blade. /ʂ, ɕ/ had a more advanced tongue root and a more raised tongue body.

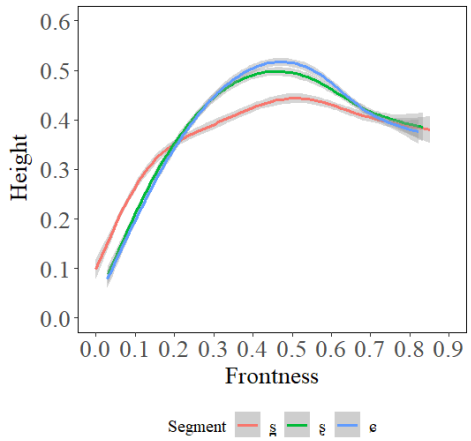


Figure 1: GAMM tongue contours for /s̪/ (red), /ʂ/ (green), /ɕ/ (blue) for A-level learners. Tongue tip is on the right.

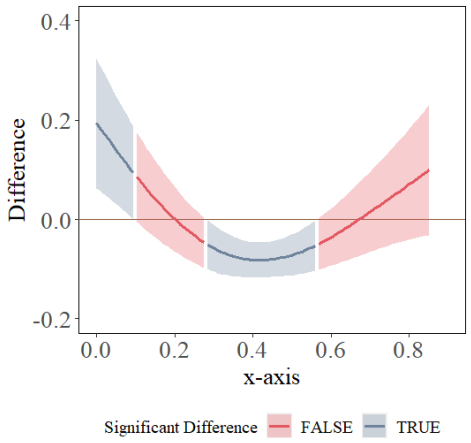


Figure 2: Estimated difference in contours between /ʂ/ and /ɕ/ for A-level learners. Areas with red indicate no significant differences between the two contours.

The comparison of Czech /ʃ/ and Polish /ʂ, ɕ/ revealed a significant effect for interaction between

the x-axis and /ʃ/ [F(1.8, 1.9) = 23; p < 0.01], /ɕ/ [F(2.3, 2.5) = 18; p < 0.01], and /ʂ/ [F(3.1, 3.4) = 40; p < 0.01]. The R^2 was 0.83. The observable difference was a slightly lowered tongue body for /ʂ, ɕ/, although the overall tongue contour was comparable. Figure 3 presents the comparison of /ʃ, ʂ, ɕ/.

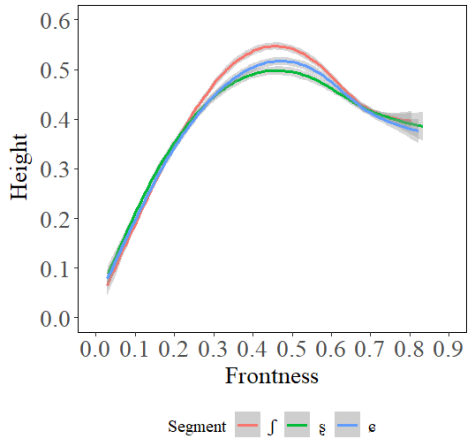


Figure 3: GAMM tongue contours for /ʃ/ (red), /ʂ/ (green), /ɕ/ (blue) for A-level learners of Polish. Tongue tip is on the right.

3.2 C-Level Learners

The analysis for C-level learners of Polish revealed significant effects for the smoothing terms for the interaction between the x-axis and /ʂ̻/ [F(5, 5) = 10; p < 0.01], /ʂ/ [F(5, 5) = 267; p < 0.01], and /ɕ/ [F(3.9, 4.3) = 191; p < 0.01]. The R^2 value for the model was 0.67. Figure 4 presents the GAMM contours for /ʂ̻, ʂ, ɕ/. Figure 5 presents the estimated differences between /ʂ, ɕ/.

The GAMM contours for C-level learners revealed a distinct three-way sibilant contrast. /ʂ̻/ had a retracted tongue dorsum compared to /ɕ/, and a relatively low and flat tongue body/blade compared to /ʂ, ɕ/. /ʂ/ had the most retracted tongue dorsum, with a tongue body/blade raised in between /s, ɕ/. Finally, /ɕ/ had the most advanced tongue dorsum and the most fronted and raised tongue body/blade.

The comparison of Polish /ʂ, ɕ/ and Czech /ʃ/ revealed a significant effect for the interaction between the x-axis and /ʃ/ [F(5, 5) = 193; p < 0.01], /ʂ/ [F(3.3, 3.8) = 201; p < 0.01], and /ɕ/ [F(3.6, 3.1) = 189; p < 0.01]. The R^2 was 0.70. The comparison of /ʃ, ʂ/ revealed differences in the tongue dorsum and the tongue body/blade, with /ʂ/ having a more retracted tongue dorsum and a lowered tongue body. The only area of significant difference between /ʃ, ɕ/ was in the most posterior regions of the tongue dorsum, indicating a slightly more retracted dorsum for /ʃ/.

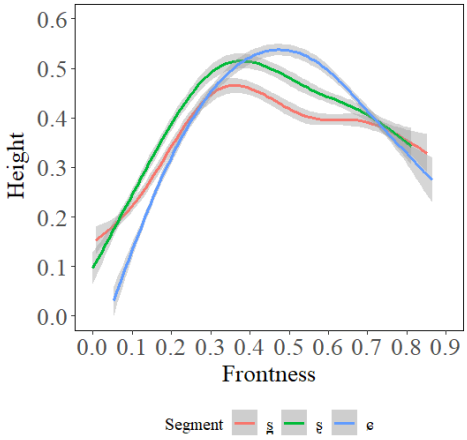


Figure 4: GAMM tongue contours for /ʂ̻/ (red), /ʂ/ (green), /ɕ/ (blue) for C-level learners of Polish. Tongue tip is on the right.

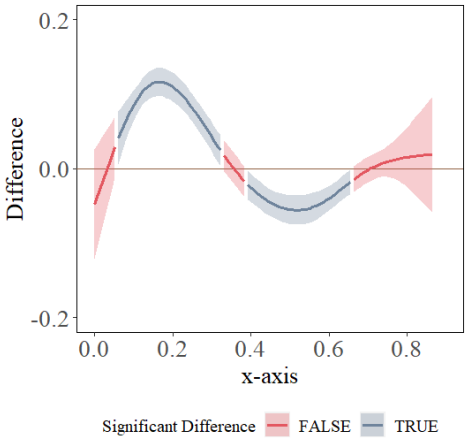


Figure 5: Estimated difference in contours between /ʂ/ and /ɕ/ for C-level learners. Areas with red indicate no significant differences between the two contours.

Figure 6 presents the comparison of Czech /ʃ/ and Polish /ʂ, ɕ/.

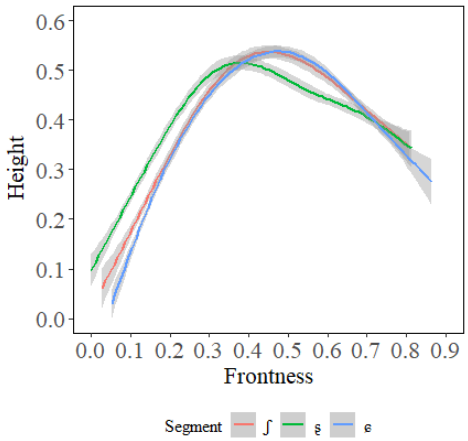


Figure 6: GAMM tongue contours for /ʃ/ (red), /ʂ/ (green), /ɕ/ (blue) for C-level learners of Polish. Tongue tip is on the right.

4. DISCUSSION

In this project, we hypothesized that /ʂ, ɕ/ would initially develop as linked to the /ʃ/ category, and

that with experience novel category formation for /ʂ, ɕ/ would occur. We found evidence for early learning, even at the A-level, with differences in tongue body height and tongue dorsum retraction for /ʂ, ɕ/ compared to /ʃ/. Additionally, we found small differences between /ʂ, ɕ/. C-level learners had successfully split the articulation of /ʂ, ɕ/, although we found that /ɕ/ had the same posture as /ʃ/, aside from a slightly more advanced tongue dorsum.

The data when taken together implicates that acquisition and learning for perceptually linked segments happens early [10], but that it is a slow process [4]. It takes at least past the A-level to acquire the three-way sibilant contrast in Polish. Additionally, /ɕ, ʃ/ merge later in the acquisition process, not at the onset. This indicates that even small acoustic-perceptual differences lead to differences in articulation, no matter how minor. The learning trajectory observed here suggests that perceptually similar segments are used to "piggyback" novel articulatory routines. L2 leaners make use of learned patterns and constriction locations and modify them during the acquisition process to eventually develop novel articulatory representations. However, in cases of segments with high goodness-of-fit, it appears that they become articulatorily linked with time [3], which suggests merging of L1 and L2 phonetic spaces.

The fact that our experiment examined FLA acquisition and not SLA acquisition does mean it does not have a direct bearing on the theoretical implications of models like the PAM-L2. However, Tyler [10] does discuss the implications for the PAM-L2 in an FLA context. He suggests that single-category assimilation (i.e., two L2 segments assimilating to the same L1 segment) will yield an even lower probability of acquisition due to increased poverty of stimulus. However, our research finds that this is only partially correct. In single-category assimilation, the segment with the lower goodness-of-fit can and will be acquired in the FLA context, but that the segment high goodness-of-fit will remain articulatory and potentially cognitively linked to the L1 segment. Variation observed in L1 articulation of /ʃ/ may partially account for this effect, as the internalized acoustic-perceptual representation of /ʃ/ may intersect with those same properties for /ɕ/.

7. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101018840.

8. REFERENCES

[1] Best, C. T. (1995). A direct realist view of cross-language speech perception. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross Language Research*, pp. 171-204. Baltimore: York Press.

[2] Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In Munro, M. J. & Bohn, O. S., (eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, pp. 13-34. Amsterdam: John Benjamins.

[3] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pp. 233-276. Baltimore: York Press.

[4] Flege, J. E. & Bohn, O.S. (2021). The revised speech learning model (SLM-r). In Wayland, R. (ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*, pp. 3-83. Cambridge University Press.

[5] Mathis, A., Mamidanna, P., Cury, K. M., Abe, Taiga, Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 1281-1289.

[6] R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[7] Shoemaker, E. (2016). The development of perceptual sensitivity to Polish sibilants at first exposure. In Albright, A. & Fullwood, M. A. (eds.), *Proceedings of the 2014 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America.

[8] Stefano, C. (2022). *tidymv: Tidy Model Visualisation for Generalised Additive Models*. R package.

[9] Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. New York: CRC Press.

[10] Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In Nyvad, A. M., Hejná, M., Højen, A., Jespersen, A. B. & Sørensen, M. H. (Eds.), *A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*, pp. 607-630. Dept. of English, School of Communication & Culture, Aarhus University.

[11] Wickham. H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[12] Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1): 3-36.

[13] Wrench, A. & Balch-Tomes, J. (2022). Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut. *Sensors*, 1133.

The Impact of Cognitive Load on Speech Production in German-English Bilinguals

Jessica Göbel[1,2], Kathleen M. McCarthy[2]

[1] New York University and [2]Queen Mary University of London
jessica.goebel@nyu.edu, k.mccarthy@qmul.ac.uk

ABSTRACT

This study investigates cognitive load effects on L1 transfer and style-shifting in ten German L1-English L2 bilinguals. Participants were asked to complete two speech tasks, reading and recalling a text, under varying cognitive load conditions. The lower load condition consisted of speakers simply reading or recalling a text out loud. The higher load condition included a distractor task whereby speakers were instructed to read or recall a text out loud while simultaneously concentrating on arithmetical calculations. Both variables examined – word-final light /l/ and word-final devoiced /d/ – showed a very consistent trend towards an increase in the German-accented variant in the greater load condition. In other words, speakers appear to default to a more typical German-accented style when attention is diverted. Few studies have explicitly examined cognitive load effects on bilingual speech production, thus, this paper sheds light onto the cognitive demands that influence linguistic variability for multilinguals.

Keywords: cognitive load, language transfer, bilingualism.

1. BACKGROUND

Language transfer is a relatively well-studied phenomenon in bilingual speech research [9]. However, few studies have explored how fluctuations in cognitive demands can impact the speech production of bilinguals. We know that divided attention, for example, can interrupt the speech processes of monolinguals. Sharma and McCarthy [18] found that an increase in cognitive load can lead to an increase in vernacular features. Furthermore, previous sociophonetic research on monolinguals has found that an increased cognitive load can disrupt speech convergence processes [1; see e.g., 16 for a review]. The current study asks how attentional multi-tasking affects bilingual speech, specifically how it affects the type and rate of L1 transfer in formal speech. Our research question is: Does an increase in cognitive load result in an increase in L1 transfer?

Language transfer is when a linguistic feature is applied from one language to another by a multilingual speaker [9]. Forward transfer, the process whereby a speakers' first language (L1) influences their second language (henceforth L1 transfer) is well-attested in speech production research [e.g., 9; 10]. Multiple factors, such as age of acquisition, proficiency, or the structural differences between the L1 and L2, have been shown to influence the extent of L1 transfer [2; 8].

Another potential factor at play is formality. Passoni et al. [14] found that formality predicted pitch range in Japanese L1-English L2 bilinguals. Bilinguals significantly varied their pitch range according to level of formality in Japanese, but not English. The authors explained this finding in terms of indexical differences between the two languages: pitch is used to index politeness and formality in Japanese but not in English. The current paper aims to build on Passoni et al. [14] by examining bilinguals across formal read speech and casual interview speech, thus furthering understanding of the formality association of non-native accented English.

The role of formality on speech-style has been thoroughly studied in style-shifting research. According to Labov's [11] attention-paid-to-speech model, casual conversational speech has the lowest attention paid to speech while formal read speech has the highest attention paid to speech. Specifically, in casual conversations, speakers tend to revert to their first learned speech style [12]. Importantly, Labov's style shifting model does not unambiguously differentiate between social and cognitive factors that could affect style shifting as interview speech differs from read speech both socially and cognitively.

Recent work on style control has found that cognitive factors, such as cognitive load, might play a key role in style shifting. Sharma and McCarthy [18] examined the effect of cognitive load on style control in monolinguals. They employed a design that allowed them to test the impact of cognitive load while eliminating contextual differences across speech tasks by instructing participants to use a formal news report style throughout. The study found that an increase in cognitive load led to higher rates of vernacular features, thus inhibiting the speakers' ability to maintain their potentially later learned style.

For bilingual speakers, increasing the cognitive load when speaking their L2 could result in an increase in L1 transfer. Sorace [19] found that sentence ambiguities affect L2 syntactic processing. Specifically, sentence ambiguities led to a greater than normal processing load and induced L2 learners into shallow processing, resulting in more syntactic errors. At the phonetic level, Pajak et al. [13] found that that L2 learners have difficulty producing similar-sounding words. They propose that producing highly similar sounding words with novel sound contrasts not present in speakers' L1 increases the cognitive load. This in turn leads to higher rates of non-target-like productions.

The above literature suggests that increased cognitive load impacts language production and perception for bilinguals as it does for monolinguals. However, studies which examine cognitive load in a controlled experimental setting for bilinguals are lacking. This underscores the need for the present paper. In the current study we focused on the speech of German L1-English L2 bilinguals. Based on previous findings, we hypothesize that a higher cognitive load would lead to significantly greater rates of German-accented realisations and L1 transfer. Ten German L1-English L2 bilinguals were asked to complete two speech tasks, reading and recalling a text, under varying cognitive load conditions. The primary aim is to examine cognitive load effects on rates of German-accentedness in order to shed light on how cognitive load can influence language transfer. A secondary aim is investigating how level of formality affects rates of German-accentedness.

Two variables believed to be sensitive to transfer were considered: word-final light /l/ and word-final devoiced /d/. Standard Southern British English has a sharp light /l/ dark /l/ contrast, whereby /l/ is realised as [ɫ] word-finally and as [l] word-initially [21]. In comparison, Standard German lacks this light /l/ dark /l/ contrast. In Standard German, /l/ is always realised as light, making this variable ideal for an analysis of L1 transfer [17].

The second variable, word-final /d/, varies across English and German. Final obstruent devoicing is a highly regular phonological process that takes place in German and does not exist in English. Özaslan and Gabriel [23] find that final obstruent devoicing is transferred from German to English by L2 learners, yielding high rates of non-target-like productions. Although final obstruent devoicing applies to plosives, fricatives, and affricates, the current study will limit its focus to word-final /d/.

2. METHOD

2.1 Participants

The speakers are ten German L1-English L2 bilinguals between the ages of 20 and 29. All are Standard German speakers who considered German to be their first and dominant language. The participants learned English as an L2 in elementary school (except for one speaker who learned English before elementary school, but not from birth). At the time of recording, the speakers lived in Germany or moved out of Germany relatively recently. Age of English acquisition varied from 3 to 12, with an average age of English acquisition of 8.5 years**.**

2.2. Materials and procedure

The experimental design follows Sharma and McCarthy's [18] study on cognitive load and style shifting in monolingual speakers. To maintain the same speech style across the different tasks and cognitive load conditions, participants were instructed to speak as if they were presenting a radio news report across the different tasks. Speakers were given a practice report to read aloud at the start of the experiment to enable them to establish their newsreader style.

The experiment consisted of two tasks. Task 1 (Read Speech) involved reading a text out loud in a formal newsreader style. The task consisted of two conditions. Condition A (lower cognitive load) consisted of speakers simply reading a text out loud. Condition B (higher cognitive load) included a distractor task where speakers were asked to read a text out loud and simultaneously concentrate on an audio stream of numbers (inter-stimulus interval [ISI]: 2 seconds). Speakers were instructed to add up the numbers while they read and report the total sum after giving their report [see 22 for this design].

Task 2 (Recall Speech) consisted of reading a news report and being asked to memorize it as best as they could before giving the report from memory. This task also involved two cognitive load conditions. Condition A (lower cognitive load) simply involved speakers reading, memorizing, and then recalling the text out loud. Condition B (higher cognitive load) also used a distractor task where speakers were asked to read, memorize, and then recall a text while simultaneously concentrating on an audio stream of numbers (ISI: 2 seconds). Speakers were asked to report the first five numbers they heard after completing their report [see 7 for this design]. Following Sharma and McCarthy [18], an easier distractor task was used for recall speech as simultaneous arithmetical calculations were too

difficult due to working memory effects of memorizing a text. Out of ten participants, two speakers correctly solved one of the number tasks in the higher cognitive load condition.

To control for confounding effects of cognitive load order of the tasks, the order of the cognitive load conditions (high or low) was counterbalanced across speakers. To control for confounding effects of task type (i.e, read vs recall), the order of the read and recall speech was counter-balanced across speakers.

After the experimental component was complete, the first author conducted casual sociolinguistic interviews with participants to gather information about speakers' subjective responses to the different tasks, their language background, as well as getting a baseline of casual speech. The interviews and experiment were conducted remotely using Zoom and speakers were asked to record themselves on their phone to ensure that internet connectivity issues would not affect the audio quality. Gorilla Experiment Builder [3] was used to design the experiment so that participants could read and recall texts and listen to the audio distractor at the same time. The interviews were conducted in English, though all speakers knew that the interviewer was a German-English bilingual.

2.3 Scripts and variables

All scripts were adapted from Sharma and McCarthy [18]. The reading scripts were on science news stories to target a formal style but avoid political material that may prompt affect. Texts were edited to include tokens for three variables thought to be sensitive to L1 transfer: word-final /l/ and /d/. Care was taken to control for clausal context, lexical frequency, and phonetic context in the texts as much as possible. Following and preceding phonological contexts of both variables were additionally considered in statistical analyses. Variables were coded auditorily by the first author, consulting the spectrogram in Praat [6] for ambiguous tokens. A random sample of 20% of the tokens were coded by the second author and yielded an 84% inter-rater agreement.

For ambiguous /l/ tokens, the F1 and F2 contours were considered. If the F1 was relatively high accompanied by a very low F2 towards the end of the lateral, the token was classed as dark. If the F1 and F2 contour remained relatively flat, the token was classed as light. For the analysis of word-final /l/ – whether it is realized as dark [ɫ] or light [l] – only non-cluster contexts were considered, following [8]. This meant that word final /l/ could be followed by a non-/l/ consonant, vowel in the following word, or pause.

For /d/, the presence (or absence) of a continuous voice bar determined whether the item was coded as voiced (or devoiced). Coding for word-final /d/ – whether voiced [d] or devoiced [t] – also followed previous work [23]. Word-final and word medial contexts were considered, and the following phonological environments included non-/d, t/ consonants, vowels, or pauses. Preceding contexts included voiced consonants and vowels.

3. RESULTS

To analyse which constraints are the most accurate predictors for L1 transfer, mixed-effect logistic regression models were created in R [15] using the lme4 package [5]. The models included cognitive load (high or low) and task type (read, recall, or casual speech) as fixed factors, and the speaker and word as random intercepts.

3.1 Overall distribution

Table 1 shows the overall proportion of variants for all 10 speakers, across all conditions. In total, 970 tokens were collected, an average of around 50 tokens per speaker per variable. Both variables show nearly a 50-50 split between variants, with rates of word-final light /l/ at 49%, and rates of word-final devoiced /d/ at 54%, which suggests a high degree of variability among speakers.

	Word-final /l/		**Word-final /d/**	
Variant	light*	dark	devoiced*	voiced
N	212	217	292	249
%	49%	51%	54%	46%

Table 1. Overall distribution of all variants: German-accented English* vs. British English realisations

3.2 Cognitive load and task effects

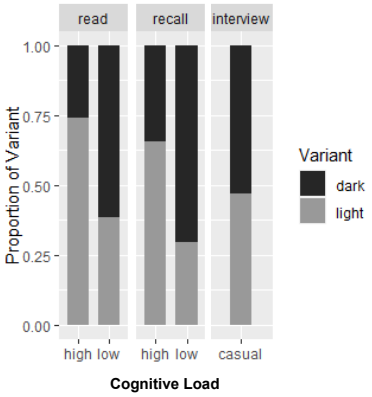


Figure 1. Rates of word-final light /l/ according to task

Figure 1 shows the proportion of word-final light /l/ (the German-accented variant) according to task type and cognitive load. Greater cognitive load elicited

higher rates of German-accented /l/ in both read and recall speech than the lower cognitive load. Additionally, casual speech (interview) had higher rates of the German-accented variant than read and recall speech in the low cognitive load condition. A main effect of cognitive load (Std. Error = 0.428; z = -4.402; $p < 0.001$) and task type (Std. Error = 0.472; z-value = -2.310; $p = 0.021$) was found, whereby interview speech had significantly higher rates of the German-accented variant than read speech. The interaction between task type and cognitive load (Std. Error = 0.722; z-value = -0.441; $p = 0.659$) was not significant.

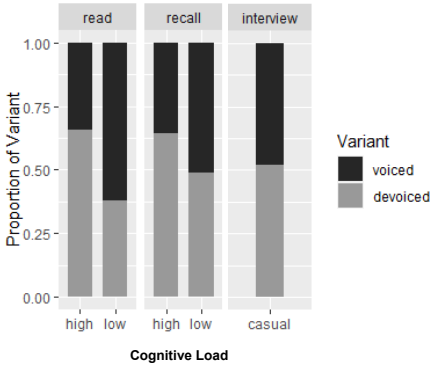


Figure 2. Rates of word-final devoiced /d/ according to task

Figure 2 shows the proportion of word-final devoiced /d/ (the German-accented variant) according to task type and cognitive load condition. Greater cognitive load led to higher rates of the German-accented variant. Additionally, we can see that casual interview speech had the highest rates of German-accented /d/, while formal read speech (in the low load condition) had the lowest rates of German-accented /d/, with recall speech patterning somewhere in between. The mixed model found a main effect of cognitive load (Std. Error = 0.439; z-value = -3.881; $p < 0.001$) and task type (Std. Error = 0.423; z-value = -2.800; $p = 0.005$), whereby interview speech had significantly higher rates of the German-accented variant than read speech. The interaction between task type and cognitive load was not significant (Std. Error = 0.556; z-value = 1.251; $p = 0.211$).

4. DISCUSSION

The main aim of our paper was to examine cognitive load effects on L1 transfer. Both variables show the same cognitive load pattern – greater load conditions led to significantly higher rates of the German-accented variant. As task type and cognitive load were fully controlled, the findings illustrate that speakers were more likely to default to a more typical German-accented style when attention is diverted. A reversion to vernacularity or a first-learned speech style in higher cognitive load conditions has been reported for monolinguals [e.g. 11; 18]. However, to our knowledge, cognitive load has never been examined in a bilingual context in a controlled experimental setting. Thus, this paper furthers understanding of how cognitive load affects L2 speech production and shows that diverted attention can lead to an increase in L1 transfer and cause speakers to default to a more German-accented style. This is potentially due to the cognitive primacy of speakers' first language [11]. Note, however, that the small participant sample renders these results less statistically robust.

Some speakers informally reported that they felt their speech was more German-accented in the greater cognitive load condition. One speaker noted that "in the last reading task it was less of a German accent and the first one was more just because […] you need to focus on the numbers and reading and also […] on your accent, so this is three things at the same time." This provides tentative evidence that some speakers have an awareness of their varying rates of German-accentedness and presents some qualitative evidence for the quantitative patterns observed.

A secondary aim of the current study was to examine the formality-associations of German-accented speech through a consideration of both casual interview and formal newscaster speech. We found that casual interview speech elicited significantly higher rates of German-accented /d/ and /l/ than formal read or recall speech. This finding suggests that German-accented speech is considered less formal than certain native varieties of English. The finding that word-final light /l/ and devoiced /d/ are sensitive to task type (read or interview) provides at least some evidence that rates of German-accentedness is somewhat driven by formality-associations. This finding is not altogether surprising as previous research has found that people associate non-native accents with a lack of linguistic competence and a lack of formality [4; 20].

The effect of task type (read, recall, or interview) and cognitive load (high or low) indicates that rates of L1 transfer are both socially and cognitively driven, whereby rates of German-accentedness increase in casual speech and in higher cognitive load conditions. Note that the effect of cognitive load was stronger than the effect of formality.

In sum, our findings suggest that cognitive factors may play a role in the variation of L2 speech production patterns and should therefore be considered alongside other well-researched social factors (e.g., motivation, attitudes) when conducting multilingual speech research.

5. REFERENCES

[1] Abel, J.,Babel, M. 2017. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech* 20, 479–502.

[2] Abrahamsson, N., Hyltenstam, K. 2009. Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, 59(1), 249–306.

[3] Anwyl-Irvine, A., Massonié, J., Flitton, A., Kirkham, N., Evershed, J. 2019. Gorilla in our midst: an online behavioural experiment builder. Behavior Research Methods. Doi: https://doi.org/10.3758/s13428-019-01237-x

[4] Ballard, L., Winke, P. 2017. Students' Attitudes Towards English Teachers' Accents: The Interplay of Accent Familiarity, Comprehensibility, Intelligibility, Perceived Native Speaker Status, and Acceptability as a Teacher. In T. Isaacs & P. Trofimovich (Eds.), *Second Language Pronunciation Assessment: Interdisciplinary Perspectives* 1(107), 121–140.

[5] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48. doi:10.18637/jss.v067.i01.

[6] Boersma, P, Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.40, retrieved 27 February 2021 from http://www.praat.org/.

[7] Christodoulides, G. 2016. Effects of Cognitive Load on Speech Production and Perception. Doctoral dissertation, Université catholique de Louvain.

[8] de Leeuw, E., Menne, I., Scobbie, J. 2012. Dynamic systems, maturational constraints and L1 phonetic attrition. *International Journal of Bilingualism* 17(6), 683–700.

[9] Flege, J. 1995. Second-language speech learning: Theory, findings, and problems. In: Strange, W. (ed.) *Speech perception and linguistic experience: Issues in cross-language research*, 229–273. Timonium, MD: York Press.

[10] Kim, K., Joo, K. 2021. L1 transfer in the interpretation of L2 reflexive pronouns by child learners of Korean. *International Journal of Bilingualism* 25(6), 1529–1544.

[11] Labov, W. 1966. *The Social Stratification of English in New York C*ity. Cambridge: Cambridge University Press.

[12] Labov, William. 2013. The Language of Life and Death: The Transformation of Experience in Oral Narrative. Cambridge: Cambridge University Press.

[13] Pajak, B., Creel, S. C., Levy, R. 2016. Difficulty in learning similar-sounding words: A developmental stage or a general property of learning?. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(9), 1377.

[14] Passoni, E., de Leeuw, E., Levon, E. (2022). Bilinguals Produce Pitch Range Differently in Their Two Languages to Convey Social Meaning. *Language and Speech*, *65*(4), 1071–1095. https://doi.org/10.1177/00238309221105210

[15] RStudio Team. 2019. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

[16] Roelofs, A., Piai, V. 2011. Attention demands of spoken word planning: A review. *Frontiers in psychology* (2), 1–14.

[17] Recasens, D. 1995. Velarization degree and coarticulatory resistance for /l/ in Catalan and German. *Journal of Phonetics*, 23, 37–52.

[18] Sharma, D., McCarthy, K. 2018. Attentional load and style control. *University of Pennsylvania Working Papers in Linguistics* 24(2), 127–136.

[19] Sorace, A. 2006. Possible manifestations of shallow processing in advanced second language speakers. *Applied Psycholinguistics,* 27, 88–91.

[20] Subtirelu, N. 2015. "She does have an accent but...": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society* 44, 35–62.

[21] Wells, John Christopher. 1982. *Accents of English* (Vols.1&2). Cambridge: Cambridge University press.

[22] Yin, B., Ruiz, N., Chen, F., Khawaja, A. 2007. Automatic cognitive load detection from speech features. In Proceedings of OzCHI 2007.

[23] Özaslan, M., Gabriel, C. 2019. Final obstruent devoicing in English and French as foreign languages: Comparing monolingual German and bilingual Turkish-German learners. In S. Thiele & C. Gabriel (Ed), *Romanische Sprachen in ihrer Vielfalt: Brückenschläge zwischen linguistischer Theoriebildung und Fremdsprachenunterricht*, 177–209. Stuttgart: Ibidem.

The acoustic realization of L2 Spanish phonetic categories and allophonic alternations by English-speaking immigrants

Mark Amengual

University of California, Santa Cruz
amengual@ucsc.edu

ABSTRACT

This experiment analyzes the acoustic realization of English and Spanish voiceless stops /p,t,k/ (*phonetic categories*) and Spanish voiced stops/approximants /b,d,g/-[β,ð,ɣ] (*allophonic alternations*) produced by native English Second Language (L2) learners of Spanish and their Spanish/Catalan/English trilingual children. Fifteen British-born, native English-speaking immigrants, who have resided in Spain for +30 years and fifteen English heritage speakers, who were raised and educated in the same Spanish/Catalan bilingual community, completed a reading-aloud task in which they produced Spanish and English /p,t,k/ and Spanish /b,d,g/-[β,ð,ɣ]. The acoustic analyses suggest that the voiced stop/approximant allophonic alternation in Spanish may be relatively easier to acquire than voiceless stops for highly proficient L2 learners that are long-term residents in the L2-speaking community. These findings add to our understanding of the sources of variation in L2 pronunciation providing insight into the differences in the acquisition of phonetic categories and allophonic alternations in the L2.

Keywords: L2 phonetic categories; Allophonic alternations; Ultimate attainment; Speech production

1. INTRODUCTION

It is well-known that the probability of acquiring a near-native or native-like accent in a foreign language diminishes as we age [1, 23, 24]. Thus, hearing and speaking with a "foreign" or "First Language (L1)-influenced" accent is minimized the earlier the language is acquired [9, 15]. However, we still do not have a clear understanding of the speech production and perception patterns of adults who learn a Second Language (L2) at a later stage in life but who have many years of experience with the language as a result of long-term and continued residence in the L2-speaking community. This brings us to the following question: What does the ultimate L2 attainment (i.e., the acquisitional "end state") look like for highly-experienced long-term immigrants? It may be reasonable to expect that these individuals, who in many cases have spent most of their lives with intensive exposure to their L2 and are immersed in a community where their L2 is exclusively used, are those who are better equipped to overcome age of acquisition effects in their L2 speech. Several recent studies that investigate the speech patterns of long-term L1 English speakers that have immigrated from the USA to Spain [12, 13, 14] suggest that this may not be the case, showing uneven rates of success in the acquisition of a variety of sound classes in Spanish.

Additionally, most research on L2 speech has exclusively examined the acquisition of phonemic contrasts but the phonetic knowledge of L2 learners goes beyond the acquisition of two phonological systems with contrasting sounds. In comparison to how much we know about the acquisition of language-specific phonetic categories in L2 speech, we still know relatively little about the implementation of language-specific phonological processes, such as allophonic alternations and neutralization processes [4, 7, 10, 16, 21, 26]. The present study bridges this gap by comparing the acoustic realization of voiceless stops /p, t, k/ in L1 English and L2 Spanish (*phonetic categories*) and the L2 Spanish voiced stop /b,d,g/-approximant [β,ð,ɣ] (*allophonic alternation*) by advanced L1 English late L2 learners of Spanish and their early simultaneous Spanish-Catalan-English trilingual children.

In order to examine the acquisition of *new phonetic categories* we analyze the Voice Onset Time (VOT) of voiceless stops. This phonetic variable was selected because word-initial Spanish /p,t,k/ have a short VOT and are always unaspirated [p,t,k], with Spanish VOT ranging between 0 and 30 ms [2, 17, 22] whereas word-initial English voiceless stops are produced with a long lag and are aspirated [p^h,t^h,k^h], resulting in a VOT from 30 to 120ms [8, 22]. To explore the acquisition of a *new allophonic alternation*, this experiment investigates a Spanish-specific lenition process (i.e., spirantization), wherein the voiced stops /b,d,g/ in Spanish are articulated with full occlusion only phrase initially, after a homorganic nasal, and for /d/ only when there is a preceding /l/. In all other positions, including in intervocalic position, these voiced stops are commonly realized as voiced approximants [β,ð,ɣ] [18].

2. METHOD

2.1. Participants

The participants in the present study comprise 15 female British-born, native English-speaking long-term immigrants in Spain who are L2 Spanish speakers (Age: M=56.6, SD=7.7) and 15 female early simultaneous Catalan-Spanish-English trilinguals (Age: M=26.4, SD=7.3) who reside in the same Spanish-Catalan bilingual community. Participants reported normal speech and hearing and normal or corrected-to-normal vision, and received a stipend for their participation in the study.

The British-born L1-English-L2 Spanish group consisted of British Expatriates (BE) who were born, raised, and educated in Scotland or England. Most of these native English speakers immigrated to Spain in their early 20s, settled on the island of Majorca, and married a native Spanish-Catalan bilingual partner. At the time of testing they reported (i) that they had been residing in Spain for an average of 32.6 years (SD=10.1), (ii) that they spoke in Spanish with their partners and in the community and English with their children, (iii) that even though they understood Catalan they did not regularly speak Catalan, and (iv) that they were not native speakers of any other language. The early simultaneous trilingual group consisted of English Heritage Speakers (EHS) who were raised, and educated in Majorca (Spain). They had grown up in a trilingual environment speaking English with their mothers, Catalan with their fathers, and Spanish and Catalan in the community.

2.2. Materials and Recording Procedure

The production of the target Spanish and English voiceless stops /p,t,k/ and Spanish /b,d,g/-[β,ð,ɣ] was elicited in a reading-aloud task. The materials consisted of 30 word-initial voiceless stops (10 /p/, 10 /t/, 10 /k/) for each language, in stressed position and preceding a low vowel (e.g., *cama* 'bed', *casa* 'house', cat, catch) and 30 word-initial and 30 intervocalic (10 bilabial, 10 dental, 10 velar) Spanish voiced stops in stressed position and preceding a mid vowel (e.g., *boca* 'mouth', *jabón* 'soap'). Each target item was embedded in a carrier phrase: *I say* TARGETWORD *today* (English), *Puedo decir* TARGETWORD *también* (Spanish). Participants produced two repetitions of each experimental item.

The production task was conducted in a quiet room in the home of each participant, who was comfortably seated in front of a computer display. Participants were told that the study involved reading sentences on a computer screen and that their speech would be recorded for subsequent acoustic analysis. Each sentence was presented on a computer screen and participants were asked to read the sentences clearly and with a natural pace, producing "clearly enunciated speech" [20]. The speech samples were recorded using a head-mounted microphone (Shure SM10A) and a solid-state digital recorder (Marantz PMD660), digitized (44kHz, 16-bit quantization), and computer-edited for subsequent acoustic analysis.

2.3. Acoustic Analysis

The VOT values of the voiceless stops in Spanish and English were obtained from the waveform in Praat [6] by measuring the time interval between the stop release and the onset of voicing as discerned on the waveform as periodic (repeating) cycles. The measurement (rounded to the nearest decimal) was determined from the beginning of the burst (identified by a sharp spike where the waveform changes from quiescent to transient) to the beginning of the first regularly repeating voicing cycle. The point in the first glottal cycle that was counted as the onset of voicing was the initial zero crossing in the waveform.

Two gradient measures of intensity from the target word-initial and intervocalic Spanish /b,d,g/-[β,ð,ɣ] were taken using Praat. The first was the lowest intensity point of the voiced stop/approximant, and the second was the intensity peak of the following vowel. For each item, the valley of the consonant and the peak of the following vowel on the intensity curve were annotated. All points were manually marked, and a Praat script was run to extract all points and subtract all the valleys from their corresponding peaks. The degree of spirantization was determined by the difference between the intensity minimum of the consonant and the intensity maximum of the following tautosyllabic vowel (*IntDiff*; [19]). The more lenited the consonant, the smaller the *IntDiff* (dB) is expected to be with respect to the following vowel. Previous studies have reliably used this measure to determine the degree of spirantization in the production of approximants in Spanish [3, 11].

Each participant produced 120 target voiceless stops and 120 voiced stops/approximants, for a total of 7,200 tokens. Twelve voiceless stops and 54 voiced stop/approximant tokens were excluded due to mispronunciations or recording errors. As a result the dataset comprised 3,588 VOT (ms) and 3,546 intensity (dB) measurements.

3. RESULTS

3.1. Voice Onset Time

In order to compare the VOT values of BE and EHS as a function of their language and point of articulation, a dataset was created including the

average over subjects as a condition of language (Spanish, English) and phoneme (/p/, /t/, /k/). As shown in Figure 1, both groups maintain language-specific VOT values in their production of voiceless stops in each language. In other words, they produce distinct VOT values in English and Spanish.

The dataset was submitted to a mixed-model ANOVA, which was performed using R (R Core Team, 2022), with *Speaker Group* as between-subjects factor, *Language* and *Phoneme* as within-subjects factors, and *Subject* as the random term. The mixed-design ANOVA yielded significant main effects of *Speaker Group* ($F(1,28) = 22.3$, $p<0.001$), *Language* ($F(1,140) = 573.7$, $p<0.001$) and *Phoneme* ($F(2,140) = 53.1$, $p<0.001$). In addition, there was a significant interaction between *Language* and *Phoneme* ($F(2,140) = 4.1$, $p<0.05$).

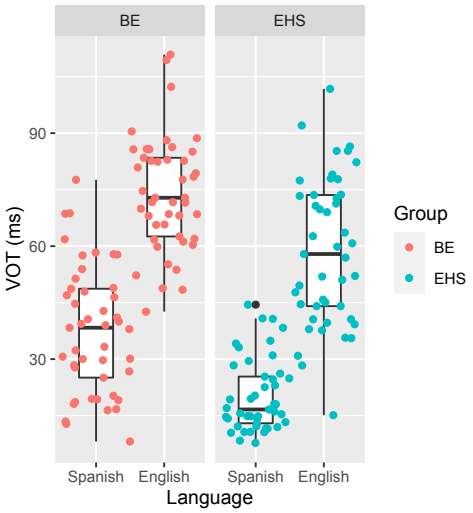


Figure 1: VOT values as a function of group and language.

Figure 2 presents the Spanish VOT values (ms) for each phoneme produced by the BE and EHS. As the main focus of this study is the acoustic realization of L2 Spanish voiceless stops, pairwise comparisons between both participant groups were performed using Tukey's HSD test. In all cases, paired t-tests and effect sizes calculated by means of paired Cohen's *d* are reported.

An analysis of the Spanish VOT data demonstrates that the BE group is far from producing VOT values in the Spanish target range, as shown in the production of the EHS group. The pairwise comparisons revealed a significant difference between the VOT values produced by BE and EHS for /p/ (diff. = 14.69, $t(14) = 4.58$, $p<0.001$, Cohen's $d=1.18$), /t/ (diff. = 19.3, $t(14) = 5.2$, $p<0.001$, Cohen's $d=1.34$), and /k/ (diff. = 18.19, $t(14) = 3.99$, $p<0.001$, Cohen's $d=1.03$).

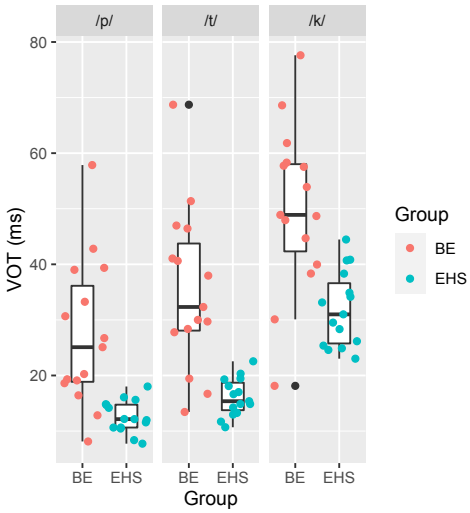


Figure 2: VOT values of Spanish /p,t,k/ by BE and EHS.

3.2. Spirantization

The Intensity Difference (*IntDiff*) data are plotted in Figure 3. The figure includes two panels: one for each speaker group with a boxplot that represents the post-pausal phrase-initial productions, and the other that corresponds to the intervocalic realizations. These data show that *IntDiff* is consistently larger for phrase-initial voiced stops than for intervocalic voiced stops, demonstrating that both groups are applying the spirantization rule in Spanish.

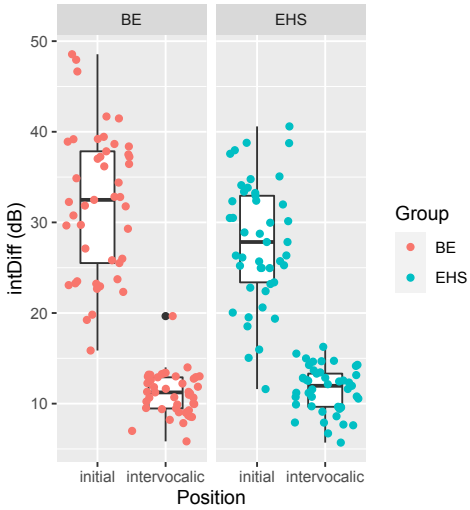


Figure 3: *IntDiff* of word-initial and intervocalic /b,d,g/-[β,ð,ɣ] as a function of place of articulation.

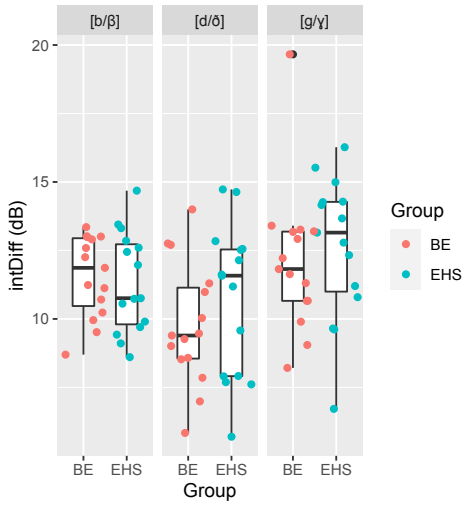


Figure 4: *IntDiff* of intervocalic /b,d,g/-[β,ð,ɣ] as a function of place of articulation.

The *IntDiff* data were also analyzed through mixed-model ANOVAs with *Speaker Group* as between-subjects factor, *Phoneme* ([b/β, d/ð, g/ɣ] and *Word Position* (word-initial, intervocalic) as within-subjects factors, and *Subject* as the random term. The mixed-design ANOVA yielded significant main effects of *Phoneme* ($F(2,140) = 9.7$, $p<0.001$), and *Word Position* ($F(2,140) = 882.2$, $p<0.001$). In addition, there were significant interactions between *Speaker Group* and *Word Position* ($F(1,140) = 14.4$, $p<0.001$) and between *Phoneme* and *Word Position* ($F(2,140) = 6.3$, $p<0.001$). The model did not yield a significant effect of *Speaker Group*.

4. DISCUSSION AND CONCLUSIONS

This study examined the acoustic realization of L2 Spanish voiceless stops /p,t,k/ (*new phonetic categories*) and the L2 Spanish voiced stop /b,d,g/-approximant [β,ð,ɣ] allophonic distribution (*new allophonic alternation*) produced by advanced late learners of Spanish, who are British Expatriates (BEs) and their Spanish-Catalan-English trilingual children, who are English Heritage Speakers (EHSs).

In terms of their acoustic realization of voiceless stops /p,t,k/, the results indicate that even though both groups produced acoustically distinct VOT values in each language, the Spanish VOT values of the voiceless stops produced by the BEs converge towards the English-like range, with intermediate ("compromise") values between the VOTs in Spanish and English as produced by the EHSs. With respect to the acquisition of the voiced stop/approximant allophonic alternation in Spanish, both groups produced a more lenited consonant in intervocalic position than in phrase-initial position, showing that they had acquired the Spanish spirantization rule, producing comparable degrees of spirantization in intervocalic position. These results indicate that BEs seem to fall short of target-like performance with regards to their VOT values but they more closely approximate the *IntDiff* values of their native Spanish-speaking children.

With respect to their ultimate attainment in the acquisition of L2 Spanish voiceless stops (*phonetic categories)* and spirantization in intervocalic voiced stops (*allophonic alternation*), on the one hand, the VOT data appear to show signs of fossilization (arrested development) in the acquisition of L2 learners who are truly exceptional in the amount of extended Spanish immersion and usage they have had during their lives. On the other hand, their acoustic realization of the allophonic alternation between /b,d,g/-[β,ð,ɣ] in Spanish indicates a different, more target-like, acquisitional path. As the data might suggest, are allophonic alternations "easier" to acquire than language-specific phonetic categories? If so, why would this be the case?

The Second Language Perceptual Assimilation model (PAM-L2; [5]) postulates that there are at least two levels involved in cross-linguistic sound interactions: the phonetic and the phonological level. In the acquisition of L2 sounds, the notion of phonetic category coexists with that of the complementary distribution found in allophonic alternations, such as spirantization. The results of this experiment suggest that, at least for experienced late L2 Spanish learners, [β,ð,ɣ] may be easier to acquire than /p,t,k/ because intervocalic articulatory reduction (lenition) can be predicted by a phonological process that is highly frequent in discourse.

Because L2 phonetic categories, such as voiceless stops for English-Spanish bilinguals, consist of cross-linguistically 'similar' sounds but with subtle, fine-grained differences (a tight connection with an L1 category), this may impede 'target-like' acquisition. In contrast, sounds that are only loosely connected with L1 categories (such as [β,ð,ɣ]) may be particularly easier to learn because they are in complementary distribution with /b,d,g/ and the bilingual individual may be more likely to notice, over time, that this phonological process is specific to the L2, and sufficiently different from comparable sounds in the same phonological context in their L1. This predictability and high frequency may attract attention, which would be conducive to phonetic learning.

In closing, two questions can be posed for future research: how close to a 'native-like' range must a learner realize these L2 segments for the acoustic differences to be inconsequential in contributing to a foreign accent? How are these sounds perceived in tandem with other segments that are similarly prone to cross-linguistic influence in L2 speech acquisition?

5. ACKNOWLEDGMENTS

I would like to express my gratitude to Magaly Gastélum, Rebecca Hernández, Serene Tseng, Lindsey Archer, Daisy Soto, Caitlyn Welsh, Katrina Escher, Richard Bibbs, and Chase Rodriques from the UCSC Bilingualism Research Lab for their assistance with the acoustic analyses. This research was supported by the Hellman Fellows fund.

6. REFERENCES

[1] Abrahamsson, N., Hyltenstam, K. 2009. Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning*, *59*(2), 249-306.

[2] Abramson, A. S., Lisker, L. 1973. Voice-timing perception in Spanish word-initial stops. *Journal of Phonetics*, 1, 1–8.

[3] Amengual, M. 2019. Type of early bilingualism and its effect on the acoustic realization of allophonic variants: early sequential and simultaneous bilinguals. *Int. Journal of Bilingualism,* 23(5), 954-970.

[4] Amengual, M., Simonet, M. 2020. Language dominance does not always predict cross-linguistic interactions in bilingual speech production. *Linguistic Approaches to Bilingualism,* 10(6), 141-166.

[5] Best, C.T., Tyler, M.D. 2007. Nonnative and second-language speech perception: commonalities and complementarities. In: Bohn, O.-S., Munro, M. (eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*, 13-34. Amsterdam: John Benjamins Publishing.

[6] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program], Version 6.1.39.

[7] Braun, B., Lemhöfer, K., Mani, N. 2011. Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*(1), 376-387.

[8] Cho, T., Ladefoged, P. 1999. Variation and universals in VOT: Evidence from 18 languages, *Journal of Phonetics 27*(2), 207–229.

[9] DeKeyser, R. 2000. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–534.

[10] Dmitrieva, O., Jongman, A.,Sereno, J. 2010. Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics*, *38*(3), 483-492.

[11] Eddington, D. 2011. What are the contextual phonetic variants of in colloquial Spanish? *Probus, 23*(1), 1–19.

[12] Face, T. 2021. Ultimate attainment of Spanish laterals by native English-speaking immigrants in Spain. *Journal of Second and Multiple Language Acquisition*, 9(1), 167-180.

[13] Face, T. L. 2018. Ultimate attainment in Spanish spirantization: The case of US-born immigrants in Spain. *Spanish in Context*, *15*(1), 27-53

[14] Face, T. L., Menke, M. R. 2020. L2 acquisition of Spanish VOT by English-speaking immigrants in Spain. *Studies in Hispanic and Lusophone Linguistics*, *13*(2), 361-389.

[15] Flege, J. E. 2003. Assessing constraints on second-language phonetics and phonology. In: Meyer, A. Schiller, N. (eds.), Phonetics and phonology in language comprehension and production, differences and similarities, 319–355. Berlin, Germany: Mouton de Gruyter.

[16] Flege, J. E., Bohn, O. S. 1989. An instrumental study of vowel reduction and stress placement in Spanish-accented English. *Studies in Second Language Acquisition*, 11, 35-62.

[17] Flege, J., Eefting, W. 1986. Linguistic and developmental effects on the production and perception of stop consonants. *Phonetica*, 43, 155–171.

[18] Harris, J. W. 1969. *Spanish phonology*. Cambridge, MA: MIT Press.

[19] Hualde, J. I., Simonet, M., Nadeu, M. 2011. Consonant lenition and phonological recategorization. *Laboratory Phonology*, 2(2), 301–329.

[20] Kessinger, R.H., Blumstein, S.E. 1997. Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, 25, 143-168.

[21] Lee, B., Guion, S. G., Harada, T. 2006. Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition*, 487-513.

[22] Lisker, L., Abramson, A. 1964. A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, 20, 384–422.

[23] Moyer, A. 1999. Ultimate attainment in L2 phonology: The critical factors of age, motivation, and instruction. *Studies in Second Language Acquisition*, 21(1), 81-108.

[24] Patkowski, M. 1990. Age and accent in a second language: A reply to James Emil Flege. *Applied Psycholinguistics*, 11, 73–89.

[25] R Core Team (2022). R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria), http://www.R-project.org/ (Last viewed November 14, 2022).

[26] Smith, B. L., Hayes-Harb, R., Bruss, M., Harker, A. 2009. Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German. *Journal of Phonetics,* 37(3), 257-275.

EXPLORING THE PROBLEM IN NATIVE DANISH LISTENERS' PERCEPTION OF THE ENGLISH *HOT* AND *HUT* VOWELS

Jonas Villumsen[1,a], Ocke-Schwen Bohn[2,a], Camilla S. Horslund[3,b]

[a)] Department of English, Aarhus University, Denmark
[b)] Department of Linguistics, Cognitive Science, and Semiotics
[1)] jv@cc.au.dk, [2)] ocke.bohn@cc.au.dk, [3)] horslund@cc.au.dk

ABSTRACT

Native Danish speakers tend to neutralise the vowels /ɒ, ʌ/ of Standard Southern British English in production, but research on the cause of this is sparse. This study examined the identification of these vowels by Native Danish and Native English listeners. As expected, Native Danish listeners identify /ɒ/ and /ʌ/ less accurately than Native English listeners. Exploration of by-token accuracy in F1–F2 space suggests that the location of the boundary between /ɒ/ and /ʌ/ differs for Native Danish and Native English listeners. Comparison with previous research indicates the contrast may span multiple Native Danish vowel categories. Taken together, these findings imply that inaccurate perception of the vowels by Native Danish listeners at least partially explains the vowels' neutralisation in production.

Keywords: second language speech learning, cross-language vowel perception, English, Danish

1. INTRODUCTION

Native Danish (ND) speakers have been observed to pronounce the Standard Southern British English vowels /ɒ, ʌ/ identically [5, 8, 9]. As identified by Native English (NE) listeners, both vowels sound like /ʌ/ when produced by ND speakers [5]. No explanation has yet been posited as to why ND speakers exhibit this exact pattern of mispronunciation. This paper explores the problem.

Perception of this vowel contrast likely plays an important role in its production problems. The /ɒ, ʌ/ vowels occupy an area of the vowel space which is sparsely populated in Danish (see Figure 1). This means that ND listeners will have encountered few vowel tokens in this area in native speech perception, which may have desensitised ND listeners to vowel contrasts in that area [4, 6]. It further adds to the problem that ND listeners are exposed to heterogenous varieties of English whose phonetic realisations of /ɒ, ʌ/ are diverse [5].

When ND listeners categorise /ɒ, ʌ/ in terms of Danish vowel categories, results suggest English /ɒ/ assimilates perfectly to Danish /ɔ/, whereas /ʌ/ assimilates inconsistently to Danish /ɔ/ as well as Danish /æ/ [7]. Viewed from the Perceptual Assimilation Model [2], /ɒ, ʌ/ are in a categorised–uncategorised relationship.

Few other tests of ND listeners' perception of /ɒ, ʌ/ have been conducted. Those that exist show great variation in performance. The vowels are discriminated in the range between 50–70% correct, depending on the order of vowel presentation [12]. Identification ranges between 65–85% [10], but those data were sampled from students of English at a university rather than from the general population.

The present paper reports on a study in which ND listeners and NE listeners identified /ɒ/ and /ʌ/ tokens presented in a /hVt/ environment, comparing the accuracy of the two groups. We also compare two types of ND listeners on the same task: high school students and first semester English students at a Danish university. ND students of English have been sampled previously in investigations into the /ɒ, ʌ/ problem [10]. Comparing their performance to that of a more general sample (high school students) tests the validity of that practice. Finally, as a step towards explaining the perceptual problem, this study explores whether misidentification of /ɒ, ʌ/ by ND

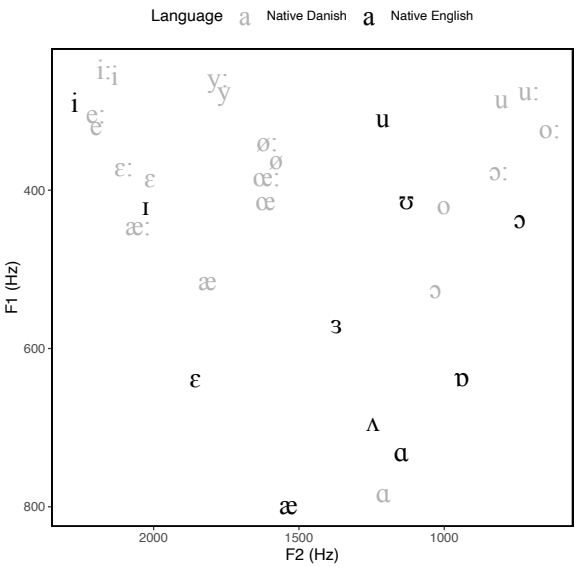


Figure 1. F1–F2 plot of L1 Danish vowels (grey) and L1 English vowels (black) based on data from [14].

listeners is related to the acoustic properties of individual vowel tokens.

2. METHODS

41 ND speakers and 18 L1 British English speakers participated in the study. Of the ND participants, 18 were first semester students of English at Aarhus University and 23 were senior year high school students. The age range of the participants was 18 to 26 years.

The stimuli comprised 48 unique tokens of *hot* and *hut* produced four times each by six male native talkers of Standard Southern British English. *Praat* [3] was used to analyse the first and second formant frequency (F1 and F2) of the stimuli at the midpoint of each vowel token.

Participants performed a two-alternative forced choice identification task with a total of 96 trials. Each of the 48 unique stimuli was presented twice, in a pseudo-randomised order. In a trial, participants heard either *hot* or *hut*, respectively containing /ɒ/ and /ʌ/. They then responded by pressing one of two buttons labelled *hot* and *hut*, with an inter-trial-interval of 1200 ms. The position of the buttons remained constant throughout the task, which lasted 4 to 7 minutes.

3. RESULTS

The data were modelled using mixed effects logistic regression, implemented in *R* [13] using the package *lme4* [1]. Model predictions were derived using the *emmeans* package [11]. Model selection proceeded from the maximal structure with stepwise reduction in random-effects until convergence. The final model estimated the correctness of per-trial responses as a function of Group (High school students, University English students, and NE listeners). Group was dummy coded with University English students as the reference level. The selected method for obtaining *p*-values was likelihood ratio tests. The final *lme4* formula, including random effects, was as follows:

```
Response ~ 1 + Group +
    (0 + Vowel | Participant) +
    (1 + Vowel + Group | Talker)
```

3.1 Model effects

The fixed effects of the levels of Group were as follows: University English students (intercept) were estimated at 1.84 log odds accuracy. High school students performed worse than University English students, $\beta = -0.90$, $SE = 0.30$, $p < .01$. NE participants perform considerably better than University English students, $\beta = 2.76$, $SE = 0.51$, $p < .001$. Model predictions in probabilities are plotted in Figure 2.

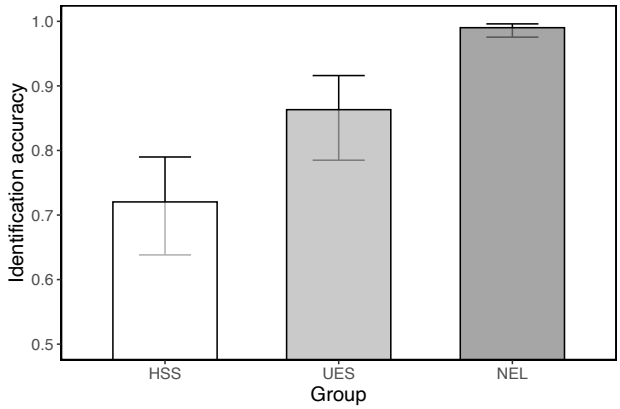


Figure 2: Identification accuracy by Native English listeners (NEL), university English students (UES), and high school students (HSS). Error bars: 95% confidence intervals.

In the random effects, there was variation between participants on their accuracy on each vowel, $SD_{hot} = 0.78$, $SD_{hut} = 0.85$. Accuracy on each vowel correlated highly within participants, $r = 0.91$. Within stimulus talkers, there was variation in accuracy on *hot*, $SD = 0.78$, with high variation in *hut*, $SD = 1.26$. Importantly, the intercept–slope correlation of Vowel within Talker was strongly negative, $r = -0.87$. This means that the difference between accuracy on *hot* and *hut* tokens from the same talker can be predicted from the accuracy of *hot* tokens: Higher accuracy on *hot* implies a negative difference—relatively lower accuracy—for *hut* from the same talker; lower accuracy on *hot* implies a positive difference—relatively higher accuracy—for *hut* from the same talker. This will be further explored in the next section.

3.2 Data exploration: accuracy by vowel token

The model's random effects suggested unexpected systematic variation in accuracy in vowel accuracy by talker. To explore what the model reflects, Figure 3 plots the averaged listener accuracies on both vowels within each talker. Note that the plotted values are not estimates from the model but manually computed averages. For ND listeners, the figure shows that most talkers' *hut* tokens are identified less accurately than their *hot* tokens; but tokens from talkers 4 and 5 exhibit the reverse relationship. This caused the model to estimate a high intercept–slope correlation in the random-effect term of vowel within talker; the lower the intercept (i.e., accuracy for *hot*), the higher the slope for vowel (i.e., accuracy for *hut*). The fact that NE listeners do not show the same pattern

suggests that the reversal is not due to talker idiosyncrasy. Rather, the cause of the reversal is likely to be due to the acoustic properties of the vowels themselves, as well as ND listeners' use of these as perceptual cues.

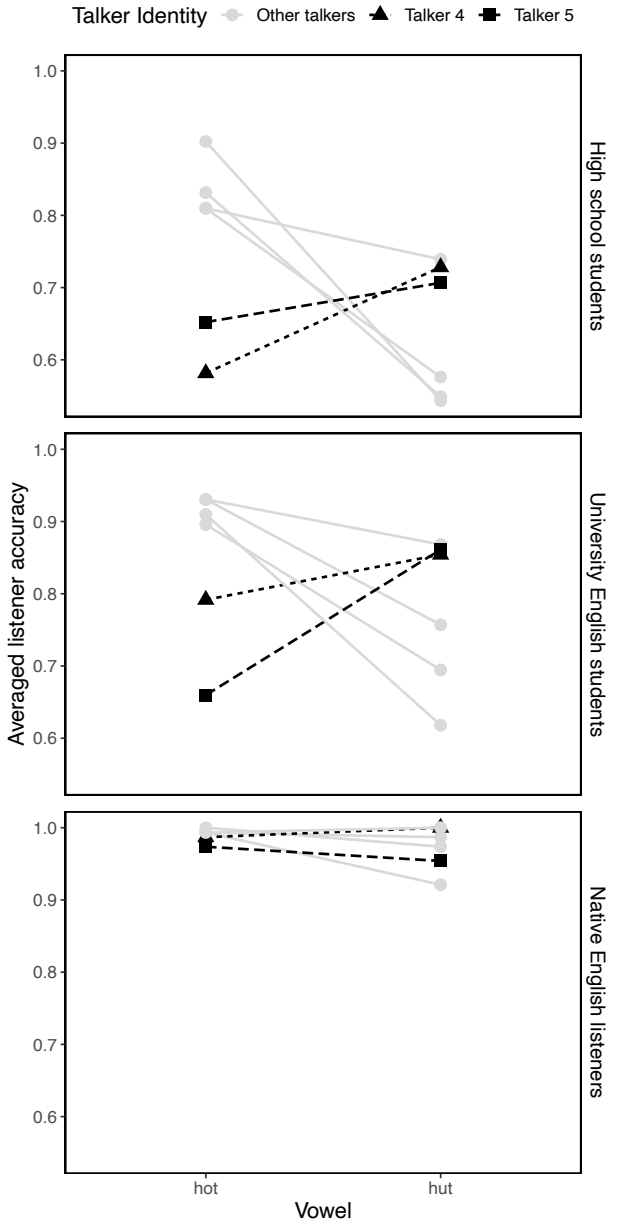


Figure 3. Group identification accuracy on vowels from each talker. To illustrate the reverse vowel relationship (see text), talkers 4 and 5 are highlighted.

Figure 4 plots all unique vowel tokens in the stimuli in F1–F2 space. The tokens are grey-scaled by average identification accuracy. Both ND groups exhibit the same pattern: With both vowels, some tokens are perceived highly accurately while others are perceived less accurately with a few being close to chance-level identification. Accuracy on each vowel appears to depend slightly differently on F1 and F2. Accuracy on /ɒ/-tokens appears to depend on F2, where most are identified accurately until F2 exceeds 945 Hz, at which point an abrupt decline in accuracy occurs. The /ɒ/-tokens where F2 > 945 Hz are exactly those produced by talkers 4 and 5 for whom the relative listener accuracies of /ɒ, ʌ/ were reversed (see Figure 3). Conversely, accuracy on /ʌ/-tokens does not appear to be sensitive to F2 variation. Instead, a small decrease in accuracy occurs when F1 surpasses about 675 Hz. Figure 5 depicts these relationships explicitly with smoothing lines (see especially the middle two panels). It appears that, in the case of /ɒ/-tokens with F2 > 945 Hz, ND listeners experienced an abrupt boundary shift at which point they became uncertain as to what they were perceiving. In the case of /ʌ/-tokens, these listeners experienced slightly more uncertainty as F1 increased past 675 Hz.

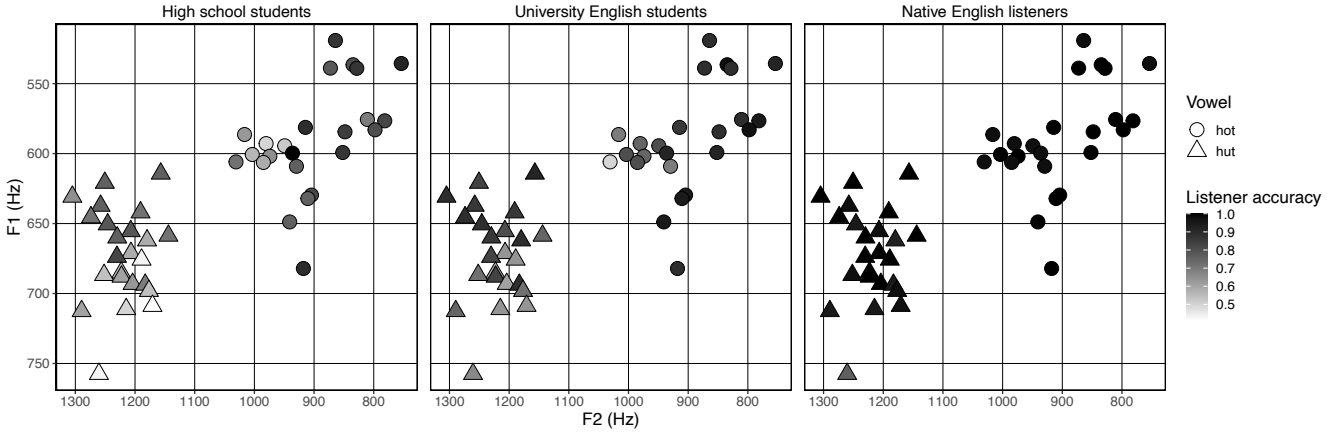
4. DISCUSSION AND CONCLUSION

The present study investigated ND and NE listeners' identification of Standard Southern British English /ɒ/ and /ʌ/. It furthermore compared university English students to high school students to test whether the former constitute an adequately random sample. A difference was found in both comparisons. ND university English students perform worse than NE listeners but better than ND high school students. Thus, University students of English constitute a poor stand-in for the general Danish population, at least when it comes to the present vowel contrast.

The study also examined whether the first and second formant frequency influenced ND listeners' responses. We found that (1) ND listeners identify some tokens of both vowels near ceiling, (2) they identify a few vowel tokens within each category with much lower accuracy, sometimes at chance-level, and (3) low-accuracy tokens tend to cluster in the same areas in F1–F2 space within each category.

When these results are compared to those of NE listeners, it is clear that ND listeners rely on F1 and F2 as cues for the two vowels in slightly different ways from NE listeners.

Foremost, the abrupt shift in accuracy on /ɒ/ at F2 > 945 Hz is only observed with ND listeners. This indicates that ND listeners perceive a category boundary at F2 ≈ 945 Hz. That category could represent an emerging second language vowel. Alternatively, it could represent a pre-existing native language vowel because ND listeners have previously been shown to categorise EN /ɒ/ highly consistently as DA /ɔ/ [7].



(Above) Figure 4. Vowel tokens in F1–F2 space. Dots are *hot* tokens and triangles are *hut* tokens. Each token is shaded according to its average proportion correct score (black is 1), split into the three groups

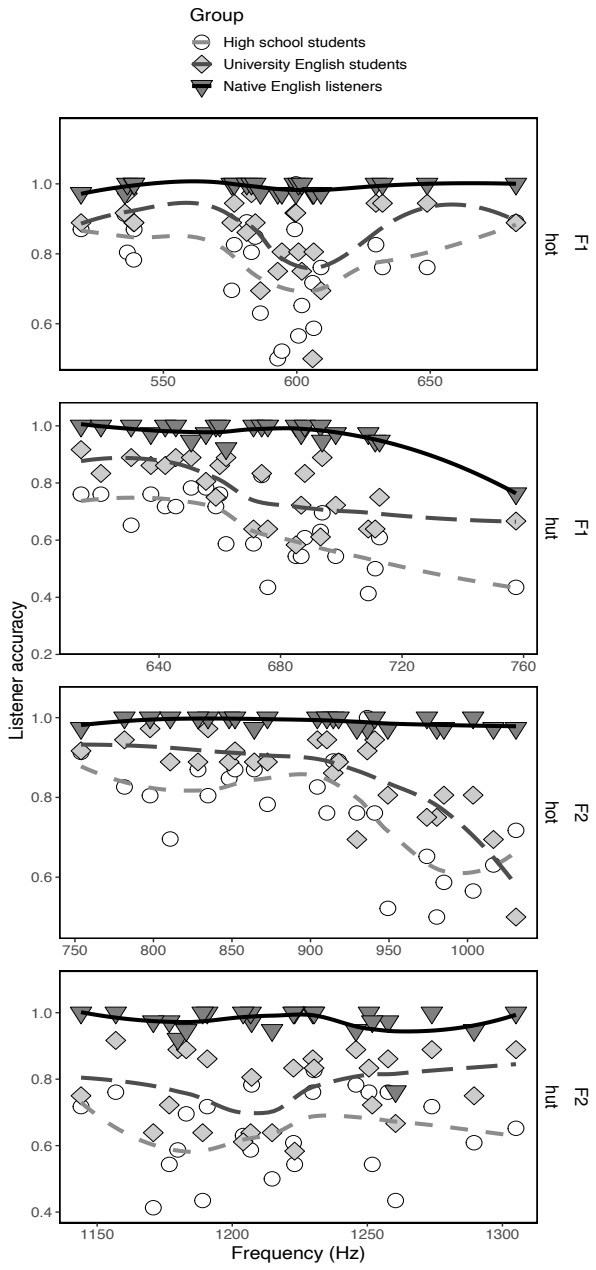


Figure 5. Listener accuracy by formant frequency for each vowel, fitted with separate local polynomial regression lines per group. Note that the label for each plot appears on the right.

It is possible that the acoustic properties of tokens in the /ɒ/–/ʌ/ contrast span multiple ND vowel categories, or even emerging L2 EN vowel categories, at various levels of goodness of fit. In the cross-language identification task reported in [7], ND listeners categorised EN /ʌ/ inconsistently as either DA /ɔ/ or DA /æ/. In the task employed in the present study, chance-level identification for any particular token could indicate one of two things: Either the token is truly intermediate between the two response categories, or it is a poor fit for either of the categories. In the present results, the /ʌ/ tokens that are identified at chance level may be a poor fit for either *hot* or *hut*. These tokens could be perceived as instances of DA /æ/. Future research should explore which categories ND listeners perceive in this area of F1–F2 space by presenting them with tokens from an array of orthogonally varied formant frequencies.

In conclusion, this study demonstrated that ND listeners perceive Standard Southern British English /ɒ, ʌ/ differently than NE listeners according to the acoustic properties of the vowel tokens. More research is needed to determine precisely why. With regard to ND speakers' propensity to neutralise the two vowels in production, the present results make it seem likely that at least part of problem is perceptual in nature.

6. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67 (1). https://doi.org/10.18637/jss.v067.i01.

[2] Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues*

in Cross-Language Research (pp. 171–204). York Press.

[3] Boersma, P., & Weenik, D. (2022). *Praat* (Version 6043) [Windows/MacOS]. http://www.fon.hum.uva.nl/praat/

[4] Bohn, O.-S. (2019). Cross-Language Phonetic Relationships Account for Most, But Not All L2 Speech Learning Problems: The Role of Universal Phonetic Biases and Generalized Sensitivities. In *Approaches to the Study of Sound Structure and Speech*. Routledge.

[5] Bohn, O.-S., & Bundgaard-Nielsen, R. L. (2008). Second language speech learning with diverse inputs. In T. Piske & M. Young-Scholten (Eds.), *Input Matters in SLA* (pp. 207–218). Multilingual Matters. https://researchdirect.westernsydney.edu.au/islandora/object/uws%3A25770/

[6] Bohn, O.-S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, *11*(3), 303–328. https://doi.org/10.1017/S0142716400008912

[7] Bohn, O.-S., & Steinlen, A. K. (2003). Consonantal Context Affects Cross-Language Perception of Vowels. *ICPhS 15*, 4.

[8] Davidsen-Nielsen, N. (1970). *Engelsk Fonetik*. Gyldendal.

[9] Davidsen-Nielsen, N., Færch, C., & Harder, P. (1982). *The Danish Learner*. Antony Taylor.

[10] Kristiansen, A.-M. (2000). *The perception and production of the British English vowel contrast hot-hut by native Danish learners* [Unpublished MA thesis, English Department, Aarhus University].

[11] Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (R package version 1.5.4). https://CRAN.R-project.org/package=emmeans

[12] Polka, L., & Bohn, O.-S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, *39*(4), 467–478. https://doi.org/10.1016/j.wocn.2010.08.007

[13] R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.0.4). R Foundation for Statistical Computing. https://www.R-project.org/

[14] Steinlen, A. K. (2005). The influence of consonants on native and non-native vowel production: A cross-linguistic study (Vol. 30). Gunter Narr Verlag.

DYNAMIC TONGUE MOVEMENTS IN L1 JAPANESE AND L2 ENGLISH LIQUIDS

Takayuki Nagamine

Lancaster University
t.nagamine@lancaster.ac.uk

ABSTRACT

Dynamic tongue movements of intervocalic English and Japanese liquids are analysed based on ultrasound data obtained from 17 L1 Japanese and 12 L1 North American English speakers. Principal Component Analysis (PCA) identified three key articulatory properties, including variation in tongue retraction and height as well as tongue tip movement. The time-varying changes showed different magnitude and timing for tongue retraction for Japanese speakers compared to English speakers. Substitution of English liquids with Japanese /r/ was not clearly observed in the tongue movement pattern. The findings highlight the complexity involved in the articulation of English liquids in L2 speech and the usefulness of the finer-grained articulatory analysis in understanding the particular challenges L2 learners face in producing English liquids in L2 speech.

Keywords: L2 speech production, liquids, ultrasound, Principal Component Analysis

1. INTRODUCTION

The acquisition of English liquids /l/ and /ɹ/ presents particular challenges to L2 learners, notably to L1 Japanese speakers [1]. English /l/ and /ɹ/ can be considered gesturally complex sounds because multiple articulatory gestures need to be coordinated spatially and temporally [2, 3]. English /l/ requires two lingual gestures, tongue coronal and dorsal gestures, which are patterned differently depending on the syllable position. Similarly, English /ɹ/ requires coordination of labial, tongue anterior and posterior gestures, and the tongue shape shows substantial cross-speaker variation [4, 5].

While it is generally agreed that L2 learners' difficulty in L2 speech production is rooted in the perception of L2 sounds [1], adult L2 learners' established articulatory routines that are optimised for L1 production could also constrain the accurate production of L2 sounds, especially gesturally complex sounds such as English liquids [6]. L2 learners struggle to produce articulatory gestures that are absent in their L1 [7]. Previous articulatory studies suggest that Japanese speakers' tongue dorsum movement may be the key for them to produce L2 English liquids accurately. Less advanced L1 Japanese learners of English exhibit little movement in the tongue posterior compared to the more experienced learner or the L1 English speaker [8]. Also, the substitution of English liquids with Japanese /r/ was not clearly observed in articulatory data even for less advanced Japanese learners of English [9]. These studies illustrate a learning scenario of L1 Japanese learners of English; while they attempt to differentiate English liquids from Japanese /r/, they struggle to realise the dorsal gesture in producing English liquids.

Japanese has one liquid category /r/, canonically realised as alveolar taps or flaps [ɾ] [10]. In previous research, whether a dorsal gesture is actively involved in Japanese /r/ remains unclear. The tongue dorsum in plain taps [ɾ] shows a stronger coarticulatory effect with the vowels than the palatalised taps [ɾʲ] [11]. Compared to alveolar stops, however, the tongue dorsum in alveolar taps/flaps is more retracted and stabilised [12, 13]. It would, therefore, be useful to compare the tongue movement between English and Japanese liquids in discussing the articulatory L1 influence in L2 speech, especially with regard to the tongue dorsum gesture.

Identifying specific articulatory difficulties in L2 speech also has theoretical implications. The Speech Learning Model (SLM) [14] hypothesises that articulatory realisation rules for L2 sounds are specified in the phonological representation. The Perceptual Assimilation Model for L2 learning (PAM-L2) [15] posits that L2 learners perceive articulatory gestures directly, and they assimilate L2 sounds into the L1 phonological categories based on the gestural information. The case of Japanese speakers' acquisition of English liquids could be a good testing ground as to whether and what articulatory information is important for successful L2 speech learning.

Building on the previous research, the current

study aims to identify and compare key articulatory properties involved in English liquids produced by L1 English and L1 Japanese speakers. I particularly focus on the tongue dorsum movement for which I expect to observe the L1 influence carried over from Japanese /r/. Methodologically, the use of ultrasound tongue imaging would complement the findings of previous qualitative articulatory work.

2. METHOD

2.1. Participants

Data from 29 speakers are analysed in the study, including 17 L1 Japanese (eight females and nine males, M_{age} = 19.76 years, SD_{age} = 0.97) and 12 L1 North American English speakers (ten females and two males, M_{age} = 29.08 years, SD_{age} = 6.30).

The Japanese speakers represent the typical English-as-a-foreign-language (EFL) learner population in Japan. They are university students from two universities located in Western and Central Japan who studied English mostly through the school and the university curriculum. They had little experience of a long-term stay outside Japan, except for six who had stayed in an English-speaking country for up to four months.

The North American English speakers were recruited in London and Lancaster in the UK, eight of whom came from the US and four from Canada. Although one of the Canadian speakers was originally born in Poland, she moved to Canada early in her childhood and thus considered herself an L1 Canadian English speaker. No participants reported any history of speech or hearing impairments.

2.2. Materials

Intervocalic English and Japanese liquids are analysed, elicited with the words *believe* /biˈliːv/, *bereave* /biˈɹiːv/, and *biribiri* /biɾibiɾi/ (びりびり). These words are a subset of a larger data collection session. The flanking vowel /i/ is chosen because the vowel quality is similar in Japanese and English and is appropriate for cross-linguistic comparisons [10]. *biribiri* is a Japanese mimetic word that describes the sound and/or the situation of paper being torn. The intervocalic environment most likely yields the canonical tap/flap realisations of Japanese /r/ as it is subject to allophonic variations in other environments [16]. The Japanese speakers were asked to read the Japanese *biribiri* in the LHHH accent so that the liquids in both languages appeared as an onset consonant in an accented syllable.

2.3. Data collection

Participants wore an ultrasound stabilisation headset to stabilise the ultrasound probe under their lower jaw [17]. At the beginning of the recording, they were asked to bite a plastic plate to measure their occlusal plane [18]. Then, the participants read aloud the target words in isolation, resulting in 271 tokens for analysis. The Japanese participants' language modes were controlled by changing the language of instructions and including an English conversation activity between the Japanese and English recording blocks.

Ultrasound data were obtained using a Telemed MicrUs system, with a 64-element probe of 20 mm radius, recorded with the Articulate Assistant Advanced (AAA) software version 220.4.1 [19]. Midsagittal tongue views were imaged with a fixed probe frequency between 2-4 MHz, 80 mm depth, 100% field of view and 64 scan lines, resulting in a framerate of ca. 80 per second. Simultaneous acoustic signals were collected with the signal pre-amplified and digitised using a USBPre2 audio interface, and then recorded onto a laptop computer at 44.1 kHz with 16-bit quantisation. Side-profile lip images and participants' perceptual identification accuracy data for English /l/ and /ɹ/ were also collected but will not be presented here.

2.4. Data analysis

Tongue spline data were exported from AAA using the DeepLabCut (DLC) plug-in [20]. DLC estimates the tongue shape based on 11 key points along the tongue contours in the ultrasound videos based on the trained models [21].

The current study takes the dynamic measurements throughout the vowel-liquid-vowel(V_1LV_2) interval in the target words (i.e., ***belie**ve*, ***berea**ve*, and ***biri**biri*) under the assumption that English liquids exhibit dynamic changes in the acoustic and articulatory realisations, and it is often difficult to specify a particular time point that best represents the liquid quality [22, 23]. Segmentation was carried out based on the acoustic signals using Montreal Forced Aligner [24], with the V_1 onset and the V_2 offset marked at the point where periodic cycles began or ended in waveforms and where the formant structures were clearly visible.

Principal Component Analysis (PCA) has been performed in order to summarise the tongue spline data into a manageable number of key articulatory dimensions that allow for cross-speaker comparisons [25]. Prior to running PCA, data from each speaker were normalised into *z*-scores to allow for

comparisons across speakers. Then, PCA was run based on all the *x* and *y* coordinates of the 11 points along the tongue surface at 11 equidistant time points during the V_1LV_2 interval produced by all speakers. PCA was performed using the *princomp* function in R [26]. The code and data for analysis are available online at https://osf.io/29tac/.

3. RESULTS

3.1. Identifying key articulatory properties using PCA

PCA identified three key dimensions that account for 88.33% of the data, with each dimension exceeding the 5% threshold suggested in the literature [27]: PC1: 58.03%, PC2: 22.65%, and PC3: 7.65%. In order to make the PCs interpretable, the loadings of each PC are plotted against the mean tongue shape using the standard deviation information [25]. Fig 1 shows key dimensions involved in the production of the V_1LV_2 sequence concerning the degrees of tongue retraction (PC1: left), tongue height (PC2: middle), and the variation in the tongue tip (PC3: right). Given the focus of the current study being the tongue posterior movement, I will focus on PC1 in the following sections.

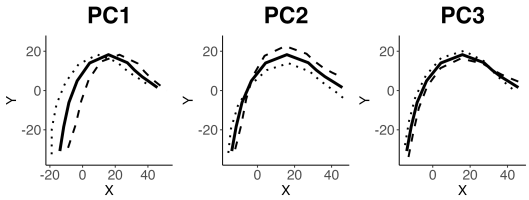


Figure 1: Variation captured in PCs 1 to 3. The thick line represents mean tongue for all of the data for all speakers, with the dimension of variation captured by that PC shown in the dashed and dotted lines.

3.2. Dynamic changes in tongue retraction

Fig 2 presents the time-varying changes of the PC1 values (i.e., tongue retraction) during the V_1LV_2 intervals aggregated for the L1 English and Japanese speakers, where larger PC values indicate more tongue fronting. The overall shape of the PC1 trajectories shows a somewhat similar front-back tongue movement for English liquids for both English and Japanese speakers. The variation associated with the degree of tongue retraction, however, is smaller for the Japanese speakers than the English speakers during the production of *believe* (blue). Regarding *bereave* (yellow), despite a similar degree of tongue retraction between the two speaker populations, the Japanese speakers seem to show a different timing in which they achieve tongue retraction later during the interval than the English speakers do. Finally, changes in PC1 for Japanese /r/ in *biribiri* (grey) were different from either English /l/ or /ɹ/.

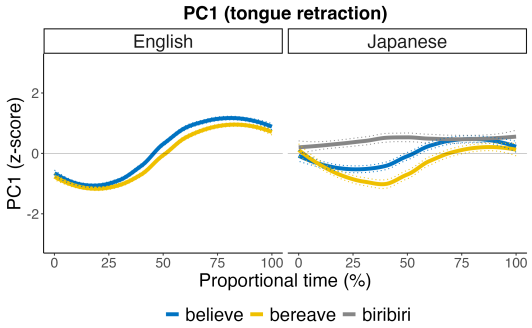


Figure 2: Time-varying changes of the tongue retraction (PC1).

In light of the articulatory patterns identified by PCA, three speakers have been selected to compare the articulatory differences in detail: a female speaker of Canadian English ('English A'), a male and a female EFL Japanese learner of English ('Japanese A' and 'Japanese B', respectively). The midsagittal tongue splines extracted at 11 equidistant points during the V_1LV_2 intervals are presented in Fig 3 and the time-varying changes in PC1 in Fig 4. The two Japanese speakers are chosen based on the auditory impression of the author.

Japanese A maintains an auditory three-way contrast among the three liquids, which is also obvious in the midsagittal tongue shapes. Changes in the PC1 values, however, suggest that the front-back tongue movement would be different from that of English A. Similarly, despite her clear substitution in the auditory analysis and the similarity between English and Japanese liquids in the tongue shapes, the PC1 movement for Japanese B's English /l/ and /ɹ/ is different from that for Japanese /r/.

4. DISCUSSION

The main findings of the current study include that English and Japanese speakers differed in the magnitude of tongue retraction for *believe* and timing for *bereave*. The results could be taken as evidence that the tongue posterior movement may impose difficulty on Japanese EFL learners in articulating English /l/ and /ɹ/ accurately.

While only one vowel environment has been investigated in this study, the current results could

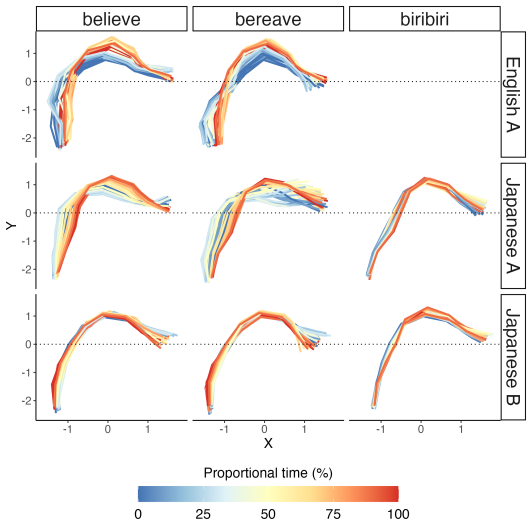


Figure 3: Midsagittal tongue shapes during the V_1LV_2 intervals for a Canadian English speaker (top) and two Japanese speakers (middle and bottom). Tongue tip to the right.

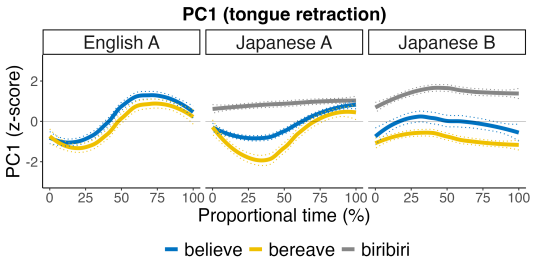


Figure 4: Time-varying changes of the tongue retraction (PC1) during the V_1LV_2 interval for a L1 English and two L1 Japanese speakers.

nevertheless characterise the front-back dimension of the tongue movement for Japanese /r/ compared to the English liquids. The changes in PC1 in Fig 2 suggest minor movements of the tongue posterior over the course of the V_1LV_2 sequence. This agrees with the previous claims that the tongue dorsum for the alveolar taps/flaps shows a substantial coarticulatory effect with the flanking vowels [11] and the dorsal 'stabilisation' strategy [12].

The time-varying trajectory of the front-back movement for *believe* in Fig 2 is 'flatter' for the Japanese speakers than for the English speakers. This may indicate the dorsal stabilisation carried over from Japanese /r/ to the English liquids. In addition, the trajectories for *bereave* (yellow in Fig 2) suggest different gestural timing patterns between Japanese and English speakers' production, such that the Japanese speakers achieved tongue retraction later than the English speakers. Overall, these results could provide evidence for the previous claim that L1 influence is observed in the tongue posterior movement in L1 Japanese speakers' production of English liquids [8].

The individual data replicate the previous findings [9] that a less advanced EFL learner (i.e., Japanese B) does not substitute English liquids with Japanese /r/ completely given the clear differences in the trajectory height (see Fig 4). SLM might explain that she forms separate articulatory realisation rules for English /l/, /ɹ/ and Japanese /r/. Under PAM's account, she might not yet be fully capable of using the dorsal gesture in learning the L1-L2 contrast of liquids. Given that dynamic information of segments might need to be part of phonological representation [28], the dynamic approach in this research is worth pursuing, especially for English liquids that show dynamic changes in articulation.

Finally, the interpretation of the liquid quality in the dynamic approach used in the current study may be subject to the articulatory realisations of vowels. This could be true of V_1; The preliminary acoustic analysis suggested that the second formant frequencies for V_1 seem to differ by 500 Hz between the two participant populations, meaning that the tongue retraction effect could not only be due to the articulatory realisations of English liquids but also the carry-over coarticulatory effects from V_1.

5. CONCLUSIONS

The study analysed dynamic tongue movements in L2 English liquids produced by L1 Japanese speakers. Dynamic analyses of the principal components show a smaller tongue dorsum movement for English liquids produced by L1 Japanese speakers compared to L1 English speakers. Future research could analyse liquids in other vowel environments and the intergestural timing to better generalise this assumed dorsal coarticulatory effect. The participants' perceptual identification accuracy data will also provide further theoretical insights into the nature of L2 speech learning.

6. ACKNOWLEDGEMENTS

Thank you to all the participants for their time and efforts. I thank Dr Claire Nance and Dr Sam Kirkham for their comments and support. Prof. Noriko Nakanishi, Prof. Yuri Nishio, and Dr Bronwen Evans helped me with data collection. The research is financially supported by Graduate Scholarship for Degree-Seeking Students by Japan Student Services Organization and the 2022 Research Grant by the Murata Science Foundation.

7. REFERENCES

[1] J. E. Flege, K. Aoyama, and O.-S. Bohn, "The Revised Speech Learning Model (SLM-r) Applied," in *Second Language Speech Learning*, 1st ed., R. Wayland, Ed. Cambridge University Press, Feb. 2021, pp. 84–118.

[2] D. Recasens, "A cross-language acoustic study of initial and final allophones of /l/," *Speech Communication*, vol. 54, no. 3, pp. 368–383, 2012.

[3] M. Proctor, R. Walker, C. Smith, T. Szalay, L. Goldstein, and S. Narayanan, "Articulatory characterization of English liquid-final rimes," *Journal of Phonetics*, vol. 77, p. 100921, 2019.

[4] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/," *Language*, vol. 92, no. 1, pp. 101–140, 2016.

[5] H. King and E. Ferragne, "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics*, vol. 80, p. 100978, 2020.

[6] B. Gick, P. Bacsfalvi, B. M. Bernhardt, S. Oh, S. Stolar, and I. Wilson, "A motor differentiation model for liquid substitutions in children's speech," in *Proceedings of the Meeting Acoustics*, vol. 1, Salt Lake City, Utah, 2008, p. 060003.

[7] S. Harper, L. Goldstein, and S. S. Narayanan, "L2 Acquisition and Production of the English Rhotic Pharyngeal Gesture," in *Interspeech 2016*. ISCA, Sep. 2016, pp. 208–212.

[8] G. N. Zimmermann, P. Price, and T. Ayusawa, "The production of English /r/ and /l/ by two Japanese speakers differing in experience with English," *Journal of Phonetics*, vol. 12, no. 3, pp. 187–193, 1984.

[9] J. Moore, J. Shaw, S. Kawahara, and T. Arai, "Articulation strategies for English liquids used by Japanese speakers," *Acoustical Science and Technology*, vol. 39, no. 2, pp. 75–83, 2018.

[10] T. J. Vance, *The Sounds of Japanese*. Cambridge University Press, 2008.

[11] N. Yamane, P. Howson, and W. Po-Chun (Grace), "An ultrasound examination of taps in Japanese," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Aug. 2015.

[12] M. Proctor, "Towards a gestural characterization of liquids: Evidence from Spanish and Russian," *Laboratory Phonology*, vol. 2, no. 2, pp. 451–485, 2011.

[13] M. Morimoto, "Geminated Liquids in Japanese: A Production Study," Ph.D. dissertation, University of California Santa Cruz, Mar. 2020.

[14] J. E. Flege, "The intelligibility of English vowels spoken by British and Dutch talkers," in *Studies in Speech Pathology and Clinical Linguistics*, R. D. Kent, Ed. Amsterdam: John Benjamins Publishing Company, 1992, vol. 1, pp. 157–232.

[15] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, O.-S. Bohn and M. J. Munro, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 13–34.

[16] T. Arai, "On Why Japanese /r/ Sounds are Difficult for Children to Acquire," in *Interspeech 2013*, Lyon, France, 2013, pp. 2445–2449.

[17] L. Spreafico, M. Pucher, and A. Matosova, "UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1517–1520.

[18] J. Scobbie, E. Lawson, S. Cowen, J. Cleland, and A. Wrench, "A common co-ordinate system for mid-sagittal articulatory measurement," *QMU CASL Working Papers*, vol. 20, pp. 1–4, 2011.

[19] Articulate Instruments, "Articulate Assistant Advanced version 220," Articulate Instruments, Edinburgh, 2022.

[20] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[21] A. Wrench and J. Balch-Tomes, "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[22] J. Ying, J. A. Shaw, C. Kroos, and C. T. Best, "Relations Between Acoustic and Articulatory Measurements of /l/," in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Sydney, Dec. 2012, pp. 109–112.

[23] S. Kirkham, C. Nance, B. Littlewood, K. Lightfoot, and E. Groarke, "Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 784–794, 2019.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 498–502.

[25] C. Nance and S. Kirkham, "Phonetic typology and articulatory constraints: The realisation of secondary articulations in Scottish Gaelic rhotics," *Language*, pp. 419–460, 2022.

[26] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2022.

[27] R. H. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, 2008.

[28] G. Schwartz and K. Kaźmierski, "Vowel dynamics in the acquisition of L2 English – an acoustic study of L1 Polish learners," *Language Acquisition*, vol. 27, no. 3, pp. 227–254, Jul. 2020.

IS THE EFFECT OF L2 VOCABULARY SIZE ON LEXICAL ENCODING MODULATED BY L1-L2 SIMILARITY?

Miquel Llompart[13], Brian Rocca[2], Isabelle Darcy[2]

[1]Universitat Pompeu Fabra, Spain, [2]Indiana University, USA, [3]Friedrich Alexander University, Germany
miguel.llompart@upf.edu, brocca@iu.edu, idarcy@indiana.edu

ABSTRACT

Previous research suggests that second language (L2) vocabulary size is a key predictor of how robust the phonological encoding of challenging L2 sounds into lexical representations is. Here we test whether this is also the case when the native language (L1) and the L2 are orthographically and lexically dissimilar. Korean learners of English completed an auditory lexical decision task (LDT) with phonological substitutions, a categorization task assessing their perception of the English /ɛ/-/æ/ distinction, and a vocabulary test. Results showed that vocabulary scores predicted /ɛ/-/æ/ nonword (e.g., m[ɛ]tch) rejection rates in the LDT task and that this effect was larger for participants with better perception skills for the target contrast. These results further support the role of L2 vocabulary size as a major determiner of phonological improvements in the representation of L2 words and speak against an account merely based on the mediating role of L1-L2 orthographic and lexical similarity.

Keywords: L2 learning, speech perception, lexical encoding, vocabulary size, L1-L2 similarity

1. INTRODUCTION

Learning a second language (L2) requires familiarizing oneself with a phonological inventory containing sounds and sound contrasts that are not part of the native language (L1) while simultaneously being exposed to vast amounts of L2 words. Therefore, learners need to acquire the ability to perceptually identify and produce the L2 categories that are not part of the L1 inventory. At the same time, they must create lexical representations for many newly-learned words and store them in long-term memory.

While there is much research on perceptual improvements in L2 learning, less is known about the development of L2 lexical representations and about how perception and word learning interact at different stages in the acquisition process. A recurrent finding of recent studies has been that, even in cases in which a robust perceptual identification of challenging L2 categories is in place, learners still struggle during spoken word recognition if recognition is dependent on such sounds [1, 2]. This is thought to be due to a combination of increased processing demands in word recognition tasks, unreliable acoustics-to-lexicon mappings and phonologically imprecise lexical representations [3, 4].

Regarding the phonological form of lexical representations, it has been suggested that improvements in perceptual ability are crucial and necessary for an accurate encoding of L2 sounds at the lexical level (henceforth *lexical encoding*), yet not sufficient on their own [5]. Crucially, recent studies have additionally shown that, at least for proficient learners, the precision of lexical encoding is connected to their lexical knowledge (operationalized as vocabulary size), above and beyond sheer perceptual ability [5, 6]. In [5], the lexical encoding of English /ɛ/ and /æ/ by German learners of English was assessed, showing that, while the ability to perceptually identify the vowels in a phonetic task predicted lexical encoding for learners of intermediate proficiency, it was vocabulary size that did so for the most advanced learners in the sample. Similarly, in [6], vocabulary size was the main predictor of the lexical encoding of the challenging Spanish segments /r/, /ɾ/ and /d/ for English speakers of intermediate-to-high proficiency in Spanish.

These findings have been crucial to highlight the need for a deeper understanding of the interplay between phonetic, phonological and lexical knowledge in L2 speech learning. In particular, more research is needed to determine the extent to which the relationship between vocabulary and lexical encoding holds across different combinations of L1s and L2s. This is especially relevant because the findings of previous studies were obtained for languages with a common script and a large overlap in their vocabularies (i.e., German-English and English-Spanish). Consequently, it cannot be ruled out that the effect of L2 vocabulary size is actually directly mediated to some extent by the L1 vocabulary and/or by the phonological and orthographic similarities between the L1 and L2 vocabularies.

Considering this, in the present study we test L1-Korean intermediate to advanced learners of English on their perception and lexical encoding of the /ɛ/-/æ/ contrast, which is known to also pose difficulties for this population (see [3]), and measure their

vocabulary size in an attempt to replicate the findings in [5]. Crucially, this means that now the L1 and L2 of these learners are not only expected to share fewer cognates and phonologically similar words than in [5] (see [7]), but also have different orthographic scripts. This constitutes a stringent test for the effects of L2 vocabulary previously observed.

For the German learners in [5], L1-L2 lexical similarity could have contributed to a tight coupling between L2 vocabulary and lexical encoding both indirectly through the degree of congruency in grapheme-to-phoneme correspondences (GPC) across languages with the same script and directly through the existence of many phonologically-similar lexical items. L1-congruent GPCs can be beneficial in the establishment of phonological contrasts in the lexicon [8]. In a similar way, cognates can highlight the existence of particular contrasts if a similar phonological contrast can already be inferred from the cognates' L1 form.

For instance, for /ɛ/ and /æ/, which in English are most often orthographically represented as <e> and <a>, the GPC of German for these same symbols would point towards the existence of a phonological contrast. In German <e> and <a> are mostly linked to the phonological categories /e/ or /ɛ/, and /a/, respectively. These similarities in GPC would be particularly obvious, and possibly also particularly useful for learning, in phonologically similar words such as cognates and near cognates. For example, *panda* is written in the same way in English and German and, in both languages, the first vowel, albeit different in English (/æ/) and German (/a/), would sound different from that of the word *pencil* (/ɛ/).

Therefore, if one expects that L1-L2 lexical similarity mediates the effect of L2 vocabulary on lexical encoding, this should be much weaker, if not non-existent, for Korean learners of English. If these learners, whose L1 is orthographically represented through a different script and its lexical similarity to English is considerably lower than that of German [7], do indeed show an effect of vocabulary size on lexical encoding, this will be taken as crucial evidence for a major role of L2 lexical knowledge in the process of establishing phonologically-robust L2 lexical representations.

2. METHODS

2.1. Participants

Thirty-five Korean learners of English (7 males) were recruited to participate in the experiment. To be included, participants had to be native speakers of Korean as well as intermediate or advanced L2 learners of English. To meet the proficiency criterion, participants had to have taken at least one course at a university or college that used English as the primary language of communication. Table 1 provides the participants' mean age, age of onset, years spent in an English-speaking country, current English usage at home and at university/work, and self-reported proficiency for English comprehension and speaking.

Measure	Mean (SD)
Age	29.4 (5.4)
Age of onset (speaking)	8.8 (5.1)
Years in English-speaking country	5.4 (4.9)
Current L2 use at home	2.3 (1.7)
Current L2 use at university/work	4.7 (1.6)
English comprehension	4.0 (0.8)
Spoken English	3.9 (0.8)

Table 1: Learner group characteristics. Note: L2 use ratings were given on a 1-6 scale (1=no English, 6=only English); proficiency ratings were given on a 1-5 scale (1=no ability, 5=perfect).

2.2. Materials and procedure

The three experimental tasks in this study were programmed in PsychoPy 3 [9], and participants were tested individually via the experimental platform Pavlovia. All participants wore headphones, used a keyboard, and met with one of the researchers remotely on Zoom during the experiment.

2.2.1. Lexical decision task

This task assessed the lexical encoding of English words containing /ɛ/ and /æ/. Participants were auditorily presented with real words and nonwords with phonological substitutions. The materials were 328 English words. 64 of these words were mono- or disyllabic words that contained the critical /ɛ/-/æ/ contrast—32 for each vowel. The rest of the words targeted other contrasts that can be considered fillers for the purposes of this study. Half of the /ɛ/-/æ/ items were real words and half nonwords. Nonwords were created by switching the two target sounds in each contrast. For example, an /æ/ real word "match" was transformed into the /æ/ nonword "m[ɛ]tch". The stimuli were recorded by a native speaker of American English. First and second formant values for /ɛ/-/æ/ items are provided in Table 2.

	Mean F1 (SD)	Mean F2 (SD)
/æ/ nonword	881 (54)	1947 (100)
/æ/ real word	1060 (69)	1854 (87)
/ɛ/ nonword	1024 (48)	1845 (64)
/ɛ/ real word	842 (50)	1958 (101)

Table 2: Vowel formants for /ɛ/-/æ/ real words and nonwords.

On each trial, participants heard an auditory stimulus (i.e., a word or a nonword) and indicated whether it was a real word or not by pressing 1 or 0 on the keyboard, such that nonword responses were always provided with the dominant hand.

2.2.2. Perceptual categorization task

A two-alternative forced-choice categorization task (2AFC) on the steps of a *bet-bat* continuum was used to assess the sharpness of participants' perceptual categories for /ɛ/ and /æ/. The words "bet" and "bat" were recorded by the same speaker who recorded the words for the LDT. These natural productions were then used to create a 21-step continuum in Matlab using the STRAIGHT morphing algorithm [10].

On each trial, participants saw a picture representing "bet" on the left side of the computer screen and a picture representing "bat" on the right and heard one of the steps of the continuum. They were asked to press 1 or 0 to indicate whether they heard "bet" or "bat", respectively. The 21 steps from the continuum were presented 10 times (210 trials) in a pseudorandomized order.

2.2.3. Vocabulary test

The vocabulary component of Shipley-2 [11] was administered to assess participant's English vocabulary. This is a multiple-choice test with 40 items of differing lexical frequencies. A word is presented in capitals and four response options are provided. Participants are asked to select the response that they believe is a synonym of the capitalized word. The items were presented in the fixed order of the test.

3. RESULTS

Before entering any analyses, data from 5 participants were excluded, leaving a total of 30 participants. One participant was excluded because their accuracy on the LDT control items was more than 2.5 SDs below the group mean and two more were excluded because part of their data did not save correctly. Two other participants were excluded because their categorization function in the perception task was the opposite from what was expected (i.e., had clearly negative slopes), suggesting that they had reversed the response keys. LDT items that were responded to with a mean accuracy 2.5 SDs below the mean were also removed from the analysis for all participants. Finally, items which individual participants identified as unfamiliar in a subsequent questionnaire were also excluded on an individual basis.

For the LDT, the average accuracy for /ɛ/ and /æ/ real words was 95% (SD=22%) and 97% (SD=18%), respectively, while the average score for /ɛ/ and /æ/ nonwords was 28% (SD=45%) and 14% (SD=35%), respectively. Figure 1 shows the mean accuracy by participant by condition.

Perception was measured by calculating the steepness of the /ɛ/-/æ/ categorization curve in the 2AFC task. Following [5], individual slopes were calculated by submitting the categorization data to a generalized linear mixed-effects regression model (GLMM) with a logistic linking function (lme4 package) with Response (coded as 0 and 1) as the categorical dependent variable, an intercept term, and a random slope for Continuum step over Participants. The slope coefficient for each participant was extracted from the model. This coefficient quantifies the increase in log-odds of a "bat" response as a function of an increase of one unit for continuum step. Hence, the higher the slope coefficient, the steeper the slope of the categorization function is. The average slope on the perception task was 0.46 (SD=0.26) with a range of -0.1 - 1.08.

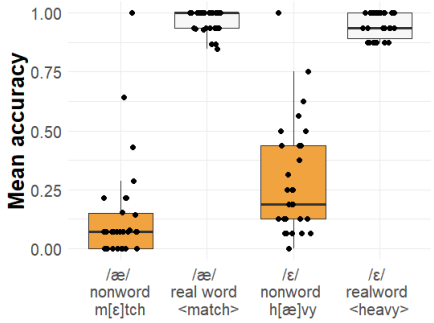


Figure 1: Mean accuracy by participant in the LDT as a function of lexical status and vowel.

For the vocabulary test, the number of correct responses out of 40 was calculated for each participant. The average percentage correct was 72.5% (SD=0.14) with a range of 47.5-92.5%. Figure 2 shows the performance for each participant in the perception and vocabulary tasks.

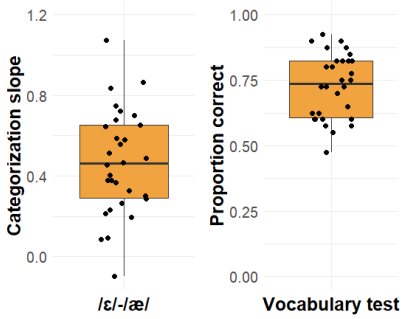


Figure 2: /ɛ/-/æ/ categorization slopes (left) and vocabulary test scores (right) by participant.

Because of the clear ceiling effects with real words (see Figure 1), statistical analyses on LDT data were limited to /ɛ/-/æ/ nonword trials. These data were submitted to a GLMM with a logistic linking

function.[i] The categorical dependent variable was Response (0=incorrect, 1=correct), and the independent variables were Perception score, Vocabulary size, Vowel, and the interaction between Perception score and Vocabulary size. The variable Vowel was contrast-coded with /æ/ as -0.5 and /ɛ/ as 0.5 and Perception score and Vocabulary size were centered and scaled. The random-effects structure included random intercepts for Participants and Items. A random slope for Vowel over Items was not included because it did not improve the model's fit. The results of the model are provided in Table 3.

Predictor	*b*	std error	*z*	*p*
Intercept	-2.13	0.28	-7.68	<.005
Vowel	1.24	0.38	3.30	<.005
Perception	0.00	0.28	0.02	0.99
Vocabulary	0.97	0.24	4.14	<.005
Perception x Vocabulary	0.56	0.25	2.27	< .05

Table 3. Results of the GLMM on the effects of Vowel, Perception, Vocabulary, and the Perception x Vocabulary interaction on LDT nonword rejection accuracy.

The model revealed significant effects of Vowel and Vocabulary as well as a significant Perception x Vocabulary interaction. Hence, participants were significantly more accurate in rejecting /ɛ/ nonwords than /æ/ nonwords and, crucially, having a larger vocabulary size led to higher accuracy in the LDT. Interestingly, the effect of vocabulary was modulated by an additive interaction with perception such that the effect of vocabulary size was larger for participants with better perception abilities. This interaction is illustrated in Figure 3.

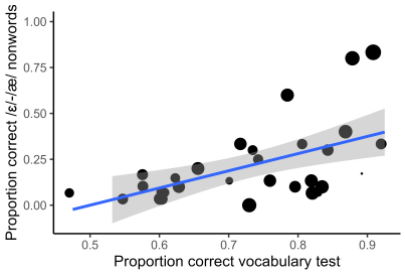


Figure 3: Scatterplot of the relationship between vocabulary size and /ɛ/-/æ/ nonword rejection in the LDT. Dot size conveys perceptual accuracy, where a larger dot means a steeper categorization slope.

4. DISCUSSION

In this paper, we aimed at a conceptual replication of [5] with an L2 learner group whose native language (i.e., Korean) is orthographically and lexically less similar to English than German. The main question we asked is whether an effect of L2 vocabulary size on the lexical encoding of two sounds that constitute a difficult L2 contrast could also be found for this new population. Results showed that, indeed, scores in an English vocabulary test predicted Korean learners' performance in a lexical decision task assessing their lexical encoding of English /ɛ/ and /æ/. In addition, this effect was qualified by an additive interaction with perception, indicating that the effect of vocabulary size was larger for learners with better perceptual abilities for the target contrast.

The across-the-board effect of vocabulary size therefore replicates the findings in [5] and [6] and suggests that it is not only when L1 and L2 are highly similar in terms of the orthographic and phonological forms of their words that knowledge of L2 vocabulary plays a major role in the development of phonologically-robust L2 lexical representations. In fact, a coarse comparison by means of Pearson product-moment correlations between vocabulary scores and LDT nonword rejection scores reveals an *r* of .55 for the Korean learners of English here and one of .39 for the whole sample of native German speakers in [5]. This indicates that a lower L1-L2 lexical similarity does not entail a looser relationship between vocabulary and lexical encoding. If anything, the coefficients appear to suggest the opposite. Hence, our results for Korean learners indirectly provide evidence against the possibility that an effect of L2 vocabulary is just a by-product of how existing vocabulary in the native language mediates the acquisition of L2 lexical items.

The second key finding of this study is the significant interaction between vocabulary and perception (see Figure 3). This interaction points towards the idea that a larger vocabulary size in the L2 is particularly beneficial when accompanied by sufficient ability in perceptual categorization for the target contrast. These results fit nicely with the proposal in [5] that learners need to reach a threshold of perceptual acuity (what could perhaps be referred to as "*good enough perception*") before they can leverage their L2 lexical knowledge and undergo substantial improvements in their lexical encoding.

In sum, the present study showed that L2 vocabulary predicts how learners phonologically encode challenging L2 categories into lexical representations even when the L1 and L2 of the learners are much more dissimilar than in previous studies on the same topic. This starkly challenges the concern that this relationship may simply surface because it is overwhelmingly mediated by the relationship between L1 and L2 words in terms of orthography and phonology, and strengthens the notion that L2 vocabulary is indeed a key predictor of essential improvements in the phonological form of L2 words containing challenging L2 sounds.

5. REFERENCES

[1] Darcy, I., Daidone, D., Kojima, C. 2013. Asymmetric lexical access and fuzzy lexical representations in second language learners. *The mental lexicon* 8(3), 372-420.

[2] Llompart, M., Reinisch, E. 2021. Lexical representations can rapidly be updated in the early stages of second-language word learning. *Journal of Phonetics* 88, 101080.

[3] Barrios, S., Hayes-Harb, R. 2021. L2 processing of words containing English /æ/-/ɛ/ and /l/-/ɹ/ contrasts, and the uses and limits of the auditory lexical decision task for understanding the locus of difficulty. *Frontiers in Communication* 6, 144.

[4] Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., Scott, J. H. 2012. Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research* 28(1), 5-40.

[5] Llompart, M. 2021. Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and Cognition* 24(3), 481-496.

[6] Daidone, D., Darcy, I. 2021. Vocabulary size is a key factor in predicting second-language lexical encoding accuracy. *Frontiers in psychology* 12, 2769.

[7] "Similar languages," ezglot.com. https://www.ezglot.com/most-similar-languages.php (accessed December 15, 2022)

[8] Escudero, P., Hayes-Harb, R., Mitterer, H. 2008. Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36(2), 345-360.

[9] Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. 2019. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

[10] Kawahara, H., Masuda-Katsuse, I., De Cheveigne, A. 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication* 27(3-4), 187-207.

[11] Shipley W. C., Gruber, C.P., Martin, T.A., Klein, A.M. 2009. *Shipley-2 manual.* Western Psychological Services.

[i] The dataset and the code to reproduce the main analyses of this paper are available at https://osf.io/e637k/ (Open Science Framework)

REVISING PARAMETERS FOR PREDICTING L2 SPEECH FLUENCY AND PROFICIENCY

Heini Kallio & Mikko Kuronen

University of Jyväskylä
heini.h.kallio@jyu.fi, mikko.j.kuronen@jyu.fi

ABSTRACT

The aim of the study was to investigate whether integrating parameters based on pause location improve the prediction of fluency and proficiency in L2 Finnish monologic speech. To answer this question, multiple linear regression models were fitted using two data sets containing L2 Finnish speech and expert assessments of fluency and oral proficiency. Separate models were derived for fluency and proficiency using combined data as well as the two separate data sets. The comparison of the models indicate that pause-by-location parameters can improve the prediction of L2 fluency and proficiency, but the relevant parameters and their significance in the regression models depend on the speech data. Parameters with low incidence work only in longer speech samples, while parameters with frequent occurrence can be used even in shorter samples. The results have implications for improving automatic assessment of L2 speech especially in low-resource languages.

Keywords: L2 speech fluency, L2 proficiency, pause locations, automatic assessment

1. INTRODUCTION

Speech fluency parameters have been investigated as predictors of fluency and proficiency for several second or foreign languages (L2). However, research on L2 fluency in low-resource languages is limited. The current study extends the research to a scarcely studied L2, Finnish, and explores new fluency parameters for automatic assessment of fluency and proficiency in L2 Finnish.

Speech fluency is one of the strongest indicators of L2 proficiency and often used as a diagnostic measure in language assessment. Previous studies that have attempted to quantify L2 speech fluency have measured features related to speed, pausing, and repair [1]. Especially the ones of speed and pausing have proven to be good predictors of L2 fluency. Recent findings of [2] support the importance of speed measures as fluency indicators in spontaneous L2 Finnish speech, but they also found the durations of so-called composite breaks and filled pauses as well as creaky voice to affect human ratings of L2 fluency and proficiency. However, there are conflicting results on the significance of some fluency parameters, such as pause rate and pause duration [3, 4, 5]. Reasons for the differing results can be looked for in variation in the number and type of parameters used, but also in the analyzed data.

The occurrence of disfluencies in speech can depend on various factors. For example, [4] found that the predictive power of different acoustic measures varied not only between beginner and intermediate level speakers, but also between speech styles: filled pauses seem to be more typical for spontaneous speech than read speech. The use of silent and filled pauses can also be language specific [6, 7]. Although studies on speech fluency in both L1 and L2 Finnish are limited, the findings of [8, 9] indicate that Finnish speakers may use silent pauses considerably longer in duration than speakers of other more frequently studied languages (English, French, Italian, German, and Spanish as studied in [7]). It is therefore important to study speech fluency in different languages in order to find out, which fluency features are global and which ones language-dependent.

The role of pause location to L2 fluency have been investigated only in the recent years. These studies suggest that pause location is an important aspect in perceived L2 fluency [10, 11], and pause locations have also been integrated into automatic assessment of L2 fluency [12]. However, only silent pauses have been taken into account with respect to their location. Moreover, pause locations have been generally defined in terms of grammatical clauses, which may limit the usefulness of the measure: in our recent study we found that silent and filled pauses within grammatical *phrases* were better indicators of (dis)fluency than pauses within clauses in L2 Finnish [13]. Pauses after incomplete words also significantly reduced the perceived fluency of L2 Finnish speech [13]. Based on these results we now investigate whether pause-

by-location parameters can improve the prediction of perceived fluency and proficiency in L2 Finnish (RQ1). We also study the possible differences in predictive fluency parameters between two similar data sets (RQ2). This study is among the first ones to integrate pause-by-location parameters into predicting L2 proficiency, and the first one to acknowledge the location of filled pauses in predicting L2 fluency and proficiency.

2. DATA AND METHODS

2.1. Speech data and human assessments

In this study we used two speech and assessment data sets provided by the DigiTala project [14]: one consisting of semi-spontaneous narratives produced by 147 adult (aged 18–61) and another from 53 younger learners of Finnish (high school students aged 15–21), both with related expert assessments of language skills. The data sets are described in more detail in [2] and [15]. Both data sets consist of responses to narrative tasks, but the instructive response times and speakers' proficiency and fluency distributions differed between the data sets. The adult L2 Finnish speakers were instructed to speak for 1.5 minutes, and the mean duration of their speech samples was 83.6 seconds. The young L2 speakers were instructed to speak for one minute on the given topic, and the mean duration of the samples was 42.6 seconds. In the adult data, ratings are skewed towards lower proficiency and fluency levels, while in the young data the ratings are skewed towards higher proficiency and fluency levels. The two data sets were analyzed separately but also combined to get a more balanced distribution in terms of speaker proficiency and fluency.

Expert ratings were collected for the adult and young speech data using the same experimental settings, assessment criteria, and rating scales [2, 15]. For the current study, assessments of overall proficiency level (holistic scale 0–6) and fluency (analytic scale 1–4) were used. In [2] and [15], the inter-rater reliability was tested and found sufficient for research purposes.

2.2. Fluency parameters

The speech samples were annotated to word and utterance levels using previously prepared transcriptions and the online alignment tool WebMAUS [16]. Annotations were revised and disfluencies marked manually in Praat [17] by a phonetic expert and checked by another to avoid mistakes and differing interpretations of disfluencies.

We computed 44 fluency parameters from extracted annotation intervals using R [18]. 21 general fluency parameters included the following: speech and articulation rate; rate, ratio, and mean duration of silent pauses (SP) 50–250 ms and > 250 ms; rate, ratio, and mean duration of filled pauses (FP) as well as corrections or repetitions; mean length of run in words; rate and mean duration of utterances; rate and mean duration of utterance breaks; pause ratio; disfluency ratio. An utterance was defined as a continuous speech run, which is separated from the next by a break of 250 ms or longer. The utterance break, in turn, could contain silent or filled pauses, hesitations, corrections, or repetitions. Disfluency ratio refers to the relative proportion of all pauses and disfluencies in the response.

23 parameters were based on pause and disfluency location and included SPs and FPs further defined by their syntactic location: between clauses (BC), within clauses (WC), between phrases (BP), and within phrases (WP). The duration threshold of pauses-by-location was set to 250 ms, following [19, 5, 3, 20, 2]. We defined clause in Finnish as a constituent that links a predicate to a subject or object [21]. Phrase, in turn, was defined as a word or group of words that act together as a grammatical unit but do not necessarily include a predicate [21]. In addition, pauses within words, or between an incomplete and a corrected word (WW), were measured. All fluency parameters were further z-scored to 0 for statistical analysis.

2.3. Statistical analysis

The effects of the fluency parameters to human ratings were studied using multiple linear regression models with average ratings of fluency or proficiency as the dependent variable and fluency parameters as predictors. The simplest models were derived with a stepwise feature selection method using the stepAIC algorithm (implemented in the R package MASS [22]). To study the contribution of the pause-by-location parameters in predicting L2 fluency and proficiency ratings we fitted separate model with only general fluency parameters (referred to as Model 1) and with both general and pause-by-location fluency parameters (Model 2). The two models were compared with respect to their predictive power as well as with likelihood ratio tests (RQ1). To study the role of data in predicting L2 fluency and proficiency, the models were also fitted separately for the adult and young data sets (RQ2). For brevity, we do not report the full models in this paper but focus on the most

relevant results in terms of our RQs.

3. RESULTS

Using the combined data, pause-by-location parameters improved the predictive power of the fluency model from 59% (Model 1) to 60% (Model 2). The likelihood ratio tests showed that the fits of Model 1 and Model 2 did not differ significantly (Chi-Squared test-statistic = 9.6, p-value > 0.05). For proficiency, integrating pause-by-location parameters to the model improved the predictive power of the model from 59% (Model 1) to 62% (Model 2). The fits of Model 1 and Model 2 differed significantly (Chi-Squared test-statistic = 24.2, p-value < 0.01).

Of the 44 fluency parameters, 12 were included in the final model predicting fluency ratings. Five of these were pause-by-location parameters: rate of filled pauses within phrases (WP-FP rate, $p < 0.05$), ratio of silent pauses between an incomplete and a corrected word (WW-SP ratio, $p < 0.05$), mean duration of silent pauses within phrases (mean WP-SP, $p < 0.1$), mean duration of silent pauses between phrases, (mean BP-SP, $p > 0.1$), and ratio of filled pauses within clauses (WC-FP ratio, $p > 0.1$). 15 parameters were included in the final model predicting proficiency ratings, and seven of these were pause-by-location parameters: WC-FP ratio ($p < 0.01$), WP-FP ratio ($p < 0.01$), WW-SP ratio ($p < 0.01$), mean duration of filled pauses within clauses (mean WC-FP, $p < 0.01$), mean WP-SP ($p < 0.01$), ratio of filled pauses between phrases (BP-FP ratio, $p < 0.1$) and rate of silent pauses within phrases (WP-SP rate, $p < 0.1$). Speech rate was expectedly the most significant predictor of both fluency and proficiency ratings ($p < 0.001$).

The pause-by-location parameters had differing improvements depending on the data: the predictive power of the fluency model improved from 68% to 81% for the young data set (Model 1 – Model 2 Chi-Squared test-statistic = 54.7, p-value < 0.001), while the predictive power of the fluency model for the adult data set improved from 58% to 60% (Model 1 – Model Chi-Squared test-statistic = 18.6, p-value < 0.05). For proficiency, the predictive power improved from 50% to 72% for the young data set (Model 1 – Model 2 Chi-Squared test-statistic = 49.7, p-value < 0.001) and from 61% to 66% for the adult data set (Model 1 – Model 2 Chi-Squared test-statistic = 31.2, p-value < 0.001).

The final fluency model for the adult data set included 22 fluency parameters, 10 of which were based on pause locations. The final fluency model for the young data set also included 22 parameters, 12 of which were pause-by-location parameters. The final proficiency model for the adult data set included 18 fluency parameters, 8 of which were based on pause locations. The final proficiency model for the young data set included 19 parameters, 8 of which were based on pause locations.

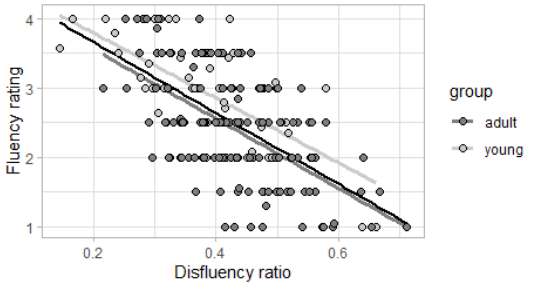


Figure 1: The relationship between disfluency ratio and fluency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

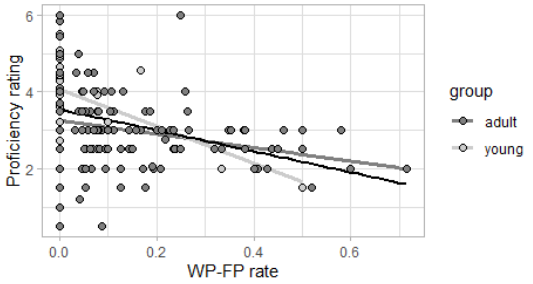


Figure 2: The relationship between WP-FP rate and proficiency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

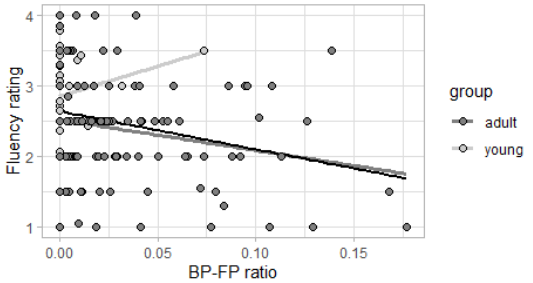


Figure 3: The relationship between BP-FP ratio and fluency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

Disfluency ratio proved to be a significant predictor of fluency in both data sets but was not included in the models predicting proficiency. Figure 1 illustrates the linear relationship between

disfluency ratio and fluency ratings of the two data sets. Of the pause-by-location parameters, WP-FP rate and WW-SP ratio had significant negative effects on the fluency as well as proficiency ratings. In the model for the combined data the effect was stronger for proficiency. Figure 2 shows the linear relationship between proficiency ratings and WP-FP rate for both data sets.

The significant parameters (excluding speed related parameters) and the predictive power of the models differed notably between the two data sets. For example, repair-related parameters (rate, ratio, and mean duration of corrections and repetitions) were significant only for the fluency model of the young data set, while mean length of run contributed only in the prediction models of the adult data set. Some parameters common for the models of adult and young data sets provided opposite effects: for example, ratio of silent pauses >250 ms had significant positive effects for the fluency and proficiency in the adult data set but significant negative effects in the young data set. BP-FP ratio, in turn, had a significant negative effect on the fluency ratings in the adult data set but significant positive effect in the young data set. The reason for the conflicting results might be due to rare occurrence of the parameterized features in the data. Figure 3 illustrates the relationship between BP-FP ratio and proficiency ratings of the two data sets.

4. DISCUSSION

The main aim of this study was to investigate whether integrating pause-by-location parameters improve models predicting fluency and proficiency of L2 Finnish monologic speech (RQ1). We also investigated the differences in prediction models between two similar data sets (RQ2).

Integrating pause-by-location parameters improved the predictive power of the fluency model only one percentage point, and the revised Model 2 (including pause-by-location parameters) did not differ significantly from Model 1 (including only general fluency parameters). The predictive power of the proficiency model, in turn, improved three percentage points and the revised Model 2 differed significantly from Model 1. Our results indicate that pause-by-location parameters are more strongly related to the perceived proficiency than fluency in L2 Finnish. This is an important finding, because pausing patterns have previously been studied in relation to perceived fluency rather than proficiency. It is also noteworthy that filled pauses within phrases contributed significantly to the prediction of both fluency and proficiency ratings. In previous studies, only silent pause locations have been under investigation. Our results encourage to study the role of filled pauses in L2 speech in more detail.

Of the general fluency measures, speech rate was expectedly the most significant predictor of both fluency and proficiency, regardless of the data. The less common parameter, disfluency ratio, proved significant in fluency prediction, which suggests that fluency models could be simplified by combining silent and filled pauses as well as corrections and repetitions into one parameter. However, the role of pauses and disfluencies seems to depend on how they occur in the data at hand. Although we studied two seemingly similar data sets (spontaneous L2 Finnish narratives), the final models of these data sets differ in the predictive parameters. This indicates that there are differences in the occurrence of disfluencies between the two data sets. These differences may also affect the results of the models for the combined data, as some parameters gained opposing effects in the two data sets. One reason for the differences can be looked for in the duration of the responses: the mean duration of the young L2 Finnish learners' responses was 42.6 seconds, while the mean duration of the adult L2 Finnish learners' duration was 83.6 seconds. The probability of disfluencies, especially filled pauses and repairs, increases with the duration of the response. The differences between the data sets manifest clearly in the occurrence of these disfluencies defined by location: for example, mean frequency of filled pauses within phrases was 0.13 for the young L2 learners and 2.4 for the adult L2 learners. Since the young data set is smaller (N = 53) and the speech samples are shorter than in the adult data set, some disfluencies may occur with only few speakers, which can distort the effect of these disfluencies, as shown in Figure 3. Therefore the statistical results for parameters with sparse occurrence must be interpreted with caution. With regards to pause locations, research on the use of silent and filled pauses in native Finnish speech is needed.

The current study contributed to the research of fluency parameters for automatic assessment of L2 fluency and proficiency in a low-resourced language, Finnish. The results support further investigation of pause-by-location parameters in predicting oral proficiency in L2 Finnish. Data-specific features may be an issue for training automatic assessment systems: using the same parameters regardless of the speech type and language can leave the prediction models weak or even biased.

5. ACKNOWLEDGEMENTS

The authors would like to thank the following people: Rosa Suviranta and Liisa Koivusalo for their help in annotating and analyzing the speech data, all the expert assessors, the Finnish National Certificates of Language Proficiency for their help in enabling the use of part of the speech data and recruiting expert assessors, and Aalto University's DigiTala team for their help in collecting human ratings. The DigiTala project is funded by the Academy of Finland and the research consortium includes University of Helsinki (grant number 322619), Aalto University (grant number 322625), and University of Jyväskylä (grant number 322965).

6. REFERENCES

[1] P. Tavakoli and P. Skehan, "Strategic planning, task structure and performance testing," in *Planning and Task Performance in a Second Language*, R. Ellis, Ed. John Benjamins, 2005, pp. 239–273.

[2] H. Kallio, R. Suviranta, M. Kuronen, and A. von Zansen, "Creaky voice and utterance fluency measures in predicting perceived fluency and oral proficiency of spontaneous L2 Finnish," *Speech Prosody: Proceedings of the 11th International Conference on Speech Prosody*, pp. 777–781, 2022.

[3] H. R. Bosker, A.-F. Pinget, H. Quené, T. Sanders, and N. H. De Jong, "What makes speech sound fluent? The contributions of pauses, speed and repairs," *Language Testing*, vol. 30, no. 2, pp. 159–175, 2013.

[4] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[5] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[6] D. Duez, "Silent and non-silent pauses in three speech styles," *Language and Speech*, vol. 25, no. 1, pp. 11–28, 1982.

[7] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech prosody 2002, international conference*, 2002.

[8] M. Toivola, M. Lennes, and E. Aho, "Speech rate and pauses in non-native Finnish," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[9] N. Penttilä, A.-M. Korpijaakko-Huuhka, and R. D. Kent, "Tavallista sujumattomuutta: Aikuisten puheen sujuvuuden kvantitatiivinen analyysi. [Normal disfluency: Quantitative analysis of adults' speech fluency.]," *Puhe ja kieli*, no. 3, pp. 153–173, 2018.

[10] J. Kahng, "The effect of pause location on perceived fluency," *Applied Psycholinguistics*, vol. 39, no. 3, pp. 569–591, 2018.

[11] K. Saito, M. Ilkan, V. Magne, M. N. Tran, and S. Suzuki, "Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency," *Applied psycholinguistics*, vol. 39, no. 3, pp. 593–617, 2018.

[12] C.-N. Hsieh, K. Zechner, and X. Xi, "Features measuring fluency and pronunciation," in *Automated Speaking Assessment*, K. Zechner and K. Evanini, Eds. Routledge, 2019, pp. 101–122.

[13] H. Kallio, M. Kuronen, and L. Koivusalo, "The role of pause location in perceived fluency and proficiency in L2 Finnish," *Proceedings of ISAPh 2022, 4th International Symposium on Applied Phonetics*, pp. 22–27, 2022.

[14] M. Kautonen and A. von Zansen, "DigiTala research project: Automatic speech recognition in assessing L2 speaking," *Kieli, koulutus ja yhteiskunta*, vol. 11, no. 4, 2020.

[15] L. Koivusalo, "Phonetic Fluency in Finnish as a Second Language: Acoustic Analysis of High School Students' Spontaneous Speech," *Master's Thesis, University of Helsinki*, 2022.

[16] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer. [Computer program]. Version 6.0. 19," *Online: http://www. praat. org*, 2016.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[19] R. Towell, R. Hawkins, and N. Bazergui, "The development of fluency in advanced learners of French," *Applied Linguistics*, vol. 17, no. 1, pp. 84–119, 1996.

[20] M. Kautonen and M. Kuronen, "Kvantitativa perspektiv på L2-tal på olika färdighetsnivåer [Quantitative perspectives on L2 speech on different skill levels]," *Folkmålsstudier*, vol. 59, pp. 11–40, 2021.

[21] A. Hakulinen, M. Vilkuna, R. Korhonen, V. Koivisto, T. R. Heinonen, and I. Alho, *VISK: Iso suomen kielioppi. Helsinki: Verkkoversio [The Great Grammar of Finnish. Online version]*. Suomalaisen Kirjallisuuden Seura, 2004, http://scripta.kotus.fi/visk.

[22] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, "Package 'mass'," *Cran r*, vol. 538, pp. 113–120, 2013.

L2-Mandarin regional accent variability facilitates Mandarin-naïve English listeners' learning of Mandarin tone-words

Yanping Li[1], Michael D. Tyler[2], Denis Burnham[1], Catherine T. Best[1]

[1]The MARCS Institute, Western Sydney University, Australia [2]Independent researcher
yanping.li@westernsydney.edu.au, tylerspeechscience@gmail.com
denis.burnham@westernsydney.edu.au, c.best@westernsydney.edu.au

ABSTRACT

English learners have difficulties using tones in Mandarin words [1]. As high phonetic variability facilitates second language (L2) word learning [2], we investigated how L2-Mandarin accent variability in minimal-tone-contrast word training affects Mandarin-naïve English participants' tone-word learning. 48 English learners completed 6 training sessions on 16 Mandarin pseudowords spoken by 12 Beijing talkers or four talkers each from Beijing, Yantai, and Guangzhou. Participants took word identification tests after each session, and generalisation tests to novel talkers with familiar and unfamiliar accents after session 6. Growth curve analysis on word identification accuracy and response times across training sessions revealed no significant differences between training conditions. However, linear mixed effects modelling of the generalization results showed that while both groups succeeded on novel talkers with familiar and unfamiliar accents, the multiple accent group identified words significantly faster than the Beijing accent group, indicating that L2-Mandarin accent variability facilitates English learners' tone-word learning.

Keywords: High variability word training; L2-Mandarin regional accents; tone-word identification, talker and accent generalization

1. INTRODUCTION

Lexical tones are used as phonological contrasts at the segmental level in tone languages [3] such as Mandarin Chinese. There are four Mandarin lexical tones, i.e., T1 with a high-level contour, T2 high-rising, T3 low-dipping, and T4 high-falling [4]. Imposing the four tones on the identical consonant-vowel [CV] syllable /*ma*/ yields four different words in Mandarin Chinese: level = *mother*, rising = *hemp*, dipping = *horse*, and falling = *curse*. Non-tone language listeners (e.g., English and French) are able to rely on their native intonation patterns to discriminate dissimilar tones, e.g., dipping vs. level, dipping vs. falling, and rising vs. falling [5, 6] and to improve categorization between similar tones, e.g., level vs. rising, level vs. falling, and rising vs. dipping [7]–[9]. However, perceiving Mandarin tone contours as intonation patterns does not help English listeners to overcome persistent difficulties with lexical tones in learning and using words, especially in larger utterances in communicative contexts [1, 10, 11].

Word training is a possible solution to establish L2 phonological representations and connect them to lexical meanings [e.g., 12]. Phonological constancy, which maintains word recognition across lexically irrelevant talker and accent variation [13], appears important to such training. Talker variability facilitates word learning in non-tone languages, presumably by providing greater variability in phonetic realizations of words, thereby yielding more robust L2 word representations. For example, native (L1) English learners trained on Spanish words with high talker variability had better accuracy and response time for identifying the newly learned words than those trained with low talker variability [2]. High talker variability word training has also been applied to non-tone language learners during tone language acquisition, to direct their attention to lexical tones [10, 14, 15]. For example, naïve English participants exposed to high talker variability in training on minimal-tone-contrast Mandarin words performed better on word identification than those exposed to low talker variability [14].

Studies with high talker variability have rarely if ever examined the effects of *accent* variability on minimal-tone-contrast word training; as talkers have nearly always been selected from the same region, usually an accepted standard variety of the target language. The present study examines accent variability effects on Mandarin tone word learning. Standard Mandarin (Mandarin, henceforth) developed historically from the dialect spoken in Beijing, China. Talker(s) from regions outside of Beijing acquire Mandarin as an L2, given that their L1 Chinese dialects are mutually unintelligible languages [16]. Thus, there is variability in L2-accented Mandarin tones, triggered by similarities and dissimilarities between their native dialect tone systems and Mandarin. For example, Yantai, Guangzhou, and Shanghai dialects' tone systems differ from each other and from Mandarin, resulting in L2-Mandarin tone pronunciations that deviate from native Mandarin in the slopes for rising (T2) and falling (T4) tones, and in the depth of the dip in dipping (T3) tone [17] (see also [18] for acoustic details on L2-accented Mandarin tones).

This study adapted the high talker variability training paradigm in [10] to manipulate accent variability during naïve English participants' learning of minimal-tone-contrast Mandarin words, by using stimuli produced by multiple talkers from either (i) Beijing (single accent condition: talker-only variability) or (ii) Beijing, Yantai and Guangzhou (multiple accent condition: talker and accent variability). Learners in both training conditions should be able to learn words and to generalize recognition of the newly learned words to novel talkers, given the high talker variability in both training conditions. In addition, the same words spoken in an unfamiliar accent (Shanghai) should be more difficult for both conditions to identify than those with the familiar (trained) Beijing accent, resulting in slower response times to the unfamiliar accent talker than to the familiar Beijing accent talker. However, we expected the multiple accent training condition learners to generalize words to novel talkers in both accents faster and/or more accurately than those in the single accent condition.

2. EXPERIMENT

2.1. Method

2.1.1. Participants

Mandarin-naïve Australian English participants (n = 48) were recruited online (see [19] for details), and randomly assigned to the single (n = 24, M_{age} = 24.5 years, SD = 5.8, 14 females) vs. multiple accent (n = 24, M_{age} = 25.5 years, SD = 5.1, 15 females) training conditions.

2.1.2. Stimuli

There were 16 Mandarin pseudowords based on four CV syllables (/ba/, /di/, /du/, /gu/). Imposing the four tones on each syllable yielded 16 Mandarin real words. But artificial meanings were assigned for this study, using 16 frequent English words from [20], which were represented with 16 grey-scaled drawings selected from the Multilingual Picture database [21].

Training stimuli were produced by female talkers, either 12 from Beijing (single accent condition) or four each from Beijing, Yantai, and Guangzhou (multiple accent condition). Both training groups thus heard 12 talkers (i.e., high talker variability) in total. There were four tokens for each word, resulting in 768 stimuli (12 talkers × 16 words × 4 tokens) in each training condition. Generalization stimuli (n = 128, 2 talkers × 16 words × 4 tokens) were the newly trained words but produced by a novel female talker (19.0 years old) with a familiar Beijing accent, and by a novel female talker (24.0 years old) with an unfamiliar Shanghai accent.

2.1.3. Procedure

This study was run remotely in a quiet room with E-Prime Go. The experimenter conducted it via ZOOM meetings with the participants to ensure data quality. Participants in both training conditions completed six training sessions on the 16 words in six consecutive days with a picture-to-word paradigm [10], followed immediately by two generalization tests.

There were four blocked talkers of the same accent in each training session (45 min), yielding three sets of 12 talkers, which were counterbalanced across participants; those three sessions were repeated a second time for a total of six sessions. Correspondingly, the four talkers each of the Beijing, Yantai, and Guangzhou Mandarin accents in the multiple accent condition were blocked by session. Accent order of the sessions was counterbalanced across participants.

There were 256 randomized trials (4 talkers × 4 syllables × 4 tones × 4 tokens) in each session with 16 words blocked by talker (n = 4) to optimise word learning [22]. For the same reason, talker blocks were further subdivided by syllable (i.e., by minimal-tone-contrast word set), yielding 16 randomized trials per word set (4 words × 4 tokens). Talker orders and syllable orders within talker were randomized across participants. Immediately after training on words for a given syllable by talker, participants completed a quiz with feedback on the four words in that syllable set, to optimize the learning on the just-trained words [23]. This word training + quiz cycle was repeated for the other three syllables produced by the same talker for all 16 target words by that talker. Participants then learned the 16 words again produced by each of the other three talkers with the same cycle.

On each word training trial, a drawing was displayed, and the paired word was played out; participants were instructed to remember as many audio-visual pairs as possible. After training on the four minimal-tone-contrast words of a syllable set by a given talker, they completed the corresponding 16-trial quiz. In each quiz trial, they heard one target word and had to choose the correct item in the display of 4 drawings. Correct answers were followed by a green tick on the screen, incorrect answers with a red cross mark. If the participant gave no response within 5.5 s, a reminder appeared asking them to respond more quickly. Following wrong or no answers, the correct drawing and audio word were presented as feedback.

To track learning progress across the six sessions, participants completed word identification tests at the end of each session. Each post-session test had 64 randomized trials (16 words × 4 trained talkers × 1 token; token per talker per word was counterbalanced across participants), and the procedure was identical to the quizzes except that no feedback was given and

there was no time-out, i.e., participants could take as long as needed to respond. Generalization tests to novel talkers with a familiar Beijing and an unfamiliar Shanghai accent were run, in that order, with the same procedure, after session 6. They included 128 randomized trials (2 tests × 16 words × 4 tokens), blocked by talker, and took ~ 7 min total to complete.

2.2. Results

2.2.1. Identification tests across training sessions

Indexed by their identification tests in each training session, English learners in both training conditions improved more than 30% from the first (M_{single} = 44.54%, SD = 22.50; $M_{multiple}$ = 55.33%, SD = 23.33) to the sixth session (M_{single} = 85.75%, SD = 13.52; $M_{multiple}$ = 86.96%, SD = 9.36), with greater improvement during the first three sessions than the last three, in which they asymptoted at 75-80% accuracy.

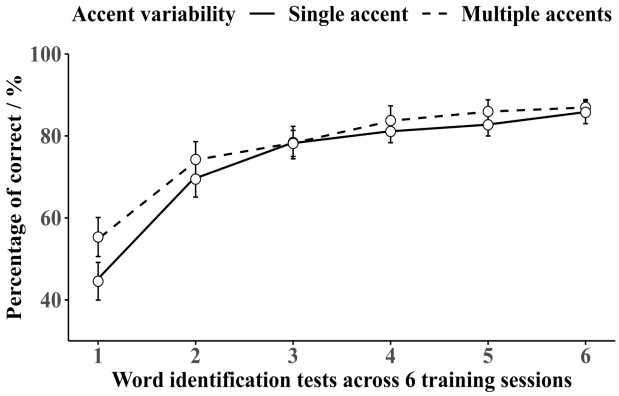


Figure 1: Mean percent correct word identifications across training session tests in single and multiple accent conditions. Error bars = ± 1 standard error of the mean.

Response times (RTs) for correct responses are shown in Figure 2. Values beyond ± 2.5 standard deviations from the mean for each participant were excluded – 449 occurrences, 3.22% of trials across participants, as recommended by [24]. The final RT data (n = 13475) were log transformed for data visualization and statistical analyses. Higher log values indicate slower responses. RTs in both training conditions decreased more sharply in the first three training sessions than in the last three (asymptote ~7.8).

Both accuracy and log-transformed RTs across the six training sessions were subjected to growth curve analysis [25], which captures changes in longitudinal data. Given that the curves in Figures 1 and 2 each had a single inflection, growth curve data were modelled using the *lmer* function from package *lmerTest* in *R* version 4.2.1 [26] including up to second-order orthogonal polynomials: intercept (mean) and linear (+/- slope) and quadratic terms (degree of inflection, i. e., the depth of peak or valley). Two mixed-effects models (for accuracy and RTs) were built with training condition as a fixed effect, and participants as a random effect. While learners in the multiple accent training condition appear somewhat more accurate and faster than those in the single accent condition (see Figures 1 & 2), neither groups' accuracy nor RT differed significantly in mean, linear, or quadratic trends, suggesting that talker variability during training is already optimally effective, with neither a deficit nor further benefit due to accent variability.

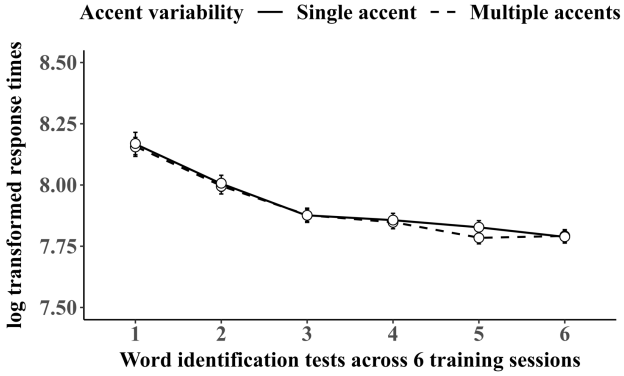


Figure 2: Log transformed RTs for correctly identified words across training sessions in single and multiple accent conditions. Error bars = ± 1 standard error of the mean.

2.2.2. Generalization tests

Both training groups also retained accuracy in generalization tests to novel talkers with familiar Beijing (M_{single} = 85.13%, SD = 13.57; $M_{multiple}$ = 91.17%, SD = 7.21) and unfamiliar Shanghai accents (M_{single} = 79.58%, SD = 14.60; $M_{multiple}$ = 86.04%, SD = 10.33). However, the multiple accent group slightly outperformed the single accent group on the generalization tests (see Figure 3).

One linear mixed-effects model was built on each participant's accuracy (%) with training conditions and generalization tests as fixed effects, and participants as random effects, using the *lmer* function from package *lme4*. We used the Kenward-Roger degrees of freedom approximation to calculate p values for the fixed-effects factors and the *Anova* function from package *car* to calculate F. Pairwise comparisons were conducted with *lsmeans* in *R* when necessary.

The main effect of generalization tests was significant, $F(1, 47)$ = 21.87, p < .001, and that for training conditions was marginally significant, $F(1, 46)$ = 3.81, p = 0.05. No interactions were significant; pairwise comparisons showed that participants in both training conditions correctly identified more words with the familiar Beijing accent than those with the unfamiliar Shanghai accent, i.e., Single: Estimate = 5.54, SE = 1.63, $t(46)$ = 3.40, p = 0.007; Multiple: Estimate = 5.13, SE = 1.63, $t(46)$ = 3.15, p = 0.01.

Figure 4 displays log transformed RTs for correct

identifications in the two generalization tests by participants from the single (familiar: M = 7.75, SD = 0.46, vs. unfamiliar: M = 7.75, SD = 0.47) and multiple accent (familiar: M = 7.70, SD = 0.44, vs. unfamiliar: M = 7.70, SD = 0.47) conditions.

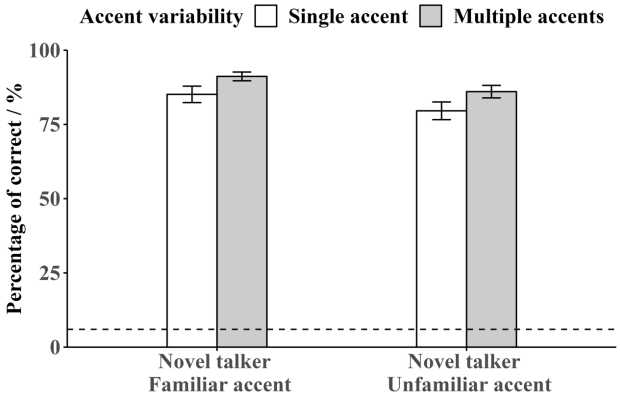


Figure 3: Mean percent correct responses in word generalization tests to novel talkers of familiar Beijing or unfamiliar Shanghai accents. Error bars = ± 1 standard error of the mean. Dashed line = chance level (1/16 × 100 = 6%).

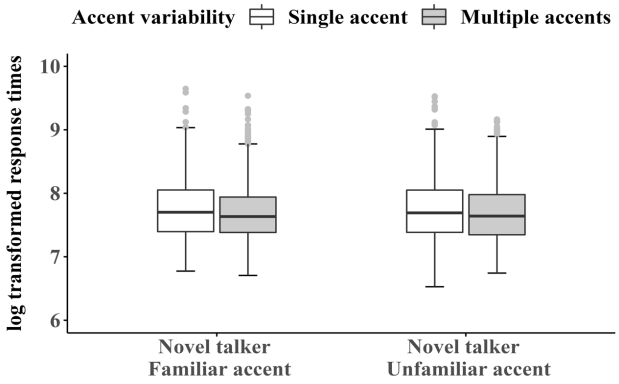


Figure 4: Log transformed RTs for correctly identified words in generalization tests for familiar and unfamiliar accents. Error bars = 95% confidence interval, top and bottom of the box are 25% and 75% percentiles; line inside box is 50% percentile (median). Grey points are outliers.

Another linear mixed-effects model was built for RTs with training conditions and generalization tests as fixed-effects, and participants and target words as random effects. Calculation of p and F values was the same as in the accuracy model. While there was no significant main effect of generalization tests, the main effect of training conditions, $F(1, 5046) = 14.29$, $p < .001$, and the interaction of training conditions × generalization tests, $F(3, 5042) = 4.82$, $p = 0.002$, were both significant. To tease the interaction apart, pairwise comparisons were conducted to assess differences in RTs between the two training conditions for each generalization test. While there were no significant differences between the two generalization tests in each training condition, the participants in the multiple accent condition identified the words produced by novel talkers with Beijing, Estimate = -0.05, SE = 0.02, $t(5042)$ = -2.73, p = 0.03, and Shanghai accents, Estimate = -0.05, SE = 0.02, $t(5043)$ = -2.63, p = 0.04, significantly faster than those in the single accent condition.

3. DISCUSSION

This study investigated effects of L2-Mandarin accent variability on naïve English participants' learning of Mandarin minimal-tone-contrast pseudowords. The single and multiple accent condition learners were comparable in identifying tone-words during training. However, participants in the multiple accent condition identified words in the two generalization tests more quickly than those in the single accent training condition, indicating that accent variability during training on minimal-tone-contrast words established more robust phonological representations of the four Mandarin tones, which facilitated their generalization of the learned words to novel talkers and a novel accent.

For each training condition, word identification accuracy in the generalization tests was significantly lower for the novel talker with the unfamiliar Shanghai accent than for the novel talker with the familiar Beijing accent, suggesting that both groups experienced difficulties with the unfamiliar accent. That inference would thus predict that RTs for words with the unfamiliar accent should be greater than for a novel talker with a familiar accent. However, RTs did not differ in either training condition for identification of words produced by novel talkers with familiar versus unfamiliar accents. Given that identifying words spoken by a novel talker with an unfamiliar accent was less accurate than for words with a familiar accent, it is possible that the learners spent more time identifying words spoken by the novel talker with the familiar accent, thereby achieving higher accuracy at a cost of RTs for that accent. This suggests that achieving phonological constancy across talker variability, and perhaps especially across accent variability, may require additional cognitive effort, i.e., pose additional cognitive load, in tone-word identification (e.g., [27]).

4. ACKNOWLEDGEMENT

Supported by a Western Sydney University (WSU) – China Scholarship Council (CSC) joint scholarship, Candidature Research Funds from the MARCS Institute, WSU, and partly by the Australian Linguistic Society Research Grants, which were all awarded to the first author. We are grateful for online recruitment assistance of Careers at Australian universities, technical support from Johnson Chen and Chris Wang at MARCS, and above all to the participants.

5. REFERENCES

[1] E. Pelzl, E. F. Lau, T. Guo, and R. DeKeyser, "Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioural and event-related potentials experiments," *Stud. Second Lang. Acquis.*, vol. 43, no. 2, pp. 268–296, 2021, doi: 10.1017/S027226312000039X.

[2] J. Barcroft and M. S. Sommers, "Effects of acoustic variability on second language vocabulary learning," *Stud. Second Lang. Acquis.*, vol. 27, no. 03, 2005, doi: 10.1017/S0272263105050175.

[3] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[4] Chao Y.-R., "A system of tone letters," *Maître Phon.*, vol. 45, pp. 24–27, 1930.

[5] C. K. So and C. T. Best, "Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties," *Poznan Stud. Contemp. Linguist.*, vol. 47, no. 1, pp. 133–145, 2011, doi: 10.2478/psicl-2011-0011.

[6] P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *J. Phon.*, vol. 32, pp. 395–421, 2004, doi: 10.1016/S0095-4470(03)00016-0.

[7] S. Wiener, K. Ito, and S. R. Speer, "Effects of multitalker input and instructional method on the dimension-based statistical learning of syllable-tone combinatons," *Stud. Second Lang. Acquis.*, vol. 43, no. 1, pp. 155–180, 2021, doi: 10.1017/S0272263120000418.

[8] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, "Training American listeners to perceive Mandarin tones," *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3649–3658, 1999, doi: 10.1121/1.428217.

[9] G. Shen and K. Froud, "Categorical perception of lexical tones by English learners of Mandarin Chinese," *J. Acoust. Soc. Am.*, vol. 140, no. 6, pp. 4396–4403, 2016, doi: 10.1121/1.4971765.

[10] P. C. M. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Appl. Psycholinguist.*, vol. 28, no. 4, pp. 565–585, 2007, doi: 10.1017/S0142716407070312.

[11] Jie Xi, Hongkai Xu, Ying Zhu, Linjun Zhang, Hua Shu, and Yang Zhang, "Categorical perception of Chinese lexical tones by late second language learners with high proficiency: Behavioral and electrophysiological measures," *J. Speech Lang. Hear. Res.*, vol. 64, no. 12, pp. 4695–4704, 2021, doi: 10.1044/2021_JSLHR-20-00210.

[12] D. J. Bolger, M. Balass, E. Landen, and C. A. Perfetti, "Context variation and definitions in learning the meanings of words: An instance-based learning approach," *Discourse Process.*, vol. 45, no. 2, pp. 122–159, 2008, doi: 10.1080/01638530701792826.

[13] C. T. Best, "Devil or angel in the details? : Perceiving phonetic variation as information about phonological structure," in *Phonetics-Phonology Interface: Representations and Methodologies*, J. Romero and M. Riera, Eds., 2015, pp. 3–31.

[14] M. Sadakata and J. M. McQueen, "Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training," *Front. Psychol.*, vol. 5, no. NOV, pp. 1–15, 2014, doi: 10.3389/fpsyg.2014.01318.

[15] H. Dong, M. Clayards, H. Brown, and E. Wonnacott, "The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones," *PeerJ*, vol. 7, p. e7191, 2019, doi: 10.7717/peerj.7191.

[16] C. N. Li and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, 1989.

[17] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "Regionally accented Mandarin lexical tones," *J. Acoust. Soc. Am.*, vol. 148, no. 4, pp. 2474–2475, 2020, doi: 10.1121/1.5146856.

[18] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "Tone variations in regionally accented Mandarin," in *Interspeech 2020*, ISCA, 2020, pp. 4158–4162. doi: 10.21437/Interspeech.2020-1235.

[19] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "L2-Mandarin regional accent variability during lexical tone word training facilitates naive English listeners' tone categorization and discrimination," in *Proceedings of SST 2022*, R. Billington, Ed., 2022, pp. 156–160.

[20] W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-UK: A new and improved word frequency database for British English," *Q. J. Exp. Psychol.*, vol. 67, no. 6, pp. 1176–1190, 2014, doi: 10.1080/17470218.2013.850521.

[21] J. A. Duñabeitia *et al.*, "The multilingual picture database," *Sci. Data*, vol. 9, no. 1, Art. no. 1, 2022, doi: 10.1038/s41597-022-01552-7.

[22] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *J. Acoust. Soc. Am.*, pp. 461–472, 2011, doi: 10.1121/1.3593366.

[23] M. A. Pyc and K. A. Rawson, "Why testing improves memory: mediator effectiveness hypothesis," *Science*, vol. 330, no. 6002, p. 335, 2010, doi: 10.1126/science.1191465.

[24] R. K. Chan and J. H. Leung, "Why are lexical tones difficult to learn?: Insights from the incidental learning of tone-segment connections," *Stud. Second Lang. Acquis.*, vol. 42, no. 1, pp. 33–59, 2020, doi: 10.1017/S0272263119000482.

[25] D. Mirman, *Growth Curve Analysis and Visualization Using R*. in The R Series. Boca Raton, Florida: CRC Press, 2014.

[26] R Core Team, "R: The R Project for Statistical Computing." 2022. [Online]. Available: https://www.r-project.org/

[27] T. Bent and R. Frush Holt, "The influence of talker and foreign-accent variability on spoken word identification," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1677–1686, 2013, doi: 10.1121/1.4776212.

HOW JAPANESE L2 LISTENERS PERCEIVE ENGLISH SILENT-CENTRE FRONT VOWELS

Shinichi Tokuma

Chuo University
tokuma@tamacc.chuo-u.ac.jp

ABSTRACT

This study examines the perception of three American English front vowels, /iː/, /ɪ/, /e/, by Japanese listeners in 'Silent-Centre' (SC) paradigm. A perceptual experiment was conducted to investigate whether English L2 vowel perception by Japanese listeners is affected by the SC paradigm, using the set-up utilised by previous research. The results demonstrate that: (1) the Japanese sound system, in particular the importance of duration on vowels and consonantal geminates as a perceptual cue, plays a significant role in the perception of English SC /iː/ and /ɪ/; and (2) certainly for some vowels, duration-neutralised tokens create some perceptual confusions, but the influence of Japanese sound system outweighs the perceptual effect of neutralised duration.

Keywords: L2 vowel perception, Silent-Centre vowels

1. INTRODUCTION

Anybody with good knowledge of acoustic phonetics knows that the F1/F2 formant patterns of the vowels flanked by consonants show great variability in comparison with the values when the vowels are produced in isolation. Listeners somehow cope with this variability and find no difficulty in perception.

This phenomenon, vowel formant undershoot [1] and its perpetual overshoot [2], had been a central issue in early days. However, it was discovered later that vowel inherent duration and vowel dynamic spectral change carries more weight in perception. One kind of strong evidence comes from what is called a 'silent-centre' (SC) vowel paradigm, suggested by Strange and her colleagues (e.g. [3]), where listeners are asked to identify the vowel in /CVC/ stimuli whose entire central portion was silenced while durational information is kept intact. Their results demonstrate that "vowels in SC syllables were identified remarkably well, despite the total absence of vocalic nuclei. In fact, when only a single speaker's utterances were represented, errors on SC syllables were no higher than on unmodified control stimuli." ([3] p.2186)

This perceptual independence on static vowel information proposes one intriguing question: do L2 listeners show the similar perceptual behaviour when they listen to SC vowels, because it is well known that L2 listeners sometimes rely on different acoustic cues for perception (e.g. perception of English /l/-/ɹ/ by Japanese listeners in [4])?

This L2 perception of SC vowels has been studied previously. For example, Rogers & Lopez [5] investigated the perception of American English vowels in SC paradigm by L1 and Spanish-speaking L2 listeners. Their results demonstrate that the L2 listeners who started learning English after the age of 18 showed the identification rate consistently lower than the L1 listeners, while the L2 listeners with the age of onset of immersion of 12 years or earlier identified the whole-word and 40ms duration-preserved syllables as accurately as the L1 listeners, but identified the silent-centre syllables significantly with less accuracy overall. Also, Schwartz and his colleagues ([6] [7]) studied how some SC vowels of British English were perceived by L1 and L2 Polish listeners, the latter of whom were at different English proficiency levels. They discovered that depending on proficiency levels, the listeners adopted a more dynamic approach to vowel identification and showed higher accuracy rates on the SC vowels: they heed more attention to dynamic formant cues, or vowel inherent spectral change.

However, no studies were found which deal with perception of SC English vowels by Japanese L2 listeners, and it would be of particular interest to investigate the perception of English vowels which cause a great confusion among Japanese listeners. such as /iː/-/ɪ/, and compare the results with those of /e/ that causes less confusion (see [8] [9]). As in other SC studies, the effect of neutralised durational difference is also examined: /iː/-/ɪ/, in this study.

Hence, this study investigates the perception of English SC vowels /iː/, /ɪ/, /e/ by Japanese listeners. The experimental set-up used in [5] is followed because their L1 and L2 data can be compared with the data obtained in this study. The main objectives of this study are: (1) to compare the perceptual data with those obtained in the previous studies and see if there is any effect by Japanese sound system (2) to compare perceptual patterns, particularly in difficult

conditions like neutralised or /iː/-/ɪ/ vowel pairs that are likely to cause confusions.

2. EXPERIMENT

2.1. Materials

Six American English /bVb/ and /dVd/ words were used as stimuli, where /V/ is /iː/, /ɪ/ or /e/: in the English writing system, they are written as 'Beeb', 'bib', 'bebb' (a nonsense word), 'deed', 'did' and 'dead'.

The vowels /iː/ and /ɪ/ were selected since it has been reported (e.g. [9]) that Japanese listeners of English show great perceptual confusion between these two vowels even in ordinary listening environments. The rest of the vowels, /e/, which shows less perceptual confusion among Japanese listeners, was selected for comparison.

Rogers & Lopez [5] explored three more American English vowels, but this study focuses on the vowel pair /iː/-/ɪ/ which demonstrates more significant confusions by Japanese listeners than other vowels. In contrast, although Rogers & Lopez [5], without no apparent reason, investigated only the /bVb/ consonantal frame, two consonants, /b/ and /d/, were examined in this paper since, as stated in [8], the perception of of /ɪ/ or /e/ by Japanese listeners is affected by the types of flanking consonants.

The target /CVC/s for editing were recorded as follows. A native speaker of General American (male in his 50s, with no speech impairment) was asked to read, at a comfortable speed, the target /CVC/s in the frame sentence used in [5], "I say _____ on the tape." The sentences were printed on a sheet, in the English orthography, described above. He was asked to read the sheet twice, and this session was repeated. Therefore, four tokens per each /CVC/ type were recorded.

The left-channel monaural recording was made in a quiet room with Sony Electret Condenser Microphone ECM-PCV80U attached to a Windows laptop PC, and the sound was transferred to the PC by Praat (ver. 6.2.10) [10], at a sampling frequency of 44100 Hz.

Out of four tokens for each /CVC/ types recorded as above, one optimal token for stimulus manipulation was selected based on the spectral analyses and the auditory impression. The target /CVC/s were excised from the frame sentences and they were manipulated, by Praat, to create SC vowels.

The procedures were as follows: first, the onset and offset of the vowel in the /CVC/ token were determined by waveform and spectrograph analysis. Then, four types of /CV/ and /VC/ transition, whose vowel duration is 10/20/30/40ms from the onset/offset, were created and the central parts of the vowel were silenced, by manipulating Intensity tier on Praat. After that, the edges of the initial /CV/ and final /VC/ (i.e. just before and after the central silence part) were smoothed by applying a 2ms linear filter ramp, as in [5], in order to minimise the auditory impression of clipping. Also, as a control group, the non-manipulated (i.e. no SC) /CVC/ tokens were presented to the listeners.

Furthermore, as in the previous studies, duration-neutralised (DN) tokens of /iː/ and /ɪ/ were also created and presented to participants since the importance of vowel intrinsic duration in SC paradigm has been demonstrated in the previous studies ([3] [5], amongst others), and in particular, it is demonstrated ([11], also amongst others) that the duration is a main factor that affects the substantial perceptual confusion in non-SC English /iː/ and /ɪ/ by Japanese listeners. Henceforth, for clarification, tokens without durational manipulation are called duration-preserved (DP).

The duration of DN tokens was calculated as follows. First, the mean duration was obtained across four tokens for each /CVC/ types. Then the mean of the two was calculated for the duration of DN tokens. The actual values are shown in Table 1 below.

	/biːb/	/bɪb/	/diːd/	/did/
Mean duration of 4 /CVC/ tokens (ms)	195.8	130.4	228.2	168.0
Duration of DN tokens (ms)	163.1		198.1	

Table 1: Mean duration of 4 /CVC/ tokens and the DN duration.

Manipulating Duration tier on Praat, the DN tokens were created by editing the central silent part, or the central steady part in the case of non-SC vowels, without changing the /CV/ and /VC/ transitions, as in the previous studies. Note that this process created DN tokens for each consonant and vowel type (i.e. one for /biːb/, another for /bɪb/) .

This process created 30 token types for DP (2 consonants x 3 vowels x 5 duration patterns of transition), and 20 for DN (2 consonants x 2 vowels x 5 duration patterns of transition). The edited tokens were put back in the original frame sentence and presented to the participants.

2.2. Participants

Twenty-four Japanese first-year and second-year students of Chuo University participated in the experiment. None of them had lived or studied abroad, and had a hearing impairment. They learnt English at

school for 6 years (first-years) or 7 years (second-years), and their mean TOEIC L&R score was 615. This, and informal evaluation by the author, established their English proficiency level around CEFR A2-B1. They received a small monetary compensation for their participation.

2.3. Procedure

The participants were tested in a quiet Language Laboratory room at Chuo University. They were seated individually in front of laptop PCs, and their task was to listen to the stimulus token through covered-ear headphones, and using a mouse, click the target SC word that appeared on the screen.

The locations of the words on the screen were randomised, and each stimulus token was played in a random order. The stimuli were played through covered-ear headphones at a sound level adjusted by them, and none of them reported that their attention had been diverted by extraneous noises or by the presence of other participants. The whole experiment process was controlled by Praat, utilising Experiment MFC objects.

Since the experiment had to be conducted during one 100-minute class, the twenty-four participants were randomly assigned to two groups of 12 participants, and one group listened to /bVb/ DP + DN, while the other to /dVd/ DP + DN. This arrangement reduced the number of stimulus presentation to 90 (15 token types x 6 target word position on the screen) + 40 (20 token types x 2 target position on the screen) =130 and minimalised the effect of reduced concentration. All the members of each group were tested at the same time.

The experiment was preceded by a task demonstration by the author designed to make the participants familiar with the experimental setting and the nature of the stimuli. Care was taken that the participants knew that they must not click a word on their screen before it was played. After that, the participants joined the trial session of three test stimuli, and the questions raised were answered at this stage. In the main session, short break was inserted after every 10 stimulus presentations. At the end of the experiment, the general purpose of experiment was explained to the participants but no individual feedback was given.

2.4. Results

Figures 1 and 2 are for DP results. The vertical axis represents the identification rate, while the horizontal axis is for the transition duration ('full' means there is no SC). In these Figures, /i:/ tokens show a steady decline in % correct of identification as the duration of transitions become shorter, and also there is a big drop in 10ms for /e/. These observations are in concordance with [5]. However, in these Figures, /ɪ/ shows an overall rise in % correct, and this rise was not observed in the previous studies. This pattern is replicated for DN results in Figures 3 and 4, where /i:/ tokens shows a steady decline in contrast with the rise of /ɪ/ tokens. This rise was also unreported in [5], where the performance of the participants shows no significant difference or fall.

Furthermore, there is a difference in the response patterns of /ɪ/ between DP and DN tokens. For DP, the rise occurs later in Figures 1 and 2: when the duration of the /CV/ and /VC/ transition is shorter, around 10ms or 20ms, while for DN, in Figures 3 and 4, the rise occurs earlier: when the duration of the transition is still 40ms. Furthermore, overall identification rate for 'full' (i.e. no SC) /ɪ/ is lower in DN than DP. (N.B. The full token has no SC but the duration has been manipulated.)

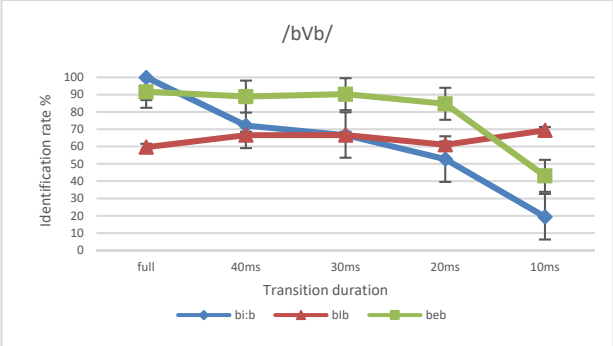


Figure 1: Results of DP /bVb/

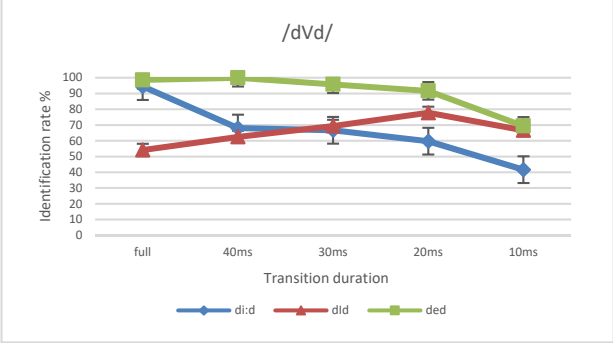


Figure 2: Results of DP /dVd/

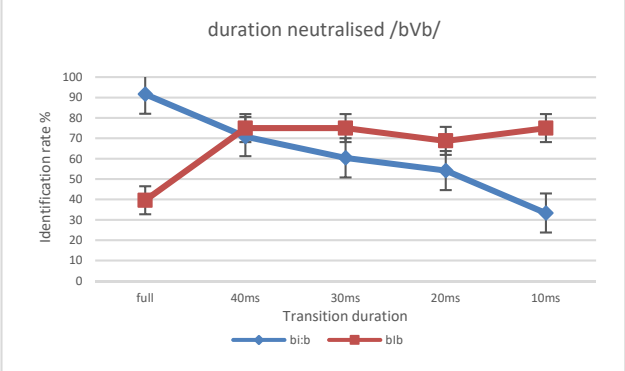


Figure 3: Results of DN /bVb/

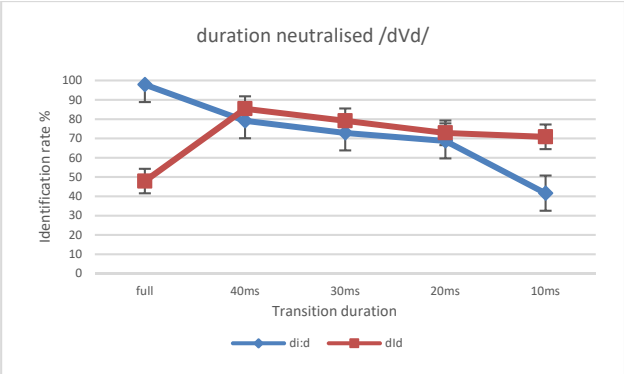


Figure 4: Results of DN /dVd/

3. DISCUSSION

Overall, as mentioned in Section 2.4, there are some concordances in the results between [5] and this study: the falling trend of /i:/ and /e/ in DP and DN conditions, and the sudden dip in 10ms for /e/ (See Figure 4 in [5]). However, the rise pattern of /ɪ/, which was not reported in their studies or others, as well as the difference of the point where /ɪ/ rises between DP and DN (or the higher % correct in full DN tokens), was observed in this study. This issue needs to be discussed in detail.

As mentioned before, Japanese listeners are heavily dependent on durational cues when they listen to /i:/-/ɪ/ difference, and they hardly pay any attention to the spectral cues. When the DN stimuli were created, to neutralise the durational difference, the nucleus duration of /ɪ/ was elongated (See Table 1), and consequently, this longer nucleus duration was likely to evoke a perceptual preference for longer /i:/, hence the lower % correct for /ɪ/.

Another explanation comes from the influence of Japanese sound system. In Standard Japanese, there is a phonological contrast based on consonantal gemination: *saka* 'slope' and *sakka* 'writer'. Many loan words from English /CVC/ containing /i:/ are written in Japanese with non-gemination and a long /i:/, hence 'sheep' having become *shiipu*, while those containing /ɪ/ with gemination and a short /i/, thus 'ship' having become *shippu*. It is known that the Japanese consonantal geminate gives an auditory impression of vowel clipping. Vance [12] illustrates how a Japanese vowel in an emphatic word, followed by a glottal stop (with auditory clipping impression) is spelt with a letter representing geminates.

Although the extra care was taken to minimise the clipping impression, the finally produced SC tokens were not free from it, and therefore they tend to be perceived as having consonantal geminates, providing an auditory bias towards /ɪ/. This influence of clipping impression also explains why there is a rise in /ɪ/ as the duration of the /CV/ and /VC/ transitions becomes shorter: the shorter transitions give a stronger impression of clipping because of the shorter vowel duration.

Moreover, the fact that the % correct is higher for DN tokens than DP ones is also explained by this process. If the duration of the English vowel /ɪ/ is longer and closer to that of /i:/, Japanese listeners, according to Morrison [11], will more likely to perceive it as /i:/, but the results were that they preferred /ɪ/ in DN, which has a longer duration.

The length of the silence is crucial for the perception of Japanese plosive geminates (e.g. [13]). When creating the DN stimuli, the duration of /ɪ/ was increased. This means that the SC interval between the transitions is longer in DN than in DP. The longer silence duration certainly evokes the perception of geminates, and this contributes to the perceptual preference of /ɪ/. Interestingly, the peak of % /ɪ/ in /dVd/, in Figure 3, falls in 20ms stimuli, while in Figure 5, it is in 40ms stimuli. The 20ms shift in peak could be explained by the durational difference between DN and DP: 197.1(DN)-168.0(DP)=20.1ms. However, this hypothesis requires further verification.

In the perception test, the data for the goodness-of-fit for each token was also taken but not yet analysed. The scope of the future research includes its analysis and the investigation of individual difference in performance in correlation of their English proficiency.

4. CONCLUSION

This study examined the perception of English SC vowels /i:/, /ɪ/, /e/ by Japanese listeners. The results demonstrate that: (1) the Japanese sound system, in particular the importance of duration on vowels and consonantal geminates as a perceptual cue, plays a significant role in English SC perception of /i:/ and /ɪ/; and (2) certainly for some vowels, DN tokens creates some perceptual confusions, but the influence of Japanese sound system outweighs the perceptual effect of neutralised duration.

5. ACKNOWLEDGEMENT

This research is partially funded by Chuo University Personal Research Grant.

6. REFERENCES

[1] Lindblom, B. 1963. Spectrographic study of vowel reduction. *J. Acoust Soc. Am.* 35, 1773–1781.

[2] Lindblom, B., Studdert-Kennedy, M. 1967. On the role of formant transitions in vowel recognition. *J. Acoust Soc. Am*. 42, 830–843.
[3] Strange, W. 1989. Dynamic specification of coarticulated vowels spoken in sentence context. *J. Acoust Soc. Am*. 85, 2135-2153.
[4] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87, B47–B57.
[5] Rogers, C., Lopez, A. 2008. Perception of silent-center syllables by native and non-native English speakers. *J. Acoust Soc. Am*. 124, 1278-1293.
[6] Schwartz, G., Aperliński, G., Jekiel, M., Malarski, K. 2016. Spectral Dynamics in L1 and L2 Vowel Perception. *Research in Language*. 14, 61-77.
[7] Schwartz, G., Dzierla, J. 2018. Polish listeners' perception of vowel inherent spectral change in L2 English. *Poznan Studies in Contemporary Linguistics.* 54. 307-332.
[8] Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S., Nishi, K. 2001. Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *J. Acoust Soc. Am*. 109. 1691-1704.
[9] Nozawa, T., Wayland, R. 2012. Effects of Consonantal Contexts on the Discrimination and Identification of American English Vowels by Native Speakers of Japanese. *Kotoba-no-Kagaku-Kenkyu.* *（ことばの科学研究）* 13. 19-40.
[10] Boersma, P., Weenink, D. Praat: doing phonetics by computer [computer programme]. Version 6.2.10, retrieved on 20 March 2022 from http://www.praat.org/
[11] Morrison, G. 2002. *Effects of L1 duration experience on Japanese and Spanish Listeners' perception of English high front vowels*. Unpublished MA dissertation. Dept. of Linguistics, Simon Fraser University.
[12] Vance, T. 2008. *The Sounds of Japanese.* Cambridge University Press.
[13] Kawahara, S. 2015. The phonetics of sokuon, or geminate obstruents. In: Kubozono, H. (ed). *Handbook of Japanese Phonetics and Phonology*. De Gruyter Mouton, 43-78.

NUCLEAR CONTOURS OF SPANISH ECHO QUESTIONS PRODUCED BY PROFICIENT CHINESE LEARNERS OF SPANISH: A DYNAMIC ANALYSIS

Peng Li [1], Xiaotong Xi [2]
[1] Center for Multilingualism in Society across the Lifespan (MultiLing), University of Oslo, Norway
[2] Department of Translation and Language Sciences, Universitat Pompeu Fabra, Spain
[1]*peng.li@iln.uio.no*, [2]*xiaotong.xi@upf.edu*

ABSTRACT

This study investigates Spanish nuclear contours produced by Chinese students and the influence of lexical stress position on these contours. In a Discourse Completion Task, 16 Chinese students and nine Spanish natives produced 450 Spanish yn- and wh- questions with various pragmatic functions (information seeking, clarification echo, and counter expectation echo). The nuclear words were stressed in initial, medial, or final positions. Dynamic analyses using Generalized Additive Mixed Models revealed that (a) clarification echo questions were the most challenging for Chinese students, who tended to produce a high boundary tone (H%) instead of the native pattern (L%), and (b) word-final stress was the least favorable position for producing L2 nuclear contours, especially for the clarification echo questions. These results provide direct evidence of cross-linguistic influence on prosody and support the L2 Intonation Learning Theory, emphasizing the importance of prosody in L2 teaching practice.

Keywords: Nuclear contour, dynamic analysis, L2 prosody acquisition, Spanish, Chinese students

1. INTRODUCTION

Compared to second language (L2) speech sound learning, relatively fewer studies focused on L2 speech prosody, such as intonation, due to its abstract and dynamic nature. Only a few models have tackled L2 prosody learning. For instance, the L2 Intonational Learning Theory (LILt) [1] predicts that the first language (L1) prosody is the key to predicting the learning outcome. Some empirical studies supporting LILt found that tone language speakers showed general advantages in learning L2 prosody [2], [3]. However, others suggest that the positive transfer is domain-specific and conditioned by individual differences [4], [5]. This study thus further explores L2 prosodic learning with Spanish prosody produced by Chinese learners and Spanish natives.

As an "intonation language", Spanish varies pitch at sentence level to shape intonation. By contrast, "tone languages", like Chinese, vary pitch on lexical level to distinguish words [6]. Therefore, Chinese speakers use multiple prosodic cues for intonational meanings [7], resulting in a more complex tone-intonation interaction than Spanish [8]. Previous research showed that Chinese students realized Spanish lexical stress as Tone 2 (rising tone) [9], manipulated pitch more than Spanish natives to contrast stress [10], and tended to replace the low nuclear accent (L) with high (H) or rising (L+H) tones on intonational level [8]. Therefore, the complex interplay between lexical stress and intonation should be considered in L2 prosody research. A stressed syllable may trigger unexpected high boundary tones or rising pitch accents for Chinese students.

Turning to specific sentence types, Spanish questions are said to be more challenging than statements for Chinese learners [11]. Recent research has investigated the Chinese students' production of Spanish *yn*- and *wh*- questions varied in pragmatic functions (i.e., information-seeking vs. confirmation-seeking) and disjunctive questions [8]. However, many other question types were left out in previous research, for example, echo questions. Compared to information-seeking (INF) questions, echo questions are biased towards the proposition *p* as conveying pragmatic meanings of either clarification (CLA, i.e., *Did you say p?*) or counter expectation (EXP, i.e., *Are you saying p?*) [12], and are thus more difficult than neutral questions. Within echo questions, CLA would be more difficult than EXP for Chinese learners for the following reasons. Spanish CLA bears a low boundary tone [12], but the underlying form of Chinese CLA is a *yn*-question with a high boundary tone [13]. By contrast, EXP shows a high boundary tone in both Spanish [12] and Chinese [14].

Therefore, this study selected six question types produced by Chinese students and Spanish natives from a large corpus: CLA-*yn*, EXP-*yn*, CLA-*wh*, EXP-*wh*, INF-*yn*, and INF-*wh*. Since nuclear configurations are key for intonation types [15] and the stress position of the sentence-final words largely affects the surface nuclear curves [8], we mainly focused on the nuclear contours while considering the stress position of the sentence-final nuclear words.

Furthermore, traditional analytic frameworks on prosody (e.g., ToBI systems for Spanish [16] and Chinese [17]) categorized intonation elements on phonological level [18]. However, when comparing

intonation curves produced by L2 learners and native speakers, a dynamic approach is necessary to estimate the exact locations of differences. To this end, we used Generalized Additive Mixed Models (GAMM) to compare the estimated intonation curves at group level (e.g., Chinese students vs. Spanish natives) and the traditional Sp_ToBI [16] to phonologically label the intonation curves.

We hypothesized that:

- Spanish CLA would be the most challenging sentence type for Chinese students since the low boundary tone is contrary to Chinese CLA.
- The stress position of nuclear words would affect the surface nuclear contours, with sentence-final stress being the least favorable since it would trigger unexpected high or rising boundary tones.

2. METHOD

2.1. Participants

We recruited 16 Chinese learners of Spanish (female = 12, aged 22-33 years) and 9 monolingual speakers of peninsular Spanish (female = 6, aged 18-28 years) from Spain. They signed consent forms to allow the researchers to process their data. The Chinese students started learning Spanish in their early adulthood (*M* = 19.3 years old, *SD* = 1.9). They had learned Spanish with formal instructions for 4.1 years (*SD* = 1.3) in China, had lived in Spanish-speaking countries for 5.4 years (*SD* = 3.3), and had studied various subjects instructed in Spanish for 2.4 years (*SD* =2.2) in Spain. They all had passed the DELE test (Diplomas of Spanish as a Foreign Language). Five of them were qualified as B2/advanced level and the rest (11), C1/high advanced level. Therefore, the Chinese students were proficient late adult learners with extensive exposure to the target L2.

2.2. Materials and procedure

The experiment was a Discourse Completion Task. We designed 54 scenarios to elicit Spanish utterances with different prosodic structures. The scenarios were organized in a PowerPoint presentation, with each slide containing one scenario followed by a prompt of the target sentence. The participants had to read each scenario carefully and orally read the prompts in Spanish. The participants' speech outcomes were recorded with a Zoom H4n Pro in a soundproof room. The current study targeted 18 scenarios (6 sentence types × 3 stress positions). That is, for each sentence type, we designed three scenarios, with each eliciting a sentence ending with one of the three target words, varied in stress positions: ***vino*** 'wine' (word-initial), *Marina* (word-medial), and *Milán* 'Milan' (word-final). In total, we analyzed 450 utterances (25 participants × 18 scenarios). See (1)-(3) for English-translated examples of the scenarios for each pragmatic type, with the target sentences in Spanish and English translation. The target word is "vino".

(1) In a noisy party you seemed to hear Lola drinks wine. But you didn't understand well, so you asked: *¿Lola bebe vino?* 'Lola drinks wine?' [CLA-*yn*]

(2) You're told that Elena, who's allergic to alcohol drinks wine. You were surprized and asked: *¿Elena bebe vino?* 'Elena drinks wine?' [EXP-*yn*]

(3) In a store, you want to ask if they have wine: *¿Tienes vino?* 'Do you have wine?' [INF-*yn*]

2.3. Data coding and analysis

We manually segmented the target words at the syllable level using Praat [19] and extracted 10 regularly spaced F0 samples from each syllable to generate a time-normalized pitch contour. The raw values were transformed into *z*-scores.

To estimate and compare the pitch contours of the target words across groups and functions, we built six GAMMs using the *bam()* function from the *mgcv* package [20] in R for the two structures (*yn*- vs. *wh*-questions) across the three target words ("vino", "Marina", "Milán"). For all the GAMMs, the fixed factors were time (the normalized time point), gender (m vs. f), function (CLA vs. EXP vs. INF), group (Chinese students vs. Spanish natives), and a two-way interaction of Function × Group. The smooth terms included a smooth curve for the Function × Group interaction and a by participant (id) random smooth to account for individual differences. The data distribution was specified as "scaled-t".

Finally, as we are interested in the differences in nuclear pitch contours as a function of speaker group and pragmatic meanings, we only report the findings on smooth terms. Figs 1-2 visualize the groupwise comparisons of pitch contours.

3. RESULTS

Table 1 summarizes the GAMM results for smoothed nuclear pitch contours with initial, medial, and final stress. A boldfaced smooth term means that the estimated curve significantly differs from a straight line as measured by the estimated degree of freedom (*edf*), where *edf* =1 means a straight line. Since only a few cases showed non-significant results, we will mainly interpret the post-hoc comparisons of the estimated pitch curves between groups.

Table 1: Smooth terms of the GAMMs for the *z*-scored nuclear F0 of Spanish *yn*- and *wh*-questions produced by Chinese students (ch) and Spanish

natives (sp) with various pragmatic functions (CLA, EXP, and INF) divided by the word stress positions (initial, medial, and final).

	Vino (initial)		Marina (medial)		Milán (final)	
	edf	*F*	*edf*	*F*	*edf*	*F*
***Yn* questions**						
s(time): interaction (group, function)						
ch.CLA-*yn*	**4.6**	**66.4**	**6.0**	**81.9**	**5.3**	**32.5**
sp.CLA-*yn*	1.0	1.0	**5.5**	**28.5**	**2.0**	**8.0**
ch.EXP-*yn*	**4.8**	**95.9**	**6.1**	**81.9**	**6.3**	**58.0**
sp.EXP-*yn*	**4.8**	**24.8**	**5.9**	**28.8**	**5.6**	**16.9**
ch.INF-*yn*	**4.6**	**71.7**	**6.0**	**79.8**	**5.8**	**52.8**
sp.INF-*yn*	**4.6**	**38.7**	**5.8**	**55.2**	**5.5**	**18.7**
s(id, time)	**22.5**	**67.4**	**22.7**	**118.2**	**22.6**	**96.1**
***Wh*-questions**						
s(time): interaction (group, function)						
ch.CLA-*wh*	**1.0**	**4.1**	**5.6**	**48.5**	**5.1**	**40.8**
sp.CLA-*wh*	2.5	2.4	**3.4**	**8.4**	**5.1**	**21.0**
ch.EXP-*wh*	**3.2**	**5.1**	**5.9**	**113.4**	**5.2**	**45.8**
sp.EXP-*wh*	**4.8**	**10.5**	**5.1**	**44.5**	**5.6**	**37.9**
ch.INF-*wh*	1.0	2.1	**4.2**	**4.8**	**3.1**	**3.7**
sp.INF-*wh*	1.9	1.1	**2.6**	**2.6**	2.1	2.3
s(id, time)	**22.5**	**68.7**	**22.5**	**55.2**	**22.6**	**97.2**

3.1. Results of yn-questions

For the initial stressed "vino" (Fig. 1), Chinese students produced significantly higher pitch than Spanish natives on the second syllable. In terms of pitch contours, both groups showed similar patterns in INF-*yn* and EXP-*yn* (L* H%), but in CLA-*yn*, Chinese students produced a L* H% pattern in contrast with the native L* L% pattern.

As for the medial stressed "Marina" (Fig. 1), the significant contrasts in pitch height between groups were mainly on the first syllable, "ma" in CLA-*yn* and EXP-*yn*. The two groups produced similar contour patterns for all three functions (L* H%).

Regarding the final stressed "Milán" (Fig. 1), the most important difference lay in the contours of CLA-*yn,* with Chinese learners showing a L* H% and Spanish natives showing a L* L%. As for EXP-*yn*, Spanish natives showed a L* H% pattern, but Chinese students had a significantly higher boundary tone (L* ¡H%). No significant contrast was found for INF-*yn*, with both groups showing a L* H%.

3.2. Results of wh-questions

First, the initial stressed "vino" (Fig. 2) revealed no significant difference in pitch height between Chinese learners and Spanish natives in any of the three functions. Noteworthily, although Chinese students seemed to show a different pitch contour from Spanish natives, the difference observed in Fig. 2 was not meaningful since no significant contrast was found. Therefore, both groups produced INF-*wh* and CLA-*wh* as L* L% and EXP-*wh* as L* HL%.

Second, the medial stressed "Marina" (Fig. 2) revealed a significant difference in CLA-*wh,* with Spanish natives showing a L* L% while Chinese students showing a L* H%. However, in EXP-*wh*, although Spanish natives produced "ma" with a significantly higher pitch than Chinese students, the

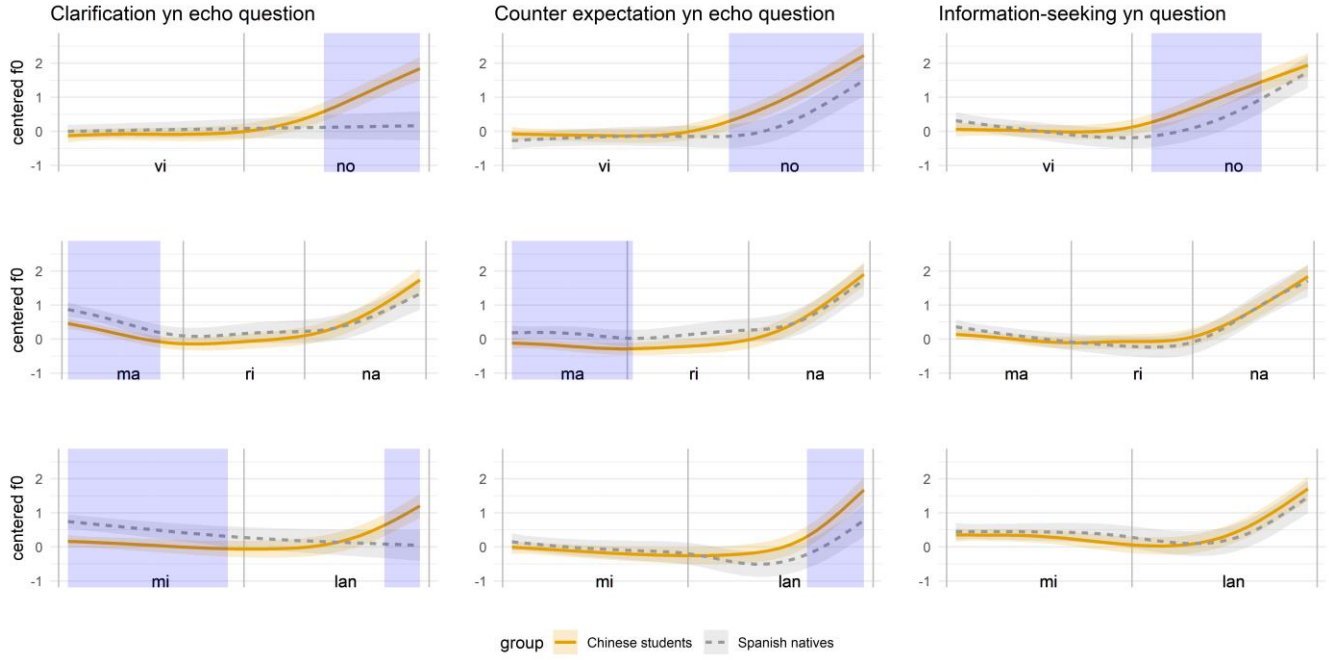


Fig. 1. Groupwise comparisons of F0 contour of the nuclear words estimated by GAMM for each pragmatic function of *yn*-questions, from left to right: CLA-*yn*, EXP-*yn*, and INF-*yn*, varied across stress patterns, from top to bottom, initial stressed, medial stressed, and final stressed. The shaded area of each contour paints the 95% Confidence Interval. The purple squared shades illustrate significant contrasts between the intonation contours produced by Chinese students and Spanish natives.

pitch contours of both groups were L* H%. Again, no significant contrast was found for INF-*wh* (L* L%).

Finally, the final stressed "Milán" (Fig. 2) revealed significant differences in CLA-*wh*, with the Chinese students producing a L+H* H% while Spanish natives showed a H+L* L%. No significant group difference was found in INF-*wh* (L* L%) and EXP-*wh* (L* H%).

4. DISCUSSION AND CONCLUSION

This study compared the nuclear contours produced by proficient Chinese learners of Spanish and Spanish natives for six Spanish question types (INF-*yn*, INF-*wh*, CLA-*yn*, CLA-*wh*, EXP-*yn*, and EXP-*wh*) in a Discourse Completion Task. The target nuclear words varied in stress position (initial, medial, and final).

Our first hypothesis, that CLA would be the most challenging for Chinese students, was confirmed. In INF and EXP, Chinese students showed on-target contours in most cases, but they tended to produce high boundary tones (H%) for CLA-*yn* ending with "vino" and "Milán", and CLA-*wh* with "Marina" and "Milán", which was inconsistent with Spanish natives (L%). That is, Chinese students could not prosodically mark the clarification echo questions.

Our second hypothesis, that the word-final stress would be the least favorable for producing L2 nuclear contours, was also confirmed. In INF and EXP, "vino" and "Marina" showed subtle differences in pitch height but not many in contour shapes between groups. However, the word-final stressed "Milán showed an unexpected high boundary tone (H%) in CLA produced by Chinese students. Especially in CLA-*wh*, they showed a rising nuclear accent (L+H*), which contrasted with the native speakers' falling pattern (H+L*). Also, in EXP-*yn*, even though the contours of "Milán" were similar between groups, Chinese students had a higher boundary tone (¡H%) than Spanish natives (H%).

Importantly, Chinese students performed well in EXP, which suggested a positive cross-linguistic influence since EXP is also marked by a high boundary tone in Chinese [14], [21]. By contrast, Chinese students failed to produce on-target prosodic patterns for CLA because Chinese CLA resembles the intonation of *yn*-question, which bears a high boundary tone [13]. Our results thus support the predictions by LILt [1] that the interaction between L1 and L2 prosodic categories is driven by the (dis)similarities of the two languages. In addition, the dynamic analysis revealed pitch height differences between groups even though the overall pitch contours were similar (e.g., EXP-*yn* with "vino"). This suggests a need for dynamic approaches in L2 prosodic research.

To conclude, this study, for the first time, dynamically modeled Chinese students' nuclear contours of Spanish. For the six question types in our corpus, the clarification echo question seemed to be the most challenging for Chinese students. Also, stress position affected nuclear pitch contours, with the word-final position being the most problematic. The findings thus call for attention to prosody in L2 teaching practice.

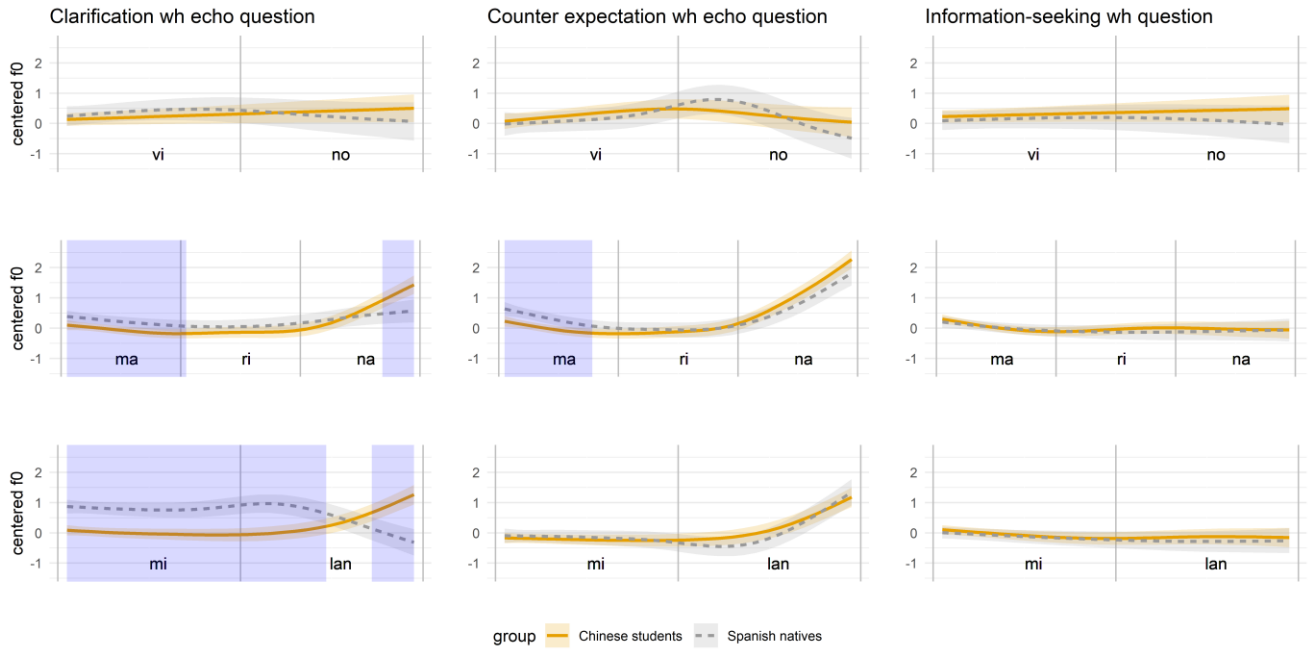


Fig. 2. Groupwise comparisons of F0 contour of the nuclear words estimated by GAMM for each pragmatic function of *wh*-questions, from left to right: CLA-*wh*, EXP-*wh*, and INF-*wh*, varied across stress patterns, from upper to bottom, initial stressed, medial stressed, and final stressed. The shaded area of each contour paints the 95% Confidence Interval. The purple squared shades illustrate significant contrasts between the intonation contours produced by Chinese students and Spanish natives.

5. ACKNOWLEDGMENT

The first author is supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265]. The second author is supported by the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the European Social Fund under the Grant for the recruitment of early-stage research staff [2021FI_B 00137]. We thank Dr. Peizhu Shang (Beijing Institute of Technology) for her comments on the first draft of this paper.

6. REFERENCES

[1] I. Mennen, "Beyond Segments: Towards a L2 Intonation Learning Theory," in *Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations*, E. Delais-Roussarie, M. Avanzi, and S. Herment, Eds. Berlin, Germany: Springer, 2015, pp. 171–188. doi: 10.1007/978-3-662-45168-7_9.

[2] V. Schaefer and I. Darcy, "Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones," *Lab. Phonol.*, vol. 5, no. 4, pp. 489–522, 2014, doi: 10.1515/lp-2014-0016.

[3] J. Feng, S. Tao, X. Wu, K. Alsbury, and C. Liu, "The effects of amplitude and duration on the perception of English statements vs questions for native English and Chinese listeners," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. EL449–EL455, 2019, doi: 10.1121/1.5109046.

[4] P. Shang, W. Elvira-García, and X. Li, "Cue weighting differences in perception of Spanish sentence types between native listeners of Chinese and Spanish," in *Proc. Speech Prosody 2022*, 2022, pp. 644–648. doi: 10.21437/SpeechProsody.

[5] T. Laméris, "The effect of lexical status of pitch in the L1 and extralinguistic factors on L2 tone acquisition," in *Proc. Speech Prosody 2022*, 2022, pp. 708–712. doi: 10.21437/speechprosody.2022-144.

[6] S. Zerbian, L. J. Downing, and F. Kügler, "Introduction: Tone and intonation from a typological perspective," *Lingua*, vol. 119, no. 6, pp. 817–826, 2009, doi: 10.1016/j.lingua.2007.10.024.

[7] Y. Chen and C. Gussenhoven, "Emphasis and tonal implementation in Standard Chinese," *J. Phon.*, vol. 36, pp. 724–746, 2008, doi: 10.1016/j.wocn.2008.06.003.

[8] P. Shang and W. Elvira-García, "Second language acquisition of Spanish prosody by Chinese speakers: Nuclear contours and pitch characteristics," *Vigo Int. J. Appl. Linguist.*, vol. 19, pp. 129–176, 2022, doi: https://doi.org/10.35869/vial.v0i19.3762.

[9] Y. Chen, "From tone to accent: the tonal transfer strategy for Chinese L2 learners of Spanish," in *Proceedings of the 16th International Congress of Phonetic Science*, Saarbrücken, Germany, 2007, no. August, pp. 1645–1648.

[10] P. Li and X. Xi, "Spanish lexical stress produced by proficient Mandarin learners of Spanish," in *Proceedings of the 4th International Symposium on Applied Phonetics*, Lund, Sweden, 2022.

[11] M. Cortés Moreno, "Análisis experimental del aprendizaje de la acentuación y la entonación españolas por parte de hablantes nativos de chino," *Phonica*, vol. 1, 2005, doi: 10.1344/phonica.2005.1.%25p.

[12] E. Estebas-Vilaplana and P. Prieto, "Castilian Spanish intonation," in *Transcription of Intonation of the Spanish Language*, P. Prieto and P. Roseano, Eds. LINCOM publishers, 2010, pp. 17–48.

[13] J. Shao, *Xiandai Hanyu Yiwenju Yanjiu [Interrogative Sentences in Modern Chinese]*. Beijing, China: The Commercial Press, 2014.

[14] Y.-Y. Chuang and J. Fon, "Production and perception of incredulity in yes-no question intonation in Taiwan Mandarin," in *Proceedings of the International Conference on Speech Prosody*, Baixas, France, 2016, vol. 2016, pp. 854–858. doi: 10.21437/speechprosody.2016-175.

[15] P. Prieto and P. Roseano, "Current issues and challenges in Spanish intonational research," in *The Routledge Handbook of Spanish Phonology*, S. Colina and F. Martínez-Gil, Eds. London: Routledge, 2019, pp. 222–233. Accessed: Dec. 22, 2022. [Online]. Available: 10.4324/9781315228112-11

[16] J. I. Hualde and P. Prieto, "Intonational variation in Spanish: European and American varieties," in *Intonation in Romance*, S. Frota and P. Prieto, Eds. Oxford: Oxford University Press, 2015, pp. 450–391.

[17] S. Peng, M. K. M. Chan, C. Tseng, T. Huang, O. J. Lee, and M. E. Beckman, "Towards a Pan-Mandarin System for Prosodic Transcription," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford, U.K: Oxford University Press, 2005, pp. 230–270. doi: 10.1093/acprof:oso/9780199249633.003.0009.

[18] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Massachusetts Institute of Technology, 1980. [Online]. Available: https://dspace.mit.edu/handle/1721.1/16065

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2020. [Online]. Available: http://www.praat.org

[20] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[21] O. J. Lee, "The prosody of questions in Beijing Mandarin," The Ohio State University, 2005. [Online]. Available: https://etd.ohiolink.edu/!etd.send_file?accession=osu1122332580&disposition=attachment

Exploring the Phonetic Context Effects on the Production of Portuguese /ɾ/ by L1-Mandarin Learners

Yidan Cao, Chao Zhou & Gueorgui Hristovsky

Center of Linguistics, School of Arts and Humanities, University of Lisbon
madalena970603@163.com, zhouchao@edu.ulisboa.pt, ghristovsky@campus.ul.pt

ABSTRACT

In this study, we examined whether L1-Mandarin learners' production of European Portuguese tap in onset cluster /CɾV/ is conditioned by the preceding consonant and the following vowel. 14 native Mandarin speakers first learned to associate 12 Portuguese pseudo-words with /CɾV/ cluster to 12 pictures and then performed a picture-naming task. L2 production data were judged by 25 Portuguese native raters and analysed with a logistic mixed-effects regression.

Results indicate that the preceding stop affects L2 production of /ɾ/ as a function of place of articulation (/tɾ/ > /kɾ/ > /pɾ/). We argue that this may be accounted for by different degrees of articulatory effort, rather than by phonological markedness or lexical frequency. An effect of the following vowel was also found (/i/ facilities the rhotic production more than /a/ does), which might also be articulatory-based. The implication of these results on the theoretical development of L2 speech research is discussed.

Keywords: Phonetic context, Onset cluster, European Portuguese, Mandarin Chinese, L2 speech.

1. INTRODUCTION

The European Portuguese (EP) tap /ɾ/ is notoriously difficult for L1-Mandarin learners. It is very often confused with an L1 sound category, /l/, both in speech perception and production (see [32] for an overview). Recent empirical studies even show that the complexity in acquiring this novel sound goes beyond segmental confusability ([29, 31, 32]). For instance, Zhou [31] observed that the EP tap of an onset cluster was produced by L1-Mandarin learners more accurately when preceded by an alveolar stop (e.g. /dɾ/ or /tɾ/) than by an bilabial one (e.g. /pɾ/). Zhou [31] reasoned that this result may stem from the fact that the most prevalent repair strategy employed by L1-Mandarin learners (/ɾ/ → [l]) is blocked when the preceding consonant is an alveolar stop. This is because such replacement would give rise to a cross-linguistically marked structure ([dl] or [tl]) (e.g. [11, 12]). Examining L2 perception of EP /ɾ/ by Mandarin-speaking learners, Vale [29] found that, in intervocalic onset position, the adjacent vowel /a/ helps L1-Mandarin participants identify /ɾ/ (/aɾa/), compared to /i/ (e.g. /iɾi/). Zhou [32] further reported a similar facilitating effect of adjacent vowel /a/ over /i/ in a perceptual discrimination task, where L1-Mandarin learners were asked to discern the difference between /ɾ/ and its confusable category /l/. According to Zhou [32], this vowel effect might be due to the fact that [a] leads to more salient F2 transition slopes than [i] does, thus empathizing the acoustic difference between [l]-[ɾ]. Taken together, these results suggest that L2 acquisition of EP /ɾ/ does not proceed equally in all contexts, i.e. it is subject to its surrounding phonetic contexts.

In spontaneous speech, the perception [26] and articulation [20] of a given speech sound are considerably shaped by its adjacent context. Since an L2 sound is normally acquired within a word or larger prosodic unit, it is not surprising that contextual effects on L2 segmental learning have long been documented (e.g. [16, 17, 22]). Prior research on such effect either targets the preceding or the following context, while, in the case of EP tap in onset cluster (/CɾV/), the acquisition of /ɾ/ may be prone to contextual influences from both directions. In this study, expanding on previous studies, we aim to assess the contextual effects on the L2 acquisition of EP /ɾ/ by L1-Mandarin learners, by exploring both the role of the preceding consonant and also that of the following vowel.

2. PHONETICS AND PHONOLOGY OF EP /ɾ/

The EP /ɾ/ is most frequently realised as a tap. Depending on the adjacent segment and syllable position, other phonetic variants have also been reported [25]. In coda position, when preceding a stop (e.g., *carta*, "letter"), it is very often produced with an epenthetic vowel, while a realization as fricative is more common, if followed by another fricative (e.g., *curva*, "curve"). In word-final position, /ɾ/ is often produced as a voiceless fricative [14] and can even be omitted, especially when the following word starts with a consonant (e.g. [19, 21]).

Regarding its distribution, /ɾ/ may occupy all syllabic and word-level positions, except word-initially [18]. It may form an onset cluster with both

stops (/b, p, d, t, g, k/) and labial-dental fricatives (/f, v/), although clusters /C$_{\text{frictive}}$ɾ/ have a more restricted distribution [18]. The vocalic element of /CɾV/ cluster can be any of EP phonological vowels /a, ɛ, e, i, o, ɔ, u/.

3. CURRENT STUDY

In order to explore whether and how the preceding and following contexts influence the L2 acquisition of EP /ɾ/, we assessed the production of /CɾV/ cluster by L1-Mandarin learners via a picture-naming task. We deem this task more suitable for the present study than other production tasks, because picture-naming does not require the activation of orthographic representations (as in the case of a reading task) and it avoids misperception effect (as in the case of an imitation task), for instance. Both orthographic and perceptual influences have been previously reported on L1-Mandarin learners' acquisition of EP /ɾ/ [32].

On the basis of Zhou [31]'s results, we predicted that a preceding alveolar stop will promote the production of /ɾ/, namely /tɾ/ > /pɾ, kɾ/. Following Vale [29] and Zhou [32], we further expected to observe that the production accuracy of /ɾ/ is higher when followed by /a/ than by /i/.

3.1. Participants

Fifteen female Mandarin native speakers (mean age: 20 years old), all second-year Portuguese majoring students at the XXXX University, were recruited and tested in China. The production data of one participant was excluded from the analysis, because her L1 Mongolian has a trill /r/ in its inventory [27], which may promote the acquisition of EP /ɾ/. None of the participants has fluency in or regular use of another language than English (college-level). None has reported hearing or speech impairment.

3.2. Stimuli

Twelve dissyllabic pseudowords were created with onset cluster /CɾV/ in penultimate stressed syllable. For assessing the role of the preceding consonant, 9 items contain clusters starting with stops varying in place of articulation (labial, alveolar and velar), namely 3 with /pɾa/, 3 with /tɾa/ and 3 with /kɾa/. Only voiceless initial stops were chosen as the voicing feature is quite problematic for L1-Mandarin learners (e.g. [23, 30]). In order to explore the effect of the neighboring vowel, another 3 items with onset cluster /pɾi/ were added to contrast with those with /pɾa/. Here, the bilabial stop /p/ was chosen as the pre-rhotic consonant to minimize the carry-over coarticulation effect between the initial consonant and /ɾ/. Finally, 3 items with onset cluster /pla/ were included as fillers.

Fifteen coloured pictures representing different kinds of objects (e.g. 3 different types of knives; 3 different types of boats) were selected and were randomly associated with 15 test items with /CɾV/ cluster. By doing so, we assume that L2 learners can hardly know the name of all subcategories of a given object, so that it is possible to adopt novel word learning paradigm. See the supplementary material for a complete list of pictures and items (osf.io/4fuec/?view_only=82fdd9ffaf8d4f39abfbe9b2aa5ab33e).

3.3. Task and procedure

The entire experimental session consists of two parts. The first is a novel word learning phase, during which the participants were given the full list of test items with associated pictures. To guarantee the lexicalisation of these items, they were given a week to memorize them and were told that they can review the list as many times as they want before the testing phrase. This time interval was defined, based on the necessary time for lexical consolidation reported in the literature (e.g. [7, 13]).

During the testing phase, each participant performed a picture naming task, administrated in Microsoft PowerPoint. The testing material contains a total of 45 slides, 3 repetitions of each stimulus. The presentation order was totally randomised. The task on average took about 5 minutes. Recordings were made in a quiet room in the library of Beijing Jiaotong University to a Tascam DR100mkIII recorder and were digitized at an audio sampling rate of 44.1 kHz. All recorded sound files were adjusted to the average intensity of 70 dB in Praat [3].

3.4. Data preparation and analysis

A total of 630 tokens (3 items × 5 conditions × 14 speakers × 3 repetitions) were collected, among which 25 tokens were excluded (3.96%) from analysis, due to segmental changes either in first syllable (target form: /ˈtɾaʃo/; learner form: [ˈpɾaʃo]) or the second syllable (target form: /ˈpɾita/; learner form: [ˈpɾiɦi]). The occurrence of segmental replacement led us to consider that the learners have not lexicalised these items.

In order to determine whether the tap was produced accurately by L1-Mandarin participants, 25 native speakers of EP were recruited and were asked to identify the target segment in each production. The remaining 605 tokens were divided into 5 subtests. Each subtest contains the production by 3 informants and was judged by 5 native raters. During the judgment task, native raters listened to one auditory token each time and was asked to choose from three available response options representing the target tap

("r"), the most frequent replacement ("l") and other production deviants ("neither r nor l"). The learners' production of /ɾ/ was regarded as correct, if a rater's response for that production is "r".

A practice session with five test items produced by a native EP speaker was included before the main session to help native judges understand the procedure. Another five items produced by the native speaker were placed randomly in the main session serving as control items. For these 5 control items, if a listener corrected on at least 4, her responses would be considered valid, which was the case for all 25 native judges.

4. RESULTS

The L2 production accuracy of the EP /CɾV/ clusters judged by 25 native raters were visualised in Figure 1.

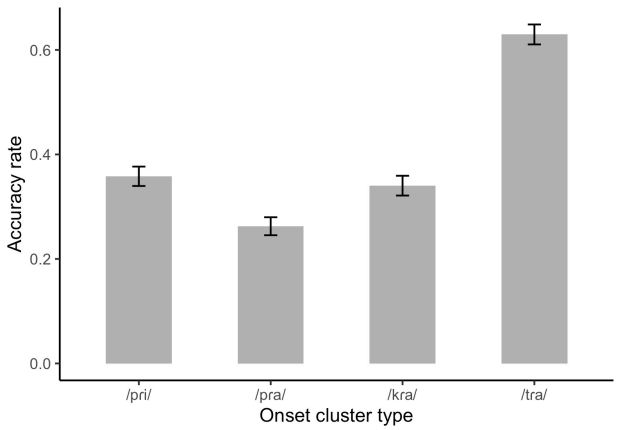


Figure 1: L1-Mandarin learners' production accuracy as a function of EP onset cluster type.

A visual inspection suggests that both the preceding consonant (the comparison between "pra" and "cra" and "tra") and the following vowel (the comparison between "pri" and "pra") seem to have an influence on the L1-Mandarin learners' production of /ɾ/. To explore these effects, we fitted a mixed-effects logistic regression model, which has *Context* as predictor (four levels: "pri", "pra", "cra" and "tra"). The predictor was coded using Backward Difference Coding, which allows us to compare each level with its adjacent one. Both random intercepts and slopes by participant and random intercepts by test item were included in the model. The model's results are summarised in Table 1.

Contrast	*b*	*SE*	*p*.value	95% CI
pri - pra	-1.66	0.66	0.012	[-2.95, -0.37]
pra - cra	1.85	0.68	<0.01	[0.51, 3.2]
cra - tra	4.32	0.73	<0.001	[2.89, 5.75]

Table 1: Results of the mixed-effects logistic regression model.

The statistical analysis confirms the visual inspection. In contrast to previous studies (e.g. [29, 32]), the results of this study reveal that L2 production of EP /ɾ/ is promoted when the tap is followed by /i/ than by /a/. Regarding the preceding consonantal context, the alveolar stop /t/ indeed helped L2 learners produce more target-like tap, in line with Zhou [31]; moreover, the accuracy rate of producing /kɾ/ is higher than that of /pɾ/ clusters, implying the following acquisition order /tɾ/ > /kɾ/ > /pɾ/.

5. DISCUSSION

In the current study, we explored phonetic context effects on L2 segmental learning, by studying how the preceding consonant and the following vowel affect L2 production of EP /ɾ/ in onset cluster. 14 native Mandarin speakers first learned to associate 12 EP pseudo-words containing /CɾV/ clusters to different pictures and then performed a picture-naming task. The L2 production data, judged by 25 EP native raters, indicate that both the preceding and the following phonetic contexts have an impact on the L2 production of EP /ɾ/.

Regarding the preceding consonant, we found an effect of its place of articulation (/tɾ/ > /kɾ/ > /pɾ/). The fact that the L2 production of /ɾ/ is promoted when preceded by an alveolar stop is in line with our prediction. According to Zhou [31], L1-Mandarin learners tend to use [l] when they fail to produce /ɾ/ in onset cluster (/pɾ/ as [pl]), but they rarely do so when the preceding stop is alveolar (/d/ or /t/). This is because the replacement with [l] would yield a cross-linguistically marked structure [dl] or [tl]. The phonological markedness may force L2 learners to rely less on such repair strategy, which may in turn lead to fast acquisition of /tɾ/. However, it should be noted that the phonological markedness does not account for the difference with respect to the production accuracy between /kɾ/ and /pɾ/, since the resulted repair forms [kl] and [pl] are both allowed in the target language EP.

Another plausible explanation for the observed preceding consonant effect pertains to lexical frequency. Although all test items produced by L1-Mandarin speakers are newly learned words, prior research suggests that, as long as these new words become engaged in the mental lexicon, the processing of these novel lexical items can be influenced by

previously learned words [28]. Given that the Chinese participants were colleague students that mainly received Portuguese input in the classroom setting, the frequency of EP onset clusters in the L2 lexicon was estimated on the basis of their coursebook [5], instead of a database of the native Portuguese lexicon. It turns out that the estimated frequency (/pɾ/ > /tɾ/ > /kɾ/) does not really resemble the observed acquisition order.

Instead of phonological markedness and lexical frequency, we argue that the preceding consonant effect on L2 production of /ɾ/ may stem from the articulatory representation of the EP onset clusters. Following the articulatory-based featural representation for onset clusters proposed by Duanmu [6], one can assume that the EP /tɾ/ cluster is represented by one articulator [*Coronal], while two articulators are needed for both /kɾ/ ([*Dorsal, Coronal]) and /pɾ/ [*Labial, Coronal]. Consequently, the simpler representation for /tɾ/ may imply less articulatory effort in its production (see also Espiga [8]), in comparison to other clusters. The articulation-based view also offers a plausible account for the order /kɾ/ > /pɾ/. Although, to our best knowledge, no articulatory evidence exists for the EP onset clusters, a recent study on the Spanish ones shows that the latency from the end of the target gestural plateau for the stop to the onset of that for /ɾ/ is longer in /kɾ/ than in /pɾ/ [10]. This difference might imply an increased articulatory effort for L2 learners in the production of /pɾ/ clusters, where they have to realise the complex and novel gestural coordination of /ɾ/ (a tongue back-lowering followed by a tongue top raising) in a more limited temporal space.

Turning to the effect of the following vowel (/pɾi/ > /pɾa/), we speculate that the nature of this vocalic effect might be articulatory as well. The articulation of [ɾ] has been shown to be highly influenced by its surrounding vocalic context [4]. For instance, when surrounded by /i/, the tongue tip of [ɾ] is more raised and advanced toward the alveolar region than in any other vocalic context (e.g. /a/). Accordingly, the vocalic context /i/ might have helped L2 learners realise this tongue tip gesture, which is crucial for the articulation of a tap.

The vowel effect observed in this study goes in the opposite direction to what has been reported previously (e.g. [29, 32]). Two methodological factors may underlie this asymmetry, namely the stimulus type and the examined speech modality. In this study, the vocalic context can only exert backward effect on the production of /ɾ/, as the EP rhotic only occurs in onset cluster (/CɾV/), while both forward and backward influences are expected in Vale [29] and Zhou [32], where the EP /ɾ/ was assessed in intervocalic onset position (/V.ɾV/). Moreover, the current study targets the contextual effects in L2 speech production, while Vale [29] and Zhou [32] examined L2 speech perception. Many (psico)linguistic models assume explicitly that speech perception and production involve different representational levels and paralinguistic processes [2], which seems to be supported by an increasing number of studies showing that L2 learners' production may not mirror their perceptual performance (e.g. [15, 24]). Future studies targeting both speech modalities are needed, in order to gain a better understanding of the surrounding vowel effect on L2 acquisition of EP /ɾ/.

In our view, the current findings regarding the phonetic context effects on L2 segmental learning have important implications for the development of L2 speech theories. Some major L2 speech models (e.g. [1, 9]) are mainly concerned with the acquisition of a phonetic or phonological category in isolation. It is unclear how the phonetic context effects can be accounted for by these models. Nevertheless, L2 phonetic context effects can be nicely integrated, if one assumes an explicit distinction between phonetics and phonology, as in the *Bidirectional Phonology and Phonetics Model* [2]. In particular, the phonetic form necessarily involved in speech production is a continuous representation, consisting of articulatory gestures, e.g. tongue and lip movements, jaw depression. This phonetic representation is subject to a set of articulatory constraints, which may reflect the articulatory effort during speech production. For instance, an OT-like constraint ranking instantiating articulatory effort *LIPS$_{ART}$ > *TONGUE TIP$_{ART}$, borrowed from Boersma (2011), may account for the acquisition order /tɾ/ > /pɾ/, because the articulatory form of /pɾ/, which involves both tongue lip and tongue tip gestures violates both constraints, while the articulatory form of /tɾ/ is only penalised by the lower-ranked constraint that militates against tongue tip gesture.

6. ACKNOWLEDGEMENTS

This work was partially funded by grant UIDB/00214/2020 from the Portuguese Foundation of Science and Technology (FCT).

7. REFERENCES

[1] Best, C. T., Tyler, M. D. 2007. Nonnative and second-language speech. *Language experience in second language speech learning: In honor of James Emil Flege*, 13-34.

[2] Boersma, P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. In *Bidirectional Optimality Theory*, eds.

Anton Benz and Jason Mattausch, 33–72. Ams- terdam: Benjamins.
[3] Boersma, P., Weenink, D. 2019. Praat. Computer program, version 6.0.48. Retrieved from www.praat.org.
[4] Celata, C., Vietti, A., Spreafico, L. 2018. An articulatory account of rhotic variation in Tuscan Italian: Synchronized UTI and EPG data, *in* Mark Gibson, and Juana Gil (eds), *Romance Phonetics and Phonology*. Oxford Academic.
[5] Coimbra I., Coimbra, O. M. 2018. *Novo Português sem Fronteiras*. LIDEL.
[6] Duanmu, S. 2002. Two theories of onset clusters. *Chinese Phonology*. 11, 97-120.
[7] Dumay, N., Gaskell, M. G. 2007. Sleep-associated changes in the mental representation of spoken words. *Psychological Science*. 18(1), 35-39.
[8] Espiga, J. W. D. R. 1997. *Influência do espanhol na variação da lateral posvocálica no português de fronteira*. MA dissertation, Universidade Católica de Pelotas.
[9] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*. 92, 233-277.
[10] Gibson, M., Sotiropoulou, S., Tobin, S., Gafos, A. I. 2019. Temporal aspects of word initial single consonants and consonants in clusters in Spanish. *Phonetica*. 76(6), 448-478.
[11] Hallé, P. A., Segui, J., Frauenfelder, U., Meunier, C. 1998. Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 592–608.
[12] Hallé, P. A., Best, C. T. 2007. Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters. *The Journal of the Acoustical Society of America*, *121*(5), 2899–2914.
[13] Hayes-Harb, R., Masuda, K. 2008. Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research*. 24(1), 5-33.
[14] Jesus, L. M., Shadle, C. H. 2005. Acoustic analysis of European Portuguese uvular [χ, ʁ] and voiceless tapped alveolar [ɾ̥] fricatives. *Journal of the International Phonetic Association*. 35(1), 27-44.
[15] de Leeuw, E., Stockall, L., Lazaridou-Chatzigoga, D., Gorba Masip, C. 2019. Illusory vowels in Spanish–English sequential bilinguals: Evidence that accurate L2 perception is neither necessary nor sufficient for accurate L2 production. *Second Language Research*, *37*(4), 587–618.
[16] Levy, E. S. 2004. *Effects of language experience and consonantal context on perception of French front rounded vowels by adult American English learners of French.* PhD dissertation, City University of New York, USA.
[17] Levy, E. S., Strange, W. 2008. Perception of French vowels by American English adults with and without French language experience. *Journal of phonetics*. 36(1), 141-157.
[18] Mateus, M. H.M., d'Andrade, E. 2000. *The phonology of Portuguese*. OUP Oxford.
[19] Mateus, M.H.M., Rodrigues, C. 2003. A vibrante em coda no português. *Teoria Lingüística. Fonologia e outros temas*, 181-199.
[20] Recasens, D. 2022. Acoustic characteristics and placement within vowel space of full schwa in the world's languages: A survey. *Journal of the International Phonetic Association*, *52*(1), 59-94.
[21] Rodrigues, C. 2003. *Lisboa e Braga: Fonologia e Variação*. Ph.D. Dissertation, University of Lisbon, Portugal.
[22] Ogden, D. C. 2015. Phonetic context effects in second-language phonetic category discrimination. *The Journal of the Acoustical Society of America*. 137(4), 2381–2381.
[23] Oliveira, D. 2021. *Auditory selective attention and performance in high variability phonetic training: The perception of Portuguese stops by Chinese L2 learners*. PhD thesis, University of Minho, Portugal.
[24] Sakai, M. 2016. (Dis)Connecting perception and production: Training native speakers of Spanish on the English /i/-/I/ distinction. PhD dissertation). Georgetown University, USA.
[25] Silva, A. 2014. *Análise Acústica da Vibrante Simples do Português Europeu*. Dissertação de Mestrado, Universidade de Aveiro.
[26] Stilp, C. 2020. Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*. 11(1), e1517.
[27] Svantesson, J. O., Tsendina, A., Karlsson, A., Franzén, V. 2005. *The phonology of Mongolian*. OUP Oxford.
[28] Tamminen, J., Gaskell, M. G. 2008. Newly Learned Spoken Words Show Long-Term Lexical Competition Effects. *Quarterly Journal of Experimental Psychology*, *61*(3), 361–371.
[29] Vale, A. A. 2020. *Perceção das Consoantes líquidas /ɾ/ e /l/ do Português Europeu sob Influência do Mandarim L1*. MA dissertation, University of Minho, Braga, Portugal.
[30] Yang, S. 2014. *Perceção das consoantes oclusivas de Português L2 sob a influência de Mandarim L1*. MA dissertation, University of Minho, Braga, Portugal.
[31] Zhou, C. 2017. *Contributo para o estudo da aquisição das consoantes líquidas do português europeu por aprendentes chineses*. MA dissertation, University of Lisbon, Lisbon, Portugal.
[32] Zhou, C. 2021. L2 speech learning of European Portuguese /l/ and /ɾ/ by L1-Mandarin learners: Experimental evidence and theoretical modelling. *Language Acquisition*. DOI: 10.1080/10489223.2021.1952591

PERCEPTION OF VOWEL AND CONSONANT QUANTITY CONTRASTS BY CANTONESE, ENGLISH, FRENCH, AND JAPANESE SPEAKERS

Albert Lee[1], Yasuaki Shinohara[2-4], Faith Chiu[5], Tsz Ching Mut[1]

[1]The Education University of Hong Kong, [2]Waseda University, [3]City University of New York, [4]University of Delaware, [5] Glasgow University Laboratory of Phonetics, University of Glasgow
albertlee@eduhk.hk, y.shinohara@waseda.jp, faith.chiu@glasgow.ac.uk, tcmut@eduhk.hk

ABSTRACT

This paper reports on the findings of a study which investigated Cantonese, Japanese, English and French listeners' ability to perceive non-native quantity contrasts. In these languages, duration is used to mark phonemic quantity contrasts to different degrees. We had native listeners of these languages listen to resynthesized Estonian nonce word stimuli in AXB discrimination and identification tasks. The stimuli contrasted in consonant and vowel quantity. The results showed that Japanese listeners, who have short vs. long contrasts in both vowels and consonants, outperformed the other listeners in discrimination and identification. However, their identification accuracy for *overlong* Estonian vowels and consonants was not as high as that for *long* Estonian vowels and consonants. Meanwhile, French listeners, who have no quantity contrasts in their L1 phonology, did not perform worse than the other groups. The theoretical implications of these findings are discussed.

Keywords: feature hypothesis, L2 phonology, phonemic quantity contrasts

1. INTRODUCTION

1.1. Background

The issue of L2 phonemic quantity contrasts has received increasing attention in recent years as it holds the key to answering some theoretical questions in L2 phonological acquisition. Specifically, this line of research asks whether learners' difficulty in learning L2-specific phonetic contrasts is affected by L1-L2 differences at the level of discrete sound categories (i.e. discrete segments) or continuous phonetic dimensions (or 'features') (see review in [1], §2.5).

Existing experimental results have painted a mixed picture of the relationship between L1 phonology and L2 quantity acquisition. Although generally learners seem to be able to tell apart short vs. long in the target language at least to some extent (e.g. [2] for Cantonese), they also deviate from native speakers in numerous ways (e.g. ability to exaggerate short vs. long differences at slower speech rates [2]), making answering the category vs. feature question less than straightforward. A further complication is that these studies often compared only one language pair at a time (e.g. [3] who looked at Mandarin-speaking learners of Italian), with learners of varying proficiency levels, using different experimental methods. Moreover, the languages tested also differed in terms of orthographic depth, with some where quantity is clearly marked in the writing system (e.g. Japanese) whereas for others the same is not true (e.g. Cantonese). Thus, to examine the relationship between L1 phonology and L2 quantity acquisition, ideally one should compare participants from L1 backgrounds that make quantity distinctions to varying degrees. Other factors should also be held constant, such as proficiency (e.g. naïve) and orthographic depth (e.g. naïve participants without literacy of the target language). This study is an attempt to make such direct comparison.

1.2. Phonemic quantity in Estonian

Estonian has three-way quantity distinctions in both consonants and vowels. There are three levels of quantity, namely short, long, and overlong. The short vs. long distinction is faithfully represented in the writing system, e.g. *sada* 'hundred' and *saada* 'send', but the long vs. overlong difference is not [4]. Table 1 (in §2 below) shows relative segment durations of typical disyllabic Estonian words of different syllable structures.

1.3. Line-up of L1 backgrounds

Here we tested native listeners of Japanese, English, Cantonese, and French, who use duration in their L1 phonology to different degrees. Japanese has systematic short vs. long differences in both vowels and consonants [5]. Both obstruent consonants (e.g. /kita/ 'came' vs. /kitta/ 'cut') and vowels (e.g. /obasan/ 'aunt' vs. /oba:san/ 'grandmother') phonologically contrast in quantity, with duration being the primary acoustic cue [5]. English has short vs. long vowels (e.g. *bit* vs. *beat*) although duration is only one of the acoustic cues (alongside vowel quality) [6]. English also has false geminates at word boundaries such as *cat tail*. Cantonese has short vs.

long vowels but only limited to a small set of pairs (e.g. /ʉi/ vs. /aːi/). It also has false geminates at morpheme boundaries like in English. French has no phonemic quantity contrasts [7], and is said to be 'quantity insensitive'.

1.4. Experimental findings on L2 quantity acquisition

Experimental findings have pointed to a clear relationship between L1 phonology and mastery of L2 quantity distinctions. In [8], Estonian, English, and Spanish speakers were assessed on their mastery of Swedish two-way quantity distinctions (short vs. long). Spanish speakers do not use duration to mark phonemic quantity contrasts even as a secondary cue, unlike English. The results demonstrated that although both the English and Spanish speakers identified Swedish vowels less well than the Estonian speakers, the English speakers showed slightly better performance than the Spanish speakers. This shows that an L1 background with more extensive use of duration as a quantity cue is beneficial for L2 quantity acquisition.

For those from a 'quantity-sensitive' L1 background (e.g. Japanese), acquiring more complex L2 quantity contrasts can be challenging. In [9] (replicated in [10]) Japanese learners of Estonian were able to distinguish between Estonian long vs. overlong consonants, but not long vs. overlong vowels, though both are absent in Japanese.

Interestingly, recent studies ([3], [11]) showed that arguably "quantity-insensitive" Mandarin learners were able to tell apart Italian short vs. long consonants. Similarly, Cantonese learners were found to be able to distinguish between Japanese short vs. long vowels and consonants in their production [2]. All in all, it seems that insofar as their ability to tell apart short vs. long is concerned, with enough exposure learners can acquire quantity distinction without major problems. What is interesting, however, is how participants from different L1 backgrounds *relatively* perform in a direct comparison, where known confounds are controlled.

1.5. Hypotheses

Our overarching goal is to understand the relationship between how far duration is used in L1 as a quantity cue and how successfully one can acquire L2 quantity contrasts. To this end, it is necessary to line up L1 backgrounds that make quantity distinctions to varying degrees. The stimuli and L2 proficiency of participants should also be controlled.

Here we tested the following hypotheses: (H1) Japanese listeners' perception accuracy in both discrimination and identification is the highest; (H2) French listeners' perception accuracy is the lowest; (H3) Japanese listeners would not be able to discriminate or identify the long vs. overlong contrast as well as they do the short vs. long contrast in Estonian.

2. METHODOLOGY

2.1. Participants

Twenty native listeners of Cantonese (11 female, aged 19 to 50), 20 Japanese (11 female, aged 18 to 25), 20 English (10 female, aged 19 to 49) and 15 French (14 females, aged 18 to 24) were recruited. They had no (history of) hearing or language impairments. All Cantonese participants spoke English and Mandarin as L2 with varying proficiency levels. The Japanese participants had been learning English at school but did not speak it in their daily lives, nor did they have experience living outside Japan for over two months. The French participants were university students studying in the United Kingdom. No participant reported to speak any other language.

2.2. Stimuli

Synthetic Estonian nonce word stimuli were generated using VocalTractLab 2.2 [12]. We chose to use synthetic stimuli to control for non-durational secondary cues that could influence perception, e.g. pitch [13]. There were 75 (nonce) words (15 CVCV base real words × 5 quantity conditions: CVCV, CVVCV, CVVVCV, CVCCV, CVCCCV). These were spoken in three synthetic voices, differing in fundamental frequency (male 110 Hz, male 150 Hz, female 200 Hz), vocal tract length, and voice quality. The actual duration of each segment is listed in Table 1 (based on [2]). For all data collection sites, we used e-Prime to present stimuli and record responses.

	CVCV	CVVCV	CVVVCV	CVCCV	CVCCCV
C1	80				
V1	90	170	240	90	
C2	80			140	200
V2	140				

Table 1. Segment duration of Estonian stimuli (in ms)

2.3. Procedures

2.3.1 AXB discrimination task

In the AXB discrimination task, participants heard a sequence of three speech stimuli through a pair of headphones, and judged whether the middle one was the same as the first or the last one. They were asked to respond as quickly as possible.

2.3.2 Identification task

In each trial of the identification task, participants heard one speech stimulus, and identified the quantity of vowels and consonants. In the vowel block, participants chose from CVCV, CVVCV, and CVVVCV (e.g. pada, pa ː da, pa ː ː da); in the consonant block, they chose from CVCV, CVCCV, and CVCCCV (e.g. pada, pad:a, pad::a). The order of the two blocks was counterbalanced across participants.

3. RESULTS

3.1 AXB discrimination task

Figure 1 displays the AXB discrimination accuracy of the Estonian nonce word stimuli by Cantonese, English, French and Japanese listeners. We fitted logistic mixed effects models to the correct/incorrect binomial responses using the R package lme4 [14]. The model included the fixed factors of participants' L1 (Cantonese, English, French, Japanese), stimulus pair (short-long, long-overlong) and their interaction. Orthogonal contrasts were set for these categorical variables. Random intercept for participant was also included.

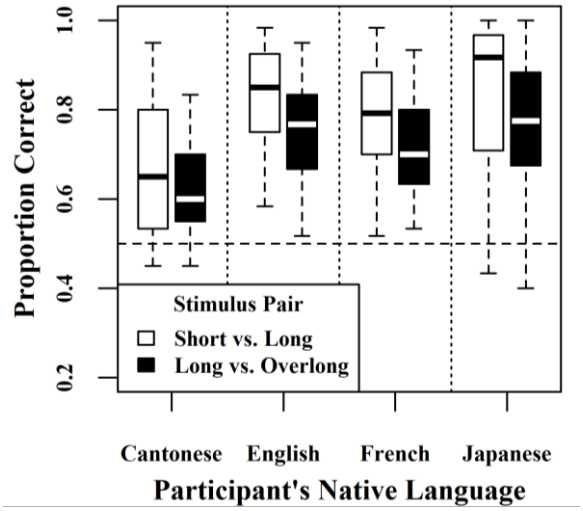


Figure 1. Discrimination accuracy of Estonian vowel and consonant quantity contrasts by Cantonese, English, French, and Japanese speakers. The horizontal dashed line represents 50% chance level.

The results demonstrated that Japanese listeners were significantly better at discriminating Estonian quantity contrasts than other groups of listeners, β = 0.14, SE = 0.05, z = 3.07, p < 0.01. English listeners performed significantly better than Cantonese listeners, β = -0.40, SE = 0.11, z = −3.66, p < 0.001, whereas French speakers did not perform significantly worse than Cantonese or English speakers, β = 0.04, SE = 0.07, z = 0.55, p > 0.05. The stimulus pair effect (short-long vs. long-overlong discrimination) was significant, β = -0.21, SE = 0.02, z = -11.28, p < 0.001, suggesting that the short vs. long contrast was easier to discriminate than the long vs. overlong contrast across L1 groups.

The two-way interaction of participant's L1 and stimulus pair demonstrated that the effect of stimulus pair was larger for Japanese speakers than other language groups, albeit marginally significant, β = -0.02, SE = 0.01, z = -1.90, p = 0.058. This means that, compared to other language speakers, the short vs. long contrast was easier for Japanese speakers to discriminate than the long vs. overlong contrast. In addition, the stimulus pair effect was significantly larger for English speakers than Cantonese speakers, β = 0.08, SE = 0.02, z = 3.44, p < 0.001.

Further analysis was conducted for each language group. The results demonstrated that the long vs. overlong discrimination was significantly more difficult than the short vs. long discrimination for all language groups, p < 0.01.

3.2 Identification task

Figure 2 shows the identification accuracy of Estonian vowels and consonants quantity contrasts by the four language groups. The logistic mixed effects model included the fixed factors of participants' L1, quantity condition (short vs. long, long vs. overlong), and their interaction. The random effects were participant and talker voice.

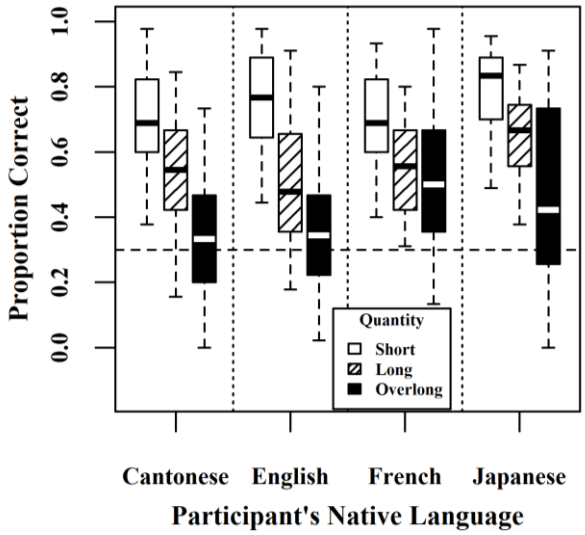


Figure 2. Identification accuracy of Estonian vowels and consonants in the quantity contrasts by Cantonese, English, French, and Japanese speakers. The horizontal dashed line indicates 33.3% chance level.

The logistic mixed effects model demonstrated that Japanese listeners were better at identifying the Estonian quantity contrasts than all other groups of listeners, β = 0.07, SE = 0.03, z = 2.64, p < 0.01. The identification accuracy of short stimuli was higher

than that of the long stimuli, $\beta = 0.44$, SE = 0.07, $z = 6.38$, $p < 0.001$, and the overlong stimuli were less accurately identified than the other two quantity conditions (short & long), $\beta = 0.36$, SE = 0.04, $z = 10.05$, $p < 0.001$.

The two-way interactions of participant's L1 and quantity condition demonstrated that, although the quantity effect (overlong vs. short & long) was not significantly larger for Japanese speakers than other language groups, $\beta = 0.003$, SE = 0.01, $z = 0.55$, $p > 0.05$, that was larger for Cantonese and English groups than for French speakers, $\beta = -0.08$, SE = 0.01, $z = -8.53$, $p < 0.001$. Similarly, the quantity effect of the short vs. long contrast was larger for Cantonese and English groups than for French, $\beta = -0.04$, SE = 0.02, $z = 2.64$, $p < 0.01$, and that was larger for English than Cantonese group, $\beta = 0.11$, SE = 0.03, $z = 4.23$, $p < 0.001$.

Further analysis for each language group was conducted. For all L1 groups the identification accuracy of short stimuli was significantly higher than the long stimuli, $p < 0.001$, and that of overlong stimuli was significantly lower than the other two quantity conditions (short and long), $p < 0.001$.

5. DISCUSSION

This study set out to test the following hypotheses: (H1) Japanese listeners' perception accuracy is the highest; (H2) French listeners' perception accuracy is the lowest; (H3) Japanese listeners would not be able to discriminate or identify the overlong vowels and consonants as well as they do the long vs. short contrast. We found that Japanese listeners outperformed Cantonese, English and French listeners in both discrimination and identification, supporting H1. On the other hand, French listeners were not found to perform worse than other groups in any of the tasks, thus refuting H2. Discriminating and identifying overlong stimuli were more difficult than the short and long stimuli for Japanese speakers, and that was common for all language speakers, supporting H3.

Having systematic two-way quantity contrasts in their L1 may have allowed the Japanese listeners to discriminate similar two-way quantity contrasts of a non-native language. This is reminiscent of a previous study [15], in which a non-native vowel quantity contrast was perceived just as well as the native consonantal quantity contrast.

For the Cantonese listeners, their partial use of duration to mark vowel quantity contrasts (in only a small subset of vowels) in their L1 may have helped them discriminate non-native three-way quantity contrasts but not to the same extent as for the Japanese listeners. However, they performed less well in discrimination than English listeners, to whom duration is only one of multiple acoustic cues to vowel quantity. Note also that the English listeners were the only monolingual group not reporting to speak any L2. Our findings may appear to suggest that having many vowel pairs contrasting in quantity (i.e. English), even if duration is but one of multiple cues, is more beneficial to acquiring non-native quantity contrasts than having few quantity-contrasting vowel pairs in L1 (i.e. Cantonese). Needless to say, this speculation needs to be verified with more empirical evidence.

What is puzzling is why the French listeners performed much better than expected, despite the fact that their L1 is often deemed 'quantity-insensitive' [7]. As all the participants were naïve listeners of the target language, it is unclear what their good performance in the present study can be attributed to. The only conceivable factor that may set them apart from the other L1 groups is that they were immersed in an L2 at the time of testing, though how this might have contributed to the current findings is equally unclear. Also interesting is the significant interactions between L1 and quantity contrast, where the differences in identification accuracy of numerous quantity contrasts were significantly smaller for French than for other L1 groups. Further investigation is needed.

Although we have lined up multiple L1 backgrounds and compared listeners' perception accuracy in non-native word stimuli, we found that only Japanese listeners unambiguously outperformed the others. Meanwhile, the relative performance of Cantonese and English (partial quantity distinctions) as well as French ('quantity-insensitive') in different tasks does not seem to be easily attributable to their respective use of duration as a quantity cue (*contra* [8]). Although recent experimental findings (looking at two languages at a time) have improved our understanding of L2 quantity acquisition, the present direct comparison of *four* language backgrounds has shown that the picture is far from clear. A production study with these four listener groups is currently underway to shed further light on this.

6. ACKNOWLEDGEMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China awarded to AL (Project No. ECS 28605120). We thank Mr. Mingyu Weng for assistance with creating the stimuli.

7. REFERENCES

[1] N. Jiang, *Second language processing: An introduction*. New York, NY: Routledge, 2018.

[2] A. Lee and P. K. P. Mok, "Acquisition of Japanese quantity contrasts by L1 Cantonese speakers," *Second Lang. Res.*, vol. 34, no. 4, pp. 419–448, 2018.

[3] Q. Feng and M. G. Busà, "Mandarin Chinese-speaking learners' acquisition of Italian consonant length contrast," *System*, vol. 111, no. 102938, pp. 1–13, 2022.

[4] I. Lehiste, "The function of quantity in Finnish and Estonian," 1965.

[5] Y. Hirata, "Effects of speaking rate on the vowel length distinction in Japanese," *J. Phon.*, vol. 32, no. 4, pp. 565–589, 2004.

[6] A. S. House, "On vowel duration in English," *J. Acoust. Soc. Am.*, vol. 33, no. 9, pp. 1174–1178, 1961.

[7] P. A. Hallé, R. Ridouane, and C. T. Best, "Differential difficulties in perception of Tashlhiyt Berber consonant quantity contrasts by native Tashlhiyt listeners vs. Berber-naïve French listeners," *Front. Psychol.*, vol. 7, no. 209, pp. 1–16, 2016.

[8] R. McAllister, J. E. Flege, and T. Piske, "The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian," *J. Phon.*, vol. 30, no. 2, pp. 229–258, 2002.

[9] E. Meister, R. Nemoto, and L. Meister, "Production of Estonian quantity contrasts by Japanese speakers," *Eesti ja Soome-Ugri Keeleteaduse Ajak.*, vol. 6, no. 3, pp. 79–96, 2015.

[10] A. Lee, Y. Shinohara, and T. C. Mut, "Non-native length contrast perception by Japanese and Cantonese speakers," *Proc. Meet. Acoust.*, vol. 45, no. 060003, pp. 1–9, 2022.

[11] Q. Feng and M. G. Busà, "Acquiring Italian stop consonants: A challenge for Mandarin Chinese-speaking learners," *Second Lang. Res.*, 2022.

[12] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS One*, vol. 8, no. e60603, pp. 1–17, 2013.

[13] Y. Minagawa, K. Maekawa, and S. Kiritani, "日本語学習者の長 / 短母音の同定におけるピッチ型と音節位置の効果 [Effects of pitch accent and syllable position in identifying Japanese long and short vowels: Comparison of English and Korean speakers]," *J. Phonetic Soc. Japan*, vol. 6, no. 2, pp. 88–97, 2002.

[14] D. M. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[15] H. Altmann, I. Berger, and B. Braun, "Asymmetries in the perception of non-native consonantal and vocalic length contrasts," *Second Lang. Res.*, vol. 28, no. 4, pp. 387–413, 2012.

ASSESSING INTELLIGIBILITY OF L2 SPEECH BY NATIVE AND NON-NATIVE LISTENERS: A LONGITUDINAL STUDY

Alejandra Pesantez[1], Ronaldo Lima Jr[2]

[1]University of Zurich, [2] Federal University of Ceará
alejandra.pesantezpesantez@uzh.ch; ronaldojr@letras.ufc.br

ABSTRACT

This study investigated the development of the English vowel contrast /i/-/ɪ/ by Ecuadorian learners over three semesters (T1, T2, T3). Development was examined by means of local intelligibility via native English, Swiss German, and Ecuadorian Spanish listeners' words transcribed correctly. It also explored the Interlanguage Speech Intelligibility Benefit (ISIB) between talkers and listeners from the same native language (L1) background. Speech samples collected from Ecuadorians (EC) with a Spanish L1 background (N=24) were presented to the three groups of listeners (N=15). Results of a Bayesian hierarchical model showed a tendency of improvement of the EC learners from T1 to T2, but not to T3. There was also evidence of ISIB for listeners, with EC listeners, in general, assessing a higher rate of intelligibility to EC speakers compared to native English and Swiss German listeners.

Keywords: Development, L2 speech, vowel, intelligibility, listeners.

1. INTRODUCTION

The majority of students in the world learn English as a second language (L2), in most cases in a formal instruction setting. In this kind of context, most L2 learners have limited access to native English speakers with more exposure to foreign-accented speech, receiving only a few hours of instructional input per week, with not much opportunity to use the target language outside of the classroom [1], [2]. Typical L2 learners of English in Ecuador, the population of interest in our study, are mainly exposed to Spanish-accented English during their years of classroom-based learning. This language environment can provide L2 speakers with different L2 phonological representations from those exposed to native English [3]. Furthermore, during the learning process of a language, it is common to observe some pronunciation errors, and segmental errors appear to be very common in L2 speech. Data from several studies suggest that some L2 pronunciation errors can be overcome during the first year of L2 immersion while others appear impervious [4], [5]. At the same time, results from previous studies have proposed some hierarchies of difficulty in vowel production and perception based on the L1 of learners [4], [6]. For example, L1 Spanish speakers have difficulties producing tense/lax vowel differences [7]. Moreover, not all vowel contrasts are equally important for communication, and teachers should know which contrasts are problematic for L2 English speakers [8]. [9]For example, for L1 Spanish speakers, the English vowel /ɛ/ does not have a big impact on vowel intelligibility, even though it is sometimes pronounced as /e/ [4]. Conversely, the vowel contrast /i/ -/ɪ/ has a tremendous influence on intelligibility, and their mispronunciation can trigger misunderstandings. For this paper, intelligibility refers to "the extent to which a speaker's message is actually understood" [10]. Furthermore, the role of the listener in the intelligibility of speech is also crucial for effective communication [11], and the number of non-native L2 English listeners is more prevalent than that of native-English speakers (NEs).

Numerous studies on the assessment of L2 speech intelligibility have focused on native English (NE) listeners' ratings of vowels or word intelligibility [10]. However, there is some evidence to suggest that there is a benefit of speakers and listeners sharing a language, indicating that L2 speech with relatively low intelligibility for NE listeners could be highly intelligible for non-native English listeners who share the same L1 background [11], [12], [3]. For instance, the study by Uzun [12] demonstrated that the English words produced by Turkish speakers were more intelligible to Turkish listeners than NE listeners. This talker-listener relationship advantage was described by [13], [14] as the Interlanguage Speech Intelligibility Benefit (ISIB). In contrast to these findings, however, no evidence of ISIB was detected by other studies [15], [16]. For example, Munro et al. [15] did not find a consistent intelligibility benefit for languages shared between listeners with Cantonese, Japanese, Mandarin, and English backgrounds and speakers of Cantonese, Japanese, Polish, and Spanish. The explanation for these conflicting results might be due to the different kinds of elicitation tasks for measuring L2 intelligibility (global vs local) used, as well as the non-native listeners' proficiency levels.

The goal of the present study is two-fold. Firstly, we aimed to track the early stages of L2 development

of the English vowel contrast /i/-/ɪ/ produced by 24 Ecuadorian speakers. Development was analyzed by means of local intelligibility to investigate L2 learning processes and speaker errors that lead to problems for listeners [17]. Secondly, we aimed to investigate ISIB via native English, Swiss German, and Ecuadorian listeners' transcriptions of EC and NE speakers' recorded words.

2. METHOD

2.1. Speakers

This data is part of a longitudinal study about foreign language acquisition conducted by the first author and used in different publications with different purposes. The study started with 24 Spanish-speaking Ecuadorian learners of English (EC) aged between 18 and 28, but for the following two recordings only 21 participants showed up. The EC speakers had not lived in an English-speaking country for the purpose of studying and are L1 Spanish monolinguals. However, three participants reported having studied French, Portuguese, and Quechua respectively, reaching a basic level in these languages. Participants were majoring in English language teaching at a state university in Ecuador. They were in the third, fourth, and fifth levels of their studies in the three recordings (T1, T2, and T3) that were analyzed. In the fourth semester, they studied phonetics, and in the fifth phonology. They also had experience learning English for four hours per week during their secondary school in Ecuador with non-native English input most of the time. Furthermore, in the English program at the university, most of the English instructors were non-native English speakers. All participants gave their written informed consent for each recording session and were paid for their participation. Furthermore, 4 American speakers of English (NE) (mean age = 19.75), were used as a control group for comparative purposes in the intelligibility assessment. Participants in the NE group were two female and two male students from different parts of the United States of America, who were in Ecuador studying Spanish as a second language. They were recorded at T1, and their participation was voluntary.

2.2. Speech production

The EC and NE speakers produced 40 isolated monosyllabic words containing the following English vowel contrasts /iː/-/ɪ/; /uː/- /ʊ/; /ɛ /-/æ/; /ʌ/-/ ɑ/) in a CVC or CVCC context through a picture-naming task. For this study, we only present words containing the /i/-/ɪ/ contrast because it has a high functional load in English, such as the word pairs of this research: cheap-chip, keys-kiss, feet-fit, seat-sit, and sheep-ship [18]. Second, previous studies have suggested that speakers from diverse L1 backgrounds (e.g., Catalan, Cantonese) have difficulties to produce the /i/-/ɪ/ pairs. [6], [7], [19]. Participants were recorded at six months intervals, upon finishing each semester.

The recordings were conducted in the radio station of a state university in Ecuador, and the audio was captured using a Zoom H2n handy recorder at a 44.1 kHz sampling rate, and a 16-bit quantization rate. Before starting the recordings, participants were familiarized with the task in a preliminary trial, with instructions written in English and appearing on the first slide. No recording was performed during the preliminary phase. All the pictures were presented in a random order with their Spanish translation next to them to avoid the effect of orthography [4] in the production of the English segments. Participants repeated each word twice to avoid mispronunciation or hesitation during the first production. They were asked to say the words in a natural way and at a normal volume. NE speakers followed the same procedures.

2.3. Listeners

Fifteen listeners participated in the current study. Most of them live in Ecuador or were visiting the country at the moment of the judgement sessions. 5 were American speakers of English (NEL) aged between 22 and 45, two of whom had lived in Ecuador for over 4 years, and two lived in Central America for the purpose of studying Spanish for a mean of 1.65 years. There were also 6 Ecuadorian (two English teachers, one tourist guide and three university students) and 4 Swiss German listeners, who will be referred to as ECL and GL respectively, and who were fluent in English. Most of the GL were English teachers (primary, secondary, or university) in their respective countries. The members of the ECL group were aged between 22 and 38. Two of the participants had studied in an English-speaking country (mean=1 year). The members of the GL group were aged between 26 and 45. Three of them were Swiss German speakers and one was a German speaker fluent in Swiss German. All the GL reported having also studied French. One GL had lived in Ecuador for around 9 years, being fluent in Spanish as well. Prior to participating, they declared not to have hearing problems. A language questionnaire was administered to each listener prior to the listening tasks. Their participation was voluntary.

2.4. Procedure

Participants were seated in front of a computer monitor using headphones in a quiet room during the

sessions. Using Praat software [20], each listener pressed a button to play the target word and heard the randomly ordered presentations of the stimuli twice through headphones. They heard the words containing the /i/-/ɪ/ vowel contrast produced by the EC during the three semesters and the words produced by NE at T1, T2, and T3 which were presented during the three rating sessions as a control group. They were instructed to write out exactly what was said using a computer keyboard. Before starting the task, listeners were trained with four trials and regulated the loudness of the stimuli. Each session lasted from 45 to 50 minutes.

2.5. Statistical analyses

EC and NE individual words were scored as correct or incorrect based on whether orthographic transcriptions by listeners matched the target words produced by the speaker [21]. This data was used to fit a Bayesian mixed-effects logistic regression model, with *vowel* ([i] or [ɪ]), *time* (T1, T2 and T3), *speaker nationality* (EC or NE) and *listener nationality* (ECL, NEL and GL) as predictor variables, with varying intercepts for word and listener, and varying slopes for speaker as a function of time.

3. RESULTS

3.1. Aim 1: To track early stages of L2 development of English vowel contrast /i/-/ɪ/ produced by 24 Ecuadorian speakers.

We fitted a Bayesian logistic regression model[1] (with flat priors) to the data using the brms package [22] for R, whose coefficients are presented in Table 1. Instead of providing point estimates with the probability of the data given the null hypothesis (i.e., p-values), Bayesian models provide probability distributions of the parameters given the data. Figure 1 presents the graphs for the probability distributions from Table 1. The corresponding measurement of error model yields an estimated posterior mean of 0.37, with a 95% credible interval [0.12, 0.65] showing the increase in the probability of correct vowels for T2 in comparison with T1 (the intercept), but not so much for T3 in comparison with T2 with an estimated posterior mean of 0.48, with 95% credible interval [0.23, 0.73]. It also shows that the model predicts a credible higher probability of correct vowels for NE speakers.

Predictors	**correct.code** *Log-Odd*	*CI (5%)*
Intercept	0.11	-0.50 – 0.71
speaker_nationality:NE	0.72	0.10 – 1.37
time: T2	0.37	0.12 – 0.65
time: T3	0.48	0.23 – 0.73
vowel: ɪ	-0.10	-0.81 – 0.67
listener_nationalityGL	-0.22	-0.62 – 0.20
listener_nationalityNEL	-0.22	-0.60 – 0.17
listener_nationalityNE:listener_GL	1.33	0.76 – 1.95
listener_nationalityNE:listener_NEL	2.43	1.59 – 3.54
timeT2: listener_nationalityGL	0.16	-0.09 – 0.42
timeT3: listener_nationalityGL	0.41	0.15 – 0.69
timeT2: listener_nationalityNEL	0.24	-0.03 – 0.53
timeT3: listener_nationalityNEL	0.46	0.17 – 0.76
vowelɪ: listener_nationalityGL	0.64	0.43 – 0.85
vowelɪ: listener_nationalityNEL	0.61	0.40 – 0.83

Table 1: Coefficients for Regression Model.

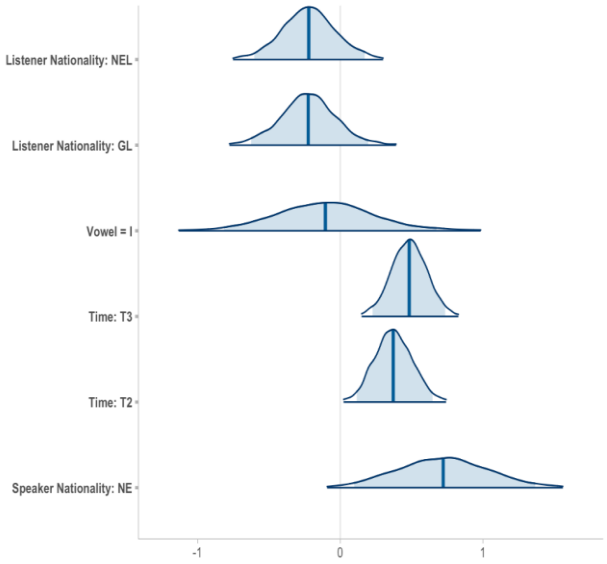


Figure 1: Posterior probability distributions of the model's coefficients.

3.2. Aim 2: To investigate ISIB via native English, Swiss German, and Ecuadorian Spanish listeners' transcriptions of Ecuadorian learners' words.

The effects of vowel and listener nationality, however, were not found to be credible. For vowel, the corresponding measurement error model yields an estimated posterior mean of -0.10, with 95% credible interval [-0.81, 0.67] (with parts of the distribution in the negative and positive sides). For GL, the corresponding measurement error model yields an

[1] Model: correct.vowel ~ speaker_nationality + time + vowel + listener_nationality + speaker_nationality:listener_nationality + time:listener_nationality + vowel:listener_nationality + (time|ID) + (1|word) + (1|listener)

estimated posterior mean of -0.22, with 95% credible interval [-0.62, 0.20]. Similarly, for NEL the corresponding measurement error model yields an estimated posterior mean of -0.22, with 95% credible interval [-0.60, 0.17] though the latter yielded some insightful interactions for speaker nationality and time (see Figs 3a, 3b,) respectively. Figure 3a shows that ECL listeners rated EC speakers slightly higher than the other listeners, but rated NE speakers much lower than the other listeners. Finally, Fig 3b shows that ECL listeners assigned higher rates for /i/ and lower rates for /ɪ/ when compared to the other listeners.

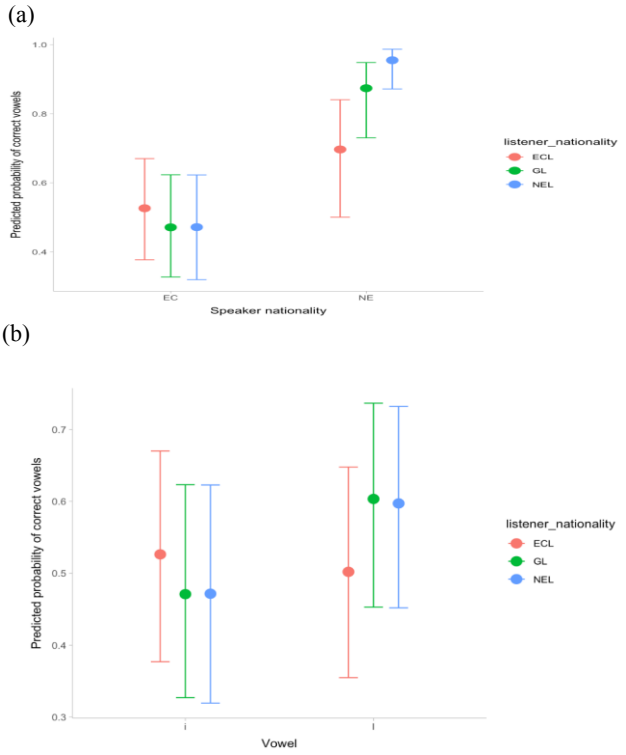


Figure 2: Probabilities predicted by the model.

4. DISCUSSION

Several reports have shown that L2 learners can accurately produce some L2 phonemes over time [4]. In our study, EC speakers started with clear lower proportions of correct vowels and tended to improve as time increased. The biggest improvement was observed from T1 to T2 and with increased probability of correct vowels. One possible explanation for this pronunciation improvement could be, apart from exposure, the positive effect of the phonetics classes, which might lead L2 speakers to establish representations of the vowel categories in long-term memory. From T2 to T3 little increased probability of correct vowel was observed. It could be possible that L2 speakers showed an initial rapid improvement of the phonetic learning which followed a plateau [9]. Previous studies have shown that L1 Spanish speakers have difficulties differentiating tense-lax vowel contrasts. However, in our study, EC speakers seemed to have slightly better proportions for lax [ɪ] than for tense [i].

The second aim was to investigate ISIB via native English, Swiss German, and Ecuadorian Spanish listeners' transcriptions of Ecuadorian and NE speakers' words. In general, a benefit between listeners and speakers who share the same L1 background was slightly observed, with non-native speakers being more intelligible to ECL than to GL and NEL, though with no robust credibility. This finding goes in the direction of Uzun's [12], in which a benefit from an L1 shared between Turkish listeners and speakers was observed on intelligibility through a transcription task. We also found that NE speakers got the lowest intelligibility score by the ECL. This result could be related to the ECL experience since three of the participants did not use English every day compared to the others. Moreover, ECL, GL and NEL differed on how they perceived word intelligibility produced by the EC speakers. ECL could hear more accurately than NE listeners words containing the vowel /i/. However, GL and NEL were more accurate when transcribing words containing /ɪ/. A possible explanation could be the differences between the acoustic information that NEL and GL rely on when listening to NE speech and the information available in the Spanish-accented English speech.

5. CONCLUSIONS

The small listeners' sample size did not allow to conclude an ISIB on talker-listener relationship advantage. However, we consider that our study contributes data to an important line of understandings listeners' perception of word intelligibility on the /i/-/ɪ/ vowel contrast and EC speakers' improvement over time. For future research, it is important to see different listening conditions and different methods of evaluation including a greater number of native and non-native listeners. Moreover, it is difficult to conclude that ECL cannot distinguish well the NE speakers' vowels due to the lower number of NE speakers than the EC speakers. However, this study should be replicated with homogeneous number of native and non-native speakers to observe if a poor vowel intelligibility rating by non-native listeners is the result of the lack of distinguishing tense/lax contrast. Finally, the study should be repeated with L2 speakers' English teachers rating their students' utterances to provide more insights of the familiarity of the L1 speaker's background.

6. REFERENCES

[1] C. Muñoz, "Contrasting effects of starting age and input on the oral performance of foreign language learners," *Appl Linguist*, vol. 35, no. 4, pp. 463–482, Sep. 2014, doi: 10.1093/applin/amu024.

[2] C. Muoz, "Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning," *Appl Linguist*, vol. 29, no. 4, pp. 578–596, Dec. 2008, doi: 10.1093/applin/amm056.

[3] X. Xie and C. A. Fowler, "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English," *J Phon*, vol. 41, no. 5, pp. 369–378, 2013, doi: 10.1016/j.wocn.2013.06.003.

[4] J. E. Flege, O.-S. Bohn, and S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *J Phon*, vol. 25, no. 4, pp. 437–470, Oct. 1997, doi: 10.1006/JPHO.1997.0052.

[5] W. Baker and P. Trofimovich, "Perceptual paths to accurate production of L2 vowels: The role of individual differences," *IRAL - International Review of Applied Linguistics in Language Teaching*, vol. 44, no. 3, pp. 231–250, Nov. 2006, doi: 10.1515/IRAL.2006.010.

[6] J. C. Mora and N. Fullana, "Production and perception of English /i:/-/ɪ/ and /ae/-/ʌ/ in a formal setting: Investigating the effects of experience and starting age," 2007. [Online]. Available: www.icphs2007.de

[7] J. Cebrian, "OLD SOUNDS IN NEW CONTRASTS: L2 PRODUCTION OF THE ENGLISH TENSE-LAX VOWEL DISTINCTION." [Online]. Available: www.icphs2007.de

[8] J. M. Levis, "What Should and Should Not Be Taught," in *Intelligibility, Oral Communication, and the Teaching of Pronunciation*, Cambridge University Press, 2018, pp. 239–264. doi: 10.1017/9781108241564.013.

[9] J. M. Levis, "Changing Contexts and Shifting Paradigms in Pronunciation Teaching," 2005. [Online]. Available: https://www.jstor.org/stable/3588485?seq=1&cid=pdf-

[10] J. E. Flege, M. J. Munro, and I. R. A. MacKay, "Factors affecting strength of perceived foreign accent in a second language," *J Acoust Soc Am*, vol. 97, no. 5, pp. 3125–3134, 1995, doi: 10.1121/1.413041.

[11] P. Lochland, "Intelligibility of L2 speech in ELF," *Australian Journal of Applied Linguistics*, vol. 3, no. 3, pp. 196–212, Dec. 2020, doi: 10.29140/ajal.v3n3.281.

[12] T. Uzun, "The Interlanguage Speech Intelligibility Benefit for Turkish Speakers of English," *Dil Eğitimi ve Araştırmaları Dergisi*, Apr. 2022, doi: 10.31464/jlere.1024616.

[13] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *J Acoust Soc Am*, vol. 114, no. 3, pp. 1600–1610, Sep. 2003, doi: 10.1121/1.1603234.

[14] R. Hayes-Harb, B. L. Smith, T. Bent, and A. R. Bradlow, "The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts," *J Phon*, vol. 36, no. 4, pp. 664–679, Oct. 2008, doi: 10.1016/j.wocn.2008.04.002.

[15] M. J. Munro, T. M. Derwing, and S. L. Morton, "THE MUTUAL INTELLIGIBILITY OF L2 SPEECH," *Stud Second Lang Acquis*, vol. 28, no. 01, pp. 111–131, Mar. 2006, doi: 10.1017/S0272263106060049.

[16] R. C. Major, S. F. Fitzmaurice, F. Bunta, and C. Balasubramanian, "The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment," 2002.

[17] M. J. Munro and T. M. Derwing, "Intelligibility in research and practice: Teaching priorities," in *The Handbook of English Pronunciation*, wiley, 2015, pp. 375–396. doi: 10.1002/9781118346952.ch21.

[18] A. Sewell (2017). "Functional load revisited: Reinterpreting the findings of 'lingua franca'intelligibility studies". *Journal of Second Language Pronunciation*, vol. 3, no 1, pp. 57-79, 2017, doi: 10.1075/jslp.3.1.03sew

[19] M. J. Munro, "On the difficulty of defining "Difficult" in second-language vowel acquisition". *Frontiers in Communication*, *6*. Aug. 2021, doi: 10.3389/fcomm.2021.639398

[20] Boersma, P. Weenink, D. (1992). *PRAAT: Doing phonetics by computer*. (6.1.02) [Computer software]. http://www.praat.org/

[21] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang Learn*, vol. 49, no. PART 1, pp. 285–310, 1999, doi: 10.1111/0023-8333.49. s1.8.

[22] H. Badr. P.Bürkner Paul, (2021) "*bayesian: Bindings for Bayesian TidyModels*.," doi:10.5281/zenodo.6654386, https://hsbadr.github.io/bayesian/.

EFFECTS OF DIALECT EXPOSURE ON THE PROCESSING OF ARGENTINE REGIONAL PHONES BY ENGLISH-SPEAKING L2 LEARNERS OF SPANISH

Lauren B. Schmidt

San Diego State University
lschmidt@sdsu.edu

ABSTRACT

Previous research shows that regional phonetic variation may impede comprehension for both native and L2 listeners, but that experience with regional varieties can lead to changes in the perception of dialectal sounds. This study investigates how exposure to a novel dialect of the target language through a 4-week study abroad immersion experience in Buenos Aires affected the processing of Argentine phonetic variants by English-speaking L2 learners of Spanish. Learners completed a lexical decision task in the first and fourth weeks abroad, which presented Spanish stimuli with Argentine assibilated palatal and aspirated-/s/ variants. Initially, learners were less accurate in recognizing known Spanish words articulated with the Argentine assibilated palatal, and also had slower recognition reaction times for both the assibilated palatal and aspirated-/s/. By the end of four weeks, effects of the assibilated palatal on lexical decision accuracy and processing speed disappeared; however, effects of the aspirated-/s/ on processing speed persisted.

Keywords: L2 speech perception, regional variation, Spanish, study abroad, lexical decision

1. INTRODUCTION

The objective of the current study is to examine how regional phonetic variation in the speech signal affects L2 speech processing, and how exposure to a regional dialect of the target language may mediate such effects over time.

1.1. Effects of dialectal speech on perception and comprehension

Previous research shows that dialectal speech affects comprehension and intelligibility for both native and L2 listeners [e.g., 3, 8, 12, 22], with greater comprehension of some regional or social varieties over others. However, in the case of L2 learning, such dialectal effects have been found to diminish when explicit instruction and awareness drawing to dialectal variants of the target language is provided [20] or learners are exposed to dialectal speech through an immersion or study abroad experience [18, 21].

Dialectal phonetic variants have also been shown to affect performance on perception tasks for both native and nonnative listeners. Several studies document evidence of differences in how regional sounds are identified and processed by native speakers according to the linguistic experiences of the listener, such as the variants produced in one's own speech [7, 22] or the region in which one was raised [6, 2]. A growing body of studies have begun to test how L2 learners categorize and discriminate dialectal sounds in the L2, and these have shown that the learners' perceptual patterns can change according to the varieties exposed to in a study abroad program [1, 4, 5, 19], or via dialectal social contacts [19]. It is still unclear, however, how dialectal speech affects speech processing in a L2, such as in terms of lexical retrieval or processing time, and how such processing effects may or may not change with learner experience with dialectal speech.

1.2. Argentine Spanish phonetic features

Two sociophonetic features characteristic of Buenos Aires Argentine Spanish were targeted in the current study: the assibilated pre-palatal [ʃ] and aspiration [h] of syllable-final /s/.

In many of the varieties and pedagogical models of Spanish that L2 learners are commonly exposed to in the US learning context (for example, Mexican or Castilian varieties), the phonemic category corresponding to orthographic 'y' and 'll' and that category corresponding to orthographic 's' (as well as 'c' before 'e' or 'i' and 'z' in non-Peninsular varieties) are realized as a palatal fricative [ʝ], or as an alveolar fricative sibilant [s], respectively [17]. However, in the River Plate region, including Argentina and Uruguay, 'y' and 'll' are realized as a voiceless or voiced assibilated pre-palatal fricative [ʃ ʒ], with the voiceless variant the preferred form amongst the recent generations in the capital region [17], as in *playa* 'beach' [ˈpla-ʃa]. This variant is specific primarily to the River Plate region, and US-based L2 learners of Spanish may have limited exposure to speakers who use these phones. The second feature, lenition of syllable-final /s/, is

likewise characteristic of Argentine Spanish, but also occurs in the speech of Spanish speakers of many other varieties, including Caribbean and coastal varieties [13] and is variable in nature, subject to social and stylistic factors such as speaker gender or speech style [15]. In Buenos Aires, the most frequently produced lenited form is the aspirated glottal fricative [h] (as in *costa* 'coast' [ˈkoh-ta]), although speakers variably produce the full sibilant [s] and elided forms Ø as well [17], depending on stylistic and sociolinguistic factors.

1.3. Research questions

The current study is guided by the following research questions:

1. What are the effects of Argentine phonetic variants on L2 speech processing, in terms of lexical retrieval and processing time?
2. Does dialect exposure through an immersion experience mediate these effects, and if so, in what ways and to what extent?

2. METHODS

2.1. Participants

Twenty-two English-speaking second language learners of Spanish from various regions of the United States participated in the study during a short-term summer study abroad program in Buenos Aires, Argentina. The learners, aged 19-21, were third- and fourth-year North American undergraduate students enrolled in Spanish language, literature and history courses taught by local Argentine instructors and taken with other international students. They also participated in homestays with Argentine families. The majority of the participants self-rated their Spanish language proficiency as high intermediate or advanced. While most of the learners had prior travel or study experiences in other Spanish-speaking regions and contact with speakers of a variety of other Spanish dialects (primarily in and from Spain, Mexico, Central America, and Spanish-speaking Caribbean countries), the learners had very little to no prior exposure to Argentine Spanish. Only 4 reported previous contact with speakers of Argentine Spanish, and none had previously traveled to Argentina.

2.2. Stimuli and procedure

Participants completed a lexical decision task (LDT) within the first week of arrival in Buenos Aires, and again after the fourth week of their program. They heard common disyllabic Spanish words as well as invented words that followed Spanish phonotactic patterns, and responded as quickly as possible as to whether the item heard was an existing Spanish word or not. Following the LDT, subjects completed a Vocabulary Familiarity Task to ensure familiarity with all of the real word task items, as well as a Language Background Questionnaire.

A total of 189 items were presented in the LDT, balanced across nonwords (n = 99) and real words (n = 90). All target and distractor stimuli were produced by a female, university-educated speaker from Buenos Aires living in the U.S., with target items articulated with either those variants learners were previously most exposed to in the North American L2 classroom – the palatal fricative [ʝ] and sibilant [s] –, or with the Argentine regional phones – the assibilated palatal [ʃ] and aspiration [h] (see Table 1).

Feature	Real word	Nonword
dis-tractor	*leche* 'milk' (n = 40)	*dalpa* (n = 59)
[s] variant	*tri[s]te* 'sad' (n = 13)	*ga[s]po* (n = 10)
[h] variant	*chi[h]te* 'gossip' (n = 13)	*bi[h]ca* (n = 10)
[ʝ] variant	*toa[ʝ]a* 'towel' (n = 12)	*dre[ʝ]o* (n = 10)
[ʃ] variant	*pla[ʃ]a* 'beach' (n = 12)	*me[ʃ]o* (n = 10)

Table 1: Examples of the LDT Spanish stimuli.

2.3. Analysis

Lexical judgments and RTs were recorded. Judgments were coded as correct if the participant accurately accepted a Spanish real word or accurately rejected a nonword. Any lexical items not marked as known in the Vocabulary Familiarity Task were excluded from the analysis as were data that exceeded three *SD*'s in either direction from an individual's mean RT. Lexical judgment accuracies (0-100%) of the real word stimuli data were then calculated for each participant according to feature and test time by dividing the number of correct judgments by the total number of judgments for each feature at each test time. Mixed model analyses were then conducted for both the variants of Spanish 's', 'ce/ci', and 'z' (here-on referred to as the 'S context') and for the variants of Spanish 'y' and 'll' (the 'Y context'), with word judgment accuracies and reaction times set as dependent variables, phonetic feature and test time as fixed effects, and subject as a random effect.

3. RESULTS

3.1. Lexical decision accuracies

Mean recognition accuracies of the Spanish real word stimuli according to phonetic feature and test time are presented in Table 2. Table 3 shows the results of the mixed models run for each of the two contexts.

Context	Variant	Pretest	Posttest
S	[s]	.94 (.077)	.95 (.069)
	[h]	.93 (.098)	.91 (.110)
Y	[ʝ]	.95 (.084)	.94 (.096)
	[ʃ]	.89 (.097)	.96 (.074)

Table 2: Mean LDT accuracy rates for acceptance of real word stimuli by test time and variant (*SD*).

There were no significant main effects nor interactions for feature or test time in the statistical model for the S context, which indicated that the regional [h] variant had no effect on the learners' abilities to accurately identify known Spanish lexical items, even at the beginning of the study abroad program (during week 1).

However, the mixed model for the Y context revealed a different picture: an interaction between feature and test time, approaching significance. First, Bonferroni posthoc comparisons showed that at time 1, learners were significantly worse at recognizing words realized with Argentine [ʃ] (89%) than those articulated with the previously familiar [ʝ] variant (95%) (p = .036). By time 2, however, this difference disappeared (i.e., was no longer significant; p = .575). Moreover, within category comparisons further showed that learners became significantly more accurate in identifying known Spanish words with the [ʃ] variant at time 2 (96%) when compared to time 1 (89%) (p = .020). No such differences were found for lexical recognition accuracies of the [ʝ] variant across time (p = .747).

Context	Effect	*F (df)*	*p*
S	feature	2.012 (1, 84)	.160
	test time	.795 (1, 84)	.795
	feature*time	.271 (1, 84)	.604
Y	feature	1.232 (1, 84)	.270
	test time	2.095 (1, 84)	.151
	feature*time	3.632 (1, 84)	.060(*)

Table 3: Results of mixed models for LDT accuracy.

3.2. Reaction time latencies

Table 4 presents the RT latencies for the identification of the Spanish real word items across feature and time. Results of the mixed models for RTs for each of the two contexts are also shown below in Table 5.

Context	Variant	Pretest	Posttest
S	[s]	1.33 (.199)	1.28 (.191)
	[h]	1.43 (.193)	1.38 (.217)
Y	[ʝ]	1.32 (.155)	1.26 (.158)
	[ʃ]	1.54 (.306)	1.32 (.165)

Table 4: Mean RTs in seconds by test time and variant (*SD*).

For the S context, a significant main effect of phonetic feature was found for learner RTs. The learners were significantly slower overall in making lexical judgments when the words were articulated with the regional aspirated [h] variant (M = 1.40 sec., SD = .205) than with the sibilant [s] variant (M = 1.30 sec., SD = .194) (p = .021). This effect was constant across test times.

The statistical model for the Y context found significant main effects of both phonetic feature and test time, and a significant interaction between the two. Bonferroni posthoc comparisons revealed that only at time 1 did the L2 learners make lexical decisions significantly slower in the presence of the Argentine [ʃ] variant (1.54 sec.) as opposed to the previously familiar [ʝ] variant (1.32 sec.) (p < .001). By time 2, there was no significant difference in processing speed between the two variants (p = .378). Furthermore, the posthoc analyses of the within category comparisons showed no change from time 1 to time 2 in the response speed for the identification of the [ʝ] items (p = .359), while they revealed a significant increase in processing speed (i.e., faster RTs) for the identification of the [ʃ] items from time 1 (1.54 sec.) to time 2 (1.32 sec.) (p < .001).

Context	Effect	*F (df)*	*p*
S	feature	5.517 (1, 84)	.021*
	test time	1.343 (1, 84)	.250
	feature*time	.010 (1, 84)	.919
Y	feature	9.559 (1, 84)	.003**
	test time	9.983 (1, 84)	.002**
	feature*time	3.382 (1, 84)	.069(*)

Table 5: Results of mixed models for RTs.

4. DISCUSION AND CONCLUSIONS

This study examined how previously unfamiliar regional dialectal variation affects processing of speech for second language learners, and how exposure through a month-long study abroad immersion experience in the target dialectal region may mediate such effects of dialectal phonetic variation on L2 speech processing.

Prior to extensive exposure to Argentine Spanish (during week one of arrival to Buenos Aires), the L2 learners' abilities to accurately comprehend and retrieve known Spanish words were impeded by the presence of one of the dialectal phones, of the assibilated pre-palatal fricative [ʃ]. The presence of the [ʃ] phone also resulted in greater processing costs in terms of longer processing times to make lexical decisions. The second regional feature studied, aspiration of syllable-final /s/, did not similarly impede lexical retrieval (lexical judgments), but [h] did likewise result in longer processing times for the learners in making lexical decisions at week one.

Interestingly, after only four weeks living and studying in Argentina, the processing effects of the Argentine [ʃ] variant observed initially on lexical retrieval and processing speed completely disappeared: that is, after one month, learners were equally accurate and equally fast at identifying known Spanish words realized with the Argentine dialectal [ʃ] phone as with the previously familiar [j] variant. This was not the case, however, for the dialectal lenited [h] phone, which, still at week four, continued to cause slowed processing as compared to the processing of the previously familiar (and more acoustically robust) full sibilant form [s].

A possible explanation for the disparity observed across the two regional variants is the role of perceptual saliency of the two phones, which, in turn influences attention, which then influences the encoding and representing of new episodic instances into the L2 mental lexicon upon exposure to the dialectal forms. Under an Exemplar Model approach [9, 11], listeners are understood to encode acoustically detailed and socially and lexically indexed information from the speech signal in cognitive representations of speech [11, 16]. Speech is then recognized by comparisons with these stored exemplars [11]. Pierrehumbert [16] stresses, however, that in order for exposure be "effective" in the encoding and storing of this detailed information, cognitive factors such as attention and memory are also at play. That is, those episodic events that are more attended to are more likely to be remembered [16]. As [ʃ] is a distinctive (phonemic) feature in the L1, English (albeit associated with a different phonetic category), this phone should be particularly perceptually salient to this learner group, and thus – while it may be more disruptive in speech processing upon initial exposure (time one) – this saliency also allows for greater awareness and drawing of attention to the form during exposure, eventually allowing for greater effects for the encoding and remembering of episodic instances of the new variant stored with lexical information over time. On the other hand, as shown previously in [19], Spanish syllable-final aspiration [h] is often 'ignored' by English-speaking learners of Spanish, interpreted as part of the preceding vowel. As such, the presence of this phone may be less disruptive in terms of impeding lexical recognition (recognition accuracies). However, the learners are more likely to have less awareness and less attention drawn to the syllable-final [h] phone, and this could result in less "effective exposure" and less opportunities for changes to the cognitive system even after a month of exposure. This could explain the effects of dialectal [h] on processing speed persisting into time two. Future research is needed to further explore the role of attention and awareness – for example testing effects of explicit instruction of and awareness drawing toward dialectal phones – in the processing of dialectal forms in the L2.

5. REFERENCES

[1] Baker, W., Smith, L. C. 2010. The impact of L2 dialect on learning French vowels: Native English speakers learning Quebecois and European French. *The Canadian Modern Language Review* 66, 711–738.

[2] Delgado-Díaz, G., Galarza, I. 2015. ¿Qué comiste [x]amón? A closer look at the neutralization of /h/ and posterior /r/ in Puerto Rican Spanish. In: Willis, E. W., Butragueño, P. M., Herrera Zendejas, E. (eds), *Selected Proceedings of the 6th Conference on Laboratory Approaches to Romance Phonology*. Cascadilla, 70–82.

[3] Eisenstein, M., Berkowitz, D. 1981. The effect of phonological variation on adult learner comprehension. *Studies in Second Language Acquisition* 4, 75–80.

[4] Escalante, C. 2018. ¡Ya pué[h]! Perception of coda-/s/ weakening among L2 and heritage speakers in coastal Ecuador. *E-JournALL* 5, 1–26.

[5] Escudero, P., Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26, 551–585.

[6] Evans, B. G., Iverson, P. 2004. Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America* 115, 352–361.

[7] Evans, B. G., Iverson, P. 2007. Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America* 121, 3814–3826.

[8] Floccia, C., Girard, F., Goslin, J., Konopczynski, G. 2006. Does a regional accent perturb speech

processing? *Journal of Experimental Psychology: Human Perception and Performance* 32, 1276–1293.
[9] Goldinger, S. D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*, 1166–1183.
[9] Hualde, J. I. 2005. *The Sounds of Spanish*. Cambridge.
[11] Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K. & Mullenix, J. (eds), *Talker variability in speech processing*. Academic Press, 145–166.
[12] Kaisse, E. M. 1996. The prosodic environment of s-weakening in Argentinian Spanish. In: Zagona, K. (ed), *Grammatical Theory and Romance Languages Vol. 13*. John Benjamins, 123–143.
[13] Lipski, J. 1994. *Latin American Spanish*. Longman.
[14] Major, R. C., Fitzmaurice, S. M., Bunta, F., Balasubramanian, C. 2005. Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning* 55, 37–69.
[15] Núñez-Méndez, E. 2022. Variation in Spanish /s/: Overview and new perspectives. *Language* 7.
[16] Pierrehumbert, J. B. 2006. The next toolkit. *Journal of Phonetics 34*, 516–530.
[17] Rohena-Madrazo, M. 2015. Diagnosing the completion of a sound change: Phonetic and phonological evidence for /ʃ/ in Buenos Aires Spanish. *Language Variation and Change* 27, 287–317.
[18] Schmidt, L. B. 2009. The effect of dialect familiarity via a study abroad experience on L2 comprehension of Spanish. In: Collentine, J., García, M., Lafford, B., Marín, F. M. (eds), *Selected Proceedings of the 11th Hispanic Linguistics Symposium*. Cascadilla: 143–154.
[19] Schmidt, L. B. 2018. L2 development of perceptual categorization of dialectal sounds: A study in Spanish. *Studies in Second Language Acquisition 40*, 857–882.
[20] Schoonmaker-Gates, E. 2017. Regional variation in the language classroom and beyond: Mapping learners' developing dialectal competence. *Foreign Language Annals 50*, 177–194.
[21] Schoonmaker-Gates, E. 2018. Dialect comprehension and identification in L2 Spanish: Familiarity and type of exposure. *Studies in Hispanic and Lusophone Linguistics* 11, 193–214.
[22] Sumner, M., Samuel, A. G. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory & Language 60*, 487–501.

L1 EFFECTS ON NAIVE PERCEPTION AND PRODUCTION OF [ʉ] BY MANDARIN AND ITALIAN SPEAKERS

Xuanda Chen[1,3] Meghan Clayards[1,2,3] Heather Goad[1,3] Donghyun Kim[4]

[1]Department of Linguistics, [2]School of Communication Sciences and Disorders,
[3]CRBLM, McGill University [4]Kumoh National Institute of Technology, South Korea
{xuanda.chen, meghan.clayards, heather.goad}@mcgill.ca, heydonghyun@gmail.com

ABSTRACT

Accurately predicting the difficulty of learning novel sounds based on linguistic experience is a challenge. This paper explores the effectiveness of two approaches, the spatial and dimension approaches, in predicting naive perception and production of novel sounds. The spatial approach measures the gradient acoustic-phonetic distance, while the dimension approach leverages discrete sound constructs. We investigate Mandarin and Italian speakers' performance on discrimination and imitation tasks involving the novel sound [ʉ]. The results show that Mandarin speakers discriminate [ʉ] better than Italian speakers, but both groups have similar performance in imitation. Our findings suggest that the dimension approach is more effective in capturing early learning difficulty of novel sounds, and that listeners perform better when the dimensions used to construct the novel sounds are contrastive in their native language.

Keywords: perception; imitation; similarity

1. INTRODUCTION

Cross-linguistic studies revealed that linguistic experience strongly conditions listeners' perception and production of novel sounds [1]. Novel sounds have been observed to range in difficulty for speakers of different languages. For example, it is well-known that Japanese speakers have trouble discriminating between English /l/ and /r/ [2], but Chinese speakers do not [3]. Thus, given the sound structure of one's native language, what predictions can we make about which novel speech sounds will be hard to perceive and produce? A large number of studies have appealed to the notion of 'similarity' to answer this question. A novel sound should cause difficulties when it is similar to existing L1 sounds [4]. However, how 'similarity' between novel sounds and L1 sounds should be defined is still a matter of debate. This paper compares two different approaches to quantifying sound similarity: the spatial approach and the dimension approach.

The spatial approach, exemplified by the Speech Learning Model (SLM), defines similarity between sounds based on their acoustic-phonetic distance. SLM predicts that a greater distance between an L2 and the nearest L1 sound results in easier acquisition of the L2 sound [4]. The SLM makes the correct predictions for Japanese listeners' acquisition of the English /l/-/r/ distinction [5]. While the SLM makes no prediction for naive learners, if the SLM's prediction holds at the onset of learning, sounds that are closer in acoustic-phonetic distance to L1 sounds will be harder to perceive and produce than those that are further away.

The dimension-based approach quantifies sound similarity in cross-language perception by using dimensions to classify sounds. These dimensions, such as vowel backness or roundedness, describe how a language contrasts pairs of phonemes. Distinctive feature theory [6] shares similar concepts, but unlike features, dimensions do not aim to capture a language's phonological processes.

These dimensions can be expressed acoustically, such as using F3 (e.g. [7]), or articulatorily, like lip rounding (e.g. [8]). We assume that if a dimension is contrastive, it can be dissociated from other dimensions and repurposed to build new sounds (cf. e.g. [3, 9]). Thus, speakers of that language should be better able to perceive and produce novel combinations using that dimension than speakers of a language that does not use that dimension contrastively.

The two approaches predict naive perception and production based on different aspects of linguistic experience. For example, Mandarin has three high vowels, /i/, /u/ and /y/, while Italian only has two, /i/ and /u/. The phonetic realization of the two peripheral vowels is similar in the two languages, with the only difference being the presence or absence of front rounded /y/. Suppose that Mandarin and Italian speakers are presented with the novel high central rounded vowel [ʉ]. Which language group will have better naive perception and production?

Under the spatial approach, the crucial distance metric for Mandarin is [y]-[ʉ], while for Italian, it is [u]-[ʉ]. The acoustic distance in Mandarin is smaller than in Italian, and thus, Italian speakers, due to the greater distance between [ʉ] and its nearest L1 sound, should have better naive perception and production than Mandarin speakers. The dimension approach makes the opposite prediction. Italian does not have contrasts in roundedness (rounding is bound to the backness dimension). By contrast, Mandarin speakers have a /y/-/i/ contrast, meaning that the roundedness dimension is contrastive in this language. Thus, Mandarin speakers should have better naive perception and production of [ʉ] because contrastive roundedness is available to capture [ʉ].

Our research question is which approach will be borne out? We use an AX discrimination task to assess perception and an imitation task to assess production. We use an ABX identification task to confirm that the crucial distance metric is different between Mandarin and Italian, and it can be used to test the spatial approach's assumptions.

2. METHODS

2.1. Participants

We recruited 27 Mandarin speakers (mean age = 22.5, standard deviation = 3.14) and 27 Italian speakers (mean age = 26.2, standard deviation = 6.60). All participants were monolingual speakers of either Mandarin or Italian. None reported any speech or hearing impairments.

2.2. Stimuli

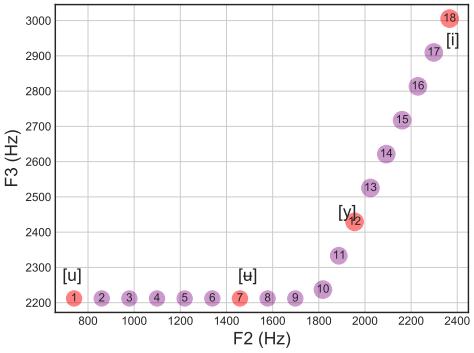


Figure 1: The high vowel stimuli.

The stimuli were 60 msec isolated high vowels on the [i]-[y]-[u] continuum following [10], synthesized in Praat [11]. We first selected four sound steps from the continuum to represent the categories of [i], [y], [ʉ], and [u] (see Figure 1). F1 was 300 Hz for all stimuli. F2 values were manually determined, and F3 values were automatically calculated using the equation $F3 = 1.4 * (F2 - 220)$ [12]. Two endpoints, representing [i] (step 18) and [u] (step 1) categories, were selected based on the estimation of mean values for Mandarin and Italian prototypical [i] and [u] [13, 14]. We chose [y] (step 12) based on the judgment of the first author, a native speaker of Mandarin, and [ʉ] (step 7) based on the first recording of [ʉ] on the course webpage [8], available at https://linguistics.berkeley.edu/

acip/course/ipachart/vowels.html. Equidistant steps were created to fill in the gaps between categories for a total of 18 stimuli. The stimuli and related files are available at the project page https://osf.io/7k2j8/?view_only=a26c2961e5cf4d6a99ccf50c23bceee3.

2.3. Procedure

The experiment was performed using an online experiment service implemented in jsPsych [15]. Participants accessed the experiment in their own homes and used their own recording devices and headphones (they had to pass a headphone test before the main experiment [16]). They received written instructions about the experiment in their native language, and were required to give responses by pressing buttons on the keyboard. Each participant completed three tasks, and each task contained a practice session.

ABX Task Participants had to decide whether the last sound they heard (sound steps between 1 and 12) was the same as the first ([y] step 12) or the second ([u] step 1) [17]. Participants were tested on 40 trials using stimuli from the [y]-[ʉ]-[u] range, with an inter-stimulus interval (ISI) of 1500 msec [18].

AX Task Participants had to determine whether two stimuli in a trial were the same or different, with an ISI of 1500 msec. We tested all possible different sound pairs within 9 steps (e.g., step 2-10 but not step 2-11), and 36 same pairs (e.g., step 1-1) for a total of 144 trials in three blocks.

Imitation Task Participants were instructed to produce a sound as close as possible to the presented stimulus. 18 stimuli were presented 4 times in random order. In each trial, the stimuli were played twice, with an ISI of 1500 msec. We used the default sampling rate of 22.05 kHz during recording.

2.4. Measurements

For the ABX task, we used a logistic mixed-effect regression to model the response pattern, i.e., to determine if [ʉ] was categorized as an instance of [y] or [u] [19]. The normalized predictors included continuous VOWELS and categorical LANGUAGE. The model included random intercepts for participants with p-values calculated by *lmerTest* [20]. For the AX task, goodness of sound discrimination was measured as the average accuracy of response to a sound pair for each stimulus across languages. We performed ANOVA tests to examine the difference in [ʉ] accuracy across the two languages. For the imitation task, each imitation sample was manually measured at the temporal midpoint to obtain formant values (Hz) in Praat. Goodness of imitation was measured as the Euclidean distance between production and target. Imitation distance was normalized across participants with the Lobanov method [21]. We also performed ANOVAs to examine the difference in imitation in [ʉ] across the two languages.

3. RESULTS

Due to limited space, we will focus on results for [ʉ] only. Figure 2 shows GLM fits to the response data from the ABX task. Step

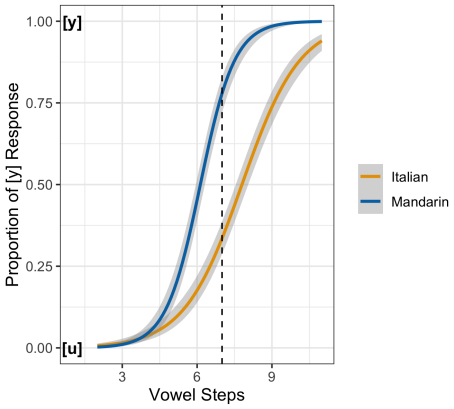


Figure 2: Vowel categorization

7, [ʉ], is marked with a dashed line. The model found that participants categorized the stimuli more often as [y] as the stimuli shifted from [u] to [y] [$\hat{\beta}$=7.154, σ=0.377, p<0.005] and that Mandarin speakers were more likely to categorize the stimuli as [y]

than Italian speakers [$\hat{\beta}$=2.058, σ=0.344, p<0.005]. From the plot we see that [ʉ] is categorized as [y] more than 50% of the time for Mandarin speakers and as [u] more than 50% of the time for Italian speakers. The results confirm that the crucial distance metric is different for Mandarin and Italian participants, and that [ʉ] is closer to a native sound for Mandarin than for Italian listeners.

Figure 3 provides boxplots for the discrimination data from the AX task and imitation accuracy for the two languages. The ANOVA test showed that Mandarin participants have higher discrimination accuracy for [ʉ] than Italian participants [F(1, 52)=8.812, p=0.004]. The ANOVA also found that the two groups of speakers do not significantly differ from each other [F(1, 192)=0.002, p=0.965] in imitation.

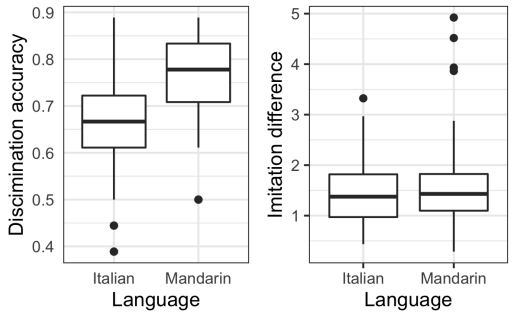


Figure 3: Discrimination and imitation of [ʉ]

4. DISCUSSION AND CONCLUSION

The goal of the present study was to test which of the two different approaches to calculating similarity makes correct predictions for perception and production on first exposure to a non-native sound, [ʉ]. The categorization results show that Mandarin participants tend to categorize [ʉ] as /y/, while Italian participants tend to categorize it as /u/. This confirms the assumption that the crucial distance metric is [u]-[ʉ] for Italian and [y]-[ʉ] for Mandarin, allowing us to test the spatial approach (gradient acoustic-phonetic distance). Mandarin speakers have higher discrimination accuracy than Italian speakers, which is not predicted by the spatial approach but supports the dimension approach. The imitation results do not support one approach over the other.

Our results suggest that acoustic-phonetic distance may not be sufficient to explain all sound similarity problems. On first exposure to a novel sound, speakers may benefit from a contrastive dimension in their language to perceive a novel sound that involves that dimension, though this is not guaranteed to facilitate production. This suggests that speakers learn representations for sounds with discrete parts. Similar ideas can be found in the dimension-based perceptual interference account [7] and in the Redeployment Hypothesis (RH) [9]. The RH states that a phonological feature [F] can be redeployed from the L1 grammar to represent a non-native contrast by combining with another feature [G] to form [F, G]. Both the dimension approach and the RH suggest that success in naive perception depends on speakers' knowledge of sound constructs and their ability to reuse these components of sounds, either dimensions or phonological features.

In conclusion, we have argued that the impact of linguistic experience on perceiving and producing a novel sound can be characterized as experience in using particular dimensions in L1. If the novel sound includes a contrastive dimension in L1, perceiving a novel sound combination using that dimension on first exposure may be less difficult. Conversely, perceiving it may be more challenging if the dimension is absent. The imitation results are particularly intriguing as they indicate that neither approach provides accurate predictions. It is possible that additional factors, such as retrieval of incorrect motor plans, influence the imitation of novel sounds. Future research should investigate the underlying reasons for these findings, as well as the connection between perception and production.

5. REFERENCES

[1] K. Miyawaki, J. J. Jenkins, W. Strange, A. M. Liberman, R. Verbrugge, and O. Fujimura, "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," *Perception & Psychophysics*, vol. 18, no. 5, pp. 331–340, 1975.

[2] H. Goto, "Auditory perception by normal Japanese adults of the sounds 'L' and 'R'," *Neuropsychologia*, 1971.

[3] C. Brown, "The interrelation between speech perception and phonological acquisition from infant to adult," *Second language acquisition and linguistic theory*, vol. 1, pp. 4–64, 2000.

[4] J. E. Flege, "Assessing constraints on second-language segmental production and perception," *Phonetics and phonology in language comprehension and production*, vol. 6, pp. 319–355, 2003.

[5] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/," *Journal of Phonetics*, vol. 32, no. 2, pp. 233–250, 2004.

[6] N. Chomsky and M. Halle, *The sound pattern of English*. New York: Harper and Row, 1968.

[7] P. Iverson, P. K. Kuhl, R. Akahane-Yamada, E. Diesch, A. Kettermann, C. Siebert *et al.*, "A perceptual interference account of acquisition difficulties for non-native phonemes," *Cognition*, vol. 87, no. 1, pp. B47–B57, 2003.

[8] P. Ladefoged and K. Johnson, *A course in phonetics [6th ed.]*. Boston, Massachusetts: Wadsworth Publishing, 2010.

[9] J. Archibald, "Second language phonology as redeployment of phonological features," *Canadian Journal of Linguistics*, vol. 50, no. 1-4, pp. 1000–1030, 2005.

[10] B. L. Rochet, "Perception and production of second-language speech sounds by adults," *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*, pp. 379–410, 1995.

[11] P. Boersma, *Praat: doing phonetics by computer*, 2018. [Online]. Available: http://www.praat.org/

[12] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2088–2113, 1989.

[13] P. M. Bertinetto and M. Loporcaro, "The sound pattern of standard Italian, as compared with the varieties spoken in florence, milan and rome," *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 131–151, 2005.

[14] J. Yang, *Acoustic properties of vowel production in Mandarin-English bilingual and corresponding monolingual children*. The Ohio State University, 2014.

[15] J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, 2015.

[16] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[17] W. Strange, V. L. Shafer *et al.*, "Speech perception in second language learners: The re-education of selective perception," *Phonology and Second Language Acquisition*, vol. 36, pp. 153–192, 2008.

[18] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Perception & Psychophysics*, vol. 37, no. 1, pp. 35–44, 1985.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[21] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, 1971.

NEURAL PREDISPOSITION FOR L2 STRESS LEARNING IN FRENCH AND GERMAN LISTENERS

Sandra Schwab[1,2], Lea Jost[1], Michael Mouthon[1], Faustine Etter[3], Julie Kamber[2], Lars Rogenmoser[4] & Jean-Marie Annoni[1]
University of Fribourg[1]; University of Bern[2]; University of Zurich[3]; Friedrich Schiller University Jena[4]
{Sandra.Schwab; Julie.Kamber}@unibe.ch; {Lea.Jost; Michael.Mouthon; Jean-Marie.Annoni}@unifr.ch; Faustine.Etter@uzh.ch; Lars.Rogenmoser@uni-jena.de

ABSTRACT

The present research examines the neural predisposition for L2 stress learning, and investigates whether this predisposition is driven by the learners' native language (free- vs. fixed-stressed). fMRI and behavioral experiments were run with French- and German-speaking listeners. Participants performed stress-related tasks in Spanish (an unknown language), and underwent a perceptual training on Spanish stress contrasts. Results showed a different relationship between the training effect and the degree of neural activation for both groups. German listeners with larger activation in fronto-temporal areas were more successful in learning L2 stress, whereas such relationship could not be observed for French listeners.

Keywords: L2 stress perception, neural predisposition, language-dependent, fronto-temporal regions.

1. INTRODUCTION

Besides cognition- and rhythm-related factors (e.g., working memory or phonological awareness; musical expertise; [1], [2]), the listeners' native language (L1), and more specifically, its stress properties, have a considerable influence on the listeners' ability to process stress in a foreign language ([1], [4], [5]). Native speakers of languages with predictable (also called 'fixed') stress (e.g., French) have been shown to experience larger difficulties in discriminating lexical stress contrasts in a foreign/second language (hereafter 'L2') (e.g., en. import versus import; es. número versus numero)[1] than speakers of languages with variable (also called 'free') stress (i.e., English, German) ([3], [4]).

With the specific goal to improve the listeners' ability to perceive accurately L2 sounds or contrasts, perceptual training is used for learning not only segmental, but also suprasegmental features. Importantly for the present study, recent research has shown that learners were able to partly overcome their L2 stress detection difficulties thanks to specific perceptual training, although with a certain amount of interindividual variability ([6]). Besides the behavioral factors that might influence L2 stress perception and learning ability, the existence of neural predisposition for prosodic learning is also conceivable.

To date, neuroimaging studies on a possible neural predisposition for prosodic learning are still very scarce. However, there seems to be evidence that working memory, encompassing its two key components phonological and executive working memory, is a strong predictor of language aptitude ([7], [8]). Furthermore, neuroimaging studies show that phonetic learning performance is associated with activation in the inferior frontal gyrus (IFG) and temporal areas, including the superior temporal gyrus (STG) ([9], [10], [11]), which seems to be involved in high-order auditory processing of speech ([12]). Of particular interest for the present study is the research carried out by [11] on L2 tone perception, showing brain activation differences in superior temporal regions even before training between 'successful' (i.e., participants who profited more from training) and 'less successful' learners.

The present research is the first attempt to examine the neural predisposition for L2 stress learning, and to investigate whether this predisposition is driven by the learners' native language (free- vs. fixed-stressed). The question to be answered is whether, as previously shown for L2 tone learning, there is a relationship between the training effect and the activation in frontal and temporal brain areas during L2 stress processing, and to what extent this relationship depends on the learners' native language.

2. METHOD

2.1. Participants

Two groups of participants took part in the study. Forty-six participants were native speakers of French (hereafter 'FR') aged between 18 and 28 years (M = 23.1, SD = 2.41; 27 females). Thirty-five participants were native speakers of (Swiss) German (hereafter 'DE') aged between 19 and 28 years (M =

[1] The underlined syllable in these examples, and in the rest of the paper, corresponds to the stressed syllable.

22.6, *SD* = 2.34; 26 females). All participants were students from Fribourg or Bern universities. They had no knowledge of Spanish, Italian, or Portuguese (i.e., free-stress romance languages). Bilingual participants, who learned another language than French or (Swiss) German before the age of 6, were not included. However, since French, German and English are mandatory disciplines in the Swiss educational system. DE had school knowledge of French and English and FR school knowledge of German and English. Participants were paid for their participation. The study was approved by the Ethics Committee of the Psychology Department of University of Fribourg.

2.2. Experimental design

Participants took part in four experimental sessions (see Figure 1). In sessions 1 and 2, they performed an fMRI experiment and a behavioral experiment (i.e., pretest), respectively. In session 3, participants were administered a perceptual training on Spanish stress contrasts. In session 4, they performed the behavioral experiment again (i.e., posttest).

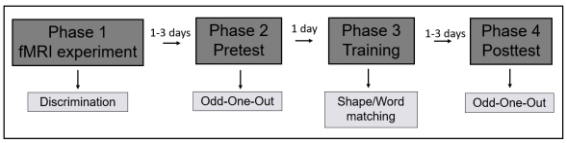


Figure 1: Experimental design.

2.3. fMRI experiment

2.3.1. Material and procedure

In the fMRI experiment, participants performed the discrimination task described in [13]. Participants were presented with pairs of auditory trisyllabic Spanish words produced by a female native speaker of Castilian Spanish. They were asked to indicate whether the two Spanish words were the same or different. In the 'different' pairs, items differed either in the final vowel of the word (e.g., valoro versus valore) or in the position of word stress (e.g., valoro versus valoró). In the 'vowel' deviating pairs, one of the items was always the 1sg present indicative of the verb (e.g., valoro, en. I value) and the other item the 1sg or 3sg present subjunctive of the same verb (e.g., valore, en. that I/he/she value). The 'stress' deviating pairs were composed of the 1sg present indicative of the verb (e.g., valoro, en. I value) and 3sg simple past tense of the same verb (e.g., valoró, eng. he/she valued). The usage of identical verbs for the vowel and stress conditions ensured similar lexical properties across both conditions (bypassing the potential confounding factors of lexical frequency, phonological neighborhood and cognateness). The vowel and stress conditions were presented alternatively using a block design. The vowel condition served as a control condition to be compared with the stress condition.

2.3.2. Data acquisition and pre-processing

MRI data collection included two fMRI runs of 9 minutes that were acquired successively with a 3T GE scanner (acquisition details: TR = 2000 ms, TE = 30 ms, flip angle = 85°, voxel size = 2.3 x 2.3 x 3 mm). MRI data preprocessing and analyses were conducted with SPM12 (UCL, London). Functional images were preprocessed with a standard pipeline (slice timing, spatial realignment and unwarping, co-registration of anatomic scan on fMRI, Unified-segmentation of anatomic scan, normalization, smoothing). The resulting images were analyzed at the individual subject level using a general linear model (GLM). fMRI signal was modelized as condition-specific block of 22.63s of duration convolved with the hemodynamic response function (HRF). A high-pass filter with a 1/128 Hz threshold was applied at time series to remove low frequency noise and signal drifts and an autoregressive function (AR(1)) was implemented to correct temporal correlations between neighboring voxels.

2.3.3. Data analysis

The contrast between Stress and Vowel blocks was sent to a one-sample t-test to study the difference of brain activity between the two conditions. Based on previous research ([13], [14], [15]), brain activation differences were extracted in the following regions of interest: left and right inferior frontal gyrus (IFG) as well as left and right superior temporal gyrus (STG). The higher the value, the larger the activation difference between Stress and Vowel.

2.4. Behavioral experiment

2.4.1. Pre-/posttests

In the pre- and posttests, participants performed the Odd-One-Out task described in [6]. They heard trials of three trisyllabic Spanish words (e.g., validó-validó-valido) and had to indicate which of the three words differed regarding the stress pattern. The pretest included 216 trials comprising different degree of talker and intonation variability (see [6] for details), while the posttest included the same trials as in the pretest as well as 108 additional trials composed of novel words produced by two novel female speakers. For each participant, percent correct for pre-, posttest as well as for novel items was collected, and the training effect was calculated by subtracting pretest percent correct to posttest percent correct. We also computed the difference

between correct percent at posttest with novel items and pretest percent correct.

2.4.2. Training

Training focused only on the perception of stress contrasts in Spanish. We used the non-explicit training developed by [6] that was divided into eight 30-minute sessions and administered with Praat ([16]) over two weeks, at the participants' place. Six Spanish words, produced by two Castilian Spanish female speakers, were used: two first-syllable (i.e., cáscara and género), two second-syllable (i.e., cascara and genero) and two third-syllable stressed words (i.e., cascará and generó). We also created six shapes (similar to Tetris shapes), each one associated to each of the six words. It is important to mention there was no direct relationship between the stress patterns of the words and the shapes (i.e., a word with stress on the first syllable was not necessarily associated with a shape with a 'peak' in its left part). Without receiving previous metalinguistic explanation about Spanish accentual system, the participants performed a shape/word matching task during the entire duration of the training. In this task, they heard one of the six Spanish words and four shapes appeared on the screen. They had to click on the shape they thought corresponded to the word they heard. After giving their response, they received feedback: they heard the word again and only the correct shape stayed on the screen. The feedback allowed the participants to learn the correspondence between the words and the shapes. The outcome measures of the training were not further analyzed.

2.5. Data analysis

Unpaired t-tests were used to test the difference between FR and DE regarding training effect, and neural activation in the four brain regions under study. One-sample t-tests were used to test the training effect on novel trials in FR and DE. Finally, correlations were run separately for FR and DE between training effect and neural activation in each region of interest.

3. RESULTS

3.1. Effect of L1 on training effect with validation

As shown in Figure 2, training effect did not differ between FR and DE (M = 8.036, SD = 5.872 and M = 7.865%, SD = 7.068, respectively; t(79) = 0.115, p = .908), although FR performed more poorly than DE at pretest (M = 50.936%, SD = 15.862 and M = 74.643%, SD = 9.834 respectively; t(79) = -8.262, p <.001). The difference between percent correct at pretest and at posttest with novel items was significantly different from 0 for FR (M = 11.051; SD = 8.828; t(45) = 8.490, p < .0001) and for DE (M = 9.166; SD = 87.355; t(34) = 7.373, p < .0001) This indicates that FR and DE's performance improved after training even on items that were new to the listeners, and thus validates the efficiency of the training method.

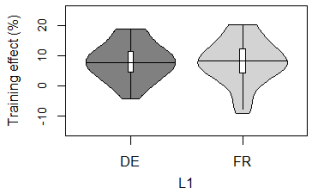


Figure 2: Training effect (%) as a function of L1.

3.2. Effect of L1 on neural activation

As observed in Figure 3, FR exhibited stronger activation than DE in the left IFG (M = 0.359, SD = 0.258 and M = 0.231, SD = 0.214 respectively; t(79) = 2.370, p = .020) as well as in the right IFG (M = 0.567, SD = 0.334 and M = 0.341, SD = 0.264, respectively; t(79) = 3.300, p = .001). In contrast, the activation for FR and DE did not differ in left STG (M = 0.112, SD = 0.194 and M = 0.109, SD = 0.177, respectively; t(79) = 0.055, p = .956) and the right STG (M = 0.190, SD = 0.242 and M = 0.121, SD = 0.195, respectively; t(79) = 1.373, p = .174).

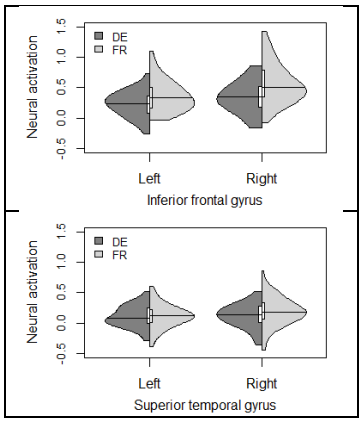


Figure 3: Neural activation in bilateral IFG (top) and STG (bottom) as a function of L1.

3.3. Relationship between neural activation and training effect in DE and FR

As apparent in Figure 4, DE listeners presented a (marginal) relationship between neural activation and training effect for both left IFG (r(33) = .313, p = .067) and right IFG (r(33) = .318, p = .062): the stronger the activation in the bilateral IFG, the larger the training effect. No such link was confirmed for FR listeners, neither in left IFG (r(44) = .192, p = .202), nor in right IFG (r(44) = -.03, p = .841).

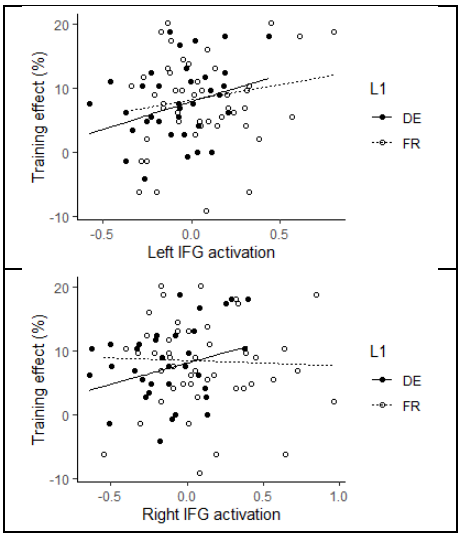


Figure 4: Training effect (%) as a function of neural activation in left and right IFG for DE and FR listeners.

As can be seen in Figure 5, the relationship between the left STG activation and the training effect was significant for DE ($r(33) = .442$, $p = .008$), but not for FR ($r(44) = -.028$, $p = .855$). The same pattern was observed for the right STG (DE: ($r(33) = .382$, $p = .023$); FR: $r(44) = -.038$, $p = .8$), confirming that the training effect increased with larger STG activation in DE, but not in FR.

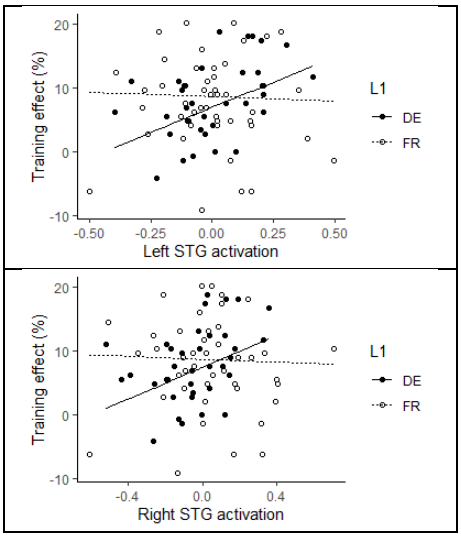


Figure 5: Training effect (%) as a function of neural activation in left and right STG for DE and FR listeners.

4. DISCUSSION

The present research aimed to examine the neural predisposition for L2 stress learning in French- and German-speaking listeners. Results from our behavioral experiment showed that although the French group performed overall more poorly at the pretest than the German group, a similar training effect was found for both groups. Such a training effect confirms previous results with French listeners ([6]). Furthermore, it suggests that not only speakers of languages with predictable stress but also speakers of variable stress languages benefit from a short perceptual training to improve their stress discrimination abilities in a foreign language with variable stress.

As for neural activation, French listeners showed a higher degree of activation in the left and right IFG than German listeners. In line with [15], the involvement of areas associated with cognitive control and working memory ([17]) was stronger for French than German listeners, suggesting larger cognitive resource demands for participants of a fixed-stress native language. On the other hand, neural activation in the bilateral STG did not differ between both language groups, suggesting similar auditory and word stress processing ([18], [19]) for French and German listeners.

Interestingly, our findings showed a different relationship between the training effect and the degree of neural activation in the two language groups. The stronger the engagement in areas associated with cognitive control and working memory (i.e., IFG), the larger the amount of improvement after training for L1 German participants, but not for L1 French participants. As for temporal activation, results showed a similar pattern. For L1 German listeners only, the larger the activation related to stress auditory processing (i.e., STG), the larger the training effect. No such relationship between strength of activation and training effect was found for L1 French listeners.

Taken together, our findings suggest a neural predisposition for L2 stress learning for L1 German listeners. German listeners with larger activation in fronto-temporal areas were more successful in learning L2 stress. In contrast, such a neural predisposition could not be observed for French listeners. The relationship between the amount of improvement after training and the degree of neural activation during L2 stress processing seems thus to be driven by the learners' native language. Our results therefore suggest that the neural predisposition for L2 stress learning comes along with the fact of having a native language with variable stress. Such an assumption however needs to be further explored with native learners of other free- and fixed-stress languages.

The present research reveals that the link between neural correlates of L2 stress processing and the amount of improvement after training might be specific to the learners' native language (fixed- vs. free-stress), even in case of similar average progression after training.

5. ACKNOWLEDGMENTS

This research was funded by SNF programs (IZSEZ0_180523/1; 10001F_200824), Pool recherche 2019 from University of Fribourg and Fonds du Centenaire (no 797), University of Fribourg. We would like to thank Prof. Nathalie Giroud (University of Zurich) and Prof. Rachel Theodore (University of Connecticut, USA) for their help in designing the fMRI paradigm. Thanks also to Laura Canedo for recording the stimuli.

6. REFERENCES

[1] Honbolygó, F., Kóbor, A., Csépe, V. 2019. Cognitive components of foreign word stress processing difficulty in speakers of a native language with non-contrastive stress. *International Journal of Bilingualism*, 23(2), 366–380. https://doi.org/10.1177/1367006917728393

[2] Degrave, P. 2017. Can music help learners and teachers in word stress perception? *Travaux du Cercle Belge de Linguistique*, 11, 1-20.

[3] Dupoux, E., Pallier, C., Sebastián, N., Mehler, J. 1997. A destressing 'deafness' in French? *J. Memory Lang.*, 36(3), 406-421,

[4] Schwab, S., Dellwo V. 2017. Intonation and talker variability in the discrimination of Spanish lexical stress contrasts by Spanish, German and French listeners. *J. Acoust. Soc. Am.*, 142 (4), 2419–2429, 2017.

[5] Ortega-Llebaria, M., Gu, H., Fan, J. 2013.English Speakers' Perception of Spanish Lexical Stress: Context-Driven L2 Stress Perception. *J. Phon.*, 41, 186–97, 2013.

[6] Schwab, S, Dellwo, V. 2021. Explicit versus non-explicit prosodic training in the learning of Spanish L2 stress contrasts by French listeners. *Journal of Second Language Studies*, 2021. http://doi.org/10.1075/jsls.21017.sch.

[7] Grymska, B. 2016. New Conceptualizations of Language Aptitude—The Potential of Working Memory in Second Language Acquisition (SLA). *Theory and Practice of Second Language Acquisitio*n, 2(1), 103-118.

[8] Wen, Z. E. 2019. Working memory as language aptitude: The phonological/executive model. In *Language Aptitude* (pp. 187-214). Routledge.

[9] Callan, D. E., Jones, J. F., Callan, A. M., Akahane-Yamada, R. 2004. Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models, *NeuroImage*, 22 (3), 1182-1194.

[10] Golestani N., Zatorre, R.J. 2004. Learning new sounds of speech: reallocation of neural substrates. *Neuroimage*, ;21(2):494-506. doi: 10.1016/j.neuroimage.2003.09.071.

[11] Wong, P. C. M, Perrachione, T. K., Parrish T. B. 2007. Neural characteristics of successful and less successful speech and word learning in adults. *Human Brain Mapping*; 28(10), 995–1006.

[12] Mesgarani, N., Cheung, C., Johnson, K., Chang, E. F. 2014. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343 (6174), 1006-1010.DOI: 10.1126/science.1245994

[13] Schwab, S., Mouthon, M., Jost, L., Salvadori, J., Stefanos-Yakoub, I., Ferreira da Silva, E., Giroud, N., Perriard, B., Annoni, J.-M. in press. Neural correlates of lexical stress processing in a foreign free-stress language. Brain & Behavior.

[14] Klein, E., Domahs, U., Grande, M., Domahs, F. 2011. Neuro-cognitive foundations of word stress processing—Evidence from fMRI. *Behavioral and Brain Functions*, 7(1), 15. https://doi.org/10.1186/1744-9081-7-15

[15] Rogenmoser, L., Mouthon, M., Etter, F., Kamber, J., Annoni, J.-M., & Schwab, S. (2022) The processing of stress in a foreign language modulates functional antagonism between default mode and attention network regions. bioRxiv; doi: https://doi.org/10.1101/2022.12.21.521156

[16] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5:9/10, 341-345.

[17] Fedorenko, E., Blank, I. A. 2020. Broca's area is not a natural kind. *Trends in Cognitive Sciences*, 24(4), 270-284.

[18] Aleman, A., Formisano, E., Koppenhagen, H., Hagoort, P., de Haan, E. H. F., Kahn, K. 2005. The functional neuroanatomy of metrical stress evaluation of perceived and imagined spoken words. *Cereb.Cortex,* 15, 221-228.

[19] Domahs, U., Klein, E., Huber, W., Domahs, F. 2013. Good, bad and ugly word stress – fMRI evidence for foot structure driven processing of prosodic violations. *Brain and Language*, 125(3), 272-282. https://doi.org/10.1016/j.bandl.2013.02.012.

Word-final devoicing in Brazilian Portuguese and L2 English

Wellington Mendes Junior, Thaïs Cristófaro Silva

Federal Center for Technological Education of Minas Gerais, Federal University of Minas Gerais
wellington@cefetmg.br, thaiscristofaro@gmail.com

ABSTRACT

This study examines word-final devoicing in Brazilian Portuguese (BP) and L2 English. It aims at assessing whether an ongoing sound change related to devoicing in the L1 affects the production of L2 forms. We considered the production of plural suffixes comprised of a (stop + sibilant) cluster (e.g. *ringues* [hĩgs] ~ [hĩks] 'rings' in BP and *bags* [bægz] ~ [bæks] in English). Harmonics-to-noise ratio was used to measure the degree of sibilant voicing. Results showed that devoicing was significantly attested in both L1 and L2. This provides evidence that word-final devoicing, which is an ongoing sound change in the L1, affects the production of L2 forms. Moreover, L2 English sibilant voicing is influenced by the following phonological context, a pattern which also seems to be imported from the L1. We show that not only segments, but also fine phonetic detail is transferred to the L2 as suggested by Exemplars Model.

Keywords: Word-final devoicing; Brazilian Portuguese; L2 English; Exemplars Model.

1. INTRODUCTION

This paper examines word-final devoicing (WFD) in Brazilian Portuguese (e.g. *ringues* [hĩgs] ~ [hĩks] 'rings') and in English as a Second Language (e.g. *bags* [bægz] ~ [bæks]). It intends to contribute to the understanding of L2 phonology by examining how emerging sound patterns from the L1 are adopted in the L2. The analysis is grounded on the premises of Exemplars Model which propose that fine phonetic detail is part of phonological representations (c.f. [3], [4], [10]).

In BP, word-final devoicing has been associated with an ongoing sound change that favors the production of voiceless consonants word-finally (e.g. *ringues* [hĩgs] ~ [hĩks] 'rings'). This paper intends to be a contribution to examining the earlier stages of WFD in BP phonology, a phenomenon which occurred in several Indo-European languages (c.f. [2], [8], [9], [11]).

It has been noticed that unstressed high front vowels in BP are weakened and lost when flanked between a consonant and a word-final sibilant. For example, *lápis* ['la.pɪs] > [laps] 'pencil' (c.f. [12]). This pattern applies to plural forms, as in *ringues* ['hĩ.gɪs] > [hĩgs] ~ [hĩks] 'rings' and *cheques* ['ʃɛ.kɪs] > [ʃɛks] 'cheques'. The word-final sibilant is always voiceless in BP, regardless of the voice quality of the preceding consonant (c.f. [5]). Across morphemes or word-boundaries, a regressive assimilation rule leads the sibilant to be voiced when it is followed by a vowel or a voiced consonant: [hĩgs] *ringues* 'rings' > [hĩgz#a.ma.'ɾɛ.lʊs] *ringues amarelos* 'yellow rings' and [ʃɛks] *cheques* 'cheques' > [ʃɛkz#a.ma.'ɾɛ.lʊs] *cheques amarelos* 'yellow cheques'.

In this paper, we examine the ongoing sound change in BP which is related to the alternation between voiced and voiceless consonants (e.g. [hĩgz#a.ma.'ɾɛ.lʊs] ~ [hĩks#a.ma.'ɾɛ.lʊs]) and we also consider whether this sound change influences the production of L2 English plural forms (e.g. *bags* [bægz] ~ [bæks], *bags are* [bægz#ar] ~ [bæks#ar]).

In English, the regular plural suffix and 3rd person singular present is posited to be /z/ (c.f. [7]). A progressive assimilation rule predicts that if a vowel or a voiced consonant precedes /z/, the output is [z], as in *dogs* [dɔgz], *trees* [triːz] and *pies* [paɪz]. On the other hand, if a voiceless consonant precedes /z/, it surfaces as [s], as in *cups* [kʌps], *cats* [kæts] and *ducks* [dʌks]. Finally, if a sibilant or an affricate precedes the sibilant, the outcome is [ɪz], as in *buses* [bʌsɪz], *quizzes* ['kwɪz.ɪz] and *watches* ['wɔtʃ.ɪz]. Whereas in English sibilants are prone to progressive assimilation, in Brazilian Portuguese sibilants undergo regressive assimilation.

The main question we posit is whether WFD in PB influences the production of L2 English plural forms. If that is the case, we expect that English plural forms produced by BP speakers will favor voiceless consonants word-finally, as it is the more robust pattern in L1, inhibiting the English progressive assimilation rule that applies to plural forms (e.g. [bæks] instead of [bægz] *bags*). We also posit that voiced sibilants will be favored when followed by a word-initial vowel (e.g. [bægz#ar] *bags are*). This follows from the BP regressive assimilation rule which leads sibilants to be voiced when followed by a vowel.

In sum, we will investigate whether word-final devoicing takes place in both languages and

whether sibilant voicing can be influenced by the following phonological environment. An experiment was designed to test the production of (stop + sibilant) clusters in English spoken by Brazilian speakers. Harmonics-to-noise ratio (HNR) was used to measure the degree of sibilant voicing, aiming at offering an accurate evaluation of the experimental data carried out by acoustic analysis.

2. METHODOLOGY

A group of twenty Brazilians studying at the Federal Center for Technological Education of Minas Gerais took part in this study. All participants were high school students who had been taking English classes as part of the school's curriculum for at least one year. Half of the participants displayed basic levels of L2 English (A1 or A2) whereas the other half displayed pre-advanced (B2) to advanced (C1) levels, according to the Common European Framework of Reference for Languages.

A set of 36 plural nouns ending in a (stop + sibilant) cluster were considered in BP, e.g. *cheques* [ʃɛks] 'cheques' and *botes* [bɔts] 'boats'. For the L2 English case study, another set of 36 words were selected, e.g. *bags* [bægz] and *boats* [bəʊts].

The experiment was comprised of two trials. The first one involved a picture-counting task in which participants were asked to count and name the elements shown in the pictures. This would lead to the production of the target words in their plural forms (e.g. *two cakes*). In a second application of the picture-counting task, participants were asked to pronounce a short carrier-sentence following the sibilant (e.g. *two cakes are seen*). The goal was to assess whether they would voice the word-final sibilant due to the presence of a word-initial vowel in sequence. The second trial involved sentence reading. The number of sentence syllables, intonation patterns and words' parts of speech were controlled. The BP study took place after the L2 English study and considered the same phonological environments and task types.

Due to the COVID-19 pandemic, data collection was performed remotely. Experiments were recorded with the Open Broadcaster Software Studio at 48 kHz sampling rate. The obtained recordings were converted into WAVEform audio format by the software Adobe Premiere 2020, which was able to maintain the same sampling rate as the original files. The average time to complete the experiment was 1 hour and 20 minutes. A total of 2833 tokens were considered for the L2 English study. For the BP study, 2831 tokens were considered. Samples were edited and manually annotated using Praat TextGrids (c.f. [1]).

This research mainly considered the voice quality of word-final sibilants. In BP, only voiceless sibilants occur word-finally, unless a vowel follows it, to which a voiced sibilant emerges. In English, voiced and voiceless sibilants occur word-finally. When a vowel follows the sibilant, the voice quality remains as it formerly was (rather than changing as it occurs in BP). Thus, we posited that word-final voiceless sibilants would be favored in L2 English, as it is the more robust pattern in L1. We also posited that voiced sibilants occur at higher rates when followed by a word-initial vowel.

Voicing was measured under Harmonics-to-noise ratio, which represents the degree of acoustic periodicity. Each token was extracted to a separate sound object and a harmonicity object was created, from which the mean harmonicity was calculated, hereafter the HNR. The details of its calculation can be found in [1]. Based on the discussion from Praat's manual [1], higher values of HNR should correspond to higher voicing rates. This is consistent with prior research which has shown that voiceless intervocalic sibilants in English ([s] and [ʃ]) typically have HNR values below 5 dB, whereas voiced sibilants ([z] and [ʒ]) fall within the 5-10 dB range [13].

3. RESULTS

Results showed that Brazilian Portuguese speakers tend to present a voiceless sibilant word-finally *cheques* [ʃɛks#], even in contexts where a voiced counterpart would be expected. i.e., *cheques amarelos* [ʃɛkz#amaɾɛlʊs]. Consider Figure 1.

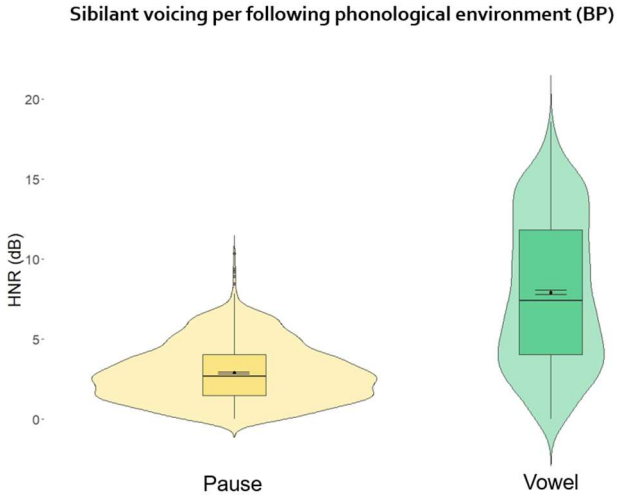


Figure 1: Sibilant voicing per following phonological environment in Brazilian Portuguese.

Figure 1 displays the rates of sibilant voicing per following phonological environment. Sibilants followed by a pause (left violin plot) were expected

to display low rates of HNR and sibilants followed by a vowel (right violin plot) were expected to be voiced, thus displaying higher rates of HNR.

As expected, lower voicing rates were attested when the sibilant was followed by a pause (e.g. *cheques* [ʃɛks#], mean = 2.8 dB), whereas higher voicing rates were attested when the sibilant was followed by a word-initial vowel (e.g. *cheques amarelos* [ʃɛkzamaɾɛlʊs], mean = 7,9 dB). Wilcoxon test results revealed that this factor is statistically significant (W = 340138, p-value < 0.01). We highlight the fact that sibilants followed by vowels showed considerable variability, with a standard deviation of 4.5. This indicates that such sibilants were produced with variable degrees of voicing, some of them with unexpectedly low rates of HNR, which indicates a tendency towards the production of [s]. Spectrographic data analysis corroborates this observation. Consider Figure 2.

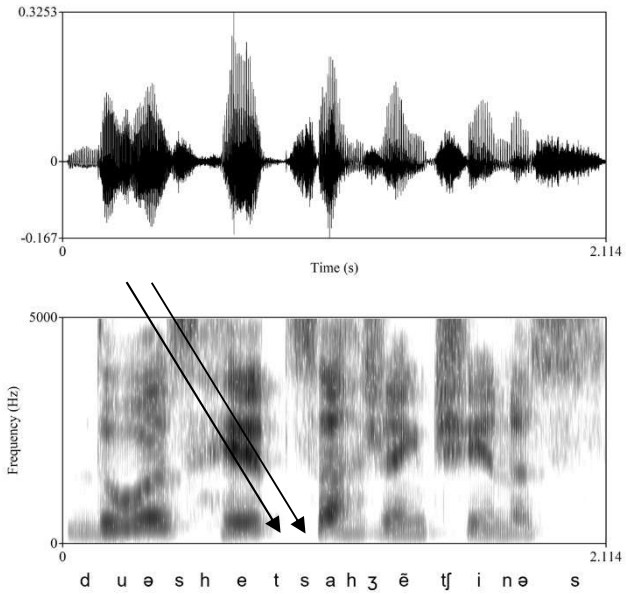


Figure 2: Spectrogram analysis of word-final devoicing in BP.

Figure 2 shows the spectrogram of the sentence *duas redes argentinas* 'two argentine hammocks', produced by a 17 year-old male participant. A voiced sibilant was expected as the following context is a vowel. However, two arrows indicate the absence of voicing during the production of the (stop + sibilant) sequence in the target word. If there was voicing, the bottommost part of the spectrogram would be dark instead of being white.

Additionally, Figure 2 shows that not only did the sibilant remain voiceless, but the preceding stop – which was formerly voiced - also shared the same property, being voiceless. Thus, [hets ah.ʒẽ.'ʧi.nəs] instead of [hedz ah.ʒẽ.'ʧi.nəs] was attested. Our data showed that word-final devoicing occurred with all cluster types followed by vowels assessed in this study ([ps, ts, ks, bs, ds, gs]), pointing out to an ongoing sound change in BP which triggers only voiceless consonants word-finally. This tendency was attested in BP even with words to which word-final devoicing leads to contrast loss: *grades* 'grids' and *grátis* 'free' [gɾats], *sedes* 'headquarters' and *setes* 'seven' [sɛts], *ringues* 'rings' e *rinques* [hĩks] 'rinks'.

It appears that word-final devoicing in BP might be associated with an Indo-European historical trend, as observed in languages such as Dutch [11], German [2], French [9] and Romanian [8]. This study thus contributes to examining the earlier stages of word-final devoicing in BP.

Let us now examine WFD in L2 English. An important question that arises is whether the ongoing sound change in BP related to word-final devoicing influences the production of L2 English plural forms. Consider Figure 3.

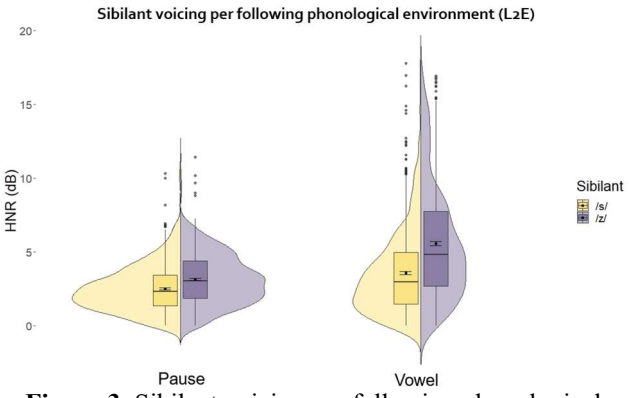


Figure 3: Sibilant voicing per following phonological environment in L2 English.

Figure 3 displays sibilant voicing rates per following phonological environment in L2 English by BP speakers. Sibilants followed by a pause (left violin plots) were expected to display low rates of HNR and sibilants followed by a vowel (right violin plots) were expected to be voiced, thus displaying higher rates of HNR. Yellow plots represent tokens whose voiceless sibilant [s] was expected (e.g. *cups* [kʌps] and *cats* [kæts]), whereas purple plots represent tokens whose voiced sibilant [z] (e.g. *jobs* [dʒɔbz], *beds* [bɛdz]) was expected. Results show that, as we hypothesized, lower voicing rates were attested when the sibilant was followed by a pause (e.g. *cups*, *jobs*) – with an average of 3.0 dB - and that occurred regardless of the consonant that precedes the sibilant. On the other hand, higher rates of voicing were attested when there was a vowel following the sibilant (e.g. *cups are*, *jobs are*) – with an average of 4.3 dB. Wilcoxon test results revealed that the following phonological

environment is statistically significant in L2 English, both within the group whose sibilant [s] is expected (W = 203088, p-value < 0, 01) and within the group whose sibilant [z] is expected (W = 155599, p-value < 0.01).

Notice that in L2 English voiceless consonants occur regardless of the preceding phonological context - which is actually what triggers voicing in L1 English due to progressive assimilation. The predominance of voiceless clusters in L2 English points out to the influence of L1 sound patterns for the construction of L2 phonological representations (c.f. [6]).

Moreover, when comparing Figures 1 and 3, we observe that learners transfer an allophonic pattern from L1 BP (i.e. regressive assimilation) to L2 English. Our results additionally suggest that L2 learners not only import allophonic sound patterns from the L1 – i.e. regressive assimilation -, but also emerging sound patterns – i.e. word-final devoicing. Once again, consider the data represented in purple in Fig. 3. These data involve words whose sibilant [z] was expected (e.g. [bægzɑ:r] *bags are*), not only due to progressive assimilation attested in English but also due to BP's regressive assimilation. Note that even though most sibilants followed by a vowel had higher rates of HNR, some tokens were associated with low rates (standard deviation = 3.5). The L2 English data in purple correspond to the BP data reported in Figure 1: in both languages the production of voiced clusters occurs only to some extent when there is a vowel following the sequence. This offers evidence that word-final devoicing, a phenomenon that still emerging in L1 BP, is also aggregated in L2 English. Consider Figure 4.

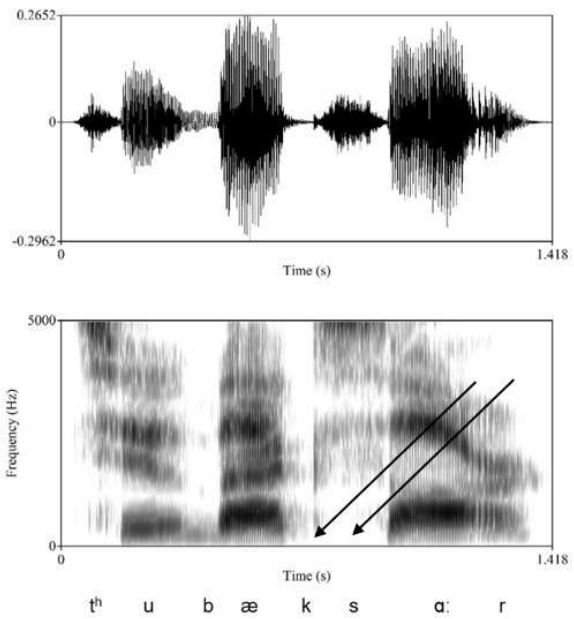


Figure 4: Spectrogram analysis of word-final devoicing in L2 English.

Figure 4 shows the spectrogram of the sentence "two bags are [...]", produced by the same participant depicted in Figure 2. The arrows indicate the absence of voicing in the final consonant cluster. It is clear that there was the production of [bæksɑ:r] instead of [bægzɑ:r], in which both the stop and the word-final sibilant were devoiced. Notice that WFD occurred with all voiced cluster types assessed in this study (i.e. [bz, dz, gz]), not only when the sibilant was followed by a pause (e.g. *two labs* [tu: læps]) but also when the sibilant was followed by a vowel (e.g. *two labs are* [tu: læps ɑ:r]). Alike our BP data, this was attested even in words which word-final devoicing leads to contrast loss: *coats* and *codes* [kəʊts], *sites* and *sides* [saɪts], *laps* and *labs* [læps] *kits* and *kids* [kɪts], *pigs* and *pix* [pɪks]. This suggests that not only segments, but also fine phonetic detail – associated with a phonological pattern that is still emerging in the L1 – is transferred to the L2.

4. CONCLUSIONS

Our results show that word-final devoicing is widely attested in both L1 BP and L2 English. This provides evidence that an ongoing sound change in the L1 affects the production of L2 forms. We also found that not only do the final sibilants devoice but also the preceding stops are also devoiced. Moreover, L2 English sibilant voicing is influenced by the following phonological context, a pattern which also seems to be imported from the L1. We argue that exemplars associated with an allophonic sound pattern (production of voiced clusters followed by vowels) are in competition with an emerging sound pattern (devoicing of word-final clusters). This offers evidence that fine phonetic detail is part of phonological representations, as posed by the Exemplars Model (c.f. [3, 4, 10]). Finally, this paper shows that word-final devoicing, an emergent phonological pattern in L1 BP, is also aggregated in English L2.

5. REFERENCES

[1] Boersma, P., & Weenink, D. (2020). Praat manual. *Amsterdam: University of Amsterdam, Phonetic Sciences Department*.

[2] Brockhaus, W. (2012). Final devoicing in the phonology of German. In *Final Devoicing in the Phonology of German*. Max Niemeyer Verlag.

[3] Bybee, J. (2001). Frequency effects on French liaison. *Typological studies in language*, *45*, 337-360.

[4] Bybee, J. (2008). Usage-based grammar and second language acquisition. In *Handbook of cognitive linguistics and second language acquisition* (pp. 226-246). Routledge.

[5] Cristófaro-Silva, T, Mendes-Jr., W. (2022). Plural formation in English: a Brazilian Portuguese case study. In: *Second Language Pronunciation: Different Approaches to Teaching and Training*. De Gruyter Mouton.

[6] Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3-83.

[7] Hayes, B. (2011). *Introductory phonology*. John Wiley & Sons.

[8] Hutin, M., Niculescu, O., Vasilescu, I., Lamel, L., & Adda-Decker, M. (2020, maio). Lenição e fortificação de stop codas em romeno. In *SLTU-CCURL*.

[9] Jatteau, A., Vasilescu, I., Lamel, L., Adda-Decker, M., & Audibert, N. (2019, September). " Gra [f] e!" Word-Final Devoicing of Obstruents in Standard French: An Acoustic Study Based on Large Corpora. In *INTERSPEECH* (pp. 1726-1730).

[10] Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology*, 25-40.

[11] Simon, E. (2010). Phonological transfer of voicing and devoicing rules: evidence from L1 Dutch and L2 English conversational speech. *Language Sciences*, *32*(1), 63-86.

[12] SOARES, V. H. M. (2016). *Encontros consonantais em final de palavra no português brasileiro* (Master's dissertation, Universidade Federal de Minas Gerais.

[13] MANIWA, K., JONGMAN, A., & WADE, T. (2009). Acoustic characteristics of clearly spoken English fricatives. In *The Journal of the Acoustical Society of America*, 125(6), 3962-3973.

Characterization of Mandarin-accented French across three different speaking styles: a corpus-based study

Wenxun Fu, Martine Adda-Decker, Barbara Kühnert

Laboratory of Phonetics and Phonology, UMR 7018, CNRS - Paris 3 Sorbonne Nouvelle University, France
(wenxun.fu, martine.adda-decker, barbara.kuhnert)@sorbonne-nouvelle.fr

ABSTRACT

This paper investigates the production of oral vowels and voiced fricatives in Mandarin L2 French across different speaking styles. We analyzed a corpus of 5 hours' speech from fifteen Mandarin learners of French performing three different production tasks of increasing complexity: isolated word reading, text reading and spontaneous conversations. Similar data from one hour's recording of five French natives were used for comparison. Results show that rounded vowels, although part of the Mandarin inventory, tend to be less rounded than those of natives, whereas voiced fricatives are poorly voiced due to their absence from the Mandarin system. Results also reveal challenges in the continuous speech task, highlighting the importance of specific training for L2 learners. The main differences found between natives and L2 speakers lay in the amount of vowel reduction and the degree of voicing of the voiced fricatives as speech changes from reading to spontaneous speaking style.

Keywords: Mandarin learners, L2 French, vowel reduction, fricatives, speaking styles.

1. INTRODUCTION

1.1 Variation and reduction in spoken languages

It is well established that the quality of foreign accent highly depends on the learner's native language. It is therefore important for L2 teachers to put much emphasis on those sounds that are different from the L1 language. However, it is not enough to be able to pronounce L2 sounds correctly only in the base form of the word; after all, seldom do we speak in the same way as does an electronic dictionary. Variation and reduction are omnipresent in spoken language production and highly dependent on factors such as speaking style, environment, social context and the relationship between the speaker and the listener [1].

While a native speaker usually switches between different degrees of variation and reduction without effort, it can be challenging for L2 learners to produce the reduced forms of the target language appropriately [2]. For instance, Gut [3] showed a lack of sufficient vowel reduction in German by English learners. Meanwhile, variation and reduction are often related to the rate of speech and the duration of segments, and L2 learners are known to have a slower speech rate in general.

Recently, more L2 oral corpora have been created for studying L2 phonetics and phonology, together with thorough annotations, permitting a more comprehensive description of L2 learners' pronunciation and accent. Aside from its linguistic purpose, the study of L2 oral corpora can serve as a valuable reference for the design of specific pronunciation training programs. L2 pronunciation training is supposed to be most effective when conducted with material appropriately adapted to the specific L1 difficulties. In our study, we have based our analyses on a newly collected oral corpus, which includes the reading of word lists, the reading of texts and spontaneous conversations by Mandarin learners of French (ChiLFreSC).

1.2. Some differences between the French and Mandarin phonological systems

One of the most notable differences between the Mandarin and French phonological systems is the occurrence of voiced occlusive and fricative consonants. Although rich in both fricatives and affricates, Mandarin possesses only one *voiced* fricative /ʐ/, and the main opposition for the occlusives is made by the distinctive feature of aspiration rather than voicing. While previous studies [4, 5, 6] investigated the problems of Mandarin learners of French when producing voiced occlusives, the production of voiced fricatives has not yet been studied extensively.

The presence and frequency of nasal and rounded oral vowels are further distinguishing features of the French phonological system when compared with the Mandarin inventory. In French, six out of ten oral vowels (excluding the central vowel /ə/) and all of the back vowels /u, o, ɔ/ are rounded

vowels. In Mandarin, there are two confirmed rounded vowels /y/ and /u/. The roundedness of the middle vowels is as debatable as the inventory of the middle vowels themselves. Five variants of mid-vowels have been noted by Xu [7], namely [o], [E], [ɤ], [e] et [ə]. The author believes at the same time that [o], in the syllable [wo], is not completely rounded, and according to Duanmu [8] the unrounded [ɤ] is the most spoken variant in Mandarin, a vowel which is basically distinguished from [ə] more by its duration than by its quality. The mid rounded vowels in French /ø, œ/ are therefore challenging for Mandarin learners for they are not only constantly rounded but also different from each other with respect to the degree of tongue aperture, as shown in the study of Wang [9].

2. OBJECTIVES AND HYPOTHESES

The present study wants to gain a more comprehensive understanding of the production strategies of Mandarin learners of French on the segmental level. In order to do so, our corpus includes all the oral vowels (excluding schwa) and all the fricatives of the French system. Moreover, we aim to describe the variation and reduction of the L2 speech in comparison to the L1 speech across three speaking styles, namely, isolated word reading, text reading and spontaneous conversations. We hypothesize the degree of vowel centralization and reduction to be smaller in L2 than in L1 speakers. With respect to the production of the fricatives /v, z, ʒ/, less voicing should be found in L2 speech, and depending on the position of the fricative and the speaking style, the quality of voicing might be different. The overall results of the study will serve as the main reference for an on-going pronunciation training program targeted at Mandarin learners of French from the Beijing Language and Culture University.

3. METHODS

2.1. Participants and recordings

Fifteen female Mandarin learners of French (mean age 22, SD: 2.1) participated in the experiment that consisted of reading 50 isolated words, then reading three different texts corresponding to the level between B1 and B2 of the CEFR [10], and two short conversations with the investigator. At the moment of recording, all Mandarin participants had learned French for at least one and a half years with a level between B1 and C1 based on the results from standard tests such as DELF and DALF [11]. According to their auto evaluations, 8 participants believed that they had a strong to very strong foreign accent. Five French female native speakers from the Parisian region (mean age 23, SD: 1.4) completed the same tasks.

2.2. Data processing

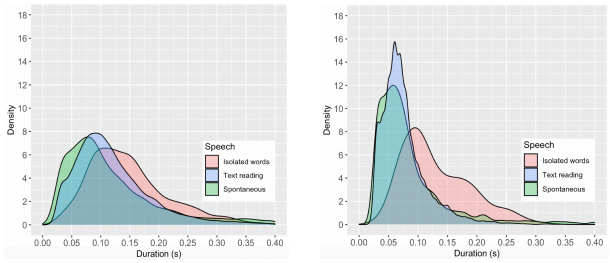
All data were processed with the help of the LIMSI forced alignment system [12] before a manual correction. The acoustic parameters were extracted using Praat [13]. For each parameter, a single value was obtained by averaging measurements made at 1/3, 1/2, and 2/3 of the segment. 70647 and 25498 segments were collected from the L2 and L1 speech respectively, among which 24531 and 8809 are oral vowels, 3233 and 1330 are voiced fricatives. The duration and the first three formants were analyzed for oral vowels. The analysis of the voiced fricatives was based on the f0 ratio.

4. RESULTS

4.1. Oral vowels

4.1.1. Duration

While the mean duration of vowels is shortest in text reading for both groups (L2: mean=123 ms, SD=76 ms; L1: mean=76 ms, SD=40 ms), L2 learners show longer vowel durations in each of the three speaking styles, in particular in text reading (L2:L1=1:0.61) and spontaneous speech (L2:L1=1:0.68). For both L1 and L2 speakers, the highest variability of vowel duration occurs in spontaneous speech, which explains its longer mean duration in comparison to text reading, even though more short vowels (duration < 50ms) occur in spontaneous speech.



,

4.1.2. F1/F2 space

Figure 2 shows that Mandarin learners occupy a larger F1/F2 space compared to the L1 speakers, and that less vowel centralization can be observed from isolated word reading to spontaneous speech for L2 than for L1. The two-way ANOVA with

independent factors *Speaker* and *Phoneme* shows a significant difference between L1 and L2 for F1 ($p<0.001$) but not F2. The higher F1 suggests that L2 learners produced French oral vowels with a lower tongue position and a more open mouth.

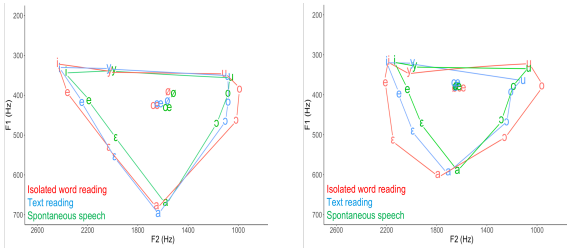


Figure 2: F1/F2 plot of French oral vowels produced by Mandarin learners (left) and French natives (right) by speaking style.

4.1.3. Roundedness

We grouped the oral vowels into rounded vowels /u, o, ɔ, ø, œ, y/ and unrounded vowels /i, e, ɛ, a/. Figure 3 compares the difference between F3 and F2 (F3-F2) according to the roundedness of the vowels. The F3-F2 difference value for rounded vowels is much higher for Mandarin learners than for native speakers, while the value for unrounded vowels is rather similar for the two groups. Since a lower F3 is associated with more roundedness, the results suggest that the Mandarin learners produced the French rounded vowels with less lip rounding in all three speaking styles.

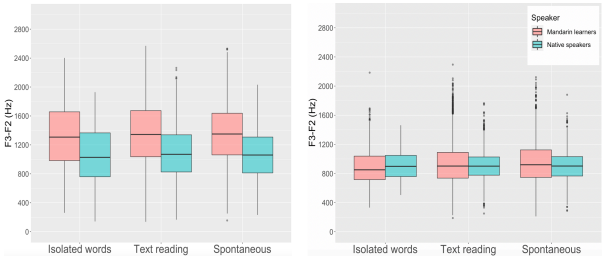


Figure 3: Difference between F3 and F2 in Hz for the rounded vowels /u, o, ɔ, ø, œ, y/ (left) and the unrounded vowels /i, e, ɛ, a/ (right) for Mandarin learners (red) and French natives (green).

The rounded vowels /u, y, o, ɔ/ occur in the Mandarin vowel inventory (though /o, ɔ/ are questionable). However, Figure 4 shows that F3-F2 difference for these four vowels is even larger between the two groups of the speakers ($p<0.001$), suggesting that in Mandarin the rounded vowels are less rounded as compared to the French vowels of the same category. The largest F3-F2 variability is found in text reading for the Mandarin learners (mean =1432; σ=566), while for French speakers, no significant F3-F2 differences have been observed.

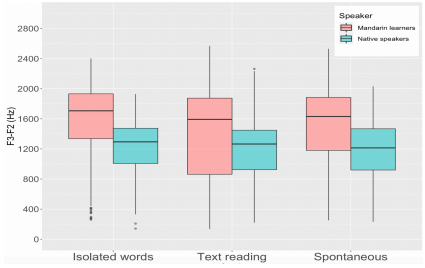


Figure 4: Difference between F3 and F2 in Hz for the rounded vowels /u, o, ɔ, y/ for Mandarin learners (red) and French natives (green).

4.1.4. Diffusion

/i, y, e, ɛ/ are intrinsically diffuse vowels as the distance between F2 and F1 (F2-F1) is larger compared to the compact vowels /u, o, ɔ, a/. For Mandarin learners, the compact vowels are more compact ($p<0.001$), and the diffuse vowels are more diffuse ($p<0.001$) in comparison to French natives in the three speaking styles (see Figure 5). For Mandarin learners there is no significant difference of the diffusion in the compact vowels between the three speaking styles, and the significant difference of the diffusion in the diffuse vowels is only found between isolated word reading and spontaneous speech ($p<0.001$). The French native speakers, however, show a clear distinction in the diffusion of the diffuse vowels for the three speaking styles ($p<0.001$), with the mean value of F2-F1 being 1789 Hz in isolated word reading, 1681 Hz in text reading and 1629 Hz in spontaneous speech.

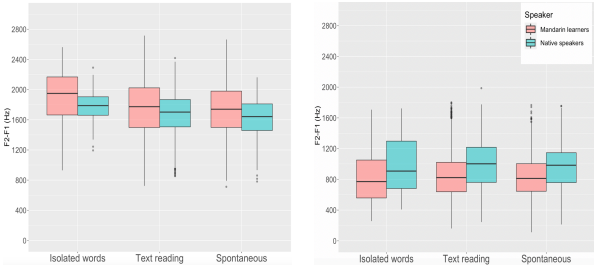


Figure 5: Difference between F2 and F1 in Hz for the diffuse vowels /i, y, e, ɛ/ (left) and compact vowels /u, o, ɔ, a/ (right) for Mandarin learners (red) and French natives (green).

4.2. Voiced fricatives

Among 11624 fricatives collected in both L1 and L2 speech, 4563 were the voiced fricatives /v, z, ʒ/. 1410 were positioned at the beginning of words, 2146 in the middle, and 1000 at the end. Overall, the t-test shows significant differences of f0 ratio ($p < 0.001$) between the Mandarin learners and the French native speakers in all positions. As shown in Table 1, the mean f0 ratio of the voiced fricatives produced by the Mandarin learners is systematically lower when compared to native speakers. The largest difference occurs in spontaneous speech, in which the mean f0 ratio is 0.83 for the native speakers and only 0.59 for the Mandarin learners. No significant effect of speaking style on the f0 ratio is revealed for Mandarin learners, while French speakers show a significant difference between word reading and spontaneous speech ($p<0.05$). Both groups voiced the fricatives the most when they occurred inter-vocalically; the vowels may have facilitated the voicing of the embedded fricatives. With respect to position, the most significant difference between the two groups is found in word-final position, with on average, less than half of the segments being voiced for the Mandarin learners. The French subjects have approximately the same voicing ratio in word-initial and word-final position.

	Mandarin learners			French native speakers		
	Beginning	Center	Final	Beginning	Center	Final
Total	0.65 (0.35)	0.77 (0.32)	0.49 (0.37)	0.79 (0.31)	0.95 (0.16)	0.81 (0.31)
Word reading	0.48 (0.30)	0.70 (0.36)	0.31 (0.30)	0.58 (0.26)	0.90 (0.17)	0.50 (0.31)
Text reading	0.67 (0.34)	0.79 (0.32)	0.53 (0.37)	0.81 (0.31)	0.97 (0.14)	0.83 (0.31)
Spontaneous	0.62 (0.37)	0.71 (0.34)	0.42 (0.34)	0.80 (0.31)	0.91 (0.22)	0.78 (0.31)

Table 1: Mean f0 ratio (SD) of the voiced fricatives /v, z, ʒ/ by speaking style.

Looking at the consonants individually, Figure 6 shows that for French natives, /z/ is the fricative with the strongest voicing. This may be partially explained by the much higher rate of produced liaisons when compared to the Mandarin learners. Most of /ʒ/ occurrences in final position feature an f0 ratio of around 20% for the Mandarin learners, while for the natives more /ʒ/ occurrences are found with higher f0 ratios.

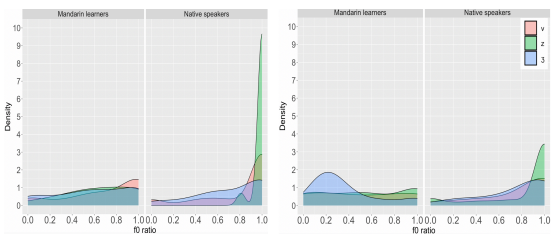


Figure 6: The density of f0 ratio of the voiced fricatives in initial (left) and final position (right).

5. DISCUSSION AND CONCLUSION

Our study confirmed the difficulties of pronouncing rounded vowels and voiced fricatives for Mandarin learners of French in all three examined speaking styles. Interestingly, compared to the rounded vowels which were unknown and new to the Mandarin learners, those present in the Mandarin inventory were not necessarily produced better; rounded vowels in French are likely to be more rounded than the rounded vowels in Mandarin. In addition, the commonly studied F1/F2 dimension may not be sufficient to measure the degree of foreign accent during French vowel productions, and the analysis of F3 should be considered for lip rounding. The examination of the F1/F2 space nevertheless revealed the varying degrees of centralization between L1 and L2 speakers with a decrease in formality. For the Mandarin speakers, the difficulty lies particularly in continuous speech where natives realize shorter vowel durations and larger degrees of reduction.

Though Mandarin learners showed a larger voicing ratio for the voiced fricatives in text reading and spontaneous speech than isolated word reading, the difference of the voicing ratio between the two groups is also larger in the two types of continuous speech. This phenomenon again confirms the major challenge that continuous speech presents for L2 learners. It also should be noted that while for the French speakers centralization and reduction seem the most important characteristics of vowel production in continuous speech, a continuous increase, rather than weakening, of the voicing feature occurred during the production of voiced fricatives. Therefore， continuous speech may not only be marked by “reduction”, and more studies should be devoted to language-specific rules of connected speech production for the L2 learners. Meanwhile, since our study has been purely based on the analyses of acoustic parameters, future research could involve perception tests of native speakers to evaluate whether the difficulties encountered by Mandarin learners in continuous speech are associated with a more noticeable foreign accent.

6. REFERENCES

[1] Lancien, M., Côté, M. H. 2018. Phonostyle et réduction vocalique en français laurentien. *In SHS Web of Conferences,* Vol. 46, p. 09003. EDP Sciences.

[2] Spilkov, H., van Dommelen, W. A. 2010. English of in L1 and L2 speakers' read and spontaneous speech. Working papers/Lund University, Department of Linguistics and Phonetics, 54, 91-96

[3] Gut, U., Trouvain, J., Barry, W. J. 2007. Bridging research on phonetic descriptions with knowledge from teaching practice-The case of prosody in non-native speech. *Trends in linguistics studies and monographs,* 186, 3. Monton

[4] Gabriel, C., Kupisch, T., Seoudy, J. 2016. VOT in French as a foreign language: A production and perception study with mono-and multilingual learners (German/Mandarin-Chinese). *In SHS web of conferences,*Vol. 27, p. 09002. EDP Sciences.

[5] Wang, R. "The categorical perception of voicing of French occlusives by learners of Mandarin." M.S.thesis, Dept.Phonetics, Sorbonne-Nouvelle Univ., Paris, France, 2017

[6] Zhang, y. "The production and perception of French consonants by Mandarin EFL learners of phonetic and non-phonetic background." Dept.Phonetics, Sorbonne-Nouvelle Univ., Paris, France, 2016

[7] Xu,S. 1980. *The phonology of standard Chinese*. Wenzi Gaige Beijing.

[8] Duanmu, S. 2007. *The phonology of standard Chinese.* Oxford: Oxford University Press.

[9] Wang, X., Hueber, T., Badin, P. 2014. On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. *In 10th international seminar on speech production (issp 2014)* p. 449-452.

[10] Council of Europe (CEFR) (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. New York: Cambridge University Press.

[11] DELF and DALF
http://www.delfdalf.fr

[12] Gauvain, J.L., Lamel, L. Adda, G. 2002. *The Limsi Broadcast News Transcription System, Speech Communication,* 37(1-2):89-108.

[13] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

7. ACKNOWLEDGEMENTS

This work was supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083) and ANR CLD2025.

THE LINK BETWEEN PERCEPTION AND PRODUCTION OF ENGLISH VOWEL CONTRASTS IN NON-NATIVE SPEAKERS

Chih-Chao Chang[1], Chun-Ting Chien[2], Yu-An Lu[1]

National Yang Ming Chiao Tung University[1], National Taipei University of Nursing and Health Sciences[2]
c.c.willchang@gmail.com; alixchien.ct@gmail.com; yuanlu@nycu.edu.tw

ABSTRACT

This study investigates the link between L1-Mandarin speakers' perception and production of English /i, ɪ/ (high) and /ɛ, æ/ (mid) vowels. Through perceptual categorization and phonetic imitation experiments, we found that L2 speakers have difficulty integrating spectral and durational cues for English vowel categorization, as indicated by either a trade-off or a null correlation between the perceptual weights of the two cues at the individual level. For the perception-production link, the imitation of English high vowels was constrained by the reliance on durational cues in perception and mediated by L1 categories at the phonological level. In contrast, the perception-production link for English mid vowels was only established at the acoustic-phonetic level for vowel quality but not for vowel duration. Our findings provide a look into the level of abstraction of L2 sound categories in non-native speakers across different phonetic aspects and vowel contrasts.

Keywords: L2 perception-production link, English tense-lax vowels, imitation, vowel categorization

1. INTRODUCTION

While some second-language (L2) learning models propose a strong connection between speech perception and production [4], other models do not make a clear association between the two [1, 2]. Most L2 studies explored the two modalities separately in different tasks, yielding discordant claims about the perception-production link. To rectify this, the current study uses perceptual categorization and phonetic imitation experiments to investigate the perception and production of English /i, ɪ/ and /ɛ, æ/ vowel contrasts by L1-Mandarin speakers and the perception-production link at the individual level through engaging the two modalities simultaneously.

In English, tense vowels have more peripheral F1/F2 values and longer durations than lax vowels [6]. Perceptually, however, English native speakers rely more on spectral than durational cues for categorizing high (/i, ɪ/) and mid (/ɛ, æ/) vowels [10]. Furthermore, the perceptual weights of spectral and durational cues are positively correlated at the individual level in mid vowels [11], whereas the correlation is underexplored in high vowels. In addition, [11] used a phonetic imitation task to examine the perception-production link for mid vowels in English native speaker. They found less faithful spectral imitation in ambiguous tokens than in unambiguous /ɛ/ and /æ/ tokens, suggesting that native phonological categories constrain phonetic imitation. In contrast, duration, the secondary cue in L1-English speakers' vowel categorization, was imitated gradiently. At the individual level, the speakers with a stronger reliance on spectral cues in perception exhibited a greater degree of duration imitation in production, implying that some English native speakers were more sensitive to acoustic-phonetic information in general, which was reflected in their duration imitation and showed the perception-production link at the acoustic-phonetic level.

In contrast, previous studies showed that L1-Mandarin speakers heavily rely on duration to distinguish English high vowels /i, ɪ/, while the role of duration is less clear for mid vowels /ɛ, æ/ [7, 8]. This highlights the fact that L1-Mandarin speakers use acoustic-phonetic cues to categorize English vowels differently from native speakers. The correlation between spectral and duration cue weights in L1-Mandarin speakers, however, remains unclear. Regarding the L2 perception-production link, most studies correlate perception and production measures through separate tasks and thus lead to mixed conclusions [5, 8, 17]. In addition, in terms of phonology, Mandarin lacks tense-lax vowel contrasts but has /i/ and [ɛ] (an allophone of /a/) in its inventory; also, Mandarin contrasts vowels in spectral but not durational aspects [13]. These indicate a possible modulation of L1 linguistic knowledge in the L2 perception-production link [5, 8, 17].

Given the differences in phonetic cue weightings and native phonological categories compared to English native speakers, we aim to investigate the perception-production link in L1-Mandarin speakers for different L2 English vowel contrasts.

2. METHODS

2.1. Participants

Twenty-five L1-Mandarin speakers were recruited. However, 9 and 6 participants were excluded from the

data of high and mid vowels, respectively, due to their mislabeling responses in the perception experiment (reversed pattern of identification), resulting in a sample of 16 participants for high vowels (8F, 8M; *M* = 24.3) and 19 participants for mid vowels (11F, 8M; *M* = 24.5). All participants were native Mandarin speakers and had learned English as a second language. Their self-reported English proficiency was 4.38 and 4.33 on a scale of 1 to 7 (1 being poor and 7 being proficient), and the onset of L2 exposure was 6.25 and 5.94 years of age, for the high vowel group and mid vowel group, respectively. None of the participants reported speech or hearing impairments.

2.2. Stimuli

2.2.1. Perceptual categorization

Bit-beat [bɪt-bit] and *bet-bat* [bɛt-bæt] continua were created from natural tokens produced by a phonetically trained male native English speaker in a sound-attenuated booth. To resynthesize the tokens, we first used Tandem-Straight [9] to manipulate spectral F1/F2 values of the vowels and created a seven-step continuum from /bɪt/ to /bit/ and from /bɛt/ to /bæt/. Next, the PSOLA technique was implemented in Praat [3] on each spectral step to manipulate the vowel durations in seven equal steps (40 ms/step) from 100 ms to 340 ms. In total, 49 stimuli (7 spectral steps x 7 duration steps) were created for each of the high and mid vowel pairs for the vowel categorization task (Fig.1, left panel).

2.2.2. Phonetic imitation

We included a subset of stimuli for the vowel categorization (steps 1, 4 and 7 at 140ms, 220ms and 300ms) along with additional perceptually salient tokens (steps 1, 4, 6 at 60ms and 380ms). In total, 15 stimuli (3 spectral steps [steps 1, 4, 7] x 5 duration steps [60ms, 140ms, 220ms, 300ms, 380ms]) were created for each of the high and mid vowel pairs for the phonetic imitation task (Fig.1, right panel).

2.3. Procedure

The participants completed the two-alternative forced choice perceptual vowel identification task before the phonetic imitation task. For the perception task, they were asked to categorize the stimuli they heard as *bit/beat* or *bet/bat* upon hearing stimuli randomly sampled from the *bit-beat* and *bet-bat* continua via E-Prime [15] (49 tokens x 5 repetition = 245 trials). For the imitation task, they were asked to imitate the stimuli presented to them twice in each trial. To establish the degree of imitation for each participant, baseline productions of the target words *bit/beat* and *bet/bat* were collected prior to the experiment for high and mid vowel groups, respectively (6 tokens from the 8 repetitions for each word).

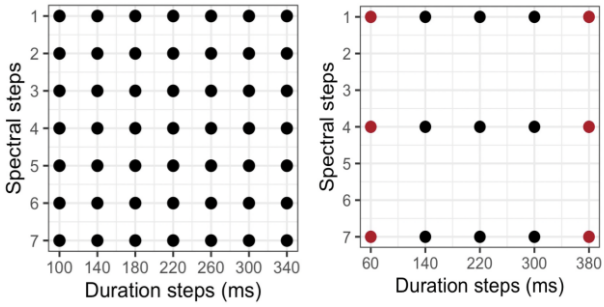


Figure 1: The stimuli for the perceptual categorization (left) and the phonetic imitation (right) task

3. RESULTS

3.1. Perceptual categorization

Participants' responses of *bit/bet* (lax vowels) were coded as 0 and *beat/bat* (tense vowels) as 1, as shown in Fig. 2. A number closer to 1 thus indicates a higher proportion of *beat/bat* responses. To examine the participants' responses as a function of spectral and durational steps at the group level, the data were submitted to two mixed-effect logistic regression models, one for high vowels and the other for mid vowels, using the *glmer* function in the *lme4* package in R [14]. The results showed that L1-Mandarin speakers' perceptual categorization of high vowels was significantly predicted by both spectral (β = 1.85, p <.01) and durational steps (β = 1.48, p <.001), with sharper discontinuities along the durational continuum; in contrast, mid vowel categorization was only significantly predicted by spectral steps (β = 2.46, p <.001; durational steps: β = -0.16, p =.47).

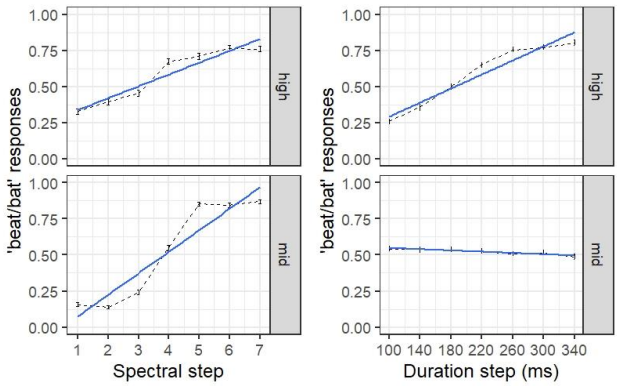


Figure 2: Proportion (dashed black lines) and estimation (solid blue lines) of the participants' *beat* responses (upper panel, high vowel) and *bat* responses (lower panel, mid vowel) along the spectral (left panel, step 1 = *bit/bet*; step 7 = *beat/bat*) and duration continua (right panel)

Next, we extracted by-participant spectral and durational coefficients from the omnibus model to

serve as the cue weights in perception at the individual level (Fig. 3). Further inspection revealed a *trade-off* relationship between the two cues for the high vowels in which the *more* the speakers relied on durational cues to categorize /i/ and /ɪ/, the *less* they relied on spectral cues (r = -.61, p < .05). No direct link was found between the two cues in /ɛ, æ/ categorization (r = .09, p = .69). These findings suggest the L1-Mandarin speakers' failure to integrate the two cues in English vowel perception.

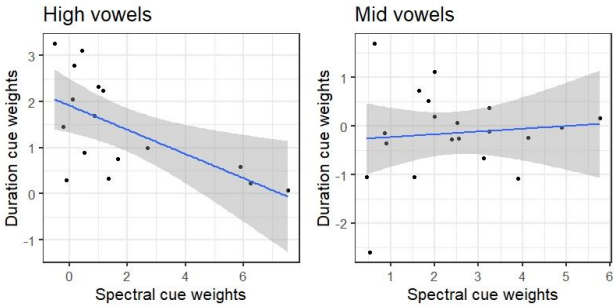


Figure 3: Correlation between individual spectral and duration cue weights in perception of high (left) and mid (right) vowels

3.1. Phonetic imitation

Considering the gender effect on formant values and the distribution of duration data, we standardized the raw F1/F2 values by participant and converted the values into Bark scales, and performed log-transformation for millisecond values of duration data for the subsequent analyses. In the speakers' imitated speech, ambiguous high vowels were produced categorically as /i/ indicated by the overlapping *beat* and *ambiguous* vowel ellipses, while ambiguous mid vowels were imitated gradiently along the spectral steps (Fig. 4). This can be interpreted as an L1 phonological effect: Mandarin /i/ warped the imitation of the ambiguous high vowels, but no such native magnet was available for the mid vowels. Duration, on the other hand, was imitated gradiently for both high and mid vowels, suggesting a faithful imitation of the durational cues in the stimuli (Fig. 5).

We further calculated the degrees of imitation for vowel quality (F1-F2 Euclidean distance) and duration, using the formula $|X_{target}-X_{baseline}| - |X_{target}-X_{imitation}|$ [16]. At the individual level, the speakers who weighted duration cues for the English high vowels more heavily in perception were able to imitate spectral (Fig. 6B, t = 2.64, p < .01) and duration (Fig. 6D, t = 3.91, p < .001) information more faithfully; in contrast, the spectral cue weights did not correlate with either spectral (Fig. 6A, t=0.82, p=.41) or duration (Fig. 6C, t=-1.45, p=.15) imitation. This suggests that both spectral and duration imitation of English high vowels seems to be tied to the perceptual sensitivity in duration, rather than spectral cues, in L1-Mandarin speakers, indicating that the formation of L2 /i/ and /ɪ/ categories by L1-Mandarin speakers is based on duration to a larger extent.

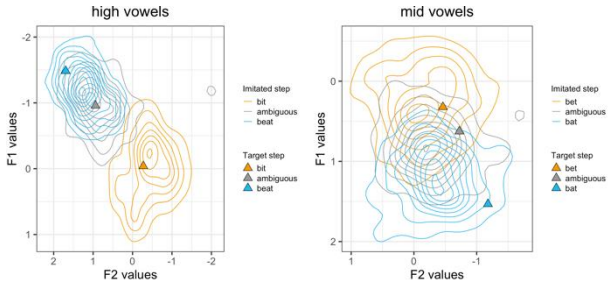


Figure 4: Imitated (colored ellipses) and target F1/F2 values (triangles) of high (left) and mid (right) vowels

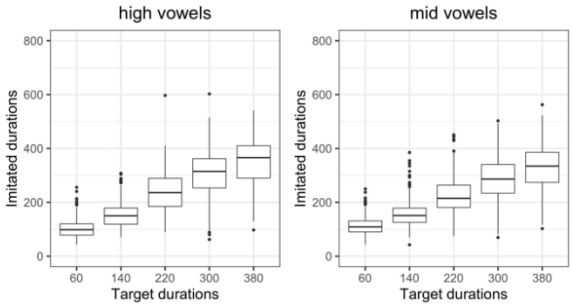


Figure 5: Imitated and target duration values of high (left) and mid (right) vowels

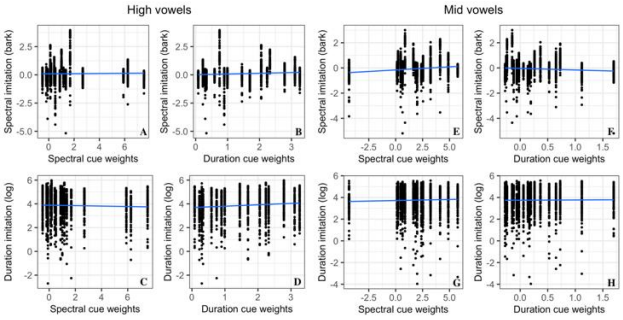


Figure 6: Correlation between individual perceptual cue weights (spectral/duration) and imitation (spectral/ duration) for high (left, A-D) and mid vowels (right, E-G)

For the mid vowels, however, the speakers who weighted spectral cues more heavily in perception displayed a higher degree of spectral imitation (Fig. 6E, t = 5.58, p <.001), while those who relied more on duration in perception showed lower degree of spectral imitation (Fig. 6F, t = -3.28, p < .01); on the other hand, the duration imitation of the mid vowels was not predicted by either spectral (Fig. 6G, t = 1.47, p = .14) or duration (Fig. 6H, t = 0.39, p = .70) cue weights. These results indicate that, for English mid vowels /ɛ/ and /æ/, the two acoustic-phonetic cues in perception constrained the production of vowel quality in a trade-off manner.

6. DISCUSSION

The current study adopted perceptual categorization and phonetic imitation tasks to examine the link between perception and production of English high (/i, ɪ/) and mid (/ɛ, æ/) vowel pairs in L1-Mandarin speakers.

In perception, L1-English speakers rely primarily on spectral cues across vowel contrasts, and less on durational cues. In contrast, we found that (1) L1-Mandarin speakers rely more on durational cues for the /i, ɪ/ contrast, with spectral cues being secondary, and (2) they rely more exclusively on spectral cues to contrast /ɛ, æ/ vowels. It appears that the L2 speakers developed contrast-specific phonetic cue weighting for vowel perception in a non-native way. Furthermore, at the individual level, we observed a trade-off relationship between the perceptual weights of spectral and durational cues for the high vowels, while a null relationship between the two was revealed for the mid vowels. Together, these results imply that L2 speakers may have difficulty simultaneously integrating various acoustic-phonetic cues to map sounds onto L2 categories.

With regard to the perception-production link, our findings suggest that the L1-Mandarin speakers' spectral imitation of L2 English high vowels was mediated by their L1 phonemic categories. During phonetic imitation, L1-Mandarin speakers may perceptually map the L2 English high vowel /i/ onto the native sound category /i/ in Mandarin [5, 8, 17] and thus condition the phonological contrast in their production, leading to the assimilated imitation of the ambiguous and /i/ vowel tokens. On the other hand, the speakers were able to imitate the spectral information of the mid vowels in a gradient manner. This is presumably due to the lack of influence from L1 phonology given no corresponding phonemic categories for both /ɛ/ and /æ/ in Mandarin, though [ɛ] sound does appear as an allophone under certain contexts. The vowel durations, on the other hand, were faithfully imitated for both high and mid vowels, which is in line with previous findings suggesting that suprasegmental properties of speech can be picked up more easily for imitation [11].

At the individual level, for high vowels, the speakers who relied more on durational cues in perception were better able to imitate *both* durational and spectral information of targets; in contrast, spectral cue weights did not correlate with either duration or spectral imitation. This suggests that the imitation of English high vowels, in terms of both vowel quality and duration, is linked to the sensitivity of the primary duration cue, but not that of the secondary spectral cues, in perception. Partly supporting phonological imitation hypothesis claiming all phonologically-relevant phonetic cues to be enhanced in production, our finding exhibits asymmetries between perception and imitation (cf. [12]). Despite the fact that spectral information serves as a secondary cue in L1-Mandarin speakers' perception and is thus encoded to some extent in their English /i, ɪ/ categories, duration, the primary cue, is better represented to contrast the /i, ɪ/ categories. As such, a greater perceptual sensitivity to duration, but not spectral cues, may help activate the L1-Mandarin speakers' L2 representations more robustly and lead to more accurate L2 imitation across different phonetic aspects.

For the mid vowels, however, the speakers who relied more on spectral cues in perception displayed a higher degree of spectral imitation, while those with larger reliance on durational cues imitated the spectral information more poorly. This finding suggests a direct perception-production link at the acoustic-phonetic level for the vowel quality of English mid vowels (cf. [11]); that is, greater sensitivity to spectral cues in perception leads to better imitation of vowel quality in the same phonetic aspect. On the other hand, if the speakers attended more to duration, they were less capable of imitating the spectral information closely. These results provide further evidence that L1-Mandarin speakers utilize primary and non-primary cues in a trade-off manner. In addition, the speakers' duration imitation was not predicted by either of the perceptual cue weights, which may reflect the minor role of vowel duration for the perceptual categorization of English mid vowels and hence result in an absence of perception-production link.

Taken together, L2 English vowel production in L1-Mandarin speaker is linked to perception only when the corresponding phonetic aspect(s) is critical for their perceptual categorization. For English high vowels, both the vowel quality and duration in production were related to the efficiency in using the primary (duration) cue in perception that is more robustly represented to distinguish L2 sounds; in addition, the imitation of spectral information, which is contrastive between Mandarin vowels, was further mediated by the L1 categories at the phonological level. On the contrary, for English mid vowels, the perception-production link is only established in vowel quality, directly constrained by a trade-off in perceptual sensitivity to the primary (spectral) and non-primary (duration) cue at the acoustic-phonetic level. Our findings highlight the roles of phonetic cue weightings and native phonological categories in shaping the abstraction of L2 sound categories during second language acquisition, leading to disparate perception-production links across various phonetic aspects in different L2 vowel contrasts.

7. ACKNOWLEDGEMENTS

We would like to thank Meng-Hsuan Lin and Yin-Ching Chang for their help on collecting data and data annotation. This work was supported by Ministry of Science and Technology of Taiwan (MOST110-2628-H-A49-001-MY2) to Yu-An Lu.

8. REFERENCES

[1] Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 171-204). Baltimore: York Press.

[2] Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

[3] Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.26). Retrieved from www.praat.org.

[4] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-276). Baltimore: York Press.

[5] Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

[6] Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of Acoustic Society of America*, 97(5), 3099-3111.

[7] Hsieh, B.-r., & Pan, H.-h. (2010). L2 experience and non-native vowel categorization of L1-mandarin speakers. Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

[8] Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, 119(2), 1118-1130.

[9] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. Paper presented at the Acoustics, Speech and Signal Processing, Las Vegas, NV.

[10] Kim, D., Clayards, M. & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. Linguistics Vanguard, 3(1), 20160025.

[11] Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, 34(6), 769-786.

[12] Kwon, H. (2019). The role of native phonology in spontaneous imitation: Evidence from Seoul Korean. Laboratory Phonology: *Journal of the Association for Laboratory Phonology*, 10(1).

[13] Lin, Y.-H. (2007). *The Sounds of Chinese*. Cambridge, UK: Cambridge University Press.

[14] R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria. URL http://www.R-project.org/: R Foundation for Statistical Computing.

[15] Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

[16] Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, 6, 546.

[17] Wang, X. (1997). The Acquisition of English Vowels by Mandarin ESL Learners: A Study of Production and Perception. Master's thesis, Simon Fraser University, Burnaby, BC, Canada.

A preliminary study of Mandarin neutral tone production by Japanese and Korean L2 learners

Tong Shu, Peggy Pik Ki Mok

The Chinese University of Hong Kong
tongshu@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

This study investigates the production of Mandarin neutral tone by L2 learners whose L1s were Japanese and Korean. Mandarin neutral tone (T0) is characterized by its underspecified pitch contours and shorter duration compared with the four lexical tones (T1-T4). These two important acoustic correlates of Mandarin neutral tone, pitch and duration, are phonemic in Japanese (lexical pitch accent and vowel length contrast) while non-phonemic in Korean. Following the Feature Hypothesis, Japanese speakers were predicted to outperform Korean speakers in producing the two features of neutral tone. 10 Japanese and 10 Korean speakers with intermediate proficiency in L2 Mandarin participated in a reading task. The results showed that both groups had little difficulty producing the context-conditioned pitch patterns of neutral tone, whereas Japanese speakers produced more native-like duration patterns of neutral tone than Korean speakers did, which partially supported the Feature Hypothesis.

Keywords: L2 production, Mandarin neutral tone, Japanese, Korean, Feature Hypothesis

1. INTRODUCTION

The acquisition of L2 stress has been extensively studied. Stress is a suprasegmental feature realized by configuring several acoustic correlates such as F0, duration, and vowel quality. Previous studies on the phonetic implementation of L2 stress found that L2 learners have difficulties producing each acoustic correlate of L2 stress, which seems to be influenced by how these acoustic correlates are adopted in their native prosodic systems [1]–[3]. McAllister et al. (2002) [4] proposed the Feature Hypothesis concerning L1 transfer at the featural level. According to the Feature Hypothesis, if an L2 contrast is based on a phonetic feature which is not prominent in the L1, L2 learners will have difficulty perceiving and producing this contrast [4]. In their study, this is supported by the finding that Estonian speakers, whose L1 has phonemic vowel length contrast, produced L2 Swedish vowel length contrast in a native-like manner, compared with English and Spanish speakers whose L1s had limited or non-contrastive use of duration. Following the Feature Hypothesis, Lee and Guion (2006) [1] investigated the production of English unstressed vowels by early and late Japanese- and Korean-English bilinguals. Their study found that Japanese speakers produced comparable duration ratios of unstressed to stressed vowels with native English speakers, while Korean speakers did not. This result corresponded with the role of duration in their L1s: Japanese has phonemic vowel length contrast based on duration [5] whereas Korean has lost this contrast [6]. However, both Japanese and Korean groups were native-like in producing the F0 peak ratio of unstressed to stressed vowels, which did not echo the role of pitch in the two languages, i.e., phonemic in Japanese (lexical pitch accent) [7] but not in Korean [8].

In Mandarin, there is a 'toneless' category called the neutral tone. It resembles English unstressed syllables in that it has restricted pitch patterns, shorter duration, and vowel reduction [9], [10]. However, the acquisition of Mandarin neutral tone by L2 learners has received little attention [11], [12]. Therefore, the current study compared the production of L2 Mandarin neutral tone by L1-Japanese and L1-Korean speakers, whose L1s differed in the phonological status of duration and pitch, the two acoustic correlates of Mandarin neutral tone.

1.1. Types and acoustic correlates of neutral tone words

The neutral tone in Mandarin cannot occur independently and usually follows lexical tones. There are three typical types of disyllabic neutral tone words [13]: 1) the suffix type: some suffixes are usually produced in neutral tone, such as *-de* in *dui4de0* 'right', *-le* in *zou3le0* 'gone' and so on; 2) the reduplication type: words derived from reduplication usually have the second syllable in neutral tone, such as *ge1ge0* 'brother'; 3) the lexeme type: words that are conventionally pronounced with the second syllable in neutral tone, such as *zhi1shi0* 'knowledge'; Among the three types, neutral tone in the lexeme type is unpredictable and must be learned in a word-by-word manner, whereas the reduplication type and suffix type are predictable.

The acoustic correlates of neutral tone are F0, duration, and vowel quality [9], [10], [14]. Its pitch contour is determined by the preceding lexical tones,

which is falling after T1/2/4 but with different pitch heights and level/rising after T3 [15]. The duration of neutral tone is about 50~70% of the preceding tone, and some studies found that neutral tone after T3 tends to have larger duration ratios [13], [14]. Neutral tone showed varying degrees of vowel reduction compared with lexical tones, but the role of vowel reduction in native perception is less studied compared with F0 and duration. The acoustic properties of neutral tone reported in the above studies are summarized below.

Preceding lexical tone	Neutral tone	
	Pitch	Duration
T1 [55]	mid falling [41]	50~60%
T2 [35]	high falling [52]	50~60%
T3 [21]	mid level/rising [33/34]	~70%
T4 [53]	low falling [21]	50~60%

Table 1: Pitch patterns and duration of neutral tone.

Following the Feature Hypothesis and previous studies [1], Japanese speakers were predicted to 1) better distinguish the pitch contours of neutral tone following different lexical tones, and 2) produce more native-like duration patterns of neutral tone than Korean speakers would, given the more prominent role of pitch and duration in Japanese than in Korean.

2. METHOD

2.1. Participants

Ten L1-Japanese (NJ) and 10 L1-Korean (NK) L2 Mandarin learners, as well as 10 native Beijing Mandarin (NM) speakers participated in this experiment (see Table 2 for details). All the L2 learners were intermediate learners with equivalent proficiency of HSK-3~4 level, but an independent t-test indicated that the NK group had a significantly larger mean length of learning (LoL) than the NJ group, $t(16.2) = -2.65$, $p = .02 < .05$.

Group	Number	Age in yrs	LoL in yrs
NJ	6F, 4M	18.8 (1.03)	0.97 (1.69)
NK	7F, 3M	22.4 (3.41)	2.71 (1.19)
NM	7F, 3M	27.0 (7.60)	NA

Table 2: Summary of participant information. Inside brackets are standard deviations.

2.2. Materials

56 common disyllabic neutral tone words, including the three typical types introduced in Section 1.1, were selected as stimuli. All words are high-frequency words selected from common Chinese textbooks for L2 learners.

2.3. Procedures

Due to the COVID-19 pandemic, face-to-face experiments were difficult. Recent studies [16], [17] have suggested the feasibility of remote data collection via video conferencing software like ZOOM [18] and smartphone recordings. For duration and F0, the major concerns of this study, previous studies have shown that the two measurements remain stable across different devices [19]–[23]. Therefore, the current study combined ZOOM and smartphone recordings for data collection. The production stimuli were randomized and shown to the participants in PowerPoint Slides via the share-screen function in ZOOM. The participants were instructed to record themselves on their smartphones. All the stimuli were embedded in a carrier phrase *wo3du2 __ san1ci4* 'I read __ three times' to avoid final lengthening. Both simplified Chinese characters and pinyin were presented, and the participants were corrected when they mispronounced the preceding tones. Before the formal experiment, the participants were given careful instructions about the procedure, and they were required to record a demo for us to check and optimize the recording quality.

The preceding tone and the neutral tone in each token produced by the L2 learners were identified by two native Beijing Mandarin speakers. When the two judges give contradictory judgments, the first author would listen to the confusing token again and make the final judgment. Only tokens identified as carrying neutral tone were included in subsequent analysis, which made up 86.6% of the original dataset (2915 out of 3360 tokens).

2.4. Acoustic analysis

First, the onset and offset of the neutral tone and its preceding tone were marked manually in Praat [24]. The onset was identified as the beginning of the vowel, and the offset was identified as the end of the voiced interval. Second, the pitch and duration information were extracted using ProsodyPro [25]. For pitch analysis, the F0 values of neutral tone syllables at 10 evenly distributed time points were obtained and converted to z-scores for further visualization. For duration analysis, the absolute duration of the neutral tone and its preceding tone was converted into a duration ratio = neutral tone duration (ms) / preceding tone duration (ms) as in [13].

3. RESULTS

3.1. Pitch patterns of neutral tone

Figure 1 shows the pitch contour of neutral tone (T0 hereafter) following different lexical tones produced

by the NJ, NK, and NM groups visualized via the Smoothing Spline ANOVA model using the "gss" and "ggplot" packages in R [26], [27].

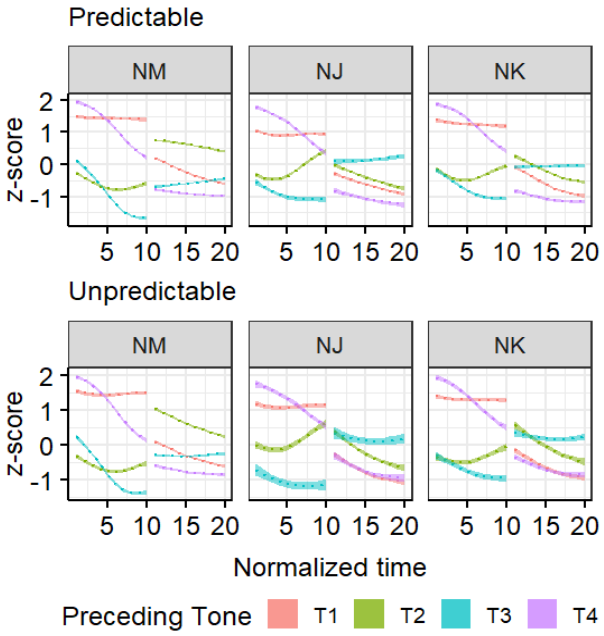


Figure 1: The pitch patterns of neutral tone after the four lexical tones produced by the NM, NJ and NK groups (Timepoint 1~10 represent the preceding lexical tone, and 11~20 represent the following neutral tone).

The pitch patterns of T0 produced by the NM speakers were consistent with previous studies (summarized in Table 1) for both the predictable and unpredictable types. Similarly, T0 produced by the NJ and the NK groups also clearly showed the falling vs. non-falling distinction between T0 after T1/2/4 vs. T3.

However, we can still observe several differences between the L2 learners and the NM speakers. First, the distribution of T0 after the four lexical tones in the NJ and the NK groups was generally less dispersed than in the NM group. Second, the relative position of T0 after T3 and the other T0s was different. In the NM group, T0 after T3 was placed in the middle position among all the T0s, while it had a higher overall pitch than the other T0s in both the NJ and the NK groups. Third, in terms of the relation between T0 and its preceding tone, in the NM group, the onset pitch of T0 after T2 was notably higher than the offset of its preceding tone, and the overall pitch level of T0 was higher than the preceding T2 as well. However, in both groups of L2 learners, the gap between the offset pitch of T2 and the onset pitch of the following T0 was much smaller than in the NM group.

Considering the effect of context, a difference in the shape of T0 after T3 can be observed between different contexts for both the NJ and the NK groups. It was level/rising in the predictable context while showing a slightly concave shape in the unpredictable context. Also, T0 after T1 and T4 largely overlapped in the unpredictable context, while they were better distinguished in the predictable context.

3.2. Duration ratios of neutral tone

A three-way mixed ANOVA was performed to examine the effects of Preceding Tone (T1 vs. T2 vs. T3 vs. T4), Type (predictable vs. unpredictable), and L1 (NJ vs. NK vs. NM) on the mean duration ratios of T0, with L1 as the between-subject factor, Preceding Tone and Type as the within-subject factors. The three-way interaction was significant, $F(6, 81) = 2.700$, $p = .02 < .05$. The simple two-way interaction between Group and Preceding Tone was significant for both the predictable type, $F(3.65, 49.29) = 5.954$, $p < .001$, and the unpredictable type, $F(6, 81) = 4.331$, $p < .001$. The simple one-way main effect of Group on mean duration ratios of neutral tone was statistically significant for the predictable type when the preceding tone was T2, $F(2, 27) = 11.960$, $p < .001$; T3, $F(2, 27) = 14.521$, $p < .001$; and T4, $F(2, 27) = 6.702$, $p = .004$. The simple one-way main effect of Group on mean duration ratios of neutral tone was statistically significant for the unpredictable type when the preceding tone was T2, $F(2, 27) = 6.673$, $p < .004$; T3, $F(2, 27) = 10.550$, $p < .001$; and T4, $F(2, 27) = 6.585$, $p = .005$.

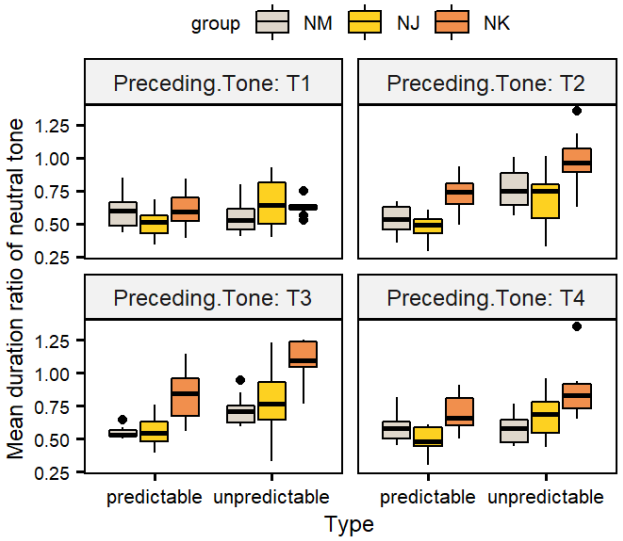


Figure 2: Mean duration ratios of neutral tone after four lexical tones produced by the three groups.

Pairwise comparisons (BH-adjusted) between groups were run where significant main effects were found. For T2 as the preceding tone, the NK group ($Mean = 0.717$, $SD = 0.133$) had significantly larger mean ratios than both the NM ($Mean = 0.531$, $SD = 0.111$) and the NJ group ($Mean = 0.475$, $SD = 0.101$), $p = .002 < .05$ and $p < .001$ respectively, while there was no significant difference between the NM and the NJ groups, $p = .29 > .05$. This result indicated that both the NJ and NM groups shorten the T0 syllable to

a similar degree while the NK speakers had longer T0 syllables. When the preceding tone was T3, the NK group (*Mean* = 0.841, *SD* = 0.202) had larger mean ratios than both the NM (*Mean* = 0.546, *SD* = 0.046) and the NJ group (*Mean* = 0.560, *SD* = 0.118), *p*s < .001, while the difference between the NM and the NJ groups was not significant, $p = .82 > .05$. When the preceding tone was T4, there was no significant difference between the NM (*Mean* = 0.587, *SD* = 0.117) and the NJ group (*Mean* = 0.491, *SD* = 0.102), $p = .09 > .05$, or between the NM and the NK groups (*Mean* = 0.689, *SD* = 0.140), $p = .09 > .05$. However, the NK group had significantly larger mean ratios than the NJ group, $p = .003 < .05$.

For the unpredictable type, when the preceding tone was T2, the NK group (*Mean* = 0.983, *SD* = 0.197) had larger mean ratios than both the NM (*Mean* = 0.762, *SD* = 0.162) and the NJ groups (*Mean* = 0.689, *SD* = 0.201), $p = .02 < .05$ and $p = .005 < .05$ respectively, while there was no significant difference between the NM and NJ groups, $p = .39 > .05$. When the preceding tone was T3, the NK group (*Mean* = 1.08, *SD* = 0.178) had larger mean ratios than both the NM (*Mean* = 0.715, *SD* = 0.112) and the NJ groups (*Mean* = 0.787, *SD* = 0.255), $p < .001$ and $p = .003 < .05$ respectively, while the difference between the NM and the NJ groups was not significant, $p = .41 > .05$. When the preceding tone was T4, the NK group (*Mean* = 0.859, *SD* = 0.203) had larger mean ratios than both the NM (*Mean* = 0.581, *SD* = 0.113) and the NJ groups (*Mean* = 0.681, *SD* = 0.191), $p = .004 < .05$ and $p = .045 < .05$ respectively, while the difference between the NM and the NJ groups was not significant, $p = .21 > 0.5$.

To conclude, for both types of T0 syllables, the NJ group's mean duration ratios of T0 were closer to the NM group than the NK group.

4. DISCUSSION

This preliminary study examined the production of L2 Mandarin neutral tone by Japanese and Korean speakers. In terms of pitch patterns, both the NJ and the NK speakers showed the context-conditioned pitch patterns of neutral tone like the NM speakers did, despite some target-deviant patterns observed in both groups. For example, the pitch patterns of neutral tone were consistent in the NM group regardless of context. In contrast, in both the NJ and the NK groups, the distribution of different neutral tones was less dispersed in the unpredictable context than in the predictable context. It was reasonable that neutral tone in the unpredictable context was more difficult, as it must be learned in a word-by-word manner, and was more likely to receive influence from its citation tone, compared with the other types, such as suffixes that do not have citation tones. Under the Feature Hypothesis [4], Japanese speakers were predicted to produce more accurate pitch contours of neutral tone than Korean speakers did because of the phonemic status of pitch in their L1. However, similar to Lee and Guion (2006)'s study [1], little difference was found between the NJ and the NK groups in terms of pitch. Note that there was a potential influence from the discrepancy in the length of learning between the two groups. The NK group had a longer mean length of learning than the NJ speakers did, which may obscure the cross-linguistic influence from native prosodic systems. However, as the NK group did not do better than the NJ group, the impact of the longer mean length of learning should be limited.

In terms of duration, the NJ speakers showed an advantage over the NK speakers in both the predictable and the unpredictable context. The NJ speakers' advantage over NK speakers in the predictable context was consistent with the Feature Hypothesis [4]. It is likely that NJ speakers' extensive experience manipulating duration to signify L1 vowel length contrast facilitated their production of the duration patterns of L2 Mandarin neutral tone. In addition, English learning experience may also facilitate the production of L2 Mandarin neutral tone. As mentioned in the Introduction, neutral tone resembles English unstressed syllables in terms of restricted F0, shorter duration, and vowel reduction. Therefore, they may adapt the duration pattern of English unstressed vowel to Mandarin neutral tone based on their similarity. It is possible since all the participants in the NJ and NK groups reported English learning experiences prior to Mandarin.

Only the hypothesis for duration but not pitch was supported in our study, which suggests that featural transfer based on phonemic status is not the whole story. More studies are needed to explore the potential factors modulating the transfer, such as the relative complexity of the prosodic contrast signified by the target feature in the L1 and L2. For example, in the current study, the L2 Mandarin tone has more extensive pitch variations than the L1 Japanese lexical pitch accent does. Thus, L1 experience with phonemic pitch may not necessarily facilitate the acquisition of a more complicated L2 contrast.

The present study has several limitations. First, we only compared the duration of neutral tone with its preceding tones, not with lexical tones occurring in the same positions, i.e., disyllabic-final. Second, we did not look at the influence of the onset type on the duration of neutral tone, which can be improved in future studies. Third, the sample size for each group was limited. To obtain a more comprehensive picture, more data from Japanese and Korean speakers with intermediate and higher proficiency is being collected.

5. ACKNOWLEDGEMENTS

The first author would like to thank Xiang Min, Zhao Ziyi, Zoe Cheung Sheung Yu for the help in data collection and annotation, Xu Huan for the guidance on statistical analysis, and all the participants who gave their time to participate in this study.

6. REFERENCES

[1] B. Lee, S. G. Guion, and T. Harada, "Acoustic analysis of the production of unstressed english vowels by early and late Korean and Japanese bilinguals," *Stud. Second Lang. Acquis.*, vol. 28, no. 3, pp. 487–513, 2006.

[2] W. Zuraiq and J. A. Sereno, "English Lexical Stress Cues in Native English and Non-Native Arabic Speakers," *ICPhS XVI*, no. August, pp. 829–832, 2007.

[3] Y. Zhang, S. L. Nissen, and A. L. Francis, "Acoustic characteristics of English lexical stress produced by native Mandarin speakers," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4498–4513, Jun. 2008.

[4] R. McAllister, J. E. Flege, and T. Piske, "The influence of L1 on the acquisition of Swedish quantity by native speaker of Spanish, English and Estonian," *J. Phon.*, vol. 30, no. 2, pp. 229–258, 2002.

[5] Y. Hirata, "Effects of speaking rate on the vowel length distinction in Japanese," *J. Phon.*, vol. 32, no. 4, pp. 565–589, 2004.

[6] Y. Kang, T.-J. Yoon, and S. Han, "Frequency effects on the vowel length contrast merger in Seoul Korean," *Lab. Phonol.*, vol. 6, no. 3–4, pp. 469–503, 2015.

[7] S. Kawahara, "The phonology of Japanese accent," in *Handbook of Japanese phonetics and phonology*, M. Shibatani and T. Kageyama, Eds. De Gruyter, 2015, pp. 445–492.

[8] H.-S. Jeon, "Prosody," in *The handbook of Korean linguistics*, L. Brown and J. Yeon, Eds. John Wiley & Sons, Inc., 2015, pp. 41–58.

[9] M. Lin and J. Yan, "Beijinghua qingsheng de shengxue xingzhi[Acoustic properties of neutral tone in Beijing Mandarin]," *Fangyan*, no. 3, pp. 166–178, 1980.

[10] J. F. Cao, "Putonghua qingsheng yinjie texing fenxi[Analysis of the characteristics of Mandarin neutral tone syllables]," *Yingyong shengxue*, vol. 5, no. 4, pp. 1–6, 1985.

[11] C. B. Chang and Y. Yao, "Production of neutral tone in mandarin by heritage, native, and second language speakers," in *the 19th International Congress of Phonetic Sciences*, 2019, pp. 2291–2295.

[12] P. Tang, "Riben gaoji hanyu xuexizhe hanyu qingsheng yunlv xide pianwu fenxi[A study of prosodic errors of Chinese neutral tone by advanced Japanese students]," *Huawen jiaoxue yu yanjiu*, no. 4, pp. 39–47, 2014.

[13] P. Tang, I. Yuen, N. Xu Rattanasone, L. Gao, and K. Demuth, "Acquisition of weak syllables in tonal languages: Acoustic evidence from neutral tone in Mandarin Chinese," *J. Child Lang.*, vol. 46, no. 1, pp. 24–50, Jan. 2018.

[14] J. F. Cao, "On neutral-tone syllables in Mandarin Chinese," *Can. Acoust.*, vol. 20, no. 3, pp. 49–50, 1992.

[15] W.-S. Lee and E. Zee, "Prosodic characteristics of the neutral tone in Beijing Mandarin," *J. Chinese Linguist.*, vol. 36, no. 1, pp. 1–29, 2008.

[16] N. H. Hilton and A. Leemann, "Editorial: Using smartphones to collect linguistic data," *Linguist. Vanguard*, vol. 7, no. s1, pp. 1–7, 2021.

[17] A. Leemann, P. Jeszenszky, C. Steiner, M. Studerus, and J. Messerli, "Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing," *Linguist. Vanguard*, vol. 6, no. s3, 2020.

[18] I. Zoom Video Communications, "ZOOM cloud meetings." 2021, [Online]. Available: https://zoom.us/.

[19] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements," in *Proceedings of INTERSPEECH 2021*, 2021, pp. 3984–3988.

[20] P. De Decker and J. Nycz, "For the record: Which digital media can be used for sociophonetic analysis?," *Univ. Pennsylvania Work. Pap. Linguist.*, vol. 17, no. 2, p. Article 7, 2011.

[21] E. U. Grillo, J. N. Brosious, S. L. Sorrell, and S. Anand, "Influence of Smartphones and Software on Acoustic Voice Measures.," *Int. J. Telerehabilitation*, vol. 8, no. 2, pp. 9–14, 2016.

[22] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 3910–3916, 2021.

[23] Y. Guan and B. Li, "Usability and Practicality of Speech Recording by Mobile Phones for Phonetic Analysis," *2021 12th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2021*, 2021.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." 2021, [Online]. Available: http://www.praat.org/.

[25] Y. Xu, "ProsodyPro - A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 2013, pp. 7–10.

[26] C. Gu, "Smoothing spline ANOVA models: R package gss," *J. Stat. Softw.*, vol. 58, pp. 1–25, 2014.

[27] H. Wickham, *ggplot2: elegant graphics for data analysis*. New York: Springer, 2009.

THE CLASSIFICATION OF NATIVE AND NON-NATIVE PRODUCTIONS OF ITALIAN CONSONANT LENGTH CONTRAST

Qiang Feng, M. Grazia Busà

Università degli Studi di Padova
qiang.feng@phd.unipd.it, mariagrazia.busa@unipd.it

ABSTRACT

It has been suggested that the duration ratio between Italian consonants and preconsonantal vowels (C/pre-V ratio) is a potentially good candidate for automatic classification of Italian singleton vs. geminate consonants produced by native speakers. The optimal C/pre-V ratio threshold value for discriminating between these two categories of consonants is about 1. This triggers the question of whether this threshold also applies to the classification of non-native productions of Italian consonant length contrast. To answer this question, in this study we classified Italian singleton vs. geminate consonants produced by Chinese speakers and native Italian speakers using two approaches, namely (a) relying on native listeners' perceptual identification and (b) using the above threshold. A comparison of the classification results shows significant differences between the two approaches, suggesting that, for both native and non-native productions, using the C/pre-V ratio threshold value of 1 cannot always accurately classify Italian singleton vs. geminate consonants.

Keywords: Consonant length, singleton, geminate, Italian, Chinese.

1. INTRODUCTION

Italian consonants (i.e., /b, d, g, p, t, k, tʃ, dʒ, f, v, s, m, n, l, r/) can contrast in phonemic length. Some examples are *pena* 'penalty' vs. *penna* 'pen', *papa* 'pope' vs. *pappa* 'mush', etc. Previous investigations on the status of Italian consonant length contrast [1]–[3] show that (i) Italian geminate consonants are longer than their singleton counterparts, and (ii) Italian pre-geminate vowels are shorter than the corresponding pre-singleton ones because of an anticipatory durational compensation. Take the minimal pair of *pena* vs. *penna* as an example: the "nn" in *penna* is longer than the "n" in *pena*, and the "e" in *penna* is shorter than the "e" in *pena*.

However, what noted above seems to hold only within a given speech rate. This is because Italian consonant length is found to vary with speaking rates, and a singleton consonant produced at a slow speaking rate may be as long as a corresponding geminate produced at a fast speech rate [4]. Therefore, consonant duration may not be a suitable criterion for the classification of Italian singleton vs. geminate consonants across speech rates. By contrast, according to [4], the duration ratio between Italian consonants and preconsonantal vowels (i.e., C/pre-V ratio) may serve as a better candidate for the classification of Italian consonant length contrast across speech rates. This was confirmed by [1], [2], who showed that the optimal C/pre-V ratio threshold value for discriminating between Italian singleton and geminate consonants is about 1. Specifically, if the C/pre-V ratio is greater than 1, a geminate consonant will be detected; otherwise, it will be a singleton.

This triggers an interesting question. That is, could the C/pre-V ratio threshold value of 1 that is optimal for discriminating between Italian singleton and geminate consonants produced by native Italian speakers also apply to non-native productions?

This question is raised for two reasons. First, according to the cue-weighting theory, in determining the perceptual identity of a sound, some acoustic dimensions play a greater role than others [5]. In our case, the C/pre-V ratio works better than the absolute consonant duration in classifying native Italian speakers' singleton vs. geminate consonants. Thus, it is reasonable to speculate that Italian listeners, in perceptually identifying native productions of Italian consonant length contrast, may pay more attention to the duration interplay between Italian consonants and preconsonantal vowels than to the absolute consonant duration. However, we do not know whether Italian listeners also apply this cue-weighting strategy to their perception of non-native productions of Italian consonant length contrast. Moreover, as shown in previous studies [6]–[12], Italian singleton and geminate consonants produced by non-native speakers often differ from those produced by native Italian speakers. Therefore, it seems interesting to examine whether the ratio threshold above is also valid for the classification of non-native productions of Italian consonant length contrast.

Second, empirical studies on L2 speech acquisition often rely on the perception of native speakers to assess non-native productions. For example, to know whether an Italian consonant produced by a non-native speaker is a singleton or geminate, the most reliable way is to find out how native Italian speakers identify it perceptually.

However, the recruitment of native speakers and the development of perceptual identification tasks can be rather effortful and time-consuming. Thus, if using the above threshold could represent an alternative, more economic approach for easily and accurately classifying Italian singleton vs. geminate consonants produced by non-native speakers, future research might become less burdensome. For this reason, we wanted to test whether the above threshold can be considered valid for the classification of non-native productions of Italian consonant length contrast.

In this study, we will use both native Italian speakers' perceptual identification and the above-mentioned C/pre-V ratio threshold to classify the production of Italian singleton vs. geminate consonants by Chinese speakers and native Italian speakers. Through a comparison of the classification results of these two approaches, we aim to assess whether the C/pre-V ratio threshold value of 1 can be used to accurately classify non-native (as well as native) productions of Italian consonant length contrast.

2. METHOD

Three perceptual identification tasks were developed using the ExperimentMFC 7 in Praat [13]. The tokens to be identified in these tasks were as described in [6]. In short, they were 5 Italian minimal pairs contrasting in consonant length produced by 30 Chinese speakers and 10 native Italian speakers, both in isolation and in carrier sentences. In total, 1600 tokens were elicited (40 tokens × 40 speakers = 1600).

2.1. Listeners

Three native Italian-speaking undergraduates (one 23-year-old female and two 24-year-old males) were recruited as listeners. They were from the Veneto region in the North-East of Italy, the Abruzzo region in central Italy, and the Puglia region in the South of Italy, respectively. The reason for recruiting listeners from three different regions in Italy was to minimize the dialectal influence on the perceptual identification tasks.

2.2. Stimuli

We divided the above 1600 tokens into 3 parts. Each part served as the stimuli for one identification task. Table 1 shows the stimulus composition of each task. As the table shows, each task consisted in the identification of a set of tokens produced by 12 Chinese speakers and 4 Italian speakers; 12 of them were different across the three tasks and the remaining 4 (in bold) were identical. So, each task had 640 stimuli (40 tokens × 16 speakers = 640), 480 of which (40 tokens × 12 speakers = 480) were different across tasks and the remaining 160 (40 tokens × 4 speakers = 160) were identical.

Task	Stimuli: tokens produced by
1	CS-01, CS-02, CS-03, CS-04, CS-05, CS-06, CS-07, CS-08, CS-09, **CS-28, CS-29, CS-30,** IS-01, IS-02, IS-03, **IS-10**
2	CS-10, CS-11, CS-12, CS-13, CS-14, CS-15, CS-16, CS-17, CS-18, **CS-28, CS-29, CS-30,** IS-04, IS-05, IS-06, **IS-10**
3	CS-19, CS-29, CS-21, CS-22, CS-23, CS-24, CS-25, CS-26, CS-27, **CS-28, CS-29, CS-30,** IS-07, IS-08, IS-09, **IS-10**

Table 1: Stimulus composition of the identification tasks (CS=Chinese speaker, IS=Italian speaker).

At a perceptual examination, 43 tokens were found to contain a conspicuous misreading of the target words (e.g., *papa* 'pope' misread as *papà* 'father') and therefore were discarded. Thus, rather than 640 the actual number of stimuli in Task 1, 2, and 3 was 624, 622, and 625 respectively; and the actual number of identical stimuli in the three tasks was 157 (rather than 160).

2.3. Procedure

The listeners were randomly assigned to one of the three identification tasks. They completed the tasks with headphones on in front of a computer in the Language and Communication Lab at the University of Padova in Italy.

Each task consisted of two parts. In the first part, the tokens produced in isolation were presented in random order. Two words (i.e., a minimal pair) were shown on the computer screen as options, along with an "oops" button. The listeners were asked to click on the option that corresponded to the word heard. If they accidentally clicked on the undesired option, they could go back to the word heard previously by clicking on the "oops" button. Each word appeared only once and could be listened to up to three times when it appeared. After every 30 words there was a break.

The second part was structured similarly to the first one, except that the stimuli were produced in carrier sentences. The listeners' task was to determine which of the two options on the computer screen appeared in the sentences heard.

The tasks were self-paced, and each lasted about 25 minutes. At the end of the tasks, the listeners' responses that contained target geminate consonants (e.g., *penna*, *pappa*) were extracted as "geminate", and those containing target singletons (e.g., *pena*, *papa*) were extracted as "singleton".

2.4. Data preparation

In the first place, we had to determine the responses to the 157 stimuli that were identical across the three identification tasks. This is because all three listeners had identified them, and thus we had three different sets of responses to the stimuli. When the listeners' responses were divergent, the determination was made following the "majority rule". For example, if a stimulus was classified as "singleton" by two listeners and as "geminate" by one listener, the response would be counted as "singleton". In total, 35 stimuli required the application of the "majority rule". Afterwards, we merged the listeners' responses to the 157 stimuli with their responses to the other stimuli. In this way, we obtained the first set of classification results, which was based on the native Italian listeners' perceptual identification.

Secondly, we measured the durations of the target consonants and preconsonantal vowels in each token. Based on the duration values, we calculated the C/pre-V ratios for all tokens. The C/pre-V ratios of 27 tokens could not be calculated because the target consonants and/or preconsonantal vowels were unmeasurable. The tokens with a C/pre-V ratio smaller than or equal to 1 were classified as "singleton", and those with a C/pre-V ratio greater than 1 were classified as "geminate". In this way, we obtained the second set of classification results, which was based on the tokens' C/pre-V ratios.

2.5. Analyses

We first calculated the interrater reliability using the irr package 0.84.1 [14] in R 3.6.3 [15] based on the three listeners' responses to the 157 stimuli that were identical across the three identification tasks.

Then, we calculated the overlap rate between the two sets of classification results. By overlap we refer to the cases in which the two approaches classified one token in the same way. For example, if a token was classified as "geminate" by both approaches, there was an overlap between the two sets of classification results.

Afterwards, we calculated the Chinese speakers' and native Italian speakers' production accuracy rates. Specifically, if the speakers' target consonants were produced as intended (e.g., a "*pena*" was classified as "singleton"), they would be noted as correct. Otherwise (e.g., a "*pena*" was classified as "geminate"), they would be noted as incorrect.

Finally, we applied a generalized linear mixed-effects model (GLMM) with a binomial link function to the classification results (i.e., singleton vs. geminate) using the lme4 package 1.1.26 [16] in R 3.6.3 [15]. The GLMM had Approach (two levels: relying on native Italian listeners' identification ["by identification" for short] vs. using the C/pre-V ratio threshold ["by ratio" for short]), Group (two levels: native vs. non-native), Consonant type (two levels: singleton vs. geminate), and the interaction between them as fixed factors, and Subject and Item as random intercepts. The main effects of the fixed factors were assessed with the Type II Wald chi-squared tests using the car package 3.0.10 [17]. Post-hoc pairwise comparisons with FDR (false discovery rate) correction were implemented using the emmeans package 1.5.3 [18].

3. RESULTS

The interrater reliability was moderately high (Fleiss $k = 0.7$), indicating that though the three native Italian listeners were from different regions of Italy, they generally agreed with each other in identifying Italian singleton/geminate consonant contrast perceptually. Thus, the first set of classification results obtained on the basis of the native listeners' perceptual identification is reliable.

The overlap rate between the two sets of classification results reached 89.65% and 77.07% for the tokens produced by the native Italian speakers and the Chinese speakers, respectively. This indicates that there was more agreement between the two approaches in classifying native productions than in classifying non-native ones.

The Chinese speakers' and native Italian speakers' production accuracy rates for Italian singleton and geminate consonants are shown in Table 2. As can be seen, the production accuracy rates calculated by the two classification approaches reveal some common tendencies. That is, the Chinese speakers had lower production accuracy for either Italian singleton or geminate consonants as compared to the native Italian speakers; and the Chinese speakers' production accuracy was higher for Italian singleton consonants than for geminates. However, it should also be noted that the accuracy rates calculated by the two classification approaches are not exactly the same. Specifically, the accuracy rates calculated on the basis of the Italian listeners' identification were almost always higher than those calculated on the basis of the C/pre-V ratio threshold.

		By identification	By ratio
Non-native	Singleton	66.67%	71.50%
	Geminate	46.63%	38.51%
Native	Singleton	95.48%	88.94%
	Geminate	96.00%	89.50%

Table 2. Chinese speakers' (non-native group) and Italian speakers' (native group) production accuracy rates for Italian singleton and geminate consonants

Table 3 shows the results of the GLMM applied to the classification results of the native and non-native productions of Italian consonant length contrast. As can be seen, the GLMM yielded a significant main effect on the interaction between Approach, Group and Consonant type. Since we were particularly interested in whether there were any differences between the two classification approaches, we conducted post-hoc pairwise comparisons. The results, in Table 4, show that in the classification of native and non-native productions of both Italian singleton and geminate consonants, the two approaches performed significantly differently.

Fixed effects	χ^2	*df*	*p*
Approach (A)	13.63	1	0.0002*
Group (G)	3.85	1	0.0498*
Consonant type (C)	6.10	1	0.0135*
A×G	1.23	1	0.2669
A×C	1.45	1	0.2288
G×C	232.58	1	<.0001*
A×G×C	8.97	1	0.0027*

Table 3: Results of the GLMM applied to the classification results of the native and non-native productions of Italian consonant length contrast.

By identification vs. By ratio					
		Est.	*SE*	*z*	*p*
Non-native	Singleton	0.37	0.15	2.53	0.0115*
	Geminate	0.41	0.14	2.94	0.0032*
Native	Singleton	-0.91	0.42	-2.17	0.0303*
	Geminate	1.06	0.44	2.39	0.0171*

Table 4: Results of the post-hoc pairwise comparisons of the two classification approaches.

4. DISCUSSION AND CONCLUSIONS

In this study, we classified native and non-native productions of Italian consonant length contrast using two approaches, namely (a) relying on native listeners' perceptual identification and (b) using the consonant to preconsonantal vowel duration ratio (C/pre-V ratio) threshold (i.e., 1). The results are as follows.

Regarding native productions, there was a large degree of overlap between the classification results obtained with the two approaches, suggesting that the C/pre-V ratio does serve as an important acoustic cue in Italian listeners' perceptual categorization of Italian consonant length contrast. Nonetheless, the differences between the two approaches were significant, indicating that the (b) approach cannot always accurately classify Italian consonant length contrast produced by native Italian speakers. This is in line with [1], [2], who showed that when the C/pre-V ratio threshold value of 1 is applied to the classification of native Italian speakers' singleton vs. geminate consonants, certain error rates exist. We argue that this may be because the C/pre-V ratio is an important but not the only acoustic cue that could determine the perceptual identification of Italian consonant length category. In some cases, other acoustic cues (e.g., absolute consonant duration) may determine the perceived category membership.

As for the classification of non-native productions, the differences between the two approaches were also significant. Moreover, the differences were more evident than in the classification of native productions. This suggests that for Italian listeners the C/pre-V ratio plays a less important role in their perceptual identification of non-native productions than in their perceptual identification of native ones. We argue that this is because the Chinese speakers' productions of Italian consonant length contrast differed greatly from the native Italian speakers'. Specifically, the Chinese speakers completely ignored the duration interplay between Italian consonants and preconsonantal vowels in production [6]. Maybe for this reason, native Italian listeners are forced to rely more on other acoustic cues (e.g., absolute consonant duration) in perceptually classifying non-native productions than in perceptually categorizing native ones. In other words, Italian listeners apply somewhat different cue-weighting strategies to their perception of non-native productions of Italian consonant length contrast.

To conclude, for both native and non-native productions, using the C/pre-V ratio threshold cannot always accurately classify Italian singleton vs. geminate consonants. Therefore, for future research aiming to classify Italian consonant length contrast produced by non-native speakers, this approach cannot replace native speakers' perceptual identification. However, this study only classified Italian singleton vs. geminate consonants produced by Chinese speakers. To consolidate the present conclusions, future studies including non-native productions by speakers of other languages are needed.

5. REFERENCES

[1] M. G. Di Benedetto and L. De Nardis, "Consonant gemination in Italian: The nasal and liquid case," *Speech Communication*, vol. 133, pp. 62–80, 2021, doi: https://doi.org/10.1016/j.specom.2021.07.006.

[2] M. G. Di Benedetto and L. De Nardis, "Consonant gemination in Italian: The affricate and fricative case," *Speech Communication*, vol. 134, pp. 86–108, Nov. 2021, doi: 10.1016/j.specom.2021.07.005.

[3] A. Esposito and M. G. Di Benedetto, "Acoustical and perceptual study of gemination in Italian stops," *The Journal of the Acoustical Society of America*, vol.

106, no. 4, pp. 2051–2062, Oct. 1999, doi: 10.1121/1.428056.
[4] E. R. Pickett, S. E. Blumstein, and M. W. Burton, "Effects of Speaking Rate on the Singleton/Geminate Consonant Contrast in Italian," *Phonetica*, vol. 56, no. 3–4, pp. 135–157, Dec. 1999, doi: 10.1159/000028448.
[5] L. L. Holt and A. J. Lotto, "Cue weighting in auditory categorization: Implications for first and second language acquisition," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3059–3071, 2006.
[6] Q. Feng and M. G. Busà, "Mandarin Chinese-speaking learners' acquisition of Italian consonant length contrast," *System*, vol. 111, p. 102938, Dec. 2022, doi: 10.1016/j.system.2022.102938.
[7] C. Celata and L. Costamagna, "Timing delle geminate nell'italiano L2 di apprendenti estoni," *Quaderni del laboratorio di linguistica*, vol. 10, no. 1, pp. 1–27, 2011.
[8] B. Kabak, T. Reckziegel, and B. Braun, "Timing of second language singletons and geminates," in *Proceedings of the 17th International Congress of the Phonetic Sciences (ICPhS XVII)*, W.-S. Lee and E. Lee, Eds., Hong Kong: City University of Hong Kong, 2011, pp. 994–997.
[9] S. D'Apolito and B. G. Fivela, "L2 Pronunciation Accuracy and Context: A Pilot Study on the Realization of Geminates in Italian as L2 by French Learners.," in *INTERSPEECH*, 2019, pp. 1706–1710.
[10] B. De Clercq, E. Simon, and C. Crocco, "Rosa versus rossa: The acquisition of Italian geminates by native speakers of Dutch," *Phrasis: Studies in Language and Literature*, vol. 2, no. 2, pp. 3–29, 2014.
[11] Q. Feng and M. G. Busà, "How Do L1 Mandarin Chinese Learners Produce L2 Italian Singleton and Geminate Consonant Contrasts?," *Pronunciation in Second Language Learning and Teaching Proceedings*, vol. 12, no. 1, 2022.
[12] V. Galatà, G. Meneguzzi, L. Conter, and C. Zmarich, "Primi dati sull'acquisizione fonetico-fonologica dell'italiano L2 in prescolari rumeni," in *La voce nelle applicazioni*, A. Paoloni and M. Falcone, Eds., Roma: Bulzoni Editore, 2012, pp. 35–50.
[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2020. [Online]. Available: http://www.praat.org/
[14] M. Gamer, J. Lemon, and I. Singh, "irr: Various Coefficients of Interrater Reliability and Agreement." 2019. [Online]. Available: https://CRAN.R-project.org/package=irr
[15] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/
[16] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.
[17] J. Fox and S. Weisberg, "An {R} Companion to Applied Regression." Sage, Thousand Oaks CA, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
[18] R. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2020. [Online]. Available: https://CRAN.R-project.org/package=emmeans

ACQUISITION OF THE LOW VOWEL IN L2 SPANISH BY CHINESE SPEAKERS

Cao, Yongfa[1], Font-Rotchés, Dolors[1], Rius-Escudé, Agnès[1]

[1] Universitat de Barcelona
caoyf1@outlook.com; dolorsfont@ub.edu ; agnesrius@ub.edu

ABSTRACT

This research aims to characterise the timbre of the low vowel [a] in the spontaneous speech of the Spanish interlanguage spoken by Chinese people at different learning stages and to determine the convergent and divergent features of Peninsular Spanish. In addition, we will assess variables including L1, gender, level of Spanish proficiency and (a)tonicity of the vowel to see the extent to which these factors influence the acquisition of the sound. Using methods of quantitative linguistic typology, calculation of *accuracy* and *precision*, the results indicate that in the interlanguage (IL) the [a] is more open than in Spanish; that the IL is a dynamic system that improves the timbre of the vowel [a] with increased Spanish input; that gender and (a)tonicity factors influence learning; and that L1 influences the vowel [a] in the IL for beginner learners, with fossilisation occurring at higher levels.

Keywords: Low vowel, acoustic analysis, L2 Spanish, spontaneous speech, Chinese speakers

1. INTRODUCTION

1.1. Theoretical Framework

The phonetic categories that make up the subsystems of L1 and L2 interact dynamically and they are updated when the L2 input increases statistically and the L1-L2 combined categories (diaphones) change [1]. The phonetic differences between an L2 sound and an L1 sound are closer because the quantity and quality of L2 input affect the formation of new L2 phonetic categories [2].

The interlanguage (IL) [3], a language generated in the process of L2 acquisition, which is the natural result of blending the learner's L1 and the target language, can be described as systematic, dynamic, reduced and variable, and can present backsliding [4]. Fossilisation is another important phenomenon of the IL process, according to which there is a permanent cessation of the development of a linguistic structure [5].

Many phoneticians think that Spanish and Mandarin Chinese share the same vowel system (e.g [6]) but one of the most relevant differences between the two is that Mandarin is a tonal language [7]. Therefore, Spanish vowels are not "identical" to the Mandarin category but are "similar" to Mandarin, they are assimilated into the L1 category. These similarities can lead to difficulties because, despite the phonetic differences between L2 and L1 sounds, the learner continues to use the L1 category in both perception and production [8].

Specifically, the [a] in Mandarin, has two variants: [a] and [ɑ], but neither phoneme corresponds exactly to the Spanish [a] [9]. The timbre of the Spanish IL [a] spoken by Chinese speakers is more unstable than that of L1 Spanish [10] and is more open [11] [12] (opposing view, e.g. [13]), with no distinction between tonicity and atonicity and with a greater tendency to internal oscillation [14].

To date, research on IL Spanish spoken by Chinese speakers has been based on limited corpora of informants with basic or independent linguistic competence in controlled speech, ignoring the dynamic nature of the IL and the naturalness of human communication — spontaneous speech —, which is very different from controlled speech[15].

1.2. Aims of the Study

In this study, we aim to characterise the timbre of the vowel [a] in the spontaneous speech of IL Spanish spoken by Chinese speakers at different learning stages. Based on previous literature, we established the following objectives:

a. Determine the differences and similarities between the [a] sound in Mandarin, the IL and Spanish.
b. Analyse whether gender, level of Spanish proficiency and (a)tonicity are influential factors in L2 acquisition.
c. Determine whether there is L1 interference in the IL.

2. CORPUS AND METHODOLOGY

2.1. Corpus

Three corpora of spontaneous speech were created for this research:

A corpus of interlanguage Spanish spoken by Chinese speakers of Mandarin (IL). This is formed by

thirty-six native Chinese speakers (F=18, M=18, age=24 ± 6) studying Spanish philology at a university in China. A total of 814 [a] vowels have been obtained from monosyllables (CV, VC, CVC, CCV, CCVC) uttered by three groups of speakers with different levels, according to CEFR[16]: group A basic, group B independent and group C proficient.

A corpus of L1 Mandarin (L1). This is formed by ten native Chinese speakers (F=5, M=5, age=25 ± 5), from whom sixty [a] vowels have been obtained.

A corpus of Spanish (Sp). This has been obtained from the [17,18] corpus, with seventy-nine Peninsular Spanish informants (F=35, M=44) of different ages, from which we have analysed 298 [a] vowels.

2.2. Methodology

The acoustic analysis of F_1 and F_2 of low vowel sounds was carried out using the Praat program [19] and their values were obtained from the central value of the formant if it was stable, or from the mean value of the formant, if it was not.

First, the Mandarin (L1) values were compared with the IL and Spanish (Sp) values using multiple comparison tests: the Scheffe test was applied if the data complied with the normality test and homogeneity of variance test, if not, the Games-Howell test was applied.

The three IL groups were then compared with the target language based on two analysis models: *accuracy* and *precision*. To measure *accuracy*, we chose the ANOVA test, using the SPSS program, version 26, and for *precision,* we compared the centrality (dispersion) of the results for L1 Spanish and the IL, according to the [20] formulae:

(1)

$$p1(p2) = \frac{\mathrm{F1(F2)Max.\,IL} - \mathrm{F1(F2)Min.\,IL}}{\mathrm{F1(F2)Max.\,L1} - \mathrm{F1(F2)Min.\,L1}}$$

(2) $$p = p1 \times p2$$

From this methodology, the results indicate that the multiple relationships between the degree of dispersion (p) of an IL vowel and that of Spanish: $p1$ indicate the difference in the openness between the IL and the L1, and p*2* reflects the difference in the frontness.

3. RESULTS

3.1. The [a] Sound of Mandarin, IL and Spanish

In this section, we will look at the tonic and atonic variants of [a] uttered by males and females, given the significant differences found in the F_1 [21].

Concerning males (graph 1), the differences between L1 Chinese and the IL are always significant, in both F_1 (p<.001 stressed/tonic (S) and unstressed/atonic (U)) and F_2 (p<.001, S and U). Differences are also significant between Mandarin and Spanish in F_1 and F_2, for tonic and atonic vowels (p<.001).

On the other hand, differences between the IL and Spanish are shown in the F_1 of tonic vowels and the F_2 of atonic vowels (p<.001) and are similar in the F_1 of atonic vowels (p=.082) and the F_2 of tonic vowels (p=.249).

The [a] vowel of Mandarin is more open (upper F_1) and more back (lower F_2) than that of the IL, and the [a] of IL is also more open and back compared to that of Spanish. The F_1 and F_2 show an evolution of the sound in the IL towards the L2 Spanish but without achieving it: the tonic [a] of the IL should be more closed and the atonic, more front.

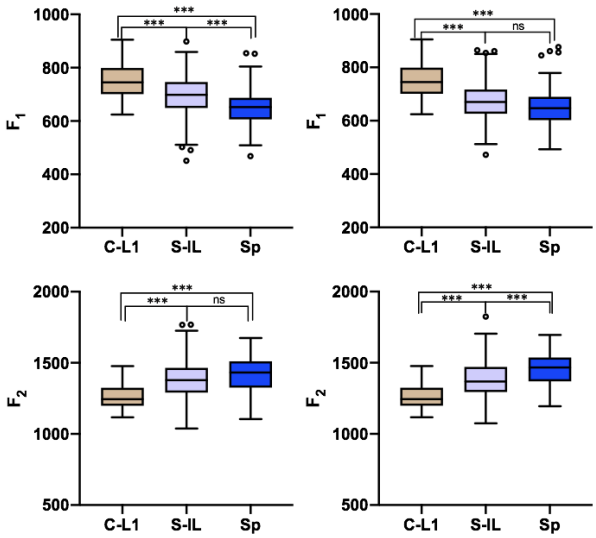


Figure 1: [a] of Chinese (L1), interlanguage (IL) and Spanish (Sp) in the male voice (S: left; U: right).

In the results for females (Figure 2), the differences between L1 Chinese and L1 Spanish are also significant, both in F_1 — p=.024 (S); p=.009 (U) — and in F_2 — p<.001(S and U).

In F_1, the differences between L1 Chinese and the IL are not significant but there are differences between the IL and Spanish. This means that Chinese speakers maintain the [a] very open, typical of Mandarin, in tonic and atonic vowels in the IL (p=.772 and p=.579, respectively), which separates them from Spanish. In contrast, in terms of the F_2, there are significant differences, in tonic and atonic vowels (p<.001), between Chinese and the IL because the [a] of the IL is more fronted and very similar to that of Spanish.

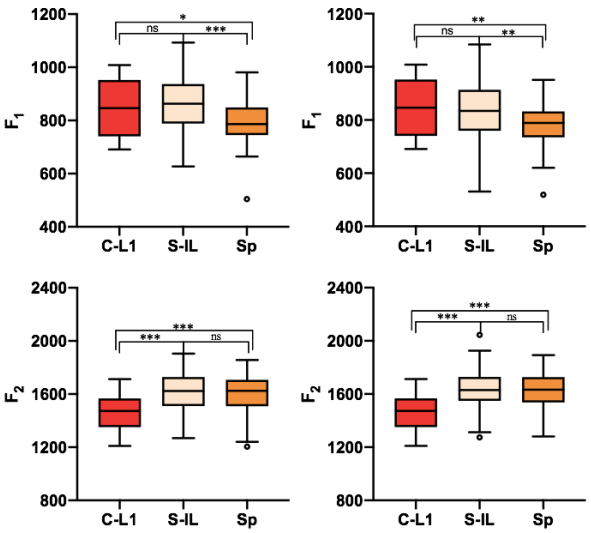


Figure 2: [a] of Chinese (L1), interlanguage (IL) and Spanish (Sp) in the female voice (S: left; U: right).

3.2. Comparison of Groups with Different Levels of Spanish Proficiency and Spanish

3.2.1. Comparison of accuracy

The differences between the Spanish (Sp) male group and the IL groups are significant in F_1 in the tonic vowel (A v. native, p=.027; B v. native, p=.043; C v. native, p<.001) because they articulate the tonic [a] more open (Figure 3). In contrast, in the atonic position, there are no significant differences between the IL and Spanish groups.

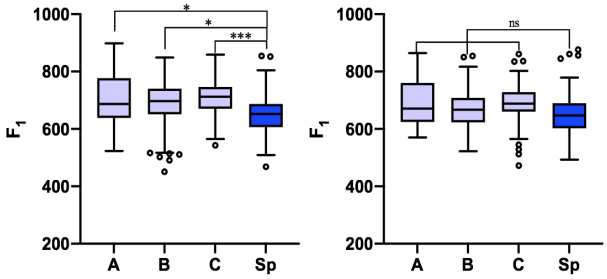


Figure 3: F_1 of [a] of the interlanguage (IL) and Spanish (Sp) in the male voice (S: left; U: right).

In F_2 (Figure 4), there are no significant differences between group C and native speakers, either in tonic (p=.252) or atonic (p=.594) vowels. Proficient-level informants have a similar position to the target language when articulating this vowel.

The differences between group A and the target language are not significant (p=.507) in the results for tonic vowels but they are significant in the atonic vowels (p<.001). As for the comparison between group B and the target language, there are significant differences in tonic (p=.004) and atonic (p<.001) vowels.

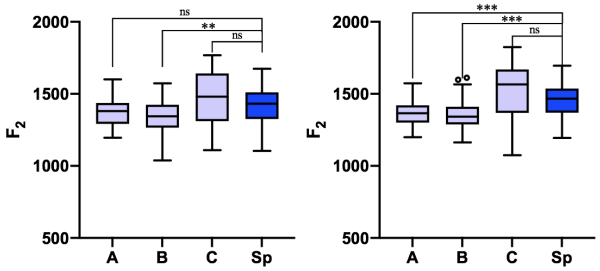


Figure 4: F_2 of [a] of the interlanguage (IL) and Spanish (Sp) in the male voice (S: left; U: right).

For females, in F_1, the differences between Spanish tonic vowels and those of the three IL groups are significant (A v. native, p<.001; B v. native, p=.041; C v. native, p<.001). The F_1 in atonic vowels only shows significant differences between group A and native speakers (p<.001) and no significant differences between group B and C and native speakers (p=.577; p=.219, respectively). According to Figure 5, females, like males, produce the most open IL tonic vowels. As for F_2, there are no differences between native speakers and the three IL groups, either in tonic or atonic vowels.

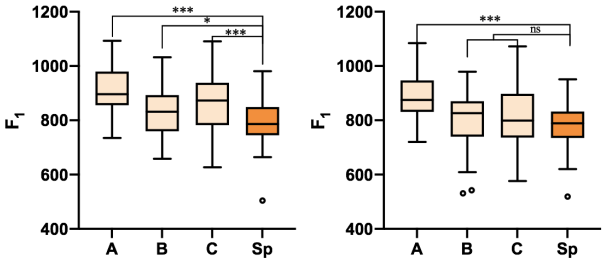


Figure 5: F_1 of [a] of the interlanguage (IL) and Spanish (Sp) in the female voice (S: left; U: right).

3.2.2. Comparison of precision

As can be seen in Tables 1 and 2, the *p*-values of the IL [a] are less than 1 in both tonic and atonic vowels, except for the *p*-value of group C in the atonic vowel, which means that, generally, male Chinese speakers produce this sound in a very stable way. The results are very interesting because, logically, we would expect the *p*-value of the L1 to be less than 1 and the *p*-value of the IL to be greater than 1 and consequently the IL realisations would be less stable.[19]

In Table 1, comparing the *p*-values of the three groups A, B and C, the *p*-value of group A is lower and that of group C is higher, both in the tonic (A 0.692; C 0.947) and atonic (A 0.572; C 1.518) vowels, implying that pronunciation of the vowel [a] by male learners in the initial IL stage is very stable, in other words, close to their L1, Mandarin [8]. As they progress through the levels (from B in tonic vowels and C in atonic vowels), a new category starts to be formed for the L2.

	p1	p2	p
A	0.972	0.712	0.692
B	1.031	0.939	0.968
C	0.819	1.156	0.947

Table 1: *p* values of the tonic vowel [a] in the male voice of the IL.

	p1	p2	p
A	0.768	0.745	0.572
B	0.867	0.894	0.775
C	1.016	1.494	1.518

Table 2: *p* values of the atonic vowel [a] in the male voice of the IL.

In females, similar to males, the *p*-values of the tonic [a] vowel are lower than the standard value 1 (A 0.626; B 0.715; C 0.846), which explains why the realisations of this phoneme by female Chinese speakers are very stable (see Table 3). In the unstressed position, from the independent level onwards, there is already a change of category (B 1.023), which progresses in C (1.373). Findings show that there is a relevant interference of Mandarin in the learning of Spanish in the initial stage and that it has more influence on tonic than atonic vowels (see Table 4).

	p1	p2	p
A	0.751	0.833	0.626
B	0.784	0.913	0.715
C	0.973	0.870	0.846

Table 3: *p* values of the tonic vowel [a] in the female voice of the IL.

	p1	p2	p
A	0.843	0.985	0.831
B	1.037	0.987	1.023
C	1.148	1.196	1.373

Table 4: *p* values of the atonic vowel [a] in the female voice of the IL.

4. DISCUSSION

In the IL, for both genders, we have observed that, in general, they produce a more open [a] vowel than native Spanish speakers, particularly in the tonic position, supporting the conclusions of other authors [10][11][12] and, in contrast, with [13]. These results seem reasonable if we consider that the vowel of L1 Mandarin, more open than Spanish, may influence the IL. As for the tongue position (F_2) of [a] in the IL, females produce it in a similar way to Spanish, more fronted than in Mandarin, in tonic and atonic vowels, and males only in tonic vowels. These results lead to the observation that females acquire the tongue position earlier in the production of the sound. However, the opening is difficult for both genders.

Analysing the learning process, we have observed that in terms of the acquisition of the vowel opening (F_1), there is more difficulty in the tonic realisations of the IL than atonic (Fig. 3, 4 and 5), a result contrary to the statements of [22], for the vowels of the IL, and of [23], for the [e]. According to [22,23], the realisation of tonic vowels is easier than the atonic vowels. In terms of F_2, referring to the position of the tongue and as we have said, women have less difficulty than men.

In group A, both genders have a more stable tonic [a] sound, more typical of Mandarin. In other words, we can confirm that there is a tendency among learners to continue using the L1 category, which essentially fossilises in the tonic vowels of groups B and C, which was also noted by [5], without considering the tonicity factor. In line with [8], we believe that the factor causing these difficulties in learning the sound could be because the sounds of the L1 and the target language are similar. However, this is not the case with the atonic variants, which evolve as they receive more L2 input. We understand that this input encourages the formation of a new category, close to L2 at a proficient level (C) and, in some cases, also at the independent level (B), as stated by [2]. When we distinguish the (a)tonicity factor, we see that this change is more marked in atonic than tonic vowels (Tables 1, 2, 3, 4), a factor which [2, 8] did not consider and which we determine to be relevant.

5. CONCLUSIONS

The results of this research have shown that in the IL, the [a] sound is made more open than in Spanish and similarly or more closed than in Mandarin. As for the tongue position, the sound [a] in the IL is more fronted than in Mandarin and, therefore, more similar to Spanish. The interlanguage is in a centralised position between the L1 and the L2 with which it shares some features.

We have verified that gender, level of Spanish proficiency and (a)tonicity are influential factors in the acquisition of the L2.

We have observed L1 interference in the IL in group A and also the appearance of the fossilisation phenomenon in groups B and C.

The data obtained allow us to better understand the process of Spanish acquisition by Chinese speakers of different levels and to understand the similarities and differences between the low vowel of the IL and that of the target language. These data may constitute an important basis for the preparation of pronunciation teaching activities adapted to the needs of the learners.

6. REFERENCES

[1] Flege, J., Bohn, O. 2021. The revised speech learning model (SLM-r). In Wayland, R. (eds), *Second language speech learning: Theoretical and empirical progress*. Cambridge University Press, 3–83.

[2] Flege, J., Aoyama, K., Bohn, O. 2021. The Revised Speech Learning Model (SLM-r) Applied. In Wayland, R. (eds), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge University Press, 84–118.

[3] Selinker, L. 1972. Interlanguage. *IRAL* 10, 209–232.

[4] Saville-Troike, M., Barto, K. 2016. *Introducing second language acquisition*, 3rd ed. Cambridge University Press.

[5] Han, Z., Odlin, T. 2016. *Studies of fossilization in second language acquisition*, Multilingual Matters.

[6] Cortés-Moreno, M. 2009. Chino y español: un análisis contrastivo. In Sánchez Griñán, J., Melo, M. *Qué saber para enseñar a estudiantes chinos*. Voces del Sur, 175–200.

[7] Norman, J. 1988. *Chinese*. Cambridge University Press.

[8] Flege, J. 1995. Second language speech learning: Theory, findings, and problems. In Strange, W. (eds), *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, 233–277.

[9] Varela Rivera, N. 2014. La pronunciación del chino para hispanohablantes. *Asiadémica: revista universitaria de estudios sobre Asia Oriental* 4, 36–50.

[10] Poch, D., Igarreta, A. 2014. Tender puentes: distancia lingüística y pronunciación. In Ferrús, B., Poch, D. *El español entre dos mundos*, Vervuert Verlagsgesellschaft, 139–158.

[11] Jiménez J., Tang, A. 2018. Producción del sistema vocálico del español por hablantes de chino. *RLA* 56 (1), 13–34.

[12] López, M. 2012. *Estudio de los problemas de pronunciación de los estudiantes sinohablantes que aprenden español*. Universidad de Cantabria-Fundación Comillas.

[13] Igarreta Fernández, A. 2019. *El comportamiento fónico de los sinohablantes ante las vocales del español: efectos de la distancia lingüística sobre el proceso de adquisición*. Universidad Autónoma de Barcelona.

[14] Pérez García, R. 2020. Producción de las vocales tónicas en español por sinohablantes. *SinoELE* 20, 105–120.

[15] Cantero, F. 2015. De la fonética del habla espontánea a la fonología de la complejidad. *Normas* 5 (1), 9-19.

[16] Consejo de Europa. 2012. *Marco común europeo de referencia para las lenguas*. Anaya.

[17] Alfonso Lozano, R. 2010. *El vocalismo del español en el habla espontánea*. [PhD Thesis], Universitat de Barcelona.

[18] Alfonso Lozano, R. 2014. El vocalismo del español en habla espontánea. *Phonica*, 9-10, 3-12.

[19] Boersma P., Weenink, D. 2021. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.39, retrieved 10 February 2021 from http://www.praat.org/.

[20] Wen, B., Deng, D., Shi, F., 2016. *Research on Chinese phonetics acquisition [汉语语音习得研究]*. Nankai University Press [南开大学出版社].

[21] Cao Y., Rius-Escudé, A. 2019. Caracterización acústica de las vocales del español hablado por chinos. *Phonica* 15, 3-22. https://doi.org/10.1344/phonica.2019.15.3-22

[22] Cortés Moreno, M. 2003. Producción y adquisición de la acentuación española en habla espontánea: el caso de los estudiantes taiwaneses. *Estudios de Fonética Experimental,* XII, 61-89.

[23]. Cao Y., Font-Rotchés, D., Rius-Escudé, A., 2023. Front vowels of Spanish, a challenge for Chinese speakers. *Open Lingüistics*, 9(1). https://doi.org/ 10.1515/opli-2022-0230

PHONETIC TRANSCRIPTION OF ATYPICAL SPEECH: SECOND OR A FOREIGN LANGUAGE

Ines Carović

Department of Phonetics
Faculty of Humanities and Social Sciences
University of Zagreb
icarovic@ffzg.hr

ABSTRACT

The paper shows the results of significant research to "create a large set of new speech sound categories and to expand the listener's perceptual system to detect differences between that expanded set of sounds" through training [1]. The research used phonetic transcriptions of 360 first year BA students in Phonetic Studies after 16 and after 60 hours of training. The students had to transcribe a foreign language speaker and results of the study show that there is inadequate time for correct transcription. In addition, a questionnaire given to students show that it was easier to write broad transcription rather than narrow; but when it came to the accuracy of pronunciation or being able to hear and understand, it was more helpful to write narrow transcription. Also, vowels were more difficult to transcribe than consonants. The research showed that the multilingual students were better in phonetic transcription than their monolingual peers.

Keywords: phonetic transcription, transcription training, perception of foreign language

1. INTRODUCTION

Lee [2] analysed articles and research on the importance of listener training for perceptual evaluation of speech. [1, 3, 4, 5, 6, 7, 8] which dealt with the training of speech perception and transcription. They all agree with Stemberger & Bernhardt [1] that it is most important to "create a large set of new speech sound categories and to expand the listener's perceptual system to detect differences between that expanded set of sounds" through training and that both monolingual and multilingual contexts are important for perception training, as they expand perceptual sensitivity to fine distinguishing of phonetic nuances. However, Müller & Papakyritsis
[5] note that we should "not let a 'segmental mindset' become a tyrant dominating our thinking about speech, whether normal or disordered". The paper researched how students listen to, transcribe, and pronounce specific speech and the differences after training in phonetic transcription (PT).Some authors believe that the best transcription comes with video and audio recordings [1, 3] and supplemented with instrumental techniques [1, 9], such as the ultrasound [10, 11] and the spectrogram [12]. Although after the completion of lectures in articulatory and acoustic phonetics, the students might use spectrograms based on acoustic recordings, this was not applied in the present research. Only two students stated in the questionnaire that in the recording of the Chinese speaker they wanted to see whether the spectrograms showed that all voiceless phonemes were indeed voiceless, and they should be voiced in Croatian language. Regarding the time spent on PT training, different authors state different times required for high-quality PT: "100 hours … received by some European speech and language therapist (SLT) students is probably insufficient" [3], there should be more practice [6], "consistent practice" [8], training spread over the years [1, 8], and some mentioned the need of lifelong learning – Refresher Course for Researchers & Clinicians [1, 7]. The most important goal of the course, but also of the study of phonetics is "accuracy, being able to hear and understand, knowledge of symbols for speech sounds" [12]. The authors dealt with the effect of concurrent speech disorders on auditory-perceptual evaluation: effect of pitch, vocal intensity, breathiness, hoarseness (or roughness), articulation on judgements of hypernasality; effect of hypernasality on judgements of articulation, intonation, effect of pitch, vocal intensity on judgements of breathiness, harshness, hoarseness [14]; however, this research omits all labels for pitch, intensity, supralaryngeal placements, suprasegmental features – unless this was extensively noted. Nelson et al [7] mention a broad and narrow PT ("Broad – recording of phonemes which is language dependent, does not contain detail"; "Narrow – detail of pronunciation is recorded"). A question arises: in PT training, how difficult is it for students to perceptually notice the differences between a language which is their mother tongue and the pronunciation of that language as a second or foreign one? The assumption is that after the training it will be easier for them to choose the symbol for perceptual difference although, phonemically, that phoneme is a part of realisation, and this is the precise reason why they will more easily opt for narrow PT

when they are asked to repeat identical pronunciation as well. With phonemes and differences not in their language system or the system of the language they learned, they will find it difficult to agree on the PT symbol with their partner. Main research question for this study was how good are students results of PT after 60 hours of training. The study goal is to compare this training results with results from different areas of research [1, 3, 6, 7, 8].

2. RESEARCH

The research analysed 360 transcribed texts based on 22 audio recordings of foreign student speeches who were learning Croatian and who had a different levels of Croatian language knowledge. Transcriptions were annotated in pairs by 1st year undergraduate students (average age of 19,7) of enrolled in a Phonetic Transcription Coursein the Department of Phonetics, Faculty of Humanities and Social Sciences, in the 4th and 15th week of the semester, , after 16 and 60 hours of training, respectively. The same students received different audio recordings of speakers twice. Mother tongues of speakers were : English, Chinese, Japanese, Vietnamese, Spanish, Arab, Danish, Hungarian, German, Malay, Bulgarian, Russian, Turkish, Catalan, Portuguese, Slovenian and Italian. Some languages had speakers from different areas. For example, some English speakers were from USA, South Africa, and Australia, while Spanish speakers were from Chile, Spain, and Argentina. All speakers read the fable by Aesop, The North Wind and the Sun [15], which was very familiar in Croatian to the students. The students were advised to listen to the audio recording with their headphones on and not to focus on the voice quality, but only on consonants and vowels. The instruction was to agree in pairs on transcription symbols and use the symbols to read the text as the speaker they analysed. At the beginning of the semester, the students received 16 hours of introduction to PT and transcription practice on the example of Croatian standard language (typical speech), as well as the perceptual speech analysis, annotation of transcription symbols of broad PT for vowels and consonants in Czech and Catalan.

At the end of the semester, after 60 training hours in PT, the students transcribed in pairs the new acoustic example. Following both transcriptions, the subjects presented their results in a group setting, received a model PT to check the transcription of the recording they analysed and received feedback of the whole group on whether they agree with the transcription. After the phonetic transcription, the students/subjects received a questionnaire on their impressions and the transcription process.

The analysis of the transcribed texts was made in relation to the transcription model made by expert phoneticians.

3. RESULTS

3.1. Questionnaire

In 97% of cases, respondents felt that PT was easier at the end of semester, but their comments from the questionnaire indicate that at the end of semester they are aware they are not accurate. Respondents are also not confident they are accurate in PT. Students responded that it was easier to write broad transcription rather than narrow, but when it came to the accuracy of pronunciation or being able to hear and understand, it was more helpful to write narrow transcription. Also, vowels were more difficult to transcribe than consonants. Table 1 shows answers to questionnaire questions after each PT. The scale for answers was from 1 to 5. 1 – I do not agree at all; 5 – I completely agree.

Question	Results after the 1st PT	Results after the 2nd PT
1. We agreed easily in pair on the PT symbol.	3.5	3.7
2. I am completely certain regarding the PT.	2.1	3.0
3. It is easy to transcribe the broad PT.	3.6	3.2
4. It is easy to transcribe the narrow PT.	3.3	3.6
5. It is useful to transcribe the broad PT.	3.0	3.5
6. It is useful to transcribe the narrow PT.	4.1	4.5
7. I am certain regarding the consonant PT.	3	3.9
8. I am certain regarding the vowel PT.	2.2	2.4
9. I have studied the mother tongue of the speaker for PT.	2.1	1.9
10. I have listened to the recording more than 100 times.	4.4	4.6

Table 1: Answers to the questionnaire. The scale for answers was from 1 to 5. 1 – I do not agree at all; 5 – I completely agree.

3.2. Transcription accuracy at the beginning and at the end of the semester

Although the subjects did not learn most of speakers' mother tongues, the number of deviations and potential changes in the PT in relation to Croatian standard language was very different with different speakers, so a percentage of accurate PTs in relation

to the overall number of differences in comparison with standard Croatian was calculated, as shown in Chart 1.

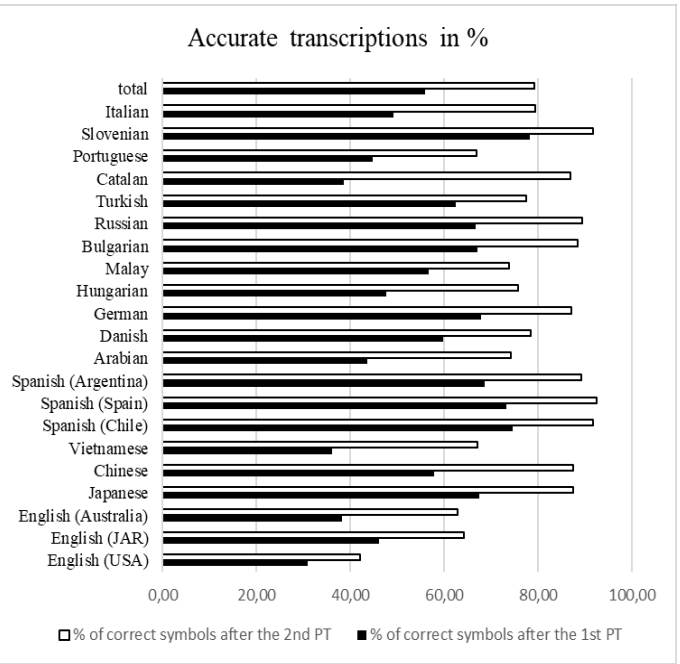


Chart 1: The percentage (%) of accurate transcriptions (that is, changes in relation to the Croatian text).

The subjects made fewer errors in the second PT, and Russian and Spanish were the languages that had fewest differences between the model and the subjects.

It is notable that the percentage of correct PT symbols increased in the second PT after 60 hours of training, to 79,2%, however, it is still not enough for a certain PT. We can conclude that the students would definitely need more practice and training, especially for atypical speeches or speeches they have not been in contact with.

In the second PT, the students who evaluated speakers of Spanish, Slovenian and Russian as mother tongue, performed the best transcription; the reason for this could be that Slovenian and Russian are Slavic languages, so they are closer to the mother tongue of subjects, while Spanish is a language that does not have a large number of deviations in the phonemic system from the Croatian language. Also, although the subjects did not learn it within an institution, Spanish was a language they were most exposed to at an earlier age, through the television. The success of English speakers' PT is surprisingly low. After the analysis, it was established that the level of accuracy was most influenced by the level of learning Croatian as a foreign language, regardless of the fact that these languages were closer to the students as phonemic systems.

The subjects estimated they had more confidence in PT at the end, rather than at the beginning of the semester, but that they cannot be 100% sure for some transcribed elements of the speech. The results of the analysis show that there are no 100% correct answers in PT, but that incorrect answers pose a problem to most listeners as well after the presentation of the results. Figures 1 and 2 show PT examples for speakers of English and Russian as mother tongue in the end of the semester.

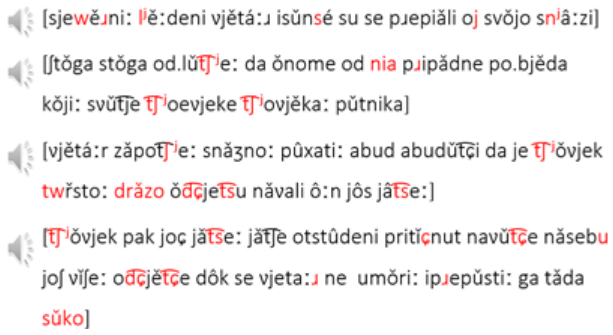


Figure 1: Transcription example of a speaker of English as mother tongue.

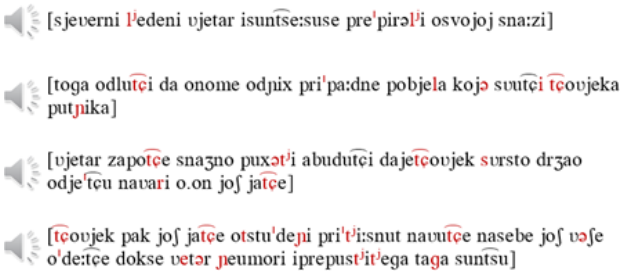


Figure 2: Transcription example of a speaker of Russian as mother tongue.

Some "incorrect" transcriptions were annotated as such as they were annotated in too much detail in the narrow PT. Therefore, it is important to be aware of the agreement regarding the final decision on the PT and the explanation for future collaborators as well. Howard & Heselwood [3] state that "there are no right answers, only more or less likely answers". Agreement of transcriptions: "fail to produce an accurate reflection of the similarities between listeners' perceptions precisely because they rely on scores of symbol agreement, rather than symbol equivalence" [3], and "do not reflect the range of discrepancies that might occur between two transcribers" [4].

3.3. Students who knew the mother tongue of speakers who learned Croatian as a second or foreign language

The fewest deviations from the model PT (phonetic expert) were expected from students who learned the mother tongue of speakers from the recording, and especially the students who studied that language. Given that they were familiar with the phonetic features of the language, and they wrote in the questionnaires that it was easier to "listen" to them as they could transcribe with more certainty what they heard, this was not completely confirmed. However, it was confirmed that the students who knew more than one language had more success in the PT. The most likely reason for this is a more developed phonological awareness and perception for fine differences. Students who did not know the mother tongue wrote that they had to listen many times to the recording to be certain of they heard, but that after many instances of listening they would more easily create a repetitive system, that is, a system of errors in foreign language. It is difficult to state the causes for some errors in PT made by students who knew the mother tongue of the speaker they analysed. Some authors [3, 4] state that in phonetic transcription "predictions of expected speech sounds through knowing the intended target form also have a part to play in producing unreliable transcriptions".

4. CONCLUSION

We can conclude from this research that after the PT training, students become better with perception, transcription symbol annotation and the pronunciation of these symbols. However, 60 hours of training are not enough for obtaining a unified group of reliable listeners. Howard and Heselwood have mentioned 100 hours of training which "is probably insufficient" [3], Knight [6] have noticed there should be more practice and Shaw & Yanushevskaya [8] talk about consistent practice and training spread over the years as well as Stemberger & Bernhardt [8]. Phonetic transcription, articulatory and acoustic phonetics and more importantly - practice and exercises are good way to improve results in PT. Howard & Heselwood [3] talk about the linguistic background: "a native speaker- hearer's perceptual system will constrain perception of speech data", and [3, 16] agree that "listeners tend to force the new material through the perceptual grid of the phonological categories of their own language". Therefore, it is important to expose the subject to a larger number of different stimuli to develop a bigger sensitivity to small differences, and to categories they are familiar with in their own linguistic system. The present research has confirmed this. As Lee [2] concludes in her work, "perceptual evaluation is the gold standard for assessing speech. We need effective listener training programmes for professional training and continuous professional development". The program at the Department of Phonetics at the University of Zagreb does not yet have research on the PT of atypical speech, although graduate students do practice with annotating and analysing atypical speech. PT is a start for perceptual analysis, error systems in a foreign language and phonetic correction of speech, analysis of developmental children's speech, annotation of typical and atypical speech, and it is necessary if someone wants to do phonetics. Lee [2] concludes that "it is necessary to develop a training programme that suits the curriculum and further investigate the impact of different factors on perceptual evaluation for improving training programmes". Results in this research indicate that 60 hours is too little time for making transcription perfect. Research like this is only an indicator of the need for a systematic training in PT, perceptual evaluation of speech, and perceptual judgements.

Acknowledgements. Research infrastructure and research costs were supported by project *Coarticulation in Croatian speech: instrumental investigation (CROCO)* funded by Croatian Science Foundation and by institutional support (Faculty of Humanities and Social Sciences University of Zagreb) for project *Fluency in speech production.* We would like to thank reviewers for taking the time and effort necessary to review the paper. We sincerely appreciate all valuable comments and suggestions, which helped us to improve the quality of the paper.

5. REFERENCES

[1] Stemberger, J. P., Bernhardt, B. M. (2020). Phonetic transcription for speech-language pathology in the 21st century. *Folia Phoniatrica et Logopaedica, 72*(2), 75-83.
[2] Lee, A. 2022. Listener Training for Perceptual Evaluation of Speech. Plenar talk presented at the Speech Research 2022, 8-10 December 2022, Zagreb, Croatia.
[3] Howard, S. J., Heselwood, B. C. (2002). Learning and teaching phonetic transcription for clinical purposes. *Clinical Linguistics & Phonetics, 16*(5), 371-401.
[4] Ball, M., Müller, N., Klopfenstein, M., Rutter, B. (2009). The importance of narrow phonetic transcription for highly unintelligible speech: Some examples. *Logopedics Phoniatrics Vocology, 34*(2), 84-90.
[5] Müller, N., Papakyritsis, I. (2011). Segments, letters and gestures: Thoughts on doing and teaching phonetics

and transcription. *Clinical Linguistics & Phonetics, 25*(11- 12), 949-955.
[6] Knight, R. A. (2010). Sounds for study: Speech and language therapy students' use and perception of exercise podcasts for phonetics. *International Journal of Teaching and Learning in Higher Education, 22*(3), 269-276. https://files.eric.ed.gov/fulltext/EJ938562.pdf
[7] Nelson, T. L., Mok, Z., Ttofari Eecen, K. (2020). Use of transcription when assessing children's speech: Australian speech-language pathologists' practices, challenges, and facilitators. *Folia Phoniatrica et Logopaedica, 72*(2), 131-142.
[8] Shaw, A., Yanushevskaya, I. (2022). Students' views and experiences of the training and use of phonetic transcription in speech and language therapy – The Irish perspective. *Clinical Linguistics & Phonetics, 36*(2-3), 276-291.
[9] Shriberg, L. D., Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical*
[10] *Linguistics & Phonetics, 5*(3), 225-279.
[11] Cleland, J., Lloyd, S., Campbell, L., Crampin, L., Palo, J. P., Sugden, E., Zharkova, N. (2020). The impact of real-time articulatory information on phonetic transcription: Ultrasound-aided transcription in cleft lip and palate speech. *Folia Phoniatrica et Logopaedica, 72*(2), 120-130.
[12] Sugden, E., Cleland, J. (2022). Using ultrasound tongue imaging to support the phonetic transcription of childhood speech sound disorders. *Clinical Linguistics & Phonetics, 36*(12), 1047-1066.
[13] Sullivan, K., Czigler, P. (2002). Maximising the educational affordances of a technology supported learning environment for introductory undergraduate phonetics. *British Journal of Educational Technology, 33*(3), 333-343. https://doi.org/10.1111/1467-8535.00268
[14] Speights Atkins, M., Bailey, D. J., Seals, C. D. (2022). Implementation of an automated grading tool for phonetic transcription training. *Clinical Linguistics & Phonetics*. Advance online publication.
https://doi.org/10.1080/02699206.2022.2048314
[15] Lee, A., Potts, S., Bressmann, T. (2020). Speech- language therapy students' auditory-perceptual judgements of simulated concurrent hypernasality and articulation disorders. *Clinical Linguistics & Phonetics, 34*, 479- 492.
[16] International Phonetic Association. 2009. *Handbook of the International Phonetic Association*. London: Cambridge University Press.
[17] Laver, J. (1991). *The gift of speech: Papers in the analysis of speech and voice*. Edinburgh University Press.

INVESTIGATING PREDICTORS OF FOREIGN ACCENTEDNESS IN L3 ACQUISITION

Magdalena Wrembel, Kamil Kaźmierski, Katarzyna Dziubalska-Kołaczyk, Jarosław Weckwerth, Zuzanna Cal

Adam Mickiewicz University, Poznań
magdala@amu.edu.pl, kamil.kaźmierski@amu.edu.pl, dkasia@amu.edu.pl, wjarek@amu.edu.pl, zuzanna.cal@amu.edu.pl

ABSTRACT

Research into predictors of foreign accentedness investigated from the perspective of third language acquisition (TLA) has been rather scarce to date. The study aims to explore to what extent a holistic assessment of global accent in the third language is correlated with the learners' general proficiency level, oral fluency, fine-grained phonetic performance and the frequency of L3 use. Accentedness and comprehensibility ratings were performed on samples of L3 read speech. The results demonstrated that the two rating parameters are significantly correlated, yet the impact of particular variables varied, with the rater's nativeness status being the most important predictor.

Keywords: foreign accentedness, comprehensibility, predictors, raters, L3 acquisition.

1. INTRODUCTION

Ratings of perceived global foreign accent have been widely applied in second language acquisition (SLA) research [1, 2, 3]; however, this phenomenon has been less frequently explored from the multilingual acquisition perspective [4, 5]. Further, the most recent L3 studies focus mostly on heritage speakers [6, 7]. Previous research on factors contributing to a perception of accentedness has identified the amount of L1/L2 use, the age of onset or age of arrival in an L2-speaking country and the presence of non-native segmental features in rated samples as the most significant predictors (see [8] for a detailed overview).

The present study forms one part of a large project investigating L3 development longitudinally. It aimed to explore to what extent a holistic assessment of global accent in the third language is correlated with the learners' general proficiency level, oral fluency and fine-grained phonetic performance. The theoretical framework embraced in this paper is the Natural Growth Theory of Acquisition (NGTA) [9]. NGTA's main assumptions include a gradual dynamic emergence of foreign language phonology, shaped by the input from L1 and other L2/L3/Ln, influenced by typology, universal preferability generalizations and context of use. NGTA relies on principled explanations as well as inductive, data-driven accounts and proposes a hierarchy of linguistic and extralinguistic factors to account for the process of multilingual acquisition of speech in a given acquisition situation (see [6] for a detailed discussion). The results of the current study are used to further inform the hierarchy of variables as stemming from the NGTA framework.

2. STUDY DESIGN

The participants were 24 speakers of L1 Polish–L2 English–L3 Norwegian, aged 21, after 8 weeks of intense initial exposure to the L3 in a formal academic setting. They performed a Norwegian placement test as a measure of proficiency and completed the Language History Questionnaire LHQ [10]. The language material used in the rating study involved text reading of a Norwegian version of *The North Wind and the Sun*. Oral reading fluency was expressed as the number of words per minute (wpm). Fine-grained phonetic performance was assessed based on the reading of a word list in L3 including /p, t, k/ stop tokens in stressed onset positions (controlled for vocalic context), where Norwegian, unlike Polish, displays long-lag VOT durations [11, 12, 13]. We also calculated the amount of L3 use based on the self-declared detailed responses to the LHQ.

In the rating study, approximately 20-second-long samples were extracted from the recordings of the read text and normalised for loudness. Twenty-three raters, half of whom were Norwegian native speakers and the remaining half were non-native speakers highly proficient in Norwegian, rated the samples for the degree of foreign accentedness and comprehensibility on a 9-point scale, using a Qualtrics online survey. The raters had moderate to considerable amount of previous experience with foreign accented speech in Norwegian. The survey included 30 randomised samples, featuring the 24 L3 learners and 6 Norwegian controls.

2.1. Measures

- L3 Proficiency – Norwegian placement test
- Amount/frequency of L3 use – a composite score based on self-declared answers in LHQ

- Oral reading fluency: number of words per minute (wpm)
- Fine-grained phonetic performance: VOT durations in /p, t, k/ in word list reading in L3
- Rating parameters: degree of foreign accentedness and comprehensibility on a 9-point scale

2.2 Research questions

RQ1: Do the rating parameters (i.e., accentedness and comprehensibility) correlate?
RQ2: Does perceived global accent correlate with the learners' proficiency level, their oral fluency and fine-grained phonetic performance in L3 Norwegian?
RQ3: Does perceived comprehensibility correlate with the learners' proficiency level, their oral fluency and fine-grained phonetic performance in L3 Norwegian?
RQ4: Do rater-related variables influence the ratings?

3. RESULTS

Statistical analysis was run in R [14]. For significance testing, mixed-effects ordinal regression modeling was performed using the 'clmm' function from the *ordinal* R package [15]. All models included by-speaker and by-rater random intercepts.

Descriptive statistics including mean values and SD for Accentedness and Comprehensibility ratings, fine-grained phonetic performance (i.e. VOT durations for /p, t, k/ in L3 Norwegian, oral reading fluency (words per minute) and L3-frequency of use is presented in Table 1.

Parameters	Experimental group *M (SD)*	Control group *M (SD)*
Accentedness (1–9)	6.72 (1.8)	1.5 (1.5)
Comprehensibility (1–9)	6.03 (2.3)	7.8 (2.7)
Oral fluency (wpm)	0.05 (0.01)	–
VOT /p/ (ms)	44 (14)	–
VOT /t/ (ms)	62 (15)	–
VOT /k/ (ms)	74 (18)	–
Norwegian use (hrs/week)	4.2 (4.6)	–

Table 1. Descriptive statistics.

The control samples received lower accentedness ratings both from Norwegian native ($b = -14.24$) and non-native ($b = -6.47$) raters. The control samples also received higher comprehensibility ratings from both Norwegian ($b = 8.06$) and non-native ($b = 0.9$) raters. All four effects were statistically significant, with effect sizes smaller for non-native raters than for Norwegian native raters in both cases.

In turn, the test samples of L3 Norwegian speakers were judged as more accented by non-native raters than by Norwegian native raters (see Fig. 1). Similarly, the L3 Norwegian speakers were rated as fairly comprehensible; a little less comprehensible to native Norwegian raters than to non-native raters (Fig. 2).

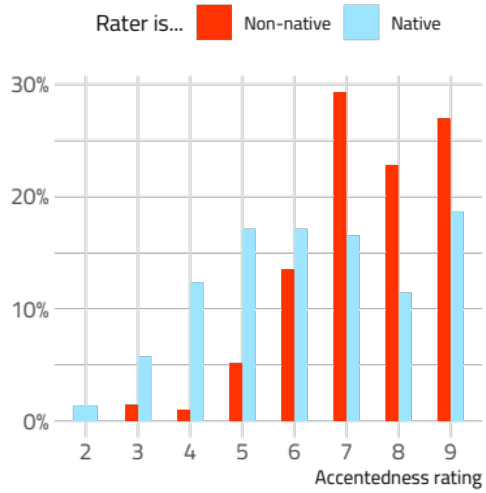


Figure 1: Effects of rater status for Accentedness (1 – no foreign accent, 9 – strongly accented).

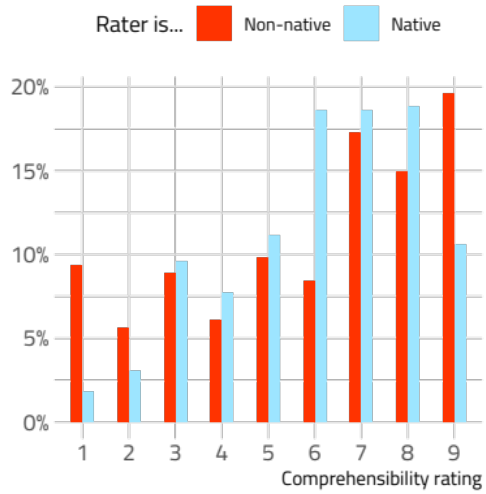


Figure 2: Effects of rater status for Comprehensibility (1 – not comprehensible, 9 – fully comprehensible).

For the first rating parameter, the Pearson's correlation between Accentedness and Comprehensibility ($r = -0.77$) was found to be strong and significant demonstrating that the stronger the accent, the lower the comprehensibility of the speech sample (see Fig. 3).

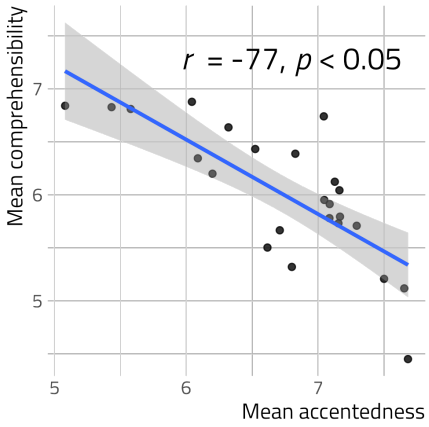


Figure 3: Correlation between Accentedness and Comprehensibility.

The Pearson's correlation between Accentedness and L3 Norwegian Proficiency was significant and moderate in strength (r = −0.59); similarly to that between Accentedness and oral fluency (r = −0.59), thus showing that the higher the speech rate, the less accented it is perceived to be. No correlations were found between perceived foreign accent and VOT measures.

For the second rating parameter, the Pearson's correlation between Comprehensibility and L3 Norwegian Proficiency was found to be significant and moderate (r = 0.41), demonstrating that more proficient learners were perceived as more comprehensible by the raters. As far as Comprehensibility and oral fluency is concerned, the correlation between these two measures proved to be significant and moderate (r = 0.51), showing that the higher the speech rate, the higher the comprehensibility rating. No correlations were found between Comprehensibility and VOT measures.

The analysis of rater variables pointed to statistically significant differences between native vs. non-native raters for Accentedness but not Comprehensibility. A mixed-effects ordinal logistic regression model was fitted with Accentedness as a function of nativeness of rater, with Norwegian proficiency as control, and by-speaker and by-rater random intercepts. The results showed a statistically significant effect of the rater's nativeness status as well as Norwegian proficiency of the speakers for the performed ratings of Accentedness. It was found that non-native raters proficient in Norwegian tended to give higher (i.e., harsher) accentedness ratings than native Norwegian raters who tended to be more lenient (b = −1.8, p = 0.03). Higher L3 Norwegian proficiency of the speakers was also linked to lower accentedness ratings (b = −0.14, p < 0.001). However, no statistically significant effect of rater nativeness was found in the case of the second parameter, i.e., Comprehensibility.

The Interrater reliability was calculated using Cronbach's alpha; for Accentedness, α = 0.89; and for Comprehensibility, α = 0.87.

The relative importance of predictors was compared by means of random forests: the forests were fitted and variable importance was calculated with the *party* R package [16]. For both Accentedness and Comprehensibility, the influence of nativeness status of the rater (variable importance: 0.026 for Accentedness and 0.003 for Comprehensibility) by far outweighs the influence of all other predictors, i.e. L3 Norwegian proficiency (0 and 0), overall oral fluency (0 and 0), amount of Norwegian use (−0.002 and −0.001), and VOT values of /p (0 and 0), t (0 and 0), k (0 and 0)/. Figures 4 and 5 show representative Conditional Inference Trees.

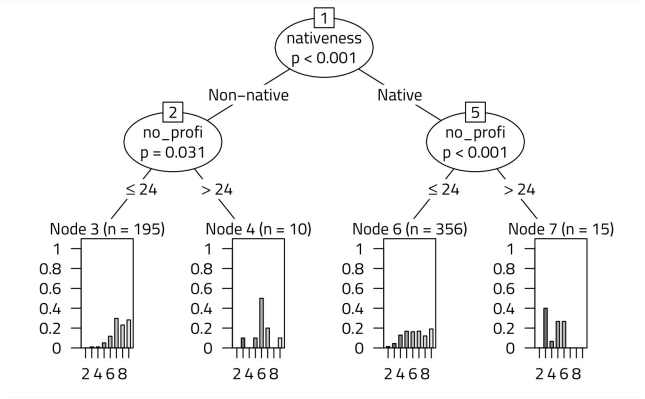


Figure 4: Conditional importance of predictors for Accentedness.

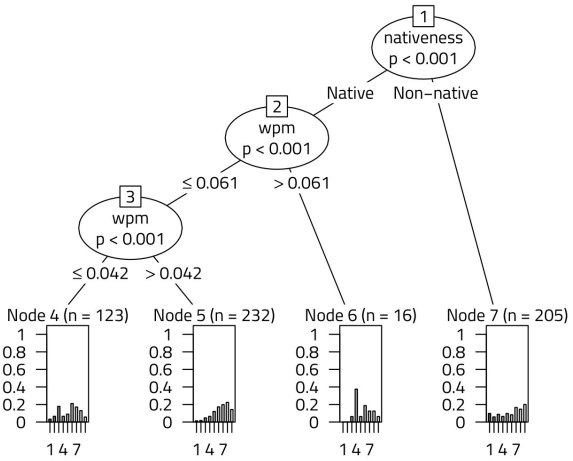


Figure 5: Conditional importance of predictors for Comprehensibility.

4. DISCUSSION

RQ1 asked whether rating parameters (i.e., accentedness and comprehensibility) correlate, and this was corroborated by our results. As predicted, a heavier accent was perceived to be less comprehensible. Moderate correlation patterns were observed for both parameters, unlike in some of previous L2 rating studies where even more accented samples were perceived as still comprehensible [1, 2]. RQ2 explored the relationship between perceived global accent and the learners' proficiency level, their oral fluency as well as their fine-grained phonetic performance in L3 Norwegian. Our predictions were partially confirmed as the holistic perceptual assessment correlated with how proficient the participants were in their L3 and with how fluent their reading production was. However, we did not find any correlations for the VOT durations even though it could have been expected that these fine-grained measures of phonetic performance should be an important predictor of global perceived foreign accentedness [5].

In turn, RQ3 enquired if perceived comprehensibility correlated with the variables under investigation. Similar patterns emerged as in RQ2 since comprehensibility ratings tended to correlate with the level of L3 proficiency as well as oral fluency measures, which was not the case for VOT performance of L3 speakers. Consequently, our predictions that all the investigated performance measures should align with the global perception of comprehensibility were partially confirmed.

Finally, RQ4 asked if rater variables influenced the ratings of L3 speech. The results indicate that rater nativeness status (native vs. non-native) proved to be the most significant predictor, as pointed out by the random forest analysis. Interestingly, the non-native raters proficient in Norwegian appear harsher in their ratings of both the degree of foreign accentedness as well as comprehensibility, which corroborates also some previous research both from L2 and L3 accent rating studies, e.g. [5, 8], and L2 accent attitude studies, e.g. [17]. These systematic differences in rating patterns resulting in the more lenient performance of the Norwegian native raters could be partially ascribed to the predicted higher degree of familiarity with accented speech as well as a higher predicted level of metalinguistic awareness on the part of non-native raters.

The study findings can also be used to further elaborate the hierarchy of variables as proposed by the NGTA framework [9]. As far as the linguistic factors are concerned, the raters' status (native vs. non-native) was the most significant predictor of L3 speech ratings pointing to the influence of language-specific systems; whereas fine-grained phonetic performance (i.e., VOT durations) was not. The raters' status also implied the potential significance of extralinguistic factors mentioned above, i.e., familiarity with accented speech and metalinguistic awareness, both expected to be greater in non-native speakers, but not measured explicitly in this study. As to further extralinguistic factors, reading fluency and L3 proficiency predicted accentedness and comprehensibility, but the amount and frequency of L3 use, and speaker individual differences were found to be of lesser importance. Recapitulating, based on the performed analysis of L3 speech ratings, the NGTA hierarchy of factors in the acquisition situation under investigation here is depicted in Figure 6. The most important factor – the influence of the raters' language-specific systems – was linguistic, but two extralinguistic factors were also implicated, i.e. metalinguistic awareness and familiarity with foreign accent. Oral fluency and proficiency turned out to be the most important measured extralinguistic factors, which were classified as such as they both are consequence of language use by speakers as a result of their linguistic competence. Variables of lesser importance included the frequency of input and use, and individual factors (see Fig. 6).

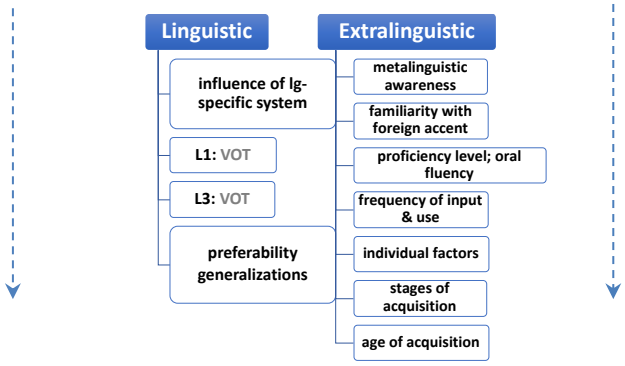


Fig. 6: Hierarchy of linguistic and extralinguistic variables in NGTA.

5. CONCLUSION

This study aimed to contribute to a growing body of research on the acquisition of L3 phonetics and phonology [18]. In particular, we investigated predictors of foreign accentedness from a lesser studied perspective of third language acquisition (TLA). Moreover, we attempted to further verify the hierarchy of variables as proposed by NGTA [9].

In future studies, we intend to explore how foreign accentedness and comprehensibility ratings are correlated between the same speakers' samples in their first and second foreign languages (i.e., L2 and L3) to tap into more intricate interactions between all the languages in multilingual learners' repertoires.

6. ACKNOWLEDGEMENT

This research is supported by Norway funds/ NCN project grant GRIEG-1 (UMO-2019/34/H/HS2/00495) ADIM "Across-domain investigations in multilingualism: Modeling L3 acquisition in diverse settings".

7. REFERENCES

[1] Flege, J. E. 1988. Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America* 84. 70–79.

[2] Piske, T., MacKay, I. R. A., Flege, J. E. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics* 29. 191–215.

[3] Wrembel, M., Marecka, M., Szewczyk, J., Otwinowska, A. 2019. The predictors of foreign-accentedness in the home language of Polish-English bilingual children. *Bilingualism: Language and Cognition* 22(2). 383–400.

[4] Wrembel, M. 2012. 'Foreign accentedness in Third Language Acquisition; the case of L3 English', in J. Cabrelli Amaro, S. Flynn and J. Rothman (eds) *Third Language Acquisition in Adulthood*. Amsterdam: John Benjamins, 281–309.

[5] Wrembel, M. 2015. *In search of a new perspective: Cross-linguistic influence in the acquisition of third language phonology*. Poznań: Wydawnictwo Naukowe UAM.

[6] Lloyd-Smith A., Gyllstad H., Kupisch T. 2017. Transfer into L3 English: Global accent in German-dominant heritage speakers of Turkish. *Linguistic Approaches to Bilingualism* 7: 131–162.

[7] Lloyd-Smith, A. 2021. Perceived foreign accent in L3 English. The effects of heritage language use. *International Journal of Multilingualism.*

[8] Kiełkiewicz-Janowiak, A., Wrembel, M. 2020. L1 Foreign Accentedness in Polish Migrants in the UK: An Overview of Linguistic and Social Dimensions In: Wrembel, M., Kiełkiewicz-Janowiak, A., Gąsiorowski, P. (eds.) *Approaches to the Study of Sound Structure and Speech*, New York: Routledge, 185–197.

[9] Dziubalska-Kołaczyk, K., Wrembel, M. 2022. Natural Growth Theory of Acquisition (NGTA): Evidence from (mor)phonotactics. In Sardegna, V., Jarosz, A. (eds): *Theoretical and Practical Perspectives on English Pronunciation Teaching and Research.* Springer.

[10] Li, P., Sepanski, S., Zhao, X. 2006. Language history questionnaires: A web-based interface for bilingual research. *Behavior Research Methods* 38(2): 202–210.

[11] Keating, P., Mikoś, M., Ganong III, W. 1981. A cross-language study of range of VOT in the perception of stop consonant voicing. J. Acoust Soc. Am. 70. 1260–1271.

[12] Ringen, C., & Dommelen, W.A. 2013. Quantity and laryngeal contrasts in Norwegian. *J. Phonetics*, 41, 479–490.

[13] Czarnecki, P. 2016. *The Phonology of Quantity in Icelandic and Norwegian*. Wydział Neofilologii UAM w Poznaniu.

[14] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[15] Christensen, R. H. B. 2022. ordinal – Regression Models for Ordinal Data. R package version 2022.11-16. https://CRAN.R-project.org/package=ordinal.

[16] Strobl, C., Boulesteix, A-L., Kneib, T., Augustin, T., Zeileis, A. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9(307).

[17] Hendriks, B., van Meurs, F., Usmany, N. 2021. The effects of lecturers' non-native accent strength in English on intelligibility and attitudinal evaluations by native and non-native English students. *Language Teaching Research* Online First. https://doi.org/10.1177/1362168820983145.

[18] Wrembel, M., Cabrelli Amaro, J. (eds.). 2018. *Advances in the Investigation of L3 Phonological Acquisition*. London: Routledge.

PERCEPTION OF ENGLISH LEXICAL STRESS IN DIFFERENT INTONATIONS BY MANDARIN LISTENERS

Tzu-Hsuan Yang[1], Annie Tremblay[2]

University of Kansas[1], University of Texas at El Paso[2]
thyang@ku.edu[1], actremblay@utep.edu[2]

ABSTRACT

The present study seeks to elucidate the nature of first language transfer effects on the perception of English lexical stress in different intonational contexts. We examine whether Mandarin learners of English assume a one-to-one relationship between pitch and lexical stress in English by testing their stress perception in four different intonations: H*L–L%, L*H–H%, H*H–H%, and L*L–L%, where F0 cues to stress were realized differently in each intonation. The results of a stress identification task showed that Mandarin listeners falsely associated higher F0 with stress and lower F0 with the absence of stress, and their performance did not improve as they became more proficient in English. The findings provide corroborating evidence for the Cue-Weighting Transfer Hypothesis by showing that the use of a suprasegmental cue can transfer from one phonological category (tone) to another (stress).

Keywords: Second language acquisition, cue-weighting transfer, lexical stress, lexical tone

1. INTRODUCTION

Languages with lexical stress differ in how acoustic cues signal stress. One approach that seeks to explain how differences in cue weighting affect second-language (L2) speech perception and processing is the Cue-Weighting Transfer Hypothesis [1]–[5]. This hypothesis predicts that listeners can transfer the weighting of an acoustic cue (e.g., F0) from one phonological phenomenon in the first (i.e., native) language (L1; e.g., lexical tones) to another in the L2 (e.g., lexical stress). To illustrate, in English, the stressed syllable is produced with a full vowel, longer duration, and greater intensity [6]–[10]. Crucially, pitch (F0) can only serve as a *surface* cue to lexical stress: Different pitch accents (e.g., H*, L*) can be aligned with the stressed syllable, and the pitch contour that is realized over the word is created by the combination of a pitch accent, a phrase accent (e.g., H– or L–, and a boundary tone (e.g., H% or L%) [11]–[13]. Unlike English, Mandarin is a tonal language in which lexical tones, conveyed primarily by F0 cues, serve to distinguish word meanings [14], [15]. Given the great functional load of pitch for signaling lexical contrasts in Mandarin, Mandarin listeners are expected to rely on F0 cues when distinguishing lexical stress contrasts in English as well.

This prediction was largely confirmed in previous studies. For example, Qin et al. [3] compared learners of English whose L1 was Standard Mandarin (SM), which has lexical stress contrasts signaled primarily by duration cues, or Taiwan Mandarin (TM), which does not have stress contrasts. The authors found that both groups of Mandarin listeners could use F0 to perceive lexical stress in English, but only SM listeners could use duration cues. This provides clear evidence that listeners can transfer the use of acoustic cues from one phonological phenomenon (lexical tones) to another (lexical stress) in speech perception.

Crucially, though, these studies all examined the perception or processing of stress in the canonical intonational context—that is, when the stressed syllable had a H* pitch accent and thus was realized with higher pitch relative to the unstressed syllable. To investigate whether and how L2 learners' processing of English lexical stress is affected by different sentence intonations, Ou [16] tested L1 Taiwan Mandarin listeners' perception of English stress in two intonations: a falling intonational contour where the stressed syllable is realized with H* (H*L–L%) and a rising intonational contour where the stressed syllable is realized with L* (L*H–H%). The results of the discrimination tasks revealed that overall, English native listeners showed significantly greater sensitivity to stress than L2 learners. Crucially, the difference between native and non-native listeners was greater in the rising contour than in the falling contour, indicating that L2 learners had more difficulty perceiving stress in this context.

Adopting a similar design as [16], Choi [17] investigated the perception of English lexical stress by L1 Cantonese L2 learners of English in the falling contour and the rising contour. The results of the discrimination task showed that when lexical stress contrasts were not cued by vowel quality changes, Cantonese listeners outperformed English listeners in the falling intonational context. Nonetheless, L2 learners' discrimination ability significantly worsened in the rising contour, and an advantage over native listeners was no longer found.

However, one important limitation arises in the interpretation of [16]–[17] due to the authors' adoption of discrimination tasks: The use of a discrimination task may have encouraged listeners to rely on the acoustic realization of the words at a surface level, leaving open the question of whether listeners can correctly assign nouns and verbs to different categories despite variability in their intonational realizations. In addition, [16]–[17] did not investigate whether L2 learners can rely on non-tonal cues to English lexical stress in contexts where the intonation is relatively flat.

Adopting an identification task, the current study seeks to extend [16]–[17] by addressing the research questions of whether Mandarin listeners assume a one-to-one relationship between pitch and lexical stress in English, and whether they can rely on non-tonal (suprasegmental) cues when the pitch cues heard over the target word are ineffective.

2. METHODS

2.1. Participants

Participants were 30 Mandarin-speaking L2 learners of English from Taiwan (14 female, 15 male, 1 preferred not to answer, aged 18–31, M = 25.3) and 29 native speakers of American English (20 female, 9 male, aged 18–29, M = 20.6). The English proficiency of the L2 listeners was assessed using the LexTALE test [18]. Mandarin listeners' averaged LexTALE score was 65 out of 100 (SD = 14), which roughly amounts to an intermediate level of proficiency.

2.2. Materials

A two-alternative forced-choice identification task was conducted to examine Taiwan Mandarin listeners' perception of lexical stress. The target words included ten disyllabic minimal pairs with stress on either the first (noun) or the second (verb) syllable (Table 1). As the focus of the present study was to examine how Mandarin listeners utilize suprasegmental cues to process lexical stress, stimuli were selected so that the unstressed syllable of these disyllabic words would not undergo vowel reduction.

Trochaic (noun)	Iambic (verb)
PERmit	perMIT
IMpact	imPACT
IMport	imPORT
DEcrease	deCREASE
INcrease	inCREASE
INsult	inSULT
DIScharge	disCHARGE

Table 1: Stress minimal pairs used in the stress identification task.

The target words were elicited in four intonations: H*L–L% (i.e., a H pitch accent followed by a L phrase accent then by a L boundary tone), L*H–H%, H*H–H%, and L*L–L% (or HLL, LHH, HHH, and LLL hereafter). The auditory stimuli were produced and then extracted from two carrier sentences: *Mary said _____ before* or *Mary say _____ before?*, with the former used to elicit the HLL and LLL contours and the latter to elicit LHH and HHH. Note that the target words were elicited in non-sentence-final position to decrease the amount of word- and sentence-final lengthening. All four contours were naturally occurring intonations in English [13], [19]. The pitch contours of the target words in each intonation are shown in Figure 1.

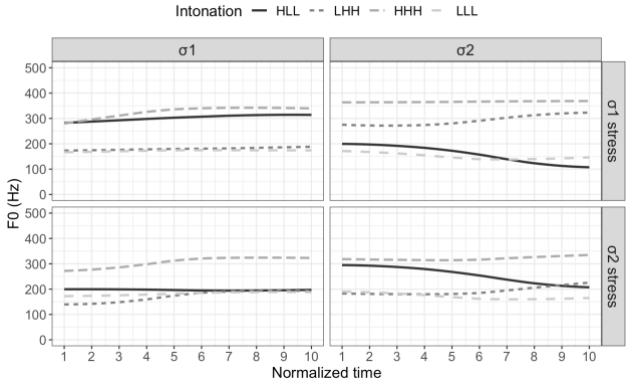


Figure 1: Pitch contours of the stimuli (sample word: *insult*).

These intonations were chosen so that the F0 cues to stress would be realized differently in each intonation, and to allow for an examination of the use of non-tonal prosodic cues to English lexical stress (duration and intensity cues, which were not teased apart in this study) by systematically manipulating the pitch contour. In the HLL condition, a H* was aligned with the stressed syllable for both trochaic and iambic words. The stressed syllable was thus always realized with higher F0, which was consistent with the canonical association expected from Mandarin L2 learners of English. However, in the LHH condition, a L* was aligned with the stressed syllable, and H phrase and boundary tones surfaced on the second syllable for both stress patterns. In other words, the second syllable was higher in pitch for both stress patterns, possibly preventing Mandarin listeners from resorting to their canonical association to perceive lexical stress. On the other hand, the HHH and LLL conditions contained relatively neutral F0 cues, as the pitch contours that were heard over the target word were relatively flat. Crucially, in these two conditions, words with stress on the first or second syllable were realized with very similar pitch contours, and therefore F0 cues were ineffective for

participants to distinguish between the stress minimal pairs. Table 2 summarizes the acoustic information of the stimuli in each intonation condition.

The task included a total of 160 stimuli (10 pairs × 2 stress positions × 4 intonations × 2 talkers). All stimuli were randomized.

Intonation	Stressed syllable	Duration ratio of σ1 to σ2	Mean F0 ratio of σ1 to σ2
HLL	First	0.47	1.62
	Second	0.35	0.78
LHH	First	0.49	0.63
	Second	0.35	0.87
HHH	First	0.53	0.94
	Second	0.35	0.94
LLL	First	0.5	1.08
	Second	0.36	1.07

Table 2: Acoustic measurements of the stimuli. Values were averaged across all tokens in each intonation condition.

2.3. Procedure

The experiment was conducted online via Gorilla Experiment Builder [20]. Participants were told that they would hear English words that can either be a noun or a verb depending on which syllable is stressed. They were instructed to press the left arrow key if they perceived the stress to be on the first syllable (if they heard a noun) and the right arrow key if on the second syllable (if they heard a verb). The experiment took approximately 25 minutes.

3. RESULTS

The raw accuracy data were converted into d-prime scores [21]. Figure 2 presents L1 and L2 listeners' d-prime scores in each intonation. Mixed-effects linear regression models were fitted to participants' d-prime scores in R [22]. Table 4 summarizes the results of the model with the best fit. We also correlated L2 learners' z-transformed LexTALE scores with their d-prime scores to determine whether L2 listeners' sensitivity to lexical stress was modulated by their proficiency in English.

The results showed that both Mandarin and English listeners were significantly more sensitive to stress in the HLL intonation than in the other intonations (demonstrated by simple effects of Intonation with negative coefficients, which were also found in the releveled model with English as the baseline of L1). The model also revealed a significant simple effect of L1. The negative coefficient indicates that in the HLL intonation, Mandarin listeners' sensitivity to lexical stress was in fact higher than that of native English listeners, suggesting that Mandarin listeners were better able to use the higher F0 to facilitate their perception of stress in the falling intonation. Additionally, we found a significant interaction between Intonation (LHH) and L1, suggesting that the difference between HLL and LHH was larger for Mandarin listeners than for English listeners. Collectively, these findings indicate that Mandarin listeners made a stronger association between higher F0 and lexical stress than English listeners did.

To examine listeners' performance when F0 cues were not as effective (relatively flat), we releveled the Intonation variable with HHH as the baseline. The releveled model reported a non-significant intercept (β = 0.217, SE = 0.151, t = 1.437, p > .1). As a d-prime score of 0 signals no sensitivity, this finding showed that Mandarin listeners' performance in the HHH condition was at chance level. There was also no significant simple effect of Intonation (LLL) (β = 0.032, SE = 0.185, t = <|1|, p >.1), suggesting that L2 listeners' performance was similar across HHH and LLL conditions. Taken together, these results imply that Mandarin listeners were not able to utilize non-tonal cues (i.e., duration and intensity) to distinguish words with stress on the first or second syllable when F0 cues were not informative. The lack of simple effect of L1 (β = –0.174, SE = 0.216, t = <|1|, p > .1) and Intonation (LLL)-by-L1 interaction (β = 0.301, SE = 0.264, t = 1.14, p > .1) indicated that English listeners patterned similarly to Mandarin listeners.

Finally, Figure 3 shows the relationship between learners' d-prime scores and their L2 proficiency (measured by LexTALE). The results demonstrate that when Mandarin listeners encountered difficulty in identifying stress, that is, when their association between higher F0 and lexical stress was violated (i.e., LHH, HHH, and LLL), their performance did not improve with increasing proficiency in English.

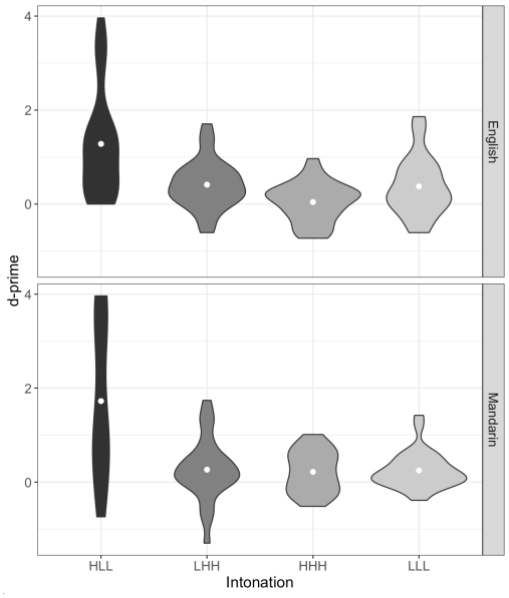


Figure 2: Participants' d-prime scores in each intonation (the white dots represent the mean).

	β	SE	t	p
(Intercept)	1.724	0.151	11.394	< .001
Intonation (LHH)	–1.459	0.185	–7.888	< .001
Intonation (HHH)	–1.507	0.185	–8.147	< .001
Intonation (LLL)	–1.475	0.185	–7.976	< .001
L1 (English)	0.444	0.216	–2.054	< .05
Int. (LHH) × L1 (Eng)	0.592	0.264	2.243	< .05
Int. (HHH) × L1 (Eng)	0.269	0.264	1.02	0.309
Int. (LLL) × L1 (Eng)	0.569	0.264	2.16	< .05

Table 3: Summary of fixed-effects for the model on participants' d-prime scores. Formula: dprime ~ intonation*language + (1|Participant), with HLL being the baseline for Intonation and Mandarin being the baseline for L1.

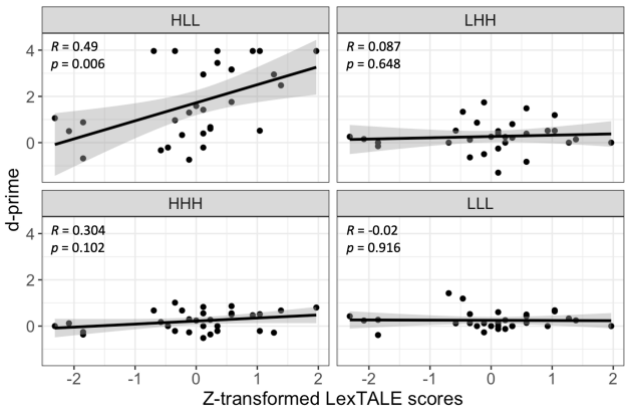


Figure 3: Relationship between L2 learners' d-prime scores and their English proficiency.

4. DISCUSSION

The results of the stress identification task provided clear evidence for the hypothesis that Mandarin listeners assume a one-to-one relationship between pitch and lexical stress in English. Mandarin listeners showed the greatest sensitivity to stress in the HLL condition, where the stressed syllable was always realized with a higher F0. In contrast, their performance worsened in the LHH contour where this association was violated, regardless of their L2 proficiency. These results were consistent with the findings of previous studies [16], [17] and support the proposal of the Cue-Weighting Transfer Hypothesis [1]–[5] that L2 learners can transfer the use of prosodic information (here, F0) from a different phonological category in the L1 (here, lexical tones) to lexical stress in the L2, even if their L1 does not contain lexical stress contrasts.

An interesting observation arises from the results: In the falling contour, L2 learners outperformed native listeners of English. The low accuracy of L1 listeners is attributed to the lack of vowel quality cues in the stimuli. Given that English listeners rely on vowel quality among all acoustic cues to stress [1], [5], it might have been difficult for them to distinguish stress contrasts using suprasegmental cues only [2]. Furthermore, in recent years there has been emerging evidence showing that in the absence of vowel quality cues, native English listeners show less sensitivity to stress compared to L2 learners of English whose L1 uses pitch to signal lexical contrasts [2], [17]. Mandarin listeners' advantage over native English listeners has been ascribed to their positive transfer of F0 processing skills: As F0 is the primary acoustic correlate of lexical tones, it carries greater functional weight in Mandarin, and thus Mandarin listeners are more sensitive to it than English listeners.

To examine how Mandarin listeners weighted non-tonal cues when F0 was an ineffective cue to stress, we examined L2 learners' performance in the HHH and LLL conditions. In these intonations, the words with stress on the first and second syllables were realized with very similar contours, and therefore participants had to rely on duration and intensity cues to distinguish stress contrasts. The results revealed that L2 learners had almost zero sensitivity to stress placement in these conditions, suggesting that they were not able to use the duration and intensity cues to identify the position of stress, even as they became increasingly proficient in English. These results corroborated [3], who reported that Taiwan Mandarin listeners were unable to use duration as a cue to stress due to it not being a primary cue to any lexical contrasts in their L1.

Note that English listeners' sensitivity was very low in the HHH and LLL conditions as well. This, again, stems from their reliance on the vowel quality cues. The results were consistent with the findings of [23] that without vowel quality cues, L1 and L2 learners of English (whose L1 is Cantonese) performed equally bad in a discrimination task when the F0 information was neutralized and only duration and intensity cues were available.

5. CONCLUSION

The present study examined the perception of English lexical stress by L1 Taiwan Mandarin L2 learners of English. The results of the stress identification task showed that Mandarin listeners associated higher F0 with the presence of stress and lower F0 with the absence of it, and they failed to make use of non-tonal suprasegmental cues to process stress in the L2 when F0 cues were ineffective. The findings provide robust evidence for the Cue-weighting Transfer Hypothesis by showing that the use of a suprasegmental cue can transfer from one phonological category to another and that how L2 learners use this cue to process lexical stress is affected by how much weight it carries in their L1.

6. REFERENCES

[1] A. Chrabaszcz, M. Winn, C. Y. Lin, and W. J. Idsardi, "Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1468–1479, 2014.

[2] H. Kim and A. Tremblay, "Korean listeners' processing of suprasegmental lexical contrasts in Korean and English: A cue-based transfer approach," *J Phon*, vol. 87, p. 101059, 2021.

[3] Z. Qin, Y.-F. Chien, and A. Tremblay, "Processing of word-level stress by Mandarin-speaking second language learners of English," *Appl Psycholinguist*, vol. 38, no. 3, pp. 541–570, 2017.

[4] A. Tremblay, M. Broersma, Y. Zeng, H. Kim, J. Lee, and S. Shin, "Dutch listeners' perception of English lexical stress: A cue-weighting approach," *J Acoust Soc Am*, vol. 149, no. 6, pp. 3703–3714, 2021.

[5] Y. Zhang and A. Francis, "The weighting of vowel quality in native and non-native listeners' perception of English lexical stress," *J Phon*, vol. 38, no. 2, pp. 260–271, 2010.

[6] M. E. Beckman and J. Edwards, "Articulatory evidence for differentiating stress categories," *Phonological structure and phonetic form: Papers in laboratory phonology III*, pp. 7–33, 1994.

[7] T. Gay, "Physiological and acoustic correlates of perceived stress," *Lang Speech*, vol. 21, no. 4, pp. 347–353, 1978.

[8] B. Lindblom, "Spectrographic study of vowel reduction," *J Acoust Soc Am*, vol. 35, no. 11, pp. 1773–1781, 1963.

[9] D. B. Fry, "Experiments in the perception of stress," *Lang Speech*, vol. 1, no. 2, pp. 126–152, 1958.

[10] M. Liberman, *The Intonational System of English*. Cambridge: MIT Press, 1975.

[11] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology*, vol. 3, pp. 255–309, 1986.

[12] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[13] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Doctoral dissertation, Massachusetts Institute of Technology, 1980.

[14] Y. R. Chao, *A grammar of spoken Chinese*. Univ of California Press., 1965.

[15] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.

[16] S. Ou, "Perception of English lexical stress with a marked pitch accent by native speakers of Mandarin," *Taiwan Journal of Linguistics*, vol. 14, no. 2, pp. 1–31, 2016.

[17] W. Choi, "Theorizing positive transfer in cross-linguistic speech perception: The Acoustic-Attentional-Contextual hypothesis," *J Phon*, vol. 91, p. 101135, 2022.

[18] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," *Behav Res Methods*, vol. 44, no. 2, pp. 325–343, 2012.

[19] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," *The OSU Research Foundation*, vol. 3, no. 30, pp. 255–309, 1997.

[20] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav Res Methods*, vol. 52, no. 1, pp. 388–407, 2020.

[21] N. A. Macmillan and C. D. Creelman, "1991. Detection theory: a user's guide," *Cambridge University PressMacmillanDetection theory: a user's guide1991*, 2005.

[22] R Core Team, "R: A language and environment for statistical computing; 2018." 2018.

[23] W. Choi, X. Tong, and A. G. Samuel, "Better than native: Tone language experience enhances English lexical stress discrimination in Cantonese-English bilingual listeners," *Cognition*, vol. 189, pp. 188–192, 2019.

ORTHOGRAPHY AND THE MENTAL LEXICON: THE EFFECTS OF ENGLISH SILENT LETTERS ON FRENCH LEARNERS

Marine Mouquet[1] & Paolo Mairano[2]

[1] Aix-Marseille Université (France), LPL UMR 7309
[2] Université de Lille (France), STL UMR 8163
marine.mouquet@univ-amu.fr ; paolo.mairano@univ-lille.fr

ABSTRACT

Previous L2 research has shown that direct exposure to orthographic input increases orthographic effects in L2 speech. In our previous work, we demonstrated that French learners of English produce intrusive consonants in correspondence of silent letters when directly exposed to orthographic input. The present study investigates whether this persists when orthography is not provided. 91 French undergraduate students specialising in English and enrolled in various French universities performed a picture naming task and an AB audio preference task in which a phoneme had been added to each stimulus reflecting grapheme-phoneme correspondence (e.g., *[ˈsæ**l**mən] for *salmon*). Results clearly confirm the strong effects of spelling on L2 pronunciation and phonological representations, therefore disclosing that: (i) orthographic input modulates L2 phonological knowledge and (ii) that orthography is directly activated when producing and listening to speech in a second language.

Keywords: second language acquisition; effects of orthography; silent letters; speech production; L2 phonology

1. INTRODUCTION

It is common knowledge that language teachers rely on orthographic input as well as spoken input [1], and recently there has been growing interest in the interplay between orthography and phonology in second language (L2) acquisition. Recent research provides evidence that orthographic representations can affect – positively or negatively – L2 learners' speech and perception [2]. In this respect, various experiments have shown that providing orthographic input in addition to acoustic input can lead to more target-like sounds than providing acoustic input only [3], it can assist in discriminating phonological contrasts [4], and enhance word learning [3]. Nonetheless, it has been shown that it can also lead to non-target-like productions that would not have been produced as such without exposure to orthography [3], including phone additions, omissions and substitutions (cf. [2], [5] and [6], among others). These studies provide clear evidence that exposure to the orthographic forms can affect L2 production [7], metalinguistic awareness [8] and perception: work by Bassetti and colleagues has shown that orthography can cause the importation of a phonological category of the L1 (gemination) into the L2. Henceforth, it seems important to contribute to this field to comprehend orthographic effects on L2 speech production, perception and mental representations of L2 phonological forms.

Our previous work [9] replicated Bassetti & Atkinson's study [5] on silent letters but focussed on French learners of L2 English (N=110). We demonstrated that French learners' productions contained intrusive consonants in correspondence of silent letters in spelling (e.g.*[ˈkəʊm**b**], *[ˈkɒm**b**] or *[ˈkɔ̃**b**] for *comb*) in 47% of productions elicited via orthographic input during a read-aloud task.

The present study expands on our previous work and turns to the effect of silent letters on L2 learners' mental representations. In order to test whether silent letters affect L1 French learners' mental representations of L2 English words, we ran two tests: (i) a production task in which participants were not exposed to silent letters, and (ii) an AB audio preference task where participants had to select an A or B sound corresponding to a word illustrated via an image. A and B were either the standard spoken form of a word (e.g. [ˈbɒm] for *bomb*), or a version of the word with an added consonant reflecting grapheme-phoneme correspondence (e.g., *[ˈbɒm**b**] for *bomb*).

Additionally, we analyse the interplay of orthographic effects with acquisitional variables (English proficiency, stay abroad and length of study abroad, onset of L2 acquisition, number of languages spoken and the preferred English variety) and lexical variables (written vs spoken frequencies and cognate effects).

Our main hypothesis is the following:

1. intrusive consonants in correspondence of silent letters will be present, but less frequently than in our previous study, because this time orthography is not displayed.

Additionally, we also expect that:

2. words with high spoken frequency should be pronounced less often with intrusive

consonants in correspondence of silent letters, since learners will have heard more often the pronunciation of such words and consequently will have had more chances of storing the correct phonological version;

3. cognate words will not have any evident effects on the production of intrusive consonants in correspondence of silent letters, reflecting our previous findings;
4. the number of intrusive consonants pronounced by each participant in the production task in correspondence of silent letters will be correlated to the number of stimuli with intrusive consonants selected in the AB audio preference task;
5. acquisitional variables could interact with orthographic effects. We predict that less proficient learners and those who have had less exposure to native English input will more often pronounce intrusive consonants and select stimuli with intrusive consonants.

2. DATA AND METHODS

Our tests were hosted on a website and run remotely. This enabled us to: (i) gather a significant sample of participants (N=91), (ii) recruit participants from different parts of France, and (iii) compare the results with our previous study, which was run online on a similar platform.

2.1. Participants

Participants were 91 French undergraduate students (women=73, men=18; these figures are indicative of a gender bias typically present in the Departments of Foreign Languages in France) specialising in English in various French universities at BA level: 30 first-year, 19 second-year and 42 third-year students with no reported language or reading impairments, aged 18 to 65 (M=21, MD=20, SD=5.17). The median self-reported CEFR level of English was B2 and ranged from B1 to C2. 2 participants had one English-native speaker as a parent, but none of them claimed to be early bilinguals. On average, participants started to learn English at the age of 8 (MD=7, SD=2.46), and had been studying English for approximately 13 years. 12 participants (13%) had lived in an English-speaking country (the UK, the US, and New Zealand) with an average duration of 8.42 months (MD=7, SD=6.5). On average, participants spoke 3.37 languages (including French and English, MD=3, SD=1.05). Participation was voluntary and unpaid.

2.2. Tasks

Two tasks were developed for this study: a production task (a picture naming task, inspired by [10]) and a listening task (an AB audio preference task, inspired by [11], in which a phoneme had been added to each stimulus, e.g., *[ˈsæ**l**mən] for *salmon* via Mbrola resynthesis [12]). Before each task, written instructions were given in English.

In the picture naming task, participants were given an image and were asked to name it by recording themselves within a timeframe of 3 seconds. In order to make sure that participants would recognise each word correctly, a pilot study (N=10) was conducted. Considering that the experiment was performed remotely, a 'Retry' button was inserted, to be used in case of technical problems (but participants were advised to be spontaneous and refrain from using it for other than technical problems). A bar at the top tracked their progress through each task. Once they had named each of the forty stimuli, they could continue with the second task.

In the AB audio preference task, participants were provided with the same pictures along with two audio stimuli: one with the native pronunciation and one with an added or deleted consonant in correspondence of a silent letter (e.g., *[ˈbɒm**b**] for *bomb*, or *[ˈiːtʃ] for ***b**each*) via Mbrola resynthesis (using the British English voice). Stimuli such as *beach* with a deleted phoneme served as distractors. As for target words, we added a sound taking into account the duration of the preceding or following consonant (e.g., the added /l/ of *half* had the same duration as the adjacent /f/).

Participants were instructed to select the sound corresponding to the word pictured in the image. In order to do so, they had to press <w> on their keyboards to select the first sound or <n> for the second sound.

Finally, a follow-up test was performed to check the familiarity of participants with the orthography of target words: they were provided with the images along with the audios and had to spell out the target stimuli.

2.3. Stimuli

The target stimuli (cf. Table 1) were chosen considering three aspects: words with which French learners struggled the most in our previous work [9], that is, words in <mb> (e.g., *thumb*) and <al + C> (e.g., *almond*), the frequency of words containing silent letters and congruency with French cognates (e.g., *bomb*(EN)*/bombe*(FR)).

In addition, we considered lexical frequency in COCA (Am. Eng.) and BNC (Br. Eng.) because the

degree of exposure to different English varieties varies significantly among our participants.

Target	COCA	BNC
ha**l**f	219.35	294.66
i**s**land	99.49	99.1
clim**b**	50.17	59.81
bom**b**	53.33	43.33
knee	51.67	42.97
mus**c**le	42.81	37.66
colum**n**	36.82	42.41
wrap	44.76	25.62
knife	36.13	32.63
autum**n**	9.74	38.22
lam**b**	12.31	21.45
thum**b**	18.85	14.29
psychologist	18.43	13.23
sa**l**mon	14.29	14.07
bisc**u**it	5.35	15.62
tom**b**	7.91	8.75
com**b**	8.2	7.63
a**l**mond	5.87	5.49
cha**l**k	5.11	10.18
plum**b**er	3.27	2.86

Table 1: The series of 40 chosen words for the two tasks classified by total frequency extracted from COCA (Corpus of Contemporary American English) and the BNC (British National Corpus) (in lemma, per mil).

Distractors were chosen to include phonological contrasts with which French learners of L2 English are known to struggle (e.g., *hungry/angry*), and homophones used in Bassetti & Atkinson [5], in order to mislead participants' attention from silent letters: *hungry, angry, flour, flower, write, right, through, squirrel, beach, sandwich, clothes, jewellery, photographer, always, medal, salt, hospital, fruit, stomach,* and *apricot.*

2.4. Data analysis

The recordings of the twenty target words in the production task were examined for each student. Based on auditory and spectrographic analysis via Praat [13], each target consonant was coded as realised or as silent. As for the AB preference task, the test recorded the selected stimulus and the reaction time (RT) for each pair.

3. RESULTS

Results reveal that although participants were not exposed to orthography during the task, 31% of the target words were produced with an intrusive consonant during the picture naming task and 27.4% of the words selected during the AB audio preference task included an intrusive consonant. 97% of participants pronounced at least one intrusive consonant in the picture naming task and 98% of them selected at least one stimulus with an intrusive consonant in correspondence of a silent letter. As seen in Figure 1, silent letters with which participants struggle the most were in the context of <al + C> (*salmon, almond*) and <mb> (*thumb, bomb*, *plumber, comb*, *lamb* and *climb*).

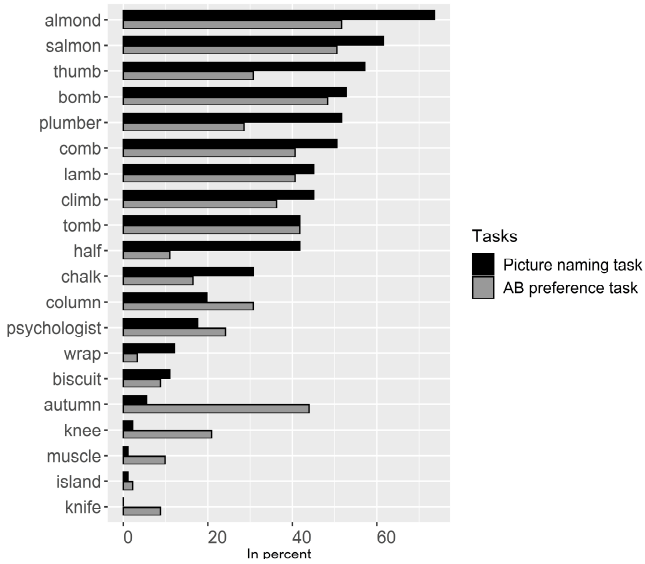


Figure 1: A clustered column chart showing the percent of silent letters realised in the picture naming task and selected in the AB preference task.

The correlation between the percentage of intrusive consonants produced in the first task and selected in the second task by each *participant* was fairly low and barely tended to significance ($r \approx 0.19$, $p \approx 0.07$), revealing that participants who produced more intrusive consonants in the first task were not necessarily the same participants who selected more stimuli with intrusive consonants in the second task. We also ran a linear mixed-effects model with lme4 ([14]) on R to predict the rate of silent consonants that were pronounced in the picture naming task vs selected in the audio preference task, with word and participant as random effects with by-task random slopes: the effect of task turned out to be non-significant, suggesting that on average participants did not perform better or worse in one task than in the other.

Nonetheless, the correlation of intrusive consonants produced in the first and selected in the second task for each *item* was high and significant ($r \approx 0.72$, $p < .001$), suggesting that words for which the silent letter was realised most often in the production

task tended to be selected with an intrusive consonant in the AB audio preference task.

As for learner-level variables, a Pearson test disclosed that participants' self-reported proficiency correlated with the number of realised or selected intrusive consonants in the two tasks ($r\approx-0.225$, $p\approx0.032$; $r\approx-0.29$, $p<.001$), suggesting that L2 proficiency has an effect on performance. Nonetheless, no other correlation was significant for learner-level variables such as length of stay abroad ($r\approx0.085$, $p\approx0.79$; $r\approx-0.37$, $p\approx0.24$), number of languages spoken ($r\approx-0.115$, $p\approx0.278$; $r\approx-0.1$, $p\approx0.322$), and onset of L2 acquisition ($r\approx-0.08$, $p\approx0.42$; $r\approx-0.11$, $p\approx0.28$).

In addition, congruency status with French cognates did not affect the number of intrusive consonants produced or selected ($r\approx0.084$, $p\approx0.818$ for the first task; $r\approx-0.385$, $p\approx0.271$ for the second task). However, a Spearman test revealed significant lexical frequency effects for the picture naming task (written frequency: $r\approx-0.442$, $p\approx0.05$ and spoken frequency: $r\approx-0.451$, $p\approx0.046$) and for the AB audio preference task for spoken frequency only ($r\approx-0.516$, $p\approx0.02$). No significant written frequency effect was instead found for the AB audio preference task ($r\approx-0.317$, $p\approx0.173$).

4. DISCUSSION

The present study have investigated the effects of orthography on the acquisition and processing of L2 words and provides further evidence for orthographic effects in L2 production and mental representations. The findings add support to Bassetti & Atkinson [5] and our own previous study [9]. While previous results showed that Italian and French learners of English produced a significant number of intrusive consonants during a reading aloud task (47% of French learners, 80% of Italian learners), we found that orthographic effects were reduced in a picture naming task where spelling is not shown (31% of the target words were produced) and that stimuli with intrusive consonants were selected as preferred in 27.4% of the target stimuli.

Two (possible) competing explanations could justify such effects [15]. One argues that learning a written language influences the structure of the system of spoken language and that consequently, orthographic representations are automatically activated during the processing of spoken language, as demonstrated for the L1 ([16]). Another concentrates on the learning process, during which learners are exposed to orthographic forms, therefore influencing phonological representations. Connine & Ranbom [11], Whatmough et al. [17], Ehri & Wilce [18] and Taft & Hambly [19] argue that L1 phonological and orthographic representations in the mental lexicon are intrinsically linked; our findings show that this is also the case in an L2. The present results provide evidence that orthographic and phonological information are co-activated during lexical retrieval. The lexicon of L2 learners is nonetheless orthographically biased: orthographic representations modulate phonological knowledge, therefore directly affecting the mental representations of L2 learners and resulting in non-target-like L2 pronunciations (as also seen in [8]). As Young-Scholten [20] and Bassetti [2] also suggested, it is as if "the orthographic input has somehow moved from the page to the mind of the learner" ([2], p7). Future studies could investigate whether this could be avoided when L2 learning is not orthographically-based (e.g., in a non-formal context, as in [15]).

These results provide further support that although French has an opaque orthography and includes silent letters, L1 French learners of L2 English are nonetheless affected by the opaqueness of the English orthographic system, and in particular, their silent letters, as it was already shown by [9] with direct exposure to orthography. This further blurs the hypothesis [21] that L2 learners with an L1 opaque writing system tend to rely less on L2 orthographic input than those with an L1 transparent writing system. In this view, the decoding skills acquired in the L1 do not seem to affect the way L2 words are learned, retained, and produced as opacity and French cognates did not yield significant effects on our results. This further stresses the importance of orthographic input in a L2 phonological context in which acquisition mechanisms should be better apprehended.

These findings add further implications for L2 phonology acquisition theory and for language teaching practices. While exposure to orthography cannot always be avoided [22] and can also be beneficial [3], rules for grapheme-phoneme conversion during classes may not always be effective ([23] and [9]). From a theoretic perspective, we once more argue for the need to include the effects of orthography within L2 phonological models (which currently do not consider spelling, e.g. [24] and [25]).

Other factors such as lexical frequency and proficiency also seem to play a role, and many learner-individual variables that we have not tested may also have an effect, such as the type of memory used by learners (e.g., visual or auditory [3]). Future studies could investigate the influence of visual memory on L2 production and perception with respect to orthographic effects.

5. REFERENCES

[1] Hayes-Harb, R., Barrios, S. 2021. The influence of orthography in second language phonological acquisition. *Language Teaching*, 54(3), 297-326, doi: 10.1017/S0261444820000658.

[2] Bassetti, B. 2008. Orthographic input and second language phonology. In: Piske, T. and Young Scholten, M. (eds.), *Input matters in SLA*, Clevedon, UK: Multilingual Matters, 191-206.

[3] Bürki, A., Welby, P., Clément, M., Spinelli, E. 2019. Orthography and second language word learning: Moving beyond "friend or foe?". *The Journal of the Acoustical Society of America*, 145(4), EL265-EL271, doi: 10.1121/1.5094923.

[4] Escudero, P., Simon, E., Mulak, K. E. 2014. Learning words in a new language: Orthography doesn't always help. *Bilingualism: language and cognition*, 17(2), 384-395.

[5] Bassetti, B., Atkinson, N. 2015. Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics*, 36(1), 67–91. doi: 10.1017/S0142716414000435.

[6] Bassetti, B., Sokolović-Perović, M., Mairano, P., Cerni, T. 2018. Orthography-induced length contrasts in the second language phonological systems of L2 speakers of English: Evidence from minimal pairs. *Language and Speech*, 61(4), 577-597.

[7] Bassetti, B., Escudero, P., Hayes-Harb, R. 2015. Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics*, *36*(1), 1-6.

[8] Bassetti, B., Mairano, P., Masterson, J., Cerni, T. 2020. Effects of Orthographic Forms on Second Language Speech Production and Phonological Awareness, With Consideration of Speaker-Level Predictors. *Language Learning*, *70*(4), 1218-1256.

[9] Mouquet, M., Mairano, P. Forthcoming. Effects of silent letters on the L2 English pronunciation of L1 French learners.

[10] Cerni, T., Bassetti, B., Masterson, J. 2019. Effects of orthographic forms on the acquisition of novel spoken words in a second language. *Frontiers in Communication*, 4, 31.

[11] Ranbom, L. J., Connine, C. M. 2011. Silent letters are activated in spoken word recognition. *Language and Cognitive Processes*, *26*(2), 236-261.

[12] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vreken, O. 1996. The MBROLA project. *Proc. ICSLP '96 Philadelphia*, 1393-1396.

[13] Boersma, P., Weenink, D. 2019. Praat: doing phonetics by computer [Computer program]. Version 6.0.49, from http://www.praat.org/

[14] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[15] Mitterer, H. 2021. The role of orthography in learning a second language: Evidence from Maltese English. *Xjenza Online*, 9, 162–172, doi: 10.7423/XJENZA.2021.3.09.

[16] Pattamadilok, C., Perre, L., Dufau, S., Ziegler, J. C. 2009. On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience*, *21*(1), 169 – 179.

[17] Whatmough, C., Arguin, M., Bub, D. 1999. Cross-modal priming evidence for phonology-to-orthography activation in visual word recognition. *Brain and Language*, 66(2), 275-293.

[18] Ehri, L. C., Wilce, L. S. 1980. The influence of orthography on readers' conceptualization of the phonemic structure of words. *Applied Psycholinguistics*, 1(4), 371-385.

[19] Taft, M., Hambly, G. 1985. The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, 24(3), 320-335.

[20] Young-Scholten, M. 2002. Orthographic input in L2 phonological development. *An integrated view of language development: Papers in honor of Henning Wode*.

[21] Erdener, V. D., Burnham, D. K. 2005. The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning*, *55*(2), 191-228.

[22] Thornbury, S. 1999. How to teach grammar (Vol. 3). Longman Harlow.

[23] Bassetti, B., Cerni, T., Masterson, J. 2022. The efficacy of grapheme-phoneme correspondence instruction in reducing the effect of orthographic forms on second language phonology. *Applied Psycholinguistics*, 43(3), 683-705.

[24] Best, C. 1995. A direct realist view of cross-language speech. In W.Strange (Ed.), *Speech perception and linguistic experience*, 171-204. Baltimore: York Press.

[25] Flege, J. 1995). Second language speech learning: Theory, findings and problems. *Speech perception and linguistic experience: Issues in cross-language research* 92, 233-277.

INFLUENCES OF MULTI-LANGUAGE EXPERIENCE ON MANDARIN CONSONANT PERCEPTION

Shuwen Chen[1], Janice Wing Sze Wong[2]

[1]Institute of Linguistics, Chinese Academy of Social Sciences, [2]Hong Kong Baptist University
[1]chensw@cass.org.cn, [2]janicewong@hkbu.edu.hk

ABSTRACT

Cantonese speakers from Hong Kong learn Mandarin under the influence of both Cantonese and English. This study investigates the influence of Cantonese and English on Mandarin consonant perception. Nineteen Cantonese speakers were tested in an identification task followed by a cross-language mapping and goodness-rating task when perceiving Mandarin fricatives, affricates, liquids and the alveolar nasal. The results showed that Cantonese-English bilinguals mapped Mandarin consonants onto both Cantonese and English consonants. Language dominance and perceptual similarity both play a role in cross-language assimilation. The cross-language mapping pattern, however, only partially predicts the identification results. In addition, mapping Mandarin consonants onto English sound categories has a facilitative effect on the perception of Mandarin consonants.

Keywords: Speech perception, Mandarin consonants, Cantonese-English bilinguals

1. INTRODUCTION

An important question in L2 speech acquisition is how cross-language similarities and differences affect and predict the acquisition of L2 sounds. Perceptual Assimilation Model - L2 (PAM-L2) [1] predicts that if two non-native sounds are assimilated to two different L1 categories, the perception of the two sounds will be good. If two non-native sounds are assimilated to the same L1 category, the discrimination of the two sounds will be poor. If two non-native sounds are assimilated as a good version and a poor version of the same category, the accuracy is predicted to be intermediate. Previous studies on L2 vowel and consonant perception have provided some support for this hypothesis [2, 3]. A relevant and intriguing question is whether the link between cross-language similarities and perception results can be extended to the scenario where a third sound system is learned. Previous studies suggested that the first two sound systems can both have an impact on the production and perception of the third sound system [4, 5]. But will the sound categories in the third system be mapped onto the sound categories in the first or the second sound system? Can cross-language mapping patterns predict perceptual accuracy when three sound systems interact?

To answer these questions, the current study examined the perception of Mandarin consonants by Cantonese-English bilinguals. Cantonese and Mandarin are both Chinese languages, but their phonological systems are quite different. Mandarin has a rich inventory of sibilants. It contrasts three manners of articulation for dental sounds (/s, ts, tsʰ/), retroflex sounds (/ʂ, tʂ, tʂʰ/) and alveolo-palatal sounds (/ɕ, tɕ, tɕʰ/). In contrast, Cantonese has a much smaller inventory of sibilants (/s, ts, tsʰ/). English has six alveolar and post-alveolar sibilants (/s, z, ʃ, ʒ, tʃ, dʒ/). In addition, both Mandarin and English contrast /n/, /ɹ/, and /l/, while Cantonese does not have a /ɹ/ phoneme and is gradually losing /n/-/l/ contrast [6]. Given the differences in consonant inventories, it is interesting to examine how Cantonese-English bilinguals map Mandarin consonants onto Cantonese or English consonants. Does English facilitate Mandarin consonant perception? Can cross-language assimilation patterns predict the accuracy of Mandarin consonant identification?

2. METHOD

2.1. Participants

Nineteen undergraduate students (15 females and 4 males) from Hong Kong Baptist University were invited to participate in the experiment. Their average age was 20 years old. They were born and grew up in Hong Kong. They all speak Cantonese as their first language, and learned English and Mandarin later. They started learning English at an average age of three and Mandarin at an average age of four. In the language background survey, participants reported that they used Cantonese and English more frequently than Mandarin in daily communication and in school, and had a higher proficiency in Cantonese and English than in Mandarin. All participants formally learned *Pinyin* before participating in the experiment.

2.2. Stimuli

The target consonants included Mandarin alveolar fricative and affricates /s, ts, tsʰ/, retroflex fricative

and affricates /ʂ, tʂ, tʂʰ/, and palatal fricative and affricates /ɕ, tɕ, tɕʰ/, approximants /l, ɹ/ and alveolar nasal /n/. The fricatives and affricates are followed by the high front vowel /i/ or the apical vowel /ɿ, ʅ/, and the liquids and the nasal are followed by /i/ and /a/ (see Table 1). The stimuli were read by two Mandarin native speakers from Beijing (one male and one female) and recorded in a sound-treated room. The target syllables were embedded in the carrier phrase "wo ʂwo __ tʂɤ kɤ tsʅ" (I say ___ this word). After the recording, the target syllables were cut out to be used in the experiment.

Target consonant	Stimuli	Target consonant	Stimuli
ts	tsʅ	n	ni
tsʰ	tsʰʅ		nan
s	sʅ	l	li
tʂ	tʂʅ		lan
tʂʰ	tʂʰʅ	ɹ	ɹʅ
ʂ	ʂʅ		ɹan
tɕ	tɕi		
tɕʰ	tɕʰi		
ɕ	ɕi		

Table 1: Stimuli used in the experiment.

2.3. Procedure

The participants completed an identification task, and a cross-language mapping and goodness rating task in a classroom. The stimuli were presented through the loudspeakers in the classroom, and the subjects were asked to respond using online questionnaires. The stimuli were played in a random order using Praat [7]. In the identification task, the subjects heard the stimuli and they were asked to choose which sound it was in *Pinyin*. The 30 stimuli (15 words × 2 speakers) were repeated three times, resulting in a total of 90 trials. In the second task, participants listened to the stimuli and were asked to classify the sounds they heard as one of the Cantonese categories (/s, ts, tsʰ/) or English categories (/s, z, ʃ, tʃ, dʒ/). English /ʒ/ was not included because it cannot occur in the syllable initial position. They were then asked to rate the similarity between the word they heard and the sound they chose on a scale of 1 (unlike) to 5 (very similar). They were allowed to choose one sound (either Cantonese or English), or two sounds (one from Cantonese and one from English). Each stimulus was identified and rated two times, resulting in a total of 60 trials.

2.4. Statistical analysis

For the identification results, identification accuracy for each sound was calculated, and confusion matrices were summarized. For the cross-language mapping and goodness rating task, the percentage of identification for each category in Cantonese and English was calculated. Following Guion et al. [8], the "fit index" was calculated to combine the identification percentage and goodness rating into one single matrix. It is calculated by multiplying the percentage of identification by the goodness rating of that identification. For example, if the percentage of identifying Mandarin /ts/ as Cantonese /ts/ is 24%, and the goodness rating of this identification is 1.67, then the fix index of this classification is 0.40.

3. RESULTS

3.1. Identification results

Figure 1 shows the accuracy of identifying Mandarin consonants by Cantonese-English bilinguals. Mandarin /ɹ/ and /n/ were identified with the highest accuracy. The accuracy of /tɕʰ/ (88%) and /tɕ/ (87%) was highest among the fricatives and affricates. The accuracy of /tsʰ/ (62%) and /s/ (71%) was the lowest among all the tested consonants.

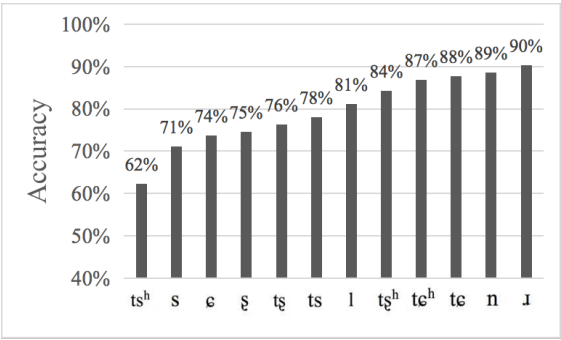


Figure 1: Accuracy of identifying Mandarin consonants

Table 2 is the confusion matrix of Mandarin fricatives and affricates. First, it is easy for the speakers to confuse Mandarin /ɕ/ and /s/. It is also hard for them to distinguish the three retroflex sounds / ʂ, tʂ, tʂʰ/, where /tʂ/ and /ʂ/ were often identified as /tʂʰ/. In addition, Mandarin /tsʰ/ was often identified as /tɕʰ/, /tʂʰ/ or /ts/.

		Responses								
		ts	tsʰ	s	tʂ	tʂʰ	ʂ	tɕ	tɕʰ	ɕ
Stimuli	ts	89	5	1	7	0	0	8	3	1
	tsʰ	10	71	1	0	10	0	2	18	2
	s	2	10	81	2	0	3	0	2	14
	tʂ	5	1	1	87	14	0	5	1	0
	tʂʰ	0	5	0	5	96	4	0	4	0
	ʂ	0	2	6	3	16	85	0	0	2
	tɕ	8	3	0	1	1	0	100	1	0
	tɕʰ	1	5	2	1	6	0	0	99	0
	ɕ	1	7	15	1	0	3	0	3	84

Table 2: Confusion matrix of Mandarin fricatives and affricates.

Table 3 is the confusion matrix of Mandarin nasal /n/ and approximants /l, ɹ/. The most common confusion was between /n/ and /l/. The confusion between /l/ and /ɹ/ can also be seen, but was less frequent.

		Responses		
		ɹ	l	n
Stimuli	ɹ	206	13	9
	l	12	185	31
	n	3	23	202

Table 3: Confusion matrix of Mandarin nasal and approximants.

3.2. Perceptual assimilation results

Table 4 and 5 summarized the mean percent identification and goodness rating of Mandarin consonants in terms of Cantonese and English sound categories. When compared to Cantonese categories, Mandarin /ts/, /tʂ/ and /tɕ/ were most similar to Cantonese /ts/, Mandarin /tsʰ/, /tʂʰ/ and /tɕʰ/ were most similar to Cantonese /tsʰ/, and Mandarin /s/ were most similar to Cantonese fricative /s/. When compared to English categories, Mandarin /s/ and /ɕ/ were assimilated to English /s/, Mandarin /ts/ to English /z/, Mandarin /ʂ/ to English /ʃ/, Mandarin /tsʰ/, /tʂʰ/ and /tɕʰ/ to English /tʃ/, Mandarin /tʂ/ and /tɕ/ to English /dʒ/.

As to the liquids and the nasal, Mandarin /l, ɹ, n/ were mapped onto English /l, ɹ, n/ respectively. When compared to Cantonese sounds, Mandarin /l, ɹ, n/ were all assimilated to Cantonese /n/.

	Cantonese			English			
	n	l	None of them	l	n	ɹ	None of them
l	**8 (1.38)**	4 (2)	64	**26 (3.04)**	10 (2.6)	1 (2)	39
n	**9 (1.44)**	3 (2.67)	64	3 (2.33)	**33 (3.06)**	1 (2)	39
ɹ	**3 (1.67)**	2 (3)	71	1 (1)	3 (3)	**9 (1.33)**	63

Table 5: Mean percent identification and goodness rating (in parentheses) of Mandarin nasal and approximants in terms of Cantonese and English categories. Boldfaced values indicate the most commonly identified responses in each language.

To compare the cross-language mapping pattern more directly, the fit index of the most commonly identified Cantonese and English sounds was calculated and summarized in Table 6. A higher score of fix index indicates more similarity between Mandarin consonants and the identified sound. According to Table 6, among the 12 Mandarin consonants, 6 were mapped onto Cantonese consonants, and 6 were mapped onto English consonants. Mandarin /tɕ/ and /tʂ/ were mapped onto Cantonese /ts/, Mandarin /s/ and /ɕ/ were mapped onto Cantonese /s/, Mandarin /tsʰ/ and /tɕʰ/ were mapped onto Cantonese /tsʰ/. The rest of the Mandarin consonants were mapped onto English sounds. Mandarin /tʂʰ/ and /ʂ/ were mapped onto English /tʃ/ and /ʃ/ respectively. As expected, Mandarin /n, ɹ, l / were mapped onto English /n, ɹ, l/. But unexpectedly, Mandarin /ts/ was mapped onto English /z/ rather than Cantonese /ts/.

	Cantonese				English					
	ts	tsʰ	s	None of them	s	z	ʃ	tʃ	dʒ	None of them
ts	**18 (1.67)**	0	2 (2)	56	0	**24 (2)**	0	2 (2)	6 (1.67)	44
tsʰ	3 (1.33)	**19 (1.42)**	0	54	1 (1)	4 (2.5)	0	**9 (1.78)**	1 (1)	61
s	2 (1.5)	2 (2.5)	**21 (2.14)**	51	**9 (2)**	6 (2.67)	1 (2)	1 (1)	0	59
tʂ	**21 (2)**	0	1 (1)	54	0	5 (1.4)	0	4 (2)	**9 (1.78)**	58
tʂʰ	1 (1)	**17 (1.82)**	1 (2)	57	1 (3)	4 (2)	0	**16 (2.06)**	0	55
ʂ	0	6 (2.17)	**7 (1.14)**	63	1 (1)	1 (1)	**14 (1.43)**	9 (1.78)	1 (1)	50
tɕ	**56 (3.48)**	5 (2.4)	3 (1.33)	12	3 (2)	21 (3.43)	0	4 (2.75)	**31 (2.97)**	17
tɕʰ	3 (2.67)	**54 (3.35)**	1 (1)	18	1 (2)	8 (3)	1 (1)	**23 (2.65)**	2 (2.5)	41
ɕ	1 (4)	2 (1.5)	**51 (2.96)**	22	**47 (3.15)**	6 (2.5)	1 (3)	1 (1)	0	21

Table 4: Mean percent identification and goodness rating (in parentheses) of Mandarin fricatives and affricates in terms of Cantonese and English categories. Boldfaced values indicate the most commonly identified responses in each language.

		Cantonese				English				
		Most common identification	Proportion of identification	Goodness rating	Fit index	Most common identification	Proportion of identification	Goodness rating	Fit index	Most fit sound category
Mandarin consonants	ts	ts	24%	1.67	0.40	z	32%	2.00	0.64	English /z/
	tsʰ	tsʰ	25%	1.42	0.36	tʃ	12%	1.78	0.21	Cantonese /tsʰ/
	s	s	28%	2.14	0.60	s	12%	2.00	0.24	Cantonese /s/
	tʂ	ts	28%	2.00	0.56	dʒ	12%	1.78	0.21	Cantonese /ts/
	tʂʰ	tsʰ	22%	1.82	0.40	tʃ	21%	2.06	0.43	English /tʃ/
	ʂ	s	9%	1.14	0.10	ʃ	18%	1.43	0.26	English /ʃ/
	tɕ	ts	74%	3.48	2.58	dʒ	41%	2.97	1.22	Cantonese /ts/
	tɕʰ	tsʰ	71%	3.35	2.38	tʃ	30%	2.65	0.80	Cantonese /tsʰ/
	ɕ	s	67%	2.96	1.98	s	62%	3.15	1.95	Cantonese /s/
	l	n	11%	1.38	0.15	l	34%	3.04	1.03	English /l/
	n	n	12%	1.44	0.17	n	43%	3.06	1.32	English /n/
	ɹ	n	4%	1.67	0.07	ɹ	12%	1.33	0.16	English /ɹ/

Table 6: Fit indexes for Mandarin consonants in terms of Cantonese and English categories.

4. DISCUSSION

Based on the results of the cross-language mapping task, the listeners map Mandarin consonants onto both Cantonese and English sound categories. This is consistent with Wrembel et al. [9] which shows that the phonemes in the third sound system are mapped onto categories in the other two sound systems. Our results also suggest that both language dominance and perceptual similarity play a role in cross-language mapping. Although Cantonese has a much smaller inventory of sibilants than English does, the bilinguals still map six Mandarin sibilants onto Cantonese, and mapped the other three Mandarin sibilants onto three English sibilants. It should also be noted that in the perceptual assimilation task, the majority of responses were "none of them" for many consonants. It seems to indicate that the participants are aware of the cross-linguistic difference between the three sound systems, and try to keep them separate.

The cross-language mapping pattern only partially predicts Mandarin identification results. According to PAM-L2, if two sounds are mapped onto the same L1 category, it would be hard for learners to distinguish the two sounds. According to the mapping results, Mandarin /s/ and /ɕ/ were both mapped onto Cantonese /s/, and Mandarin /tsʰ/ and /tɕʰ/ were both mapped onto Cantonese /tsʰ/. In the identification task, we did find that participants often mixed /s/ with /ɕ/, and /tsʰ/ with /tɕʰ/. However, there are some exceptions. Mandarin /tɕ/ and /tʂ/ were both mapped onto Cantonese /ts/, but Mandarin /tɕ/ were seldom mixed with /tʂ/, and were often identified as /ts/. Mandarin /tʂ/ was more often identified as /tʂʰ/ than /tɕ/.

In terms of identification accuracy, it is expected that Mandarin consonants that are mapped onto the same Cantonese or English category would have lower accuracy because learners tend to mix those consonants, while consonants that are mapped onto separate categories would have higher accuracy. As expected, we did find that the identification accuracy of Mandarin /ɹ/ and /n/ was high, and Mandarin /s/ and /ɕ/ had a low accuracy. Contrary to the prediction, the identification accuracy of Mandarin /tɕ/ and /tɕʰ/ was high although they were mapped onto Cantonese /ts/ and /tsʰ/ with other Mandarin sounds. It suggests that the relationship between cross-language mapping and perception accuracy is more complex than PAM predicts when there are three sound systems.

The results also show that English consonants facilitate the discrimination of Mandarin consonants for Cantonese-English bilinguals. Cantonese does not contrast /l/ and /ɹ/, but the learners showed high accuracy in identifying the two phonemes. The cross-language mapping pattern has shown that this is because the bilinguals mapped Mandarin /l/ and /ɹ/ onto English /l/ and /ɹ/. A similar effect can be found in /n/-/l/ contrast. Many speakers do not distinguish Cantonese /l/ and /n/ in daily communication, but the identification accuracy of /n/ is still high. This might be the result of a positive transfer from English. This facilitative effect is also reported in Onishi [10].

5. CONCLUSION

In the current study, the perception of Mandarin consonants by Cantonese-English bilinguals was tested. The results showed that the bilinguals mapped Mandarin consonants onto both Cantonese and English categories. Cross-language assimilation is influenced probably by both perceptual similarity and language dominance. Mapping Mandarin sounds onto English sound categories can facilitate the identification of Mandarin consonants when English has contrasts that Cantonese does not have. However, cross-language mapping patterns only partially predict the identification results. It suggests that the relationship between cross-language assimilation and perceptual accuracy is more complex when there are three sound systems.

7. REFERENCES

[1] Best, C. T., Tyler M. D. 2007. Non-native and second language speech perception: Commonalities and complementaries. In Munro, M., Bohn, O-S. (eds), *Second Language Speech Learning*. Amsterdam: John Benjamins Publishing, 13-34.

[2] Tyler, M., Best, C., Faber., A, Levitt, A. 2013. Perceptual assimilation and discrimination of non-native vowel contrast. *Phonetica* 71, 4-21.

[3] Best, C. T., McRoberts, G. W., & Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am*. 109(2), 775–794.

[4] Zhu, Y., Chen, A., Sudhoff, S., Mok, P.P. 2019. Third language prosody: Evidence from Cantonese-English-German Trilinguals. *Proc. 19th ICPhS*, Melbourne.

[5] Balas, A., Kopečková, R., Wrembel, M., Mickiewicz, A. 2019. Perception of rhotics by multilingual children. *Proc. 19th ICPhS*, Melbourne.

[6] Yip, V., Matthews, S. 2000. *Basic Cantonese: A Grammar and Workbook*. London: Routledge.

[7] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer. http://www.praat.org/.

[8] Guion, S.G., Flege, J.E., Akahane-Yamada, R., Pruitt, J. 2000. An investigation of current models of second language speech perception: the case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am*. 107(5), 2711-2724.

[9] Wrembel, M., Marecka, M., Kopečková, R. 2019. Extending perceptual assimilation model to L3 phonological acquisition. *International Journal of Multilingualism* 16(4), 513-533.

[10] Onishi, H. 2016. The effects of L2 experience on L3 perception. *International Journal of Multilingualism* 13(4), 459-475.

Distributional learning and overnight consolidation of a difficult non-native tone contrast

[1]Yin-To CHUI, [2]Zhen QIN

The Hong Kong University of Science and Technology
[1]ytchuiac@connect.ust.hk, [2]hmzqin@ust.hk

ABSTRACT

Previous research has demonstrated the success of distributional training for the acquisition of both segmental and suprasegmental categories, but no study has tested the distributional learning effect of a difficult tone contrast after overnight consolidation. The perception of Mandarin level tone (T1) – falling tone (T4) contrast has been shown to be especially difficult for native (L1) Cantonese speakers. The present study addresses whether this tonal contrast is successfully learned with distributional training, and whether this training effect is enhanced after overnight consolidation. 62 L1-Cantonese speakers were divided into two groups, where they were exposed to two different frequency distributions of the T1-T4 continuum (bimodal vs unimodal). Results showed slightly better post-training performance by the bimodal group, and this difference is enhanced after overnight sleep. This shows that the effect of overnight consolidation extends to statistical learning of lexical tones.

Keywords: Distributional training, overnight consolidation, lexical tone, Mandarin

1. INTRODUCTION

The distributional learning paradigm first appeared in [1] and followed a line of research on statistical learning mechanisms of language [2, 3]. Distributional learning refers to learning to discriminate between two phonetic categories (e.g. /d/ vs /t/) purely by implicit exposure to distributional information along a phonetic continuum (e.g. Voice onset time, VOT). As illustrated in Figure 1, a bimodal distribution consisting of more typical tokens of the two categories was found to benefit subsequent discrimination performance, as opposed to a unimodal distribution consisting of ambiguous tokens in between.

This ability to learn implicitly from frequency distributions has been shown to apply to infants [1, 4] and adults [5, 6], where it was demonstrated for consonants [1], vowels [4, 5], and lexical tones [6]. However, this seemingly ubiquitous nature of distributional learning is in fact not universal, as it has been demonstrated to be modulated by factors like learners' linguistic background. For example, the distributional learning of a Dutch vowel contrast /ɑ/-/a:/ might work for L1-Spanish speakers [5] but not L1-English speakers [7].

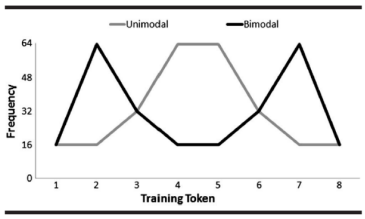


Figure 1: **Frequency counts of each token as heard in the unimodal and bimodal conditions [6]**

One aim of the present study is to see whether the distributional learning of tonal contrasts may also be subject to the influence of prior linguistic experience. Specifically, while previous research has demonstrated the success of distributional training of non-native tonal categories for non-tonal speakers (i.e. L1-English) [6, 8], the same might not be concluded for tonal speakers (e.g. L1-Cantonese), since the two populations exhibit differential processing difficulties of non-native tones. For example, the Mandarin level tone (T1) – falling tone (T4) contrast has been shown to be especially difficult for L1-Cantonese speakers, even more so than L1-English speakers [9]. Specifically, both tones in the contrast are perceptually assimilated to the same Cantonese high level tone, falling into the [Single Category] or [Category Goodness] classification according to the Perceptual Assimilation Model (PAM), which PAM posits to be hard to distinguish. This difficulty was verified empirically in a series of perception and production tasks, where L1-Cantonese speakers performed worse on this contrast than L1-English speakers [9].

The first goal of the present study, therefore, is to establish whether there is a distributional learning effect of the Mandarin T1–T4 contrast for L1 speakers of Cantonese. While previous studies have shown successful distributional learning of this contrast for L1-English speakers [8], it may turn out to fail for L1-Cantonese speakers given the added difficulty that arises from linguistic background.

Another unanswered question is whether the benefit of distributional training, if any, is enhanced after overnight consolidation. Memory consolidation refers to the stabilization of newly learned memory traces by transformation into long-term memory. Literature on sleep and learning suggests that sleep facilitates learning a new linguistic rule [10]; specifically, sleep that comes shortly after training facilitates the stabilization of newly-learned novel tonal categories, for example Cantonese tones by L1-Mandarin speakers [11]. Previous studies on distributional learning have focused on measuring only immediate performance after training. However, it is arguable that successful language learning should be indexed not only by immediate performance, but also whether the performance can be retained, so it is beneficial to identify whether consolidation benefits distributional learning as well where statistical learning mechanisms come into place. So far, only one distributional training study tested the longitudinal learning effect by retesting participants at 6- and 12-month intervals [12]. The results showed a lasting effect of distributional learning even in delayed intervals, but the influence of natural exposure outside the laboratory cannot be ruled out in that study.

The second goal of the present study, therefore, is to test whether the distributional learning effect (i.e. the performance of the bimodal over the unimodal group) is enhanced after overnight sleep. The benefit of sleep has been documented in other statistical learning paradigms [13], and the results of the current study can inform whether overnight consolidation extends to distributional learning.

2. METHODS

2.1. Participants

62 young adult native speakers of Cantonese were recruited in Hong Kong. Initial screening ensured that they had less than 9 years of Mandarin training and a self-reported proficiency of 'intermediate' or below. A cognitive battery was administered for matching purposes, which tested their pitch threshold [14], pitch memory [15], musical aptitude [16], working memory [17], and Mandarin vocabulary size [18]. A summary is given in Table 1.

	Unimodal	Bimodal
No. of participants	30 (9 males)	32 (15 males)
Age (year)	19.4 (1.37)	20.0 (1.49)
Pitch threshold test – Glide (in semitones)	4.6 (4.2)	4.3 (3.4)
Pitch memory task (in number of pitch notes)	5.4 (1.2)	5.2 (1.4)
MBEA (Musical aptitude)	0.83 (0.07)	0.85 (0.06)
Operation Span (Working memory)	60.9 (14.4)	63.2 (7.7)
PPVT(Mandarin-vocabulary size)	0.83 (0.07)	0.82 (0.08)

Table 1: Means (and standard deviations, in parentheses) of cognitive battery variables; MBEA: Montreal Battery of Evaluation of Amusia; PPVT: Peabody Picture Vocabulary Test

2.2. Stimuli

2.2.1. Discrimination Pre-test/post-test

A female and male native Mandarin speaker were invited to provide multiple recording tokens of two novel syllables /nua/ and /fao/ in Mandarin T1 and T4 (i.e. solid lines in Figure 2). Three tokens of each gender-syllable-tone combination were used in the discrimination task (see Procedure). The stimuli were novel to minimize the effect of prior exposure, and multiple syllables and genders were included to test participants' generalization of their discrimination performance to untrained settings (since only female /nua/ was used in the training phase).

2.2.2. Distributional training

The eight-step continuum used to construct the bimodal/unimodal distributions (Figure 1) was created using Praat. Only female /nua/ was used in training. Pitch contours of the naturally produced tones constituted the two end points of the continuum, and then 6 equidistant intermediate steps were synthesized in between using the ProsodyPro script in Praat [19], with the following procedure: 10 interpolation points were identified along the pitch contours (at 0%, 10%, …, 90%), then the distances between corresponding points were divided into seven equal spaces (Figure 2). The duration was normalized to the mean duration of the naturally produced tones, and the intensity was normalized to 65 dB SPL. Five native Mandarin speakers listened to the stimuli and confirmed that they are acceptable tokens of the syllable, and categorized the endpoints as good exemplars of T1 and T4. In addition to the syllable tokens, 32 sine-wave beeps were interspersed between the tokens for attended listening (see Procedure).

2.3. Procedure

The experiment followed a pre-test, training, and post-test procedure. Besides the T1-T4 pair, T1-T2

(i.e., a level-rising tonal contrast) was also tested for control purposes (stimuli created in the same way as described in section 2.2), and the results are not reported here given the space limit.

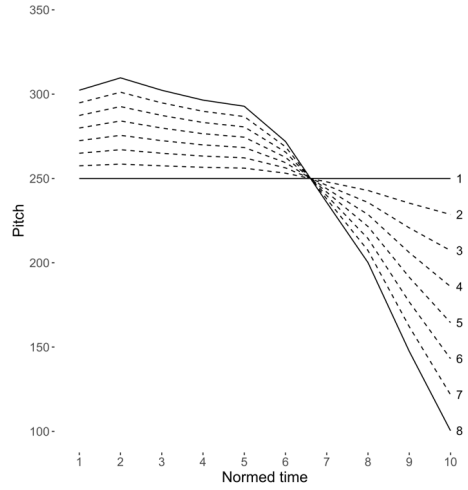


Figure 2: Mandarin T1-T4 continuum. Dashed line = intermediate tokens. Numbers = step number.

Participants first performed ABX tone discrimination tasks for the T1-T4 pair (order counter-balanced). In this task, subjects were exposed to natural T1 and T4 reference tokens A and B, followed by the target token X (taken from another recording) and was asked to identify whether X is the same tonal category as A or B in a 2-alternative forced choice (2AFC) task (Figure 3). All gender-syllable combinations were used here in a randomized, unblocked design. The inter-stimulus interval was chosen to be 1000ms, consistent with previous literature that this duration targets phonological categorization and processing. There were altogether 64 trials in each test. Participants' discrimination accuracy was chosen as the target measure.

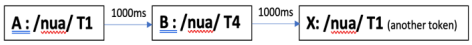


Figure 3: Example trial of the ABX discrimination task

Participants, matched in the cognitive tests, were then pseudo-randomly assigned to either the bimodal training group or the unimodal training group, where they were exposed to a bimodal tonal continuum or unimodal tonal continuum of /nua/ produced by the female talker. The syllable tokens were presented in a randomized order, with a total of 256 trials. In addition, 32 sine-wave beeps were interspersed in between the tokens, in which participants were instructed to press a keyboard response every time it appeared. This was to ensure the participants were actively attending to the stimuli, which has been shown to be necessary to induce distributional learning [6]. Interstimulus interval was set at 750 ms, and each training phase took around 10 minutes to complete.

Immediately after training, the discrimination task was assigned again (i.e. post-test 1). They were then instructed to wear a Fitbit device to record their sleep quality for the night, and they came back the next morning, after 12 hours, for the third round of ABX discrimination tasks (i.e. post-test 2).

2.4. Data analysis

Mixed-effects logistic regression models were performed on participants' response accuracy in the ABX discrimination tasks (binary-coded, 1 for correct and 0 for incorrect). Non-responded trials were excluded. The models were fitted in R using the lme4 package. Fixed effect is distribution (two levels: bimodal vs unimodal; deviation coding: -0.5, 0.5). To compare the performance of the bimodal and unimodal groups for each session (pre-test, post-test 1, and post-test 2), three nested models were performed with the effect of distribution nested within each session [20]. In all models, trial was entered as random effect.

3. RESULTS

Accuracy of the two groups in the ABX discrimination tasks is illustrated in Figure 4. The graph shows a steadily improving performance by the bimodal group across sessions, and a stable performance for the unimodal group.

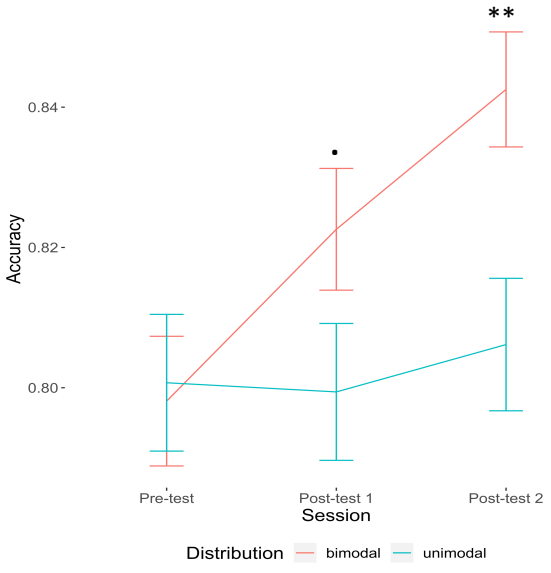


Figure 4: Accuracy (and standard errors) in ABX discrimination task by distribution and by session

A summary of the three nested models is presented in Table 2. The first nested model shows that the two groups' difference in accuracy in the pre-test did not reach significance, showing that the bimodal and unimodal groups had comparable baseline performance before distributional training. The second nested model shows that in post-test 1, immediately after training, the performance of the two groups started to diverge with marginal significance, with the bimodal group performing better than the unimodal group. This difference became highly significant in post-test 2 after overnight consolidation.

Session	Variable	Est.	SE	*z*	*p*
Pre-test	Distribution	0.016	0.084	0.195	0.845
Post-test 1	Distribution	-0.151	0.085	-1.78	**0.075**
Post-test 2	Distribution	-0.252	0.086	-2.916	**<0.01**

Table 2: Mixed effects logistic models on all participants' accuracy in the ABX discrimination task

4. DISCUSSION AND CONCLUSION

This study investigated whether L1-Cantonese speakers can benefit from distributional training of a Mandarin tone pair that has been shown to be difficult due to the influence of existing Cantonese tonal categories, and also whether any distributional learning effects can be enhanced after overnight consolidation. Overall results suggest a weak improvement after distributional training, and a stronger, enhanced improvement after overnight sleep.

In terms of immediate distributional learning performance, results only yielded a marginal improvement of the bimodal group over the unimodal group. When compared with the more robust success of distributional training exhibited in other adult studies testing different phonetic categories [5], or with a similar study testing the same Mandarin tonal contrast on another cohort (L1-English speakers without any prior tonal experience) [8], the weakened distributional learning performance by L1-Cantonese learners may indeed lend further support that distributional learning is modulated by prior linguistic experience, which, in the current case, might have arisen from the relative perceptual difficulties experienced by different L1 speakers as suggested by PAM. To make a stronger conclusion for this case, a direct comparison study of two participant groups of different linguistic backgrounds (i.e. L1-Cantonese vs L1-English) is needed in future research. It is also worth noting that we allowed our participants to have up to 9 years of Mandarin training, and amount of L2 experience may have played a role in the results, as participants' baseline performance may have been higher than naïve learners to begin with, making immediate improvement more difficult. More research is needed to confirm the effect of perceptual assimilation on the weakened effect of distributional learning for naïve learners, as it has been suggested that naïve learners will show an even stronger effect of perceptual assimilation [21] than experienced learners [9, 22], since naïve learners may rely purely on acoustic similarity, but experienced learners may pay more attention to the phonemic status of the target contrast and find it easier to distinguish between them [23].

In terms of performance after overnight consolidation, results showed that only the bimodal group continued to improve overnight, while the unimodal group stagnated at a stable performance level. This shows that looking at immediate performance may not enough to conclude the presence or absence of a distributional learning effect, as it is shown that even the weakened immediate performance is enhanced and reaches significance at a delay. The bottom line is that the distributional learning effect is there, just that it needed consolidation to make it visible. It will be interesting to see if the lack of effect in other studies can be recuperated if the learners are allowed to consolidate the newly-learned categories through an overnight interval. Current results also imply that overnight consolidation, while demonstrated in other more general training paradigms [11], may extend to the distributional learning of lexical tones.

Overall, the present findings suggest that Cantonese speakers exhibit a weak distributional learning effect of Mandarin tones, which might be due to influence by prior linguistic experience, and that overnight consolidation extends to the distributional learning setting and may be able to uncover distributional learning effects that were thought to be absent.

ACKNOWLEDGEMENTS

The project was funded in part by the Postgraduate Research Grant 2021-22 at the Hong Kong University of Science and Technology. We thank members of our research team, Ms. Ruofan Wu, Ms. Yuqi Wang for helpful comments and suggestions.

5. REFERENCES

[1] Maye, J., Werker, J. F., Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3), B101-B111.

[2] Saffran, J. R., Aslin, R. N., & Newport, E. L. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294), 1926-1928.
[3] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255(5044), 606-608.
[4] Wanrooij, K., Boersma, P., & Van Zuijen, T. L. 2014. Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in psychology* 5, 77.
[5] Escudero, P., Benders, T., & Wanrooij, K. 2011. Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America* 130(4), EL206-EL212.
[6] Ong, J. H., Burnham, D., & Escudero, P. 2015. Distributional learning of lexical tones: A comparison of attended vs. unattended listening. *PloS one* 10(7), e0133446.
[7] Terry, J., Ong, J. H., & Escudero, P. 2015. Passive distributional learning of non-native vowel contrasts does not work for all listeners. *Proc. 17th ICPhS*, Glasgow, UK.
[8] Liu, L., Yuan, C., Ong, J. H., Tuninetti, A., Antoniou, M., Cutler, A., & Escudero, P. 2022. Learning to Perceive Non-Native Tones via Distributional Training: Effects of Task and Acoustic Cue Weighting. *Brain Sciences* 12(5), 559.
[9] Hao, Y. C. 2012. Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of phonetics* 40(2), 269-279.
[10] Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. 2014. Sleep facilitates learning a new linguistic rule. *Neuropsychologia* 65, 169-179.
[11] Qin, Z., & Zhang, C. 2019. The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PloS one* 14(12), e0221498.
[12] Escudero, P., & Williams, D. 2014. Distributional learning has immediate and long-lasting effects. *Cognition* 133(2), 408-413.
[13] Simon, K. N., Werchan, D., Goldstein, M. R., Sweeney, L., Bootzin, R. R., Nadel, L., & Gómez, R. L. 2017. Sleep confers a benefit for retention of statistical language learning in 6.5 month old infants. *Brain and Language* 167, 3-12.
[14] Qin, Z., Zhang, C., & Wang, W. S. Y. 2021. The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *The Journal of the Acoustical Society of America* 149(1), 435-446.
[15] Williamson, V. J., & Stewart, L. 2010. Memory for pitch in congenital amusia: Beyond a fine-grained pitch discrimination problem. *Memory* 18(6), 657-669.
[16] Peretz, I., Champod, A. S., & Hyde, K. 2003. Varieties of musical disorders: the Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences* 999(1), 58-75.
[17] Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. 2005. An automated version of the operation span task. *Behavior research methods* 37(3), 498-505.
[18] Tsoi, E. Y. L., Yang, W., Chan, A., & Kidd, E. (2019). Mandarin–English speaking bilingual and Mandarin speaking monolingual children's comprehension of relative clauses. *Applied Psycholinguistics* 40(4), 933-964.
[19] Xu, Y. 2013. ProsodyPro—A tool for large-scale systematic prosody analysis. Laboratoire Parole et Langage, France.
[20] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. 2009. Mixed Effects Modelling for Nested Data. In: *Mixed effects models and extensions in ecology with R.* Springer New York, 101-142.
[21] So, C. K., & Best, C. T. 2010. Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and speech* 53(2), 273-293.
[22] Leung, A. 2008. Tonal assimilation patterns of Cantonese L2 speakers of Mandarin in the perception and production of Mandarin tones. *Proc. 2008 CLA annual conference*.
[23] Wu, X., Munro, M. J., & Wang, Y. 2014. Tone assimilation by Mandarin and Thai listeners with and without L2 experience. *Journal of Phonetics* 46, 86-100.

STATIC AND DYNAMIC FEATURES OF ENGLISH MONOPHTHONGAL VOWELS PRODUCTION BY HIJAZI ARABIC L2 LEARNERS

Wael Almurashi,[1] Jalal Al-Tamimi,[2] and Ghada Khattab[3]

[1]Taibah University, Saudi Arabia; [2]Université Paris Cité, LLF, France; [3]Newcastle University, United Kingdom
wmurashi@taibahu.edu.sa; jalal.al-tamimi@u-paris.fr; ghada.khattab@newcastle.ac.uk

ABSTRACT

Dynamic cues in monophthongal vowels—in particular, vowel inherent spectral change (VISC)—have been found to add to our understanding of vowel identity and of crosslinguistic differences in vowel patterning. Less work has been carried out on the role they play in understanding second language vowel production. This study investigated the role of static versus dynamic cues (along with additional cues: vowel duration, F0, and F3) in the production patterns of English monophthongal vowels by Hijazi Arabic L2 learners (HA L2) compared with first language speakers of each of Hijazi (HA L1) and English (NE). Data were collected from three groups (20 HA L1, 20 HA L2, and 20 NE) producing target monophthong vowels in a word list with varied consonantal contexts. Results show that dynamic measures and additional cues provide insights into L2 production patterns that are not normally gleaned from static measures or F1 and F2 correlates alone.

Keywords: Static cues, dynamic cues, discriminant analysis, second language learning, Hijazi Arabic.

1. INTRODUCTION

Common practice in the investigation of second language (L2) vowel production relies on measuring the first two formants (F1 and F2) at the monophthong vowel's midpoint and making conclusions about the degree of second-language attainment. For some time, this static approach to vowel measurement was believed to provide an insight into the optimal acoustic characteristics of monophthong vowels [23]. However, many acoustic studies have since reported that measuring vowel formants from multiple locations (e.g., dynamic cues—in particular, vowel inherent spectral change [VISC] models) can provide more information and lead to better identification when using discriminant analysis, a statistical method that is used to predict listeners' vowel categorisation patterns [15]; [16]; [20]; [21]. The investigation of dynamic properties of vowel production remains understudied in the domain of L2 speech acquisition despite the fact that it constitutes a more sophisticated method for exploring the degree to which learners achieve target-like vowel production in their L2 [17].

VISC can be defined as the 'relatively slowly varying changes in formant frequencies associated with vowels, even in the absence of consonantal context' [21]. It is taken between two locations over the full duration of the vowel: one near the vowel's onset (at around 20%) and the other near the vowel's offset (at around 80%), to minimise the effects of surrounding consonants. VISC has been quantified through three approaches, namely: the offset model, which investigates the degree of spectral change; the slope model, which investigates the rate of spectral change; and finally, the direction model, which investigates the direction of spectral change [20].

In terms of offset, [20] found that speech dynamics are greater for speakers of languages that have a sparse monophthong vowel system (e.g., Chinese) than for those that have a dense vowel system (e.g., Korean and English) possibly due to the former having more variability in production [19]. Additionally, the production of English monophthongal vowels produced by Korean and Chinese L2 learners was found to be influenced by their L1 VISC (e.g., similar amount of VISC). The slope and direction models have also been found to provide a characterization of dynamic cues of monophthongal vowels and have been used to examine the intrinsic dynamic of monophthong vowels in more detail [4]; [5]; [10]; [14]. For instance, [10] concluded that vowels can be reliably distinguished if formant contour is considered, and such a model can help to separate those vowels whose formant values are very close and similar.

Many researchers have built on the direction model and reported that monophthongs can be characterised effectively when their formants are taken from three locations (at 20%, 50%, and 80%), e.g. [12], [16], among others. Another line of research building on the direction model takes the VISC measurement to an advanced level by measuring multiple samplings to represent detailed information of the entire formant trajectories (e.g., [1]; [6]; [11]; [14]; [22]). These studies suggest that using multiple measurements (as well as additional cues such as vowel duration, fundamental frequency [F0], and third formant frequency [F3]) are not only useful in

terms of describing monophthongs but also in terms of classifying monophthongal vowels using discriminant analysis.

Despite the importance of this research for more detailed crosslinguistic comparisons on vowel production, the role of dynamic correlates in L2 vowel production has not been fully investigated [24]. English and Arabic provide a good testing ground for the role of dynamic cues in vowel identity and L2 vowel learning, but we are not aware of research that has examined the dynamic properties of English monophthong vowels by Arabic L2 learners (see [2]; [3], among others, for examples of L2 production studies by Arabic L1 speakers using static models). This study constitutes the first examination of the dynamic production patterns of English vowels by Hijazi Arabic L2 (HA L2) learners and is also the first research contribution to examine L2 production comprehensively in terms of VISC. It examines the production patterns of Standard Southern British English (SSBE) vowels by HA L2 learners compared with English (NE) speakers and L1 patterns. It evaluates the relative importance of static and dynamic cues, particularly VISC, in describing and classifying SSBE vowel production by HA L2; it also explores to what extent vowel duration, F0, and F3 act as additional cues to classification accuracy.

2. METHODOLOGY

Data were collected from 60 participants (20 HA L1, 20 HA L2 learners, and 20 NE balanced by gender). To note, HA L2 learners were intermediate classroom English students in a foreign language setting and had no direct access to native L2 input. Recordings were made on a Roland Edirol R-09 recorder and Audio Technica Cardioid stereo microphone with a sampling rate of 44,100 Hz and 16-bit quantization. The target vowels were examined in a list with varied consonantal contexts. Each HA participants produced 48 words with one of eight HA vowels (/i iː eː a aː oː u uː/) and each L2 and NE participants produced 60 words with one of 10 English monophthong vowels (/ɪ iː ɛ ɔː ɑː ʊ uː æ ɒ ʌ/), with three repetitions (a total of 10,080 tokens). Vowel duration, the first three formant and F0 values were extracted from one location (50% for the static model) and multiple locations (two [20% and 80%]), three [20%, 50%, and 80%], and seven locations [20%, 30%, 40%, 50%, 60%, 70%, and 80%] for the dynamic models) over the course of the vowel duration.

Acoustic analysis was conducted using Praat [9]. The F0 settings were speaker-dependent and formant tracks were obtained using a 0.025 s window length, 50 Hz pre-emphasis, and a maximum frequency of 5,000 Hz for males and 5,500 Hz for females. For the offset model, we obtained the amount of a vowel's spectral changes by calculating the differences for all three formant and F0 values between the two vowel's positions. For the slope model, we obtained the vowel's rate of changes by using the offset value (see above) and then dividing them by the vowel duration. For the direction model, we computed the vowel's spectral shifts by tracking the first three formant and F0 values from two samples (for the two-point model), three samples (for the three-point model), and seven samples (for multiple points). All F0 and formant values were checked manually to ensure the accuracy of the results.

Two types of statistical techniques were used to evaluate the differences in the data—namely, pairwise comparisons (post-hoc tests) after a linear mixed effects regression, to determine the statistical significance of the study results [14]. Then a discriminant analysis (qda function) with a leave-one-out cross-validation, or 'jackknife' [5]; [16] was used as a classification tool to evaluate the extent to which the static and dynamic models and other acoustic feature sets (F0, F1, F2, F3, and vowel duration) improve vowel classification.

3. RESULTS

Beginning with the static model, HA L2 speakers were found to produce some SSBE vowels significantly (e.g., $p < 0.0001$) differently from the NE group, namely, /ʌ/, /ʊ/, /ɒ/, /æ/, /ɪ/, and /ɑː/ in terms of F0; /ʊ/ and /ɪ/ in terms of F1; /ɒ/, /ʌ/, /ɛ/, and /æ/ in terms of F2; and /ɛ/, /æ/, /ɔː/, /ʊ/, /uː/, and /ɒ/ in terms of F3. In contrast, HA and HA L2 groups produced vowels in a similar way (e.g., Figure 1 for F1/F2 estimates).

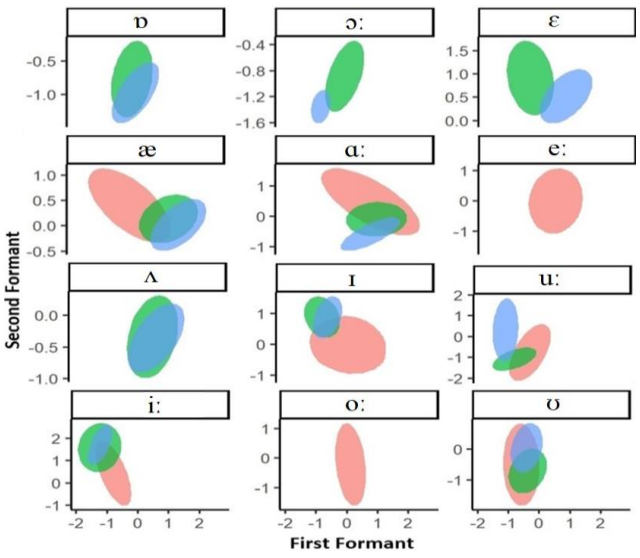


Figure 1: Scatter plot of the normalised midpoints (LOBANOV Z-scores) of the first two formant values of the vowels produced by HA (in red), HA L2 (in green), and NE (in blue) participants.

Regarding dynamic cues, particularly the offset model, HA L2 speakers were found to produce SSBE vowels with a great amount of spectral movement in terms of F1, F2, and F3, but fewer spectral changes in F0 (e.g., see Figure 2 for F2 changes). The results of the offset for HA and HA L2 speakers revealed that HA L2 speakers produced SSBE vowels with a similar amount of VISC as HA L1 speakers (e.g., a great amount of in F1, F2, and F3). HA L2 participants were found to produce most SSBE vowels significantly differently from the NE group: /ʊ/ in terms of F0; /ʌ/, /iː/, /ɔː/, /æ/, /ɛ/, and /ɑː/ in terms of F1; /ʊ/, /ɔː/, and /ɑː/ in terms of F2; and /ʌ/, /ʊ/, /æ/, and /ɑː/ in terms of F3.

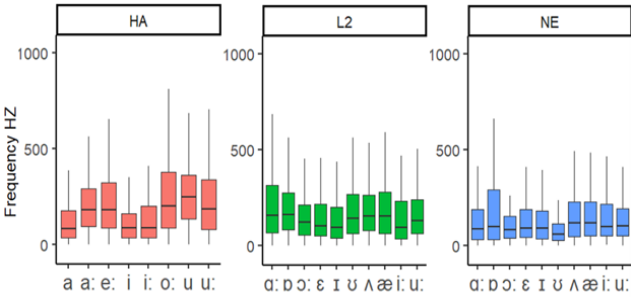


Figure 2: Box plot of the F2 offset model of the vowels produced by HA, HA L2, and NE participants.

In terms of the slope model, HA L2 participants were found to produce SSBE vowels significantly differently from the NE group, particularly for F0, and with some vowels in F1 (e.g., /uː/ and /ɛ/), F2 (e.g., /ɒ/ and /iː/), and F3 (e.g., /iː/). Once again, HA and HA L2 mostly had similar production of vowels, (e.g., see Figure 3 for F0 results).

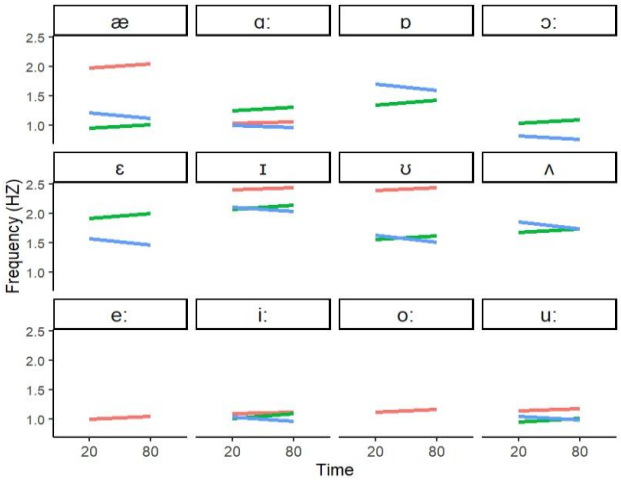


Figure 3: Results of the F0 slope model (measured at two points) of the vowels produced by HA (in red), HA L2 (in green), and NE (in blue) participants.

Regarding the direction of the two-point model, HA L2 was found to produce most SSBE vowels significantly dissimilar from the NE group (e.g., /ʌ/, /ʊ/, /ɒ/, /æ/, and /ɑː/ in terms of F0; /ɪ/ in terms of F1; /ɒ/ in terms of F2; and /ɔː/ in terms of F3) but similarly to their L1 production. The direction from three-point model showed the same results from the direction model from three-point measure, however we noted few more significant results between the production of HA L2 and NE (e.g., /ɪ/ in terms of F0, /ʌ/ and /ɛ/ in terms of F2, /uː/ and /æ/ in terms of F3. Using multiple points model (e.g., seven measurements) showed similar findings to the three-point model, and indeed, more significant results of the production of the L2. The most consistent result pertains to the HA L2 participants' production, which approximated the L1 (HA) participants in terms of the relative position of the vowels in the acoustic space, the direction, and the amount of the spectral change (e.g. Figure 4 for formant direction results).

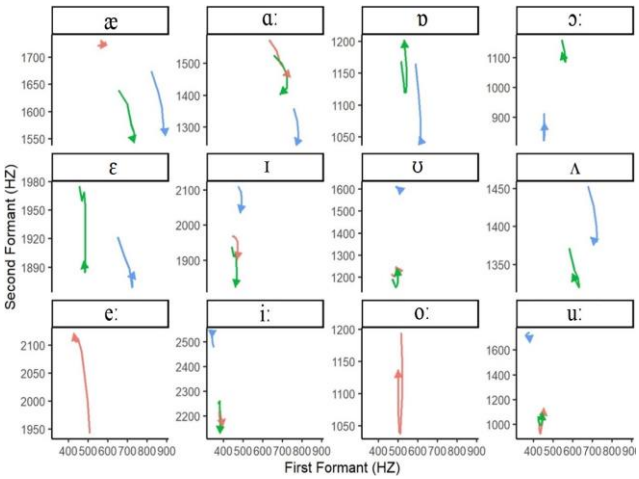


Figure 4: Vowel formant trajectories in F1-F2 space (measured at seven points) of the vowels produced by HA (in red), HA L2 (in green), and NE (in blue) participants; arrows represent the direction of formant movement.

In terms of classification. the QDA showed that using dynamic cues for all three groups (in particular, the seven-point model) with F0, F1, F2, and F3 (without and with the duration) resulted in the highest classification accuracy (with an average of 85% for HA; 80% for NE; and 61% for HA L2), followed by the three-point model (with an average of 76% for HA; 70.5% for NE; and 51% for HA L2), then by two-point model (with an average of 75% for HA; 68.5% for NE; and 50% for HA L2), and finally by the static model (with an average of 71% for HA; 65% for NE; and 48.5% for HA L2). With respect to the additional cues' results, vowel duration was found to play a significant role in the classification accuracy for HA and HA L2, while F0 was the most important additional cue for accurately classifying NE vowels. (see Table 1 for percentage improvement).

	The average improvement percentage		
	Vowel duration	F0	F3
HA L2	5.8	2.8	1.3
HA	11.8	2.1	1.6
NE	4.6	10	5.5

Table 1: The average improvement percentage in the discriminant classification accuracy of the three groups for all models with the addition of the vowel duration, F3, and F0 as additional cues.

4. DISCUSSION AND CONCLUSION

The data on the acoustic correlates of the production of SSBE by HA L2 speakers using the midpoint model showed that most of the SSBE vowels were produced differently from native-like productions, particularly for vowels that do not exist in the learner's L1 (HA), such as /ɔ:/, /ʊ/, /ɒ/, /ɛ/, and /ʌ/, suggesting a difficulty in producing what are often referred to as 'new' vowels in the literature in a target-like manner [13]. Furthermore, the HA L2 static results for 'similar' vowels generally revealed that HA speaker produced many vowels of SSBE with similar patterns to their L1.

The effect of L1 on the performance of L2 was also found in the dynamic cues. For example, for the offset model, HA L2s were found to produce SSBE vowels with a degree of vowel-inherent spectral change that is similar to their HA L1 (e.g., greater spectral shifts in F1, F2, and F3). Such a result is expected due to speakers of low-density languages (e.g., HA) having more freedom and space to produce their vowels compared to high-density languages (e.g., SSBE; [7]; [17]; [19]). Similarly, the slope result showed that HA L2s had mostly similar slope values to their L1 (positive slopes in most cases related to faster spectral changes of HA monophthongal vowels during the vowel duration) and different from NE speakers. Interestingly each of the static, offset and slope models showed differences in terms of which vowels exhibited significant differences in their patterns between HA L2 and NE productions, and in which formants. This highlights the importance of considering measurement method before making firm conclusions about which vowels L2 speakers will find challenging to produce in a target-like manner. Conversely, all three methods stressed the influence of L1 on the production of L2.

Regarding the direction result (from two, three, or seven points), the HA L2 group once again produced most SSBE vowels in similar patterns of formant trajectories to those of their L1 and different from NE speakers. While this supported results from the other models, the data of the direction model revealed that using dynamic measurements provided significantly more information and differences about the production of vowels than the static model did. Moreover, more measuring points from the vowel duration provided richer information about the fuller extent of the vowel spectral changes that might remain unnoticed when formant values are taken from fewer locations. This result supports the necessity of investigating monophthongal vowels by using dynamic measurements (multiple points) [1]; [14]; [22], among others. Overall, such findings confirm that the VISC models could be used as another perspective to examine more closely the extent to which L2 learners are influenced by their L1 [20].

With respect to the classification accuracy of vowels, the QDA results suggest that extracting two measurement points (or more) from the vowel formants shows notable improvements compared with taking a single point. This finding provides support for the dynamic approach [4]; [5]; [12]; [16]. Among the dynamic measurements, the seven-point measure is the most accurate for classifying the vowels' production for HA L2. Such a finding indicates that entire trajectory of the formants might contain additional vowel identity information over and above that captured in two- or three-point model [20]. We also noted that the QDA outcome's average rate of the classification accuracy for HA L2s was very low compared to that of NE speakers which is expected due to the noticeable variations in producing SSBE vowels in all acoustic models used. Despite the efficiency of the F1 and F2 values in identifying vowels, this study highlights the role of additional cues in providing more insights and increased separation of vowels. For example, HA L2s relied mostly on vowel duration (similarly to their L1), and this can be explained by considering the phonological role of vowel duration as a cue to distinguishing short and long vowels in Arabic vowels [4]; [5]; [8]; [18]. In addition, the use of F0 and F3 by each of HA L2 participants in SSBE vowel production, exhibited influence from the speakers' L1. Taken Together, these findings reveal that cues to vowel identification are not expressible in one time location and that the transitional changes perform significant functions in terms of describing and classifying monophthongal vowels [4]; [5]; [6]; [15]; [16]; [20]; [21]. These can be missed when crosslinguistic vowel comparisons are reduced to single points, and in turn concealing potentially important differences between L1 and L2 speakers' production of these vowels.

To sum up, our findings are consistent with dynamic theories of vowels and heed recent calls to expand the sociophonetic toolkit beyond single-point measures and beyond the first two formants to represent more information and reflect the dynamic complexity of actual vowel movement. As quoted from [25], 'traditional two-dimensional vowel charts are not sufficient, and … a deeper understanding is gained when dynamic patterns of formant movement are documented'; we believe we underscored this point in this paper.

5. ACKNOWLEDGMENTS

This work was supported by Taibah University to the first author (WA) and partially supported by the French Investissements d'Avenir—Labex EFL program (ANR-10-LABX-0083), contributing to the IdEx Université Paris Cité—ANR-18-IDEX-0001 to the second author (JAT). We thank all of our subjects who participated in this study.

6. REFERENCES

[1] Adank, P., Van Hout, R., and Smits, R. 2004. An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America* (*J. Acoustic. Soc. Am*), 116(3), 1729-1738.

[2] Ali, E. 2013. Pronunciation problems: Acoustic analysis of the English vowels produced by Sudanese learners of English. *International Journal of English and Literature*, 4(10), 495-507.

[3] Almbark, R. 2014. Perception and production of SSBE vowels by foreign language learners: Towards a foreign language model. In *Proceedings of the International Symposium on the Acquisition of Second Language Speech: Concordia Working Papers in Applied Linguistics*, 5, Montreal, Canada, 3-18.

[4] Almurashi, W., Al-Tamimi, J., and Khattab, G. 2019. Static and dynamic cues in vowel production in Hijazi Arabic. In *Proceedings of the 19th ICPhS*, Melbourne, Australia, 3468-3472.

[5] Almurashi, W., Al-Tamimi, J., and Khattab, G. 2020. Static and dynamic cues in vowel production in Hijazi Arabic. *J. Acoustic. Soc. Am*, 147(4), 2917-2927.

[6] Al-Tamimi, J. 2007. Static and dynamic cues in vowel production: a cross dialectal study in Jordanian and Moroccan Arabic. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 541-544.

[7] Al-Tamimi, J., and Ferragne, E. 2005. Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2465-2468.

[8] Al-Tamimi, J., and Khattab, G. 2015. Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *J. Acoustic. Soc. Am*, 138(1), 344-60.

[9] Boersma, P., and Weenink, D. 1992–2022. Praat: Doing phonetics by computer.

[10] Chladkova, K., and Hamann, S. 2011. High vowels in Southern British English: /u/-fronting does not result in merger. In *Proceedings of the 17th ICPhS*, Hong Kong, China, 476-479.

[11] Darcy, I., and Mora, J. C. 2015. Tongue movement in a second language: the case of Spanish/ei/-/e/ for English learners of Spanish. In *Proceedings of the 18th ICPhS*, Glasgow, UK.

[12] Ferguson, S. H., and Kewley-Port, D. 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoustic. Soc. Am*, 112(1), 259-271.

[13] Flege, J., and Bohn, O. S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (3-83). Cambridge University Press.

[14] Fox, R. A., and Jacewicz, E. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *J. Acoustic. Soc. Am*, 126(5), 2603-2618.

[15] Hillenbrand, J., Clark, M., and Nearey, T. M. 2001. Effects of consonant environment on vowel formant patterns. *J. Acoustic. Soc. Am*, 109(2), 748-763.

[16] Hillenbrand, J., Getty, L. A., Clark, M., and Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoustic. Soc. Am*, 97(5), 3099-3111.

[17] Jin, S. H., and Liu, C. 2013. The vowel inherent spectral change of English vowels spoken by native and non-native speakers. *J. Acoustic. Soc. Am*, 133(5), 363-369.

[18] Khattab, G. 2007. A phonetic study of gemination in Lebanese Arabic. In *Proceedings of the 16th ICPhS*, Saarbrücken, Germany, 153-158.

[19] Meunier, C., Frenck-Mestre, C., Lelekov-Boissard, T., and Le Besnerais, M. 2003. Production and perception of vowels: Does the density of the system play a role? In *Proceedings of the 15th ICPhS*, Barcelona, Spain, 723-726.

[20] Morrison, G. S., and Assmann, P. 2013. *Vowel inherent spectral change*. Springer Science and Business Media.

[21] Nearey, T. M., and Assmann, P. F. 1986. Modeling the role of inherent spectral change in vowel identification. *J. Acoustic. Soc. Am*, 80(5), 1297-1308.

[22] Neel, A. T. 2004. Formant detail needed for vowel identification. *Acoustics Research Letters Online*, 5(4), 125-131.

[23] Peterson, G. E., and Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoustic. Soc. Am*, 24(2), 175-184.

[24] Rogers, C.L., Glasbrenner, M.M., DeMasi, T.M., and Bianchi, M. 2013. Vowel inherent spectral change and the second-language learner. In G. S. Morrison, and P. F. Assmann (Ed.), *Vowel inherent spectral change* (231-259). Springer, Berlin.

[25] Schwartz, G., Aperliński, G., Kaźmierski, K., and Weckwerth, J. 2016. Dynamic targets in the acquisition of L2 English vowels. *Research in Language*, 14(2), 181-202.

THE ROLE OF INTONATION TYPE AND STRESS POSITION IN NATIVE MANDARIN SPEAKERS' PERCEPTION OF ENGLISH LEXICAL STRESS

Weijing Zhou[1], Francis Nolan[2], Jiawei Jing[3], Jian Gong[4]
[1,3] Yangzhou University, [2] University of Cambridge , [4] Jiangsu University of Science and Technology
[1] zhouwj@yzu.edu.cn, [2] fjnolan@gmail.com, [3] 674581920@qq.com , [4] habaogj@hotmail.com

ABSTRACT

This is an original empirical study investigating the extent to which intonation type (IT) and stress position (SP) influence the perception of L2 English lexical stress by native Mandarin speakers. The stimuli were 16 sets of non-vowel-reducing disyllabic English heteronyms, which differs in stress positions (trochaic vs. iambic) and intonation types (read with falling vs. rising contours). Thirty Mandarin speakers majoring in English completed a lexical stress test using two-alternatives forced-choice identification paradigm, and the accuracy rates (AR) and reaction times (RT) of their lexical stress perception were compared. The results reveal that both IT and SP had significant main effects on subjects' lexical stress judgment, that IT exhibited a more robust role than SP according to effect sizes, and that there are genuinely significant interaction effects between IT and SP signifying a complex interactive network existing in native Mandarin speakers' perception of L2 English lexical stress.

Keywords: intonation type; stress position; English lexical stress perception; native Mandarin speakers.

1. INTRODUCTION

Stress plays a pivotal role in word recognition and speech communication in English, not only because English is a typically 'stress-timed' language in rhythm, but also that stress embodies both segmental and suprasegmental features. It is acoustically manifested in multiple dimensions, especially pitch (f0), duration, intensity, and vowel quality [1] [2] [3] [4] [5] [6] [7] [8]. However, the pattern or weighing of these acoustic cues varies from lexical stress languages (like English) to languages that arguably have no stress or no easily perceived stress in their word phonology (like Mandarin) [6] [9] [10] [11] [12] [13]. Furthermore, variation exists within the family of lexical stress languages [11] [14] [15] [16]. There is also a clear bias in how native English listeners decide that a syllable is stress-bearing with an important cue being a full vowel, but with other features playing an important role [16] [17]. Such complexity is very prone to make it difficult and problematic, directly or indirectly, for non-naïve users to produce and perceive English lexical stress [12] [18] [19] [20] [21].

Research on the perception of L2 English word stress by native French listeners identified so-called 'stress deafness' among native French speakers, as stress (or accent) in French is often regarded as a phrasal phenomenon and consistently falls on the last syllable of words or phrases [19] [20]. Quite a few follow-up cross-linguistic studies supported the claim of the Stress Parameter Model (SPM) [22] [23] that typological stress similarity between L1 and L2 may determine whether the acquisition of L2 word stress is attainable, i.e., if L1 exhibits irregular and contrastive stress patterns, its native speakers tend to have an easy time to encode L2 English lexical stress; otherwise they will encounter obstructions like native French speakers' 'stress deafness'.

However, less research addresses the questions of whether learners whose L1 has a radically different type of lexical prosody (like Mandarin) can encode L2 English stress phonologically and use it for word recognition [20], and whether their L2 English stress perception is influenced by or interacts with some suprasegmental properties of words which constrain lexical access in their L1 (e.g., tones in Mandarin and pitch accents in Japanese). Both issues are worthy of investigation for researchers and practitioners to have a better understanding of native Mandarin speakers' acquisition of English lexical stress as well as the improvement on their English speech performance.

Although vowel quality has been proved to be the most important variable in processing English word stress among both native and non-native English speakers [12] [14] [15] [16] [20] [21], most previous studies have neglected stress contrasts lacking vowel reduction (e.g., PROtest /'prəʊtest/ & proTEST /prə'test/) thus leaving the picture quite incomplete.

Mandarin, as a typical tone language, does not straightforwardly have lexical stress but uses tones to distinguish lexical items, such as 妈 (/mā/ mother)，麻 (/má/ hemp)，马 (/mǎ/ horse)，骂 (/mà/ scold). A number of studies have verified that native Mandarin learners of English as a foreign language (EFL) run into considerable difficulties and pitfalls in perceiving English lexical stress due to the absence of lexical stress in L1 Mandarin word phonology – so-called negative L1 transfer [20] [21] [24] [25] – yet none employed stimuli well controlled in vowel quality, nor

were they intended to explore fundamental issues such as whether intonation has a role in native Mandarin speakers' processing of L2 English lexical stress. Due to the theoretical and practical significance of these unattended questions, the present study examines the role of intonation type, stress position, and their interaction in native Mandarin speakers' perception of English lexical stress.

2. METHODS

2.1. Participants

Thirty English majors who speak Mandarin as their L1 were recruited to participate in this study. They were first-year undergraduates majoring in Business English and had just finished a compulsory term course in English Pronunciation before the experiment. They were all born in China and had never been to any English-speaking countries. Their average age was 18.5 years, and the ratio of male to female is 17% (N=5) to 83% (N=25). As their English scores in both the National College Entrance Examination and self-reported overall English proficiency (including listening, speaking, reading, writing and translation) were 'good' and 'excellent', they could be assumed to be near-advanced speakers of English with respect to the whole population of college students. All had normal or corrected-to-normal vision as well as normal hearing.

2.2. Stimuli

Table 1: The stimuli for the perception test of English lexical stress

IT	Falling ↘		Rising ↗	
SP	Trochaic	Iambic	Trochaic	Iambic
1	**DI**gest	di**GEST**	**DI**gest	di**GEST**
2	**IN**crease	in**CREASE**	**IN**crease	in**CREASE**
3	**IN**sult	in**SULT**	**IN**sult	in**SULT**
4	**SUR**vey	sur**VEY**	**SUR**vey	sur**VEY**
5	**IN**cense	in**CENSE**	**IN**cense	in**CENSE**
6	**TOR**ment	tor**MENT**	**TOR**ment	tor**MENT**
7	**IM**port	im**PORT**	**IM**port	im**PORT**
8	**TRANS**form	trans**FORM**	**TRANS**form	trans**FORM**
9	**IM**print	im**PRINT**	**IM**print	im**PRINT**
10	**IN**cline	in**CLINE**	**IN**cline	in**CLINE**
11	**IN**tern	in**TERN**	**IN**Tern	in**TERN**
12	**FORE**arm	fore**ARM**	**FORE**arm	fore**ARM**
13	**MIS**fire	mis**FIRE**	**MIS**fire	mis**FIRE**
14	**TRANS**fer	trans**FER**	**TRANS**fer	trans**FER**
15	**MIS**hit	mis**HIT**	**MIS**hit	mis**HIT**
16	**TRANS**plant	trans**PLANT**	**TRANS**plant	trans**PLANT**

Notes: IT=intonation type SP=stress position

The stimuli included 16 sets of disyllabic heteronyms. Each set of words had identical vowels in stressed and unstressed syllables (e.g., /aɪ/ & /e/ in **DI**gest /ˈdaɪdʒest/ vs. di**GEST** / daɪˈdʒest/), but differed in stress positions (trochaic vs. iambic / stress located on the 1st vs. 2nd syllable) and intonation types (read with falling vs. rising contours) (see Figure 1). Therefore, the whole experiment consisted of 64 trials for identification. They were read and recorded by the second author, a native speaker of British English and phonetician. The sampling rate of recording was 44.1kHz.

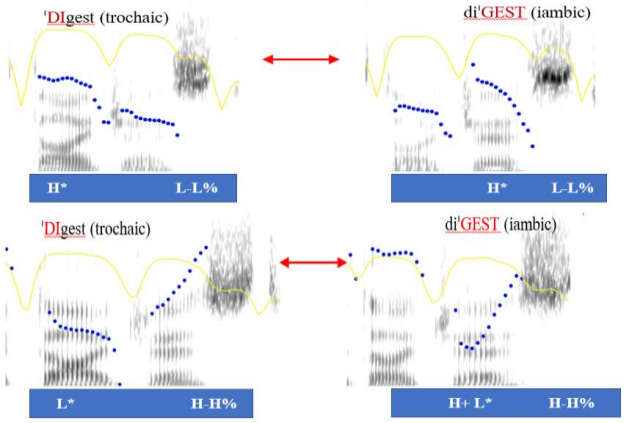


Figure 1:The Intonation patterns of *digest* with stress position on the 1st vs. 2nd syllable.

2.3. Data collection and analysis

The experiment was conducted in Yangzhou University Laboratory of Phonetics, Hearing & Cognitive Science, with the participants seated in a comfortable chair at around 70cms from an LCD monitor. The stimulus presentation program was written with E-Prime (Version 3.0). The auditory stimuli were presented binaurally through headphones at 70dB SPL. And the test follows a two-alternatives forced-choice paradigm.

The experiment consisted of 4 steps: **Step 1**: ***Introduction***. Prior to the experiment, the instructions on how to do the test and requirements for doing the task were delivered on the screen. **Step 2**: ***Rehearsal***. A practice test of eight self-paced trials plus a response after each trial was carried out so as to ensure all the participants knew exactly the procedure and the task. **Step 3**: ***Main Test***. The test consisted of 64 trials, with each trial presented only once. The presentational order of the trials was randomized across participants so that they were unlikely to be able to anticipate the upcoming trial. Each trial began with a red cross on the screen and an auditory instruction "ding" sound for 1000 milliseconds. Then a target word was binaurally delivered over the headphones and the participants were required to identify the word with perceived pattern of intonation type (IT) or stress position (SP) shown on the screen as quickly as possible without sacrificing accuracy. The data of accuracy rates (AR) and reaction time（RT）were generated and stored onto the computer right after each trial. **Step 4**: ***Ending***.

A thank-you note was presented on the screen when all the trials were over.

The data of AR and RT were then checked and calculated via Excel (Version 2017) before being statistically analysed via Jamovi (Version 1.2).

3. RESULTS

3.1. Effects of intonation type

As shown in Table 2 and Figure 2, the average RTs to perceive all the target lexical stresses in falling vs. rising intonation are 2989ms and 3260ms respectively with a gap of 271ms, indicating that the participants took longer or more effort to discriminate lexical stress embedded in rising intonation than in falling intonation, though the RT differences are not statistically significant ($F_{(29)} = -1.76$, P=.090, Effect Size = -0.320).

Table 2: Paired-sample T-test results: The AR & RT means of the perception of target lexical stress read in falling vs. rising intonation.

	N	M	SD	T	df	P	Effect Size
F-RT (ms)	30	2989	928	-1.76	29	.090	-0.320
R-RT (ms)	30	3260	926				
F-AR (%)	30	70.71	15	4.66		.000	0.851
R-AR (%)	30	54.17	6				
OvrM-AR(%)	30	62.44	11				

Notes: F=falling intonation; R=rising intonation; M=mean
RT=reaction time; AR=accuracy rate; OvrM-AR=overall Mean of accuracy rate.

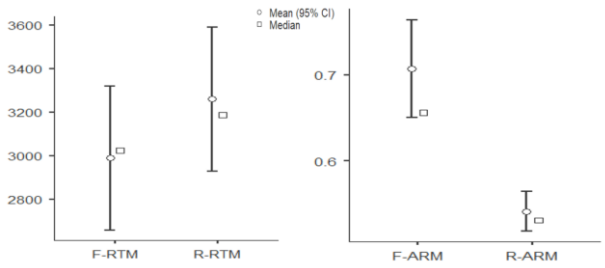


Figure 2:The AR & RT means of perception of target lexical stress read in falling (F) vs. rising (R) intonation.

As for the correct perception rates, the overall mean of correct perception rate is 62.44%. Specifically, the average perceptual AR in falling intonation is 70.71%, whereas that in rising intonation is 54.17%. The gap is 17.54%, and the AR differences are overwhelmingly significant ($F_{(29)} = 4.66$, P< .001, Effect Size = 0.851). This proves that, although the overall average AR range is fairly moderate, the participants' perception of target lexical stress embedded in falling intonation significantly outperformed that in rising intonation, and that there seem to be a falling-intonation bias in Mandarin speakers' distinguishing lexical stress in L2 English. Also, it could be assumed that the lower AR of perceived English lexical stress in rising intonation corroborates the longer RT during the perception.

3.2. Effects of stress position

Table 3 and Figure 3 show the data of participants' perception of target lexical stress located on the 1st vs. 2nd syllable for falls, rises, and for falls and rises combined respectively. The overall RTs to discriminate the target lexical stress located on both 1st and 2nd syllables are 3106ms and 3143ms respectively with a gap of 37ms, and the RTs to judge lexical stress in sub-categorical 1st vs. 2nd syllables read in falling and rising intonation contours are 2981ms vs. 2997ms embedded in falling intonation and 3230ms vs. 3290ms in rising intonation with gaps of 16ms and 60ms respectively. The overall and sub-categorical RT differences are not statistically significant, indicating that the time the participants employed in discriminating the lexical stress is not significantly different for stress either located on the 1st syllable or on the 2nd syllable. As for the correct perception rates,

Table 3 Paired-sample T-test results: The AR & RT means of the perception of target lexical stress located on the 1st vs. 2nd syllable.

	N	M	SD	T	df	P	Effect Size
S1F-RT (ms)	30	2981	1232	-0.082	29	0.935	-0.015
S2F-RT (ms)	30	2997	869				
S1R-RT (ms)	30	3230	886	-0.416	29	0.680	-0.076
S2R-RT (ms)	30	3290	1112				
S1-RT (ms)	30	3106	928	-0.451	29	0.655	-0.082
S2-RT (ms)	30	3143	776				
S1F-AR (%)	30	72.71	23	0.558	29	0.581	0.102
S2F-AR (%)	30	68.75	26				
S1R-AR (%)	30	42.29	32	-2.269	29	**0.031**	-0.414
S2R-AR (%)	30	66.04	26				
S1-AR(%)	30	57.50	11	-2.030	29	**0.042**	-0.371
S2-AR(%)	30	67.40	18				

Notes:S1F/S2F= stress located on the 1st / 2nd syllable with falling intonation
S1R/S2R= stress located on the 1st / 2nd syllable with rising intonation
RT=reaction time AR=accuracy rate M=mean `

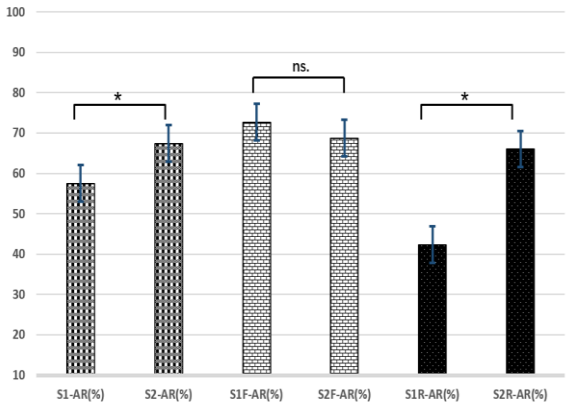


Figure 3: Paired-sample T-test results: The AR means of the perception of target lexical stress located on the 1st vs. 2nd syllable.

the average ARs of lexical stress located on the 1st vs. 2nd syllable are 57.50% and 67.40% respectively, and the gap between the two reaches significance ($F_{(29)}$ = -2.030, P= .042, Effective Size= -0.371), exhibiting a fairly certain iambic over trochaic advantage for Mandarin speakers when they do English word stress judgement. In addition, a further comparison between the perceptual ARs of word stress located on the 1st and 2nd syllable with falling intonation (72.71% & 68.75%) and those with rising intonation (42.29% & 66.04%) reveals that the AR differences in falling intonation is not significant ($F_{(29)}$ = 0.558, P= .581), whereas the difference in rising intonation is truly significant ($F_{(29)}$ = -2.269, P= .031). This not only provides robust evidence for the falling-intonation advantage in Mandarin speakers' doing English lexical stress judgement task, but also suggests there could be a potential interaction effect between intonation type and stress position in Mandarin speakers' perception of English lexical stress such that low pitch on a stressed first syllable of a disyllable causes problems.

3.3 Interaction of intonation type & stress position

Table 4: GLM repeated measures results of IT, SP and their interaction in the perception of target lexical stress.

Tests of Within-Subjects Effects						
Source	Type III Sum of Squares	df	Mean Square	F	P	η^2p
IT	37749	1	37749	643.651	.000	.936
Error(IT)	1759	29	58			
SP	1841.	1	1841	18.464	.000	.302
Error (SP)	2992	29	99			
IT * SP	2430	1	2430	25.553	.000	.460
Error (IT*SP)	2853	29	95			

Notes: IT=intonation type SP=stress position

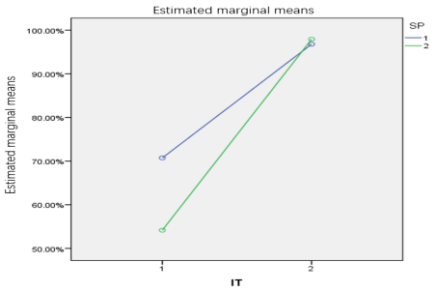


Figure 4:Within-Subjects Effects of IT & SP in the perception of target lexical stress.

To verify the possible interaction effects between intonation type (IT) and stress position (SP) in participants' lexical stress perception, GLM repeated measures were conducted to test the within-subjects effects of 2 types of IT (falling vs. rising) and 2 patterns of stress position (on the 1st vs. 2nd syllable) on the perception of target English lexical stress.

As displayed in Table 4 and Figure 4, both intonation type and stress position display significant main effects on the participants' perception of target lexical stress (IT: $F_{(1,29)}$ = 643.65, P=.000, η^2p= .936; SP: $F_{(1, 29)}$ =18.464, P= .001, η^2p= .302), and the effect of intonation pattern is more robust than that of stress position according to their effect sizes. This could be considered as confirmative evidence for the effective roles of both factors in Mandarin speakers' perception of English lexical stress. Moreover, there is a genuinely significant interaction effect between intonational type and stress position ($F_{(1, 29)}$ =25.553, P= .000, η^2p= .460), demonstrating a complex interactive network existing in native Mandarin speakers' perception of English lexical stress.

4. DISCUSSION

The implications of these findings are several. First, the overall perception AR of English lexical stress embedded in 16 sets of English heteronyms (OvrM-AR=62.44%) proves that native Mandarin speakers, whose L1 does not have lexical stress, can encode L2 English stress phonologically and use it for word recognition to a great extent. This encouraging outcome could be accounted for partly by the participants' near-advanced overall English proficiency and partly by their exposure to one term of systematic instruction on English pronunciation, which helped develop their theoretical knowledge and practical competence in perceiving and producing English lexical stress. Secondly, when judging L2 English lexical stress in falling vs. rising intonation, the participants did significantly better with falling intonation than with rising intonation. Such a falling-intonation bias demonstrates the influence of L1 falling intonation in L2 lexical stress perception. Thirdly, when judging L2 English lexical stress located on the 1st vs. 2nd syllable, the participants' performance did not match native speakers' clear bias towards the initial syllable. Instead, they did better both in the 1st syllable with falling intonation and in the 2nd syllable with rising intonation, indicating native speakers' stress bias is mediated by L2 learners' suprasegmental properties of their L1 word phonology, such as lexical tones in Mandarin. Specifically, given that the learners' weakest performance occurred when stress position was on a low-pitch first syllable followed by a high-pitch terminal (S1R), Mandarin EFL learners are doomed to find it hard to associate stress with a syllable lacking high pitch and to adapt to their perception by seeking help from other parameters, such as duration and intensity. Given the predominance of falling accents in English, an association of higher pitch with stress is a reasonable generalisation and a lack of high pitch on the stressed first syllable of a disyllable causes problems for Mandarin EFL learners. That's exactly why their S1R-AR% was so strikingly low.

5. REFERENCES

[1].Beckman, M. E. 1986. *Stress and non-stress accent*. Dordrecht: Foris.

[2]. Fry, D. B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of Acoustic Society of America, 27*(4), 155-158.

[3] Fry, D. B. 1958. Experiments in the perception of stress. *Language and Speech, 1*, 126- 152.

[4] Ladefoged, P. 2016. *A Course in Phonetics*. Boston, MA: Thomson/Wadsworth.

[5] Lehiste, I., and Fox, R. A. 1992. Perception of prominence by Estonian and English listeners. *Language and Speech, 35*(4), 419-434.

[6] Lehiste, I. 1970. *Suprasegmentals*. Cambridge: MA: MIT Press.

[7] Sluijter, A. M.C., & van Heuven, V. J. 1996a. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4), 2471–2485.

[8] Sluijter, A. M. C., & van Heuven, V. J. 1996b. Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings of the fourth international conference on spoken language processing* 2, Philadelphia, 630–633.

[9] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25, 61–83.

[10] Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, 55, 179–203.

[11] Cutler, A.2005. Lexical tress. In: *The Handbook of Speech Perception,* D. B. Pisoni and R. E. Remez (eds.), 246-289, Orford: Blackwell.

[12] Cutler, A. 2015. Lexical stress in English pronunciation. In: *The Handbook of English Pronunciation.* R. Marnie and J. M. Levis (eds.),106-124, Oxford: Blackwell.

[13] Duanmu, S. 2000. *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

[14] Cooper, N., Cutler, A., and Wales, R. 2002. Constraints of lexical stress on lexical access in English: evidence from native and non-native listeners. *Language and Speech*, 45: 207-228.

[15] Slowiaczek, L. M. 1990. Effects of lexical stress in auditory word recognition. *Language and Speech* 33: 47-68.

[16] Slowiaczek, L. M.1991. Stress and context in auditory word recognition. *Journal of Psycholinguistic Research 20*: 465-481.

[17] Fear, B. D., Cutler, A., and Butterfield, S. 1995. The strong/weak syllable distinction in English. *Journal of Acoustical Society of America*, 97, 1893-1904.

[18] Archibald, J. 1997. The acquisition of English stress by speakers of non-accentual languages: lexical storage versus computation of stress. *Linguistics*, 35, 167-181.

[19] Dupoux, E., Pallier, C., Sebastian, N., and Mehler, J. 1997. A destressing ‘‘deafness’’ in French?. *Journal of Memory and Language*, 36, 406–421.

[20] Lee, G., Shin, D, and Garcia, M.T.M., 2019. Perception of lexical stress and sentence focus by Korean-speaking and Spanish-speaking L2 learners of English. *Language Sciences*, 72: 36-49.

[21] Zhang, Y., Nissen, S. L., and Francis, A. L. 2008. Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *Journal of the Acoustical Society of America,* 123 (6), 4498-4513.

[22] Peperkamp, S.A., 2004. Lexical exceptions in stress systems: arguments from early language acquisition and adult speech perception. *Language*, 80 (1), 98– 126.

[23] Peperkamp, S., Dupoux, E., 2002. A typological study of stress ‘deafness’, *Laboratory phonology*, 7, 203–240.

[24] Zhang, Y., and Francis, A. 2010. The weighting of vowel quality in native and non-native listeners’ perception of English lexical stress. *Journal of Phonetics*, 38 (2), 260– 271.

[25] Lado, R. 1957. *Linguistic Across Cultures*, Ann Arbor: University of Michgan Press.

THE EFFECT OF LEARNING CONTEXT ON MANDARIN LISTENERS' PERCEPTION OF ENGLISH VOWELS

Jiayi Lu[1] & Mitsuhiko Ota[2]

[1]Hong Kong Polytechnic University, [2]University of Edinburgh
jiayilu1998@163.com, mota@ed.ac.uk

ABSTRACT

Do developmental patterns in second language speech perception differ depending on whether the learning takes place in a predominantly naturalistic setting (second language acquisition [SLA]) or an instructed setting (foreign language acquisition [FLA])? We addressed this question by testing the predictions of the extended Perceptual Assimilation Model (PAM-L2) in both settings. Native adult speakers of Mandarin learning English in the UK (SLA learners; N=15) and in China (FLA learners; N=15) performed an ABX discrimination task on 15 pairs of British English vowels. The vowel pairs were classified into six PAM assimilation categories based on goodness-of-fit ratings between English and Mandarin vowels obtained from Mandarin speakers with little exposure to English (N=15). The order of discrimination accuracy for PAM categories did not differ between the two groups (i.e., SLA and FLA), demonstrating that learning context does not affect the relative difficulty with which adult learners perceive non-native vowels.

Keywords: vowel perception, second/foreign language learning, Perceptual Assimilation Model.

1. INTRODUCTION

1.1. Background

Attainment levels of second language (L2) speech are known to differ depending on whether the learning takes place in a predominantly naturalistic context (second language acquisition [SLA]), where the target language is the language of the community, or an instructed context (foreign language acquisition [FLA]), where the target language is not the community language [1, 2]. Such differences in attainment levels are usually attributed to the limited quantity and quality of the input FLA learners receive, their lack of opportunities to practice the L2, and the formal nature of the instruction they receive.

A less well understood question is whether different learning contexts, such as SLA and FLA, also lead to different developmental patterns in speech perception. More specifically, does the relative difficulty with which learners perceive L2 sounds depend on their contexts of learning? This issue has been raised by Best and Taylor [1] in their adaptation of the Perceptual Assimilation Model to the L2 context (PAM-L2). According to PAM-L2, a pair of L2 sounds that are assimilated to separate native categories (two-category [TC] assimilation) is easier to discriminate than a pair of L2 sounds that are assimilated to a single native category with different degrees of goodness-of-fit (category-goodness [CG]). A CG pair, in turn, is easier to discriminate than a pair of L2 sounds that are perceived to be equally good exemplars of a single native category (single-category [SC]).

These predictions have been confirmed separately in SLA [3, 4] and FLA [2, 5, 6, 7]. However, few studies have systematically tested the predictions of PAM's discrimination hierarchy in both SLA and FLA within a single study, nor have studies tested the perceptual accuracies of assimilation pairs other than TC, CG, and SC in these contexts.

Moreover, two studies investigating vowel perception in FLA have yielded results inconsistent with PAM. In Sun and van Heuven [8], Mandarin speakers learning English in an FLA context showed very good discrimination of an SC contrast (/æ-ɑː/), against PAM's prediction. Lai [9], who also studied Mandarin FLA learners of English, found that a vowel pair (/æ-ɛ/) classified as two uncategorized L2 sounds (uncategorized-uncategorized [UU]) was better discriminated than six other pairs classified as one uncategorizable and one native exemplar (uncategorized-categorized [UC]), despite PAM's prediction that UC should show better discrimination than UU. These results can be interpreted in different ways. It is possible that Mandarin learners of English in an FLA context do not follow PAM, thus presenting counter-evidence to the applicability of the model to FLA. Lai's [9] results can be partly interpreted as evidence that PAM makes incorrect predictions for the relative difficulties of UC and UU. Finally, we need to consider the possibility that these discrepancies are due to methodological idiosyncrasies in the two studies [8, 9].

The purpose of the current study is to directly compare the relative difficulties with which Mandarin learners in FLA and SLA contexts perceive English vowels belonging to different PAM assimilation types. In order to test a wide range of assimilation

types, we include vowel pairs that fall into the categories of UC and UU in addition to those belonging to TC, CG, and SC. These classifications were carried out using a goodness-of-fit task with naïve Mandarin-speaking listeners of English. Furthermore, we tested both SLA and FLA learners using the same methods. If L2 perception develops in a similar way regardless of the learning contexts, the same order of perceptual difficulty should be observed in both SLA and FLA groups.

2. EXPERIMENT I. GOODNESS-OF-FIT RATING

2.1. Participants and stimuli

The participants in Experiment 1 were 15 naïve Mandarin speakers living in China (age range = 20–25 years, mean age = 23.4 years, 7 females). They were all Mandarin monolinguals with limited exposure to English.

The perceptual stimuli were the following British English vowels: /ɪ, e, æ, ɒ, ʌ, ʊ, ɑː, ɜː, iː, ɔː, uː, ə, aɪ, eɪ, əʊ, aʊ, ɔɪ, ɪə, eə, ʊə/ [10]. The vowels were recorded by three female native speakers of Southern Standard British English (mean age = 20.7 years) in a /h_t/ context. The stimuli were denoised and normalized to an amplitude of 70 dB, after which the vowel portions were extracted as the final stimuli.

2.2. Procedure

The experiment was conducted online using Qualtrics. All participants were asked to complete the test alone in a quiet room using a laptop and wearing headphones with sound levels adjusted to a comfortable level. They also completed a language background questionnaire after the rating task.

In the rating task, the participants were asked to listen to each English vowel and choose the Mandarin vowel that they thought was closest to the English vowel. The Mandarin vowels provided were the following: /i, ə, ɚ, ɤ, a, ɔ, u, y, ai, au, ei, ie, ia, ou, ua, uo, ye, uei/ [11, 12, 13]. The option "none of the above" was also provided. After they selected a Mandarin vowel, they were asked to rate its goodness-of-fit with the English vowel on a scale of 1 (not similar) to 7 (very similar).

2.3. Results

For each English vowel, the proportion of the Mandarin vowel that was most commonly selected (modal assimilation proportion) and its goodness-of-fit rating were multiplied to calculate the fit index (FI). Table 1 presents the modal assimilation proportion, goodness-of-fit score, and FI of each English vowel, shown in descending order of the FI within each assimilation set. An assimilation set is defined as a collection of L2–L1 pairs in which two or more L2 sounds are assimilated into a single L1 sound category [8]. For example, the first four L2–L1 pairs in Table 1 involve different English vowels that are all assimilated into the same Mandarin vowel (/ei/). Vowels that do not belong to any assimilation set are listed at the end of the table in descending order according to the FI. The English vowels are divided into different categories based on the FI using the standard deviation (SD) as the criterion. English vowels with an FI more than 1 SD (1.0) above the mean FI (2.5) were classified as "good" instances of a Mandarin category. Those with an FI between the mean and 1 SD above the mean were classified as "fair" instances of a Mandarin category. Those with an FI between the mean and 1 SD below the mean had a "poor" fit. Finally, those with an FI lower than 1 SD below the mean were considered "uncategorized."

English-Mandarin vowel assimilation	Modal proportion	Rating score	Fit index	
eɪ-ei	0.93	4.5	4.2	Good
e-ei	0.73	4.4	3.2	Fair
eə-ei	0.67	4.4	3.0	Fair
ɪ-ei	0.47	4.3	2.0	Poor
æ-a	0.93	4.5	4.2	Good
ɑː-a	0.93	3.1	2.9	Fair
ʌ-a	0.47	3.3	1.6	Poor
ə-ɤ	0.60	3.0	1.8	Poor
ɜː-ɤ	0.47	3.3	1.5	Poor
ʊə-ɤ	0.47	3.3	1.5	Poor
ɒ-ɔ	1.00	3.1	3.1	Fair
ʊ-ɔ	0.27	2.8	0.8	Uncategorized
iː-i	1.00	3.1	3.1	Fair
aɪ-ai	1.00	3.1	3.1	Fair
uː-y	0.93	3.2	3.0	Fair
aʊ-au	0.93	3.2	3.0	Fair
ɪə-ie	0.87	3.2	2.8	Fair
əʊ-ou	0.67	3.1	2.1	Poor
ɔː-uo	0.40	3.5	1.4	Uncategorized
ɔɪ-uei	0.40	2.8	1.1	Uncategorized

Table 1: Assimilation sets and fit indexes (FIs) for modal assimilated English–Mandarin vowel pairs.

Based on these, we selected the English contrast /eɪ-æ/ as a case of TC contrast, /æ-ɑː/, /æ-ʌ/, and /ɑː-ʌ/ as CG contrasts, and /e-eə/ and /ɜː-ʊə/ as SC contrasts. The pair /ɔː-ɒ/ involves one uncategorized sound and one categorized sound but, as the latter elicited an L1 response that overlaps with another uncategorized sound (/ɔ/), it was classified as UC+OL (UC with an overlap). Other UC pairs, i.e., /ɔː-eɪ/, /ɔː-iː/, /ɔː-ə/, /ɔɪ-eɪ/, and /ʊ-ʌ/, were classified as UC-OL (UC without an overlap). Contrasts involving two uncategorized sounds, i.e., /ʊ-ɔɪ/, /ɔː-ʊ/, and /ɔː-ɔɪ/, were classified as UU-OL (UU without an overlap).

3. EXPERIMENT II. AXB DISCRIMINATION TASK

3.1. Participants and stimuli

The participants in Experiment 2 were 15 FLA Mandarin speakers (age range = 19–23 years, mean age = 20.8, 7 females) and 15 SLA Mandarin speakers (age range = 21–23 years, mean age = 22.5, 8 females). All FLA participants were non-English majors, and none had any experience of going abroad or receiving English lessons from a native English speaker. Both FLA and SLA participants in this study had studied English in the classroom for several years (FLA: mean = 12.2 years, SD = 2.8; SLA: mean = 12.6 years, SD = 1.6); in addition, the SLA group had attended a university in the UK for an average of 2.8 years (SD = 0.8). The FLA learners' mean exposure to English in the classroom was 3.1 hours per week, with little to no exposure to English outside the classroom. The SLA learners' mean English exposure in the classroom was 22.5 hours per week, with their exposure outside the classroom approximately half of what they received in class, although with large individual differences (mean = 11.8 hours/week, SD = 11.4).

The stimuli were the vowels recorded in Experiment 1 in the context of [h_t] denoised and normalized. Each AXB trial contained three vowels produced by three different speakers. The A, X, and B components were concatenated with a 1.5-second inter-stimulus interval.

3.2. Procedure

The participants were instructed to decide whether the first or third sound was more similar to the middle sound in each trial. They were told that different speakers produced the three sounds in each trial and that they needed to ignore the differences in the speaker's voice. There were 120 trials in total (15 pairs * 8 trials). The fifteen vowel pairs were presented in a random order across subjects.

3.3. Results

Figure 1 shows the discrimination results of the 15 English vowel pairs in the two learning context groups.

A mixed two-way ANOVA was conducted with group (two levels) as a between-subjects factor and vowel pair (15 levels) as a within-subjects factor. The results show a main effect of vowel pair, $F(14, 392) = 44.116$ ($p < 0.001$), but no main effect of group or an interaction.

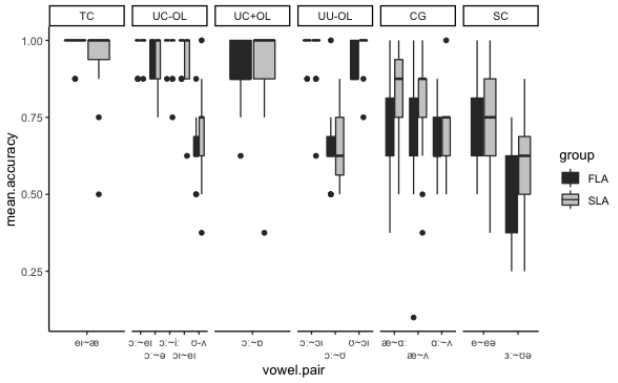


Figure 1: Discrimination accuracy of the 15 British English vowel pairs by SLA learners (N = 15) and FLA learners (N = 15).

A Tukey multiple pairwise-comparisons test was conducted to identify the sources of the vowel pair effect. The results are summarized in Figure 2.

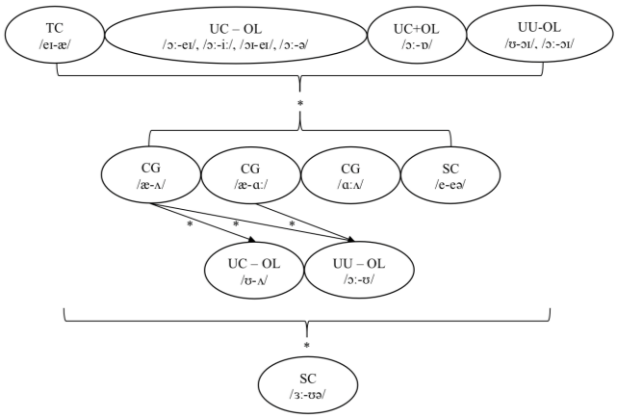


Figure 2: Summary of Tukey test results (* $p < 0.05$).

Discrimination levels were significantly higher for the TC pair /eɪ-æ/, most UC–OL pairs (/ɔː-eɪ/, /ɔː-iː/, /ɔɪ-eɪ/, /ɔː-ə/), the UC+OL pair /ɔː-ɒ/, and most UU–OL pairs (/ʊ-ɔɪ/, /ɔː-ɔɪ/) than all CG pairs (/æ-ɑː/, /æ-ʌ/, /ɑː-ʌ/), all SC pairs (/e-eə/, /ɜː-ʊə/), the UU–OL pair /ɔː-ʊ/, and the UC–OL pair /ʊ-ʌ/. Among the three CG pairs, only /æ-ʌ/ was significantly higher than the UC–OL pair /ʊ-ʌ/, and only /æ-ʌ/ and /æ-ɑː/ were significantly higher than the UU–OL pair /ɔː-ʊ/. The three CG pairs were all significantly higher than the SC pair /ɜː-ʊə/. Finally, the UU–OL pair /ɔː-ʊ/ and the UC–OL pair /ʊ-ʌ/ were significantly higher than the SC pair /e-eə/.

4. DISCUSSION

This study systematically examined the relative difficulties with which Mandarin listeners perceive English vowel contrasts depending on whether they are exposed to English in a predominately naturalistic context (SLA) or an instructed context (FLA). Based

on their discrimination of vowel pairs belonging to a range of PAM assimilation categories, no difference between the SLA and FLA groups emerged in our results, indicating that L2 learners follow similar developmental patterns in speech perception regardless of their learning contexts.

Our results also show that the majority of PAM-L2 predictions were borne out in both learning contexts. Pairs classified as TC, CG, or SC contrasts largely followed the predicted discrimination accuracy order of TC > CG > SC. Thus, the core of PAM-L2 applies not only to SLA but also to FLA. Note here that the contrast /æ-ɑː/, which was treated as a case of SC in Sun and van Heuven [8], was classified as CG in our study based on the outcome of Experiment 1. In [8], the accuracy level of that contrast was seen to be too high for an SC contrast, but this apparent counter-evidence to PAM-L2 is likely to be due to misclassification of /æ-ɑː/ for Mandarin listeners.

Not all PAM-L2 predictions were confirmed, however. Against the model, and as in Lai [9], two UU contrasts (/ʊ-ɔɪ/, /ɔː-ɔɪ/) were discriminated better than a UC contrast (/ʊ-ʌ/). Another point of discrepancy can be found in the UC contrasts with overlap (UC+OL) and without overlap (UC-OL). Recent work [14] proposes that UC+OL should have poorer discrimination than UC-OL. However, in the current study, the discrimination score for the UC+OL contrast /ɔː-ɒ/ was better than the UC-OL contrast /ʊ-ʌ/. If these findings are found to be generalizable to other L1–L2 pairs, they force a re-examination of the UC and UU categories in PAM-L2.

Finally, we note that we found no between-group difference in the overall discrimination accuracy of the target English contrasts. This is somewhat surprising considering that the SLA participants had lived in a target language environment for nearly three years on average, and we would, therefore, expect them to exhibit higher levels of attainment in their speech learning than the FLA participants. This suggests that the SLA learners in this study had not yet seen the benefits of an immersive linguistic experience (perhaps due to lower-than-expected use of English with native speakers); this also raises the possibility that SLA learners with longer or more intensive naturalistic target language exposure will exhibit a speech perception profile that is qualitatively different from that of FLA learners. Further work is required to determine long-term effects of learning contexts on L2 speech perception patterns.

5. REFERENCES

[1] Best, C.T., Tyler, M.D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities, In: O.-S. Bohn & M. J. Munro (eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege.* Amsterdam: John Benjamin, 13-34.

[2] Grimaldi, M., Sisinni, B., Gili Fivela, B., Invitto, S., Resta, D., Alku, P., Brattico, E. 2014. Assimilation of L2 vowels to L1 phonemes governs L2 learning in adulthood: a behavioral and ERP study. *Frontiers in Human Neuroscience*, 8.

[3] Bundgaard-Nielsen, R. L., Best, C. T., Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, 33(3), 433-461.

[4] Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J. C. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*, 107(5), 2711–2724.

[5] Ghaffarvand-Mokari, P., Werner, S. 2017. Perceptual assimilation predicts acquisition of foreign language sounds: The case of Azerbaijani learners' production and perception of Standard Southern British English vowels. *Lingua*, 185, 81–95.

[6] Lengeris, A., Hazan, V. (2007). Cross-language perceptual assimilation and discrimination of southern British English vowels by Greek and Japanese learners of English. *Proc. 16th ICPhS Saarbrücken*, 1641-1644.

[7] Rallo Fabra, L., Romero, J. 2012. Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, 40(3), 491–508.

[8] Sun, L., van Heuven, V. J. 2007. Perceptual assimilation of English vowels by Chinese listeners. *Linguistics in the Netherlands*, 24, 150–161.

[9] Lai, Y. H. 2010. English vowel discrimination and assimilation by Chinese-speaking learners of English. *Concentric: Studies in Linguistics*, 36(2), 157-182.

[10] Ogden, R. 2017. *Introduction to English Phonetics.* Edinburgh university press.

[11] Jia, G., Strange, W., Wu, Y., Collado, J., Guan, Q. 2006. Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, 119(2), 1118.

[12] Lee, W. S., Zee, E. 2003. Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1), 109-112.

[13] Wan, I. P., Jaeger, J. J. 2003. The phonological representation of Taiwan Mandarin vowels: A psycholinguistic study. *Journal of East Asian Linguistics*, *12*(3), 205-257.

[14] Faris, M. M., Best, C. T., Tyler, M. D. 2016. An examination of the different ways that non- native phones may be perceptually assimilated as uncategorized. *The Journal of the Acoustical Society of America*, 139(1), EL1–EL5.

EXPLORING SECOND LANGUAGE ATTITUDINAL PROSODY WITH A MACHINE LEARNING APPROACH

Xiaolin Xu, Izabelle Grenon

University of Tokyo
shelynxu@gmail.com

ABSTRACT

This research explored the language universal and specific effects in the second language (L2) prosody. Our multivariate analysis covers over 20 prosodic features from phonetics (e.g. F0) and phonology(e.g. pitch pattern) aspects. The importance of features was defined by a linear Support Vector Machine (SVM) model performing classification on both speakers' groups and attitudes. Speech data were recorded from Japanese speakers (n=34) and English speakers (n=16) having semi-spontaneous question-reply conversations with neutral, strict, and gentle attitudes. Our model successfully classified the gentle attitude from neutral as well as the two speaker groups. The feature weights indicate a significant role of phonetic features with attitudinal prosody. Japanese speakers with higher English proficiency tend to change sentence-level F0 and intensity with a gentle attitude similar to the English natives. While the lower English proficiency group prefers syllable-level features which is the same as the case in L1 Japanese.

Keywords: second language acquisition, attitudinal prosody, SVM,L1 transfer, pitch

1. INTRODUCTION

Human beings have developed the ability to produce various prosody and intonation patterns throughout evolution, with which we can directly express pragmatic information such as attitudes (e.g. strict, friendly) and emotions (e.g. comfort, distress)[1]. This crucial ability enriched our social interactions and facilitates close relationships between individuals[2]. With the promotion of globalization and cultural exchange, cross-language pragmatics receives more attention. Research on prosody cross-language not only helps us to avoid misunderstandings but also deepens our knowledge of language acquisition, and supports automatic speech modeling technologies development.

1.1. language-universal aspect of prosody

To a large extent, prosodic features are considered universal across different languages. For instance, people can distinguish a question from a statement (sentence mood) when listening to languages they have no proficiency in [3]. The theory of the Biological Codes proposed by Gussenhoven in 2004 with the previous work of [4] indicates that different languages, even of separate typologies, share unified functions of reflecting attitudes. Within this framework, the low-level fundamental frequency (F0) and the falling-shaped pitch contour are associated with an unfriendly attitude while the high pitch level and rising contour tend to be produced when conveying a friendly attitude. It is supported by previous research on many languages including English and Japanese [5].

1.2. language-specific differences of prosody

On the other hand, the specific differences between the two languages are supposed to cause difficulty in L2 prosody acquisition. Previous studies reported the first language (L1) transfer due to language-specific differences. In the Intonation learning theory (LILt) [6], the differences can be categorized as phonological, and phonetic dimensions. Phonologically, when a particular pitch pattern of L2 doesn't exist in L1, the learners tend not to learn that L2 pitch pattern [7][8][9]. Phonetically, F0 peaks produced by learners are either higher or lower than those produced by native speakers; Learner's steepness of pitch contour is either greater or smaller than native speaker's [10][11].

1.3. machine learning (ML) models

Models such as attention-long Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Hidden Markov model(HMM) and Support Vector Machines (SVM) were widely adopted on robotic speech emotion recognization[12][13], automatic L2 English proficiency evaluation [14]

with Test of English as a Foreign Language (TOEFL) database and Clinical detection[15] of Depression and Autism [16]with utterance level acoustic features. Most of the models reached a high accuracy with a full aspect of features covering intensity, F0, speech rate reflected by duration, voice quality, and zero crossing rate features.

1.4. The significance of this study

Most of the linguistic studies on prosodic features focus on the syllable level in order to interpret each specific feature's effect in a clear domain. Multiple phonetic and phonological features have been proven significant in L2 prosody acquisition in separate research but difficult to rank which feature matters the most in each case. In contrast, studies with machine learning models are able to make accurate predictions with a massive amount of features at once. But they focus more on accuracy increase in dealing with big-size utterance corpus. As a drawback, background factors like scenarios could not be controlled without an experimental design. In addition, it is difficult to interpret each feature's effect with black box models. ML models on speech emotion detection are mostly trained on monolingual data, and models involving L2 barely cover pragmatic functions of conveying emotions or attitudes. Particularly noteworthy is the study of cross-language prosody of Autism[16] pointing out the language-specific prosody differences in Autism classification. It found that classifications using F0 values were significant for English, but not for Cantonese speakers, which is potentially due to long exposure to their native tone language. Regarding the comparison of previous studies above, it's significant for our research to bridge the gap between L2 prosody research and acoustic ML model development territories. To achieve this goal, we classify attitudes and speaker groups, utilizing an ML approach on sentence and syllable-level prosodic features from a controlled experimental design. We focus on two languages of separate typologies, English and Japanese, and particularly choose a linear SVM model. The SVM models perform well and were frequently adopted by classification tasks with acoustic features [16][14], and the linear SVM is rather transparent (c.f. black box) since it outputs a weight interpreting the importance of each feature[17].

1.5. Research question

Our research questions can be narrowed down to:

- What are the key important prosodic features when conveying different attitudes?
- Do these important features vary when produced with a second language?
- If the features varies, could it be explained by a language-specific difference (L1 transfer)?

For the record, the feature's importance is defined at the ML model level, which should be treated differently from the human perception level.

2. HYPOTHESIS AND METHOD

2.1. Hypothesis

English is a stressed-time language while Japanese is a pitch-accent language where the specific pitch patterns are decided by a particular lexical word. The two languages differ both phonologically and phonetically: Phonologically, English interrogative utterance typically has a rising pitch pattern over the focused word with the Yes/No question, and a falling pitch pattern with the Wh question. Japanese doesn't have such variation. When expressing an attitude, English speakers use the pitch pattern over the focused word's syllable. In Japanese, pitch pattern over the final syllable (Boundary pitch movement, BPM) of utterance functions to convey attitudes[18]. Phonetically the Japanese overall pitch level is higher than in English[19]. According to [11], the English interrogative rising pitch contour produced by Japanese learners has a steeper slope than native speakers. Additionally, Japanese has a higher speech rate generally[20]. 1. Based on the language-universal Biological code framework, we first expect that pitch features (measured by F0 values of focused syllable and utterance domains) are important when conveying positive and negative attitudes in both L1 English and L1Japanese. 2. L1 Japanese will have a higher pitch level and a steeper slope than L1 English with a higher speech rate (measured by focused syllable duration, sentence duration, and duration-per-syllable over utterance) 3. When classifying Japanese L2 English from native speakers' L1 English, the language-specific difference of hypothesis 2 should be more important than the language-universal features of hypothesis 1.

2.2. Participants and procedure

34 participants speak in Tokyo Japanese accent. Their English proficiency ranges from Elementary (n=12), Intermediate(n=15) and Advanced (n=6).16 English speakers are native to North American English. They were traveling in Tokyo at the time of data collection. All participants produced 10 Yes/No and 10 Wh questions, respectively in

Neutral, Strict, and Gentle attitudes from slides. Pictures of faces of corresponding attitudes were also shown to speakers to better trigger attitudes. The focus word is underlined on slides (half in the middle of the sentence and half at the end). To simulate natural conversations, utterances were produced by finishing semi-spontaneous production tasks, in which the participants asked each question as having a real conversation with the researcher after memorizing a short question from the slide and moving to the next question after having a short answer. Japanese speakers asked the same questions translated into Japanese in the same procedure.

2.3. Features extraction

The phonological pitch pattern is annotated as rising (R), falling (F), level (L), and combinations of them (e.g.RF stands for rising-falling pattern) within focus syllable using the original python script. Then we visually re-check the annotation on Praat.Phonetic features of F0 and intensity's max, min, range, and mean values were extracted from Praat on syllable and utterance levels. The range, slope, and deviation values were also calculated as subfeatures. We also recorded the duration, intensity, jitter, and zero crossing rate values and input them into the following model together with focus word position information (middle or final in a sentence.)

2.4. SVM classification

As the figure of the modeling process shows, three types of classifications were conducted to explore how language-universal and specific prosodic features work together on L2 production. We first classify utterances under Gentle and Strict from Neutral ones with L1 Japanese and L1 English together and separately, to see the common prosodic cues used in English and Japanese with and without an attitude.Secondly, we seek the specific features that distinguish the two languages with and without an attitude.Finally, speaker groups of English/Japanese, as well as Japanese speakers of three English proficiencies were classified with L1 + L2 English speech data together. The important features extracted based on SVM weights were compared with other important features from the first two classification types. Outliers were filtered by rescale feature data within the 1st quartile and the 3rd quartile (75th quantile) with RobustScalar of the Sklearn library, then standardized with Sklearn's StandardScaler function before input into mode. Since feature data consisted of continuous data with different units and scales (e.g. F0, intensity, duration), as well as binary data of focus position sentence mood, and nominal data of pitch patterns. The important feature was extracted from classification models that were evaluated as effective based on 10-fold cross-validation.

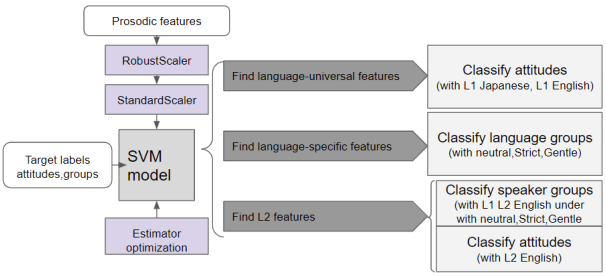


Figure 1: The SVM classification process.

3. RESULTS

The following chart shows the top-weighted features of each effective (Accuracy> 0.6, accuracy - null accuracy >0.1, cross-validation AUC >0.6) classification case. These values are the average from 1000 times ML sampling iterations.

Classification types	Data scope	Top weighted features
Attitude Gentle/ Neutral	L1 Japanese	Nmax_F0 Nmean_F0 Nduration
	L1 English	Smean_F0 Smean_intensity Sdevi_F0
	L2 English Advanced	Sdevi_F0 Smax_intensity Smean_F0
	L2 English Elementary	Nmean_F0 Nmin_F0 Sdevi_pitch
Languages	Neutral	Smean_intensity Nmean_intensity Nunvoiced
Speaker groups L1/L2 English	Neutral	Smean_F0 Sdevi_F0 Nmax_F0

Figure 2: SVM models chart of means values of 1000 times itterations

The classification of languages (L1 Japanese from L1 English) is with the highest accuracy of over 0.9. The results indicate the intensity feature on both

nucleus syllable (Nmean intensity) and sentence levels (Smean intensity) is the most important prosodic feature to distinguish the two languages with the SVM model. Attitudes classification was only on Gentle and Neutral utterances only. English speakers' sentence-level F0 and intensity features (Smean F0, Smean intensity, Smax intensity) seem to be more dominant. In L1 Japanese, on the other hand, nucleus syllable level F0 and speech rate features (Nmax F0, Nmean F0, Nduration) are most important, for instance when conveying a Gentle attitude. They tend to produce longer duration and lower speech rate. The phonological features also contribute to the model, Gentle utterances were produced with more rising nucleus patterns (PitchPattern R) in English as well as more rising BPM (Tone R) in Japanese after ruling out the unvoiced data. But still, the pitch pattern features are not as highly weighted as phonetic features. This set of results also presents the primary features of Japanese speakers' L2 English compared with native English speakers and L1 Japanese. L2 English of all three proficiency groups doesn't have an effective classification though the Advanced and Elementary English level groups do. The two proficiency groups have opposite tendencies. Japanese speakers with higher English proficiency tend to change sentence-level features (Sdevi F0, Smax intensity, Smean F0) with a gentle attitude similar to the English natives. While the lower English proficiency group prefers syllable-level features which is the same as the case in L1 Japanese. Finally, the SVM can classify L1 and L2 English groups with both sentence and syllable level F0 features (Smean F0, Sdevi F0, Nmax F0), which is consistent with the previous study that Japanese speakers' overall pitch level is higher than English. Our results show clear language-universal use of higher pitch levels associated with a positive attitude and language-specific differences in syllable and sentence domains.

4. DISCUSSION

Our research re-examined the language-universal and specific differences effects on attitudinal prosody. Overall, partially as the Biological code has summarized, both Japanese and English adopt higher F0 with positive attitudes. It can be utilized by ML models. But we also find that F0 ques in the nucleus syllable domain work better with Japanese while the sentence domain features are more primary in English. And Japanese speakers are able to learn much differently when proceeding

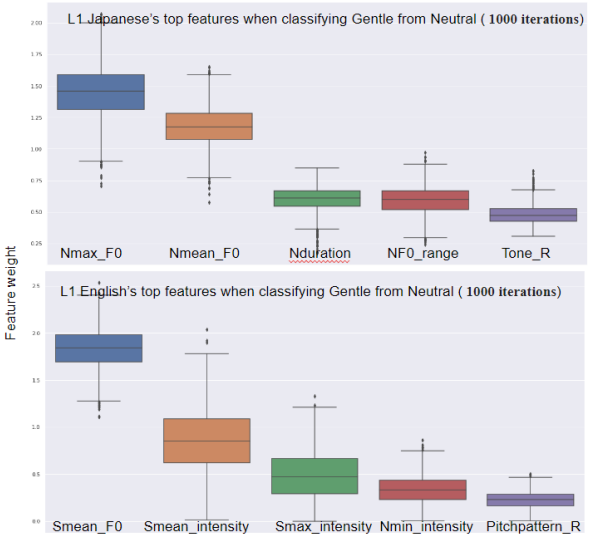


Figure 3: Different primary prosodic features of English and Japanese.

to a higher level. These findings not only help ML models recognize human speech better but also provide an interpretation from an L1 transfer perspective. One question is why the negative attitude was not effectively classified.[21] probably shows us the answer that negative emotions like anger can be divided into hot anger and cold anger. When someone is suppressing their anger, the pitch and intensity will go into low-level, so-called cold anger. The explosive type of anger should be the opposite. Therefore it was difficult to classify the two subcategories mixed together. Another problem with our research is gender unbalanced data. We find that when restricted to female, the model accuracy tends to be higher. Gender should be considered an important factor in future studies. More for the futural direction, it would be meaningful to investigate the difference between the ML model and human speech perception, which features that matter might be quite different. And mental disorder diagnosis using prosody may be conducted with a different standard with blingos regarding the cross-language differences found in our research.

5. CONCLUSION

Our research shows that the phonetic features of F0 are significant during the ML classification, but still, they need to work together with intensity speech rate and phonological pitch pattern to achieve the best predictions. When classifying second language prosody, the effect from L1 can not be ignored. The syllable level features are more salient in the case of Japanese speakers producing L2 English.

6. REFERENCES

[1] E. Altenmuller, S. Schmidt, and E. Zimmermann, *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man*. OUP Oxford, 2013.

[2] J. K. Hall, J. Hellermann, and S. P. Doehler, *L2 interactional competence and development*. Multilingual Matters, 2011, vol. 56.

[3] C. Gussenhoven and A. Chen, "Universal and language-specific effects in the perception of question intonation," in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. 91–94.

[4] J. J. Ohala, "Cross-language use of pitch: an ethological view," *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983.

[5] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge University Press, 2004.

[6] I. Mennen, "Beyond segments: Towards a l2 intonation learning theory," in *Prosody and language in contact*. Springer, 2015, pp. 171–188.

[7] E. Grabe and M. Karpinski, "Universal and language-specific aspects of intonation in english and polish," in *Proceedings of the 15th International Congress of Phonetic Sciences*, vol. 39, 2003.

[8] A. Chen, *Universal and language-specific perception of paralinguistic intonational meaning*. Utrecht: LOT, 2005.

[9] M. Zięba, *The acquisition of English intonation by Polish adult learners*. Wydawnictwo Naukowe Państwowej Wyższej Szkoły Zawodowej w Nowym Sączu, 2013.

[10] M. Jilka, "The contribution of intonation to the perception of foreign accent," Ph.D. dissertation, Universitat Stuttgart, 2000.

[11] M. Ueyama and S.-A. Jun, "Focus realization of japanese english and korean english intonation," *UCLA Working Papers in Phonetics*, pp. 110–125, 1996.

[12] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[13] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*. IEEE, 2013, pp. 86–91.

[14] O. Kang, D. O. Johnson, and A. Kermad, *Second Language Prosody and Computer Modeling*. Routledge, 2021.

[15] M. Higuchi, M. Nakamura, S. Shinohara, Y. Omiya, T. Takano, D. Mizuguchi, N. Sonota, H. Toda, T. Saito, M. So *et al.*, "Detection of major depressive disorder based on a combination of voice features: An exploratory approach," *International journal of environmental research and public health*, vol. 19, no. 18, p. 11397, 2022.

[16] J. C. Lau, S. Patel, X. Kang, K. Nayar, G. E. Martin, J. Choy, P. C. Wong, and M. Losh, "Cross-linguistic patterns of speech prosodic differences in autism: A machine learning study," *PloS one*, vol. 17, no. 6, p. e0269637, 2022.

[17] D. Mladeni, J. Brank, a. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 234–241.

[18] Igarashi, "Handbooks of japanese language and linguistics.chapter 13: Intonation," pp. 525–568, 2015.

[19] Y. Ohara, "Gender-dependent pitch levels: A comparative study in japanese and english," in *Locating Power. Proceedings of the Second Berkeley Women and Language Conference, 1992*, 1992.

[20] C. Coupé, Y. M. Oh, D. Dediu, and F. Pellegrino, "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche," *Science advances*, vol. 5, no. 9, p. eaaw2594, 2019.

[21] F. Biassoni, S. Balzarotti, M. Giamporcaro, and R. Ciceri, "Hot or cold anger? verbal and vocal expression of anger while driving in a simulated anger-provoking scenario," *Sage Open*, vol. 6, no. 3, p. 2158244016658084, 2016.

NATIVENESS AND INTELLIGIBILITY OF JAPANESE ACCENTED ENGLISH CONSONANTS BY FRENCH LISTENERS

Rubén Pérez-Ramón[1], Mariko Kondo[1], Sylvain Detey[1], Lionel Fontan[2], Maelle Amand[3], Takeki Kamiyama[4]

[1]Waseda University, [2]Archean Labs, [3]CeReS Université de Limoges, [4]TransCrit Université Paris 8 Vincennes - Saint-Denis
rperez@aoni.waseda.jp, mkondo@waseda.jp, detey@waseda.jp, fontan.lionel@gmail.com, maelle.amand@unilim.fr, takeki.kamiyama@univ-paris8.fr

ABSTRACT

The presence of a Foreign Accent (FA) in the communicative process may interfere with the correct transmission of a message. This interference can arise from various traits, such as rhythm, intonation and speech rate. This can affect perception by both L1 and L2 listeners. Therefore, in this study we focused on the segments of Japanese FA in English as perceived by French listeners. Seven consonants ([l, ɹ, t^h, k^h, θ, f, v]) were extracted from the beginning of CVC English words. Using sound manipulation techniques, they were transformed into their most expected Japanese realisation ([ɾ, ɾ, t, k, s, ɸ, b] respectively) to generate a new set of words in which only the first segment was accented. The French cohort showed clear differences in perceived intelligibility of rhotic consonants and the [f/ɸ] contrast compared with previously analysed groups of Japanese and American listeners.

Keywords: perception, foreign accent, intelligibility, consonants

1. INTRODUCTION

There are many factors involved in second language (L2) phonological acquisition, such as native-likeness, fluency, intelligibility and comprehensibility. Acquiring "nativelike" pronunciation used to be considered the key issue in teaching or learning a foreign language, which meant that having a foreign accent was considered to be a failure of L2 acquisition. However, now the focus of L2 teaching and learning has shifted to intelligibility and comprehensibility [1]. So, while speech with a strong foreign accent may not always be easy to follow, studies such as [2] have shown that simply having a foreign accent does not necessarily hinder intelligibility or comprehensibility. In addition, since English has become a lingua franca, there are many varieties of English throughout the world, and the number of L2 English users has surpassed first language (L1) English users [3, 4]. This means that the majority of English users now have a foreign accent.

In terms of L2 perception, some studies have found that segmental accuracy is more important than suprasegmental accuracy in relation to correct word recognition, i.e., intelligibility [5, 6, 7], while other studies have shown that suprasegmental accuracy has a greater effect on native speakers' judgement of L2 speakers' fluency and nativeness of speech [8, 9].

Most assessments of learner pronunciation are done by native speakers of the target language. However, it is also important to consider how speakers of other languages perceive L2 pronunciation. Earlier studies [10, 11] compared assessments of Japanese accented English by assessors of 10 different L1s. It was found that assessors' L1 had a significant influence on their fluency assessment scores. However, more studies are needed to better understand how listeners' L1 affects perception of foreign accent.

Therefore, in this study, we examine the assessments of nativeness and intelligibility of Japanese-accented English – especially on the consonantal level – by native speakers of French and compare it with a prior assessment by native speakers of American English and Japanese.

2. JAPANESE AND FRENCH CONSONANTAL SYSTEMS

In the study, we focused on the following English consonants, [l, ɹ, t^h, k^h, θ, f, v], in the word initial position of a monosyllabic consonant-vowel-consonant (CVC) word. These segments were chosen because they are the most troublesome segments for native Japanese speakers in terms of their English production and perception [12, 13, 14, 15]. The important considerations regarding Japanese are as follows. First, Japanese has only one liquid consonant, /r/, and there is no phonemic distinction between /r/ and /l/. The most common realisation of /r/ is an alveolar tap [ɾ] [16, 17], although many other variants including [l] also occur [18, 19]. Japanese does not have the labio-dental fricatives /v/ and /f/. The most common substitutions are [b] for /v/ and bilabial fricative [ɸ] for /f/. Furthermore, Japanese lacks the voiceless inter-dental fricative /θ/ which is usually substituted with [s]. Finally, while the

plosives /t/ and /k/ are common consonants in Japanese, they are produced with less aspiration in all positions including the word-initial position that is the focus of this study.

Next, let us consider the French consonant inventory, and the main potential interferences with English [20] and Japanese [21, 22] consonants described in the literature. It is relatively similar to the English consonant system, but there are differences which may affect the evaluation of Japanese accented English by French speakers. For example, like Japanese, /θ/ is absent in French, and the most common substitute used by French speakers is [s], just like Japanese speakers [23, 12, 13]. While the other consonants examined in this study are found in French, the phonetic realisations can be quite different from English. For example, the French plosives /t/ and /k/ are unaspirated in word initial position, unlike their English equivalents, in word-initial position [24, 13]. In addition, while French has the voiceless labio-dental fricative /f/, the bilabial fricative [ɸ], which is a common Japanese substitution for /f/, does not occur in French. The French /v/ is similar to English /v/. With regard to lateral consonants, /l/ is always clear in present-day French ([l]). However, French /r/ is either a uvular fricative [ʁ] or slightly fronted velar [ɣ] in prevocalic position in most accent varieties in France, both of which are different from the English approximant [ɹ] which can be perceived by French speakers as [w] [25]. Therefore, such phonemic and phonetic differences between English, Japanese and French may influence the French speakers' judgement on Japanese accented English.

3. METHODOLOGY

3.1 Tokens

Seven target English consonants were selected according to their known impact on intelligibility when accented by Japanese speakers. The rationale for this selection is detailed in [26]. For each of these seven segments, three words were selected based on their structure and frequency, i.e., all of them had a consonant-vowel-consonant structure and ranked among the 1000 most frequent words in English according to the n-gram analysis of the Google corpus (Table 1).

A bilingual speaker was recruited based on her nativeness in both American English and Japanese, as judged by native speakers of each language [27]. In addition to the recorded English words, the speaker also recorded a set of non-words with a Japanese accent, one for each English word. The accented segment of each non-word was placed in the same (or similar) phonetic context to its equivalent English word. For example, for the English word *fish* [fɪʃ], the Japanese non-word *fishi* [ɸiɕi] was recorded, so [f] and [ɸ] would be in a similar context [_/i/]. Afterwards, by applying the splicing technique [28], a new token was generated in which only the initial consonant was pronounced with a Japanese accent, while the rest of the word remained with an unaltered English pronunciation (in the case of the example given above, this new word would be [ɸɪʃ]). Finally, the gradation technique [29] was applied, and, for each pair of generated tokens, a 7-step continuum was generated in which the accent of the target consonant varied gradually from fully accented to fully native. The final number of tokens was, therefore, 147 (7 consonants × 3 words × 7 steps).

Native segment	Foreign segment	English word	Japanese non-word
l	ɾ	leave	rifi
		league	rigi
		long	rogo
ɹ	ɾ	run	rana
		reach	richi
		roof	rufu
t^h	t	tag	taga
		teach	tichi
		took	tuku
k^h	k	keep	kipi
		come	kama
		call	koro
θ	s	thick	shiki
		thought	soto
		thumb	sama
f	ɸ	feet	fiti
		fish	fishi
		food	fudu
v	b	vet	bete
		van	bana
		ville	biri

Table 1: Target segments in English, Japanese segment for the accented version generation and selected English words for the experiment. Note that phonotactics in Japanese do not allow for a combination of [s]+[i], so the accented version of *thick* is closer to [ɕik].

3.2 Participants

Four cohorts were recruited for the experiment. A group of native American English participants were recruited using the Amazon Mechanical Turk tool (n=14). All of them were native speakers of English,

born in the United States and reported no contact with native speakers of any other language or from any other linguistic backgrounds. Two cohorts of Japanese participants, all members of the student population at Waseda University (Tokyo, Japan), also took part in the experiment: one group had high proficiency in English (n=25) and the other had low proficiency in English (n=26). These two groups were analysed separately because proficiency in a language has been demonstrated to have a significant effect on the perception of a foreign accent [29], and the Japanese cohort was the only one with enough heterogeneity to create two separate groups. Proficiency was self-reported and standardised according to the Common European Framework [30]. Finally, a group of French listeners was recruited from a body of university students in France (n=7). None of the non-native participants were bilingual in any language and had not lived for long periods of time in foreign countries. All subjects were paid for their participation.

3.3 Tasks

All participants underwent two tasks, one for intelligibility and one for nativeness, which were part of a bigger experiment.

In the intelligibility task, the participants were asked to listen to tokens and then type the word they heard on a screen. Once written, the screen instructed them to press a button with the text "Next", and the following token played automatically, leading to a self-paced task. They were asked to type always real English words.

The nativeness task consisted of choosing between two buttons in which the labels "Native" and "Foreign" could be read. Participants were asked to press the appropriate button based on their perception of every token. In this task, the word appeared written on the screen in order to avoid the Ganong effect [31], i.e., so listeners can decide whether a token that sounds like [bæn] is a good exemplar or not of the word *van* without ascribing it to the word *ban*. Therefore, this task was conducted last, so that any learning effect would not affect the intelligibility task.

For both tasks, stimuli were pseudo-randomised to ensure that no two tokens from the same target consonant were presented consecutively. The experiment was designed using the platform for online experimenting Gorilla [32], and participants completed all the tasks in their own house with their own equipment, following two requirements: to complete the tasks in a quiet setting and to use headphones.

4. RESULTS

Results for the Native American English (N) and the Japanese cohorts (HP for High Proficiency and LP for Low Proficiency) have been reported in [26]. In this study, we will focus on the information collected from the French cohort (F) and compare it with the other groups.

4.1 Preprocessing and statistical analysis

Before proceeding with the analysis of the results, answers deemed too fast (< 0.3s) or too slow (> 5s) were removed from the dataset. All statistical analysis was carried out with R [33]. General linear mixed models were calculated using the `lme4` package [34], and significant effects were measured using the *Anova* function of the `car` package [35]. Post-hoc pairwise comparisons were conducted using the `emmeans` package [36].

4.2 Nativeness results

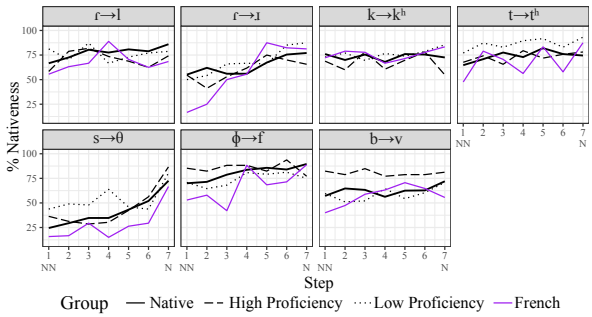


Figure 1: Results for the nativeness task for the four groups. Results are presented from the most accented end of the continuum (NN) to the most native-like token (N). Error bars are omitted for clarity.

A visual inspection of the results reveals some difference of the French cohort with the other groups. In [ɸ→f] and [ɾ→ɹ], a somewhat lower nativeness was reported, meaning that the accented realisation of these tokens was more harshly judged by the French cohort than the other groups. Additionally, in [s→θ], the French listeners seemed to perceive the change in category later than the other groups.

A general linear mixed model with CONSONANT and COHORT as fixed factors and PARTICIPANT as a random factor revealed a significant interaction between CONSONANT and COHORT ($p < .001$) as well as a significant effect of CONSONANT ($p < .001$). Averaged across steps, the French group showed a significant difference with the HP group for the [b→v] continuum ($p < .05$) and with the LP group for the [t→tʰ] and [s→θ] continua ($p < .05$). Marginal differences were found with the LP group for the [ɸ→f] continuum. Individual analysis carried out for

each continuum with p-values corrected for multiple comparisons revealed a significant difference with the HP group for the first 3 steps of the [ɸ→f] continuum ($p < .05$) and for step 4 of the [s→θ] and [t→tʰ] continua ($p < .05$).

4.3 Intelligibility results

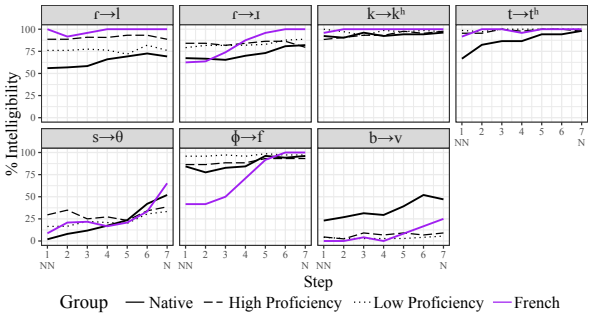


Figure 2: Results for the intelligibility task for the four groups. Results are presented from the most accented end of the continuum (NN) to the most native-like token (N). Error bars are omitted for clarity.

The replacement of [f] with [ɸ] seems to have affected the French listeners significantly more than any of the other three groups. There is also a clearly different pattern of the [ɾ→ɹ] continuum, which follows a trend similar to [ɸ→f], specifically, a sharp change in perceived category at the central part of the continuum (i.e., categorical perception).

As with the nativeness task, the results from the intelligibility task were used to generate a similar general linear mixed model. The model showed a significant interaction between sound and cohort ($p < .001$) and a significant effect of sound ($p < .001$). The analysis averaged across steps revealed significant differences of the French group with the native, high proficiency ($p < .05$) and low proficiency ($p < .001$) groups for the [ɸ→f] continuum. Similarly, for [ɾ→l], a significant difference was found between the French and the other three groups (N, LP: $p < .001$; HP: $p < .05$). The analysis of individual sounds only showed significant differences for the [ɸ→f] continuum, with the native cohort at the first step ($p < .05$) and with the other two cohorts at steps 1, 2 and 3 (HP: $p < .05$; LP: $p < .01$).

5. DISCUSSION AND CONCLUSION

The main goal of this study was to compare the perception of Japanese accented English consonant segments by French listeners with that of a native cohort (American English) and a matched non-native cohort (Japanese). For the latter, results from two sub-groups were collected, attending to their proficiency in English.

Previous research [26] analysed the impact of a foreign accent on perception by both native speakers of the target language and the accent of the speaker. In [37], the concept of matched interlanguage speech intelligibility benefit was coined to refer to the fact that listeners whose L1 is the same as the speaker report a less extreme drop in intelligibility. Moreover, the authors also suggested a mismatched interlanguage speech intelligibility benefit in the sense that listeners with an L1 that differs both from the target language and the accent of the speaker could experience a similar "benefit".

However, this mismatched interlanguage benefit in intelligibility (which is the case of the French cohort analysed in this study) has been disputed [38] or, at least, not found for other language pairs. In our study, we have found that this advantage in intelligibility is not only dependent on the language of the speaker, but also on the phonotactic strategy that leads to a certain transformation when speaking with a foreign accent.

For French listeners, the changes in [ɸ→f] and [ɾ→ɹ] involve unfamiliar (Japanese) units, which may explain the noticeable difference in the intelligibility patterns. In these cases, the French cohort perceived a more abrupt change in the category along the continuum, suggesting that a new segment conveys a bigger loss in intelligibility as the listener is not able to assign the proper category.

On the other hand, [s→θ] and [t→tʰ] are changes that are also expected among French listeners (though the latter is seen more often in low proficiency speakers with little metaphonological awareness). Our results confirm that, when this is the case (i.e., when the transformation of accentedness is similar in two languages), intelligibility follows a similar pattern.

In conclusion, our study complements the classical theories of the mismatched interlanguage speech intelligibility benefit by providing detailed information about individual segments. By applying the segmental FA generation technique and the gradation technique, we can conclude that patterns of perception are not dependent on the L1 of the listener, but are probably related to the phonological transformations that arise in their mind at the time of speaking a non-native language.

7. REFERENCES

[1] Levis, J. 2020. Revisiting the intelligibility and nativeness principles. *J. of Second Language Pronunciation*, *6*(3), 310-328.

[2] Munro, M. J. & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *J. Second Language Pronunciation*, *6*(3), 283-309.

[4] Crystal, D. 2003. *English as a Global Language*. CUP.

[5] Ethnologue Languages of the World, 2022. https://www.ethnologue.com/
[6] Fayer, J. M. & Krasinski, E. 1987. Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.
[7] Koster, C. J. & Koet, T. 1993. The evaluation of accent in the English of Dutchmen. *Language Learning,* 43, 69-92.
[8] Rognoni, L. & Busà, M. G. 2014. Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating: An experiment using prosody transplantation. *Proc. the International Symposium on the Acquisition of Second Language Speech, Concordia Working Papers in Applied Linguistics*, Vol. 5. 547-560.
[9] Anderson-Hsieh, J., Johnson, R. & Koehler, K. 1992. The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42(4), 529-555.
[10] Gut, U. 2003. Non-native speech rhythm in German. *Proc. 15th ICPhS,* Barcelona, 2437-2440.
[11] Kondo, M. & Konishi, T. 2017. Tsuujiru Eigo no tameno hatsuon kyouiku (What we need to teach for better communication in English)," *Conference Handbook of the 35th Conference of the English Linguistic Society of Japan*, Sendai, Japan, 232-237.
[12] Kondo, M., Fontan, L., Le Coz, M., Konishi, T. & Detey, S. 2020 "Phonetic fluency of Japanese learners of English: automatic vs native and non-native assessment", *Proc. Speech Prosody 2020,* 784-788.
[13] Yavaş, M. 2011. *Applied English Phonology (2nd ed.)*, Wiley-Blackwell.
[14] Collins, B. & Mees, I. M. 2013. *Practical Phonetics and Phonology - A resource book for Students* (*3rd ed.*). London: Routledge.
[15] Rogerson-Revell, P. 2011. *English Phonology and Pronunciation Teaching*. Continuum International Publishing Group.
[16] Kondo, M., Tsubaki, H. & Sagisaka, Y. "Segmental Variation of Japanese Speakers' English: Analysis of "the North Wind and the Sun" in AESOP Corpus", J. of Phonetic Society of Japan, 19(1), 3-17.
[17] Saitô, Y. 2006. *Nihongo Onseigaku Nyûmon* (*Introduction to Japanese Phonetics*) (2nd ed.). Sanseido.
[18] Vance, T. J. 2008. *The Sounds of Japanese*. CUP.
[19] Ohnishi, M. 1987. Dôteki jinkô kôgai ni yoru nihongo shiin no kôsatsu [A study of Japanese consonants by means of an artificial palate]. Onsei Gakkai Kaihô 186: 19–24.
[20] Arai, T. (1999), A case study of spontaneous speech in Japanese, *Proc.14th ICPhS*, San Francisco. 615-618.
[21] Tennant, J. 2016. Les anglophones. In S. Detey, I. Racine, Y. Kawaguchi & J. Eychenne (Eds.), *La pronunciation du français dans le monde : du natif à l'apprenant*. CLE international, 97-102.
[22] Kamiyama, T., Detey, S. & Kawaguchi, Y. 2016. Les japonophones. In S. Detey, I. Racine, Y. Kawaguchi & J. Eychenne (Eds.), *La pronunciation du français dans le monde : du natif à l'apprenant*. CLE international, 155-162.
[23] Detey, S., Fontan, L., Le Coz, M., Barcat, C. & Kawaguchi, Y. (in press). Identifying segmental substitutions in spontaneous speech of L3-French/L1-Japanese learners: a corpus-based pilot study. *J. of the Phonetic Society of Japan.*
[24] Tranel, B. 1987. *The Sounds of French*: An Introduction, CUP.
[25] Hallé, P., Best, C. T., Levitt, A. 1999. Phonetic vs. Phonological influences on French listeners' perception of American English approximants, J. *of Phonetics*, 27, 281-306.
[26] Kondo, M. & Pérez-Ramón, R, (in press) Perception of Japanese Accented English segments in Words, *J. of Second Language Studies.*
[27] Pérez-Ramón, R. & Kondo, M. 2021."Relationship between reaction time and nativeness rating", Proc. of 35th Annual Meetings of the Phonetic Society of Japan, 95-100.
[28] García Lecumberri, M. L., Barra-Chicote, B., Pérez-Ramón, R., Yamagishi, J. & Cooke, M. 2014. Generating segmental foreign accent. Proc. INTERSPEECH 2014, 1303-1306.
[29] Pérez-Ramón, Rubén and Kondo, Mariko (2021) "Relationship between reaction time and nativeness rating", The proceedings of the 35th Japan Phonetic Society Annual Meeting, 95-100, 25-26 September, 2021.
[30] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
[31] Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, *6*(1), 110.
[32] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388-407.
[33] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
[34] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
[35] John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage.
[36] Russell V. Lenth (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.4.
[37] Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600-1610.
[38] Stibbard, R. M., & Lee, J. I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *The Journal of the Acoustical Society of America*, *120*(1), 433-442.

FRENCH /ø-u/ CONTRAST IN JAPANESE LEARNERS WITH AND WITHOUT GESTURE FEEDBACK IN MONOSYLLABLES: A DESCRIPTIVE ACOUSTIC STUDY

Claire Pillot-Loiseau[1], Takeki Kamiyama[2, 1], Ayako Kitamura[3]

[1]Laboratoire de Phonétique et Phonologie (LPP) UMR 7018 CNRS / Sorbonne Nouvelle University, France
[2]UR TransCrit (EA1569), Paris 8 Vincennes - Saint-Denis University, France [3]Sophia University, Tokyo, Japan
claire.pillot@sorbonne-nouvelle.fr, takeki.kamiyama@univ-paris8.fr, ayakokitamura@sophia.ac.jp

ABSTRACT

This study investigated French /ø/-/u/ contrast produced by 2 native speakers (NS) and 12 Japanese-speaking learners; 7 of them received 6 lessons on these vowels with hand and arm gesture feedback (GG), and the others 6 lessons with articulatory explanation (AG). Acoustic measurements (F1, F2, F3, F1/F2 distance, Euclidean distance - ED) of 2 repetitions of /dø/, /ʁø/, /du/ and /ʁu/ are compared to /ø/ and /u/ in isolation, before, after and two months after the lessons. F2 of /u/ decreases more for GG. F3 of /ø/ decreases more for AG. F1/F2 distance is smaller for /ʁø/ and /ʁu/ than for the other monosyllables, while it is smaller and closer to NS for /u/ produced by 4 learners in GG than the others. F2 of /u/ improved due to visual and kinesthetic (gesture) modalities.

Keywords: L2 vowel learning, French, Japanese, gesture as feedback, acoustics.

1. INTRODUCTION AND GOAL

The vowel system of the first language (L1) has a significant impact on the acquisition of the vowels of second languages (L2). For example, front rounded vowels are known to cause difficulty to speakers of languages without such vowels ([1-4], *inter alii*).

In present-day Parisian French, the close-mid front rounded /ø/ and the high back /u/ are phonemically contrasted. Achieving this contrast is thus of great importance for learners of L2 French and often poses a challenge to learners without such vowels in L1, as is the case with Japanese-speaking learners of French (JSLF). Native speakers of Tokyo Japanese tend to produce the French /u/ with a higher F2 than native speakers of French (NSF), resulting in a sound that is perceived mostly as /ø/ by native listeners of French [5]. This stems from the high non-front /u/ in Tokyo Japanese, commonly transcribed [ɯ]. Acoustically, it shows a higher F2 (> 1000 Hz) than the French /u/; articulatorily, the tongue is less retracted and the lips are less rounded than for French /u/ ([6-7]). Besides, it has been observed that JSLF sometimes realize /ø/ as [ɯ] [8], with F2≈1800Hz, F3≈2900Hz for beginners [9], whereas NSF produce this vowel with lower formants (F2≈1600 Hz and F3≈2700Hz for female speakers: [10], *inter alii*). These tendencies lead to the difficulty that JSLF face in realizing the phonemic contrast /ø/-/u/ [9], which is, in PAM-L2 terms [11], a case of single-category or category-goodness assimilation.

Spoken language proficiency is a multimodal task involving auditory, kinesthetic and visual perception; gestures are also part of it. It also involves the production of complex articulatory gestures. The advantage of hand and arm gesture feedback is that it requires no additional equipment, and it can be used in distance learning. This type of embodied feedback was mainly used to improve L2 prosody [12], but little for segmental contrasts [13, 14]. It has been successfully applied to the training of consonants in L2 teaching [13], and to the case of Japanese short and long vowels acquired by Catalan-speaking learners with and without hand gestures [15]. However, to our knowledge, the effect of training to produce the French vowel contrast /ø/-/u/ by JSLF with and without hand and arm gestures has never been studied. Our goal is to explore this effect for monosyllables: does hand gesture feedback improve vowel production over time more than articulatory explanation? What are the differences between learners?

2. METHOD

12 Japanese-speaking learners of French as a Foreign Language FFL (3 male, 9 female) took part in a series of 6 weekly individual lessons of 20 minutes each by videoconference (Zoom) given by an experienced instructor of FFL. An auditory identification test of minimal-pair monosyllables, longer words and phrases with /ø/ and /u/, a reading of similar monosyllables, words and phrases, and a survey (linguistic profile questionnaire at the beginning, satisfaction survey at the end) were administered before (Pre), immediately after (Post), and, for the auditory identification test and the reading, 2 months after the lessons (2M).

Participants were divided into two groups matched for the results of the first perception test, for their age, proficiency level of French and length of stay in

France. 7 learners in the "gesture" group (GG) received the training sessions with the movement of the arms and hands associated with the articulatory features of /ø/ and /u/: G1 (19 y.o., B1 in the Common European Framework of Reference for Languages), G2 (33, C1) G3 (31, A1), G4, (41, B2), G5 (31, B1), G6 (♂, 27, B1) and G7 (65, CEFR level unknown). 5 learners in the "articulatory" group (AG) had training sessions with articulatory explanations on /ø/ and /u/ provided in Japanese but without using gestures: A1 (♂, 62 y.o., B2 in CEFR), A2 (♂, 23, B2), A3 (39, CEFR level unknown), A4 (31, C1) and A5 (38, B1). Seven of them were born in Tokyo (G2, G3, A2, A4) or near Tokyo (G1, G7, A3). They started learning French as adults (learning experience of French: 3 months to 6 years), and had been living in France between 3 months and 21 years at the time of the study. One male (MN, 54 y.o) and one female native speaker (FN, 52 y.o.) of French participated as a reference for comparison.

All training sessions were devoted to vowel(?) production. The training began with the repetition of CV monosyllables (e.g., /ku/, /kø/), followed by the repetition of bisyllabic words containing /ø/ and /u/ in phrase-final (phrase-accented) syllables and non-final syllables (e.g., *couteau* /k**u**to/, *jeudi* /ʒ**ø**di/). Next, learners repeated di- and trisyllabic words and longer phrases containing two tokens of the same target vowel (e.g., *heureux* /**øʁø**/, *kangourou* /kɑ̃g**u**ʁ**u**/) and then two different target vowels (e.g. *coûteux* /k**u**t**ø**/, session 3). They then repeated minimal pairs (e.g., *bourrer* /b**u**ʁe/; *beurrer* /b**ø**ʁe/, session 4). In session 5, they read words and phrases introduced in the first 4 sessions. Finally, they read the lyrics of a song and a short text containing the target vowels (session 6). During the last two sessions, the subjects in GG read the words and phrases first with gestures and then without them. The arm and hand movements of the selected gestures, inspired by [16], are aimed to represent the articulatory characteristics of the target vowel (e.g., tension, tongue position). When pronouncing /u/, the elbows are raised laterally to shoulder height, with the hands and arms placed horizontally (palms down), then the elbows are pulled back vigorously. This gesture is intended to encourage Japanese-speaking learners to place their tongue in a posterior position. For the target vowel /ø/, the arms are stretched forward (hands with palms up and down), as if to hold a long pipe between the two outstretched arms. As articulating /ø/, the arms and hands are moved slightly forward. This gesture is intended to encourage Japanese speakers to move their tongue and lips forward.

The recorded material for each test and each subject included /ø u ɑ̃ ɔ̃/ in isolation and 24 CV monosyllables where V=/ø u ɑ̃ ɔ̃/ and C=/b d g z ʒ ʁ/, presented to the participants in a random order and read aloud by them. The list of these 28 items were repeated twice and a total of 56 monosyllable tokens were obtained. A list of 40 minimal-pair words and phrases of 2 to 5 syllables containing the target vowels were also read aloud twice during this test. Only the monosyllables "eu"/ø/, "ou"/u/ (in isolation), "deu"/dø/, "dou"/du/ (dental context) and "reu"/ʁø/, "rou"/ʁu/ (uvular context) are analysed in this article (12 items per subject).

The acoustic signals, recorded at 44100 Hz, 16 bits, were subjected to the analysis of the first three formants, semi-automatically measured at 50% of vowel duration, using Praat [17]. This allowed measuring the Euclidean distance (ED) between /ø/ and /u/ in order to quantify the degree of achievement of the phonemic contrast, and calculating the distance between F1/F2 for [u] [10]. French focal vowels [18] in isolation are known to be characterized by a distance of less than 800Hz [10] between the neighbouring formants: F1 and F2 for /u o ɔ (ɑ)/.

The present study is descriptive. ANOVA tests were only applied to the formant frequency measurements of the productions of all same-sex learners in a group (GG or AG, Fig. 1). The small number of measurements per subject and per monosyllable did not allow further acoustic analyses (Fig. 2 to 4).

3. RESULTS

3.1. Formant frequencies

Figure 1 shows the evolution, throughout Pre, Post, and follow-up 2M, of the average frequencies of F1 to F3 of all learners of the same sex combined, broken down into groups ("articulatory" AG vs. "gesture" GG), all monosyllables of the same target vowel combined. The formant values of /ø/ do not change significantly overall and remain close to those of native speakers. Although we observe that from Pre to 2M, F1 of /ø/ and /u/ increases only for the male learner of GG and approaches that of MN, this increase is not statistically significant ($F_{(3,44)}$=0.21; p=0.889 for /ø/; $F_{(3,44)}$=0.21; p=0.929 for /u/). There was a very slight increase in F2 for /ø/ for women and men in GG (respectively $F_{(3,224)}$=0.59; p=0.62, and $F_{(3,44)}$=0.18; p=0.899), while there was a decrease in F3 for /ø/ and /u/, especially for AG. The most salient formant change observed (yellow rectangles, Fig. 1) is a significant decrease in F2 of /u/ for GG, especially between Post and 2M ($F_{(3,224)}$=9.7; p<0.0001 for female and $F_{(3,44)}$=5.71; p=0.0018 for male learners in GG).

3.2. F1-F2 distance for /ø/ and /u/

Figure 2 shows the evolution, throughout Pre, Post and 2M, of the average F1-F2 distance for each speaker for “eu” /ø/, “deu” /dø/ and “reu” /ʁø/. F1-F2 distance of “reu” is smaller than that of “eu”, and “deu” for 10 out of the 12 learners. Moreover, the evolution is variable according to the subjects: there is little difference between the two groups. It does not result in a convergence towards the values of the native subjects (dotted lines). In addition, G1, G2, G4 and G5 had F1-F2 distance close to that of the native subjects already in Pre. Overall, more subjects in GG show F1-F2 distance as small as 800Hz to 1200Hz already in Pre than those in AG (higher values for this latter group).

Figure 3 shows the evolution, throughout Pre, Post and 2M, of the average F1-F2 distance for each speaker for “ou” /u/, “dou” /du/ and “rou” /ʁu/. F1-F2 distance of /u/ in "rou" is smaller than that of "ou" and "dou" for 8 out of the 12 subjects. Moreover, the evolution is variable from one learner to another: the variability is remarkable for subjects G1 and G6 with a decrease of this distance in Post and 2M, except for "dou" for G1. A2 progressed by decreasing his F1-F2 distance by 30% (-400Hz) between Pre and Post, but lost this gain for "ou" and "rou" in 2M. In addition, G2 for "rou", G3, G4 for "dou", G5 and G6 for "rou" showed small F1-F2 distance, close to the native values already in Pre. Overall, more subjects in GG showed F1-F2 distances around 400Hz (value of "rou" for both natives and "ou" for the female native), especially for "rou", than subjects of Group AG.

3.3. Euclidean Distance (ED)

Figure 3 shows the evolution, throughout Pre, Post, and follow-up 2M, of the average Euclidean Distance (ED) for each speaker for “eu-ou” /ø-u/, “deu-dou” /dø-du/ and “reu-rou” /ʁø-ʁu/ contrasts. A larger ED suggests that /ø/ and /u/ were produced more distinctly. There is an increase of ED for G1 for “eu-ou”, and from Pre to Post for “deu-dou” and “reu-rou”, G04 (for “reu-rou” and “eu-ou” at 2M), G6 (for “eu-ou” and “deu-dou” at 2M), G7 for “deu-dou” at 2M), but also for “reu-rou” for A2. However, there is a decrease in ED for G3 (except for “deu-dou”), G4 for “deu-dou”, G5 (except for “eu-ou”) and G7 for “reu-rou”. Except for subjects G1 for “eu-ou”, G3 for “eu-ou” and “reu-rou”, G6 for “eu-ou” and “deu-dou” at 2M, ED is not greater for participants in GG than for those in AG. In addition, subjects A1 and A2 increase their ED for vowels in isolation from Pre to Post, and then to 2M.

4. DISCUSSION

The results for ED show that the use of arm and hand gestures moderately helps Japanese learners to better distinguish /ø/ and /u/, compared to learners in AG. However, 4 subjects out of 7 in GG decreased F2 and F1-F2 distance for /u/ in isolation and in dental and uvular contexts. The preceding consonant influences F1-F2 distance of /u/ more for AG than for GG, but differently depending on the subject. In GG, F1-F2 distance of "reu" and "rou" is globally smaller due to the uvular context that retracts the vowels /ø/ and /u/.

The gestures used for L2 Spanish acquisition by native Dutch speakers improved only /u/, not /θ/ [14]. Prosodic (not segmental) gestures improved French /y, ø, œ/ for 57 Catalan learners [19]. However, in both of these studies, learners only *looked* at the gesture. In our study, learners were additionally asked to *produce* this gesture at the same time as they pronounced the target vowels.

The moderate effect of the gesture action is probably due to the fact that the training was mainly based on repetition rather than reading. In addition, each training session was short (20mn). However, subjects were more motivated by gesture instruction, even when following the lessons through Zoom.

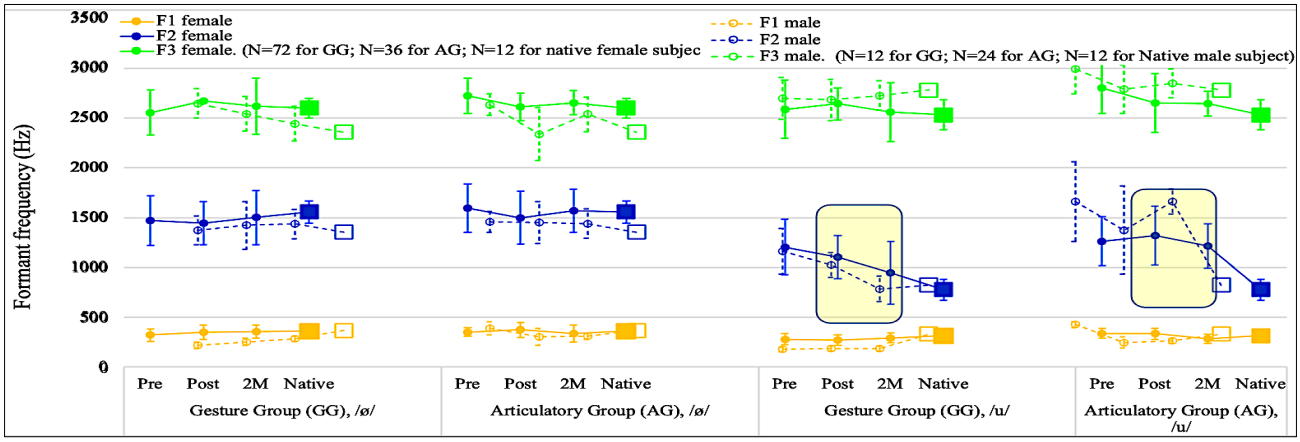


Figure 1: F1 to F3 frequencies before (Pre), just after (Post) and 2 months after (2M) the 6 lessons (all monosyllables included, MN and FN, and all subjects of Groups GG and AG). Light yellow rectangles: the most relevant changes between GG and AG (F2 lowering for /u/).

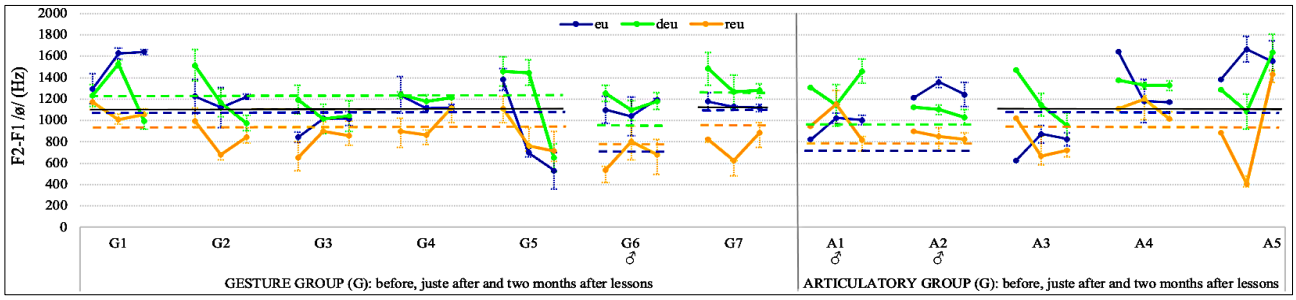


Figure 2: Mean (N = 2 for each monosyllable) F1-F2 distance and standard deviation (SD) for each speaker for "eu" /ø/, "deu" /dø/ and "reu" /ʁø/. Within each speaker, the lines connect Pre, Post and 2M. Horizontal dotted lines: values of MN and FN for the same monosyllables. Black horizontal solid lines: value of /ø/ by 40 native French-speaking women [10].

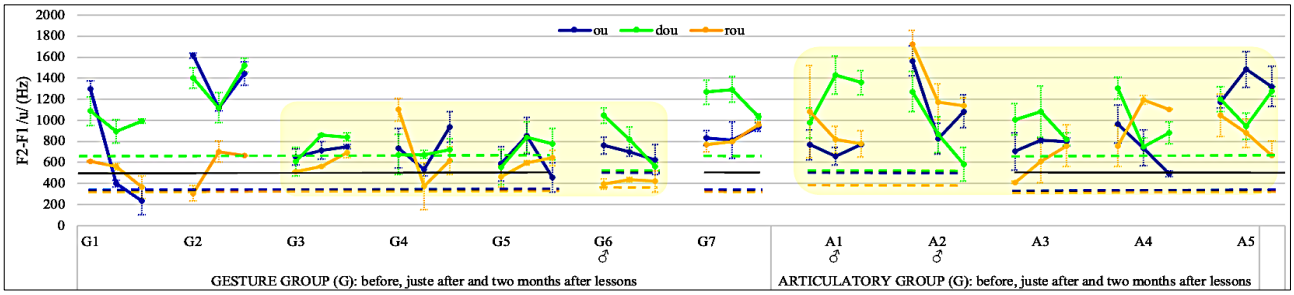


Figure 3: Mean (N = 2 for each monosyllable) F1-F2 distance and SD for each speaker for "ou" /u/, "dou" /du/ and "rou" /ʁu/. Within each speaker, the lines connect Pre, Post and 2M. Horizontal dotted lines: values of MN and FN for the same monosyllables. Black horizontal solid lines: value of /u/ by 40 native French-speaking women according to [10]. Light yellow rectangles: most relevant changes between GG and AG (smaller F1-F2 distance for most subjects in GG).

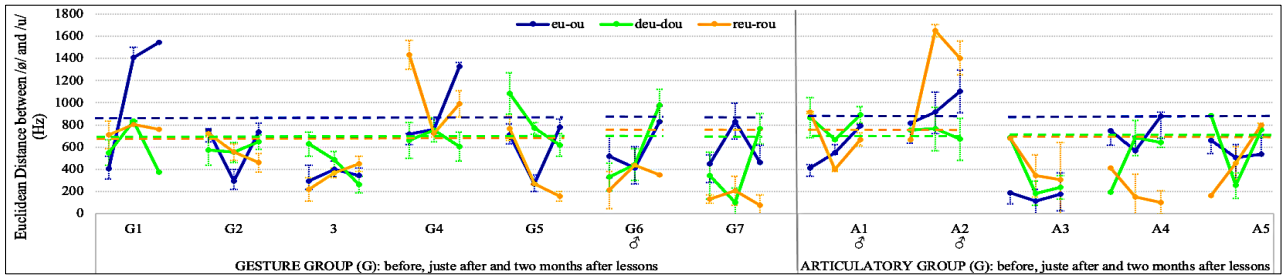


Figure 4: Mean (N = 2 for each monosyllable) ED and SD for each speaker between "eu-ou" /ø u/, "deu-dou" /dø du/ and "reu-rou" /ʁø ʁu/. Within each speaker, the lines connect Pre, Post and 2M. Horizontal dotted lines: values of MN and FN for the same monosyllables.

Furthermore, although the learners in GG were more satisfied with the training than those in AG (satisfaction survey), motivation and the ability to reproduce the gesture faithfully to the model varied among learners, according to the instructor who led their training (the third author). The nature of the gesture can also influence the production of the vowel phoneme: the gesture of /u/ insists more on the posterior position of the tongue than on the rounding and protrusion of the lips: this may be the reason why we do not observe a drop in F3 of this vowel for the learners of GG, contrary to those of AG: only the latter received an articulatory explanation on the lip rounding of /u/ and /ø/.

The learner's individual profile is another effect to be considered, as reported by [20]: for example, learner A2 is fluent in English, Spanish and French (which is his 4th language). Subject G7 is 65 y.o. and expressed great difficulty in pronouncing /ø/ and /u/. She started learning French after the age of 40.

5. CONCLUSION AND PERSPECTIVES

Such a multimodal approach can train the ear to perceive unfamiliar speech features by presenting them through visual and kinesthetic modalities. In addition, controlling pronunciation through hand and arm gestures bypasses ingrained patterns in the natural voice that may be difficult to correct [21].

This preliminary research needs to be complemented by examining more data from these monosyllables and other elements of the corpus already recorded (other monosyllables and phrases); studying vowel stability taking into account formant measures at 25% and 75% of each vowel; conducting a perceptual evaluation of these learners' productions by native listeners of French.

6. ACKNOWLEDGMENTS

This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

7. REFERENCES

[1] Flege, J. E. 1987. The production of new and similar phones in a foreign language: evidence for the effect of equivalence classification. *J. of Phonetics* 15, 47-65.

[2] Levy, E. S., Law II, F. F. 2010. Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *J. Acoust. Soc. Am.* 128(3), 1290-1305.

[3] Rochet, B. L. 1995. Perception and production of second-language speech sounds by adults. In: Strange, W. (ed), *Speech perception and linguistic experience: Issues in cross-language research*. Baltimore: York Press, 379-410.

[4] Lauret, B. 2007. *Enseigner la prononciation du français : questions et outils*. Paris: Hachette.

[5] Kamiyama, T., Vaissière, J. 2009. Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners. *Acquisition et interaction en langue étrangère Aile... Lia* 2, 9-41.

[6] Bothorel, A., Simon, P., Wioland, F., Zerling, J. P. 1986. *Cinéradiographie des voyelles et consonnes du français*. Strasbourg : Publications de l'Institut de Phonétique de Strasbourg.

[7] Uemura, Y., Takada, M. (Kokuritsu Kokugo Kenkyûjo [National Institute for the Japanese Language]) 1990. *Nihongo no boin, shiin, onsetu: chouon undou no jikken-onseigakuteki kenkyû* [Vowels, consonants and syllables in Japanese: an experimental phonetic study on articulatory gestures]. Tokyo: Shûei shuppan.

[8] Kamiyama, T., Detey, S., Kawaguchi, Y. 2017. Les japonophones. In: S. Detey, I. Racine, Y. Kawaguchi, J. Eychenne (eds), *La prononciation du français dans le monde : du natif à l'apprenant*. Hachette, 155-161.

[9] Marushima, N., Detey, S., Kawaguchi, Y. 2010. Caractéristiques phonétiques des voyelles orales arrondies du français chez des apprenants japonophones. *Flambeau* 36, 53-72.

[10] Georgeton, L., Paillereau, N., Landron, S., Gao, J., Kamiyama, T. 2012. Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE. *Proc. Journées d'Étude sur la Parole-TALN-RECITAL* Grenoble, 145–152.

[11] Best, C. T., Tyler, M. D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: M. J. Munro & O.-S. Bohn (eds), *Language experience in second language speech learning: In honor of James Emil Flege*. John Benjamins, 13-34.

[12] Baills, F., Alazard-Guiu, C., Prieto, P. 2022. Embodied Prosodic Training Helps Improve Accentedness and Suprasegmental Accuracy. *Applied Linguistics* 43(4), 776-804.

[13] Xi, X., Li, P., Baills, F., Prieto, P. 2020. Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features. *Journal of Speech, Language, and Hearing Research* 63(11), 3571-3585.

[14] Hoetjes, M., Van Maastricht, L. 2020. Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, *11*, 575032.

[15] Li P, Baills F, Prieto P. 2020. Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition* 42(5),1015-1039.

[16] Gladić, V. A. 1982. *Le graphisme phonétique : du geste phonétique au geste graphique*. Nathan.

[17] Boersma, P., Weenink, D. 2010. *Praat: doing phonetics by computer* [Computer program]. Version 6.2.23, http://www.praat.org/

[18] Schwartz, J.-L., Boë, L.-J., Vallée, N., Abry, C. 1997. The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25(3), 255-286.

[19] Li, P., Baills, F., Baqué, L., Prieto, P. 2022. The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second Language Research*, 02676583221124075.

[20] Li, J. J., Ayala, S., Harel, D., Shiller, D. M., McAllister, T. 2019. Individual predictors of response to biofeedback training for second-language production. *J. Acoust. Soc. Am.* 146(6), 4625-4643.

[21] D'alessandro, C., Rillard, A., Lebeux, S. 2011. Chironomic stylization of intonation. *J. Acous. Soc. Am.* 3129, 1594–1604.

PERCEPTION OF SPANISH DECLARATIVE QUESTIONS AND STATEMENTS BY L2 SPANISH SPEAKERS

Izaro Bedialauneta Txurruka
University of Illinois at Urbana-Champaign
Izarob2@illinois.edu

ABSTRACT

The distinction between declarative questions and statements in Spanish is often found only in prosody. However, question contours vary from one Spanish variety to another. In Central Peninsular Spanish (PS) neutral questions tend to end with a rise while Argentinian Spanish (AS) shows a circumflex pattern. Here we investigate the perception of 20 L2 speakers whose L1 is American English identifying Spanish declarative questions with rising vs circumflex contours. The goal is to investigate the influence of primary and secondary cues to interrogation by comparing original and manipulated F0 contours. In general, all participants identified original sentences accurately. Contrary to our predictions, circumflex contours were more consistently perceived as questions than rising contours. We speculate that this may be due to the presence of secondary cues for interrogative contours.

Key words: intonation, perception, declarative sentences, L2 Spanish speakers

1. INTRODUCTION

Declarative questions are a type of yes/no interrogative sentences that are lexically and syntactically identical to declarative statements. For example, the sentences "he is outside" and "he is outside?" are identical except for their intonation. Even though the intonational contours that differentiate these two sentences may be language and dialect-specific, universal tendencies have been identified for the expression of interrogativity, such as questions having an overall higher pitch, and higher onset and final pitch than statements (Ohala [1]; Henriksen [2]; Gussenhoven [3]; Murphy [4]; Chen [5]; Grabe [6]).

In American English (AE) there are two types of yes/no questions: a) Is it cold? (polar interrogative), b) It is cold? (declarative question). Not only do they have differences in their syntactic constructions but they also differ in pragmatics and usage. According to Gunlogson [7], "declarative questions are not appropriate in situations where the questioner is supposed to be neutral or ignorant". Unlike in AE, the use of neutral meaning of declarative questions is acceptable in Spanish.

In Spanish declarative questions and statements are lexically and syntactically identical: a) *bebe agua*. 'he/she drinks water', b) *¿bebe agua?* 'does he/she drink water?'. The listener must rely on prosody to differentiate questions from equivalent declarative statements, especially if the subject is left unexpressed (Face [8]; Henriksen [1]; Trimble [9]).

Moreover, the use of prosodic cues to mark interrogativity may vary across dialects. For instance, in the standard Spanish variety of Peninsular Spanish (PS) interrogativity is expressed by a rising boundary tone (H%) at the end of the interrogation and there is a fall (L%) in statements (Face [8]). Furthermore, it has been observed that the fundamental frequency (F0) of the first peak in questions is another cue that this variety of Spanish employs to distinguish the two types of sentences. According to Face [8], the final rise is the primary marker for interrogation, whereas the first peak is a secondary cue. Declarative questions start high at the beginning of the utterance, followed by a gradual descent and a rise at the end of the sentence (Navarro Tomás [10]). Additionally, if there is a pitch accent in sentence medial position, the rise of F0 is found in declaratives, but in questions it is not generally that consistent (Face [8]).

In Spanish from Buenos Aires, Argentina (AS), there is a height difference of F0 between declarative statements and questions. However, unlike PS, questions in AS have a circumflex contour with a falling boundary tone (HL%) at the end of the utterance [Prieto et al. [11].

In order to identify the primary and secondary cues to interrogation in Spanish, Face [8] explored the perception of interrogative sentences by L1 speakers using both original and manipulated sentences of their native dialect. The results showed that the primary cue for native speakers is found in the last part, in the boundary tone, of the utterance.

Regarding the perception of Spanish intonation by L2 speakers, Sanchez points out that "the study of the acquisition of intonation by L2 speakers is still an understudied topic within the field of Second Language Acquisition" (Sanchez [12]). Nevertheless, there are studies that are worth mentioning that have investigated the production of declarative statements and questions in bilinguals (Zárate-Sández [15]) and L2 speakers (Gabriel et al. [14]; Sanchez [12]; Trimble [9]; Yuan et al. [15], Brandl et al. [16], Henriksen, et al. [17], Zárate-Sández [18], Méndez

Seijas [19]); and there are also studies on perception of bilinguals (Nibert [20]; Sanchez [12]; Timble [9]; Casillas et al. [21]).

For example, Trimble [9] aimed to analyze the perception by L1 American English L2 Spanish speakers of declarative statements and questions, produced by a Castilian speaker, a Venezuelan speaker, and a proficient L2 Spanish speaker. The three speakers differed in the contours of the questions. The Castilian and the L2 speakers produced the question with a H% boundary tone, whereas the Venezuelan speaker sentences had a HL% boundary. The results showed only 44% of the Venezuelan interrogative utterances with the HL% boundary were perceived as question by the L2 speakers. This indicates that L2 speakers are sensitive to the boundary tones of the utterances when perceiving interrogation. Most of them agreed that H% signifies interrogation and L% finality, while there was less agreement in the meaning of HL%.

Here we aim to contribute to the field by examining the perception by L2 Spanish speakers of declarative questions and statements of PS and AS to better understand the acquisition of the intonation of Spanish as a second language across dialects. These two varieties were specifically chosen due to their differences in the acoustic contours in questions.

2. RESEARCH QUESTIONS AND HYPOTHESES

The study will address the following questions: Q1) How do L1 English L2 Spanish speakers interpret questions that do not end in a rise? Q2) Do listeners use both primary and secondary cues to interpret interrogative sentences? Q3) Do L1 and L2 Spanish speakers use the same cues?

L2 speakers should be able to identify correctly the Peninsular original statements and questions based on the universal cues and rising boundary tone. On the other hand, based on the results of Trimble [9] on the perception of the circumflex contour in questions, in interrogative sentences, we expect to observe variation when interpreting the Argentinian questions. Moreover, we expect that the main cue for questions to be the final boundary tone, even in the manipulated sentences, based on Face [8] and Trimble [9] for both L1 and L2 speakers.

3. METHODS

3.1. Participants

In total, 41 individuals participated in our perception study: 13 Peninsular Spanish-speakers (control group for PS sentences), 14 Argentinian Spanish-speakers (control group for AS sentences), and 20 L1 American English L2 Spanish-speakers. All L2 speakers had experience learning Spanish for over 5 years and their home language was English.

3.2. Stimuli

Participants heard 27 sentences. 18 of them were produced by a female native speaker of Central Peninsular Spanish: 6 original declarative questions (PS-OQ) and their counterpart statements (PS-OS), and 12 manipulated sentences. In the case of AS, only questions were recorded by a male speaker, since the two varieties differ mainly in the contour of interrogative sentences. There were 9 utterances, 3 of which were original (AS-OQ) and 6 were manipulated (AS-MQ).

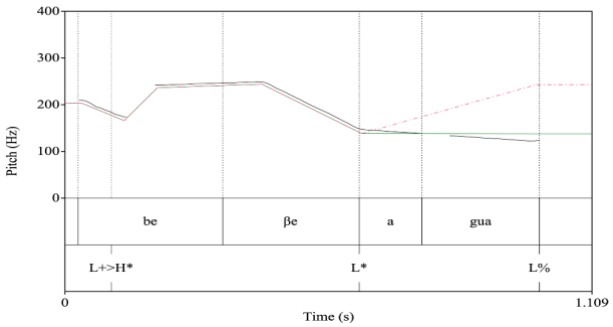


Figure 1: Manipulation from original statements. Pink line represents the rising boundary tone, the green baseline position and the black line represents the original contour.

Manipulations were made using the program Praat [22]. AS and PS contours were manipulated separately and all the utterances were resynthesized.

From the statements produced by the Peninsular Spanish speaker, two other manipulated sentences were created (see Figure 1), changing the boundary tone to two positions, high (PS-MSH) and at the baseline (PS-MSB) of the utterance. The high boundary tone was created considering the high peak of the lexical stress of the first content word making the sentence have a raise at the end of the utterance. The baseline position, on the other hand was in the same height as the L*, to ensure that there was no rising or falling boundary tone.

Similarly, from original questions two other sentences were created (see Figure 2). This time, since the questions already have a H% boundary tone, the manipulated sentences that were created had a falling tone (PS-MQL) or they ended in a baseline position (PS-MQB). In order to create the L% in the same way in all the stimuli, the height of L* was taken into account as a reference and afterwards, the contour was lowered 40 Hz.

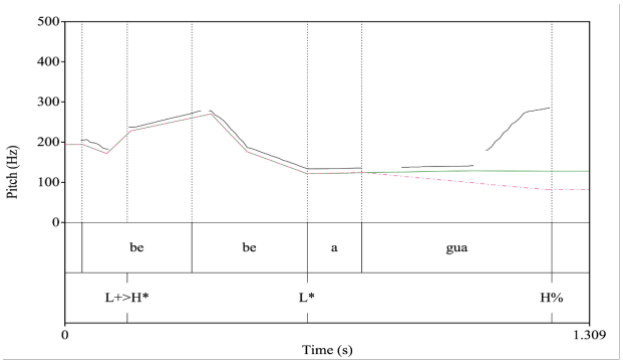


Figure 2. Manipulation from original questions. Pink:falling boundary tone; Green: baseline position; Black=the original contour.

The first manipulation of AS questions was done by deleting the peak of the HL% ending in M% (AS-MQB). The second manipulation was conducted in the same way as it was done in the Peninsular Spanish stimuli. First, the height of L* was taken into account as a reference and afterwards, the boundary tone was lowered 40 Hz, creating a L% (AS-MQL). Pitch was lowered by 40 Hz for both speakers (PS and AS) to have the same tonal difference from the last lowest point before the start of the rise of the final boundary tone.

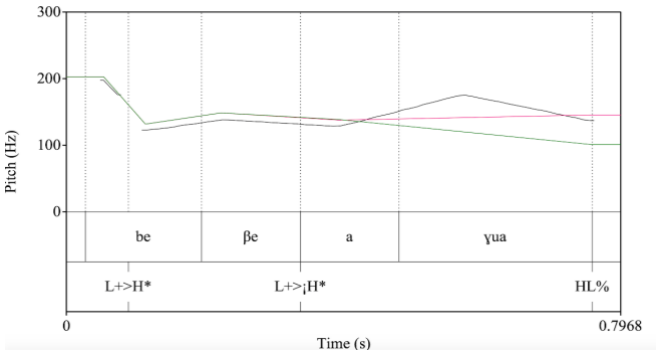


Figure 3: Manipulation from original questions. Pink: the contour after the circumflex was deleted; Green falling boundary; Black: the original contour.

3.3. Task

The study was conducted by using the online platform Qualtrics. The participants heared the stimuli, one at a time, and only once. After listening to each token, they were asked if the sentence was a question or not by pressing a button on the screen.

4. PREDICTIONS AND RESULTS

4.1. Perception of sentences from Peninsular Spanish

4.1.1. Predictions

P1: *OQ as questions and OS as statements*. Questions end with a H% and statements in a L%. As these patterns align with the intonation patterns of the L2 participants' L1 (AE), L2 listeners should accurately perceive original statements as statements and original questions as questions.

P2: *MQL as statements and MQH as questions*. L2 speakers will be guided by the manipulated boundary tone when interpreting the original statements that were manipulated to have a H% boundary tone and the manipulated questions with the L% tone. This may lead to the perception of manipulated statements as questions and manipulated questions as statements.

P3: *MQB as questions and MSB as statements.* In the absence of the primary cue, the prediction for questions and statements that end in the baseline is listeners will rely on a secondary cue and will perceive statements as statements and questions as questions.

4.1.2. Results

The results are shown in Figure 4. The performance of the L2 speakers was compared to the performance of the L1 speakers of PS.

P1: *OQ as questions and OS as statements.* This prediction was confirmed for original statements and to lesser extent for questions.

P2: P3: *MQL as statements and MQH as questions.* This prediction was confirmed for original statements and questions.

P3: *MQB as questions and MSB as statements*. This prediction was only partially confirmed for manipulated statements by the PS group. Only a few participants appear to have used other secondary cues as markers of interrogativity.

Lastly, we conducted t-tests using two-sample t-test for proportions to test whether L2 performed significantly different from the PS group in each sentence. For all the sentence types, expect for OS, the p-value was smaller than .05. This shows that PS and L2 perceived PS intonation contours in different ways, including the original questions.

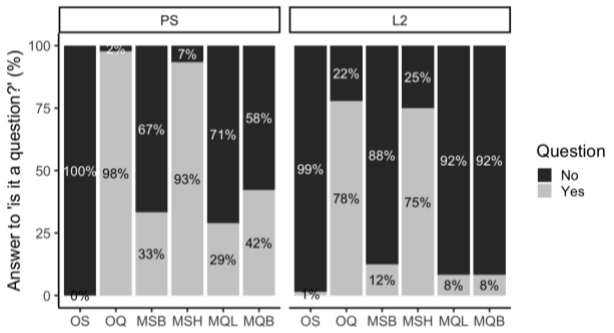


Figure 4: Perception of PS stimuli. Percentage of tokens identified as questions and statements.

4.2. Perception of sentences from Argentinian Spanish.

4.2.1. Predictions

P1: *Varation perceiving OQ*. Questions end in HL% boundary tone and since this is not part of the listeners L1, they will show variation in their responses.

P2: *MQL as statements*. Manipulated questions ended in L% will be perceived as statements, since the fimary cue is the boundary tone.

P3: *MQB as question*. L2 speakers will use secondary cues to perceive manipulated questions ended in M% as questions in the absence of the primary cue.

4.2.2. Results

The results are shown in Figure 5. The performance of the L2 speakers was compared to the performance of the L1 speakers of AS.

P1: *Varation perceiving OQ*. This prediction was not confirmed because L2 speakers interpreted the original Argentinian questions as questions.

P2: *MQL as statements.* It was confirmed as L2 speakers were generally sensitive to the final L% boundary tone and perceived the manipulated questions as statements.

R3: *MQB as question*. The was confirmed. They identified the manipulated sentences as questions in the majority of cases.

Again, we conducted T-tests for group comparisons. All the p-values were larger than .05 Thus, we can say AS and L2 perceived AS intonation contours in similar ways.

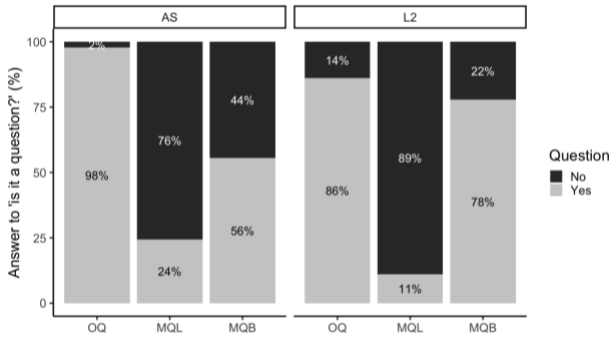


Figure 5: Results of AS sentences. Percentage of tokens identified as questions.

5. DISCUSSION AND CONCLUSION

Overall, this study has shown that L2 speakers relied on universal tendencies to interrogation when perceiving the sentences. They perceived stimuli with a H% boundary tone as questions and stimuli with a final L% boundary tone as statements. Moreover, final stimuli with a high tone on the last stressed syllable and a final circumflex were also perceived as questions even if in the participants' native language this contour is not usually found in such contexts.

Regarding the first research question, original statements and questions from PS were perceived quite accurately. In the case of declarative questions, accuracy was somewhat lower, and their performance was significantly different from the control group, but the majority of participants perceived the unmanipulated interrogative sentences as questions. In the case of the perception of the Argentinian utterances, addressing the second research question, all the groups showed sensitivity to the presence of the circumflex contour and to the high tone at the end of the utterance. In fact, AS questions were perceived slightly more accurately than the PS questions and L2 speakers performed similarly to the control group. Therefore, we may conclude that the main cue to interrogation is the presence of a high tone at the end of utterance, regardless of the final low boundary tone.

Regarding the third research question, in general, L1 and L2 speakers of Spanish appeared to have relied on the same cues when interpreting the declarative sentences. Natural stimuli ending either in a high boundary tone or an upstepped circumflex tended to be interpreted as a questions and stimuli with a final falling contour were mostly interpreted as statements. Manipulation of the final contour resulted in more ambiguous stimuli.

This project has contributed to understanding how the intonation of the declarative questions and statements of Spanish differing solely in the prosody. It also showed how intonation is perceived by L2 speakers compared to L1 Spanish speakers. In order to create stimuli, two variants of Spanish were chosen: Argentinian and Peninsular.

The main difference between the two variants is in the boundary tone of the questions, H% for PS stimuli, and HL% tone for AS questions. Since the latter boundary tone is not usually found in American English, it can be challenging to acquire for L2 speakers, as was shown in Trimble [5]. However, this study showed different results, since L1 and L2 speakers interpreted HL% in a similar way.

Both the Venezuelan questions in Trimble [9] and our Argentinian questions end in HL%, but these two studies show different results. We surmise that the difference may be due to differences between Venezuelan and Argentinian interrogative circumflex contours. It is likely that the relative height of the last accentual peak before the final fall conditions perception of interrogativity for L1 English speakers. Alternatively, the relative height of the first peak in the sentence may also play a role (Face [19]).

6. ACKNOWLEDGMENTS

I would like to thank those who have guided and given me feedback in the processes of completing this study, specially to José Ignacio Hualde. Moreover, I am greateful for those who participated willingly and for the two speakers who helped me create the stimuli.

7. REFERENCES

[1] Ohala, J. J. 1984. An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, *41*(1), 1-16.

[2] Henriksen, N. C. 2012. The intonation and signaling of declarative questions in Manchego Peninsular Spanish. *Language and Speech*, *55*(4), 543-576

[3] Gussenhoven, C. 2002. Phonology of intonation. *Glot International*, *6*(9/10), 271-284

[4] Murphy, K. 2013. Universals and variation in question intonation: A comparative study of Hawaiian and HCE speech melodies. *Proc. Austin Symposium about Language and Society,* University of Texas.

[5] Chen, A. 2005. *Universal and language-specific perception of paralinguistic intonational meaning*. Utrecht.

[6] Grabe, E. 2002. Variation adds to prosodic typology. *Proc. International Conference Speech Prosody,* Aix-en-Provence.

[7] Gunlogson, C. 2002. Declarative questions. In *Semantics and linguistic theory* (Vol. 12, pp. 124-143).

[8] Face, T. 2007. The role of intonational cues in the perception of declaratives and absolute interrogatives. Castilian Spanish. *Estudios de fonética experimental*, 186-225.

[9] Trimble, J. 2013. *Acquiring variable L2 Spanish intonation in a study abroad context* [Doctoral dissertation, University of Minnesota]

[10] Navarro Tomás, T. 1944. *Manual de entonación española*. Madrid: Guadarrama, 1974.

[11] Prieto, P., & Roseano, P. 2010. *Transcription of intonation of the Spanish language*. Munich: Lincom Europa.

[12] Sanchez, C. 2020. *The Production and Perception of Subject Focus Prosody in L2 Spanish*. [Doctoral Dissertation, University of Massachusetts].

[13] Zárate-Sández, G. 2015. *Perception and production of intonation among English-Spanish bilingual speakers at different proficiency levels*. [Doctoral dissertation, Georgetown University].

[14] Gabriel, C., & Kireva, E. 2014. Prosodic transfer in learner and contact varieties: Speech rhythm and intonation of Buenos Aires Spanish and L2 Castilian Spanish produced by Italian native speakers. *Studies in Second Language Acquisition*, *36*(2), 257-281.

[15] Yuan, C., González-Fuente, S., Baills, F., & Prieto, P. 2019. Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition*, *41*(1), 5-32.

[16] Brandl, A., González, C., & Bustin, A. 2020. The development of intonation in L2 Spanish: A perceptual study. *Hispanic linguistics: Current issues and new directions*, 12-31.

[17] Henriksen, N. C., Geeslin, K. L., & Willis, E. W. (2010). The development of L2 Spanish intonation during a study abroad immersion program in León, Spain: Global contours and final boundary movements. *Studies in Hispanic and Lusophone linguistics*, *3*(1), 113-162.

[18] Zárate-Sández, G. 2018. Production of final boundary tones in declarative utterances by English-speaking learners of Spanish. In Proceedings of the 9th International *Conference on Speech Prosody. International Speech Communication Association (ISCA) Online Archive* (pp. 927-31).

[19] Méndez Seijas, J. 2019. *Second Language Spanish Intonation: Systemic and Realizational Dimensions of Its Acquisition*. Georgetown University.

[20] Nibert, H. J. 2005. The acquisition of the phrase accent by intermediate and advanced adult learners of Spanish as a second language. *Proc. 6th Acquisition of Spanish and Portuguese as first and second languages* Somerville, 108-122.

[21] Casillas, J. V., N. Rodríguez, L. F. Arroyo, J. J. G. Pozú, K. Parrish, R. Esposito, K. Gómez, and I. 2022. In Chang *Using intonation to disambiguate meaning: The role of empathy and proficiency in L2 perceptual development*. [Under review]

[22] Boersma, P.& D. Weenink, 1992-2007. Praat: doing phonetics by computer. http://www.praat.org/

SCHWA IN FRENCH OBSTRUENT-LIQUID CLUSTERS BY NATIVE SPEAKERS AND MANDARIN LEARNERS

Minmin Yang, Rachid Ridouane

Laboratoire de Phonétique et Phonologie (UMR7018, CNRS/Sorbonne Nouvelle, Paris)
minmin.yang@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

ABSTRACT

This study investigates the production of French obstruent-liquid clusters in word-initial and -final positions by 5 native French speakers and 10 Mandarin speakers learning French. The frequency of schwa within these clusters and its temporal and formant characteristics are analyzed based on word position, liquid type, obstruent place of articulation, and voicing. Results show that schwa is most frequent in clusters containing a rhotic liquid for both groups. Additionally, both word position and voicing have a significant impact on the rate of schwa occurrence and on its duration. Formants of schwa vary according to position, place of obstruent articulation, and liquid type. A notable difference was found between natives and learners in terms of word-final schwa production, with native French rarely producing it and Mandarin learners producing it more frequently. It is suggested that while schwas within consonant clusters may have phonetic origins, word-final schwas are driven by phonological factors.

Keywords: Epenthetic schwa, French consonant clusters, Mandarin speakers, duration, quality

1. INTRODUCTION

Consonant sequence acquisition in non-native languages has been extensively studied over the years, with a particular focus on how learners repair illegal consonant sequences [1, 2, 3, 4]. There are various strategies learners can use to do this, including deletion, insertion, and substitution, with vowel insertion being the most common [1, 5]. This type of insertion, which often involves adding a schwa vowel between two consonants, has traditionally been seen as a phonological process used to repair an inappropriate syllable structure. However, the Articulatory Phonology perspective suggests that the inserted schwa may actually result from reduced gestural overlap between the surrounding consonants due to articulatory constraints [6, 7].

The current study investigates the insertion of the vocalic elements in the production of French obstruent-liquid (OL) sequences by Mandarin speakers learning French and compares it to the production of native French speakers. The focus is on the occurrence of these sequences at word-initial and word-final positions. Mandarin prohibits consonant sequences, including OL [4,8], while French has a high frequency of consonant clusters in various positions within words, with OL being the most common sequence [9].

The insertion of vocalic elements within clusters has been documented in various languages around the world [10]. However, there have been relatively few studies on schwa intrusion in French consonant clusters. [11] conducted an acoustic analysis of Quebec French and Argentinian Spanish OL clusters and found that epenthetic vowels were commonly used in French. The frequency of this epenthesis depended on the type of liquid and the similarity of the cluster members in terms of manner, place, and voicing, with liquid type and voicing having the greatest influence. Epenthetic vowels were observed in almost all voiced obstruent-rhotic clusters for all speakers. [12] also examined the insertion of schwa in OL clusters in Quebec French and found fewer, but longer schwas compared to [11]. They compared the spectral characteristics of schwa and the following vowel and found similar F1 and F2 values, suggesting a vowel harmony process between the two vowels. Unlike [11], [12] did not view schwa insertion as a means of simplifying clusters, but rather tentatively attributed it to articulatory or perceptual constraints specific to these OL clusters.

To the best of our knowledge, no study has previously compared the production of schwas in French obstruent-liquid clusters by native French speakers and Mandarin speakers learning French. The current study aims to address this gap. Specifically, our objectives are to 1) compare the frequency, temporal characteristics, and formant frequencies of schwa in the two groups, and 2) analyze how the quantitative and qualitative characteristics of schwa vary across word positions, liquid types, and obstruent place and voicing.

2. METHOD

2.1. Participants

The production experiment included five native French speakers and ten Mandarin speakers learning

French. The native French speakers (F) were female university students with a mean age of 22.2. The Mandarin learners of French were divided into two groups: a beginner level group of five subjects with a mean age of 18 (M1) and an advanced level group of five subjects with a mean age of 20.5 (M2). All of the Mandarin learners of French were female university students studying French at Leshan Normal University in China.

2.2. Stimuli

The stimuli used in the study were 20 French words containing the target OL clusters (Table 1). Mostly common words were included, but some were rare (e.g., *blase* 'name', or *bagle* 'computer worm'), and their impact was not examined. The stimuli were selected to control for the type of liquid (/l/ or /ʁ/), the place and voicing of the obstruent, and the position of the cluster within the word. All onset clusters were followed by the vowel /a/. The stimuli were presented in isolation and within a carrier sentence, but only the isolated words were analyzed for the purposes of this study. Each word was repeated seven times in a randomized order, resulting in a total of 2100 tokens.

	#CC	CC#		#CC	CC#
/pr/	Prague	câpre	/br/	brade	sabre
/tr/	trappe	quatre	/dr/	drague	cadre
/kr/	crabe	sacre	/gr/	grâce	tigre
/pl/	place	Naples	/bl/	blase	câble
/kl/	classe	tacle	/gl/	glace	Bagle

Table 1: Lexical items used for the experiment.

We recorded the 15 subjects using a Roland Rubix24 sound card and AKGC 520L microphone in a soundproof room at the Laboratoire de Phonétique et Phonologie in Paris (CNRS/Sorbonne Nouvelle) for the French subjects and at Leshan Normal University for the Mandarin speakers.

2.3. Segmentation

The acoustic data were analyzed using Praat [13] software. The determination of the presence of schwa within an OL cluster was not always straightforward and required a two-step process. A first annotator initially labeled all instances of non-ambiguous schwas and marked the ambiguous cases. These cases were then reviewed by two annotators, and a final decision was made on the presence or absence of schwa based on a combination of cues. The annotators considered as evidence of the presence of schwa an interval between two consonants with periodic vibrations and a local increase in signal energy at the consonant release, and/or an interval after the consonant release with formant structure or energy in the F2/F3 region. These criteria for schwa identification are consistent with previous research (e.g., [14,15]). In the example presented in Fig. 1, the presence of schwa was determined based on the increase of signal amplitude after the stop release and an interval with a formant structure within the OL cluster.

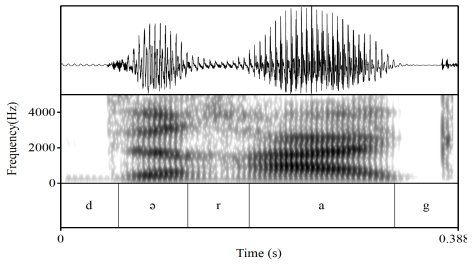


Figure 1: Waveform and spectrogram of the word <drague> 'dredger' with schwa between /d/ and /ʁ/.

We used Praat to analyze target clusters in three different contexts: within onset cluster (*onset*, e.g., [dəʁag]), within coda cluster (*coda1*, [takəl]), and after coda cluster (*coda2*, e.g., [baglə]). We extracted the duration and the formant frequencies for each inserted schwa. Only the formant frequency values obtained from the middle of the schwa are reported, despite measurements being taken at the beginning, middle, and end. We compared the presence/absence, duration, and formant structure of schwa between the group of native French speakers and the two groups of Mandarin learners, considering variations in word position, liquid type, and obstruent place and voicing.

3. RESULTS

Fig. 2 illustrates the rate of occurrence of schwa in the production of natives and the two groups of learners. There was a variation in the occurrence of schwa across the three groups, with a higher rate for M1 (79%) and M2 (67%) compared to F (35%).

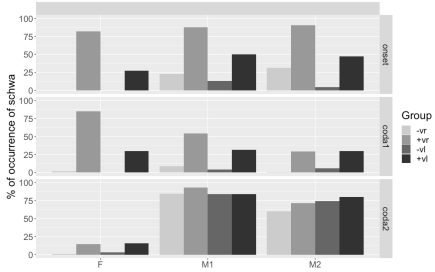


Figure 2: Percentage of occurrence of schwa for French (F) and Mandarin learners (M1 & M2) (-/+v = -/+voiced).

The frequency of schwa varied among the three groups based on several factors, including the voicing

of the obstruent and the type of liquid. The highest frequency occurred when the obstruent was voiced and the liquid was /ʁ/, with rates of 60% for native French speakers, 78% for beginner Mandarin learners, and 64% for advanced Mandarin learners. Additionally, the frequency of schwa differed based on word position, with Mandarin learners producing more schwas in coda2 (79%), compared to coda1 (21%) and onset (46%), while French speakers showed the opposite pattern (8% for coda2, 32% for coda1, and 30% for onset). The place of articulation of obstruents had a less important impact, although there was a trend towards /bl/ sequences attracting fewer schwas in onset and coda1.

In the following sections, we will explore the temporal and spectral properties of schwa sounds as produced by native French speakers and Chinese-speaking learners. For the purposes of this analysis, we will focus specifically on clusters containing voiced obstruents, as clusters with voiceless obstruents exhibit very few instances of schwa in the speech of French native speakers. The data were analyzed using the ANOVA function in R [16].

3.1. Schwa in native speakers' productions

3.1.1 Durational properties of schwa

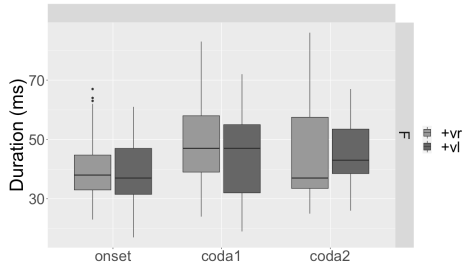


Figure 3: Schwa duration in different word positions for French speakers.

As shown in Fig. 3, the duration of schwa varied depending on its position within a word. The shortest duration was observed in the onset position, with values of 39 ms, 47 ms, and 46 ms for onset, coda1, and coda2, respectively. The Tukey HSD test showed that these differences were significant between onset and coda1 ($p<0.001$) and between onset and coda2 ($p<0.01$). In contrast, the manner of articulation of liquids and the place of articulation of obstruents did not have a significant impact on schwa duration.

3.1.2. Formant frequencies of schwa

Fig. 4 and Fig. 5 present the formant frequencies for the different clusters. The analyses revealed that word position significantly impacted F1, F2, and F3 (all $p<0.001$). Specifically, coda2 had a higher F1 and lower F2 and F3 compared to the other positions. The type of liquid also significantly influenced all three formants, with /r/ clusters exhibiting higher F1, lower F2, and higher F3 than /l/ clusters (all $p<0.001$).

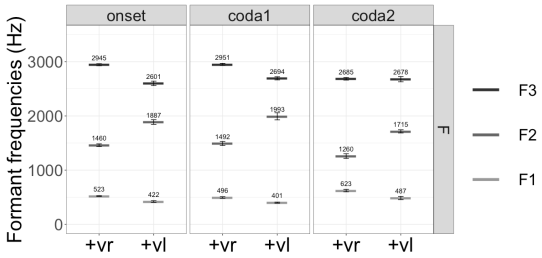


Figure 4: Schwa formant frequencies in different contexts of French speakers.

The place of articulation of the obstruent also significantly affected F1, F2, and F3 (all $p<0.001$). /gr/ and /gl/ clusters had lower F1 than /br/ and /dr/, and /bl/. F2 increased as the place of articulation moved further back in the vocal tract, and F3 displayed higher frequencies in /dr/ compared to /br/ and /gr/ ($p<0.001$).

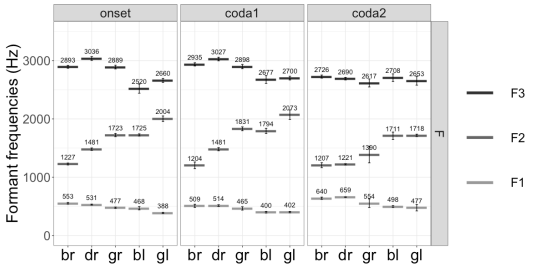


Figure 5: Schwa formant frequencies in different clusters for French speakers.

3.2. Schwa in Mandarin learners' productions

3.2.1 Durational properties of schwa

As for French native speakers, the duration of schwa varied depending on its position (Fig. 6). But unlike French, the coda2 position had the longest mean duration (89 ms), while onset and coda1 displayed similar durations (45 ms and 47 ms, respectively).

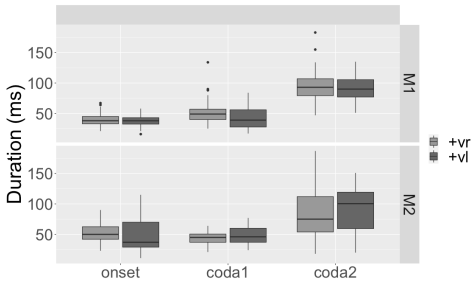


Figure 6: Schwa duration in different word positions for Mandarin speakers.

Statistical analysis revealed the effect of position on schwa duration (F (2,669) =280, p<0.001). The Tukey HSD test showed significant differences in duration between coda2 and onset (p<0.001) and between coda2 and coda1 (p<0.001). The type of liquid did not have an effect on schwa duration, but significant differences were found in relation to place of articulation (F (2,442) =4, p<0.05), with /gr/ being longer than /br/ (p<0.05) and /dr/ (p<0.05). Significant differences were also observed for /gl/ which is longer than /bl/ (p<0.01).

3.2.2. Formant frequencies of schwa

Fig. 7 and Fig. 8 present the frequencies for the different clusters. The F1 of schwa was affected by word position (F (2, 669) = 6, p<0.001), with coda1 displaying a lower F1 than onset (p<0.001) and coda2 (p<0.05). The place of articulation and type of liquid did not have a significant effect. F2 was also affected by position (F (2, 669) = 14, p<0.001), with schwas having higher F2 in onset compared to coda1 and coda2 (all p<0.001). Place of articulation of the obstruent also had an effect (F (2, 442) = 44, p<0.001), with significant differences found between /br/ and /gr/ and /dr/ (p<0.001). Similar to French, schwa in the context of /r/ displayed lower F2. F3 was marginally affected by word position (F (2, 669) = 4, p<0.05), type of liquid (F (1, 669) =24, p<0.001), and place of articulation (F (2,442) = 4, p<0.05).

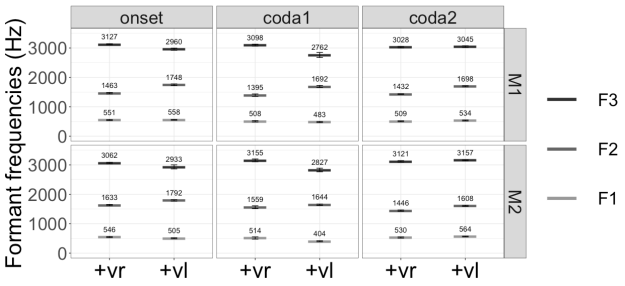


Figure 7: Schwa formant frequencies in different word positions for Mandarin speakers.

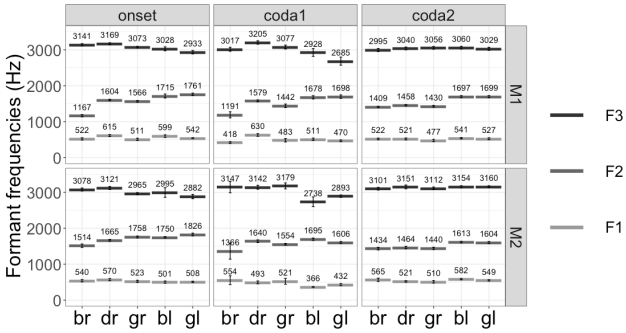


Figure 8: Schwa formant frequencies in different clusters for Mandarin speakers.

4. DISCUSSION AND CONCLUSIONS

This study explored French obstruent-liquid clusters produced by French native speakers and two Chinese learner groups, focusing on schwa presence and characteristics. Our findings revealed that all three groups produced schwas in the clusters we analyzed, but with varying frequencies. As expected, Chinese learners used schwas at high rates, with 79% of beginner learners and 67% of advanced learners inserting them. It was also notable that native French speakers frequently used schwas, though at a lower rate. This suggests that the presence of these vocalic elements may be a common feature in the production of these consonant clusters, regardless of the speaker's language background. Notably, the highest rate of schwa production occurred for both native French speakers and Chinese learners when the obstruent was voiced and the liquid was /ʁ/. The frequent occurrence of schwa in this position prompts the question of its origin. From the perspective of Articulatory Phonology [17], this vocalic element may be seen as a transitional vocoid that does not have its own distinct articulatory gesture, but rather is a result of the way in which the consonant gestures are phased and articulated. The fact that the spectral characteristics of schwa are very much influenced by the place of articulation of the obstruent and the manner of articulation of the liquid may be considered as evidence that schwa in this context does not have its own distinct articulatory gesture.

The fact that schwas are more frequent in /Cr/ compared to /Cl/ may be due to differences in the way the gestures associated with these consonants overlap. /Cr/ sequences tend to have a more sequential articulation of gestures with no overlap, while laterals tend to have overlapping and blending gestures [18]. There was a higher prevalence of schwa insertion at word-finally among Mandarin learners. Only 8% of schwas were produced in this way by native speakers, a figure that is similar to the 6% reported in [19] which analyzed the production of French CR clusters in a corpus of about four hours of speech. Word-final schwas in Mandarin speech can be explained by the prohibition of codas (except /n/ and /ŋ/) and coda clusters. Some suggest that the presence of schwa in this position in French may be due to the need to retain the cluster by resyllabifying it as a complex onset [20]. The results of this study can guide future experiments with more speakers and consonant clusters to expand our understanding of the topic.

5. ACKNOWLEDGEMENTS

This study was supported by the French ANR (ANR-10-LABX-0083, Labex-EFL).

6. REFERENCES

[1] Davidson, L. 2006a. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34(1), 104-137.

[2] Davidson, L. 2006b. Phonotactics and articulatory coordination interact in phonology: Evidence from nonnative production. *Cognitive Science 30*(5), 837-862.

[3] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Mehler, J. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance* 25(6), 1568.

[4] Zhao, X., Berent, I. 2016. Universal restrictions on syllable structure: Evidence from Mandarin Chinese. *Journal of Psycholinguistic Research* 45(4), 795-811.

[5] Gafos, A. I. 2002. A grammar of gestural coordination. *Natural language & linguistic theory* 20(2), 269-337.

[6] Browman, C. P., Goldstein, L. 1990. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics* 18(3), 299-320.

[7] Davidson, L., Stone, M. 2003. Epenthesis versus gestural mistiming in consonant cluster production: an ultrasound study. *Proc. of the West Coast Conference on Formal Linguistics* UC San Diego, 165-178.

[8] Duanmu, S. 2007. *The phonology of standard Chinese*. OUP Oxford.

[9] Malécot, A. 1974. Frequency of occurrence of French phonemes and consonant clusters. *Phonetica* 29(3), 158-170.

[10] Hall, N. 2006. Cross-linguistic patterns of vowel intrusion. *Phonology* 23(3), 387-429.

[11] Colantoni, L., Steele, J. 2005. Phonetically-driven epenthesis asymmetries in French and Spanish obstruent-liquid clusters. In: Gess, R., Rubin, E. J. (eds), *Theoretical and Experimental Approaches to Romance Linguistics: Selected papers from the 34th Linguistic Symposium on Romance Languages (LSRL), Salt Lake City, March 2004*. Amsterdam: John Benjamins, 77–96.

[12] Lancien M., Côté M-H. 2019. Intrusive vowels preceding /R/ in Quebec French. *Proceedings of the 19th International Congress of Phonetic Sciences* Melbourne, 3473-3477.

[13] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.02, http://www.praat.org/.

[14] Ridouane, R., Fougeron, C. 2011. Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology* 2(2), 275-300.

[15] Kirby, J. 2014. Acoustic transitions in Khmer word-initial clusters. *Proceedings of the 10th International Seminar on Speech Production* Cologne, 234-237.

[16] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[17] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4), 155-180.

[18] Bradley, T. G. (2004). Gestural timing and rhotic variation in Spanish. *Laboratory approaches to Spanish phonology* 7, 195.

[19] Guan, Q., Wu, Y., Chitoran, I. 2022. A corpus-based study of /CR/ and /RC/ clusters in French: Prosodic and segmental effects. *Speech Prosody 2022* Lisbon, 244-248.

[20] Noske, R. 1996. Is French optimal? A question concerning phonological process order. In: Durand, J., Bernard, L. (eds.), *Current Trends in Phonology*. CNRS, Université Paris-X & University of Salford: University of Salford publications, 485-507.

REVISITING INTELLIGIBILITY: THE ROLE OF LISTENERS' L1 AND PROFICIENCY IN PERCEIVING CHINESE EFL LEARNERS' SPEECH

Yating Cao[1,2] Hua Chen[1]

[1]Nanjing University [2]University of Warwick
yatingcao@smail.nju.edu.cn; chenhua@nju.edu.cn

ABSTRACT

This study examines how listeners' first language and English proficiency level affect the intelligibility of Chinese EFL learners' oral production. Native English speakers, intermediate-level Chinese EFL learners, and high-level Chinese EFL learners dictated 18 sentences, which were analyzed through both form- and meaning-based understanding and interview. Results showed that the accuracy of both form- and meaning-based comprehension by the intermediate-level Chinese EFL listeners was significantly lower than that of high-level Chinese EFL listeners and native English speakers, but there was no significant difference between the latter two groups. Further, the type and position of segmental features and stressed information were crucial for all the three groups; speech rate had a complex relationship with intelligibility and significantly affected intermediate-level Chinese EFL learners; However, native English listeners were more sensitive to tone. The findings will offer insights into teaching English pronunciation and improving the communicative ability of L2 learners.

Keywords: intelligibility, L1 background, English proficiency level

1. INTRODUCTION

As English has been used as a Lingua Franca, communication between non-native speakers has become more frequent [1] and non-native speakers could reach mutual intelligibility even with a perceptible foreign accent [2]. Therefore, the traditionally-advocated nativeness principle has been shifted to the intelligibility principle, which sets more realistic teaching and learning goals and prioritizes mutual intelligibility [3].

[4] conceptualized intelligibility as *accentedness*, *comprehensibility,* and *intelligibility*, referring to the degree of deviation from specific phonetic standards, the perceived degree of difficulty of understanding speech, and the actual understanding of utterances, respectively. Studies have established that accented speech does not necessarily lower intelligibility, but there is still dispute over which characteristics and how they affect intelligibility [5-6]. [2] proposed the Lingua Franca Core model, which listed some segmental features as the priority to maintain mutual intelligibility. Other research found that prosody, such as stress [7-8], tone choices [9], and speech rate [10], is also essential to intelligibility. Furthermore, recent studies have begun to examine the contribution of various linguistic features to intelligibility using similar methods [e.g., 5-6, 11]. Speakers were asked to deliver an impromptu speech describing an eight-frame image, and then their speeches were evaluated by listeners to further investigate which linguistic characteristics were associated with accentedness and comprehensibility, respectively.

The listener factor is also critical to intelligibility [12]. Interlanguage Speech Intelligibility Benefit Hypothesis (ISIB) developed by [13] believed that L2 listeners have an advantage in understanding speech by speakers sharing the same language background, which has gained empirical support from quantitative [14] and qualitative analyses [6, 11].

However, there are still some gaps to fill. First, more studies on the role of the listener factor are called. On the one hand, previous research has primarily depended on the perceptions of native English speakers, which is inconsistent with the fact that non-natives constitute the bulk of international communication. On the other hand, results regarding the ISIB hypothesis that whether listeners and speakers sharing the same L1 have advantages in understanding are conflicting [15]. It is assumed that the function of proficiency level accounts for this inconsistency [16]. [17] reported that only the low-proficient listeners recognized low-proficient accented speech better than native English speakers did when speakers and listeners shared the same first language. However, this aspect has not received due attention [18].

Second, research should reconsider the methods concerning measuring intelligibility. The research conducted by the Trofimovich team [5-6, 11] equated *comprehensibility* and *intelligibility* by using the self-reported difficulty of comprehending L2 speech, disregarding whether listeners understood the speech or not. Further, previous studies have frequently adopted the traditional word-by-word transcription to calculate the accuracy [e.g., 19, 20]. In fact, this method is more likely to emphasize listeners' form-based comprehension, and more importantly, the large proportion of function words included in the analysis plays a limited role in meaning construction [21]. In contrast, fewer studies have used idea-unit as

a standard to examine meaning-based comprehension. Such method could evaluate listeners' understanding of the key information of utterances to better examine intelligibility-related features [22]. Therefore, it is necessary to combine both form-based and meaning-based comprehension to better reflect the concept of intelligibility. Besides, this line of research usually adopts the quantitative analysis, while the qualitative evidence of how listeners perceive L2 speech and their feelings is limited, which is called for more studies [6, 11].

Therefore, this study attempts to investigate the role of both listeners' L1 and English proficiency level in the intelligibility of Chinese EFL learners' accented speech to further testify the ISIB hypothesis. Also, combining both quantitative data from form- and meaning-based comprehension and qualitative interview data will advance the understanding of speech intelligibility.

2. METHODS

2.1. Data source

60 online audio clips of English public speech given by Chinese college students were selected as the candidate data source. All of these speeches were persuasive speech [23], and topics including culture, sports, and education and others related to college life. Three experts in L2 phonetic acquisition and spoken English teaching were invited to rate the accentedness and comprehensibility via a 100-point scale (1=very accented/very difficult to understand, 100 = no accent/very easy to understand). And the inter-rater reliability indicated that expert raters achieved a high degree of agreement (*Accentedness*: Cronbach's α = .821, *Comprehensibility*: Cronbach's α = .834). Then, clips were ranked based on their mean ratings of accentedness and comprehensibility, respectively, with the first 20 marked as less accented/more comprehensible and the last 20 being more accented/ less comprehensible. Taking the two rankings, syntactic structures, prosodic features, and gender balance of each audio clip into account, the current study chose 18 clips as the final data source (7.85s/sentence; 17.87 words/sentence), among which 15 clips matched in terms of their accentedness and comprehensibility ranks, and three clips did not match.

2.2. Listeners

This study included 29 Chinese college students and 10 native English speakers (NS). In particular, Chinese college students were from the same university and comprised 17 high-level English learners (HL) and 12 intermediate-level English learners (IL). The HL group was either senior English majors or had excelled in English speech contests at the national level. Also, they received high scores on standardized English tests (TOEFL: M = 114.4, SD = 1.41; IELTS: M =7.7, SD = 0.27). By contrast, the IL group were freshman non-English majors. They were classified as intermediate level in the university displacement test, and had no prior experience with public speech contests nor taking any standardized English tests before. For the NS group, they were all international exchange students (from UK) of the same university. They lived in China for only one month when the experiment was carried out. No hearing or any other health issues were reported.

2.3. Procedure

All the participants were instructed to listen to the 18 audio clips in a random order and write down what they heard with a 30-second break every five clips. Every clip was played four times with a 10-second interval. Before the experiment, participants were shown three identical audio clips that were not included in the analysis to familiarise them with the procedures. This task lasted for 30 minutes.

After the sentence dictation task, 22 listeners were randomly selected to the semi-structured interview (9 HL, 7 IL, 6 NS). The questions included but were not limited to the following:

(1) What's the meaning of this sentence?
(2) What's your biggest difficulty during this task?
(3) Why did you fail to write down this part?

2.4. Analysis

The quantitative analysis includes the accuracy of both form- and meaning-based comprehension. For form-based comprehension, it is the percentage of words correctly transcribed as traditional method did [19, 20]. And meaning-based comprehension refers to the percentage of idea-units correctly written down. [24] believed that idea-unit is related to intonation unit and pause of the speech. Considering the possibility of improper use of pause and/or intonation unit by Chinese EFL learners, the present study also considered the syntactic structure as proposed by [25] when identifying idea-units. In total, 89 idea-units were identified. For instance:

(1) For most ordinary people // the lives of the famous stars // are very mysterious

The above sentence contains 13 words and can be divided into three idea-units. This part of analysis was conducted by the first author and then a Ph.D. student in linguistics analyzed 30% of randomly-selected data. And the Cohen's Kappa value (k=0.85) indicated the high inter-annotator reliability. Then, the between-group comparison was conducted.

For the qualitative analysis, all the interview data were transcribed and then classified according to different linguistic features and reasons mentioned by participants. Then, we analyzed the impact of certain phonetic features on intelligibility and the effects of listener factors.

3. RESULTS AND DISCUSSION

3.1. The effect of listener factors on the form and-meaning-based comprehension

It can be seen in Table 1 that the form-based comprehension accuracy of the three groups exceeded 70%, among which the HL and NS groups reached 93%, and the meaning-based comprehension accuracy exceeded 70%. It suggested that the overall intelligibility of Chinese EFL learners' speech is acceptable, corroborating [20]'s findings. Moreover, the one-way ANOVA test found that the group had a significant main effect ($F_{form}(2, 701) = 38.88, p<.001$; $F_{meaning}(2, 701) = 27.46, p<.001$). A further Tukey HSD post-hoc test revealed that the accuracy of two indices of the IL group was significantly lower than that of the HL and NS group ($p<.0001$), but that there was no significant difference between the latter two groups ($p=.963$).

Table 1: The accuracy of form- and meaning-based comprehension.

	HL	IL	NS
Form-based comprehension	93.6 (10.98)	84.02 (17.58)	93.73 (10.95)
Meaning-based comprehension	87.1 (20.25)	74.45 (24.72)	87.62 (17.92)

First, the results revealed that the listeners' first language background did affect the intelligibility of Chinese EFL learners' speech. Regardless of form- or meaning-based comprehension, the accuracy of NS group was higher than the other two Chinese listener groups, which appears to contradict the shared-L1 advantage observed in earlier research [e.g., 6, 11, 13]. The discrepancies in research designs may account for these contradictory outcomes. Specifically, [13] used the phoneme identification task to test intelligibility, and [6, 11] used the subject ratings to explore accentedness and comprehensibility, without tapping into the third aspect of intelligibility (i.e., the actual understanding of utterances). However, the present study focused on listeners' real understanding, which was involved with phoneme identification, spoken word recognition, contextual information, etc. This task placed a greater demand on listeners' linguistic ability. Thus, the interlanguage speech intelligibility benefit might be weakened.

The present study also provided additional evidence for extending the ISIB hypothesis [13] that listeners' proficiency level was an important factor. Studies like [26] have shown that the more exposure to particular sounds, the higher listeners' sensitivity and general perceptual adaptability would become. In this study, the high-level group had greater exposure to and higher familiarity with English compared with the intermediate-level group, and their language skills were more comprehensive [21]; therefore, they were more likely to perceive the deviated sounds of the examined speech. This might effectively reduce cognitive load and processing difficulty, which results in better comprehension than the IL group.

Besides, the accuracy of form- and meaning-based comprehension was also compared. Results revealed that all the listener groups achieved significantly lower accuracy in meaning-based comprehension, indicating that listeners did not fully understand the core information of the speech even if they could write down the individual words. The findings may be pertinent to the analysis method. Since form-based comprehension analysis took every individual word into account, some function words, which did not convey actual discourse meaning, were also included. Consequently, its proportion was inflated and did not reflect the real comprehension [27]. Another possibility is related to the listening strategy. [19] found that listeners preferred to employ the bottom-up strategy (i.e., from isolated words to discourse meaning) when perceiving accented speech rather than the top-down approach (i.e., from key information and whole discourse meaning to individual words). And this is particularly the case for non-native English speakers [2]. This evidence also shows the advantages and necessity of adopting idea-unit to measure meaning-based comprehension in this line of research.

3.2. The effect of listener factors on the features safeguarding intelligibility

To further find out which and how phonetic features affected different listener groups, the present study combined both quantitative and interview data.

3.2.1. Segmental features

There were only 19 segmental deviations found throughout 18 sentences, taking up 3.6% of all the phonemes. Further analysis showed that listeners had increased difficulty in deviations of vowels contrasts /æ/-/ɑ:/, /aɪ/-/ʌ/, /æ/-/e/, /ei/-/æ/, /ʌ/-/ɜ:/ and consonant contrasts /θ/-/z/, /k/-/g/, and /l/-/r/, corroborating the findings of prior study on the intelligibility of Chinese EFL learners' speech [19, 20]. Moreover, results also found that listeners relied on the initial syllables to recognize words. That is, listeners might lose the key anchor to search for the mental lexicon when the

phonetic deviation occurs at the initial syllable [28], as listener 23 (IL) said "Sometimes if I can only clearly hear the initial part, I can still guess the meaning, but not vice versa. For example, I heard *pros* first, then I can quickly think of *prosperity*. But for the word *valuable*, the vowel of the initial syllable was very unclear, so I could not search for the word."

The interview data also revealed that more proficient learners might had higher adaptability of the deviated segmental features, which could reduce the processing difficulty and thus improve the intelligibility, as described by listener 6 (HL) that "Even though the speaker did not pronounce *awareness* correctly, I could still understand because the speaker possesses the typical Chinese accent."

3.2.2. Sentence stress

It was found that the stimuli exhibited a typical syllable-timed rhythmic pattern of Chinese EFL learners' speech [19]. Specifically, the speaker in sentence 9 stressed every single word and frequently had a long pause between each word without clear communicative intention. All the listeners considered this sentence as the most difficult one for them to understand. Moreover, listeners reported that they "would rely on the stressed information to predict or organize the discourse; therefore, it is hard to distinguish the major and minor information when speakers emphasized every word" (listener 20, IL).

In fact, the existing literature has shown that sentence stress is vital for speech comprehension, particularly for native English speakers [29]. The present study offers empirical support for the findings from Chinese EFL learners as listeners.

3.2.3. Speech rate

The average speech rate of the current study was 2.17 syllables/s. However, [10] found that the optimal speech rate for accentedness rating was 4.76 syllable/s, and that for comprehensibility rating was 4.23 syllables/s, which was faster than that of current study. Combined with the interview data, it was shown that the IL group was strongly impacted by the fast speech rate, as 15.7% mentioned factors leading to misunderstandings were attributed to speech rate. For instance, listener 25 (IL) said that "I cannot grasp this sentence and it became more difficult to write it down as the speaker spoke faster. So I could only write down what I have heard." Again, the qualitative data triangulated with the quantitative findings that IL group tended to use the bottom-up strategy [2, 19]. Under such situation, the faster speech rate taxed more processing difficulty for them to catch the key information.

Interestingly, the speech rate of sentence 10 was 4.8 syllables/s, which was faster than the average speech rate and the optimal rate found by [10], but the accuracy achieved the ceiling effect. The interview elucidated the reasons, such as "fast yet clear pronunciation" (listener 29, IL), "like a native speaker and I can adapt to that rate quickly" (listener 13, HL), and "maybe a little fast but the pronunciation and stress were perfect to the point" (listener 30, NS). The verbal reports showed the intricate interplay between speech rate and other phonetic features – clear pronunciation and stress served as the basis, and speech rate worked as the further safeguard for intelligibility.

3.2.4. Tone

Despite the fact that 13 out of 18 speakers in the present study overused the falling tone which made the speech monotonous, the quantitative data indicated that such a tone choice did not impair intelligibility. However, results showed that sentence 15 negatively affected native English speakers' understanding, which was also the only sentence where they underperformed than the other two groups. As stated by listener 34 (NS), the tone was "unnatural and abruptly changing", which led to the failure of understanding sentence and speaker's intention.

Such group difference might result from the different sensitivity to tone information. Chinese EFL learners are found to be less sensitive to the intonational changes as influenced by their L1–a tonal language. Moreover, Chinese EFL students are taught intonation through rote memorization, which might influence their production and perception of intonation [30]. Thus, the abrupt changes in tone choices had less negative impact on HL and IL groups.

4. CONCLUSION

This present study extends the ISIB hypothesis by providing empirical evidence of the importance of listeners' proficiency level and offered qualitative data on how certain phonetic features influence listeners' comprehension. Also, the methods of measuring of intelligibility through both form- and meaning-based comprehension merits further research to get a deeper and more valid understanding of intelligibility. Future studies can explore the effects of more linguistic features on intelligibility and include more listeners' and speakers' language backgrounds to advance this line of research.

5. ACKOWNLEDGEMENTS

This study was supported by the National Social Science Foundation of China (20AYY013), and China Scholarship Council (202206190122).

6. REFERENCES

[1] Crystal, D. 2003. *English as a Global Language (2nd ed)*. Cambridge: Cambridge University Press.

[2] Jenkins, J. 2002. A sociolinguistically based empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics* 23, 83–103.

[3] Levis, J. 2005. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39, 369–377.

[4] Munro, M. J., Derwing, T. M. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language Learning* 45, 73-97.

[5] Trofimovich, P., Isaacs, T. 2012. Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15, 905-916.

[6] Foote, J.A., Trofimovich, P. 2018. Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review* 74, 253-278.

[7] Hahn, L. 2004. Primary stress and intelligibility: Research to motivate the teaching of supra-segmentals. *TESOL Quarterly* 38, 201–223.

[8] Field, J. 2005. Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39, 399–423.

[9] Pickering, L. 2001. The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly* 35, 233-255.

[10] Munro, M. J., Derwing, T. M. 2001. Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition* 23, 451-468.

[11] Crowther, D., Trofimovich, P., Saito, K., Isaacs, T., 2015. Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly* 49, 814-837.

[12] Munro, M J. 2008. Foreign accent and speech intelligibility. In: Hansen, J. G., Zampini, M.L. (eds.). *Phonology and Second Language Acquisition*. Amsterdam: John Benjamins.

[13] Bent, T., Bradlow, A. R. 2003. The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America* 114, 1600-1610.

[14] Kennedy, S., Trofimovich, P., 2008. Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64, 459-489.

[15] Munro, M. J., Derwing, T. M., Morton, S. L. 2006. The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition* 28, 111-131.

[16] Xie, X., Fowler, C. A. 2013. Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics* 41, 369-378.

[17] Hayes-Harb, R., Smith, B. L., Bent, T., Bradlow, A. R. 2008. The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetic* 36, 664-679.

[18] Kang, O., Rubin, D., Kermad, A. The effect of training and rater differences on oral proficiency assessment. *Language Testing* 36, 481-504.

[19] Deterding, D. 2010. ELF-based Pronunciation Teaching in China. *Chinese Journal of Applied Linguistics* 33, 3-15.

[20] Zhang, L. L. 2015. An empirical study on the intelligibility of English spoken by Chinese university students. *Chinese Journal of Applied Linguistics* 38, 36-53.

[21] Field, J. 2008. Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly* 42, 411-432.

[22] Winke, P., Gass, S. 2016. Using free recall and idea units for evaluating second language comprehension: Methodological choices and issues.http://newsmanager.commpartners.com/tesolalis/issues/2016-11-04/5.html.

[23] Lucas, S. E. 2020. *The Art of Public Speaking (13th ed)*. Boston: Mc. Graw Hill.

[24] Chafe, W. L. 1998. The deployment of consciousness in the production of narrative. In: Chafe, W.L. (ed.). *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood. NJ: Ablex.

[25] Kroll, B. 1997. Combining ideas in written and spoken English: A look at subordination and coordination. In: Ochs, E., Bennett, T. (eds.). *Discourse across Time and Space.* Los Angeles: Southern California Occasional Papers in Linguistics, 69-108.

[26] Best, C. 1995. A direct realist view of cross-language speech perception. In: Strange, W. (ed.). *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Timonium: York Press.

[27] Isaacs, T. 2008. Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review* 64, 555-580.

[28] Weber, A., Scharenborg, O. 2012. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 3, 387-401.

[29] Cutler, A. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge: MIT Press.

[30] Chen, H., Li, J. 2017. Reflections on the current situation of phonological assessment: A survey among raters of standardized oral tests in China. *Foreign Languages and Their Teaching 5*, 81-87.

THE PERCEPTION OF JAPANESE LEXICAL CONTRASTS BY L2 LEARNERS: THE ROLE OF THE LEXICON

Marco Fonseca

University of Illinois at Urbana-Champaign
marcofon@illinois.edu

ABSTRACT

This paper examines the perception of three types of Japanese lexical contrasts by L2 learners native speakers of English: words with pitch accent on the penultimate syllable versus words with pitch with on the last syllable, voiceless stops versus voiced stops, and phonemically long versus short vowels. An ABX discrimination task and a lexical assignment task were conducted. A group of 22 L2 learners and a control group of 17 native speakers of Tokyo Japanese were recruited. The results of this study show that although L2 learners are able to discriminate the three types of contrast as well as native speakers, they find it difficult to assign meaning to different contrasts, especially for pitch accent. It is argued that this is evidence that the lexical pitch accent contrast is "weakly" encoded by L2 learners.

Keywords: speech perception, lexical contrast, second language acquisition, pitch accent, Japanese.

1. INTRODUCTION

This paper examines how second language learners native speakers of English (henceforth L2 learners) perceive three different types of Japanese lexical contrasts: voiceless versus voiced stops: *kakkou* "outfit" versus *gakkou* "school"; phonemically long versus short vowels: *biru* "building" versus *biiru* "beer"; and words with lexical pitch accent on their last syllable versus first syllable: *háshi* "chopsticks" versus *hashí* "bridge". The following research questions are proposed: are pitch suprasegmental contrasts more difficult to be perceived than segmental contrasts (RQ1)? Can L2 learners encode (i.e., assign meaning to) these different types of lexical contrast (RQ2)? The motivation for choosing English and Japanese and the importance of these two research questions are as follows. The phonetic implementation of these three type of contrasts in Japanese is different from English: English uses a lax-short vowel distinction [1], is a stressed-timed language [2], and has longer voice onset time than Japanese [3]. Moreover, previous studies have generally focused on a specific type of contrast, such as vowel length [4, 5], voicing of stops [6, 7], or pitch accent [8, 9]. In this study, three different types of contrasts were analyzed. This allows us to determine the relative difficulty in the acquisition of prosodic contrasts (e.g., pitch accents) with respect to segmental contrasts (voicing of stops and vowel length). Furthermore, previous studies have focused on the perception of phonetic differences and on the influence of one's native language phonological categories on the perception of L2 categories [10, 11, 12]. Fewer studies have paid attention to whether L2 learners can encode contrastive L2 categories in their mental lexicon [13, 14, 15]. Therefore, the importance of the research questions is two-fold: it will be tested whether different contrasts are perceived differently by L2 learners, since previous studies tend to focus solely on one type of contrast. It will also be tested the role of lexical encoding (i.e., whether L2 learners are able to assign meaning to different minimal pairs of different contrast types or not).

If the answer to RQ1 is no and even so L2 learners have a hard time encoding different lexical contrasts (RQ2), perceiving an L2 contrast does not imply that the contrast will be learned, as evidenced by Amengual [13] and Ota et. al [14]. This would be evidence that the phonology of an L2 is acquired differently from that of an L1, since native speakers do not seem to find difficulty in encoding different lexical items. To answer these two research questions an ABX discrimination task and a lexical assignment task were conducted. If pitch accent contrasts are more difficult to be acquired than segmental contrasts, participants will perform better with stimuli containing minimal pairs differing in pitch accent than the other contrasts. Furthermore, if L2 learners find it more difficult to encode pitch accent in comparison to stop and vowel contrasts, they will find it more difficult to assign meaning to minimal pairs differing in pitch accent than minimal pairs differing in vowel length and stop voicing.

2. METHODS

Thirty-nine participants were recruited. From these, 22 were advanced L2 learners native speakers of English (7 females and 15 males). Seven were recruited at the University of Illinois and 15 were recruited remotely and performed the experiment online. From these 15, 13 were currently residing in the Tokyo region and 2 had recently returned to the U.S. from a year abroad studying in the Tokyo region. All participants residing in the U.S. had spent at least one year studying in Japan at the university level. The range of all L2 learners' time studying Japanese was 2-18 years (M: 6.42, SD: 4.75). Their age range was 21-42 years old (M: 27.95, SD: 6.7). The age that they started studying Japanese ranged from 13-25 years old (M: 18.77, SD: 3.22). For the participants living in Tokyo, their age of arrival ranged from 20 to 31 years (M: 24, SD: 3.015). Eighteen participants were from the U.S., 3 were from the U.K. and 1 was from Australia. All of them self-rated as being advanced learners[1]. The control group was comprised of 17 native speakers of Tokyo Japanese (5 females and 12 males). Six of them were recruited at a research university in the U.S. and the remaining 12 were recruited remotely and were living in the Tokyo region. All native speakers were university students. Their age range was 21-40 (M: 28.94, SD: 5.56). None of the participants reported having been diagnosed with a hearing impairment.

A total of 60 words, comprised of 10 minimal pairs for each type of contrast (pitch accent, stops, and vowel length) were selected. Pitch accent words minimal pairs were accented in their first syllable. Some words were unaccented rather than second accented. However, because of the initial lowering rule in Tokyo Japanese, and because they were phrase-final, without any following particles, their phonetic realization was the same as second accented words. Eight of the minimal pairs were disyllabic. The remaining two were trisyllabic. Each minimal pair of voiced/voiceless stops had the same accent pattern. For the stop contrasts, only minimal pairs of voiceless/voiced velars /k, g/ were selected. Nine of the minimal pairs had the velar stop in the onset of the first syllable. The remaining one had the velar stop in the onset of the second syllable. Words of a previous study by Hirata [4] were selected for long and short vowel minimal pairs.

The minimal pairs were recorded by two native speakers of Tokyo Japanese who resided in the U.S. at the time of the recording. One participant self-identified as a female and the other one as a male. Native speakers of different genders were recorded to increase the difficulty of the task: in the ABX, A and B were produced by the male participant and X was produced by the female participant. The female participant was in her early thirties at the time of the recording while the male participant was in his late twenties. The native speakers were recorded at a sound-attenuated booth. An AKG C520 head-mounted microphone positioned approximately 2 inches away from the participants' mouth was used. The audio was captured by a Marantz PMD570 solid state recorder with a sampling rate of 48 kHz and a sample size of 16 bits. Participants were asked to produce each word 5 times and the researcher selected the token that sounded most representative of the category.

An ABX discrimination task and a lexical assignment task were conducted. For the participants recruited at the research university, the software PsychoPy2 (version 1.85.4) and Sony Dynamic Stereo Headphones MDR-7508 were used. For the participants recruited online, the online platform Qualtrics was employed. It was asked for the participants to conduct the experiment with ear/headphones. In the ABX, participants heard the first word ("A"), the second word ("B"), and then the third word ("X"). Then the following sentence would appear on the screen: "Is the last word you heard the same as the first or as the second?". If the third word was the same as the first one, participants would press the left arrow key. If the third was the same as the second one, they would press the right arrow key. Before starting the task, a quick training was performed in order to get the participants familiar with the experiment. After the training session, they heard the 30 minimal pairs in 4 word orders (ABA, ABB, BAA, and BAB), for a total of 120 items. There was an interval of 2 seconds between each word. In the lexical assignment task, participants heard the same words that they heard in the ABX task (produced by the male speaker), one at time. Along with the audio, (e.g. */háɕi/*), they were presented two options (e.g. *chopsticks* or *bridge*), and were asked to select the appropriate meaning of the word[2]. The words that they reported not knowing were deleted from the analysis. Each participant took approximately 60 minutes to complete the tasks.

3. RESULTS

A total of 4560 responses were analyzed in the ABX task. From these, 2040 (120 tokens * 17 participants) were obtained from native speakers

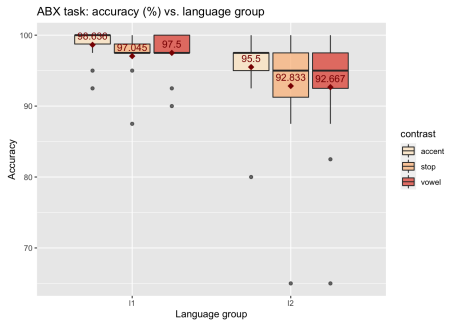


Figure 1: Accuracy by language group and type of contrast for the ABX task. Red numbers in the boxes are the means for each group.

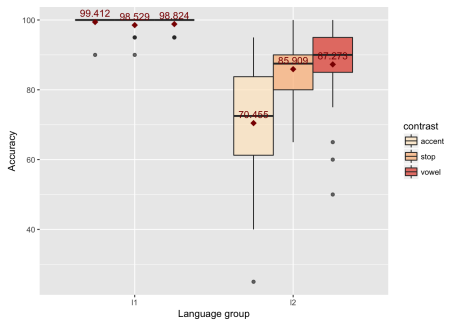


Figure 2: Accuracy by language group and type of contrast for the lexical assignment task.

while 2520 (120 tokens * 21 participants) were obtained from L2 learners. For each type of contrast, 840 (2520 responses /3 types of contrasts) correct responses were possible for the L1 group and 280 (840 responses /3 types of contrasts) correct responses were possible for the L2 group. The results of the ABX task are illustrated in Figure 1.

Both language groups performed at ceiling in the three types of contrast tested. The data were fit into a linear mixed effects logistic regression (function *glmer*, package *lme4* [16]) in R [17]. P-values were obtained with the function *mixed* from the package *afex* [18]. The dependent variable was response (two levels: wrong/right); the predictors were language group (two levels: L1/L2), type of contrast (three levels: accent/stops/ vowels) and their interaction, experimental environment (two levels: in person/remote); random effects were participant and item. All of the mixed effects regressions presented in this paper had their random effects maximized [19]. Considering a statistical significance threshold of $p < 0.05$, no significance was found for any of the predictors [3].

For the lexical assignment task, 2340 responses were obtained (60 items * 39 participants). From these responses, 1020 (60 items * 17 participants) were from native speakers and 1320 were from L2 learners (60 items * 7 participants). For each type of contrast, 340 (1020 responses / 3 types of contrasts) correct responses were possible for the L1 group and 440 (1320 responses / 3 types of contrasts) correct responses were possible for the L2 group. The distribution of correct responses is illustrated in Figure 1. In this task, the L1 group performed at ceiling. For L2 learners, on the other hand, a greater discrepancy was found. The lowest accuracy was observed for accent contrast (70%), followed by stops (86%), and vowels (87%). Similarly to the ABX, the responses were fit into a linear mixed effects logistic regression. The dependent variable was response (two levels: wrong/right); the predictors were language group (two levels: L1/L2), type of contrast (three levels: accent/stops/vowels) and their interaction, experimental environment (two levels: in person/remote); random effects were participant and item. There was an effect of language ($\chi^2(1) = 71.12$, $p < 0.0001$) and its interaction with contrast ($\chi^2(2) = 8.06$, $p = 0.02$). Post-hoc comparisons were conducted using the *emmeans* function from the package *emmeans*. Almost all of the comparisons (p-values ranging from < 0.0001 to 0.0041) were significant. All estimates of all contrasts for the native speakers were higher than the estimates of L2 learners. This indicates that native speakers performed better than L2 learners across the three contrasts tested in this experiment. For L2 learners the comparison between the accent contrast versus the vowel contrast was significant, and the estimate of the former contrast (the baseline) was lower than the latter contrast (β = -1.23, p = 0.004). This indicates that the vowel contrast was easier to be discriminated than pitch accent. The same tendency was observed in the pitch accent versus stop contrast for L2 learners (β = -1.13, p = 0.01). Thus, the statistics indicate that the accent contrast is the most difficult one to be identified in the lexical assignment task for L2 learners.

4. DISCUSSION

The first RQ proposed by this paper is whether suprasegmental contrasts are more difficult to be perceived than segmental contrasts by L2 learners. The results of the ABX task showed that L2 learners are able to perceive the difference between segmental vs. suprasegmental contrasts as well as native speakers. The second RQ is whether L2 learners encode these different types of lexical contrast. The results of the lexical assignment task showed that although L2 learners can accurately assign meaning to minimal pairs differing in stop and vowel contrasts as well as native speakers, they

find it difficult to assign meaning to pitch accent minimal pairs. The results of the present paper have implications for models on the acquisition of L2 phonology. This section discusses some of these models, and whether their predictions can explain the results of the present paper.

Both the Native Language Magnet Model (NLM, [20]) and the Speech Learning Model (SLM, [21, 11]) take into account the similarity of phonetic properties of an L1 sound and an L2 sound. According to NLM, an L1 sound whose phonetic properties are similar to an L2 sound would work as a prototype. This prototype would act as a perceptual magnet that attracts L2 sounds, which would be categorized based on this prototype. As L1 sound categories develop through adulthood, the model proposes that these categories will become more powerful attractors of L2 categories. According to SLM, the sounds that comprise an L1 and L2 phonetic system of a bilingual share a common phonological space. Consequently, these two systems will influence one another. SLM proposes that speech learning abilities remain intact throughout a learner's lifespan. However, age effects arise due to how the L1 and L2 phonetic systems interact. A revised version of SLM (SLM-r, [22]) switches the focus from age effects to the precision of the L1 category: through time an L1 category becomes more robust and therefore it will more likely affect an L2 category. These two models take into account phonetic similarities between two languages, however, encoding of lexical items is outside their scope and therefore they cannot fully explain the results of the present paper.

Another prominent model regarding the perception of nonnative speech categories is the Perpectual Assimilation Model—PAM-L2 [23, 24, 25]. Similarly to NLM and SLM, PAM-L2 argues that not all sound categories of a nonnative language are acquired equally. However, a major difference between PAM-L2 and the other models is that PAM-L2 proposes that listeners are able to detect articulatory gestures. Within articulatory phonology [26], gestures are considered the atomic unit of representation. Thus, while SLM and NLM focus on the phonetic similarities between L1 and L2 sounds, PAM-L2 taps into the phonological level. Although PAM-L2 does not explicitly discuss lexical encoding, positing the importance of the phonological level provides it a level of abstraction.

The Second Language Linguistic Perception Model—L2LP [27, 28] also takes into account the learner's native phonological system. This model posits that L2 learners copy their native

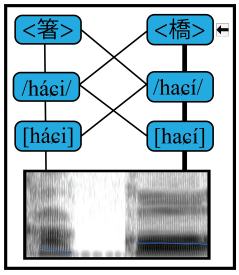


Figure 3: Levels of L2 sound categories acquisition as proposed by the revised L2LP. This diagram illustrates the correct acquisition of 橋 /haɕí/ "bridge", rather than 箸 /háɕi/ "chopsticks".

phonological system and use this copy in order to develop new sound categories. From there, listeners might create a new sound category for the L2 sound if it is considered to be too different from L1 categories. It might also be the case that a category present in the L1 phonological system is reused in order to categorize the L2 sound. A revised version of this model further proposes that learning an L2 sound is divided into different levels. According to the revised L2LP, learning of an L2 phonetic category is meaning-driven and occurs sequentially: acoustic form → [phonetic form] → /phonological form/ → <lexical form>. Such levels are connected and input strengthens these connections, as shown in Figure 3. It is argued that this model can best explain the results of the present paper.

The results of the perception experiment show that as argued by these models, the perception of lexical contrasts varies between native speakers and L2 learners, and that such differences might be because of phonetic differences. Furthermore, meaning also plays a role in the accuracy of this perception, highlighting the importance of the meaning-driven learning of L2 phoneme categories.

5. CONCLUSION

The primary finding of this study is that even if L2 learners have learned to attend to L2 contrasts that have a phonetic implementation different from their L1, they may still fail to fully use that distinction in their mental lexicon. The present study provided evidence that even if L2 learners can hear the difference between a lexical contrast that are not present in their native language, it does not mean that their lexical representation is as robust as the lexical items of their native language. Such differences could be because pitch accent differences are not commonly taught in the classroom, or because Japanese orthography does not indicate it. These may be worthwhile factors to explore in future research.

REFERENCES

[1] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[2] M. E. Beckman and J. Edwards, "Articulatory evidence for differentiating stress categories," *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form*, pp. 7–33, 1994.

[3] T. J. Riney, N. Takagi, K. Ota, and Y. Uchida, "The intermediate degree of VOT in japanese initial voiceless stops," *Journal of Phonetics*, vol. 35, no. 3, pp. 439–443, 2007.

[4] Y. Hirata, "Effects of speaking rate on the vowel length distinction in Japanese," *Journal of Phonetics*, vol. 32, no. 4, pp. 565–589, 2004.

[5] M. Nakagawa and T. Funamura, "Shokyuu nihongo gakushuusha no choutan boin no ninshiki keikou to jizoku jikan [auditory perception of long and short vowels by beginning learners of Japanese in relation to duration time]," *Journal of International Student Center, Hokkaido University*, vol. 4, pp. 18–37, 2000.

[6] N. Yamada, "Nihongo akusento shuutoku no ichi dan kai - gaikokujin gakushuusha no baai [The steps of acquisition of Japanese accent - the case of foreign learners]," *Nihongo Kyouiku*, vol. 83, pp. 108–120, 1994.

[7] M. Fukuoka, "Pekingo bogo washa to shanhaigo bogo washa o taishou to shita nihongo no yuusuei musei haretsuon no oudanteki shuutoku kenkyuu [a longitudinal study in the acquisition of voiced and voiceless stops by native speakers of Beijing Mandarin and Shanghainese]," *Nihongo Kyouku*, no. 87, pp. 40–53, 1995.

[8] T. Ayusawa, "Gaikokujin gakushuusha no nihongo akusento intoneeshon shuutoku (acquisition of Japanese accent and intonation by foreign learners)," *Journal of the Phonetic Society of Japan*, vol. 7, no. 2, pp. 47–58, 2003.

[9] J. J. Atria and V. L. Hazan, "Development of accentual categories in japanese as a second language," *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[10] J. E. Flege, "The production of new and similar phones in a foreign language: Evidence for the effect of equivalence classification," *Journal of phonetics*, vol. 15, no. 1, pp. 47–65, 1987.

[11] ——, "Language contact in bilingualism: Phonetic system interactions," *Laboratory Phonology*, vol. 9, pp. 353–382, 2007.

[12] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*, vol. 1334, 2007.

[13] M. Amengual, "The perception and production of language-specific mid-vowel contrasts: Shifting the focus to the bilingual individual in early language input conditions," *International Journal of Bilingualism*, vol. 20, no. 2, pp. 133–152, 2016.

[14] M. Ota, R. J. Hartsuiker, and S. L. Haywood, "The KEY to the ROCK: Near-homophony in nonnative visual word recognition," *Cognition*, vol. 111, no. 2, pp. 263–269, 2009.

[15] J.-W. Van Leussen and P. Escudero, "Learning to perceive and recognize a second language: The L2LP model revised," *Frontiers in psychology*, vol. 6, p. 1000, 2015.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[17] R. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[18] H. Singmann, B. Bolker, J. Westfall, and F. Aust, *afex: Analysis of Factorial Experiments*, 2016, r package version 0.16-1. [Online]. Available: https://CRAN.R-project.org/package=afex

[19] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[20] P. K. Kuhl and P. Iverson, "Linguistic experience and the perceptual magnet effect," *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, pp. 121–154, 1995.

[21] J. E. Flege, C. Schirru, and I. R. MacKay, "Interaction between the native and second language phonetic subsystems," *Speech Communication*, vol. 40, no. 4, pp. 467–491, 2003.

[22] J. E. Flege and O.-S. Bohn, "The revised speech learning model (SLM-r)," *Second language speech learning: Theoretical and empirical progress*, pp. 3–83, 2021.

[23] C. T. Best, "The emergence of native-language phonological influences in infants: A perceptual assimilation model," *The development of speech perception: The transition from speech sounds to spoken words*, vol. 167, no. 224, pp. 233–277, 1994.

[24] ——, "A direct realist view of cross-language speech perception," *Speech perception and linguistic experience: Issues in cross-language speech research*, pp. 171–206, 1995.

[25] J. Chen, C. T. Best, M. Antoniou, and B. Kasisopa, "Cognitive factors in perception of Thai tones by naïve Mandarin listeners," *Proceedings of the 19th ICPHS, Melbourne*, pp. 1684–1688, 2019.

[26] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[27] P. Escudero, "Linguistic perception of similar L2 sounds," *Phonology in Perception*, vol. 15, pp. 152–190, 2009.

[28] R. Mayr and P. Escudero, "Explaining individual variation in L2 perception: Rounded vowels in English learners of German," *Bilingualism: Language and Cognition*, vol. 13, no. 3, pp. 279–297, 2010.

[1] Logistically, it is very difficult to find L2 learners of Japanese who are native speakers of English, hence the somewhat heterogeneous L2 group. Several statistical analyses considering the possible differences in this group were conducted. No difference was found between such differences. These analyses are available in my GitHub account: https://github.com/marcomzp/ICPhSdata

[2] Participants were asked at the end of the experiment if there were words that they did not know. The words were chosen in consultation with instructors of Japanese who taught Japanese to some of the participants at the time

[3] All the data and analyses are also available in my GitHub account.

FACTORS DETERMINING PERCEPTUAL AND ACOUSTIC SIMILARITY BETWEEN NATIVE AND NON-NATIVE VOWELS

Anna Balas, Magdalena Wrembel, Kamil Kaźmierski, Jarosław Weckwerth

Faculty of English, Adam Mickiewicz University, Poznań

abalas@amu.edu.pl, magdala@amu.edu.pl, kamil.kaźmierski@amu.edu.pl, wjarek@amu.edu.pl

ABSTRACT

This study investigates perceptual assimilation of Norwegian vowels to Polish vowels, based on a task completed by 15 learners of Norwegian. The perceptual similarity is juxtaposed with the acoustic similarity operationalized as the Euclidean distance. The so-called marked lip rounding (as in the case of Norwegian front and central rounded vowels) and vowel length are examined as factors which may potentially influence perceptual assimilation. AIC comparison suggests that a model including Euclidean distances based on F1 and F2 only is better than a model including F3 as well. Neither vowel length nor marked lip rounding turned out to be significant in predicting assimilation count, but the interaction of lip rounding with Euclidean distance proved to be significant, meaning that for vowels with unmarked lip rounding, there is a stronger effect of Euclidean distance on assimilation. The conclusion for non-native vowel perception research is that marked lip rounding could be a factor to analyze when considering assimilation patterns.

Keywords: perceptual and acoustic similarity, non-native vowels, Euclidean distance, Norwegian.

1. INTRODUCTION

So far a number of studies have been conducted to test the perceptual assimilation of non-native vowels (for example, [1]), but relatively few studies have addressed the relationship between vowel perception and acoustic parameters. Strange et al. [2] examined acoustic and perceptual similarity of North German and American English vowels, to find that acoustic similarity did not always predict perceptual similarity, especially for front rounded vowels. Vowel duration differences did not affect perceptual assimilation patterns and spectral differences were more important than duration cues. [3] showed that detailed acoustic comparisons between native (Peruvian Spanish and Australian English) and non-native (Dutch) vowels predicted perception patterns more accurately than overall comparisons of inventory size. Although both [2] and [3] included front rounded vowels and observed peculiar assimilation patterns, they did not specifically investigate the role of lip rounding or F3 in the perceptuo-acoustic relationship and this is the point the present paper aims at addressing.

1.1. Perception of front rounded vowels

In perception studies it has been shown that speakers of languages that do not have front rounded vowels and are not experienced in contrasting them with front unrounded and back rounded vowels, find it challenging to perceive front rounded vowels, which seem to be more marked (i.e. less frequent among world languages, and also more difficult to learn).

[4] showed relatively poor discrimination by inexperienced English listeners of the Norwegian high front unrounded versus outrounded /i–y/ in /bV/ syllables. Further studies [5] tested the perception of Norwegian vowels by English, French and Danish listeners and concluded that the results were largely in line with the phonologically contrastive and non-contrastive phonetic-articulatory properties of the listeners' L1s.

Most of previous research on this topic concentrated on perceptual differentiation of French or German front rounded vowels by English listeners. [6] pointed to the challenges that American English learners of French and American English speakers who had not learned French, had in comparison to native speakers of French when discriminating between French front rounded vowels and when identifying mid front rounded vowels.

In [7] Canadian English listeners had more difficulty discriminating German lax vowels /ʏ/ and /u/ than the tense contrast /y:/ and /u:/. A vowel identification and rating task using English keywords clearly showed that the four German vowels were distinguished from native English categories, but all the vowels were assimilated to English /u:/ and /ʊ/, pointing to the lip rounding being more important in assimilation decisions than the front position of the tongue.

1.2. Norwegian and Polish vowel systems

The two vocalic systems selected for scrutiny in the present study differ considerably. Polish has a simple six-vowel inventory: /i, ɨ, ɛ, a, ɔ, u/ with no

distinctions in rounding, tenseness or duration (though some of the vowels have nasalized variants). Norwegian, on the other hand, has a rich a complex vocalic system with duration and rounding contrasts: /iː, i, yː, y, eː, e, øː, ø, ɑː, ɑ, oː, o, uː, u, ʉː, ʉ/, not to mention excessive dialectal variability.

2. THE PRESENT STUDY: RESEARCH QUESTIONS AND PREDICTIONS

L1 Polish listeners learning Norwegian, a language with rounding contrasts, but where rounding is not as strong as for example in German or French, seem to offer a good testing ground for investigating the relationship between perceptual and acoustic similarity and the role of lip rounding and F3. In the present experiment, perceptual assimilation of 16 Norwegian vowels to Polish vowel categories by Polish learners of Norwegian was investigated. Examining perceptual assimilation and juxtaposing it with the Euclidean distances between Norwegian vowel stimuli and reference values for Polish vowels (cf. [8]) permitted four research questions to be asked and predictions to be made.

RQ1: Is the Euclidean distance measure based on F1, F2 and F3 better suited to predict assimilation counts than the Euclidean distance based on F1 and F2 only?

As F3 is related to lip rounding, and lip rounding plays an important role in many Norwegian vowels, including the front high and central rounded vowels, we expect the Euclidean distance measure including F3 to be able to predict assimilation counts more precisely than the Euclidean distance measure including F1 and F2 only.

RQ2: Does the smaller Euclidean distance between a Norwegian vowel stimulus and a Polish vowel category enhance the likelihood of assimilating a given Norwegian vowel to a given Polish vowel?

We hypothesize that acoustic similarity operationalized as the Euclidean distance will correspond to perceptual similarity operationalized as assimilation count.

RQ3: Is the impact of Euclidean distances stronger for vowels without marked lip rounding vs. vowels with marked lip rounding, i.e. /y(ː), ø(ː), ʉ(ː)/?

In the case of front and central rounded vowels, listeners with an L1 that has only less marked combinations of rounding in the case of high and mid back vowels, probably experience tension between choosing the assimilation target on the basis of the standard F1 and F2, or tongue height and advancement or F3 more related to lip rounding, when these features are incongruent in comparison to the standard unmarked configurations familiar from the L1. Therefore, in the case of Norwegian vowels with more marked lip rounding we expect the assimilation results to be influenced by the conflict between the position of the tongue typical for front vowels and lip rounding typical for back vowels in the L1, and less related to the Euclidean distance.

RQ4: Is the impact of Euclidean distances different for short and long vowels?

We hypothesize that vowel length may modulate the perceptuo-acoustic relationship, yet its relative influence remains to be established, as so the results of previous studies are inconclusive (cf. [2] where American English listeners relied on spectral cues when duration and spectral similarity were in conflict, and [9] where desensitization hypothesis emphasizes the role of duration in non-native vowel perception, but does not make claims regarding assimilation).

3. METHODS

3.1. Participants

15 participants, aged 21, 9 females, 6 males, who were 3rd year students in a Norwegian modern language BA programme, participated in the study. They filled in the Leap-Q Language Experience and Proficiency Questionnaire (Marian et al. 2007). Their language inventory included L1 Polish, L2 English and L3 Norwegian. Both L2 and L3 were acquired in a classroom setting. The exposure to Norwegian during the three years prior to the experiment was approximately 17 hours a week of speaking and listening and the subjects spent similar amount of time reading and writing in Norwegian. Their proficiency level in Norwegian was B2.

3.2. Stimuli

The stimuli were recorded by a Norwegian female native speaker and included 16 Norwegian vowels /iː, i, yː, y, eː, e, øː, ø, ɑː, ɑ, oː, o, uː, u, ʉː, ʉ/. They were embedded in nonce words in a /dVd/ framework, e.g. *dåd*, *dedd*, *did*.

3.3. Procedure

Due to the COVID-19 pandemic, the experiment was conducted online in June 2021. In addition to the perception experiment reported on here, the session included a battery of tasks in speech production and syntax. The part devoted to speech perception included assimilation and goodness of fit rating tasks presented to the participants in an online session using PsychoPy [10]. Prior to the audio stimulus, the participants were presented with a fixation point, at the onset of the audio stimulus (one of the 16 Norwegian vowels embedded in a /dVd/ nonce word), the participants saw a screen with a question in

Polish: *Which Polish vowel is the vowel you have just heard most similar to?* The participants were also presented with orthographic labels representing the six Polish vowels, as Polish vowel orthography is transparent, matched with numeric values to press on the keyboard (1, 2, 3 and 5, 6, 7 as these were deemed to be the most convenient to be used with both hands – the participants were instructed to keep the fingers on the keyboard). Once they made their choice, the participants saw another screen asking them to rate how well the vowel they heard fits the category they chose on a 7-point scale.

4. RESULTS, ANALYSIS AND DISCUSSION

4.1 Perceptual assimilation results

Table 1 presents the results of the behavioral tests.

Norwegian stimulus	Polish vowel targets					
	<i> /i/	<y> /ɨ/	<e> /e/	<a> /a/	<o> /ɔ/	<u> /u/
/i:/ TID /	95.6 (5.6)		2.2 (3)			
/i/ FIN	35.6 (4.3)	57.8 (5.1)	6.7 (4.3)			
/y:/ LYS	84.1 (4.6)	11.4 (3.6)			2.3 (1)	2.3 (3)
/y/ SYND	8.9 (3.3)	80 (4.7)			2.2 (1)	8.9 (4.5)
/e:/ STED	2.2 (4)		95.6 (5.1)	2.2 (7)		
/e/ BEST	4.4 (4)		93.3 (5.8)	2.2 (5)		
/ø:/ LØP		20 (4)	24.4 (2.9)		46.7 (3.7)	6.7 (4.7)
/ø/ SØNN		27.3 (4.4)	20.5 (3.2)		40.9 (4.2)	9.9 (4)
/ɑ:/ DAG				100 (5)		
/ɑ/ TAKK				100 (5.1)		
/o:/ RÅD	2.2 (7)				97.8 (5.1)	
/o/ NOK					100 (5.5)	
/u:/ BOK					22.2 (4.7)	77.8 (4.5)
/u/ ROM	2.2 (5)				64.4 (4.6)	33.3 (5.3)
/ʉ:/ GUD		16.7 (4)				83.3 (3.9)
/ʉ/ SLUTT		15.9 (4.3)	2.3 (2)		6.8 (4.3)	75 (4.2)

Table 1: Confusion matrix with mean per cent categorization and goodness rating (in parentheses) of Norwegian vowel stimuli, presented here as phonetic symbols and keywords, in terms of Polish vowel categories, presented here as graphemes and phonetic symbols. The goodness of fit ratings are based on a scale from 1 *unlike* to 7 *identical*.

4.2 Analysis and discussion

Statistical analysis was run in R [11]. For significance testing, initially Poisson regression models with assimilation count (the number of times a given Norwegian vowel was categorized as a given Polish vowel) as the response variable were used, as they are restricted to non-negative values, and assimilation counts cannot be negative. Further, Poisson regression captures the association on a logarithmic scale, which seems appropriate for the data. Since the Poisson models suffered from overdispersion, negative binomial regression models were fitted using the 'glm.nb' function from the *MASS* R package [12].

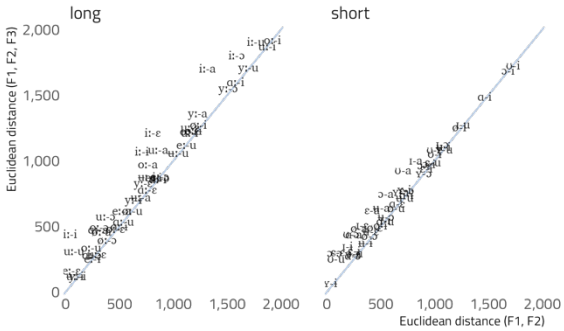


Figure 1: Distances between Norwegian and Polish vowels in terms of Euclidean distances based on F1, F2 and F3, and F1 and F2 only measures. Left vowel in each pair is Norwegian.

We fitted a model to examine assimilation count as a function of Euclidean distance, vowel length, unmarked vs. marked status with respect to lip rounding, and the interaction between the Euclidean distance and marked lip rounding. In order to answer the first research question, whether the Euclidean distance measure based on F1, F2 and F3 is better suited to account for assimilation counts than the Euclidean distance measure based on F1 and F2 only, the analysis was performed for the Euclidean distance defined in terms of F1 and F2 only and F1, F2, and F3. We then compared the models based on the Akaike Information Criterion. AIC comparison led to somewhat unexpected results, in the light of the fact that F3 was expected to be a proxy for lip rounding and enhance the model. Comparing the F1, F2, and F3 model (ac ~ euc_f1_f2_f3 * length) to an F1, F2 model (ac ~ euc_f1_f2 * length), we found that the F1 and F2 only model is lower by three AIC units (440 vs. 443, 7df) so it is a better model. It needs to be

noted that both Euclidean distance measures (F1, F2 and F3 and F1 and F2) are highly correlated (r=0.98, p=0.001). For some vowel pairs (those above the diagonal in Fig. 1) there is a greater F1, F2 and F3 difference than F1 and F2 difference. For many vowel pairs, on the diagonal in Fig. 1, there is no difference.

Term	Estimate	p-value
(Intercept)	2.778	<0.05***
euc_f1_f2	-0.001	1
lengthshort	0.139	0.758
non_marked_rounded	0.571	0.452
euc_f1_f2:lengthshort	-0.001	0.553
euc_f1_f2: marked_rounding	-0.002	0.029*

Table 2: Coefficient table of the negative binomial model of the assimilation count. *p*-values of fixed effects calculated with likelihood-ratio tests.

Let us now inspect the negative binomial model of assimilation count in Table 2. Answering the second research question, the negative estimate for euc_f1_f2 indicates that the higher the Euclidean distance between a given pair of vowels, the lower the number of times the Norwegian vowel is predicted to be assimilated to the Polish vowel, but the effect is not statistically significant (p=1).

To answer the third research question and check whether the effect of the Euclidean distance is stronger in the case of vowels with unmarked lip rounding vs. with marked lip rounding, we included marked lip rouding as a predictor in the model. It did not turn out to be significant on its own, but its interaction with Euclidean distance proved to be (b = -0.002, p = 0.029), meaning that for vowels with unmarked lip rounding, there is a stronger effect of Euclidean distance on assimilation count than for vowels with marked lip rounding. Looking at the empirical data in Fig. 2, we can see that usually a given Norwegian vowels is assimilated to the closest Polish category as determined by the Euclidean distance. On the plots in Fig. 2 such cases exhibit a standard decreasing pattern (as for vowels /iː, yː, e/ and /uː/). We can, however, also observe a few examples with rather low assimilation counts in the beginning, a rise afterwards, followed by a drop, as in the case of /ʉ(ː), ø(ː), oː/ and /u/.

As to the fourth research question, there are negative, but statistically insignificant estimates for short vowels (p=0.758) and the interaction between Euclidean distance and vowel length (p=0.553), which means that vowel length does not play a role in perceptual assimilation. This finding is in line with [2], but the duration variable was considered worth checking, because L1 Polish in the present study does not use any duration cues, in contrast to English in [2].

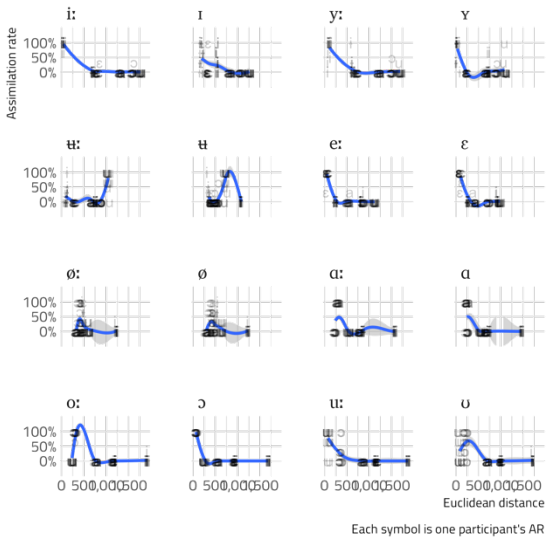


Figure 2: Plots of assimilation rates as a function of the Euclidean distance for each Norwegian vowel stimulus.

5. CONCLUSION

In this experiment, involving Polish advanced learners of Norwegian, a model with Euclidean distances based on the first two formants appeared to be better than a model with the three formants. The association between Euclidean distances and assimilation counts was not statistically significant, but this finding will need to be further tested among larger pools of listeners and also the level of proficiency will need to be looked at, as both [2, 3] referred to naïve listeners, whereas advanced learners may base their assimilation on other features in addition to formant similarity. There were no differences between assimilation predictions for long or short vowels. Finally, Euclidean distance was found to be related to vowels without marked lip rounding more directly than to front and central rounded vowels, absent from Polish, and more marked/less frequent among world languages. In conclusion, this study contributed to a better understanding of the role of Euclidean distance, marked lip rounding and vowel length in determining the perceptuo-acoustic similarity between foreign and native vowels, and motivates more research to elucidate this intricate phenomenon further.

6. ACKNOWLEDGEMENT

This research is supported by a project grant of the Polish National Science Centre (NCN) OPUS-19-HS (UMO-2020/37/B/HS2/00617) CLIMAD “Cross-linguistic influence in multilingualism across domains: Phonology and syntax.”

7. REFERENCES

[1] Tyler, M. D., Best, C.T., Faber, A., Levitt, A.G. 2014. Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica* 71, 4-21. DOI: 10.1159/000356237

[2] Strange, W., Bohn, O-S., Trent, S.A., Nishi, K. 2004. Acoustic and perceptual similarity of North German and American English vowels. *J. Acoust. Soc. Am*.115, 1791-1807. DOI: 10.1121/1.1687832

[3] Alispahic, S., Mulak, K.E., Escudero, P. 2017. Acoustic properties predict perception of unfamiliar Dutch vowels by adult Australian English and Peruvian Spanish listeners. *Frontiers in Psychology* 8, 52. DOI: 10.3389/fpsyg.2017.00052

[4] Best, C.T., Faber, A., Levitt, A. 1996. Assimilation of non-native vowel contrasts to the American English vowel system. *J. Acoust. Soc. Am*. 99, 2602.

[5] Best, C.T., Halle, P., Bohn, O-S., Faber, A. 2003. Cross-language perpcetion of non-native vowels: Phonological and phonetic effects of listeners' native languages. In: Solé, M.J, Recasens, D., Romero, J.(eds.), *15th International Congress of Phonetic Sciences, Barcelona, Spain, August 3-9, 2003*. ICPhS Archive, 2889-2892.

[6] Gottfried, T.L. 1984. Effects of consonant context on the perception of French vowels. *Journal of Phonetics* 12, 91-114.

[7] Polka, L. 1995. Linguistic influences in adult perception of non-native vowel contrasts. *J. Acoust. Soc. Am*. 97, 1286-1296.

[8] Weckwerth, J., Balas, A. 2020. Selected aspects of Polish vowel formants. In: Wrembel, M, Kiełkiewicz-Janowiak, A., Gąsiorowski, P. (eds), *Approaches to the study of sound structure and and speech: Interdisciplinary work in honour of Katarzyna Dziubalska-Kołaczyk.* Routledge: 338-348.

[9] Bohn, O-S. 1995. Cross-language speech perception in adults: First language transfer doesn't tell it all. In: Strange, W. (ed), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research.* York Press, 279-300.

[10] Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, Sogo, H., Kastman, E., Lindeløv, J.K. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51, 195-203.

[11] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[12] Venables, W.N., Ripley, B.D. 2002. *Modern Applied Statistics with S.* Springer.

PHONETIC REALIZATION OF NARROW FOCUS BY CHINESE EFL LEARNERS

Xiaoyu Zhou, Binbin Sun, and Hui Feng

School of Foreign Languages, Tianjin University
zxy8899000@126.com, sunbinbin_tju@163.com, fenghui@tju.edu.cn

ABSTRACT

Focus bears the function of information transmission and is regarded as a notion closely related to the discourse function of prosody. This study takes 20 Chinese EFL learners and 8 native English speakers as the subjects to investigate whether the difference between Mandarin and English focus realization patterns affect the narrow focus realization of Chinese EFL learners. Findings are as follows. (1) Learners of lower English proficiency exhibit similar phonetic realization to Mandarin Chinese, with a post-focus intensity decrease in both initial and medial focus sentences and a significant post-focus duration reduction than English speakers. (2) Learners show new features not found in English narrow focus phonetic realization, that is, the loss of post-focus pitch range compression (PFC) in lower English proficiency learners, and the significantly higher pitch rising of final-focus words than the other focus positions in higher English proficiency learners.

Keywords: Focus, Phonetic realization, Chinese EFL learners, L2 Intonation

1. INTRODUCTION

As a notion closely related to the discourse function of prosody [1], the phonetic realization of focus embodies the combination of semantics and prosody.

Previous studies have observed the several cues related to focus realization within the phonetic level and discovered the existence of a close relationship between on-focus constituents and higher F0, longer duration and stronger intensity [2-3]. Based on detailed F0 analyses, Xu and Xu [4] revealed that the narrow focus is realized by expanding the pitch range of the on-focus stressed syllables and suppressing the pitch range of post-focus syllables, namely "PFC" phenomenon. The existence of PFC can be found in many languages and dialects, for instance, English [5], Korean [6], Beijing dialect [7], Uyghur [8], and Lhasa Tibetan [9].

Though there have been numerous constructive explorations on the way native speakers realize the focus, studies concerning language learners are fewer. The differences in focus realization between learners and native speakers are mainly reflected in the different roles of acoustic cues, the less obvious changing range, and the loss of PFC (eg., [10-12]). Besides, L2 focus realization have been discussed from the perspective of L1 influence (eg., [13-15]).

Based on the research background, the present study attempts to detect the phonetic realization of narrow focus by Chinese EFL learners, and explore cross-language distinction on the narrow focus realization.

2. RESEARCH DESIGN

2.1. Research questions

In an effort to detect the phonetic realization of narrow focus by Chinese EFL learners, this study aims to anwser the following questions:

1) Does the difference between Mandarin and English focus realization patterns affect the narrow focus realization of Chinese EFL learners?

2) Do different proficiency levels affect the phonetic realization strategies of narrow focus by Chinese EFL learners?

2.2. Stimuli

A total of 28 subjects participated in the experiment, including 20 Chinese and 8 Native American English speakers. The 20 Chinese learners are divided into two groups with equal numbers of higher and lower proficiency levels based on the CEFR grade derived from Dialang vocabulary test results [16].

To ensure that speakers naturally produce target declarative sentences with narrow focus, this experiment adopts the combination of guiding questions and corresponding target declarative sentences as the main form of pronunciation stimuli.

The English stimuli design refers to [4], which contains short declarative sentences with keywords based on sonorant onsets and limited coda consonants (reproduced in Table 1 in the Appendix). The pairs of English sentences are divided into four groups according to various focus conditions.

The Chinese recording materials are selected from the stimuli of [17]. Each Chinese sentence contains three keywords with the same number of Chinese characters and consistent internal tones (reproduced in Table 2 in the Appendix). They are produced under

the four focus conditions described above. Thus, a total of 1980 tokens are taken into consideration in the following experiment process, with 28 subjects * 3 keywords * 21 sentences for English stimuli and 8 subjects * 3 keywords * 9 sentences for Chinese stimuli.

2.3. Data collection and processing

During recording process, the recording materials were presented to the subjects in turn through the software Speech Recorder [18] and the subjects were required to operate the recording independently. After recording, all recordings were manually marked under ProsodyPro script files [19]. On this basis, ProsodyPro automatically modified pitch contour and extracts F0, duration and intensity.

Due to the variety of pronunciation styles and physical properties of speech signals, normalization is needed to eliminate interpersonal random differences and extract constant parameters to obtain information of linguistic significance [19-20]. Therefore, Formula (1) is used for F0 values to convert Hz to semitones (st) with 1Hz as reference value.

(1) $F_{st} = 12 \times \log_2(F0)$.

As for the parameters of duration and intensity, in order to explore the impact of discrete variables on duration and intensity and eliminate individual differences, the analysis of duration and intensity in this study mainly focuses on comparison rather than conducting quantitative analysis of absolute values. Hence in terms of the statistical analysis, this paper adopts the processing method of time length and intensity data normalization by Zhao et al. [21], and takes the duration ratio of syllables and relative intensity ratio as the statistical analysis parameter. The formulas are presented below:

(2) $D_r = D/\overline{D_i}$
(3) $INT_r = INT/\overline{INT_i}$

where D_r and INT_r refer to the duration ratio and the relative intensity ratio of the stressed syllable within a keyword respectively, and they are calculated based on the ratio between *D/INT* (the duration/intensity of the syllable) and $\overline{D_i}$ /INT_i (the average syllable duration/intensity of the sentence in which the syllable is located).

3. RESULTS AND DISCUSSION

3.1. Comparison of focus realization between English and Chinese

3.1.1. Pitch analysis

As is illustrated in Fig. 1, American English speakers (AE) show in-focus MaxF0 values increase (t=2.058, p=0.049), and in-focus pitch range expansion (t=3.298, p=0.013) when realizing narrow focus. Moreover, there are post-focus pitch drop and pitch range compression for initial (t=-4.977, p=0.002; t=-3.237, p=0.013) and medial focus sentences (t=-2.325, p=0.053; t=-2.74, p=0.029). Pre-focus pitch range compression also exists, but within medial focus sentences only (t=-3.723, p=0.008). Besides, one-way ANOVAs with focus location as influencing factors confirms that in-focus MaxF0 rising values are significantly larger than that of post-focus and pre-focus words (F(2,23)=5.352, p=0.013).

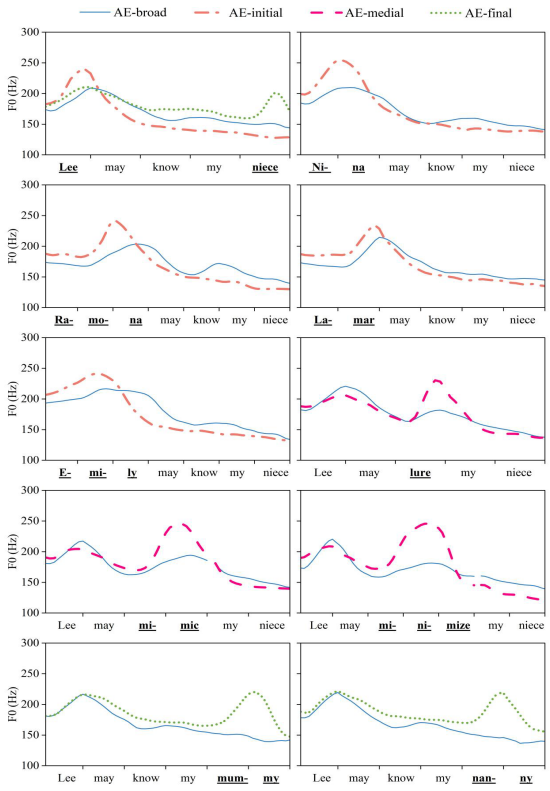


Figure 1: Mean F0 contours of native American English speakers.

Similar to AE, as is reflected in Fig. 2, in-focus pitch rise (by 1.37 semitones, t=2.158, p=0.049) and pitch range expansion (by 0.77 semitones, t=2.607, p=0.035) are both confirmed among Mandarin Chinese speakers (MC). Besides, the post-focus pitch drop (t=5.281, p=0.001; t=-3.278, p=0.014) and pitch range compression (t=-6.436, p=0.000; t=-2.909,

p=0.023) are inspected significantly from data analysis. However, paired-samples t-tests' results suggest that the pitch changing values of MC are larger in post-focus words than in-focus (by 1.95 semitones, p=0.000) and pre-focus words (by 2.78 semitones, p=0.000), which is distinct from AE.

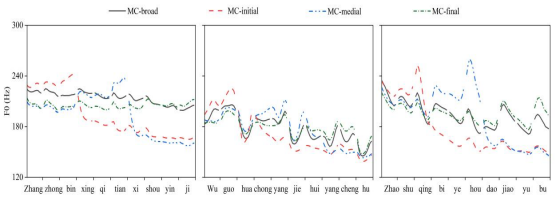


Figure 2: Mean F0 contours of Mandarin Chinese speakers.

3.1.2. Duration analysis

As is shown in Fig. 3, the duration extension was exhibited by both native AE and MC under all focus conditions. Besides, as is proved by the data analysis, the in-focus rise of MC is significantly greater than that of AE (t=-2.769, p=0.015), and post-focus decline of AE is lower than MC (t=4.207, p=0.001). It is also observed that pre-focus declining amplitude in AE is larger than MC (t=-2.393, p=0.031).

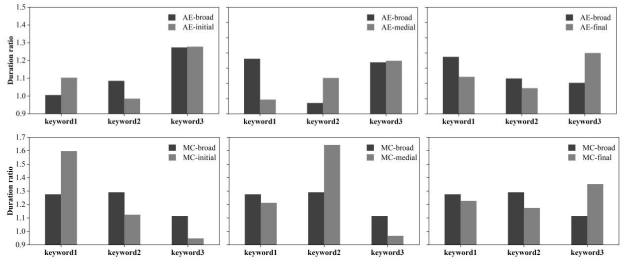


Figure 3: Mean duration ratio of AE and MC.

3.1.3. Intensity analysis

As illustrated in Fig. 4, intensity ratio increase is found in native AE speakers. Moreover, in-focus intensity increase is also found in MC. The pre-focus intensity ratio changes in AE are also significantly lower than that of the MC (t=-4.041, p=0.001).

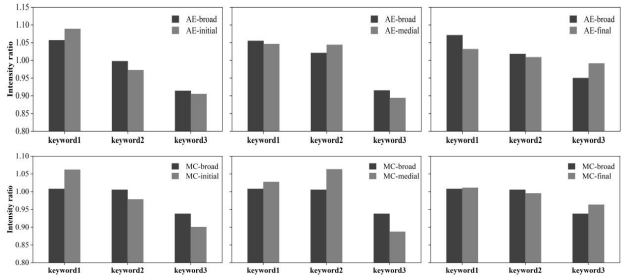


Figure 4: Mean intensity ratio of AE and MC.

3.2. Analysis of focus realization in Chinese EFL learners

3.2.1. Pitch analysis

As is shown in Fig. 5, similar to AE, in-focus MaxF0 rise, pitch range expansion and post-focus pitch drop were visually observed and statistically confirmed to be significant in both levels of Chinese learners.

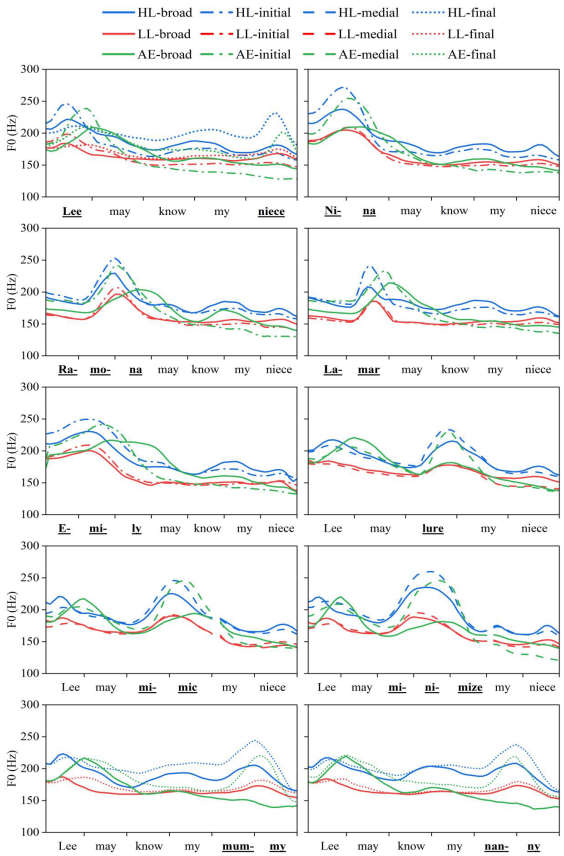


Figure 5: Time-normalized F0 contours (Hz) by HL, LL and AE group.

However, the magnitude of pitch rise does not appear to be consistent across different focus positions when higher proficiency learners (HL) use pitch as the clue to realize narrow focus. As is observed in Fig. 5, the in-focus rising range under final focus condition shown by the distance between dotted blue lines and solid blue lines appeare to be greater than that under the other two focus conditions. To statistically verify the influence of focus positions on in-focus pitch rising values, one-way ANOVAs were conducted and results showed that the main effect of focus positions is significant on the in-focus increase magnitude of MaxF0 values (F(2,29)=5.157, p=0.013), manifesting that in-focus rising values in final focus sentences was significantly higher than that in initial focus by 2.05 semitones (p=0.007) and

than in medial focus sentences by 1.86 semitones (p=0.014).

As for F0 contours of lower proficiency learners (LL) (red lines in Fig. 5), the post-focus pitch range compression was not found either in initial or in final focus sentences. Paired-sampled t-tests confirmed the absence of post-focus pitch range compression in initial focus (t=-1.782, p=0.108) and medial focus sentences (t=0.493, p=0.634). Hence it is speculated that lower proficiency learners lost PFC in the process of learning English as a foreign language, which proved that PFC was not easy to be transferred from one language to another [20]. Although pitch range was found to have insignificant changes in post-focus words produced by LL, it exhibited significant difference in pre-focus words.

3.2.2. Duration analysis

As is illustrated in Fig. 6, and is proved by repeated ANOVAs and post-hoc independent t-tests, the in-focus duration change ratio of AE was significantly higher than that of HL and LL (a near significant difference between HL and AE (t=-2.001, p=0.063) and between LL and AE (t=-1.192, p=0.064)). For post-focus duration changing, significant difference was found between LL and AE (t=-2.490, p=0.024).

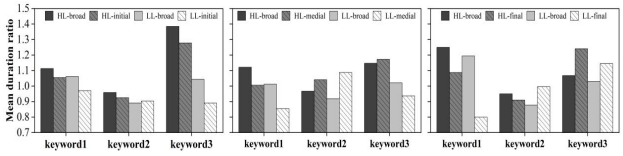


Figure 6: Mean duration ratio of higher proficiency (HL) and lower proficiency learners (LL) .

3.2.3. Intensity analysis

As is illustrated in Fig. 7, and is verified by a set of repeated measures ANOVA, both HL and LL showed in-focus intensity increase, but the in-focus intensity ratio changes (t=-2.709, p=0.015 for HL; t=-2.638, p=0.018 for LL) were significantly lower than AE.

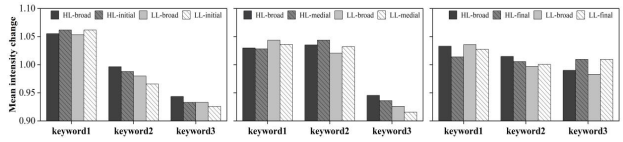


Figure 7: Mean intensity ratio of higher proficiency and lower proficiency learners.

Besides, HL showed significant post-focus intensity decrease in initial focus sentences (t=-2.858, p=0.019) and pre-focus intensity decrease in finial focus sentences (t=-3.020.p=0.014). LL showed significant post-focus intensity decrease in both initial (t=-4.001, p=0.003) and medial (t=-2.421, p=0.039) focus sentences. However, the pre-focus intensity change was found not significant both in medial (t=-1.658, p=0.132) and final focus conditions (t=-0.640, p=0.538) among LL.

4. CONCLUSION

This study aims to explore the the narrow focus realization of Chinese EFL learners under the influence of the differences between Mandarin and English focus realization patterns and English proficiency.

In terms of the differences between English and Mandarin focus realization, it is found that English mainly uses in-focus cues rather than those before and after it, and have a larger pitch rise, smaller duration extension for in-focus words, and less pitch drop, more duration reduction, smaller intensity declines for post-focus words than Mandarin Chinese.

As to focus realization by Chinese EFL learners, it is revealed that higher proficiency learners tend to use the consistent parameter adjusting methods as native American English speakers, while lower proficiency learners tend to use the categorical strategies and gradient features similar to Mandarin Chinese.

It is also found that both speakers of higher and lower proficiency show new features different from their L1 and L2 focus phonetic realization. Learners of lower proficiency lose PFC, the important categorical strategy in Mandarin Chinese, when learning English as a foreign language, which proved that PFC is easy to lose but difficult to learn [12]. Learners of higher proficiency show significantly higher in-focus pitch rising values in final focus sentences than initial and medial focus sentences. In terms of the in-focus duration extension, learners' value is significantly less than American English speakers, while the duration extension of American English speakers is less than Mandarin Chinese speakers, which indicates the overuse of L2 strategies.

Future works are needed to prove the effect of perception performance on the phonetic realization of narrow focus and to conduct long-term research to further confirm the positive development as their English proficiency increased.

5. ACKNOWLEDGEMENTS

This research is partly supported by National Major Social Sciences Foundation of China under Grant 15ZDB103, the Innovation Program of Chinese Academy of Social Sciences, and the Social Science Foundation of Tianjin, China (No. TJWW19-009), and the National Key R&D Program of China under Grant 2018YFB1305200.

6. REFERENCES

[1] Chun, D. M. 2002. *Discourse intonation in L2: From theory and research to practice,* Vol. 1. John Benjamins Publishing.

[2] Cooper, W. E., Eady, S. J., Mueller, P. R. 1985. Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77* 6, 2142-2156.

[3] Wells, W. H. 1986. An experimental approach to the interpretation of focus in spoken English. In *Intonation in discourse,* 53-76. Routledge.

[4] Xu, Y., Xu, C. X. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33* 2, 159-197.

[5] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech communication*, *46* 3-4, 220-251.

[6] Lee, Y. C., Xu, Y. 2010. Phonetic realization of contrastive focus in Korean. In *Speech Prosody 2010-Fifth International Conference*.

[7] Liu, F. 2009. *Intonation systems of Mandarin and English: A functional approach*. The University of Chicago.

[8] Çetinkaya, E. 2022. The linguistic realization of focus in Uyghur: can the two focusing strategies be used interchangeably? *Poznan Studies in Contemporary Linguistics*.

[9] Zhang, X. X., Wang, B. 2018. The prosodic coding of focus and question in Tibetan Lhasa dialect. *Journal of Tsinghua University (Science and Technology)*, *58* 4, 368-373.

[10] Ueyama, M., Jun, S. A. 1996. Focus realization of Japanese English and Korean English intonation. *UCLA Working Papers in Phonetics*, 110-125.

[11] Yeou, M. 2004. The realization of accentual focus in Moroccan learners of English. In *Speech Prosody 2004, International Conference*.

[12] Chen, Y., Xu, Y., Guion-Anderson, S. 2015. Prosodic realization of focus in bilingual production of Southern Min and Mandarin. *Phonetica*, *71* 4, 249-270.

[13] Nava, E., Zubizarreta, M. L. 2010. Deconstructing the nuclear stress algorithm: Evidence from second language speech. *The sound patterns of syntax*, 291-316.

[14] Zubizarreta, M. L., Nava, E. 2011. Encoding discourse-based meaning: Prosody vs. syntax. Implications for second language acquisition. *Lingua*, *121* 4, 652-669.

[15] Li, A., Wan, X., Zhao, C., Zhu, L. 2020. Phonological and phonetic realization of narrow focus in declarative sentences by Jinan EFL learners. *In proceedings of the 10th International Conference on Speech Prosody*, 876-880.

[16] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

[17] Chen, Y., Shi, F. 2011. Pitch performance of intonation of focus sentences in Putonghua. *Nankai Linguistics*, 1, 12.

[18] Draxler, C., Jänsch, K. 2020. *SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software* (Version 6.2.6). [Computer software].

[19] Xu, Y. 2013. ProsodyPro—A tool for large-scale systematic prosody analysis. *Laboratoire Parole et Langage*, France.

[20] Zhu, X. 2004. Fundamental frequency normalization - how to deal with random differences in pitch?. *language Science*, *3* 2, 3-19.

[21] Zhao, J., Yang, X., Yang Y. 2012. The correspondence between focus and stress in Chinese -- a preliminary study based on Corpus. *Studies in Language and Linguistics*, 32 4, 5.

APPENDIX

Focus	Type	Sentences
Broad Focus	Question	What did you say?
	Answer	*See listed sentences below.*
Initial Focus	Question	Who may know your niece?
	Answer	***Lee / Nina / Lamar / Emily / Ramona*** may *know* my *niece*.
Medial Focus	Question	What may Lee do to your niece?
	Answer	*Lee* may ***lure / mimic / minimize*** my *niece*.
Final focus	Question	Who may Lee know?
	Answer	*Lee* may *know* my ***niece / nanny / mummy***.

Table 1: Prompt questions and answers for eliciting focus in the English stimuli (reproduced from [4])

Focus	Tones	*Sentences*
Broad Focus	high level	*Zhangzhongbin Xingqitian Xiu Shouyinji. (Zhangzhongbin fixes the radio on Sunday.)*
	high rasing	*Wuguohua Chongyangjie Hui Yangchenghu. (Wuguohua returns to Yangcheng Lake on the Double Ninth Festival.)*
	high falling	*Zhaoshuqing Biyehou Dao Jiaoyubu. (Zhaoshuqing went to the Ministry of Education after graduation.)*
Initial Focus	high level	***Zhangzhongbin*** *Xingqitian* Xiu *Shouyinji.*
	high rasing	***Wuguohua*** *Chongyangjie* Hui *Yangchenghu.*
	high falling	***Zhaoshuqing*** *Biyehou* Dao *Jiaoyubu.*
Medial Focus	high level	*Zhangzhongbin* ***Xingqitian*** Xiu *Shouyinji.*
	high rasing	*Wuguohua* ***Chongyangjie*** Hui *Yangchenghu.*
	high falling	*Zhaoshuqing* ***Biyehou*** Dao *Jiaoyubu.*
Final focus	high level	*Zhangzhongbin Xingqitian* Xiu ***Shouyinji***.
	high rasing	*Wuguohua Chongyangjie* Hui ***Yangchenghu***.
	high falling	*Zhaoshuqing Biyehou* Dao ***Jiaoyubu***.

Table 2: Target declarative sentences for eliciting focus in Chinese stimuli (reproduced from [17])

Note. The words in italics are keywords, and the non-italic words are non-keywords in this experiment. The in-focus words are located in bold.

TONGUE CONTOURS FOR THE JAPANESE MORAIC NASAL BY SPEAKERS OF STANDARD CHINESE

Ai Mizoguchi[1, 2], Maho Morimoto[3, 4], Li Weiyu[3], Takayuki Arai[3]

[1]Maebashi Institute of Technology, [2]NINJAL, [3]Sophia University, [4]JSPS
aimizoguchi@maebashi-it.ac.jp, maho.morimoto.jp@gmail.com, w-li-5x4@eagle.sophia.ac.jp, arai@sophia.ac.jp

ABSTRACT

The Japanese moraic nasal /N/ has been transcribed as a uvular nasal [ɴ] in the utterance final position, although variability in the place of articulation by speaker has been reported. For L2 learners whose L1 is English, /N/ is often produced as [n], avoiding an unfamiliar place of articulation. The current study investigated the tongue contours for /N/ by native speakers of Standard Chinese learning Japanese using ultrasound to analyze whether the same avoidance strategy is applied for learners with different L1s.The results showed that two of four speakers used [n] for /N/, while one speaker created a tongue blade constriction and one showed a tongue back gesture. It was unclear whether proficiency level affected their articulation because even an advanced learner showed /n/ substitution. English and Chinese speakers seem to have different acquisition processes and strategies for unfamiliar phonemes.

Keywords: Japanese moraic nasal, Standard Chinese, L2, articulation, ultrasound.

1. INTRODUCTION

A Japanese coda nasal, also called a moraic nasal, /N/, which forms a mora on its own, has often been transcribed as a uvular nasal, [ɴ], in the utterance final position [1–3]. Recent articulatory studies have revealed the variability in its place of articulation by speakers from the alveolar ridge to the uvula [4–6], although most speakers produce /N/ as a velar or uvular nasal.

By contrast, for learners of Japanese, it has been observed that /N/ is produced as an alveolar nasal, [n], by native speakers of English [7]. This is probably because there is no uvular sound in the English phoneme inventory and thus native English speakers use a peripheral variation of /N/, which is used in their L1. As all three speakers including an advanced learner have been found to use [n] for /N/ in [7], this seems to be a robust characteristic for learners of Japanese.

To investigate whether this strategy of avoiding an unfamiliar place of articulation is universal across languages, the current study observed the tongue contours for /N/ uttered by native speakers of Standard Chinese, which also has no uvular sound.

Standard Chinese has two coda nasals, /n/ and /ŋ/. The coda /n/ is often produced with a weaker constriction than the onset /n/, resulting in an incomplete closure at the alveolar ridge [8, 9]. The current study first investigated the tongue contours for the Chinese coda nasals and then compared them with the tongue contours for the Japanese moraic nasal uttered by the same speakers. It was hypothesized that Standard Chinese speakers substitute /n/ for the Japanese /N/, as English speakers do.

2. METHODS

2.1. Participants

Six native Standard Chinese speakers (3 males and 3 females) who had been learning Japanese participated in the experiment. Of these, four participants were included in the current analysis to compare learners of different proficiency levels. Two were advanced learners (ACF02 and ACM02), while the other two self-reported as basic-level learners (ACF03 and ACM03). The advanced learners had lived in Japan more than three years and reached the highest level (N1) in the Japanese-Language Proficiency Test. The basic-level learners had lived in Japan for fewer than four months. Two speakers were from Beijing (ACF03 and ACM03), one speaker was from Dalian (ACF02), and one speaker was from Inner Mongolia (ACM02). A speaker ID prefix ‘ACF’ was used for female speakers and ‘ACM’ was used for male speakers. In addition, three native speakers of Japanese (2 females and 1 male) participated in the experiment for comparison purposes, although the articulatory analysis was not conducted for the current study.

The participants were recruited by the authors and were compensated for their participation in the study. They consented to the experimental procedure approved by the ethics committee of Sophia University. They were in their 20s, had no history of speech/hearing impairment, and had corrected-to-normal vision. Language background information about the participants was collected using a questionnaire.

2.2. Speech materials

The speech materials consisted of 29 Chinese words and 15 Japanese words including nasal consonants preceded and followed by a low-vowel /a/ with and without oral consonants designed for a broad analysis. For the current investigation, four Chinese words and one Japanese word were analyzed. The target phonemes were the Chinese onset /t/ in /an.ta/ (安打 [ān.tǎ] 'a hit'), onset /n/ in /an.na/ (安娜 [ān.nà] 'Anna'), coda /n/ in /an.an/ (暗暗 [àn.àn] 'secretly'), coda /ŋ/ in /aŋ.aŋ/ (昂昂 [áŋ.áŋ] 'brave-looking'), and the Japanese coda /N/ in /kaN.aN/ (勘案 ([kaɴ.aɴ] 'consideration'). The Chinese onset /t/ and onset /n/ were used for comparison purposes. The participants read aloud the words displayed on the screen one at a time. The Chinese words were written in Chinese characters and the Pinyin and Japanese words were written in Japanese characters (Kanji and Hiragana). The words were randomized and each word was repeated 10 times. The Chinese trials followed the Japanese trials after a short break.

2.3. Data collection

The audio and ultrasound recordings took place in a soundproof room. The audio signal was digitally recorded monaurally at 22,050 Hz using a RODE-NT2-A microphone through an audio interface (Focusrite, Scarlett Solo 2nd Gen.), connected to a laptop computer.

Real-time mid-sagittal images of the oral cavity were recorded with an ultrasound system (MicrUs, EXT-1H) using a microconvex probe (MC10-5R10S-3). The probe was placed under the participant's chin with an UltraFit Headset [10] stabilizing the relative position of the probe and head. The video frame rate was 113 fps. The tongue images were recorded while the speakers uttered the words and the palate images were recorded while they swallowed water. The audio and ultrasound video were recorded and synchronized using AAA software [11]. Nine tokens were excluded from the analysis due to a synchronization error.

2.4. Analysis

The midpoint of each target phoneme from the audio recording was measured using Praat software [12]. The tongue contours at the acoustic midpoint and palate edge were manually traced using GetContours [13] and interpolated to 100 points on an x/y coordinate for each frame. The origin of the ultrasound fan shape was set to zero. If the tongue images at the acoustic midpoint were unclear, then an image up to three frames before the midpoint was used. One token was not included in the analysis due to poor image quality. The number of analyzed tokens was 190 (5 target phonemes × 10 repetitions × 4 speakers − 10 excluded tokens).

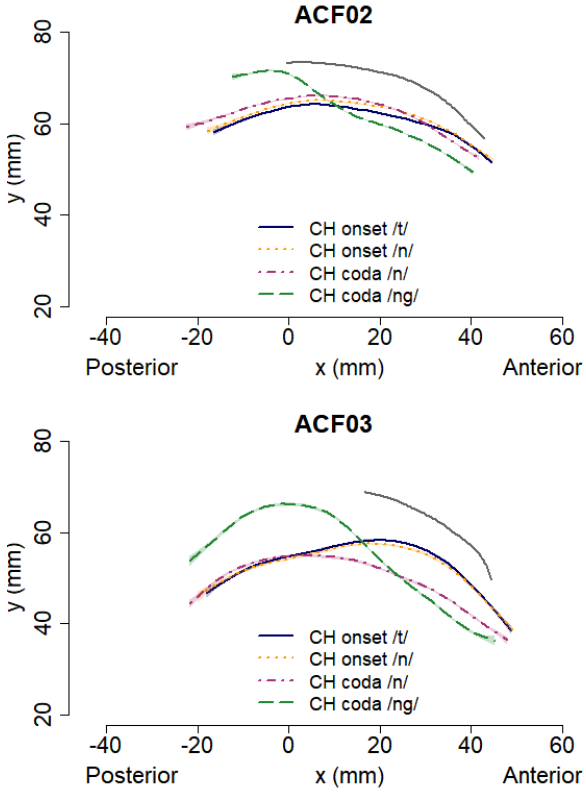
The predicted tongue contours from the repetitions were plotted by speaker for each phoneme using a generalized additive model (GAM) [14]. The contours were trimmed at the observed maximum and minimal points on the x-axis.

3. RESULTS

3.1. Tongue contours for the Chinese phonemes

Fig. 1 shows the predicted tongue contours with 95% confidence intervals for the Chinese phonemes: the onset /t/, onset /n/, coda /n/, and coda /ŋ/ by speaker. All the speakers showed a tongue front gesture for the coda /n/ and a tongue back gesture for the coda /ŋ/.

For three of the speakers (ACF02, ACF03, and ACM03), the onset /n/ showed a significant difference around the tongue tip from the coda /n/, but no significant difference from the onset /t/. As it has been reported [8, 9], the constrictions for the coda /n/ seemed to be weaker than those for the onset /n/ for these three speakers.



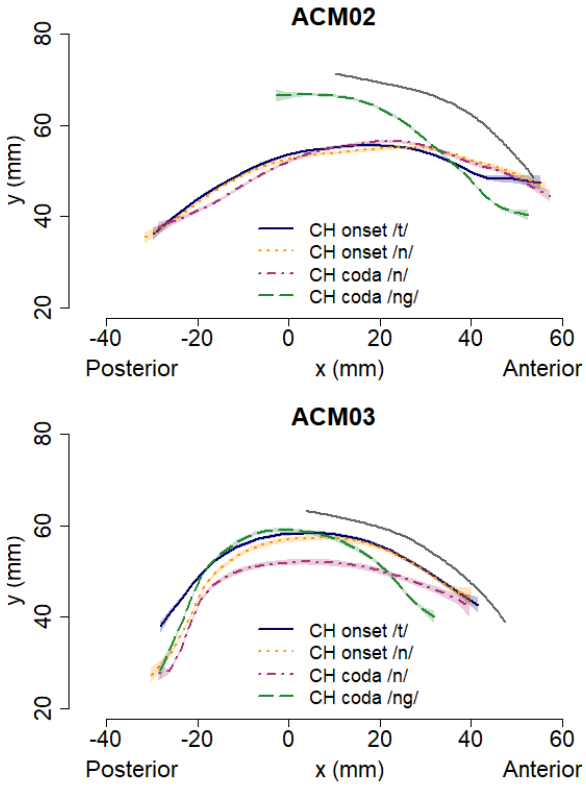
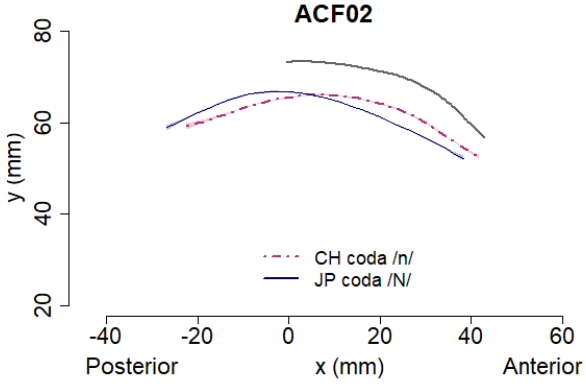


Figure 1: The GAM contours with 95% confidence intervals for the four Chinese phonemes: the onset /t/, onset /n/, coda /n/, and coda /ŋ/ by speaker. The gray line above the tongue contours shows the palate contour.

3.2. Tongue contours for the Japanese moraic nasal

Fig. 2 shows the predicted tongue contours with 95% confidence intervals for the Chinese coda /n/ and Japanese coda /N/ by speaker.

ACF02 and ACF03 showed different tongue shapes between the Chinese coda /n/ and Japanese coda /N/, while the other two speakers showed similar tongue shapes for the two phonemes. ACF02 showed a tongue front gesture for the Chinese /n/ and a tongue back gesture for the Japanese /N/. ACF03, on the contrary, showed tongue front gestures for both the Chinese /n/ and the Japanese /N/, although the tongue blade positions differed.



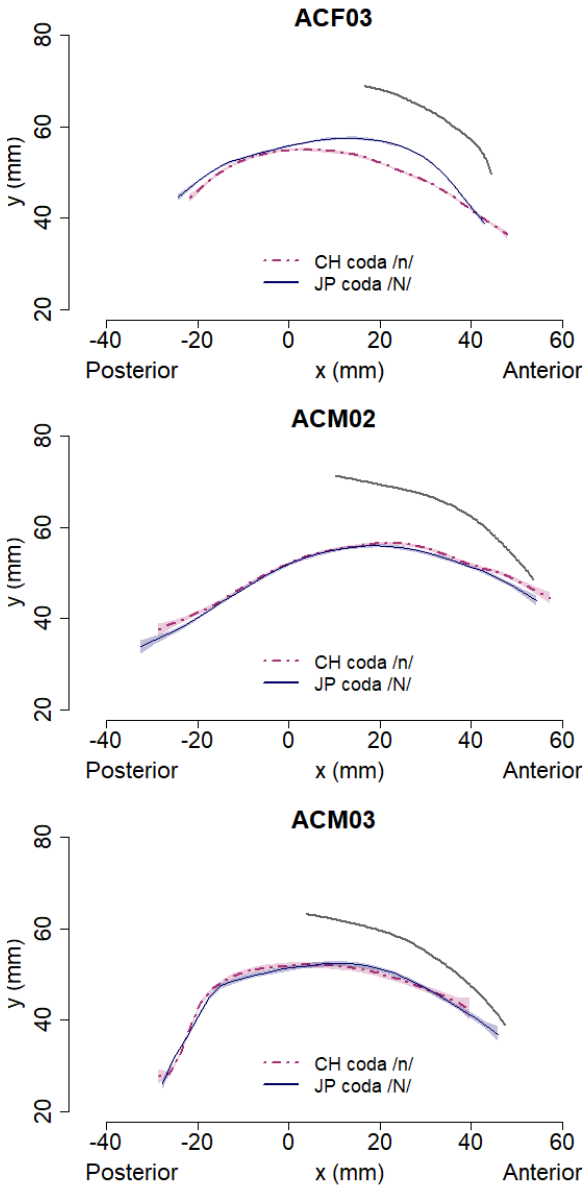


Figure 2: The GAM contours with 95% confidence intervals for the Chinese coda /n/ and Japanese coda /N/ by speaker. The gray line above the tongue contours shows the palate contour.

4. DISCUSSION

The tongue contours for the Chinese coda nasals /n/ and /ŋ/ and the Japanese moraic nasal /N/ by native speakers of Standard Chinese were investigated using ultrasound imaging. It was confirmed that the Standard Chinese speakers used the tongue front gesture for /n/ and tongue back gesture for /ŋ/. Moreover, the coda /n/ had a weaker constriction than the onset /n/ for three of the speakers, in line with [8, 9].

Two speakers (ACM02 and ACM03) showed a tongue front gesture for /N/ that was similar to their /n/ articulation. One of these two speakers was an advanced learner and the other was a basic-level learner. This /n/ substitution for /N/ has been

observed in native speakers of English [7] regardless of their proficiency level. This suggests that, as seen in English speakers, the Japanese moraic nasal was not realized as a back nasal, as most Japanese speakers do, by some of the Standard Chinese speakers, including one advanced learner.

However, one basic-level learner (ACF03) showed a tongue blade gesture for /N/ dissimilar from /n/ and /ŋ/. This speaker appeared to notice some difference in /N/ but could not achieve a tongue back constriction. It is hard to conclude whether this speaker misperceived /N/ as a front nasal less anterior than /n/ or perceived /N/ as a back nasal but had difficulty to articulate as a uvular nasal.

Furthermore, one advanced learner (ACF02) showed a clear tongue contour difference between the Chinese coda /n/ and Japanese moraic /N/. This suggests that the Japanese /N/ could be acquired like native speakers of Japanese by Chinese speakers and the /n/ substitution is not robust, as seen in English speakers. However, it will be necessary to collect data from a sufficient number of speakers to draw a conclusion.

The reasons why /N/ is substituted by /n/ and not /ŋ/ remain unclear. One possible reason is orthography, as [7] pointed out. Since /N/ is often written with 'n' in the Roman alphabet, learners are already used to it when beginning to learn the Japanese writing system. Some Japanese language textbooks for Chinese learners also use 'n' for the utterance final /N/ to explain the sound. However, it seems less common to write Japanese in the Roman alphabet than in English settings. Rather, typing letters on computers might play a role because either 'n' or 'nn' is required to type /N/ in Hiragana.

Another possible reason for the /n/ substitution is the preceding vowel quality. In Chinese, /a/ in /an/ is a front vowel [a], whereas /a/ in /aŋ/ is a back vowel [ɑ] [9]. By contrast, /a/ in Japanese has been described as a centralized vowel [ä] [2, 3] and few descriptions have been reported in relation to the following consonant. If the difference in vowel qualities between /an/ and /aŋ/ are salient enough for Chinese speakers to distinguish the two nasals, vowel quality could be a cue to identify the Japanese /aN/, whose coda consonant is unfamiliar to them. In other words, if the Japanese /a/ in /aN/ is perceived as [a] by Chinese speakers, they would identify the sound as /an/ and thus produce /an/ for /aN/.

To test this hypothesis, we conducted a post-hoc acoustic analysis. Fig. 3 shows the F2 values for /a/ before /n/, /ŋ/, and /N/ by the two native speakers of Chinese (ACF02 and ACM02) and two Japanese speakers (1 female and 1 male). /an/ and /aŋ/ were produced in the speakers' first languages.

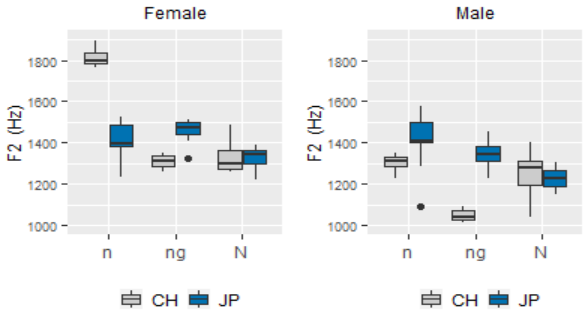


Figure 3: F2 values for /a/ in /an/, /aŋ/, and /aN/.

As shown in Fig. 3, the /a/ uttered by the Chinese speakers had higher F2 values before /n/ than /ŋ/. A linear mixed-effects model analysis [15] and a pairwise post-hoc test [16] revealed that the differences were significant (female: β=499.52, t=17.74, p<0.01; male: β=255.40, t=6.94, p<0.01). By contrast, for the /a/ uttered by the Japanese speakers, no significant differences between /an/ and /aŋ/ were found (female: β=−46.02, t=−1.63, p=1.00; male: β=59.60, t=1.62, p=1.00). However, there were significant differences between /aŋ/ and /aN/ (female: β=132.37, t=4.70, p<0.01; male: β=118.4, t=3.21, p=0.03). A difference between /an/ and /aN/ was significant for the male speaker (β=178.00, t=4.83, p<0.01) and that for the female speaker approached statistical significance (β=86.35, t=3.07, p=0.0508), but did not reach the conventional threshold of 0.05.

The Chinese female speaker, who was an advanced learner and could differentiate /an/ and /aN/ in articulation (Fig. 2, ACF02), had significantly different F2 values between them (β=478.47, t=16.99, p<0.01) and similar values for /aŋ/ and /aN/ (β=−21.05, t=−0.75, p=1.00). The Chinese male speaker, who was an advanced learner but could not differentiate /an/ and /aN/ well in articulation (Fig. 2, ACM02), had similar F2 values for them (β=48.3, t=1.31, p=1.00) and different F2 values between /aŋ/ and /aN/ (β=−207.1, t=−5.62, p<0.01).

The clear difference of the nucleus vowel quality between /an/ and /aŋ/ in Chinese implies that vowel quality could be a cue to perceive these syllables. However, a lack of systematic differences in the Japanese /a/ depending on the coda consonants could result in variability among the speakers to realize the Japanese /aN/, making it harder for learners to use such information.

5. CONCLUSION

/n/ substitution for /N/ by native speakers of Standard Chinese was observed in two of the four speakers. It is necessary to evaluate more speakers to understand their acquisition processes and strategies for unfamiliar phonemes compared with speakers of other languages.

6. ACKNOWLEDGMENTS

This study was supported by JSPS KAKENHI Grant Numbers JP19K13254, JP22J01381, JP23K12212, and Sophia University Special Grant for Academic Research (Research in Priority Areas).

7. REFERENCES

[1] Maddieson, I. 2011. Uvular consonants, In: Dryer, M. S., Haspelmath, M. (eds), *The World Atlas of Language Structures Online.* Max Planck Digital Library.

[2] Okada, H. 1999. Japanese. In: International Phonetic Association. (ed), *Handbook of the International Phonetic Association*, Cambridge University Press, 117–119.

[3] Vance, T. J. 2008. *The Sounds of Japanese*. Cambridge University Press.

[4] Mizoguchi, A., Tiede, M. K., Whalen, D. H. 2022. Inter-speaker variability of articulation for the Japanese moraic nasal : An ultrasound study. *Phonological Studies*, 25, 121–132.

[5] Yamane, N., Gick, B. 2010. Speaker-specific place of articulation: Idiosyncratic targets for Japanese coda nasal, *Can. Acoust.*, 38, 3, 136–137.

[6] Maekawa, K. 2021. Production of the utterance-final moraic nasal in Japanese: A real-time MRI study, *J. Int. Phon. Assoc.*, 1–24.

[7] Mizoguchi, A., Tiede, M. K., Whalen, D. H. 2019. Production of the Japanese moraic nasal /N/ by speakers of English: An ultrasound study. *Proc. 19th ICPhS* Melbourne, 3493–3497.

[8] Dow, F. D. M. 1972. *An Outline of Mandarin Phonetics*. Australian National University Press.

[9] Duanmu, S. 2007. *The Phonology of Standard Chinese*. Oxford University Press.

[10] Spreafico, L., Pucher, M., Matosova, A. 2018. UltraFit: A speaker-friendly headset for ultrasound recordings in speech science. *Proc. 19th Interspeech 2018* Hyderabad, 1517–1520.

[11] Articulate Instruments Ltd. 2022. Articulate Assistant Advanced (AAA).

[12] Boersma, P., Weenink, D. 2021. Praat: Doing phonetics by computer. Version 6.1.55, retrieved October 2021 from http://www.praat.org/.

[13] Tiede, M. K. 2022. GetContours. GitHub repository, https://github.com/mktiede/GetContours.

[14] Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R*. chapman and hall/CRC.

[15] Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting linear mixed-effects models using lme4, *arXiv Prepr. arXiv1406.5823*.

[16] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R package.*

SUPRASEGMENTAL CUES TO PERCEPTION OF ENGLISH LEXICAL STRESS BY MANDARIN SPEAKERS

Mosi He, Ting Zhang, Bin Li,

City University of Hong Kong, Hong Kong SAR, China
mosihe2-c@my.cityu.edu.hk, tzhang343-c@my.cityu.edu.hk, binli2@cityu.edu.hk

ABSTRACT

Lexical stress in English is signaled by a combination of suprasegmental cues, including pitch, duration and intensity. Previous research reported that non-native listeners might not use all cues when perceiving English stress. The current study investigated how pitch and duration cues affected stress identification in resynthesized English disyllabic nonwords by native speakers of English and Mandarin speakers who learn English as a foreign language (NE and EFL speakers). Within each stimulus, the pitch and/or duration cue suggested either a strong-weak or weak-strong stress pattern. Results showed that both listener groups were able to exploit pitch and duration cues to perceive lexical stress. NE speakers used pitch and duration cues similarly for both stress patterns. EFL speakers, however, relied more on the pitch cue than the duration cue, especially for stimuli with a strong-weak pattern. The perceptual differences were attributed to influences of EFL speakers' first-language phonology.

Keywords: lexical stress, speech perception, EFL, Mandarin Chinese, phonetic cue

1. INTRODUCTION

Lexical stress in English is encoded by a combination of pitch, duration and intensity cues [1-3]. Previous research has shown that non-native listeners might not use all cues when perceiving lexical stress due to influences of their L1. Native speakers of non-tone languages, such as Russian speakers, used duration and intensity as cues to lexical stress in English and disregarded pitch cues [4]. In contrast, native speakers of tone languages, such as Vietnamese [5], Mandarin Chinese [6-8], and Cantonese [9], exhibited stronger pitch perception for lexical stress. It is unclear regarding the extent of the effect of L1 on their weighting of pitch and duration cues. The current study aims to investigate how tonal speakers use pitch and duration as cues to English lexical stress.

As a tone language, Mandarin Chinese uses pitch variations in tones to distinguish lexical meanings, while English uses pitch in word and sentence stresses. In addition to pitch, lexical stress in English is indexed by intensity and duration which only serve as secondary cues to tones in Mandarin [10, 11].

The predominant role of pitch in L1 influences perception of lexical stress by Mandarin speakers. In a stress identification task, Mandarin speakers relied on pitch differences for perception of stress in English disyllabic nonwords [6]. Similar findings were reported in the other studies [7, 8]. Comparison between Mandarin speakers' and English speakers' use of duration as cues to stress, on the other hand, showed mixed results. Specifically, Qin et al. [8] reported that speakers of Standard Mandarin used more duration cues than speakers of Taiwan Mandarin in a sequence recall task of English disyllabic nonwords, though both Mandarin groups used fewer duration cues than native speakers of English did. In contrast, Lai [12] found that beginning Mandarin L2 learners of English showed comparable use of duration cues to native speakers of English in identification of lexical stress.

The weighting of pitch and duration to stress perception may be affected by stress pattern. In English, the position of stress within a word indicates word class and syllable structure [13]. Pitch cues were found to facilitate perception of stress when they indicate a weak-strong pattern [4], whereas the findings in [6] showed that a word-initial syllable with higher pitch was more likely to be perceived as the stressed syllable.

To provide further insights into tonal speakers' perceptual cues to non-native lexical stress, the current study examines the effects of pitch and duration cues, and stress pattern on stress identification patterns of resynthesized English disyllabic nonwords by native speakers of English and Mandarin speakers who learn English as a foreign language (NE and EFL speakers). It intended to answer the following research questions: 1) What suprasegmental cues do EFL speakers employ to perceive English lexical stress? 2) Is the cue weighting to English lexical stress similar of different between EFL and NE speakers? 3) Do strong-weak and weak-strong stress patterns affect EFL speakers' perception of English lexical stress?

2. METHODS

2.1. Stimuli preparation

The stimuli were two minimal stress pairs of English disyllabic nonwords (e.g., /'tu.ku/ vs. /tu.'ku/) with a CV.CV syllable structure (C = consonant, V = vowel). The four nonwords were produced four times in a carrier sentence *I said ___ twice.* by a female native speaker and a male native speaker of American English. Two of the four repetitions for each nonword and each speaker, with natural and correct stress placement, were selected and manipulated in Praat [14] in two steps.

Firstly, intensity of all vowels was normalized to 70 dB to minimize the influence of intensity differences on perception of stress. Intensity was excluded from examination because the auditory treatment of the intensity, loudness, is dependent on duration factors [15]. Secondly, F0 and duration of vowels were manipulated so that the stressed-unstressed difference was co-indicated by both cues (higher F0 and longer duration, "F0+Dur" condition) or by a single cue (higher F0 or longer duration, "F0_only" condition or "Dur_only" condition). Specifically, the duration parameters in F0_only condition and the F0 parameter in Dur_only condition were normalized to the average values of all vowels. In addition, the ratios between vowels of the first and second syllables (V_1/V_2) were determined with reference to those in [4, 12]. A ratio larger and smaller than 1 indicates a strong-weak stress and weak-strong pattern, respectively.

Table 1 shows the acoustic values for the resynthesized stimuli across cue conditions and stress patterns from the recordings of the male speaker. The strong-weak (SW) pattern was represented by higher F0 and/or longer duration on the vowel of the first syllable, whereas the weak-strong (WS) pattern was indicated by higher F0 and/or longer duration on the vowel of the second syllable.

2.1. Participants

Thirty-five native speakers of Mandarin Chinese (mean age = 21.1, SD = 1.4) including 20 females and 15 males, participated in the study. They were born and raised in northern China. All were undergraduate students at a university in Guangzhou, China. Self-reported language proficiency and results of a cloze test showed that they were intermediate to advanced EFL speakers. Twenty-one native speakers of American English (mean age = 32.5, SD = 9.6), including 11 females and 10 males, were recruited as a control group. None of the participants reported any speech, hearing, or visual disorders.

Table 1: Pitch and duration parameters for the resynthesized stimuli across cue conditions and stress patterns from the recordings of the male speaker.

			Cue condition		
			F0+Dur	F0_only	Dur_only
F0 (Hz)	SW	V_1	212	212	165
		V_2	133	133	165
		ratio	**1.61**	**1.61**	**1**
	WS	V_1	146	146	165
		V_2	168	168	165
		ratio	**0.87**	**0.87**	**1**
Dur (ms)	SW	V_1	187	160	188
		V_2	127	160	123
		ratio	**1.47**	**1**	**1.53**
	WS	V_1	119	160	113
		V_2	221	160	217
		ratio	**0.54**	**1**	**0.52**

2.4. Procedure

An identification task was conducted online using the Psytoolkit platform [16]. To identify inattentive participants, four attention checks were embedded in the task. Participants took the task in an uninterrupted quiet place, using a laptop or a computer. Those who used devices other than laptops/computers were automatically blocked by the platform.

Before the formal test session, participants took a practice session, where pairs of English real words such as "IMport" /ˈɪm.pɔːrt/ and "imPORT" /ɪm.ˈpɔːrt/ were presented. They listened to one stimulus at a time and indicated whether the stress was on the first or the second syllable by pressing a corresponding button on the keyboard. Feedback was given immediately with the sign "correct" or "wrong". There was a 1500 ms interval between stimuli. If no response were made within 3000 ms, the next audio would be played after 1500 ms.

The same procedure was used in the formal test session, except that there was no feedback given in the test session. There were two blocks in the test session. All stimuli across three cue conditions were presented randomly in each block where the genders of the speakers were counterbalanced.

3. RESULTS

In total, 1344 responses were collected (56 participants × 4 stimuli × 3 cue conditions × 2 blocks). One EFL and two NE participants failed 50% of the attention checks and were excluded from further analysis.

Responses were marked as correct if they matched the stress patterns indicated by the cue conditions. For each participant, accuracy rates were calculated by dividing the number of correct responses by the total number of trials in each cue condition and each stress pattern. Figure 1 shows the average accuracy for the two listener groups across three cue conditions. Both listener groups had accuracy rates above 60% across three cue conditions. But the perceptual patterns varied between the two listener groups. For the EFL group, F0_only condition and F0+Dur condition yielded similar accuracy, whereas lower accuracy was found in Dur_only condition. The NE group, however, had the lowest accuracy in F0_only condition.

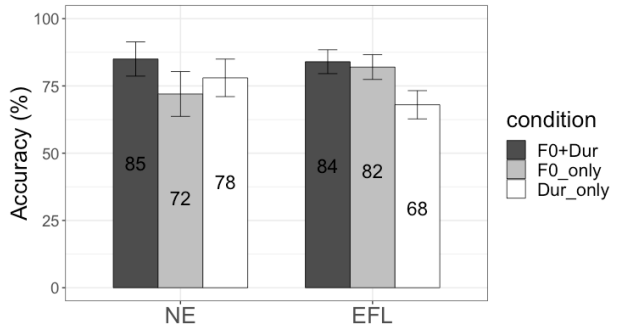


Figure 1: Mean accuracy rates for two listener groups by three conditions

A logistic regression mixed effects model (glmer) was fit to the data using the lme4 package [17] in R [18]. The dependent variable was accuracy, with correct responses coded as "1" and incorrect responses as "0". *Group* (NE vs. EFL), *cue condition* (F0+Dur, F0_only, Dur_only), *pattern* (SW vs. WS), and their interactions were entered as fixed factors. *Participants* and *items* were entered as random intercepts. All fixed factors were coded with sum coding. The Anova() function in the *car* package [19] was used to obtain significance of main effects and interactions. Post-hoc pairwise comparisons, wherever needed, were conducted with the *emmeans* package with Tukey adjustment [20].

Statistical results revealed that accuracy significantly differed across cue conditions [$\chi^2(2) = 12.8, p < .002$]. There was no significant difference in *group* [$\chi^2(1) = .31, p = .58$] or *pattern* [$\chi^2(1) = 2.30, p = .13$]. Two significant interactions were observed, *group* × *cue condition* [$\chi^2(2) = 7.20, p < .05$] and *pattern* × *cue condition* [$\chi^2(2) = 15.10, p < .001$], which will be elaborated in Sections 3.2 and 3.3 below. No other significant interaction was found.

3.1. Cue condition effect

Due to the significant difference among the three cue conditions, pairwise comparisons using Tukey adjustment were carried out. Results showed that F0+Dur condition brought about significantly higher accuracy than Dur_only condition ($b = .67, z = 3.58, p = .001$), while no significant difference was found between F0+Dur condition and F0_only condition, or F0_only condition and Dur_only condition (both bs $<.38$, zs < 1.99, ps $> .11$).

3.2. Group and condition interaction

The significant *group* × *cue condition* interaction indicated that the effect of cue condition varied between the NE and EFL groups. Post-hoc pairwise comparisons revealed that the EFL group had higher accuracy in conditions with F0 cue than Dur_only condition (both bs $> .75$, zs > 3.45, ps $< .002$), whereas the NE group had higher accuracy in conditions with duration cues than F0_only condition, as shown in Figure 1. But the differences did not reach statistical significance (all $|b|$s $< .57$, $|z|$s < 1.95, ps $> .13$).

Additionally, comparison within cue conditions showed that, in the F0_only condition, the EFL group had higher accuracy than the NE group, although this was not statitically significant ($b = .57, z = 1.79, p = .07$). No significant difference between the two groups was found in other cue conditions. Taken together, it indicated that the EFL group employed more F0 cues than the NE group did to encode stress in English nonwords.

3.3. Pattern and condition interaction

There was also a significant interaction between *pattern* and *condition*, suggesting that accuracy among cue conditions was influenced by stress patterns of stimuli in that condition. When averaged over *group*, for F0_only condition, the accuracy of SW nonwords was significantly higher than WS nonwords ($b = .80, z = 3.07, p < .01$). Contrary to F0_only condition, the accuracy in Dur_only condition was lower for SW nonwords than WS nonwords ($b = -.52, z = -2.15, p < .05$).

In addition, the effect of *cue condition* also varied within stress patterns. For SW nonwords, accuracy in conditions with F0 cue was significantly higher than Dur_only condition (both bs $> .95$, zs > 3.68, ps $< .001$). For WS nonwords, accuracy in F0+Dur condition was higher than F0_only condition, but it was not statistically significant ($b = .57, z = 2.27, p = .06$). There was no significant difference between F0_only and Dur_only conditions ($b = -.37, z = -1.49, p = .29$). Overall, listeners tended to identify a first

syllable with higher F0 as stressed and a second syllable with longer duration as stressed.

Although there was no significant three-way interaction among *pattern*, *condition*, and *group*, the two groups showed different perceptual patterns across stress patterns and cue conditions. As shown in Figure 2, the NE group used more F0 cues for SW nonwords and more duration cues for WS nonwords. The EFL group, on the other hand, constantly used more F0 cues than duration cues regardless of stress patterns, and more so for SW nonwords. In other words, EFL speakers showed similar use of F0 cues between the two stress patterns.

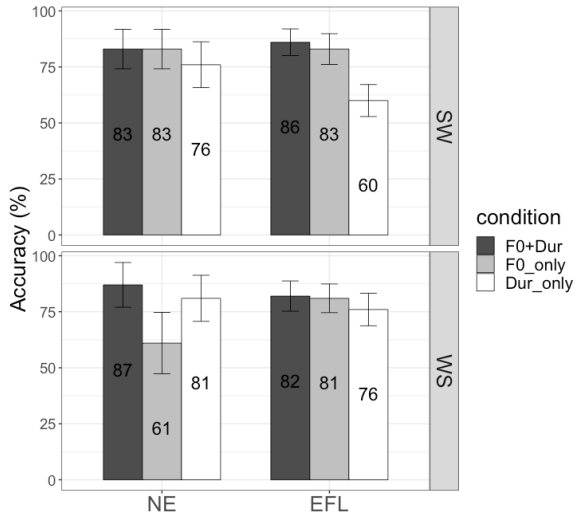


Figure 2: Mean accuracy rates for three conditions and two listener groups by two stress patterns

4. DISCUSSION

The current study examined stress identification in resynthesized English disyllabic nonwords by the NE and EFL speakers. Results showed that both listener groups were able to exploit pitch and duration cues to perceive lexical stress, which suggested that listeners did not attend exclusively to one cue but rather used a variety of cues at the same time [1]. The perceptual pattern, however, appears to vary between the two listener groups.

Firstly, the NE speakers used pitch and duration cues similarly, while the EFL speakers relied more on pitch cues than duration cues. These results are in line with those in [6] and [8]. EFL speakers' reliance on pitch cues to encode lexical stress in English could be explained by the influence of cue-weighting in their L1, where pitch variations in Mandarin contrasting lexical meanings. In addition, the NE speakers used more duration cues than the EFL speakers did, which suggested that lexical stress in English is encoded by the intergral of intensity and duration [15].

Secondly, the NE and EFL speakers employed duration cues to perceive stress in stimuli with the weak-strong pattern in a similar way, suggesting that duration cues also facilitated EFL speakers' stress perception when they indicated a weak-strong pattern. A possible explanation would be that the EFL speakers were influenced by the preference for a weak-strong pattern between syllables at the phonetic level in Mandarin [21, 22]. Specifically, a word-final syllable with sufficient duration and wide pitch range was perceived as more stressed [23]. The current findings contradict those in [8], which could be accounted for by different experimental paradigms. A sequence recall task was used in [8], which required a higher memory load.

Thirdly, the NE speakers alternated perceptual cues for different stress patterns. The EFL speakers, on the other hand, showed reliance on pitch cues irrespective of stress patterns. Similarly, Mandarin speakers in [12] were found to adopt a fixed and inflexible pattern between trochaic and iambic words, while NSE participants altered the weighting of cues accordingly. This could be attributed to a word-by-word learning strategy used by Mandarin speakers when processing stress patterns [24, 25]. Rather than associating stress patterns with grammatical functions, Mandarin speakers might associate an initial stress with the high-level tone and a final stress with the high-falling tone in Mandarin [6, 12]. Taken together, although both two groups used pitch cues to identify stress in English nonwords, the cross-language different weighting of pitch and duration across stress patterns indicated an overall influence of typological differences between tone languages and stress languages.

In conclusion, this study investigated the effects of pitch and duration, and stress pattern on perception of lexical stress in English by NE and EFL speakers. Results revealed the EFL speakers' reliance on pitch cues and inflexible use of cues between stress patterns, suggesting certain influences of their L1 tonal system. The findings contribute to L2 speech acquisition and studies of speech perception from a typological perspective. In future research, we will compare the perception of lexical stress by naïve native speakers of Mandarin to experienced EFL speakers.

5. ACKNOWLEDGEMENT

The study was supported by EDB(LE)/P&R/EL/175/12 funded by SCOLAR of Hong Kong and by TDG #6000807 funded by City University of Hong Kong.

6. REFERENCES

[1] Fry, D. B. 1958. Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

[2] Lieberman, P. 1960. Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, 32, 451–454.

[3] Lehiste, I. 1970. *Suprasegmentals*. MIT Press.

[4] Chrabaszcz, A., Winn, M., Lin, C. Y., Idsardia, W. J. 2014. Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers. *Journal of Speech, Language, and Hearing Research*, 57, 1468–1479.

[5] Nguyễn, T., Ingram, C. L. J. 2005. Vietnamese Acquisition of English Word Stress. *TESOL Quarterly*, 39(2), 309–319.

[6] Wang, Q. 2008. *Perception of English stress by Mandarin Chinese learners of English: An acoustic study*. Unpublished Ph.D. dissertation, University of Victoria.

[7] Yu, V. Y., Andruski, J. E. 2010. A Cross-Language Study of Perception of Lexical Stress in English. *Journal of Psycholinguistic Research*, 39(4), 323–344.

[8] Qin, Z., Chien, Y. F., Tremblay, A. 2017. Processing of word-level stress by Mandarin-speaking second language learners of English. *Applied Psycholinguistics*, 38(3), 541–570.

[9] Lai, W. W. 2017. *The Production and Perception of English Lexical Stress by Hong Kong Cantonese Learners of English*. Unpublished Ph.D. dissertation, University of Hong Kong.

[10] Lin, T., Wang, L. J. 1992. *Course in Phonetics*. Beijing University Press.

[11] Liu, S., Samuel, A. G. 2004. Perception of mandarin lexical tones when F0 information is neutralized. *Language and Speech*, 47(2), 109–138.

[12] Lai, Y. W. 2009. *Acoustic Realization and Perception of English Lexical Stress by Mandarin Learners*. Unpublished Ph.D. dissertation, University of Kansas.

[13] Guion, S., Clark, J. J., Harada, T., Wayland, R. 2003. Factors Affecting Stress Placement for English Nonwords include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words. *Language and Speech*, 46(4), 403–427.

[14] Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer. Version 6.1.08. http://www.praat.org/

[15] Turk, A. E., Sawusch, J. R. 1996. The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America*, 99 (6), 3782-3790.

[16] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. Teaching of Psychology, 44(1), 24-31.

[17] Bates, D., Maechler, M., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[18] R Core Team, R: A language and environment for statistical computing. Vienna, Austria. URL: http://www.R-project.org/, 2020.

[19] Fox, J., Weisberg, S. 2019. An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA.

[20] Lenth, R. V. 2018. *emmeans*: Estimated Marginal Means, aka Least- Squares Means. R package version 1.3.0. https://CRAN.R- project.org/package=emmeans.

[21] Chao, Y. R. 1968. A Grammar of Spoken Chinese. Reprinted in 2011. The Commercial Press.

[22] Yu, V. Y. 2021. Effects of Syllable Position, Fundamental Frequency, Duration and Amplitude on Word Stress in Mandarin Chinese. *Journal of Psycholinguistic Research*, 50(2), 293–312.

[23] Lin, M. C., Sun, G. H. 1984. Preliminary study on stress of disyllabic words in Beijing Mandarin. *Dialect*, (1), 57–73.

[24] Archibald, J. 1997. The acquisition of English stress by speakers of nonaccentual languages: lexical storage versus computation of stress. *Linguistics*, 35(1), 167–181.

[25] Wayland, R., Landfair, D., Li, B., Guion, S. G. 2006. Native Thai Speakers' Acquisition of English Word Stress Patterns. *Journal of Psycholinguistic Research*, 35(3), 285–304.

Regional Dialect Influence on L2 English Production: BATH-STRUT-LOT Triangle for L1 Turkish Speakers

Bahar Aksu

Lancaster University
baharaksuu9@gmail.com

ABSTRACT

Studies examining the role of regional dialects on L2 speech production are scarce. This study examines the role of Turkish regional dialects on L2 English /ɑː/ - /ʌ/ - /ɒ/ production. Acoustic realization of the /ʌ/ phoneme varies between İstanbul and Trabzon Turkish. I test the hypothesis that the regional dialect difference in L1 will influence the L2 English production of /ʌ/ vowel as well as the non-native phonemes (/ɑː/ - /ɒ/) in L2 English among these regional dialect speakers. Fourteen speakers from each region (Trabzon, İstanbul, and Southern England) aged 18-35 were recorded reading a text. Vowels were acoustically analysed in word-medial position and mixed-effect modelling was used to test the regional dialect influence on L2 speech. The findings show that the existing regional dialect variation in /ʌ/ does not transfer to L2 English, while regional dialect speakers significantly differ in /ɑː/ - ɒ/ English production. The results also reveal crosslinguistic variation between SSBE and L1 Turkish speakers.

Keywords: L2 speech production, regional dialects, vowels, Turkish

1. INTRODUCTION

Second Language (L2) speech models acknowledge the influence of L1 phonology on the perception and production of non-native sounds [2,9]. While there is a solid theoretical foundation for how L1 phonology affects L2 speech production and perception, less is known about how inter-speaker variance in the L1 affects the L2. These models do pose the hypothesis that within-L1 variation that can affect the L2 speech learning [6,9] and one way of examining within-L1 variation is to look at L1 regional dialect variation on L2 speech. Recent studies on L2 speech perception found that L1 regional dialect variation can lead to variation in L2 perception. For example, Escudero et al. [7] found that Dutch speakers of different regional dialects varied in their perception of L2 English vowel contrasts. Research on the influence of L1 regional dialects on L2 speech production, however, has only found L1 dialect partial effects on the L2 [11,13]. Therefore, further research into the specific dynamics of how structured L1 variation affects the results of L2 speech production is clearly necessary.

Another scarcity in studies on regional dialect and L2 speech processing is typological variation between the target L1 and L2. Turkish, a member of the Altaic language family, is more typologically from English than many other Indo-European languages, such as Spanish or Dutch, which are common examples in L1-L2 English speech studies. Turkish has eight phonologically contrastive (front-back, high-low, rounded-unrounded) vowels. Among these, /ʌ/ is phonologically classified as a low back vowel [10], while some research describes it as low central phoneme [14]. While phonetic documentation of regional dialects in Turkish is another neglected subfield, a recent study showed that Trabzon and İstanbul dialects differ in the production of /ʌ/ in L1 due to the influence of neighbouring languages in the regions [1].

This study aims to examine how regional dialect variation of /ʌ/ in Turkish influences the L2 English speech production of /ʌ/ - /ɑː/ - /ɒ/ vowels in an EFL setting. In order to test this, I examine the acoustic realization of these vowels produced by İstanbul and Trabzon Turkish dialect speakers in L2 English. As the acoustic realization of /ʌ/ is different between İstanbul and Trabzon dialect in Turkish, I predict that this difference can lead to variation in L2 English between these groups. It is worth noting that /ɑː/ – /ɒ/ is not a native contrast for L1 Turkish speakers. Since there is not any available acoustic data on the allophonic variation of these sounds in L1 Turkish, I hypothesise the production of these non-native vowels in L2 English will behave according to the *similar sound* scenario of SLM-r [9]. While I do not predict any regional dialect influence on the production of /ɑː/ – /ɒ/ non-native vowels, I hypothetically claim that the acoustic realization of these non-native sounds may merge with /ʌ/ for L1 Turkish speakers due the equivalence classifications as framed by SLM-r. I do make the clear prediction that L2 English production of /ʌ/ will be different between the regional dialect speakers.

2. METHODS

2.1. Participants

Twenty-eight participants aged 18-35 from Trabzon and İstanbul regions were recruited for the speech production experiment. In addition, 14 Standard Southern of British English speakers were recruited to compare L2 English production of Turkish speakers with native speakers. All Turkish participants received English as a Foreign Language in classroom from year 2 to the end of high school, however the proficiency level varies between speakers, and it is not controlled between groups and speakers. Participants had not spent longer than 6 months in any English-speaking country and SSBE speakers also reported that they did not spend more than one year in a different country.

2.2. Speech Materials

The text "The Boy Who Cried Wolf" was selected for speech data collection [4]. This text presents all English monophthongs in word-medial position. Two repetitions for each vowel were obtained from the text in word-medial position except for /ɑ:/, which was observed only once in word-medial position.

2.3. Procedure

Participants were given a consent form explaining the experimental procedure and data protection policy prior to data collection. Participants were recorded in quiet rooms in the related regional dialect locations, namely Trabzon, Istanbul, and England. The text was presented to the participants on a computer in a PowerPoint presentation. The session was audio recorded using a Zoom H1 at 44,100 Hz and a 16-bit quantization, which files saved as WAV files. Participants were requested to read aloud the text on the screen at their normal pace.

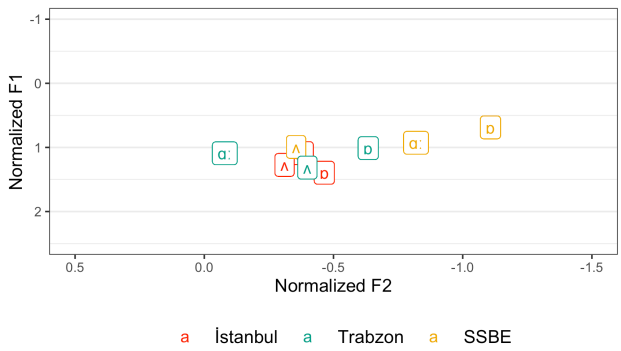
2.4. Acoustic Analysis

The audio files were annotated in ELAN [5] and divided into tokens and then each token was tagged for participant, phoneme, word, and dialect. This was followed by the acoustic coding of phonemes in Praat. The onset and offset of the vowels were determined based on periodicity on the waveform and onset/offset of the formants on the wideband spectrogram. All tokens were downsampled to 11,025 kHz and static formant values at midpoint of vowels were obtained through custom scripts in Praat. Acoustic results were then transferred to R [12] for statistical analysis. First, all vowels were normalized using within-speaker Lobanov *z*-scoring. Mixed-Effect modelling with a restricted maximum likelihood ratio were fitted to the data. Models were fitted separately to each lexical set to simplify the model. 'Dialect' was set as a fixed effect, and 'word' and 'speaker' were set as random intercepts. The random variable 'word' was removed from the model for lexical set /ɑ:/, as to the observation of the vowel only in one word for each speaker. Due to small sample size of the dataset, random slopes were not included as they caused convergence issues in the models.

3. RESULTS

The results showed that both İstanbul and Trabzon regional dialect speakers produce English /ʌ/ in similar acoustic values to SSBE speakers, also showing that the L1 regional dialect difference of this phoneme is not observed in L2 English. However, both regional dialect speakers exhibit different production patterns for the non-native /ɑ:/ – /ɒ/ contrast in L2 English (See Figure 1). Trabzon dialect speakers produce a front-back contrast for /ɑ:/ – /ɒ/, whereas the two are nearly merged with /ʌ/ for İstanbul speakers.

Figure 1: F1~F2 plot of STRUT-BATH-LOT vowels across dialect groups

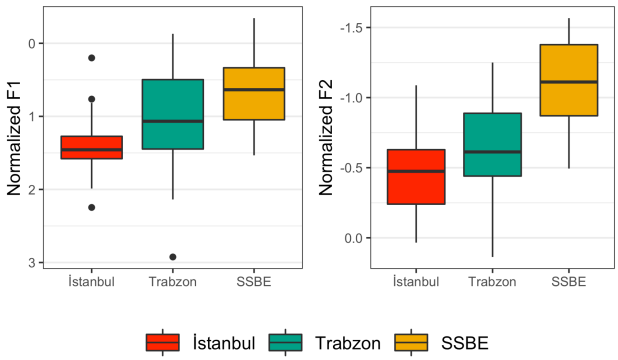


Statistical results revealed that there was no regional dialect effect on F1 ($\chi2$ (2) = 5.539, p =.062), F2 ($\chi2$ (2) =1.461, p = .481), F3 ($\chi2$ (2) =0.716, p = .698), or duration ($\chi2$ (4) = 4.496, p = .105) for /ʌ/. In addition, there was not any cross linguistic difference in the production of /ʌ/ as L1 Turkish and SSBE speakers produced this vowel very similarly. With regard to /ɑ:/, a non-native phoneme for L1 Turkish speakers, the results showed that there was a significant regional dialect influence on F2 ($\chi2$ (2) =-3.887, p = .0001), and duration ($\chi2$ (2) =-7.472, p = .0001). However, there was no significant effect of dialect on F1 ($\chi2$ (2) =-0.246, p = .537) and F3 ($\chi2$ (2) = - 1.772,

p = .309). SSBE speakers produced the longest and the least fronted /ɑ:/ compared to L1 Turkish speakers of İstanbul and Trabzon dialects.

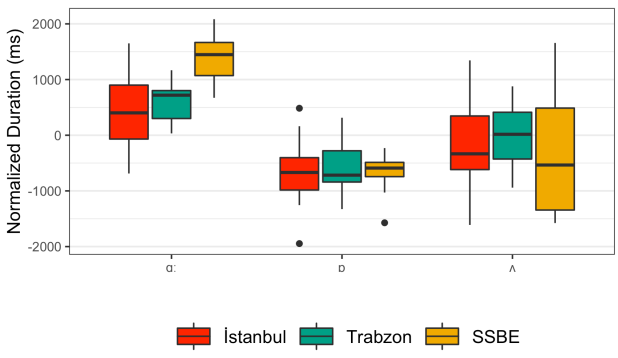
Lastly, LRT model comparison showed that there was a significant effect of dialect on F1 (χ2 (2) =-18.80, p = .0001) and F2 (χ2 (2) =-38.26, p = .0001) in the production of /ɒ/, while there was no significant effect of dialect on F3 (χ2 (2) =1.582, p = .453), and duration (χ2 (2) = 0.690, p = .708). Figure 2 demonstrates that Trabzon and İstanbul speakers produced /ɒ/ differently in terms of F1, while Trabzon and SSBE speakers showed similarity. In terms of F2, SSBE speakers significantly differed from L1 Turkish speakers of both dialects by producing it as a back vowel.

Figure 2: F1 and F2 values of /ɒ/ among regional dialect speakers



In summary, these results revealed an influence of regional dialect on the production of non-native phonemes /ɑ:/ and /ɒ/, while the existing regional dialect difference in /ʌ/ is observed to diminish in L2 English production of L1 Turkish speakers. In terms of duration, it is found that there is an L1 difference on the production of /ɑ:/, while the other vowels produced similarly between the regional dialect and SSBE speakers (See Figure 3).

Figure 3: Duration Boxplot of /ɑ:/ - /ɒ/ - /ʌ/ across dialect groups



4. DISCUSSION AND CONCLUSION

This paper aimed to examine the regional dialect influence on L2 English speech production among L1 Turkish speakers. The focus was on low vowels due to existing regional dialect variation of /ʌ/ in L1, and its relation to the acquisition of non-native /ɑ:/ – /ɒ/ contrast in English. Although the acoustic realization of /ʌ/ is different between Trabzon and Istanbul regional dialects, statistical analysis revealed that neither regional dialect nor L1 affected the production of /ʌ/ in L2 English. /ʌ/ is found to be the most fronted among SSBE and İstanbul speakers, whereas it is /ɑ:/ that is most fronted for Trabzon Turkish speakers. In addition, we observed that F2 and duration of /ʌ/ are similar across all dialect groups. It is somewhat surprising that L1 Turkish speakers produce /ʌ/ similarly in L2 English, despite the regional dialect difference. One potential reason for this could be the change in the phoneme inventory size of the L1 and L2. According to the SLM-r, L2 English learners of L1 Turkish may be able to easily recalibrate the phonemic category of /ʌ/ since it represents an '*identical sound*' scenario in which L2 speakers can produce the target sound with a high accuracy due to direct transfer from L1.

On the other hand, both /ɑ:/ and /ɒ/ create '*similar sound*' scenarios of SLM-r. That is, L2 leaners may not be able to discriminate the phonetic differences of non-native sounds if they are categorically very similar to one another. /ɑ:/ – /ɒ/ do not form a phonemic contrast in Turkish. According to SLM-r, producing non-native L2 sounds as a separate category depends on the distance of these sounds from its nearest category in L1. The results in this study revealed that L1 regional dialects lead to variation in L2 English production of /ɑ:/ and /ɒ/ among Turkish speakers, despite the fact that these phonemes do not exist in L1 Turkish. /ɑ:/ is produced the most fronted by Trabzon dialect speakers while its production is nearly similar to /ʌ/ among İstanbul speakers. In addition, as can be seen in Figure 3, SSBE speakers produce /ɑ:/ significantly longer than L1 Turkish speakers of both dialects. These results overall showed that F2 and duration are the key discriminators between İstanbul and Trabzon dialect speakers for the production of /ɑ:/ in L2 English.

With regard to /ɒ/, İstanbul speakers produced a fronted /ɒ/ while Trabzon and SSBE speakers produced this vowel similarly (less fronted). The regional dialect difference is also observed in F1 that /ɒ/ is produced the highest among İstanbul speakers and the lowest among Trabzon speakers, which caused İstanbul speakers' production to be significantly different than SSBE speakers.

Overall, the findings of this study demonstrate that L1 Turkish regional dialect speakers of L2 English rely on different acoustic correlates for each novel L2 sound. Trabzon dialect speakers produce an open-

mid, back /ɒ/ and a significantly fronted /ɑː/, while İstanbul dialect speakers tend to merge /ɑː/ – /ɒ/ with /ʌ/. According to the 'category precision' hypothesis of SLM-r, it may be argued that İstanbul speakers were not able to form new phonetic categories for these new L2 sounds due to their phonetic closeness to their equivalent in the L1 (i.e., /ʌ/). However, Trabzon speakers formed a moderate front-back contrast for /ɑː/ – /ɒ/ in L2 English. İstanbul and Trabzon speakers produce /ʌ/, which is categorically different across dialects in L1, similarly in L2 English while they differ in the acoustic mapping of /ɑː/ and /ɒ/, which are both allophones of the /ʌ/ vowel in Turkish. These findings, taken together, suggest that learning non-native contrasts in L2 can benefit from both phonemic and acoustic realizations of allophonic diversity found in regional dialects of L1. That is, L1 regional dialects may not only results in one-to-one phonemic change in L2 speech variation, but also can affect the phonetic re-mapping of non-native sounds. What caused this regional dialect influence on the production of /ɑː/ and /ɒ/ in L2 English remains unclear as the allophonic variation of /ʌ/ in Turkish is not well-documented.

Aside from the regional dialect influence, it is worth mentioning that these results showed a crosslinguistic difference that L1 Turkish speakers produced /ɑː/ shorter than SSBE speakers. Although the duration of /ɑː/ is significantly longer than other L2 sounds, L1 Turkish speakers produce this non-native phoneme shorter than SSBE speakers. Duration is not a phonemic contrast in Turkish, but these results suggest that Turkish learners are able to employ duration as an acoustic cue in L2 English, though the degree of the duration is not fully in line with the production patterns of SSBE speakers.

In conclusion, these findings suggest that the acquisition of non-native contrasts in the L2 is sensitive to regional dialect variation in the L1. In addition, it demonstrates that the regional dialect may not directly lead to one-to-one L2 speech variation. Instead, it may play a role in the phonetic re-mapping of non-native sounds. Further research is needed to elucidate the role of regional dialects on L2 perception and production as well as taking L1 allophonic variation into consideration.

5. ACKNOWLEDGEMENTS

This research was supported by the Turkish National Ministry of Education. I thank the anonymous reviewers for their time and input. I also thank Dr Sam Kirkham and Dr Luke Harding for their valuable comments in the previous version of this paper.

6. REFERENCES

[1] Aksu, B. "*The influence of Turkish Regional Dialects on L2 English Speech Production*" Ph.D. dissertation, Dept. of Ling. and Eng., University of Lancaster, Lancaster, 2022

[2] Best, C. T., & Tyler, M. D. 2007. Nonnative and second-language speech perception. Commonalities and complementarities. In: M. J. Munro and O. S. Bohn (eds.), *Second language speech learning: The role of language experience in speech perception and production*. Amsterdam: John Benjamins. 13 -34

[3] Chládková, K., & Jonáš Podlipský, V. 2011. Native dialect matters: Perceptual assimilation of Dutch vowels by Czech listeners. *The Journal of the Acoustical Society of America,* 130(4), EL186-EL192.

[4] Deterding, D. 2006. The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, 36(2), 187-196

[5] ELAN (Version 5.8, 2019) [Computer software]. 2019. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

[6] Escudero, P. "*Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization.*" Ph.D. dissertation,University of Utrecht, Utrecht: LOT. 2005.

[7] Escudero, P., Simon, E., & Mitterer, H. 2012. The perception of English front vowels by North Holland and Flemish listeners: Acoustic similarity predicts and explains cross-linguistic and L2 perception. *Journal of Phonetics*, 40(2), 280- 288

[8] Escudero, P., & Williams, D. 2012. Native dialect influences second-language vowel perception: Peruvian versus Iberian Spanish learners of Dutch. *The Journal of the Acoustical Society of America,* 131(5), EL406-EL412.

[9] Flege, J., & Bohn, O. 2021. The Revised Speech Learning Model (SLM-r). In: R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge: Cambridge University Press. 3 -83

[10] Göksel, A., Kerslake, C. 2004. *Turkish*. Routledge.

[11] Marinescu, I. "*Native Dialect Effects in Non-native Production and Perception of Vowels*." Ph.D. Dissertation, University of Toronto, Toronto, ON 2012

[12] R Core Team 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

[13] Simon, E., Debaene, M., & Van Herreweghe, M. 2015. The effect of L1 regional variation on the perception and production of standard L1 and L2 vowels. *Folia Linguistica*, 49(2), 521-553.

[14] Zimmer, K., Orgun, O. 1992. Turkish. *Journal of the International Phonetic Association*, 22(1-2), 43-45.

GEMINATE REARTICULATION IS PSYCHOACOUSTICALLY ROBUST: PHONETIC IMITATION OF POLISH GEMINATES BY SPANISH-CATALAN BILINGUALS

Arkadiusz Rojczyk[1], Lucrecia Rallo Fabra[2]
[1]Speech Processing Laboratory, University of Silesia in Katowice
arkadiusz.rojczyk@us.edu.pl
[2] Interdisciplinary Speech & Language Acquisition Laboratory
Universitat de les Illes Balears
lucrecia.rallo@uib.es

ABSTRACT

The study investigates to what extent rearticulation observed in Polish geminates may be imitated by speakers of Catalan, a language with single-articulated geminates. Polish geminates have optional rearticulation, which appears to be a unique feature of Polish compared to other languages with consonantal gemination. To our knowledge, this is the first study to investigate this phonetic property in phonetic imitation. A total of 20 participants (10 speakers of Catalan and 10 speakers of Polish) took part in the experiment which included three blocks: baseline (read the words); imitation (imitate the model talkers' productions); post-test (read the words again). The results revealed that Catalan speakers imitated rearticulation very effectively, matching the native Polish participants. Moreover, the post-test analyses showed that they retained rearticulation in more than 50% of their productions. We interpret these results in terms of the psychoacoustic robustness of geminate rearticulation and phonetic imitation in a non-native language.

Keywords: geminate consonants, rearticulation, phonetic imitation, second-language speech

1. INTRODUCTION

The accumulating studies on phonetic imitation in a second or foreign language provide evidence that speakers are able to imitate phonetic properties that are absent in their native language after being exposed to a model talker of a foreign language. This effect has been found for multiple spectral and temporal features such as VOT [1, 2, 3], vowel duration [4, 5], formant frequencies of vowels [6, 7, 8, 9], suppression of the release in stop consonants [10], t-glottaling [11], or suprasegmentals [12, 13]. The robustness of imitation is measured as the magnitude of emergence or suppression of a given feature in productions that shadow a model talker in relation to default productions in a baseline condition in which the speakers read a set of words or phrases without any audio prompts. Moreover, the weight of feature retention, sometimes referred to as a carry-over effect, is measured in post-imitation productions. The results from these studies show that L2 learners may temporarily bypass their L1 pronunciation habits and converge with a model talker towards an L2 native-like norm by modifying phonetic properties of their speech. The analysis of post-imitation productions suggests that at least some of the imitators retain the imitated features in subsequent non-imitative productions.

The common denominator of a large number of previous studies on L2 phonetic imitation is that they use English as a target language and they recruit learners, usually of higher proficiency, of the imitated language. Accordingly, relatively little is known about how phonetic imitation works for features that are absent in English and how they are shadowed by imitators who have had no prior exposure to the imitated language. In the current study we fill this gap by investigating the magnitude of phonetic imitation of the phenomenon of rearticulation in Polish consonantal geminates by Spanish-Catalan speakers without any prior formal learning or communicative command of Polish.

2. REARTICULATION IN POLISH GEMINATES

The production of Polish consonantal geminates is characterised by optional rearticulation. Rearticulation in Polish appears to be a rare, if not unique, feature, since it is, to our knowledge, unattested word-internally in any other geminating language. There is a significant between-speaker variation in the articulation type in that some speakers will tend to single-articulate while others will produce most tokens as rearticulated [14, 15]. The rearticulation rate may also depend on the consonantal group, the largest rate was found for affricates, followed by stops, nasals and fricatives [16].

Rearticulation in Polish geminates is manifested as an audible and observable central disruption or discontinuity in the geminate portion. The result is a

sequence of two homorganic consonants, each having its own release burst, rather than one uniform long consonant as in the case of single-articulated geminates. Figure 1 shows visualisation of a single-articulated velar stop geminate in the word *lekka* 'light'. The occlusion portion is undisrupted, indicating the production of the /kk/ as one long consonant. Figure 2 shows the rearticulated /kk/ in the same word *lekka* 'light'. Rearticulation emerges as the visible release of the first /k/.

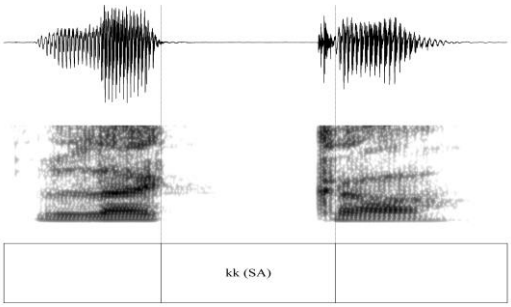


Figure 1: Single-articulated production of the word *lekka* 'light'. The geminate does not have any discontinuities in its total portion

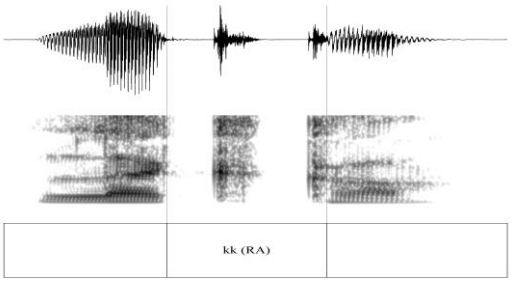


Figure 2: Rearticulated production of the word *lekka* 'light'. Observable release of the first consonant in the central part of the geminate occlusion.

Similar to Polish, Catalan has consonantal geminates, especially sonorant geminates are quite common both in Peninsular Catalan and Majorcan Catalan [17, 18], however, unlike in Polish, they do not form lexical minimal pairs with their singleton counterparts and they are never rearticulated. Accordingly, we hypothesised that Majorcan Catalan speakers would be able to imitate geminate consonants of an unknown language such as Polish, provided that gemination would be signalled in the spelling.

3. THE CURRENT STUDY

In this study we test whether the Catalan speakers would be able to imitate rearticulation of geminates in Polish. More specifically, we address the following research questions:

1. What is the magnitude of imitation of rearticulation by speakers of Catalan?
2. How does the imitative performance of Catalan speakers compare to the one by native speakers of Polish?
3. What is the magnitude of post-imitative retention of rearticulation in subsequent non-imitative productions?

3.1. Participants

Ten native Spanish-Catalan bilinguals (9 females and 1 male) were recruited for the Catalan group. They were all students and staff at the University of the Balearic Islands in Palma de Mallorca, ranging in age from 20 to 28 years (M = 21.5). For 6 participants Catalan was a dominant language, 4 participants reported that it was their second language. The Polish group consisted of 10 native speakers of Polish (7 females and 3 males) with the mean age of 24.6 years (range: 23 to 28). They were all students of the University of Silesia in Katowice. All of them were born and resided in Poland. None of the speakers in either group reported any speech or hearing disorders nor had any indication of such.

3.2. Materials

The materials consisted of 10 two-syllable singleton-geminate pairs with stop and nasal consonants. In order to maintain maximum spelling-to-production clarity for Catalan speakers, some of them were existing lexical units in Polish and some of them were phonologically potential words. Slight spelling changes were made to assist Catalan speakers, for example the Polish orthographic representation 'w' for a voiced labiodental fricative /v/ was altered into 'v'. Table 1 shows the word tokens used in the experiment.

singleton	geminate	singleton	geminate
leka	*lekka*	*seni*	*senni*
pana	*panna*	*dana*	*danna*
vana	*vanna*	*maki*	*makki*
leki	*lekki*	*rana*	*ranna*
maka	*makka*	*paka*	*pakka*

Table 1: Singleton-geminate pairs used in the experiment.

Two male model talkers (native speakers of Polish, qualified phoneticians) recorded the audio prompts for the imitation task. Each of them produced two types or articulation for each geminate: rearticulated and single-articulated. They were recorded in a sound-proof booth in the Speech Processing

Laboratory, University of Silesia in Katowice using a headset dynamic microphone Sennheiser HMD 26 powered by a USBPre2 (Sound Devices) preamplifier at 44 100 Hz. Visualisations of the recorded prompts confirmed that the targeted rearticulated productions had expected discontinuities in the geminate portion.

3.3. Procedure

The experiment consisted of three blocks: (1) baseline; (2) imitation; (3) post-imitation. In the baseline block, the Catalan imitators were informed that they would be presented with words from a foreign language and their task was to show how they thought these words should be pronounced. None of the Catalan participants had any difficulties with confident productions owing to clear spelling-to-sound representations. The Polish imitators were instructed to read the words in Polish in their default pronunciation. In the imitation block, both groups were instructed to shadow the pronunciation of the voices (2 model talkers) they would hear as faithfully as they could immediately after the audio prompts. The rearticulated productions were counterbalanced with an equal number of single-articulated productions. Orthographic representations were provided to facilitate the correct production of whole word forms. In the post-imitation block, the participants were asked to read the tested words again from orthographic representations.

The words were presented in black 80-point font against white background on a laptop (Catalan group) and monitor (Polish group). The items were randomised for each imitator in each block. The audio prompts were delivered through headphones integrated with a headset microphone at a comfortable listening level of 70 dB. The Catalan group was tested in a quiet room in the Interdisciplinary Speech & Language Acquisition Laboratory, University of the Balearic Islands in Palma de Mallorca. The Polish group was tested in the Speech Processing Laboratory, University of Silesia in Katowice. Both groups were recorded with a dynamic headset microphone Sennheiser HMD 26 fed by a USBPre2 (Sound Devices) amplifier with the signal captured at 44 100 Hz. The microphone was positioned approximately 5 cm directionally from the speaker's mouth.

3.4. Measurement and analysis criteria

The collected geminate productions were binary coded as rearticulated or single-articulated. Rearticulation was defined as an observable disruption or discontinuity in the geminate portion indicating the release of the first consonant. There were 40 geminate word forms analysed for each speaker (baseline: 10; imitation 20 (2 model talkers); post-imitation: 10), giving a total of 800 analysed productions. The binary codes were recalculated as proportions of rearticulations for each speaker and each task ranging from 0 to 1. The Mixed Model ANOVA was designed with speaker as a random effect, task and language as fixed effects, and the rearticulation rate as a dependent variable. This model estimates the variance of random factors by constructing sums of squares and cross product matrix for independent variables by using Satterthwaite's method of denominator synthesis.

4. ANALYSIS AND RESULTS

The analysis of the Catalan group showed a robust effect of imitation – the speakers in this group imitated rearticulation at the mean rate of 0.83 (*SE* = 0.1) compared to 0 in the baseline condition. Rearticulation was retained in more than half of the productions (*M* = 0.55; *SE* = 0.09) in the post-test task. There was a significant main effect of task on the rearticulation rate [F (2, 18) = 40.94, p < 0.001, η_p^2 = 0.92] (Figure 3). Post hoc Fisher LSD tests revealed that the post-imitation rearticulation rate was significantly higher than in the baseline condition [p < 0.001], suggesting the effective retention of this feature in subsequent non-imitative productions.

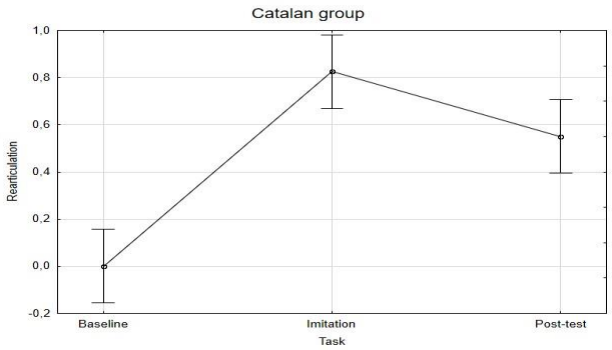


Figure 3: Rearticulation rate in the three tasks (baseline, imitation, post-test) in Catalan speakers' productions.

The speakers of Polish had a mean rearticulation rate of 0.56 (*SE* = 0.12) in the baseline condition. There was an increase to 0.81 (*SE* = 0.08) in imitation, followed by the mean of 0.66 (*SE* 0.11) in the post-test non-imitative productions. The main task effect was also significant in this group [F (2, 18) = 4.3, p = 0.03, η_p^2 = 0.32] (Figure 4). The post-hoc tests showed that the rate difference between the baseline task and the post-test task was not statistically significant [p = 0.5].

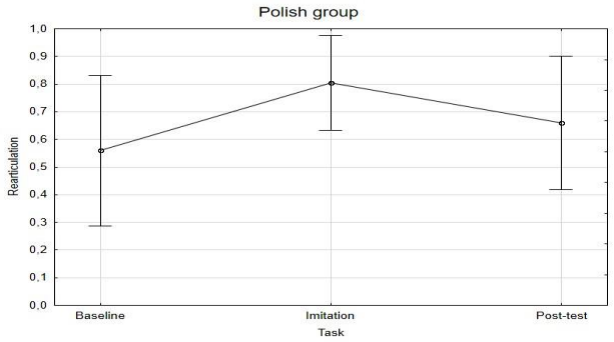


Figure 4: Rearticulation rate in the three tasks (baseline, imitation, post-test) in Polish speakers' productions.

The interaction model showed that the task type interacted significantly with the language group in their effect on the rearticulation rate [F (2, 54) = 5.67, p = 0.006, η_p^2 = 0.17] (Figure 5). The post-hoc tests revealed that the interaction was mainly contributed to by a significant difference between Catalan and Polish speakers in the baseline task (Catalan: *M* = 0; Polish *M* = 0.56; p < 0.001). The was no significant difference between Catalan and Polish speakers for either imitation (Catalan: *M* = 0.83; Polish: *M* = 0.81; p = 0.88) or post-test (Catalan: *M* = 0.55; Polish: 0.66; p = 0.39).

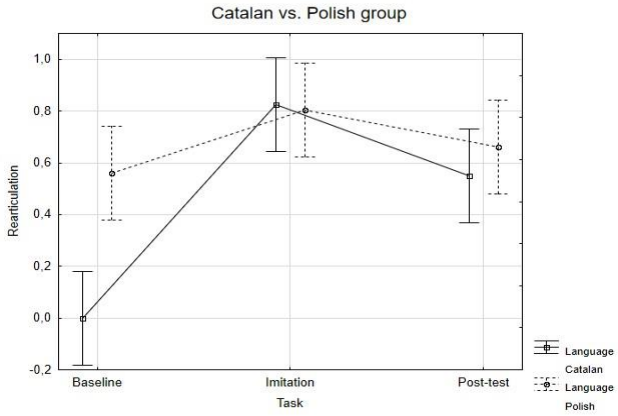


Figure 4: Interaction between task and language. Catalan speakers: solid line, Polish speakers: dotted line.

5. DISCUSSION

The analysis of the results allows us to address the formulated research questions.

What is the magnitude of imitation of rearticulation by speakers of Catalan?

The magnitude of imitation of rearticulation in Polish geminates by speakers of Catalan is high, at the rate of 0.83. It suggests that the Catalan speakers detected this feature and implemented it in their shadowed productions. Such a high rate implies that rearticulation is a psychoacoustically robust feature and may be very effectively imitated even by speakers of the language that has only single-articulated geminates.

How does the imitative performance of Catalan speakers compare to the one by native speakers of Polish?

The performance of Catalan and Polish speakers is very similar with the imitation rate of 0.83 for the former and 0.81 for the latter group. It suggests that rearticulation is effectively imitated independent of whether or not it is implemented in the native language.

What is the magnitude of post-imitative retention of rearticulation in subsequent non-imitative productions?

The retention of rearticulated productions is high with more than half of the rearticulated non-imitative productions. It is especially evident in the Catalan speakers who never produced rearticulated geminates in the baseline task and had a rearticulation rate of 0.55 in the post-imitation productions. Not only was this feature robust in direct imitation but it was also encoded and stored in later productions, suggesting that phonetic imitation might facilitate learning of L2 sounds that do not have an equivalent in the L1.

The rearticulation rates obtained by the Catalan speakers are in line with previous studies exploring phonetic convergence in non-interactive settings [19, 20, 21]. However, the possibility exists that speakers converged on rearticulation while diverging on other dimensions such as VOT or closure durations.

Phonetic convergence has been hypothesized to work as an implicit and automatic mechanism [9]. In non-interactive settings, it involves low-level cognitive processes that prompt the listener to imitate the model speaker. Future research should address whether the same behaviour will transfer to conversational interactions.

Finally, the results reflect the cognitive processes connecting speech perception and production skills in L2 learning, in so far as the Catalan listeners were able to perceive the articulatory gestures of the Polish stops, which were primed during the shadowing and retention tasks.

5. CONCLUSIONS

This study on phonetic imitation of Polish rearticulated geminates by speakers of Catalan showed that:

1. Rearticulation in Polish was very effectively imitated by speakers of Catalan.

2. Rearticulation was retained in more than half of the Catalan speakers' post-imitative productions.
3. There were no significant differences in imitation and post-imitation productions between speakers of Catalan and control native speakers of Polish.

6. ACKNOWLEDGEMENTS

Research supported by the National Science Centre Poland grant *Phonetic imitation in a native and nonnative language* (2019/35/B/HS2/02767) to the first author and grant PID 2020-117804 GB-100 from the Spanish Ministry of Science and Innovation to the second author.

7. REFERENCES

[1] Flege, J. E., Eefting, W. 1988. Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *J. Acoust. Soc. Am.* 83(2), 729–740.

[2] Rojczyk, A. 2012. Phonetic and phonological mode in second-language speech: VOT imitation. Paper presented at EUROSLA 22, September 5–8, Poznań, Poland.

[3] Hashimoto, D., Gomi, K., Shiraishi, R. 2022. Updating L2 phonetic memories: Shadowing training vs. listening training. *Journal of the Phonetic Society of Japan* 26, 13–26.

[4] Zając, M., Rojczyk, A. 2014. Imitation of English vowel duration upon exposure to native and non-native speech. *Poznan Studies in Contemporary Linguistics* 50(4), 495–514.

[5] Podlipský, V. J., Šimáčková, Š. 2015. Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *Proc. 18th ICPhS* Glasgow.

[6] Rojczyk, A. 2013. Phonetic imitation of L2 vowels in a rapid shadowing task.. In: Levis, J., LeVelle, K. (eds), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*. Iowa State University, 66–76.

[7] Jia, G., Strange, W., Wu, Y., Collado, J., Guan, Q. 2006. Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *J. Acoust. Soc. Am.* 119(2), 1118–1130.

[8] Llompart, M., Reinisch, E. 2018. Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech* 62(3), 594–622.

[9] Jiang, F., Kennison, S. 2022. The impact of L2 English learners' belief about an interlocutor's English proficiency on L2 phonetic accommodation. *Journal of Psycholinguistic Research* 51, 217–234.

[10] Rojczyk, A., Porzuczek, A., Bergier, M. 2013. Immediate and distracted imitation in second-language speech: Unreleased plosives in English. *Research in Language* 11, 3–18.

[11] Šturm, P., Przedlacka, J., Rojczyk, A. 2022. Phonetic imitation of t-glottaling by Czech speakers of English. *Linguistica Pragensia* 32(1), 142–165.

[12] Hao, Y. C., de Jong, K. 2016. Imitation of second language sounds in relation to L2 perception and production. *J. Phon.* 54, 151–168.

[13] Ulbrich, C. 2021. Phonetic accommodation on the segmental and the suprasegmental level of speech in native – non-native collaborative tasks. *Language and Speech*. Online First: https://doi.org/10.1177/00238309211050094

[14] Rojczyk, A., Porzuczek, A. 2019. Rearticulated geminates are not sequences of two identical sounds: Evidence from Polish affricate geminates. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proceedings of the 19th ICPhS*. Australasian Speech Science Technology Association Inc., 3671–3675.

[15] Rojczyk, A. 2022. The perception of rearticulated and single-articulated geminates in Polish. *Canadian Journal of Linguistics* 67(1-2), 118–126.

[16] Rojczyk, A., Porzuczek, A. 2019. Durational properties of Polish geminate consonants. *J. Acoust. Soc. Am.* 146(6), 4171–4182.

[17] Recasens, D. 1993. *Fonètica i Fonologia.* Enciclopèdia Catalana, Barcelona.

[18] Wheeler, M. W. 2005. *The Phonology of Catalan*. Oxford University Press.

[19] Pardo, J. S. 2013. Measuring phonetic convergence in speech production. *Front. Psychol.* 4.

[20] Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., Ward, M. 2018. A comparison of phonetic convergence in conversational interaction and speech shadowing. *J. Phon.* 69, 1–11.

[21]Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., Lemhöfer, K. 2021. Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *J Phon.* 88, 101076.

The effect of L2 experience on Mandarin speakers' production of Australian English short and long vowels

Shuting Liu, Olga Maxwell, Janet Fletcher

University of Melbourne
shutingl5@student.unimelb.edu.au, o.maxwell@unimelb.edu.au, j.fletcher@unimelb.edu.au

ABSTRACT

This study explores the effect of L2 experience on Mandarin speakers' production of Australian English short/long vowel pairs: /ɪ - iː/, /ʊ - ʉː/, and /ɐ - ɐː/. Vowel quality (F1/F2 estimates, spectral overlap) and vowel duration are examined. /ɪ - iː/ is found to be produced with the greatest spectral overlap, followed by /ɐ - ɐː/ and then /ʊ - ʉː/, where /ʉː/ is fronted compared to /ʊ/, indicative of the potential influence from Australian English. Durational differences are significant only for /ɪ - iː/. Speakers who have more experience with Australian English (AusE) make greater spectral and durational distinction between the short/long vowels. The results reveal a more nuanced picture with different L2 sources influencing the speakers' vowel production in addition to AusE. This calls for more studies that examine Mandarin L2 English within the 'New Englishes' paradigm.

Keywords: L1 Mandarin, Australian English, L2 experience, vowels, spectral overlap.

1. INTRODUCTION

Adult second-language (L2) speakers are often reported to have difficulties producing L2 vowel contrasts and may rely on partial or different acoustic cues compared to first-language (L1) speakers of the target L2 [1, 2, 3]. L1 Mandarin speakers in the U.S.A. are reported to produce the tense-lax English vowel contrasts, /i - ɪ/, /ɛ - æ/ and /u - ʊ/, which have both spectral (F1/F2) and durational differences, with extensive spectral overlap, and durational contrast maintained for only the high front vowels /i - ɪ/ [4, 5, 6, 7, 8, 9]. Similarly, proficient L1 Mandarin speakers of English, who were born, raised, and educated in Mainland China and mostly exposed to American/British English, are found to spectrally merge tense-lax English vowel contrasts [10, 11, 12], with some studies finding a lack of durational distinction [10, 11] whereas others show the opposite [12]. The term 'China English' (CE) has been used to refer to this emerging English variety, shaped by multiple sources of influence, e.g., language teaching, the media, in China [12, 13, 14, 15].

Extensive research has shown that the phonological system of L1 (acquired very early in life), in combination with L2 experience, critically shapes L2 vowel production [1, 16]. The term 'L2 experience' may be used in diverse ways in the field, but generally indicates the cumulative L2 speech input learners have received during the acquisition process. There is wide agreement that L2 production changes as a function of L2 experience [1, 4]. L2 experience, however, has proved to be difficult to measure [1].

Most previous studies have quantified L2 experience using the index, 'length of residence' (LOR), reported by learners in the format of years of residence in an L2-prevalent environment [1, 4, 17]. LOR, however, may not be able to provide an accurate estimate of the quantity of L2 input learners have received, as it is unlikely that the learners have been exposed to the L2 in a uniform way within a given time interval [1]. Full-time equivalent (FTE) years of L2 input, estimated as LOR multiplied by the proportion of L2 use, is therefore proposed to be a better quantitative measure of L2 experience [1, 18]. L2 experience also varies qualitatively, which may account for the difference or similarity between the performance of L2 learners and has been largely ignored by previous studies [1, 19]. For instance, L1 Spanish speakers, who learned English since childhood and often heard Spanish-accented English, resembled other adult Spanish speakers in producing English /p t k/ with shorter voice onset time compared to L1 English speakers [19].

Research on the effect of L2 experience on L1 Mandarin speakers' English production shows that more experienced learners, measured by LOR in an English-prevalent environment, produced the vowel pairs, /i - ɪ/ and /ɛ - æ/, with greater spectral distinction than less experienced speakers, whereas the two groups did not differ significantly in the temporal distinction they made for these contrasts [4, 17, 20]. The research, however, has been mainly carried out in the U.S.A., where the L2 speakers were exposed to American English, an English variety already familiar to them back in China via education and media, etc. [12, 15]. It is therefore unsurprising that L1 Mandarin speakers of English recorded in China and the U.S.A. had consistent vowel production patterns (i.e., spectrally merged tense-lax vowels), given the same

L1 background and the exposure to shared and similar English varieties. This leads to the question of how exposure to another variety of English, which is less familiar to Mandarin learners from China, would influence their use of acoustic cues associated with the production of varied English vowel categories.

To fill this gap, the present study examines English vowel production of L1 Mandarin speakers who migrated to Australia as adults and had limited exposure to Australian English (AusE) before their arrival to Australia. The L2 production is assumed to be influenced by L1 Mandarin, the mainstream varieties of English the speakers were exposed to before and after the migration (American/British English), and AusE. Given that research on CE commonly factors in the influence of Mandarin and the combined effect from American/British English, characteristics of CE vowels, along with features of Mandarin and AusE vowels, could be used as comparison baselines to reveal the effect of experience with AusE on the production. The effect of L2 experience is also examined quantitatively, using both LOR and FTE as measures. Further, most findings reported to-date are limited to vowels in stressed monosyllabic words, but varied stress patterns may be associated with different acoustic cue use [21, 22]. The present study, therefore, examines three AusE vowel contrasts, /ɪ - iː/, /ʊ - ʉː/, and /ɐ - ɐː/, in trochaic- and iambic-stress pattern words.

1.1. Mandarin and China English vowels

The present study follows [23]'s analysis of Mandarin vowels as a six-vowel system (see Fig. 1 - left), including two high front vowels /i y/, one high back vowel /u/ and an open front vowel /a/. Duration is not phonemically contrastive for Mandarin vowels [23, 24], but it is claimed to be a universally salient cue that could be easily employed to make a vowel distinction [16]. A lack of durational difference in L2 vowel production, therefore, may indicate the influence from L1 Mandarin.

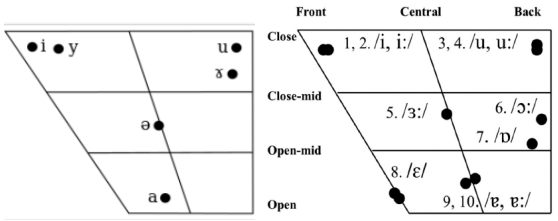


Figure 1: Vowel charts for monophthongs in Mandarin (left, [24]) and China English (right, [12]).

Based on previous research, [12] proposes a ten-vowel inventory for CE (see Fig. 1 - right), with the high front vowel pair /i - iː/ and the back vowel pair /u - uː/ showing a high degree of quality merging. The open central vowel pair /ɐ - ɐː/, in contrast, shows less degree of quality merging, with /ɐː/ being further back than Mandarin /a/ but more front than American English /ɑ/ [9]. All these vowel pairs are contrastive in length.

1.2. Australian English vowels

AusE has 12 monophthongs (see Fig. 2), most of which are distinguished in both quality and quantity [25]. There are six short vowels and six long vowels, with short vowels being approximately 60% the length of the long vowels in the /hVd/ context [26]. The three short/long vowel pairs we selected for the current study, /ɪ - iː/, /ʊ - ʉː/, and /ɐ - ɐː/, have a varied degree of spectral overlap/separation in AusE. Specifically, /ʊ - ʉː/ has the greatest spectral difference, followed by /ɪ - iː/ and then /ɐ - ɐː/, with /ɐ - ɐː/ lacking spectral differentiation [25].

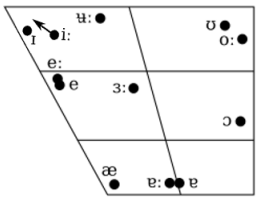


Figure 2: Vowel chart for Australian English ([25]).

1.3. Research questions and hypotheses

The research questions are: 1) What are the durational and spectral characteristics of the short/long vowels produced by L1 Mandarin speakers residing in Australia? 2) What is the effect of experience with AusE on the durational and spectral cues in the L2 vowel production?

We assume greater influence of AusE if the speakers show the tendency: 1) to make greater spectral separation for /ʊ - ʉː/ than /ɪ - iː/, and the least spectral distinction for /ɐ - ɐː/, with all three vowel pairs contrasting in duration; 2) to produce a more fronted high back vowel, given the presence of the back vowel /u/ and the absence of a high central vowel in Mandarin, CE, or American/British English. In AusE, the /ʉː/ vowel is central/fronted compared to other mainstream English varieties [25].

2. METHOD

2.1. Participants

Eight (4F; 4M) L1 Mandarin speakers were audio-recorded. At the time of recording, they aged 24 to 38 years, with an LOR in Australia between 2-10 years (M = 5.50, SD = 3.16) and a self-reported L2 use

proportion between 40%-90%. Four (2F; 2M) speakers had significantly longer LOR (M = 8.75) than the other four (M = 2.75; p = 0.005), but their proportional L2 use was not significantly different, so the group with longer LOR also have significantly longer FTE (M = 5.20) than the other group (M = 1.78; p = 0.013).

2.2. Materials and procedure

Twelve disyllabic words containing the target vowels were selected for the study. As shown in Table 1, 50% of the words have the main stress on the initial syllable (trochaic, Ss) and 50% have the stress on the final syllable (iambic, sS). These words were embedded in short declarative sentences. The speakers were asked to produce the sentences, displayed on a computer screen one by one, as answers to pre-recorded questions, so that the target words were in narrow focus. The recordings took place in the Recording Studio at the University of Melbourne, Australia, using a Zoom H4N recorder at 44.1kHz. Each carrier sentence was elicited 5 times per speaker, with the sentences presented in randomised order. The four most natural productions were selected for subsequent analyses.

Vowel Contrasts	Stress Pattern	
	Trochaic (Ss)	Iambic (sS)
/ɪ - iː/	L**i**ly - L**ee**na	conv**i**nce - conv**e**ned
/ʊ - ʉː/	w**o**man - r**u**mor	af**oo**t - rec**ou**ped
/ɐ - ɐː/	r**u**nner - ll**a**ma	red**o**ne - emb**a**lmed

Table 1: Words containing the target vowels.

2.3. Analysis

The recordings (384 sentences) were automatically segmented through WebMAUS [27], followed by manual correction of the boundaries of target words and vowels in Praat [28]. Vowel duration (ms) and formant (F1/F2) estimates (Hz) taken at vowel midpoint were extracted. Vowel duration, F1 and F2 values were fit to separate linear mixed-effect models using the lme4 package in R [29] to examine the effects of vowel pairs, phonemic length, stress pattern, L2 experience and their interactions. Post-hoc analyses were carried out with the *emmeans* function in the emmeans R package [30]. Pillai scores were calculated to explore the degree of spectral overlap (ranging from 0-1 with 0 representing total overlap and 1 for total separation) of the vowel pairs [31]. Spearman correlation tests were run between the speakers' individual Pillai scores and their LOR/FTE data to examine whether longer LOR/FTE predicts less spectral overlap of the vowels.

3. RESULTS

3.1. Vowel duration

Normalised vowel duration values are plotted in Fig. 3. Significant differences were found only for /ɪ - iː/ and /ʊ - ʉː/: 1) /iː/ was significantly longer than /ɪ/ (t = -6.21, p < 0.001); 2) /ʉː/ was longer than /ʊ/ in trochaic-stress pattern words (t = -2.55, p = 0.011), but shorter than /ʊ/ in iambic words (t = 6.53, p < 0.001). Vowels in sS words were significantly longer than in Ss words (F(1,354) = 128.96, p < 0.001), indicating a final lengthening effect. Overall, the phonemically long vowels (/iː, ʉː, ɐː/) had longer duration than the short vowels (/ɪ, ʊ, ɐ/) (F(1,354) = 6.46, p = 0.011), and the distinction was greater for the more experienced group than the less experienced group (F(1,354) = 4.78, p = 0.030).

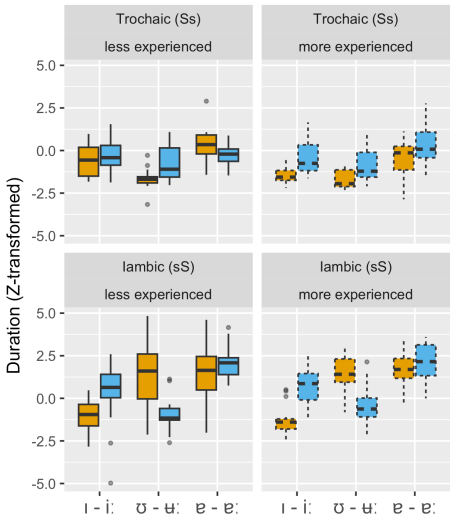


Figure 3: Normalised vowel duration in trochaic- (upper) and iambic-stress (lower) pattern words by less (left) and more (right) experienced speakers.

3.2. Vowel quality

3.2.1. First and second formants

Fig. 4 illustrates the F1/F2 vowel space for the vowels (with Lobanov-normalised F1/F2 values). The three vowel pairs form a triangle in the vowel space, with /ɪ - iː/ being most front along the F2 dimension, followed by /ɐ - ɐː/ and /ʊ - ʉː/ (F(2,354) = 778.16, p < 0.001). /ʊ - ʉː/ were highest along the F1 dimension, followed by /ɪ - iː/ and /ɐ - ɐː/ (F(2,354) = 910.94, p < 0.001). The height difference between /ʊ - ʉː/ and /ɪ - iː/ was only significant in sS words (F(2,354) = 46.46, p < 0.001), which may be an effect of the selected lexical items. Looking at vowels in each pair, no significant F1/F2 differences were found between the front or the open vowels. /ʊ - ʉː/ showed no height

difference either, but /ʉː/ was more front than /ʊ/ along the F2 dimension (t = -3.71, $p < 0.001$).

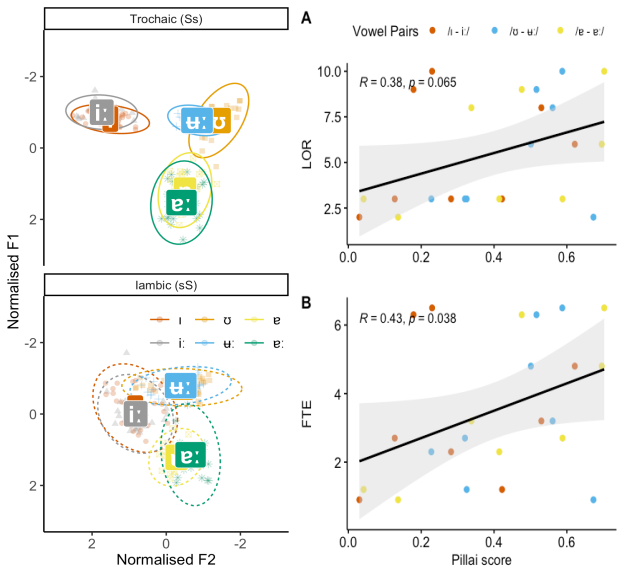


Figure 4 (left): F1/F2 plots with ellipses covering 95% CI from the mean value for each vowel. **Figure 5** (right): Correlation between Pillai scores and LOR (A) / FTE (B).

Fig. 4 also shows that the F1/F2 vowel space is smaller when the stress pattern is iambic compared to trochaic. Specifically, in sS words compared to Ss words: 1) /ɪ - iː/ were more centralised along both F1 (t = -10.92, $p < 0.001$) and F2 dimension (t = 10.39, $p < 0.001$). The F2 centralisation was more prominent for the less experienced group than the more experienced group (t = 3.44, $p < 0.001$); 2) /ɐː/ (but not /ɐ/) was raised along the F1 dimension (t = 3.26, $p = 0.001$); 3) /ʊ - ʉː/ were produced as front as /ɐ - ɐː/ (t = -0.07, $p > 0.05$) along the F2 dimension (t = -8.91, $p < 0.001$).

3.2.2. Spectral overlap

The vowel pair /ɪ - iː/ was produced with the greatest spectral overlap (Pillai = 0.02), followed by /ɐ - ɐː/ (Pillai = 0.08) and /ʊ - ʉː/ (Pillai = 0.10). Vowel category was a significant predictor of the F1/F2 variation for /ɐ - ɐː/ ($p = 0.008$) and /ʊ - ʉː/ ($p = 0.001$), but not for /ɪ - iː/ ($p > 0.05$), confirming that /ɪ - iː/ were produced with the least spectral difference.

As shown in Fig. 5, the Pillai score from individual speakers for each vowel pair was found to correlate with their FTE positively and significantly ($r = 0.43$, $p = 0.038$), suggesting that more L2 experience with AusE is associated with greater vowel spectral separation. In comparison, the Pillai score was positively correlated with the LOR at a marginally significant level ($p = 0.065$) and with a smaller effect size ($r = 0.38$), suggesting that FTE is a stronger predictor of L2 vowel production variation than LOR.

4. DISCUSSION

L1 Mandarin speakers in this study are found to produce the AusE vowel pair /ɪ - iː/ with extreme spectral overlap and salient durational distinction, which is in line with previous findings [4, 5, 6, 7] and with the features reported for the high front vowels in CE [12, 15]. In contrast, /ʊ - ʉː/ show the greatest spectral separation among the three vowel pairs with /ʉː/ being more fronted than /ʊ/, suggesting an effect of AusE on L2 production. The pair /ʊ - ʉː/ is also produced with significant durational difference but in the opposite direction (/ʊ/ longer than /ʉː/) in iambic-stress pattern words, suggesting that the L2 speakers may still be in the process of acquiring the durational cue for this vowel contrast. /ɐ - ɐː/ is produced with somewhat greater degree of spectral separation than /ɪ - iː/, especially in the iambic-stress condition, exhibiting a pattern more characteristic for CE. This vowel pair, however, is produced without significant durational difference, reflecting the influence from Mandarin. The findings suggest that the examination of Mandarin L2 English within the 'New Englishes' paradigm could be a more useful approach as opposed to simply looking at L1-L2 interaction.

Our study also shows that speakers with more L2 experience, which is better indicated by FTE compared to LOR, make greater distinction between the contrasted vowels by producing less spectral overlap and by manipulating vowel duration compared to less experienced speakers. Stress pattern also influences vowel quality and quantity, where vowels have greater spectral overlap and are produced in a more centralised F1/F2 region with greater duration values in iambic words compared to trochaic words. Future vowel studies need to examine vowel quality in varied stress patterns.

5. CONCLUSION

Mandarin L2 English speakers produce the short/long vowel pairs with a varied degree of spectral overlap and durational contrast, not only reflecting influence from L1 Mandarin and exposure to AusE, but also shows the vowel characteristics of 'China English'. More experienced L2 learners, better indicated by FTE, make greater spectral and durational distinction between the vowels. With some findings potentially limited by the selected speech material and the relatively small data set, this study hopefully could shed light on future research direction.

6. REFERENCES

[1] Flege, J. E., Bohn, O.-S. 2021. The Revised Speech Learning Model (SLM-r). In: Wayland, R. (ed), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge University Press, 3–83.

[2] Munro, M. J. 2021. On the difficulty of defining "difficult" in second-language vowel acquisition. *Frontiers in Communication*, 6, 639398.

[3] Tsukada, K. 2009. Durational characteristics of English vowels produced by Japanese and Thai second language (L2) learners. *Australian Journal of Linguistics*, 29(2), 287–299.

[4] Flege, J. E., Bohn, O.-S., & Jang, S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4), 437–470.

[5] Wang, X., Munro, M. J. 1998. English vowel production by native speakers of Beijing Mandarin. *The Journal of the Acoustical Society of America*, 103(5), 3089.

[6] Chen, Y., Robb, M., Gilbert, H., Lerman, J. 2001. Vowel production by Mandarin speakers of English. *Clinical Linguistics & Phonetics*, 15(6), 427–440.

[7] Chen, Y. 2006. Production of tense-lax contrast by Mandarin speakers of English. *Folia Phoniatrica et Logopaedica*, 58(4), 240–249.

[8] Barkana, B. D., Patel, A. 2020. Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems. *Applied Acoustics*, 162, 107203.

[9] Gong J., Yang, Z., Bellamy W., Wang F., Ji X. 2020. Effect of perceptual training on the production of English tense-lax vowels by native Chinese speakers. *2020 International Conference on Asian Language Processing (IALP)*. Kuala Lumpur, Malaysia.

[10] Hung, T. N. 2002. Phonological features of Chinese varieties of English. *The 7th English in Southeast Asia Conference*. Hong Kong Baptist University.

[11] Ao, R., Low, E. L. 2015. A description of the Yunnan English accent. *World Englishes*, 34(3), 336–354.

[12] Wang, Q. 2022. An instrumental investigation of the vowel inventory of China English. *Asian Englishes*, 1-22.

[13] Wang, R. 1991. Zhongguo yingyu shi keguan cunzai (China English is an objective reality). *Jiefangjun Waiyu Xueyuan Xuebao* (*Journal of PLA Foreign Languages Institute*), 1, 1–8.

[14] Xu, Z. 2006. Rectifying "Chinese English". In: Hashim, A., Hassan, N. (eds), *Varieties of English in Southeast Asia and Beyond*. University of Malaya Press, 283–291.

[15] Xu, Z. 2010. *Chinese English, features, and implications*. Open University of Hong Kong Press.

[16] Bohn, O.-S. 1995. Cross language speech production in adults: First language transfer doesn't tell it all. In: Strange, W. (ed), *Speech perception and linguistic experience: Issues in cross-language research.* Baltimore: York Press, 279-304.

[17] Flege, J. E., Liu, S. 2001. The effect of experience on adult's acquisition of a second language. *Studies in second language acquisition,* 23(04), 527-552

[18] Flege, J. E. 1991. Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *The Journal of the Acoustical Society of America*, 89(1), 395–411.

[19] Flege, J. E., Wayland, R. 2019. The role of input in native Spanish late learners' production and perception of English phonetic segments. *Journal of Second Language Studies*, 2(1), 1–44.

[20] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.

[21] Byers, E., Yavas, M. 2017. Vowel reduction in word-final position by early and late Spanish-English bilinguals. *PLoS ONE*, vol. 12, no. 4.

[22] Silber-Varod, V., Khorshidi, N., Levi, L., Amir, N. 2019. The influence of lexical stress on formant values in spontaneous Hebrew speech. *The 19th International Congress of Phonetic Sciences (ICPhS)*, 3538-3542.

[23] Howie, J. M. 1976. *Acoustical studies of Mandarin vowels and tones,* vol. 6. Cambridge University Press.

[24] Lee, W.-S., Zee, E. 2003. Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1), 109-112.

[25] Cox, F. 2012. *Australian English Pronunciation and Transcription*. Cambridge University Press.

[26] Cox, F. 2006. The Acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics*, vol. 26, no. 2, 147–179.

[27] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, vol. 45, 326–347.

[28] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.50, http://www.praat.org/.

[29] Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[30] Lenth R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.0.

[31] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, vol. 34, no. 4, 458–484.

DOES VISUOSPATIAL WORKING MEMORY PREDICT L2 PERCEPTUAL LEARNING FROM PHONETIC TRAINING WITH HAND GESTURES?

Xiaotong Xi[1], Peng Li[2], Pilar Prieto[3,1]

[1]Universitat Pompeu Fabra, [2]University of Oslo, [3]Catalan Institution for Research and Advanced Studies
xiaotong.xi@upf.edu, peng.li@iln.uio.no, pilar.prieto@upf.edu

ABSTRACT

This study assessed whether visuospatial working memory (VSWM) can predict L2 perceptual learning through audiovisual phonetic training with or without hand gestures. Ninety-nine Catalan speakers were trained on the perception of English vowel pairs /æ-ʌ/ and /i-ɪ/ under one of the following three conditions: training without gestures, with hand gestures cueing lip shape, or with hand gestures cueing tongue position. We assessed participants' VSWM via a symmetry span task and vowel perception accuracy through a word identification task before and after training. Results showed that VSWM positively predicted the perceptual learning of /i-ɪ/ only in the no gesture condition, while no such relationship was found in the two gesture conditions or in the learning of /æ-ʌ/. This suggests that VSWM resources might be recruited during audiovisual phonetic training for processing of subtle visual articulation (e.g., the lip spreading of /i-ɪ/ in the no gesture condition) rather than visually salient articulation (e.g., the lip aperture of /æ-ʌ/ in the no gesture condition) or hand gestures movements.

Keywords: Hand gestures, L2 perception, VSWM, phonetic training, English vowel

1. INTRODUCTION

While phonetic training has shown benefits in L2 perceptual phonological learning [1], recent studies showed that training L2 pronunciation with multimodal input was more effective. On top of auditory input, audiovisual training that provides visual input of articulatory information could facilitate L2 pronunciation more than auditory training (e.g., [2]). Moreover, providing kinesthetic information such as hand gestures in audiovisual training showed further benefits for the non-native speech sound production [3]–[5]. However, less strong effects have been found on the perceptual learning of difficult L2 contrasts. For example, hand gestures encoding durational features could help improve the perception accuracy of L2 Japanese long and short vowels, but the effects were not greater than training without hand gestures (e.g., [4], [6]). Similar findings were also obtained when using gestures encoding aspiration feature to train L2 Mandarin aspirated plosives [3], [7].

It might be that individual differences accounted for the null results found in gestural training on L2 speech perception. In terms of auditory training, individual differences such as perceptual ability, phonological awareness [8], auditory selective attention, attention switching [9], and phonological working memory [10] were shown to predict learning outcomes. However, few studies have investigated whether individual differences in visuospatial processing abilities could predict the effectiveness of gestures in L2 phonetic training.

While working memory has been considered an important factor for various aspects of L2 development [11], the visuospatial sketchpad which stores visual and spatial information [12] may be crucial for learners to process multimodal information during learning. A recent study showed that learners with high visuospatial working memory (VSWM) improved more in math through training with hand gestures, whereas no such effect was found for training without hand gestures [13]. Nevertheless, to the best of our knowledge, no previous study has looked into whether VSWM could predict the learning outcome of gestural training in L2.

This study thus investigates the role of VSWM in the perceptual learning of difficult L2 vowel contrasts through audiovisual phonetic training with or without gestures. Two pairs of English vowels /æ-ʌ/ and /i-ɪ/ were chosen. Both pairs are challenging for Catalan learners of English, as they tend to perceive /æ/ and /ʌ/ as Catalan /a/ [14] and rely more on durational cues in distinguishing /i/ and /ɪ/ than native English speakers [15]. In addition, the vowels also differ in articulation. Specifically, /æ/ is produced with a larger lip aperture and more fronted tongue position than /ʌ/ [16], and /i/ shows wider lip spreading and higher tongue position than /ɪ/ [17]. Two types of hand gestures were designed to highlight these key articulation differences for each vowel pair (Figure 1). The "lip hand gesture" illustrated the differences in lip aperture for /æ-ʌ/ and lip spreading for /i-ɪ/ by the distance between the thumb and the four fingers; the "tongue hand gesture" mirrored the mouth roof by one hand and the tongue movement by the other hand to show the tongue positions for /æ-ʌ/ and /i-ɪ/.

Based on previous studies [13], we hypothesized that VSWM abilities would positively predict the perceptual learning of L2 sound contrasts in the phonetic training with hand gestures, while it does not predict the learning outcome with no gestures.

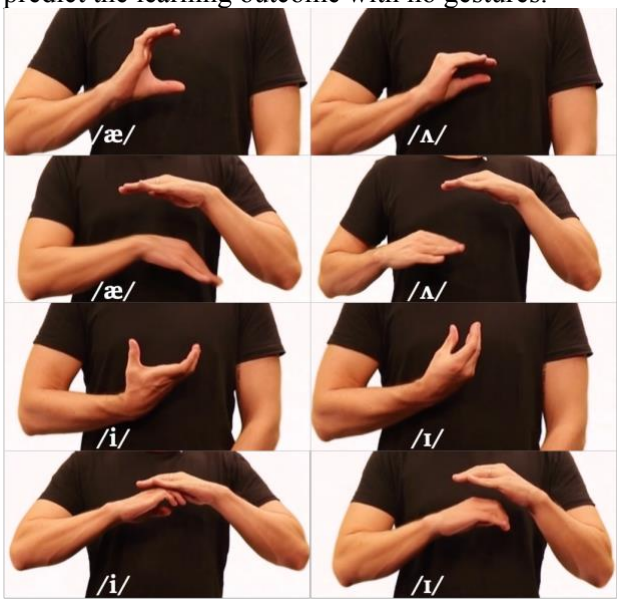


Figure 1: From top to bottom: lip hand gestures for /æ/-/ʌ/, tongue hand gestures for /æ/-/ʌ/, lip hand gestures for /i/-/ɪ/, and tongue hand gestures for /i/-/ɪ/.

2. METHOD

2.1. Participants

Ninety-nine Catalan/Spanish bilingual learners of English (female = 84, M_{age} = 19.7, SD = 1.8) were recruited from a public university in Barcelona. They reported having an intermediate English level. They were randomly assigned to one of the three training conditions, with each condition having 33 participants: No Gesture (NG, 26 females), Lip Hand Gesture (LG, 28 females), and Tongue Hand Gesture (TG, 30 females). Participants within each condition received two training sessions, one for /æ-ʌ/ and the other one for /i-ɪ/, in counterbalanced order.

2.2. Materials

2.1.1. Phonetic training

An American English male speaker was video recorded in a broadcasting studio while producing the materials of the two vowel pairs (/æ-ʌ/ and /i-ɪ/) for the three conditions (i.e., NG, LG, and TG). Each training session consisted of a familiarization phase and a training phase. For the familiarization phase, the instructor introduced the vowel pairs explaining the articulatory differences. In addition, the instructor presented the hand gestures and explicitly explained what they represented for the two gesture conditions.

For the training phase, 6 minimal pairs of English CVC words were selected for each vowel pair (e.g., *cat-cut* /kæt-kʌt/, *beat-bit* /bit-bɪt/). Then, for each training word, a short sentence was created using the word (e.g., A CAT walks by). For the NG condition, the instructor produced the training stimuli without any hand movements. For the LG and TG conditions, the instructor performed the corresponding lip or tongue hand gestures while producing target vowels.

In the end, the familiarization and training videos were uploaded to the Tobii Pro Lab software to create six training videos (2 vowel pairs × 3 conditions). In each video, the familiarization video preceded the training video clips which contained 3 repetitions of training words and sentences. In total, the duration of each training video was about 15 minutes.

2.1.2. Perception test: identification task

A word identification task was used to measure participants' perception accuracy of the two vowel contrasts before, right after training, and one week after training. For each vowel contrast, a total of 18 CVC words were selected. The instructor recorded a total of 36 words and the audio files were uploaded to Alchemer (www.alchemer.com) to build 2 surveys, one for each vowel contrast.

2.1.3. VSWM

We created a symmetry span task adapted from [18] on the platform Pavlovia (https://pavlovia.org) to test participants' VSWM capacity. The task consisted of the spatial recall task and the symmetry judgment task. In the spatial recall trials, a 4 × 4 matrix with a square filled with red was presented. In the follow-up symmetry judgment trial, another 4 × 4 matrix was presented with some squares filled in black and two response options of 'yes' and 'no' below. After 2 to 6 sequences of spatial recall and symmetry judgment, a 4 × 4 matrix was presented. Figure 2 exemplifies a trial of 2 sequences.

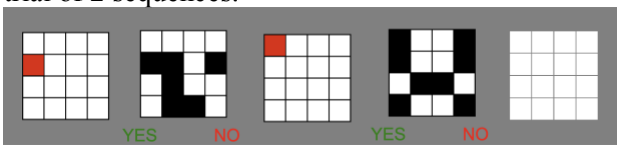


Figure 2: Example symmetry span trial of 2 sequences.

2.3. Experimental procedure

All participants signed a consent form prior to the experiment. They received the training and tests individually in a soundproof room. Half of the participants received the /æ-ʌ/ training first, the other half were trained on /i-ɪ/ first. Before each training session, participants were assessed on their perception of the target vowel pair through the word

identification task. They listened to each of the testing words and had to choose the correct answer from one of the two options which were minimal pair words contrasting only in /æ-ʌ/ or /i-ɪ/. Participants then watched training videos. The LG and TG group saw the instructor produce the gestures accompanying the speech, whereas the NG group only saw the instructor produce the same speech. Participants performed the identification task immediately after the training session. At the end of the experiment, participants completed the VSWM task in which they were instructed to recall the red squares in the correct location and serial order and to decide whether the black-square design was symmetrical or not in between. One week later, participants performed the delayed posttest by using the same identification task. Note that the testing words of the identification task were identical across all three tests, with the order being automatically randomized by the platform.

2.4. Data coding and statistical analyses

For the identification task, participants' correct responses scored 1, and incorrect responses scored 0. The pretest, posttest, and delayed posttest scores of /æ-ʌ/ and /i-ɪ/ were exported from Alchemer. Then, the improvement scores of each vowel pair were calculated for each participant. Specifically, the immediate improvement score was obtained by subtracting the identification sum score of the posttest from that of the pretest and the sustained improvement score was calculated by subtracting the identification sum score of the delayed posttest from that of the pretest. The VSWM score was calculated for each participant by summing up the number that they recalled the position of the red squares in the correct order [18]. One participant from the TG condition scored 0 due to a misunderstanding of the instruction. Thus, the VSWM score of this participant was removed from the database.

In order to check whether participants' VSWM capacity could predict L2 perceptual learning through phonetic training, we ran two Linear Mixed Models (LMMs) using the *lme4* package [19] for the immediate improvement score and sustained improvement score. The fixed effects included Condition (3 levels: NG, LG, and TG), Vowel Pair (2 levels: /æ-ʌ/ and /i-ɪ/), VSWM score, and their interactions. Random factors included a by-participant intercept. The significance for the fixed effects was calculated with Type II Wald Chi-squared tests using the *car* package [20]. For the post-hoc analysis, when significant interactions were observed with VSWM, we examined the slopes of VSWM and conducted pairwise comparisons with Bonferroni adjustment using *emtrends* function from the *emmeans* package [21].

3. RESULTS

Table 1 shows the descriptive data of the perceptual improvement scores across groups and vowel pairs.

	/æ-ʌ/	/i-ɪ/
Posttest – Pretest		
NG	0.1 (1.2)	2.0 (2.2)
LG	0.2 (1.6)	2.4 (2.4)
TG	0.4 (1.8)	2.2 (2.3)
Delayed posttest - Pretest		
NG	-0.3 (2.1)	1.6 (2.2)
LG	0 (1.7)	2.1 (2.2)
TG	0.1 (1.6)	1.5 (2.0)

Table 1: Mean (SD) of perceptual improvement scores across three groups and vowel pairs.

3.1. Improvement from pretest to posttest

Results of the LMM for the immediate improvement revealed a significant main effect of vowel pair (χ^2 = 136.3, $p < .001$) and a significant 3-way interaction of Condition × Vowel Pair × VSWM (χ^2 = 12.9, p = .002). Figure 3 plotted the model prediction on the immediate perceptual improvement from the 3-way interaction. The post-hoc results revealed a significantly positive slope of VSWM in the NG condition when learning /i-ɪ/ (β = 0.06, 95% CI = [0.01, 0.11], p = .028), suggesting a positive relationship between VSWM and L2 perceptual learning of /i-ɪ/ when gesture was absent. Pairwise comparisons showed that the positive slope of VSWM in the NG for the learning of /i-ɪ/ significantly differed from that in the LG (Δ = 0.11, p = .043).

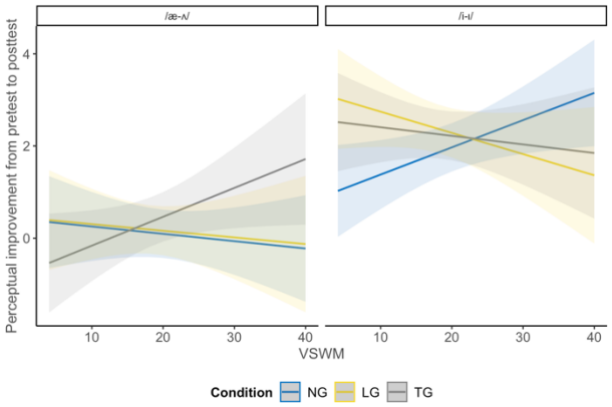


Figure 3. LMM prediction of the perceptual improvement from pretest to posttest across condition, vowel pair, and VSWM. The bands represent the 95% CI.

3.2. Improvement from pretest to delayed posttest

Results of the LMM for the sustained improvement revealed a significant main effect of vowel pair (χ^2 = 89.0, p < .001), a significant 2-way interaction of Condition × VSWM (χ^2 = 6.8, p = .033), and a significant 3-way interaction of Condition × Vowel Pair × VSWM (χ^2 = 7.3, p = .026). Figure 4 showed the model prediction on the sustained improvement from the 3-way interaction. The post-hoc analysis of the 3-way interaction showed that VSWM was a significant predictor of perceptual learning of /i-ɪ/ in the NG condition (β = 0.08, 95% CI = [0.02, 0.13], p = .008). These results suggest that better VSWM abilities predict better sustained improvement in the perception of /i-ɪ/ contrast in the NG condition. Pairwise comparisons revealed that the positive slope observed in the NG condition was significantly higher than those in the LG (Δ = 0.15, p = .006) and TG conditions (Δ = 0.12, p = .017).

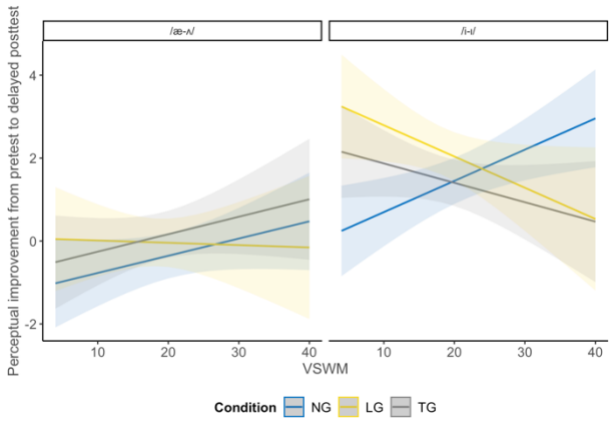


Figure 4. LMM prediction of the perceptual improvement from pretest to delayed posttest across condition, vowel pair, and VSWM. The bands represent the 95% CI.

4. DISCUSSION

This study examined whether VSWM predicts the perceptual learning of L2 English vowels /æ-ʌ/ and /i-ɪ/ through audiovisual phonetic training with or without hand gestures cueing articulation. We assessed the perceptual learning outcome through an identification task and learners' VSWM via a symmetry span task. The hand gestures cued either the lip shape or the tongue movement information. We found that VSWM predicted the learning outcome of /i-ɪ/ in the no gesture condition.

First, contradictory to our hypothesis, VSWM could not significantly predict the identification accuracy of the pair of /i-ɪ/ sounds in audiovisual phonetic training with either of the two types of hand gestures. This result is inconsistent with the results in [13], which might be due to the gesture type and function. Our gestures iconically represented the articulatory movements of the speech sounds but the pointing gestures in [13] pointed to the learning target *per se*. Therefore, if students were good at processing spatial information, gestures would help the learning process by pointing to the target. In our case, the learning target contained abstract knowledge. The processing of this information might require cognitive abilities beyond visuospatial processing abilities. Therefore, VSWM was not a significant predictor of the gesture conditions.

Second, VSWM was not a significant predictor of the learning of /æ-ʌ/ in any of the conditions, including the no gesture condition. A pilot study with 15 participants found that native English speakers distinguished /æ-ʌ/ better than /i-ɪ/ (85% vs. 60%) through visual-only input. Therefore, we hypothesized that the visually subtle articulation differences between /i/ and /ɪ/ would require more VSWM resources. Another possible explanation could be due to the difference in improvement between the two vowel pairs. As noted in Table 1, learners' perception accuracy of /i-ɪ/ improved more than /æ-ʌ/. With such a small improvement and variation (indicated by mean and SD) in the learning of /æ-ʌ/, it could be hard to observe a clear correlation between improvement and VSWM.

This is an initial attempt to assess the role of VSWM in learning L2 speech sounds through hand gestures. There might be several factors that can affect the results. For example, the way in which we test the VSWM (e.g., spatial span task, Corsi's block-tapping task, etc.), the measure for assessing the learning (e.g., perceptual measure or productive measure, subject rating or acoustic analysis, etc.).

In conclusion, this study shows that VSWM can predict perceptual learning of L2 sounds in situations where the visual information is subtle, e.g., in the no gesture conditions where the visual difference of /i-ɪ/ lies in the subtle variation in lip spreading. VSWM resources may thus be recruited in more challenging phonetic training tasks but might not be needed in the processing of visually salient hand gestures cueing articulation during phonetic training.

5. ACKNOWLEDGMENTS

This study was supported by the Ministry of Science, Innovation, and Universities, State Research Agency and European Regional Development Fund [PGC2018-097007-B-I00], and by the Ministry of Science and Innovation [PID2021-123823NB-I00]. The first author is supported by the Ministry of Research and Universities of Catalonia and the European Social Fund [2022FI_B2 00209]. The second author is supported by the Research Council of Norway [223265]. We thank Patrick Louis Rohrer for helping with the creation of materials.

6. REFERENCES

[1] R. I. Thomson, "High Variability [Pronunciation] Training (HVPT)," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 208–231, Dec. 2018, doi: 10.1075/jslp.17038.tho.

[2] V. Hazan, A. Sennema, M. Iba, and A. Faulkner, "Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English," *Speech Commun.*, vol. 47, no. 3, pp. 360–378, 2005, doi: 10.1016/j.specom.2005.04.007.

[3] X. Xi, P. Li, F. Baills, and P. Prieto, "Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features," *J. Speech, Lang. Hear. Res.*, vol. 63, no. 11, pp. 3571–3585, 2020, doi: 10.1044/2020_JSLHR-20-00084.

[4] P. Li, F. Baills, and P. Prieto, "Observing and Producing Durational Hand Gestures Facilitates the Pronunciation of Novel Vowel Length Contrasts," *Stud. Second Lang. Acquis.*, vol. 42, no. 5, pp. 1015–1039, 2020, doi: 10.1017/S0272263120000054.

[5] M. Hoetjes and L. van Maastricht, "Using gesture to facilitate L2 phoneme acquisition: the importance of gesture and phoneme complexity," *Front. Psychol.*, vol. 11, p. 575032, 2020, doi: 10.3389/fpsyg.2020.575032.

[6] Y. Hirata and S. D. Kelly, "Effects of lips and hands on auditory learning of second-language speech sounds," *J. Speech, Lang. Hear. Res.*, vol. 53, no. 2, pp. 298–310, 2010, doi: 10.1044/1092-4388(2009/08-0243).

[7] P. Li, X. Xi, F. Baills, and P. Prieto, "Training non-native aspirated plosives with hand gestures: learners' gesture performance matters," *Lang. Cogn. Neurosci.*, vol. 36, no. 10, pp. 1313–1328, 2021, doi: 10.1080/23273798.2021.1937663.

[8] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *J. Acoust. Soc. Am.*, vol. 130, no. 1, pp. 461–472, 2011, doi: 10.1121/1.3593366.

[9] I. Mora-Plaza, M. Ortega, and J. C. Mora, "High-Variability Phonetic Training Under Different Conditions: Individual Differences in Auditory Attention Control," in *Theoretical and Practical Developments in English Speech Assessment, Research, and Training*, V. G. Sardegna and A. Jarosz, Eds. Springer Nature Switzerland AG, 2022, pp. 241–260.

[10] C. Aliaga-García, J. C. Mora, and E. Cerviño-Povedano, "L2 speech learning in adulthood and phonological short-term memory," *Poznań Stud. Contemp. Linguist.*, vol. 47, no. 1, pp. 1–14, Jan. 2011, doi: 10.2478/psicl-2011-0002.

[11] J. A. Linck, P. Osthus, J. T. Koeth, and M. F. Bunting, "Working memory and second language comprehension and production: A meta-analysis," *Psychon. Bull. Rev.*, vol. 21, no. 4, pp. 861–883, 2014, doi: 10.3758/s13423-013-0565-2.

[12] A. Baddeley, "Working memory: Theories, models, and controversies," *Annu. Rev. Psychol.*, vol. 63, pp. 1–29, 2012, doi: 10.1146/annurev-psych-120710-100422.

[13] M. Aldugom, K. Fenn, and S. W. Cook, "Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity," *Cogn. Res. Princ. Implic.*, vol. 5, p. 27, 2020, doi: 10.1186/s41235-020-00215-8.

[14] J. Cebrian, "Perception of English and Catalan vowels by English and Catalan listeners: A study of reciprocal cross-linguistic similarity," *J. Acoust. Soc. Am.*, vol. 149, no. 4, pp. 2671–2685, 2021, doi: 10.1121/10.0004257.

[15] J. Cebrian, "Experience and the use of non-native duration in L2 vowel categorization," *J. Phon.*, vol. 34, no. 3, pp. 372–387, 2006, doi: 10.1016/j.wocn.2005.08.003.

[16] J. P. Zerling, "Frontal lip shape for French and English vowels," *J. Phon.*, vol. 20, no. 1, pp. 3–14, 1992, doi: 10.1016/s0095-4470(19)30249-9.

[17] S. A. J. Wood, "A radiographic and model study of the tense-lax contrast in vowels," in *Phonologica 1988*, 1992, pp. 283–291.

[18] K. J. Blacker, S. M. Weisberg, N. S. Newcombe, and S. M. Courtney, "Keeping track of where we are: Spatial working memory in navigation," *Vis. cogn.*, vol. 25, no. 7–8, pp. 691–702, 2017, doi: 10.1080/13506285.2017.1322652.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[20] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*. SAGE, 2019.

[21] R. V Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2022.

EFFECTS OF LANGUAGE TRAINING ON MANDARIN LEXICAL TONE PERCEPTION BY JAPANESE SPEAKERS

Wu Qi

University of Tsukuba
wuqi920524@yahoo.co.jp

ABSTRACT

In this study, we investigated the perception of Mandarin tones with different registers and syllable durations among Japanese speakers with different kinds of language training. The results showed that language training type and registers affected tone perception: the longer the language training, the closer the tone perception was to that of native Mandarin speakers. Register and syllable duration also affected the perception of Mandarin tones: the perception was more difficult in the low register, particularly for less experienced learners. Less experienced learners performed better on the shortest syllable duration. Also, for both learners and native speakers of Mandarin, the threshold values of ΔF0 (|onset – offset|) necessary to identify Tone 4 were larger than that of Tone 2.

Keywords: tone, years of language training, register, syllable duration, perception.

1. INTRODUCTION

There are four phonemic tones in Mandarin. Tone 1 (T1 hereafter) has a high-level pitch, Tone 2 (T2 hereafter) has a high-rising pitch, Tone 3 (T3 hereafter) has a low-dipping pitch, and Tone 4 (T4 hereafter) has a high-falling pitch [3]. According to [5] and [6], native Mandarin listeners identify tones by employing multiple acoustic cues. Furthermore, perceptual studies have suggested that F0 and F0 contour are the fundamental perceptual cues of Mandarin tones, e.g. [7].

How Mandarin learners identify tones is also a topic of interest. Previous studies have conducted experiments on native speakers of English, e.g. [4], but only a few studies have been conducted on Japanese Mandarin learners whose native language is Japanese which is non-tonal, e.g. [8] or [9]. In this study, we investigated the effects of three factors (length of language training, register, and syllable duration) on Japanese learners' tone perception and compared them with that of native speakers.

2. METHOD

2.1. Stimuli

Speech materials were developed from Mandarin bā (T1), produced by a single female native speaker. The native speaker pronounced it with three other tones at her comfortable pitch range. It was recorded through Praat [2] with 44.1 kHz, 16-bit digitization. The female voice was chosen as the original stimulus so that stimuli with a wide range of pitch can be created without the loss of naturalness.

Fig. 1 is a schematic representation of synthesized stimuli for this study where the onset and offset for each stimulus were varied systematically. All stimuli were synthesized based on stimuli0 in Fig. 1 whose onset and offset were 280.5Hz. VocalShifter LE [1] was used for the synthesis.

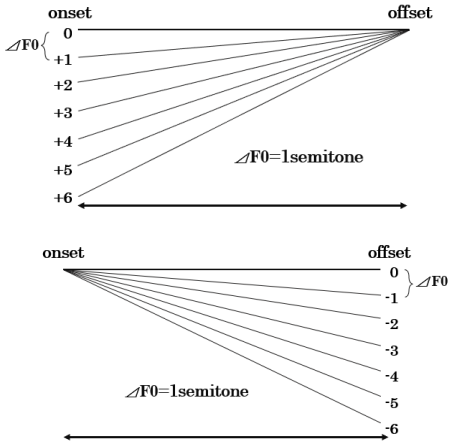


Figure 1: Schematic diagram of synthesized F0 patterns

The onset F0 and stimulus durations were manipulated subsequently to create stimulus sets with three F0 ranges (“register” hereafter), and three durations. F0 range was set to 25-31 semitone in the high register, 12-18 semitone in the medium register, and 5-11 semitone in the low register relative to 100Hz. Three durations, 350ms, 550ms, and 750ms, were used. In each register and syllable duration, thirteen stimuli with stepwise rising or falling pitch contours as seen in Fig.1 were included, resulting in 117 stimuli in total.

2.2. Participants

16 native speakers (NS) and three groups of 26 Japanese speakers participated in this study (Table 1). The first group (G1) learned Mandarin for 2 months to 1 year, the second group (G2) for 1.5 to 3 years, and the third group (G3) for 3.5 to 6 years. Participants learned Mandarin in regular undergraduate classes, and some continued their studies after graduation. We classified the learners based on the length of training rather than proficiency test scores because the test scores were not available for many of the participants. All participants were reported to have normal hearing.

Group	Length of Training	Gender	Age
G1	2 months - 1 year	2M, 12F	20.71
G2	1.5 years - 3 years	2M, 6F	22.38
G3	3.5 years - 6 years	2M, 2F	22.25
NS	Native speaker	6M, 10F	25

Table 1: Participants.

2.3. Procedures

This study used a four-alternative forced-choice task. Each participant listened to a stimulus and selected a tone they perceived, from "bā", "bá", "bǎ" and "bà" written in pinyin (the Romanized spelling system of Chinese characters with tone symbols indicated by diacritics). A calculation task was inserted after each response to avoid biases from the previous stimulus, the stimuli were presented in three blocks by syllable durations (350ms, 550ms, 750ms). The order of stimuli was randomized within each block. Each participant listened to the stimuli at his/her comfortable level over headphones on a computer. The participants were allowed to take a break after each block. All participants completed the experiment in 30 to 45 minutes.

3. RESULTS

Fig. 2 shows the responses in ratio for each stimulus by groups, tallied across syllable durations and registers. Lines connecting symbols represent responses to the four tones in Mandarin: red for T1, green for T2, yellow for T3, and blue for T4. The judgment curves by NS showed sharp boundaries between T1 and T2 as well as T1 and T4. Near 100% responses were observed for T1, T2, and T4 at some points along the stimulus continuum, indicating clear categorical perception. Responses for T3 did not show the clear peak seen in T1, T2, and T4, and the stimuli were rarely identified as T3 by NS. This indicates the stimulus continuum in this study did not include a stimulus that can be perceived clearly as T3 by native Mandarin speakers[1].

The judgment curves by G1 were less steep and the response ratio for T1, T2, and T4 did not reach near-100 percent anywhere along the continuum. It is of interest that G1 had T3 responses rather often compared with other groups. The judgment curves of G3 have distinct areas with high identification rates for T1, T2, and T4, similar to those of NS.

The maximum responses along the continuum (maximum identification rates, hereafter) for T1, T2, and T4 (i.e., the peaks in Fig. 2) increased with years of language training. For T1 and T4, the maximum identification rates were near 100% in G3 and NS but were lower in G1. Notice that for T2, the maximum identification rate occurred at S+2 in G1 but S+6 in G3 and NS.

3.1. Response of T1, T2 and T4

Table 2 shows the maximum identification rates for T1, T2, and T4 by registers and groups. Data were tallied across durations.

For T1, the maximum identification rates of G1 learners were 52% to 57%, which were substantively lower than those of the other groups in all registers.

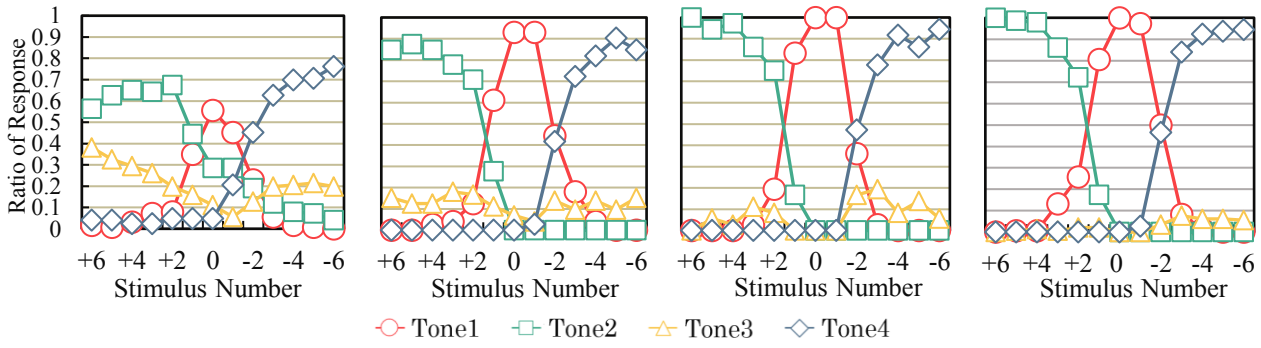


Figure 2: Responses in ratio for each tone and stimulus. From left to right, each figure depicts results of G1, G2, G3 and NS.

The maximum identification rates for G2, G3, and NS were near 100% across registers.

For T2, the maximum identification rates were 100% across registers in G3 and NS. However, G2 showed a lower maximum identification rate at the low register compared to the high and medium registers. In G1, the maximum identification rate declined as the register became lower. All groups showed a decline in maximum identification rates at the low register for T4.

Table 3 summarizes the maximum identification rates of T1, T2, and T4 by groups and by syllable durations. For T1, T2, and T4, the effects of language training were observed in all durations. No apparent effect of syllable duration was observed in any groups for T1. For T2 and T4, the highest maximum identification rates in G1 and G2 were achieved at T=350ms, with substantively lower rates at longer durations.

3.2. Perceptual boundaries between lexical tones

Phonological perceptual boundaries were calculated for each participant in each register and syllable duration. Results from T1 responses were used. G1 data were not included in this analysis because they did not demonstrate sufficient categorical perception.

First, the perceptual judgment curve was approximated by a logistic function (see equation (1) below) for each participant in each register and duration. Second, the locations where the curve crossed the 50% identification threshold were identified. Two locations, representing T1T2 and T1T4 boundaries, were identified for each curve and defined as the phonological perceptual boundary values. The horizontal axis in Fig. 3 shows the phonological perceptual boundary values and the left and right panels show the T1T2 and T1T4 boundaries, respectively. HR, MR, and LR indicate high, medium, and low registers, respectively. Each symbol represents the mean phonological value for the specified group, register, and syllable duration. The error bars depict one standard deviation around the mean for each group, register, and duration.

$$f(x) = \frac{1}{1+e^{ax+b}} \quad (1)$$

The T1T2 boundaries are closer to S0 than the T1T4 boundaries. This means the amount of descent required for the perception of T4 is greater than the rise required for the perception of T2.

For each of the T1T2 and T1T4 boundaries, a three-way ANOVA was conducted with group, register, and syllable duration as explanatory variables and perceptual boundaries as the objective variable.

For T1T4 boundaries, the main effects of group, register, and syllable duration were not significant and there were no significant interactions. For T1T2 boundaries, the main effect of the register ($F(2, 50)=17.854$, $p<.05$) and of syllable duration ($F(2, 50)=3.827$, $p<.05$) were significant. There was a significant interaction between syllable duration and register ($p<.05$). Post-hoc analysis revealed that the

Response / Register	T1				T2				T4			
	G1	G2	G3	NS	G1	G2	G3	NS	G1	G2	G3	NS
High register	57	100	100	100	86	100	100	100	81	100	100	100
Medium register	57	100	100	100	74	100	100	100	76	96	100	100
Low register	52	92	100	100	52	71	100	100	71	75	83	83

Table 2: The maximum identification rates for T1, T2, and T4 by groups and registers.

Response / Duration	T1				T2				T4			
	G1	G2	G3	NS	G1	G2	G3	NS	G1	G2	G3	NS
T=350ms	57	96	100	100	74	96	100	100	81	100	92	96
T=550ms	55	96	100	100	64	83	100	100	76	88	100	96
T=750ms	55	96	100	100	69	88	100	100	76	83	92	94

Table 3: The maximum response rates for T1, T2, and T4 by durations.

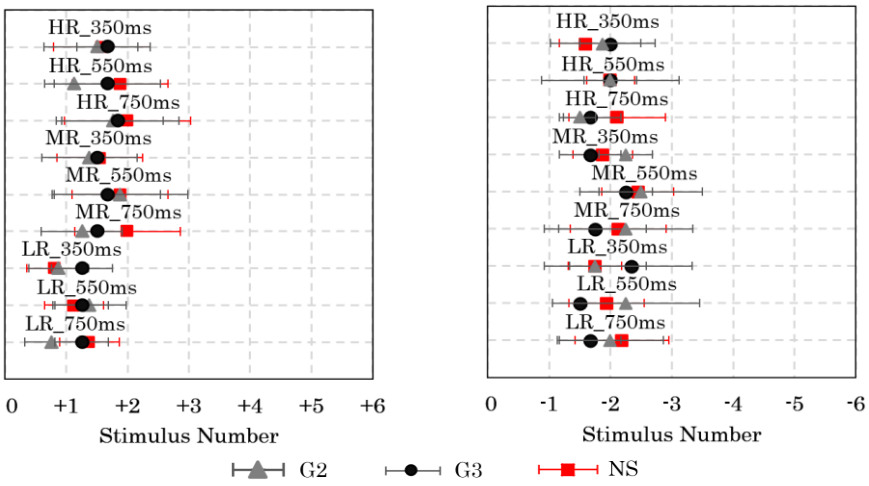


Figure 3: Relationship of perceptual boundary and stimulus number from participants.

differences between T=350ms vs. T=550ms and T=350ms vs. T=750ms were significant (p<.05). T1 T2 boundary was closest to S0 in T=350ms. The differences between the low vs. high and the low vs. medium register were significant (p<.05): the T1T2 boundary was closest to S0 in the low register. In other words, the lower the register, the smaller the rise required for a stimulus to be perceived as T2, and the shorter the syllable duration, the smaller the rise required to be perceived as T2.

4. DISCUSSION

The results of the present study suggest that the longer the language training, the closer the judgment curves were to those of native Mandarin speakers. Learners with language training lasting more than 3.5 years (G3) may have formed a system of categorical perception of Mandarin tones similar to that of native speakers.

Voice register is also likely to have effects on the perception of these tones: learners with a shorter period of language training, such as G2 and G1, did not identify T2 as well as G3 and NS in the low register. Syllable durations also appeared to have effects, suggesting that perception was easier with short syllable duration for listeners with less language training.

Examination of perceptual boundaries revealed that the threshold values of ΔF0 (|onset – offset|) necessary for identifying T4 were larger than that for T2. In other words, listeners were more sensitive to the rising tone than to the falling tone. It was also found that the location of the T1T4 boundary was affected by registers and by syllable durations. Taken together, the results provide quantitative evidence of the acquisition of tone perception over the course of language learning among Japanese speakers.

Although stimuli that can be perceived as typical T3 (a low-dipping tone) were not present in this study, some participants, especially G1 listeners, perceived T3 occasionally. The criteria for judging T3 is an issue for future research.

5. ACKNOWLEDGEMENTS

I am grateful to Prof. Zhu Chunyue and Yasunori Takahashi for their insightful comments. Additionally, I extend my gratitude to the anonymous reviewers for their detailed and valuable comments. This research was supported by JSPS KAKENHI Grant Number JP22K13161.

6. REFERENCES

[1] AckieSound. 2016. VocalShifter LE (Version 3.1) [Software].

[2] Boersma, P., & Weenink, D. 2018. Praat (Version 6.0.42) [Software].

[3] Chao, Y. R. 1948. Mandarin primer. Cambridge, MA: Harvard University Press.

[4] Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. 2015. Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. Journal of Second Language Pronunciation, 1(1), 86–114.

[5] Gandour, J. 1984. Tone dissimilarity judgments by Chinese listeners. Journal of Chinese Linguistics, 12, 235–261.

[6] Massaro, D. W., Cohen, M. M., & Tseng, C. 1985. The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. Journal of Chinese Linguistics, 13, 267–290.

[7] Wayland, R. P., & Guion, S. G. 2004. Training English and Chinese listeners to perceive Thai tones: A preliminary report. Language Learning, 54(4), 681–712.

[8] Zhang Linjun. 2010. Riben liuxuesheng hanyu shengdiao de fanchouhua zhijue. *Yuyan Jiaoxue yu Yanjiu* (in Chinese.). 3, 9–15.

[9] Zhu Hong. 2013. Nihonjin gakushusha ni yoru Chugokugo seichou no shuutoku no kenkyu (in Japanese.). Ph.D. Dissertation. Tohoku University.

[1] This may be because the stimuli in this experiment, as seen in Fig. 1, did not contain the low-dipping pitch pattern that is a key feature of T3.

PERCEPTION AND PRODUCTION OF ENGLISH TENSE AND LAX VOWELS BY CHINESE EFL LEARNERS

Lijuan Wang, Jiwei Guo, and Hui Feng

School of Foreign Languages, Tianjin University
ewlj316393@163.com, 13354301576@163.com, fenghui@tju.edu.cn

ABSTRACT

Dynamic spectral patterns are important in the English tense-lax contrasts for native speakers; however, researches on EFL learners focus more on duration and static spectral features. The paper, based on the Revised Speech Learning Model (SLM-r), investigated the acquisition of /i-ɪ/, /u-ʊ/, /ɑ-ʌ/, /æ-ɛ/ by 114 EFL learners and examined L2 perception-production relationship. Findings suggest: (1) Participants weighted duration more heavily than spectrum for /i-ɪ/ and /ɑ-ʌ/, and gave equal values for /u-ʊ/ and /æ-ɛ/. (2) Dynamically, Chinese EFL learners used similar vowel inherent spectral change patterns, especially the similar directions of formant movement, in realizing tense-lax vowel contrasts, which shortened the spectral distance in the vowel space. (3) The durations of /i-ɪ/ and /ɑ-ʌ/ showed consistency in perception and production, unlike /u-ʊ/ and /æ-ɛ/. Such findings supported SLM-r's predictions relating to the perceived cross-language dissimilarity and cue weighting, but did not fully corroborate the hypothesis of perception-production coevolution.

Keywords: Tense-lax vowel contrast; The Revised Speech Learning Model; Chinese EFL learners

1. INTRODUCTION

Chinese consists of five basic vowel phonemes /ɑ, e, i, u, y/[1][1]. English has a larger vowel inventory than Chinese, with /i, ɑ, æ, u, ɔ, eɪ, oʊ/[2] as tense vowels and /ɪ, ʌ, ɛ, ʊ/ as lax vowels [2-4]. English tense vowels differ from lax vowels in duration and vowel quality. When producing a tense vowel, the strengthened subglottal air pressure needs longer time [5]. Bohn and Flege [6] found that English speakers relied more on formant frequency than duration when discriminating between tense and lax vowels. Nearey and Assmann [7] found that trajectory of tense vowels moved towards the periphery of the formant space while that of lax vowels moved towards the centre.

Researchers have also explored cross-linguistic acquisition of tense and lax vowels. Flege, Bohn, and Jang [8] found that native-like spectral distance could be produced for certain pairs such as /æ-ɛ/ by Chinese EFL learners. Some research has indicated a close link between perception and production. Training on the perception of English /r-l/ contrast positively transferred to production improvement among Japanese English learners [9].

So far, some areas still need more research. First, the weight of duration and spectrum in tense-lax distinction for non-native speakers is still under controversy. Second, few studies have paid attention to the pattern of dynamic change in vowels. Third, for the correlation between perception and production, the detailed acoustic cues employed by L2 learners have not received sufficient attention.

Speech Learning Model revised (SLM-r), the improved version of SLM [10], argues that formation of a new L2 category depends on the cross-language dissimilarity, the nativelike use of cues, and cue weights, and that segmented production and perception evolve [11].

Therefore, this study will explore how Chinese EFL learners perceive and produce English tense-lax contrasts, and how perception and production are related.

2. METHODOLOGY

2.1. Participants

A total of 114 subjects participated in the experiments, including 40 English learners (EFL group) and 74 native English speakers (NE group). For the NE group, 40 speakers (gender-balanced) in the production experiment are from the database of Hillenbrand et al [12], and the other 34 speakers participated in the perceptual experiment.

2.2. Stimuli

The perceptual assimilation task is used to assess the perceived dissimilarity between English and Chinese vowels. Eight English stimuli were produced by a male native English speaker and three closest Chinese vowels were produced by a male Chinese speaker.

For the identification task, which is aimed at investigating the acoustic cues used by Chinese EFL learners when discriminating between tense and lax vowels, with each continuum (/ɪ-i/, /ʊ-u/, /ɛ-æ/, /ʌ-a/) contained 25 tokens varying along 5 spectral steps (SP1-SP5) and 5 durational steps (D1-D5). A total of

100 tokens (4 continua*25 stimuli) were interpolated in the /hVd/ context.

For the production experiment, which is designed to investigate how Chinese EFL learners realize tense-lax contrasts from duration, formant frequency, and tongue body movement, the material contains four pairs of English vowels /ɪ/-/i/, /ʊ/-/u/, /ɛ/-/æ/ and /ʌ/-/ɑ/ in the /hVd/ context, and three Chinese vowels /i, u, ɑ/ in the /hVdə/ context. A total of 2,280 tokens were produced containing 1,920 tokens of English vowels (80 participants*8 vowels*3 repetitions) and 360 tokens of Chinese vowels (40 participants*3 vowels*3 repetitions).

2.3. Procedures

The perception experiment and the production experiment are conducted separately. The perception experiment, which includes the perceptual assimilation task and the identification task, aims to investigate how EFL listeners discriminate between tense and lax vowels in terms of duration and spectrum, and then to determine cue weighting.

Firstly, the 40 EFL learners were told to hear random English vowels /ɪ, i, ʊ, u, ɛ, æ, ʌ, ɑ/ and choose the Chinese response closest to the English stimuli from a seven-alternative forced-choice task, including six Chinese vowels /ɑ, o, e, i, u, y/ and a "none" option for detecting EFL group's perception of "new" non-native sounds. Then, participants judged the degree of dissimilarity between the sound heard and the option selected from a seven-point scale (1 to 7 corresponded to *the least similar* to *the most similar*). Next, 34 English listeners were asked to hear three Chinese stimuli /i, u, ɑ/ and identify the English stimulus they heard from an eight-alternative forced-choice task, including eight English vowels /ɪ, i, ʊ, u, ɛ, æ, ʌ, ɑ/. They then judged the dissimilarity between the stimulus with the same procedure as EFL group.

Secondly, the 40 EFL learners and 34 English listeners were told to hear four continua and choose the response that best matched the word they heard from a two-alternative forced-choice task. The two options were "heed" and "hid" for /i-ɪ/ continuum, "head" and "had" for /ɛ-æ/ continuum, "hood" and "who'd" for /ʊ-u/ continuum, "Hudd" and "hod" for /ʌ-ɑ/ continuum.

Finally, the 40 EFL learners read the materials three times at random, including 20 English words (8 vowels and 12 fillers) and 3 Chinese words. The recordings of 40 native American English speakers were obtained from the Hillenbrand database[3].

2.4. Acoustic measurements

The cue weighting according to Morrison [13] was calculated by Equation (1),

(1) Cue weighting = $\beta_{spectrum} / (\beta_{duration} + \beta_{spectrum})$

where $\beta_{duration}$ and $\beta_{spectrum}$ refer to the dependence on duration and spectrum, respectively.

To describe the acoustic differences between tense and lax vowels, the durational ratio (DR) between tense-lax contrast was calculated by Equation (2).

(2) $DR = Duration_{tense} / Duration_{lax}$

3. RESULTS AND DISCUSSION

3.1. Results of the production experiment

3.1.1. The role of duration

To describe the temporal contrast between tense and lax vowels, the mean duration ratio (DR) of the tense to lax vowels for each pair was plotted in Fig. 1. A t-test revealed significant differences of DR between EFL and NE group for all tense-lax pairs ($t(78)=-8.38$, $p < 0.05$ for /i-ɪ/; $t(78)=-7.69$, $p < 0.05$ for /u-ʊ/; $t(78)=3.21$ $p < 0.05$ for /æ-ɛ/; $t(78)=-3.42$, $p < 0.05$ for /ɑ-ʌ/). Such a result indicates that EFL learners overemphasized the durational differences between the tense-lax contrast for each pair except /ɛ-æ/. The result replicated previous findings that native Chinese and native Korean learners of English failed to produce sufficient temporal differences between /ɛ/ and /æ/ compared to native English speakers (cf. [8, 14-15]. The excessive use of durational cues in tense-lax production for the EFL group supported the *full access hypothesis* of SLM-r that L2 speakers employ non-native features to develop L2 categories [11].

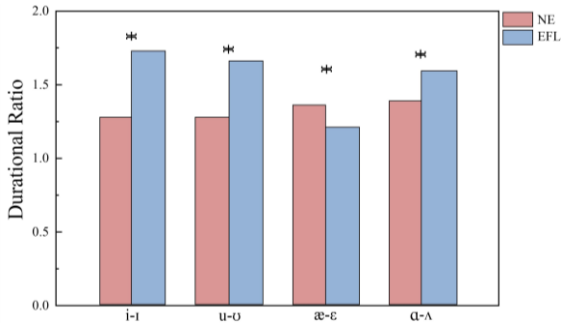


Figure 1: DR for each contrast by NE and EFL groups.

3.1.2. Dynamic spectral patterns

A bidimensional phonetic space of English vowels produced by the EFL and NE groups was illustrated in Fig. 2. For the spectral distance, the tense-lax contrast produced by the NE group (the red line) showed larger spectral differences than the EFL group (the blue line). The solid line connected tense vowels and the dotted line connected lax vowels. A multivariate analysis of variance revealed significant

differences in each pair between the EFL and NE groups ($F(2,77)=16.80$, $p < 0.01$ for /ɑ/; $F(2,77)=66.73$, $p < 0.01$ for /ʌ/; $F(2, 77)=26.65$, $p < 0.01$ for /æ/; $F(2,77)=70.47$, $p < 0.01$ for /ɛ/; $F(2,77)=35.09$, $p < 0.01$ for /i/; $F(2, 77)=170.42$, $p < 0.01$ for /ɪ/; $F(2, 77)=17.38$, $p < 0.01$ for /u/; $F(2,77)=187.56$, $p < 0.01$ for /ʊ/), indicating a deviation from the norm for all the tense and lax vowels produced by Chinese EFL learners.

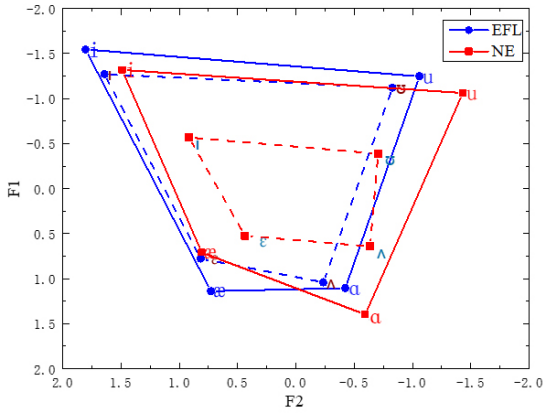


Figure 2: F1-by-F2 plot produced by NE and EFL groups.

Fig. 3 depicted the dynamic formant trajectories in the F1-F2 space (Barked and Z-normalized) of each tense-lax contrast and the closest Chinese vowel. The arrow represents the moving direction from 20% to 80% point on the vowel portion. For /i-ɪ/, NE group's trajectory length of formant movement was larger than EFL group, suggesting a small range of tongue body movement for EFL learners. Also, EFL learners' trajectory of /i/ and /ɪ/ moved downwards, indicating a lowering tongue body.

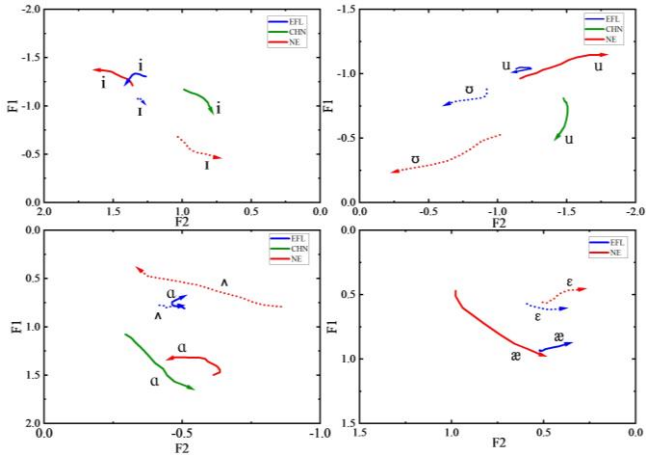


Figure 3: Dynamic formant trajectories of English vowels and the closest Chinese vowels.

Similarly, the pair /u-ʊ/ by NE group showed a larger trajectory length than EFL group. For EFL learners, both vowels moved downwards to the left, indicating a falling and advancing tongue body. The similar directions of /u-ʊ/ movement were probably influenced by L1, in that their trajectory directions approximated the direction of the closest Chinese vowel /u/.

In the case of /ɑ-ʌ/, the trajectory directions of the NE and EFL groups were opposite on the horizontal axis, with the former moving to the left and the latter moving to the right. This indicates NE group realized /ʌ-ɑ/ with an advanced tongue body while EFL group realized /ʌ-ɑ/ with a retracted tongue body. The Chinese /ɑ/ closest to /ʌ-ɑ/ contrast also moved to the right on the horizontal axis, indicating L1 influence on the L2 articulation.

For the pair /ɛ-æ/, EFL trajectories moved in a converging direction while NE trajectories moved in a diverging direction, suggesting that NE group strengthened the spectral contrast. Specifically, native English speakers realized /ɛ/ with a retracting and rising tongue body while EFL learners produced /ɛ/ with a retracting and falling tongue body.

3.2. Results of the perception experiment

3.2.1. Perceived dissimilarity between Chinese and English vowels

In Table 1, English tense vowels /i, u, ɑ/ were consistently identified as Chinese /i, u, ɑ/ respectively at a high frequency (95% - 97.5%) and with high goodness fitting scores above 5 out of 7. A reverse perceptual assimilation experiment performed by native English speakers confirmed that Chinese /i, u, ɑ/ were identified as English /i, u, ɑ/ respectively.

Similarly, the English lax vowels /ɪ, ʊ, ʌ/ were assimilated to the Chinese responses /i, u, ɑ/ respectively (82.5% - 92.5%) but with lower scores ranging from 3.4 to 4.4. English /ɛ/ was identified as "new" by EFL learners (87.5%). Nearly two-thirds of listeners identified /æ/ as an instance of Chinese /ɑ/ with the goodness fitting score of 4.1 out of 7.

		English target vowels							
		/i/	/ɪ/	/u/	/ʊ/	/ɑ/	/ʌ/	/æ/	/ɛ/
Chinese responses	/ɑ/				2.5 (2.0)	**97.5 (5.2)**	**90.0 (4.3)**	**60.0 (4.1)**	2.5 (2.0)
	/o/			2.5 (6.0)	5.0 (4.0)				
	/e/		15.0 (3.0)			2.5 (3.0)	7.5 (3.7)	5.0 (2.0)	10.0 (2.3)
	/i/	**97.5 (5.6)**	**82.5 (3.4)**						
	/u/	2.5 (2.0)		**95.0 (5.3)**	**92.5 (4.4)**		2.5 (2.0)		
	/ü/		2.5 (2.0)	2.5 (3.0)					
	none							35.0	**87.5**

Table 1: Mean percentage assimilation and mean goodness rating scores (in parenthesis) of English target stimuli to Chinese vowels (EFL group).

3.2.2. Vowel identification of synthetic continua

A binary logistic regression analysis was conducted to quantify the dependence on duration or spectrum for participants. A cue weighting value above 0.5 represents more use of spectrum than duration. An independent-samples t-test was conducted to examine the significant differences between the cue weighting value and 0.5.

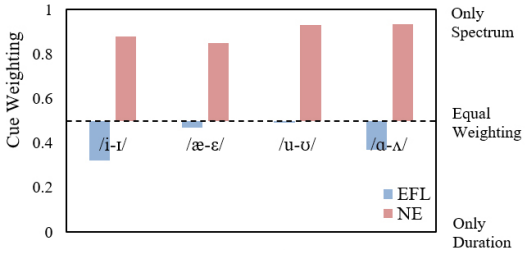


Figure 4: Mean cue weighting for tense-lax contrasts.

Results of mean cue weighting values of the EFL and NE groups were shown in Fig. 4. Cue weighting of all the four pairs was significantly larger than 0.5 for the NE group, which indicates the NE group weighted the spectrum more strongly than duration for each pair. For the EFL group, the mean cue weighting value of /i-ɪ/ and /ɑ-ʌ/ was significantly smaller than 0.5 ($t(78)=-3.89$, $p < 0.05$; $t(78)=-3.013$, $p < 0.05$). No significant differences between the cue weighting value and 0.5 were found for /u-ʊ/ and /æ-ɛ/. Such results suggest that the EFL group gave more values to duration than to spectrum when distinguishing /i-ɪ/ from /ɑ-ʌ/, but weighted duration and spectrum equally for /u-ʊ/ and /æ-ɛ/.

3.3. Relationship between perception and production

A significant relationship between perception and production was found across the four pairs on the durational dimension rather than the spectral dimension, which is consistent with previous study [16]. Fig. 5 (left panel) provided the mean cue weighting for each pair in the perception and the mean durational radio (DR) in the production. A one-way ANOVA revealed a main effect of vowel pair [$F(3, 36)=30.80$, $p < 0.05$]. Except for /u-ʊ/, the other three pairs, namely /i-ɪ/ /ɑ-ʌ/ and /ɛ-æ/ showed the pattern that the more strongly temporal cue was used for one contrast during the perception, the more durational differences were produced in the same contrast during production.

Fig. 5 (right panel) depicted the cue of spectrum in the perception and production. A one-way ANOVA revealed no significance of vowel pair [$F(3, 36) = 0.17$, $p > 0.05$]. This indicates that changes in the weight of spectral cues in perception do not affect the amount of spectral difference in production across the four pairs. Such findings reflect that the relationship between perception and production was of great complexity and the result varied according to different L2 contrasts and different assessment parameters.

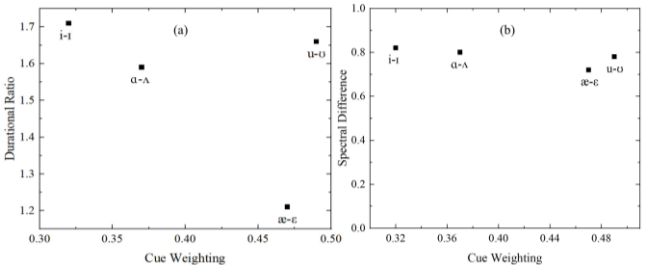


Figure 5: Mean cue weighting value for each contrast by the EFL group.

4. CONCLUSION

The paper investigated the acquisition of /i-ɪ/, /u-ʊ/, /ɑ-ʌ/, /æ-ɛ/ by 114 Chinese EFL learners and especially examined the dynamic spectral patterns of EFL learners and native English speakers.

For the cue weighting in perception, EFL learners weighted duration more heavily than spectrum for /i-ɪ/ and /ɑ-ʌ/, but gave equal values to duration and spectrum for /u-ʊ/ and /æ-ɛ/.

In the dynamic learning process, EFL learners exaggerated durational differences and weakened spectral differences in /i-ɪ/, /u-ʊ/, and /ɑ-ʌ/ (except /æ-ɛ/), whereas native speakers presented an opposite duration-vowel quality relationship with stronger use of spectral differences. What's more, EFL learners applied their L1 spectral change patterns of vowels, especially the formant movement directions, to the similar vowels (/i-ɪ/, /u-ʊ/, /ɑ-ʌ/) they perceived in L2, and realized tense-lax vowel contrasts by shortening the spectral distance in the vowel space.

The research partially supported the prediction of SLM-r relating to the L2 perception-production relationship. For acoustic cues, the predominant cues for perception were not always consistent with that for production within each pair. The contrast more dependent on duration in perception showed more durational differences in production, but this correlation was not found for the cue of spectrum.

In conclusion, the findings will be helpful for Chinese EFL learners for native-like perception and production, such as emphasizing the differences between the L1 and L2 vowel system, highlighting the native-like acoustic cues, and strengthening the articulatory process. For future research, different types of formant movement could be considered for the synthesized stimuli to investigate the role of formant movement in the perception of English tense and lax vowels.

5. ACKNOWLEDGEMENTS

This research is partly supported by National Major Social Sciences Foundation of China under Grant 15ZDB103, the Innovation Program of Chinese Academy of Social Sciences, the Social Science Foundation of Tianjin, China (No. TJWW19-009), and the National Key R&D Program of China under Grant 2018YFB1305200. We thank the anonymous reviewers for their suggestions.

6. REFERENCES

[1] Cheng, R. L. 1966. Mandarin phonological structure. *Journal of Linguistics*, 2, 135—158.

[2] Ladefoged, P., & Johnson, K. 2015. *A Course in Phonetics (7th Edition)*. Cengage Learning.

[3] Strange, W., Bohn, O. S., Trent, S. A., & Nishi, K. 2004. Acoustic and perceptual similarity of north German and American English vowels. *J. Acoust. Soc. Am.* 115(4), 1791–1807.

[4] Wang, H.-Y., & Van Heuven, V. J. 2006. Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. *Linguistics in the Netherlands.* Advance online publication.

[5] Jakobson, R., Font, C. G., & Halle, M. 1952. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. The MIT Press.

[6] Bohn, O. C., & Flege, J. E. 1990. Interlingual identification and the role of foreign language experience in L2 vowel perception. A*pplied Psycholinguistics*, 11(3), 303-328.

[7] Nearey, T. M., & Assmann, P. F. 1986. Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am*. 80, 1297–1308.

[8] Flege, J. E, Bohn, O. S., Jang, S.-Y. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4), 437-470.

[9] Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., Tohkura, Y. 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *J. Acoust. Soc. Am*. 101(4), 2299–2310.

[10] Flege, J. E. 1995. Second-language speech learning: Theory, findings, and problems. In: W. Strange (eds.), *Speech Perception and Linguistic Experience: Issue in Cross-Language Research*. Timonium, MD: York Press, 229–273.

[11] Flege, J. E., Aoyama, K., Bohn, O. S. 2021. The Revised Speech Learning Model (SLM-r) Applied. In: R. Watland (eds.), *Second Language Speech Learning Theoretical and Empirical Progress*. Cambridge University Press, 84-119.

[12] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am*. 97(5), 3099-3111.

[13] Morrison, G. S. 2007. Logistic regression modelling for first- and second-language perception data. In: M. J. Solé, P. Prieto, J. Mascaró (eds.), *Segmental and Prosodic Issues in Romance Phonology*. Amsterdam: John Benjamins, pp. 219–236.

[14] Chen, Y. 2006. Production of Tense-Lax Contrast by Mandarin Speakers of English. *Folia Phoniatrica et Logopaedica*, 58(4), 240-249.

[15] Kim, D., Clayards, M., Goad, H. 2018. A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics*, 67, 1-20.

[16] Qin Y. & Archibald, J. 2022. Modified Input Training and Cue Reweighting in Second Language Vowel Perception. *Frontiers in Educational Research*, 5(6), 65-75.

[1] The number of monophthongs in Chinese varies according to different analyses, ranging from 3 to 12, with 5 or 6 being the most common without taking allophones into account [2].

[2] The classification of /æ/ is in dispute. The study followed the criterion of Strange et al and categorized /æ/ in the tense set [3]. This is because the duration of /æ/ is longer and its location is more peripheral in the vowel space compared with /ɛ/ despite the fact that it occurs only in closed syllables [4].

[3] Available at http://homepages.wmich.edu/~hillenbr/voweldata.html.

INTELLIGIBILITY AND RELATED FEATURES OF ENGLISH SOUNDS PRODUCED BY YOUNG JAPANESE LEARNERS

Akiyo JOTO, Yuri NISHIO

Prefectural University of Hiroshima, Meijo University
joto@pu-hiroshima.ac.jp, ynishio@meijo-u.ac.jp

ABSTRACT

This study investigated the intelligibility of young Japanese learners' pronunciations of English alphabet letter names and clarified acoustic features related to the varying intelligibility levels of their productions. Eight native English speakers of four different nationalities evaluated productions made by 10 Japanese third-graders. We performed acoustic analysis to elucidate the features of their productions as well as those of native English speakers and Japanese sounds for comparison.

Young learners' pronunciations of L, P, T, V, and Z were less intelligible. We attributed lower intelligibility to the shorter voice onset time (VOT) of P and T, the shorter duration of the fricative consonants V and Z, and the different formant patterns of L compared to the productions of native English speakers, along with other features affected by Japanese sounds. We offer some pedagogic suggestions based on the results to prevent the stabilization of English pronunciation with a Japanese accent.

Keywords: intelligibility, young Japanese learners, English pronunciation, English alphabet letter names

1. INTRODUCTION

A new system of early English education was started in public elementary schools in 2020 after a 2-year transitional period in Japan, under which the alphabet, an important part of learning English, is first taught to children in third grade: they learn to pronounce the name of, and identify, each letter. Teaching and learning the English alphabet is crucial not only orthographically but also phonetically, especially in early English education in Japan, because many of the sounds in the English alphabet letter names, which account for about half of English phonemes, do not exist in the Japanese sound system; it is presumed that young Japanese learners (JLs) will find it difficult to produce them. To effectively teach young learners how to pronounce each alphabet letter name, the features of JLs' English pronunciations need to be studied. However, compared with a large number of studies on teaching and learning the orthography of the English alphabet in early English education in Japan, there are few objective studies of the features of pronunciations produced by JLs [1], [2], and very few focus on the productions of alphabet letter names [3].

To better understand the English pronunciation of young JLs, particularly aspects that have yet to be fully explored, we conducted a study to assess the intelligibility of the English alphabet letter names spoken by elementary school students in Japan (investigation 1), and attempted to determine acoustic features of their productions related to the intelligibility levels in comparison with the productions of native English speakers and Japanese sounds (investigation 2). The findings provide suggestions to improve the teaching of English pronunciation, in which many elementary school teachers lack confidence.

2. INVESTIGATION 1

Native speakers of English evaluated the intelligibility levels of the pronunciations of English alphabet letter names produced by JLs.

2.1. Materials

The stimulus corpus for the evaluation comprised 260 English productions; 10 Japanese third-graders individually pronounced each alphabet letter name. Each production was sampled at 44.1 kHz, randomly ordered, and filed on a computer.

2.1.1. Speakers

The speakers were 10 Japanese third-graders at a public elementary school (5 boys and 5 girls) who had no experience living outside Japan. They were asked to pronounce each English alphabet letter name in order after receiving four classroom lessons on the alphabet taught by a Japanese teacher.

2.1.2. Evaluators

Eight native English speakers of four different nationalities participated in the experiment: two Americans, two British, two Canadians, and two Australians, chosen to reflect the present situation in Japanese public elementary schools where students

have the opportunity to learn English from assistant language teachers (ALT) of these nationalities.

All evaluators teach English in Japan at the preschool, elementary school, or university level. The length of time they have lived in Japan ranges from 16 to 37 years; all of them can identify English with a Japanese accent because of their daily experiences speaking with Japanese people in English. None of the evaluators had received phonetic training.

2.2. Procedure

Each evaluator listened to 260 productions by JLs on a computer and rated the intelligibility level of the alphabet letter names on a 5-point Likert scale, from 1 *very poor* to 5 *excellent*. They were provided with the alphabet letters whose names the students intended to pronounce when rating each JL production. They could listen to each sound as many times as they wanted and rated the productions at their own pace. They could also write comments about each pronunciation if they desired.

After giving a rating, they listened to the same productions by the JLs just one or two times each and typed the letter whose name they heard in an Excel spreadsheet. If they were unable to identify a production as any letter name, they typed a question mark. They could write comments about each production if they wished.

2.3. Results

We averaged the rating scores assigned by the eight different evaluators to each production. Table 1 shows the mean rating scores and identification rates for each letter. The inter-rater reliability was high according to Cronbach's alpha: α=0.919 for the intelligibility rating and α=0.990 for letter identification. A strong positive correlation was observed between the rating scores and the identification rates (r=0.878); productions with higher intelligibility tended to be more identifiable.

Table 1 indicates that the overall intelligibility level of the productions by the JLs was "fair" (3.74) but that the levels differed across letters: E was the most intelligible (4.66), whereas V, Z, P, and L were less intelligible; for these letters, the rating scores were below 3 (*fair*), with the lowest being L at 2.14.

As outlined in Table 1, the productions for E, H, U, and X were highly intelligible and identified by all evaluators. On the other hand, V, Z, P, L, and T were less identifiable. The intelligibility of T was "fair" (3.01); its identification rate was the third lowest of all (57.5%) and lower than L (58.8%).

To investigate acoustic features related to the various intelligibility levels, we focused on the JLs' productions of V, Z, P, T and L whose intelligibility levels were below 3.0 and/or identification rates were below the rate for V (65.0%).

Table 1: The rating scores of intelligibility levels and the identification rates for the pronunciations of English alphabet letter names produced by the young Japanese learners.

Letter name	Mean of rating scores	Identification rate (%)	Letter name	Mean of rating scores	Identification rate (%)
E	4.66	100.0	Q	3.86	95.0
X	4.51	100.0	A	3.79	93.8
U	4.49	100.0	F	3.69	92.5
I	4.38	97.5	K	3.54	87.5
D	4.36	98.8	M	3.54	78.8
G	4.35	97.5	W	3.54	97.5
H	4.19	100.0	N	3.45	85.0
Y	4.18	98.8	T	3.01	57.5
J	4.14	98.8	R	3.00	80.0
S	4.11	96.3	V	2.85	65.0
B	4.08	80.0	Z	2.73	51.3
O	4.08	97.5	P	2.70	42.5
C	4.01	97.5	L	2.14	58.8
		Mean		3.74	86.4

3. INVESTIGATION 2

We performed acoustic analyses of the JLs' and NSs' productions of V, Z, P, T, and L to determine the features of individual JLs' productions with varying levels of intelligibility. To compare the JLs' and NSs' productions, we also acoustically analyzed the Japanese equivalents of the English letter names.

3.1. Materials

The JLs' productions were the same as in investigation 1. The English materials were English letter names produced by eight native NSs: two male Americans (General American), two female British (one from the Midlands and the other from the middle north), two female Canadians, and two Australians (General Australian), one male and one female. They were the same speakers as the evaluators for investigation 1 except for one Canadian and one American speaker.

The Japanese sound materials were produced by 8 Japanese third-graders (5 males and 3 females) of the 10 participants in investigation 1. They were asked to read Japanese words written in Katakana characters, specifically the Japanese equivalents of English alphabet letter names transcribed into Katakana in the

way they are used and pronounced in daily life in Japan. Each production was sampled at 44.1 kHz and filed on a computer.

3.2. Procedure

We conducted acoustic measurements for all the productions of JLs, NSs, and Japanese sound materials using Praat [4]. We measured the VOT and intensity of stop consonants for P and T, and the duration of the fricatives /v/ and /z/ (V and Z). As for L, we measured three formants of /l/.

3.3. Results

Table 2 shows the mean values of the VOTs of /p/ and /t/ for P and T produced by the 10 JLs and 8 NSs, as well as Japanese stops. We observed a temporary rapid rise of intensity during the VOT of /p/ and /t/ for almost all NSs' productions. We call this the "intensity boost" (henceforth, IB). The temporal point of the IB peak is also shown in Table 2. An example of the IB, due to strong air pressure of aspirated voiceless stop consonants, is indicated by the arrow in Figure 1. Table 2 presents the alphabet letters identified by the evaluators for the JLs' productions of P and T: the left side shows the most frequently identified letters, while the right side displays the second most identified letters. The numbers in brackets indicate the percentage of identification.

Table 2: VOT and the temporal point of the IB of stops for P and T produced by JLs and NSs, as well as Japanese stops, and letters identified for P and T.

Stops	Productions (number of JLs)	VOT (ms)	Temporal point of IB (ms)	Identified letter and identification rate (%)
/p/	NS	65.7	24.0	–
	JL (HI) (4)	37.2	–	P (87.5), B (6.3)
	JL (LI) (6)	15.4	–	B (68.8), P (12.5)
	Japanese	21.5	6.3	–
/t/	NS	81.2	28.1	–
	JL (HI) (5)	40.5	43.3	T (92.5), D (7.5)
	JL (LI) (5)	22.3	–	D (70.0), T (22.5)
	Japanese	36.5	12.0	–

For both stops, a positive correlation was found between the intelligibility rating scores and the VOT: $r = .509$ for P and $r = .819$ for T. The VOT was shorter for the JLs' productions than for the NSs' ones, even for the productions with high intelligibility (HI) rated as 3.0 or higher, but much shorter for productions with low intelligibility (LI) rated lower than 3.0 and the Japanese equivalents. In addition, the IB did not appear in the JLs' stops, but did occur in all NSs' productions except those of two British speakers for P and one British speaker for T. The weak IB appeared in only one JL's production with HI rated as 5 for T and only one Japanese equivalent of each stop consonant. The IB occurred within an average of 30 ms from the release of the stop consonants /p/ and /t/ for most of the NSs' productions.

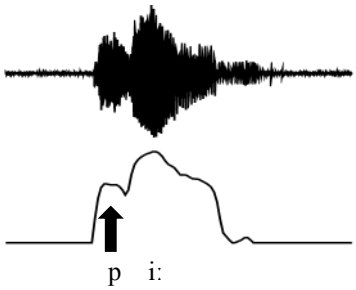


Figure 1: The intensity boost for /p/ produced by a native speaker of American English.

Table 3 displays the mean values of the duration of the fricative consonants and the identified letters for V and Z as pronounced by the JLs and the NSs, as well as the Japanese equivalents. The Japanese equivalent of V is /bɯi/ and that of Z is /ʣi:/ or /ʣetto/. Since there is no /v/ sound in Japanese, the duration of the Japanese equivalent of V was not measured.

Table 3: The duration of the fricatives and identified letters for the V and Z productions of the JLs and NSs, as well as Japanese fricatives.

Fricatives	Productions (number of JLs)	Duration (ms)	Identified letter and identification rate (%)
/v/	NS	139.1	–
	JL (HI) (4)	127.7	V (100)
	JL (LI) (6)	44.1	V (41.7), B (35.4)
/z/	NS	158.1	–
	JL (HI) (5)	97.4	Z (77.5), G (12.5)
	JL (LI) (5)	–	? (27.5), Z/G (25.0)
	Japanese	41.3	–

All the JLs' productions with high intelligibility of V had a /v/ sound; the mean duration was almost as long as that of the NSs' fricative and longer than one JL's production with low intelligibility. The other five with low intelligibility productions had /b/ for /v/.

The acoustic analysis for Z showed a release-burst of /ʣ/ or /ʤ/ in all JLs' productions and all Japanese equivalents except for one JL production with high intelligibility and one Japanese sound, both without a release-burst.

The correlation coefficient was measured between the intelligibility rating scores and the durations of JLs' /v/ and /z/. JLs' pronunciations of V and Z with higher intelligibility were correlated with longer duration of /v/ and /z/, respectively: $r = 0.96$ for V and $r = 0.766$ for Z.

Table 4 presents the mean values of F1, F2, and F3 of /l/ for L produced by the JLs and NSs. The Japanese equivalent of L is /ɛɾɯ/: instead of the approximant /l/, the tap or flap sound is used in Japanese, in which formants cannot be measured. We could measure three formants for L in three male productions.

Table 4: The mean frequencies (Hz) of three formants of /l/ in the JLs' and NSs' productions and the Japanese equivalent, and letters identified for L.

Productions (number of JLs)	F1	F2	F3	Identified letter and identification rate (%)
NS	594.6	1072.5	3022.0	–
JL (LI) (10)	507.3	1886.4	2515.9	L (58.8), ? (18.8)

All JLs' pronunciations were of low intelligibility rated below 3.0. F2 is lower in English dark /l/ than light /l/ due to velarization [5], [6]. Notably, regardless of the speaker's age or gender, the JLs' productions have higher F2 and lower F3 than NSs'. The continuous descent of F2 from the vowel observed in NSs' dark /l/ [5] was not present in all three JLs' /l/ sounds. Due to the limited number of JLs' productions available for statistical analysis, we did not measure the correlation coefficient between intelligibility and formant patterns.

There was some variation in the JLs' pronunciations. Seven of the ten students added the Japanese vowel /ɯ/ at the end, and half of them pronounced /ɾ/ for /l/. We also observed the slightly retroflexed lateral /l̻/, the weak intervocalic /ɹ/, and the weak light /l/; for all of these sounds, a higher F2 or descending F3 is characteristic [6], which affected the acoustic features of JLs' productions for L.

4. DISCUSSION

As presented in Table 1, there were differences in intelligibility across the 10 JLs' productions for each of the five letter names. Most evaluators heard low-intelligibility pronunciations of P or T as B or D, respectively, attributable to the shorter VOT of the voiceless consonants /p/ and /t/. Table 2 shows that the VOTs of JLs' low intelligible productions are less than 25 ms, which suggests English voiced plosives [7]. The VOTs of the Japanese voiceless stops are also shorter than the English ones. This means that negative transfer of Japanese sounds affected the JLs' pronunciations of P and T.

Negative language transfer also occurred in the JLs' pronunciations of V and Z: /b/ was used instead of /v/ and /dʑ/ or /dz/ instead of /z/. In Japanese, the affricate /dʑ/ and the fricative /ʑ/ occur before /i/; /dʑ/ or /dz/ usually appear word-initially and /ʑ/ or /z/ word-medially [8]. Thus, it is difficult for Japanese speakers to pronounce /z/ [9], especially in word-initial position. In their comments, some of the evaluators described Z as a "strong /d/ sound mixed in," "hard to decide G or Z," and "the sound in between G and Z." This confusing pronunciation is reflected in the results shown in Table 3; the native English speakers heard the JLs' productions of Z with low intelligibility as Z or G at the same percentage.

The dark /l/ was the most difficult sound for the JLs to pronounce. A strong Japanese accent in the JLs' pronunciation was mentioned in most comments, partly due to the additional vowel at the end and partly because of /l/ sounds different from the dark /l/. In order to gain a deeper comprehension of the dark /l/ pronunciation among young JLs, it is imperative to undertake more detailed and extensive research (quantitative and qualitative).

The JLs' difficulties in pronouncing the above five English consonants have also been observed among adult Japanese speakers. Nishio and Joto [10] reported that over 50% of university students participating in their experiment made errors in pronouncing the letter names of L, P, T, V, and Z. The VOT was shorter in their productions of P and T than in native English speakers' ones. Aspiration of voiceless stops is listed as one of the Lingua Franca Core crucial for international communication [11].

There are some pairs of phonemic contrast among the English letters: B-P, B-V, T-D, and Z-G (in American English). These contrasts carry a higher functional load [12] and need to be taught to young Japanese learners with careful attention to avoid the stabilization of pronunciation. From a pedagogic point of view, we suggest that P and T be pronounced with strong aspiration, and V and Z with longer friction in initial position. As for L, the tongue tip should touch the alveolar ridge while the lips make the gesture of the Japanese vowel /o/ in order to realize velarization of dark /l/, and should be pronounced without an additional vowel at the end.

5. CONCLUSIONS

The intelligibility levels of young JLs' pronunciations of English alphabet letter names differed based on the letter in question as well as the student. Acoustic features related to intelligibility provide suggestions for teaching English pronunciation to young JLs.

6. ACKNOWLEDGMENTS

This study was supported by JSPS KAKENHI Grant numbers JP18K00787, JP20H01295. We would like to extend our gratitude to Mr Kazuhiro Toi, currently vice-principal at an elementary school in Hiroshima City, for collaborating with elementary school students on their recordings.

7. REFERENCES

[1] Akatsuka, M. and Yamami, Y. 2016. Nihonjin shogakusei niokeru eigo no hatsuon to kokusaiteki rikaido no ichi-kosatsu [English pronunciation of elementary school students and their international intelligibility: focusing on word front consonants]. *Journal of the School of Contemporary International Studies, Nagoya University of Foreign Studies*, 12, 51-61.

[2] Kato, T. and Yamamoto, S. 2020. Nihonjin shogakusei no eigo onsei no jizokuteki shuroku to eigo tanboin no onkyoteki jikanteki bunseki [Longitudinal collection of English speech produced by Japanese elementary school children and acoustic and temporal analysis of their vowel production]. *Journal of the Phonetic Society of Japan*, Vol. 24, 71-83.

[3] Joto, A. and Nishio, Y. 2021. The phonetic features of English diphthongs produced by Japanese elementary school students [Online poster presentation], *The 19th World Congress of Applied Linguistics (AILA2020),* Groningen, the Netherlands.

[4] Boersma, P. and Weenink, D. 2022. Praat, ver. 6.3.03, https://www.fon.hum.uva.nl/praat/

[5] Olive, J. P., Greenwood, A., and Coleman, J. 1993. *Acoustics of American English Speech: A Dynamic Approach*. Springer.

[6] Ladefoged, P. 2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Blackwell.

[7] Baken, R, J. 1987. *Clinical Measurement of Speech and Voice*. Taylor & Francis Ltd.

[8] The Japan Foundation. 2009. *Onsei o Oshieru* [Teaching Pronunciation], The Japan Foundation Teaching Japanese Series 2. Hitsuji Shobo.

[9] Shirota, S. 1995. *Nihongo no Oto* [The Sounds of the Japanese Language]. Hitsuji Shobo.

[10] Nishio, Y. and Joto, A. 2023. Improving fossilized English pronunciation by simultaneously viewing a video footage of oneself on an ICT self-learning system. In: Albuquerque, J. I. A. and Alves, U. K. (eds), *Second Language Pronunciation: Different Approaches to Teaching and Training*. De Gruyter/Mouton, 249-286.

[11] Jenkins, J. 2000. *The Phonology of English as an International Language*. Oxford University Press.

[12] Cruttenden, A. 2014. *Gimson's Pronunciation of English*, 8th ed. Routledge.

L3 NORWEGIAN /ʉː/ and /uː/ IN L1 POLISH/L2 ENGLISH LEARNERS: DIFFERENT PATTERNS OF CROSS-LINGUISTIC INTERACTIONS

Jarosław Weckwerth, Magdalena Wrembel, Kamil Kaźmierski and Anna Balas

Adam Mickiewicz University, Poznań, Poland
{wjarek;magdala;kamil.kazmierski;abalas}@amu.edu.pl

ABSTRACT

The spectral properties of Norwegian /ʉː/ and /uː/ were studied in a population of L1 Polish/L2 English speakers acquiring Norwegian as an L3 in an instructed setting in Poland. The vowel /ʉː/ is a rounded high central vowel, displaying a combination of features nominally absent from the speakers' L1 and L2. In addition to descriptive formant measurements, two measures of overlap were used to ascertain if a separate category can be argued to develop for /ʉː/. The results indicate that a distinction is formed against Polish /u/ but not necessarily against English /uː/. Change over time is inconclusive at these early stages of acquisition. Interestingly, /uː/ seems to display influence of L2 acoustics and L1 orthography.

Keywords: Vowel formants; Norwegian; Polish; English; vowel overlap; cross-linguistic influence.

1. INTRODUCTION

1.1. Theoretical background, previous research

Third and additional language acquisition has become recognized as an independent field, based on growing empirical evidence that it is quantitatively and qualitatively different from second language acquisition mainly due to broadened phonetic repertoires, enhanced perceptual sensitivity, higher metalinguistic awareness, language-learning experience and strategies, among others [1, 2]. Systematic enquiries into L3/Ln phonological acquisition have focused on the sources and directionality of cross-linguistic influence (CLI) between the speaker's L1, L2, L3/Ln subsystems, attesting to different types (i.e. progressive, regressive, combined) of multidirectional CLI (L1 ↔ L2, L1 ↔ L3, L2 ↔ L3, etc) modulated by an interplay of factors (see [3] for an overview).

Previous studies on L3 vowel quality acquisition have been rather limited to date [4, 5, 6]. Missaglia [4] demonstrated a bilingual advantage for the acquisition of L3 English vowels in Italian-German bilingual children. Hypothesizing a multilingual system as a 'global entity', Sypiańska [5] showed that multilinguals' vowel space is subject to reshaping in the component languages, consequently becoming less peripheral, and divergent from monolingual baseline data. She also found evidence of combined CLI from L1 Polish and L2 Danish on L3 English vowel production, in line with previous findings. Kopečková et al. [6] investigated the interactions between three vocalic systems in young speakers of L1 German, L2 English and L3 (heritage) Polish. The findings demonstrated high individual variability in vowel production in all three languages, with language status and use as one of the conditioning factors.

The present study aims to contribute further to investigations of the complexities of potential interactions of the speakers' multiple languages while zooming in on selected vowel qualities.

1.2. A basic comparison of the three languages

Polish has a simple six-vowel oral vowel system with three high vowels, /i/, /ɨ/ and /u/. To facilitate transcription-agnostic reference, we will refer to them as BITY, BYTY, BUTY, on the model of the Wellsian keywords for English [7]. Only the latter vowel is rounded, with an F2 in the region of 900 Hz in females [8]. There is no phonological length.

English has a larger set of high vowels comprising FLEECE, KIT, FOOT and GOOSE, traditionally symbolized phonemically as /iː/, /ɪ/, /ʊ/, and /uː/. At first glance, the rounding vs. backness situation may seem similar to that in Polish, with front unrounded and back rounded vowels. However, in many modern English accents, including General British, GOOSE and FOOT have undergone significant fronting in the last 50+ years, and in younger speakers may be regarded as at least central (Deterding [10] citing female F2 values around 1400 Hz, and Bjelaković [18] – around 1800 Hz). This change has been robust enough to be recognized by reference works such as Cruttenden [9] and has seen considerable research from various angles, including its availability to L2 learners [11]. Length is usually taken not to be contrastive, with quality being the chief distinctive factor.

Norwegian has the most complex high vowel system of the three languages, including front unrounded /i(ː)/ (keywords TID, MITT), front (weakly) rounded /y(ː)/ (LYS, NYTT), central rounded /ʉ(ː)/ (GUD, SLUTT) and back rounded /u(ː)/ (BOK, BORT) [12]. Length is phonemic, with only small spectral differences between short and long vowels at each position.

In this paper, we zoom in on the production of Norwegian GUD and BOK, English GOOSE, and Polish BUTY. This is motivated by the combination of height, rounding and backness in Norwegian GUD which is nominally absent from English and Polish. In the light of the fronting of GOOSE, this opens up a number of logical possibilities of cross-linguistic interactions.

Against this background, a few tentative working hypotheses may be formulated:

H1: New categories in L3 Norwegian could form if they are sufficiently dissimilar from L1 Polish and L2 English (cf. Flege [13]).

H2: We hypothesise considerable L1→L3 CLI due to automatized neuro-motor articulatory routines of the native language.

H3: We also expect L2–L3 interactions based on the intrinsic phonetic similarities between English and Norwegian high vowel systems as well as a frequently attested 'foreign-language effect' [1].

H4: We predict developmental changes in spectral overlap as a function of time and learning experience (L3→L1).

2. METHODS

2.1. Participants

Ten female students (mean age = 20) majoring in Norwegian at two Polish tertiary education institutions took part in a larger longitudinal study. They were selected from a larger group that also included six male speakers. They had started their programme one month before the first recording session, and were fresh starters without any prior proficiency in Norwegian. All had Polish as their L1, with no language deficits reported in a background questionnaire; and all reported English at the CEFR level B2. Their English proficiency was tested using a placement test as part of the first data collection session.

2.2. Recording sessions

The recording sessions were part of three data collection rounds, abbreviated T1, T2 and T3. T1 took place in November 2021, T2 in March 2022 and T3 in June 2022. Norwegian was collected on the first day of each round, while English and Polish were collected on the second day. Material for other parts of a larger project (including a battery of production and perception tasks, written grammaticality judgements and survey data) was also obtained. The audio recordings accounted for about half of the data collection time.

2.3. Stimuli

Read speech was collected using slides presented on a computer screen at a comfortable pace controlled by the experimenter. We collected all the monophthongs in the three languages, embedded in priming words and target non-words inserted into carrier sentences. These had the general form of *There is the same vowel in 'bit' and 'dit'. The first one is 'bit', and the second one is 'dit'*. In this way, four instances of each vowel (here, English KIT) were collected from each pair of sentences. Additionally, material for other parts of the larger project (e.g. related to plosive aspiration) was collected.

2.4. Recordings, data processing, measurements

The recordings were made in quiet rooms in university offices. Audio was captured using a Shure SM-35 unidirectional cardioid head-worn condenser microphone. It was digitized at 48 kHz, 16 bit using a portable Marantz PMD661 solid state recorder.

The recordings were subsequently force-aligned using the online interface provided by BAS Munich [14]. The alignments were corrected manually by three trained phoneticians, with target vowel boundaries aligned on the basis of presence of periodic waveform, F2 energy, and amplitude relative to the neighbouring consonants. A Praat [15] script was used to obtain readings of the first three formants of each target vowel, along with the duration. All measurements falling more than a standard deviation from the mean for each vowel were verified manually and corrected.

The statistical analyses were run in R using the *lme4* package for mixed effects regression modelling [16].

3. RESULTS

3.1. General overview

Fig. 1 and Table 1 provide a general overview of the results. From a visual inspection, Norwegian GUD did seem to form a category separate from Polish BUTY and BITY already at T1 but it was less clear what its relation was to English GOOSE. We decided to investigate the overlap between the L2 and L3 vowels.

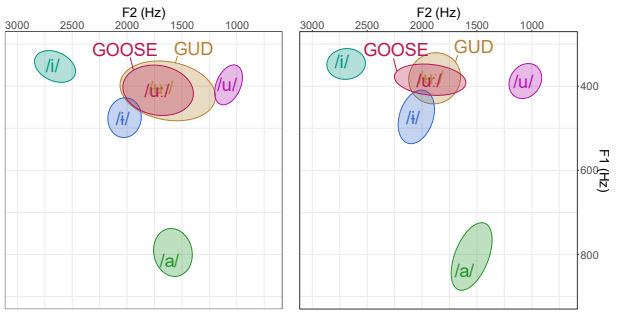


Fig. 1: Aggregate raw Hz data for T1 (left) and T3 (right). Polish /i/, /ɨ/, /u/ and /a/ included as anchors. Ellipses at a 0.5 confidence level.

Vowel	F	T1	T2	T3
BUTY	F1	400	402	388
GOOSE	F1	409	387	384
GUD	F1	406	390	376
BUTY	F2	1068	1053	1059
GOOSE	F2	1727	1769	1926
GUD	F2	1646	1981	1898

Table 1: Raw Hertz group mean formant values for the three data collection times. BOK excluded due to extreme inter-speaker variability.

3.2. Assessing overlap

Two methods were used to assess spectral overlap at the three data collection times: Pillai scores and Mahalanobis distances (as recommended by [17]).

The Pillai score is an output of a MANOVA model that represents the distance between two distributions on a scale from 0 (total overlap) to 1 (complete separation). The Pillai scores for GUD vs. BUTY increased over time, in particular between T1 and T2, indicating improved separation, presumably due to the fronting of GUD/GOOSE visible in Fig. 1; importantly, the scores for most speakers were high at T1, suggesting that even at the very start of their Norwegian course, they had already established a category for the L3 vowel that was separate from their L1.

The situation is far less clear for L3 GUD vs. L2 GOOSE. There was evident overlap between the two vowels at all three Ts, and they both increased in F2 (thus improving separation from BUTY). There was considerable inter-speaker variability. For several speakers, the scores were very low at T1 (indicating substantial overlap), but even the top scores were lower than for GUD vs. BUTY. At T2 and T3, the variability remained, with the Pillai scores increasing for some speakers but decreasing for others. (See Fig. 2.)

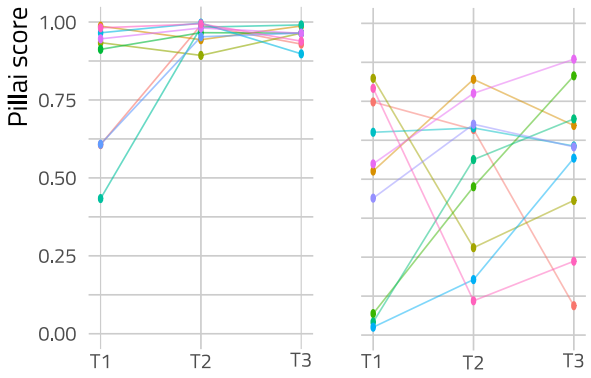


Fig. 2: Pillai scores for GUD–BUTY (left) and GUD–GOOSE (right) at T1, T2 and T3. Each line is one speaker.

A similar picture emerges from the second analysis using Mahalanobis distances. The Mahalanobis distance is the distance between a point and a distribution; in vowel merger studies, it is the distance between a token of one category and the distribution of another category. The greater the score, the greater the distance (thus less overlap).

We calculated distances for each vowel pair in both directions for each speaker. The mean distances between GUD and BUTY increased over time for all but two speakers, even though the main change occurred between T1 and T2, with some apparent reversal between T2 and T3 (see Fig. 3). The picture for BUTY vs. GUD was similar, but with less inter-speaker variation – presumably due to lower variability within the L1 category.

The patterns of developmental change in the mean distances between L3 GUD and L2 GOOSE were again far less clear. At least for some speakers, the mean distance seemed to decrease over time, in particular in the GUD–GOOSE direction, suggesting decreased separation. The apparent anomaly at T2, especially for GOOSE–GUD, may be due to the forward 'leap' of GUD visible from the means in Table 1, which increased distance from GOOSE; at T3, GOOSE seems to have caught up, resulting in a 'reversal'. Fig. 3 summarizes the analyses.

The developmental changes were assessed using mixed-effects regression models fitted to each vowel pair, with Mahalanobis distance as the response variable, *time* as the treatment-coded categorical fixed predictor of interest (levels: 't1', 't2', 't3', reference: 't2') and a by-speaker random intercept. Statistical significance was assessed with likelihood ratio tests. They retuned time as a significant factor in all models, i.e. there were significant differences between the three data collection points.

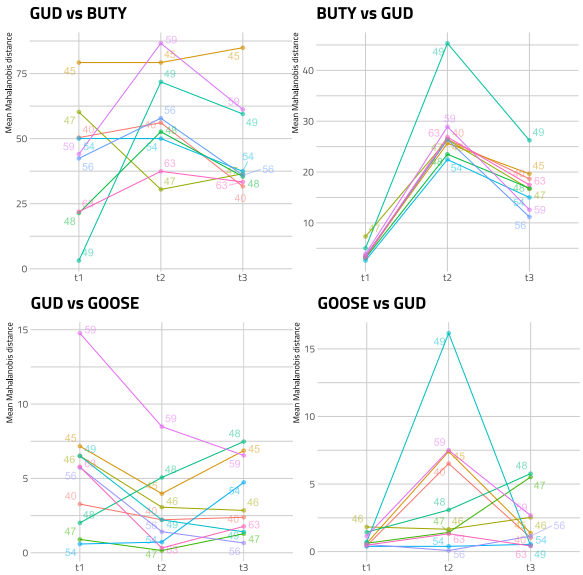


Fig. 3: Mahalanobis distances for T1, T2, and T3. Each line represents one speaker.

3.3. Variability and relation to spelling

We found considerable inter- and intra-speaker variability in the treatment of BOK. In a subgroup of our speakers, BOK was mapped onto Polish BUTY, as expected on the basis of its phonology/phonetics. This may be seen as corroborating our H2 (of the facilitative effect of automated L1 articulatory routines). However, most speakers showed two other patterns.

In some speakers, BOK was mapped onto Polish /ɔ/ (BOTY). We think that the obvious explanation of interference from L1 orthography is sufficient in view of the fact that the data in the present study come exclusively from a reading task.

Very interestingly, other speakers realized BOK with qualities similar to their GUD/GOOSE. We would be inclined to name this a 'foreign /u/ effect' (cf. the 'foreign language effect' in [1]): the fronter qualities shown in GUD/GOOSE may be seen as 'allophonic' 'foreign' realizations of an underlying /u/ category, and once a phone is assigned to this category, the realization applies even if it is not required by the phonetics of the Ln phone; in other words, 'BOK is /u/, but foreign, and foreign /u/ is [ʉ]'. This effect may be seen to also confirm our H3: the 'foreign language effect' means that the phonetic realization of an L3 phone is guided by an L2 phone (GOOSE in our case), resulting in an apparent merger whose phonetic quality derives from the L2 rather than L1. Additional evidence for this pattern is provided by a perception study from this same project, in which GUD and BOK were, somewhat surprisingly, perceptually assimilated to Polish /u/ at similar rates by Polish listeners.

Finally, in some speakers, there was evidence of inter-speaker variability with two or three realizations. This is summarized in Fig. 5.

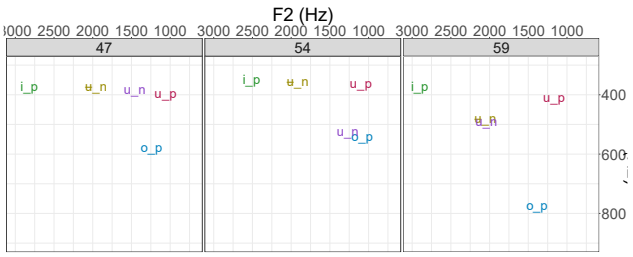


Fig. 4: Inter-speaker differences for BOK (u_n). Polish /i/, /u/, /ɔ/ shown as i_p, u_p, o_p; GUD as ʉ_n. Each panel is one speaker. Separate high back (Speaker 47); merged with /ɔ/ (Sp. 54); merged with GUD (Sp. 59).

4. DISCUSSION

Overall, GUD and GOOSE were well differentiated from BUTY. This seems to partially confirm our H1 in that there must be sufficient acoustic separation between the native and foreign languages to warrant formation of a new category in the L2/L3.

We do not have English production data from our speakers that would predate their Norwegian course. Consequently, we cannot state with any certainty that these speakers had acquired appropriately 'modern' central qualities of GOOSE first (i.e. before starting their Norwegian). However, existing literature, e.g. Šimačková & Podlipský [11], seems to suggest that this is indeed a possibility. In any case, what *can* be stated with relative confidence is that that the speakers were sensitive to the spectral similarity between GUD and GOOSE. Also, the two vowels seemed to drift towards higher F2 vowels jointly over time. Overall, this seems to, at least partially, confirm our H3.

Interestingly, the acoustic characteristics of the vowels seemed to override any putative phonological assignment of GOOSE as back, and the combination of non-backness with rounding was not a complicating factor. This is despite the unanimous assignment of English GOOSE borrowings in Polish to Polish /u/ (and despite the <u> spelling in Norwegian GUD). Thus, L1 articulatory routines did not play a major role for these two vowels, contra our H2.

However, we do have evidence that the spectral characteristics are only one possible source of cross-linguistic influence. Given the spectral similarity between Polish BUTY and Norwegian BOK, one would be inclined to predict that the acquisition of the latter should pose no problems for Polish learners. This is, in fact, *not* the case in our data (see 3.3 above).

The extremely varied developmental paths visible in both the raw data and the Pillai and Mahalanobis results for the individual speakers serve as confirmation of our H4 (of substantial developmental changes). They point to considerable inter-speaker differences in the attested patterns of development, to the extent that straightforward interpretation of the direction of change is difficult. In particular, the apparent reversal of some trends in some speakers between T2 and T3 requires more analysis.

5. CONCLUSION AND OUTLOOK

We have observed a complex pattern of cross-linguistic interactions and developmental changes in trilingual acquisition of typologically untypical vowels. However, several important avenues remain open for further research.

Currently, we are investigating naturalistic acquirers in the same configuration of languages. In addition to their realizations of GUD (where its dissimilation from GOOSE will be the focus), we are also looking at BOK in spontaneous speech, where the influence of spelling is less strong., We are also collecting relevant data from Norwegian English and Polish English to apply a subtractive L1–L2 design on the model of Westergaard et al. [19].

6. ACKNOWLEDGEMENTS

This research is supported by Norway funds/NCN project grant GRIEG-1 (UMO-2019/34/H/HS2/00495) ADIM "Across-domain investigations in multilingualism: ModelingL3 acquisition in diverse settings".

7. REFERENCES

[1] De Angelis, G. 2007. *Third or additional language acquisition*. Multilingual Matters.
[2] Gut, U. 2010. Cross-linguistic influence in L3 phonological acquisition. *International Journal of Multilingualism* 7(1), 19–38.
[3] Cabrelli Amaro J, Wrembel, M. 2016. Investigating the acquisition of phonology in a third language – a state of the science and an outlook for the future. *International Journal of Multilingualism* 13(4), 395–409.
[4] Missaglia, F. 2010. The acquisition of L3 English vowels by infant German-Italian bilinguals. *International Journal of Multilingualism* 7(1), 58–74.
[5] Sypiańska, J. 2016. Multilingual acquisition of vowels in L1 Polish, L2 Danish and L3 English. *International Journal of Multilingualism*, 13(4), 476–495.
[6] Kopečková, R., Marecka, M., Wrembel, M., Gut, U. 2016. Interactions between three phonological subsystems of young multilinguals: The influence of language status. *International Journal of Multilingualism* 13 (4), 426–443.
[7] Wells, J. 1982. *Accents of English*. CUP.
[8] Weckwerth, J., Balas, A. 2020. Selected aspects of Polish vowel formants. In: Wrembel, M., Kiełkiewicz-Janowiak, A., Gąsiorowski, P. Approaches to the study of sound structure. Routlegde, 338–348.
[9] Cruttenden, A. 2014. *Gimson's pronunciation of English*. (8th edn.) Routledge.
[10] Deterding, D. H. 1997. The formants of monophthong vowels in standard southern British English pronunciation. *Journal of the International Phonetic Association* 27, 47–55.
[11] Šimáčková, Š., Podlipský, V. 2017. Foreign language learners acquire L2 phonetic detail: GOOSE and FOOT fronting in non-native English. Research in Language 15(4), 358–401.
[12] Kristoffersen, G. 2000. *The phonology of Norwegian*. OUP.
[13] Flege, J.E. 1995. Second language speech learning: Theory, findings and problems. In: Strange, W. (ed.), *Speech perception and linguistic experience: Issues in cross-language research*. York Press, 233–277.
[14] Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
[15] Boersma, P., Weenink, D. 1992–2022. Praat: doing phonetics by computer [Computer program]. https://www.praat.org.
[16] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
[17] Nycz, J., Hall-Lew, L. 2014. Best practices in measuring vowel merger. *Proc. Meet. on Acoustics* 20, 060008.
[18] Bjelaković, A. 2017. The vowels of contemporary RP: vowel formant measurements for BBC newsreaders. *English Language & Linguistics* 21(3), 501–532.
[19] Westergaard, M., Mitrofanova, N., Mykhaylyk, R., Rodina, Y. 2017. Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model. *International Journal of Bilingualism* 21(6), 666–682.

f0 enhancement in Japanese voicing contrast by Japanese native speakers and its implication to L2 perception/production

Keiji Iwamoto

Indiana University Bloomington
kiwamoto@iu.edu

ABSTRACT

This study investigated how f0 was enhanced when Japanese native speakers try to disambiguate minimal pairs of voicing contrast (i.e., the directions of f0 enhancement: up or down). A phonological focus task was utilized (e.g., it's not ***t****enki*, 'weather', but ***d****enki*, 'electricity'). Speakers should maximize phonological contrast when disambiguating the minimal pair. The results showed that f0 was raised for voiceless items and depressed for voiced items when the items were in focus (i.e., when they tried to maximize the phonological contrast) in comparison to 'unfocus' items. The results indicate that speakers have control over the manipulation of f0 (support for the concept of controlled phonetics [8]). These findings also contribute to our understanding of L2 perception and production, especially of learners whose L1 are 'aspirating languages' (e.g., Mandarin Chinese).

Keywords: voicing contrast, phonological focus, Japanese

1. INTRODUCTION

The motivation for the current study is to investigate the acquisition of the L2 phonological system. However, this study aims to examine what the target language speakers do in the first place. The linguistic target is Japanese voicing contrast and the target learner's L1 is Mandarin Chinese. First, the relationship between Japanese voicing contrast and f0 will be reviewed and the results of the L1 production study (the current experiment) will be discussed. Second, based on the results and the past literature, some implications for L2 perception of Japanese voicing contrast will be discussed.

Japanese voiced stops have pre-voicing (negative VOT), whereas voiceless stops have short-lag aspiration (positive VOT). It might be this difference in VOTs that Japanese speakers rely on to distinguish the voicing contrast in Japanese. However, studies of modern Japanese have indicated that the phonetic cue of pre-voicing is playing a less important role in voicing contrast due to an overlap of the VOTs between voiced and voiceless sounds. That is, there has been no prominent difference between voiced and voiceless stops in terms of VOTs [11]. If it is the case that pre-voicing does not help much in terms of Japanese people's perception of voicing contrast, there must be other acoustic cues that Japanese people recruit to distinguish voiced and voiceless sounds. Previous studies have suggested that in addition to VOT, other acoustic cues, such as consonantally induced f0 (CF0) might be what has been employed in voicing contrast by Japanese native speakers [2, 4], especially in the word-initial contexts.

This study investigated the importance of the f0 cue for Japanese voicing contrast by having speakers disambiguate minimal pairs of /t/ and /d/, where they must maximize the voicing contrast. The phonological focus task was employed [3]. The phonological focus task has subjects disambiguate minimal pairs differing only in one phoneme (e.g., *ma****d*** vs. *ma****t***). The task has been used to diagnose which acoustic cues or phonetic features are employed for the phonemic contrast in a language [7]. This method allows us to examine which acoustic cues would be perceived as more important by speakers.

There are two conditions: focus and unfocus. In the focus condition, they need to disambiguate minimal pairs, whereas, in the unfocus condition, they do not need to differentiate minimal pairs. The followings are examples.

(1) Focus
It's not ***t***enki but ***d***enki.
It's not ***d***enki but ***t***enki.

(2) Unfocus
She didn't say ***t***enki but ***he*** did.
She didn't say ***d***enki but ***he*** did.

Since devoicing of voiced stops has been observed in modern Japanese, the pre-voicing cue is less robust. In addition, positive VOT cue also seems to be overlapping between voiced and voiceless stops. Therefore, f0 might play a key role in the maximization of the voicing contrast.

The previous studies on CF0 revealed that phonologically voiceless stops have higher CF0 than voiced counterparts [6]. This study assumes the controlled phonetics that speakers have the capability to control f0 manipulation [8]. That is, f0 is employed

intentionally by speakers to maximize the voicing contrast. The question is whether there would be a difference in f0 between the focus and unfocus conditions.

If native speakers would like to manipulate f0 to maximize the contrast, the direction of the f0 enhancement would be *raising* (up) after voiceless stops, whereas it would be *depressing* (down) after voiced stops. The degree of f0 enhancement would be larger in the focus condition than in the unfocus condition.

2. METHODS

2.1. Speakers

Three monolingual native Japanese speakers (3 males) with a Tokyo dialect were recruited remotely. Because the VOT shift is happening among younger generations [11], participants were between the ages of 22 and 32. They had not been exposed to extensive second language training and linguistics. This was to avoid L2 influence on the L1.

2.2. Items

Two minimal pairs of /t/ and /d/ were used: *tenki*, 'weather', & *denki*, 'electricity', and *temae*, 'front', & *demae*, 'catering'. Both /t/ and /d/ occur in the word-initial context where f0 is most likely to be used [4]. The former /t/ and /d/ occur in the syllable with a high tone, whereas the latter pair occurs in the syllable with a low tone. As well as the two minimal pairs of /t/ and /d/, minimal pairs of nasals (/m/ and /n/ i.e., sonorants) for each tonal context were employed to establish a baseline to which we can compare f0 values of post-plosive vowels. Sonorants, unlike obstruents, do not create f0 perturbation effects in the following vowel [4]. The minimal pairs of /m/ and /n/ were these two pairs: *maasu*, 'Mars' & *naasu*, 'nurse' (word-initial & high tone), and *makeru*, 'to lose' & *nakeru*, 'to make/let you cry' (word-initial & low tone). The following vowels were not matched between the voicing items and nasal pairs because of the author's mistake. However, since the purpose of having nasal pairs is to set the baseline of the *directions* of f0 enhancement, the following vowel should not play a significant role.

2.3. Task

As discussed earlier, the phonological focus paradigm was employed in the current study [3]. The lexical items were placed in carrier sentences. In addition to making the onsets and closures of the target plosives more distinguishable, reciting words in a carrier sentence makes the overall acoustics of the sentence (and thus the items) more stable. Two types of sentences are used to elicit different focusing effects. First is the lexical focus condition in which the speaker corrects a word from the list of minimal pairs (i.e., the focus is on the target word; hereafter referred to as 'focus'). Second is the postnuclear focus condition in which the speaker also employs correction, but in a frame preceding the target word (hereafter referred to as 'unfocus'). This condition will act as a baseline for the lexical focus condition. The sentence structures are simply Japanese translations of de Jong's [3]; in both English and Japanese, the position of the target words and emphases of these sentences are approximately the same.

(3) The sentence frame for 'focus' conditions
これは X じゃなくて、Y です。
'This is not X, it is Y.'
これは Y じゃなくて、Xです。
'This is not Y, it is X.'

(4) The sentence frame for 'unfocus' conditions
彼女じゃなくて彼が X と言った。
'She didn't say X, but he did.'

Data collection was done remotely, without supervision. Participants were given instructions prior to the start of the experiment, such as recording in a quiet environment with minimal electronic noise. The recording was done by the participants themselves on their phones or laptop, and the audio files were sent directly to the researcher upon completion.

2.4. Measurement

Praat was used to analyze the speech data. After annotating, f0 on post-stop vowels was extracted through Praat scripts [12]. Semitone was used as a unit for f0. f0 differences in hertz are not a good measurement because f0 needs to be perceived on a relative scale. For example, absolute value means nothing when male and female speech are compared (male voice is inherently lower than female voice). It is said that a logarithmic scale of f0 can approximate human perception and the semitone is used as a unit for the logarithmic scale of f0 [1]. One semitone difference can be consciously perceived by people.

3. RESULTS

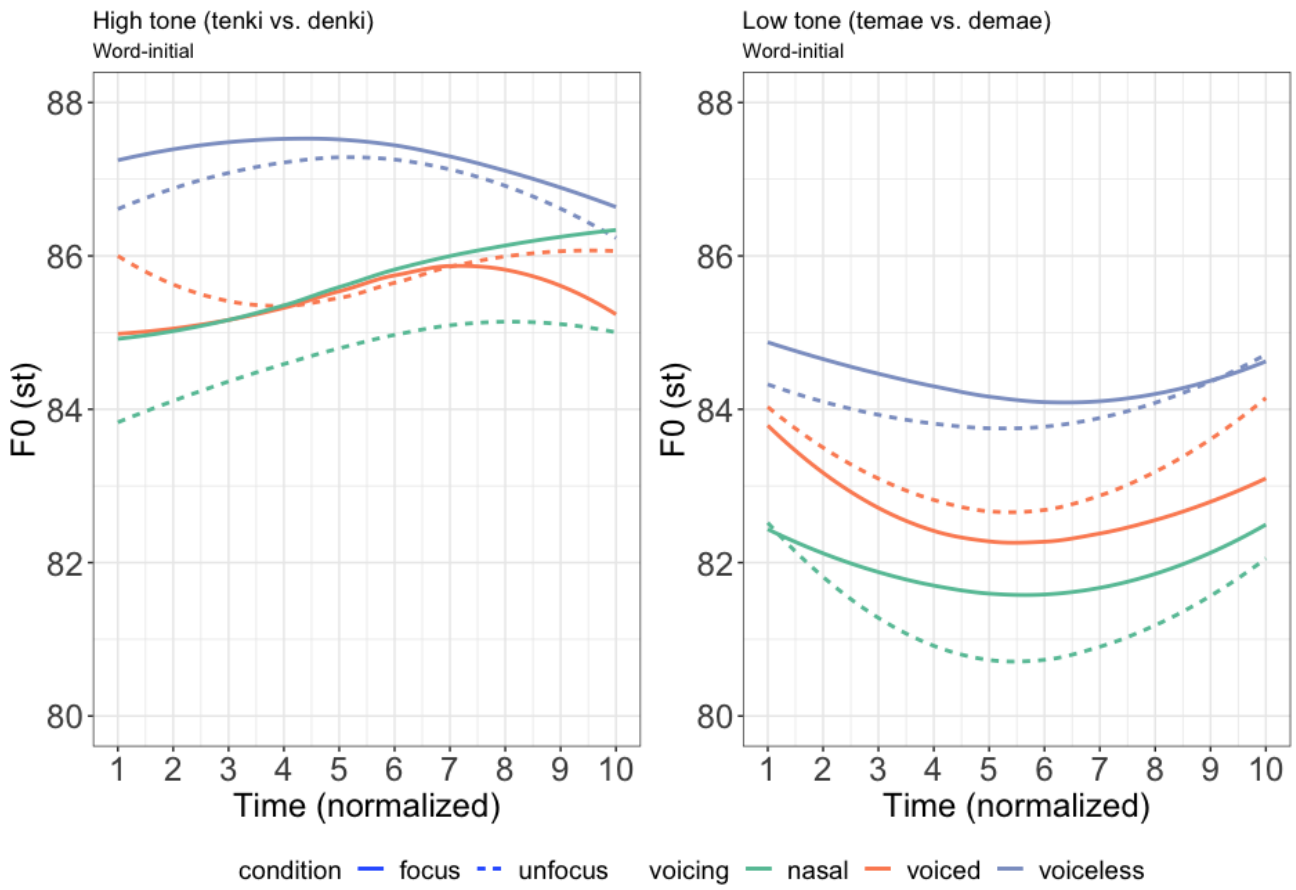


Figure 1: The f0 contours on post-stop vowels in semitone (st)

Figure 1 shows the f0 contours on post-stop vowels. Y-axis shows f0 in semitone and X-axis shows normalized time. There are two conditions: focus (Solid line) and unfocus (Dashed line). What we are interested in is the directions of f0 enhancement (up or down).

The *directions* of f0 enhancement would be analyzed; f0 depression for voiced stops and f0 raising for voiceless stops. Speakers are expected to enhance f0 in this manner so that they can maximize the contrast. The direction of f0 enhancement for nasals between focus and unfocus conditions will also be analyzed to check the default f0 enhancement effect due to the focus effect. The f0 direction of the nasals will be the baseline of the analysis.

First, let's look at the nasals (baseline). If we compare the focus and unfocus conditions, both plots show that when items are in focus, f0 tends to be raised. In a high-tone context, the difference between the conditions is clear, whereas, in a low-tone context, there is no difference at the f0 onset but there is a clear difference at the middle time points. In sum, when items are in focus, f0 is raised (even when there is no CF0 effect).

The left plot shows a high-tone context (*tenki* vs. *denki*). As expected, there are visually significant differences between voiced and voiceless stops (i.e., CF0 effect), especially at the f0 onset. There is also a focus effect. The direction of the f0 enhancement is 'up' for voiceless items, as predicted. Since the direction of the enhancement for nasals is also 'up', we are still not sure if the enhancement is the enhancement for voicing contrast or for the overall saliency of the syllable. However, the direction is 'down' for voiced items, which is different from the baseline. Therefore, this f0 depression due to the voiced items could be interpreted as an intentional manipulation by speakers. They might have depressed f0 to maximize the voicing contrast on purpose.

The right plot shows a low-tone context (*temae* vs. *demae*). The same tendency can be observed. The direction of the f0 enhancement is 'up' for voiceless items and 'down' for voiced items.

4. DISCUSSION

4.1. Support for Controlled Phonetics

This study investigated the directions of f0 enhancement. The target context was the word-initial context. In both high-tone and low-tone contexts, the

directions were the same: f0 rising for voiceless stops and f0 depression for voiced stops.

The results provide support for controlled phonetics [8]. It seemed that speakers deliberately depressed f0 in the post-voiced-stop vowels (and possibly raise f0 in the post-voiceless-stop vowels) to maximize the voicing contrast.

4.2. Implications to L2 Perception/Production

Based on the past literature and the results of the current study, some implications can be made for second language acquisition.

4.2.1 Implications to L2 Perception

The results suggest that f0 seems to play a role in Japanese voicing contrast when speakers try to disambiguate the contrast. In addition, the results of the past literature indicate that positive VOT cues and pre-voicing cues are less robust for the voicing contrast.

We are not sure if the f0 cue is a primary cue or a secondary cue (i.e., a redundant cue) for the perception of Japanese native speakers, but it seems that f0 is used when VOT and pre-voicing cues are ambiguous [5]. However, the redundant cue could be a primary cue for L2 learners, especially for L2 learners whose L1 does not employ pre-voicing as a phonologically distinctive cue (e.g., Mandarin Chinese). This is because Mandarin Chinese, for example, employs aspiration for the voicing contrast. L1 Mandarin Chinese learners of L2 Japanese have difficulty in the Japanese voicing contrast [9]. The previous research on the L2 perception of Japanese voicing contrast did not manipulate f0 but given the results of the current study and the past literature [5], further L2 research with the variable of f0 is warranted. Having Mandarin Chinese learners of L2 Japanese as participants would be interesting because it employs f0 for lexical contrast (i.e., tone) and thus they are sensitive to f0 even in non-native languages [10].

The hypothesis is that although VOT and pre-voicing cues may not be readily available to L1 Mandarin Chinese learners of L2 Japanese, they may be able to distinguish the Japanese voicing contrast when f0 is perceptually different enough, because of their sensitivity to f0 [10]. If this is the case, Mandarin learners should be able to distinguish Japanese voicing contrast in the word-initial contexts where f0 seems to be employed for the contrast [4], whereas they should have difficulty in the intervocalic (word-medial) contexts where f0 is not employed for the contrast [4].

4.2.2 Implications to L2 Production

In a similar vein, the implications for L2 production can be discussed too. If Mandarin learners of L2 Japanese perform the same task (the phonological focus task), and if they are sensitive to f0 in L2 perception (in other words, the input to learners would lead to learning of L2 perception), they should be able to produce the Japanese voicing contrast distinctively in the word-initial contexts (f0 is available), whereas they may not be able to in the intervocalic (word-medial) contexts (f0 is not available). Other learners whose L1 employs aspiration for L1 voicing contrast but is not a tonal language (e.g., English) can be recruited for the comparison.

5. CONCLUSION

In conclusion, this study aimed to examine what Japanese native speakers do in their production of the voicing contrast to draw implications for L2 production/perception, especially for Mandarin Chinese learners of L2 Japanese. More specifically, this study investigated whether f0 is enhanced when Japanese native speakers attempt to disambiguate minimal pairs of /t/ and /d/. The results showed the deliberate manipulation of f0 to maximize the contrast, suggesting that f0 plays a role in the voicing contrast, at least in the word-initial contexts. Based on the results, hypotheses, and predictions were generated regarding L2 perception and production

These hypotheses on L2 perception and production should be applicable to other typologically similar language pairs. Specifically, when tonal and 'aspirating' language speakers learn 'true voicing' languages (e.g., Japanese, Russian, Spanish, Swedish, etc.), their sensitivity to f0 differences learned in their L1 may play a role in their L2 perception and production. A seemingly redundant cue (i.e., f0) may become a precious cue when viewed from a different perspective (i.e., L1).

6. ACKNOWLEDGEMENTS

I thank Lily Kawaoto for help as a team member of the Japanese L1 production study, and Professor Kenneth de Jong for his guidance. The shortcomings in this paper are certainly my own.

6. REFERENCES

[1] Baart, J. 2010. *A field manual of acoustic phonetics*. SIL International.

[2] Byun, H. 2021. Acoustic characteristics for Japanese stops in word-initial position: VOT and post-stop f0. *Journal of Phonetic Society of Japan,* 25, 41-63.

[3] de Jong, K., J. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics*, 32. 493-516.

[4] Gao, J., & Arai, T. 2019. Plosive (de-)voicing *f*0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics,* 77, 1-33.

[5] Gao, J., Yun, J., Arai, T. 2019. VOT-F0 coarticulation in Japanese: Production-biased or misparsing?. *Proc. 19th ICPhS* Melbourne, 210–214.

[6] Guo, Y., Kwon, H. 2022. Production and Perception of Mandarin Laryngeal Contrast: The Role of Post-plosive F0. *Frontiers in Communication,* 7, 1-15.

[7] Kim, S., Kim, J., Cho, T. 2018. Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics,* 71, 65-80.

[8] Kingston, J., Diehl, R. L. 1994. Phonetic knowledge. *Linguistic Society of America* 70(3), 419-454.

[9] Liu, J., Zeng, T., Lu, X. 2019. Challenges in multi-language pronunciation teaching: A cross-linguistic study of Chinese students' perception of voiced and voiceless stops. *Círculo de Lingüística Aplicada a la Comunicación,* 79. 99-118.

[10] Schaefer, V., Darcy, I. 2014. Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology,* 5(4), 489–522.

[11] Takada, M., Kong, E.J., Yoneyama, K., Beckman, M.E. 2015. Loss of prevoicing in modern Japanese /g, d, b/. *Proc. 18th ICPhS* Glasgow.

[12] Xu, Y. 2013. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France. 7-10.

PHONETIC REALIZATION OF MULTIPLE STRESS LEVELS BY SPEAKERS OF A NON-STRESS LANGUAGE: A CASE OF JAPANESE-ACCENTED ENGLISH

Keiichi Tajima[1], Mafuyu Kitahara[2], and Kiyoko Yoneyama[3]

[1]Hosei University, [2]Sophia University, [3]Daito Bunka University
tajima@hosei.ac.jp, mafuyu@sophia.ac.jp, yoneyama@ic.daito.ac.jp

ABSTRACT

The present study investigated how well Japanese learners of English (JE) can differentiate stressed and unstressed syllables in English, by analyzing ten word pairs and five word triplets that exhibit stress shift, e.g., *compúter-còmputátion*, produced by high-proficiency (JEH) and mid-proficiency (JEM) learners and native American English speakers (AE). Results were as follows. First, JEH and JEM frequently produced devoiced vowels in unstressed syllables, e.g., *còmp[ɟʊ]tátion*, approximately in the same phonetic environments where vowels are devoiced in Japanese. Second, vowel duration and first-formant frequencies of primary-stressed, secondary-stressed, and unstressed syllables differed significantly across AE, JEH, and JEM. In particular, the acoustic differences between secondary-stressed and unstressed syllables were smaller for JE than for AE, but those between primary- and secondary-stressed syllables were larger for JE than for AE. Altogether, JE diverge from AE in their phonetic realization of English stress, showing native-language influences and difficulties with more than two stress levels.

Keywords: vowel reduction, vowel devoicing, vowel duration, vowel quality, stressed/unstressed syllables

1. INTRODUCTION

Speakers of a non-stress language such as Japanese often have difficulty producing distinctions in lexical stress in languages such as English. In particular, they may struggle with the phonetic realization of unstressed vowels, for instance, when producing the first vowel in the word *compúter*. Past studies that investigated the duration, pitch, intensity, and formant frequencies of weak syllables produced by second-language (L2) learners revealed that vowel reduction was highly correlated with English proficiency [1]. However, vowel quality adjustment was found to be difficult even for highly proficient Japanese learners of English, often reflecting a strong influence of orthography [2-4].

While theories vary in the treatment of English stress, English is commonly described as having two levels of stress: primary and secondary. Previous studies have focused on primary (1-stress) and unstressed (0-stress) vowels, but much less is known about how L2 learners produce secondary-stressed (2-stress) vowels.

Several preliminary studies have investigated how Japanese learners of English (JE) produce syllables with different degrees of stress by analyzing their production of English words that exhibit stress shift, e.g., *phótogràph-photógraphy* [5-7]. Results showed that JE often produced 0-stress vowels as devoiced vowels [5-7], approximately in the same phonetic environment in which short high vowels are devoiced in Japanese [8]. Furthermore, compared to 1-stress vowels, 0-stress vowels produced by JE were shorter in duration, but not lower in first-formant frequencies (F1), which would indicate a smaller jaw opening [6]. Finally, duration and F1 differed between 2- and 0-stress for vowels produced by native American English speakers (AE), but not by JE [6], suggesting that JE may not reliably distinguish between 2- and 0-stress vowels. Comparison between 1- and 2-stress vowels led to mixed results due to data limitations [7].

The purpose of the present study is to systematically compare the phonetic realization of three levels of stress, 1-, 2-, and 0-stress, produced by JE and AE, through acoustic analysis of 15 sets of English words that exhibit stress shift. Three hypotheses are examined. (1) JE do not differentiate 2- and 0-stress vowels, while AE do. (2) JE reduce 0-stress vowels more strongly in terms of duration than F1. (3) High-proficiency JE show characteristics more similar to AE than do mid-proficiency JE.

2. METHODS

2.1. Participants

Sixteen native Japanese speakers (JE) were recruited (5 males, 11 females). They were students at universities in the Tokyo area who had no experience of living in English speaking countries for more than 3 months. Their mean age was 21.5 years (SD: 1.55). JE were divided into two groups of eight participants each, high-proficiency (JEH) and mid-proficiency (JEM) learners, based on TOEIC scores. The average scores for JEH and JEM were 823 (SD: 64.6) and 531

(SD: 65.6), respectively. Eight native speakers of General American English (AE) living in the US were also recruited as a control group (4 males, 4 females). Their average age was 23.6 years (SD: 3.07). All participants had no history of hearing or speaking problems.

2.2. Material

Thirty-five derivationally related English words with alternating stress positions were chosen for the material. For example, *còmpútátion-compúter* are a pair of words whose stress patterns are [2-0-1-0] and [0-1-0], respectively. This pair contains a 2-0 contrast in the first syllable and a 0-1 contrast in the second syllable. In this design, target vowels shared the same context; that is, consonants before and after the target vowel were consistent across pairs.

In addition to 10 such pairs of words, 5 triplets were included in the material. For example, *demócracy-démocràt-dèmocrátic* are a triplet of words whose stress patterns are [0-1-0-0], [1-0-2], and [2-0-1-0], respectively. Focusing on 1-0 contrast, there is one in the first syllable (*démocràt* vs. *demócracy*) as well as in the second syllable (*demócracy* vs. *démocràt-dèmocrátic*). One 2-0 contrast (*dèmocrátic* vs. *demócracy*) and two 1-2 contrasts (*démocràt* vs. *dèmocrátic* and *dèmocrátic* vs. *démocràt*) appear in this triplet as well.

2.3. Procedure

Participants were instructed to read target words in isolation presented on a computer screen. They read aloud the 35 words along with other filler words in one pseudo-random order and then in another order. The order of lists was counter-balanced across participants. Trained phoneticians checked their pronunciation on-site and asked for repetition when they made errors. For JE participants, phonetic symbols in IPA and stress positions were provided in addition to orthography.

For JE speakers, recordings were done in a sound-proof room. For AE speakers, recordings were done either in a quiet room or a sound-proof room. All the recordings were digitized at 44.1 kHz and 16 bits.

2.4. Analysis

Acoustic analyses were conducted using Praat [9]. Three acoustic properties were measured: (a) presence vs. absence of vowels, where a vowel was defined as present if it possessed two or more glottal cycles and had a clear vowel-like formant structure, (b) vowel duration (neighboring liquids which were hard to segment from the vowel were included in the vowel duration), and (c) first formant frequency (F1) as an indicator of jaw opening either at the center of the vowel, or at the 1/3 point when an inseparable liquid came after the vowel, or at the 2/3 point when it came before the vowel.

Statistical analyses were conducted on the *z*-score of either duration or F1 as a dependent variable and stress and speaker groups as independent variables. Linear mixed effects model implemented in the lme4 package [10] on R [11] was used. Speakers' intercept was the only random effect in the model.

3. RESULTS

3.1. Devoicing of unstressed vowels

As witnessed in previous studies, JE produced some words with devoiced vowels, i.e., vowels that did not meet the criterion in (a) above. Table 1 shows cases where "devoicing rate" was 25% or greater for one or more speaker groups. "Devoicing rate" is the percentage of productions among all tokens in which a particular vowel was devoiced.

Word	AE	JEH	JEM
còmpu̲tátion	0.0	62.5	50.0
cónti̲nent	12.5	12.5	37.5
cònti̲néntal	25.0	0.0	0.0
di̲plomátic	0.0	0.0	25.0
hóspi̲tal	0.0	37.5	37.5
hòspi̲tálity	0.0	25.0	37.5
imàgi̲nátion	0.0	25.0	42.9
nátio̲nal	0.0	50.0	25.0
nàtio̲nálity	12.5	50.0	14.3
orìgi̲nálity	0.0	37.5	12.5
phò̲tográphic	0.0	25.0	25.0
pho̲tógraphy	12.5	25.0	12.5
políti̲cal	0.0	37.5	37.5

Table 1: Vowels whose "devoicing rates" were 25% or greater for one or more speaker groups. Target vowels are underlined.

Table 1 shows that devoicing rate for JE was relatively high (37.5% or higher) for vowels in words such as *còmpu̲tátion*, *hóspi̲tal*, and *políti̲cal*. Note that these vowels are flanked by voiceless consonants. Such phonetic environments are similar to those in which vowels undergo devoicing in Japanese. Table 1 further shows that devoiced vowels also appeared in contexts where the target vowel is not surrounded on both sides by voiceless consonants, e.g., *cónti̲nent*, *imàgi̲nátion*, and *nátio̲nal*. Moreover, devoiced vowels occurred occasionally in 2-stress vowels, e.g., *di̲plomátic* and *phò̲tográphic*.

Devoicing rate for AE was not zero (12.5–25.0%) for some words. Rather than being "devoiced" per se, these vowels were likely deleted or reduced, as often seen in casual speech [12]. Such instances occurred

	0-1 contrast		0-2 contrast		1-2 contrast	
	Est.	*t*-value	Est.	*t*-value	Est.	*t*-value
			--- duration ---			
Intercept	-0.570	-4.324***	-0.657	-4.308***	0.462	2.513*
JEH	0.054	0.291	0.305	1.419	0.283	1.089
JEM	0.186	0.992	0.493	2.289*	0.134	0.513
stress	1.393	17.049***	0.951	7.301***	-0.651	-5.217***
JEH:stress	-0.024	0.204	-0.674	-3.660***	-0.590	-3.326***
JEM:stress	-0.370	-3.149***	-0.924	-4.974***	-0.462	-2.543*
			--- F1 ---			
Intercept	-0.458	-2.378*	-0.312	-1.455	-0.334	1.435
JEH	0.101	0.369	-0.105	-.347	-0.036	-0.110
JEM	0.352	1.287	0.007	0.024	0.005	0.002
stress	0.786	9.325***	0.876	7.974***	0.188	-1.700
JEH:stress	-0.045	-0.377	-0.611	-3.935***	-0.405	-2.593**
JEM:stress	-0.150	-1.237	-0.754	-4.793***	-0.408	-2.537*

*$p < .05$ **$p < .01$ ***$p < .001$

Table 2: Linear mixed effects model results for duration (top) and F1 (bottom) for 0-1 contrasts, 0-2 contrasts, and 1-2 contrasts. "Est" = parameter estimate. For speaker group, AE was the baseline condition.

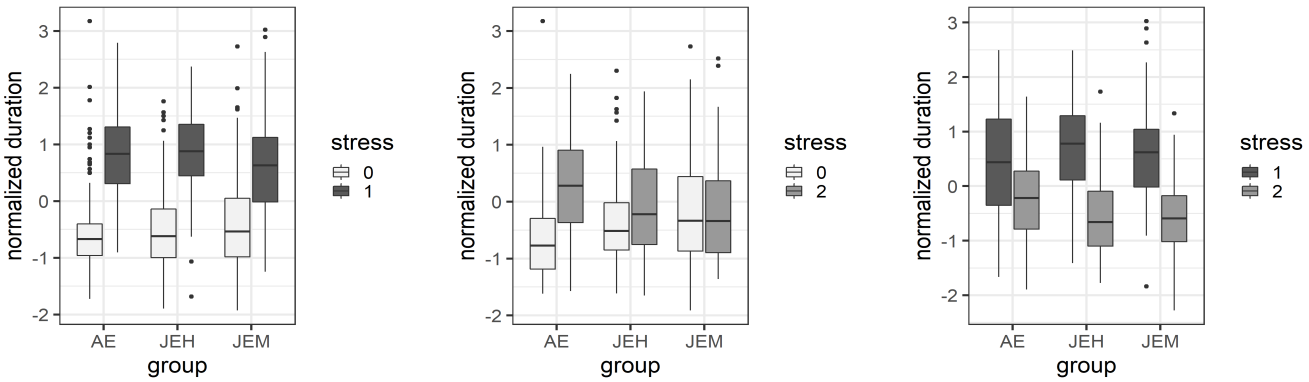


Figure 1: Boxplots of normalized vowel duration as a function of speaker group and stress level. Left: comparison between unstressed and primary-stressed vowels. Middle: comparison between unstressed and secondary-stressed vowels. Right: comparison between primary- and secondary-stressed vowels.

in fewer words and were less frequent in AE than JE. In fact, most words that exhibited vowel devoicing by JE were never produced with devoiced vowels by AE.

3.2. Vowel duration

To examine the effect of stress on vowel duration, vowel duration was compared for vowel pairs with different contrasts. Vowels in word-final syllables were excluded from duration comparisons because their duration may be adversely affected by final lengthening. A total of 17 vowel pairs were compared for 0-1 contrasts. Since the vowels vary in vowel identity and phonetic environment across different word pairs, vowel durations were normalized (*z*-transformed) for each pair and speaker group, and then aggregated for analysis. In addition to 0-1 contrasts, 10 vowel pairs each were compared for 0-2 and 1-2 contrasts.

Figure 1 shows boxplots of normalized vowel duration as a function of speaker group and stress level. The left panel illustrates results for 0-1 contrasts. Statistical analysis in Table 2 indicate that 0-stress vowels were significantly shorter than 1-stress vowels for all speaker groups, but that the magnitude of the distinction was smaller for JEM than for JEH and AE.

The middle panel of Figure 1 shows results for 0-2 contrasts. In addition to Table 2, individual comparisons within each group indicated that 2-stress vowels were significantly longer than 0-stress vowels for AE ($p < .001$) and JEH ($p < .05$), but not for JEM. These results indicate that AE produced systematically longer 2-stress vowels than 0-stress vowels, whereas JEM did not make such distinctions. JEH showed intermediate results.

Finally, the right panel of Figure 1 shows results for 1-2 contrasts. Individual comparisons per group indicated that 1-stress vowels were significantly longer than 2-stress vowels for all groups ($p < .001$).

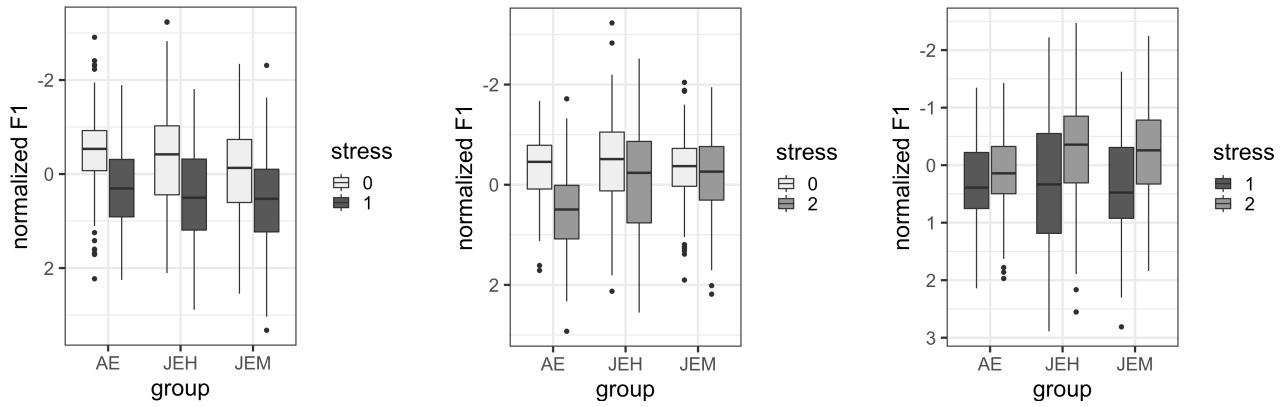


Figure 2: Boxplots of normalized F1 as a function of speaker group and stress level. The vertical axis is reversed so that downward direction indicates greater jaw opening. The left, middle, and right panels show the same pairwise comparisons of stress levels as Figure 1.

Together with Table 2, these results indicate that the duration difference between 1- and 2-stress syllables is greater for JEH and JEM than for AE.

3.3. Vowel formant frequencies

To examine the effect of stress on vowel quality, F1 was analyzed including vowels in word-final syllables, for a total of 22 pairs for 0-1 contrasts, 11 pairs for 0-2 contrasts, and 13 pairs for 1-2 contrasts.

Figure 2 shows boxplots of *z*-transformed F1 as a function of speaker group and stress level. The left panel shows results for 0-1 contrasts. Statistical analyses in Table 2 shows that 0-stress vowels were produced with significantly less jaw opening than 1-stress vowels for all speaker groups.

The middle panel of Figure 2 illustrates results for 0-2 contrasts. Results of statistical analysis matched those of duration. Individual comparisons for each group also indicated that 2-stress vowels were significantly more open than 0-stress vowels for AE ($p < .001$) and JEH ($p < .05$), but not for JEM. These results demonstrate that AE produced more open 2-stress vowels than 0-stress vowels, whereas JEM did not make such distinctions. JEH showed intermediate results.

The right panel of Figure 2 illustrates results for 1-2 contrasts. Individual comparisons per group showed some differences from those of duration: AE showed no significant difference in F1 between 1- and 2-stress vowels, while both JE groups produced significantly more open vowels for 1- than 2-stress vowels ($p < .001$). These results suggest that AE do not differentiate 1- and 2-stress vowels in terms of F1, but JE realize a systematic difference in F1.

4. DISCUSSION AND CONCLUSION

The present paper investigated acoustic correlates of different stress levels in English as produced by JE and AE. Results indicated that JE differentiated 1- and 0-stress vowels as well as AE did for the most part, in general agreement with past studies [1-4]. As for 2-stress vowels, AE clearly distinguished them from 0-stress vowels, and did not distinguish them from 1-stress vowels in F1. In contrast, JE showed a much smaller distinction, if at all, between 2- and 0-stress vowels in both duration and F1. Vowel devoicing was also observed in both 0-stress and 2-stress vowels. Together, these results support hypothesis (1), and suggest that JE may not be fully aware of the contrast between 2- and 0-stress vowels.

Analysis of vowel duration and F1 in sections 3.2 and 3.3 indicated many similarities and few differences between these two acoustic parameters. JE showed systematic differences in both duration and F1 for 0-1 and 1-2 contrasts, but not for 0-2 contrasts. These results do not provide strong support for hypothesis (2), nor do they agree with studies that reported greater difficulty with vowel quality than duration for L2 learners of English [1-7]. Since only normalized F1 values were analyzed in the present study, it remains to be seen whether inclusion of other parameters, e.g. F2, would lead to alternative results.

Finally, some differences were found between the two JE groups, such as a larger 0-1 contrast in duration for JEH than JEM, and a somewhat larger 2-0 contrast in duration and F1 for JEH than JEM. These results support hypothesis (3) to some extent, and suggest that increased English proficiency, as measured by TOEIC scores, may be correlated with improved production of English stress by JE.

In conclusion, the present study suggests that JE can generally realize the distinction between 1- and 0-stress vowels, but they are not fully aware of 2-stress vowels, and often produce them as if they are 0-stress vowels. Insofar as such tendencies lead to intelligibility and communication problems, production of stress levels should be an integral element of L2 training and education.

5. ACKNOWLEDGMENT

This research was supported by JSPS Kakenhi Grant Numbers: 18K00549, 18K00662, 19K00585, 22K00625, 23K00489. We are grateful to Maria Arvelo, Prof. Miwako Hisagi, and Prof. Benjamin Munson for their help coordinating experiments in the US.

6. REFERENCES

[1] Lee, B., Guion, S. S., Harada, T. 2006. Analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition* 28, 487–513.

[2] Kondo, Y. 2000. Production of schwa by Japanese speakers of English: An acoustic study of shifts in coarticulatory strategies from L1 to L2. In: Broe, M. B., Pierrehumbert, J. B. (eds), *Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge University Press, 29–39.

[3] Kondo, M. 2009. Is acquisition of L2 phonemes difficult? Production of English stress by Japanese speakers. *Proc. 10th GASLA 2009*, 105–112.

[4] Yazawa, K., Ozaki, Y., Short, G., Kondo, M., Sagisaka, Y. 2015. A study of the production of unstressed vowels by Japanese speakers of English using the J-AESOP corpus. *The 18th Oriental COCOSDA /CASLRE*. DOI: 10.1109/ICSDA.2015. 7357872.

[5] Kitahara, M., Tajima, K., Yoneyama, K. 2022. Phonetic Realization of Vowel Reduction by Japanese Learners of English: A Preliminary Corpus Analysis. In Kubozono, H., Morimoto, M. (eds), *Prosody Kenkyu no Shintenkai (New Developments in Prosody Research)*, Kaitakusha, 258–278 (in Japanese).

[6] Tajima, K., Kitahara, M., Yoneyama, K. 2022. Vowel reduction by speakers of a non-stress language: A preliminary corpus analysis of Japanese-accented English. Oral presentation at *New Sounds 2022*, held at Universitat de Barcelona, Apr. 20–22, 2022.

[7] Yoneyama, K., Kitahara, M., Tajima, K. 2022. Acoustic analysis of English vowel reduction by Japanese college students. In *Proc. 2022 General Meeting of the Phonetic Society of Japan*, 109–114 (in Japanese).

[8] Kondo, M. 1997. Mechanisms of vowel devoicing in Japanese. Doctoral Thesis, University of Edinburgh.

[9] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.17, retrieved from http://www.praat.org/.

[10] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

[11] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-roject. org/.

[12] Tajima, K. and Shattuck-Hufnagel, S. 2012. Phonetic characteristics of syllable reduction and enhancement in American English. *Journal of the Acoustical Society of America.* 132, 2005.

ACCOMMODATION TO PASSIVE EXPOSURE IN THE L2

Julika Weber, Christopher Geissler

Heinrich Heine University Düsseldorf
julika.weber@hhu.de, geisslc@hhu.de

ABSTRACT

Speakers reliably exhibit phonetic accommodation to more closely resemble their interlocutors. The aim of this study is to determine whether accommodation extends to passive exposure to media, specifically for L2 speakers. We hypothesized that L2 speakers would accommodate and produce more native-like tokens as a result of this exposure.

L2 English speakers (L1 German) recorded themselves reading a wordlist and answering open-ended questions, then watched an episode of an English-language television show before recording themselves again. Recordings were analyzed for phonetic features characteristic of phonetic transfer by German learners of English.

Results show that speakers consistently exhibited more native-like formant values of the TRAP vowel after exposure, consistent with the hypothesis. However, phonetic correlates of word-final consonant voicing were not consistently affected. These findings suggest that some L1 transfer effects are attenuated after passive exposure to native speech, with implications for phonetic representation and L2 acquisition.

Keywords: accommodation, L2 acquisition

1. INTRODUCTION

1.1. Approaches to accommodation

Generally, accommodation effects have been mainly investigated in a native speaker setting and results have either been interpreted in a sociolinguistic framework or with a psycholinguistic approach. The former is based on the Communication Accommodation Theory [1, 2] that characterizes assimilation as a sociolinguistic process where speakers highlight common ground (convergence) or distance (divergence) themselves from interlocutors, often linked to issues like social identity. In contrast, the "interactive alignment" account suggests that assimilation is instead automatic in nature and mainly a result of priming [3]. Another approach to explain frequency, recency and social effects of accommodation is Exemplar Theory, in which utterances are stored with phonetic and social information, ready to be selected and reproduced [4].

1.2. L2 accommodation

Most previous studies on accommodation effects in L2 speech have focused on an interactive type of exposure, e.g. in the form of a conversation partner or map tasks [5, 6, 7, 8, 9, 10, 11]. However, similar results have been found even in cases of passive media exposure [12], which [13] discusses with reference to exemplar models. Such models can account for passive exposure effects since exemplars can be stored and made available for a faithful reproduction irrespective of their source. This type of accommodation is of particular interest for L2 learners, as travel and in-person access to native speakers are not required to consume music, television, films, and newer media.

1.3. German and English Phonology

Linguistically, there are two phonological differences between the languages that we focus on in this study and that are likely to be transferred from L1 German to L2 English. First, one highly regular phonological process in German is the devoicing of voiced word-final obstruents [14]. English, on the other hand, not only preserves the voiced-voiceless contrast lexically but also phonologically. Second, German lacks the English phoneme /æ/ (the TRAP vowel) in its inventory and speakers are likely to substitute a close native equivalent, like /ɛ/ instead

2. METHOD

2.1. Participants

Nine native German speakers, four women and five men, between the age of 23 and 29 participated in the study. All speakers stated that they consume media in English regularly and all except one studied or had studied English literature and linguistics at university.

2.2. Material and procedure

Speakers conducted the experiment on their own at home following written instructions that were given beforehand in English. The experiment consisted of three parts: in the first participants had to introduce themselves and read a word list, in the second part speakers had to watch episode nine of the second season of the TV show *Friends*, the third part was again a recording of the same wordlist as in part one and questions about the plot of the episode. The wordlist consisted of 100 words, of which 60 words ending in voiced plosives: 20 in /g/, 20 in /b/ and 20 in /d/. A total of 29 items contained the the TRAP vowel /æ/ followed by a range of consonants [1] .

The questions, especially for after exposure, were designed to elicit certain items from the episode like <cab> and <dad>, containing target sounds.

2.3. Data analysis

Recordings were manually transcribed, then force-aligned using the Montreal Forced aligner [15] as implemented by DARLA [16].

Measurements were taken using a Praat [17] script based on [18] as well as [19]. For vowel measurements, these included recording the first and second formants from the middle third of each vowel; for word-final stops, these included duration of the preceding vowel, closure duration, and burst duration. Additionally, word-final stops were manually annotated for the presence of voicing in the closure. The selection of these phonetic correlates of voicing were based on [20] and [21].

3. RESULTS

3.1. TRAP vowel

Results confirmed the predictions made for the TRAP vowel: participants exhibited significantly higher F1 after exposure (with no difference for the ninth). As shown in Fig. 1, participants also produced higher F1 in the wordlist reading task as compared to the free-response questions, which is consistent with more careful speech in reading. The higher F1 corresponds to a lower vowel, more distinct from the neighboring DRESS vowel.

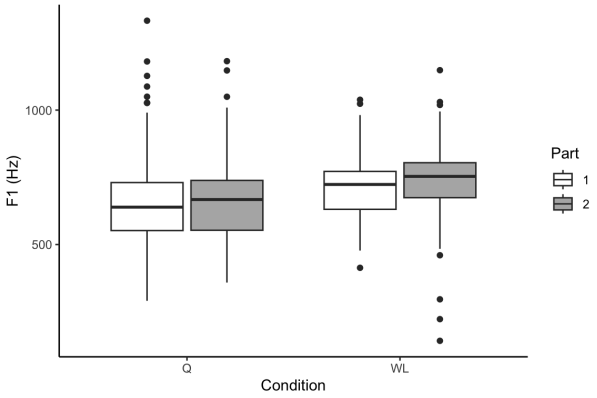


Figure 1: F1 of TRAP vowel by condition (Question-answer, Word List) before and after exposure

In contrast, no effect was observed for F2, which is unsurprising given that both TRAP and DRESS are generally similar front vowels. This can be seen in Fig. 2.

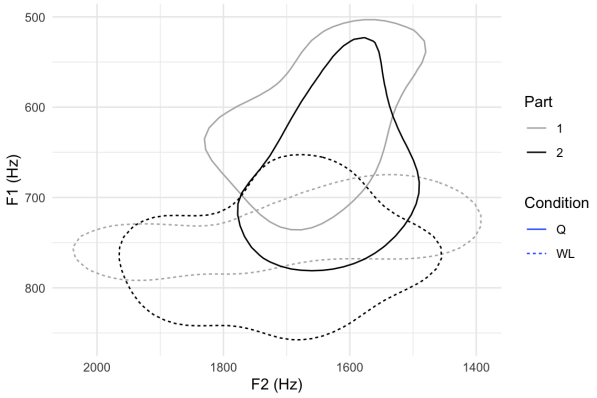


Figure 2: F1/F2 plot of TRAP, indicating pre-exposure values in grey and post-exposure values in black. Lines contain the middle 50% of tokens in each Part*Condition group

To test the significance of the observed differences of these observations, we fit linear mixed-effects models to z-scored F1 and F2 data. A baseline model included a random effects of lexical item and fixed effects of speaker, condition (wordlist/questions), and speech rate. This was compared to a second model with an additional effect of the exposure ("Part"), which significantly improved fit over the baseline model. Additional models that included interactions between Part and Condition or Part and Speaker did not further improve fit. The model comparison using ANOVA is summarized in Table 1.

Model	AIC	BIC	logLik
F1∼Base	2666.7	2732.6	-1320.3
F1∼Base+Part	2661.0	2731.9	-1316.5
F2∼Base	2841.9	2907.8	-1408.0
F2∼Base+Part	2842.8	2913.8	-1407.4

Table 1: Comparison of models predicting TRAP F1 and F2, with and without an effect of Part (before/after exposure)

3.2. Consonant voicing measures

3.2.1. Word-final burst and closure durations

We predicted that speakers would produce word-final stops with shorter closure and burst durations after exposure. Results show that burst durations did not change after exposure, but that there was closure durations were affected. Specifically, closure durations slightly decreased in the wordlist-reading task, but *increased* in the question-answer condition (free speech). This is shown in Fig. 3.

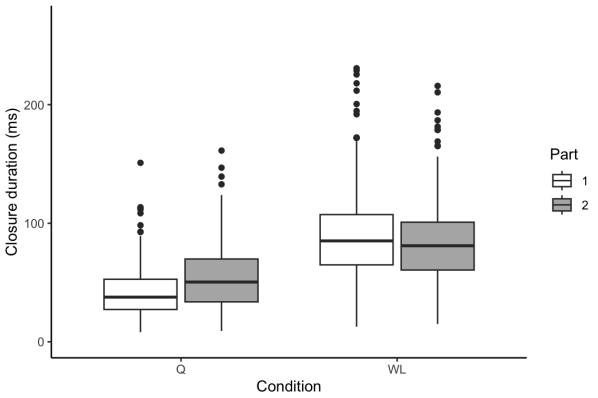


Figure 3: Closure duration of word-final stops by condition (Question-answer, Word List) before and after exposure

This result was substantiated through comparison of linear mixed-effects models with ANOVA, presented in Table 2. A baseline model included a random effects of lexical item and fixed effects of participant, Condition (wordlist/questions), and speech rate. This was compared to a second model with an additional effect of the exposure ("Part"), and a third model with an interaction of Part and Condition. This interaction effect was not expected, and we interpret this at least partially as a task effect.

3.2.2. Voicing in closure

There was overall no significant change in the rate at which speakers produced word-final /b d g/ tokens

Model	AIC	BIC	logLik
Burst duration			
BD∼Base	3222.5	3298.2	-1596.3
BD∼Base+Part	3222.9	3302.6	-1595.0
BD∼Base*Part	3223.4	3309.2	-1594.7
Closure duration			
CD∼Base	2588.9	2664.5	-1279.4
CD∼Base+Part	2584.4	2665.1	-1276.2
CD∼Base*Part	2580.6	2666.3	-1273.3

Table 2: Comparison of models predicting burst and closure duration in word-final voiced stops.)

with voicing during the closure (74.0% before, 77.3% after, $\chi^2(1, N = 1105) = 1.57$, $p = .21$). Responses varied by speaker, as shown in Fig. 4, but no consistent effect was observed.

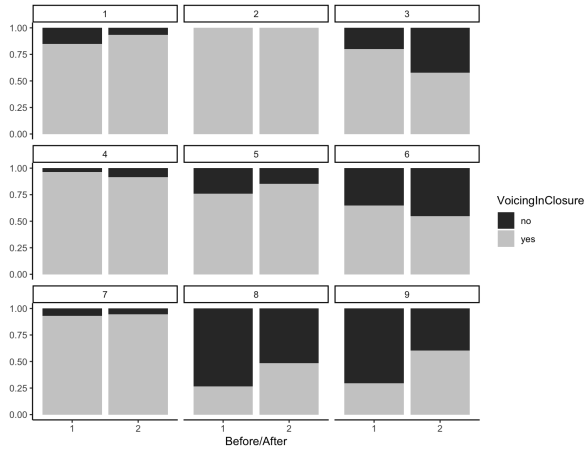


Figure 4: Voicing in closure before and after exposure, by speaker.

3.2.3. Duration of preceding vowel

We predicted that participants would produce longer vowels preceding word-final phonologically-voiced stops after exposure. Results appear to show that preceding vowels were longer after exposure in the question-response condition, but not in the wordlist-reading task. This is shown in Fig. 5.

However, this effect is not supported by a comparison of linear mixed-effects models with ANOVA, presented in Table 3. A baseline model included a random effects of lexical item and fixed effects of participant, Condition (wordlist/questions), and speech rate. This was compared to a second model with an additional effect of the exposure ("Part"), and a third model with an interaction of Part and Condition. Despite the increased AIC and BIC in the interaction

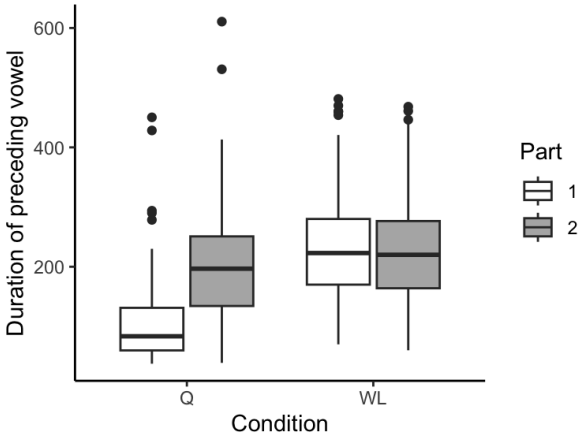


Figure 5: Duration of vowels preceding word-final voiced stops, by condition (Question-answer, Word List) before and after exposure

model, we still find the differences in means to be noteworthy. The lack of improvement in the interaction model may result from the fact that lexical items were not not often repeated in the question-answer condition, particularly in Part 1.

Model	AIC	BIC	logLik
Vdur∼Base	1475.6	1636.3	-703.81
Vdur∼Base+Part	1475.0	1640.3	-702.48
Vdur∼Base*Part	1475.2	1645.3	-701.59

Table 3: Comparison of models predicting duration of vowels preceding word-final voiced stops)

4. CONCLUSION

This study has investigated the effect of passive exposure on non-native phonetic performance with L1 German/ L2 English speakers. The exposure did not impact variables to the same degree. Our first hypothesis was confirmed and speakers produced more native like /æ/ tokens but results for the second hypothesis regarding final devoicing were mixed and further influenced by the type of condition. After exposure, preceding vowel duration was not increased in the wordlist, and may not have increased in the question-answer condition. The lack of increase in the wordlist may also be explained by a ceiling effect since realizations in the wordlist condition were longer–more native-like–from the beginning. Preceding vowel durations over 0.2 seconds are associated with voicing, so there was little to no accommodation possible [20, 22]. The observation that participants produced more native-like tokens from the beginning only in the wordlist could reflect the more formal nature of the condition and increased attention to speech, usually leading to the production of more prestigious forms [23]. The lack of uniformity in the results indicates that there are not only frequency and recency effects at play as exemplar models would suggest, but that such effects vary by the type of feature.

A variety of factors influence how difficult it is to acquire features in the L2 [24], and this study extends such findings to accommodation. Shifting realization of the TRAP vowel is an example of a relatively simple case, being a change in one phonetic parameter (F1). In contrast, the final plosive voicing involves a combination of phonetic cues, a neutralizing phonological process, and typological markedness [25]. The present study provides evidence for structural constraints in short-term passive accommodation mirroring those in long-term L2 acquisition.

5. REFERENCES

[1] H. Giles, "Accent mobility: A model and some data," *Anthropological linguistics*, pp. 87–105, 1973.

[2] H. Giles, N. Coupland, and I. Coupland, "Accommodation theory: Communication, context, and consequence," in *Contexts of accommodation: Developments in applied sociolinguistics*, H. Giles, J. Coupland, and N. Coupland, Eds. Cambridge University Press, 1991, vol. 1.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 02, Apr. 2004.

[4] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Typological Studies in Language*, J. L. Bybee and P. J. Hopper, Eds. Amsterdam: John Benjamins, 2001, vol. 45, p. 137.

[5] L. Burin, "Accommodation of L2 Speech in a Repetition Task: Exploring Paralinguistic Imitation," *Research in Language*, vol. 16, no. 4, pp. 377–406, Dec. 2018.

[6] M. A. Wagner, M. Broersma, J. M. McQueen, S. Dhaene, and K. LemhÃ¶fer, "Phonetic convergence to non-native speech: Acoustic and perceptual evidence," *Journal of Phonetics*, vol. 88, p. 101076, Sep. 2021.

[7] J. Hwang, S. E. Brennan, and M. K. Huffman, "Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design," *Journal of Memory and Language*, vol. 81, pp. 72–90, May 2015.

[8] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," *Laboratory Phonology*, vol. 2, no. 1, Jan. 2011.

[9] C. Ulbrich, "Phonetic Accommodation on the Segmental and the Suprasegmental Level of Speech in Native–Non-Native Collaborative Tasks," *Language and Speech*, Oct. 2021.
[10] N. Lewandowski and M. Jilka, "Phonetic Convergence, Language Talent, Personality and Attention," *Frontiers in Communication*, vol. 4, p. 18, May 2019.
[11] N. Lewandowski, "Sociolinguistic factors in language proficiency: Phonetic convergence as a signature of pronunciation talent," in *Language Talent and Brain Activity*, ser. Trends in Appied Linguistics, G. Dogil and M. S. Reiterer, Eds. Walter de Gruyter Berlin, 2009, vol. 1.
[12] A. E. Hutchinson and O. Dmitrieva, "Exposure to speech via foreign film and its effects on non-native vowel production and perception," *Journal of Phonetics*, vol. 95, 2022.
[13] T. MatanoviÄ, "Sociophonetics and Exemplar Theory: A study of second-language accent modification through exposure," Ph.D. dissertation, Heinrich Heine University DÃ¼sseldorf, DÃ¼sseldorf, Germany, 2020.
[14] E. KÃ¶nig and V. Gast, *Understanding English German Contrasts*, 3rd ed. Berlin: Schmidt, 2012.
[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.
[16] S. Reddy and J. Stanford, "A Web Application for Automated Dialect Analysis," in *Proceedings of NAACL-HLT 2015.*, 2015.
[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer software]. Version 6.0. 43," 2018.
[18] C. DiCanio, "Pitch Dynamics Script for Praat, version 6," 2012. [Online]. Available: https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_duration_2.0.praat
[19] ——, "Vowel Acoustics Script for corpus data analysis, v2.0," 2013. [Online]. Available: https://www.acsu.buffalo.edu/~cdicanio/scripts/Vowel_Acoustics_for_corpus_data_2.praat
[20] B. L. Smith, R. Hayes-Harb, M. Bruss, and A. Harker, "Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German," *Journal of Phonetics*, vol. 37, no. 3, pp. 257–275, Jul. 2009.
[21] T. Roettger, B. Winter, S. Grawunder, J. Kirby, and M. Grice, "Assessing incomplete neutralization of final devoicing in German," *Journal of Phonetics*, vol. 43, pp. 11–25, Mar. 2014.
[22] K. R. Kluender, R. L. Diehl, and B. A. Wright, "Vowel-length differences before voiced and voiceless consonants: an auditory explanation," *Journal of Phonetics*, vol. 16, no. 2, pp. 153–169, Apr. 1988.
[23] W. Labov, "The social stratification of English in New York City, 3rd printing," *Washington: Center for Applied Linguistics*, 1982.
[24] R. C. Major, *Foreign accent: The ontogeny and phylogeny of second language phonology*. Routledge, 2001.
[25] F. R. Eckman, "Typological markedness and second language phonology," in *Phonology and second language acquisition*, J. G. H. Edwards and M. L. Zampini, Eds. John Benjamins, 2008, vol. 36, p. 95.

[1] Stimuli and other materials are available online at https://osf.io/r2nca/?view_only=d1363fcfd2df4660afd29323305811c3

LEARNING TO DISTINGUISH THE THREE-WAY LARYNGEAL CONTRAST OF KOREAN PLOSIVES BY NATIVE ENGLISH SPEAKERS

Tyler K. Perrachione[1,2], John D.E. Gabrieli[2], & Amy S. Finn[2,3]

[1]Department of Speech, Language, and Hearing Sciences, Boston University, USA
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA
[3]Department of Psychology, University of Toronto, Canada
tkp@bu.edu

ABSTRACT

Second-language acquisition often entails learning to use unfamiliar speech sounds to distinguish words, which may require learning to perceive an unfamiliar phonetic feature or accommodate a new category boundary along a known phonetic continuum. Here, native English speakers (*N*=37) learned an 18-word vocabulary comprised of six minimal triplets based on the Korean three-way plosive contrast (fortis, lenis, aspirated), defined by a trading relation between two phonetic features (onset pitch and voice onset time) that conflicts with how these cues structure listeners' native categories. Variation in these features across talkers and place of articulation further obfuscates this contrast. Learning outcomes were highly variable: Mixture model analysis suggested two groups of learners who differed primarily in whether they learned to distinguish lenis from aspirated stops. Both groups learned these contrasts best for bilabials and least accurately for alveolar stops. These results underscore the challenge of overcoming native category structure in second-language learning.

Keywords: second language acquisition, voice onset time

1. INTRODUCTION

A key challenge in acquiring a second language in adulthood is learning to perceive and produce novel phonological categories. Listeners begin to commit their perceptual and neural pathways to processing the sound structure of their native language during the first years of life [14], and thereafter it becomes increasingly challenging to learn to perceive or produce contrasts from other languages [2].

Two prominent factors constrain speech-sound learning in adulthood: *(i)* The extent to which a novel contrast's category boundary conflicts with a native boundary on the same phonetic continuum (e.g., stop categories with different voice onset time (VOT) boundaries, or vowels with different distributions in the F1×F2 space) [4], and (*ii*) the extent to which a novel contrast requires attending to a phonetic feature that is not contrastive in one's native language (e.g., F3 in the case of Japanese learners of English /r/-/l/ [13], or pitch contour direction in the case of English learners of Mandarin lexical tones [5]).

An anecdotally challenging contrast for native English speakers is the Korean three-way laryngeal contrast for onset plosives. This contrast distinguishes words like /ppul/ 뿔 "horn"; /bul/ 불 "fire"; and /pul/ 풀 "grass." In word onset, the Korean three-way stop consonant contrast can be characterized via a trading relation between three levels of positive VOT (short for fortis vs. intermediate for lenis vs. long for aspirated sounds) and onset f_0 (low for lenis vs. higher for fortis and aspirated sounds) (Fig. 1).

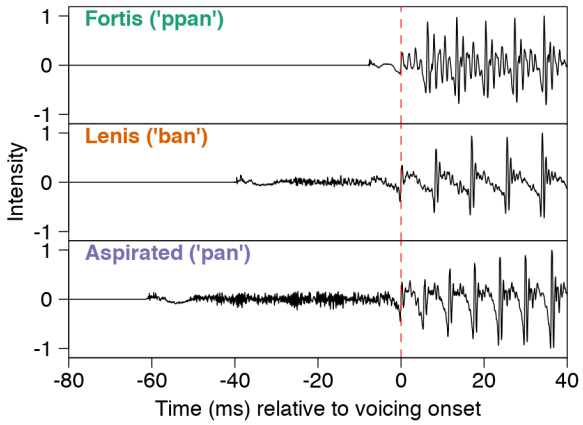


Figure 1: Acoustics of Korean plosive onset contrasts, aligned to voicing onset (dashed line). Note differences in both VOT and onset glottal period (f_0) across categories.

While naturally occurring minimal triplets are rare in Korean, there are multitudinous instances of both two-way minimal contrasts. Phonetic descriptions of these contrasts have defied simple characterization as a function of only voice onset time (VOT) or aspiration [1], and debate about the articulatory distinctions among the three contrasts [7] has led to inconsistencies in their transcription via the IPA. A straightforward articulatory or acoustic classification is further complicated by allophonic variation in the realization of these stops (particularly the lenis contrast) as a function of syllable position [16] and by variation in their realization across dialects [15]. To avoid mechanistic claims about articulation implicit in IPA transcriptions, in this report we use the standard Revised

Romanization transliterations: *fortis* /pp/, /tt/, /kk/; *lenis* /b/, /d/, /g/; and *aspirated* /p/, /t/, /k/.

Cross-linguistic work has historically focused on Korean learners of English [20], but the rapid global ascent of Korean pop culture has catalysed new interest in learning this language [24]. However, little empirical work has examined the interesting challenges that the fortis-lenis-aspirated contrast poses for English learners of Korean: First, its division of the VOT continuum into two aspirated categories conflicts with English listeners' expectation that aspiration reliably signals a voiceless (vs. voiced) word-initial stop. Second, onset stops in English (like many languages) covary in VOT and onset f_0, such that onset f_0 is low for voiced sounds like /b/ and high for voiceless sounds like /p/ [10,12]. This VOT-onset f_0 pattern is violated by Korean lenis stops, where a voiceless, slightly aspirated stop is realized with low onset f_0.

It is unclear how English learners will cope with the VOT-onset f_0 trading relation in Korean stops. Some have speculated that the fortis-lenis contrast will be hardest for English speakers, since the intermediate VOT and low onset f_0 of lenis stops may be confusable for English voiced stops [18]. Others have suggested that the lenis-aspirated contrast will be the most challenging [19], because aspiration unambiguously signals voiceless stops in English. Furthermore, despite the relationship between VOT and onset f_0 in English stops, English listeners are notoriously poor at learning to use pitch in phonological contrasts [5,22,23]. However, the atypically low onset f_0 of the lenis stop provides a correlated cue that may instead help listeners overcome their native VOT boundary.

2. METHODS

In this study we trained native-English speakers with no prior Korean experience to use the three-way Korean plosive onset contrast to identify a vocabulary of pseudoword triplets based on those contrasts.

2.1. Participants

Native English-speaking monolingual adults ($N = 37$, 25 female, 12 male, age 18-33, mean = 23.1 years) completed this study. Participants had minimal foreign language experience: none had more than two years of study, no earlier than in college. No participant had any prior experience with Korean. This study was approved by the IRB (COUHES) at the Massachusetts Institute of Technology.

2.2. Stimuli

Training stimuli consisted of 18 monosyllabic Korean pseudowords (Table 1). These words were organized into six minimal triplets (e.g., 'ppan', 'ban', 'pan'). There were two triplets per place of articulation (bilabial, alveolar, and velar). Familiar vowel and coda consonants were chosen to reduce learning demands.

빤 ppan (cow)	똑 ttok (bell)	깻 kkaet (car)
반 ban (seashell)	독 dok (sock)	갯 gaet (parrot)
판 pan (hammer)	톡 tok (grapes)	캣 kaet (camera)
삠 ppim (bus)	뗍 tteop (brush)	꿍 kkung (chair)
빔 bim (lamp)	덥 deop (box)	궁 gung (hat)
핌 pim (desk)	텁 teop (fish)	쿵 kung (fork)

Table 1: Vocabulary of minimal triplet Korean pseudowords (and target images).

High quality sound recordings were obtained from four adult native Korean speakers (two male, two female) of the Seoul dialect. Stimuli were RMS amplitude normalized to 65 dBA. As a control, two additional native Korean speakers confirmed the naturalness and distinctiveness of these stimuli and achieved perfect accuracy on the vocabulary training.

2.3. Acoustic measurements

We measured the key phonetic features of the three-way contrast (VOT and onset f_0) for all tokens in the vocabulary to validate the contrastiveness of our talkers' phonetic categories. VOT was measured as the latency between the release burst and the onset of phonation (the point on the waveform with the first high-amplitude, periodic deflection from zero). Onset f_0 was measured over the first five glottal cycles.

Each talker produced clear distinctions among the contrasts (Fig. 1, left). However, overlap between categories was also evident due to variation in mean f_0 and idiosyncratic VOT across talkers. In addition to being a critical feature of this contrast, VOT also varies for stops as a function of place of articulation, with longer VOTs for more posterior constrictions [1,18].

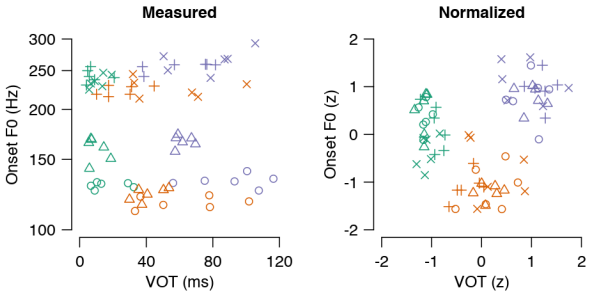


Figure 2: Phonetics of the vocabulary. Acoustics (left) show the challenge of overlapping category boundaries across talker and place of articulation (normalized at right). Symbols denote each talker, colours denote each contrast (green: fortis, orange: lenis, purple: aspirated).

A linear mixed-effects model of laryngeal contrast and place of articulation effects on VOT (by-talker random intercepts) revealed significant effects of laryngeal contrast ($F_{2,60} = 132.7$, $p < 0.0001$; aspirated > lenis > fortis) and place of articulation ($F_{2,60} = 20.8$, $p < 0.0001$; velar > labial > alveolar), but their interaction was marginal ($F_{4,60} = 2.20$, $p = 0.08$).

The corresponding model for onset f_0 showed a significant effect of laryngeal contrast ($F_{2,60} = 56.7$, $p < 0.0001$; aspirated > fortis > lenis) but not place of articulation ($F_{2,60} = 1.51$, $p = 0.23$) nor their interaction ($F_{4,60} = 0.45$, $p = 0.77$).

Normalizing VOT and onset f_0 within talker and place of articulation clearly reveals the three-way category structure in our vocabulary (Fig. 2, right).

2.4. Training procedure

Participants learned to associate each of the 18 Korean pseudowords with the photograph of a unique, familiar, natural object (Table 1). Images were shown in isolation on a white background. Participants completed four days of training, each consisting of a training and a test phase (Fig. 3). During the training phase, participants passively listened to the word-object pairings and actively practiced matching spoken words to the corresponding objects while receiving corrective feedback. During the test phase, participants actively matched spoken words to the corresponding objects without feedback [23].

Stimulus delivery was controlled using PsychoPy [21]. Participants completed each day's training and testing by themselves in a quiet room. Audio stimuli were delivered via Sennheiser HD-202 headphones at a comfortable volume selected by the participant.

During the training phase, participants completed interleaved blocks of passive familiarization and active practice. Words were trained in minimal triplets to emphasize the target contrasts. During passive familiarization, participants heard each word while the corresponding object was shown for 2000 ms (Fig. 3, left). Participants were instructed to attend to the stimuli and remember each word-object pairing. Each familiarization block contained 24 trials (3 triplet words × 4 talkers × 2 repetitions). Recordings from all four talkers were used during training, as high-variability training contributes to category learning [11], but stimuli were blocked by talker to reduce processing costs incurred by talker changes [9,22].

Participants next actively practiced matching the words in the triplet to the corresponding picture (Fig. 3, middle). While all three objects were displayed on the screen, participants heard each of the 12 recordings (3 words × 4 talkers) in a random order. Their task was to click on the corresponding picture using a mouse. Feedback was given immediately indicating whether they had chosen correctly or what the correct answer should have been. This self-paced procedure was repeated for a total of 24 active practice trials.

The interleaved familiarization-practice cycle of the training phase was repeated six times, once for each of the minimal triplets, in a random order. Thus, participants underwent 144 passive familiarization trials and 144 active practice with feedback each day.

After completing the training phase, participants' vocabulary learning was assessed in the Test Phase (Fig. 3, right). In this phase, an array of all 18 objects was shown. Participants heard all 72 recordings (18 words × 4 talkers) in a random order, selecting the corresponding object by clicking on it with a mouse. The test phase was self-paced and no feedback was given. Performance on the daily test phase was used as the dependent variable operationalizing learning.

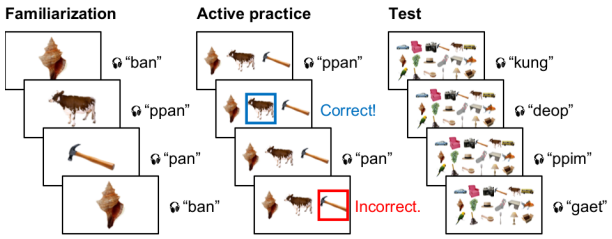


Figure 3: Vocabulary training procedure. Participants completed interleaved familiarization and active practice blocks for each triplet, followed by a test of all 18 items.

3. RESULTS

All participants improved with training, but learning outcomes on Day 4 were highly variable, with word-identification accuracy ranging from 22% to 69%. Mixture-model analysis of Day 4 accuracy indicated learning outcomes were best described by two subpopulations (Fig. 4): low-performance learners (LPL) comprised the bottom two-thirds of participants with average word learning of 37% ± 7%, while high-performance learners (HPL) comprised the top third of learners with average word learning of 63% ± 5%.

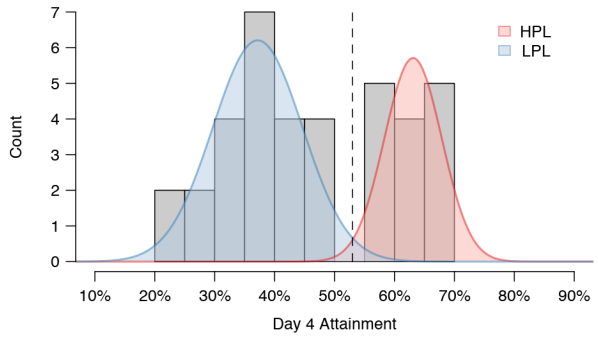


Figure 4: Mixture-model analysis revealed that learning outcomes were best captured by two distributions, one for high- and one for low-performing learners (HPL, LPL).

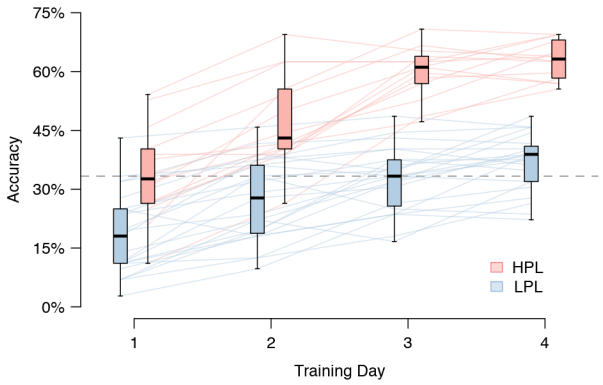


Figure 5: Learning progress by group. Lines show individual learning trajectories. Chance performance is shown for distinguishing between minimal triplets (1/3).

The HPL group learned faster than the LPL group even on Day 1, when they had already mastered the basic vocabulary triplets and were constrained by within-triplet phonetic contrast (chance level = 1/3), whereas the LPL group did not master between-triplet word-object associations until Day 4 (Fig. 5).

Patterns of word identification confusion can reveal the development of learners' novel phonological category structures. Plotting the patterns of confusion on Day 4 (Fig. 6) revealed that the HPL group had largely mastered the fortis contrast (83% ± 12% correct identification of fortis contrast words) and was making progress towards being able to use the lenis (55% ± 12%) and aspirated (51% ± 11%) contrasts to distinguish words. The LPL group, however, had not yet mastered using the fortis contrast to recognize words (50% ± 17% word identification) and was essentially at chance for identifying words distinguished by the lenis (32% ± 11%) and aspirated (30% ± 15%) contrasts. In general, both groups found the fortis stop most distinctive and had the most difficulty distinguishing lenis from aspirated onsets.

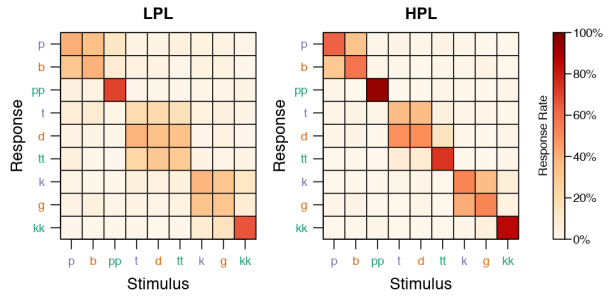


Figure 6: Onset consonant confusion matrices by group. More accurate learning of phonetic categories is captured by greater response clustering along the diagonal.

Both groups were most accurate at identifying words based on the three-way contrast involving bilabial stops (HPL: 70%, LPL: 44%), then velar stops (HPL: 63%, LPL: 44%), and were least accurate for alveolar stops (HPL: 57%, LPL: 22%).

4. DISCUSSION

Learning the Korean three-way laryngeal contrast for onset plosives is particularly challenging for native English speakers because (*i*) it divides a familiar phonetic continuum (VOT) in a way that conflicts with native voicing contrasts [1,4,18] and (*ii*) it violates the expected relationship between VOT and onset f_0 [10] in a way that requires listeners to pay attention to a perceptually challenging phonetic feature [6,17,23].

Participants varied widely in mastering the vocabulary involving these contrasts after four days of training. About one in three learners mastered word identification based on the fortis contrast (which is most similar to English /b/, /d/, and /g/) but still made mistakes identifying words based on the lenis vs. aspirated contrasts (which are both perceptually similar to English /p/, /t/, and /k/ in word onset). This pattern of learning is consistent with the prominence of aspiration in English voiceless stops [1,8], which appears to dominate perception even when onset f_0 provides a potentially disambiguating cue. Despite the VOT-onset f_0 relationship in English [10], it is not clear that listeners actually use this cue to perceive word-initial stops, and they may need to be explicitly instructed to attend to pitch in order to learn to use this feature [6].

Interestingly, the LPL group did not master even the fortis stop, despite its similarity to English /b/. It may be that partial overlap in the VOT distributions of the fortis and lenis stops across talkers led listeners to discount the contribution of this otherwise reliable cue, yielding a haphazard pattern of responses.

Also unexpected is learners' difficulty with the alveolar stops. On Day 4, even the HPL group was confusing fortis /tt/ for aspirated /t/ nearly as often as lenis /d/. This may be due to the unfamiliar pattern of VOT-by-place of articulation in this vocabulary: Whereas English onset stops tend to have increasing VOT with more posterior place of articulation, here alveolar stops' VOT was shorter than bilabial stops (lenis: /d/ 32 ms vs. /b/ 41 ms; aspirated: /t/ 59 ms vs. /p/ 74 ms). Combined with an unfamiliar amount of within-talker variation in VOT (resulting from lenis vs. aspirated contrast), these unexpected patterns of VOT may have upended English-speaking listeners' ability to generalize systematicity in VOT across talkers [8], further undermining their learning.

Overall, these results provide new empirical data to reaffirm that a core tenet of second-language speech-sound learning holds true even when a secondary cue is available to reinforce learning: The more that a novel phonological contrast conflicts with native categories, the harder it will be to learn [4].

5. REFERENCES

[1] Abramson, A.S. & Whalen, D.H. 2017. Voice onset time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *J. Phonetics* 63, 75-86.

[2] Baese-Berk, M., Chandrasekaran, B., & Roark, C.L. 2022. The nature of non-native speech sound representations. *J. Acoust. Soc. Am.* 152, 3025-3034.

[3] Barcroft, J. & Sommers, M.S. 2005. Effects of acoustic variability on second language vocabulary learning. *Stud. Second Lang. Acq.* 27, 387-414.

[4] Best, C.T. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words* 167, 233-277.

[5] Chandrasekaran, B., Sampath, P.D., & Wong, P.C.M. 2010. Individual variability in cue-weighting and lexical tone learning. *J. Acoust. Soc. Am.* 128, 456-465.

[6] Chandrasekaran, B., Yi, H.-G., Smayada, K.E., & Maddox, W.T. 2016. Effect of explicit dimensional instruction on speech category learning. *Attention, Perception, & Psychophysics* 78, 566-582.

[7] Cho, T., Jun, S.-A., & Ladefoged, P. 2002. Acoustic and aerodynamic correlates of Korean stops and fricatives. *J. Phonetics* 30, 193-228.

[8] Chodroff, E., Godfrey, J., Khudanpur, S., & Wilson, C. 2015. Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. *18th International Congress of Phonetic Sciences* (Glasgow, August 2015).

[9] Choi, J.Y., Hu, E.R., & Perrachione, T.K. 2018. Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics* 80, 784-797.

[10] Dmitrieva, O., Llanos, F., Shultz, A.A., & Francis, A.L. 2015. Phonological status, not voice onset-time, determines the acoustic realization of onset *f*0 as a secondary voicing cue in Spanish and English. *J. Phonetics* 49, 77-95.

[11] Flege, J.E. 1995. Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguist.* 16, 425-442.

[12] Haggard, M., Ambler, S., & Callow, M. 1970. Pitch as a voicing cue. *J. Acoust. Soc. Am.* 47, 613-619.

[13] Iverson, P., Kuhl, P.K., Akahane-Yamada, R., Deisch, E., Tohkura, Y., Ketterman, A., Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47-B57.

[14] Kuhl, P.K. 2004. Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831-843.

[15] Lee, H., Politzer-Ahles, S., & Jongman, A. 2013. Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *J. Phonetics* 41, 117-132.

[16] Lee, H.B. 1993. Korean. *J. Int. Phon. Assoc.* 23, 28-31.

[17] Lim, S.-J., & Holt, L.H. 2011. Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science* 35, 1390-1405.

[18] Lisker, L. & Abramson, A.S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384-422.

[19] Oh, E. 2011. Effects of speaker gender on voice onset time in Korean stops. *J. Phonetics* 39, 59-67.

[20] Park, H. & de Jong, K.J. 2008. Perceptual category mapping between English and Korean prevocalic obstruents: Evidence from mapping effects in second language identification skills. *J. Phonetics* 36, 704-723.

[21] Peirce, J.W. 2007. PsychoPy - Psychophysics software in Python. *J. Neurosci. Meth.* 162, 8-13.

[22] Perrachione, T.K., Lee, J., Ha, L.Y.Y., & Wong, P.C.M. 2011. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J. Acoust. Soc. Am.* 130, 461-472.

[23] Wong, P.C.M. & Perrachione, T.K. 2007. Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguist.* 28, 565-585.

[24] Zabel, S. 2021. "Squid Game" could inspire a new wave of Korean language learners. Available online: https://blog.duolingo.com/squid-game-could-inspire-a-new-wave-of-korean-language-learners/

ARTICULATORY TIMING OF THE JAPANESE SINGLETON AND GEMINATE /t/ PRODUCED BY SPEAKERS OF STANDARD CHINESE

Maho Morimoto[1, 2], Ai Mizoguchi[3, 4], Li Weiyu[1], Takayuki Arai[1]

[1]Sophia University, [2]JSPS, [3]Maebashi Institute of Technology, [4]NINJAL
maho.morimoto.jp@gmail.com, aimizoguchi@maebashi-it.ac.jp, w-li-5x4@eagle.sophia.ac.jp, arai@sophia.ac.jp

ABSTRACT

Consonant length contrast is known to be problematic to acquire for L2 learners of Japanese. Previous studies pointed out that the difficulty in producing the contrast can largely be attributed to the incomplete mastery of the timing control in the target language. In this study, we conducted an ultrasound experiment to investigate the realization of the consonant length contrast by L2 learners of Japanese with an L1 Standard Chinese background.

We report the results from acoustic and articulatory analyses of the contrast between /t/ and /t:/ with special focus on the gestural duration and articulatory timing. While certain acoustic durational aspects of the contrast were successfully learned by the L2 speakers examined in this study, the difference in the tongue tip gesture duration observed in the native speakers was not present for some of the learners. We discuss the relationship between the articulatory and acoustic timing control.

Keywords: Japanese geminates, Standard Chinese, L2, articulation, ultrasound.

1. INTRODUCTION

For L2 learners of Japanese, learning the quantity contrast in the language can be problematic, especially if their L1 does not make use of such contrasts [1, 2, 3]. Previous acoustic analyses of L2 productions of Japanese singletons (short consonants, as in /kata/ 'shoulder') and geminates (long consonants, as in /kat:a/ 'win-PAST') have revealed that issues in the realization of the contrast largely arise due to the difficulty in controlling the duration of the segments, especially the constriction duration which is known to be the primary perceptual cue to the contrast [1, 4, 5]. L2 learners are often reported to underdifferentiate singletons and geminates by producing either geminates that are too short in constriction duration, or singletons that are too long in duration [4, 6].

Issues in manipulating the segment duration extend to the duration of the vowels surrounding the singletons and geminates [4, 6]. Especially, in Japanese, vowels preceding geminates (V_1) are phonetically longer than vowels preceding singletons [7, 8]. This can be a notable challenge for learners with various L1 backgrounds, as it goes against the cross-linguistic tendency whereby the nucleus of a closed syllable shortens [9].

Previous studies on the articulation of geminate stops by native speakers of Japanese have discussed the duration of the closure gesture and its timing as a possible cause for the phonetic lengthening of V_1 [8, 10]. Several studies have shown that the tongue raising gesture at the offset of V_1 is longer in duration for geminates than for singletons [11, 12, 13, 14]. In addition, the onset of the tongue raising gesture relative to the acoustic onset of the consonant closure occurs earlier in geminates than in singletons [11]. Based on these findings, it has been pointed out that slower consonantal gestures allow for a more articulated and prolonged production of V_1 [8, 10].

While various acoustic studies have investigated the timing control in the realization of the consonant length contrast by L2 learners of Japanese, how the manipulation of segmental duration is implemented articulatorily has not been addressed directly. In this study, we report on a production experiment using ultrasound in order to investigate the lingual gestures and their relation to the acoustic durations in the singleton and geminate /t/ as produced by native speakers and Chinese-speaking learners of Japanese.

2. METHODS

2.1. Participants

We recruited 10 native speakers of Standard and Mandarin Chinese (CH) and 10 native speakers of Japanese (JP). In this study, we report the results from six of these speakers (three from each group), as shown in Table 1. We include three advanced learners (N1 level in the Japanese Language Proficiency Test) identifying themselves as native speakers of Standard Chinese (BCF01 and BCF04) or Mandarin Chinese (BCF05), and three native speakers of Japanese from Tokyo (BJF03 and BJF06) or an adjacent area (BJF07). All of them are female speakers in their 20s with normal hearing and speaking ability, residing in Japan at the time of the experiment. The Standard and Mandarin Chinese speakers were chosen for two reasons. First, their L1 phonology does not allow for a non-nasal obstruent coda, thus lacking obstruent

geminates; second, Chinese speakers constitute one of the largest Japanese learner populations [15, 16]. Advanced learners were chosen to ensure that they learned the consonant length distinction to some extent.

Instructions were provided in written and spoken Japanese and Chinese as necessary, and the participants consented to the experimental procedure approved by the Sophia University Committee for Ethics. A language background questionnaire was administered prior to participation. They were compensated for their participation.

Table 1: Linguistic background of the six speakers reported in this study.

Speaker	Age	Region	JP level
BCF01	24	Shandong/Tianjin	N1
BCF04	21	Shanghai	N1
BCF05	22	Beijing	N1
BJF03	20	Tokyo	Native
BJF06	22	Tokyo	Native
BJF07	21	Kanagawa	Native

2.2. Speech materials

The speech materials consisted of 16 Japanese bisyllabic words including /t/, /k/, /s/, and /ʃ/ as the target consonant (C_2 in $C_1V_1C_2V_2$). The words were varied in terms of the length of the target consonants and lexical pitch accent, and were embedded in a carrier phrase (/korewa ___ to i:masu/ 'this is called ___'). They were manually presented on a screen one at a time in semi-randomized order, using Japanese characters (both Hiragana and Kanji). We obtained 10 repetitions for each word. For the current study, we report the results for two words, namely, /hata/ 'field (as a surname)' for singletons (S), and /hat:a/ 'crawl-PAST' for geminates (G), both of which have a falling pitch accent.

2.3. Data collection

Audio and ultrasound recordings were made simultaneously for each speaker in a soundproof room, as they read the presented phrases aloud.

The acoustic signal was digitally recorded monaurally at 22,050 Hz to a laptop computer through an audio interface (Focusrite, Scarlett Solo 2nd Gen), using a RODE-NT2-A microphone.

Real-time mid-sagittal images of the oral cavity were recorded with an ultrasound system (MicrUS, EXT-1H) using a microconvex probe (MC10-5R10S-3). The probe was placed under the participant's chin, and an UltraFit Headset [17] was used to stabilize the relative position of the probe and the participant's head. The tongue images were recorded during the utterances.

The audio and ultrasound video were recorded and synchronized using AAA software [18]. The frame rate of the obtained ultrasound videos was 113 fps. Four tokens were excluded from the analysis due to synchronization error. Altogether, 116 tokens (2 consonant length × 10 repetitions × 6 speakers, minus the 4 excluded tokens) were subjected to acoustic and articulatory analyses.

2.4. Analysis

Data analysis was implemented using R [19], based on the acoustic and articulatory landmarks obtained through the following annotation procedures.

The audio recordings were annotated using Praat [20]. The segmental boundaries were identified in the waveform and spectrogram display based on the periodic cycles of the vowels.

Ultrasound images were annotated using the GetContours program [21]. Two articulatory landmarks were identified based on the tongue tip (TT; the most anterior point of the visible tongue contour): *GestOns* (onset of the TT raising gesture, or the frame in which the TT starts to rise) and *GestMax* (onset of the TT maximal constriction, or the frame in which the TT ceases to rise). In addition, an acoustic landmark, *CloOns* (acoustic onset of the consonant closure), was imported from the Praat TextGrids. The time difference between *GestOns* and *GestMax* was calculated to be analyzed as the duration of the TT raising gesture, or *TTGest*. The duration between *GestOns* and *CloOns*, as well as the duration between *CloOns* and *GestMax* were calculated to examine the timing of the TT raising gesture.

In order to evaluate the effect of consonant length on the durational variables, we performed a linear mixed effects model analysis [22] for each speaker group, with consonant length (*hata* or *hatta*) as the fixed effect and speakers as the random effect.

3. RESULTS AND DISCUSSION

First, we confirmed through auditory impression that the learners' productions of /t:/ were perceived as geminates (see 3.3. for an acoustic analysis).

3.1. TT gestural duration

Table 2 summarizes the durational results related to the TT raising gesture. Overall, the TT raising gesture was longer for geminates than for singletons among the Japanese speakers (β=24.95, t=4.53, p<.001), in accordance with previous literature. However, this trend was not observed among the Chinese speakers (β=4.12, t=0.9, p=0.37), suggesting that the learners did not vary the duration of the TT raising gesture across consonant length. Based on the view that V_1

lengthening is a by-product of the slower TT gesture duration, these results tentatively suggest that the L2 learners in this study have not mastered the timing control for Japanese geminates, including V_1 lengthening (however, see 3.3).

In the meantime, Fig. 1 shows that the TT gesture duration in singletons and geminates vary among the learners. Specifically, BCF05 showed the same tendency as the Japanese speakers.

Table 2: Average duration in milliseconds (sd) and geminate-to-singleton ratios (GSR) for *TTGest*, *GestOns~CloOns*, and *CloOns~GestMax*, per native language (NL) and consonant length (L).

NL	L	*n*	*TTGest*	*GestOns ~CloOns*	*CloOns~ GestMax*
CH	S	30	80.21 (19.55)	54.35 (12.27)	25.85 (12.66)
	G	29	83.89 (21.66)	56.53 (16.23)	27.36 (15.69)
GSR			1.05	1.04	1.06
JP	S	29	85.41 (24.72)	66.57 (18.34)	18.85 (13.06)
	G	28	110.81 (36.66)	77.44 (26.21)	33.37 (18.64)
GSR			1.30	1.16	1.77

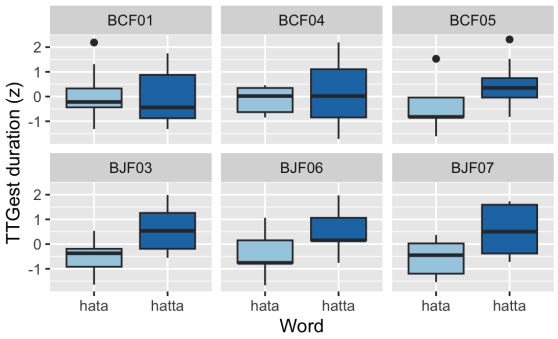


Figure 1: *TTGest* duration per word and speaker.

3.2. TT gestural timing

A similar trend is observed in the timing of the TT raising gesture relative to the acoustic onset of the closure. While the duration for *GestOns~CloOns* differed significantly for singletons and geminates among the Japanese speakers (β=10.52, t=2.34, p<.05), no such difference was observed among the Chinese speakers (β=2.2, t=0.6, p=0.55). The trend among the Japanese speakers conforms to the findings in [11], in that there was a greater overlap between V_1 and the TT raising gesture for geminates than for singletons. On the contrary, the timing with which the Chinese speakers initiated the TT raising gesture was the same across consonant length.

Fig. 2 again shows some interspeaker variability, but this time for both speaker groups. Early gestural onset for geminates was observed in learner BCF05, but not for the native speaker BJF03. The exceptional behavior of BCF05 is consistent with the distinction in the TT gestural duration, and suggests that she has acquired the gestural timing for Japanese geminates.

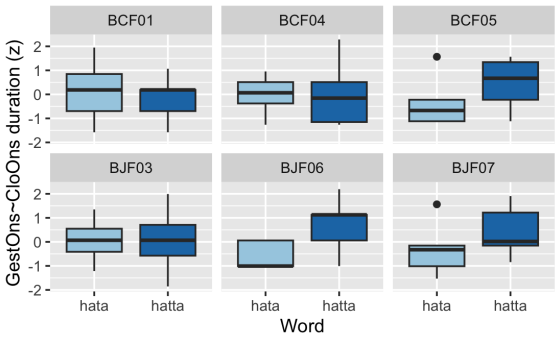


Figure 2: *GestOns~CloOns*, per word and speaker.

Comparison of the *CloOns~GestMax* durations reveals that overall, the TT maximal height was achieved later for geminates among the Japanese speakers (β=14.45, t=4, p<.001), but not among the Chinese speakers (β=1.83, t=0.62, p=0.53). Fig. 3 shows that the covert TT raising gesture continued after the acoustic onset of the consonant, and its duration was slightly longer for geminates than for singletons in native speakers.

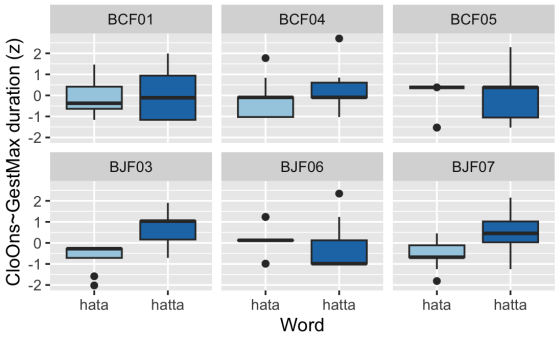


Figure 3: *CloOns~GestMax*, per word and speaker.

The difference was the largest for BJF03 (3.94 times longer for geminates on average), suggesting that the timing of the TT raising gesture onset in her geminate productions was no different from that of her singletons, but that completing the gesture took much longer. While BJF03's tendency deviates from those of the other Japanese speakers reported in this study, some variability in the timing control for geminates among native speakers has been previously reported [10]. Once a fuller analysis with more Japanese speakers is implemented, it may provide an insight into the typology of L1 articulatory strategies for the consonant length.

3.3. Acoustic duration

Table 3 summarizes the acoustic durations. The first point to note is the considerable difference in the overall speaking rate between the two groups, as

exemplified in the word duration (β=−280.48, t=−16, p<.001). Second, the acoustic results reveal that the Chinese speakers in this study were able to differentiate singletons and geminates in their productions at the segmental level. The geminate-to-singleton ratios for each segment were largely in line with those of the native speakers, and their C_2/Word ratios, known to have high classification accuracy at the boundary ratio of 0.35 [23], were 0.33 for singletons and 0.46 for geminates.

Table 3: Duration in milliseconds (sd) and geminate-to-singleton ratios (GSR) for word, V_1, C_2, and V_2 per native language (NL) and consonant length (L).

NL	L	*n*	Word	V_1	C_2	V_2
CH	S	30	489.53 (49.79)	51.3 (9.67)	159.83 (22.26)	141.96 (29.68)
	G	29	663.86 (62.43)	64.77 (14.2)	309.72 (66.52)	171.88 (35.84)
GSR			1.36	1.26	1.94	1.21
JP	S	29	249.29 (30.09)	47.92 (9.79)	82.05 (15.56)	60.1 (9.33)
	G	28	341.21 (45.39)	67.42 (10.4)	161.13 (27.24)	50.74 (8.13)
GSR			1.37	1.40	1.96	0.84

In addition, the learners have acquired the pre-geminate V_1 lengthening (CH: β=13.12, t=5.54, p<.001; JP: β=19.24, t=7.32, p<.001). This was unexpected given our findings on the TT gesture duration. However, not only the native speakers and BCF05, the exceptional learner, but also the other two Chinese speakers with a shorter TT gesture duration have mastered pre-geminate V_1 lengthening (Fig. 4).

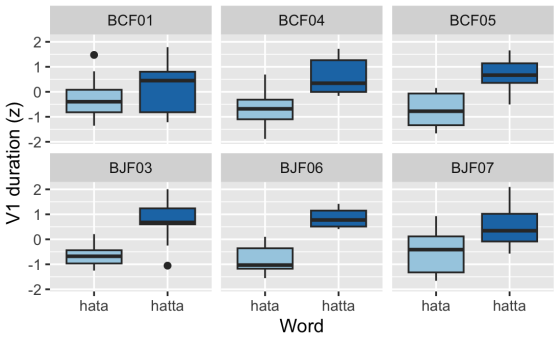


Figure 4: V_1 duration per word and speaker.

Meanwhile, although clear segmental distinctions were made by the learners, some of the durational correlates were not adequately acquired, as illustrated in Fig. 5. For example, the slight shortening of V_2 among the Japanese speakers (β=−9.58, t=−5.2, p<.001) was not shown by the Chinese speakers. Rather, those learners lengthened V_2 after geminates (β=29.87, t=4.31, p<.001). Furthermore, the ratio of the acoustic closure duration and V_1 duration was strikingly different between the two groups. This pattern is reminiscent of the over-exaggerated geminates in L2 speech reported in [1]. The slower speaking rate and the V_1 lengthening effect suggest that TT gesture duration was short for Chinese speakers.

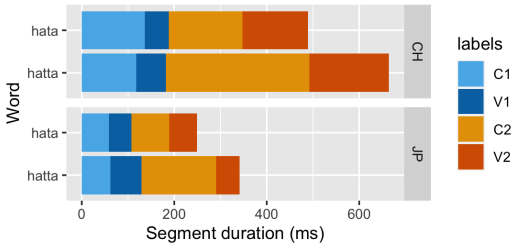


Figure 5: Bar chart for average segment duration.

4. CONCLUSION

The duration and timing of the TT raising gesture in the singleton and geminate /t/ were investigated, in relation to the acoustic segmental durations. Results showed that while native speakers of Japanese employed a longer TT raising gesture for geminates than for singletons, such an articulatory distinction was not made by the Chinese speakers except for one (BCF05). In addition, while the TT raising gesture started early relative to the acoustic consonantal onset in geminates for two out of the three Japanese speakers and the learner BCF05, that was not the case for the other two Chinese speakers and the other Japanese speaker (BJF03). This suggests that it is possible for the learners to acquire the gestural timing control for Japanese geminates, and that there may be some interspeaker variability among native speakers.

In the meantime, the acoustic V_1 lengthening in geminates was seen robustly across speaker groups. While this can be accounted for by the high proficiency of the speakers included in the current analysis, it also shows that it is possible for advanced learners of Japanese with an L1 Chinese background to manipulate the segmental duration to realize the perceptually relevant acoustic distinction, even without acquiring a native-like gestural timing. This finding is in support of the view that learners employ different strategies from that of native speakers to achieve timing control, possibly governed by a different rhythmic unit [4, 6, 8].

How the distinction is achieved articulatorily should be addressed in more details in future research. In particular, this study focused on the TT gesture of a limited number of advanced learners. Further analysis should include the articulation of the tongue as a whole, in more tokens produced by learners with various proficiency levels. Of special interest is the tongue positions during V_1, to address the magnitude of articulatory gestures before geminates and their relation to gestural duration.

5. ACKNOWLEDGMENTS

This study was supported by JSPS KAKENHI Grant Numbers JP19K13254, JP22J01381, JP23K12212, and Sophia University Special Grant for Academic Research (Research in Priority Areas). We thank the anonymous reviewers for their comments.

6. REFERENCES

[1] Toda, T. 2003. Acquisition of special morae in Japanese as a second language. *Journal of the Phonetic Society of Japan* 7(2), 70–83.

[2] Tsukada, K., Yurong. 2022. Non-native perception of the Japanese singleton/geminate contrast: Comparison of Mandarin and Mongolian speakers differing in Japanese experience. *Interspeech 2022* 3068–3072.

[3] Lee, A., Mok, P. 2018. Acquisition of Japanese quantity contrasts by L1 Cantonese speakers. *Second Language Research* 34(4), 419–448.

[4] Toda, T. 1994. Interlanguage phonology: Acquisition of timing control in Japanese. *Australian Review of Applied Linguistics* 17(2), 51–76.

[5] Han, M. 1992. The timing control of geminate and single stop consonants in Japanese: A challenge for nonnative speakers. *Phonetica* 49, 102–127.

[6] Yamakawa, K., Amano, S., Kondo, M. 2021. Mispronunciation of Japanese singleton and geminate stops by Korean and Taiwanese Mandarin speakers. *Acoust. Sci. & Tech.* 42(2), 73–82.

[7] Idemaru, K., Guion, S. 2008. Acoustic covariants of length contrast in Japanese stops. *J. Int. Phon. Assoc.* 38, 167–186.

[8] Fujimoto, M., Maekawa, K. 2014. Effects of sokuon on adjacent vowel duration: An analysis of the Corpus of Spontaneous Japanese. *Journal of the Phonetic Society of Japan* 18(2), 10–22.

[9] Maddieson, I. 1985. Phonetic cues to syllabification. In: Fromkin, V. A. (ed.), *Phonetic Linguistics*. New York: Academic Press, 203–221.

[10] Fujimoto, M., Funatsu, S., Hoole, P. 2015. Articulation of single and geminate consonants and its relation to the duration of the preceding vowel in Japanese. *Proc. of the ICPhS 2015*.

[11] Takada, M. 1985. Sokuon no chouon jo no tokucho ni tsuite [Articulatory characteristics of sokuon]. *Kenkyu Hokokushu* 6, 17–40.

[12] Ishii, T. 1999. A study of the movement of the articulatory organs in Japanese geminate production: An X-ray microbeam analysis. *J. Otolaryngol. Jpn.* 102, 622–634.

[13] Löfqvist, A. 2007. Tongue movement kinematics in long and short Japanese consonants. *J. Acoust. Soc. Am.* 122(1), 512–518.

[14] Fujimoto, M. 2013. Timing differences in articulation between single and geminate voiceless stop consonant: An analysis of cine-MRI data. Paper presented at the 3rd International Conference on Phonetics and Phonology, Tachikawa, Japan.

[15] Agency for Cultural Affairs, Government of Japan. 2021. Kokunai no nihongo kyoiku no gaiyo [Summary of domestic Japanese language education]. *Nihongo Kyoiku Jittai Chosa Hokokusho* https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/r03/pdf/93791201_01.pdf.

[16] The Japan Foundation. 2018. *Survey Report on Japanese-Language Education Abroad 2018*. https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey2018/Report_all_e.pdf.

[17] Spreafico, L., Pucher, M., Matosova, A. 2018. UltraFit: A speaker-friendly headset for ultrasound recordings in speech science. *Interspeech 2018* 1517–1520.

[18] Articulate Instruments Ltd. 2022. Articulate Assistant Advanced (AAA). Version 2.19.08, http://www.articulateinstruments.com/.

[19] R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

[20] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. Version 6.2.19, retrieved September 2022 from http://www.praat.org/.

[21] Tiede, M. K. 2022. *GetContours.* GitHub repository. https://github.com/mktiede/GetContours.

[22] Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting linear mixed-effects models using lme4. *arXiv Prepr. arXiv1406.5823*.

[23] Hirata, Y., Whiton, J. 2005. Effects of speaking rate on the single/geminate stop distinction in Japanese. *J. Acoust. Soc. Am.* 118, 1647–1660.

COMPARISON OF L2 KOREAN PRONUNCIATION ERROR PATTERNS FROM FIVE L1 BACKGROUNDS BY USING AUTOMATIC PHONETIC TRANSCRIPTION

Eun Jung Yeo*, Hyungshin Ryu*, Jooyoung Lee, Sunhee Kim, Minhwa Chung

Seoul National University
{ej.yeo, rhss10, excalibur12, sunhkim, mchung}@snu.ac.kr

ABSTRACT

This paper presents a large-scale analysis of L2 Korean pronunciation error patterns from five different language backgrounds, Chinese, Vietnamese, Japanese, Thai, and English, by using automatic phonetic transcription. For the analysis, confusion matrices are generated for each L1, by aligning canonical phone sequences and automatically transcribed phone sequences obtained from fine-tuned Wav2Vec2 XLS-R phone recognizer. Each value in the confusion matrices is compared to capture frequent common error patterns and to specify patterns unique to a certain language background. Using the Foreign Speakers' Voice Data of Korean for Artificial Intelligence Learning dataset, common error pattern types are found to be (1) substitutions of aspirated or tense consonants with plain consonants, (2) deletions of syllable-final consonants, and (3) substitutions of diphthongs with monophthongs. On the other hand, thirty-nine patterns including (1) syllable-final /l/ substitutions with /n/ for Vietnamese and (2) /ɯ/ insertions for Japanese are discovered as language-dependent.

Keywords: Comparative analysis, Pronunciation error patterns, Automatic phonetic transcription, L1 influence, L2 Korean speech

1. INTRODUCTION

Computer Assisted Pronunciation Training (CAPT) has emerged as an effective tool for non-native speakers, offering cost-effective feedback while overcoming the time and location constraints of traditional language learning [1, 2]. With the advancement of deep learning, novel architectures have been proposed to improve the detection and diagnosis performance of the CAPT system [2, 3, 4]. Linguistic studies were conducted with a different point of view, where the characteristics of non-native speech are explored [5, 6, 7, 8, 9, 10, 11].

The influence of the first language (L1) on the second language (L2) pronunciation has long been investigated. However, due to the requirement of time- and labor-intensive human transcription, analyses were often limited to certain acoustic phenomena from a small number of participants. For example, 127 obstruent-initial words were collected from 22 English-speaking learners to analyze their realizations of Korean stops [5], while 11 words with monophthongs or diphthongs were recorded from 53 Egyptians to understand their Korean vowel perception and production [6]. On the other hand, some studies have taken advantage of deep learning methods including forced alignments [7, 8] or feature extraction [9, 10, 11], which allowed an analysis of larger materials or participants. Nevertheless, their analyses were often limited to a small subset of phone inventories, such as fricatives [7], nasals [9], and vowels [8, 9, 10, 11].

To mitigate the limitation of human transcriptions especially on large-scale analysis, utilizing automatic transcription can be one possible option. Model-based transcriptions have advantages in terms of time- and cost-efficiency, consistency, and objectivity [12]. With self-supervised-based speech models (SSL) showing state-of-the-art performances in various speech tasks including automatic speech recognition [13, 14], this work proposes to employ automatic phonetic transcription for L2 mispronunciation analysis.

This study conducts a large-scale analysis on L2 Korean pronunciation error patterns of speakers from five different L1 backgrounds, Chinese, Vietnamese, Japanese, Thai, and English. Confusion matrices are generated by aligning the canonical and automatically transcribed phones obtained from fine-tuned SSL Wav2Vec2 XLS-R recognizer [13]. The error patterns for each non-native speech are inferred from the confusion matrix, and are scrutinized with respect to L1 background. We expect our analysis to bring new insights into L1-influenced pronunciations and provide linguistic clarity in developing the CAPT system.

*Equal contributions.

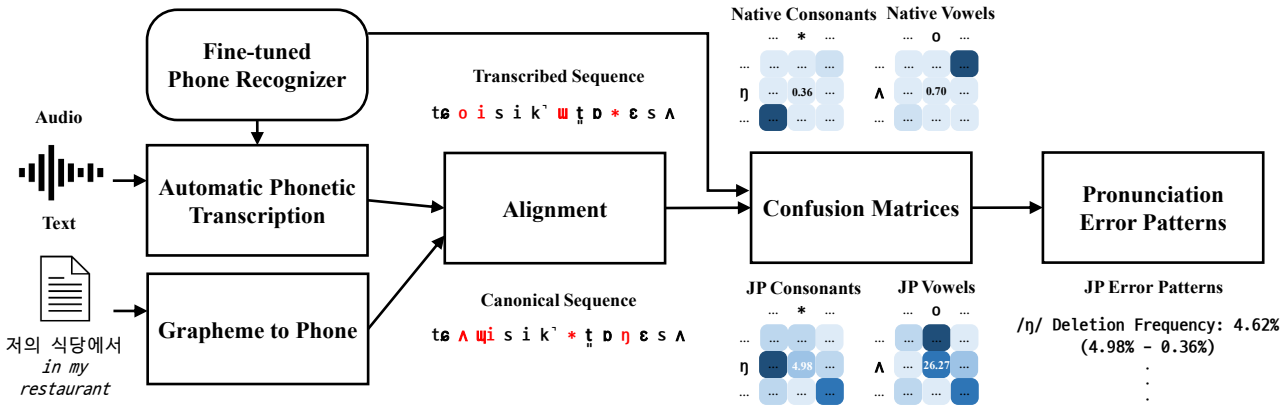


Figure 1: Overview of the pronunciation error patterns analysis using automatic phonetic transcription

2. METHOD

2.1. Corpus

Foreign Speakers' Voice Data of Korean for Artificial Intelligence Learning [15] is an open-source dataset designed to train automatic speech recognition systems for Korean learners. The dataset consists of L2 Korean speech from 1911 speakers with 80 different L1s, where each speaker recorded both read-speech and spontaneous speech. In this paper, we restricted the analysis to five L1 backgrounds, where the most abundant recordings were collected - Chinese (ZH), Vietnamese (VI), Japanese (JP), Thai (TA), and English (EN). In addition, we examined only the read-speech to exclude the influence of speech fluency and focus on pronunciation error patterns. Participants without official scores on the Test of Proficiency in Korean (TOPIK) are also excluded from the analysis. As Table 1 presents, the dataset includes beginners, intermediate and advanced learners for all five language backgrounds.

Table 1: Number of speakers and utterances

Language	Beginner		Intermediate		Advanced		Total Hrs
	spk	utt	spk	utt	spk	utt	
Chinese	2	513	56	18599	291	127811	487.41 h
Vietnamese	12	2837	85	36100	144	74921	412.91 h
Japanese	3	1474	27	12004	140	87915	326.35 h
Thai	44	19773	57	26377	44	26688	260.57 h
English	14	938	15	1682	12	4272	24.34 h

2.2. Analysis

Figure 1 is the overview of our analysis. First, **(1) automatic phonetic transcription** on non-native speech is performed by using the fine-tuned Wav2Vec2 XLS-R phone recognizer. Next, **(2) confusion matrices** for both native and non-native speech are created by aligning the canonical phone sequences and automatically transcribed phone sequences. Finally, **(3) pronunciation error patterns** for non-native speech are extracted from confusion matrices and analyzed regarding five L1 backgrounds. We release the source code and the fine-tuned model for the ease of reproduction.[1]

2.2.1. Automatic phonetic transcription

For automatic phonetic transcription, we choose to utilize Wav2Vec2 XLS-R-300m model, which is pre-trained on 128 different languages including Chinese, Vietnamese, Japanese, Thai, and English [13]. The model is fine-tuned with a private, phonetically balanced native Korean read-speech corpus, with a split of 108 hours of the training set (54000 audio files) and 12 hours of the test set (6000 audio files). The transcription used for fine-tuning is generated by trained linguists with the phone set introduced in [16]. Thus the same set is used for error pattern analysis in this study. The phone set includes 40 phones, consisting of 19 consonants, 17 vowels, and 4 allophones. As for allophones, [pˀ],[tˀ],[kˀ] are stops at codas and [ɾ] is /l/ in syllable-initial position. Our model showed a Phone Error Rate (PER) of 3.88% on the test set, which verifies the model's performance.

2.2.2. Confusion matrices

Confusion matrices are created for the native dataset, the non-native dataset as a whole, and each L1 group. Canonical phone sequences are generated with the grapheme to phone module in Montreal Forced Aligner [16] and are aligned with the decoded phone sequences from the fine-tuned model, by using SCTK package offered at Kaldi [17]. Figure 2 demonstrates the vowel confusion matrix for Japanese L2 Korean speech which refers

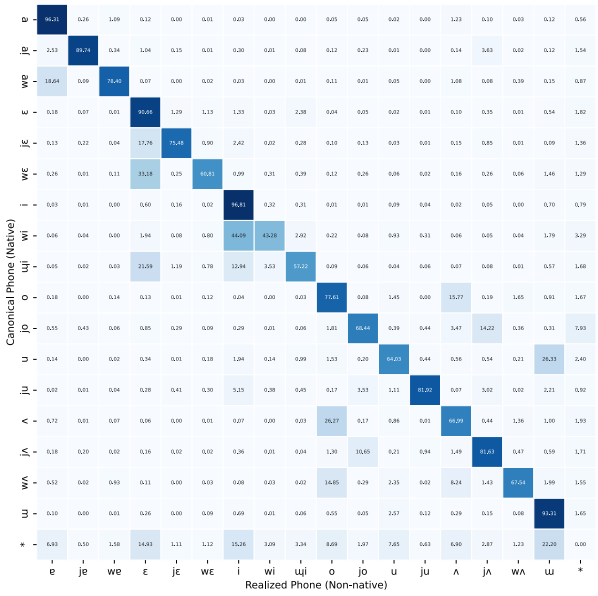


Figure 2: Confusion matrix for L2 Korean vowel pronunciation of Japanese speakers

to L2 error patterns. * represents insertion and deletion. Considering each value as the percentage of observations of the canonical phone row, each row sums to 100.

2.2.3. Pronunciation error patterns

Pronunciation error patterns for each L1 background are inferred from the corresponding confusion matrix. Error patterns include substitution, insertion, and deletion errors. We propose two types of analyses to scrutinize L1-influenced error patterns.

In the first analysis, we focus on determining the common error patterns shared across five non-native groups. These common patterns can reveal the general characteristics of L2 Korean speech and be used for core feedback in pronunciation training. The first step involves choosing phones for each L1 background that have an accuracy lower than the average accuracy. Different thresholds are used because the performance of the phone recognizer can vary depending on the L1 background. Then, a list of pronunciation error patterns is created for each L1 background, which includes the three most common error patterns for each selected phone. Lastly, the common error patterns are identified by comparing the five error pattern lists and determining which error patterns are shared.

As for the second analysis, we aim to find unique patterns for each L1. Language-dependent error patterns can give guidelines for providing learners with customized feedback. The Chi-square test of independence is conducted to compare each L1 group to the average of five L1 groups to look at the association of L1 with each error pattern. The four most frequently observed patterns for each canonical phone are only tested, leading to a total of 840 error patterns (5 L1 backgrounds $\times$ 42 canonical phones (40 phones, insertion, deletion) $\times$ 4 realized phones) being statistically analyzed.

For both analyses, native confusion matrix error values were subtracted from corresponding non-native confusion matrix values beforehand, as they include not only the errors of the phone recognizer itself but also the pronunciation variation of native speakers.

3. RESULTS

3.1. Common error patterns

Three types of error patterns are observed to be commonly shared across five languages: two types for consonants and one type for vowels. For consonants, seven error patterns are related to the first type, substitution errors of tense or aspirated with plain consonants: /t^{h}/to/t/, /k^{h}/to/k/, /t͡ɕ^{h}/to/t͡ɕ/; /p͈/to/p/, /t͈/to/t/, /t͡ɕ͈/to/t͡ɕ/, /s͈/to/s/. The second error type is deletion errors of syllable-final, which includes [t$^{\neg}$] and /l/ coda deletions. As for vowels, substitution errors from diphthongs to monophthongs are found: /wi/to/i/, /ɯi/to/i/, /wʌ/to/ʌ/.

3.2. Language-dependent error patterns

According to the statistical results, only 39 out of 840 pronunciation patterns showed L1 association. 25 patterns had L1 causing speakers with higher mispronunciation rates than the average, and the rest 14 had L1 being helpful to better pronunciations. In this paper, we focus on the former 25 pronunciation error patterns as presented in Table 2 and show the full results in our repository[1]. *Canon* and *Real* each refers to canonical and realized phones, with *Average Freq* referring to the confusion matrix value averaged across the five L1s. For example, row 1 can be interpreted that the [k$^{\neg}$] accuracy of Chinese learners was lower than the average of five L1s.

Chinese-speaking learners have 3 language-dependent error patterns including [k$^{\neg}$] mispronunciation, [k$^{\neg}$] deletion, and /ju/to/jo/ substitution. **Vietnamese**-speaking learners have 3 error patterns which includes substitution from /t͡ɕ/to/t͡ɕ^{h}/, /l/ mispronunciation, and the following /l͈/to/n/ substitution. **Japanese**-speaking learners have the most various association with L1, with 18 errors showing statistical significance. The error

patterns include mispronounced aspirated stops (/pʰ/,/tʰ/,/kʰ/) and their substitution to tense stops (/p͈/,/t͈/,/k͈/). Mispronunciation of /ŋ/ and its substitution to /n/ are more frequent than average learners. Insertion of /ɾ/ and /ɯ/ and substitution of diphthongs to monophthongs are also observed to be significantly frequent (/wɐ/,/wɛ/,/wi/). Lastly, the learners often mispronounce monophthongs (/u/,/ʌ/) to different monophthongs (/ɯ/,/o/). For **Thai**-speaking learners, frequent substitution of /ʨʰ/to/s/ are found. No phones are significantly mispronounced by **English**-speaking learners only. This result may imply that English-speaking learners have advantages in learning Korean compared to other learners. However, to validate this claim, further research using a more balanced dataset is essential, considering the limited number of English-speaking participants in our dataset.

Table 2: Frequency values (%) of confusion matrices. * < .05, ** < .01, *** < .001

L1	Canon	Real	L1 Freq. (%)	Average Freq. (%)	P-value
ZH	[kˀ]	[kˀ]	58.87	**74.95**	*
	[kˀ]	*	**22.79**	7.32	**
	/ju/	/jo/	**23.50**	11.61	*
VI	/ʨ/	/ʨʰ/	**18.73**	7.99	*
	/l/	/l/	56.54	**73.23**	*
	/l/	/n/	**23.20**	7.52	**
JP	/pʰ/	/pʰ/	51.22	**70.70**	**
	/pʰ/	/p/	**20.72**	7.33	*
	/tʰ/	/tʰ/	44.40	**62.24**	*
	/tʰ/	/t/	**20.51**	5.99	**
	/kʰ/	/kʰ/	54.21	**70.20**	*
	/kʰ/	/k/	**31.15**	12.74	**
	/ŋ/	/ŋ/	55.22	**82.67**	***
	/ŋ/	/n/	**31.22**	10.10	***
	*	[ɾ]	**10.43**	1.17	*
	*	/ɯ/	**8.98**	0.07	**
	/wɐ/	/ɐ/	**14.58**	3.96	*
	/wɛ/	/ɛ/	**26.69**	9.42	**
	/wi/	/wi/	43.28	**58.58**	*
	/wi/	/i/	**32.05**	11.49	***
	/u/	/u/	64.03	**77.79**	*
	/u/	/ɯ/	**24.72**	7.08	**
	/ʌ/	/ʌ/	66.99	**81.78**	*
	/ʌ/	/o/	**25.58**	10.46	**
TA	/ʨʰ/	/s/	**22.05**	7.33	**

4. DISCUSSION

Based on the observations in Section 3, we uncover the three L1-influence factors attributing to the L2 Korean pronunciation error patterns. The result of our analysis is found to be further supported by the conclusions from various previous literature.

First is differences in **phoneme inventory** [18, 19]. Substitution errors of three-way contrasted consonants can be explained by languages not distinguishing consonants as plain-tense-aspirated [20]. The commonly found substitutions to plain consonants further support the analysis, regarding each five languages includes plain consonants. Results also reveal an interesting finding of Japanese learners' consistent tensification of aspirated stops, alongside the tendency to substitute aspirated stops to plain. Previously, word-medial tense and aspirate distinction in perception have been reported to be difficult for Japanese speakers [21]. Our study implies that the production of aspirated stops may also be related to this phenomenon.

Error patterns can also be attributed to differences in **syllable structure**, especially in the syllable-final position of languages [22]. While Korean allows /l/ as coda, four languages excluding English do not allow syllable-final /l/. Although English allows /l/ as the final consonant, the allophone /ɬ/ at the word-final position sounds different from Korean /l/. Vietnamese learners especially had a higher chance of mispronouncing /l/, in particular as /n/ that has a similar place of articulation [23, 24]. Chinese learners show a significant pattern of [kˀ] deletion, which was previously explained in [25] with syllable structure. Japanese learners consistently insert /ɯ/ as well, due to their syllable-based constraints with the tendency to open-syllabification [26].

Lastly, **pronunciation rule** can explain the frequent deletion of coda [tˀ]. The corresponding rule is tensification, where the plain consonants are pronounced as tense in a stop-plain consonant sequence. While the sequence must be acoustically realized as long closure, many L2 Korean learners are reported to produce shorter closures compared to the natives, because they are either unaware of or unfamiliar with the tensification rule [27, 28]. The mispronounced sounds are often perceived as deletion of coda [27], which is observed in our automatic phone transcriptions as well.

5. CONCLUSION

This paper presents a large-scale linguistic analysis on L2 Korean pronunciation error patterns from speakers of five different L1s, Chinese, Vietnamese, Japanese, Thai, and English, by using the automatic phonetic transcription. Comparative analyses regarding L1 background reveal common error patterns and L1-dependent error patterns, which can be further explained by three types of L1 influences: phoneme inventory, syllable structure, and pronunciation rule. This study contributes not only in integrating and expanding the previous knowledge of L2 Korean error patterns on a large-scale basis, but also validating the possibilities of automatic phonetic transcriptions on non-native speech analysis. Future work includes considering L1 and uncategorizable phones to yield a better and deeper understanding of L1-influenced non-native pronunciation error patterns.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01833) supervised by the IITP (Institute for Information&Communications Technology Planning&Evaluation)

6. REFERENCES

[1] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (capt): Current issues and future directions," *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.

[2] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *2022 APSIPA ASC*, 2022, pp. 575–582.

[3] M. Yang, K. Hirschi, S. D. Looney, O. Kang, and J. H. L. Hansen, "Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment," in *Interspeech*, 2022.

[4] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhang, "A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis," in *Interspeech*, 2021.

[5] E. J. Kong, S. Kang, and M. Seo, "The acoustic cue-weighting and the l2 production-perception link: A case of english-speaking adults' learning of korean stops," *Phonetics and Speech Sciences*, vol. 14, no. 3, pp. 1–9, 2022.

[6] S. Benjamin and H.-Y. Lee, "The perception and production of korean vowels by egyptian learners," *Phonetics and Speech Sciences*, vol. 13, no. 4, pp. 23–34, 2021.

[7] A. Jatteau, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Final devoicing of fricatives in french: Studying variation in large-scale corporawith automatic alignment," 2019.

[8] J. Yang and R. Fox, "L1-l2 interactions of vowel systems in young bilingual mandarin-english children," *J. Phonetics*, vol. 65, pp. 60–76, 2017.

[9] J. Yuan, H. Lin, and Y. Liu, "Nasal coarticulation in l1 and l2 english speech: A large-scale study," *Training*, pp. 96–6, 1974.

[10] S. Shi and C. Shih, "Acoustic analysis of l1 influence on l2 pronunciation errors: A case study of accented english speech by chinese learners," *ICPHS, Melbourne, Australia*, 2019.

[11] S. Shi, C. Shih, and J. Zhang, "Capturing l1 influence on l2 pronunciation by simulating perceptual space using acoustic features," in *Interspeech*, 2019, pp. 2648–2652.

[12] C. Cucchiarini and H. Strik, "Automatic phonetic transcription: An overview," in *ICPHS*, 2003, pp. 347–350.

[13] A. Babu *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022.

[14] K. Choi and H.-M. Park, "Distilling a pretrained language model to a multilingual asr model," in *Interspeech*, 2022.

[15] "Foreign speakers' voice data of korean for artificial intelligence learning," https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=505, 2020.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[17] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[18] L. Wang, X. Feng, and H. Meng, "Mispronunciation detection based on cross-language phonological comparisons," in *ICALIP*, 2008, pp. 307–311.

[19] S. Khanal, M. T. Johnson, and N. Bozorg, "Articulatory comparison of l1 and l2 speech for mispronunciation diagnosis," in *SLT*, 2021, pp. 693–697.

[20] J. J. Holliday, "A longitudinal study of the second language acquisition of a three-way stop contrast," *Journal of Phonetics*, vol. 50, pp. 1–14, 2015.

[21] T. Yasuta, *Stop perception in second language phonology: Perception of English and Korean stops by Japanese speakers*. University of Hawaii, 2004.

[22] Y. He, "Production of english syllable final/l/by mandarin chinese speakers." *Journal of Language Teaching & Research*, vol. 5, no. 4, 2014.

[23] K. B. Choe, J. Y. Song, and S.-G. Park, "Analysis of pronunciation errors of korean syllable-final liquid /ㄹ/ by vietnamese learners of korean," *Urimal*, no. 64, pp. 181–209, 2021.

[24] M.-k. Hwang, "Analysis of the Korean liquid pronunciation of Vietnamese entry-level Students," *Journal of Ehwa Korean Language and Literature*, vol. 39, pp. 49–87, 2016.

[25] E. J. Lee and I. H. Woo, "A study on an educational plan for the pronunciation of the final consonants /ㄱ, ㄷ, ㅂ/ in korean for chinese korean learners -by applying the syllable structure of korean to that of chinese reversely-," *The Academy for Korean Language Education*, no. 97, pp. 327–359, 2013.

[26] E. K. Yoon, "An experimental perceptual study of different syllable structures in l1 & l2," *Journal of Dong-ak Language and Literature*, no. 64, pp. 409–438, 2015.

[27] L. J and P. K, "A study on the korean learners' realization aspect and teaching methods of glottalization after obstruent," *The Korean society of bilingualism*, no. 71, pp. 223–248, 2018.

[28] K. K H *et al.*, "A study on the final consonant pronunciation errors of chinese korean learners," Ph.D. dissertation, Jeju University, 2019.

[1] https://github.com/eunjung31/L2K-PronunciationError

Learning phonotactically complex L3 words: Are bilinguals more successful?

Adriana Hanulíková

University of Freiburg, Germany
Basque Center on Cognition, Brain and Language (BCBL), Donostia-San Sebastián, Spain
Adriana.Hanulikova@germanistik.uni-freiburg.de

ABSTRACT

Bilinguals are often claimed to acquire new languages more easily than monolinguals due to enhanced meta-/linguistic abilities driven by prior experience with languages. When it comes to learning how to pronounce foreign words, empirical evidence is inconclusive and mostly limited to analyses of single speech sounds. Less is known about the acquisition of complex phonotactics in a foreign language during adulthood. The focus of this study is on the very initial phase of foreign language production, specifically examining whether the frequency of dominant language use predicts the ease of imitation of phonotactically complex words in a delayed repetition task. To this end, 251 Spanish and Spanish/Basque speakers imitated Slovak words (e.g. [stɾ̩t͡ʃ] "stick"). Preliminary analyses of vowel epenthesis revealed more insertions with higher language use, supporting the notion that sensorimotor and perceptual flexibility improves with frequent and sustained experience with different languages.

Keywords: phonotactics, epenthesis, L3 acquisition, production, bilingualism, individual differences

1. INTRODUCTION

Many learners of a foreign language struggle to pronounce unfamiliar speech sounds so that proficient listeners of that language can easily understand them. Yet some learners achieve this with apparent ease. What is responsible for such individual differences? Here we explored the extent to which bilingualism contributes to such differences by focusing on production of phonotactically complex words.

It is well known that the acquisition of an additional language (L2/L3) later in life is often accompanied by the presence of an audible accent in the pronunciation. Not only production is affected, but also the perception often remains imprecise. This is the case because learners hear new languages through the ears trained for their first or dominant (L1) language [1]. Researchers have long sought the answer to questions such as a) why mastery in an additional language perception and production is so difficult in adulthood compared to childhood, and b) why there is such great variability in performance.

A number of both linguistic and non-linguistic factors determine acquisition success as well as the individual differences. Linguistic factors such as markedness, universal preferences, characteristics of the L1 compared to the L2/L3, and language transfer can explain part of the difficulties [2-4]. Non-linguistic factors such as motivation, quality of training, amount of L2/L3 input, socio-psychological effects, personality, and intelligence can explain some of the individual differences between learners [4-6]. A vast number of cross-disciplinary studies have claimed neurocognitive plasticity and age of acquisition (AoA) to be among the main contributors to the decay of learning abilities in adulthood [for a review, see 7]. Native phonetic categories are believed to be less influential in childhood than in adulthood, resulting in better perception and production abilities in a child's L2 [4, 8, 33]. However, there is no strict age limit to speech-learning abilities [9]. Adults can exhibit exceptional phonetic abilities as shown in reports on (10 to 20% of) late learners who sound almost native-like in the newly acquired language [10-12]. Thus considerable plasticity is available for adult speech learning, and further factors such as experience need to be considered [33].

The majority of research within this area has focused on ultimate proficiency attainment after several years of language acquisition, which has the disadvantage of being strongly determined by diverse non-linguistic factors. Other studies focus on training perception or production of phonetic contrasts or on individual differences in consonant clusters perception and production after only a very short exposure to an unfamiliar language [13]. [13] report large individual differences but the underlying cause for performance variation remains unclear.

Apart from the factors mentioned above, some studies suggest that bilingualism provides learners with benefits when learning an additional language. Previous studies have demonstrated that bilinguals outperform monolinguals across various language related tasks, such as learning new words [14, but see 31], phonetic contrasts [15], and pronunciation [16]. Studies that address the initial stages of phonotactic acquisition are mostly limited to late bilingual populations [13]. It thus remains unclear to what extent bilingualism and other linguistic factors provide learners with benefits in word learning, and

more specifically in complex phonotactics. In addition, most studies do not consistently isolate the influence of language use and treat bilingualism categorically. An alternative and perhaps better way to capture the influence of bilingualism on learning and linguistic skills is the amount of active use or exposure to one or more languages. We therefore operationalized language use as a continuous variable [for a similar approach, see 29-31].

Phonetic production ability for newly-learned consonant clusters requires sensorimotor flexibility to minimize the influence of native language articulation, and this ability is likely to be more limited in adults compared to children. Individual differences in sensorimotor flexibility could have a genetic basis or arise from experiential factors, such as growing up speaking different languages [see also 17, for vocal flexibility of singers], and may be related to phonological working memory [18]. Similarly, phonetic perception ability may depend on perceptual plasticity or be shaped by experience with diverse phonologies [20, 33]. It has been suggested that speech-specific capabilities could partly explain individual differences in phonetic mastery of early bilingual speakers [19].

Whether active bilinguals who grew up speaking two languages of comparable complexity in terms of consonant clusters will exhibit better imitation skills than active monolinguals is yet unclear. The present study addresses this issue by a) exploring the mastery of complex words at the very onset of exposure to a new language, and b) including speakers of typologically unrelated languages with relatively comparable consonant clusters and phonology (Basque and Castilian Spanish). This allows us to examine the relationship between active bilingualism and L3 acquisition in cases where the quantity and quality of the phoneme repertoire is less likely to provide bilinguals with benefits in speech processing (as shown for L3 perception [20]).

2. THE PRESENT STUDY

The objective of this study was to explore the performance of monolingual and bilingual speakers in imitating words with complex consonant clusters in a foreign language (Slovak). The analysis reported here focuses on the question of whether language habits predict successful production skills, as measured by vowel insertions. If bilinguals exhibit fewer insertions compared tomonolinguals, it would suggest that sensorimotor flexibility and perceptual flexibility are influenced by the amount of experience in multiple languages. This result would shed light on the interface between bilingualism and language learning, and would provide insights into whether there is a physiological basis for speech production difficulties in late learners.

To examine individual differences in L3-speech learning, we administered a delayed imitation task using auditorily presented Slovak words to participants who had not learned Slovak before. In addition, we used a Spanish/Basque nonword-repetition task to control for variability in phonological short-term memory, which is known to differ among learners. This task assessed the ability to store and repeat unfamiliar phonological sequences.

Slovak is well suited for these purposes due to its relatively complex syllabic structure [for more details, see 21, 22]. In Slovak, consonant clusters of up to four consonants are allowed in onset position (e.g. *pstruh* [pstrʊ̯x] ‘trout’). In contrast, Spanish and Basque have a maximum of two onset consonants [for more details on combinatory restrictions, see 23, 24] and tend to prefer CV syllables over consonant clusters in onset positions [24, 25]. Word initially, Castilian Spanish permits clusters of a stop or /f/ followed by a liquid, and /f/ is the only possible fricative as the first cluster member. Spanish also avoids impermissible /s/ + obstruent clusters through epenthesis in word-initial and word-medial positions (e.g. stop /es.ˈtop/). Despite many shared phonological characteristics, Basque has a larger set of sibilants and palatals compared to Spanish.

It is expected that there will be individual variation among speakers, with participants who use their L1 more frequently imitating phonotactically complex structures with less ease, resulting in more insertions, compared to active bilinguals who use their L1 less often. In the preliminary analysis reported here, imitation success was measured by schwa and /e/ insertion at the cluster onset and within the word.

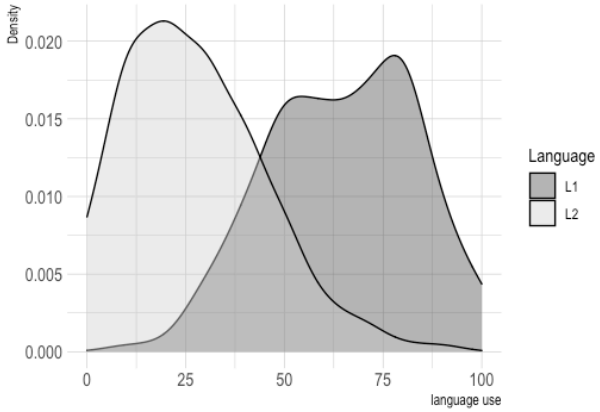
3. METHODS

4.1. Participants

A total of 251 participants with self-reported normal hearing were recruited from the BCBL subject pool and received payment for their participation. We excluded 28 adults due to double/missing/noisy recordings or incomplete questionnaires. Participants who did not indicate Spanish and/or Basque as their L1/L2 were excluded from the analysis, leaving a final sample of 189 (mean age 22.9, SD 4.7, range 15-43, 99 women, 90 men, 106 simultaneous bilinguals who learned both languages from birth, 142 Spanish dominant, 47 Basque dominant). The amount of self-reported daily L1 use was used to calculate active bilingualism, as shown in Figure 1, which displays a density plot of the relative proportion of L1 and L2 use on a scale from 0 to 100,

where 0 refers to no L1/L2 use. On average, speakers used their L1 64.8% of the time (SD 18.7) and L2 27% of the time (SD 17.7). None of the participants reported knowledge of any Slavic language. All but 17 of the participants indicated English as their dominant L2 or L3, all but two had knowledge of Basque, and all but six adults reported skills in an L3, using it 7% of the time (SD 8.1).

Figure 1: Density plot for the variable language use.



4.2. Material

The stimuli for the delayed repetition task comprised seven Slovak words (the dot indicates syllable boundaries): strč [str̩t͡ʃ] "stick", žblnkot [ˈʒbl̩n.kot] "fizz", šmrnc [ʃmr̩n̩t͡s] "twist", žgrloška [ˈʒgr.lɔʃ.ka] "curmudgeon", poštrngať [ˈpɔ.ʃtr̩n.gac] "clink glasses", tvŕdza [ˈtvr̩.d͡za] "dire straits", zvlchčiť [ˈzvl̩x.t͡ʃi̯c] "moisten". A female Slovak speaker recorded the words in a soundproof booth at 44.1 kHz sampling rate with 16-bit resolution. She read the words in a clear speaking style at a normal speech rate with a flat intonation. For the nonword repetition task, five Spanish/Basque nonwords were selected from the Syllabarium database [27]. All nonwords were phonotactically permissible in both languages. A female Spanish/ Basque speaker recorded the nonwords and the instructions. The recordings were then cut into single speech files, normalized for amplitude, and put in one audio file together with the Spanish instructions using the Praat speech editor [26]. There was one experimental list with an identical randomisation across all participants.

4.3. Procedure

Participants were tested individually in a lab. Prior to the testing session, which took about five minutes, they filled in an extensive language background questionnaire, and most of them also completed language interviews to assess their Basque and/or Spanish skills. This was the standard procedure for participants registered in the subject pool.

In the first task, participants heard nonwords of increasing syllable length hrough head-phones, one at a time. In the delayed imitation task, participants were asked to listen to each Slovak word presented to them and repeat it twice after a short instruction presented in Spanish, which delayed the imitation. After the task, participants were asked to guess the language of the speaker, with only three participants guessing correctly.

4.4. Data Coding and Analysis

One Slovak speaker and one phonetically trained research assistant transcribed the imitations of the second repetition of each word. The agreement was low (78.6%), and a third rater is currently evaluating the data. Here, we focus on schwa and /e/ insertions, which were annotated by a third rater for a subset of 102 participants. The agreement was high (98.59%), and the first author resolved the remaining cases. Two research assistants transcribed the nonword-repetition task, and a third rater resolved 4.6% cases of disagreement. We calculated the average accuracy of correctly remembered items per participant.

The analysis focuses on vowel insertion at word onsets and within words, in which insertions were expected (e.g. in /s/ + obstruent clusters). One word that did not start with a cluster (poštrngať) was excluded from the analysis (there were no onset insertions and only 10% word-medially, with /u/ being the predominant insertion in this word). Epenthesis was taken as the dependent variable and was coded in a binary fashion (present or not). The variable L1 use was the main predictor (note that it strongly correlated with L2 use: Spearman's $r(1321) = -.86$, $p < .001$). A mixed effect logistic regression model for binary responses was fitted to the data (glmer, implemented in the R package lme4, R Core Team, 2018 [28]). The full model included the main predictor and the control variables age and nonword repetition as fixed effects. Random intercepts for participants and items were included in both models. All continuous variables were centred and scaled.

5. RESULTS

Across all participants, the proportion of epenthesis was 48.2% (SD 23). The statistical model showed a main effect of L1 use (see Table 1), in that insertions increased with higher L1 use. None of the other two variables reached significance but there was a tendency for more insertions with increasing age. The result is in line with the hypothesis that bilingualism-related factors may predict production accuracy.

Fixed effects	*b*	*SE*	*z*	*p*
Intercept	0.281	0.451	0.622	0.534
L1 use	0.241	0.116	2.127	0.033
Age	0.229	0.117	1.955	0.051
Nonword rep.	0.075	0.115	0.658	0.510
Random effects				
σ^2	3.29			
$\tau_{00\ Subject.ID}$	1.46			
$\tau_{00\ word}$	1.14			
ICC	0.44			
N_{word}	6			
$N_{Subject.ID}$	189			
Observations	1134			
Marginal R^2 / Conditional R^2	0.019 / 0.404			

Table 1: Summary of the mixed effect model.

6. DISCUSSION

Foreign speech-sound learning exhibits substantial individual variability in adulthood. The aim of this study was to explore such differences during the very initial exposure to an L3 across Spanish and Spanish/Basque speakers. Schwa and /e/ insertions at the onset of a cluster and word-medially were used as proxies for articulation accuracy. The findings indicate an inverse relationship between L1 use and the number of vowel insertions, even after controlling for age and short-term memory, showing more insertions with an increasing L1 use. This suggests that active bilingualism may contribute to more accurate imitation of phonotactically complex words in an L3. Short-term memory did not affect vowel insertions, but there was a tendency for more insertions with increasing age, which is consistent with previous findings [4, 8].

The superior imitation skills during the initial stage of L3-phonetic learning as a result of active bilingualism cannot be easily explained by the linguistic and non-linguistic factors discussed in the introduction. Although learners' motivation or AoA were not accounted for in the present analysis, it is reasonable to assume that imitation skills are influenced by a complex interaction of multiple factors. In this study, the factor of interest was self-reported frequency of language use, which was operationalized as a continuous variable. Some may argue that such estimations are unreliable. However, recent studies on bilingual speech perception and production have shown that such approaches are valid. Language use has often been operationalized categorically using self-reports or by performance-based tasks, setting widely disparate grouping criteria. An alternative approach to capture the influence of additional languages on speech-related performance is to move away from categorical measures and instead examine input characteristics and the amount of use of different languages and varieties [30-32].

One advantage of the present design is that the effect of bilingualism may go beyond direct transfer from the bilingual's languages, because Spanish and Basque have similar phonological characteristics despite being typologically unrelated. Both languages prefer simple onsets with the maximal onset consisting of two CCs. However, there are some combinatorial restrictions within each language that require closer inspection, as they might affect overall performance. Dominant Basque speakers may have had advantages due to a larger number of sibilants and palatals compared to Castilian Spanish. Moreover, Castilian Spanish modifies /s/ + obstruent clusters by epenthesis in onset positions. Three of the Slovak clusters started with /s/ or /ʃ/, which could explain the higher frequency of epenthesis in speakers who primarily use Spanish. Indeed, an analysis limited to the onset epenthesis confirmed the effect of L1 use ($z = 2.55$, $p = .01$), refuting one of the reviewer's concerns that the present result are solely due to Slovak showing a tendency for an epenthetic schwa in clusters with syllabic consonants [34], which may be perceived vocalically by naïve listeners. All participants should be equally likely to perceive a schwa, but this was not the case.

Due to the time-intensive nature of transcriptions, the present analysis was limited to schwa and /e/ insertions. Further analyses will focus on phonetic distances based on features and factors such as proficiency in the additional languages of the participants. Replicating the effect reported here would provide stronger evidence for the role of bilingualism on imitation and would be consistent with previous studies that have shown facilitative effects of bilingualism on word learning [14-16]. It is reasonable to assume that active and proficient bilinguals have better imitation skills for L3 phonotactically complex words that arise as a result of experiential factors. The active use of multiple languages may enhance sensorimotor and perceptual flexibility, minimizing the influence of the other language articulatory habits. This explanation aligns with studies on foreign-accent imitation, where singers outperform instrumentalists due to their experience related to greater vocal flexibility [17].

In conclusion, these results suggest that imitation skills for complex words in an L3 may be predicted by bilingualism, as measured by daily use of phonologically but not typologically related languages.

7. ACKNOWLEDGMENT

This work was supported by the European Commission under the Marie Curie Action: Intra-European fellowships for career development awarded to AH. I would like to thank all my research assistants and

Sophia Wulfert for their invaluable help with data transcriptions and data pre-processing.

8. REFERENCES

[1] Strange, W. (ed.). 1995. *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*. Timonium, MD: York Press.

[2] Odlin, T. 1989. *Language Transfer: Cross Linguistic Influence in Language Learning.* Cambridge University Press.

[3] Eckman, F.R. 1977. Markedness and the contrastive analysis hypothesis. *Language Learning*, 27, 315-330.

[4] Ellis, R. 1994. *The Study of Second Language Acquisition.* Oxford University Press.

[5] Gardner, R.C. 1985. *Social Psychology and Second Language Learning. The Role of Attitudes and Motivation.* Edward Arnold.

[6] Guiora, A. 1990. A psychological theory of second language pronunciation. *Toegepaste Taalwetenschap in Artikelen* 37, 15-23.

[7] Birdsong, D. 2006. Age and second language acquisition and processing: A selective overview. *Language Learning* 56, 9-49.

[8] Flege, J.E. et al. 1999. Age constraints on second-language acquisition. *Journal of Memory and Language* 41, 78-104.

[9] Flege, J.E. 1995. Second language speech learning: theory, findings and problems. In W. Strange (eds.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Timonium, 233-277.

[10] Abu-rabia, S., Kehat, S. 2004. The critical period for second language pronunciation: Is there such a thing? *Educational Psychology* 24/1, 77-97.

[11] Birdsong, D. 2007. Nativelike pronunciation among late learners of French as a second language. In O.-S. Bohn & M. Munro (eds.), *Language Experience in Second Language Speech Learning*. Benjamins, 99-116.

[12] Bongaerts, T. 1999. Ultimate attainment in L2 pronunciation: The case of very advanced late learners. In D. Birdsong (ed.), *Second Language Acquisition and the Critical Period Hypothesis*. Erlbaum, 133–159.

[13] Hanulíková, A. et al. 2012. Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning* 62, 79–109.

[14] Kaushanskaya, M., Marian, V. 2009. The bilingual advantage in novel word learning. *Psychonomic Bulletin and Review* 16(4), 705–710.

[15] Antoniou, M. et al. 2015. The bilingual advantage in phonetic learning. *Bilingualism: Language and Cognition* 18(4), 683–695.

[16] Kieseier, T. 2021. Bilingual advantage in early EFL pronunciation accuracy of German 4th-graders. *International Journal of Bilingualism* 25(3), 708–726.

[17] Christiner, M., Reiterer, S. M. 2015. A Mozart is not a Pavarotti: Singers outperform instrumentalists on foreign accent imitation. *Frontiers in Human Neuroscience*, 9.

[18] Marini, A. et al. 2019. Impact of early second-language acquisition on the development of first language and verbal short-term and working memory. *International Journal of Bilingual Education and Bilingualism* 22, 165-76.

[19] Diaz, B. et al. 2008. Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *PNAS* 105/42, 16083-1688.

[20] Hanulíková, A., Ekström, J. 2017. Lexical adaptation to a novel accent in German: A comparison between German, Swedish, and Finnish listeners. *Proceedings of Interspeech* 2017, Stockholm, Sweden, 1784-1788.

[21] Hanulíková, A., Hamann, S. 2010. Illustrations of Slovak IPA. *Journal of the International Phonetic Association* 40, 373–378.

[22] Hanulíková, A., Dietrich, R. 2008. Die variable Coda in der slowakisch-deutschen Interimsprache. In M. Tarvas (Ed.), Tradition und Geschichte im literarischen und sprachwissenschaftlichen Kontext. Bern: Peter Lang, 119-130.

[23] Colina, S. 2012. Syllable structure. In Hualde, J. I. et al. (eds.), *The Handbook of Hispanic Linguistics*, Blackwell, 133-151.

[24] Egurtzegi, A. 2013. Phonetics and phonology. Basque and Proto-Basque. Language-internal and typological approaches to linguistic reconstruction. *Mikroglottika* 5.

[25] Malmberg, B. 1965. *Estudios de fonética hispánica*. Instituto Miguel de Cervantes.

[26] Boersma, P., Weenink, D. (2014). *Praat: Doing Phonetics by Computer* [Computer program]. Version 5.3.63, http://www.praat.org/.

[27] Duñnabeitia, J. A. et al. 2010. SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish syllables. *Behavioral Research Methods* 42, 118–125.

[28] Bates, D. et al. 2019. Package lme4 (Version 1.1-21) [Computer software]. https://cran.r-project.org

[29] De Cat, C. 2020. Predicting language proficiency in bilingual children. *Studies in Second Language Acquisition* 42, 279–325.

[30] Levy, H., Hanulíková, A. 2019. Variation in children's vowel production: Effects of language exposure and lexical frequency. *Laboratory Phonology* 10(1): 9.

[31] Levy, H., Hanulíková, A. 2023. Spot It and Learn It! Word Learning in Virtual Peer-Group Interactions Using a Novel Paradigm for School-Aged Children. *Language Learning* 73, 197-230.

[32] Levy, H. et al. 2019. Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience. *Journal of Child Language* 46(2), 368-392.

[33] Flege, J. E., Bohn, O-S. 2021. The revised Speech Learning Model (SLM-r). In R. Wayland (ed.), *Second language speech learning: Theoretical and empirical progress*. Cambridge University Press, 3– 83.

[34] Pouplier, M., Beňuš, Š. 2011. On the phonetic status of syllabic consonants: Evidence from Slovak. *Laboratory Phonology* 2, 243-273.

Perceptual representation and production of American English simplex vowels by adolescent Palestinian-Arabic learners

Vincent J. van Heuven[1, 2, 3] & Bashar M. M. Farran[1]

[1] Multilingual Doctoral School, University of Pannonia, Veszprém, Hungary
[2] Leiden University Centre for Linguistics, Leiden, The Netherlands
[3] Frisian Academy, Leeuwarden, The Netherlands
v.j.j.p.van.heuven@hum.leidenuniv.nl; bfarran89@gmail.com

ABSTRACT

To establish their perceptual representation of the American English (AE) vowels, 40 adolescent Palestinian-Arabic (3 long + 3 short vowels) learners of English as a foreign language and 20 American native controls identified each of 86 artificially generated vowel sounds (in /mVf/ nonsense items) as one of the 11 simplex vowels of AE (forced choice). F1 was varied in 7 steps of 1 Bark, F2 in 9 steps (20 impossible F1-F2 combinations were excluded). Vowel durations were either 200 or 300 ms. Results reveal large deviation of vowel centroids in the L2 responses (from L1 control centroids), and much greater overlap of spreading ellipses. The effect of duration was stronger in L2 than in L1 responses. The perceptual representations matched the locations of the AE vowels in the acoustic vowel diagram based on the same learners' production data, suggesting a close link between vowel perception and production.

Keywords: Perceptual representation, synthesized vowels, L1 interference, Arabic, American English.

1. INTRODUCTION

1.1. Importance of perceptual studies for L2 learning

When comparing the vowel systems of native (L1) and foreign (L2) languages, the usual procedure is to record speakers of the language variety of interest and then measure the lowest two to four resonances in the speaker's vocal tract as an indication of how each vowel is pronounced. The lowest resonance (formant F1), corresponds to the openness of the vowel, while the second-lowest formant (F2) is an indication of backness and lip rounding [1, 2]. Plotting the F1 and F2 values as vowel coordinates in a two-dimensional map then gives a good impression of the general organization of the vowel system. The mean (also called centroid) of the dispersion cloud of each vowel is taken as the most representative ('prototypical') realization of the particular vowel type. Vowel duration is often added as a third parameter to define the vowel space of the language (variety).

By comparing the target system with the learner's L1, differences and similarities in the organization of the respective vowel systems can be illustrated, potential learning problems can be identified, and instructions can be formulated inform the learner how to modify their native vowel category so as to articulate a more authentic vowel in L2 (e.g., [3]).

It is insufficiently realized, however, that studying the production of the vowel systems per se does not reveal the full organization of a vowel system, and – more importantly – does not reveal the (often incorrect) perceptual representation of the vowel system of the L2 in de mind of the learner. What is needed to appreciate the representation of the vowel system in the mind of the learner (and of the native speaker) is a perceptual mapping. Using perceptual techniques allows the researcher to establish so-called trading relationships between the parameters that define the individual listener's vowel space.

The present paper therefore aims to establish the perceptual representation of the vowels of American English (AE) in the minds of adolescent (secondary school) intermediate-level Palestinian-Arabic learners of English as a foreign language (EFL). We compare the perceptual representation with vowel production data from the same speakers. We test the hypothesis that the perceptual representation matches L2 production, and that L2 perception and production deviate in the same way from L1 norm data.

1.2. Vowels of Palestinian Arabic vs American English

The vowel system of Modern Standard Arabic (MSA) is generally analysed with 6 phonemes [4], i.e., corner vowels /i, a, u/ and contrastive length (Figure 1, left). Additionally, Palestinian Arabic (PA) has mid-vowel allophones [e, eː, o, oː], which may affect EFL learning [5: 171, 6: 529, 7: 1, 8]. American English is analyzed with 11 simplex vowels, i.e., 4 lax/short vowels /ɪ, ɛ, ʊ, ʌ/ and 7 tense/long vowels /i, e, æ, ɑ, ɔ, o, u/, of which /e/ and /o/ diphthongize (but not enough to consider them complex vowels [9: 114-115]), with positions in the IPA vowel diagram as in Figure 1, right. The short/lax vowels (joined by the inner quadrilateral in Figure 1) are articulated closer to the center of the vowel space.

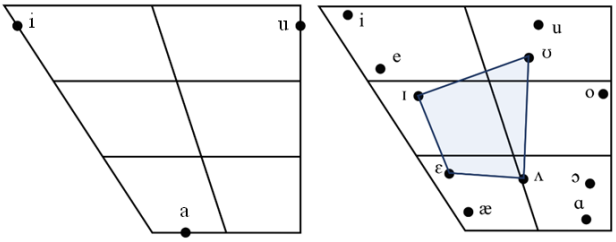


Figure 1: Vowel diagrams for Arabic (left, copied from [4]) and American English (right, after [10]).

2. METHODS

2.1. Stimuli

We used the universal reference set of synthesized vowel tokens proposed by [11] to map out the (oral) monophthongs of any language. The vowel space is defined by three parameters, i.e., vowel height (F1), vowel backness/rounding (F2) and length (vowel duration). The vowels were synthesized in an acoustic vowel triangle defined by the F1-by-F2 coordinates of extreme, prolonged vowel tokens pronounced as /mif/, /maf/ and /muf/, technically not cardinal vowels but contexted tokens which are a close approximation to the cardinals. To ensure auditorily uniform inter-vowel spacing, F1 and F2 stepsizes were always 1 Bark. In the grid, F1 values ran from 2.5 to 8.5 Bark, in 7 steps. F2 values were varied in 9 steps between 6 and 14 Bark. This yields 9 × 7 = 63 vowel spectra in a rectangular matrix (Figure 1). Twenty F1-by-F2 combinations sound inhuman (grey in Figure 2). When these are eliminated, a subset of 43 remain, which were synthesized with (ecologically valid) 200 and 300-ms vowel durations (for background and details, see [11, 12]).

F1			F2 (step number, Hz, Bark)								
			1.	2.	3.	4.	5.	6.	7.	8.	9.
			2357	2031	1746	1497	1278	1086	915	764	628
#	Hz	Bk	14.0	13.0	12.0	11.0	10.0	9.0	8.0	7.0	6.0
1.	237	2.5									
2.	339	3.5									
3.	447	4.5									
4.	565	5.5									
5.	694	6.5									
6.	838	7.5									
7.	998	8.5									

Figure 2: Steady-state F1 and F2 values for 43 reference vowel qualities.

2.2. Participants and procedure

Twenty male and 20 female native speakers of PA were selected as intermediate learners of English. They were secondary school pupils with a mean exposure to (American) English of roughly 6 years in a school setting. These participants listened to the 86 synthesized /mVf/ vowel stimuli over good-quality headphones in individual sessions. The participant saw 11 response buttons on a computer screen, each of which exemplified one of the 11 response vowels by a single keyword, which were supposed to be well-known to the students.

The participants decided, for each /mVf/ token, which of the 11 response vowels came closest to the sound they had just heard, by clicking on the corresponding response button. Stimulus presentation and data collection were controlled by a computer script written for the ExperimentMFC module in the Praat software [13, 14].

In an immediately preceding perceptual assimilation task, the PA participants had been exposed to two tokens of the 11 AE vowels in /hVd/ context spoken by two male native speakers of American English (4 tokens per vowel; see [12] for details).

The procedure was repeated with 20 native listeners of American English, bachelor students of linguistics at the University of Southern California in Los Angeles (7 males) with ages between 19 and 22.[1]

2.3. Collection of production data

The same 40 PA EFL learners produced the 11 AE vowels in everyday keywords as well as in /hVd/ contexts (twice) in a fixed carrier *Now say ... again*. Recordings were made directly (44.1 kHz, 16 bit) on a silent notebook computer in a quiet room, using a Sennheiser PC131 headset microphone. In the results section, we only use the recordings of the /hVd/ words: *heed*, *hid*, *hayed*, *head*, *had*, *hud*, *hod*, *hawed*, *hoad*, *hood*, *who'd*, for details see [14].

3. RESULTS

3.1. Centroids and ellipses in F1-by-F2 vowel space

The results of the identification of the 86 reference vowels in terms of the simplex vowels of AE are shown in Figure 3A-D for the L1 control listeners (top), and for the PA EFL learners (bottom), separately for the 43 short (left, panels A, C) and 43 long (right, B, D) stimulus vowels. The phonetic symbols mark the approximate location of the vowel centroid, while the dispersion ellipses are drawn at ± 1 SD away from the centroid along the first two principal components of the scatter cloud, theoretically including the 46% most typical responses. This presentation of the results facilitates direct comparison of the perceptual representation and the acoustic vowel production. A more traditional presentation of the results in terms of the modal response vowel for each of the stimuli (as used in, e.g., [15, 16]) can be found in [14].

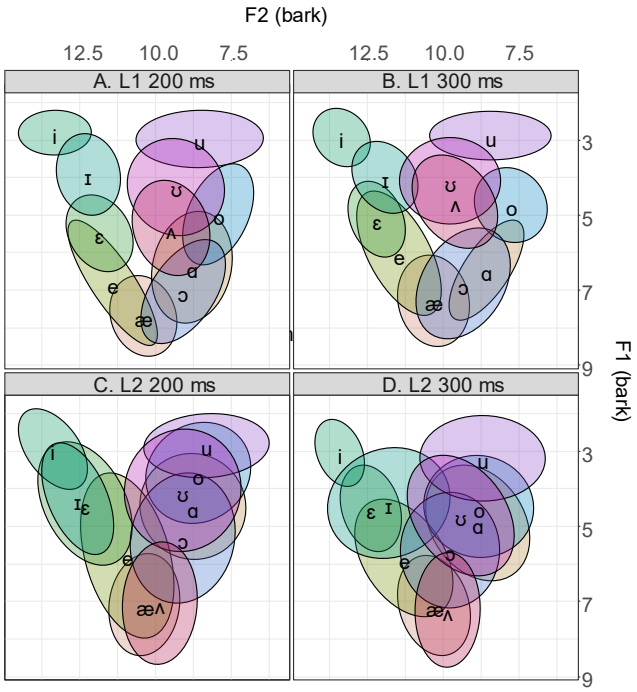


Figure 3. Centroids and dispersion ellipses in F1 by F2 (Bark) plot of American English simplex vowels as identified by American L1 listeners (top panels) and Palestinian Arabic EFL learners (bottom) for 43 synthesized reference vowels with durations of 200 ms (left panels) and 300 ms (right panels). Figure produced with Visible Vowels [17].

The L1 listeners maintain a clear difference between the tense-lax counterparts, /i/-/ɪ/ and /u/-/ʊ/. There is a large distance between the centroids associated with the members of these pairs, and there is only relatively little overlap between the associated spreading ellipses. The centroids of the lax vowels, especially of /ɪ, ʊ, ʌ/, are rather centralized when computed for the stimulus vowels with long duration, but assume more peripheral locations (closer to the tense counterparts) when heard with short duration. This would be a first indication that vowel quality and duration are in a trading relationship in the native speakers' mental representation of the tense-lax vowels. For a phonetically long vowel to be perceived as a lax member of a contrast, it has to be very clearly centralized. When a less centralized vowel is short (enough), it will still be perceived as lax. It is also apparent that the L1 listeners do not maintain distinct vowel categories for pair of tense vowels /ɑ-ɔ/ and for the lax vowels /ʊ-ʌ/. Many speakers, especially in California, no longer distinguish between /ɔ/ and /ɑ/, characteristic of the *cot–caught* merger; also, /ʊ/ is moving towards [ʌ] so that, for example, *book* and *could* start to sound like *buck* and *cud* (e.g., [1]: 212–213).

The L2 participants correctly placed the point vowels /i, u, æ/ at the left-top, right-top and mid-bottom parts of the F1-by-F2-plane, respectively. These vowel qualities were correctly interpreted, presumably because they have suitable near-equivalents in Arabic. Notice that, in the L2 identifications, the centroid for tense /u/ is further back (lower F2 value) than for all other back vowels. In the L1 data, the /u/ centroid is rather more fronted, which corresponds well with the literature on AE.

While the L1 vowel centroids are more or less evenly distributed over the F1-by-F2 space (with the exceptions noted above), substantial clustering of vowels is observed in the L2 identifications. The lax vowels /ɪ, ɛ/ are virtually indistinguishable, as are the members of the /æ, ʌ/ pair. Moreover, the remaining back vowels, /ɑ, ɔ, o, ʊ/ tend to cluster. Typically, the locations of the /ɑ, ɔ/ centroids are too high, while the lax vowel centroids for /ʊ, ʌ/ are located close to the back edge of the vowel space.

3.2. Vowel duration

To obtain an indication of the difference in weight attached to the duration of the stimulus vowel in the perceptual identification, we computed the relative frequencies of the 11 response vowel categories for short (200-ms) and long (300-ms) stimulus vowels (across all 43 synthesized qualities), expressed as a percentage of the total number of responses, i.e., 860 (43 × 20) for L1 responses and 1720 (43 × 40) for L2 responses. These percentages are shown in Figure 4.

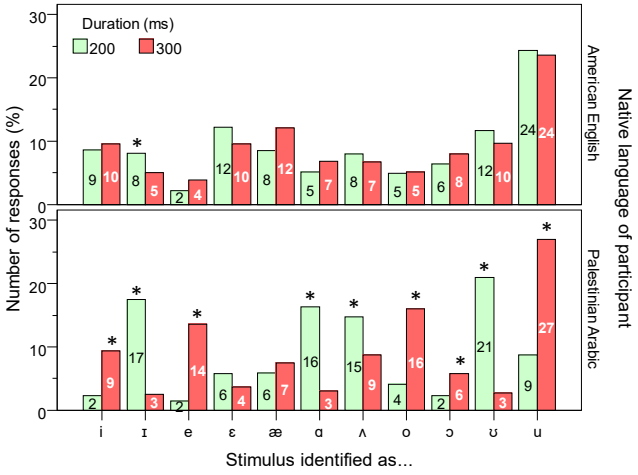


Figure 4. Frequency (%) of 11 response categories for short (200 ms) and long (300 ms) stimulus vowel durations, irrespective of formant structure, for L1 AE (upper panel) and L2 PA listeners (bottom panel). See text for more information.

A significant difference due to stimulus length is seen only for one vowel in the L1 responses, i.e., where the number of /ɪ/ responses is significantly smaller for long stimuli than for short stimuli (all discrepant pairs are marked by an asterisk in Figure 4; binomial test, assuming a 50-50% distribution, $\alpha \leq .05$). For all other vowels, stimulus duration does not affect the response frequencies significantly. For the PA L2

listeners, however, the difference in duration has a significant effect for 8 out of 10 response categories, i.e., for all vowels except /ɛ/ and /æ/.

3.3. Comparison with production data

Figure 5A-D shows the location of the centroids in the F1-by-F2 space found in the perceptual representation in section 3.1, and in the production data recorded from the same PA participants (see section 2.3). The L1 production data have been copied from the literature [18, 19] and, though fully comparable, were produced by different individuals than the participants in the perception experiment.[2]

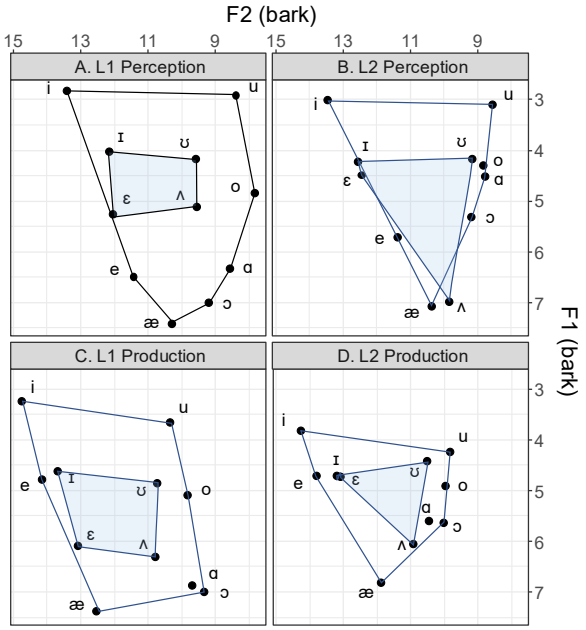


Figure 5. Centroids of 11 AE vowels in F1-by-F2 plane (Bark) found for perceptual representation (top panels) and in speech production (bottom panels) for L1 speakers (left) and for PA EFL learners (right). See text for more information.

The vowel configuration in the production data is shifted away from the origin of the plot relative to the perceptual representation. The latter is based on listeners' responses to a synthesized male voice, whereas the production centroids have been averaged over equal numbers of male and female speakers, which increases the F1 and F2 frequencies due to the shorter female vocal tract. In spite of this, the two L1 configurations (panels A, C) are quite similar. The four short/lax vowels are the corner points of an inner quadrilateral with clear distances from each other and from the nearest long/tense vowels. In the PA EFL configurations, the configurations are rather more triangular. The /ɪ-ɛ/ contrast is absent in both the perceptual representation and in production. Also, The L2 vowel space is compressed in the height dimension in comparison with the L1 configuration, in panel B, and especially in the L2 vowel production (panel D). The compression is most likely caused by the absence of contrastive mid vowels in the EFL learners' native language. Lack of separation between adjacent back vowels is seen in both L2 perception and production.

4. CONCLUSIONS AND DISCUSSION

This study has shown that the perceptual representation of the simplex vowels of American English can be adequately mapped out by having listeners identify a small set of artificial (contexted) vowel sounds varying systematically in F1, F2 and duration. The PA non-native perceptual representation deviates strongly from the L1 configuration, and the non-native categories are less clearly delineated (larger spreading ellipses). The L1 listeners identify the AE vowels mainly on the basis of quality (F1, F2) and tend to ignore vowel duration, as has been shown earlier by artificially lengthening and shortening natural AE vowels [20]. The PA non-native listeners were strongly led by vowel duration, which is contrastive in Arabic but not in English.

As predicted, we found a strong match between the perceptual representation of the AE vowels and the acoustic properties in their production, both for the L1 and for L2 participants at the group level.[3]

The present study was not set up to shed light on the question whether perception leads production or *vice versa*. This would have required following the participants longitudinally. Still, our study shows that production studies should be complemented by research on the perceptual representation of target sounds. The greater reliance on vowel duration by the non-natives cannot be seen in the L2 production data.

The L2 vowel durations (not presented in this paper but see [12, 14]) show less contrast between the AE lax/short vs tense/long vowels than is seen in the L1 production data. Moreover, the L2 speakers produced much shorter vowel durations overall than the L1 speakers, most likely because the L2 speakers failed to apply vowel lengthening before the voiced coda consonant in the /hVd/ words. Similar shorter overall vowel duration in L2 AE speech has been observed for EFL learners whose L1 is Javanese or Sundanese [21, 22], or Persian L1 [23] or some other variety of Arabic than PA [24].

We end with a word of caution. Our perceptual labelling method will be rather more complicated when also used to study dynamically changing vowel sounds such as diphthongs. Diphthongal trajectories can be synthesized with no problem but the number of stimuli will soon become unmanageable.

7. REFERENCES

[1] Ladefoged, P., Johnson, K. 2011. *A course in Phonetics*. Wadsworth.

[2] Ladefoged, P., Disner, S. 2012. *Vowels and consonants*. Wiley-Blackwell.

[3] Yavaş, M. 2011. *Applied English phonology*. Wiley-Blackwell.

[4] Thelwall, R., Sa'Adeddin, M. A. 1999. Illustrations of the IPA: Arabic. *Handbook of the IPA*. Cambridge University Press, 51–54.

[5] Palva, H. 1965. *Lower Galilean Arabic: An analysis of its anaptyctic and prothetic vowels with sample texts*. Studia Orientalia.

[6] Shahin, K. 2011. Palestinian Arabic. In: Versteegh, K. Eid, M., Elgibali, A., Woidich, M., Zaborski, A. (eds.), *Encyclopedia of Arabic language and linguistics, Vol.* 3. Brill, 526–538.

[7] Amir, N., Amir, O., Rosenhouse, J. 2014. Colloquial Arabic vowels in Israel: A comparative acoustic study of two dialects. *J. Acoust. Soc. Am.* 136, 1895–1907. doi: 10.1121/1.4894725

[8] Hall, N. 2017. Phonetic neutralization in Palestinian Arabic vowel shortening, with implications for lexical organization. *Glossa* 2, 1–23. doi: 10.5334/gjgl.257]

[9] Celce-Murcia, M., Brinton, D. M., Goodwin, J. M. Griner, B. 2010. *Teaching pronunciation: A course book and reference guide*. Cambridge University Press.

[10] Ladefoged, P. 1999. Illustrations of the IPA: American English. *Handbook of the IPA*. Cambridge University Press, 41–44.

[11] van Heuven, V. J., Afshar, N., Disner, S. 2020. Mapping perceptual vowel spaces in native and foreign language: Persian learners of English compared with American native speakers. In: Bátyi, S., Lengyel, Z. (eds.), *Bilingualism: Hungarian and non-Hungarian context. Studies in honor of Judit Navracsics*. Pannon Egyetem, 113–130. http://hdl.handle.net/1887/3146383

[12] van Heuven, V. J., Farran, B. M. (2022). The role of duration in the articulation and mental representation of American English vowels by Palestinian Arabic learners of English. 23rd Summer School of Psycholinguistics, Veszprém, www.researchgate.net/publication/361570474

[12] Afshar, N., van Heuven, V. J. 2022. Perceptual assimilation of English vowels by monolingual and bilingual learners in Iran. *Argumentum* 18, 172–191. doi: 10.34103/ARGUMENTUM/2022/9

[13] Boersma, P., Weenink, D. 2019. Praat, A system for doing phonetics by computer. www.praat.org

[14] Farran, B. M. 2022. The production and perception of American English sounds by Palestinian Arabic adolescents. PhD. diss., University of Pannonia. doi: 10.18136/PE.2022.820

[15] Hombert, J.-M., Puech, G. 1984. Espace vocalique et structuration perceptuelle: application au Swahili. *Pholia* 1, 199–208. http://www.ddl.cnrs.fr/Download/Pholia/Pholia_N-1.pdf

[16] van Zanten, E., van Heuven V. J. 1984. The Indonesian vowels as pronounced and perceived by Toba Batak, Sundanese and Javanese speakers. *Contr. Royal Inst. Anthrop*. 140, 497–521. doi: 10.1163/22134379-90003411

[17] Heeringa, W., Van de Velde, H. 2018. Visible Vowels: A tool for the visualization of vowel variation. *Proc. CLARIN 2018*, 120–123. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf

[18] Wang, H. 2007. *English as a lingua franca. Mutual intelligibility of American, Chinese and Dutch speakers of English*. LOT. https://www. lotpublications.nl/documents/147_fulltext.pdf

[19] Wang, H., van Heuven, V. J. 2006. Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. In: van de Weijer, J. M., Los, B. (eds.), *Linguistics in the Netherlands 2006*. John Benjamins, 237–248. doi: 10.1075/avt.23.23wan

[20] Hillenbrand, J. M., Clark, M. J., Houde, R. A. 2000. Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.* 108, 3013–3022. doi: 10.1121/1.1323463

[21] Perwitasari, A., Klamer, M., and Schiller, N. O. 2017. Quality of Javanese and Sundanese vowels. *J. Southeast Asian Ling. Soc.* 10, 1–9. http://hdl.handle.net/10524/52406.

[22] Perwitasari, A. 2018. *The acquisition of English vowels by Javanese and Sundanese native speakers*. LOT. https://www.lotpublications.nl/Documents/532_fulltext.pdf

[23] Afshar, N. 2022. Production and perceptual representation of American English vowel sounds by monolingual Persian and early bilingual Azerbaijani-Persian adolescents. PhD. diss., University of Pannonia. doi: 10.13140/RG.2.2.36272.02563

[24] Munro, M. J. 1993. Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Lang. Speech* 36, 39–66. doi: 10.1177/002383099303600103

[1] The authors are grateful to professor Sandra Ferrari Disner of the Linguistics Department at USC for making the materials and software to do the vowel identification experiment available to her students, running a quick survey of the participants' linguistic background and for collecting and transmitting the data files.

[2] We thank professor Hongyan Wang of Shenzhen University (P. R. China), who made her vowel data available to us. No results for the tokens of *hawed* are given in [18, 19] but the formant and duration values for the *hawed* tokens were included in the underlying dataset.

[3] Correlation of production and perceptual representation was considerably poorer at the level of individual speakers (for details and statistical tests, see [14])

Different realizations of German underlying vs. derived diphthongs between two generations of Cantonese L1 speakers

Michelle Vuong, Marieke Einfeldt, Bettina Braun

Department of Linguistics, University of Konstanz, Germany
michelle.vuong@uni-konstanz.de, marieke.einfeldt@uni-konstanz.de, bettina.braun@uni-konstanz.de

ABSTRACT

The theory of cross-linguistic influence predicts that languages influence each other, leading to phonetic realizations that differ between monolingual und bilingual speakers. This paper tested heritage speakers, examining the influence of Cantonese on the realization of two kinds of German diphthongs, underlying ([aɪ̯], [aʊ̯]) and derived ([iːɐ̯], [uːɐ̯], due to /ʁ/-vocalization), by first- and second-generation Cantonese speakers living in Germany (Mannheim area). A delayed imitation paradigm was used to reduce effects of orthography. We predicted that underlying diphthongs would result in smaller differences across generations than derived diphthongs. Formant tracks of first and second formants were analyzed using general additive mixed models. Results largely confirmed the predictions (significant differences between generations in F2 only for underlying diphthongs, but both for F1 and F2 for derived diphthongs). Second generation heritage language speakers were more similar to a German control group from the same region for some diphthongs.

Keywords: diphthongs, formants, Cantonese, German, general additive mixed models

1. INTRODUCTION

Heritage speakers (HSs) are early bilinguals who acquire two first languages (L1): the majority language (ML) of the society they live in and a minority language, also called heritage language (HL) [1]. HSs form a heterogenous group as they often differ from one another in terms of age of onset, proficiency, language use, etc. Most studies on HLs have investigated HSs' knowledge and performance in the HL [2–5]. Less research focused on the effects of the HL on the ML, and if it did, the ML was often English. This paper investigates the influence of Cantonese (HL or L1) on the ML German for Cantonese L1 speakers in Germany. While there are a number of studies on the influence of Cantonese on English (and vice versa) [5, 6, 7], most probably because of the language situation of Hong Kong (L1 Cantonese L2 English), the language combination Cantonese–German is largely underexplored.

To track the developmental trace of HSs, we compared two generations of Cantonese L1 speakers living in Germany, with a focus on diphthongs. Cantonese has a large inventory of vowels, with eight contrastive monophthongs and ten contrastive diphthongs (/eɪ̯, aɪ̯, aːɪ̯, ɔɪ̯, uɪ̯, aʊ̯, aːʊ̯, iʊ̯, oʊ̯, œʏ̯/) [8]. The German inventory consists of fifteen contrastive monophthongs and three contrastive, closing, diphthongs (/aɪ̯, aʊ̯, ɔʏ̯/). We focus on two kinds of German diphthongs, those that are part of the phonological system of German (*underlying*, phonemic /aɪ̯/ and /aʊ̯/, which are more frequent than /ɔʏ̯/, cf. [9]) and those with a reversed vowel quality that are derived by a phonological rule of /ʁ/-vocalization. This rule changes underlying "r" in coda-position to [ɐ̯], e.g. "*vier*" [fiːɐ̯] 'four', which is diphthongized with the preceding vowel in many German dialects. We call this class *derived* diphthongs and test [iːɐ̯] and [uːɐ̯].

We hypothesize that the two generations of Cantonese L1 speakers realize the underlying diphthongs more similar to each other than the derived diphthongs. This hypothesis is based on two considerations: First, the closing diphthongs occurring in the underlying category (in which the second vowel target has a higher tongue position than the first, e.g. [aɪ̯]) are universally less marked than the opening diphthongs in the derived category (e.g. [iːɐ̯]). English and Cantonese, for instance, lack opening diphthongs. Research on the role of markedness in cross-linguistic influence has shown that markedness can affect rate and difficulty of acquisition: Generally, unmarked structures are acquired easier and faster (leading to positive cross-linguistic influence) than marked structures that tend to not be transferred at all [10]. Second, underlying diphthongs are represented in German orthography ([aɪ̯]: "ei", [aʊ̯]: "au", [ɔɪ̯]: "eu"), while derived diphthongs are not, which may influence language acquisition [11].

We analyzed the differences in the realization of the two kinds of German diphthongs between the two generations of Cantonese speakers using general additive mixed models (GAMMs). GAMMs allow a direct comparison between formant contours since they can model non-linear dependencies over time using smooth functions [12-17].

2. EXPERIMENT

We collected productions using a delayed imitation paradigm with a 2500ms interval between stimulus and imitation. After this delay, the phonetic trace has decayed, and participants need to access phonological representations [18, 19].

2.2. Methods

2.2.1 Participants

Thirty L1 speakers of Cantonese in Germany participated in the experiment. They were divided into four categories, split by gender and generation (Gen 1, Gen 2), see Table 1. Gen 1 included L1 speakers of Cantonese who were born in Vietnam (N = 11) or Hong Kong (N = 1); they all immigrated to Germany around puberty (mean age of immigration = 16.0, SD = 3.4 years). Gen 2 are HSs of Cantonese who were all born and raised in Germany by the Gen 1 participants. For them, German is the ML.

Gen 1 consists of six female (mean age = 54.0, SD = 2.1 years) and six male speakers (mean age = 58.0, SD = 2.4 years), while Gen 2 consists of seven female (mean age = 26.0, SD = 4.2 years) and eleven male speakers (mean age = 23.0, SD = 4.0 years). The majority of Gen 1 and Gen 2 participants were from Rhineland-Palatinate (N = 25). The remaining participants were from Baden-Württemberg (N = 3), Hesse (N = 1) and North Rhine-Westphalia (N = 1). The areas the participants came from all possess the process of coda /ʁ/-vocalization.

Each of the four speaker groups was matched with a baseline of two native speakers of German (GER 1, GER 2), who did not acquire a second language before the age of six and grew up in the same area, see Table 1. The GER 1 speakers had an average age of 54.3 years, SD = 1.9, the GER 2 speakers an average of 21.5 years, SD = 2.1. We could not control the educational status between Gen 1 and Gen 2 (higher degrees for Gen 2 than for Gen 1).

Group	Female	Male	**Total**
Gen 1	6	6	**12**
Gen 2	7	11	**18**
GER 1 (control)	2	2	**4**
GER 2 (control)	2	2	**4**
Total	**17**	**21**	**38**

Table 1: Distribution of participants across groups.

2.2.2 Material

A total of 24 items were constructed, half containing the underlying diphthongs [aɪ̯] and [aʊ̯], half the derived diphthongs [iːɐ̯] and [uːɐ̯]. For each of the four diphthongs, six non-words (e.g. ['paɪ̯pɐ]) were constructed and embedded in the same carrier sentence: *Er hat Peiper gesagt* ("He said Peiper"). The non-words were constructed to fit German and Cantonese phonotactics. Non-words were chosen over words to avoid effects of familiarity and lexical frequency, which may influence productions [17, 20]. The diphthongs were flanked by voiceless plosives (/p, t, k/) to ease segmentation.

The stimuli for imitation were recorded by a male speaker of GER 1. He read each of the 24 test sentences out loud three times. The production with the clearest pronunciation was chosen. Care was taken to have similar intonation contours across items (L+H* on target word, followed by low boundary tone L-%).

2.2.3 Procedure

The only task for participants from GER 1 and 2 was reading the list of target sentences out loud, recorded directly onto an Acer Switch Alpha 12 computer using an AT2020USB+ microphone (16 Bit, 44.1 kHz). Participants from Gen 1 and Gen 2 first filled in a background questionnaire programmed via *SoSci Survey* [21], ran on an in-house server. They were asked for their age, year of immigration to Germany (if applicable), place of residence, and education. They were then instructed to rate their proficiency in Standard High German, the Palatinate (or other) dialect(s), Cantonese and Vietnamese. Lastly, the participants were asked which languages they speak at home and outside home and to what proportion. They could take part in a lottery for reimbursement.

The main part of the study was the delayed imitation experiment programmed using the software ***Presentation***® (Neurobehavioral Systems), installed on an HP NB ProBook 650 G2. The majority of participants (N = 26) was tested in a quiet environment in a private setting in presence of the first author. Participants' imitations were recorded using an AT2020USB+ microphone. Four participants took part remotely as they were not available for testing on-site. In those cases, a video of the experiment was screenshared and imitations were recorded on smartphone microphones. This process was monitored remotely by the first author. Participants were instructed to imitate the utterance in their usual voice. Four practice trials familiarized the participants with the task. The main experiment consisted of 24 trials. Each trial started with a 500ms sine tone played at 300Hz, followed by 1500ms silence. Next, the sentence was played three times with 1500ms silence between each time. This was followed by 2000ms silence and a 500ms sine tone played randomly at either 150Hz or 450Hz to avoid

predictability and to overwrite acoustic traces, as in [22, 23]. Participants then pressed "A" on the laptop keyboard to record themselves and "L" to stop the recording. After stopping the recording, participants were automatically directed to the next trial and the recording was saved onto the laptop (16 Bit, 44.1 kHz). The order of trials was randomized.

2.2.4 Data Treatment and Statistical Analysis

First, all 912 sound files (38 participants x 24 imitations) were annotated semi-automatically using *Web-MAUS* [24]. Next, diphthong boundaries were manually corrected using a wide-band spectrogram and following standard segmentation criteria [25]. We used *emuR* [26] to create a database combining all annotated items and merged the database with the participants' meta data. The function get_trackdata() extracted the formants (separately for male and female participants); these were time-normalized to get 10 formant points per diphthong. The formants were converted to bark [27].

The basic model included a smooth over normalized time and smooth terms for participants and items [28, 29]. Next, we included smooth terms for the factors *gender* and *generation* and kept them if this improved the model fit. Model comparison was done using the function compareML() [13, 15, 30]. Since there is a strong autocorrelation between subsequent formant points, we used the autocorrelation parameter *rho* to correct for this. *Rho* was determined by the acf_resid()-function [30]. Models were checked using gam.check() to see if the number of knots was appropriate and the residuals normally distributed. All F2 models were re-run with the scaled-t-distribution (family="scat"), because the default gaussian distribution resulted in a highly tailed distribution of residuals [14]. Difference plots (across generations) are used to determine and visualize the parts of the diphthong in which significant differences occurred.

2.3 Results

The smooth terms for gender and generation always improved the model fit. Model comparisons showed no significant improvement for adding an interaction smooth between generation and gender. Figure 1 shows the averaged difference curve (solid line) for F1 and F2 for the **underlying diphthongs** ([aɪ̯] top panels, [aʊ̯] bottom panels) between Gen 1 and Gen 2. The difference curve shows the subtraction of the Gen 1 contour from the Gen 2 contour, revealing the parts in which they significantly differ from each other (areas in which the gray shading of the 95% Confidence Interval (CI) excludes 0). These areas are additionally marked by the vertical red lines. F1 of [aɪ̯] and [aʊ̯] showed no significant differences, while F2 of [aɪ̯] is significantly different from 0.17 to 0.28 and from 0.61 to 1 of the normalized time. Main significant differences in F2 of [aʊ̯] are from 0 to 0.08 and from 0.66 to 0.85 normalized time.

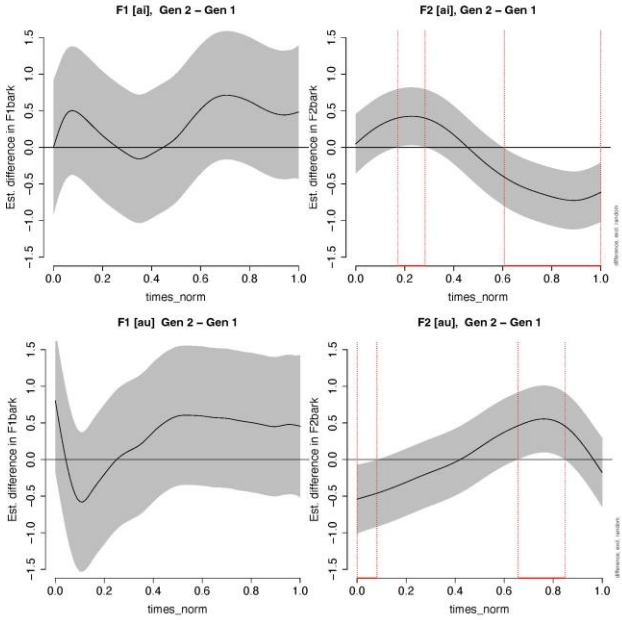


Figure 1: Estimated differences between Gen 2 and Gen 1 for F1 and F2 of [aɪ̯] (top panels) and [aʊ̯] (bottom panels). Red areas show significant differences.

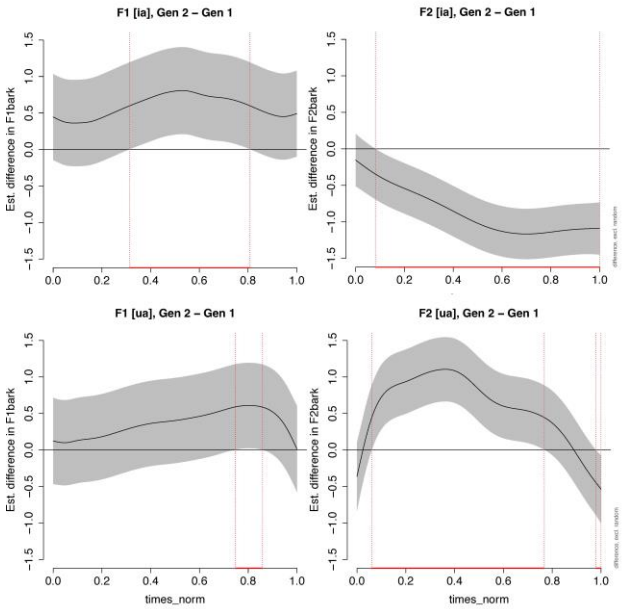


Figure 2: Estimated differences between Gen 2 and Gen 1 for F1 and F2 of [iːɐ̯] (top panels) and [uːɐ̯] (bottom panels). Red areas show significant differences.

Figure 2 displays the averaged difference curve between Gen 1 and Gen 2 for F1 and F2 of the **derived diphthongs** ([iːɐ̯] top panels, [uːɐ̯] bottom panels). [iːɐ̯] and [uːɐ̯] show significant differences in both F1 and F2: from 0.31 to 0.81 for F1 and 0.08 to 1 for F2 of [iːɐ̯] and from 0.75 to 0.86 for F1 and from 0.06 to 0.77 and 0.98 to 1 for F2 of [uːɐ̯]. Comparison

between Figures 1 and 2 suggest that F2 differences between generations are considerably larger in derived than in underlying diphthongs.

Figure 3 compares the L1 Cantonese speakers to the monolingual German control group (GER, merging data from GER 1 and GER 2). This comparison may be informative to investigate whether the Gen 2 speakers are closer to the German speakers than Gen 1 speakers (due to the small sample of GER speakers, a statistical analysis with GAMMs could not be conducted). Squares with dotted lines highlight formant trajectories, for which Gen 1 speakers were closer to the German controls than Gen 2 speakers (n = 3), squares with dashed lines highlight formant trajectories for which Gen 2 speakers are closer to German controls (n = 5). For the other formant trajectories, no pattern could be discerned. This exploratory analysis suggests that Gen 2 speakers were closer to the monolingual Germans than Gen 1.

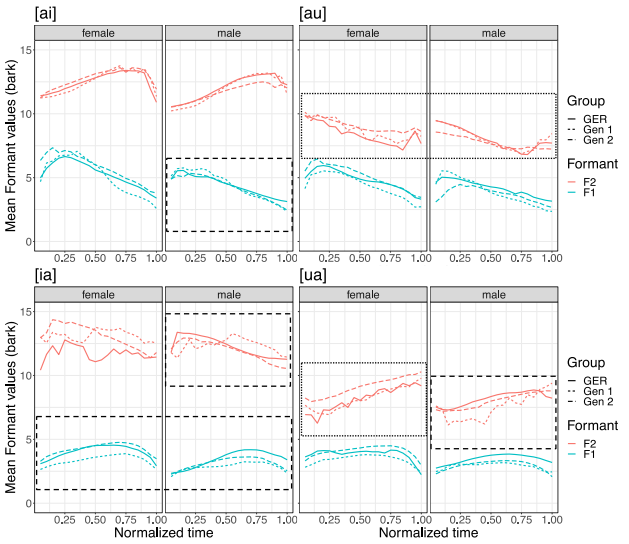


Figure 3: F1 and F2 contours of Gen 1, Gen 2 and GER for female and male for [aɪ̯] (top left panel), [aʊ̯] (top right), [iːɐ̯] (bottom left) and [uːɐ̯] (bottom right).

3. DISCUSSION

We examined inter-generational differences between Cantonese L1 speakers of German in their realizations of German underlying ([aɪ̯], [aʊ̯]) vs. derived ([iːɐ̯], [uːɐ̯]) diphthongs. We found significant differences between generations for both kinds of diphthongs in the realization of F2 (which mostly represents the horizontal tongue position, front-back). Moreover, there were fewer differences for the underlying diphthongs (differences only in F2 trajectory) than for the derived diphthongs (differences in both F1 and F2 trajectories). Furthermore, differences in F2 were larger and spanned a longer time interval of the diphthong for the derived compared to the underlying diphthongs. Comparison between Gen 1, Gen 2 and the GER baseline showed that Gen 2 was more similar to GER than Gen 1 in more cases, but there were too little data for statistical comparison.

Our findings support the hypothesis that the two generations of L1 Cantonese speakers in Germany realize the underlying diphthongs more similar to each other than the derived ones. This hypothesis was based on the fact that the closing diphthongs ([aɪ̯], [aʊ̯]) are less marked than the opening diphthongs ([iːɐ̯], [uːɐ̯]). Moreover, the German derived opening diphthongs [iːɐ̯], [uːɐ̯] are generally marked and Cantonese does not have opening diphthongs at all (and does not possess /ʁ/-vocalization). At the same time, the German underlying, closing diphthongs [aɪ̯], [aʊ̯] may be similar to the Cantonese diphthongs /ai̯, aːi̯, au̯, aːu̯/, possibly leading to positive influence from Cantonese to German.

The research hypothesis was also based on the fact that underlying diphthongs are represented in German orthography ([aɪ̯]: “ei”, [aʊ̯]: “au”), whereas the grapheme-phoneme correspondence of derived diphthongs is more opaque (e.g. [iːɐ̯]: “ier”, “ir”, “ihr”). Research on the effect of orthography on L2 production has shown that L2 orthographic forms influence the acquisition of L2 phonology. For instance, L2 speakers of American English pronounce the flap /ɾ/ as [t] when spelled with <t> (*beauty*) and as [d] when spelled with <d> (*lady*) [31].

From a perception perspective, many of the derived diphthongs ([iːɐ̯], [uːɐ̯]) by Gen 1 participants sounded like monophthongs ([i] and [u] respectively). This observation during annotation would need to be confirmed in a proper perception experiment. This would yield a more complete picture for the question how foreign language learners of German (here Gen 1) internalize and process /ʁ/-vocalization. Conceivably, differences in acquisition may also be expected for other phonological rules which might have implycations for acquisition theories and teaching.

4. CONCLUSION

Two generations of Cantonese L1 speakers of German were tested on the realization of two kinds of German diphthongs (underlying vs. derived). Results showed fewer and smaller differences between generations for the underlying, closing diphthongs ([aɪ̯], [aʊ̯]) than for the derived, opening diphthongs.

5. ACKNOWLEDGEMENTS

We thank Oleksiy Bobrov for programming the experiment and Bernhard Brehmer for advice and literature suggestions on heritage languages.

6. REFERENCES

[1] Kan, R. T. Y. 2021. Phonological Production in Young Speakers of Cantonese as a Heritage Language. *Language and speech* 64, 73-97.

[2] Kupisch, T. et al. 2014. Foreign Accent in Adult Simultaneous Bilinguals. *Heritage language journal* 11, 123–150.

[3] Lloyd-Smith, A., Einfeldt, M., Kupisch, T. 2020. Italian-German bilinguals: The effects of heritage language use on accent in early-acquired languages. *International Journal of Bilingualism* 24, 289–304.

[4] Cummings Ruiz, L. D. 2019. North Midland /u/-fronting and Its Effects on Heritage Speakers of Spanish. *Proc. 19th ICPhS* Melbourne, 1099–1103.

[5] Tse, H. 2016. Variation and change in Toronto heritage Cantonese. *Asia-Pacific Language Variation* 2, 124–156.

[6] Tse, H. 2019. *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Dissertation, University of Pittsburgh, Pittsburgh.

[7] Soo, R., Monahan, P. J. 2022. Phonetic and Lexical Encoding of Tone in Cantonese Heritage Speakers. *Language and speech*.

[8] Chan, A. Y. W., Li, D. C. S. 2000. English and Cantonese Phonology in Contrast: Explaining Cantonese ESL Learners' English Pronunciation Problems. *Language, Culture and Curriculum* 13, 67–85.

[9] Pätzold, M., Simpson, A. P. 1997. *Acoustic analysis of German vowels in the Kiel Corpus of Read Speech*. University of Kiel.

[10] Jarvis, S., Pavlenko, A. 2010. *Crosslinguistic influence in language and cognition*. London, New York: Routledge.

[11] Mitterer, H. 2021. *The role of orthography in learning a second language: Evidence from Maltese English*. Malta Chamber of Scientists.

[12] Baayen, R. H., van Rij, J., de Cat, C., Wood, S. 2018. Autocorrelated Errors in Experimental Data in the Language Sciences: Some Solutions Offered by Generalized Additive Mixed Models. In: D. Speelman, K. Heylen, D. Geeraerts (eds.) *Mixed-Effects Regression Models in Linguistics*, Springer International Publishing, 49–69.

[13] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70, 86–116.

[14] van Rij, J. , Hendriks, P., van Rijn, H., Baayen, R. H., Wood, S. N. 2019. Analyzing the Time Course of Pupillometric Data. *Trends in hearing* 23.

[15] Sóskuthy, M. 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics* 84.

[16] Voeten, C. C. 2021. Regional variation in ongoing sound change: The case of the Dutch diphthongs. *Journal of Linguistic Geography* 9, 162–177.

[17] Beaman, K. V., Tomaschek, F. 2021. Loss of Historical Phonetic Contrast across the Lifespan. in Buchstaller, I., Beaman, K. V. (eds.) *Language Variation and Language Change Across the Lifespan*, Routledge, 209–234.

[18] Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

[19] Baddeley, A. D. (2003). Working memory and language: An overview. *J. Commun. Disord.* 36, 189–208. doi: 10.1016/S0021-9924(03)00019-4.

[20] Hay, J., Jannedy, S., Mendoza-Denton, N. 1999. Oprah and /ay/: Lexical frequency, referee design and style. *Proc. 14th ICPhS* San Francisco, 1389–1392.

[21] Leiner, D. J. 2019. *SoSci Survey* (Version 3.2.55 [Computer software] Available at https://www.soscisurvey.de.

[22] Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., Braun, B. 2022. Three Kinds of Rising-Falling Contours in German wh-Questions: Evidence From Form and Function. *Frontiers in Communication* 7.

[23] Zahner-Ritter, K., Zhao, T., Einfeldt, M. Braun, B. 2022. How experience with tone in the native language affects the L2 acquisition of pitch accents. *Frontiers in psychology* 13.

[24] Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer speech & language* 45, 326–347.

[25] Turk, A., Nakai, S., Sugahara, M. 2006. Acoustic Segment Durations in Prosodic Research: A Practical Guide. In: *Methods in empirical prosody research*, New York, Berlin, 1–28.

[26] Winkelmann, R., Jaensch, K., Cassidy, S., Harrington, J. 2021. *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.3.0.

[27] Traunmüller, H. 1990. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America* 88, 97–100, 1990.

[28] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 3–36.

[29] Wood, S. N. 2017. *Generalized additive models: an introduction with R,* 2nd ed. London, New York, Boca Raton: CRC Press, Taylor & Francis Group.

[30] van Rij, J., Wieling, M., Baayen, R., van Rijn, H. 2022. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.4.1.

[31] Vokic, G. 2011. When alphabets collide: Alphabetic first-language speakers' approach to speech production in an alphabetic second language. *Second language research* 27, 391–417.

EFFECTS OF STRESS AND PROMINENCE ON SPANISH STOPS AND LENITION IN L2 SPEECH OF PROFICIENT MANDARIN LEARNERS OF SPANISH

Xiaotong Xi[1], Peng Li[2]

[1]Department of Translation and Language Sciences, Univeristat Pompeu Fabra, Spain
[2]Center for Multilingualism in Society across the Lifespan (MultiLing), University of Oslo, Norway
[1]xiaotong.xi@upf.edu, [2]peng.li@iln.uio.no

ABSTRACT

This study assessed the effects of lexical stress and sentence prominence on Chinese students' phonetic realization of Spanish stops by measuring the Voice Onset Time (VOT) and the allophonic variants of voiced stops. Sixteen Chinese students and twelve Spanish natives participated in reading Spanish sentences containing /b, d, g, p, t, k/ in different utterance positions (initial vs. non-initial), lexical stress (stressed vs. unstressed) and sentence prominence (prominent vs. non-prominent). Results showed that (a) lexical stress showed no effect on learners' VOT; (b) learners could produce longer prevoicing of voiced stops in prominent positions; (c) stress and prominence did not show clear effects on learners' allophonic realization of voiced stops due to the limited cases of lenition. These results suggest that prosodic structure could potentially aid in the production of challenging L2 sounds, and that teachers may benefit from incorporating speech prominence in L2 teaching practice for training voiced stops.

Keywords: stop, lenition, prominence, L2 Spanish, Chinese students

1. INTRODUCTION

When learning voicing languages like Spanish, aspiration language speakers, such as Mandarin speakers, tend to assimilate both voiced (with negative voice onset time [VOT]) and voiceless (short positive VOT) stops to unaspirated stops (short positive VOT), as predicted by SLM-r [1] and validated by empirical studies [2], [3]. Another challenge arises from consonant lenition [3], [4]. Spanish natives often lenite voiced plosives to approximants in intervocalic positions, within a word or across word boundaries [5], but L2 learners tend to retain the plosive manner of articulation [3].

Despite the challenges faced by L2 learners at the segmental level, prosodic structures can influence the phonetic details of segments [6]. For instance, the Hyperarticulation hypothesis posits that sounds in stressed positions are hyperarticulated [7]. Empirical studies have shown that prosodically strong positions can affect the articulatory features of stops, such as shortening the VOT of voiceless stops [8] and lengthening the prevoicing period of voiced stops [9], [10]. In the context of L2 speech acquisition, lexical stress has been found to favor the production of voiced stops by English learners of Spanish [11]. However, few studies have assessed the effects of stress at both the word (lexical stress) and sentence levels (sentence prominence) on L2 Spanish stops, especially in learners whose native language lacks lexical stress and voicing contrast, such as Mandarin. Thus, it is worth exploring whether Mandarin-speaking learners would produce more nativelike Spanish stops under lexical stress or sentence prominence. Furthermore, if prosodically strong positions can strengthen the articulation of plosives, it is plausible that prosodically weak positions may facilitate lenition. Hence, would Mandarin-speaking learners lenite Spanish stops at unstressed or non-prominent positions?

Based on the literature reviewed above, this study formulates two hypotheses:

H1: Mandarin speakers would produce more native-like Spanish voiced stops, reflected by longer prevoicing, in stressed and prominent positions compared to unstressed and non-prominent positions.

H2: Mandarin speakers would be more likely to exhibit lenition of Spanish voiced stops in unstressed and non-prominent positions compared to stressed and prominent positions.

Since short-lag /p, t, k/ are shared phonemes in both Spanish and Mandarin, no specific hypothesis was formulated, but we included them in the design.

2. METHOD

2.1. Participants

We recruited 16 Mandarin learners of Spanish (12 females, M_{age} = 26.9, SD = 3.0) and 12 Spanish natives (8 females, M_{age} = 23.5, SD = 3.7). All participants signed the consent form to allow us to process their data. The Chinese students were late adult learners who began learning Spanish at the age of 18.7 years (SD = 1.6) and had studied Spanish for

4.3 years (*SD* = 1.1). They had formal Spanish language instruction in Spanish-speaking countries for 1.9 years (*SD* = 1.7) and had resided in Spanish-speaking countries for 5.3 years (*SD* = 2.9). Thus, they are proficient Spanish learners with substantial exposure to Spanish through study and living abroad.

2.2. Materials and procedure

We created 17 four-sentence dialogues. The first and third sentences were prompt questions to induce an answer that would have a narrow focus on the subject or object. The answers (second and fourth sentences) were two identical SVO sentences, with the target words embedded in the subject and object positions. For each of the six phonemes /b, d, g, p, t, k/, we selected four target words. The target phonemes always occurred at the beginning of the target words, with half of them being stressed and the other half unstressed. This resulted in a 2 × 2 × 2 design for each phoneme, varying in prominence (prominent vs. non-prominent), stress (stressed vs. unstressed), and sentence position (initial vs. non-initial). See (1) for an example, with stressed syllables underlined, target phonemes in boldface, and focus in uppercase.

(1) *¿Quién come la piña?* 'Who eats the pineapple?'
***B**EL**É**N come la **p**iña.* 'Belen eats the pineapple.'
¿Belén come la pera? 'Belen eats the pear?'
*No, **B**elén come la **P**IÑA.* 'No, Belen eats the pineapple.'

The speech data were recorded in a soundproof booth using a Zoom H4n Pro recorder and a SHURE SM35 microphone, with a sampling rate/resolution of 44.1kHz/16bits. Participants read all the sentences, but only the target sentences were analyzed.

2.3. Data coding and analysis

We obtained a total of 1,344 tokens. The first author manually annotated the VOT of all the plosives at initial and voiceless stops at non-initial positions as well as the length of the target words using Praat [12]. To avoid the effects of speech rate on the VOT production [13], we calculated the ratio of VOT to word length. However, we will plot and report the raw VOT values in the "Results" as descriptive statistics.

To test hypothesis 1, we built two Linear Mixed-Effects models (LMM) using the *lme4* package [14] in R to examine whether the two groups of speakers exhibited differences in the realizations of VOT for voiced and voiceless stops, depending on lexical stress and sentence prominence. The dependent variables for the two LMMs are the VOT ratio of sentence-initial stops and the VOT ratio of non-initial voiceless stops. Random structures were by-subject and by-item intercepts. The fixed effects for the first LMM involved voicing (voiced and voiceless), prominence (prominent and not prominent), stress (stressed and unstressed), group (Chinese students and Spanish natives), and their interactions. The second LMM included prominence, stress, group, and their interactions as fixed effects.

(2) *lmer(vot_ratio ~ voicing * prominence * stress * group + (1 | subject) + (1 | item)*
(3) *lmer(vot_ratio ~ prominence * stress * group + (1 | subject) + (1 | item)*

For the sentence-medial voiced /b, d, g/, the first author labeled the realizations by assessing the waveforms and spectrogram. Specifically, a sound was annotated as (a) voiceless plosive if the release burst was present with the aspiration phase, (b) voiced plosive if a voicing bar preceded the release burst, (c) fricative if there was a chaotic mix of random frequencies, and (d) approximant if vowel-like formant patterns could be identified. To test hypothesis 2, we ran a multinomial logistic regression (MLR) using *nnet* package [15] to examine whether the two groups of speakers differed in the allophonic realizations depending on the lexical stress and sentence prominence. The dependent variable was the allophonic realization (voiceless plosive, voiced plosive, fricative, and approximant) and the fixed effects were prominence, stress, group, and their interactions.

Significance were tested with Type II Wald chi-squared tests using *Anova()* from *car* package [16], and the group comparisons were carried out using the *emmeans* package [17] adjusted with the Bonferroni.

3. RESULTS

3.1. Effects of stress and prominence on VOT

The LMM for the VOT ratio of sentence-initial stops showed significant 3-way interactions of Voicing × Stress × Group (χ^2 = 5.7, p = .017) and Voicing × Prominence × Group (χ^2 = 11.3, p < .001), suggesting that the way in which the two groups distinguished voiced and voiceless stops by VOT differed across prominence or stress conditions. The LMM for the non-initial voiceless stops only revealed a significant 2-way interaction of Prominence × Group (χ^2 = 4.9, p = .027), suggesting that the VOT ratio of voiceless stops differed across group and prominence. In what follows, we will report the post-hoc results of the significant interactions related to the effects of stress and prominence.

3.1.1. Effects of stress on VOT

Regarding sentence-initial stops (Figure 1), stress had a significant effect only on the VOT ratio of voiced stops produced by Spanish natives, with the stressed

position showing longer prevoicing than the unstressed position (t = 2.5, p = .027). As for the comparisons between groups, Chinese students showed significantly shorter prevoicing of voiced stops than Spanish natives in both stressed (t = 6.0, p < .001) and unstressed (t = 4.1, p < .001) positions. Moreover, Chinese students produced significantly shorter VOT ratio for voiceless stops than Spanish natives in stressed positions (t = 2.7, p = .009), but not in unstressed positions (t = 1.5, p = .138). In addition, both groups exhibited clear VOT ratio contrasts between voiced and voiceless stops regardless of stress condition (all p < .05).

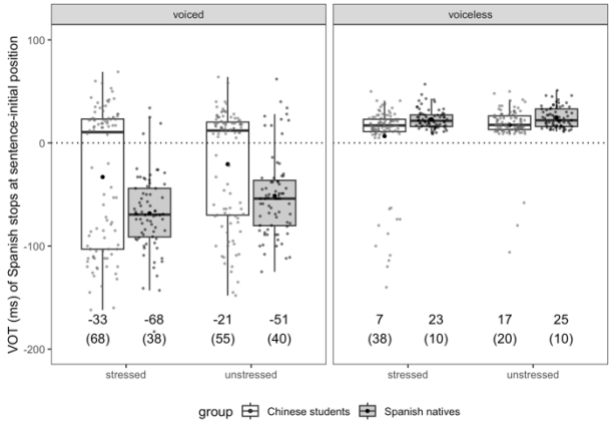


Fig. 1: Boxplots of VOT for the sentence-initial stops across group and stress conditions. Jittered points illustrate individual data points and means (SD) are labeled at the bottom.

In terms of non-initial voiceless stops (left panel in Figure 2), we did not find significant stress (χ^2 = 1.3, p = .262) or Stress × Group interaction (χ^2 = 0.1, p = .817). Both groups produced similar VOT ratio irrespective of the stress condition.

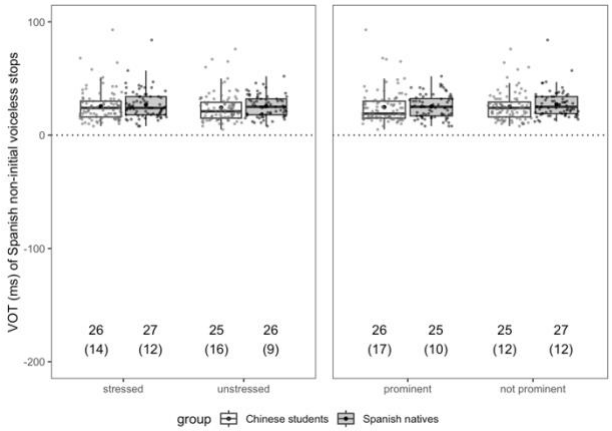


Fig. 2: Boxplots of VOT for non-initial voiceless stops across group and stress conditions (left panel) and across group and prominence conditions (right panel). Jittered points illustrate individual data points and means (SD) are labeled at the bottom.

Taken together, Chinese students' VOT was less affected by lexical stress and proficient learners still exhibited significant differences in VOT compared to Spanish natives.

3.1.2. Effects of prominence on VOT

At sentence-initial positions (Figure 3), Chinese students produced significantly longer prevoicing of voiced stops in prominent positions than in non-prominent positions (t = 5.4, p < .001). By contrast, Spanish natives did not show such a difference (t = 0.2, p = .874). As for voiceless stops, prominence did not show significant effects on VOT ratio in either group (all p > .05). In comparison with Spanish natives, Chinese students showed significantly shorter prevoicing than Spanish natives in both prominent (t = 3.4, p = .001) and non-prominent positions (t = 6.8, p < .001). Interestingly, for voiceless stops, Chinese students showed significantly shorter VOT ratio than Spanish natives in non-prominent positions (t = 2.6, p = .012) but not in prominent positions (t = 1.6, p = .114). Again, the voiced-voiceless contrast in VOT ratio was clear across group and prominence conditions (all p < .05).

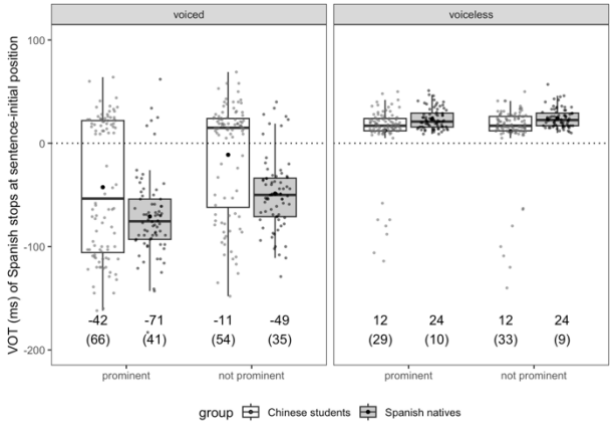


Fig. 3: Boxplots of VOT for sentence-initial stops across group and prominence. Jittered points illustrate individual data points and means (SD) are labeled at the bottom.

As for the non-initial voiceless stops (right panel in Figure 2), Spanish natives produced longer VOT ratio in prominent position (t = 4.2, p < .001), but this was not the case for Chinese students (t = 1.5, p = .131). In terms of group-level comparisons, similar to the initial voiceless stops, Chinese students showed significantly shorter VOT ratio than Spanish natives in non-prominent positions (t = 3.6, p = .001) but not in prominent positions (t = 1.5, p = .129).

3.2. Effects of stress and prominence on lenition

The MLR revealed two significant main effects. First, there was a significant main effect of prominence (χ^2 = 13.3, p = .004). Post-hoc pairwise comparisons showed that sentence non-initial voiced stops were realized more as approximants at non-prominent

positions compared to prominent positions (t = 3.7, p = .001), regardless of group. Second, the main effect of group was significant (χ^2 = 178.0, p < .001). Compared to Spanish natives, Chinese students produced a smaller portion of approximants (t = 16.3, p < .001) and a larger portion of voiced stops (t = 3.5, p = .002) and voiceless stops (t = 12.9, p < .001). No other significant main effects or interactions were found. Descriptive data can be found in Figure 4.

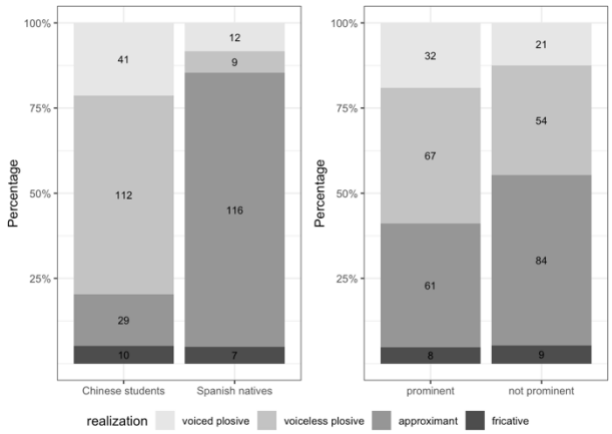


Fig. 4: Portion of allophonic realizations of the non-initial voiced stops divided by group (left panel) and prominence (right panel), with counts labeled on the bars.

4. DISCUSSION AND CONCLUSION

This study examined the impact of lexical stress and sentence prominence on Chinese students' VOT and allophonic realizations of Spanish stops using a dialogue-reading task. We found that (a) lexical stress did not show clear effects on Chinese students' prevoicing or lenition; (b) Chinese students produced longer prevoicing of sentence-initial voiced stops in prominent positions but prominence did not affect lenition as such; and (c) no clear effects of prominence or stress were observed for voiceless stops. Chinese students distinguished the voiced and voiceless stops by different VOT values, but they still showed differences from Spanish natives.

Our first hypothesis was partly confirmed that Mandarin speakers could produce more native-like L2 Spanish voiced stops in prominent positions. Firstly, unlike Spanish natives who produced longer prevoicing in stressed positions, Chinese students' production of voiced stops was not influenced by stress. This finding contrasts with [11] which showed that English speakers could increase the amount of Spanish prevoicing in stressed positions, similar to native speakers. However, it is important to note that our Chinese students had a much shorter length of residence compared to the participants in [11] (5.3 years vs. 36 years). Furthermore, English has lexical stress, and some English speakers indeed use prevoiced stops in their L1 [18], [19], but Mandarin shows little evidence for Spanish-like lexical stress or prevoicing. Therefore, the effects of stress on L2 speech may vary depending on learners' length of residence and L1 background. Secondly, Chinese students showed better pronunciation of the nonnative voiced stops under speech prominence. This finding adds to the growing body of evidence supporting the relationship between prosody and segment [8]–[10], and extends this research to L2 speech, where similar effects on Spanish vowels have been observed [20].

Notably, for the supposedly less challenging voiceless stops, we did not find significant effects of sentence prominence or lexical stress on Chinese students' VOT. In contrast, Spanish natives produced shortened VOT of voiceless stops in prominent positions, consistent with previous research [8]. Furthermore, Chinese students showed significantly shorter VOT of voiceless stops than Spanish natives in non-prominent positions. This finding suggests that although Chinese students were able to contrast voiced stops and voiceless stops by VOT values, they still showed some nonnative production patterns in the "less challenging" voiceless stops.

Our second hypothesis was not confirmed, as we did not observe more occurrence of lenition in L2 speech at unstressed and non-prominent positions compared to stressed or prominent positions. In general, more approximants were produced in non-prominent positions than in prominent positions. However, Chinese students showed a significantly smaller portion of lenition (15%) compared to Spanish natives (66%). Therefore, the effects of prominence on lenition were not evident in L2 speech. Moreover, previous research has reported a larger proportion of lenition within words by Chinese students [3], as opposed to our study where lenition occurred across word boundaries. This suggests that the lenition in L2 Spanish speech may be frequency-driven as approximants are more frequent within words than across word boundaries [5].

In conclusion, this study provides evidence that L2 learners can produce longer provoicing of Spanish voiced /b, d, g/ in sentence prominent positions, which highlights the prosody-segment interaction in L2. In practice, L2 teachers may manipulate the prosodic conditions to train the production of Spanish stops. Finally, further attention should be given to the lenition of voiced stops in L2 Spanish learning.

5. ACKNOWLEDGEMENTS

This study was supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 223265.

6. REFERENCES

[1] J. E. Flege and O. S. Bohn, "The Revised Speech Learning Model (SLM-r)," in *Second language speech learning: Theoretical and empirical progress*, R. Wayland, Ed. Cambridge University Press, 2021, pp. 84–118.

[2] Z. Liu and M. J. M. Ayuso, "Effects of attention on the production of Spanish consonants by SFL learners," *Estud. Fon. Exp.*, vol. 30, pp. 209–224, 2021.

[3] X. Xi and P. Li, "Spanish Stops And Their Allophones Produced By Proficient Mandarin Learners Of Spanish," in *Proceedings of the 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Nov. 2022, pp. 152–156, doi: 10.1109/O-COCOSDA202257103.2022.9997932.

[4] C. L. Nagle, "Individual developmental trajectories in the L2 acquisition of Spanish spirantization," *J. Second Lang. Pronunciation*, vol. 3, no. 2, pp. 218–241, 2017, doi: 10.1075/jslp.3.2.03nag.

[5] T. L. Face and M. R. Menke, "Acquisition of the Spanish voiced spirants by second language learners," in *Selected proceedings of the 11th Hispanic Linguistics Symposium*, 2009, pp. 39–52.

[6] D. Mücke and M. Grice, "The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation?," *J. Phon.*, vol. 44, pp. 47–61, May 2014, doi: 10.1016/j.wocn.2014.02.003.

[7] K. J. de Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 491–504, Jan. 1995, doi: 10.1121/1.412275.

[8] T. Cho and J. M. McQueen, "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," *J. Phon.*, vol. 33, no. 2, pp. 121–157, 2005, doi: 10.1016/j.wocn.2005.01.001.

[9] E. Wojtkowiak and G. Schwartz, "Prosody-segment interaction in the acoustics of Polish plosives," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, 2019, pp. 3285–3289.

[10] M. Simonet, J. Casillas, and Y. Díaz, "The effects of stress/accent on VOT depend on language (English, Spanish), consonant (/d/, /t/) and linguistic experience (monolinguals, bilinguals)," in *Speech Prosody 2014*, May 2014, pp. 202–206, doi: 10.21437/SpeechProsody.2014-28.

[11] T. L. Face and M. R. Menke, "L2 Acquisition of Spanish VOT by English-Speaking Immigrants in Spain," *Stud. Hisp. Lusoph. Linguist.*, vol. 13, no. 2, pp. 361–389, Oct. 2020, doi: 10.1515/shll-2020-2034.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022.

[13] A. M. Schmidt and J. E. Flege, "Speaking rate effects on stops produced by spanish and english monolinguals and spanish/english bilinguals," *Phonetica*, vol. 53, no. 3, pp. 162–179, 1996, doi: 10.1159/000262196.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[15] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*. Springer, 2002.

[16] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*. SAGE, 2019.

[17] R. V Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2022.

[18] J. E. Flege, "Laryngeal timing and phonation onset in utterance-initial English stops," *J. Phon.*, vol. 10, no. 2, pp. 177–192, 1982, doi: 10.1016/s0095-4470(19)30956-8.

[19] L. Hunnicutt and P. A. Morris, "Prevoicing and Aspiration in Southern American English," *Univ. Pennsylvania Work. Pap. Linguist.*, vol. 22, no. 1, pp. 215–224, 2016.

[20] P. Li and X. Xi, "Spanish lexical stress produced by proficient Mandarin learners of Spanish," in *ISAPh 2022, 4th International Symposium on Applied Phonetics*, 2022, pp. 40–45, doi: 10.21437/ISAPh.2022-8.

AUTOMATIC FEEDBACK ON PRONUNCIATION AND ANOPHONE: A TOOL FOR L2 CZECH ANNOTATION

RICHARD HOLAJ[1] – PETR POŘÍZKA[2]

[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic
[1]holaj@mail.muni.cz, [2]petr.porizka@upol.cz

ABSTRACT

This paper introduces a research project that represents an innovative approach to e-learning applications targeting automatic feedback on the pronunciation of non-native speakers based on computer speech recognition (specifically for Czech). We have collected data from 187 speakers of different pronunciation levels from 36 languages, conducted a pilot project, and developed the first version of an attributive annotation system based on tagging isolated speech sounds. We briefly mention the results of this stage (especially the success rate of the trained model), which led us to change our strategy and move to the next phase of the development of the automatic speech recognition tool. In this article, we present the current and next project phases: the Anophone annotation tool, a new annotation system based on whole-word tagging (two- to four-syllable words). The result is a measurable improvement in both the model and the success rate of speech recognition.

Keywords: automatic feedback on pronunciation, speech recognition, annotation, Czech, e-learning

1. INTRODUCTION

Today, there are some e-learning tools available for L2 competency in various languages. However, they prevailingly focus on lexical and grammatical aspects, and not much attention is paid to pronunciation. When they do address pronunciation, they only offer playback of pre-recorded vocabulary (words and phrases) or, exceptionally, provide the user with the possibility to record a short text sample and then compare this recorded sample with the original pre-recording. Thus, there is still a lack of applications providing an evaluation and automatic feedback on pronunciation to help users improve or achieve correct pronunciation in the foreign language they are learning. For Czech, we are aware of only two exceptions so far: the CzechME application [1] with several lessons focused on sound discrimination (differentiation), but the current version of this tool does not provide feedback on the user's pronunciation. The second application is Duolingo [2], where an automatic speech recognition (ASR) system transcribes the recording into text and compares with the text that was supposed to be spoken. However, ASR technology is designed to "understand;" thus, even if the pronunciation is more or less incorrect, it uses a language model [3] to estimate the intended linguistic meaning. That is a problem, because we receive feedback that our pronunciation is correct (successful in communication), even though it may be more than slightly mispronounced.

To our knowledge, probably the most technologically advanced e-learning system for L2 pronunciation is currently the English-focused ELSA Speak mobile app [4]. It includes its own proprietary solution for evaluating pronunciation and providing feedback to users in the form of the sound they were supposed to pronounce and the sound the user pronounced. However, there are still a few shortcomings. It is limited to only a segmental level of pronunciation and the feedback is restricted to the sound inventory of English, even though the speech sounds pronounced by the learners often do not correspond to any correct speech sound in the target language (i.e., English).

This situation led us to the idea that more than just the sound inventory of the target L2 (in case of ELSA, English; in our case, Czech) is needed to create a successful system for providing feedback on L2 pronunciation. Moreover, according to [5], the system "often misidentifies incorrect sounds as correct," so the problem of ASR technology remains. It is also important to mention that detection and evaluation of mispronunciations without accurate feedback is much more advanced, and there are already promising results (see [6], [7]), but none of them, to our knowledge, is able to provide accurate feedback in the form of which non-native sound was pronounced and how it differs from the speech sound that should have been pronounced.

2. L2 ANNOTATION SYSTEM AND NON-NATIVE SPEECH RECOGNITION OF CZECH

The aim of our research project is to create an e-learning tool for automatic pronunciation feedback for non-native speakers learning Czech. Due to the

limited scope of this paper, it is not possible to present all key aspects of our research project. Therefore, for this reason, we refer here to the previous article [8] where all key elements are described.

We have completed the first (pilot) phase of the project, which is described and summarized in [8], during which we collected recordings a total of 187 non-native speakers of 36 different languages across all learning levels (using the CEFR scale from A0 to C1) and in different age categories from 18 to 73 years old, with the largest group being speakers under the age of 40 (73%; cumulative frequency: 136 speakers).

The key idea of our approach is to include non-native sounds (phonetic features from the non-native sound inventory) into the speech recognition system inventory, so that we are not limited to the most similar speech sound from the target language and can obtain less biased (more accurate) results from the actual pronunciation. We then want to provide the learners with relevant linguistic information through a mobile app not only by outputting the pronounced versus correct sound, but also by more explicitly describing what was wrong in their pronunciation (e.g., presence of aspiration, nasalized pronunciation, etc.) and how to fix it (e.g., more open mouth, vocal cords in action, etc.).

In the first phase of our research project, we tested this approach on a manually annotated sample of just under 4,000 isolated speech sounds from 32 non-native speakers of Czech using the attributive annotation system described in [8], which we developed for this purpose and which allows us to label not only standard pronunciation but also various deviations from standard pronunciation (see 2.2). Other important phenomena from the project are presented in [8], such as the characterization of the data, the methodological framework that reflects the phonetic basis of Czech, the relevant phonetic specificities of foreign languages, and the most common pronunciation errors of non-native speakers learning Czech.

2.1. Test models for the speech recognition of non-native speakers of Czech

For a project based on ASR technology and data (segmented and annotated audio recordings) that contain "incorrect speech sounds" of non-native speakers, we needed an annotation system that allows us to distinguish the differences between the prompt for the desired speech sound and the user's resulting attempt at making the speech sound. Likewise, we needed a tool based on annotated data and capable of recognizing from the recordings of the speech sound and corresponding annotation.

For this purpose, we have developed a tool for speech recognition of individual speech sounds: a Python script for ASR and model evaluation based on the Persephone library [9] and, more recently, the Anophone annotation tool (see 2.3). Our data model consists of three parts in the following ratio: (a) training set – 90%, (b) validation set – 5%, and (c) test set – 5%.

Persephone contains a tool which extracts several audio features, and in both the old and new experiment we used the LMFB (Log Mel Filterbank) with delta and delta-delta features. Extracted features along with their corresponding labels were then split by library into non-intersecting train (90 %), validation (5 %) and test (5 %) sets (see above our data model). "The underlying model used is a long short-term memory (LSTM) recurrent neural network [10] in a bidirectional configuration [11]. The network is trained with the connection's temporal classification (CTC) loss function [12]." [9]

In both the old and new models, we used the same default three-layered architecture with 250 hidden nodes (cf. [8]). As we mentioned in [8], it has to be "trained with pre-processed data for at least 30 epochs. Training stops when one of those conditions is met:

(a) training LER (learning error rate) is lower than 0.1% and the validation LER is lower than 1%
(b) validation LER has not improved in the last 10 epochs
(c) after 100 epochs

In the last step, we test our trained model against the test data set."

2.2. Pilot project results (annotation of isolated speech sounds)

In the first phase of the research project, we developed an attribute–value annotation system that works with manually annotated isolated speech sounds and is based on systematically categorized pronunciation errors from different languages or language groups (see [8] for more detail).

This annotation system specifies two groups of attributes: (1) fixed, with a binary value of 0 or 1 for phonological features (such as quantity, voicing, etc.), and (2) variable, with the possibility to add additional values as needed (phonetic features such as palatalization, nasalization, etc.). It also includes a special tag for replacing one speech sound with another.

The annotation tag (label) is divided by a colon into two main parts: (1) the part before the colon indicates the sound to be pronounced; (2) the attributes after the colon indicate deviations in

pronunciation from the standard/correct phonetic form of the speech sound (using possible attribute values). If the pronunciation is correct, only the part before the colon is used. In case of incorrect or non-standard pronunciation, any number of attributes may follow the colon (see examples below).

a [standard pronunciation of vowel *a*]
a:k1 [short vowel *a* pronounced as long]
a:k1vN [short vowel pronounced as long and nasalized]
E:A [*ɛ* pronounced like *a*]

Explanatory notes:
k = quantity (0 = short, 1 = long).
v = non-standard pronunciation variants (N = nasalization)

The pilot experiment [8], included 3,717 labelled sounds from 32 non-native Czech speakers with two versions of annotation AV1 and AV2 (improved by the identification of duplicate tags and some corrections of system errors in the annotation).

The results of those annotation models, which work with isolated speech sounds, are summarized in Table 1:

error rate	training	validation	test
model AV1	43%	42%	51%
model AV2	15%	37%	41%

Table 1: Old models – based on the attributive system and annotation of isolated speech sounds (approx. 3,700 samples) [6]

The error rate for AV1 was huge, while adjusted annotation model AV2 was more successful, but error rates (especially validation and test) were still quite big, see [8] for more detailed description.

2.3. Anophone annotation system and preliminary results of the new ASR model (word annotation)

The results of the pilot test led us to change our strategy and the whole annotation system to achieve more satisfactory outputs and a more reliable speech recognition model. This change meant moving from an attributive annotation system to a different approach and way of processing speech data. For the next phase of testing, we created new annotated data that annotated whole words instead of isolated speech sounds, and most importantly, we created a new annotation system, Anophone [13].

This is a very flexible tool that allows the uploading of audio segments to a database for further processing: it allows not only the annotation of data, but also the creation of new sets of tags or modifications of existing tagsets according to the user's needs, also assigning these tags to the data (audio segments) through a web interface. In addition, Anophone allows one to build a new annotation task independently of an existing one (for example, for the needs of another project or language). Thus, this tool can be used by other researchers for similar projects: to create their own annotation tasks, and upload and annotate data.

Anophone works with four datasets:
(1) Recordings: segmented sounds for annotation, which can be further filtered by four categories – language, speaker, repetition, and word.
(2) Tasks: annotation tasks, the way we annotate the data – e.g., at the level of vowels, phonetic features, etc.
(3) Labels: the tags (set of tags) that we create for a particular annotation task(s) and through which we annotate the data.
(4) Annotations: custom/individual annotation labels assigned to audio segments in the database (stored under a unique annotator ID).

Anophone selects data to annotate randomly, even repeatedly with the same sounds, to ensure multiple annotations of the same word (segment) by different annotators. Alternatively, intentional filtering of the data by a defined category/attribute (language, speaker, repetition, word)[1] can be used in combination with regular expressions. Thus, we can select only speakers of a particular language or language group, or all speakers regardless of native language, etc. In the administration menu, we can perform a final annotation of the audio segment, which is intended for export to machine learning, considering the existing annotation variants (the annotator is in the system under a certain ID, e.g., to detect systematic annotation errors, to track the correspondence between annotators, etc.).

For this phase of testing the annotation model, we selected more homogeneous data, limited to 4 groups of speakers, namely German, Russian, Ukrainian, and Vietnamese speakers. Apart from the homogeneity of the data, the reason for the choice was that these are the largest groups of non-native speakers in the Czech Republic (we did not include Slovaks because of their interlingual proximity to Czech).

We manually annotated whole words (not isolated speech sounds), specifically more than 6,500 samples (from 47 speakers) of two- to four-syllable words in which the given speech sound occurred in different positions (initial, middle, final) and in different phonemic contexts (vowels, obstruents, sonorants). Each segment was assigned metadata: speaker code, gender, age, nationality, native language, level of

Czech. The sub-segment (word) is named according to the specific meaning of the word and this label is then compared with the text output of the manual annotation, which represents the actual pronunciation of the segment, in the next phase of testing the annotation model. Thus, the word label (file name) is compared with the annotated segments (pronounced sounds, including non-standard variants or phonetic features from foreign languages). For the annotation, two sets of sound-level labels are defined in Anophone: (1) from the standard Czech inventory, (2) or containing various non-standard sounds produced by non-native speakers (in the label set color-coded into two categories: (ad 1) native vs. (ad 2) non-native sounds or features such as aspiration, etc.).

In the current phase we have created three models with different training batch size (one of the neural network parameters). We call those models B72 (batch size 72), B36 (batch size 36), and B18 (batch size 18).

The results of the current phase, including the success rate of the trained model, are summarized in Table 2 (cf. Table 1 and the success rate of the pilot model):

error rate	training	validation	test
model B72	5%	23%	24%
model B36	1%	23%	23%
model B18	11%	25%	25%

Table 2: New models – based on Anophone and whole word annotation (approx. 6,500 samples)

Despite fact that the task in the current phase is more difficult than in the initial phase (individual sounds vs. whole words), we see that the results are better, since we achieved much lower error rate in all models. From three tested models, B36 with a medium batch size is the most promising.

2.3.1. Detailed evaluation

Since evaluation is phoneme based (correct vs. incorrect sound in output), to truly see the quality of the tool, we need to look closer at the incorrectly transcribed words.

Table 3 shows a few of those words. In the left part there are mistakes in validation phase, while in the right part there are mistakes in test phase. Both parts consist of expected output on the left side and real output on the right side.

VALIDATION		TEST	
expected	**output**	**expected**	**output**
b ou̯ r̥ k a	p ɔ ʃ k a	m a t k a	s n a p k a
ʃ tʲ ɛ s tʲ i	ʃ ɛ s ɟ ɪ	b ɔ : r ʃ k a	b ou̯ : r ʃ k a
k ɔ : l ɔ	k ɔ l ɔ	s k ou̯ ʃ k a	s p u ʃ k a
ʒ a k	ʐ̩ a k	m a t k a	a k a

Table 3: Incorrectly transcribed words

In Table 3, both in test and validation, mistakes are often make in between two words with perceptive similarity, often interchanging similar speech sounds, differences in length, voice or manner of articulation. Missing segment errors are also common. This implies that since the evaluation does not make any difference for the type of mistake, the results are actually slightly better than we would expect from an error rate alone, since the errors encountered are usually less severe.

3. CONCLUSION

Our system has promising results that can lead to a tool which will be able to automatically transcribe the speech of non-native speakers of Czech and thus provide a base to create better feedback for learners of L2 Czech. Although results are getting better, the model is still not good enough to be applied in an e-learning app. To change this, we need to collect more data and optimize our annotation and model. Furthermore, we need to include L2 Czech speakers from another native languages.

We also want to experiment with tier-based annotation which would work with different phonetic tiers and annotate them separately (including tones, stress, and another suprasegmental features that can have many effects on individual segments or on the speech production of non-native speakers in general).

Following our current findings, we believe that a reliable tool for automatic L2 Czech pronunciation feedback will be possible within few years.

7. REFERENCES

1. EVE Technologies, s.r.o. 2021. CzechME (1.0.5) [Mobile app]. https://play.google.com/store/apps/details?id=cz.evetech.czechme.
2. Duolingo, Inc. 2021. Duolingo (5.1.5) [Mobile app]. https://play.google.com/store/apps/details?id=com.duolingo
3. Kuhn, R., De Mori, R. 1990. Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6), 570–583.

4. ELSA Co, Ltd. 2021. ELSA Speak: Online English Learning & Practice App (6.2.1) [Mobile app]. https://play.google.com/store/apps/details?id=us.nobarriers.elsa.
5. Becker, K., Edalatishams, I. 2019. ELSA Speak – Accent Reduction [Review]. In: J. Levis, C. Nagle, and E. Todey (eds), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA: Iowa State University, 434–438.
6. Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Calamaro, S., and Kostek, B. 2021. Weakly-supervised word-level pronunciation error detection in non-native English speech. *Proceedings of Interspeech 2021*, 4408–4412.
7. Yang, L., Fu, K., Zhang, J., and Shinozaki, T. 2021. Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning. *Neural Networks*, 142, 597–607.
8. Holaj, R., Pořízka, P. 2021. L2 Czech Annotation for Automatic Feedback on Pronunciation. *Jazykovedný časopis / Journal of Linguistics* 72 (2), 510–519.
9. Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., et al. 2019. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *LREC 2018 (Language Resources and Evaluation Conference)*, Miyazaki, Japan, 3356–3365.
10. Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8), 1735–1780.
11. Schuster, M., Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
12. Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd international conference on Machine Learning*, 369–376.
13. Anophone: an annotation tool for phonemes. [Software]. https://anophone.evetech.cz/#/intro.

[1] Readings: the texts were read twice by non-native speakers, first with and then without instructor assistance. The segments were then matched accordingly with the assigned markers v1 (repetition after instructor) vs. v2 (learners pronouncing without any support). Pronunciation quality for most recorded speakers differs noticeably between versions.

EXPLORING THE RELATIONSHIP BETWEEN INDIVIDUAL DIFFERENCES IN CROSS-LANGUAGE PERCEIVED SIMILARITY AND L2 VOWEL PERCEPTION

Juli Cebrian, Celia Gorba

Universitat Autònoma de Barcelona
juli.cebrian@uab.cat, celia.gorba@uab.cat

ABSTRACT

This paper explores the relationship between cross-linguistic perceived similarity and L2 vowel identification and discrimination at the individual level. The study also investigates if the perception of cross-linguistic similarity changes as a result of a 6-session perceptual training regime found to improve L2 perception. A group of Spanish-speaking learners of English in an instructional setting completed a perceptual assimilation task before and after training. The results indicated that learners varied in their choice of L1 match for English /ɪ/ and /ɑː/, but assimilation patterns were consistent for the remaining vowels. The outcome of a series of correlations and multiple regression analyses yielded some evidence that individual differences in L2 to L1 assimilation patterns predicted L2 perception. However, this was only found in a subset of cases. Further, perceptual training generally had little effect on cross-linguistic perceived similarity, except for changes involving /ɪ/, possibly related to increased metalinguistic awareness.

Keywords: L2 speech, cross-linguistic similarity, individual differences, L2 vowel perception.

1. INTRODUCTION

According to current second language (L2) speech theories the likelihood of target-like categorization of L2 speech sounds is determined by L2 learners' ability to detect differences between native and target language sounds [1, 2]. Hence, assessing the degree of similarity between native and non-native sounds is crucial in order to make predictions about the difficulty of perceiving and producing specific L2 phones. Cross-linguistic similarity is often measured perceptually [3, 4], e.g. by means of a cross-language categorization task or perceptual assimilation task (PAT). In this task, listeners categorize L2/non-native stimuli in terms native categories and provide goodness of fit ratings for each categorization. However, predictions based on group results often overlook the fact that individual learners sharing the same L1 may differ in their perception of similarity between native and non-native sounds (e.g., [5]). Hence the need to explore if individual differences in cross-language perception of similarity may result in differences in L2 development, as has sometimes been reported [6].

Furthermore, according to current models of L2 speech, greater experience with the L2 may result in increased ability to perceive phonetic differences between L1 and L2 sounds [7]. Still the evidence from previous research is inconsistent. Some studies show that experience results in a greater ability to distinguish L1-L2 sounds [8, 9]. By contrast, other studies report no effect of experience on cross-linguistic mappings, e.g., comparing similarity judgements from learners varying in length of residence in an L2 setting (1.8 vs 7 years, [10]) or the same group of learners tested a year apart [11].

On the other hand, it is well-known that phonetic training approaches such as high variability phonetic training (HVPT) have a positive effect on L2 learners' ability to perceive and produce target language sounds accurately (see [12, 13] for reviews). However, few studies have investigated if the perception of cross-linguistic similarity changes as a result of phonetic training (e.g., [14]).

Thus the current study aims to examine the relationship between perceived cross-language similarity and L2 perception by exploring whether individual differences in perceived similarity between target and native vowels are reflected in the individuals' ability to identify and discriminate L2 vowels. In addition, the study investigates if cross-linguistic perception of similarity can change as a result of L2 exposure through high variability phonetic training.

2. METHODOLOGY

2.1. Participants

Twenty-eight Spanish-speaking learners of English participated in the study. They were all university undergraduate students in their first year of English Studies at Universitat Autònoma de Barcelona (UAB), Spain. They started learning English at school but had scarce experience in a naturalistic setting, as their time spent in an English-speaking country was

limited to short summer stays. Their average age was 19.6.

2.2. Tasks

Participants were tested on L2 perception as well as on L1-L2 perceived cross-linguistic similarity before and after completing the training program. Specifically, learners completed an identification task, a categorical AX discrimination (same/ different) task, and a perceptual assimilation task (PAT). The identification task involved the British English vowels /æ, ʌ, ɪ, iː, ɜː, e, ɑː/ (as in 'bat,' 'but,' 'bit,' 'beat,' 'Bert,' 'bet' and 'Bart'), totalling 104 trials (four words per vowel per talker and several repetitions). The response options were pairs of two common English words containing and representing each of the vowels tested. The discrimination task involved the following vowel contrasts: /æ-ʌ/, /ɪ-iː/, /ɜː-e/, /ɜː-ɑː/. The total number of trials was 96 (48 same, 48 different). Vowel stimuli were presented in CVC non-sense words produced by two different talkers.

As for cross-linguistic perception, in the PAT learners were presented with L2 (English) stimuli that had to be identified in terms of L1 (Spanish) categories and then rated in terms of their goodness of fit on a 7-point scale. The stimuli included the English vowels /iː ɪ e ɜː æ ʌ ɑː/, which had been elicited from three male native speakers of British English in bVt sequences, and edited so that each stimulus contained from the release of the /b/ to the beginning of the /t/ closure. The response alternatives involved the Spanish vowel set (/i e a o u/) and the diphthongs /ai ei oi/. The total number of trials was 108 (9 vowels x 3 talkers x 2 tokens x 2 repetitions).

2.3. Procedure

The participants took part in a HVPT program and were tested on their L2 perception as well as on cross-linguistic perceived similarity before and after training (pretest and posttest). The training program consisted of six 30-minute sessions over several weeks and was carried out at a UAB's speech laboratory. The training tasks (a 7-alternative forced choice identification task or a categorical AX discrimination task) were similar to the pretest and posttest tasks, although different stimuli were used and corrective feedback was provided after each trial during training.

2.4. Data analysis

In order to examine the relationship between learners' cross-linguistic similarity judgements and their ability to identify and discriminate L2 vowels, two different types of analyses were carried out. First, a series of Spearman correlations were conducted involving the identification and discrimination accuracy and the modal responses obtained in the PAT, at both testing times. The measures used to quantify the PAT results were the assimilation scores as well as the goodness of fit rating (GR). In addition, a fit index score (FI), a composite score resulting from multiplying the assimilation percentage by the corresponding GR, was also calculated [15]. Next, a series of multiple regression analyses were conducted so as to determine whether the PAT results (independent variables) predicted identification and discrimination results (dependent variables) at both testing times. Separate analyses were carried out for high and low vowels, for pre and posttest results. All mathematical assumptions, including no perfect collinearity, were met.

3. RESULTS

3.1. Pretest and posttest results

The results of the PAT are shown in Table 1, which presents the main assimilation patterns observed at pretest and at posttest. As can be observed, every English vowel received assimilation scores equal to or higher than a 70% categorization threshold [16]. The English to Spanish assimilation modal responses were: /æ/-/a/, /i/-/i/, /e/-/e/, /ʌ/-/a/, /ɑː/-/a/, /ɪ/-/e/ and /ɜː/-/e/. Still, some English vowels received non-modal responses ranging from 9% to 29%, namely /ɪ/ as /i/ (29%), /ɑː/ as /o/ (22%), and /ɜː/ as /o/ (14%) or /a/ (13%) and /ʌ/ as /o/ (9%). Particularly in the case of English /ɪ/, the different assimilation patterns corresponded to patterns observed for different individuals. Thus, at pretest 14 listeners assimilated /ɪ/ to Spanish /e/, six to /i/ and for the remaining eight /ɪ/ was uncategorized (i.e., did not reach 70% assimilation to any L1 vowel). At posttest, the numbers were 11 (/ɪ/ to Spanish /e/), 10 (/ɪ/ to /i/) and 6 (uncategorized).

The results of the discrimination and identification pretest and posttest are presented in Tables 2 and 3. The English vowel pairs /ɜː/-/e/ and /ɜː/-/ɑː/ were the most accurately discriminated, followed by /iː/-/ɪ/, and /æ/-/ʌ/, which obtained the lowest scores. In terms of identification accuracy, English /e/, /ɪ/, /iː/ and /æ/ obtained the highest scores, followed by /ɑː/, while /ʌ/ and /ɜː/ were the most challenging.

As can be observed, overall there is an improvement in identification and discrimination from pre to posttest (these results are addressed in a separate paper). The focus of the current paper is on the effect of training on cross-language perception and its relationship with L2 perception at the

individual level. Regarding changes in cross-language perception from pretest to posttest, Table 1 shows that most assimilation patterns remain stable. The only vowels that show some change are /ɪ/, with an increase in assimilations to Spanish /i/ from 29% to 45%, and, to a lesser extent, /ɑ:/ as Spanish /o/ from 22% to 30%. These changes may be attributable to increased metalinguistic awareness, as discussed in section 4. The results of the correlational analysis and the multiple regression analyses relating the PAT and L2 perception results are reported the next section.

Target vowel	L1 vowel	Pretest %	Pretest GR	Posttest %	Posttest GR
/æ/	/a/	99%	6.2	100%	6.4
/ʌ/	/a/	90%	5.8	88%	5.1
	/o/	9%	4.5	10%	3.6
/i:/	/i/	93%	6.5	92%	6.5
/ɪ/	/e/	**70%**	5.8	**55%**	5.7
	/i/	**29%**	5.5	**45%**	5.5
/ɜ:/	/e/	70%	4.0	75%	3.5
	/a/	13%	3.2	10%	2.3
	/o/	14%	3.7	15%	2.6
/e/	/e/	99%	5.2	99%	5.5
/ɑ:/	/a/	74%	4.8	70%	4.7
	/o/	22%	4.7	30%	4.7

Table 1: PAT results at pretest and posttest showing for each target English vowel the corresponding L1 match(es); %. = % assimilation, GR = goodness of fit rating.

Vowel contrast	Pretest	Posttest
/æ/-/ʌ/	54 (19)	63 (21)
/i:/-/ɪ/	62 (20)	68 (21)
/ɜ:/-/ɑ:/	71 (23)	73 (27)
/ɜ:/-/e/	80 (18)	93 (12)

Table 2: Discrimination of each target vowel at pretest and posttest (SD are given in parentheses).

	Pretest	Posttest
/æ/	59 (21)	59 (19)
/ʌ/	31 (24)	47 (26)
/i:/	64 (18)	71 (19)
/ɪ/	63 (24)	78 (19)
/ɜ:/	36 (20)	51 (25)
/ɑ:/	50 (17)	68 (27)
/e/	67 (25)	89 (15)

Table 3: Identification of each target vowel contrast at pretest and posttest (SD are given in parentheses).

3.2. Relationship between PAT and L2 perception

Overall, the results of the correlational analyses did not show a clear relationship between the accuracy scores of the perception tests and the PAT results given that few outcomes were correlated. The significant results obtained are presented in Table 4, which are limited to some identification results and goodness ratings (GR) obtained in the PAT. The negative correlations indicate that the lower the GRs for English /ɪ/ as a match to Spanish /i/, the more accurately English /i:/ was identified; and also the lower the GRs in the assimilation of English /ʌ/ to Spanish /a/, the better that both /æ/ and /ʌ/ were identified. No significant correlations were obtained involving discrimination results.

Ident. result	PAT result	ρ	p	N
/i:/ in NW	GR /ɪ/ as /i/:	-.581	.009	19
/æ/ in NW	GR /ʌ/ as /a/:	-.527	.006	26
/ʌ/ in NW	GR /ʌ/ as /a/:	-.501	.009	26

Table 4: Significant Spearman's correlations between identification and PAT measures.

Regarding the multiple regression analyses, the individual results obtained in the PAT predicted the perception accuracy scores on very few occasions. Regarding high vowels, cross-linguistic perception did not significantly predict any of the identification or discrimination results at pretest, as shown both by the regression equations ($p > .05$). As for low vowels, neither the regression equations nor the individual coefficients revealed that the PAT results predicted ID or DIS scores ($p > .05$).

At posttest, the multiple regression analyses did not reveal any significant result for high vowels, indicating that PAT results did not significantly predict ID or DIS accuracy scores. Some significant results were found for low vowels. Specifically, the ID of /æ/ was found to be significantly predicted by the PAT results. In this case, the PAT's FI score for English /ʌ/ as Spanish /a/ and for English /ʌ/ as Spanish /o/ negatively predicted the identification of English /ʌ/ ($F(5, 20) = 2.638$, $p = .055$, $R^2 = .397$; /ʌ/ as /a/: $\beta = -1.103$, $p = .035$; /ʌ/ as /o/: $\beta = -1.0447$, $p = .030$). This indicates an unclear relationship between PAT results and identification accuracy, as both assimilation patterns are inversely related to identification accuracy. No significant results were obtained regarding PAT results and discrimination results.

4. DISCUSSION AND CONCLUSIONS

This study set out to examine if differences across individuals in their perceived similarity between native and target language vowels was reflected in the individuals' ability to identify and discriminate L2 vowels. The results of the PAT showed that similarity judgements were very consistent across individuals for most vowels, but individuals varied in their cross-linguistic categorization of particularly English /ɪ/, and to a lesser extent English /ɑː/ and /ʌ/. For example, some learners assimilated English /ɪ/ to Spanish /e/ while others assimilated English /ɪ/ to Spanish /i/. Given that English /iː/ was predominantly assimilated to Spanish /i/, it could be hypothesized that learners who assimilate each English vowel to a different L1 vowel (two-category assimilation in PAM-L2's terms [1]) will be more successful at identifying and discriminating English /iː/-/ɪ/ than those learners who assimilate both English /iː/ and /ɪ/ to the same Spanish vowel (single-category or category goodness difference assimilation to Spanish /i/). The same applies to the English pairs /æ/-/ʌ/ and /ɑː/-/ʌ/. English /æ/ was consistently assimilated to Spanish /a/, but both /ʌ/ and /ɑː/ obtained assimilations to Spanish /o/ in addition to Spanish /a/.

In order to examine these predictions, a series of correlations and multiple regressions were conducted, investigating if assimilation patterns were related to and predicted identification and discrimination accuracy. The results, however, found that a learner's L2 perception was generally unrelated to the assimilation preference, with very few exceptions. It was found that the lower the GRs given to English /ɪ/ as a match to Spanish /i/, the better English /iː/ was identified. Similarly, the lower the GRs obtained for English /ʌ/ as Spanish /a/, the better English /æ/ and /ʌ/ were identified. However, the assimilation of English /ɪ/ to either Spanish /i/ or Spanish /e/ did not appear to affect the identification of /ɪ/ or the discrimination between /ɪ/ and /iː/. The lack of a stronger relationship between cross-linguistic similarity differences and identification and discrimination accuracy thus runs counter to what the theoretical models would suggest (e.g., [7]). It is possible that the measures obtained in a PAT reflect a perceptual sensitivity that does not play a role in the categorization of L2 sounds. As discussed below, other factors such as the learners' metalinguistic and orthographic knowledge of the L2 may interfere with the categorization of the target sounds. Along the same lines, it is also possible that a closer relationship between PAT measures and L2 perception may be obtained from a naïve population or from learners at a more initial stage in the acquisition of the L2.

Finally, while a 6-session perceptual training regime was effective in improving identification and discrimination of L2 vowels, it was insufficient to affect cross-linguistic similarity relations, as no consistent change in perceived similarity between L1 and L2 vowels was observed from pretest to posttest. Nevertheless, there were some changes in the perceived similarity for those sounds for which there was more variability in perceived similarity. For example, at pretest English /ɪ/ was perceived as closest to Spanish /e/ 70% of the time and to Spanish /i/ 29% of the time; at posttest the percentages were 55% and 45%, respectively. Thus, there was an increase in the assimilation of English /ɪ/ to Spanish /i/. This change can be related to the claim that perceived similarity may respond to acoustic-phonetic similarity at initial stages of learning and may become more phonological as learners gain experience with the L2 [17]. In other words, enhanced metalinguistic knowledge of the L2 (in this case through perceptual training) may go beyond the raw acoustic information of the L2 and may be taken into account when judging cross-linguistic similarity. This may be particularly relevant in cases like English /ɪ/ due to familiarity with orthography (/ɪ/ is typically spelled with letter <i>, which stands for the vowel /i/ in Spanish) as well as to increased awareness of the English vowel system as a result of exposure, instruction and training (e.g., the functional load of the /iː/-/ɪ/ contrast, and instructional biases that refer to this contrast as a contrast between "long and short /i/").

In brief, L2 vowel identification and discrimination was not found to be generally affected by individual differences in cross-linguistic similarity, with very few exceptions. In addition, perceptual training was not observed to alter cross-linguistic perceived similarity for the most part. It appears that factors other than pure perceptual sensitivity to acoustic differences between native and non-native vowels may also play a role in the categorization of L2 sounds, as illustrated by the changes observed in the perceptual assimilation of vowels like English /ɪ/.

This research was supported by Grants No. FFI2017-88016-P and PID2021-122396NB-I00 from the Spanish Ministry of Science and Innovation and Grants No. 2017SGR34 and 2021SGR00544 from the Catalan Government.

5. REFERENCES

[1] Best C.T., Tyler M.D. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In Bohn, O-S, Munro, M. J. (eds),

Language Experience in Second Language Speech Learning. Amsterdam: John Benjamins, 13–34

[2] Flege, J. E., Bohn, O. S. 2021. The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: theoretical and empirical progress,* 3–83. New York, NY: Cambridge University Press.

[3] Bohn, O.-S. 2002. On phonetic similarity, in An Integrated View of Language Development: Papers in Honor of Henning Wode, edited by P. Burmeister, T. Piske, and A. Rohde (Wissenschaftlicher Verlag, Trier), pp.191-216.

[4] Strange, W. 2007. Cross-language similarity of vowels, in Language Experience in Second Language Speech Learning: In Honor of James Flege, edited by M. Munro, and O.-S. Bohn (John Benjamins, Amsterdam), pp. 35-55

[5] Escudero, P., Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. *Studies in second language acquisition* 26(4), 551-585.

[6] Mayr, R., Escudero, P. 2010. Explaining individual variation in L2 perception: Rounded vowels in English learners of German. *Bilingualism: Language and Cognition* 13(3), 279-297.

[7] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In: W. Strange (ed), *Speech perception and linguistic experience: Issues in cross-language researc*h. Timonium, York Press, 223-277.

[8] Ingram, J.C.L., Park, S.-G.. 1997. Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, 25, 343-370.

[9] Rallo Fabra, L. (2005). Predicting ease of acquisition of L2 speech sounds: A perceived dissimilarity test, *Vigo Int. J. Appl. Linguist*., 2, 75–92

[10] Flege, J. E., Munro, M. J., Fox, R. A. (1994). Auditory and categorical effects on cross language vowel perception, *J. Acoust. Soc. Am.*, 95, 3623–3641

[11] Imai, S., Flege, J. E., Wayland, R.. 2004. Perception of cross-language vowel differences: A longitudinal study of native Spanish adults learning English. Poster presented at 9th Conference on Laboratory Phonology, University of Illinois at Urbana-Champaign.

[12] Thomson, R. I. 2018. High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, 4(2), 208-231.

[13] Sakai, M., Moorman, C. 2017. Can perception training improve the production of second language phonemes? A meta-analytic review of twenty-five years of perception training research. *Applied Psycholinguistics*, 1-36.

[14] Gong, J., García-Lecumberri, M. L., Cooke, M. (2017). Ab initio perceptual learning of foreign language sounds: Spanish consonant acquisition by Chinese learners. *System*, 66, 142-155.

[15] Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J. S. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America,* 107, 5, 1, 2711–2724.

[16] Tyler, M. D., Best, C., Faber, A., and Levitt, A. 2014. Perceptual assimilation and discrimination of non-native vowel contrasts, *Phonetica*, 71, 4-21.

[17] Chang, C. B. 2019. The phonetics of second language learning and bilingualism. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics*. Abingdon, UK: Routledge, 427–447.

SECOND LANGUAGE LEARNERS' ACQUISITION OF MANDARIN FRICATIVE CONTRASTS: AN ULTRASOUND STUDY

Madeleine Oakley

North Carolina State University
mjoakley@ncsu.edu

ABSTRACT

This study examines how second language learners transfer articulatory and acoustic categories to produce non-native fricative contrasts. It is hypothesized that L1 English-L2 Mandarin learners will have difficulty producing the Mandarin /ʂ/-/ɕ/ contrast because both sounds are perceptually assimilated to English /ʃ/ [1]. Results from multi-taper spectra show that learners who produce Mandarin /ʂ/ and /ɕ/ similarly also tend to produce them similarly to English /ʃ/. However, ultrasound tongue imaging results reveal that these same learners do produce a difference in tongue position between Mandarin /ʂ/ and /ɕ/. Furthermore, L1 English-L2 Mandarin learners produce all English fricatives with a higher tongue body than Mandarin fricatives, showing learners are not transferring articulatory gestures to produce these fricatives. The results of this study show that learners are producing the non-native fricative contrast articulatorily, although this finding is obscured when only observing the acoustic data.

Keywords: L2 phonetics, articulation, ultrasound, fricatives, sibilants

1. INTRODUCTION

The primary objective of this study is to examine how learners produce the articulatory contrast between non-native consonant that are predicted to be perceived as phonetically similar to the same L1 consonant category. When acquiring a non-native contrast, it is predicted that learners will assimilate non-native phones to L1 sound categories if the two sounds are perceived as phonetically similar, but will create new categories for L2 sounds if that sound is not perceived as phonetically similar to any L1 sound [2]. If the L2 phone is perceived as similar to an L1 category, learners are expected to produce the L2 phone as if it were an L1 phone [2, 3]. However, learners often use different strategies than native speakers to produce non-native contrasts, which may obscure acquisition of a contrast [4]. In a study examining the perception and production of second language Korean fricatives, learners did not produce a target-like acoustic difference between L2 Korean alveolar fricatives but did perceive the contrast, suggesting the acoustic results obscure information about whether learners have acquired the contrast [5]. Learners may be using different articulatory strategies to produce a contrast than native speakers, which can lead to non-target like, or 'merged', acoustic productions. Thus, there is a possibility that learners are not transferring L1 categories to produce L2 phones, but only using acoustic data and native-like comparisons misinforms theories of category formation.

This study examines both the articulatory and acoustic properties of L1 English-L2 Mandarin learners' fricative productions. Mandarin has a three-way contrast between the fricatives /s/, /ʂ/, and /ɕ/ [6, 7]. It has been shown that L1 English learners of Mandarin have difficulty perceiving the difference between the Mandarin fricatives /ʂ/ and /ɕ/, and it is predicted that this is because they are assimilating the two Mandarin fricatives to their L1 English fricative /ʃ/ [1]. Because learners are predicted to have difficulty producing the difference between a contrast if they are perceived as the same L1 phoneme [2, 3], learners are expected to have difficulty producing the contrast between /ʂ/ and /ɕ/. Mandarin /s/, on the other hand, is predicted to be perceived as similar to English /s/, and thus produced as such. This study uses ultrasound tongue imaging to examine whether learners transfer L1 English acoustic or articulatory categories to produce Mandarin /ʂ/ and /ɕ/, and whether this differs from patterns of transfer for Mandarin /s/.

2. METHODS

2.1. Participants

8 L1 English-L2 Mandarin learners (2 male, 6 female, mean age=22) participated in a production task in English and in Mandarin. All learners were enrolled in intermediate Chinese language courses at a university in the U.S. at the time of

study. 2 L1 Mandarin speakers also completed the Mandarin production task (1 male, 1 female, mean age=25.5). Native speakers reported only speaking the Mandarin dialect of Chinese and speak Chinese on a regular basis with family in China and friends in the U.S.

2.2. Materials

The target Mandarin fricatives /s, ʂ, ɕ/ were presented in the onset of 6 Mandarin words, repeated twice for a total of 12 tokens of each target sound. The words were embedded in the carrier phrase [wə tɕei də __ xən xau], meaning "I think __ (is) very good". Majority of the words were bisyllabic words of the shape CVCV. Following vowel quality and tone were controlled for to the extent possible, while only using lexical items familiar to intermediate learners. All target items were chosen from a vocabulary list given to them in class. The target English fricatives /s, ʃ/ were also presented in the onset position of 6 bisyllabic English words, repeated twice for a total of 12 tokens of each target sound. The words were embedded in the carrier phrase "give me a __ again". Following vowel context was controlled for across the English fricatives, and were similar to the environment of the Mandarin following vowel contexts.

2.3. Procedure

Participants wore a head stabilization headset manufactured by Articulate Assistant Advanced (AAA) [8], attached to an ultrasound probe and a video camera. Participants also wore a head-mounted microphone. All acoustic and articulatory data were recorded using AAA. Learners completed the Mandarin task first, followed by the English task. The Mandarin words were written in Chinese characters.

2.4. Analysis

First, each speakers' audio prompts were aligned to transcriptions using MFA [9]. The aligned textgrids were hand-corrected in Praat and re-uploaded to AAA for the articulatory analysis. For the acoustic analysis, a Praat script was used to extract all the fricatives from the target phrases. Spectra were measured using multi-taper spectral analysis in the middle 20 ms portion of each fricative with the spectRum package for R [10, 11]. Spectral center of gravity and spectral speak were measured for each fricative, following from work that argues these acoustic measures correlate to the Mandarin three-way fricative contrast [12, 7].

For the articulatory analysis, splines were semi-automatically fitted to the ultrasound images for each speaker. Splines were extracted out of AAA at the acoustic midpoint using polar coordinates. The midpoint is chosen for analysis to avoid the coarticulatory effects of preceding or following vowel context as much as possible. For the L1 Mandarin speakers, the tongue positions of Mandarin fricatives /s, ʂ, ɕ/ were compared using SSANOVAs in order to confirm how speakers produce the articulatory contrast between the target segments. Next, the learner productions of Mandarin /s, ʂ, ɕ/ were compared within each speaker to see whether learners were using the same articulatory strategies as native speakers to produce this contrast. Finally, each learners' productions of Mandarin /s, ʂ, ɕ/ were compared to their English productions of /s, ʃ/ using SSANOVAs to determine whether learners were transferring L1 articulatory gestures to produce the L2 fricatives. All analyses were conducted within participants (rather than across participants) in order to show how each speaker is producing the contrast between Mandarin fricatives, and to examine whether learners were transferring their L1 categories to produce Mandarin fricatives.

3. RESULTS

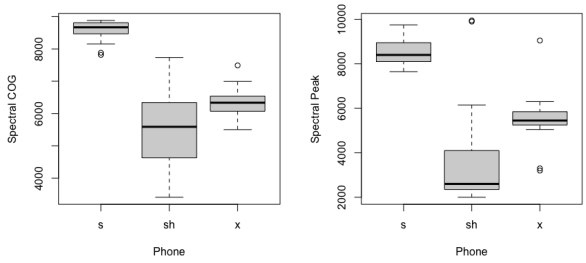
3.1. Acoustic Results

First, separate one-way ANOVAs for each native speaker with spectral COG as the dependent variable confirms that there is a significant difference in COG between the target fricatives (results for NM1 are shown in Figure 1a; F=29.38, p=3.77e-10***).

Similarly for spectral peak, separate ANOVAs confirm that there is a significant difference in spectral peak for native speaker productions' of Mandarin fricatives (F=51.08; p=1e-14***, shown in Figure 1b). Post-hoc Tukey HSD tests confirm that COG and spectral peak are different for all three fricatives.

Next, learner acoustic productions of the Mandarin fricative contrast were compared using one-way ANOVAs. Again, results are shown within-participant to confirm how learners are producing the non-native contrast. One learner's results are shown here as representative of group results, and other learner results are summarized.

Results show that all learners produce a difference in COG between the Mandarin fricatives. However, Tukey HSD tests confirm that while all learners produce a difference in COG between Mandarin

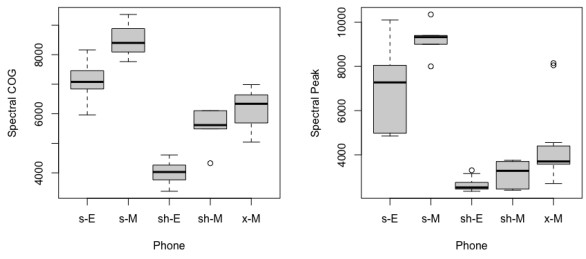


(a) Center of Gravity (Hz) **(b)** Spectral Peak (Hz)

Figure 1: Native speaker NM1's spectral measurements of Mandarin fricatives; Mandarin fricatives /ʂ/ and /ɕ/ are represented by their Pinyin orthography 'sh' and 'x'

/s/ and /ʂ/ and Mandarin /s/ and /ɕ/, only half of the learners produce a difference in COG between Mandarin /ʂ/ and /ɕ/. Results for learner LM4, who does not produce a difference between Mandarin /ʂ/ and /ɕ/, are presented in Figure 2a (F=92.08, p<2e-16***).

For spectral peak, separate one-way ANOVAs for each learner show that learners do produce a difference in spectral peak for the Mandarin fricatives. Post-hoc Tukey HSD tests show that all learners produce a difference between /s/ and /ɕ/, and again, half of the learners produce a difference between /ʂ/ and /ɕ/. Results for LM4, who does not produce a difference in spectral peak between Mandarin /ʂ/ and /ɕ/, are presented in Figure 2b; F=38.77, p=1.78e-14***). All learners except one (LM1) produce a difference between Mandarin /s/ and /ʂ/.



(a) Center of Gravity (Hz) **(b)** Spectral Peak (Hz)

Figure 2: LM4's spectral measurements of English and Mandarin fricatives

Finally, each learners' productions of their Mandarin fricatives were compared to their English fricatives using separate one-way ANVOAs for each acoustic measurement with fricative category as the dependent variable. There is a significant difference in COG between fricatives for all learners. Tukey HSD tests show that most learners do not produce a difference in COG between English /s/ and Mandarin /s/, or between English /ʃ/ and Mandarin /ʂ/. LM4 (Figure 2a) is the only speaker who produces a difference in COG between these fricatives. All learners produce a difference in COG between English /s/ and Mandarin /ʂ/, and English /s/ and Mandarin /ɕ/. Half of the learners produce a difference in COG between English /ʃ/ and Mandarin /ɕ/.

All learners produce a difference in spectral peak between fricatives as well. Post-hoc Tukey HSD tests show that no learners (except LM4, Figure 2b) produce a difference in spectral peak between English /s/ and Mandarin /s/. No learners produce a difference between English /ʃ/ and Mandarin /ʂ/. Half of the learners produce a difference English /ʃ/ and Mandarin /ɕ/. All learners do produce a difference between English /s/ and Mandarin /ʂ/, and English /s/ and Mandarin /ɕ/.

To summarize the acoustic results: native speakers produce a difference in COG and spectral peak between the three Mandarin fricatives /s/, /ʂ/, and /ɕ/. L1 English-L2 Mandarin learners do produce a difference in COG and spectral peak between the Mandarin fricatives /s/ and /ʂ/, and /s/ and /ɕ/, but only half produce a difference between Mandarin /ʂ/ and /ɕ/. Turning to the cross-language results, most learners produce Mandarin /s/ with the same COG and spectral peak as English /s/. Most learners also produce Mandarin /ʂ/ with the same COG and spectral peak as English /ʃ/. About half of the learners produce a difference in COG and spectral peak between English /ʃ/ and Mandarin /ɕ/. Many of these learners produce the Mandarin fricatives /ʂ/ and /ɕ/ with the same COG and spectral peak.

3.2. Ultrasound results

Native speaker NM1's SSANOVA results for the production of Mandarin fricatives are shown in Figure 3. There is a significant difference in native speakers' productions of the Mandarin fricatives at the constriction location.

Next, learner productions of the Mandarin fricatives were compared (again, within speaker) using SSANOVAs. Results for LM5's productions of Mandarin and English fricatives are shown in Figure 4a. As can be seen, this learner does not produce a difference in tongue position for Mandarin /s/ and /ʂ/. Half of the learners in this study do not produce Mandarin /s/ and /ʂ/ with significantly different tongue positions (similar to LM5), while half do produce a difference in tongue position (see Figure 4b for learner LM4's

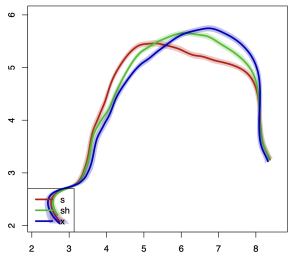
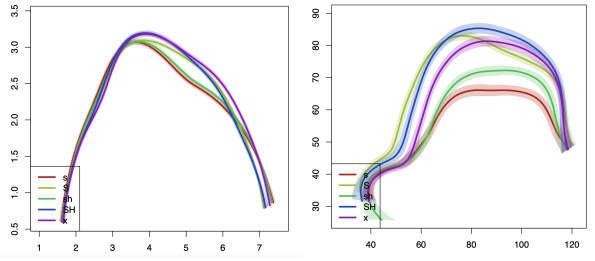


Figure 3: NM1's SSANOVA results (tongue tip to the right) for the production of Mandarin fricatives

SSANOVA results). However, Mandarin /ɕ/ is produced with a different tongue position than either of the other two Mandarin fricatives for learners. All learners in this study do produce a significant difference between Mandarin /s/ and /ɕ/ and Mandarin /ʂ/ and /ɕ/ at least along the tongue front.



(a) LM5's SSANOVA results **(b)** LM4's SSANOVA results

Figure 4: Learner SSANOVA results (tongue tip to the right) for the production of Mandarin and English fricatives (note that English fricatives /s/ and /ʃ/ are indicated as capital letters S and SH while Mandarin fricatives /s/, /ʂ/ and /ɕ/ are indicated by their Pinyin orthography s, sh, and x)

Turning to the cross-language results, all learners tend to produce a significant difference in tongue position between Mandarin /s/ and English /s/, and Mandarin /ʂ/ and English /ʃ/. Again, this is shown in Figures 4a and 4b. Interestingly, most learners produce their English fricatives with a higher position than their Mandarin fricatives. Finally, all learners do produce a difference between Mandarin /ɕ/ and English /ʃ/. This is surprising given the acoustic findings that half of the learners produce Mandarin /ɕ/ with similar spectral measures to English /ʃ/.

4. DISCUSSION

This study examines how L1 English-L2 Mandarin learners acquire the non-native contrast between the Mandarin fricatives /s/-/ʂ/-/ɕ/, and whether learners transfer acoustic categories from L1 fricatives to produce the non-native fricatives, or articulatory gestures. The results show individual differences in patterns of transfer. Importantly, learners who do not produce an acoustic difference between the Mandarin fricatives /ʂ/-/ɕ/ do produce an articulatory difference between these sounds, and do not transfer L1 articulatory gestures to produce these fricatives.

Half of the learners in this study do not produce a difference in spectral measures between Mandarin /ʂ/ and /ɕ/. These sounds are both acoustically similar to English /ʃ/ (although this varies across learners). This result may lead to the conclusion that the learners who do not produce the difference between these sounds have not acquired the contrast (or assimilated them to the same L1 category). However, articulatory results show that /ʂ/ and /ɕ/ are produced with different tongue positions by all speakers, including those who do not produce these fricatives with different spectral measures.

The lack of spectral difference between L2 Mandarin /ʂ/-/ɕ/ for many learners may be caused by learners producing this contrast as a difference in tongue height rather than constriction location. The spectral measures included in this study are correlated to the size of the front cavity between the lingual constriction and the lip opening, which does not reflect how these learners are producing the contrast. Importantly, native Mandarin speakers do produce a difference in COG and spectral peak between /s/ /ʂ/ and /ɕ/. The results presented here show that there are a variety of individual strategies learners use to produce the contrast, but these are often different articulatory strategies than native speakers. Only observing the acoustic results obscures the fact that learners do appear to have the contrast.

Finally, it is not the case that learners transfer L1 articulatory gestures to produce L2 Mandarin fricatives. There is no evidence of articulatory transfer, even for similar phones /s/ and /s/. Again, this was not seen in the acoustic results.

The results of this study highlight the fact learners often use different strategies to produce L2 contrasts from native speakers, and including both acoustic and articulatory data in the study of category formation can reveal covert contrasts. This study also shows that learners do not transfer L1 articulatory gestures to produce L2 fricatives, which suggests L1 transfer may not be as straightforward as assumed from models based solely on acoustic data.

5. REFERENCES

[1] X. Wang and J. Chen, "English speakers' perception of mandarin consonants: The effect of phonetic distances and l2 experience," *ICPhS 2019 Proceedings*, pp. 1–5, 2019.

[2] J. E. Flege, "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," *Journal of phonetics*, vol. 15, no. 1, pp. 47–65, 1987.

[3] ——, "Origins and development of the speech learning model," *Retrieved December*, vol. 13, p. 2005, 2005.

[4] R. McAllister, "Strategies for realization of l2-categories," *Language experience in second language speech learning: In honor of James Emil Flege. John Benjamins Publishing*, vol. 17, pp. 153–166, 2007.

[5] S. Y. Cheon, "Production of korean fricatives in second language acquisition: Acoustic characteristics," *Korean Linguistics*, vol. 13, no. 1, pp. 17–48, 2006.

[6] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.

[7] F. Li, J. Edwards, and M. Beckman, "Spectral measures for sibilant fricatives of english, japanese, and mandarin chinese," in *Proceedings of the 16th International Congress of Phonetic Sciences*. Pirrot, Dudweiler Germany, 2007, pp. 917–920.

[8] AAA, "Articulate instruments ltd.: Articulate assistant advanced user guide: Version 218.04." 2019.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[10] P. F. Reidy, "spectRum," 2013, r package. [Online]. Available: https://github.com/patrickreidy/spectRum

[11] R Core Team, "R language definition," 2015.

[12] S.-I. Lee, "Spectral analysis of mandarin chinese sibilant fricatives." in *ICPhS*, 2011, pp. 1178–1181.

IDENTIFICATION OF NON-NATIVE ENGLISH SPEAKERS' L1s VIA PATTERNS OF PROSODIC FEATURE DEVIANCE FROM NATIVE SPEAKER NORMS

Jim Talley

Linguistic Computing Systems, Austin, TX USA
talley@lingcosms.com

ABSTRACT

This study is an exploratory data science-based look at the question of whether the first language (L1) of non-native speakers of English can be identified from only a few simple syllable- and utterance-level prosodic features of their speech. Simple machine learning (ML) modeling on these loudness, pitch, and duration cues yields imperfect, but much better than chance, discrimination (1) between each individual L1 and General American English (GAE), and (2) between the five studied L1s.

The described modeling is based upon "atypicality scores" (*a-Scores*) for the prosodic features, representing the degree to which features deviate, or not, from GAE native speaker norms. The prosodic features, their normalizations, and the *a-Score* characterizations are discussed.

Finally, ML-based feature selection analysis examines the individual prosodic features' relative importances for the individual L1 vs. GAE discrimination tasks and for the 5-way, forced-choice L1 classification task.

Keywords: prosody, L2 English, feature atypicality, machine learning, corpus phonetics

1 INTRODUCTION

Prominences are fundamental in speech, serving to mark focused words, to aid in speech segmentation, to support lexical access, to bootstrap first language (L1) learning [1], and more. There are multiple means employed by languages to signal prominence – increases in duration, significant changes in pitch, increases in loudness, and (non-)reduction of vowels are all markers of stressed syllables in English [2]. For an L2 learner of English to achieve native-like prosody, he/she must master all four of those aspects in dynamic combination (even in trading relations with each other [3]). English is, of course, not alone in richly marking prominent words/syllables, but in some languages, the four recognized correlates of stress that English uses may be "occupied" by other linguistic functions – for example, in languages with lexical tone (*e.g.*, Vietnamese), speakers may be constrained with respect to pitch variations for stress marking, or in languages with phonemic length (*e.g.*, Finnish), duration changes to mark stress may lead to lexical confusions. This raises the question of whether ingrained prosodic habits related to the L1 of a non-native speaker (NNS) significantly, and predictably, influence his/her L2 prosody.

Such cross linguistic influence (CLI) certainly plays some role in L2 spoken language acquisition, though there has been robust discussion regarding the centrality or significance of that role [*e.g.*, 2,4,5]. This exploratory study is not an attempt to advocate for, or against, a strong position regarding the role of CLI. It merely asks, via a corpus and machine learning approach, if there are, indeed, reliably detectable manifestations of non-English L1 prosody on English L2 speech prosody.

2 THE DATA

The on-line Speech Accent Archive (SAA) [6] of English language recordings of the short "Stella" passage, read by thousands of native and non-native speakers, is a rich resource for addressing that question. This study selects SAA speakers from six L1s: General American English (**en**), Andean Latin American Spanish (**es**), Japanese (**ja**), Polish (**pl**), Russian (**ru**), and Mainland Mandarin (**zh**).

2.1 Speaker Selection

The SAA has an impressive variety of speakers, but, as a crowd-sourced resource, it also has considerable diversity with respect to the quality (and the equipment and environments) of recording. The L2 speakers' varying degrees of command of English is both an advantage and a challenge – false starts, self corrections, long pauses, and repetitions are abundant, which causes difficulties for automatic processing. Also, in SAA, L1s are not broken down regionally, so for example, native English speakers from the U.S., Ireland, Australia, India, England, etc. are lumped into a single English L1 pool, though metadata can be accessed to help tease them apart.

Most L1s in SAA have only a small set of speakers, and some of the more abundant languages suffer from the lack of homogeneity mentioned above. For this study, an average of 32 speakers were selected from the six L1s (**en**, **es**, **ja**, **pl**, **ru**, & **zh**)[1]. Regional homogeneity was enforced, but there was not an extensive effort to balance genders. Recordings were further pre-screened for noise, echo, or generally poor quality, and speakers who spent their early school years in an English-speaking country were also filtered out. Prosodically typologically diverse languages – *e.g.*, lexical tone, syllable timing, phonemic length – was also of interest, in the sense that we might expect more differentiated CLI.

2.2 Features

Given a backdrop of interest in ESL/EFL pedagogical issues, a primary criterion in selection of prosodic features for this study was intuitiveness – that is, that they be amenable to actionable interpretation in a pedagogical context. There are a number of reasonably well-established prosodic metrics, *e.g.*, the openSMILE suite [7] and the PVI variants [8]. However, these were generally formulated for technical purposes, not for interpretability by/for general L2 learners. We have, therefore, chosen instead to use a small, relatively simple set of utterance and syllable characterizations (4 and 10, respectively) which can be explained to an L2 student (*e.g.*, lBSNGapDur is linking).

2.2.1 Basic features

The fourteen basic loudness (red), pitch (green), and duration (blue) features are listed and described in Table 1. Note that these (automatically extracted) basic features are normalized in several ways to compensate for cross-speaker (and recording) variability. First of all, units (mels and dB) are chosen to provide approximately linear scaling perceptually. Then, they are normalized against utterance means (and, in the case of pitch, also by standard deviation). These transformations are important for limiting idiosyncratic variance for modeling. The utterance level features, in turn, are intended to capture potentially significant aspects of between-speaker variation, such as pitch dynamism.

2.2.2 Atypicality scores (a-Scores)

In addition to the normalizations applied in deriving the basic prosodic features, we convert those basic features to "atypicality scores" (*a-Scores*) for modeling by norming them against our population of GAE native speakers (NS). The *a-Scores* represent the degree to which prosodic features deviate (or not) from the normal variation observed among NS.

Sym	Short Name	Type	Description
	dBStdDev	Utt, Loud	Utterance loudness variability; more specifically, the stddev of the cNDur values from the utterance
	f0StdDev	Utt, Pitch	Utterance pitch variability; more specifically, the stddev of the utterance's syllables' nuclear f0 means (in mels)
	syllsPerSec	Utt, Dur	Speaking rate – the total utterance length (not including initial and final silences) divided by the number of syllables in the utterance
	pau2Spch	Utt, Dur	Utterance fluency – the ratio of silence time to speech time within the utterance (*i.e.*, not including initial and final silences)
	lDeltNLoud	Syll, Loud	Loudness change from the center syllable to its left neighbor, *i.e.*, $cNLoudDB_{i-1}$ - $cNLoudDB_i$
	cNLoudDB	Syll, Loud	Mean loudness (in dB) computed over the nucleus of the ith syllable, then divided by the mean of utterance's syllables' loudness values
	rDeltNLoud	Syll, Loud	Loudness change from the center syllable to its right neighbor, *i.e.*, $cNLoudDB_{i+1}$ - $cNLoudDB_i$
	lDeltZNF0	Syll, Pitch	Pitch change from the center syllable to its left neighbor, *i.e.*, $cZNormF0_{i-1}$ - $cZNormF0_i$
	cZNormF0	Syll, Pitch	Mean f0 (in mels) computed over the nucleus of the ith syllable, then z-scaled using the utterance's f0 mean and stddev
	rDeltZNF0	Syll, Pitch	Pitch change from the center syllable to its right neighbor, *i.e.*, $cZNormF0_{i+1}$ - $cZNormF0_i$
	lBSN-GapDur	Syll, Dur	Duration of the leftward between-syllable gap (*i.e.*, the length of pause, if any, between the $(i-1)^{th}$ and i^{th} syllables) divided by the mean duration of the utterance's syllables
	lDeltNDur	Syll, Dur	Duration change from the center syllable to its left neighbor, *i.e.*, $cNDur_{i-1}$ - $cNDur_i$
	cNDur	Syll, Dur	Duration of the ith syllable divided by the mean duration of the utterance's syllables
	rDeltNDur	Syll, Dur	Duration change from the center syllable to its right neighbor, *i.e.*, $cNDur_{i+1}$ - $cNDur_i$

Table 1: Prosodic features used in this study

For v_x, a basic feature value from language x, we first calculate its normed value z_x (Equation 1) using (Yeo-Johnson transformed [9]) **en** sample statistics:

$$z_x = \frac{v_x - \mu_{en}}{\sigma_{en}} \quad (1)$$

Then, z_x is "soft capped" (logarithmically squashed) if its magnitude exceeds a threshold of c=3 standard deviations (Equation 2) to produce the *a-Score*, a_x:

$$a_x = \begin{cases} z_x & \text{if } -c \le z_x \le c \\ sign(z_x) \times (c + \log_e(1 + (|z_x| - c))) & \text{otherwise} \end{cases} \quad (2)$$

The logic behind the squashing is that more than $\pm c$ stddevs qualifies as very atypical, and, while we would still like monotonic increase (decrease) in value, any (possibly large and possibly spurious) amount beyond c should have limited influence in combinations with other features' *a-Scores*.

3 MODELING

Exploration with respect to L1 transfer of prosodic characteristics is carried out via ML model training on the *a-Score* features. Ensembles of small, simple models (5 utterance and 77 syllable models, for 82 in all) are trained to produce independent likelihood estimates. The likelihoods are then fused (by simply averaging them) and converted into L1 classifications for the speakers (forced-choice – highest mean likelihood wins). Simple ML methods with few trainable parameters are used to avoid overfitting on the rather small data set.

All results mentioned in this paper are derived from linear regression estimators (LREs)[2] as implemented in Scikit-Learn [10]. The LRE models, importantly, are trained with sample weighting (to offset class size imbalances). Furthermore, in all cases, estimation is done using the leaving-one-out method – that is, when training/evaluating on a 200-item sample set, 200 models are trained, each being trained on 199 samples and each producing a single estimation for the one sample which was left out of its training set.

4 CLASSIFICATION RESULTS

We directly consider how distinguishable the 5 NNS L1s are from each other (given the *a-Score* featurization of the prosodic cues) via a 5-way forced-choice classification between them. The results (displayed in the normalized[3] confusion matrix (CM) of Figure 1) are far from perfect at 58.04% correct overall, but also far above chance (~20%). Results are fairly uneven across the five L1s – **ru** is commonly misidentified as **pl** (and also **es**), and **ja** is confused as **zh** about a quarter of the time.

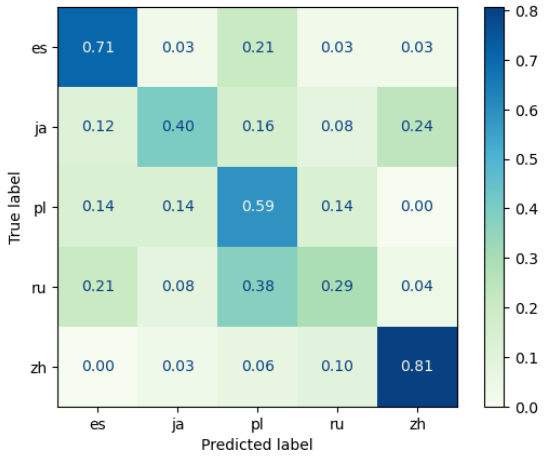


Figure 1: 5-way, forced-choice L1 classification

To test the conjecture that the "non-nativisms" of the L1s were interfering with each other, thus impacting achievable accuracy, we also modeled each L1 separately vs. **en** – *i.e.*, asking whether speakers' L1s were reliably identifiable, vis-a-vis the native speakers, based on their prosody.

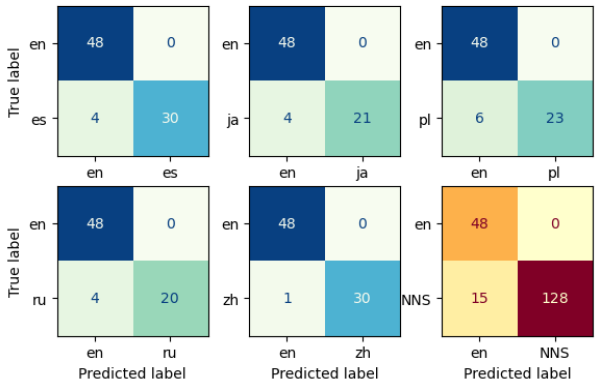


Figure 2: Separately distinguishing each NNS L1 from **en** (NS), plus **en** vs. the pooled NNS L1s

Figure 2 shows the results of the set of 5 binary L1 models (unnormalized CMs), plus the results of a binary model built from all of the NNSs pooled together vs. **en**. The **en**-NNS model was correct 92.15% of the time, while the **en**-L1 models averaged 95%.[4] So, clearly, the prosodic features do support distinguishing the L1s from GAE (**en**).

5 FEATURE IMPORTANCE RESULTS

Given demonstration of the capacity to more or less identify L1s from these data, we would also like to

know what prosodic features actually matter for performing that task. One approach to estimating the relevance of the various prosodic cues (as manifested by the data features used in the study) is to look at how their presence, or absence, affects the ability to carry out an identification task. Ideally, we would like to consider every one of the possible combinations of features in that assessment[5], but instead opted for the common, but sub-optimal, compromise of doing greedy search.

Greedy feature selection starts with single features – training models with them individually and picking the feature which provides the best performance on the task. It then extends to two features by considering remaining features in combination with the 1st ranked feature, and so on.

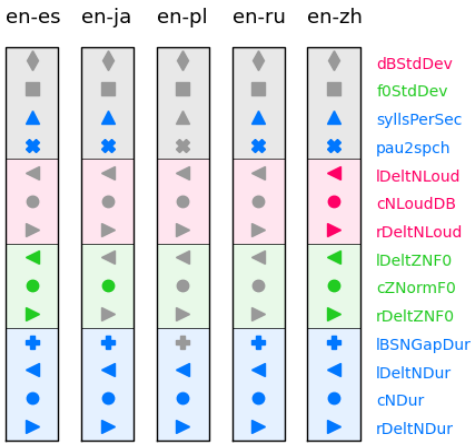


Figure 3: Features, by L1, which individually can discriminate the L1 from **en** at 80% or higher

It would be of definite interest to be able to derive robust, ordered lists of the most important features distinguishing each of the L1s from **en**. Unfortunately, there was too much of a ceiling effect with that analysis – they almost all approached 100% accuracy with only one or two features (with many features tied for the top rankings). Figure 3 is a compromise. It shows the features (colored) which are (individually) most effective for the **en**-L1 binary classifications. Note the predominance of (blue) duration features. Figure 3 also provides some potential insight with respect to performance on the 5-way classification task – *i.e.*, it suggests that there may be clashes among the various L1s' most informative features. Best performing **zh** notably has a number of unique (clash-free) top features.

Table 2 is an ordered list of the top features found for the 5-way L1 discrimination task. Once again, duration cues are prevalent, even though cross-L1 clashes may be limiting their effectiveness.

Rank	5-way, Forced-choice between L1s	
	Feature	*%Corr*
1	syllsPerSec	29.37
2	pau2Spch	37.76
3	f0StdDev	40.56
4	dBStdDev	36.36
5	cZNormF0	34.27
6	cNDur (rDeltNDur)	44.76
7	rDeltZNF0	52.44
8	rDeltNDur	55.94
9	lBSNGapDur	57.34

Table 2: Feature importance (greedy search)

It is probably unwise to put too much stock in the specific selection of features here. The automatically extracted data (from the challenging SAA recordings) are still fairly rough. To make more confident conclusions, we would want to work with a significantly larger, cleaner corpus.

6 SUMMARY

This study was primarily conducted to prototype a new methodology, including: 1) choosing a set of features with pedagogical potential; 2) defining a robust representation scheme for those features (the *a-Scores*); and, 3) validating an ensemble-of-limited-experts approach to small parameter space ML. The overarching question being asked of the corpus with this methodology was whether prosodic footprints of L1s are detectable in L2 English.

The task – roughly the equivalent of guessing the L1 of a speaker after hearing the melody of a subject's speech played on a piano – would likely be very challenging for humans. With noisy data, an impoverished representation of the speech (lacking, presumably useful, segmental information) and weak modeling, expectations for the study were low.

It is quite clear from the successful binary **en**-L1 and **en**-NNS results that non-native speakers are distinguishable from native speakers – *i.e.*, NNS footprints exist in the data. The 5-way forced-choice task speaks more directly to the question of whether characteristic L1 footprints also exist. That question is less conclusively answered. While the accuracy exceeded expectations, coming in well above chance, it was somewhat uneven across L1s and clearly more difficult than **en** vs. NNS classification.

7 REFERENCES

[1] Gervain, J. 2022. Word frequency and prosody bootstrap basic word order in prelexical infants. Speech Prosody 2022.

[2] Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. 2014. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, 57(4), 1468-1479.

[3] Howell, P. 1993. Cue trading in the production and perception of vowel stress. *J. Acoust. Soc. Am*. (Oct) 94(4), 2063-2073.

[4] Munro, M. 2018. How well can we predict second language learners' pronunciation difficulties? *The CATESOL Journal* 30.1, 267-281.

[5] Wardhaugh, R. 1970. The Contrastive Analysis Hypothesis. TESOL Quarterly, 4, 123.

[6] Weinberger, S. 2015. Speech accent archive. George Mason University. http://accent.gmu.edu/.

[7] Coutinho, E., Hoenig, F., Zhang, Y., Hantke, S.; Batliner, A., Noth, E., & Schuller, B. 2016. Assessing the prosody of non-native speakers of English: Measures and feature sets. *LREC* 2016.

[8] Low, E.L., Grabe, E., & Nolan, F. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43 (4), 377–401.

[9] Yeo, I.K. & Johnson, R.A. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.

[10] Pedregosa *et al*., 2011. Scikit-learn: Machine Learning in Python, *JMLR* 12, 2825-2830.

---

[1] The selected SAA speaker numbers were: **en** (english<*n*>) – 9, 10, 32, 36, 39, 49, 51, 54, 55, 59, 60, 62, 63, 65, 68, 71, 74, 75, 78, 81, 82, 88, 90, 92, 96, 99, 106, 107, 114, 117, 118, 121, 123, 124, 126, 128, 133, 138, 143, 150, 157, 161, 162, 163, 165, 169, 171, 173; **es** (spanish<*n*>) – 3, 6, 7, 11, 13, 14, 18, 21, 23, 32, 42, 52, 53, 64, 67, 68, 70, 75, 78, 79, 80, 82, 83, 87, 88, 96, 97, 100, 102, 108, 111, 115, 123, 124; **ja** (japanese<*n*>) – 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29; **pl** (polish<*n*>) – 1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 31, 32, 34, 35, 36; **ru** (russian<*n*>) – 1, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 24, 30, 34, 36, 37, 41, 45, 47, 49, 50; **zh** (mandarin<*n*>) – 1, 2, 3, 4, 5, 6, 8, 14, 17, 19, 21, 22, 25, 26, 28, 31, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 51, 52, 53

[2] Experiments were done with other modeling methods (k-Nearest Neighbors regressors, small neural networks (MLPs),...) but none were found to give accuracy advantages over the linear models, and training/evaluation time for the linear models was the fastest of the considered ML methods.

[3] The confusion matrix of Table 1 is normalized to fractions of the true speakers (rows).

[4] Individual **en**-L1 accuracies: **en-es**=95.12%, **en-ja**=94.52%, **en-pl**=92.21%, **en-ru**=94.44%, and **en-zh**=98.73%

[5] Given the structure of the modeling in this work, the 5-way, forced-choice task alone would require training ~1.4B models for exhaustive feature combination exploration, while greedy search reduces that to a more manageable ~7.3M models trained.

IS RHOTICITY ON THE TIP OF YOUR TONGUE? INVESTIGATING TONGUE SHAPES FOR ENGLISH /r/ IN FRENCH LEARNERS WITH ULTRASOUND

Alice Léger, Hannah King, Emmanuel Ferragne

Université Paris Cité, CLILLAC-ARP, F-75013 Paris, France
alice.leger@etu.u-paris.fr, hannah.king@u-paris.fr, emmanuel.ferragne@u-paris.fr

ABSTRACT

English /r/ takes different tongue shapes from one speaker to another. It is well-established that *tip-down* and *tip-up* shapes produce perceptually similar outputs. However, it remains unclear why speakers intuitively acquire one type or another. The present study considers the hypothesis that rhotic and non-rhotic varieties of English may influence the acquisition of different tongue shapes. We provide articulatory data on the pronunciation of English /r/ by 19 French learners of English, 10 with rhotic and 9 with non-rhotic accents. Ultrasound tongue images were recorded for onset and coda /r/ in various vocalic contexts and were classified as either tip-up or tip-down. Although rhoticity as a predictor of tongue shape does not reach statistical significance, we found a tendency for rhotic speakers to use a higher proportion of tip-down shapes. We conclude that while rhoticity may partly influence tongue shape, other factors are also at play, including co-articulatory constraints.

Keywords: rhotic, articulation, English, French learners

1. INTRODUCTION

The English post-alveolar approximant /r/ has intrigued researchers for its unusual articulatory variability. Native speakers produce equivalent acoustic outputs, at least up to the third formant [1], with various degrees of retroflex and bunched tongue shapes [2, 3]. Retroflex configurations are *tip-up* and raised towards the hard palate, while bunched configurations are *tip-down*, with the dorsum retracted. However, it remains unclear why individuals develop tip-down and/or tip-up shapes. Across studies, some consistent coarticulatory patterns of tongue shape distribution emerge, with tip-down shapes typically occurring with high front vowels [3, 4, 5] and in coda position [6]. Physiological differences such as oral cavity size or palate domedness have also been suggested to affect articulatory variability [7, 8, 9]. Sociolinguistic factors are reported in Scottish English, with tip-down and tip-up variants prevailing in middle-class and working-class speakers, respectively [10]. However, given the lack of perceptual distinction between tongue configurations in American English [11], the social distribution of /r/ may rather be due to a perceptible temporal contrast between tongue shapes specific to Scottish English [12].

The present study examines Heyne et al.'s [13] hypothesis that accent rhoticity may influence tongue shape acquisition. Most research on /r/ has focused on rhotic Englishes [2, 3, 10], but more recent studies on non-rhotic Englishes [13, 4] indicate a stronger preference for tip-down patterns in rhotic than in non-rhotic accents. As tip-down shapes are preferred in postvocalic position, non-rhotic varieties, which do not have /r/ in that context, may therefore discourage the development of a tip-down variant [13].

To investigate the influence of rhoticity on tongue shape acquisition, we looked at the articulation of French learners of English. Despite non-rhotic Standard Southern British English being traditionally taught in French schools [14], we observe both rhotic and non-rhotic accents among students. Learners' articulations of English /r/ are largely understudied, except for Polish [15] and Mandarin learners [16]. And yet, with its complex set of three constrictions at the lips, palate and pharynx [2, 17, 18], English /r/ is likely a challenging new sound for learners. Indeed, /r/ is one of the latest sounds to be acquired in native English-speaking children [19, 20]. As suggested by the Speech Learning Model (SLM) [21, 22] and the Perceptual Assimilation Model (PAM) [23, 24], influence from the L1 system is expected when learning a non-native sound. Nevertheless, the French rhotic /ʁ/ appears phonetically distinct enough from English /r/ to prevent the assimilation of the two sounds in French learners. Substitutions for [w], closer in articulation, also remain rare [25].

As language classes in France rarely include

pronunciation training, the acquisition of lingual gestures for /r/ in French learners appears rather implicit. Interestingly though, King and Ferragne [4] were not able to find mentions of a tip-down variant in pronunciation manuals. Thorough descriptions of tip-up constrictions alone are provided, with Ashton and Shepherd [26] notably considering it the 'correct position' relative to the tip-down alternative.

We explored the tongue shape patterns acquired by French learners of English in order to determine whether rhoticity contributes to the acquisition of a specific lingual gesture. If rhoticity indeed influences acquisition, we predict to find more tip-down patterns in learners with a rhotic accent and more tip-up patterns in learners with a non-rhotic accent. In addition, we expect learners to develop similar preferences to native speakers in the distribution of tongue shapes. That is, a stronger preference for tip-down shapes (i) surrounding high-front vowels, and (ii) in coda position.

2. METHODOLOGY

2.1. Participants

Data were collected from 20 French learners of English (13F, 7M), aged between 18 and 32. All were students at Université Paris Cité and had a rhotic (R) or non-rhotic (NR) accent. Rhoticity was perceptually assessed prior to the experiment based on audio recordings of 12 English sentences containing 22 possible productions of coda /r/. One subject's data was excluded due to ultrasound visualisation issues. We provide results from 10 rhotic and 9 non-rhotic learners. The study was granted ethical approval by the Université Paris Cité's ethics board.

2.2. Procedure

Participants read a randomised list of 14 monosyllabic words in isolation, each one repeated three times. /r/ occurred in word-initial and word-final position in various vocalic contexts, before /iː, ɪ, e, æ, ɑː, ɒ, ʌ, ɔː, uː/ and after /ɪ, e, ɑː, ɜː, ɔː/. Ultrasound tongue images were collected using an Echo Blaster 128 unit with a 5-8MHz transducer along with audio data in the Articulate Assistant Advanced (AAA) software [27]. The probe was stabilised relative to the head using an UltraFit Headset [28]. Before recording, a bite plate was used to adjust the probe-to-chin angle to each speaker's occlusal plane to improve interpretation of tongue position [29]. A Focusrite Scarlett 6i6 sound card and a Sennheiser t.bone Earmic500 condenser microphone were used for audio recording.

2.3. Data processing

For each /r/ produced, ultrasound frames were manually extracted at the point of maximal constriction and were visually classified as tip-up (TU) or tip-down (TD). Sequential frames were holistically examined to better assess the position of the tongue tip, as in [3, 4]. Tongue contours in the frames depicting maximal constriction were tracked automatically in AAA and manually corrected when necessary. The resulting splines were extracted in Cartesian coordinates, yielding 42 data points per token. Missing values at either end of a tongue contour were removed whereas missing values inside contours were reconstructed based on a moving average filter with a span of 4 data points. We present tongue contour data from 17 speakers as technical issues prevented us from tracking the contours in 2 speakers. Among the remaining 714 contours, 119 were discarded as no data point could be extracted. The size of our final 595 tokens ranged from 12 to 32 data points ($\mu = 24.16, \sigma = 3.43$).

One of the statistical models we used (Section 2.4) included discrete cosine transform (DCT) coefficients (estimated from the Y coordinates of each contour) as predictors. We computed these in an attempt to summarise overall tongue shapes with as few parameters as possible, which might constitute an alternative characterization of tongue shapes when data points at the tip are not available. Earlier uses of the DCT in phonetics have applied to the modelling of formant trajectories [30]. Here, based on data from speaker 128, we observed that an inverse transform with the first four coefficients offered a good approximation of the original contours and therefore kept these four coefficients for our model.

2.4. Statistical analysis

A binary decision tree was fitted to the data with speaker, speaker rhoticity, word, and syllable position (onset vs coda) as predictors and tongue shape (TU vs TD) as response. The accuracy of the model was computed using 10-fold cross-validation. The split criterion was Gini's diversity index. The maximum number of splits was constrained to 4 in order to make for optimal interpretability.

Tongue shape classification was further assessed using binomial generalised linear mixed-effects models (GLMM). The maximal set of successfully converging random slopes and intercepts were

included for speakers and where possible, words. The significance of main effects was tested using likelihood ratio tests. Indications of significance in the final models were calculated with Satterthwaite's approximations for degrees of freedom.

3. RESULTS & DISCUSSION

Figure 1 displays the tongue contours in 17 speakers. TU shapes appear in blue and TD in yellow. Speakers are sorted according to rhoticity (R vs NR). We observed more TU than TD shapes overall. Just two speakers (151, 214) systematically used TD shapes, both of whom belonged to the rhotic group.

In total, 666 ultrasound images of /r/ (252 in non-rhotic speakers; 414 in rhotic) were classified as either TU or TD. 3.2 % and 34.1 % of the tokens were classified as TD in the non-rhotic and rhotic group, respectively. As non-rhotic speakers rarely produced coda /r/ (11 tokens), the proportion of TD shapes in onset /r/ was also evaluated. Onset /r/ was classified as TD in 3.3 % of the tokens in non-rhotic and 24.4 % in rhotic speakers. These results follow our predictions that TD shapes are more common in rhotic than in non-rhotic speakers and that within rhotic speakers, there is a stronger preference for TD shapes in coda than in onset position.

The cross-validated binary decision tree model reached 96 % accuracy. As Figure 2 shows, the model picked speaker and word as relevant predictors and left out speaker rhoticity and syllable position. Although the model did not use rhoticity as a predictor, we note that all the speakers predicted to use a TD shape (128, 151, 214, 308) were in the rhotic group. This suggests that while having a rhotic accent does not mean that a speaker will systematically use a TD shape, TD shapes are much more typical of rhotic than of non-rhotic speakers. In the two speakers with variable tongue shapes (128, 308), TD shapes are predicted to occur in coda and only with high front vowels (/iː, ɪ/) in onset.

As rhotic and non-rhotic speakers all produced onset /r/, we ran a GLMM with tongue shape as the response variable on the onset data, regressed against speaker rhoticity and vocalic context. Random intercepts for subjects were included. While the main effect of rhoticity was not significant [$\chi^2(1) = 0.12$, $p = 0.729$], vocalic context was significant [$\chi^2(8) = 89.93$, $p < .001$]. The final model indicated that for an average speaker, a TD tongue shape is significantly more likely to occur in the context of /iː/ than it is for /e, æ, ɑː, uː, ʌ, ɔː, ɒ/. No significant difference was observed between /iː/ and /ɪ/. Although rhoticity failed to reach significance, these results follow our prediction that TD shapes are more likely to occur with high-front vowels, as observed in previous studies.

To assess the impact of syllable position, a further GLMM was run with tongue shape as the response variable on the data from rhotic speakers only. Main effects included syllable position and vocalic context. Random intercepts were included for subjects and words. Both main effects were significant [position: $\chi^2(1) = 11.84$, $p < .001$; vowel: $\chi^2(8) = 15.65$, $p < .05$]. The resulting model showed that for the average rhotic speaker, TD tongue shapes are significantly more likely to occur in coda than in onset, following our initial prediction. Regarding the effect of vocalic context, the same results as in the previous onset model were observed, again following our intial prediction that surrounding high-front vowels (/iː/ and /ɪ/) are more likely to result in TD shapes than the other vowels in the dataset.

As discussed in Section 2.3, DCT coefficients were computed in an attempt to summarise overall tongue shapes. Tongue tip height was also calculated by extracting the Y value of the rightmost Cartesian coordinate in each tongue contour. We then explored how these five numerical variables (DCTs 1-4 & tip height) may predict our manual coding of tongue shape in a final GLMM. Tongue shape as the response variable was regressed against the five numerical variables, which were converted to z-scores. Random intercepts were included for speakers. The addition of random intercepts for words failed to converge. The following main effects were significant: DCT1 [$\chi^2(1) = 20.99$, $p < .001$]; DCT3 $\chi^2(1) = 47.29$, $p < .001$]; DCT4 [$\chi^2(1) = 17.02$, $p < .001$] and tip height [$\chi^2(1) = 28.52$, $p < .001$]. While DCT2 failed to reach significance [$\chi^2(1) = 0.43$, $p = 0.514$]. Tokens with higher DCT1, higher DCT3, lower DCT4 and lower tongue tip values were significantly more likely to be coded as TD. With regards to tongue tip height, these results are coherent, as TD tongue shapes should naturally have a lower tongue tip than TU shapes. Higher DCT1 - sometimes referred to as the zeroth coefficient [30] - values suggest that the mean Y value in TD is higher. The absence of a difference in DCT2 tells us there is no evidence supporting a potential distinction in terms of magnitude and direction of change from the mean. The difference in DCT3 indicates that the degree of U-shaped curvature is greater in TD contours. And lower DCT4 values in TD show that TU contours contain energy in the higher frequencies, i.e. potential abrupt changes in trajectories. While our use of the

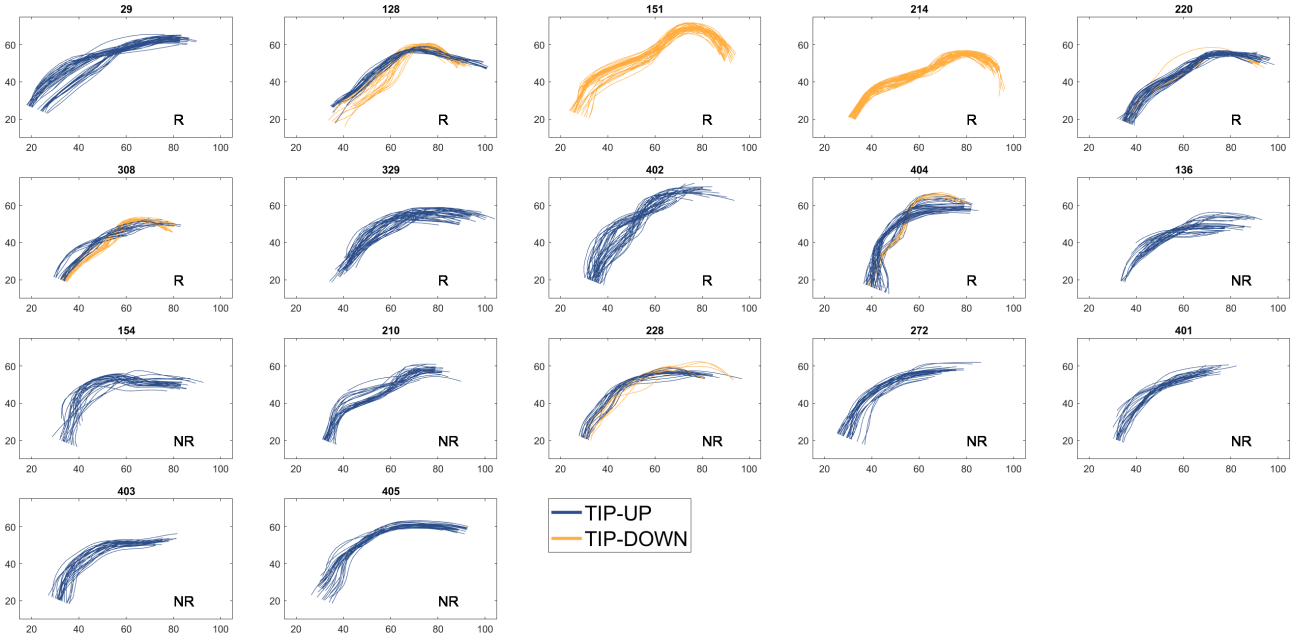


Figure 1: Tongue shapes in 17 participants. R: rhotic; NR: non-rhotic.

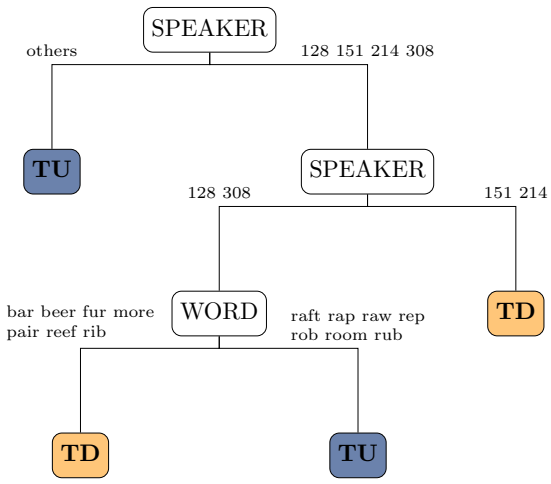


Figure 2: Classification tree.

DCT here remains exploratory, and more evidence has yet to be provided as to the benefits of such an approach, we must note that the DCT produced a compact and potentially interpretable representation of our contours.

4. CONCLUSION

This study aimed to evaluate if rhotic and non-rhotic accents constrain lingual gesture acquisition of /r/ in French learners of English. Although our results predict no significant effect of rhoticity, rhotic learners used proportionally more tip-down configurations than non-rhotic learners. In addition, speakers who used a fully tip-down pattern all belonged to the rhotic group. Tip-down shapes were significantly more frequent in coarticulation with high front vowels and in coda position, thus mirroring tendencies found in native speakers. We conclude that while having a rhotic accent does not mean that a speaker will systematically use tip-down shapes, tip-down shapes occur more frequently in rhotic than in non-rhotic speakers. This native-like production of a non-native sound suggests that the acquisition of /r/ lingual gestures is highly constrained by mechanical or physiological factors. Given that tip-up shapes prevailed across groups regardless of these preferences, we suggest, following Mielke et al. [3], that tip-up shapes may be the default variant, and tip-down gestures context-specific.

5. REFERENCES

[1] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of retroflex and bunched American English /r/," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4466–4481, 2008.

[2] P. Delattre and D. C. Freeman, "A dialect study of American r's by X-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[3] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex

/ɹ/," *Language*, vol. 92, no. 1, pp. 101–140, 2016.
[4] H. King and E. Ferragne, "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics*, vol. 80, p. 100978, 2020.
[5] D. Ong and M. Stone, "Three dimensional vocal tract shapes in /r/ and /l/: A Study of MRI, Ultrasound, Electropalatography, and Acoustics," *Phonoscope*, vol. 1, no. 1, pp. 1–13, 1998.
[6] J. M. Scobbie, E. Lawson, S. Nakai, J. Cleland, and J. Stuart-Smith, "Onset vs. coda asymmetry in the articulation of English /r/," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, 2015.
[7] J. Brunner, S. Fuchs, and P. Perrier, "On the relationship between palate shape and articulatory behavior," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3936–3949, 2009.
[8] S. Bakst, "Palate shape influence depends on the segment: Articulatory and acoustic variability in American English /ɹ/ and /s/," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 960–971, 2021.
[9] D. Dediu and S. R. Moisik, "Pushes and pulls from below: Anatomical variation, articulation and sound change," *Glossa: a journal of general linguistics*, vol. 4, no. 1, 2019.
[10] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "The social stratification of tongue shape for postvocalic /r/ in Scottish English," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 256–268, 2011.
[11] A. Twist, A. Baker, J. Mielke, and D. Archangeli, "Are "covert" /ɹ/ allophones really indistinguishable?" *Penn Working Papers in Linguistics*, vol. 13, no. 2, p. article 16, 2007.
[12] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of scottish sociophonetic variation," *Journal of Phonetics*, vol. 41, no. 3-4, pp. 198–210, 2013.
[13] M. Heyne, X. Wang, D. Derrick, K. Dorreen, and K. Watson, "The articulation of /ɹ/ in New Zealand English," *Journal of the International Phonetic Association*, vol. 50, no. 3, pp. 366–388, 2020.
[14] A. Przewozny-Desriaux, "Research Environments For Pedagogical Tools: Teaching and Learning with Authentic Oral English," *Recherche et pratiques pédagogiques en langues de spécialité*, no. Vol. 37 N°2, 2018.
[15] P. Lyskawa, "The ultrasound study of /ɹ/ in non-native speakers," in *Proceedings of International Congress of Phonetic Sciences*. London, UK: International Phonetic Association, 2015.
[16] S. Chen and P. P. K. Mok, "Speech production of rhotics in highly proficient bilinguals: Acoustic and articulatory measures," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019.
[17] A. Alwan, S. Narayanan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1078–1089, 1997.
[18] M. K. Tiede, S. E. Boyce, C. K. Holland, and K. A. Choe, "A new taxonomy of American English /r/ using MRI and ultrasound," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2633–2634, 2004.
[19] A. B. Smit and L. Hand, "The Iowa Articulation Norms Project and its Nebraska Replication," *Journal of Speech and Hearing Disorder*, vol. 55, pp. 779–798, 1990.
[20] L. Cialdella, H. Kabakoff, J. L. Preston, S. Dugan, C. Spencer, S. Boyce, M. Tiede, D. H. Whalen, and T. McAllister, "Auditory-perceptual acuity in rhotic misarticulation: Baseline characteristics and treatment response," *Clinical Linguistics & Phonetics*, vol. 35, no. 1, pp. 19–42, 2021.
[21] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.
[22] J. E. Flege and O.-S. Bohn, "The revised speech learning model (slm-r)," *Second language speech learning: Theoretical and empirical progress*, pp. 3–83, 2021.
[23] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception, commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*. John Benjamins Publishing, 2007, pp. 13–34.
[24] C. T. Best, "A direct realist view of cross-language speech perception," *Speech perception and linguistic experience*, pp. 171–206, 1995.
[25] I. Trifu-Dejeu, N. Herry-Bénit, and S. Lopez, "Le /r/ de l'anglais dans le corpus IPCE-IPAC : Évaluation quantitative et qualitative des productions orales des apprenants francophones," *Anglophonia*, no. 32, 2021.
[26] H. Ashton and S. Shepherd, *Work on your accent*, 2nd ed. Glasgow: Collins, 2020.
[27] Articulate Instruments Ltd., *Articulate Assistant Advanced User Guide: Version 2.19*. Edinburgh, UK: Articulate Instruments Ltd., 2012.
[28] L. Spreafico, M. Pucher, and A. Matosova, "UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science," in *Proceedings of Interspeech*, Hyderbad, India, 2018, pp. 1517–1520.
[29] E. Lawson, J. Stuart-Smith, and L. Rodger, "A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4363–4381, 2019.
[30] J. Elvin, D. Williams, and P. Escudero, "Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 576–581, 2016.

L1 PHONOTACTIC CONSTRAINTS MEDIATE PHONOLOGICAL AWARENESS OF NON-NATIVE GEMINATION

Paolo Mairano[1], Fabián Santiago[2]

[1]UMR 8163 STL – CNRS & Lille University, [2]UMR 7023 SFL – CNRS & Paris 8 University
paolo.mairano@univ-lille.fr, fabian.santiago.ling@gmail.com

ABSTRACT

Previous studies have revealed that L2 learners whose L1 has contrastive gemination may produce long consonants in an L2 which does not have gemination, due to orthography (e.g., in words with double consonants, such as 'Finnish' vs 'finish'). For the same reason, learners may reject rhymes such as 'very' and 'cherry'. In this study, we confirm that gemination affects the phonological awareness of L1 Italian learners of L2 French, but we find that it is mediated by L1 phonotactics. 24 Italian learners of French and 24 French control speakers participated in a discrimination test where the duration of target consonants was manipulated. Italian learners tended to rate stimuli with a lengthened consonant as different (non-)words, but only in contexts where gemination is licit in Italian. Vice versa, in contexts where gemination is illicit, responses by Italian learners do not differ from those of control French native speakers.

Keywords: second language acquisition, gemination, phonotactics, consonantal length, L2 perception.

1. INTRODUCTION

Phonological contrasts in quantity (i.e., gemination) for consonants exist in languages such as Italian, Arabic, and Japanese. Consonant duration is known to be the main cue of gemination, though secondary cues have been uncovered for various languages. In Italian, geminate consonants have been found to be approx. twice as long as singleton consonants in isolated words [1], but only approx. 1.7 times longer in running speech [2]. Geminating languages often have geminate consonants in intervocalic position, but far more rarely within clusters and in word-initial or word-final position [3]; gemination in these contexts is therefore considered as typologically marked.

Recent studies have uncovered that second language speakers whose L1 has contrastive gemination may produce long and short consonants in an L2 even if gemination is not lexically contrastive in the target language. We refer to this as *non-native gemination*. This phenomenon has been observed for Italian$_{L1}$ [4] and Japanese$_{L1}$ [5] learners of English$_{L2}$, as well as Italian$_{L1}$ [6] and Arabic$_{L1}$ [7] learners of French$_{L2}$. The presence of non-native geminate consonants is attributed to the effect of orthography: Bassetti et al. [4] demonstrated that English homophones such as '*finish*' and '*Finnish*' are pronounced respectively with a short vs long consonant by Italian$_{L1}$ learners following Italian grapheme-phoneme correspondence rules, since these words are spelled with a single vs double letter. The authors argue that such minimal pairs are evidence of the contrastive role of gemination in the mind of Italian$_{L1}$ learners, and also find that single vs double consonants in spelling affect rhyming judgments [8] and perception [9]. Mitterer [10] did not find comparable results with Maltese$_{L1}$ learners of English$_{L2}$, whose acquisition is less based on orthography (English$_{L2}$ is used outside the classroom), and concluded that the orthographic effect found for Italian$_{L1}$ learners is due to focus on orthography in formal education.

Interestingly, it seems that Italian$_{L1}$ learners may not only produce longer consonants for words like '*Finnish*', but also (though far more moderately) for words where gemination is not possible in Italian and typologically marked, namely in clusters ('*guessed*') and word-final position ('*add*') in English$_{L2}$ [4] and French$_{L2}$ ('*patte*' [pat], [6]). Although it is well-known that L1 phonotactics can have an impact on L2 production (e.g., Mandarin$_{L1}$ and Japanese$_{L1}$ learners produce an epenthetic / excrescent vowel for complex consonant clusters in English$_{L2}$, cf. [11, 12]) and L2 perception [13], the results on non-native gemination mentioned above suggest that the orthographic effect of double consonants override at least partially the phonotactic effect restricting contexts for gemination.

In this study, we aim to better understand the phonetic and phonological properties of non-native gemination, making abstraction of its roots in spelling. On the one hand, we aim to gather further evidence that gemination can have a contrastive role in the mind of Italian$_{L1}$ learners by demonstrating that French (non-)words are judged as different words if pronounced with a short vs long consonant. On the other hand, we wish to examine more closely the interplay of phonotactics on this phenomenon, by establishing whether such judgments differ in contexts where geminate consonants do vs do not comply with Italian phonotactic restrictions.

2. DATA AND METHODS

2.1. Stimuli

In view of developing an auditory discrimination test (cf. 2.2) for Italian$_{L1}$ learners of French$_{L2}$, we devised a set of 60 disyllabic stimuli, all of which were French non-words in French (as well as in Italian), corresponding to 3 different sets (20 stimuli x 3 sets), as shown in Table 1.

1. Stimuli in the first set were C_1V_1ˈ$C_2V_2C_3$ non-words, where C_2 is a geminable consonant in a ***legal context*** (intervocalic) by Italian phonotactic rules, and V_1 and V_2 are French vowels that have a direct counterpart in Italian (***native***) (e.g., [vaˈpiʁ]).
2. Stimuli in the second set were C_1V_1ˈC_2V_2 non-words, where C_2 is a geminable consonant in a ***legal context*** (intervocalic) by Italian phonotactic rules, and V_1 and V_2 are French vowels that do not have a direct equivalent in Italian (***foreign***) (e.g., [vyˈpɔ̃]).
3. Stimuli in the third set were $C_1V_1C_2$ˈC_3V_2 non-words, where C_3 is a geminable consonant in an ***illegal context*** (preceded by another consonant) by Italian phonotactic rules (e.g., [vɔʁˈpe]).

	[p]	[t]	[l]	[m]
legal	[vaˈpir]	[vaˈtir]	[daˈlir]	[baˈmir]
	[baˈpir]	[naˈtir]	[naˈlir]	[kaˈmir]
	[laˈpir]	[laˈtir]	[gaˈlir]	[gaˈmir]
	[ʃaˈpir]	[zaˈtir]	[maˈlir]	[taˈmir]
	[faˈpir]	[faˈtir]	[faˈlir]	[faˈmir]
foreign	[vyˈpɔ̃]	[vyˈtɔ̃]	[dyˈlɔ̃]	[byˈmɔ̃]
	[byˈpɔ̃]	[nyˈtɔ̃]	[nyˈlɔ̃]	[kyˈmɔ̃]
	[gyˈpɔ̃]	[gyˈtɔ̃]	[gyˈlɔ̃]	[gyˈmɔ̃]
	[ʃyˈpɔ̃]	[zyˈtɔ̃]	[zyˈlɔ̃]	[tyˈmɔ̃]
	[fyˈpɔ̃]	[fyˈtɔ̃]	[fyˈlɔ̃]	[ʒyˈmɔ̃]
illegal	[vɔrˈpe]	[vɔrˈte]	[vɔrˈle]	[bɔrˈme]
	[bɔrˈpe]	[bɔrˈte]	[nɔrˈle]	[kɔrˈme]
	[lɔrˈpe]	[lɔrˈte]	[gɔrˈle]	[gɔrˈme]
	[ʃɔrˈpe]	[ʃɔrˈte]	[mɔrˈle]	[tɔrˈme]
	[nɔrˈpe]	[nɔrˈte]	[fɔrˈle]	[zɔrˈme]

Table 1: The 60 experimental stimuli. For each stimulus, the target consonant is underlined.

The stimuli were recorded by a native speaker of French (a phonetician), in a sound-proof booth. Subsequently, every non-word was manipulated in *Praat* [14] by artificially lengthening the duration of the target consonant by steps of +30%, thereby obtaining 5 different stimuli covering the full range from singletons to geminate consonants: +0% (original recording), +30%, +60%, +90%, +120%. No artefacts of the manipulation were audible.

Additionally, 60 French non-words were recorded to be used as distractors. Each distracting non-word differed from a corresponding experimental non-word by one phoneme; for example, the distractor [vaˈmiʁ] was recorded to match the experimental item [baˈmiʁ].

2.2. AX auditory discrimination test

The stimuli described above were used for an auditory discrimination test with 120 experimental trials (and 4 training trials) in *PsychoPy2* [15]. The format of the test is illustrated in figure 1 and ran as follows: after a short fixation (0.4 sec.), participants heard a first stimulus (A), then a second one (X), separated by an interval of 1.2 seconds in order to trigger phonological rather than acoustic processing (cf. [16, 17]). Participants were told they would hear pairs of extremely rare French words that would probably be unknown to them; their task was to choose whether the two words were in fact the same, or different words, by clicking 's' or 'l' on the keyboard. They were exposed to audio only (the screen was blank) while listening; once the second stimulus had been played, a one-liner appeared on the screen reminding participants which keys should be pressed. Once they had provided a response, they were immediately taken to the following trial. They could not change their response, nor go back, nor re-play stimuli. *PsychoPy2* recorded participants' responses and response times.

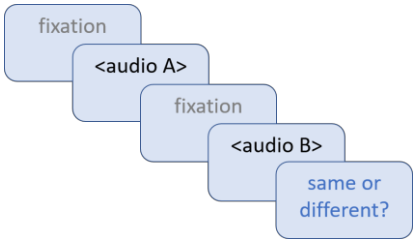


Figure 1. Auditory discrimination test.

60 trials tested experimental stimuli: 12 trials tested stimuli with the target consonant duration at +0% (our baseline, correct response = 'same'), 12 trials tested items with +30% duration, 12 trials tested items with +60% duration, 12 trials tested items with +90% duration, 12 trials tested items with +120% duration. We used a Latin square design with 5 presentation lists, so that participants heard each non-word only once (i.e., in only one of the 5 experimental conditions). The other 60 trials were control trials: participants heard an experimental item (A) and a distractor (X), so that the expected response was always 'different'. Trials testing experimental items and those testing distractors were randomised.

2.3. Participants

We recruited 24 Italian$_{L1}$ learners of French$_{L2}$ (henceforth IT) and 24 French$_{L1}$ native speakers (henceforth FR). IT participants were students at the Faculty of Foreign Languages in Turin. 15 of them were born in the local area and had grown there, while 9 were born in other parts of Italy and had moved to Turin for their studies. Despite claims of regional variation for geminate consonants across Italy, recent large-scale studies have revealed that, due to the progressive standardisation of the language, speakers (and especially younger generations) do not show relevant regional differences [2, 18] (except for sandhi gemination, a.k.a. *raddoppiamento fonosintattico*, not relevant for our study). Among IT participants, 20 identified as women, 4 as men (average age: 25.1, *SD* = 3.7), reflecting the gender imbalance found among students of Languages. The average age of first contact with French was 12 (range: 6 - 21). The self-declared level of French$_{L2}$ ranged from B1 to C1; 6 participants had been in Erasmus programmes in France, and 13 others had been at least once to a French-speaking country (median: 2 weeks, range: 1 week - 9 months). A larger number of participants claimed to regularly read and listen to French (*n* = 17 and 16, respectively), than write and speak (*n* = 11 and 9). Among FR participants, 16 identified as women, 8 as men (average age = 23.2, *SD* = 2.6). They were students at the Faculty of Linguistics at the University of Paris 8 and lived in the Paris area at the time of recording.

2.4 Procedure

Participants were asked to take the test in the university premises, either in a sound-proof booth, or in a silent room, depending on availability. They sat in front of a Mac with a AKG HSC 271 headset and ran the test on *PsychoPy2*. A training session of 4 trials preceded the real test, which was conducted without interruption and lasted approximately 10 minutes. Participants were instructed to take the test as spontaneously as possible, providing immediate and non-pondered responses. The test was taken within a larger data collection project, so participants also performed other production and perception tasks for a total of approximately 75 minutes.

3. RESULTS

3.1 Analysis of responses

In total, we obtained 2880 responses for experimental stimuli (60 x 48 participants). The data were analysed on R, building generalised linear mixed-effects models with *lme4* [19], *p* values were obtained with *lmerTest* [20]. We built a binomial model to predict participants' responses ('same' or 'different') on the basis of consonant duration (+0% to 120%), context (legal, foreign, illegal) and L1 (Italian, French). We included participant and consonant (/p/, /t/, /l/, /m/) as random effects, with random intercept and random slopes: *Response ~ ConsonantDuration * L1 * Context + (ConsonantDuration + Context | Participant) + (Step + L1 | Consonant)*. In order to deal with convergence issues, we set the *bobyqa* optimizer to run up to 100.000 function evaluations. It has to be noted that, due to the continuous nature of acoustic durations, consonant duration was coded as numeric, ranging from 1 to 2.2, despite the fact that we tested only 5 steps on the continuum (1 = +0%, 1.3 = +30%, 1.6 = +60%, 1.9 = +90%, 2.2 = +120%). To be on the safe side, we also ran the analysis with step as a 5-level factor instead of duration, and obtained comparable results. Model predictions extracted via *ggeffects* [21] are shown in figure 2.

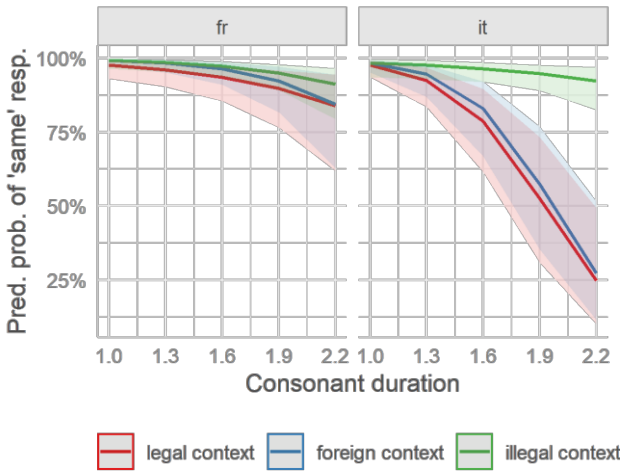


Figure 2: Predicted probabilities for responses as 'same' by L1 (fr, it) and consonant duration (ranging from 1 = +0%, to 2.2 = +120%).

The plots clearly show differences across groups: responses by FR participants do not seem to be heavily affected by the target consonant duration, while responses by IT participants are heavily affected by duration in contexts where gemination is phonotactically legal. The summary of our model revealed a significant effect of consonant duration on responses (p = .005), reflecting the overall decrease of 'same' responses for higher durations. Additionally, we observed a significant two-way interaction of consonant duration x L1 (p = .003), and a significant three-way interaction of consonant duration x L1 x context. We ran post-hoc pairwise comparisons with Holm correction via *emmeans* ([22]) at each of the 5 steps, confirming what had been inferred from the plots: responses by FR

participants for the legal, foreign and illegal contexts do not significantly diverge from each other at any of the 5 steps (all adj. p values > .273). Instead, responses by IT participants significantly diverge for the illegal versus legal and foreign contexts at the 1.6 step and higher (adj. p values < .001 for the 1.6, 1.9 and 2.2 steps), while responses for the legal and foreign contexts do not significantly diverge from each other (adj. p = 1 at all steps). Additionally, responses given by IT participants in the illegal context do not significantly diverge from those given by FR participants (adj. p = 1 at all steps).

3.2. Analysis of response times (RTs)

The analysis of normalised (log-transformed) RTs was performed in a similar way. After eliminating outliers (RTs > 2 SDs of the mean, n = 103), we built a linear mixed-effects model to predict RTs: *log(RT) ~ ConsonantDuration * L1 * Context + (ConsonantDuration + Context | Participant) + (Step + L1 | Consonant)*. Model predictions are shown in figure 3.

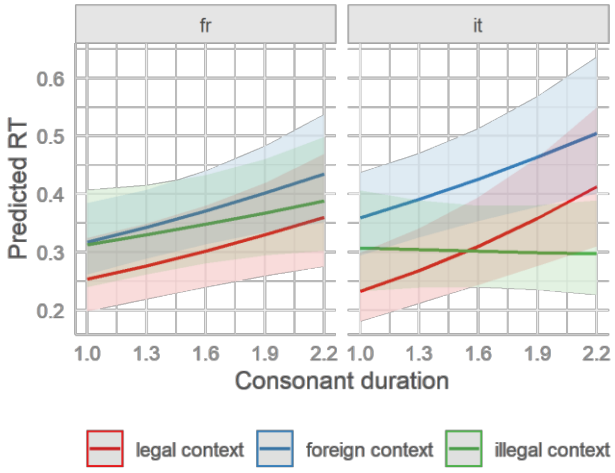


Figure 3: Predicted response time by L1 (fr, it) and consonant duration (ranging from 1 = +0%, to 2.2 = +120%).

The plots illustrate a general trend of RTs to increase with consonant duration for both groups, except for IT when gemination is illegal. The model summary confirms that the effect of consonant duration is significant (p = .002), and suggests the presence of a mild three-way interaction of consonant duration x L1 x Context (p = .054). Similar to above, we ran post-hoc pairwise comparisons with Holm correction at each step: RTs by FR participants for the legal, foreign and illegal contexts do not significantly diverge from each other at any step (all adj. p values > .087). Instead, RTs by IT participants significantly diverge for the foreign vs legal context (all adj. p values < .042, except at the last step), probably reflecting the higher cognitive demand in processing non-native sounds. RTs by IT participants in the illegal context are shorter than in the foreign context (adj. p values < .001 at steps 1.6 and higher).

4. CONCLUSIONS

The results of the auditory discrimination test revealed that IT participants respond to manipulated consonant durations in French$_{L2}$ differently from FR participants: when the target consonant duration is manipulated to sound like a geminate consonant (+60% and higher), IT participants tend to judge it to be a different word than the same stimulus with non-manipulated duration. We do not know if this effect is simply phonetic (e.g., longer durations directly activate gemination even though it does not exist natively in the L2), or if it is mediated by a reconstructed orthographic form (i.e., when listening to such stimuli, participants imagine it as spelled with a double consonant and hence judge it to be a different word). At any rate, it seems reasonable to conclude that gemination affects Italian$_{L1}$ leaners' phonological awareness of French$_{L2}$: for them, a stimulus with a long consonant is a different word than the same stimulus with a short consonant: clearly, gemination can have a contrastive role in their interlanguage (cf. also [4, 8, 9]).

The main result of this study is that non-native gemination seems to be mediated by phonotactic constraints: when the target consonant is preceded by another consonant (and therefore gemination is not possible in Italian), IT participants are not affected by consonant duration and respond similarly to FR participants; even their RTs are not affected by consonant duration in this condition. We think that various explanations are plausible. On the one hand, learners may simply transfer L1 phonotactic rules to the L2, and consider gemination as impossible in such contexts. Alternatively, their responses may be due to a phonotactically-conditioned length-deafness; in other words, Italian$_{L1}$ speakers may be deaf to consonant lengthening in contexts where gemination cannot exist in their L1. Such length-deafness may be caused by phonotactic rules themselves (i.e., L1 phonotactic restrictions make speakers deaf to consonant lengthening in other contexts), or driven by markedness (gemination being typologically marked in these contexts). While our present data do not allow us to definitively settle the exact cause, the fact that IT participants' RTs are not affected by consonant durations in the illegal context seems to suggest that these speakers do not notice variation in duration, thereby potentially hinting at a phonotactically-conditioned length-deafness. Future studies will address this issue more specifically.

5. ACKNOWLEDGMENTS

We are grateful to all Italian and French participants who kindly accepted to take the test, and to Rémi Godement-Berline for sitting in a sound-proof booth and recording the stimuli for the experiment. We would also like to thank members of the LFSAG lab in Turin for helping us recruit participants.

6. REFERENCES

[1] Esposito, A., Di Benedetto, M. G. 1999. Acoustical and perceptual study of gemination in Italian stops. *Journal of the Acoustical Society of America*, 106(4), 2051-2062.

[2] Mairano, P., De Iacovo, V. 2020. Gemination in northern versus central and southern varieties of Italian: A corpus-based investigation. *Language and Speech*, 63(3), 608-634.

[3] Ridouane, R. 2007. Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association*, 37(2), 119-142.

[4] Bassetti, B., Sokolović-Perović, M., Mairano, P., Cerni, T. 2018. Orthography-induced length contrasts in the second language phonological systems of L2 speakers of English: Evidence from minimal pairs. *Language and Speech*, 61(4), 577-597.

[5] Sokolović-Perović, M., Bassetti, B., Dillon, S. 2020. English orthographic forms affect L2 English speech production in native users of a non-alphabetic writing system. *Bilingualism: Language and Cognition*, 23(3), 591-601.

[6] Mairano, P., Santiago, F., Delais-Roussarie, E. 2018. Gémination non-native en français d'apprenants italophones. *Actes des JEP 2018*, 657-665.

[7] Nawafleh, A. 2022. L'interférence de la gémination dans la prononciation des étudiants arabophones apprenant le français. *Electronic Journal of Foreign Language Teaching*, 19(1), 85-98.

[8] Bassetti, B., Mairano, P., Masterson, J., Cerni, T. 2020. Effects of Orthographic Forms on Second Language Speech Production and Phonological Awareness, With Consideration of Speaker-Level Predictors. *Language Learning*, *70*(4), 1218-1256.

[9] Bassetti, B., Masterson, J., Cerni, T., Mairano, P. 2021. Orthographic forms affect speech perception in a second language: Consonant and vowel length in L2 English. *Journal of Experimental Psychology: Human Perception and Performance*, 47(12), 1583-1603.

[10] Mitterer, H. 2021. The role of orthography in learning a second language: Evidence from Maltese English. *Xjenza Online*, 9, 162-172.

[11] Shibuya, Y., Erickson, D. 2010. Consonant cluster production in Japanese learners of English. *Proc. of Second Language Studies: Acquisition, Learning, Education and Technology*, 1-4.

[12] Radant, H. L. H. J., Huang, H. L. 2009. Chinese phonotactic patterns and the pronunciation difficulties of Mandarin-Speaking EFL learners. *The Asian EFL Journal*, 11(4), 115.

[13] Weber, A., Cutler, A. 2006. First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America*, 119(1), 597-607.

[14] Boersma, P. & Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.2, retrieved 15 Nov 2021 from http://www.praat.org/

[15] Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51, 195-203.

[16] Werker, J. F., Logan, J. S. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37(1), 35-44.

[17] Flege, J. E., MacKay, I. R. 2004. Perceiving vowels in a second language. *Studies in second language acquisition*, 26(1), 1-34.

[18] Giordano, R., Savy, R. 2012. Sulla standardizzazione del consonantismo dell'italiano: consonanti geminate, rafforzate e fricative alveolari in contesto intervocalico. *Atti del Congresso SILFI* (vol. 2), 431–445.

[19] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

[20] Kuznetsova, A, Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82 (13), 1-26.

[21] Lüdecke D. 2018. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software*, 3 (26), 772.

[22] Lenth R. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.3.

PAPER TEMPLATE FOR ICPhS 2023 PRAGUE

XXX

XXX
XXX

ABSTRACT

This French-Czech contrastive study deals with the duration of French vowels in the production of French speakers and Czech learners of French. The considerable differences between the vowel duration in the Czech and French language systems led to four research hypotheses. Based on recordings of the text being read, the analysis examines not only the influence of primary and secondary accent (final accent, initial accent and *accent d'insistance*) on the duration of French vowels, but also the type of syllable in which these vowels are found. The results confirm that the phonetic-phonological habits that Czech speakers apply in their mother tongue have a considerable influence on the production of the language being acquired, here French.

Keywords: vowel duration, final accent, initial accent, closed syllable, open syllable

1. INTRODUCTION

Czech and French are phonetically and phonologically very different languages, both from the phonemic and the prosodic points of view. One of the many differences is the vowel duration.

In Czech, vowel duration has a phonological function, in that it can distinguish the meaning of words. This phonological opposition of vowel duration, which does not depend on the accent or the position of the vowel within the word, is one of the linguistic features of Czech language [1].

In French, the parameter of vowel duration is an important part of the accent. In French, the vowel duration has a phonotactic function [2], as it indicates the appropriate rhythm of a sentence and thus the utterance and its comprehension. It is not strictly connected to the lexical unit because the word is not an independent sound unit in French. In standard French it no longer has a distinctive validity. The number of French words where vowel duration could have a distinctive function is very small. The distinctions between minimal pairs (e.g., *l'être x lettre, maître x mettre, pâte x patte*) are blurred in Standard French; they are maintained only in some regional varieties of French and are highly unstable. The very few cases of minimal pair oppositions that still exist in Standard French today are realized by a difference in timbre, and these may be complemented by a difference in vowel duration (e.g., *cote/côte, pomme/paume*) [3].

Authors whose research focuses on French vowel duration agree that long vowels occur in accented positions. Some authors approach the issue not only from the perspective of accented/unaccented position, but also take into account the type of syllable [4, 5, 6, 7, 8, 9, 10, 11].

Note that different functions of vowel duration in the two language systems mentioned above (L1) can lead to interference in the production of the acquired language (L2).

2. OBJECTIVES AND HYPOTHESES

The aim of this study is to research the influence of primary (final) accent and secondary accent (initial accent and *accent d'insistance*) on the duration of French vowels in text being read by French and Czech speakers. The phonetic-phonological differences between Czech and French, mentioned in the introduction, led to the formulation of the following hypotheses:

1) Czech learners of French produce more accented syllables than French speakers.

2) Czech learners of French produce more syllables with an initial accent than French speakers.

3) Initial accent and *accent d'insistance* have a quantitative effect on the duration of French vowels for Czech learners of French, but not for French speakers.

4) The difference between vowel duration in the final accent and those in unaccented positions is more significant for French speakers than for Czech learners of French.

3. CORPUS AND RESEARCH PARTICIPANTS

For the purpose of this experiment, we searched for a text that would reflect the prosodic features of French, and which would not contain too many words unknown to Czech speakers. For these reasons, we have chosen the beginning of the fairy tale *The Little Red Riding Hood* - namely its narrative part, without dialogue lines. Ten French speakers and ten Czech learners of French were recorded. The French speakers are university students originally from Paris,

Brittany and Burgundy, living in Paris for at least 6 years. In phonetic-phonological terms, they do not show any significant differences to Standard French. In the case of the Czech learners of French, they are first-year students of the Faculty of Education at Charles University in Prague. Their language proficiency in French according to the Common European Framework of Reference is B1. At the time of the recordings, none of the respondents were aware of the purpose of the experiment.

4. METHODOLOGY AND MEASURING

Vowel duration measurements were performed manually using WinPitch software. The marking of accented syllables was based on consensus between two independent speakers. Changes in speech rate (rate coefficient k) for each speaker were normalized using the following equation:

$$k = \frac{D_m}{d_i} \quad (1)$$

Dm = the average duration of the sentence in which the vowel is found,
di = the individual duration of a given sentence for a given speaker.

The speed coefficients of the individual speakers did not show much variation; thus, the speed of the speakers while reading the excerpts is rather constant. To increase the relevance of the normalization of the duration of the measured vowels, the coefficient was calculated separately for each sentence of the corpus.

To better compare the differences and similarities in the production of French speakers and Czech learners of French, all the syllables in the corpus have been labelled using the following abbreviations. These will be used in the following presentation of the analyses results:

N/o – a vowel in an open syllable, in an unaccented position;

N/f – a vowel in a closed syllable, in an unaccented position;

AF/f – a vowel in a closed syllable, with the final accent;

AF/o – a vowel in an open syllable, with the final accent;

AI/f – a vowel in a closed syllable, with the initial accent or *accent d'insistance*;

AI/o – a vowel in an open syllable, with the initial accent or *accent d'insistance*.

5. RESULTS

The graph below illustrates the differences in the French vowel duration for French speakers and Czech learners of French.

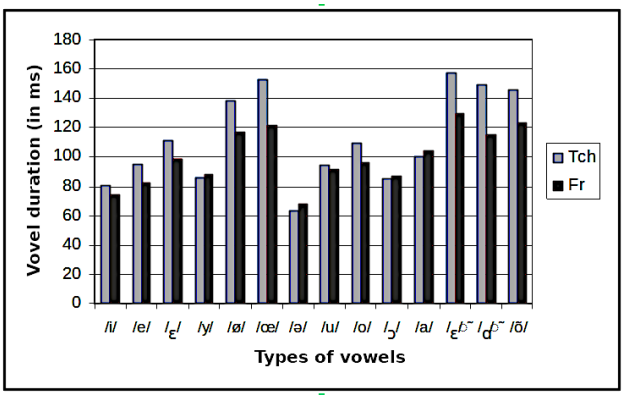


Table 1: French vowel duration in the production of French speakers and Czech learners of French

Note that the biggest difference in the vowel duration between the French speakers and Czech learners of French is in the product ion of such vowels which do not occur in the phonemic system of Czech language. These are primarily front rounded vowels /ø œ/ and nasal vowels /ɛ̃ ɑ̃ õ/.

As far as nasality is concerned, the influence of this articulatory feature on the corresponding oral vowels in closed and open syllables in accented and unaccented positions has already been analysed by Delattre and Monnot [12]. The authors found that differences in vowel duration caused by nasality can vary by up to 42%. The corpus for this experiment contains vowels in different accented positions and different syllable types, we can thus subject our measurements to the same analysis and observe differences and similarities in both groups of speakers.

As for the vowels /ɛ/ and /ɛ̃/, the measured values show that for French speakers the differences in duration between the oral and nasal vowels in open syllables in the unstressed position are 28%, and 16% in open syllables carrying the final accent. For Czech learners of French, these differences in the same positions and the same syllable types are even larger: 93% and 31%, respectively.

As for the differences between the vowel duration of /a/ and /ɑ̃/, the biggest differences for French speakers are observed in the unaccented position in closed syllables (62%), with the final accent in open syllables (48%) and with the secondary accent in open syllables (48%). In the case of Czech learners of French, the biggest difference in the duration of these two vowels was found in open syllables with the secondary accent (78%). In case of final accent, we observe an opposite effect for Czech learners than for

French speakers: oral vowels are longer than the corresponding nasal vowels, both in closed and open syllables.

As for the vowels /o/ and /õ/ in the production of French speakers, we observe the biggest differences in their duration in the unaccented position in open (48%) and closed (48%) syllables. The same is true for Czech learners of French, for whom these differences in duration in the same positions come to 58% and 26% respectively.

The comparison of the measured values shows that in case of French speakers, the differences in duration between the corresponding oral and nasal vowels can reach up to 62%. The values are even higher for Czech learners: up to 93%. Our data therefore reveal even greater differences than those observed by Delattre and Monnot [12].

We now turn our attention to the differences in the number of syllables carrying final and secondary accents for the two groups of speakers. It has been observed that Czech learner of French produce on average 63 accented syllables (out of 221 total), 38 of which are with the final accent and 25 with the secondary accent (initial or *accent d'insistance*). French speakers produce on average 51 accented syllables (out of 218 total), 41 of which are affected by the final accent and 10 by the secondary accent. Czech speakers learning French seem to produce more accented syllables than native speakers. The number of secondary accents is also higher for Czech speakers (39.6%) than for native speakers (19.6%). This phenomenon is very likely due to the influence of accenting in the native language, in this case Czech, on the acquired foreign language (French).

The following graph shows that the secondary accent in the production of Czech learners has also a quantitative effect on the duration of French vowels (vowels affected by this accent are on average 15.8% longer than vowels in unaccented positions). For French speakers, the opposite phenomenon can be observed: vowels carrying the secondary accent are even shorter than vowels in unstressed positions (on average 15% shorter). The secondary accent is therefore a quantitative factor for the group of Czech learners of French, but not for French speakers.

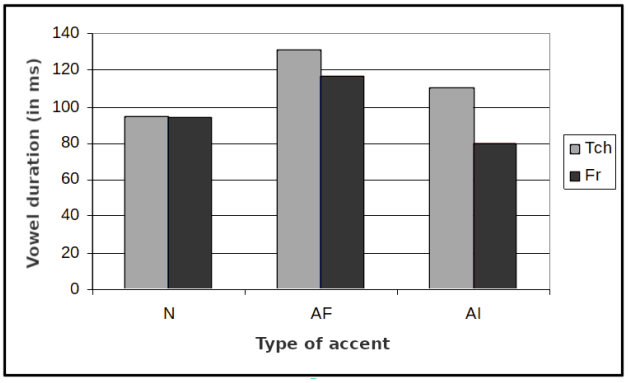


Table 2: Vowel duration in the unaccented position, with final and secondary accents in the production of French speakers and Czech learners of French

The graph also shows that for Czech learners, vowels with final accent are on average 38% longer than vowels in unstressed position. For native speakers, vowels with final accent are on average only 24% longer than vowels in unstressed positions. The results have therefore not supported the fourth hypothesis, where we expected the duration of vowels carrying final accent to be longer for French speakers than for Czech speakers.

Looking at the measured values for the oral and nasal vowels separately, at the type of syllable in which they occur, and also at the presence (or absence) of an accent and its type, we get much more accurate results. The following graph shows that for oral vowels the results for the two groups of speakers do not differ significantly. Here again, we observe a significant lengthening of the vowels carrying the final accent for both groups of speakers (for Czech learners of French again more significant than for French speakers) and a lengthening effect of the secondary accent in the production of Czech learners.

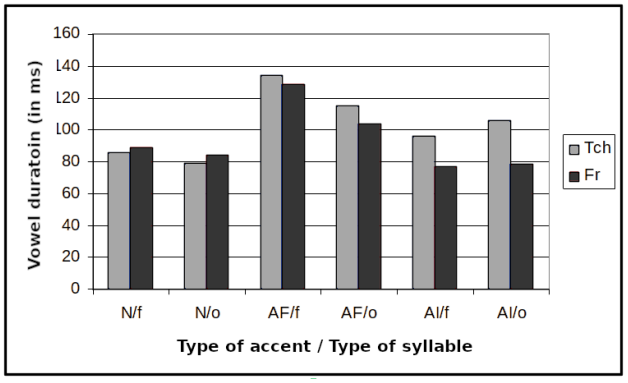


Table 3: The duration of French oral vowels in all possible positions and for all syllable types in the production of French speakers and Czech learners of French

As for nasal vowels (see the graph below), we can observe an absence of influence of final and

secondary accents on their duration for French speakers. For Czech speakers, however, it shows that the most lengthened vowels are those in closed syllables with the secondary accent and those with the final accent.

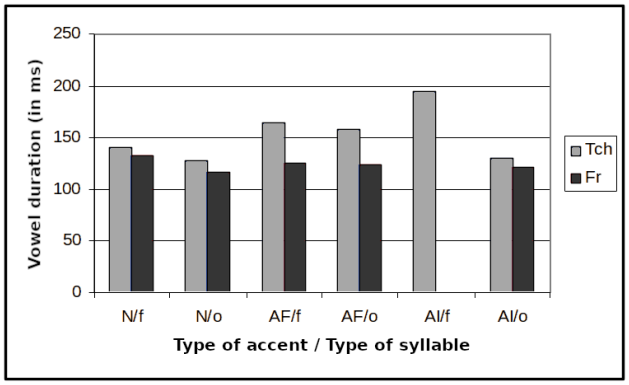


Table 4: Duration of French nasal vowels in all possible positions and syllable types in the production of French speakers and Czech learners of French

As for the effect of open and closed syllables on the duration of French vowels, in the case of oral vowels we observe a significant lengthening in closed syllables carrying the final accent for both groups of speakers, as well as in open syllables carrying the secondary accent for the Czech speakers (see Table 3).

As for nasal vowels, we do not observe any vowel lengthening that would be caused by syllable or accent type in case of French speakers. However, in the production of Czech learners of French we can find lengthening of nasal vowels carrying final accent and also longer duration of French nasal vowels in closed syllables with secondary accent. Unfortunately, we cannot compare these results with those of French speakers, because our corpus does not contain any occurrence of this syllable type or accent type in their production.

6. DISCUSSION AND CONCLUSION

This contrastive study reveals the existence of differences and similarities in the production of French vowel duration by French speakers and Czech learners of French.

The analysis of the results shows that the most problematic for Czech learners in terms of vowel duration production are the front rounded vowels /ø œ/ and nasal vowels /ɛ̃ ɑ̃ ɔ̃/.

The effect of syllable type on vowel duration was also studied. The data analysis shows that the factor does not play a significant role for Czech learners of French. For French speakers, however, it is the closed syllables in which the duration of vowels is longer.

Furthermore, we followed the study of Delattre and Monnot [12] who investigated the effect of nasality on vowel duration in closed and open syllables and in accented and unaccented positions. As for the French speakers, the most significant differences between oral vowels and their nasal counterparts were found in the case of /a/ and /ɑ̃/: 62% in closed syllables in the unaccented position, in open syllables with secondary accent (48%), and in open syllables with final accent (47%). Almost the same values, 48% and 60%, were measured for the vowels /o/ and /ɔ̃/ in the unaccented position in open (48%) and closed (48%) syllables, respectively. For Czech learners of French, these differences reach even higher values: 78% for /a/ and /ɑ̃/ in open syllables with an adjacent accent, and 93% for /ɛ/ and /ɛ̃/ in open syllables in unstressed positions. In general, these vowels are the longest when carrying final accent for both groups of speakers; however, for the Czech learners of French the longest vowels are found also with the secondary accent.

Based on the data analysis, we can confirm or disconfirm the research hypotheses:

Hypothesis 1 - Czech learners of French produce a greater number of accented syllables than French speakers - has been confirmed. The data show that the occurrence of accented syllables is higher for Czech speakers. In general, it can be pointed out that Czech speakers produce shorter rhythmic groups (i.e., a higher number of rhythmic groups) in French than native speakers, which can probably be explained by their intermediate language proficiency.

Hypothesis 2 - Czech learners of French stress initial syllables more than native speakers - has also been confirmed. The data show that the occurrence of secondary accents is higher for Czech speakers (39.6%) than for native speakers (19.6%). The influence of the native language of Czech learners on L2 production is therefore evident.

Hypothesis 3 - the initial accent and the *accent d'insistance* have a quantitative effect on the realization of vowel duration for Czech learners of French, but not for French speakers - has also been confirmed. The lengthening of vowels in syllables carrying a secondary accent is produced by Czech learners of French, but not by French speakers. The opposite tendency seems to be valid here.

Hypothesis 4 - the difference between the duration of vowels carrying final accent and those in unstressed positions is greater for French speakers than for Czech learners of French – has not been confirmed. The data show that both groups of speakers produce longer vowels in positions with final accent, with Czech speakers producing even longer vowels than French speakers.

6. REFERENCES

[1] Dohalská, M. 2006. Trvání českých samohlásek – funkce a forma. In: Palková, Z., Janoušková, J. (eds), *Kapitoly z fonetiky a fonologie slovanských jazyků: příspěvky z pracovního vědeckého setkání na XVI, zasedání Komise pro fonetiku a fonologii slovanských jazyků při Mezinárodním komitétu slavistů*. Praha.

[2] Duez, D. 1987. Hiérarchisation des paramètres acoustiques et identification des frontières. *Actes des XVIièmes Journées d'Etudes sur la Parole*, 221-223.

[3] Astésano, C. 2001. *Rythme et Accentuation en Français: Invariance et Variabilité Stylistique*. L'Harmattan.

[4] Nyrop, K. 1963. *Manuel phonétique du français parlé.* Alf Lombard Publisher.

[5] Delattre, P. 1959. Rapports entre la durée vocalique, le timbre et la structure syllabique en français. *The French Review*, vol. 32, no. 6, 547-552.

[6] Delattre, P. 1966. Durée vocalique et consonnes subséquentes. In: Delattre, P. (eds), *Studies in French & comparative phonetics* London. Mouton & Co.

[7] Carton, F. 1974. *Introduction à la phonétique du français*. Bordas.

[8] Zichová, M. 1967. Contribution à l'étude de la durée des voyelles françaises. *Phonetica Pragensia*, *Philologica*, vol. 6, 113-124.

[9] Fouché, P. 1969. *Traité de prononciation française*. Klincksieck.

[10] Wioland, F. 1991. *Prononcer les mots du français : des sons et des rythmes*. Hachette.

[11] Léon, P. 2009. Nouveau regard sur la phonostylistique. *La Linguistique*, vol. 45, fasc.1., 159-170.

[12] Delattre, P., Monnot, M. 1968. The role of duration in the identification of French nasal vowels. *International Review of Applied Linguistic in Language Teaching*, vol. 6, no. 3, 267-288.

Individual differences in VOT realizations of English plosives by Palestinian pupils

Nour Tamim, Silke Hamann, Paul Boersma

University of Amsterdam
{n.j.s.tamim, silke.hamann, paul.boersma}@uva.nl

ABSTRACT

Palestinian Arabic (PA) learners of English meet with two types of challenges when learning English plosives in initial position. The first challenge is that English initial /p, t, k/ are aspirated while PA /t, k/ are not, and that English /b, d, g/ have short-lag VOT while PA /b, d/ are fully voiced. The second challenge is that PA lacks /p/ and (to some extent) /g/. This paper focuses on individual differences between learners in how they cope with the two types of challenges. As could be expected, some learners have fully voiced /b, d, g/ in their English, and/or unaspirated /p, t, k/. As could also be expected, some learners pronounce /p/ and /b/ in the same way. We found no evidence of an implicational hierarchy between these challenges, e.g. whether learners who conflate /p/ and /b/ also have fully voiced /b, d, g/ in their English.

Keywords: VOT, second language acquisition, variation, Palestinian Arabic, English.

1. INTRODUCTION

There are several differences between English and Palestinian Arabic (PA) when it comes to the structure and implementation of the plosive voicing contrasts in these two languages.

As far as structure is concerned, English contrasts voiced and voiceless plosives at three places of articulation, namely bilabial (/p, b/), alveolar (/t, d/), and velar (/k, g/). Palestinian Arabic (PA), on the other hand, contrasts denti-alveolar stops (/t, d/) and their pharyngeal counterparts (/$t^{ˤ}$, $d^{ˤ}$/), but at the bilabial and velar places of articulation, there are gaps: while PA has /b/ and /k/, it lacks /p/ and /g/ (though these do appear in loanwords, and /g/ in some of the dialects). Both gaps are quite common cross-linguistically (see [1] for an overview).

English and PA also lie in different corners of the typology of the phonetic realization of voicing contrasts established by Lisker & Abramson [2], as measured on the voice onset time (VOT) continuum. In initial position, PA employs voicing lead for /b, d/ and short lag for /t, k/ [3], while English employs short lag (with occasional prevoicing) for /b, d, g/ and long lag for /p, t, k/ [2]. In other words, PA contrasts prevoiced plosives with unaspirated plosives, whereas English contrasts unaspirated plosives with aspirated plosives.

Differences in structure and implementation of voicing may provide obstacles in second-language learning. Incomplete acquisition of an L2 voicing contrast was observed in learners of English whose L1 utilizes prevoicing to distinguish voiced plosives from voiceless ones. Such learners typically employ their L1 implementation of voicing (namely, voicing lead) when realizing English voiced plosives; this has been found for Saudi Arabic speakers [4], Dutch speakers [5], and Italian speakers [6].

The present study looks at inter-speaker variation in the production of initial English plosive voicing by Palestinian Arabic teenagers who have learned English from a native English teacher at school: how do they handle the difference between the voicing structures of PA and English, how do they handle the difference between the phonetic realization of voicing in PA and English, and how is their handling of structure related to their handling of phonetic realization?

2. VOICING MEASUREMENT

It would be ideal if the voicing continuum were simply one-dimensional, i.e. if the degree of voicing could be determined by measuring a single VOT value. Davidson [7] pointed out that this will not work for plosives that have voicing during their closure period: cases of a true negative VOT, i.e. a closure that may start out as voiceless, then switches to voiced, and stays voiced during the remainder of the closure, turn out to be rare in English. Instead, for instance, *voicing bleed* from the preceding vowel (or other sonorant) turns out to be common. Davidson therefore proposed a detailed classification scheme, which takes into consideration partial voicing in the constriction interval of plosives.

We loosely follow Davidson's classification procedure, with the goal of establishing an ordering of phonetic voicedness (disregarding aspiration), as follows.

For each token of an initial plosive (spoken by a PA learner of English) we label the beginning and end of the closure phase manually in Praat [8]. The start of the closure is determined by the offset of the

preceding vowel, and its end by the start of the release of the plosive. We then divide the closure into three equal parts. For each part we determine the percentage of voicing with the help of Praat's Voice Report (*fraction of locally unvoiced frames*), which uses Boersma's [9] pitch detection algorithm, with the silence threshold set to 0.02 and the voicing threshold set to 0.6 (to avoid that occasional background noise is interpreted as voicing).

We consider a closure part "voiced" if the Voice Report reports 50 or more percent voicing, and "voiceless" otherwise. This yields the eight possible voicing patterns listed in Table 1. To establish an ordering to work with, we classify the eight patterns into three degrees of closure voicing (similarly though not identically to [7]): we classify a closure as *voiced* if more than 50% voicing is detected in all three parts or in the first two parts (the latter are cases where vocal fold vibration ceases in the last third, usually due to lack of transglottal pressure differences [10]); we classify a closure as *voiceless* if less than 50% is detected in all three parts or in the last two parts (in the latter case, the voicing in the first part is considered a short bleeding effect from the preceding vowel); we classify the remaining four cases as *partially voiced*. For a full overview, see Table 1.

first	second	third	class label
V	V	V	voiced
V	V	–	
V	–	V	partially voiced
–	V	V	
–	V	–	
–	–	V	
V	–	–	voiceless
–	–	–	

Table 1: Classification based on percentage voicing in the first, second, and third part of the closure phase; "–" indicates less than 50% voicing in the part, "V" 50% or more voicing.

3. EXPERIMENT

3.1. Participants

Participants were 20 pupils of the 10th grade (16 years of age) at a school in Nablus where English is taught by native English speakers. They had had 10 years of English classes at school.

(The present results are part of a larger study that looks at the influence of input by native versus non-native English teachers on the English production of Palestinian students [11].)

3.2. Material and procedure

18 English words that students are well-familiar with were used as stimuli in a picture naming task. All words had an initial plosive, and were followed by one of three vowels, see the overview in Table 2.

	/b/	/d/	/g/	/p/	/t/	/k/
/aː/	basket	dark	garden	palm	tart	car
/uː/	boots	do	goose	pool	two	cool
/iː/	bee	D	geese	peas	tea	key

Table 2: English test words with initial plosives in three different following vowel contexts.

Speakers were presented with the picture of a word three times, they thus produced three repetitions of each word and a total of 54 stimuli. The stimuli and 26 filler words were presented in randomized order. Each word had to be produced in the carrier sentence *This is a* _, causing an intervocalic context for the plosive.

Instructions were given in English. The recordings took place in a quiet room with an Edirol R-1 solid-state recorder, an Intricon HT-747 super-cardioid electret condenser head-mounted microphone, and a sample frequency of 44.1 kHz.

4. RESULTS

Of the total of 1080 tokens (18 test words × 20 speakers × 3 repetitions), 16 tokens had to be excluded, mainly because of background noise. The distribution of closure voicing classes for the remaining 1064 tokens is given in Table 3. We can see that the class "partially voiced" was rare for all plosives.

Figures 1–3 provide a visual overview of the speaker-specific use of the three voicing classes.

classes	**/b/**		**/d, g/**		**/t, k/**		**/p/**	
voiced	147	83%	253	71%	10	3%	37	21%
partially voiced	8	4%	13	4%	24	7%	9	5%
voiceless	23	13%	90	25%	320	90%	130	74%
total	**178**		**356**		**354**		**176**	

Table 3: Occurrence of different classes of voicing realization during closure phases.

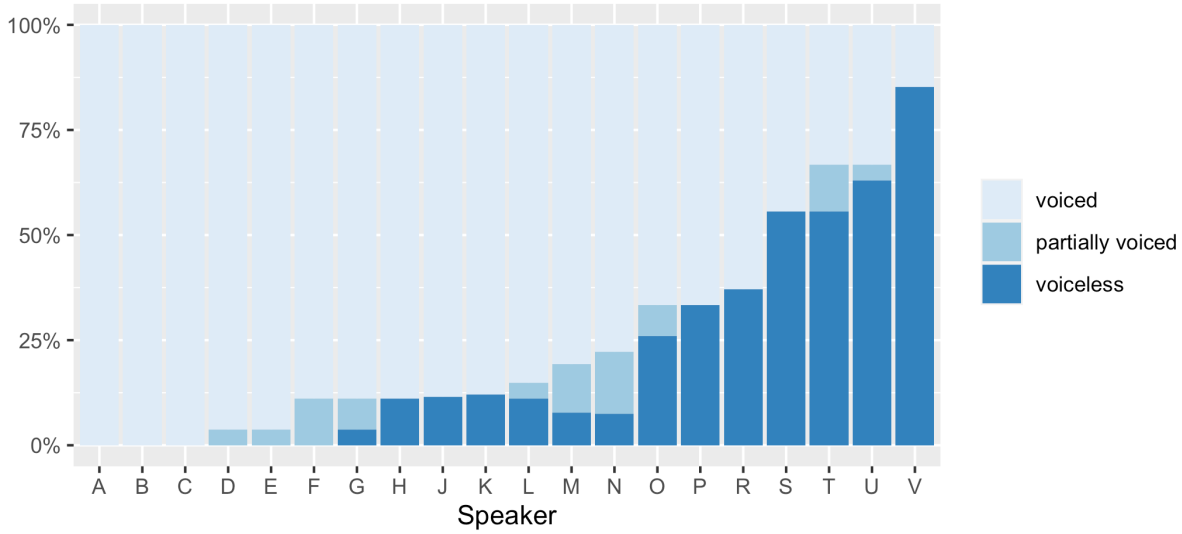


Figure 1: Speaker-specific realizations of the closure phase in English voiced plosives.

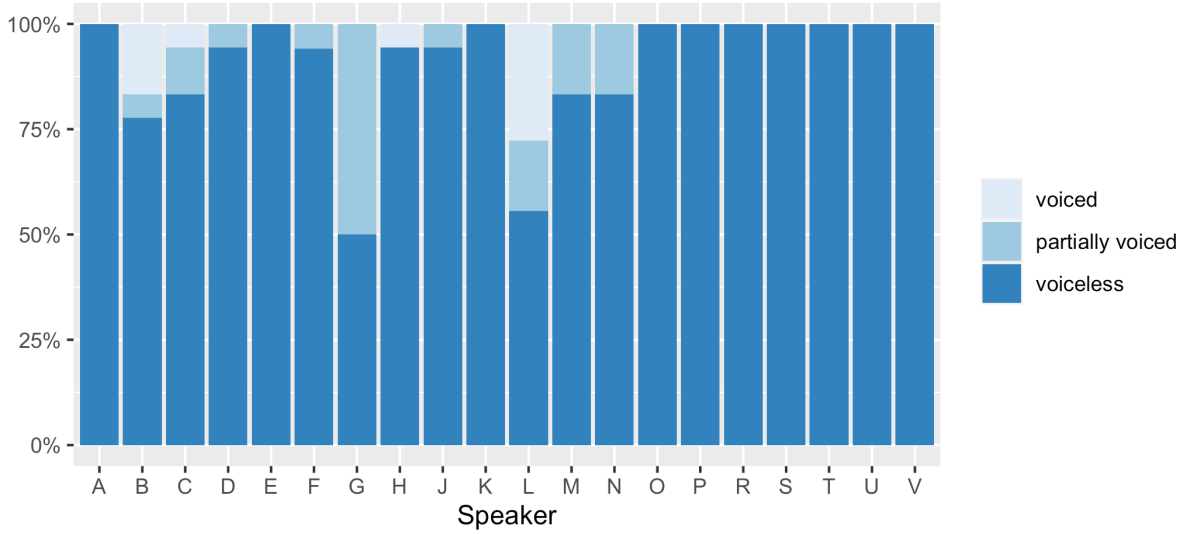


Figure 2: Speaker-specific realizations of the closure phase in English /t/ and /k/.

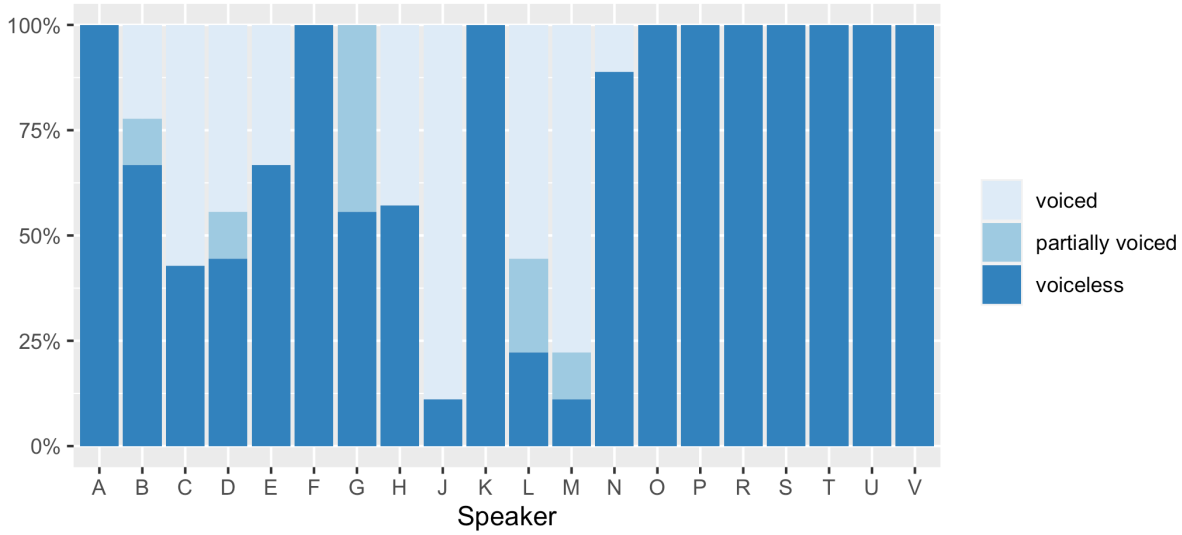


Figure 3: Speaker-specific realizations of the closure phase in English /p/.

For **statistical analysis**, we constructed a mixed-effects model in *R* [12] using the package *lme4* [13], with the *closure voicing* class as an ordered dependent variable that we assumed linear: we coded *voiceless* as −0.5, *partially voiced* as 0, and *voiced* as +0.5. The model contained two nominal within-speaker fixed effects: the target (English) phonological features [*place*] (of articulation) and [*voice*]. For [*place*], the research question about the lack of a phonetic voicing distinction between /b/ and /p/ involves a ternary contrast between *labial* (coded as +2/3) on the one hand, and *coronal* and *dorsal* (both coded as −1/3) on the other hand, supplemented with a binary contrast between *coronal* (coded as +0.5) and *dorsal* (coded as −0.5). For [*voice*], there is only a binary contrast between [–*voi*] (coded as −0.5) and [+*voi*] (coded as +0.5). All contrasts are thus orthogonal and sum to zero, and they differ by 1.0 so that the model's estimates will be interpretable as effect sizes. Computing the model involved fitting the following parameters: four fixed effects (intercept, [*place*], [*voice*], and the interaction of [*place*] and [*voice*]); per speaker a random intercept as well as random slopes for [*place*], [*voice*], and [*place*] × [*voice*]; and per word a random intercept.

As expected (and included here solely as a sanity check), the model reported an estimated **[*voice*] effect** of +0.65, i.e. the difference between [+*voi*] and [–*voi*] consonants amounts to 65 percent of the closure voicing scale. Since $t = 10.155$, we conclude that (for 16-year-old PA learners of English with native-English teachers) the [+*voi*] consonants /b, d, g/ tend to have more closure voicing (as a group) than the [–*voi*] consonants /p, t, k/ (Satterthwaite's approximation yields $df = 27.25$, $p = 9.2 \cdot 10^{-11}$, and 95% CI = +0.52 .. +0.78). This means that the visual distinction between Figure 1 on the one hand and Figures 2 and 3 on the other hand reflects a true population effect, and that this effect is large (at least 52 percent of the closure voicing scale).

As also expected, /p, b/ taken together have more closure voicing than /t, d, k, g/ taken together. The interval estimate for the **ternary [*place*] effect** runs from +0.03 to +0.27, i.e. we can conclude that labials on average have between 3 and 27 percent points more closure voicing than non-labials (point estimate +0.15, $t = 2.660$, $df = 20.52$, $p = 0.015$). This means that the visual distinction between Figures 2 and 3 (which according to Fig. 1 cannot exist to this extent for the [+*voi*] plosives) reflects a true population effect, and this effect can be small or medium (between 3% and 27% of the closure voicing scale).

If this labiality effect reflects the existence of a smaller distinction between /b/ and /p/ than between /d/ and /t/ or between /g/ and /k/, we can hope for a negative interaction effect between the ternary [*place*] contrast and [*voice*]. The model did not, however, reliably detect this (CI = −0.36 .. +0.23).

Differences between coronals and dorsals were detected neither for the binary [*place*] effect (CI = −0.12 .. +0.13) nor for its interaction with the [*voice*] effect (CI = −0.21 ..+0.27).

Individual differences. When visually comparing all Figures 1–3, it seems that speakers J and M, and perhaps L, implement /p/ as equally voiced as the [+*voi*] plosives. The only way in which our mixed-effects model can hint at the existence of such differences of strategies between speakers lies in the standard deviations of the random effects: the between-speaker standard deviation is highest (48 percent of the closure voicing scale) for the interaction between ternary [*place*] and [voice], and lower (between 8 and 21 percent) for the other five by-speaker random effects.

5. DISCUSSION

Besides an obvious expected result ([+*voi*] plosives have more closure voicing than [–*voi*] plosives), we have found several effects that can be ascribed to differences between Palestinian Arabic and English.

The structural challenge of PA lacking /p/ could lead to /p/ merging into /b/ for some learners. This phenomenon can lie behind the detected ternary [*place*] effect ($p = 0.015$), as well as (tentatively) behind the large between-speaker standard deviation for the interaction between [*voice*] and ternary [*place*]. Future modelling with a latent predictor (L2 phonological merger of /p/ with /b/) may shed more light on this issue. It is already clear from the 10 speakers who totally devoice /p/ in Figure 3, several of whom must have full closure voicing for all [+voi] plosives according to Figure 1, that there are speakers who do master the English /p/~/b/ opposition.

The only phonetic implementation challenge we studied was the English devoicing of initial [+voi] plosives. According to Figure 1, a few PA learners master this, but most don't, even after 10 years of native input.

We found no evidence for a hierarchical or conditional or implicational relationship between these two challenges. Both challenges might be remedied by specific minimal-pair training [14, 15].

We conclude that Palestinian Arabic learners of English vary in the extent to which they carry over their native sound structure to their L2 (some merge /p/ with /b/, some don't), as well in the extent to which they carry over their native phonetic implementation.

6. REFERENCES

[1] Seinhorst, K. 2021. *The Learnability and Complexity of Phonological Patterns: Simulations, Experiments, Typology*. Doctoral thesis, University of Amsterdam.

[2] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *WORD*, 20(3), 384–422.

[3] Tamim, N., Hamann, S. 2021. Voicing contrasts in the singleton stops of Palestinian Arabic: Production and perception. *Proc. Interspeech 2021* Brno, 401–405. doi: 10.21437/Interspeech.2021-1079

[4] Flege, J.E. & Port, R. 1981. Cross-language phonetic interference: Arabic to English. *Language and Speech* 24(2), 125–146.

[5] Simon, E. 2009. Acquiring a new second language contrast: An analysis of the English laryngeal system of native speakers of Dutch. *Second Language Research* 25(3), 377–408.

[6] MacKay, I.R.A., Flege, J.E., Piske, T., Schirru, C. 2001. Category restructuring during second-language speech acquisition. *J. Acoust. Soc. Am.* 110(1), 516–528.

[7] Davidson, L. 2016. Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics* 54, 35–50.

[8] Boersma, P., Weenink, D. 1992–2022. *Praat: Doing Phonetics by Computer*. https://www.praat.org.

[9] Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences* Amsterdam 17, 97–110.

[10] Klatt, D.H. 1975. Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research* 18(4), 686–706.

[11] Tamim, N., Hamann, S. 2023. The influence of native speaker input on the production of the English voicing contrast in stops by Palestinian Arabic pupils. Manuscript, University of Amsterdam.

[12] R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org.

[13] Bates, D., Mächler, M., Bolker, B, Walker, S. 2015. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software* 67(1), 1–48.

[14] Barlow, J.A., Gierut, J.A. 2002. Minimal pair approaches to phonological remediation. *Seminars in Speech and Language* 23(1), 57–68.

[15] Hamzah, M. H., Bawodood, A. M. A. 2019. Teaching English sounds via minimal pairs: The case of Yemeni EFL learners. *Veda's Journal of English Language and Literature* 6(3), 97–102.

The Effect of Age of Entry and Contact on Backward Transfer of Glaswegian English (L2) on Hindi and Indian English (L1)

Divyanshi Shaktawat

Glasgow University Laboratory of Phonetics, University of Glasgow
d.shaktawat.1@research.gla.ac.uk

ABSTRACT

This study investigated phonological backward transfer in the ethnolinguistic minority of first-generation bilingual immigrant Indians in Glasgow 'Glaswasians' [1] in relation to Flege's Speech Learning Model [2, 3], which predicts `assimilation' and `dissimilation' of sound categories. The study explored whether and how sounds of Glaswasian native language (Hindi) and dialect (Indian English) are influenced by sounds of the dominant host language/dialect (Glaswegian English). Age of Entry (AoE) in Glasgow and amount of Indian Contact were also examined. Two control groups (Indians and Glaswegians) and an experimental group (Glaswasians) were recorded reading in English and Hindi words containing two phones (/t/ - VOT, /l/- F2-F1 difference). In both languages, Glaswasian VOT became more Glaswegian-like (assimilation), whereas F2-F1 difference diverged to an even greater degree than the Indians (dissimilation). Increasing AoE and Indian Contact decreased assimilation of /t/ or increased dissimilation of /l/.

Keywords: Glasgow Indians, Indian English, Hindi, Glaswegian English, Backward Transfer

1. INTRODUCTION

Languages in the bilingual mind interact and influence each other [2, 4]. This influence can be from the native language (L1) to the second language (L2) (*Forward Transfer)* [5] or from the L2 to L1 (*Backward Transfer*) [6 – 9]. The focus of the present study is the latter. In the wider context of Second Language Acquisition, much evidence has been found for phonological backward transfer across languages [6 – 9]. However, it is much less studied in dialects, and further still, even less is known about how backward transfer across languages is different from backward transfer across dialects.

The present study investigates first-generation bilingual Indian immigrants in Glasgow who are native speakers of Hindi and Indian English and are termed 'Glaswasians' [1]. It examines how certain sounds of speakers' native and heritage language (Hindi) and dialect (Indian English - IE) have changed after coming into contact with a new language and dialect (Glaswegian English - GE), which is the host language and majority variety in their current environment (Glasgow).

An explanation for backward transfer comes from the Speech Learning Model (SLM) [2, 3] which proposes that all speech sounds of a bilingual's languages exist in the same phonetic space. This allows for these phonological systems to interact and mutually influence each other which causes bidirectional transfer. With increased L2 input, an L2 learner may sense enough dissimilarity between two perceptually similar L1-L2 categories, to seek to maintain a contrast between these L1 and L2 categories in the common phonetic space. One way to maintain this contrast is by deflecting the categories away from each other. This is called *phonetic category dissimilation* [2] and may lead to non-native like pronunciation of the native sound.

However, as the age of acquisition increases, this ability to distinguish between phonetically similar sounds is reduced by the *Perceptual Mechanism of Equivalence Classification* [2]. Such perceptually linked L1 and L2 sounds will then over time come to resemble one another and processed as the same sound category. Thus, a new L2 category is not formed, but the existing L1 category is appropriated. This may happen because in late L2 learners, the phonological categories of their L1 are so well established, that instead of creating new categories for the similar L2 sound, they rather assimilate it to the existing similar L1 category. A similar argument is made by the Perceptual Assimilation Model [10], where L2 learners perceptually assimilate non-native phones to native phone categories based on perceived similarity between them.

In contrast, early learners of L2, are exposed to both the host language (which in this case is GE and is both a new language and a new dialect) and heritage languages from an early age. Thus, they acquire the linguistic features of both the host and heritage languages in childhood when their categories are more plastic and still developing [2, 11]. Furthermore, early L2 learners may have more exposure to the host language by way of education and communicating with multiple social groups and peer networks. Thus, their L2 may even become dominant allowing transfer to similar L1 categories. Additionally, the quality of input received by early

and late L2 learners (here, first-generation immigrants) may also be different. While late first-generation speakers are more likely to learn L2 from speakers of their own community, who are themselves L2 speakers of the host language, early first-generation speakers are more likely to learn L2 from those who speak it is as a native language in the host country [11]. Thus, early or late Age of Entry (AoE) in the L2 country can also have implications for transfer in first-generation immigrants [11].

In addition to the AoE in the host country, much research has found evidence that the amount of contact with the host or heritage community also has implications for transfer, such that higher contact with the host community is associated with more transfer from the host language [12 – 15].

In light of these findings, the current study seeks to better understand the nature of backward transfer from host to heritage languages, and further asks whether the patterns of transfer across languages are replicated in patterns of transfer across dialects. It focusses on the following three research questions.

1.1. Is there a backward transfer of GE on Hindi and IE? If yes, does it appear as assimilation or dissimilation?

Transfer was studied in two phonetic features: Acoustic darkness of /l/ and aspiration (VOT) of /t/.

GE has darker /l/, and therefore smaller F2-F1 difference [16, 17]. IE and Hindi, on the other hand, are known to have clearer /l/, therefore, larger F2-F1 difference [17, 18]. In case of assimilation, Glaswasian /l/ will become more Glaswegian-like with smaller F2-F1 difference (darker /l/). However, if there is dissimilation, Glaswasian will develop clearer /l/ as compared to Indians, to contrast it from darker Glaswegian /l/. If there is no transfer, then Glaswasian /l/ will have similar F2-F1 difference as Indians.

GE /t/ has a denti-alveolar realization and in word-initial position shows aspiration, reflected in longer lag VOT [19]. On the other hand, /t/ in IE has a retroflex realization and is not aspirated; Hindi shows dental and retroflex stops, and a retroflex stop is usually used for English /t/. In both Hindi and IE, this stop has shorter VOT than GE [18, 20]. In case of assimilation, Glaswasians will develop Glaswegian-like longer VOT. In case of dissimilation, Glaswasians will develop even shorter VOT than Indians to contrast it from longer Glaswegian VOT. If there is no transfer, then Glaswasians will have similar VOT as Indians.

1.2. Is there an effect of AoE and amount of Indian Contact on this transfer?

For AoE, I predicted that early arrival in Glasgow (lower AoE) will be associated with higher transfer from GE. For Indian Contact, I predicted that less Indian Contact will be associated with higher transfer from GE. These predictions apply to both phone categories. Furthermore, with respect to the SLM, transfer could appear as assimilation or dissimilation.

1.3. Does IE (as compared to Hindi) receive more transfer from GE on the account that they are both dialects of the same language?

Previous research has suggested that in cases of language contact, mutually intelligible dialects influence each other [8, 21]. Based on this, it is predicted that as they are dialects of the same language, IE will receive more transfer from GE than Hindi, which is a different language altogether.

2. METHODOLOGY

Three groups of speakers were recruited: two controls groups (Glaswegians and Indians) and one experimental group (Glaswasians). Glaswegians (n = 31) comprise native speakers of GE who reside in Glasgow. Indians (n = 31) comprise native speakers of Hindi and IE who reside in India and have never had any contact with GE. Finally, Glaswasians (n = 41) comprise first-generation Indian immigrants in Glasgow. They are native bilinguals in Hindi and IE and have been residing in Glasgow for at least 3 years. The three groups were recorded reading sentence lists in English and Hindi (only in English for 'Glaswegians'). There were 10 words per target sound and the target word was embedded in the carrier frame of *Say __ again*. in English and *kəha: _ a:pne?* (direct translation: Said__you?) in Hindi.

Using a questionnaire, data was collected on AoE and amount of Indian Contact only from Glaswasians. AoE is the age (in years) at which the Glaswasian participant arrived in Glasgow (*min*. = 12, *max*. 36, *mean* = 26.19, *SD* = 6.38). Indian Contact represents the amount of contact with the participants' own ethnic group (Indian). There was a total of 11 questions such as "How often do you spend free time with your close Indian friends?". Participants responded on a scale ranging from 'never' (1) to 'almost always' (5). A higher value represents higher contact with Indians (*min*. = 20, *max*. 50, *mean* = 37.16, *SD* = 7.35; score out of 57).

The data is analysed in two steps using the lme4 package (version 1.1.29) [22] in R (version 3.6.3) [23]. First, the speech data is acoustically analysed and compared across the three groups separately for each phonetic feature. Second, if backward transfer is found for any phonetic feature,

it is then analysed as a function of AoE and Indian Contact.

3. RESULTS

3.1 Acoustic analysis across groups

3.1.1 Word-initial /l/: F2-F1 difference

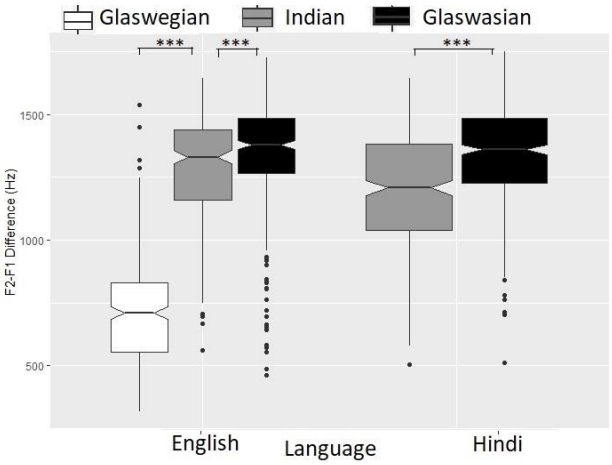


Figure 1: F2-F1 difference across Group and Language

Linear mixed modelling revealed that, consistent with previous research, Glaswegians had smaller F2-F1 difference (darker /l/) than Indians who had larger F2-F1 difference (clearer /l/) in English (β = -568.97, $t(604)$ = -39.11, p < .001) (Figure 1). Subsequent analysis indicated that in English, Glaswasians had higher F2-F1 difference (clearer /l/) than Indians, who had smaller F2-F1 difference (darker /l/) (β= 52.25, $t(1341)$ = 3.33, p < .001). This is consistent with the pattern of dissimilation, where Glaswasians are exaggerating the native F2-F1 difference to maintain a contrast with the smaller F2-F1 difference in GE. Furthermore, the difference between Indians and Glaswasians in Hindi was larger than in English (β= 86.34, $t(1341)$ = 4.01, p < .001). This indicates that Hindi has undergone more dissimilation than English.

3.1.2 Word-initial /t/: log VOT

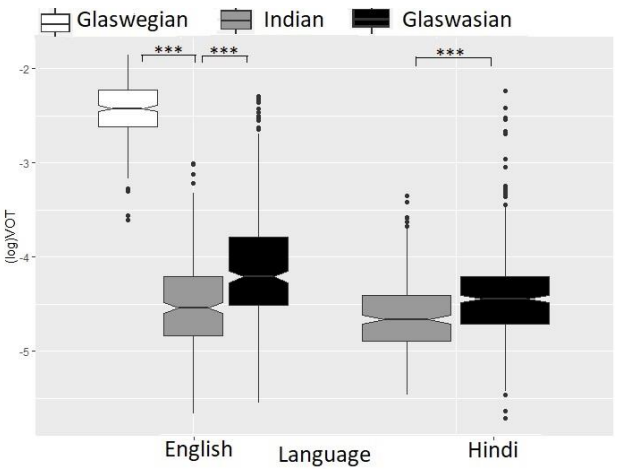


Figure 2: log VOT across Group and Language

Linear mixed modelling revealed that Glaswegians had longer VOT than Indians in English (β = 2.05, $t(589)$ = 76.31, p < .001) (Figure 2). Further analysis showed that Glaswasians had longer VOT than Indians (β= 0.35, $t(1272)$ = 9.58, p < .001) in English. This is indicative of assimilation, where Glaswasian VOT has become longer, that is, more Glaswegian-like. Additionally, the difference between Glaswasians and Indians was much bigger in English than in Hindi (β = -0.18, $t(1272)$ = -3.57, p < .001). That is, English underwent more assimilation than Hindi for VOT.

3.2 Analysing the effect of AoE and amount of Indian Contact on transfer

As transfer was found in both /l/ and /t/, I examined whether AoE and Indian Contact affected the degree of transfer across languages and accounted for the variability in the data (Figures 1 and 2).

3.2.1 AoE

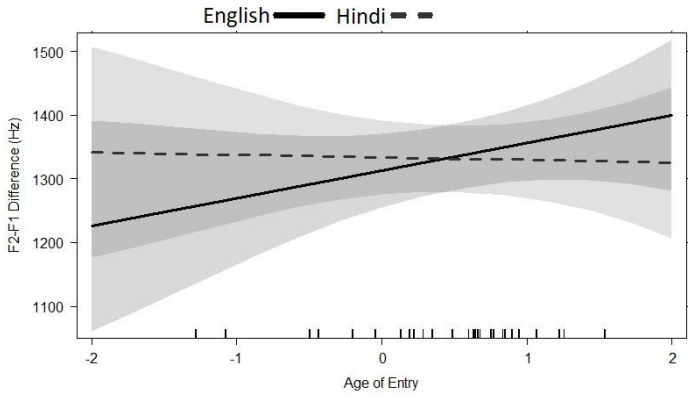


Figure 3: Interaction effect of AoE and Language on F2-F1 difference

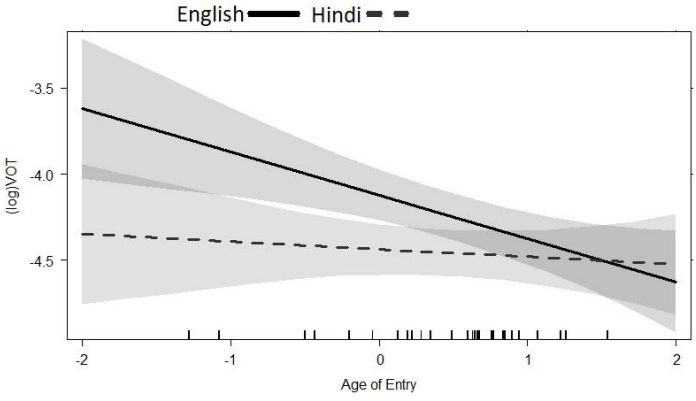


Figure 4: Interaction effect of AoE and Language on log VOT

F2-F1 difference and log VOT were analysed in separate linear mixed models and significant

interactions between Language and AoE were found for both (see Figures 3 and 4). For F2-F1 difference, higher AoE had no effect on Hindi, whereas it increased dissimilation in English (β = -47.31, t(565) = -2.74, p = .006). For VOT, as the AoE increased, log VOT in English became shorter, whereas it was unaffected for Hindi. In other words, there was decreasing assimilation in English as the AoE increased, but no change in Hindi.

3.2.2 Indian Contact

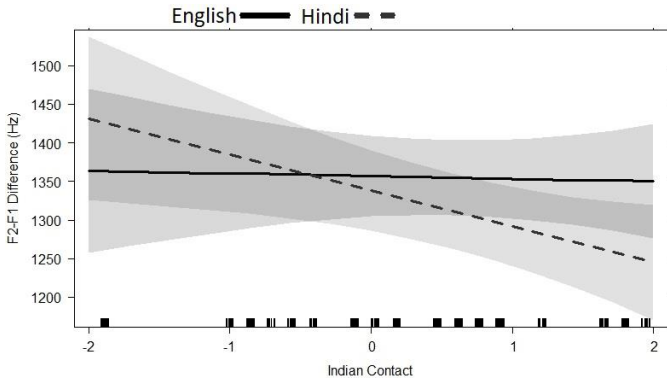


Figure 5: Interaction effect of Indian Contact and Language on F2-F1 difference

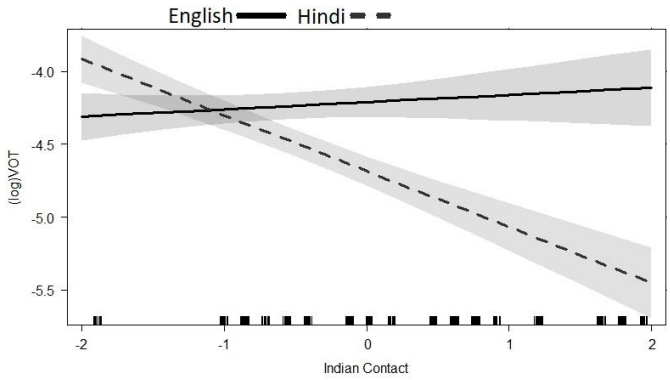


Figure 6: Interaction effect of Indian Contact and Language on log VOT

In separate linear mixed models, F2-F1 difference and log VOT were analysed, and significant interactions effects were found between Language and Indian Contact for both (see Figures 5 and 6). For F2-F1 difference, an increase in Indian Contact had no effect in English, whereas it was associated with a decrease in F2-F1 difference (lesser dissimilation) in Hindi (β = -43.13, t(424) = -2.23, p = .026). For VOT, an increase in Indian Contact had no effect on log VOT in English, whereas it led to a steeper decrease in log VOT (lower assimilation) in Hindi (β= 0.21, t(546) = 4.67, p < .001).

4. DISCUSSION AND CONCLUSION

This study investigated phonological backward transfer in first-generation immigrant Indians in Glasgow. Specifically, it asked three questions: (1) Is there phonological backward transfer from GE to Hindi and IE, and if yes, does that appear as assimilation or dissimilation, (2) Is there an effect of AoE and Indian Contact on transfer, and (3) Whether IE will receive more transfer from GE on account that both are dialects of the same language.

With respect to the first research question, I found backward transfer of GE on both Hindi and IE /l/ [7] and /t/ [11]. It manifested as dissimilation in /l/ (higher dissimilation in Hindi than in English) and assimilation in /t/ (higher assimilation in English than in Hindi). It may be possible that formant patterns are more vulnerable to dissimilation, whereas VOT is more susceptible to assimilation. However, more research is required to make such a claim.

With respect to the second question, the results on the effect of AoE and Indian Contact on transfer support previous research [12 – 15], but also add to our knowledge of how it might be different across languages and dialects. In relation to the effect of AoE, two points can be concluded. First, English was more vulnerable to AoE effects. This was found for both /l/ and /t/. Second, as AoE increased, the values became more Indian-like (exaggeratedly so in case of /l/). Contrary to this, those who arrived early in Glasgow had Glaswegian-like darker /l/ and longer VOT in English. Two such similar conclusions can also be drawn for the effect of Indian Contact. First, Hindi was more vulnerable to the effects of Indian Contact in /l/ as well as /t/, whereas there was no effect of it in English. Second, as the amount of Indian Contact increased, the values became more Indian-like. This is even argued for /l/ as the range of Glaswasian F2-F1 difference is similar to that of Indians, not Glaswegians.

Finally, in relation to the third question, it is not clear if IE is more susceptible to transfer from GE. This is because Hindi underwent more transfer for F2-F1 difference, but English underwent more transfer for VOT. However, this much is clear that English was more vulnerable to AoE effects, and Hindi to the effects of Indian Contact. This may be because AoE represents influence from the host language (GE), whereas Indian Contact represents influence from the heritage language (Hindi). In this light, it makes sense that AoE affected Indian English more than Hindi but Indian Contact affected Hindi more than Indian English.

REFERENCES

[1] Alam, F. 2006. Language and Identity in `Glaswasian' Adolescents. University of Glasgow.

[2] Flege, J. E. 1995. Second Language Speech Learning Theory, Findings, and Problems. In Strange, W. (ed), *Speech perception and linguistic experience. Issues in Cross-Language Research*. York Press, 233–277.

[3] Flege, J. E., Bohn, O. 2021. The Revised Speech Learning Model (SLM-r). In: Wayland, R. (ed.), *Second Language Speech Learning:Theoretical and Empirical Progress*. Cambridge University Press, 3–83.

[4] Weinreich, U. 1953. *Languages in Contact: Findings and Problems*. Linguistic Circle of New York.

[5] Lado, R. 1957. *Linguistic Across Cultures*. University of Michigan Press.

[6] Chang, C. 2012. Rapid and multifaceted effects of second-language learning on first-language speech production. *J. Phon.* 40, 249–268.

[7] Bergmann, C., Nota, A., Sprenger, S. A., and Schmid, M. 2016. L2 immersion causes non-native-like L1 pronunciation in German attriters. *J. Phon.*, 58, 71–86.

[8] de Leeuw, E. 2009. When your native language sounds foreign: A phonetic investigation into first language attrition. Queen Margaret University.

[9] Lev-Ari, S., Peperkamp, S. 2013. Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *J. Phon.*, 41, 320–331.

[10] Best, C. T. 1995. A direct realist view of cross-language speech perception. In Strange, W. (ed), *Speech perception and linguistic experience. Issues in Cross-Language Research*. York Press, 171–199.

[11] McCarthy, K., Evans, B. G., Mahon, M. 2013. Acquiring a second language in an immigrant community : The production of Sylheti and English stops and vowels by London-Bengali speakers. *J. Phon.*, 41, 344–358.

[12] de Leeuw, E., Schmid, M. S., Mennen, I. 2010. The effects of contact on native language pronunciation in an L2 migrant setting. *Bilingualism*, 13, 33–40.

[13] Torgersen, E., Kerswill, P., Fox, S. 2006. Ethnicity as a source of changes in the London vowel system. In: Hinskens, F. (ed.), *Language Variation - European Perspectives*. Benjamins, 249–263.

[14] Fox, S. 2010. Ethnicity, Religion and Practices: Adolescents in the East End of London. In: Llamas, C., Watt, D. (eds). *Language and Identities*. Edinburgh University Press, 144–156.

[15] Stuart-Smith, J., Timmins, C., Alam, F. 2011. Hybridity and ethnic accents: A sociophonetic analysis of 'Glaswasian'. In: Gregersen, F., Parrott, J., Quist, P. (eds.), *Language Variation - European Perspectives III: Selected papers from the 5th International Conference on Language Variation in Europe.* John Benjamins Publishing Company, 43–58.

[16] Stuart-Smith, J. 2004. Scottish English: phonology. In: Kortmann, B., Schneider, E.W. (eds.), *A Handbook of Varieties of English: 1: Phonology*. Mouton de Gruyter, 47–67.

[17] Wells, J. C. 1982. *Accents of English Vol. 3*. Cambridge University Press.

[18] Gargesh, R. 2008. Indian English: phonology. In: Mesthrie, R., Kortmann, B., Schneider, E. W. (eds.), *A Handbook of Varieties of English*. Mouton de Gruyter, 231-243.

[19] Stuart-Smith, J., Sonderegger, M., Rathcke, T., Macdonald, R. 2015. The Private Life of Stops: VOT in a Real-Time Corpus of Spontaneous Glaswegian. *Lab. Phon.* 6, 505–549.

[20] Kaur, D. 2020. A Phonetic Comparison Between Indian English and British English. University of Barcelona.

[21] Trudgill, P. 1986. *Dialects in Contact*. Basil Blackwell.

[22] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67.

[23] R Core Team. 2020. R: A language and environment for statistical computing." R Foundation for Statistical Computing. https://www.r-project.org/.

FRENCH PROSODIC STRESS PRODUCED BY TWI SPEAKERS COMPARED TO FRENCH SPEAKERS: AN ACOUSTIC INVESTIGATION

Kofi Adu Manyah
Rejoice Owusua Amponsah

Kwame Nkrumah University of Science and Technology
kofi.adu6@gmail.com amponsahnezem@gmail.com

ABSTRACT

French Prosodic Stress is acoustically related to a combination of duration, fundamental frequency (f0) and intensity. This paper presents a comparative study investigating the use of these correlates by French native speakers and Twi speakers learning French. French speakers and students studying French were selected for this study. In a production experiment, their readings were recorded in a studio and analyzed with a sound Editor. Durational, intensity and fundamental frequency values were extracted for each of the selected target syllables. Results show that French and Twi speakers do not use these acoustic correlates in similar manner. Duration seems to be the preferred, reliable cue used by Twi speakers acquiring French prosodic stress, unlike the French speakers who combine all three cues. Acoustic results further show which syllable types were more stressed.

Keywords: Duration, intensity, fundamental frequency, syllabic stress.

1. INTRODUCTION

French prosodic stress operates within the framework of the rhythmic unit therefore it is imperative to recognize the strong and weak positions that determine pronunciation habits. Wioland [32] showed that syllables occupy 3 positions: (i) an accented/stressed position, called the rhythmic stress, the final position of a rhythmic unit, is a position to be privileged par excellence and on which the main part of teaching of pronunciation and acquisition of prosodic stress must be focused; (ii) a stressable position which, in spite of the brevity of its realization, keeps a vocalic timbre comparable to that of the stressed position; (iii) an unstressable position whose syllabic structure, mostly open, does not require the differentiation of timbres for certain vowels. A syllable is in the unstressed position because it is not the final syllable of the rhythmic unit. A two stress system, Initial Accent (IA) and Final Accent (FA) [6, 10, 11], penultimate accentuation in certain varieties of French [27], and a 'stressless?' type [25] have been accounted for the French language. This study will focus on the first category i.e. the syllable in the final position of a rhythmic group, which is a group of words spoken in one breath, where there is no pause between words. Punctuation marks help to distinguish rhythmic groups in a long sentence.

The place of the French stress is always on the last syllable of the word or group of words. In other words, the French stress strikes the final syllable of any group of words linked by meaning and not separated by a pause. The place of the stressed syllable therefore varies from one statement to another. The final syllable of each rhythmic group is pronounced with a longer duration than the syllables preceding it. For a consonant, the final position of the syllable, whatever the place of the syllable in the rhythmic unit, is a weak position which leads to a systematic relaxation of its articulation.

Duration: Syllabic duration/lengthening i.e. short vowels vs. long vowels is the quality of a sound which is related to the time it takes to produce a phoneme. In French, the open or closed structure of the syllable determines timbre and vowel duration. Variations in duration of stressed vowels comprise three types of duration: unmarked/short duration, marked/long duration and very marked/very long duration [32]. Short duration includes not only open syllables but also closed syllables with voiceless consonants [p t k f s ʃ] and [l] e.g. bref [ˈbʁɛf], tasse [ˈtas].

Long duration is the intermediate between short and very long duration. They have closed syllables i.e. oral vowels plus [b d g m n ɲ j]. Very long duration exhibits closed syllables with oral vowels plus these voiced consonants [z v ʒ ʁ] and [vʁ] e.g. brève [ˈbʁɛːv], tard [ˈtaːʁ] père [ˈpɛːʁ], livre [ˈliːvʁ].

A stressed syllable has a relatively longer duration, as well as a higher intensity and a higher fundamental frequency (f0). In this study, we dwell on the third category: very long duration.

Aim

This study reports observations and results obtained in an experiment on the comparison of the production of French prosodic stress by French native speakers

and learners of French as a foreign language viz. Twi speakers.

1. f0 values of the target syllables (final) and the preceding syllables (penultimate) were obtained in order to verify that prosodic stress contrasts indeed rely on factors pertaining to difference in pitch.
2. f0 values were calculated for the penultimate and final syllables.
3. Intensity values were calculated for the two juxtaposing syllables.
4. Durational values were taken for the two syllables.

Statistical analyses (ANOVAs) were carried out on all measures obtained from the speakers ($p \leq 0.01$).

2. METHOD

Two adult male and female native French Speakers (FS) and three male and four female native Twi Speakers (TS), with no speech or hearing impairment, served as speakers. Twi is a register tone language of the Kwa Group, spoken in Ghana. It is one of the languages spoken by the Akan including people in the Asante Region, parts of the Western, Eastern, Central, Volta, Bono, Bono East and Ahafo Regions [2]. All Akan languages are mutually intelligible. The data consisted of acoustic productions i.e. reading a text, at a self-selected conversational rate. Speakers were not given any indication of the position of the stress on the syllables. The reading was produced at least 3 times by each speaker. Material: 11 syllables of 20 (bold) were selected from *Cauchemar*, Paul Verlaine's *Poèmes saturniens*. The poem [31] contains stress patterns and challenges in varying degrees which permitted to make a reasonably satisfactory analysis: CAUCHE**MAR**

J'ai vu passer dans mon **rêve**
Tel ouragan sur la **grève**,
D'une main tenant un **glaive**
Et de l'autre un sablier,
Ce cavalier
Des ballades d'Allemagne
Qu'à tra**vers** ville et campagne,
Et du **fleuve** à la montagne,
Et des forêts au vallon,
Un étalon
Rouge-flamme et **noir** d'ébène,
Sans bride, ni **mors**, ni rêne,
Ni hop ! ni cravache, entraine
Parmi des râlements **sourds**
Tou**jours** ! tou**jours** !
Un grand feutre a longue plume
Ombrait son œil qui s'allume
Et s'éteint. Tel, dans la brume,
Éclate et **meurt** l'**éclair** bleu
D'une arme à feu.
Comme l'aile d'une orfraie
Qu'un subit orage effraie,
Par **l'air** que la **neige** raie.
Son manteau se soulevant
Claquait au vent,
Et montrait d'un **air** de **gloire**
Un **tor**se d'ombre et d'i**voire**,
Tandis que dans la nuit **noire**
Luisaient-en des cris stridents
Trente-deux dents.
Paul Verlaine, Poèmes saturniens

By means of PRAAT sound editor, fundamental frequency (f0) and Intensity measurements were taken for the target stressed syllables and the corresponding preceding syllables. Durational measurements of the stressed syllables were also done for the:

a) target syllable
b) preceding syllable

Intensity, f0 and Duration are key because it is widely known that a prominent syllable is marked by variations in four acoustic cues: fundamental frequency (pitch), amplitude (volume), duration (length), and formant structure (a different timbre of the vowel). Of all the cues f0 seems the most important in terms of perception in French [13]. In the Twi language, like many tone languages in E. Asia and Africa, there is also a one-to-one association between tone and syllable i.e. each monosyllabic word or morpheme is associated with a tone [1, 34, 35].

A stressed syllable, like a high tone, is said to be distinguished not only by acoustic features but also has to be considered as a perceptual phenomenon. Therefore, in defining it one must account not only for its production but also for its perception by the interlocutor [13]. How do Twi learners (perceive and) produce the unfamiliar prosodic stress distinctions of the French language? Given that the amount of linguistic experience in stress use does not guarantee a facilitating effect for the learners of another foreign language. A study [3] on production of English lexical stress by TS revealed that f0 is one of the determining factors in distinguishing stress differences in English, but the key factor is intensity. The core question of the present study, therefore, was how French prosodic stress may be identified and produced by Twi learners of the French language, in terms of rhythmic groups and at the sentence level.

Experimental Design

This experiment was performed to find an answer to the question. In this preliminary study, we sought to compare native speaker stress production and the language learner production of the same text. Two French (1 male, 1 female) native speakers and seven (3 males and 4 females) Ghanaian final year students pursuing a four-year undergraduate program in French were recruited to accomplish the task. To best enable us obtain answers to the questions, Purposive

Sampling was used in selecting the subjects. The students had previous exposure to the Oral French Language and its prosodic structure.

To detect the presence of differences in stress, we juxtaposed duration, intensity and fundamental frequency values of all preceding (penultimate) syllables of the target (final) syllables produced by each speaker.

Procedure: Acoustic data were recorded in an anechoic room for all speakers at TekTV Recording Studio, KNUST. A subject read the poem (3 times). DAW (digital audio workstation) used in executing the production is Steinberg Cubase 5 for both recordings and mixing. The audio sample rate of the project was set to 44100k Hz with Bit Depth of 16 Bit. A short training session consisting of reading examples preceded the test.

Results: Conforming to standards, French prosodic stress production by TS using one of the three cues (duration) was displayed in this experiment. The performance for duration was very high (100% of the time) in the production of stressed syllables by both French and Twi speakers (Tables 2 and 3). The second most dominant parameter was the fundamental frequency for the two categories of Speakers: FS (69%), TS (15.8%). The performance rate for the intensity cue, was the poorest, the least used by native FS (36%) and TS (5.3%) to distinguish the two classes of syllables: unstressed (penultimate) and stressed (absolute final). The production rate of the prosodic stress in duration was therefore higher than f0 and intensity (Table 1).

3. RESULTS AND DISCUSSION

Table 1 below illustrates the general tendency of the results i.e. the success rate of production of duration, f0 and intensity of the stressed syllables by the two groups of speakers. Data were obtained by averaging over the individual values of those measurements for all eleven tokens of the corpus. The overall data indicate that the most important parameter for determining prosodic stress assignment contrasts, i.e. relative duration, is highly significant ($p<0.001$).

In absolute terms, the general tendency i.e. performance rate of 100% of duration of the target syllables were greater than the penultimate syllables for both native FS and TS. The second most dominant parameter was the f0 with a performance rate of 32%. On intensity, the most stressed syllable was /ʒu:ʁ/, the least being /vwa:ʁ/ and /su:ʁ/ by TS. Intensity values in absolute terms for the target syllables were higher than those of the penultimate syllables for the FS compared to TS.

Native speakers combined the 3 acoustic cues in varying degrees with duration, f0 and intensity whereas Twi learners relied on duration (100%) but less on f0 and intensity, like the native speakers. Here, the most stressed syllables are /su:ʁ/, and /glɛ:v/, the least stressed being /nwa:ʁ/ for the TS. On the f0, the most stressed syllable was /ʒu:ʁ/ for the TS. The most stressed syllable is /ʒu:ʁ/, the least being /vwa:ʁ/ and /su:ʁ/ for the TS in intensity. All in all, the Ghanaian learners (TS) stress syllables in final position lesser/shorter than the FS, as far as duration is concerned. The marked stress in the study is evident in Tables 1, 2, 3. Syllable type and syllable position are factors that influenced the production of prosodic stress. Many syllables that are in the final positions tend to be very long e.g. /ʒu:ʁ/ and /glɛ:v/.

	French Speakers (FS)			Twi Speakers (TS)		
syl	duration (ms)	f0 (Hz)	intensity (dβ)	duration (ms)	f0 (Hz)	intensity (dβ)
/ma:ʁ/	386	197	68	402	196	65
/ʁɛ:v/	453	200	69	392	173	63
/gʁɛ:v/	497	187	66	380	164	62
/glɛ:v/	524	200	66	419	183	63
/nwa:ʁ/	287	191	70	386	188	66
/mɔ:ʁ/	421	199	72	425	210	66
/su:ʁ/	500	314	66	439	185	63
/ʒu:ʁ/	531	199	70	392	212	66
/ʒu:ʁ/	289	174	66	372	162	63
/glwa:ʁ/	520	189	67	414	180	64
/vwa:ʁ/	451	199	66	421	170	64
AVER	442	206	68	404	184	64
STDEV	89	39	2	22	17	2

Table 1: Average and Standard Deviation for Duration, Fundamental Frequency and intensity: French and Twi speakers.

Of the three cues, f0 seems the most important in terms of perception in French [13]. The acoustic results of this study, involving native and non-native speakers suggest that, of the three cues, duration seems to be the most important in terms of production in French (Tables 1, 2, 3). However, intra-speaker comparison of production reveals that speakers' performance rate is not the same for the three cues. In fact, the performance rate is higher in f0 for the female French speaker (82%) than the male speaker (55%). The performance rate is 46% female and 36% male respectively with regards to intensity (Table 2). Durational values range between 287 ms and 524 ms (FS) and 372 ms and 419 ms (TS) with a very low standard deviation of 22%. The corresponding figures for intensity ranges from 66 dβ and 72 dβ (FS) and 62 dβ and 66 dβ (TS) with a very low standard deviation of 2% for both FS and TS.

TS seem to lengthen /ma:ʁ/ and /mɔ:ʁ/ more than the FS. The results differ from one stressed syllable to another. The French production of 7 syllables

/ʁɛːv/, /gʁɛːv/, /glɛːv/, /suːʁ/, /ʒuːʁ/, /glwaːʁ/ and /vwaːʁ/ are all higher than those of the Ghanaians for all three acoustic parameters namely, duration, fundamental frequency and intensity. The average production for all syllables follows the same trend, duration is 442 ms for FS and 404 ms for TS, f0: 206 Hz and 184 Hz, Intensity: 68 dβ and 64 dβ for FS and TS respectively, with relatively low standard deviations (Table 1).

	syllable	duration (ms)	intensity (dβ)	f0 (Hz)
Penultimate Sy	/koʃ/	242	72.1	324.4
Stressed Sy	/maːʁ/	391	68.3	259.4
Difference +/-		+149	-3.8	-65.0
Penultimate Sy	/mɔ̃/	203	69.4	200.5
Stressed Sy	/ʁɛːv/	470	67.5	217.2
Difference +/-		+267	-1.9	+16.7
Penultimate Sy	/la/	111	66.0	169.8
Stressed Sy	/gʁɛːv/	521	66.3	209.2
Difference +/-		+410	+0.3	+39.4
Penultimate Sy	/ ɛ̃/	194	66.1	176.8
Stressed Sy	/glɛːv/	552	67.3	247.0
Difference +/-		+358	+1.2	+70.2
Penultimate Sy	/e/	79	71.0	224.9
Stressed Sy	/nwaːʁ/	312	69.9	222.2
Difference +/-		+233	-1.1	-2.7
Penultimate Sy	/ni/	154	74.1	251.1
Stressed Sy	/mɔːʁ/	441	74.9	253.9
Difference +/-		+287	+0.8	+2.8
Penultimate Sy	/mɑ̃/	172	69.2	221.0
Stressed Sy	/suːʁ/	430	66.1	225.5
Difference +/-		+258	-3.1	+4.5
Penultimate Sy	/tu/	120	73.5	238.3
Stressed Sy	/ʒuːʁ/	462	72.9	264.3
Difference +/-		+342	-0,6	+26.0
Penultimate Sy	/tu/	133	64.1	167.2
Stressed Sy	/ʒuːʁ/	312	67.1	216.0
Difference +/-		+179	+3.0	+48.8
Penultimate Sy	/də/	132	71.7	214.2
Stressed Sy	/glwaːʁ/	451	69.7	223.4
Difference +/-		+319	-2.0	+9.2
Penultimate Sy	/di/	112	69.1	231.3
Stressed Sy	/vwaːʁ/	434	70.2	257.4
Difference +/-		+322	+1.1	+26.1

Table 2: Duration, intensity, fundamental frequency values for French native speaker (female)

The closed syllable type influenced the speakers' production of stress. When the final phonemes /ʁ/ and /v/ were presented to the FS subjects and the TS subjects, they consistently assigned final stress (duration) to the word 100% of the time. It would be worthwhile to investigate the performance rate of speakers in other post-vocalic consonantal /z, ʒ, vʁ/ environment.

	syllable	duration (ms)	intensity (dβ)	f0 (Hz)
Penultimate Sy	/koʃ/	311	61.1	251.1
Stressed Sy	/maːʁ/	505	63.1	238.1
Difference +/-		+194	+2.0	-13.0
Penultimate Sy	/mɔ̃/	252	60.0	196.1
Stressed Sy	/ʁɛːv/	329	60.7	208.0
Difference +/-		+77	+0.7	+11.9
Penultimate Sy	/la/	256	61.0	202.6
Stressed Sy	/gʁɛːv/	414	62.1	204.5
Difference +/-		+158	+1.1	+1.9
Penultimate Sy	/ ɛ̃/	276	66.1	221.7
Stressed Sy	/glɛːv/	584	64.1	210.2
Difference +/-		+308	-2.0	-11.5
Penultimate Sy	/e/	130	64.8	228.6
Stressed Sy	/nwaːʁ/	523	66.5	234.0
Difference +/-		+393	+1.7	+5.4
Penultimate Sy	/ni/	262	64.8	234.3
Stressed Sy	/mɔːʁ/	389	65.9	231.1
Difference +/-		+127	+1.1	-3.2
Penultimate Sy	/mɑ̃/	270	63.9	214.8
Stressed Sy	/suːʁ/	491	61.3	226.3
Difference +/-		+221	-2.6	+11.5
Penultimate Sy	/tu/	178	65.2	257.3
Stressed Sy	/ʒuːʁ/	357	64.4	238.0
Difference +/-		+179	-0.8	-19.3
Penultimate Sy	/tu/	191	61.9	240.9
Stressed Sy	/ʒuːʁ/	462	63.1	216.5
Difference +/-		+271	+1.2	-24.4
Penultimate Sy	/də/	120	64.1	204.0
Stressed Sy	/glwaːʁ/	391	61.1	200.6
Difference +/-		+271	-3.0	-3.4
Penultimate Sy	/di/	118	61.9	226.5
Stressed Sy	/vwaːʁ/	489	61.1	197.3
Difference +/-		+371	-0.8	-29.2

Table 3: Duration, intensity, fundamental frequency values for Twi speaker 7 (female)

4. CONCLUSIONS

This investigation has shown, on the basis of the selected corpus and evidence from our acoustic data that the most important acoustic parameter in producing French prosodic final stress, by both native speakers and Twi learners of French, is duration which is consistent. Pitch difference (f0) and intensity are not the key determining factors in distinguishing prosodic stress differences between penultimate and final syllables. Intra syllabic phonemic comparison, suggests that French Speakers' (f0) performance rate is higher than the Twi Speakers in distinguishing stress differences between the two syllables. The intensity cue reveals that Twi Speakers' performance rate is very low compared to the French Speakers.

Further investigations will focus on identification performances for non-poetic corpus on prosodic stress contrasts since different tones, stress patterns could be used when reading poems compared to daily conversation. Data will also be based on more and same number of female and male native FS and TS.

5. REFERENCES

[1] Adu Manyah K. 2019. Non-native Speakers' Identification of Lexical Tone Contrasts. *19th ICPhS.* Melbourne. 2782-2786.

[2] Adu Manyah K. 2010. *Parlons Twi*. Langue et culture. Paris: l'Harmattan.

[3] Adubofour I. 2019. *Lexical Stress in Ghanaian English,* MPhil Thesis, KNUST, Kumasi.

[4] Aoyama K. 2007. Prosody in second language acquisition. *Language experience in second language speech learning. 281-297.*

[5] Astésano, C. et al. 2007. Structural influences on initial accent placement in French. *PhonlACog- the role of the initial accent in prosodic structuring in French- from phonology to speech processing. Language and speech 50 (Pt 3): 423-46.*

[6] Astésano. C. 2016. Accentuation et niveaux de constituance en français: enjeux phonologiques et psycholinguistiques. *Langue Française, Paris, Armand Colin.*

[7] Beaugendre F. et al. 2001, Accentuation boundaries in Dutch, French and Swedish. *Speech Communication 33(4):305-318.*

[8] Best C. T., Bohn P. O, Faber A. 2003. Cross Language perception of non-native vowels: Phonological and phonetic effects of listeners' native languages, *15th ICPhS.* Barcelona. 2889-2892.

[9] Conlen, M. 2016. *A Linguistic Comparison: Stresses-timed and syllable timed languages and their impact on their impact on second language acquisition*. Honors College Theses, 30.

[10] Di Cristo, A. 2008. Vers une modélisation de l'accentuation du français : Première partie. *Journal of French Language Studies.*

[11] Di Cristo, A. 2010. *Vers une modélisation de l'accentuation du français : Seconde partie* Cambridge, Cambridge University Press.

[12] Engel, H. 2017. *Accentuation et transfert en français langue seconde.* Université de Stockholm. URN: urn:nbn:se:su:diva-155916.

[13] Frost, D. 2011. Stress and cues to relative prominence in English and French: A perceptual study. *J. of the Int. Phonetic Assoc*. Vol. 41, No. 1, 67-84.

[14] Lunden, A. 2018. Explaining word-final stress lapse. *From Part I- Phonetic correlates and prominence distinctions. The study of Word Stress and Accent: theories, methods and data. 76-101.*

[15] Martin, P. 2006. Phonologies and Phonetics of French Prosody. *UFR linguistique Université Paris 7 Denis Diderot. ISCA Archive Speech Prosody 2006, Dresden.*

[16] McRoberts, G.W. 1995. The role of fundamental frequency in signalling linguistics stress and affect: Evidence of dissociation. *Perception & Psychophysics 57. 159-174.*

[17] Moyer, A. 2004. Age, accent and experience in second language acquisition. *Second language acquisition 7, Dublin : Trinity College.*

[18] Obin, N., Rodet X., Lacheret-Dujour A. 2008. Un modèle de durée des syllabes fondé sur les propriétés syllabiques intrinsèques et les variations locales de débit. *HAL Id: halshs-00334347, version 1.*

[19] Patalauskaitė, I. 2005. Comparaison de l'accentuation des deux langues: français et lithuanien. *VDU, ASU ir LEU iki/ until 2018.*

[20] Piat M., Fohr D., Illina I. 2008. Identification de l'origine des locuteurs non natifs en utilisant des paramètres prosodiques. *JEP'08, Avignon, France.*

[21] Quemoun, F. R. 2003. Le système accentuel du français et sa valeur stylistique. *Vol. 2. 405-418.*

[22] Hagberg, R. L. 2006. An autosegmental theory of stress. *SIL e-Books, 3, Dallas, SIL International.*

[23] Rietmolen, N. 2019. Neural signature of metrical stress processing in French. *Université Toulouse le Mirail- Toulouse II English. version 1.*

[24] Roosman L. M. 2009. Dutch word stress as pronounced by Indonesian students. *Wacana*, Vol. 11 No. 2: 241-256.

[25] Rossi, M. 1980. Le français, langue sans accent ? *L'accent en français contemporain, Studia Phonetica, 15 Ottawa : Didier, 13-52.*

[26] Schwab, S. 2013. Apprenants hispanophones de FLE et accentuation en français. *Travaux neuchâtelois de linguistique, 59, 71-86.*

[27] Schwab, S. et al. 2012. An acoustic study of penultimate accentuation in three varieties of French. *Laboratoire d'Analyse et de Traitement du Langage (LATL).*

[28] Schweitzer C., Dodane C. 2020. Description de l'accent en français : des premiers grammairiens aux premiers phonéticiens (XVIe-début du XXe siècle) *SHS Web Conferences. Volume 78, 7e Congrès Mondial de Linguistique Française.*

[29] So, C. K. & Best, C. T. 2014. Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories. *Studies in Second Language Acquisition, 36, 195-221.*

[30] Vaissière, J. 2002. Cross-linguistic prosodic transcription: French Vs English. *St. Petersburg State University Press, FLE, Vol. 1.*

[31] Verlaine, P. 1866. Cauchemar *Poèmes saturniens/ Eaux-fortes II Alph Lemerre Librairie-Éditeur.*

[32] Wioland, F. 1991. *Prononcer les mots du français. Des sons et des rythmes*. Paris. Hachette.

[33] Woehrling C., Boula P., Adda-Decker M. 2008. Aspects prosodiques du français parlé en Alsace, Belgique et Suisse. *LIMSI-CNRS, Orsay Cedex.*

[34] Xu, Y. 2004. Understanding tone from the perspective of Production and Perception. *Language and Linguistics*. 5.4: 757-797.

[35] Xu, Y. 1994. Production and perception of coarticulated tones, *J. Acoustical Society Of America*. 2240-2253.

Native-Language Phonotactic Processing in Bilinguals

Max R. Freeman[1] & Viorica Marian[2]

[1]St. John's University, [2]Northwestern University
freemanm@stjohns.edu, v-marian@northwestern.edu

ABSTRACT

Bilinguals are prone to competition from similar sounding words across their two languages (e.g., *strict* activates Spanish *estricto*). In three experiments, we relied on phonotactic constraints (i.e., rules for combining speech sounds) to examine how experience with multiple languages transforms the way in which speech sounds are processed. In Experiment 1, we found that Spanish-English bilinguals accessed the L1 Spanish vowel+s+consonant constraint (Spanish *estricto*) during L2 English processing (English *strict*) when performing a cross modal phonological priming lexical decision task. In Experiments 2 and 3, we found that bilinguals processed L2 auditory and/or visual inputs in line with the L1 constraint (*strict* as *estrict*) in a vowel detection task and a sound recognition task (visual world paradigm). We conclude that bilinguals access L1 phonotactic constraints during L2 processing and that L1 representations for sounds and words influence how L2 input is perceived.

Keywords: bilingualism, phonotactic constraints, phonology, visual world paradigm, speech perception

1. INTRODUCTION

What if we hear a sequence of phonemes that conflicts with how sounds are typically combined within our language? For example, a rule for combining speech sounds in Spanish is that a vowel must precede word-initial s+consonant clusters (s+c), as in *estricto* (English: "strict"). In English, however, s+c onsets with and without an initial vowel are abundant. For a native Spanish speaker learning English, s/he may experience competition from the Spanish vowel+s+consonant cluster (v+s+c) rule when speaking English. Herein lies the reason native Spanish speakers often produce English s+c words with a vowel at the onset, such as *estudy* and *espring* [26]. And, while prothesis (i.e., the addition of a vowel at a word's onset) is commonly observed in Spanish-English bilinguals' accented English speech, the current investigation examines whether bilinguals implement rules of their native language (L1) when processing their second language (L2). Do non-native speakers perceive speech differently than L1 speakers of a given language?

Bilingualism provides a unique tool to examine perception of non-native sounds that conflict with L1 rules. Interestingly, when monolinguals hear nonsense sounds that contrast with the rules of their language, they repair the sound sequences to conform to the rules [8,11,12,14]. For example, Spanish monolinguals repair the Spanish-like non-word *special* (/spesjal/) to *especial*, the latter conforming to Spanish's v+s+c rule [8,14]. Spanish-dominant bilinguals also repair sounds that do not align with their L1 when they are in an L1 testing environment [5,6]. Here, we examine how bilinguals process L2 sound sequences that conflict with L1 rules. Bilinguals have demonstrated parallel activation of the L1 when comprehending in the L2, across phonotactic-constraint [13], phonological-word [2,3,9,20], lexical [1,18], semantic [21], and syntactic levels [16,18]. The current experiments provide further evidence for parallel processing in bilinguals, with perception of L1 phonotactics during L2 processing.

Moreover, this investigation provides insight into the structure of acoustic space and the phonological system within the bilingual mind by characterizing new cross-linguistic interactions at the sub-lexical level. If bilinguals process L2 input (e.g., English s+c word or non-word: *strict*/*spelg*) in accordance with L1 phonotactic constraints (e.g., Spanish-like v+s+c word or non-word: *estrict*/*espelg*), then bilinguals might perceive an illusory vowel onset due to L1 constraints on how phonetic categories are represented.

The purpose of this investigation is twofold. First, we examine *whether* bilinguals access L1 phonotactic constraints during L2 comprehension. When examining between-language co-activation in bilinguals, previous studies have identified that bilinguals process auditory and visual input through a combination of bottom-up (e.g., *plum* activates Spanish *pluma* (English: *pen*) and top-down pathways (e.g., inhibitory control of the irrelevant language allows for lexical selection) [2,3,20]. As speech unfolds through time, for example, hearing the word *strict*, neighbors that share phonology become activated (e.g., within-language neighbor: *string*, between-language Spanish neighbor *estudio*). Second, we investigate *the extent to which* bilinguals

process L1 phonotactic constraints while listening to and/or reading the L2. For example, a Spanish-English bilingual hearing and/or reading *strict* may process it as *estrict*, since their L1 (Spanish) contains the v+s+c rule. We tested the following two hypotheses: (1) Bilinguals experience L1 cross-linguistic influences during L2 sub-lexical processing, and (2) Bilinguals access L1 sounds during L2 audio-visual and visual-only (orthographic) processing.

2. EXPERIMENT 1

Learning a new set of phonotactic constraints is challenging for L2 learners, especially when an L2 phonological structure is not present in the L1 [4]. For example, in Spanish, s+cs cannot exist at word onsets and a vowel, usually 'e' must be added (e.g., English *strict,* Spanish *estricto*). In language production, native-Spanish speakers may at times produce English words such as *estudy* ("study"), adding an 'e' onset [26], which suggests that bilinguals access and apply Spanish constraints when speaking English. During receptive language processing, Spanish monolinguals report hearing the 'e' onset when primed with a Spanish word that has the 'e' removed (e.g., *stricto*) [14]. For bilinguals, when hearing *strict,* they may activate phonological cohorts that overlap with Spanish through phonotactic constraints and phonological form (e.g., *estándar*/standard) and potentially even cohorts that overlap with Spanish through phonotactic constraints only (e.g., *edad*/age) through cross-linguistic activation. Therefore, it was hypothesized that Spanish-English bilinguals accessed Spanish (L1) phonotactic constraints during English (L2) comprehension.

2.1. Method

In Experiment 1, bilingual Spanish-English (n=22) adult participants were tested. Proficiency was measured with *Language Experience and Proficiency Questionnaire* [19]. Bilinguals acquired English at age 6 or later and were exposed to Spanish a minimum 30% daily. In a cross-modal phonological priming lexical decision task, English primes included words with 'st' and 'sp' onsets, and fillers, controlled on lexical characteristics (e.g., frequency). Lexical decision targets included English words, or English-like non-words with 'es' and 'e' onsets, control non-words, and fillers. Participants heard English primes with initial s+c (*strict*) and controls without initial s+c (*workers*). Immediately after hearing the prime, participants performed a lexical decision on visual targets: English-like 'es' non-words (*estomb*), 'e' non-words (*entaty*), non-word controls (*atters*), and English word controls (*rising*). If English primes containing s+c onsets activated Spanish phonological rules, then faster responses to 'e' onset target non-words were expected.

2.2. Results

A 2(prime: s+c, control) x 4(target: 'es' non-word, 'e' non-word, non-word control, word control) repeated-measures ANOVA was conducted on lexical decision targets. There was a main effect of target condition, $F(3,129)=16.02$, $p<.001$, $\eta_p^2=.27$. Bonferroni-corrected pairwise post hoc comparisons indicated that bilinguals were faster to respond to 'es' overlap non-word targets (*estomb*) *and* 'e' non-word targets (*entaty*) preceded by s+c primes (*strict*; M = 876.33ms, SE = 61.85; M = 881.14ms, SE = 62.67, respectively) than to non-word controls preceded by s+c (M = 944.39ms, SE = 72.11), $t(21) = -4.63$, $p < 0.001$; $t(21) = -3.56$, $p < 0.01$, respectively. The results demonstrate significant effects of Spanish phonotactic constraint activation during English comprehension. Bilinguals demonstrated faster reaction times, relative to control conditions, to 'es' overlap non-words and 'e' overlap non-words when primed with s+c onset words.

2.3. Discussion

Experiment 1's findings demonstrate that during single-language comprehension, bilinguals access phonotactic constraints of the other language (Spanish) to comprehend words in the relevant language (English). Bilinguals were faster to respond to 'e' onset non-words than to controls without 'e' onsets when primed with s+c words, suggesting that bilinguals co-activated English and Spanish lexica containing words with 'e' onsets. We extend previous findings examining cross-linguistic phonological access in bilinguals [2,3] to suggest that phonotactic constraints of the non-target language, in addition to phonological representations, are accessed during comprehension. Experiment 1 established that bilinguals access L1 phonotactic constraints during L2 processing. Experiments 2 and 3 examine the extent to which parallel processing of phonotactic constraints occurs.

3. EXPERIMENT 2

To follow up on the findings of Experiment 1, as well as on the findings of Carlson et al. [6], Freeman et al. [13], Lentz and Kager [17], and Weber and Cutler [25], we examined the extent to which bilinguals engaged parallel processing of L1 phonotactic constraints during L2 comprehension. Our objective was to determine whether bilinguals *perceived* L1 phonotactics when hearing L2 sound sequences

(auditory input). Previous evidence demonstrates that bilinguals perceptually repair L2 syllable sequences that are illicit in the L1 to conform to the L1 [25]. The hypothesis was that Spanish-English bilinguals relied on parallel processing to perceive English and English-like sounds as Spanish-like, assimilating English (L2) sound sequences to Spanish (L1) phonotactic constraints.

3.1. Method

Participants included a new group of 26 Spanish-English bilinguals (native Spanish speakers), ages 18-34. The vowel detection task measured perception of the 'e' onset in s+c words and non-words. Perception of this phonotactic constraint was examined in the presence of s+c that conflicted with the Spanish 'e' onset vowel rule. Accuracy and reaction times to identifying whether a vowel was present were measured. We used a mix of words and non-words that conflicted with the 'e' onset constraint. With a non-word, participants may have been more likely to report hearing a vowel, as lexical status, and therefore the language to which the non-word could belong, were unclear.

In the vowel detection task, we asked participants if they heard a vowel at the onset of English s+c words (*strict*) and English-like s+c non-words (*spelg*), as well as word and non-word controls (*can, nulse*). Bilinguals were expected to demonstrate differences in reaction times to s+c words and non-words relative to control words and English-like non-words.

3.2. Results

Two 2(onset type: s+c onset, control) by 2(lexical status: word, non-word) mixed effects logistic regression models were used to analyze accuracy and reaction time data. The accuracy model failed to converge due to ceiling effects; therefore, we only report reaction time data. There was a main effect of Onset Type, $\beta = -0.04$, $SE = 0.02$, $t = -2.00$, $p < 0.01$; and a marginal interaction between Lexical Status and Onset Type, $\beta = 0.06$, $SE = 0.02$, $t = 1.50$, $p = 0.07$. Bilinguals responded more slowly to non- words ($M = 1133.11$ms, $SE = 30.76$) than to words ($M = 1076.03$ms, $SE = 22.88$), $\beta = 0.05$, $SE = 0.02$, $t = 2.40$, $p = 0.02$. Bilinguals showed slower response times to s+c words ($M = 1112.85$ms, $SE = 30.29$) than to control words ($M = 1039.09$ms, $SE = 28.98$), $\beta = -0.07$, $SE = 0.03$, $t = -2.4$, $p = 0.03$. Bilinguals' response times to s+c non-words ($M = 1136.03$ms, $SE = 35.45$) relative to control non-words ($M = 1130.17$ms, $SE = 47.47$) were similar, $\beta = -0.01$, $SE = 0.03$, $t = -0.42$, $p = 0.10$. Therefore, bilinguals demonstrated slower response times to Spanish-conflicting words relative to control words, suggesting L1 cross-linguistic influence on L2 s+c words. This pattern did not hold true for Spanish-conflicting non-words relative to control non-words.

3.3. Discussion

In vowel detection, which is an explicit measure of vowel perception, Spanish-English bilinguals perceived an illusory 'e' onset when listening to L2 words that conflicted with the L1 v+s+c rule. An explanation as to why there were differential effects for words versus non-words on vowel detection is that with words, bilinguals recruited top-down perceptual knowledge of phonotactic constraints, as well as top-down lexical knowledge. This finding slightly contrasts with the stronger phonotactic effects for non-words in Experiment 1. Only non-words were used as visual targets in Experiment 1, and it appears that when the onset sound of a stimulus is the explicit focus (Experiment 2), a stronger perceptual representation exists for words than for non-words, likely due to differential recruitment of top-down processes. Interestingly, the result for words but not non-words suggests that Spanish-like phonotactic processing affected word learning in earlier stages of acquiring English, but it may no longer affect the learning of new s+c words in English for this population [10]. Experiments 1 and 2 established that bilinguals process L2 input in line with L1 phonotactics through audio-visual and auditory-only input. Experiment 3 further examines the extent to which bilinguals experience parallel processing of the L1 during L2 comprehension in the visual-only modality.

4. EXPERIMENT 3

When comprehending words in one language, words from the other, irrelevant language may be simultaneously accessed through parallel activation [18,20]. This cross-language interactivity within bilinguals is surprising given that individuals can only speak in one language at a time. In Experiment 3, we again examined whether bilinguals were influenced by L1 phonotactic constraints when processing the L2, but this time we further investigated the role of input modality. In Experiments 1 and 2, bilinguals processed the L2 according to L1 phonotactics audio-visually (Experiment 1) or auditorily (Experiment 2). In Experiment 3, the Spanish phonotactic violation occurred in the visual domain with eye-tracking to different words on a visual display (visual world paradigm). The visual target word was identified by hearing only the word onset (e.g., "Click on /s/" for *spa*), while viewing three other words as well (an e-onset competitor: *egg*, and two fillers: *work* and *can*).

It was hypothesized that activation of the Spanish (L1) phonotactic constraint could occur independently of auditory input, as parallel activation does not always rely on hearing words [7,15,24]. In addition, because bilinguals were tested in English, their L2, it was predicted that those with lower L2 proficiency would experience increased L1 phonotactic interference during L2 processing than bilinguals with higher L2 proficiency.

4.2. Method

Participants included 33 Spanish–English bilingual adults, ages 18-34. In the sound recognition task, Spanish-English bilinguals saw four words on a visual display while eye movements were tracked. Participants identified the target word by hearing its onset (i.e., "Click on /s/", target = *spa*). On critical trials, items included an English target word that conflicted with the Spanish phonotactic constraint of a vowel onset at the beginning of s+c words (*spa*) and a competitor word containing the presumably activated vowel onset (*egg*). Two filler words were also present that did not conflict or overlap with the Spanish constraint (*work* and *can*). If bilinguals accessed Spanish phonotactics during English comprehension, then more looks to 'e' onset competitors (*egg*) than fillers (*work*/*can*) were expected when presented with s+c onset targets.

4.2. Results

Growth-curve analyses (GCA; [22]) of fixation proportions were employed to examine the time course of phonotactic-constraint activation during visual word processing. A composite score was created for L2 proficiency comprising of objective (PPVT-3 standard score) and self-report measures (LEAP-Q averaged speaking, understanding, and reading proficiency ratings) and entered into the model. Visual fixations were analyzed from the auditory prompt onset until the point at which fixations to the target peaked, indicating final target selection, which was around 1100ms post-sound onset. The model revealed main effects of proficiency on the intercept term, $\beta = 0.69$, $SE = 0.37$, $t = -2.48$, $p = 0.01$ and on the quadratic term (i.e., the rise and fall rate of fixation proportions in the model curve), $\beta = 0.51$, $SE = 0.28$, $t = 2.53$, $p = 0.01$. There were also significant interactions of word type by proficiency on the intercept, $\beta = 0.32$, $SE = 0.12$, $t = 2.67$, $p < 0.0$, and on the quadratic terms, $\beta = 0.62$, $SE = 0.28$, $t = -2.51$, $p = 0.01$. The model demonstrated that bilinguals with lower L2 (English) proficiency produced a greater proportion of fixations to the "e"-onset word relative to filler words than did higher L2 proficiency bilinguals.

4.3. Discussion

Experiment 3 results suggest the Spanish phonotactic constraints influence bilinguals relatively early on during English comprehension, without hearing the conflicting word (in the visual modality). In addition, bilinguals with lower L2 proficiency were more likely to activate the L1 phonotactic constraint when viewing L2 words. In other words, decreased L2 proficiency results in increased L1 interference. Experiment 3 was unique in that, for the first time, evidence was found for phonotactic-constraint activation of the unintended language during visual word processing without auditory input of the constraint-conflicting structure (i.e., s+c; *spa*).

5. GENERAL DISCUSSION

The current investigation demonstrates that native-language phonotactic constraints influence how bilinguals processe their second language. The purpose of these experiments was to further characterize and understand the involvement of the L1 during L2 comprehension. Typically, as auditory input unfolds through the acoustic stream, bilinguals activate neighboring words within and between their languages [23]. As each phone is heard, neighboring words are eliminated that do not coincide with the input until the target representation is reached. This process of elimination explains how phonologically competing words are activated and suppressed. However, the findings of the current investigation suggest that when the L2 acoustic stream conflicts with L1 phonotactic constraints, or rules, then words that conform to this rule are activated as well.

Moreover, the findings that bilinguals access phonotactic constraints from the unintended language when receiving auditory and visual input (Experiments 1 and 2) and only visual input (Experiment 3) suggest that models of bilingual language activation should include phonotactic constraints as further evidence for the extent to which cross-linguistic structures can be activated. These results underscore the dynamic connections within the bilingual language system[27]. Bilinguals activate and perceive L1 sub-lexical structures, such as phonotactic constraints, during L2 processing. These results also contribute to understanding how the language system is organized in the bilingual brain.

6. REFERENCES

[1] Bartolotti, J., Marian, V. 2012. Language learning and control in monolinguals and bilinguals. *Cog. Sci.* *36*(6), 1129–47.

[2] Blumenfeld, H. K., Marian, V. 2007. Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Lang. Cog. Proc. 2*(5), 633-660.
[3] Blumenfeld, H. K., Marian, V. 2013 Parallel language activation and cognitive control during spoken word recognition in bilinguals. *J. Cog. Psy. 25*(5), 547-567.
[4] Broersma, M., Cutler, A. 2011. Competition dynamics of second-language listening. *Q. J. Exper. Psy. 64*(1), 74-95.
[5] Carlson, M. T. 2018. Making room for second language phonotactics: Effects of L2 learning and environment on first language speech perception. *Lang. Spch. 61*(4), 598-614.
[6] Carlson, M. T., Goldrick, M., Blasingame, M., Fink, A. 2016. Navigating conflicting phonotactic constraints in bilingual speech perception. *BLC, 19*(5), 939-954.
[7] Chabal, S., Marian, V. 2015. Speakers of different languages process the visual world differently. *JEP. Gen. 144*(3), 539.
[8] Cuetos, F., Hallé, P. A., Dominguez, A., Segui, J. 2011. Perception of prothetic /e/ in #sC utterances: Gating data. *Proc. 17th ICPhS* 540-543.
[9] Darcy, I., Park, H., Yang, C. L. 2015. Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learn. Indiv. Diffs. 40*, 63-72.
[10] Darcy, I., Thomas, T. 2019. When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Lang. Cog. 22*(5), 1141-1159.
[11] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Mehler, J. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *JEP. HPP. 25*(6), 1568–1578.
[12] Flege, J. E. 2003. Assessing constraints on second-language segmental production and perception. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities.* Mouton de Gruyter, 319–355.
[13] Freeman, M. R., Blumenfeld, H. K., Marian, V. 2016. Phonotactic constraints are activated across languages in bilinguals. *Fron. Psy. 7*(702).
[14] Hallé, P. A., Dominguez, A., Cuetos, F., Segui, J. 2008. Phonological mediation in visual masked priming: Evidence from phonotactic repair. *JEP. HPP. 34*, 177–192.
[15] Kaushanskaya, M., Marian, V. 2007. Non-target language recognition and interference: Evidence from eye-tracking and picture naming. *Lang. Learn. 57(1),* 119–163.
[16] Kootstra, G. J., Van Hell, J. G., Dijkstra, T. 2012. Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *BLC. 15*(04), 797-819.
[17] Lentz, T. O., Kager, R. W. 2015. Categorical phonotactic knowledge filters second language input, but probabilistic phonotactic knowledge can still be acquired. *Lang. Spch. 58*(3), 387-413.
[18] Linck, J. A., Hoshino, N., Kroll, J. F. 2008. Cross-language lexical processes and inhibitory control. *Ment. Lex. 3*(3), 349-374.
[19] Marian, V., Blumenfeld, H. K., Kaushanskaya, M. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *JSLHR. 50*(4), 940-967.
[20] Marian V, Spivey M. 2003. Competing activation in bilingual language processing: Within-and between-language competition. *BLC. 6*, 97–115.
[21] Martín, M. C., Macizo, P., Bajo, T. 2010. Time course of inhibitory processes in bilingual language processing. *Brit. J. Psy. 101*(4), 679–693.
[22] Mirman, D., Dixon, J. A., Magnuson, J. S. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *JML. 59*(4), 475-494.
[23] Shook, A., Marian, V. 2013. The bilingual language interaction network for comprehension of speech. *BLC. 16*(2), 304-324.
[24] Thierry, G., Wu, Y. J. 2007. Brain potentials reveal unconscious translation during foreign language comprehension. *Proc. Nat. Acad. Sci. 104,* 12530–12535.
[25] Weber, A., Cutler, A. 2004. Lexical competition in non-native spoken-word recognition. *JML. 50*, 1–25.
[26] Yavas, M., Someillan, M. 2005. Patterns of acquisition of /s/-clusters in Spanish-English bilinguals. *J. Multiling. Comm. Dis. 3*(1), 50-55.
[27] Marian, V. 2023. *The Power of Language: How the Codes We Use to Think, Speak, and Live Transform Our Minds.* New York: Dutton.

Mid vowel and nasal vowel production in young adult French speakers on Mayotte Island in the Indian Ocean

Miki Mori

CUFR Mayotte, UMR 5267 CNRS Praxiling
miki.mori@univ-mayotte.fr

ABSTRACT

Phonetic and phonological variation in French spoken around the world has been well documented over the last few decades thanks to the Phonology of Contemporary French research project. However, little is known of the variety spoken on Mayotte Island, which was recently made a French department. This is despite its rich linguistic landscape, with two distinct languages: Shimaore, a Sabaki-Bantu language and Kibushi, an Austronesian language. For example, whereas French has upwards of 13 oral vowels and 4 nasal vowels, Shimaore has 5 oral vowels and 3 nasal vowels: a, i, u, e, o, ĩ, ẽ, and õ. It is unclear how Shimaore speakers adapt the complex and varying French vowel system, particularly for the mid vowels and considering the 'loi de position.' This study looks at mid vowels and nasal vowels spoken by eight speakers in Mayotte. It looks at the vowel space, overlap and distance.

Keywords: loi de position, French, Mayotte, vowel space, nasals

1. INTRODUCTION

Thanks to the Phonology of Contemporary French research project, phonetic and phonological variation in French spoken around the world has been well documented over the past few decades (see [1]). From prosody to phonetics, the ways in which French speakers differ by variety is better understood, including aspects concerning language contact. However, there are still French speaking regions yet to be explored. Mayotte Island, which was integrated in 2011 as the nation's most recent *department*, is one location where little to nothing is known about its spoken French. Located in the Mozambique Channel in the Indian Ocean, Mayotte is a small tropical island with a rich linguistic and cultural landscape. Two principal local languages exist: Shimaore, a Sabaki-Bantu language and Kibushi, an Austronesian language. Dialects of both languages are observed and are often distinguished by village and region. Shimaore and its dialects from neighbouring Comorian islands is the most dominant, whereas only 15% of locals speak a variety of Kibushi [2].

French is the language of instruction and administration according to law. In practice, a sizable portion of the population does not speak or understand French [3, 4]. With younger generations attending school from 3 years of age, this proportion will certainly decrease. Much discussion has been made over how to label the French language in Mayotte: Is it a second language or a foreign language? Most agree that it is the language of schooling, in which first contact with it often comes from within the classroom walls.

Indeed, any well-attuned ear can hear unique aspects of the French spoken on the island, be it related to syntax, vocabulary, prosody, or phonetics. Yet, little to nothing is known about Maore French, particularly its pronunciation. This despite the language contact with Shimaore and Kibushi resulting in potential phonological and phonetic variation. For example, French has upwards of 13 oral vowels and 4 nasal vowels: a, i, u, e, o, y, ɛ, ø, œ, ɔ, ɑ, ɛː, ə, ɛ̃, ɔ̃, ɑ̃, and œ̃. However, Shimaore has 5 oral vowels and 3 nasal vowels: a, i, u, e, o, ĩ, ẽ, and õ. It is unclear how Shimaore speakers adapt the complex and varying French vowel system, particularly for the mid vowels and considering the 'loi de position' in which mid vowels are open in a closed syllable position and closed when in an open syllable [5]. Thus, it may be interesting to explore the vowel pairs /e/~/ɛ/, /ø/~/œ/, and /o/~/ɔ/ (E, Ø and O, respectively) [6] when used by Maore. In addition, given the presence of some nasal vowels in Shimaore, it remains unclear how the nasal vowels from the two languages map onto each other.

This study looks at mid vowels and nasal vowels in the spoken French of Mayotte. It particularly looks at the vowel space as well as overlap and distance.

2. METHOD

The project was inspired by the Phonology of Contemporary French research project. It adopted their protocol for data elicitation, including the reading of a wordlist, paragraphs, as well as a semi-directive interview. Eight participants, of which four women, were recorded in 1m^2 soundproof booth at the Centre universitaire de formation et de recherche de Mayotte using an Apex 435B condenser microphone, Presonus audiobox with a sampling rate of 44,100 Hz. Participants were undergraduate students who had gone through the French public education system on the island. All but one participant spoke Shimaore or a variety of it. Speaker 00M25 only spoken French.

Participation was voluntary.

Analyses of vowel space and Euclidean distance were done using R [7] with the help of PhonR [8]. Norm was used to normalized formants by Lobanov methods [9]. For overlap, Spectral Overlap Assessment Metric (SOAM) measurements were obtained using Wassink's VOIS3D [10].

The following PFC words were analysed:

/e/~/ɛ/ aspect, baignoire, beauté, bouleverser, bêtement, cinquième, des genêts, des jeunets, dégeler, déjeuner, épais, épais, épier, épée, étrier, étriller ex-femme, ex-mari, explosion, extraordinaire, fêtard, fête, fêter, infect, liège, lierre, miette, millionnaire, mouette, muette, médecin, niais, nier, nièce, piquais, piquer, piquet, piqué, pécheur, pêcheur, quatrième, relier, rhinocéros, scier, trouer

/o/~/ɔ/ agneau, beauté, botté, explosion, extraordinaire, gnôle, millionnaire, paume, pomme, rauque, rhinocéros, roc, socialisme

/ø/~/œ/ creuse, creux, des jeunets, déjeuner, feutre, jeune, jeûne, meurtre, peuple, pécheur, pêcheur

/ɛ̃/ brin, cinquième, infect, influence, intact, médecin,

/ɔ̃/ compagne, compagnie, blond, explosion, million,

/ɑ̃/ bêtement, blanc, influence, vous prendriez

/œ̃/ brun

3. RESULTS

As can be seen in Figure 1, participants tend to reduce vowel spaces for the /o/- /ɔ/ pairs while maintaining differentiation for the anterior /e/~/ɛ/ pairs, with the latter sound having a larger variation space. As for /ø/~/œ/, the two vowels tend to be similar in terms of formants, with /œ/ having more variation in height realization. The only significant difference found for formants and vowel pair was for F1 in the /e/~/ɛ/ pair ($\chi 2 = 473.06$, df = 344, p-value = 4.56e-06)

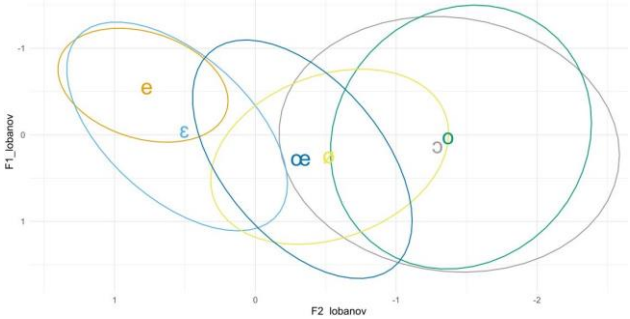


Figure 1: Lobanov-transformed oral mid-vowel space for all participants

Individual differences can be noted in Figure 2, particularly for /e/~/ɛ/ and /ø/~/œ/. Some participants distinguish between these pairs, whereas others tend to merge them. While varying slightly in degree, participants do not strongly distinguish between the open and closed O. No significant differences were found per individual.

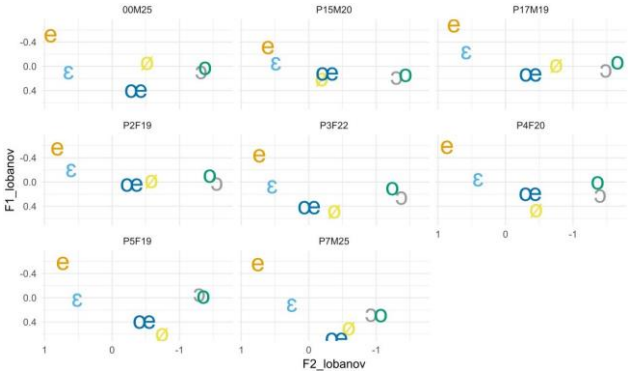


Figure 2: Lobanov-transformed oral mid-vowel plots per participant

In fact, looking at measures of distance and overlap, we can see that for /o/~/ɔ/, Euclidean distances are relatively close and SOAM measures indicate a near overlap of vowel space. On the other hand, for /e/~/ɛ/, Euclidean distances and SOAM measures show a more distinct vowel space for many of the participants, such as 00M25 and P3F22. Participants P7M25 and P4F20, however, have the vowels that share vowel space. The same can be said for /ø/~/œ/, where some participants pronounce these vowels similarly (P7M25 & P4F20) whereas others maintain a difference (00M25 & P17M19).

Vowel pair	Participant	Euclidean distance	SOAM score
/o/- /ɔ/	00M25	24.18	.850
	P2F19	56.86	.974
	P3F22	65.03	.973
	P4F20	42.87	.926
	P5F19	29.33	.820
	P7M25	50.19	.911
	P15M20	50.05	.870
	P17M19	76.93	.837
/e/~/ɛ/	00M25	139.96	.417
	P2F19	137.81	.690
	P3F22	137.21	.641
	P4F20	262.09	.742
	P5F19	156.84	.897
	P7M25	203.87	1.00
	P15M20	53.48	.970
	P17M19	106.73	.671
/ø/~/œ/	00M25	94.27	.532
	P2F19	150.55	.692
	P3F22	161.24	.911
	P4F20	69.03	.912
	P5F19	137.46	.932
	P7M25	70.12	1.00

	P15M20	32.08	.862
	P17M19	168.57	.362

Table 1: Euclidean distances and SOAM scores per participant and vowel pair

Figure 3 offers a visualization of SOAM scores to demonstrate the variation in overlap among the three vowel pairs and among speakers. As can be seen the greatest difference for /e/~/ɛ/ is seen with 00M25 and for /ø/~/œ/, P17M19 distinguishes them the most. P7M25 tends to merge all three vowel pairs.

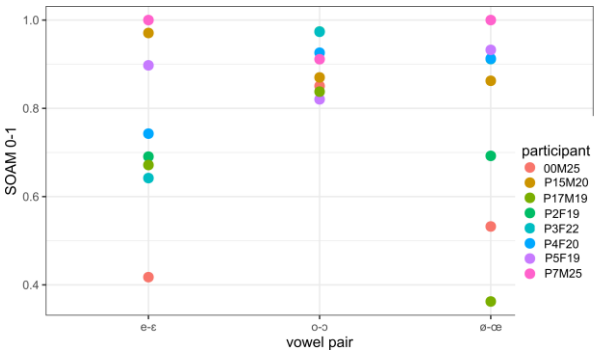


Figure 3: SOAM scores by vowel pair and participant

As for the nasal vowels Figure 4 illustrates the overall averages. As can be seen, there is much overlap in the vowel spaces as well as a nesting tendency, in which /œ̃/ production occurs within the vowel spaces of the other three nasals. /ɔ̃/ has the most variation in its elliptic space, occurring in posterior, anterior, open, and closed positions.

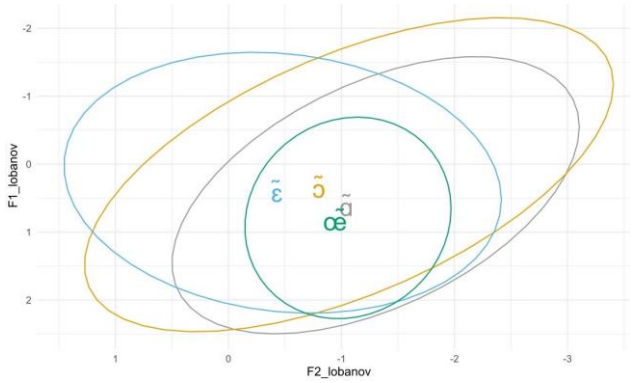


Figure 4: Lobanov-transformed nasal vowel plots averages

As for individual differences, Figure 5 shows variation among the eight speakers. For example, P2F19, P3F22 and P4F20 (middle row) have distributions that are close, dispersed, and overlapping, respectively. Curiously, some participants have an /ɔ̃/ that is anterior and/or open. The same can be said of the height of some participants' /ɑ̃/, which are closed rather than open.



Figure 5: Lobanov-transformed nasal vowel plots per participant

4. DISCUSSION

Looking at the overall group and individual differences concerning the three mid-vowel pairs /e/~/ɛ/, /ø/~/œ/, and /o/~/ɔ/, it is apparent that this latter vowel pair is merging for the speakers, such that little to no distinction is made between the two. The same can be said for /ø/~/œ/, except that variation among speakers reveals that some merge the sounds more than others. /e/~/ɛ/ appear to mostly maintain distinction in open and closed syllables per 'loi de position.' Given Mayotte's unique linguistic landscape, it is difficult to directly compare these results with other French speaking regions, including in the Indian Ocean, such as the fact that there are no Creoles on the island, unlike La Réunion Island or the Seychelles Islands. Nevertheless, a recent case study of Mauritius French showed that speakers maintain distinction between the mid-vowels in question according to the 'loi de position' [11]. A case study on the French spoken in Bangui in the Central African Republic, although it is noted that other speakers tend to not make a distinction between the open-mid and closed-mid vowels [12]. As for the nasal vowels, it appears that speakers vary in their pronunciation of them, with some nearly completely merging the four. It cannot be clearly stated that Maore French has three or four nasals, as is the case of La Réunion Island and Guadeloupe Island, respectively [13].

Considering that participants read from a list, it is assumed that the controlled environment would have elicited careful, enunciated speech from speakers. That is, the vowel spaces would be as large as possible. Given this assumption, it appears that some vowels in question are not distinguished for speakers and that the 'loi de position' is not always adhered to. Much is yet needed to understand this phenomenon, including a look at less controlled speech, speakers with other demographic profiles, as well as the other vowels in the French language inventory.

5. REFERENCES

[1] Detey, S., Durand, J., Laks, B., Lyche, C. (eds). 2016. *Varieties of spoken French*. Oxford University Press.

[2] Jamet, R. 2016. *Dictionnaire kibushi-français. Dzoumogné*. Mayotte: Association SHIME.

[3] Insee. 2007. Mayotte : les langues en 2007. https://www.insee.fr/fr/statistiques/2569783#consulter

[4] Insee Mayotte Infos. 2014. Les difficultés face à l'écrit en langue française. Quatre jeunes sur dix en grande difficulté à l'écrit à Mayotte. *Population* 70, 1-6.

[5] Detey, S., Racine, I., Kawaguchi, Y. Eychenne, J. (eds) 2016. La pronunciation du français dans le monde. Du natif à l'apprenant. CLE International.

[6] Scherrer, Y., Boula de Mareüil, P., Jean-Philippe Goldman, J.P. 2015. Crowdsourced mapping of pronunciation variants in European French *Proc. 18th ICPhS* Glasgow, 1-5.

[7] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[8] McCloy, D. R. 2016. phonR: tools for phoneticians and phonologists. R package version 1.0-7.

[9] Norm, the Vowel Normalization and Plotting Suite http://lingtools.uoregon.edu/norm/norm1_help.php

[10] Wassink, A. B. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three varieties, *Journal of the Acoustical Society of America,* 119(4), pp 2334-2350.

[11] Ledegen, G., Lyche, C. 2016. French in Mauritius. A speaker from Quatre-Bornes. In S. Detey, J. Durand, B. Laks, C. Lyche. (eds). 2016. *Varieties of spoken French*. Oxford University Press, 256-267.

[12] Bordal Steien, G., Boutin, A. B., Beyom, R. 2016. French in the Central African Republic. A speaker from Bangui. In S. Detey, J. Durand, B. Laks, C. Lyche. (eds). 2016. *Varieties of spoken French*. Oxford University Press, 236-246.

[13] Pustka, E., Ledegen, G. 2016. Le français dans les DROM. In S. Detey, I. Racine, Y. Kawaguchi & J. Eychenne (eds), *La pronunciation du français dans le monde. Du natif à l'apprenant*. CLE International, 69-72.

THE RELATIONSHIP BETWEEN L2 ACQUISITION AND L1 ATTRITION IN THE PHONETIC DOMAIN

Lisa Kornder[1], Ulrich Reubold[1], Robert Mayr[2], Ineke Mennen[1]

[1]University of Graz, [2]Cardiff Metropolitan University
lisa.kornder@uni-graz.at, johann.reubold@uni-graz.at, rmayr@cardiffmet.ac.uk, ineke.mennen@uni-graz.at

ABSTRACT

This study explored the relationship between second language (L2) acquisition and first language (L1) attrition of speech in late consecutive English-German bilinguals living in an L2 country (Austria). We acoustically examined the bilinguals' realizations of the lax vowel /ɪ/ in both of their languages and compared their productions to monolingual productions of English and German, respectively. Despite some individual-speaker variation, results showed that those speakers who produced the L2 target vowel in a native-like manner, as manifested in a tensing of /ɪ/ towards /i:/, showed a tendency towards tensing the /ɪ/ vowel also in their L1 English, suggesting L1 attrition. This indicates that successful L2 acquisition of a pronunciation feature might entail attrition of the corresponding feature in speakers' L1.

Keywords: vowels, first language attrition, second language acquisition, Austrian German, English.

1. INTRODUCTION

A number of studies show that the first language (L1) phonetic/phonological system may change under the influence of a second language (L2) system, which often results in modifications of L1 pronunciation features in the direction of L2 norms [1–4]. Changes in L1 pronunciation experienced by late bilinguals who are long-term residents in an L2 environment are commonly referred to as *phonetic attrition* [5].

While the occurrence of L2-induced changes in speakers' native pronunciation system have been attested for segmental [1,2] and prosodic [3,4] variables, findings concerning the actual *relationship* between L2 acquisition and L1 attrition of speech are still rather inconclusive. We might expect that the longer a speaker has lived in an L2 country, the more likely their L1 pronunciation system is to be affected by L2 influences, given that length of residence is often associated with increased L2 experience and proficiency. While it has indeed been shown that advanced L2 *proficiency* might cause L1 pronunciation features to move closer to the norms of an L2 in experienced bilinguals living in a migration setting [6], length of residence has so far not been established as a reliable variable predicting attrition effects in bilinguals [7]. Furthermore, L2 influences on L1 pronunciation features are not ultimately limited to advanced, long-term immersed bilinguals, but also occur in inexperienced speakers with short-term L2 exposure at the onset of L2 learning [8,9] as well as in intermediate L2 learners [10]. Such L1 modifications are, however, not likely to represent long-term changes in a speaker's pronunciation system, but may be either completely or partially reversed [11,12] and do therefore not represent instances of attrition.

Despite the steady proliferation of studies exploring L1 attrition in the phonetic domain, only few studies *compare* segmental speech production in the L1 *and* L2 of late bilinguals immersed in an L2 setting [1,2;13−15], with only two studies to date examining L1 and L2 vowel productions [1,2]. However, to improve our understanding of how L2 acquisition of pronunciation and potential modifications of L1 pronunciation are related, we need more studies that closely investigate speech production in both languages.

The present study aims to shed light on the question of how L2 acquisition and L1 attrition are related by directly comparing L1 and L2 realisations of a specific vowel pair, i.e., the tense-lax pair /i:/–/ɪ/, in a group of L1 English speakers who acquired L2 Austrian German in adulthood upon migrating to Austria. Their L1 and L2 vowel productions were compared to monolingual productions of English and Austrian German, respectively, with the aim to identify if and to what extent the bilinguals' L2 productions have reached native (monolingual) Austrian German norms and how this has affected their L1 English.

2. METHODS

2.1. Speakers

The main group of speakers included English-Austrian German bilinguals (*BIL*, $N = 8$, 4 females, 4 males), who were raised as monolingual speakers of Standard Southern British English (SSBE) and moved to Austria in adulthood where they acquired Austrian German as an L2. Two monolingual groups served as controls: these were monolingual speakers of SSBE (i.e., MONO English) and Austrian German

(i.e., MONO German) residing in England or Austria and who have never lived outside their home counties (N = 8, 4 females, 4 males per group). They reported no more than high-school level knowledge of other languages.

The bilinguals had all participated in an earlier global accent rating experiment [16] in which their L1 speech was rated by monolingual SSBE listeners in terms of perceived nativeness. The eight bilinguals selected for the present study were perceived as moderately accented in their L1 in the rating study, suggesting that listeners perceived deviations from the SSBE norm in the bilinguals' L1 pronunciations. As previous studies show that bilinguals who are rated as sounding non-native in their L1 do not *inevitably* also show divergences of L1 segments [17], modifications of the bilinguals' L1 vowels were not expected from the start in the present investigation.

2.1. Speech Materials

The selection of materials was based on the nativeness ratings in [16] which were accompanied by listeners' comments on the pronunciation features that made them believe that a speech sample sounded non-native. The most frequently mentioned segmental comment had been on /ɪ/ (Wells' [18] standard lexical set: KIT), which was described as being too /iː/-like (i.e., approaching the standard lexical set FLEECE). A known feature of Austrian Standard German (in opposition to German Standard German) is the neutralization of quality, but not quantity of high vowel pairs [19], irrespective of surrounding consonantal context [20]. This suggests a difference in the implementation of the tensity contrast between Austrian and German Standard German, with the latter being very similar to the same contrast in English, which maintains a tense-lax distinction [21, 22]. To investigate the acoustic properties that might cause the auditory perception of /ɪ/ taking on the quality of /iː/, and in order to compare tense and lax high front unrounded vowels acoustically in both languages, we analysed English and German words containing both vowels. We chose instances of /ɪ/ occurring in stressed syllables in 30 English and 10 German words that were recorded embedded in the carrier sentences "Say [target word] again" and "We said [target word] together" (target words were: *bit, bitter, chill, filled, fin, hid, his, hiss, Jill, lip, miss, in, myth, Phil, picks, pigs, pill, rich, ship, sick, sin, sinking, sit, thick, thin, things, thinking, wick, will,* and *winter*) in the English set and in the carrier sentence "Ich habe [target word] gesagt" (target words were: *Knittelfeld, Krippe, Lippe, Mitte, Stiftung, Winter, bitter, Innkreis, Ritter,* and *bitte*) for the German set.

Additionally, other materials from the same recording sessions were used in order to obtain words with the tense, high front unrounded vowel /iː/, and the tense, low back unrounded vowel /ɑː/, both occurring as syllable nuclei of the stressed syllables. These vowels were later used as anchor vowels (see section 2.3).

2.2. Recording Procedure and Pre-processing

Bilinguals recorded both sets, with the English set always occurring before the German one, while monolinguals only recorded a single set in their respective language. Each word was read aloud twice in each carrier sentence, resulting in four repetitions per vowel, context and speaker. Speech recordings took place online due to Covid-19 restrictions, using the software Wikispeech [23]. This system allows participants to complete unsupervised online recordings, using their own recording devices. Participants were instructed to avoid noisy environments and adjust recording levels after listening to the first few test items to avoid poor audio quality. All recordings were prepared for further analyses by normalizing maximal amplitude and sampling size (16kHz). The texts of the prompts and the audio data were then sent to BAS WebServices [24], a service for automatic forced alignment; the resulting segment boundaries were manually corrected. These corrections and the analyses were conducted in EMU-SDMS [25]. Formants were calculated with *Praat*'s [26] *To Formant (burg)* method [27], with standard settings.

2.3. Data analysis

We measured F1 and F2 in /ɪ/, /iː/ and /ɑː/ at the temporal midpoints of the vowels. A method called orthogonal projection [28,29] was used, which normalizes the vowel spaces by using the vowels /iː/ and /ɑː/ as anchor vowels that represent the upper left and lowest right corners of each speaker's vowel space, allowing for normalization of the vowel spaces, and for a quantification of the closeness of the lax high front vowel to the tense high front vowel. The closeness of /iː/ to /ɪ/ was then determined by calculating the relative position of every /ɪ/ vowel between the anchor vowels /iː/ and /ɑː/ using the orthogonal projection parameter *op*, the so-called orthogonal projection ratio using (1):

$$op(\vec{\imath}) = 1 - 2\frac{(\vec{\imath}-\vec{c}_{i:}) \odot (\vec{c}_{i:}-\vec{c}_{\alpha:})}{(\vec{c}_{i:}-\vec{c}_{\alpha:}) \odot (\vec{c}_{i:}-\vec{c}_{\alpha:})}. \quad (1)$$

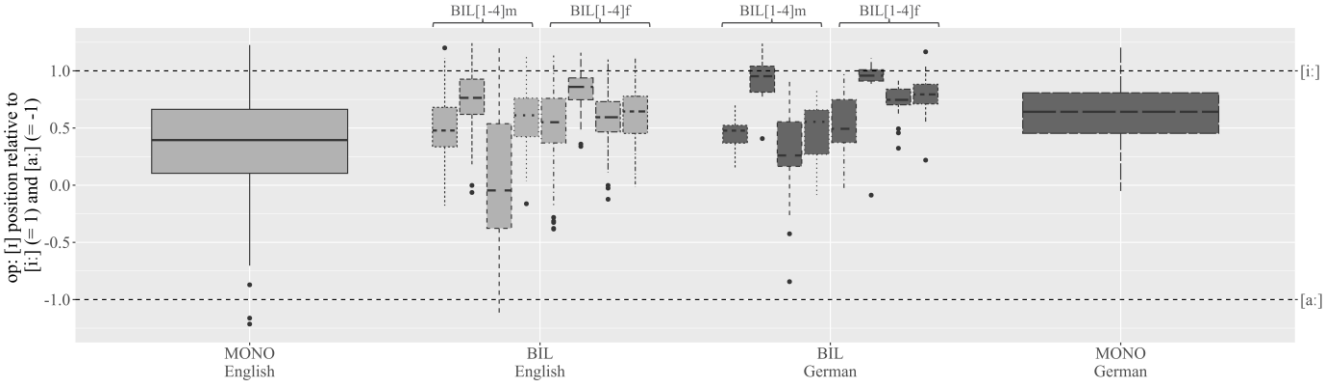


Figure 1: Distributions of the *op* values for [ɪ] in the different speaker groups. *Note: m=male; f=female.*

In a two-dimensional acoustic space (formed from the acoustic measurements for F1 and F2), $\vec{\imath}$ is the position of a given target vowel of the category [ɪ], $\vec{c}_{i:}$ and $\vec{c}_{\alpha:}$ are the centroids (means) of the two anchor vowels respectively, i.e., of [i:] and [ɑ:], and $\odot$ is the scalar (inner) product of two vectors. This means that a value of -1 was assigned to the formant values corresponding to the speaker-specific means of F1 and F2, respectively, of /ɑ:/, and a value of 1 was assigned to formant values corresponding to speaker-specific means of F1 and F2 of /i:/. The mid-point between the two would then be reflected by the value of 0 (which roughly corresponds to F1 and F2 of a prototypical schwa vowel).

2.4. Statistical analysis

We applied linear mixed models with *op* values as dependent variable, *Group/Language* (four levels: *MONO English*, *MONO German*, *BIL English*, *BIL German*) as fixed factor, and *Word* as random factor (including random slopes for the *Group/Language* combinations). We did this on a.) the group-level (i.e., comparing eight speakers per Group/Language level), and b.) on a bilingual-speaker-specific level (i.e., comparing eight *MONO English* and *MONO German* speakers with one *BIL English* and *BIL German* speaker, respectively). We added an *emmeans* [30] analysis as a post-hoc comparison and then calculated whether the mean *op* values per speaker in the bilingual group for English were correlated with the same speaker's German *op* values by means of a simple linear model.

3. RESULTS

Figure 1 shows the distributions of the *op* values for [ɪ] (that is, the relative positions of [ɪ] vowels between speaker-specific mean positions of [i:] and [ɑ:]) in English and German words produced by English and Austrian German monolinguals (*MONO English*, *MONO German*), respectively, and by English-German bilinguals in both languages (*BIL English*, *BIL German*). A linear mixed model revealed a significant main effect [$F(3, 26.1) = 31.4$, $p < 0.001$) for *Group/Language*. Estimated marginal means analyses showed significant differences between *MONO English* vs. *MONO German* ($p < 0.05$), *BIL German* ($p < 0.01$), and also *BIL English* ($p < 0.001$). There were no significant differences, though, between the *MONO German* values and *BIL German* or *BIL English*, or between *BIL German* and *BIL English*.

As can be seen in Figure 1, there are individual differences in the BIL speaker groups. The post-hoc analysis revealed that 5 out of 8 BIL speakers (BIL4m, BIL1f, BIL2f, BIL3f, BIL4f) showed no significant differences in their L2 German vowels compared to the MONO German speakers. That is, these BIL speakers showed a tendency towards tensing their L2 /ɪ/ vowel in the direction of /i:/, as shown in Figure 1. At the same time, the BIL's English productions were shown to be significantly different from MONO English (BIL4m, BIL2f, BIL3f, BIL4f: $p < 0.0001$; BIL1f: $p < 0.05$), i.e., the BIL's L1 /ɪ/ has also moved closer towards L2 /i:/ (Figure 1).

The analysis of the remaining three speakers (BIL1m, BIL2m, BIL3m) shows significant differences between both their L1 English (BIL1m: $p < 0.05$; BIL2m, BIL3m: $p < 0.0001$) and their L2 German (BIL1m, BIL2m: $p < 0.05$; BIL3m: $p < 0.01$) as compared to MONO English and MONO German, respectively. An inspection of Figure 1 reveals the following: Speaker BIL1m shows a tensing of L1 English /ɪ/ in the direction of German /i:/ and at the same time, he is closer to L1 English /ɪ/ in his L2 German. Speaker BIL2m, by contrast, shows a tendency to produce tense /i:/ in both English and German, but significantly overshoots the MONO German target in his L2. Finally, speaker BIL3m's L2 productions are closer to L1 English /ɪ/ while his L1

productions suggest that he overshoots the MONO English norm in the direction of /ɪ/.

According to a linear model, the English mean *op* values of the bilinguals are positively correlated with the same speakers' German mean *op* values (Figure 2: Adjusted R^2 = 0.51, F[1 ,6] = 8.3, p < 0.05), although the relationship is not 1-to-1 (slope of the model: 0.61).

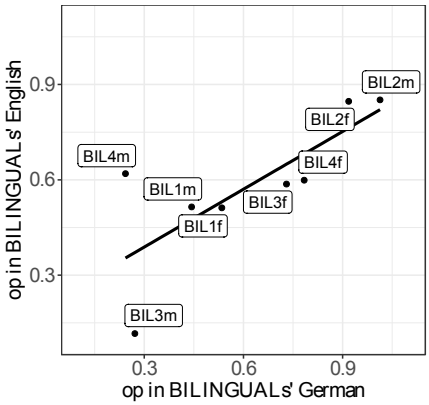


Figure 2: Speaker-specific mean *op* values in BIL English as a function of their BIL German mean *op* values.

4. DISCUSSION

In order to more closely characterize the relationship between L2 acquisition and L1 attrition of speech, we examined how late consecutive bilinguals, living in Austria, produced the tense-lax pair /iː/–/ɪ/ in their L1 SSBE and their L2 Austrian German as compared to monolingual speakers of SSBE and Austrian German, respectively.

In terms of the monolingual productions, our data revealed that the Austrian German speakers produced significantly less of a quality contrast between /iː/ and /ɪ/ compared to SSBE speakers. That is, in line with what has previously been reported for Austrian German [19] and English [21,22], Austrian German speakers' productions of /ɪ/ showed a tensing in the direction of /iː/, although there was no evidence of a merger.

The individual speaker analyses revealed that five out of eight bilinguals produced a native-like tense /iː/ in their L2 German, which suggests that these bilinguals have successfully acquired native (monolingual) Austrian German norms in their L2. At the same time, their L1 English /ɪ/ – similar to their German /ɪ/ – has moved closer towards L2 /iː/. This indicates that their L1 category has assimilated to the L2. Consequently, these five speakers' vowels have *approximated L2 Austrian German norms in both the L1 and the L2*. Similar assimilatory patterns have been observed in [1] and [2], where the bilinguals under investigation produced some of their L1 Austrian German [1] or Dutch [2] vowels acoustically closer to the L2 English target.

The analysis of the remaining three speakers reveals a different, less systematic pattern: Their vowel productions suggest that they *deviate from both L1 and L2* norms, i.e., these speakers show signs of attrition in their L1 vowels while not achieving monolingual values in their L2. In some of the speakers, we observed a tendency to *overshoot* either the L1 or the L2 monolingual target, i.e., their vowels were either too lax for English or too tense for German. Speaker BIL2m, for instance, produced L1 English vowels closer to German /iː/ (suggesting attrition) while at the same time significantly overshooting L2 German monolingual values (i.e., his L2 productions were even more /iː/-like). Speaker BIL3m, by contrast, was observed to produce English-like /ɪ/-vowels in his L2 German (showing an L1-to-L2 influence) and to overshoot monolingual English values in his L1, i.e., producing more /ɪ/-like vowels than the monolingual speakers. This might reflect an attempt to keep the L1 and L2 vowels apart, assuming that – as suggested by the Speech Learning Model [31] – a bilingual's phonetic categories exist in a shared phonetic space.

Taken together, these findings indicate that successfully acquiring an L2 pronunciation feature, in this case the tensing of /ɪ/, might entail non-native realizations of the corresponding feature in the L1. This does, however, not seem to be the case in all bilinguals, with some speakers showing a tendency to overshoot monolingual targets in their L1 or L2, resulting in non-native-like productions in both languages.

5. CONCLUSION

The findings of the present study suggest that the ability to produce an L2 pronunciation feature in a native-like fashion – as measured against monolingual norms – might impede successful L1 maintenance. In order to shed light on how individual speaker variation in L2 acquisition and L1 attrition of pronunciation can be explained, future studies may want to explore the relative contribution of predictor variables such as length of residence or age of arrival as well as the relation between nativeness ratings and the reduced distance between [ɪ] and [iː].

6. ACKNOWLEDGEMENTS

We would like to thank the Austrian Science Fund (FWF) for their financial support under grant number P33007-G.

7. REFERENCES

[1] Kornder, L., Mennen, I. 2020. Longitudinal developments in bilingual second language acquisition and first language attrition of speech: the case of Arnold Schwarzenegger. *Languages* 6, 1–25.

[2] Mayr, R., Price, S., Mennen, I. 2012. First language attrition in the speech of Dutch-English bilinguals: the case of monozygotic twin sisters. *Biling.: Lang. Cogn.* 15, 687–700.

[3] Mennen, I., Reubold, U., Endes, K., Mayr, R. 2022. Plasticity of native intonation in the L1 of English migrants to Austria. *Languages* 7, 1–27.

[4] de Leeuw, E., Mennen, I., Scobbie, J.M. 2012. Singing a different tune in your native language: First language attrition of prosody. *Int. J. Biling.* 16, 101–116.

[5] Köpke, B., Schmid, M.S. 2004. Language attrition: The next phase. In Schmid, M.S., Köpke, B., Keijzer, M., Weilemar, L. (eds), *First language attrition: Interdisciplinary perspectives on methodological issues.* John Benjamins, 1–43.

[6] Flege, J.E. 1987. The production of 'new' and 'similar' phones in a foreign language: evidence for the effect of equivalence classification. *J. Phon.* 15, 47–65.

[7] Hopp, H., Schmid, M.S. 2013. Perceived foreign accent in first language attrition and second language acquisition: the impact of age of acquisition and bilingualism. *Appl. Psycholinguist.* 34, 361–394.

[8] Chang, C.B. 2013. A novelty effect in phonetic drift of the native language. *J. Phon.* 41, 520–533.

[9] Kartushina, N., Hervais-Adelman, A., Frauenfelder, U.H., Golestani, N. 2016. Mutual influences between native and non-native vowels in production: evidence from short-term visual articulatory feedback training. *J. Phon.* 57, 21–39.

[10] Kupske, F.F. 2021. Destabilizing effects of L2 explicit pronunciation instruction on L1 speech: Voice Onset Time production by Brazilian intermediate users of English. *Gradus: Revista Brasileira de Fonologia de Laboratorio* 6, 32–49.

[11] Kartushina, N., Martin, C.D. 2019. Third-language learning affects bilinguals' production in both their native languages: A longitudinal study of dynamic changes in L1, L2 and L3 vowel production. *J. Phon.* 77, 1–21.

[12] Chang, C.B. 2019. Language change and linguistic inquiry in a world of multicompetence: Sustained phonetic drift and its implications for behavioral linguistic research. *J. Phon.* 74, 96–113.

[13] Stoehr, A., Benders, T., van Hell, J.G., Fikkert, P. 2017. Second language attainment and first language attrition: The case of VOT in immersed Dutch-German late bilinguals. *Second Lang. Res.* 33, 483–518.

[14] Alharbi, A.S., Foltz, A., Kornder, L., Mennen, I. 2022. L2 acquisition and L1 attrition of VOTs of voiceless plosives in highly proficient late bilinguals. *Second Lang. Res.*, 1–31.

[15] Major, R. 1992. Losing English as a first language. *Mod. Lang. J.* 76, 190–208.

[16] Ditewig, S., Reubold, U., Mayr, R., Mennen, I. Under review. The relation between perceived non-native features and deviations from the L1 norm in adult sequential bilinguals.

[17] Bergmann, C., Nota, A., Sprenger, S.A., Schmid, M.S. 2016. L2 immersion causes non-native-like L1 pronunciation in German attriters. *J. Phon.* 58, 71–86.

[18] Wells, J. 1982. *Accents of English*. CUP.

[19] Brandstätter, J., Kaseß, C., Moosmüller, S. 2015. Quality and quantity in high vowels in Standard Austrian German. In: Leeman, A., Kolly, M.-J., Schmid, S., Dellwo, V. (eds), *Trends in Phonetics and Phonology: Studies from German-speaking Europe.* Peter Lang, 79–92.

[20] Cunha, C., Harrington, J., Moosmüller, S., Brandstätter, J. 2015. The influence of consonantal context on the tense-lax contrast in two standard varieties of German. In: Leeman, A., Kolly, M.-J., Schmid, S., Dellwo, V. (eds), *Trends in Phonetics and Phonology: Studies from German-speaking Europe.* Peter Lang, 65–77.

[21] Lehiste, I., Peterson, G.E. 1961. Transitions, glides and diphthongs. *J. Acoust. Soc. Am.* 33, 268–277.

[22] Venemann, T. 2000. From quantity to syllable cuts: on so-called lengthening in the Germanic languages. *Ital. J. Linguist.* 12, 251–292.

[23] Draxler, C., Jänsch, K. 2008. WikiSpeech – A content management system for speech databases. *Proc. 9th Interspeech* Brisbane, 22–26.

[24] Kisler, T., Reichel, U.D., Schiel, F. 2017. Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347.

[25] Winkelmann, R., Jänsch, K., Cassidy, S., Harrington, J. 2021. emuR: main package of the EMU Speech Database Management System. R package version 2.3.0.

[26] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer Program]. Version 6.2.23, retrieved from http://www.praat.org/

[27] Childers, D.G. 1978. *Modern spectrum analysis.* IEEE Computer Society Press, 252–255.

[28] Stevens, M., Harrington, J., Schiel, F. 2019. Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa: a journal of general linguistics* 4, 1–31.

[29] Harrington, J., Reubold, U. 2021. Accent reversion in older adults: evidence from the Queen's Christmas broadcast. In: Beaman, K., Buchstaller, I. (eds), *Language variation and language change across the lifespan: theoretical and empirical perspectives from panel studies.* Routledge, 119–137.

[30] Length, R. 2022 emmans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.1-1, <https://CRAN.R-project.org/package=emmeans>.

[31] Flege, J.E., Bohn, O.-S. 2021. The revised Speech Learning Model (SLM-r). In: Wayland, R. (ed), *Second language speech learning: Theoretical and empirical progress.* CUP, 3–83.

Prosodic cue weighting in the perception of lexical stress by French-English bilinguals: Preliminary behavioural data

Annie C. Gilbert[1,2], Claire T. Honda[1,2], Shari R. Baum[1,2]

[1]Centre for Research on Brain, Language, and Music, Montreal, Canada
[2]School of Communication Sciences and Disorders, McGill University, Montreal, Canada
annie.c.gilbert@mail.mcgill.ca , claire.honda@mail.mcgill.ca , shari.baum@mcgill.ca

ABSTRACT

Learning to adapt one's prosodic processing to a second or non-dominant language can be difficult, particularly if the second language uses different prosodic constructs than the native language. In the present study, we examine how French-English bilinguals with varied language experience (native language, relative language dominance, proficiency) assign weight to F0 and syllabic duration cues during the perception of English lexical stress. Participants were asked to report, via key press, which stress pattern they perceived when presented with words produced with either their original prosody or with F0 or duration edited. Preliminary analyses of the results from 16 participants suggest that language experience has a significant impact on the weight assigned to F0 cues in perceiving English lexical stress, but not on the weight assigned to duration cues. These results align with previous findings suggesting that French-English bilinguals have an easier time producing L2-specific duration cues than L2-specific F0 cues.

Keywords: bilingualism, speech perception, prosody, lexical stress, acoustic cue weighting

1. INTRODUCTION

Learning a second language (L2) is not a trivial task. Beyond learning to recognize new speech sounds, L2 learners must also learn to interpret prosodic cues appropriately. In many cases, this process requires the learner to assign new meaning to acoustic cues, because the same cue can play different roles during native language (L1) versus L2 perception. For instance, while rises in fundamental frequency (F0) are often associated with lexical stress in English [1], they are often associated with phrase boundaries in French [2], where they co-occur with phrase-final lengthening. Of note, lexical stress and phrase boundaries both provide important information about how to segment the speech stream, with English lexical stress tending to fall on word onsets [3] and French phrase boundaries falling on word offsets. Thus, the misinterpretation of the F0 cue in L2 (as lexical stress instead of phrase boundary, or vice-versa) could lead to segmentation errors, hindering speech comprehension.

Past research with bilingual participants has revealed that successful learning of L2 prosodic processing varies as a function of learners' experience and of the specific characteristics of the languages involved. For example, in terms of language experience, more native-like use of L2 or non-dominant language cues was predicted by greater length of immersion in an L2-speaking environment, higher L2 or non-dominant language proficiency, and relative language dominance [4-7].

Moreover, research has shown that it is easier to learn L2-specific prosody processing when both L1 and L2 make use of the same cue. For instance, one study compared English-French and Dutch-French bilinguals' use of F0 rise as a cue to word offset in French (their L2) [8]. English and Dutch both differ from French in that they typically use F0 rise to signal word onset rather than word offset; however, they also differ from each other in that the F0 rises associated with lexical stress/word onset in English generally co-occur with changes in vowel quality, which is not the case in Dutch. Thus, F0 rises have less functional weight in English than in Dutch. Interestingly, compared to the English-L1 speakers, the Dutch-L1 speakers showed earlier and greater use of the F0 rise cue in French [8]. This finding suggests that it was easier for Dutch-L1 listeners to successfully reallocate F0 rises to a new function in their L2 (i.e., identifying word offset instead of onset), compared to English-L1 listeners who usually encounter F0 rises paired with an additional cue (vowel change). It therefore appears that the more functional weight a cue has in one's L1, the easier it is to use the cue in an L2.

Thus, previous research has established the factors modulating learners' ability to learn to segment an L2 (or any non-dominant language) in a native-like manner, but questions remain as to how learners associate different acoustic cues to prosodic constructs in their L2—particularly when such constructs do not exist in their L1. Various studies have provided evidence that it is difficult to perceive and produce lexical stress when it is not used in one's L1 (e.g., [4, 6, 9]). As such, French-L1 learners of English have the challenging task of learning not only

to assign new meaning to prosodic cues to identify lexical stress instead of phrase boundaries, but also of learning to use lexical stress in the first place. The present study is part of a larger project aimed at clarifying the processes involved in such prosodic learning. We focus here on lexical stress perception, and on how non-native listeners assign weight to F0 and duration cues during lexical stress perception.

In the present study, we examined how French-English bilinguals with varied language backgrounds (native language, relative language dominance, English proficiency) assign weight to F0 and syllabic duration during the perception of English lexical stress. To do so, we created stimuli with conflicting F0 and syllabic duration cues, allowing us to observe on which cue participants rely. Based on previous research on speech segmentation in French-English bilinguals [4], we expected participants' use of F0 and duration cues to be modulated by their language experience.

We also expected that French-L1 participants might rely less heavily on F0 cues than native listeners to assign lexical stress in English because F0 rises often coincide with lengthened syllables in French, and therefore likely have a different weight in French compared to English.

2. METHOD

2.1. Participants

Sixteen French-English bilinguals were recruited from the Montreal area (18 to 41 years old, $M = 25$). Five reported French as their L1, five reported English as their L1, and six reported having been exposed to both languages from birth. Participants' proficiency in each language was assessed using a verbal fluency task. During this task, participants had one minute to name as many words as possible belonging to a specific category (one semantic category like "fruits" or "musical instruments" and three orthographic categories with words starting with specific letters per language.) The total number of words produced across trials in one language was used as a proficiency index for that language, and a relative language dominance index was computed by dividing the total number of English words produced by the total number of French words produced (see [4] and [5] for similar computations). Table 1 presents language proficiency scores of participants from each L1 group separately. Kruskal-Wallis Tests yielded significant differences across L1 groups in relative language dominance ($p = 0.024$) and English proficiency ($p = 0.049$). Pairwise comparisons revealed significant differences in relative language dominance only between English-L1 and French-L1 participants ($p = 0.023$), while English proficiency scores different significantly between English-L1 and French-L1 participants ($p = 0.021$) and between French-L1 and simultaneous bilinguals ($p = 0.049$).

	Language proficiency score		Relative language dominance
	English	French	English / French
	M (*SD*)	M (*SD*)	M (*SD*)
English-L1	63.8 (*11.8*)	41.2 (*10.5*)	1.61 (*0.40*)
Simultaneous	64.3 (18.0)	54.3 (18.4)	1.28 (0.46)
French-L1	45.0 (8.2)	46.6 (10.1)	0.99 (0.23)

Table 1: Average language proficiency (English and French) and relative language dominance scores as a function of participant native language.

2.2. Stimuli

Stimuli consisted of five bisyllabic base words that can be produced as either nouns or verbs depending on the prosodic structure used (trochaic nouns vs. iambic verbs; *implant*, *protest*, *refund*, *survey*, *torment*). The five words were recorded as both nouns and verbs by a native speaker of English in a sound-attenuated booth. Recordings were manually annotated by a trained phonetician to extract duration and mean F0 values for each vowel. F0 and duration ratios were computed for each version of the base word (trochaic noun and iambic verb) by dividing the value from the second syllable by the value from the first syllable. For example, a duration ratio above 1 indicates that the vowel of the second syllable was longer than the vowel of the first syllable while a ratio below one indicates that the first vowel was longer than the second (see [4] and [5] for a similar use of F0 and duration ratios).

Original recordings were then edited using *Praat* version 6.1.26 [10] to match the F0 or vowel duration ratio of the other stress pattern, leading to the creation of two edited conditions per stress pattern, as illustrated in Figure 1.

Stimuli were presented within three blocks following a practice block. The first block consisted of five repetitions of the original recordings of each base word produced with both trochaic and iambic stress patterns (as nouns or verbs). Stimulus order was

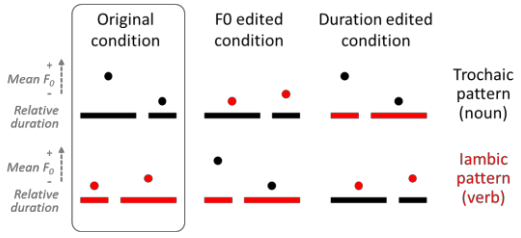


Figure 1: Schematic representation of conditions.

pseudorandomized to avoid consecutive repetitions of the same base word (50 trials in total). Results from this block are used a reference condition to which the other conditions are compared. The other two blocks (order counterbalanced across participants) consisted of five repetitions of each base word in the six conditions represented in Figure 1 (for a total of 150 trials per block). Stimulus order was again pseudorandomized to avoid consecutive repetitions of the same base word and the same condition. Participants had the opportunity to take up to three short breaks during each block as well as a break between blocks.

2.3. Procedure

Participants sat in a quiet room about 60 cm from a computer screen and listened to the stimuli through insert earphones (EARTONE 3A) at 66 dBA. They were instructed to listen to the word presented in isolation and to simply indicate which stress pattern they perceived via button press. Special emphasis was placed on the fact that participants should respond based on their first impression, even if this was not a speeded task. Half the participants signalled perceiving stress on the first syllable (trochaic pattern or noun) with the left hand and half with the right hand. Visual reminders of response key assignment were provided onscreen where stress was represented by capital letters. For instance, when hearing a trial involving the base word "protest", participants would see the written forms "PROtest" and "proTEST" above their assigned response buttons.

2.4. Statistical analyses

Responses were coded as "1" if the participant assigned the original stress pattern to the word regardless of the edited F0 or duration, and as "0" if they assigned the opposite stress pattern. Responses were analysed using a series of generalized linear mixed effects models (GLME) comparing results from the edited conditions to results from the first block of original recordings. Such comparison allowed us to determine if participants relied on the edited F0 or duration ratios to assign stress to the presented words. Each edited condition and original stress pattern was analysed separately.

Moreover, the models examined whether individual differences in language experience impacted reliance on F0 and duration during stress assignment. Based on previous publications, we investigated the potential impact of relative language dominance (log transformed), English proficiency (scaled), and L1 (as reference group = English-L1) [4, 5, 11]. In cases where a significant effect of L1 was observed to affect non-native (French-L1) participants but not simultaneous bilinguals, a follow-up model was fitted using French-L1 as reference group to compare French-L1 and simultaneous bilinguals. The random structure of the models took into account participants (intercept) and base words (intercept; [Response ~ Condition * L1 + (1 | Participant)] + (1 | Base word)], family = binomial). GLME models were implemented in RStudio version 3.2.4 [12], using the lme4 library, version 1.1-7 [13], and estimates of p-values were obtained using the lmerTest package version 2.0-29 [14].

3. RESULTS

3.1. Trochaic pattern (nouns)

3.1.1. F0 edits vs. originals

The model examining the effect of L1 on the use of F0 in stress assignment revealed a significant interaction between L1 and condition (original vs. edited F0), where French-L1 listeners presented a smaller difference between conditions than native listeners (English-L1 and simultaneous bilinguals; ($|z| > 2.972$, $p < 0.003$). Replacing L1 by relative language dominance yielded no significant effect of dominance as a main effect or as part of an interaction ($|z| < 1.698$, $p > 0.089$).

On the other hand, the model taking English proficiency into account revealed a significant interaction between condition (original vs. edited F0) and English proficiency ($z = 2.962$, $p = 0.003$). Namely, participants who named more English words during the verbal fluency task showed a greater difference between conditions than participants who named fewer words. This pattern of results suggests that higher-proficiency English listeners rely more heavily on F0 when assigning stress than lower-proficiency English listeners.

3.1.2. Duration edits vs. originals

The models yielded no significant effect of individual differences as either main effects or as part of interactions ($|z| < 1.647$, $p > 0.099$). Only the main effect of condition remained significant across comparisons ($|z| > 4.272$, $p < 0.001$), with participants reporting having heard the original stress pattern less often in the duration edited condition than in the original condition.

3.2. Iambic pattern (verbs)

3.2.1. F0 edits vs. originals

As observed earlier, the model examining the effect of L1 on the use of F0 in stress assignment revealed a

significant interaction between L1 and condition. That is, replacing the original F0 ratio by the ratio from the trochaic condition had a smaller impact on stress assignment for French-L1 than for English-L1 participants (z = -2.133, p = 0.033). The model investigating the impact of relative language dominance yielded no significant effect of dominance (either as a main effect or part of an interaction; $|z| < 0.461$, $p > 0.645$).

Interestingly, the model investigating the effect of English proficiency on the use of F0 in stress assignment yielded no significant interaction between English proficiency and condition (z = 1.294, p = 0.196), but revealed a main effect of English proficiency, where less proficient English listeners were less likely to recognize the iambic stress pattern even in the original condition (z = 1.999, p = 0.046).

3.2.2. Duration edits vs. originals

As observed for the trochaic pattern, the models yielded no significant effect of individual differences as either main effects or part of interactions ($|z| < 1.939$, $p > 0.053$) while the main effect of condition remained significant across comparisons ($|z| > 4.117$, $p < 0.001$), suggesting that listeners' stress assignment was affected by modulation of duration patterns regardless of their linguistic experience.

4. DISCUSSION / CONCLUSION

The goal of the present study was to determine how much weight French-English bilinguals assign to different acoustic cues (F0 and duration) during lexical stress perception in English, and whether individual differences in language experience affect this weighting.

Our preliminary results suggest that experience did affect the use of F0 as a cue to lexical stress. Specifically, French-L1 bilinguals assigned less weight to F0 compared to English-L1 or simultaneous bilinguals. This finding held across both trochaic (noun) and iambic (verb) conditions. The use of duration as a cue, on the other hand, was not affected by experience. All participants assigned similar weight to duration regardless of their L1.

These results are in line with work that compared French-English bilinguals' prosodic cue weighting during the production of English sentences [4], which had found that bilinguals showed difficulty adapting their F0 production to the specifics of their L2, whereas they showed no difficulty adapting their duration production—in fact, some participants even overcompensated for their inability to adapt F0 production by relying more heavily on duration as a cue. It appears that, at least for French-L1 learners of English, duration is an easier prosodic cue to adapt than F0.

Interestingly, the effect of F0 on stress perception depended not only on language background, but also in part on language proficiency. The more proficient the bilinguals were in English, the more they used F0 to perceive lexical stress in nouns. During verb perception, there was only a main effect of proficiency rather than a proficiency-by-condition interaction as had been found for noun perception. This pattern could have arisen for a couple of reasons. The F0 values for English verbs (F0 rise on the second syllable) are more similar to the values that would be found for French words, so perhaps this made it more complicated for French listeners to use F0 in an English-like way in the verb condition, regardless of proficiency. It could also be that, because the lexically stressed syllable was less marked in the verb condition than in the noun condition (i.e., the noun stimuli had greater F0 and duration differences across syllables compared to the verbs), proficiency provided less of an advantage when recognizing the more subtle stress cues in verbs.

Contrary to the results for F0, the effect of duration on lexical stress perception did not depend on language proficiency or language dominance. This is again in line with the previous finding that duration is an easier prosodic cue to adapt to one's L2 [4].

Our preliminary results also suggest that French-English bilinguals are capable of perceiving lexical stress, even if they are native or proficient speakers of French (which does not use lexical stress). Earlier work had suggested that French speakers have great difficulty with lexical stress perception (e.g., [9]), but more recent work corroborates our finding that French-L1 bilinguals are capable of showing similar lexical stress perception to native listeners [15].

Overall, we found that even though English and French are languages involving different weighting and different use of the same acoustic cues, French-English bilinguals are able to adapt their prosodic cue use to the language being perceived (in this case, English). For example, French-L1 speakers—who would normally use F0 as a cue to phrase boundaries—were generally able to use F0 as a cue to English lexical stress, albeit to a lesser extent than native listeners. This ability to adapt cue use appears to develop to some extent with increased L2 proficiency, in line with other findings [8].

By manipulating two acoustic parameters and studying a group of bilinguals with varied language experience, we helped to elucidate the acoustic and experiential factors involved in stress perception. Future work could further investigate the extent to which language background affects the use and adaptation of various prosodic cues from L1 to L2.

7. REFERENCES

[1] Morton, J., Jassem, W. 1965. Acoustic Correlates of Stress. *Lang. Speech.* 8, 159–181.

[2] Jun, S.-A., Fougeron, C. 2002. Realizations of accentual phrase in French intonation. *Probus* 14, 147-172.

[3] Cutler, A., Norris, D. 1988. The role of strong syllables in segmentation for lexical access. *J. Exper. Psych.: Human Percep. Perform.* 14, 113-121.

[4] Gilbert, A. C., Wolpert, M., Saito, H., Kousaie, S., Itzhak, I., Baum, S. R. 2019. Adaptive and selective production of syllable duration and fundamental frequency as word segmentation cues by French-English bilinguals. *J. Acoust. Soc. Am.* 146, 4255-4272.

[5] Gilbert, A. C. et al. 2021. Spoken Word Segmentation in First and Second Language: When ERP and Behavioral Measures Diverge. *Front. Psychol.* 12, 3974.

[6] Tremblay, A. 2008. Is second language lexical access prosodically constrained? Processing of word stress by French Canadian second language learners of English. *Appl. Psycholinguist.* 29, 553-584.

[7] Tremblay, A. et al. 2017. Experience with a second language affects the use of fundamental frequency in speech segmentation. *PLOS ONE* 12, e0181709.

[8] Tremblay, A., Broersma, M., Coughlin, C. E. 2018. The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilin.: Lang. Cogn.* 21, 640-652.

[9] Dupoux, E., Pallier, C., Sebastian, N., Mehler, J. 1997. A Destressing "Deafness" in French? *J. Mem. Lang.* 36, 406-421.

[10] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341-345.

[11] Anonymous. Submitted. Phrase parsing in a second language as indexed by the Closure Positive Shift: The impact of language experience and acoustic cue salience.

[12] R Development Core Team. 2010. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

[13] Bates, D., Maechler, M., Bolker, B., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6.

[14] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2015. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-29.

[15] Gilbert, A. C., Honda, C. T., Phillips, N. A., Baum, S. R. 2021. Near native-like stress pattern perception in English-French bilinguals as indexed by the mismatch negativity. *Brain. Lang.* 213, 104892.

ACOUSTIC ANALYSES OF MANDARIN TONES PRODUCED BY BILINGUAL ELEMENTARY STUDENTS IN CANADA

Youran Lin[1], Karen Pollock[1], Benjamin Tucker[1,2], Fangfang Li[3]

[1]University of Alberta, Canada, [2]Northern Arizona University, USA, [3]University of Lethbridge, Canada
youran1@ualberta.ca, karen.pollock@ualberta.ca, benjamin.tucker@nau.edu, fangfang.li@uleth.ca

ABSTRACT

In Canada, children can learn Mandarin in bilingual schools as a heritage language (HL) or second language (L2). Many Canadian students have greater difficulty acquiring Mandarin tones than other aspects of the language. The present study investigates bilingual students' productions of Mandarin citation tones. Pitch contours are modelled using generalized additive mixed models for HL and L2 learners across grades to examine the roles of speech input. Results suggest that early home language experiences and recent school learning experiences interact with the tone targets and impact students' tone productions differently. These results provide evidence on Mandarin phonetic learning in Canada among bilingual students from diverse backgrounds and expand the evidence for L2 phonetic learning theories in the suprasegmental domain among child learners of non-English languages.

Keywords: Mandarin, tone, acoustic, bilingual, children

1. INTRODUCTION

In Western Canada, English is the majority language, but two-way bilingual education is available in many international languages. The current study concerns a Chinese-English bilingual program in Edmonton, Alberta. This program delivers half of its academic content in Mandarin Chinese and the other half in English, thus attracting both students who learn and maintain Mandarin as a heritage language (HL) and those who learn Mandarin as a second language (L2).

It has been argued that a two-way design supports students with diverse backgrounds to develop functional bilingualism [1]. Empirical evidence from Indo-European languages suggested that English-speaking children can learn the pronunciation of a minority L2 similarly to their HL peers, despite their different home language input [2, 3]. However, little is known about how school-aged children in an English-dominant environment learn a minority language from another language family, for example, a tonal language like Mandarin.

Mandarin tones primarily use pitch contours to differentiate meanings, which can be measured acoustically by a sequence of fundamental frequency (*f0*) over time. There are four citation tones in monosyllabic productions: high-level ◌̄ (T1), mid-rising ◌́ (T2), low-dipping ◌̌ (T3), and high-falling ◌̀ (T4) tones. Pitch contours of Mandarin tones are often described in a 5-degree convention, i.e., T1 [55], T2 [35], T3 [214], and T4 [51], where a larger number represents a higher pitch. The developmental order is typically T1 > T4 > T2 > T3, where T2 and T3 are confusable due to their phonetic similarity [4, 5].

Not only the phonetic characteristics of the targets but also the amount and quality of speech input play a role in L2 phonetic learning, as suggested by the Speech Learning Model (SLM) [6]. Furthermore, early input and recent input may contribute differently, which could be attributed to differences in neurocognitive maturation [7] or differences in the development of phonetic categories [6]. The Native Language Magnet theory (NLM) claims that the perception of phonetic categories is biased by language-specific input early in life, which impacts perceptual learning of other languages later in life [8]. Therefore, we hypothesize that bilingual students' tone learning is influenced differently by early home language input and recent school input, as well as by the specific tone targets.

In this paper, we examine the four citation tones produced by students in a Chinese bilingual program in Canada. We compare pitch contours between HL and L2 groups and between grades 1 and 3 to determine the roles of early home input and recent school input: HL and L2 students differ in their home language input; Grade 1 and grade 3 students differ in their amount of Mandarin input at school. By examining the learning of Mandarin tones, this paper expands the evidence for L2 speech learning theories in child learners of a non-English language. To eliminate the impact of tone errors on pitch contours, this study only examines the correct productions, thus investigating the phonetic refinement of tones instead of the perceived accuracy.

2. METHOD

2.1. Participants

HL and L2 students were recruited from grade 1 (age (months) $\mu = 77.35$, $\sigma = 3.65$) and grade 3 ($\mu = 103.16$,

σ = 4.13). Parent questionnaires [9, 10] indicated that the HL group (N = 26) has strong home input in Mandarin and a wide range of English exposure before school, and the L2 group (N = 30) receives predominantly home input in English. Table 1 depicts the differences between groups with regard to the onset of exposure to both languages, the time their parents spend speaking both languages, and their parents' proficiencies in both languages.

	HL	L2
Onset age of regular exposure to Mandarin (months)	0 – 1	37 – 96
Onset age of regular exposure to English (months)	0 – 93	0 – 1
Parent % of time speaking Mandarin*	50 – 100	0 – 15
Parent % of time speaking English	0 – 100	80 – 100
Parent self-reported Mandarin proficiency on a scale of 0-5*	4 – 5	0 – 3
Parent self-reported English proficiency on a scale of 0-5	1 – 5	4 – 5
* The parent who is stronger at Mandarin.		

Table 1: A comparison of the home language environment of HL and L2.

In addition to the students described above, we recruited 12 Chinese teachers from the bilingual program to provide their speech samples as a reference. Among them, seven were native (L1) speakers of Mandarin, five were L1 speakers of another Chinese language and started learning Mandarin at school age, and two were born in Canada and graduated from the bilingual program.

2.2. Procedures

2.2.1 Speech sample collection

A picture-based elicitation task [11, 12] was used to elicit a list of 72 words. Three Mandarin-native speakers who were proficient in English administered the task. Spontaneous productions were encouraged but imitative models were provided when necessary. Two examples for each tone were selected based on the spontaneity of production (Table 2). Therefore, in total 544 productions of HL, L2, and teachers were collected (68 speakers × 8 words = 544 words).

Tone	Word	Pinyin	IPA	Meaning	Spontaneity
T1 ˉ	三	sān	[san^{55}]	three	98%
	八	bā	[pa^{55}]	eight	96%
T2 ˊ	蓝	lán	[lan^{35}]	blue	92%
	鱼	yú	[y^{35}]	fish	86%
T3 ˇ	手	shǒu	[ʂo͡ʊ^{214}]	hand	95%
	狗	gǒu	[ko͡ʊ^{214}]	dog	92%
T4 ˋ	二	èr	[ɐ˞^{51}]	two	98%
	绿	lǜ	[ly^{51}]	green	94%

Table 2: Selected target words for the four tones.

2.2.2 Transcription

Speech samples were transcribed by four Mandarin-speaking researchers [13]. Each tone production was transcribed as either one of the four citation tones, an inappropriate allophonic variation [5], or an uncategorizable production. Spontaneity was coded. In the whole wordlist, 23% of the samples were transcribed by a second transcriber and reached 90% inter-transcriber reliability. The first author re-examined the selected words and identified questionable transcriptions. Transcribers reached full consensus through blinded voting, revising 44 out of the 47 questionable transcriptions. Only spontaneous productions that were recognized as correct were included, which left 406 productions, 317 from students and 89 from teachers. The accuracy of each tone by each group is presented in Figure 1.

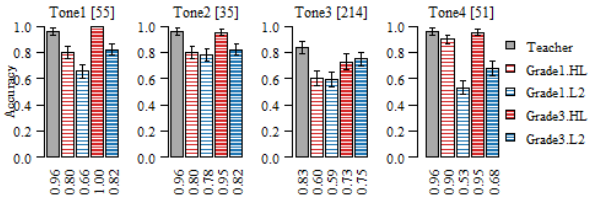


Figure 1: Accuracy of the four lexical tones among teachers and students with 95% confidence intervals.

2.2.3 f0 extraction

The onset and offset of pitch contours were labelled, including the nucleus and voiced coda. The first and last four cycles were excluded to avoid perturbations [4]. ProsodyPro was used to extract *f0* [14], where automatically recognized voicing pulses were examined and manually adjusted. Ten *f0* values with equal time intervals were extracted for each syllable. Trimming and smoothing functions were disabled to obtain raw *f0* values. The *f0* values were normalized into T values within each speaker according to Function (1) [15]. It logarithmically compressed a speaker's *f0* productions into a 5-degree scale [16].

(1) $T = 5 \times \frac{log_{10}^{x_i} - log_{10}^{x_{min}}}{log_{10}^{x_{max}} - log_{10}^{x_{min}}}$.

2.2.4 Statistical analysis

To model pitch contours, generalized additive mixed models (GAMMs) [17] were conducted using `mgcv` and `itsadug` in R [18, 19]. GAMMs were used due to the inclusion of parametric (linear), smooth (nonlinear), and random terms. The nonlinear terms were especially suitable to model the nonlinear pitch contours of Mandarin tones.

3. RESULTS

3.1 Model comparisons

Pitch contours were modelled with the fixed effects of group (HL and L2), grade (ordered, 1 and 3, and 0 for teachers), and tone (T1 to T4), as well as random smooths of speaker and word token. Interaction effects were implemented with contrast coding to allow interactions with time (the x-axis of contours). Models were compared using the `compareML` function [19]. The selected model with the lowest AIC included parametric and smooth terms of the 3-way interactions among tone × group × grade and the random smooths. This model suggested that pitch contours were impacted by the interactions among tones, groups, and grades. Therefore, a model was built for each tone to examine group and grade effects.

3.2 Group and grade effects

Each tone's model included group × grade interaction as parametric and smooth terms, as well as the random smooths. As an example, in Figure 2 for T1 model, the top row shows predicted contours of T1 by subgroups. The following rows include difference plots to examine the contrasts of interest: The middle row presents in grade 1, how HL and L2 differed from teachers and how L2 differed from HL. The bottom row presents the same for grade 3. In a difference plot, a section above zero suggests a higher pitch contour in the first group and *vice versa*. The percentage of sections of significant differences can be calculated to indicate the extent of differences.

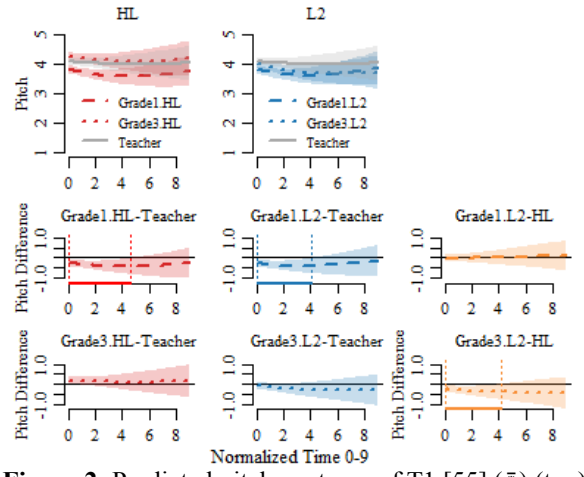


Figure 2: Predicted pitch contours of T1 [55] (◌̄) (top), and group differences in grade 1 (middle) and grade 3 (bottom). HL is coded in red, L2 is in blue, and L2-HL difference is in orange. Grade 1 is coded in dashed lines, grade 3 is in dotted lines, and teachers are in grey solid lines. Ribbons describe 95% confidence intervals. Vertical lines mark sections of significant differences.

For the high-level (◌̄) T1 (Figure 2), all subgroups produced high-pitch contours. In grade 1, HL's and L2's contours were different from teacher models (the percentage of sections with significant difference to the full range of normalized time for HL was 52%; L2: 45%), both with a lower pitch onset. This means both groups did not acquire the "level" feature in grade 1 and produced a slightly rising contour. In grade 3, both groups showed no significant difference from teachers, although L2's onset was significantly lower than HL (46%). This suggests both groups were able to acquire T1's features in grade 3, although the two groups' productions were slightly different phonetically.

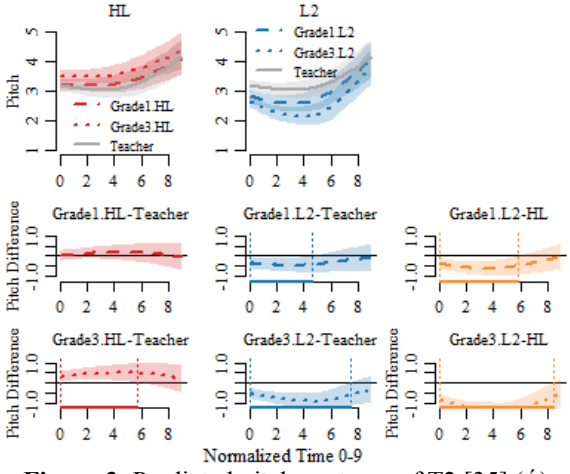


Figure 3: Predicted pitch contours of T2 [35] (◌́), differences between students and teachers, and group differences in both grades.

For the mid-rising (◌́) T2 (Figure 3), in grade 1, L2's pitch contour was significantly different from teacher models (52%) with a lower pitch onset. Unsurprisingly, L2's contour was also different from their HL peers in grade 1 (65%). In grade 3, the two groups produced opposite patterns: HL's contour was significantly different from teachers' (64%) with a higher onset. On the other hand, L2's contour was also different from teachers' (83%) but showed a lower onset and an even lower middle section of the pitch contour. Consequently, HL's and L2's difference was more distinct in grade 3 (94%). This suggests that while HL produced a mid- to high-rising contour in grade 3, L2 produced a mid to low onset with a dipping contour. It indicates that L2's T2 productions in grade 3 could be more subjected to being misrecognized as T3 by listeners, as a dipping contour with a falling portion is a characteristic feature of T3 compared to T2 [20, 21]

For the low-dipping (◌̌) T3, all subgroups produced a low-dipping pitch contour (Figure 4). Compared to the teacher models, HL in grade 1 produced a slightly higher offset (5%); HL in grade 3 produced a slightly higher onset (26%); L2's contours were not different from teacher models in both grades. Furthermore, L2's and HL's pitch contours were

significantly different in grade 3 (34%). However, given the complex phonetic specifications of T3, it remains a question whether these differences were perceptually significant.

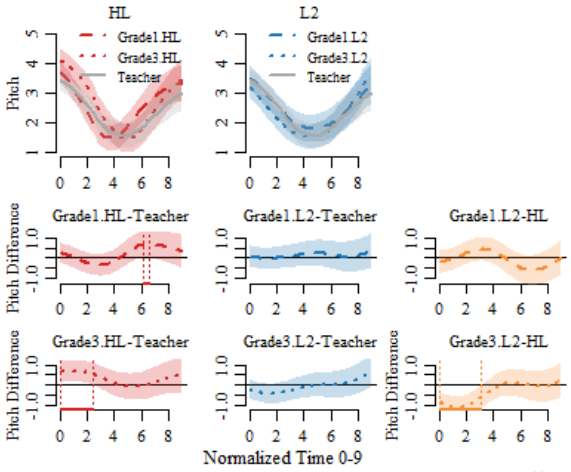


Figure 4: Predicted pitch contours of T3 [214] (˘), differences between students and teachers, and group differences in both grades.

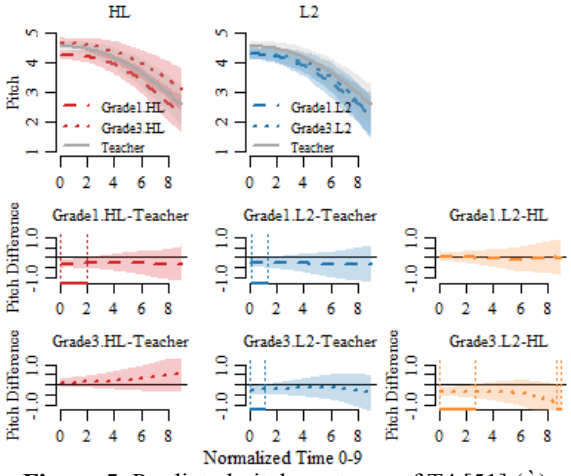


Figure 5: Predicted pitch contours of T4 [51] (ˋ), differences between students and teachers, and group differences in both grades.

For the high-falling (ˋ) T4, all subgroups produced a falling pitch contour (Figure 5). In grade 1, both HL's (22%) and L2's (13%) contours were significantly different from teacher models with a lower pitch onset, and the two groups showed no difference between each other. This suggests that both groups did not fully acquire the "high" feature of T4 in grade 1. However, in grade 3, HL's contour had no difference from teacher models, while L2's contour still had a lower onset than teachers' (12%), causing a significant difference between the HL and L2 groups (32%). This suggests that HL was able to acquire the high feature in grade 3, but L2 still had marginal difficulty producing a high-pitch onset. Similar to T3, the short sections of differences in T4 might be perceptually insignificant.

4. DISCUSSION

Both HL and L2 students produced some key features of citation tones as early as in grade 1, suggesting the efficiency of the two-way bilingual program [20]. However, their phonetic specifications sometimes differed from teacher models, suggesting a process of phonetic learning [4]. In this process, the impacts of speech input interacted with tone targets. Specifically, when the target was simple, bilingual students followed the same developmental trajectory despite their different backgrounds. However, when the target was complex, bilingual students' phonetic learning was impacted by their home input.

As a reminder, L1 children's developmental order of tones is T1 > T4 > T2 > T3 [4, 12]. For the earliest-developmental high-level T1, five-year-old L1 children's productions tend to be less level [4, 5]. Bilingual students demonstrated a similar pattern: Both HL and L2 produced the high feature in grade 1 and acquired the level feature later in grade 3. For the high-falling T4, young L1 children's productions tend to show a falling trend that is not steep enough [4, 5]. HL and L2 both produced the falling feature in grade 1, but only HL produced a high onset in grade 3, therefore a steeply falling contour. The mid-rising T2 is perceptually and productively confusable with the low-dipping T3, because both contours include a rising section [4, 21, 22]. The current study shows that HL produced the rising feature in grades 1 and 3, while L2 in grade 3 produced a dipping contour. It seems that the increasing input at school did not help L2 students refine their T2 productions. Instead, they were at an increased risk of T2-T3 convergence. A possible explanation is that L2 students were less sensitive to pitch due to limited early exposure and did not form the same phonological representations as HL students [8], which hindered their phonetic refinement of complex tone targets. No conclusive patterns were identified for the latest-developmental T3. A closer look at the tone errors is warranted.

These results differ from the evidence in Indo-European languages that home input difference can be levelled out by school learning [2, 3]. The current study indicates that the role of home input in L2 phonetic learning is impacted by the targets [6], with the learning of late-developmental targets more impacted by early home input. Moreover, it provides evidence in the suprasegmental domain for how speech input impacts phonetic learning [6] and how early input might tune learners' speech perception and impact future learning [8]. It needs to be noted that this study focused on phonetic specifications of productions judged as accurate, and future studies will further examine the relationships between acoustic measurements and perceived accuracy.

5. ACKNOWLEDGEMENTS

This paper draws on research supported by the Social Sciences and Humanities Research Council (SSHRC). The authors would like to thank the participating schools, students, parents, and teachers, as well as the research assistants who helped with transcription and tone labeling.

6. REFERENCES

[1] Cummins, J. 1979. Linguistic Interdependence and the Educational Development of Bilingual Children. *Rev. Educ. Res.* 49(2), 222–251.

[2] Menke, M. R. 2017. Phonological development in two-way bilingual immersion. *J. Second. Lang. Pronunciation*. 3(1), 80–108.

[3] Nance, C. 2020. Bilingual language exposure and the peer group: Acquiring phonetics and phonology in Gaelic Medium Education. *Int. J. Biling.* 24(2), 360–375.

[4] Wong, P. 2012. Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *J. Phon.* 40(1), 141–151.

[5] Rattanasone Xu, N., Tang, P., Yuen, I., Gao, L., Demuth, K. 2018. Five-year-olds' acoustic realization of mandarin tone sandhi and lexical tones in context are not yet fully adult-like. *Front. Psychol.* 9.

[6] Flege, J. E., Bohn, O.-S. 2021. The revised Speech Learning Model. In: Chen, S. (ed), *Second language speech learning: Theoretical and empirical progress*. Cambridge University Press, 3–83.

[7] DeKeyser, R. M. 2000. The robustness of critical period effects in second language acquisition. *Stud. Second. Lang. Acquis.* 22(4), 499–533.

[8] Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., Nelson, T. 2008. Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Phil. Trans. R. Soc. B*. 363, 979–1000.

[9] Marian, V., Blumenfeld, H. K., Kaushanskaya, M. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *J. Speech. Lang. Hear. Res*. 50, 940–967.

[10] Paradis, J. 2011. Individual differences in child English second language acquisition. *Linguist. Approaches. Biling*. 1(3), 213–237.

[11] Zhao, J., Bernhardt, B. M. 2012. Mandarin Single-Word Elicitation Tool for Phonology. http://phonodevelopment.sites.olt.ubc.ca/mandarin-picture-elicitation_2012/.

[12] Zhu, H. 2002. *Phonological Development in Specific Contexts: Studies of Chinese-speaking Children*. Multilingual Matters.

[13] Hedlund, G., Rose, Y. 2020. Phon 3.1. https://phon.ca.

[14] Xu, Y. 2013. ProsodyPro-A Tool for Large-scale Systematic Prosody Analysis. *TRASP* Aix-en-Provence, 7–10.

[15] Shi, F., Wang, P. 2006. A statistic analysis of tone groups in Beijing Mandarin. *Contemporary Linguistics* 8(4), 379–380.

[16] Chao, Y.-R. 1968. *A Grammar of Spoken Chinese*. University of California Press.

[17] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phon*. 70, 86–116.

[18] Wood, S. N. 2017. *Generalized additive models: An introduction with R (2nd ed)*. CRC Press.

[19] van Rij, J., Wieling, M., Baayen, R., van Rijn, H. 2022. itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.4.1.

[20] Xu, Y. 1997. Contextual tonal variations in Mandarin. *J. Phon.* 25, 61–83.

[21] Wang Y., Li, M. 2010. The effect of tone pattern and register in perceptions of Tone 2 and Tone 3 in Mandarin. *Acta. Psychol. Sin.* 42(9), 899-908.

[22] Wang, Y., Jongman, A., Sereno, J. A. 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *J. Acoust. Soc. Am*. 113(2), 1033–1043.

HESITATING WITH AND WITHOUT LANGUAGE HERITAGE - PROSODIC ASPECTS OF FILLER PARTICLES IN THE RUEG CORPUS

Marlene Böttcher, Margaret Zellers

ISFAS, Kiel University, Germany
mboettcher@isfas.uni-kiel.de, mzellers@isfas.uni-kiel.de

ABSTRACT

This study investigates the frequency and prosodic form of filler particles (FP, such as *uh* and *uhm*) produced by speakers of English, with and without a heritage language (HL) background. This is done based on semi-spontaneous corpus data. A subset of 24 speakers with Russian, German or no HL background, each producing 2 narrations, was analyzed in this study. There were 304 FPs realized in the 50 min of analyzed speech. Overall English speakers' productions show close similarity in FP form. The analysis shows no effect of HL status for FP frequency and frequency of different FP forms (*uh* vs *uhm*) as well as FP duration and prosodic phrasing. The acoustic analysis of pitch slope, however, showed differences between speaker groups with different and without HL background. Interesting results emerged when including gender into the analysis, which will be addressed in future work.

Keywords: filler particles, hesitation, spontaneous discourse, heritage speakers, bilingualism

1. INTRODUCTION

Spontaneous discourse differs from prepared or read speech in many ways. As part of its unplanned nature, it contains many disfluencies like repetitions, repairs and fillers. Such disfluencies can comprise up to ten percent of words in natural conversations [1]. One type of filler in English are the constituents *uh* and *uhm* (and their variants). Different terms have been used in the literature (e.g. *hesitation, disfluency, error* or *filled pause, discourse particle*) reflecting researchers' approaches on fillers as either symptoms [2, 3], or signals in discourse [4, 5, 1] or stressing their similarity with silent pauses [6, 7]. Following [8] we use the term *filler particle* (FP) to refer to lexically underspecified, non-inflectable entities which are frequently produced in naturally occurring speech, in an effort to be agnostic about the status and function of FPs. Generally, FPs are produced preceding new thought units. Their use is therefore connected to speech planning and language processing as well as discourse organisation [4]. Speech planning is an interesting area in bilingual research, since the availability of two languages and the suppression of one of them depending on the context of use demands a higher cognitive load in bilinguals' speech [9]. Higher cognitive load can then be related to slower speech rate and an increase of hesitation phenomena [9, 10]. Hence, the use of FPs has been found to be an interesting area of research in the study of bilingualism [11, 12] and heritage language use [13].

2. BACKGROUND

2.1. Filler particles

In their frequently cited work, [4] were among the first to classify FPs like *uh* and *uhm* as words "with conventional phonological shapes and meanings [...] governed by the rules of syntax and prosody" [4, p.75]. In English a central vowel is often followed by a nasal[1]. The dichotomy between a vocalic (V, *uh*) and a nasalized variant of FPs (VN, *uhm*) is well established [4, 5, e.g.]. and the use of the two forms has been related to gender-specific use patterns: while female speakers tend to produce more VN variants, male speakers seem to produce more V variants [14]. With regard to their prosody, FPs have been described with reduced fundamental frequency (F0) and a level or gradually falling F0 contour [1, 8]. Their phonological form of a central vowel optionally followed by a nasal and their reduced pitch characteristics leads to FPs having relatively low perceptual prominence [15]. FPs have been interpreted as symptoms of difficulties in speech planning [6] and have been used as an indicator for difficulties in bilingual speech due to higher cognitive load [10]. While FPs are connected to speech planning and processing, their occurrence is rule based and follows specific functions in spontaneous discourse [5]. FPs can be seen as one of many aspects of language learners have to acquire. Since FPs are perceptually not prominent they may pose challenges in language acquisition. That is, phenomena at the edge of our consciousness

are more difficult to learn than more salient aspects of language. This might be especially relevant in bilingual contexts with limited input such as in heritage language acquisition.

2.2. Filler particles in bilingual speech

Investigations of FPs in bilingual speech have shown deviances from monolingual FP use. In a study on English speakers of French, [11] found that bilinguals overly produce FPs along with other non-lexical markers compared to lexical discourse markers. These speakers also overused the vocalic *uh* form compared to *uhm* [11]. Similarly, a study by [16] on English speakers of Dutch found more FPs and silent pauses in their L2 speech compared to L1 [16]. While these findings suggest the acquisition of FP frequency and form to be a challenge for L2 learners other research suggests differently. In a study on English and te reo Māori bilinguals, [17] showed that these bilingual speakers display a language-specific use of FPs in both their languages. The acquisition of an appropriate FP use might therefore be related to the amount of exposure and the language status.

2.3. Bilingual heritage speakers

Heritage speakers (HS) represent a specific case of bilinguals speaking a heritage language and a majority language. The majority language (ML) is typically the dominant language of these bilinguals and used in everyday and most areas of the public sphere, e.g. work and education. The heritage language (HL) as a minority language in the larger society is acquired in specific contexts and typically spoken at home, e.g. with relatives and friends [13]. The use of the HL is therefore limited to certain interlocutors, genres and registers. Prior research has shown HL acquisition to be an interesting area of research regarding language acquisition, language contact phenomena and language change [18].

2.4. Research aim and hypotheses

This study addresses the productions of FPs in the speech of English speakers with Russian, German and no HL background. Specifically, the aspects of FP frequency, FP segmental form (vocalic *uh* (V) vs vocalic-nasal *uhm* (VN)) and the prosodic realisation of FPs are analysed. Following prior findings on FP use in bilingual speech, the hypotheses are:

1. Bilingual speakers of English produce FPs more frequently than monolingual speakers since the cognitive load of monitoring two languages is higher in bilingual speech planning [9].
2. Bilingual speakers deviate from the monolingual preference for a VN form of FPs resulting in a transfer of a V preference from heritage Russian [19].
3. Bilingual speakers deviate from monolingual realisations in duration, prosodic phrasing and slope resulting in transfer of discourse particle intonation from their HL (e.g. rising *nu* in Russian [20]).

3. DATA ANALYSIS

3.1. The corpus

The hypotheses presented above are investigated using the RUEG corpus, a corpus of semi-spontaneous speech including data from both mono- and bilingual speakers with different language backgrounds [21]. The narrations within the RUEG corpus were elicited by means of the Language Situation Method [22]: A video of a car accident was used to prompt participants to explain what happened in two situations. Participants were asked to provide a police report in form of a voice message on the phone, yielding a formal register. Participants were also asked to describe the incident to a friend by means of a voice message, yielding an informal situation. The hypotheses in this study do not make any predictions regarding possible register differences.

3.2. Acoustic analyses

For the acoustic analysis a subset of 24 speakers from the RUEG corpus data was selected (8 monolinguals (E), 8 bilinguals with Russian HL (R), 8 bilinguals with German HL (D); 11 male; mean age 16 y.) Since each speaker produced 2 narrations there were 48 narrations comprising 50 minutes of monologues. In this data, 303 FPs were realised, which were then further analysed for the acoustic parameters duration, pitch level and pitch slope. Both mean pitch and F0 maxima and minima were measured within the segmented FPs in Praat [23]. The speaker's mean pitch over the whole narration was also measured. The slope of the FP was then calculated as the difference between F0 maximum and F0 minimum in semitones (st). Slopes larger than 40 st were excluded as measurement errors considering physiological constraints [24]. Additionally, FPs were annotated for their segmental form and their prosodic phrasing based on universal

phonetic cues [25]. Based on the F0 measurements pitch contour was categorized as *level, falling* and *rising* with excursion size values larger than 1 semitone for the latter two categories. Statistical analyses were carried out in R using the lme4 package [26].

4. RESULTS

4.1. FP frequency

In the data analyzed here, 303 FPs were realised by 21 speakers in 39 narrations (i.e. not all narrations by all speakers contained FP productions). Given the different length of narrations, the frequency of FPs was normalized per 100 words. There were on average 4.6 FPs per 100 words produced in the data set (D: 4.3; R: 4.5; E: 4.6). A linear regression analysis revealed no effect of the speakers' language background or gender (all p-values > 0.30).

4.2. FP form

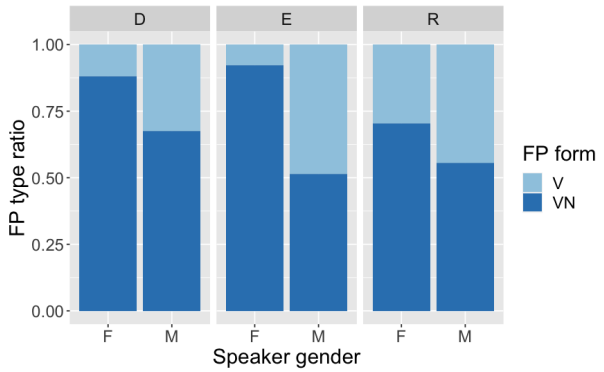


Figure 1: Ratio of vocalic (V) and vocalic-nasal (VN) filler particle (FP) forms across speakers with different language background (D: German heritage, E: monolingual, R: Russian heritage) and gender (F: female, M: male)

Overall the most frequent FP segmental form was vowel + nasal (VN: n=199) followed by vocalic forms (V: n=100). There were only few cases of nasal forms (N: n=3) and one instance of a glottalized variant; the latter two forms will be excluded in the following analysis. The preference for the VN variant is shared by all speaker groups with an average of 79% VN forms of FPs produced by speaker (D: 85%; R: 72%; E: 79% VN forms). Another observation relates to gender differences in the FP form preference. While all speakers prefer the VN form of FPs, female speakers show a higher use of VN forms and rarely V forms. The mean ratio of VN forms produced by speakers is 86% for female and 69% for male speakers. The gender of the speaker had a significant main effect on the mean VN proportion of FPs. A post-hoc Tukey pairwise comparison revealed a higher proportion of VN forms produced by female compared to male speakers (female: β = 17.13, SE = 6.89, df = 13.95, t = 2.48, p < 0.05). There seems to be a tendency for a less robust gender difference within the Russian heritage speaker group (cf. Figure 1). Yet no firm conclusions can be drawn from this small data set with only four female speakers in each group.

4.3. FP duration and phrasing

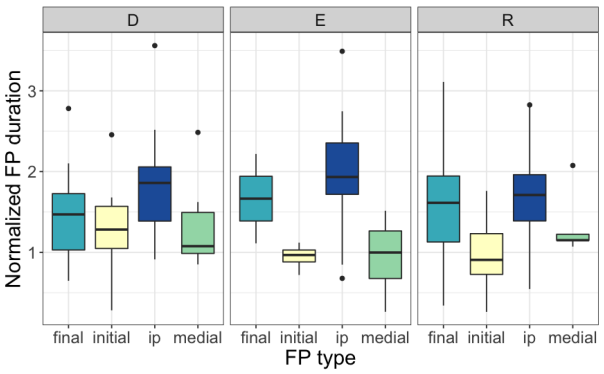


Figure 2: Normalized filler particle (FP) duration related to prosodic phrasing (ip: separately phrased) across speakers with different language background (D: German heritage, E: monolingual, R: Russian heritage)

The mean duration of FPs in the analysed data was 0.34 s with VN forms longer ($\bar{x}$= 0.38 s) than V forms ($\bar{x}$= 0.26 s). To compare FP duration across prosodic phrasing and language background the measured duration was normalized by the speech rate of the speaker. Both FP form (VN vs. V) and prosodic phrasing had a significant main effect on normalized FP duration. Post-hoc Tukey pairwise comparisons revealed a longer duration of FPs produced in final position and phrased separately compared to phrase initial position (final: β = 0.47, SE = 0.09, df = 287.29, t = 4.99, p<.001; initial: β = -0.57, SE = 0.08, df = 288.37, t = -7.51, p<.001) while separately phrase FPs (ip) were also longer than FPs in phrase medial position (ip: β = 0.33, SE = 0.12, df = 280.41, t = 2.83, p<.001). Additionally, a post-hoc Tukey pairwise comparison revealed a shorter duration of V forms compared to VN forms (V: β = -0.46, SE = 0.06, df = 289.81, t ratio = -7.90, p<.001). This is true for all speaker groups as is illustrated in Figure 2.

4.4. Filler pitch

Overall, the pitch level of FPs tends to be lower than the speaker's mean ($\bar{x}$= -1.35 st). The most frequent intonation contour produced on FPs was falling (n= 198; D: n= 72, 71%; E: n= 25, 54%; R: n= 101, 65%) while level (n= 49, D: n= 13, 13%; E: n= 4, 8%; R: n= 32, 21%) and rising contours (n= 56; D: n= 16, 16%; E: n= 17, 4%; R: n= 23, 15%) were comparably frequent. The slope in st per second was overall greater for rising contours ($\bar{x}$= 21 st/s) than for falling contours ($\bar{x}$= 12 st/s). Both language background and contour type are relevant factors for slopes of FPs. The linear mixed model for the slope of the pitch also showed a significant interaction for Russian HL speakers. Post-hoc Tukey pairwise comparisons revealed that FPs produced with a rising contour had a larger slope compared to FPs produced with falling intonation (falling: β = -6.31, SE = 1.82, df = 250.73, t ratio = -3.47, p < 0.01; level: β = -14.33, SE = 2.14, df = 253.65, t ratio = -6.69, p < 0.01) as well as larger slope of FPs produced with falling compared to level intonation (β = 8.02, SE= 1.48, df = 259.72, t ratio = 5.42, p < 0.01).

5. DISCUSSION

The results of this study illustrate FP productions by multilingual English speakers for the case of HSs living in the US with either German or Russian heritage background.

The analysed data does not provide evidence for an increased FP use due to higher cognitive load in bilingual speakers. The first hypothesis therefore has to be rejected contrary to prior findings [9, 10, 11]. This is in line, however, with their status as majority speakers of the analysed language.

Further evidence for this is provided in the analysis of segmental form and duration. Againn no group difference was found. FPs are produced by ML speakers of English similarly to other words, as suggested by [4]. As words they follow prosodic considerations such as pre-boundary lengthening [27] and are lengthened in phrase-final position or if phrased separately. The differences in duration found for V and VN forms can be related to the respective number of segments. The HL background does not influence the use of FPs as words in ML English.

Overall the data analysed in this study replicates the VN preference of FPs in English as well as the gender difference of VN preference in female speakers [14]. This is further evidence for the linguistic and cultural dominance in the ML of HS. That is, HS seem to acquire sociolinguistic and pragmatic aspects of language use in their ML. The second hypothesis, therefore, also has to be rejected. However, the observed difference between genders is less strong in the speakers of English with Russian HL. It would be insightful to also investigate their FP behavior in their HL to see whether this is a trend within this group which translates to the HL as well or whether it results from individual variation.

Prosodically the FPs in the analysed data were produced in a reduced manner, with low F0 and predominantly falling intonation. Unexpectedly, FPs were also frequently produced with rising contours. While in the distribution there was no effect of language background, speakers with Russian HL produced significantly larger slopes, especially with rises. This could be related to the influence of Russian on the intonation of these speakers' English due to typological differences in the two languages [28] or a transfer of intonational patterns from other discourse particles in Russian (e.g. *nu* [20]). The third hypothesis could, therefore, be partially confirmed. Further analysis is necessary to see whether the intonation of FPs reflects yet another area of cross-linguistic influence as shown in prior studies on bilingual intonation (e.g. [29]). The surprisingly frequent occurrence of rising FPs overall could alternatively be connected to discourse functions which are specifically used in these types of narrations.

6. CONCLUSION AND OUTLOOK

This study investigated the productions of FPs in ML English speakers with and without HL background. In the analyzed data speaker groups differed in the prosodic realisation of FPs which could be related to a possible influence of the HL. These findings could also be related to different signalling functions of FPs [30] used to different degrees by different speakers. The similarity in frequency and segmental form supports the native speaker status of HS in their ML [18]. Future research will address the FP use in the HS use of FPs in their HL hoping to provide a broader picture on FP use in bilingual speech.

7. ACKNOWLEDGEMENTS

Thanks to the phonetics group at Kiel University for helpful feedback and to the RUEG group for support and information on using their corpus data.

8. REFERENCES

[1] E. Shriberg, "To 'errrr'is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[2] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[3] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[4] H. H. Clark and J. E. FoxTree, "Using *uh* and *um* in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[5] G. Kjellmer, "Hesitation. in defence of ER and ERM," *English Studies*, vol. 84, no. 2, pp. 170–198, 2003.

[6] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous English speech," *Word*, vol. 15, no. 1, pp. 19–44, 1959.

[7] S. R. Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, vol. 2, pp. 51–81, 1973.

[8] M. Belz, *Die Phonetik von äh und ähm*. Springer Berlin Heidelberg, 2021.

[9] J. F. Kroll and T. H. Gollan, "Speech planning in two languages: What bilinguals tell us about language production," *The Oxford Handbook of Language Production*, pp. 165–181, 2013.

[10] N. H. De Jong, "Fluency in second language testing: Insights from different disciplines," *Language Assessment Quarterly*, vol. 15, no. 3, pp. 237–254, 2018.

[11] G. Gilquin, "Hesitation markers among EFL learners: Pragmatic deficiency or difference?" *Pragmatics and corpus linguistics: A mutualistic entente*, no. 2, pp. 119–149, 2008.

[12] M. M. de Boer and W. F. L. Heeren, "Cross-linguistic filled pause realization: The acoustics of "uh" and "um" in native Dutch and non-native English," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3612–3622, 2020.

[13] M. Polinsky, *Heritage languages and their speakers*. Cambridge University Press, 2018, vol. 159.

[14] E. K. Acton, "On gender differences in the distribution of um and uh," *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, 2011.

[15] O. Niebuhr and K. Fischer, "Do not hesitate!-unless you do it shortly or nasally: how the phonetics of filled pauses determine their subjective frequency and perceived speaker performance." in *Proc. of INTERSPEECH*, 2019, pp. 544–548.

[16] N. H. de Jong, "Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency," *International Review of Applied Linguistics in Language Teaching*, vol. 54, no. 2, pp. 113–132, 2016.

[17] S. Wong and V. Papp, "Transferability of non-lexical hesitation markers across languages: Evidence from te reo Māori-English Bilinguals," in *Proc. of the 26th IAFPA*, Huddersfield, 2018, pp. 35–36.

[18] H. Wiese, A. Alexiadou, S. Allen *et al.*, "Heritage speakers as part of the native language continuum," *Frontiers in Psychology*, vol. 12, 2022.

[19] S. Stepanova, "Some features of filled hesitation pauses in spontaneous Russian," in *Proceedings of ICPhS*, vol. 16, 2007, pp. 1325–1328.

[20] A. Kuosmanen, "On the relationship between the melodical structure and discourse functions of the particles nu and vot in Spontaneous Russian," in *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*, 1999.

[21] H. Wiese, A. Alexiadou, S. Allen *et al.*, "RUEG corpus," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.3765218

[22] H. Wiese, "Language situations: A method for capturing variation within speakers' repertoires," *Methods in Dialectology XVI*, vol. 59, pp. 105–117, 2020.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer. version 6.2.03," http://www. praat. org/, 2021.

[24] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 3, pp. 1399–1413, 2002.

[25] N. P. Himmelmann, M. Sandler, J. Strunk, and V. Unterladstetter, "On the universality of intonational phrases: A cross-linguistic interrater study," *Phonology*, vol. 35, no. 2, pp. 207–245, 2018.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using eigen and s4," 2014. [Online]. Available: https://CRAN.R-project.org/package=lme4

[27] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[28] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 2008.

[29] Y. Zuban, T. Rathcke, and S. Zerbian, "Intonation of yes-no questions by heritage speakers of Russian," in *Proc. of 10th International Conference on Speech Prosody*, 2020, pp. 96–100.

[30] T. Pistor, "Prosodic universals in discourse particles," in *Proc. of Speech Prosody*, vol. 178, Boston, USA, 2016, pp. 869–872.

SWITCHING BETWEEN PHONOLOGICAL BIASES IS NOT FREE: EVIDENCE FROM A MULTILINGUAL RECONSTRUCTION TASK

Adam A. Bramlett, Seth Wiener

Carnegie Mellon University
abramlet@andrew.cmu.edu, sethw1@cmu.edu

ABSTRACT

Language users develop phonological biases in response to the acoustic-phonetic and phono-lexical characteristics of the language being learned. English speakers, for example, develop a consonantal bias (C-bias) whereas Mandarin speakers develop a vocalic bias (V-bias). L1 Mandarin-L2 English multilinguals must switch between an L1 V-bias and an L2 C-bias during L1 and L2 lexical processing. This exploratory study uses the word reconstruction task to examine the potential switch-cost associated with two different phonological biases. Seventeen participants were asked to orally change non-words into words by replacing either the consonant or vowel. Half the participants first did the task in English (C-bias) followed by Mandarin (V-bias) while the other half did the task first in Mandarin (V-bias) followed by English (C-bias). A switch-cost was only found in the Mandarin results in which a V-bias was attenuated. Switching between phonological biases can therefore potentially affect lexical processes in multilinguals.

Keywords: Phonological biases, Switch-cost, Learning, Multilingualism, L2 speech processing

1. INTRODUCTION

All documented spoken languages have consonants and vowels [8]; for this reason, researchers posited that speech sounds must play language-independent roles during language processing [12]. The 'division-of-labor hypothesis' stated that vowels carry a higher functional load in syntactic and prosodic processing, while consonants carry a higher functional load in lexical processing [13]. However, developmental findings tell a more nuanced story, where a C-bias or V-bias is believed to develop in early infancy (e.g., [7], [3], [11], [18]).

For example, [6] used a word learning task to examine French infants' phonological biases. Infants were presented with two objects named by two non-words that involved a one-feature consonantal (e.g., /paʃ-baʃ/) or vocalic contrast (e.g., /ʃyl/-/ʃul/). French infants demonstrated a C-bias: they successfully learned non-words that had consonant contrasts but unsuccessfully learned non-words that had vowel contrasts. In contrast, Danish infants performing a similar task demonstrated a V-bias: they successfully learned words contrasted by a vowel rather than a consonant [7].

For adults, one of the many experimental tasks used to assess phonological biases is the reconstruction task [14]—the task used in the present study. In this task, participants hear a non-word, such as /wɛsk/, and must orally report a real word by changing a single consonant or vowel. For example, by changing the final consonant from /k/ to /t/ for the word /wɛst/, or by changing the vowel from /ɛ/ to /ɪ/ for the word /wɪsk/. Participants are also given a free choice condition in which they can choose either V- or C- substitution. The non-word in this task acts as a "perceptual template" [14] of multiple real words. Participants' correct reaction time (RT) and accuracy serve as the dependent variables. [14]'s English-speaking participants were both faster and more accurate in the vowel change condition (e.g., /wɛsk/ to /wɪsk/) than the consonant change condition (e.g., /wɛsk/ to /wɛst/), and changed vowels more often and faster than consonants in the free change condition. English (and Spanish [2]) listeners are biased towards preserving consonants and thus demonstrate a C-bias.

[21] extended the reconstruction task to Mandarin, thereby adding a fourth condition: tone change. Similar to V- and C- reconstruction, T- reconstruction requires the participant to change a tone to make a non-word into a real word. For example, /taɪ/-*dai2* is a non-word in Mandarin that can be changed into a real word by substituting a vowel (e.g., /tɤ/-*de2*), consonant (e.g., /laɪ/-*lai2*), or tone (e.g., /taɪ/-*dai4*). [21] found a V-bias for Mandarin listeners. That is, Mandarin listeners were least accurate and slowest in vowel reconstruction as compared to tone or consonant reconstruction.

Thus, listeners develop language specific phonological biases. L1 Mandarin [21] and

Cantonese [5] adult listeners, as well as infant L1 Danish [7] listeners, demonstrate a lexically related V-bias whereas L1 English [14], French [6], and Spanish [2] listeners demonstrate a lexically related C-bias. These biases are the result of the acoustic-phonetic and phono-lexical characteristics of the language environment: Danish consonants exhibit a high degree of lenition, which weakens the functional load of consonants [15] whereas Mandarin and Cantonese vowels carry lexically contrastive fundamental frequency information [4]. This tonal information plays a crucial role in lexical access, which increases the functional load of vowels (e.g., [10], [20]).

More recently, [19] examined whether L1 Mandarin listeners (L1 V-bias) demonstrate a new bias in their L2 English reconstruction behavior (i.e., an L2 C-bias). Participants performed the reconstruction task in their L2 and indeed showed faster RTs and lower error rates in the vowel change condition than the consonant change condition. [19] also tested L1 English-L2 Mandarin listeners engaged in classroom L2 learning. These L2 listeners had a reduced lexicon and limited experience with Mandarin; yet, they demonstrated a robust V-bias by changing consonants faster and more accurately than vowels. Lexically related phonological processing biases are not determined solely by a listener's L1, but rather, are at least partially determined by the phono-lexical features of the language.

Whereas the learning of phonological bias happens quickly in the L2, the processing cost of switching between V- and C- bias languages is unknown. In this exploratory study, we examine the effects of switching from a C-bias language to a V-bias language (and vice-versa) using the word reconstruction task. Here we ask two research questions:

1. To what extent is reconstruction accuracy affected by language task order?
2. To what extent is reconstruction reaction time affected by language task order?

We have no a priori reason to believe that doing the task in one language is qualitatively different than doing the task in another language. Naturally, performing the task in the L2 (here, English) may result in slower RTs and lower accuracy scores relative to the group performing the task in their L1. We are interested in comparing performance between subjects who perform the task first in English (C-bias) and second in Mandarin (V-bias) and subjects who perform the task in the opposite order: Mandarin-English. Does a switch-cost exist?

2. METHODS

2.1. Participants

Eighteen L1 Mandarin speakers (3 male; 15 female; mean age = 21.8; SD = 2.0) from Mainland China took part in the study. All participants had normal hearing and speech, had completed up to high school in China, and were studying at an American university. All participants reported Mandarin as the only Chinese language spoken or understood. One participant's data was removed for failing to follow the task's instructions.

2.2. Materials

For the English stimuli, 60 non-words were taken from [14]. Each non-word could be changed into a real word by either changing a vowel or a consonant. The position of the substitution was controlled with 35 consonant changes happening before vowel changes (e.g., /wɛmən/ to /lɛmən/ or /wʊmən/) and 25 vowel changes occurring before consonant changes (e.g., /wɛsk/ to /wɪsk/ or /wɛst/). An additional 70 non-words were created as fillers along with 12 practice items. For the Mandarin stimuli, 64 CV and CVV non-words were taken from [21]. Each non-word could be made into a real word by changing either consonant, vowel, or tone. In contrast to the English stimuli, the Mandarin stimuli always had consonant changes before the possible vowel changes. This is due to the limited syllable structure of Mandarin words [19]. 32 CVN nasal final words were used as fillers and 16 more were used as practice items. All stimuli were recorded at 44.1 kHz and 16-bit resolution in a sound attenuated booth. All stimuli, data, and R code are available on the Open Science Framework: https://osf.io/9zgd3/.

2.3. Procedure

Participants were tested individually in a lab setting using headphones. First, participants answered a language background questionnaire, then were given printed and oral instructions in the target language. Order of language was counterbalanced: nine performed the English version then the Mandarin version; nine performed the Mandarin then the English version. Within each language, order of the change conditions (e.g., C-, V-) was counterbalanced. In each change condition, participants were asked to change a speech sound so that a real word could be created. Participants were given 4 practice trials with feedback. Participants were told to focus on the word's sound and not the

spelling, and to say the word into a microphone as quickly and clearly as possible. Stimuli were presented using E-Prime [16]. Each trial was separated by a button click controlled by the participant with a 2 second ISI and a 10 second time limit per trial. Verbal responses and reaction times were recorded at word onset using Chronos [1]. Participants were unaware that they would be doing multiple conditions and the task in different languages. Oral responses were transcribed by two bilingual Mandarin-English speakers.

2.4. Results

~10.78% of all responses were removed due to timing out at 10 seconds (Mandarin: ~9.07%, English: ~12.31%). All single changes that were appropriate for the respective task were marked as correct. Responses that had more than one change (e.g., /wɛsk/ to /tɛst/) or had an incorrect change for the respective condition were marked incorrect. Three English words were removed from analysis for low accuracy (/mɪtən/, /lɒvəl/, /gruːn/). Median absolute deviation was used for removal of log transformed reaction times [9] removing 47 items. Mean reaction times of correct responses and error rates of each task can be found in Table 1. Table 1 shows that in the Mandarin task, on average, vowel changes were the slowest and had the highest error rate (i.e., listeners showed a V-bias); whereas, in the English task, on average, consonant changes were the slowest and had the highest error rate (i.e., listeners showed a C-bias).

Task Language	Reconstruction Condition	Mean RT(ms)	Mean % error
Mandarin	Vowel	4934.5	35.37
Mandarin	Tone	2562.1	19.23
Mandarin	Free	2888.0	19.14
Mandarin	Consonant	3637.0	29.04
English	Vowel	3237.0	30.49
English	Free	3029.1	34.36
English	Consonant	3348.7	45.90

Table 1: Mean correct reaction times (ms) and percent error for each task.

To analyze switch-cost, correct reaction time and accuracy were compared across language testing orders for English and Mandarin. To do this, we created a variable for the order in which the task was done (i.e., whether the English/Mandarin task was done first or second). For accuracy, (in)correct answers were analyzed as a binary variable. For reaction time, only correct reaction times were analyzed, following [14].

To test whether there is a switch-cost associated with doing the task in English (C-bias) or Mandarin (V-bias) and vice-versa, two generalized mixed effects models (GMEM) were built to compare accuracy and two linear mixed effects models (LMEM) were used to compare log transformed reaction time results in R [17]. Models were built using the *lme4* package. The models included two fixed effects: task order (first or second), condition (vowel, tone, consonant, free) and their interaction. Task order (first) and condition-consonant were set as reference levels. All variables were coded with effects coding. Random intercepts for item and participant were included, as were random slopes for conditions but these slopes were removed after non-convergence. An effect of condition would indicate a C- or V-bias whereas an effect of task order and/or an interaction would indicate a switch-cost.

Accuracy results from the English task GMEM found null results at an alpha-level of .05. Accuracy results from the Mandarin task GMEM found a significant effect in the free (β = 1.00, SE = .42, z = 2.35, p < .05), tone (β = .99, SE = .43, z = 2.30, p < .05), and vowel conditions (β = -1.44, SE = .41, z = -3.45, p < .001), indicating more accurate reconstruction for free and tone conditions and less accurate vowel reconstruction in comparison to consonants. Additionally, a significant interaction was found between the vowel condition and task order (β = 1.10, SE = .47, z = 2.33, p < .05), indicating that there is a switch-cost, which results in more accurate responses when the Mandarin vowel task is performed after the English tasks.

RT results from the English task LMEM found null effects. The Mandarin RT LMEM found effects for the free condition (β = -.35, SE = .07, t = 5.17, p < .001), tone condition (β = -.57, SE = .69, t = -8.21, p < .001), and vowel condition (β = .79, SE = .08, t = 10.56, p < .001), indicating that vowel responses had significantly slower RTs while tone and free responses had significantly faster RTs (relative to consonant responses). Additionally, an interaction between task order and the free condition was found (β = -.18, SE = .09, t = -2.17, p < .05), indicating that doing the free condition in Mandarin second allowed for faster reaction times.

3. DISCUSSION

Like the results of [19] and [21], our Mandarin reconstruction results offer additional evidence of Mandarin speakers demonstrating a V-bias: L1 Mandarin speakers responded slower and less

accurately towards vowel reconstruction than consonant reconstruction. Unexpectedly, the C-bias that was reported in [14] was not replicated. Numerically, however, English consonant accuracy (and RT) was lower than vowel accuracy (and slower than vowel RT). Keeping in mind that our participants were doing the English task in their L2, their means (Table 1) are roughly comparable with [14]'s L1 English participants' means (vowel: 2,217 ms, 28%; consonant: 2,412 ms, 42%). The error rates are nearly identical while our L2 participants were about a second slower, on average. The null effect in our English task is most likely due to the low power of this study (17 participants). Additionally, 15% of the variability found in the English accuracy model was due to participant variability. We are in the process of resuming in-person testing in order to obtain a larger sample.

With respect to switch-cost, we found interactions in the Mandarin accuracy and RT models. When performing the Mandarin reconstruction task after the English reconstruction task participants became more accurate at vowel changes and faster at free choice changes. Figure 1 shows this switch-cost in terms of the interaction between task order and vowel condition accuracy. Figure 2 shows this switch-cost in terms of the interaction between task order and free condition RT. These figures show model outputs by displaying the standard error with horizontal blocks and a vertical center line which marks coefficient estimates. Teal and yellow bars represent coefficient estimates that are lower (teal) and higher (yellow) than reference levels (i.e., condition: consonant, task order: Mandarin-first), respectively. Significant results are marked with red asterisks ($p < .05$ = *, $p < .001$ = ***). In general, doing a task second appears to lower reaction time in the free choice condition (a practice effect), which is demonstrated by the significant effects found in the Mandarin models (Fig 2) and numerical difference in the English task. In contrast, the significant interaction between vowel condition and task order in accuracy suggests that the V-bias demonstrated in Mandarin tasks is somewhat attenuated when a participant first does the English tasks (Fig 1). Doing the English task second leads to a numerically greater error in the vowel change condition, though this effect is not significant at an alpha-level of .05.

Additionally, eight post-hoc analyses were performed by task order and language for both accuracy and RT. In Mandarin reconstruction, vowels were significantly slower despite task order (i.e., V-bias persisted across task order). In contrast, vowels were only significantly less accurate

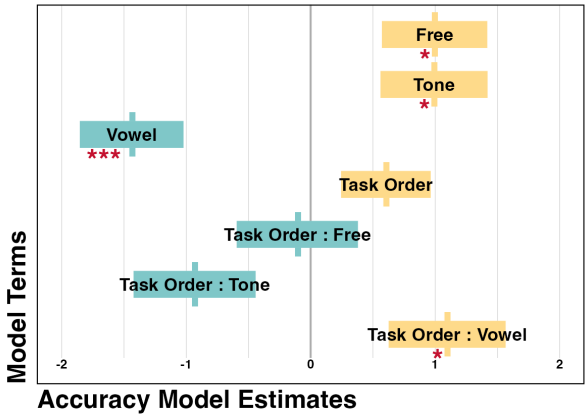


Figure 1: Mandarin accuracy model output

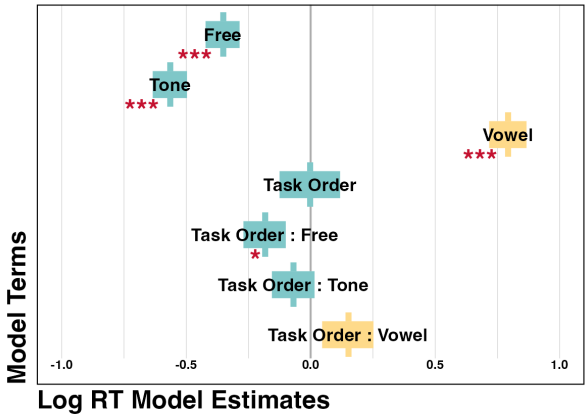


Figure 2: Mandarin Log RT model output

when doing Mandarin first. For English-first participants, however, no V-bias was found, which may indicate an increase in flexibility for lexical retrieval. English RT post-hoc analyses showed null effects. Accuracy results in English-first condition, however, indicated that vowels are more accurate than consonants (i.e., showing a C-bias; [14]). The increased accuracy for vowels in the English-first group is lost in the post-hoc analysis for English-second where no significant difference is found. This change indicates a potential loss of flexibility for vowel-driven lexical access for those that do Mandarin first then English.

In sum, this exploratory study demonstrated a processing cost for multilinguals switching between phonological biases. Listeners demonstrate language-specific patterns of biases, and yet these patterns can change given the previously processed language. To what degree these findings impact theories of multilingual lexical representations and processing, and whether these findings can improve adult language learning remain open questions for future research to address.

REFERENCES

[1] D. L. Babjack, B. Cernicky, A. J. Sobotka, *et al.*, "Reducing audio stimulus presentation latencies across studies, laboratories, and hardware and operating system configurations," *Behav. Res. Methods*, vol. 47, pp. 649–665, 2015.

[2] A. Culter, N. Sebastián-Gallés, O. Soler-Vilageliu, and B. Van Ooijen, "Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons," *Mem. Cognit.*, vol. 28, no. 5, pp. 746–755, 2000.

[3] C. Floccia, T. Nazzi, C. Delle Luche, S. Poltrock, and J. Goslin, "English-learning one- to two-year-olds do not show a consonant bias in word learning," *J. Child Lang.*, vol. 41, no. 5, pp. 1085–1114, 2014.

[4] J. Gandour, "Tone perception in far eastern-languages," *J. Phon.*, vol. 11, no. 2, pp. 149–175, 1983.

[5] D. M. Gómez, P. Mok, M. Ordin, J. Mehler, and M. Nespor, "Statistical speech segmentation in tone languages: The role of lexical tones," *Lang. Speech*, vol. 61, no. 1, pp. 84–96, 2018.

[6] M. Havy and T. Nazzi, "Better processing of consonantal over vocalic information in word learning at 16 months of age," *Infancy*, vol. 14, no. 4, pp. 439–456, 2009.

[7] A. Højen and T. Nazzi, "Vowel bias in danish word-learning: Processing biases are language-specific," en, *Dev. Sci.*, vol. 19, no. 1, pp. 41–49, 2016.

[8] P. Ladefoged and S. F. Disner, *Vowels and Consonants*, 3rd ed. Chichester, England: Wiley-Blackwell, 2012.

[9] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[10] J. G. Malins and M. F. Joanisse, "The roles of tonal and segmental information in mandarin spoken word recognition: An eyetracking study," *J. Mem. Lang.*, vol. 62, no. 4, pp. 407–420, 2010.

[11] T. Nazzi and S. Poltrock, "Consonant bias in the use of phonological information during lexical processing: A lifespan and cross-linguistic perspective," in *Speech perception and spoken word recognition*, J. Gaskell, Ed., London, England: Psychology Press, 2016, pp. 37–54.

[12] T. Nazzi and A. Cutler, "How consonants and vowels shape spoken-language recognition," *Annu. Rev. Linguist.*, vol. 5, no. 1, pp. 25–47, 2019.

[13] M. Nespor, M. Peña, and J. Mehler, "On the different roles of vowels and consonants in speech processing and language acquisition," *Lingue e linguaggio*, vol. 2, pp. 203–230, 2003.

[14] B. van Ooijen, "Vowel mutability and lexical selection in english: Evidence from a word reconstruction task," *Mem. Cognit.*, vol. 24, no. 5, pp. 573–583, 1996.

[15] N. Pharao, "Plosive reduction at the group level and in the individual speaker," in *ICPhS*, XVII, Hong Kong, 2011.

[16] Psychology Software Tools, *E-Prime*, Pittsburgh, PA.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2022. [Online]. Available: https://www.r-project.org/.

[18] L. Singh, H. H. Goh, and T. D. Wewalaarachchi, "Spoken word recognition in early childhood: Comparative effects of vowel, consonant and lexical tone variation," *Cognition*, vol. 142, pp. 1–11, 2015.

[19] S. Wiener, "Second language learners develop non-native lexical processing biases," *Bilingualism*, vol. 23, no. 1, pp. 119–130, 2020.

[20] S. Wiener and K. Ito, "Do syllable-specific tonal probabilities guide lexical access? evidence from mandarin, shanghai and cantonese speakers," *Lang. Cogn. Neurosci.*, vol. 30, no. 9, pp. 1048–1060, 2015.

[21] S. Wiener and R. Turnbull, "Constraints of tones, vowels and consonants on lexical selection in mandarin chinese," *Lang. Speech*, vol. 59, pp. 59–82, 2016.

ASSESSING THE EFFECTS OF AUDITORY FEEDBACK PERTURBATION ON PERCEPTION OF SIMILAR VOWELS IN A DIFFERENT LANGUAGE

Madeleine Rees[1], Abbie Bradshaw[2], Matthew H. Davis[2], Brechtje Post[1]

[1]Dept. of Theoretical & Applied Linguistics, University of Cambridge [2]MRC Cognition and Brain Sciences Unit, University of Cambridge
mljr2@cam.ac.uk

ABSTRACT

Auditory feedback perturbation of vowel formants in real time causes the speaker to perceive a change in their own voice, and to change their production to offset the perceived error. This response, known as adaptation, has been shown to influence the perceived category boundary of the perturbed vowel for monolingual English speakers. Here, we added a bilingual dimension by applying formant perturbation to tokens of "head" and "dedo" produced by 23 English-Spanish bilinguals, which moved the intended vowel /ɛ/ or /e/ closer to /i/. The feedback perturbation task was preceded and followed by a forced choice vowel identification task in the language not used in the production task. Results indicate significant adaptation responses to formant perturbation in English dominant speakers, producing English words. However, we observed null effects of auditory feedback perturbation on categorical perception of similar vowels in the other language. The roles of vowel inventory and experimental design are discussed.

Keywords: speech perception, speech production, feedback perturbation, bilingualism

1. INTRODUCTION

The relationship between speech perception and speech production is unclear and highly-debated. One paradigm which can uncover functional links between production and perception is auditory feedback perturbation (AFP). AFP is an experimental technique in which a computer modifies the acoustic cues of produced speech in near-real-time (typically less than 50ms delay). The altered feedback causes the speaker to perceive a change in their own voice and adjust their speech to offset the effects of the perturbation, a process termed adaptation [5]. For example, a speaker instructed to say "head" and given a downshift perturbation in F1 would hear themselves producing an aberrant vowel with a lowered F1 (i.e. "hid"). Over repeated trials, the adaptation response would consist of speakers producing a sound with a higher F1, closer to "had", opposing the perturbation.

The adaptation response is predicted by models like DIVA [16] and SFC [6], in which a difference between expected speech acoustic cues (in accordance with an auditory target for speech production) and observed speech acoustic cues leads to corrective motor commands, in turn generating the adaptation response.

Feedback-induced changes in production of vowel formants have been found in turn to influence vowel category boundaries in a vowel perception task [9]. Adaptation to a downshift of F1 of the /ɛ/ vowel (in the word "head") was associated with an increase in hearing "had" rather than "head" in a follow-up 2AFC vowel identification task. This finding suggests that adaptation to the altered feedback is associated with a reconfiguration of the perceptual boundaries of the perturbed vowel and surrounding vowels. By extension, the change in production induced by AFP appears to influence phonemic representations and auditory targets of the perturbed and nearby vowels.

Production and perception tasks in [9] were carried out in English. So far, no study has performed this experiment across languages with bilingual speakers. Testing bilingual transfer of adaptation allows for further exploratory investigation of whether the adaptation response modifies the perceptual boundary of solely the perturbed vowel, or reconfigures the vowels in the surrounding vowel space more generally and across linguistic dimensions. Nearby vowels in the other language with similar acoustic characteristics may be susceptible to the effects of adaptation on categorical perception if they are also reconfigured during speech motor learning. However, other research indicates that speech motor learning is 'local' [17]. In this study, speech motor learning leading to changes in production of the /iæ/ diphthong did not generalise to different words. This result suggests that the effects of adaptation are highly specific and may not extend to affect similarly close vowels in another language. However, the study in [17] was a jaw perturbation study rather than an AFP study, so the two paradigms are not directly comparable. More information about the reconfiguration of phonemic representations during AFP is a useful addition to speech production models like DIVA and SFC.

The vowel inventories of English and Spanish contain a similar but not identical mid/mid-high front vowel (/ɛ/ and /e/ respectively). Therefore, an appropriate speaker group in which to compare cross-linguistic perceptual boundary change is English-Spanish bilinguals. Here, we report a null effect of feedback perturbation of English /ɛ/ on categorical perception of the /e/-/a/ continuum in Spanish, and of perturbation of Spanish /e/ on perception of the /ɛ/-/æ/ continuum in English.

2. METHODS

2.1. Participants

Participants were 23 English-Spanish bilinguals (12 (Southern British) English L1, 11 Spanish L1, aged 19-47). All of the Spanish L1 participants spoke Peninsular Spanish, except for one Mexican and one Chilean Spanish-speaking participant. Language dominance was calculated using the Bilingual Language Profile [1]: henceforth *language-dominant* will be used to refer to each speaker group. Ethical approval was granted by the departmental ethics committee. Participants were compensated for their time.

2.2. Production experiment design

The experiment took place in the soundproof booth of the Phonetics Lab of the University of Cambridge. Feedback perturbation was configured in MATLAB (MathWorks) using Audapter [3] to introduce a consistent perturbation of -130 mel in F1 and +130 mel in F2.

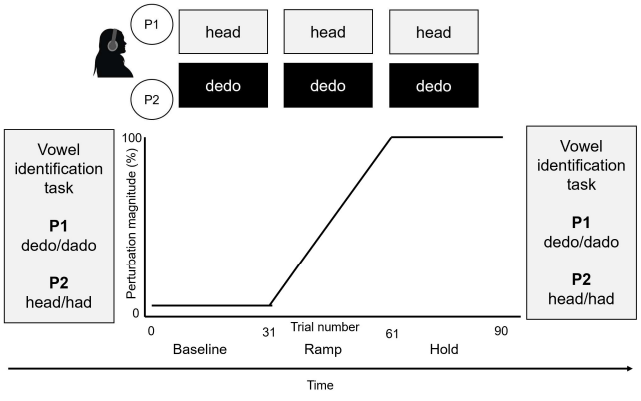


Figure 1: Experimental design showing task order over time and perturbation magnitude over each phase of the production test. Session 1 (noted as P1) was performed first (perception-production-perception) followed by session 2 (P2). The language of the feedback perturbation task was the opposite of the language of the vowel identification task (i.e. task language was switched twice in each session).As participants produced speech, they heard their own voice presented binaurally through headphones in real time. The participant's own voice was mixed with 60dB masking noise to reduce bone conduction.

The English-Spanish task was conducted first. After a pre-test assessing categorical vowel perception in Spanish, the experiment started with a baseline phase in which participants produced 30 "head" utterances with no formant perturbation. Subsequently, a ramp phase introduced the perturbation gradually over 30 trials until the maximum perturbation was reached. Then, in the Hold phase, the maximum perturbation was held constant over 30 trials. At this point, the participant undertook a categorical perception test in Spanish. The Spanish-English task followed the same protocol, with switched language order (see Fig. 1) and with the production stimuli being "dedo". For 16 participants, both English-to-Spanish and Spanish-to-English task runs were conducted in the same session, with a 10–15-minute break between tasks to wash out the effects of the first perturbation. For 6 participants, the two task runs were conducted in separate sessions typically a week apart, due to logistical constraints. Instructions were given in the perturbed language.

Speech was recorded with a Sennheiser cardioid microphone approximately 10cm from the speaker. The audio signal was processed with a SoundDevices MixPre6 integrated soundcard and mixer connected to an ASUS Zenbook Pro laptop via USB-C. The average latency of the Audapter setup was between 16-26ms, with <50ms typically considered acceptable latency for feedback perturbation experiments [15].

2.3. Categorical perception tests

Vowel stimuli in English and Spanish were taken from recordings of "head", "had", "dedo" *(finger)* and "dado" *(dice)* produced by a male and female native speaker of British English and a male and female native speaker of Peninsular Spanish. Two continua of 10 resynthesised vowel tokens between "head" and "had" and "dedo" and "dado", with identical durations, were generated using Tandem-STRAIGHT morphing [7]. Morphing percentages ranged from 0% to 100% in equidistant 11.1% intervals.

Vowel perception tests were run in MATLAB using PsychToolBox-3 [2]. The 10 vowel stimuli were presented 10 times each binaurally through headphones with an ITI of 1s. The gender of the stimuli speaker was matched to the gender of the participant, to ensure that the perception stimuli sounded as similar as reasonably possible to the

participant's own production stimuli. Responses were assessed by keyboard press.

2.4. Production data analysis

Adaptation response magnitude was assessed by projecting the vector of each token in the Hold phase onto the vector directly opposing the perturbation (i.e. the vector of perfect compensation).

2.5. Perception data analysis

Categorical perceptual boundary (CPB) midpoints (i.e., the point of subjective perceptual equality) between vowels were calculated in R [14]. Logistic functions were fitted to perception test response data. CPB midpoint was calculated as the number of the vowel token (between 1-10) at which the proportion of responses was 0.5 (i.e. the point of subjective equality).

3. RESULTS

3.1. Feedback perturbation task results

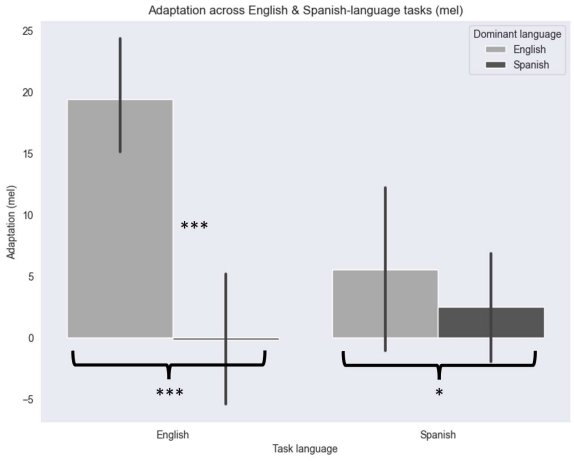


Figure 2: Adaptation responses to feedback perturbation in English (left) and Spanish (right). Asterisks indicate statistical significance at <0.001 (***) and marginal 0.053 (*).

Adaptation magnitude significance (i.e., difference from 0) was assessed with one-tailed two-sided t-tests in R. English-dominant speakers showed significant positive adaptation in English (t = 8.4046, df = 325, p = <0.001) but non-significant adaptation in Spanish (t = 1.6628, df = 328, p = 0.097). Spanish-dominant speakers showed non-significant adaptation in English (t = -0.068297, df = 297, p = 0.9456) and non-significant adaptation in Spanish (t = 1.0525, df = 320, p = 0.29). Linear mixed effects modelling using *lmerTest* [8] in R was used to determine the statistical significance of the differing adaptation magnitudes across the two task languages and speaker groups. *Task language (TL)*, *dominant language (DL)* and *task language*dominant language (TL*DL)* were fixed effects, with *participant* as a random effect. TL of Spanish was associated with significantly lower adaptation magnitude (t = -4.827, df = 1250, p = < 0.001). There was a significant interaction between TL and DL (t = 5.078, df = 1252.325, p = <0.001). Pairwise comparisons were calculated using *emmeans* [11] and corrected using the multivariate t-distribution method. English-dominant speakers showed greater adaptation response magnitude in English than in Spanish (t = 4.218, df = 1308, p = 0.0001). Spanish-dominant speakers showed greater adaptation response magnitude in English than in Spanish, although the difference was marginally non-significant (t = 2.42, df = 1254.4, p = 0.053).

3.2. Categorical vowel perception task results

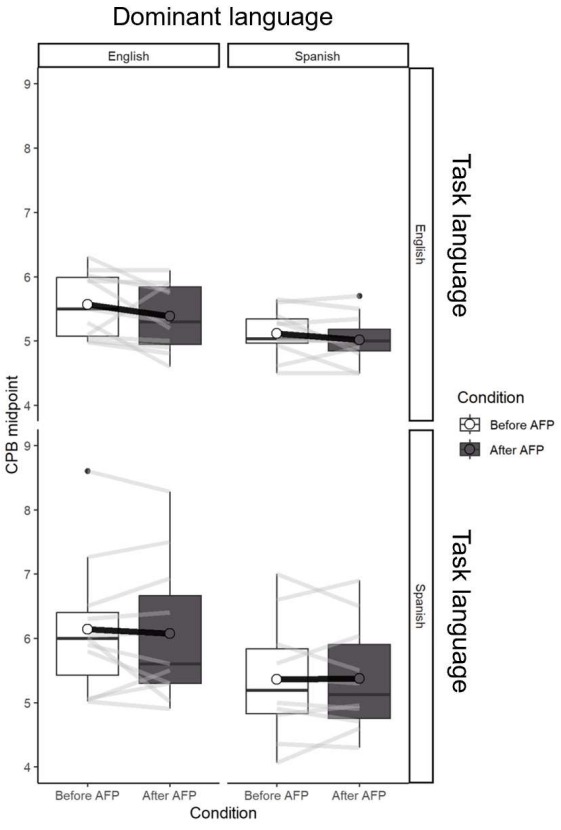


Figure 3: Left: Categorical perceptual boundary midpoints in English before and after feedback perturbation in Spanish-dominant group, in English (above) and Spanish (below). Right: Categorical perceptual boundary midpoints in Spanish before and after feedback perturbation in English-dominant group, in English (above) and Spanish (below).

Following calculation of category perceptual boundaries (CPB) for each continuum, participant and test phase, linear mixed effects modelling was used to assess the statistical significance of the change i

n the perceptual boundary location before and after A FP in each language (see Fig. 3). Fixed effects were *c ondition* (before or after AFP), with participant as a r andom effect. Overall, the effect of condition was no n-significant (t = 0.536, df = 63.21264, p = 0.594). U ltimately, even one robust adaptation response was n ot associated with statistically significant change in t he location of the categorical boundary.

4. DISCUSSION

Only one feedback perturbation condition, English-dominant speakers producing English, led to expected changes in formant frequencies indicative of adaptation. This result suggests that while compensatory responses to altered feedback are apparent for vowels in the dominant language, the direction of adaptation is less reliable in a second language, and less reliable for an otherwise comparable vowel in Spanish than in English.

These differences between Spanish and English support previous research in suggesting that adaptation is stronger at category boundaries [12]. Unlike English-dominant speakers, Spanish speakers showed no significant adaptation in their dominant language; although formant changes were in the expected direction, effect sizes were significantly reduced and non-significant. This result hints that the vowel inventory of the task language may influence adaptation, alongside language dominance. The English vowel inventory is denser than the Spanish vowel inventory: in a more crowded vowel space, the same feedback perturbation may have a more pronounced effect since greater sensitivity to auditory targets and boundaries must be used to maintain differences between vowel categories. An effect of density also supports a result in which both English and Spanish speakers opposed the perturbation in their dominant languages, but only the English speakers did so significantly.

In none of our conditions did we observe any reliable change in categorical perception boundaries, even though participants produced robust adaptation responses in one of the conditions (English-dominant speakers producing English words). These null effects suggest that perceptual change previously shown to stem from speech motor learning may be restricted to the perturbed vowel itself and not applied to a similar vowel in the surrounding vowel space.

This result has implications for the perception-production relationship. Post-AFP after-effects in speech production ([4],[9],[10]) indicate that adaptation to perturbation persists beyond the removal of altered feedback. Meanwhile, the lack of after-effects in speech perception observed here, in contrast to monolingual results in [9], suggest either: (1) that perception is not necessarily altered by a change in production, and therefore that speech production representations are not a key source of information in speech perception, or (2) that the underlying representations are sufficiently category-specific that transfer between even highly similar vowels in L1 and L2 is absent. Further research will be required to disentangle these two possibilities.

It is possible that the experimental characteristics and task order (English tasks first) had a detrimental effect on adaptation in the later Spanish-language tasks, and that this lack of adaptation response was associated with a subsequent lack of change in the categorical boundary. Adaptation after-effects from feedback perturbation can linger throughout the washout phase [13] and may have required a more targeted task to ensure that the after-effects of the perturbation were fully eradicated before starting the new block.

In order to correct for this possibility, we are currently collecting more data with a different task design: categorical perception data is only collected in one language (the unperturbed language) with one week in between different language sessions. This weeklong break between AFP sessions should ensure that adaptation after-effects are completely washed out before starting a new task. We are also collecting more data with an improved counterbalancing, where the first perturbed language (English or Spanish) varies for odd and even-numbered participants. This improved experimental design will allow us to more accurately capture and assess cross-language transfer in perception and production and determine whether these processes are present.

5. CONCLUSIONS

The findings shown here provide evidence for a mediating role of task language in adaptation to formant perturbations. Furthermore, this experiment did not find a significant effect on categorical perception boundaries after a feedback perturbation task in the other language. This result held regardless of task language or the speaker's dominant language. It is possible that this result arises due to a disconnection between production change and perceptual boundary change when the test crossed languages, even in a condition where adaptation magnitude was significantly higher than 0. These findings suggest that changes in production of one vowel do not automatically entail changes in categorical perception of nearby vowels in another language.

6. REFERENCES

[1] Birdsong D., Gertken L. M., Amengual M. (2012, January 20). Bilingual Language Profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin [online]*. URL: https://sites.la.utexas.edu/bilingual/

[2] Brainard, D. H. 1997. The Psychophysics Toolbox. Spatial Vision 10, 433-436.

[3] Cai, S., Boucek, M., Ghosh, S.S., Guenther, F.H., Perkell, J.S. 2008. A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In: *Proceedings of the 8th Intl. Seminar on Speech Production, Strasbourg, France, Dec. 8 - 12, 2008*, 65-68.

[4] Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., & Eisner, F. 2019. Consistency influences altered auditory feedback processing. *Quarterly Journal of Experimental Psychology* 72(10), 2371-2379.

[5] Houde, J. F., Jordan, M. I. 1998. Sensorimotor adaptation in speech production. *Science* 279(5354), 1213-1216.

[6] Houde, J. F., & Nagarajan, S. S. 2011. Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*, 82.

[7] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3933-3936.

[8] Kuznetsova A, Brockhoff P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26.

[9] Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., Ostry, D. J. 2014. Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience* 34(31), 10339-10346.

[10] Martin, C. D., Niziolek, C. A., Duñabeitia, J. A., Perez, A., Hernandez, D., Carreiras, M., Houde, J. F. 2018. Online adaptation to altered auditory feedback is predicted by auditory acuity and not by domain-general executive control resources. *Frontiers in Human Neuroscience* 12, 91.

[11] Lenth, R. 2022. _emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.8.0, <https://CRAN.R-project.org/package=emmeans>.

[12] Niziolek, C. A., Guenther, F. H. 2013. Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *Journal of Neuroscience* 33(29), 12090-12098.

[13] Parrell, B., Agnew, Z., Nagarajan, S., Houde, J., Ivry, R. B. 2017. Impaired feedforward control and enhanced feedback control of speech in patients with cerebellar degeneration. *Journal of Neuroscience* 37(38), 9249-9258.

[14] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[15] Stuart, A., Kalinowski, J., Rastatter, M. P., Lynch, K. 2002. Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America* 111(5), 2237-2241.

[16] Tourville, J. A., & Guenther, F. H. 2011. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952-981.

[17] Tremblay, S., Houle, G., Ostry, D. J. 2008. Specificity of speech motor learning. *Journal of Neuroscience* 28(10), 2426-2434.

L2-INDUCED PHONETIC DRIFT IN L1 POLISH VOWEL PRODUCTIONS

Ewelina Wojtkowiak

Adam Mickiewicz University in Poznań, Poland
ewelina.wojtkowiak@amu.edu.pl

ABSTRACT

An acoustic study sought to investigate the effects of intensive phonetic instruction in L2 English on L1 Polish productions of vowels /ɛ/ and /a/ in order to attest possible effects of L2-induced phonetic drift. The data from first year students of the English programme who took part in a longitudinal experiment were compared with comparison groups from second and third years as well as a group of monolinguals.

The results showed some lowering of both /ɛ/ and /a/ as training progressed, peaking in the productions of second year students. Some fronting was found in /ɛ/ but no effects on the advancement of /a/ were observed. Overall, the Polish vowels appeared to move towards more peripheral positions, possibly to accommodate the new L2 categories being acquired in the common phonological space.

Keywords: phonetic drift, Polish, vowels, Speech Learning Model, speech production

1. INTRODUCTION

The effects that one's native language (L1) can exert on the pronunciation in one's second language (L2) have been studied quite thoroughly. The opposite, however, that is the influence that L2 may have on the native productions has only been brought to the forefront of attention in the last decade or so, beginning with a series of studies authored by Charles B. Chang [1]. He coined the term *phonetic drift* to refer to short-term changes in the acoustics of L1 resulting from recent exposure to L2. Ever since then, more and more language pairs have been investigated, with the purpose of learning which acoustic parameters, and to what extent, are subject to drift.

As a result, L2-induced drift in L1 has been attested in both production and perception. Production experiments on consonantal drift have focused on parameters such as VOT (e.g., [2], [3]), pitch at vowel onset [2], and F1 and F2 values of vowel segments [4]. In turn, perception studies looked mostly at perceptual shifts (e.g., [5], [6]), and reaction times [7].

Furthermore, studies have shown that phonetic drift can be expected in both L2-immersion (e.g., [8], [9]) and L1-dominant (e.g., [10], [11]) settings, and in both novice [2] and proficient [12] learners.

In general, the effects of learning L2 on L1 Polish productions remain understudied. The few existing studies have investigated the effects of multilingualism, rather than bilingualism, on L1 [13], [14]. To date, only two studies ensured that no possibly confounding effects of Ln were present in their data [15], [16], and those focused on stop consonant productions. The present experiment sought to fill the research gap with regards to the effects on learning L2 English on L1 Polish vowels – to establish whether they would be attested in this language pair.

2. PREVIOUS STUDIES ON VOCALIC DRIFT IN L1

The overview of selected previous studies on vocalic drift can be divided into two parts, depending on the context of L2 acquisition.

Beginning with L2-immersion environment, [2] investigated the productions of 19 American English speakers enrolled in a language course in South Korea (i.e. novice learners of Korean). While the differences were subtle, English vowels were found to drift to approximate Korean norms, with respect to vowel height and but not advancement. Moreover, the drift effects were observed to be systemic, thus targeting the entire vocalic inventory rather than individual segments. [13] studied the vowel productions of 32 Japanese-English bilinguals (16 adults and 16 children) living in Texas, USA for a year and compared them with a control group of monolingual English speakers from the same area. By means of a picture naming task the authors obtained acoustic data which showed that while no changes over time were detected in the adult learners' production of Japanese vowels, Japanese children showed drift effects for the vowels /i/ and /a/. Importantly, it was the group of the children that improved most in their L2 productions, while no changes in the adults' productions of English vowels between the two testing times were found. This suggests that the broader the experience in the L2, the more salient its influence on the L1.

In L1-dominant settings, [17] studied the productions of Catalan-Spanish bilinguals and the extent to which extensive exposure to L2 Spanish (in particular its one vowel /e/) influences the correct

realisation of the Catalan /e/-/ɛ/ contrast. The authors observed that the group of speakers with high weekly usage of Catalan produced the difference between this particular vowel pair more robustly (in minimal pairs, cognates, and non-cognates alike) than the speakers with higher degree of exposure to Spanish and lower weekly usage of Catalan. Therefore, greater experience in Spanish contributed to more noticeable changes in L1 Catalan production. [18] compared early, mid, and late Quichua-Spanish bilinguals and the effects of Spanish vowels on their L1 Quichua /i, a, u/. The results of a delayed repetition task revealed that the speakers who successfully acquired L2 vowels showed more noticeable drift effects in vowel height. [19] additionally looked at whether articulatory training in L2 has any effects on L1 vowel production. In order to do so, 20 monolingual L1 French speakers were trained on two non-native vowels: Danish /ɔ/ and Russian /ɨ/. Even brief phonetic training on novice learners triggered the occurrence of phonetic drift in L1 vowel formants.

3. METHOD

3.1. Materials

Polish is a language with a relatively small vocalic inventory: /a, ɛ, i, ɨ, ɔ, u/, in contrast to a relatively rich vowel system in English. The Polish word list consisted of 121 words in total, out of which 54 were included in the analysis with the rest of them being fillers. The /b, d, ɡ/ and /p, t, k/ items were mono- and di-syllabic words, followed by vowels [a, ɛ, ɔ]. The dataset included 17 labial, 19 coronal, and 18 velar onsets. 29 of them were voiceless while 25 of them voiced. With a total of 65 speakers (out of whom 20 were recorded three times), a database of 5670 Polish recorded items was comprised. In this paper we focus on the words where the voiced and voiceless plosives were followed by the unrounded vowels /ɛ/ and /a/.

3.2. Participants

The participants that have taken part in the present study can be divided into two major groups: the main group of students taking part in the longitudinal study (henceforth 1BA[1]; N=20) and the comparison group, comprising three smaller subgroups: second year students (=2BA, N=15), third year students (=3BA, N=15), as well as Polish "quasi-monolinguals" (N=15). The choice of the population of Polish English majors is by no means accidental. Their motivation to acquire native-like proficiency in English is assumed to be high – they are studying to use English professionally at work, be it teaching, translation, or other endeavours.

The group of students were subject to intensive phonetic training over the course of their studies. 1BA students were taking part in practical phonetics classes (3hrs per week; mostly drilling exercises devised to master all segments of British/American English, taught by trained in phonetics Polish teachers) and theoretical courses in English phonetics and phonology (2hrs per week). Prior to their enrolment in the English programme, they had never had any experience with pronunciation training.

2BA students, having undergone 1BA courses, were enrolled in practical phonetics classes in their second year (1.5hrs per week), where they were trained on phonostylistics and intonation.

3BA students have attended all 1BA and 2BA phonetics courses, but while still taking courses in English in their third year, they were no longer being actively trained in pronunciation.

Polish "quasi-monolinguals" comprised speakers who do have a history of learning foreign languages in school, but they claim not to be fluent in any of them, at least as far as production is concerned. They could, then, be described as "functional monolinguals" [20].

Proficiency in any other language was ruled out on the basis of a language background questionnaire administered in all of the groups.

3.3. Procedure, acoustic measures, and statistical analysis

1BA students were the only that have participated in the longitudinal experiment. They were tested three times during the first year of their university education. The sessions were held in October (henceforth T1), February (=T2), and June (=T3). The first batch of recordings was made within the first two weeks of October. Therefore, it can be assumed that the session was held early enough for the effects of phonetic training not to have affected the students' productions yet. The second session was held in February, during the winter exam session, while the third took place in the last two weeks of June, right before the summer exam session, therefore towards the end of the first, second, and third year of studies for 1BA, 2BA, and 3BA, respectively.

The recordings were made in a sound-attenuated booth at a Polish university, directly onto laptop, using a condenser microphone and a USB interface. The words were shown to the participants on a

[1] "BA" refers to the programme in which the students were enrolled as they studied to obtain their Bachelor of Arts degree.

monitor via PowerPoint slides in a pseudo-randomised order (the same for each participant), with the pace controlled by the researcher sitting in another room.

In the case of the monolingual group, the recordings were made in a quiet classroom in the private language school which they were attending. They were recorded directly onto laptop, using a head-mounted microphone and a USB interface.

As far as the acoustic measurements go, the vowel was measured in Praat [21] from the onset of voicing associated with vowel production (thus, excluding burst and release noise) until the point in which F2 and F3 were no longer visible. Using a Praat script, the mean values of F1 (difference between F1- f_0; Bark normalised) and F2 (difference between F3-F2; Bark normalised) from the middle 20% of the vowel were taken to be included in the analysis.

The Bark difference measures used in the Syrdal and Gopal [22] normalisation method have been found to better reflect the auditory properties of phonological categories associated with vowels, such as height and backness, than raw values of single formants [23], [24]. While Syrdal and Gopal's normalisation has been observed to be slightly less successful than vowel-extrinsic methods in classification studies dealing with sociolinguistic variation, the present study does not focus on speech categorisation. For the purposes of the current experiment, the Bark-difference normalisation method allows us to control for speaker-dependent differences, while yielding a single-measure suitable for cross-speaker comparison.

The statistical analysis was run in SPSS [25]. Two Generalised Linear Mixed Models were built, one with F1 (F1- f_0) and the other with F2 (F3-F2) as the dependent variables. In both models, the main interaction of interest was Vowel*[Recording]Session. Speaker and Item were included as random factors. The results below show pairwise comparisons provided by the model.

3.4. Results

In total, 4179 tokens were analysed, out of which /ɛ/ comprised 990 items and /a/ – 3189 items. In what follows, we will discuss each vowel individually.

3.4.1. Vowel height

Let us begin with the higher of the two vowels, that is /ɛ/. Out of the 990 obtained productions, Group 1 provided 584 items (194 at T1: *M*=4.5 Bark, *SD*=.55; 196 at T2: *M*=4.57 Bark, *SD*=.62; and 194 at T3: *M*=4.74 Bark, *SD*=.62), Group 2 – 147 items: *M*=4.85, *SD*=.59; Group 3 – 144: *M*=4.70 Bark, *SD*=.48; while Group 4 – 115: *M*=4.47 Bark, *SD*=.64.

The changes in the height of the vowel /ɛ/ are visualised in the line graph in Fig. 1. The *y* axis has been reversed for the differences with respect to height to be clearer.

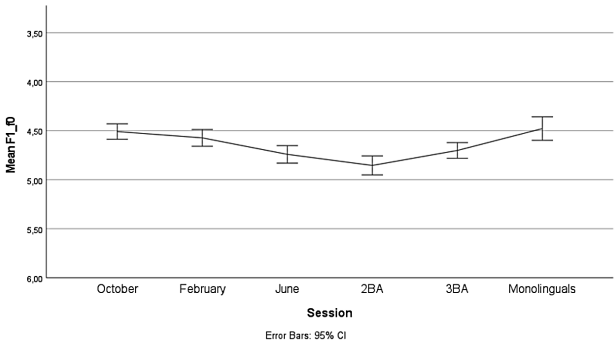


Figure 1: Mean F1 (F1-f_0; Bark normalised) values of the vowel /ɛ/ sorted for Group/Session.

While the differences are very small, we can observe some lowering taking place, with the F1-f_0 difference getting bigger, over the course of the phonetic instruction in L2. The analysis revealed that there was a significant difference in 1BA students between T1 and T3 (*B*=-.223; *t*=-4.487; *p*<.001) and T2 and T3 (*B*=-.164; *t*=-3.325; *p*=.001). T1 was also different from 2BA (*B*=-.324; *t*=-2.212; *p*=.027), while 2BA students differed from the monolingual controls (*B*=.368; *t*=2.323; *p*=.02). No other contrasts turned out to be significant. Therefore, the most visible effects of phonetic drift on the height of /ɛ/ were present in after eight months and two years of pronunciation instruction.

Now let us turn to the vowel /a/. Out of the 3189 obtained productions, Group 1 provided 1854 items (618 at T1: *M*=5.58 Bark, *SD*=.64; 612 at T2: *M*=5.75 Bark, *SD*=.65; and 624 at T3: *M*=5.88, *SD*=.57), Group 2 – 467 items: *M*=5.91 Bark, *SD*=.57; Group 3 – 468: *M*=5.58 Bark, *SD*=.61; while Group 4 – 400: *M*=5.36 Bark, *SD*=.75. Fig. 2. traces the changes in the mean height values of the vowel /a/ over time and compares them to the monolingual norms. Once again, the *y* axis has been inverted for ease of interpretation.

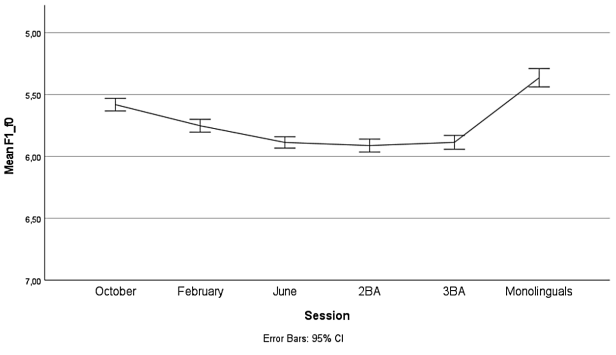


Figure 2: Mean F1 (F1-f_0; Bark normalised) values of the vowel /a/ sorted for Group/Session.

Again, as we can see the vowel appears to get progressively lower (with the F1-f_0 difference getting bigger) over the course of training.

In the case of first year students, the progressive lowering of /a/ turned out to be significant across all testing times: T1 vs. T2 (B=-.169; t=-6.051; p<.001), T2 vs. T3 (B=-.131; t=-4.707; p<.001), as well as T1 and T3 (B=-.299; t=-10.798; p=.000). 1BA students at T1 differed also from 2BA (B=-.3.22; t=-2.306; p=.021) and Group 3 (B=-.292; t=-2.084; p=.037). Additionally, aside from 1BA students at T1 (B=.232; t=1.654; p=.098), all groups were significantly different from the monolingual controls.

3.4.2. Vowel advancement

The mean values of vowel advancement for /ɛ/ presented themselves as follows. In 1BA productions, at T1: M=2.74 Bark, SD=.60; T2: M=2.61 Bark, SD=.63; T3=2.69 Bark, SD=.59. In 2BA: M=2.59 Bark, SD=.51, while in 3BA: M=2.41 Bark, SD=.69. The monolingual controls average was 3.00 Bark, SD=.60.

A visual representation of the changes with respect to the advancement of /ɛ/ is shown in Fig. 3 The x and y axes have been swapped for the reader to clearly see the changes in fronting across the groups.

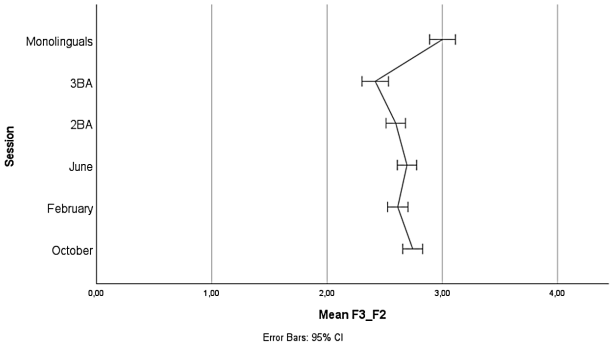


Figure 3: Mean F2 (F3-F2; Bark normalised) values of the vowel /ɛ/, sorted for Group/Session.

As can be observed, /ɛ/ appears to move towards more front realisations as the phonetic instruction progresses, peaking in the productions of 3BA.

The statistical analysis revealed that 1BA at T1 differed significantly from both 1BA at T2 (B=.126; t=2.042; p=.041) and 3BA (B=.331; t=1.981; p=.048). In turn the productions of the monolingual group were significantly less front than 1BA at T2 (B=-.384; t=-2.271; p=.023), and both 2BA (B=-.402; t=-2.228; p=.026) and 3BA (B=-.589; t=-3.261; p=.001). No other contrasts turned out to be significant.

Decidedly less variation was present in the central vowel /a/ (1BA T1: M=3.77 Bark, SD=.75; T2: M=3.80 Bark, SD=.82; T3: M=3.79 Bark, SD=.84; 2BA: M=3.81 Bark, SD=.81; Monolinguals: M=3.75, SD=.73). As can be seen, the differences in F3-F2 distance are very small in magnitude, indicating very little change in the advancement of /a/. This is illustrated also in Fig. 4. Once again, the x and y axes have been swapped to better illustrate the changes in fronting.

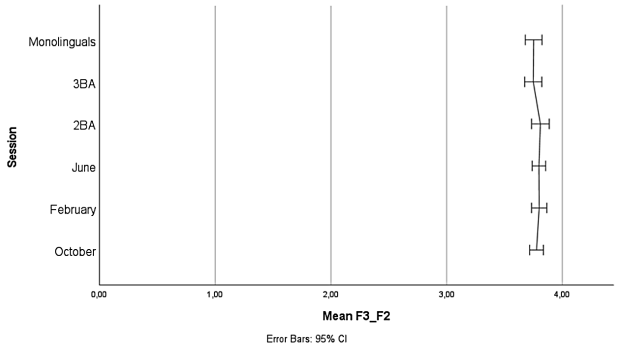


Figure 4: Mean F2 (F3-F2; Bark normalised) values of the vowel /a/, sorted for Group/Session.

The statistical analysis revealed no significant differences across groups for the advancement of /a/.

4. DISCUSSION

The present study sought to attest possible phonetic drift effects in the acoustics of vowels (namely vowel height and vowel advancement). Two non-high unrounded vowels were considered, that is /a/ and /ɛ/. While the differences were overall rather small, the results indicate that the quality of proficient Polish vowels was affected by phonetic training in L2 English, despite the fact that they were living in an L1-dominant environment. The effects were most striking productions of the students who were finishing their second year of phonetic instruction.

Both Polish vowels investigated here appeared to move towards more peripheral positions, which goes in line with the postulate of common phonological space, known from the Speech Learning Model (SLM) [26]. It claims that all L1 and L2 sounds co-exist, therefore the vowel space needs to be progressively expanded so that the new L2 categories being acquired can be accommodated. In Polish, this seems to be done by means of /ɛ/ fronting and the lowering of both vowels.

The results of the current experiment yield empirical support also to the SLM's assumption of the bi-directional interaction between the languages and the remark found in the revised version of the SLM [27] about how L1 categories are malleable across the lifespan of the user.

5. REFERENCES

[1] Chang, C. B. 2010. First language phonetic drift during second language acquisition. [Unpublished PhD thesis, University of California, Berkeley.].

[2] Chang, C. B. 2012. Rapid and multifaceted effects of second-language learning on first-language speech production, *JPhon* 40, 249-268.

[3] Podlipský, V. J., Šimáčková, Š., Chládková, K. 2020. Phonetic drift reveals interconnected phonological representations in simultaneous bilinguals: a case study of English and Czech stop consonants, *International Journal of Bilingualism* 6: 271-289.

[4] Baker, W., Trofimovich, P. 2005. Interaction of native- and second-language vowel system(s) in early and late bilinguals, *Language and Speech* 48: 1-27.

[5] Gorba, C. 2019. Bidirectional influence on L1 Spanish and L2 English stop perception: the role of L2 experience, *J. Acoust. Soc. Am.* 145: 1-7.

[6] Cabrelli, J., Luque, A., Finestrat-Martinez, I. 2019. Influence of L2 English phonotactics in L1 Brazilian Portuguese illusory vowel perception, *J Phon* 73: 55-69.

[7] Osborne, D. M. 2016. The acquisition of fine phonetic detail in a foreign language: perception and production of stops in L2 English and L1 Portuguese. [Unpublished PhD dissertation. University of Arizona.].

[8] Harada, T. 2003."L2 influence on L1 speech in the production of VOT. *Proc. 15th ICPhS*, Barcelona.

[9] Dokovova, M. 2015. Bulgarian voicing contrast: at home and abroad. [Unpublished MA thesis, University of Edinburgh].

[10] Herd, W., Walden, R.L., Knight W. L., Alexander, S. 2015. Phonetic drift in a first language dominant environment, *Proceedings of Meetings on Acoustics 23.*

[11] Yang, J., Fox, R. A.. 2017. L1-L2 interactions of vowel systems in young bilingual Mandarin-English children, *J Phon* 65: 60-76.

[12] Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., Yamada, T. 2011. A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting, *J Phon* 39, 156-167.

[13] Wrembel, M. 2011. Cross-linguistic influence in third language acquisition of voice onset time, *Proc. 17th ICPhS*, Hong Kong, 2157-2160.

[14] Sypiańska, J. 2017. *Cross-linguistic influence in bilinguals and multilinguals.* Poznań: Wydawnictwo Naukowe UAM.

[15] Schwartz, G. 2020. Asymmetrical cross-language interaction – phonological implications, *Linguistic Approaches to Bilingualism* 12, 103-132.

[16] Sypiańska, J. Production of voice onset time (VOT) by senior Polish learners of English. *Open Linguistics* 7: 316-330.

[17] Mora, J. C., Nadeu, M. 2012. L2 effects on the perception and production of a native vowel contrast in early bilinguals, *International Journal of Bilingualism* 16: 484-500.

[18] Guion, S. G. 2003. The vowel systems of Quichua-Spanish bilinguals. Age of acquisition effects on the mutual influence of the first and second languages, *Phonetica* 60, 98-128.

[19] Kartushina, N., Hervais-Adelman, A., Fauenfelder, U. H., Golestani, N. 2016. Mutual influences between native and non-native vowels in production: evidence from short-term visual articulatory feedback training", *JPhon* 57, 21-39.

[20] Bohn, O-S., Best C. T. 2012. Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners, *J Phon* 40, 109-128.

[21] Boersma, P. 2001. Praat, a system for doing phonetics by computer, *Glot International* 5, 341-345.

[22] Syrdal, A. K., Gopal, H. S. 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels, *J. Acoust. Soc. Am.* 79, 1086–1100

[23] Chistovich, L. A., Lublinskaya, V. V. 1979. The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli, *Hearing Research* 1: 185-195.

[24] Hoemeke, K. A., Diehl, R. 1994. Perception of vowel height: the role of F1–F0 distance, *J. Acoust. Soc. Am* 96: 661-674.

[25] IBM Corporation. 2020. *IBM SPSS Statistics for Windows, Version 27.1*. Armonk, New York: IBM Corp.

[26] Flege, J. E. 1995. Second language speech learning: theory, findings, and problems. In: Strange, W. (ed.). *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, Timonium, MD: York Press, 233-276.

[27] Flege, J., & Bohn, O. (2021). The Revised Speech Learning Model (SLM-r). In: Wayland, R. (ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge: Cambridge University Press, 3-83.

EFFECTS OF COGNITIVE LOAD ON VOWEL PRODUCTION IN BILINGUAL SPEAKERS OF HIGH AND LOW GERMAN

Marina Frank, Marina Rohloff, Jörg Peters

Carl von Ossietzky University Oldenburg
{marina.frank, marina.rohloff1, joerg.peters}@uni-oldenburg.de

ABSTRACT

Cognitive load from speaking a foreign language has an influence on vocal measures such as pitch. Cognitive load induced by increasing task difficulty was found to affect the vowel space in some studies, however, the results are inconclusive. This study examines the influence of increased cognitive load from speaking a less dominant language on vowel formants in 95 bilingual speakers of High German and Low German. F1 and F2 values and vowel durations were extracted from the vowels /a/, /i/, and /u/ in stressed syllables in a route description task. We found gender-dependent differences related to language dominance for F1 and vowel duration of /a/ and for F2 of /i/. The other results point more towards language-specific differences. We conclude that cognitive load from speaking a less dominant language has some effect on vowel realizations, however, language-specific differences prevail even for very similar languages.

Keywords: Vowels, cognitive load, bilingualism, High German, Low German.

1. INTRODUCTION

Various acoustic measures are related to speech under increased cognitive load [1]–[4]. Cognitive load induced by task difficulty has been associated with an increase in disfluencies and filled pauses [5]. An increase of f0 has been reported in multiple studies, with a steeper increase for real-life stress (e.g., airplane crash) than for laboratory induced stress [6]–[11]. Additionally, acoustic measures related to the glottal signal, such as jitter and shimmer, are reduced under increased cognitive load [12]. Cepstral peak prominence has also been correlated with speech production under increased cognitive load [13].

Previous studies on the influence of cognitive load on vowel formant frequencies show contradictory results. For real-life stress in airplane incidences, [6] found higher F1 values for one pilot while [10] found no significant changes in F1 for the pilot but a significant increase in F1 for /i/ and a decrease for /u/ for the copilot. For F2, the pilot showed an increase for /a/, /e/, and /ɛ/, while the co-pilot did not show any significant differences. In studies like these, the sample size is naturally very small, i.e., only one or two pilots are investigated. In a flight simulator study with 13 subjects, [14] found increased F1 values and decreased F2 values for front vowels and increased F1 and F2 values for back vowels under cognitive load, i.e., more centralized vowels.

For laboratory induced stress, [15] found no significant differences for F1 and F2 in a dual-task condition compared to the control condition. In a Stroop test analysis of one speaker, [10] found a decrease in F1 for /u/ and for /ø/ and increased F1 and F2 values for /ɛ/ with increasing cognitive load. In their analysis of a larger Stroop test database, [16] found a decrease in F2 for /ao/; however, they did not find an overall smaller vowel space for the higher cognitive load condition. With increased cognitive load, vowels showed longer durations, potentially because subjects spoke more slowly [16]. Cognitive load was also found to affect the vowel space in anxiety-denying women (centralization), i.e., these effects varied depending on speaker sex and coping style [17]. In three different cognitive load conditions, [18] found increased F1 and F2 values compared to a control condition. In conclusion, both centralization and peripheralization of the vowel space have been observed with regard to an increase in cognitive load due to an increase in task difficulty.

It is not clear whether effects of cognitive load on vowel formants also occur in the less dominant language of bilingual speakers. Cognitive load from speaking a foreign language is reported to have an influence on f0 level and f0 dispersion ([19], [20] (English/Finnish), [21] (High/Low German)). These results are in line with the results regarding cognitive load from task difficulty. There is ample research on the influence of the vowel system of the more dominant language on the vowel system of the less dominant language (e.g., [22]–[24]). However, there have been no studies on the effect of cognitive load caused by speaking the less dominant language on vowel formants. In this case, vowel formant analysis is difficult because the acoustic characteristics of vowels often differ between two languages. For this reason, we focus on two closely related languages in this paper.

We report results of a study that examines the influence of cognitive load on vowel formants and vowel duration in bilingual speakers of High German

and Low German. Low German is a regional language spoken in northern Germany that nowadays is no longer acquired monolingually, but always together with High German. Since the use of Low German is declining in younger speakers [25], we expect that they are more dominant in High German than older speakers. We expect that these speakers exhibit larger differences in F1 and F2 between the two languages than more balanced bilinguals, i.e., the middle-aged and older speakers. We also expect the speakers who are more dominant in High German to have longer vowel durations in Low German than the more balanced bilingual speakers.

2. METHOD

2.1. Participants and data collection

A total of 95 participants, aged 15 to 88 years, took part in this study. All subjects (47 female, 48 male) grew up in the municipality of Krummhörn in East Frisia, in the northwest of the federal state of Lower Saxony in Germany. The subjects were native speakers of East Frisian Low German (LG) and the regional standard variety of High German (HG). We had to exclude 4 participants because of missing data, retaining 91 subjects for analysis.

All subjects verbally completed a questionnaire consisting of 55 questions on socio-demographics as well as on age of acquisition, frequency of language use, and self-attributed language proficiency (adopted from [25–27]). Of all questions, 28 were used to calculate global language scores for each subject and language (based on [26] and [27]). To obtain the language dominance score used in our analysis, we subtracted the HG values from the LG values. Negative values indicate a dominance of HG over LG (= HG dominant), positive values a dominance of LG over HG (= LG dominant). A value around 0 indicates balanced bilingualism. Possible values range from -174 to +174; the values in our sample range from -119 to +58. The younger speakers show the highest dominance of HG while the older speakers exhibit balanced dominance scores. For a more thorough analysis of the correlation between age and dominance score in our subjects see [28].

Speech recordings were collected using a head-mounted omnidirectional microphone (DPA 4066) and a portable digital recorder (Tascam DR-100 MKIII). The recordings were digitized at 48 kHz sampling rate with 24 bits/sample quantization.

2.2. Procedure and data selection

During the experiment, subjects completed seven tasks in each language. The order of the languages was randomized per subject, the order of the tasks was kept stable. One of these tasks was giving a route description, for which we varied the level of task difficulty to induce different levels of cognitive load. In the more difficult condition, the participants received a map without any labels or landmarks. In the easier condition, labels and landmarks were added to the map. The routes were different in each condition.

We manually segmented the point vowels /i/, /u/, and /a/ in stressed syllables in the route description tasks. We excluded vowels in the context before /r/, as post-vocalic /r/ is vocalized in LG and in HG resulting in a diphthong [29][30]. For /i/ and /u/, we do not expect any language-specific spectral differences [29][30]. For /a/, we expect a central open realization for HG [29] and a more retracted and slightly rounded realization for LG [30]. Regarding vowel duration, we do not expect language-specific differences for all three vowels.

The duration of the route description tasks ranged from 15.87 s to 176.31 s ($\bar{x}$ = 68.93 s). In total, the sample contained 2,470 /a/ tokens, 2,343 /i/ tokens, and 1,409 /u/ tokens.

2.3. Acoustic analysis

F1 and F2 values were measured at vowel midpoint using Praat [31]. We calculated five formants using the Burg LPC method. The formant ceiling was set to 5000 Hz for male speakers and to 5500 Hz for female speakers. Window length was 25 ms and pre-emphasis from 50 Hz. Additionally, we extracted the duration for each vowel. All recordings were downsampled to 16 kHz before analysis. During the inspection of the automatic formant extraction, we found obvious measurement errors. To reduce these, we excluded the 5% outermost values for each vowel and each formant, leaving 2,235 /a/ tokens (HG: 1,207; LG: 1,263), 2,115 /i/ tokens (HG: 1,294; LG: 1,049), and 1,286 /u/ tokens (HG: 370; LG: 1,039) for analysis.

2.4. Statistical analysis

We fitted LMMs in R [32] using the glmmTMB package [33]. The dependent variables were F1, F2, and Duration of each vowel. As fixed effects, we added the within-subject effect Language (HG vs. LG) and the between-subjects effects Dominance Score and Gender (female vs. male) to the model as well as all possible interactions. We used random intercepts for Subjects as well as by-subject random slopes for Language. Since the residuals for the dependent variables were not normally distributed, we used the Gamma distribution and log-link function in our models. Interaction results are reported using the emtrends function from the emmeans package [34]. The level of significance was set to $p = 0.05$. Figures were created using tidyverse packages [35].

We included the main effects for Gender and Dominance Score only as control variables, which is why we do not report these results below. The factor Task Difficulty (easy vs. complex) was left out of the models because the effects on vowel production were inconsistent and difficult to interpret. This could probably be because the difficulty levels did not differ sufficiently from each other to reliably affect vowel realizations.

3. RESULTS

3.1. F1

Figure 1 displays F1 values for each vowel as a function of Dominance Score per Gender.

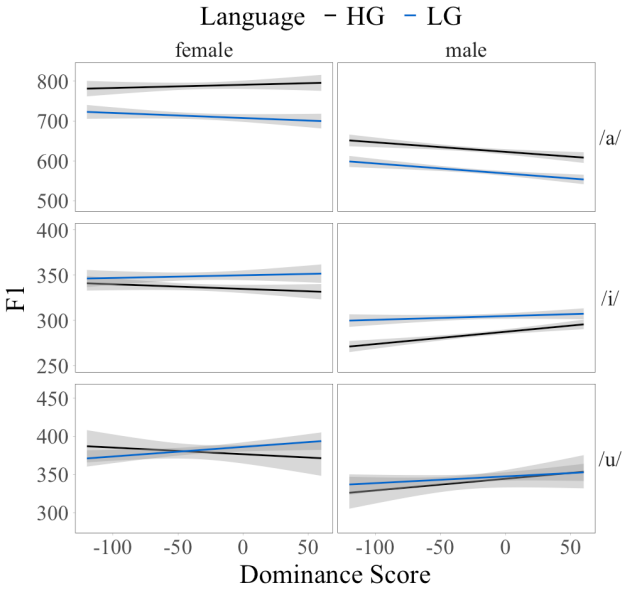


Figure 1: F1 values. Shading indicates 95% confidence intervals.

For /a/, statistical analysis revealed significantly lower F1 values for LG compared to HG ($\chi^2 = 185.691$, $p < .001$).

For /i/, statistical analysis revealed significantly higher F1 values for LG compared to HG ($\chi^2 = 63.394$, $p < .001$). In addition, we found a three-way interaction between Language, Gender, and Dominance Score. Contrast analysis revealed a significantly steeper, positive slope for HG compared to LG for male speakers (estimate = -0.0005, $SE = 0.0002$, z-ratio = -2.051, $p < .05$), suggesting that LG dominant speakers exhibit smaller differences in F1 between the two languages than HG dominant speakers.

For /u/, statistical analysis revealed significantly higher F1 values for LG compared to HG ($\chi^2 = 4.484$, $p < .05$).

3.2. F2

Figure 2 displays F2 values for each vowel as a function of Dominance Score per Gender.

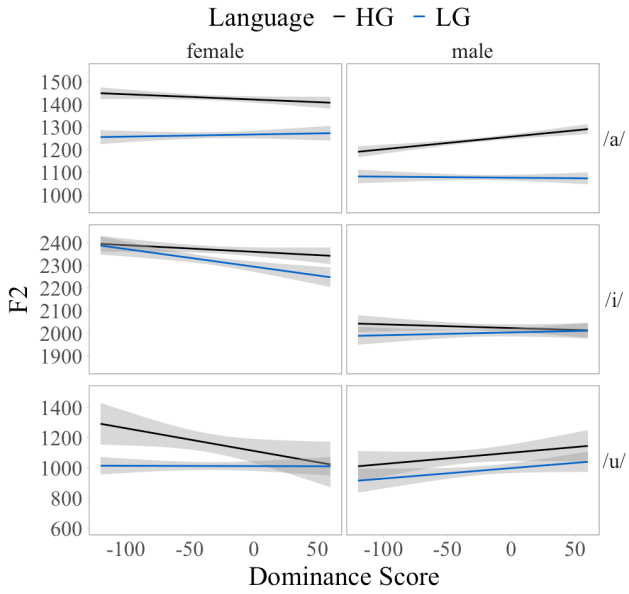


Figure 2: F2 values. Shading indicates 95% confidence intervals.

For /a/, statistical analysis revealed significantly lower F2 values for LG compared to HG ($\chi^2 = 440.83$, $p < .001$). Additionally, we found a three-way interaction between Language, Gender, and Dominance Score. Contrast analysis revealed a significantly steeper, positive slope for HG compared to a slightly negative slope for LG for male speakers (estimate = -0.0005, $SE = 0.0002$, z-ratio = -2.299, $p < .05$), suggesting that LG dominant speakers exhibit larger differences in F2 between the two languages than HG dominant speakers.

For /i/ and /u/, statistical analysis revealed significantly lower F2 values for LG compared to HG (/i/: $\chi^2 = 12.766$, $p < .001$; /u/: $\chi^2 = 16.784$, $p < .001$).

3.3. Duration

Figure 3 displays Duration values for each vowel as a function of Dominance Score per Gender.

For /a/, statistical analysis revealed significantly longer vowels for LG compared to HG ($\chi^2 = 29.353$, $p < .001$). Additionally, we found a three-way interaction between Language, Gender, and Dominance Score. Contrast analysis revealed a significantly steeper, positive slope for HG compared to a negative slope for LG for female speakers (estimate = -0.0013, $SE = 0.0006$, z-ratio = -2.340, $p < .05$), suggesting that LG dominant speakers exhibit a smaller difference in Duration between the two languages than HG dominant speakers.

For /i/ and /u/, statistical analysis revealed significantly longer vowels for LG compared to HG (/i/: $\chi^2 = 58.117$, $p < .001$; /u/: $\chi^2 = 17.348$, $p < .001$).

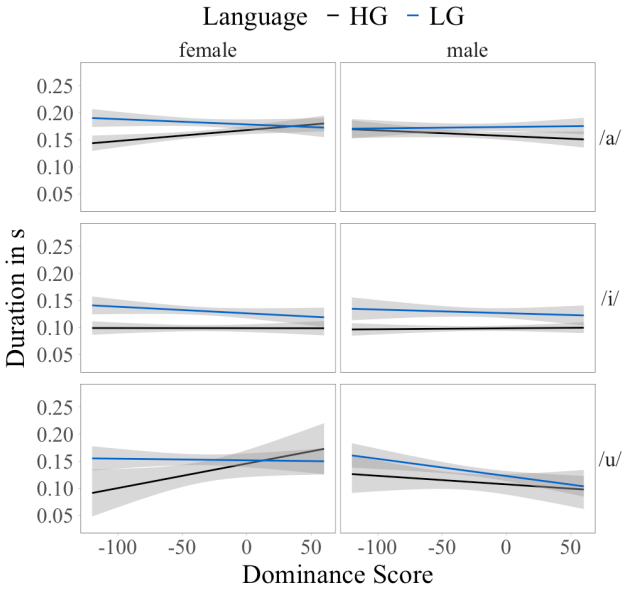


Figure 3: Duration values. Shading indicates 95% confidence intervals.

4. DISCUSSION

In this study, formant values and vowel durations of the point vowels /a/, /i/, and /u/ were analyzed from route descriptions provided by 91 bilingual speakers. Our hypothesis was that HG dominant speakers, i.e., mostly younger speakers, would show larger differences between LG and HG than LG dominant speakers, i.e., middle-aged and older speakers, as for the former the use of the less dominant language, LG, is associated with increased cognitive load. We conducted nine analyses, including F1, F2, and vowel duration for each of the three vowels, and found that language dominance had a partial effect on male speakers' F1 values for /i/ and F2 values for /a/, as well as on female speakers' durational values for /a/.

The lower F1 and F2 values for /a/ in LG suggest that /a/ is pronounced as a more closed back vowel in LG compared to HG. These results can directly be attributed to language-specific differences [30]. Additionally, a three-way interaction between Language, Gender, and Dominance Score showed a larger difference in F2 for LG dominant male speakers than for HG dominant male speakers. This effect is the contrary of what we hypothesized. Moreover, the difference can be attributed to a change in F2 in HG while the F2 values in LG remain relatively stable across speakers with varying language dominance.

For /i/, we found higher F1 values and lower F2 values in LG compared to HG. In combination, these values indicate a less peripheral production of /i/ in LG. Furthermore, a three-way interaction between Language, Gender, and Dominance Score showed a smaller difference in F1 for LG dominant male speakers than for HG dominant male speakers, as we hypothesized. However, the difference can be attributed to a change in F1 in HG while the F1 values in LG remain relatively stable across speakers independent of their dominance score.

The higher F1 values for /u/ in LG compared to HG indicate a more open production of /u/ in LG. In contrast, lower F2 values in LG indicate a further back production of /u/ in LG. As we find no interactions with language dominance for /u/, we cannot decide whether these effects arise from increased cognitive load in speaking the less dominant language or from language-specific differences.

Regarding the durational values, we found longer vowel durations in LG than in HG for /a/, /i/, and /u/. These results are in line with [16], who found it an important measure for increased cognitive load. For /a/, a three-way interaction between Language, Gender, and Dominance Score showed converging durational values for LG dominant female speakers. However, for the most part we cannot decide whether these are language-specific or cognitive load-related differences.

In only three cases, our results include a partial effect of language dominance; and even these cases are contradictory. Considering the results at hand, the question arises whether vowel formants and durations are suitable indicators for cognitive load in bilingual speech. The inconsistent results of our study are in line with previous research. One possible explanation would be that different coping styles as reported in [17] have led to opposing effects that have largely cancelled each other out.

5. CONCLUSION

Our study revealed only a few effects of increased cognitive load from speaking a less dominant language on vowel production. Most of our findings could also be attributed to language-specific differences. Nevertheless, this study adds to the existing literature by investigating cognitive load in spontaneous speech of bilingual speakers.

6. ACKNOWLEDGMENTS

The research presented here was funded by the German Research Foundation (DFG), grant number PE 793/3–1. We thank our participants for their time and effort, our student assistants for help with data cleaning and analysis, and our recording assistants for their help with data collection.

7. REFERENCES

[1] Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., Winter, A. S. 2013. Vocal Indices of Stress: A Review. *J. Voice* 27, 390.e21–390.e29.

[2] Kirchhübel, C., Howard, D. M., Stedmon, A. W. 2011. Acoustic correlates of speech when under stress: Research, methods and future directions. *Int. J. Speech Lang. Law* 18, 75–98.

[3] Moore, R. K. (ed.) 1996. Speech under Stress [Special Issue]. *Speech Commun.* 20.

[4] Van Puyvelde, M., Neyt, X., McGlone, F., Pattyn, N. 2018. Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Front. Psychol.* 9, 1994.

[5] Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F. 2001. Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. *User Modeling: Proceedings of the Eighth International Conference*, 24–33.

[6] Benson, P. 1995. Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency. *ESCA-NATO Workshop on Speech under Stress*, 61–64.

[7] Brenner, M., Shipp, T., Doherty, E. T., Morrissey, P. 1985. Voice measures of psychological stress: laboratory and field data. In: Titze, I. R., Scherer, R. C. (eds.), *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*. Denver Center for Performing Arts, 239–248.

[8] Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T. 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Appl. Ergon.* 42, 348–357.

[9] Jessen, M. 2006. *Einfluss von Stress auf Sprache und Stimme: Unter besonderer Berücksichtigung polizeidienstlicher Anforderungen*. Schulz-Kirchner.

[10] Ruiz, R., Absil, E., Harmegnies, B., Legros, C., Poch, D. 1996. Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions. *Speech Commun.* 20, 111–129.

[11] Scherer, K. R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T. 2002. Acoustic correlates of task load and stress. *7th ICSLP 2002 Denver Colo. USA Sept. 16-20 2002*, 2017–2020.

[12] Boyer, S., Paubel, P.-V., Ruiz, R., El Yagoubi, R., Daurat, A. 2018. Human Voice as a Measure of Mental Load Level. *JSLHR* 61(11), 2722–2734.

[13] Yap, T. F., Epps, J., Ambikairajah, E., Choi, E. H. C. 2015. Voice source under cognitive load: Effects and classification. *Speech Commun.* 72, 74–95.

[14] Huttunen, K. H., Keränen, H. I., Pääkkönen, R. J., Päivikki Eskelinen-Rönkä, R., Leino, T. K. 2011. Effect of cognitive load on articulation rate and formant frequencies during simulator flights. *J. Acoust. Soc. Am.* 129, 1580–1593.

[15] Lively, S. E., Pisoni, D. B., Van Summers, W., Bernacki, R. H. 1993. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences *J. Acoust. Soc. Am.* 93, 2962–2973.

[16] Yap, T. F., Epps, J., Ambikairajah, E., Choi, E. H. C. 2011. Formant Frequencies under Cognitive Load: Effects and Classification. *EURASIP J. Adv. Signal Process.* 2011, 219253.

[17] Tolkmitt, F. J., Scherer, K. J. 1986. Effect of Experimentally Induced Stress on Vocal Parameters. *J. Exp. Psychol. Hum. Percept. Perform.* 12(3), 302–313.

[18] Meinerz, C. 2010. *Effekte von Stress auf Stimme und Sprechen: Eine phonetische Untersuchung auf der Grundlage ausgewählter akustischer und sprechdynamischer Parameter unter Berücksichtigung verschiedener Stressklassen*. Books on Demand.

[19] Järvinen, K., Laukkanen, A.-M. 2015. Vocal loading in speaking a foreign language. *Folia Phoniatr. Logop.* 67, 1–7.

[20] Järvinen, K., Laukkanen, A.-M., Aaltonen, O. 2013. Speaking a foreign language and its effect on F0. *Logoped. Phoniatr. Vocol.* 38, 47–51.

[21] Peters, J., Frank, M., Rohloff, M. 2020. Pitch range variation in High German (L1) and Low German (L2). *Proceedings of the 10th International Conference on Speech Prosody*, 650–654.

[22] Mayr, R., Morris, J., Mennen, I., Williams, D. 2017. Disentangling the effects of long-term language contact and individual bilingualism: The case of monophthongs in Welsh and English. *Int. J. Biling.* 21, 245–267.

[23] Schoormann, H. E., Heeringa, W., Peters, J. 2019. Standard German vowel productions by monolingual and trilingual speakers. *Int. J. Biling.* 23, 138–156.

[24] Yang, J. 2021. Vowel development in young Mandarin-English bilingual children. *Phonetica* 78, 241–272.

[25] Adler, A., Ehlers, C., Goltz, R., Kleene, A., Plewnia, A. 2016. *Status und Gebrauch des Niederdeutschen 2016: Erste Ergebnisse einer repräsentativen Erhebung*. Institut für Deutsche Sprache.

[26] Birdsong, D., Gertken, L. M., Amengual, M. 2012. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. sites.la.utexas.edu/bilingual/.

[27] Poarch, G. J., Vanhove, J., Berthele, R. 2019. The effect of bidialectalism on executive function. *Int. J. Biling.* 23, 612–628.

[28] Peters, J., Frank, M., Rohloff, M. 2023. Vocal fold vibratory patterns in bilingual speakers of Low and High German. *Proceedings of the 20th International Congress of Phonetic Sciences.*

[29] Kleiner, S., Knöbl, R., Dudenredaktion, Institut für Deutsche Sprache (eds.). 2015. *Duden: Das Aussprachewörterbuch*. Dudenverlag & Institut für Deutsche Sprache.

[30] Lücht, W. 2016. *Ostfriesische Grammatik*. Ostfriesische Landschaft.

[31] Boersma, P., Weenink, D. 2022. Praat: Doing Phonetics by Computer [Computer program]. V. 6.2.14., retrieved 24 May 2022. www.fon.hum.uva.nl/praat/.

[32] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. www.r-project.org

[33] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., Bolker, B. M. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 9, 378–400.

[34] Lenth, R. V. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.5. https://cran.r-project.org/package=emmeans.

[35] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. 2019. Welcome to the tidyverse. *JOSS* 4, 1686.

VOWEL REDUCTION IN SPONTANEOUS BULGARIAN JUDEO-SPANISH

Jonas Grünke[1], Mitko Sabev[2], Christoph Gabriel[1], Bistra Andreeva[2]
[1]Johannes Gutenberg University Mainz, [2]Saarland University
jgruenke@uni-mainz.de, msabev@lst.uni-saarland.de, christoph.gabriel@uni-mainz.de, andreeva@lst.uni-saarland.de

ABSTRACT

This study compares spontaneous speech data from Judeo-Spanish as spoken in Bulgaria (BJS) and from Bulgarian (BG) with respect to spectral and durational reduction of unstressed vowels, as well as the ensuing height neutralizations. It is shown that BJS speakers, who are all bilingual with BG, largely follow Bulgarian reduction patterns, i.e., unstressed underlyingly non-high vowels raise considerably and tend to merge with their high counterparts. At the same time, however, the dialectal background of the speakers emerges as a relevant factor since neutralizing reduction of front vowels is only present in Eastern Bulgarian dialects. In contrast, raising is gradient in (close-to-standard) western dialects, where the raising of unstressed /ɛ/ to [i] is highly stigmatized.

Keywords: vowel reduction, vowel merger, incomplete neutralization, Judeo-Spanish, Bulgarian

1. INTRODUCTION

Judeo-Spanish (JS) refers to the varieties of Spanish spoken by the Sephardic Jews in their new areas of settlement (mostly in the former Ottoman Empire and North Africa) after their expulsion from the Iberian Peninsula at the end of the 15th century. From that point onwards, it developed independently from other Spanish varieties but in close contact with the respective surrounding languages, among them Greek, Turkish, and Bulgarian. The Bulgarian variety of JS addressed in this paper is still spoken today by a rather small group of probably less than 200 native speakers, the youngest of whom were born in the 1960s ([1]). All speakers are at least bilingual and dominant in Bulgarian (BG). The use of JS is nowadays restricted to informal communication within the community.

So far, the literature on JS phonology is rather sparse. However, recently both the segmental and the prosodic properties of the variety of JS spoken in Istanbul (Turkey) and in Sofia (Bulgaria) have been investigated ([2–11]).

As regards its vowel system, JS generally continues to exhibit the three-level system inherited from Spanish consisting of the five phonemes /i e u o a/ ([10]). In Bulgarian JS, /ɤ/ may be added to these as it can occur in loans from BG according to [4]. Furthermore, the quality of the mid-level vowels favours their classification as /ɛ/ and /ɔ/, making the BJS vowel system strongly resemble the Bulgarian one, which contains the six contrastive stressed vowels /i ɛ a ɤ ɔ u/. In opposition to Mainstream Spanish, however, where (spectral) reduction of unstressed vowels is virtually absent ([12–13]), it is generally accepted that in Bulgarian the six stressed vowels /i ɛ a, ɤ ɔ u/ are reduced to a subsystem of four (three in some dialects) /i (ɛ) ɤ u/ in unstressed positions. Opinions differ, however, as to the exact nature of the reduction process. [14] and [15] identify an intermediate quality for the reduced vowels, with /ɛ/ and /i/ neutralizing to [e], /ɔ/ and /u/ neutralizing to [o] and /a/ and /ɤ/ neutralizing to [ə], which implies additional defined targets for the unstressed vowels. Contrary to these claims, recent studies show evidence that the general pattern of the Bulgarian vowel reduction is one of low and mid vowel raising ([16–19], cf. also [20]). This results in the neutralization of /a/ and /ɤ/ to [ɤ], /ɔ/ and /u/ to [u], and in some (eastern) dialects /ɛ/ and /i/ to [i]. The latter merger is strongly stigmatized in the variety spoken in the capital Sofia.

As for BJS, recent studies suggest that, presumably as a consequence of its long-lasting and intense contact with Bulgarian, the phonological system has largely converged with the one of the contact language (cf. [5] for rhythm, [6–7] for intonation). Regarding its vowel system, [5] have analysed read data to show that in JS as spoken in Sofia unstressed /a/ and /ɔ/ tend to raise following the BG pattern, although to a somewhat lesser extent (n = 420). Regarding unstressed /ɛ/, an auditory analysis by [21] revealed that speakers stemming from Eastern Bulgaria present raising to [i] when spontaneously speaking BJS (to an extent of roughly 24 to 50%, n = 647) but clearly avoid this in their reading pronunciation as well as in (spontaneous and read) BG.

The present study is concerned with the analysis of spectral and durational unstressed vowel reduction (UVR) in the spontaneous speech of bilingual female speakers of BJS and BG and monolingual speakers of BG and addresses the following research questions (RQ):

RQ 1: How do the vowel spaces shrink in unstressed positions?
RQ 2: Which is the relative weight or importance of the acoustic parameters F1, F2, and duration in distinguishing stressed from unstressed vowels?
RQ 3: What is the extent of contrast loss between low and high vowels in unstressed positions?

2. METHODS AND PROCEDURE

We collected data from four female bilingual BJS-BG speakers, aged 80 to 88 (recordings Sofia, September 2012). They were born and raised in different places in Bulgaria (Kyustendil, Pazardzhik, Kazanlăk, and Karnobat) and had used BJS on a regular basis in family situations along with BG during their childhood. Nevertheless, BG became their dominant language when they moved to Sofia for study purposes between 1947 and 1950. Having lived there for more than 70 years at the time of recording, they spoke the BG variety of the capital. The bilingual subjects were recorded in both of their languages (BJS and BG_B(ilingual)). Four same-aged monolingual BG (BG_M) speakers born and raised in Sofia (all females, aged 79–86) served as control group (recordings Sofia, September/October 2016). All participants held an academic degree.

The material gathered for the analysis of UVR consisted of extracts from narrative interviews conducted in both BJS and BG (net amount of speaking time, excluding all pauses, ca 3 min per speaker and variety). The interviews were semi-focused in that all speakers were asked to retell their life story and family history and to speak freely about their daily lives and, only in the case of the bilinguals, about their language use. The data were recorded with a Marantz hard disk recorder (PMD671) and a Sennheiser microphone (ME64).

Variety	Speakers	Vowel tokens
BJS	4	2438
BG_B	4	2842
BG_M	4	2290

Table 1: Speakers and vowel tokens per variety.

The vowels were manually segmented in Praat [22], on the basis of the synchronized spectrogram, waveform, and audio signal. Vowel boundaries were determined by the presence of clear formant structure and sharp changes in intensity. Material produced with creaky voice or disfluencies was excluded from the acoustic analysis. Since unstressed vowels preceding pauses are prone to be lengthened cross-linguistically [23], we decided to analyse tokens in phrase-final position as a separate category. A Praat script was used to extract vowel duration and midpoint F1 and F2 frequencies. Formant values were normalized using Lobanov (*z*-transform) as implemented in the R package phonR [24]. Next, outliers, defined as values 1.5 times beyond the interquartile range, were removed (F1 = 2.5%, F2 = 4.3%, duration = 3.5%).

A series of MANOVAs were performed to measure UVR, on the one hand, and height contrast, on the other. In the former set, we examined the effect of stress and location (i.e., phrase-final vs non-final) on F1, F2, and duration, while vowel height was the input variable in the latter set. Pillai's trace (Λ), which is part of the output of the MANOVAs, was used as a measure of separation, values approaching Λ = 1 indicating no overlap and values approaching Λ = 0 indicating complete overlap. The strength of contrast between high and low vowels was studied in stressed and unstressed position, and a measure of contrast loss was computed as the difference in contrast between stressed and unstressed position relative to contrast in stressed position (i.e., $\Lambda_{stressed}$–$\Lambda_{unstressed}$)/ $\Lambda_{stressed}$) [18]. MANOVA, and in particular Pillai's trace, have been used increasingly in phonetic and phonological research to measure the extent of vowel mergers (including mergers resulting from vowel reduction) [18, 19, 25–31].

Finally, the parallel discriminant ratio coefficient derived from Discriminant Function Analysis (DFA) was used as an estimate of the relative importance, or weight, of each acoustic parameter (i.e., F1, F2, and duration) in discriminating between stressed and unstressed, or high and non-high vowels, respectively. The role of a parameter was considered to be negligible if its coefficient was lower than half the mean of the absolute coefficient values for all three parameters. For a detailed discussion of this measure, cf. [18, 32].

3. RESULTS

Table 2, below, shows the means for F1, F2, and duration of stressed and unstressed vowels in different positions in the three varieties under concern in this paper. Means for phrase-final unstressed /u/ are only given for BG_M due to an insufficient number of tokens in the bilingual varieties. Note, in this respect, that there are no native words ending in unstressed /u/ in BJS.

The lower means of the F1 values indicate that the non-high unstressed vowels /a ɛ ɔ/ are clearly raised as compared to their stressed counterparts in all varieties. The high vowels /i u/, and in BG /ɤ/, on the contrary, largely remain in place. This is illustrated in Figure 1, below, where the unstressed vowel spaces are noticeably reduced as compared to the stressed ones. Final unstressed vowels, however, are somewhat less reduced than non-final unstressed vowels, which present the smallest vowel spaces across varieties.

Regarding duration, non-final unstressed vowels present markedly shorter means than stressed vowels. Again, this difference is more prominent in non-high vowels. Final unstressed vowels, as expected, showed the longest duration, which in some cases doubled the duration of stressed vowels.

Vowel	Position	F1			F2			duration		
		BSJ	BG_B	BG_M	BSJ	BG_B	BG_M	BSJ	BG_B	BG_M
a	stressed	816	802	717	1599	1626	1663	147	127	110
	unstr. final	637	584	546	1688	1601	1608	144	189	173
	unstr. non-final	628	517	488	1701	1775	1651	81	61	66
ɛ	stressed	596	591	548	2182	2123	1923	135	109	104
	unstr. final	450	482	490	2223	2179	1971	197	164	178
	unstr. non-final	438	428	455	2167	2155	1868	78	63	64
ɤ	stressed		498	483		1646	1608		99	81
	unstr. final		544	438		1589	1396		85	79
	unstr. non-final		479	482		1634	1524		59	69
i	stressed	377	364	407	2525	2450	2133	126	90	83
	unstr. final	349	398	396	2563	2491	2182	263	173	175
	unstr. non-final	344	372	406	2354	2330	2079	83	67	67
ɔ	stressed	589	611	597	1058	1086	1237	137	124	106
	unstr. final	426	413	459	1071	983	1140	158	193	181
	unstr. non-final	440	391	429	1203	1156	1262	87	65	63
u	stressed	389	383	434	1024	1078	1291	111	87	78
	unstr. final	-	-	423	-	-	1127	-	-	134
	unstr. non-final	381	342	394	1263	1065	1231	83	72	68

Table 2: Means of F1, F2 (in Hz) and duration (in ms) across varieties and positions.

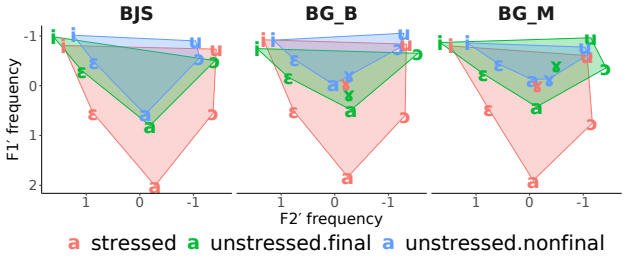


Figure 1: Mean F1/F2 vowel spaces across varieties.

The MANOVAs computed for each vowel phoneme confirmed the significant effect ($p < 0.05$) of stress on the three acoustic parameters taken together, except for BG_M /u/ ($p = 0.2745$). Pillai scores for UVR are given in Figure 2.

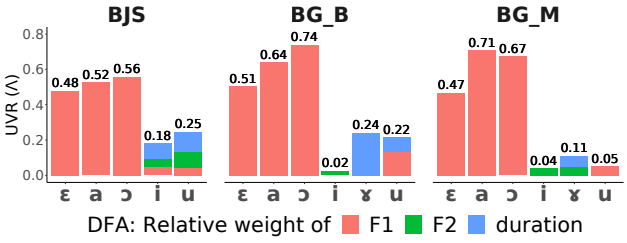


Figure 2: Extent of UVR (stressed/unstressed difference).

As can be seen, the non-high vowels are reduced only with regard to the F1 dimension, while in the high vowels, which undergo very little UVR, reduction results mainly from duration and/or F2 frequency.

MANOVAs comparing phrase-final and non-final unstressed vowels produced significant results, except for BG_M /ɤ/ (p = 0.1519). As can be seen in Figure 3, the main difference was in duration, with the formant frequencies playing practically no role.

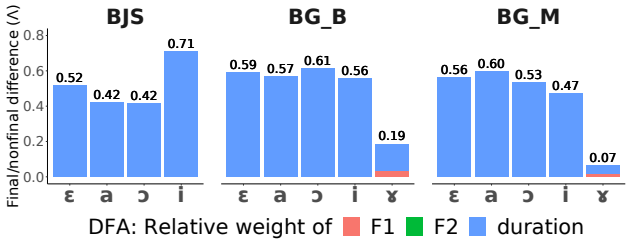


Figure 3: Extent of (unstressed) final–non-final difference (Λ) with parameter weight.

The vowel /u/ was excluded from this comparison because of the insufficient number of tokens in final position.

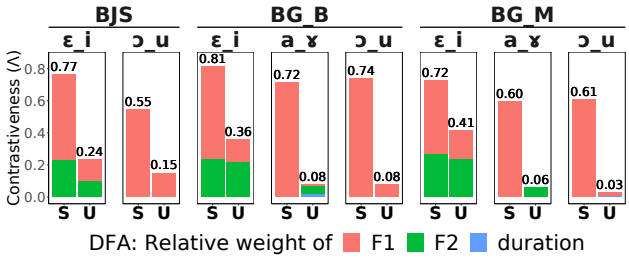


Figure 4: Height contrast with parameter weight in stressed (S) and unstressed (U) position ($p < 0.05$).

As pertains the contrast between high and non-high vowels, as expected, it was considerably stronger in stressed position (Figure 4). The respective vowel pairs were primarily distinguished by F1, while in the front vowels F2 played a secondary role. Contrast is practically lost in unstressed /a–ɤ/ and /ɔ–u/ for all varieties. In /ɛ–i/, also as expected, there is considerably less contrast loss in unstressed positions, since this

pair does not neutralize in Western Bulgarian. Interestingly, BJS shows somewhat less contrast loss in the back pair /ɔ–u/, but considerably more in /ɛ–i/ than either Bulgarian variety. Ratios of contrast loss between unstressed high and low vowels are shown in Table 3.

Variety	ɛ–i	a–ɤ	ɔ–u
BJS	69	-	73
BG_B	56	89	89
BG_M	43	90	95

Table 3: Contrast loss in per cent.

Figure 5, finally, shows the distribution of vowels across stressed and unstressed positions and hence illustrates the loss of contrast between high and low vowel pairs. As can be seen, there is almost no overlap between the distributions of the realizations corresponding the different vowel phonemes in stressed position (uppermost panel), but the overlap is substantial in unstressed positions (lower panels).

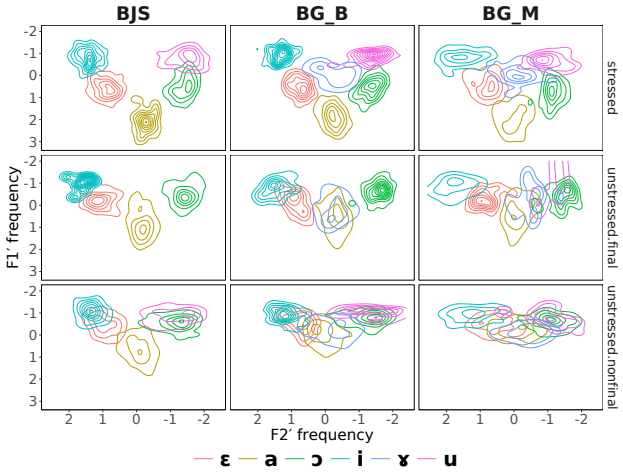


Figure 5: F1 and F2 distributions of stressed and unstressed vowels based on kernel density estimation.

4. DISCUSSION AND CONCLUSIONS

The results of the analyses show that vowel spaces were thus considerably smaller in unstressed positions across varieties (RQ 1). In this sense, BJS largely follows BG patterns with regard to UVR, i.e., the non-high vowels /ɛ a ɔ/ are considerably raised when unstressed, while other dimensions (i.e., duration and F2) do not play much of a role (RQ 2). As regards high vowels, UVR affects all dimensions, but its extent is nominal. In BG, bilingual speakers behave roughly like BG monolinguals. However, the spectral reduction in terms of F1 was found to be overall somewhat less strong in BJS than in BG. While this implies less loss of the contrasts between high and low vowels (cf. below), it could possibly be a side effect of the bilinguals' lower speech rate in their non-dominant language BJS. Findings made regarding the durational properties of vowels in different positions support this hypothesis. More particularly, unstressed vowels were generally much shorter than stressed vowels, as was expected. In phrase-final position, however, they were considerably lengthened. A secondary effect of this lengthening seems to be that that these vowels were somewhat less spectrally reduced than unstressed vowels in non-final positions. This very observation, in turn, suggests that speech rate could also have an effect on spectral UVR. As pertains the extent of contrast loss between high and low vowels (RQ 3), the raising of the latter entails that the back vowel pair /ɔ–u/ is virtually neutralized in unstressed position across varieties. The same holds true for the central pair /a–ɤ/ in BG. Both is expected in Western and Standard BG. Regarding the front pair /ɛ–i/, the situation is somewhat more complex. While unstressed /ɛ/ is clearly raised in all varieties, there is some overlap but certainly no merger of the categories in BG_M, i.e., the reduction presents itself as gradient and the quality of raised unstressed /ɛ/ rather corresponds to [e] than to [i]. As opposed to this, the extent of reduction of /ɛ/ is similar to the back vowels in BJS, with substantial overlap of the categories /ɛ/ and /i/. BG_B, finally, is in an intermediate position. A possible explanation for this finding could be that three out of the four analysed bilinguals originally stem from Eastern Bulgaria, where a loss of the contrast between unstressed /ɛ–i/ is expected. Yet, as they have been living in Sofia for most of their lives, where raising of /ɛ/ to [i] is severely stigmatized, it is likely that they try to avoid it when speaking BG ([19]).

In sum, UVR is a systematic process in BJS that clearly reflects patterns known from the BG dialects. This feature thus not only set this variety apart from present-day Spanish and other JS varieties, but also suggests strong influence from the contact language. The present results are thus in line with the observation that vowel systems of non-dominant languages tend to converge with the system of the environmental or dominant language (cf. [33] for Spanish as a heritage language in the US, [18, 29] for Turkish in Bulgaria).

5. ACKNOWLEDGEMENTS

This research was funded by the DFG (project 491553503, 'Judeo-Spanish in Bulgaria: a contact language between archaism and innovation') and by the Bulgarian National Science Fund (КП-06-Н40/11 'Prosodic aspects of Bulgarian in comparison with other languages with lexical stress').

6. REFERENCES

[1] Studemund-Halévy, M., Fischer, S. 2013. What happens when a language ceases to be used by its speakers? Documentation of Bulgarian Judezmo. In: Studemund-Halévy, M., Liebl, C., Vučina Simović, I. (eds), *Sefarad an der Donau. Lengua y literatura de los Sefardíes en tierras de los Habsburgos*. Tirocinio, 407–424.

[2] Hualde, J. I., Şaul, M. 2011. Istanbul Judeo-Spanish. *J. IPA* 41, 89–110.

[3] Romero, R. 2013. Palatal east meets velar west. Dialect contact and phonological accommodation in Judeo-Spanish. *SHLL* 6, 279–300.

[4] Kanchev, I. 1975, *Fonética y fonología del judeoespañol de Bulgaria*, Ph. D. dissertation. University of Sofia.

[5] Gabriel, C., Kireva, E. 2014. Speech rhythm and vowel raising in Bulgarian Judeo-Spanish. In: Campbell, N., Gibbon, D., Hirst, D. (eds), *Proc. SP 2014*. Dublin, 728–732.

[6] Andreeva, B., Dimitrova, S., Gabriel, C., Grünke, J. 2019. The intonation of Bulgarian Judeo-Spanish spontaneous speech. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proc. 19th ICPhS*. Melbourne, 3827–3841.

[7] Andreeva, B., Dimitrova, S., Gabriel, C., Grünke, J. 2021. Intonational convergence in Bulgarian Judeo-Spanish spontaneous speech. In: Teixeira Kalkhoff, A., Selig, M., Mooshammer, C. (eds), *Prosody and conceptional variation*. Lang, 171–190.

[8] Bradley, T. G. 2007. Constraints on the metathesis of sonorant consonants in Judeo-Spanish. *Probus* 19, 171–207.

[9] Bradley, T. G. 2015. Labialization and palatalization in Judeo-Spanish phonology. In: Smith, J., Ihsane, T. (eds), *Romance Linguistics 2012. Selected papers from the 42nd Linguistic Symposium on Romance Languages*. Benjamins, 63–83.

[10] Bradley, T. 2022. Judeo-Spanish. In: Gabriel, C., Gess, R., Meisenburg, T. (eds), *Manual of Romance phonetics and phonology*. DeGruyter, 808–838.

[11] Bradley, T., Lozano, C. J. 2022. Language contact and phonological innovation in the voiced prepalatal obstruents of Judeo-Spanish. *Languages* 7 (4), 313.

[12] Martínez Celdrán, E. 2007. *Análisis espectrográfico de los sonidos del habla*. Ariel.

[13] Nadeu, M. 2014. Stress- and speech rate-induced vowel quality variation in Catalan and Spanish. *J. Phon.* 46, 1–22.

[14] Tilkov, D. (ed.) 1982. *Gramatika na săvremennija bălgarski knižoven ezik*, Tom I: Fonetika. Sofia: Izd. na Bălgarskata Akademija na Naukite (*Grammar of the Modern Bulgarian Literary Language*, Vol. l: Phonetics).

[15] Ternes, E., Vladimirova-Buhtz, T. 1999. Bulgarian. In: *Handbook of the IPA*. Cambridge University Press, 55–57.

[16] Andreeva, B., Barry, W., Koreman, J. 2013. The Bulgarian stressed and unstressed vowel system. A corpus study. In: Bimbot, F., Fougeron, C., Pellegrino, F. (eds), *Proc. Interspeech 2013*. Lyon, 345–348.

[17] Dokovova, M., Sabev, M., Scobbie, J., Lickley, R., Cowen, S. 2019. Bulgarian vowel reduction in unstressed position. An ultrasound and acoustic investigation. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds), *Proc. 19th ICPhS*. Melbourne, 2720–2724.

[18] Sabev, M. 2020. *Spectral and durational unstressed vowel reduction. An acoustic study of monolingual and bilingual speakers of Bulgarian and Turkish*, DPhil thesis. University of Oxford.

[19] Sabev, M. Forthcoming. Unstressed vowel reduction and contrast neutralisation in western and eastern Bulgarian: A current appraisal. *J. Phon*.

[20] Scatton, E. 1975. *Bulgarian phonology*. Cambridge, MA: Slavica Publishers.

[21] Gabriel, C., Grünke, J. 2018. Is there a standard pronunciation of Bulgarian Judeo-Spanish? Evidence from /e/-to-[i] raising in read and spontaneous speech. In: Muhr, R., Meisnitzer, B. (eds), *Pluricentric languages and non-dominant varieties worldwide. New pluricentric languages – old problems*. Lang, 367–381.

[22] Boersma, P., Weenink, D. 1992–2022. *Praat. Doing phonetic by computer*, v. 6.2. http://www.fon.hum.uva.nl/praat/.

[23] Paschen, L., Fuchs, S., Seifart, F. 2022. Final lengthening and vowel length in 25 languages, *J. Phon.* 94, 101179.

[24] McCloy D. R. 2016. *phonR. Tools for phoneticians and phonologists.* R package version 1.0-7.

[25] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *J. Phon.* 34, 458–484.

[26] Hall-Lew, L. 2010. Improved representation of variance in measures of vowel merger. *Proc. Meetings on Acoustics* 9, 1–10.

[27] Amengual, M., Chamorro, P. 2015. The Effects of Language Dominance in the Perception and Production of the Galician Mid Vowel Contrasts. *Phonetica* 72, 207–236.

[28] Renwick, M. E. L., Olsen, R. M. 2017. Analyzing dialect variation in historical speech corpora. *JASA* 142, 406–421.

[29] Sabev, M., Payne, E. 2019. A cross-varietal continuum of unstressed vowel reduction: evidence from Bulgarian and Turkish. In: S. Calhoun, P. Escudero, M. Tabain, P. Warren (eds.), *Proc. 19th ICPhS*. Melbourne, 1164–1168.

[30] Wang, Y., Sabev, M. Forthcoming. Rime changes and mergers in Beijing retroflex suffixation: An acoustic study. In: *Proc. 20th ICPhS*. Prague.

[31] Stanley, J. A., Sneller, B. 2023. Sample size matters in calculating Pillai scores. *JASA* 153, 54–67.

[32] Thomas, D. R. 1992. Interpreting discriminant functions. A data analytic approach. *Multivariate Behavioral Research* 27(3), 335–362.

[33] Ronquest, R. E. 2013. An acoustic examination of unstressed vowel reduction in heritage Spanish. In: Howe, C., Blackwell, S. E., Lubbers Quesada, M. (eds). *Selected Proc. 15th Hispanic linguistics symposium*. Somerville, 157–171.

Crosslinguistic influence and language-specific strategies in bilingual clear speech

Ye-Jee Jung, Olga Dmitrieva

Purdue University
jung292@purdue.edu, odmitrie@purdue.edu

ABSTRACT

The current study investigated how late Korean-English bilinguals produced clear speech in their L1 (Korean) and L2 (English). Specifically, we examined the acoustic enhancement of laryngeal contrasts in English and Korean clear speech. Thirty Korean-English bilinguals produced casual and clear speech in each of the two languages by reading Korean and English word-lists. In addition, monolinguals of the two languages (20 in each group) produced casual and clear speech under the same experimental conditions, for comparison.

The results showed that bilinguals used onset f0 to enhance the laryngeal categories in a distinctly language-specific way in English and Korean. The use of VOT as a correlate of laryngeal categories in bilingual clear speech suggested mutual crosslinguistic influence between English and Korean. Therefore, the acoustic structure of bilingual clear speech resembled the monolingual counterparts but also carried unique acoustic characteristics.

Keywords: Korean, English, bilingual speech, clear speech, crosslinguistic influence

1. INTRODUCTION

The H & H theory [12] postulates that speech is produced on a continuum between hypo-articulation (less articulatory efforts) and hyper-articulation (more articulatory efforts), depending on the communicative goals and settings. Clear speech is one form of hyper-articulated speech, and it is a distinctive speaking style directed at increasing speech intelligibility in communicatively compromised settings (e.g., the presence of background noise). Increased acoustic difference between phonological categories contributes to speech intelligibility and is typically found among the features of clear speech. Because languages differ with respect to the types and acoustic implementation of phonological categories, clear speech strategies are by necessity at least partly language-specific.

For example, in English, word-initial stop voicing distinction is realized primarily via the voice onset time (VOT) difference: longer VOT for voiceless stops, shorter VOT for voiced stops. This VOT difference is further enhanced in clear speech, typically via the asymmetrical lengthening of voiceless VOT [13]. In Korean, the three-way contrast between lenis, aspirated, and fortis stops is realized via VOT and/or onset f0. One or the other cue becomes more important depending on the specific binary contrast. For example, the lenis-aspirated contrast in modern Seoul Korean is believed to rely primarily on the onset f0 difference. Correspondingly, this cue is further enhanced in clear speech, where lenis and aspirated stops are even more distanced via onset f0 [9].

While these speech behaviours have been relatively well researched in monolingual speech, considerably less is known about clear speech produced by bilingual second language speakers. There are at least two main questions that can be asked with respect to bilingual clear speech. First, to what extent do bilingual speakers maintain language-specific clear speech strategies in both of their languages. And second, to what extent is clear speech subject to bidirectional crosslinguistic influence in bilingual speakers predicted by the Speech Learning Model (SLM/SLM-r) [5, 6]. In order to address these questions, the present study examined how laryngeal contrasts were realized in casual speech and enhanced in Korean and English clear speech produced by Korean (L1) – English (L2) bilinguals, in comparison to Korean and English monolingual speakers.

To our knowledge, previous research on bilingual clear speech has been limited. For example, native Croatian speakers seemingly transferred their L1 clear speech strategies to their L2 (English) [14]. Particularly, they did not lengthen the VOT of English /p/ in clear speech, instead tending towards the enhancement of prevoicing for English /b/. In another study, native Finnish speakers increased the VOT of the English /p/ in English clear speech, enhancing the acoustic distance between the English /p/ and /b/, similarly to native speakers of English [8]. Therefore, existing research suggests that bilingual clear speech can be both language-appropriate and subject to crosslinguistic influence.

Accordingly, we hypothesize that Korean bilinguals can produce English and Korean clear speech where the laryngeal categories are enhanced in a language-specific manner, at least to some extent. At the same time, we also expect to see some evidence of crosslinguistic influence, such that English laryngeal categories are enhanced in a

somewhat Korean-like fashion (e.g. with a significantly greater reliance on onset f0 and weaker reliance on VOT than in monolingual English speech), and vice versa (e.g. Korean laryngeal categories' enhancement primarily via VOT).

2. METHODS

2.1. Participants

Three speaker groups participated in the present study: 20 native American English monolinguals (4 male, mean age = 24.95_{yrs}, SD = 9.30), 30 late Korean-English bilinguals residing in the United States at the time of recording (20 male, mean age = 29.73_{yrs}, SD = 3.48), and 20 Korean monolinguals living in South Korea (8 male, mean age = 27.40_{yrs}, SD = 5.52). Korean and English monolinguals served as control groups for each language of the bilingual speakers.

2.2. Procedures

Participants performed a word-list reading task, with words displayed one by one on the computer screen in random order. Each word remained on the screen for 1.8 seconds. Korean words were presented in Korean orthography. In each speaking style, three repetitions of each word were recorded for each participant. For casual speech, participants were instructed to read each target word casually, as if they were talking to their friends or family members. For clear speech, they were instructed to read each word clearly, as if their interlocutors were hearing-impaired or elderly. Casual speech recording preceded clear speech with a short self-timed break provided between the two recording sessions. Recordings were performed in a sound-attenuated booth in a speech laboratory or in a quiet classroom using professional quality recording equipment.

2.3. Stimuli

For English, six minimal pairs that differed in voicing of the word-initial alveolar stops were used (e.g. *tab* vs. *dab*). For Korean, we used six triplets/near triplets differing in the laryngeal category of the word-initial stops adopted from [9], for example, *t^{h}ant^{h}anhata* 'to be solid' (aspirated), *tantanhata* 'to be hard' (lenis), and *t*ant*anhata* 'to be hard/strong (fortis). Additional distractor items (16 English and 18 Korean) with similar syllabic structure were included in each word-list, e.g. *pan* and *manmanhata*. As a result, 72 English target words (168 including fillers) were and 108 Korean target words (216 including fillers) were recorded per participant per language.

2.4. Acoustic measurements

For each target word, VOT (in milliseconds) of the word-initial stops and onset f0 (in hertz) at the onset of the vowel following the word-initial stop vowel were annotated and measured using custom scripts in Praat [3]. To control for individual variability, including anatomic differences, in f0 production, raw onset f0 values were normalized into semitones (ST) using the formula in [4].

2.5. Statistical analysis

A series of linear mixed-effect models was implemented in R using the *lme4* package [1]. VOT and onset f0 were the dependent variables analysed in separate models. The following sum-coded fixed effects were used for English models: Speaker Group (English monolinguals and Korean-English bilinguals), Speaking Style (casual and clear), and Stop Type (voiced and voiceless) and all two- and three-way interactions. By-participant intercepts and random slopes for Speaking Style and by-item random intercepts were also included. For Korean data, a structurally identical model was used, except Stop Type in this model had three levels (aspirated, lenis, and fortis). Therefore, for the Stop Type effect and every interaction including the effect, the *emmeans* package [11] was implemented to perform pairwise post-hoc analyses. Statistical significance of test and interactions was established with ANOVA type III tests implemented using the *car* package [7]. In reporting the interactions, we focus on those that included Stop Type. For significant effects and interactions, we calculated partial eta squared (η_p^2) using the *effectsize* package as an estimate of effect size [2].

3. RESULTS

3.1. English clear speech

The VOT results for each speaker group and each speaking style are illustrated in Figure 1. The significant effects of Speaking Style and Stop Type were observed: VOT was longer in clear speech ($\chi2$ (1) = 51.43, p < .001, η_p^2 = 0.52) and in voiceless stops ($\chi2$ (1) = 6040.74, p < .001, η_p^2 = 1.00).

All two-way and three-way interactions including Stop Type were also significant. Most importantly, the English group made a greater VOT distinction between two voicing categories than the bilingual group (Stop Type by Group interaction, $\chi2$ (1) = 58.16, p < .001, η_p^2 = 0.02). The VOT distinction between voiced and voiceless stops was greater in clear speech than in casual speech (Stop Type by Speaking Style interaction, $\chi2$ (1) = 362.76, p < .001,

η_p^2 = 0.10). As Figure 1 demonstrates, the enhancement was achieved via the asymmetrical lengthening of voiceless stops' VOT. On the other hand, VOT of voiced stops remained stable across speaking styles in both speaker groups. Finally, the magnitude of clear speech voicing enhancement via VOT was greater for the English group compared to the bilingual group (Stop Type by Group by Speaking Style interaction, $\chi 2$ (1) = 101.71, $p < .001$, η_p^2 = 0.03).

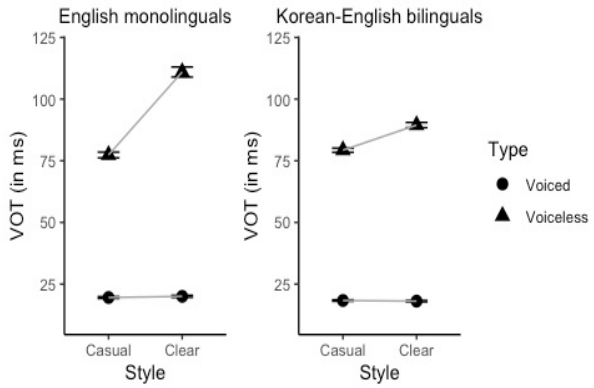


Figure 1: VOT of English stops produced by native English monolinguals (left) and Korean-English bilinguals (right) in each speaking style.

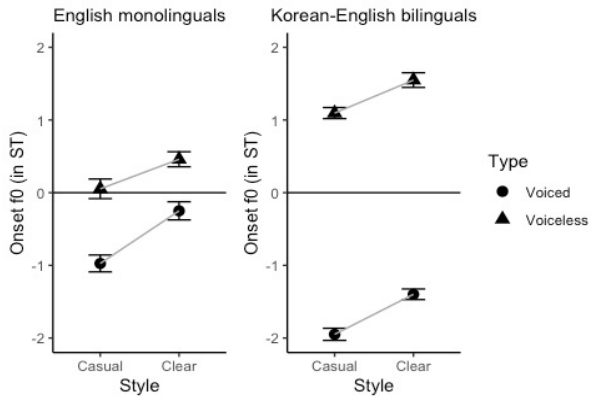


Figure 2: Onset f0 of English stops produced by native English monolinguals (left) and Korean-English bilinguals (right) in each speaking style.

Figure 2 demonstrates the onset f0 distinction between the two stop types in each speaker group and each speaking style. Speaking Style and Stop Type effects were significant: onset f0 was higher in clear speech ($\chi 2$ (1) = 56.53, $p < .001$, $\eta_p^2 = 0.02$) and after voiceless stops ($\chi 2$ (1) = 254.01, $p < .001$, $\eta_p^2 = 0.96$).

The significant interaction between Speaker Group and Stop Type was attested due to the fact that bilingual speakers made a greater onset f0 distinction between voiced and voiceless stops than English monolinguals ($\chi 2$ (1) = 224.44, $p < .001$, $\eta_p^2 = 0.06$), as shown in Figure 2. However, no other interactions were significant, indicating that onset f0 difference between voiced and voiceless stops was not enhanced in clear speech, compared to casual speech, in either of the speaker groups.

3.2. Korean clear speech

Figure 3 shows the VOT results for Korean casual and clear speech for each speaker group. The effect of Stop Type was significant ($\chi 2$ (2) = 239.74, $p < .001$, $\eta_p^2 = 0.98$). Post-hoc tests indicated that the VOT of aspirated and lenis stops was significantly longer than the VOT of fortis stops [aspirated-fortis: $\beta = 52.43$, $p < .001$; fortis-lenis: $\beta = -49.48$, $p < .001$]. Lenis and aspirated stops did not differ from each other in terms of VOT. Average VOT was also longer in clear speech than in casual speech ($\chi 2$ (1) = 15.12, $p < .001$, $\eta_p^2 = 0.24$).

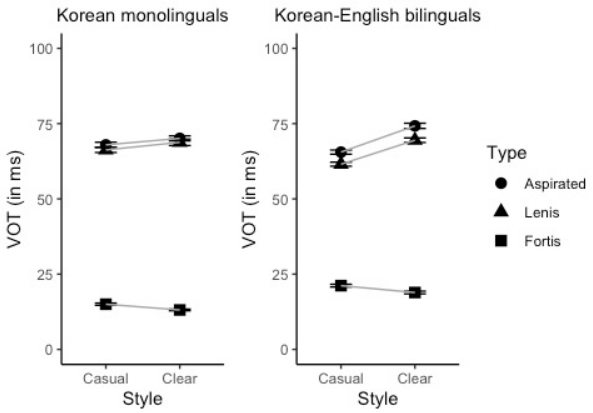


Figure 3: VOT of Korean stops produced by Korean monolinguals (left) and Korean-English bilinguals (right) in each speaking style.

A number of significant interactions indicated variability in the way each speaker group realized and enhanced laryngeal categories. First, aspirated and lenis VOT was lengthened in clear speech while fortis VOT was shortened in clear speech (Stop Type by Speaking Style interaction, $\chi 2$ (2) = 106.75, $p < .001$, $\eta_p^2 = 0.02$). Second, bilinguals made a smaller VOT difference between laryngeal categories than monolingual speakers (Group by Stop Type interaction, $\chi 2$ (2) = 98.02, $p < .001$, $\eta_p^2 = 0.02$). Finally, the enhancement of VOT differences between the laryngeal categories was more pronounced in bilingual than in monolingual clear speech (Stop Type by Speaking Style by Group interaction, $\chi 2$ (2) = 19.98, $p < .001$, $\eta_p^2 = 0.00$).

As Figure 4 illustrates, Korean stop types were well-dispersed in terms of onset f0 for both speaker groups. The main effects of Stop Type ($\chi 2$ (2) = 2139.26, $p < .001$, $\eta_p^2 = 1.00$) and Speaking Style ($\chi 2$

(1) = 214.56, $p < .001$, $\eta_p^2 = 0.04$) were significant. Post-hoc tests indicated that onset f0 of aspirated stops was higher than that of both lenis and fortis stops (aspirated-fortis: $\beta = 2.63$, $p < .001$; aspirate-lenis: $\beta = 4.63$, $p < .001$) while lenis f0 was lower than fortis f0 ($\beta = 2.00$, $p < .001$). Average onset f0 was also higher in clear than in casual speech.

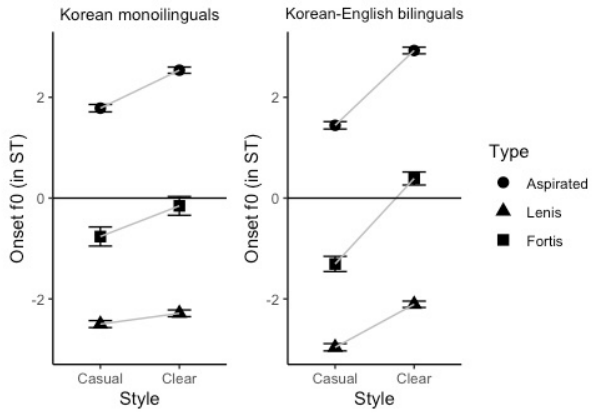


Figure 4: Onset f0 of Korean stops produced by Korean monolinguals (left) and Korean-English bilinguals (right) in each speaking style.

There was a significant two-way interaction between Speaking Style and Stop Type ($\chi 2$ (2) = 19.77) = 9,87, $p < .001$, $\eta_p^2 = 0.00$). Onset f0 of every stop type was raised in Korean clear speech, but the magnitude of the increase was greater for aspirated and fortis stops than for lenis stops, resulting in an enhanced f0 difference between lenis and other types of stops in clear speech.

4. DISCUSSION AND CONCLUSIONS

The presented study investigated native and non-native clear speech produced by late Korean-English bilinguals, focusing on language-specific strategies and crosslinguistic influence. The results, first and foremost, indicated that the experiment was successful in eliciting clear speech. As could be expected based on previous research [9, 10, 13], average VOT was lengthened and average f0 was raised in both Korean and English clear speech across speaker groups.

Even more importantly, for both cues, this increase was often asymmetric: VOT was lengthened for voiceless but not voiced stops in English and for lenis and aspirated stops but not fortis stops in Korean. Similarly, onset f0 was raised for aspirated and fortis stops to a greater extent compared with lenis stops in Korean. As a result of these manoeuvres, the difference between contrastive phonological categories in both languages was acoustically enhanced in clear speech.

Finally, as predicted, we observed that bilingual clear speech was both language-specific/language-appropriate and subject to crosslinguistic influence.

We can deduce language-specificity by observing what bilinguals did differently across their two languages, especially if the difference is in line with the corresponding monolingual behaviours. The clearest example is the use of onset f0 to enhance laryngeal contrasts in clear speech. Bilinguals did not increase onset f0 difference between voiced and voiceless English stops in clear speech, similarly to native speakers of English. Yet bilinguals did increase onset f0 difference between Korean laryngeal categories, just as monolingual speakers of Korean did. These findings suggest a great sensitivity on the part of bilingual speakers to the way the same cue is used in different languages and awareness of the potential utility of this cue in increasing speech intelligibility in each language.

Crosslinguistic influence can be established by observing what bilinguals did differently from monolingual speakers of the same language, especially if this difference is in line with the behaviour of monolinguals of another language. At least two findings to that effect can be noted. First, bilingual speakers realized a significantly greater onset f0 difference between voiced and voiceless English stops than English monolinguals (although this difference was not further enhanced in clear speech). This behaviour is explainable by the fact that a subset of Korean laryngeal distinctions is cued by rather sizeable f0 differences.

Second, in clear speech, bilinguals produced *less* enhancement of VOT differences between English voicing categories than English monolinguals, but *more* enhancement of VOT differences between Korean laryngeal categories than Korean monolinguals. As such, bilinguals' clear speech enhancement behaviour with respect to VOT was 'intermediate' between the corresponding monolingual patterns. Such 'intermediate' patterns are often found in the phonetics of bilingual speech. This finding is explainable by the fact that VOT plays a more important role in cueing and enhancing laryngeal categories in English than in Korean.

It remains to be understood why onset f0 was used in a language-specific fashion while the use of VOT was subject to crosslinguistic influence. Notwithstanding, bilingual results indicate remarkable sensitivity to crosslinguistic differences, as well as sophistication and flexibility in navigating these differences in adaptive speech behaviours.

5. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.
[2] Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815.
[3] Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03
[4] Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset
time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77-95.
[5] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research, 92*, 233-277.
[6] Flege, J., & Bohn, O. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3-83). Cambridge: Cambridge University Press.
[7] Fox J, Weisberg S. (2019). An R Companion to Applied Regression, Third edition. Thousand Oaks, CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
[8] Granlund, S., Hazan, V., & Baker, R. (2012). An acoustic–phonetic comparison of the clear speaking styles of Finnish–English late bilinguals. *Journal of Phonetics, 40*(3), 509-520.
[9] Kang, K. H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *The Journal of the Acoustical Society of America*, *124*(6), 3909–3917.
[10] Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, *115*(1), 362-378.
[11] Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0. https://CRAN.R-project.org/package=emmeans.
[12] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. *Speech production and speech modelling*, 403-439.
[13] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 29*(4), 434-446.
[14] Smiljanic, R., & Bradlow, A. R. (2009). Native and non-native clear speech production. *The Journal of the Acoustical Society of America*, *125*(4), 2753-2753.

Cross-linguistic influence in the mid-vowels of Hausa/Ekhwa speakers[i]

Emmanuel Bawa[1], Rachel Steindel Burdin[2] & Taylor Cassiani[2]

[1]The Graduate Center, CUNY| [2]University of New Hampshire
ebawa@gradcenter.cuny.edu, rachel.burdin@unh.edu, taylor.cassiani@unh.edu

ABSTRACT

This paper examines contact between the vowel systems of Hausa (Afro-Asiatic) and Ekhwa (Niger-Congo). Hausa has ten monopthongs, including a phonemic length contrast, and previous descriptions of Ekhwa suggest 7 monopthongs, with a potential tense/lax distinction in the mid vowels. We examine Hausa as it is spoken by Ekhwa speakers, using wordlist data gathered remotely, focusing on the Hausa mid vowels /e, e:, o, o:/, and purported Ekhwa /e, ɛ, o, ɔ/, measuring F1, F2, and duration. We find first that there does not appear to be phonetic evidence for a tense/lax distinction in this variety of Ekhwa. Secondly, while there is limited evidence for a spectral distinction between the long and short mid vowels in this variety of Hausa, durational differences were mixed. Finally, to the extent that there are spectral differences between the long and short vowels in Hausa, they are made within the Ekhwa vowel space.

Keywords: Hausa, Ekhwa, vowel length, language contact

1. INTRODUCTION

1.1 Vowel spaces of bilinguals

It has long been observed that the phonetic and phonological systems of bilinguals, particularly late bilinguals, often differ from those of monolinguals. This is explained in terms of interference from the L1 and/or more dominant language [1], [2]. The degree of interference is influenced by many factors, with age of acquisition being a major factor. For example [3] showed that early Quichua/Spanish bilinguals not only were able to produce a distinction between Spanish /i/ and /e/ (a contrast not present in Quichua), but also their Spanish /i/ and Quichua /ɪ/. However, those who learned Spanish later in life showed a near-overlap between Spanish /i/, Quichua /ɪ/, and Spanish /e/.

A series of studies have also studied distinctions in the vowel spaces of Catalan/Spanish bilinguals. Catalan has four mid vowels, /ɛ/, /e/, /ɔ/ and /o/, vs. only two, /e/ and /o/, in Spanish. [4] found differences in the phonetic implementation of these vowels within early bilinguals, with those who reported using Catalan more than Spanish having a larger phonetic distinction between their Catalan mid vowels compared to those who reported using Spanish more. These studies show that, even in communities with long term bilingualism, where many speakers acquire both languages relatively early in life, there may still be cross-linguistic influence in the phonetic space.

This study builds on this past work by focusing on a relatively neglected field in contact studies: contact between indigenous languages in Africa. While there are some areas of intensive research (e.g., creole studies) [5] notes contact in Africa is understudied and calls for more work in this area, especially involving undocumented languages. This study, in addition to providing data on the linguistic behaviour of early bilinguals more broadly, helps address this gap in the literature, and provides data from two understudied varieties: a non-standard variety of Hausa, and Ekhwa.

1.2 Hausa and Ekhwa in Kaduna State

Hausa is a West Chadic language of the Afro-Asiatic language family. It is one of Nigeria's regional languages, spoken mainly in northern Nigeria by Hausa people and members of other ethnic groups, including in Kaduna state, where this research is situated. Hausa has been well-documented, with at least two extensive grammars in English [6], [7].

In Kaduna state, apart from L1 Hausa speakers, other groups speak Hausa as a second language, including the Adara people. Ekhwa, also called Ehwa, Ankwa, or Iku-Gora-Ankwa, spoken mainly in the Kachia district area, is a variety of Adara, which is a cluster of related languages in the Niger-Congo family. The exact classification and status of Ekhwa is unclear. Adara appears to be a cluster of at least five varieties, with a range of mutual intelligibility and relatedness, with Ekhwa being more divergent from the other varieties [8], [9]. Ekhwa has not been well-studied or documented, apart from some wordlist data, and brief overviews in larger projects [8]–[10].

Hausa is politically dominant in the region, and thus, other groups have adopted the language in a wide variety of spheres outside the home, including for worship, business, entertainment, news, and education. However, Hausa as spoken by Ekhwa speakers shows some clear differences from

Standard Hausa, which appear to be contact-induced effects from Ekhwa. For example, emphatic stops, which Ekhwa lacks, are often produced as plain stops, e.g., [wúk'á:] 'knife' becomes [wúká:]. Hausa has two rhotic sounds, [r] and [ɽ]; Ekhwa appears to have a single, non-trilled rhotic, and thus, the rhotics are merged, e.g., [k'ù:rá:] 'dust' becomes something like [kù:ɽa:] or [kù:ɾá:]. Additionally, there are influences from Hausa into Ekhwa, particularly loanwords, such as [ɔ̃kɔ́fâ] 'door', from Hausa <kofa>. Given these facts, we might also expect to see contact effects in the vowel space.

1.3 Hausa vowels

Standard Hausa has ten monopthongal vowels /a, a:, e, e:, i, i:, e, e:, o, o:, u, u:/. Length distinctions are only present in open syllables. Additionally, the distinction between /e, e:/ and /o, o:/ mostly appears word finally; only /e:/ and /o:/ can appear word medially (with the exception of word medial closed syllables). The length distinction between vowels has been described as being produced with both spectral and durational differences [7]. [11], in a study of /i, i:, u, u:, a, a:/ found both durational and spectral differences between the long and short vowels. /i:/ and /a:/ were about 50% longer than /i/ and /a/ in open syllables before voiceless consonants (106 ms vs. 67 ms, and 118 vs. 71ms, respectively). [12], in a study focused on pre-pausal productions of /a, a:/ and /o, o:/, also found durational differences of about the same magnitude, with long vowels between 45-50% longer than short vowels.

The impetuous for this study was a Field Methods course focusing on Hausa, in which the first author, a speaker of Ekhwa and Hausa, produced long /e:/ and /o:/ that, to the second author, an American English speaker, sounded almost /i/ and /u/-like, e.g., <na gode> 'thank you' sounded like [ná gúdì]. As far as we could tell, this raising process had not been previously noted, and given other features of his Hausa which appeared to be influenced by Ekhwa, as noted above, we decided to investigate the production of long and short vowels in Hausa as spoken by Ekhwa speakers.

1.4 Ekhwa vowels

Ekhwa, and Adara in general, have had minimal descriptive linguistic work. [8] provides a wordlist collected by Alex Maikarfi, but cautions that "it does not seem worth attempting an analysis of the phonology" (pg. 2). The transcriptions of the words include seven vowels, [a, e, ɛ, i, o, ɔ, u], with occasional doubled vowels, including [oo], [ɛɛ], and [uu]. A preliminary writing system for Ekhwa proposes <a, e, i, o, u, ẹ, ọ>, with <ẹ> described as sounding as in English 'slept', 'kept', and 'get', and <ọ> as in English 'pot' and 'hot' [13]. [14] suggests a 7 vowel system for Adara, with two mid vowels, which we take as our starting point; however, the first author's intuitions suggest a 5 vowel system instead. The first goal of this work, then, was to investigate the presence or absence of a contrast in the mid vowels in Ekhwa, and from there, given this system, investigate how the length contrast in Hausa mid vowels was produced by Ekhwa speakers.

2. METHODOLOGY

2.1 Data collection

A list of target words was created for both Hausa and Ekhwa. The 29 Hausa target words contained the vowels /a, a:, e, e:, i, i:, o, o:, u, u:/ in medial and final positions, with both high and low tone. The 32 Ekhwa target words contained the vowels /a, i, u/, as well as the mid vowels, in word initial, medial, and final position, with both low and high tone where possible. We will be focusing on the mid vowels.

Data were collected, with assistance from a local contact, remotely over the internet, using Cleanfeed, in Fall 2021. Cleanfeed allows for recording directly from a participant's device, avoiding the effects of compression of files from services like Zoom. Remote data collection was necessary due to COVID-19 travel restrictions between the United States and Nigeria, as well as travel restrictions within Nigeria due to on-going conflict in the northern part of the country.

The local contact travelled to three locations to recruit participants in southern Kaduna State. Participants, who were asked if they spoke Ekhwa as their tribal language and Hausa, were recorded in relatively quiet locations in each, using a laptop. Participants were compensated with a cell phone data plan credit worth approximately 5000 NGN.

The words were shown to the participants in a slideshow with a picture of the intended target word, and a carrier phrase (*Nache X a Hausa* "I say X in Hausa"; *Ega X mi piya* "I call this X in my dialect" in Ekhwa) with the target word written beneath the picture. There is not a standard writing system for Ekhwa, although at least one systems has been proposed [13], and people can and do read and write Ekhwa using Roman characters. Written Hausa does not mark tone or vowel length, and including pictures helped distinguish heteronyms in isolation. The pictures also allowed for participation from non-literate participants.

Participants recited the words in the carrier phrase three times, starting first with Hausa. Participants were then asked demographic

information, and about their acquisition and use of Ekhwa, Hausa, and English in different domains. This portion of the interview was conducted in English if the participant was comfortable in that language, or in Hausa or Ekhwa if they weren't.

10 female and 10 male participants were recorded (average age: 32). 19 were from Kachia or nearby. 16 reported learning Ekhwa at home; the rest reported learning Ekhwa in the community, but 3 out of these remaining 4 did not report learning another language in the home. All 20 reported speaking Hausa, and learning it either in the community or at school; 18 reported also speaking English, with most (17) reporting learning it at school. Most (14) reported that Ekhwa was the language they used most at home, followed by Hausa (5), and Ejegha (another Adara variety)/Hausa (1). Hausa predominated in marketplace contexts (all 20). For work and school, English (13) and Hausa (6) predominated. In general, the participants were early sequential, bilinguals, with a variety of Ekhwa acquired, and still spoken, in the home and some community contexts, and Hausa learned slightly later, and used as a lingua franca, particularly in the marketplace, as well as at school or at work.

2.2 Data analysis

The target words were segmented using Praat [15]. A script extracted the duration as well as the F1 and F2 at vowel midpoint. F1 and F2 values that were more than 1 standard deviation from the mean F1 or F2 (by speaker and vowel) were hand checked and corrected or discarded if reliable values could not be extracted; the F1 and F2 values were then z-score normalized by speaker [16].

Participants occasionally produced a different form of a word (e.g., [járè] 'sweep' instead of [èjéré] 'broom'), forms from other Adara dialects (e.g., [ànésésén] 'white person' instead of target [ànésè]), non-target heteronyms (e.g., [bà:kí:] 'mouth' instead of [bák'í:] 'black'), along with other mispronunciations or disfluencies; these were excluded. Other tokens were excluded due to background noise, internet connectivity issues, or some type of mispronunciation not covered by the above. Out of a theoretical 8640 tokens (144 vowels x 20 participants x 3 repetitions), we were left with 5818 tokens total. From this, we examined 390 Hausa /e(:)/ tokens, 330 long, and 60 short; 363 Hausa /o(:)/ tokens, 324 long, and 39 short; 737 Ekhwa /e/ tokens, and 666 Ekhwa /o/ tokens. The long and short tokens in Hausa are unbalanced due to the phonotactic restrictions outlined above, with only long vowels appearing word medially.

3. RESULTS

3.1 Ekhwa vowels

Using wordlist data from [8] where available, words with a front mid vowel were classified as either being [e] or [ɛ]; this was 388 tokens. Linear mixed effects models were built using lmer [17] in R [18] predicting F1 and F2, with fixed effects for [e]/[ɛ] status, following segment place of articulation, and height of the following vowel for F1, and backness of the following vowel for F2. All fixed effects were treatment coded (reference levels: [ɛ], alveolar, high vowels, back vowels). None of these factors were found to be significant. As can be seen in figure 1, there is considerable overlap between the tokens tagged as [e], and those tagged as [ɛ]. A Pillai score [19] was also calculated, and was 0.14162, confirming the large degree of overlap.

Similar models were built for the back mid vowels, classifying tokens as [o] or [ɔ]; a total of 340 tokens were included. Fixed effects (all treatment coded) included [o]/[ɔ] status (reference level: [ɔ]), and following segment place of articulation (reference level: alveolar); the structure of the data was such that effects for following vowel height or backness could not be included. [o]/[ɔ] status was not significant in either the F1 or F2 model. As can be seen in figure 2, there was considerable overlap between the tokens tagged as [o] vs. [ɔ]; the Pillai score was 0.19819. These findings thus suggest only two mid vowels in this variety of Ekhwa.

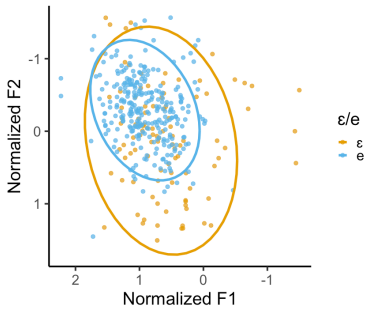


Figure 1: Ekhwa front mid vowels

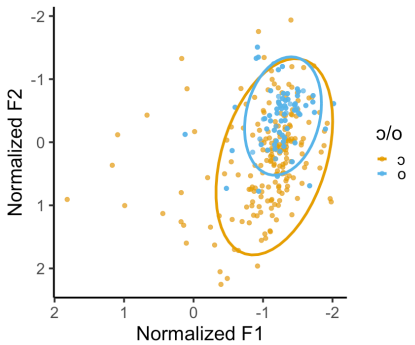


Figure 2: Ekhwa back mid vowels

3.2 Hausa vowels

Linear mixed effects models were first built predicting duration of /e, e:/ in Hausa, in ms., with length, slide number, and position in the word (medial vs. final) as fixed effects (treatment coded; reference levels: long vowels, final vowels), and random intercepts by speaker. /e/ was significantly *longer* than /e:/ ($\beta = 10.3546$, $t = 2.470$, $p < 0.05$); in addition, there was a small effect of slide number, with some lengthening as the task went on ($\beta = 0.7601$, $t = 2.317$, $p < 0.05$). For /o, o:/, /o/ was significantly shorter than /o:/ ($\beta = -24.7322$, $t = -4.088$, $p < 0.001$). However, these differences were smaller than in [11], [12]: /o:/ was, on average, 124 ms, and /o/, 94.5 ms, about a 30% increase.

For the F1 and F2 models, random intercepts were included by word and speaker; fixed effects were included as the structure of the data allowed. For F1 of /e, e:/, the fixed effects were length and place of articulation of the preceding segment (treatment coded; long vowels, alveolars). /e/ had higher F1 than /e:/ ($\beta = 0.1672$, $t = 2.139$, $p < 0.05$). For F1 of /o, o:/, there was a fixed effect of length (treatment coded; long vowels). /o/ had higher F1 than /o:/ ($\beta = 0.3556$, $t = 3.543$, $p < 0.001$). For F2 for both vowels, there was a fixed effect for length (treatment coded; long vowels). There was no significant effect of length for /e, e:/. /o/ had higher F2 than /o:/ ($\beta = 0.23911$, $t = 3.684$, $p < 0.001$).

Figure 3 shows these effects, with /e:/ being higher than /e/; and /o:/ being backer and higher than /o/. This aligns with impressionistic observations, with, e.g., <goro> 'kolanut' sounding like [gúrù].

Finally, we compared the vowel space of the Hausa vowels of these speakers vs. their Ekhwa vowels. Models were built predicting F1 and F2 based on vowel status (Ekhwa, Hausa long, Hausa short), with Ekhwa as the reference level and preceding segment place of articulation as a fixed effect (alveolar as reference level), with random intercepts by speaker and word. The Hausa vowels were not significantly different from the Ekhwa vowels in F1 or F2. As figure 4 shows, to the extent that differences in F1 and F2 are being made between /e, e:/ and /o. o:/ in Hausa, they are within the vowel space of Ekhwa /e/ and /o/.

4. CONCLUSIONS

We find first a lack of evidence for an overall phonetic distinction between tense and lax mid vowels in Ekhwa as produced by these speakers, which aligns with the intuitions of the first author. Secondly, there is mixed evidence for production of a length distinction in Hausa by these speakers. For duration, while /o, o:/ showed the expected effect for length, the magnitude of this effect was smaller than has previously been reported for Hausa; /e, e:/ showed an effect, but in the opposite direction than expected. For spectral differences, there was some evidence for raising of /e:/ and /o:/, confirming auditory impressions. This distinction was made within the Ekhwa vowel space, making these speakers look more like the late bilinguals in [3] or the Spanish-dominant bilinguals in [4]: that is, they do not have appeared to have created separate phonetic spaces for either the Hausa long or short vowels which are distinct from their Ekhwa vowels. This effect may be caused by the fact that Ekhwa was generally acquired first by these speakers; ideologies about the two languages may also be affecting the degree to which speakers might want to speak a more Standard-sounding Hausa. Future work will explore individual differences among the speakers here, as well as the rest of the vowel space.

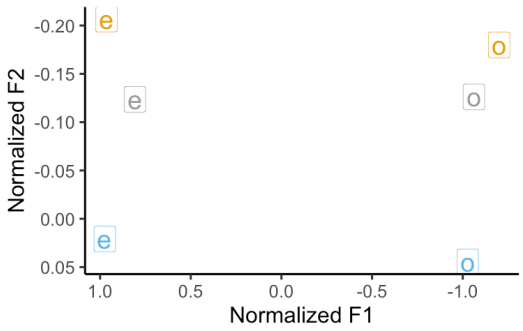


Figure 3: Means, mid vowels. Grey = Ekhwa, blue = Hausa short vowels, orange = Hausa long vowels

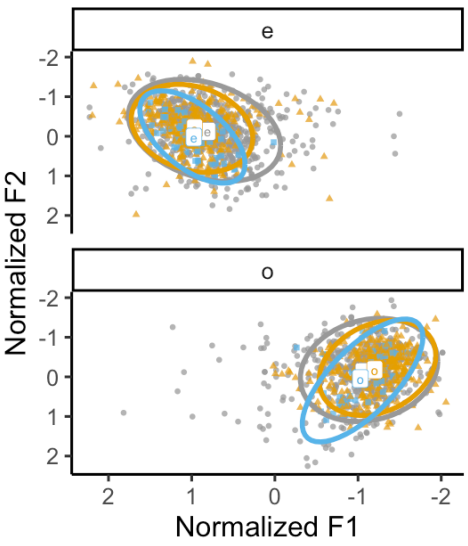


Figure 4: All mid-vowels. Grey = Ekhwa, blue = Hausa short vowels, orange = Hausa long vowels

5. REFERENCES

[1] F. van Coetsem, *Loan Phonology and the Two Transfer Types in Language Contact*. Providence R.I., USA: Foris Publications, 1988.

[2] U. Weinreich, *Languages in contact: Findings and problems*. The Hague: Mouton de Gruyter, 1968.

[3] S. G. Guion, "The Vowel Systems of Quichua-Spanish Bilinguals," *Phonetica*, vol. 60, no. 2, pp. 98–128, Jun. 2003, doi: 10.1159/000071449.

[4] M. Simonet, "Production of a Catalan-Specific Vowel Contrast by Early Spanish-Catalan Bilinguals," *Phonetica*, vol. 68, no. 1–2, pp. 88–110, Jul. 2011, doi: 10.1159/000328847.

[5] G. T. Childs, "Language contact in Africa: A selected review," in *The Handbook of Language Contact*, R. Hickey, Ed. John Wiley & Sons, 2013, pp. 695–713.

[6] P. J. Jaggar, *Hausa*. John Benjamins Publishing, 2001.

[7] P. Newman, *The Hausa language: An encyclopedic reference grammar*. 2000.Yale University Press

[8] R. Blench, "The Ehwa language of Central Nigeria." Unpublished draft manuscript. 2009.

[9] L. Hon, G. Ajaegbu, C. Magnusson, U. S. Nweke, and Z. Yoder, "A sociolinguistic survey of the Adara of Kaduna and Niger States, Nigeria." SIL International, 2018.

[10] R. Blench, "Prospecting Proto- Plateau,". Unpublished draft manuscript. 2008.

[11] M. Lindau-Webb, "Hausa Vowels and Dipthongs," *Stud. Afr. Linguist.*, vol. 16, no. 2, pp 161-182. 1985.

[12] R. M. Newman and V. J. Van Heuven, "An Acoustic and Phonological Study of Pre-Pausal Vowel Length in Hausa," *J. Afr. Lang. Linguist.*, vol. 3, no. 1, pp. 1–18, 1981, doi: 10.1515/jall.1981.3.1.1.

[13] "Reading and writing Ehua/Obai no wow Ehua: A proposal for writing the Ehua language." Author unknown. 2021.

[14] R. Blench, "The Eda [=Kadara] language of Central Nigeria." Unpublished draft manuscript. 2009.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer . Version 6.0.12." 2016. [Online]. Available: http://www.praat.org/

[16] B. M. Lobanov, "Classification of Russian Vowels Spoken by Different Speakers," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 606–608, Feb. 1971, doi: 10.1121/1.1912396.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[18] R Core Team, "R: A language and environment for statistical computing." Vienna, Austria, 2013.

[19] L. Hall-Lew, "Improved representation of variance in measures of vowel merger," *Proc. Meet. Acoust.*, vol. 9, 2010, doi: 10.1121/1.3460625.

[i] The authors would like to thank Titus Hwere for assistance with data collection. This research was partially supported by the UNH McNair Scholars program.

Foreign-accented phonetic detail in L1 word processing: L1 Spanish L2 English listeners and English-accented Spanish

Hadley Forst[1], Miquel Simonet[2]

University of Arizona[1, 2]
hforst@arizona.edu, simonet@arizona.edu

ABSTRACT

This study investigates to what extent English-like VOT in Spanish words affects lexical access for L1 Spanish listeners with English as an L2. Will English-accented Spanish words be recognized as easily as words with Spanish-like VOT? L1 Spanish L2 English bilinguals born and raised in Mexico completed an auditory lexical decision task online. Participants heard /p/-initial Spanish words and nonwords featuring English-like (long-lag) VOT or Spanish-like (short-lag) VOT and indicated whether each item was a real word or not. Results suggest that, for these listeners, English-like VOT in Spanish words causes a processing delay at best, with significantly slower response times to words with English-like VOT than words with Spanish-like VOT, and at worst prevents lexical access entirely, as Spanish words with English-like VOT were accepted as real words in fewer than half of the trials. These findings indicate that, for these bilinguals, L2 phonetic detail is not included in L1 word entries.

Keywords: accented speech perception, phonetic detail, lexical representation, bilingualism, Spanish

1. INTRODUCTION

It is well known that accented speech can be challenging to understand. A growing body of evidence suggests that phonetic detail is stored in the lexicon [1-6], which helps explain why phonetic variation, like that in accented speech, affects lexical access [7-9]. When a listener is confronted with phonetic variants which deviate from the ones they are accustomed to hearing, mismatches between the signal and information in the listener's lexical representation or expectations may cause processing delays and interfere with lexical access [10-12]. Low accuracy on word judgment tasks and delayed lexical access have been found in studies with monolinguals listening to foreign-accented speech [10, 13], L1 regional accents [6, 12] and artificially controlled accents [14]. However, when confronted with variation that is familiar, such as after exposure to a particular accent in the lab, or when listening to an accent that one is accustomed to hearing, the processing delay and interference decreases or disappears [6, 10-14].

Studies on pronunciation and dialect variant recognition have reached similar conclusions: experience with certain types of phonetic variants reduces or eliminates any processing delay present for listeners initially unfamiliar with those variants. This experience with variant phonetic forms contributes to the formation of lexical representations that are sensitive to phonetic detail [1-6]. For example, hearing common pronunciation variants of words—such as the casual pronunciation "cenner" for the English word *center*—does not cause a delay in lexical access when compared with hearing the canonical form "center" [5]. Furthermore, stronger activation of the target word may be triggered by the more frequent variant, indicating that lexical entries store detailed phonetic information, and that a single lexical entry can be linked to several phonetic forms. Similarly, there is evidence that, due to their linguistic backgrounds, speakers of some dialect varieties may have multiple representations for dialect variants (such as "slender" for General American English and "slenda" in a regional New York dialect of English) and can activate the intended word in their lexicon via either phonetic variant, whereas speakers of other dialect varieties cannot [6]. For researchers interested in bilingual speech processing and representation, these findings suggest that some bilinguals may possess multiple representations of words in their lexicons, including accented and unaccented phonetic forms, just as some bidialectal speakers do.

Research into second-language (L2) perception of accented speech also reveals a processing benefit associated with accent familiarity [15-17]. Weber, Broersma and Aoyagi [17] found that while listening in L2 English, native (L1) Dutch listeners processed Dutch-accented English words faster than Japanese-accented English words. The authors conclude that linguistic experience with the phonetic features of a particular accent, such as being a speaker and frequent overhearer of the accent in question, allows for faster lexical access and word recognition of accented speech. In a different study of bilingual word recognition, Shea [18] found that L1 English L2 Spanish bilinguals showed similar levels of lexical activation when hearing both English-accented and Spanish-accented variants of Spanish words (words

featuring either an intervocalic voiced stop or intervocalic spirant, respectively). L1 English L2 Spanish learners are known to produce the stop instead of the spirant intervocalically, at least in initial stages of acquisition, and may be accustomed to hearing it in the speech of other L2 Spanish speakers. In contrast, the native speakers in Shea's study only showed lexical activation from the Spanish-accented forms (with intervocalic spirant), and not the English-accented forms. These results also suggest that bilinguals store phonetic information in the lexicon: the L1 Spanish listeners were unable to match the signal to their representation because they had insufficient experience with the phonetic variants characterizing English-accented Spanish. In contrast, the L1 English L2 Spanish speakers have had ample experience with English-accented Spanish, being both producers and overhearers of it, and they were able to match the signal to an existing representation.

The present study examines English-accented Spanish word recognition, but with a focus on a different salient characteristic of English-accented Spanish. In word-initial position, L1 English L2 speakers of Spanish often produce an English-like [p^{h}], with aspiration or long-lag voice onset time (VOT), instead of Spanish-like [p], with short-lag VOT. Thus, L1 English speakers commonly pronounce Spanish words like *pelo* 'hair' as [p^{h}]elo instead of [p]elo.

Instead of examining the processing of L1-accented L2 speech, the present study focuses on L2-accented L1 speech and tests a new population: L1 Spanish L2 English listeners. Learning more about how bilinguals recognize their L1 as likely produced by native speakers of their L2 will help fill in a gap in the research. The participants in our study are neither producers nor frequent overhearers of English-accented Spanish, as they are L1 Spanish speakers who have lived in Mexico since birth. However, they have had exposure to English, as it is their L2, and may implicitly be familiar with some basic pronunciation differences between Spanish and English. Do L1 Spanish L2 English speakers recognize English-accented Spanish words as readily as Spanish-accented words? If so, this would indicate they had formed lexical representations including English-accented detail. Or, alternatively, will they show a processing delay upon hearing English-accented Spanish words?

L1 Spanish L2 English listeners completed an auditory lexical decision task featuring /p/-initial Spanish words differing in their phonetic detail. Recordings of critical words and pseudowords were manipulated so that each item had two phonetic realizations: one with Spanish-like, short-lag VOT, [p], and one with English-like, long-lag VOT, [p^{h}]. Due to their relatively low experience with English-accented Spanish, we hypothesized that our L1 Spanish listeners would show longer response times (RTs) for word tokens featuring long-lag VOT than for word tokens featuring short-lag VOT.

2. METHOD

2.1. Participants

Forty-five participants were recruited online through the participant recruitment platform Prolific [20]. Participants were L1 Spanish L2 English speakers who were born and raised in a monolingual Spanish environment in Mexico and had never resided outside of Mexico for more than six months. Scores from the History and Attitudes modules of the Bilingual Language Profile questionnaire [19] indicate that all participants were dominant in Spanish. Responses to the English proficiency module of the BLP [19] indicate a median self-assessment of 4.5 on a scale of 0–6 (*0 = not well*, *6 = very well*) when asked how well they speak, understand, read, and write in English (M = 4.3, SD = 0.9, *min* = 2.25, *max* = 6).

2.2. Stimuli

Disyllabic /p/-initial Spanish words with penultimate stress were used as auditory stimuli for a lexical decision task. The 20 most frequent words meeting these criteria were chosen from the stimuli search engine NIM [21], based on the relative frequency-per-million-words for each word. To create a balance between 'yes' and 'no' responses in the lexical decision task, 20 disyllabic /p/-initial pseudowords were generated, as well as 160 fillers. The fillers were evenly divided between real words and pseudowords, all disyllabic and beginning with vowels, fricatives, nasals, laterals and stops, including /b/ and /p/.

An English-dominant Spanish-English early bilingual talker (with phonetic training) was recorded in a sound-attenuated booth using professional equipment at a sampling rate of 44.1 kHz. The talker repeated each item in the carrier phrase "__ también es palabra" (__ *is also a word*) three times each in random order. After recording all items without any specific instructions regarding pronunciation, the talker was asked to produce the critical words and pseudowords an additional three times but was instructed to "exaggerate" the VOT of the initial /p/ to imitate an aspirated "English" word-initial [p^{h}].

Stimuli were resynthesized in Praat [22]. The VOT values of the initial /p/ of the 40 critical items (20 words and 20 pseudowords) were manipulated using a progressive cutback and replacement method [23]. This method begins with recorded minimal pairs, such as [p]elo and [p^{h}]elo. The VOT was then

adjusted systematically for each item such that all [p]-initial items had a VOT of 10 ms (representative of a short-lag, Spanish-like VOT) and [p^h]-initial items had a VOT of 60 ms (representing a long-lag, English-like VOT). Trials were presented in random order, with all critical words, nonwords and fillers included in the same list.

2.3. Procedure

The experimental activities were completed online by participants in Mexico. Participants were recruited through Prolific [20], and the experiment was designed with and hosted by Gorilla [24]. While the quality of online data collection cannot be guaranteed, participants were encouraged to participate from a quiet and distraction-free location and were required to use headphones, which was confirmed through a headphone screening task [25]. After completing a brief demographic screening and passing the headphone check, participants completed an auditory lexical decision task, followed by the Bilingual Language Profile questionnaire [19].

Each trial of the auditory lexical decision task began with the presentation of a red fixation cross in the center of the computer screen for 1 s. When the fixation cross disappeared, an auditory stimulus was presented, and participants responded by keypress. Participants were instructed to decide, for each stimulus, whether they believed the item was a real Spanish word. They indicated their decision by pressing the 'j' key for 'yes' or the 'f' key for 'no' and were instructed to respond as quickly and accurately as possible. Each trial ended upon the key press, or after 3 s if no response was made, and the fixation cross would reappear, signalling the start of a new trial. There were 5 practice trials, followed by 240 trials presented in 2 equal blocks, with one 30 s break in the middle.

3. RESULTS

All participants heard each critical word twice during the lexical decision task, once with short-lag VOT and once with long-lag VOT. Because hearing a word a second time results in speeded lexical access due to priming [26], only responses to the first encounter with each lexical item are discussed in this paper. Additionally, only data from participants with accuracy greater than 80% on the filler items (both words and pseudowords) was included in the following analyses to ensure participant engagement.

3.1. Word Acceptance Rates

An analysis of the proportions of 'yes' responses to real words with differing phonetic detail revealed that, overall, word acceptance rates were low for critical words. English-accented, [p^h]-initial words were judged to be real Spanish words less than half the time, with a mean word acceptance rate of 43.1%, 95% CI [37%, 48%]. Participants rejected words with English-accented VOT more often than they accepted them. For short-lag VOT words, mean acceptance was higher, at 60.8% [55%, 66%], but still rather low—filler words were accepted at a rate of 90.3% [89%, 92%]. A paired *t*-test on logit-transformed word acceptance rates yielded a significant difference between the mean acceptance rate for word tokens presented with short-lag VOT and those presented with long-lag VOT: $M_{diff} = -0.861$, 95% CI [−1.22, −0.5], $t(44) = -4.80$, $p < 0.001$, Cohen's $d_{avg} = -0.85$, 95% CI [−1.25, −0.49], $r = 0.28$. Participants reliably rejected words with English-accented VOT more often than they rejected words with Spanish-accented VOT. Note, however, that word acceptance for all critical words, including Spanish-accented tokens, was surprisingly low relative to filler words.

3.2. RT as a function of VOT

Reaction time measured from stimulus onset was analyzed for 'yes' responses to see if VOT variation affected the speed of lexical access. Responses to short-lag VOT words were faster (1279 ms, 95% CI [1236, 1323]) than responses to words with long-lag VOT (1348 ms, 95% CI [1290, 1406]). Mean RT to filler words was 1216 ms [1175, 1257]. A paired *t*-test revealed that the mean log RT to words with short-lag VOT was significantly faster than mean RT to those with long-lag VOT: $M_{diff} = 0.054$, 95% CI [0.016, 0.091], $t(44) = 2.9$, $p = 0.005$, Cohen's $d_{avg} = 0.42$, 95% CI [0.14, 0.72], $r = 0.75$. Thus, when participants heard Spanish words with long-lag VOT, they took relatively longer to find a match in their lexicon. Note, however, that RTs to filler words were on average even faster ($M = 7.08$, 95% CI [7.05, 7.12]) than to short-lag VOT words.

3.3. RT to long VOT words by word judgment

When words with long-lag VOT were accepted, listeners experienced a relative delay in lexical retrieval. However, words presented with long-lag VOT were rejected nearly 60% of the time as nonwords. Because we are interested in the full picture, it is also important to examine the speed of rejection for these words. Did listeners take just as long to reject [p^h]-initial words as they did to accept them?

Mean RT in trials where long-lag VOT words were accepted as words was slower ($M = 1358$ ms, 95% CI [1288, 1428]) than that in trials where those words were rejected as nonwords ($M = 1258$ ms, 95%

CI [1200, 1316]). A paired *t*-test revealed that mean log RT to words with long VOT was significantly slower for trials in which participants responded 'yes' than for trials in which participants responded 'no': M_{diff} = −0.075, 95% CI [−0.126, −0.023], $t(44)$ = −2.92, p = 0.005, Cohen's d_{avg} = −0.49, 95% CI [−0.85, −0.16], r = 0.35. Thus, participants were faster when rejecting long-lag VOT words than when accepting them. Not only were participants more likely to reject [p^h]-initial Spanish words than accept them—recall the acceptance rate for long VOT words was only 43%—but they also took less time to reject these items than they did to accept them.

4. DISCUSSION AND CONCLUSION

This study examined to what extent English-like VOT in Spanish words affects lexical access for L1 Spanish L2 English speakers. It was hypothesized that these listeners would show sensitivity to the differences in phonetic detail between the English-accented and Spanish-accented stimuli, and that the processing of English-accented words would result in a delay when compared with the processing of Spanish-accented words. Nevertheless, since these participants are Spanish-English bilinguals, it was possible that cross-language phonetic interactions would modulate the speed of word processing even in their L1.

An analysis of word acceptance rates revealed that items with long-lag VOT were accepted as words only 43% of the time, indicating that the presence of long-lag VOT interfered with lexical access for these listeners—most English-accented words were not accepted as words at all. In these cases, the mismatch between the phonetic detail in English-accented Spanish words and listeners' lexical representations was great enough to prevent lexical retrieval. L1 Spanish L2 English bilinguals from monolingual Spanish backgrounds who have never lived amongst L1 English L2 Spanish speakers in a bilingual environment did not easily recognize English-accented Spanish words as real words, at least not in a context like the one in our research study. Additionally, when these English-accented words were recognized as words, they were accepted far slower than their Spanish-accented counterparts, strengthening evidence of interference with lexical access.

The speed of lexical access is modulated by the phonetic detail in the acoustic signal, and these listeners do not appear to have had sufficient experience with English-accented Spanish to have formed functioning representations of Spanish words beginning with [p^h] that would have allowed for more rapid lexical access of the base word. This suggests that phonetic detail is stored in lexical entries, as processing words with unfamiliar detail delayed lexical access in comparison to processing words with familiar detail. Additionally, an analysis of responses to long-lag VOT word tokens indicated that RT in trials where participants responded 'yes' was significantly slower than RT in trials where participants responded 'no', revealing that participants were quick to reject English-accented words and slow to accept them. Thus, there was no evidence that L2 phonetic detail was associated with the phonolexical representation of L1 words for these participants, suggesting that phonetic detail is mostly language-specific for these bilinguals.

Evidence from the lexical decision task indicates that phonetic detail is indeed stored in the lexicon, results that align with previous studies [1-6]. For these listeners, the phonetic detail in the English-accented Spanish words delayed lexical access or even prevented it completely. This is not surprising, given that these listeners are L1 Spanish speakers raised and living in Spanish-dominant environments. Previous research points to the role of experience with a given accent, and while these listeners may be implicitly familiar with some phonetic characteristics of English due to their experience as L2 learners, they do not show evidence of having had enough experience with English-accented Spanish to form "accented" lexical representations for L1 words.

This raises the question of whether Spanish-English bilinguals living in bilingual environments would behave differently. L1 English L2 Spanish late bilinguals living in the US are likely to be exposed to English-accented Spanish frequently, both in the speech of their classmates and in their own productions [27]. It is reasonable to hypothesize that they might show more acceptance of English-accented Spanish words if they have had the opportunity to form representations of this type of phonetic variants. There are also many Spanish-English early bilinguals living in the US who were raised bilingual and have had ample experience with the phonetic features of both Spanish and English, with frequent code-switching. Would this population have additional representations of English-accented Spanish words and accept them? Or would they show processing delays? The present study has outlined a background against which to compare the behaviour of these bilingual populations.

5. ACKNOWLEDGMENTS

This work was funded in part through a University of Arizona *Graduate and Professional Student Council Research and Project* (ReaP) *Grant* awarded to the first author.

6. REFERENCES

[1] Bürki, A., Ernestus, M., Frauenfelder, U. H. 2010. Is there only one 'fenêtre' in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *J. Mem. Lang*.

[2] Connine, C., and Pinnow, E. 2006. Phonological variation in spoken word recognition: Episodes and abstractions. *The Linguistic Review*. 23(3), 235–245.

[3] Connine, C., and Ranbom, L. 2004. Production frequency effects in perception of phonological variation. *J. Acoust. Soc. Am.*

[4] Pinnow, E., Connine, C., Ranbom, L. 2017. Processing pronunciation variants: the role of probabilistic knowledge about lexical form and segmental co-occurrence. *J. of Cognitive Psychology.*

[5] Ranbom, L., Connine, C. 2007. Lexical representation of phonological variation in spoken word recognition. *J. Mem. Lang.* 57(2), 273–298.

[6] Sumner, M., Samuel, A.G. 2009. The effect of experience on the perception and representation of dialect variants. *J. Mem. Lang.* 60(4), 487–501.

[7] Dahan, D., Magnuson, J.S., Tanenhaus, M.K. 2001. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language*.

[8] Ju, M., Luce, P.A. 2006. Representational specificity of within-category phonetic variation in the long-term mental lexicon. *J. Exp. Psychol. Hum. Percept. Perform.* 32(1), 120–138.

[9] McMurray, B., Tanenhaus, M.K., Aslin, R.N. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition.* 86(2), B33-42.

[10] Clarke, C.M., and Garrett, M.F. 2004. Rapid adaptation to foreign-accented English. *J. Acoust. Soc. Am.* 116(6), 3647–3658.

[11] Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., Floccia, C. 2012. Linguistic processing of accented speech across the lifespan. *Front. Psychol.* 3(479).

[12] Floccia, C., Goslin, J., Girard, F., Konopczynski, G. 2006. Does a regional accent perturb speech processing? *J. Exp. Psychol. Hum. Percept. Perform.* 32(5), 1276–1293.

[13] Bradlow, A., Bent, T. 2008. Perceptual adaptation to non-native speech. *Cognition*. 106(2), 707–729.

[14] Maye, J., Aslin, R.N., Tanenhaus, M.K., The weckud wetch of the wast: lexical adaptation to a novel accent. *Cogn. Sci.* 32(3), 543–562.

[15] Larraza, S., Samuel, A.G., Oñederra, M.L. 2016. Listening to accented speech in a second language: First language and age of acquisition effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 42(11), 1774–1797.

[16] Larraza, S., Best, C.T. 2018. Differences in phonetic-to-lexical perceptual mapping of L1 and L2 regional accents. *Bilingualism: Language and Cognition*. 21(4), 805–825.

[17] Weber, A., Broersma, M., Aoyagi, M. 2011. Spoken-word recognition in foreign-accented speech by L2 listeners. *J. Phon.* 39(4), 479–491.

[18] Shea, C. 2017. L1 English / L2 Spanish: Orthography–phonology activation without contrasts. *Second language research*. 33(2), 207–232.

[19] Birdsong, D., Gertken, L.M., Amengual, M. 2012. Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism. *COERLL, University of Texas at Austin*.

[20] Prolific. www.prolific.co.

[21] Guasch, M., Boada, R., Ferré, P., Sánchez-Casas, R. 2013. NIM: A Web-based Swiss Army knife to select stimuli for psycholinguistic studies. *Behav. Res. Methods*. 45, 765–771.

[22] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer.

[23] Winn, M.B. 2020. Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *J. Acoust. Soc. Am.*, 147(2), 852-865.

[24] Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N. Evershed, J.K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods*. 52(1), 388–407.

[25] Milne, A.E. *et al.* 2021. An online headphone screening test based on dichotic pitch. *Behav. Res. Methods*. 53(4), 1551–1562.

[26] Forster, K., Davis, C. 1984. Repetition priming and frequency attenuation in lexical access. *J. Exp. Psychol. Learn. Mem. Cogn.* 10(4), 680–698.

[27] Tyler, M.D. 2019. PAM-L2 and phonological category acquisition in the foreign language classroom. In: Nyvad, A.M., Hejná, M., Højen, A., Jespersen, A.B., Sørensen, M.H. 2019. *A Sound Approach to Language Matters- In honor of Ocke-Schwen Bohn*. Aarhus: Aarhus University, 607–630.

L2 PROFICIENCY PREDICTS L1 ACCENTEDNESS AND COMPREHENSIBILITY

Kakeru Yazawa[1], Rubén Pérez-Ramón[2], Mariko Kondo[2]

[1]University of Tsukuba, [2]Waseda University
yazawa.kakeru.gb@u.tsukuba.ac.jp, rperez@aoni.waseda.jp, mkondo@waseda.jp

ABSTRACT

Learning a second language (L2) can change the pronunciation of one's first language (L1), but its implications for communicative dimensions such as accentedness and comprehensibility remain unclear. In this study, L1 recordings of 183 Japanese learners of English living in Japan were rated for accentedness and comprehensibility by a cohort of Japanese listeners. The resulting scores were compared with those of a previous similar assessment, in which the L2 recordings of the same speakers were rated for nativelikeness by English listeners. Statistical analyses revealed that those who were judged to be more nativelike in the L2 tended to be perceived as more foreign-accented in the L1, but surprisingly, their L1 speech was also perceived as more comprehensible. Theoretical implications and future directions are discussed.

Keywords: foreign accent, comprehensibility, proficiency, phonetic drift, nativelikeness

1. INTRODUCTION

The acquisition of a second language (L2) and its deployment in actual communicative scenarios have a deep and meaningful impact on the way phonological processing is shaped in the mind of the speaker. Traditional studies have shown a clear shift in a 'forward' way, that is, a transfer of traits from the first language (L1) to the spoken realization of the L2 [1, 2]. This imposition of one's own native language on the target language gives rise to what is known as a foreign accent [3].

There is, however, a possibility that the opposite also occurs: as the particularities of the L2 are instilled in the mind of the speaker, the pronunciation of the L1 may change as these new traits transfer 'backward' [4]. This phenomenon, often referred to as L1 phonetic drift, has been examined by an increasing number of studies lately (see [5] for a review). According to these studies, L2 learners can show evidence of phonetic drift in the form of L1 sound categories either assimilating to [6, 7] or dissimilating from [8, 9] the L2 ones, both of which are predicted by the Speech Learning Model (SLM) [10, 11]. The likelihood and magnitude of phonetic drift appear to be influenced by how proficient the learner is in the L2 and/or how dominantly the learner uses it, but since these two factors often go along in real-world learning scenarios, their own effects remain to be elucidated.

In this study, we will tackle the implications of the above phenomenon for communication. While studies on L1 phonetic drift abound, the majority of them have examined the acoustic-phonetic aspects of it, including the VOT of stops, spectral characteristics of vowels, approximants, and fricatives, f_o level, and f_o alignment [12]. Very few studies have focused on the consequence of the observed drift, namely how the drifted speech is perceived by listeners. One such study [13] examined L1 German speakers who had lived in Canada or the Netherlands for an average of 37 years, finding that their native German speech had a foreign accent perceptible to monolingual German listeners, with some speakers even being perceived as non-native speakers of German. A similar and perhaps more iconic case is that of Arnold Schwarzenegger, who was born in Austria in 1947, moved to the US in 1968 and has lived there since then. According to a longitudinal study of his speech [14], his recent L1 German (2010 - 2017) has a perceptible foreign accent that was absent in the early stage of his career (1977 - 1989).

The main research question we will address in this study is how L2 proficiency (defined as the ability to speak in a native-like fashion) relates to L1 accentedness. Specifically, we will examine a population of Japanese learners of English who are based in Japan and use Japanese as the primary means of communication but show diverse levels of L2 nativelikeness. This would help us disentangle the effects of L2 proficiency and L2 use, in contrast to the two previous studies mentioned above [13, 14] which examined a migrant population who were not only proficient but also dominant in the L2. We will also examine L1 comprehensibility, which, to the best of our knowledge, is a novel attempt in the field.

2. METHODS

2.1. Speakers

The speakers to be assessed are 183 L1 Japanese learners of L2 English (115 female, 68 male) from the J-AESOP corpus [15, 16, 17]. All of them were undergraduate or graduate students at universities in Tokyo and surrounding areas (mean age = 20.3). Most of the students started learning English at the age of 13 as part of their compulsory education in Japanese schools, while others were first exposed to English at an earlier age in Japan or overseas. Approximately one-third of the speakers ($n = 63$) had an experience of living in an English-speaking country (Australia, Canada, New Zealand, the US, and the UK) for various periods of time: 13 speakers for 1 to 6 months, 25 for 9 to 18 months, 15 for 2 to 5 years, and 10 for 6 to 11 years.

The corpus also includes 20 L1 English learners of L2 Japanese (15 female and 5 male) who were visiting, exchange, or degree-seeking students at one of the universities mentioned above (mean age = 20.7). Their speech will also be assessed as a distractor and reference.

2.2. Materials

There are eight types of recording tasks in J-AESOP (see [15] for details). This study uses the data from Task 6, in which the speakers read aloud the Japanese and English versions of "The North Wind and the Sun" [18].

2.3. Assessment of English speech

As part of the development of the corpus, all speakers' perceived nativelikeness in English had previously been assessed by four phonetically trained judges whose L1 is American English (see [19] for detailed procedures). The judges listened to each speaker's recorded sample of the English version of "The North Wind and the Sun" and evaluated its nativelikeness on a scale of 1 ("strongly foreign-accented") to 10 ("free of foreign accent"). Inter-judge consistency was very high, with a Cronbach's alpha of 0.98. The resultant scores of L1 Japanese speakers, when averaged across judges, ranged from 1.33 to 10.00 (mean = 5.44, median = 5.00, standard deviation = 1.98), suggesting that the speakers' impression ranged from heavily Japanese-accented to practically native English-like. The average scores of L1 English speakers ranged from 8.92 to 10.00 (mean = 9.83, median = 10.00, standard deviation = 0.29).

2.4. Assessment of Japanese speech

2.4.1. Listeners

Ten female L1 Japanese listeners, all of whom were undergraduate or graduate students at the University of Tsukuba (mean age = 20.8), participated in the assessment of the Japanese speech for this study. Most of them started learning English at the age of 13, while a few started earlier. Six had not lived outside of Japan for more than one month, while four had studied overseas for less than a year: one in the UK for 2 months, one in Australia for 2 months, one in the US for 7 months, and one in Germany for 10 months. All reported normal hearing.

2.4.2. Stimuli

The stimuli were the recorded samples of the Japanese version of "The North Wind and the Sun" as read by the 183 Japanese and 20 English speakers. Each audio sample's intensity was scaled to have a peak of 60 dB and noise-reduced on Praat [20].

2.4.3. Procedure

Gorilla Experiment Builder [21] was used to create an online experimental platform that the listeners could access from their own devices. They were encouraged to wear earphones or headphones throughout the experiment. After a brief tutorial, they engaged in a series of trials for assessment.

In each trial, the listeners were first presented with one of the 203 stimuli in random order. They then assessed the accentedness and comprehensibility of the speech, using one response slider per criterion.[1] The slider for accentedness had a "strong foreign accent" label on one end and a "no foreign accent" label on the other, while the slider for comprehensibility had "difficult to understand" on one end and "easy to understand" on the other. The internal value of each slider ranged from 0 to 100, where a larger value would indicate a higher level of perceived accentedness or comprehensibility, respectively.

A pause screen was displayed between trials so that the listeners could take a short break at any time during the experiment. This was because the experiment was estimated to take three to five hours to complete. The listeners were also instructed to take a long break after no more than one hour of assessment. All listeners completed their tasks within one week. They received monetary compensation for their time and effort.

3. RESULTS

3.1. Accentedness

Figure 1 shows the overall relationship between the Japanese accentedness score and the English nativelikeness score across all speakers. The accentedness scores were highly consistent across listeners, with a Cronbach's alpha of 0.94.

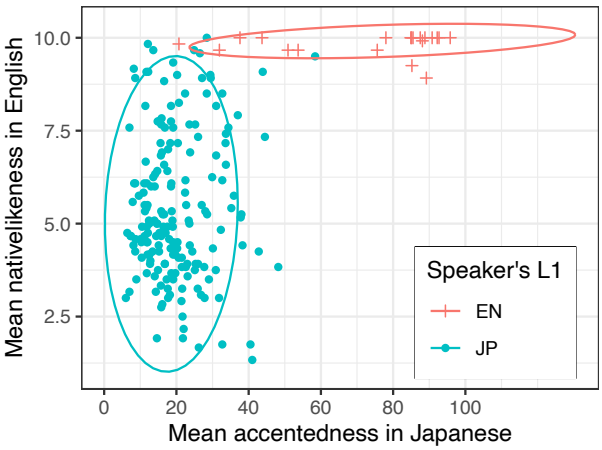


Figure 1: Mean accentedness in Japanese and nativelikeness in English per speaker.

The main question here is whether L1 Japanese speakers who are more nativelike (i.e., proficient) in L2 English tend to be perceived as more foreign-accented in the L1. Even though such a tendency is not evident in the figure, a linear mixed effects (LME) model with the following structure would reveal the effect of L2 proficiency:

lmer(Acc ~ Ntv + (1 | judge) + (1 | listener), data = L1.Japanese.speakers)

where *Acc* is the accentedness score in Japanese and *Ntv* is the nativelikeness score in English. Random intercepts were fitted for *judge* (of nativelikeness scores) and *listener* (of accentedness scores). The result of the model according to the *lmerTest* function [22] of R [23] is shown in Table 1.

	Estimate	*s.e.*	*t*	*p*
Intercept	18.505	6.493	2.850	.019
Ntv	0.273	0.083	3.306	$< .001$

Table 1: Results of the LME model on accentedness.

Ntv is a statistically significant predictor of *Acc*, and its positive estimated coefficient suggests that those who are more proficient in the L2 tend to be perceived as more foreign-accented in their L1.

3.2. Comprehensibility

Figure 2 shows the overall relationship between the Japanese comprehensibility score and the English nativelikeness score across all speakers. The listeners' assessment was again highly consistent, with a Cronbach's alpha of 0.92.

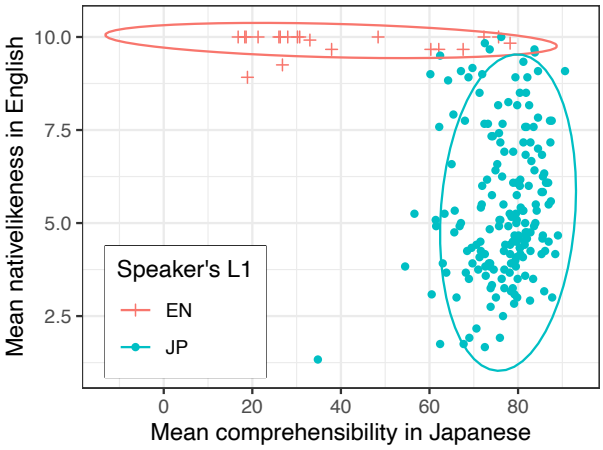


Figure 2: Mean comprehensibility in Japanese and nativelikeness in English per speaker.

The question here is whether L1 Japanese speakers who are more nativelike (i.e., proficient) in L2 English tend to be perceived as less comprehensible in the L1, which would be a plausible hypothesis given the result in Section 3.1. In order to test this, another LME model was fitted with the following structure:

lmer(Cmp ~ Ntv + (1 | judge) + (1 | listener), data = L1.Japanese.speakers)

where *Cmp* is the comprehensibility score in Japanese. The result of the model is shown in Table 2.

	Estimate	*s.e.*	*t*	*p*
Intercept	74.142	5.334	13.898	$< .001$
Ntv	0.477	0.074	6.413	$< .001$

Table 2: Results of the LME model on comprehensibility.

Ntv is again a statistically significant predictor of *Cmp*, but its estimated coefficient is unexpectedly positive. This indicates that those who are more proficient in the L2 are *easier* to understand in their L1. This would align with the ellipse for Japanese speakers in Figure 2 that is slightly tilted rightwards.

4. DISCUSSION

This study explored the potential effects of L2 proficiency on perceived L1 accentedness and comprehensibility using a corpus of Japanese learners of English. While previous studies have found perceptible foreign accents in the L1 speech of highly proficient L2 learners [13, 14], it was unclear whether the accent was due to their high proficiency per se or to their dominant use of the L2. It is possible that these migrant learners had experienced a significant reduction in L1 use and a consequent decline in linguistic abilities in the L1, commonly referred to as L1 attrition [24]. The current study, therefore, eliminated the possibility of L1 attrition caused by insufficient L1 use by focusing on a population who live where the L1 is spoken and use the L1 as their primary language but differ in L2 proficiency levels. The analysis revealed that those who were judged to be more nativelike or proficient in L2 English tended to be perceived as more foreign-accented in L1 Japanese, confirming a sole effect of L2 proficiency on L1 accentedness, perhaps for the first time.

Another novel finding of this study was the relationship between L2 proficiency and L1 comprehensibility. If L2 proficiency leads to L1 accentedness, one might expect that it would also lead to less comprehensibility of the L1 speech, given the close connection between accentedness and comprehensibility [25]. The result was in fact the opposite, as higher L2 proficiency predicted higher L1 comprehensibility. This surprising result may be explained by the "category precision" hypothesis of the revised SLM [11]. According to this hypothesis, individuals having relatively precise L1 phonetic categories will be better able to discern phonetic differences between an L2 sound and the closest L1 sound. This will increase their likelihood of forming new phonetic categories for L2 sounds and thus becoming nativelike in the L2. In other words, individuals whose L1 speech is clear and precise (i.e., comprehensible) are likely to attain a high level of L2 proficiency [26]. There is, in theory, another possibility that higher L2 proficiency makes L1 speech more comprehensible, but no current models of L2 speech acquisition provide a plausible explanation for why this would be the case.

One caveat with the obtained results, however, is that the overall magnitude of the effects of L2 proficiency on nativelikeness or comprehensibility was not that large. Specifically, the estimate of the effect of *Ntv* (which itself ranged from 1 to 10) on *Acc* was 0.273, meaning that the predicted difference between the least and most accented speakers is less than 3 on a scale of 0 to 100. Likewise, the estimated effect of *Ntv* on *Cmp* was 0.477, meaning that the predicted difference between the least and most accented speakers is less than 5. Thus, while both effects were statistically significant, it must be kept in mind that the differences were only subtle.

The results of the current study are admittedly preliminary, and more can be done to extend the findings. For example, since the L1 Japanese speakers show some degree of variation in their age of learning (AOL) and length of residence (LOR) in an English-speaking country, these factors can be additionally examined. Of particular importance is AOL, as early, late, and very late bilinguals may show different patterns of L1 phonetic drift [5]. Testing more L1 Japanese listeners would help to consolidate the obtained results, while at the same time, a stricter control of the listeners' language learning background would be desirable because it can affect the evaluation of L1 accented speech [27].

Another important avenue for future research is to identify what kind of segmental characteristics contribute to the overall perceived accentedness and comprehensibility of the L1 Japanese speech. As mentioned earlier, previous studies on L1 phonetic drift have tended to focus on the acoustic-phonetic aspect of it, and its relevance to communicative dimensions is still unclear. Previous studies on segmentally manipulated L2 speech have also found non-uniform effects of different vowels and consonants on perceived accentedness and intelligibility [28, 29], which would most likely apply to L1 accentedness and comprehensibility as well. Further investigation of the corpus data in relation to the obtained accentedness and comprehensibility scores may help shed further light on this matter.

5. ACKNOWLEDGMENTS

This study was supported by JSPS Grant-in-Aid for Early-Career Scientists (20K13144), for Exploratory Research (19K21638), and for Scientific Research (B) (21H00533).

6. REFERENCES

[1] J. Leather and A. James, "The acquisition of second language speech," *Studies in Second Language Acquisition*, vol. 13, no. 3, pp. 305–341, 1991.

[2] L. Colantoni, J. Steele, and P. Escudero, *Second Language Speech: Theory and Practice*. Cambridge: Cambridge University Press, 2015.

[3] T. Piske, I. R. A. MacKay, and J. E. Flege, "Factors affecting degree of foreign accent in an L2: A review," *Journal of Phonetics*, vol. 29, no. 2, pp. 191–215, 2001.
[4] V. Cook, Ed., *Effects of the Second Language on the First*. Bristol, Blue Ridge Summit: Multilingual Matters, 2003.
[5] N. Kartushina, U. H. Frauenfelder, and N. Golestani, "How and when does the second language influence the production of native speech sounds: A literature review," *Language Learning*, vol. 66, no. S2, pp. 155–186, 2016.
[6] C. B. Chang, "Rapid and multifaceted effects of second-language learning on first-language speech production," *Journal of Phonetics*, vol. 40, no. 2, pp. 249–268, 2012.
[7] ——, "A novelty effect in phonetic drift of the native language," *Journal of Phonetics*, vol. 41, no. 6, pp. 520–533, 2013.
[8] J. E. Flege and W. Eefting, "Production and perception of English stops by native Spanish speakers," *Journal of Phonetics*, vol. 15, pp. 67–83, 1987.
[9] ——, "Cross-language switching in stop consonant perception and production by Dutch speakers of English," *Speech Communication*, vol. 6, pp. 185–202, 1987.
[10] J. E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 233–277.
[11] J. E. Flege and O.-S. Bohn, "The revised Speech Learning Model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*, R. Wayland, Ed. Cambridge: Cambridge University Press, 2021, pp. 3–83.
[12] C. B. Chang, "Phonetic drift," in *The Oxford Handbook of Language Attrition*, M. S. Schmid and B. Köpke, Eds. Oxford: Oxford University Press, 2019, pp. 91–203.
[13] E. D. Leeuw, M. S. Schmid, and I. Mennen, "The effects of contact on native language pronunciation in an L2 migrant setting," *Bilingualism: Language and Cognition*, vol. 13, no. 1, pp. 33–40, 2009.
[14] L. Kornder, *Arnold Schwarzenegger Now and Then: A Longitudinal Case Study of Schwarzenegger's English and German Pronunciation*, ser. Gewi Ausgezeichnet: Abschlussarbeiten. Graz: Graz University Library Publishing, 2022, no. 4.
[15] M. Kondo, H. Tsubaki, and Y. Sagisaka, "Segmental variation of Japanese speakers' English: Analysis of "the North Wind and the Sun" in AESOP corpus," *Journal of the Phonetic Society of Japan*, vol. 19, no. 1, pp. 3–17, 2015.
[16] H. Meng, C.-Y. Tseng, M. Kondo, A. Harrison, and T. Visceglia, "Studying L2 suprasegmental features in Asian Englishes: A position paper," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association*. Brighton, UK: International Speech Communication Association, 2009, pp. 1715–1718.
[17] T. Visceglia, C.-Y. Tseng, M. Kondo, and Y. Sagisaka, "Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)," in *Proceedings of 2009 Oriental COCOSDA International Conference on Speech Database and Assessments*. Beijing: Institute of Electrical and Electronics Engineers, 2009, pp. 52–57.
[18] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to Use the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.
[19] T. Konishi, "A corpus-based study on Japanese English rhythm," Ph.D. dissertation, Waseda University, 2022.
[20] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org/
[21] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020.
[22] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[23] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2022.
[24] A. Stoehr, T. Benders, J. G. van Hell, and P. Fikkert, "Second language attainment and first language attrition: The case of VOT in immersed Dutch-German late bilinguals," *Second Language Research*, vol. 33, no. 4, pp. 483–518, 2017.
[25] T. M. Derwing and M. J. Munro, "Second language accent and pronunciation teaching: A research-based approach," *TESOL Quarterly*, vol. 39, no. 3, pp. 379–397, 2005.
[26] N. Kartushina and U. H. Frauenfelder, "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," *Frontiers in Psychology*, vol. 5, p. 1246, 2014.
[27] L. Kornder and I. Mennen, "Listeners' linguistic experience affects the degree of perceived nativeness of first language pronunciation," *Frontiers in Psychology*, vol. 12, 2021.
[28] R. Pérez-Ramón, M. Cooke, and M. L. G. Lecumberri, "Is segmental foreign accent perceived categorically?" *Speech Communication*, vol. 117, pp. 28–37, 2020.
[29] ——, "Foreign accent strength and intelligibility at the segmental level," *Speech Communication*, vol. 137, pp. 70–76, 2022.

[1] The speakers' linguistic background was not disclosed to the listeners until the end of the experiment.

DEMYSTIFYING HERITAGE ACCENT: HERITAGE ACCENT PERCEIVED BY HERITAGE SPEAKERS

Joo Kyeong Kim, Ji Young Kim, John Carter, & Jenny Eonsuh Choi

University of California, Los Angeles
anitajookyeong@g.ucla.edu, jiyoungkim@g.ucla.edu, jackcarter1868@g.ucla.edu, choijen@g.ucla.edu

ABSTRACT

Heritage speakers (HSs) generally sound more native-like than L2 learners (L2s), despite having an accent that is different from non-heritage native speakers (NHNS) or a "heritage accent." However, there is little understanding on the relative impact of segments and prosody on heritage accent. Moreover, how the heritage accent is perceived by their peer HSs is not well understood. The present study examines the relative contribution of segments and prosody in the perception of heritage accent by two groups of raters, HSs and NHNSs of Mexican Spanish. Semi-spontaneous speech samples were collected from three speaker groups (i.e., NHNS, HS, and L2), which were used as auditory stimuli for an accent rating task. HSs and NHNSs evaluated the nativeness of the speech samples presented in three different conditions (i.e., original, segment-only, and prosody-only). The results suggest that for both rater groups, prosody plays a bigger role in heritage accent perception, while L2 foreign accent is mainly perceived in the segments.

Keywords: heritage accent, accent rating, resynthesized speech, heritage speaker raters

1. INTRODUCTION

Heritage languages are minority languages acquired naturalistically in a bilingual or multilingual environment, such as diasporic languages spoken by immigrants and their children [14, 19]. The speakers of heritage languages, or heritage speakers, are generally considered to sound more native-like than second language learners, despite having an accent that diverges from that of non-heritage native speakers [1, 9, 12, 15]. However, the phonetic features that are responsible for the so-called "heritage accent" [16] are not well understood. While some studies have shown that deviation in prosody from the monolingual norm is associated with heritage accent [8, 20, 22], it is unclear whether prosody contributes more to heritage accent than segments, given that the two are inevitably intertwined. Moreover, research on heritage accent has largely been based on non-heritage native speakers' impression [1, 9, 12, 15], while there is little understanding on how heritage speakers evaluate other heritage speakers' speech.

In order to fill these gaps in the literature, we conducted an online accent rating task using resynthesized speech with both heritage and non-heritage native listeners as raters. Specifically, we seek to elucidate (1) which of the two aspects of speech, segments or prosody, contribute more to perceived heritage accent and (2) whether heritage speakers consider other heritage speakers' speech native-sounding.

2. METHODS

2.1. Participants

2.1.1. Speakers

Semi-spontaneous speech samples were recorded from 18 adult speakers of Spanish in their twenties. The speakers consisted of six heritage speakers (HSs), six second language learners (L2s), and six non-heritage native speakers (NHNSs). The L2s and the NHNSs were included as control groups that represent two extreme points of the native-non-native accent continuum [9]. All three groups were balanced by gender. The HSs were second-generation Mexican Americans, whose parents were from Mexico and moved to the US as adults. All the HSs were born and raised in Southern California. Three L2s were native speakers of American English and the other three L2s were native speakers of Korean. Lastly, the NHNSs were born and raised in Mexico.

2.1.2. Raters

Fifty-four raters participated in the accent rating task. Half of them were NHNSs (mean age: 23.0 years; 8 female, 19 male), who were all born and raised in Mexico. The other half were HSs residing in Southwestern US (24 California, 2 Texas, and 1 Nevada) and at least one of their parents was from a Spanish-speaking country. For 21 out of 27 HS raters, either one or both parents were from Mexico. For the remaining 6 raters, the parents were from other Spanish-speaking countries: Argentina (1), Colombia (1), Ecuador (1), Guatemala (1), and El Salvador (2). Since dialect may have an effect on accent ratings, in

this study, we only report data of the 21 HS raters with Mexican Spanish input (mean age: 25.6 years; 12 female, 9 male), consistent with other participants.

All the NHNS raters and 13 HS raters were recruited via Prolific (https://www.prolific.co) which is an online research participant recruitment platform. The other 8 HSs were recruited at the University of California, Los Angeles (UCLA).

2.2. Stimuli

To elicit semi-spontaneous speech, we asked the speakers to describe and narrate a wordless picture book, "Frog, Where Are You?" [13]. Two ten-second samples were extracted from each speaker, resulting in a total of 36 speech samples. All the samples were free of morphosyntactic errors. Pauses longer than 0.3 seconds were manually trimmed and speech rate was controlled across the groups using Praat [3]. Then, the samples were resynthesized to test two additional conditions (i.e., segments-only and prosody-only) other than the original condition.

For the segments-only condition, we first controlled for the rhythm so that all the samples represented the rhythm of a non-heritage native speaker. For this purpose, we had a non-heritage native speaker of Mexican Spanish, who did not participate in this study as a speaker, read aloud the transcripts of the 36 speech samples as naturally as possible. The duration of each segment in the native speaker samples was measured and then superimposed onto the test stimuli using Praat [3]. Then, we monotonized the samples using the 'Monotonize' function in Praat Vocal Toolkit [4]. This command monotonizes the pitch contour to the pitch median of the selected audio file.

For the prosody-only condition, we used the Hann Band filter in Praat [3] to remove formant values below a certain cut-off point. The frequency range was set from 0 to 320 Hz for male speakers and from 0 to 400 Hz for female speakers, and smoothing was set to 50 Hz [7, 20, 21]. These low-pass filtered samples solely retained prosodic properties. After the resynthesis, we were left with a total of 108 stimuli (= 36 samples * 3 conditions).

2.3. Procedures

The accent rating task was administered on SurveyMonkey (https://www.surveymonkey.com). The experiment was conducted in Spanish for the NHNS raters and in English for the HS raters. The experiment consisted of a background questionnaire and three blocks of accent rating tasks (i.e., one block for each stimulus type). The background questionnaire was used to ensure that participants fit the criteria for each group. For the NHNS raters, we only included individuals born and raised in Mexico. For the HSs, we only included individuals born and raised in the US or those that were born in Mexico and moved to the US before receiving formal education in Spanish (i.e., at or before the age of 6).

The participants were presented first with the original stimuli, followed by the segments-only stimuli and the prosody-only stimuli. For each item, the raters were asked to evaluate the nativeness of the speaker on a scale of 1 to 6 (= completely native). There was a training period with two speech samples prior to all three blocks. These two samples were not part of the experimental stimuli. One sample was from a NHNS and the other was from an L2.

For the segments-only stimuli, the raters were instructed that they would be hearing speech sounding like unsophisticated robots or androids. For the prosody-only stimuli, they were informed that the speech would sound as if someone was talking in an adjacent room, but with a closed door. They were also told that they would not be able to recognize the words. Instead, a transcript corresponding to the speech sample was presented on the screen and the raters were instructed to read it aloud before listening to the audio file. The entire study took around 45 minutes to complete.

3. RESULTS

A total of 5,112 responses from 48 raters were collected in the accent rating task. As for one HS rater, only the ratings of the original stimuli were included in the analysis for giving consistent ratings (1 = completely non-native) for the other two conditions with resynthesized stimuli. Table 1 demonstrates the descriptive statistics of accent ratings by stimulus type, by speaker group, and by rater group.

Stimulus Type	Mean	Median	SD
Original	4.19	5	1.77
Segments-only	3.91	4	1.63
Prosody-only	4.08	4	1.49
Speaker Group	**Mean**	**Median**	**SD**
HS	4.13	4	1.40
L2	2.95	3	1.62
NHNS	5.11	5	1.07
Rater Group	**Mean**	**Median**	**SD**
HS	4.18	5	1.56
NHNS	3.97	4	1.69

Table 1: Descriptive statistics of accent ratings (1 = completely non-native, 6 = completely native). HS: heritage speakers, L2: L2 learners, NHNS: non-heritage native speakers.

We performed a mixed effects linear regression analysis using the *lmer*() function in the *lme4* package [2] in R [17] to examine whether stimulus type, speaker group, and rater group have an effect on accent ratings. As fixed effects, we entered stimulus type (reference level = original), speaker group (reference level = HS), and rater group (reference level = HS), which were contrast-coded using simple coding. Participant was included as a random effect. The best fitting model selected through backward elimination included an intercept for participant with by-participant random slope for speaker group. Statistical significance of the fixed effects was analysed through likelihood ratio tests, which were performed using the *anova*() function in the *car* package [5].

Adding the rater group did not improve the model fit (p = 0.08). Thus, in the full model, we only included stimulus type, speaker group, and the interaction between the two fixed effects. We found a significant main effect of stimulus type ($\chi^2(2)$ = 41.828, $p < 0.001$) and a significant main effect of speaker group ($\chi^2(2)$ = 108.9, $p < 0.001$). Moreover, there was a significant interaction between the two fixed effects ($\chi^2(4)$ = 333.47, $p < 0.001$).

We obtained the summary of the results of the full model using the *summary*() function in the base package of R [17] and performed post-hoc pairwise multiple comparisons between levels using the *emmeans*() function in the *emmeans* package [11]. Results of the effect of stimulus type showed that the original stimuli were rated significantly higher than the segments-only stimuli (β = 0.287, $p < 0.001$) and the prosody-only stimuli (β = 0.118, $p < 0.05$). Post-hoc pairwise comparisons revealed that the prosody-only stimuli were also rated significantly higher than the segments-only stimuli (β = –0.168, $p < 0.001$). That is, removing the segments from the original stimuli did not affect the overall accent ratings as much as removing the prosody did.

As for the effect of speaker group, results demonstrated that HSs' speech was rated significantly higher than that of the L2s (β = 1.172, $p < 0.001$) and significantly lower than that of the NHNSs (β = −0.975, $p < 0.001$). Post-hoc pairwise comparisons confirmed that L2s' speech was rated significantly lower than that of the NHNSs (β = −2.147, $p < 0.001$). This suggests that, to both the HS and the NHNS raters, the HSs sounded more native-like than the L2s and less native-like than the NHNSs.

Now we turn to the interaction between stimulus type and speaker group. Figure 1 demonstrates the effect plot of the interaction between stimulus type and speaker group, which was created using the *allEffects*() function in the *psych* package [18].

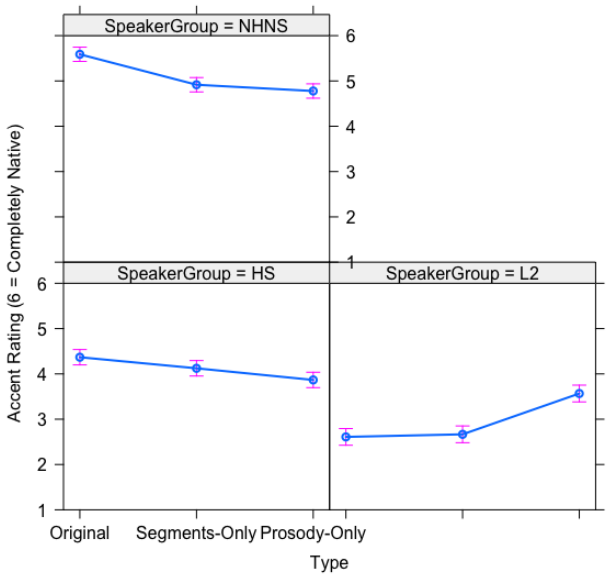


Figure 1: Effect plot of the interaction between stimulus type and speaker group.

Post-hoc pairwise comparisons showed that the above-mentioned pattern of HSs' speech being rated significantly higher than L2s' speech and significantly lower than NHNSs' speech was consistent with the original stimuli and with the segments-only stimuli (ps < 0.001) but not with the prosody-only stimuli. NHNSs' prosody-only stimuli received significantly higher ratings than those of the HSs and the L2s, while the difference between HSs' and L2s' ratings in this condition did not reach statistical significance (p = 0.089) (NHNS > HS = L2).

With regard to the role of segments and prosody in accent ratings, we found different patterns across speaker groups. HSs' segments-only stimuli were rated significantly lower than the original stimuli (p < 0.05), and the ratings of prosody-only stimuli were even lower than those of the segments-only stimuli (p < 0.05). Regarding L2s' speech, the accent ratings of the segments-only stimuli did not significantly differ from those of the original stimuli (p = 1), but both stimulus types were rated significantly lower than the prosody-only stimuli (ps < 0.001). Lastly, the accent ratings of NHNSs' speech demonstrated that the segments-only stimuli and the prosody-only stimuli were rated lower than the original stimuli (ps < 0.001). The accent ratings of the two resynthesized stimulus types did not significantly differ from each other (p = 0.645). These findings are summarized in Table 2.

Speaker Group	Accent Rating
HS	O > S > P
L2	O = S < P
NHNS	O > S = P

Table 2: Summary of the findings (higher/lower accent rating = higher/lower perceived nativeness). O = original, S = segments-only, P = prosody-only.

4. DISCUSSION

Our data demonstrated that HSs sound more native-like than L2s, but less native-like than NHNSs. This finding is consistent with previous research on perceived nativeness of heritage speakers with NHNS as raters [1, 9, 12, 15]. It is noteworthy that these patterns were also found in HS raters. HS raters' average accent rating of HSs' unaltered speech was 4.21 (SD = 1.31) and their average accent rating of NHNSs' unaltered speech was 5.11 (SD = 0.98). That is, while HSs may consider their linguistic variety (i.e., heritage Spanish) to be native, they evaluate it as sounding less native-like than non-heritage Mexican Spanish, which is likely to be associated with the standard language ideology that excludes and devalues heritage language varieties [6, 10]. Future research should examine whether HSs indeed recognize other speakers of heritage Spanish and, if so, tap into their explicit and implicit attitudes toward this linguistic variety.

In this study, we also found that the NHNS > HS > L2 pattern was observed in the original and the segments-only stimuli but not in the prosody-only stimuli where HSs' ratings did not significantly differ from L2s' ratings (NHNS > HS = L2). Moreover, among the three stimulus types, the HSs received the lowest accent ratings in the prosody-only stimuli, whereas the L2s received the highest ratings in this condition. In other words, divergence from non-heritage native accent in HSs' and L2s' speech is likely to derive from prosody and segments, respectively. These findings suggest that heritage accent is not only quantitatively different from L2 foreign accent (i.e., heritage accent sounds more native-like than L2 foreign accent), but also qualitatively different from it; heritage accent is a type of native accent that is different from non-heritage native accent mainly in the prosody, while L2 foreign accent is a non-native accent that diverges from heritage/non-heritage native accent mainly in the segments.

Future research should examine which prosodic feature in HSs' speech and which segmental feature in L2s' speech play the strongest role in perceived heritage and L2 foreign accents.

5. REFERENCES

[1] Au, T. K., Oh, J. S., Knightly, L. M., Jun, S.-A., Romo, L. F. 2008. Salvaging a childhood language. *Journal of Memory and Language* 58(4), 998–1011.

[2] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[3] Boersma, P., Weenink, D. 2021. *Praat: Doing phonetics by computer* (6.1.38) [Computer software]. http://www.praat.org/

[4] Corretge, R. 2020. *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com

[5] Fox, J., Weisberg, S. 2019. *An {R} Companion to Applied Regression*, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[6] Gasca Jiménez, L., Adrada-Rafael, S. 2021. Understanding heritage language learners' critical language awareness (CLA) in mixed language programs. *Languages* 6, 37.

[7] Jilka, M. 2000. *The contribution of intonation to the perception of foreign accent* [Unpublished doctoral dissertation]. Stockholm University.

[8] Kan, R. T. Y. (2020). Suprasegmental and prosodic features contributing to perceived accent in heritage Cantonese. *Proceedings of the 10th International Conference on Speech Prosody*, 101–105.

[9] Kupisch, T., Barton, D., Klaschik, E., Lein, T., Stangen, I., van de Weijer, J. 2014. Foreign accent in adult simultaneous bilinguals. *Heritage Language Journal* 11(2), 123–150.

[10] Leeman, J. 2012. Investigating language ideologies in Spanish as a heritage language. In: Beaudrie, S., Fairclough, M. (eds), *Spanish as a Heritage Language in the United States: The State of the Field*. Georgetown University Press, 43–59.

[11] Lenth, R. V. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.2. https://CRAN.R-project.org/package=emmeans

[12] Lloyd-Smith, L., Einfeldt, M., Kupisch, T. 2020. Italian-German bilinguals: The effects of heritage language use on accent in early-acquired languages. *International Journal of Bilingualism* 24(2), 289–304.

[13] Mayer, M. 1969. *Frog, where are you?* Dial Press.

[14] Montrul, S., Polinsky, M. 2021. Introduction: Heritage languages, heritage speakers, heritage linguistics. In: Montrul, S., Polinsky, M. (eds), *The Cambridge Handbook of Heritage Languages and Linguistics*. Cambridge University Press, 1–10.

[15] Oh, J. S., Jun, S.-A., Knightly, L. M., Au, T. K. 2003. Holding on to childhood language memory. *Cognition* 86, B53–B64.

[16] Polinsky, M., Kagan, O. 2007. Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass* 1, 368–395.

[17] R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[18] Revelle, W. 2022. *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.

[19] Rothman, J. 2009. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism* 13(2), 155–163.

[20] Shin, E. 2005. The perception of foreign accents in spoken Korean by prosody: Comparison of heritage and non-heritage speakers. *The Korean Language in America* 10, 103–118.

[21] Trofimovich, P., Baker, W. 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28(1), 1–30.

[22] Wrembel, M., Marecka, M., Szewczyk, J., Otwinowska, A. 2019. The predictors of foreign-accentedness in the home language of Polish-English bilingual children. *Bilingualism: Language and Cognition* 22(2), 383–400.

The production of Mandarin /r/ by early and late Japanese-Mandarin bilinguals: articulatory and acoustic findings

Zhiqiang Zhu, Peggy Pik Ki Mok

The Chinese University of Hong Kong
zhiqiangzhu@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

The acquisition of English /r/ by Japanese learners is a well-known issue in speech acquisition. However, few studies have been done on how Japanese learners acquire rhotic sounds in other languages. This study aims to fill this gap by collecting both articulatory and acoustic data of Mandarin /r/ sounds by Japanese learners and native Beijing Mandarin speakers. Past studies reported that early language experience could benefit the speakers in the long run. This study also focuses on comparison between early bilinguals and late bilinguals. The results suggested that the early bilinguals had an overall advantage over the late bilinguals. The late bilinguals tended to use similar articulation gestures for all allophones of Mandarin /r/, have longer duration for postvocalic /r/ and syllabic /r/, and produce Mandarin /r/ with higher F2 and F3. The results will be discussed under the theoretical frameworks of the SLM, PAM, and NLM-e.

Keywords: Mandarin /r/, Japanese learners, simultaneous bilingual, articulation, speech acquisition

1. INTRODUCTION

Since the 1980s, the acquisition of English /r/ by Japanese learners has become a well-known topic in second language acquisition [1], [2], which directly influenced the development of many popular speech acquisition theories, e.g., SLM [3], [4], PAM [5], [6], NLM [7], [8]. There are two main reasons why Japanese learners have difficulty learning English /r/. First, rhotic sounds do not exist in standard Japanese. Second, Japanese learners typically tend to assimilate both English /r/ and /l/ into a single sound category, the Japanese /ɾ/ [9]. However, the assimilation is asymmetric, Japanese learners are more likely to assimilate English /l/ than English /r/ to Japanese /ɾ/ [10], [11] .

Despite intensive studies on the acquisition of English /r/ by Japanese, only a few studies have been done on Japanese learners' acquisition of rhotic sounds in other languages [12], for example, Mandarin Chinese /r/. Table 1 compares the Mandarin /r/, English /r/, and Japanese /ɾ/ in terms of syllable position, acoustic feature, and articulatory feature. Given the similarities and differences between Mandarin /r/ and English /r/, it would be theoretically interesting to examine Japanese learners' production of Mandarin /r/.

Another focus of this study is the comparison between early language learners (Japanese-Mandarin simultaneous bilingual, SB speakers) and late language learners (advanced Japanese learners of Mandarin, AJ speakers). The differences between early and late language learners in speech acquisition are a long-debated issue. Compared with late bilinguals, cross-linguistic influences between the two languages start early on and last for a longer time for early bilinguals, which can lead to non-native performance in vulnerable domains [15]. However, some studies have shown that early language experience, even over-hearing, can significantly boost a speaker's production and perception of that language later in life compared to L2 learners with no prior experience [16], [17]. Nevertheless, it is still unclear whether such an advantage can be extended to articulation. Using ultrasound imaging, this study can help to see, first, if early bilinguals generally perform better than the late bilinguals in producing Mandarin /r/; second, whether such an advantage of early bilinguals can be seen in Mandarin /r/ articulation.

	Mandarin /r/	**English /r/**	**Japanese /ɾ/**
Syllable position	Prevocalic, syllabic, postvocalic.	Prevocalic, syllabic, postvocalic.	Only prevocalic.
Acoustic feature	Low F3, but relatively higher than English /r/.	Low F3.	High F2 and F3 [13].
Articulatory feature	Prevocalic /r/: only bunched gesture. Syllabic and postvocalic /r/: both bunched and retroflex gestures [14].	Bunched and retroflex gestures are used in all syllable positions.	Apico-alveolar tap.

Table 1: Comparisons between Mandarin /r/, English /r/, and Japanese /ɾ/.

For theoretical predictions, the SLM and the PAM differ regarding Japanese learners' articulation of Mandarin /r/ sounds. The SLM claims that once the learners can discern the difference between an L2 sound from its closest native sound, they can establish a new sound category. However, sound categories are established based on auditory cues; therefore, Japanese learners may not acquire the subtle articulatory variation within Mandarin /r/. However, the PAM claims that articulatory gestures serve as the primitives for speech perception; thus, the Japanese learners can learn articulatory variation for the Mandarin /r/ sounds. In addition, the NLM-e claims that simultaneous bilinguals can successfully map phonetic features of two languages onto separate perceptual spaces like the monolinguals did for their one native language. Therefore, compared with the AJ speakers, the SB speakers have already well-established the Mandarin /r/ category in their mental representation and should have a more nativelike production performance than the AJ speakers.

2. METHOD

2.1. Participants

This study consists of three groups: 8 advanced L1 Japanese learners of Mandarin (AJ speakers), 8 Japanese-Mandarin simultaneous bilinguals (SB speakers) and 8 native Beijing Mandarin speakers (NM speakers). Details of the participants are listed in Table 2. Notably, the AJ speakers (late bilinguals) acquired Mandarin after adulthood, they were classified as "advanced" by two main criteria: they must have over one year of immersion in Beijing and have passed the HSK-6 test, which is the highest level for Mandarin learners. This study chose "advanced" learners to minimize language proficiency's influence on comparison between early and late bilinguals. The SB learners are early bilinguals of both Japanese and Mandarin because at least one of the SB speakers' parents was a native Mandarin speaker. The SB learners were exposed to both Japanese and Mandarin from an early age.

Group of speakers	Details
Advanced L1 Japanese speakers (AJ speakers)	8 participants (2M, *Mean age* = 31, *SD* = 5.55), > 1-year immersion in Beijing and with HSK-6 level.
Japanese-Mandarin simultaneous bilingual speakers (SB speakers)	8 participants (4M, *Mean age* = 24, *SD* = 2.98). All of them were exposed to both Japanese and Mandarin from an early age, with their mothers being native Mandarin speakers. All with HSK-6 level.
Native Beijing Mandarin speakers (NM speakers)	8 participants (3M, *Mean age* = 21, *SD* = 2.06). All of them were born and raised in Beijing.

Table 2: Information of participants.

2.2. Stimuli and Procedure

Three sets of target stimuli were designed for the Mandarin /r/ sounds, including 14 tokens for Chinese prevocalic /r/ (e.g., "如, [ɹu]), 3 tokens for Chinese syllabic /r/ (e.g., "儿," [ɚ]), 32 tokens for Chinese postvocalic /r/ (e.g., "皮儿," [pʰiɹ]). All the stimuli were produced together with a carrier phrase, "他答__吧" (he answered__). All the stimuli were randomized with three repetitions. Thus, for each participant, there were (14+3+32) tokens × 3 repetitions = 147 tokens in total.

The recording sessions took place at a linguistic laboratory in Beijing. The articulatory data was collected with the Telemed Echo B ultrasound system (framerate at 81.6 fps, probe field of view at 92 degrees, depth at 80mm). The acoustic data was then gained by synchronizing with the ultrasound video using the AAA software (44.1kHz/16-bit sampling rate) [18].

The participants were seated on a chair with their jaws resting on the ultrasound probe during experiment. The participants also wore a specially designed helmet for ultrasound stabilization. Target words embedded into carrier phrases were shown in PPT slides on a laptop screen in front of the participants. The participants were instructed to read sentences at a normal speech rate without any pauses. Before the recording, the experimenter would fine-tune the placement of the ultrasound settings for an optimal image.

2.3. Measurement and data analysis

The articulatory analysis included both tongue shape categorization and tongue contour comparison. Following the tongue shape categorization criteria of previous studies [19], [20], this study classifies the Mandarin /r/ into two types, retroflex gesture (the tongue tip curling up) and bunched gesture (no sign of retroflex, the tongue dorsum is bunched). The tongue splines were drawn manually at the key frame where the maximal constriction of Mandarin /r/ occurred. 42 equally spaced data points were exported as x–y Cartesian coordinates without rotation for each spline. The tongue contours comparison was then made via smoothing spline ANOVA using the "ggplot2" and "ggs" packages in R [21], [22]

For the acoustic data, the /r/ sound, together with its adjacent vowel, was labelled in Praat, and the acoustic target of the Mandarin /r/ was identified at the minimum F3 point. Duration of the /r/ sound, F1, F2, F3, and F3-F2 of the acoustic target was modelled in linear mixed effect (LME) models, respectively [23]. Due to page limits, only the results of duration, F2, and F3 are reported here.

In addition, four native Beijing Mandarin speakers (4F, mean age = 22.5, SD = 1.91) were invited on giving ratings to all the target stimuli. The native judges were university students without hearing impairment. The target stimuli with their carrier phrase were randomized across all the speakers. Before the experiment, the native judges were told that they would hear a series of utterances produced by either native Beijing Mandarin speakers or Japanese learners. The native judges were then instructed to give ratings to the target stimuli on goodness and accentedness based on 7-point scales, from 1 (very good; no accent) to 7 (very bad; heavy accent). The ratings averaged across the four native speakers were reported for each group.

3. RESULTS

3.1. Articulatory results

The results of tongue shape categorization showed that the SB and NM groups have similar patterns for the Mandarin /r/ gestures, which differed from the AJ speakers. For the SB and NM groups, only bunched gesture is used for prevocalic /r/. Both retroflex and bunched gestures are used for syllabic /r/ and postvocalic /r/. However, retroflex gestures and bunched gestures are found in all phonological positions by the AJ speakers; four out of eight AJ speakers have used retroflex gesture for the prevocalic /r/. For illustration, Figure 1 shows the raw ultrasound images of Mandarin /ri/ by three speakers in separate groups. It is evident that AJ3 curled up the tongue to produce the Mandarin prevocalic /r/ (/ri/), while SB2 and NM1 both used bunched gestures for the Mandarin /ri/.

Figure 2 shows the tongue contours of the Mandarin /r/ sounds by the AJ2, SB2, and NM2 speakers, respectively. Due to page limits, this study cannot list all the participants' SSANOVA tongue contour images. The AJ2, SB2 and NM2 speakers were selected as typical productions from

speakers to illustrate the different tongue configurations. The results also yielded similar patterns between the SB and NM groups but not for the AJ group. As displayed in Figure 2, the SB and NM speakers (represented by SB2 and NM2) employed different articulatory gestures to produce the Mandarin /r/ sounds. They adopted similar tongue gestures for syllabic /r/ and postvocalic /r/, but not for prevocalic /r/. However, five out of eight AJ speakers, like AJ2, used similar tongue gestures for the Mandarin /r/ sounds at all syllable positions.

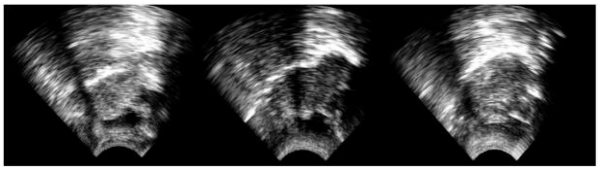


Figure 1: Raw ultrasound images of Mandarin /ri/ by the AJ3, SB2 and NM1 speakers respectively.

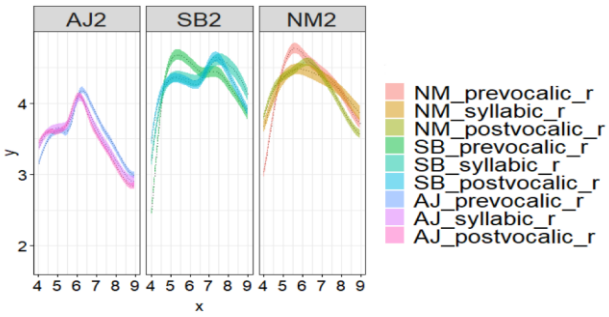


Figure 2: Tongue contours of the Mandarin /r/ sounds by the NM2, SB2, and AJ2 speakers, respectively.

3.2. Acoustic results

Duration: Figure 3 shows the duration of Mandarin /ɹ/ sounds at different syllable positions by the three groups. The best model for duration included Group, Syllable position, and the interaction between Group and Syllable position as fixed effects, and Participant and Utterance as random intercepts, with a random slope for Participant and Utterance on Syllable position, respectively. The model results showed that there was a main effect on Group and a main effect on Syllable position. The model results also showed a significant two-way interaction between Group and Syllable position. Post-hoc pairwise comparisons were performed on the duration in each syllable. The results showed that, in the prevocalic position, the duration between each two groups remained non-significant (p > .05). In the syllabic position and the postvocalic position, the duration of the AJ group was significantly longer than that of the NM group (for the syllabic position, Estimate =61.259, SE = 22.5, t = 2.717, p = .027; for the postvocalic position, Estimate =76.632, SE = 21.3, t = 3.600, p = .003). However, in the syllabic and postvocalic positions, the duration between the SB group and the NM group remained non-significant (p > .05).

F2: Figure 4 shows the F2 values of Mandarin /ɹ/ sounds at different syllable positions by the AJ, SB, and NM groups. The best model for F2 included Group, Syllable position, and the interaction between Group and Syllable position as fixed effects, and Participant and Utterance as random intercepts, with a random slope for Participant and Utterance on Syllable position, respectively. The model results showed that there was a main effect on Group and a main effect on Syllable position. The model results also yielded a significant two-way interaction between Group and Syllable position. Post-hoc pairwise comparisons were performed on formant values in each syllable. The results suggested that the F2 values of the AJ group were significantly higher than that of the NM and SB groups in all three syllable positions (compared with the NM group, in the prevocalic position, Estimate =1.079, SE = 0.238, t = 4.530, p < .001; in the syllabic position, Estimate =0.675, SE = 0.267, t = 2.528, p =.038; in the postvocalic position, Estimate =0.655, SE = 0.235, t = 2.785, p =.024). However, the F2 values remain non-significant between the SB and the NM group in all syllable positions (p > .05).

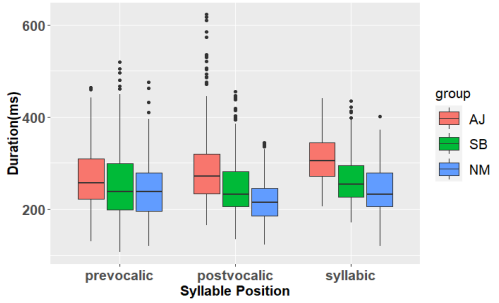


Figure 3: Duration of Mandarin /ɹ/ sounds at different syllable positions by the AJ, SB, and NM groups.

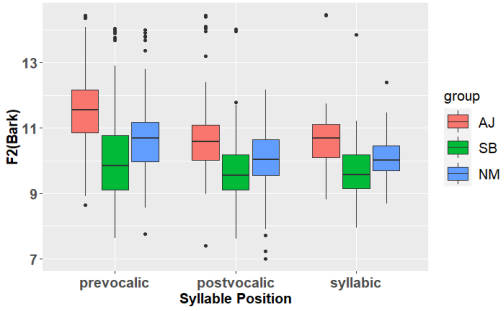


Figure 4: F2 values of Mandarin /ɹ/ sounds at different syllable positions by the AJ, SB, and NM groups.

F3: Figure 5 shows the F3 values of Mandarin /ɹ/ sounds at different syllable positions by the AJ, SB, and NM groups. The best model for F3 included Group and Syllable position as fixed effects, and Participant and Utterance as random intercepts, with a random slope for Participant and Utterance on Syllable position, respectively. The model results showed that there was a main effect on Group and a main effect on Syllable position. Post-hoc pairwise comparisons were performed on formant values in each syllable. The results were similar to the F2 values, which suggested that the F3 values of the AJ group were significantly higher than that of the NM and SB groups in all three syllable positions (compared with the NM group, in the prevocalic position, Estimate =1.042, SE = 0.245, t = 4.261, p < .001; in the syllabic position, Estimate =0.918, SE = 0.256, t = 3.582, p = .003; in the postvocalic position, Estimate =1.103, SE = 0.244, t = 4.526, p < .001). However,

the F3 values remain non-significant between the SB and the NM group in all syllable positions ($p > .05$).

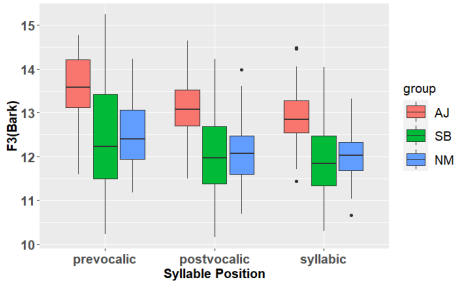


Figure 5: F3 values of Mandarin /ɹ/ sounds at different syllable positions by the AJ, SB, and NM groups.

3.3. Ratings by native speakers

Figure 6 shows the averaged ratings of the target stimuli by the four native Mandarin speakers. Higher scores in the accentedness and goodness ratings indicated that the target stimuli were judged as more accented and less perfect, respectively. For the goodness rating, the scores in all syllable positions ranked as AJ group > SB group > NM group. It was the same case for the accentedness rating in all syllable positions. The rating results suggested that the native Mandarin speakers gave better scores to the SB group than that of the AJ group for all allophones of Mandarin /r/ sounds.

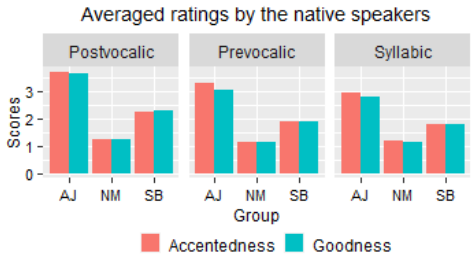


Figure 6: Averaged ratings of goodness and accentedness by the native Mandarin speakers.

4. DISCUSSION

This study examines the production of Mandarin /r/ by Japanese-Mandarin simultaneous bilinguals and advanced Japanese learners of Mandarin, which confirms that simultaneous bilinguals (early bilinguals, SB) have an overall advantage over advanced Japanese learners (late bilinguals, AJ). Compared with the SB and NM groups, the production differences of the AJ group are mainly reflected in four aspects: 1) using the retroflex gesture for prevocalic /r/, 2) employing similar gestures for all allophones of Mandarin /r/, 3) having excessively longer duration for syllabic /r/ and postvocalic /r/, 4) producing Mandarin /r/ sounds with significantly higher F2 and F3. Also, the ratings by the native Mandarin speakers showed that scores of the SB group were better than that of the AJ group in both goodness and accentedness. The advantage of the early bilinguals over the late bilinguals corroborated the NLM-e, which claims that simultaneous bilinguals can successfully acquire phonetic features in both languages. This study demonstrates that the SB group successfully acquired phonetic features of Mandarin /r/, while the AJ group showed more production biases. The following discussion will compare the production performance between the AJ and SB groups in terms of articulation, duration, and formant values, respectively.

Articulation: the articulatory patterns suggested that the SB and NM speakers could discern subtle articulatory variations between Mandarin prevocalic /r/ and Mandarin syllabic/postvocalic /r/, while some AJ speakers could not. Therefore, most of the AJ speakers used similar gestures to produce all the allophones of Mandarin /r/ sounds. In addition, half of AJ speakers transferred their retroflex gesture for the syllabic/postvocalic /r/ into the prevocalic /r/. The results partially supported both the PAM and the SLM. The PAM was supported by the SB speakers and a few AJ speakers that employed different gestures for the Mandarin /r/ sounds. The SLM was supported by most of the AJ speakers that they could not acquire the subtle articulatory variations within Mandarin /r/.

Duration: the AJ group demonstrated significantly longer durations in producing the syllabic /r/ and the postvocalic /r/ compared to the SB and NM groups. Excessively longer durations in segments are a common production bias observed in second language acquisition [24]. It is possible that the AJ speakers exaggerated the duration of these sounds to intensify the rhoticity of the syllabic /r/ and postvocalic /r/. However, the SB and NM groups did not use this technique. Instead, the rhoticity of the Mandarin /r/ for these groups was primarily reflected in the low F3 and small distance between F3 and F2.

Formant values: compared with the SB and NM speakers, the AJ speakers produced Mandarin /r/ in all syllable positions with higher F2 and F3. Since a low F3 is an essential indicator of rhotic sounds [25], a higher F3 suggested that the Mandarin /r/ sounds produced by the AJ speakers were less rhotic compared with that of the SB and NM groups. In addition, Japanese /ɾ/ is characterized by high F2 and F3. Compared with the SB speakers, the Mandarin /r/ sounds produced by the AJ speakers were more heavily influenced by their Japanese /ɾ/. This finding is also compatible with the claims of NLM-e, because the SB speakers had better performance in acquiring the phonetic features of the Mandarin /r/ than the AJ speakers.

In conclusion, this study demonstrates that the Japanese-Mandarin simultaneous bilinguals (early bilinguals) outperformed the advanced Japanese of Mandarin (late bilinguals) in Mandarin /r/ production. The production biases of the late bilinguals were manifested in articulation, duration, and formant values. The results of this study corroborated the NLM-e, while the SLM and the PAM were only partially supported. This study has three main implications for future research. First, to further investigate the assimilation relationship between Mandarin /r/-/l/ to the Japanese /ɾ/, the production of Mandarin /l/ and Japanese /ɾ/ by Japanese learners and native speakers should be examined as well. Second, more participants should be recruited for both Japanese-dominant simultaneous bilinguals and Mandarin-dominant simultaneous bilinguals to investigate the effect of language dominance. Third, perception experiments should be conducted to test perceptual relationship between Mandarin /r/-/l/ and Japanese /ɾ/.

5. REFERENCES

[1] K. Miyawaki, J. J. Jenkins, W. Strange, A. M. Liberman, R. Verbrugge, and O. Fujimura, "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," *Percept. Psychophys.*, vol. 18, no. 5, pp. 331–340, 1975.

[2] S. G. Guion, J. E. Flege, R. Akahane-Yamada, and J. C. Pruitt, "An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants," *J. Acoust. Soc. Am.*, vol. 107, no. 5, pp. 2711–2724, 2000.

[3] J. E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 233–277.

[4] J. E. Flege and O.-S. Bohn, "The revised Speech Learning Model (SLM-r)," doi: 10.13140/RG.2.2.27529.06249.

[5] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. York Press, 1995, pp. 171–204.

[6] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language experience in second language speech learning: In honor of James Emil Flege*, O.-S. Bohn and M. J. Munro, Eds. John Benjamins, 2007, pp. 13–34.

[7] P. K. Kuhl and P. Iverson, "Linguistic experience and the 'perceptual magnet effect,'" in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 121–154.

[8] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, "Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)," *Philos. Trans. Biol. Sci.*, vol. 363, no. 1493, pp. 979–1000, 2008.

[9] N. Takagi and V. Mann, "The limits of extended naturalistic exposure on the perceptual mastery of English/r/and/l/by adult Japanese learners of English," *Appl. Psycholinguist.*, vol. 16, no. 4, pp. 380–406, 1995.

[10] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese/r/and English/l/and/r," *J. Phon.*, vol. 32, no. 2, pp. 233–250, 2004.

[11] K. Hattori and P. Iverson, "English/r/-/l/category assimilation by Japanese adults: Individual differences and the link to identification accuracy," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 469–479, 2009.

[12] J. Larson-Hall, "Predicting perceptual success with segments: a test of Japanese speakers of Russian," 2004.

[13] P. Iverson, V. Hazan, and K. Bannister, "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3267–3278, 2005.

[14] S. Chen and P. P. K. Mok, "Articulatory and acoustic features of Mandarin/ɹ: A preliminary study," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[15] V. Yip and S. Matthews, *The bilingual child: Early development and language contact*. Cambridge University Press, 2007.

[16] L. M. Knightly, S.-A. Jun, J. S. Oh, and T. K. Au, "Production benefits of childhood overhearing," *J. Acoust. Soc. Am.*, vol. 114, no. 1, pp. 465–474, 2003.

[17] S. A. Montrul, "Incomplete Acquisition in Bilingualism: Re-examining the Age Factor," vol. 39, Sep. 2008, doi: 10.1075/SIBIL.39.

[18] Articulate Instruments Ltd, "Articulate Assistant Advanced User Guide: Version 2.16." Edinburgh, UK, 2012.

[19] H. B. Klein, T. M. Byun, L. Davidson, and M. I. Grigos, "A multidimensional investigation of children's/r/productions: Perceptual, ultrasound, and acoustic measures," *Am. J. Speech-Language Pathol.*, vol. 22, pp. 540–553, 2013.

[20] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[21] H. Wickham, *ggplot2: elegant graphics for data analysis*. New York: Springer, 2009.

[22] C. Gu, "Smoothing spline ANOVA models: R package gss," *J. Stat. Softw.*, vol. 58, pp. 1–25, 2014.

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2014.

[24] R. E. Baker, M. Baese-Berk, L. Bonnasse-Gahot, M. Kim, K. J. Van Engen, and A. R. Bradlow, "Word durations in non-native English," *J. Phon.*, vol. 39, no. 1, pp. 1–17, 2011.

[25] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of American English/r," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 343–356, 2000.

PERCEPTION-PRODUCTION MISMATCH IN L1 DRIFT

Jolanta Sypiańska[1] and Zuzanna Cal[2]

[1]University of Szczecin, Poland; [2]Adam Mickiewicz University, Poznań, Poland
jolanta.sypianska@usz.edu.pl, zuzanna.cal@amu.edu.pl

ABSTRACT

Most studies on the relationship between perception and production show an interdependency of the two modalities. We investigated whether the perception-production link is also present in L1 drift. L1 Polish/L2 English speakers were asked to decide whether they heard a voiced or voiceless stop in a two-alternative forced-choice (2AFC) task. The stimuli were obtained from VOT continua of word-initial (/p,b,t,d,k,g/). We expected that L1 speakers of a true voicing may experience L1 perceptual drift from the aspirating L2 English by placing the boundary between voiced and voiceless at higher VOT values. The results showed that the speakers perceived the voiceless-voiced boundary at 0ms. However, they produced a great amount of L1 aspirated stops and realized some L1 voiced stops either as voiceless or with limited prevoicing. We conclude that, in L1 drift situations, there may be a perception-production mismatch that defines L1 drift as an element of execution without influencing phonological representations.

Keywords: L1 drift, perception, production, aspirating language, true voicing language.

1. INTRODUCTION

There is a body of research that supports the interdependence of perception and production both in the native [1] and foreign languages [2,3]. The two modalities must share or have access to at least some part of the sounds' representation. Moreover, the perception-production link defines perception as a precondition to production. Speakers first possess the ability to recognise that the sound belongs to a particular category, only then can they produce the sound accurately. Widely accepted models of speech production in foreign language acquisition, such as the Speech Learning Model [4] or the Perceptual Assimilation Model [5], advocate the link by stressing that accurate production of an L2 sound is possible only when the learner establishes an L2 sound category that is sufficiently different from a similar L1 sound category. However, some studies have found a perception-production mismatch where either perception exceeded production [2,6] or production but not perception was on target with regard to native speaker norms [7].

This paper focuses on the perception and production of voiced-voiceless stops. The differences in voice onset time (VOT) that lead to differences in fortis and lenis stop realisation in true voicing languages versus aspirating languages have been a frequently discussed issue in L1 drift literature. In terms of perception, [8] found that L1 English, L2 French novice learners shifted their boundary between voiced and voiceless stops towards the more French-like values after intensive exposure to L2 French in L2 immersion. [7] compared L1 Spanish, L2 English and L1 English, L2 Spanish groups with different lengths of residence in the L2 country and showed that each group had less native-like L1 category boundaries. However, in an L1 setting, no statistically significant shift in the boundary between the voiced and voiceless bilabial stops was observed [9].

In production, [10] found that L1 Japanese, L2 English speakers prolong their L1 Japanese VOT in Japanese short-lag stops under the influence of L2 English long-lag stops. Interestingly, [11] describe an asymmetry in the susceptibility to L1 drift between the voiced and voiceless stop series in which the L1 Polish voiced stops showed more drift than the voiceless stops under the influence of L2 English. Also, [12] noticed that only voiced stops in both languages of two English-Czech simultaneous bilinguals drifted towards more negative, so more Czech-like values, after an extended stay in Czechia. There was no change in the voiceless stops for either language. However, in a study on L1 Polish senior learners of L2 English, it was mostly the voiceless series that underwent L1 drift and the speakers largely maintained prevoicing in the voiced series [13].

In terms of the perception-production link in L1 drift setting, [14] showed that the L1 English, L2 Spanish group with the greatest amount of L2 experience produced short VOT in L1 voiceless stops (more Spanish like) but placed the perceptual boundary for voiced and voiceless at higher VOT values in comparison to other groups. No statistically significant correlations between production and perception were found in other groups in the study. These results suggest that there may be no clear link between the two dimensions.

The perception-production link has been explored in an L1 drift setting only to a limited extent. Thus, the aim of the paper is to investigate L1 drift in perception and production of voiced-voiceless stops, an element that has previously been shown to undergo drift in both modalities but was mostly investigated separately. The first hypothesis is that L1 drift will be manifested in both modalities of L1 Polish fortis and lenis stops under the influence of L2 English. The second hypothesis is based on the perception-production link and states that the perception of L1 Polish stops will be reflected in L1 production. The manifestation of the link should be visible in a higher VOT category for fortis and lenis stops reflected in a tendency to produce aspirated voiceless fortis stops and voiceless lenis stops in L1 Polish.

2. METHOD

2.1. Participants

Twenty-four L1 Polish, L2 English speakers took part in the experiment including twenty-three females and one male participant. They were all Polish students of English studies at Adam Mickiewicz University in Poznań whose age ranged from 19 to 27. They were advanced learners of English and the L2 proficiency was measured with the Lexical Test for Advanced Learners of English (LexTale) [15]. Their self-reported average of English instruction per week was 20 hours. Most of them reported past exposure to German but currently claimed having no instruction or other forms of exposure in an aspirating language.

2.2. Procedure

The experiment consisted of four parts. First of all, the participants were asked to fill in a questionnaire, which included questions about their biodata as well as language learning history and language use. After that, they completed LexTale [15] as a measure of English proficiency. Having completed the test, they participated in the perception experiment and later took part in a recording session.

2.2.1. Perception task

The perception part of the experiment involved preparation of VOT continua based on one-syllable minimal pairs with word initial stop sounds (/p,b,t,d,k,g/). The continua ranged from -96 ms to +96 ms with 17 steps each of 12 ms and were prepared using one prevoiced and one aspirated stop produced by a trained phonetician with Polish as the native language. Three places of articulation namely bilabial, dental and velar were taken into consideration. The tokens were presented in one syllable words including par, bar, tam, dam, kas, gaz. Participants took part in a two-alternative forced-choice (2AFC) task programmed in the E-Prime 3.0 software [16]. After hearing a sound, they were asked to decide which word, out of two presented on a screen, they heard. The presented words differed only by voicing of the initial stop sounds. Each sound from the continua appeared twice in the experiment, thus the participants were presented with 102 tokens altogether.

2.2.2. Production task

In the production task, the participants were recorded reading the minimal-pair words used in the perception experiment with word-initial voiced and voiceless stops (par, bar, tam, dam, kas, gaz) mixed with filler words. The participants were presented with the words on a computer screen via a PowerPoint presentation and were asked to read words as they appeared. Finally, in order to check whether the participants themselves produce aspirated and voiceless short-lag stops in their L2, they were asked to produce English words with initial voiced and voiceless plosives in an analogical context to the L1 vocabulary used for the E-prime experiment (pat, bat, tan, Dan, cast, gas), also together with filler words. The recording session took place in a quiet room and the task was approximately 5 minutes long.

3. RESULTS

3.1. Perception

To calculate the location of perceptual boundaries in each place of articulation (PoA), the data was converted using a logistic regression function in SPSS [17] and the boundary location was computed with the use of the following formula: $-LN(b0)/LN(b1)$, where $b0$ corresponded to the constant and $b1$ to the slope of the function. Figure 1 presents mean values of boundary location three PoA - bilabial, coronal and velar.

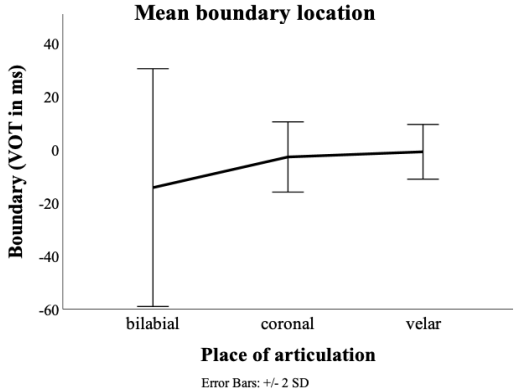


Figure 1. Mean perceptual boundary locations (in ms) for all three places of articulation

The bilabial continuum turned out to have the earliest boundary yielding on average -14.22 ms (SD = 22.38). In coronal stops the mean boundary location was -2.68 ms (SD = 6.61) and the velar continuum appeared to have a slightly higher boundary, that is -0.74 ms (SD = 5.15).

3.2. Production

The participants' stop realisations (N=283, 142 fortis sounds, 141 lenis sounds) were analysed in Praat by measuring VOT manually. The productions were then coded into five categories, two for fortis stops: aspirated and voiceless, and three for voiced stops: prevoiced, voiceless and mixed (Fig.2). The last category refers to tokens of voiced stops that include portions of voiceless signal at some point between the closure and the release as opposed to prevoiced tokens which maintain voicing throughout the duration of the closure. The results show that the speakers have acquired aspiration and use it as a main type of realisation for fortis L2 English stops (almost 100% correctness, Fig.3). Aspirated stops also dominate the L1 fortis stop production which is why more than two-thirds of the Polish fortis stops are produced incorrectly. Lenis stop production is more varied for both languages. In L1 Polish, the correct prevoiced realisation is present in 60% of tokens whereas the incorrect productions belong to the voiceless or mixed categories. Finally, prevoiced and mixed tokens in L2 English productions amount to approximately 50% of incorrect realisations (Fig.3.).

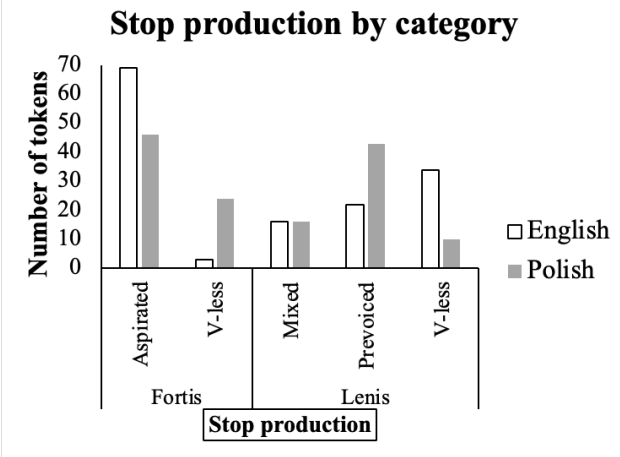


Figure 2. Number of tokens in Polish and English produced in each category for fortis stops: aspirated, voiceless, and lenis stops: mixed, prevoiced and voiceless

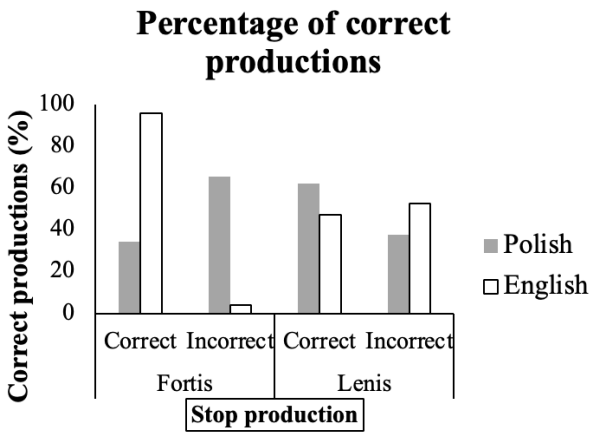


Figure 3. Percentage of correct fortis and lenis stop productions in Polish and English

3.3. Perception-production link

A linear regression was run to examine the link between perception and production. The boundary (dependent variable) was regressed on predicting categorical variables of fortis realisation (voiceless, aspirated) and lenis realisation (voiceless, prevoiced, mixed). The analysis was conducted separately for each continuum. Table 1 provides the results of the statistical analysis.

Dependent variable: Perceptual boundary (VOT in ms)

Independent variable	Coefficient (B)	Standard Error	t-Statistic	P-value
PoA: bilabial				
Fortis[a]				
Voiceless	7.909	11.553	.685	.504
Lenis[b]				
Voiceless	7.428	17.967	.413	.686
Mixed	7.979	17.967	.444	.664
PoA: coronal				
Fortis[a]				
Voiceless	-.785	3.309	-.237	.815
Lenis[b]				
Voiceless	7.985	6.815	1.172	.258
Mixed	3.564	3.713	.960	.351
PoA: velar				
Fortis[a]				
Voiceless	-1.307	2.564	-.510	.616
Lenis[b]				
Voiceless	2.467	2.907	.849	.408
Mixed	4.776	2.569	1.859	.080

[a] reference factor = Aspirated
[b] reference factor = Prevoiced

Table 1. Results of linear regression across places of articulation

No effect of fortis realisation was found in all three places of articulation, that is bilabial (F=.469, p=.504), coronal (F=.56, p=.815) and velar (F=.260 p=.616). Similarly, lenis realisation was not found to

be a predicting variable for the perceptual boundary location in all places of articulation, namely bilabial (F=.162, p=.825), coronal (F=1.031, p=.378), velar (F=1.728, p=.207).

4. DISCUSSION AND CONCLUSION

The perception data showed that the participants placed the perceptual boundary between voiced and voiceless coronal and velar stops in their L1 Polish at approximately 0ms that is within the same 12ms step of the VOT continuum. The boundary for the bilabial stop was placed one step earlier in the prevoicing part of the stop. These results indicate that the speakers did not experience perceptual drift which would be manifested in a category boundary at higher (positive) VOT values.

The production data from the L2 showed that the speakers have acquired aspiration in the L2, and at least partly, learned to suppress prevoicing in the L2 lenis stops. This gives a basis for cross-linguistic influence from the L2 to the L1. Furthermore, L1 production data revealed that aspiration was present in the majority of L1 tokens resulting in an error rate of more than 60%. However, lenis stops were produced with greater variation and only 60% of the tokens were correctly prevoiced.

The first hypothesis regarding L1 drift under the influence of the L2 in both L1 perception and production has only partly been corroborated. With a basis for cross-linguistic influence from the L2, it can be stated that L1 drift in Polish stemmed from L2 English. It can also be concluded that L1 drift in production was manifested in both voiced and voiceless stops but was greater in the later series. However, there was no perceptual L1 drift.

Lack of changes in the speakers' L1 perceptual boundary may be attributed to the fact that, unlike previous studies on perceptual drift in VOT [8,14], the current speakers resided in the L1 country before and at the moment of data gathering. Thus, they may not have had enough exposure to the L2 to trigger changes in L1 perception.

The results give further evidence of L1 drift in the form of prolonging L1 VOT in short-lag stops under the influence of L2 English long-lag stops as in [10,13]. However, they do not substantiate the asymmetry between voiced and voiceless series found in [11,12]. The differences in the results could be task-related as both [11,12] used a variation of an imitation protocol as opposed to words in isolation that are not previously heard or read as part of the study procedure.

Finally, the results of the regression revealed that neither fortis nor lenis realisations were statistically significant predictors of the perceptual boundary between voiced and voiceless stops. Although there was a tendency to produce aspirated voiceless fortis stops and voiceless lenis stops in L1 Polish, the category boundary was not placed at higher (positive) VOT values. Since our data exhibit drift in production but not in perception, we conclude that there is a mismatch between the two modalities. This partly confirms [7] whose results indicate no perception-production link apart from one group who showed L1 drift in production but not in perception. This result contradicts theories of speech perception in which perception precedes production, and accurate production is a consequence of correct perception.

All in all, we conclude that, in L1 drift situations, there may be a perception-production mismatch that could define L1 drift as an element of execution without influencing phonological representations.

6. REFERENCES

[1] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M. 2004. The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of. the Acoustical Society of America* 116, 2338–2344.

[2] Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., Yamada, T. 2004. Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics* 32, 233-250.

[3] Baker, W., Trofimovich, P. 2006. Perceptual paths to accurate production of L2 vowels: The role of individual differences. *International Review of Applied Linguistics* 44, 231-250.

[3] Best, C., Tyler, M. 2007. Nonnative and second-language speech perception. In: Bohn, O.-S., Munro, M. J. (eds), *Language experience in second language speech learning*. John Benjamins, 13–34.

[4] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In:. Strange, W. (ed), *Speech perception and linguistic experience: Issues in cross-linguistic research*. York Press, 233-277.

[5] Best, C., Tyler, M. 2007. Nonnative and second-language speech perception. In: Bohn, O.-S., Munro, M. J. (eds), *Language experience in second language speech learning*. John Benjamins, 13–34.

[6] Chan, A. Y. W. 2014. The perception and production of English speech sounds by Cantonese ESL learners in Hong Kong. *Linguistics* 52, 35-72.

[7] Darcy, I., Kruger, F. 2012. Vowel perception and production in Turkish children acquiring L2 German. *Journal of Phonetics* 40, 568-581.

[8] Tice, M., Woodley, M. 2012. Paguettes & bastries: Novice French learners show shifts in native phoneme boundaries. *LSA Annual Meeting Extended Abstracts* 3, 1-5.

[9] Osborne, D.M. 2016. "The acquisition of fine phonetic detail in a foreign language: perception and production of stops in L2 English and L1 Portuguese." Ph.D. dissertation, University of Arizona.

[10] Harada, T. 2003. L2 influence on L1 speech in the production of VOT. *Proc. 15th ICPHS* Barcelona, 1085-1088.

[11] Schwartz, G., Dzierla, J., Wojtkowiak, E. 2019. Laryngeal phonology and asymmetrical cross-language phonetic influence. In: Wrembel, M., Kiełkiewicz-Janowiak, A., Gąsiorowski, P. (eds), *New Approaches to the study of sound structure and speech: Interdisciplinary Work in Honour of Katarzyna Dziubalska-Kołaczyk*. Routledge, 316-325.

[12] Podlipský, V.J., Šimáčková, Š., Chládková, K. 2020. Phonetic drift reveals interconnected phonological representations in simultaneous bilinguals: A case study of English and Czech stop consonants. *International Journal of Bilingualism* 25, 483-832.

[13] Sypiańska, J. 2021. Production of voice onset time (VOT) by senior Polish learners of English. *Open Linguistics* 7, 316-330.

[14] Gorba, C. 2019. Bidirectional influence on L1 Spanish and L2 English stop perception: the role of L2 experience. *The Journal of the Acoustical Society of America* 145, 1-7.

[15] Lemhöfer, K., Broersma, M. 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44, 325-343.

[16] Psychology Software Tools, Inc. [E-Prime 3.0]. 2016. Retrieved from https://support.pstnet.com/.

[17] IBM Corp. Released 2021. IBM SPSS Statistics for Macintosh, Version 28.0. Armonk, NY: IBM Corp.

Does gemination resist linguistic attrition? A study on Italian migrant speech in Melbourne Australia

Valentina De Iacovo[1], Paolo Mairano[2], John Hajek[3]

[1]Università di Torino, [2]Université de Lille, [3]The University of Melbourne
valentina.deiacovo@unito.it, paolo.mairano@univ-lille.fr, j.hajek@unimelb.edu.au

ABSTRACT

In this study, we explore the possible effect of linguistic attrition on consonant gemination among speakers of Italian living in Australia. We analyse a corpus of spontaneous speech for a group of native Italian speakers who emigrated to Melbourne in Australia between the 1950s and 1970s. Given gemination is a fully lexicalised and stable phenomenon in Italian but not in English, we investigate whether, after many years of living in Australia, the phonological contrast between singleton and geminate consonants in Italian is preserved, or undergoes phonetic shift due to long-term contact with English. An analysis of acoustic durations reveals that all participants produced a significant difference in singleton and geminate consonant length, albeit with variable lengthening ratios. A post-hoc qualitative analysis seems to point to shorter geminate durations for speakers who have lived longer in Australia, but this will need to be supported by further research.

Keywords: gemination, L1 attrition, speech production, bilingualism, Italian.

1. THEORETICAL BACKGROUND

1.1 The Italian migrant community in Australia

The city of Melbourne has long been home to the largest Italian community in Australia as a result of mass migration from Italy that took place mainly between the 1950s and 1970s [1]. First generation migrants are today typically aged between 60 and 90 years old, and present a peculiar situation in terms of multilingualism/triglossia: they usually speak, as a result of simultaneous acquisition, (a) a regional dialect as spoken in their original village or hometown, (b) a regional variety of standard Italian supported by schooling in Italy before departure, as well as by interactions with other Italians in Australia, in addition to (c) an Australian-based variety of English which they have learnt later in life in order to integrate into Australian society. Subsequent generations mainly speak Australian English, and potentially also Italian, and, less frequently dialect, as heritage language(s) (HL) [2, 3]. The general picture described in the literature (e.g. [4]) reveals a complex and multifaceted linguistic situation, whereby the L1 Italian of these migrants can undergo different degrees of attrition, seen in a gradual decline in native language proficiency [5] according to the degree of influence of their later acquired L2 English, which also depends on social factors such as age of acquisition, level of education, family background, the use of English at work or for university studies. The linguistic studies carried out so far on Italians in Australia have often adopted a contrastive approach across generations [6] taking into account linguistic change and adaptation, with a particular focus on language shift [2] and code-switching [7]. While they are often far-reaching in scope, existing studies have noticeably paid little or no attention to the consequences of long-term language contact and shift on the phonetics and phonology of the Italian spoken by the first generation. The few exceptions to explore phonetic and/or phonological aspects of potential long-term attrition in L1 Italian outside Europe include [8], [9], [10] and [11]. The present contribution aims to add to this research by investigating potential phonetic and phonological attrition through the lens of a specific phonological feature of Italian: contrastive consonant gemination.

1.2 Linguistic attrition from a phonetic and phonological perspective

While linguistic attrition can impact on all areas of language proficiency and use, we focus our attention here on the phonetic and phonological aspects of it [12, 13, 14]. The term ‘phonetic attrition’ refers to L2-influenced phonetic change in an individual’s L1 system over the long-term, most typically with respect to migrants living in an L2 context [12]. The related concept of ‘phonological attrition’ refers to a reduction in the phonetic robustness (erosion) and even loss of a contrast [8, 13] and is particularly relevant here given our focus on the impact of long-term L2 contact on the maintenance or otherwise of the consonant length contrast in Italian.

1.3 Consonant gemination in Italian vs English

Gemination consists of a temporal extension of a sound. It is considered as typologically marked, since it is phonologically contrastive in relatively few languages of the world. One of these languages is Italian, where gemination is distinctive within and across morphemes for 15 consonants [15], namely /p/, /b/, /t/, /d/, /k/, /g/, /tʃ/, /dʒ/, /f/, /v/, /s/, /m/, /n/, /l/, /r/; additionally, it is allophonic for 5 other consonants (/ts/, /dz/, /ʃ/, /ʎ/, /ɲ/), while the 3 remaining consonants (/z/, /j/, /w/) are always short. An illustrative minimal pair *cane* (dog) /ˈkane/ [ˈkaːne] and *canne* (canes) /ˈkanːe/ [ˈkanːe] shows the gemination contrast. The duration of the geminate consonant is greater than that of singleton by a ratio of approx. 2:1 [16], with some variation depending on the category of sounds, [17, 18]; large-scale studies on running and spontaneous speech found a smaller ratio of approx. 1.7:1 [19]. Gemination also affects the preceding vowel, which is shortened before a geminate consonant (but only in nuclear position, cf. [19] and [20]). Additionally, gemination can also occur in standard Italian as an external sandhi phenomenon (referred to as *raddoppiamento fonosintattico*), but this phenomenon is not present in all varieties of Italian and will not be analysed in our study.

English, on the contrary, does not have a lexicalized long-short consonant contrast within morphemes, although in more careful styles there may be so-called *fake* gemination due to the concatenation of two morphemes ending and starting with the same sound, such as in *unnatural*. These fake geminates, not characteristic of all speakers, are also typically reduced in spontaneous speech (see [21] for detailed discussion and overview).

1.4 Gemination and attrition

Since gemination is typologically marked, and as such is restricted or absent in many languages, it is not unreasonable to hypothesise that it may undergo some degree of erosion for Italian migrants who have lived for many years in Australia. In fact, the erosion of the phonological contrast between singleton and geminate consonants in migrants and heritage speakers, as a result of contact with English, has been investigated in a series of production studies for L1 Farsi and L1 Arabic speakers living in Canada [22 and 23 respectively] and in the US [24]: their results point to evidence of attrition for this phonological contrast and a gradual shift towards English, whereby the acoustic duration of L1 geminate consonants is on average still longer than for singletons, but shorter than for geminate consonants produced by monolingual Farsi and Arabic speakers. Interestingly, generation was found to be a predictor of the degree of erosion, showing that it increased in successive generations (for example, [22] considered generations 1, 1.5 and 2). On the other hand, other studies have not uncovered evidence of geminate attrition in migrant speech: [25] found no differences in geminate production for HL speakers of Italian in Germany, while [26] did not find changes in the perception of geminates for L1 Hungarian speakers living in Germany and the US. [8] explored (de-)gemination in Italian migrants and heritage speakers in the US, and more particularly the perception of geminates by members of the Italian community from Lucca now living in San Francisco, CA. The authors found evidence of a progressive impairment in the perception of the length feature across generations, as well as differences between first generation immigrants and Lucchese speakers still living in the area around Lucca. No study has yet examined the production of geminate consonants in L1 Italian migrants or heritage speakers.

1.5 Aim of this study

In this contribution we present the results of a study aimed at exploring potential erosion patterns of gemination for L1 Italian migrants in Australia. As noted above, the existing studies focusing on the erosion of L1 gemination are scarce and report conflicting results: evidence of erosion is found for L1 Farsi [22], L1 Arabic ([23], [24]), L1 Italian [8] speakers, but not for Hungarian [26] speakers. For this preliminary study, we focus on whether first generation L1 Italian migrants in Australia produce geminates at all, and if so, whether geminate durations are comparable to those produced by L1 Italian speakers living in Italy.

2. DATA AND METHODS

2.1 Corpus

For this study 10 first-generation bilingual Italian-Australian speakers living in Melbourne (Australia) were recruited. The speakers (6 females and 4 males) are aged between 61 and 91 and have lived in Australia around 43 and 67 years. They were born in different parts of Italy and for the most part grew up there. They moved, at different ages, to Australia, where they learned English. Only two participants arrived relatively early in Australia (ages 6 and 7 respectively) and belong to what is sometimes referred to as generation 1.5. Upon arrival in Australia, four of them pursued their education/studies in English, while all participants worked in a mixed Italian-English environment, maintaining contact with members of the Italian

community and with their families. All participants have continued to speak Italian, as well as English, on a regular basis during their time in Australia. The speech material used for analysis is now contained within the so-called *SEIM* corpus, consisting of semi-spontaneous audio-visual interviews recorded in Italian by the first author with Italian bilinguals and heritage speakers with the aim of creating an oral archive for speech analysis. The interviews were conducted by following a set of questions addressing specific topics relating to their experience as migrants (arrival in Australia, lifestyle, use of Italian). Most speakers were interviewed individually by the first author, while three of them (speakers BE, AN, MA) were interviewed in each case with a partner – but their partner's speech was not analysed. The recordings were made using the microphone of an *IPad mini* at a sampling rate of 48 kHz. They have variable durations, mostly around 30-40 minutes. For the present analysis we considered the first 10 minutes of speech for each speaker.

2.2 Acoustic measurements

The audio files were extracted in wav format and resampled at 16kHz. The sound material was first transcribed orthographically manually, then phonetized and forced aligned using WebMaus [27], resulting in *textgrids*. Segmentation was subsequently checked manually in *Praat* [28]. The duration values of the target sounds were then extracted for all 15 geminating consonants (/tʃ, dʒ, p, t, k, b, d, g, m, n, f, s, v, r, l/), in word-medial contexts where gemination may potentially occur according to phonotactic restrictions of Italian, i.e. preceded by a vowel, and followed by a vowel, liquid or approximant. Consonants with few geminate occurrences were eliminated from the analysis (namely affricates, fricatives and voiced plosives, thereby leaving voiceless plosives, nasals and liquids). Due to limitations of space, we report only on the total number of tokens analysed per speaker (table 1) without further distinction by manner or other segment type.

Speaker	Singletons	Geminates
AD	442	86
AC	503	112
AP	298	49
AM	387	103
BM	463	121
CP	441	138
CM	365	74
CT	549	120
GB	486	105
MP	547	147

Table 1: Number of tokens analysed per speaker.

3. RESULTS

The data were then imported into *R* [29], and analysed with linear mixed-effects models with the *lme4* [30] library, *lmeRtest* [31] was used to obtain *p* values. The structure of random slopes was kept maximal.

Inspection of the data (figure 1) immediately illustrates the presence of geminate consonants in the speech of our speakers. This was also confirmed by a first generic model testing durational differences between singletons and geminates in our data, including speaker, sound and word as random factors: *duration ~ isGeminate + (isGeminate|speaker) + (isGeminate|sound) + (1|word)*. Sound was included as a random effect to account to the fact that different consonants have intrinsically different durations. This model clearly illustrated that geminates are longer than singletons by 31 ms on average (SE = .004, t = 8.097, p < .001) for our speakers. A second generic model with the same structure tested the duration of the preceding vowel, but in this case no significant effect was found. This is not surprising and simply reflects results on running speech by previous studies (e.g. [19]).

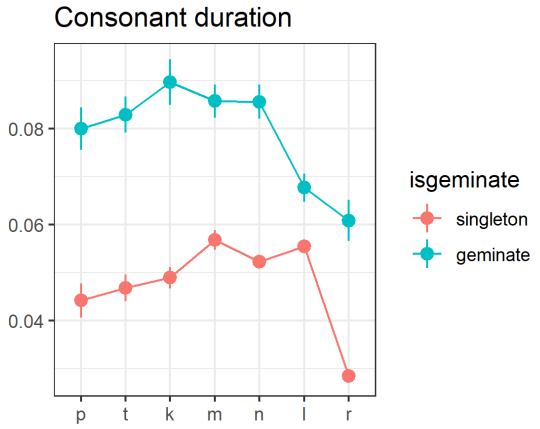


Figure 1: Mean consonant durations averaged across speakers.

Subsequently, we built a third model to investigate whether all our speakers produce geminate consonants, and with which ratios. With this in mind, we transformed speaker into a fixed (rather than random) variable, in order to be able to perform post-hoc analyses on each speaker, and ran the following model on consonant durations: *duration ~ isGeminate * speaker + (isGeminate|sound) + (1|word)*. An Anova of the model confirmed that the main effect of geminate status on consonant durations was highly significant (F = 88.28, p < .001), but also that consonant durations varied by speaker (F = 34.14, p < .001), and that there was an interaction between geminate status and speaker (F = 9.43, p < .001). A post-hoc analysis with Holm correction for multiple

comparisons revealed that the singleton-geminate difference in duration was significant for every single participant (all p values < .001). The predicted values for singletons and geminates for each speaker were then extracted from the model and are plotted in figure 2: these values illustrate the existence of a clear-cut distinction between geminates and singletons for every speaker.

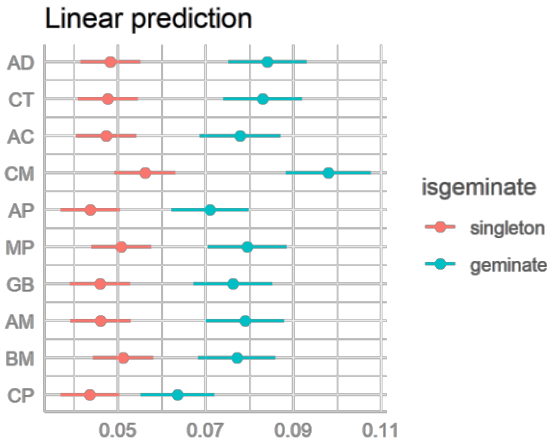


Figure 2: Linear prediction of geminate and singleton durations in sec. for every speaker.

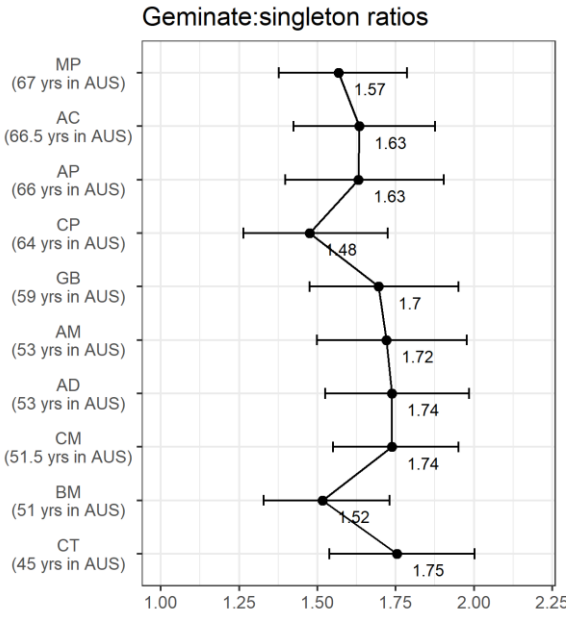


Figure 3: By-speaker geminate:singleton ratios computed on linear predictions for consonant durations.

Since the plot in figure 2 shows that the magnitude of lengthening is not the same for all participants, we used these values to compute by-speaker geminate:singleton ratios. These are shown in figure 3, where speakers are sorted by length of residence in Australia. This plot shows that most speakers have a geminate:singleton ratio between 1.6 and 1.7, which is comparable to what was found by [19] for running speech involving native speakers of Italian in Italy. Only two speakers stand out as having smaller ratios (1.52:1 and 1.48:1), namely CP and BM. This may be an indication of erosion for these two speakers. Interestingly, the two 1.5 generation speakers (AC and AD) do not stand out; instead, they behave in a similar fashion to first generation speakers in terms of gemination.

4. FINAL DISCUSSION

Our analysis of singleton and geminate consonant durations of L1 Italian migrants in Australia revealed little if any attrition: (1) our speakers maintain the geminate contrast, and (2) they generally do so with a magnitude of differentiation comparable to native speakers in Italy, i.e. with geminate:singleton ratios of 1.6:1 and 1.7:1 (similar to results obtained for running speech by [19]). This also includes our two 1.5 generation speakers, who do not show any patterns of gemination specific only to them. These last results appear to be consistent with [25], who report that geminates produced by their German-dominant Italian HL speakers were reliably longer than their singletons. That said, we do observe smaller ratios of gemination for two other speakers in our sample, which may be due to linguistic attrition; this may have various causes, and may depend on specific linguistic behaviours or on different permeability to linguistic erosion [12].

Finally, we also note that speakers who have lived in Australia for less than 60 years tend to have ratios of 1.7 or higher (with the exception of BM), while speakers with more than 60 years of residence show ratios of less than 1.7. This may be the indication of a trend to decrease the magnitude of lengthening for geminate consonants after many years of residence in Australia. The correlation between the number of years in Australia and geminate:singleton ratios is however non-significant (R = -.54, p = .107) and this trend could be an artefact due to the small number of participants available. Future larger-scale and more controlled studies may give us a deeper insight into this issue. For now, we are able to conclude that L1 Italian speakers in Australia produce geminate consonants with patterns that are comparable to those of native speakers in Italy, with lengthening ratios of 1.6-1.7. If erosion does apply to geminate consonants at all, then it may do so only at very late stages, or only for a minority of speakers.

5. ACKNOWLEDGEMENTS

We are grateful to all speakers, to people who collaborated in organising and setting up the interviews, and to the CO.AS.IT. of Melbourne.

6. REFERENCES

[1] Clyne, M., Hajek, J., Kipp, S. 2008. Tale of two multilingual cities in a multilingual continent. *People and Place* 16, 1–9.

[2] Bettoni, C. 1989. Language Loss Among Italians in Australia. *International Journal of Applied Linguistics* 83(1), 37–50.

[3] Clyne, M. 1987. Bilingualism and community languages: The educational consequences. *L'italiano in Australia* 31(1-2), 57–71.

[4] Bettoni, C. 1989. Maintaining and developing Italian in Australia. *Vox, The Journal of the Australian Advisory Council on Languages and Multicultural Education* 3, 48–53.

[5] Schmid, M. S. 2009. On L1 attrition and the linguistic system. *EuroSLA yearbook* 9(1), 212–244.

[6] Bettoni, C., Gibbons, J. 1988. Linguistic purism and language shift: a guise-voice study of the Italian Community in Sydney. *International Journal of the Sociology of Language* 72, 15–35.

[7] Rubino, A. 2004. Trilingual women as language mediators in the family: a Sicilian-Australian case study. *Australian Review of Applied Linguistics* 18, 25–49.

[8] Celata, C., Cancila, J. 2010. Phonological attrition and the perception of geminate consonants in the Lucchese community of San Francisco (CA). *International Journal of Bilingualism* 14, 1–25.

[9] Tordini, O., Galatà, V., Avesani, C., Vayra, M. 2018. Sound maintenance and change: Exploring inter-language phonetic influence in first-generation Italo-Australian immigrants. Il parlato nel contesto naturale: Modelli e metodi per lo studio del parlato in reali condizioni comunicative—Speech in the natural context. Models and methods for the analysis of speech under real communicative conditions. *Collana Studi AISV* 4, 77–98.

[10] Nodari, R., Celata, C., Nagy, N. 2019. Socio-indexical phonetic features in the heritage language context: Voiceless stop aspiration in the Calabrian community in Toronto. *Journal of Phonetics* 73, 91–112.

[11] Frontera, M. 2020. Calabrian migrants in Argentina: assessing first language phonetic attrition. *Proc. 16th AISV Conference* Rende (CS), 73–94.

[12] De Leeuw, E. 2019. Phonetic attrition. In: Schmid, M. S., Köpke, B. (eds), *The Oxford handbook of language attrition*. Oxford University Press, 204–217.

[13] Celata, C. 2019. Phonological attrition. In: Schmid, M. S., Köpke, B. (eds), *The Oxford handbook of language attrition*. Oxford University Press, 218–227.

[14] Chang, C. B. 2019. Phonetic drift. In: Schmid, M. S., Köpke, B. (eds), *The Oxford Handbook of Language Attrition*. Oxford University Press, 191–203.

[15] Loporcaro, M. 1996. On the analysis of geminate in Standard Italian and Italian dialects. In: Hurch, B., Rhodes, R. (eds), *Natural phonology: The state of the art.* Mouton de Gruyter, 153–187.

[16] Esposito, A., Di Benedetto, M. G. 1999. Acoustical and perceptual study of gemination in Italian stops. *The Journal of the Acoustical Society of America* 106(4), 2051–2062.

[17] Di Benedetto, M. G., De Nardis, L. 2021. Consonant gemination in Italian: The nasal and liquid case. *Speech Communication* 133, 62–80.

[18] Di Benedetto, M. G., De Nardis, L. 2021. Consonant gemination in Italian: The affricate and fricative case. *Speech Communication* 134, 86–108.

[19] Mairano, P., De Iacovo, V. 2020. Gemination in Northern versus Central and Southern Varieties of Italian: A Corpus-based Investigation. *Language and Speech* 63(3), 608–634.

[20] Bertinetto, P. M., Dell'Aglio, M., Agonigi, M. 2008. Quali fattori influenzano maggiormente la durata vocalica e consonantica in italiano? Un'indagine mediante l'algoritmo di decisione C&RT. *La comunicazione parlata*, 1000–1026.

[21] Kaye, A.S. 2005. Gemination in English. *English Today* 21(2), 43–55.

[22] Rafat, Y., Mohaghegh, M., Stevenson, R.A. 2017. Geminate attrition across three generations of Farsi-English bilinguals living in Canada: An acoustic study. *Ilha do Desterro Investigating Second Language Speech Learning* 71(3), 151–168.

[23] Alkhudidi, A., Rafat, Y., Stevenson, R. 2018. Geminate Attrition in the Speech of Arabic- English Bilinguals Living in Canada. *Heritage Language Journal* 17(1), 1–37.

[24] Hanini, R., Alkhudidi, A., Rafat, Y., Spinu, L. 2019. Geminate Attrition in the Speech of Arabic-English Bilinguals Living in the United States. *178th meeting of Acoustical Society of America* San Diego, CA.

[25] Einfeldt, M., van de Weijer, J., Kupisch, T. 2018. The production of geminates in Italian-dominant bilinguals and heritage speakers of Italian. *Language, Interaction and Acquisition* 10(2), 177–203.

[26] Neuberger, T. 2022. Effects of L2 exposure on the perception of the singleton/geminate contrast in Hungarian. *Alkalmazott Nyelvtudomány* 22(1), 191–207.

[27] Kisler, T., Uwe, R., Schiel, F. 2017: Multilingual processing of speech via web services, *Computer Speech & Language* 45, 326–347.

[28] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/)

[29] R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[30] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1–48.

[31] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 1–26.

Do Chinese-English Bilinguals Speak English Words with Lexical Tone in Mind?

Qing Yang & Yiya Chen

Centre for Linguistics, Leiden University; Leiden Institute for Brian and Cognition, Leiden University
q.yang@hum.leidenuniv.nl; yiya.chen@hum.leidenuniv.nl

ABSTRACT

Although it is well known that words of bilinguals' two languages interact extensively, whether and how language-specific suprasegmental features interact in bilingual lexical access remains unclear. This study investigated whether lexical tone affects pitch processing during English word production. Using the picture-word interference paradigm, we asked Chinese-English bilinguals and English monolinguals to name pictures in English while ignoring simultaneously played auditory Standard Chinese distractors. Crucially, these distractors are cross-language homophones to the English target names, which have a falling or a rising lexical tone in Standard Chinese. Naming latency results showed that cross-language homophones with rising-tone facilitated picture naming more than their falling-tone counterparts for the bilinguals. This effect was not found with English monolinguals. Such a difference suggests a significant influence of lexical tone on pitch processing during spoken word production even in these bilinguals' non-tonal language, lending evidence to the interaction between bilinguals' two languages at the suprasegmental level.

Keywords: lexical tone; spoken word production; the bilingual lexicon; picture-word interference paradigm

1. INTRODUCTION

The functional role of pitch variation differs across languages. In lexical tone languages such as Standard Chinese (hereafter, SC), pitch contour plays a crucial role in differentiating morpheme meanings, just as consonants and vowels (e.g., *ma* with a rising pitch contour means "hemp" but "scold" with a falling contour). For words in non-tonal languages such as English, pitch contour serves as a cue (together with cues such as duration and intensity) to distinguish a limited number of words, known as lexical stress [1]. For both types of languages, pitch variation also serves to signal utterance-level information such as sentence mode and information status [2]. For example, in most varieties of English, "Mary" can be uttered with a rising pitch contour to signal a question and a falling contour to signal a statement. There is a probabilistically stable mapping between pitch contour shape and sentence mode. In Standard Chinese, however, pitch variation for sentence mode is constrained by the pitch contours of lexical tones [3]. Such cross-language differences between SC and English in the form and function of pitch variations offer a unique case for investigating pitch processing in bilingual lexical access.

One widely accepted assumption is that bilinguals' two languages interact extensively [4]. For instance, the phonemes of bilinguals' two languages are found to be co-activated and compete for selection during speech production [5], [6]. However, most evidence for language interaction came from studies on segmental processing. Only a limited number of studies examined whether bilingual lexical access is influenced by suprasegmental properties such as lexical tone. Most of them have focused on subtle differences in tone perception between native and non-native tonal speakers [7], [8]. There is a surprising paucity of empirical research on how speaking a tonal language may affect bilinguals' pitch processing in the non-tonal language they command.

As one of the few studies that tapped into this issue, Ortega-Llebaria and her colleagues [9], [10] asked SC-English bilinguals and English monolinguals to perform a primed-lexical decision task. The prime and target were manipulated to fully match (e.g., "rice" with a falling-f0 – "rice" with a falling-f0), fully mismatch (e.g., "gold" with a rising-f0 – "rice" with a falling-f0), mismatch in segment (e.g., "mice" with a falling-f0 – "rice" with a falling-f0), or mismatch in pitch (e.g., "rice" with a falling-f0 – "rice" with a falling-f0). Results of reaction time showed that only SC-English bilinguals experienced larger facilitation across conditions when the targets were produced with a falling f0 than with a rising f0. This "falling-f0 bias" in SC-English bilinguals led the authors to propose that English words with falling f0 are closer to English lexical representations than those with rising f0 in the bilingual lexicon; therefore, English words with falling f0 were easier to access than their rising-f0 counterparts. Moreover, only SC-English bilinguals manifested the "falling-f0 bias" indicating the impact of long-term experience with lexical tone on bilinguals' pitch processing even in their non-tonal spoken word recognition.

Despite evidence from the comprehension domain on the effect of lexical tone on bilingual lexicon and

spoken word recognition, to our knowledge, no study has investigated whether and, if so, how lexical tone affects non-tonal lexical access during speech production. To develop a more comprehensive account of the role of lexical tone in cross-language interaction, empirical data from non-tonal spoken word production is still needed.

In this study, we aimed to fill this knowledge gap. Following Ortega-Llebaria and her colleagues [9], [10], one may hypothesize that the "falling-f0 bias" of SC-English bilinguals also plays a role in bilingual spoken word production. Using the picture-word interference paradigm (hereafter, PWI) [11], we asked native SC-English bilinguals and English monolinguals to name pictures in English while ignoring simultaneously played SC auditory distractors. Crucially, for the same target word (e.g., *lung*), there were four types of distractors (as illustrated in Table 1): 1) the target's cross-language homophone with a falling tone (CH_F; e.g., "lang4*", wave*); 2) the target's cross-language homophone with a rising tone (CH_R; e.g., "lang2*", wolf*); 3) unrelated distractor with a falling tone (UN_F; e.g., "you4*", right*); 4) unrelated distractor with a rising tone (UN_R; e.g., "you2*", oil*).

English Target	SC Distractors			
	Cross-language Homophone		Unrelated	
	Rising (CH_R)	Falling (CH_F)	Rising (UN_R)	Falling (UN_F)
fei4 **"lung"**	***lang2*** "wolf"	***lang4*** "wave"	***you2*** "oil"	***you4*** "right"

Table 1: A set of sample stimuli.

Previous bilingual PWI studies have found robust facilitation effects of cross-language homophones (e.g., [12]). We, therefore, expected to observe significant facilitation effects in both the CH_F and CH_R conditions, compared with the UN_F and UN_R conditions for both SC-English bilingual and monolingual speakers. Secondly, if lexical tone indeed shapes pitch processing in English spoken word production, the influence of falling vs. rising-tone SC homophone distractors on English picture naming was expected to differ between SC-English bilinguals and native English monolinguals. Furthermore, the "falling-f0 bias" of SC-English bilinguals, if at play in spoken word production, would lead to processing differences between the two homophone (i.e., CH_F vs. CH_R) conditions.

2. METHOD

2.1. Participants

48 SC-English bilinguals and 48 American English monolinguals participated in this study. All SC-English bilingual participants are native SC speakers who grew up in Beijing and speak no regional dialect. All participants learned English at an average age of 5.8 (SD=2.34). Their English proficiency level was accessed with an adapted LEAP-Q questionnaire [13] and the multilingual naming test (MINT; [14]). Using a Likert scale from one to ten, participants' self-rated frequency is 8.5 (SD = 1.4) in reading, 6.7 (SD = 1.8) in speaking, and 7.1 (SD = 1.8) in listening. The average correct response of MINT is 43% (SD = 5.1%). The English monolingual participants had no previous exposure to Mandarin or any other tonal languages. None of the participants had a history of language disorder. All participants provided informed consent and were compensated for their participation.

2.2. Materials

24 sets of critical stimuli were included Each set consists of an English target word, an SC cross-language homophone distractor with a falling lexical tone (CH_F), a cross-language homophone distractor with a rising tone (CH_R), an unrelated distractor with a falling tone (UN_F), and an unrelated distractor with a rising tone (UN_R). 12 sets of fillers which have no phonological overlap were also added. All target words are English picturable monosyllabic nouns. All distractor words are SC monosyllabic morphemes. Lexical frequency of distractors, as computed with SUBTLEX-CH [16], are balanced across conditions ([$F(3, 92) = 1.97$, $p = 0.13$]). Homophone density, as computed with DoWLS-MAN [17], was also controlled ([$F(3, 92) = 0.855$, $p = 0.47$]). The target pictures, which are black and white line drawings, are selected and adapted from the IPNP database (Bates et al., 2003) and the BOSStimuli database [19]. All spoken stimuli were produced by a male native SC speaker who was born and grew up in Beijing. The recording was done at the Phonetics Lab of the Leiden University Centre for Linguistics, through a Sennheiser MKH416T microphone (44.1 kHz, 16-bit). All stimuli were normalized for a duration of 400ms and intensity at 70dB in Praat [20].

2.3. Procedure

Participants took part in the experiment online using Gorilla (www.gorilla.sc). All participants were required to wear headphones and sit in a quiet room. Participants were only allowed to join the experiment

if they were using laptops. Before the experiment, a headphone check based on the dichotic pitch [21], a microphone check, and an auto-play check were run to screen participants' equipment. All the instructions were given in English. Before the naming task, there was a familiarization session, during which participants were shown all target pictures with their corresponding English names printed underneath for 1500ms. After the name disappeared, participants were asked to type in the picture's English name. If participants did not respond accurately, the intended name would be shown for 1500ms again. After the familiarization session, participants were asked to complete four practice trials with an option to practice more if they wish.

The picture naming task started after the familiarization and practice sessions. Within each trial, a fixation was first displayed in the centre of the screen for 500ms, followed by a picture and a simultaneously played auditory distractor (SOA = 0ms). Participants were asked to name the picture as quickly and as accurately as possible while ignoring the distractor. The picture remains on the screen for 2000ms. Response times were measured from picture onset until naming onset using Chronset [22]. If participants did not respond within the 2000ms interval, the next trial began automatically. Between each trial, there was a blank screen of 1000ms.

In total, there were 96 (24 x 4) critical trials and 48 (12 x 4) filler trials. All trials were equally distributed into four blocks with a Latin Square design so that participants only see each target picture once in one block. Between the blocks, participants were encouraged to take a short break. After the naming task, participants were asked to complete a language background survey (i.e., the adapted LEAP-Q questionnaire) and the language proficiency MINT test. In total, the experiment took about 30 minutes.

3. RESULTS

Trials with incorrect (3.2%) and empty responses (5.5%) were excluded from the data analysis. Table 2 summarizes the mean naming latency (ms) for each experimental condition. As we can see, English monolingual participants took about 40ms longer to name pictures with unrelated distractors (UN_R and UN_F) than with cross-language homophone distractors (CH_R and CH_F). Moreover, either with unrelated distractors or cross-language distractors, there was almost no difference (~5ms) between naming with falling-tone vs. rising-tone distractors. As for SC-English bilinguals, the overall naming latency was more than 70ms longer than English monolinguals in each condition. Naming latencies with cross-language homophone distractors (CH_R and CH_F) were shorter than that with unrelated distractors (UN_R and UN_F). While there was no naming latency difference between rising-tone vs. falling-tone unrelated distractors, there was an average difference of 22ms between the rising and falling-tone cross-language homophone distractors.

	SC-English Bilingual		English Monolingual	
	Mean	*SD*	*Mean*	*SD*
CH_R	797	218	725	179
CH_F	819	234	729	209
UN_R	852	251	763	210
UN_F	852	253	768	206
CH_F – CH_R	22		4	
UN_F – UN_R	0		5	

Table 2: Mean and SD of raw picture naming latency (ms).

Response times were analysed using the generalized linear mixed-effects model (GLMM) with inverse Gaussian distribution [23]. All the statistical analyses were run in R with the package *lme4* [24]. Given that error rates were low in each condition, no further analysis of response accuracy was conducted. A maximum model includes fixed effects of distractor type (CH_R, CH_F, UN_R and UN_F), participant groups (SC-English bilinguals and English monolinguals), the interaction between distractor type and group, by-subject and by-item random intercept, and random slopes for each fixed term were constructed first. Each term was then tested for exclusion. When the model failed to converge, we first increased the number of iterations and then simplified the model by removing correlation parameters in the random structures [25]. The final GLMM consists of fixed effects of distractor type, the interaction between distractor type and group, random intercepts for subject, and random intercepts for item. As there was a significant interaction between groups and distractor type (p = 0.030), pairwise comparisons between group and distractors were also computed using the *multcomp* package [26]. Holm–Bonferroni method was implemented to correct family-wise errors [27].

According to the model estimations (see Table 2), both English monolinguals and SC-English bilinguals took longer to name targets with UN_R and UN_F than CH_R and CH_F respectively (English monolinguals: $p < 0.000$; $p < 0.000$; SC-English bilinguals: $p < 0.000$; $p = 0.002$). These results confirm the facilitatory effects of cross-language phonological overlap. For English monolinguals, there was no significant difference between the distractors UN_R vs. UN_F ($p = 0.903$) and CH_R vs.

CH_F (p = 0.596). For SC-English bilinguals, there was a significant difference between CH_R and CH_F distractors (p = 0.022), but no significant difference between UN_R and UN_F distractors (p = 0.925). This suggests that only SC-English bilinguals, but not English monolinguals, responded to the rising-tone cross-language homophone distractors faster than their falling-tone counterparts.

4. DISCUSSION

Our study investigated the effect of SC lexical tone in bilingual lexical access during English spoken word production. Within the picture word inference paradigm (PWI), both SC-English bilinguals and English monolinguals were asked to name pictures in English while ignoring SC distractors that were either cross-language homophones with the target names or unrelated. Furthermore, we manipulated the lexical tone of the distractors to be either rising or falling. Our naming latency results showed that, regardless of the pitch shape difference, both SC-English bilinguals and English monolinguals took less time to name pictures when the distractors are cross-language homophones, in comparison to the unrelated distractors. Our results thus confirmed the cross-language phonological facilitation effect and showed that our participants, despite their on-line participation, behaved similarly to participants in lab experiments [12].

Furthermore, while both rising and falling-tone cross-homophones were equally facilitative for English monolinguals, SC-English bilinguals took significantly longer time to name pictures with falling-tone cross-language homophone distractors than with their rising-tone counterparts. Such a distinction indicates a significant pitch processing difference between SC-English bilinguals and English monolinguals during English spoken word production. Our results thus echoed with the bilingual comprehension studies [8-9] that lexical tone plays a significant role in non-tonal lexical access. It is, however, important to note that our findings showed an opposite pattern with regard to the direction of pitch processing advantages in bilingual speakers. Ortega-Llebaria and her colleagues found a "falling-f0 bias" [9], [10] in SC-English bilingual speakers' lexical decision, we observed a reversed "rising-f0 bias" in SC-English picture naming. There are a few possible explanations for such a difference.

First, it might be due to different task requirements. The presence of phonologically similar words is known to elicit inhibitory effects in comprehension tasks (e.g., [28]), but facilitatory effects in production tasks (e.g., [29]). This is because, in comprehension tasks, which are phonologically driven, phonologically similar words often introduce speech ambiguity and thus slow down lexical access; while in speech production tasks, which are semantically driven, phonologically similar words often help resolve semantic competition and facilitate the processes of phonological encoding and phonetic spell-out.

Alternatively, the falling vs. rising-bias contrast may reflect cross-language interference effects. Ortega-Llebaria and her colleagues [9], [10] used English intonational falling vs. rising contrasts in the stimuli, but we selected SC cross-language homophones with lexical rising and falling tones as distractors. We know that the phonetic details of a lexical falling tone in SC and an intonational falling tone in English are different. If English words are indeed represented with an intonational falling tone, as proposed by Ortega-Llebaria and her colleagues, SC cross-language homophones with a lexical falling tone may cause more cross-language interference than their rising-tone counterparts, and therefore become less facilitative for English target naming.

A third possibility is related to the acoustic saliency of the pitch contour in the rising-tone distractors, which could have elicited faster responses, and, in turn, facilitated word production. A followup question is why such rising pitch saliency does not promote lexical decision.

The three possibilities may not be mutually exclusive; task requirements of the PWI, robust cross-language interference effect introduced by the SC distractors, and greater acoustic salience of the rising pitch contour may have played an interactive role, resulting in a relatively less facilitative effect of the falling-tone cross-language homophones, in comparison with their rising-tone counterparts. Further research is needed to investigate these possibilities.

5. CONCLUSION

Overall, the SC-English bilingual speakers in our study showed a robust facilitation effect of SC cross-language homophones. Furthermore, the rising lexical tone introduced a greater facilitation effect than a falling lexical tone. No such facilitation effects were observed in monolingual English speakers. Our findings lend novel evidence to the role of lexical tone in bilinguals' non-tonal lexical access during spoken word production, complementary to earlier findings on the impact of lexical tone during spoken word recognition. Further research is needed for a refined and comprehensive account of an interactive bilingual mental lexicon for speech processing.

6. ACKNOWLEDGEMENT

This research was supported by the China Scholarship Council for the State Scholarship Fund (No.201707110012) and the Netherlands Organization for Scientific Research (VI.C.181.040).

6. REFERENCES

[1] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3, no. 1, Dec. 2017, doi: 10.1515/lingvan-2017-0007.

[2] M. Ernestus and Y. Chen, "Message-Related Variation: Segmental Within-Speaker VariationTonal Variation," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds., Oxford University Press, 2011, p. 0. doi: 10.1093/oxfordhb/9780199575039.013.0006.

[3] M. Liu, Y. Chen, and N. O. Schiller, "Context Matters for Tone and Intonation Processing in Mandarin," *Lang Speech*, vol. 65, no. 1, pp. 52–72, Mar. 2022, doi: 10.1177/0023830920986174.

[4] J. F. Kroll and F. Ma, "The bilingual lexicon," in *The handbook of psycholinguistics*, in Blackwell handbooks in linguistics. Wiley-Blackwell, 2018, pp. 294–319.

[5] À. Colomé, "Lexical Activation in Bilinguals' Speech Production: Language-Specific or Language-Independent?," *Journal of Memory and Language*, vol. 45, no. 4, pp. 721–736, Nov. 2001, doi: 10.1006/jmla.2001.2793.

[6] P. Macizo, "Phonological coactivation in the bilinguals' two languages: Evidence from the color naming task*," *Bilingualism: Language and Cognition*, vol. 19, no. 2, pp. 361–375, Mar. 2016, doi: 10.1017/S136672891500005X.

[7] J. Gandour *et al.*, "Temporal integration of speech prosody is shaped by language experience: An fMRI study," *Brain and Language*, vol. 84, no. 3, pp. 318–336, Mar. 2003, doi: 10.1016/S0093-934X(02)00505-9.

[8] B. Chandrasekaran, A. Krishnan, and J. T. Gandour, "Mismatch negativity to pitch contours is influenced by language experience," *Brain Research*, vol. 1128, pp. 148–156, Jan. 2007, doi: 10.1016/j.brainres.2006.10.064.

[9] M. Ortega-Llebaria, M. Nemogá, and N. Presson, "Long-term experience with a tonal language shapes the perception of intonation in English words: How Chinese–English bilinguals perceive 'Rose?' vs. 'Rose,'" *Bilingualism: Language and Cognition*, vol. 20, no. 2, pp. 367–383, Mar. 2017, doi: 10.1017/S1366728915000723.

[10] M. Ortega-Llebaria and Z. Wu, "Chinese-English Speakers' Perception of Pitch in Their Non-Tonal Language: Reinterpreting English as a Tonal-Like Language," *Lang Speech*, p. 0023830919894606, Jan. 2020, doi: 10.1177/0023830919894606.

[11] R. R. Rosinski, R. M. Golinkoff, and K. S. Kukish, "Automatic semantic processing in a picture-word interference task," *Child Development*, pp. 247–253, 1975.

[12] A. Costa, À. Colomé, O. Gómez, and N. Sebastián-Gallés, "Another look at cross-language competition in bilingual speech production: Lexical and phonological factors," *Bilingualism: Language and Cognition*, vol. 6, no. 3, pp. 167–179, Dec. 2003, doi: 10.1017/S1366728903001111.

[13] M. Kaushanskaya, H. K. Blumenfeld, and V. Marian, "The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later," *Bilingualism: Language and Cognition*, vol. 23, no. 5, pp. 945–950, Nov. 2020, doi: 10.1017/S1366728919000038.

[14] K. Antoniou and N. Katsos, "The effect of childhood multilingualism and bilectalism on implicature understanding," *Applied Psycholinguistics*, vol. 38, no. 4, pp. 787–833, Jul. 2017, doi: 10.1017/S014271641600045X.

[15] T. H. Gollan, G. H. Weissberger, E. Runnqvist, R. I. Montoya, and C. M. Cera, "Self-ratings of Spoken Language Dominance: A Multi-Lingual Naming Test (MINT) and Preliminary Norms for Young and Aging Spanish-English Bilinguals," *Biling (Camb Engl)*, vol. 15, no. 3, pp. 594–615, Jul. 2012, doi: 10.1017/S1366728911000332.

[16] Q. Cai and M. Brysbaert, "SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles," *PLOS ONE*, vol. 5, no. 6, p. e10729, Jun. 2010, doi: 10.1371/journal.pone.0010729.

[17] K. D. Neergaard, H. Xu, J. S. German, and C.-R. Huang, "Database of word-level statistics for Mandarin Chinese (DoWLS-MAN)," *Behav Res*, vol. 54, no. 2, pp. 987–1009, Apr. 2022, doi: 10.3758/s13428-021-01620-7.

[18] E. Bates *et al.*, "Timed picture naming in seven languages," *Psychonomic Bulletin & Review*, vol. 10, no. 2, pp. 344–380, Jun. 2003, doi: 10.3758/BF03196494.

[19] M. B. Brodeur *et al.*, "The bank of standardized stimuli (BOSS): comparison between French and English norms," *Behavior Research Methods*, vol. 44, no. 4, pp. 961–970, Dec. 2012, doi: 10.3758/s13428-011-0184-7.

[20] P. Boersma, "Praat: doing phonetics by computer [Computer program]," *http://www. praat. org/*, 2011.

[21] A. E. Milne *et al.*, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, 2020, doi: 10.3758/s13428-020-01514-0.

[22] F. Roux, B. C. Armstrong, and M. Carreiras, "Chronset: An automated tool for detecting speech onset," *Behav Res*, vol. 49, no. 5, pp. 1864–1881, Oct. 2017, doi: 10.3758/s13428-016-0830-1.

[23] S. Lo and S. Andrews, "To transform or not to transform: using generalized linear mixed models to analyse reaction time data," *Frontiers in Psychology*, vol. 6, Aug. 2015, doi: 10.3389/fpsyg.2015.01171.

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *Journal of Statistical Software*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[25] M. Brauer and J. J. Curtin, "Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items," *Psychological Methods*, vol. 23, no. 3, pp. 389–411, 2018, doi: 10.1037/met0000159.

[26] T. Hothorn *et al.*, "Package 'multcomp,'" *Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria*, 2016.

[27] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[28] S. Dufour and R. Peereman, "Inhibitory priming effects in auditory word recognition: when the target's competitors conflict with the prime word," *Cognition*, vol. 88, no. 3, pp. B33–B44, Jul. 2003, doi: 10.1016/S0010-0277(03)00046-5.

[29] A. S. Meyer, "The time course of phonological encoding in language production: Phonological encoding inside a syllable," *Journal of Memory and Language*, vol. 30, no. 1, pp. 69–89, Feb. 1991, doi: 10.1016/0749-596X(91)90011-8.

Acquiring new contrasts in L2 vs. L3: Chinese dialectal speakers' acquisition of Mandarin and English phonetic contrasts

Tianqi Ju, Beiwei Gu, Fumo Huang, Xinyi Liu, Cen Yue, Yao Yao

The Hong Kong Polytechnic University
tianqi.ju@connect.polyu.hk, beiwei.gu@connect.polyu.hk, flora.huang@connect.polyu.hk,
xinyisein.liu@connect.polyu.hk, cen-mia.yue@connect.polyu.hk, ctyaoyao@polyu.edu.hk

ABSTRACT

A central question in L3 phonology is whether the acquisition of L3 sounds is more influenced by the sound systems of L1, L2, or a combination of the two. The current study investigates two groups of Chinese dialectal speakers' acquisition of new phonemic contrasts in Mandarin (L2) and English (L3). Both dialectal groups (Northeastern, Southwestern) have a neutralized /s/-/sh/ contrast in their native dialects and have to acquire the contrast for L2 Mandarin (/s/-/ʂ/) and L3 English (/s/-/ʃ/). The Southwestern speakers' native dialects also lack the /n/-/l/ distinction, which is present in both L2 and L3. Although an L2-to-L3 transfer would be facilitative in all three cases, we only found definitive evidence for L2-to-L3 transfer in the production of the /s/-/sh/ contrast by Southwestern speakers, mainly among the male speakers. We discuss the results of selective L2-to-L3 transfer with the L2 Status Factor Model and the Scalpel Model.

Keywords: L3 phonology, Chinese dialects, phonetic transfer, L2 Status Factor Model, Scalpel Model.

1. INTRODUCTION

When a learner sets out to acquire a third language (L3), the acquisition could be influenced by previously acquired first (L1) and/or second (L2) languages. Various proposals have been made to address which language system(s) would serve as the primary source of linguistic transfer to the L3: the language typologically closer to the L3 (the Typological Primacy Model (TPM); [1]–[4]), the later acquired language (the L2 Status Factor Model; [5]), the language that would lead to facilitative transfer (the Cumulative-Enhancement Model (CEM); [6]), the language with a similar linguistic property (the Linguistic Proximity Model (LPM); [7]), or the language selected by a number of cognitive and experiential factors (the Scalpel Model; [8]). While both the TPM and the L2 Status Factor Model assume wholesale transfer from one of the existing languages to the L3, the latter three models all posit property-by-property transfer, where different linguistic properties can be transferred from different language systems. The LPM and the Scalpel Model also allow the possibility of non-facilitative transfer, while CEM considers facilitation a necessary condition.

In this study, we examine two groups of Chinese dialectal speakers, who speak a regional Chinese dialect as L1, Standard Mandarin (or "Mandarin" in short) as L2 and English as L3. The two groups differ in the L1 dialect: one group speak Northeastern Chinese dialects, and the other group speak Southwestern dialects. The dialects spoken in Northeastern China (Heilongjiang, Jilin, and Liaoning provinces), which belong to the Mandarin dialectal family, are highly similar to Standard Mandarin in pronunciation except for certain consonants and tones ([9]). Anecdotal evidence suggests that many Northeastern speakers are not aware of the phonetic differences between their native dialect and Standard Mandarin and consider themselves native speakers of Standard Mandarin. Southwestern dialects, on the other hand, are a much more mixed category. The Southwestern speakers in this study mainly come from Sichuan, Hunan, and Hubei provinces, where the regional dialects include both Southwest Mandarin (a member of the Mandarin dialectal family) and non-Mandarin dialects (e.g., Xiang dialects). The Southwestern dialects are overall farther from Standard Mandarin—in pronunciation, lexicon, syntax, *etc.*—than the Northeastern dialects are.

We focus on the acquisition of phonetic contrasts present in both L2 (Mandarin) and L3 (English) but not in L1. Specifically, we examine the acquisition of the /s/-/sh/ contrast by both dialectal groups and the acquisition of /n/-/l/ by the Southwestern group. Regarding the /s/-/sh/ contrast, Standard Mandarin has a phonemic contrast between alveolar and retroflex fricatives/affricates (i.e., /s, ts, tsʰ/ vs. /ʂ, tʂ, tʂʰ/), and English has a similar contrast between alveolar and postalveolar voiceless fricatives (i.e., /s/ vs. /ʃ/). But the contrast is to a large degree neutralized in both Northeastern and Southwestern dialects. Some Northeastern dialects lack the retroflex series altogether, while others may have both series but use them in free variation. Southwestern dialects in general lack the retroflex series. As a result, both Northeastern and Southwestern speakers are known to have difficulty with the /s/-/sh/ contrast when speaking Standard Mandarin and English ([10]–[12]).

In addition, the Southwestern dialects also lack the /n/-/l/ contrast, which is present in both Mandarin and English, leading to difficulty in L2/3 acquisition ([13]).

Since the contrasts under investigation are present in both L2 and L3 but not in L1, influence from L2 would facilitate the acquisition of the contrast in L3, while influence from L1 would not. The potential facilitative transfer from L2 may not necessarily result in higher production accuracy in L3 (compared to L2), because L3 is usually lower in proficiency. However, if the speaker does perform better in L3 than in L2, that would provide definitive evidence for the facilitative transfer from L2 to L3, which is not available during the acquisition of L2. If the direction of transfer is conditioned by typological distance, we expect to see similar results between the two speaker groups, given the similarity of their language backgrounds. If transfer happens in a wholesale manner, results regarding the two phonetic contrasts in Southwestern speakers should be similar.

2. METHODS

2.1. Subjects

Two groups of Chinese speakers participated in this study: 25 from Northeastern China (15F, 10M; 18-30 y.o.) and 28 from Southwestern China (16F, 12M; 18-30 y.o.). All the participants spent at least 15 years before the age of 18 in their home province and had continuous exposure to the regional dialect both at home and in the public when growing up. All the participants identified the regional dialect as their L1. The Mandarin exposure started before the age of six for all participants; English was acquired as a foreign language later in primary and secondary schools. At the time of the study, all the participants had recently (within the past two years) relocated from Mainland China to Hong Kong to study in a university. None of the participants had studied linguistics before; none reported any history of language or speech disorders.

2.2. Stimuli

The critical stimuli include 16 monosyllabic words for each target phoneme (/s/, /sh/, /n/, /l/) in each language (Mandarin, English), totalling 128 critical items. The critical phoneme always appears in the onset. A wide range of vowel contexts (e.g., /a/, /i/, /u/, /ɔ/, /ə/, /ɤ/, /aj/, /aw/, /ow/) loosely shared between Mandarin and English were used to increase variability and facilitate cross-language comparison. One might notice that when followed by a high front vowel like [i], the English /ʃ/ is likely to be assimilated to the alveolo-palatal /ɕ/ in Mandarin, which is acoustically more similar to /s/ than to /ʃ/ ([14]). The current word list has two such English words (*ship*, *sheep*); results reported in the following remain unchanged when tokens of these two words are excluded from analysis. Words with multiple pronunciations and (English) words that were unfamiliar to Chinese learners were avoided.

The 128 critical stimuli contain 26 minimal pairs in the critical phoneme (5-7 pairs per contrast per language); overall, more than 70% of the critical words had a minimal pair neighbour in the critical position in the language. Another 64 monosyllabic words (32 Mandarin, 32 English), which do not contain any of the critical phonemes, were used as fillers.

2.3. Procedure

The experiment was a word reading task conducted with DMDX ([15]) in a soundproof phonetic lab. Mandarin and English word stimuli were presented in separate blocks. In the Mandarin block, the participant saw both the Chinese character and the pinyin form of the target word, in order to minimize confusion about the citation form due to imperfect lexical knowledge of Mandarin; in the English block, only the spelling form was provided. In each trial, the participant was asked to read the word on the screen in a carrier phrase ([wɔ3 kʰan4 taw4] "*I saw* __]" for Mandarin and "*I read* __" for English). The stimuli within a block were presented in a randomized order, with each item occurring twice; the order of language blocks was balanced across participants, with a 5-minute break between the blocks. After completing both blocks, the participant would complete a survey about their demographic and language backgrounds. A complete experimental session took about 40 minutes.

A total of 6400 (64 words × 2 repetitions × 50 speakers) [n/l] tokens and 6400 [s/sh] tokens were collected. Two of the authors independently listened to all the [n/l] tokens in a randomized order and annotated each as [n] or [l] without accessing any information about the speaker or the target item. A small number of the [n]/[l] tokens may sound ambiguous; in cases of annotation discrepancy (N=118), additional judgement was sought from a third author. The [s/sh] tokens were first annotated manually for the critical fricative interval in Praat ([16]); a Praat script adapted from [17] was then used to automatically measure duration and spectral mean (i.e., spectral center of gravity) of each annotated interval. After excluding tokens of misidentified targets or corrupted recordings, 6386 tokens remained in the n/l dataset and 6361 in the s/sh dataset.

2.4. Statistical analysis

The production data are analyzed separately by phonetic contrast (/s/-/sh/, /n/-/l/) and dialectal group (Northeastern, Southwestern). Mixed-effects models are built to examine the production of each contrast by each speaker group: in each model, fixed-effects include target phoneme, language, speaker sex, and their interactions, and random effects include by-speaker and by-item intercepts. The outcome variables are production accuracy (both contrasts; generalized mixed-effects models) and acoustic measures (log duration and spectral mean (converted to ERB scale); the /s/-/sh/ contrast only; linear mixed-effects models). Production accuracy of [s/sh] tokens is coded using a raw spectral mean of 5000 Hz as the boundary between [s] and [sh], following the norms of Mandarin and English sibilants ([14], [18]). All the models are built with the lme4() package ([19]) in R ([20]). Reporting in the next section focuses on language-related effects.

3. RESULTS

3.1. /s/-/sh/ contrast

Both dialectal groups have significantly less accurate productions of [sh] than [s] (*p*s <. 001; see Fig 1). Northeastern speakers tend to produce the /sh/ sound better in Mandarin (mean accuracy: 76.5% (F); 88.1% (M)) than in English (69.7% (F); 87.3% (M)), but the trend does not reach significance (p >.05). On the contrary, Southwestern speakers have overall more accurate [sh] productions in English (74.6% (F); 72.4% (M)) than in Mandarin (67.4% (F); 58.6% (M)), but the trend does not reach significance (p >.05).

The models on (log) sibilant duration reveal that both dialectal groups significantly lengthen the English sibilants compared to the Mandarin counterparts (*p*s < .01), suggesting greater effort in reading L3 English words. As predicted, Northeastern speakers' model of spectral means shows a significant lowering of spectral means in [sh] compared to [s] ($\beta_{target=sh}$ = –4.73; p < .001), and the difference is greater for male speakers than female speakers ($\beta_{target=sh}$ = –0.47; p = .002). The /s/-/sh/ difference does not interact with language.

The model of Southwestern speakers' spectral means also shows significantly lower spectral means for [sh] than for [s]. However, the size of the /s/-/sh/ difference interacts with both speaker sex and language. Female Southwestern speakers are predicted to have a difference of around 5.02 ERB in spectral mean between their [s] and [sh] of both languages ($\beta_{target=sh}$ = –5.02; p < .001). Male Southwestern speakers have significantly smaller /s/-

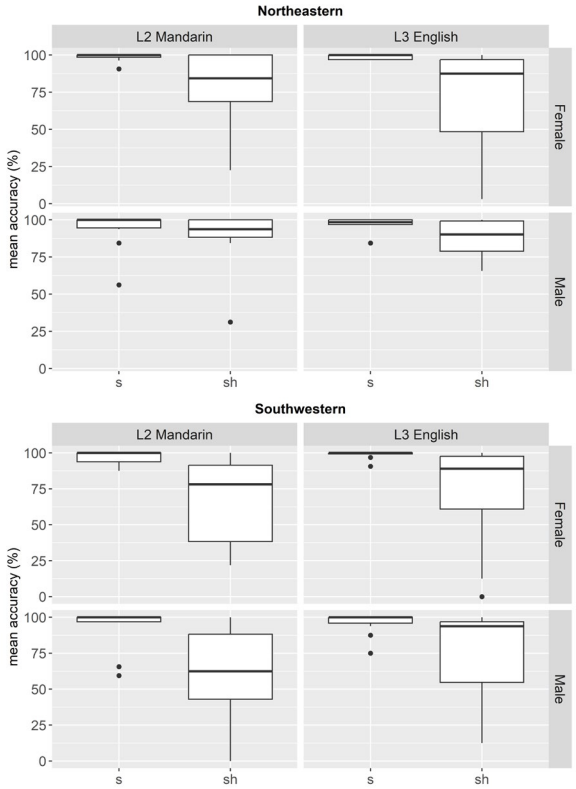


Figure 1: Production accuracy of Northeastern (top) and Southwestern (bottom) speakers' productions of the /s/-/sh/ contrast, separated by language and sex.

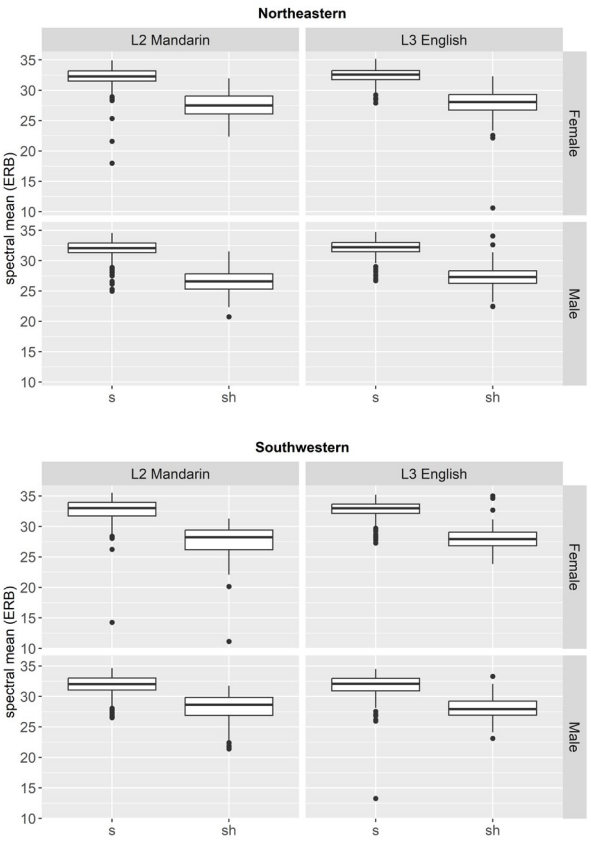


Figure 2: Average spectral means of Northeastern (top) and Southwestern (bottom) speakers' productions of the /s/-/sh/ contrast, separated by language and sex.

/sh/ difference in Mandarin ($\beta_{target=sh:sex=male}$ = 1.41; p<.001) compared to female Southwestern speakers, but they show a marginally significant trend to increase the /s/-/sh/ difference in English ($\beta_{target=sh:sex=Male:language=ENG}$ = –0.35; p = .06). Of all four region×sex groups, male Southwestern speakers have the smallest /s/-/sh/ distinction (in terms of spectral means) in Mandarin, which is consistent with their low production accuracy with Mandarin [sh] sound. Meanwhile, male Southwestern speakers are also the only group that show a significantly greater spectral distinction between [s] and [sh] in English than in Mandarin.

3.2. /n/-/l/ contrast

Northeastern speakers, whose native dialects have the /n/-/l/ contrast, produce a clear /n/-/l/ distinction in both languages (mean accuracy > 98%). Southwestern speakers, on the other hand, have more difficulty with the contrast, especially the [l] sound (Fig 3). Both male and female Southwesterners tend to produce the [l] sound less accurately in English (F: 96.3%; M: 82.6%) than in Mandarin (F: 98.2%; M: 89.9%), although the cross-language difference is not significant (p > .05).

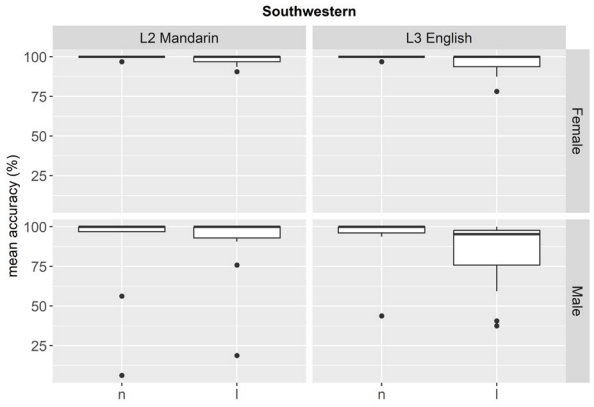


Figure 3: Production accuracy of Southwestern speakers' production of the /n/-/l/ contrast, separated by language and sex.

Taken together, our results show that male Southwestern speakers show a robust pattern of pronouncing the /s/-/sh/ contrast with a greater spectral distinction and higher accuracy in L3 English than in L2 Mandarin. Female Southwestern speakers also tend to pronounce the /s/-/sh/ contrast with higher accuracy in L3 English than in L2 Mandarin, but with no significant spectral enhancement. However, this L3-better-than-L2 pattern is not observed in the other cases examined in this study. Northeastern speakers' production of the /s/-/sh/ contrast shows no evidence of enhancement in L3 over L2, despite the greater articulatory effort as shown in the hyperarticulation of L3 English words. Southwestern speakers' own production of the /n/-/l/ distinction also lacks the L3-better-than-L2 pattern; if anything, the speakers tend to have lower—instead of higher—accuracy of [n/l] tokens in L3 English than in L2 Mandarin.

4. DISCUSSION

As stated above, better performance in L3 than in L2 would provide definitive evidence for facilitative transfer from L2 to L3 in this study. Current results only show facilitative transfer from L2 to L3 in Southwestern Chinese speakers' acquisition of the /s/-/sh/ contrast. There are two possible reasons why Northeastern speakers did not show the L3-better-than-L2 pattern. First, Southwestern speakers are more likely to recognize Mandarin as an L2, whereas Northeastern speakers may not consider Mandarin as distinctly different from their L1, given the close resemblance between the two. Following the L2 Status Factor Model, Southwestern speakers are thus more likely to transfer from Mandarin to English than Northeastern speakers. Second, maybe the Northeastern speakers in this study happen to have lower English proficiency, which would cancel out the facilitation from the transfer from Mandarin. However, male Northeastern speakers actually had much higher /s/-/sh/ production accuracy (>87%) than Southwesterners (52.5-77.5%) in both languages, which undermines the account based on L3 proficiency.

Another question regards why we didn't see the L3-better-than-L2 pattern in Southwestern speakers' production of the /n/-/l/ contrast. Obviously, this cannot be explained by L2 status, L3 proficiency or typological proximity between languages. We think that the answer is probably related to some linguistic, cognitive, or experiential factors, following the Scalpel Model. The discrepancy between the /s/-/sh/ contrast and the /n/-/l/ contrast in Southwestern speakers' data may be explained by articulatory or perceptual features of the two contrasts or the speaker's awareness about these contrasts in the L2 and L3.

To summarize, our results showcase the selectivity of facilitative transfer from L2 to L3 in L3 phonology. Our findings are most compatible with the L2 Status Factor Model and the Scalpel Model. We hope to further investigate the effects of cognitive and experiential factors in L3 acquisition in future research.

7. REFERENCES

[1] Rothman. J. 2011. L3 syntactic transfer selectivity and typological determinacy: The typological

primacy model. *Second Lang. Res.*, 27 (1), 107–127.
[2] Rothman, J. 2010. On the typological economy of syntactic transfer: Word order and relative clause high/low attachment preference in L3 Brazilian Portuguese. *IRAL - Int. Rev. Appl. Linguist. Lang. Teach.*, 48 (2–3), 245–273.
[3] Rothman, J. 2013. Cognitive economy, non-redundancy and typological primacy in L3 acquisition: Initial stages of L3 Romance and beyond. In: S. Baauw, F. Drijkoningen, L. Meroni, and M. Pinto (eds) *Romance Languages and Linguistic Theory 2011. Selected papers from 'Going Romance' Utrecht 2011*. John Benjamins, 217–247.
[4] Rothman, J. 2015. Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered. *Biling. Lang. Cogn.*, 18 (2), 179–190.
[5] Bardel, C. and Falk, Y. 2007. The role of the second language in third language acquisition: The case of Germanic syntax. *Second Lang. Res.*, 23 (4), 459–484.
[6] Flynn, S., Foley, C., Vinnitskaya, I. 2004. The Cumulative-Enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *Int. J. Multiling.*, 1 (1), 3–16.
[7] Westergaard, M, Mitrofanova, N., Mykhaylyk, R. Rodina, Y. 2017. Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model. *Int. J. Biling.*, 21 (6), 666–682.
[8] Slabakova, R. 2017. The scalpel model of third language acquisition. *Int. J. Biling.*, 21 (6), 651–665.
[9] Liu, B. 2009. *Shenyang fangyan de yuyin fenxi* [Phonetic analysis of Shenyang dialect]. *Wenjiao Ziliao*, 2009.April, 54–55.
[10] Gong, C, Tian, Y. 2008. *Dongbei fangyan dui yinghan yuyin xide chansheng de yingxiang* [Influence of Dongbei dialect on the acquisition of English and Mandarin pronunciations]. *Jiaoyan Tansuo*, 10, 30–31.
[11] Zhang, L. 2018. An acoustic study on the English fricative production by Chinese Northeastern EFL learners. Thesis.
[12] You, S., Kan, P., Wu, J.-Y., Yang, Z.-Y. 2009. On the language transfer of Chinese dialects to English pronunciation: Comparison of Jinan, Chengdu and Pingxiang Dialects. *J. Gannan Norm. Univ.*, 2009.5, 54–58.
[13] Zhang, W., Levis, J. M. 2021. The Southwestern Mandarin /n/-/l/ Merger: Effects on Production in Standard Mandarin and English. *Front. Commun.*, 6 (August), 1–22.
[14] Chang, C. B., Yao, Y., Haynes, E. F., Rhodes, R. 2011. Production of phonetic and phonological contrast by heritage speakers of Mandarin. *J. Acoust. Soc. Am.*, 129 (6), 3964–3980.
[15] Forster, K., Forster, J. C. 2003. DMDX: A Windows display program with millisecond accuracy. *Behav. Res. Methods, Instruments, Comput.*, 35 (1), 116–124.
[16] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer.
[17] DiCanio, C. 2021. Spectral moments of fricative spectra script in Praat.
[18] Jongman, A., Wayland, R., Wong. S. 2000. Acoustic characteristics of English fricatives. *J. Acousitical Soc. Am.*, 108 (3), 1252–1263.
[19] Bates, D. M., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 67, 1–48.
[20] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Crosslinguistic influence on perceptual cue-weighting in Korean-English bilinguals

Yuhyeon Seo, Olga Dmitrieva

Purdue University
seo86@purdue.edu, odmitrie@purdue.edu

ABSTRACT

The current study investigates L2 crosslinguistic influence on the perceptual weighting of the cues to L1 phonological categories. Two groups of Korean-English bilinguals participated in the study: (1) long-term immigrants in the US (LTIs, n = 25) and (2) Korean heritage speakers (HSs, n = 25). Participants completed a three-alternative forced-choice identification task using a set of resynthesized stimuli containing Korean stops varying in two acoustic parameters: onset f0 and voice onset time (VOT). The results of mixed-effects logistic regression models suggested that HSs' L1 perception was influenced by their L2, as indicated by a lesser reliance on onset f0 as a cue to some Korean laryngeal contrasts, and the overall 'regularization' of the use of both cues, compared to LTIs. HSs' perception of Korean stop contrasts was also less categorical than that of LTIs.

Keywords: bilingual speech, cue-weighting, crosslinguistic influence, phonetic drift, heritage speakers

1. INTRODUCTION

Crosslinguistic influence in bilingual speech is known to be bidirectional – the first language (L1) and the second language (L2) mutually affect each other. A substantial body of research documented crosslinguistic influence in bilingual speech production [1]–[5]. Moreover, recent work suggested that bilingual speech perception is also subject to crosslinguistic influence [6]–[8]. The present study continues this line of inquiry by investigating crosslinguistic influence in Korean–English bilinguals' perception of the two acoustic cues to Korean laryngeal contrasts in stop consonants. Two groups of bilingual speakers participated in the study: (1) a group of heritage speakers of Korean (HS) and (2) a group of long-term immigrants (LTI) to the United States.

LTIs were operationalized as bilinguals who had immigrated to an L2-speaking country at the age of 18 or later and had resided there for at least 12 years. While few investigations of L1 perception by LTIs have been conducted to date, existing research suggests that L1 perceptual strategies can be affected by an L2. For instance, [6] found that in the perception of word-final stop voicing in Russian, Russian LTIs in the US relied on a cue-weighting pattern similar to that of L1 American English speakers and distinct from that of L1-immersed Russian speakers, suggesting L2 influence on L1 perception.

HSs were operationalized as bilinguals who speak their parents' language (L1) at home, while speaking an L2 elsewhere in which they are more proficient and dominant than in the L1 [9]. One of the defining characteristics of HSs' linguistic development is the dominance reversal they typically undergo during early childhood, when an L2 starts overtaking their L1 in the amount of use and exposure. The National Heritage Language Survey [10] reports that almost 99% of Korean HSs in the US become English (L2) dominant by the age of 17.

This dominance pattern combined with early and substantive L2 exposure suggests that L2 influence on HSs' L1 perceptual strategies should be even more evident than for LTIs. Nevertheless, previous studies suggested that L1 phonetics of HSs could be immune to crosslinguistic influence from the L2, especially in the perceptual domain [1], [8], [11]. For instance, studies on the perceptual discrimination of Korean stops by HSs showed accuracy comparable to that of L1-immersed native speakers, while outperforming L2 learners [8], [11]. However, we hypothesized that the methodological choices were partially responsible for these findings. Specifically, we posited that discrimination tasks are not sensitive enough to detect L2 influence on L1 perception, as they do not illuminate the pathway to discrimination decisions.

In the present study, we adopt a cue-weighting approach to investigate the degree of an L2 crosslinguistic influence on L1 speech perception in HSs compared to LTIs. While both groups had a significant amount of exposure to the L2, they differed in the timing of L2 exposure (early childhood for HSs vs. early adulthood for LTIs), as well as the quality and quantity of L1 exposure during childhood. HSs acquired Korean from birth but a large share of their linguistic exposure and use began to be dedicated to the L2 already in early childhood, eventually leading to dominance reversal. LTIs, who came to the US in early adulthood, were exclusively Korean-dominant for a considerably longer period of their life. Therefore, the current study pursues the

hypothesis that HSs' L1 perception should be affected by the L2 more than LTI's L1 perception.

The present study focuses on the two acoustic properties, onset f0 and voice onset time (VOT), as acoustic cues to laryngeal contrasts in both Korean and English stop. As shown in Table 1, the Korean language exhibits a three-way distinction in stop consonants (lenis-aspirated-fortis, /p/-/p^{h}/-/p*/) [12], while English has a two-way voicing contrast in stops (/p/-/b/) [13]. In the word-initial position, both Korean and English stops are distinguished by the same two acoustic parameters, but the two languages differ in their relative weighting of these cues. While VOT plays a primary role in the voicing contrast in English stops, onset f0 is the primary cue to the lenis-aspirated contrast in modern Seoul Korean [12], [14]–[16].

Therefore, we hypothesized that L2 influence on the L1 will manifest itself as a greater reliance on VOT and a lesser reliance on onset f0 in the identification of Korean stops, given the difference in the functional load of these cues in English and Korean laryngeal phonology. Therefore, HSs are predicted to rely on VOT to a greater extent and on onset f0 to a lesser extent compared to LTIs.

Korean	VOT	Onset f0
Lenis	Long-lag	Low
Aspirated	Long-lag	High
Fortis	Short-lag	High
English		
Voiceless	Long-lag	High
Voiced	Short-lag	Low

Table 1: Stop contrasts in Korean and English.

2. METHODS

2.1. Participants

A total of 50 Korean-English bilinguals were recruited in the Midwestern areas of the US. The LTI group consisted of 25 Korean immigrants who moved to the US from South Korea at the age of 18 or later and were, on average, 39 years old at the time of the participation (SD = 6.2). They had resided in the US for an average of 12 years by the time of the experiment. The HS group consisted of 25 speakers who were born and raised in the US by Korean parents on both sides. HSs were, on average, 27.3 years old at the time of the experiment (SD = 6.7). Care was taken to only enroll participants whose parents were born in Korea after 1950 since the sound change that shifted the burden of differentiating lenis and aspirated stop from VOT to onset f0 is believed to have begun by 1940 [12], [16]. Participants completed a Bilingual Language Profile [17] to verify their language dominance and proficiency. The self-reports of Korean and English proficiency indicated that LTIs were more balanced bilinguals than HSs in speaking, listening, reading, and writing. Most HSs were more proficient and dominant in English than in Korean. HSs showed greater variability in terms of Korean use in daily life compared to LTIs, who used Korean more frequently than HSs. HSs also indicated a significantly less formal education of Korean than LTIs ($t(38.7) = 8.35$, $p < .001$) (HS: M = 3.3 years, SD = 3.5; LTI: M = 14.8, SD = 5.9).

2.2. Stimuli

To implement a three-alternative forced-choice (3AFC) identification task, 64 unique stimuli were created by resynthesizing a real-word minimal pair, /pul/, 'fire' and /p^{h}ul/, 'grass' produced by a male Korean speaker in his twenties, using Praat [18]. The initial stops in the stimuli differed orthogonally in terms of VOT and onset f0, based on an eight-by-eight continuum with log-scaled VOT steps (10, 14, 19, 27, 38, 53, 74, and 104 ms) and equidistant f0 steps in 10 Hz steps ranging from 120 to 190 Hz. The f0 values were resynthesized to be consistent throughout the vocalic part in each stimulus.

2.3. Procedures

All participants completed a 3AFC identification task, using PsychoPy [19] in either a speech lab or at a quiet place of their preference. Participants were wearing headphones throughout the task. In this task, a single stimulus was played on each trial, and participants identified it by clicking with a mouse on one of the three response options on a monitor screen: /pul/, 'fire,' /p^{h}ul/, 'grass,' and /p*ul/, 'horn'. Trials were self-timed. The following trial was presented only after a response was provided in the preceding trial. The response options were presented in Korean orthography, and the order of the response options was counterbalanced across trials. The full set of stimuli was randomized and repeated six times to each participant. Each participant took approximately 15 minutes to complete a total of 384 trials.

2.4. Analysis

Participants' responses were paired by stop contrast, and the resulting three pairs (lenis-aspirated, fortis-aspirated, and lenis-fortis, and) were submitted as binary dependent variables to three separate mixed-effects logistic regression analyses implemented in R (version 4.2.1) [20] using the 'lme4' package [21]. Each model was sum-coded and included Group, VOT, Onset f0, and interactions of Group with VOT and with Onset f0 as fixed effects. The models also

included by-subject intercepts and slopes as random effects to account for individual sensitivity to each acoustic cue.

3. RESULTS

3.1. Lenis-aspirated contrast

Lenis and aspirated stops in modern Seoul Korean are believed to be differentiated primarily on the basis of onset f0 values, and, to a lesser extent, VOT [12], [14]–[16]. The statistical results indicated that while both cues were significant predictors of the identification response (onset f0; $\beta = 1.20$, SE =.06, $z = 19.23$, $p < .001$; VOT; $\beta = .99$, SE =.05, $z = 19.35$, $p < .001$), onset f0 was a primary cue, as indicated by a greater coefficient than for VOT, which played a secondary role for both groups. However, the interaction between Group and Onset f0 indicated that the two groups differed in terms of their reliance on onset f0 ($\beta = -.28$, SE =.06, $z = -4.48$, $p < .001$). Specifically, HSs assigned lower weight to the vocalic cue of onset f0 contrast than LTIs. This effect is illustrated in Figure 1 (left panel), where the identification slope of HSs is visibly less steep than that of LTIs. On the other hand, the absence of an interaction between Group and VOT indicated that participant groups were not significantly different from each other in their use of VOT ($p = .84$).

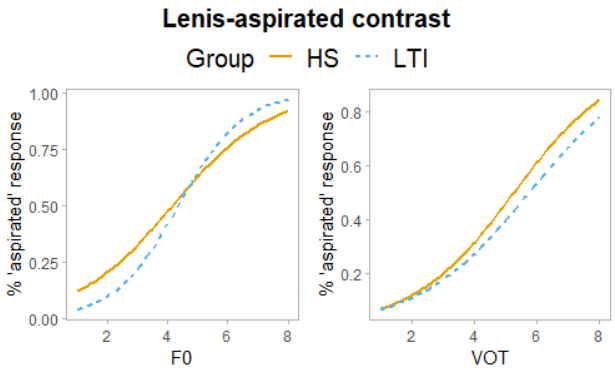


Figure 1: Identification curves for the lenis-aspirated stop contrast by group and by cue.

3.2. Lenis-fortis contrast

Lenis and fortis stops in Korean are contrasted via both acoustic correlates, onset f0 and VOT; lenis stops are characterized by longer VOT and lower onset f0 values than fortis stops [15]. The logistic regression analysis confirmed that both VOT and onset f0 were significant predictors of lenis vs. fortis responses (onset f0; $\beta = .91$, SE =.06, $z = 16.07$, $p < .001$; VOT; $\beta = -.67$, SE =.05, $z = -13.5$, $p < .001$). In addition, both interaction terms (Group by VOT and Group by Onset f0) were significant. These interactions were due to the fact that HSs assigned lower weight to onset f0 than LTIs ($\beta = -.27$, SE =.06, $z = -4.88$, $p < .001$), while assigning greater weight to VOT than LTIs ($\beta = .17$, SE =.05, $z = 3.51$, $p < .001$). This means that the two participant groups differed in whether they used VOT or onset f0 as the primary cue to the lenis-fortis contrast. These differences are illustrated in Figure 2. For example, LTIs have a visibly steeper identification curve as function of onset f0 (left panel).

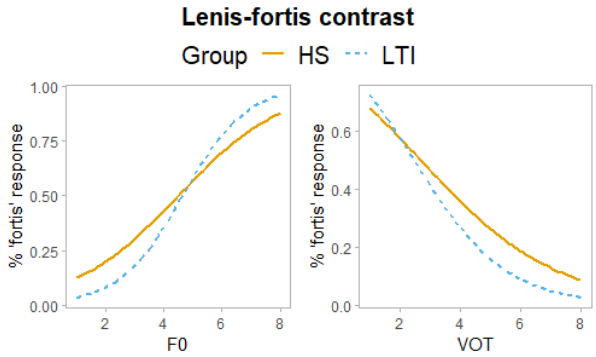


Figure 2: Identification curves for the lenis-fortis stop contrast by group and by cue.

3.3. Fortis-aspirated contrast

Fortis and aspirated stops in Korean are distinguished mainly via VOT, but not onset f0 [16]. Nevertheless, the statistical analysis demonstrated that both VOT and onset f0 were significant predictors of the response (onset f0; $\beta = .24$, SE =.04, $z = 6.19$, $p < .001$; VOT; $\beta = 1.67$, SE =.08, $z = 19.99$, $p < .001$), with VOT being the primary cue as expected. Additionally, a significant interaction of group and onset f0 ($\beta = .12$, SE =.04, $z = 3.21$, $p = .001$) showed that HSs relied on onset f0 more (by a factor of 0.25) than LTIs. In fact, as seen in Figure 3 (left panel), LTIs did not use f0 to any appreciative degree. While the HSs' identification curve went up as f0 increased, LTIs function remained almost level. In contrast, the significant interaction between Group and VOT was due to a greater reliance of LTIs than HSs on VOT.

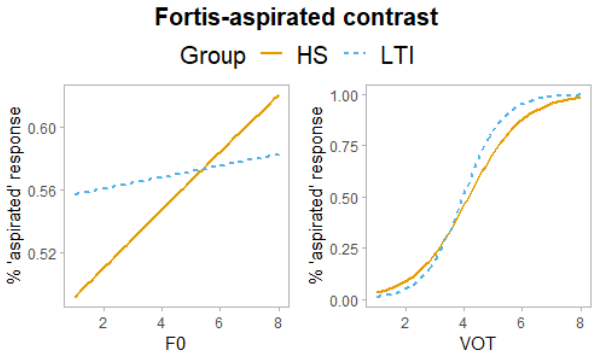


Figure 3: Identification curves for the fortis-aspirated stop contrast by group and by cue.

4. DISCUSSION

We hypothesized that L2 influence would be imposed on HSs' L1 perception to a greater degree than on LTIs' L1 perception, due to the differences between the two groups in terms of the timing, quantity, and quality of exposure to both L1 and L2. Specifically, we predicted that HSs would be less reliant on onset f0 and more reliant on VOT in identifying Korean stop categories, compared to LTIs. This hypothesis was based on previous studies demonstrating that onset f0 plays a secondary role in cueing laryngeal stop categories in English unlike in Korean [7].

The results partially supported these hypotheses. First and foremost, HSs presented a markedly different cue-weighting pattern in identifying the three Korean stop categories, as compared to LTIs. Second, as predicted, in two out of three binary contrasts, HSs relied on onset f0 significantly less than LTIs in making their perceptual decisions.

Specifically, the lenis-aspirated contrast was perceived primarily on the basis of onset f0 by both groups, indicating a presence of tonal contrast for both bilingual groups. Nevertheless, HSs relied on onset f0 significantly less than LTIs.

The lenis-fortis contrast is cued by both parameters in Korean stops [15]. The results confirmed this cue-assignment in both participant groups to a significant degree in identifying stops. Nevertheless, HSs used onset f0 significantly less than LTIs.

These findings are compatible with the prediction that lower functional load of onset f0 in English laryngeal phonology would lead to a lesser reliance of HSs on f0 when identifying Korean stops. Nevertheless, an alternative explanation should be acknowledged. HSs' native language input differed from that of the LTIs. Critical to our investigation is the possibility that participants' parents did not participate in the sound change that increased reliance on onset f0 as a cue to laryngeal categories, or that they spoke a dialect of Korean that did not undergo this sound change [16]. Moreover, the parents' Korean speech could already be influenced by English at the time of its transmission to our participants. While we cannot rule out the latter, some of the results discussed below speak against the sound change non-participation hypothesis.

Regarding the use of VOT, contrary to the expectation, there was little evidence that HSs were more reliant on VOT as a cue to laryngeal categories than LTIs, despite the central role that VOT plays in cuing English voicing. Thus, the overall pattern suggests HSs rely less on onset f0 without increasing their reliance on VOT, as opposed to LTIs.

A puzzling case is presented by the fortis-aspirated contrast, normally cued exclusively by VOT in Korean [16]. LTIs behaved accordingly, relying exclusively on VOT in making the fortis-lenis decisions. Surprisingly, HSs made use of onset f0, in contrast to LTIs. While this finding continues the pattern of a distinct perceptual behavior by HSs compared to LTIs, it does not align with the hypothesis that influence from English will lead to pervasive underutilization of onset f0 in L1 Korean. It also does not support the hypothesis that parents' non-participation in the relevant sound change leads to underuse of f0. Instead, HSs 'regularized' their cue-weighting approach to all Korean laryngeal contrasts, always relying to some degree on both cues, VOT and onset f0. To a certain extent, this strategy could also be viewed as English-based. Perceptual studies suggest that, in English, both VOT and onset f0 are often attended to by native listeners when making voicing decisions, albeit to a different degree, [7], [13], unlike in Korean, where, depending on the contrast, listeners are supposed to attend to one cue while largely ignoring the other one [14].

Finally, it should also be noted that, in most cases and independently of the cue, HSs' identification curves were shallower than those of LTIs, suggesting a less categorical perception overall. This observation is also in accordance with the hypothesis that HSs' early exposure and greater dominance in their L2, English, has consequences for their L1 perception. In this case, the consequence is the decreased overall sensitivity to Korean laryngeal contrasts.

Therefore, the results presented here provide evidence for the hypothesis that bilinguals' L2 can affect their L1 perception and that earlier exposure and greater dominance in L2 lead to stronger crosslinguistic influence on L1 perception.

This study is among the first few to uncover the effects of bilingualism on the perception of native speech [1], [4], [6]. The results confirm that HSs' L1 perception is not immune to the influence of L2, similarly to other aspects of their native language competence [22]. Finally, the present findings fit well within theories of L2 acquisition, such as SLM-r [23], which postulates that L1 and L2 sound categories co-exist in the same phonetic space and, therefore, are predicted to affect each other in a bidirectional manner. While in the past such predictions were made and tested almost exclusively on the basis of speech production data, the recent revision of SLM (SLM-r) proposed that production and perception co-evolve during L2 acquisition. The present findings lend support to this theoretical assumption.

5. ACKNOWLEDGEMENT

We are very grateful to Dr. Jennifer Cabrelli (University of Illinois at Chicago) for invaluable assistance with data collection in Chicago, Illinois. We also thank the anonymous reviewers for constructive suggestions.

6. REFERENCES

[1] Chang, C. B. 2016. Bilingual perceptual benefits of experience with a heritage language, *Bili.: Lang. and Cogn* 19, 791–809.

[2] Chang, S.-E., Weiss-Cowie, S. 2021. Hyper-articulation effects in Korean glides by heritage language learners, *International. J. of Bilingualism.* 25, 3–20.

[3] Flege, E. 1987. The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification, *J. of Phon.* 15, 47–65.

[4] Baker W., Trofimovich, P. 2005. Interaction of native-and second-language vowel system (s) in early and late bilinguals, *Lang. Speech.* 48, 1–27.

[5] Cheng, A. 2017. VOT merger and f0 contrast in Heritage Korean in California, *UC Berkeley PhonLab Annu. Rep.* 13, 1.

[6] Dmitrieva, O. 2019. Transferring perceptual cue-weighting from second language into first language: Cues to voicing in Russian speakers of English, *J. of Phon.* 73, 128–143.

[7] Dmitrieva, O., Llanos, F., Shultz, A. A., Francis, A. L. 2015. Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English, *J. of Phon.* 49, 77–95.

[8] Seo, Y., Dmitrieva, O., Cuza, A. 2022. Crosslinguistic influence in the discrimination of Korean stop contrast by heritage speakers and second language learners, *Languages* 7, 6.

[9] Valdés, G. 2005. Bilingualism, heritage language learners, and SLA research: Opportunities lost or seized? *Modern Language* 89, 410–426.

[10] Carreira, M., Kagan, O. 2011. The results of the National Heritage Language Survey: Implications for teaching, curriculum design, and professional development, *Foreign Language. Annals* 44, 40–64.

[11] Oh, S., Jun, S.-A., Knightly, L. M., Au, T. K. 2003 Holding on to childhood language memory, *Cognition* 86, B53–B64.

[12] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287–308.

[13] Abramson A. S., Lisker, L. 1985. Relative power of cues: F0 shift versus voice timing, *Phonetic Linguistics Essays Honor Peter Ladefoged*, 25–33.

[14] Kong, E., Holliday, J., H. Lee, H. 2022 Post-adolescent changes in the perception of regional sub-phonemic variation, *J. of Phon.* 90.

[15] Kang, K.-H., Guion, S. G. 2006. Phonological systems in bilinguals: Age of learning effects on the stop consonant systems of Korean-English bilinguals, *J. Acoust. Soc. Am.* 119, 1672–1683.

[16] Kang, Y., Nagy, N. 2016. VOT merger in Heritage Korean in Toronto. *Language Variation and Change* 28, 249–272.

[17] Birdsong, D., Gertken, L. M., Amengual, M. 2012. Bilingual language profile: An easy-to-use instrument to assess bilingualism, *COERLL Univ. Tex. Austin*.

[18] Boersma, P., Weenink, D. 2012. *Praat Doing Phon. by Computer*.

[19] Peirce, J. W. 2007. PsychoPy—psychophysics software in Python. *J. of Neuroscience Methods* 162.

[20] R Core Team. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/.

[21] Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting linear mixed-effects models using lme4, *ArXiv Prepr. ArXiv14065823*.

[22] Polinsky, M. (2018). *Heritage languages and their speakers* (Vol. 159). Cambridge University Press.

[23] Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). Second Language Speech Learning: Theoretical and Empirical Progress, 3–83.

INDIVIDUAL VARIABILITY IN L1 CATEGORY COMPACTNESS ON L2 PRODUCTION COMPACTNESS AND ACCURACY

Weiyi Zhai[a,c], Meghan Clayards[a,b,c], Heather Goad[a,c]

[a] Department of Linguistics, McGill University; [b] School of Communication Sciences and Disorders, McGill University; [c] Centre for Research on Brain, Language and Music, Montreal, Canada

weiyi.zhai@mail.mcgill.ca; meghan.clayards@.mcgill.ca; heather.goad@mcgill.ca

ABSTRACT

Individuals with the same native language (L1) differ in how successfully they acquire the sounds of a second language (L2). One source of individual differences explored in recent studies is L1 production variability. In this study, we examined whether individual variation in the compactness of Japanese L1 categories can be directly correlated with the compactness and accuracy of speakers' L2 English categories. The F2 and F3 values of 30 Japanese speakers' productions of L1 /ɾ/ were used to calculate individual L1 compactness. The F2 and F3 values of the same speakers' productions of L2 English /l/ and /ɹ/ were used to calculate speakers' L2 category compactness and production accuracy (as compared to L1 English). We found no correlation between speakers' L1 category compactness and L2 production compactness or accuracy.

Keywords: category compactness, production variability, second language acquisition

1. INTRODUCTION

1.1. Background

The speech of second language (L2) learners is often marked by non-native-like perception and pronunciation. Theories of L2 acquisition attribute learners' difficulties acquiring the sound system of the target language to cross-language interference where speakers' native language (L1) sound systems act as a filter that influences the production and perception of novel L2 sounds [5, 3]. However, it has been shown that speakers with the same L1 background and with similar levels of L2 proficiency can vary greatly in how native-like their L2 productions are (e.g. [9]). Research has also shown that L1 speakers differ from each other in L1 production [12, 13, 11].

Kartushina and Frauenfelder [8] investigated the role of L1 category variability in the production and perception of L2 vowels. They quantified variability in production as a single "Compactness Score" (CS). CS is calculated as the area of an ellipse using the standard deviation of the F1 and F2 values of a vowel as the two axes. Speakers who produced more variable, less consistent L1 vowels had higher compactness scores. The authors focused on native Spanish speakers' production of the French /e/ - /ɛ/ contrast which is assimilated to a single /e/ category in Spanish. Spanish speakers who produced more compact L1 /e/ also produced the French /e/ - /ɛ/ contrast more accurately as defined by a smaller acoustic distance between the L2 and native French speakers' productions of the same vowel in F1-F2 space. This study demonstrated that for native-like production of L2 phones, it is advantageous for L2 speakers to have more compact L1 categories.

Huffman and Schuhmann [7] showed that the link between L1 category variability and L2 production accuracy extends beyond the F1-F2 space to voice onset time (VOT). They measured English speakers' production of VOT in L1 word-initial stops and L2 Spanish stops. They measured the standard deviation of learners' L2 VOT over the course of a semester in an introductory Spanish class. They found that the learners' L1 English production compactness in stop VOT at the beginning of the semester correlated with the speakers' production accuracy of L2 Spanish VOT at the end of the semester as in [8]. Additionally, they found that speakers with more accurate L2 VOTs are also more compact in their L2 VOT productions.

Kartushina and Frauenfelder [8] suggested two possible explanations for the positive correlation between compact L1 categories and accurate L2 productions. First, having more compact categories could mean a greater distance in acoustic or articulatory space between L1 categories and new categories, making the new categories easier to learn [5]. Second, individuals' L1 category compactness could reflect cross-language articulatory skill. More compact categories could demonstrate higher precision and higher articulatory skill, and that this skill can vary across individuals. Articulatory precision should give speakers an advantage in learning any novel L2 phones.

If articulatory precision were the explanation, we would expect to find a correlation between speakers' L1 and L2 category compactness in addition to L1 compactness and L2 accuracy. The study of VOT [7] found that speakers with more compact L1 categories are more accurate in L2 productions and that speakers with more accurate L2 VOTs are also more compact

in their L2 VOTs, providing some support for this hypothesis. However, [8] tested the correlation between L1 and L2 category compactness directly and did not find a link. Therefore, it is still unclear whether there is such a link, and thus what the nature of the relationship is between L1 compactness and L2 production accuracy.

1.2. Current study

The current study aims to further explore the relationship between individual-specific L1 category compactness and L2 production. We do so by replicating the production experiment in [8] to extend it to the F2 - F3 space for consonants. We will test whether the compactness of native Japanese speakers' productions of the L1 flap /ɾ/ category is linked to production accuracy and compactness in the L2 English /l/ - /ɹ/ contrast by measuring F2 and F3. This contrast is notoriously difficult for Japanese L1 learners [1, 14, 6] and while large individual differences have been observed [6], the source(s) of these differences remain unclear. In addition to shedding light on sources of individual variability in Japanese learners of English, this will also allow us to replicate the link between L1 compactness and L2 accuracy found for vowels [8] and VOT [7] with a new set of categories and contrast. Secondly, we will test whether there is a direct link between L1 compactness and L2 compactness which would support a role for generalized articulatory skill in L2 learning. Additionally, as a preliminary exploratory analysis, we will look for any possible effects of coarticulation from the phonetic context on the compactness and accuracy of L2 productions. Since previous research did not control for the phonetic environment of the productions [3, 7], participants could have more or less compact categories due to more or less coarticulation. By exploring coarticulatory influences, we can determine whether speakers who are less compact in a category are so due to them having a general lack of articulatory precision, or having distinct productions of the same category depending on the phonetic context.

2. METHODS

2.1. Participants

After applying exclusion criteria, productions from 31 native Japanese speakers (19M, 12F) currently residing in Japan and 5 native English speaking controls from the United States were included in the analysis (3M, 2F). As a measure of proficiency, Japanese participants were asked to read a short English passage. Two sentences from the middle of the recording were used to assess global accent. Sentences were rated by 10 native English speakers on a scale from least (1) to most (9) fluent, with participants receiving an average of 5.7 (SD = 1.5) for English fluency.

None of the native English speakers self-reported any prior experience with Japanese or fluency in another language.

2.2. Stimuli and procedure

To elicit L1 production of Japanese liquids, a word list was created with 10 /ɾ/-initial two-mora target words. The full word list is available at the OSF: https://osf.io/wu7rq/?view_only=3735fd1c2fed46e4a525824007715207. The word list included the target L1 phone in a variety of following vowel environments. All words had an initial high pitch accent. Similarly, to elicit productions of the English liquids for the native Japanese and native English speakers, a word list was created with 20 /ɹ/-initial and 20 /l/-initial one-syllable words. The target liquids were initial singletons to match the profile of the Japanese targets and to control for any effects due to syllable position. Additionally, 30 Japanese and 94 English words were included as fillers for the Japanese and English production tasks respectively. Filler words matched the profile of the target words.

The experiment was created and hosted on Gorilla Experiment Builder [2]. All participants completed the experiment on their own laptops or desktop computers. Prior to completing experimental tasks, all participants filled out a language questionnaire. Native English participants completed the English production task. Native Japanese participants first completed the Japanese production task, followed by the English passage reading task, and finally the English production task.

The production task was disguised as a picture-naming memory task to avoid participants putting any deliberate focus on the target sounds. Each trial began with a familiarization phase in which participants were shown three picture-word pairings and were asked to memorize the pairings. In each trial, there was one target word and two filler words, and the position of the target and filler words was randomized between trials. The words were presented only orthographically in the familiarization phase. Participants were then presented with each of the three pictures individually in random order and asked to name the picture out loud.

For L1 productions, participants had 5s to familiarize themselves with the three picture-word pairings and 5s to record each word. For the L2 English productions, Japanese participants had 15s to familiarize themselves with the picture-word pairings and 5s to record each word.

2.3. Measurement and normalization

Formant values and duration data were extracted from the recordings with Praat [4]. For each speaker's productions, F1, F2, and F3 values of the liquid productions were measured at the earliest measurable point in the steady state of the approximant. To compare speakers across genders and different vocal tract sizes while preserving formant variation, Nearey Normalization [10] was used on all formant values.

2.3.1. Computation of category compactness

Compactness scores (CS) were calculated for the target liquids for each participant in their L1 and L2. Because the distribution of target phones was assumed to be elliptical and normally distributed, the formula for compactness scores (CS) was derived from the equation for the area of an ellipse in (1) following [8].

(1) $$\text{area} = ab\pi$$

a = ½ the length of the ellipse's major axis, b = ½ the length of the ellipse's minor axis

This is translated into F2-F3 space in (2).

(2) $$CS = \sigma F2 \sigma F3 \pi$$

σF2 = 1 standard deviation of the mean of normalized F2, σF3 = 1 standard deviation of the mean of normalized F3

The area calculated in CS corresponds to a 66% confidence ellipse: the smallest ellipse that captures 66% of the data. Two sets of compactness scores were calculated for each of the Japanese participants. This includes two L1 compactness scores (L1CS) for the L1 /ɾ/ and two L2 compactness scores (L2CS) for the L2 /ɹ/ and /l/ categories. Higher CS values indicate less compact categories and lower CS values indicate more compact categories.

2.3.2. Computation of L2 production accuracy

Following [8], Japanese participants' L2 production accuracy was calculated as the Euclidean distance (EuD) in the F2-F3 space between each Japanese speakers' production and the average value of the native English-speaking controls' approximant productions in the same target word. This allowed us to control for any co-articulatory effects from the surrounding context. The EuD is the distance from one point to another in a two-dimensional space, calculated as in (3).

(3) $$x = \sqrt{(F2_i - F2_j)^2 + (F3_i - F3_j)^2}$$

Here, x is the EuD, F2i and F3i represent the Nearey normalized second and third formant values of a given Japanese participant's production, F2j and F3j represent the normalized formant values of the monolingual English participants. Greater EuD values represent lower production accuracy and smaller EuD values represent higher production accuracy.

3. RESULTS

For Japanese L1 productions, there were a total of 277 tokens of word-initial /ɾ/. The L1 compactness score (L1CS) for the Japanese liquids ranged from 0.0013 to 0.4361 (SD = 0.075).

For English L1 productions, there were a total of 106 tokens of word-initial /l/ and 110 tokens of word-initial /ɹ/. Figure 1 shows the F2 and F3 values for the L1 Japanese and L1 English productions.

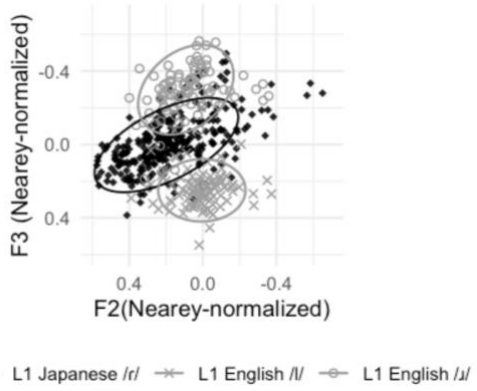


Figure 1: Scatter plot of L1 liquid productions.

For English L2 productions, 566 /ɹ/-initial and 533 /l/-initial L2 English productions were measured. Figure 2 shows the F2 and F3 values of all L2 English productions.

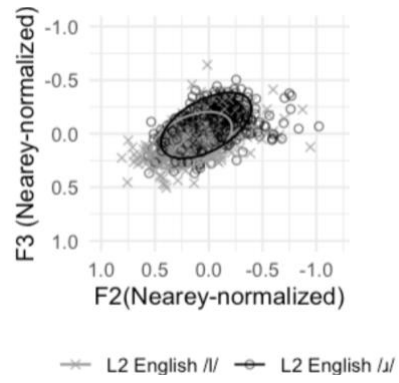


Figure 2: Scatter plot of L2 productions of English liquids.

Paired Pearson correlation analysis between speakers' L1 compactness scores and L2 EuD scores did not find evidence of a correlation ($r(28) = 0.46$, p

= 0.65). No correlation was found after excluding 5 data points that are visually identified as outliers ($r(23) = 0.17$, $p = 0.86$; Fig. 3).

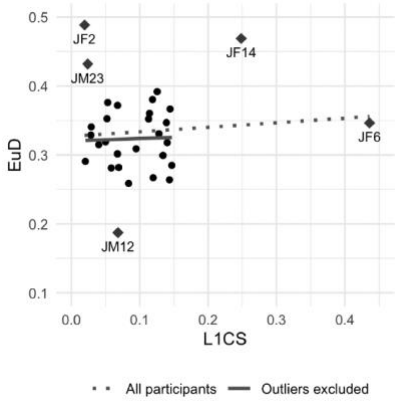


Figure 3: Correlation between L1CS and EuD with and without outliers. Each point in the graph represents a participant. Outliers are labeled.

We also did not find a correlation between speakers' L1 compactness scores and their L2 compactness scores ($r(28) = 0.52$, $p = 0.61$). No correlation was found after excluding 4 data points identified as outliers by visual inspection ($r(24) = 0.40$, $p = 0.693$).

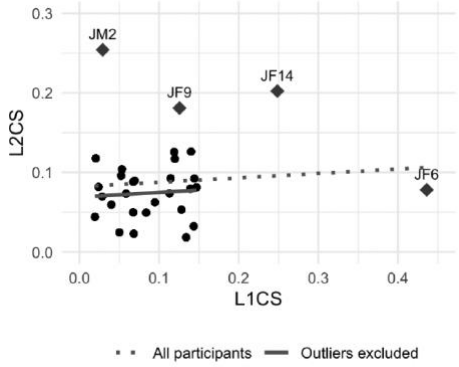


Figure 4: Correlation between L1CS and L2CS with and without outliers. Each point in the graph represents a participant. Outliers are labeled.

In addition, we did not find correlations between Japanese speakers' English fluency in the global accent test and L2 production accuracy ($r(28) = -0.04$, $p = 0.83$) or L2 compactness ($r(28) = 0.04$, $p = 0.39$).

A generalized linear mixed-effects model was built with L2 speakers' EuD values as the response variable. The model found that Japanese speakers produce more accurate English liquids when the liquid is followed by a back vowel compared to a non-back vowel ($\beta = -0.04$, $\sigma = 0.02$, $p < 0.05$). In addition, we also found a significant interaction between L1CS and following vowel backness ($\beta = -0.19$, SE = 0.06, $p < 0.005$). This shows that the effect of L1CS on EuD is modulated by the backness of the following vowel.

4. DISCUSSION AND CONCLUSIONS

Our first research question aimed to assess whether individual differences in native Japanese speakers' L1 category variability impact the production accuracy of L2 liquid consonants. We measured native Japanese speakers' L1 category compactness for /ɾ/ in the F2-F3 space. The results showed no evidence of a direct relationship between L1 category compactness and production accuracy in the L2. This result is inconsistent with what was found in [7] and [8]. This may be because having a more compact L1 /ɾ/ category in Japanese does not provide an advantage for native-like production of L2 English liquids, or we may have failed to find a link for other reasons. The second research question was whether there is a direct correlation between L1 and L2 category compactness. In [7], an indirect link was found between L1 compactness and L2 accuracy and between L2 accuracy and L2 compactness. However, like in [8], we did not find a correlation between L1 and L2 category compactness in the current study.

While we must be cautious in interpreting these null results, they could indicate that articulatory skill may not be the best explanation for the individual variation observed in L1 and L2 category compactness. It is possible that compactness in F2 - F3 space should be measured differently from compactness in the F1 - F2 space in [8]. Future research can explore different ways of measuring compactness for different categories.

In terms of coarticulatory influence of the following vowel backness, we found that L2 productions are more accurate when the liquid is followed by a back vowel. This is because English liquids have lower F2 compared to the Japanese flap [Fig. 1]. When the F2 of L2 productions are lowered due to coarticulation, the productions are also more accurate. The interaction between L1CS and vowel backness shows that for speakers with more compact L1 categories, the effect of back vowel coarticulation is small and speakers with less compact L1 categories are more prone to the effect of coarticulation. This result suggests that one reason behind speakers' category compactness in both L1 and L2 is speakers' resistance to coarticulation. Although this result does not explain the possible relationship between L1 compactness and L2 accuracy, it is an interesting area for future follow-up studies.

In conclusion, the results of this study indicate no direct relationship between L1 category compactness and L2 production accuracy. However, category compactness might be influenced by the degree of coarticulation that varies on an individual level.

5. REFERENCES

[1] Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. 2004. Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32*(2), 233-250.

[2] Anwyl-Irvine, A.L., Massonié, J., Flitton, A., Kirkham, N.Z., Evershed, J.K. 2019. Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods.*

[3] Best, C. T., & Tyler, M. D. 2007. Nonnative and second-language speech. *Language experience in second language speech learning: In honor of James Emil Flege, 13*-34.

[4] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.12, retrieved 17 April 2022 from http://www.praat.org/

[5] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research, 92*, 233-277.

[6] Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *J. Acoust. Soc. Am., 125*(1), 469-479.

[7] Huffman, M. K., Schuhmann, K. S. 2020. The relation between L1 and L2 category compactness and L2 VOT learning. *In Proceedings of Meetings on Acoustics 179ASA* (Vol. 42, No. 1, p. 060011). Acoustical Society of America.

[8] Kartushina, N., Frauenfelder, U. H. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology, 5*, 1246.

[9] Moyer, A. 2004. Age, accent and experience in second language acquisition. Multilingual Matters.

[10] Nearey, T. M. 1978. Vowel space normalization in synthetic stimuli. *J. Acoust. Soc. Am. 63*(S1), S5-S5.

[11] Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am., 109*(3), 1181-1196.

[12] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. J. Acoust. Soc. Am., 116(4), 2338-2344.

[13] Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., ... & Guenther, F. H. (2004). The distinctness of speakers'/s/—/ʃ/contrast is related to their auditory discrimination and use of an articulatory saturation effect.

[14] Sheldon, A., & Strange, W. 1982. The acquisition of/r/and/l/by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied psycholinguistics, 3*(3), 243-261.

A preliminary study on Cantonese vowel development by young heritage children

Chen Lan, Peggy Mok

The Chinese University of Hong Kong
lchsapphire@gmail.com, peggymok@cuhk.edu.hk

ABSTRACT

This study investigated the developmental pattern of Cantonese vowel system by young heritage speakers (HSs). Eighteen HSs from Vancouver aged 2;0, 4;0, and 6;0 participated in a picture-naming task; seven long vowels and four short vowels were extracted and measured. Results of acoustic analysis suggested that there was no clear distinction among the vowels produced by HSs at age 2;0, displaying a pattern of neutralization to the central position of the vowel space. Vowels became more distributed as age increased, but 6-year-old HSs' production was still more centralized compared to the reference pattern. Further comparison regarding the acquisition of vowel length contrast among HSs, homeland children, and native teenagers indicated that while reference speakers distinguish [aː] vs. [ɐ] in both vowel quantity and quality in all the environments (diphthong and close syllable), homeland children and HSs made use of the cues inconsistently according to environments.

Keywords: heritage language, heritage speakers, Cantonese vowel, language acquisition

1. INTRODUCTION

Heritage language (HL) is a newly recognized form of bilingualism, which has attracted growing research interest in recent years. Heritage speakers (HSs) hear and speak their HL as well as the majority language sequentially or simultaneously in early childhood. Several studies investigated HL phonology in Chinese regarding segments and tones, but most of them were on Mandarin (e.g. [1, 2]) and primarily focused on teenagers or adult speakers (e.g. [3, 4]). One published research investigated young children at the early stages of language development, and it was on Cantonese tones [5]. Little is known about the acquisition of segments in Cantonese by young HSs.

The Government of Canada reports that 208,935 people emigrated from Hong Kong to Canada during the period of 1980 – 2016, and Cantonese is the second most spoken first language among the non-official languages in Canada [6, 7]. The most recent census indicates that 191,940 of 565,270 Canadians who speak Cantonese as their mother tongue are in Vancouver, providing a valuable pool of potential participants for studies on the acquisition of Cantonese by HSs [7]. This study made use of the Vancouver data pool to investigate the production of Cantonese vowels by young HSs, aiming to explore the developmental pattern of vowels and compared it to those reported among the monolingual speakers in Hong Kong.

1.1. Acquisition of Cantonese vowels

The Cantonese vowel system contains seven long vowels [iː yː uː ɛː œː ɔː aː] and six short vowels [ɪ ʊ e ø o ɐ] [8, 9]. Considering [e] and [ɛː] as well as [o] and [ɔː] are variants of the same phonemes, this study analysed them as /ɛː/ and /ɔː/ respectively. Vowels in Cantonese contrast in length and lip rounding. Vowel quality (formant) and quantity (duration) are the significant cues for the distinction of vowel length contrasts. Considering that only the /aː/ vs /ɐ/ pair contrasts in all environments (in diphthongs, before stops and nasal codas) with a high functional load [10, 11], this study only examined this contrastive vowel pair. Regarding vowel roundedness, Cantonese has two pairs of vowels differing in lip roundedness: /iː/ vs. /yː/ and /ɛː/ vs. /œː/.

Previous studies on the acquisition of Cantonese vowels suggested that monolingual children are able to correctly produce most of the vowels and use them contrastively by age 2;3, and the acquisition is completed at age 5;0 [12, 13, 14]. Bilingual pattern was comparable to the monolingual norm [15]. Adult HSs aged between 20 and 87 maintained the 7 vowel contrasts while having inter-generational vowel shift (i.e., the retraction of [yː] and the fronting of [iː]) [16]. Vowel changes were also observed in adult Cantonese speakers (i.e., lowering of [ɪ] and [ʊ]) [16].

1.2. Phonological development of heritage language

To date, there is no framework that specifically targets the speech acquisition of young HSs. A developing predictive model of HL competence on sound system suggested that the phonologically distinct contrasts in HL should be maintained or even overregularized by HSs, and HSs may not simply develop a merged system [17, 18]. Based on this preliminary model, it can be predicted that the HL sound system will be less ambiguous, more regular, and with less innovative merging phenomenon. The Speech Learning Model (SLM [19]) was also utilized

to predict the development of a HL sound system, proposing that the interaction between the sound system of the majority language and the HL may contribute to an HL sound system differing from the monolingual norm [20]. Previous studies on HSs supported that the productions of HL segments by HSs, even at pre-schooler age, were significantly different from those of monolinguals [e.g., 21]. In addition, HL acquisition may relate to language variation and change [16]. For instance, the retraction of Cantonese [y:] found in adult HSs did not exist among Hong Kong Cantonese speakers.

Based on these models, we predicted that the development of Cantonese vowels by young HSs may differed from that of their monolingual peers, with a slower acquisition progress. Also, some distinctive vowel contrasts in Cantonese may be overregularized by HSs.

2. METHOD

2.1. Participants

Eighteen HSs aged 2;0 (6, 3f), 4;0 (6, 2f), and 6;0 (6, 4f) who were born and raised in Vancouver, Canada, participated in the production experiment. The age of each participant is more or less no more than 5 months than age 2;0, 4;0, and 6;0. All the participants reported having no hearing or speech problems. Formant frequency data of native female speakers aged between 18 and 21 reported in [9] was used as the reference for comparison. For the acquisition of vowel length contrast, eighteen Hong Kong Cantonese children (homeland children) aged 2;0 (6, 2f), 4;0 (6, 4f), and 6;0 (6, 3f) and nine native speakers (7f) aged between 15 and 18 from [22] were used for comparison.

2.2. Materials

The vowel production experiment was a picture-naming task, which was adopted from the Hong Kong Cantonese Articulation Test (HKCAT) [23], a standardized Cantonese articulation test with norm reference on the acquisition of consonants, vowels, and tones in Hong Kong. All words, such as "faa1 *flower*, ce1 *car*, so1 *comb*, fu3 *a pair of pants*, jyu5 *rain*, daai6 *big*", were familiar for Cantonese-speaking children. Fifty-two colored pictures, each accompanied by its name in Chinese character(s), were used to elicit 63 Cantonese words. Fourteen of the 63 words were chosen to investigate the 7 long vowels in open syllable and closed syllables, 4 were used to examine the 4 short vowels in closed syllables, and 2 with diphthong were selected to explore the participants' ability to produce the vowel length contrast (/a:/ vs. /ɐ/).

2.3. Procedure

All participants were paid to do the experiment in a quiet room with parental consents. A detailed language background questionnaire was filled in by the parents of the children. An experimenter, who was a native speaker of Cantonese with phonetical training, facilitated the test. Following the instructions in HKCAT, the experimenter elicited the production by asking questions such as "*ni1 go3 mat1 ye5 lei4 gaa3? What is this?*" while pointing to the picture. Children were instructed to name the target words in isolation twice. No time restriction was imposed. If the participants really could not produce the item, they would repeat the target word after the experimenter. The two clearest repetitions of each item were chosen for analysis. In total, each participants produced 40 words for comparison.

2.4. Data analysis

The acoustic measurements were done in Praat using FormantPro [24]. The first, second and the third formant frequencies of all the tokens were measured, and the average values of 11 vowels were plotted on the F1-F2 and F1-F3 panel. Further evaluation on the vowel length contrast (/a:/ vs. /ɐ/) involved both vowel quantity and vowel quality. The duration of each target vowel production was extracted, measured in millisecond (ms), and reported in second (s). The formant frequencies of the relevant tokens were measured at 10 equidistant points along the vowel interval. Manual check and correction were done. The raw data of formant frequencies in Hz were converted to Bark for normalization [25].

3. RESULTS

3.1. Vowel space

Figure 1 displays the positions of the 7 long vowels produced by the HSs at ages 2;0, 4;0, and 6;0 relative to the reference positions in the F1-F2 panel. In general, there is no clear distinction among the vowels produced by HSs at age 2;0, displaying a pattern of neutralization to the central position of the vowel space. Vowels became more distributed as age increased, but 6-year-old HSs' production was still more centralized compared to the reference pattern. The high vowels [i: y: u:] and the low vowel [a:] were acquired first at age 2;0, with significant difference in F1 respectively between [a:] and [i:] ($p < .001$), [y:] ($p < .001$), and [u:] ($p < .001$). There was no significant F1 difference between the low vowel [a:] and the vowels posited in the middle except for [œ:] ($p = .009$). For F2, significant difference was only found between [y:] and [u:] ($p < .001$). The

development of vowel production by HSs from age 2;0 to 4;0 was inconsistent, with more expanded movement in F1 but not in F2. For example, F1 of [ɛː, iː, uː, œː, yː] decreased as age increased, resulting in the 4-year-old vowel space more contrastive in height. However, there was no significant change in F2 for all vowels; thus, the front and back vowels remained non-distinctive. The vowels produced by HSs at age 6;0 was observed to be more distributed compared to the vowel space at younger ages, and the relative position of each long vowel on the F1-F2 panel was similar to the reference. Statistical analysis further supported that [aː] had a significantly higher F1 frequency while [iː yː uː] had a significantly lower F1 than all the other vowels except for [œː] ([uː] vs [œː], p = .09), suggesting that HSs have acquired the vowel height of the primary vowels, although the space was still centralized relative to the reference pattern. Regarding vowel frontness, [iː, ɛː, yː] had a significantly higher F2 frequency while [uː] had a significantly lower F2 than all the other long vowels except for [ɔː]. Therefore, a vowel system with distinctive long vowels had been preliminarily formed at age 6;0 by HSs, with the high front vowels [iː, yː], a high back vowel [uː], a mid front vowel [ɛː], and a low central vowel [aː]. The vowel height of [œː] and the backness of [ɔː] were still developing. Furthermore, [iː] was expected to have a relatively lower F1 and higher F2 than [yː] and [ɛː]; however, there was a retraction of [iː] produced by HSs even at age 6;0, reflected by the overlap between [iː] and [yː] as shown in the F1-F2 panel.

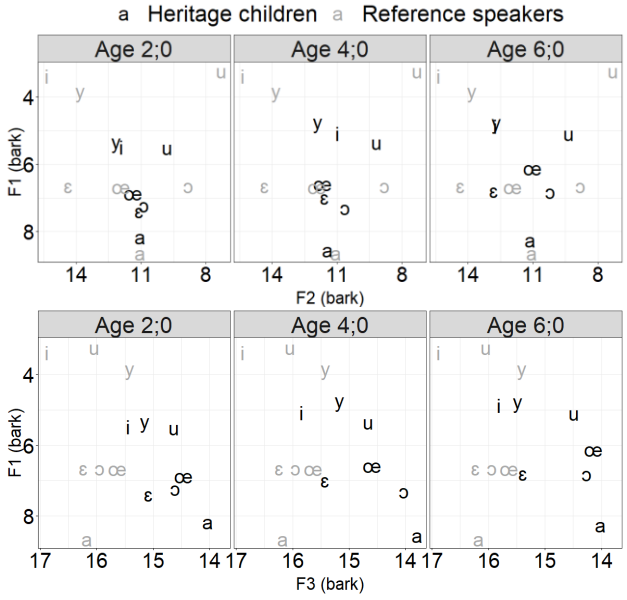


Figure 1: Mean F1, F2, and F3 values (in bark) of 7 long vowels produced by HSs and reference speakers.

3.2. Vowel roundedness contrast

Lip-rounding results in lowered F2 and F3 and is therefore useful in distinguishing between rounded and unrounded sounds. Cantonese has two pairs of vowels differing in lip roundedness: [iː] vs. [yː] and [ɛː] vs. [œː]. As seen in Figure 1, compared to the unrounded vowel [ɛː], the rounded vowel [œː] had a significantly lower F2 (p = .01) and F3 (p = .02) frequencies, revealing a significant distinction in lip-rounding between these two vowels by HSs at 6;0. The F3 distance between these two mid vowels was larger for HSs than for the reference speakers. Regarding the [iː yː] pair, HSs distinguished them in F3 with a lower frequency for [yː] than for [iː], although the difference was smaller than the reference. There was no significant difference between these two vowels in F2.

3.3. Vowel length contrast

Table 1 summarizes the durational differences between the mean value of [aː] and [ɐ] produced by the HSs, homeland children, and reference speakers in different environments. There was a significant effect for Age (F (3,915) = 21.254, p < .001), for Group (F (1,915) = 19.541, p < .001) and for Environment (F (1,915) = 888.094, p < .001). Evaluations regarding Age and Environment indicated that both HSs and homeland children utilized duration to distinguish [aː] and [ɐ] in pre-nasal environment since age 2;0. However, for the vowel contrast in diphthong, significant durational difference between these two vowels was only found in the homeland productions at age 6;0. There was no significant durational difference in each age group of HSs.

Table 1: The durational differences (in s) between the mean values of [aː] and [ɐ] by young HSs, homeland children, and reference speakers

	Diphthong	Pre-nasal
Heritage	0.08	0.17
Homeland	0.13	0.20
Reference	0.03	0.12

With respect to vowel quality, the formant data were analyzed using Gu's [26] Smoothing Spline ANOVA (SSANOVA). Simply put, any non-overlapping area represents significant difference between the formant measurements. Because of the page limit, only figures indicating the F1 and F2 trajectories of the youngest and the oldest target age groups were displayed here (Figure 2). The reference speakers produced [aː] and [ɐ] in diphthong [aːi ɐi] significantly differing in both F1 and F2, while HL and homeland children partially utilized the formant

cue. HSs (Figure 2a) and homeland children (Figure 2c) were able to significantly distinguish the vowel length contrast since age 2;0, but the pattern of F2 was a reverse of those of the reference even at 6;0 (Figure 2b, 2d). For pre-nasal environments, while the reference speakers significantly produced the difference in both F1 and F2 between [a:] and [ɐ], HL and homeland children had a less contrastive pattern. HSs at age 2;0 significantly produced the difference in F1, but not in F2 (Figure 2e). Homeland children performed better than the HSs at age 2;0, with significant difference in both F1 and F2 (Figure 2g). However, no significant difference was found in F1 nor F2 among HSs (Figure 2f) and homeland children (Figure 2h) aged 4;0 and 6;0, suggesting that HL and homeland children could not distinguish [a:] and [ɐ] with pre-nasal coda in vowel quality.

To sum up, while reference speakers distinguish [a:] vs. [ɐ] in both vowel quantity and quality in all the environments, homeland children and HSs made use of the cues inconsistently according to environments.

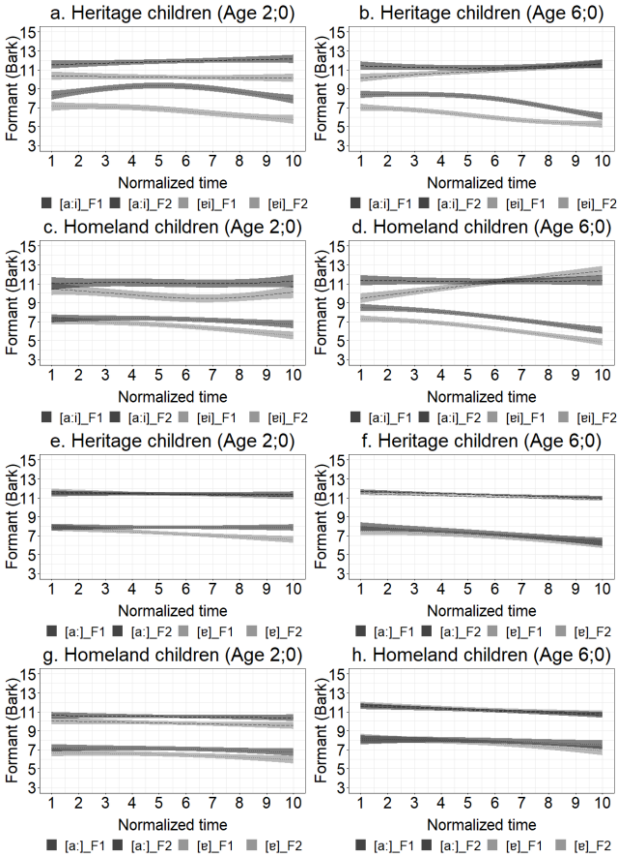


Figure 2: F1 and F2 trajectories for [a:] and [ɐ] in diphthongs (a-d) and before nasal codas (e-h) produced by HSs and homeland speakers.

4. DISCUSSION

This study investigated the acoustic development of 11 Cantonese vowels by young HSs. HSs mastered the distinction regarding vowel height earlier than vowel frontness. Unlike monolingual children who are reported to acquire the vowels by age 5;0, the acquisition of Cantonese vowels by HSs is not completed by age 6;0. For example, the vowels [œ:], [ɔ:], and [ɵ] were not mastered by 6-year-old HSs. It can be expected since even though the HSs had similar acquisition context as their monolingual peers at the early stage, they were exposed to the majority language since they attended schools at age 3;0. The interaction between two sound systems may influence the development of HL. For example, the absence of [œ:] and [ɵ] in Toronto English vowel system reduced the amount of inputs of these vowels to the children, which may further slow their development down.

Regarding the vowel length contrast, HSs did not distinguish [a:] vs. [ɐ] in diphthongs in duration but partially in vowel quality. It can be expected since vowel quality functions as the primary cue in English tense/lax contrast [27]. For homeland children, they could distinguish [a:] vs. [ɐ] in both diphthongs and the pre-nasal environment by utilizing vowel quantity by age 6;0. Interesting pattern was found regarding the formant pattern of [a:] vs. [ɐ] in the pre-nasal environment. HSs did not have significant distinction in F1 and F2 since age 2;0, showing that they have not acquired this cue until age 6;0. However, for homeland children, significant differences in both F1 and F2 were found in their production at age 2;0, but the distinctions became weaker as they got older. Instances of [a:]-[ɐ] merger-in-progress have been found among native Hong Kong Cantonese speakers, in particular the younger generation [28], South Asian speakers [22], and Mandarin speakers [16]. A possible explanation is that, as age increased, the homeland children may communicate with a larger amount of people whose Cantonese may have such a merging phenomenon. Another interesting finding is with respect to the vowel roundedness contrast by HSs. While L2 learners found [ɛ:] vs [œ:] more difficult to distinguish than [i:] vs [y:] [22], HSs performed better in the distinction of the former pair. The extensively lower F3 for the rounded vowels relative to the unrounded vowel revealed that the HSs may overregularize the lip roundedness to discriminate the rounded vowel from the unrounded vowel. This finding provides evidence for the predictive model on the competence of HL sound system, supporting the principle that it is possible for the HSs to maintain or over-mark the phonologically distinct contrasts in HL. This study provided preliminary findings on the acquisition of vowels by young HSs. Data from more young HSs and from monolingual children at comparable ages will be adopted for a more comprehensive understanding.

5. ACKNOWLEDGEMENT

We would like to thank Professor Molly Babel and her RAs at the University of British Columbia for facilitating our data collection.

6. REFERENCES

[1] C. B. Chang, Y. Yao, E. F. Haynes, and R. Rhodes, Production of phonetic and phonological contrast by heritage speakers of Mandarin. *The Journal of the Acoustic Society of Ameirca, 129*(6), 3964 – 3980, 2011.

[2] C. B. Chang and Y. Yao, Toward an understanding of heritage prosody: Acoustic and perceptual properties of tone produced by heritage, native, and second language speakers of Mandarin. *Heritage Language Journal, 13*(2), 134-160, 2016.

[3] W. M. Lam, Perception of lexical tones by homeland and heritage speakers of Cantonese. *University of British Columbia*. 2018.

[4] R. T. Y. Kan and M. S. Schmid, Development of tonal discrimination in young heritage speakers of Cantonese. *Journal of Phonetics, 73*, 40-54, 2018.

[5] C. Lan and P. K. Mok, A preliminary study on Cantonese tone production by young heritage speakers. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, 106-110. Tokyo, Japan. 2020.

[6] Statistics Canada. *Immigration and ethnocultural diversity highlight tables*. 2016 Census. Statistics Canada Catalogue no. 98-402-X2016007. Ottawa. 2017.

[7] Statistics Canada. *Canada [Country] and Vancouver, CY [Census subdivision], British Columbia (table). Census Profile.* 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. 2017.

[8] E. Zee. An acoustical analysis of the diphthongs in Cantonese. In *Proceedings of the XIVth International Congress of Phonetic Sciences*. 1999.

[9] E. Zee, Frequency Analysis of the Vowels in Cantonese from 50 Male and 50 Female Speakers. *Proc. 15th ICPhS*, Barcelona, 1117-1120. 2003.

[10] H.-N. S. Cheung. *A grammar of Cantonese as spoken in Hong Kong*. Hong Kong: The Chinese University of Hong Kong. 1972.

[11] O. Yue-Hashimoto. *Phonology of Cantonese*. Cambridge: Cambridge University Press. 1972.

[12] L. K. H. So and B. J. Dodd. The acquisition of phonology by Cantonese-speaking children. *Child Lang. 22*, 473-495. 1995.

[13] S. F. Stokes and I. M. Wong. Vowel and diphthong development in Cantonese-speaking children. *Clinical Linguistics and Phonetics, 16*, 597-617. 2002.

[14] C. To, P. Cheung and S. McLeod. A Population Study of Children's Acquisition of Hong Kong Cantonese Consonants, Vowels, and Tones. *Journal of speech, language, and hearing research: JSLHR, 56*. 2012.

[15] A. Holm and B. Dodd, A longitudinal study of phonological development of two Cantonese-English bilingual children. *Applied Psycholinguistics, 20*(3), 349-76. 1999.

[16] H. Tse, Beyond the monolingual core and out into the wild: A variationist study of early bilingualism and sound change in Toronto heritage Cantonese. Unpublished doctoral dissertation, University of Pittsburgh, 2019.

[17] M. Polinsky and G. Scontras. Understanding heritage languages. *Bilingualism: Language and Cognition 23*, 4–20. 2020.

[18] T. Kupisch. Towards modelling heritage speakers' sound systems. *Bilingualism: Language and Cognition 23*, 29–30. 2020.

[19] J. E. Flege. Second-language speech learning: theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp.233-278). Baltimore, MD: York Press. 1995.

[20] R. T. Y. Kan. Phonological Production in Young Speakers of Cantonese as a Heritage Language. *Language and Speech, 64*(1), 73–97. 2021.

[21] A. Stoehr, T. Benders, J. G. Van Hell and P. Fikkert. Heritage language exposure impacts voice onset time of Dutch – German simultaneous bilingual preschoolers. *Bilingualism: Language and Cognition*. 2017.

[22] P. Mok, C.W. Leung, C. Lan and A.C.L. Yu. The acquisition of Cantonese vowel length contrast and vowel rounding contrast by South Asian students in Hong Kong. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)*. Melbourne. 2019.

[23] P. Cheung, A. Ng, and C. K. S. To, *Hong Kong Cantonese Articulation Test*. City University of Hong Kong, Hong Kong, 2006.

[24] Y. Xu and H. Gao. FormantPro as a tool for speech analysis and segmentation. *Revista de Estudos da Linguagem*. 2018.

[25] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America, 88*(1), 97–100, 1990.

[26] C. Gu. *Smoothing spline ANOVA models*. New York: Springer New York, 2013.

[27] H. J. Giegerich. *English phonology: An introduction*. Cambridge, UK: Cambridge University Press. 1992.

[28] S. Y. R. Fung and C. K. C. Lee. Cantonese vowel merger-in-progress. In S. Calhoun, P. Escudero, M. Tabain , & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, Australia (pp. 3225-3229). 2019.

LEXICAL STRESS PERCEPTION IN INDIAN ENGLISHES

Elinor Payne[a], Olga Maxwell[b], Robert Fuchs[c] and Yizhou Wang[b]

[a]University of Oxford, [b]University of Melbourne, [c]University of Hamburg
elinor.payne@phon.ox.ac.uk, omaxwell@unimelb.edu.au, robert.fuchs@uni-hamburg.de,
yizhouw4@student.unimelb.edu.au

ABSTRACT

We report an experiment investigating the relative weighting of acoustic cues (vowel quality, intensity, duration and f0) in lexical stress perception in Indian English (IndE), compared with Southern Standard British English (SSBE). GLMM modelling of responses shows both similarities (e.g. vowel quality was by far the most important cue for both IndE and SSBE) and differences (IndE listeners were less sensitive to all cues except duration, and made least use of f0).

Differentiating IndE participants according to L1 background (Indo-Aryan vs Dravidian), however, reveals a finer-grained picture, with L1 Indo-Aryan listeners exhibiting cue hierarchy and degree of cue strength that are closer to SSBE listeners. For L1 Dravidian listeners, while vowel quality remains the most important cue, the strength of this cue, and that of intensity, are significantly lower than for L1 Indo-Aryan and SSBE listeners. At the same time, duration ranks more highly for these listeners.

Keywords: Stress, perception, Indian English, SSBE, acoustic cues.

1. INTRODUCTION

The perception of stress is complex and subject to cross-linguistic variation that remains poorly understood. English is considered to be a stress language, with lexical stress cued by a combination of acoustic parameters that make stressed syllables more salient to listeners [1]. While the main acoustic cues to stress in SSBE and General American are reported as being higher f0, longer duration, greater intensity and full vowel quality [e.g. 2], there is little consensus over relative cue weighting and interaction. Some [e.g. 3] claim relative pitch to be dominant, while others assert its unreliability [4], or propose duration instead [5]. Furthermore, findings for one variety of English are not necessarily generalisable to another, as has been found for e.g. Welsh English [6].

Previous limited research suggests that speakers of new ('non-settler') Englishes spoken in multilingual societies, e.g. Singapore or Indian English, may lack a robust distinction between stressed and unstressed syllables, and/or may employ different acoustic cues, due to the influence of other languages in sustained *historic* contact during the development of these varieties. In addition, for many speakers, even with native-like command, English may not be the predominant or first language, and thus there may also be *ongoing* contact influence from another language. With regard to L2 perception, extensive research has shown that the phonological system of the L1, acquired very early in life [7], critically shapes the perception of sounds in L2, and the influence is greater the later the acquisition of L2; for example, segmental contrasts that do not exist in L1 are difficult for adult learners to acquire [8,9].

L1-mediated perceptual bias also affects prosody. Research has shown that adult listeners' ability to perceive and learn lexical stress is closely linked to the mechanisms used to signal stress in their L1 [10]. In a study on the influence of typologically diverse L1s (English, Russian and Mandarin) in the perception of lexical stress in English, Chrabaszcz et al. [11] found that while vowel quality was the strongest cue for all groups, pitch was the second strongest cue for English and Mandarin listeners, but virtually disregarded by Russian listeners who, in turn, used duration and intensity cues to a greater extent. This is despite Russian being a stress language, like English, and unlike Mandarin, which is a tonal language.

1.1 Indian English

'Indian English' (IndE) refers to the variety(/ies) of English used by speakers in India and by some in the global Indian diaspora. English is one of the official languages of India and is used in a wide range of domains (e.g. the media, education, business, government, literary writing), but coexists with a large range of diverse Indian languages. IndE is usually an L2, with the vast majority of speakers in India also speaking at least one and often two or three Indian languages. In the 2011 census [12], around 10% of inhabitants declared English as their L2 or L3, and a tiny minority (0.02%) as their L1.

Lexical stress in IndE is of particular interest because most, if not all, Indian languages lack

contrastive lexical stress [e.g. 13], and any word-level prominences are typically determined by structural properties such as syllable structure or position in the word [14]). Previous research on IndE phonological and phonetic features more generally indicates strong evidence for the influence of indigenous languages [e.g. 15,16,17,18], leading to identifiable accents of IndE, either through historical influence or from continued contact in a multilingual society. However, pan-Indian features also arise, through serendipitous cross-L1 commonalities ('areal features', see [19-20]), or through standardisation and homogenisation of IndE [21] due in part to its 'self-replicating nature', where English is taught to Indians *by* Indians [22]. Hence, the question arises as to whether and how this complex multilingual backdrop shapes the perception of lexical stress in IndE.

A further factor in the strength of ongoing L1 influence concerns educational background, known to play a significant role in distinguishing basilectal and acrolectal speakers of IndE [23]. While IndE is "primarily being transmitted through school" [17], the Indian education system comprises different types of schools providing varying levels of exposure to varying kinds of English. Government schools (especially in remote areas) are less likely to be English-medium, and more likely to have teachers who are either less fluent in English, or whose IndE is more influenced by a local or state language. By contrast, 'public' (private) schools are more likely to be English-medium, provide full immersion in English, and with teachers more proficient in English. There has also been a rapid rise in the number of international schools, with curricula closely modelled on British or US educational systems, some with British or US English-speaking teachers.

1.2 Stress/prominence in Indian L1s

The notion of 'stress' is an elusive phenomenon in South Asian languages, with speakers reporting difficulty in its auditory identification with disagreement over the presence and type of acoustic cues, and over the location of prominence [e.g. 24]. Unlike English, all Indian languages lack lexically contrastive stress, although prominences at the word level do exist, and these are derived from language-specific phonological and syllable structure considerations, and appear to serve what may be considered a post-lexical role, viz. the existence of a smaller, word-like prosodic constituent in Indian languages (and the higher tonal density that results). Auditory studies have claimed that speakers transfer these properties onto IndE, by accenting most words (c.f. [25]), and in L1-specific ways. Investigating phonetic cues in IndE speakers' (with Hindi or Telugu L1) production of accentual and focal prominence, Moon [26] showed that neither IndE groups relied on durational cues. However, while the two groups used similar values for maximum f0, IndE speakers with a Hindi L1 background showed greater lowering of f0 at the beginning of the accented syllable, possibly reflecting differences in pitch accent type and/or alignment in their L1.

Given the controversial status of 'stress' in Indian languages, and their very close contact with IndE, we might reasonably expect L1 properties to shape perceptual sensitivity of IndE listeners to lexical stress cues. While we might expect this to be strongly the case for multilingual speakers with IndE as L2, it could also be true for those who learn IndE at a very early age and have a 'native' command, since even if the hypothesised differences historically arose through contact, they could now be an intrinsic property of IndE, and thus be present even in monolingual speakers. We hypothesise lower sensitivity to stress cues in general, compared to SSBE listeners, and, following [26]'s results, that duration play little to no role at all in stress perception by IndE listeners. Given variable usage of acoustic cues across L1s for post-lexical prominences, and the scant research on this area for Dravidian languages in particular, potential differences in relative cue weighting are harder to predict. This paper reports a preliminary study exploring these questions.

2. METHOD

2.1 Perceptual experiment

To investigate this, we used stimuli from Chrabaszcz et al. [11], which were modified natural recordings of the disyllabic nonword '*maba'*, produced by a male speaker of North American English (see [11: 5] for a description of how pitch, duration, intensity and vowel quality were manipulated). Stimuli were varied according to a 2-factorial manipulation for each variable on each syllable, resulting in 256 unique tokens.

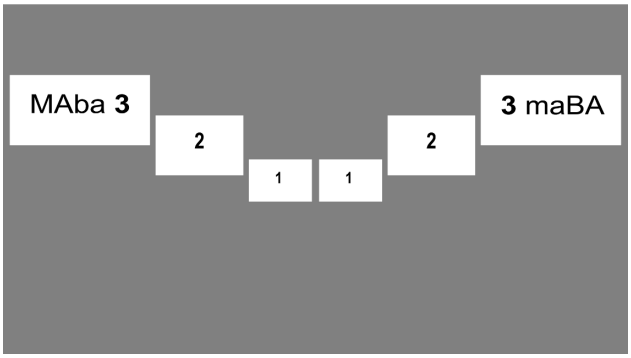


Figure 2 Screen layout for experiment in PsychoPy.

All participants performed a forced-choice auditory identification (using PsychoPy [27] and hosted online using *Pavlovia*) of the location of the stressed syllable by clicking on corresponding buttons on a computer screen with a computer mouse (see Fig. 2.1), a selection that also indicated degree of confidence. Reaction times (RTs) were also recorded. In addition, participants received 16 practice trials to familiarise themselves with the procedure. Participants were instructed to use headphones for completing the task.

Participants consisted of 44 IndE speakers, resident in India, and 29 SSBE speakers, resident in the UK. Due to COVID-restrictions, the experiment was carried out online. Participants were aged 18-34 years and either currently in higher education (HE), or already in possession of an HE qualification. The SSBE cohort were all monolingual speakers of English (with varying foreign language knowledge from school/university). The IndE cohort were all highly proficient in English, and chosen on the basis of their L1s representing two major Indo-Aryan languages, Bengali (n=13) and Hindi (n=10), and four major Dravidian languages, Telugu (n=8), Malayalam (n=5), Tamil (n=4), and Kannada (n=2).

We additionally collected detailed information on the Indian participants' sociolinguistic background, including medium of instruction in school, onset of exposure to English during childhood, and school type (state, private). Space constraints prevent us from exploring all these variables here. However, based on previous research and our initial analysis of the data, we focus here on L1 background and explore the difference between IndE speakers with Dravidian and Indo-Aryan L1s.

2.2. Analysis

After inspecting the original dataset, we removed one IndE participant who responded randomly during the task, and any trials with a very short (< 100 ms) or very long (> 10,000 ms) RT. This left us with 96.2% of the total trials. The perception data were analysed using generalised linear mixed-effects models (binomial GLMM) with response (first vs second syllable stressed) as dependent variable, acoustic cues, i.e. vowel, intensity, duration, and pitch, and language background (Model 1: SSBE vs IndE; Model 2: SSBE vs IndE Indo-Aryan vs IndE Dravidian) as fixed factors. Specific trials and individual participants were kept as random factors. Subsequently, we interpreted the coefficients of acoustic cues as indicators of the relative weighting of these cues in stress perception.

3. RESULTS

3.1 Indian English and British English

In the first model, we contrasted IndE speakers with SSBE speakers. Results indicate that vowel quality is by far the most important acoustic cue for speakers of both dialects and that IndE listeners are significantly less sensitive to all acoustic cues except duration, when compared to SSBE listeners (p's < .001). As a consequence, the cue hierarchies for stress perception differ to some extent between the two groups, principally due to a lower ranking of pitch for IndE listeners. Specifically, SSBE listeners followed a cue hierarchy of vowel > intensity > pitch > duration, while IndE listeners followed a somewhat different hierarchy of vowel > intensity/duration > pitch (see Fig. 3.1 for coefficients derived from the GLMM).

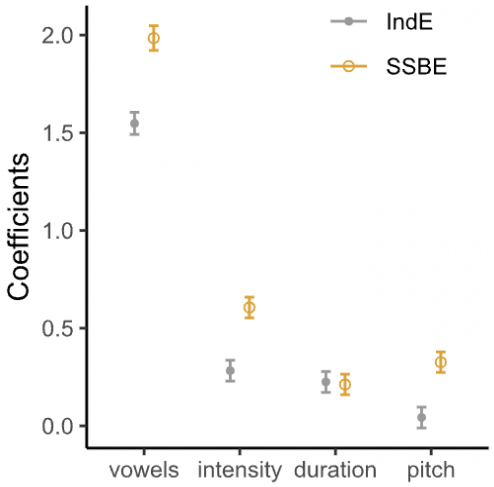


Figure 3.1: Cue-weighting in stress perception in IndE and SSBE participants (with standard errors).

3.2 Indian English L1 Background

To further investigate the influence of L1 background, we differentiated the IndE participants based on their L1 language family (Indo-Aryan vs Dravidian) and contrasted these two groups again with SSBE, using a GLMM. While vowel quality still emerges as the most important cue for all groups, results indicate that L1 language family background influences both cue hierarchy and cue strength in stress perception. The cue hierarchy is vowel quality > duration > pitch/intensity for L1 speakers of Dravidian languages, and vowel quality > intensity > duration (and no use of pitch) for L1 speakers of Indo-Aryan languages - see Fig. 3.2 for coefficients derived from the GLMM.

Furthermore, differences in cue strength are significantly influenced by background L1 family: cue strength for both vowel quality and intensity is

significantly lower for IndE listeners with Dravidian L1s compared to both the Indo-Aryan and the SSBE groups ($p<0.001$), with no significant difference between the latter two groups. By contrast, while cue strength for duration is somewhat higher for the IndE Dravidian compared to the Indo-Aryan group, and the same holds for pitch, these differences do not reach statistical significance. In sum, analysis by background L1 family indicates that acoustic cues to stress perception for IndE listeners with Indo-Aryan L1s are more similar to SSBE listeners than to IndE listeners with Dravidian L1s.

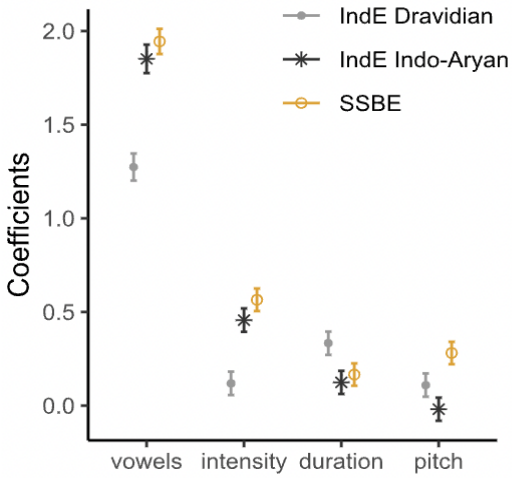


Figure 3.2: Cue-weighting in stress perception in SSBE, Indo-Aryan and Dravidian IndE participants (with standard errors).

4. DISCUSSION AND CONCLUSION

The results in part confirm our hypothesis that IndE speakers would show a different perceptual response to acoustic cues to lexical stress in English from speakers of SSBE. Though on first inspection, IndE speakers as a group appeared to be less sensitive to all cues except duration, further analysis distinguishing IndE speakers based on L1 revealed that L1 Indo-Aryan speakers actually behave very similarly to SSBE speakers, with regard to vowel quality, intensity and duration, and cue hierarchy, and differ only for pitch. In contrast, Dravidian speakers of IndE showed a clearly different perceptual response from SSBE speakers, with significantly less sensitivity to vowel quality and intensity differences, and a different cue hierarchy. At the same time, Dravidian speakers also showed a slightly (but not significantly) higher sensitivity to duration, and therefore a different cue hierarchy.

One possible explanation for these differences between IndE cohorts is that they are attributable to differences in L1 structural or phonetic properties between the two language families. The limited research on such properties across the relevant languages suggests varying uses of cues to mark post-lexical prominences, but without any obvious robust alignment according to language family. For example, in Telugu, Malayalam and Kannada, phonetic cues of non-lexical prominences are reported to include vowel duration [28, 29, 30, 31]. However, [24] reports Telugu listeners to be highly variable in their judgments, and for Malayalam, [32] reports listeners perceived stress to fall on the first syllable regardless of vowel duration. Furthermore, there are reportedly no reliable durational cues to prominence placement in Tamil [13].

The use of a particular acoustic cue for unrelated structural properties may be a relevant factor. Notably, only Dravidian languages have lexically contrastive vowel duration, and thus Dravidian speakers may already be more attuned to differences in vowel duration. This could explain the different positioning of duration in their cue hierarchy. With this in mind, a further step would be to differentiate L1 Hindi and Bengali speakers too. While neither language has contrastive vowel duration, there is potentially a predisposition in Hindi to associate a shorter vowel duration and more reduced quality with a lack of stress (at least in open syllables). In contrast, Bengali has no schwa and no indirect weak correlation between stress and vowel duration.

A second possibility is that these differences reflect variation in the participant profile for the two language family cohorts. Initial exploration of background variables reveals a small but significant difference in the onset of their learning English, with the Dravidian speakers on average starting at a slightly later age. This may explain the relative weakness of cue sensitivity in this cohort, compared with the Indo-Aryan speakers.

Besides these differences between the two IndE groups, a couple of similarities are noteworthy. Firstly, both groups clearly show much greater sensitivity to vowel quality cues than to any other cue, and this is also true for SSBE. Secondly, for both groups pitch ranks as the weakest (or, for Dravidian speakers, the joint weakest) cue, in contrast with the SSBE cohort, for whom it is weak but not the weakest. One possible explanatory avenue to explore might lie in pan-L1 interference from the preponderance of pitch accents in Indian languages (higher tonal density, low pitch word-initially).

In addition to further exploration of background variables, a next step will be to analyse production data for the same participants and to investigate whether these cue strengths and hierarchies are replicated in their own speech (for example the use of vowel reduction to signal prosodic prominence).

5. REFERENCES

[1] Culter, A. 2005. Lexical stress. In: Pisoni, D., Remez, R. (eds) *The Handbook of Speech Perception.* NY, USA: Blackwell Publishing, 264-289.

[2] Lehiste, I. 1970. *Suprasegmentals*. Cambridge MIT Press.

[3] Beckman, M. 1986. *Stress and non-stress accent*. Dordrecht: Foris.

[4] Sluijter, A., van Heuven, V. 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. Proceedings of *Proceedings of the International Congress of Spoken Language Processing,* Philadelphia, PA, 630-633.

[5] Okobi, A. 2006. *Acoustic correlates of word stress in American English*. Unpublished PhD dissertation, MIT.

[6] Mennen, I., Kelly, N., Mayr, R., Morris, J. 2020. The effects of home language and bilingualism on the realization of lexical stress in Welsh and Welsh English. *Front Psychol.* Jan 22. Doi: 10.3389/fpsyg.2019.03038.

[7] Werker, J., Tees, R. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development* 7(1), 49–63.

[8] Best, C., Tyler, M. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, O.-S., Munro, M. (eds), *Language experience in second language speech learning: In honour of James Emil Flege*. John Benjamins, 13–34.

[9] Flege, J., MacKay, I. 2004. Perceiving vowels in a second language. *Studies in Second Language Acquisition* 26, 1-34.

[10] Tremblay, A. 2021. The past, present, and future of stress in second-language word production and recognition. In: Wayland, R. (ed) *Second language speech learning: Theoretical and empirical progress*. Cambridge: Cambridge University Press, 175–192.

[11] Chrabaszcz, A., Winn, M., Lin, C. Y., Idsardi, W. J. 2014. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *J. Speech. Lang. Hear. Res.* 57, 1468–1479.

[12] Census of India. 2011. Registrar General and Census Commissioner of India.

[13] Keane, E. 2006. Prominence in Tamil. *Journal of the International Phonetic Association*, 36(1), 1-20.

[14] Mehrotra, R. C. 1965. Stress in Hindi, *Indian Linguistics,* 26, 96-105

[15] Bansal, R.K. 1970. A phonetic analysis of English spoken by a group of well-educated speakers from Uttar-Pradesh. *CIEFL Bulletin (Hyderabad)* 8, 1-11.

[16] Wiltshire, C. 2005. The "Indian English" of Tibeto-Burman language speakers. *English World-Wide* 26(3), 275–300.

[17] Wiltshire, C. 2020. *Uniformity and Variability in the Indian English Accent*. (Elements in World Englishes). Cambridge: Cambridge University Press.

[18] Maxwell, O., Fletcher, J. 2009. Acoustic and durational properties of Indian English vowels. *World Englishes* 28(1), 52-70.

[19] Masica, C. 2005. *Defining a linguistic area: South Asia*. New Delhi: Chronicle Books.

[20] Khan, S. D. 2016. The intonation of South Asian languages. *FASAL-6*, Umass Amherst.

[21] Sirsa, H., Redford, M. 2013. The effects of native language on Indian English sounds and timing patterns. *Journal of Phonetics* 41(6), 393-406.

[22] Wiltshire, C., Harnsberger, J. 2006. The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. *World Englishes* 25(1), 91-104.

[23] Pandey, P. 2016. Indian English prosody. In: Leitner, G., Hashim, A., Wolf, H.-G. (eds) *Communicating with Asia*, CUP: Cambridge, 56-68.

[24] Lisker, L., Krishnamurti, Bh. 1991. Lexical stress in a 'stressless' language: judgments by Telugu- and English-speaking linguists. *Proceedings of ICPhS1991*, Université de Provence, 90–93.

[25] Gargesh, R. 2004. Indian English: Phonology. In: Schneider, E., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (eds) *A handbook of varieties of English: A multimedia reference tool* (Vol. 1). Berlin, Germany: Mouton de Gruyter, 992-1002.

[26] Moon, R. 2002. *A comparison of the acoustic correlates of focus in Indian English and American English*. Unpublished Master's thesis, University of Florida.

[27] Peirce, J. 2009. Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.* 2, 10.

[28] Balusu, R. 2001. Acoustic correlates of stress and accent in Telugu. *21st South Asian Languages Analysis Roundtable*, University of Konstanz, 7-19.

[29] Asher, R. E. & Kumari, T. C., 1997. *Malayalam*, Routledge Descriptive Grammars, London: Routledge.

[30] Narayan, S., Mahesh, S., Veeramani, P. 2019. Acoustic cue for stress perception in Kannada speaking children. *Research and Reviews: Journal of Neuroscience* 9(3), 16-23.

[31] Savithri S.R. 1999. Perception of word stress. *Proceedings of the Madras India Regional Conference of the Acoustical Society of America* (2). 110–113.

[32] Terzenbach, L. 2011. *Malayalam prominence and vowel duration: listener acceptability*. Unpublished Masters dissertation, UT, Austin.

Phonetic integration in Heritage Korean language-mixing

Nicoline Butler

University of Hawai'i at Mānoa
butlern@hawaii.edu

ABSTRACT

Phonetic integration, the application of native phonological features to non-native utterances, is frequently operationalized to differentiate between loanwords and other categories of language mixing. Amongst bilingual communities however, recent studies have reported variability of phonetic integration across all types of language mixing, arguing that factors such as phonological segment serve as stronger predictors of integration.

This study analyzes social and linguistic factors which motivate phonetic integration amongst Heritage Korean-English bilinguals and explores the interaction between phonetic integration and speaker identity. Counter to previous studies, a linear regression model shows language mixing type to have the greatest effect on integration, with social factors such as speaker generation, language preference and cultural identity also surfacing as significant. These findings suggest that rather than merely linguistic diagnostic, phonetic integration can also function as an indicator of the unique identity and culture of bilingual speakers.

Keywords: phonetic integration, language contact code-switching, Heritage Korean, corpus linguistics

1. INTRODUCTION

Language mixing, the alternation of two languages within a single discourse, sentence, or constituent [1], is a common language practice among bilinguals, and has received attention not only for how it is used to signal identity, but also for its potential to drive language change. Within bilingual research, the distinction between established loanwords and other spontaneous language-mixing has been particularly polarizing, and phonetic integration, the application of recipient language phonology to donor language words, has frequently been operationalized as a key diagnostic.

In stark contrast to these long-held assumptions, recent studies have demonstrated that phonetic integration is not a reliable metric for delineation. They report that, in fact, that amongst bilingual French-English speakers of Quebec, phoneme itself is the most significant predictor of phonetic integration [2].

These findings lead to interesting questions about the role of phonetic integration in language mixing strategies:

i. Does the primacy of phoneme in integration apply to other bilingual language-mixing communities or is it exclusive to French-English speakers of Quebec?

ii. are there other linguistic and social features not considered in previous studies that factor into integration? If so, how significant is their contribution to this phenomena?

Poplack and colleagues themselves stress that patterns of phonetic integration should not be assumed, but rather "established segment by segment, word by word, speaker by speaker, [and] community by community" [2, p. 152]. While integration may, in some cases, be used to signal loanword status, it is likely that other factors, both linguistic and social, condition phonetic integration amongst bilingual speech communities.

This study endeavors to pursue these answers by investigating the phonetic integration of bilingual Heritage Korean speakers of the Greater Toronto Area. Applying methods of quantifying integration from previous research to sociolinguistic interviews with bilingual speakers of Heritage Korean, this study explores what social and/or linguistic factors may condition phonetic integration across English language-mixing types.

We begin by providing an overview of phonetic integration and then outline language-mixing habits amongst heritage bilingual communities reported in previous work. We then present the data, design, and results of the current study. In discussing the methodology used to address the above questions, this paper also highlights challenges inherent to investigating heritage languages as a distinct variety, rather than simply as an extension of their homeland counterpart.

2. BACKGROUND

2.1. Language Mixing and Phonetic Integration

Language mixing strategies of bilingual speakers have long-held the attention of linguists. While qualitative discussions of language mixing behaviors and their influence on language variation are numerous, quantitative analyses supporting these findings, particularly of the phonetic variety, are more difficult to come by.

Previous investigations have demonstrated that phonetic integration in language-mixing exhibits

significant variability [3, 2], and can function independently from morphosyntax of the same phenomena [1, 5, 2]. Of most significance to the current study are the findings of Poplack *et al.*, who assert that the assessment of integration at any level "depends of the successful identification of conflict sites - elements that can be unambiguously be characterized as belonging to [the donor] or [recipient language]" [2, p. 133].

In their work on French-English language mixing Québécois bilinguals, they report variability across all factors of analysis. They summarize that "intra-individual variation overwhelms inter-individual variation", and that conflict phonemes themselves are the most significant factor, each exhibiting different amenability to integration.

2.2. Heritage Korean and language mixing

Discussions on the contact between English and Korean have been widely represented in the literature. As one of the first languages analyzed in loanword phonology, Korean has been a popular choice for investigating the phonological processes involved in adapting donor tokens into the recipient language [6, 7, 8]. Additionally, there have been numerous studies on Korean-English code-switching, with particular emphasis on its use in constructing identity and positioning oneself vis-à-vis their interlocutor [see 9, 10, and 11]. Previous studies on linguistic integration and language mixing have also found that a speaker's cultural identity and generational heritage were the most significant predictors of language mixing [12].

An investigation into phonetic integration in Heritage Korean-English language mixing provides an opportunity to examine the findings of previous research within the context of a bilingual community consisting of languages with different societal status in the country. Previous studies have focused on bilingual speakers of French and English, both official languages of Canada which enjoy relatively similar prestige. It could be argued that speakers who display a working proficiency of both English and French enjoy higher prestige than their monolingual counterparts, due to the socioeconomic benefits afforded to bilinguals. In contrast, Heritage Korean speakers do not experience the same level of social status when speaking Korean as they do when speaking English, and many interviews coded in this study included accounts of racial discrimination experienced when speaking Korean. Comparing the integration patterns of Heritage Korean speakers and French-English bilinguals could offer valuable insights into the relationship between phonetic integration and extralinguistic social factors.

3. MATERIALS AND METHODS

This study makes use of sociolinguistic interviews from the Heritage Language Documentation and Change Project, hereon HLVC [13]. Typically 30-40 minutes in length, these interviews were conducted in Korean between 2009 and 2010. During interviews, participants were asked questions pertaining to their cultural identity, social lives, and experiences based on questionnaires adapted from [14] and [15].

Interviews were transcribed in Hangul by various members of the HLVC project, with instances of language mixing indicated by a shift to English transcription. (see [12] for more details on the interview procedures and transcription procedures).

Twenty-eight interviews were included in the study, and instances of language mixing were segmented and coded in ELAN (see Table 1 for demographics).

Gen.	Korean	Kor-Can	Canadian	*total*
1st	6	8	1	*15*
2nd	2	7	4	*13*
total	*8*	*15*	*5*	*28*

Table 1: Speakers included in study by generation of immigration and cultural orientation. Ages range from 39-85 (avg. 60) in first generation speakers and 18 to 42 (avg. 27) in second.

3.1. Quantifying Integration

Phonetic integration was quantified using conflict phonemes, sounds that do not occur in the recipient language and are easily identifiable as belonging to the donor language [16]. The fricatives /z v f θ, ð, and s/ which occur in English are not native to Korean phonology, and the application of Korean phonological processes, when adapted, is predictable (illustrated in Table 2).

PHONEME	ENGLISH DONOR		KOREAN ADAPTATION	
/z/ → [ts]	/zig.zag/	zigzag	[tsi.kɨ.tsæ.kɨ]	지그재그
/v/ → [p]	/vik.to.ri/	victory	[pik.to.ɹi]	빅토리
/f/ → [pʰ]	/al.fa.bet/	alphabet	[ɑɭ.pʰɑ.pɛt]	알파벳
/θ/ → [tʰ]	/kaθ.lik/	Catholic	[kʰɑ.tʰoɭ.ɭik]	가톨릭
/ð/ → [t]	/ɹi.ðum/	rhythm	[ɹi.tɨm]	리듬
/s/ → [ʃ]	/bei.sik/	basic	[pɛ.i.ʃik]	베이식

Table 2: Phonological processes by which conflict phonemes are adapted into Korean.

This predictability allows a clear coding system for instances of language mixing, where instances containing [f, v, θ, ð, z and s] are assigned a zero (*unintegrated*), and instances in which they are not are assigned a one (*integrated*).

3.2. Categorizing Language Mixing

Poplack *et al.* [2] identifies three types of language-mixing: established loanwords, nonce borrowings, and codeswitches. In order to better suit an investigation of Heritage Korean, delineations for each have been adapted as follows:

(1) *loanwords*: words adopted into the language.

(2) *nonce borrowings*: single speaker borrowings.

(3) *codeswitching* : multi-word switches.

In comparison to speech communities in previous studies, identifying established loanwords in Heritage Korean is less straightforward. The use of existing Korean dictionaries as the diagnostic for determining the status of Heritage Korean loanwords is reductive, possibly perpetuating the false narrative that homeland Korean is 'proper' Korean, and limits our understanding of Heritage speakers. For this reason, loanwords in this study are considered "adopted into the language" when they have been uttered more than once by more than one speaker across all interviews (1).

Additionally, nonce borrowings (2) are extended from "single-word utterance used once by a single speaker", to any single word used by a single speaker, regardless of frequency. This adjustment is made to account for the style of interview utilized in the HLVC, which frequently results in the repetition of certain nonce-borrowings. The definition of code-switching given in (3) remains unchanged.

3.3. Coding and factors

Approximately 14.5 hours of transcribed data was coded, resulting in a total of 2642 unique instances of language mixing. To produce a multi-use dataset, each of these instances were coded for language mixing-type and grammatical category, and well for heritage generation, language preference, and cultural orientation. 943 language-mixing utterances containing the fricatives /f, v, θ, ð, z and s/ were selected extracted for analysis. These tokens were also coded to include integration, conflict phoneme, and segment position.

4. DATA ANALYSIS

Of the 943 coded tokens, 295 phonemes were phonetically integrated to Korean, resulting in a mean of 33 tokens per speaker and a total integration average of 30.7%.

While working with naturalistic spontaneous speech has many benefits, it also poses difficulties for creating balanced data sets. Amongst the interviews selected for this study, not all speakers produced every target phoneme, /θ/ and /ð/ being the most difficult to come by (13 and 18 interviews respectively). Speakers do, however, all exhibit the ability to produce both integrated and unintegrated conflict phonemes.

TYPE	θ	ð	v	z	s	f	TOTAL
loanwords	0	101	109	18	61	42	*331*
nonce borrowing	5	124	81	38	86	38	*372*
codeswitches	41	59	50	21	61	8	*240*
TOTAL	*46*	*77*	*88*	*208*	*240*	*284*	*943*

Table 3: Distribution of data by language-mixing type and phoneme.

Across tokens, utterances containing /f/ occurred most, while /θ/ occurred least. When looking at integration, /s/ is integrated most and /θ/ is integrated least. The distribution of each conflict phoneme is outlined in Table 3.

4.2. Linguistic factors

Language-mixing type: Also presented in Table 3 is the total number of tokens observed in each language mixing type, as well as their occurrence with each conflict phoneme. Most frequent was nonce borrowings, representing approximately 39% of all data, while codeswitching occurred the least, representing only 25% approximately. Phonemes occurring in loanwords were integrated most often (50.4% of all realizations), while phonemes contained in codeswitches integrated least often (only approx. 13.7%).

Position of phoneme: Amongst the phonemes observed in this dataset, most occurred word-medially (406 phonemes), some occurred word-initially (312 phonemes), and phonemes occurring word-finally were the most rare (225 phonemes). Word-medial phonemes also integrated the most (36.2% of instances), while word-initial phonemes integrated the least (22.8% of instances).

Grammatical category: Grammatical category is by far the most disproportionate of all factors. Conflict phonemes occurring in nouns represent approximately 64% of all data, while sentences represent only around 25%. All other categories including adjectives and verbs each represent 6% or less of the remaining data. Phonemes occurring in nouns also integrated the most (37.4%), while phonemes occurring in sentences integrated the least (14.7%).

4.3 Social factors

Generation: Despite the higher number first generation speakers, second generation speakers made up a larger proportion of the data (55.5% of all instances) and exhibited a higher number of tokens per speaker. Conversely, despite lower total instances of language mixing (44.5%; avg. of 28 tokens per speaker), first generation speakers integrate much

more frequently than second (47.4% and 14.9% respectively).

Cultural orientation: While the largest number of speakers identified as Korean-Canadian, speakers who identified as Canadian and Korean both display a higher average number of tokens per speaker (39 and 37 tokens respectively). In addition, Korean speakers integrated the most frequently (47.4% of the time), while Canadian speakers integrated the least (14.9% of the time).

Language Preference: Although most speakers indicate a preference for speaking English or Korean, speakers who have no preference (both) tended to integrate the most (45.6% of the time).

5. REGRESSION ANALYSIS

Logistic regression modeling [17] was used to test the comparative influence of each factor on phonetic integration. Anova and *phia* [18] were used to conduct post-hoc analyses. The results of these tests are summarized in Table 4.

Mixing Type (p < .000)	[DF]	[chi.sq.]	[Pr(>|z|)	
nonce borrowing	1	15.56	< .000	***
codeswitch	1	12.72	< .001	***
loanword	1	0.00	0.949	
Language Preference (p < .000)				
English	1	20.38	< .000	***
Korean	1	30.03	< .000	***
both	1	**3.30**	0.069	
Speaker Generation (p < .000)				
1st	1	2.86	0.091	
2nd	1	39.55	< .000	***
Conflict Phoneme (p < .000)				
f	1	31.23	< .000	***
ð	1	10.58	0.004	**
θ	1	9.63	0.005	**
v	1	5.87	0.016	*
s	1	6.98	0.016	*
z	1	16.37	< .000	***
Cultural Identity (p = .02)				
Korean-Canadian	1	22.20	< .000	***
Canadian	1	22.33	< .000	***
Korean	1	8.86	0.003	**

Table 4: results of fixed effects of logistic regression model. glm (integration ~ 1 + mixing type + generation + phoneme + language preference + cultural identity + segment position + grammatical category, family = "binomial")

Based on this model, mixing-type and language preference appear to be particularly strong predictors of integration, with second generation speakers and those who identify as Canadian also showing higher levels of integration. No significant effects were found for grammatical category or segment position. Overall, these results suggest that the motivations behind phonetic integration amounts Heritage Korean speakers are complex and multifaceted.

6. DISCUSSION

The results of this study confirm that phonetic integration is not realized uniformly across language-mixing types. However, extra-linguistic factors such as language preference and speaker generation demonstrate a more pronounced effect on integration. The significance of these social factors challenges the assertions of previous studies, which suggested that "extralinguistic characteristics do not clarify the motivations behind phonetic realization" [2, p.147].

The study also raises important questions about the complex relationship between language, culture, and identity. While phonetic integration may be a matter of choice for many bilingual speakers, for Heritage speakers this choice may hold more consequences beyond mere ease of speech.

Pronunciation has long been held as the strongest hallmark of proficiency, and for Heritage speakers, the decision to integrate may be strongly motivated by how their proficiency will be perceived by interlocutors. This presents an interesting but complex conundrum: Do speakers measure their own proficiency to decide on language preference and/or cultural identity? What other criteria are considered when making this decision? Heritage language and culture in the diaspora is an incredibly complex and nuanced topic that undoubtedly goes beyond the scope of this paper, but highlights a need for future research on the nature of language contact and change in multilingual communities

This study is not without its limitations. Phonetic integration in Korean-English language mixing cannot be fully captured by diagnostics of this study. To more accurately account for phonetic integration as a whole, future studies could explore a more complex diagnostic that considers vowels, syllable structure, and suprasegmental features. Additionally, future studies could also consider speaker proficiency and the possibility of L1 intrusions in 1st get speakers [19]. Finally, another area of interest wound be to observe the effect of phonetic convergence and the cultural identity of the interlocutor on an individual's propensity to integrate.

Overall, this study provides a preliminary account of phonetic integration amongst speakers of Heritage Korean in the Greater Toronto area. Results speak to an interaction between social factors and language-mixing behaviors, and reinforce the long-held understanding that language does not exist in isolation, but rather is intricately interwoven with culture and identity.

7. REFERENCES

[1] Poplack, S. 1988. Contrasting patterns of code-switching in two communities. Codeswitching: anthropological and sociolinguistic perspectives 48, 215-244.

[2] Poplack, S., Robillard, S., Dion, N., Paolillo, J. 2020. Revisiting phonetic integration in bilingual borrowing. Language 96(1), 126-159.

[3] Fought, C. 2010. Ethnic Identity and Linguistic Contact. In The Handbook of Language Contact, Wiley-Blackwell, pp. 282–298.

[4] Bullock, B. E. 2009. Phonetic reflexes of code-switching. In B. E. Bullock, & A. J. Toribio (Eds.), The Cambridge handbook of linguistic code-switching (pp. 163–181). Cambridge: Cambridge University Press.

[5] Muldner, K., Hoiting, L., Sanger, L., Blumenfeld, L., and Toivonen, I. 2007. The phonetics of code-switched vowels. International Journal of Bilingualism, vol. 23, no. 1, pp. 37–52, Jun. 2017.

[6] Kenstowicz, M. 2005. The phonetics and phonology of Korean loanword adaptation. In: Rhee, S.-J. (ed), Proceedings of the First European Conference on Korean Linguistics. Seoul: Hankook Publishing Co., 17-32.

[7] Kang, Y. 2010. The emergence of phonological adaptation from phonetic adaptation: English loan-words in Korean. Phonology 27(1), 225-253.[8] Kent, D. B. 1999. Speaking in tongues: Chinglish, Japlish and Konglish. *KOTESOL Proceedings of PAC2.* Seoul, 197-209.

[8] Lee, D., Kim, D., Yun, S., Kim, S. 2021. Phonetic variation modeling and a language model adaptation for Korean English code-switching speech recognition. Applied Science 11.

[9] Kent, D. B. 1999. Speaking in tongues: Chinglish, Japlish and Konglish. KOTESOL Proceedings of PAC2. Seoul, 197-209.

[10] Kang, A. 2003. Negotiating conflict within the constraints of social hierarchies in Korean American discourse. Journal of Sociolinguistics 7(3), 299–320.

[11] Song, J. 2009. Bilingual creativity and self negotiation: Korean–English bilingual children's language socialization into address terms. In: Reyes, A., Lo, A. (eds), Beyond Yellow English: Towards Linguistic Anthropology of Asian Pacific America. New York: Oxford University Press, 213–232.

[12] Chung, S. 2010. Code-switching as a marker of cultural identity among Koreans in Toronto. LIN 497. University of Toronto: unpublished assignment.

[13] Nagy, N. 2009. Heritage language and variation change in Toronto. http://individual.utoronto.ca/ngn/research/heritage_lgs.htm#objectives

[14] Keefe, S., Padilla, A. 1987. Chicano ethnicity. Albuquerque, NM: University of New Mexico Press.

[15] Labov, W. 1984. Field methods of the project on linguistic change and variation. University of Pennsylvania.

[16] Poplack, S., Meehan, M. 1998. How languages fit together in codemixing. International Journal of Bilingualism 2, 127-138.

[17] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[18] De Rosario-Martinez, H. (2015). phia: Post-Hoc Interaction Analysis. R package version 0.2-1. https://CRAN.R-project.org/package=phia.

[19] Kirkham, S. 2017. "Ethnicity and phonetic variation in Sheffield English liquids" in Journal of the International Phonetic Association, 47(1): 17—35

THE ROLE OF PSYCHOTYPOLOGY IN THE ACQUISITION OF RHOTIC CONSONANTS BY ADOLESCENT EMERGENT MULTILINGUALS

Iga Krzysik

Adam Mickiewicz University, Poznań
iga.krzysik@amu.edu.pl

ABSTRACT

Psychotypology has been postulated as one of the factors shaping multilingual acquisition and affecting cross-linguistic influence across various domains of language, including phonetics and phonology. Despite its explanatory potential, psychotypology was rarely quantified [1], [2], [3] and correlated with phonological data. To investigate the relationship between psychotypology and multilingual acquisition of speech, this exploratory study assessed the perception of language distance by 20 adolescent emergent sequential multilinguals (L1 Polish, L2 English, L3 German) through the ViLDiM measure [4]. The results were correlated with the L2 and L3 speech production scores obtained in a delayed repetition task focused on rhotic consonants. Although there was no statistically significant relationship between the L2 and L3 rhotic realisations and perceived language distance, possibly meaningful interactions were observed on an individual case basis.

Keywords: multilingual phonology, multilingual acquisition, psychotypology.

1. INTRODUCTION

Psychotypology is a term proposed by Kellerman [5], [6] describing individual perception of distances between languages. Psychotypology focuses on the perceived differences and similarities determining language proximity and, unlike typology, it is considerably more fluid, subjective and prone to change throughout language learning [7]. Psychotypology has been positioned as a significant factor in the process of bilingual and multilingual acquisition. Kellerman [5], [6] recognised the potential relationship between the perceived language distance in the likelihood of cross-linguistic influence (CLI) from L1 to L2. Odlin [8] suggested that individual perception of relations between linguistic structures may affect the probability of CLI. De Angelis [9] acknowledged the influence of psychotypology on CLI patterns in multilingualism. In the Typological Primacy Model, Rothman [10], [11], [12] postulated that typology and psychotypology influence the perception of language proximity by the parser, which in turn affects the characteristics of CLI in the process of multilingual acquisition.

Despite the assumptions concerning the role of psychotypology in language acquisition, a considerable number of previous studies did not quantify it as a variable (e.g. [5] [6] [7], [13], [14]). Other studies measured psychotypology through questionnaires, think-aloud protocols or magnitude estimations (e.g. [1], [2], [3]), relaying predominantly on a binary understanding of the relationship between the languages. This issue was addressed by Nelson et al. [4], who devised a measure of language distance by representing all languages in the repertoire within the same visual space.

The relationship between psychotypology and multilingual phonology has rarely been explored. In their case studies of phonological development, Williams and Hammarberg [15] and Hammarberg [16] indicated that speakers who detect similarities and parallels between the languages experience some facilitative influence. Wrembel [2] investigated the cognitive processes behind metalinguistic awareness related to phonological performance in a third language. The results indicated that typology and psychotypology had a particularly strong presence in the cross-linguistic awareness comments of the participants and might have a substantial influence on their performance.

Psychotypology might be positioned as one of the crucial factors associated with multilingual development. However, the studies of its role are in an early, exploratory phase. The current study aims to quantify psychotypology and examine its role in the acquisition of rhotics in multilingual speech production data. Moreover, through the application of an innovative visual measure of language distance, it intends to account for the dynamic perception of the relationships between the languages in the multilingual mind.

2. METHOD

2.1. Aims

The present study intends to examine the relationship between psychotypology and the acquisition of

multilingual speech, with the former operationalised as ViLDiM scores and the latter as a target or non-target-like renditions of rhotic consonants in the L2 English and L3 German of the participants. Consequently, the research question posed in the study is as follows: what is the relationship between psychotypology and speech production of adolescent multilinguals?

It is postulated that the participants who perceive a given pair of languages as closer, might display related cross-linguistic influence (CLI) in the production of rhotic consonants. For instance, the participants who see L1 Polish and L3 German as closer might exhibit more L1-like realisations in L3 German, whereas the ones who perceive L2 English and L3 German as closer might have more L2-like realisations in L3. However, it should be noted that long exposure to a given language may affect psychotypology; in more experienced learners the distance between the languages might decrease due to the increasing language familiarity.

The focal features in this study include rhotic consonants in L2 and L3; their diversity might be particularly interesting in the context of studying CLI in multilingual speech acquisition and its potential interaction with psychotypology. The realisations of rhotic consonants differ across the three languages tested in the present study In Polish, it is predominantly realised as an alveolar trill [17] with possible realisations as an alveolar flap [18]. In L2 English (General British), a rhotic consonant is realised as a post-alveolar approximant [19]. In L3 German (Standard German), the realisations are quite diverse including uvular trill, uvular fricative or a vocoid in the syllable-final position [20] [21].

2.2. Participants

The participants were 20 adolescent emergent multilinguals (mean age=12.65, *SD*=0.48, male to female ratio – 10:10) with L1 Polish, learning L2 English (age of language learning onset=7) and L3 German (age of language learning context=12) in the context of formal instruction in a Polish public primary school. The participants had the same amount of language classes and were taught by the same L2 and L3 teachers. They had also the same order of L2 and L3 acquisition and had no additional languages in their repertoire. The participants were tested individually in a quiet room in a school library. The administered tasks are described in the following sections.

2.3. Psychotypology task

Psychotypology was assessed through the Visual Language Distance Measure (ViLDiM) [4]. The measure employs a visual paradigm to illustrate the perception of distances between the languages. ViLDiM utilises a dotted sheet of A3 paper and transparent circles marked with the names of the languages. Such a framing of language distances accounts for the multidimensional and dynamic nature of the relationships between the languages in the multilingual repertoire. The participants are requested to arrange the circles on the sheet of paper to indicate the perception of distance between the languages. In the current study, the participants were specifically instructed to pay attention to the way the languages in their repertoire sound. The raw distances between the circles were measured and normalised following the formula outlined by Nelson [4].

$$NV_{L1-L2} = \frac{RV_{L1-L2}}{RV_{L1-L2} + RV_{L1-L3} + RV_{L2-L3}} \quad (1)$$

Figure 1: Score normalisation formula for psychotypology task (ViLDiM). NV- normalised distance, RV – raw distance (in cm).

2.4. Speech production task

The production of rhotic consonants was tested through a delayed repetition task, administered separately for L2 and L3. Before completing the task, the experimenter attempted to induce a given language mode [22] through a short conversation in L2 or L3 respectively.

Both language versions of the delayed repetition task included a sequence of dialogues, consisting of two sentences. The first sentence featured the target word with the embedded target feature (a rhotic consonant). The second sentence was an intervening material, intended to minimise the short-term recall and imitation to possibly access the actual representations of the features. To resemble a conversational pace, the interval between the sentences in each dialogue was 400 ms. The participants were requested to listen to the dialogues and recall the first sentence after the presentation of each sequence. The pace of the task was adjusted for each participant; after the reproduction of the first sentence, the experimenter moved to the next one.

Each language version included 6 target words with an embedded rhotic consonant (in the initial or medial position). The responses of the participants were recorded through a dynamic microphone plugged to a laptop via an external audio interface (16bit, 44.1 kHz).. To control for rhotacism, the L1 Polish production was recorded and analysed by the experimenter.

3. RESULTS

3.1. Psychotypology – perception of language distances

The normalised measurements of language distance indicated that the participants perceived the L2 English-L3 German pair as the closest in their repertoire (M=0.28, SD=0.09), followed by L1 Polish-L2 English (M=0.34, SD=0.08) and L1 Polish-L3 German (M=0.38, SD=0.08) as the most distant. Table 1 presents the normalised measures of language distance.

Lg pair	N	95% CI	M	SD	min	max
L1-L2	20	0.306-0.381	0.34	0.08	0.20	0.50
L1-L3	20	0.339-0.418	0.38	0.08	0.21	0.49
L2-L3	20	0.234-0.320	0.28	0.09	0.15	0.50

Table 1: The normalised measurements of language distance.

3.2. Delayed repetition task in L2 and L3

The recordings of rhotics production obtained through L2 and L3 delayed repetition tasks were rated by three independent raters per language, near-native speakers of either English or German with L1 Polish who completed phonetics and phonology courses during their BA studies. The raters received detailed instructions concerning the characteristics of rhotic consonants in the three languages of the participants and a list of rating categories. The production of the participants was rated according to the following categories specified in the rating instructions: (1) target-like realisation, (2) L1 realisation, (3) L2/L3 realisation, (4) a hybrid realisation, (5) realisation as a non-rhotic sound, (6) unintelligible pronunciation. Each rater analysed a total of 120 items per language. Inter-rater reliability (IRR) was computed according to a solution suggested for nominal datasets where each participant is rated by the same raters (more than 2) [23]. Cohen's Kappa was computed for each rater pair (rater 1 - rater 2, rater 1 – 3, rater 2 – rater 3) and subsequently averaged to produce an overall estimate of IRR for each language. The resulting kappa for the L2 English rating indicated a moderate agreement (approaching the benchmark for substantial), k=.609. The kappa for L3 German indicated a substantial agreement, k=.692 between the raters.

The rating of L2 production indicated that L2 target-like realisations were dominant (67.50%), with the presence of L1 Polish CLI in L1/L2 hybrid realisations (22.50%). L1-like realisations were rather infrequent (6.67%), followed by unintelligible realisations (3.33%), see Fig. 1.

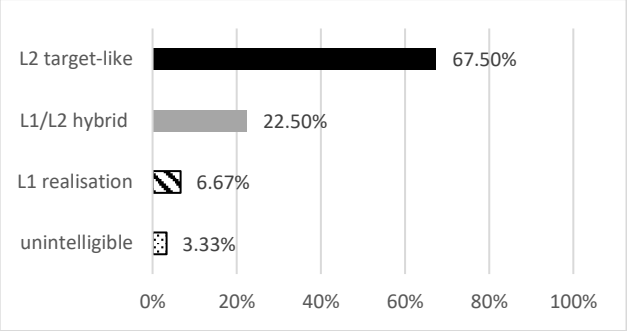


Figure 1: The realisations of L2 rhotic consonants.

L3 production was dominated by L1 realisations of rhotic consonants (33.33%), with frequent replacements of the rhotic consonants with non-rhotic ones (/v/ and /x/ in particular) (21.67%) as well as unintelligible renditions (19.17%). 10.83% of realisations were characterised by L2 target-like production and 10.00% by L3 target-like. Occasional L2/L3 and L1/L3 hybrid realisations were observed (4.17% and 0.38% respectively), see Fig. 2.

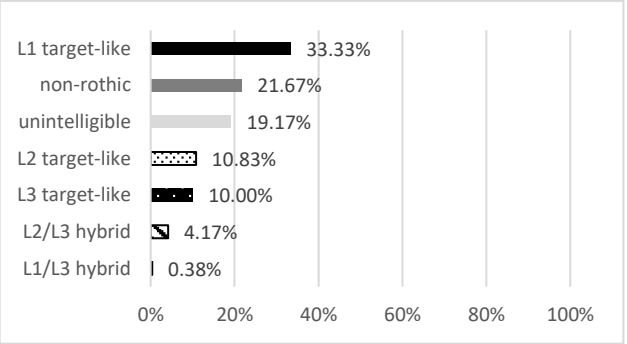


Figure 2: The realisations of L3 rhotic consonants.

3.3. Delayed repetition task in L2 and L3

The decisions about correlations between the particular scores of speech production and psychotypology were based on the tendencies observed in speech production. As for L2, the L2 target realisations were correlated with L1-L2 distance to see the directionality of the relationship between these variables. Pearson correlation coefficient indicated no relationship between the two variables (r=-0.53, N=20, p=0.82). Moreover, the hybrid L1/2 realisations were correlated with L1/L2 distance to investigate if the participants with higher number of hybrids perceived the two languages as closer. Pearson correlation coefficient indicated no relationship between the two variables (r=0.83, N=20, p=0.73). As for L3, the high number of L1-like realisations was correlated with the L1-L3 distance to investigate whether the participants with closer L1-L3 proximity produced more L1-like rhotics in L3.

Further, Pearson correlation coefficient indicated no relationship between the two variables (r=-0.93, N=20, p=0.69).

Despite the lack of significant relationships on a global level, some interesting relations between psychotypology and rhotic production were observed for individual cases. For instance, the two participants with the greatest distance between L1 and L2 (NV=0.500) demonstrated high L2 target-like production, which might suggest that for some speakers the perception of proximity might increase with proficiency. Conversely, the participant with the closest L1-L2 proximity exhibited a high percentage of hybrid realizations and errors (66.66%). In L3, the participant with the second closest L2-L3 proximity (NV=0.164) was characterized by a high percentage of L2/L3 hybrids and L2 realizations in L3 (66.66%). Such findings may indicate a relationship between CLI and L2-L3 perception. The participant with the third closest L1-L3 proximity produced L3 rhotics with strong L1 CLI and /v/ or /x/ replacements (0% of target-like realisations).

4. DISCUSSION

The study investigated the relationship between psychotypology and the acquisition of multilingual speech. Sequential adolescent multilinguals completed a ViLDiM task [4] to measure their perception of language distance as well as a delayed repetition task in their L2/L3 rhotics production. The results indicated that the participants perceived their L2 English and L3 German as the closest, followed by L1 Polish and L2 English and the most distant L1 Polish and L3 English. Despite different characteristics of rhotics, the L2 and L3 were perceived as more similar than L1-L2 and L1-L3 language pairs. Such results may be driven by the overall typology, genetic relatedness and the perception of both sounds as belonging to the domain of foreign language.

As far as multilingual speech production is concerned, the participants produced more target-like rhotics in the more proficient L2 than in the early-stages L3. The L2 rhotics were predominantly realised as postalveolar approximants and L1/L2 hybrids. The L3 realisations were substantially influenced by L1. Moreover, the high percentage of non-rhotic substitutions and unintelligible realisations indicates that L3 rhotics posed a challenge in perception and production.

The overall look at the relationship between psychotypology and multilingual acquisition revealed that the more proficient L2 was not considerably closer to L1 compared to the newly acquired L3. Consequently, the perception of language distance may not be affected by the length of the language learning experience. Additionally, the perceived closeness of L2 and L3 (despite a short period of L3 learning) supports the assumptions that phonological features of L2 and L3 are either perceived from a more typological perspective or classified as foreign and therefore assessed as close.

The correlations between psychotypology and rhotic production in L2 and L3 did not indicate any significant relationships. However, the individual cases revealed some links between CLI and multilingual speech production. Such findings correspond with the predictions concerning psychotypology by Bardel and Lindqvist [24]. They suggest that it is best observed and explored on an individual case basis rather than at a group level. However, it still would be worthwhile to extend the scope of the current study to a larger pool of participants and more tokens and tested features as well as perception tests to increase the possibility of revealing significant group tendencies. Additionally, different language repertoires and acquisition contexts could contribute to further understanding of these relationships.

5. CONCLUSION

The current study constituted one of the first attempts to investigate the relationship between quantified psychotypology and multilingual speech production. There were no significant correlations between psychotypology and rhotics production, however, there were some meaningful interactions between CLI and language distance on an individual level. Additionally, the findings indicate that psychotypology might be informed by typology and language status (native vs. foreign). Further research is needed to explore the relationship between psychotypology and multilingual acquisition of phonology from an individual and group perspective.

6. ETHICS STATEMENT AND ACKNOWLEDGEMENTS

Informed consent was obtained from the participants and their parents / legal guardians. The study was positively evaluated by the ethics committee at Adam Mickiewicz University, Poznań.

This research was funded by the National Science Centre, Poland, (grant reference number: 2017/27/N/HS2/00285).

7. REFERENCES

[1] Jessner, Ulrike. 2006. *Linguistic awareness in multilinguals. English as a third language*. Edinburgh: Edinburgh University Press.

[2] Wrembel, M. 2015. *In search of a new perspective: Cross-linguistic influence in the acquisition of third language phonology*. Poznań: Wydawnictwo Naukowe UAM.

[3] Xia, Chris M. 2017. Psychotypology of Chinese learners of English and its influence on the acquisition of metaphorical expressions: An online study, *Cambridge Occasional Papers in Linguistics* 101: 237-255.

[4] Nelson, Christina, Iga Krzysik, Halina Lewandowska and Magdalena Wrembel. 2021. "Multilingual learners' perceptions of cross-linguistic distances: a proposal for a visual psychotypological measure", *Language Awareness* 30, 2: 176-194.

[5] Kellerman, E. 1978. Giving learners a break: Native language intuitions as a source of prediction about transferability. *Working Papers on Bilingualism*, 15, 59–92.

[6] Kellerman, E. 1983. Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning*. Newbury House, , 112–134.

[7] Ringbom, H. 2007. *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.

[8] Odlin, T. 1989. *Language Transfer: Cross Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

[9] De Angelis, G. 2007. *Third or additional language acquisition.* Clevendon: Multilingual Matters Ltd.

[10] Rothman, J. 2011. L3 syntactic transfer selectivity and typological determinacy: the typological primacy model, *Second Language Research* 27, 1: 107-127.

[11] Rothman, J. 2013. Cognitive economy, non redundancy and typological primacy in L3 acquisition: Evidence from initial stages of L3 Romance", in: Baauw, S., Dirjkoningen, F., Pinto, M., Meroni, L. (eds.), *Romance languages and linguistic theory 2011,* Amsterdam: John Benjamins Publishing Company, 217-247.

[12] Rothman, J. 2015. Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered, *Bilingualism: Language and Cognition* 18, 2: 179-190.

[13] De Angelis, G. 2005. Interlanguage transfer of function words, *Language Learning* 55, 3: 379-414.

[14] Singleton, D., Ó Laoire, M. 2006. Psychotypologie et facteur L2 dans l'influence translexicale: une analyse de l'influence de l'anglais et de l'irlandais sur le français L3 de l'apprenant. Acquisition et interaction en langue étrangère, 24, 101–117.

[15] Williams, S., Hammarberg, B. 1998. Language switches in L3 production: Implications for a polyglot speaking model, *Applied Linguistics*, 19: 295-333.

[16] Hammarberg, B. 2001. Roles of L1 and L2 in L3 production and acquisition, in: Cenoz, J., Hufeisen, B., Jessner, U. (eds.), *Cross-linguistic influence in third language acquisition*: *Psycholinguistic perspectives*. Clevedon: Multilingual Matters, 21-41

[17] Gussmann, E. 2007. *The phonology of Polish*. Oxford: Oxford University Press.

[18] Jassem, W. 2003. "Polish", Journal of the International Phonetic Association 33, 1: 103-107.

[19] Cruttenden, A. 2014. *Gimson Pronunciation of English* (8th ed.). Oxon: Routledge.

[20] Boase-Beier, J., Lodge, K. 2003. *The German language: A linguistic introduction.* Oxford: Blackwell.

[21] Kopečková, R. 2016. The bilingual advantage in L3 learning: A developmental study of rhotic sounds, *International Journal of Multilingualism* 13, 4: 410-425.

[22] Grosjean, F. 1998. Studying bilinguals: Methodological and conceptual issues, *Bilingualism: Language and Cognition* 1: 131-149.

[23] Hallgren, A. K. 2012. Computing Inter-Rater Reliability for Observational Data. An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology* 8 (1): 23-34.

[24] Bardel, C., & Lindqvist, C. 2007. The role of proficiency and psychotypology in lexical cross-linguistic influence. A study of a multilingual learner of Italian L3, *Atti del VI Congresso Internazionale dell'Associazione Italiana di Linguistica Applicata*, 2007, 123-145.

REVISITING THE VOWELS OF HONG KONG ENGLISH – THE POST-HANDOVER GENERATION

Hoi Ling Chan & Robert Fuchs
University of Hamburg
hollychan0603@gmail.com, robert.fuchs@uni-hamburg.de

ABSTRACT

Previous research investigating the vowels of Hong Kong English (HKE), a postcolonial variety of English, has suggested mergers between several pairs of vowels. This evidence stems from speakers who grew up before the 1997 handover of Hong Kong to the People's Republic of China. Given the enormous political and social changes since then, we investigate whether we can detect recent diachronic change in the vowels of HKE.

Wordlist vowel data (in carrier phrases) sampling 14 lexical sets from 20 L1 Cantonese-speaking university students was analysed with regard to F1 and F2 (Lobanov-normalised) and duration. Potential vowel mergers were assessed through Support Vector Machines (SVM). Results confirm previous research with regard to mergers between DRESS-TRAP, THOUGHT-LOT and FLEECE-KIT due to great overlap in F1, F2 and duration. However, GOOSE and FOOT are relatively distinct, in contrast to previous research, possibly indicating recent diachronic change.

Keywords: Vowels, monophthongs, Hong Kong English, diachronic change, World Englishes

1. HONG KONG AND HONG KONG ENGLISH

The territory of Hong Kong (HK) was ceded in the 19th century to British colonial rule and subsequently saw rapid economic development and population increase [1]. Its current population of 7 million consists mainly of Cantonese speakers, but also speakers of American and British English, a long-standing South Asian community as well as more recent Filipino, mostly non-permanent, residents [2]. In 1997, authority over HK was handed over to the People's Republic of China, which has ruled it since then as a Special Administrative Region with a degree of autonomy in local matters.

As in other colonial contexts, a local second language (L2) dialect of English emerged in HK over time, shaped by (i) cross-linguistic influence from the L1 Cantonese, (ii) general language learning mechanisms and (iii) the population structure during the early colonial period, with influence from other British colonies, e.g. in South Asia and Malaya (cf. founder effect; [3]).

With the 1997 handover, HK entered an era of profound political, cultural and linguistic realignment. British (English) influence may continue to be relevant due to its historical influence on HKE and in so far as it is ingrained in local education, while American (English) influence is commercially relevant. Perhaps the greatest force driving the development of HKE post-1997 may be internal dynamics (i.e. endonormative forces; [2-3]). A wealth of studies has documented the distinctiveness of HKE on the levels of syntax [4-5], pragmatics [6-7], lexis [8] and phonology [9-10].

Against these fundamental political changes, we ask whether there has been any recent diachronic change in HKE since 1997. For reasons of space, we focus on one particular area of phonology that is often investigated in sociophonetic research, monophthongs.

2. THE PHONOLOGY OF HKE

The present section summarises key results on the monophthongs of pre-handover HKE. Throughout the paper, Wells' lexical sets notation [14] is used, though we also provide each vowel's articulatory value in British Eng. at the first mention.

The absence of a contrast between several pairs of vowels in HKE has been reported based on impressionistic evidence [11]. This concerns FLEECE /i/ and KIT /ɪ/ as well as DRESS /e/ and TRAP /æ/. Moreover, the long (or tense) vowels THOUGHT /ɔ:/ and GOOSE /u:/ were claimed to be shortened in HKE.

Acoustic evidence presented by Hung [12, pp. 340] similarly found an absence of a distinction between FLEECE – KIT, DRESS – TRAP, GOOSE – FOOT and THOUGHT – LOT. Additional acoustic evidence by Deterding, Wong & Kirkpatrick [henceforth DWK] [13] broadly confirmed this assessment, with the modification that THOUGHT – LOT may not be "fully merged" and that GOOSE and FOOT showed substantial fronting.

These features are generally thought to derive from cross-linguistic influence from Cantonese [1], although this influence is not deterministic or

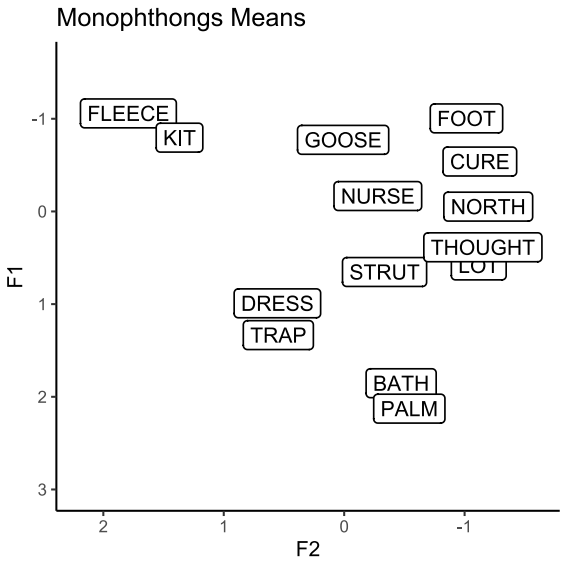


Figure 1. Mean position for all monophthongs.

straightforward. Individual speakers may have different conscious or unconscious orientations to emerging endocentric (local) pronunciation norms or different exocentric norms (such as British or American English).

3. AIMS AND RESEARCH QUESTIONS

To summarise, the existing acoustic evidence on the vowels of HKE comes from two studies, which rely on speakers who were born before the 1997 handover [12-13]. Against the fundamental cultural and political changes in HK since then, we ask whether the vowel system of HKE shows evidence of recent diachronic change. We will contrast the existing evidence, stemming from speakers born and raised before 1997, with new data from speakers who were born between 1995 and 2001. These speakers essentially lived their entire lives, and received their education, in post-independence HK. They are thus part of HK's post-handover generation.

More broadly, we also aim to provide an up-to-date acoustic analysis of the vowels of HKE. To this end, we will improve upon existing research by analysing vowel quantity (duration) alongside vowel quality (formants). Moreover, formant data will be normalised to control for anatomical variation in vocal tract length.

4. DATA AND METHODS

4.1 Speakers and Data

The 20 (13 f, 7 m) participants of this study were born and raised in HK, spoke Cantonese as their L1, were proficient in English and, at the time of

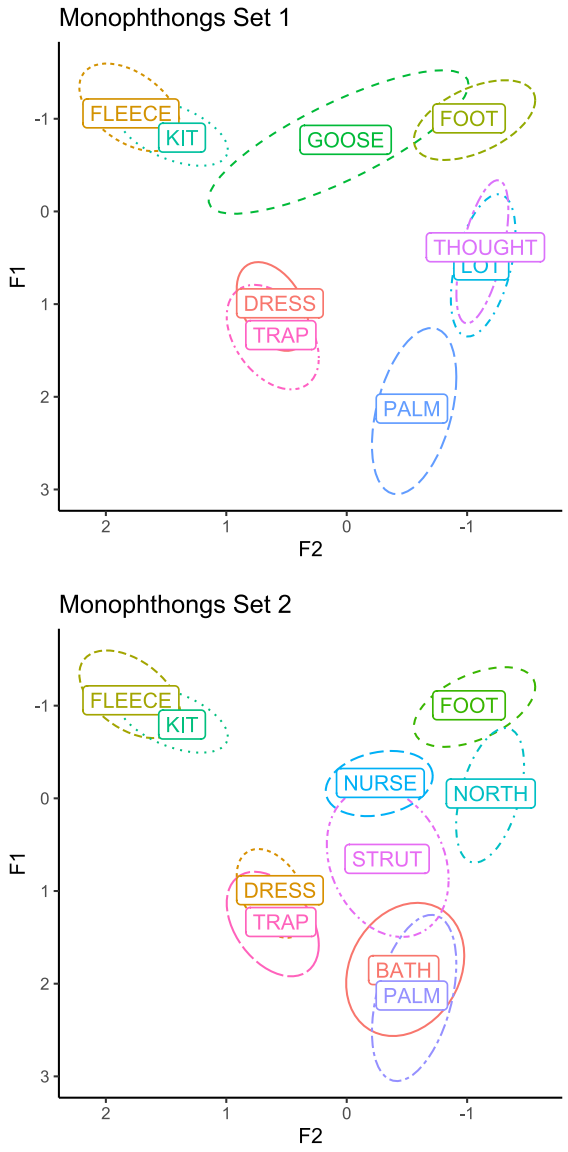


Figure 2a and b. Means and 66 % ellipses for all monophthongs, separated into two sets.

recording, were undergraduate students at several universities in HK (in English-medium programmes, with two exceptions). We elicited read speech in the form of a wordlist in carrier phrases and a ten-minute informal conversation between each participant and the interviewer (the first author, native to HK). For reasons of space, we here focus on the wordlist data (duration 55 - 139 s.). The wordlist items were chosen from Wells' lexical sets [14, pp. 127-169], supplemented by further English words. 14 lexical sets were elicited with at least two distinct keywords (except for three lexical sets, which were elicited only once). In total, each speaker produced 26 lexical items.

Recordings were made with a Zoom H1n handheld recorder. Participants filled in a questionnaire to provide linguistic and demographic

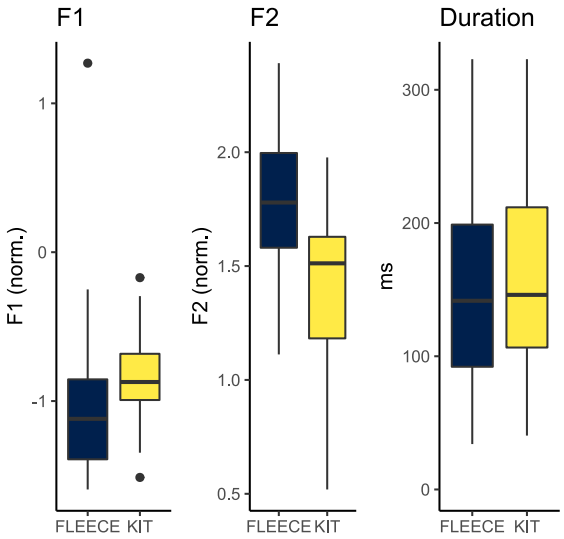


Figure 3. Distribution of F1, F2 and duration for FLEECE and KIT.

background information and consented to the research procedure prior to data collection.

4.2 Analysis

The recordings were phonemically annotated through forced alignment with Webmaus [15] and subsequent manual correction. Of the 520 recorded words, 10 were removed due to mispronunciations (mostly as initialisms). For the remaining items, duration as well as the first and second formants were extracted at 50 % and 80 % of vowel duration through a Praat script. Formants were then normalised with the vowels package in R, using the Lobanov method [16]. All vowel charts display normalised formant frequencies at vowel midpoint and were plotted with ggplot2 in R [17].

We further assess potential vowel mergers with the machine learning method Support Vector Machines (svm) in R (package e1071) [19]. For each potentially merged pair of vowels, F1, F2 and duration are used to predict the vowel in a binary classification task. Over 50 iterations, the data is each time split randomly into a training set (two thirds) and the remaining one third is held out as test set. We average the classification statistic kappa over all 50 iterations, with kappa (accuracy) ranging from 0 to 1.

5. RESULTS

6.1. Overall distribution of monophthongs

The overall distribution of the 14 monophthongs is shown in Figure 1 as means. In addition, Figures 2a and 2b indicate variability with the help of ellipses (shown in two separate charts to avoid crowding). Visual inspection of the formant charts indicates clear overlap between THOUGHT – LOT and BATH - PALM, some overlap between DRESS – TRAP and FLEECE – KIT and little overlap between GOOSE – FOOT.

A comparison of these results with (somewhat older) data for Southern Standard British English (BrE) [18] indicates the following: There appears to be a merger between THOUGHT and LOT in our HKE data, whereas in BrE THOUGHT has lower F1 and F2 (articulation higher and further back). Further, HKE DRESS only has marginally lower F1 (i.e. slightly higher articulation) than TRAP. By contrast, DRESS is clearly differentiated from TRAP in BrE, where the former has a lower F1 and a higher F2 than the latter (articulation higher and fronter). The degree of overlap between FLEECE and KIT in our HKE sample seems not unlike that found in BrE, except that in the present data the two vowels are mainly differentiated in F2 (frontness) whereas in BrE they are mainly differentiated in F1 (height). Finally, FOOT appears to be articulated with somewhat lower F1 (higher point of articulation) in our HKE data compared to BrE. GOOSE partly has a somewhat higher F1 (lower point of articulation) in HKE than in BrE. Other than that, the degree of fronting of GOOSE in our data to some extent differentiates it from FOOT and is commensurate with the youngest BrE age group (20 – 25) in [18].

6.2. Potential Mergers

In order to investigate potential mergers more closely, we compare F1, F2 and duration in order to assess the distinctness of the vowel pairs. The classification statistic kappa ranges from 0 to 1 (0 = complete merger, 1 = completely separate)

The data for FLEECE – KIT indicate a very limited degree of differentiation (see Fig. 3). Especially noteworthy is that duration is virtually identical. Kappa = 0.26 shows very limited separability, i.e. it suggests a merger.

For DRESS – TRAP, the overlap is even clearer, with just a slight difference in F1 (see Fig. 4). The lack of a difference in duration is once again particularly clear. Kappa is 0.32, i.e. again indicating poor separability.

Finally, THOUGHT – LOT show no difference in F1 or F2, and a slight tendency for a longer duration in THOUGHT (see Fig. 5). For this pair, kappa = 0, i.e. these vowels are not statistically separable.

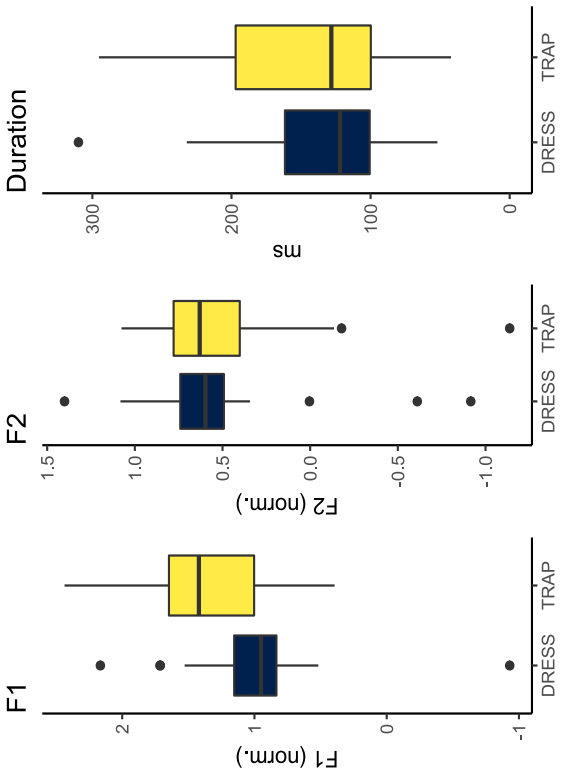


Figure 4. Distribution of F1, F2 and duration for DRESS and TRAP.

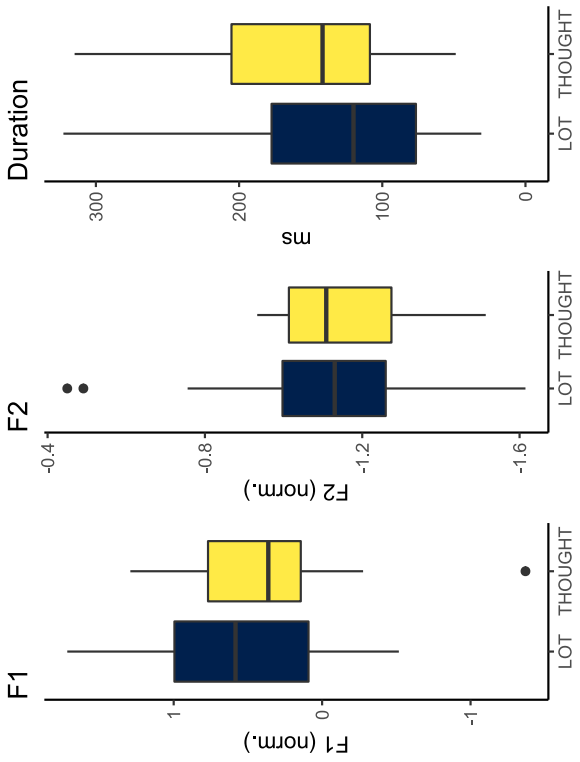


Figure 5. Distribution of F1, F2 and duration for LOT and THOUGHT.

7. DISCUSSION

The aim of this paper was to assess potential recent diachronic change in the vowel system of HKE. Our data, drawn from the post-1997 generation, provides some evidence of continuity with pre-1997 patterns. Specifically, we find a virtually complete DRESS – TRAP merger. Against DWK's [13] doubts, but confirming Hung [12], we also find a complete THOUGHT – LOT merger.

Our findings also refine and partly diverge from previous results. Specifically, we do not find a complete FLEECE – KIT merger among our HKE speakers, who produce an F1/F2 contrast that is at least in magnitude comparable to that found in BrE. This result contradicts Hung, but may be partly compatible with DWK, given that they only present mean formant values. However, given that HKE FLEECE and KIT do not differ in duration (unlike in BrE), they are not clearly separable. Finally, like [13], but unlike [12], we find evidence of a considerable degree of GOOSE fronting. However, our speakers produce only a much slighter degree of FOOT fronting. In this, our evidence goes against DWK, but parallels to some extent younger speakers of BrE [18].

Notably, our approach is arguably more sensitive to variability and promises greater technical accuracy than previous research, given (i) our reliance on not only formants but also duration, (ii) the use of vowel normalisation in order to account for anatomical variation in vocal tract length and (iii) the assessment of variability within the data, instead of just relying on mean values. This approach also allowed us to assess to what degree duration may sustain any of the vowel contrasts under investigation. Overall, there appears to be almost no difference in duration between long/short (tense/lax) vowel pairs.

8. CONCLUSION

This paper set out to analyse recent diachronic change in the monophthongs of HKE, without finding clear evidence supporting this hypothesis. This result contradicts our initial assumption that the political and social changes following the handover of HK to the People's Republic of China may have accelerated linguistic change. Nevertheless, our findings indicate that HKE maintains its distinctiveness from other varieties of English such as British and American English. In other words, this result suggests that HKE continues to rely on distinct norms, i.e. it is testament to its endonormativity and linguistic vitality (see also [2]). Finally, in terms of methodology, we

showed how Support Vector Machines can aid the analysis of potential vowel mergers.

8. BIBLIOGRAPHY

[1] J. Setter, C. S. P. Wong, and B. H. S. Chan, *Hong Kong English.* Edinburgh, England: Edinburgh University Press, 2010.

[2] R. Fuchs, "Hong Kong English: Structural features and future prospects," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai,* P. Siemund and J. R. E Leimgruber, Ed., London: Routledge, 2021, pp. 285-302.

[3] W. E. Schneider, *Postcolonial English: Varieties around the world*. New York: Cambridge, 2007, pp. 1-97,133-172.

[4] R. Fuchs, "The frequency of the present perfect in varieties of English around the World," in *Re-Assessing the Present Perfect*, V. Werner, E. Seoane, Ed., Berlin: de Gruyter, 2016, pp. 223-258.

[5] M. Wong, "Verb-preposition constructions in Hong Kong English: A cognitive semantic account," *Linguistics*, vol. 52, no.3, pp. 603-635, 2014.

[6] H. Parviainen and R. Fuchs, "'I don't get time only': an apparent-time investigation of clause-final focus particles in Asian Englishes," *Asian Englishes.*, vol. 21, no. 3, pp. 285-304, 2019.

[7] M. Wong, "Expressions of gratitude by Hong Kong speakers of English: research from the International Corpus of English in Hong Kong (ICE-HK)," *Journal of Pragmatics*, vol. 42, no. 5, pp. 1243-1257, 2010.

[8] H. Wolf and T. Chan, "Understanding Asia by means of cognitive sociolinguistics and cultural linguistics - the example of GHOSTS in Hong Kong English," in *Communicating with Asia: the future of English as a global language*, G. Leitner, A. Hashim and H. Wolf, Ed., Cambridge: Cambridge University Press, 2016, pp. 249-266.

[9] A. Y. W. Chan and D. C. Li, "English and Cantonese Phonology in Contrast: Explaining Cantonese ESL Learners' English Pronunciation Problems," *Lanuage, Culture and Curriculum,* vol. 13, no. 1, pp. 67-85, 2000.

[10] A. Sewell and J. Chan, "Patterns of variation in the consonantal phonology of Hong Kong English," *English World-Wide*, vol. 31, no.2, pp. 138-161, 2010.

[11] K. Bolton and H. Kwok, "The dynamics of the Hong Kong accent: Social identity and sociolinguistic de-scription," *Journal of Asian Pacific Communication 1*, pp. 147-72, 1990.

[12] T. T. N. Hung, "Towards a phonology of Hong Kong English," *World Englishes*, vol. 19, no. 3, pp. 337-356, 2000.

[13] D. Deterding, J. Wong and A. Kirkpatrick, "The pronunciation of Hong Kong English," *English World-Wide*, vol. 29, no. 2, pp. 148-175, 2008.

[14] J. C. Wells, *Accents of English 1: An introduction*, Great Britain: Cambridge University Press, 1982.

[15] T. Kisler, U. D. Reichel and F. Schiel, "Multilingual processing of speech via web services," *Computer, Speech and Language*, vol. 45, pp. 326-347, Sept. 2017.

[16] T. Kendall and Thomas E. R., *vowels: Vowel Manipulation, Normalization, and Plotting. R package version 1.2-2 (2018)*. [Online]. Available: https://CRAN.R-project.org/package=vowels

[17] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.

[18] S. Hawkins and J. Midgley, "Formant frequencies of RP monophthongs in four age groups of speakers," *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 183-199, 2005

[19] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*, TU Wien. 2021. R package version 1.7-5. https://CRAN.R-project.org/package=e1071

Lexically-guided phonetic recalibration transfers across languages in French-English bilinguals

Tiphaine Caudrelier [1,2], Clara D. Martin [2,3], Arthur G. Samuel [2,3,4], Marie-Michèle Beausoleil [5], Mark Tiede [6], Lucie Ménard [5]

[1] LPNC, Univ. Grenoble Alpes, CNRS UMR 5105, Grenoble, France,
[2] Basque Center on Cognition, Brain and Language, 20009 Donostia-San Sebastián, Spain
[3] IKERBASQUE, Basque Foundation for Science, Spain
[4] Stony Brook University, Dept. of Psychology, Stony Brook, NY, United States of America,
[5] Département de linguistique, Université du Québec à Montréal, Montréal, Québec, H3C 3P8, Canada,
[6] Haskins Laboratories, New Haven, Connecticut, United States of America

Tiphaine.caudrelier@univ-grenoble-alpes.fr

ABSTRACT

Lexically-guided phonetic recalibration results from exposing listeners to an ambiguous sound (e.g., a sound between /f/ and /s/) embedded in real words disambiguating the sound (as in 'beautiful'). After repeated exposure to the sound in various /f/-words, listeners shift their perception even in the absence of disambiguation [1]. Recalibration was shown to transfer across Germanic languages in advanced L2 learners [2]. Here we replicate this result in French-English bilinguals.

Fifty French-English bilinguals from Montreal, Canada, were tested online. Groups 1 and 2 had exposure in English toward /f/ and /s/ respectively, while Groups 3 and 4 had exposure in French. The perceptual shift was then assessed in both languages.

A perceptual shift occurred in all groups and transferred across languages. This suggests that French-English bilingual listeners have one common phonemic category for similar sounds across languages.

Keywords: bilingualism; perceptual learning; phonetic recalibration; speech perception; transfer;

1. INTRODUCTION

The lexical recalibration paradigm results from exposing a listener to an ambiguous sound, hereafter /?/ - for instance a sound between /f/ and /s/ - embedded in real words (e.g., "beautiful") that disambiguate the sound [1]. The ambiguous sound will not only be perceived as /f/ in this context but even in the absence of disambiguation [1], [3]. For instance, a repeated exposure to the ambiguous sound /?/ replacing the sound /f/ in real words will produce a shift of the perceptual boundary between /f/ and /s/. The boundary is expected to shift toward /s/, resulting in more /f/ responses in a subsequent identification task. By contrast, s-words like the item "curren/?/y" will be expected to trigger a shift in the opposite direction, as "currency" is an existing word while "currenfy" is not.

Previous work has shown that lexically-guided phonetic recalibration in advanced L2 learners transfers from L2 English to L1 Dutch [2] or L1 German [4]. This suggests that the phonetic systems of these listeners are shared between their languages or at least linked. In the present study we investigated whether this result could be replicated with French-English bilinguals. These languages are more distant topologically as they belong to different language groups (English being part of Germanic languages while French belongs to the Romance language family).

2. METHOD

2.1. Participants

Fifty French-English bilinguals (age: 18-40; French AoA=0; English AoA range = 0-12 years-old) from Montreal, Canada, were tested online. They were screened for their language experience and proficiency with a subset of the LEAP-Q (Language Experience and Proficiency Questionnaire - Marian, Blumfield, & Kaushanskaya, 2007). There was a headphone screening before the experiment to make sure they were wearing headphones or earbuds [6].

2.2. Design

Participants were divided into four groups. Groups 1 and 2 had exposure in English toward /f/ and /s/ respectively, while Groups 3 and 4 had exposure in French toward /f/ and /s/ respectively.

2.3. General procedure

The procedure consisted of two phases: an exposure phase and a post-test.

The exposure phase was an auditory lexical decision task in which stimuli differed across groups as explained in the Design section. On each trial, participants heard an auditory stimulus and they had to indicate if what they heard was a real word or not by pressing one of two keys on their keyboard. The critical items containing the ambiguous /f/-/s/ sound (either f-words or s-words) were presented among fillers.

The post-test aimed to characterize the category boundary between /f/ and /s/ in both French and English. It was an identification (categorization) task, in which participants heard tokens on a continuum between two words constituting a minimal pair (e.g., "found" and "sound"). Participants heard a token and had to indicate whether they heard /found/ or /sound/ by pressing a key. The post-test consisted of four blocks of 40 trials each (4 minimal pairs * 5 steps on the continuum * 2 repetitions), alternating between French and English. The language of the first block was counterbalanced across participants.

2.4. Stimuli

All stimuli in French and English were recorded by the same French-English bilingual talker, recruited in the Montreal area.

The stimuli of the exposure phase consisted of 36 critical items (18 /f/-words and 18 /s/-words), 36 fillers matched in length and frequency with critical items, and 72 pseudo-words. There were thus 144 stimuli per language.

Critical items were selected with the following criteria: /f/ or /s/ had to appear as late as possible in the word (ideally as the onset of the 3rd or 4th syllable); /f/ or /s/ should not appear in a consonantal cluster (for instance /fr/ or /sl/); words had to be as frequent as possible to maximize recalibration; cognates (words that are phonologically similar between French and English) had to be avoided. A f-version and a s-version of each critical item were recorded (e.g. 'beautiful' and 'beautisul', or 'currecy' and 'currenfy'). A 10-step continuum was created for each pair using a Matlab script that synthesizes weighted combinations of LPC-characterized end-points [7]. The fricatives were extracted, made equal in duration and then merged by morphing LPC coefficients. Formant transitions in the vowel following fricatives were also merged so that resulting formants were the weighted average of formants from both f-version and s-version recordings. These tokens were piloted to find the most ambiguous token in each continuum - the one for which the proportion of /f/ responses is the closest to 0.5. Six French native speakers performed an identification task on these continua, where they had to say if they heard /f/ or /s/, with 4 repetitions by token. English tokens were selected the same way with six English native speakers.

English f-words	French f-words
barefoot	*agrafer*
battlefield	*analphabète*
beautiful	*anglophones*
benefit	*approfondir*
counterfeit	*bénéfique*
delightful	*biographie*
identify	*confiture*
manifold	*décoiffer*
pocketful	*ébouriffé*
purify	*emberlificoter*
qualify	*haltérophile*
quantify	*imparfait*
radiography	*indifférent*
ungrateful	*mammifère*
uniform	*montgolfières*
unprofitable	*nénuphar*
waterfall	*réconfort*
wonderful	*triomphant*

English s-words	French s-words
accuracy	*angoissant*
countersign	*compréhensible*
cumbersome	*déboussolé*
currency	*débroussailler*
democracy	*déconcerté*
innocent	*dépoussiérer*
legacy	*éclaboussure*
noticeable	*éclaircir*
policy	*endurci*
precursor	*épanouissant*
pregnancy	*garçon*
producer	*intarissables*
pronouncing	*menacer*
reconcilable	*raccourci*
reducing	*ralentisseur*
rehearsal	*ramasser*
troublesome	*reconnaissant*
undersea	*remerciement*
underside	*tournesol*

Table 1: Critical items used in the exposure.

For the post-test, four minimal pairs that differ only on the /f/-/s/ contrast with various vocalic contexts were selected in both languages (in English: *feel/seal, for/soar, fine/sign, found/sound* and in French: *feuille/seuil, fil/cil, fort/sort* and *fond/son*). A 20-step continuum was created for each pair using the same procedure as for critical items [7].

The stimuli were also piloted with the same native speakers to find the most ambiguous token in each continuum - the one for which the proportion of /f/ responses is the closest to 0.5. The 5 steps selected for the post-test were the most ambiguous stimuli (MAS) along with MAS -5, MAS -2, MAS+2 and MAS+5. For instance, if the most ambiguous stimulus was step 8, then the selected steps were steps 3, 6, 8, 10 and 13 on the 20-step continuum. In the Results section these steps are relabelled from 1 to 5.

The pseudowords were generated using the multilingual pseudoword generator Wuggy [8]. They were matched in length and syllable frequency with real words (critical items and filler words). All exposure items contained no fricative except a single /f/ or /s/ in each critical items.

2.5. Data analysis

Data was analysed using General Linear Mixed Models with the *glmer* function from the *lme4* package in R. The variable *resp* is a binary variable which equals 1 when participants press the /f/ button and 0 when the /s/ button is pressed. This variable was modeled as follow:

*resp ~ recalibration_language * direction * test_language + step + (test_language|Participant) + (1|minimal_pair)*

recalibration_language (English or French), and *direction* (recalibration toward /f/ or /s/) are between-subject variables while *test_language* (English or French) and *step* (from 1 to 5) are within-subject variable. *participant* was included as a random factor. We followed a backward deletion approach [9] to select the best model using likelihood ratio tests. Post hoc tests were adjusted for multiple comparisons, with the *glht* function from the *multcomp* package of R [10].

2.6. Operational hypothesis

We expected that recalibration would occur in all groups. We thus predicted that groups exposed to f-words with an ambiguous sound between /f/ and /s/ (Groups 1 and 3) would perceive more /f/ in the post-test (with continua made from minimal pairs) than Groups exposed to ambiguous s-words during the recalibration phase.

We wanted to investigate whether there would be some transfer across languages, and if so, whether there would be full transfer? Transfer across languages occurs when there is a significant shift even when the test language is not the same as the recalibration language. A full transfer would occur if the magnitude of the shift is the same regardless of whether test language and recalibration are the same or different.

3. RESULTS

Figure 1 represents the proportion of /f/ responses across Groups (recalibration language and direction being between-subjects factors) in the post-test over the 5-step f-s continua. It shows that Groups 1 and 3, which were exposed to ambiguous sounds in f-words, tend to give more f responses than groups 2 and 4 which were exposed to ambiguous sounds in s-words.

The model that best fitted the data was *resp ~ direction + test_language + step* There was an effect of the direction of the recalibration (toward /f/ or toward /s/) and an effect of the language in the post-test with no interaction between those factors.

However there was no effect of the recalibration language (p=0.57). This absence of effect suggests that there was a transfer of recalibration across languages and that this transfer was nearly a full transfer. Figure 2 clearly illustrates that the recalibration is equivalent when post test language is the same as exposure or not.

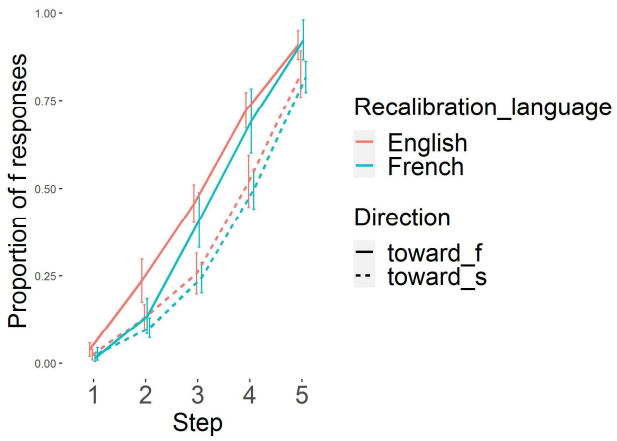


Figure 1: Proportion of /f/ responses across groups in the post-test. The x-axis represents the 5 steps on the /f/-/s/ continuum. The error bars give standard errors.

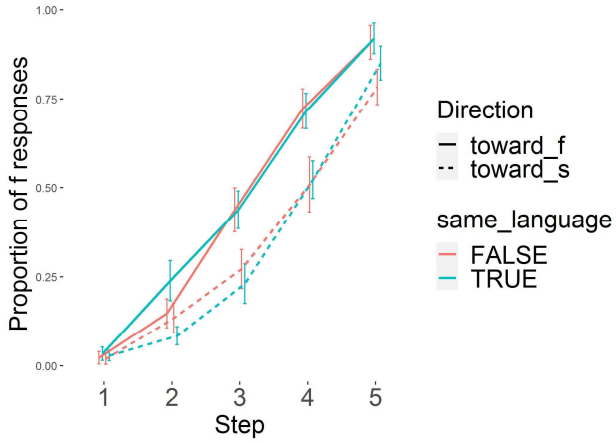


Figure 2: Proportion of /f/ responses given the direction of the recalibration and whether the post-test language is the same as the recalibration language or not. The gap between the two blue lines represents the recalibration in the same language while the gap between the red lines represents the transfer of recalibration from one language to another. The x-axis represents the 5 steps on the /f/-/s/ continuum. The error bars give standard errors.

Post hoc comparisons revealed that participants who had exposure toward /f/ gave significantly more /f/ responses than participants who had exposure toward /s/ (0.52±0.18; z=2.94; p=0.006). This main effect confirms that recalibration occurred. Post hoc comparison also indicate that the proportion of /f/ responses was higher when tested on French minimal pairs than when tested on English minimal pairs (0.43±0.07; z=6.36; p<0.001).

4. DISCUSSION

In the present study, we wanted to investigate whether perceptual learning, more specifically lexically-guided phonetic recalibration, transfers across languages in French-English bilinguals.

We had four Groups with different exposure conditions. Groups 1 and 3 had exposure toward /f/ while Groups 2 and 4 had exposure toward /s/. Groups 1 and 2 had exposure in French while Groups 3 and 4 had exposure in English. All groups were then tested on their perception of /f/-/s/ continua in both English and French minimal pairs, to assess their perceptual boundary between /f/ and /s/ in each language.

There was no effect of the exposure language (French vs English) on the proportion of /f/ responses. This confirmed our prediction: there was a perfect transfer of recalibration from English to French and vice versa. This replicates the study by c which observed similar magnitude in the shift triggered in L1 by exposure in L2 and the shift induced by L2 on L1. This result indicates that French-English bilinguals may have common phoneme categories for /f/ and /s/ which are similar across languages. Our study extends their conclusion to languages that are part of distinct languages families and which differ a lot in terms of phoneme inventory.

In addition, results indicate that we succeeded in inducing recalibration in our participants, since participants who had exposure toward /f/ produced more /f/ responses in the post-test than participants who had exposure toward /s/. This confirms that our stimuli were adequate to trigger lexically guided phonetic recalibration. This bilingual set of stimuli (about 300 stimuli, including 80 /f/-/s/ continua in various words or minimal pairs) can therefore be used in future experiments based on lexically guided phonetic recalibration.

Finally, there was an effect of the test language, and this effect was equivalent across groups (there was no interaction). Although we piloted our stimuli to try to find the middle of each continuum - the most ambiguous point where participants would respond half /f/ half /s/ - there was still some bias and the bias differed across languages (and even across minimal pairs). This is quite common in the literature to have a bias in the results toward one end point of the continuum and it does not affect our conclusion.

In conclusion, the present study indicates that lexically-guided phonetic recalibration may transfer across languages in French-English bilinguals although these languages are quite distant from a phonetic standpoint. It also provides a robust set of stimuli that can be used for future research in the domain of perceptual learning in bilingual speakers.

7. ACKNOWLEDGMENTS

This research was supported by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S. The research was also supported by the Spanish Ministry of Economy and Competitiveness (PSI2017-82941-P to C.D.M.), the Spanish Ministry of Science and Innovation through the Juan de la Cierva Research Fellowship (FJC2019-039005-I awarded to T.C.), the Spanish Ministry of Science and Innovation through Grant (# PID2020-113348GB-I00 awarded to A.D.S), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 819093 to C.D.M.). Besides, support for this project was provided by Economic and Social Research Council (UK) Grant #ES/R006288/1.

8. REFERENCES

[1] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cogn. Psychol.*, vol. 47, no. 2, pp. 204–238, 2003.

[2] E. Reinisch, A. Weber, and H. Mitterer, "Listeners retune phoneme categories across languages.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 39, no. 1, pp. 75–86, 2013.

[3] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychon. Bull. Rev.*, vol. 13, no. 2, pp. 262–268, 2006.

[4] K. Schuhmann, "Cross-linguistic perceptual learning in advanced second language listeners," *Proc Ling Soc Amer*, vol. 1, no. Article 31, pp. 1–32, 2016.

[5] V. Marian, H. Blumfield, and M. Kaushanskaya, "The Language Experience And Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," *J. Speech, Lang. Hear. Res.*, vol. 50, pp. 940–967, 2007.

[6] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, Psychophys.*, vol. 79, no. 7, pp. 2064–2072, 2017.

[7] T. M. Byun and M. Tiede, "Perception-production relations in later development of American English rhotics," *PLoS One*, vol. 12, no. 2, pp. 1–22, 2017.

[8] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behav. Res. Methods*, vol. 42, no. 3, pp. 627–633, 2010.

[9] R. Mundry and C. L. Nunn, "Stepwise model fitting and statistical inference: turning noise into signal pollution.," *Am. Nat.*, vol. 173, no. 1, pp. 119–123, 2009.

[10] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical J.*, vol. 50, no. 3, pp. 346–363, 2008.

EXAMINING THE EFFECTS OF STRESS ON VOWEL PRODUCTION IN HERITAGE VS. MONOLINGUAL SPANISH SCHOOL-AGED CHILD SPEAKERS

Gemma Repiso-Puigdelliura[1], Daniel Pape[1]

[1]McMaster University
repisopg@mcmaster.ca, paped@mcmaster.ca

ABSTRACT

While Spanish adult heritage speakers (i.e., HSs) show vocalic reduction in unstressed position, it is still not known whether vowel reduction is already present as HSs develop their language speech-specific patterns. We examined the effect of lexical stress on 24 primary school-aged child HSs (mean age = 8;11 ± 2) and compared their non-high vowel productions to those of 12 Spanish monolingual speakers (mean age = 8;12 ±1;6). Our results showed that non-high vowels in disyllabic words are produced higher in unstressed positions when compared to stressed positions. In addition, *stress* interacted with *group* (HS vs monolingual) for the /e/ vowel. Along the front-back dimension, our findings only show a cross-over interaction between *stress* and *group* for /a/. For all other vowels, no significant differences were found between HSs and monolinguals. The lack of vowel reduction during childhood may be due to overall late development of vowel reduction patterns.

Keywords: lexical stress, child heritage speakers, vowel reduction

1. INTRODUCTION

Heritage speakers (HSs) are early bilinguals that are exposed to their family language (i.e., heritage language) simultaneously or sequentially with the majority language of the society [1]. While speech production in HSs presents more target-like characteristics than that of L2ers [2], HSs' productions have been shown to diverge from canonical descriptions of the language [3]. It is still unclear, however, whether such divergences arise during speech development due to early grammar interaction or whether HSs undergo attrition as their input to the heritage language is reduced [4]. To investigate this question, it is important to examine HSs' speech production during their primary school years as HSs gain systematic exposure to the majority language and may undergo a shift in their dominance from the heritage to the majority language [5].

Early accounts of the Spanish vowel system (i.e., /a e i o u/) describe it as stable and relatively unaffected by lexical stress [6]–[8]. It is worth mentioning, however, that studies on dialectal variation have found that in dialects such as Andean Spanish, Mexican Spanish, or Peninsular Spanish unstressed vowels can vary from stressed vowels with respect to duration and vowel quality [9]–[11].

Unlike in Spanish, in English vowel reduction is a phonological phenomenon, and vowel quality is, therefore, more consistently affected by stress [12]. Unstressed English vowels are more central than stressed vowels and are produced with less effort [6], [12], [13]. When in contact with English, Spanish HSs show vowel centralization in unstressed positions [14], [15]. Ronquest [14] found that unstressed vowels are shorter than stressed vowels and demonstrate a more reduced vocalic space. This reduction, however, is not in the direction of the centralized schwa. The vowels /e a o/ in the above-mentioned study had lower F1 values and thus were produced higher in unstressed position. All unstressed vowels except for /a/ showed centralization in the F2 dimension. Similarly, Elias et al. [15] demonstrated that unstressed /a/, /e/, and /o/ have lower F1 values than their stressed counterparts. Along the F2 dimension, their results show an interaction between lexical stress and vowel phoneme, indicating that stressed /e/ and /i/ are more fronted than unstressed /e/ and /i/, whereas stressed /o/ and /u/ are more posterior than their unstressed counterparts. Moreover, vowels produced in code-switched mode are, overall, more centralized than vowels produced in monolingual mode. To summarize, adult Spanish HSs present a more reduced vowel space in unstressed position, which is characterized by lower F1 values overall and more compressed F2 values, with respect to the stressed position. To our knowledge, however, no studies have investigated whether a similar reduction is present earlier in childhood during the development of the HSs' two phonologies.

In this study, in order to examine whether the reduction patterns found in previous studies on adult HSs would appear already during the primary school years, we explore the production of /a/, /o/, and /e/ in the semi-spontaneous speech of a group of child HSs and compare the results to that of a group of monolingually-raised Spanish speakers.

2. METHODS

2.1. Participants

24 child HSs (12 M, 12 F, mean age = 8;11 ± 2, age range = 5;3 to 11;11) raised in California with two caregivers from Mexico (except for 2 participants with one caregiver from Mexico and one English-speaking caregiver [Australia and Los Angeles], and 3 participants with one caregiver from Mexico and one caregiver from El Salvador) and 12 monolingually-raised child Spanish speakers from Mexico (i.e., SpanMonoSs, 8 M, 4 F, mean age = 8;12 ±1;6, age range = 6;1 to 11;1) participated in this study. A parental background questionnaire showed that the child HSs and the SpanMonoSs significantly differed in the amount of Spanish relative to English to which they were exposed at home and at school ($t(30.87) = -4.93$, $p < 0.001$), and in the amount of Spanish relative to English that they produced at home ($t(26.53) = -7.22$, $p < 0.001$). In particular, relative to English, SpanMonoSs were exposed to more Spanish at home and at school (M = 85.9%, SD = 14.6%) than child HSs (M = 55.65%, SD = 21.8%), and produced more Spanish at home (M = 96.9%, SD = 5.7%) than child HSs (M = 53.0%, SD = 28.7%).

2.2. Task

Semi-spontaneous speech was elicited using the wordless book *Frog, where are you?* [16]. Participants were asked to narrate the story as naturally and detailed as possible. The session was conducted on Zoom by the first author. During the session, caregivers were asked to record the participants' speech using a smartphone positioned 4 inches from the children's mouths. Recordings were done using the ShurePlusMotiv, an app that records with a sampling rate of 48.1 kHz and a sample size of 16 bits in an uncompressed audio format.

2.3. Formant extraction

Disyllabic content words[1] were extracted from the data and the formant frequencies F1 and F2 were measured at the acoustic midpoint of the vowels using PRAAT's formant extraction algorithm ("To Formant... burg") with adapted parameter settings for children's vocal tracts (max. formant: 8000 Hz; max. number of formants: 5; window length: 25 ms; pre-emphasis from 50 Hz) as a first approximation of the correct formant value. The automatically extracted spectral maxima (i.e., formant candidate) of the described Praat algorithm were then carefully manually corrected as needed based on the full LPC spectra (as main source), but also guided by FFT broadband spectra and the FFT broadband spectrograms. Thus, for each timepoint, the automatic measurement value (obtained by Praat's (adapted) automatic formant extraction algorithm) was manually checked and, if necessary, corrected, based on (1) the temporal formant movement information visible in the spectrogram, (2) the spectral information visible in the broadband FFT spectral slice (5ms window length, identical to the spectrogram settings) and (3) the spectral information visible in a generated LPC spectral slice (Praat LPC object with LPC order = 16, sampling rate of the underlying audio file = 16000 Hz). This hybrid approach of measuring formants for notoriously difficult child speech data based on the combination of the three methods, but mainly relying on the LPC spectra and maxima, ensures that eventual automatic formant tracking errors are noticed and thus manually corrected. Vowels were then normalized[2] using the Lobanov method of vowel normalization using the R package phonR [17].

4. RESULTS

A total of 3226 non-high vowels[3] were extracted from the semi-spontaneous speech. Six linear mixed effects models for the vowels /a/ /e/ /o/ in the Lobanov transformed F1 and F2 dimensions were run using the lme4 package [18] and p-values were obtained with the lmerTest package [19]. The variables *stress* (i.e., stressed vs. unstressed) and *type of speaker* (child HSs vs. SpanMonoSs) were entered as fixed effects, and the intercepts for participant and word were allowed to vary.

4.1. Results for /a/

With regard to /a/ (See Fig. 1), the model for F1 showed that unstressed vowels were produced with lower values of F1 (i.e., higher) than stressed vowels ($\beta = -0.32$, $SE = 0.12$, $t = -2.59$, $p < 0.001$, CI 95% [-0.57, -0.08]). No main effect for *type of speaker*, or interaction between *type of speaker* and *stress* was found. The model for F2 showed a significant effect for *stress* ($\beta = -0.43$, $SE = 0.13$, $t = -3.34$, $p < 0.001$, CI 95% [-0.68, -0.18]) and a significant cross-over interaction was found between *stress* and *type of speaker* ($\beta = 0.68$, $SE = 0.14$, $t = 4.71$, $p < 0.001$, CI 95% [0.09, 0.34]), indicating that the direction of the effect of *stress* on F2 depends on the variable *type of speaker*: In the child SpanMonoSs, /a/ is more fronted in the stressed condition than in the unstressed condition, whereas in the child HSs /a/ is more fronted in the unstressed condition than in the stressed condition.

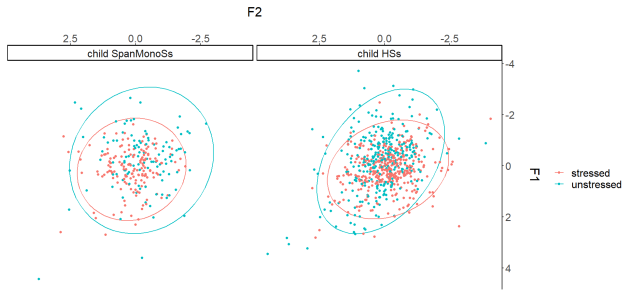


Figure 1: Lobanov-normalized F1 and F2 values for /a/ by variables *type of speaker* and *stress*. The data ellipses expect to contain 95% of the population distribution.

4.2. Results for /e/

As for /e/ (See Fig. 2), the model for F1 showed that unstressed vowels were produced with lower F1 values (i.e., higher) than stressed vowels (β = -0.78, SE = 0.21, t = -3.56, p < 0.001, CI 95% [-1.21, -0.39]), and that *stress* interacted with *type of speaker* (β = -0.47, SE = 0.22, t = 2.15, p = 0.03, CI 95% [0.02, 0.87]), but no significant effect of *type of speaker* was found. Post-hoc analyses indicated that while *stress* significantly shifted F1 on the child SpanMonoSs (β = 0.78, SE = 0.22, t = 3.48, p < 0.001), it did not do so for the child HSs (β = 0.31, SE = 0.19, t = 1.61, p = 0.10). No main effect of *type of speaker* was found. The model for F2 did not show any significant effects for any of the variables or their interaction.

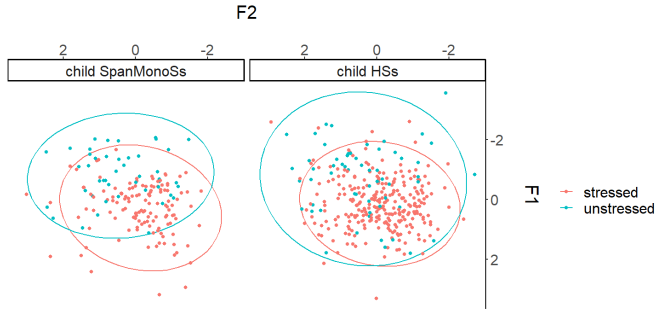


Figure 2: Lobanov-normalized F1 and F2 values for /e/ by *type of speaker* and *stress*.

4.3. Results for /o/

Concerning the results for /o/ (See Fig. 3), the model for F1 demonstrated that unstressed vowels were produced with lower F1 values (i.e., higher) than stressed vowels (β = -0.28, SE = 0.09, t = -2.88, p = 0.004, CI 95% [-0.49, -0.07]). No effect of *type of speaker* nor interaction between *type of speaker* and *stress* was found. No significant differences were found along the F2 dimension.

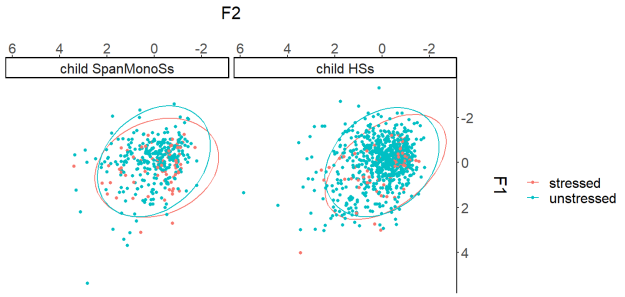


Figure 3: Lobanov-normalized F1 and F2 values for /o/ by *type of speaker* and *stress*.

5. DISCUSSION

5.1. General effects of stress on F1 and F2

Similar to previous studies on HSs' vowel production [14], [15], our results demonstrate that, overall, non-high vowels (i.e., /a e o/) in unstressed position are produced with lower F1 values (i.e., higher) when compared to those in stressed position. Unlike Ronquest [14] and Elias et al., [15], however, the front-back dimension (i.e., F2) was not significantly affected by *stress* in our data.

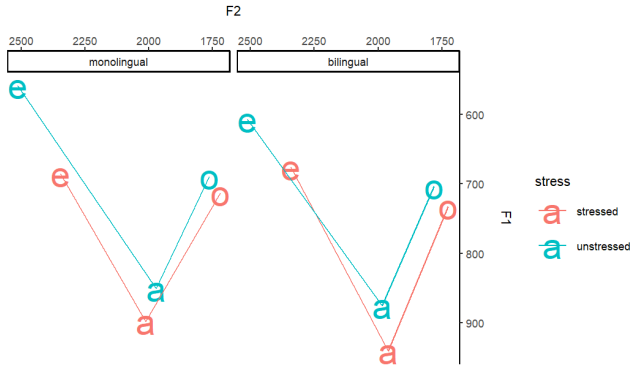


Figure 4: Mean Lobanov-normalized F1 and F2 values by vowel, *stress*, and *type of speaker*

Despite our unstressed vowels shifting towards a more reduced acoustic space (see Fig. 4), the lower F1-values for unstressed /e/ and /o/ and the lack of significant movement along the F2 dimension do not suggest that Spanish-speaking children centralize non-high vowels to /ə/. Instead, it is possible that F1 effects are a result of prominence-related universal phonetic principles. Prominent vowels are associated with a lower jaw position [20] and more extreme articulatory movements [21], [22], which could lead to more open vowel productions in general and more peripheral F1 values [21], [23]. Two strategies support an enhanced production of accented vowels. First, the 'sonority expansion strategy' states that speakers adopt a more open vocal tract to create a larger oral passage which results in louder and thus more open vowels [22]. The 'localized hyperarticulation', in turn, posits that prominence

affects tongue movement, by which vocal targets become more distinct and contrasts between vowels are strengthened [24]. In sum, we propose that the higher F1 values in stressed vowels compared to unstressed vowels, seen for both monolingual and HSs groups, could be a result of increased intensity in prominent vowels attained through larger jaw opening and lowering of the tongue dorsum.

5.2. Child heritage speakers compared to Spanish monolinguals

Contrary to our predictions, our findings did not show that child HSs produced unstressed-stressed vowels with a greater difference than that of child SpanMonoSs. Our model for /a/ F2 showed an interaction between the type of speaker and stress, suggesting that, while unstressed vowels produced by child HSs move forward on the F2 dimension, unstressed vowels produced by SpanMonoSs move backward. It is possible that the Spanish /a/ from the child HSs is affected by the English vocalic space, in that /ɑ/ (i.e., stressed) is more backward than the Spanish /a/ [25], and /ə/ is more forward than its stressed counterpart. In addition, the fronted English /æ/ could further restrict the available acoustic space that child HSs have in Spanish. Our model for /e/ F1 showed that only the child SpanMonoSs, but not HSs, produced lower stressed /e/ vowels compared to unstressed ones, with no significant effect between the two populations. It is hypothesized that an expected prominence effect is constrained in the child HSs due to the existence of /ɛ/ in their English acoustic vowel space. In principle, we should see the same interaction for the /o/ vowel[4]. However, the COT-CAUGHT merger in Californian English [26], [27] and the limited number of stressed /o/ tokens in our data set (N = 147) could explain the lack of a significant interaction. Our results, hence, do not provide evidence for early transfer of vowel reduction from English to Spanish in HSs. To explain why our results diverge from those of adult HSs, we could posit that English reduction patterns are still in development among the child HSs in this study (mean age = 8;11 ± 2), and thus, cannot be transferred to the HL at this point in their development. In fact, patterns of reduction are characteristic of expert motor control [28], [29] and some of them can extend until the primary school years. For example, Goffman [30] found that function words are less reduced in children up to the age of 7 than in adults. Future research should include an examination of the English vowels of the HSs to determine whether the children at the time of testing have acquired English-like vowel reduction patterns.

Alternatively, our findings could be a result of the articulatory costs of vowel reduction. The late-developing nature of reduction patterns may render vowel centralization taxing from a processing and articulatory perspective during early speech production. In turn, recent accounts on the structure of heritage language grammars suggest that processing costs may determine whether an area of the HL is vulnerable to language transfer or grammar restructuring [31]. Given the potentially higher costs for reducing vowels, a possible pressure of the majority language grammar will be less likely to reorganize the Spanish phonologies into adopting vowel reduction from the HSs' English phonologies. Future research should investigate vowel reduction using a longitudinal design to better understand the period at which HSs start showing the patterns of vowel reduction found in adult HSs.

6. REFERENCES

[1] G. Valdés, "Heritage language students: Profiles and possibilities," in *Handbook of Heritage, Community, and Native American Languages in the United States: Research, Policy, and Educational Practice*, T. G. Wiley, J. K. Peyton, S. C. K. Moore, D. Christian, and N. Liu, Eds., Routledge, 2014, pp. 41–49. doi: 10.4324/9780203122419.

[2] L. M. Knightly, S.-A. Jun, J. S. Oh, and T.-K. Au, "Production benefits of childhood overhearing," *J Acoust Soc Am*, vol. 114, no. 1, pp. 465–474, 2003, doi: 10.1121/1.1577560.

[3] R. Ronquest and R. Rao, "Heritage Spanish phonetics and phonology," in *The Routledge Handbook of Spanish as a Heritage Language*, K. Potowski, Ed., London/New York: Routledge, 2018, pp. 164–177.

[4] M. Polinsky, "Bilingual children and adult heritage speakers: The range of comparison," *International Journal of Bilingualism*, vol. 22, no. 5, pp. 547–563, 2018, doi: 10.1177/1367006916656048.

[5] S. Montrul, "Heritage language development: Connecting the dots," *International Journal of Bilingualism*, vol. 22, no. 5, pp. 530–546, Oct. 2018, doi: 10.1177/1367006916654368.

[6] P. Delattre, "An acoustic and articulatory study of vowel reduction in four languages," *IRAL - International Review of Applied Linguistics in Language Teaching*, vol. 7, no. 4, 1969, doi: 10.1515/iral.1969.7.4.295.

[7] T. Navarro Tomás, *Manual de pronunciación española*, 12th ed. Madrid: CSIC, 1918.

[8] A. Quilis and M. Esgueva, "Realización de los fonemas vocálicos españoles en posición fonética normal," in *Estudios de fonética*, M. Esgueva and M. Cantarero, Eds., Madrid: Consejo Superior de Investigaciones Científicas, 1983, pp. 159–251.

[9] A.-M. Delforge, "Unstressed vowel reduction in Andean Spanish," in *Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology*, L. Colantoni and J. Steele, Eds., Sommerville, MA: Cascadilla Proceedings Project, 2008, pp. 107–124.
[10] R. Marín Gálvez, "La duración vocálica del español," *E.L.U.A.*, vol. 10, pp. 213–226, 1994.
[11] J. M. Lope Blanch, "En torno a las vocales caedizas del español mexicano," *Nueva Revista de Filología Hispánica (NRFH)*, vol. 17, no. 1/2, 1964, doi: 10.24201/nrfh.v17i1/2.1507.
[12] M. Fourakis, "Tempo, stress, and vowel reduction in American English," *Journal of the Acoustical Society of America*, vol. 90, no. 4, 1991, doi: 10.1121/1.401662.
[13] W. R. Tiffany, "Nonrandom sources of variation in vowel quality.," *J Speech Hear Res*, vol. 2, 1959, doi: 10.1044/jshr.0204.305.
[14] R. Ronquest, "An Acoustic Examination of Unstressed Vowel Reduction in Heritage Spanish," in *Selected Proceedings of the 15th Hispanic Linguistics Symposium*, C. Howe, S. Blackwell, and M. Lubbers Quesada, Eds., Cascadilla Proceedings Project, 2013, pp. 151–171.
[15] V. Elias, S. McKinnon, and Á. Milla-Muñoz, "The effects of code-switching and lexical stress on vowel quality and duration of heritage speakers of Spanish," *Languages*, vol. 2, no. 4, p. 29, 2017, doi: 10.3390/languages2040029.
[16] M. Mayer, *Frog Where Are You?* New York: Dial Press, 1969.
[17] D. R. Mccloy, "Vowel normalization and plotting with the phonR package," *Technical Reports of the UW Linguistic Phonetics Laboratory*, vol. 3, 2012.
[18] D. Bates, D. Sarkar, M. D. Bates, and L. Matrix, "The lme4 Package," *October*, vol. 2, no. 1, 2007.
[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J Stat Softw*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/JSS.V082.I13.
[20] M. E, Beckman, and J. Edwards, "Articulatory evidence for differentiating stress categories," in *Phonological Structure and Phonetic Form*, 2010. doi: 10.1017/cbo9780511659461.002.
[21] D. Erickson, "Articulation of Extreme Formant Patterns for Emphasized Vowels," 2002.
[22] K. de Jong, M. E. Beckman Ohio, and J. Edwards, "The interplay between prosodic structure and coarticulation," 1993.
[23] P. A. Keating, "Comments on the jaw and syllable structure," *J Phon*, vol. 11, no. 4, 1983, doi: 10.1016/s0095-4470(19)30839-3.
[24] K. J. Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, no. 1, 1995, doi: 10.1121/1.412275.
[25] A. R. Bradlow, "A comparative acoustic study of English and Spanish vowels," *Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1916–1924, 1995, doi: 10.1121/1.412064.
[26] P. Eckert, "Where do ethnolects stop?," in *International Journal of Bilingualism*, 2008. doi: 10.1177/13670069080120010301.
[27] A. A. Metcalf, "Directions of Change in Southern California English," *J Eng Linguist*, vol. 6, no. 1, 1972, doi: 10.1177/007542427200600104.
[28] M. A. Redford, "The Acquisition of Temporal Patterns," in *The Handbook of Speech Production*, 2015. doi: 10.1002/9781118584156.ch17.
[29] V. Kapatsinski, S. Easterday, and J. Bybee, "Vowel reduction: A usage-based perspective," *Italian Journal of Linguistics*, vol. 32, no. 1, 2020, doi: 10.26346/1120-2726-146.
[30] L. Goffman, "Kinematic differentiation of prosodic categories in normal and disordered language development," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 5, 2004, doi: 10.1044/1092-4388(2004/081).
[31] M. Polinsky and G. Scontras, "Understanding heritage languages," *Bilingualism: Language and Cognition*, vol. 23, no. 1, pp. 4–20, Jan. 2020, doi: 10.1017/S1366728919000245.
[32] J. I. Hualde, "Unstressed words in Spanish," *Language Sciences*, vol. 31, no. 2–3, pp. 199–212, Mar. 2009, doi: 10.1016/j.langsci.2008.12.003.

[1] In Spanish, some function words are considered to be unstressed (e.g., *menos amigos* 'except for friends') [32]. Thus, to avoid confounds, we decided to exclude function words from our data. Similarly, Elias et al. [15] excluded high-frequency words (definite articles, possessives, prepositions, conjunctions).

[2] In order to determine if vowel normalization would give different results compared to non-normalized children formant data, we run all analyses of this paper also on non-normalized data. The results were all identical. Reasoning to either include or not include vowel normalization was that at the mean age of nine years no relevant differences between vocal tract sizes (i.e., male/female but also across children of the same gender) were assumed to be relevant and thus it was not clear whether speaker normalization would be necessary for this data. Since there were no differences, we report the normalized data here.

[3] Only non-high vowels were included in the analysis given the small number of tokens for /i/ (N of unstressed = 23, N of stressed = 515) and /u/ vowels (N of stressed =26, N of unstressed = 29).

[4] That is, /o/ compared to /ɔ/ for the English grammars of the HSs. Despite the fact that [26] reports a rising in the merger, it is still possible that a raised merger does not exert the same influence on the Spanish acoustic space as /ɛ/ could.

DIPHTHONG ACOUSTICS AND ARTICULATORY UNIFORMITY

Fang Hu

University of CASS and the Institute of Linguistics, Chinese Academy of Social Sciences
hufang@hotmail.com

ABSTRACT

This paper examines acoustics and articulation in the production of diphthongs in Ningbo Chinese. The acoustic study is based on audio recordings from 20 speakers, and the articulatory data is based on EMA recordings from 6 speakers. Formant data indicate that there is a typological difference between falling and rising diphthongs. Falling diphthongs seem to have a dynamic spectral target, while rising diphthongs are sequences of two spectral targets. Articulatory kinematics reveals an articulatory uniformity in diphthong production. In conclusion, the observed acoustic difference between falling and rising diphthongs should be attributed to the nonlinear articulatory-to-acoustic relations.
Keywords: diphthong production, falling diphthong, rising diphthong, acoustics, articulatory uniformity.

1. INTRODUCTION

The Chinese language is well-known for its simple syllabic structure and complex tonal system. The complexity of vowels is essential to the understanding of syllables in Chinese dialects and related minority languages. A Chinese syllable is composed of an initial consonant (C), a glide (G), a vowel nucleus (V), and a coda (Co), which could be represented as CGVCo, where only the nucleus V is obligatory while the other elements are all optional. There is a consensus that Chinese syllable doesn't allow consonant clusters. That is, the glide G is treated as a vowel, rather than a consonant. However, there are different analyses of Chinese vowel elements and different proposals for syllable structure [1, 2, 3, 4, 5, 6, 7, 8]. It is quite natural that different phonological analyses yielded non-unique solutions [9]. The fact is that GV forms a rising diphthong, VCo forms a falling diphthong if Co is a vocalic element, and GVCo forms a triphthong, no matter G and vocalic Co is treated as a vowel or an approximant.

This paper focuses on controversy regarding the nature of diphthongs. There are two kinds of viewpoints. Some scholars argued that diphthongs are an articulatory event with a dynamic target; that is, a diphthong is a single phonological vowel with a complex phonetic nucleus [10, 11, 12]. The others simply treated diphthongs as being composed of two articulatory events, transiting from a static target to the other; that is, a diphthong is a sequence of two monophthongal vowels [13, 14]. This paper examines spectral properties and articulatory kinematics of diphthong production in Ningbo Chinese, a language with both falling and rising diphthongs, and aims to provide fine-grained phonetic details to the understanding of diphthong production in general and the understanding of syllable structure of Chinese in particular.

2. METHODOLOGY

In acoustic studies, 20 average adult speakers, 10 male and 10 female were recorded. All speakers were born and raised up locally and have no reported speech and hearing disorders. Monosyllabic words containing target vowels or diphthongs were used as test words. Each test word X was placed in the middle of a carrier frame. Audio sound were recorded directly into a laptop through a DMX 6 Fire USB soundcard with a Shure SM 86 or 58 microphone. Audio recordings were conducted in quite rooms during the fieldwork trips. The sampling rate is 11,025 Hz or 22,050 Hz, 16 bit, and five repetitions were recorded. Monophthongs were annotated as one segment; diphthongs are typically composed of two steady states and a transition, and were annotated as three segments, i.e., an onset, an offset, and the transition connecting them. The lowest three formants were extracted from the spectrogram in the mid-point of each target segment.

6 native adult speakers were recorded for the articulatory study by using the Carstens 2D AG100 or AG200 system. Sensors were attached at the midline on the upper and lower lip, the gum ridge at the lower teeth (jaw), and the three points of the tongue, namely the tongue tip (TT), tongue mid (TM), and tongue dorsum (TD). In addition, two receiver coils were attached to the bridge of the nose and the gum ridge at the upper teeth respectively, serving as the two reference points. The original 500 Hz EMA data were down sampled to 250 Hz and smoothed with the sensor on the lower lip being smoothed with a filter that has a cut-off frequency at 40 Hz and the rest of the sensors with a filter that has a cut-off frequency at 25 Hz. The positions of the tongue, jaw, and lip transducers at the target point in a diphthong element were extracted from

the EMA data by using the tangential velocity minima criterion [15, 16]. The criterion was applied with reference to the sampled principal lingual point, that is, TM for diphthong elements in [ai] and [ia], and TD for diphthong elements in [au] and [ua].

Due to the space limit, the discussion was focused on the two pairs of falling versus rising diphthongs, i.e., [ai] versus [ia] and [au] versus [ua], for comparison's sake.

3. RESULTS

3.1. Specteral properties

Chinese dialects have rich inventories of diphthongs. There are usually both falling and rising diphthongs. Northern Wu dialects are famous for the richness in monophthongs and the lack of falling diphthongs [17, 18]. The Ningbo dialect is a typical northern Wu. There are 12 monophthongs [ɿ ʯ i y ʏ e ø ɛ a ɔ o u] including the two apical vowels, 3 falling diphthongs [ai au œy], 6 rising diphthongs [ia ie io yo ua uaʔ uɛ], and 1 triphthong [uai] [19].

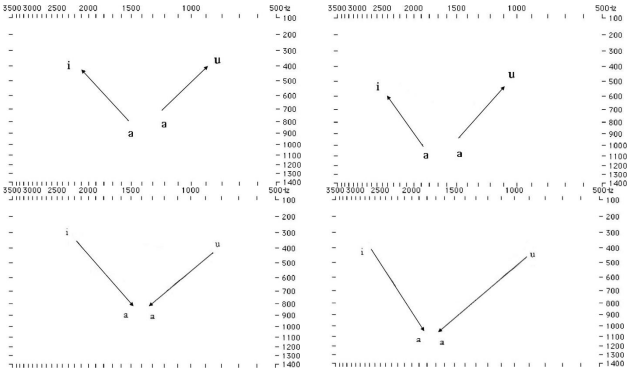


Figure 1: The paired falling diphthongs [ai au] (upper) and rising diphthongs [ia ua] (lower): averaged formant data in Hertz from 10 males (left) and 10 females (right).

Figure 1 plotted the paired falling diphthongs [ai au] (upper) and rising diphthongs [ia ua] (lower) in a two-dimensional acoustic space (F2 against F1) with the origin of the axes to the top right of the plot. The arrow connected diphthong onset to offset, which were averaged from 5 repetitions across 10 male (left) or female (right) speakers. The scaling relation of the axes has been converted to the bark to better approximate the perceptual distances. And the scale on the ordinate is double that on the abscissa to give appropriate prominence to F1 and make the plots more in accord with auditory judgments of vowels. However, the values along the axes still correspond to the original values in Hertz. The F1/F2 vowel plane establishes a good correlation with linguistic vowel features such as height and backness, namely F1 with vowel height and F2 with vowel backness.

Figure 2 compares diphthong elements of falling diphthongs with their monophthongal counterparts [i a u]. Confidence ellipses with radii of two standard deviations were drawn along axes oriented along the principal components of each vowel cluster. The descriptive statistics of the 2-sigma vowel ellipse predicts about 86% of the distribution of data points, and thus provides a straightforward way to examine the variability of diphthong elements vis-à-vis their monophthongal counterparts. It is apparent from the figure that diphthong onsets behave differently to diphthong offsets. And data are consistent across male and female speakers. On one hand, as compared with the monophthongal [a], [a] in [ai] becomes more front and [a] in [au] more back, due to the coarticulatory effect of the offset [-i] and [-u] respectively. But on the other hand, both diphthong elements and the monophthong [a] are all located in the height of low vowel, and their ellipses overlap with each other to certain extent. Diphthong offsets are different. Both [-i] in [ai] and [-u] in [au] don't reach their target high vowel positions: [-i] in [ai] stops at mid-high [e] or even mid-low [ɛ] region, and accordingly [-u] in [au] stops at mid-high [o] or even mid-low [ɔ] region. The ellipse for [-i] in [ai] even doesn't overlap with the monophthongal [i]; the ellipses for [-u] in [au] and the monophthongal [u] overlap marginally. And the apparently bigger ellipses for diphthong offsets, especially in female speakers, also suggest greater variability for diphthong offsets than for their monophthongal counterparts. In summary, formant data suggest that the onsets are better controlled than the offsets in the production of falling diphthongs [ai au]. It seems that in a falling diphthong, only the onset has a spectral target, and the offset could be viewed as a consequence of articulatory movement and is constrained by diphthong dynamics.

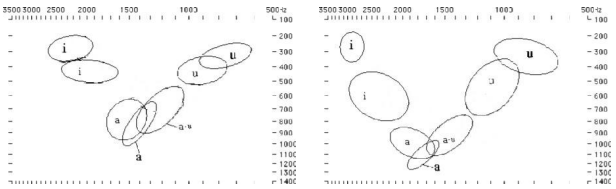


Figure 2: Comparison between diphthong elements of falling diphthongs (small IPA) and their monophthongal counterparts [i a u] (large IPA): 2-sigma confidence ellipses from 10 males (left) and 10 females (right).

Figure 3 compares diphthong elements of rising diphthongs in small IPAs with their monophthongal counterparts [i a u] in large IPAs. As shown in the figure, although there is certain coarticulatory effect, both the onset and offset in a rising diphthong have comparable vowel heights and elliptic distributions

with their monophthongal counterparts in the acoustic F1/F2 plane, suggesting that diphthong elements are as well controlled as are the corresponding monophthongs. In short, formant data suggest that both the onset and offset seem to have spectral targets in rising diphthongs [ia ua].

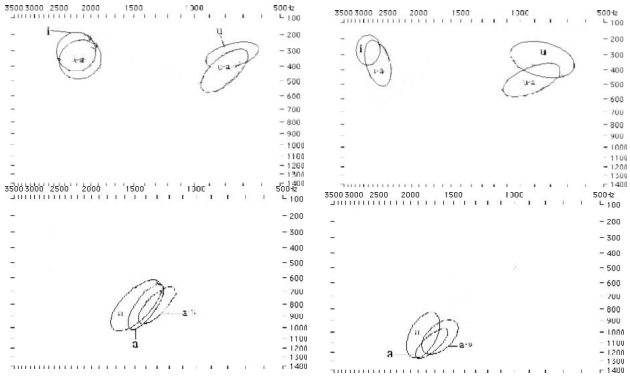


Figure 3: Comparison between diphthong elements of rising diphthongs (small IPA) and their monophthongal counterparts [i a u] (large IPA): 2-sigma confidence ellipses from 10 males (left) and 10 females (right).

3.2. Articulatory uniformity

Diphthong production could be understood as the course of spectral movement from its onset to offset. It has been shown so far that both onset and offset in [ia ua] have comparable spectral targets as their monophthongal counterparts do, whereas [ai au] seem to have a dynamic spectral target since their onset is much better controlled than their offset. A further step inquiring into the production mechanism of diphthongs is to examine the kinematics of diphthong articulation, especially the lingual articulation. Figure 4 and 5 show mean articulatory positions in millimetre (n=5) for the sampled articulators in the production of falling diphthongs [ai au] and rising diphthongs [ia ua] from 6 speakers. Diphthong elements were denoted by large IPA symbols and the three tongue points were connected by solid lines; the corresponding monophthongs were denoted by small IPA symbols and the three tongue points were connected by dot lines. Speakers are facing left in the figures.

Although there are inter-speaker variations, the onset in falling diphthongs [ai au] has articulatory positions comparable to those for the monophthong [a] in general, in terms of both lingual articulation and jaw position. However, there are a few exceptions. For instance, [a] in [ai au] has a higher jaw position than the monophthong [a] in female speaker 1. Unlike the onset, the offset in [ai au] apparently doesn't reach the articulatory positions for the monophthong [i] or [u], in terms of either lingual articulation or jaw position. The articulatory data are consistent with formant data discussed in Section A. It seems that in the production of falling diphthongs [ai au] in Ningbo Wu, the onset has both articulatory and acoustic targets, whereas the offset is probably a dynamic consequence of diphthong movement, as it has no specific articulatory or acoustic targets.

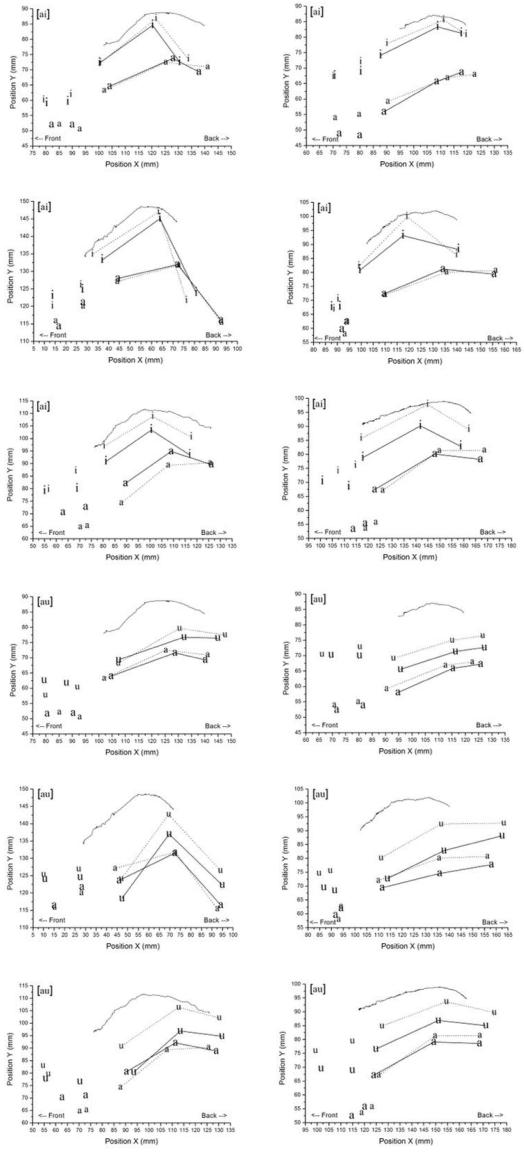


Figure 4: Mean lower lip, jaw, and three points on the tongue in mm (from left to right) in the production of falling diphthongs [ai au].

The articulation of the rising diphthongs [ia] demonstrates a similar pattern to that of falling diphthongs. First, [i] in [ia] has generally comparable articulatory positions to the monophthong [i] in terms of both lingual articulation and jaw position. This corroborates the formant data, suggesting a good control for the onset target. But what is not expected is that [a] in [ia] has apparently different articulatory positions with the

monophthong [a] in most speakers. The speakers don't seem to employ a specific articulatory strategy for the realization of the offset spectral target in the production of rising diphthong. Instead, there is articulatory uniformity regarding both falling and rising diphthongs.

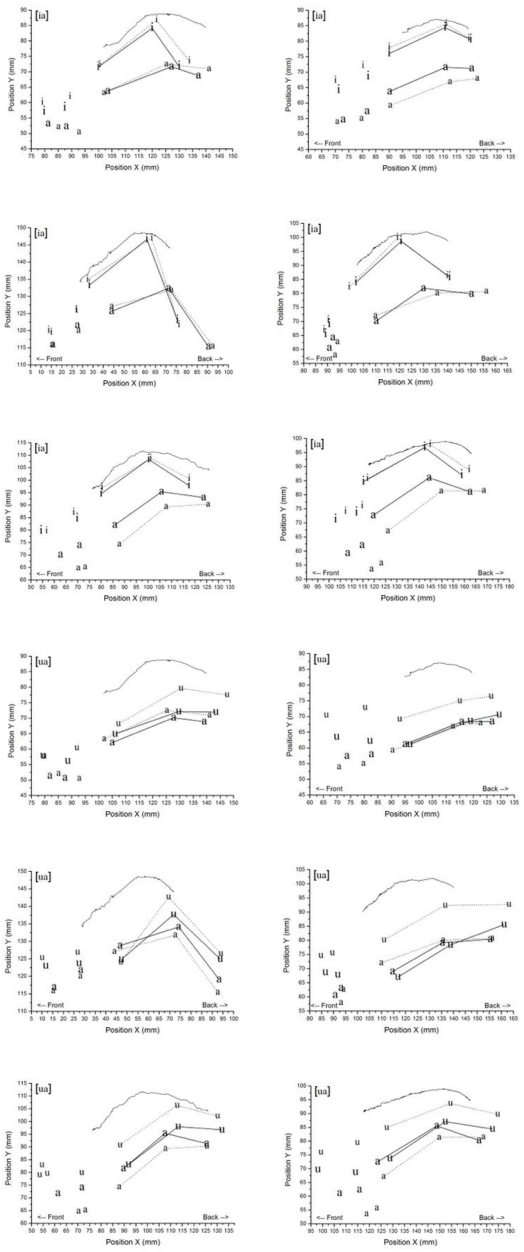


Figure 5: Mean lower lip, jaw, and three points on the tongue in mm (from left to right, n=5) in the production of rising diphthongs [ia ua].

The question is why a rising diphthong has a better offset controlled spectral offset. This could possibly be explained by the articulatory-to-acoustic relations. EMA recording sampled three tongue points in the mouth. But the constriction location for the low vowel [a] is in lower pharynx [20, 21]. So, it is reasonable to speculate that the observed variability in lingual articulation doesn't affect the realization of articulatory target in pharynx, since a general low lingual position, rather than a precisely controlled specific position, facilitates the constriction in pharynx. The fact that the offset [a] in a rising diphthong has great variability in articulation but a good controlled spectral target corroborates the quantal nature of the low vowel [a] [20, 22, 23, 24, 25]. Regarding the other rising diphthong [ua], the onset [u] in [ua] demonstrates anticipatory effect of coarticulation, namely it has apparently lower lingual and jaw positions than the monophthong [u]. As shown in Fig.3 in Section A, [u] in [ua] also have a lower distribution than the monophthong [u] in the acoustic F1/F2 vowel plane. However, they are both located in the high back region in the vowel plane. The offset [a] in [ua] behaves similar to [a] in [ia].

In summary, articulatory uniformity is observed for the production of both falling and rising diphthongs. Speakers seem to employ a uniform articulatory strategy to produce a diphthong. First, no matter in a falling or rising diphthong, diphthong onset is better controlled than diphthong offset both articulatorily and acoustically, although it subject to certain coarticulatory effect. Second, both falling and rising diphthongs are constrained by articulatory kinematics as well as the inherent sluggishness of articulators [26, 27]. Third, the difference in acoustic outputs for falling and rising diphthongs is constrained by the articulatory-to-acoustic relation and the quantal nature of the three point vowels [i a u]. As a result, falling diphthongs have a dynamic spectral target, while rising diphthongs are composed of two static spectral targets.

4. CONCLUSION

Acoustic data suggest a typological difference between falling and rising diphthongs. Falling diphthongs seem to have a dynamic spectral target, while rising diphthongs are sequences of two spectral targets. But articulatory kinematics reveals an articulatory uniformity in diphthong production, and the observed spectral difference between falling and rising diphthongs could be attributed to the articulatory-to-acoustic relations and the quantal nature of the three point vowel targets [i a u].

Results support a dynamic view of vowel production. That is, the fundamental unit of vowel could be dynamic. In Ningbo Chinese that has been examined, a falling diphthong has a dynamic spectral target, and should thus be treated as one vowel phoneme; a rising diphthong has two static spectral targets, and can thus be viewed as a sequence of two vowel phonemes.

7. REFERENCES

[1] Hartman, L. M. 1944. The segmental phonemes of the Peiping dialect. *Language* 20, 28-42.

[2] Luo, C., Wang, J. 1957. *An Outline of General Phonetics* [in Chinese]. Science Press, Beijing.

[3] Cheng, R. L. 1966. Mandarin phonological structure. Journal of Linguistics, 2, 135-158.

[4] Chao, Y.-R. 1968. *A grammar of spoken Chinese*. University of California Press, Berkeley.

[5] Cheng, C. C. 1973. *A synchronic phonology of Mandarin Chinese*. Monograph on Linguistic Analysis, No. 4. Mouton.

[6] You, R., Qian, N., Gao, Z. 1980. On the phonemic system of Putonghua [in Chinese], *Chinese Language* 5, 328-334.

[7] Duanmu, S. 2008. *Syllable Structure: The Limits of Variation*. Oxford University Press.

[8] Duanmu, S. 2017. Syllable and syllable structure in Chinese. In: Sybesma, R. (editor-in-chief), Behr, W., Gu, Y., Handel, Z., Huang, C.-T. J., Myers, J. (eds.) *Encyclopedia of Chinese Language and Linguistics* Brill, 230-236.

[9] Chao, Y.-R. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology*, Academia Sinica, 4, 363-397.

[10] Malmberg, B. 1963. *Structural linguistics and human communication*. Springer-Verlag.

[11] Abercrombie, D. 1967. *Elements of general phonetics*. Edinburgh University Press.

[12] Catford, I. 1977. *Fundamental problems in phonetics*. Edinburgh University Press.

[13] Sweet, H. 1877. *A handbook of phonetics including a popular exposition of the principles of spelling reform*. Clarendon Press, Oxford.

[14] Jones, D. 1922. *Outline of English phonetics*, 2nd Ed. E. P. Dutton, New York.

[15] Löfqvist, A., Gracco, V. L., Nye, P. W. 1993. Recording speech movements using magnetometry: One laboratory's experience. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* 31, 143-162.

[16] Löfqvist, A. 1999. Interarticulator phasing, locus equations, and degree of coarticulation. *The Journal of the Acoustical Society of America* 106, 2022-2030.

[17] Chao, Y.-R. 1928. *Studies in the Modern Wu Dialects*. Peking: Tsing Hua College Research Institute.

[18] Yuan, J. et al. 1960. *An outline of Chinese dialects* [in Chinese]. Script Reformation Press, Beijing.

[19] Hu, F. 2014. *A Phonetic Study on the Vowels in Ningbo Chinese*. China Social Sciences Press.

[20] Stevens, K. N. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In: David, E. E. Jr. & Denes, P. B. (eds.). *Human Communication: A Unified View*. McGraw-Hill, 51-66.

[21] Wood, S. 1979. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics* 7: 25-43.

[22] Stevens, K. N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17: 3-45.

[23] Perkell, J. S., Nelson, W. L. 1982. Articulatory targets and speech motor control: A study of vowel production. In Grillner, S., Persson, A., Lindblom, B. Lubker J. (eds.). *Speech Motor Control*. Pergamon, 187-204.

[24] Perkell, J. S., Nelson, W. L. 1985. Variability in production of the vowels /i/ and /a/. *The Journal of the Acoustical Society of America* 77: 1889-1895.

[25] Perkell, J. S., Cohen, M. H. 1989. An indirect test of the quantal nature of speech in the production of the vowels /i/, /a/ and /u/. *Journal of Phonetics* 17: 123-133.

[26] Saltzman, E. L., Kelso, J. A. S. 1983. Skilled actions: A task dynamic approach. *Haskins Laboratories Status Report on Speech Research* SR-76: 3-50.

[27] Kelso, J. A. S., Saltzman, E. L., Tuller, B. 1986. The dynamical perspective on speech production: Data and theory. *Journal of Phonetics* 14: 29-59.

Vowels and Diphthongs in Chengdu Mandarin

Nan Li[1] & Fang Hu[1, 2]

[1]University of CASS, [2]Institute of Linguistics, Chinese Academy of Social Sciences
linanholmes@163.com, hufang@hotmail.com

ABSTRACT

This paper gives an acoustic phonetic analysis of the monophthongs and diphthongs in Chengdu Mandarin. Acoustic data suggest a typological difference between falling diphthongs and rising diphthongs. Rising diphthongs are sequences of vowels, since both onset and offset elements have comparable spectral targets as do their monophthongal counterparts. By contrast, falling diphthongs are dynamic units of vowels, since they are featured by the dynamicity rather than could be decomposed to a sequence of two spectral targets. In conclusion, fine-grained phonetic details support a dynamic view of vowel production. And falling diphthongs should be grouped with monophthongs in the account of vowel phonology in Chengdu Mandarin.

Keywords: vowel, monophthong, diphthong, Chengdu Mandarin, southwestern Mandarin.

1. INTRODUCTION

Chengdu is the capital city of Sichuan Province in southwestern China, and the Chengdu dialect is the representative dialect of southwestern Mandarin [1]. The segmental phonology of Chengdu Mandarin resembles that of Standard Chinese or Beijing Mandarin. A major difference lies in that there is no retroflex sibilants in the consonant inventory of Chengdu Mandarin. In this connection, the segmental inventory of Chengdu Mandarin is like a simplified version of Standard Chinese. And previous phonetic researches on Chengdu Mandarin mainly focused on tones and tone sandhi [2, 3, 4].

Various solutions have been proposed for the analysis of syllable structure for Standard Chinese or Beijing Mandarin [5, 6, 7, 8, 9, 10, 11, 12]. And the major issues are general to all dialects in the Mandarin dialect family. Scholars disagree regarding the number of vowel phonemes. There are basically two different approaches. The first approach is to take vowel elements as is transcribed. In a Mandarin syllable CGVCo, a glide and a vowel GV form a rising diphthong, a vowel and a vocalic coda VCo form a falling diphthong, and the three components GVCo form a triphthong. It is not a difficult task to account for vowel contrasts from a structural distributional analysis of GVCo sequences in Mandarin dialects. Therefore, a five-vowel proposal is favoured in the literature: in addition to the three high vowels [i y u], there are a mid-vowel [e/o/ə/ɤ] and a low vowel [a] [7, 8, 11, 12, 13, 14].

The second approach considers the nature of diphthongs. The unit of vowel could be dynamic and a dynamic vowel phoneme is composed of more than one element. For instance, [10] considered falling diphthongs in Standard Chinese as one phoneme; and [15] even included all monophthongs, diphthongs and triphthongs in the vowel inventory of Standard Chinese.

However, previous studies are mostly phonological analyses focusing on Standard Chinese or Beijing Mandarin. This paper examines vowel production in Chengdu Mandarin, and aims to provide acoustic phonetic details for the analysis of Mandarin vowel phonology. As shown in Table 1, Chengdu Mandarin has 8 monophthongs [ɿ i u y a e o ɚ], 4 falling diphthongs [ai au ei ou] and 6 rising diphthongs [ia ie io ua ue ye]. The rhoticized vowel [ɚ] only appears in open syllables. Due to the space limit, discussion is focused on monophthongs and diphthongs in open syllables.

vowels	(C)(G)V	(C)(G)VN
Monophthongs	ɿ i u y a e o ɚ	ã ən in yn aŋ oŋ
Rising diphthongs	ia ie io ua ue ye	iã uã yã uən iaŋ uaŋ yoŋ
Falling diphthongs	ai au ei ou	
Triphthongs	uai uei iau iou	

Table 1: Vowel inventory of Chengdu Mandarin.

2. METHODOLOGY

10 native adult speakers, 5 male and 5 female, provided the speech data during a fieldwork in 2015. All speakers are born and raised up in Chengdu and have no reported speech and hearing disorders.

Meaningful monosyllabic words were used as test words. Each test word contains a target vowel with an initial of labial stop. Each test word was placed in a carrier sentence [X, tu X san pian] (X, read X three times). 5 repetitions were recorded.

The 16-bit audio sound was recorded directly into a laptop PC through a TerraTec DMX 6Fire USB sound card with a SHURE SM86 microphone. The sample rate is 11,025 Hz.

The annotation work is done in Praat 6.0.40 [16]. Diphthongs were annotated as being composed of three segments, namely an onset, a transition and an offset. The lowest four formants were extracted in the midpoint of each target element and the duration of each segment was also measured.

3. RESULTS

3.1 Monophthongs

Figure 1 shows the distribution of the 8 Chengdu monophthongs [ɿ i u y a e o ɚ] for male (upper) and female (lower) speakers respectively in a two-dimensional acoustic plane (F1 against F2) with the origin of the axes to the top right of the plot. The ordinates are bark scaled, but the values along the axes are still labelled in Hz. Each 2-sigma ellipse is based on 50 data points (5 repetitions × 5 speakers × 2 test syllables).

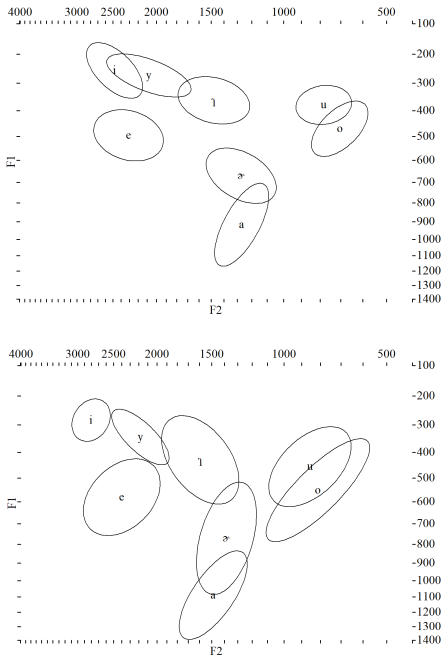


Figure 1: 2-sigma ellipses for the Chengdu monophthongs in males (upper) and females (lower).

The 8 monophthongs of Chengdu have a triangular distribution in the acoustic vowel space. There is a 3-way vowel height, a 2-way backness, and a 2-way lip rounding. [ɿ i y u] are high vowels, [e o] are mid vowels, and [a] is low; [i y e] are front, and [u o] are back; [ɿ i e a] are unrounded, and [y u o] are rounded. However, only front high vowels distinguish in lip rounding, and all back vowels are rounded by default. In addition to the major vowel features, [ɚ] contrasts with other vowels in rhoticity.

It should be noted that the ellipses for the two back vowels [u o] overlap with each other. This is a characteristic that is also attested in other southwestern Mandarin dialects such as Longchang [17] and Changde [18], but is not common for other dialects such as Hangzhou Wu [19] or Xupu Xiang [20]. This is probably due to the fact that in southwestern Mandarin, [u] is usually pronounced as an approximant [ʊ] or sometimes even featured with a slight friction.

3.2 Diphthongs

3.2.1. Temporal structure

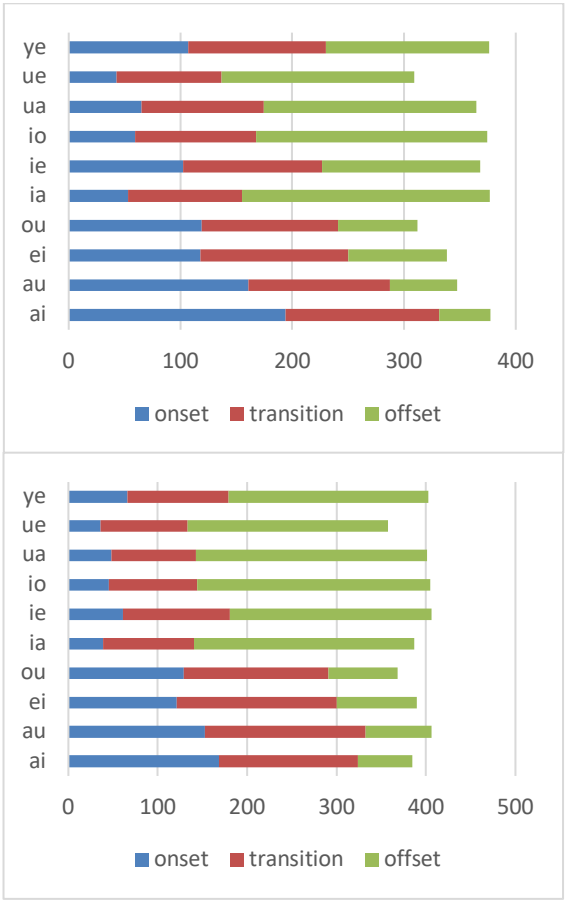


Figure 2: Temporal structure in millisecond of diphthongs in males (upper) and females (lower).

Figure 2 and 3 show the temporal structures in millisecond and in percentage of Chengdu diphthongs respectively. There is an apparent difference between the two types of diphthongs. The onset is the dominating component in falling diphthongs while the offset is the dominating one in rising diphthongs. Mandarin falling diphthongs are known for their smooth gliding transitions. That is, the transition

plays an important role, which is demonstrated by a long duration in the figure. This is quite different from the dialects such as Taiyuan Jin [21] whose falling diphthongs have abrupt gliding transitions with a short duration.

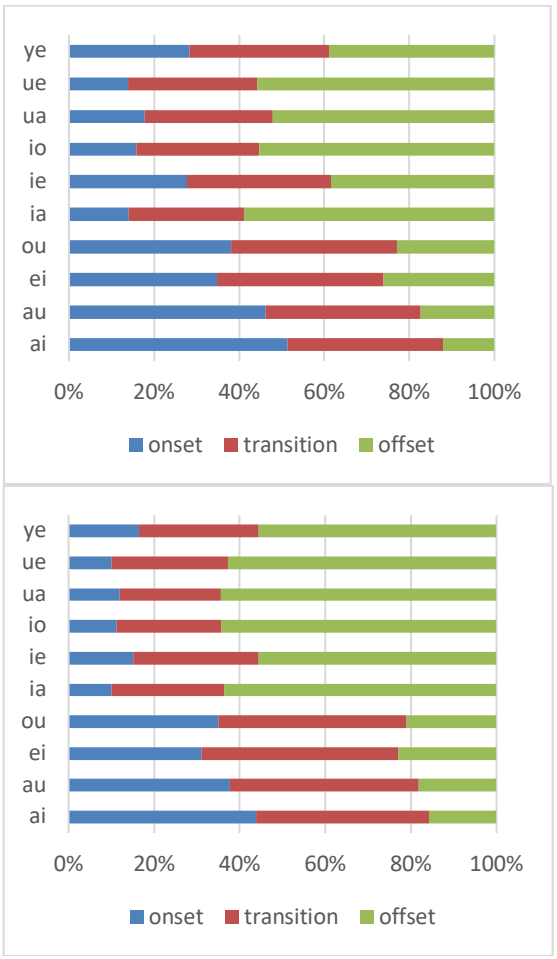


Figure 3: Temporal structure in percentage of diphthongs in males (upper) and females (lower).

3.2.2 Spectral properties

Figures 4 and 5 compare 2-sigma ellipses of the onset and offset elements in falling diphthongs and their monophthongal counterparts in acoustic F1/F2 vowel planes. A diphthong element is denoted by the other element in parenthesis.

The onsets in [ai au] exhibit comparable variability with the corresponding monophthongs, and their ellipse extensively overlap with each. In contrast, the offsets in [ai au] are distributed apparently different to their monophthongal counterparts, and exhibit more variability. That is, only the onsets in [ai au] have comparable spectral targets as do their monophthongal counterparts, but the offsets do not.

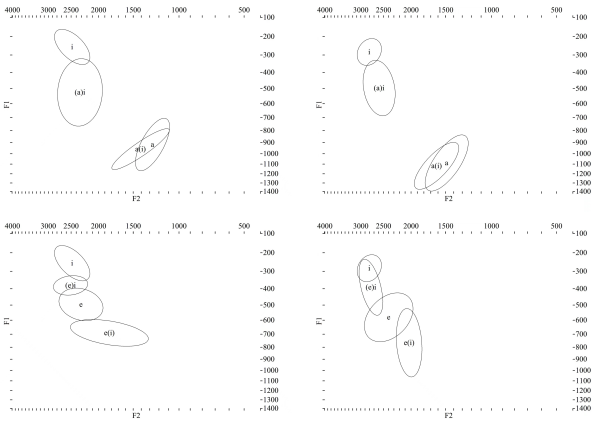


Figure 4: 2-sigma ellipses for the onset and offset of [ai ei] and corresponding monophthongs in males (left) and females (right).

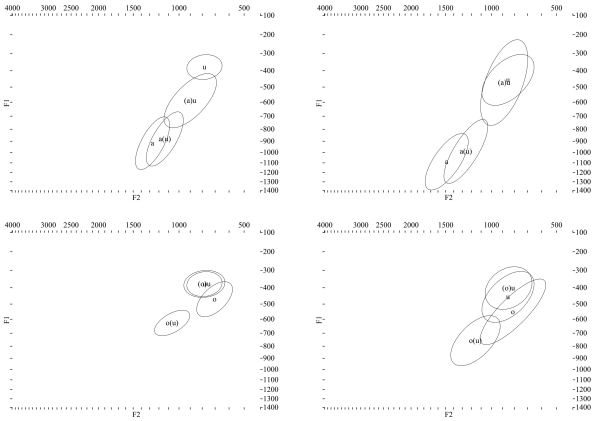


Figure 5: 2-sigma ellipses for the onset and offset of [au ou] and corresponding monophthongs in males (left) and females (right).

The other two falling diphthongs [ei ou] demonstrate a totally different pattern. The production of [ei ou] begins from a lower region and ends at a higher region than the corresponding onset monophthong [e o] respectively. A comparison between diphthong elements and their monophthongal counterparts is not a viable way to evaluate the formant data, since the ellipses for the onsets in both diphthongs [ei ou] and the offset in [ou] have spectral distributions apparently different to their monophthongal counterparts but the ellipses for the offset in [ou] and the monophthong [u] happen to overlap extensively. What is crucial here is that [ei ou] are dynamic high vowels, which contrast with dynamic low vowels [ai au] respectively on the one hand, and contrast with static mid-high vowels [e o] on the other hand. In summary, it is not justifiable to treat falling diphthongs as sequences of vowels, as not all diphthong elements have comparable spectral distribution as do their monophthongal counterparts. Rather, falling diphthongs should be treated as

dynamic vowels. Therefore, accepting the dynamicity as a vowel feature help understand distinctive relations of vowel phonology in Chengdu Mandarin as well as in other Chinese dialects in general.

Figures 6 to 8 compare 2-sigma ellipses of the onset and offset elements in rising diphthongs and their monophthongal counterparts in the acoustic F1/F2 vowel plane.

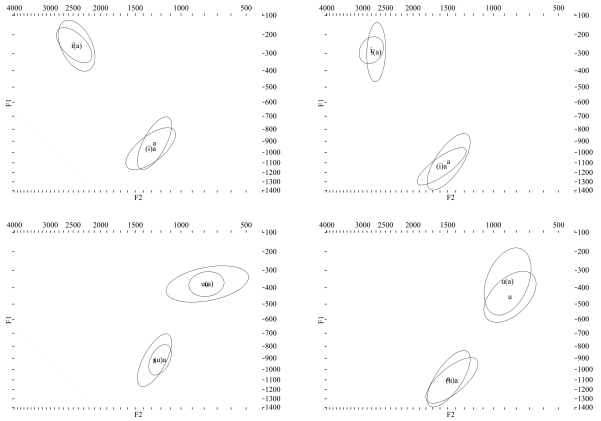


Figure 6: 2-sigma ellipses for the onset and offset of [ia ua] and corresponding monophthongs in males (left) and females (right).

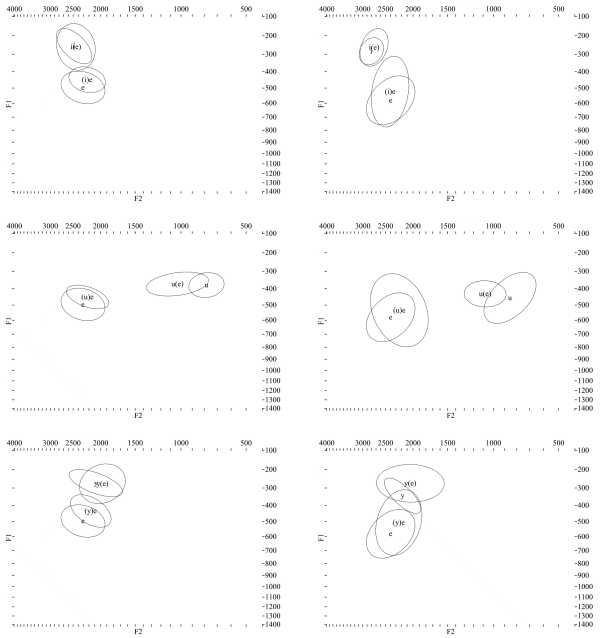


Figure 7: 2-sigma ellipses for the onset and offset of [ie ue ye] and corresponding monophthongs in males (left) and females (right).

Unlike in falling diphthongs, both onset and offset elements in rising diphthong have spectral distributions comparable to their monophthongal counterparts, since their ellipses overlap with each other and exhibit comparable variabilities in general. Therefore, rising diphthongs could be treated as sequences of vowels, since both onsets and offsets have comparable spectral targets as do the corresponding monophthongs. There are surely certain degree of coarticulatory effect on diphthong elements. For instance, it seems that the onset of [io] demonstrates an apparent effect of coarticulation, and its distribution is shifted to a spectral region posterior to the monophthong [i].

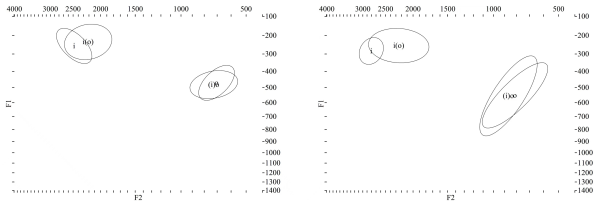


Figure 8: 2-sigma ellipses for the onset and offset of [io] and corresponding monophthongs in males (left) and females (right).

In summary, formant data suggest a typological difference between falling diphthongs and rising diphthongs in Chengdu Mandarin. Rising diphthongs are sequences of vowels whereas falling diphthongs should better be understood as a dynamic unit of vowel.

4. CONCLUSION

This paper gives an acoustic phonetic analysis of the monophthongs and diphthongs in Chengdu Mandarin. Both durational data and spectral data reveal a typological difference between falling diphthongs and rising diphthongs. Fine grained phonetic details support a dynamic view of vowel production [22, 23]. Rising diphthongs are vowel sequences, as both onset and offset elements have spectral targets comparable to the corresponding monophthongs. Falling diphthongs are dynamic units of vowels. The production of falling diphthongs are characterized by the feature of dynamicity rather than an onset target and an offset target, since not all onset and offset elements have spectral targets comparable to their monophthongal counterparts in Chengdu Mandarin.

Therefore, both monophthongs and falling diphthongs should be taken into account for vowel phonology of Chengdu Mandarin. For instance, in addition to the three static levels of distinction in vowel height, there are two more dynamic levels of height distinction. That is, Chengdu Mandarin has a 5-way distinction of vowel height, namely [i u] are static high vowels, [e o] are static mid vowels, [a] is the static low vowel, [ei ou] are dynamic high vowels, and [ai au] are dynamic low vowels.

REFERENCES

[1] Li, L. 2009. Classification/Distribution of Southwest Mandarin, *Dialect*, vol. 31, no. 01, 72–87.

[2] Cun, X., Zhu, X. 2013. Tonotypes of Chengdu-Chongqing Mandarins, *Studies in Language and Linguistics*, 1–11.

[3] He, W. 2015. The Social Linguistic Research of the Monosyllabic Tones of Chengdu Dialect, *Papers on Languages and History*, 178–198.

[4] Guo, W. 2020. A Study on Sandhi of Chengdu Dialect, Master's Theses of Nanjing Normal University.

[5] Hartman, L. M. 1944. The segmental phonemes of the Peiping dialect. *Language* 20, 28-42.

[6] Luo, C., Wang, J. 1957. *An Outline of General Phonetics* [in Chinese]. Science Press, Beijing.

[7] Cheng, R. L. 1966. Mandarin phonological structure. *Journal of Linguistics*, 2, 135-158.

[8] Chao, Y.-R. 1968. *A grammar of spoken Chinese*. University of California Press, Berkeley.

[9] Cheng, C. C. 1973. *A synchronic phonology of Mandarin Chinese*. Monograph on Linguistic Analysis, No. 4. Mouton.

[10] You, R., Qian, N., Gao, Z. 1980. On the phonemic system of Putonghua [in Chinese], *Chinese Language* 5, 328-334.

[11] Duanmu, S. 2007. *The phonology of Standard Chinese*, 2nd edition. Oxford University Press.

[12] Duanmu, S. 2017. Syllable and syllable structure in Chinese. In: Sybesma, R. (editor-in-chief), Behr, W., Gu, Y., Handel, Z., Huang, C.-T. J., Myers, J. (eds.) *Encyclopedia of Chinese Language and Linguistics* Brill, 230-236.

[13] Wan, I.-P., Jaeger, J. 2003. The phonological representation of Taiwan Mandarin vowels: A psycholinguistic study. *Journal of East Asian Linguistics*, 12(3), 205–257.

[14] Lin, Y.-H. 2007. *The sounds of Chinese*. Cambridge University Press.

[15] Lee, W.-S., Zee, E. 2003. Standard Chinese, *Journal of the International Phonetic Association* 33, 109-112.

[16] Boersma, P., Weenink, D. 2017. *Praat, a system for doing phonetics by computer (Version 6.0. 28)*, https://www.fon.hum.uva.nl/praat/.

[17] Qiu, Y., Hu, F. 2013. Longchang vowels, *Bulletin of Linguistic Studies* 10: 38-51.

[18] Zhang, Z., Hu, F. 2019. Vowels and diphthongs in the Changde Mandarin Chinese. *Proceedings of ICPhS 2019*, Melbourne, Australia.

[19] Yue, Y., Hu, F. 2018. Vowels and diphthongs in Hangzhou Wu Chinese dialect, *Proceedings of Interspeech 2018*, Hyderabad, 207-211.

[20] Zhang, Z., Hu, F. 2019. Vowels and Diphthongs in the Xupu Xiang Chinese Dialect. *Proceedings of Interspeech 2019*, Graz, 3960–3964.

[21] Xia, L., Hu, F. 2016. Vowels and Diphthongs in the Taiyuan Jin Chinese Dialect. *Proceedings of Interspeech 2016*, San Francisco, 993–997.

[22] Hu, F. 2013. Falling diphthongs have a dynamic target while rising diphthongs have two targets: Acoustics and articulation of the diphthong production in Ningbo Chinese, *Journal of the Acoustical Society of America*, 134, 4199.

[23] Hu, F. 2020. *The Vowel: A General Introduction with Reference to Chinese Data*. Foreign Language Teaching and Research Press.

PATTERNS OF THE SYLLABLE-FINAL CONSONANTS IN CHINESE DIALECTS

Wai-Sum Lee; Eric Zee

Department of Linguistics and Translation, City University of Hong Kong
w.s.lee@cityu.edu.hk

ABSTRACT

This paper presents the patterns of the syllable-final consonants in the present-day Chinese dialects. It reports the changes in the historical final consonants, -**p* -**t* -**k* -**m* -**n* -**ŋ*, taken place in the dialects, resulting in the formation of the different patterns. The study is based on the consonant inventories of a representative sample of 70 dialects of the 11 Chinese dialect groups. The findings are, (i) there is a total of 15 patterns of the syllable-final consonants; (ii) many patterns are defective, where one or more of the six historical final consonants, -**p* -**t* -**k* -**m* -**n* -**ŋ*, are missing; (iii) the changes in the syllable-final consonants are conditioned by the preceding vowel type; (iv) the changes are bi-directional in terms of place of articulation; and (v) the changes result in the emergence of nasal vowels and syllable-final glottal stop in the dialects.

Keywords: patterns of syllable-final consonants, consonant change, Chinese dialects.

1. INTRODUCTION

The paper presents (i) the patterns of the syllable-final consonants, which include the unreleased oral stops [-p -t -k], glottal stop [-ʔ] and nasals [-m -n -ŋ], in the present-day Chinese dialects, and (ii) the formation of the patterns. The syllable-final consonants are the reflexes of the six historical final consonants, -**p* -**t* -**k* -**m* -**n* -**ŋ*, that occurred in the sound systems of Old Chinese and Middle Chinese ([1, 2, 3]). The study is based on the consonant inventories of a representative sample of 70 genetically and areally balanced dialects of the 11 Chinese dialect groups, including *Mandarin*, *Jin*, *Hui*, *Gan*, *Xiang*, *Wu*, *Min*, *Yue*, *Kejia*, *Pinghua*, and *Tuhua*. Following the quota rule proposed in [4], a dialect is selected from each one of the 70 subgroups of the 11 Chinese dialect groups. The reference sources on the sound systems of the present-day Chinese dialects for investigation include (i) journal articles in *Fangyan* (Dialect) published by the Institute of Linguistics at the Chinese Academy of Social Sciences, Beijing, (ii) book chapters, (iii) monographs, and (iv) dissertations on the Chinese dialects. The classification of the dialect groups and their subgroups follows the proposals in the published dialect map, LANGUAGE ATLAS OF CHINA ([5]), and the journal article, '*Classification of the Chinese dialects*' ([6]).

2. PATTERNS

Table 1 presents a total of 15 patterns of the syllable-final consonants in the 70 dialects of the 11 Chinese dialect groups. In the column under 'Dialect groups', the number in parentheses denotes the number of dialects of a dialect group, in which a particular pattern occurs. 'Freq[1]' refers to the frequency of occurrence of the 15 individual patterns, and 'Freq[2]' refers to the frequency of occurrence of the individual syllable-final stops and nasals.

Patterns								Freq[1]	Dialect groups
#1	-p	-t	-k		-m	-n	-ŋ	15	*Yue* (8), *Kejia* (4), *Min* (2), *Pinghua* (1)
#2						-n	-ŋ	14	*Mandarin* (5), *Gan* (3), *Xiang* (2), *Hui* (2), *Min* (1), *Tuhua* (1)
#3				-ʔ			-ŋ	9	*Jin* (3), *Wu* (3), *Mandarin* (1), *Kejia* (1), *Min* (1)
#4							-ŋ	9	*Mandarin* (2), *Wu* (2), *Gan* (2), *Min* (1), *Xiang* (1), *Tuhua* (1)
#5				-ʔ		-n	-ŋ	6	*Wu* (3), *Kejia* (2), *Jin* (1)
#6						-n		3	*Xiang* (2), *Hui* (1)
#7				-ʔ				3	*Jin* (2), *Hui* (1)
#8		-t	-k			-n	-ŋ	2	*Kejia* (1), *Gan* (1)
#9	-p		-k		-m		-ŋ	2	*Min* (2)
#10		-t		-ʔ		-n	-ŋ	2	*Kejia* (1), *Gan* (1)
#11	-p	-t	-k	-ʔ	-m	-n	-ŋ	1	*Min* (1)
#12	-p			-ʔ	-m	-n	-ŋ	1	*Gan* (1)
#13					-m		-ŋ	1	*Min* (1)
#14				-ʔ	-m	-n		1	*Hui* (1)
#15								1	*Pinghua* (1)
Freq[2]	19	20	20	23	21	45	62	70	

Table 1: Patterns of the syllable-final consonants in the 70 dialects of the 11 Chinese dialect groups.

The 15 patterns of the syllable-final consonants (Table 1) may be grouped into five types. <u>*Type I*</u> consists of six patterns, [-p -t -k -m -n -ŋ] (15, i.e.,

occurring in 15 dialects), [-t -k -n -ŋ] (2), [-p -k -m -ŋ] (2), [-t -ʔ -n -ŋ] (2), [-p -t -k -ʔ -m -n -ŋ] (1), and [-p -ʔ -m -n -ŋ] (1), each of which contains one or more syllable-final oral stops and two or more syllable-final nasals, and also a syllable-final glottal stop in some patterns. *Type II* consists of three patterns, [-ʔ -ŋ] (9), [-ʔ -n -ŋ] (6), and [-ʔ -m -n] (1), that contain a syllable-final glottal stop and one or two syllable-final nasals, but not the syllable-final oral stops. *Type III* consists of a single pattern, [-ʔ] (3), that contains only a syllable-final glottal stop. *Type IV* consists of four patterns, [-n -ŋ] (14), [-ŋ] (9), [-n] (3), and [-m -ŋ] (1), that contain only the syllable-final nasals. *Type V* consists of a single pattern (1) that contains a zero syllable-final consonant ('Ø'). It occurs in *Zhongshan* (*Pinghua*), the only one of the 70 dialects in this study that lacks all the syllable-final consonants. There are no occurrences of the patterns that contain only a syllable-final glottal stop with the addition of one or more oral stops, such as ˣ[-ʔ -p], ˣ[-ʔ -p -t] and ˣ[-ʔ -p -t -k], nor the patterns that contain only the syllable-final oral stops, such as ˣ[-p], ˣ[-t], ˣ[-k], ˣ[-p -t], ˣ[-p -k], ˣ[-t -k] and ˣ[-p -t -k]. In essence there are no patterns that contain only the syllable-final oral stops, but patterns that contain only a glottal stop [-ʔ] (*Type III*) or only the syllable-final nasals (*Type IV*).

Table 1 also presents the frequencies of occurrence of the syllable-final oral stops, glottal stop and nasals. In order of decreasing frequency of occurrence, they are [-ŋ] (62) > [-n] (45) > [-ʔ] (23) > [-m] (21) > [-t] (20) = [-k] (20) > [-p] (19). The velar nasal [-ŋ] occurs in 62 of the 70 dialects. It is more frequent than the other types of syllable-final consonants. Of the syllable-final stops, the glottal stop [-ʔ] is more frequent than the oral stops [-p], [-t], [-k]; and of the oral stops, [-p] is less frequent than [-t] and [-k]. The bilabials, [-m] and [-p], have a lower frequency of occurrence than the alveolar and velar nasals or oral stops. They may be said to be less stable than the other nasal and stop types, and the velar nasal [-ŋ] is more resistant to the effects of the diachronic process of final consonant elision compared to the other types of syllable-final consonants.

3. FORMATION OF PATTERNS

The various patterns of the syllable-final consonants are formed as a result of the changes in the six historical syllable-final stops and nasals, -**p* -**t* -**k* -**m* -**n* -**ŋ*, taken place in the present-day dialects.

3.1. Changes in -**p* -**t* -**k*

This section presents the changes in the three historical syllable-final stops, -**p* -**t* -**k*, that have contributed to the formation of the *Type I*, *Type II* and *Type III* patterns. As will be shown, the path to the realization of the patterns that contain the syllable-final stops involves more changes than just the drop or elision of some of the historical syllable-final stops -**p* -**t* -**k*.

Type I patterns consist of [-p -t -k -m -n -ŋ] (15), [-p -t -k -ʔ -m -n -ŋ] (1), [-p -ʔ -m -n -ŋ] (1), [-t -k -n -ŋ] (2), [-p -k -m -ŋ] (2), and [-t -ʔ -n -ŋ] (2). In both patterns, [-p -t -k -m -n -ŋ] (15) and [-p -t -k -ʔ -m -n -ŋ] (1), -**p* -**t* -**k* are preserved. The pattern [-p -t -k -ʔ -m -n -ŋ] (1) occurs in a single dialect, *Xiamen*, of the *Min* dialect group. The extra syllable-final glottal stop [-ʔ] results from the debucaalization of -**p* -**t* -**k* in some rimes; in other rimes, -**p* -**t* -**k* are preserved. In the patterns, [-p -ʔ -m -n -ŋ] (1), [-t -k -n -ŋ] (2), [-p -k -m -ŋ] (2), and [-t -ʔ -n -ŋ] (2), as can be seen there are missing syllable-final oral stops. These are the defective patterns, as they do not contain all the reflexes of -**p* -**t* -**k*. In the pattern, [-p -ʔ -m -n -ŋ] (1), which occurs in *Lichuan* (*Gan*), while -**p* is preserved, both -**t* -**k* have turned into [-ʔ] (< -**t*, -**k*), rather than being elided. The pattern [-t -k -n -ŋ] (2) occurs in *Yugan* (*Gan*) and *Wengyuan* (*Kejia*). In *Yugan*, while -**t* -**k* are preserved, -**p* has merged into [-t] (< -**p*). In *Wengyuan*, -**p* has bifurcated into [-t] (< -**p*) and [-k] (< -**p*), and -**k* is preserved in some rimes; in other rimes it has turned into [-t] (< -**k*). Despite the similarity between the patterns on the surface level in the two dialects, the changes that lead to the formation of the patterns differ. This is also true for the patterns [-p -k -m -ŋ] (2) and [-t -ʔ -n -ŋ] (2), each occurring in two dialects. The pattern [-p -k -m -ŋ] (2) occurs in *Ningde* (*Min*) and *Leizhou* (*Min*). In *Ningde*, -**t* has bifurcated into [-p] (< -**t*) and [-k] (< -**t*). In *Leizhou*, *-*t* is either dropped, resulting in changing CVS (where S = stop) to CV syllables, or it has merged into [-k] (< -**t*). The pattern [-t -ʔ -n -ŋ] (2) occurs in *Yongding* (*Kejia*) and *Nanchang* (*Gan*). In *Yongding*, -**p* -**k* have turned into [-ʔ] (< -**p*, -**k*). In *Nanchang*, *-*p* has merged into [-t] (< *-*p*) and -**k* has bifurcated into [-ʔ] (< -**k*) and [-t] (< *-*k*). While [-ʔ] occurs in both dialects, it is a derivative of both -**p* and -**k* in *Yongding*, but only -**k* in *Nanchang*.

Type II patterns consist of [-ʔ -ŋ] (9), [-ʔ -n -ŋ] (6), and [-ʔ -m -n] (1) that contain only a syllable-final glottal stop with the additional one or two syllable-final nasals, occurring in 16 dialects of six Chinese dialect groups. In all the dialects, with the exception of two, in which the three patterns occur, the syllable-final oral stops have turned into [-ʔ] (< -**p*, -**t*, -**k*). In *Yudu* (*Kejia*), only -**k* has turned into [-ʔ] (< -**k*), while -**p* -**t* are dropped; and in *Jiyuan* (*Jin*), -**p* -**t* -**k* have turned into [-ʔ] in some rimes; in other rimes, -**p* -**t* -**k* are dropped.

Type III pattern consists of [-ʔ] (3) that contains only a single syllable-final glottal stop. It occurs in

Datong (*Jin*), *Shenmu* (*Jin*) and *Jixi* (*Hui*). In *Jixi*, -**p* -**t* -**k* have turned into [-ʔ] (< -**p*, -**t*, -**k*). In *Datong* and *Shenmu*, -**p* -**t* -**k* have turned into [-ʔ] (< -**p*, -**t*, -**k*) in some rimes; in other rimes, -**p* -**t* -**k* are dropped.

Summary. The changes in the three historical syllable-final stops, -**p* -**t* -**k*, that contribute to the formation of the Type I, Type II and Type III patterns include (i) elision of -**p* -**t* -**k*, resulting in changing CVS to CV syllables; (ii) debucaalization of -**p* -**t* -**k*; (iii) bi-directional consonant shift in place of articulation, including [-t] < -**p*, [-k] < -**p*, [-p] < -**t*, [-k] < -**t*, [-t] < -**k*, but not ×[-p] < -**k*; and (iv) bifurcations of -**p* into [-t] and [-k], -**t* into [-p] and [-k], and -**k* into [-t] and [-ʔ]. The changes (ii), (iii) and (iv) are in a large majority of cases conditioned by the vowel type that precedes the syllable-final stops.

3.2. Changes in -**m* -**n* -**ŋ*

This section presents the changes in the three historical syllable-final nasals, -**m* -**n* -**ŋ*, that have contributed to the formation of the Type I, Type II and Type IV patterns. As is the case with the syllable-final stops, the path to the formation of the various patterns that contain the syllable-final nasals involves more than just the drop of some of the three historical syllable-final nasals, -**m* -**n* -**ŋ*.

Type I patterns consist of [-p -t -k -m -n -ŋ] (15), [-p -t -k -ʔ -m -n -ŋ] (1), [-p -ʔ -m -n -ŋ] (1), [-t -k -n -ŋ] (2), [-p -k -m -ŋ] (2), and [-t -ʔ -n -ŋ] (2) that contain two or more syllable-final nasals. In the patterns [-p -t -k -m -n -ŋ] (15) and [-p -ʔ -m -n -ŋ] (1), -**m* -**n* -**ŋ* are preserved in 16 dialects. The pattern [-p -t -k -ʔ -m -n -ŋ] (1) occurs only in *Xiamen* (*Min*). In the dialect, -**m* -**n* -**ŋ* are preserved in some rimes; in other rimes, they are dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*, -**ŋ*). The pattern [-t -k -n -ŋ] (2), occurs in *Wengyuan* (*Kejia*) and *Yugan* (*Gan*). In *Wengyuan*, -**m* has bifurcated and merged into [-n] (< -**m*) and [-ŋ] (< -**m*). In *Yugan*, -**m* has merged into [-n] (< -**m*). The pattern [-p -k -m -ŋ] (2), occurs in *Ningde* (*Min*) and *Leizhou* (*Min*). In both dialects, -**m* is preserved. In *Ningde* -**n* has merged into [-ŋ] (< -**n*). In *Leizhou*, -**n* is dropped in some rimes; in other rimes it has merged into [-ŋ] (< -**n*). As for -**ŋ*, it is preserved in some rimes; in other rimes it is dropped. The pattern [-t -ʔ -n -ŋ] (2) occurs in *Yongding* (*Kejia*) and *Nanchang* (*Gan*). In both dialects, -**n* and -**ŋ* are preserved. In *Yongding*, -**m* has merged into [-ŋ] (< -**m*), while in *Nanchang*, -**m* has merged into [-n] (< -**m*).

Type II patterns consist of [-ʔ -m -n] (1), [-ʔ -n -ŋ] (6), and [-ʔ -ŋ] (9), that contain one or two syllable-final nasals with the additional syllable-final glottal stop. Type IV patterns consist of [-n -ŋ] (14), [-ŋ] (9), [-n] (3), and [-m -ŋ] (1), that contain one or two syllable-final nasals. Details of the changes in respect to the patterns of Type II, but not Type IV (due to space limitations), are given below.

The pattern [-ʔ -m -n] (1) occurs in *Chun'an* (*Hui*). In the dialect, -**m* -**n* -**ŋ* are dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*, -**ŋ*) in some rimes; in other rimes, -**ŋ* has bifurcated into [-m] (< -**ŋ*) and [-n] (< -**ŋ*), -**m* has merged into [-n] (< -**m*) and -**n* is preserved. Thus, [-m] in the dialect is not the descendant of -**m* but -**ŋ*. The pattern [-ʔ -n -ŋ] (6) occurs in six dialects of three Chinese dialect groups, *Guidong, Chengdu-Longtansi* (both Kejia), *Teqing, Suzhou*, *Wenling* (all *Wu*), and *Jiyuan* (*Jin*). In *Guidong* and *Chengdu-Longtansi*, -**m* -**n* -**ŋ* are dropped in some rimes with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*, -**ŋ*); in other rimes -**n* and -**ŋ* are preserved. In *Teqing*, *Suzhou* and *Wenling*, -**m* is dropped. -**n* is dropped in some rimes; in other rimes it is preserved. -**ŋ* is preserved in some rimes; in other rimes, it is dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**ŋ*). In *Jiyuan*, -**m* has merged into [-n] (< -**m*) and -**n* and -**ŋ* are preserved. The pattern [-ʔ -ŋ] (9) occurs in nine dialects of five Chinese dialect groups, *Taiyuan*, *Changzhi* and *Lishi* (all *Jin*), *Ningbo*, *Yunhe* and *Jinhua* (all *Wu*), *Taixin* (*Mandarin*), *Yudu* (*Kejia*), and *Putian* (*Min*). In all the dialects, -**ŋ* is preserved. In *Taiyuan*, *Changzhi* and *Lishi*, -**m* -**n* are dropped in some rimes; in other rimes they have merged into [-ŋ] (< -**m*, -**n*). In *Taiyuan*, the drop of -**m* -**n* results in the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*). In *Ningbo*, -**m* has merged into [-ŋ] (< -**m*). -**n* has merged into [-ŋ] (< -**n*) in some rimes; in other rimes, it is dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**n*). -**ŋ* is dropped in some rimes with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**ŋ*). In *Yunhe,* -**m* -**n* are dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*). -**ŋ* in some rimes is dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**ŋ*). In *Jinhua*, -**m* -**n* are dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*) in some rimes; in other rimes, they have merged into [-ŋ] (< -**m*, -**n*). In both *Taixing* and *Yudu*, -**m*, -**n* are dropped with the concomitant change of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**m*, -**n*). In *Taixing*, -**m* in some rimes has merged into [-ŋ]. In *Yudu*, -**ŋ* in some rimes is dropped with the concomitant change

of the preceding vowel into a nasal vowel ([Ṽ] < 'Ø' < -**ŋ*). In *Putian*, -**m* -**n* are dropped.

Summary. The changes in the three historical syllable-final nasals, -**m* -**n* -**ŋ*, that contribute to the formation of the *Type I*, *Type II* and *Type IV* patterns include (i) elision of -**m* -**n* -**ŋ*, resulting in changing CVN (where N = nasal) to CV syllables; (ii) occurrence of nasal vowels; (iii) bi-directional consonant shift in terms of place of articulation, including [-n] < -**m*, [-ŋ] < -**m*, [-ŋ] < -**n*, [-m] < -**ŋ*, [-n] < -**ŋ*, but not ×[-m] < -**n*; and (iv) bifurcations of -**m* into [-n] and [-ŋ] and -**ŋ* into [-m] and [-n]. The changes (ii), (iii) and (iv) are in a large majority of cases conditioned by the vowel type that precedes the syllable-final nasals.

4. DISCUSSION

(1) Chen ([7]) proposes a theory of diachronic change of the historical syllable-final consonants, -**p* -**t* -**k* -**m* -**n* -**ŋ*, taken place in the Chinese dialects and postulates the successive stages of the change of the historical sounds. The sequence of the stages of the change in the syllable-final stops is, Stage 1: -**p* -**t* -**k* → Stage 2: [-t, -k] ([-t] < -**p*; -**t* -**k* unchanged) → Stage 3: [-k] ([-k] < -**t*; -**k* unchanged) → Stage 4: [-ʔ] ([-ʔ] < [-k]) → Stage 5: 'Ø' ('Ø' < [-ʔ], i.e., [-ʔ] dropped). The parallel sequence of the stages of change in the syllable-final nasals is, Stage 1: -**m* -**n* -**ŋ* → Stage 2: [-n, -ŋ] ([-n] < -**m*; -**n* -**ŋ* unchanged) → Stage 3: [-ŋ] ([-ŋ] < -**n*; -**ŋ* unchanged) → Stage 4: [Ṽ] (< 'Ø' < [-ŋ], i.e., [-ŋ] is dropped with the concomitant change of the preceding vowel into a nasal vowel) → Stage 5: V (< [Ṽ], i.e., vowel de-nasalization). Chen's postulation suggests that (i) the changes in the syllable-final consonants are unidirectional with respect to the place of articulation, that is, from front to back, [-p] > [-t] > [-k] > [-ʔ] > 'Ø' and [-m] > [-n] > [-ŋ] > [Ṽ] > [V], and (ii) the successive stages are unalterable and unskippable. The language data presented in the present study show that (a) the changes in -**p* -**t* -**k* and -**m* -**n* -**ŋ* taken place in the dialects are bi-directional consonant shift in terms of place of articulation, such as [-t] < -**p*, [-p] < -**t*; [-k] < -**t*, [-t] < -**k*; and [-ŋ] < -**m*, [-m] < -**ŋ*; [-ŋ] < -**n*, [-n] < -**ŋ*; (b) the successive stages are alterable, e.g., [-p] < -**t* and [-t] < -**k*; and skippable, e.g., [-k] < -**p* and [-m] < -**ŋ*. In the case of the syllable-final stops, there is no phonetic or phonological justification that -**p* or -**t* must first turn into [-k] before turning into [-ʔ]. Articulatorily, for -**p*, -**t* or -**k* to turn into [-ʔ], the release of the oral closure suffices to accomplish the change. As reported in Iwata, et al. ([8, 9]), the laryngoscopic data reveal that the production of the syllable-final applosives in the Chinese dialects, Fukienese and Cantonese, is glottalized, characterized by a laryngeal constriction with a closed glottis as observed in the production of the glottal stop. The glottalization prevents the vocal folds from vibrating at the vowel offset and creates the phonatory condition for effectively producing the unreleased final stops.

There are parallels between the changes in the syllable-final stops and nasals. (a) It is unnecessary for -**m* and -**n* to turn into [-ŋ] before the occurrence of [Ṽ], as the changes, [Ṽ] < 'Ø' < -**m* and [Ṽ] < 'Ø' < -**n*, occur in many dialects; (b) the diachronic changes in the historical syllable-final nasals are bi-directional, such as [-ŋ] < -**m*, [-m] < -**ŋ*; [-ŋ] < -**n*, [-n] < -**ŋ*; and (c) the successive stages are also alterable and skippable. The data in the present study thus call into question on the validity of Chen's theory of diachronic change of the historical syllable-final stops and nasals, -**p* -**t* -**k* -**m* -**n* -**ŋ*, that have taken place in the Chinese dialects.

(2) In this study, the syllable-final consonants in descending frequency of occurrence are [-ŋ] (62) > [-n] (45) > [-ʔ] (23) > [-m] (21) > [-t] (20) = [-k] (20) > [-p] (19). The number of occurrences of the syllable-final nasals, [-m] [-n] [-ŋ], is larger than that of the syllable-final oral stops, [-p] [-t] [-k], at the same place of articulation. This may be because the nasal murmur and nasality on the V-to-N transition are more perceptible than the unreleased oral stops, contributing to nasal identification and nasal place distinction ([10, 11]). As for the larger number of occurrences of [-ŋ] than that of [-m -n], it may be because in the dialects [-ŋ] occurs more frequently after the vowels [a ɑ ɔ], which have a higher intensity level ([12, 13]), contributing to a more perceptible V-to-N transition.

(3) In this study, there are many cases in which the syllable-final stops have turned into the glottal stop and the syllable-final nasals are dropped resulting in the change of the preceding vowel into a nasal vowel. It is believed that the occurrence of [-ʔ] following the loss of [-p] [-t] [-k] serves to prevent the disappearance of checked syllables and checked tones at the time of change, thus continuing to maintain the contrast between CV and CVS syllables. The same can be said about the function of the nasal vowels that occur following the loss of [-m] [-n] [-ŋ].

5. CONCLUSION

The change in the syllable-final consonants is a link in the chain of sound change. It is triggered by the change in the preceding vowel, and it in turn brings about the changes in syllable type and tone type and the appearance of the nasal vowels and glottal stop, which contribute to the shaping of the sound systems in the Chinese dialects.

6. REFERENCES

[1] Pulleyblank, E.G. 1962. The consonant system of Old Chinese. *Asia Major* 9, 58-114, 206-265.

[2] Pulleyblank, E.G. 1977-1978. The final consonants of Old Chinese. *Monumenta Serica 33*, 180-206.

[3] Pulleyblank, E.G. 1984. *Middle Chinese: A Study in Historical Phonology*. UBC Press.

[4] Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press.

[5] Wurm, S.A., Li, R. 1987. *Language Atlas of China*. Longman.

[6] Xiong, Z.H, Zhang, Z.-X. 2008. Hangyu fangyande fenqu (Classification of the Chinese dialects). *Fangyan 2*, 97-108.

[7] Chen, M. 1973. Cross-dialectal comparison: a case study and some theoretical considerations. *Journal of Chinese Linguistics 1*, 38-63.

[8] Iwata, R., Sawashima, M., Hirose, H., Niimi, S. 1979. Laryngeal adjustment of Fukienese stops - initial plosives and final applosives. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics* 13, 61-81.

[9] Iwata, R., Sawashima, M., Hirose, H. 1981. Laryngeal adjustments for syllable-final stops in Cantonese. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics* 15, 45-54.

[10] Racasens, D. 1983. Place cues for nasal consonants with special reference to Catalan. *Journal of the Acoustical Society of America* 73, 1346-1353.

[11] Kurowski, K., Blumstein, S.E. 1984. Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America* 76, 383-390.

[12] Peterson, G.E., Barney, H.L. 1952. Control methods used in a study of vowels. *Journal of the Acoustical Society of America* 24, 175-184.

[13] Lehiste, I., Peterson, G.E. 1959. Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America* 31, 428-435.

SPATIAL SOUND SYMBOLISM: A CROSS-LINGUISTIC STUDY

Alexandra Wikström, Lari Vainio & Martti Vainio

University of Helsinki, Finland
firstname.lastname@helsinki.fi

ABSTRACT

Recent behavioral evidence shows that the spatial concepts of *front* and *back* are sound symbolically tied to vowels that are pronounced either at the front or the back of the mouth, respectively. With such effects found from behavioral experiments, we examined whether such phenomenon could be found cross-linguistically from words pertaining to the concepts of front and back. The frequency counts of vowels of such words from 266 languages were extracted and analyzed. The analysis showed statistically significant differences in frequency counts of vowels between the *back* and *front* concepts. These findings suggest that the phenomenon of vowel articulation connecting to word meaning does have some cross-linguistic basis and hints at a common motor-sensory foundation for spatial sound symbolic effects.

Keywords: sound symbolism, spatial sound symbolism, vowel space, cross-linguistic research

1. INTRODUCTION

Sound symbolism assumes a non-arbitrary connection between the phonetic form and conceptual meaning of a word when it comes to certain concepts. Empirical studies have shown the effects of sound symbolism in multiple aspects, such as round and pointy shapes and large and small objects connecting to differing speech sounds ([1], [2]). Recent experimental data has also expanded this to the modality of the spatial concepts of *front* and *back* in addition to the more dynamic concepts of *forward* and *backward*. For example, [3] discovered the connection between front vowels and front concepts and back vowels and back concepts. Earlier evidence comes from Tanz's work [4] who studied the occurrence of front and back vowels in word forms indicating both spatial and temporal distances in different languages. Most commonly sound symbolism has been behaviorally researched with forced-choice tests ([1], [5]) as well as with cross-linguistic methods focusing on the universality of the phenomenon, usually by studying the lexicons of the world ([4], [6]).

With the empirical and earlier cross-linguistic research suggesting sound symbolic effects in vowels in the modality of spatial positions, we set out to discover if the vocabularies of the world would show preference of front or back vowels when it comes to the words presenting the spatial concepts of *front* and *back*, respectively.

2. METHODS

The phonetic representations of words pertaining to the concepts of *front* and *back* were extracted from the Database of Cross-Linguistic Colexifications (CLICS) [7] which includes word lists from varying datasets. All the words contained in the database have been connected to particular, catalogued, meanings. More notes on the utilizing of this database can be found in section 5.1.

Data was gathered from 266 languages, with the distribution of data points within language families shown in Table 1 and the language-typological distribution of the data in Figure 1. The dataset ended up consisting of words with the meanings of *front*, *in front of* and *forward* in the *front* category and *back*, *behind* and *backward* in the *back* category. Altogether 713 words were included, with 355 words connected to the *front* concepts and 358 words connected to the *back* concepts. Multiple words from a single language connected to a concept were included in the analysis taken that the words for a particular concept were different from each other (e.g., *behind* vs. *backward* pertaining to the concept of *back* in English), and would thus offer more varied phonetic representations, while still representing the two separate spatial concepts.

3. ANALYSIS

Out of the gathered words, the frequency counts of all vowels were extracted, and each vowel was given an (x,y) coordinate based on the location of said vowel in the IPA vowel diagram (x values 0-8, y values 0-6). This helped to create a vowel space heat map to visualize the distribution of the vowels in the contrasting concepts (Figure 2). Out of all the words, 1821 vowels were extracted, 890 from the



Figure 1: A map of the languages included in the study, locations based on Glottolog language coordinates [8].

back concepts and 931 from the *front* concepts.

Statistical analysis of the vowel frequency counts was carried out using RStudio [9]. Considering the variation in how many data points were observed due to the nature of certain vowels being more common than others, for analysis purposes the frequency counts of vowels containing zero samples in either of the front or back categories were excluded from the analysis. A Fisher's Exact Test for count data was carried out to determine the statistical significance of the differences in vowel frequency counts of the samplings of *front* and *back* concepts altogether. To analyze the groupings in more detail, a Pearson's Goodness-of-Fit Test was computed on the accumulated frequency counts of vowel groups of front vowels and back vowels. Lastly, to better study the effects of frontness/backness in individual vowels that show seemingly robust effects, individual Pearson's Goodness-of-Fit Tests were computed on the frequency counts of individual vowels that had frequency counts higher than 5 in both the *front* and the *back* condition.

4. RESULTS

The frequency count distribution among vowels is shown in Table 2. The representation of different vowels is quite varied in the data, as is their appearance in the world's languages themselves: the vowels [ɜ], [ɞ] and [ɶ] do not appear in the data at all and other rarer vowels such as [y], [ɘ] or [œ] appear only a handful of times, whereas the more common vowels, e.g. [i], [a] and [u] are heavily present [10]. It is useful to note that these more common vowels also represent clear articulatory and acoustic contrasts and are at the edges of the articulatory vowel space, and are thus an opportune

Language family	**# of data points**
Indo-European	174
Tungusic	16
Turkic	32
Basque	6
Uralic	117
Sino-Tibetan	95
Yeniseian	6
Ainu	4
Mongolic	15
Dravidian	16
Afro-Asiatic	17
Chukotko-Kamchatkan	8
Nakh-Daghestanian	25
Eskimo-Aleut	9
Yukaghir	8
Burushaski	2
Abkhaz-Adyge	8
Koreanic	4
Kartvelian	3
Japonic	4
Nivkh	1
Nuclear Trans New Guinea	18
Austronesian	8
Kehu	1
Anim	6
Lakes Plain	13
Geelvink Bay	1
Kaure-Narau	1
Tor-Orya	2
Atlantic-Congo	10
Cariban	5
Mosetén-Chimané	2
Zuni	1
Tsimshian	4
Nubian	2
Hmong-Mien	45
Pama-Nyungan	2
Saharan	2
Karok	2
Otomanguean	1
Mayan	2
Uru-Chipaya	1
Zamucoan	1
Nuclear-Macro-Je	1
Chonan	2
Araucanian	2
Kunza	3
Tupian	5

Table 1: Distribution of data points (words) per language family.

place for such sound-spatial effects to emerge that are of interest to this research.

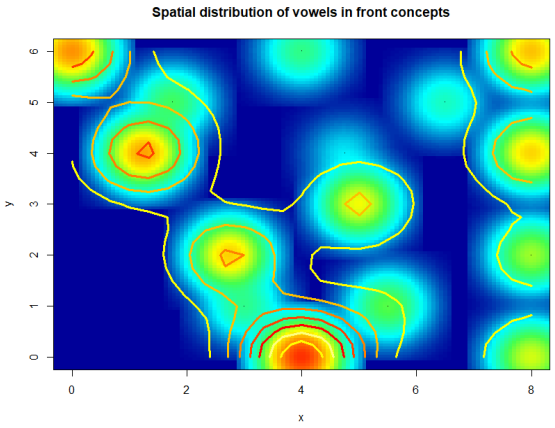
From the heat maps in Figure 2 some differences in data distribution can be observed, also visible in the frequency count data of 2. Most prominently the vowel coordinates of i/y, e/ø, ɯ/u and ɤ/o show more densely in the front concepts than in the back concepts, whereas the vowel coordinate of ɑ/ɒ is more prominent in the back concepts. Most of the data points in both back and front concepts are located in the a/ɶ coordinate.

From the statistical tests we see some significant differences between the frequency counts. With the Fisher's Exact Test a p-value of 0.009 was observed (Monte Carlo simulation, 2000 rounds). This indicates that the altogether uneven distribution of frequency counts between the two categories is

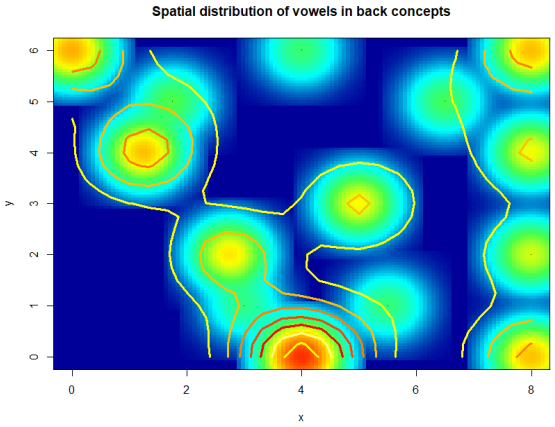
Vowel	Back	Front	Total	p-value
i	96	120	216	0.103
y	4	4	8	NA
ɨ	15	12	27	0.564
ʉ	1	2	3	NA
ɯ	23	10	33	0.024 *
u	68	82	150	0.253
ɪ	19	19	38	1
ʏ	3	0	3	NA
ʊ	19	8	27	0.034 *
e	85	108	193	0.098 .
ø	4	3	7	NA
ɘ	0	1	1	NA
ɵ	0	2	2	NA
ɤ	0	8	8	NA
o	56	72	128	0.157
ə	58	58	116	1
ɛ	66	78	144	0.317
œ	1	5	6	NA
ɜ	0	0	0	NA
ɞ	0	0	0	NA
ʌ	5	4	9	NA
ɔ	42	33	75	0.299
æ	13	13	26	1
ɐ	15	24	39	0.150
a	208	216	424	0.698
ɶ	0	0	0	NA
ɑ	83	45	128	0.0008 ***
ɒ	6	4	10	NA

Table 2: Frequency counts of vowels in the front and back concepts and the p-values of Pearson's Goodness-of-Fit Tests on the significance of frequency count differences between the concepts. Significant values are marked with *, vowels with too little data are marked NA.

statistically significant. With a Pearson's Goodness-of-Fit Test on the vowel groupings of Front vowels (N=499 back concepts, N=599 front concepts) and Back vowels (N=302 back concepts, N=266 front concepts), we obtain a p-value of 0.04 for the Front vowels group and a p-value of 0.131 for the Back vowels group. This is to say that the front vowels are significantly more frequent in the *front* concepts, and that the back vowels are more frequent in the *back* concepts, but not significantly. A Goodness-of-Fit Test was also computed for the groupings of rounded and unrounded vowels to identify any effect that the feature of roundness might have on the frequency counts. No significant differences in frequency between the groups was observed (Unrounded vowels N=628 back concepts, N=658 front concepts, p=0.403; Round vowels N=204 back concepts, N=215 front concepts, p=0.591).



(a) Vowels in front concept words



(b) Vowels in back concept words

Figure 2: Distribution of vowels in a) front concepts and b) back concepts presented in the vowel space. The rounded and unrounded versions of a vowel are represented by the same coordinate.

Looking at Table 2 with the p-values acquired from the Pearson's Goodness-of-Fit Tests on individual vowels we see that the only significant differences in frequency counts of individual vowels between the two concepts occur with the back vowels [ɑ], [ʊ] and [ɯ] which show more prominently in the *back* concepts, although the front vowel [e] comes close with a value of p=0.098 and showing prominence in the *front* concepts.

5. DISCUSSION

The results revealed both expected and unexpected outcomes. The *back* concept was mostly connected to the back vowels [ɑ], [ʊ] and [ɯ]. The *front* concept seemed to have vowels connected to it more prominently ([i, e, o u]), both front and back, but of which the back vowels were rounded. One

explanation for the emergence of [o] and [u] in the *front* words could be that lip protrusion can be seen as a way of pointing forward to an object [11]. In addition, for example the unrounded counterpart of [u], [ɯ], was significantly more frequent in the *back* concepts than in the *front* concepts. When comparing the groups of front and back vowels only the connection of front vowels to *front* concepts was significant, although there were more back vowels in the *back* concepts as well. This is in accordance with the fact that the back vowels [u] and [o] appeared more strongly, although not significantly so, in the *front* concepts and thus would not contribute to the *back* concepts as much.

Conducting cross-linguistic experiments in a way that quantifies the vowel space in the way that this study did still needs to be improved upon. Although this study concentrated on the dimensions of *front* and *back*, future analysis will also be done on other spatial concepts such as *high* and *low*, and *up* and *down*, as well as on the distribution of consonants between the two categories.

5.1. Notes on utilizing the CLICS database

The CLICS database used in this study, although fairly large, has data of varying phonetic transcription quality. For example, the NorthEuraLex database [12] and the Lalo Regional Varieties database [13] have data with concise and exact phonetic transcriptions, while the data of some other datasets such as the Intercontinental Dictionary series [14] has both phonetic and orthographic forms and had to be separately validated before being added to the data pool. The quality fluctuations of the original datasets also affected the distribution of data points among language families: for some of the languages simply more vocabulary was available, and many of the datasets focused on a particular region, adding to the varying quality of data available for a certain language family.

The CLICS database also lacks linkages of recorded languages to Glottolog [8] or to any other official cataloging of languages, which meant that for mapping purposes, the languages presented in CLICS had to sometimes be linked manually via given name and coordinates to an official name of the language visible in Glottolog.

6. SUMMARY

The emergence of front and back vowels in the word forms of *front* and *back* concepts was examined across the world's languages. The vowel frequency counts gathered from the data showed differences in frequency distributions of the vowels between the two concepts. Most prominently, the vowels [ɑ], [ʊ] and [ɯ] showed significantly more in the *back* concepts. Also, the front vowels [i] and [e] were more frequent in the *front* concepts, although not significantly so. In addition the rounded vowels [o] and [u] showed prominence in the *front* concept data, although not significant either. The group comparison of front and back vowels showed that there are more back vowels in *back* concepts and significantly more front vowels in *front* concepts. These results would suggest that a sound symbolic effect of frontness and backness can be observed in the word inventories of the world, despite all the change that is imposed on languages over time.

Although the data indicate roundness having an effect on how frontal a vowel is perceived as, more research could be done on how roundness affects the emergence of spatial sound symbolic effects. Lip pointing as a conversational gesture is a widespread phenomenon among the world's populations [15] and therefore it would not be surprising for it to be embedded in how frontal a vowel is perceived as depending on its roundness, which could in turn affect the prevalence of rounded vowels in the *front* concepts researched here.

7. ACKNOWLEDGEMENTS

This research was conducted with funding from the Academy of Finland under grant agreement number 342627.

8. REFERENCES

[1] E. Sapir, "A study in phonetic symbolism," *J. Exp. Psychol.*, vol. 12, pp. 225–239, 1929.

[2] V. S. Ramachandran and E. M. Hubbard, "Synaesthesia–a window into perception, thought and language," *J. Conscious. Stud.*, vol. 8, pp. 3–34, 2001.

[3] L. Vainio, M. Kilpeläinen, A. Wikström, and M. Vainio, "Sound-space symbolism: Associating articulatory front and back positions of the tongue with the spatial concepts of forward/front and backward/back," *Journal of Memory and Language*, vol. 130, p. 104414, 2023.

[4] C. Tanz, "Sound symbolism in words relating to proximity and distance," *Lang. Speech*, vol. 14, pp. 266–276, 1971.

[5] M. Imai, S. Kita, M. Nagumo, and H. Okada, "Sound symbolism facilitates early verb learning," *Cognition*, vol. 109, pp. 54–65, 2008.

[6] D. E. Blasi, S. Wichmann, H. Hammarström, P. F. Stadler, and M. H. Christiansen, "Sound-meaning association biases evidenced across thousands of

languages," *PNAS*, vol. 113, no. 39, pp. 10 818–10 823, 2016.
[7] C. Rzymski, T. Tresoldi *et al.*, "The database of cross-linguistic colexifications, reproducible analysis of cross- linguistic polysemies," 2019. [Online]. Available: clics.clld.org
[8] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. (2022) Glottolog 4.6.
[9] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/
[10] P. Ladenfoged and I. Maddieson, "Vowels of the world's languages," *J. Phon.*, vol. 18, pp. 93–122, 1990.
[11] A. Wallace, "Review of Tylor's anthropology," *Nature*, vol. 24, pp. 242–245, 1881.
[12] J. Dellert and G. Jäger. Northeuralex (version 0.9). Tübingen: Eberhard-Karls University.
[13] C. Yang. (2011) Lalo regional varieties: Phylogeny, dialectometry and sociolinguistics.
[14] M. R. Key and B. Comrie, Eds., *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2015.
[15] N. Enfield, "Lip-pointing," *Gesture*, vol. 1, no. 2, pp. 185–212, 2001.

RECONSIDERING FAVORED PHONOTACTIC PATTERNS FOR EJECTIVE AND IMPLOSIVE CONSONANTS

Shelece Easterday

University of Hawai‘i at Mānoa
shelece@hawaii.edu

ABSTRACT

This study quantitatively examines an influential generalization from Greenberg (1970): that syllable-initial position is favored for glottalic consonants. In a diverse sample of 106 languages having ejectives and/or implosives, the phonotactic distribution of these consonants was characterized. While Greenberg's generalization holds, it does not capture some stark crosslinguistic differences in the distributional patterning of the two consonant types: for example, while only 4/25 (16%) languages with implosives and word-final stops feature implosives in this position, over half the languages with ejectives and word-final stops feature ejectives in this position (33/55, 60%). Further qualitative comparison of reported phonetic processes affecting ejectives and implosives suggests that a plausible interpretation of Greenberg's generalization – that glottalic consonants are frequently subject to neutralizing deglottalization in codas – should be reconsidered.

Keywords: ejectives, implosives, typology, sound change

1. INTRODUCTION

Approximately 28% of the world's spoken languages have glottalized consonants, which involve a "[tight] constriction of the vocal folds and/or a movement of the larynx a short distance up or down in the throat" [1]. Since Greenberg's influential 1970 study [2], two of these consonant types, ejectives and implosives (henceforth 'glottalic consonants'), have been the subject of synchronically- and diachronically-oriented crosslinguistic investigation [3, 4, 5, 6].

Most of the synchronic generalizations in [2] illustrate how ejective and implosive consonants differ from one another in their crosslinguistic patterning with respect to voicing, place of articulation preferences, and their relation to other series of consonants within phonological inventories. One of the few generalizations made by [2] concerning the two consonant types jointly is the following: "Syllabic initial position is favored for glottalic consonants in general. In almost every language studied they occur in syllable initial position, while in many they do not occur in syllable final even when the corresponding plain consonants do." While this generalization is often cited in works on glottalic consonants, it leaves out key details. The generalization was not quantitatively characterized in [2], and it is unclear from its formulation whether ejective and implosive consonants are purported to favor initial position to the same extent, or disfavor non-initial positions to the same extent.

Further, the above generalization has been interpreted to suggest that glottalic consonants are likely to be deglottalized through active processes of phonological neutralization in coda position [4]. Some crosslinguistic properties of the deglottalization and loss of implosives were explored in [2] and a study on glottalic consonants in Austronesian [3], but almost entirely with respect to place of articulation preferences in implosive inventory structure. These works do not explicitly examine ongoing phonetic processes or historical sound changes involving the deglottalization of implosives with respect to the stated preference for initial over final position or discuss them with respect to putative coda neutralization patterns. Processes of ejective deglottalization have been studied in a sample of moderate size in [4] with an eye towards investigating laryngeal neutralization in coda position. To date no study has considered and compared the properties of deglottalization processes affecting implosives and those affecting ejectives.

2. RESEARCH QUESTIONS

The current study seeks to address the following open questions regarding the synchronic and diachronic details of the distributional properties of glottalic consonants: (1) To what extent do ejectives and implosives exhibit a similar (dis)preference for initial and final positions within the syllable and the word? (2) How common is complete active phonological neutralization of ejective or implosive consonants in coda position? (3) To what extent are variable phonetic processes which deglottalize ejective or implosive consonants conditioned by position in the syllable or word?

These research questions are addressed within a survey of genealogically and geographically diverse languages which are reported to have ejective and/or implosive consonants.

3. METHOD

3.1. Language sample

The language sample consists of 106 languages reported to have contrastive ejective consonants, contrastive implosive consonants, or both (Table 1). The languages represent 84 families according to the classifications in [7], and diverse subgroupings wherever a family is represented by more than one language. Altogether, 72 languages in the sample are reported to have ejectives and 51 are reported to have implosives. Both consonant types are present in 17 languages. The sample was constructed to maximize geographic diversity; however, given the well-known skewed regional distributions of both consonant types [1, 5], the ejective and implosive portions of the sample represent largely non-overlapping regions of the world. Over half of the languages with ejectives belong to small language families in North and South America, while over half of the languages with implosives belong to families of large and moderate size in Africa and Southeast Asia and Oceania.

EJECTIVES (72 lgs)																
Language	**(i)**	**(ii)**	**(iii)**	**(iv)**	**Language**	**(i)**	**(ii)**	**(iii)**	**(iv)**	**Language**	**(i)**	**(ii)**	**(iii)**	**(iv)**
Abkhaz	✓	✓	✓	✓	Itonama	✓	✓			Shabo	✓	✓	-	-
Berta	✓	✓	✓		Klamath-Modoc	✓	✓	nt.	nt.	Shasta	✓	✓	✓	✓
Caddo	✓	✓	-	-	Komo (Sudan)	✓	✓	nt.		S. Haida	✓	✓	-	✓
C. Aymara	✓	✓	✓	✓	Kutenai	✓	✓	✓	✓	S. Nambikuára	✓	✓		-
C. Mazahua	✓	✓		-	Lake Miwok	✓	✓	-	✓	Squamish	✓	✓	✓	✓
Chimariko	✓	✓	✓	✓	Lakota	✓	✓	-	-	Taos N. Tiwa	✓	✓	-	
Chipaya	✓	✓	-	-	Lezgian	✓	✓	✓	✓	Tehuelche	✓	✓	✓	✓
Coahuilteco	✓	✓	✓	-	Me'en	✓	✓	-	-	Tigre	✓	✓	✓	✓
Cusco Quechua	✓	✓	-	-	Molale	✓	✓	-	-	Tlingit	✓	✓	✓	✓
Dawro	✓	✓			Mpade	✓	✓			Tojolabal	✓	✓	✓	-
Digor Ossetian	✓	✓	-	-	Nisga'a	✓	✓	✓	✓	Tol	✓	✓	✓	✓
Dizin	✓	✓	✓	-	Nivaclé	✓	✓	nt.	nt.	Trumai	✓	✓	✓	✓
E. Oromo	✓	✓	✓	✓	N.E. Maidu	✓	✓	-	-	Ts'ixa	✓	✓		
E. Pomo	✓	✓	✓	✓	N. Gumuz	✓	✓	✓		W. !Xoon	✓	✓		
Emberá-Catío	✓	✓			N. Yokuts	✓	✓	✓	✓	W. Itelmen	✓	✓	✓	✓
Ganza	✓	✓	✓	✓	N. Yukian	✓	✓	nt.	nt.	W. Keres	✓	✓		
Georgian	✓	✓	✓	✓	N.W. Sahaptin	✓	✓	✓	✓	W. Niger Fulfulde	✓	✓		
Hadza	✓	✓			Nuu-chah-nulth	✓	✓	✓	✓	Witsuwit'en	✓	✓	-	-
Hamer-Banna	✓	✓	✓		Patwin	✓	✓	-	-	Yana	✓	✓	✓	✓
Hanis	✓	✓	✓	-	Piaroa	✓	✓			Yapese	✓	✓	✓	-
Hausa	✓	✓	-	-	Puelche	✓	✓	-	-	Yuchi	✓	✓		
Huehuetla Tepehua	✓	✓	-	-	Qawasqar	✓	✓	-	-	Yurok	✓	✓	✓	✓
Ik	✓	✓			Quileute	✓	✓	✓	✓	Zulu	✓	✓		
Ineseño	✓	✓	✓	nt.	Sandawe	✓	✓			Zuni	✓	✓		-
IMPLOSIVES (51 lgs)																
Language	**(i)**	**(ii)**	**(iii)**	**(iv)**	**Language**	**(i)**	**(ii)**	**(iii)**	**(iv)**	**Language**	**(i)**	**(ii)**	**(iii)**	**(iv)**
Aja	✓	✓			Keiga	✓	✓	-	✓	N.E. Maidu	✓	✓	-	-
Berta	✓	✓	✓		Khmer	✓	✓	-		N. Burun	✓	✓	-	
Brao	✓	✓	-		Khmu	✓	✓	-	-	N. Gumuz	✓	✓	✓	
C. Mazahua	✓	✓		-	Kimaragang	✓	✓	-	-	Otoro	✓	-		
Dar Daju Daju	✓	✓		-	Komo (Sudan)	✓	-	-		Paumari	✓	✓		
Dawro	✓	✓			Kwaza	✓	✓			Pwo N. Karen	✓	✓		
Doyayo	✓	✓	-	-	Lakkia	✓	✓	-	-	Sabanê	✓	✓		
E. Oromo	✓	✓	-	-	Lele	✓	✓			Shabo	✓	✓	-	-
Ese Ejja	✓	✓			Ma'di	✓	✓			Sindhi	✓	✓		
Gabogbo	✓	✓			Mandari	✓	✓	-		S. Kisi	✓	✓		
Gbagyi	✓	✓			Mann	✓	✓			Tese	✓	✓	-	-
Hamer-Banna	✓	✓	✓		Me'en	✓	✓	-	-	Tojolabal	✓	✓	✓	-
Hausa	✓	✓	nt.	✓	Movima	✓	✓	-	-	Tsat	✓	✓	-	-
Huehuetla Tepehua	✓	✓	-	-	Mpade	✓	✓			Tsou	✓	✓		✓
Ik	✓	✓			Mubi	✓	✓	nt.	nt.	Tukang Besi N.	✓	✓		
Kambera	✓	✓			Ngambay	✓	✓	-	-	W. Niger Fulfulde	✓	✓		✓
Karajá	✓	✓			Ngbaka Manza	✓	✓			Zulu	✓	✓		

Table 1: Languages in the sample with glottalic consonants, coded for properties (i)-(iv) (see 3.2). ✓ = Glottalic Cs occur in this position; **-** = Glottalic Cs do not occur in this position; **nt.** = Glottalic Cs are deglottalized through complete active neutralization in this position; shaded cell = non-glottalic counterparts do not occur in this position.

3.2. Data collection and coding

The inventory of contrastive ejective and/or implosive consonants in each language was gathered from phonological descriptions. For each language, it was noted whether the phonotactics permit corresponding non-glottalic consonants (usually stops) to occur word-finally or in word-medial codas.

For each language, it was noted whether any of the relevant glottalic consonant types occur in the following positions: (i) word-initially; (ii) word-medially in intervocalic position; (iii) word-finally; and (iv) in a word-medial coda. It was also noted whether glottalic consonants are analyzed as being actively neutralized to non-glottalic counterparts in positions (iii) or (iv), as supported by evidence from morphophonemic alternations in the description.

The coding for the language sample can be found in Table 1. It is exemplified as follows: In Ma'di, which has an implosive inventory of /ɓ ɗ ƒ/ and no codas of any kind, implosives may occur (i) word-initially, /ɓà/ 'home', and (ii) word-medially in intervocalic position, /màɗí/ 'Madi' [8]. In Eastern Pomo, which has an ejective inventory of /p' t̪' t' ts' tʃ' k' q'/ and allows stops/affricates in coda positions, ejectives occur (iii) word-finally, /t͡sét͡s'/ 'dandelion', and (iv) in word-medial codas, /xa.ʃaːk'.le/ 'killed, they say' [9]. In Klamath-Modoc, which has an ejective inventory of /p' t' tʃ' k' q'/ and word-medial and word-final stop/affricate codas, all word-final stops are actively neutralized to aspirated voiceless stops: cf. /nt͡ʃ'ek'/ [nt͡ʃ'ekʰ] 'in little bits' and [nt͡ʃ'ek'aːni] 'small, little' [10, 11].

In addition to the above properties, processes reported in the reference materials which have the effect of deglottalizing ejective or implosive consonants were noted. These include variable phonetic processes conditioned by the phonological environment, speech rate, or social factors, and fully neutralizing phonological processes that affect some or all members of the ejective/implosive inventory.

4. RESULTS

4.1. Distributional properties of glottalic consonants

The distributional properties of glottalic consonants in the language sample are shown in Figure 1. The figure shows the percentage of relevant languages in which ejectives and implosives occur, do not occur, or are subject to complete active neutralization in positions (i)-(iv) (see 3.2 for definitions). Note that only languages with the relevant phonotactic possibilities are represented in Figure 1: for example, the percentages for (i) are calculated over all 72 languages with ejectives and all 51 languages with implosives; however, for (iii) the percentages are calculated over only the languages in which non-glottalic counterpart consonants, usually stops, can occur in word-final position (55 languages with ejectives and 25 languages with implosives).

In this sample, ejectives and implosives can occur word-initially in all languages which have them. This is nearly the case for word-medial intervocalic position, as well: the exceptions are Komo, in which implosives are limited to word-initial position [12], and Otoro, in which the sole implosive /ɗ/ is reported to be extremely rare and is illustrated only in word-initial contexts in [13].

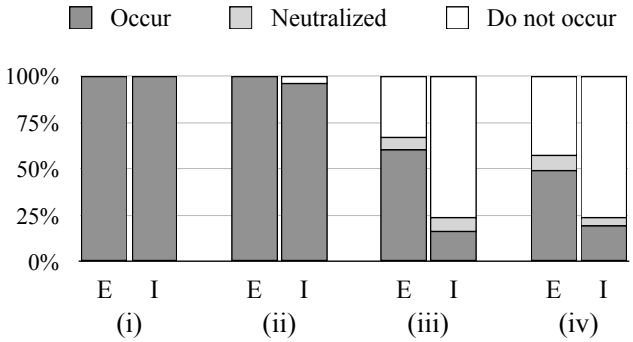


Figure 1: Percentage of relevant languages in which ejectives (E) and implosives (I) (do not) occur or are neutralized, by position in syllable or word (see 3.2).

Ejectives occur in word-final position in 33/55 (60%) of the languages for which that pattern is relevant. In four other languages – Klamath-Modoc [10], Komo [12], Nivaclé [14], and Northern Yukian [15] – ejectives are reported to be subject to complete active neutralization in word-final position. Implosives occur word-finally in only 4/25 (16%) of the relevant languages: Berta [16], Hamer-Banna [17], Northern Gumuz [18], and Tojolabal [19]. Word-final implosives are neutralized to plain voiceless stops in Hausa [20] and to unreleased voiceless stops in Mubi [21].

Ejectives occur in word-medial coda position in 26/53 (49%) of the languages with the relevant phonotactic patterns. Ejectives are fully neutralized in this position in four other languages: Ineseño [22], Klamath-Modoc [10], Nivaclé [14], and Northern Yukian [15]. Implosives occur in word-medial codas in 4/21 (19%) of the relevant languages: Hausa [20], Keiga [23], Tsou [24], and Western Niger Fulfulde [25]. Implosives are neutralized to unreleased voiceless stops in this position in Mubi [21].

To summarize, word- and syllable-initial positions are clearly crosslinguistically favored for glottalic consonants. However, among languages in which non-glottalic counterparts may occur word- and syllable-finally, ejectives are much more likely to occur in these positions than implosives.

4.2. Processes deglottalizing ejectives

Variable phonetic processes which deglottalize ejectives are reported in 14/72 languages with ejectives. None of these occur exclusively in word- or syllable-final environments.

The most common environment for the variable deglottalization of ejectives is in fact intervocalic, or more generally word-medial, position (five languages). For example, all ejectives can be realized as plain voiced stops intervocalically in Komo [12]. Other phonological factors conditioning variable deglottalization include an adjacent glottalized consonant (Central Aymara [26]), and strikingly, initial position: in Tol, a glottalized stop onset of a stressed syllable varies with a plain stop followed by a laryngealized vowel [27]. Variable deglottalization of ejectives is often not phonologically conditioned: it is conditioned by factors such as speech rate and speaker age in five languages, including Ik [28]. Unconditioned variable deglottalization of ejectives is reported for Ineseño [22] and Qawasqar [29].

Apart from the complete neutralization processes affecting entire ejective inventories in coda position discussed in 4.1, near-neutralization to slightly aspirated voiceless stops is reported for word-final ejectives in Nisga'a [30]. In three additional languages, a subset of the ejective series is neutralized in some environment, but just one such process is conditioned by coda position: in Ganza, word-final ejective stops, but not fricatives, are neutralized to their plain voiceless counterparts [31].

4.3. Processes deglottalizing implosives

Variable phonetic processes which deglottalize implosives are reported for 9/51 languages with implosives. These are conditioned by a syllable- or word-final environment in just one language: in Northern Gumuz, /ɗ/ tends to weaken to a flap post-vocalically, and /ɓ/ tends to weaken to an unreleased voiced bilabial stop word-finally [19]. The seven other languages which have coda implosives (Table 1) are not reported to have phonetic processes which deglottalize them in this position.

Variable deglottalization of implosives is conditioned by intervocalic position in four languages, including Sabanê [32]. Variable deglottalization occurs in unstressed syllables in Kwaza [33] and in postconsonantal position in Southern Kisi [34]. Speech rate and age (partially) condition variable deglottalization of implosives in four languages, including Hamer-Banna [17].

As discussed in 4.1, Hausa and Mubi have complete neutralization processes which deglottalize all implosives in coda positions. No additional languages here were reported to have deglottalization processes which fully neutralize a subset of the implosive inventory.

5. DISCUSSION

The results of the crosslinguistic survey fill in important details of Greenberg's generalization that syllable-initial position is preferred for glottalic consonants, but they also complicate its usual diachronic interpretation.

While ejective and implosive consonants pattern very similarly in that they occur in word-initial and word-medial intervocalic position in nearly all languages which have them, they differ starkly in their distribution in word-medial and word-final codas. In languages in which non-glottalic counterparts occur in these positions, ejectives are much more likely than implosives to be permitted in codas. Thus, while the generalization in [2] holds with respect to initial position, the dispreference for final position may be somewhat overstated for ejectives. Here it is worth noting that languages with ejectives and those with implosives tend to have different phonotactic profiles in general: this is apparent from the patterning of the shaded cells in Table 1. Further, a strong correlation has previously been established between syllable complexity and the presence of contrastive ejectives [35], but the precise motivation behind this correlation is not yet known.

Although the generalization in [2] has been interpreted to suggest that entire glottalic consonant series are susceptible to active phonological neutralization processes in final position, only seven languages in the sample show this pattern. Even disregarding languages without stops in the coda, this low rate may seem surprising given how prominently the phenomenon of laryngeal neutralization features in the phonological literature [36, 37, 38]. As noted by [4], complete neutralization is a powerful theoretical device entailing a long diachronic process and is therefore not necessarily straightforward to glean from synchronic descriptions. However, a survey of reported variable processes deglottalizing ejective and implosive consonants in the language sample turned up very few which target these consonants in final positions. Deglottalization processes most often apply intervocalically, a position in which the distribution of glottalic consonants is mostly unrestricted. Additionally, many reported variable deglottalization processes are not phonologically conditioned. These results, gathered from qualitative descriptions, should be verified by systematic phonetic studies, but they suggest that assumptions regarding the susceptibility of word- and syllable-final glottalic consonants to neutralizing deglottalization should be reconsidered.

7. REFERENCES

[1] Maddieson, I. 2013. Glottalized consonants. In: Dryer, M. S., Haspelmath, M. (eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/7, Accessed on 2023-01-02.)

[2] Greenberg, J. 1970. Some generalizations concerning glottalic consonants, especially implosives. *International Journal of American Linguistics* 36, 123–145.

[3] Blust, R. 1980. More on the origins of glottalic consonants. *Lingua* 52, 125-156.

[4] Fallon, P. D. 2002. *The synchronic and diachronic phonology of ejectives*. Routledge.

[5] Urban, M., Moran, S. 2021. Altitude and the distributional typology of language structure: Ejectives and beyond. *PLoS ONE* 16: e0245522.

[6] Easterday, S., Bybee, J. In press. Diachronic phonological typology: understanding inventory structure through sound change dynamics. *Linguistic Typology* 27. (Published online ahead of print 2023).

[7] Hammarström, H., Forkel, R., Haspelmath, M., Bank, S. 2022. *Glottolog 4.7*. Max Planck Institute for Evolutionary Anthropology. (Available online at http://glottolog.org, Accessed on 2023-01-02.)

[8] Blackings, M., Fabb, N. 2003. *A grammar of Ma'di* (Mouton Grammar Library, 32). Mouton de Gruyter.

[9] McLendon, S. 1975. *A grammar of Eastern Pomo.* (University of California Publications in Linguistics, 74.) University of California Press.

[10] Barker, M. A. R. 1963. *Klamath dictionary*. (University of California Publications in Linguistics, 31.) University of California Press.

[11] Blevins, J. 1993. Klamath laryngeal phonology. *International Journal of American Linguistics* 59, 237-279.

[12] Otero, M. A. 2019. A historical reconstruction of the Koman language family. (Ph.D. dissertation, University of Oregon.)

[13] Stevenson, R. C. 2009. A grammar of Otoro. In Schadeberg, T. C. (ed.), *Tira and Otoro: Two Kordofanian grammars*. Rüdiger Köppe Verlag, 115-326.

[14] Gutiérrez, A. 2019. Nivaĉle (shiichaam lhavos variety). *Journal of the International Phonetic Association* 49, 401-417.

[15] Balodis, U. I. J. 2011. Yuki grammar in its area context with sketches of Huchnom and Coast Yuki. (Ph.D. dissertation, University of California at Santa Barbara.)

[16] Andersen, T. 1993. Aspects of Berta phonology. *Afrika und Übersee* 76, 41-80.

[17] Petrollino, S. 2016. A grammar of Hamar, a South Omotic language of Ethiopia. (Ph.D. dissertation, Rijksuniversiteit te Leiden.)

[18] Ahland, C. A. 2012. A grammar of Northern and Southern Gumuz. (Ph.D. dissertation, University of Oregon.)

[19] Supple, J., Douglass, C. M. 1949. Tojolabal (Mayan): Phonemes and verb morphology. *International Journal of American Linguistics* 15, 168-174.

[20] Newman, P. 2000. *The Hausa language: An encyclopedic reference grammar*. Yale University Press.

[21] Prickett, D. 2012. Phonology and morphology of verb forms in Mubi. (MA thesis, University of North Dakota.)

[22] Applegate, R. B. 1972. Ineseño Chumash grammar. (Ph.D. dissertation, University of California at Berkeley.)

[23] Reh, M. 1994. A grammatical sketch of Deiga. *Afrika und Übersee* 77, 197-261.

[24] Wright, R., Ladefoged, P. 1994. A phonetic study of Tsou. *UCLA Working Papers in Phonetics* 87, 67-89.

[25] Alkali, A. A. 2019. A description of the phonology and morphology of Sokoto Fula. (Ph.D. dissertation, University of Hong Kong.)

[26] Coler, M. 2014. *A Grammar of Muylaq' Aymara: Aymara as spoken in Southern Peru*. Brill.

[27] Holt, D. 1999. *Tol (Jicaque)*. (Languages of the World/Materials, 170.) Lincom Europa.

[28] Schrock, T. 2014. A grammar of Ik (Icé-tód): Northeast Uganda's last thriving Kuliak language. LOT. (Ph.D. dissertation, Rijksuniversiteit te Leiden.)

[29] Clairis, C. 1985. *El qawasqar: Lingüística fueguina, teoría y descripción.* (Estudios Filológicos: Anejo, 12.) Universidad Austral de Chile.

[30] Tarpent, M.-L. 1989. A grammar of the Nisgha language. (Ph.D. dissertation, University of Victoria.)

[31] Smolders, J. 2016. A phonology of Ganza (Gwàmì Nánà). *Linguistic Discovery* 14, 86-144.

[32] Antunes de Araujo, G. 2004. A grammar of Sabanê: A Nambikwaran language. 94. LOT. (Ph.D. dissertation, Vrije Universiteit Amsterdam.)

[33] van der Voort, Hein. 2004. *A grammar of Kwaza.* (Mouton Grammar Library, 29.) Mouton de Gruyter.

[34] Childs, G. T. 1995. *A grammar of Kisi: A Southern Atlantic language.* (Mouton Grammar Library, 16.) Mouton de Gruyter.

[35] Easterday, S. 2019. *Highly complex syllable structure: A typological and diachronic study* (Studies in Laboratory Phonology 9). Language Science Press.

[36] Lombardi, L. 1991. Laryngeal features and laryngeal neutralization. (Ph.D. dissertation, University of Massachusetts, Amherst.)

[37] Steriade, D. 1997. Phonetics in phonology: the case of laryngeal neutralization. Unpublished ms, UCLA.

[38] Blevins, J. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Effect of vowel context on stop place identification in Yoruba

Roslyn Burns and Jason Shaw

Yale University
roslyn.burns@yale.edu jason.shaw@yale.edu

ABSTRACT

Vowel context is a known factor in palatalization, in part because of how co-articulation impacts consonant place cues. While much is known about palatalization triggers and simplex place targets, complex places remain understudied. Here, we investigate the perception of Nupe labial-velars by Yoruba listeners across vowel contexts in an identification task. Results show that labial-velar confusability follows the hierarchy of palatalization triggers established for simplex segments. Labial-velars exhibit two different outcomes in palatalizing environments, one in which the dorsal component is interpreted as a coronal, but also one in which the dorsal component is ignored leading to confusion with labials.

Keywords: labial-velars, palatalization, sound change, Yoruba, Nupe, sound change

1. INTRODUCTION

Vowel context has been shown to influence the perception of velar stops /k/ and /g/ in the environment of /i/, where they become more confusable with the anterior segments /tʃ/ and /dʒ/ respectively. The increased confusability is attributed to CV co-articulation which brings the tongue dorsum forward, thus reducing the perceptual distinctiveness between velars and anterior affricates [1]. Our study is primarily concerned with whether a similar perceptual effect of vowel place can be observed with double closure labial-velar stops, which are understudied with respect to perception.

There are some reasons to suspect that vowel context may impact labial-velar stop perception. First, labial-velar stops are under-attested preceding high front vowels in some West African languages, which is observable in both Igbo and Yoruba dictionaries [2, 3, 4]. Second, co-articulation with a high front vowel could reduce articulatory movements that distinguish labial-velars from other stop places. There are two considerations. First, the labial component of labial-velars can differ from simplex labials in how the jaw contributes to the constriction and release; jaw lowering just before labial release can give labial-velars an implosive quality [5], which may differentiate them from simplex labials. Second, the velar component of labial-velars can differ from simplex velars, with the tongue dorsum being more retracted in labial-velars than simplex velar counterparts [5, 6, 7]. These two articulatory enhancements (jaw lowering and tongue dorsum retraction) may be suppressed by co-articulation with a high front vowel, which requires jaw raising and tongue fronting, possibly leading to increased perceptual confusion with other places.

To evaluate the effect of vowel context on labial-velar stop perception, we ran a cross-language perception study. Native speakers of Yoruba (Niger-Congo: Yoruboid) categorized stops, including labial-velars, produced by a Nupe (Niger-Congo: Nupoid) speaker in different vowel contexts. Both languages have labial-velar stops, but we decided to use Nupe items as stimuli because labial-velars are particularly robust in this language. Nupe exhibits a full six-way contrast in voicing across labial-velar, velar, and labial places of articulation: /k͡p/, /k/, /p/, /g͡b/, /g/, /b/ [8]. Yoruba only contrasts /k͡p/, /k/, /g͡b/, /g/, and /b/ [9]. We used Yoruba as the listener language because of the accessibility of online participants and because Yoruba is the most widely used Niger-Congo language in Nigeria (and globally) with an estimated 42,000,000 native users.

In addition to the /i/ context, we also included /e/, /u/, /a/ contexts in our experiment. Our predictions for the effect of these vowel contexts on stop place identification draw from the literature on sound change. Cross-linguistic surveys show that simplex velars are likely to become either alveolar or palatal in the environment of /i/, i.e., “palatalization” [10, 11, 12]. The vowels /e/ and /u/ can also condition the same sound change but to a lesser extend than /i/. The least likely trigger is /a/. Notably, if vowels other than /i/ trigger palatalization, they are subject to the implicational hierarchy in (1) [10, 11].

(1) a. Low front vowels do not trigger palatalization unless high front vowels do.
b. High back vowels do not trigger palatalization unless high front vowels do.

By including all four vowels, /i/, /e/, /u/, /a/, we test whether perceptual confusions vary across vowel contexts in the way that reflects the likelihood of sound change from velar to a different (more anterior) lingual place of articulation.

3. METHODS

3.1 Nupe stimuli design

A native speaker-linguist of Nupe, Ahmadu Ndanusa Kawu, produced 96 disyllabic nonce word targets of the shape C1V1C2V2 where C1 varied in voicing (voiced, voiceless) and place (labial-velar, dorsal, labial, coronal), V1 varied in tone (H, L) and quality ([i], [u], [e], [a]), C2 was always coronal, and V2 was always [a] and matched the tone of V1. Labial-velar-initial words had three variants of C2 ([t], [d], [s]) whereas all other places had only one ([t]).

The speaker produced the CVCV sequences in two Nupe carrier phrases *wun ka* ___ *be* 'he wrote ___ again' and *wun ganan* ___ *be* 'he said ___ again' for a total of 192 utterances. All utterances were recorded on a Zoom Hn4 recorder at a sample rate of 44.1 khz. These utterances were forced aligned with the Montreal Forced Aligner using a Nupe dictionary created by the researchers and an English pronunciation model. Boundaries were checked by hand in Praat [13]. The C1 boundary was moved to the nearest zero-crossing preceding the stop burst for obstruents and to the onset of F2 rise for [w]. The end of the V1 boundary was shifted to the nearest zero crossing at the offset of vowel voicing.

Of the 192 recorded utterances, 96 target items were selected for inclusion in the experiment. Audio stimuli were selected by identifying the most peripheral vowel production for each target. Once we selected a representative token for each item, we excised the initial CV portion of each target word, based on the boundaries described above, to serve as stimuli for the experiment. The stimuli were then normalized to 70dB in Praat. Formant values of the stimuli were extracted at seven evenly spaced intervals using an Inverse Filter Control script [14, 15]. Burst frequencies were extracted in Praat with a 300 hz high pass filter.

3.2 Yoruba participants

49 native speakers of Yoruba who grew up in Nigeria until age 18 were recruited on Prolific.co. Everyone reported speaking, listening, reading, and writing proficiency in Yoruba and proficiency in English. Participants ranged in age from 18 to 59.

3.3 Procedure

As labial-velars across West Africa are phonetically heterogeneous [5] and this is the first study of its kind, we opted for an open response task instead of, e.g, forced choice. Participants were directed to an online experiment written in Yoruba (translation by Oluwaseyi Fasunhan) on the Gorilla platform. Participants were told that they would hear the first part of a word and were instructed to type what they heard as a response. Prior to the 96 experimental trials, there were 7 practice trials of stimuli not used in the actual experiment.

3.3 Response coding

Since we were primarily interested in whether vowel context impacted the perception of labial-velar consonant place of articulation, we coded responses according to this dimension. Place was coded into one of four major place categories, labial, coronal, which collapsed over alveolar and palatal, dorsal, and labial-velar, based on the orthographic representation provided by the participant. In line with previous studies involving non-phonemic palatal contrast [1], we did not distinguish between palatal and alveolar responses since both involve fronting. A fifth category, neither, was coded for non-buccal segments and vowels. As with other perceptual studies, these places represent phonological categories as opposed to exact phonetic places of articulation (i.e. unless the orthography encodes allophones, then allophones are not coded).

Due to variation in how responses were entered (e.g. additional words, providing entire words, etc.) all tokens were coded for place by hand (n= 4705). 83 trials were discarded for failing to respond, containing uninterpretable place properties, and technical errors. Because voiced phonemes in Yoruba exhibit the full set of contrasts across the four investigated places we restricted our analysis to voiced stimuli (n= 2312).

In many Nigerian languages, <kp> represents [k͡p], <p> represents [p], and <gb> represents [g͡b]. Nigerian Yoruba uses <p> for both [k͡p] and [p], which are not contrastive, and lacks <kp>. Yoruba is, however, spoken in countries which impose national orthographies that distinguish between <kp> [k͡p] and <p> [p] [16]. In some cases, participants provided voiceless responses to voiced stimuli. In these cases, <p> was coded as both labial-velar and labial (n = 264) as orthographically it represents both, and <kp> was coded as labial-velar (n = 43).

After coding for place, responses were coded as

either "correct" or "incorrect" based on whether the response place matched the original stimuli's coded place. Because the Nupe speaker produced /gu/ as [wu], both <g> and <w> were accepted as correct responses. The reported findings in section 3.2 are made on the basis of incorrect responses.

3. RESULTS

3.1 Nupe stimulus items

Table 1. summarizes the stimuli properties grouped by consonant type (columns) and vowel type (rows). Table 1a on the top shows the highest amplitude frequency of the stop burst, Table 1b in the middle shows the F2 measurement at the first of seven evenly spaced time intervals in the vowel (T1=0%), and Table 1c on the bottom shows the change in F2 between the first and second time intervals (Δ 17%-0%).

	coronal		dorsal		labial		labial-velar	
	mean	sd	mean	sd	mean	sd	mean	sd
a. Maximum amplitude frequency of stop burst								
/a/	1357	355	1389	55	1336	50	1185	239
/e/	3274	3014	2584	870	1637	187	1404	207
/i/	3058	1708	4533	1734	1379	141	1314	165
/u/	3015	3618	1163	87	1120	122	1644	1576
b. F2 at the onset of voicing								
/a/	1671	75	1456	60	1353	49	1283	60
/e/	1656	390	1741	88	1727	65	1602	83
/i/	1594	276	1641	130	1652	322	1551	194
/u/	1479	151	1143	278	1221	171	1143	114
c. F2 change at the onset of the vowel (first 17%)								
/a/	-50	12	28	54	5	50	70	52
/e/	-141	225	-68	193	-39	176	-29	74
/i/	-52	74	-24	88	1	139	-63	118
/u/	-110	66	-81	150	-66	68	-86	68

Table 1: Stimuli properties in hz

Coronal stimuli often have the highest burst frequency followed by dorsals, labials, and then labial-velars. Before /e/, dorsals have higher burst frequencies than coronals. Before /u/, all places except for coronals exhibit low burst frequencies.

On average, labial-velars exhibit the lowest F2 values at T1 in all vowel contexts, except for /u/, while coronals exhibit the highest values.

In the F2 change over the vowel onset, negative numbers indicate a decrease in frequency; positive numbers indicate an increase; and 0 indicates no change. Before /u/, coronals have the greatest F2 decrease followed by labial-velars. Before /i/, F2 increases the most in segments articulated with the lips. Before /e/, coronals exhibit the greatest decrease in F2.

Although burst frequencies are cues that can potentially help listeners distinguish labial-velars from other places (Table 1a), if listeners rely on F2 transition cues (Table 1b-c), labial-velars may be more confusable with other places of articulation in palatalizing contexts.

3.2 Yoruba perception

Participants mostly responded with the correct answer (mean= 0.65) but there is a wide amount of variability (sd= 0.47). Fig. 1. shows the error rate across the 49 participants in different consonant and vowel contexts. The standard error estimates the variability across individual speakers.

As shown in Fig. 1., consonant places exhibit different error patterns across vowels. To evaluate the significance of the error patterns, we fit nested linear mixed effects models (LMER) to the data. The baseline model included consonant place and vowel (fixed) and participant (random intercept). To this baseline, we added a place*vowel interaction. An ANOVA comparison showed that the model with the interaction term was significantly better than the baseline (AIC 2785.9 vs. baseline AIC 2911.7, $\chi 2=144$, $p < 0.05$), indicating that the differential effect of vowel on consonant confusions is reliable.

Labial-velar errors appear to follow the vowel place dependencies outlined in (1). That is, among front vowels error rates were highest for /i/ followed by /e/ then /a/. Among high vowels, error rates were higher for /i/ than for /u/. We evaluated this trend with a LMER model of the labial-velar data predicting accuracy by vowel (fixed) and participant (random intercept). Table 2 shows the fixed effects.

	Estimate	Std. Error	pr(>|t|)
/g͡bi/ (intercept)	0.66	0.04	< 2e-16
/g͡ba/	0.16	0.03	3.33e-07
/g͡be/	0.06	0.03	0.03926
/g͡bu/	0.12	0.03	0.00023

Table 2: Voiced labial-velar response fixed effects.

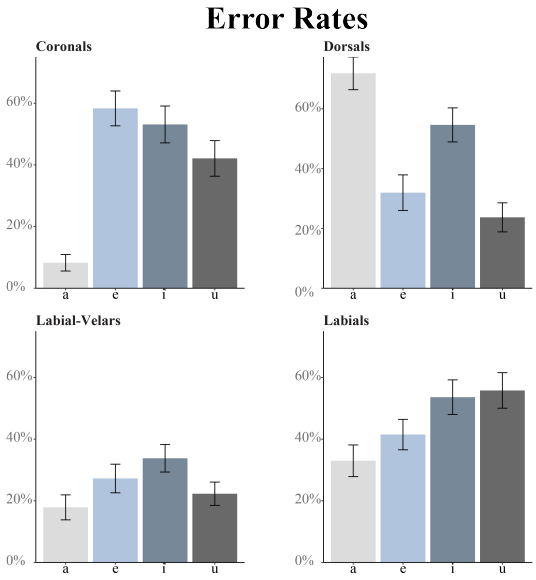


Figure 1: Error rate averaged across subjects

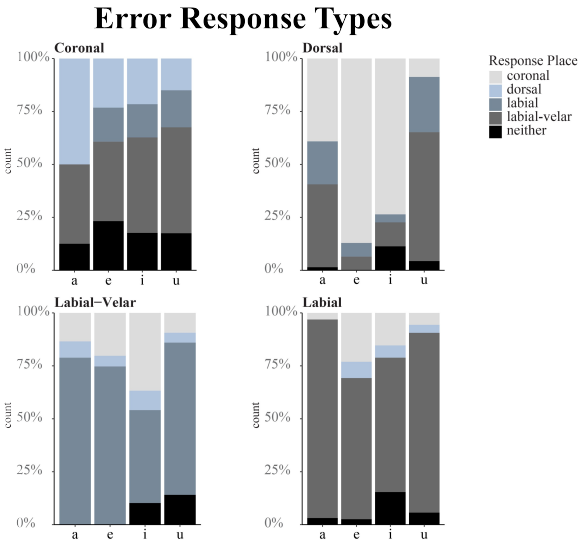


Figure 2: Error response types

Each vowel significantly differs from /i/ (intercept) and accuracy increases from /i/ to /e/, /u/, and then /a/. To directly compare /e/ and /a/, we made /a/ the intercept [17]. In this model, the /e/ coefficient was -0.1 ($p < 0.05$). Thus, the trends corresponding to (1a) and (1b) among labial-velars are statistically supported.

Fig. 2. shows the distribution of erroneous responses by consonant and vowel. As shown in Fig. 2., /g͡b/ is mostly mistaken as labial and /b/ is mostly mistaken as labial-velar. This is likely because, as shown in Table 1, acoustic cues for labial-velars and labials are highly similar to each other. In palatalizing contexts, /g͡b/ is increasingly mistaken as coronal following the hierarchy in (1). Notably, /b/ lacks increased confusability with coronals in this same context implicating the dorsal constriction as the likely source of the increased coronal confusability for labial-velars.

Comparison between /g͡b/ and /g/ suggests that the dorsal closures have similar confusability profiles save for the high error rate of /ga/ in an apparent violation of (1a). Fig. 2. reveals that /ga/ compounds errors involving labial closures (simplex and complex) and those involving coronal closures. Focusing on just the misperceptions of /g/ as coronal, /ga/ has lower error rates than both /gi/ and /ge/ which conforms to (1a); labial-velars show the same pattern. One exception is that /ge/ has a higher error rate than /gi/. This possibly because velars in the context of /e/ have higher F2 values than velars in the context of /i/, which may be an artifact of our stimuli. The above exceptions notwithstanding, the confusions for both /g͡b/ and /g/ are relatable to the palatalization hierarchy, but the confusions for /b/ are not.

In line with cross-linguistics studies of sound change [10, 11, 12], velars in our speech perception task were more likely to be misidentified as coronals in the contexts identified in (1) but notably labial-velars were also more likely to be identified as coronal in the same contexts. Unlike velars, however, labial-velars are also highly confusable with labials. Despite the increased confusability with coronals in palatalizing contexts, labial-velars are still frequently (mis)identified as labials, but notably labials do not follow the same vowel-induced pattern of confusability with coronals. These findings have implications for theorizing about sound changes. When the cues to labial-velar stops are compromised in palatalizing contexts, labial-velars may be perceived as velars. If language users prioritize the maintenance of cues which distinguish labial-velars from velars, labial-velars may undergo simplification by losing the dorsal component in palatalizing contexts (i.e. labial-velar > labial) thus leading to sound change.

4. CONCLUSION

Our study found an effect of vowel context on the perception of labial-velar stops. The likelihood of being misperceived as coronal follows the palatalization hierarchy outlined in previous work on simplex velars. This is likely due to the fact that in palatalizing contexts, the cues which make labial-velars unique deteriorate due to co-articulation with the neighboring vowel. This suggests the potential for diachronic instability of labial-velars in palatalizing contexts.

5. REFERENCES

[1] Guion, S. 1998. The role of perception in the sound change of velar palatalization. *Phonetica* 55(1–2): 18–52.

[2] Echeruo, M. J. C. 1998. *Igbo-English dictionary : a comprehensive dictionary of the Igbo language with an English-Igbo index*. New Haven: Yale University Press.

[3] Igwe, G. E. 1999. *Igbo-English Dictionary*. Ibadan: University Press.

[4] 2001. *A dictionary of the Yoruba language*. Ibadan: University Press.

[5] Ladefoged, P. 1968. *A Phonetic Study of West African Languages: an auditory-instrumental survey*. Cambridge: Cambridge University Press.

[6] Connell, B. 1994. The structure of labial-velar stops. *Journal of Phonetics* 22(4): 441–476.

[7] Hudu, F., Miller, A., & Pulleyblank, D. 2009. Ultrasound imaging and theories of tongue root phenomena in African languages. In Austin, P. K., Bond, O., Charette, M., Nathan, D., & Sells, P. (eds.), *Proceedings of Conference on Language Documentation & Linguistic Theory* 2, 153–163. London: SOAS.

[8] Pulleyblank, D. 1984. Yoruba. In Comrie, B. (ed.), *The World's Major Languages 2nd edition*, 866–882. London: Routledge.

[9] Moran, S. & McCloy, D (eds.) 2019. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. http://phoible.org

[10] Bateman, N. 2007. *A Crosslinguistic Investigation of Palatalization.* Dissertation: University of California San Diego.

[11] Bateman, N. 2011. On the typology of palatalization. *Language and linguistics compass* 5(8): 588–602.

[12] Bhat, D. NS. 1978. A general study of palatalization. In Greenberg, J. (ed.), *Universals of Human Language* vol. 2, 47–92. Stanford: Stanford University Press.

[13] Boersma, P. & Weenink, D. 2019. *Praat: doing phonetics by computer* v. 6.0.46. praat.org

[14] Sprouse, R. & Johnson, K. 2016. The Berkeley phonetics machine. *UC Berkeley PhonLab Annual Report* 12(1).

[15] Ueda, Y., Hamakawa, T., Sakata, T., Hario, S., & Watanabe, A. 2007. A real-time formant tracker based on the inverse filter control method. *Acoustical science and technology* 28(4): 271–274.

[16] Tchitchi, T. & Hazoumé, M. L. 1983. Le dévelopment d'un alphabet nacional dans une communauté multilingue. In Winter, W. (ed.), *Writing in Focus*, 187–208. Berlin: Mouton.

[17] Baayen, R.H. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

BORA /ɨ/ IS PRODUCED WITH A LOW TONGUE DORSUM AND A DENTAL CONSTRICTION: AN ULTRASOUND STUDY

Jeff Mielke[1], Madeleine Oakley[1], Alonso Vásquez-Aguilar[2], Roberto Zariquiey[3], and Steve Parker[4]

[1]North Carolina State University, [2]University of California, Santa Barbara,
[3]Pontifical Catholic University of Peru, [4]Dallas International University
jimielke@ncsu.edu, mjoakley@ncsu.edu, alonsovasquez@ucsb.edu, rzariquiey@pucp.edu.pe, steve_parker@diu.edu

ABSTRACT

We present an ultrasound study of the vowels of Bora, an Amazonian language spoken by about 850 people in Peru and Colombia. Bora's phonemic vowels have been traditionally described as /i ɨ ɯ ɛ o a/, where all three high vowels are unrounded, and /ɨ/ and /ɯ/ are very close in F1-F2 space. It has recently been documented that /ɨ/ is produced with lingual-dental contact. Other details of the tongue posture used to produce it have been unknown until now. On the basis of ultrasound recordings of Bora and Spanish words, we find that Bora /ɯ/ is indeed a high back vowel, but /ɨ/ is produced with a low tongue body. Its principal constriction is between the tongue blade and teeth.

Keywords: Bora, dental vowel, ultrasound, vowels, articulatory

1. INTRODUCTION

Bora is a language spoken by about 750 people in Peru and 100 in Colombia. Its phonemic vowels have been described as shown in Table 1, where both /ɨ/ and /ɯ/ are unrounded [1].

Table 1: The vowel inventory of Bora as traditionally described [1]

	front	central	back
high	i	ɨ	ɯ
mid	ɛ		o
low		a	

A contrast between a central and a back vowel which are otherwise identical is theoretically significant since it implies that the binary feature [±back] is too weak to encode all phonological contrasts along the front/back dimension. Previously, Bora vowels have been described in acoustic terms with measurements of F1-F3 based on audio recordings, indicating that /ɨ/ is farther front than /ɯ/ [2]. [3] show that only /o/ is produced with lip rounding and that /ɨ/ is produced with visible contact between the tongue blade and the upper and lower incisors, as shown in Figure 1.

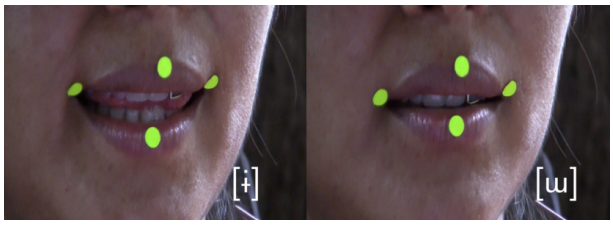


Figure 1: Representative lip/jaw/tongue positions for /ɨ/ and /ɯ/ for one female speaker [3].

This evidence suggests that /i ɨ ɯ/ are not distinguished by lip rounding, but that /ɨ/ is distinguished from the other two by its regular dental contact and apparent lateral airflow. Unlike the apical vowels of Mandarin [4], Bora /ɨ/ is an inherently dental vowel, regardless of context. It is also produced without detectable frication [3]. Its dorsal posture has been undocumented.

2. METHODS

We made recordings of Bora and Spanish words using audio recording and ultrasound video of the tongue. The Spanish /i/ and /u/ provide reference points of prototypical front and back high vowels to compare with the high vowels of Bora.

2.1. Participants

Participants were six native speakers of Bora, aged 25-55 (two female, four male). All speakers were fluent in Spanish as well. Recordings were made in a quiet room at a medical clinic in Iquitos, Peru.

2.2. Materials

Stimuli consisted of 63 Bora words from the dictionary of [5] occurring in the carrier phrase '[tíɲɛ] _____' 'Say _____' as well as 30

Spanish words in the context '[dime] _____' 'Say to me _____'. Phrases were presented up to six at a time, in three predetermined random orders. Each word was presented five times. The Bora recordings were made in a single block first, followed by the Spanish recordings.

2.3. Data collection

Participants wore a Shure Beta 53 head-mounted condenser microphone and an Articulate Instruments UltraFit headset. Audio and lip video were recorded using Articulate Assistant Advanced software (AAA) [6], which was also used to present stimuli orthographically. Since lip positions are relatively well documented for Bora vowels [3], we do not analyze the lip video here.

Lingual ultrasound images were recorded using a GE LOGIQ S8 ultrasound machine with an IC5-9 microconvex transducer via screen capture to a laptop at 60 frames per second. The actual rate of new scans being output by the ultrasound machine was 23 frames per second. The transducer was held in place submentally by the UltraFit headset.

2.4. Data analysis

Audio recordings were force-aligned using the Montreal Forced aligner (MFA) [7]. Spanish was aligned with the Spanish MFA acoustic models [8]. Since no Bora acoustic models exist, we used Kyrgyz models [9], which performed the best among several models we tried based on languages with phoneme inventories similar to Bora's, substituting Kyrgyz phones for Bora before alignment and then back to Bora after alignment. Phone boundaries were corrected manually in Praat [10] after forced alignment.

The screen capture audio was used to synchronize the ultrasound videos with the audio/video recordings using cross-correlation, but only the audio recorded in AAA was used for analysis. Formant frequencies were measured at vowel midpoints using Praat [10]. We note that all of the target vowels are long and most of them have a rising tone, but this does not appear to be accompanied by a change in vowel formant frequencies.

Tongue contours were traced in ultrasound videos in DeepLabCut [11]. We traced the tongue contour in the portions of the videos containing speech using models trained on a diverse set of tongue videos [12]. We then adapted the models to our dataset by relabeling 20 outlier frames per speaker and tracing our images again using the adapted models.

Obvious errors in tongue traces and formant measurements were removed. Tongue traces extracted from vowel midpoints were rotated to make the occlusal plane horizontal and compared using polar Smoothing Spline ANOVA [13, 14, 15] using the `gss` package in R [16, 17].

3. RESULTS

Figure 2 shows the mean acoustic vowel spaces for all six speakers, for all vowels in word-initial position before a labial. This includes the Bora words /iípa/ 'small ash-colored deer', /ɪ́pa/ 'shad fish species' *(sábalo)*, /ɯ́ɯ́pa/ 'species of worm' /ɛ́ɛ́pa/ 'that one (drum, guava, etc.)', /áápa/ 'to be weak or soft (voice/fire)', and /óópa͡kpa/ 'species of monkey', and the Spanish words *iba*, *Eva*, *haba*, *ova*, and *uva*. Bora and Spanish /a/ and /i/ are quite similar in formant frequency, and the Spanish mid vowels are slightly higher than the Bora mid vowels. The Bora vowels /ɨ/ and /ɯ/ are about 1/3 and 2/3 of the way across the top of the F2 vowel space between the /i/s and Spanish /u/, respectively.

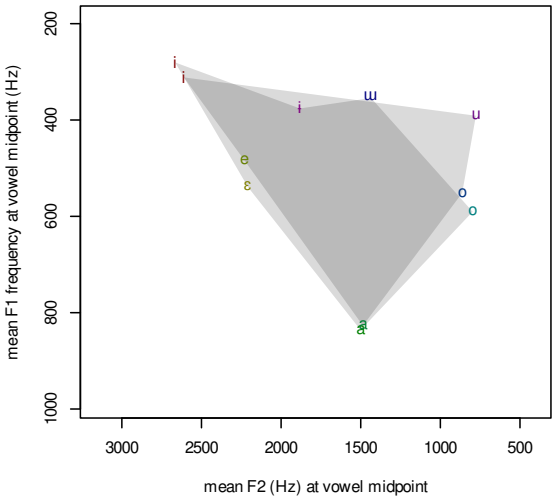
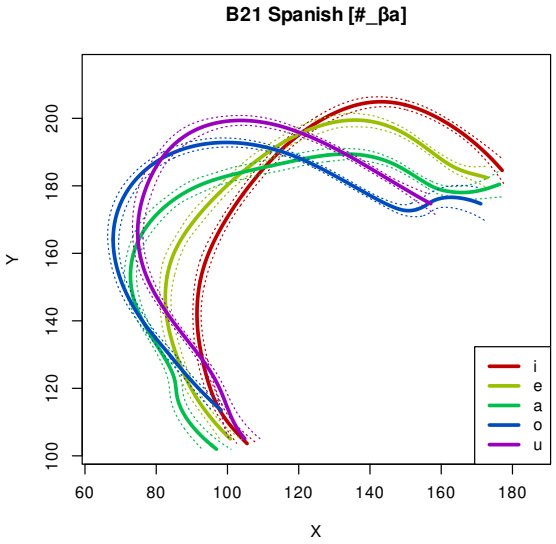


Figure 2: Mean Spanish and Bora acoustic vowel spaces for six speakers. The polygon with /i ɨ ɯ ɛ o a/ is for Bora and the polygon with /i u e o a/ is for Spanish.

Bora /ɨ ɯ/ have been shown not to be distinguished by lip rounding, but rather differ in jaw height and lingual-dental contact [3]. However, the lingual differences between Bora /ɨ ɯ/ and Spanish /u/ (all of which have been described as high non-front vowels) is a major research question which this paper seeks to resolve.

Figure 3 shows sample ultrasound SSANOVA comparisons for one representative male speaker's vowels. Figure 4 shows the same data grouped



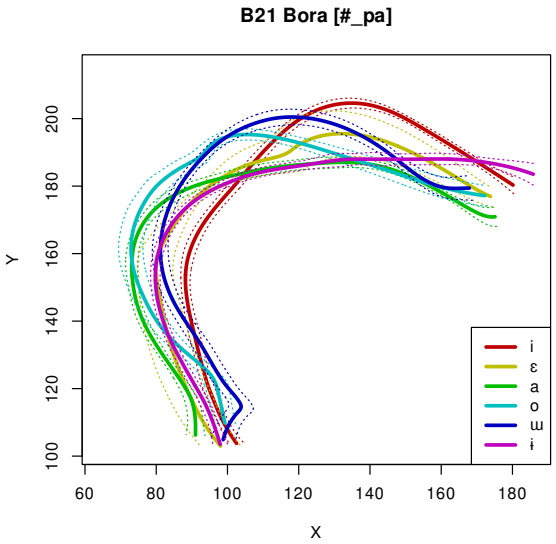
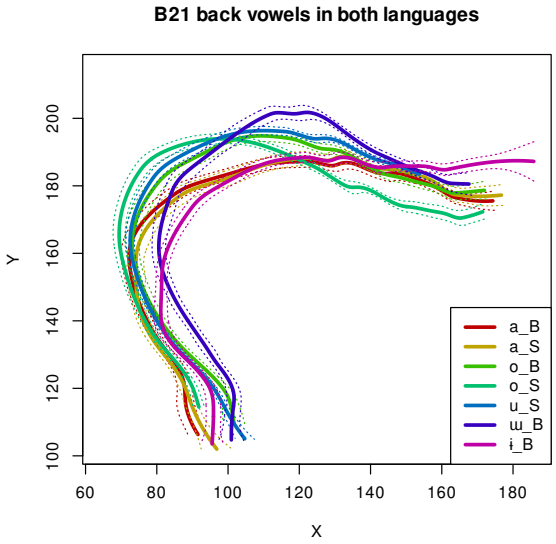


Figure 3: Spanish and Bora SSANOVA comparisons for one male speaker. The tongue tip points to the right.



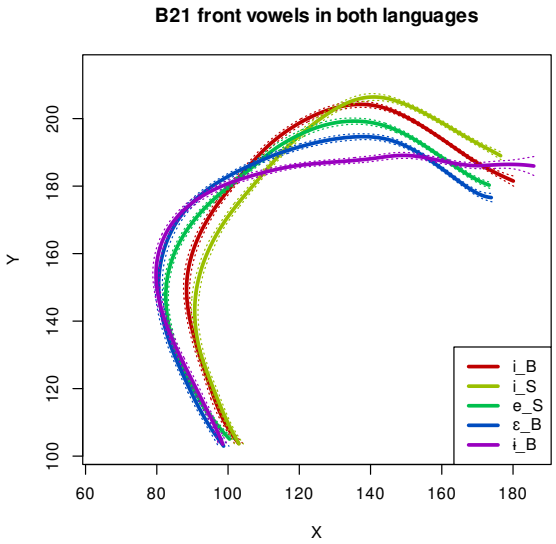


Figure 4: Front and back vowels from both languages for one male speaker. /ɨ/ is included in both sets.

into front and back vowel sets combining both languages, where /ɨ/ is included in both sets. The top panel in Figure 3 shows the expected difference in anteriority between Spanish /i/ and /u/. /e/ has a lower tongue dorsum and more retracted tongue root than /i/. /o/ is produced lower and further back in the vocal tract than /u/, consistent with an upper pharyngeal constriction. /a/ has the lowest tongue dorsum and appears to have its greatest constriction in the lower pharynx, consistent with a low vowel. The bottom panel of Figure 3 shows a similar configuration of five of the Bora vowels /i ɯ o ɛ a/. Surprisingly, the tongue dorsum for the sixth vowel /ɨ/ is as low as for /a/. The blade of the tongue is higher for /ɨ/ than for any of the other vowels, consistent with this vowel having a dental constriction that the others lack.

Figure 4 directly compares the front and back vowels across languages. /ɨ/ is included in both sets since its status is arguable, and the main focus of this study. The top panel shows that the Spanish /u/ in *uva* is articulated more posteriorly than the Bora /ɯ/ in /ɯúpa/. That is, although Bora /ɯ/ exhibits a higher F2 than Spanish /u/, this is not just due to /ɯ/ being unrounded; it is also a more central vowel for this speaker (and three of the other five). The /o/s in both languages are very similar, with interspeaker variation in which is more retracted (posterior). Bora and Spanish /a/ appear identical. For four of the six speakers, Bora /ɯ/ is central in relation to the Spanish back vowel /u/ (as seen for speaker B21 in Figure 4). For the other three (B7,

and B8), the tongue positions for the two vowels are similar, and the observed acoustic difference is likely due primarily to lip rounding in /u/, which lowers F2. For five of the six speakers, Bora /o/ is higher than Bora /a/ (as seen for B21). For one speaker (B7), the tongue positions for these vowels are very similar, suggesting that they are distinguished primarily by lip rounding.

The bottom panel of Figure 4 shows that Spanish /i/ in *iba* is the highest of the front vowels, followed by Bora /i/ (/iípa/), Spanish /e/ (*Eva*), and then Bora /ɛ/ (/ɛɛ́pa/). Bora /ɨ/ is lower than all of the back vowels and virtually all of the front vowels in both languages. The tongue root position for /ɨ/ is similar to the high back/central /u ɯ/ and the mid front /e ɛ/. The tongue height hierarchy observed for B21 is replicated to some extent by the other five speakers, as summarized in Table 2.

Table 2: Tongue front height hierarchies among front vowels for all six speakers. Bora front vowels are shown as /i_B ɛ ɨ/ and Spanish front vowels are shown as /i_S e/

speaker	front vowel height hierarchy
B2	i_S > i_B > e > ɛ > ɨ
B21	i_S > i_B > e > ɛ > ɨ
B20	i_S > i_B > e = ɛ > ɨ
B1	$i_S = i_B$ > e > ɛ > ɨ
B8	i_B > ɛ > i_S = e = ɨ
B7	i_S = e > i_B > ɨ > ɛ

All six speakers show a higher tongue blade for /ɨ/ than for any of the other vowels. In addition, four of the six speakers show significant raising/retraction of the tongue dorsum in /ɨ/, higher than in any of the other front vowels. B21 shows a similar shape that is not significantly higher than /ɛ e/, and B20 has no difference in posterior tongue dorsum height between /ɨ ɛ e/.

4. DISCUSSION

In terms of formant values, Bora's /ɨ/ and /ɯ/ are acoustically intermediate between /i/ and Spanish /u/. /ɨ/ and /ɯ/ both have low F1 frequencies consistent with high vowels, and intermediate F2 frequencies. With respect to /ɯ/, the explanation for its observed acoustic properties is straightforward: since it is produced with unrounded lips [3], it has somewhat higher F2 than Spanish /u/, regardless of whether it is produced with a constriction location similar to /u/ or with a somewhat more anterior constriction. Both types of articulations are observed in our data.

/ɨ/ is articulated with a very low tongue body, typically lower than for the high vowels /i/ and /ɯ/ and typically also lower than the Spanish /e/ and/or Bora /ɛ/. Although the tongue body is low for /ɨ/, the tongue blade is raised, which is consistent with the previously observed lingual-dental contact illustrated in the left panel of Figure 1. The previously analyzed video data have indicated central dental contact and lateral airflow for /ɨ/ [3]. For many of the speakers, /ɨ/ could be characterized as /ɛ/ with tongue blade raising. [18]'s reconstruction posits that Bora /ɨ/ is a reflex of Proto-Bora-Muinane /*e/, suggesting that this vowel underwent a change of dentalization rather than raising and retraction. The raised/retracted tongue dorsum is consistent with a lateral sonorant (expanding the tongue in the mid-sagittal plane as it narrows laterally). Tongue dorsum retraction is also consistent just with a dental articulation, since narrowing the tongue laterally is one way to achieve tongue blade fronting.

Given that /ɨ/ is produced with such a low tongue dorsum, it is potentially surprising that it has such a low F1 and has been classified on this basis as a high vowel. F1 in high vowels is typically a Helmholz resonance of the back cavity and the tongue passage. For Bora /ɨ/, the entire pharynx and oral cavity constitute the back cavity, and the passages around the sides of the tongue blade constitute the neck(s) of the Helmholz resonator. [19] describes such a configuration for the prototypical front rounded vowel [y], where the tongue passage and lip constriction form a single narrow tube, without any front cavity.

5. CONCLUSION

Ultrasound images of Bora vowels confirm that /ɯ/ is a high back or central vowel, whereas /ɨ/ is produced with a very low tongue body, meaning that its primary constriction is between the tongue blade and the teeth. It may still be considered a high vowel in acoustic terms, but articulatorily its status as a high vowel depends on how vowel height is defined. In terms of the relationship between the tongue dorsum and the palate, /ɨ/ is a low front vowel, but in terms of the relationship between the tongue blade and teeth, it is a high super-front (dental) vowel. This finding is consistent with the conclusions in [3], who additionally present phonological evidence demonstrating that /ɨ/ patterns in Bora as a front vowel which is not quite as high as /i/.

6. REFERENCES

[1] W. Thiesen and D. Weber, *A grammar of Bora: with special attention to tone*. Dallas: SIL International, 2012.

[2] S. Parker, "The acoustic qualities of Bora vowels," *Phonetica*, vol. 58, no. 3, pp. 179–195, 2001.

[3] J. Mielke and S. Parker, "Bora high vowels involve a two-way dental contrast, not a three-way backness contrast," *Laboratory Phonology*, accepted.

[4] S.-I. Lee-Kim, "Revisiting mandarin 'apical vowels': An articulatory and acoustic study," *Journal of the International Phonetic Association*, vol. 44, no. 3, pp. 261–282, 2014.

[5] W. Thiesen and E. Thiesen, "Diccionario bora-castellano, castellano-bora," 1998.

[6] A. Wrench, *Articulate Assistant Advanced User Guide (Version 2.17.02)*. Edinburgh: Articulate Instruments Ltd, 2017.

[7] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[8] M. McAuliffe and M. Sonderegger, "Spanish MFA acoustic model v2.0.0," https://mfa-models.readthedocs.io/acoustic/Spanish/SpanishMFAacousticmodelv2_0_0.html, Tech. Rep., Mar 2022.

[9] E. Ahn and E. Chodroff, "VoxCommunis corpus," https://osf.io/t957v, Jan 2022.

[10] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, 2007, [Computer program]. [Online]. Available: http://www.praat.org

[11] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[12] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using deeplabcut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[13] C. Gu, *Smoothing spline ANOVA models*. New York: Springer-Verlag, 2002, springer Series in Statistics.

[14] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[15] J. Mielke, "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons," *Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2858–2869, 2015.

[16] C. Gu, "Smoothing spline ANOVA models: R package gss," *Journal of Statistical Software*, vol. 58, no. 5, pp. 1–25, 2014.

[17] R Core Team, "R language definition," 2015.

[18] F. Seifart and J. A. Echeverri, "Proto Bora-Muinane," *LIAMES*, vol. 15, no. 2, pp. 279–311, 2015.

[19] G. Fant, *Acoustic theory of speech production*. The Hague: Walter de Gruyter, 1970.

GHOST SEGMENTS IN THE FLEMISH TUSSENTAAL

Mishko Bozhinoski & Paul Boersma

University of Amsterdam

ABSTRACT

The Flemish informal standard (*Tussentaal*) exhibits a word-final elision process whereby words such as 'wat' and 'met' are realised without their final plosive: [βɑ] and [mɛ]. This elided segment triggers devoicing in following fricatives. This process is not derivable from phonological context alone, and the current literature has not explained its distribution. This study therefore set off to (1) document this elision process, and (2) account for its distribution.

Using Python and the corpus engine OpenSonar, audio tokens of the fifty most common Dutch words containing final coronal plosives were scraped from the Corpus Gesproken Nederlands. Joint results of an algorithmic and a manual analysis reveal that this elision occurs in the present tense marker /-t/ and most function words. All exceptions found have historically undergone word-final schwa deletion, suggesting that final coronal plosive elision may be older. Evidence from 13th century texts supports this claim.

Keywords: Franconian dialectology, ghost segments, sound change, corpus linguistics, sandhi

1. INTRODUCTION

A peculiar and often ridiculed feature of colloquial Standard Flemish (henceforth *Tussentaal*) is the elision of expected final coronal plosives in words such as *wat*, *niet* and *met*, whereby they are regularly realized without their final plosive: [βɑ], [ni] and [mɛ]. The result of such elision frequently results in word-final lax vowels and monomoraic phonological words, which are impermissible in Standard Dutch [1, p.26]. It also results in a devoicing of following voiced fricatives, as if this dropped *-t* was still there [2]:

Ghost *-t*	Pronounced *-t*	No *-t*
<met zand>	<nat zand>	<blauw zand>
[mɛ **s**ɑnt]	[nɑt **s**ɑnt]	[blɑu **z**ɑnt]
'with sand'	'wet sand'	'blue sand'

Table 1: Ghost *-t* and its effect on adjacent consonant

Additionally, in cases of hiatus across word boundaries, whereby the first word ends in an otherwise elided coronal plosive, a [d] surfaces in between the vowels [3]:

Ghost *-t*	Pronounced *-t*
<wat een>	<vat een>
[βɑ**d**ə(n)]	[vɑ**t**ən]
'what a...'	'catch a...'

Table 2: Liaison of ghost *-t*

The distribution of this ghost segment, henceforth transcribed as /ᵗ/ or [ᵗ][1], seems on the surface to be phonologically unpredictable from a synchronic perspective:

No *-t*	Ghost *-t*	Pronounced *-t*
[mu] 'tired'	[muᵗ] 'must'	[mut] 'courage'
[zɛː] 'said'	[zɛːᵗ] 'are'	[tɛːt] 'time'
[ɣaː] '(I) go'	[ɣaːᵗ] 'goes'	[laːt] 'let'

Table 3: Unpredictable distribution of ghost *-t*

While the existence of this phenomenon has been documented in the literature [2, 3, 4], accounts for its distribution, either synchronically or diachronically, are cursory at best. The present research aims to fill this gap through an exploratory corpus study, with the following research questions:

1. What is the synchronic distribution of /ᵗ/?
2. Is there a diachronic account for /ᵗ/?

2. METHOD

2.1. Corpus

The present study is based on corpus data provided by the Corpus Gesproken Nederlands [5], henceforth CGN, consisting of speech collected in various contexts, ranging from spontaneous face-to-face conversations to TV and radio interviews to religious services and speeches. The data pool used in the present study is only that of face-to-face conversations in order to ensure an informal colloquial register as opposed to one in which Standard Dutch may be more appropriate.

The corpus is accessed using OpenSoNaR [6], a corpus search engine used alongside both CGN as

well as a written corpus that is not relevant to the current research. CGN is annotated for phonetic and orthographic information in the *TextGrid* format, chiefly employed by *Praat* [7], in which each word is encapsulated within an interval associated with plain text. This facilitates the query of data based on orthographic (or phonetic) information.

2.2. Conditions and coding

Individual tokens examined by this study consisted of word pairs in which the first word ends (orthographically) in *-t* or *-d*, and the word right after begins with an underlyingly voiced fricative. The word pairs were selected on the basis of two conditions: the first word, which ends in *-t* or *-d*, and the fricative that the following word begins with.

2.2.1. Word conditions

The word condition is filled by the fifty most frequent words (orthographically) ending in *-t* or *-d*, as per *A Frequency Dictionary of Dutch* [8]. To limit the scope of the current study, nouns containing a maximum of one full (non-schwa) vowel are to be used. This restriction also minimizes potential stress-related effects. Thus, words such as /minyt/ 'minute', /rezʏltaːt/ 'result' and /ɪnhɑud/ 'content' are not allowed, while /ɣəbid/ 'area' and /ɣəlœyd/ 'sound' are. No such restriction was applied to non-nouns. The words were divided into four categories: verbs, adjectives/adverbs, nouns and other (prepositions, determiners and conjunctions).

2.2.2. Fricative conditions

This study concerns itself with the phonetic realization of the ghost segment /ᵗ/, and its realization when adjacent to a fricative. Namely, what is relevant is (a) whether the segment is realized and (b) whether the following underlyingly voiced fricative in question is voiceless. As Dutch contains three native voiced fricatives, namely /v/, /z/ and /ɣ/, each word undergoes three batches of data collection: one in which it is followed by /v/, one by /z/ and one by /ɣ/.

2.3. Data collection

To limit the time required to manually process the data, each word–fricative pair was limited to ten tokens. This means that the ideal number of tokens collected is 1500 (3 fricatives $\cdot$ 50 words $\cdot$ 10 tokens). The collection of the relevant data was carried out using a process by which web-hosted data are extracted and saved locally, a process known as web-scraping. This was carried out using a script created for this study written in the scripting language *Python*, and chiefly using the web-automation library *splinter* [9] and the audio processing library *pydub* [10]. The program's workflow can be summarized as follows:

1. Send a corpus query to OpenSoNaR per element of the cartesian product $W \times F$, where W consists of all target words and F consists of the three voiced fricatives.
2. Discard all but one result per audio file. This is to guarantee that each audio clip comes from a different speaker.
3. Retrieve the first 10 results (where available) of each query, as well as the TextGrid boundaries of the beginning and end of the word pair.
4. Download the audio file of each query and cut it in accordance with the boundaries acquired in step 2.

2.4. Analysis

Using *Praat* [7], an acoustic analysis was performed per word pair with respect to two variables: the presence of an alveolar plosive in the first word, and the voicing of the initial fricative in the second word. Due to the unreliable quality of the audio, plosive presence was analyzed using a hybrid automated–manual approach.

2.4.1. Plosive presence

The detection of the presence of a plosive was conducted using *Praat*. Firstly, the data were manually annotated so that the whole span of the fricative was tightly encapsulated within a TextGrid interval. These intervals were additionally marked with *1* or *0*, representing respectively the perceived presence or absence of a plosive before them. This binary marking is henceforth referred to as T. Then, a script was written to execute the following tasks:

1. For each fricative interval, extract a clip running from 100 ms to 0 ms before the beginning of the fricative.
2. Cut off all frequencies below 500 Hz. This was done due to the presence of unwanted noise, presumed to be a result of the relatively low recording quality.
3. Take the difference between the highest and lowest points of intensity ΔI within the clip.

It was expected that the presence of a plosive would result in a high ΔI, and inversely that the lack of a plosive would show a low ΔI.

2.4.2. Fricative voicing

Cues for fricative voicing in Dutch have been documented to include periodicity, duration and to a lesser degree intensity [11]. The cue used to identify fricative voicing within this study was that of duration, as it is least dependent on the quality of the data. First, a control study was conducted to establish whether a duration cue for voicing is reliably found in the data. Data were scraped using the same script as in §2.3, with the word conditions all ending in a vowel or sonorant, and the fricative conditions containing both underlyingly voiced and voiceless fricatives. The extracted audio files were then annotated in the TextGrid format, such that the span of each target fricative corresponds to the span of an interval. The annotation was carried out manually. There was no way to guarantee unbiased annotation within the temporal constraints of the study.

Using a Praat script, the duration of each interval and the consonant it corresponds to was extracted and saved onto a plain text file, such that each token interval is represented by its duration d and a binary voicing parameter v, such that <z>, <v> and <g> are represented by '+v', and <s>, <f> and <ch> are represented by '-v'. A total of 66 token fricatives were scraped and annotated. Finally, with the help of R [12], the data were fit onto a linear model of $\ln d$ as a function of the binary voicing parameter v. This yielded that the average duration of voiceless fricatives was longer than that of voiced fricatives (v-coef. $= 0.752$; conf. interval $= 0.592 \ldots 0.911$; $t(64) = 9.452$; $p = 9.4 \cdot 10^{-17}$).

2.5. Data processing

The majority of the data was formatted and partially processed using *Python* and *NumPy* [13]. Per word, the following values were calculated:

- RT = the ratio of tokens with deleted [t] over all tokens of a given word
- n = the number of tokens per word
- ΔI_m = the mean of ΔI over n

The values of T, ΔI, and d per token were also saved to a text file to be fit onto several linear models using *R*. This is the per-token analysis.

3. RESULTS

3.1. Per token

In total, 848 tokens from 39 words were collected. Data were not collected for any pair for which fewer than three tokens were available.

3.1.1. Computation of ΔI versus manual marking

To check whether the manual annotation T – the manual binary annotation of 1 (final plosive present) and 0 (final plosive absent), agrees with the automated measurement ΔI – the difference between the highest and lowest point of intensity of a token, they were fitted onto a linear model of ΔI as a function of T. The value of ΔI was an estimated 5.21 dB higher (as expected in the case of a present plosive) when $T = 1$ than when $T = 0$. This effect was statistically significant (95% conf. interval $=$ 4.01 dB ... 6.41 dB; $t(363) = 8.66$; $p = 1.6 \cdot 10^{-16}$).

For §3.1.2, in which a binary decision of voicing is needed, we use the manual voiced/voiceless marking rather than the continuous ΔI. Using ΔI would require a conversion to a binary value that the reliability of ΔI cannot afford.

3.1.2. Voicing of fricative after ghost t

To determine the ghost -*t*'s effect on fricative voicing, the durations of fricatives that follow a phonetically absent -*t* was compared to those of known voiced and voiceless fricatives, as established in §2.4.2. A data set was constructed consisting of all known voiced (marked 'v') and voiceless (marked 'u') tokens alongside all tokens of fricatives directly following a ghost -*t* (marked 'a'). The values of d (fricative duration) and V (ternary marking of 'a', 'u' and 'v') were fit onto a linear model of $\ln d$ as a function of V, with an orthogonal contrast measuring the distance of $\ln d$ when $V =$ 'a' from the average of $\ln d$ of fricatives following known voiceless and voiced plosives. The value of $\ln d$ when $V =$ 'a' was shown to be significantly greater than the average $\ln d$ of voiced vs voiceless fricatives (est. distance from mean $= 0.267$; 95% conf. interval 0.171 ... 0.363; $t(187) = 5.46$; $p = 1.5 \cdot 10^{-7}$). We conclude that fricatives that follow a ghost -*t* are on average closer to postvocalic voiceless than to postvocalic voiced fricatives (any remaining difference with the postvocalic voiced fricatives can be due to lexicalized cases of variation, which there is no room to discuss here).

3.2. Per word

The per-word analysis is taken as relevant only if the word in question has ten or more tokens representing it. This adds up to thirty words. Below is a summary of those thirty words depending on their value of *RT* – the ratio of tokens which the manual data marks as having a deleted final [t]. In bold are function words as well as conjugational affixes.

- $RT \geq 0.5$: ***dat***, ***doet***, ***niet***, ***omdat***
- $0.2 \leq RT < 0.5$: ***gaat***, *goed*, ***met***, *staat*, ***wat***
- $RT < 0.2$: ***moet***, *weet*, ***had***, *zat*, *altijd*, *bed*, *groot*, *kwaad*, *kwijt*, *laat*, ***net***, *nooit*, *ooit*, *stad*, *straat*, *tijd*, ***tot***, ***uit***, *zet*, *zit*

4. DISCUSSION

4.1. Effect of ghost *-t* on following fricatives

The results of §3.1.2. show us that the ghost *-t* have a duration (which we take to be a voicing cue) that is significantly more similar to voiceless fricatives than to voiced fricatives, suggesting that the ghost *-t* has a devoicing effect on following fricatives. This is in-line with the observations of Camerman [3] and De Schutter [2].

4.2. Distribution of ghost *-t*

The distribution of *t*-deletion seems to be limited to function words, with the potential exception of *goed* 'good, well' – which is a high frequency content word, although its vague meaning and modal usage may qualify it as a function word. While *t*-deletion seems to only happen in function words, being a function word alone is not enough to predict *t*-deletion. Namely, the words, *had* 'had', *tot* 'to, until', *uit* 'from, out' and *net* 'just' do not show any *t*-deletion in the dataset.

The lack of deletion in *net* can be explained by considering that it may be a loan from the standard language. This is supported by its lack in dialect dictionaries such as the Antwerps Woordenboek [14] (in Antwerp Brabantian, for instance, we see the word *zjust*, from French *juste*, covering the semantic range of *net*).

A possible explanation for the lack of *t*-deletion in *had*, *tot* and *uit* is the presence of a schwa following the final obstruent in their older forms. Namely, *had*, *tot* and *uit* appear as *had(d)e*, *tote* and *uite* in their Middle Dutch forms. This suggests that this *t*-deletion may have occurred before word-final schwa was deleted, and indeed there is evidence of this in the corpora. See the below examples sourced from Corpus Gysseling Vol I [15]:

(1) (0295, Oudenbiezen, 1280 april 22–28)
Des lants hef si geloft te mestene no hore magt, en **dassi** (< dat zij) mest do salsi VI joer derno 15ane hebben. Die selue Mente heft XV ruden bamt ombe VII vaet roghen Tungers ter seluer getonst **dassi** (< dat zij) dlant heft.

(2) (0729, Brugge, 12 juli 1287)
en*de* in ke*n*nesse van desen sticke*n* dat sie vast bliue*n* en*de* ghestade hebbe*n* vorseid*e* redenars dese lett*re*n gheseghelt **me** (< met) onsen zeghell*en* huthan*g*hen*d*e dit was ghedaen sat*er*daghes vore alf hoimaent.

5. CONCLUSION

Pertaining to the topic of the current study, one may take away the following.

Firstly, the ghost segment /^{t}/ is limited to function words that historically end in /t/ or /d/, represented orthographically by <t> and <d> respectively. Function words that synchronically end in /t/ or /d/ robustly realized as [t] seem to comprise words that historically end in a schwa instead of a plosive, suggesting that this phonetic weakening of final coronal plosives chronologically precedes word-final schwa-deletion. This is supported by historical corpus data dating from the 13th century.

Secondly, the ghost *-t*, despite its phonetic absence, causes devoicing in fricatives that directly follow it.

Finally, there are many gaps in our knowledge pertaining the ghost segment /^{t}/, gaps which hide critical insight into both synchronic and diachronic Franconian linguistics. These gaps may viably be explored using computational methods, but our current spoken language corpora are thusfar insufficient for these purposes. We plan to address this in a future project, which will recreate this study with a corpus at least three orders of magnitude the size of the one of the present study.

6. REFERENCES

[1] G. E. Booij, *The phonology of Dutch*, 1st ed., ser. The phonology of the world's languages. Amsterdam: Oxford Univ. Press, 1999.

[2] G. De Schutter, "Het Antwerps. een schets van zijn evolutie tussen 1898 en 1998," *Honderd jaar stadstaal.*, pp. 301–315, 1999.

[3] F. Camerman, *Antwerps schrijven : spelling en grammatica van het 21e-eeuwse Antwerps*. Antwerpen: de Vries-Brouwers, 2007.

[4] K. R. Haeseryn, K. Romijn, G. Geerts, J. de Rooij, and M. van den Toorn, "1.5.2.6 Deletie van t," 2020, retreived January 03, 2023. [Online]. Available: https://e-ans.ivdnt.org/topics/pid/topic-16007638353447182

[5] L. van Eerten, "Corpus gesproken Nederlands," *Nederlandse taalkunde (Groningen)*, vol. 12, no. 3, pp. 193–214, 2007.

[6] N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman, "The construction of a 500-million-word reference corpus of contemporary written Dutch," in *Essential speech and language technology for Dutch*. Springer, Berlin, Heidelberg, 2013, pp. 219–247.

[7] Boersma, Paul and Weenink, David, "Praat: doing phonetics by computer," 2022. [Online]. Available: http:\\praat.org

[8] C. Tiberius and T. Schoonheim, *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge, 2013.

[9] A. Medina, B. Barreto, D. Fernandes, D. S. de Andrade, D. Camata, R. Goncalves, E. Flores, F. M. Costa, F. Missi, F. Ribeiro, L. Geron, B. Heynemann, H. L. Tavares, H. S. Antunes, J. Bochi, L. Lemos, M. de Oliveira, G. Rezende, R. Caricio, E. B. d. Luz, T. Azevedo, T. Al-Chueyr, and T. Santos, "splinter (version 0.17)," 2021. [Online]. Available: https://github.com/cobrateam/splinter

[10] J. Robert, M. Webbie *et al.*, "Pydub," 2018. [Online]. Available: http://pydub.com/

[11] de Jonge, Mirjam and Aligbeh, Isaura, "Duration as a voicing cue in Dutch," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, 2019.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[13] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[14] F. Camerman, "Antwerps.be," 2009. [Online]. Available: https://www.antwerps.be/

[15] M. Gysseling, "Corpus van Middelnederlandse teksten (tot en met het jaar 1300), mmv Willy Pijnenburg," *Reeks I: Ambtelijke bescheiden. 's-Gravenhage, Nijhoff*, 1977.

[1] The transcription /^{t}/ is not meant to imply that the ghost *-t* is underlyingly voiceless. It is a largely arbitrary choice motivated by its orthographical representation and its correspondence to historical /t/ in most cases.

INTERACTION OF QUANTITY, FOOT STRUCTURE, AND STRESS IN THE 2ND/3RD SYLLABLE SONORANTS OF SOIKKOLA INGRIAN TRISYLLABLES

Natalia Kuznetsova[1,2], Irina Brodskaya[3], Elena Markus[4,5]

[1]Università Cattolica del Sacro Cuore; [2]Institute for Linguistic Studies, Russian Academy of Sciences; [3]Queen Margaret University; [4]University of Tartu; [5]Institute of Linguistics, Russian Academy of Sciences
natalia.kuznetsova@unicatt.it, i.brodskaya@gmail.com, elmarkus@ut.ee

ABSTRACT

The talk describes interaction between phonological quantity (originally, binary) and the compensatory effects of foot structure in sonorants at the 2nd and the 3rd syllable boundary (C_3) in the trisyllabic foot of vanishing Soikkola Ingrian (Finnic). Such interaction had been previously observed as a phonetic effect in the 1st syllable segments and as a phonological impact on the 2nd syllable vowels [1]. C_3 sonorants are also affected, both phonologically and phonetically, and their binary length contrast is almost lost. Depending on the foot structure, original C_3 sonorant geminates are either retained, or nearly or completely shortened to singletons. This result is placed within other results on the trisyllabic foot, which appears to be a coherent metrical unit different from both the disyllabic foot and the bifoot trisyllabic word in Soikkola Ingrian.

Keywords: quantity contrast loss, trisyllabic foot, sonorants, temporal compensation, Soikkola Ingrian.

1. INTRODUCTION[i]

The Soikkola dialect of the Ingrian language (Finnic), spoken at the Russian-Estonian border, manifests the trisyllabic metrical foot, e.g. in *murǩkina* [ˈmurkˑina] 'breakfast'. The trisyllabic foot shows phonetic and phonological differences from both the disyllabic foot, e.g. *kurki* [ˈkurĝi] 'crane', and the combination of a disyllabic and a monosyllabic foot, e.g. *murkinā* [ˈmurĝiˌna:] 'breakfast:PRT'. In particular, in a major phonetic study on 22 Soikkola trisyllables [1], the trisyllabic foot showed stronger compensatory effects than the disyllabic foot. Compensatory effects imply a shortening of segments as a function of an increase in the quantity and number of adjacent segments.

In the first syllable segments (V_1 — the 1st syllable vowel, C_2 — a consonant at the 1st and the 2nd syllable boundary, R — an optional sonorant between them, as in *murǩkina*), all compensatory effects were purely phonetic. However, in the second syllable unstressed vowel (V_2), they led to a merger of the binary length contrast in trisyllables, unlike in the disyllabic foot, which still retains the contrast of short and long V_2.

The present study explores the influence of foot structure on the quantity contrast in C_3 consonants, which follow V_2. In C_2, the quantity contrast of consonants in Soikkola Ingrian is ternary, which is an extreme phonological rarity [2]–[4]: *ka**n**a* [ˈkana:] 'hen' (singleton) — *ka**nn**ā* [ˈkanˑa:] 'hen:PRT' (short geminate) — *li**nn**ā* [ˈlin:a:] 'city:PRT' (long geminate). In C_3, the contrast is only binary (singleton vs. geminate). C_3 geminates historically correspond to C_2 long geminates. Hence the typical transcriptions of the C_3 length contrast in older Soikkola sources [5] as *sūri**m**a* 'grain' vs. *sūvi**mm**a* 'judge:1PL' (sonorants), *hōva**tt**a* 'relax:INF' vs. *sīvo**tt**a* 'cattle' (stops).

However, geminates in Finnic languages, including Ingrian, are shorter after an unstressed syllable (i.e. as C_3) than after a stressed syllable (i.e. as C_2) [6, p. 274], [7]. Moreover, sonorant geminates differ from stops in that they tend to lose gemination and become singletons in the C_3 position. The shortening of C_3 sonorant geminates has been observed as complete in Estonian and as ongoing in Finnish dialects [6, pp. 273–277] and in Ingrian dialects, including Soikkola Ingrian [8]–[10], but never experimentally studied.

In our dataset from [1][ii], the shortening of C_3 geminate sonorants in Soikkola Ingrian can be studied on the first plural verbal forms with the suffix *-mma/-mmä* and the nominal forms in allative with the suffix *-lle* (or *-lla/-llä*), e.g. *sāta-**mm**a* 'get-1PL', *palka-**ll**e* 'salary-ALL'. Original geminate **l:* (160 tokens) can be also directly compared to singleton *l* in forms like *katti**l**a* 'cauldron' (828 tokens). In turn, the forms with original C_3 geminate **m:* (988 tokens) do not have any counterparts with singleton *m* in these data. Datasets on **l:-l* and **m:* are imbalanced also in other respects (see 3.1). However, the overall variety of the structural types represented in these data is still sufficient for making certain conclusions about the process of shortening of the C_3 sonorant geminates.

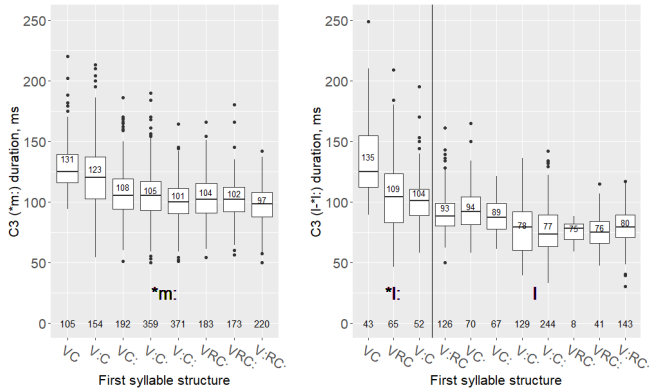
We find that C_3 long sonorants are durationally at most short geminates. We confirm an ongoing merger of C_3 singleton and geminate sonorants. This process has a compensatory nature. It is largely regulated by the 1st syllable structure: the more complex the latter, the shorter any C_3. The C_3 geminates, like the V_2 short vowels, are the longest in the "light" foot with the 1st stressed light syllable (C)V: *otaǩma* [ˈod**a**:**m**ˑa]. This is an effect of stress-induced lengthening, untypically located in the 2nd unstressed syllable of the light foot in Finnic languages and affecting also C_3 duration.

2. DATA AND METHODS

Phrase-final tokens from [1], used for this study, were collected in the Kingisepp district (Russia) in 2014-2016 from five female speakers of north Soikkola Ingrian born in 1929-1936. Speakers were presented with a phrase spoken in Russian and asked to translate it into Ingrian. Mixed effects linear regression models are fitted in R with *lme4* on raw durations to study the relations between the structural parameters in the first two syllables of the foot and the duration of C_3 sonorants. The quantity contrast in V_2 has been lost and is excluded from all models. The fixed effects are V_1 length (*V* vs. *V:* — short vowel vs. long vowel), C_2 length (*C* vs. *C·* vs. *C:* — singleton vs. short geminate vs. long geminate), and the presence of a sonorant R (*yes* vs. *no*), in various interactions. Models also include random by-speaker and by-word intercepts.

3. RESULTS

3.1. Simplification of the datasets on C_3 sonorants



(a) **m: as C_3* (b) **l:-l as C_3*

Figure 1: Mean durations in ms of C_3 (a)**m:* and (b)**l:-l* grouped by the combinations of V_1 length, C_2 length, and the presence of a sonorant R represented in our data.

			(a) *m*-dataset	(b) *l*-dataset	
V_1	**R**	**C_2**	**C_3**	**C_3**	
			****m:***	****l:***	***l***
short	no	sing	105	43	0
		gem	190	0	137
	yes	sing	0	65	0
		gem	358	0	49
long	no	sing	155	52	0
		gem	730	0	373
	yes	sing	0	0	126
		gem	220	0	143

Table 1: Numbers of tokens per combinations of V_1 length, C_2 length (singleton vs. any geminate), and the presence of R in our data on (a)**m:* and (b)**l:-l*.

Figures 1a-b represent the duration of original C_3 geminates **m:* and **l:* and C_3 singletons *l* as a function of combinations represented in the dataset from [1] of the levels of the three fixed effects in the 1st syllable. Raw means are reported within boxes; the number of tokens is given below each box at the bottom.

These data contain structural gaps, because the data from [1] were not intended for studying C_3. However, this can be partially remediated by conflations along certain parameters. Preliminary statistical exploration of data showed that short geminate and long geminate C_2 did not significantly differ in their impact on C_3 duration in any fitted model. In Tables 1a-b, they are, therefore, conflated into one category of geminates (*"gem"*), as opposed to C_2 singletons (*"sing"*).

In the model on **m:* (3.2), also the factor of R is excluded, as it was not significant for those pairs of structures for which it could be studied: V:C·– V:RC·, VC: – VRC:, cf. Fig. 1a. This makes the V_1 length * C_2 length matrix of the *m*-dataset full, cf. Table 1a.

In the *l*-dataset, C_3 length (singleton *l* vs. geminate **l:*) is collinear to the unique combinations of the levels of the three structural factors in the 1st syllable. Therefore, the factor of C_3 length itself cannot be included in a model. However, C_3 length can still be explored in the post-hoc pairwise comparisons of means obtained with a model containing the ternary interaction of the factors in the 1st syllable (see 3.3).

3.2. Results of modelling on **m:* (m-dataset)

Fig. 2 shows the results of C_3 **m:* duration modelling according to the following formula: C_3 duration ~ V_1 length * C_2 length + (1 | speaker) + (1 | word), data = 1758 tokens. The interaction is insignificant but is kept in the model, because this helps to study the relationships between the light foot (in bold in Fig. 2) and the heavy foot (any other type of foot).

A post-hoc pairwise comparison of the modelled means in *emmeans* shows no significant differences in C_3 between the structures with a short vs. a long V_1 when C_2 is a singleton: [ˈoda:**m**·a] — [ˈša:da**m**·a] (9 ms, SE = 4.6, df = 50, t = 2, p = 0.052; a black line). In both structures, C_3 **m:* still retain gemination.

When C_2 is a geminate, there is also no difference in C_3 between the structures with a short vs. a long V_1 (4 ms, SE = 1.9, df = 53, t = 2, p = 0.06; a red line), but both C_3 **m:* are almost shortened. Their durations of 101 and 106 ms are similar to the duration of word-initial *m* in our dataset (106 ms, 338 tokens). All other differences in Fig. 2 are significant at $p < 0.0001$.

The results on **m:* show that C_3 gemination is kept when C_2 is a singleton, but lost when C_2 is a geminate: *otaműma* [ˈoḑa:**m**·a] 'take:1PL', *sātaműma* [ˈša:ḑa**m**·a] 'get:1PL', but [ˈpałk:ama] 'hire:1PL' (<**palkkāmma*), [ˈłak:ama] 'sweep:1PL' (<**lakkāmma*), [ˈkä:nt·ima] 'turn:1PL' (<**kāntt̆īmmä*), [ˈšu:t·ima] 'judge:1PL' (<**suutt̆īmma*), [ˈleik:ama] 'cut:1PL' (<**leikkāmma*), [ˈhülk·äma] 'abandon:1PL' (<**hülkk̆āmmä*).

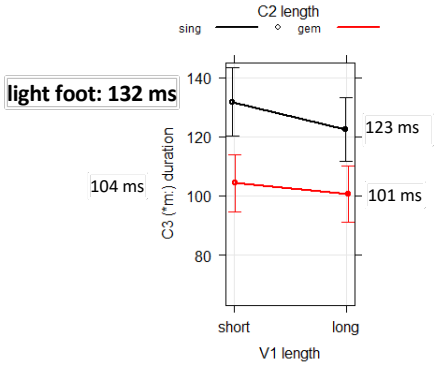


Figure 2: Modelled least squares mean duration of C_3 **m:* as a function of V_1 length and C_2 length (Satterthwaite's approximation of df and a confidence level of 95%).

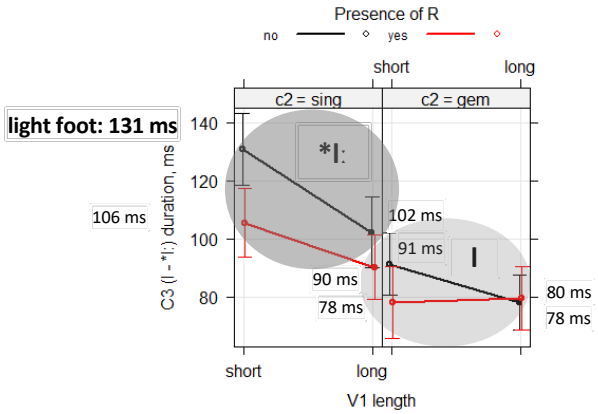


Figure 3: Modelled least squares mean duration of C_3 **l:-l* as a function of V_1 length, C_2 length, and the presence of R (Satterthwaite's approximation of df and a CL of 95%).

3.3. Results of modelling on **l:-l* (l-dataset)

Fig. 3 shows the results of C_3 **l:-l* duration modelling according to the following formula: C_3 duration ~ V_1 length * C_2 length * Presence of R + (1 | speaker) + (1 | word), data = 988 tokens. The structures which contain original C_3 **l:* are shaded dark grey, while the structures with original short *l* as C_3 are in light grey.

C_3 **l:*, similarly to **m:*, has the longest duration of 131 ms in the light foot (marked in bold in Fig. 3; *akaľle* [ˈåga:**l**ˑe] 'old_woman:ALL'). It is significantly different at $p < 0.0001$ from any other type in Fig. 3.

The remaining two types of **l:* in Fig. 3 (**palkalle* 'salary:ALL' and **koikalle* 'bed:ALL') are shorter: 106 ms and 102 ms. Statistically they do not differ from each other, but differ from all types of originally short **l* as C_3, including durationally the nearest types of **l* in *kattila* [ˈkat:i**l**a] 'cauldron' (91 ms) and *kāntelō* [ˈkä:nde̯**l**ö:][iii] (90 ms), although only at $p < 0.05$ or at $p < 0.01$ (cf. also with modelled 89 ms of word-initial *l;* 436 tokens). The difference is anyway too small to be perceived by humans [11], [12]. Gemination of **l:* in these two structures can be, therefore, considered nearly or entirely lost: [ˈpałg̊a**l**(ˑ)e] and [ˈkoi̯g̊a**l**(ˑ)e].

In the duration of short *l*, we observe the same kind of compensatory trend which distinguished the light foot from the heavy foot in case of **l:*. The duration of 91 ms of *l* in *kattila* with a short V_1 and no sonorant significantly differs from that of *l* in structures with a long V_1 (78 ms; *rīt̆tele* [ˈri:tˑe**l**e] 'argue:IMP', *Mīkkula* [ˈmi:k:u**l**a] 'St. Nicholas'), with a sonorant (78 ms; *kant̆tele* [ˈkantˑe**l**e] 'kantele', *küntteli* [ˈkünt:e**l**i] 'candle'), or both (80 ms; *vān̆teli* [ˈvä:ntˑe**l**i] 'turn:IPF:3SG'), but only at p<0.05 or at p<0.01. C_3 duration before a long vowel in a bifoot structure *kāntelō* 'turn:3SG' (90 ms; cf. Ftn. *ii*) is left aside here.

This is clearly just a phonetic shortening of *l* — a manifestation of how durations in the foot shorten in an inverse relation to the growing number or quantity of segments within the foot, further discussed below.

4. DISCUSSION

4.1. Similarities and differences between **m:* and **l:-l*

Our data confirm the ongoing shortening of **m:* and **l:* as C_3, i.e. after an unstressed syllable, in the trisyllabic foot. The earliest data from 1884 [8] did not yet report this process. Data from the 1930s [13] did not attest it for **m:* but reported the shortening of **l:* in the allative suffix, although only as C_2 in some personal pronouns (e.g. *miu-lle* ~ *miu-**l**e* 'me:ALL', p. 78). However, the data on the allative suffix and the verbal 1st plural from the 1960s [9] and from the 2000s [10] already reflect a wide variation in **m:* and **l:* length as C_3 in the trisyllabic foot (and as C_2 after a long vowel, like in the personal pronouns above). The present data, although limited, allow us to clarify some mechanisms of the shortening of C_3 sonorants.

Geminate sonorants **m:* and **l:* undoubtedly retain gemination in the light foot. However, their duration as C_3 (~130 ms) does not reach the duration even of short geminate stops in C_2 (modelled 177 ms in [1]; cf. with 234 ms for long geminate C_2 stops). Shorter duration of C_3 geminates than of C_2 geminates was expected on the basis of earlier Ingrian and Finnish data. Synchronic C_3 sonorant geminates can be phonologically considered as short geminates /Cˑ/.

In this, they might potentially differ from C_3 stops, which impressionistically do not show any signs of phonological shortening in Soikkola Ingrian (in words like *sīvo**tt**a* 'cattle'), but this is yet to be verified on the newly collected data.

In all studied types of heavy foot, **m:* and **l:* have almost or entirely shortened, apart from the type with **m:* and a long V_1 plus singleton C_2 (i.e. "V:C", *sāta**m̆m**a* 'get:1PL'; ~120 ms), where the duration of C_3 **m:* does not significantly differ from that of C_3 **m:* in the light foot. Still, in the same type of foot, **l:* has shortened to ~100 ms (**koikalle* > *koika**l**e*).

To investigate whether the difference in the "V:C" foot between **m:* and **l:* is systematic or due to some possible gaps in the data, we considered inter-speaker differences in the structures with a singleton C_2 for **m:* and **l:*. Broadened inter-speaker variability between competing phonological variants is typical of an unfinished sound change [14]. Fig. 4 shows the results of modelling on *m*- and *l*-datasets by a formula: C_3 duration ~ First syllable structure * Speaker + (1 | word). In Fig. 4a, all the types from Fig. 1a where **m:* occurs after any type of a geminate are conflated into *"*m:_after_gem"*. In Fig. 4b, all types from Fig. 1b with a short *l* are conflated into *"short_l"*. Fig 4a-b show consistent short geminates in the light foot and fully short consonants in the conflated types. There is, however, a large inter-speaker variability of **m:* and **l:* in heavy feet with a singleton C_2 (highlighted by circles in Fig. 4). In particular, in the "V:C" foot, C_3 **m:* has actually shortened in 2 out of the 5 speakers.

It appears that C_3 **m:* and **l:* retain gemination in the light foot, are in a process of shortening in heavy feet with singleton C_2, and are (almost) shortened in other heavy feet. C_3 **m:* is a bit more conservative than C_3 **l:*. In both cases, we anyway see the effects of the general compensatory trend attested also for other foot segments in [1] (see Introduction). In 4.2, we discuss the relevance of our findings on trisyllabic temporal compensation for stress and foot typology.

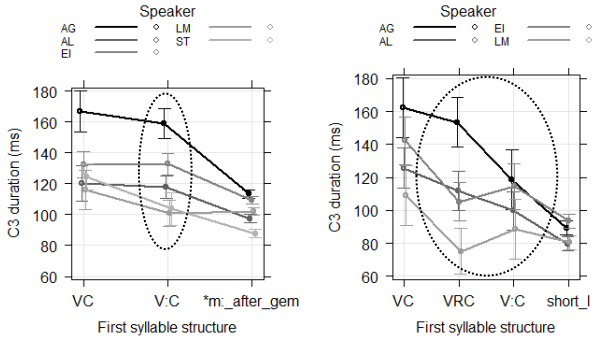


Figure 4: Modelled durations of (a) **m:* (left) and (b) **l:-l* (right) as C_3 with a focus on structures with singleton C_2

4.2. Significance of results for stress and foot typology

Our most robust finding as regards the C_3 geminate sonorants is their clear preservation in the light foot, as opposed to their shortening in the heavy foot. This fact can be also related to a phonetic lengthening of the preceding short V_2 in the light foot (modelled 127 ms — durationally the longest V_2 type in [1]).

This cross-linguistically untypical phenomenon of lengthening of unstressed V_2 rather than of stressed V_1 in the light foot in Finnic languages has been long known as a "half-long vowel" [15], [16]. For Standard Finnish, the "half-long vowel" has been explicitly considered as a manifestation of stress [17].

Our present data show that this lengthening in the 2nd syllable of the light foot spreads not only to the syllable nucleus (i.e. V_2) but also to the coda. In the structures considered, the coda consists of the first part of C_3 geminates, e.g. *o-**tam̆**-ma* [ˈoo̯a:mˑa]. Stress-related lengthening in the 2nd syllable likely prevents C_3 geminates from phonological shortening.

In the heavy foot, which is in turn characterised by the metrical weakening of the 2nd unstressed syllable, both long V_2 [18], [1] and long C_3 shorten, e.g. **lak-**kām-ma*** > [ˈłak:**am**a] 'sweep:1PL'. Both processes are observed as still ongoing in Soikkola Ingrian and complete in more prosodically innovative Estonian.

Table 2 draws a comprehensive picture of the compensatory processes affecting segmental duration and quantity within the trisyllabic foot in Soikkola Ingrian on the basis of this study and findings in [1]. Compensatory processes represent an interaction of the foot structure (first of all, light vs. heavy foot, with finer gradations within heavy feet) and segmental quantity. Word-initial consonant C_1 is not affected, neither the 3rd syllable vowel V_3 is. The latter is always short in the trisyllabic foot (trisyllables with long V_3 are bifoot words). Quantity contrasts in the 1st syllable are influenced by compensatory shortening only phonetically. The duration and length of V_2 and C_3 are affected both phonetically and phonologically.

The concept of ternary rhythm is contradictory in modern formal phonology. Some deny its relevance altogether and re-analyse known cases of trisyllabic feet through regular di- and monosyllabic foot types [19]. Others propose a recursive (two-layered) foot, where a "minimal" disyllabic foot is embedded into a "maximal" disyllabic foot [20], [21].

Our phonological and phonetic data on Soikkola Ingrian rather suggest that the trisyllabic foot is a coherent metrical whole in this variety. It is a type of a rhythmic group on its own, different from both the disyllabic foot and the bifoot structure. Especially our present data on C_3 geminate sonorants, which are at the boundary of the 2nd and the 3rd syllable but are still compensatorily affected by the structure of preceding syllables, might pose challenges for the above-mentioned formal accounts on ternary rhythm.

Factor	C_1	V_1	R	C_2	V_2	C_3	V_3
Original length contrast	–	binary	–	ternary	binary	binary	–
Contrast preservation		+		+	–	-/+	
Compensatory impact of the first syllable elements (V_1, R, C_2)	–	phone-tic	phone-tic	phone-tic	phono-logical, phonetic	phonological, phonetic	–
Affecting compensatory factors		C_2 length (* V_1 length)	V_1 length, C_2 length	V_1 length (* C_2 length), R	V_1 length * R * C_2 length	***m**: — C_2 length; ***l**:-**l** — V_1 length * R * C_2 length	

Table 1: Prosodic structure of the Soikkola Ingrian trisyllabic foot, as observed in [1] and in this study.

7. REFERENCES

[1] N. Kuznetsova, I. Brodskaya, and E. Markus, "Compensatory effects of foot structure in segmental durations of Soikkola Ingrian di- and trisyllables," *Journal of Phonetics*, Accepted.

[2] E. Markus, P. Lippus, K. Pajusalu, and P. Teras, "Three-way opposition of consonant quantity in Finnic and Saamic languages," in *Nordic prosody: Proceedings of the XIth Conference, Tartu 2012*, E. L. Asu-Garcia and P. Lippus, Eds., Frankfurt am Main: Peter Lang Edition, 2013, pp. 225–234. doi: 10.3726/978-3-653-03047-1.

[3] N. Kuznetsova, "Two phonological rarities in Ingrian dialects," in *New trends in Nordic and general linguistics*, M. Hilpert, J.-O. Östman, C. Mertzlufft, M. Riessler, and J. Duke, Eds., in Linguae & litterae, no. 42. Berlin; Boston: De Gruyter, 2015, pp. 91–117. doi: 10.1515/9783110346978.91.

[4] N. Kuznetsova, "Typology, chronology, and phonetic mechanisms of Finnic secondary gemination in the light of Soikkola Ingrian acoustic data," in *Proceedings of the 13th International Conference of Nordic Prosody: Phonetics, phonology, and technology for and of Scandinavian languages*, O. Niebuhr and M. Svensson Lundmark, Eds., University of Southern Denmark, Center for Industrial Electronics: Sciendo, Submitted. doi: https://doi.org/10.0000/9788366675728.

[5] R. E. Nirvi, *Inkeroismurteiden sanakirja [The dictionary of Ingrian dialects]*. in Lexica Societatis Fenno-Ugricae, no. 18. Helsinki: Suomalais-Ugrilainen Seura, 1971.

[6] M. Rapola, *Suomen kielen äännehistorian luennot [Lectures on the history of sounds of the Finnish language]*. in Suomalaisen kirjallisuuden seuran toimituksia, no. 283. Helsinki: Suomalaisen kirjallisuuden seura, 1966.

[7] M. Leppik, "On the non-phonological character of consonant gradation in Proto-Finnic," *Soviet Fenno-Ugric Studies*, vol. IV, pp. 1–12, 1968.

[8] V. Porkka, *Ueber den Ingrischen Dialekt: Mit Berücksichtigung der übrigen finnisch-ingermanländischen Dialekte [On the Ingrian dialect: With consideration of other Finnic dialects of Ingria]*. Helsingfors: J. C. Frenckell & Sohn, 1885.

[9] A. Laanest, *Ižorskije dialekty: Lingvogeografičeskoje issledovanije [Ingian dialects: Lingvo-geographic study]*. Tallin: Akademija nauk Estonskoj SSR [Academy of Sciences of the Estonian Republic], 1966.

[10] N. V. Kuznecova, "Fonologičeskije sistemy ižorskih dialektov [Phonological systems of Ingrian dialects]," Diss. ...kand. filol. nauk [Candidate thesis], Institut lingvističeskih issledovanij RAN [Institute for Linguistic Studies, Russian Academy of Sciences], Sankt-Peterburg, 2009.

[11] I. Lehiste, "Isochrony reconsidered," *Journal of Phonetics*, vol. 5, no. 3, pp. 253–263, 1977, doi: 10.1016/S0095-4470(19)31139-8.

[12] J. Obleser and F. Eisner, "Pre-lexical abstraction of speech in the auditory cortex," *Trends in Cognitive Sciences*, vol. 13, no. 1, pp. 14–19, 2009, doi: 10.1016/j.tics.2008.09.005.

[13] A. Sovijärvi, *Foneettis-äännehistoriallinen tutkimus Soikkolan inkeroismurteesta [Phonetic and historical-phonological study on the Soikkola Ingrian dialect]*. in Suomi-kirjan eripainoksia, no. 5. Helsinki: Suomalaisen kirjallisuuden seura, 1944.

[14] J. Harrington, F. Kleber, U. Reubold, F. Schiel, and M. Stevens, "The phonetic basis of the origin and spread of sound change," in *The Routledge Handbook of Phonetics*, W. F. Katz and P. F. Assmann, Eds., Abingdon, Oxon; New York, NY: Routledge, 2019, pp. 401–426. doi: 10.4324/9780429056253-15.

[15] L. Hakulinen, "Pari suomen murteiden vokaalinkestoseikkaa [A couple of dialectal vowel duration phenomena]," *Virittäjä*, vol. 26, pp. 49–59, 1922.

[16] J. Lehtonen, *Aspects of quantity in standard Finnish*. in Studia philologica Jyväskyläensia, no. 6. Jyväskylä: University of Jyväskylä, 1970.

[17] K. Suomi and R. Ylitalo, "On durational correlates of word stress in Finnish," *Journal of Phonetics*, vol. 32, no. 1, pp. 35–63, 2004, doi: 10.1016/S0095-4470(03)00005-6.

[18] N. Kuznetsova, "Evolution of the non-initial vocalic length contrast across the Finnic varieties of Ingria and adjacent areas," *Linguistica Uralica*, vol. 52, no. 1, pp. 1–25, 2016, doi: 10.3176/lu.2016.1.01.

[19] C. Golston, "No stress system requires recursive feet," *Catalan Journal of Linguistics*, vol. 20, pp. 9–35, 2021, doi: 10.5565/rev/catjl.333.

[20] V. Martínez-Paricio and R. Kager, "The binary-to-ternary rhythmic continuum in stress typology: layered feet and non-intervention constraints," *Phonology*, vol. 32, no. 3, pp. 459–504, 2015, doi: 10.1017/S0952675715000287.

[21] V. Martínez-Paricio and R. Kager, "In favour of layered feet: a response to Golston," *Catalan Journal of Linguistics*, vol. 20, pp. 37–55, 2021, doi: 10.5565/rev/catjl.352.

[i] Author contributions: The design and the background of the study were developed by NK; the data were collected by NK, IB, and EM, segmented by IB with a help of EM, extracted and processed by IB. NK is responsible for the statistical analysis and visualisation of the data presented in this paper. The draft text was written by NK, revised and edited by NK with a help of EM. NK is the corresponding author and responsible for the definitive version of the text.

[ii] The full dataset, including the data used in this paper, will appear at https://osf.io/rbw3m/.

[iii] This type with a final V: was the only bifoot structure among our original 22 trisyllabic types, to cover the structure "V:RC" absent from the trisyllabic feet with C_2 stops; cf. Footnote (c) to Appendix 1 in [1].

Predicting sound change: acoustic evidence from a longitudinal study of /s/-retraction in English /str/

Mary Stevens

Institute for Phonetics and Speech Processing (IPS), LMU Munich
mes@phonetik.uni-muenchen.de

ABSTRACT

This study is concerned with the predictability of regular sound change at the population level and within individual grammars. Taking /s/-retraction in Australian English /str/ as a case study, it presents an acoustic analysis of longitudinal data from production experiments with eighteen speaker participants from a single speech community. Coarticulatory /s/-retraction was evident for all speaker participants in /str/ at the baseline. After eight intervening years, incremental shifts in the same direction (towards /ʃ/) were evident for most speakers. Comparison with the outcome of simulated interaction in an agent-based model suggests that regular sound change is predictable at the population level but not at the level of the individual speaker.

Keywords: sound change, longitudinal studies, individual variation, agent-based models, /s/-retraction

1. BACKGROUND

Population-level sound change can be examined by comparing speech produced by older and younger speakers of the same variety, who are understood to represent earlier and later stages, respectively, in the progression of a sound change (e.g., [2, 9, 12]). Such apparent time studies rest on the assumption that pronunciations are stable throughout adulthood – and can therefore be compared between generations. However, there is evidence from case studies that adult speakers' pronunciations can change incrementally in the direction of sound changes going on in the language spoken around them (e.g., [10]), as well as due to biological ageing (e.g., [15]) – thus potentially complicating comparisons between different age groups. There is also evidence that the tendency to shift in the direction of population-level sound changes varies between individuals. Longitudinal studies of sound changes in progress involving multiple speaker participants are rare, but [17] analysed Montreal French speakers who were recorded once in 1971 and again thirteen years later, during which time Montreal French was undergoing a sound change from [r] -> [R]. Based on the authors' auditory categorization, most speakers' rhotic pronunciation was found to have remained stable – only nine (28%) speakers were found to have significantly increased their use of innovative [R] in the later recordings, even though a sound change was in progress at the population level. The relative rarity of change within individual participants' lifespans may be due to the nature of the [r] -> [R] sound change in Montreal French, which has its source in dialect contact (rather than coarticulatory variation, as is instead the case for regular sound change).

The present study concentrates on the effects of regular sound change on speakers' production repertoires in adulthood. The sound change chosen for investigation was /s/-retraction in English /str/, whereby e.g., the sibilant in *street* comes to resemble that in *sheet*. /s/-retraction in /str/ has been shown to have its origins in coarticulation ([18]) and has been the subject of several phonetic investigations in various varieties of English (e.g., amongst others, [2, 3, 16, 18, 20]). This sound change has not yet been the subject of a longitudinal study with real speakers, but in [19] interaction was simulated amongst a group of speakers (agents) in an agent-based model (ABM). Before simulated interaction, the speakers' /str/ pronunciations showed coarticulatory /s/-retraction in /str/; afterwards, there was acoustic evidence of a population-level shift in the direction of /ʃ/. This result suggests that regular sound change is the result of the gradual accumulation of synchronic variation, in combination with density of communication [5]. However, the question remains as to whether the outcomes of simulated interaction match longitudinal changes within real speakers' production targets, i.e., whether regular sound change proceeds via incremental changes to individual grammars, even in adulthood (which would conflict with an assumption of the apparent time method, as noted above). The present study addresses this question with an acoustic analysis of sibilants in /str/ that were produced by the same speakers in two separate recording sessions, eight years apart. In section 4.2, the predictability of regular sound change is addressed by comparing our results concerning the effects of real-world interaction on speakers' /str/ pronunciation with those reported in [19], in which interaction was simulated between the same speakers (agents) in an agent-based model.

2. METHODS

2.1 Participants and recording setup

Participants were all adult monolingual speakers of Australian English and long-term residents of Braidwood, a semi-rural town in New South Wales. They were originally recruited for production experiments conducted in 2014. The present study is based on data for eighteen speakers who took part in a follow-up recording session in early 2022. Participants were paid for their participation in both recording sessions. The participants (thirteen female, five male) were aged between 29 and 49 years at the time of the first recordings. They were mostly known to each other and some reported interacting with each other daily. In terms of contact with non-local varieties, only one participant reported having spent more than one of the intervening eight years (in total) outside of the state of New South Wales.

2.2 Recording procedure and stimuli

All recording sessions took place in the same room in a private house. Participants wore a headset microphone and read words aloud as they appeared on a laptop computer screen. The words in Table 1 were presented to participants in isolation (no carrier phrase) in a randomized order, with ten repetitions in the first recording session and nine repetitions in the second. The segmental and prosodic contexts for the target sibilant were matched as closely as possible across the three (/s/ and /ʃ/, and /str/) word types. Due to a small number of speech errors and recording errors in both recording sessions, 7369 sibilant tokens from the first recording session and 6640 sibilant tokens from the second recording session (of a possible 7380 and 6642, respectively) were available for analysis.

/s/	*assault, assembly, fascinating, messy, Minnesota, motorcycle, policy, possible, sane, seem, soak*
/ʃ/	*dishevelled, information, machine, overshadowed, passionate, perishable, polishing, sheep, Shane, show, tissue*
/str/	*administrate, astringent, astronaut, astronomy, catastrophe, catastrophic, chemistry, claustrophobic, destroy, district, gastro, oestrogen, pastrami, pedestrian, orchestra, stream, restrict, strong, strut*

Table 1: Target words containing prevocalic /s/ or /ʃ/, or a /str/ cluster.

2.3 Acoustic analysis

The signal data from each recording session were labelled semi-automatically using WebMAUS (e.g., [11]) and converted into two EMU databases ([21]). Segment boundaries were corrected manually. The first spectral moment (M1), which indexes place of articulation, was calculated as a trajectory between the temporal onset and offset of each sibilant token. M1 trajectories were high pass filtered to remove all spectral information below 500 Hz and were z-score transformed (across all data frames for all sibilants, separately by speaker and recording session) using the speaker-normalization method in [13]. Rather than tracking the absolute location of sibilants in acoustic space using e.g., the M1 trajectories, /s/-retraction was measured relatively, consistent with other recent studies ([2, 19]). More specifically, it was measured in terms of the *orthogonal projection* (OP), which tracks the relative location of any sibilant token between the same speaker's /s/-words (centred at +1) and /ʃ/-words (centred at –1). The OP can be interpreted similarly to M1 or spectral centre of gravity, i.e., higher/lower OP indicates a more /s/- or /ʃ/-like pronunciation, respectively. To obtain the OP for each sibilant token, the speaker normalized M1 trajectories were parameterised into three cepstral coefficients using Discrete Cosine Transformation (DCT) (e.g. [7]), corresponding to their height, slope, and curvature. Each sibilant token's position in this three-dimensional DCT space on a line passing through speaker-specific centroids for /s/- and /ʃ/-words, namely the OP, was then calculated. Note that the OP was calculated within each recording session separately, that is, the position of each sibilant token was calculated relative to centroids for /s/- and /ʃ/-words produced during the same recording session (and by the same speaker).

3. RESULTS

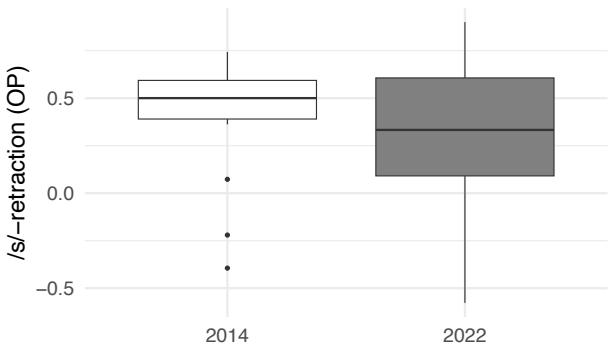


Figure 1: OP for sibilants in /str/, by recording session. Each data point depicts one speaker's mean OP in all /str/-words and repetitions. Lowering of OP in 2022 indicates increased /s/-retraction.

Figure 1 shows the mean OP for sibilants in /str/ in the two recording sessions. Lowering of OP in 2022 (dark grey) indicates increased acoustic proximity to /ʃ/. A mixed effects model was run to test the effect of the intervening years on sibilants in /str/ with OP as the dependent variable, Recording session (two levels, 2014 or 2022) and Gender (two levels: male or female) as fixed factors, and Word (41 levels) and Speaker (18 levels) as random factors. The results showed a significant effect for Recording session (Chi Square = 4.8, $p < 0.05$) which confirmed that OP was significantly lower (= increased acoustically measurable /s/-retraction) in 2022. Neither the fixed factor Gender nor its interaction with Recording session was found to have a significant effect on OP. It is evident in Figure 1 that there was an increase in the interquartile range from 2014 to 2022, which suggests divergence between speaker participants over time. To examine individual speakers' behaviour more closely, the data are shown separately by speaker in Figure 2.

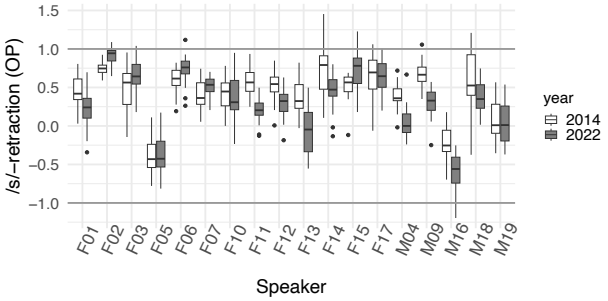


Figure 2: The same data from Figure 1, by speaker. The two thick horizontal lines indicate the OP for sibilants in /s/-words and /ʃ/-words, centred at 1 and -1, respectively.

First, all participants' OP for sibilants in /str/ was lower than that for pre-vocalic /s/, which was centred at 1 (cf. section 2.3) and is indicated by the upper horizontal line in Figure 2. Second, most – but not all – speakers' median OP showed a further downwards shift towards /ʃ/ (indicated by the lower horizontal line) in 2022. This shift towards /ʃ/ is consistent with the group-level pattern seen in Figure 1. In other words, all speakers showed acoustically measurable /s/-retraction in /str/ in 2014 and most speakers showed an increase in the degree of /s/-retraction over time. However, upward shifts in OP (i.e., increased proximity to /s/) can be seen for five speakers (F02, F03, F06, F07, F15) and two further speakers' (F05, M19) OP in /str/ remained almost identical over time. Thus, it appears that some individual speakers did not participate in the sound change evidently underway at the population level (Fig. 1), i.e., in the language spoken around them.

Note that the five male speakers (M04-M19, on the right in Fig. 2) all showed a downwards shift from 2014 to 2022, whereas female speakers (F01-F17) differed in terms of whether they showed an increase or a decrease in the degree of /s/-retraction over time. This dataset was not balanced for speaker gender, with thirteen females but only five males, and it is unclear whether resistance to shifts in the direction of /s/-retraction might be specific to female speakers or simply an artefact of our unbalanced dataset with too few male participants. In terms of the literature on /s/-retraction in English /str/, patterns differ between varieties and it is not clear whether males or females typically lead the change (see e.g., [20]). The tendency for some individuals to resist shifting in the direction of the population-level sound change is discussed further in section 4.

Some phonetic models of sound change propose that participation in a sound change involves a categorical change to an individual's phonetic repertoire, replacing a conservative variant with an innovative variant (e.g., [3, 14]). Patterns in Fig. 2 show that distributions for /str/ in the two recording sessions mostly overlapped and there was no evidence of a speaker making a categorical change from a conservative pronunciation (in this case very /s/-like) in 2014 to an innovative pronunciation (in this case very /ʃ/-like) in 2022. Thus /s/-retraction in this variety appears to involve incremental changes to speakers' pronunciation targets, in line with the idea that sound change involves a gradual shift from one gestural coordination pattern to another (e.g., [6]) and with usage-based models (e.g., [4]) in which change is incremental as language users' representations are updated over time.

It has also been suggested [3] that the tendency to shift in the direction of /s/-retraction might be related to the degree to which a speaker coarticulates in their own speech production. However, in the present dataset, speakers with relatively similar degrees of /s/-retraction in 2014, such as F01 and F07, did not behave similarly over time: a downwards shift was evident for speaker F01, for example, while speaker F07 shifted upwards; see also the speaker pair F07 and F13, or the speaker pair F05 and M16. Thus, there is no evidence in Figure 2 that the tendency for an individual to participate in the sound change involving /s/-retraction depended on the degree of coarticulation in their own speech production.

4. DISCUSSION

4.1 Longitudinal change in real speakers' /str/

This study explored whether a coarticulatory tendency towards /s/-retraction that was common to speaker participants in 2014 increased over a period

of eight intervening years. Acoustic results showed evidence of a population-level shift in the direction of /s/-retraction which was also evident within most individual speakers. The relatively high proportion of speakers whose /str/ pronunciation shifted with the (shifting) population-level norm differs from [17] in which only a minority of speakers were found to have significantly increased their use of innovative [R] over an intervening 13-year period. Indeed Sankoff [17, pp. 576)] concluded that the sound change [r] -> [R] in Montreal French is *not* driven by "speakers slowly increasing their use of [R] across their lifetimes", but rather through new speakers (with more innovative pronunciations) being added to the population. These differences in terms of the proportion of individuals whose production repertoires were found to have changed might best be attributed to the different sound changes under investigation: regular sound change with its origins in coarticulatory variation (here) vs. sound change due to dialect contact [17].

In the present study, the five speakers whose /str/ pronunciation shifted in the opposite direction to the population-level shift were all female, but a higher number of male participants would be necessary to determine whether participation in this sound change is conditioned by speaker gender (and see e.g. [20]).

In order to better understand the factors that shape population-level sound change, future studies might also explore why real speakers differed in terms of their longitudinal production behaviour. For example, resistance to shift in the direction of population-level sound change may be governed by awareness of the sound change (e.g., [1]), the ability to detect /s/-retraction in perception or to control one's own speech production.

4.2 Effects of real vs. simulated interaction

Here we compare our longitudinal data for real speakers with the outcomes of simulated interaction in an agent-based model (ABM) [19]. The input data to the ABM were the parameterized M1 trajectories (cf. section 2.3) from the 2014 recording session, with additional data from one male speaker who was not available for the follow-up recordings in 2022. Each speaker was represented in the ABM as an agent, with their own sibilant production data stored in memory. The ABM involved 60,000 speaker-listener interactions and was repeated 100 times; here we consider the outcome of a randomly chosen simulation run. Figure 3 shows the mean OP for sibilants in /str/ in 2014 (on the left), after eight years of real-world interaction (2022, in the middle) and after simulated interaction (ABM, on the right).

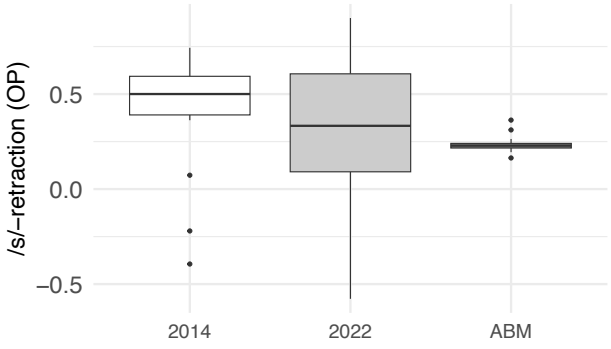


Figure 3: The same data as in Figure 1, with the addition of the right-hand box showing the outcome of simulated interaction in an agent-based model (ABM). Lowering of OP indicates increased /s/-retraction. 19 speaker-agents interacted in the ABM, data for 18 of these are shown at the baseline (2014) and after eight years (2022).

It is evident from Figure 3 that both simulated (ABM) and real-world interaction (2022) resulted in a downwards shift for sibilants in /str/, compared to the baseline (2014). Thus, the ABM accurately predicted the direction of the population-level shift that took place from 2014 to 2022. This pattern is consistent with results reported in [8], which also compared the effects of simulated versus real-world interaction. In that study, population-level vowel shifts resulting from interaction amongst an isolated group of speakers were – to a certain extent – accurately predicted by the ABM. However, most of the changes observed in [8] were found to be due to the influence of a small number of speakers (with different accents). Here, on the other hand, we have seen that tendencies that are common amongst speakers (like coarticulatory /s/-retraction) can accumulate via interaction – real and simulated, causing population-level shifts. Compared to the baseline (i.e., 2014), the interquartile range increased in the real world (2022) but decreased in the ABM. This indicates that there was divergence between real speakers but convergence between speaker-agents in the ABM (each data point in Fig. 3 shows the mean OP for one speaker/agent's sibilants in /str/). In other words, the ABM accurately predicted the direction of the population-level change that occurred between 2014 and 2022, but not the changes that took place within individual grammars. The evident convergence between individuals in the ABM is primarily due to the structure of the model: agents interacted only with each other and exchanged only the lexical items in Table 1, whereas real interactions would have involved a larger number of interlocutors and lexical items. As noted in section 4.1, a fruitful approach for future research might be to explore the reasons why some real speakers' /str/ pronunciations did not shift with the (shifting) population-level norm.

5. ACKNOWLEDGMENTS

This research was supported by German Research Council Grant No. 387170477.

6. REFERENCES

1 Babel, A. 2016. *Awareness and control in sociolinguistic research.* Cambridge, Cambridge University Press.

2 Bailey, G., Nicholls, S., Turton, D., Baranowski, M. 2022. Affrication as the cause of /s/-retraction: Evidence from Manchester English. *Glossa: a journal of general linguistics* 7, (1): 1–30.

3 Baker, A., Archangeli, D., and Mielke, J. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change* 23: 347–374.

4 Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. 2009. Language is a complex adaptive system: position paper. *Language learning* 59, (s1): 1–26.

5 Bloomfield, L. 1933. *Language.* George Allen & Unwin.

6 Browman, C. and Goldstein, L. 1992. Gestural structures and phonological patterns. In Mattingly, I.G., and Studdert-Kennedy, M. (Eds.) *Modularity and the Motor Theory of Speech Perception*. Erlbaum.

7 Harrington, J. 2010. *The phonetic analysis of speech corpora*. Wiley-Blackwell.

8 Harrington, J., Gubian, M., Stevens, M., and Schiel, F. 2019. Phonetic change in an Antarctic winter. *Journal of the Acoustical Society of America* 146: 3327–3332.

9 Harrington, J., Kleber, F., and Reubold, U. 2008. Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *Journal of the Acoustical Society of America* 123, (5): 2825–2835.

10 Harrington, J., Palethorpe, S., and Watson, C.I. 2000. Does the Queen speak the Queen's English? *Nature* 408: 927.

11 Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer, Speech and Language* 45, 326–347, doi:10.1016/j.csl.2017.01.005.

12 Kuang, J., and Cui, A. 2018. Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics* 71. 194-214.

13 Lobanov, B.M. 1971. Classification of Russian Vowels Spoken by Different Speakers. *Journal of the Acoustical Society of America* 49: 606.

14 Ohala, J. 1993. The phonetics of sound change. In Jones, C. (Ed.) *Historical Linguistics: Problems and Perspectives.* Longman: 237-278.

15 Reubold, U., Harrington, J. 2017. The influence of age on estimating sound change acoustically from longitudinal data. In Wagner, S.E. and Buchstaller, I. (Eds.) *Panel Studies of Language Variation and Change*. Routledge. 129-152.

16 Rutter, B. 2011. Acoustic analysis of a sound change in progress: The consonant cluster /stɹ/ in English, *Journal of the International Phonetic Association* 41. 27 – 40.

17 Sankoff, G., and Blondeau, H. 2007 'Language change across the lifespan: /r/ in Montreal French'. Language 83, 3. 560-588.

18 Stevens, M., and Harrington, J. 2016. The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English. *Journal of Phonetics* 58. 118-134.

19 Stevens, M., Harrington, J., and Schiel, F. 2019. Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa: a journal of general linguistics* 4 (1). 1–30.

20 Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., and Thomas, E. 2019 Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction'. *Proc. ICPhS 2019*, Melbourne.

21 Winkelmann, R., Harrington, J., and Jänsch, K. 2017. EMU-SDMS: advanced speech database management and analysis in R. Computer Speech & Language 45. 392-410.

INDIVIDUALS' PERCEPTUAL RETUNING PREDICTS ARTICULATORY ACCOMMODATION: IMPLICATIONS FOR SOUND CHANGE

Patrice Speeter Beddor[1], Andries Coetzee[1], Ian Calloway[2], Stephen Tobin[3] and Ruaridh Purse[1]

[1]University of Michigan, MI, USA, [2]ALEX-Alternative Experts, VA, USA, [3]Universität Potsdam, Germany
beddor@umich.edu, coetzee@umich.edu, icalloway@alexinc.com, stephen.tobin@uni-potsdam.de, rupurse@umich.edu

ABSTRACT

Individuals differ from each other, as speakers and listeners, in the extent to which they adapt to unfamiliar speech patterns. The hypothesis that listener-specific perceptual adjustments for an unfamiliar pattern are reflected in the listener-turned-speaker's imitation of the pattern was tested for raising of English /æ/ before /g/ (e.g., [beɪg] *bag* but [bæk] *back*). For 37 American English participants, perceptual learning and spontaneous imitation of raised /æ(g)/ were assessed using eye-tracking and ultrasound imaging, respectively. Results support the hypothesis that perceptual retuning predicts, in part, articulatory accommodation: the more an individual perceptually adapted to raised /æ(g)/ (e.g., used [beɪg] to rapidly disambiguate *back-bag* trials), the more that individual imitated raised /æ(g)/. These findings are viewed as relevant to the spread of (coarticulatorily motivated) change in that community members who attend particularly closely to innovative interlocutors' novel forms may be especially likely to converge towards those forms in their subsequent productions.

Keywords: sound change, imitation, perceptual learning, production-perception link

1. INTRODUCTION

Speakers' productions and listeners' percepts are malleable and dynamic. For example, speakers modify their speech in ways that converge towards the patterns of an interlocutor [18] or a model talker [10, 22]. Listeners also adapt; their perceptual decisions show sensitivity to community speech norms [12] and to the patterns of a specific idiolect [16]. Yet despite malleability, adaptation is highly variable across individuals: speakers differ in how much and how accurately they imitate [14] and listeners differ in the extent to which they learn a new phonetic contrast or native-like perceptual weights [9, 20]. These individual differences have been found to be reliable, with individual differences in degree of convergence towards a talker, for example, being stable across test sessions [25].

This study tests the hypothesis that an individual's perceptual adjustments for a novel, coarticulatory speech pattern will be reflected in their imitation of that pattern, that is, that perceptual retuning predicts, in part, articulatory accommodation. Theoretically, our approach is motivated by our interest in testing accommodation in the laboratory as a vehicle for understanding the spread of coarticulatorily motivated change in a speech community. In particular, we aim to better understand the phonetic behaviors of early adopters of an incipient change in a speech community—community members who, arguably, are especially likely to converge towards an innovative interlocutor's novel forms. Empirically, the expectation of a relation between an individual's production and perception of coarticulated speech is consistent with recent evidence that individuals who produce more extensive coarticulation (e.g., vowel nasalization) are also more efficient users of that information [3, 4, 27, 28]. Yet not all studies find a comparable production-perception relation [11, 20]. Moreover, when perception and production involve adaptation to an unfamiliar pattern, there are social, attitudinal, and other factors [1, 2] that might mitigate against the hypothesized link.

The targeted novel pattern is raising of /æ/ towards [eɪ] before /g/ (but not /k/; as in [beɪg] *bag* vs. [bæk] *back*), a vowel shift found for speakers in parts of the Northern U.S. and Canadian Prairies [21]. Mielke et al.'s [17] ultrasound study of English speakers who do and do not produce raised /æ/ before /g/ (henceforth, raised /æ(g)/) showed that, even for non-raisers, the tongue root was more advanced and tongue body more fronted for /g/ than /k/ following /æ/, consistent with enlarging the supralaryngeal cavity to sustain voicing during /g/. This more fronted dorsal constriction for /g/ presumably underlies, or has at least contributed to, /æ(g)/ raising.

In this study, participants unfamiliar with this pattern were exposed, using a visual world task, to a model talker's raised /æ(g)/ (and unraised /æ(k)/) (see [5, 23] for a similar approach). Participants' spontaneous imitation of that pattern was assessed in a subsequent ultrasound session. This approach allowed us to study the time course of participants' reorganization of their perceptual and articulatory spaces as they learn an unfamiliar phonetic variant. Our main interest is in the *relation* between perceptual and articulatory learning and reorganization for individual listener-speakers.

2. PERCEPTUAL ADAPTATION

The perception task assesses whether, over the course of an eye-tracking experiment, learning novel raised /æ(g)/ facilitates participants' recognition of (unraised) /æk/-final words but slows recognition of (phonetically similar) /eɪk/-final words.

2.1. Methods

Auditory stimuli for the audio-visual trials were 10 sets of minimal C(C)VC triplets where V = /æ/ or /eɪ/ and final C = /g/ or /k/ (e.g., *back-bag-bake*) produced by a model talker with native /æ(g)/ raising. This talker's diphthongal /æ(g)/ had an F1/F2 trajectory similar to that of /eɪ(k)/ but with a higher F1 frequency. Because /æ/ was longer before /g/ than /k/, minor stimulus editing was done to reduce the average /æ(g)/-/æ(k)/ difference to about 35 ms (and thus to reduce duration as information for the voicing contrast). Visual stimuli were black and white line drawings corresponding to each word. Each audio-visual stimulus for a trial consisted of one auditory stimulus and two side-by-side visual images, one target (corresponding to the auditory stimulus) and one competitor. Visual targets and competitors were all six possible combinations of stimuli for a triplet (e.g., *back-bag*, *back-bake*, *bake-bag*, *bake-back*, *bag-back*, *bag-bake*). Auditory stimuli were presented in a blocked design in which words for the first 30 trials ended only in /æk/ or /eɪk/. Following this pre-exposure block was a 150-trial exposure block containing all target words, including /æg/-final words. Across the full experiment, each audio-visual pairing occurred three times.

Participants were 37 listeners whose variety of American English does not exhibit /æ(g)/ raising. For each trial, participants heard the recorded instruction "Look at the pictures;" 2 s later a fixation cross appeared on the computer screen and participants heard "Fixate cross. (pause) Now look at [target word])." Eye movements over the course of each trial were monitored using an EyeLink 1000 Plus (SR Research) remote eye-tracker.

2.2. Predictions

Trials with target /Cæg/ were learning trials; our main predictions are for trials with competitor /Cæg/, which test effects of that exposure. For target *back* | competitor *bag*-type trials, participants should fixate on *back* earlier during the exposure block than during pre-exposure because /æ(g)/ raising should reduce *back-bag* competition. The reverse pattern (later target fixation) should hold for target *bake* | competitor *bag*-type trials because of increased *bake-bag* competition due to raised /æ(g)/. A further prediction is that individual listeners will differ in whether, and the extent to which, the novel pattern influences their lexical decisions.

2.3. Results

Analyses of participants' fixations were conducted using Generalized Additive Mixed Modeling in R [19], as implemented in the *mgcv* [26] and *itsadug* [24] packages. Target fixation proportions from 200-1200 ms after the target vowel onset were extracted for analysis, allowing 200 ms to plan and launch a saccade [15]. The fixation proportions (binary responses) were computed with a logit link function with the following model structure: experimental block (pre-exposure / exposure), stimulus competitor type, and elapsed time over the course of a trial (as well as their interactions) were modeled as fixed effects; by-participant random smooths for elapsed time were also included.

The modeled fixation results in Figure 1, left panel, show that, when the target was /Cæk/, that is, when raised /æ(g)/ provided early disambiguating information during the vowel portion of the trial (*back-bag*), listeners fixated earlier and more often on the target during the exposure block than during pre-exposure. However, when the target was /Ceɪk/ (Figure 1, right), that is, when /æ(g)/ raising increased ambiguity (*bake-bag*), fixations were later and less frequent during the exposure block. Thus, aggregate results support predictions (see also [23]).

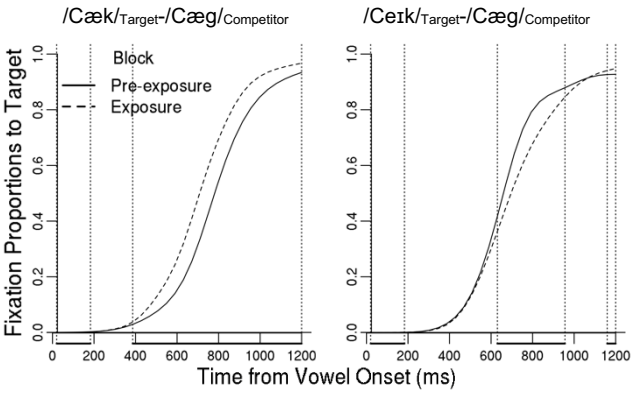


Figure 1: Perception: model-derived fixations, according to block, on /Cæk/ (left) and /Ceɪk/ (right) targets when competitor = /Cæg/. (Dotted lines: region of significant difference between blocks)

To quantify *individual* differences in perceptual retuning, we calculated, for each listener, the average difference between exposure and pre-exposure blocks in time to first target fixation for trials with /Cæg/ as the visual competitor. In this paper, we focus on target /Cæk/ | competitor /Cæg/ trials where, the larger the difference score, the greater the reduction in lexical competition for that listener—and, arguably, the

greater the learning. The median difference score was 43 ms, with 30% of participants' scores falling within ±20 ms of the median. However, scores ranged from -46 ms to 203 ms, indicating that some participants' looks to target were substantially faster after hearing raised /æ(g)/ whereas those of others slowed somewhat. In subsequent models (section 3), each participant's "difference to first target fixation" score serves as the *perception score* used to test our hypothesis that perceptual adjustments for a novel speech pattern are predictive of individuals' imitation of that pattern.

3. ARTICULATORY ACCOMMODATION

3.1. Methods

The same 37 participants returned about 1-2 weeks later for the production study. Auditory and visual stimuli were those used in the perception study. Production trials were of two types and used an alternating talker-participant naming task (see [6]). In "talker" trials, a talker icon appeared with the stimulus image, indicating that the participant should simply listen to the image name (e.g., *Say bag*) produced by the model talker. In "participant" trials, a microphone icon appeared with the stimulus image, indicating that the participant should name the image.

Table 1 gives the structure of the experiment. In blocks 1 (pre-exposure) and 3 (post-exposure), the participant produced words of all three types but did not hear the talker say /Cæg/ words. In block 2 (exposure), both participant and talker produced all word types. To maximize imitation, in all blocks half of the trials were repetition trials in that the word the participant was prompted to produce was the same as the immediately preceding word spoken by the talker.

	/Cæg/		/Cæk/		/Ceɪk/	
Block	T	P	T	P	T	P
1 (pre-exposure)	-	40	40	20	40	20
2 (exposure)	60	100	50	30	50	30
3 (post-exposure)	-	40	40	20	40	20

Table 1: Number of each trial type produced by the model talker (T) and participant (P) in the imitation study, according to block.

Tongue contour data were collected using a Zonare z.one ultrasound system with a p4-1c phased-array transducer. The probe was held in place by a custom-made, lightweight ultrasound stabilizer [7].

3.2. Results

Ultrasound sequences were submitted to MTracker [29] for automated tongue contour extraction. Vowel raising (i.e., imitation) is assessed by comparing, across blocks, the frame-wise maximum y-coordinate (normalized within participants) of participants' tongue contours from vowel onset to offset. These y-coordinate values were submitted to a GAMM in R again using the *mgcv* [26] and *itsadug* [24] packages. To investigate our hypothesis that perceptual and articulatory accommodation are linked at the level of the individual, the model included as fixed effects the Perception Scores (difference in first fixation for /Cæk/-/Cæg/ eye-tracking trials; section 2.3), the interaction factor of Rhyme (/æg/, /æk/, /eɪk/) x Block, and smooths for Normalized Time x Perception Score for each level of Rhyme x Block. Random smooths over Normalized Time for each level of Participant x Rhyme x Triplet were also included. Only /æg/ results are reported here.

Figure 2, left panel, gives the modeled aggregate results for maximum normalized tongue height for /æ(g)/ across the time course of the vowel for the three blocks. As would be expected for pre-velar /æ/, the tongue dorsum is highest at vowel offset (at onset of /g/ closure). However, across much of the duration of the vowel, tongue body height raises with increasing experience with the model talker's raised /æ(g)/ (from blocks 1 to 2 and 3).

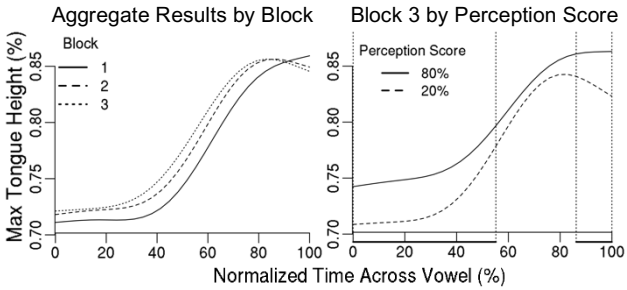


Figure 2: Production: model-predicted maximum tongue height. Left: aggregate results according to block. Right: block 3 (post-exposure) results for participants who adapted more (80th percentile) and less (20th) in perception.

Figure 2, right, assesses whether individuals' perceptual retuning predicts imitation and gives, for the post-exposure block, modeled /æ(g)/ tongue height values for individuals who fall near the 80th and 20th percentiles of the perception scores. Consistent with a retuning-imitation link, maximum tongue height was higher across the initial half of the vowel for individuals who adapted more in perception (80th percentile) than those who adapted less (20th).

Although Figure 2 shows that stronger perceptual adapters had overall higher /æ(g)/ realizations at the end of the task, further inspection of imitation results by block reveals a more nuanced pattern. Figure 3 shows that perceivers who adapted more (top panels) imitated quickly (from blocks 1 to 2, left) and then

plateaued (blocks 2 to 3, right) whereas those who adapted less (bottom) showed increasing tongue height across the task.

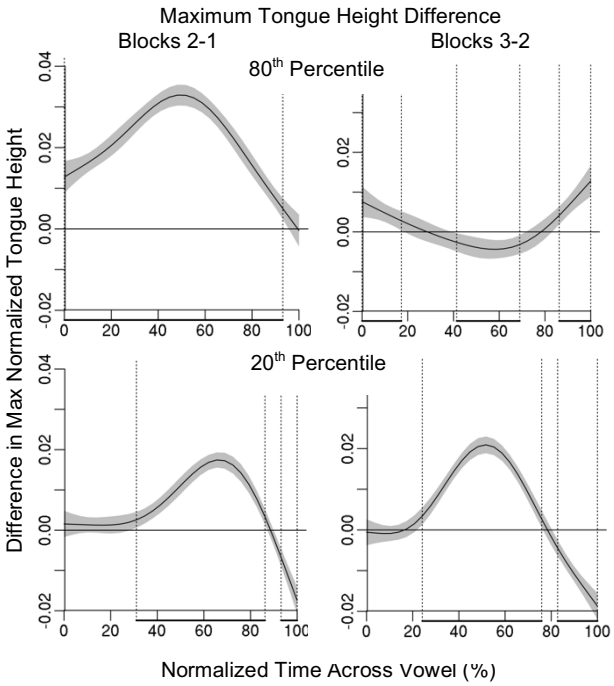


Figure 3: Model-predicted differences in maximum tongue height between Blocks 2-1 (left panels) and Blocks 3-2 (right panels) for participants whose perception scores fell near the 80th (top) and 20th (bottom) percentiles.

4. DISCUSSION

This study investigated American English-speaking participants' perceptual learning of, and articulatory accommodation to, an unfamiliar, coarticulatorily motivated pattern of raised /æ/ before voiced velar /g/. Consistent with previous work [5, 23], listeners exposed to this pattern over the course of an eye-tracking task showed, on average, faster recognition of minimal pair competitors ending in /æk/ (e.g., *back* vs. *bag*) but slower recognition of minimal pair competitors ending in /eɪk/ (*bake* vs. *bag*), indicating that, as listeners learned more information about the model talker's vowel patterns, they perceptually recalibrated in ways that influenced their lexical decisions in real time. These same participants also accommodated to this pattern as speakers, with ultrasound imaging data showing that average maximum tongue body height for /æ(g)/ was higher after participants had been exposed, in a "take turn" spontaneous imitation task, to the same model talker's raised /æ(g)/.

Despite these clear aggregate patterns, individual participants differed from each other, both as listeners and speakers, in degree of accommodation to the model talker. (Responses to a post-test survey provide no evidence that these individual differences in accommodation can be attributed to differences in participants' prior familiarity with the novel pattern.) We hypothesized that the perceptual and articulatory differences are linked: that an individual's degree of perceptual retuning for a novel pattern would predict, to some extent, their imitation of that pattern. In support of this hypothesis, the results of a production model in which each participant's degree of perceptual learning (for target /Cæk/ | competitor /Cæg/ trials) was included as a predictor for imitation showed that, the more an individual perceptually adapted to raised /æ(g)/ (i.e., used raised /æ(g)/ to rapidly disambiguate *back*-*bag*-type trials), the more that individual imitated raised /æ(g)/. Indeed, these stronger perceptual adapters were both more accurate imitators (more closely approximating the model talker's yet more raised /æ(g)/) and more rapid imitators, exhibiting especially large imitative tongue height positions during the exposure block.

That the degree of an individual's perceptual accommodation to a novel pattern is linked to their articulatory accommodation may seem unsurprising from the perspective that articulatory accommodation has perceptual and articulatory components: the novel pattern must be a sufficiently salient perceptible difference to be imitated. These data, though, indicate that the link is tighter than "sufficiently salient" and rather show that perceptual *retuning* is predictive of imitation.

We conclude by considering the implications of these findings for theories of the spread of sound change. To illustrate, we take the incipient change of spread of raised /æ(g)/ through a speech community. Mielke et al.'s [17] data show that even speakers from communities without this pattern tend to produce fronted /g/ relative to /k/ constrictions. Thus, English-speaking communities will tend to have non-random directional /g/ ~ /k/ variation (with /g/ having a more front articulation) and, at least in some cases, co-occurring directional /æ/ variation (higher tongue body before /g/ than before /k/). A challenge for sound change researchers is to explain why the more innovative forms of this systematic variation (i.e., especially fronted /g/ and raised /æ(g)/) might become the new phonetic norm in a speech community (e.g., [8], [13]). By hypothesis, our data shed some light on the phonetic behaviors of early adopters of these more innovative raised /æ(g)/ variants—and, more generally, of early adopters of innovative coarticulatorily motivated phonetic forms. The findings raise the possibility that these early adopting listeners-turned-speakers are perhaps especially attentive to the new variants, adjust their perceptual space (or repertoire) accordingly, and mirror those adjustments in their own, now converging productions.

5. ACKNOWLEDGMENTS

This material is based on work supported by NSF Grant BCS-1348150 to Patrice Beddor and Andries Coetzee; any opinions, findings, and conclusions are the authors' and do not necessarily reflect the views of the NSF.

6. REFERENCES

[1] Babel, M. 2010. Dialect divergence and convergence in New Zealand English. *Language in Society* 39, 437-456.

[2] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40, 177-189.

[3] Beddor, P.S., Coetzee, A. W., Boland, J. E., McGowan, K. B., Styler, W. 2018. The time course of individuals' perception of coarticulatory information is linked to their production: implications for sound change. *Language* 94, 1-38.

[4] Coetzee, A. W., Beddor, P. S., Styler, W., Tobin, S., Bekker, I., Wissing, D. 2022. Producing and perceiving socially structured coarticulation: coarticulatory nasalization in Afrikaans. *Laboratory Phonology* 13, 1-43.

[5] Dahan, D., Drucker, S. J., Scarborough, R. A. 2008. Talker adaptation in speech perception: adjusting the signal or the representations? *Cognition* 108, 710-718.

[6] Delvaux, V., Soquet, A. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica* 64, 145-173.

[7] Derrick, D., Carignan, C., Chen, W. R., Shujau, M., Best, C. T. 2018. Three-dimensional printable ultrasound transducer stabilization system. *J. Acoust. Soc. Am.* 144, EL392-EL398.

[8] Garrett, A., Johnson, K. 2013. Phonetic bias in sound change. In: Yu, A. C. L. (ed), *Origins of Sound Change: Approaches to Phonologization*. Oxford University Press, 51-97.

[9] Golestani, N., Zatorre, R. J. 2004. Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494-506.

[10] Goldinger, S. D. 1998. Echo of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251-279.

[11] Grosvald, M. 2009. Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics* 37, 173-188.

[12] Hay, J., Warren, P. Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34, 458-484.

[13] Harrington, J., Kleber, F., Reubold, U., Schiel, F., Stevens, M. 2018. Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in Cognitive Science* 10, 707-728.

[14] Honorof, D. N., Weihing, J., Fowler, C. A. 2011. Articulatory events are imitated under rapid shadowing. *Journal of Phonetics* 39, 18-38.

[15] Hutton, S. B. 2008. Cognitive control of saccadic eye movements. *Brain and Cognition* 68, 327-340.

[16] Kraljic, T., Brennan, S. E., Samuel, A. G. 2008. Accommodating variation: dialects, idiolects, and speech processing. *Cognition* 107, 54-81.

[17] Mielke, J., Carignan, C., Thomas, E. R. 2017. The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: an ultrasound study. *J. Acoust. Soc. Am.* 142, 332-349.

[18] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* 119, 2382-2393.

[19] R Core Team. 2021. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http:/www.R-project.org/

[20] Schertz, J, Cho, T., Lotto, A., Warner, N. 2015. Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics* 52, 183-204.

[21] Stanley, J. A. 2022. Regional patterns in prevelar raising. *American Speech* 97, 374-411.

[22] Tobin, S. J. 2022. Effects of native language and habituation in phonetic accommodation. *Journal of Phonetics* 93, 101148.

[23] Trude, A. M., Brown-Schmidt, S. 2012. Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes* 27, 979-1001.

[24] van Rij, J., Wieling, M., Baayen, R. H., van Rijn, H. 2020. itsadug: interpreting time series and autocorrelated data using GAMMs. R package version 2.4.

[25] Wade, L., Lai, W., Tamminga, M. 2021. The reliability of individual differences in VOT imitation. *Language and Speech* 64, 576-593.

[26] Wood, S. 2019. mgcv: mixed GAM computation vehicle with automatic smoothness estimation. R-package version 1.8–31.

[27] Yu, A. C. L. 2019. On the nature of the perception-production link: individual variability in English sibilant-vowel coarticulation. *Laboratory Phonology* 10, 1–29

[28] Zellou, G. 2017. Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics* 61, 13–29.

[29] Zhu, J., Styler, W., Calloway, I. 2019. A CNN-based tool for automatic tongue contour tracking in ultrasound images. *arXiv preprint arXiv:1907.10210.*

PRE-/l/ VOWEL CHANGE IN AUSTRALIAN ENGLISH *POOL-PULL*

Tünde Szalay[1], Titia Benders[2], Felicity Cox[3], Michael Proctor[3]

[1]University of Sydney, [2] University of Amsterdam, [3]Macquarie University
tuende.szalay@sydney.edu.au

ABSTRACT

Contrast reduction between Australian English pre-lateral /ʉː-ʊ/ (*pool-pull*) compared to other environments might indicate an ongoing sound change resulting in merger. Our apparent-time study explores this merger by comparing the *F*2 characteristics of /ʉː-ʊ/ produced in pre-lateral and pre-obstruent contexts by 9 older and 8 younger female speakers. Differences between older and younger speakers and potential merger within younger speakers were tested using Bayesian methods. Consistent with a pre-lateral vowel change, younger speakers showed reduced pre-lateral vowel contrast compared to older speakers. There was no conclusive evidence for vowel merger, as *F*2 targets of /ʉː-ʊ/ remain contrastive in the pre-lateral environment for some young speakers.

Keywords: sound change, pre-lateral vowels, change by coarticulation

1. INTRODUCTION

Systematic and directional coarticulation leading to contrast reduction is often implicated in the initiation of sound change and merger [1, 2]. Yet, not all coarticulation leads to sound change, and not all contrast reduction leads to merger [1, 2, 3]. In the Interactive Phonetic (IP) model, a sound change may be initiated when highly coarticulated realisations of one phoneme become acoustically similar to another phoneme, and a merger is complete when contrast between two categories is lost [2].

Pre-lateral vowels show contrast reduction and/or merger due to the coarticulatory influence of /l/ in several varieties of English. For example, vowel contrast reduction between pre-/l/ /iːl-ɪl/ has been observed in some dialects of American English as /iːl/ shifts toward /ɪl/ [4]. Southern British English /uːl-ʊl/ show an acoustic merger, as pre-/l/ /uːl/ is produced with an *F*2 similar to pre-/l/ /ʊl/ [5]. In Melbourne Australian English, /el-æl/ show merger through the lowering of /el/ [6]. Members of the Australian English (AusE) vowel pairs /iː-ɪ, ʉː-ʊ, æɔ-æ,/ and /əʉ-ɔ/ (e.g., *feel-fill, pool-pull, howl-Hal, dole-doll*) show acoustic contrast reduction in the pre-lateral position, while their pre-obstruent allophones remain distinct [7, 8, 9, 10, 11]. The AusE contrast reduction is consistent with an ongoing sound change and merger brought about through coarticulation. However, only the /æɔ-æ/ (*howl-Hal, owl-Al*) contrast has been studied using apparent-time methods: sound change is shown in younger speakers producing smaller /æɔl-æl/ contrast compared to older speakers, and merger is shown in younger male speakers' increased similarity between pre-/l/ /æɔl-æl/ and pre-/d/ /æɔ/ [12].

	1967 [13]	1999 [7]	2004 [8]	2021 [10]
Gender	M	M	F	F
ʊd	950	1000	1100	1132
ʉːd	1550	1750	2200	2197
ʊl	NA	NA	1000	937
ʉːl	NA	NA	1300	983

Table 1: Changes in AusE /ʉː-ʊ/ over time: mean *F*2 for speakers of different ages. Columns: publication date and speaker gender [13, 7, 8, 10].

Sound change and merger has not been thoroughly explored for pre-lateral /ʉː-ʊ/ (*pool-pull*), although the reduced contrast may be consistent with a contextual vowel merger [9, 10]. Due to the *F*2 drop in /ʉːl/, /ʉːl/ becomes spectrally similar to /ʊl/, which may contribute to loss of spectral differentiation (Table 1). Therefore, we examine whether the reduced contrast between pre-/l/ /ʉː-ʊ/ [10] is indicative of a vowel change and merger in AusE. We hypothesised that younger speakers would (1) preserve spectral /ʉː-ʊ/ contrast in the pre-obstruent environment; (2) show smaller spectral contrast between pre-lateral allophones of the members of the vowel pair /ʉː-ʊ/ than older speakers; and (3) not spectrally differentiate between /ʉːl/ and /ʊl/.

2. METHODS

2.1. Speakers

Data were extracted from AusTalk, an AusE speech corpus recorded between 2011 to 2015 [14]. Recordings of 8 younger (age = 20 – 29, mean = 24.4) and 9 older (age = 54 – 80, mean = 66.6) native female speakers of AusE were selected. Speakers were born and educated in the Greater Sydney Metro Region with at least one Australian-born parent. The speakers did not report any reading, speaking, or hearing difficulties.

2.2. Material and procedure

The stressed vowels /ʉː-ʊ/ were produced in two monosyllabic paradigms, /hVd/ and /pVl/ (*who'd-hood, pool-pull*), in a single-word production task. Speakers read 322 isolated words, including the four target words, as they were presented orthographically on a computer monitor in random order. The task was recorded during three separate sessions, each using a different order of words. Each speaker produced up to three repetitions of each lexical item; the number of repetitions differs between participants, as not all participants attended all sessions.

2.3. Phonetic analysis

200 tokens were analysed (4 target words × 17 speakers × 3 maximal repetitions - 4 missing repetitions). Segment boundaries were located automatically using the MAUS forced aligner with the AusE grapheme-to-phoneme converter [15, 16, 17], and corrected manually in a Praat interface when a boundary was misplaced by more than 20 ms [18, 19]. Vowel onset was determined based on voicing onset and sudden increase in amplitude. Vowel offset in the /d/ context was determined based on amplitude drop. Rime offset in the /l/ context was determined based on voicing offset. As there is no discernible boundary between the vowel and the following /l/ in /pVl/ words, the entire /pVl/ rime was analysed (Fig. 1). Vowel duration was not measured due to the lack of discernible boundary.

Formant trajectories in pre-/d/ vowels and lateral-final rimes were extracted automatically and corrected manually in Praat [18]. Formant frequencies were estimated at every 5 ms throughout a 25 ms formant analysis window using a 50 ms Gaussian window with 75% overlap and with 50 dB dynamic range and a pre-emphasis filter increasing spectral slope above 100 Hz by 6 dB/octave. To optimise formant settings, four to five formants were tracked up to 4500 Hz ceiling in tokens with comparatively lower $F2$ and $F3$, or up to a maximum frequency of 6000 Hz in tokens with a comparatively higher $F2$ or $F3$ trajectory. Formant trajectories were manually corrected using a Praat-based interface that superimposed formant estimates over a broadband spectrogram calculated over 5 ms windows with 40% overlap, allowing for corrections of estimates that did not align with the visible formants. After hand-correction, $F1$–$F3$ trajectories for every word were visually inspected; values 1.5 times above or below the interquartile range for each formant in each vowel × coda × age group were rechecked.

Acoustic targets were located automatically in the corrected $F2$ trajectories using the $F2$ maxima for central /ʉː/ and the $F2$ minima for back /ʊ/. Pre-/d/ targets were located in the vowel trajectory and pre-/l/ targets in the first half of the rime (Fig. 2). Phonetic analysis was conducted by the first author.

2.4. Statistical analysis

The $F2$ difference between coda contexts and age groups was tested using a Bayesian Multilevel Model (BMLM). In Bayesian modelling, our certainty regarding the estimate of each factor prior to modelling the data is quantified by a prior probability distribution of the estimate [20, 21]. The prior distributions are combined with the data, generating a posterior distribution of estimates for each factor [20]. The interval containing 95% of all posterior estimates is referred to as the Credible Interval (CI). The CI is compared to the Region Of Practical Equivalence (ROPE), defined by the researcher as the "effect smaller in magnitude than the minimally interesting effect" [21]. When the CI falls outside the ROPE, the two groups are considered to be different; when it falls within the ROPE, the two groups are considered to be equivalent; and when the 95% CI spans values both outside and within the ROPE, the results are considered to be inconclusive [21].

Factor	Level	Estimate	SD
Age	Younger	100	100
Vowel	/ʉː/	750	100
Coda	/l/	-200	100
Age:Vowel	Younger:/ʉː/	250	100
Age:Coda	Younger:/l/	-100	100
Age:Vowel:Coda	Younger:/ʉː/:/l/	-300	300

Table 2: Informed priors in the Bayesian models

We constructed a BMLM with the dependent variable $F2$ Target, and the independent variables Vowel, Coda, and Age (interacting, dummy coded, comparing /ʉː/ to the baseline /ʊ/, /l/ to /d/, younger to older) using the brms library [22, 23, 24]. The model included a random intercept for Speaker with slope for Coda. Informed priors are reported in Table 2 [7, 8, 10]. The ROPE was defined as -76 to +76 Hz: a factor of two larger than the perceptual threshold of noticeable vowel difference, comparable to the smallest distance between two adjacent vowels in American English [25]. Pillai-scores were calculated for each participant using MANOVA [26, 27]. All data analysis was conducted in R [28].

3. RESULTS

Older speakers produced /ʉː/ with a higher $F2$ than /ʊ/ (β = 1115, est. error = 33.34, CI = 1051 – 1182), with the CI exceeding the ROPE, providing strong evidence for distinct pre-/d/ vowels. Older speakers produced /ʊ/ with a lower $F2$ in the /l/ than in the /d/-context (β = -136, est. error = 47.16, CI = -232 – -45). The evidence for older speakers backing

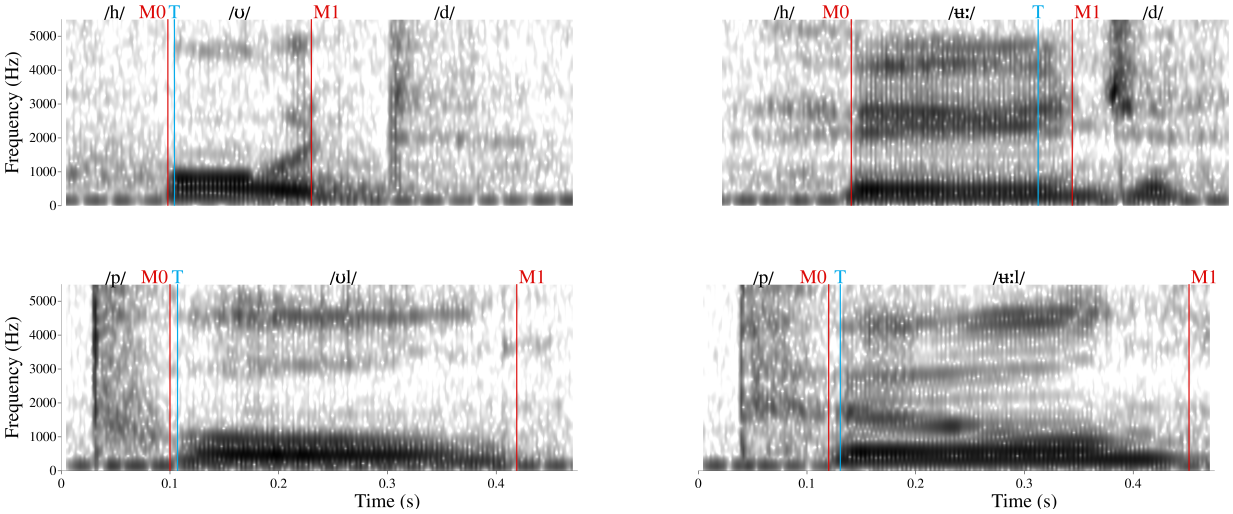


Figure 1: *Vowel (top) and rime (bottom) boundaries (start:M0, end: M1) with vowel target (T) by an older speaker.*

pre-/l/ /ʊ/ is inconclusive, as the CI spans interesting values and the ROPE. There is strong evidence for older speakers backing /ʉː/ more than /ʊ/ (β = -670, est. error = 44.4, CI = -760 – -585), as the CI exceeds the ROPE.

Younger speakers might have produced /ʊ/ with a higher *F*2 than older speakers (β = 71, est. error = 50.94, CI = -33 – 168). The Age:Vowel interaction (β = -16, est. error = 49.27, CI = -112 – 81) indicates that younger speakers, compared to older speakers, might not have fronted /ʉː/ as much as they fronted /ʊ/ (Fig. 2). The Age:Coda interaction (β = -13, est. error = 66.61, CI = -136 – 125) indicates that the difference between pre-obstruent and pre-lateral /ʊ/ might have been similar for younger and older speakers. The evidence for age-related vowel change in pre-/d/ vowels and pre-/l/ /ʊ/ is inconclusive as the CIs span both interesting values and the ROPE (Fig. 2). The negative three-way interaction (β = -344, est. error = 66.63, CI = -475 – -212) provides strong evidence that the *F*2 decrease in /ʉː/ associated with /l/ is greater for younger than older speakers, as the CI exceeds the ROPE.

Planned comparisons show a categorical difference for older speakers with the CI exceeding the ROPE in both coda contexts (/d/: β = 1117, CI = 1050 – 1181; /l/: β = 446, CI = 382 – 513). For younger speakers, *F*2 difference exceeded the ROPE before /d/ only (β = 1099, CI = 1027 – 1171). Pre-/l/ results are inconclusive, as the CI spans interesting values and the ROPE (β = 85, CI = 11 – 158).

Pillai-scores show distinct vowel categories in both contexts for older speakers (/d/-context: mean = 0.97, sd = 0.05; /l/-context: mean = 0.90, sd = 0.14). Younger speakers maintain contrast in the /d/-context (mean = 0.99, sd = 0.004), but reduce it before /l/ (mean = 0.45, sd = 0.312) (Fig. 4).

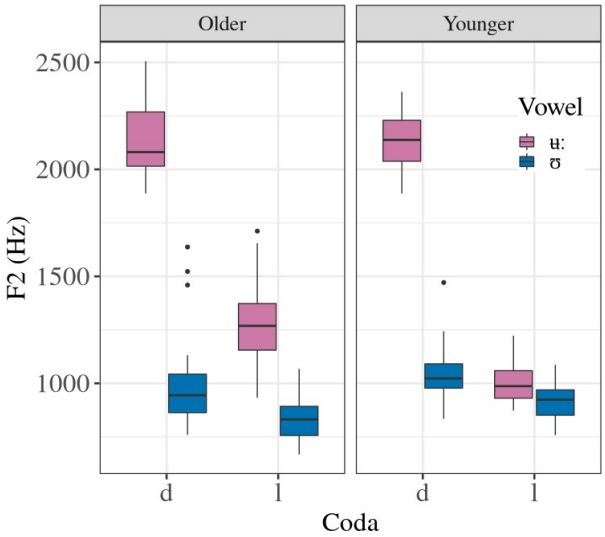


Figure 2: *F2 at vowel target*

4. DISCUSSION

H(1) predicted that vowel contrast would be preserved in the /d/-context. H(1) holds, as both younger and older speakers maintain /ʉː-ʊ/ contrast in the pre-/d/ context through higher *F*2 in /ʉː/ and distinct /ʉː-ʊ/ categories in the vowel space (Figs. 3-4). Although young speakers may produce /ʊ/ with a marginally higher *F*2 than older speakers, results on an ongoing /ʊ/-fronting are inconclusive.

H(2) predicted that younger speakers would produce pre-/l/ allophones of /ʉː-ʊ/ with smaller contrast than older speakers. H(2) holds, as younger speakers produce pre-/l/ /ʉː-ʊ/ with a smaller *F*2 contrast and a larger overlap in the *F*1-*F*2 space (Figs. 2–4). Younger speakers' larger overlap is primarily attributed to their larger *F*2 drop in /ʉːl/ compared to older speakers'. Younger speakers may produce /ʊ/ with an overall higher *F*2 in both contexts, further reducing the pre-/l/ /ʉː-ʊ/ contrast.

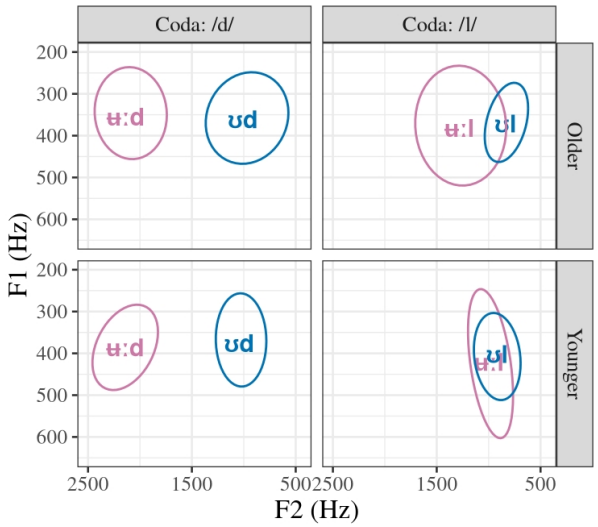


Figure 3: *F1 and F2 at vowel target*

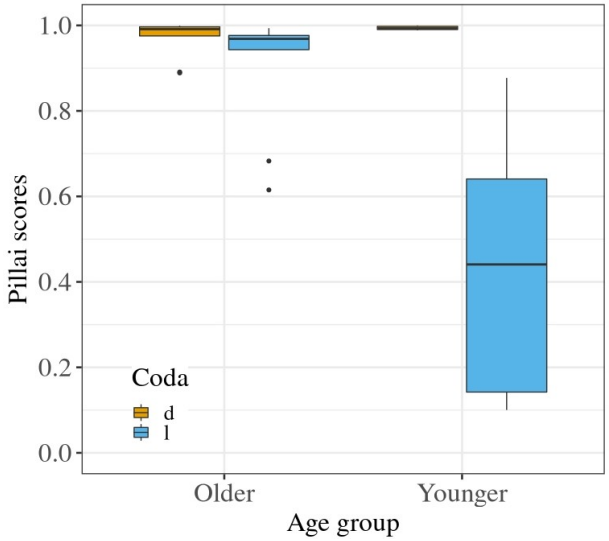


Figure 4: *Pillai-scores for the /ʊ-ʉː/ contrast*

Increased contrast reduction is consistent with an ongoing pre-lateral sound change in AusE during which the pre-/l/ allophone of /ʉː/ shifts towards pre-/l/ /ʊ/ in the acoustic space. That is, contrast reduction caused by coarticulatory variation in the /l/-context has become a sound change in AusE. This sound change can be represented in the IP model [2]: as listeners and speakers interact over time, coarticulated /ʉːl/ realisations are incorporated into listeners' representation of /ʊl/, shifting /ʉːl/ closer to /ʊl/ [2].

Pre-/l/ vowel contrast reduction due to *F*2 drop and coarticulatory vowel backing is observed in many varieties of English [29, 6, 30]. In Melbourne Australian English, *F*2 drop in /e/ before /l/, together with an *F*1 increase, contribute to an /el-æl/ merger [6]. *Pool* shows an *F*2 lowering in Ohio-, British-, and Yorkshire English, reducing *pool-pull* contrast [5, 31, 30, 32]. However, tongue body contrast is maintained for British English /uː-ʊ/, showing that an *F*2 drop does not always indicate tongue backing [5]. Thus, an articulatory study on AusE *pool-pull* contrast is required.

H(3) predicted that younger speakers would not contrast pre-lateral /ʉː-ʊ/. The evidence for H(3) is inconclusive: younger speakers may or may not produce pre-/l/ /ʉː-ʊ/ with a contrastive *F*2. Our results regarding a pre-/l/ /ʉː-ʊ/ merger may be inconclusive due to the large inter-speaker variation in younger speakers. Pillai scores, calculated on a speaker-by-speaker basis, reveal that younger speakers exhibit a range from near-maximal to near-minimal contrast (Fig. 4). Variation among younger speakers is consistent with the spread of a sound change during which some speakers are innovators merging pre-lateral /ʉːl-ʊl/, while others lag behind and maintain a contrast similar to older speakers [33]. However, younger speakers producing small contrast at the target may preserve duration contrast or contrast at some other point in the formant trajectories.

The inconclusive results on the /ʉːl-ʊl/ merger are consistent with the variation in young listeners' ability to distinguish /ʉːl/ and /ʊl/ in perception [11]. As spectral difference may fall above or below the just-noticeable difference, listeners are still able to discriminate between the two vowels, albeit less accurately [11]. Speakers producing /ʉːl-ʊl/ with a smaller spectral contrast may be less accurate at perceiving the same contrast, as speakers who produce a contrast less robustly are less likely to perceive it accurately [30]. More research is required to explore the relationship between listeners' perceptual accuracy with respect to the size of the spectral difference and its relation to just noticeable difference. Perceptual contrast reduction, a key marker of vowel merger [1, 2], may also be caused by contrast reduction in *F*1, in dynamic formant properties, and/or in durational differences [12]. Future research may explore the link between acoustic and perceptual contrast reduction using a combination *F*1, *F*2, durational values, and formant trajectories.

5. CONCLUSION

Pre-lateral /ʉːl-ʊl/ show an ongoing sound change in AusE, as younger speakers produce the pre-/l/ allophones with smaller *F*2 contrast than older speakers. Contrast reduction is caused by /ʉːl/ shifting towards /ʊl/. Yet, it is not clear if the *F*2 contrast reduction reaches the threshold of a merger or if the vowels are still contrastive.

6. ACKNOWLEDGEMENTS

This research was supported in part by the Australian Research Council DE150100318 and FT180100462 grants.

7. REFERENCES

[1] Ohala, J. J. 1989. Sound change is drawn from a pool of synchronic variation. In: Breivik, L. E., Jahr, E. H. (eds), *Language change: Contributions to the study of its causes*. Mouton de Gruyter, 173–198.

[2] Harrington, J., Kleber, F., Reubold, U., Schiel, F., Stevens, M. 2018. Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in cognitive science* 10(4), 707–728.

[3] Garrett, A., Johnson, K. 2013. Phonetic bias in sound change. In: Yu, A. C. L. (ed), *Origins of Sound Change*. Oxford, UK: Oxford University Press, 51–97.

[4] Labov, W., Ash, S., Boberg, C. 2008. *The Atlas of North American English, Phonetics, Phonology and Sound Change*. Berlin, Boston: Gruyter Mouton.

[5] Strycharczuk, P., Scobbie, J. M. 2017. Fronting of Southern British English high-back vowels in articulation and acoustics. *JASA* 142(1), 322–331.

[6] Loakes, D., Hajek, J., Fletcher, J. 2017. Can you t[æ]ll I'm from M[æ]lbourne? *English World-Wide* 38(1), 29–49.

[7] Cox, F. 1999. Vowel change in Australian English. *Phonetica* 56, 1–27.

[8] Cox, F., Palethorpe, S. 2004. The border effect: Vowel differences across the NSW-Victorian border. Moskovsky, C. (ed), *Proc. Conference of Australian Linguistics Society* Newcastle, Australia.

[9] Palethorpe, S., Cox, F. 2003. Vowel modification in pre-lateral environments. *Int. Seminar on Speech Production* Macquarie University, Sydney.

[10] Szalay, T., Benders, T., Cox, F., Palethorpe, S., Proctor, M. 2021. Spectral contrast reduction in Australian English /l/-final rimes. *JASA* 149(2), 1183–1197.

[11] Szalay, T., Benders, T., Cox, F., Proctor, M. 2021. Perceptual vowel contrast reduction in Australian English /l/-final rimes. *LabPhon* 12(1).

[12] Szalay, T., Benders, T., Cox, F., Proctor, M. 2022. Vowel merger in Australian English lateral-final rimes: /æɔ-æ/. Billington, R. (ed), *Proc. 18th Australasian Int. Conf on Speech Science and Technology* Canberra, Australia, 106–110.

[13] Bernard, J. R. 1967. *Some measurements of some sounds of Australian English*. PhD thesis.

[14] Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., others, 2011. Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box. ISCA.

[15] Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

[16] Schiel, F. August 1999. Automatic Phonetic Transcription of Non-Prompted Speech. Ohala, J. J., Hasegawa, Y., Ohala, M., Granville, D., Bailey, A. C. (eds), *Proceedings 14th ICPhS* San Francisco, CA, USA, 607–610.

[17] Schiel, F. 2015. A statistical model for predicting pronunciation. *Proc. ICPhS*.

[18] Boersma, P., Weenink, D. 2021. Praat 6.1.39.

[19] Cosi, P., Falavigna, D., Omologo, M. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *EuroSpeech*.

[20] Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., others, 2018. Bayesian inference for psychology. part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review* 25(1), 35–57.

[21] Dienes, Z. 2021. How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice* 8(1), 9.

[22] Bürkner, P.-C. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1), 1–28.

[23] Bürkner, P.-C. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1), 395–411.

[24] Bürkner, P.-C. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software* 100(5), 1–54.

[25] Kewley-Port, D., Zheng, Y. 1999. Vowel formant discrimination: Towards more ordinary listening conditions. *JASA* 106(5), 2945–2958.

[26] Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics* 26(1), 117 – 121.

[27] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of phonetics* 34(4), 458–484.

[28] R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[29] Giles, S., Moll, K. 1975. Cinefluorographic study of selected allophones of English /l/. *Phonetica* 31(3-4), 206–227.

[30] Harrington, J., Kleber, F., Reubold, U. 2008. Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *JASA* 123(5), 2825–2835.

[31] Wade, L. 2017. The role of duration in the perception of vowel merger. *LabPhon* 8(1).

[32] Gorman, E., Kirkham, S. 2020. Dynamic acoustic-articulatory relations in back vowel fronting: Examining the effects of coda consonants in two dialects of British English. *JASA* 148(2), 724–733.

[33] Rogers, E. M., Singhal, A., Quinlan, M. M. 2014. Diffusion of innovations. In: *An integrated approach to communication theory and research*. Routledge, 432–448.

DIPHTHONGISATION OF /iː/ IN WEST FRISIAN

Martijn Kingma[1,2], Paul Boersma[2], Hans Van de Velde[1,3], Arjen Versloot[2,4]
[1]Fryske Akademy, Leeuwarden, The Netherlands
[2]Universiteit van Amsterdam, Amsterdam, The Netherlands
[3]Universiteit Utrecht, Utrecht, The Netherlands
[4]Rijksuniversiteit Groningen, Groningen, The Netherlands

mkingma@fryske-akademy.nl; paul.boersma@uva.nl; hvandevelde@fryske-akademy.nl; a.p.versloot@uva.nl

ABSTRACT

Contemporary West Frisian (ISO 639-2 fry; henceforth: Frisian) has a rich and symmetrical vowel system, with nine short and nine long monophthongs as well as numerous rising, falling and centring diphthongs. However, the vowel system of this minority language is changing, some of these changes are induced by contact with Dutch, others are more likely to be internal changes. One such sound change in progress is observed in the long high vowel /iː/, which younger speakers tend to realise as a centring diphthong [iˑə]. Consequently, the traditional minimal pair *wiid* /ʊiːt/ 'wide' ~ *wiet* /ʊiˑət/ 'wet' becomes homophonic [ʊiˑət].

We present the first analysis of this sound change in progress. Based on acoustic analyses of the trajectory of /iː/ of 51 speakers from Boarnsterhim, we show that the youngest speakers use this centralising diphthongisation more often than older speakers. Moreover, analyses show that diphthongisation occurs more often before alveolar consonants.

Keywords: Sound change, diphthongisation, Frisian, sociophonetics, vowels

1. INTRODUCTION

1.1. The Frisian language situation

Frisian is a West Germanic language spoken by approximately 450 thousand people in the province of Fryslân in the north of The Netherlands [1]. It is recognised as the official second language in the province and spoken by around three quarters of its inhabitants. Though the language is supported in more formal domains, it is strongest in the domains of family, work and village communities [2]. Traditionally, three dialects are distinguished: Clay, Wood and Southwestern. The Clay dialect provides the basis for the official standard, though this standard is not often adhered to in less formal domains.

For centuries, Frisian has been in direct contact with its neighbouring languages, especially Dutch. Yet, the influence of Dutch on Frisian has increased since the advent of mass media and the increase of mobility from the 19th century onwards [1, 2]. This intensive contact has left its mark on Frisian syntax (e.g., verb order), lexicon, morphology (e.g., diminutive formation) and phonology (e.g., realising /a/ as [a] instead of rounded [ɔ] before dentals) [3].

1.2. The Frisian vowel system

The Frisian vowel system is quite a large one [4]. Each monophthong displayed in Table 1 – except schwa – has a long counterpart that has phonemic status (e.g., *wer* /ʊɛr/ 'again' ~ *wêr* /ʊɛːr/ 'where' and *kat* /kɔt/ 'cat' ~ *kâld* /kɔːt/ 'cold'). Thus, there are 19 monophthong phonemes in Frisian.

	front	central	back
close	i iː y yː		u uː
close–mid	e eː ø øː		o oː
mid		ə	
open–mid	ɛ ɛː		ɔ ɔː
open		a aː	

Table 1: Vowel chart of Frisian.

Frisian also has three types of diphthongs, which can be split into rising (e.g., *ljocht* /ljɔχt/ 'light'), falling (e.g., *rein* /rajn/ 'rain') and centring (e.g., *beam* /beˑəm/ 'tree'). The centring diphthongs play a role in one of the most characteristic processes of Frisian phonology: breaking [5]. In complex words, they change to a combination of a glide + vowel (e.g., *stien* /stiˑən/ 'stone' becoming *stiennen* [stjenn̩] 'stones' or *stientsje* [stjentsjə] 'small stone').

As Dutch pressure on Frisian increases, the vowel system of Frisian may also be affected. One sound change brought forward as Dutch interference is the diphthongisation of long high vowels [3]. Though Sipma already noted slight diphthongisation in all long vowels as early as 1913 [6], Frisians born after circa 1980 are said to realise the long high vowels as centring diphthongs instead, especially before dental and alveolar consonants [3, 7]. These novel realisations indicate a merger between two vowel

phonemes, as *wiid* /ʊiːt/ ‘wide’ and *wiet* /ʊiˑət/ ‘wet’ are both increasingly realised as [ʊiˑət]. The current diphthongisation can also be seen as a language-internal process, as since the Early Modern Period, a context-conditioned reshuffle of /i/, /iː/ and /iˑə/ of various etymological origins has been in progress in Frisian [8].

1.3. The Boarnsterhim Corpus

The Boarnsterhim Corpus (BHC) provides an excellent tool to study sound change in Frisian.[1] This corpus comprises roughly 120 hours of speech in both Frisian and Dutch by bilinguals from the former municipality of Boarnsterhim. It is claimed that the (Clay) Frisian variety from this centrally located area is rather conservative and close to the standard (like Oxford for English or Haarlem for Dutch). The corpus is unique in its design, as multiple generations of a total of 37 families were recorded in both Frisian and Dutch at two points in time (1982–1984 and 2017–2019).

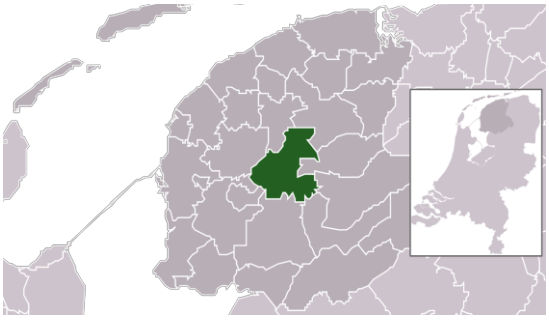


Figure 1: Map showing the Boarnsterhim area (1984–2012) [9].

All 113 speakers in the BHC were recorded reading a passage of 20 sentences and a short story of roughly 5 minutes. Following this, a semi-structured interview with a native speaker of the language in question ranging between 15 and 40 minutes is included. Recordings were made in the living rooms of participants to create a natural language setting for the use of Frisian.

In this paper, we use data from the BHC to look for acoustic confirmation that younger Frisian speakers show more centralising diphthongisation of /iː/ and make some first steps toward relating this sound change to linguistic factors.

2. METHOD

2.1. Materials

A total of 51 speakers recorded in the 2017–2019 section of the BHC were analysed. There were 20 male speakers and 31 female speakers. The youngest was born in 2001, while the eldest was born in 1929 (M = 1963, SD = 22). Thus, this study spans a birth cohort of 72 years.

All instances of /iː/ in closed, stressed syllables of content words in both read and spontaneous speech were traced using the Frisian orthographic annotation (i.e., <ii>) in the textgrids accompanying the audio files. After excluding 6 tokens with excessively foregrounded background noise, a total of 370 tokens were analysed.

2.2. Data processing

The start and end times of the vowels were manually segmented in a textgrid in Praat [10]. Neighbouring consonants were also annotated. The first three formants were estimated using a Praat script at 25%, 50% and 75% of the vowel duration with the Burg algorithm (for data files and scripts, see fon.hum.uva.nl/archive). While the time step (10 msec), number of formants (5), window length (25 msec) and pre-emphasis (50 Hz) were kept constant, the formant ceiling was adapted per speaker following the best fit provided by the FormantPath function in Praat in the read word *wiif* [ʊiːf] ‘woman’. For 2 speakers, the read word *siik* [siːk] ‘sick’ was used due to interfering background noises in *wiif*. Outliers that were further than 1.5 times the interquartile range from the lower/upper quartiles at 25, 50 and 75 percent were manually corrected based on changed LPC settings that best tracked the first three formants upon visual inspection.

2.3. Data analysis

To analyse the effect of the fixed factor *birth year* on diphthongisation of speakers, the lme4 package [11] in R [12] was used, as well as the lmertest package [13] to compute *p*-values. Diphthongisation was operationalised as change in (natural) log-transformed F1 between 25%, 50% and 75% of vowel duration. The three points of measurement, i.e., the within-speaker factor, were coded as the factor *position* (1, 2 and 3 respectively). The factor *speaker* was included as a random effect. The factor *birth year* was centred (byc) by subtracting the mean birth year (1963) individual values, and position by subtracting 2 from each position value (psc). The model specification was as follows in R (for more details, we refer to our analysis scripts):

(1) logf1~byc*psc+(psc|speaker)

The interaction effect between position and birth year will answer our primary research question. Inspection of the residuals showed no deviations of normality such as skewness or long tails. F1–F2 graphs were created in the web tool Visible Vowels [14].

3. RESULTS

Figure 2 shows the untransformed mean formant values, pooled over both genders and all following consonants and split by birth year (before and after 1980). This year was chosen because it represents the start of the fourth generation of BHC speakers. Note that due to this pooling, the figure is best interpreted as heuristic. As expected, younger speakers show a centralising realisation against a closing trajectory for older speakers. This is visualised by a higher F1 and lower F2 at 75% than at the previous two points. Older speakers on average exhibit a trajectory to a more peripheral position, similar to (short) [i] in Dutch [15].

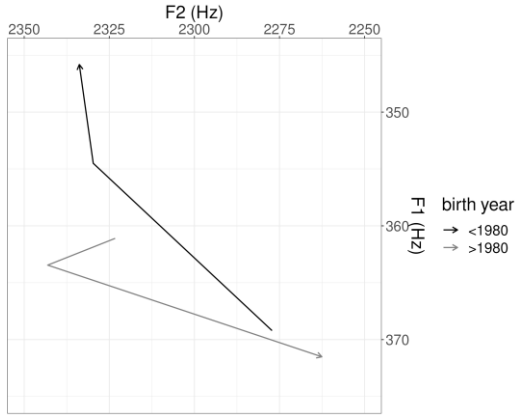


Figure 2: Average trajectory of [i:] split by generation.

	$\hat{\beta}$	SE	*t*	*p*
intercept	5.874	0.012	477.386	<0.001
byc	$-3.6 \cdot 10^{-5}$ /yr	$56.4 \cdot 10^{-5}$ /yr	-0.064	0.95
psc	-0.0254 /qrtile	0.0036 /qrtile	-7.009	<0.001
byc × psc	$9.5 \cdot 10^{-4}$ /qrtile/yr	$1.7 \cdot 10^{-4}$ /qrtile/yr	5.497	<0.001

Table 2: Summary of the linear-mixed effects model for logF1, referenced to *position* 2 (psc) and *birth year* 1963 (byc).

Table 2 shows the results of our model, specified in section 2.3. The interaction between *birth year* and *position* shows that when birth year increases, the slope for position becomes less negative (i.e., the fall of F1 becomes shallower by 0.095 percent per quartile per year, i.e., by 0.19 percent from first to third quartile in the vowel per year, e.g., by 9.5 ± 3.4 percent in 50 years). It can thus be concluded that younger speakers centralise more than older speakers, as such a diphthongised realisation goes hand in hand with an increased F1.

3.1. Effect of following consonant

In order to explore the effect of place of articulation of the *following consonant* on diphthongisation, it is included as a fixed effect (folc). In our data set, possible assimilation was not accounted for, meaning for instance that all tokens of 'tiid' were treated as [ti:t] and possible cases of assimilation of voice (e.g., [ti:d]) were ignored. As there were four consonants, we made a four-way distinction with three contrasts: alveolar (+0.5) vs non-alveolar (-0.5), labiodental [f] (-0.5) vs velar [k] (+0.5) and the two alveolars [s] (-0.5) vs [t] (+0.5). The model specification in R was:

(2) logf1~byc*psc*folc+(psc*folc|speaker)

	$\hat{\beta}$	SE	*t*	*p*
intercept	5.877	0.012	464.820	<0.001
byc	$1.9 \cdot 10^{-5}$ /yr	$58.1 \cdot 10^{-5}$ /yr	0.033	0.97
psc	-0.0228 /qrtile	-0.0045 /qrtile	-4.987	<0.001
folc-fk+st	0.0381	0.0137	2.787	0.0078
folc-f+k	0.0124	0.0123	1.011	0.32
folc-s+t	-0.0128	-0.0185	-0.695	0.50
byc × psc	$8.6 \cdot 10^{-4}$ /qrtile/yr	$2.1 \cdot 10^{-4}$ /qrtile/yr	4.070	<0.001

Table 3: Summary of the linear-mixed effects model for logF1, including *following consonant.*

As can be observed in Table 3, there is a significant effect for the following consonant. Non-significant interactions are not mentioned in the table. Alveolarity influenced F1 positively, increasing roughly 3.8% per quartile (measurement position). This rise in F1 can be interpreted as a lowering of the tongue, indicating a centring realisation. Thus, it seems that diphthongisation of /i:/ is encouraged by the alveolarity of the following consonant, confirming [3, 7]. No significant differences between the alveolar fricative and plosive, or between velar [k] and labiodental [f], were detected. The fact that these non-alveolars mostly come from read speech, while the alveolars all came from spontaneous speech, may have influenced these results.

3.2. Individual variation

The average trajectories in Figure 2 are, due to the averaging, not very insightful in the nuances of the sound change in question. Following the theories on language change [16], some speakers lead while others follow. Therefore, we explore the individual variation displayed regarding diphthongisation of [iː] in this section.

Figure 3 shows two extremes, exemplified by the average trajectories of the two young male speakers in the data set: speaker 73 (14 tokens) and speaker 113 (5 tokens). The former can be labelled as conservative, while the latter innovatively diphthongises his /iː/. As can be observed, 73 shows a centralising trajectory, realising a centralising offglide typical to Frisian [iˑə] [17, 18].

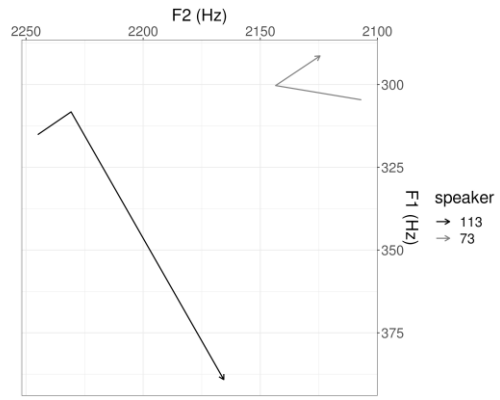


Figure 3: Average trajectory of [iː] of a conservative (073) and an innovative (113) speaker.

To further explore the individual variation of each speaker, we added up the fixed effect of position, plus each speaker's age effect ($\hat{\beta}_{byc \times psc}$ · (birth year – 1963)), plus each speaker's random slope for position, and multiplied the result by 2 to obtain the difference between the first and third quartile. Figure 4 shows a histogram of these total effects, or diphthongisation indexes, of all speakers.

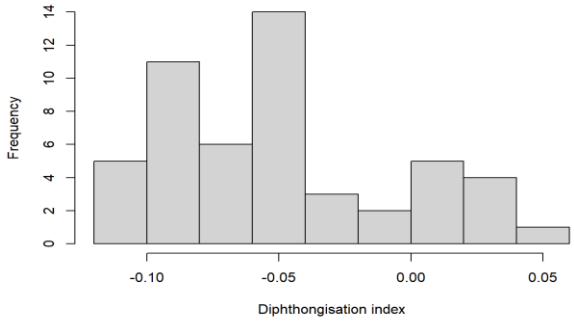


Figure 4: Histogram showing the individual variation in diphthongisation of [iː]. (n=51)

In Figure 4, we find the people that diphthongise their /iː/ to [iˑə] in the rightmost peak (0–0.05), as their realisations show a trajectory of rising F1 values instead of the conventional lowering (see also Figure 2). Speakers on the negative side of the x-axis more often have a closing trajectory typical of [iː], while speakers on the positive side more often show a centring trajectory. The fact that most speakers display a negative diphthongisation index shows us that the ongoing sound change is still in its incipient phase.

4. CONCLUSION

In conclusion, young Frisian speakers diphthongise /iː/ to [iˑə] more than older speakers, though not every young speaker participates equally in this sound change. On average, centring diphthongisation (with regard to F1) increased between 6.1 and 12.9 percent over half a century in apparent time. This diphthongisation is especially apparent before alveolar [t] and [s]. The findings regarding alveolarity ought to be viewed as exploratory, as the data were not equally spread across speech styles.

Although diphthongisation of long vowels in Frisian was already observed in 1913 [6], it is mainly present in the speech of speakers from Boarnsterhim born after 1980. Though the actuation and diffusion of sound change lie beyond the scope of this paper, we hypothesize that the (minority) language situation in Fryslân may play a role. For Frisian, both laymen and linguists commonly point to Dutch as the cause of all changes. As Dutch lacks the phonemes /iː/ and /iˑə/, and as a matter of fact lacks centring diphthongs altogether [19], attributing this sound change purely to language contact seems unsatisfactory. The sound change fits a long-standing shift in realisations of high vowels in Frisian [8]. Yet, the merger of two non-Dutch phonemes to one may be an indirect result of Dutch influence on Frisian. We plan to investigate this in the future, on the basis of the complete BHC. This will also include more analyses on the graduality of the sound change in question.

As far as other future research is concerned, most opportunities with the BHC lie in including more vowels and more consonantal contexts. Speakers recorded in both time periods of the corpus could also be investigated in order to see whether change in real-time can be observed. Nevertheless, this study serves as a stepping stone for additional research into the sound changes currently occurring in this minority language.

5. REFERENCES

[1] Hilton, N. H., Jensma, G., Visser, W. 2020. Frisian. In: Grenoble, L., Lane, P., Royneland, U. (eds.), *Linguistic minorities in Europe online*. De Gruyter, https://doi.org/10.1515/lme.10739543

[2] Gorter, D., Van der Meer, C., Riemersma, A. 2022. Frisian in the Netherlands. In: Grenoble, L., Lane, P., Royneland, U. (eds.), *Linguistic minorities in Europe online*. De Gruyter, https://doi.org/10.1515/9783110208351.3.185

[3] De Haan, G.J. 1997. Contact-induced change in modern West Frisian. *Us Wurk,* 46, 61-89.

[4] Visser, W. 2020. Segment inventory. *Taalportaal*. https://www.taalportaal.org/taalportaal/topic/pid/topic-14248766305431313

[5] Visser, W. 2020. Breaking. *Taalportaal*. https://www.taalportaal.org/taalportaal/topic/pid/topic-14358346024113628

[6] Sipma, P. 1913. *Phonology and grammar of modern West Frisian*. Oxford University Press.

[7] Visser, W. 1997. *The syllable in Frisian*. PhD dissertation. Vrije Universiteit Amsterdam.

[8] Hoekstra, J. 2001. An outline history of West Frisian. In: H.H. Munske et al. (eds.), *Handbuch des Friesischen.* Max Niemeyer Verlag, 722-734.

[9] CBS. 2009. Map municipality code 0055. https://commons.wikimedia.org/wiki/File:Map_-_NL_-_Municipality_code_0055_(2009).svg

[10] Boersma, P., Weenink, D. 2022. *Praat, doing phonetics by computer*. Praat.org. Version 6.3.03.

[11] Bates, D., Maechler, M., Bolker, B., Walker, S. 2022. Lme4: linear mixed-effects models using Eigen and S4. *R package*. Version 1.1-31.

[12] R Core Team. 2022. R: a language and environment for statistical computing. R-project.org. Version 4.2.2.

[13] Kuznetsova, A. Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82, 1-26.

[14] Heeringa, W., Van de Velde, H. 2018. Visible Vowels: a tool for the visualisation of vowel variation. *Proc. CLARIN* Pisa, 120-123.

[15] Van der Harst, S. 2011. *The vowel space paradox*. PhD dissertation. Radboud Universiteit Nijmegen.

[16] Labov, W. 1991. *Principles of linguistic change, vol. 2: social factors.* Wiley.

[17] De Graaf, T. Meinsma, G. 1978. Ien lûd, twa lûden, twalûden: oanrin ta fonetyske stúdzje [Monophthong, two sounds, diphthong: towards phonetic study]. *Us Wurk,* 27, 81-88.

[18] De Graaf, T., Meinsma, G. 1980. De brekking fan sintralisearjende twalûden yn it Frysk [The breaking of centralising diphthongs in Frisian]. *Us Wurk,* 29, 77-82.

[19] Van Oostendorp, M. 2020. Segment inventory. *Taalportaal*. https://www.taalportaal.org/taalportaal/topic/pid/topic-14020545827588509

[1] The first author is currently working on publishing an updated version of the corpus online.

CHANGE OVER TIME IN [ɛ] AND [æ] IN HAWAIʻI CREOLE

James Grama

Sociolinguistics Lab, University of Duisburg-Essen
james.grama@uni-due.de

ABSTRACT

The paper presents the first quantitative investigation of two vowels in Hawaiʻi Creole, [ɛ] and [æ] (referred to here as JRES and CHRAEP). While these vowels have been described both as distinct and overlapping in the literature, the present paper adds differentiated evidence of external and internal factors that bear on their realizations. Using a corpus of 32 speakers born between 1896 and 1988, 2,311 tokens of the target vowels are acoustically analyzed. Results indicate phonological effects, in particular, that pre-lateral JRESS is merged with CHRAEP. In addition, changes are observed over time; whereas older speakers show notable overlap between JRES and CHRAEP, this has waned over time due to the lowering and retraction of CHRAEP. Despite this change, speakers show synchronic variation within age group in JRES-CHREAP overlap. These findings bear on models of language change in Hawaiʻi including decreolization and the feature pool.

Keywords: language variation, vowel change, Hawaiʻi Creole, Pidgin, phonological conditioning.

1. INTRODUCTION

Hawaiʻi Creole (or Pidgin, as it is locally known; ISO 693-3 hwc) is an English-lexified creole spoken by ca. 700,000 people, the vast majority of whom reside on the Hawaiʻi archipelago [1]. As a plantation-based creole, Pidgin has its roots in colonialism. After initial European contact in 1778, the Kingdom of Hawaiʻi underwent rapid and sweeping social changes. Increased foreign presence, in particular by Christian missionaries in the 19th century, contributed to both the steady decline of *ʻōlelo Hawaiʻi* (the Hawaiian language) and the systematic disenfranchisement of the *Kānaka Maoli* (Hawaiian people) [2-3]. These changes were spurred on by the establishment of sugarcane plantations in 1835, which drew on a globally sourced labor force, including workers from Portugal, Japan, and Cantonese-speaking parts of China. Plantation workers initially used a Hawaiian-lexified pidgin (Pidgin Hawaiian; PH) as a means of communication; however, by the 1880s, a small, but increasingly powerful English-speaking minority dominated Hawaiian economic spheres, motivating the development of a new, English-lexified contact variety, Hawaiʻi Pidgin English (HPE), spoken especially widely in the urban center of Honolulu [4]. HPE gradually displaced PH on plantations, as English systematically replaced ʻōlelo Hawaiʻi, and by 1910, assisted by the US's illegal seizure of Hawaiʻi in 1896, HPE was the *lingua franca* of Hawaiʻi [4]. After approximately three generations (between 1920-1930), HPE had creolized into Pidgin (Hawaiʻi Creole), and was the dominant language of the majority of locally-born children in Hawaiʻi [4-5].

A noteworthy feature of Pidgin is the status of the front vowels [ɛ] and [æ], referred to here as JRES and CHRAEP (this convention is based on Wells' lexical sets [6], but uses Pidgin cite words for *dress* and *trap*, spelled in Odo orthography [7] to refer to vowel categories; compare [8-9]). Various analyses claim that Pidgin shows no distinction between JRES and CHRAEP, with both vowels converging on [æ̝], particularly in basilectal varieties [7], and in the presence of non-obstruents (e.g., /l, n/) [10]. By contrast, mesolectal Pidgin speakers are reported to be more likely to produce a distinction between the two vowels, paralleling their realizations in Hawaiʻi English (compare [8] and [11]). However, example sentences from other descriptions (e.g., [12]) imply phonemic contrast between JRES and CHRAEP in a way that does not map cleanly onto English categories.

While acoustic phonetic analysis has been conducted on the local dialect of English spoken in Hawaiʻi (see [11, 13]), most descriptions of Pidgin have either not undertaken a direct analysis of the vowel system [1, 12], or they have relied entirely on auditory accounts [7, 10]. Thus, little is known about the acoustic patterning of Pidgin vowels, how phonological constraints bear on their realizations, and whether (and, if so, how) JRES and CHRAEP have changed over time. This paper begins to fill these gaps by conducting acoustic analysis on real-time spontaneous speech from a stratified Pidgin corpus.

2. METHODS

2.1. Speech corpora and participants

Participants were drawn from two speech corpora housed on Kaipuleohone at the University of Hawaiʻi at Mānoa [14]: the Bickerton Collection [10], recorded in the 1970s, and the Influences and Variation in Hawaiʻi Creole English collection,

recorded in the 2000s [15]. These corpora represent two independent samplings at 30-year intervals of the Pidgin-speaking community, with relatively older and younger speakers at each time point, allowing for real- and apparent-time comparisons (for more, see [16]). Given that the breadth of speaker birthdates (1896-1988) represents the entire life-span of Pidgin as a contact language, the corpora offer a thorough picture of how the language has changed over time.

A balanced number of speakers were drawn across relevant demographic categories: *corpus* (1970s v. 2000s), *relative age* (old v. young), and *speaker sex* (women v. men). All participants (n=32) were recorded with other Pidgin speakers and resided on Oʻahu, Hawaiʻi (the Big Island), Kauaʻi, or Maui. Regional differences in vowel realizations were tested for, but none arose (see [9]). Table 1 presents a breakdown of the participants across the aforementioned demographic categories.

Corpus	Age (YOB)	Women	Men
1970s	Old (1896-1922)	4	4
	Young (1927-1946)	4	4
2000s	Old (1947-1967)	4	4
	Young (1983-1988)	4	4

Table 1: Distribution of participants.

2.2. Data preparation

The sociolinguistic-style interviews analyzed here were transcribed in Transcriber [17], totaling some 11 hours, or 66,000 words of spontaneous speech. These data were force-aligned using LaBB-CAT [18] housed on the Sociolinguistics Server at the University of Hawaiʻi at Mānoa. Boundaries surrounding all target vowels were manually checked and, if necessary, hand-corrected in Praat [19]. Grammatical words and unstressed tokens were excluded from analysis; no more than five instances of any lexeme were extracted for analysis from any one speaker. Measurements of F1 and F2 were taken at the vowel midpoint, and outliers greater than three standard deviations were removed. The resulting measurements were normalized following [20] on the basis of the entire vowel space (n=11,191). This yielded a total of 1,158 tokens of JRES and 1,153 tokens of CHRAEP for analysis.

The data were analyzed graphically and with linear mixed-effects models in R [21]. Separate models were fit to normalized F1 and F2 of JRES and CHRAEP using *lme4* [22], with significance assessed via Kenward-Roger approximation with *sjPlot* [23] (following recommendations in [24]). These models initially included age group, speaker sex, and phonological context in a three-way interaction, which was simplified if this did not improve model fit (assessed by comparing AIC). Preceding segment did not improve model fit, and so investigations of phonological context are restricted to following environment; this includes pre-obstruent (JRES=641, CHRAEP=714), pre-nasal (JRES=343, CHRAEP=409), and pre-lateral tokens (JRES=174, CHRAEP=30). Given the effect of duration on formant behavior [25], all models include vowel duration as a fixed effect. Speaker and word were random intercepts. Age group was chosen over year of birth in order to capture non-linear relationships over time. Alpha was set at 0.05.

3. RESULTS

3.1. Global phonological effects

Effects of following phonological context are shown in Figure 1. In the aggregate, JRES and CHRAEP exhibit some overlap in pre-obstruent and pre-nasal positions. However, the vowels show completely overlapped distributions when followed by /l/. While pre-lateral CHRAEP could not be modeled due to low token numbers and uneven representation across age group, the effect is clear for JRES, which is significantly lower (β=0.65, t=5.87, p<0.001) and backer (β=-0.41, t=5.96, p<0.001) than JRES in pre-obstruent position. A comparison of pre-lateral JRES and CHRAEP in Figure 2 appears to confirm this effect; pre-lateral JRES exhibits clear overlap with CHRAEP, suggesting that the distinction between JRES and CHRAEP is neutralized in pre-lateral position.

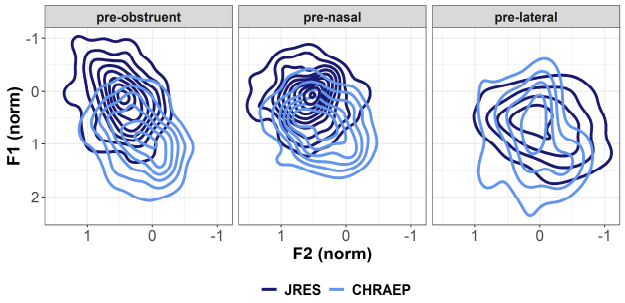


Figure 1: Two-dimensional kernel density plots of JRES (dark) and CHRAEP (light) across phonological context.

The two vowels show disparate effects when followed by a nasal. In JRES, a following nasal does not motivate significantly different realizations in either formant dimension (F1: β=-0.12, t=-1.31, p=0.192; F2: β=0.08, t=1.65, p=0.110). However, pre-nasal CHRAEP is significantly fronter (β=0.20, t=4.03, p<0.001) and numerically higher (β=-0.11, t=-1.90, p=0.058) than pre-obstruent CHRAEP. Of note, this effect in Pidgin does not motivate the extreme differences seen in the split-nasal system of some US varieties (where pre-nasal TRAP is raised and fronted; see [26]).

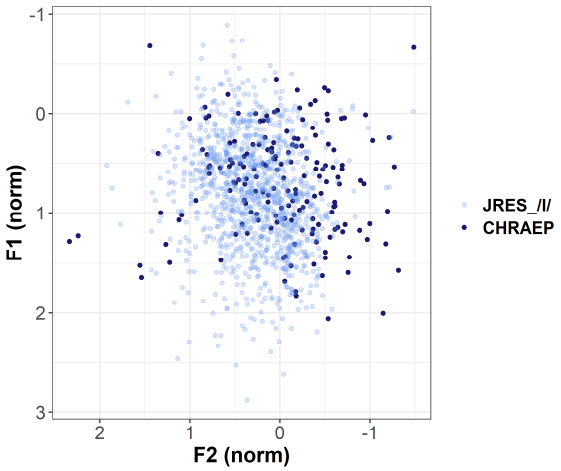


Figure 2: Scatterplot of pre-lateral JRES (dark) and CHRAEP (light); pre-lateral CHRAEP excluded.

3.2. Changes over time

Results indicate that the front vowel space has undergone significant reorganization over time, chiefly due to motion observed in CHRAEP (see Figure 3). In the oldest group (1970s old), CHRAEP and JRES show a large degree of spectral overlap, a relationship that persists into the 1970s young group. In the more recent corpus, both the 2000s old (β=0.34, t=3.03, p<0.01) and 2000s young (β=0.28, t=2.45, p<0.05) show significant lowering relative to the oldest cohort, with the 2000s young also showing significant retraction (β=-0.51, t=-3.38, p<0.01). By contrast, JRES is stable in F1 and F2 across age group.

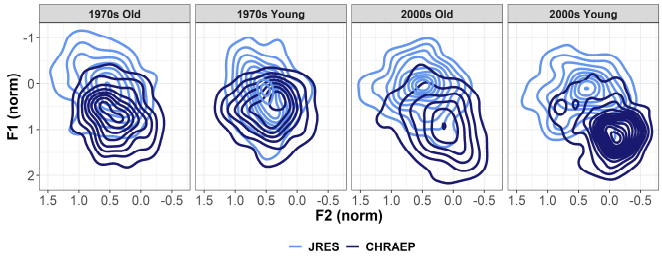


Figure 3: Two-dimensional kernel density plots of JRES (light) and CHRAEP (dark) over time; all pre-lateral tokens excluded.

The observed changes over time are mediated by phonological context. These effects are represented in the model estimates in Figure 4. While pre-lateral JRES generally remains lower than JRES in other contexts over age group, there is some variability by sex across age group. Both old 2000s men (β=-0.37, t=-2.12, p<0.05) and young 2000s women (β=-0.45, t=-2.54, p<0.05) show significantly raised pre-lateral JRES. This effect disappears for men in the subsequent generation, and pre-lateral JRES remains significantly backer and numerically (but not significantly) lower than pre-obstruent JRES for both old 2000s men (F1: β=0.16, t=1.02, p=0.562; F2: β=-0.66, t=-7.34, p<0.001) and young 2000s women (F1: β=0.20, t=1.38, p=0.354; F2: β=-0.45, t=-5.18, p<0.001). Given these differences, it is likely that the behavior seen in F1 for pre-lateral JRES points to variability in the sample rather than genuine change over time.

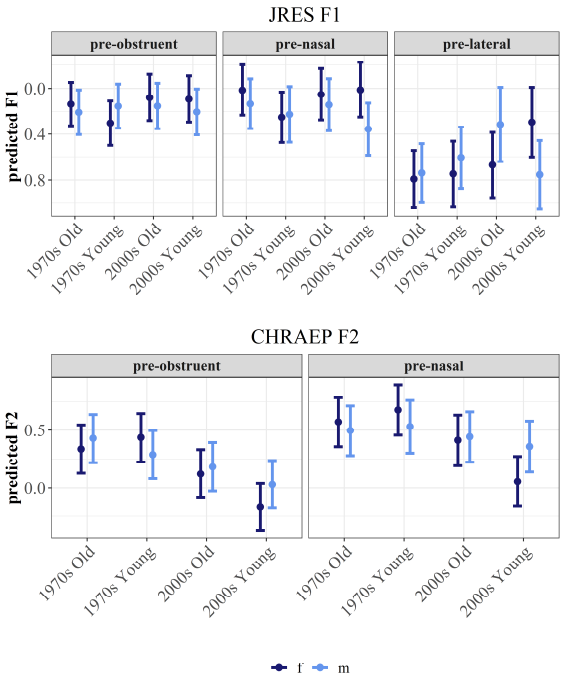


Figure 4: Predicted effects of phonological context and speaker sex (men=light; women=dark) by age.

Changes to CHRAEP over time are also mediated by phonological context and speaker sex. In particular, pre-nasal realizations of CHRAEP lag in their retraction over time, especially among men, an effect that is greatest in young 2000s men (β=-0.33, t=-5.02, p<0.001).

Together, these results suggest that older Pidgin speakers are more likely to exhibit a system where JRES and CHRAEP are overlapping, as described in [8]. Changes in speaker Pillai scores over time (a measure of vowel overlap; see [27]) support this observation (see Figure 5). Old 1970s speakers show a relatively tight grouping of scores (range: 0.03-0.22), which broadens somewhat in young 1970s speakers (range: 0.01-0.50). The old 2000s speakers show a relatively raised floor and ceiling (range: 0.12-0.62), a trend which continues into the young 2000s speakers (range: 0.18-0.70). Following the formula in [28] for calculating Pillai thresholds by sample size, only six speakers (all from the 1970s corpus) meet the criteria for a lack of distinction between JRES and CHRAEP (MrMonitz, Mrs Kodama, JA, George, GN, Allen).

Despite these trends, speakers clearly show variability in JRES-CHRAEP overlap in each corpus. Several speakers in the 1970s corpus (MrDias, NK, Kapeka, SA) show relatively high Pillai scores,

suggesting that the two vowels are distinct. In fact, it is these speakers who motivate the apparent aggregate differences evident in Figure 3 in each 1970s age group. And while most 2000s speakers exhibit clearly distinct JRES-CHRAEP distributions, some (Bula, Darryl, Summer), show stronger evidence of overlap. This synchronic variation indicates that, far from being complete, JRES-CHRAEP overlap remains in flux in modern Pidgin speakers.

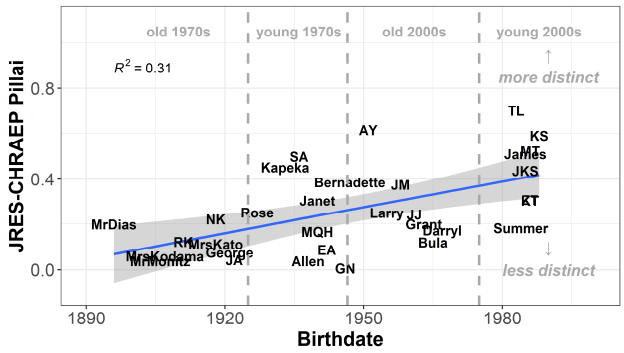


Figure 5: JRES-CHRAEP Pillai scores by birthdate with regression line; pre-lateral tokens excluded.

4. DISCUSSION

These results add considerable nuance to the phonetic relationship between JRES and CHRAEP, categories which the Pidgin literature has characterized as both distinct and overlapping. Results from this study demonstrate that CHRAEP has lowered and backed away from JRES over time, increasing the spectral distance between the two vowels. Evidence for JRES-CHRAEP overlap is found in Pidgin speakers recorded in the 1970s, while younger speakers are more likely to exhibit greater spectral distance between the two vowels. These changes over time are mediated by phonological context. For CHRAEP, a following nasal inhibits retraction (an effect that is clearest in young 2000s men), leading to a system where pre-nasal CHRAEP is somewhat more peripheral than pre-obstruent realizations (an effect that is nowhere near as extreme as the split-nasal system observed in some North American varieties; see [26]). In addition, pre-lateral JRES is considerably overlapped with CHRAEP, suggesting a contextual merger in Pidgin. This merger bears surface similarity to similar phenomena in Australia [29] and New Zealand [30], and is likely due to coarticulatory effects of the following /l/ rather than any direct contact between the varieties. Indeed, coda /l/ is velarized in Pidgin (or vocalized entirely; see [8-9]), which, in varieties of English, frequently motivates vowel retraction [6]. Importantly, the observed lowering and retraction does not generalize to other sonorant-adjacent contexts (compare [10]).

It is impossible to discuss the observed changes without also contending with the literature on decreolization. Though decreolization has chiefly been argued on the basis of morpho-syntactic variables, it has been the main explanatory model of change in Pidgin since the 1970s (see discussion in [31]). If, as [8] argues, JRES-CHRAEP overlap is a feature of basilectal Pidgin, then results from the present study would suggest that older speakers represent more basilectal varieties than younger speakers, and hence, be evidence for phonological decreolization (although the present study offers no operationalization of speakers' lectal status).

However, it is important to remember that English and Pidgin are two varieties that, in Hawai'i, are synchronically linked, despite their distinct grammars and histories. Pidgin and English have always been in direct, localized contact with one another in Hawai'i; the social hegemony established on the plantations (which reified English at the expense of Pidgin) continues to be reproduced today [3, 31], and many Pidgin speakers are bilingual in English [8]. In this way, decreolization fails to capture "the tensions inherent in post-colonial capitalist societies" [31]; instead, a model that draws on the importance of variables from both languages—the feature pool—is more appropriate to characterizing the socio-indexical landscape of Hawai'i [32-33]. Drager et al. [13] use apparent time data to demonstrate that TRAP is retracting in Hawai'i English without concomitant change in DRESS, similar to the pattern identified here for Pidgin. The authors argue that this asymmetry cuts against claims of a chain shift in Hawai'i English (compare the Low-Back-Merger Shift [34]). They instead identify self-reported ability to speak Pidgin as an important axis of variation in these vowels. Young Hawai'i English speakers who reported that they did *not* speak Pidgin exhibited lower TRAP onsets than those who reported that they *did* speak Pidgin, suggesting that English-Pidgin bilinguals showed more Pidgin-like TRAP realizations. Importantly, the Pidgin speakers considered in the present study show variability in the degree of JRES-CHRAEP overlap even in the youngest speakers, where CHRAEP retraction is greatest. This raises the possibility that the *extent* of JRES-CHRAEP overlap (or, whether a specific token is realized as either [ɛ] or [æ]) is a feature available for meaning making in Hawai'i for speakers of both Pidgin and Hawai'i English.

An investigation of the complex indexical meanings of JRES and CHRAEP is necessarily left to future work; however, the current study sets the groundwork by offering the first acoustic analysis of these vowels in Pidgin. Specifically, there is evidence (i) that pre-lateral JRES is merged with CHRAEP, and (ii) of a change in progress in Pidgin, whereby CHRAEP retraction away from JRES has increased the spectral differentiation between the two vowels.

5. ACKNOWLEDGMENTS

Funding was provided by the Russel J. and Dorothy S. Bilinski Fellowship Award and the University of Hawai‘i at Mānoa. Thanks to Melody Ann Ross, Isabelle Buchstaller, and two anonymous reviewers for their helpful comments, and to the participants for lending their voices to the corpora mentioned herein.

6. REFERENCES

[1] Velupillai, V. 2013. Hawai‘i Creole structure dataset. In S. M. Michaelis, P. Maurer, M. Haspelmath, & M. -Huber (eds.) *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

[2] Stueber, R. K. 1964. *Hawai‘i: A Case Study in Development Education, 1778-1960*. PhD dissertation. University of Wisconsin Madison.

[3] Kawamoto, K. Y. 1993. Hegemony and language politics in Hawai‘i. *World Englishes* 12(2), 193-207.

[4] Roberts, S. J. 2005. The emergence of Hawai‘i Creole English in the early 20th century: The sociohistorical context of creole genesis. PhD dissertation. Stanford University.

[5] Sakoda, K & Siegel, J. 2003. *Pidgin Grammar: An introduction to the creole language of Hawai‘i*. Honolulu: Bess Press.

[6] Wells, J. C. 1982. *Accents of English*. Cambridge: C.U.P.

[7] Odo, C. 1975. Phonological processes in the English dialect of Hawaii. PhD dissertation. University of Hawai‘i at Mānoa.

[8] Sakoda, K. & Siegel, J. 2008. Hawai‘i Creole: phonology. In B. Kortmann & E. W. Schneider (eds.), 729-749, *A Handbook of Varieties of English, Volume 1: Phonology*. Berlin: Mouton de Gruyter.

[9] Grama, J. 2015. Variation and change in Hawai‘i Creole vowels. PhD dissertation. University of Hawai‘i at Mānoa.

[10] Bickerton, D & Odo, C. 1976. *Change and Variation in Hawaiian English I: General Phonology and Pidgin Syntax.* Final report on NSF Project No. GS-39748.

[11] Kirtley, M. J, Grama, J., Drager, K., & Simpson, S. 2016. An acoustic analysis of the vowels of Hawai‘i English. *J. Int. Phon. Assoc.* 46(1), 79-97.

[12] Velupillai, V. 2003. *Hawai‘i Creole English: A Typological Analysis of the Tense-Mood-Aspect System*. Palgrave MacMillan: Hampshire, New York.

[13] Drager, K. Kirtley, M. J., Grama, J., & Simpson, S. 2013. Language variation and change in Hawai‘i English: KIT, DRESS, and TRAP. *Penn. Working Papers in Linguistics* 19(2), 41-50.

[14] Berez, A. 2013. The digital archiving of endangered language oral traditions: Kaipuleohone at the University of Hawai‘i and C’ek’aedi Hwnaz in Alaska. *Oral Tradition* 28(2), 261-270.

[15] Siegel, J. 2004. External influence and interval variation in current Hawai‘i Creole English. *National Science Foundation (Grant Number: 0345959)*.

[16] Grama, J. 2022. Managing legacy data in a sociophonetic study of vowel variation and change. In A. L. Berez, B. Mcdonnell, E. Koller, & L. B. Collister *The Open Handbook of Linguistic Data Management*, 221-236. MIT Press.

[17] Barras, C., Geoffrois, E., Wu, Z., Liberman, M. 2001.Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication* 33(1-2), 5-22.

[18] Fromont, R. & Hay, J. 2012. LaBB-CAT: An annotation store. In *Proc. of the Australasian Language Technology Workshop* 10, 113-117.

[19] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

[20] Lobanov, B.M. 1971. Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am.* 49(2b), 606-608.

[21] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL https://www.R-project.org/.

[22] Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1-48.

[23] Lüdecke 2022. sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.10.

[24] Luke, S. G. 2017. Evaluating significance in linear-mixed effects models in R. *Behavior Research Methods* 46(4), 1494-1502.

[25] Lindblom, B. 1963. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35(11), 1773-1781.

[26] Labov, W., Ash, S. & Boberg, C. 2006. *The Atlas of North American English*. New York: Mouton de Guyter.

[27] Nycz, J. & Hall-Lew, L. 2013. Best practices in measuring vowel merger. *J. Acoust. Soc. Am. – Proc. Meet. Acoust.* 20, 060008.

[28] Stanley, J. A. & Sneller, B. 2023. Sample size matters in calculating Pillai scores. *J. Acoust. Soc. Am.* 153(1), 54-67.

[29] Schmidt, P., Diskin-Holdaway, C., & Loakes, D. 2021. New insights into /el/-/æl/ merging in Australian English. *Australian Journal of Linguistics* 41(1), 66-95.

[30] Thomas, B. 2003. A study of the /el/-/æl/ merger in New Zealand English. *New Zealand English Journal* 17, 28-44.

[31] Sato, C. J. 1994. Language change in a creole continuum: Decreolization? In K. Hyltenstem & A. Vivey (eds.), 122-143, *Progression and regression in language*. Cambridge: C.U.P.

[32] Mufwene, S. S. 2001. *The Ecology of Language Evolution*. Cambridge: C.U.P.

[33] Cheshire, J, Kerswill, P, Fox, S, Torgersen, E. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *Journal of Sociolinguistics* 15(2), 151-196.

[34] Becker, K. 2019. *The Low-Back-Merger-Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and Short Front Vowel Shifts across North America*. Publication of the American Dialect Society, 104. Durham: Duke University Press.

FROM PASTA TO P^{H}IZZA: VARIATION AND CHANGE IN VOT IN SWISS GERMAN

Carina Steiner[1,2], Péter Jeszenszky[3], Adrian Leemann[1,2]

[1]University of Bern, Switzerland; [2]University of Zurich, Switzerland, [3]Independent researcher
carina.steiner@unibe.ch, pjeszenszky@gmail.com, adrian.leemann@unibe.ch

ABSTRACT

In contrast to Standard German (StG) or English, word-initial bilabial and alveolar stops are traditionally not aspirated in Swiss German (SwG), and typically exhibit short voice onset times (VOT; e.g. StG [ˈpʰitsa] vs. SwG [ˈpitsa]). Throughout the last century, however, more and more aspirated forms have been reported for SwG, but empirical evidence on the factors driving this change is thin. This study is the first to investigate SwG VOT in a large and balanced sample of 1000 speakers from 125 localities. The results suggest an apparent-time change, depending on factors related to the specific word as well as to the region. Furthermore, our models revealed that more frequent use and more favorable attitudes towards StG trigger longer VOT. This effect is particularly interesting against the backdrop of the special diglossic situation in German-speaking Switzerland and suggests that linguistic practices and attitudes can be important motivators of sound change.

Keywords: VOT, aspiration, voiceless stops, sound change, dialect-standard convergence

1. INTRODUCTION

The time interval between the release of a stop and the onset of periodicity in the following vowel (voice onset time; VOT) is a distinguishing typological feature. Across different languages, VOT has been found to increase as a function of front to back in place of articulation ([1–3]; but cf. also [4]). Further, VOT may vary in terms of prosodic features related to the position in a phrase [5, 6] or lexical frequency [6–8].

Two recurring sociolinguistic factors which have been found to cause variation in VOT are age and linguistic repertoire, encompassing the range of a person's linguistic competences and resources. Concerning age, previous studies have yielded mixed results [e.g. 3, 9]. Regarding linguistic repertoire, interesting insights come from [10], who reported convergence trends in VOT in the spontaneous code-switching of Spanish-English bilinguals (cf. also [11]).

In Standard German, VOT has been described as the most important cue to contrast between word-initial voiceless and voiced stops [12][1]. In German dialects spoken in Switzerland, however, fortis and lenis are traditionally not contrasted by VOT, but by closure duration and potential F0 effects in the following vowel, as shown in [13] (cf. [4, 14] for similar phenomena in other German varieties and [3, 15] for other dialect–standard contexts). The following example illustrates the /d, t/ contrast in both varieties in *danken* 'to thank' vs. *tanken* 'to refuel':

- Standard German (StG): [ˈd̥aŋkn̩] vs. [ˈtʰaŋkn̩]
- Swiss German (SwG): [ˈd̥ɑŋk͡xə] vs. [ˈtɑŋk͡xə]

Thus, except for some rare minimal pairs (cf. e.g. [13]), word-initial /p, t/ are generally unaspirated in SwG.[2] In recent years, however, it has been argued that the dialects might show convergence trends to StG in that – depending on a set of intra- and extralinguistic factors – longer VOTs in word-initial bilabial and alveolar fortis stops may occur.

The main *intralinguistic* factors driving this change are assumed to be related to lexical criteria. For example, [8] report aspirated stops in loan words from StG (e.g. [tʰeː] 'tea') as well as from English (e.g. [ˈpʰɑrti] 'party', cf. also [16]), and in proper nouns (e.g. [pʰɑʊl] 'Paul'). Based on an analysis of grammars, linguistic atlases, and personal field notes, [2] also documents more aspirated forms in proper nouns, referring not only to persons, but also to places (e.g. *Thun* [ˈtʰuːn]). In relation to the last example, [2] argues that the graphic representation <Th> might also trigger aspiration. Furthermore, according to [13], the letter names of <p> and <t> are nowadays usually aspirated. Due to the strong lexical determination of aspiration in SwG, [13] even discuss a potential slow lexical diffusion of a sound change which might lead to a phonological contrast as in StG or English in the distant future.

Regarding *sociolinguistic* variation in VOT in SwG, regional stratification has been investigated to some degree, while other factors have remained empirically uncharted. Previous reports point towards a North–South divide with speakers from urban, Northern regions exhibiting longer VOTs than those from the Alpine South [2, 17]. Besides region, [2] further speculates that aspiration may convey cosmopolitanism, urbanity, and education, which leads to the conclusion that style and social meaning

may be even more important than regional origin in explaining variation in VOT.

In sum, while VOT has been broadly investigated in various languages, empirical evidence for variation and change due to sociolinguistic factors in SwG is thin. The goal of this study is to bridge this gap by investigating VOT in word-initial /p, t/ in a controlled sample of 1000 speakers from 125 localities, taking into account large-scale metadata. The special diglossic situation and the atypically high prestige attributed to the dialect (cf. e.g. [18]) further make SwG an interesting context in which to explore effects of linguistic practices and attitudes on VOT.

2. METHODS

The current study draws on data from the SDATS corpus [19]. Its materials, speakers, and procedures are detailed in sections 2.1–2.3.

2.1. Materials

Table 1 presents the four stimuli selected for the current study, two with a word-initial /p/ and two with a /t/, in both isolated and phrase-embedded contexts.

Item	Place o.A.	Context
Parlament	bilabial	isolated
Patrick	bilabial	in phrase
Thermometer	alveolar	isolated
Tee	alveolar	in phrase

Table 1: Properties of the four VOT stimuli.

Besides place of articulation and co-articulation context, the items were selected due to lexical properties discussed in the literature: first, they vary in lexical frequency. Second, *Tee* was chosen as a typical example of a StG loanword, as opposed to *Parlament* as a French borrowing; *Patrick* represents a proper noun, and with *Thermometer*, graphemic influences beyond place names can be tested.

2.2. Speakers

The SDATS database comprises a total of 1000 speakers from 125 localities in German-speaking Switzerland, balanced across age (four younger and four older speakers) and gender (two female and two male speakers per cohort) in each locality. All participants are autochthone speakers of the local dialect with a comparatively low long-term mobility and are competent and daily users of (Swiss) StG, while their English skills vary.

For the current analysis, 11 speakers were excluded due to missing data, reducing the sample to 989 speakers.

2.3. Procedures

Data collection took place in 2020 and 2021. The four VOT items were collected as part of 2–3h oral dialect interviews. *Parlament* and *Thermometer* were elicited via picture-naming tasks, while *Patrick* and *Tee* were embedded in sentence translation tasks.

VOT was coded manually in Praat [20]. The period of aperiodic friction between stop release and the initiation of periodicity was identified based on acoustic and visual inspection of the oscillogram and spectrogram. To normalize VOT for articulation rate, the time between stop release and the end of the following vowel (RVT) was coded and normVOT was calculated as the relative proportion of RVT (normVOT = VOT/RVT*100).

Statistical analyses were conducted in R [21]. After descriptive and visual data inspection, mixed effects modeling was performed using lme4 [22] with normVOT as outcome and the predictors listed in Table 2 as fixed effects. Random intercepts were allowed for speakers. We started with a model comprising all main effects plus potential two-way interactions, of which only those reaching $p<.01$ were retained in the final model.

Fixed Effects	Specifications
Item	4 levels, cf. Table 1
Age Cohort	2 levels: older (60+), younger (20–35)
Region	8 levels, clustered based on [23]
DSP	numeric -10 to 10: Dialect Standard Profile (cf. below)
EN Competence	Numeric from 1–7, based on self-assessments
Education	4 levels: (voc.) baccalaureate, secondary voc. education, tertiary voc. education, university degree
Gender	2 levels: female, male

Table 2: Structure of fixed effects. EN=English, voc.=vocational.

A note on the DSP in Table 2: the Dialect Standard Profile (DSP) was designed following the principles of the Bilingual Language Profile (BLP [24]), an established language dominance index for bilingual speakers. Analogous to the BLP, the DSP was based on questionnaire data about productive (e.g. varieties spoken with family, friends, etc.) and receptive use (e.g. TV, radio), as well as attitudes towards both varieties (for more details, cf. analysis report: https://osf.io/rq5gk/).

3. RESULTS

Our final model revealed significant effects for item, age, region, DSP, and education, as well as item*age and item*region interactions (the full model output can be consulted in the analysis report, cf. link in 2.3). All effects are presented in more detail in sections 3.1–3.4. Figures 1–3 do not represent raw data, but predicted values based on the final model with all fixed and random effects taken into account (formula: normVOT ~ Item + Age Cohort + Region + DSP + EN Competence + Education + Gender + Age Cohort:item + item:Region + (1 | Speaker)).

3.1. Item and age

In general, normVOT was higher in the younger cohort (1.84(±0.61), t=3.03, p<0.01) and higher for alveolar compared to bilabial stops (most pronounced in the case of *Thermometer* compared to *Parlament*, 23.26(±.93), t=25.11, p<0.001).

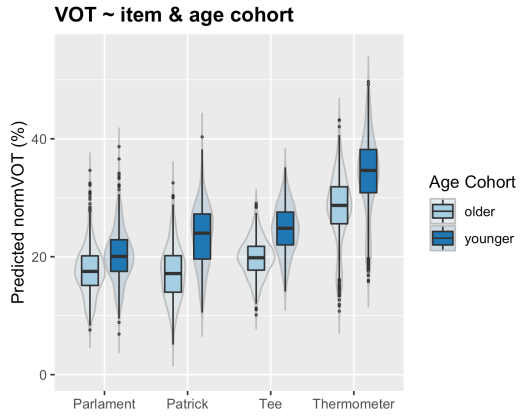


Figure 1: Predicted VOT according to item and age.

However, age and item interacted, as shown in Figure 1: while only slight age-related differences were identified in *Parlament*, the gap widened in *Tee* and *Thermometer* and was the biggest in the case of *Patrick*, where the highest between-subject variation within the younger cohort was also observed.

3.2. Item and Region

Figure 2 illustrates the interaction between item and region. With regard to *Parlament*, the North-Western region showed the highest predicted normVOT, and for the remaining territory, a slight increase from west to east was revealed. While *Patrick* also exhibited the highest values in the North-West, normVOT was predicted to be higher in central regions compared to the Alpine South. Almost no regional differences were identified for *Tee*. A sharp North–South contrast, paired with the highest normVOTs overall, was revealed for *Thermometer*.

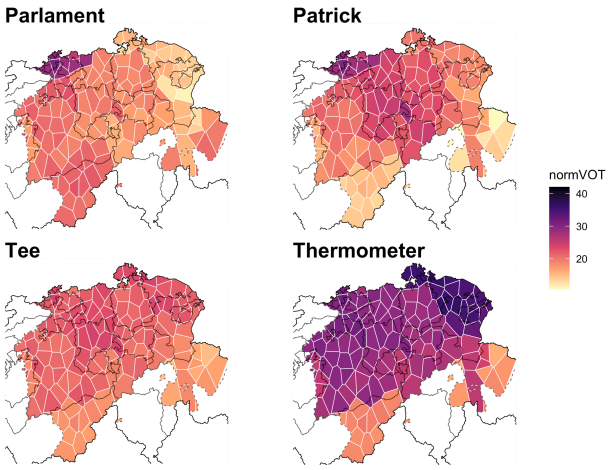


Figure 2: Predicted VOT according to region and item. Darker colors suggest higher values.

3.3. Language attitudes and use

The effect for DSP (cf. Figure 3) can be read as follows: The stronger the dialect orientation in terms of attitudes and use, the lower the normVOT (-0.23(±0.10), t=-2.41, p=0.016). Higher normVOT, in turn, was associated with more positive attitudes towards and more frequent use of StG.

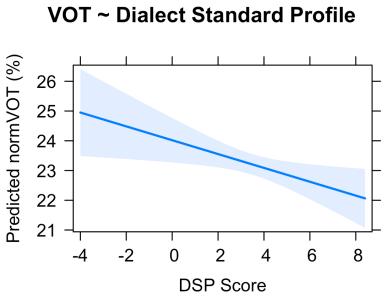


Figure 3: Predicted VOT according to DSP.

Further, our model revealed a positive trend slightly above the p=.05-level for higher English competence increasing normVOT (0.26(±0.14), t=1.86, p=0.063).

3.4. Education and gender

While no gender effects were detected, speakers with a (vocational) baccalaureate were shown to exhibit slightly higher normVOTs compared to those with other educational backgrounds; the effect was most pronounced in comparison with speakers with secondary vocational education (-1.14(±0.50), t=.2.27, p=0.024).

4. DISCUSSION

4.1. Intralinguistic effects

Our study corroborates previous findings in that VOT is longer in alveolar than in bilabial stops [1–3]. Furthermore, the notably longer VOTs in *Thermometer* support the argument presented in [2], showing that the graphemic representation may trigger aspiration, regardless of whether it is a place name or not. Our study is also in line with previous work regarding longer VOTs in lexical borrowings from StG, and – comparing the two bilabial stops – longer VOTs for *Patrick* as a proper noun [8, 13] (cf. also interaction with age discussed in section 4.2 below). Lastly, frequency effects discussed in [6–8] might be at play, since the longest VOTs were identified for the least frequent word according to [25] (*Thermometer*; 3/7) and consecutively longer VOTs for the more frequent *Tee* (4/7) and *Parlament* (5/7) (*Patrick* was not integrated since no frequency statistics existed for this proper noun). This result suggests that, while more commonly used words are pronounced in a rather traditional manner, less typical words might trigger more standard-like, aspirated forms.

While these are the main factors discussed in previous literature, they only account partly for variation in VOT and further specificities related to the particular words might be at play (e.g. syllabic structure or position, cf. also 4.3. limitations).

4.2. Sociolinguistic effects

While age effects generally showed the expected standard-convergent apparent-time trend and did not interact with other social variables (as e.g. in [3, 9]), the age*item interaction requires further discussion. The fact that the biggest age difference as well as the highest variability among the younger speakers were related to *Patrick* suggests that – in line with [2, 8] – the ongoing change might be most obvious in proper nouns. The reason for the least age-related differences being found in *Parlament* might be twofold: The high frequency discussed in section 4.1 above might not only trigger the shortest VOT, but also the slowest change. Second, the short-lag VOT might also be explained by the fact that this lexeme was borrowed from French rather than StG or English.

Regarding regional variation, the North–South divide documented in [2, 17] was replicated for *Thermometer* and – to some degree – for *Patrick*, whereas the other items exhibited different regional distributions. Similar VOTs across all regions in *Tee* suggests that, while the first instances of aspirated forms were reported at the beginning of the 20th century [2], this sound change has advanced to all regions in German-speaking Switzerland. At the same time, the apparent-time effect indicates that it has not yet completed. A surprising regional effect was related to lower VOTs in *Parlament* in the East, because standard convergence trends are usually most obvious in these regions.

The effect related to the Dialect Standard Profile showed that cross-linguistic influences in VOT may not only be at play among different *languages* in a repertoire (as in [10, 11]), but also among different *varieties*. For example, a person who frequently watches StG TV shows and who feels comfortable speaking the standard language may aspirate word-initial stops more often. More generally speaking, this effect suggests convergence to StG for speakers with more favorable attitudes towards and more frequent productive and receptive use of the standard language. The importance of this index as a predictor in our model, as well as the identified positive trend for English competence, suggests that it would be worthwhile for future studies to investigate influences of language competence, practices, and attitudes in dialect–standard situations more deeply.

Lastly, while – as expected – no differences between male and female speakers arose, the effect for education is surprising to a certain extent, given that SwG is usually not expected to vary in terms of educational background. However, this effect reflects an age trend, i.e., speakers with a (vocational) baccalaureate in our study are also the youngest ones.

4.3. Limitations and conclusion

The main limitation of the current study is that more items would be needed to allow for a more in-depth investigation of the phonetic environment. Furthermore, the four items were coded by four different people, and although the coding procedure was objective and problematic cases were cross-checked by a second coder, influences on between-item variation cannot be ruled out completely.

Notwithstanding these limitations, this is the first large-scale multilocality study on variation and change in VOT in SwG, yielding generalizable results due to its big sample size, controlled speaker selection, and supervised data collection. We showed that increasing VOT in word-initial stops can be seen as a standard-converging sound change in progress, which, on the one hand, may heavily depend on word-specific characteristics, and, on the other hand, is influenced by the speakers' age, regional origin, and their linguistic practices and attitudes. However, if this phonetic change may lead to a potential standard-like phonological contrast, as discussed in [13], remains open for future research.

5. ACKNOWLEDGEMENTS

This research was funded by the Swiss National Science foundation (Grant no. PCEFP1_181090).

6. REFERENCES

[1] Cho, T., Ladefoged, P. 1999. Variations and universals in VOT: Evidence from 18 languages. *J. Phon.*, vol. 27, 207–229.

[2] Schifferle, H. 2010. Zunehmende Behauchung. Aspirierte Plosive im modernen Schweizerdeutsch. In: Christen, H., German, S., Haas, W., Montefiori, N., Ruef, H. (eds), *Alemannische Dialektologie: Wege in die Zukunft. Beiträge zur 16. Tagung für alemannische Dialektologie in Freiburg/Fribourg vom 07.–10.09.2008*, 43–55.

[3] Stuart-Smith, J., Sonderegger, M., Rathcke, T., Macdonald, R. 2015. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Lab. Phonol.* 6(3–4), 505–549. doi: 10.1515/lp-2015-0015.

[4] Hödl, P. 2019. Perception of voice onset time by Austrian and German listeners, in *Proc. 19th ICPhS* Melbourne, 3686–3690.

[5] Cole, J., Kim, H., Choi, H., Hasegawa-Johnson, M. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *J. Phon.*, 35(2), 180–209. doi: 10.1016/j.wocn.2006.03.004.

[6] Yao, Y. 2009. Understanding VOT Variation in Spontaneous Speech. *UC Berkeley PhonLab Annu. Rep.* 5(5), 29–43.

[7] Yu, A. C., Abrego-Collier, C., Sonderegger, M. 2013. Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PloS One* 8(9), e74746.

[8] Fleischer, J., Schmid, S. 2006. Zurich German. *J. Int. Phon. Assoc.* 36(2), 243–253. doi: 10.1017/S0025100306002441.

[9] Torre, P., Barlow, J. A. 2009. Age-related changes in acoustic characteristics of adult speech. *J. Commun. Disord.* 42(5), 324–333. doi: 10.1016/j.jcomdis.2009.03.001.

[10] Balukas, C., Koops, C. 2015. Spanish-English bilingual voice onset time in spontaneous code-switching. *Int. J. Biling.* 19(4), 423–443. doi: 10.1177/1367006913516035.

[11] Sancier, M. L., Fowler, C. A. 1997. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *J. Phon.* 25(4), 421–436.

[12] Beckman, J., Jessen, M., Ringen, C. 2013. Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *J. Linguist.* 49(2), 259–284. doi: 10.1017/S0022226712000424.

[13] Ladd, D. R., Schmid, S. 2018. Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops. *J. Phon.* 71, 229–248. doi: 10.1016/j.wocn.2018.09.003.

[14] König, W. 2001. Ein Lautwechsel T>K. Oder wie die Tartoffel zur Kartoffel wurde. *Ger. Pragensia* XVIII, 95–101.

[15] Scobbie, J. M. 2006. Flexibility in the face of incompatible English VOT systems. *Lab. Phonol.* 8, 367–392.

[16] Dalcher, P. 1986. Anglicisms in Swiss German: The evaluation by computer of a survey conducted in 1964/5. In: Viereck, W., Bald, W.-D. (eds), *English in Contact with other Languages*. Budapest: Akadémiai Kiadó, 179–206.

[17] SDS = Hotzenköcherle, R., Baumgartner, H. (eds). 1962–2003. *Sprachatlas der deutschen Schweiz*. Bern / Basel: Francke.

[18] Berthele, R. 2004. Vor lauter Linguisten die Sprache nicht mehr sehen – Diglossie und Ideologie in der deutschsprachigen Schweiz, In: Christen, H. (ed), *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum. Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen*. Wien: Praesens, 111–136.

[19] Leemann, A., Jeszenszky, P., Steiner, C., Messerli, J., Studerus, M. 2020. *SDATS Corpus – Swiss German dialects across time and space*. [Online]. Available: osf.io/s9z4q.

[20] Boersma, P.,Weenink, D. 2022. *Praat: doing phonetics by computer*. Version 6.3.02. [Online]. Available: http://www.praat.org/.

[21] R Core Team. 2022. *R: A language and environment for statistical computing*. Version 4.2.2. R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/.

[22] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1), 1–48. doi: 10.18637/jss.v067.i01.

[23] Scherrer, Y., Stoeckle, P. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectol. Geolinguistica* 24(1), 92–125.

[24] Birdsong, D., Gertken, L. M., Amengual, M. 2022. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. Center for Open Educational Resources and Language Learning (COERLL), University of Texas Austin. [Online]. Available: https://sites.la.utexas.edu/bilingual/.

[25] DWDS = Berlin-Brandenburgischen Akademie der Wissenschaften (eds). 2022. *DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. [Online]. Available: https://www.dwds.de/.

[1] Findings from [4, 13], however, suggest that the reality is more complex than a mere dichotomy based on VOT.

[2] Velar stops are not considered in this case since they follow different dialectal distributions, e.g. word-initial /k/ is realized as [χ] in most SwG dialects.

STOP VOICING IN DREHU: EFFECTS OF PLACE OF ARTICULATION, SPEAKER SEX, AND LANGUAGE ATTITUDES

Catalina Torres[1], Jeremy Steffman[2]

[1]University of Zurich, [2]The University of Edinburgh
[1]catalina.torres@ivs.uzh.ch, [2]jeremy.steffman@ed.ac.uk

ABSTRACT

This study investigates a potential ongoing sound change in the language Drehu [d͡ʒehu], spoken by a small community in the South Pacific. We focus on the voiceless and voiced stop series in the language. Data from teenage female and male speakers was analysed in relation to voice onset time (VOT), percent voicing, and self reported language proficiency and attitudes. We find that the velar stops show substantial aspiration and de-voicing, led by female speakers. The voiceless velar stop shows the greatest VOT duration whereas voiced velars display the greatest devoicing. Interestingly, female and male speakers' increased use of VOT and devoicing are differently correlated to their language attitudes. This shows that in this small and rural community, the same features may have different socio-linguistic associations.

Keywords: stop voicing, voice onset time, Oceanic languages, language attitudes, sound change

1. INTRODUCTION

Drehu is an Austronesian language spoken by a population of around 15000 on the island of Lifou in New Caledonia. Like closely related languages Nengone and Iaai, Drehu is an Oceanic language of the Southern Melanesian group of the Loyalty Islands (LI) [1]. Grammatical descriptions [2, 3, 4, 5] have documented the phonological systems of LI-languages. However, their phonetic structure remains under-studied. Only few studies have dealt with their segmental inventories, thus showing a need for more instrumental work [6, 7]. Arguably, in Lifou, there are increasing linguistic pressures arising from bilingualism (with French) [8] and language contact, potentially influencing Drehu.

Drehu's phonetic structures are the perhaps best studied, especially considering its prosodic system [9, 10, 11]. Little is known on the acoustics of segments in the language, but a more recent study showed interesting results regarding earlier impressionistic descriptions. The previously documented retroflex stops (/ɖ, ʈ/) [12] were not confirmed in acoustic investigations [6] which instead reported affricate consonants [d͡ʒ, t͡ʃ]. An investigation a putative word-initial stress pattern [2, 13] failed to find evidence in favour of it, reporting a word-final prominence pattern instead [14]. While instrumental studies of Drehu have provided acoustic evidence deviating from early phonological analyses, the reasons for these observations are poorly understood. It is difficult to ascertain whether the results of these studies are due to prior description carried out impressionistically by fieldworkers without precise or quantitative approaches [15] or whether they are the result of ongoing sound changes. As [15] points out, small speech communities differ from large communities in that they have different social structures possibly leading to different phonetic and cognitive influences on sound change. This could, for example, influence the way sound change diffuses when there are fewer speakers in the community. This study investigates stop consonants in Drehu. We focus on the speech production of young (teenage) bilingual speakers while considering self reports on language proficiency and attitudes. With this we hope to contribute to a better understanding of how a potential sound change originates and spreads in a speech community of a small size while considering its particular ecology.

1.1. Stop consonants in the Loyalty Islands

Modern day Drehu is described as having a series of phonologically voiceless /p, t, k/ and voiced stop consonants /b, d, g/. Historically, the series of voiceless preceded that of voiced stops which were largely introduced through contact with neighbouring languages and borrowings from English [2, 12]. Regarding their phonetic characteristics, the bilabial /p/, alveolar /t/, and velar /k/ stops are considered to be voiceless and unaspirated, although [2] notes that voiceless

stops can sometimes be voiced intervocalically. Apart from an impressionistically recorded voicing distinction, no other feature is said to contrast the voiceless and voiced stop consonants. Since voiced stops were introduced to Drehu mostly through borrowings this explains the rather asymmetrical distribution between the voiceless and voiced stops, whereby voiceless stops are more common. One study [7] reports a series of voiceless /p, t̪, k/ and voiced /b, bʷ, d̪, g/ stop consonants in Iaai. Interestingly, the authors report evidence for retroflex stops (/ɖ, ʈ/) based on linguagrams. There are no instrumental studies dealing with the segmental inventory of Nengone but a relatively recent summary collected by a native speaker [16] provides an updated assessment. In Nengone, in addition to the voiced /b, d, g/ and voiceless /p, t̪, k/ contrast, there is a series of voiceless aspirated stops /pʰ, t̪ʰ, kʰ/. Additionally, the retroflex stops (/ɖ, ʈ/) also seem to be preserved. Finally a word on French, the contact language: lead voicing is used in voiced stops /b, d, g/ and voiceless stops aren't aspirated /p, t, k/ [17], though they can have lag VOT up to approximately 50 ms in some cases [18].

2. RESEARCH AIMS

We examine the acoustics of a series of voiceless /p, t, k/ and voiced /b, d, g/ stops. We focus on the measurements (i) voice onset time (VOT) and (ii) percentage of voicing during closure. We explore how place of articulation and speaker sex impact voicing and aspiration. We further investigate how self reported language proficiency and attitudes relate to acoustic realisations for female and male participants, suggesting a possible change in progress.

3. METHODS

3.1. Participants

Eight (4 female) teenage speakers (15-17 years) of Drehu took part in a speech production experiment. All are bilingual in French and Drehu, reported growing up speaking both languages from childhood, and self identify as Kanak. Participants were recruited at a local high school on the island of Lifou, where the speakers reported originating from. This is the only high school in the LI meaning that speakers of Iaai and Nengone are also represented.

3.2. Materials and procedures

A list of words, from which target tokens were selected, was created with the help of a native Drehu language expert. Participants were recorded individually performing a controlled reading task using experimental materials written in Drehu only. Stimuli appeared on a slide as individual sentences written in Drehu orthography. Target tokens were placed in the frame sentence *eni a qaja x me x nge x hmaca* 'I say x and x and x again'. The order in which sentences appeared was randomised. A head mounted microphone together with a Zoom H6 recorder were used. Recordings were made at a sampling rate of 44.1 kHz and 24 bit. Participants responded to an adapted version of the BLP [19], a questionnaire that collects self reported evaluations of bilinguals language proficiency and attitudes and provides scores for these categories.

3.3. Data processing, analysis, and statistics

Recordings were manually transcribed and forced aligned using the web MAUS interface [20]. Boundaries for segments were hand corrected. To identify and mark VOT we used AutoVoT [21] which was first trained on a subset of the Drehu data and was then used to mark VOT in all the data. Percent voicing in voiced stops was measured using the voice report functionality in Praat [22].

We used Bayesian mixed-effects regression analyses with the R package brms [23]. We report the median estimate from the posterior and the "probability of direction" computed with bayestestR [24]. This metric indicates the percentage of the posterior with a particular sign, ranging between 50 (posterior centered directly on zero) and 100 (posterior entirely excluding zero). The measure helps quantify with what certainty can we say that an effect shows a particular directionality. We take pd > 95 to be robust evidence for an effect, though we consider values approaching this threshold to provide weaker evidence for effect existence (e.g., if pd = 92 we can be 92% sure that an effect exists with a particular directionality). Models were run with four chains, an adapt-delta value of 0.99, and 4,000 iterations per chain, and a burn-in period of 1,000 iterations.

4. RESULTS

4.1. VOT measurements

The VOT model was run with the log-normal family to account for the log-normal nature of the VOT

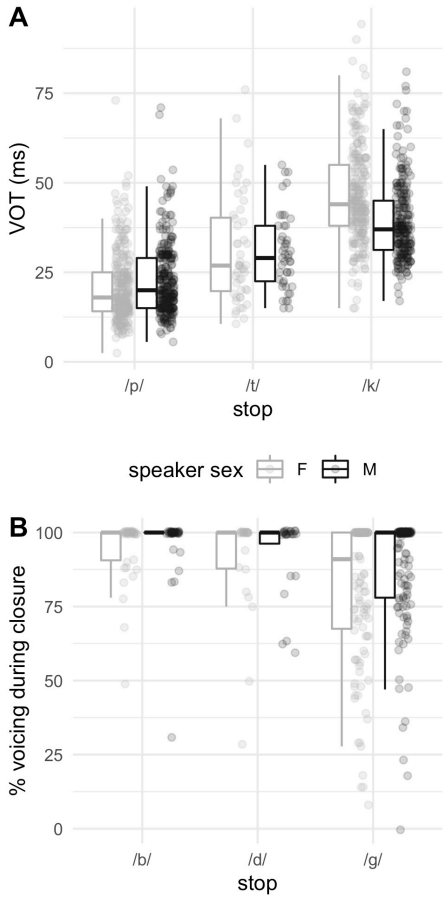


Figure 1: Top: VOT measurements in milliseconds for three voiceless stops /p, t, k/ for female (grey) and male (black) speakers. Bottom: % voicing for three voiced stops /b, d, g/ for female and male speakers.

data. With /p/ as the reference level in the model, there was a main effect of place of articulation, whereby /k/ had longer VOT (β = 0.73, pd = 100; β in logged values). There was also a credible difference between /t/ and /p/, with VOT in /t/ being longer (β = 0.35, pd = 100). We compared the VOT for /t/ and /k/ using the emmeans package [25] to compute the marginal effect between /t/ and /k/, finding that /k/ has credibly longer VOT than /t/ (β = 0.38, pd = 100). There was no main effect of sex in the model (pd = 82), however there was a credible interaction between sex and the place of articulation level /k/ (β = -0.26, pd = 100). There was no interaction between the place of articulation level /t/ and sex (pd = 80). As shown in Figure 2(A), this interaction comes from the fact that, only for /k/, female speakers produce longer VOT than male speakers. We examined the interaction further with emmeans, testing the effect of sex within each place of articulation, and finding that for /k/ there was some evidence for a difference across speaker sexes (β = 0.16, pd = 92), while for other places of articulation there was not evidence for a difference (pd = 82 for /p/, pd = 58 for /t/), cf. [26, 27].

4.2. Percent voicing

We used a zero-one inflated beta model for the % voicing data for /b, d, g/. This model family is appropriate for proportion or percent data in which there are some 0% and 100% measures. We used weakly informative priors. The data was converted to proportions, and was predicted as a function of place of articulation, speaker sex and their interaction. With /b/ as the reference level, there was a main effect of place of articulation, whereby /g/ showed a lower percentage of voicing during the stop (β = -0.53, pd = 98), with no credible evidence for a difference between /b/ and /d/ (pd = 88). Comparison with emmeans showed no credible difference between /d/ and /g/ as well (β = -0.15, pd = 75), indicating the /d/ is effectively in between /b/ and /g/ in terms of % voicing, with more devoicing clearly evident in /g/ versus /b/. There was no main effect of speaker sex (pd = 53), and only very weak evidence for an interaction of speaker sex with the /g/ place of articulation (β = 0.47, pd = 87). The directionality of the interaction is such that it suggests female speakers are producing more devoicing in /g/ than male speakers. Tests of the effect of sex within each place with emmeans corroborate this, showing weak evidence for a difference between female and male speakers in /g/ (β = -0.44, pd = 92), while for other places of articulation there was not evidence for a difference (pd = 52 for /b/, pd = 61 for /d/). Here we note that de-voicing generally occurred at the end of the stop closure, consistent with the aerodynamic voicing constraint [28].

4.3. Language proficiency and attitudes

A Pearson correlation coefficient was computed to assess the linear relationship between self-reported language proficiency in Drehu and duration in VOT. In the female speakers, there was a positive correlation between the two variables, (r(241) = .22, p = .0004). This correlation was only marginally present in the male speakers (r(160) = .04, p = .5 (n.s)). An additional correlation coefficient was calculated to assess the relationship between language attitudes in Drehu and % voicing. In the female speakers, there was a positive correlation between the two variables, (r(109) = 0.29, p = .002). The same test yielded an opposite result

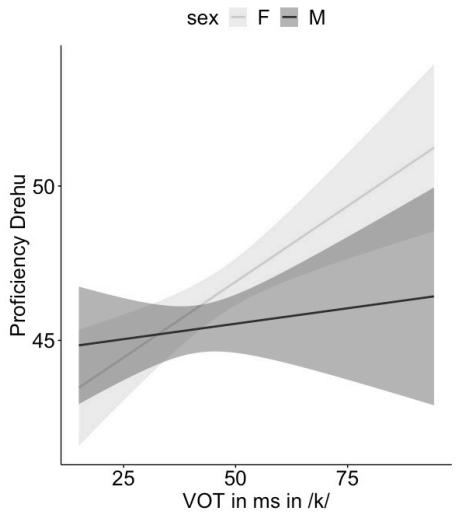


Figure 2: Linear relationship between language proficiency and VOT in /k/, for female (grey) and male (black) speakers.

for the male speakers (r(109) = -.32, p = .002). Finally, we were also interested in assessing whether language attitudes in French and % voicing could be correlated. There was a negative correlation, for the female speakers (r(109) = -.33, p = .0004). The test did not reach a significance level for male speakers in this case, although a negative trend was observed (r(86) = -.15, p = .1).

5. DISCUSSION AND CONCLUSION

This study examined the acoustic properties of voiceless and voiced stop consonants of teenage Drehu speakers. The results show that voiceless stops are produced with longer VOT, particularly /k/: mean VOT in /k/ is 47 ms for female speakers and 40 ms for males, making /k/ "slightly aspirated" in Cho and Ladefoged's taxonomy [29]. In parallel, we observe that phonologically voiced stops, particularly /g/, are sometimes partially devoiced. This may be evidence that cues to the contrast are changing, in line with the literature showing that sound change is a gradual phenomenon that first affects parts of the segmental inventory and not all segments of a series at once [30]. We find robust evidence for VOT in voiceless velar consonants /k/ to be longer than for other stops in the series /p, t/. Place of articulation is a possible factor influencing VOT duration, though notably [18, 31] find small and inconsistent differences between unaspirated /p/ and /t/. We find that young females show more aspiration and de-voicing. An interaction between place of articulation (for /k/) and female speakers was found showing that they produce longer VOT in the velar stop only. We further investigated the linear relationship of VOT with self reported language proficiency in Drehu. A positive correlation was found for female speakers. This suggests that female speakers who affirm having a higher proficiency in Drehu produce /k/ with longer VOT. This pattern was not found for male speakers. An analysis of % voicing in voiced stops showed that there is robust evidence for the voiced velar /g/ displaying greater devoicing than the bilabial and alveolar stops /b, d/. Effects of sex were also tested showing weak evidence that female speakers produce more devoicing than male speakers in the velar stop /g/ only, in parallel to longer VOT in /k/. An examination of a linear relationship between language attitudes towards Drehu and % voicing was carried out. In the male speakers we found a negative correlation whereby the more positively they view the Drehu language the less voicing they produce. There was a positive relationship in female speakers, suggesting that the more positively they view the language Drehu the more voicing they use. Additionally, in the female speakers it was found that there was a negative relationship between language attitudes in French and % voicing, showing that the more positively the speakers view the French language the less voicing they produce. Similar to observations in Western societies [32], we find that young females are leading a potential sound change. Although both young female and male speakers are involved in this potential sound change they seem to make somewhat different associations with the linguistic features involved. We find that in female speakers a better knowledge of Drehu is linked with increased VOT. Additionally, a positive attitude towards French shows a stronger correlation with devoicing of /g/ than a positive attitude towards Drehu. In contrast, male speakers show greater devoicing in relation to more positive language attitudes towards Drehu. Considering that French canonically uses lead voicing in voiced stops and short lag VOT for voiceless stops, this suggests speakers could be marking a stronger difference between Drehu and French. We speculate that aspiration, which is present in the neighbouring language Nengone, could be viewed as a more 'Kanak' feature whereas voicing could mark a more 'European' feature. It remains to be seen how speakers produce these stop series in French and how this aligns with their language attitudes. Further comparison to older speakers will be necessary to confirm our hypothesis of a change in progress. Finally, a more nuanced approach that considers social factors beyond sex will be of relevance too.

6. REFERENCES

[1] T. Crowley, J. Lynch, and M. Ross, *The Oceanic languages*. Routledge, 2011.

[2] D. T. Tryon, *Dehu grammar*. Australian National University, 1968.

[3] ——, *Nengone grammar*, ser. Pacific Linguistics, Series B. Canberra: Australian National University, 1967, vol. 6.

[4] ——, *Iai Grammar*, ser. Pacific Linguistics, Series B. Canberra: Australian National University, 1968, vol. 8.

[5] F. Ozanne-Rivierre, *Le iaai: langue mélanésienne d'Ouvéa, Nouvelle-Calédonie: phonologie, morphologie, esquisse syntaxique*. Peeters Publishers, 1976, vol. 20.

[6] J. Monnin, "Influence de la langue ambiante sur l'acquisition phonologique: une comparaison du français et du drehu," Ph.D. dissertation, Université de Grenoble; Université de Nouvelle Calédonie, 2010.

[7] I. Maddieson and V. B. Anderson, "Phonetic structures of iaai," *UCLA Working Papers in Phonetics*, vol. 87, pp. 163–182, 1994.

[8] C. Torres, J. Fletcher, and G. Wigglesworth, "Fundamental frequency and regional variation in Lifou French," *Language and Speech*, 2020.

[9] ——, "Investigating word prominence in Drehu," in *Proceedings of the 17th Australasian International Speech Science and Technology Conference*, 2018, pp. 141–144.

[10] C. Torres, "Pitch range modulations in an edge-marking language," *Proceedings of Speech Prosody 2022*, pp. 832–836, 2022.

[11] C. Torres and J. Fletcher, "The alignment of F0 tonal targets under changes in speech rate in Drehu," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2947–2958, 2020.

[12] C. Moyse-Faurie, *Le drehu, langue de Lifou (Iles Loyauté). Phonologie, morphologie, syntaxe*. Langues et Cultures du Pacifique Ivry, 1983.

[13] M.-H. Lenormand, "La phonologie du mot en lifou (îles loyalty)," *Journal de la Société des Océanistes*, vol. 10, no. 10, pp. 91–109, 1954.

[14] C. Torres and J. Fletcher, "Phrase-level and edge marking in Drehu," *Glossa: a journal of general linguistics*, vol. 7, no. 1, 2022.

[15] G. Zellou, C. DiCanio, A. Pycha, and A. C. L. Yu, "Introduction to sound change in endangered or small speech communities," *Linguistics Vanguard*, vol. 8, no. s5, pp. 491–497, 2022. [Online]. Available: https://doi.org/10.1515/lingvan-2021-0095

[16] S. Bearune, "L'expression linguistique de l'espace en nengone (Nouvelle-Calédonie)," Ph.D. dissertation, Paris, INALCO, 2012.

[17] C. Fougeron and C. L. Smith, "Illustration of the ipa: French," *Journal of the International Phonetic Association*, vol. 23, pp. 73–76, 1993.

[18] T. M. Nearey and B. L. Rochet, "Effects of place of articulation and vowel context on vot production and perception for French and English stops," *Journal of the International Phonetic Association*, vol. 24, no. 1, pp. 1–18, 1994.

[19] L. M. Gertken, M. Amengual, and D. Birdsong, "Assessing language dominance with the Bilingual Language Profile," *Measuring L2 proficiency: Perspectives from SLA*, pp. 208–225, 2014.

[20] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[21] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.52)[computer program]," 2017.

[23] P.-C. BÃ¼rkner, "Advanced Bayesian Multilevel Modeling with the R Package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018. [Online]. Available: https://doi.org/10.32614/RJ-2018-017

[24] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework," *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019.

[25] R. Lenth, "emmeans: Estimated marginal means, aka least-squares means. R package version 1.7.1-1," 2021.

[26] H. M. Hanson, *Glottal characteristics of female speakers*. Harvard University, 1995.

[27] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, and T. Yamasoba, "Age-and gender-related difference of vocal fold vibration and glottal configuration in normal speakers: analysis with glottal area waveform," *Journal of Voice*, vol. 28, no. 5, pp. 525–531, 2014.

[28] J. J. Ohala, "Aerodynamics of phonology," in *Proceedings of the Seoul International Conference on Linguistics*, vol. 92. Linguistic Society of Korea Seoul, 1997, p. 97.

[29] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *Journal of phonetics*, vol. 27, no. 2, pp. 207–229, 1999.

[30] J. Kuang and A. Cui, "Relative cue weighting in production and perception of an ongoing sound change in southern yi," *Journal of Phonetics*, vol. 71, pp. 194–214, 2018.

[31] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[32] W. Labov, "The intersection of sex and social class in the course of linguistic change," *Language variation and change*, vol. 2, no. 2, pp. 205–254, 1990.

SIMULTANEOUS VOWEL CHANGES IN BRITISH COLUMBIAN ENGLISH

Amanda Cardoso, Suyuan Liu, Robert Pritchard, Molly Babel

University of British Columbia
amanda.cardoso@ubc.ca, suyuan.liu@ubc.ca, bob@mail.ubc.ca, molly.babel@ubc.ca

ABSTRACT

System-wide vowel investigations acknowledge the interrelationship of sounds, and recognize that changes often involve multiple sounds within the systems. When vowels change together this may be driven by structural relationships and/or social meaning; the cause is not always apparent. We introduce the term simultaneous changes to describe vowels changing together without implying a causal relationship. We explore whether simultaneous changes in the British Columbia English vowel system are present across apparent time using a novel method involving Generalized Additive Mixed Models and Principal Component Analysis [1]. While we find evidence of simultaneous changes, where adjacent vowels with structural relationships pattern together, we also see non-adjacent vowels co-patterning. Framing sound co-patterning as simultaneous changes allows us to sidestep the insurmountable challenges of diagnosing the impetus of these changes.

Keywords: chain shifts, vowel systems, English

1. INTRODUCTION

Vowels do not behave independently. Rather, individual vowels are part of a larger structured system. Within a system, small individual changes may have knock-on effects that lead to changes in the wider system over time. Vowel patterns may shift over time, moving into new phonetic space or impinging on other vowel categories, which may result in mergers of categories or additional changes within the vowel space. Sets of vowel changes that have been causally linked (e.g., through structural relationships) have been dubbed push or pull chain shifts, depending on whether the first change in the shift was one that "pushed" another category into new phonetic space or "pulled" another category into a vacated position [2, 3]. Other structurally-linked vowel changes are suggested to happen in parallel, driven by shared features, such as the fronting of GOAT and GOOSE vowels in North American Englishes [4]. There is also evidence that non-adjacent, seeming unrelated changes occur. For example, the fronting of GOOSE and the lowering of DRESS have been observed in, for example, California English [5, 6]. Generally, providing strong evidence for the changes' impetus is difficult (e.g., [7, 6]. We introduce the term simultaneous changes to convey the coordination of such changes without implying a mechanistic connection between the vowels. Their coordinated behavior may be rooted in either shared social meaning or phonetic/phonological structure. While these patterns are all observable from a bird's eye view of the system, the quantification of change on the system level has been a challenge, often amounting to simple correlations between variables.

A method for studying vowels together as a system in a more unified quantitative framework is introduced in [1]; see the methodological details in 2.2. In an exploratory analysis, we apply these methods to 11 monophthongs in British Columbia English (henceforth BCE), the western-most variety of Canadian English. The BCE vowel space is ideal for investigating changes involving the coordination of sets of vowels, as it has been described to be undergoing two sets of changes in the vowel space that involve multiple vowels: the Canadian Vowel shift (hereafter CVS [8, 9, 10]) and a set of changes in the front lax vowels preceding /ɡ/ and /ŋ/ [11, 12]. The changes simultaneous with the CVS are lowering and/or retraction of KIT, DRESS, and TRAP, and fronting or centralization of STRUT, with the impetuous for the CVS suggested to be the merger of THOUGHT and LOT. KIT, DRESS, and TRAP have also been suggested to be undergoing raising, fronting, and/or diphthongization before voiced velars. In other words, the same "vowels" (but different allophones) are participating in different changes. The fact that there have been such strong suggestions of coordinated behaviour of vowels in the BCE vowel system makes this an ideal place to explore system-wide co-variation of vowel categories. We examine to what extent these changes are simultaneous changes – that is, are they patterning seemingly in concert? To be clear, the goals of this set of analytic methods is for identifying patterns of co-variation in vowel formant structure, as opposed to providing

descriptions of specific vowel characteristics.

2. METHODS

2.1. Data

Data were collected via a browser plugin that saved recordings as .wav files as part of the Determining Regional Accents With Literature (DRAWL) project (https://blogs.ubc.ca/drawl/). This project asks participants from British Columbia, Canada to read a short story composed to include tokens of interest for local regional variation. Audio files were high quality at 44.1kHz sampling rate and 16 bit rate, but subject to variation in individuals microphones, sound cards, and recording environment. Here we analyze the data from 170 participants. Participants self-categorized their "voice type" as female or male (126 females, 44 males) and provided age (in years: M = 41.38, SD = 15.97, range = 14-78).

Recordings were orthographically transcribed, force-aligned with the Montreal Forced Aligner [13], and hand-checked for major errors. Files were downsampled to 16kHz and formants were estimated using emuR [14]. Formants were estimated separately for female and male voices by changing the gender parameter, which changes the nominal F1 value from 500 Hz (for male) to 560 Hz (to female) and affects the formant frequency range table [15].

F1 and F2 midpoint values for stressed FLEECE, KIT, FACE, DRESS, TRAP, GOOSE, GOAT, NURSE, STRUT, THOUGHT, and LOT were normalized for apparent talker size using the [16] algorithm. Local speech rate was estimated for each vowel using the average segment duration in the word in which the vowel occurred.

2.2. Analysis

Our analysis follows closely the methods described in [1], and we direct readers to that paper for a more verbose description of the methods. The basic steps of this analysis are as follows: 1) separate GAMMs for each formant by vowel category are run with random smooths for speakers and words; local speech rate is a control. Models include smooth terms using additive smooths for speech rate, reader age, and reader age by voice type. There were smooth and parametric terms for voice type. 2) Speaker intercepts are extracted from the models; 3) a PCA is run on the speaker intercepts taken from each of the vowel formant GAMM models.

The vowels included in the analysis are: FLEECE, KIT, FACE, DRESS, TRAP, GOOSE, GOAT, NURSE, STRUT, THOUGHT, and LOT. KIT, DRESS and TRAP before voiced velars are coded separately with BIG, EGG, and BAG, respectively as their lexical sets. After outlier removal processes, 44,217 tokens were analyzed.

3. RESULTS

Figure 1 visualizes the trajectories of the vowel categories over apparent time. The arrows start from the estimated values of the oldest speakers and end at the estimated values of the youngest speakers. From this we can observe that KIT, BIG, DRESS, and EGG are mostly lowering, while BAG is raising and fronting. We also see fronting of GOOSE, and backing of GOAT, LOT, and THOUGHT. In the middle of the vowel space we see NURSE also appears to be lowering. We see the maintenance of an allophonic contrast between BIG and KIT and EGG and DRESS, despite the observation that these vowels are lowering over time. While the CVS is often characterized by a backing of TRAP, that is not observed, which may suggest it was already backed by our oldest speakers. Contrary to descriptions, we also see LOT and THOUGHT midpoints are distinct, and STRUT and LOT are close together in the F1 by F2 space for older speakers.

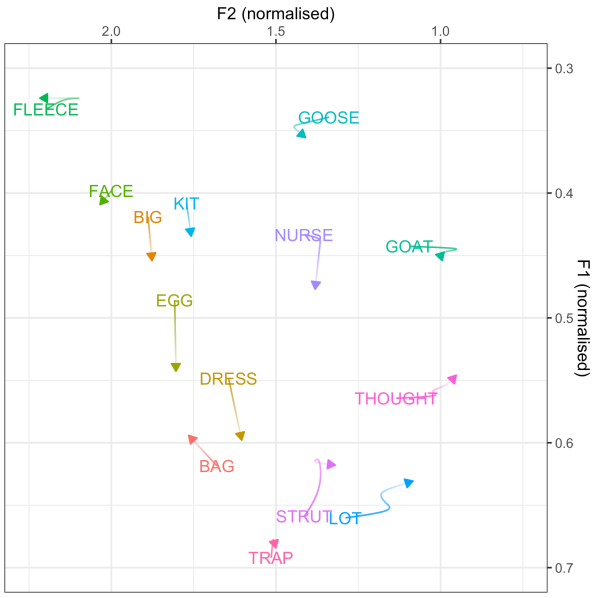


Figure 1: Trajectories of the vowels under investigation based on the GAMM smooths of reader age (older speakers = labels; younger speakers = arrows).

The group-level observations gleaned from this static figure obscures the extent to which any individual participates in these changes. Unlike [1], we are primarily interested in the co-variation of vowels and system-wide effects rather than

individuals that are more or less advanced in the sound changes. Our focus is less on who are the leaders and laggers in these changes and more on characterizing which vowel specific formants pattern together within individuals.

The PCA on the by-speaker intercepts returned with two PCs accounting for more than 10% of the variance (PC1 = 44% of the variance, PC2 = 13.1% of the variance). Given space limitations, we only describe the patterns within PC1. In contrast to [1] where a few vowels loom large in their PC1, we see in Figure 2 that many vowels contribute to the first PC. In this figure, the vowels and formants are plotted in their order of contribution to the PC. Their sign (+/-) of the loadings is arbitrary, but vowel formant combinations that pattern with the same sign pattern in the same direction. For example, DRESS F1 and KIT F1 (+ sign) and EGG F2 (- sign) all contribute strongly to PC1 and so are co-varying, but DRESS and KIT F1 are patterning in the same direction – we can interpret from Figure 3 that they are increasing, while EGG F2 is patterning in the opposite direction, i.e., decreasing. The vowel formants and their PCA values are grayed-out once 50% of the variance within PC1 has been accounted for. This cut-off at contributing to 50% of the variance within PC1 seems more arbitrary in our dataset than [1], as the first two grayed-out vowel/formants – F1 of STRUT and F1 of FACE – are nearly equivalent in their PC value to the last bolded one – F2 of EGG. F2 of KIT, FACE, DRESS, EGG, and BAG all pattern together (with an arbitrary - sign) and F1 of KIT, DRESS, NURSE, GOAT, and LOT all pattern together (with an arbitrary + sign).

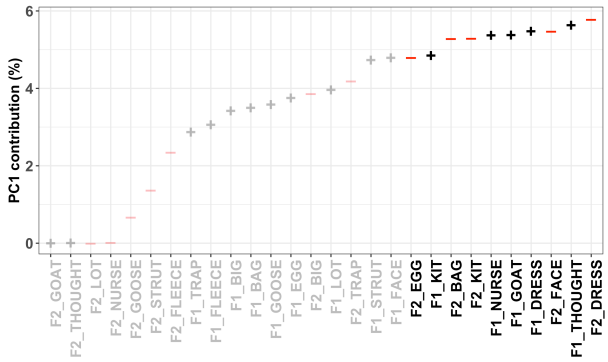


Figure 2: The contributions of each vowel/formant to PC1. The sign +/-, redundantly colour-coded, indicates the (arbitrary) direction of the loadings. The bold-faced formants contribute to the first 50% of the variance for PC1.

Figure 3 presents the relationship between the speaker intercepts from the GAMM models and the speaker PC values by reader age for all of the vowels and formants. Positive speaker PC signs from the PCA are in green and negative signs are in black. The vowel-formant combinations that contributed to more than 50% of the variance in PC1 are boxed in yellow. The percent of variance accounted for by each vowel formant is reported; the yellow box's brightness indicates the contribution of that vowel formant to the PC. The red line in each of the panels is the sound change smooth from the GAMM, and one observes that it clearly separates the green and black speaker PCA points in most panels. The individuals with data points above the line are more innovate, while those below are more conservative.

Due to the large number of vowel formants contributing to this first PC, an eloquent synopsis of the data is challenging. There is clearly structured variation in that the regression line bifurcates green (+) and black (-) data points. Starting with BAG, we see a pattern where the F2 value becomes higher over time, indicating a fronting pattern; this is accompanied by a lowering of F1, indicating a raising in the vowel space. For TRAP, we see a lower, backer vowel in comparison to BAG, with little change over time, but substantial variation. This may suggest that TRAP had already reached its retracted position as part of the CVS for even the oldest speakers in this sample. Individuals who have more fronted and raised TRAP are more likely to have a BAG that is fronted and raised. Such a pattern holds across the front vowels in voiced velar environments: individuals who have more fronted and raised TRAP, DRESS, and KIT have more fronted and raised BAG, EGG, and BIG. These same speakers also have fronter and higher FACE vowels, and to some extent FLEECE too. These same speakers have higher THOUGHT, NURSE, and GOAT vowels as well. While this analysis is not set up to directly quantify change over time, we can observe some patterns from the regression lines. For example, there is an indication that FACE is fronting and GOOSE has fronted over time, with a clear elbow in the GAMM prediction. LOT and THOUGHT are backing, which is related to PC2 and not discussed in detail here. NURSE and GOOSE are lowering. EGG and BIG are lowering, while DRESS lowers and retracts. BAG is fronting and raising.

4. DISCUSSION AND CONCLUSION

An individual speaker's vowels are part of a vowel system, which is, in turn, structured by the patterns of a speech community and the phonetic and

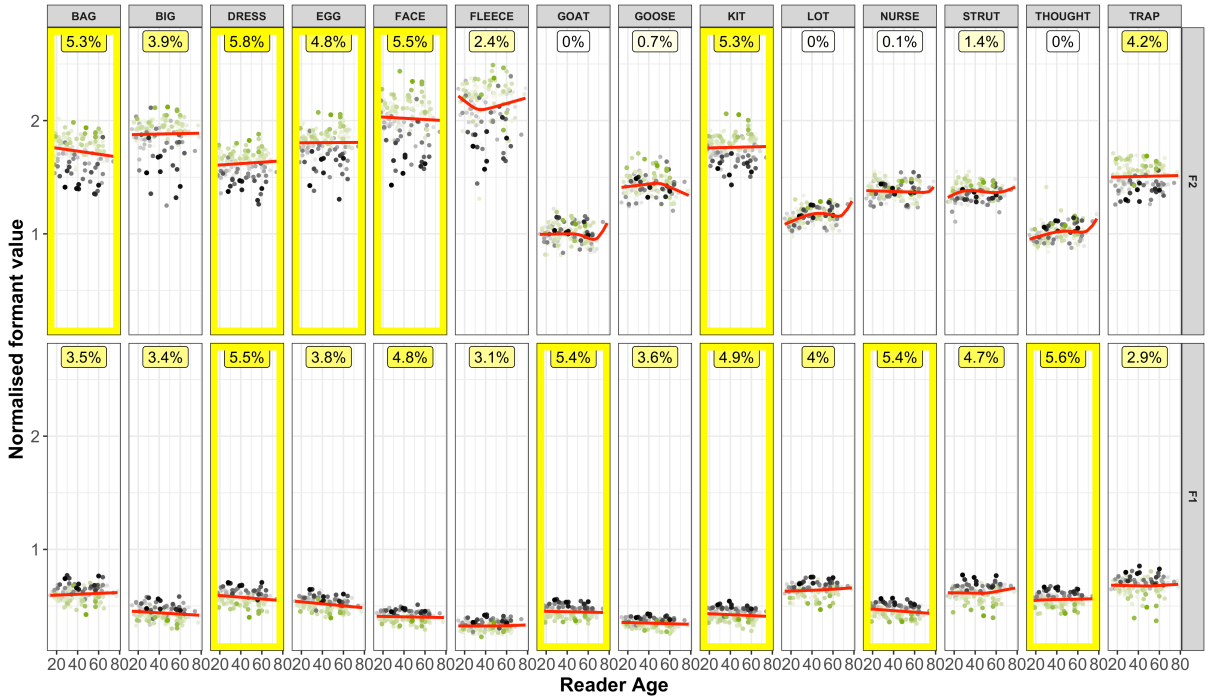


Figure 3: The trajectories of vowel formants by reader age (younger = left; older = right). Speaker intercepts are presented as points. The point color denotes the direction of the speaker's PC score with green as positive and black as negative. The contribution of each vowel formant is provided with the yellowness of the box indication strength of contribution.

phonological structure of the vowels themselves. The application of this GAMM + PCA analysis method allows us to see how vowel variation across a large number of vowels is structured within individuals in a speech community and to some extent across apparent time. We find that 10 vowel-specific formants simultaneously pattern together in individuals' productions, contributing over 50% of the variation in PC1, which itself accounted for 44% of the variance in the data. We see evidence of expected vowel patterns for the BCE vowel system, such as BIG, EGG, and BAG being separate from KIT, DRESS, and TRAP in F1/F2 space and GOOSE fronting. However, we also find unexpected patterns, such as THOUGHT and LOT's midpoints not occupying the same F1/F2 space, and NURSE lowering.

As the use of this vowel system analysis is still novel, we draw attention to the fact that our PC1 accounted for 44% of the variance in our data, while PC1 in [1] accounted for just 17.2% of the variance in their New Zealand English data. This suggests that the vowels we examine in BCE are behaving more uniformly across speakers. There are several reasons to suspect this is the case. Our data are operating on a more contracted time depth, with ages ranging from 14 to 80, compared to the 118 year time depth in [1]. Their work uses spontaneous speech from multiple regions in NZ, compared to our read speech from BC only, which not only elicits a more constrained style, but is just a single style, as opposed to the myriad styles that may be present in the ONZE corpus. It is also possible that the vowels of BCE are less variable than those of NZE.

Scholars often entertain causal explanations related to phonetic or phonological structural relationships when adjacent vowels shift in the F1/F2 space. While we lack evidence for any causal mechanism, we see co-patterning of both adjacent and non-adjacent vowels. The term simultaneous changes allows us to qualify the observation that particular phonetic variation across vowels co-occurs without proposing the causation. One potential explanation for any coordinated variation is its indexing of social meaning. The methods deployed here do not allow for the assignment or interpretation of social meaning, but having identified the coordinated loci of phonetic variation, we can now begin hypothesis-driven work that examines structural relationships between co-patterning vowels and how the array of observed phonetic variation cues listeners to interpret particular social meaning(s).

5. ACKNOWLEDGMENTS

Thank you to Kyra Hayter and Kaining Xu for their work on this project. Thank you to Márton Sóskuthy for guidance on the analysis techniques.

6. REFERENCES

[1] J. Brand, J. Hay, L. Clark, K. Watson, and M. Sóskuthy, "Systematic co-variation of monophthongs across speakers of new zealand english," Journal of Phonetics, vol. 88, p. 101096, 2021.

[2] A. Martinet, "Function, structure, and sound change," WORD, vol. 8, no. 1, pp. 1–32, 1952. [Online]. Available: https://doi.org/10.1080/00437956.1952.11659416

[3] A. Dinkin, "Toward a unified theory of chain shifting," in The Oxford Handbook of the History of English. Oxford University Press: Oxford, 2012.

[4] W. Labov, Principles of linguistic change. Vol. 1: Internal factors. Oxford: Blackwell, 1994.

[5] L. Hinton, B. Moonwomon, S. Bremner, H. Luthin, M. Van Clay, J. Lerner, and H. Corcoran, "It's not just the valley girls: A study of california english." 1987.

[6] A. D'Onofrio, T. Pratt, and J. Van Hofwegen, "Compression in the california vowel shift: Tracking generational sound change in california's central valley," Language variation and change, vol. 31, no. 2, pp. 193–217, 2019.

[7] M. J. Gordon, "Methodological and theoretical issues in the study of chain shifting," Language and Linguistics Compass, vol. 5, no. 11, pp. 784–794, 2011. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2011.00310.x

[8] S. Clarke, F. Elms, and A. Youssef, "The third dialect of english: Some canadian evidence," Language Variation and Change, vol. 7, no. 2, p. 209–228, 1995.

[9] E. Sadlier-Brown and M. Tamminga, "The canadian shift: Coast to coast," in Proceedings of the 2008 annual conference of the Canadian Linguistic Association, vol. 4, 2008, pp. 1–14.

[10] C. Boberg, "A closer look at the short front vowel shift in canada," Journal of English Linguistics, vol. 47, no. 2, pp. 91–119, 2019.

[11] ——, "Regional phonetic differentiation in standard Canadian English," Journal of English Linguistics, vol. 36, no. 2, pp. 129–154, 2008.

[12] Anonymous, "Regional variation in Canadian English in British Columbia," Journal of Linguistic Geography, Under Review.

[13] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in Interspeech, vol. 2017, 2017, pp. 498–502.

[14] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, emuR: Main Package of the EMU Speech Database Management System, 2021, r package version 2.3.0.

[15] L. Bombien, R. Winkelmann, and M. Scheffers, wrassp: an R wrapper to the ASSP Library, 2022, r package version 1.0.2.

[16] K. Johnson, "The delta f method of vocal tract length normalization for vowels," Laboratory Phonology, vol. 11, no. 1, 2020.

SOCIOLINGUISTIC FACTORS PREDICTING SOUND CHANGE IN SWISS GERMAN

Adrian Leemann[1], Carina Steiner[1,2], Beat Siebenhaar[3], Péter Jeszenszky[4]

[1]University of Bern, [2]University of Zurich, [3]University of Leipzig, [4]Independent researcher
adrian.leemann@unibe.ch, carina.steiner@unibe.ch, siebenhaar@uni-leipzig.de, pjeszenszky@gmail.com

ABSTRACT

Swiss German dialects have changed substantially over the past few centuries. Previous literature provides ample evidence of sound change. However, most previous studies are limited either in the number of localities or variables, or in the sociolinguistic metadata obtained from participants. In the present contribution we provide a large-scale analysis of sound change, examining ten phonetic variables from 1000 speakers who come from 125 representative localities in German-speaking Switzerland. Factors such as age, mobility, and dialect identity appear to be the driving forces behind sound change, which is further patterned regionally and constrained by between-item variation. We discuss reasons for the sound change observed.

Keywords: sound change, Swiss German, sociophonetics, sociolinguistics

1. INTRODUCTION

Switzerland has played a leading role in dialect documentation. Most noteworthy is the world-renowned Linguistic Atlas of German-speaking Switzerland [1, henceforth SDS], documenting Swiss German (SwG) dialects in the middle of the 20th century. In recent decades, a series of studies have observed that this documentation is no longer up to date and have given potential sociolinguistic reasons for sound change in progress.

In the domain of consonants, for example, Schifferle [2] reports that the areal distribution of aspirated plosives has expanded since the creation of the SDS, which is explained by speakers converging towards Standard German in word-initial plosives. Several studies have further examined the diffusion of /l/-vocalization. This feature, which was originally most likely Bernese, has been spreading particularly towards Southwest and Central Switzerland [3–5]. Causes for this wave-like diffusion are manifold: /l/-vocalization may have become a local identity marker since it leads to a stronger distinction from Standard German [6]; at the same time, speakers have highly positive associations with Bern German, which is likely to support diffusion of this feature [7].

In the domain of vowels, Leemann and Kolly [8] show that regional distributions of Old Upper German <iu> have changed dramatically compared to the SDS [1]: the Zürich variant [ty:f] appears to have spread hierarchically towards Central Switzerland and the Southwest. Leemann and Kolly [8] argue that commuters between Central Switzerland and Zurich, as well as between the Southwest and Zurich, are likely to have brought about diffusion: the Zurich variant was first adopted in larger hubs like Zug and Schwyz and, from there, diffused to smaller neighboring towns and valleys. In addition to these studies which focused more on a multi-locality, geolinguistic dimension, studies such as Eckhardt; Fleischer and Schmid; Hofer; Russ; and Siebenhaar [9–13] have also focused on change in multiple variables, but only in one specific locality.

The studies presented so far either focus on large-scale, multi-locality analyses of change in one or a few variables or else they revolve around single-locality studies that examine multiple variables concurrently but lack geolinguistic scale. What is currently missing is a combination of a large-scale multi-locality approach with an examination of multiple variables. Further – and perhaps most importantly – none of the above-mentioned studies collected a large amount of sociolinguistic metadata from their participants, which prevented a thorough analysis of sociolinguistic factors (potentially) predicting sound change. The present study attempts to fill these gaps.

In a large-scale survey involving 1000 participants from across 125 representative localities, the current study investigates sound change by examining ten exemplary phonetic variables (five consonants and five vowels). The 1000 participants provided an unprecedented amount of metadata (such as personality traits, dialect identity, political leaning, specific education backgrounds, mobility behavior etc.) that enable an explanation of sociolinguistic factors affecting sound change. Given the previous literature, we expected sound change to have occurred, particularly in central Switzerland and the Southeast, and we expected dialect identity (i.e., orientation towards the local vernacular) to play a change-impeding role (cf. Werlen [7]).

2. METHODS

2.1. Materials

Ten items were investigated in the current study: five exemplary vowels and five exemplary consonants. Table 1 shows the variables: type, variable, and the item elicited containing the variable in question, written in Standard German (NB: participants recorded their dialectal variant, cf. Section 2.3 Procedures).

Type	Variable	Standard German
Vowel	MHG <u>	*Rücken* 'back' (ana.)
Vowel	OUG <iu>	*tief* 'deep'
Vowel	MHG <æ>	*Käse* 'cheese'
Vowel	Rounding of MHG <e>	*Apfel* 'apple'
Vowel	MHG <e>	*Bett* 'bed'
Consonant	MHG -nt	*Hund* 'dog'
Consonant	/n/ before fricative	*Zins* 'interest (fin.)'
Consonant	Gemination of MHG <nn>	*Tanne* 'fir'
Consonant	Germ. <-k>	*Kind* 'child'
Consonant	MHG <-hs>	*Sechs* 'six'

Table 1: Ten variables of the current study (MHG = Middle High German, OUG = Old Upper German, Germ. = Germanic).

2.2. Speakers

A total of 1000 speakers from the SDATS database [14] – a contemporary database of Swiss German compiled during the COVID-19 pandemic in 2020–2021 – were investigated. Participants came from 125 localities across German-speaking Switzerland. Eight participants per locality took part in the survey: four females and four males. Two age cohorts were included: 500 speakers aged 60 and over and 500 speakers aged 20–35 years old. Eight larger dialect regions were established: Bern, Central Switzerland, Fribourg Valais and Ticino, Grisons, Northeastern Switzerland, Northwestern Switzerland, Zurich, and Aargau (cf. Hotzenköcherle [15]).

2.3. Procedures

2.3.1. Data collection

Due to the COVID-19 pandemic, data was mostly collected remotely via smartphones and Zoom (76.2% was collected remotely, cf. [14]). Participants were connected with the investigator via Zoom while recording answers to prompts viewed on a custom-built app on their smartphones [16]. The investigator, in real time, checked the quality of the uploaded audio files. Most items were elicited via picture prompts (e.g., *Rücken* 'back', *Käse* 'cheese', *Apfel* 'apple' etc.), a few were elicited via text prompts (e.g., 'What is the opposite of high? – *tief*, 'deep'). Following the ~2h supervised recording session, participants filled in a metadata questionnaire online, without supervision, providing information on their mobility behavior [17] and dialect identity (an index was calculated on the basis of five questions that capture the participant's degree of local affinity), dialect background, social networks, education, Big Five personality questionnaire [18] etc. Completing this questionnaire took another ~45min. Participants were compensated with 100 CHF in total.

2.3.2. Data coding and modeling

Data was coded auditorily by five human annotators and compared to the historic SDS [1]. In case of uncertainty, codings were double-checked between the annotators. Change was binary coded for each variable (0 vs. 1), and – in a second step – proportional mean change was calculated (e.g., change in five out of ten items results in a mean change of 0.5). Linear regressions were used to model the effects of sociodemographic (e.g., age: older (60+) vs. younger (18-35) cohort), regional (e.g., eight larger regions mentioned earlier), attitudinal (e.g., local orientation), and personality-related factors (e.g., Big Five personality traits) on sound change. Mixed models [19] with all of the above-mentioned predictors as fixed effects and random intercepts for speakers and item did not converge. Between-item variation was instead examined by comparing the raw relative change across all items in both age cohorts (cf. Section 3.4 below).

3. RESULTS

The full model output is displayed in Table 2.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.143302	0.045820	3.127	0.001816	**
Age_Cohortyounger	0.082343	0.023444	3.512	0.000465	***
GenderM	-0.006309	0.007026	-0.898	0.369426	
Edu_4levelsSecondary voc. education	0.026500	0.013915	1.904	0.057149	.
Edu_4levelsTertiary voc. education	0.022728	0.019281	1.179	0.238783	
Edu_4levelsUniversity degree	0.004225	0.016071	0.263	0.792659	
LMI_D	0.042870	0.007733	5.544	3.81e-08	***
RegionBern	0.007603	0.019687	0.386	0.699425	
RegionCentral Switzerland	0.010811	0.020438	0.529	0.596954	
RegionFribourg, Valais, Ticino	0.031973	0.021058	1.518	0.129260	
RegionGrisons	0.047078	0.024515	1.920	0.055108	.
RegionNortheastern Switzerland	-0.023503	0.019647	-1.196	0.231903	
RegionNorthwestern Switzerland	-0.007559	0.023829	-0.317	0.751154	
RegionZurich	-0.060790	0.026329	-2.309	0.021159	*
Aff_identity	-0.008365	0.003872	-2.160	0.030986	*
extra	0.010322	0.008151	1.266	0.205735	
gewiss	0.001656	0.008731	0.190	0.849563	
offen	-0.014745	0.008219	-1.794	0.073125	.
Age_Cohortyounger:RegionBern	-0.050643	0.027799	-1.822	0.068799	.
Age_Cohortyounger:RegionCentral Switzerland	-0.036989	0.028956	-1.277	0.201755	
Age_Cohortyounger:RegionFribourg, Valais, Ticino	-0.058250	0.029794	-1.955	0.050861	.
Age_Cohortyounger:RegionGrisons	0.029430	0.034932	0.842	0.399734	
Age_Cohortyounger:RegionNortheastern Switzerland	-0.032148	0.027747	-1.159	0.246911	
Age_Cohortyounger:RegionNorthwestern Switzerland	-0.060514	0.033738	-1.794	0.073182	.
Age_Cohortyounger:RegionZurich	-0.075509	0.037211	-2.029	0.042706	*
Edu_4levelsSecondary voc. education:LMI_D	-0.028664	0.009517	-3.012	0.002663	**
Edu_4levelsTertiary voc. education:LMI_D	-0.026351	0.013102	-2.011	0.044571	*
Edu_4levelsUniversity degree:LMI_D	-0.026203	0.010634	-2.464	0.013905	*

Table 2: Full model output (model prompt: lm(mean change ~ Age Cohort + Gender + Education + Mobility + Region + Identity + Extraversion + Conscientiousness + Openness + Age Cohort:Region + Education:Mobility).

The linear regression revealed significant effects of age, region, mobility, and identity, as well as interactions between age*region and

education*mobility. In what follows, we present results for each of the significant effects along with their interactions.

3.1 Age*region

Figure 1 is a bar chart of change by age (younger cohort: dark blue, older cohort: light blue) and region.

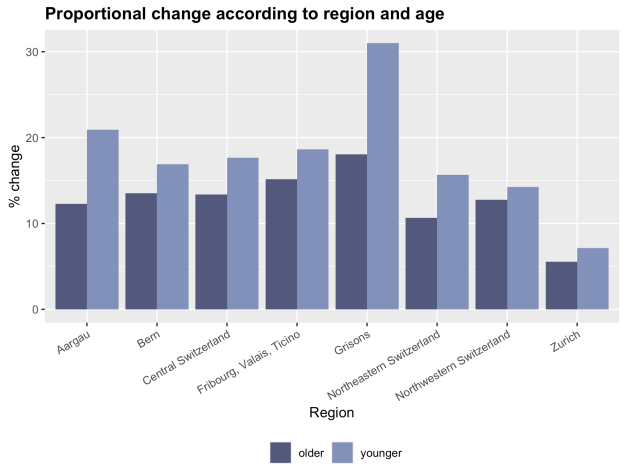


Figure 1: Bar chart of change by region crossed by age cohort.

The younger cohort shows more change on average (18%) than the older cohort (13%). The figure further shows Zurich exhibiting the least change compared to the other regions. The figure also displays the interaction between age and region: for Grisons, for example, the change is particularly substantial in the younger cohort (31% vs. 18%), while for Fribourg, Valais, and Ticino the difference between the two age cohorts is much smaller (18% vs. 15%). To get a better idea of the regional effects, Figure 2 shows the regional distribution for the two age cohorts (older: top, younger: bottom).

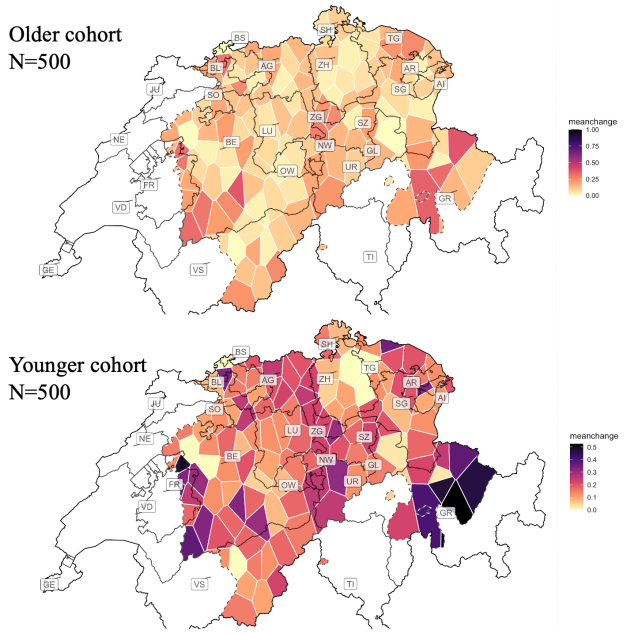


Figure 2: Regional variation of sound change across all ten items by age cohort: older (top), younger (bottom).

Both maps illustrate the least change for the canton of Zurich (ZH) and the most change for the canton of Grisons (GR). The change in Grisons is particularly pronounced in the younger cohort (hence the interaction age*region).

3.2 Mobility*education

The results further revealed that the more mobile the participant, the more likely they are to exhibit sound change (0.04(±0.008), t=5.5, p<0.001). This effect is constrained by the education level of the participant, however: the effect is much more pronounced if the participant's highest current degree is a (vocational) Baccalaureate; the effect is less pronounced for the other education degrees.

3.3 Dialect identity

There was also an effect of dialect identity. The higher the participants' score on the identity index (i.e., the more they exhibited local affinity), the less change they are likely to exhibit (-0.008(±0.004), t=-2.2, p=0.03). This means the more they identify with their local heritage, the more likely they are to exhibit little change.

3.4 Item

Finally, on a descriptive level, we explored change by item, crossed by age cohort; see the bar chart in Figure 3.

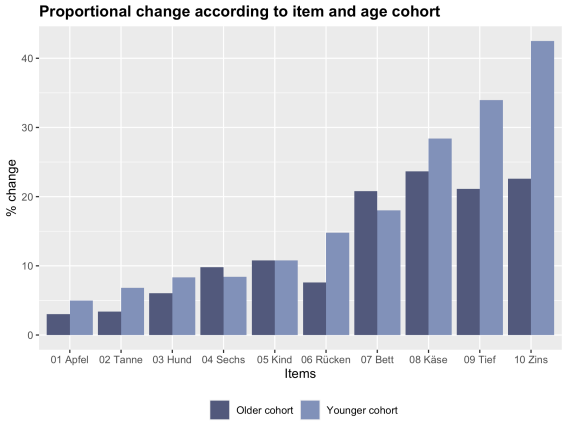


Figure 3: Bar chart of change by item crossed by age cohort.

Figure 3 reveals that the most change can be found in /n/ before a fricative as in *Zins* (42% young cohort, 22% old cohort), and the least change in rounding of MHG <e> as in *Apfel* ('apple') (5% vs. 3%). For most items, we find apparent-time change. This is not the case for the variables MHG <e> as in *Bett* ('bed') and MHG <-hs> as in *sechs* ('six'), however. In both of these variables, it is the older cohort that actually shows slightly more change compared to the SDS.

4. DISCUSSION

We will begin by discussing the main effects of age and region, before moving on to mobility, dialect identity, and item effects. An effect of age was expected. Younger speakers typically lead sound change. Going into the study, we predicted that most change would happen in the Southeast and in Central Switzerland (cf. Leemann and Kolly; Eckhardt; Fleischer and Schmid; Hofer; Russ; Siebenhaar [8–13]). Our prediction, based on the findings of previous studies, was accurate regarding change in the Southeast. Perhaps this is diffusion of the Chur (capital of that region) variant towards these localities – some of which, historically, happen to be localities where Walser (a historically Southwestern variety) dialects were spoken. This change has progressed particularly swiftly in the younger cohort. Change in this cohort may have been sped up by the fact that the Southeast has, to a large part, a bilingual population – Romansh and German. Romansh has been increasingly pushed aside by German influence, particularly for younger speakers, thus creating further linguistic instability in this cohort which may cause an acceleration of sound change.

Regarding mobility, we found that the more mobile the person (as quantified by their exposure to other dialects), the more sound change they exhibit. Mobility and the intensity of interactions as an explanatory force for language change have been shown in Jeszensky et al.; Hernández-Campoy; and Beaman [17, 20, 21]. The fact that this effect is particularly pronounced for participants with a (vocational) Baccalaureate (hence the interaction with education degree*mobility) may be because they are of the younger cohort.

Moving on to dialect identity, our results showed that strong regional association can impede sound change. This finding makes sense intuitively: the more locally attuned speakers are, the prouder they are of their dialect, and the more likely they are to retain phonetic features of that region. This has been shown previously by Beaman and Tomaschek [22], who examined the retention of traditional Swabian features in relation to how closely oriented towards Swabia the participants were (cf. also Steiner et al. [23] for a similar phenomenon in the morphosyntactic domain).

Finally, regarding the differences between items in relation to change, we can point out a couple of trends. The items with the most change are /n/ before a fricative (as in *Zins*, 'interest fin.') and OUG <iu> (as in *tief*, 'deep'). Regarding the former, there appears to be a convergence towards Standard German, with lots of speakers – particularly younger speakers – moving towards *Zins*. Looking at our data for OUG <iu>, the Zurich variant [ty:f] is spreading in virtually every direction, which has been previously reported by Leemann and Kolly [8]. We speculate that there may be lexical factors at play: the retention of the omission of /n/ before a fricative, for example, was already present at the time of the SDS [1]: words like *Gans* ('goose') already showed convergence towards the standard in the 1950s, whereas *Zins* barely did so. Nowadays, however, variants like [ts͡eis] for *Zins* are gradually retreating, but at a much slower rate than similar n-less variants of *Gans* did. Regarding items with very little change, such as <nn> and rounding of MHG <e>, we can only speculate. The little change in MHG <nn> is somewhat surprising, given that the geminate realization has been largely abandoned in both Standard German and German dialects, even in the South (cf. BSA and SSA [24, 25]). The little change in the rounding of MHG <e> may again be due to lexical constraints, showing very little change in words like *Apfel* ('apple') but – so we speculate – demonstrating substantial change in words like *Löffel* ('spoon'), where most regions as well as Standard German use this rounded MHG <e> variant.

In the future, further phonetic variables (all Swiss German vowels, for example) will be analyzed and compared to historical data. This will be further placed into the context of change at other linguistic levels, such as lexis, syntax, and morphology – which will reveal stability or change in the phonetic domain in a much broader linguistic context.

5. ACKNOWLEDGMENTS

This research was funded by the Swiss National Science Foundation (Grant no. PCEFP1_181090).

6. REFERENCES

[1] *SDS: Sprachatlas der deutschen Schweiz.* 1962–2003. Vols. 1–8. Francke.

[2] Schifferle, H. P. 2010. Zunehmende Behauchung. Aspirierte Plosive im modernen Schweizerdeutsch. In Christen, H., Germann, S., Haas, W., Montefiori, N., Ruef, H. (eds), *Alemannische Dialektologie: Wege in die Zukunft. Beiträge zur 16. Tagung für alemannische Dialektologie in Freiburg/Fribourg vom 07.– 10.09.2008* (ZDL-Beiheft 141). 43–55.

[3] Haas, W. 1973. Zur l-Vokalisierung im westlichen Schweizerdeutschen. In Bausinger, H. (ed.), *Dialekt als Sprachbarriere: Ergebnisbericht einer Tagung zur alemannischen Dialektforschung*. Tübinger Vereinigung für Volkskunde, 63–70.

[4] Christen, H. 1988. *Sprachliche Variation in der deutschsprachigen Schweiz. Dargestellt am Beispiel der /l/-Vokalisierung in der Gemeinde Knutwil und in der Stadt Luzern*. Steiner.

[5] Leemann, A., Kolly, M.-J., Werlen, I., Britain, D., Studer-Joho, D. 2014. The diffusion of /l/-vocalization in Swiss German. *Lang. Var. Change* 26(2), 191–218.

[6] Christen, H. 2001. Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz. *Zeitschrift für Dialektologie und Linguistik* 1, 16–26.

[7] Werlen, I. 1985. Zur Einschätzung von schweizerdeutschen Dialekten. In Werlen, I. (ed.), Probleme der schweizerischen Dialektologie. 2. *Kolloquium der schweizerischen. Geisteswissenschaftlichen Gesellschaft. Problèmes de la dialectologie suisse. 2e Colloque de la Société suisse des sciences humaines*. Editions Universitaires Fribourg, 195–266.

[8] Leemann, A., Kolly, M.-J. 2016. Big Data for analyses of small-scale regional variation: A case study on sound change in Swiss German. In: *Proceedings of Phonetik & Phonologie 2016*.

[9] Eckhardt, O. 2016. *Alemannisch im Churer Rheintal: Von der lokalen Variante zum Regionaldialekt*. Steiner.

[10] Fleischer, J., Schmid, S. 2006. Zurich German. *J. Int. Phon. Assoc.* 36(2), 243–253.

[11] Hofer, L. 1997. *Sprachwandel im städtischen Dialektrepertoire: Eine variationslinguistische Untersuchung am Beispiel des Baseldeutschen*. Francke.

[12] Russ, C. V. J. 2002. *Die Mundart von Bosco Gurin: Eine synchronische und diachronische Untersuchung*. Steiner.

[13] Siebenhaar, B. 2000. *Sprachvariation, Sprachwandel und Einstellung: Der Dialekt der Stadt Aarau in der Labilitätszone zwischen Zürcher und Berner Mundartraum*. Steiner.

[14] Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguist. Vanguard* 6(s3).

[15] Hotzenköcherle, R. 1984. *Die Sprachlandschaften der deutschen Schweiz: Herausgegeben von Niklaus Bigler und Robert Schläpfer*. Sauerländer (= Reihe Sprachlandschaften, 1).

[16] Leemann, A., Jeszenszky, P. Steiner. C. 2020. SDATS App. https://apps.apple.com/us/app/sdats/id1516597765, https://play.google.com/store/apps/details?id=ch.sdats&hl=en&gl=US

[17] Jeszenszky, P., Steiner, C., Leemann, A. (revisions) Effects of mobility on dialect change: Introducing the Linguistic Mobility Index. *PLOS ONE*.

[18] Tupes, E. C., Christal, R. E. 1961. *Recurrent Personality Factors Based on Trait Ratings* (USAF ASD Tech. Rep. No. 61-97). Lackland Air Force Base, U.S. Air Force.

[19] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1). https://doi.org/10.18637/jss.v067.i01

[20] Hernández-Campoy, J. M. 2003. Exposure to contact and the geographical adoption of standard features: Two complementary approaches. *Lang. Soc.* 32(2), 227–255.

[21] Beaman, K. V. 2019. Identity and mobility in linguistic change across the lifespan: The case of Swabian German. In Ziegler, A., Edler, S., Oberdorfer, G. (eds), *Urban Matters: Current Approaches of International Sociolinguistic Research* (Studies in Language Variation), 1–42.

[22] Beaman, K. V., Tomaschek, F. 2021. Loss of historical phonetic contrast across the lifespan: Articulatory, lexical, and social effects on sound change in Swabian 1. In: *Language Variation and Language Change Across the Lifespan*. Routledge, 209–234.

[23] Steiner, C., Jeszenszky, P., Leemann, A. 2022. Variation and change in Swiss German agreement morphology: Spatial, social, and attitudinal effects. *Journal of Linguistic Geography*, 1–17. doi: 10.1017/jlg.2022.12.

[24] *BSA: Bayerischer Sprachatlas*. Hinderling, R., König, W., Eichinger, L. M., Eroms, H.-W., Munske, H. H., Wolf, N. R. (eds).

[25] *SSA: Südwestdeutscher Sprachatlas* (1989–2012). Steger, Hugo (ed.). Elwert.

PHONOTACTICALLY DRIVEN CUE WEIGHTING IN A SOUND CHANGE IN PROGRESS: ACOUSTIC EVIDENCE FROM WEST CENTRAL BAVARIAN

Katharina Thon, Felicitas Kleber

Institute for Phonetics and Speech Processing (IPS), LMU Munich
[k.thon|kleber]@phonetik.uni-muenchen.de

ABSTRACT

An analysis of acoustic recordings of fortis and lenis stops in pre- and postvocalic position in a variety of real words obtained from 23 speakers evenly distributed across younger and older West Central Bavarian (WB) as well as Standard German (SG) speakers showed that closure duration, the hitherto primary cue to the fortis/lenis contrast in WB becomes less important in younger WB speakers and in phonotactically previously illegal post-vocalic contexts while VOT – the primary cue to this contrast in SG – becomes more important in legal post-vocalic contexts. These results indicate that the apparent simultaneous change in acoustic cues occurs in different contexts and non-linearly at different paces. In initial position, older and younger WB speakers use VOT similarly to signal the fortis/lenis contrast but to a significantly lesser extent than SG speakers suggesting incomplete neutralization without indications of an ongoing change in this position in terms of apparent-time differences.

Keywords: German varieties, dialect levelling, acoustic cue change, incomplete neutralization

1. INTRODUCTION

The two main aims of the present paper are to investigate (i) a suspected change in cue weighting in the production of the word-medial fortis/lenis contrast in West Central Bavarian (WB) spoken in south-eastern Germany and (ii) the potential emergence of this contrast in initial position in WB. Both changes are presumed to occur as a result of dialect levelling towards Standard German (SG) [1].

The voicing or fortis/lenis contrast in SG occurs in both syllable-initial and postvocalic word-medial position and is signaled primarily by the length of positive voice onset time (VOT; lenis: short lag, fortis: long lag) values [2]. In medial position, proportional vowel duration in the vowel+stop closure sequence serves as a secondary cue to the contrast with longer vowels and simultaneously shorter stop closure durations signaling the three lenis stops /b, d, ɡ/ and shorter vowels preceding stops with a longer closure duration cueing the three fortis counterparts /p, t, k/ [2, 3]. In this position the acoustic cues are not only differently weighted (cf. [4]) but they are also in a trading relationship, by which one cue may be offset by the other [5, 6]. In WB, on the other hand, the contrast is considered neutralized towards the lenis variant in initial position [7] and primarily cued by closure duration [8] in medial position. VOT plays no role [6]. A further difference between the two varieties is that phonemically long and short vowels can be freely combined with lenis and fortis stops in SG while in WB lenis stops (L) are always preceded by long vowels (V:) and fortis stops (F) by short vowels (V) [9–10] (cf. [11] for a discussion of allophonic vowel quantity). Because of this restriction some SG combinations are phonotactically illegal in WB and adjusted accordingly (e.g., V:F in SG /ˈfaː.tɐ/, ‘father’ > V:L in WB /fˈɔː.dɐ/; VL in SG /pˈʊd̥ɪŋ/ ‘pudding’ > VF in WB /pˈʊt̥ːɪŋ/) while others are legal (e.g., VF in SG /ˈpapa/ ‘dad’) but implemented differently in terms of acoustic cues.

The present study builds upon recent apparent-time results from two generations of the same speech community (cf. [6]) that suggest the gradual emergence of a VOT-based fortis/lenis contrast in Central Bavarian varieties spoken in Austria [12] and Germany [6], the leveling out of the phonotactic restrictions in medial position in WB [6, 9, 13, 14], and the often reported change in cue weighting in connection with diachronic sound change [15–17]. Regarding WB, younger speakers were shown to produce fortis stops with closure duration values in between the higher values produced by older WB and the substantially lower values found for SG speakers. Preliminary, though less robust results further suggest that VOT is used to a greater extent in medial position by younger vs. older WB speakers [6, 13]. The present study investigates whether these observations hold true for other lexical items produced by different WB speakers and whether the change robustly affects the merged contrast in initial position. In other words, this study is concerned with a generalization of this gradual sound change in terms of positional and lexical effects. Another issue is whether the suspected change in cue weighting affects word-medial stops differently as a function of phonotactic restrictions.

2. METHOD

2.1. Participants and speech materials

The analysis includes data from a total of 23 speakers who were assigned to one of three speaker groups: seven older WB speakers (aged 52–65, mean 57.3, SD 4.0; 4 female), eight younger WB speakers (aged 19–30, mean 25.7, SD 4.1; 4 female), and eight SG speakers (aged 21–82, mean 42.1, SD 20.43; 4 female). SG speakers were from Munich and its urban surroundings and self-identified as non-dialect speakers who only speak the in this area prevalent southern SG variety. Dialect speakers were from the rural parts of Upper Bavaria where WB is common and acquired as the first variety. Their dialect competence was assessed by the examiners who were native speakers of WB. None of the participants reported any hearing or speaking impairments.

Five monosyllabic words of a CVC(C) structure with stops in initial position and 35 disyllabic trochees of the form C'VCVC(C) with stops in initial, (prevocalic) and word-medial (postvocalic) position were analyzed. Only words with alveolar stops were selected for analysis as this place of articulation is not affected by phonotactic restrictions regarding the cooccurrence of vowel length and subsequent stop voicing in SG [18]. The 40 stops in total were evenly distributed across position and the underlying voicing category (i.e., 10 lenis and fortis stops per position) as well as lexical frequency (high and low; which, however, is not analyzed as a predictor variable in the present study). Within the target words with medial stops, an equal number of long and short vowels preceded both lenis and fortis stops.

All target words are part of both the SG and the WB lexicon and were embedded in different semantically meaningful carrier declarative sentences. Target words preceded the phrase-final past participle, a position that usually triggers a nuclear pitch accent. While carrier sentences were presented in standard orthography to SG speakers (e.g. *Er will die Bar verkaufen.* 'He wants to sell the bar'), a non-standardized but commonly used form (e.g. in text messages) was used for WB speakers (e.g. *Ea mog de Bar verkaffa.*).

2.2. Recording procedure and data preprocessing

Speakers were recorded using the SpeechRecorder software [19] and a head-mounted microphone either in a sound-attenuated at the Munich Institute of Phonetics and Speech Processing or online via WikiSpeech [20]. The audio signal was digitized at a minimum sampling rate of 44.1 kHz (48 kHz in some cases). Participants read each sentence, presented in randomized order, silently on a screen and reproduced it loudly from memory after the screen turned grey before the next sentence was presented. In case of an obvious mispronunciation the participants were asked to repeat the entire sentence. Out of three repetitions per sentence only the second repetition each was included in the present analysis (23 speakers × 40 target words = 920 tokens).

All utterances were automatically segmented using WebMAUS [21] and stored as an EMU speech database [22]. The relevant segment boundaries were subsequently checked and adjusted manually. These were the on- and offset of the utterance and the target word, respectively, and the segment boundaries of each phoneme within the target word. The target stops were further subdivided into closure and aspiration (corresponding to VOT) phase by manually adding an additional boundary right after the stop's burst.

2.3. Data analysis

To factor out potential between group differences in speech rate, we normalized each stop's closure and aspiration duration separately for word duration (with target stop duration being excluded). We refer to them as pClosure and pVOT, respectively. These proportional durations were the dependent variables in a total of four linear mixed-effects models (LMM) fitted separately to the data with stops in initial and medial position. In each of the four analyses Group (three ordered levels: SG > younger WB > older WB) and Stop (two levels: fortis and lenis) served as fixed factors and speaker and word as random factors. Because of the phonotactic restriction in WB regarding stops in medial position the underlying length of the preceding Vowel (two levels: short vs. long) was included as a third fixed factor in the LMMs fitted to postvocalic stops. Pairwise comparisons were computed in case of significant interactions.

All statistical analyses were run in R (v. 4.2.2, [23]) using tidyverse (v. 1.3.1, [24]), lmerTest (v. 3.1-3, [25]), and emmeans (v. 1.8.2, [26]).

3. RESULTS

3.1 Postvocalic stops in word-medial position

The LMM with pClosure as the dependent variable showed significant main effects for Group ($F[2, 25] = 4.5, p < .05$), Vowel ($F[1, 21] = 30.3, p < .001$), and Stop ($F[1, 24] = 24.6, p < .001$) as well as a significant interaction effect between Group and Vowel ($F[2, 21] = 5.2, p < .05$). Commensurate with Fig. 1, however, the speaker group dependent differences in pClosure emerged in particular in fortis stops and in a variety of ways that depended among others on Vowel, i.e., on its underlying quantity.

		Contrast	Estimate	*SE*	*df*	t.ratio	*p value*
pClosure	V:F	Standard – younger Bavarian	-0.046	0.043	28.4	-1.051	0.896
		Standard – older Bavarian	-0.117	0.047	28.5	-2.480	0.164
		Younger Bavarian – older Bavarian	-0.072	0.039	22.1	-1.839	0.463
	VF	Standard – younger Bavarian	-0.187	0.510	28.1	-3.705	0.011*
		Standard – older Bavarian	-0.170	0.055	29.5	-3.113	0.043*
		Younger Bavarian – older Bavarian	0.017	0.048	23.1	0.364	0.999
pVOT	V:F	Standard – younger Bavarian	0.051	0.015	21.9	3.342	0.031*
		Standard – older Bavarian	0.052	0.016	22.8	3.243	0.037*
		Younger Bavarian – older Bavarian	0.001	0.015	21.6	0.016	1.000
	VF	Standard – younger Bavarian	0.042	0.021	22	1.900	0.428
		Standard – older Bavarian	0.078	0.023	21.8	3.468	0.023*
		Younger Bavarian – older Bavarian	0.037	0.022	21.6	1.642	0.582

Table 1: Pairwise comparisons of estimated marginal means for pClosure and pVOT separately for long vowel plus fortis stop (V:F) and short vowel plus lenis stop (VF) sequences.

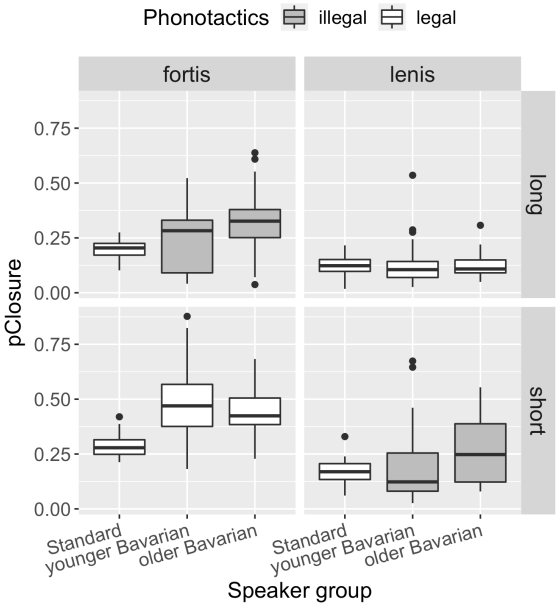


Figure 1: Proportional closure duration separately for the three speaker groups and the underlying stop and vowel category, respectively.

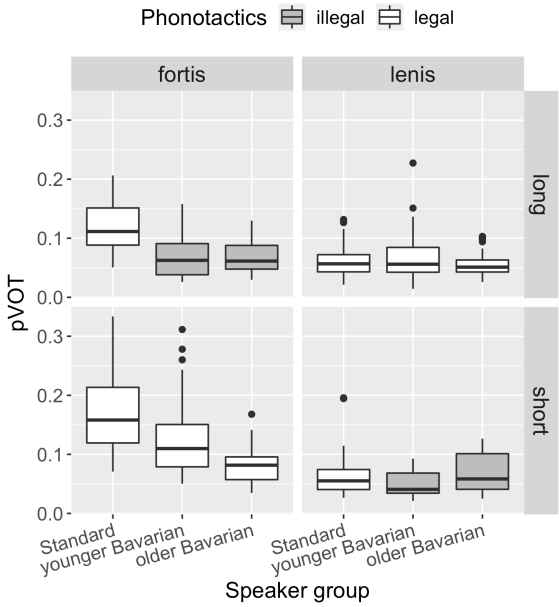


Figure 2: Proportional VOT separately for the three speaker groups and the underlying stop and vowel category, respectively.

SG speakers used closure duration to a lesser extent to signal the fortis/lenis contrast in particular compared to older WB speakers (cf. Fig. 1 and the significant contrast in the post-hoc pairwise comparison in Tab. 1). Interestingly, younger WB speakers took up an intermediate position in the production of V:F-sequences but not in that of VF sequences (cf. Tab. 1). This suggests that the importance of the closure duration cue diminishes in younger WB speakers but only in the sequence that is considered phonotactically illegal in WB. No such group or vowel quantity differences emerged for words with lenis stops.

The LMM with pVOT as the dependent variable showed again significant main effects for Group ($F[2, 21] = 3.9$, $p < .05$), Vowel ($F[1, 17] = 8.1$, $p < .05$), and Stop ($F[1, 22] = 22.2$, $p < .001$). This model further revealed significant interaction effects for Group and Stop ($F[2, 23] = 10.1$, $p < .001$), Vowel and Stop ($F[1, 15] = 5.8$, $p < .05$), and Group, Vowel and Stop ($F[2, 323] = 6.7$, $p < .01$). Commensurate with Fig. 2, SG speakers realized the fortis/lenis contrast on the basis of VOT and much more so than WB speakers (as seen by the significantly larger pVOT values for fortis stops in Fig. 2). In this analysis, too, younger WB speakers took up an intermediate position in the production of some words, most interestingly, however, not in that of V:F-sequences but in that of VF sequences (cf. Tab. 1). This suggests that VOT becomes more important for younger WB speakers to signal the fortis/lenis contrast but at first only in those sequences that are considered legal in Bavarian. Post-hoc pairwise comparisons revealed that the speaker group differences described above occur systematically only in words with fortis stops.

3.2 Prevocalic stops in word-initial position

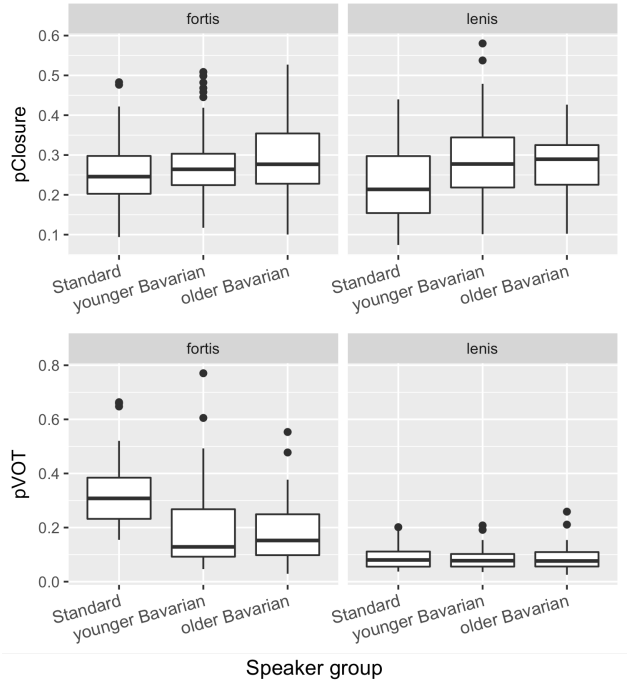


Figure 3: Proportional closure duration (top) and proportional VOT (bottom) separately for the three speaker groups and the underlying stop category.

Commensurate with the top row of Fig. 3, pClosure was significantly affected by Group (F[2, 21] = 3.9, $p < .05$) but not by Stop: WB speakers produced longer closure durations than SG speakers and in particular in lenis stops as shown by a significant interaction effect between the two fixed factors (F[2, 374] = 4.3, $p < .05$).

The LMM with pVOT as the dependent variable revealed significant main effects for Group (F[2, 25] = 5.8, $p < .01$) and Stop (F[1, 30] = 39.3, $p < .001$) as well as a significant interaction between these two (F[2, 26] = 7.3, $p < .01$). Commensurate with Fig. 3 all three speaker groups use VOT to produce the fortis/lenis contrast in initial position but the contrast appears more pronounced in SG than in WB as the significant difference between SG and both WB groups suggests. No such difference was found between the two WB groups.

4. DISCUSSION AND CONCLUSION

The present results are in line with previous findings on WB which indicated that the implementation of the post-vocalic fortis/lenis contrast becomes more SG-like [6, 9, 13, 14]. They further support our initial assumption that this gradual sound change in progress involves a reweighting of cues as documented for other changes in progress [15–17]. Interestingly, the present change in cue weighting appears to be conditioned by the underlying phonotactic differences between WB and SG.

While distinct closure durations, thus far the primary cue to distinguish postvocalic fortis and lenis stops in WB, diminish in contexts with preceding long vowels in younger compared to older WB speakers, neither of the two age groups use VOT in this hitherto in WB phonotactically illegal sequence. However, VOT differences, the primary cue to this contrast in SG, increase in younger compared to older WB speakers' realizations of words with fortis stops following upon short vowels, i.e., in sequences that are phonotactically legal in both varieties. Closure duration again remains a steadily strong cue in this context and this group of speakers. Thus, cues change non-linearly in production at different paces.

A post-hoc test of the relation between closure duration and VOT in postvocalic fortis stops at the group level using Pearson's r [27] revealed positive relationships in all three speaker groups with pVOT increasing in tandem with pClosure (WB young: r =0.58, WB old: r = 0.39, SG: r = 0.43). Contrarily to the trading relation between the two cues described in [5], this suggests the presence of cue enhancement [28], though on both dimensions and in particular in younger WB speakers [15]. Future studies should further investigate the nature and the role of enhancement during this particular sound change in progress (also by including further places of articulation).

The present results are less clear with respect to the second aim of this study. On the one hand, WB speakers contrasted lenis and fortis stops in word-initial position by means of clear differences in positive VOT, but to a lesser extent compared to SG speakers. This points towards incomplete [29] instead of complete neutralization [7] of the fortis/lenis contrast in initial position in WB. However, and in contrast to [6, 12], the present results do not support the initial assumption that this dialectal near-merger is currently being reversed as younger and older WB-speakers did not differ in this apparent-time comparison. This may be due either to differences in the lexical sets investigated (e.g., in place of articulation and lexical frequency) or the fact that reversals of mergers (as in initial position) are generally rare [30], also in comparison to changes in cue weighting regarding existing contrasts (as in medial position).

5. ACKNOWLEDGEMENTS

This research was supported by DFG grant number KL 2697/1-2 awarded to the second author. The R analysis script and the corresponding data frames are permanently available at [31].

5. REFERENCES

[1] Auer, P. 1998. Dialect levelling and the standard varieties in Europe. *Folia Linguistica* 32, 1–9.

[2] Jessen, M. 1998. *Phonetics and Phonology of Tense and Lax Obstruents in German.* John Benjamins Publishing.

[3] Kohler, Klaus J. 1977. The production of plosives. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel* 8, 30–110.

[4] Schertz, J., Clare, E. J. 2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science* 11, e1521.

[5] Harrington, J., Kleber, F., Reubold, U. 2012. The production and perception of coarticulation in two types of sound change in progress. In: Fuchs, S., Weirich, M., Pape, D., Perrier, P. (eds), *Speech planning and dynamics*. Peter Lang, 39–62.

[6] Kleber, F. 2018. VOT or quantity: What matters more for the fortis/lenis contrast in German regional varieties? Results from apparent-time analyses. *Journal of Phonetics* 71, 468–486.

[7] Wiesinger, P. 1990. The central and southern Bavarian dialects in Bavaria and Austria. In: Russ, C. V. J. (eds), *The dialects of modern German: A linguistic survey*. Routledge, 438–519.

[8] Seiler, G. 2005. On the development of the Bavarian quantity system. *Interdisciplinary Journal of Germanic Linguistic and Semiotic Analysis* 10, 103–129.

[9] Bannert, R. 1976. Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage. In: Malmberg, B., Hadding, K. (eds), *Travaux de l'institut de linguistique de Lund*. CWK Gleerup.

[10] Maddieson, I. 1997. Phonetic universals. In: Laver J., Hardcastle, W. (eds), *Handbook of phonetic sciences*. Blackwell, 619–639.

[11] Kleber, F. 2020. Complementary length in vowel-consonant sequences: Acoustic and perceptual evidence for a sound change in progress in Bavarian German. *Journal of the International Phonetic Association* 50, 1–22.

[12] Luef, E. M. 2020. Development of voice onset time in an ongoing phonetic differentiation in Austrian German plosives: Reversing a near-merger. *Zeitschrift für Sprachwissenschaft* 39, 79–101.

[13] Jochim, M., Kleber, F. 2022. Fast-speech-induced hypoarticulation does not considerably affect the diachronic reversal of complementary length in Central Bavarian. *Language and Speech*, doi: 10.1177/00238309221127641.

[14] Wolfswinkler, K., Harrington, J. 2021. The influence of Standard German on the vowels and diphthongs of West Central Bavarian. *Journal of the International Phonetic Association.* doi:10.1017/S0025100321000232

[15] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., Yoon, T. J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics* 66, 120–144.

[16] Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., Wissing, D. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics* 66, 185–216.

[17] Kuang, J. & Cui, A. 2018. Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics* 71, 194–214.

[18] Kleber, F., John, T., Harrington, J. 2010. The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics*, 38, 185–196.

[19] Draxler, C., Jänsch, K. 2004. SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. *Proc 4th International Conference on Language Resources and Evaluation* Lisbon, 559–562.

[20] Draxler, C., Jänsch, K. 2008. WikiSpeech – A Content Management System for Speech Databases. *Proc. of Interspeech* Brisbane, 1646–1649.

[21] Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., Pörner, N. 2016. BAS Speech Science Web Services - an Update of Current Developments. *Proc. 10th LREC* Slovenia.

[22] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45, 392– 410.

[23] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[24] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Yutani, H. 2019. *Welcome to the tidyverse. Journal of Open Source Software* 4 (43), 1686.

[25] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 1–26.

[26] Lenth, R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. https://CRAN.R-project.org/package=emmeans.

[27] Levshina, N. 2015. *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.

[28] Kirby, J. 2013. The role of probabilistic enhancement in phonologization. In: Yu, A. C. L. (eds), Origins of sound change: *Approaches to phonologization*. Oxford University Press, 228–246.

[29] Port, R. F., O'Dell, M. L. 1985. Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13, 455–471.

[30] Duncan, D. 2022. Merger reversal in St. Louis: Implementation and implications. *Journal of English Linguistics* 50, 72-105.

[31] Thon, K., Kleber, F. 2023. Phonotactically driven Cue weighting in a sound change in progress: Acoustic evidence from West Central Bavarian. Open Data LMU. 10.5282/ubm/data.378

POSSIBLE BACK PRELATERAL MERGERS IN OKLAHOMA

Valerie Freeman[1] & Molly Landers[2]

[1]Oklahoma State University, [2]Michigan State University
valerie.freeman@okstate.edu, lander15@msu.edu

ABSTRACT

Prelateral mergers occur in American English with a loss of vowel distinction before /l/, often between tense-lax pairs (/il-ɪl/ *feel-fill*, /el-ɛl/ *sale-sell*, /ul-ʊl/ *pool-pull*). This study examined non-low back/central prelateral vowels in Oklahoma, where plain counterparts /u, o, ʊ, ʌ/ are fronted and earlier work found /ul/-lowering. Formants (F1, F2) from 113 adults reading words and passages were measured mid-vowel. All four prelaterals were backed, without /ul/-lowering but with /ʊl/ varying in height between /ol/ and /ul/, and /ʌl/ varying between plain /ʌ/ and /ol/ or higher. While many speakers overlapped /ʊl/ and/or /ʌl/ with /ol/, none showed definitive three-way /ʊl-ʌl-ol/ merger. Instead, a correlation between /ʊl/- and /ʌl/-raising suggests a chain shift that might allow only one /ol/-merger per speaker – or two lax-tense mergers (/ʊl-ul/ *pull-pool*, /ʌl-ol/ *hull-hole*). However, wide individual variation is not well explained by demographics, leaving the direction and social meaning of these changes unsettled.

Keywords: Prelateral merger, vowel shift, sound change, American English, variationist dialectology

1. INTRODUCTION

Conditioned vowel mergers are sound changes over time in which phonemes lose distinction in certain phonetic environments, in this case before /l/. Several prelateral mergers have been described as advancing through generations of various American English regions, often between tense-lax pairs, with front pairs laxing in the South (/il-ɪl/ *feel-fill*, /el-ɛl/ *sale-sell*) but back vowels often tensing in the Midland and West (/ul-ʊl/ *fool-full*, /ʊl-ol/ *bull-bowl*, /ʌl-ol/ *hull-hole*, /ʌl-ol-ʊl/ *pulp-pole-pull*, /ʊl-ol-ul/ *pull-pole-pool*) [1, 2, 3, 4, 5, 6]. Oklahoma is in a dialect transition zone with a mix of regional features, including Southern /u, o/-fronting, Midland /o, ʊ, ʌ/-fronting, and Western /a-ɔ/ *caught-cot* merger [1, 7]. Thus, we might expect any combination of prelaterals to shift toward each other, with age differences indicating change over time and gender, rurality, or register indicating social meaning.

Past studies in Oklahoma reported the laxing of three prelaterals /il, el, ul/ advancing over time [1, 7, 8, 9, 10], but they either had very few participants or did not include lax counterparts, and none detailed any mid-back prelaterals. The present study provides a deeper examination of four back prelaterals, seeking to describe their arrangements and identify possible mergers in progress. The youngest speakers in earlier work would now fit into the two older age ranges in the present study (Table 1), and speakers in the most recent work [7] overlap the three oldest age ranges here, providing a continuation in apparent time. Age, gender, rurality, and reading style will be explored as factors that often differentiate stages of change.

2. METHODS

2.1. Participants

Speakers were 113 adults aged 18–92 ($M = 33$) who were native speakers of English raised in Oklahoma with no speech disorders post-childhood. Males and females were sought from four age groups (college-age, early adult, middle-aged, senior; see Table 1), and three city sizes: City (urban/suburban Oklahoma City and Tulsa), Town (populations > 20,000), and Country (rural). Ethnicity was not considered in recruitment but roughly reflected the overall makeup of Oklahoma: most were White, with 17 identifying as Native American and 4 as Hispanic, Asian, or Middle Eastern (solely or in addition to White).

Table 1: Numbers of participants by age range, city size, and gender (Male, Female, Other).

	City		Town		Country		
Age	M	F	M	F	M	F	O
18–24	9	17	4	9	2	13	
25–39	3	5	3	3	8	7	1
40–59	2	4		3	3	5	
60+	2	1		4	2	3	

City: Tulsa/Oklahoma City metro; Town: pop. >20,000.

In preliminary analyses, the two younger and two older age groups each patterned together, as did the Town and Country speakers (similar to past work in Oklahoma [8, 9]), so these groups were collapsed for the present results. As seen in Table 2, distributions were skewed toward people under 40, women, and rural residents. (With only one non-binary participant, analyses with gender as a factor only included male and female speakers.)

Table 2: Total participants by age group, rurality, and gender.

	Urban		Rural			Totals		
Age Group	M	F	M	F	O	M	F	Age
Younger <40	12	22	17	32	1	29	54	84
Older >40	4	5	5	15		9	20	29
Totals	16	27	22	47	1	38	74	
	43		70					113

Urban = City; Rural = Town + Country.

2.2. Materials and procedures

Most participants recorded themselves using personal devices (phones, laptops) in 2020; 14 were recorded in a sound-attenuated booth before the COVID-19 pandemic. Each read a 129-item word list from which 69 monosyllabic words were used here, with three for each American English monophthong before /t, d/ or no coda, plus 3–6 words for each of the eight non-low vowels before /l/ [11, 12]. They also read three short passages (69-word "Please Call Stella" [13] and two new stories, 140 words each), from which 78 total stressed vowels were measured (2–8 per vowel-environment, except /ʌl/, which was mistakenly omitted from the text).

Participants were recruited primarily through a university email list and the first author's students forwarding emails to personal contacts. Consent, demographics, and instructions were presented/ collected in Qualtrics (qualtrics.com; consent was collected on paper for those recorded in person). Participants provided informed consent and received $5 cash or gift card in compensation. Students in the first author's phonetics course could opt for extra credit instead of payment (or do an alternative extra credit activity). All procedures were approved by the university's institutional review board.

2.3. Measurements and analysis

Transcripts of the word list and passages were run through DARLA [14] for forced alignment. The vowel boundaries in the resulting TextGrids were examined in Praat [15] and corrected when necessary. The aligner's method of separating vowel and lateral appeared to be consistent across vowel qualities and was therefore left unadjusted. A simple Praat script measured the first and second formants (F1, F2) at 50% of vowel duration for plain vowels and at 35% for prelaterals to reduce the effects of lateral coarticulation. The settings for most speakers were 5 formants and a maximum formant value of 5000 Hz, with some speakers adjusted to 6 formants or 5500 Hz to improve formant tracking. Each speaker's raw formants were plotted in NORM [16] with ellipses of 1 *SD* around vowel means to enable quick visual identification of outliers, which were then verified or remeasured. All speakers' formants from all vowels in both tasks were normalized together using the Nearey 2 formula in the R package phonR [17, 18].

To support descriptions of prelaterals as backed and/or raised from plain counterparts, linear mixed-effects (LME) models were run for each normalized formant of each vowel using the R package lme4 [19], with Environment (plain, prelateral) as the fixed effect and Speaker and Word as random effects.

Pillai scores were used to characterize the degree of overlap between various pairs of vowels. Pillai-Bartlett statistics ('Pillai score' for short [20]) are output from multivariate analyses of variance (MANOVA) and indicate a degree of distinction between distributions, accounting for multiple dependent variables (F1, F2). They range from 0 to 1, with lower scores indicating greater similarity. However, there are no standard cut-offs to determine merger status, and it can be difficult to define degrees of shifting between distributions that are clearly separate. Thus, indices were created to quantify the relative locations of /ʊl/ and /ʌl/, which varied between speakers.

The WOOL Raising Index (WRI for /ʊl/ in a Wells-style convention of using non-minimal pairs to label vowel classes) was calculated for each speaker as a proportion of the Euclidean distance between their mean /ul/ and /ol/, with 0 as the midpoint, +1 as even with /ul/, and –1 as even with /ol/ [21]. The same was done for /ʌl/ to create a GULF Shifting Index (GSI) between 0 at plain /ʌ/ and 1 at /ol/. (Note: three speakers with no measurable /ʌl/ tokens were excluded from GSI calculations.) The respective three vowels did not always appear in a straight line, so a perpendicular line was drawn from the target vowel to a line connecting the anchors, with the index calculated at the intersection as a proportion of the length of the anchor-line (see [11] for examples).

ANOVAs tested effects of categorical factors on WRI and GSI: Age (Younger/Older than 40), Gender (Male, Female), Rurality (Urban, Rural), and for WRI, also Style (Reading Passage, Word List) (/ʌl/ only appeared in the word list). Spearman correlations were run between WRI and GSI, and between various Pillai scores, to look for possible chain relationships between moving vowels.

3. RESULTS

Figure 1 plots all speakers' plain vowels with ellipses around vowel means. It is essentially a West-Midland arrangement, with low-back /a-ɔ/ merger, raised /e/, and notable fronting of all four /u, o, ʊ, ʌ/ [1]. (Not shown: Southern *pin-pen* merger, /aɪ/ glide loss.)

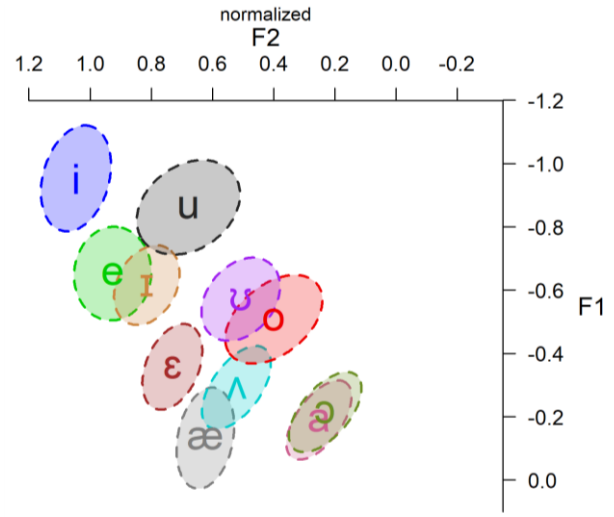


Figure 1: Plain vowels, 0.5-*SD* ellipses

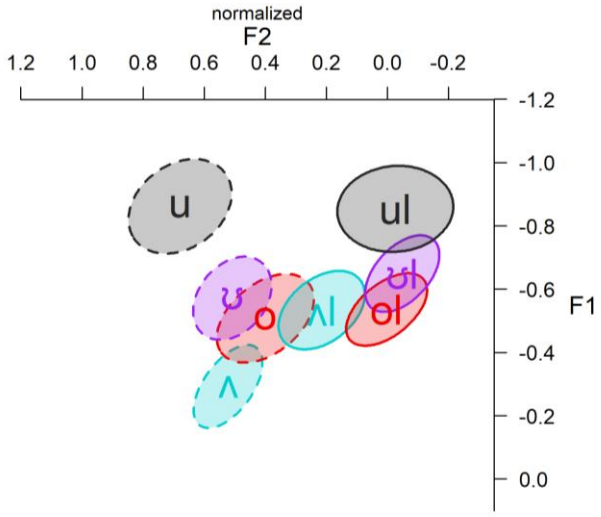


Figure 2: Non-low back/central vowels, 0.5-*SD* ellipses (dotted: plain; solid: prelateral).

Figure 2 plots all non-low back and central vowels in plain vs. prelateral environments. In line with recent work in Oklahoma [7], the prelaterals were not fronted (as in the Midland but not the South [1]). Contrary to earlier work reporting /ul/-lowering [1, 8, 9, 10], tense /ul, ol/ were nearly straight back from their plain fronted counterparts, differing from them only in F2 (/u/: $\beta = -.68$, $SE = .06$, $t(1971) = -12.25$, $p < .001$; /o/: $\beta = -.38$, $SE = .04$, $t(2184) = -9.13$, $p < .001$). Lax /ʊl/ was also slightly raised, appearing between /ol/ and /ul/ and differing from plain /ʊ/ in both F1 ($\beta = -.08$, $SE = .03$, $t(1086) = -2.95$, $p < .05$) and F2 ($\beta = -.56$, $SE = .02$, $t(1086) = -28.08$, $p < .001$). Notably, /ʌl/ was partially backed but also raised, landing front of /ol/ and differing from plain /ʌ/ in both F1 ($\beta = -.24$, $SE = .02$, $t(878) = -12.85$, $p < .001$) and F2 ($\beta = -.30$, $SE = .03$, $t(878) = -10.09$, $p < .001$). This backing was much greater than occurred with front prelaterals (not shown), which were only slightly backed and lowered among these speakers, maintaining partial to substantial overlap with their plain counterparts [12].

With all speakers together (Figure 2), /ʊl, ʌl/ were raised, but there were no clear mergers among back prelaterals. Amounts of overlap between /o/ and its neighbors were similar between environments (Pillai scores: /o-ʊ/ .30 vs. /ol-ʊl/ .28 and /o-ʌ/ .59 vs. /ol-ʌl/ .54), and raised /ʊl/ overlapped /ul/ only somewhat (Pillai .54). However, between speakers, the height of /ʊl/ and degree of shifting of /ʌl/ varied considerably.

Figure 3 shows the WOOL Raising Index (WRI) divided by speaker age, rurality, and reading style. An ANOVA with these and gender as independent variables found no significant effects, but Younger speakers showed greater individual variation, with wide distributions extending through the full range of heights between /ul/ (1 "SPOOL" in the graph) and even lower than /ol/ (–1 "GOLD"). (Additionally, 7 Younger outliers are not shown: 2 RP WRI > 2; 4 RP and 1 WL WRI < 2). In contrast, Older speakers' /ʊl/ distributions were more tightly centered a bit below 0, the midpoint between /ol, ul/, with only a few Older Rural Word Lists reaching either /ol/ or /ul/.

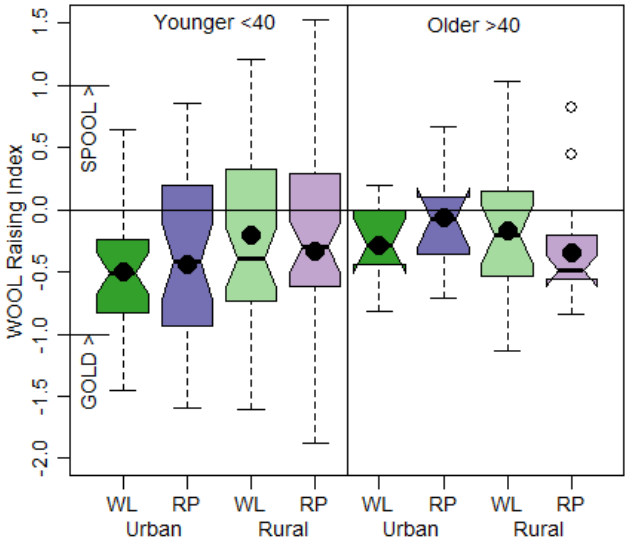


Figure 3: WOOL Raising Index (WRI, /ʊl/) by age, rurality, style (RP: passage, WL: word list; dots: means, notches: 95% CIs; 7 outliers not shown).

Figure 4 shows the effects of age, gender, and rurality on the GULF Shifting Index (GSI). An ANOVA with all three independent variables found a main effect on GSI for Gender ($F(1) = 10.27$, $p < .01$) and a three-way interaction ($F(1) = 9.83$, $p < .01$). In Tukey post-hoc comparisons, mean GSI differed significantly between Males ($M = .58$) and Females ($M = .72$; $p < .01$, 95% CI = [–.22, –.05]). This effect might be driven by Rural speakers, as mean GSI also differed significantly between Rural Males ($M = .57$) and Rural Females ($M = .74$; $p < .05$, 95% CI = [–.32, –.02]). The three-way interaction was driven by the gender difference among Older Rural speakers, the most and least raised on average (but recall the small

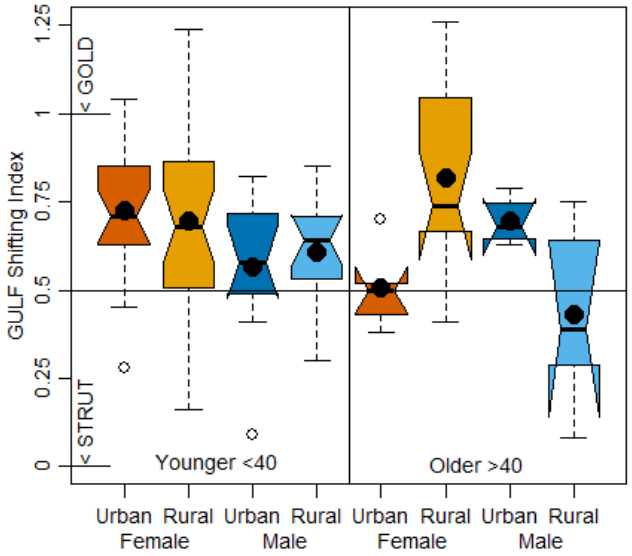


Figure 4: GULF Shifting Index (GSI, /ʌl/) by age, gender, rurality (dots: means, notches: 95% CIs).

number of Older speakers overall) (Male $M = .43$; Female $M = .82$; $p < .05$, 95% CI = [–.73, –.05]). In addition, the /ʌl/ distributions of Female speakers, especially Rural Females, extended the highest, sometimes surpassing the height of /ol/.

In testing for relationships between /ʊl/- and /ʌl/-raising, the correlation between WRI and GSI was weak but significant ($\rho = .26$, $p < .01$). However, any parallels in raising did not indicate equivalent paths toward /ul/ and /ol/ mergers, as there was no correlation between /ʊl-ul/ and /ʌl-ol/ Pillai scores.

4. DISCUSSION

This study illustrates the acoustic positions of four back prelaterals among adults in Oklahoma. To our knowledge, it is the first to identify variable raising of lax /ʊl, ʌl/, which could predict multiple possible *tensing* mergers (/ʊl-ul/, /ʌl-ol/, /ʌl-ol-ʊl/), rather than the /ul/-laxing found previously.

While the West-Midland configuration of plain monophthongs matched prior work in Oklahoma [1, 7, 8, 9, 10], prelaterals did not. Mainly, no speaker group laxed (lowered) /ul/. Instead, tense /ul/ and /ol/ were straight back from their plain fronted counterparts /u, o/, and lax /ʊl, ʌl/ raised to various degrees toward/between them. This contradicts the direction of change reportedly moving steadily across past generations, with the then-innovative /ul/-laxing led by the young, urban, educated, and female, culminating in high proportions of /ul/-laxing among speakers who would now be in this study's oldest age groups [8, 9]. Now, the apparent lack of /ul/-laxing, and /ʊl-ul/ overlap being more prevalent in rural speakers and casual styles [21], may indicate a reversal, possibly even within those generations as they aged (age grading).

However, with wide individual variation, about a third of speakers (39) across demographics did lower /ul/ in one or both reading tasks. Less than half of these overlapped /ʊl-ul/, mostly by raising /ʊl/ so the two 'met in the middle.' Only 5 speakers lowered /ul/ to an unraised /ʊl/, the pattern suggested by past descriptions of *pool-to-pull* laxing. However, most past work on Oklahoma [7, 8, 9] did not include /ʊl/ tokens (or had only 2-4 speakers with variation between them [1, 10]), so it is not clear if (near-) merger was truly achieved through /ul/-lowering.

Prior work also did not include mid /ʌl, ol/, so past states are unknown, but as younger, urban, and female speakers often lead sound changes, the patterns here suggest trends in more /ʌl/-raising but less /ʊl/-raising, which could lead to convergence around /ol/ (as found in Kansas City and Utah [2, 5]), possibly someday a three-way merger. However, only two speakers showed much /ʊl-ʌl/ overlap (Pillai < .5), none showed definitive three-way merger (all three pairs with similarly low Pillai scores), and there was no correlation between /ʊl-ol/ and /ʌl-ol/ Pillai scores to suggest convergence within speakers.

Instead, most speakers kept /ʊl/ and /ʌl/ apart, even when one or both overlapped /ol/ (usually on different sides, as in Figure 5a, 5f). Although weak, the correlation between /ʊl/ and /ʌl/ indices suggests a possible chain shift. A raising shift could keep all four prelaterals separate (Figure 5a) or result in two lax-tense mergers (/ʊl-ul/ *pull-pool*, /ʌl-ol/ *hull-hole*, Figure 5b). Or, opposition to /ʊl/-raising could lead to /ʊl-ol/ *bull-bowl* merger (Figure 5d, [3, 4]). With /ʌl/ apparently trending toward raising, it could stay close to /ol, ʊl/ to form a three-way near-merger (Figure 5e, 5f, found in Utah [5]) or even surpass the two in height (Figure 5c, 5f). All these patterns are attested in the data with varying frequencies similar to recent findings in Midland Kansas City [2].

So much variation, particularly in younger ages, indicates a community in flux. Future work will examine trajectories, lip rounding, and social factors like education, SES, and rootedness [11] to explain patterns of change in prelateral shifting.

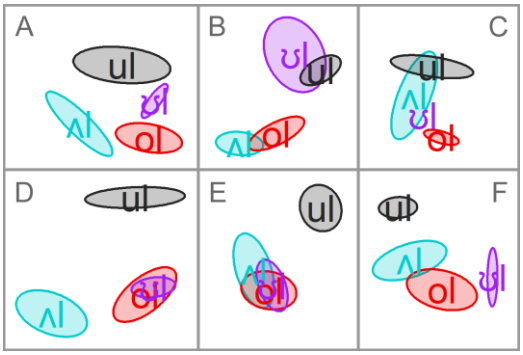


Figure 5: Example prelateral patterns, individual speakers (all Young, B-F Female, C-E Urban).

7. REFERENCES

[1] Labov, W., Ash, S., Boberg, C. 2006. *The atlas of North American English: Phonetics, phonology, and sound change.* Mouton de Gruyter.

[2] Strelluf, C. 2016. Overlap among back vowels before/l/in Kansas City. *Language Variation and Change 28*(3), 379–407.

[3] Squizzero, R. 2018, Apr. Bull or bowl? A production study of prelateral back vowel mergers in Pacific Northwest English. Cascadia Workshop in Sociolinguistics, Portland, OR.

[4] Stanley, J. 2017, Jan. The perception and production of two vowel mergers in Cowlitz County, Washington. American Dialect Society, Austin, TX.

[5] Stanley, J., Morgan Johnson, L. 2022, Jan. Vowels can merge because of changes in trajectory: Prelaterals in rural Utah English. Linguistic Society of America, Washington, DC.

[6] Arnold, L. 2015. Multiple mergers: Production and perception of three pre-/l/ mergers in Youngstown, Ohio. *Penn Working Papers in Linguistics 21*.

[7] Bakos, J. 2013. *Comparison of the speech patterns and dialect attitudes of Oklahoma*. Doctoral dissertation, Oklahoma State University.

[8] Bailey, G., Wikle, T., Tillery, J., Sand, L. 1993. Some patterns of linguistic diffusion. *Language Variation and Change 5*, 359–390.

[9] Bailey, G., Wikle, T., Tillery, J., Sand, L. 1996. The linguistic consequences of catastrophic events: An example from the American Southwest. In Arnold, J. et al. (eds), *Sociolinguistic Variation: Data, Theory, and Analysis,* 435–451.

[10] Thomas, E. R. 2001. *An acoustic analysis of vowel variation in new world English.* Duke University Press.

[11] Landers, M. 2022. Rootedness and pre-lateral mergers in Oklahoma. Masters thesis, Oklahoma State University.

[12] Freeman, V., Landers, M. 2020, Sept. Prelateral mergers in Oklahoma. Linguistic Association of the Southwest, Online.

[13] Weinberger, S. 2015. Speech Accent Archive. George Mason University. http://accent.gmu.edu.

[14] Reddy, S., Stanford, J. 2015. A Web application for automated dialect analysis. *Proc. NAACL-HLT 2015*.

[15] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer, version 6.1.16. http://www.praat.org.

[16] Thomas, E. R., Kendall, K. 2007. NORM: The vowel normalization and plotting suite, version 1.1. http://lingtools.uoregon.edu/norm/norm1.php

[17] R Core Team. 2020. R: A language and environment for statistical computing, version 4.0.2. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

[18] McCloy, D. 2016. phonR: tools for phoneticians and phonologists. http://drammock.github.io/phonR/. [R package v. 1.0-7.]

[19] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Statistical Software 67*, 1–48. [R package v. 1.1.27.1.]

[20] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *J. Phonetics 34*, 458–484.

[21] Freeman, V., Landers, M. 2021, Oct. Back prelateral mergers in Oklahoma: Variation in production. New Ways of Analyzing Variation, Online

Investigating the Irish English PEN-PIN merger: patterns and origins

Kate Tallon

Phonetics and Speech Laboratory, Trinity College Dublin
tallonka@tcd.ie

ABSTRACT

This study investigates the status and sociolinguistic patterning of the PEN-PIN merger in the English of L1 Irish speakers. An Irish origin has been suggested for the merger, which has been reported for (south)western Irish English varieties. By quantitatively investigating the merger in L1 Irish speakers' English alongside (i) their Irish and (ii) monolingual English of the same region, this study aims to shed light on its possible Irish influence. English DRESS tokens (n=530) from 12 L1 Irish speakers were acoustically and statistically analysed alongside (i) the same speakers' Irish /ɛ/ tokens (n=142) and (ii) English DRESS tokens (n=503) of 12 monolingual English speakers. Results show that the bilingual speakers demonstrate a more advanced PEN-PIN merger than monolinguals. A similar merger in the bilingual speakers' Irish was not found. These findings advance our knowledge of this understudied English variety and challenge an Irish source theory for the PEN-PIN merger.

Keywords: Irish English, PEN-PIN merger, DRESS-raising, language contact, sound change.

1. INTRODUCTION

A PEN-PIN merger, in which the DRESS vowel is raised in pre-nasal position, has been reported for western and southwestern varieties of Irish English (IrE). Descriptions of these varieties dating back as far as the 15th century [1] indicate that DRESS-raising was once an unconditioned process that has gradually become nasally-conditioned. Although this is generally acknowledged, there is a lack of modern quantitative data confirming this pattern.

Regarding the social conditioning of the PEN-PIN merger in western IrE, Peters [1] found it to be more common among older speakers, indicating its recession. However, this observation was based on a small dataset (3 speakers) and, as acknowledged by Peters, remains to be verified with additional data. With the exception of Peters' study, it is unclear to what extent social factors may condition the merger. Additionally, further descriptions of the PEN-PIN merger in IrE are largely anecdotal with little by way of quantitative analysis.

As is the case with many characteristic features of IrE, the origins of the PEN-PIN merger in remain elusive. Kallen [2] and Henry [3] highlight the possibility that it is a retention from Early Modern English (EModE) which, according to Wyld [4], demonstrated raising of /ɛ/ before /s, l, n/ + consonant. He supports this claim with EModE spellings such as <Gintlemen> and <till> (for 'tell'). Other linguists [5] [6] have attributed the merger to an Irish transfer source; Connacht (western) and Munster (southern) dialects of Irish reportedly have raising of mid vowels in nasal environments [5] e.g. *seinm* [ʃɪnʲəmʲ] ('playing') [7]. The mirrored geographical patterning of raised /ɛ/ between Irish and IrE adds further weight to the possibility of an Irish source.

The possibility that the merger is an independent development (external to contact-induced factors) must also be considered. Hickey [8] highlights the possibility of a more generally universal phonetic process, whereby the nasal consonant triggers vowel raising (lowering of F1) due to its formant structure, i.e. the vowel undergoes anticipatory coarticulation with its following nasal consonant.

Notably, other varieties of English, such as those of Southern US and Californian English, also display the PEN-PIN merger. The question remains as to whether these dialects have the merger due to a universal phonetic influence or due to past Irish or British English input to the Southern states.

In the investigation of an Irish influence on the PEN-PIN merger one area remaining untapped is the study of L1 Irish speakers of English. The English of these speakers reportedly exhibits extensive Irish influence [9], [5], [10]. Hypothetically, if the merger is influenced or reinforced by Irish we might expect relatively advanced merger in the English of these L1 Irish speakers.

With the aims of describing the status of the PEN-PIN merger, investigating its sociolinguistic patterning and exploring its possible Irish origins, this study addresses the following research questions:

1. To what extent do L1 Irish speakers of English show a PEN-PIN merger?
1.a If the merger is present, to what extent is it predicted by social and linguistic factors?

2. In terms of the PEN-PIN merger, how does the English of L1 Irish speakers compare with (i) their Irish (ii) the English of monolingual (English) speakers?

2. METHODS

2.1. Participants

Twenty-four participants were recruited from the Connemara Gaeltacht, an Irish-speaking region on the west coast of Ireland. Twelve of these participants were bilingual L1 Irish speakers of English (6 M, 6 F; age range 20-79, median age 38, SD 19.4); and 12 were monolingual English speakers (6 M, 6 F, age range 27-66 , median age 55, SD 16).

2.2. Data collection

Participants were recorded in quiet locations within their own homes. Recordings were made at a 44.1 kHz sampling rate via Audacity [11] using a 2017 Apple Macbook Pro and an AT2020 USB microphone. Along with a social survey, all participants completed a short semi-structured interview in English and read a set of English sentences containing monosyllabic words from the DRESS (incl. PEN) and KIT lexical sets. The L1 Irish participants recorded an additional Irish dataset, comprising a short semi-structured interview and a series of sentences. The 3 resulting datasets were: (i) English and (ii) Irish of bilingual speakers; (iii) English of monolingual speakers.

2.3. Analysis

Recordings were transcribed and segmented in Praat [12]. A Praat script was written to extract F1 and F2 measurements for all vowel tokens at midpoint. For L1 Irish speakers a total of 983 English vowel tokens were collected (530 DRESS; 453 KIT), along with 281 Irish tokens (142 /ɛ/ and 139 /ɪ/). For monolingual English speakers 855 tokens were collected (503 DRESS; 352 KIT). To facilitate interspeaker comparison, the formant measurements were Z-score normalized in R [13].

The normalized F1 and F2 measurements were visualized in various iterations on 2D density plots using the *ggplot2* [14] package in R. This facilitated exploratory analysis of the data, including the identification of distributional patterns and central tendencies of the vowel categories. Three vowel categories were plotted for each dataset: (i) /ɪ/ vowels (English KIT) (ii) pre-nasal /ɛ/ (English PEN) (iii) pre-oral consonant /ɛ/ (English DRESS).

Statistical analysis was subsequently carried out to further quantify the presence of a merger. Pillai scores were calculated (via MANOVA tests in R) to measure the degree of overlap between vowel categories for each speaker [15]. Pillai scores range from 0 to 1 with a lower score indicating greater overlap between categories, i.e. a more advanced merger.

To investigate the social patterning of the merger, multiple linear regression models, using *lme4* [16] were run for each dataset. By-speaker Pillai scores were analysed as the dependent variable with predictors of age, sex, local attachment score (LAS), linguistic profile score (LPS). The latter variable was created to account for variation present in speakers' language use, proficiency and dominance beyond a binary monolingual-bilingual measure.

3. RESULTS

As seen in Figure 1, there is a clear difference in the height of pre-nasal /ɛ/ (PEN) vowels between bilingual (L1 Irish) and monolingual (English) groups. The former group show more advanced pre-nasal /ɛ/-raising (i.e. a PEN-PIN merger).

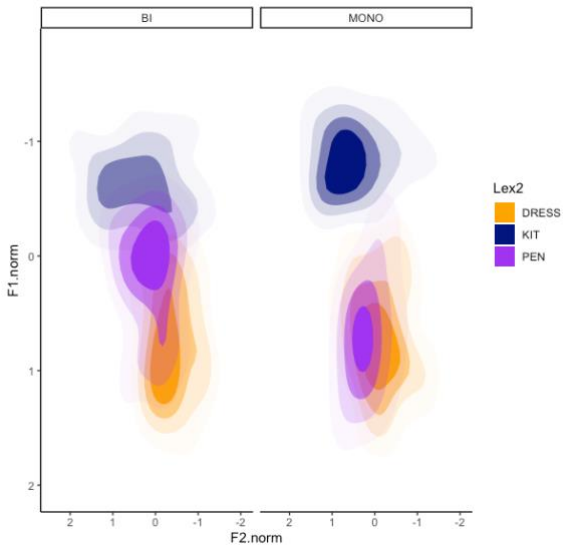


Figure 1. Distribution of English DRESS, PEN and KIT for bilingual (Irish-English) and monolingual (English) speakers.

This variation was reflected in the by-speaker Pillai scores which ranged from 0.02 (bilingual speaker with advanced merger) to 0.71 (monolingual speaker

with distinct vowel categories). On average, bilingual speakers scored 0.21 and monolinguals 0.49, i.e. the PEN-PIN merger was notably more advanced for bilingual speakers.

Regarding the effect of social factors, the multiple linear regression model for bilingual speakers revealed both speaker sex and age to be significant predictors of merger (both $p<.05$), with no age-sex interaction effect. Males and older speakers were more likely to show merger. Overall, the model accounted for a significant amount of variance in the bilingual speakers' Pillai scores ($F(4,7)=4.13$, $p<.05$, $R^2=.70$, $R^2_{adjusted}=.53$). The positive correlation between increased speaker age and Pillai score is seen in Figure 2.

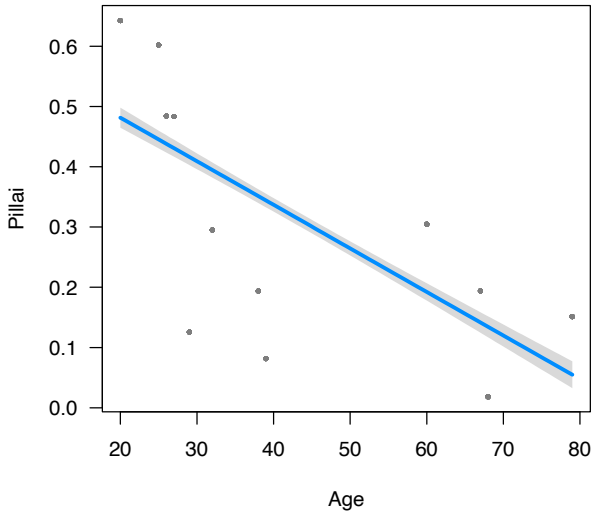


Figure 2. Relationship between bilingual speaker Pillai score and age.

For the monolingual speakers, the multiple linear regression model did not reveal any significant predictors of merger. Figure 3 demonstrates the by-sex distributional patterns of bilingual and monolingual speakers. There is a clear lack of variation between Pillai scores for monolingual males and females when compared to the bilingual speakers.

Analysis of the bilingual speakers' Irish /ɪ/ and pre-nasal /ɛ/ vowel tokens found a Pillai score of 0.3. However, when visualized, it is evident that the distributional pattern of these Irish vowels greatly differs from that of the English. For example, as seen in Figure 4, older speakers (age 50+) show prominent fronting of Irish pre-nasal /ɛ/ vowels.

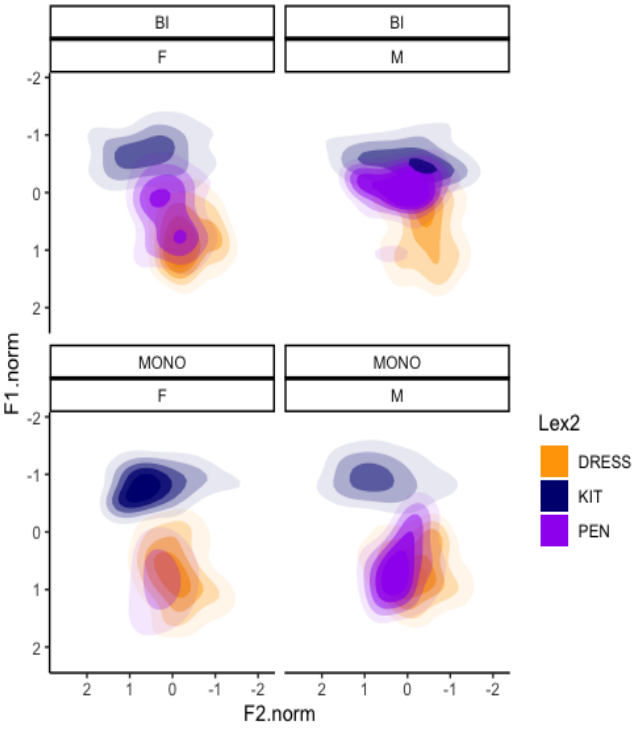


Figure 3. By-sex distribution of English DRESS, PEN and KIT for bilingual (Irish-English) and monolingual (English) speakers.

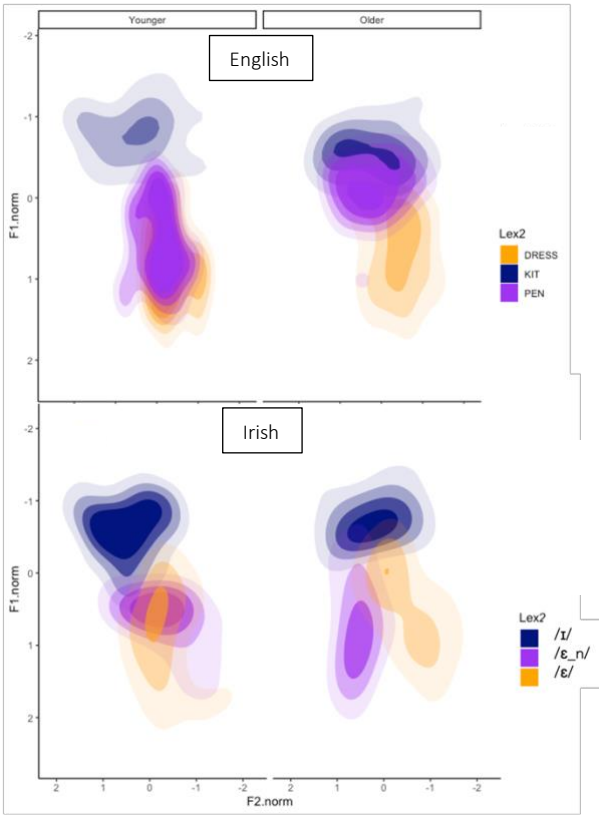


Figure 4. By-age distribution of bilingual speakers' English and Irish vowel tokens.

Additionally, the bilingual speakers' Irish vowels did not mirror their English by-sex distributional pattern. The bilingual male speakers clearly demonstrated a more advanced English PEN-PIN merger than the females. However, the Irish data suggests fronting of the pre-nasal /ɛ/ vowel for male speakers, with no clear raising pattern.

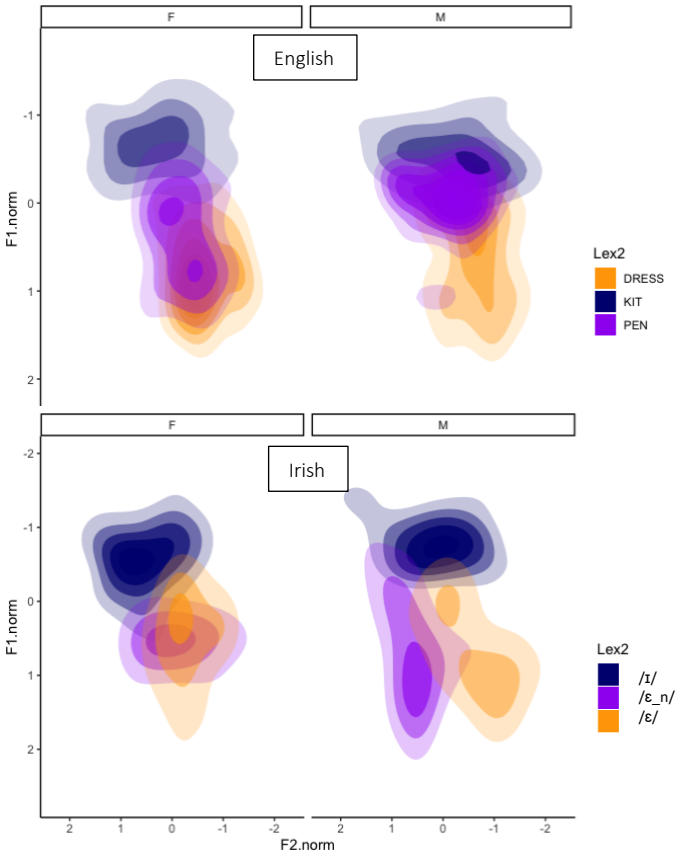


Figure 5. By-sex distribution of bilingual speakers' English and Irish vowel tokens.

4. DISCUSSION AND CONCLUSION

This study aimed to shed light on the status and sociolinguistic patterning of the PEN-PIN merger in Connemara Gaeltacht English. As the first to quantitatively analyse vowel data from Irish-English bilinguals' both languages, alongside monolingual English data, a secondary aim was to investigate the possible Irish influence on the PEN-PIN merger.

Regarding the sociolinguistic patterning of the merger, this study's findings support Peters' [1] claim that the PEN-PIN merger is a receding feature in western IrE; while monolingual speakers did not show any clear merger pattern, increasing age was found to be a significant predictor of merger for bilingual speakers. This correlation between speaker age and degree of merger is suggestive of an ongoing sound change.

Relative to monolingual English speakers, the L1 Irish bilinguals showed more advanced PEN-PIN merger overall. On first inspection, this pattern could be taken to suggest that the bilingual speakers PEN-PIN merger must be attributed to their L1 Irish. However, an investigation of these speakers' Irish vowels suggests that there is no pre-nasal /ɛ/-raising to mirror that of their English. Interestingly, there appears to be a distinct pattern of distribution within their Irish, i.e. fronting of pre-nasal /ɛ/. This finding is significant for two reasons: (i) it has not been previously reported for this dialect of Irish and (ii) it challenges the idea that Irish has played a significant role in the presence of the PEN-PIN merger in Irish English.

Although the origins of the PEN-PIN merger cannot be definitively proven, these preliminary findings cast doubt on Irish reinforcement of merger in these speakers. In departure from past claims of Irish origins, it appears that in this group of bilingual speakers, Irish does not play an essential role (e.g. reinforcement/transfer) in the presence of /ɛ/-raising in their English. This advances our knowledge of the English of L1 Irish speakers and adds much-needed data-driven findings to the existing linguistic descriptions of IrE. It is hoped that this study, as the first of its kind, will be considered a move toward more quantitative data and analyses for the field of Irish English linguistics and will encourage the analysis of Irish data alongside English in the investigation of Irish influence on Irish English.

6. REFERENCES

[1] A. Peters, *Linguistic change in Galway City English / Arne Peters*. 2016, pp. 0-0.

[2] J. L. Kallen, "Irish English Volume 2: The Republic of Ireland," in *Irish English Volume 2: The Republic of Ireland*: De Gruyter Mouton, 2013.

[3] P. L. Henry, *An Anglo-Irish Dialect of North Roscommon*. 1957.

[4] H. C. Wyld, *A history of modern colloquial English*. TF Unwin, Limited, 1921.

[5] Ó. Baoill and P. Dónall, "The emerging Irish phonological substratum in Irish English," in *Focus on Ireland*, J. Kallen Ed., 1997, pp. 73-87.

[6] M. D. O'Sullivan, *Old Galway: the history of a Norman colony in Ireland*. W. Heffer, 1942.

[7] B. Ó Curnáin, *The Irish of Iorras Aithneach County Galway*. Institute for Advanced Studies, 2007.
[8] C. Carignan, "An examination of oral articulation of vowel nasality in the light of the independent effects of nasalization on vowel quality," Associazione Italiana Scienze della Voce, 2018.
[9] R. Ó hÚrdail, "Hiberno-English: Historical Background and Synchronic Features and Variation," in *The Celtic Englishes*, H. Tristram Ed., 1997, pp. 180-99.
[10] L. O'Cuinneagain, "To What Extent Do Irish Fluency and Gender Affect Prevalence of Irish-influenced Phonological Features in Irish English?," University of York, 2019.
[11] *Audacity(R): Free Audio Editor and Recorder*. (2022). [Online]. Available: https://audacityteam.org/
[12] *Praat: doing phonetics by computer*. (2021). [Online]. Available: http://www.praat.org/
[13] *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* (2022). [Online]. Available: https://www.R-project.org/
[14] *ggplot2: Elegant Graphics for Data Analysis*. (2016). Springer-Verlag, New York. [Online]. Available: https://ggplot2.tidyverse.org
[15] J. Nycz and L. Hall-Lew, "Best practices in measuring vowel merger," in *Proceedings of Meetings on Acoustics 166ASA*, 2013, vol. 20, no. 1: Acoustical Society of America, p. 060008.
[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1 - 48, 10/07 2015, doi: 10.18637/jss.v067.i01.

AGE- AND GENDER-BASED VARIATION IN THE PERCEPTION OF VOICING CONTRAST IN TOKYO JAPANESE*

Yoonjung Kang[1], Manami Hirayama[2]

University of Toronto Scarborough[1], Seikei University[2]
yoonjung.kang@utoronto.ca, hirayama@fh.seikei.ac.jp

ABSTRACT

Recent studies report a sound change in progress in Tokyo Japanese, whereby word-initial voiced stops are frequently devoiced, and VOT alone is no longer a sufficient or reliable cue to distinguish the voicing contrast. The current study examines how Tokyo Japanese speakers of different age and gender use VOT, as well as the following vowel's pitch (F0) and voice quality (h1-h2) in voicing perception. 140 speakers of Tokyo Japanese, balanced for age and gender, participated in an online perception experiment. The majority of speakers made use of all three cues, but among the three cues, some speakers relied relatively more on VOT, while others relied more on the vocalic cues, especially F0. We found a significant interaction of F0 cue use with pitch accent, gender, and age, whereby the dominant cue for voicing perception shifts from VOT to F0, and this change is led by accented words and younger females.

Keywords: Japanese, sound change, perception, VOT, F0

1. INTRODUCTION

Recent studies on Tokyo Japanese report age- and gender-based variation in the realization of the word-initial stop voicing contrast [1-3]. Specifically, word-initial voiced stops vary between prevoiced and devoiced realizations, and the rate of devoicing is higher for younger than older speakers [1-3] and for female than male speakers [2]. This change creates an overlap in VOT (Voice Onset Time) of the voiced and voiceless categories, and the contrast cannot be reliably distinguished by VOT alone.

Production studies show that the voicing contrast is signalled by other secondary cues, namely, the pitch and voice quality of the following vowel, as well as VOT. The pitch (F0) is higher when the vowel follows a word-initial voiceless stop than a voiced stop [4-7], and the voice quality of the vowel is breathier (i.e., higher h1-h2) for voiceless than voiced stops [5, 6].

Studies have also investigated the perceptual cues that listeners use to distinguish initial stop voicing. For instance, [8] examined the perception of a monosyllabic word pair manipulated to vary in F0 and VOT (positive VOT values only) and found that the effect of F0 is only noticeable for very short VOT values. The study involved college students from four dialect regions, including Kanto, which encompasses Tokyo. Another study ([9]) created two sets of stimuli that represented different pitch accent conditions: *pasu* 'pass' vs. *basu* 'bus' for initial accented words (henceforth #HL) and *teki* 'enemy' vs. *deki* 'result' for unaccented words (henceforth #LH). The stimuli were manipulated to vary in F0 and VOT, covering both positive and negative VOT values. They found that the effect of F0 was more pronounced for the #HL than the #LH condition, which is consistent with [7]'s finding in production that initial stop voicing has a stronger effect on F0 for H-initial than L-initial words. However, it is unclear whether the interaction of pitch accent and F0 cue use is due to the larger pitch range in the #HL stimuli or whether listeners weigh the F0 cue more in the #HL condition.

Finally, one study [1] examined the age-based variation in perception of voicing contrast, using naturally produced tokens of stops. The majority of the participants in the study were from Tokyo but included speakers from other dialect regions. The study found that errors that misidentify devoiced voiced stops as voiceless were no more frequent with older listeners than younger listeners, contrary to the expected pattern given that older speakers are less likely to devoice voiced stops than younger listeners in production. Instead, overall, more errors were found for the younger speakers and their errors were more widespread, regardless of whether the stop was prevoiced or not. However, the observation about age difference is difficult to interpret as the low count and the limited distribution of errors by older listeners may be a function of the low number of older participants.

Building on these previous studies, our study examines the perceptual cues used by Tokyo speakers to identify word-initial voiced and voiceless stops, and how cue use varies as a function of pitch accent and speakers' age and gender. Given the report that the change in Tokyo Japanese is led by younger females and that the VOT cue is weakening in the speech of younger and/or female speakers, we expect the sound change to be reflected in our data. In particular, we hypothesize that female and younger speakers will exhibit more innovative cue weighting

and pay more attention to vocalic cues compared to male and older counterparts. This is based on the assumption that the speaker's own production and perception patterns are generally expected to align with each other (cf. [10]). While similar changes have been reported in other dialects, we focus on Tokyo because of a larger body of previous studies to build upon.

Regarding pitch accent, the larger F0 difference between voiced and voiceless stops in the #HL condition compared to the #LH condition presents three potential interactions between F0 and pitch accent in perception. The first possibility is that if the range of F0 variation in the perceptual stimuli is comparable across the pitch accent conditions, listeners may give equal weight to both the #HL and #LH pitch accent conditions. The second possibility is that, given the relative salience of F0 cues in the #HL condition, listeners may weigh F0 cues more for #HL words than for #LH words. The third possibility is that listeners are less sensitive to F0 cues for the #HL condition, necessitating a larger F0 difference to shift the voicing boundary for #HL words to match the large difference observed in production.

2. METHODS

The stimuli consisted of two minimal pairs, 手前 [temae] vs.出前 [demae] (unaccented, #LH) and 天使 [tenɕi] vs. 電子 [denɕi] (initial accented, #HL). The selection of coronal stop-initial words was based on the avoidance of labials, which tend to include English loans, and dorsals, which tend to exhibit less overlap in VOT between voiced and voiceless stops [2, 7]. These words are also frequently used and considered to be familiar to native speakers of (Tokyo) Japanese [11].

The stimuli words were produced by the second author, a female Tokyo Japanese speaker in her 40s, with 10 to 12 repetitions. For each word pair, four baseline tokens were generated by splicing prevoicing (for the negative VOT baselines) or aspiration (for the positive VOT baselines) with one of two base vowel tokens (one each from a voiced and a voiceless stop production).

The duration and F0 contour of the rest of the word were manipulated to match the average of all tokens for that pair, removing potential duration and pitch cues to voicing present in the natural production. The intensity of each spliced part (prevoicing, aspiration, and vowel) was also adjusted to closely match the speaker's average value.

The manipulation parameters and produced range for each acoustic dimension for each accent condition are summarized in Table 1, determined based on the stimuli talker's production. The VOT was varied in 10 steps from -60 ms to 50 ms, at 15 ms intervals for negative values and 10 ms intervals for positive values. The F0 at the following vowel onset (at 9.1% of the vowel duration) was varied in six equidistant steps from -2.5 to 2.5 in normalized semitone for each pair. h1-h2 was not directly manipulated, but baseline tokens typical for each word were chosen. The three acoustic dimensions were orthogonally varied for each word pair, creating 240 stimuli (=10 VOT steps * 6 F0 steps * 2 (h1-h2) baseline vowels * 2 pitch accent word pairs). The manipulations were conducted in Praat [12].

	Stimuli values	Produced range	
		#LH	#HL
VOT (ms)	[-60, -45, -30, -15, 0, 10, 20, 30, 40, 50]	[-64 ~ 58]	[-41 ~ 28]
F0 (st)	[-2.5, -1.5, -0.5, 0.5, 1.5, 2.5]	[-2.5 ~ 2.5]	[-4.1 ~ 4.0]
h1-h2 (dB)	#LH: [-7.1, 0.3] #HL: [-11.4, -8.5]	[-14.8 ~ 16.0]	[-18.1 ~ 2.5]

Table 1: Acoustic parameters for the perception stimuli and the ranges for the produced tokens

	60s+	50s	40s	30s	20s
Female	15	14	12	14	11
Male	14	15	16	14	15

Table 2: Age and gender breakdown of participants included in the analysis

Self-identified Tokyo Japanese speakers from Tokyo, Chiba, Saitama, or Kanagawa were recruited through an online crowdsourcing recruitment site (crowdworks.jp). A total of 172 speakers participated, out of which four were excluded for responding "no" to the question "Do you speak Tokyo-style Japanese?" Additionally, 28 speakers who answered "yes" to the same question but also listed other dialects they speak were also excluded. The breakdown of the 140 speakers included in the analysis by age and gender is presented in Table 2. The task was word identification, whereby participants heard stimuli and chose the word they heard. The full experiment, built and conducted on the Gorilla platform (gorilla.sc) [13], included informed consent, a background questionnaire, a production task, and a perception task, and it took an average of 19.5 minutes to complete. The perception experiment alone took 9.4 minutes.

For statistical analysis, we used the lme4 package [14] in R [15] to build a logistic mixed-effects regression model. The model takes the RESPONSE (voiced = 0, voiceless = 1) as the response variable and four linguistic predictors (VOT, F0, BASE.VOWEL, ACCENT), two speaker-level predictors

(year of birth (YOB), GENDER), and their full interactions as fixed effects predictors. To compare their relative strength, the phonetic variables, VOT (ms), F0 (st), and BASE.VOWEL (voiced = 0, voiceless =1), were transformed into z-scores to put them on a comparable scale. ACCENT and GENDER were sum-coded (#LH = -0.5, #HL = 0.5; F = -0.5, M = 0.5), while YOB was centered. We also included by-PARTICIPANT random intercept and random slope adjustments for the three phonetic predictors, which we explore in detail to examine cue use variation across listeners (cf. [16]). The formula is shown in (1) below. We used an alpha level of 0.05 for significance.

(1) `glmer(RESPONSE ~ VOT * F0 * BASE.VOWEL * ACCENT * YOB * GENDER + (1 + VOT + F0 + BASE.VOWEL | PARTICIPANT),"binomial")`

3. RESULTS

Figure 1 plots the proportion of voiceless responses by the four linguistic conditions (VOT, F0, BASE.VOWEL, and ACCENT) pooled across all participants. The plots illustrate, and the statistical test confirms that all three phonetic predictors have significant effects. Overall, a higher VOT, a higher F0, and a voiceless BASE.VOWEL all lead to more voiceless responses. We can also observe that these predictors interact, and the statistical model shows a significant four-way interaction (VOT * F0 * BASE.VOWEL * ACCENT). In other words, the effects of the predictors vary across different contexts.

Figure 2 displays the breakdown of the data by age and gender of the participants. Although the overall patterns are similar across different age and gender groups, there is some variation. The statistical analysis revealed that the six-way interaction of all predictors was significant, indicating that the effects of the linguistic and speaker-level predictors on the proportion of voiceless responses varied across different participant subgroups. See the full results at http://individual.utoronto.ca/yjkang/icphs_j_fit.pdf.

In this paper, we solely focus on the main effects of the three phonetic cues and their interaction with non-phonetic predictors. These interactions provide insights into how cue use changes depending on the pitch accent condition and the age and gender of the participant. Table 3 summarizes the coefficient estimates and only presents interactions with significant effects. The intercept coefficient estimate (not shown in the table) is 0.702. The scatterplots in Figure 3 display the distribution of individual participants' estimated slopes for the phonetic cues calculated from the model.

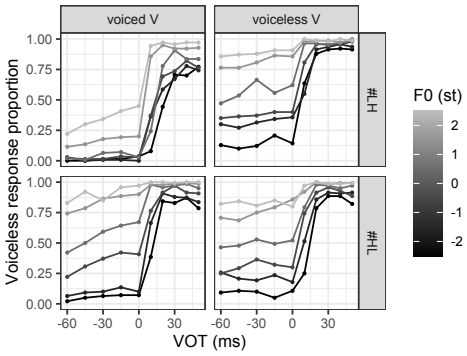


Figure 1: Proportion of voiceless responses by VOT, F0, BASE.VOWEL, and ACCENT, across all participants

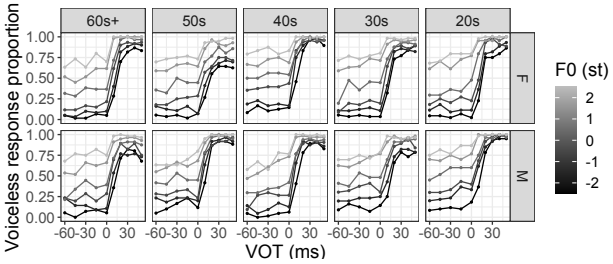


Figure 2: Proportion of voiceless responses by VOT and F0, separated by AGE and GENDER of participants.

	VOT	F0	VOWEL
main effects	2.098	1.591	0.647
2-way interaction			
* accent (#HL-#LH)	-0.573	0.326	-1.631
* gender (M-F)	(-0.054)	-0.239	0.137
* year of birth (yob)	(-0.007)	(0.053)	(0.012)
3-way interaction			
* accent * gender	(-0.165)	-0.241	(-0.011)
4-way interaction			
* acc. * gender * yob	(-0.093)	-0.193	(-0.158)

Table 3: Coefficient estimates of the main effects of the three phonetic predictors (VOT, F0, BASE.VOWEL) and interactions with non-phonetic predictors. Parentheses indicate non-significance effects.

All three main effects of the phonetic predictors are significant, but they interact significantly with ACCENT, suggesting that cue weights vary depending on the pitch accent condition. This can also be seen from the comparison of the top vs. bottom panels of Figure 3. For the #LH word pair [temae/demae], on average, VOT is a much stronger cue ($\beta = 2.098 + (-0.573 * -0.5) = 2.385$) than F0 ($\beta = 1.428$) or BASE.VOWEL ($\beta = 1.463$). Also note that in the top panels of Figure 3, all participants' coefficients are above 0, which means that everyone used all three cues in the expected direction. For the accented #HL

word pair [tenɛi]-[denɛi], on the other hand, the average coefficients are comparable between VOT (β = 1.812) and F0 (β = 1.754), but they differ in variability. In the bottom panels of Figure 3, individual slopes for F0 are tightly clustered around 2, while VOT coefficients are more variable, with some even falling below 0. BASE.VOWEL coefficients are distributed around 0, indicating that the cue does not have a consistent effect and that many participants use the cue in the opposite direction.

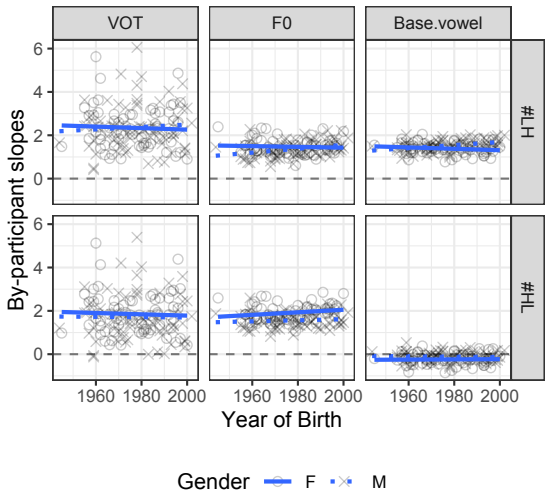


Figure 3: By-participant slope estimates for phonetic predictors, plotted by ACCENT and by the participants' GENDER and YOB. The dashed horizontal lines mark the slope of 0 and the solid and dotted lines are linear smooths for female and male participants.

Regarding the speaker-level predictors, AGE and GENDER, if individuals' perception reflects the sound change in progress where female and younger speakers are more likely to devoice voiced stops and reduce the effectiveness of the VOT cue in their speech than male and older speakers, we expect female and younger speakers to weight vocalic cues (F0 in particular) more than male and older speakers respectively, while VOT cues should show the opposite trend. This prediction holds for F0. The significant interactions of F0 * GENDER, F0 * ACCENT * GENDER, and F0 * ACCENT * GENDER * YOB show that overall, the F0 cue is stronger for females than males, and this interaction is stronger for the #HL [tenɛi]~[denɛi] pair and for younger participants. On the other hand, VOT does not show significant interactions with YOB or GENDER.

Turning to individuals' use of cues, we can classify the participants according to the cue they pay the most attention to, that is, the cue with the highest coefficient. Table 4 tabulates the number of participants by their dominant cues. For example, among the female participants in their 20s (n = 10), VOT was the best cue for four, and F0 was the best cue for the other six. None of the 10 participants weighted BASE.VOWEL as the most important cue.

The age- and gender-based variation in dominant cues at the individual level mirrors the significant interactions of F0 with GENDER, AGE and ACCENT in the model. For the unaccented (#LH) pair, VOT is still the dominant cue regardless of participants' age and gender, while for the accented (#HL) pair, which overall shows raised sensitivity to F0 and reduced sensitivity to VOT, proportionally more females use F0 as the dominant cue (57.6% = 38 out of 66) than males (50.0 % = 37 out of 64). Also, note that the majority dominant cue shifts from VOT to F0 as we move from older to younger speakers. In other words, we see the cue shifting from VOT to F0 for the younger speakers, and the change is more advanced in female participants and for accented words.

	Age	Female			Male		
		VOT	F0	vowel	VOT	F0	vowel
#LH	60s+	13*	2	0	10*	2	2
	50s	9*	3	2	12*	0	3
	40s	10*	0	2	14*	1	1
	30s	11*	2	1	9*	2	3
	20s	8*	1	2	11*	2	2
#HL	60s+	8*	7	0	8*	6	0
	50s	5	9*	0	10*	5	0
	40s	4	8*	0	9*	7	0
	30s	6	8*	0	4	10*	0
	20s	4	6*	0	6	9*	0

Table 4: The distribution of participants' dominant phonetic cue based on their AGE, Gender, and the ACCENT. The asterisk (*) denotes the majority pattern for each demographic subgroup.

4. DISCUSSION

In this paper, we investigated the perceptual cue use in the context of sound change in progress for Tokyo Japanese stop voicing. Our findings suggest that pitch accent plays a significant role in the cue weighting pattern, which is consistent with the production differences observed in the stimuli talker's speech (Table 1) and reflects the speech patterns of Tokyo Japanese speakers in general [7]. Specifically, F0 differed more by voicing for the #HL pair, while VOT and BASE.VOWEL quality differed more for the #LH. We also found listeners' age and gender had an effect on perception, as expected. For the #HL pair only, younger and female listeners relied more heavily on F0 than older and male listeners. This interaction of F0 with age and gender affected the relative importance of VOT, which was relied on relatively less by younger female listeners compared to F0. Overall, our results suggest that the dominant cue for the initial voicing contrast is shifting from VOT to F0, led by #HL words and by younger female listeners.

5. REFERENCES

[1] Takada, M. 2004. VOT tendency in the initial voiced alveolar plosive /d/ in Japanese and the speakers' age. *Journal of the Phonetic Society of Japan* 8(3), 57–66.

[2] Takada, M. 2011. *Nihongo-no Gotoo Heisaon-no Kenkyuu: VOT-no Kyoojiteki Bunpu-to Tuujiteki Henka* [*Research on the word-initial stops of Japanese: Synchronic distribution and diachronic change in VOT*]. Kurosio.

[3] Takada, M., Kong, E., Yoneyama, K., Beckman, M. E. 2015. Loss of prevoicing in modern Japanese /g, d, b/. *Proceedings of the 15th ICPhS.*

[4] Shimuzu, K. 1999. A study on phonetic characteristics of voicing of stop consonants in Japanese and English. *Journal of the Phonetic Society of Japan* 3(2), 4–10.

[5] Kong, E. J., Beckman, M. E., Edwards, J. 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *Journal of Phonetics* 40(6), 725–744.

[6] Takada, M., Kong, E. J., Yoneyama, K., Beckman, M. E. 2015. Do pitch and voice quality cue word-initial "voicing" in Tôhoku Japanese? *Poster presented at the 24th Japanese/Korean Linguistics Conference.*

[7] Gao, J., Arai, T. 2019. Plosive (de-)voicing and F0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics* 77, 100932.

[8] Byun, H.-G. 2021. Perception of Japanese word-initial stops by native listeners. *Phonetics and Speech Sciences* 13(3), 53–64.

[9] Gao, J., Yun, J., Arai, T. 2019. VOT-F0 coarticulation in Japanese: Production-based or misparsing? *Proceedings of the 16th ICPhS.*

[10] Schertz, J., Clare, EJ. 2020. Phonetic cue weighting in perception and production. *WIREs Cognitive Science* 11: e1521.

[11] NTT Communication Science Laboratories. 2021. *NTT Lexicon Database: Word Familiarity (2020 resurvey and enlarged edition*). NTT Printing Corporation.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.05, retrieved 5 January 2022 from http://www.praat.org/

[13] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[14] Bates, D., Mächler M., Bolker B., Walker S., Christensen, B. R. H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, Pavel N. 2022. Linear mixed-effects models using Eigen and S4. R package version 1.1-20.

[15] R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna. R Foundation for Statistical Computing.

[16] Drager, K., Hay, J. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change,* 24(1), 59-78.

* This study was supported by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209) and the Center for Asian and Pacific Studies at Seikei University.

INCREASING ASPIRATION OF SWISS GERMAN PLOSIVES: A SOUND CHANGE IN PROGRESS?

Franka Zebe, Camille Watter, and Stephan Schmid

University of Zurich
franka.zebe@uzh.ch, camille.watter@uzh.ch, stephan.schmid@uzh.ch

ABSTRACT

In Alemannic dialects of German-speaking Switzerland, the primary cue between lenis and fortis plosives is closure duration, with lenis plosives having a shorter closure than fortis plosives, while both are phonetically voiceless. Recently, it has been called into question whether there is an increasing tendency for speakers to additionally produce aspirated fortis plosives, possibly due to the contact to German Standard German. To investigate this, we recorded word-initial and word-medial fortis plosives produced by 24 older and 24 younger speakers of Zurich German and analysed their normalized VOT values.

Results show that, although the word-medial plosives can probably all be considered unaspirated, younger speakers overall produced significantly longer VOT values compared to older speakers. Word-initial plosives differed much more between age groups and also resulted in considerable variability for certain words. These new results hint at a possible sound change in progress.

Keywords: VOT, aspiration, sound change, Swiss German, Zurich German

1. INTRODUCTION

It is general consensus that the main acoustic cue to the phonemic contrast between homorganic lenis and fortis plosives in Alemannic dialects of German-speaking Switzerland is closure duration: lenis plosives display shorter closures compared to fortis plosives, while both are phonetically voiceless [1, 2, 3, 4]. Recently, it has been observed that another category is evolving, namely aspirated plosives [5, 6]. Hence, we can assume three phonetic categories of Swiss German plosives: lenis, unaspirated fortis, and aspirated fortis, where the phonemic status of the latter category is unclear [6]. Using the apparent-time construct [7, 8], the aim of this study is to investigate how older and younger speakers of Zurich German differ in their VOT productions in word-medial and word-initial fortis plosives.

1.1. Swiss German plosives

Recent studies have revealed an increasing amount of word-initial aspirated fortis plosives in Alemannic dialects of German-speaking Switzerland [5, 6]. [6] investigated a set of young speakers of Zurich German. Results show that the difference between lenis and unaspirated fortis plosives is stable in that its acoustic correlate primarily lies in closure duration, not in VOT [6]. There is, however, also evidence for a third category of aspirated fortis plosives in word-initial position. First and foremost, it is clear from their study that aspiration only occurs in a certain set of words, in particular loanwords from German Standard German (GSG) or English [6]. Furthermore, low frequency words seem to be aspirated more often than high frequency words [6]. Secondly, findings indicate a high degree of variability in the words that are aspirated [6].

The goal of the current study is to add onto the evidence from [6] by including a group of older speakers and investigating not only word-initial but also word-medial plosives.

1.2. Sound change

When dealing with a sound change in progress, the question usually arises whether the change is internally or externally driven. Traditionally, it has been argued that a regular sound change takes place in a phonetically gradual manner and is lexically independent, while an irregular sound change occurs in a phonetically abrupt manner due to the process of dialect levelling driven by external factors, e.g. when dialect speakers produce features similar to that of the standard variety [9]. These theoretical implications raise an issue for the current situation in German-speaking Switzerland. Since it has been shown that aspirated plosives occur in loanwords from GSG or even English, it would be logical to assume that the sound change is driven by external factors, since in those two languages aspiration is used as the main cue to distinguish between lenis and fortis plosives [10, 11, 12]. In addition, recent studies have shown that in other southern German varieties with high contact to GSG VOT has become increasingly important for distinguishing between

lenis and fortis plosives, namely in Standard Austrian German [13] and in Western Central Bavarian [14].

On the other hand, there is evidence for both lexical diffusion (in the sense of [15]) and a certain amount of interspeaker variability in Zurich German [6]. These two aspects would be an indication for an internally driven sound change in progress. In recent years, however, the evidence of continuous sound changes due to dialect levelling has been increasing [16, 17, 18].

Therefore, the results of the current study will have to be discussed also in relation to the general nature of this (possible) sound change in progress.

2. METHODS

2.1. Participants and stimuli

A corpus of read speech was collected from 24 older speakers (12 female) aged between 53 and 81 years (mean=66.54, SD=6.34) and 24 younger speakers (12 female) aged between 20 and 34 years (mean=27.96, SD=3.52) at the time of the recording.

Stimuli consisted of 24 words, 13 with a word-initial fortis plosive and 11 with a word-medial fortis plosive as target. Target words were either disyllabic trochees or monosyllables; thus, they did not differ substantially in phonological word length. This set of stimuli is part of a larger study with the same participants. For the current production study, however, the stimuli were limited to 24 words containing fortis plosives. All words were embedded in a carrier sentence. All in all, we recorded and analyzed a corpus of 1152 items (48 speakers x 24 stimuli/plosives).

2.2. Recording procedure

Participants were recorded in a soundproof booth at the Phonetics Laboratory of the University of Zurich, using a personal computer with the interface USBPre® 2 (Sound Devices) and the microphone NT2-A (RØDE) with the recording software SpeechRecorder [19]. The speakers were instructed to read each sentence once. All recordings had a sample rate of 16-bit/44.1 kHz and were saved as .wav-files. Speakers were reimbursed for their participation.

2.3. Data preparation and measurements

Each recorded sentence was automatically segmented using WebMAUS [20], selecting *German Dieth (CH)* in the language annotation settings. The recordings were then manually adjusted using the EMU-webApp [21] according to the the following procedure: The beginning and end of each sentence was adjusted as well as the boundaries of each segment of the target word. In addition, we annotated two phases within the plosives, i.e. the closure phase and the VOT, which was positive in all cases. The VOT values and target word durations were obtained in milliseconds using the emuR package [22] in R [23]. The VOT values were then normalized by dividing their absolute duration by the duration of the entire word to account for differences in articulation rate. The same normalization procedure has been adopted by [18]. In addition, the word frequency in terms of logSubtlex values (log lexical frequency per million) for each word were obtained from the Subtlex corpus [24].

2.4. Statistical analysis

For the statistical analysis, we fitted two linear mixed-effects models using the R packages lme4 [25] and lmerTest [26]. The first model was fitted for the word-medial plosives with the normalized VOT values *VOTnorm* as dependent variable. Fixed effects consisted of *word* and *age* (old vs. young). Lastly, random intercepts for *speaker* were added. The second model for the word-initial plosives had the same structure.

To obtain the Chi-square and p values reported in the results section, a Type II ANOVA for each model was calculated using the R package car [27]. In case of a significant interaction, pairwise comparisons using Tukey's test were calculated using the R package emmeans [28].

3. RESULTS

The word frequencies in terms of logSubtlex values are shown in Table 1a and 1b.

Word	logSubtlex
buute ('to boot')	0
Sitte ('manner')	1.292
räppe ('to rap')	1.415
Sippe ('clan')	1.531
Wappe ('coat of arms')	1.663
Matte ('mat')	1.851
Huupe ('car horn')	1.954
Latte ('lath')	1.964
Kööter ('dog')	2.121
Kaater ('tomcat')	2.297
Schatte ('shadow')	2.873

Table 1a: logSubtlex values for target words with plosives in word-medial position.

As can be seen, the values for the word-medial target plosives are slightly lower compared to those with word-initial target plosives. Still, we assume that the differences between word-medial and word-initial plosives are not due to the slightly different word frequencies.

Word	logSubtlex
Torf ('peat')	0.477
paare ('to mate')	1.748
Part ('part')	2.025
Paan (Greek god 'Pan')	2.041
Pèèrli ('couple', dim.)	2.041
Paar ('couple')	2.072
Poole ('Poland')	2.176
Tipp ('tip')	2.553
Tuur ('tour')	2.651
Päch ('bad luck')	2.786
Taate ('acts')	3.032
Tee ('tea')	3.176
Pass ('passport')	3.629

Table 1b: logSubtlex values for target words with plosives in word-initial position.

3.1. Word-medial plosives

	Chisq	Df	Pr(>Chisq)
word	177.1698	10	< 2e-16 ***
age	4.6053	1	.03187 *
word:age	15.5738	10	.11250

Table 2: Statistical output of the linear mixed-effects model for the word-medial plosives.

The linear mixed-effects model for the word-medial plosives revealed significant main effects for both word and age, while the interaction between the two did not turn out to be significant, which can also be seen in Table 2. Younger speakers produced longer VOT values compared to older speakers in all words, as is illustrated in Figure 1.

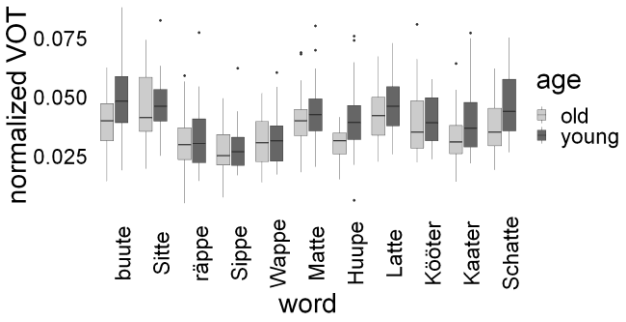


Figure 1: Normalized VOT values (y axis) for the words containing word-medial target plosives (x axis), ordered by word frequency; older speakers in light grey, younger speakers in dark grey.

Even though younger speakers produced significantly longer VOT values, it must be emphasized that these values are overall very short and can probably all be classified as unaspirated, especially when compared to the initial plosives.

3.2. Word-initial plosives

	Chisq	Df	Pr(>Chisq)
word	1061.1147	12	< 2.2e-16 ***
age	8.5101	1	.003532 **
word:age	26.8246	12	.008189 **

Table 3: Statistical output of the linear mixed-effects model for the word-initial plosives.

The model for the word-initial plosives revealed significant main effects for *word* and *age*. The interaction between the two also turned out to be significant, as is shown in Table 3. Again, younger speakers produced overall longer VOT values, although there were some exceptions.

Since the interaction turned out to be significant, pairwise comparisons between age groups for each word were calculated. The following words were significantly different between age groups: *Torf* ('peat') (z=-3.187, p=.001), *Part* ('part') (z=-3.308, p<.001), *Paan* (Greek god 'Pan') (z=-2.149, p=.032), *Poole* ('Poland') (z=-2.773, p=.006), and *Tee* ('tea') (z=-3.225, p=.001). In all cases, younger speakers produced longer VOT values compared to older speakers. This can also be seen in Figure 2.

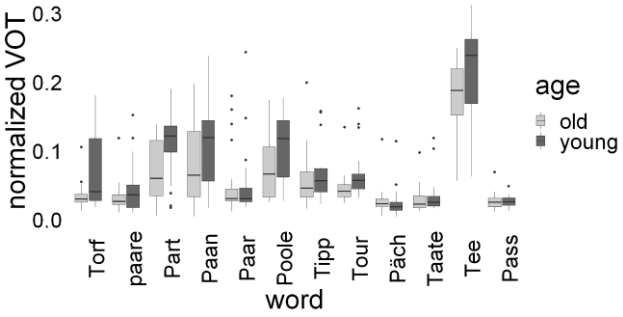


Figure 2: Normalized VOT values (y axis) for the words containing word-initial target plosives (x axis), ordered by word frequency; older speakers in light grey, younger speakers in dark grey.

It is also important to note that for the words that significantly differ between age groups (which are arguably undergoing a sound change), there is a much higher amount of variability compared to the other words, whose plosives can be considered unaspirated. Interestingly, this is not only the case for the younger but also the older speakers, indicating that the older generation is also part of the sound change in progress. Furthermore, it is a clear

indication that the change is gradual rather than abrupt.

4. DISCUSSION

4.1. Discussion of the results

First of all, our results indicate that there is a sound change in progress for word-initial fortis plosives, some of which are produced with aspiration, primarily by younger speakers. As for the word-medial plosives, although younger speakers produce significantly longer VOT values compared to older speakers, ultimately one cannot speak of a sound change in progress, as the values are overall very short. Still, word-medial plosives should also not be disregarded in future investigations.

As already pointed out by [6], the current results provide evidence for the fact that only certain words are affected by the sound change in progress. We can also confirm that, in general, lower frequency words are more aspirated than higher frequency words. Even so, it is important to note that the word frequencies of the items with initial target plosives are slightly higher with respect to those with medial target plosives. Still, within the category of word-initial plosives, words with lower logSubtlex values seem to be affected more than those with higher values. Furthermore, there is a considerable amount of variability not only for younger but also for older speakers in the words with significantly different VOT values between age groups.

This brings us back to the above mentioned question of whether this sound change in progress is internally or externally driven. Even though the change is clearly not of an abrupt manner, which would be typical for an externally driven sound change, we cannot conclude that the change is internal. The evidence that we deal with lexical diffusion seems to be the stronger argument here, leading us to the conclusion that this sound change in progress is, at least in part, due to the contact with GSG.

The issue then relates to the process of lexical diffusion itself. On the one hand, it must be emphasized that one cannot rule out completely that a sound change is both internally and externally driven. On the other hand, there is an increasing amount of evidence that the specific progress of dialect levelling is not necessarily abrupt but can also be of a gradual manner [16, 17, 18]. Based on the results of the current study, we conclude, at least for now, that the latter is also the case for the increasing aspiration in Zurich German, which allows for the possibility of a sound change being both externally driven and continuous.

4.2. Outlook

The results of the current study provide evidence on the specific sound change in progress taking place in Alemannic dialects of German-speaking Switzerland on the one hand and on the general phenomenon of sound change, specifically of the lexical diffusion type, on the other hand.

Still, there remain some important issues that need to be addressed in future research. Firstly, perceptual data is needed in order to better understand the sound change in progress in Swiss German dialects. Secondly, not only VOT but also its relationship to closure duration should be investigated. Lastly, a larger amount of words containing fortis plosives is needed both to confirm that mostly loanwords are produced with aspiration and to gather more information on how word frequency plays a role in this sound change in progress.

5. ACKNOWLEDGEMENTS

This study was funded by the Swiss National Science Foundation (SNSF), grant Nr. 190005.

6. REFERENCES

[1] Kraehenmann, A. 2001. Swiss German Stops: Geminates All over the Word. *Phonology* 18(1), 109–145.

[2] Kraehenmann, A., Lahiri, A. 2008. Duration Differences in the Articulation and Acoustics of Swiss German Word-Initial Geminate and Singleton Stops. *J. Acoust. Soc. Am*. 123(6), 4446–4455.

[3] Fleischer, J., Schmid, S. 2006. Zurich German. Journal of the International Phonetic Association 36(2), 243–253.

[4] Willi, U. 1996. *Die segmentale Dauer als phonetischer Parameter von 'fortis' und 'lenis' bei Plosiven im Zürichdeutschen: Eine akustische und perzeptorische Untersuchung*. Franz Steiner.

[5] Schifferle, H. 2010. Zunehmende Behauchung. Aspirierte Plosive im modernen Schweizerdeutsch. In: H. Christen et al. (eds.) *Alemannische Dialektologie: Wege in die Zukunft*. Steiner, 43–55.

[6] Ladd, R., Schmid, S. 2018. Obstruent Voicing Effects on F0, but without Voicing: Phonetic Correlates of Swiss German Lenis, Fortis, and Aspirated Stops. *Journal of Phonetics* 71, 229–248.

[7] Labov, W. 1994. *Principles of language change: Internal factors*. Blackwell.

[8] Bailey, G., Wikle, T., Tillery, J., Sand, L. 1991. The apparent time construct. *Language Variation and Change* 3, 241–264.

[9] Beckman, M. E., Li, F., Kong, E. J., Edwards, J. 2014. Aligning the timelines of phonological acquisition and change. *Laboratory Phonology* 5(1), 151–194.

[10] Lisker, L., Abramson, A. S. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *WORD* 20(3), 384-422.
[11] Jessen, M. 1998. *Phonetics and phonology of tense and lax obstruents in German*. John Benjamins Publishing.
[12] Beckman, J., Jessen, M., Ringen, C. 2013. Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of linguistics* 49(2), 259-284.
[13] Luef, E. M. 2020. Development of voice onset time in an ongoing phonetic differentiation in Austrian German plosives: Reversing a near-merger. *Zeitschrift für Sprachwissenschaft* 39(1), 79–101.
[14] Kleber, F. 2018. VOT or quantity: What matters more for the voicing contrast in German regional varieties? Results from apparent-time analyses. *Journal of Phonetics* 71, 468–486.
[15] Labov, W. 2007. Transmission and diffusion. *Language* 83(2), 344–387.
[16] Harrington, J., Kleber, F., Reubold, U. 2012. The production and perception of coarticulation in two types of sound changes in progress. In: S. Fuchs, M. Weirich, D. Pape, & P. Perrier (eds.) *Speech planning and Dynamics*. Peter Lang, 39–62.
[17] Bukmaier, V., Harrington, J., Kleber, F. (2014). An analysis of post-vocalic /s-ʃ/ neutralization in Augsburg German: Evidence for a gradient sound change. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.00828
[18] Jochim, M., Kleber, F. 2022. Fast-Speech-Induced Hypoarticulation Does Not Considerably Affect the Diachronic Reversal of Complementary Length in Central Bavarian. *Language and Speech*. https://doi.org/10.1177/00238309221127641
[19] Draxler, C., Jänsch, K. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). http://www.lrecconf.org/proceedings/lrec2004/pdf/242.pdf
[20] Schiel, F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. *Proc. of the 14th ICPhS*. San Francisco, 607–610.
[21] Winkelmann, R., Raess, G. 2014. *Introducing a Web Application for Labeling, Visualizing Speech and Correcting Derived Speech Signals*.
[22] Winkelmann, R., Jaensch, K., Cassidy, S., Harrington, J. 2021. *emuR: Main Package of the EMU Speech Database Management System*.
[23] RStudio Team. 2020. *RStudio: Integrated Development for R. RStudio*. PBC, Boston.
[24] Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., Böhl, A. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58(5), 412–424.
[25] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.
[26] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models, *Journal of Statistical Software* 82(13), 1–26.
[27] Fox, J., Weisberg, S. 2018. *An R companion to applied regression*. Thousand Oaks, CA: Sage publications.
[28] Lenth, R. V. 2021. *emmeans: Estimated Marginal Means, aka Least-Squares Means*.

Voice quality is not an obligatory stage in tonogenesis: A case study of Eastern Khmu

Sireemas Maspong

Department of Linguistics, Cornell University
Institute for Phonetics and Speech Processing (IPS), LMU Munich
s.maspong@phonetik.uni-muenchen.de

ABSTRACT

This paper examines the role of voice quality in the development of tonal and vowel height contrasts that accompany tonogenesis. The study focused on Eastern Khmu because it maintains a contrastive onset voicing, unlike other varieties of Khmu which have undergone tonogenesis and lost this contrast. The study shows that pitch and vowel height differences on following vowels accompany the onset voicing contrast without any concomitant voice quality difference. These findings suggest that voice quality is not an obligatory factor in deriving tonal and vowel height contrasts. Additionally, the study suggests that a tonal or vowel height language can emerge directly from a reanalysis of onset-induced acoustic perturbations such as pitch and vowel height without a stage where voice quality is contrastive.

Keywords: Tonogenesis, Register, Tone, Voice quality, Vowel height, Khmu

1. INTRODUCTION

Cross-linguistically, particularly in the Austroasiatic language family, there exist languages in which onset voicing in the proto-language corresponds to present (f0-based) tonal and vowel height contrasts. The proto-onset voicing also corresponds to present register contrasts in some other languages. Register typically consists of a binary contrast, high and low register, that encompasses laryngeal (such as voice quality and pitch) and supralaryngeal properties (such as vowel quality) [1, 2]. The contrasts are summarized in Table 1.

Contrasts	*pa	*ba	Example
Tonal (f0)	Higher f0	Lower f0	N. Khmu
Vowel height	More open vowels	More close vowels	Standard Khmer
Register	Higher f0 Modal voice More open vowels	Lower f0 Breathy voice More close vowels	Mon

Table 1: Summary of contrasts conditioned by onset voicing

The present contrasts emerge from the loss of onset voicing [3] through a process called tonogenesis [4] or registrogenesis [5]. Some scholars propose that a stage with contrastive voice quality, which is a part of register contrast, is an obligatory intermediate stage between the stage with onset voicing contrast and the stage with f0-based tonal contrast or vowel height contrast [e.g., 3, 6]. In other words, voice quality is considered an obligatory stage in emergence of pitch contrasts [7, 8] or vowel height contrasts [8]. However, recent studies have shown that a pitch contrast can derive directly from the loss of onset voicing contrast without going through an intermediate stage with a voice quality contrast [9]. Although evidence for the emergence of vowel height contrast has not been found, the possibility of deriving contrasts without an obligatory intermediate stage suggests alternative pathways, as summarized in Fig. 1.

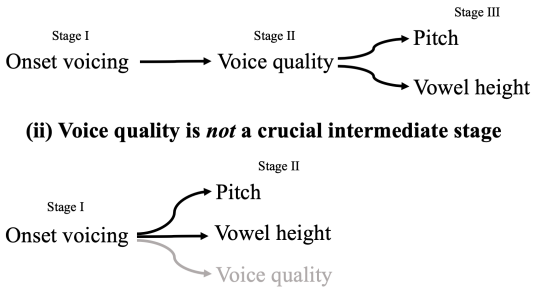


Figure 1: Two possible pathways of tonogenesis/registrogenesis

In this paper, I aim to investigate whether voice quality is an obligatory factor in deriving f0-based tonal contrast and vowel height contrast. There are two opposing hypotheses: (i) voice quality is an obligatory factor, and (ii) voice quality is not an obligatory factor.

To explore this, I focus on Eastern Khmu, an Austroasiatic language at the first stage of tonogenesis where onset voicing contrasts are still present. According to Premsrirat [10], there are Eastern Khmu dialects with onset voicing contrast (Stage I in Fig. 1-i), Northern Khmu dialects with

register (voice quality-based) contrast (Stage II in Fig. 1-i), and Northern and Western Khmu dialects with tonal (f0-based) contrast (Stage III in Fig. 1-i).

If voice quality is an obligatory factor (Fig. 1-i), we would expect that in Eastern Khmu, when the difference in voice quality of the vowel following voiced and voiceless onsets is not present, there should not be differences in pitch and vowel height. On the other hand, if voice quality is not an obligatory factor (Fig. 1-ii), the differences in pitch and vowel height can be present, even if the difference in voice quality is not present. The predictions in the case that the voice quality difference is not present are summarized in Table 2.

Differences	(i) Voice quality is obligatory	(ii) Voice quality is *not* obligatory
Voice quality	×	×
Pitch	×	✓
Vowel height	×	✓

Table 2: Predictions in the case that the voice quality difference is *not* present for the hypothesis

The findings of the acoustic study we present indicate that the predicted differences in voice quality are not present in the vowels following voiced and voiceless onsets, yet differences in pitch and vowel height are observed. These results suggest that pitch and vowel height can undergo transphonologization directly from onset voicing, bypassing an intermediate stage where voice quality is contrastive.

2. METHODS

2.1 Data collection

The study was conducted with 19 native speakers of E. Khmu (14F; 5M). All participants are bilingual in Thai. None of the participants reported any speech or hearing impairments. The target words consisted of monosyllabic or sesquisyllabic Khmu words with /iː, uː, eː, oː, aː/ vowels in the main syllables. The onsets of the target syllables were limited to alveolar and velar consonants, and included voiced stops, voiceless (unaspirated) stops, and voiceless aspirated stops. The participants were prompted to produce the target words in a carrier sentence with five repetitions, using Thai translations of the target words displayed on a screen and pronounced by a native Thai speaker. This method was chosen due to the lack of a writing system in E. Khmu and the possibility of some participants being unable to read Thai. Audio recordings of the participants were made using a Shure SM10A head-mounted microphone and an audio interface Roland OCTA-CAPTURE connected to a PC laptop, with a sampling rate of 44.1 kHz.

2.2 Data processing

The acoustic measurements extracted from the vowel of the target words were fundamental frequency (f0), First Formant (F1), and the amplitude difference between the first and second harmonics (H1-H2) and Cepstral Peak Prominence (CPP), which have been reported to successfully distinguish voice quality [11]. To extract f0, a cross-correlation algorithm in Praat [12] was used (window size: 13.3 ms for male, 10 msec for female; time step: 3.3 ms for male, 2.5 ms for female). For F1, the Berkform toolbox [13], a MATLAB toolbox for formant tracking with a Robust LPC algorithm, was used. H1-H2 was obtained from the inverse-filtered FFT spectra using a function in the Voicebox toolbox [14] in MATLAB. Finally, CPP was obtained from real cepstra extracted by a function in the Signal Processing toolbox [15] and calculated following the method in [16]. The window size and time step of F1, H1-H2, and CPP tracking were three times the size of those used for f0 tracking.

2.3 Data analysis

The acoustic trajectories of target vowels were analyzed using Generalized Additive Mixed Models (GAMMs) fitted with the *R* package `mgcv` [17]. Each acoustic measure (f0, F1, H1-H2, CPP) was used as a dependent variable, while onset voicing was the main predictor. Prior to the analysis, each acoustic trajectory was time-warped to have 40 datapoints using MATLAB's linear interpolation function. The subject was included as random smooths. Places of articulation (alveolar and velar) were included as a fixed effect in the null models with F1 as a response. For the corresponding alternative, an interaction with onset voicing was also included. Place of articulation is included only for model with vowel /iː, uː, aː/, since they have a full cross set of the onset voicing and the place of articulation.

Separate models were fitted for each vowel phoneme: /iː, uː, eː, oː, aː/. In total, there were 20 models (4 cues x 5 vowels) that were each compared with a corresponding null model without the fixed effect of onset voicing using a likelihood ratio test.

3. RESULTS

3.1 F0

Based on Table 1, we expect that vowels following voiced onsets will exhibit a lower f0 compared to vowels following voiceless onsets. The results reveal

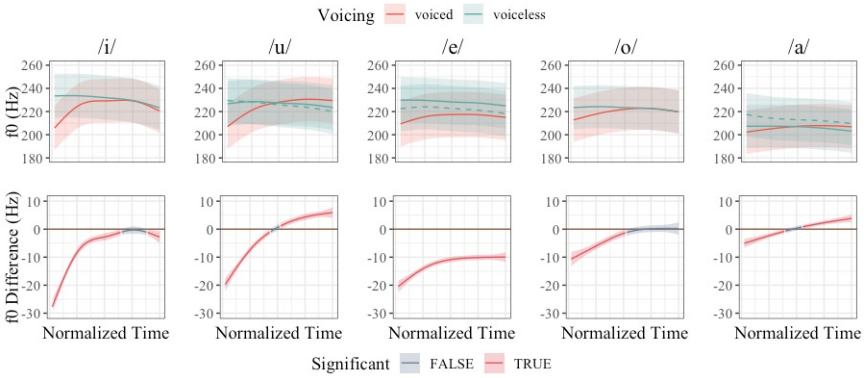


Figure 2: GAMM fitted f0 trajectories following voiced (red) and voiceless stops (green) (top) and the predicted mean difference across onset categories (bottom); dashed lines represent vowels following voiceless aspirated stops

that, as predicted, f0 values are lower following voiced onsets, at least at the onset of all vowels. Vowel /e:/ maintains the difference in f0 between following voiced and voiceless onsets throughout the entire trajectory, although the differences are more significant at the onset of the trajectories. The vowels following voiceless aspirated stops pattern with those following voiceless stops, as illustrated in Fig. 2 above.

3.2 F1

According to Table 1, vowels following voiced onset are expected to display lower F1 (higher vowel) than those following voiceless onsets. The results indicate that F1 values are lower after voiced onsets, as expected, at least at the beginning of all vowels. Vowels /i:, e:, o:/ show a sustained difference in F1 between following voiced and voiceless onsets throughout the entire trajectory, with larger differences at the onset of the trajectory. Vowels following voiceless aspirated stops display similar patterns to those following voiceless stops, as illustrated in Fig. 2.

Moreover, the effect of voicing persists despite the presence of place of articulation being included as a fixed effect. Vowels following velar onsets have lower F1 than those following alveolar onsets, but the effect of voicing remains evident, as shown in Fig. 3 below.

3.3 H1-H2

According to Table 1, we expect vowels following voiced onset to display higher H1-H2 (breathy voice) than vowels following voiceless onsets (modal voice). The results show that only vowel /o:/ displays higher H1-H2 following voiced onsets, as expected, and retains the difference in H1-H2 throughout the whole trajectory. However, unexpected patterns are observed for the other vowels. Vowels /e:, a:/ display higher H1-H2 values following voiced onsets, but not at the beginning of the trajectories. The trajectories of vowels /i:, u:/ display lower H1-H2 following voiced onsets than following voiceless onsets (breathy following voiceless onsets). The difference very small (~1–2 dB) for all vowels. Vowels following voiceless aspirated stops, again, pattern with those following voiceless stops, except for vowel /a:/. See Fig. 4 on the next page.

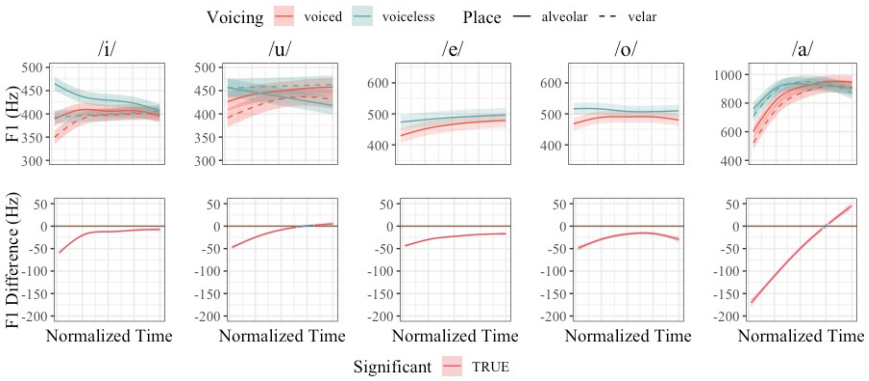


Figure 3: GAMM fitted F1 trajectories following voiced (red) and voiceless stops (green) (top) and the predicted mean difference across onset categories (bottom); dashed lines represent vowels following velar stop

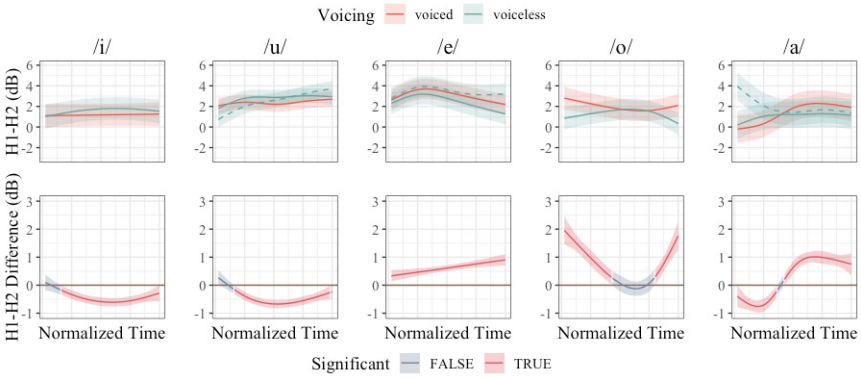


Figure 4: GAMM fitted H1-H2 trajectories following voiced (red) and voiceless stops (green) (top) and the predicted mean difference across onset categories (bottom); dashed lines represent vowels following voiceless aspirated stop

3.4 CPP

According to Table 1, it was expected that vowels following voiced onset would display lower CPP (breathy) than vowels following voiceless onsets (modal). However, the results show that only vowel /e:/ displays lower CPP following voiced onsets than following voiceless onsets, and retains the difference in CPP throughout the whole trajectory. The difference is more prominent in the middle of the trajectories. Unexpected patterns (breathy following voiceless onsets) are observed for CPP trajectories of the other vowels. These trajectories display higher CPP following voiced onsets than following voiceless onsets, either at the beginning of the trajectories or throughout the whole trajectories. The difference is minimal (~1 dB) for all vowels. The corresponding figures can be found in Fig. 5 below.

4. DISCUSSION

The findings of this study indicate that the E. Khmu variety exhibits differences in f0 and F1 values across vowels following voiced and voiceless onsets. However, while some vowels exhibit differences in voice quality across onset voicing categories, these differences are not consistent and do not align with the predictions as outlined in Table 2-ii. Thus, I conclude that voice quality is not an obligatory factor in deriving a tonal contrast based on f0 and a vowel height contrast based on F1 in this dialect.

These results suggest that an onset voicing language such as the E. Khmu could directly develop a tonal contrast or a vowel height contrast from a reanalysis of f0 and F1 differences, similar to other Austroasiatic languages, without the need for a prior contrast in voice quality. Moreover, the saliency of the observed differences at the beginning of the vowels indicates that they are likely an effect of onset voicing, a pattern that has been observed in other languages [e.g., f0: 18, 19, 20; F1: 21, 22]. Previous research has attributed f0 and F1 differences accompanying voicing to articulatory mechanisms that either facilitate or hinder voicing, such as larynx lowering, tongue root advancement [23, 24, 25, 26], or vocal fold tension [27]. We thus hypothesize that onset voicing in E. Khmu may involve one or more of these mechanisms.

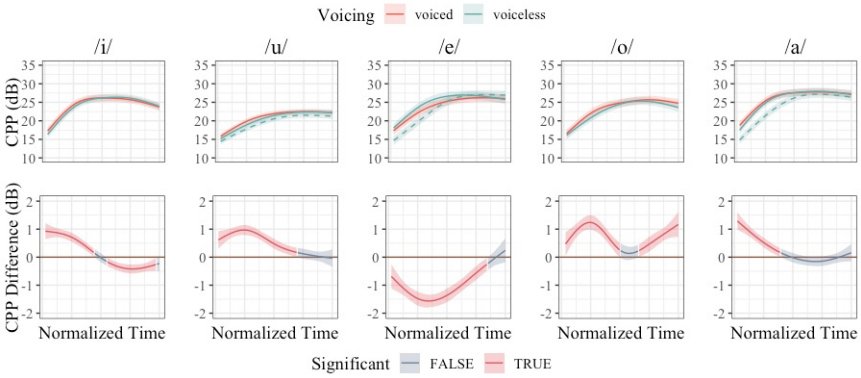


Figure 5: GAMM fitted CPP trajectories following voiced (red) and voiceless stops (green) (top) and the predicted mean difference across onset categories (bottom); dashed lines represent vowels following voiceless aspirated stops

5. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Professor Sam Tilsen, Professor James Kirby, Professor Abby Cohn, and Professor John Whitman for their valuable input and stimulating discussions. I am also thankful to the members of the Cornell Phonetics Lab for their generous support and helpful feedback throughout the research process.

6. REFERENCES

[1] Brunelle, M., Kirby, J. 2016. Tone and phonation in Southeast Asian languages. *Language and Linguistics Compass* 10(4). 191-207.

[2] Brunelle, M., Tạ, T. 2021. Register in languages of Mainland Southeast Asia: the state of the art. In: Sidwell, P., Jenny, M. (eds.), *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*. De Gruyter Mouton, 683–706.

[3] Haudricourt, A. 1965. Les mutations consonantiques des occlusives initiales en môn-khmer. *Bulletin de la Société de Linguistique de Paris* 60(1). 160–172.

[4] Matisoff, J. 1973. Tonogenesis in Southeast Asia. In: Hyman, L. (ed.), *Consonant Types and Tone*. Linguistic Program, University of Southern California, 73–95.

[5] Diffloth, G. 1982. Registres, devoisement, timbres vocaliques: Leur histoire en Katouique. *Mon-Khmer Studies* 11. 47–82.

[6] Svantesson, J., House, D. 2006. Tone production, tone perception and Kammu tonogenesis. *Phonology* 23(2). 309–333.

[7] L.Thongkum, T. 1990. The interaction between pitch and phonation type in Mon: phonetic implications for a theory of tonogenesis. *Mon-Khmer Studies* 16–17. 11–24.

[8] Thurgood, G. 2002. Vietnamese tonogenesis: Revising the model and the analysis. *Diachronica* 19. 333–363.

[9] Coetzee, A., Beddor, P., Shedden, W., Wissing, D. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *J. Phon.* 66. 185–216.

[10] Premsrirat, S. 2004. Register complex and tonogenesis in Khmu dialects. *Mon-Khmer Studies* 34. 1–17.

[11] Keating, P., Kuang, J., Garellek, M., Esposito, C. 2021. *A cross-language acoustic space for vocalic phonation distinctions*. https://linguistics.ucla.edu/people/keating/Keating-etal_ms_Nov2021.pdf

[12] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer. http://www.praat.org/

[13] Yao, Y., Tilsen, S., Sprouse R., Johnson, K. 2010. Automated Measurement of Vowel Formants in the Buckeye Corpus. *UC Berkeley PhonLab Annual Report* 6.

[14] Brooke, M. 1997. Voicebox: Speech Processing Toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[15] Signal Processing Toolbox for MATLAB. https://www.mathworks.com/products/signal.html.

[16] Hillenbrand, J., Cleveland, R., Erickson, R. 1994. Acoustic correlates of breathy vocal quality. *J. Speech and Hearing Research* 37. 769–778.

[17] Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* 73(1). 3–36.

[18] Kirby, J., Ladd, D. 2015. Stop voicing and F0 perturbations: Evidence from French and Italian. *Proc. 18th ICPhS.*

[19] Kirby, J. 2018. Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages. *J. Phon.* 71. 326–354.

[20] Schertz, J., Khan, S. 2020. Acoustic cues in production and perception of the four-way stop laryngeal contrast in Hindi and Urdu. *J. Phon.* 81. 100979.

[21] Esposito, A. 2002. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica* 59. 197–231.

[22] Kingston, J., Diehl, R., Kirk C., Castleman, W. 2008. On the internal perceptual structure of distinctive features: The [voice] contrast. *J. Phon.* 36(1). 28–54.

[23] Bell-Berti, F. 1975. Control of pharyngeal cavity for English and voiceless stop. *J. Acoust. Soc. Am.* 57(2).456–461.

[24] Ahn, S. 2018. The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese. *J. Phon.* 71. 451–467.

[25] Westbury, J. 1983. Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *J. Acoust. Soc. Am*. 73(4). 1322–1336.

[26] Westbury, J., Keating, P. 1986. On the naturalness of stop consonant voicing. *J. Ling.* 22(1). 145–166.

[27] Löfqvist, A., Baer, T., McGarr, N. 1989. The cricothyroid muscle in voicing control. *J. Acoust. Soc. Am*. 85(3). 1314–1321.

THE EFFECTS AND INTERACTION OF AGE AND REGION ON THE ACOUSTIC REALIZATION OF THE KOREAN STOP CONTRAST

Jeffrey J. Holliday[1], Eun Jong Kong[2], Hyunjung Lee[3]

[1]University of Kansas, [2]Korea Aerospace University, [3]Incheon National University
[1]jjh@ku.edu, [2]ekong@kau.ac.kr, [3]hjlee123@inu.ac.kr

ABSTRACT

Previous studies exploring phonological contrasts undergoing sound change have found that the relationship between production and perception can depend on factors such as region, age, and social setting. It remains unclear how this may unfold with multiple acoustic cues, and in stationary and mobile speakers. In this study, we measured VOT and F0 in the word-initial stops of 197 speakers of Seoul (standard) and Gyeongsang (non-standard regional variety) Korean in elementary school, high school, and university, along with Gyeongsang-to-Seoul university student transplants. In line with perception results from the same population, we found that the high school students' productions in both regions were the most regionally distinctive. While the Gyeongsang university students were no different from the high school students, the acoustic cues used by the transplants shifted toward a Seoul pattern. These results suggest that although production may be less flexible than perception, physical migration may modulate this effect.

Keywords: sound change, Korean stop production, acoustic cue weighting, mobile speakers

1. INTRODUCTION

Studies on phonological contrast undergoing a sound change have produced mixed findings regarding the relationship between perception and production. While many studies have found that perception and production are correlated to some degree, it has been reported that changes in perception precede changes in production [1, 2], that changes in production precede changes in perception [3, 4], and that perception leads at the beginning of the change but begins to lag behind production as the change nears completion [5, 6]. Sound changes spread not only over time, however, but also within a sociolinguistic context. This includes not only geographical space (e.g. from one region to another), but also speech communities and social networks within a given region (e.g. entering university after high school). In this context, we are interested in how sound changes spread into non-standard varieties. The present study therefore explored this question by investigating local and mobile speakers' use of acoustic cues to produce a contrast undergoing a sound change from below originating in the standard variety.

The sound change in the Korean three-way stop laryngeal contrast has been extensively studied in terms of both production and perception, both within Seoul and across other regions [7]. While Seoul Korean (SK) lenis and aspirated stops were historically cued by VOT (intermediate vs. long, respectively), over the past 60 years speakers have shifted to relying primarily on F0 (lenis < aspirated) in both production [8] and perception [9], effectively neutralizing VOT.

The potential for F0 to serve as the primary cue to the lenis-aspirated contrast has been questioned, however, for tonal varieties, such as Gyeongsang Korean (GK), spoken in the southeast of the country. Ostensibly because GK uses F0 to cue lexical pitch accent, studies have found that GK speakers still rely on VOT to cue the lenis-aspirated contrast [10]. But when the productions of elderly and teenage GK speakers have been directly compared, it was found that younger speakers did use F0 more [11], suggesting that the change observed in SK may be spreading to GK as well.

One recent study investigated the perception of this contrast by elementary school, high school, and university students in Seoul and Gyeongsang [12]. It was found that although the sound change is by now most likely complete in SK speakers this young, it is still in progress in GK. It was further found that the high school students in both regions exhibited perceptual cue weightings at the extreme ends of the change: SK high school students relied on VOT the least, and GK high school students relied on VOT the most. This result suggests that the high school students' perception was most tuned to the local variety, and the experience of attending university and interacting more with speakers from regions more or less advanced in the sound change could influence their perceptual cue weighting.

But given the debate over whether perception precedes production in an undergoing sound change, it is unclear whether SK and GK high school and university students would exhibit patterns of cue weighting in production that mirrored their perception. It was proposed that GK university students were more advanced than GK high school

students in the perception of the sound change, despite their older chronological age, due to their place in a speech community (i.e. the university) that includes more speakers who are more advanced in the change. The current study therefore investigates how the production of the Korean stop contrast varies in GK speakers across these same age groups (elementary school, high school, university), comparing their productions with both SK speakers, and another group of GK speakers who had moved to Seoul to attend university.

2. METHOD

2.1. Participants

The production data were collected from a total of 197 subjects from three dialect groups: 87 Gyeongsang, 90 Seoul, and 20 transplants who grew up in Gyeongsang but were attending university in Seoul. Gender in each group was split roughly equally between female and male. The number of speakers and mean age of each group is shown in Table 1. The GK and SK speakers were all born and educated in their respective dialect regions, mostly in the city of Changwon (where the South GK dialect is spoken) and the greater Seoul area, respectively. All participants, including the elementary school students, were literate in Korean. Subjects reported no language or hearing problems and were paid for their participation.

	SK	GK	Transplant
Elementary	27 (9.0)	23 (9.0)	
High School	25 (16.0)	29 (16.0)	
University	38 (24.2)	35 (23.8)	20 (19.6)

Table 1: No. of speakers (mean age) in each group.

2.2. Stimuli and recording

Each participant produced 27 disyllabic words, which targeted three initial stop phonation types (fortis, lenis, aspirated), three accent types (HL, HH, LH), and three places of articulation (bilabial, alveolar, velar). Most of the high school and university students produced three repetitions of a randomized word list presented in Korean orthography, but some of the high school students and elementary school students only recorded one or two repetitions due to fatigue. A total of 14,523 tokens were collected using a digital recorder and a Shure condenser microphone (SM81-LC) at a sampling rate of 44,100 Hz.

2.3. Acoustic measurements

The two acoustic measurements analyzed here are the VOT of the word-initial stop and the F0 of the following vowel. All measurements were done in Praat using a semi-automated script. The script stopped at each target word and let the user visually identify the stop burst and onset of voicing in the following vowel. The script then calculated the VOT as the difference between these two landmarks, and F0 as the mean over the first 15 ms after the voicing onset. For tokens with no measurable F0, which was mostly due to creakiness, the script advanced forward in 5 ms steps until F0 could be measured. If no F0 could be measured within 20 ms, the token was discarded. Lastly, F0 in Hz was converted to semitones.

2.4. Analysis

We used mixed effects logistic regression models to determine the extent to which VOT and F0 were used by speakers to signal the lenis-aspirated stop contrast. The present analysis focused only on the lenis and aspirated stops based on their observed generational and regional variation related to sound change. We also tested only the high tone-initial words (i.e. HH, HL), excluding low tone-initial words to control for the effect of lexical pitch accent in GK.

We first built separate models for female and male speakers with all groups except for the transplants. These *full models*, predicting aspirated (vs. lenis) laryngeal category, included age (elementary, high school, university) and region (Seoul, Gyeongsang) as fixed effects, with by-subject random intercepts and slopes for both VOT and F0. Age and region were treatment coded, with Seoul and high school serving as reference levels.

To analyze the effect of region more specifically, we built a second set of *university models* using only the university students, including the transplants, with transplants serving as the reference level. This model included only a three-level region factor (Seoul, Gyeongsang, transplant), and no age factor. Statistical models were run using the lme4 package [13] in R [14].

3. RESULTS

Relative use of VOT and F0: Both SK and GK speakers of all age groups produced aspirated stops with higher F0 and longer VOT than lenis stops. In terms of the relative use of these cues, F0 was more important than VOT, as the models yielded much larger coefficients for F0 than VOT. This result differs from perception patterns reported in other studies (e.g. [12]) in that F0 here served as a stronger

cue than VOT even in GK speakers' productions, and in all age groups.

	Female	Male
	Estimate (SE)	Estimate (SE)
Intercept	**-20.64 (1.93)**	**12.57 (2.17)**
VOT	0.82 (0.44)	**1.32 (0.34)**
F0	**25.40 (2.55)**	**12.56 (1.80)**
VOT×Age_{elem}	-0.18 (0.55)	*-2.21 (0.86)*
VOT×Age_{uni}	0.39 (0.61)	1.07 (0.60)
F0×Age_{elem}	-2.78 (2.29)	**19.00 (3.32)**
F0×Age_{uni}	6.27 (2.63)*	0.41 (1.53)
VOT×$Region_{GK}$	*1.80 (0.60)*	**2.41 (0.71)**
F0×$Region_{GK}$	*8.32 (2.71)*	0.13 (1.62)
VOT×Age_{elem}×$Region_{GK}$	-0.52 (0.79)	-1.60 (0.85)
VOT×Age_{uni} ×$Region_{GK}$	-0.74 (0.86)	-1.34 (1.03)
F0×Age_{elem}×$Region_{GK}$	*-10.59 (3.43)*	-3.67 (4.36)
F0×Age_{uni}×$Region_{GK}$	-9.28 (3.71)*	0.92 (2.26)

Table 2: Regression output from the *full models*. Subscripts indicate factor levels of elementary school (elem), university (uni), and Gyeongsang (GK). (**bold** $p < .001$, *italic* $p < .01$, '*' $p < .05$,)

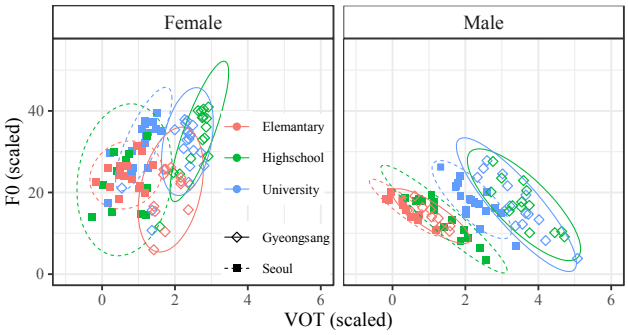


Figure 1: Female (left) and male (right) speakers' coefficient distributions for F0 and VOT from *full models*.

Figure 1 shows the distributions of individual speakers' F0 and VOT coefficients grouped by age and region, in each gender. In terms of gender, male speakers' F0 and VOT coefficients were negatively correlated (right panel; $r(92) = -.21$, $p = 0.059$), indicating a trading relationship between the two cues. In contrast, female speakers' F0 and VOT coefficients were positively correlated (left panel; $r(81) = -.46$, $p < 0.001$), suggesting that female speakers who used VOT more than others also used F0 more.

Differences between regions: As shown in Figure 1, regional differences were observed most clearly in high school speakers relative to elementary school or university speakers (see the solid and dashed green ellipses), in both genders. The regression output confirmed this trend: for males, GK high school students' VOT coefficient was significantly greater than that of their SK counterparts. For female speakers, it was not only VOT but also the F0 coefficients of GK high school students that were significantly greater than those of SK high school students. For the other age groups, regional differences were consistently found in female, but not male, speakers. GK female speakers' VOT coefficients were greater than those of SK female elementary school and university student students (solid vs. dashed blue and red ellipses). By contrast, male elementary school and university speakers did not exhibit regional differences in their VOT and F0 coefficients.

Differences among age groups: The age groups of SK and GK speakers in each gender were not consistently differentiated in terms of their VOT and F0 coefficients. While the university students were consistently different from the elementary school students (blue vs. red ellipses in Figure 1), they were not necessarily different from the high school student groups in using VOT or F0 (blue vs. green ellipses). In an analogous model run with university as the reference level age group, female university students in both regions had significantly greater F0 coefficients than the elementary student groups (SK: $\beta_{F0\times Elem} = -9.67$, SE = 2.63, $p < .001$; GK: $\beta_{F0\times Elem} = -11.01$, SE = 2.56, $p < .001$).

Similarly, male university students in both regions were differentiated from elementary school students in both acoustic dimensions, with significantly greater VOT and smaller F0 coefficients (SK: $\beta_{VOT\times Elem} = -3.47$, SE = 1.11, $p < .005$; $\beta_{F0\times Elem} = 18.96$, SE = 3.54, $p < .001$; GK: $\beta_{VOT\times Elem} = -3.70$, SE = 1.07, $p < .001$; $\beta_{F0\times Elem} = 14.5$, SE = 3.73, $p < .001$).

Between the GK university and high school students in both genders, however, neither VOT nor F0 were significantly different. In SK, however, female university students' F0 coefficient was significantly greater than those of high school students, while SK male university students' VOT coefficient was marginally greater than those of the high school students.

Differences among regions in university students: The *university model* output is given in Table 3, with the individual coefficients plotted in Figure 2. Female transplant speakers were not different from SK or GK speakers in VOT and F0. Male transplant speakers were distinguished from SK (but not GK) speakers, however, by having a significantly greater VOT coefficient.

	Female	Male
	Estimate (SE)	Estimate (SE)
Intercept	**-19.04 (2.35)**	**18.47 (1.73)**
VOT	1.78 (0.77)*	**4.40 (0.73)**
F0	**34.68 (4.57)**	**18.18 (1.93)**
VOT×$Region_{Seoul}$	-0.64 (0.92)	-2.14 (0.85)*

VOT×Region$_{GK}$	0.92 (0.89)	-0.92 (0.88)
F0×Region$_{Seoul}$	-2.74 (3.70)	-1.66 (1.57)
F0×Region$_{GK}$	-4.54 (3.60)	-2.26 (1.58)

Table 3: Regression output from *university models*. Subscripts indicate Seoul and Gyeongsang (GK) factor levels. (**bold** $p < .001$, *italic* $p < .01$, '*' $p < .05$)

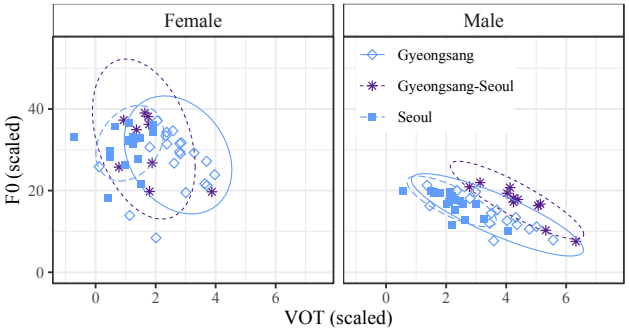


Figure 2: Female (left) and male (right) speakers' coefficient distributions of F0 and VOT from *university models*.

4. DISCUSSION

The current study examined the acoustic patterns of VOT and F0 in the Korean stops produced by Seoul and Gyeongsang speakers of three age groups: children, adolescents and young adults. We aimed to investigate whether the sub-phonemic phonetic change in the lenis-aspirated contrast would differ by age and region, and whether these results were similar to the perception patterns previously reported by [12]. We found that both the GK and SK speakers produced the newer stop variants by enhancing F0. Among the age groups, GK high school students were the most regionally distinctive by their use of VOT, but they were not the most progressive group. Notably, while the GK university students were no different from the high school students, the acoustic cues used by the transplants shifted toward a Seoul pattern.

We may draw two broader conclusions from these findings. First, the fact that the acoustic cues used in production do not necessarily mirror the acoustic cue weighting used in perception supports the idea that perception may be more flexible. [12] showed that the perception of university students seemed to reflect exposure to a wider speech community than the local variety, whether that be the standard variety SK, or a non-standard variety GK. The fact that GK university students' production did not reflect any change due to such exposure could be related to their not actually having left the GK region. This possibility is also suggested by the finding that GK transplant speakers *did* increase their use of F0 compared to local GK university students.

Second, these results lead us to consider adolescent peaks [16, 17] in sound change from a non-standard regional speaker's perspective. In the case of standard Seoul Korean, the sound change seems to be basically complete, so we may not expect an "adolescent peak" in which high school students produce the most innovative forms among age groups. But why don't we see an adolescent peak in GK, where the sound change is still ongoing? In the case of non-standard GK, high school students were not differentiated from university or elementary school students in terms of more use of F0 and/or less use of VOT. It could have to do with the fact that the sound change in question is one from another region, and the current speakers have not physically left the local area. This may be compared with the male GK transplants, who did rely on F0 more than the local GK university students. In the case of SK, the high school students did rely on VOT less than the university students, which could be interpreted as a sort of adolescent peak, although they were not differentiated from elementary school students. Thus, we find some limited evidence for adolescent peaks in the current data from multiple regions.

So, what does this say about the precedence of perception or production in an ongoing sound change? Focusing on GK, in which the sound change is still ongoing, it appears that speakers from all age groups use a more conservative acoustic cue weighting in their perception than their production. The scenario suggested by [6] (and supported by [2]) is that changes in perception will lead changes in production at the beginning of a sound change, but then as it proceeds and nears completion production will become to lead. This can be understood as speakers having advanced through the sound change but are maintaining flexible perceptual cue weightings to accommodate speakers who are more conservative. This appears to be the pattern found in the current study.

To conclude, we first observed an effect of region in the progression of the sound change in Korean stops: the sound change is mostly complete in SK, and still ongoing in GK. In terms of age, the effects are most clear when the current results are compared with findings from much older GK speakers, as in [11]. The age groups in the current study are close enough that the differences are more reflective of their immediate social setting (e.g. high school vs. university setting) than chronological age. We also saw how age and place interact in our comparison between local male GK speakers and male GK transplants living in Seoul, which supports the findings of [18]. These results suggest that although production may be less flexible than perception, physical migration may modulate this effect.

5. REFERENCES

[1] Harrington, J., Kleber, F., Reubold, U. 2008. Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *J. Acoust. Soc. Am.* 123(5), 2825-2835.

[2] Kuang, J., Cui, A. 2018. Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics* 71, 194-214.

[3] Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., Wissing, D. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics* 66, 185-216.

[4] Strickler, A. 2019. Within-speaker perception and production of dialectal /aɪ/-raising. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 3205-3209).

[5] Howe, P. 2017. *Tonogenesis in central dialects of Malagasy: Acoustic and perceptual evidence with implications for synchronic mechanisms of sound change* [Doctoral dissertation, Rice University].

[6] Pinget, A.-F., Kager, R., Van de Velde, H. 2020. Linking variation in perception and production in sound change: Evidence from Dutch obstruent devoicing. *Language and Speech* 63(3), 660-685.

[7] Lee, H., Holliday, J. J., Kong, E. J. 2020. Diachronic change and synchronic variation in the Korean stop laryngeal contrast. *Language and Linguistics Compass* 14(7), e12734.

[8] Kang, Y. 2014. Voice onset time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics* 45, 76-90.

[9] Lee, H., Politzer-Ahles, S., Jongman, A. 2013. Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *Journal of Phonetics* 41, 117-132.

[10] Holliday, J. J., Kong, E. J. 2011. Dialectal variation in the acoustic correlates of Korean stops. In W. S. Lee & E. Zee (eds.), *Proceedings of the 17th International Congress on Phonetic Sciences* (pp. 878–881).

[11] Lee, H., Jongman, A. 2012. Effects of tone on the three-way laryngeal distinction in Korean: An acoustic and aerodynamic comparison of the Seoul and South Kyungsang dialects. *Journal of the International Phonetic Association 42*(2), 145-169.

[12] Kong, E. J., Holliday, J. J., Lee, H. 2022. Post-adolescent changes in the perception of regional sub-phonemic variation. *Journal of Phonetics* 90, 101114.

[13] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67*(1), 1–48.

[14] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[15] Lee, H. 2020. Cross-generational acoustic comparisons of tonal Kyungsang Korean stops. *J. Acoust. Soc. Am. 148*(2) EL172-EL178.

[16] Tagliamonte, S., D'Arcy, A. 2009. Peaks beyond phonology: Adolescence, incrementation and language change. *Language 85*, 58-108

[17] Holmes-Elliott, S. 2016. Ladies First? Adolescent Peaks in a Male-Led Change. *University of Pennsylvania Working Papers in Linguistics 22*(2) 81-90.

[18] Kim, H., Jongman, A. 2022. The influence of inter-dialect contact on the Korean three-way laryngeal distinction: An acoustic comparison among Seoul Korean speakers and Gyeongsang speakers with limited and extended residence in Seoul. *Language and Speech 65*(3), 531-553.

INVESTIGATING SOUND CHANGE THROUGH COMPUTATIONAL AGENT-BASED MODELLING: AN R PACKAGE

Johanna Cronenberg, Michele Gubian, Jonathan Harrington, Vanessa Reichel

Institute for Phonetics and Speech Processing (IPS), LMU Munich
johanna.cronenberg | m.gubian | jmh | vanessa.reichel@phonetik.uni-muenchen.de

ABSTRACT

We introduce the R package `soundChangeR`: an implementation of a computational, agent-based model of sound change. This model rests upon ideas from exemplar theory about how the memorisation of traces of speech can result in phonetic imitation, perceptual learning, and eventually phonetic and phonological change. In theoretical views of the exemplar-based relation between speech production and perception, it has so far remained unclear whether some stored exemplars may contribute more to a listener-turned-speaker's future production than others. Using the cognitively-inspired architecture of `soundChangeR` we show that it makes a difference whether the agents preferably memorise typical and/or unambiguous exemplars. Memorising only exemplars that are typical members of their phonological class can reinforce phonetic biases, whereas memorising only unambiguous exemplars creates a repulsion between phonological classes. We argue that using such computational models can lead to new theoretical insights and we encourage readers to try `soundChangeR` on their own data.

Keywords: sound change, agent-based modelling, exemplar theory, selective memorisation

1. INTRODUCTION

Decades of research have established that sound changes are rooted in phonetic variation [1] and propagated through interactions between individuals [2, 3]. However, the exact interplay between the intra- and extralinguistic factors that may contribute to sound change at any stage of the process remains poorly understood. This is because it is impossible to know beforehand whether a sound change is going to take place, so as soon as the change is underway, it is too late to capture the circumstances which may have triggered it. Computational agent-based models (ABMs) offer an artificial world and a controlled environment which can be used to explore the cognitive, social, and linguistic factors that might play a role in the emergence and spread of a specific change that has already been observed empirically e.g. by means of apparent-time data.

The ABM presented in this paper, called `soundChangeR`, aims to link theories about the origin and spread of sound change by means of a cognitively-inspired architecture. More specifically, `soundChangeR` rests upon ideas from exemplar and usage-based models of language [4, 5, 6] about how the memorisation of traces of speech (exemplars) can result in phonetic imitation [7], perceptual learning [8], and, eventually, sound change. Therefore, and in line with other computational models of sound change [9, 10, 11, 12], agents in `soundChangeR` are the computational analogue of human speaker-listeners, i.e. they are equipped with mechanisms in order to produce and perceive exemplars of speech sounds as well as a memory that connects the two processes. In contrast to other ABMs of sound change, however, `soundChangeR` is publicly available and can be used on real speech production data to test whether the model predicts any acoustic or phonological changes. Section 2 explains the details of the agents' production-perception feedback loop and briefly mentions how the R package `soundChangeR` can be installed and used.

Exemplar-based approaches have been particularly useful in explaining phonetic shifts, such as /u/-fronting in many English varieties [13], the Northern Cities Shift [10], or the New Zealand English short front vowel shift [14]. These kinds of sound changes are hypothesised to result from updates of memorised acoustic distributions caused by consistent exposure to biased variants. However, exemplar models often remain quite vague about the relation between speech perception and speech production despite its significance in sound changes. That is, are there memorised exemplars that affect future productions more than others? For instance, do typicality and/or ambiguity in the association between exemplars and phonological classes increase the likelihood of updating learned acoustic distributions and, consequently, of shifting future productions? The

exemplar-based architecture of `soundChangeR` allows us to computationally test these assumptions about the probabilistic selection of exemplars and its relation to phonetic shifts. Three simulations on artificial data will be presented to show that it makes a difference whether agents preferably memorise typical and/or unambiguous exemplars.

2. METHOD

2.1. Model Architecture

In `soundChangeR`, each agent is typically initialised with the speech production data from a real human speaker. This data should consist of relevant acoustic parameters for the sound(s) under investigation as well as the word type in which the sound was produced (e.g. formant values for the vowel in *food*). Together, the acoustic values and the lexical entry form an exemplar which is stored in the agent's memory. Before the first interaction, and then regularly throughout the simulation, each agent derives sub-phonemic classes from the stored exemplars by means of two unsupervised machine learning algorithms. The ABM follows exemplar theory's bottom-up approach to phonological knowledge: that is, sub-phonemes (SPs) in `soundChangeR` are abstractions over clouds of stored exemplars that can be re-computed when new exemplars have been memorised [15]. The resulting classes are agent-specific and can encode both classical phonological [16] as well as sub- or quasi-phonemic knowledge [17, 18].

In an interaction, a pair of agents is chosen either randomly or based on their agent group. These groups can be defined by the user and contain any sociolinguistic information about the speakers that may play a role in the sound change (e.g. age, gender, or regional origin). The agent-speaker randomly chooses a word type from its lexicon, estimates a (Gaussian) distribution based on all memorised exemplars associated with that word type, and samples a new exemplar from that distribution. This word-based sampling procedure ensures that possible coarticulatory effects (e.g. that the vowel in *soup* is more fronted than that in *food*) are also present in the newly produced exemplar which, together with the corresponding word type, is transmitted to the agent-listener.

The task in speech perception is to decide whether or not to memorise the exemplar (but not recognising or categorising that exemplar). That is, we make explicit assumptions about which exemplars are more likely to impact the agent-listener's internal acoustic distributions and thus, their speech production in future interactions. The decision to memorise a perceived exemplar can be based on two criteria [9] pertaining to the sub-phoneme, SP, that is associated with the word type transmitted by the agent-speaker. The absolute criterion determines whether the Mahalanobis distance between the exemplar and the centroid of SP is below a threshold defined by the user. The reasoning behind this criterion is that atypical exemplars should not influence the listener-turned-speaker's future productions [19, 20].

The relative criterion determines whether the exemplar's posterior probability conditioned on SP is higher than the posterior probabilities conditioned on all other sub-phonemes in the agent-listener's memory. That is, only exemplars that are discriminable [9] (i.e. cannot be confused for a member of another sub-phoneme) are memorised. Since this criterion penalises acoustic ambiguity, it creates a repulsion between sub-phonemic classes, following the idea that phonological contrasts are usually maintained [21, 22]. However, given that sub-phonemes are regularly updated during a simulation, this repulsion does not preclude sub-phonemic classes from merging [23].

The two memorisation criteria can also be combined, i.e. the perceived exemplar is only memorised if it passes both criteria. Finally, if the agent-listener has memorised the perceived exemplar, it forgets (i.e. deletes) a random exemplar of the same word type. The process of forgetting exemplars ensures that the memory size does not grow to such an extent that newly memorised exemplars have no impact.

Simulations are run until there is no more observable acoustic change across the population. An analysis can then be undertaken of whether there is any evidence of change either in the acoustic makeup of the simulated sounds (in the case of a phonetic shift like e.g. /u/-fronting) or in how words are mapped to sub-phonemic classes (in the case of e.g. a split or a merger).

2.2. Usage

The R package `soundChangeR` can be downloaded from GitHub (see installation instructions: https://github.com/IPS-LMU/soundChangeR) and comes with full documentation in the form of help pages for all functions and a detailed vignette. The core function is `run_simulation()` which takes a number of arguments by means of which the user can adapt the simulation parameters. Importantly, users can submit their own speech production data

in any simple table format (e.g. `.txt`, `.csv`). The dataset minimally needs to contain speaker IDs (each speaker is going to be represented by one agent), numeric-continuous acoustic features of the sounds under investigation, and the word types in which the sounds were uttered. An example of a dataset that meets these requirements is `u_fronting` which is available to the user once `soundChangeR` is loaded in R. It contains DCT-parametrised F2 values for the vowels /i, u, ju/ produced by 22 Standard Southern British English speakers [13].

2.3. Testing the Effects of Exemplar Selection

Three simulations were run with `soundChangeR` in order to demonstrate how the application of the memorisation criteria (absolute only, relative only, or both) can result both in phonetic shifts and stability [12]. Figure 1 shows the artificial dataset that was generated for this purpose. It consists of two agents, A and B, both of which are initialised with 50 exemplars each of 10 word types in a two-dimensional acoustic space. For agent A, sub-phoneme 1 (SP1, solid ellipse) is elongated and overlaps with sub-phoneme 2 (SP2, dashed ellipse) at one end along its main axis. Agent A's SP2, in turn, is similar to agent B's SP2. For agent B, both sub-phonemic classes are are compact and nearly spherical.

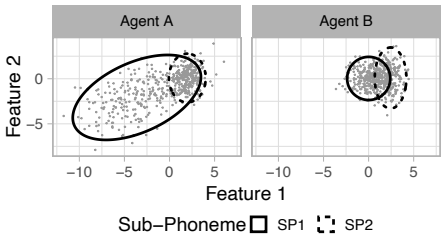


Figure 1: State of agent A's and B's representations of sub-phonemic classes SP1 (solid) and SP2 (dashed) in the acoustic space before the simulation. The dots represent stored exemplars.

The data was designed to mirror a state of a sound change at which one group of speakers (represented by agent B) has already adopted a new variant of a sound (SP1) while another group (represented by agent A) is affected by a phonetic bias that skews that sound in the direction of the change. These configurations can often be observed both in apparent-time data and across regional varieties of which one has undergone a sound change faster than another.

The two agents shown in Figure 1 interacted with each other following the architecture described in 2.1, while applying either the absolute, the relative, or both memorisation criteria in perception, until there were no more relevant acoustic changes.

3. RESULTS

Figure 2 shows the outcomes of the three simulations. When only the absolute memorisation criterion was applied (Figure 2a), agent A clearly adapted SP1 to that of agent B. This is because the exemplars of SP1 produced by agent B were very likely to fall within agent A's solid ellipse, i.e. they were typical enough to become members of A's SP1. Exemplars of SP1 produced by agent A, on the other hand, were likely to have low values of both Feature 1 and 2 (see agent A's solid ellipse in Figure 1), and thus would have been too atypical according to the absolute criterion to be incorporated into agent B's narrow SP1. Neither agents' SP2 shifted acoustically compared to the baseline depicted in Figure 1, but narrowed over the course of the simulation.

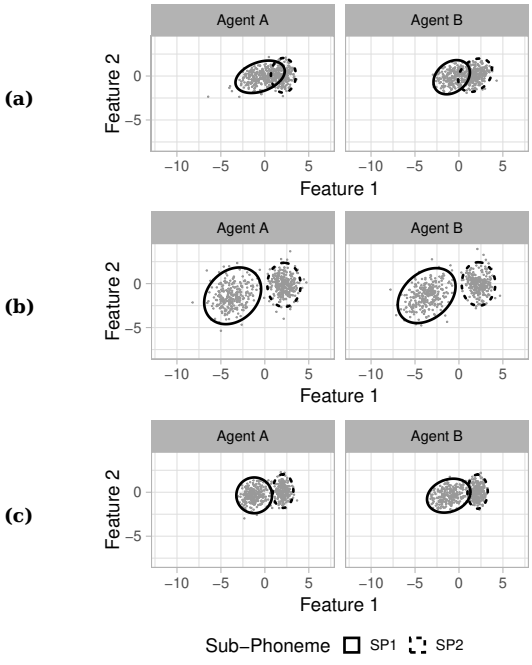


Figure 2: State of agent A's and B's representations of sub-phonemic classes in the acoustic space after the simulation. The dots represent stored exemplars. **(a)** Only the absolute criterion, **(b)** only the relative criterion, and **(c)** both memorisation criteria were applied to decide whether perceived tokens should be memorised.

When only the relative memorisation criterion

was applied (Figure 2b), both agents ended up with two non-overlapping sub-phonemic classes. In this case, the centroid of agent B's SP1 moved towards that of agent A's SP1. This is because agent B accepted agent A's exemplars of SP1 which were likely to fall into an unambiguous part of the acoustic space, i.e. a part where the acoustic distribution of SP2 is not in competition with that of SP1. Agent A, on the other hand, rejected those exemplars of SP1 produced by agent B that were located in the part of the acoustic space where agent A's SP1 and SP2 overlapped. As in the first simulation, neither agents' SP2 changed.

Finally, when both memorisation criteria had to be passed in order for an exemplar to be memorised (Figure 2c), agent A's SP1 shifted more towards agent B's SP1 than vice versa, and both the overlap between the sub-phonemes as well as their variance were reduced. Thus, this simulation displays characteristics of the previous two simulations.

4. DISCUSSION

The three simulations presented in section 3 had visibly diverse outcomes despite starting from the same input data. The only difference between the simulations was the criterion applied for the selective memorisation of perceived exemplars. Both the absolute and relative criterion can cause phonetic shifts and stability (i.e. importantly, `soundChangeR` does not inevitably predict change), but both also have side effects.

The absolute criterion can reinforce phonetic biases that are present in the input data. In the presented case, agent A's phonetic bias was reinforced through interactions with agent B whose variants of SP1 lay in the direction of the bias. This mechanism of `soundChangeR` is based on the idea that speakers that have already adopted a new variant of a sound may serve as attractors to speakers whose variant is skewed in that direction e.g. because of a phonetic bias [24]. The side effect of the absolute criterion is a reduction of within-category variance as acoustic outliers are rejected (although the severity of this side effect depends on the Mahalanobis distance threshold set by the user). Other ABMs of sound change instead opt for explicit production biases, i.e. a newly produced exemplar is shifted slightly in the direction of an articulatory bias before being transmitted to the listener [9, 11, 12]. While this reduces the risk of a return-to-the-mean effect, the constant application of the bias may shift the acoustic distributions further than it would in a real sound change.

The relative criterion results in phonetic shifts because of the repulsion created by the constant pressure against ambiguous exemplars. Since SP1 and SP2 partially overlap for both agents in Figure 1, new exemplars that are located in this area of the space are indiscriminable for the agents and therefore not memorised. When applied on its own, the relative criterion does not prevent even the furthest outliers from being memorised as long as they are unambiguous, which can lead to an expansion of the acoustic-phonetic space beyond plausible values. However, the third simulation showed that this side effect of the relative criterion can be counteracted by the absolute criterion while still maintaining the contrast between sub-phonemic classes as desired.

Phonetic stability was achieved for SP2 in all three simulations. This was because the agents had a similar representation of this sub-phoneme and thus, there was no force that could cause acoustic changes as a result of the interactions. This underlines the necessity for phonetically heterogeneous input data: none of the algorithms in `soundChangeR` can create the kinds of phonetic biases in speech production which are hypothesised to lead to change [25]. The forces that create changes in a simulation and result from the agents' production-perception feedback loop only take effect if the agents have sufficiently distinct representations of the sounds under investigation to begin with.

ABMs are a powerful tool for the investigation of so called complex adaptive systems, i.e. systems in which individual entities interact with one another according to some rules and can change as a response to their environment [6, 26]. Spoken language is well suited to be modelled as a complex adaptive system in which microscopic variation at the level of individual interactions leads to community level categorical change, as also shown by the recent upsurge in ABMs designed to answer specific questions in the field of sound change [9, 12]. The simulations presented in this paper showed that the way we handle the probabilistic selection of exemplars is critical in determining the direction of a phonetic shift. In general, ABMs like `soundChangeR` force us to be explicit about the assumptions that underlie the theoretical model to be tested. It is for this reason that we hope that `soundChangeR` will be applied to a variety of speech production data in the future: to gain a better understanding of the interplay of factors that impact the emergence and spread of many kinds of sound changes.

5. ACKNOWLEDGEMENTS

This research was supported by European Research Council Grant No. 742289 "Human interaction and the evolution of spoken accent" (2017-2023).

6. REFERENCES

[1] J. J. Ohala, "Sound change is drawn from a pool of synchronic variation," in *Language Change: Contributions to the Study of Its Causes*, ser. Trends in Linguistics, Studies and Monographs, L. E. Breivik and E. H. Jahr, Eds. Berlin: Mouton de Gruyter, 1989, no. 43, pp. 173–198.

[2] P. Trudgill, "Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation," *Language in Society*, vol. 37, pp. 241–280, 2008.

[3] W. Labov, *Principles of Linguistic Change. Volume 2: Social Factors*. Malden: Blackwell, 2001.

[4] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency Effects and the Emergence of Linguistic Structure*, J. Bybee and P. J. Hopper, Eds. Amsterdam: John Benjamins, 2001, pp. 137–155.

[5] J. Bybee, "Usage-based Phonology," in *Functionalism and Formalism in Linguistics: Volume I: General Papers*, M. Darnell, E. A. Moravcsik, M. Noonan, F. J. Newmeyer, and K. Wheatley, Eds. Amsterdam: John Benjamins, 1999, pp. 211–242.

[6] C. Beckner, R. Blythe, J. Bybee, M. H. Christiansen, W. Croft, N. C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, and T. Schoenemann, "Language Is a Complex Adaptive System: Position Paper," *Language Learning*, vol. 59, pp. 1–26, 2009.

[7] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[8] A. G. Samuel and T. Kraljic, "Perceptual learning for speech," *Attention, Perception, & Psychophysics*, vol. 71, no. 6, pp. 1207–1218, 2009.

[9] S. Todd, J. B. Pierrehumbert, and J. B. Hay, "Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model," *Cognition*, vol. 185, pp. 1–20, 2019.

[10] J. N. Stanford and L. A. Kenny, "Revisiting transmission and diffusion: An agent-based model of vowel chain shifts across large communities," *Language Variation and Change*, vol. 25, no. 2, pp. 119–153, 2013.

[11] J. Kirby, "Incipient tonogenesis in Phnom Penh Khmer: Computational studies," *Laboratory Phonology*, vol. 5, no. 1, pp. 195–230, 2014.

[12] M. Sóskuthy, "Understanding change through stability: A computational study of sound change actuation," *Lingua*, vol. 163, pp. 40–60, 2015.

[13] J. Harrington, F. Kleber, and U. Reubold, "Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2825–2835, 2008.

[14] J. B. Hay, J. B. Pierrehumbert, A. J. Walker, and P. LaShell, "Tracking word frequency effects through 130 years of sound change," *Cognition*, vol. 139, pp. 83–91, 2015.

[15] M. Gubian, J. Cronenberg, and J. Harrington, "Phonetic and Phonological Sound Changes in an Agent-Based Model," *Speech Communication*, vol. 147, pp. 93–115, 2023.

[16] P. Kiparsky, "Formal and empirical issues in phonological typology," in *Phonological Typology*, L. M. Hyman and F. Plank, Eds. Berlin: De Gruyter Mouton, 2018, pp. 54–106.

[17] J. M. Scobbie and J. Stuart-Smith, "Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English," in *Contrast in Phonology: Theory, Perception, Acquisition*, P. Avery, B. E. Dresher, and K. Rice, Eds. Berlin: Mouton de Gruyter, 2008, pp. 87–113.

[18] E. Reinisch, K. I. Juhl, and M. Llompart, "The Impact of Free Allophonic Variation on the Perception of Second Language Phonological Categories," *Frontiers in Communication*, vol. 5, p. 47, 2020.

[19] M. Sumner, S. K. Kim, E. King, and K. B. McGowan, "The socially weighted encoding of spoken words: A dual-route approach to speech perception," *Frontiers in Psychology*, vol. 4, p. 1015, 2014.

[20] Y. V. Melguy and K. Johnson, "Perceptual adaptation to a novel accent: Phonetic category expansion or category shift?" *The Journal of the Acoustical Society of America*, vol. 152, no. 4, pp. 2090–2104, 2022.

[21] B. de Boer, "Self-organization in vowel systems," *Journal of Phonetics*, vol. 28, no. 4, pp. 441–465, 2000.

[22] J. Blevins and A. Wedel, "Inhibited sound change: An evolutionary approach to lexical competition," *Diachronica*, vol. 26, no. 2, pp. 143–183, 2009.

[23] M. Stevens, J. Harrington, and F. Schiel, "Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English," *Glossa*, vol. 4, no. 1, pp. 1–30, 2019.

[24] J. Harrington, F. Kleber, U. Reubold, F. Schiel, and M. Stevens, "Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling," *Topics in Cognitive Science*, pp. 1–22, 2018.

[25] A. Garrett and K. Johnson, "Phonetic bias in sound change," in *Origins of Sound Change*, A. C. L. Yu, Ed. Oxford: Oxford University Press, 2013, pp. 51–97.

[26] T. C. Schelling, *Micromotives and Macrobehavior*. New York: Norton, 1978.

CROSS-DIALECTAL COMPARISON OF /st/ CLUSTERS IN SPANISH: IMPLICATIONS FOR SOUND CHANGE

Amber Galvano[a] & Nicholas Henriksen[b]

[a]University of California, Berkeley & [b]University of Michigan
amber_galvano@berkeley.edu, nhenriks@umich.edu

ABSTRACT

This study examines the variation present in phonological /st/ sequences for speakers of three regional varieties of Spanish: North-Central Peninsular Spanish (NCPS), Buenos Aires Spanish (BAS), and Western Andalusian Spanish (WAS). The three varieties are intended to reflect related stages of sound change in phonological /st/ clusters whereby speakers may shift from /s/-retention (NCPS) to aspiration of /s/ (BAS) to post-aspiration (WAS). We acoustically inspected the spontaneous speech of 60 speakers and coded each /st/ production for one of eight possible phonetic variants. While the results confirm expected patterns of variation based on a speaker's regional origin, they also show that long VOT is a pervasive characteristic in the realization of /st/ clusters in all varieties. We contextualize the findings within proposals that consider the phonetic pathway that gives rise to both pre- and post-aspiration of phonological /st/ in Spanish.

Keywords: sound change, actuation, propagation, Spanish, post-aspiration

1. INTRODUCTION

Spanish dialects are often classified as either 'retention' or 'lenition' varieties regarding the treatment of syllable-final /s/: while retention varieties retain the alveolar fricative, lenition varieties weaken the /s/, typically through aspiration or deletion [1]. Regarding Andalusian Spanish (spoken in southern Spain), for most speakers of this variety, pre-pausal /s/ is deleted, word-final intervocalic /s/ is aspirated or optionally deleted, and preconsonantal /s/ is aspirated [2]. Notably, in Andalusia's three westernmost provinces, pre-aspiration in /s/ + voiceless stop sequences has become highly variable such that speakers are shifting to post-aspiration in which glottal frication follows the stop closure [3, 4, 5, 6]. In other words, lexemes that were historically produced with coda retention (e.g., pasta [pasta] 'pasta') or pre-aspiration ([pahta]) now surface as post-aspirated stops, sometimes transcribed as [C^h], as in [pat^{h}a] [5, 7].

Articulatory Phonology [8] explains the shift from pre- to post-aspiration through a single realignment of the glottal-opening and oral-closure gestures [4, 9]. One question that remains unanswered, however, concerns how actuation would have unfolded among language users of WAS; in other words, how would variation present in the ambient language have instigated sound change among WAS pre-aspirating language users? Apparent-time and comparative studies converge on the idea that the sound change likely emerged in the coronal context (i.e., /st/-words) prior to undergoing transmission to the labial and velar contexts. Ruch and Peters [5] propose a three-step pathway from pre- to post-aspiration: (i) the high intraoral air pressure from the lengthened stop closure produced in pre-aspirated [hp ht hk] would result in unusually long VOT produced as [hp^h ht^h hk^h] [7]; (ii) the long VOT of such productions, especially [ht^h], would become more perceptually salient for listeners [10, 11, 12]; and (iii) actuation would materialize when WAS speakers imitate (and subsequently lengthen) the perceptually salient VOT of /st/-words.

With this background, the present study is guided by three goals. The first goal is to examine speaker data from a more informal speaking style than is commonly used in research on WAS post-aspiration. While most previous studies of the sound change elicit data in a controlled setting [3, 4, 7, 9, 13], one study that analyzes WAS spontaneous data reports a wider range of acoustic variants produced in /st/-words than straightforward aspiration after the closure [14]. Access to a wider range of variants, likely more prevalent in informal speaking contexts [15, 16], will aid in understanding the envelope of variation present in /st/-words. This range of data can provide new insight into the phonetic pathway along which speakers progress throughout the change from [st] → [ht] → [t^h].

The second goal is to investigate the linguistic and social factors that condition the attested /st/ variation in Spanish, in order to elucidate how propagation of the sound change unfolds. Previous work

suggests that factors such as the surrounding vocalic context and word position have a significant effect on WAS post-aspiration realization, while the effect of speaker sex is less certain [7, 17].

A third goal is to test the validity of Ruch and Peters' [5] proposal by comparing data produced by speakers of three regional varieties of Spanish (NCPS, BAS, and WAS), which likely reflect three stages of the sound change itself. The NCPS data are intended to reflect the ambient-language variation that is present prior to the emergence of pre-aspiration, and the BAS data are intended to reflect the variation present among pre-aspirating speakers prior to the emergence of post-aspiration (see [18]).

2. METHODS

2.1. Participants and data collection

We collected speech data from 60 speakers (age range 18-25): 19 NCPS speakers from Salamanca, Spain (9 female & 10 male); 20 BAS speakers from Buenos Aires, Argentina (10 female & 10 male); and 21 WAS speakers from Jerez, Spain (8 female & 13 male). All speakers participated in twenty-minute sociolinguistic interviews conducted in their home cities by a native speaker of their regional variety. We collected the data using a Marantz Pmd620 digital recorder and a Shure WH20 head-mounted microphone.

2.2. Analysis

The inclusion criterion was that a phonemic /st/-token should be produced between vowels, either word-medially or across word boundaries, which resulted in a total of 6,334 tokens. Using the acoustic characterization from [14] as a baseline, we coded each instance of /st/ per the following classification: (i) /s/-retention [st]; (ii) /s/-retention with long VOT [st^{h}]; (iii) /s/-deletion [t] or pre-aspiration [ht]; (iv) pre-aspiration with long VOT [ht^{h}]; (v) post-aspiration [t^{h}]; (vi) dento-alveolar affricate [t^{s}]; (vii) alveolar fricative [s]; and (viii) post-alveolar affricate [tʃ]. Based on the proposal by [5] that lengthened VOT should be present among speakers for which the post-aspiration process is not instigated, we used 29 ms as the threshold for establishing tokens with "long" VOT (per [19]), which we transcribed as [t^{h}] throughout the study. This decision is further supported by VOT findings of singleton /t/ in Spanish showing that speakers produce mean values between 15 and 20 ms [3]. Figures 1 through 5 show examples of pre-aspiration [ht], post-aspiration [t^{h}], pre- and post-aspiration [ht^{h}], the dento-alveolar affricate [t^{s}], and the alveolar fricative [s], respectively. To ensure reliability in the coding, both authors initially labeled 10% of the data. For the instances in which they did not agree, the two authors established updated labeling criteria. Following this, the first author proceeded with the labeling of the remaining files and consulted with the second author in cases of ambiguity.

For statistical analysis, we divided the data by DIALECT and, due to low token counts for some of the variants (see Table 1), we limited the comparisons to five. For BAS, we compared the variants [st] vs. [t, ht]; [st] vs. [ht^{h}]; and [t, ht] vs. [ht^{h}]. For NCPS we compared [st] vs. [st^{h}], and for WAS we compared [t^{h}] vs. [t^{s}] (see Table 2). For each of the five comparisons, we fitted a mixed logistic regression model using the package "lme4" [20] within the statistical software R, version 4.2.2 [21]. We included SEX (male, female), WORD POSITION (word-medial, cross-word), PRECEDING VOWEL, and FOLLOWING VOWEL (/a e i o u/) as fixed main effects. All models included random-intercept effects for SPEAKER to account for within-speaker dependencies. We also included by-SPEAKER random slopes for WORD POSITION, PRECEDING VOWEL, and FOLLOWING VOWEL [22]. We computed likelihood-ratio test (LRT) statistics for all fixed effects. Here, we report the significant fixed effects based on an $\alpha = 0.05$ significance level.

3. RESULTS

Table 1 presents the frequency of use for each variant based on dialect. The most common variants per dialect were as follows: /s/-retention for NCPS (72.32%), pre-aspiration (or no aspiration) for BAS (62.06%), and post-aspiration for WAS (68.56%). Regarding the second-most common variant, for WAS it was the dento-alveolar affricate (12.68%), while for both NCPS and BAS approximately 25% of the data contained long VOT above 29 ms (i.e., [st^{h}] and [ht^{h}] were second-most common, respectively). Note also that productions of [s] and [tʃ] are nearly exclusive to WAS speakers.

Table 2 presents the results from the five logistic regressions (one for NCPS, three for BAS, and one for WAS). In the NCPS model, the single significant predictor was FOLLOWING VOWEL, with /i u/ favoring [st^{h}] the most (predicted probabilities of 32% and 42%) and /a/ the least (predicted probability of 8%). For the BAS comparison of [st] vs. [t, ht], the lenited variants were favored word-medially as opposed to across word boundaries, although the estimated probabilities were very similar (99% to 98%,

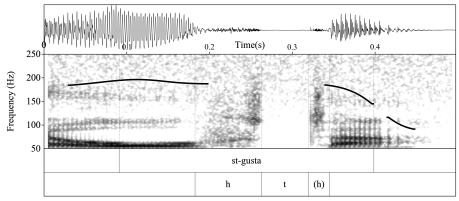


Figure 1: Sample spectrogram for pre-aspiration [ht] from a BAS speaker. VOT = 25 ms

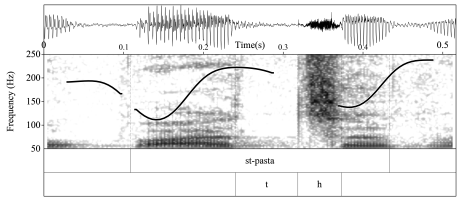


Figure 2: Sample spectrogram for post-aspiration [t^h] from a WAS speaker. VOT = 54 ms

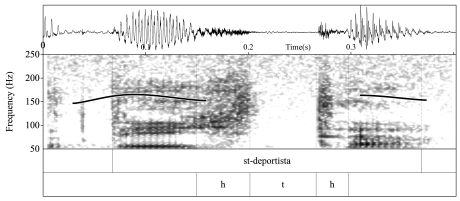


Figure 3: Sample spectrogram for pre- and post-aspiration [ht^h] from a BAS speaker. VOT = 31 ms

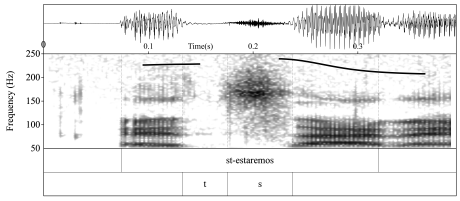


Figure 4: Sample spectrogram for [t^s] from a WAS speaker. [s] duration = 62 ms

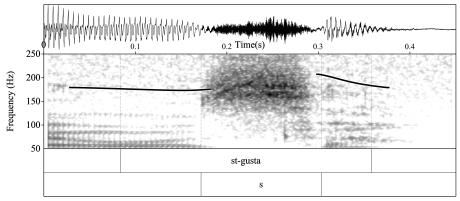


Figure 5: Sample spectrogram for the alveolar fricative [s] from a WAS speaker. [s] duration = 132 ms

respectively). For BAS [st] vs. [ht^h], the lenited variant was also favored word-medially (96%) more so than in the between-word context (72%). For the [t, ht] vs. [ht^h] comparison in the same dialect, following /i u/ favored [ht^h] (14% and 18%) compared to following /a/ (4%). Finally, in the WAS model, a following /i u/ favored [t^h] with a predicted probability of 98-100%, while for following /a/ probability was

Production type	NCPS%	BAS%	WAS%
[st]	**72.32**	5.62	0.05
[st^h]	23.22	0.61	0.35
[t, ht]	1.06	**62.06**	5.53
[ht^h]	0.16	26.96	2.41
[t^h]	0.05	3.01	**68.56**
[t^s]	0.00	0.00	12.68
[s]	3.18	1.67	8.15
[tʃ]	0.00	0.08	2.26

Table 1: Frequency of token types by dialect

92%. Male speakers also favored [t^h] (compared to [t^s]) with a probability of 97%, while female speakers did so at 92%.

Model: Comparison	Predictor	*df* & LRT-value
NCPS: [st] vs. [st^h]	FV***	4, 30.12
BAS: [st] vs. [t, ht]	WP*	1, 4.62
BAS: [st] vs. [ht^h]	WP*	1, 4.76
BAS: [t, ht] vs. [ht^h]	FV***	4, 29.96
WAS: [t^h] vs. [t^s]	FV**	4, 16.62
WAS: [t^h] vs. [t^s]	SEX**	1, 8.70

Table 2: Degrees of freedom and Likelihood-ratio test (LRT) value for sig. predictors in the five models. FV = following vowel; WP = word position; *** = $p < .001$; ** = $p < .01$; * = $p < .05$

4. DISCUSSION

Regarding this study's first goal, the spontaneous corpus yielded a richer set of variants than has been attested in the previous work on WAS post-aspiration [3, 5, 7, 9]. The more informal speech elicited in the spontaneous setting likely yielded a wider array of variants compared to methods involving controlled speech [15, 23]. The broader range of variants reflects the ways in which the sound change continues to unfold, as some WAS speakers show presumably more innovative productions than post-aspiration, such as the dento-alveolar affricate, the pre-palatal affricate, and the alveolar fricative (collectively accounting for 23.09% of the WAS corpus). Whereas Ruch [24] proposed that the shift from [t^h] to [t^s] reflects the most advanced phase of the sound change (since the [t^s] variant had not been described in traditional dialectological studies), the presence of the [s] variant in the WAS data suggests that some speakers may now be transitioning to a straightforward fricative. This would suggest a change such as: [t^h] → [t^s] → [s]. This proposal is supported by the fact that [s] tokens were notably more common among the WAS speakers than the two other groups

(cf. [25]).

In terms of the most frequent variant for each dialect, the NCPS speakers favored /s/-retention without long VOT, the BAS speakers favored /s/-weakening through pre-aspiration or coda deletion, and the WAS speakers favored post-aspiration of /st/. Notably, variants containing long VOT (i.e., values higher than 29 ms) surfaced in approximately 25% of each of the NCPS and BAS corpora. That speakers from "non-post-aspirating" varieties produced /st/ with long VOT suggests that lengthened VOT was likely present in the ambient language prior to the actuation of both pre- and post-aspiration. From this finding, we might infer that, prior to the actuation of WAS post-aspiration, language users were already attuned to the idea that the release of phonological /st/ is susceptible to lengthening in a fashion unlike phonological /sp/ or /sk/ (see [9]). WAS language users may have capitalized on this implicit knowledge by exaggerating the longer VOT of /st/, when post-aspiration was initially advancing in /st/-words.

Regarding the second goal of this study, the two most common significant predictors in the regression models were WORD POSITION and FOLLOWING VOWEL. Regarding the effect of WORD POSITION, in the BAS model, speakers favored lenited variants word-medially, but /s/-retention across word boundaries. This finding corroborates previous research on Spanish /s/-lenition showing that speakers are more likely to lenite preconsonontal /s/ word-medially (where there is no allomorphy) than across word boundaries (where word-final /s/ realization varies depending on the linguistic environment, e.g., [26, 27]). Concerning FOLLOWING VOWEL, this predictor was significant in three of the five models, and in each instance speakers favored /st/ variants containing long VOT (i.e., [st^{h}], [ht^{h}], or [t^{h}]) in the context of the following high vowels /i u/. In contrast, the variants lacking long VOT were most favored in the context of the following low vowel [a]. One possible explanation for this finding is that the transition of the vocal tract between [h] and [i, u] requires less articulatory effort than in the context of the low vowel [a] (see [28], p. 325-326). This is further supported by [29], which showed that VOT in English was longer before sonorants and high vowels than before mid and low vowels.

SEX emerged as a significant predictor in the WAS model, with female speakers favoring the presumably more advanced variant [t^{s}] in comparison to male speakers [25]. Labov's Principle II states that women (i.e., referring to all speakers assigned female at birth) are most often innovators in changes from below [30]. Our study aligns with Principle II in showing the female speakers favor the innovative [t^{s}] variant of the post-aspiration process.

Regarding this study's third goal, the findings corroborate the proposal from [5] that pre-aspirating language users are exposed to a notable proportion of /st/ tokens involving a long VOT (this was confirmed in the BAS data). Yet, in addition, the results for the NCPS speakers demonstrate that longer VOT is also present in /s/-retention varieties, which suggests that this acoustic characteristic was already present in the input prior to Andalusian speakers historically shifting to pre-aspiration. We might therefore infer that the higher intraoral air pressure of pre-aspirated [ht] (per [5]) was likely not the single source of lengthened VOT that could have triggered actuation of post-aspiration. It is nonetheless important to note the presence of pre-aspiration and/or lengthened VOT in /st/-words does not necessarily result in the development of post-aspiration in Spanish (see [31]).

Finally, regarding limitations of this study, future work might analyze VOT in more gradient terms (i.e., by extracting raw VOT values) for the sub-set of "long" VOT tokens identified here as /st/ productions with values higher than 29 ms (per [19]). A second limitation of this study is that, within each of the dialects, the frequency counts for numerous variants were too low for reliable statistical analysis. It is difficult to make generalizations about rarely occurring variants [32] and would thus be advisable for future studies to elicit such tokens in a more controlled setting [15]. Researchers interested in the replicability of our findings might also pursue automatic methods of data coding in addition to the impressionistic approach adopted here (which yielded eight variants)

5. CONCLUSION

This study analyzed the production of phonological /st/ clusters in spontaneous Spanish. The cross-dialectal results provided implications for the study of sound change, in particular, regarding the importance of analyzing data from as-yet unaffected speakers of the change in question (see [18]), in this case the NCPS and BAS participant groups. The results for the WAS speakers additionally unveiled innovative variants resulting from the change from pre- to post-aspiration, such as dento-alveolar affricates and alveolar fricatives. We proposed that, for some WAS speakers, the change may be shifting as follows: [t^{h}] → [t^{s}] → [s] (cf. [25]).

6. REFERENCES

[1] Carvalho, A. M. 2006. Spanish (s) aspiration as a prestige marker on the Uruguayan-Brazilian border. *Spanish in Context* 3(1), 85–114.

[2] Catalán, D. 1989. *El español: orígenes de su diversidad*. Thomson-Paraninfo.

[3] Henriksen, N., Galvano, A., Fischer, M. 2023. Sound change in western andalusian spanish: Investigation into the actuation and propagation of post-aspiration. *Journal of Phonetics* 98, 101238.

[4] Parrell, B. 2012. The role of gestural phasing in Western Andalusian Spanish aspiration. *Journal of Phonetics* 20, 37–45.

[5] Ruch, H., Peters, S. 2016. On the origin of post-aspirated stops: Production and perception of /s/ + voiceless stop sequences in Andalusian Spanish. *Laboratory Phonology* 7, 1–36.

[6] Torreira, F. 2006. Coarticulation between aspirated-s and voiceless stops in spanish: An interdialectal comparison. *Selected proc. 9th Hispanic Linguistics Symposium*. Cascadilla Press, 113–120.

[7] Ruch, H., Harrington, J. 2014. Synchronic and diachronic factors in the change from pre-aspiration to post-aspiration in Andalusian Spanish. *Journal of Phonetics* 45, 12–25.

[8] Browman, C. P., Goldstein, L. M. 1986. Towards an Articulatory Phonology. *Phonology* 3, 219–252.

[9] Torreira, F. 2012. Investigating the nature of aspirated stops in Western Andalusian Spanish. *Laboratory Phonology* 42, 49–63.

[10] Harrington, J. 2010. *Acoustic phonetics* volume 2, 81–129. Wiley-Blackwell.

[11] Ohala, J. 1983. *The origin of sound patterns in vocal tract constraints*, 189–216. Springer.

[12] Stevens, K. 1980. Discussion during symposium on phonetic universals in phonological systems and their explanation. *Proceedings of the 9th ICPhS* volume 2, 181–194.

[13] Gilbert, M. B. 2022. *An Experimental and Formal Investigation of Sevillian Spanish Metathesis*. PhD thesis New York University.

[14] Ruch, H. 2008. Affrication of /st/ in Western Andalusian Spanish: Variation and change from a sociophonetic point of view. Master's thesis Ludwig-Mazimilians-Universität München.

[15] Xu, Y. 2010. In defense of lab speech. *Journal of phonetics* 38(3), 329–336.

[16] Eckert, P., Rickford, J. R. 2001. *Style and Sociolinguistic Variation*. Cambridge University Press.

[17] Torreira, F. 2007. Pre- and postaspirated stops in Andalusian Spanish. *Segmental and Prosodic Issues in Romance Phonology*, 67–82.

[18] Carignan, C., Coretta, S., Frahm, J., Harrington, J., Hoole, P., Joseph, A., Kunay, E., Voit, D. 2021. Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German. *Language* 97(2), 333–364.

[19] Quilis, A. 1981. *Fonética acústica de la lengua española*. Gredos Madrid.

[20] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[21] R Core Team, 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[22] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3), 255–278.

[23] Henriksen, N. 2013. Style, prosodic variation, and the social meaning of intonation. *Journal of the International Phonetic Association* 43(2), 153–193.

[24] Ruch, H. 2010. Affrication of /st/-clusters in western andalusian spanish: variation and change from a sociophonetic point of view. *Proceedings of the Workshop "Sociophonetics, at the crossroads of speech variation, processing and communication"*.

[25] Vida-Castro, M. 2022. On competing indexicalities in southern peninsular spanish. a sociophonetic and perceptual analysis of affricate [ts] through time. *Language Variation and Change*, 1–27.

[26] File-Muriel, R. J., Brown, E. K. 2011. The gradient nature of *s*-lenition in Caleño Spanish. *Language Variation and Change* 23(2), 223–243.

[27] Horn, M. 2013. *An acoustic investigation of postaspirated stops in Seville Spanish*. PhD thesis Ohio State University.

[28] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Blackwell Publishers Ltd.

[29] Klatt, D. H. 1975. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. *Journal of Speech and Hearing Research* 18(4), 686–706.

[30] Labov, W. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change* 2(2), 205–254.

[31] Monroy, R., Hernández-Campoy, J. M. 2015. Murcian spanish. *Journal of the International Phonetic Association* 45(2), 229–240.

[32] Möbius, B. 2003. Rare events and closed domains: Two delicate concepts in speech synthesis. *International Journal of Speech Technology* 6(1), 57–71.

ACOUSTIC ANALYSIS OF VOWEL NASALIZATION IN TAIZHOU CHINESE

Qianwen Han and Wai-Sum Lee

Department of Linguistics and Translation, City University of Hong Kong
veronicahan923@hotmail.com

ABSTRACT

This study investigates the nasality of the nasal vowels, [ĩ ũ ẽ], and the nasalized V[ŋ] rhymes, [iŋ əŋ aŋ ɔŋ], in Taizhou Chinese. It also examines the weakening of the final nasal in the V[ŋ] rhymes. Praat was used to perform A1-H1, A1-P1, and MaxDA3 measurements of the test nasal vowels and V[ŋ] rhymes and their comparable oral counterparts produced by ten native Taizhou speakers. The analyzed data show that (i) [ĩ] and [ũ] are nasalized throughout; (ii) [ẽ] is partially or nearly denasalized, which is accompanied by the diphthongization of the oral counterpart [e]; (iii) in [əŋ aŋ ɔŋ], the vowel is nasalized throughout, while the final [ŋ] may be dropped when preceded by the non-back vowel [ə] or [a]; and (iv) in [iŋ], [i] is nasalized throughout when [ŋ] is retained, but nasalized partly when [ŋ] is dropped.

Keywords: Nasal vowels, nasalized rhymes, acoustic analysis, degree of nasality, Taizhou Chinese

1. INTRODUCTION

Taizhou Chinese (TZC) is spoken in Jiangsu Province on the north bank of the Yangtze River, the boundary region of Jiang-Huai Mandarin and Northern Wu [1]. Most of the dialects, including TZC, in the Jiang-Huai Mandarin group consist of nasal vowels [2, 3, 4, 5, 6]. Based on the vowel formant data in a recent study [7], the vowel system of TZC includes [i/ɿ y ɯ u e o ɚ ɑ ĩ ũ ẽ] in (C)V syllables, [i u e ɔ ə a] in (C)V[ʔ] syllables, and [i ɔ ə a] in (C)V[ŋ] syllables. Aside from the nasal vowels, [ĩ ũ ẽ], in (C)V syllables, the vowels, [i ɔ ə a], in (C)V[ŋ] syllables are naturally nasalized due to the assimilation effect from the syllable-final nasal [ŋ]. According to [4] and [5], the syllable-final [ŋ] in TZC is weakened, as the tongue dorsum may not raise high enough to make complete closure in the mouth cavity. The weakening and nasalization effect of [ŋ] may cause the syllables to change from (C)V[ŋ] to (C)Ṽ, merging into the (C)V syllables that contain a nasal vowel. Furthermore, it is pointed out in [3] and [8] that the nasality of the nasal vowel [ẽ] in (C)V syllables is weak. The denasalization of [ẽ] may cause a merger of [ẽ] and [e] in (C)V syllables.

This study investigates the nasality of the nasal vowels [ĩ ũ ẽ] in (C)V syllables and the nasalized rhymes [iŋ əŋ aŋ ɔŋ] in (C)V[ŋ] syllables of TZC through performing acoustic analysis. It also examines the weakening of the final nasal in the V[ŋ] rhymes.

2. METHOD

2.1. Speakers

Speech samples for acoustic analysis were collected from five male and five female speakers, aged 50 to 60 years, who were born and grew up in the city of Taizhou. Compared to younger speakers, older speakers are more common to use TZC as the primary language in daily communications, although they also speak Putonghua to people, in particular the non-Taizhou speakers.

2.2. Test materials

Test words with Ṽ or V[ŋ]	Test words with V or V[ʔ]
[pĩ^{44}] 變 (change) [tĩ^{44}] 電 (electricity) [iŋ^{45}] 贏 (win) [tiŋ^{44}] 定 (certainly)	[iʔ^{4}] 一 (one) [tiʔ^{4}] 跌 (tumble)
[ũ^{45}] 完 (completion) [kũ^{44}] 貫 (through)	[u^{45}] 鵝 (goose) [ku^{44}] 過 (pass)
[ẽ^{44}] 案 (case) [tẽ^{44}] 旦 (dawn)	[e^{44}] 愛 (love) [te^{44}] 帶 (bring)
[ɔŋ^{45}] 橫 (horizontal) [kɔŋ^{44}] 貢 (devotion)	[ɔ^{44}] 奧 (mystery) [kɔ^{44}] 告 (tell)
[aŋ^{45}] 昂 (raise) [kaŋ^{44}] 杠(dumbbell)	[ɑ^{45}] 牙 (tooth) [pɑ^{44}] 爸 (father)
[əŋ^{21}] 硬 (hard) [təŋ^{44}] 凳 (chair)	[ɚ^{213}] 耳 (ear) [ɚ^{45}] 兒 (kid)

Table 1: Test words for nasality measurement

The test materials included a set of monosyllabic words in TZC that contain the three nasal vowels [ĩ ũ ẽ] in (C)V syllables and the four nasalized rhymes [iŋ əŋ aŋ ɔŋ] in (C)V[ŋ] syllables. For comparison purposes, the monosyllabic words that contain the comparable oral counterparts [u e ɚ ɑ ɔ] in (C)V syllables and [i] in (C)V[ʔ] syllables were also analyzed. Note that the oral vowel [i] in (C)V[ʔ] syllables used for comparison is because [i] in (C)V syllables is apicalized to become [ɿ] [8]. As shown in Table 1, each of the target vowels or rhymes occurs in two test words which are meaningful and commonly used in TZC. The test words were presented in Chinese characters in a randomized list.

The speakers uttered the test words at a normal speech rate and repeated the word list five times.

2.3. Data analysis

The target vowels or rhymes in the test words were acoustically analyzed for the degree of nasality. There is an array of acoustic parameters commonly used to quantify the vowel nasality, such as (i) A1-H1 (difference in amplitude between the first formant A1 and the first harmonic H1), (ii) A1-P0 (difference in amplitude between the first formant and the nasal formant at 250Hz or 450Hz, whichever has a higher amplitude, and (iii) A1-P1 (difference in amplitude between the first formant and the nasal formant at 950Hz) [9, 10, 11, 12]. Based on the evaluation results in [13] and [14], A1-H1 and A1-P0 are more accurate than the other parameters to signal the nasalization of non-high vowels, whereas A1-P1 is more appropriate for quantifying the nasality of high vowels. In this study, two parameters, A1-P1 and A1-H1, were used for doing nasality measurement. For the high front vowels [i ĩ] and the rhyme [iŋ] that contains [i], A1-P1 was used. For the non-high vowels [e ẽ ɐ̃ ɑ ɔ] and the rhymes [əŋ aŋ ɔŋ] that contain [ə a ɔ], A1-H1 was used. For the high back vowels [u ũ], A1-H1, rather than A1-P1, was used. This is because the second formant of [u ũ] is close to the nasal formant at 950Hz, which may cause a boost to the amplitude of the nasal formant.

Using the Nasality Auto-measure Praat script [15], the measurements of A1-H1 and A1-P1 were made at every 10% of the duration of the vowel or rhyme in the test words. The value of A1-H1 or A1-P1 is inversely correlated to the degree of nasality, i.e., a smaller value of A1-H1 or A1-P1 suggesting a higher degree of nasality.

For the test V[ŋ] rhymes, the third formant amplitude at every 10% of the duration of the rhymes was also measured. The maximum difference in the third formant amplitude (MaxDA3) between two successive time points was taken to indicate the location of the boundary between the vowel and the final nasal in the V[ŋ] rhymes. This is in reference to the method used in [12], which is based on the visual inspection of a large reduction in energy of the vowel formants near the boundary between a preceding vowel and a following nasal consonant. In the present study, the spectrograms of the V[ŋ] rhymes, [iŋ əŋ aŋ ɔŋ], show a large reduction in energy of the third and higher formants near the end of the rhymes with a strong final nasal. Thus, the MaxDA3 is taken and it is expected to be large for the V[ŋ] rhymes as compared with the oral counterparts. In other words, a small MaxDA3 for the V[ŋ] rhymes indicates the weakening or drop of the final nasal in the rhymes.

In order to compare the degree of nasality across the different types of vowels and rhymes and across different speakers, all the measured values of A1-H1, A1-P1, and MaxDA3 of the test vowels and rhymes were normalized for each speaker by converting the values into z-score.

3. RESULTS & DISCUSSION

3.1. Nasal vowels in (C)V syllables

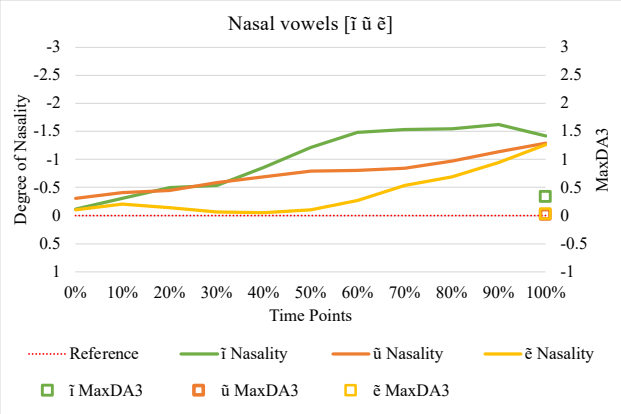


Figure 1: Degree of nasality and MaxDA3 for the nasal vowels [ĩ ũ ẽ] in (C)V syllables

Figure 1 shows the degree of nasality for the three nasal vowels [ĩ ũ ẽ] in (C)V syllables. It is computed by subtracting the mean value (n = 100, i.e., 2 test words per vowel x 5 repetitions per test word x 10 speakers) of A1-H1 or A1-P1 (in z-score) for an oral counterpart from that of the corresponding nasal vowel. The values of the oral counterparts ([i u e]) are set as '0' and used as reference (red dotted line) for comparison. Since the value is inversely correlated to the degree of nasality, the values of the nasal vowels are smaller than the reference and below zero. The figure also presents the MaxDA3 values (in z-score) of the three nasal vowels which minus those of the oral counterparts (set as '0' and used as reference). A small value of MaxDA3 suggests that the presence of a post-vowel nasal consonant is unlikely.

As shown in Figure 1, for all the three nasal vowels [ĩ ũ ẽ] (solid lines) the degree of nasality is increasing and becoming high towards to the vowel offset. The increase in nasality is not because of the presence of a post-vowel nasal as evidenced by the MaxDA3 values of the three vowels. As can be seen, the MaxDA3 is slightly above the reference '0' for [ĩ] (green square) and it is almost '0' for [ũ] (orange square) and [ẽ] (yellow square), indicating the unlikelihood to have a nasal consonant after [ĩ ũ ẽ]. A comparison of the three nasal vowels shows that the degree of nasality is lower for the mid vowel [ẽ] (yellow solid line) than the two high vowels [ĩ ũ] (green and orange solid lines). For [ẽ], the degree of

nasality is close to the reference '0' line during the first half of the vowel, while it increases during the second half of the vowel. For [ĩ] and [ũ], the degree of nasality is above '0' and increases continuously from the vowel onset to the vowel offset. Thus, the results suggest that the mid vowel [ẽ] is lesser nasalized than the two high vowels [ĩ] and [ũ]. While both [ĩ] and [ũ] are nasalized throughout, [ẽ] is partially denasalized, with no nasalization during the first half of the vowel. The data agree with the report in [3] and [8] that the nasality of the nasal vowel [ẽ] in (C)V syllables is weak.

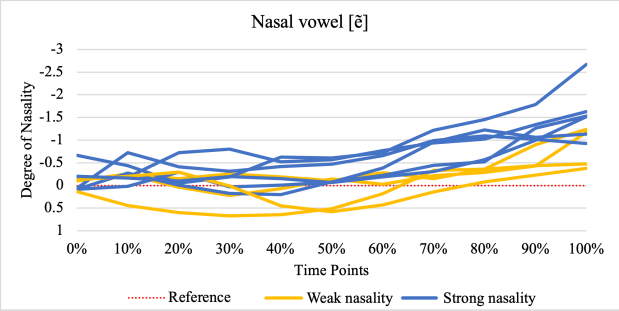


Figure 2: Degree of nasality of the nasal vowel [ẽ] for 10 individual speakers

For the vowel [ẽ], the between-speaker variation in the degree of nasality is actually large. Figure 2 presents the degree of nasality of [ẽ] for each of the ten speakers in this study. As can be seen, the nasality of [ẽ] is weaker for four speakers (yellow solid lines) the other six speakers (blue solid lines). For the former four speakers, the degree of nasality of [ẽ] is below or around the reference '0' value throughout the vowel, although a mild rise may occur near the vowel offset. The data suggest that for these speakers the vowel [ẽ] is nearly denasalized throughout. The denasalization of [ẽ] may cause a merger of [ẽ] and the oral [e] in TZC.

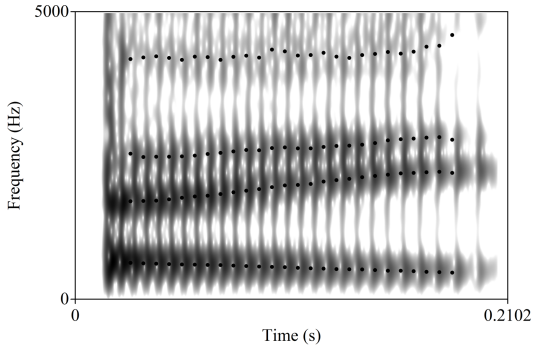


Figure 3: Superimposed formant trajectories on the spectrogram of the diphthongized [e] in [e^{44}] 愛 produced by a male speaker

Accompanying the denasalization of [ẽ], the diphthongization of the oral [e], changing from a monophthong to [eɪ], is observed. Figure 3 presents a case in which the vowel [e] is produced with the changing formants ($F_1F_2F_3$) by a male speaker. As shown in the figure, the continuous decrease in F_1 and increase in F_2 suggest that [e] changes to a high front vowel towards to the end. Table 2 presents the mean values (n = 50, i.e., 2 test words per vowel x 5 repetitions per test word x 5 speakers of the same gender) of F_1 and F_2 (in Hz) at the time points of 20% and 80% of the vowels [ẽ] and [e] for the five speakers of each gender in this study. Compared with [ẽ], there is a marked increase in F_2 from the first to the second time point for [e], which suggests [e] is fronted near the end. Also, there is a decrease in F_1 between the two time points for [e], which suggests [e] is raised, although the change is less pronounced for the male speakers due to cross-speaker and cross-token variations. In this study, no diphthongization of [e] is observed for three male speakers, while [e] is diphthongized in over 40% (one male and all five female speakers) or all 100% (one male speaker) of the test tokens for the other seven speakers. Four out of the latter seven speakers produce the diphthongized [e] and the denasalized [ẽ]. The parallel change in [ẽ] (denasalization) and [e] (diphthongization) is probably for maintaining the distinction between the two vowel phonemes.

Speakers	Formants	[ẽ]		[e]	
		20%	80%	20%	80%
Male	F_1	529	596	498	471
	F_2	2050	2008	1894	2031
Female	F_1	555	613	603	519
	F_2	2551	2484	2312	2482

Table 2: Mean values (in Hz) of F_1 and F_2 at the time points of 20% and 80% of [ẽ] and [e] for the male and female speakers in this study

3.2. V[ŋ] rhymes

Figure 4 and Figure 5 show the degree of nasality and the MaxDA3 for the test nasalized V[ŋ] rhymes in this study. Agreeing with the report in [4] and [5], this study observes that the final nasal in the V[ŋ] rhymes, [əŋ aŋ iŋ] (except for [ɔŋ]), is not always strong and the V-N boundary between the preceding vowel and the final nasal in the rhymes is not clear in some of the test tokens. These rhymes are thus divided into [əŋ aŋ iŋ] and [ə^{ŋ} a^{ŋ} i^{ŋ}]. In Figure 4 and Figure 5, the markers (circles) that represent the MaxDA3 values of [ɔŋ əŋ aŋ iŋ] with a strong final nasal are placed at the location of the V-N boundary of the rhymes, whereas the markers (squares) that represent the MaxDA3 values of [ə^{ŋ} a^{ŋ} i^{ŋ}] with a weak final nasal are placed at the end of the rhymes.

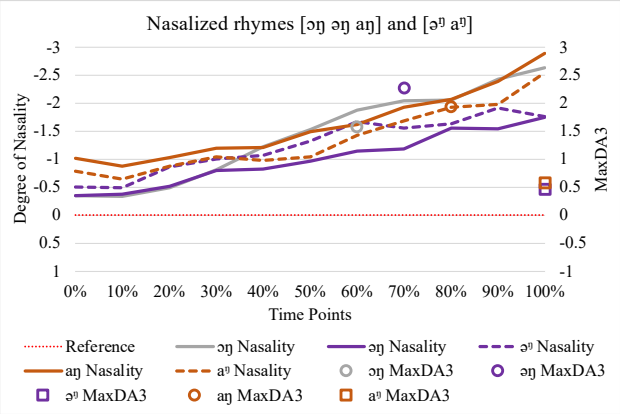


Figure 4: Degree of nasality and MaxDA3 for the nasalized rhymes [ɔŋ əŋ aŋ] and [əᵑ aᵑ]

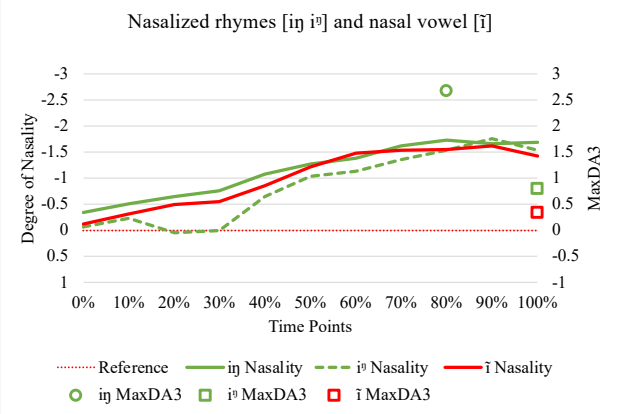


Figure 5: Degree of nasality and MaxDA3 for the nasalized rhymes [iŋ iᵑ] and the nasal vowel [ĩ]

As shown in Figure 4, the value of MaxDA3 is noticeably large for [ɔŋ əŋ aŋ], but small and close to the reference '0' for [əᵑ aᵑ]. The data prove that the final nasal is strong and retained in [ɔŋ əŋ aŋ], but weak or nearly dropped in [əᵑ aᵑ]. In spite of the difference in the final nasal between [ɔŋ əŋ aŋ] and [əᵑ aᵑ], the two types of rhymes are similar in the degree of nasality. For [ɔŋ əŋ aŋ] (solid lines) and [əᵑ aᵑ] (dashed lines), the degree of nasality of the entire rhyme is above the reference '0' and increases continuously towards to the end of the rhyme. Naturally, the increase in nasality for [ɔŋ əŋ aŋ] is related to the presence of the final nasal in the rhymes. In these rhymes, the final nasal also causes the preceding vowel to have nasalization throughout. As for [əᵑ aᵑ], while the final nasal is weakened and nearly dropped, the nasality is retained on the preceding vowel in the rhymes, although the degree of nasality is lower near the vowel onset than the vowel offset. Nonetheless, the degree of nasality is above the reference '0' throughout the vowel in [əᵑ aᵑ], and it is similar to the degree of nasality of [əŋ aŋ]. The nasality data therefore suggest that the rhymes [ɔŋ əŋ aŋ] contain a nasal vowel and they are actually [ɔ̃ŋ ə̃ŋ ãŋ]. As for the rhymes [əᵑ aᵑ], they contain a single nasal vowel [ə̃] or [ã], changing from the structure of VN to Ṽ. Since the vowel inventory of TZC does not include the nasal vowels [ə̃] and [ã], no merger is resulting from the change in the rhymes [əᵑ] and [aᵑ].

In Figure 5, the degree of nasality and the MaxDA3 values for the rhymes [iŋ iᵑ] are in comparison with those for the nasal vowel [ĩ] in (C)V syllables. As can be seen, the MaxDA3 of [iᵑ] (green square) is similar to that of the nasal vowel [ĩ] (red square) and near the reference '0', whereas the MaxDA3 is markedly large for [iŋ] (green circle). The MaxDA3 data indicate that the final nasal is weakened or nearly dropped in [iᵑ], but retained in [iŋ]. Despite the weakening of the final nasal in [iᵑ], the nasality of the rhyme is retained. As shown in Figure 5, while the degree of nasality of [iᵑ] (green dashed line) is low and close to the reference '0' during the initial 30% of the rhyme, it largely increases during the following 70% of the rhyme and becomes similar to the degree of nasality of the nasal vowel [ĩ] (red solid line). The nasality data suggest that the vowel in [iᵑ] is nasalized, but properly it is partially nasalized with no nasalization near the vowel onset, which is different from the nasal vowel [ĩ] with nasalization throughout. It follows that even the final nasal in [iᵑ] is dropped, the vowel in the rhyme is still distinguishable from the nasal vowel [ĩ] in (C)V syllables. As for [iŋ], the degree of nasality (green solid line) is above the reference '0' throughout the rhyme and it increases continuously towards to the end due to the presence of the final nasal in the rhyme. In general, the degree of nasality of [iŋ] is similar to that of the nasal vowel [ĩ] (red solid line), suggesting that [iŋ] is actually [ĩŋ], containing a nasal vowel [ĩ]. Since the final nasal in [ĩŋ] is retained, the nasalization change in the vowel of the rhyme does not result in a merger with the nasal vowel [ĩ] in (C)V syllables.

4. CONCLUSION

To sum up, the nasal vowels and the vowels in the nasalized V[ŋ] rhymes in TZC vary in the degree of nasality. For the nasal vowels, the degree of nasality is lower for the mid vowel [ẽ] than the high vowels [ĩ ũ], and denasalization of [ẽ] is observed for some speakers. In the rhymes [ɔŋ əŋ aŋ iŋ], the vowel is nasalized, and the final nasal may be dropped when preceded by the non-back vowels [i ə a]. The degree of nasality of [ɔŋ əŋ aŋ] is high, regardless of whether the final nasal is retained or dropped. As for the rhyme [iŋ], the degree of nasality is high when the final nasal is retained but reduced when the final nasal is dropped. For future studies, aerodynamic data will be obtained to further substantiate the characteristics of vowel nasalization in TZC.

5. REFERENCES

[1] Xu, X.B. 2011. Jiangsu Taizhou diqu fangyan de lishi yanbian [Historical development of the Taizhou dialect in Jiangsu Province]. *Literatures* 11, 74-75.

[2] Wu. B. 2007. *Jianghuai Guanhua Yuyin Yanjiu* [A Study on Jiang-Huai Mandarin]. Doctoral Dissertation, Fudan University, Shanghai, China.

[3] Zhang, J.M. (ed). 1991. *Taixian Fangyan Zhi* [Taixian Dialect Chronicle]. Shanghai: East China Normal University Press.

[4] Nanjing University (ed). 1998. *Jiangsu Shengzhi Fangyanzhi* [Jiangsu Province Chronicle: Dialects]. Nanjing: Nanjing University Press.

[5] Gu, Q. 2001. *Tongtai Fangyan Yinyun Yanjiu* [Phonetics and Phonology Study on Tongtai Dialect]. Doctoral Dissertation, Nanjing University, Nanjing, China.

[6] Lu, G.Y. 2003. Taizhou Fangyinshi yu Tongtai Fangyanshi Yanjiu [A study on the history of Taizhou dialect and Tongtai sector]. In Lu, G.Y. (ed), *Lu Guoyao Yuyanxue Lunwenji* [Collection of Lu's Papers on Linguistics]. Jiangsu: Jiangsu Education Press, 12-122.

[7] Han, Q. 2022. *Vowel apicalization in Taizhou Chinese* [Paper presentation]. Young Scholar Forum on Phonetics Study, Beijing.

[8] Han, Q., Lee, W-S. 2019. Formant analysis of the vowels in Taizhou Chinese. In Calhoun, S., Escudero, P., Tabain, M., Warren, P. (eds.), *Proc. 19th ICPhS* Melbourne, 3588-3592.

[9] Huffman, M.K. 1990. The role of F1 amplitude in producing nasal percepts. *J. Acoust. Soc. Am*. 88, S54.

[10] Chen, M.Y. 1995. Acoustic parameters of nasalized vowels in hearing impaired and normal-hearing speakers. *J. Acoust. Soc. Am.* 98, 2443-2453.

[11] Chen, M.Y. 1997. Acoustic correlates of English and French nasalized vowels. *J. Acoust. Soc. Am.* 102, 2360-2370.

[12] Chen, M. Y. 2000. Acoustic analysis of simple vowels preceding a nasal in standard Chinese. *Journal of Phonetics* 28, 43-67.

[13] Berger, M.A. 2007. *Measurement of Vowel Nasalization by Multi-dimensional Acoustic Analysis*. Master's Thesis, University of Rochester, Rochester, New York.

[14] Styler, W. 2017. On the acoustical features of vowel nasality in English and French. *J. Acoust. Soc. Am.* 142, 2469-2482.

[15] Nasality Auto-measure Script Package https://github.com/stylerw/styler_praat_scripts/tree/master/nasality_automeasure.

An acoustic study of the realisation of KIT in the conversational speech of young English speakers in Australia

Gerard Docherty[1], Simón Gonzalez[2], Paul Foulkes[3].

Griffith University[1], Australian National University[2], University of York[3]
gerry.docherty@griffith.edu.au; simon.gonzalez@anu.edu.au; paul.foulkes@york.ac.uk

ABSTRACT

This study sheds new light on the under-investigated Weak Vowel Merger (WVM) in Australian English whereby an historical contrast between unstressed KIT and SCHWA has been lost in favour of a central schwa-like variant. We address the following questions: (a) to what extent are unstressed vowels derived historically from KIT realised with a centralised quality as per conventional accounts of WVM; (b) does the status of the lexical item as grammatical/non-grammatical impact significantly on the realisation of the unstressed syllables; (c) to what extent is any centralisation correlated with the duration of the unstressed vowels? We present acoustic analyses of the quality and duration of 2745 tokens of historical KIT by 40 speakers (20M/F) aged 18-22 from Perth recorded for c. 30 minutes engaged in an unscripted conversational interaction. While our findings are largely consistent with the conventional WVM analysis, they point to additional complexity that warrants further investigation.

Keywords: Australian English, Weak Vowel Merger, conversational speech

1. INTRODUCTION

Descriptions of English spoken in Australia [1, 2, 3] note that it is characterised by a process of Weak Vowel Merger (WVM) whereby the historical contrast between KIT and SCHWA vowel categories is lost in unstressed syllables, resulting in a centralised realisation and homophonous pairs such as *Rosa's*/*roses* or *Lennon*/*Lenin*. This centralisation does not, however, apply uniformly to unstressed tokens of the historical KIT vowel. Exceptions include the so-called 'happY-tensing' context where a closer variant approximating [i] is the norm (e.g. *movie, cities*), and in unstressed pre-velar and pre-postalveolar contexts where centralisation does not occur (e.g. in *frantic, rubbish, singing*).

For all that WVM is cited uncontroversially as a feature of English in Australia, Cox & Palethorpe's [3] is the only study to date providing relevant quantitative data in relation to this feature. The authors of that study focused on the acoustic characteristics (F1/F2) of word-final SCHWA realisations, comparing across a range of contexts including tokens of historical KIT. The specific comparisons made were between word-final lexical tokens of SCHWA (e.g. *Rosa*), and tokens that were followed by a possessive suffix (e.g. *Rosa's*), and others that were the nucleus of a plural suffix (e.g. *Roses*), the latter being the historical KIT context. The study documented variation across the various experimental conditions, but importantly noted that in line with the predictions of the WVM there was no significant difference between the latter two contexts. However, the findings also suggest that the realisation of the nucleus in these latter two contexts yielded a more [ɨ]-like realisation that differed significantly from that of the word-final lexical SCHWA tokens. This in turn suggests that the WVM may not be straightforwardly captured as an absorption of the historical KIT tokens within the distribution of SCHWA more generally, and is reminiscent of findings from other varieties of English displaying WVM that also point to the merger of historical KIT with SCHWA being more complex than might be assumed [4, 5].

In this context, the aim of the present study is to shed new light on the acoustic properties of WVM in Australian English, departing from the sole previous study [3] in a number of important ways. First, we approach the putative merger from the point of view of the historical KIT vowel. Second, we focus on data from conversational speech. Finally, we include tokens from the high-frequency unaccented grammatical item *it* in order to test the extent to which grammatical items participate in the WVM-related variability. It was decided to focus on a single grammatical item at this point of the investigation simply to test how such items pattern in relation to the WVM, and thus to explore whether it would be important to sample from a wider range of grammatical items containing historical KIT at a future stage of this project.

The three questions that we address are (a) to what extent are unstressed vowels derived historically from KIT realised with a centralised quality as per conventional accounts of WVM; (b) does the status of the lexical item as grammatical/non-grammatical impact significantly on the realisation of the unaccented syllables; (c) to what extent is any centralisation correlated with the duration of the unstressed vowels? The findings are discussed in

relation to the specifics of WVM in English in Australia and in relation to what we know of vowel mergers more generally.

2. METHODS

The material analysed in this study was drawn from a corpus of recordings of pairs of 40 young (18-22) speakers from Perth (Western Australia) engaged in same-sex unscripted conversations each lasting around 30 minutes. All of the participants had been fully-schooled in Perth (from age 5) and there were equal numbers of males and females. While social class is not relevant to the present study, it can be noted that twenty of the speakers were residents of neighbourhoods ranked by the Australian Bureau of Statistics to be in the top socio-economic decile, and the remaining twenty were from neighbourhoods with a lower socioeconomic ranking. The majority of pairs of speakers knew each other in advance but to varying degrees. A fieldworker was present in the same room as the participants in order to initiate and conclude the conversation recording process but only intervened (rarely) if the participants' conversation subsided and was in need of a prompt.

The recordings (44Khz, 16 bit) were made using Sennheiser EW112-P-G3 lapel microphones and an Edirol R44 digital recorder. Conversations were transcribed with *Elan* [6] (starting 5 minutes in to each recording) and force-aligned within *LaBB-CAT* [7] using *HTK* [8], with a follow-up manual correction of misaligned segment boundaries. Using default settings in *Praat* [9], F1/F2 values were estimated for each vowel at the midpoint of each token (see [10] for caveats on this static approach to vowel description). Pre-/l, w, j/, pre-nasal, and post-/w, j, r/ environments were excluded, as well as post-nasal/lateral tokens where segmentation was problematic. The F1/F2 measurements were not normalised, in order to allow for comparison with previous studies (notably [3]), and consequently the findings for male and female results are reported separately below.

In order to address the research questions we created a subset of the corpus with tokens of historical KIT occurring in a number of different contexts (each identified henceforth by the relevant acronym from the list below). Three of these were contexts where descriptions of the WVM suggest that centralisation of the vowel should not be anticipated:

(a) MONO - nucleus of monosyllabic content word (e.g. *big, trip*)

(b) POLYAC - accented nucleus of a polysyllabic content word (e.g. *different, issue*)

(c) PREVEL - unaccented nucleus in polysyllabic lexical items with a following velar context (e.g. *allergic, public*)

One context was where centralised realisations were anticipated as per the WVM:

(d) UNAC - unaccented nuclei contained within a polysyllabic lexical item (e.g. *massive, habit, places*)

Two subsets of *it* were generated, the latter differentiated as a result of initial auditory analysis of the data suggesting that phrase-final position might be associated with great levels of reduction.

(e) PSINIT - tokens of grammatical item *it* occurring phrase-internally

(f) PSFINIT – tokens of grammatical item *it* occurring phrase-finally

In addition to the above, two further subgroupings of the data were identified:

(g) HAPPY_T – tokens of unstressed word-final KIT providing a context where we might anticipate that the vowel nucleus would be raised and fronted as per the pattern of happY-tensing reported for this and other varieties of English (e.g. *city, easy, movie*)

(h) FLEECE – as a point of reference we identified tokens of the FLEECE lexical set realised in monosyllabic content words. Note that in this case, in order to avoid the effect of the on-glide that can characterise the quality of this vowel for many speakers of Australian English, the formant measurements were taken at the 80% point of the vowel interval.

Table 1 provides a summary of the dataset, itemising the number of tokens in each subgroup.

	Females	Males
MONO	173	244
POLYAC	212	207
PREVEL	122	99
UNAC	356	288
PSINIT	86	79
PSFINIT	268	227
HAPPY_T	207	177
FLEECE	146	179
Total	1570	1500

Table 1: Number of tokens of each of the subgroupings of historical KIT included in the acoustic analysis (see main text for further explanation).

3. FINDINGS

In order to provide a single quantitative metric of the degree of centralisation for each token, we adopted a

method previously deployed by Labov et al. [11] and Grama et al. [12], calculating F2-(2*F1) as a derivative indicator of relative location along the front diagonal of the vowel space. We refer to this measure henceforth as F2DERIV (Hz). A higher value of F2DERIV equates to a relatively closer and fronter vowel quality.

Figure 1 shows separately for males and females the distribution of F2DERIV for each of the eight subgroupings of the dataset described above.

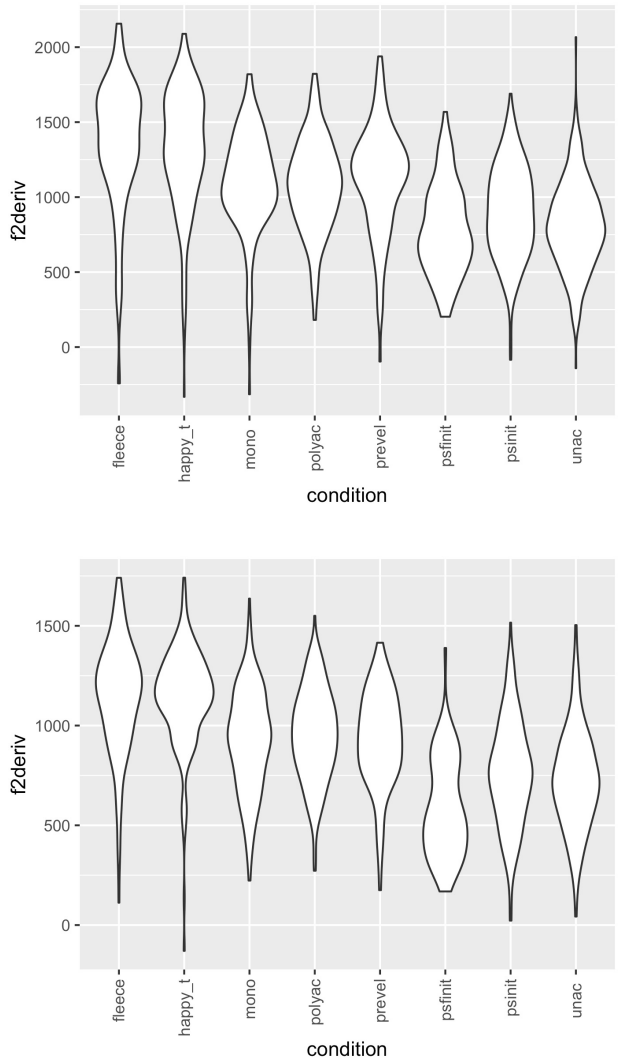


Figure 1: distribution of F2DERIV (Hz) for each of the eight subgroupings of the dataset described: females top panel; males lower panel.

The HAPPY_T and FLEECE tokens have the highest distribution of F2DERIV values, with a second cluster being formed by MONO, POLYAC and PREVEL tokens and the least close/front realisations being found for UNAC, PSINIT and PSFINIT tokens.

In order to gauge the extent to which the various distributions across the subgroupings were quantitatively different, mixed-effect models were calculated using the lmer function as part of the lme4 package [13] in R [14]. P-values were calculated using the lmerTest package [15]. For all models, F2DERIV was the dependent variable, speaker and word were included as random effects, and the subgrouping was included as a fixed effect. The data for males and females were modelled separately.

The quantitative analysis largely reflected the clustering evident from the plots. Using FLEECE as the reference grouping, the highest model estimates were found for HAPPY_T, which was not significantly different from FLEECE but different from all the other groupings. For both males and females, the UNAC grouping proved to be significantly different from MONO, POLYAC and PREVEL, but not from either of the *it* groupings. For female speakers, neither of the *it* groupings were differentiated statistically from the MONO/POLYAC/PREVEL set, whereas for male speakers, the phrase-final PSFINIT grouping was, thereby patterning with the UNAC grouping.

A similar picture is evident if the data are plotted in a conventional F1/F2 chart, although the differentiation is a little less clear to the eye (see Figure 2). This reinforces the validity of using the single F2DERIV measure as a metric for the variation that is targeted in this study.

We also examined the full set of historical KIT tokens for any relationship between F2DERIV values and the duration of vowel tokens. There was no compelling evidence of any correlation between the two. The vast majority of tokens had durations of less than 100 ms, and while there was a tendency for there to be only a few relatively long tokens in the low F2DERIV range, a subsequent statistical analysis yielded non-significant correlations for both male and female speakers.

4. DISCUSSION

In general, the findings are in line with what was anticipated, given the conditions under which WVM is said to apply and the previous acoustic study by Cox & Palethorpe [3]. The historical KIT vowels in HAPPY_T contexts are closer and fronter than those in MONO words, which in turn are fronter and closer than those found in unaccented syllables in the UNAC content words. Likewise there is a good deal of consistency in the realisation of historical KIT across MONO, POLYAC and PREVEL contexts, and more reduced variants overall are found in the UNAC and the two "*it*" groupings PSINIT and PSFINIT.

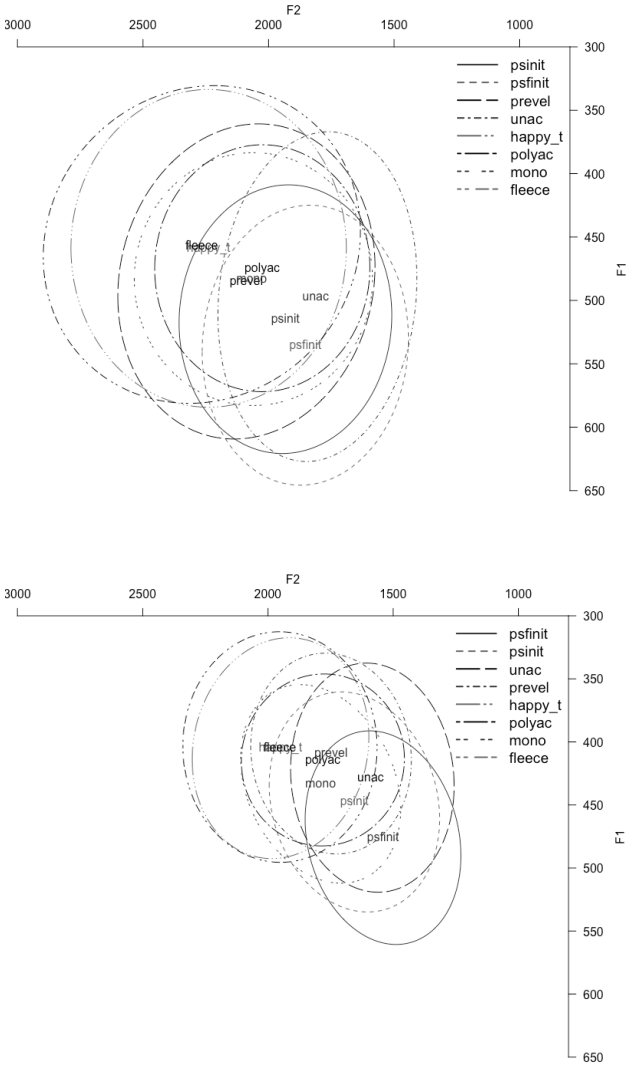


Figure 2: Distribution of realisations in F1/F2 space (Hz) for each of the eight subgroupings of the dataset described: females top panel; males lower panel; labels centred on the mean F1/F2 value with ellipses at +/- 1s.d.

The F1/F2 values for the more reduced variants, together with the extensive distributional overlap of all of the different subgroupings with each other (and in fact with FLEECE), are consistent with Cox & Palethorpe's [3]'s suggestion that WVM often results in a vowel which is more of an [ɨ] quality than a schwa. Of course, it is a point of debate whether schwa itself has a definable target that could act as a common point of reference for this sort of comparison [16], but the results reported here, together with those reported in [3], suggest that the conceptualisation of WVM as a simple merger of KIT with SCHWA in unstressed syllables may not be the best way to capture the patterns of realisation observed across this dataset. This interpretation could be tested by adding to the data sample tokens in which SCHWA figures as part of the phonological specification in order to provide the full FLEECE-to-SCHWA range in which to situation the historical KIT realisations. This forms part of the next stage of this line of investigation.

One obvious limitation of this study is that it has only tracked a single grammatical item, *it*. Further work will be needed to ascertain if *it* is typical of other unstressed grammatical words. While it appears to be patterning alongside unstressed KIT vowels in content words (excluding those with a following velar consonant), some further scrutiny is needed of what appears to be a bimodal distribution for males' realisation of *it* in phrase-final position possibly mapping to the auditory impression of occasional phrase-final tokens having particularly centralized realisations.

If the WVM is not straightforwardly a merger with SCHWA there is a question about the nature of the phonological representation of those lexical items which tend to have the [ɨ] realisation. When we refer to the KIT lexical set in English spoken by Australian speakers, do we simply need to specify that we are referring only to accented contexts? But it is a moot point whether for speakers and listeners, the vowels in *it, is, this, massive, radical* are identified in an exemplar-based representation as "the same" as the vowels in *delicious, Christmas, gifts*, in turn begging the question of whether the full range of contexts need to be included in accounts of ongoing variation and change in the vowel system. This is a significant issue to address given the high frequency of many of the relevant grammatical items. In order to resolve this, it would be instructive to undertake some auditory/ perceptual experimentation with speakers of English in Australia to gauge how they categorise those same vowels in comparison to how they categorise KIT variants in accented syllables.

Finally, we should note some non-canonical realisations that were characteristic of the conversational speech style represented by our sample; in particular, widespread reduction of *it's* to [ts] or [s], and regular elision of the vowel in [s_z] plural affix contexts (e.g. *places, faces*) with the unstressed syllable carried by continuous friction and a switch of voicing (noting the high frequency of the [s_z] plural affix). This underscores the value of investigating phonological variation within natural connected speech styles and the need to develop models of production and processing that are consistent with what we know of the phonetic properties of those styles.

Acknowledgment
Research funded by the Australian Research Council (DP130104275).

6. REFERENCES

[1] Wells, J. 1982. *Accents of English*. Cambridge University Press.

[2] Cox, F., Palethorpe, S. 2007. Australian English (Illustrations of the IPA). *Journal of the International Phonetic Association,* 37(3), 341-350.

[3] Cox, F., Palethorpe, S. 2018. Rosa's roses – unstressed vowel merger in Australian English. *Proc. 17th Australasian Int. Conf. on Speech Science and Technology* (SST17). Australasian Speech Science and Technology Association (ASSTA), 89-92.

[4] Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., Trudgill, P. 2004. *New Zealand English: Its Origins and Evolution.* Cambridge University Press.

[5] Flemming, E., Johnson, S. 2004. Rosa's roses: reduced vowels in American English. *Journal of the International Phonetic Association,* 31, 83-96.

[6] ELAN (Version 6.4) [Computer software]. 2022. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

[7] Fromont, R., Hay, J. 2012. LaBB-CAT: An annotation store. *Proc. Australasian Language Technology Workshop (ALTA),* 113-117.

[8] Young, S., Evermann, G., Gales, M. et al. 2006. *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department.

[9] Boersma, P., Weenink, D. Praat, 2018. Version 6.0.40.

[10] Cox, F. Docherty, G.J. Forthcoming. Sociophonetics and vowels. In Strelluf, C. (ed.). *The Handbook of Sociophonetics.* Routledge.

[11] Labov, W., Rosenfelder, I., Fruehwald, J. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89(1). 30–65.

[12] Grama, J., Travis, C.E., Gonzalez, S. 2019. Initiation, progression, and conditioning of the short-front vowel shift in Australia. *Proc. 19th Int. Congress Phon. Sci (ICPhS),* 1769-1773.

[13] Bates, D, Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67, 1-48.

[14] R: A language and environment for statistical computing. (2020). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org/

[15] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26.

[16] Cohen Priva, U., Strand, R. 2023. Schwa's duration and acoustic position in American English. *Journal of Phonetics,* 96, 101198.

The loss of intervocalic short /h/ in Estonian

Pire Teras

Institute of Estonian and General Linguistics, University of Tartu, Estonia
pire.teras@ut.ee

ABSTRACT

Estonian is a Finno-Ugric language that is characterised by the voicing or loss of both word-initial and intervocalic short /h/. The present paper aims to determine the acoustic characteristics of Estonian words where a short intervocalic /h/ has been lost. Disyllabic CVhV-words where the loss of /h/ has occurred are analysed. Two tendencies have been found: the loss of /h/ results either in a long monophthong or diphthong (CVV) or in a boundary between two vowels (CV.V). In most cases, the duration of V2 is longer than V1 but there are also variants where the loss of /h/ has lengthened the duration of the first vowel. Variants with a boundary are characterised by a slight dip in intensity between two vowels. Both variants' pitch contours resemble those of Estonian quantity 3 words that are characterised by an early F0 turning point and falling pitch in accented words.

Keywords: phonetics, variation, glottal fricative, sound loss, Estonian.

1. INTRODUCTION

Previous studies have shown that the voicing of short intervocalic unvoiced obstruents occurs quite frequently in Estonian [1]. Ermus [2, p. 36] pointed out that the reduction of short intervocalic plosives can result in this loss: for velar /k/ in 6% and alveolar /t/ in 4% of all analysed tokens.

Estonian /h/ is a voiceless glottal fricative. Teras [1] has suggested that the main variant of short intervocalic /h/ is its voiced variant: this occurred in 70% of analysed tokens. However, it appears that the loss of intervocalic short /h/ is also quite frequent: the reduction of /h/ resulted in a complete loss in 21% of analysed tokens and, additionally, in 3% of cases in complete loss and vowel shortening. This loss occurs first of all in informal situations, it is more frequent in the speech of male speakers and in unaccented words [1].

Voicing and loss or elision of consonants have both been connected with the term *lenition*, which means that the constriction degree or length of a consonant reduces as the result of minimising more effort-demanding articulatory gestures [3, pp. 313–314]. This phenomenon is well-known in the world's languages and the most typical context for it is intervocalic [3, p. 314]. Garellek et al. have analysed the voicing of glottal consonants and found that "both [h] and [ɦ] are likely to be as voiced as breathy vowels" [4, p. 23].

Melero García [5, p. 11] has shown that complete loss of intervocalic /d/ in Madrid Spanish leads to changes in intensity curves. Intervocalic /d/ was completely lost in 14% of analysed words, and in most cases, there was a slight dip in intensity between vowels (intensity ratio between 0.95 and 1). A study on the elision of /h/ in the Shirazi dialect of Persian shows that this loss occurs in weak positions (intervocalic, coda, clusters) and lengthens the preceding vowel [6]. Davidson [7] has shown that in Hawaiian in some cases an intervocalic glottal stop is realised – among other variants (e.g., creaky voice, full glottal closure) – as an intensity dip. The percentage of such cases was slightly greater when the flanking vowels were identical.

Recent studies on the South Estonian Leivu dialect have shown that in Leivu the loss of intervocalic short /h/ has led to the development of broken tone [8], [9]. Compared to plain tone words, broken tone words in Leivu are characterised by longer vowel duration (the first vowel is often lengthened), an intensity dip between two vowels, a more consistent early F0 turning point, and in some cases by a laryngealisation period during the second vowel. Loss of intervocalic short /h/ has also been considered to be one reason for the development of broken tone in Livonian [9], [10] and the South Estonian Lutsi dialect [11].

Teras [1] has made observations about tendencies in the pronunciation of words where /h/ has been lost: originally disyllabic quantity 1 (Q1) CVhV-words (e.g., *kohe* [koe̯ˑ] 'immediately') are pronounced either as monosyllabic quantity 3 (Q3) CVV-words, where the syllable nucleus is a long monophthong or diphthong (e.g., *kohe* [koeː]), or they are pronounced with a boundary so that the first syllable vowel lengthens (e.g., *kohe* [koː.e]). However, no closer acoustic analysis of such variants has yet been done and it is not known how frequently such variants occur and whether some other variants are possible.

The present study aims to determine what acoustically characterises the variants of CVhV-words where the loss of intervocalic short /h/ occurs. The research questions are as follows:

(1) What variants result from the loss of /h/ and what is their percentage?
(2) What characterises vowel duration and duration ratios in these variants?
(3) What can be said about the F0 contours of the variants?
(4) How often does the dip in intensity occur between two vowels?

2. MATERIALS AND METHODS

The Phonetic Corpus of Estonian Spontaneous Speech [12] was used for data retrieval. The corpus contained 77 hours of speech from 111 speakers at the time of data retrieval. The corpus is manually segmented both on the word level (annotated using Estonian orthography) and segmental level (annotated using SAMPA transcription). Disyllabic CVhV-words were searched for at the word level. The same material as in Teras [1] was used for acoustic analysis, but variants with consonant loss were reinspected and two misanalysed variants were left out.

In Estonian, there are 9 vowels (close [i y u], mid [e ø ɤ o], open [æ ɑ]) that occur in the primary stressed first syllable, and 5 of these [i u e o ɑ] occur in successive syllables. The second syllable /e/ is often pronounced lower: [ę] ~ [æ]. [13] The current material did not contain [ø] in the first syllable or [o] in the second syllable.

There are three quantity degrees in Estonian: words with short stressed syllables are Q1 and those with long stressed syllables are Q2 or Q3. All monosyllabic words are considered to be Q3, e.g., *maa* [mɑ::] 'land, earth', *kae* [kɑe:] 'cataract'. The three quantity degrees are differentiated by the duration ratios of syllables (2/3, 3/2, and 3/1, respectively), and Q2 and Q3 are additionally differentiated by the location of the F0 turning point (late or early). See, e.g., [13], [14]. CVhV-words are Q1 where both vowels are phonologically short. However, Q1 words are characterised by the lengthening of the second syllable vowel (V1/V2 ratio 0.5–0.8), e.g., *raha* [rɑhɑ̯·] 'money'.

Of all CVhV-words (1173), there were 250 tokens where a loss of /h/ occurred (see Table 1).

Table 1. The number of analysed tokens.

Variant	Number of tokens		
	No intensity dip	Intensity dip	Total
CVV	66	22	88
CV1V2	129	33	162
Total	195	55	250

These tokens were analysed acoustically using Praat [15]. The duration of all segments was measured. In the case of two different vowels, the boundary was set dividing the transition between the two vowels in half. When /h/ between two identical vowels was lost, a dip in intensity was used to mark the boundary.

Fundamental frequency and intensity were measured from the beginning of the first vowel and the end of the second vowel, and also from the turning point (TP) where the F0 or intensity started to fall. The location of the TP was calculated as a percent of the total duration of the vowels. If there was a dip in intensity during the vowels, then an intensity maximum of the first and the second vowel was measured as well as an intensity minimum during the dip. See an example of segmentation and annotation in Figure 1.

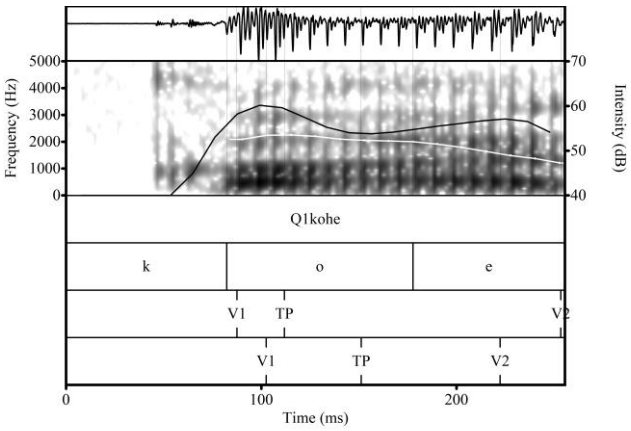


Figure 1. An example of segmentation and annotation of the word *kohe* [ko·.e] 'immediately' (a male speaker). A sound wave, spectrogram, intensity contour (solid black line), F0 contour (dotted white line, scale 60–180 Hz). The annotation shows four tiers: word, sounds, F0 (V1, V2 – the beginning and the end, TP – the turning point), and intensity (V1, V2 – maximum, TP - minimum) measurement points.

Statistical analysis was carried out in R [16]. Descriptive statistics included the mean and median values. Average F0 values were calculated separately for female and male speakers. The chi-square test was used to determine significant associations between variants and linguistic factors: accentuation (levels: accented, unaccented), V1 and V2 (levels: close, mid, open), utterance position (levels: initial, medial, final), and adjacent vowels (levels: different, identical).

3. RESULTS AND DISCUSSION

3.1. Variants and duration ratios

Intervocalic /h/ can be lost between two identical or two different vowels. It appears that the loss of intervocalic /h/ results in four variants. The average duration of segments in these variants is shown in

Figure 2 and the duration ratios of vowels in the three variants are given in Figure 3.

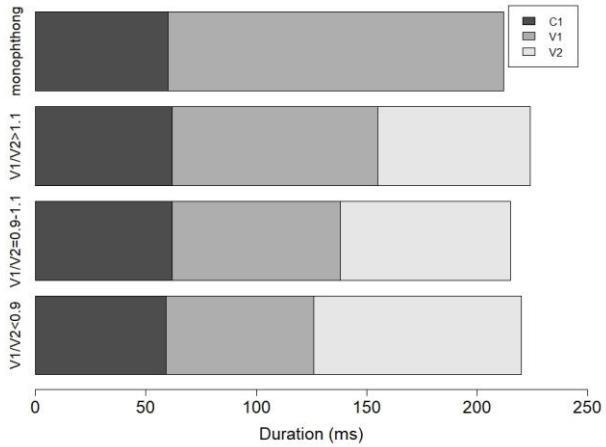


Figure 2. Sound durations (C1, V1, V2) for four variants of CVhV words where intervocalic /h/ has been lost.

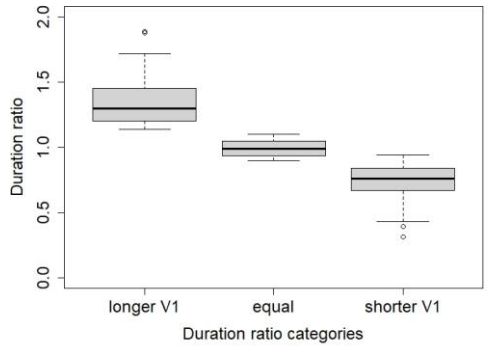


Figure 3. The duration ratio of V1/V2 for three variants: longer V1 than V2 (35 tokens), similar duration for V1 and V2 (48 tokens), shorter V1 than V2 (101 tokens).

The variants are as follows:

(1) two identical vowels are pronounced as a long monophthong (VV 152 ms);
(2) V1 is longer than V2 (V1 93 ms, V2 69 ms, V1/V2 = 1.35);
(3) two vowels have similar duration (V1 76 ms, V2 77 ms, V1/V2 = 0.99);
V1 is shorter than V2 (V1 67 ms, V2 94 ms, V1/V2 = 0.73).

Two identical vowels were pronounced in most cases as a long monophthong (66 tokens, 26% of all tokens). However, there were an additional 22 tokens with a dip in intensity between two identical vowels. These tokens were grouped into three variants.

The most frequent variant is the variant where V1 is shorter than V2 (41% of all tokens). This vowel duration ratio (0.73) is characteristic of Estonian Q1 words [14] but also of the vowel duration ratio for diphthongs in Q3 words where V2 tends to be longer than V1 [17], e.g., *lai* [lɑi:] 'broad'. It can be hypothesised that listeners would perceive these tokens as Q3 words.

In 14% of all tokens, the duration of V1 was longer than the duration of V2 (ratio: 1.35). Lengthening of the preceding vowel resulting from the loss of intervocalic /h/ has also been found in the South Estonian Leivu dialect [8] and in the Shirazi dialect of Persian [6]. In 19% of all tokens, V1 and V2 had quite similar durations (ratio: 0.99).

When the percentage of accented to unaccented tokens is similar (44% vs. 56%) for tokens with shorter V1 than V2, then there is a smaller percentage of accented than unaccented words (40% vs. 60%) for tokens with longer V1 than V2 and with a similar duration for V1 and V2. However, the chi-square test shows no significant association between the variants and accentuation or between other linguistic factors.

3.2. Pitch contours

Previous research has shown that Estonian accented Q1 and Q2 words are characterised by a late F0 TP in the stressed syllable followed by an F0 fall in the second syllable, whereas accented Q3 words have an early F0 TP followed by a fall already in the stressed syllable (e.g., [14]). Unaccented words do not have F0 TP and their contours are flat [14].

The current material contained 109 accented and 141 unaccented words. All accented words had a falling pitch contour. In unaccented words, the pitch contours were rather flat: on average there was only a slight fall from beginning to end (12–15 Hz). In 13 unaccented tokens, the pitch rose slightly (8–19 Hz).

The pitch contours of accented words will be analysed in more detail next. In Figure 4, the location of F0 TP can be seen. In accented words, in almost all cases (96 tokens, 88% of accented tokens) an early F0 TP occurred at 25% of the total vowel duration. In only 13 tokens (12% of accented tokens), the TP occurred late at 61% of the total vowel duration.

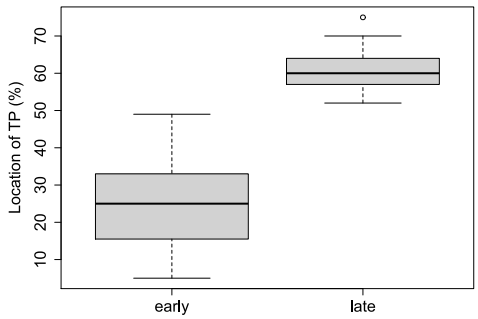


Figure 4. The location of the F0 turning point in accented words.

Figure 5 shows the average F0 contours of tokens where there was an early F0 TP (average values were calculated separately for female and male speakers). The F0 range from the beginning to the end is 32–36

Hz. It can be said that the F0 contours of these words are similar to Estonian Q3 words that are also characterised by an early F0 turning point [14].

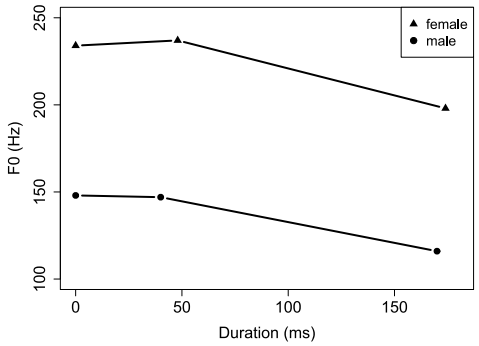


Figure 5. Average F0 contours of accented tokens with an early F0 turning point (points: the beginning of V1, TP, and end of V2).

3.3. Intensity

In the case when an intervocalic /h/ is present, there is a dip in intensity between adjacent vowels. When an intervocalic /h/ is lost, in most cases (78%, see table 1) there is no dip in intensity. In these tokens, there is in most cases (71%) an early intensity TP (occurring at 27%) and in other cases a late intensity TP (at 69%). Average intensity contours are presented in Figure 6.

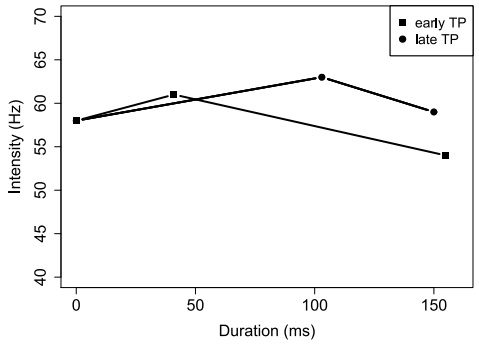


Figure 6. Average intensity contours of accented tokens with an early and late intensity turning point (points: the beginning of V1, TP, and end of V2).

In 22% of cases, there was a slight dip in intensity during two vowels. When /h/ is lost between two identical vowels, this percentage is greater (25%) than when it is lost between two different vowels (20%). The ratio of the intensity minimum (61 dB) and the quotient of V1 and the V2 intensity maximum (both 63 dB) was calculated for these tokens. The ratio varied between 0.92 and 0.99, the average ratio is 0.97. These values are similar to those found in Madrid Spanish where intervocalic /d/ was reduced [5]. Such a dip in intensity has also been found in Hawaiian when an intervocalic glottal stop is lenited, c.f. [7].

The chi-square test shows a significant association between the dip and accentuation with a small effect size ($\chi^2 = 4.67$, df = 1, $p < 0.05$, Cramer's V = 0.137), and between the dip and preceding vowel with a small effect size ($\chi^2 = 11.47$, df = 2, $p < 0.05$, Cramer's V = 0.214). Pearson's residuals suggest that the dip occurs more than expected in accented words and when preceded by an open vowel. No other significant associations were found.

4. CONCLUSION

Lenition of intervocalic obstruents is well-known in the world's languages. Its final stage is complete elision or loss. The loss of intervocalic short plosives and /h/ occurs in Estonian. When an intervocalic /h/ between two identical vowels is lost, in most cases the two vowels are pronounced as a long monophthong but in some cases, there is a dip in intensity between the vowels, and, in most cases, V2 is longer than V1. The latter is most characteristic also of variants where an intervocalic /h/ is lost between two different vowels. The ratio of V1/V2 in such variants is 0.73. Such a ratio also characterises diphthongs in Q3 words. However, the loss of /h/ can also lengthen the first syllable vowel. The two vowels can also have quite similar durations (V1/V2 ratio 0.99).

Accented words, where an intervocalic /h/ has been lost, are characterised by an early F0 TP, which is also characteristic of Estonian Q3 words. The intensity turning point is also early in most cases, but in 22% of analysed tokens, there was a slight dip in intensity between the two vowels (the ratio of the intensity minimum and two intensity maximums of vowels was 0.97). The dip in intensity occurs more in accented than in unaccented words and after an open vowel.

It can be concluded that loss of intervocalic /h/ can result in a prosodic change: originally disyllabic Q1 words are pronounced like monosyllabic Q3 words with open long syllables where the second vowel lengthens and F0 is falling. However, there are also variants where the loss of /h/ lengthens the first vowel, resulting in a different pattern of duration ratios for the two adjacent vowels and most likely indicating that there can still be a syllable boundary between the two vowels despite the loss of /h/. The dip in intensity that was found in some tokens also seems to refer to such a boundary.

5. ACKNOWLEDGEMENTS

This research has been supported by the University of Tartu, grant PHVEE19903. The author would like to thank the two anonymous reviewers for their comments on and suggestions for this article, and Uldis Balodis for his English-language editing.

6. REFERENCES

[1] P. Teras, 'The phonetic variation of short intervocalic /h/ in Estonian', in *Speech Prosody 2018*, ISCA, 2018, pp. 883–887. doi: 10.21437/SpeechProsody.2018-178.

[2] L. Ermus, 'Eesti keele lühikeste klusiilide häälduse variatsioon ja seda mõjutavad tegurid', *Mäetagused*, vol. 69, pp. 27–52, 2017, doi: 10.7592/MT2017.68.ermus.

[3] R. Kirchner, 'Consonant lenition', in *Phonetically Based Phonology*, B. Hayes, D. Steriade, and R. Kirchner, Eds., Cambridge: Cambridge University Press, 2004, pp. 313–345. doi: 10.1017/CBO9780511486401.010.

[4] M. Garellek, Y. Chai, Y. Huang, and M. Van Doren, 'Voicing of glottal consonants and non-modal vowels', *J. Int. Phon. Assoc.*, pp. 1–28, 2021, doi: 10.1017/S0025100321000116.

[5] F. Melero García, 'Variación sociofonética: comparación de análisis graduales y categóricos', *Cuad. Lingüíst. Hispánica*, no. 37, 2021, doi: 10.19053/0121053x.n37.2021.12184.

[6] N. Rashidi and M. Shokrollahi, 'Elision of /h, ?/ in the Shirazi Dialect of Persian (SHDP): An Optimality Theory Based Analysis', *Raspr. Instituta Za Hrvat. Jez. Jezikoslovlje*, vol. 36, no. 1, pp. 135–162, 2010.

[7] L. Davidson, 'Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian', *J. Phon.*, vol. 88, p. 101075, Sep. 2021, doi: 10.1016/j.wocn.2021.101075.

[8] P. Teras, 'Broken tone in Leivu CV'V-words', *J. Est. Finno-Ugric Linguist.*, vol. 12, no. 2, pp. 169–190, 2021, doi: 10.12697/jeful.2021.12.2.07.

[9] P. Teras and T. Tuisk, 'Broken tone in Livonian and Leivu', in *1st International Conference on Tone and Intonation (TAI)*, ISCA, 2021, pp. 26–30. doi: 10.21437/TAI.2021-6.

[10] T.-R. Viitso, 'Livonian and Leivu: Shared innovations and problems', *Linguist. Ural.*, vol. 45, no. 4, pp. 269–282, Dec. 2009, doi: 10.3176/lu.2009.4.03.

[11] U. Balodis, K. Pajusalu, and P. Teras, 'Broken tone in South Estonian dialects in Latvia', vol. 24, pp. 98–114, 2016.

[12] P. Lippus, K. Aare, A. Malmi, T. Tuisk, and P. Teras, 'Phonetic Corpus of Estonian Spontaneous Speech v1.2', 2021, doi: 10.23673/RE-293.

[13] E. L. Asu and P. Teras, 'Estonian', *J. Int. Phon. Assoc.*, vol. 39, no. 3, pp. 367–372, 2009, doi: 10.1017/S002510030999017X.

[14] P. Lippus, E. L. Asu, P. Teras, and T. Tuisk, 'Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian', *J. Phon.*, vol. 41, no. 1, pp. 17–28, 2013, doi: 10.1016/j.wocn.2012.09.005.

[15] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer [Computer program].' 2022. Accessed: Dec. 17, 2022. [Online]. Available: http://www.praat.org/

[16] R Core Team, 'R: A Language and Environment for Statistical Computing'. Vienna, Austria, 2022. [R Foundation for Statistical Computing]. Available: https://www.R-project.org/

[17] P. Lippus and E. L. Asu, 'The timing and quality of diphthong components in spontaneous Estonian', in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., Canberra, Australia: Australasian Speech Science and Technology Association Inc, 2019, pp. 1139–1143.

PHONETIC VARIATION OF WORD-FINAL AFFRICATES IN ROMANIAN CONNECTED SPEECH

Oana Niculescu[1] & Andrei A. Avram[2]

[1]Romanian Academy Institute of Linguistics "Iorgu Iordan – Al. Rosetti", Bucharest, Romania
[2]University of Bucharest, Department of English, Romania
oeniculescu@yahoo.com; andrei.avram@lls.unibuc.ro

ABSTRACT

Studies dealing with connected speech phenomena on Romanian data are rather scarce, mainly because of the lack of available corpora suited for analyses at the interface between phonetics and phonology. Consequently, we investigate an understudied topic related to phonetic variation of word-final affricates based on a recently developed open-access standard Romanian speech corpus.

A total of 686 tokens were manually extracted from 4hs of speech pertaining to 4 native speakers (2 female, 2 male). In terms of reduction processes, our results show that the affricates undergo coarticulation in 52% of the data. Fricativization is the preferred outcome (78%), followed by total assimilation (13%), deletion (8%), and (de)voicing (0.5% each), with male speakers showing a higher degree of coarticulation than female speakers ($\chi^2(1) = 16.1$, $p < 0.001$).

Keywords: Romanian affricates, phonetic variation, Romanian spontaneous speech, corpus linguistics.

1. INTRODUCTION

Among European languages, Romanian is classified as an under-resourced language [1]. In terms of the few topics explored having a speech technology focus, we would like to mention that of sound change in synchronic variation [2], lenition and fortition of stop codas [3], as well as the deletion of the definite article *-l* [4] and vowel hiatus reduction processes in controlled and spontaneous speech [5, 6]. Nevertheless, studies dealing with connected speech phenomena on Romanian data are still rare.

Through this research paper we aim to fill in this gap by investigating an understudied topic related to phonetic variation of word-final affricates. The analysis is carried out on a recently developed open-access standard Romanian speech corpus. We will be examining coarticulation processes conditioned by the phonological context, segment duration, voicing degree, as well as the gender of the speakers.

Our analysis centres around the three affricates present in contemporary standard Romanian:

- voiceless dental /ʦ/ (orthographic *ţ*)
  baţi 'to beat' /batsj/ [batsʲ]
- voiceless palatal /ʧ/(orthographic *c* + *e/i*)
  baci 'shepherd' /baʧj/ [baʧʲ]
- voiced palatal /ʤ/ (orthographic g + e/i)
  bagi 'put in$_{2SG}$' /baʤj/ [baʤʲ]

The voiced dental affricate /ʣ/ appears only in certain Daco-Romanian subdialects, being spread mainly in the NW (Maramureş) and SW (Banat) areas of the country [see 7, 8].

It is important to point out that, especially in word-final position, Romanian is the only Romance language to display secondary palatalization [5, p.545]. All consonants can be affected by it. In this context, Romanian /i/ does not always surface as a vowel. Under certain morpho-phonological circumstances, such as plural forms of nouns, adjectives, and certain verb forms (*face* [faʧe] 'do$_{3SG}$' – *faci* [faʧʲ] 'do$_{2SG}$'), /i/ is reduced to a palatal gesture [ʲ] [9: 537]. This morpheme, in turn, can trigger a change in PoA, assibilation or coronalization in the preceding alveolar or velar obstruent (*vag* [vag] 'vague$_{ADJ.M.SG}$' – *vagi* [vaʤʲ]'vague$_{ADJ.M./F.PL}$') [10]. Among the three affricates, only the voiceless dental /ʦ/ can occur in absolute word-final position (*hoţ* [hoʦ] 'thief' – *hoţi* [hoʦʲ] 'thief$_{PL}$'). Even though we focus on the subset of word-final affricates, these consonants can also appear in initial and medial position in the word (for a phonotactic description, see [11]).

The first studies dealing with phonetic reduction of word-final affricates in Romanian dialectal texts revealed that fricativization is the most frequent phenomena (307 cases), followed by deletion (60 cases) and (de)voicing (16 cases) [12]. The results also showed that the voiced palatal affricate did not undergo devoicing or deletion [12: 102]. The analyses later continued with pioneer acoustic studies carried out on a Kay Sonagraph investigating the overall duration of the palatal [13, 14] and dental affricates [15] in contrast to the duration of the corresponding plosive + fricative cluster duration. The data were derived from monosyllabic words ("logatoms") with the affricates in intervocalic position. The clusters scored an average duration of 270ms, 261ms and 212ms, respectively, while the affricates had a mean duration of 191ms /ʧ/, 185ms /ʦ/, and 150ms /ʤ/.

This present study complements previous research on the topic by extending the analysis toward

connected speech data. To our knowledge, this would be the first phonetic account of word-final Romanian affricates documented on naturalistic data.

2. DATA AND METHODOLOGY

2.1. Naturalistic data

This analysis is based on 4 hours of addressed monologues pertaining to 4 adult native speakers, 2 male (136 minutes) and 2 female (17 minutes), ages 33 to 37 (mean = 35.3, SD = 1.7). Participants share the same socio-economic background and are representative of the Southern dialect on which the standard language is based. The data are extracted from a larger Romanian speech corpus (12 participants, 6M/6F), *Ro-Phon*, developed and transcribed by the main author as part of her postdoctoral research project (2020 – 2022) [16]. The corpus was designed for eliciting both spontaneous speech (through a monologue task), as well as controlled speech (by means of a reading task, token words placed in carrier sentences), with recordings carried out in a sound-attenuated room, using a stand microphone connected to a laptop via an external audio interface. All recordings on connected speech data share the same three conversational topics related to past activities (memories from childhood, life lessons), present pursuits (activities, likes and dislikes, travelling), and future projects (personal as well as professional), making the data comparable. Both the audio files and the corresponding TextGrids containing the ortho- graphic and phonological transcriptions are available, alongside the necessary metadata for each speaker.

2.2. Variables studied

We focus our analysis on word-final affricate reduction processes in connected speech related to:

- fricativization /faʧj kum/ [faʃ.kum]
  /niʧj məkar/ [niʒ.mə.kar]
- total assimilation /poʦj sə/ [pos.sə]–[po.sə]
- deletion /altsj ʒukətorj/ [al.ʒu.kə.torj]
- voicing /faʧj via/ [faʤ.vi.ja]
- devoicing /aʒunʤj#/ [a.ʒunʧ]

It is important to note that these coarticulatory processes are preceded by the deletion of the word-final glide.

In relation to the phonological context, so as to make the results comparable with other studies on Romance data [17], we sorted the tokens into 5 categories depending on the following phonological context, namely when the next word begins with a: *voiceless obstruent* /ɨnvəʦ pe/, *voiced obstruent* /poʦʲ da/, *sonorant* /niʧʲ nu/, *vowel* /merʤʲ ɨn/, or *pause* (silence or breath) /minʤj#/.

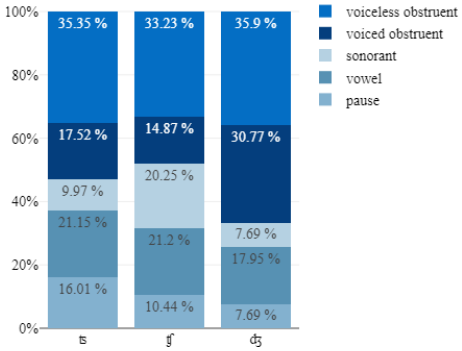


Figure 1: Percentage of word-final affricates (in group) as a function of the following context

We observe that the data are not evenly distributed across the five contexts, with voiceless obstruents having a frequency of 34% (N = 236), vowels 21% (N = 144), voiced obstruents 17% (N = 117), sonorants 15% (N = 100), and pause 13% (N = 89). Based on the phonological context, we expect to find a lower frequency of coarticulatory processes when the affricates are followed by a vowel, as opposed to a higher degree of coarticulation before word-initial obstruents.

For measuring the proportion of voicing in the constriction, we used the Voice Report in Praat [18] showcasing the fraction of locally unvoiced frames. Since the voicing ratios displayed a gradient pattern along the acoustic continuum cuing the voice contrast (Fig. 2), we opted for a ternary classification of voicing degree (similar to [19]): *voiced outputs* (when more than 90% of the interval was identified as voiced by the Voice Report), *partially voiced outputs* (between 10% and 90% of the interval was voiced), and *unvoiced outputs* (less than 10% of the interval was designated as voiced by the Voice Report in Praat).

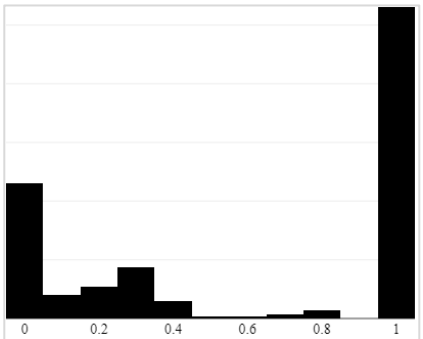


Figure 2: Voicing ratios distribution of word-final affricates undergoing fricativization

The skewness of the data can be, in part, attributed to the higher frequency of voiced segments following the word-final affricates (53%, N = 361), as opposed to 47% (N = 325) corresponding to voiceless

segments. Another explanation could be that all voiced outputs, representing 53% of the data (N = 147), have a voicing degree of 100%, with no tokens present between 90% and 99%. Unvoiced tokens have a frequency of 24% (N = 67), whereas partially voiced outputs are found in 23% of the cases (N = 63).

In relation to duration measurements (i) we predict to find a shorter duration, on average, of the voiced affricate in comparison to the voiceless affricates; (ii) depending on the context, we expect a longer duration of the outputs in pre-pausal context as a result of word-final lengthening, and (iii) in terms of voicing ratios, our predictions are that unvoiced segments will have the longest duration, followed by partially voiced and fully voiced outputs.

In this study we also wanted to test the generalizations related to male and female speech patterns in connection to reduced pronunciation variants characteristic of casual speech. Since all speakers have received higher education and the monologues produced are representative of the standard language, we expect to find an increased percentage of standard pronunciation variants.

3. RESULTS AND DISCUSSION

A total of 686 tokens were manually extracted from 4hs of speech pertaining to 4 Romanian native speakers (2 female, 2 male). The voiceless affricates have the highest frequency (/ʦ/ 48%, N = 331, with just 25 occurrences in absolute word-final position; /ʧ/ 46%, N = 316), while the voiced affricate /ʤ/ only surfaces in 39 contexts, representing 6% of the data.

In our analysis, affricates undergo coarticulation in 52% of the cases. Fricativization is the preferred outcome (78%), followed by total assimilation (13%), deletion (8%), and (de)voicing (0.5% each).

When maintained, the average duration of an affricate is as follows: /ʦ/ 166ms (SD = 101), /ʧ/ 152ms (SD = 90), and /ʤ/ 96ms (SD = 50). As expected, the consonants display a shorter average duration in connected speech compared to previous results obtained from read speech data [13, 14, 15].

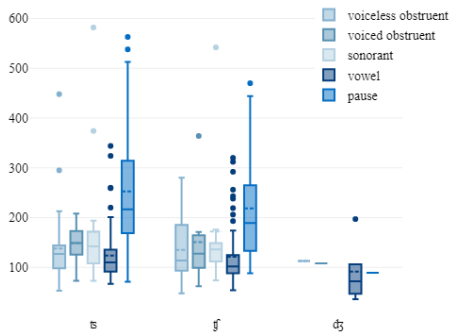


Figure 3: Duration of word-final affricates as a function of the following context

Depending on the phonological context (Fig. 3), we observe a pre-pausal lengthening effect, especially in the case of the voiceless affricates (t-test scores revealed a significant increase in the overall duration of these affricates in the presence of neighbouring pauses, namely for /ʧ/: t = 4.3, df = 36, p < 0.001; for /ʦ/: t = 2.319, df = 27, p = 0.028). On average, the lowest durations are found when the affricates are followed by a word-initial vowel or voiceless obstruent. The mean duration of each consonant as a function of the following phonological context is presented in Table 1.

	ʦ	ʧ	ʤ
voiceless obstruent	138 (68)	135 (63)	113
voiced obstruent	149 (40)	151 (93)	108
sonorant	171 (126)	172 (142)	-
vowel	124 (54)	121 (60)	92 (65)
pause	253 (122)	218 (104)	89

Table 1: Mean duration (SD) of affricates as a function of the following context

Word-final affricates undergo fricativization in various degrees. Among the three affricates, /ʧ/ has the highest rate of fricativization (54%, N = 151), followed by /ʦ/ (38%, N = 105). The voiced palatal affricate undergoes fricativization in only 8% of the cases (N = 21). These observations are in line with previous studied [13, 14, 15].

In terms of phonological context, fricativization occurs especially before an obstruent (39%, N = 107, in the case of voiceless obstruents, and 33%, N = 91, in the case of voiced obstruents). This coarticulation phenomena also occurs, to a lesser degree, before a sonorant (26%, N = 71), a vowel or a pause (1% each).

As predicted, on the basis of duration measurements and degree of vocal fold vibration, the average durations of the resulting fricatives are: unvoiced 71ms (SD = 42), partially voiced 57ms (SD = 23), and voiced 48ms (SD = 13). A t-test indicated

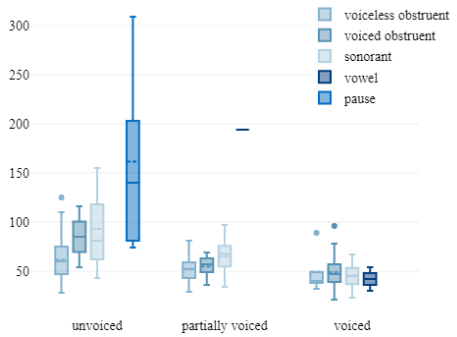


Figure 4: Duration of word-final affricates undergoing fricativization as a function of the following context

a significant decrease in duration from unvoiced outputs to partially voiced ($t = 2.392$, $df = 102$, $p = 0.018$) and, subsequently, voiced fricatives ($t = 3.108$, $df = 81$, $p = 0.002$).

As shown in Fig. 4 (and detailed in Table 2), the highest variation is found in the case of unvoiced fricatives in pre-pausal context (ranging from 74ms to 309ms, with a standard deviation of 98ms). Voiced fricatives in intervocalic position have the shortest duration.

	unvoiced	partially voiced	voiced
voiceless obstruent	62 (21)	52 (12)	50 (23)
voiced obstruent	85 (44)	55 (13)	49 (14)
sonorant	93 (57)	65 (18)	46 (11)
vowel	-	194	42 (17)
pause	161 (98)	-	-

Table 2: Mean duration (SD) of word-final affricates undergoing fricativization as a function of the following context

The second coarticulation process in terms of frequency involving word-final affricates is total assimilation. The process is highly dependent on the following context (see correlation results in Table 3). Out of the three affricates, the dental affricate is frequently targeted (85%, N = 38). The palatal affricates undergo this reduction process to a lesser degree, /ʧ/ 13% and /ʤ/ 2%.

Only 29 tokens exhibit deletion of the word-final affricate, especially before a voiceless obstruent (76%, N = 22). There are only 3 cases in which this reduction process occurs before a sonorant. The voiceless palatal affricate has the highest rate of reduction (66%, N = 19), while /ʤ/ (24%, N = 7) and /ʦ/ (10%, N = 3) are deleted exclusively when followed by a voiceless obstruent.

Devoicing occurs only in a pre-pausal context, being visible during the frication part of the affricate, while voicing was found exclusively when the voiceless palatal affricate was followed by /v/, resulting in a fully voiced output. We found only two tokens for each of the aforementioned coarticulation phenomena in the corpus. Previous results based on dialectal data [12] indicated a higher rate of deletion, especially targeting /ʧ/, and that both voiceless affricates underwent voicing.

An ANOVA test was used for each coarticulation process in correlation to the phonological context. The results show a strong correlation between the two variables particularly in the case of affricate maintenance (for all of the five contexts) on the one hand, and, on the other hand, in relation to the highest frequent coarticulation processes, namely fricativization and total assimilation (results summarised in the table below).

	maintenance	fricativization	total assimilation	deletion	voicing	devoicing
voiceless obstruent	77.242 ($p < 0.001$)	3.685 ($p = 0.0553$)	105.712 ($p < 0.001$)	23.797 ($p < 0.001$)	1.050 ($p = 0.3058$)	1.050 ($p = 0.3058$)
voiced obstruent	59.430 ($p < 0.001$)	92.799 ($p < 0.001$)	10.018 ($p = 0.0016$)	0.227 ($p = 0.6338$)	9.867 ($p = 0.0018$)	0.411 ($p = 0.5214$)
sonorant	23.955 ($p < 0.001$)	48.698 ($p < 0.001$)	8.294 ($p = 0.0041$)	0.435 ($p = 0.5100$)	0.341 ($p = 0.5592$)	0.341 ($p = 0.5592$)
vowel	243.414 ($p < 0.001$)	132.081 ($p < 0.001$)	13.000 ($p = 0.0003$)	8.117 ($p = 0.0045$)	0.532 ($p = 0.4661$)	0.532 ($p = 0.4661$)
pause	88.864 ($p < 0.001$)	55.320 ($p < 0.001$)	7.234 ($p = 0.0073$)	4.531 ($p = 0.0336$)	0.298 ($p = 0.5851$)	13.684 ($p = 0.0002$)

Table 3: Correlation (F) values and significance (p) from t-test; the grey cells highlight there is no significant correlation between the variables

When looking at the coarticulated outputs as a function of gender, we observe that male speakers show a higher degree of coarticulation than female speakers ($\chi^2(1) = 16.1$, $p = < 0.001$). The tendencies for each reduction process (male monologues vs female monologues) can be summarised as follows: 80% (N = 222) – 20% (N = 55) in the case of fricativization, 84% (N = 38) – 16% (N = 7) when the affricates undergo total assimilation, and 76% (N = 22) as opposed to 24% (N = 7) in the case of deletion. However, since the results are dependent on speakers from only one social class and age group, we acknowledge the limitations of our current study. In future socio-phonetic analyses, we consider it relevant to include speakers from various backgrounds.

Summing up, based on our data, we observed that the dental affricate /ʦ/ undergoes more coarticulation processes than /ʧ/ ($\chi^2(1) = 9.66$, $p = 0.002$), which, in turn, coarticulates more than /ʤ/ ($\chi^2(1) = 7.69$, $p = 0.006$).

5. CONCLUSIONS

This study has investigated coarticulatory processes related to word-final affricates in Romanian connected speech based on a recently developed oral corpus.

Analysing the way in which reduction processes occur in connected speech can benefit linguistics and automatic speech recognition models alike. The data acquired through this study can also contribute to the ongoing phonological investigation of fortition and lenition processes occurring in Romance languages.

With respect to future research on the topic, we plan to examine a larger sample of recorded data from the *Ro-Phon* corpus, both in terms of controlled and spontaneous speech, thus enabling an in-depth cross-linguistic analyses of affricates in Romance languages.

6. REFERENCES

[1] Trandabăț, D., Irimia, E., Mititelu, V. B., Cristea, D., Tufiș, D. 2012. *The Romanian language in the digital age*. In: Rehm, G., Uszkoreit, H. (eds), *META-NET White Paper Studies*. Berlin: Springer, 1–87.

[2] Chitoran, I., Vasilescu, I., Lamel, L., Vieru, B. 2018. Connected speech in Romanian. Exploring sound change through an ASR system. In: Recasens, D., Sánchez Miret, F. (eds), *Production and perception mechanisms of sound change*. Munich: Lincom, 129–143.

[3] Hutin, M., Niculescu, O., Vasilescu, I., Lamel, L., Adda-Decker, M. 2020. Lenition and fortition of stop codas in Romanian. *Proc. SLTU-CCURL* Marseille, 226–234.

[4] Vasilescu, I., Chitoran, I., Vieru, B., Adda-Decker, M., Candea, M., Lamel, L., Niculescu, O. 2019. Studying variation in Romanian: deletion of the definite article *-l* in continuous speech. *Linguistics Vanguard* 5(1), 1–12.

[5] Niculescu, O. 2018. *Hiatul intern şi hiatul extern în limba română contemporană. O analiză acustică* [Internal and external hiatus in contemporary standard Romanian. An acoustic analysis], PhD dissertation, University of Bucharest.

[6] Niculescu, O., Vasilescu, I., Adda-Decker, M. 2019. Duration Patterns of Internal and External Hiatus as a Function of Speaking Style. *Revue roumaine de linguistique* 64 (2), 189–197.

[7] Mărdărescu-Teodorescu, M. 1993. Structura acustică a africatei [d] în graiurile dacoromâne [The acoustic structure of the affricate [d] in Romanian dialects]. *Fonetică şi dialectologie* 12, 113-116.

[8] Rusu, V. 1984. *Tratat de dialectologie română* [Romanian dialectology treatise], Craiova, Scrisul Românesc.

[9] Renwick, M. 2021. Romanian. In: Meisenburg, T., Gabriel, C., Gess, R. (eds) *Manual of Romance Phonetics and Phonology*. Berlin: de Gruyter, 531–558.

[10] Spinu, L. 2010. *Palatalization in Romanian. Experimental and Theoretical Approaches*, PhD dissertation, University of Delaware.

[11] Roceric-Alexandrescu, A. 1968. *Fonostatistica limbii române* [Romanian Phonostatistics], Bucharest, Editura Academiei Republicii Socialiste România.

[12] Mărdărescu, M. 1975. Semioclusivele româneşti în fonetică sintactică (în sandhi) [Romanian affricates in syntactic phonetics (in sandhi)]. *Fonetică şi dialectologie* 9, 99–106.

[13] Mărdărescu, M. 1969. Durata semioclusivelor prepalatale în limba română literară [Duration of prepalatal affricates in standard Romanian]. *Fonetică şi dialectologie* 6, 59–70.

[14] Mărdărescu-Teodorescu, M. 1976. The duration of prepalatal affricates in Romanian. Auditory tests. *Revue roumaine de linguistique* 21 (1), 99–103.

[15] Mărdărescu-Teodorescu, M. 1993. Africata dentală [ţ]. Observaţii acustice [The dental affricate [ţ]. Acoustic observations]. *Fonetică şi dialectologie* 12, 33–43.

[16] Niculescu, O. 2021. Developing linguistic resources for Romanian written and spoken language, In: Rebreja, P., Onofrei, M., Cristea, D., Tufiș, D. (eds) *Proceedings of the 16th International Conference „Linguistic Resources and Tools for Natural Language Processing”*, Editura Universității „Alexandru Ioan Cuza”, Iași, 21–36.

[17] Hutin, M., Wu, Y., Jatteau, A., Vasilescu, I., Lamel, L., Adda-Decker, M. 2021. Synchronic Fortition in Five Romance Languages? A Large Corpus-Based Study of Word-Initial Devoicing. *Proc. Interspeech 2021*, Brno, 996–1000.

[18] Boersma, P. 2001. Praat, a system for doing phonetics by computer, *Glot International* 5 (9), 341–345.

[19] Davidson, L. 2015. Patterns of voicing in American English voiced obstruents in connected speech. *Proc. 18th ICPhS* Glasgow, paper number 84, https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0084.pdf [accessed 10.11.2022].

PREQUEL: SUPERVISED PHONETIC APPROACHES TO ANALYSES OF GREAT APE QUASI-VOWELS

Axel G. Ekström[1], Steven Moran[2,3], Johan Sundberg[1,4], Adriano R. Lameira[5]

[1]KTH Royal Institute of Technology, [2]University of Neuchâtel, [3]University of Miami,
[4]University College of Music Education Stockholm, [5]University of Warwick
axeleks@kth.se, steven.moran@unine.ch, jsu@kth.se, adriano.lameira@warwick.ac.uk

ABSTRACT

There is renewed interest in potential vowel production by nonhuman primates, but no agreed-upon methodologies for its estimation from real-life vocalizations. Here, we present a set of supervised approaches for estimating primate vowel-like articulation, with reference to orangutan long call pulses (N=36). We summarize our approach as a cohesive framework, the Primate Quasi-Vowel (PREQUEL) protocol. We *(1)* estimated f_0 from correlograms, *(2)* and vocal tract resonances (formants) from spectrograms, *(3)* the results of which were then compared against synthesized vowels for those frequency values; and *(4)* presented to uninformed listeners (N=16), who largely agreed on the categorization of vowel-like qualities for vocalizations (Cronbach's alpha=.701). We also provide descriptions of methods that are seemingly inadequate for formant estimation in great ape calls. We argue that a combination of phonetic methods is required to develop a science of nonhuman primate articulation.

Keywords: Primatology, Speech acoustics, Vowel space, Evolution of speech, Bioacoustics.

1. INTRODUCTION

Vowels are central to all of the world's spoken languages. Because aspects of human vocal anatomy is evolutionarily old and shared with primates and other mammals [1, 2, 3, 4], it is conceivable that comparative analyses of voiced calls by nonhuman great apes (hereafter great apes) – our closest living relatives – could inform knowledge of evolutionary origins of human vowel production capacities [5, 6, 7, 8]. Empirically, this approach has, however, grappled with major pitfalls, delaying insight towards a clearer picture of speech evolution. These include ethical and practical difficulties (e.g., apes cannot report their experience, or undergo imaging technology methods without some level of anesthesia), which may be circumvented non-invasively by approaching and analyzing great ape voiced calls as "acoustic casts" of vocal tract shape. However, because of the highly variable fundamental frequencies (f_0) characterizing various great ape calls, harmonic partials of f_0 may easily be mistaken for resonances [9] (formants) when estimations are calculated via unsupervised methods. Here, we present in detail a new approach and showcase its potential by confidently identifying [u]-like vocalizations by wild orangutans. We dub our methodological approach the primate quasi vowel (PREQUEL) protocol.

2. DATA

For the present study, we analyzed six long call pulses produced by each of six flanged male orangutans in the wild (N=36) [10] (duration M=.81s). Loud (long-distance) calls, including orangutan long calls and chimpanzee pant hoots, are a shared trait between primates that may date to the last common ancestor of primate species (around 65 Mya). In addition, due to their conspicuousness, they tend to be the best studied components of primate call repertoires. Our approach can, thus, potentially be extended to a variety of primate species. Among great apes, orangutan long calls are particularly slow paced (compared to e.g., chimpanzee pant hoots) and produced solitarily (i.e., without interference from other individuals' calls), and thus constitute a strong candidate for researchers seeking to develop methods of analysis of ape vowel-like production. We adopt the terminology of phonetics, referring to apparent vocal tract resonances as formants.

3. FUNDAMENTAL FREQUENCY

Orangutan long calls typically consist of three acoustically distinguishable phases, *(1)* the build-up grumbling phase, *(2)* climax pulses and *(3)* the bubbling tail-off. Because of apparent aperiodicity in *(1)* and *(3)*, only *(2)* was selected for acoustic

analysis.

Fundamental frequency (f_0) reflects the rate of vocal fold oscillation and corresponds to pitch in perception. In our orangutan data, selected pulses were subject to manual pitch estimation using waveform-matching correlograms [10] and verified with syntheses from the *Madde* additive voice synthesizer software [11]. This procedure was employed because high-f_0 signals are widespread in primate vocalization [12, 13] and therefore present a challenge to phonetic analyses, as the risk of biased estimates increases with f_0. Average f_0 of analyzed pulses (N=36) was 296.38Hz (SD=59.41, Min=172.76, Max=378.3), roughly corresponding to a high-pitched human voice, and within the range for a human male tenor singer. Thus, while significantly lower than those those observed in e.g., chimpanzee screams, values observed still raised the possibility of biased formant estimations, if using hand annotation or linear prediction.

4. FORMANT ESTIMATION

Problems with formant estimations are known from acoustic analyses of human speech, where $f_0 >$ 300 may result in formant estimation errors of $\pm$ 60 Hz [14]. To cancel out any such biasing effect of f_0, we initially attempted to estimate formant values by applying inverse filtering (in the *Sopran* software [11]) to all segments, resampled at 16 kHz (to smooth the spectral curve) – the main effect of which is that it allows analysis of components up to 8 kHz, well above the highest relevant spectrum components.

In speech audio analysis, inverse filtering procedures allow for the cancelling out of effects of formants on the radiated sound (the sum of voice source and vocal tract resonances). Two criteria are applied for tuning the inverse filters, each of which corresponds to a formant, a ripple-free closed phase in the waveform, and a source spectrum envelope void of local peaks and troughs near the formants (e.g., [15]). For our orangutan data, however, because inverse filtering assumes signal periodicity, satisfactory accuracy in tuning of inverse filters was not possible. Additionally, we observed that orangutan calls are breathy to the ear, possibly reflecting incomplete glottal closure in the vibration. The present work is exploratory; these initial observations indicate that a different approach may be necessary to analyze high-frequency ape calls (although higher-quality recordings may also help resolve this issue).

Instead, first and second formant frequencies (F_1, F_2) were annotated by hand; higher resonances were typically unavailable due to low signal-to-noise ratios for frequencies $>$ 2 kHz. In analyses of human voices, narrow-band analyses (45 Hz) are typically applied to reveal harmonic structure, while wide-band analyses (300 Hz) are generally used to reveal formant structures. This is so because narrow-bandwidth analysis admits one voice harmonic at a time, while wider passbands admit multiple voice harmonics at a time, rendering formants as remaining energy peaks (Fig. 1) (although higher-pitched voices may require spectrogram bandwidths at $\sim$600 Hz for intelligible rendering of formants [16, 17]). Because our data was collected in the wild, and includes background noise at higher frequencies, spectrogram bandwidths were set to 300 Hz. F_1-F_2 dispersions and were found to largely overlap with approximate human vowels [u] and [ʊ], with M=338 (SD=55.59) for F_1, and M=969 (SD=163.56) for F_2 (Fig. 2) (Tab. 1). Next, to validate estimated formant values, we sought to investigate whether the calls were indeed perceived as appropriate vowel qualities by human listeners.

Vowel quality	F_1(Hz)	F_2(Hz)
Orangutan long call	338	969
[u]	310	870
[ʊ]	450	1030
[ɔ]	590	880
[o]	360	640
[ɤ]	460	1310
[ɯ]	300	1390

Table 1: Average formant values observed for long calls, and approximate formant values for a set of adjacent (in apparent vowel quality) human vowels (male speaker) [18, 19].

To diminish the risk of mistaking harmonic partials of the f_0 for formant frequencies [9, 14], each candidate formant identified in the previous step was compared with frequencies of f_0 harmonic partials. Values were compared with output from the *Madde* software: each set of F_1-F_2 dispersions were input into *Madde* (set to exclude additional formants, and matched for pitch), to ensure that annotated quasi vowel qualities were adequate. Note that this method assumes that segment f_0 is sufficiently low that it is does not interfere with vowel perception. Next, we sought to investigate whether the identified [u]-like vocalizations were perceived as such by uninformed human listeners.

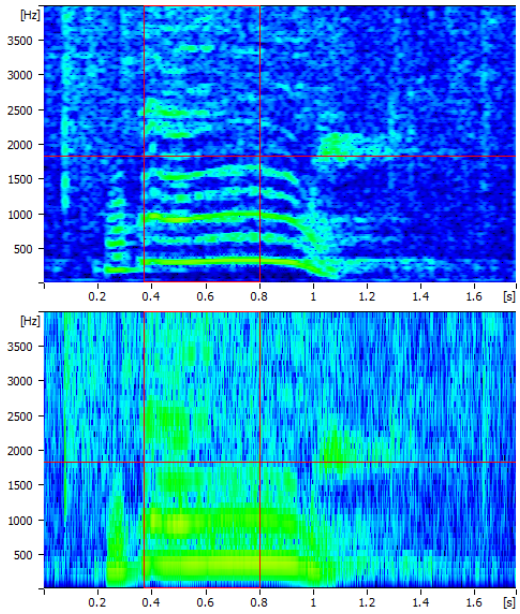


Figure 1: An orangutan long call [10], rendered as a narrow-band spectrogram (above, 45 Hz) with visible harmonic partials of the f_0; and broad-band spectrogram (below, 300 Hz), visibly dissolving partials. Spectrograms were rendered in the *Sopran* software. Length of highlighted segment is 42ms.

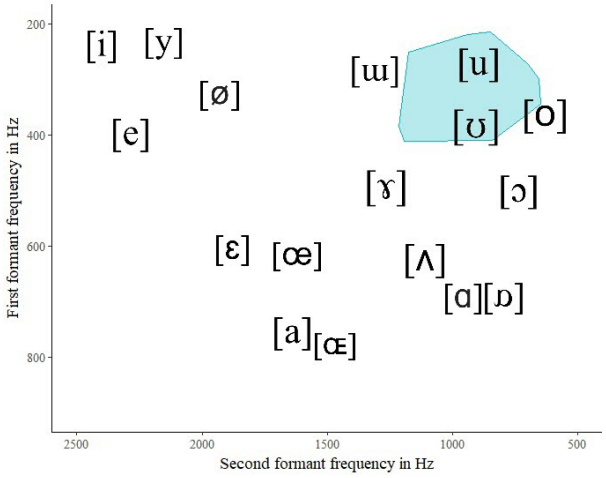


Figure 2: Orangutan [u]-like space. For comparison, average human vowel ranges (adult male speaker) are superimposed in brackets.

5. LISTENING AND SORTING TASKS

Participants (N=16, 9 women and 7 men) were presented with a visual listening sorting task using the visor software [20]. In total, participants listened to two subsets of 8 calls (randomly selected from the larger set) across two presentation sessions (Fig. 3). In the first session, participants were asked

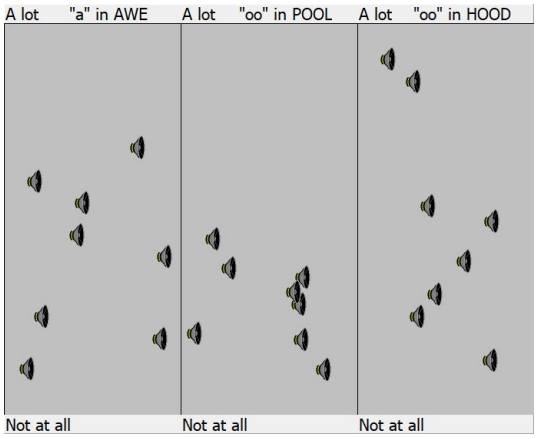


Figure 3: Listening task paradigm, set up in *Visor* software [11]. Files are repeated across each category. Participants freely listen to each sound by clicking and sort them by dragging. Starting positions are random.

to rate each of the eight calls on two dimensions, "sounds like AAH" and "sounds like OO". In the second, they were asked to rate each of 8 (different) calls on three dimensions, where the dimensions corresponded to the degree to which each sound, "sounds like 'a' in AWE [ɔ] / 'oo' in POOL [u] / 'oo' in HOOD [ʊ]". The origin of the recordings and true purpose of the ratings were withheld from transcribers until after the task was concluded. Listeners chose freely how many times to listen to each sound; on average, segments were played 2.2 times across both sessions.

Results of the ratings indicate substantive inter-rater reliability for both the first (Cronbach's alpha = .84) and second sessions (Cronbach's alpha = .701). For both sessions, a paired-samples t-test was computed to investigate whether participants perceived the segments as belonging to one category more than another. For the first session, participants indicated that the segments were more readily perceived as "OO" than "AAH" (p=.002). For the second session, three paired-sample t-tests were computed to test each of the three categories against one another. Results suggest that segments were readily perceived as /u/ compared with /ɔ/ (p=.03). No other statistically significant effects were observed; a larger-scale study should seek to investigate various factors possibly affecting the perception of great ape quasi-vowels by human listeners, as well as mapping acoustic properties corresponding to the perception of such qualities.

Finally, resynthesis of observed values may constitute an alternative method to that described here. Such syntheses can be accomplished using a variety of available software [21, 22].

6. TOWARD VOCAL TRACT MODELS

The human vocal tract was reconfigured in evolution, including changes in the marked flexure of the skull base, expansion of the pharyngeal cavity, descent of the larynx and descent of tongue root into the throat [23, 24], and concomitant "particulation" (rounding) of the tongue into independently movable sections [25]. Through its reconfiguration, the human vocal tract attains a roughly 1:1 relationship between vertical and horizontal sections, affording extreme articulations necessary for quantal vowels [24, 26]. Great apes, like other nonhuman mammals have a "flat" tongue residing almost entirely in the mouth [1, 24, 27], and narrow pharynx (i.e., back cavity), while that of (adult) humans is partially descended into the throat.

While much remains unknown about articulatory capacities by nonhuman great apes, there are currently no indications that the vocal tracts of, e.g., orangutans, are capable of achieving the extreme 1:10 midpoint discontinuities required for quantal vowels [28, 24]. In this context, it is worth noting that work by Fitch et al. is commonly claimed to have refuted this idea [8]. However, the extent of the macaque data presented by those authors were driven by outlier data observed when the animal was yawning, and also did not extend to quantal vowels [i] or [u] [29]. Thus it does not constitute evidence against purported anatomical limitations on nonhuman primates' articulatory capacities [24, 29, 30, 31].

Thus, having confidently observed [u]-like formant dispersions, we are faced with an intriguing question: given an orangutan vocal tract, how may such acoustic properties otherwise be produced? Two clues come from *(1)* visual observations of chimpanzee "hoots" – which are seemingly acoustically and perceptually comparable to orangutan long call segments analyzed above, which shows that these calls are produced with protruded rounded lips [32]. And *(2)*, previous research on the acoustic consequence of laryngeal air sacs (found in most primates, including orangutans) [33], which indicates that air sacs shift down formants, and shifts them closer together.

To verify estimated formants and putative vowel-quality identified in orangutan long calls, we therefore seek to reverse engineer a plausible vocal tract shape for the animal producing those sounds. We performed this modeling effort using the *Wormfrek* software (J. Liljencrants, KTH), which allows the systematic variation of the vocal tract area transfer function as a prediction based on a sequence of uniform tubes [34] (Fig. 4). These efforts, while tentative, are suggestive of vocal tract shapes allowing animals with "unconfigured" vocal tracts to achieve [u]-like formant dispersions and vowel-like qualities. A larger-scale study should be conducted to determine the validity of these impressions. The acquisition of real-life vocal tract area dimensions and measurements promises to improve the performance of future such efforts.

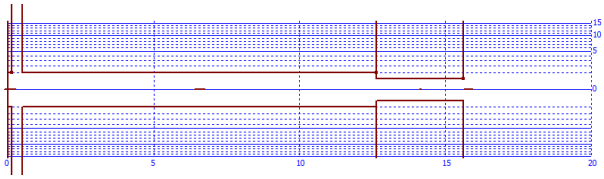


Figure 4: Tube model rendered in the *Wormfrek* software. The model assumes a vocal tract length of 12 cm, elongated with protruding (3 cm) rounded (.4 cm^2) lips, narrowing the lip passage. A .5cm cavity of 30cm^2 is assumed after a brief constriction (.125cm), simulating air sacs. The model predicts F_1=273 Hz, and F_2=1070 Hz.

7. DISCUSSION

Using a novel protocol of supervised acoustics paradigms, PREQUEL, we have confidently identified for the first time [u]-like productions in wild orangutans. Importantly, our approach allows for bypassing common hazards in related work, such as mistaking harmonic partials for formants. Though applied here to orangutans, PREQUEL can theoretically be applied to any primate call. This poses the exciting possibility for charting vowel-like articulations that, among hominids, may be older than speech itself. Moving forward, we intend to apply the PREQUEL protocol to a range of vocalizations by great apes as well as non-great ape primates, with the ultimate goal of mapping ancestral human articulatory capacities.

8. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the UK Research & Innovation, Future Leaders Fellowship (grant agreement number MR/T04229X/1; ARL), and the Swiss National Science Foundation (PCEFP1_186841; SM). The results of this work and the tools used will be made more widely accessible through the national infrastructure Sprakbanken Tal under funding from the Swedish Research Council (2017-00626).

9. REFERENCES

[1] V. E. Negus, *The comparative anatomy and physiology of the larynx*. Heinemann, 1949.

[2] D. F. N. Harrison, *The anatomy and physiology of the mammalian larynx*. Cambridge University Press, 1995.

[3] G. Kelemen, "Comparative anatomy and performance of the vocal organ in vertebrates," in *Acoustic behaviour of animals*, R. Busnel, Ed., 1963, pp. 489–521.

[4] M. J. Owren, R. M. Seyfarth, and D. L. Cheney, "The acoustic features of vowel-like grunt calls in chacma baboons (papio cyncephalus ursinus): Implications for production processes and functions," *Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2951–2963, 1997.

[5] P. H. Lieberman, D. H. Klatt, and W. H. Wilson, "Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates," *Science*, vol. 164, pp. 1185–1187, 1969.

[6] P. Lieberman, E. S. Crelin, and D. H. Klatt, "Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee," *American Anthropologist*, vol. 74, pp. 287–307, 1972.

[7] L. J. Boë, F. Berthommier, T. Legou, G. Captier, C. Kemp, . T. R. Sawallis, and J. Fagot, "Evidence of a vocalic proto-system in the baboon (papio papio) suggests pre-hominin speech precursors," *PloS one*, vol. 1, no. 12, pp. 3329–3343, 2017.

[8] W. T. Fitch, B. D. Boer, N. Mathur, and A. A. Ghazanfar, "Monkey vocal tracts are speech-ready," *Science Advances*, vol. 12, no. 2, p. e1600723, 2016.

[9] A. G. Ekström, "Ape vowel spaces remain elusive: A comment on grawunder et al. (2022)," *International Journal of Primatology*, pp. 1–3, 2022.

[10] A. R. Lameira and S. A. Wich, "Orangutan long call degradation and individuality over distance: a playback approach," *International Journal of Primatology*, vol. 29, pp. 615–625, 2008.

[11] S. Granqvist, "tolvan.com," Downloaded 2023.

[12] D. L. Bowling, M. Garcia, J. C. Dunn, R. Ruprecht, A. Stewart, K. H. Frommolt, and W. T. Fitch, "Body size and vocalization in primates and carnivores," *Scientific reports*, vol. 7, no. 2, pp. 1–11, 2017.

[13] J. C. Mitani, T. Hasegawa, J. Gros-Louis, P. Marler, and R. Byrne, "Dialects in wild chimpanzees?" *American Journal of Primatology*, vol. 27, pp. 233–243, 1992.

[14] R. B. Monsen and A. M. Engebretson, "The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction," *Journal of Speech, Language and Hearing Research*, vol. 26, pp. 89–97, 1983.

[15] J. Sundberg, "Articulatory configuration and pitch in a classically trained soprano singer," *Journal of Voice*, vol. 23, pp. 546–551, 2009.

[16] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1960.

[17] ——, "Descriptive analysis of the acoustic aspects of speech," *Logos*, vol. 5, pp. 3–17, 1962.

[18] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage Learning, 2011.

[19] J. C. Catford, *A practical Introduction to Phonetics*. Oxford University Press, 1988.

[20] S. Granqvist, "The visual sort and rate method for perceptual evaluation in listening tests," *Logopedics Phoniatrics Vocology*, vol. 28, pp. 109–116, 2003.

[21] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 3, no. 67, p. 971â995, 1980.

[22] S. Barreda, *phonTools: Functions for phonetics in R.*, 2015, r package version 0.2-2.1.

[23] J. T. Laitman and R. C. Heimbuch, "The basicranium of plio-pleistocene hominids as an indicator of their upper respiratory systems," *American Journal of Physical Anthropology*, vol. 3, no. 59, pp. 323–343, 1982.

[24] P. Lieberman, "Vocal tract anatomy and the neural bases of talking," *Journal of Phonetics*, vol. 4, no. 40, pp. 608–622, 2012.

[25] M. Studdert-Kennedy, "The particulate origins of language generativity: From syllable to gesture," in *Approaches to the evolution of language*, 1998, pp. 202–221.

[26] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 1, no. 17, pp. 3–45, 1989.

[27] E. S. Crelin, *The human vocal tract*. Vantage Press, 1987.

[28] B. De Boer and T. W. Fitch, "Computer models of vocal tract evolution: An overview and critique," *Adaptive Behavior*, vol. 1, no. 18, pp. 36–47, 2010.

[29] P. Lieberman, "Comment on "monkey vocal tracts are speech-ready"," *Science Advances*, vol. 7, no. 3, p. e1700442, 2017.

[30] H. Takemoto, "Morphological analyses and 3d modeling of the tongue musculature of the chimpanzee (pan troglodytes)," *American Journal of Primatology*, vol. 70, pp. 966–975, 2008.

[31] A. G. Ekström, "Viki's first words: A comparative phonetics case study," *International Journal of Primatology*, pp. 1–5, 2023.

[32] S. Grawunder, N. Uomini, L. Samuni, T. Bortolato, C. Girard-Buttoz, R. M. Wittig, and C. Crockford, "Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage," *Philosophical Transactions of the Royal Society B*, vol. 1841, no. 377, 2022.

[33] B. de Boer, "Acoustic analysis of primate air sacs and their effect on vocalization," *The Journal of the Acoustical Society of America*, vol. 6, no. 126, pp. 3329–3343, Dec. 2009.

[34] J. Liljencrants and G. Fant, "Computer program for vt-resonance frequency calculations," *STL-QPSR*, vol. 16, pp. 15–21, 1975.

PROSODY UNDER CONTROL: CONTROLLING PROSODY IN TEXT-TO-SPEECH SYNTHESIS BY ADJUSTMENTS IN LATENT REFERENCE SPACE

Juraj Šimko, Tuukka Törö, Martti Vainio, Antti Suni

University of Helsinki, Finland
firstname.secondname@helsinki.fi

ABSTRACT

We present a methodology for controlling prosodic characteristics of the output of an end-to-end TTS system based on direct adjustments of latent reference style conditioning vectors (style embeddings). The adjustments follow directions of maximal change calculated as gradients of regression fits of prosodic features against dimensions of the latent space. We also introduce a procedure for mitigating inter-dependencies among control in multiple prosodic dimensions based on an orthogonal projection. As the method does not rely on explicit prosodic annotations used for training, it is easy to implement, and can be applied to existing models without retraining. Our approach is evaluated on two single-speaker speech corpora that contain prosodically and stylistically rich material.

Index Terms: speech synthesis, style embeddings, prosodic features, explicit control, disentanglement

1. INTRODUCTION

As neural speech synthesis achieved human-level performance in reading aloud neutral sentences, the attention of the community has shifted towards methods of eliciting different speaking styles and rich prosodic variation ubiquitous in real human speech. Simply adding prosodic variety and different speaking styles to training data is not enough; the unaccounted variation present in the corpus is detrimental to the synthesis quality and leads to unpredictable and uncontrollable results.

In order to capture prosodic variation, utterance-level latent neural representations, *style embeddings*, are often used as an additional conditioning of the synthesis system [1, 2]. During training, the target acoustic representation identical to the systems output can be added as an additional input, and fed by a reference encoder through a heavily constrained bottleneck, before being concatenated with the textual input. The bottleneck style embeddings encode the recoverable (prosodic and stylistic) variation that is not inferrable from the text alone. Given an appropriate architecture, this approach has been shown to achieve a considerable degree of disentanglement of prosodic information from the text.

The style embedding latent space encodes the prosodic and stylistic characteristics of speech in a high-dimensional form that does not allow for a straightforward control of these characteristics in the synthesized output. Modifications to the encoder architecture [3], and learned pseudo-label approaches [4] have been applied to disentangle these complex representations or reduce the dimensionality, but robust control in terms of prosodic characteristics is still very much an open problem.

In this work, we present and evaluate a methodology for identifying the way in which relevant prosodic characteristics of speech are encoded in the latent embedding space and devise a mechanism for eliciting prosodic variation in a controlled way. The underlying assumption is that the latent space encodes the prosodic features in an entangled, but topologically coherent fashion. Broadly following the ideas presented in, e.g., [5, 6], we show that by following the gradients in the style embedding space, corresponding to systematic change in the investigated prosodic features of the encoded utterances, we can achieve a robust control of prosodic characteristics. We also present a method for "dissociating" the prosodic control, i.e., for minimizing the effects of elicited change in some prosodic features on other features of interest.

The proposed methodology is implemented using *reference prosody embeddings* trained within a Tacotron2 speech synthesis architecture as introduced in [2], and trained on two corpora of Finnish and English speech material (see Section 4).

2. PROSODIC FEATURES

Our approach is evaluated for four phonetic-prosodic characteristics of speech utterances derivable directly from speech signal: f_0 mean, f_0 standard deviation, spectral tilt, and speaking rate. Perceptually, these features correspond to overall pitch level, liveliness, voice quality, and tempo, respectively.

For utterances in the corpora as well as for synthesized utterances, the frequency related features were extracted by calling Praat [7] with the Parselmouth Python library [8]; speaking rate was approximated using text length as a proxy for the number of phonemic units. In more detail:

- f_0-MEAN and f_0-STD (standard deviation) were computed in semitones with the pitch floor at 75 Hz, ceiling at 400 Hz and a time step of 0.1;
- Spectral TILT was computed by analyzing the power spectrum to Long-Term Average Spectrum with a bandwidth of 100 Hz and computing its slope with the low band between 0 and 1 kHz and the high band between 1 and 4 kHz, using energy as the averaging method;
- Speaking RATE was calculated by dividing the orthographic length of the utterance by the total duration of the sound file.

3. PROSODIC CONTROL

Despite the complex nature of the underlying representation (see Fig. 1), the latent style embedding spaces have been shown to encode some relevant prosodic characteristics in a systematic way [5, 6, 9, 10, 11, 12]. In what follows we use a linear regression approach to devise a way to "extract" this systematicity from an embedding space produced by a trained encoder. The regression fits are used as an approximation of the relationship between the embedding space and the prosodic features of interest.

3.1. Relationship between the latent embeddings and prosodic features

Let Y be a set of n-dimensional style embedding vectors, obtained by running the training data through a trained reference encoder. In order to achieve comparable variance along all dimensions, we first Z-score normalize the vectors in Y by scaling. Let X be the resulting normalized version of Y with means equal 0 and standard deviations equal 1 along each dimension. Let *feat* be a (single dimensional) vector of phonetic-prosodic features of interest, calculated for the utterances encoded by X (e.g., an f_0-MEAN for each utterance).

A linear regression fit $feat \sim a_0 + a_1x_1 + \cdots + a_nx_n$ provides the best (in terms of RMS distance) linear approximation of the dependence of *feat* on the normalized embedding vectors. The regression coefficient vector $\mathbf{a} = (a_1, \ldots, a_n)$ is the gradient of this linear fit ($a_i \approx \frac{\partial feat}{\partial x_i}$); the vector provides the direction of the maximal slope of the fitted hyperplane, i.e., an estimate of the direction within the space X along which the feature *feat* exhibits the maximal increase.

The numerical values of the gradient vector $\mathbf{a}$ depend on the values and units of the phonetic-prosodic feature in *feat*. In order to "fit" the movement captured by vector $\mathbf{a}$ within the space X, the vector is linearly scaled so that the greatest absolute value along any dimensions equals 1 (by applying $\frac{\mathbf{a}}{\max|a_i|}$). As the standard deviation of the normalized embedding vectors is 1 along every dimension, this means the "movement" in the direction of the normalized vector $\mathbf{a}$ corresponds to maximally 1 standard deviation distance along any dimension in space X.

The normalized direction vector $\mathbf{a}$ is recast back to the original latent space Y by multiplying by the original standard deviation along every dimension. The resulting vector $\mathbf{b}$ provides an estimate of the direction in the reference prosody encoding space Y of the steepest positive change in the values of the feature of interest, and thus presumably allows for an explicit control of the synthesized speech in terms of the feature *feat* (see Section 4).

3.2. Dissociation by orthogonalization

The procedure described above provides an estimate of the direction of maximal change for a single given prosodic-phonetic feature. Given a possible entanglement of the way different features are encoded in the latent space Y (arising from potential correlations in the corpus), the process does not guarantee that the given direction does not considerably influence the other prosodic characteristics of synthesized speech. Systematically increasing mean f_0 might lead to an increase in, for example, speaking rate, if this type of relationship exist in the corpus. In order to attempt to dissociate the mutual inter-dependency of the controlled features we propose an approach based on orthogonal projection.

Let $\mathbf{a}_{feat_0}, \ldots, \mathbf{a}_{feat_m}$ be the direction vectors for $m+1$ features of interest, and let $\mathbf{F} = (\mathbf{a}_{feat_1}, \ldots, \mathbf{a}_{feat_m})$ be a matrix with columns corresponding to the feature vectors $\mathbf{a}_{feat_1}, \ldots, \mathbf{a}_{feat_m}$. Given an assumption of linear independence of these vectors, the matrix $\mathbf{F}$ is a basis of a hyperplane in the space X. The orthogonal projection $P\mathbf{a}_{feat_0} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{a}_{feat_0}$ yields the best approximation of the vector $\mathbf{a}_{feat_0}$ in the hyperplane $\mathbf{F}$. The vector $O\mathbf{a}_{feat_0} = \mathbf{a}_{feat_0} - P\mathbf{a}_{feat_0}$ is then the best approximation of the vector $\mathbf{a}_{feat_0}$ in the orthogonal complement of the hyperplane $\mathbf{F}$; it is orthogonal to all vectors in the space. In other words, the vector $O\mathbf{a}_{feat_0}$ depicts the best approximation of the direction (in the normalized space X) yielding maximal increase for the feature $feat_0$ and minimal (in terms of the linear regression fit) increase for all other features $feat_1, \ldots, feat_m$. In our subsequent analysis, we will use both these orthogonalized and non-orthogonalized direction vectors in order to dissociate the effects of prosodic control proposed here.

4. EVALUATION

In the present work we used a Tacotron 2 architecture [13], extended with a reference encoder. The implementation of the reference encoder matches the architecture described in [2]; the mel-spectrogram is processed and downsampled with six 2D convolutional layers, followed by a gated recurrent unit layer (GRU), and the final output of the GRU is taken as the style embedding. This 128 dimensional embedding is then concatenated with each character embedding of the utterance, forming the input to the decoder.

In order to evaluate the proposed prosodic control method, we trained two voices. The first voice was trained on the commonly used LJSpeech dataset [14], 24 hours of relatively lively narrative style read by an American English female speaker. The second voice was trained on a new, yet to be published Finnish speech corpus, FinSyn, of two professional female speakers, with duration of approximately 30 hours each. The corpus was designed with "ordinary" speaking style variation in mind, containing mostly continuous read speech of texts with varying degrees of formality, as well as semi-spontaneous and spontaneous speech. A 22 hour subset from a single speaker was used for training.

Both voices were trained on orthographic text, with numbers and abbreviations expanded. For the Finnish voice, we applied a version of curriculum learning, training the system first on a more formal utterances of isolated sentences until diagonal attention was achieved, after which the rest of the data was added to the training set. Both voices were trained from scratch for 200,000 steps with a batch size of 16, on a single Nvidia V100 GPU.

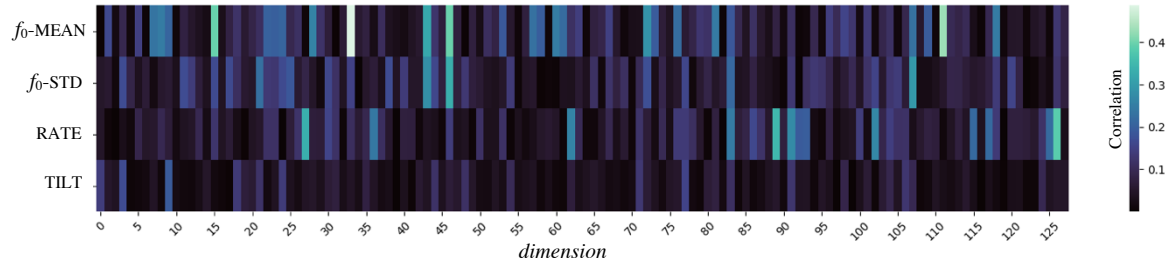


Figure 1: Correlations between the prosodic features and dimensions of the style embedding reference space.

Table 1: Slopes and r^2-values (in brackets) of the fits with scaling factor of the directional vector for the controlled features (in columns) as predictors and the feature values extracted from the synthesized utterances as dependent variables (in rows).

	non-orthogonal direction vectors				*orthogonal direction vectors*			
	f_0-MEAN	f_0-STD	TILT	RATE	f_0-MEAN	f_0-STD	TILT	RATE
Finnish corpus								
f_0-MEAN	**1.43** (*0.92*)	0.88 (*0.71*)	-0.15 (*0.25*)	0.18 (*0.45*)	**1.26** (*0.95*)	0.37 (*0.47*)	-0.20 (*0.37*)	0.18 (*0.30*)
f_0-STD	0.11 (*0.12*)	**0.65** (*0.90*)	-0.13 (*0.38*)	-0.05 (*0.09*)	0.10 (*0.12*)	**0.72** (*0.94*)	0.00 (*-0.01*)	0.17 (*0.54*)
TILT	-0.35 (*0.31*)	-0.53 (*0.50*)	**1.29** (*0.90*)	-0.05 (*0.00*)	-0.28 (*0.29*)	-0.36 (*0.33*)	**1.24** (*0.91*)	-0.25 (*0.27*)
RATE	0.05 (*0.03*)	-0.04 (*0.01*)	0.12 (*0.17*)	**0.41** (*0.66*)	0.05 (*0.02*)	0.01 (*-0.01*)	0.09 (*0.09*)	**0.37** (*0.62*)
English corpus								
f_0-MEAN	**1.53** (*0.97*)	0.38 (*0.5*)	-0.12 (*0.13*)	0.2 (*0.36*)	**1.47** (*0.96*)	0.35 (*0.45*)	-0.11 (*0.08*)	0.09 (*0.11*)
f_0-STD	0.07 (*0.07*)	**0.64** (*0.93*)	0.05 (*0.06*)	-0.06 (*0.14*)	0.01 (*0.00*)	**0.61** (*0.89*)	0.01 (*0.00*)	0.01 (*0.00*)
TILT	-0.12 (*0.02*)	0.13 (*0.03*)	**2.01** (*0.89*)	0.14 (*0.04*)	-0.1 (*0.01*)	-0.10 (*0.02*)	**2.00** (*0.89*)	0.24 (*0.11*)
RATE	0.07 (*0.03*)	-0.08 (*0.06*)	-0.01 (*-0.01*)	**0.36** (*0.48*)	0.04 (*0.01*)	-0.07 (*0.04*)	0.03 (*0.00*)	**0.30** (*0.39*)

4.1. Evaluation of prosodic control

For each of the two trained models (the Finnish and English voice), the style embedding vectors were extracted for each utterance in the training corpus using the reference encoder. Also, the values of the prosodic features were calculated for the same utterances, as described in Section 2. Fig. 1 shows the correlations between the individual dimensions of the embedding space and the prosodic features. As seen on the figure, each prosodic feature correlates with multiple dimensions of the embedding space, and several dimensions relatively strongly correlate with multiple features.

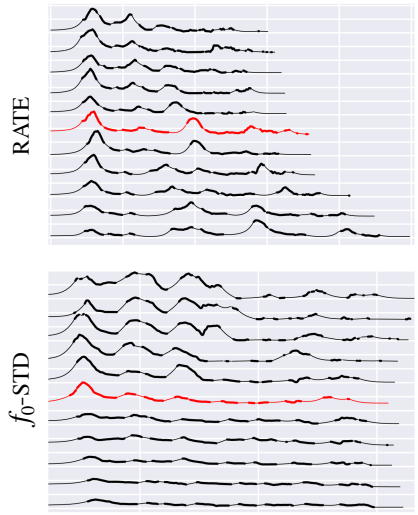


Figure 2: f_0 trajectories of utterances generated with the Finnish system, controlling speech rate and f_0 standard deviation, zero modification in red

The efficacy of the method for disentangling this complex relationship and for providing a control over the prosodic characteristics was evaluated in the following way. Both non-orthogonal and orthogonal versions of the directional vectors for the investigated features (vectors **b** from Section 3) were calculated using the style embedding spaces and prosodic features. Each direction vector was orthogonalized against the hyperspace containing the direction vectors of all three remaining features. The direction vectors were then scaled by the factors corresponding to integers between -5 and 5 (i.e., up to 5 standard deviations of the original distribution), and added to the mean vector of the entire embedding space.

For each scaled version, ten Harvard sentences/their Finnish translations were synthesized and the prosodic characteristics of the synthesized sentences were extracted (see https://tuukkaot.github.io/PhoneticFeatureTTS). Fig. 2 shows the resulting utterances obtained by scaling the direction vector for f_0-STD and RATE feature.

Fig. 3 shows the distributions of the prosodic characteristics of the test sentences (in rows) as a function of the directional vector scaling (in columns) for the Finnish corpus (the corresponding plot for the English corpus shows very similar behaviour). The plots on the diagonal, depicting the behaviour in terms of the controlled feature, show a strong albeit not always linear relationship between the scale and the controlled prosodic characteristics; increasing the scale leads to a systematic increase of the given feature. The plots off the diagonal in Fig. 3 generally show considerably weaker influence of the control of on the non-controlled prosodic characteristics.

In order to quantify these observations, a linear regression model with feature values of the synthesized utterances as the dependent variable and scaling factors as predictors were fitted for every combination captured in Fig. 3; the fits are shown as the lines in the figure. The slopes of these fits and the adjusted r^2-values of the models (depicting the quality fit) are listed in Table 1. The table summarises the models for both corpora and for both non-orthogonal and orthogonalized the direction vectors.

The quality of the fits shows that the control exerts reliable and strong influence on the controlled features. The r^2s and the slopes of the fits are greater on the diagonals (in bold) than off diagonals, i.e, we get better quality of fits and stronger relations for the controlled features than for the non-controlled ones (as the slopes depend on the units of the dependent variables, only the values in the same row are directly comparable).

Regarding the effects of orthogonalization, in the majority of cases both slopes and r^2-values off diagonals are lower for the orthogonalized directional vectors then for the non-orthogonalized ones. For example, the strong influence of f_0-STD control on the f_0-MEAN feature for the Finnish corpus, clearly visible in Fig. 3, gets substantially attenuated: the slope decreases from 0.88 to 0.37, r^2-value from 0.88 to 0.37.

For some feature combinations, primarily related to speech rate control, the orthogonalization did not work as envisaged; e.g., the effects of RATE control on TILT feature actually *increased* for both corpora.

5. DISCUSSION

The presented method of controlling prosodic features in TTS by directly manipulating style embedding space elicited considerable variation in the controlled features. Because the method constructs (approximates) the reference vectors in the existing trained latent space, the prosodic characteristics are not controlled explicitly, but the synthesis rather relies on—and replicates—the relevant variance in the training material. For example, the observed speech rate variation is augmented by appropriately inserted/removed silent pauses rather then by a uniform stretching of the synthesised output (see Fig.2). The synthesis quality is thus not seriously compromised even for the relatively high values of the scaling factors.

While we have evaluated the prosodic control only for a relatively small set of prosodic characteristics, the methodology (unlike the explicit prosodic control systems, e.g., [15, 16]) allows for a *post hoc* inclusion of any quantifiable controlled features in a straightforward way by simply calculating and scaling the appropriate direction vectors (see, e.g., [5, 6] using similar methodology). Also, we have chosen a relatively "greedy" approach to the orthogonalization of directional vectors by enforcing their disentanglement from all other investigated features. This might not be necessary in practice when the user of the TTS system might require independent control of a smaller combinations of prosodic characteristics (e.g., tempo from pitch level but not from voice quality). The proposed method accounts for this type of flexibility.

The method can be easily combined with other approaches to style control. While we added the scaled direction vectors to the mean style embedding of the corpus (equivalent to an "average" rendition), this starting point could be shifted to any point within the embedding space, for example, to a style embedding representing a particular style identified in the training material (by using appropriate explicit or learned labels).

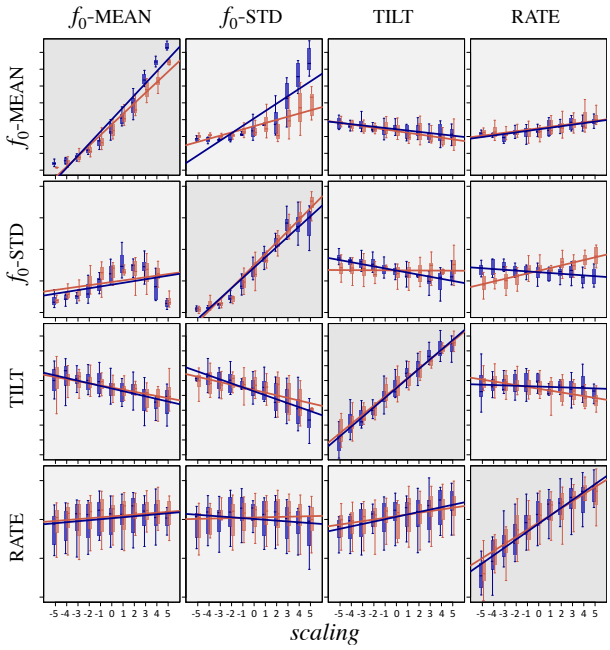


Figure 3: The distributions of the prosodic characteristics of the test sentences as a function of the directional vector scaling for the Finnish corpus. The columns correspond to the controlled features, the rows to the consequences of the control. The non-orthogonalized control in blue, orthogonalized in red.

Comparing the performance for the four tested features shows relatively less reliability of speech rate control compared to the other three features. The greater variability in RATE feature in the synthesized material (reflected by the lower r^2-values) is presumably a consequence of using the number of orthographic characters for calculating the RATE feature, and the complex relationship between the letters and durations of speech sounds. Also, in order to simplify the process, we used the duration of the sound file rather than duration of the actual utterance for the RATE feature calculation; leading and training silences (and also the pauses within the utterance) might contribute to the variance.

While the corpora used here were not explicitly designed with a vast range of variation along the controlled characteristics in mind, they do contain prosodically rich material. This richness does not compromise the synthesis quality produced by the used TTS system, but undoubtedly contributes to the degree of controllability. Design and collection of corpora containing prosodically varied material (rather than containing carefully crafted sentences and steady articulation) is one the necessary requirements for achieving high quality, and prosodically rich and controllable synthesis.

6. REFERENCES

[1] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[3] G. E. Henter, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *CoRR*, vol. abs/1807.11470, 2018. [Online]. Available: http://arxiv.org/abs/1807.11470

[4] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[5] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," *arXiv preprint arXiv:1903.11570*, 2019.

[6] N. Tits, K. El Haddad, and T. Dutoit, "Analysis and assessment of controllability of an expressive deep learning-based tts system," in *Informatics*, vol. 8, no. 4. Multidisciplinary Digital Publishing Institute, 2021, p. 84.

[7] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[8] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[9] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. Interspeech*, 2017.

[10] A. Suni, M. Wlodarczak, M. Vainio, and J. Šimko, "Comparative analysis of prosodic characteristics using wavenet embeddings," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019) Crossroads of Speech and Language*. ISCA, 2019.

[11] J. Šimko, M. Vainio, A. Suni *et al.*, "Analysis of speech prosody using wavenet embeddings: The lombard effect," in *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[12] K. Hiovain, A. Suni, S. Kakouros, and J. Šimko, "Comparative analysis of majority language influence on north sámi prosody using wavenet-based modeling," *Language and Speech*, p. 0023830920983591, 2020.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[14] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[15] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[16] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

WHO'S SPEAKING? PREDICTING SPEAKER PROFESSION FROM SPEECH

Yaru Wu[1,3,4,*], Lihu Chen[2,*], Benjamin Elie[3,5,*], Fabian Suchanek[2], Ioana Vasilescu[3], Lori Lamel[3]

[1]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[2]Télécom Paris, Institut Polytechnique de Paris, France
[3]LISN, Univ. Paris-Saclay, 91405 Orsay cedex, France
[4]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[5]LEL, School of PPLS, The University of Edinburgh; Edinburgh, Scotland, United Kingdom

yaru.wu@unicaen.fr, lihu.chen@telecom-paris.fr, benjamin.elie@ed.ac.uk, fabian.suchanek@telecom-paris.fr, {ioana.vasilescu, lori.lamel}@lisn.fr

ABSTRACT

Variations in speech can reveal the gender, birth place, age, and socio-economic level of the speaker. In this paper, we show that even the profession of the speaker can be recovered from a recording. For this purpose, we design a method that combines features from both the speech signal and the transcription. For the features from the transcription, we used pre-trained language models. This allows us to train a model that predicts the speaker profession from both signals. Our empirical results show that our model can narrow down the profession of the speakers considerably.

Index Terms: knowledge base, large corpora, multimodal representation, pre-trained language models

1. INTRODUCTION

The recording of someone's speech provides two principal types of signals: the content of the speech (*what* is being said), and the acoustic and prosodic features of the speech (*how* it is being said). The latter features cover both segmental and supra-segmental levels and are reflected in various measures related to the acoustic properties of vowels and consonants, as well as to the prosodic specificity of such units (e.g., segment duration, pitch, speech rate, etc.).

It is known that these features can help predict the background of the speakers, most notably the age, gender, and the region of birth [1, 2, 3]. In this work, we take this analysis one step further: we show that even the profession of the speaker can to some degree be determined from a recording alone. We can show that the content of the speech alone is not sufficient for this purpose. Nor are the speech-based features. However, if both features are combined, we can predict the profession of the speaker with an accuracy of up to 76%. It matters thus not only *what* the speaker says, but also *how* they say it.

Our approach takes as input a speech recording of a speaker and its transcription. It then extracts features from both the speech signal and the transcription, using pre-trained language models for the latter. We then use a machine learning model to predict the speaker profession from these features. For our experiments, we use a corpus of French oral broadcast news from radio and TV channels recorded in the 2000s. As ground truth for training and evaluation, we use profession annotations from the YAGO knowledge base [4] that have been found by previous work [5].

The ability to predict the profession of the speaker can have implications for AI applications such as speaker classification, speaker identification, dialog systems, intelligent assistants etc. It also has ramifications in the domain of privacy: it shows that machines can identify a socio-economic characteristic of speakers, possibly unbeknownst to them or against their will. It is thus important to investigate to what degree such privacy violations are possible with today's means.

2. RELATED WORK

Variation is ubiquitous in human language and is linked to both physiological and psychological properties of the individuals who communicate. According to [6], physiological properties are related to age, gender, health etc., whereas psychological ones are related to their identity as individuals and position or role within a community. Both properties translate into an array of features ranging from very basic (physical) properties of the acoustic speech signal, to high level features (the semantics of the utterances). On the speech processing side, the features cover both acoustic and prosodic levels, ranging from acoustic parameters at the phone level (for instance, the formants of vowels), to prosodic parameters such as duration, fundamental frequency and in-

*These authors contributed equally to this work.

tensity [7, 8, 9, 10, 11]. Of these features, acoustic properties of consonants and vowels as well as fundamental frequency related measures appeared to be speaker specific and have allowed to quantify inter-speaker variability [7, 11].

On the textual side, Natural language processing (NLP) methods have been used for a variety of prediction tasks about the speaker from discourse, such as political profiling [12], ideology classification [13], candidate to job matchings [14], or stance detection [15]. Closer to our approach, recent approaches combine speech and NLP features for more accurate prediction of speaker characteristics. For instance, in [16] the speaker stances in Mandarin ideological debate competitions are predicted by combining n-grams and acoustic prosodic features. In a similar direction, another approach [17] combines text and acoustics for ideology detection in Youtube data. However, the prediction of a socio-economic indicator as precise as the speaker profession has so far been out of reach.

Closest to our approach is the work of [5], whose dataset we use here. While that work found correlations between speech features and professions, it was unable to *predict* the profession from the speech features. This is because the speech features alone are not sufficient for that task, as our work shows. Only the combination with the speech content that we propose here allows a narrowing-down of the profession.

3. DATA AND METHODOLOGY

Given as input a recording of speech by a speaker with its transcription, and given a list of professions, our goal is to predict the profession of the speaker. Our approach proceeds in three steps: (1) feature extraction from the speech, (2) feature extraction from the transcription of the speech, and (3) classification of the feature vector to one of the professions. To train the model of step (3), and to evaluate our method, we need a ground truth, i.e., an annotation of speakers with professions for part of the data.

3.1. Feature extraction from speech

The first step of our method extracts features from speech signal. We focus on global features, computed at the speech chunk level. This level corresponds roughly to an utterance. This choice is made in order to avoid considering speaker-specific features that could be irrelevant for a speaker profession recognition. Computing global features at a larger time-scale also enables us to extract psychological variations of speech. The following features were extracted from the speech signal : speech rate, mean pitch, pitch span, pitch declination, normalized pitch peak extent, local peak dynamics of pitch, mean for each of the 4 formants, formant span for each of the 4 formants, voicing ratio and devoicing ratio. Speech rate is calculated on speech utterances between pauses that are equal to or greater than 200ms. The pitch and formant span are defined as the difference between the average distance between the maximums and the local mean contour and the average distance between the minimums and the local mean contour [18]. We extracted and computed the characteristics of the 4 first formant trajectories – hence the 4 mean formants and 4 formant spans in the table. Voiced and unvoiced ratio relate to non-canonical surface forms, corresponding to stops that changed their voicing category. Voicing refers to canonical /ptk/ pronounced [bdg], while unvoicing refers to canonical /bdg/ pronounced [ptk]. The estimation of non-canonical realizations is obtained with the method described in [19]. We then compute the ratio of produced non-canonical forms in each category (voicing and devoicing) by the speaker as marks for consonant reduction (voicing) or strengthening (devoicing) that may be speaker specific. Except for the voiced and unvoiced ratio, which are computed over all speech chunks of a particular speaker, features are computed at the speech chunk level, and then averaged over each speaker. Each speaker is then assigned the mean value and the standard deviation of each features. This results in a feature vector $\mathbf{s}$ of length $N = 30$ for a given speech recording.

3.2. Feature extraction from the transcription

Formally, the transcription of a speech recording is a word sequence of length l: $W = \{w_1, w_2, ..., w_l\}$. We extract a textual feature vector $\mathbf{t} \in \mathbb{R}^n$ from the sequence by using a encoder function $\phi(\cdot)$, $\mathbf{t} = \phi(W)$. The resulting vector $\mathbf{t}$ can be regarded as a document-level representation for the input transcription text. We adopt Pre-trained Language Models (PLMs) [20] as an encoder to extract textual features from transcriptions for speakers because PLMs have achieved the state-of-the-art across a series of Natural Language Processing tasks. More specifically, we use CamenBERT [21, 22], which targets French texts. The input transcription is truncated into word sequences with a maximum length of 500, and CamenBERT is applied to the input. Afterward, a pooling layer is used to the whole sequence of the last-layer hidden states to obtain the textual feature $\mathbf{t}$. We use a CamenBERT-large model[1] of dimension $n = 1048$. When a transcription is missing, we use a zero padding vector.

Profession	Politician	Journalist	Artist	Movie Person	Scientist	Lawyer	Business Person	Sports Person	All
Number of speakers	189	56	53	70	23	12	24	21	448
Utterances per speaker	505	1869	881	519	600	506	241	132	695

Table 1: Statistics of the top-8 frequent professions.

Profession	Keywords
Politician	France, politique, gouvernement, ministre, pays
Journalist	France, heures, président, Inter, soir
Artist	temps, film, films, cinéma, France
Movie Person	France, films, français, personnage, film
Scientist	France, monde, temps, Irak, Israel
Lawyer	vérité, société, France
Bussiness Person	travailler, monde, année
Sports Person	France, l'équipe, défense

Table 2: Top-k keywords extraction from all transcripts using YAKE [23]: Top-5 keywords for professions with sufficiently large samples; top-3 for those with fewer samples.

Profession	Speech signal	Text		Both	
	F1	F1	M	F1	Acc
Politician	70.9	65.9	41%	**77.2**	76.6
Journalist	22.6	38.4	29%	**44.0**	42.3
Artist	15.6	16.3	30%	**19.1**	13.6
Movie Person	49.6	8.1	74%	**60.1**	60.6
Scientist	6.7	**12.5**	26%	6.1	5.7
Lawyer	0.0	0.0	33%	**12.5**	15.4
Business Person	0.0	18.2	33%	**27.8**	22.9
Sports Person	14.8	**38.9**	14%	35.3	34.3
All	22.6	24.5	41%	**35.3**	33.9

Table 3: Performances (F1 score) of profession prediction with speech signal and textual features. Macro F1 and Accuracy for **Both** column. Column "M" gives the rate of speakers without transcriptions.

	Hit@1	Hit@2	Hit@3
Politician	66.3	92.9	96.3
Journalist	48.6	67.7	77.1
Artist	19.0	26.6	38.4
Movie Person	54.1	65.5	72.5
Scientist	22.2	30.1	30.3
Lawyer	0.0	10.2	20.1
Business Person	30.3	32.3	33.4
Sports Person	50.0	68.2	68.2
All	55.8	68.5	74.3

Table 4: Hit@K performance with combined speech and transcription features.

3.3. Model prediction

We now have, for the given speech recording, a feature vector $\mathbf{s} \in \mathbb{R}^m$ of speech features, and a feature vector $\mathbf{t} \in \mathbb{R}^n$ of text features. We combine both vectors by concatenation, and obtain a multimodal representation $\mathbf{u} \in \mathbb{R}^{m+n}$.

We then classify this vector using a Multi-Layer Perceptron (MLP) with one hidden layer [24]. We have tested different architectures and sets of input features by Leave-one-out cross-validation across speakers. The best results were obtained with a hidden layer size of 100, a RELU activation function, and a L2 penalty parameter of 1 for regularization. The number of finally selected features is reduced to 26, as some speech features are non-discriminant for this task. The rejected features are the standard deviation of the 3rd formant span, the mean value of the averaged 3rd formant, and the mean values of the averaged pitch and the pitch span. This can be explained by the fact that averaged values of the fundamental frequency and higher formants are expected to be more dependent on speaker characteristics (e.g. gender) than on speaker profession.

4. EXPERIMENTS

4.1. Dataset

We use the annotated dataset from [5]. It consists of the ESTER corpus [25] (80 hours of French formal journalistic speech, mainly from radio broadcast news) and the ETAPE corpus [26] (about 40 hours of less formal journalistic French speech data, mainly from debates and interviews). The two corpora were transcribed manually, and 66% of the utterances have an associated manual transcription. The transcription was force-aligned with the speech signal using the LISN (former LIMSI) speech transcription system [27, 28]. The speech transcription system was used to segment the speech data automatically to word and phone levels. Pauses, hesitations or breaths were detected automatically by the system. The minimum segment duration is 30ms, corresponding to 3 frames [29]. The dataset contains only utterances whose speaker could be identified. We combine all utterances per speaker, so that we arrive at one (pseudo-) recording per speaker. The dataset contains 342 distinct named speakers. For each speaker, the dataset also contains the ground truth in the form of a profession from the YAGO knowledge base [4]. Table 1 shows the statistics of the professions in our dataset.

4.2. Using only the speech signal

The performance results for profession prediction by the speech signals alone are shown in Table 3, in the column "Speech Signal". The best scores are obtained for *Politician* (F1=70.9) and *Movie Person* (F1=49.6). Both professions and in particu-

lar *Politician* are well represented in the two corpora. *Politicians* exhibit idiosyncratic patterns such as slow speech rate and high rate of disfluencies. They also have a tendency to consonant reduction through increased rate of non-canonical voiced consonants. This is in line with previous findings [5]. Beyond that, the median value of the speech rate across *Politicians* is the lowest among the professions. These trends can be associated with the profession. Reduction can be linked to a rather spontaneous and thus relaxed pronunciation and one can speculate that *Politicians* are used to speaking publicly and their speech is rather relaxed. *Movie Persons*, too, can be assimilated to trained public speakers and exhibit recognizable speaking traits resulting in a more spontaneous, relaxed speech. The scores for the other professions are much lower, indicating that speech signals alone cannot predict the profession.

4.3. Using only the speech transcriptions

The performance results for profession prediction using only the transcriptions are shown in the third column ("Transcription") of Table 3. We first observe that the textual features are more helpful than the speech features. The macro F1 has a lead of 1.9 absolute percentage points. To better understand what triggers this performance, we extracted the top keywords of each profession (Table 2). The results are consistent with the trends we saw for speech features: *Politician* and *Movie Persons* have rather profession-specific and frequent keywords, which explains the good performance of the textual model. For instance, *Politicians*, often refer to *"politique (politics)"*, *"gouvernement (government)"* and *"ministre (minister)"* in their speeches. We also notice that the performances for the *Movie Person* and *Lawyer* are extremely low. The former is caused by the high missing rate of transcriptions (74%) and the latter is a consequence of the small number of data samples (12 samples). *Artists* as broader category are also hard to distinguish from *Movie Persons*, as they use overlapping topical keywords (e.g., "film", "cinema").

4.4. Combining speech signal and transcription

We now combine the features from the speech signal and the text. The results are shown in the last two columns of the Table 3. The combination consistently performs the best, and the macro F1 across all professions is at least 10 percentage points higher than the other two (shown in the last row). This proves the interest of combining low and high level linguistic features: the integration of features extracted from both the speech signal and the text can better characterize the speaker and consequently narrow its profession. The result also suggests the reasons why human strategies are more effective: the human perception relies on a complex integration of both low and high level information and take advantage of the communicative context in a broad sense to decode speaker turns properties including the characteristics of the speaker that utter it.

The accuracy of the model is not perfect. We hypothesize that the small number of training samples constrains the performance of machine learning. Furthermore, some parts of the audio did not have manual transcripts (column M). Nevertheless, our model can correctly predict the professions for more than half of the speakers. Table 4 shows the Hit@k measure, i.e., the percentage of people for whom the correct profession is among the top k predictions. We see that, while the approach cannot outright guess the profession, the speech nevertheless gives away enough information to narrow down the set of professions to 2 or 3.

5. CONCLUSION

We have shown that it is possible to narrow down, or even predict, the profession of a speaker from a recording of their speech. This is an interesting finding that can be further deepened, e.g., by using more features from speech, using more training data, or using extrinsic features. The main finding underlines the complementarity of low and high level language features and the need to combine them in order to improve performance.

The results point to interesting avenues of research, where intrinsic characteristics (e.g. age, gender), meta-cognitive characteristics (e.g. current state of mind or subconscious attitude towards a topic), or socio-economic characteristics (e.g. social status) can be predicted from multi-level acoustic and linguistic features. Such work poses obvious challenges for the privacy of individuals. It is thus important to investigate to which degree such privacy violations are possible, and whether they require regulation. However, such work also brings opportunities: It could help dialog systems adapt their discourse to the user. For example, the explanation of an insurance case would be different for a lawyer and for a scientist. The work could also help identify fake recordings or fake videos of public personalities, if the extracted characteristics do not correspond to the known characteristics of the speaker. Finally, such analyses carry educational and scientific value, as they reveal speech characteristics that give away personal details.

6. ACKNOWLEDGEMENTS

This research was supported by the DATAIA / MSH Paris-Saclay grant "OTELO" and the ANR-20-CHIA-0012-01 ("NoRDF").

7. REFERENCES

[1] L. Lamel, J. Gauvain, and L. Canseco-Rodriguez, "Speaker diarization from speech transcripts," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA, 2004. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2004/i04_0601.html

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," 2021.

[4] T. Pellissier-Tanon, G. Weikum, and F. M. Suchanek, "YAGO 4: A Reason-able Knowledge Base," in *ESWC*, 2020.

[5] Y. Wu, F. Suchanek, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Using a knowledge base to automatically annotate speech corpora and to identify sociolinguistic variation," in *LREC*, 2022.

[6] T. Schultz, "Speaker characteristics," in *Speaker classification I*. Springer, 2007.

[7] M. Sambur, "Selection of acoustic features for speaker identification," *Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, 1975.

[8] J. Lindh and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Interspeech*, 2007.

[9] M. Ajili, J.-F. Bonastre, K. W. BEN, S. Rossato, and J. Kahn, "Comparaison des voix dans le cadre judiciaire: influence du contenu phonétique," *Journées d'Études sur la Parole*, 2018.

[10] O. Niebuhr and R. Skarnitzl, "Measuring a speaker's acoustic correlates of pitch–but which? a contrastive analysis based on perceived speaker charisma," in *ICPhS*, 2019.

[11] G. Chignoli, C. Gendrot, and E. Ferragne, "Caractérisation du locuteur par cnn à l'aide des contours d'intensité et d'intonation: comparaison avec le spectrogramme," in *Journées d'Études sur la Parole*, 2020.

[12] C. Mallavarapu, R. Mandava, S. Kc, and G. m Holt, "Political profiling using feature engineering and nlp," in *SMU Data Science Review*, 2018.

[13] A. A. Simoes and M. Castanos, "Fine-tuned bert for the detection of political ideology," 2020.

[14] T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Job prediction: From deep neural network models to applications," in *RIVF*, 2020.

[15] D. Küçük and F. Can, "Stance detection: A survey," in *ACL*, 2020.

[16] L. Li, Z. Wu, M. Xu, H. M. Meng, and L. Cai, "Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition," in *Interspeech*, 2016.

[17] Y. Dinkov, A. Ali, I. Koychev, and P. Nakov, "Predicting the leading political ideology of youtube channels using acoustic, textual, and metadata information," in *arXiv*, 2019.

[18] D. Patterson and D. R. Ladd, "Pitch range modelling: linguistic dimensions of variation," in *ICPhS*, 1999.

[19] I. Vasilescu, Y. Wu, A. Jatteau, M. Adda-Decker, and L. Lamel, "Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes," *Traitement Automatique des Langues*, vol. 61(1), 2020.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[21] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. De La Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *ACL*, 2020.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *EMNLP demos*, 2020.

[23] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, 2020.

[24] C. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[25] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Interspeech*, 2005.

[26] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *LREC*, 2012.

[27] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1, 2002.

[28] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing french broadcast news?" in *Interspeech*, 2005.

[29] M. Adda-Decker and L. Lamel, "The use of lexica in automatic speech recognition," in *Lexicon Development for Speech and Language Processing*. Springer, 2000.

ASRLUX: AUTOMATIC SPEECH RECOGNITION FOR THE LOW-RESOURCE LANGUAGE LUXEMBOURGISH

Peter Gilles, Léopold Hillah, Nina Hosseini-Kivanani

University of Luxembourg
peter.gilles@uni.lu, leopold.hillah@sproochtek.lu, nina.hosseinikivanani@uni.lu

ABSTRACT

We have developed an automatic speech recognition (ASR) system tailored to Luxembourgish, a low-resource language that poses distinct challenges for conventional ASR approaches due to the limited availability of training data and inherent multilingual nature. By employing transfer learning, we meticulously fine-tuned an array of models derived from pre-trained wav2vec 2.0 and Whisper checkpoints. These models have been trained on an extensive corpus of various languages and several hundred thousand hours of audio data, utilizing unsupervised and weak supervised methodologies, respectively. This includes linguistically related languages such as German, Dutch, and French, which expedite the cross-lingual training process for Luxembourgish-specific models.

Fine-tuning was executed utilizing 67 hours of annotated Luxembourgish speech data sourced from a diverse range of speakers. The optimal word error rate (WER) achieved for wav2vec 2.0 and Whisper models were 9.5 and 12.1, respectively. The remarkably low WERs obtained serve to substantiate the efficacy of transfer learning in the context of ASR for low-resource languages.

Keywords: Automatic speech recognition, wav2vec 2.0, Whisper, Luxembourgish

1. INTRODUCTION

Until recently, developing powerful Automatic Speech Recognition (ASR) systems for small and low-resource languages has been quite a challenging task. These languages often lack sufficient data for training and evaluating ASR systems, which makes it challenging to achieve high accuracy in recognition. It is the purpose of this study to explore and demonstrate how an ASR system for Luxembourgish can be developed, taking into account the specific constraints of a low-resource language, i.e., scarcity of training material, and the specific situation of multilingualism. In the course of the 20th century, the small language Luxembourgish evolved to the national language of the Grand-Duchy of Luxembourg [1], [2]. Luxembourgish is used by approximately 300,000 speakers, mainly as a spoken language of everyday use, but also as the only spoken language in parliament. As an identity symbol, this language is used more and more in speaking and writing due to social media and the digitization of everyday life. This has led to the need for tailored Natural Language Processing (NLP) and voice tools (especially Speech-to-Text (STT) and Text-to-Speech (TTS)). Luxembourgish is typologically quite close to German and has in fact evolved out of a German dialect, which has been standardized recently to a certain degree with regard to spelling and lexicon. In addition to that, Luxembourgish features a high degree of multilingualism: Words and phrases coming mainly from French and German are part of the regular lexicon and occur quite frequently. And it goes without saying that these multilingual insertions pose challenges for traditional approaches to ASR, which were based on monolingual systems. In recent years, several approaches to address multilingual ASR for Luxembourgish have been presented, mainly by re-utilizing training resources from related languages (see [3], [4], [5]). However, recent advances in machine learning (ML) and NLP have made it possible to build more accurate ASR systems for small and low-resource languages. In particular, these ML advances have led to the improvement of techniques (e.g., unsupervised and semi-supervised learning methods) for transferring knowledge from large, well-resourced languages to smaller languages with fewer resources. Based on recent advances in ASR research, the present study will be based on two of the most promising approaches: wav2vec 2.0 [6], developed by Meta, and Whisper [7], developed by OpenAI. For both approaches, large pre-trained model checkpoints are provided, which can be fruitfully used for adapting to a low-resource language like Luxembourgish.

2. RECENT ADVANCES IN ASR

In recent years, there have been significant advancements in the field of ASR due to the use of deep learning techniques. These techniques have proven to be particularly effective for ASR, leading to rapid improvements in the accuracy and performance of ASR systems. Currently, self-supervised transformer-based models for audio processing have significantly improved the performance of ASR systems. These models are able to handle unlabeled data effectively, which is a significant advantage over traditional ASR systems (such as HMM-based approaches like Kaldi [1]) that require a large amount of labeled data and a pronunciation dictionary for training. With the availability of pre-trained multilingual base models, it is possible to make significant progress in developing ASR systems for low-resource languages even when training data is scarce.

We will focus on two of the most effective algorithms, i.e., wav2vec 2.0 and Whisper, which are especially useful for the Luxembourgish case study because they provide pre-trained checkpoints that have been trained on multiple languages. The idea behind wav2vec 2.0 is to learn by predicting the context around a given audio segment. Wav2vec 2.0 uses a technique called "time-depth separable convolution," which involves applying a series of convolutional neural networks (CNNs) to the audio waveform data. CNNs are trained to extract features from the audio data at multiple scales.

For learning the contextual representations of speech data, both CNNs and recurrent neural networks (RNNs), which can be trained on large amounts of unannotated data, are being used. This is beneficial to use this model in low-resource languages where annotated data is often scarce. The Whisper [7] approach uses a transformer-based sequence-to-sequence (seq2seq) architecture for learning the contextual representation of speech data. This approach is robust to noise and other distortions, which is important for the data when the quality of the audio files is poor. In addition to the different architectures used in the two models, wav2vec 2.0 is designed to be trained on large amounts of unannotated audio files, while Whisper can be used on both annotated and unannotated data ('weakly supervised', see [8]). Both approaches have the potential to be used for low-resource languages such as Luxembourgish and achieve good performance with relatively little annotated data. Recently, a wav2vec 2.0 model for Luxembourgish has been presented by [9] [2], with results similar to the present study.

3. MATERIAL & METHOD

At the core of the ASRLux system under development at the University of Luxembourg is the training data, which consists of pairs of audio samples with their matching high-quality orthographic transcriptions. The 67 hours of training material come from several hundred speakers, thus representing a wide range of speaker characteristics and accents. Manual checks have been conducted to ensure that the audio and text match as precisely as possible. Table 1 provides information about the sources used.

	Datasets	**Hours**
1.	parliamentary speeches	32
2.	news commentaries	24
3.	crowd-sourced sentences [10]	9
4.	MaryLux sentences [11]	1
	Total duration	***67***

Table 1: Composition of the training data.

The data has been divided into smaller chunks of audio files with a duration ranging from 1 second to 20 seconds, with an average duration of 6 seconds (39724 chunks). Sources 2 and 3 are already available in the correct format, but larger audio files have been split into smaller chunks using a tool called MAUS (Munich Automatic Unit Segmentation), for which a reliable implementation for the Luxembourgish language is available [12]. The smaller chunks were created based on written transcripts using forced-alignment. Depending on the type of model, text normalization was applied to numbers and special characters like "%, &, +", which have been replaced by the corresponding words. Especially the parliamentary speeches and the news commentaries contain a lot of French and German words and phrases, which are then consequently included in the training.

As for the training itself, we have chosen recipes developed by Hugging Face, whose transformers framework ([13]) offers state-of-the-art scripts for the fine-tuning of wav2vec 2.0[3] and Whisper.[4] The dataset was split up into a training set of 75% and a test set of 15% and early stopping has been applied to circumvent overfitting. The fine-tuning of these highly parameterized models is still quite demanding regarding computing power. While the smaller wav2vec 2.0 checkpoint could be trained on four V100 GPUs on the University's HPC in approximately 30 hours, the large wav2vec 2.0 and

the Whisper-large checkpoints could be trained only on A4 GPUs with 46 GB of VRAM in a reasonable time.

To further improve the accuracy of the wav2vec 2.0 model, a language model has been compiled. The input data for the language model consists of text data from various sources (parliamentary debates and speeches, radio news articles and commentaries, Wikipedia, etc.) and sums up to approximately 130 million word tokens. By applying KenLM [14], the language model is constructed of n-grams with up to five words. The pyctcdecode library, a beam search decoder for CTC speech recognition, is used to attach the language model to the acoustic model. The weights of the n-grams then are used in the decoding pipeline to select certain words in cases when the probabilities of the acoustic recognition are not decisive. Using a language model improves the error rate by 2 to 5%. Note that Whisper does not currently support the attachment of a language model.

The quality of ASR systems is usually evaluated using the word error rate (WER), which measures the number of errors in the transcription, including substitutions (S), deletions (D), and insertions (I). The WER is calculated by summing up the three types of errors and dividing by the number of words (N) in the reference transcription. A lower WER indicates a higher-quality ASR system.

- $\text{WER} = \frac{\text{S+D+I}}{\text{N}} * 100$

WER can be used to compare the performance of different models and identify potential areas for improvement. In this project, we use n-gram stochastic language models with n equal to 5, considering phones as units.

4. RESULTS

We will report the results for four fine-tuned models here. Two models are based on the wav2vec 2.0 XLS-R checkpoints with 300 million and 1 billion parameters, respectively. For Whisper, we first experimented with the smaller Whisper checkpoints (small and medium), but after obtaining unsatisfactory results, we used the checkpoint Whisper-large with 1.5 billion parameters. Table 2 lists the WERs obtained for our four models and the original Whisper model. Based on the test dataset, the WERs all lie below 20 and range among rates for state-of-the-art systems for big languages. The 1 billion model for wav2vec 2.0 proves to be clearly superior to the 300 million parameter version. The additional use of a language model lowered the WER by a few points more, with a WER of 9.5 being the lowest for all models. The rates for Whisper are similar for the uncased version. For the cased version, the WER is considerably higher, which is most likely due to punctuation and capitalization issues.

Model	WER w/o language model	WER w/ language model
wav2vec 2.0 (300m)	19.1	14.5
wav2vec 2.0 (1b)	12	**9.5**
Whisper-large uncased	12.1	n.a.
Whisper-large-v2 cased	18.6	n.a.

Table 2: WERs for Luxembourgish ASR models trained with and without a language model.

To demonstrate the quality of the models, two longer examples are presented in table 3 along with their ground truth transcription. With incorrect recognized words in bold, it is obvious that both models make only a few errors, which are mostly related to rare words (e.g. *wellechen* 'some', *Ustiechlechkeet* 'contagiousness'), unstressed function words (e.g., confusion of *just* with *net*) or word endings (e.g., *nimmt* for *nimm* 'names', *bann* for *band* 'volume').

Contrary to wav2vec 2.0, the Whisper model works relatively well with multilingual speech. While the recognition of isolated French, German, or English sometimes leads to errors, longer passages in these languages (phrases, sentences) are rendered more or less correctly. The reason for this lies in Whisper's cross-lingual pre-training, which uses audio and text data and is able to manage multilingual input much better than wav2vec 2.0 Although the latter is trained on numerous languages, its multilingual capacities are linked only to the acoustic and not to the textual representation. The quality of the output thus depends on the training material used in fine-tuning. As a final example, an extract of a speech by the Grand-Duc of Luxembourg shows how well a longer stretch of French (in bold), inserted in his Luxembourgish speech, is recognized well by our Whisper model.

> *... Mir sollen houfreg sinn op d'Diversitéit an den Zesummenhalt an eise Gesellschaft.* ***A cet endroit, je voudrais remercier les non luxembourgeois qui resident ou qui travaillent à notre pays pour leur contribution précieuse à notre société. cohésion économique, mais aussi la cohésion sociale de notre pays sont des atouts qui nous appartiennent de défendre à tout prix. Ils sont au coeur de notre projet et de notre réussite. C' est notre bien commun à tous.*** *Haut, op dësem chrëschtel Wënd, wëll ech meng Unerkennung awer net nëmmen op de politesche Plang begrenzen. ...*

Wav2vec 2.0, on the other hand, is running into serious difficulties with this sample and is producing numerous errors.

One of the further advantages of Whisper's

Table 3: Examples for recognition: Ground truth and the results of wav2vec 2.0 and Whisper; recognition errors in bold face.

Ground truth	wav2vec 2.0 (1B)	Whisper-large-v2
Villmools merci, Här President. Den Avis vun den Experten huet kloer gewisen, datt d'Covid-Kris nach net eriwwer ass, dass de Risk nach ëmmer do ass an datt d'Expektative fir September, zumindest wat d'Experten ugeet, déi sinn, datt mer eventuell virun enger neier Well kënne stoen. Op wellechem Datum, wellech Variant, mat wellecher Virulenz, welleger Ustiechlegkeet, dat wësse mer an dësem Moment selbstverständlech net. Déi Zuelen, déi mer haut kennen, soen dat selwecht. De Staatsminister huet gëschter zitéiert: 1200 Infektiounen den Dag. Dat ass eppes, muss ech soen, wat eis virun enger Rei Joer, virun enger Rei Méint jo weesentlech méi erschreckt huet, wéi et eis haut erschreckt, well d'Situatioun	villmools merci här president den avis vun den experten huet kloer gewisen datt d'covidkris nach net eriwwer ass dass de risk nach ëmmer do ass an datt d'expektative fir september zumindest wat d'experten ugeet déi sinn datt mer eventuell virun enger neier well kënne stoen op wellechem datum **well ech** variant mat wellecher virulenz wellecher **ustieche keet** dat wësse mer an dësem moment selbstverständlech net déi zuelen déi mer haut kennen soen dat selwecht de staatsminister huet gëschter zitéiert den **auszweehonnert** infektiounen den dag dat ass eppes muss ech soen wat eis virun enger rei **t eis** haut erschreckt **war** d'situatioun	Villmools merci, Här President. Den Avis vun den Experten huet kloer gewisen, datt d'Covidkris nach net eriwwer ass, dass de Risk nach ëmmer do ass an datt d'Expektative fir September zumindest wat d'Experten ugeet, déi sinn, datt mer eventuell virun enger neier Well kënne stoen. Op **wellegem** Datum, **welleg** Variant, mat **welleger** Virulenz, welleger Ustiech**tegkeet**, dat wësse mer an dësem Moment selbstverständlech net. Déi Zuelen, déi mer haut kennen, soen dat selwecht. De Staatsminister huet gëschter zitéiert: 1200 Infektiounen den Dag. Dat ass eppes, muss ech soen, wat eis virun enger Rei Joer, virun enger Rei Méint jo weesentlech méi erschreckt huet, wéi et eis haut erschreckt, well d'Situatioun
Se si gesond, gesi schéin aus, schmaache gutt a schéin Nimm hunn se och nach! Ech schwätze vun den Uebst- a Geméiszorten, vun de Kraider, de Gewierzer an den Nëss. An dësem Buch huet den Zenter fir d'Lëtzebuerger Sprooch ronn 300 Nimm fir déi geleefegst Produite vun der Natur gesammelt. Dat ass natierlech just eng Selektioun, mee déi meescht Zorten, déi kann een iessen an déi et hei am Gaart oder am Buttek gëtt, fënnt een an dësem drëtte Band vun der beléifter Serie Lëtzebuerger Wuertschatz.	se si gesond gesi schéin a mech maachen a se nach ech schwätze vun den **uebstzorten deezer** an den nëss an dësem buch huet den zenter fir d'lëtzebuerger sprooch ronn dräihonnert **nimmt** fir déi **geleeft** produite vun der natur gesammelt dat ass natierlech just eng selektioun mee déi meescht zorten déi kann een iessen an déi **déi** et hei am **gaard uerder** am buttek gëtt fënnt een an dësem drëtte **bann** vun der beléifter **seng** lëtzebuerger **wirtschaft**	se si gesond, gesi schéin aus, schmaache gutt a schéin Nimm hu se och nach ech schwätze vun den Ueb**es** a Geméiszorte vun de Kräider, de Gewierzer an den Nëss**en** an dësem Buch huet de **Centre** fir d'Lëtzebuerger Sprooch ronn dräihonnert Nimm fir déi geleeftegst Produite vun der Natur gesammelt dat ass natierlech **net** eng Selektioun **wéi** déi meescht Zorten, déi kann een iessen an déi **déi** et hei am Gaart oder am Buttek gëtt, fënnt een an dësem drëtte Band vun der beléifter Serie Lëtzebuerger Wuertschatz.

pretraining on text can be seen in the integration of capitalization of nouns (Luxembourgish is here following the spelling rules of German) and of punctuation. In the Whisper examples above, the capitalization is achieved quite well, whereas the placement of period and comma is partially not yet correct. Wav2vec 2.0, on the other hand, capitalization, and punctuation have to be restored in an additional step, e.g. from the representation of a language model.

Finally, for audio recordings involving several speakers, the diarization method from pyannote.audio has been implemented [15].

5. OUTLOOK

From our evaluation and testing with various audio samples, we found that both systems produced high-quality results. The WER for all models lays clearly below 20, indicating a highly promising perspective for the ASR of Luxembourgish. Not only because of the restoration of capitalization and punctuation but mainly due to its generally better word recognition, it could be shown that Whisper and it is 'weakly supervised pretraining' outperform wav2vec 2.0.

In the next steps, it is foreseen to improve the quality and to enlarge the amount of training data, particularly by including more informal and conversational genres. Methods of audio data augmentation will be applied to make the system more robust for different levels of audio quality [16].

Regarding the applications of LuxASR, it is planned to develop an assistant system for transcribing the speeches and debates of the Luxembourgish parliament. LuxASR will also be used to extract text data from audio archives to assist in research projects in the Humanities and to develop tools for pronunciation assessment.

LuxASR can be tested for inference online on Hugging Face.[5]

6. REFERENCES

[1] P. Gilles, "Luxembourgish," in *Oxford Encyclopedia of Germanic Linguistics*, S. Kürschner and A. Dammel, Eds. Oxford: Oxford University Press, 2023.

[2] P. Gilles and J. Trouvain, "Illustrations of the ipa: Luxembourgish," *Journal of the International Phonetic Association*, vol. 43, no. 1, pp. 67–74, 2013.

[3] M. Adda-Decker, L. Lamel, and N. D. Snoeren, "Studying luxembourgish phonetics via multilingual forced alignments." in *ICPhS*, vol. 11, 2011, pp. 196–199.

[4] M. Adda-Decker, L. Lamel, and G. Adda,

"Speech alignment and recognition experiments for luxembourgish," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[5] K. Veselỳ, C. Segura, I. Szöke, J. Luque, and J. Cernockỳ, "Lightly supervised vs. semi-supervised training of acoustic model on luxembourgish for low-resource automatic speech recognition." in *INTERSPEECH*, 2018, pp. 2883–2887.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[8] J. C. Vásquez-Correa and A. Álvarez, "Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper," ENGINEERING, Preprint, Dec. 2022.

[9] L. M. Nguyen, "Improving luxembourgish speech recognition with cross-lingual speech representations," Master's thesis, Master thesis, Voice Technology (VT), University of Groningen, 2022.

[10] N. Entringer, P. Gilles, S. Martin, and C. Purschke, "Schnëssen. surveying language dynamics in luxembourgish with a mobile research app," *Linguistics Vanguard*, vol. 7, no. s1, 2021.

[11] M. Barnig, "Marylux-648-TTS-Corpus," 2022.

[12] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[14] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[15] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[16] I. Jordal, A. Tamazian, E. T. Chourdakis, C. Angonin, Askskro, N. Karpov, T. Dhyani, O. Sarioglu, Kvilouras, E. B. Çoban, F. Mirus, Jeong-Yoon Lee, K. Choi, MarvinLvn, SolomidHero, and T. Alumäe, "Audiomentations," Zenodo, Aug. 2022.

[1] http://kaldi-asr.org/doc/

[2] https://schreifmaschinn.lu

[3] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[4] https://huggingface.co/blog/fine-tune-whisper

[5] https://huggingface.co/unilux

USING TACOTRON 2 TO DEVELOP SYNTHETIC CHILD SPEECH IN SOUTH AFRICAN ENGLISH

Camryn Terblanche[1], Benjamin V Tucker[2], Tyler Schnoor[3] and Michal Harty[1]
[1]*University of Cape Town, South Africa,* [2]*Northern Arizona University, United States,*
[3]*University of Alberta, Canada*

trbcam001@myuct.ac.za/ benjamin.tucker@nau.edu/ tschnoor@ualberta.ca/ michal.harty@uct.ac.za

ABSTRACT

There is often a substantial mismatch between the synthetic voices used on augmentative and alternative communication devices, and the child user's natural voice in relation to age, gender, dialect, and language. This study aims to determine if it is possible to develop realistic child speech synthesis, using Tacotron 2, for South African English. Two hours of child speech data were manually collected from one 11-year-old male child. Following this, two existing adult models were used to "warm start" the child speech synthesis. Despite the limited child speech data, we were able to successfully create a synthesised child voice of adequate quality. When a warm start training procedure is utilised, even when there is an age and dialect mismatch, the amount of training time decreases, and the synthesised voice matches the vocal quality of the child donor's voice.

Keywords: augmentative and alternative communication (AAC), children, speech synthesis, Tacotron, text-to-speech.

1. INTRODUCTION

Children with complex communication needs do not develop their speech, language, and/or communication skills in a typical pattern, most often as a result of: a) neurological disorders (e.g., intellectual disability), b) genetic disorders (e.g., Down syndrome), or c) structural abnormalities (e.g., cleft palate) [1]. When natural speech is restricted, augmentative and alternative communication (AAC) can assist individuals with complex communication needs establish functional communication skills. AAC includes techniques, strategies, and pictorial or written symbols. Using either low- or high-tech AAC options, it can be used to supplement or if needed, replace an individual's natural speech [2]. Low-tech AAC is made up of basic communication devices (which are often paper-based) whilst high-tech AAC involves speech-generating devices such as: mobile, computer, and/or tablet-based technologies [2]. Due to advancements in technology, high-tech AAC has improved considerably in recent years. There are many commercially available speech-generating devices and mobile applications for children. However, the speech output on these devices does not always match the age, gender, personality [3], dialect, and language of the child user [4]. Tönsing *et al.* [4] state that speech-generating devices in South Africa are usually provided in English dialects that are not necessarily reflective of South African English children's speech. As AAC research and the technology developments are frequently conducted in high-income, mostly English-speaking countries, it is not surprising that US-accented English, is most often incorporated in commonly used speech-generating devices [4], [5]. It is not unusual to walk into a South African special needs classroom and see all the children making use of the same adult US-English voice. These children may have a way to communicate, but with a voice that doesn't reflect their linguistic and cultural diversity. Although an individual's right to communicate is often discussed in AAC research, the rights of individuals using AAC to communicate in whichever dialect or language they choose, has not received equal attention [4]. Thus, this study aims to determine if it is possible to develop realistic child speech synthesis, using Tacotron 2, for South African English (SAE).

Although South Africa's official language policy allows schools to select any of the 11 official languages for teaching and learning, English in education, rather than one of the African languages, is often favoured by the community [6]. Unfortunately, this means that children using a speech-generating device at school are likely using their second or third language to communicate [4]. Despite the obvious language barrier, these children also have to use synthetic voices that do not necessarily match their age, gender, or dialect. Although this is an established problem, personalised synthetic speech software is not without cost [3], [7]. Until AAC applications meet the communication needs of the child user, AAC use will likely remain limited, resulting in fewer opportunities for participation and interaction. Despite this, producing realistic synthetic child speech from text is challenging and this is largely due to the scarcity of usable child speech corpora. Further challenges are

experienced when collecting child speech data [8]. For instance, in comparison to adult speech, children's read speech is typically less fluent, their speech often includes multiple articulatory errors and as the recordings are usually conducted in schools rather than sound-attenuated booths, background noise is common [8]. When speech data is however available, Tacotron 2 [9], an open-source speech synthesis system, can produce natural synthetic speech, with high similarity to human speakers [10]. Tacotron 2 is an end-to-end neural network-based text-to-speech system that can be trained on text-to-audio pairs, without phonetic annotation. The system is also user-friendly, which means that it can be used by individuals with limited text-to-speech experience. Additionally, as Tacotron 2 allows for rich conditioning of attributes, such as speaker and language, data adaptation is possible [9].

2. METHOD

One 11-year-old typically developing male child was recruited to record a total of two hours of read speech in SAE. Each recording session was thirty minutes long, with breaks every 10-15min. Using a Zoom H1 Handy recorder (44100 Hz), the recordings were collected in a repurposed classroom. The data were manually marked up into short utterance chunks of ≤13 seconds, using a Praat textgrid [11]. If audio files are ≤13 seconds, Tacotron 2 trains relatively quickly. Fluent speech was required, so all false starts, disfluencies and misarticulations were removed. After data cleaning, there was 113.7min of speech that remained. The sound files were extracted from the textgrids, changed to mono, and downsampled to 22050 Hz. The data were then randomly divided into training (90%) and validation files (10%).

The default Tacotron 2 [9] architecture was used to create the child speech synthesis. Tacotron 2 [9] is made up of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms. A "warm start" training procedure was implemented, meaning that the child speech data was trained over a pre-existing model. Essentially, the learned voice from the pre-existing model is removed, while the linguistic characteristics of that voice remain intact. Therefore, as linguistic characteristics from the pre-existing model can be transferred to other speakers, the new model's time until convergence is reduced, which makes it particularly useful when data is limited. In the current study, two differing warm start models were compared. In *warm start A*, the published pre-trained Tacotron 2 model (tacotron2_statedict.pt) from NVIDIA, trained on the LJ dataset [12] was used. The LJ dataset consists of short recordings from one female adult North American English speaker. In total, the LJ dataset is approximately 24 hours long [12]. For *warm start B*, the child model was trained on the cleaned, resampled (22050 Hz) SAE Lwazi III text-to-speech dataset [13]. The Lwazi III dataset is made up of short and long audio clips from one female adult SAE speaker. Although the SAE dataset contains mostly SAE, there are also small amounts of additional language data. This provides phone coverage of other languages [13]. Specifically, it contained 6.5 hours of SAE, but also contained Afrikaans (2 min), isiZulu (10 min), isiXhosa (12 min), Sepedi (5 min) and Setswana (4 min) data.

Figure 1 illustrates the process used to generate the synthesis systems for the SAE child speech. As an adult SAE text-to-speech model for Tacotron 2 isn't available for download, this first had to be created. During training, the full Lwazi III dataset, including the additional language data, was used. It was theorised that the additional phone coverage may improve the pronunciation of non-English names, surnames, street and building names. Additionally, due to the frequency of loan words between languages, it was suspected that improved pronunciation may be achieved if additional language data were included. Due to the comparably restricted Lwazi III dataset, a warm start was implemented using the pre-existing Tacotron 2 model from NVIDIA. After manually cleaning the adult data and segmenting longer recordings, training took approximately 5 days. Following the creation of the SAE adult synthesis, the respective SAE adult model was used to warm start the SAE child speech synthesis. Using the generated mel-scale spectrograms, the published WaveGlow [14] model was then used as a vocoder to synthesise time-domain waveforms. WaveGlow is a flow-based generative speech synthesis program [14].

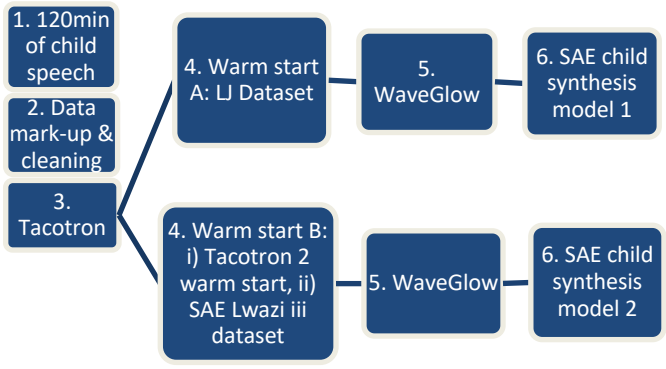


Figure 1: Process used for generating the synthesis systems for SAE child speech.

A further denoising step was included using a denoising code from the WaveGlow repository, which reduced background noise and eliminated

unwanted artefacts. In the end, naturalistic speech was created through a combination of Tacotron's prosody and WaveGlow's audio quality.

3. RESULTS

Encouragingly, despite the limited child speech data, we were able to successfully create a synthesised child voice of adequate quality in SAE. No matter which model was used to warm start the child synthesis, it took approximately 72 hours. It should be noted that the synthesised SAE adult voice was considerably less noisy than the two synthesised SAE child voices, but the additional language data used in the adult synthesis did not improve the pronunciation of non-English names and loan words to a great degree. Moreover, using a 5-point Likert Scale, where 0 was completely unnatural and 4 was completely natural, a mean opinion score method was used to gather 111 SAE listeners' (1998 responses) perceptions of the naturalness of the synthetic voices created. Results show that the adult speech ($\bar{x}$=3.38) was judged as more natural than the child speech ($\bar{x}$=2.20). Child synthesis model 2/warm start B ($\bar{x}$=2.33) was judged as more natural than model 1/warm start A ($\bar{x}$=2.06). The accent and vocal characteristics are comparable to the South African child donor, rather than the adult voice, no matter the adult model used. Figure 2 shows a spectrographic comparison between the donor child's speech and the speech output from the two SAE child models.

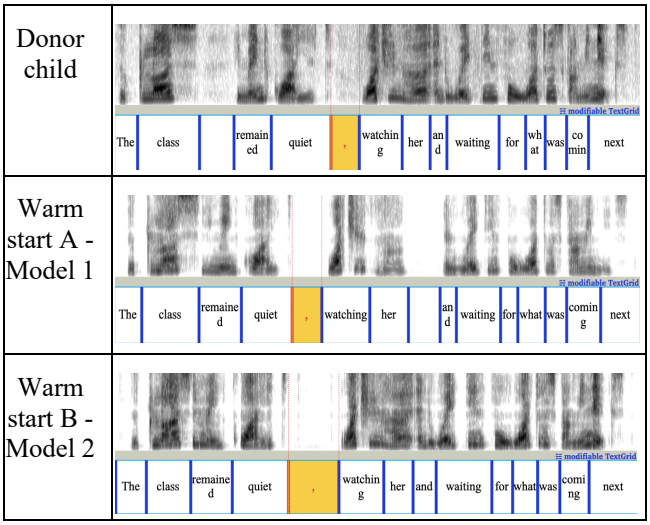


Figure 2: Spectrogram comparison between the SAE donor child and the synthesised child audio.

Alternatively, Figure 3 shows the mel-spectrograms and alignment plots for the two child models. The alignment plot is a useful tool to visualise a model's success. A straight diagonal line from the bottom left to the top right is a good indicator that the model is producing something speech-like. The pacing is largely consistent, but there are occasional arbitrary pauses. The results also show that Tacotron 2 trains well with punctuation. E.g., a pause in speech due to a comma, is highlighted in Figure 2, and circled in Figure 3. However, warm start A produced some irregularities. E.g., the word *class* has a tin-like voice quality (arrow).

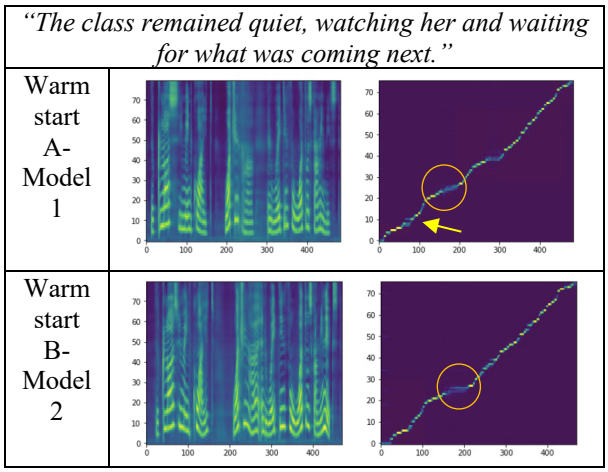


Figure 3: Comparison of the Tacotron 2 mel-spectrogram (a) and alignment (b) plots of the SAE synthesised child speech.

Model consistency was also considered. Figure 4 shows alignment plots between the two SAE child models (warm start A vs warm start B), when three sentences of differing lengths are synthesised.

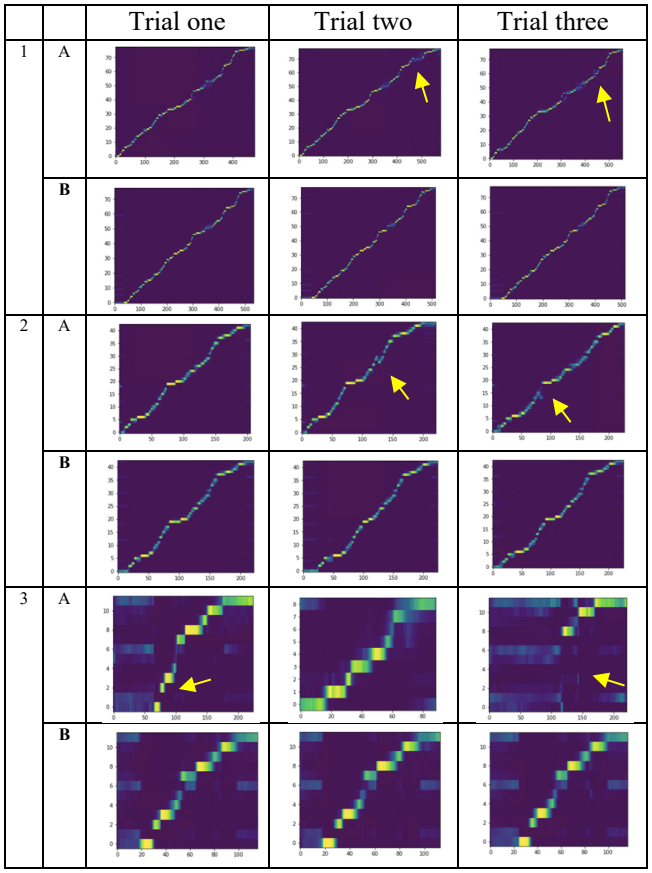


Figure 4: Alignment plots comparing the consistency of the two SAE synthesised child speech models.

It appears that the SAE child model is more consistent when warm start B is used, no matter the length of the sentence. In Figure 4, sentence 1 is long, *"Every day she woke up with a big smile on her face, because she loved her job."* Sentence 2 is medium length, *"They covered their mouths with their hands."* Sentence 3 is short, *"I like dogs."* The voices are mostly clear, but longer sentences produce better quality synthesis, as opposed to single-word or short sentences. Similarly to findings from Jain *et al.* [15], the first and last words in the phrases were more likely subject to distortions and artefacts, as compared to the middle of the phrase. The speaker's 'breathing' remained in the training data, which transferred to the synthetic models, making them appear more realistic.

4. DISCUSSION

Despite the occasional pronunciation and prosodic irregularity, adequate quality child speech synthesis was created with limited data. Both synthesised voices appear to sound like a relatively fluent South African child. Although the voices have minor distortions and some noise, the voices appear natural. As typically developing children are often hesitant when starting a phrase and may wander towards the end, it is unsurprising that these characteristics occasionally manifested in the synthetic model [15]. In fact, these characteristics likely contribute to greater naturalism as a typical child's speech patterns are less fluent, with variations in volume, pacing and emotional expressivity [15].

Firstly, it is effective to use a pre-trained North American English model to warm start a model in another English dialect. Secondly, using a warm start can greatly reduce the model's training time. This was observed when creating the SAE adult speech synthesis, and is in line with previous research [16], [17]. A warm start, using the Tacotron 2 model was utilised to create the SAE adult voice. The final SAE adult model (warm start B) was then utilised to warm start the SAE child synthesis. It was anticipated that including a warm start procedure twice, may result in the child model underperforming. Interestingly, the child model performed better with warm start B, suggesting that more than one warm start does not necessarily affect performance. Rather, the child model performance is directly proportional to the quality and the quantity of the adaptation data used. This is also true for the child speech data. In comparison, the adult model produced better synthesis quality, due to the data used. Although this study showed that only 113min of child data is sufficient, it is anticipated that if the donor child's speech were to be recorded in a sound-attenuated room, for an extended period, one would get improved child speech synthesis results.

Thirdly, the length of the text affects the speech output. Practically, this may affect the intelligibility of speech output on AAC devices. Drager [18] suggests that the context and length of the utterance play a role in the intelligibility of synthesised speech. Shorter sentences and single-word utterances are often less intelligible to the listener. Unfortunately, using a single-word AAC device is very common for new AAC users, and communication partners may find it difficult to understand the synthesised speech output, unless context is given. Fourthly, it is easier to find a high-quality adult corpus. This study has shown that by adapting adult data, using a warm start, one can successfully create a child voice that matches the vocal characteristics of the child donor. Thus, when there are limited training data, which often occurs with child speech data, along with possible computational resource constraints, incorporating a warm start, with an established model of high quality, can improve the quality of the synthesised child speech output. Although both warm start methods produced adequate quality child speech synthesis, if one were to use the voices for AAC purposes, it would be sensible to a choose a model with a consistent speech output. The fact that a warm start can be used when training data is limited, is beneficial for individuals who have complex communication needs. In the future, clinicians will be able to either collect the child's residual speech and incorporate it into the training model, or if speech is severely impaired, they will be able to use an age-matched typically developing child's voice. As Tacotron 2 [9] is open-source and produces naturalistic synthetic speech, the lack of extensive speech data should no longer limit the development of appropriate child voices. This technology can support marginalised AAC communities and develop much-needed resources in under-resourced languages.

Using adult adaptation data (warm start) to create a child voice is a viable method and could open the door for children with complex communication needs to have linguistically and culturally appropriate synthetic voices. As we successfully created a SAE child voice, future research should focus on creating child voices in other South African languages. It would also be interesting to determine if the residual speech of children with complex communication needs could be incorporated, so that the voice is a closer approximation to the child's natural voice [3]. Lastly, having children with complex communication needs select which voice they would prefer to use (i.e., an adult, a child, or a child voice that includes some of their own speech) is of interest to the authors.

5. ACKNOWLEDGMENTS

The financial assistance of the University of Cape Town and the National Research Foundation (NRF) of South Africa (DSINRF Reference Number: MND200619533947) is gratefully acknowledged. Additionally, this work was supported by Mitacs through the Mitacs Globalink Research Award Program. We also acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

6. REFERENCES

[1] M. Boesch and M. Da Fonte, *Effective Augmentative and Alternative Communication Practices: A Handbook for School-Based Practitioners*. Routledge, 2019.

[2] Y. Elsahar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor, 'Augmentative and Alternative Communication (AAC) advances: A review of configurations for individuals with a speech disability', *Sensors*, vol. 19, 2019.

[3] T. Mills, T. Bunnell, and R. Patel, 'Towards personalized speech synthesis for augmentative and alternative communication', *Augmentative and Alternative Communication*, vol. 30, no. 3, pp. 226–236, 2014.

[4] K. Tönsing, K. van Niekerk, G. Schlünz, and I. Wilken, 'Multilingualism and augmentative and alternative communication in South Africa – Exploring the views of persons with complex communication needs', *African Journal of Disability*, vol. 8, no. 0, 2019.

[5] C. Terblanche, M. Harty, M. Pascoe, and B. V. Tucker, 'A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence', *Applied Sciences*, vol. 12, no. 5623, pp. 1–17, 2022.

[6] South African Government, 'Language in Education Policy Document (LiEP)'. Pretoria: Government Printers, 1997. Accessed: Oct. 21, 2022. [Online]. Available: https://www.education.gov.za/Portals/0/Documents/Policies/GET/LanguageEducationPolicy1997.pdf

[7] C. Jreige, R. Patel, and H. T. Bunnell, 'VocaliD: Personalizing Text-to-Speech Synthesis for Individuals with Severe Speech Impairment', *Assets '09*, pp. 259–260, 2009.

[8] A. Govender, B. Nouhou, and F. De Wet, 'HMM Adaptation for child speech synthesis using ASR data', presented at the 2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Port Elizabeth, South Africa, 2015.

[9] Y. Wang *et al.*, 'Tacotron: Towards end-to-end speech synthesis', *arXiv*, 2017, [Online]. Available: https://doi.org/10.48550/arXiv.1703.10135

[10] X. Wang *et al.*, 'ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech', *arXiv*, 2020, doi: https://doi.org/10.48550/arXiv.1911.01601.

[11] P. Boersma and D. Weenink, 'Praat: Doing phonetics by computer [Computer program]'. 2022. [Online]. Available: http://www.praat.org/

[12] K. Ito and L. Johnson, 'The LJ speech dataset', 2017, [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[13] A. Louw and G. Schlünz, 'Lwazi III English TTS Corpus', vol. 1. Meraka Institute, CSIR, 2016. [Online]. Available: https://hdl.handle.net/20.500.12185/267

[14] R. Prenger, R. Valle, and B. Catanzaro, 'Waveglow: A flow-based generative network for speech synthesis', *arXiv*, 2018, [Online]. Available: https://doi.org/10.48550/arXiv.1811.00002

[15] R. Jain, M. Yiwere, Bigioi, P. Corcoran, and H. Cucu, 'A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis', *arXiv*, 2022, doi: https://doi.org/10.48550/arXiv.2203.11562.

[16] E. Cooper *et al.*, 'Zero-shot multi-speaker text-to-speech with state of the art neural speaker embeddings', presented at the IEEE (4-8 May 2020), 2020.

[17] V. Barnekow, D. Binder, N. Kromrey, P. Munaretto, A. Schaad, and F. Schmieder, 'Creation and detection of German voice deepfakes', presented at the International Symposium on Foundations and Practice of Security, 2021, pp. 355–364.

[18] K. Drager, J. Reichle, and C. Pinkoski, 'Synthesized Speech Output and Children: A Scoping Review', *American Journal of Speech-Language Pathology*, vol. 19, pp. 259–273, 2010.

USING WORD-LEVEL FEATURES FOR PROSODIC PROMINENCE DETECTION IN CONVERSATIONAL SPEECH

Julian Linke[1], Gernot Kubin[1], Barbara Schuppler[1]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[1]{linke, gernot.kubin, b.schuppler}@tugraz.at

ABSTRACT

This paper focuses on the automatic detection of prominent words in conversational speech. Most tools for prominence detection rely on prosodic features extracted at a syllable- or phone level and their accuracy thus strongly depends on the quality of the given phone-level segmentation. Given the high degree of pronunciation variation in conversational speech, automatic phonetic segmentation is not accurate enough to detect prominence reliably. Here we explore different approaches to prominence detection that require merely a prior word-level segmentation. The first experiment shows that by using word-level prosodic features cross-validation accuracies of $88\% \pm 4\%$ can be reached, and that word duration is the most important feature. The second experiment introduces entropy-based fundamental frequency and intensity features for prominence detection. Our findings suggest that entropy-based, word-level features can provide a robust approach to detecting prominent words in conversational speech.

Keywords: prosodic prominence, conversational speech, entropy, Austrian German

1. INTRODUCTION

Many tools to detect prosodic prominence (e.g., ProsAlign [1], Tamburini and Wagner [2] or Mishra et al. [3]) operate with features which rely on syllabic structure and/or phonetic annotations. These tools are motivated by linguistic studies that found that certain syllable-level characteristics of fundamental frequency (F0), intensity (RMS) and duration (e.g., speech rate, stressed syllable duration) are cues to perceived prominence, where the conclusions drawn about these cues' importance are not uniform: [4] showed that vowel duration is a more important cue for prominence perception than RMS, whereas [5, 6] found F0 to be a more important cue for prominence than syllable duration. [7] reported that RMS is a more important cue of perceived prominence than other continuous-valued prosodic variables [7]. In the mentioned studies, extracted durational features are related to vowel or syllable durations (e.g., [4, 7]). In contrast, [8] found word duration to be the strongest cue to prominence.

Given the high degree of reduction and overlapping sounds in conversational speech [9], the quality of automatically created phone segmentations has a lower accuracy than for read speech. [10] showed that, whereas for read speech, automatic phonetic annotations are good enough for a subsequent automatic prosodic boundary detection, for conversational speech, manual correction is required. In order to avoid the necessity of manual correction, we explore the use of word-level prosodic features, as the automatic detection of word boundaries showed to have a higher precision for spontaneous speaking styles than of phone boundaries.

So far, most studies on prosodic prominence analyzing F0/RMS contours considered features related to specific characteristics of those curves (i.e., mean, maximum, etc.). To the best of our knowledge, it has not yet been investigated whether entropy-based F0/RMS features, which directly relate to their distribution distinguishing prominence levels. In general, entropy-based features have broadly been used in speech science: For instance, [11] use relative entropy to measure the distance between two speech spectral distributions in concatenative synthesis applications whereas [12] showed that spectral entropy features which interpret the spectrum as a probability mass function, improved the performance of an automatic speech recognition (ASR) system. A study on voice signal characterization tested entropy measures coming from raw audio signals in order to extend voice analysis methods [13]. With respect to prosody of emotional expressions, it has been shown that the use of features capturing F0/RMS variability by calculating entropy from F0/RMS-curves helps to distinguish between arousal conditions in a free-speech setting [14].

This paper presents two approaches for automatic prominence detection in conversational Austrian German. The first experiment uses traditional F0/RMS features (i.e., maximum, mean, etc.) extracted a the word level and durational features (i.e.,

word duration and speech rate variation). The second experiment explores a new approach, i.e., the use of entropy-based F0/RMS features, which implicitly encode distribution information and thus do not require phone-level information.

2. MATERIALS AND METHODS

2.1. GRASS Corpus

The *Graz Corpus of Read and Spontaneous Speech* (GRASS) [15] contains Austrian German conversational speech from 38 Austrian speakers, containing a total of approx. 20h of speech. Word- and phone level segmentations were created by means of a forced alignment using a Kaldi-based ASR system with a lexicon containing on average $5.57-6.18$ pronunciation variants per word type [16]. Phonetically trained transcribers created prosodic annotations for a total of 5234 word tokens from 34 speakers of GRASS. The prominence annotations distinguished the prominence levels 0 (no prominence, *PL0*), 1 (weak prominence, *PL1*), 2 (strong prominence) and 3 (emphatic prominence) [17]. In this study, we combined prominence levels 2 and 3, as emphatic prominence occurred rarely (*PL2*). Annotations were created in three stages: One annotator created a first version, which later was corrected by her/him and subsequently corrected by one of the other annotators. Based on a small subset annotated by two different annotators in those stages, the inter-annotator agreement was calculated: The overall Cohen's kappa was 0.72 (598 tokens), 0.72 for level 0 vs. 1 (371 tokens), 0.92 for level 0 vs. $2/3$ (446 tokens) and 0.57 for level 1 vs. $2/3$ (275 tokens). Other studies obtained similar agreements of 0.53 [2] or 0.84 [7].

2.2. Basic Prosodic Features

84 F0 and RMS features: All features were calculated at the word level. We calculated F0 with the library AMFM decompy [18] which includes an implementation of the pitch detection algorithm YAAPT [19]. Intensity features were generated directly from the waveform by calculating the root mean square. For F0/RMS, and their respective first and second derivatives, we extracted 10 measurements: maximum, minimum, range, relative position of maximum and minimum in the word, mean, median, first and third quartile and standard deviation (60 features). Additionally, we extracted left and right slope of the maximum and minimum, absolute and relative onset and offset within the word, as well as the maximum, minimum, range and mean relative to the utterance (24 features).

12 Durational features (DUR): We extracted word duration, phrase-level speech rate (i.e., number of segments per phrase), local speech rate (i.e., the number of segments per word duration), and relative speech rates (i.e., the ratio of the local speech rate and the minimum, maximum or median of local speech rates within a phrase). Additionally, we calculated the minimum, maximum, range, mean, median and standard deviation of local speech rates within a phrase.

2.3. Entropy-based Features

Entropy measures the spread of probability distributions and provides a measure of uncertainty of a random variable X [20]. If the random variable X assumes values $x_i \in \mathscr{X}$ where $\mathscr{X}$ is a finite set, the definition of entropy can be stated as

$$(1) \quad H(X) = -\sum_i p_i \cdot \log p_i,$$

where $p_i = Pr\{X = x_i\}$ describes the probability of X taking the value x_i, assuming that $p_i \cdot \log p_i = 0$ for $p_i = 0$.

If we observe a sequence of N (non-negative) feature values $\langle f[1], f[2], \ldots, f[N] \rangle$ within a given word, we can measure the spread of these values also by a formal entropy where the (pseudo-)probability distribution is defined by normalizing the feature values

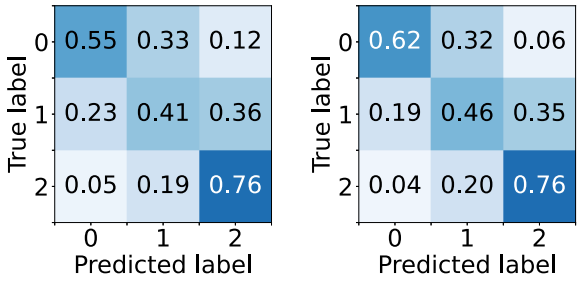
$$(2) \quad p_i = \frac{f[i]}{\sum_{i=1}^{N} f[i]}$$

such that the condition for the total probability is fulfilled: $\sum_{i=1}^{N} p_i = 1$.

With this definition, the entropy (1) achieves its maximum $H_{max} = \log N$ if the feature sequence is constant $f[1] = f[2] = \cdots = f[N] = \text{const.}$, and its minimum $H_{min} = 0$ if all probabilities according to equation (2) turn out to be close to either 1 or 0, e.g., for a very non-uniform feature sequence within the given word. Note that this entropy measures the (relative) feature variability within the word, but without accounting for the time order of the feature contour. Finally, we also experimented with a normalized entropy $\tilde{H}$ obtained from division by the sequence length N:

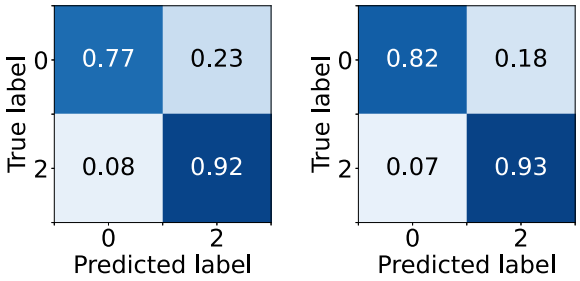
$$(3) \quad \tilde{H} = \frac{H}{N}.$$

For our experiments, we applied equation (2) to the extracted F0/RMS contours and calculated four



(a) F0+RMS **(b)** F0+RMS+DUR

Figure 1: Confusion matrices (3 classes) with F0 and RMS features **(a)** and all features **(b)**.



(a) F0+RMS **(b)** F0+RMS+DUR

Figure 2: Confusion matrices (2 classes) with F0 and RMS features **(a)** and all features **(b)**

additional entropy-based features (ENT) with equations (1) and (3) leading to two (pseudo-)entropies H and two normalized (pseudo-)entropies $\tilde{H}$ of F0/RMS.

Simulations with uniform and non-uniform distributions indicated that these entropy-based features depend primarily on the number of possible outcomes N, which in our case corresponds to word duration. Nevertheless, for words of similar lengths these measurements also encode contour variations by capturing deviations from uniform distributions.

2.4. Random Forest

We trained Random Forest classifiers (RFCs) with the *scikit learn* toolkit (version 0.21.3) [21] RFCs were built with 100 estimators, default maximum depth, a minimum samples split of 2 and the Gini impurity for measuring the quality of a split. For each of the different feature sets, we present results from two conditions, one for 2 classes (*PL0* vs. *PL2*), and one for 3 classes (*PL0* vs. *PL1* vs. *PL2*). Each classification experiment involved two steps: First, RFCs were trained with the entire feature set in order to learn about the feature's relative importance. Second, a (final) RFC was trained with the 15 most important features as given by the first step. The training and test sets were based on a random 80/20 split and we present associated F1-scores. Additionally, we provide means and standard deviations of accuracies resulting from 10-fold cross-validation experiments in order to estimate the model's generalization ability.

3. THE ROLE OF DURATIONAL FEATURES

In order to learn about the role of durational features for prominence detection, we conducted two classification experiments. While the first RFC uses all 96 F0, RMS and durational features (F0+RMS+DUR) described in Sec. 2.2, the second RFC uses the 84 F0/RMS related features only.

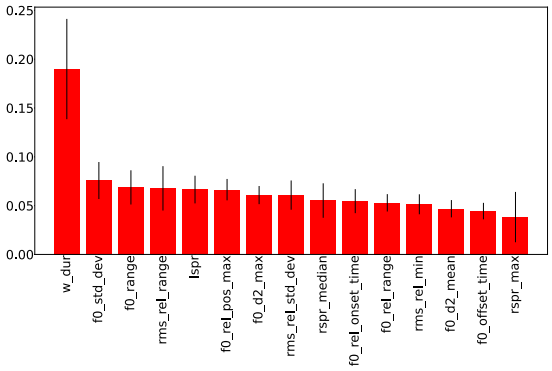


Figure 3: RFC feature importances for 3 class problem with F0, RMS and DUR features.

Fig. 1 and Fig. 2 show the confusion matrices of RFCs which were trained on the entire basic feature set (F0+RMS+DUR) and on a subset without durational features (F0+RMS). We observe that classification performance between non-prominent and highly prominent words is high in both cases, but that non-prominent words were better classified when the RFC was trained with the entire feature set than without DUR (recall 82% vs. 77%). For *PL0* the F1 increased form 80.3% to 84.4% by adding DUR, and for *PL2* from 89.8% to 91.6%. Corresponding cross-validation accuracies were 88% $\pm$ 4% (F0+RMS+DUR) and 85% $\pm$ 4% (F0+RMS).

RFCs with 3 classes (Fig. 1) showed a similar behaviour since the recall for *PL0* of 61.6% (F0+RMS+DUR) was higher than the recall of 55.2% (F0+RMS). However, in case of *PL1*, recalls of only 45.7% (F0+RMS+DUR) and 40.6% (F0+RMS) were achieved, while in both cases approx. 35% of tokens from *PL1* were predicted as *PL2*. Respective F1s of *PL0*/*PL1* were 63.1%/47.2% (F0+RMS+DUR) and 55.9%/43.2% (F0+RMS). Interestingly, recalls of highly prominent words were similar in both cases (approx. 76%). In this case, cross-validation accuracies were 63% $\pm$ 5% (F0+RMS+DUR) and 60% $\pm$ 6% (F0+RMS). Fig. 3 shows the feature ranking corre-

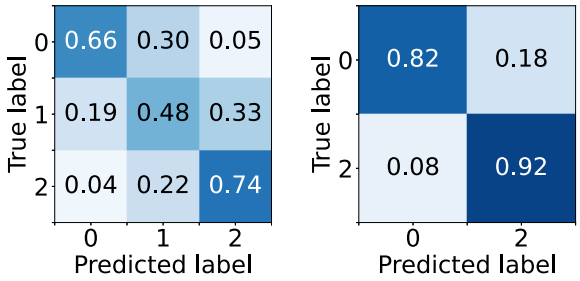


Figure 4: Confusion matrices from experiments with 15 best features (F0+RMS+DUR+ENT).

sponding to the averaged impurity decrease of the RFC for 3 classes trained with the 15 best features (F0+RMS+DUR). Word duration (w_dur) has by far the highest importance among all features, capturing almost 20% of the overall importance. Other durational features (see Sec. 2.2) like the local speech rate (lspr) or relative speech rates (rspr_median and rspr_max) were also present in the feature ranking and had similar importances as the relative relationships of the F0/RMS contours. This finding is in line with the study by [8], showing that word-duration is the most important feature for prominence detection in read speech.

Overall, durational features improve the RFC accuracy for detecting prominent words of conversational speech. Results from previous investigations pointed towards different trends. Whereas [4] found vowel duration to be an important cue to prominence in spontaneous speech, [6] concluded that F0 and [7] that RMS play a more important role. These studies, however, did not consider word duration, which in our experiments resulted to be the most important feature among all durational, F0/RMS features to classify prominence in conversational Austrian German.

4. THE ROLE OF ENTROPY-BASED FEATURES

To learn about the role of entropy-based features (ENT) for prominence detection, and whether they can complement phone-based durational features, we conducted two classification experiments. While the first RFC uses all 100 features (F0+RMS+DUR+ENT), the second RFC does not use any durational features (F0+RMS+ENT).

The RFC for 3 classes with F0+RMS+DUR+ENT features resulted in a large number of confusions of *PL1* (recall/F1: 48%/48.6%) with *PL0* or *PL2*, where approx. 19% of *PL1* was classified as *PL0* and 33% as *PL2*. In contrast, recalls/F1s of 65.6%/66.1% (*PL0*) and 74.1%/73.1% (*PL2*) indicated less confusions with others classes (i.e., only $4-5\%$ of non-prominent or highly-prominent words were classified as highly-prominent or non-prominent words). This result is to be expected, as also the inter-rater agreement showed to be lowest/highest for these classes. Overall cross-validation accuracies reached $62\% \pm 7\%$. Compared to the RFC without ENT, the classification of non-prominent words improves by adding the ENT features (recall $66\% > 62\%$). Furthermore, the comparison with the RFC trained without any durational features (F0+RMS+ENT) indicated that developed entropy-based features compensate for durational information (similar F1s for classes *PL0*/*PL2* of approx. 84%/91%). With respect to the feature importances for the RFC with F0+RMS+DUR+ENT, we observed that the 5 best features comprised word duration as well as the entropy and normalized entropy features, which all had average importances of $> 8\%$ (capturing approx. 45% of the overall importance), while all other features had importances $< 6.7\%$.

For both the 2 and the 3 class problem, prominence detection was best when adding entropy-based F0/RMS-features to the feature set. To the best of our knowledge, there exist no earlier studies on prominence detection using similar entropy-based F0 and RMS features.

5. CONCLUSION

This paper investigated different word-level features to detect prosodic prominence, to avoid the necessity of creating manual phonetic segmentations for conversational speech. Overall, the classification performances achieved with our different sets of features were in the range of the human inter-rater agreements for the respective classes. We found that durational features (incl. speech rate variations) have a higher importance than F0/RMS features, and that among them, word duration is by far the most important feature. Experiments with entropy-based F0/RMS features showed that they encode necessary durational information along with information about the features' distribution, making them useful for classifying prominence levels in conversational speech. In future, we will explore whether entropy-based F0/RMS features are also useful to capture other prosodic characteristics in speech, both with respect to speech analysis as well as in ASR.

6. ACKNOWLEDGEMENTS

This work was partly funded by grant P-32700-NB from FWF (Austrian Science Fund). We would like to thank the transcribers Nina Richter and Nikolaus Tlapak for their efforts.

7. REFERENCES

[1] Braunschweiler, N. 2003. ProsAlign - The Automatic Prosodic Aligner. *Proc. of ICPhS*, 3093–3096.

[2] Tamburini, F., Wagner, P. 2007. On automatic prominence detection for German. *Proc. of Interspeech*, 1809–1812.

[3] Mishra, T., Sridhar, V. R., Conkie, A. 2012. Word prominence detection using robust yet simple prosodic features. *Proc. Interspeech 2012*, 1864–1867.

[4] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1, 425–452.

[5] Arnold, D., Wagner, P., Baayen, R. H. 2013. Using generalized additive models and random forests to model prosodic prominence in German. *Proc. of Interspeech*, 272–276.

[6] Niebuhr, O., Winkler, J. 2017. The relative cueing power of f0 and duration in German prominence perception. *Proc. of Interspeech*, 611–615.

[7] Baumann, S., Winter, B. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgements. *J. Phon* 70, 20–38.

[8] Linke, J., Kelterer, A., Dabrowski, M. A., Zarka, D. E., Schuppler, B. 2020. Towards automatic annotation of prosodic prominence levels in Austrian German. *Proc. Speech Prosody 2020*, 1000–1004.

[9] Johnson, K. 2004. Massive reduction in conversational American English. *Proc. in Spontaneous speech: Data and analysis. Proc. of the 1st session of the 10th international symposium*, 29–54.

[10] Ludusan, B., Schuppler, B. 2022. An analysis of prosodic boundaries across speaking styles in two varieties of German. *Speech Communication* 141, 93–106.

[11] Klabbers, E., Veldhuis, R. 2001. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing* 9, 39–51.

[12] Misra, H., Ikbal, S., Bourlard, H., Hermansky, H. 2004. Spectral entropy based feature for robust ASR. *ICASSP*. IEEE, 193–196.

[13] Rogério Scalassara., P., Eugenia Dajer., M., Dias Maciel., C., Carlos Pereira., J. 2008. Voice signals characterization through entropy measures. *Proceedings of the First International Conference on Bio-Inspired Systems and Signal Processing - Volume 2: BIOSIGNALS, (BIOSTEC 2008)*. INSTICC SciTePress, 163–170.

[14] Cohen, A. S., Hong, S. L., Guevara, A. 2010. Understanding emotional expression using prosodic analysis of natural speech: Refining the methodology. *Journal of behavior therapy and experimental psychiatry* 41 2, 150–7.

[15] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: The Graz corpus of Read And Spontaneous Speech. *Proc. of LREC*, 1465–1470.

[16] Wasserfall, S. 2020. Automatic Speech Segmentation using Kaldi. Master's thesis Technical University Graz.

[17] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[18] Schmitt, B. J. B. 2018. AMFM decompy documentation 1.0.8. https://bjbschmitt.github.io/AMFM_decompy/.

[19] Zahorian, S., Hu, H. 2008. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America* 123, 4559–71.

[20] Cover, T. M., Thomas, J. A. 2006. Elements of Information Theory 2nd Edition. *Wiley Series in Telecommunications and Signal Processing*.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

DISCERNING DIMENSIONS OF QUALITY FOR STATE OF THE ART SYNTHETIC SPEECH

Fritz Seebauer[1], Michael Kuhlmann[2], Reinhold Haeb-Umbach[2], Petra Wagner[1]

Bielefeld University[1], Paderborn University[2]
{fritz.seebauer, petra.wagner}@uni-bielefeld.de, {kuhlmann, haeb}@nt.upb.de

ABSTRACT

This paper describes an approach for determining the dimensions of quality for state-of-the-art synthetic speech. We propose that current evaluation metrics do not fully capture the meaningful dimensions of text-to-speech (TTS) and voice conversion (VC) systems. In order to develop a revised paradigm for meaningful evaluation, we conducted two experiments. First, we determined descriptive terms by querying naïve listeners on their impressions of modern TTS and VC systems. In a second experiment, we refined these terms into dimensions of quality and similarity by showcasing a consolidation procedure of manual clusterings. The resulting dimensions contain the standard evaluation categories of "intelligibility" and "naturalness" for both conditions. We could additionally discern dimensions of "tempo" and "demographics" in both domains. The final two dimensions as well as the relationships between categories proved to be different between TTS and VC, suggesting the need for modified evaluation scales based on the target construct.

Keywords: voice quality, voice conversion, text-to-speech, dimensions of quality, evaluation

1. INTRODUCTION

Phonetic sciences often lean on speech technology to generate stimulus data or derive processes of speech production and perception [1, 2, 3]. It has been shown, however, that the quality of synthetic speech systems not only varies, but might also generate voice quality features that do not exist in natural voices. Consequently, there exists a need for evaluating the quality of different speech technology systems regarding their suitability in phonetic research. Traditionally, synthetic speech has been evaluated primarily on the dimensions of "intelligibility" and "naturalness" [4, 5] based on ITU-T P.85 [6]. It has been questioned whether these two dimensions are capable of exhaustively capturing the full extent of what constitutes a good synthesis [7, 8] and whether multidimensional approaches might yield more accurate representations of quality [9]. It has also been shown that the specific wording of a given evaluation task has a great influence on the resulting Mean Opinion Score (MOS) with or without context [10, 11]. Finally, it has been pointed out that the target construct might differ between different applications of speech synthesis [12]. To conciliate these findings with our research, we conducted experiments in two different domains of speech synthesis. The first one concerns pure text-to-speech, with the latent construct being "quality", the second one represents voice conversion, with the latent construct being "similarity".

2. METHODS

To elicit latent dimensions of a given construct, it is common practice to take preconceived items of potential attributes and reduce them into smaller subsets using factor analysis [8, 13]. To counteract an inherent researcher bias, we opted to follow an inductive approach of item-generation and complement it with a manual clustering of multiple participants to determine the underlying dimensions. The only step requiring expert knowledge in the newly proposed paradigm is the naming of resulting dimensions, rather than pre-selecting scales based on the generated items, as suggested in behavioral science and psychology [14, 15, 16]. The reasons for developing this method were two-fold: Firstly, the new procedure allowed us to use the whole set of terms elicited during the first experiment without introducing bias by pre-selection. Secondly, exploratory factor analysis is always dependent on the audio samples used during the weighting experiment [17]. Having participants cluster the semantic space directly instead of using an audio sample as an intermediate eliminates the chance of having an unrepresentative token.

2.1. Experiment 1: Terms of quality

80 participants (40f/40m, L1 English) were recruited over the crowdsourcing platform Prolific [18]. They

TTS-system	training corpus	gender
vits [21]	vctk	m
silero-tts [22]	private	m
tacotronDDC+hifiGAN [21]	sam	f
tacotron2+wavegrad [21]	ek1	f
glow-tts+multibandmel [21]	LJ	f
microsoft neural	private	f
google wavenet	private	m
amazon polly	private	m

Table 1: TTS system types used for generating the audio samples in the elicitation experiment.

were instructed to listen to a provided audio sample and note down terms that they felt best described its quality. There was a constraint to supply at least three different terms for each sample. The experiment was split into two subsets with 40 participants each. In the first subset the audios were constructed using 8 different neural state-of-the-art systems with varying vocoders, balanced by target gender. The different system architectures are listed in Table 1. They were chosen to generate a broad set of different architectures employed in current-day research and real-life application. Each system read the phonetically balanced "caterpillar story" [19], with four rotating sentences being presented to each participant. The experiment yielded 387 different descriptions of the quality of synthetic speech.

The second subset elicited similar terms of quality for the domain of voice conversion. Here, the data of the Voice Conversion Challenge 2020 (VCC2020) [20] served as a basis to elicit dimensions of synthetic voice similarity. The participants were asked to listen to a voice conversion sample and its corresponding natural target audio and instructed to note down terms that they felt best described *similarities* or *differences* between the samples. The dataset consisted of a subset from the VCC2020 data containing 33 different voice conversion systems. The subset was stratified by source and target gender, as well as the target content. Each participant was presented with one sample pair from each gender combination with each audio pair containing one of the 4 different contents. The voice conversion experiment yielded 359 unique terms of similarity.

2.2. Experiment 2: Clustering of terms

To aggregate the elicited terms of quality into meaningful dimensions, a second experiment was conducted. 10 Participants (5f/5m, L1 varied) were presented with a randomized list of all terms of quality for one domain. They were instructed to group those terms into 10 clusters of varying sizes on a graphical two-dimensional interface. The clusters were to be created according to the terms' semantic similarity, with closer terms being more similar. The participants were additionally asked to denote group similarity using 10 predefined colors. There were 5 participants each for both the text-to-speech and voice conversion terms. To ensure some level of English proficiency, a self-assessment pre-test was carried out based on the bilingual language profile of [23]. The experiment lasted about 2 hours, and the participants were instructed to take a mandatory break of 15min in between to avoid fatigue. The resulting clusters were evaluated both in the distance spacing and the color groupings denoting semantic similarity.

2.3. Finding latent Categories/Dimensions

The coordinate data were first normalized via min-max scaling, as some of the participants did not use the whole space. To find a representative view of their average groupings, a distance matrix was computed for each participant, denoting the relationship (distance) between two terms. Given those matrices, we obtain a mean Euclidean distance matrix by averaging across participants, from which we recalculate the plane coordinates by eigenvalue decomposition of $M_{ij} = \frac{D_{1j}^2 + D_{i1}^2 D_{ij}^2}{2}$, where D_{ij} denotes the average Euclidean distance between the i-th and j-th term [24]. This yields a single average representation of all terms in 2D space.

Regarding the color assignments, the participants each colored the terms in a self-defined order. This means that (1) they might not assign the same color to the same latent group (indicated by a large overlap of terms in the colored groups) and (2) terms might be assigned to different latent groups across participants. To solve (1), we compute the optimal cluster (color) match between participants. The problem was first turned into an assignment problem by counting how many terms were assigned the same 5-tuple of colors, and then transforming those counts into a cost matrix: Higher cost indicates better color matching. The resulting assignment problem was then solved by employing the modified Jonker-Volgenant algorithm [25]. However, the optimal assignment obtained this way is not universally valid as it may contain tuples where a specific color for a single participant is reused. To compensate for this, we used a greedy approach to prune tuples in the cost matrix by colors which already had been assigned in a previous step to a given participant. In each step, we compute the optimal assignment from the pruned cost matrix, take out the tuple with the highest cost, and perform the participant-color assignment followed by color pruning until the cost matrix

contains no more valid color tuples. To solve (2), we performed a color-matched majority voting per term: If the majority of the participants assigned the same color to the term, it was also assigned this color. If the participants could not agree on a color, the term was dropped from the average representation. Given the newly clustered and reduced set of terms, we determined the centroids of all clusters and derived a measure of confidence for each term by calculating the Euclidean distance from its group centroid. Summing these across groups and inverting them yields a measure of internal certainty. An external certainty measure is derived by summing the distance of a given centroid to all other groups. These measures are variations of the subcomponents of the Calinski-Harabasz Index for measuring cluster dispersion [26].

3. RESULTS

Figure 1 shows the average distance results for the pure text-to-speech evaluation. The majority voting procedure only kept 33% of the original 382 terms. The most distinct clusters were those pertaining to demographic information of the inferred speaker as well as those regarding human likeness. The intelligibility group exclusively contains items describing the audio quality and clarity pointing to the fact that modern TTS systems do not suffer traditional understandability problems. This clarity dimension overlaps with the one regarding speech tempo, which might reflect the fact that some of the "tempo" terms actually describe difficulties with intelligibility due to speed. The voice conversion results are visualized in Figure 2. For this domain, the majority vote left 51% of data points. Their respective distance certainty measures are shown in Table 2. As is evident, the clearest group concerned terms relating to demographic associations with the perceived speakers, such as "accent" or "gender". The worst category regarding external and internal certainty contained terms regarding voice quality like "tone", "softness" or "vocal fry". Interestingly, the participants seemed to clearly distinguish between terms of "intelligibility" and "naturalness" regarding their color assignments, but associate and conflate them in their distance markings. The same seems to be true of the categories relating to "content" and "expression".

4. DISCUSSION

The amount of variability in the participants' group assignments highlights the need for an objective procedure. They clearly differed in their category assignments as was evident by low overall Jaccard agreement scores of 32% in the TTS condition and 31% in the VC condition. This underlines our premise that having the reduction of original items into scales done by a single individual is questionable. The findings regarding the dimensions of quality are very much in line with the current status quo. The main categories we could elicit in the domain of speech technology did include the often-tested intelligibility and naturalness, while additionally uncovering that many laypeople seem to take perceived demographic information into account when judging quality. The space measures also allow us to relate the revealed dimensions to each other. For the TTS domain, we can discern that the human-likeness does seem to correspond to evoked emotional qualities and audio distortions, while intelligibility conflates with the perceived tempo and overall pleasantness. In the VC condition, we see semantic associations of naturalness, intelligibility, and voice quality features and a clear separation of tempo terms. This first step in the search for the underlying dimensions to modern day synthetic speech opens up many directions for future research. In particular, a direct comparison of our procedure and a classical Exploratory Factor Analysis is still pending. Post-experiment feedback from the participants in the second experiment revealed that they felt constrained by the two-dimensional space to accurately model the group relationships. While it would be feasible to extend the experimental setup and analysis procedure by a third dimension, care should be taken to ensure naïve participants are actually able to navigate a three-dimensional space on a digital interface. The dimensions proposed here could also be compared to quality dimensions of natural voices such as those described in [27, 28].

Domain	Category	Ext. certainty	Int. certainty
VC	expression	0.35	0.84
	intelligibility	0.05	0.22
	naturalness	0.08	0.82
	voice quality	0.0	0.0
	tempo	0.52	0.65
	content	0.54	0.85
	demographics	1.0	1.0
TTS	pleasantness	0.24	0.35
	emotion	0.0	0.07
	audio quality	0.14	0.41
	tempo	0.08	0.64
	clarity	0.11	0.0
	human-likeness	0.93	1.0
	demographics	1.0	0.54

Table 2: External and internal certainty measures for each category as calculated by the distance to other categories and averaged distance to the centroid, respectively.The values are scaled to range between 0 and 1 for easier comparison

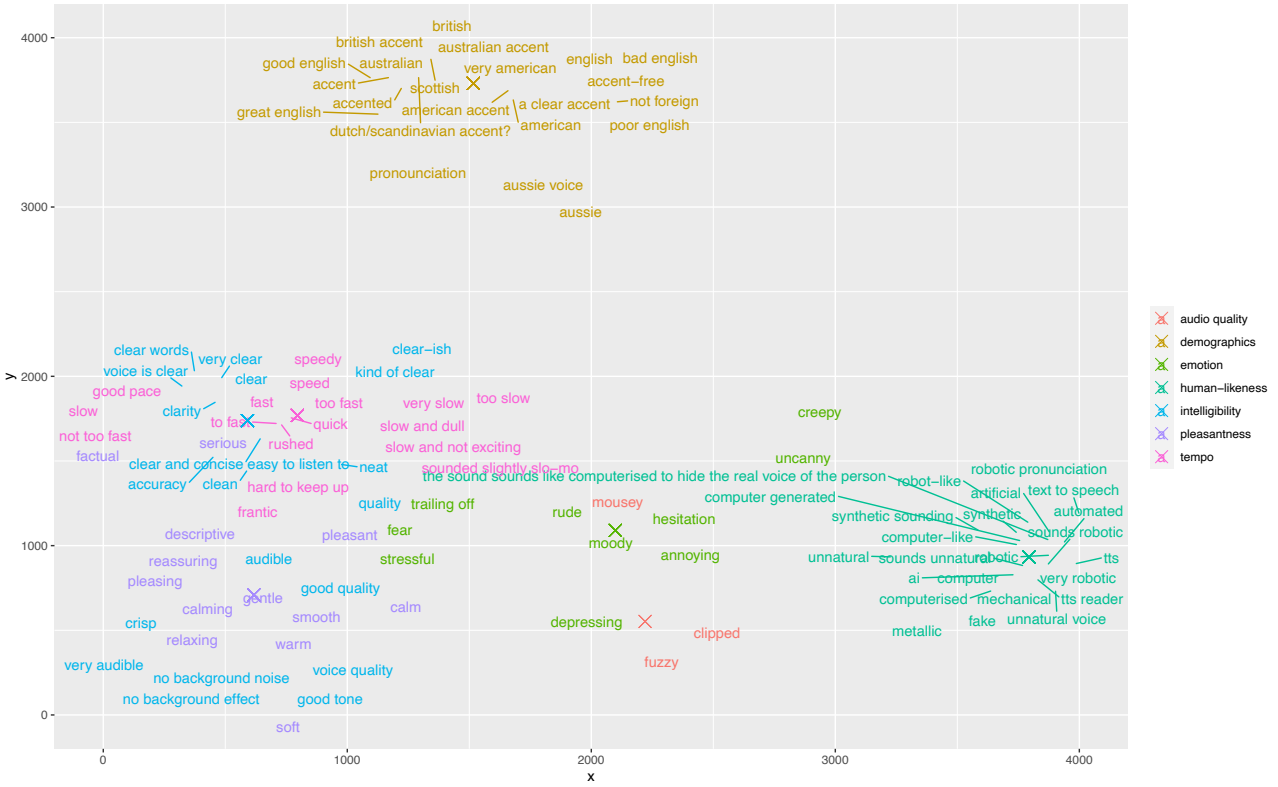


Figure 1: Relative distance assignment of text-to-speech terms averaged over all participants. Color denotes group assignment after consolidating cluster information across participants. The medians are marked in their respective colors

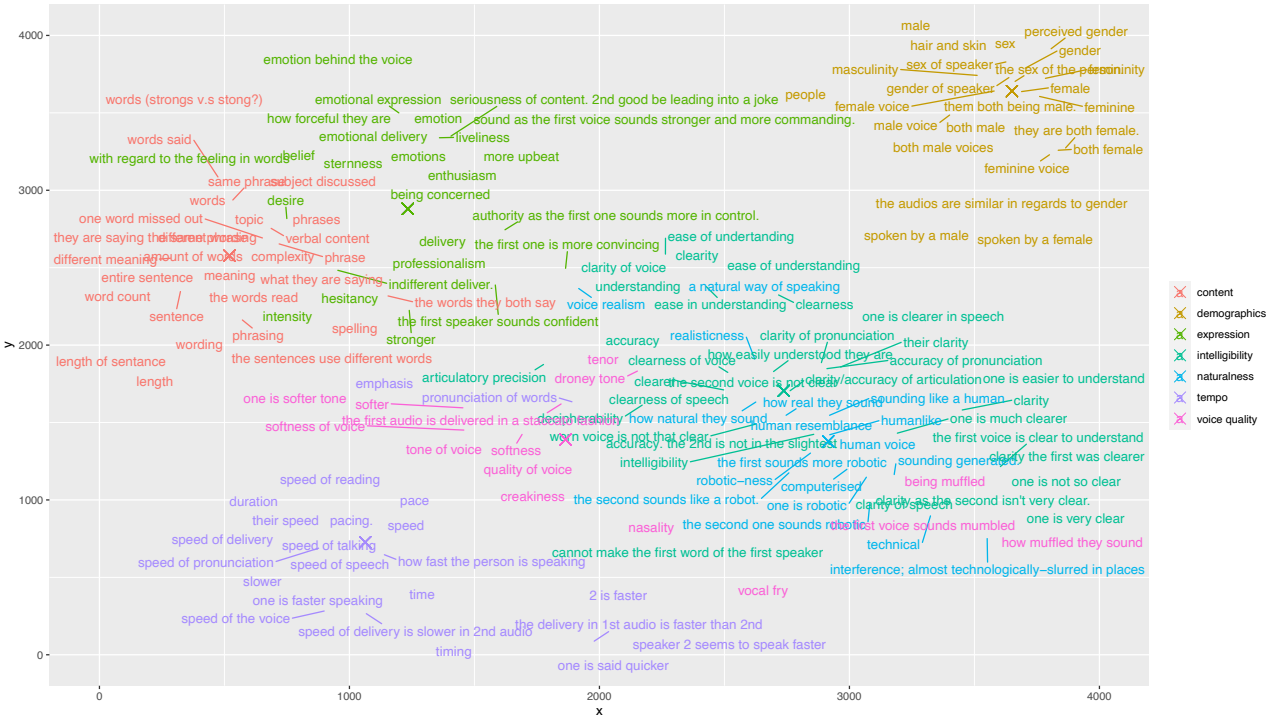


Figure 2: Relative distance assignment of voice conversion terms averaged over all participants. Color denotes group assignment after consolidating cluster information across participants. The medians are marked in their respective colors

5. REFERENCES

[1] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *International Congress of Phonetic Sciences ICPhS*, 2019, pp. 487–491.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[3] A. Cohen and S. G. Nooteboom, *Structure and process in speech perception*. Springer, 1975.

[4] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of text and speech systems*. Springer, 2007, pp. 29–64.

[5] C. Mayo, R. A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.

[6] I. T. Union, "ITU-T Rec. P.85, A method for subjective performance assessment of the quality of speech voice output devices," 1994.

[7] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale," *Computer speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[8] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech," in *Eighth ISCA Workshop on Speech Synthesis*, 2013, pp. 148–151.

[9] J. Benesty, M. M. Sondhi, Y. Huang *et al.*, "Speech quality assessment," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 84–96.

[10] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *arXiv preprint arXiv:1909.03965*, 2019.

[11] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 148–153.

[12] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Č. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation state-of-the-art assessment and suggestion for a novel research program," in *Proc. of the 10th Speech Synthesis Workshop*, 2019, pp. 105–110.

[13] S. S. Rallabandi, B. Naderi, and S. Möller, "Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 1–6.

[14] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best practices for developing and validating scales for health, social, and behavioral research: a primer," *Frontiers in public health*, vol. 6, p. 149, 2018.

[15] T. R. Hinkin, "A review of scale development practices in the study of organizations," *Journal of Management*, vol. 21, no. 5, pp. 967–988, 1995.

[16] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research." *Psychological methods*, vol. 4, no. 3, pp. 272–299, 1999.

[17] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[18] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[19] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler, and S. Russell, "'the caterpillar': A novel reading passage for assessment of motor speech disorders," *American Journal of Speech-Language Pathology*, vol. 22, no. 1, pp. 1–9, 2013.

[20] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020 database v1. 0," 2020.

[21] Coqui AI, "Coqui tts," https://github.com/coqui-ai/TTS, 2022.

[22] Silero Team, "Silero models: pre-trained enterprise-grade stt / tts models and benchmarks," https://github.com/snakers4/silero-models, 2021.

[23] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual language profile: An easy-to-use instrument to assess bilingualism," *University of Texas at Austin*, 2012.

[24] G. Young and A. S. Householder, "Discussion of a set of points in terms of their mutual distances," *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938.

[25] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.

[26] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[27] B. Weiss, D. Estival, and U. Stiefelhagen, "Non-experts' perceptual dimensions of voice assessed by using direct comparisons," *Acta Acustica united with Acustica*, vol. 104, no. 1, pp. 174–184, 2018.

[28] J. Kreiman and B. R. Gerratt, "Validity of rating scale measures of voice quality," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1598–1608, 1998.

A BIOLOGICAL INVESTIGATION OF PERFORMATIVE SPEECH THROUGH SYNTHESIS

Emily Lau[1], Brechtje Post[1], Kate Knill[2]

[1]Theoretical and Applied Linguistics and [2]Dept. of Engineering,
University of Cambridge, Cambridge, UK
{ehynl2, bmbp2, kmk1001}@cam.ac.uk

ABSTRACT

Dramatic vocal performance is a layered and nuanced phenomenon. It is difficult to pin down the exact aspects of a performance that make it appealing to the human ear. This work seeks to investigate how the Bio-informational Dimensions (BIDs) impact the judgments of dramatic performance, in particular for the emotion "anger". Test participants listened to pairs of utterances that were resynthesized along the BIDs of size projection and dynamicity to varying degrees to simulate dramatic expressions of anger, and then rated the utterances' differences in dramatic expression. Increased size projection significantly improved listener ratings, while higher dynamicity lowered them. These results indicate that size projection is very important to simulating anger, while dynamicity had a more ambiguous effect that warrants further investigation.

Keywords: speech synthesis, emotional prosody, BIDs, performance, drama.

1. INTRODUCTION

Performative speech realizes itself in varied and dynamic ways. However, even with the subjective and nuanced nature of performance quality, there is no mistaking that there are some fundamental attributes and physical manipulations to dramatic speech that distinguish it from natural speech and make it appealing to the human ear [1, 2]. Researchers in both Linguistics and Artificial Intelligence have begun to probe expressive speech, which includes the specialized nature of dramatic attitudes, which would not only help us understand the nature of human expression, but also add range of expression to today's voice agents.

With regards to expressive prosody, much of the research has focused on directly mapping acoustic parameters and F0 contours to different emotions [3, 4, 5, 6, 7]. Alternately, other work has examined emotional expression in context of the evolution of human communicative functions [8, 9, 10]. Xu et al [11] has expanded on these works and proposed that affective speech is controlled by the so-called Bio-informational Dimensions (BIDs), which manipulate the vocal signal of the speaker to influence the behavior of the listener. The BIDs most investigated so far are:

- Size projection: body size projected by the speaker, associated with median pitch shift, formant shift, and voice quality
- Dynamicity: vigorousness of the speaker's speech stream, associated with pitch range and duration

These have both been found to be influential on the perception of vocal attractiveness, "poshness" and friendliness [12, 13, 14]. Therefore, this theory seems a promising framework in which to study dramatic speech.

This paper aims to observe the effects of size projection and dynamicity on listener judgment of dramatic speech, specifically dramatically angry speech. Because anger is associated with the high end of both dimensions [11], and performative speech is found to be more exaggerated than spontaneous speech [1], it was expected that speech resynthesized to exhibit high size projection and dynamicity would be perceived as most angry.

2. METHODOLOGY

A set of listening experiments was performed to assess how listener judgment of dramatic speech would change according to modulation of speech along the BIDs.

2.1. Stimuli

The stimuli for these experiments were made to target the emotion of "anger," since that is the most identifiable emotion. A 26 year old male speaker of Standard Southern British English (SSBE) recorded the base utterances. The recording was done in a soundproof recording booth in Cambridge

University's Phonetics Laboratory. A sentence taken from the ARCTIC Corpus - "Author of 'The Danger Trail,' Philip Steels, etc." - was used as the utterance content [15]. The speaker produced a set of recordings spoken with a neutral affect but with the varying voice qualities of "modal," "tense," and "more tense."

The utterances were then resynthesized in Praat [16], using a script adapted from stimuli generator code used in Xu et al [12]. Based on the literature survey from Xu et al [11], the emotion of anger is associated with high size projection and high dynamicity. Therefore, the utterances were manipulated upwards along each of these BIDs to different degrees and in varying combinations, in order to create different degrees of "anger." This resulted in three different levels along each dimension, resulting in 3 x 3 = 9 utterances, and therefore 36 possible utterances. Each pair of utterances was tested twice for statistical stability, resulting in 36 x 2 = 72 total judgments. The pairs were presented in reverse order when repeated, to prevent listener bias towards one particular stimulus. All pairs were presented to the listeners in random order. The manipulation parameters for each BID are shown in Tables 1 and 2.

SP Manipulation	Voice Quality	Formant Shift	Pitch Shift
+2	More tense	0.8	−0.5
+1	Tense	0.9	−0.25
0	Modal	1.0	0

Table 1: Size projection Praat resynthesis parameters

Dynamicity Manipulation	Pitch Range	Duration
+2	1.2	0.8
+1	1.1	0.9
0	1.0	1.0

Table 2: Dynamicity Praat resynthesis parameters

2.2. Experimental Setup

The listening experiment was conducted online on the platform Gorilla. There were a total of 30 participants, all speakers of SSBE. There were 10 males, 19 females and 1 non-binary participant, all of whom ranged between the ages of 18 and 50, and were resident in southeast United Kingdom.

Participants were presented with stimuli pairs and asked to compare how "dramatically angry" the utterances were compared to each other by indicating their preference on this rating scale:

1. Clip A sounds much more dramatically angry.
2. Clip A sounds slightly more dramatically angry.
3. Both clips sound equally dramatically angry.
4. Clip B sounds slightly more dramatically angry.
5. Clip B sounds much more dramatically angry.

Therefore, if a listener felt the first utterance they heard was much more dramatically angry, they would select option 1, and would select option 5 if they thought that of the second sound clip.

2.3. Testing Alternate Synthesis Bases

To test how participants would react to vocal manipulations on voices that already had emotion applied, another experiment was carried out using emotional voices as a synthesis base. The same speaker recorded another set of recordings in an "angry" voice. The base utterance was similarly resynthesized in Praat, but instead of up each of the BIDs, it was manipulated one step up and one step down. These were similarly grouped into pairs and presented to participants to compare in the same manner. Voice quality was not included in the parameters included in these size projection manipulations, since it was difficult for the speaker to modulate his voice quality while speaking in an "angry" affectation. The manipulation parameters for each BID for this particular stimuli set are shown in Tables 3 and 4.

SP Manipulation	Formant Shift	Pitch Shift
+1	0.9	−0.5
0	1.0	0
−1	1.1	0.5

Table 3: Size projection resynthesis parameters in Praat for an emotional synthesis base

Dynamicity Manipulation	Pitch Range	Duration
+1	1.2	0.8
0	1.0	1.0
−1	0.8	1.2

Table 4: Dynamicity resynthesis parameters in Praat for an emotional synthesis base

3. RESULTS

A MANOVA was conducted on the results of the tests with both neutral and emotional synthesis bases. An ANOVA was also performed to observe size projection effects for specific pairs of stimuli at specific dynamicity levels, as well as an ANOVA to observe the reverse.

3.1. Neutral Synthesis Base

Figure 1 shows the effect of size projection and dynamicity on listener ratings. The listener trials were divided into three groups of size projection differences: -1 (1 size projection to 0 size projection), +1 (1 size projection to 2 size projection), and +2 (0 size projection to 2 size projection). The MANOVA showed a very significant upward effect of size projection ($F = 147.2$, $p < 2e\text{-}16$) on listener judgments, even between all three difference groups.

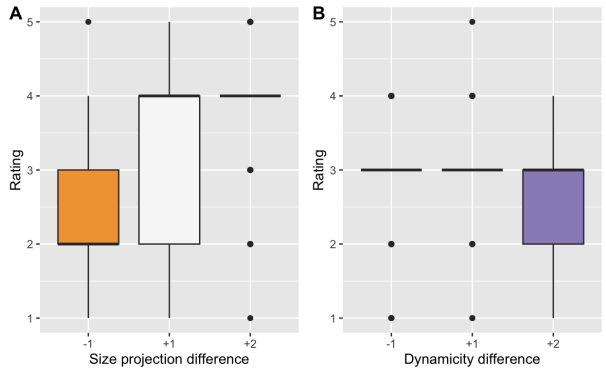


Figure 1: Ratings of dramatic anger across all trials, where A indicates the ratings for each size projection difference, while B indicates the ratings for each dynamicity difference.

The listener trials were similarly divided into three groups of dynamicity differences. This analysis showed a significant effect of dynamicity ($F = 9.715$, $p = 7.17e\text{-}05$). A closer examination using the Bonferroni correction shows no significant effect between the trials with a -1 difference versus a +1 difference ($p = 1.000$), while there was a definitely significant effect between those of a -1 difference versus a +2 difference ($p = 0.00011$). Contrary to the prediction that listener ratings would increase as dynamicity increased, the ratings decreased significantly in this case.

As expected, for the ANOVA of size projection effects at specific dynamicity levels, increased size projection did have a significant effect on listener ratings at all levels of dynamicity, and is quite apparent at the +2 dynamicity level ($F = 60.45$, $p < 2.16e\text{-}16$). As shown in Figure 2, there is a very pronounced effect between the -1 and +1 size projection levels ($p = 8.9e\text{-}14$), as well as between -1 and +2 ($p < 2.00e\text{-}16$), but interesting the effect is minimal between +1 and +2 ($p = 0.1$), which suggests there is a threshold to this effect.

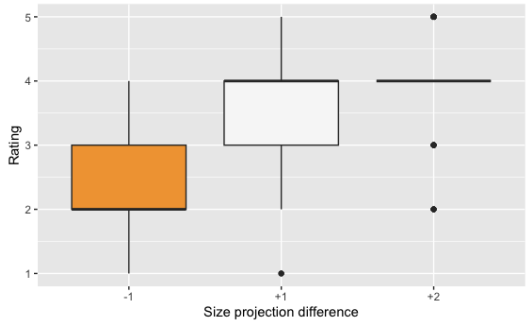


Figure 2: Ratings of dramatic anger for each size projection difference at the +2 dynamicity level.

A similar ANOVA for dynamicity showed this BID has a significant effect, but only at the +2 size projection level ($F = 7.939$, $p = 0.000499$), and is nowhere near as pronounced as the size projection effect at the same dynamicity level. As shown in Figure 3, the effect is most pronounced between the -1 and +2 size projection difference groups ($p = 0.0078$). As with the MANOVA analysis, the direction of the effect of dynamicity with regard to listener ratings was downwards rather than upwards.

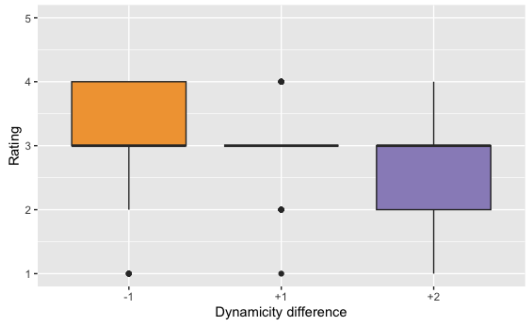


Figure 3: Ratings of dramatic anger for each dynamicity difference at the +2 size projection level.

3.2. Emotional synthesis base

Figure 4 shows the results of the listening test using stimuli resynthesized from an emotional synthesis base. As predicted, there was a substantially significant effect of size projection ($F = 147.2$, $p < 2e\text{-}16$), especially between -1 and +1 ($p < 2e\text{-}16$)

and between -1 and +2 ($p < 2e\text{-}16$) size projection differences.

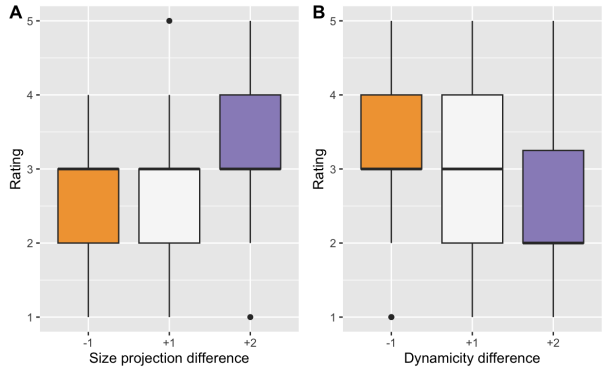


Figure 4: Ratings of dramatic anger across all trials for emotional synthesis base, where A shows the ratings for each size projection difference, while B shows ratings for each dynamicity difference.

The MANOVA for dynamicity also revealed a significant effect ($F = 9.715$, $p = 7.17e^{-5}$), most evident between trials of -1 and +2 dynamicity difference ($p = 1.1e^{-4}$), while quite minimal between trials of -1 and +1 dynamicity difference ($p = 1.000$).

4. DISCUSSION

Based on Xu et al [11] and other literature observing emotional acoustic correlates [17, 18, 19], these experiments utilized the assumption that anger is associated with higher size projection and dynamicity. Based on the study from Jurgens et al [1] observing acoustic differences between natural and acted speech, it was predicted that experimental participants would find utterances with more exaggerated features associated with anger to sound more "dramatic."

The MANOVA and individual ANOVAs for the main (neutral base) listening test indicate that size projection has a significant effect on what listeners consider to be "dramatic anger," which confirms findings in the literature [13]. Listeners consistently found utterances with higher size projection more dramatically angry. In the individual ANOVAs, however, this effect was not always linear, as listener ratings plateaued at higher dynamicity levels. This suggests that at a certain threshold, either the stimuli at higher dynamicity levels sounded similar, or listeners did not respond as well to more exaggerated acoustic features.

Dynamicity also had a significant, but not as pronounced, effect. Contrary to the prediction that listeners would rate utterances with higher dynamicity as more dramatically angry, ratings decreased when dynamicity increased. There are a few considerations for this effect. An inspection of the resynthesized stimuli, found that artificial manipulation of stimuli pitch range and duration in Praat did not consistently manipulate the pitch range, especially when it was meant to be increased. It is possible that while attempting to increase both size projection and dynamicity, lowering the pitch median cancelled out increases to pitch range. This consequence of the resynthesis process would explain why listeners seemed to find stimuli at higher dynamicity levels similar. However, it is noteworthy that when the pitch range increase did work, and the pitch range of the neutral control (508.9 Hz) is compared to that of the +2 size projection/+2 dynamicity stimulus (516.9 Hz), listener ratings still decreased, meaning that dynamicity may still have an effect on listener perception. The results of the MANOVA from the emotional synthesis base test could be interpreted to support this conclusion, as they show a significant effect of both size projection and dynamicity.

These results call into the question the assumptions around anger and the BIDs espoused at the outset of the experiment. While these may not have necessarily been wrongful, it is possible they did not correlate with listener expectations of what "dramatic anger" sounds like. It is also worth considering whether specific parameters always strictly correlate with a specific BID.

5. CONCLUSIONS

This paper attempted to discover if the BIDs of size projection and dynamicity contribute to listener judgment of performative anger when used to artificially manipulate utterances. It was predicted that listeners would find utterances with higher size projection and dynamicity more dramatically angry. While size projection and dynamicity both had significant effects, the effect of the former was larger. Higher size projection improved listener ratings showing it is quite impactful for listeners. Unexpectedly higher dynamicity lowered ratings.

The study also has interesting implications about the impact of dynamicity and how it correlates with the emotion of anger. While the resynthesis process performed this manipulation imperfectly, this calls into question listener expectations of dramatic anger and whether this was reflected in the stimuli. Further research should investigate these correlations and the nuances of different emotional subsets.

6. REFERENCES

[1] R. Jurgens, K. Hammerschmidt, and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Frontiers in Psychology*, vol. 2, pp. 1–11, 2011.

[2] R. Jurgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of Nonverbal Behavior*, vol. 39, pp. 195–214, 2015.

[3] K. R. Scherer, "Voice quality analysis of american and german speakers," *Journal of Psycholinguistie Research*, vol. 3, 1974.

[4] K. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological Bulletin*, vol. 99, no. 2, p. 143, 1986.

[5] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance." *Handbook of Social Psychophysiology*, pp. 165–197, 1989. [Online]. Available: https://psycnet.apa.org/record/1989-97735-007

[6] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, vol. 15, pp. 123–148, 1991.

[7] S. Mozziconacci, "Prosody and Emotions," in *Proc. First Int'l Conf. Speech Prosody (Speech Prosody '02)*, 2002, pp. 1–9.

[8] E. S. Morton, "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," *The American Naturalist*, vol. 111, no. 981, pp. 855–869, 1977.

[9] J. J. Ohala, "An ethological perspective on common cross-language utilization of F0 of voice," *Phonetica*, vol. 41, pp. 1–16, 1 1984. [Online]. Available: https://www.degruyter.com/document/doi/10.1159/000261706/html

[10] C. Gussenhoven, "Intonation and interpretation: Phonetics and phonology," in *Proc. Speech Prosody 2002, International Conference*, 2002.

[11] Y. Xu, A. Kelly, and C. Smillie, "Emotional expressions as communicative signals," *Prosody & Iconicity*, pp. 33–60, 2013.

[12] Y. Xu and L. Noble, "Friendly speech and happy speech-are they the same?" in *Proc. 17th International Congress of Phonetic Sciences*, 2011. [Online]. Available: https://www.researchgate.net/publication/266605583

[13] Y. Xu, A. Lee, W.-L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PLoS ONE*, vol. 8, p. e62397, 4 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0062397

[14] L. Jiao, C. Wang, C. Hsu, P. Birkholz, and Y. Xu, "Does posh English sound attractive?" in *Proc. Interspeech 2017*, 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1691

[15] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. Fifth ISCA workshop on speech synthesis*, 2004.

[16] P. Boersma, "Praat: doing phonetics by computer," 2006. [Online]. Available: http://www.praat.org/

[17] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[18] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[19] S. Chuenwattanapranithi, Y. Xu, B. Thipakorn, and S. Maneewongvatana, "Encoding emotions in speech with the size code," *Phonetica*, vol. 65, no. 4, pp. 210–230, 2008.

AUTOMATIC SOUND EVENT DETECTION AND CLASSIFICATION OF GREAT APE CALLS USING NEURAL NETWORKS

Zifan Jiang[1,2], Adrian Soldati[1,3], Isaac Schamberg[2], Adriano R. Lameira[4], Steven Moran[1,5]

[1]University of Neuchâtel, [2]University of Zurich, [3]University of St Andrews,
[4]University of Warwick, [5]University of Miami
jiang@cl.uzh.ch, {adrian.soldati, steven.moran}@unine.ch, isaac.schamberg@uzh.ch, adriano.lameira@warwick.ac.uk

ABSTRACT

We present a novel approach to automatically detect and classify great ape calls from continuous raw audio recordings collected during field research. Our method leverages deep pretrained and sequential neural networks, including wav2vec 2.0 and LSTM, and is validated on three data sets from three different great ape lineages (orangutans, chimpanzees, and bonobos). The recordings were collected by different researchers and include different annotation schemes, which our pipeline preprocesses and trains in a uniform fashion. Our results for call detection and classification attain high accuracy. Our method is aimed to be generalizable to other animal species, and more generally, sound event detection tasks. To foster future research, we make our pipeline and methods publicly available.[1]

Keywords: sound event detection, neural networks, primatology, phonetics, computational linguistics

1. INTRODUCTION

In primatology, as in documentary linguistics, the collection, annotation, and analysis of primary field data are time-consuming and expensive. The recordings of primate calls are also often undertaken in not ideal environmental conditions (e.g., in a dense and noisy forest where other vocal species are present), making the process even more challenging. Therefore it is typical that primatologists first manually annotate their recordings and then conduct acoustic analyses on their species' specific calls.

Our goal in this paper is to automatically and accurately detect and classify primate calls from raw audio data. After discussing related work (§2), we describe our data sets (§3) and present our method that leverages pretrained and sequential neural networks (§4). We carry out reliable and reproducible experiments (§5) to support the effectiveness of our approach and compare our results between different architectural choices on the data sets of three great ape lineages. Overall, our models achieve more than 80% frame-level classification accuracy and weighted F1-score. Interestingly, we find that the wav2vec 2.0 [1] model – even though pretrained on a large corpus of human speech [2] – generalizes surprisingly well as an audio feature representation layer for great ape calls without additional fine-tuning. This finding perhaps bides well with the similar vocal tract morphology between extant great apes and humans, and the larger prediction that ancestral ape-like calls evolved to become the building blocks of human speech.

2. RELATED WORK

To the best of our knowledge, there is no precedent research on the automatic detection and classification of great ape calls from raw audio. Therefore, we find our work lies in between general-purpose sound event detection and human speech recognition. We briefly discuss these also in light of existing research on animal sound classification.

Sound Event Detection (SED) aims at automatically recognizing *what* is happening in an audio signal and *when* it is happening [3]. SED tasks are usually general-purpose (e.g., birds singing or footsteps) as opposed to domain-specific tasks like human speech or music analysis, and thus encounter particular challenges, such as the *cocktail party effect* [4]. Notably, the DCASE challenge that took place in recent years involves relevant SED tasks [5]. A variant of SED is the audio tagging task [6], in which only *what* is happening in an audio signal is annotated and recognized, but not *when*.

Prominent work on both tasks [7, 8] tends to consist of common components, including: (1) an acoustic feature representation layer, either (a) traditional spectrogram-based approaches or (b) convolutional neural networks (CNN) on spectrogram features or on raw waveforms [9]; (2) a sequence modeling layer that captures temporal interaction between frame-level features, which takes the forms of mean/max pooling strategies,

data set	# audio clips	mean / call / total duration	# indiv. (♂/ ♀)	# call types (duration ratio) / units
chimpanzee	235	~ 8s / 1,955s / 1,964s	11 (11 / 0)	4 (6:3:3:1) / 686
orangutan	65	~ 74s / 2,793s / 4,817s	10 (10 / 0)	7 (700:40:200:100:20:1:1) / 9016
bonobo	28	~ 24s / 62s / 677s	7 (3 / 4)	18 (20:40:10:20:...:200:...:5:1) / 356

Table 1: Data set overview and stats. Shown in the table are the number of audio clips, mean duration of each clip, total duration of calls of all clips, total duration of all clips, number of individuals (male and female), number of call types (approximate total duration ratio of them, partially omitted for bonobo due to space limit), and number of individually annotated call units.

recurrent neural networks (RNN) or Transformers [10]; and (3) an objective function, usually cross-entropy classification on frame-level (for SED) or clip-level (for audio tagging), or occasionally CTC [11] for "sequentially labeled data" [12].

Animal sound classification involves related research that addresses animal sounds [13, 14, 15]. While this line of research uses similar acoustic feature representation techniques as seen in SED tasks, the models tend to work on top of individual short audio units of animal vocalizations that are manually selected from raw audio recordings by human experts, and thus fall short on dealing with a continuous audio signal that contains many audio events as well as noise. Instead, our method detects and classifies great ape calls directly from raw recordings ranging from seconds to minutes (theoretically recurrent models also extend to hours, but it could pose challenges in model training time).

Automatic Speech Recognition (ASR) on humans – one of the extant five great apes – has seen significant progress [16] (while animal calls remain mysterious to fully decipher), including the recent work wav2vec 2.0 [1]. We are interested in whether the wav2vec 2.0 speech representation pretrained on Librispeech [2] can be applied successfully to great ape call detection and classification.

3. DATA

Our data consists of recordings of chimpanzee pant-hoots, orangutan long calls, and bonobo high-hoots. Table 1 provides an overview of these data sets.

Chimpanzee pant-hoots are vocal sequences composed of up to four acoustically distinct phases, typically produced in this order: *introduction*, *build-up*, *climax*, and *let-down* [17]. Most pant-hoots contain two or more phases, although single phases can be produced in specific contexts [18]. The successive nature and well-balanced phase duration facilitate the training of a classifier (§5.1). We use annotated recordings of isolated pant-hoots (i.e., no temporal overlap with others' calls) produced during feeding, traveling, and resting context.

Orangutan long calls are composed by a *full pulse*, which is sub-divided into a *sub-pulse transitory element* and a *pulse body*, or into sequences of *bubble sub-pulse* or *grumble sub-pulse* [19, 20]. The complex temporal interrelationship between phases and the overlapping and class-unbalanced characteristics (some phases are extremely short) pose challenges for training a multi-class classifier. On the other hand, the duration of calls and non-calls in the data set are well balanced, which opens the gate to a binary call detection model (§5.2). Annotated recordings include spontaneous long calls and long calls produced in response to other males' vocal presence or environmental disturbances.

Bonobo high-hoots – which make up the plurality of our bonobo call data set – are loud, tonal vocalizations [21]. Unlike the chimpanzee and orangutan data sets, call types other than the homologous loud calls are also present in the data set, but are rather minor and diverse (e.g., peep-yelp, soft barks), which could be challenging to find. The total call duration is relatively short.

4. METHOD

This section introduces our method, which is illustrated in Fig. 1. We describe each step in turn.

4.1. Audio Preprocessing

First, all audio clips are converted to *.wav* format and resampled to 16 kHz sample rate. We segment them into 20ms frames and pad with zeros if necessary.

4.2. Feature and Label Extraction

Next, we extract three types of acoustic features with the *torchaudio* Python package. For an audio clip of T frames, we get a sequence of frame-level features $I_{1:T}$, in the shape of $T \times feature\_dim$:

- **Raw waveform**: the original amplitude values over time. $feature\_dim = 0.02 \times 16000 = 320$.
- **Spectrogram**: calculated then from the raw waveform with $feature\_dim = 201$.

- **wav2vec 2.0**: inferred from the raw waveform by the *WAV2VEC2_BASE* model on a CPU device with $feature\_dim = 768$.

We then extract frame-level labels from our annotations. For each annotated unit, we mark all frames inside the annotated time span with a positive class index indicating the call type (e.g., 1 for *intro*); unmarked frames are 0s by default, indicating non-calls. The labels' shape of a clip $L_{1:T}$ is $T \times 1$.

4.3. Data Split

We shuffle the clips (features and labels) of each data set and split them into 80%, 10%, and 10% for training, validation, and test sets, respectively. We make three different splits by three different random seeds 0, 42, and 3407 [22] to allow multiple experiments and to study the effect of randomness.

4.4. Sequence Modeling

Since our goal is to learn a function f from $I_{1:T}$ to $L_{1:T}$, we first input $I_{1:T}$ to a sequence modeling function f^m that outputs a hidden sequence $H_{1:T}$ ($T \times hidden\_dim$), which captures inter-frame interaction. $H_{1:T}$ then goes through a dense linear function f^d (output $D_{1:T}$) followed by a Softmax function f^s to produce the probability distribution over target classes, i.e., $P_{1:T}$ ($T \times num\_class$).

Specifically, f^m takes the following forms:

- **RNN**: a (bidirectional) LSTM with hidden size 1024. $hidden\_dim = 1024 * 2 = 2048$.
- **Transformer encoder**: a Transformer encoder with 8 heads and 6 layers. $hidden\_dim = 1024$.
- **Autoregressive model**: we optionally add autoregressive connections between time steps to encourage consistent output labels. The output of f^d at time step t, i.e., D_t is concatenated to the next time step's input I_{t+1}. By default, it indicates that the RNN layer goes unidirectionally, but one can also stack two autoregressive layers and add them up before the Softmax operation to retain bidirectionality.

Lastly, a cross-entropy loss is computed between the model output $P_{1:T}$ and the gold labels $L_{1:T}$, then backpropagated to train the models. To mitigate the class imbalance problem, we set class weights to the reciprocal of each class's occurrence times.

5. EXPERIMENTS AND RESULTS

Our experiments were done in PyTorch [23], with Python 3.8 on an Nvidia Tesla V100 GPU (32GB ram). Most models have ~40k million parameters and finish the training of up to 200 epochs with early stopping on validation F1-score within one hour (or a few hours because autoregressive recurrent models run slowly on PyTorch). Table 2 presents our results.

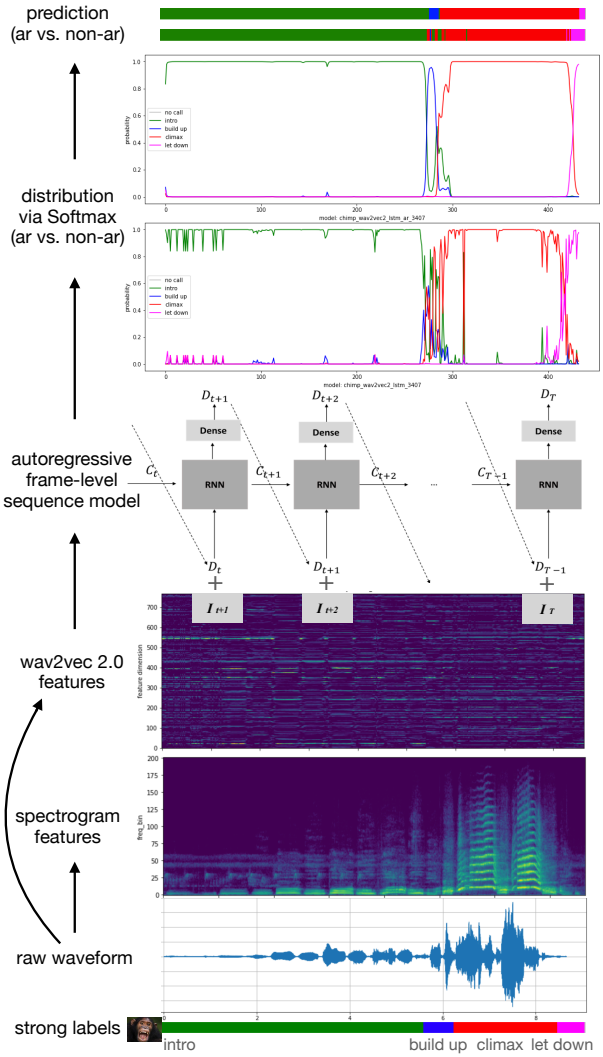


Figure 1: Our method (bottom-up). The audio clip shown in the figure is from the chimpanzee data set. *(Non-)ar* stands for (non-)autoregressive.

5.1. Initial Exploration with Chimpanzee Data

We first test the viability of our approach on the chimpanzee data in light of the simplicity of the pant-hoot annotation scheme. We start from *E1*, a simple waveform + LSTM baseline, and observe in *E2* that wav2vec 2.0 outperforms the raw waveform and spectrogram by a large margin, which demonstrates the power of transfer learning from pretraining on human speech. In *E2.1*, we find that the Transformer encoder does not outperform LSTM. Hence, we infer that Transformer's ability to capture arbitrary long-range dependency is not

ID	data	feature	model	dev acc.	dev f1	test acc.	test f1	aucpr
Explore the best feature and model combination								
E1	chimp	waveform	lstm (baseline)	51.7 ± 2.1	35.4 ± 2.5	51.0 ± 3.6	34.7 ± 4.3	-
E1.1	chimp	spectrogram	lstm	60.3 ± 1.5	55.7 ± 2.3	58.7 ± 4.7	53.9 ± 5.4	-
E2	chimp	wav2vec2	lstm	81.0 ± 4.0	79.9 ± 4.6	79.3 ± 2.3	77.9 ± 3.6	-
E2.1	chimp	wav2vec2	transformer	71.3 ± 0.6	68.3 ± 0.2	75.3 ± 0.6	72.1 ± 0.5	-
Explore the hyper-parameters								
E3.1	chimp	wav2vec2	lstm (E2 + $batch\_size = 4$)	69.7 ± 1.5	71.8 ± 2.6	67.7 ± 4.0	69.6 ± 4.0	-
E3.2	chimp	wav2vec2	lstm (E2 + $batch\_size = 8$)	63.3 ± 0.6	62.6 ± 1.0	62.0 ± 4.4	61.5 ± 4.0	-
E3.3	chimp	wav2vec2	lstm (E2 + $dropout = 0.2$)	80.7 ± 3.5	80.0 ± 4.4	78.0 ± 1.7	76.8 ± 2.7	-
E3.4	chimp	wav2vec2	lstm (E2 + $dropout = 0.1$)	81.0 ± 4.0	80.2 ± 4.8	78.7 ± 2.9	77.3 ± 3.9	-
E3.5	chimp	wav2vec2	lstm (E2 - *balance_weights*)	81.0 ± 3.6	79.6 ± 4.4	79.3 ± 2.3	78.3 ± 3.6	-
Explore autoregressive modeling								
E4	chimp	wav2vec2	lstm (E2 + autoregressive)	87.7 ± 1.2	87.1 ± 1.8	85.7 ± 2.1	85.6 ± 2.5	-
Extend to orangutan long calls and a binary setting								
E5	orang	wav2vec2	lstm (= E4)	83.0 ± 1.0	82.7 ± 1.4	81.7 ± 3.1	82.0 ± 2.6	-
E5.1	orang	wav2vec2	lstm (E5 + binary target)	92.3 ± 2.5	92.1 ± 2.5	92.0 ± 1.0	91.9 ± 1.1	0.96
Extend to bonobo calls and a binary setting								
E6	bonobo	wav2vec2	lstm (= E4)	87.0 ± 4.6	85.9 ± 6.3	83.7 ± 3.8	82.3 ± 2.2	-
E6.1	bonobo	wav2vec2	lstm (E6 + binary target)	92.0 ± 3.6	91.9 ± 3.4	87.7 ± 3.5	87.8 ± 2.9	0.87
Zero-shot transferring from orangutan to bonobo								
E7	bonobo	wav2vec2	lstm (= E5.1)	63.0 ± 13	69.2 ± 10	72.0 ± 4.0	74.2 ± 3.1	0.55

Table 2: Experimental results. We run all experiments three times based on different random seeds and report the mean and standard deviation. *acc.* stands for frame-level accuracy, *f1* stands for the frame-level average F1-score weighted by the number of true instances per class, and *aucpr* stands for the area under the precision-recall curve for the positive class in the binary case at test time when the random seed is set to 0. For hyper-parameters, we start *E1* with $batch\_size = 1$, $dropout = 0.4$ and keep them by default, if not otherwise specified in the table.

beneficial to our task and/or our small data sets. Next, we explore the hyper-parameters in the second experimental group and we find in *E3.5* that balancing class weights (§4.4) has a small impact. Lastly, we show in *E4* that the autoregressive connections are beneficial for consistent output, as illustrated by the tiny gaps in *non-ar* output in Fig. 1.

5.2. Extending to Orangutan and Bonobo Data

We successfully extend the model in *E4* to as trained on the orangutan (*E5*) and bonobo (*E6*) data sets. We note that some minority classes perform less well due to data scarcity, in contrast to the well-balanced situation in the chimpanzee data set (see Table 1).

We further reduce the task to a binary (call vs. non-call) classification that resembles voice activity detection of human speech. This is a useful tool to automatically extract calls from raw recordings for further bioacoustic analysis (e.g., studying the repertoire of a given species).

Finally, to understand the generalizability of our models, we try zero-shot transferring the model trained in *E5.1* on orangutans directly to unseen bonobo data. The results show that it is promising to build a potential general-purpose sound event detection model for all great ape calls.

6. DISCUSSION

We have addressed a gap in the bioacoustics research of non-human great apes by developing an approach for automatically identifying sound events and classifying great ape calls using a neural network architecture. Our method successfully and accurately identifies and classifies calls in three species of non-human great apes, and provides a tool for primatologists to bootstrap call identification and analysis from raw unannotated audio recordings.

Our method also shows the general applicability of the wav2vec 2.0 model trained on human speech for identifying vocalizations and call types in other species. Future work may apply our approach to more animals, as part of the goal to decode the communication systems of great apes and other non-human animals more broadly.[2]

[1] https://github.com/J22Melody/sed_great_ape

[2] For example: https://www.earthspecies.org.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from: UK Research & Innovation, Future Leaders Fellowship (MR/T04229X/1; ARL), Swiss National Science Foundation (PCEFP1_186841: SM, ZJ; 310030_185324: AS), St Leonard College (AS), and Swissuniversities (AS). We thank also Klaus Zuberbühler, Josep Call, and the field assistants of the Budongo Conservation Field Station.

8. REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[3] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[4] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[5] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[6] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[7] J. Ebbers and R. Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, 2021.

[8] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[9] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[12] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using crnn with learnable gated linear units," *arXiv preprint arXiv:1811.07072*, 2018.

[13] P. Fedurek, K. Zuberbühler, and C. D. Dahl, "Sequential information in a great ape utterance," *Scientific Reports*, vol. 6, no. 1, pp. 1–11, 2016.

[14] Y. Sun, T. M. Maeda, C. Solis-Lemus, D. Pimentel-Alarcon, and Z. Burivalova, "Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks," *arXiv preprint arXiv:2111.14971*, 2021.

[15] N. M. Recalde, "pykanto: a Python library to accelerate research on wild bird song," *arXiv preprint arXiv:2302.10340*, 2023.

[16] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[17] P. Marler and L. Hobbett, "Individuality in a long-range vocalization of wild chimpanzees," *Zeitschrift für Tierpsychologie*, vol. 38, no. 1, pp. 97–109, 1975.

[18] A. Soldati, P. Fedurek, G. Dezecache, J. Call, and K. Zuberbühler, "Audience sensitivity in chimpanzee display pant hoots," *Animal Behaviour*, vol. 190, pp. 23–40, 2022.

[19] M. Hardus, "A description of the orangutan's vocal and sound repertoire, with a focus on geographical variation," *Orangutans: Geographic Variation in Behavioral Ecology and Conservation*, pp. 49–64, 2009.

[20] A. R. Lameira and S. A. Wich, "Orangutan long call degradation and individuality over distance: a playback approach," *International Journal of Primatology*, vol. 29, no. 3, pp. 615–625, 2008.

[21] I. Schamberg, D. L. Cheney, Z. Clay, G. Hohmann, and R. M. Seyfarth, "Call combinations, vocal exchanges and interparty movement in wild bonobos," *Animal Behaviour*, vol. 122, pp. 109–116, 2016.

[22] D. Picard, "Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision," *arXiv preprint arXiv:2109.08203*, 2021.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

PRAK: AN AUTOMATIC PHONETIC ALIGNMENT TOOL FOR CZECH

Václav Hanžl[1], Adléta Hanžlová[2]

[1]Freelance researcher, [2]Institute of Phonetics, Charles University in Prague
vhanzl@gmail.com, adleta.hanzlova@gmail.com

ABSTRACT

Labeling speech down to the identity and time boundaries of phones is a labor-intensive part of phonetic research. To simplify this work, we created a free open-source tool generating phone sequences from Czech text and time-aligning them with audio.

Low architecture complexity makes the design approachable for students of phonetics. Acoustic model ReLU NN with 56k weights was trained using PyTorch on small CommonVoice data. Alignment and variant selection decoder is implemented in Python with matrix library.

A Czech pronunciation generator is composed of simple rule-based blocks capturing the logic of the language where possible, allowing modification of transcription approach details.

Compared to tools used until now, data preparation efficiency improved, the tool is usable on Mac, Linux and Windows in Praat GUI or command line, achieves mostly correct pronunciation variant choice including glottal stop detection, algorithmically captures most of Czech assimilation logic and is both didactic and practical.

Keywords: phonetic alignment, segmentation, PyTorch, Czech, Praat

1. INTRODUCTION

Labeling speech recordings and identifying phone boundaries is a significant part of phonetic research. Even though there are software tools to automate this process, many of them target only languages with great quantities of speakers and most are also not freely available, their use requires complex installation and can be restricted by license agreements. Tool choices are even more limited for less common languages like Czech.

Praat's [1] own integrated alignment tool is usable when aligning individual words or short sentences, but its performance is significantly worsened when the audio contains pauses. It also only works from the Praat *View & Edit* window menu and doesn't enable automatic alignment of larger datasets.

The Czech aligner used in academic research so far and preferred as the most useful available is Prague Labeller [2, 3] which is based on HTK GMM models and shows admirable longevity. Subsequent research used Kaldi [4]. Other ad-hoc alignment solutions were never finalized as universally usable.

As identifying words and phones and their time boundaries in a recording is very useful in phonetics not only for research purposes, but also for educational needs and general better orientation in the audio, we find it important that a cross-platform tool enabling the automation of such processes and striving to be as precise as possible is widely available and easy to access by phoneticians and students and requires no programming knowledge from the user. In this paper, we present Prak – a tool we developed with this idea in mind.

2. DESIGN GOALS AND SCOPE

Observing current practices when preparing Czech phone-aligned research data, we had several goals for our new alignment tool:

- an open-source tool free for any use – MIT license [5] for code, trained on free audio data
- functioning on Mac, Linux and Windows
- easy to install – low dependencies, only reliable dependencies which are (hopefully) here to stay
- simple architecture, preferably building on techniques from phonetics students' curricula
- usable from both GUI and command line
- using explainable and modifiable logic (rather than a trained blackbox) where possible
- automatic pronunciation variant selection

The tool expects an audio recording and text transcript as an input. While we can imagine doing the transcript via ASR, we deliberately left this task to an external tool if so desired (manual transcription is a noticeably smaller task than phone boundary adjustments). This way we can always use a state of the art ASR and combine it with our phone alignment tool, without modifying the ASR tool for phone alignment needs – merging these tasks into one tool used to be a vital approach in the GMM HTK era, but modern ASR tool construction often abandons the notion of phones altogether.

The initial release supports Czech language only, as practical help to Czech research was our primary



Figure 1: Output textgrid of Prak used in Praat.

goal. Nevertheless, we tried to make extensions to more languages easy by using a multilingual training data source and by modular design of the code.

3. DESIGN

3.1. Training data

Czech language has high quality speech corpora for acoustic model (AM) training, however, many data sets are unsuitable for a free tool. Some are commercial and costly (SpeechDAT [6], Speecon [7]), some free of cost, but for research use only. There is no LibriSpeech [8] or TEDx for Czech. Corpus of audio recordings from the Chamber of Deputies of the Parliament of the Czech Republic available freely online [9, 10] is big and free, but contains reverberations stronger than standard studio recordings. Some Czech TV recordings are available, but have a low number of speakers.

We decided to first try the CommonVoice (CV) [11] dataset, the size of which is marginal for Czech (14 hours of training data, 4 hours being silence), but it has suitable recording conditions, contains a high number of speakers, has an extremely permissive license and is available for many other languages. To our surprise, CV itself allowed us to train an AM of sufficient quality, so we initially released Prak with a model trained on CV alone.

We are also very grateful to the Institute of Phonetics, Charles University in Prague, for providing their manually labeled recordings from which we selected 5 hours of easily usable data. We did not use this data in any training to keep our tool free of additional dependencies, but we used 10% of this dataset to evaluate the tool by comparison to human labeling made by potential users themselves.

3.2. Software tools

Currently available free and precise deep learning ASR tools [12] naturally come to mind as building blocks, however, the shift towards direct mapping between text and audio omits the phone level, requiring significant modifications of these complex tools to repurpose them for phone alignment.

```
      ʔ          eu
f=ʔu:=.es=.a: pnéʊ̂matɪkɪ za_ éʊ̂ra nekóʊ̂pi:š
v=        ʔ                 _ʔ
                            |ʔ
```

Figure 2: Command line output showing phonetic transcription with pronunciation variants.

From older ASR tools, Kaldi [13] is still close to the traditional HMM-based approach of HTK and explicitly uses phones. It is however quite a complex dependency in a software system. We therefore used a modern NN toolkit PyTorch [14] for construction of the phone AM, but designed our ASR-like infrastructure from scratch in Python [15] using a classical HMM approach. As phone alignment is quite a limited subset of ASR, we expected this to be feasible.

The matrix library in PyTorch also has a very close alternative in the even more broadly used NumPy [16], allowing easy reimplementation of at least the inference part, should such need ever arise.

3.3. Integration in Praat GUI

The Praat [1] integration of Prak is designed to be simple to use without any programming knowledge. A Praat script which embeds the main Python tool is easy to add to Praat dynamic menu, so users can align text with audio in just one click – see fig. 1 for example result. The script also performs several input file checks and adds additional features such as aligning multiple sounds using one text source.

4. PRONUNCIATION GENERATOR

4.1. Phonetic alphabets

Alignment output in Praat TextGrid files uses SAMPA [17] to stay compatible with established practice. Terminal (command line) output uses much more readable IPA for Czech as used in [18], see fig. 2 for an example of generated pronunciation and its variants. Internally (in the source code) the tool uses programming-friendly encoding where each phone is represented by one Czech character.

4.2. Text cleanup

Great care was taken to be able to process all usual texts on all the supported platforms. Apart from notoriously known end-of-line encoding differences [19], there are more subtle pitfalls like invisible Unicode BOM characters, NFC and NFD forms of accented characters etc. We hopefully made it safe to process all the variants happening in practice.

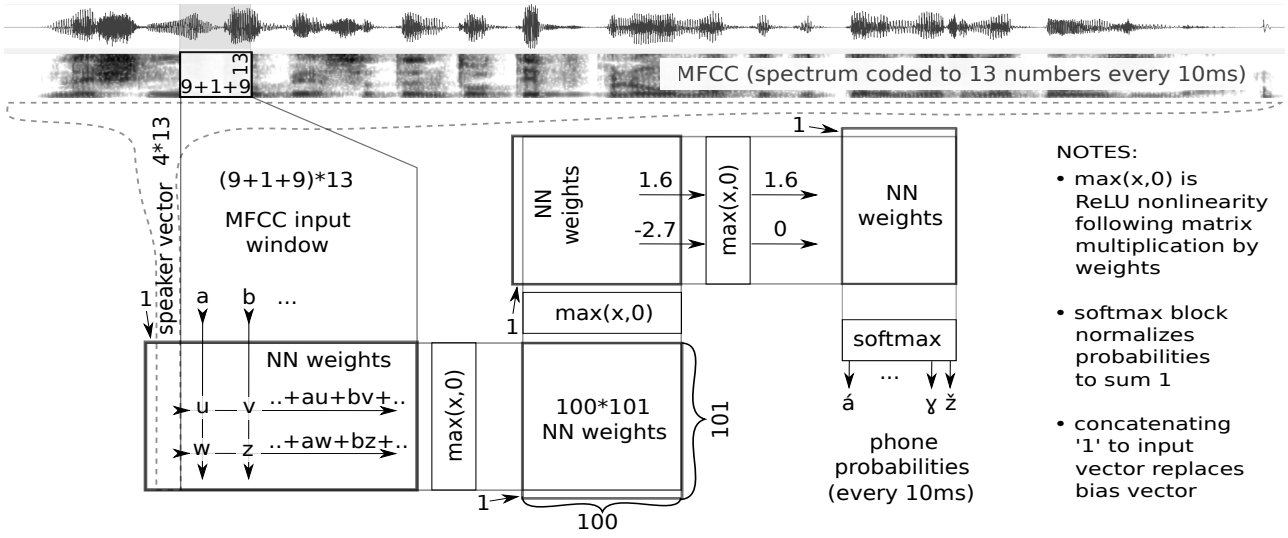


Figure 3: Acoustic model with ReLU NN stack converting MFCC window to phone probabilities.

4.3. Foreign words

Users can add exceptions in the form of replacement rules. For each word, the longest match in rules is replaced by all pronunciation variants from the rule (e.g. rule ”washington vošingtn” can do replacement in ”Washingtonu”). Replaced text is not matched anymore, but parts of words before and after replacement are subject to potential additional (shorter) matches. The resulting text eliminates differences between foreign and native Czech words.

We find the longest match priority more practical than the approach used in [3] where rules are tried in sequence, making it hard to identify the right place for adding new rules among hundreds of old ones.

4.4. Assimilation logic in backward FSTs

Many Czech assimilations can be expressed as Finite State Transducer (FST) with a small number of states (often just 2 or 3) processing the phone sequence backwards, including rather far distance effects. For example, in the sequence ”vošingtnu” (made by the replacement rule above), processing [t] changes FST state to *devoicing* and [g] turns into voiceless [k]. In another FST, [k] or [g] forces state *velarization* and affected [n] becomes [ŋ]. We use a convolution of FSTs taking care of:

- glottal stop [ʔ], intervocalic [j]
- assimilation of voicing
- consonant groups containing dtn/ďťň such as “ntní” [ntɲiː/ncɲiː/ɲcɲiː]
- bě/pě/vě/mě/fě [bjɛ, pjɛ, vjɛ, mɲɛ, fjɛ]
- velarization in nk/ng [ŋk/ŋg]

FSTs provide multiple outputs (needed esp. on word boundaries), leaving the final choice to the AM.

5. PHONE ACOUSTIC MODEL

While transformers [20, 21] of various kinds or at least convolutional structures come to mind as appropriate state of the art in speech processing, our limited goal of only phone alignment seemed achievable with much simpler architectures, significantly lowering the barrier for students trying to grasp the inner workings of the tool. We therefore started with a simple stack of layers composed of matrix multiplication and ReLU – Rectifying Linear Units, $\max(x, 0)$ – applied to each element of a vector which is passed between layers (see fig. 3).

As input, the NN gets 19 consecutive MFCC frames plus a ”speaker vector” which is an average MFCC in 4 energy bands (frames are split to above and below average energy and then sub-split again). We compute Kaldi-like MFCCs via PyTorch.

The stack of ReLU layers only maps one position in the audio to one phone hypothesis, not sequence to sequence. For this purpose, we stuck to a classical HMM structure, replacing a GMM model with ReLU NN. Our training is a cross between Baum-Welch reestimation and NN training by gradient descent, alternating phases of time re-alignment of the target phone sequence and gradient descent of the AM guessing these phones.

Phone sequences are guessed by Viterbi alignment using phone probabilities estimated by NN AM, while the NN AM itself is trained by gradient descent, using the phone sequences as training



Figure 4: Phone alignments during training, rows show train epochs. An optional second glottal stop is only found later, leading to [ʔobjɛʔu...]

targets. Tools like HTK solve similar circular dependencies in Baum-Welch reestimation by a wild initial alignment guess. What we do is similar: For the initial alignment, every phone is presupposed to be 30ms long with equal silences preceding and following speech in a recording. The system is able to converge from this (mostly false) initial alignment to a very good phone AM and very good phone alignments. Fig. 4 shows example training dynamic.

The only adjustment needed for convergence is an artificial boost of rare phones. In Viterbi decoding, we adjusted phone probabilities according to the current overall number of frames classified as this phone in the whole training dataset.

6. EVALUATION OF PHONE ALIGNMENTS

We compared the performance of Prak with Prague Labeller [3], using manually aligned data as a reference. Direct comparison is not easy as phone sets differ between aligners (e.g. Prak also detects glottal stops, uses both voiced and voiceless ”ř”, considers assimilation at word boundaries etc.). The manual alignment, while made with great effort and care, uses several slightly different approaches and contains some phone identity errors. An important parameter for practical users is the frequency of misalignments which must be corrected by moving multiple phone boundaries, therefore we tried to detect major phone boundary misplacements.

We compared pronunciations, counting phone insertions, deletions and substitutions. At places of match, we measured phone center time shifts against the manual reference (while the exact phone identity may be questioned,[1] the time positions are very exact in our manual reference data). The count of time mismatches exceeding a threshold was used as a quality measure, results are shown in tab. 1.

test	PL [3]	Prak
phone mismatch	6.61	1.88
match but misplace 0.1s+	**4.28**	**0.36**
match but misplace 0.2s+	3.22	0.09
mismatch or misplace 0.1s+	10.89	2.24

Table 1: Percentage of phone mismatch and boundary misplacement, comparing most used tool to ours.

7. FUTURE WORK

We highly emphasized simplicity, the only bigger tool being a deep learning framework, only applied to a basic ReLU stack. We avoided FST toolkit (a basic tool which Kaldi explicitly uses, unlike HTK where this functionality is present but dissolved in higher layers) and used a simple ”sausage” structure to capture pronunciation variants. There are many wonderful ready-to-use building blocks in current ASR toolkits which would be great to test in this task.

While modern ASR tools abandoned explicit design of middle layer representation like phonemes or phones, it would be possible to map text to a phone sequence and train e.g. wav2vec2 [22] architecture on phones instead of characters and extract time boundaries of phones using a method such as [23].

Another option would be to use more complex NN structures like transformers for the AM. The question is how much would time boundaries be smeared by the structure paying attention to too distant parts of audio. Such AM would perhaps only be useful as a first layer alignment anchoring audio to text on the level of words, followed by fine-grained alignment by a more local AM like ours.

Phone boundaries could be further fine-tuned. Movements of boundaries towards spectrum change might help for some phone pairs, different processing for others – in fact, the challenge here is translating [18] into a programming language.

Our speaker-vectors can certainly be replaced by i-vectors [13] or x-vectors [24], trading better results for considerably increased complexity.

We hope that our alignment tool can serve as a useful framework and baseline for such experiments.

8. CONCLUSION

We made a practical Czech phone alignment tool which significantly outperforms the best tool available so far in the boundary misalignment measure. Prak also provides new features like automatic variant selection and is more resilient to problematic input texts. We consider our biggest achievement that we gave researchers and students an open-source tool with no restrictions and simple to understand architecture, hopefully allowing others to build on it. Free download of the source code is possible here: `github.com/vaclavhanzl/prak`

9. REFERENCES

[1] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]. Version 6.1.56*, http://www.praat.org/, 2021.

[2] P. Pollák, J. Volín, and R. Skarnitzl, "Influence of HMM´s parameters on the accuracy of phone segmentation - evaluation baseline," in *Proceedings of the 16th Conference Joined with the 15th Czech-German Workshop "Speech Processing"*, vol. 1, 2005, pp. 302–309.

[3] ——, "HMM-based phonetic segmentation in Praat environment," in *The XII International Conference Speech and Computer - SPECOM*, 2007, pp. 537–541.

[4] Z. Patc, P. Mizera, and P. Pollak, "Phonetic segmentation using KALDI and reduced pronunciation detection in causal Czech speech," in *Text, Speech, and Dialogue*. Springer International Publishing, 2015, pp. 433–441.

[5] "The MIT license." [Online]. Available: https://opensource.org/licenses/MIT

[6] P. Pollák, J. Černocký, J. Boudy, K. Choukri, J. Kochanina, W. Majewski, V. Ostroukhov, M. Rusko, J. Sadowski, R. Siemund, P. Staroniewicz, M. Trnka, H. Tropf, K. Vicsi, H. Heuvel, and A. Virag, "SpeechDat(E) - Eastern European telephone speech databases," in *Proceedings LREC'2000 Satelite workshop XLDB - Very Large Telephone Speech Databases*, 2000, pp. 20–25.

[7] "Czech Speecon database." [Online]. Available: http://catalog.elra.info/en-us/repository/browse/ELRA-S0298/

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[9] M. Kopp, V. Stankov, O. Bojar, B. Hladká, and P. Straňák, *ParCzech 3.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2021. [Online]. Available: http://hdl.handle.net/11234/1-3631

[10] "Společná česko-slovenská digitální parlamentní knihovna." [Online]. Available: https://www.psp.cz/eknih/

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[15] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. CreateSpace, 2009.

[16] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[17] "Czech SAMPA." [Online]. Available: https://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm

[18] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Epocha, 2010.

[19] "CR/LF issues and text line-endings." [Online]. Available: https://portal.perforce.com/s/article/3096

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] M. Hira, "Forced alignment with Wav2Vec2." [Online]. Available: https://pytorch.org/audio/main/tutorials/forced_alignment_tutorial.html#forced-alignment-with-wav2vec2

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[1] We manually checked the test set and corrected obvious errors, however given its size (there are 20k phones in the 10% test subset of the full 5 hour set), we certainly did not correct all errors even by this additional pass of human work.

SPEAKER INTERPOLATION BASED DATA AUGMENTATION FOR AUTOMATIC SPEECH RECOGNITION

Lisa Kerle[1], Michael Pucher[2], Barbara Schuppler[3]

[1,2,3]Signal Processing and Speech Communication Laboratory, Graz University of Technology

[2]Austrian Research Institute for Artificial Intelligence

[1]lisa-kerle@gmx.at, [2]michael.pucher@tugraz.at, [3]b.schuppler@tugraz.at

ABSTRACT

In recent years, the development of automatic speech recognition systems has ensured their widespread use in a broad range of areas. Most of these systems, however, require large amounts of training data, making them less suitable for low-resourced languages and for smaller varieties of (well-resourced) languages. This paper focuses on improving automatic speech recognition for Austrian German by means of training data augmentation through neural network-based text-to-speech synthesis. For this purpose, speaker embedding vectors are extracted from an existing corpus and subsequent interpolation between these vectors is used for the generation of new voices. Synthesised speech is then used to train an automatic speech recognition system, while comparing differently large portions of synthesised speech in the training data. Overall, we find that performance improves when the ratio of real and synthesised speech is in the same order of magnitude.

Keywords: speech synthesis, automatic speech recognition, data augmentation, Austrian German

1. INTRODUCTION

Given the increasing performance of automatic speech recognition (ASR) systems, they are also used more broadly in everyday life. Modern ASR systems are based on neural networks (e.g., [1, 2]), requiring thousands of hours of speech, ideally covering a high degree of acoustic diversity [3]. When operating in a low-resource scenario, including ASR for smaller varieties of (well-resourced) languages, or for less resourced speaking styles, large performance drops have been reported [4].

In order to improve ASR performance, different approaches on data augmentation have been investigated. One method is text augmentation by means of the mix-up method, which describes the swapping of words between two or more sentences [5]. Another method is to acoustically augment the training data, for which Wang et al. [6] modified speaker characteristics using a non-autoregressive, non-parallel voice conversion model. To improve robustness in ASR training, Lam et al. [7] applied aligned data augmentation, that uses the replacement of certain tokens in an original audio-text pair and the corresponding adjustment of the audio representations.

Parallel to the development of ASR systems, also speech synthesis has advanced by integrating neural networks, leading to the generation of more naturally-sounding, human-like speech. State-of-the-art speech synthesis systems at an end-to-end level (e.g., Tacotron2) generate speech with high naturalness based on a sequence-to-sequence prediction network with attention mechanism [8, 9]. However, more conventional systems using deep neural networks for acoustic modeling (e.g., feed-forward deep neural network (FFDNN)-based synthesis) also produce intelligible speech at high quality [10].

Previous research has shown the successful combination of speech synthesis and ASR: In [3], speech is synthesised with randomly selected voice profiles and used in addition to real speech data for training an ASR system resulting in an improved ASR performance of 12.5% WER compared to speech recognition using real speech only. [11] generated speaker information by sampling from observed speaker representations to increase the speaker diversity of a training set of an ASR system and showed the effectiveness of augmenting real training data with synthesised training data (4% WER improvement).

This paper aims at investigating whether the augmentation of training data for an ASR system by means of synthesised voices improves its performance for Austrian German, a low-resourced variety of German and contains a variation of the work described in [12]. For this purpose, we generate new speaker characteristics by applying linear interpolation to speakers from an already existing corpus. The generated speaker information is passed to a speech synthesis system in order to generate speech

with new voice characteristics. Synthesised speech is subsequently used to train the ASR system, while comparing differently large portions of synthesised speech in the training data. In all experiments, we use the text from the original training, thus keeping the lexical variety equal across all experiments.

2. MATERIALS

Experiments were based on two corpora of read Austrian German. The first one, a corpus created for speech synthesis experiments, was the Wiener Corpus of Austrian Varieties for Speech Synthesis (WASS) with read speech from a total of 19 speakers of standard Austrian German (6f, 13m) [13, 14, 15]. The reading material contains, among others, sentences from the Berlin-Marburg corpus and the Kiel corpus, resulting in a total of 8293 utterances. Approx. half of the material was read by the same professional male. Second, we used the Graz corpus of Read And Spontaneous Speech (GRASS) [16, 17], which contains a total of approx. 30h of speech from 38 speakers of similar social but different regional backgrounds. In this work, we only used the read speech component of GRASS. Table 1 provides an overview of the data used in the experiments of this paper. One main difference between the WASS and the GRASS corpus is that whereas in WASS a small number of speakers produced a large number of sentences, in GRASS a larger number of speakers read a smaller number of sentences.

set	speakers		utterances	duration (s)
	m	f		
WASS				
train	10	3	6955	20483
dev	1	1	446	1433
test	1	1	446	1395
GRASS				
train	17	16	3820	19914
dev	1	1	249	1329
test	1	1	254	1157

Table 1: Training, development and test set of the WASS and GRASS corpus, with the corresponding number of utterances and duration in seconds.

3. GENERATING SYNTHESISED VOICES

3.1. Method

We used the feed-forward deep neural network speech synthesis system Merlin [10, 18]. To generate speech with specific speaker characteristics, we passed speaker information in form of speaker embedding vectors to the speech synthesis system in addition to linguistic features (i.e., quinphone, syllable, word, phrase).

We generated speaker embedding vectors from the WASS corpus using a Pytorch-Kaldi based speaker recognition system [19, 20]. The system employs a learnable dictionary encoding layer as pooling layer, after which the speaker information is extracted in form of speaker embedding vectors [21]. The quality of the speaker embedding vectors depends on the *minimum required frame length*, the *chunk size* and the *speaker embedding dimension*. First, we chose these parameters in accordance with previous work [20]. Next, we tuned these parameters by considering the *equal error rate* (EER) and the *minimum detection cost function* (minDCF) of the speaker recognition system, and additionally the visual representation of the extracted speaker embedding vectors using the *t-distributed stochastic neighbour embedding* (t-SNE). The parameter combination leading to the lowest EER (6.05%) and minDCF (0.27) and a good separation in terms of the t-SNE was used to extract the speaker embedding vectors from the WASS corpus.

For generating new speaker characteristics, we linearly interpolated between the extracted speaker embedding vectors according to:

$$v_{12} = \alpha \cdot v_1 + (1-\alpha) \cdot v_2,$$

where v_1 and v_2 describe the embedding of the original speakers and by choosing an interpolation factor $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, a vector v_{12} containing new speaker characteristics is obtained. We then passed the generated speaker embedding vectors to a previously trained average voice model (AVM), in addition to linguistic input features to generate speech with new speaker characteristics. We synthesised two different data sets: *Wsyn* and *Gsyn*. For *Wsyn*, text and interpolated speaker embeddings were based on the WASS corpus. For *Gsyn*, we used text from the GRASS corpus and speaker information was obtained from the interpolated speaker embedding vectors from the WASS corpus. We evaluated the quality of the interpolated speaker embedding vectors by passing utterances of *Gsyn* to the speaker recognition system and considering the EER and the minDCF. Additionally, the extracted speaker embedding vectors were analysed by means of the t-SNE. For example, speech synthesised by interpolation between speakers *kep* and *lsc* led to an EER of 16.05% and a minDCF of 0.53.

Figure 1 shows the t-SNE of the interpolated

speaker embedding vectors. First, speaker embeddings were interpolated, then the interpolated embeddings were used to generate synthetic test data with the AVM, and finally the dimension of the utterance embeddings was reduced using t-SNE. Figure 1 shows that the linear interpolation of speaker embeddings leads to a non-linear but continuous behaviour in the space of utterance embeddings, allowing us to create a continuum of speakers using this DNN architecture. We also performed experiments with attention-based recurrent DNN architectures that did not show such a continuous behavior in the interpolated utterance space. A further qualitative evaluation of the interpolated embedding vectors revealed that the transition from speaker *kep* to *lsc* was clearly perceptible.



Figure 1: t-SNE for utterances of *Gsyn* for speaker combinations of *kep* and *lsc* using different interpolation factors α.

4. ASR RESULTS AND DISCUSSION

4.1. Methods

For ASR, we used Kaldi, a widely used state-of-the-art speech recognition toolkit [22]. In this paper, we applied a recipe described in [23], which follows the conventional GMM/HMM approach and consists of an acoustic model, a language model and a lexicon. The acoustic model includes monophone training as well as triphone training and is based on 13-dimensional MFCCs, which we normalised using cepstral mean and variance normalisation (CMVN). The lexicon was created using a G2P online tool [24] for German German. In order to adapt this lexicon for Austrian German, some adjustments were made in the used recipe (e.g., devoicing all alveolar and postalveolar fricatives and affricates), resulting in a set of 38 phones. The language model based on the SRILM toolkit [25] used a N-gram model with order $N = 3$.

4.2. Baseline Experiments

As baseline, we used the original WASS and GRASS corpus for training, developing and testing. With both training sets, we achieved good results, i.e., a WER below 1% (cf. Table 2), whereby the ASR performed better using the WASS corpus. Although the GRASS corpus had a larger number of speakers, the hours of training data was almost equal for both corpora (cf. Table 1). The better performance of the ASR system on WASS than on GRASS may be explained by the fact that one professional speaker contributed a large proportion of the speech in WASS, i.e., about half of all utterances.

4.3. Corpus Augmentation using Synthesised Speech

We used the synthesised *Wsyn* and *Gsyn* data to augment the WASS and GRASS corpus for training the ASR system. The development and test sets remained the same as in the baseline experiment. We experimented with different amounts of synthesised utterances to find the "optimal" ratio of original and synthesised data. We successively augmented the original GRASS corpus by means of utterances from the synthesised GRASS corpus, for which the synthesised utterances were selected randomly. The amount of synthesised utterances ranged from 0% (original utterances only) to 100% (synthesised utterances only), with tighter steps in the area around 50%. To compare our results with findings from earlier studies [11], we calculated the percentage of synthesized utterances and experimented also with exclusively synthesized speech. Figure 2 shows that in the area of around 50% (i.e., when the amount of original utterances equals the amount of synthesised utterances) best results were obtained and that the performance outperformed that of the baseline experiment (WER of 0.88% is reduced to 0.64%). We made the same observations for the experiments with the WASS corpus, where at a 50% mix of real and augmented data the baseline WER of 0.55% was reduced to 0.33% (cf. Table 2). The augmentation for the WASS corpus thus leads to larger improvements than for the GRASS corpus (i.e., the augmentation is more successful if the interpolated speaker information originates from the same corpus).

To investigate which amount of synthesised data needs to be added to a training corpus to improve ASR, Rosenberg et al. [11] used a fixed amount of synthesised speech and reduced the amount of utterances from the original corpus. They report the

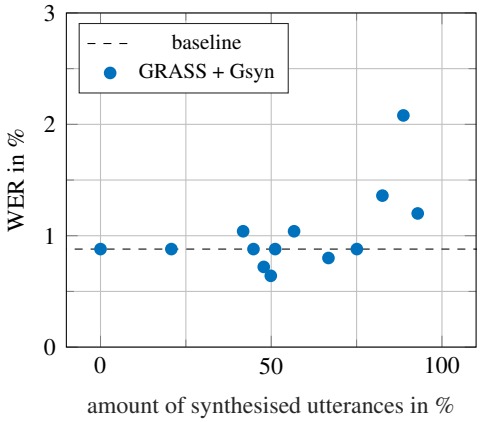


Figure 2: WER in % for different amounts of synthesised utterances of the GRASS corpus. The dashed line shows the baseline experiment.

highest performance when the amount of synthesised and original utterances were equal (at 50%), which is consistent with our observations.

4.4. Using the Synthesised Corpus for Training

We analysed the performance of the ASR system when training it on synthesised speech only (i.e., *Wsyn* and *Gsyn*). The development and test sets were the same as for the baseline experiment. Our experiments showed that ASR performance degrades drastically for *Wsyn* and for *Gsyn*, ASR was almost impossible (cf. Table 2). The poor performance for *Gsyn* might reflect the mismatch between the speakers of the training set (*Gsyn*) and the speakers of the development and test set (GRASS), as the speakers of *Gsyn* resulted from the interpolation between speakers from the WASS corpus and development and test set from speakers of the GRASS corpus. When using *Wsyn* for training, where both text and speaker information come from the same corpus (WASS) results in better ASR performance than *Gsyn*. These results indicate that synthesised speech performs better as ASR training set when generated with text and speaker information from the same corpus. The drastic degradation in ASR performance when using only synthesised material for training was observed in previous work from [11] (32.44% WER).

data		results	
train	**dev/test**	**WER**	**SER**
WASS	WASS	0.55	2.69
GRASS	GRASS	0.88	3.54
WASS + *Wsyn*	WASS	0.33	1.57
GRASS + *Gsyn*	GRASS	0.64	2.36
Wsyn	WASS	13.85	36.55
Gsyn	GRASS	72.92	95.03

Table 2: ASR results in % WER and SER, for the original training set (WASS, GRASS), the augmented training corpus with 50% synthesised speech (WASS + *Wsyn*, GRASS + *Gsyn*) and the synthesised training corpus (*Wsyn*, *Gsyn*).

5. CONCLUSION

The aim of this paper was to investigate whether augmentation of training data by means of synthesised voices improves ASR performance for read Austrian German, a variety of German lacking large amounts of annotated speech resources. We used two corpora of read Austrian German, the WASS and the GRASS corpus, which were created having different applications in mind. WASS, created for speech synthesis applications, contains a large amount of sentences read by few (partly professionally trained) speakers. GRASS, created for ASR, contains fewer different sentences, read by a larger number of speakers. For both corpora, the baseline ASR experiments (i.e., trained on the original GRASS or WASS data and tested on GRASS or WASS respectively) showed good results (WER below 1%). Next, we trained the ASR system with synthesised utterances only. For both WASS and GRASS, the WER degraded drastically, for the WASS corpus to 13.85% WER and for the GRASS corpus to 72.92%, meaning that ASR was basically impossible. Finally, we augmented the original training data with synthesised utterances, which improved the performance of the ASR system for both corpora, even more so when the amount of original utterances was in the order of magnitude of the synthesised utterances, i.e., 0.33% WER for WASS and 0.64% for GRASS. We observed a higher improvement through data augmentation for the WASS corpus, which is the data constellation where the linguistic information and the speaker information originated from the same corpus.

The experiments of this paper dealt with read speech only. As a next step, we will explore synthesis-based data augmentation for conversational speech, where ASR is even more challenging and data sparsity is even more of an issue (e.g., [26]).

6. ACKNOWLEDGEMENTS

This work was supported by grant P-32700-N and I2539-G23 from the Austrian Science Fund (FWF). We thank Julian Linke for his help with the KALDI recipes for GRASS.

7. REFERENCES

[1] Baevski, A., Zhou, Y., Mohamed, A., Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* volume 33, 12449–12460.

[2] Luo, H., Zhang, S., Lei, M., Xie, L. 2021. Simplified self-attention for transformer-based end-to-end speech recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 75–81.

[3] Fazel, A., Yang, W., Liu, Y., Barra-Chicote, R., Meng, Y., Maas, R., Droppo, J. 2021. SynthASR: Unlocking synthetic data for speech recognition. *Proceedings of Interspeech*, 896–900.

[4] Linke, J., Garner, P. N., Kubin, G., Schuppler, B. 2022. Conversational speech recognition needs data? Experiments with Austrian German. *Proceedings of LREC*, 4684–4691.

[5] Kwon, S., Lee, Y. Apr 2022. Explainability-Based Mix-up Approach for Text Data Augmentation. *ACM Transactions on Knowledge Discovery from Data*. Just Accepted.

[6] Wang, G., Rosenberg, A., Ramabhadran, B., Biadsy, F., Emond, J., Huang, Y., Moreno, P. J. 2022. Non-Parallel Voice Conversion for ASR Augmentation. *Proceedings of Interspeech*, 3408–3412.

[7] Lam, T. K., Ohta, M., Schamoni, S., Riezler, S. 2021. On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR. *Proceedings of Interspeech*, 1299–1303.

[8] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., Wu, Y. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *Proceedings of ICASSP*, 4779–4783.

[9] Lim, D., Jung, S., Kim, E. 2022. JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech. Ko, H., Hansen, J. H. L. (eds), *Proceedings of Interspeech*, 21–25.

[10] Wu, Z., Watts, O., King, S. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop*, 202–207.

[11] Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., Wu, Z. 2019. Speech Recognition with Augmented Synthesized Speech. *Proceedings of ASRU*, 996–1002.

[12] Kerle, L. K. 2022. Speaker Interpolation based Data Augmentation for Automatic Speech Recognition. Master's thesis Graz University of Technology.

[13] Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., Schmid, E. 2015. Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games. *Proceedings of Interspeech*, 1625–1629.

[14] Toman, M., Pucher, M. 2015. An Open Source Speech Synthesis Frontend for HTS. *Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302*, 291–298.

[15] Pucher, M., Rausch-Supola, M., Moosmueller, S., Toman, M., Schabus, D., Neubarth, F. 2016. Open data for speech synthesis of Austrian German language varieties. *12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 147–150.

[16] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. *Proceedings of LREC*, 1465–1470.

[17] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[18] Ronanki, S., Wu, Z., Watts, O., King, S. 2016. A Demonstration of the Merlin Open Source Neural Network Speech Synthesis System. *Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop*, 124–124.

[19] Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., Dehak, R., others, 2019. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech & Language* 101026.

[20] Cooper, E., Lai, C.-I., Yasuda, Y., Fang, F., Wang, X., Chen, N., Yamagishi, J. 2020. Zero-Shot Multi-Speaker Text-To-Speech with State-of-the-art Neural Speaker Embeddings. *Proceedings of ICASSP*, 6184–6188.

[21] Cai, W., Chen, J., Li, M. 2018. Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. *The Speaker and Language Recognition Workshop*, 74–81.

[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. Dec. 2011. The Kaldi Speech Recognition Toolkit. *Proceedings of ASRU*.

[23] Linke, J., Wepner, S., Kubin, G., Schuppler, B. 2023. Using Kaldi for automatic speech recognition of conversational Austrian German. *url: http://arxiv.org/abs/2301.06475v1*.

[24] Reichel, U. D. 2012. PermA and Balloon: Tools for string alignment and text processing. *Proceedings of Interspeech*, 1874–1877.

[25] Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. *Proceedings of Interspeech*, 901–904.

[26] Wepner, S., Schuppler, B., Kubin, G. 2022. How prosody affects ASR performance in conversational Austrian German. *Speech Prosody 2022*, 195–199.

Automatic speech recognition: system variability within a sociolinguistically homogenous group of speakers

Lauren Harrington, Vincent Hughes

Department of Language and Linguistic Science, University of York, UK
lauren.harrington@york.ac.uk, vincent.hughes@york.ac.uk

ABSTRACT

Orthographic transcription is a time-consuming task that is increasingly being automated for various applications. A growing body of work on algorithmic bias has highlighted variability in automatic speech recognition performance across demographic groups, but little research has focused on variability within a sociolinguistically homogenous group.

This paper considers this issue and explores the extent to which phonetic properties of the voice (e.g. f0, formants) can predict how well an automatic system performs. Recordings of 99 young, male Southern British English speakers were automatically transcribed using Amazon Transcribe, and word error rate was calculated.

A wide range of error rates (between 11-33%) is observed. However, of all acoustic, temporal and voice quality measures analysed, only articulation rate has a significant impact on error rate. System performance therefore remains difficult to predict within demographic groups.

Keywords: automatic transcription, automatic speech recognition

1. INTRODUCTION

Automatic speech recognition (ASR) technology is used increasingly for a variety of applications, including live captioning, virtual assistants, and transcribing professional meetings. The field of ASR has received an increasing amount of interest in recent years (see [1] for an overview) and is now a huge area of research, with modern systems, integrating state-of-the-art machine learning techniques, demonstrating consistent improvements in overall performance year-on-year. Systems tend to be evaluated on the basis of overall word error rate (WER; the percentage of errors in a transcript relative to words spoken in the audio recording), with recent reports of WERs as low as 5% [2] on the Switchboard corpus, a dataset commonly used for ASR performance analysis which contains General American English conversational telephony data.

There is now a growing focus on understanding how and why ASR systems perform in the way that they do, which has led to the development of a body of work exploring algorithmic bias in ASR systems [3,4]. Some studies have investigated system performance as a function of speaker demographic factors such as accent [5,6] and gender [7], and have demonstrated significant differences in performance across groups. The issue of variability across demographic groups is an area of concern within the field and work on mitigating the effects of performance disparities is gaining popularity [8,9].

The present study explores an issue which has received relatively little attention within the field of ASR: the extent of individual variability in system performance across a sociolinguistically homogeneous set of speakers, matched for age, gender, level of education and regional accent. Here, we keep confounding factors such as speech content and audio quality consistent, in order to isolate differences in performance due to characteristics of the speech and speaker. This work explores the extent to which linguistic properties of a speaker's voice may predict speech recognition performance, using a state-of-the-art, and widely used, commercial system, Amazon Transcribe. To our knowledge, no work has yet considered how the phonetic properties of speech may be contributing to system variability.

In this paper we address two research questions: (1) what is the range of variability in ASR performance across a set of demographically homogeneous speakers? And (2) is any of the variability predictable based on well-understood phonetic properties of a speaker's speech production?

2. METHODOLOGY

2.1. Stimuli

Data from the Dynamic Variability in Speech (DyViS) [10] project was used for this study. DyViS is a large database of speech collected under simulated forensic conditions, consisting of recordings of 100 university-educated young adult male speakers of Standard Southern British English (SSBE) taking part in a range of tasks. The speakers are demographically homogeneous and are generally very similar sounding [11]; indeed, speakers with atypical features for SSBE were removed in a screening phase when collecting the corpus [10]. DyViS task 2 was used for the present study. It

consists of a telephone conversation with an 'accomplice', played by a Research Assistant, in which the participant discusses his experience of a police interview (a previous task) so that the accomplice can 'get their story straight'. Details of the crime discussed in both the police interview and telephone conversation were presented to participants in the form of visual prompts, therefore the content of the speech is relatively controlled across speakers.

Studio recordings with a sampling rate of 44.1 kHz and minimal background noise were used for this study. We decided to use high quality samples in order to maximise the performance of the ASR system and, therefore, minimise the confounding effects of channel and background noise. Speech of the 'accomplice' and non-speech sounds, e.g. coughs and inhales/exhales, were removed such that each file contained only the speech of the participant. One speaker was removed due to issues with his data, resulting in 99 stimuli between 4 and 11 minutes in length.

2.2. Transcripts

Ground truth reference transcripts were based on orthographic transcriptions completed by researchers on the DyViS project and provided with the database in the form of Praat [12] TextGrids for each audio file. The content of the TextGrids was extracted using a Praat script and each resulting transcript was saved in .txt format and manually checked for spelling errors. Minor adjustments were made to correct spelling errors and to ensure consistency in the representation of certain words or phrases, e.g. some transcripts contained "DIY" while others contained "D I Y".

The audio files were automatically transcribed using a popular commercially-available ASR service, Amazon Transcribe via Amazon Web Services. The "en-GB" (British English) language model was used for all transcriptions. As a standard variety, we expected that the system should perform optimally with SSBE compared with non-standard varieties of British English. Again, the intention was to maximise the potential performance of the system under favourable conditions. The Amazon output for each file was saved in .txt format and minor adjustments were made to the representation of filled pauses to ensure consistency between both sets of transcripts (these are referred to below as 'hypothesis' transcripts).

2.3. Performance evaluation

Python package JiWER [13], which computes the minimum edit distance between two strings of text, was used to compare the reference and hypothesis transcripts for each audio file and to calculate the word error rate (WER) for each speaker. A lower WER demonstrates that fewer errors were made. Microsoft Azure documentation [14] posits that a WER of 5-10% is considered good quality, while error rates of over 30% signal bad quality and that the system requires further training and customisation.

A more detailed evaluation of the speakers that produced the best and worst WERs, DyViS speakers 30 and 26 respectively, was carried out to explore the types of errors produced. Reference and hypothesis transcripts were aligned on a word-level basis and word pairs were marked as a match or as one of three types of error: deletion, substitution or insertion.

2.4. Auditory and acoustic measurements

To explore the effects of phonetic properties of speech production on automatic transcription performance, a range of acoustic and auditory measurements for the speakers were collected from existing sources. These were f0, long term formant distributions (F1~F4), and articulation rate from [15] and Vocal Profile Analysis of laryngeal and supralaryngeal voice quality from [11]. One speaker was excluded from further analysis due to an issue with the audio file and inconsistency between the speech content and corresponding transcription, resulting in a total of 99 speakers.

Mean values (and standard deviation for f0) for each speaker were calculated for the acoustic measures. The VPA data provides an array of information about voice quality features from the whole vocal tract. Speakers were assigned a score from 0 to 3 for each setting, e.g. "fronted tongue body" and "raised larynx", to represent the extent to which that setting was present in the speaker's voice. For the purposes of this study, we collapsed the features to produce two voice quality categories, supralaryngeal and laryngeal, and calculated the Euclidean distance for each speaker's profile from the group mode for each category. This provides a single value indicating how unusual a speaker's laryngeal and supralaryngeal profiles are from the average for the group.

2.5. Statistical analysis

A multiple linear regression model was fitted using the *lm* function in R [16] to predict word error rate using each of the phonetic measures as fixed effects. Mixed effects regression was not conducted as there was only one data point per speaker.

3. RESULTS

3.1. Variability in system performance

Performance was measured by calculating WER for each speaker, where a higher percentage indicates a higher proportion of errors to correctly transcribed words. WERs ranged from 11.2% to 33% with a mean error rate of 20% across speakers. Figure 1 displays the variability of the word error rates, and shows a relatively normal distribution across the range. The variability itself is extremely large, especially given the favourable conditions in terms of audio quality and accent. Since the speakers are matched for multiple demographic factors and the content of the utterances is relatively controlled, the variability of WER within the group suggests that characteristics related to a specific speaker's voice must, to some extent, be responsible for the wide range of error rates.

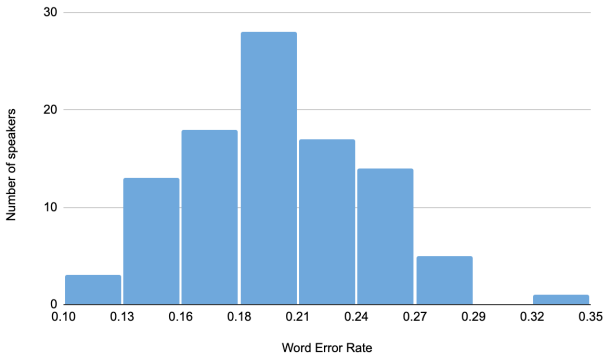


Figure 1: Distribution of word error rate across 99 DyViS speakers.

The number and types of errors most frequently made by Amazon Transcribe also varied across speakers. These can be grouped as (i) deletions (e.g. *going to the steak house → going to steak house*), (ii) substitutions (e.g. *hope avenue → pope avenue*) and (iii) insertions (e.g. *about telephone booth → about the telephone booth*). Table 1 displays the number of each type of error produced for speakers 30 and 26, representing those for whom the system performed best and worst. The total number of errors for speaker 26 is over double that of speaker 30, and the proportion of substitution errors is much higher. A higher number of the substitution errors were related to proper nouns (e.g. place names) for speaker 26 (42 compared with 23), and the system transcribed many names incorrectly for speaker 26 but correctly for speaker 30.

Speaker	WER	DEL	SUB	INS	Total
30	11.2%	115	69	13	197
26	33.0%	216	215	28	459

Table 1: Number of each error type for the speakers for whom Amazon Transcribe performed best (11.2%) and worst (33.0%).

3.2. Phonetic properties

Multiple linear regression was used to test if the phonetic properties listed in 2.4 significantly predicted word error rate. No significant effects were found for mean f0 (β = -0.00055, p = 0.14), f0 standard deviation (β = 0.00136, p = 0.22) or long-term formant distributions for any of the first four formants (F1: β = 0.00007, p = 0.55; F2: β = -0.00012, p = 0.18; F3: β = 0.00004, p = 0.42; F4: β = -0.00004, p = 0.16). The VPA scores of distance from the group mode did not significantly predict WER for either supralaryngeal (β = 0.00599, p = 0.29) or laryngeal (β = 0.00465, p = 0.44) features.

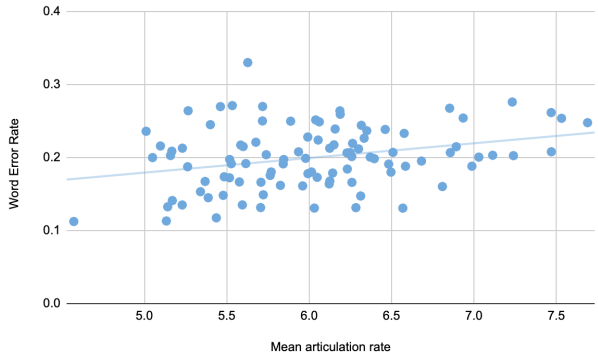


Figure 2: Relationship between word error rate and mean articulation rate (syllables uttered per second).

The only phonetic feature to significantly predict WER was articulation rate (β = 0.01844, $p < .01$) for which there was a positive correlation, demonstrated in Figure 2. Speaker 30, for whom Amazon Transcribe achieved the lowest WER, had the lowest articulation rate in the group (4.6 syllables per second) which was 0.4 lower than the second lowest articulation rate observed within the group and 3.1 lower than the highest.

4. DISCUSSION

Not much is known about why systems may perform better for some speakers than others. This study has provided an initial exploration into the amount of variability observed within a group of speakers matched for demographic factors such as age, sex and level of education, whilst controlling for audio quality and content. We find that there is a high level of variability across speakers, with word error rates

ranging from 11% to 33% and the quality of transcripts varying across speakers. The range of variability itself is worrying given the favourable conditions in the present study and raises issues about the general utility of ASR systems for many applications. Such variability also raises a question about the value of a single WER as a measure of overall system performance across lots of samples and speakers, as is typical in benchmarking systems. Minimally, it would be useful for developers to report the variability in system performance across speakers.

Furthermore, WER is unable to provide information about the types of errors made by the system and how impactful they may be to the meaning conveyed. By more closely inspecting the reference and hypothesis transcripts, it is clear that some transcripts contain many more critical errors than others. For example, the hypothesis transcript for speaker 26, for whom Amazon Transcribe achieved the highest WER, contained several substitution errors which could completely change the meaning, such as *Rose has picked me up* in place of *Rozzers* [police officers] *picked me up*. Meanwhile, the majority of substitutions for speaker 30, for whom Amazon performed best, were related to minor representational issues such as grammatical corrections (e.g. *'cause* → *because*) or contractions/expansions (e.g. *she's* → *she is*). Proper nouns, such as names of colleagues or place names, were a reliable source of errors for Amazon Transcribe. Spelling mistakes and other substitutions were common, although there were many cases where a particular name proved problematic for some but not all speakers. For example, *Weasley* was uttered in 91 stimuli and correctly transcribed at least once in 71, but for speaker 26 this name was substituted with *easily* and *weezy*.

The results in 3.2 reveal that only articulation rate is significantly correlated with WER, and other phonetic properties such as mean f0 cannot significantly predict WER. The lack of acoustic correlates with WER is worrying because there is only limited phonetic evidence as to why a given speaker may be 'easy' or 'difficult' to transcribe. State-of-the art ASR systems extract abstract features and implement deep learning within their training, contributing to the 'black box' problem (i.e. the extent to which the processing is opaque and scrutinizable). This in turn raises the general issue of interpretability; without knowing the cause of variability within a group, it is challenging to improve systems and interpret their results.

5. CONCLUSION

The results of this study show a large amount of ASR performance variability across a set of homogenous speakers, despite keeping several possible confounding factors consistent. Variability is found in overall error rates as well as the number and types of errors produced across speakers. Performance variability across speakers cannot be predicted by well-known acoustic phonetic properties, such as fundamental frequency or long-term formant distribution, or by properties related to voice quality. Only the temporal property, articulation rate, is able to significantly predict error rate, whereby more syllables uttered per second is significantly related to the ratio of errors produced relative to words in the reference transcript. In future work it would be interesting to explore whether other temporal properties, such as disfluency, may be responsible for an increase in error rates.

ACKNOWLEDGMENTS

This research was funded by the White Rose College of Arts and Humanities, a Doctoral Training Partnerships supported by the Arts and Humanities Research Council (AHRC). Thanks also to Phil Harrison for support with Praat scripting, and to Ryan Willis for technical support with Amazon Transcribe and JiWER.

6. REFERENCES

[1] Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Almojil, M. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access* 9, 131858–131876.

[2] Tüske, Z., Saon, G., Audhkhasi, K., Kingsbury, B. 2020. Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard. arXiv preprint arXiv:2001.07263.

[3] Shah, D., Schwartz, H. A., Hovy, D. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. arXiv preprint arXiv:1912.11078.

[4] Suresh, H., Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and access in algorithms, mechanisms, and optimization*, 1–9.

[5] Markl, N. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* Seoul, 521–534.

[6] DiChristofano, A., Shuster, H., Chandra, S., Patwari, N. 2022. Performance Disparities Between Accents in Automatic Speech Recognition. arXiv preprint arXiv:2208.01157.

[7] Tatman, R., Kasten, C. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Proc. Interspeech* Stockholm, 934–938.

[8] Dheram, P., Ramakrishnan, M., Raju, A., Chen, I.-F., King, B., Powell, K., Saboowala, M., Shetty, K., Stolcke, A. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. *Proc. Interspeech* Incheon, 1268–1272.

[9] Zhang, Y., Zhang, Y., Halpern, B., Patel, T., Scharenborg, O. 2022. Mitigating bias against non-native accents. *Proc. Interspeech* Incheon, 3168–3172.

[10] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & the Law* 16(1), 31–57.

[11] San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., Kavanagh, C. 2019. The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association* 49(3), 353–380.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 10 December 2022 from http://www.praat.org/.

[13] Vaessen, N. 2022. JiWER: Similarity measures for automatic speech recognition evaluation. Version 2.5.1, retrieved 16th November 2022 from https://pypi.org/project/jiwer/.

[14] Microsoft Azure. 2022. Test accuracy of a Custom Speech model. https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio.

[15] Gold, E. 2014. Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. PhD thesis, University of York.

[16] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

SLAM 3: AN UPDATED STYLIZATION MODEL FOR SPEECH MELODY

Emmett Strickland[1], Marc Evrard[2], Anne Lacheret-Dujour[1]

[1]MoDyCo, Paris Nanterre University, France
[2]LISN, Paris-Saclay University, France
emmett.strickland@parisnanterre.fr, marc.evrard@lisn.upsaclay.fr, anne.lacheret@parisnanterre.fr

ABSTRACT

This paper presents the newest version of the annotation software Stylization and Labelling of Speech Melody (SLAM), a language-independent prosodic model that automatically annotates pitch contours in linguistic units of arbitrary length. We review the core principles of SLAM before describing several shortcomings and the innovations introduced in SLAM 3 to address them. These notably include methods implemented to minimize the influence of F0 microvariations and alignment errors and to better model the perception of short-duration pitch changes. Secondly, we present additional functionality allowing speech segments to be annotated relative to the mean pitch of their nearest neighbors, reducing the influence of downdrift on annotations. Finally, we demonstrate the utility of these changes by comparing SLAM 3 against its predecessor in terms of measured distances between their stylized outputs and the natural pitch contours used as input.

Keywords: pitch contour, stylization, prosody, automatic labelling

1. FUNDAMENTALS OF THE SLAM MODEL

Stylization and **La**belling of Speech **M**elody (SLAM[1]) is a program for automatically labeling the pitch contours of any linguistic unit fixed by the user. SLAM was developed in line with other perception-based approaches to intonation [1, 2, 3], aiming to produce simplified descriptions of natural F0 contours that represent only perceptually salient variations. SLAM describes pitch contours' relative height and shape using a finite set of discrete labels. The most distinctive benefit of this model is that it can be applied to various linguistic units of any size (e.g., syllables, words or syntactic phrases).

Users are prompted to define two separate *TextGrid* tiers. The *target tier* provides a segmentation based on the linguistic unit the user wishes to study. If the target tier segments the sound file into syllables, labels will be produced for every syllable. Meanwhile, a *support tier* provides a higher-level segmentation providing a reference pitch on which these contours are computed. Users may, for example, select the syllable as the target tier, and the utterance as the support tier. In this case, syllables would be annotated according to their pitch relative to the overall F0 of the wider utterance. Annotations therefore take into account the broader context of linguistic units.

The SLAM model represents each stylized pitch contour as a textual label containing three key pieces of information:

- The initial pitch value of the segment.
- The final pitch value of the segment
- The pitch value and relative position of the most prominent internal saliency, if applicable.

At each of these three points, pitch values are represented as one of five symbolic tones ranging from *H* (extreme-high) to *L* (extreme-low). Each of these tones traditionally covers a pitch span of 4 semitones, with the medium tone being centered on the mean F0 of the support unit.

A high flat contour would therefore be described with the label *hh*, while a falling contour beginning with a medium pitch and ending with a low pitch would be assigned the label *ml*. If an extreme internal F0 value is detected, the contour is described with an additional symbolic tone, while its approximate position is specified using a numeric value ranging from 1 (beginning) to 3 (end). The contour mmh2 describes a contour beginning and ending with a medium pitch and a high pitch maximum in the middle position. These examples are illustrated in Figure 1. These contours are calculated from the raw pitch contours used as input, after an automatic preprocessing phase using the LOWESS smoothing algorithm [4].

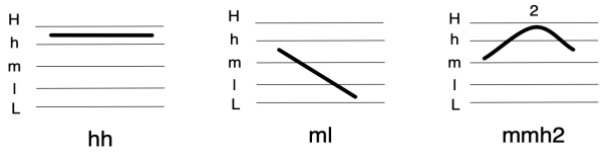


Figure 1: Examples of contour labels

SLAM+, the second major iteration of SLAM, introduced the ability to differentiate between local and global intonational registers, thus producing two

[1] https://github.com/vieenrose/SLAMplus

contours for each target. A *global contour* was based on the mean pitch of the unit located in the support tier. A *local contour* used the target segment as the support, effectively calculating the contour using the mean pitch of the target itself.

For a more detailed overview of the SLAM model, and of the core differences between the prior iterations, see [5] and [6].

2. IMPROVEMENTS IN SLAM 3

The following section presents several shortcomings of the previous iterations of SLAM, as well as the solutions introduced in SLAM 3.

2.1. Alignment errors and F0 microvariations

All stylized labels produced by SLAM include the initial and final pitch values of each alignment. The model is therefore sensitive to minor alignment errors and F0 microvariations at the edges of units. Articulatory constraints can have imperceptible impacts on F0 [7, 8]. If these occur at the edges of a target, the edge pitch values may be unrepresentative of the contour as it is perceived. To address this problem, SLAM 3 performs a linear regression, which allows a better prediction of the pitch values observed in the original contour. The endpoints of this regression are then used to determine the symbolic tones used in the stylization, as shown in Figure 2.

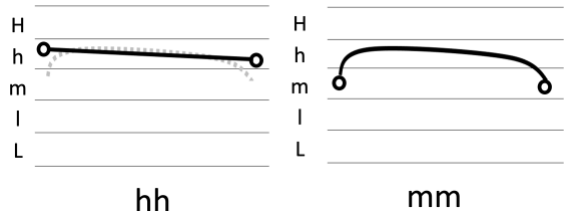


Figure 2: Improved labels enabled by linear regression (hh instead of mm)

Minor alignment errors could also cause significant inaccuracies in previous iterations of SLAM. If a target segment's alignment boundaries were slightly overstepped by a neighboring segment's F0 contour, that inaccurate information would be used in computing the target's stylized contour. This issue is typical in automatically aligned corpora. Figure 3 provides an example of a highly inaccurate contour label introduced by small misalignments at the left and right boundaries of the target.

To reduce these errors, SLAM 3 disregards continuous F0 measurements of 30 milliseconds or less, provided that they occur at the very edge of a target segment and are isolated from other voiced segments by a gap of 50 milliseconds or greater. We find this heuristic sufficient to resolve most observed alignment errors in our corpora without excluding relevant pitch information elsewhere.

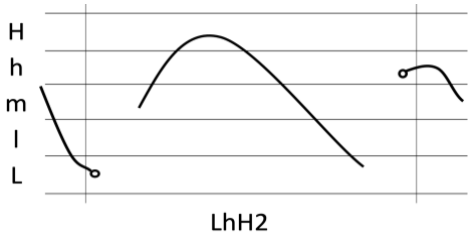


Figure 3: Annotation error caused by misalignment

2.2. Duration and the perception of pitch change

The SLAM model was designed to model pitch contours as they are perceived in linguistic units of any length. Though prior versions of SLAM could theoretically be applied to syllables and other short segments, they neglected the influence of duration in the perception of pitch changes at these levels.

Whether contours are perceived as flat tones or as glissandos is influenced not only by the pitch range traversed from start to finish, but also by the duration of the segments in question. A small pitch rise may be inaudible in a short-duration segment but easily perceived when spread over a longer timespan [9]. The *glissando threshold* refers to the rate of pitch change above which glissandos are perceived, and below which flat tones are heard. Hart et al. [1] estimated this using Formula 1, where G_t represents the threshold in semitones per second, and T the duration of the segment.

$$G_t = \frac{0.16}{T^2} \quad (1)$$

In SLAM 3, this formula is applied several times to ensure that annotations represent only pitch changes that are perceptible within a given duration. For a contour to be annotated as a rise or fall, it must now meet two conditions. As in previous versions of SLAM, its start and end points must fall within separate pitch windows associated with differing symbolic tones. However, the pitch difference between the endpoints must also exceed the glissando threshold calculated according to the timespan between those points. Otherwise, it is annotated as a flat tone whose pitch equals the mean of its endpoints.

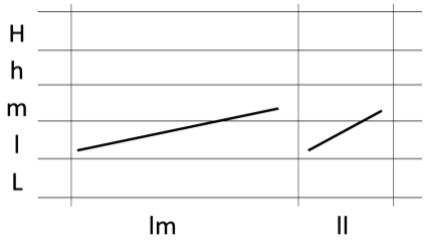


Figure 4: Differentiated annotations based on length

The same principle is used in the detection of F0 saliences, which are only included in the final annotation if the pitch changes on each side exceed the threshold, as shown in Figure 5.

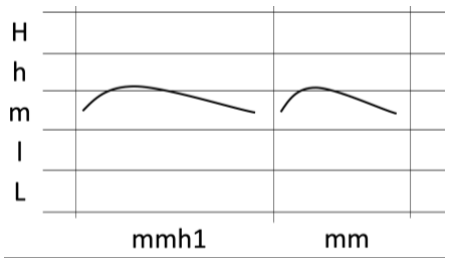


Figure 5: Differentiated annotation of pitch saliencies

In addition to annotations that better represent how short-duration F0 variations are perceived, integrating the glissando threshold also resolves a longstanding problem with the SLAM model. Associating each of the five symbolic tones with a rigid pitch range can result in flawed and inconsistent annotations in some cases. If a nearly flat F0 contour nevertheless crossed the boundary between two symbolic tones, previous versions would automatically annotate it as a glissando, even in cases where the pitch movement was imperceptibly small. This problem is thoroughly described in [10], from which Figure 6 is adapted.

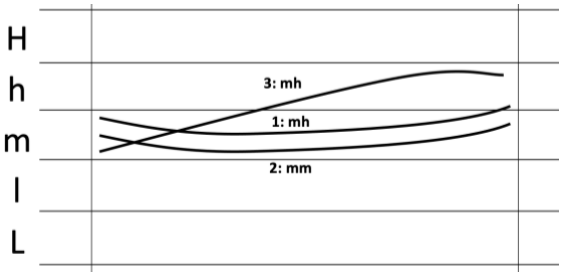


Figure 6: Inconsistent annotations from rigid categories

In this example, an imperceptible difference in height causes contour 1 and contour 2 to receive different annotations (*mh* and *mm* respectively) despite being perceptively indistinguishable. Meanwhile, contour 3 is given the same annotation as 1, despite being highly distinct from it in both shape and pitch range.

The glissando threshold ensures that contours 1 and 2 would be treated as flat tones. Meanwhile, contour 3 would be correctly annotated as a rise (*mh*), assuming its duration is long enough to permit an audible pitch change. In this specific example, linear regressions also provide additional protection against inconsistent annotations. Since contour 1 falls within the medium window for most of its duration, the associated linear regression would fall entirely within the medium window. Consequently, contours 1 and 2 would both be annotated as *mm*.

2.3. Local contours and the F0 of target neighbors

As discussed in section 1, the previous version of SLAM provided two different registers for generating stylized contours: a global register based on the mean F0 of a larger linguistic unit used as the support, and a local register that used the target as the support. Under this system, local contours accurately depicted rises, falls, and internal pitch saliencies. However, flat contours were systematically labeled *mm* regardless of their pitch values because the *m* label was centered on the mean F0 of the segment. This meant that even contours at the extremely high or low ends of the pitch spectrum were labeled *mm*, signaling a high degree of information loss regarding height.

To increase the utility of the local register, SLAM 3 introduces a sliding local support window spanning N units to the left and right of the target. Its size can be directly set by the user. When producing a local contour label for a syllable, the *m* tone label is therefore centered on the mean F0 of a larger segment comprising the target and its neighbors. If it is assigned a size of 1, the window will span from the beginning of the target's left neighbor to the end of its direct neighbor to the right. A window of size 2 will include two neighbors on each side, while a window of size 0 will function like SLAM+ by considering only the mean F0 of the target. Note that segments are only included in this window if they fall within the same support segment. This sliding support window, visualized in Figure 7, also helps to mitigate downdrift's effects on the stylization, since the pitch labels only consider the target's most immediate context.

Utterance 1	*Utterance 2 (global support)*					
...	*Syl 1*	***Syl 2***	*Syl 3*	*Syl 4*	*Syl 5*	N = 0
...	*Syl 1*	***Syl 2***	*Syl 3*	*Syl 4*	*Syl 5*	N = 1
...	*Syl 1*	***Syl 2***	*Syl 3*	*Syl 4*	*Syl 5*	N = 2

Figure 7: Target (Syl 2) and local support of size N

2.4. Customizability and ease of use

This paper has so far presented various improvements introduced in SLAM 3. While the correction of misalignments, the use of linear regressions, the glissando threshold, and the sliding window for local contours are all valuable additions, we have also aimed to make this iteration of SLAM as customizable as possible. All these additions can therefore be activated or deactivated independently in the software parameters. If all of them are deactivated, the software simply functions according to the same principles as its predecessor, SLAM+.

Finally, SLAM 3 allows annotations to be automatically exported to a *tsv* file containing each segment's duration, textual content, and stylized contours in the target tier. This lets users directly analyze results using statistical analysis packages or spreadsheet software.

3. EVALUATING SLAM 3

3.1. Data and methodology

To compare the effectiveness of SLAM 3 to that of its immediate predecessor, we produced a series of stylizations using both versions of the software. Our dataset was comprised of two random samples of ten files selected from two corpora representing languages of differing prosodic typologies: the Rhapsodie [11] corpus of French (Corpus 1), and the NaijaSynCor [12] corpus of Nigerian Pidgin (Corpus 2). In all files, word-level alignments were used as the target, with the utterance used as the support. Contours were also generated using a variant of SLAM 3 that disregarded the glissando threshold (represented as S3-G in section 3.2).

The stylizations were then converted into numerical lists representing pitch contours, from which the mean absolute error (MAE) and mean squared error (MSE) were computed between the reconstructed contours and smoothed versions of the originals. Only the global register was considered in this study, since SLAM 3 bases its local register on a sliding support that was not featured in previous versions.

3.2. Results

	Corpus 1		Corpus 2	
	MAE	**MSE**	**MAE**	**MSE**
S+	1.106	1.467	1.036	1.347
S3	1.021	**1.369**	0.975	1.278
S3-G	**1.011**	1.381	**0.943**	**1.258**

Table 1: Results of evaluation (best scores in bold)

This test yielded similar results on both corpora. SLAM 3 consistently produced a lower MAE and MSE than SLAM+. This indicates a meaningful improvement over the previous version. Excluding the glissando threshold resulted in a more modest reduction in the mean error in most cases. In Corpus 2, this yielded an approximately 3% reduction in both MAE and MSE compared to SLAM 3. The impact of this approach on Corpus 1 is negligible, with an approximately 1% reduction in MAE coupled with an equivalent increase in MSE.

4. CONCLUSION AND DISCUSSION

We have presented several changes introduced to address shortcomings in the previous versions of the SLAM prosodic modeling software. These innovations were then evaluated by comparing the performances of the most recent version of SLAM and its predecessor in a task designed to measure similarity between their stylizations and the original contours used as input.

For both corpora used in this study, SLAM 3 performed significantly better than SLAM+ at modeling the smoothed pitch contours used as inputs. These results nevertheless contain an inconsistency that merits further exploration. In Corpus 1, excluding the glissando threshold made a negligible difference in the mean error between the input contours and those reconstructed from the stylized labels. However, doing so significantly improved the performance of SLAM 3 on Corpus 2.

These varying results can be explained by fundamental differences between the two corpora. Files in Corpus 2 are characterized by a higher speech rate and more spontaneous discourse. The corpus should therefore contain more short-duration segments labeled as flat tones when the threshold is considered, leading to less accurate labels when measured in terms of objective pitch movements. Corpus 1, with its slower speech rate, would contain fewer pitch movements labeled as flat tones. Prosodic differences between French and Nigerian Pidgin are also likely to have played a role. Given the presence of lexical tone in the latter, it is logical that Corpus 2 would contain more monosyllabic words with large pitch changes.

These results highlight the utility of the features introduced by SLAM 3. However, they also underline the need to supplement our evaluations with perceptual tests. While the glissando threshold contributed to inconsistent mean error rates, it should not be concluded that it weakens the model. Indeed, this addition is intended to better represent pitch changes as they are *perceived*. The test performed in this study measures how well the model represents any change in F0, including those that are not perceptible.

In future studies of SLAM, we will provide a more comprehensive evaluation approach that combines both objective metrics and perception tests involving multiple unbiased participants. Informal explorations of our data already reveal promising results in this regard, with labels based on the glissando threshold better representing contours as we perceive them. Given the highly customizable nature of this latest version, SLAM 3 also allows users to choose the modeling approach which best suits their needs.

5. ACKNOWLEDGEMENTS

We would like to thank Luigi Liu for his technical contributions to the development of SLAM 3.

6. REFERENCES

[1] J. T. Hart, R. Collier, and A. Cohen, A perceptual study of intonation: an experimental-phonetic approach to speech melody, Cambridge: Cambridge University Press, 1990.

[2] P. Mertens, “The prosogram: Semi-automatic transcription of prosody based on a tonal perception model.,” in *Speech Prosody*, Nara, 2004.

[3] P. Mertens, “Polytonia: a system for the automatic transcription of tonal aspects in speech corpora,” *Speech Sciences,* vol. 4, no. 2, pp. 17-57, 2014.

[4] W. Cleveland, “LOWESS: A program for smoothing scatterplots by robust locally weighted regression,” *American Statistician,* vol. 35, no. 1, p. 54, 1981.

[5] N. Obin, J. Beliao, C. Veaux and A. Lacheret, “SLAM: Automatic Stylization and Labelling of Speech Melody,” in *Speech Prosody 2014*, Dublin, 2014.

[6] L. Yu-Cheng, A. Lacheret-Dujour and N. Obin, “Automatic Modelling and Labelling of Speech Prosody: What's New with SLAM+ ?,” in *International Congress of Phonetic Sciences (ICPhS)*, Melbourne, 2019.

[7] J. Kirby and R. Ladd, “ Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages,” *The Journal of the Acoustical Society of America,* vol. 140, no. 4, pp. 2400-2411, 2016.

[8] H. Torres, M. Güemes, J. Gurlekian and D. Evin, “F0 Perturbation Due to Articulatory Movements: Filtering, Characterization and Applications,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing 29,* vol. 29, pp. 1977-1986, 2021.

[9] C. d'Alessandro, S. Rosset and J.-P. Rossi, “The pitch of short-duration fundamental frequency glissandos,” *The Journal of the Acoustical Society of America,* vol. 104, no. 4, pp. 2339-2348., 1998.

[10] R. Dall and X. Gonzalvo, “JNDSLAM: A SLAM extension for Speech Synthesis,” in *Speech Prosody*, Boston, 2016.

[11] A. Lacheret-Dujour, S. Kahane, J. Beliao, A. Dister, K. Gerdes and e. al.., “Rhapsodie: un Treebank annoté pour l’étude de l’interface syntaxe-prosodie en français parlé,” in *4e Congrès Mondial de Linguistique Française*, Berlin, 2014.

[12] B. Bigi, B. Caron and A. Oyelere, “Developing Resources for Automated Speech Processing of the African Language Naija (Nigerian Pidgin),” in *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, 2017.

NEURAL SPEECH SYNTHESIS WITH CONTROLLABLE CREAKY VOICE STYLE

Harm Lameris[1], Marcin Włodarczak[2], Joakim Gustafson[1], Éva Székely[1]

[1]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Linguistics, Stockholm University, Stockholm, Sweden
lameris@kth.se, wlodarczak@ling.su.se, jkgu@kth.se, szekely@kth.se

ABSTRACT

The use of creaky voice, or vocal fry in speech has been extensively studied for its linguistic, paralinguistic, and sociolinguistic functions. However, much of the existing research on this topic is fragmented and often contradictory. In order to gain a deeper understanding of the communicative functions of creaky voice, we propose the use of comparative perceptual studies with natural sounding speech synthesis. We present a neural speech synthesizer that produces highly natural-sounding synthetic speech with controllable creaky voice styles. In a subjective listening experiment, speech experts were able to identify the presence and intensity of creaky voice produced by the synthesizer. Our results suggest that neural speech synthesis can be a valuable tool in furthering our understanding of the communicative functions of creaky voice.

Keywords: creaky voice, vocal fry, speech synthesis, TTS, spontaneous speech

1. INTRODUCTION

Creaky voice, also known as vocal fry, has been the subject of numerous linguistic, extralinguistic, and sociolinguistic studies in recent years. However, the diverse uses of creaky voice have often led to inconclusive results. Phonetically, creaky voice can refer to several types of voice production, including a low, irregular pitch with a low spectral-to-noise ratio [1]. It has a weaker first harmonic, as well as low harmonic differences, particularly between the first and second harmonic. There has been an increase in both intra-speaker usage of creaky voice in Finnish [2] and inter-speaker use in American English and Hiberno English, particularly among young women [3, 4, 5]. In terms of linguistic function, creaky voice varies widely across languages and can serve pragmatic functions such as turn-yielding in Finnish [6], signalling a phrase boundary for English [7], as well as semantic functions like providing an acoustic cue for tone in Vietnamese [8]. Creaky voice can be used to convey a wide range of attitudes and affective states [7, 9] including boredom [10], hesitation [11], sarcasm and disgust [12], and can signal romantic intentions [13]. However, not all of these results appear to be simple one-to-one mappings, with [9] pointing out that more research is needed to elucidate the relationship between creaky voice and affective state.

The sociolinguistic aspects of creaky voice for English also paint a complex and often contradictory picture. Extensive use of creaky voice carries strong social signals [14], and creaky phonation can be distinguished equally well for male and female voices by naive listeners. Despite the absence of gender disparity in the identification of creak, there appears to be a gender difference in the perception of it [3]. While earlier research indicates that the production of creaky voice by American women was perceived as "professional" and "urban-oriented" [15], more recently the perception of creaky voice has become overwhelmingly negative. In [3] vocal fry is identified as potentially harmful to women's career prospects.

The increased frequency of the occurrence of creaky voice in English, the large number of creaky voice functions, and lack of unequivocal results with regards to the sociolinguistic aspects of creaky voice highlight big gaps in the current research. This is exacerbated by the absence of a method to reliably and systematically produce creaky synthetic speech. We therefore propose a method that controls the synthesis of creaky voice ito facilitate systematic research into the sociolinguistic aspects of creaky voice. We modified a neural speech synthesizer [16], which is itself a modification of [17], to explicitly model the (non)presence of creaky voice. A corpus that extensively contains creak was analyzed for frame-level creaky phonation using DeepFry [18]. The frame-level creakiness annotations were aligned with word-level transcripts from the Montreal Forced Aligner (MFA) [19]. Two types of creak, stylistic creak and end-of-turn creak were chosen for a phonetic analysis, as these types commonly occurred in the corpus. In the analysis,

the creaky voice from the corpus closely resembled the synthesized creak for both types. Experts were asked to examine the word-level location of creaky phonation, and rate its intensity in a subjective listening experiment. While [16] implicitly model creaky voice in speech synhtesis, this is the first paper to explicitly model creaky voice using neural speech synthesis.

2. METHOD

2.1. Data

We used two speech corpora to train our models: RyanSpeech [20] and TSGD [21]. The RyanSpeech corpus is a scripted conversational corpus that was used for base training the speech synthesis model. It consists of 10 hours divided into 11,279 utterance spoken by a male speaker of American English speaking in a conversational style. The speaker reads aloud real-world conversational settings, task-oriented dialogues, supplemented by a selection of LibriTTS excerpts. This corpus contains only minute levels of creaky phonation.

Secondly, we used the audio recordings of the TSGD corpus [21], which were denoised using [22]. The corpus features a male speaker of Hiberno English and consists of 25 unscripted monologues of spontaneous speech of approximately 10 minutes each. The monologues are delivered in a colloquial style, and according to our observations the speaker makes extensive use of creaky phonation: as a stylistic feature to convey e.g. disinterest, as a phrase-final marker, and as a speech planning tool to conserve breath at the end of a respiratory cycle. The audio was transcribed and processed identically to [23, 24] by segmenting the corpus into breath groups, i.e. speech segments occurring between two breath events. The breath groups were combined into overlapping breath group bigrams to create audio files of up to 11 seconds long [25].

2.2. Creak detection

DeepFry, a neural network-based identification method for creaky voice was used to obtain the values to quantify creakiness [18]. It was chosen due to its substantially higher recall and comparable precision compared to other methods. Utterances which could not be parsed by DeepFry were excluded from further analysis. After the creak locations and durations were obtained, we used MFA to retrieve word duration alignments, which were used to calculate the percentage of creaky phonation per word, a measure we refer to as *creak value*. The *creak values* were used as additional input for the TTS model during training.

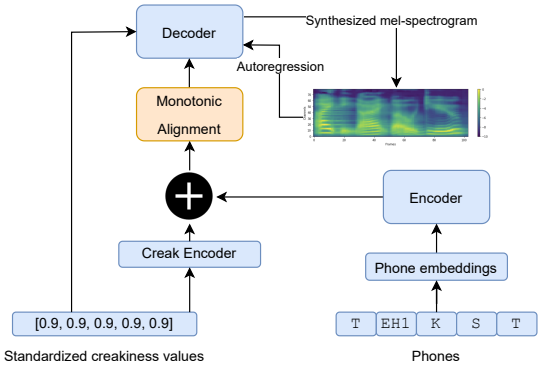


Figure 1: The model architecture

2.3. Model architecture

For the creaky voice synthesis, we modified an existing prosody-controllable sequence-to-sequence neural TTS architecture presented in [16], itself a modification of the TTS engine proposed by [17]. This architecture features an encoder-decoder model based on Tacotron 2 [26], while using a left-to-right no-skip neural Hidden Markov Model (HMM) to generate the alignments for the phone durations. The system by [16] allows for control of prosodic features such as f0 and speech rate on an utterance level. In order to achieve fine-grained control of voice quality rather than prosodic features, we modified the prosodic feature encoder from [16] to take word-level creak values as input, which we call *creak encoder*. The word-level creak values between 0 and 1 were standardized using the corpus-level mean and standard deviation, in order to create a more learnable distribution, and were copied for each phone. Copying the creakiness value for each phone instead of only the voiced phones creates a more general conditioning that allows the deep learning alogrithm to capture the general patterns of creaky phonation. In order to match the embedding size of the phone embeddings, the standardized values were passed through a feed-forward layer into a 512-dimensional control space. We used a two-step conditioning of the phone embeddings. First, we added the creak value projections to the output of the encoder. Secondly, we used a skip connection that appends the standardized creak values to each dimension of the decoder input that defines the final states of the HMM for more robust control (Figure 1).

2.4. Experimental setup

For the experiments, we trained a speech synthesizer using the speech corpora with creak annotations. As the TSGD corpus is too small for training a stable voice, we base-trained a voice on the RyanSpeech corpus for 28k iterations to increase the synthesis

stability before transfer learning on the TSGD voice for 4k iterations to learn the voice quality and creaky voice. Training on RyanSpeech does not change the voice quality after finetuning on TSGD. To analyze the synthesized creaky phonation, we performed two experiments: an acoustic analysis in which the creaky phonation present in the corpus and the creaky phonation produced by the synthesizer were compared, and a subjective listening test for which phoneticians and speech technologists were asked to mark the location and intensity of creaky phonation. To facilitate the comparison of the location of the creaky phonation for each condition, we divided each sentence in quarters based on the word count. This division was not visible to the participants. We compared the ratings for the general intensity of the creaky phonation by comparing the distribution of the intensity of the creak for the full utterances per condition. The location of the creaky phonation was compared by analyzing the creak ratings for the words in each quarter.

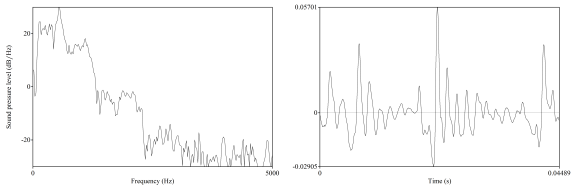
3. EXPERIMENTS

3.1. Acoustic analysis of synthesized creak

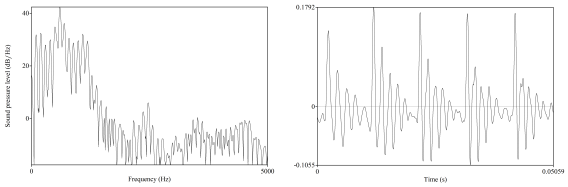
In order to compare the creaky phonation present in the corpus with the synthesized creaky phonation, two realizations of the /ɒ/ vowel preceding /t/ from the corpus with their synthesized equivalents were examined, one with end-of-phrase creak as voice quality and one with stylistic creak. The /ɒ/ vowel preceding /t/ was chosen, as it was a common combination in creaky segments. We selected a representative sample for each of the creak types and synthesized a matching sample for each creak type.

Figure 2 shows spectra and waveforms for each of the natural and synthesized vowels. Figure 2a and Figure 2c show the natural and synthesized end-of-phrase creak. The synthesized creak shows many similarities with the natural creak in the spectrum, both having weak harmonic structure and little energy for the higher frequencies. The waveform for the natural and synthesized end-of-phrase creak show high aperiodicity, which caused pitch tracking to fail for both the natural and synthesized speech.

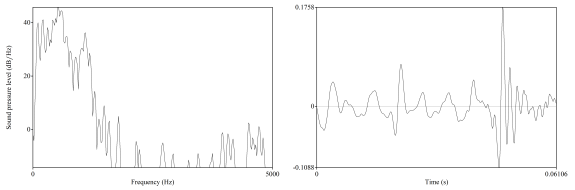
For Figure 2b and Figure 2d, showing stylistic creak on the other hand, the spectra show a clear harmonic structure with weak fundamental (i.e. low L1-L2). Additionally, it has more energy in the higher frequencies compared to the aperiodic creak. In terms of the waveform, both the natural and synthesized creak are periodic, albeit in a more regular manner for the natural creak. Pitch tracking retrieved the correct values for both examples.



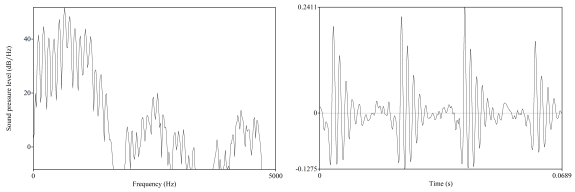
(a) The spectrum and aperiodic waveform of the natural /ɒ/ with end-of-phrase creak from [21].



(b) The spectrum and periodic waveform of the natural /ɒ/ with stylistic creak from [21].



(c) The spectrum and aperiodic waveform of synthesized /ɒ/ with end-of-phrase creaky phonation.



(d) The spectrum and periodic waveform of synthesized /ɒ/ with stylistic creak.

Figure 2: Spectra and waveforms for the natural and synthesized creaky phonation.

3.2. Subjective listening test

To examine the perception of synthesized creaky voice, 23 participants with expertise in phonetics or speech technology were presented iteratively with 36 stimuli. These stimuli comprised 12 sentences that were synthesized in three different voice qualities: no creak, i.e. modal phonation, stylistic creak with creaky phonation present throughout the utterance, and end-of-turn creak with creaky phonation present at the end of the utterance. The participants were provided with word-level transcripts for each stimulus and were asked to rate the presence of creaky phonation using a scale of 0—no creak, 1—creak, or 2—intense creak. Additionally, the participants rated the naturalness of the creaky phonation and could provide comments at the end of the experiment.

Figure 3 shows an overview of the ratings given by experts per creak condition per utterance quarter.

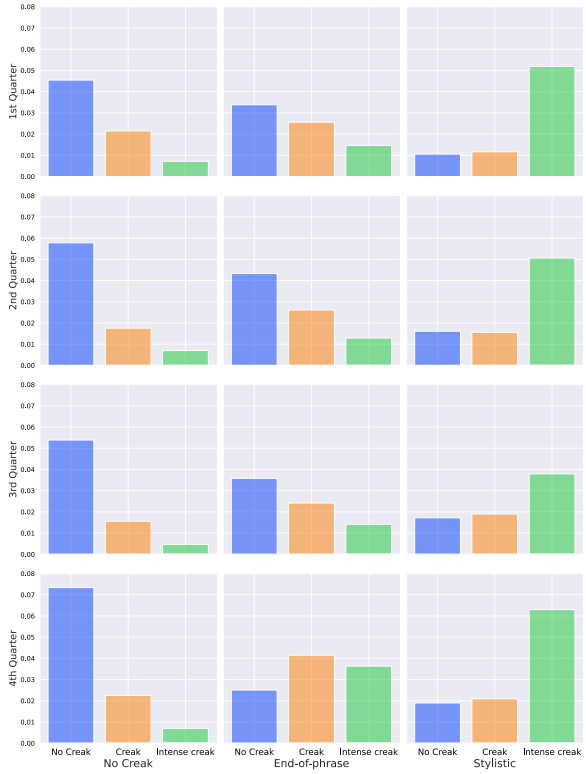


Figure 3: The distribution of the experts' ratings per condition and per quarter as percentage of the complete set of responses.

A Kruskal-Wallis H test showed significant results between the conditions for both the full utterance and for each quarter (all $p < 0.0001$). A post-hoc Dunn test with Bonferroni adjustment showed significant differences between every combination of conditions for both the full utterance and for each quarter ($p < 0.0001$). *Stylistic creak* was universally rated the strongest for each category with a median of 2 for the full utterance and each quarter ($p < 0.0001$). This was followed by *end-of-phrase creak* which had a median of 1 for the full utterance, the first, third, and fourth quarter, and a median of 0 for the second quarter, but had higher creak ratings than the no creak condition for each category ($p < 0.0001$). *No creak* had the lowest creak rating, with a median of 0 for full utterance and each quarter.

4. DISCUSSION

The acoustic analysis showed the synthesis of two types of creak: an end-of-phrase creak and a stylistic creak. The end-of-phrase creak was highly aperiodic with slow and irregular vibration. This type of creak is similar to non-constricted creak, as described by Keating [1], which also exhibits low and irregular f0, as well as slow irregular vibrations. It also showed many similarities to creaky phonation present in the corpus. The stylistic creak is more periodic, albeit with a very low f0. This creak is similar to vocal fry, as described in [1].

The ratings suggest that participants were generally able to distinguish *no creak*, *end-of-phrase creak*, and *stylistic creak*. Participants rated utterances with no creak as not creaky. Stylistic creak was rated as more intense and more present throughout the utterance than end-of-phrase creak.

Twenty-two of the experts rated the creaky phonation as natural or very natural for most utterances. Six participants noted occasional difficulty to distinguish creaky from modal phonation for stretches of speech with a high speech rate, partly due to vocoder artefacts that persist due to the nature of the data. Three participants highlighted the role of prosody in the rating of the utterances; participants had more difficulty distinguishing the phonation type for higher-pitched utterances. Two participants mentioned the influence of sociolinguistic and paralinguistic aspects in their perception of creaky voice.

5. CONCLUSION

In this paper, we present a novel approach to investigating the perception and communicative functions of creaky voice by utilizing neural speech synthesis. Our method opens up the possibility to systematically test the perception of creaky voice, and provides the opportunity to clarify, disambiguate, and build upon existing knowledge. We created a speech corpus with automatic annotation of creaky voice and modified a state-of-the-art speech synthesis architecture to explicitly model creaky voice. The synthesizer was trained on the speech corpus containing creak annotations. A phonetic analysis demonstrated that the two types of creaky phonation generated by the synthesizer closely resembled their natural counterparts found in the corpus. In a subjective listening test, experts successfully distinguished between the two creaky phonations and rated their creakiness as natural or very natural. This work reinforces the potential of neural speech synthesis as a valuable tool for advancing our understanding of the communicative functions of creaky voice and its sociolinguistic implications.

ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council projects Connected (VR-2019-05003), STANCE (VR-2020-02396), *Prosodic functions of voice quality dynamics* (VR-2019-02932), and the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

6. REFERENCES

[1] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice." in *Proc. ICPhS*, 2015, pp. 2–7.

[2] M. O'Dell, T. Nieminen, and L. Mustanoja, "Creak rate variation in individual speakers of Finnish," in *Proceedings of Nordic Prosody 2022*, in press.

[3] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatachalam, "Vocal fry may undermine the success of young women in the labor market," *PLoS one*, vol. 9, no. 5, pp. 1–8, 2014.

[4] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111–e116, 2012.

[5] R. Hickey, "Irish English in the anglophone world," *World Englishes*, vol. 36, no. 2, pp. 161–175, 2017.

[6] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[7] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech Language*, vol. 28, no. 5, pp. 1233–1253, 2014.

[8] J. Edmondson and N. V. Loi, "Tones and voice quality in modern northern Vietnamese: instrumental case studies." *Mon-Khmer Studies*, vol. 28, no. 35, pp. 1–18, 1997.

[9] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1, pp. 189–212, 2003.

[10] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[11] M. Horne, "Creaky fillers and speaker attitude: data from Swedish," *Studies in Pragmatics*, pp. 277–288, 2009.

[12] K. Fischer and O. Niebuhr, "The effects of the online visualization of acoustic-prosodic features of speech on speakers' productions," in *Studientexte zur Sprachkommunikation: elektronische Sprachsignalverarbeitung*. TUD Press, 2022, pp. 180–187.

[13] J. Jelle, "A study in Scarlett: Creaky voice and romantic intention in Spike Jonze's Her," *Leviathan: Interdisciplinary Journal in English*, no. 1, pp. 35–44, 2017.

[14] L. Davidson, "The effects of pitch, gender, and prosodic context on the identification of creaky voice," *Phonetica*, vol. 76, no. 4, pp. 235–262, 2019.

[15] I. P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[16] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous TTS with neural HMMs," *Proc. ICASSP*, 2023.

[17] S. Mehta, É. Székely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *Proc. ICASSP*. IEEE, 2022, pp. 7457–7461.

[18] B. R. Chernyak, T. Ben Simon, Y. Segal, J. Steffman, E. Chodroff, J. Cole, and J. Keshet, "DeepFry: Identifying Vocal Fry Using Deep Neural Networks," in *Proc. Interspeech 2022*, 2022, pp. 3578–3582.

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[20] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," in *Proc. Interspeech*, 2021, pp. 2751–2755.

[21] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. IVA*, 2018, pp. 93–98.

[22] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, pp. 3291–3295.

[23] J. Gustafson, J. Beskow, and É. Székely, "Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," in *Proc. SSW*, pp. 48–53.

[24] É. Székely, G. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. SSW*, 2019, pp. 245–250.

[25] É. Székely, G. Henter J., Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *Proc. ICASSP*, 2020, pp. 7649–7653.

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

RECOVERING IMPLICIT PITCH CONTOURS FROM FORMANTS IN WHISPERED SPEECH

Pablo Pérez Zarazaga, Zofia Malisz*

TMH, KTH Royal Institute of Technology, Stockholm
(pablopz,malisz)@kth.se

ABSTRACT

Whispered speech is characterised by a noise-like excitation that results in the lack of fundamental frequency. Considering that prosodic phenomena such as intonation are perceived through f_0 variation, the perception of whispered prosody is relatively difficult. At the same time, studies have shown that speakers do attempt to produce intonation when whispering and that prosodic variability is being transmitted, suggesting that intonation "survives" in whispered formant structure.

In this paper, we aim to estimate the way in which formant contours correlate with an "implicit" pitch contour in whisper, using a machine learning model. We propose a two-step method: using a parallel corpus, we first transform the whispered formants into their phonated equivalents using a denoising autoencoder. We then analyse the formant contours to predict phonated pitch contour variation. We observe that our method is effective in establishing a relationship between whispered and phonated formants and in uncovering implicit pitch contours in whisper.

Keywords: Whispered speech, Formant contours, Pitch contours, Intonation, Machine learning

1. INTRODUCTION

Pitch is an important feature in the encoding of prosody, and its variations over time are a defining characteristic of intonation. In whispered speech, fundamental frequency is not present in the signal as the glottis remains open and the vocal folds do not vibrate. In principle, this should prevent transmission of pitch and intonation via whisper.

However, several studies [1, 2, 3] have found that listeners perceive pitch and intonation effects in whispered speech. Similarly, it is evident that the whispered signal contains properties that offer cues to intonation.

Formant raising, particularly of F1, is a well-attested characteristic of whispered speech relative to phonated speech [4, 5, 6]. The higher formant frequency positions in whisper may be related to a more open configuration of the vocal tract similar to that found in e.g. Lombard speech [6, 4]. In our previous work, we showed that the jaw is more open in whispered, relative to phonated, vowels in Swedish, supporting the hypothesis that whisper is a form of hyperspeech. However, others have suggested that formant raising is related to whispered pitch perception: [3] showed that listeners were able to discriminate between low and high-pitched whispered vowels on the basis of higher F1 and F2. Additionally, raised formant frequencies seem to connect to prosody-related laryngeal activity [5]. [7] showed that in whispered speech production, laryngeal movements associated with pitch changes remain comparable to those in phonated speech. This indicates that prosody-related oral cavity shape modification takes place in whisper and hence is able to change the acoustics of the noise-like excitation travelling through the vocal tract - allowing for pitch inference in perception.

The harmonic relation between f_0 and formants has also been exploited in speech technology. In voice conversion systems, i.e. systems that reconstruct phonated speech from whispers, some solutions have tackled the problem of the missing f_0 by creating harmonic excitation for whispered speech using generative adversarial networks [8, 9]. [10] showed promising results by adding an artificial pitch model to formant structure.

In the present paper, our process runs somewhat in the opposite direction: we first aim to learn from a) the relationship between f_0 and formants in phonated speech and b) its connection to the spectral properties of whispered speech. To this end, we present a machine learning method that models the variation in f_0 present in phonated speech through the variation in whispered formant values. One goal is to use these relationships to uncover and understand the "implicit" pitch contour in whisper that allows for the perception of prosody in the absence of f_0. The other is to support technological

*This work was supported by the Swedish Research Council grant no. 2017-02861 “Multimodal encoding of prosodic prominence in voiced and whispered speech”.

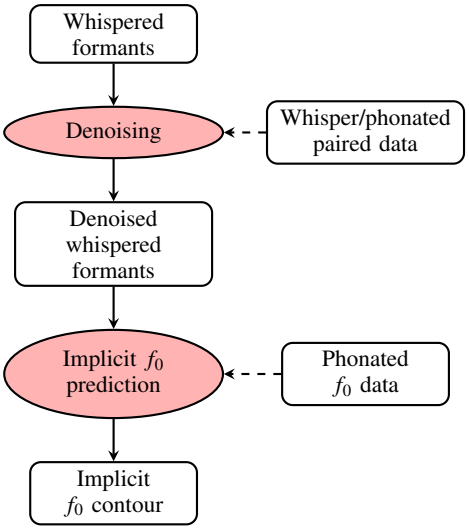


Figure 1: Steps of the proposed method. Red ellipses represent the implemented methods and white boxes represent the available data in each step. Dashed lines stand for the training process.

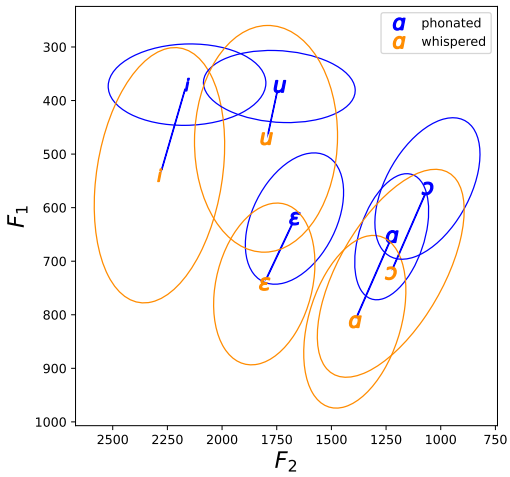


Figure 2: Vowel space representing F_1 vs F_2 for the vowel sounds [a, ɛ, i, ɔ, u] in phonated (blue) and whispered (red) speech and ellipses for their standard deviation.

applications such as voice conversion.

2. METHOD

2.1. Data

In this work, we use the CHAINS dataset [11], which contains paired recordings in both phonated and whispered speech. CHAINS contains data from 20 male and 16 female speakers with Irish (28) and American (8) accents. The dataset consists of read sentences and text fragments in different speech modes such as natural phonated and whisper. We use a subset of CHAINS with paired phonated and whispered sentences and analyse only vocalic phonemes. We transcribed the phonated audio samples using a Wav2Vec2 model [12] and copied the transcriptions into the whispered equivalents. The data was then labelled using the Montreal Forced Aligner [13] with a pre-trained model for American English. We observed that function word reductions produced event sequence mismatches between phonated and whispered pairs. They were subsequently removed from the analysis yielding an exactly matched dataset of 7549 phonated and whispered vowel phonemes.

2.2. Feature extraction and preprocessing

We extracted f_0 and formant contours from vowel phones. F_0 values were extracted using Praat, with a time step of 0.01 s and using a pitch floor and ceiling of 75 and 300 Hz respectively. To extract the formant contours, we used the Burg method in Praat with a time step of 0.01 s, a 0.025 s long analysis window and the maximum number of formants set to 5. Pre-emphasis was applied over 50 Hz. The formant ceilings were initialised to 5000 Hz for male speakers and 5500 Hz for female ones and optimised using the Escudero's method [14]. In order to normalise the length of each segment, we defined evenly distributed observation points every 10% of the total duration of the phone. The contour measurements corresponding to 0% and 100% were then discarded to avoid instabilities due to formant transitions between phones, resulting in 9 measurement points that define the pitch and formant contours. To reduce speaker variability in the formant contours, we normalised the data for each speaker following the Lobanov's normalisation method [15], where the data presents zero mean and unit variance with respect to its corresponding speaker. Additionally, pitch contours are also normalised to zero mean and unit variance relative to each speaker.

2.3. Whispered pitch prediction

We propose a machine learning method to estimate the "implicit" f_0 contour in whisper from whispered formant structure. The two-step method is depicted in Figure 1. In the first step, we leverage the relationship between whispered and phonated formants residing in paired phones to provide a de-noised representation of the whispered formant frequencies. In the second step, we predict implicit, whispered f_0 contours by modelling the corresponding phonated f_0 with the denoised formant representations.

2.4. Formant denoising

We assume that whispered formant structure is strongly correlated to its phonated equivalent. With the difference that in whisper, the formant frequencies are raised and the spectral envelope peaks are flatter manifesting a noisier behaviour. The vowel spaces obtained from the current dataset are presented in Figure 2. A high variance in some F_1 values can be observed due to the scarce representation of some phones in the dataset. The effects of whispered speech are reflected as an increase in the values of F_1 and F_2 and a higher variance in whispered, relative to phonated, formant values.

Hence, we consider whispered formant contours as a noisy representation of phonated formants and propose a denoising strategy to "transform" whispered formant contours to phonated equivalents. For this purpose, we use an autoencoder [16] that learns higher-order statistical information from the formant contours. Using whispered-phonated formant pairs as input and output of the model, we aim to map the contours from whispered speech to the formant contours of the same phones in phonated speech.

The machine leaning method consists of an symmetric encoder-decoder structure. Three 1-D convolutional layers are used as filters over the time-dependent features (16 channels, kernel length of 3). These layers, plus a fully-connected layer, generate encoder embeddings (18-sample) from whispered formant contours. The decoder uses three transposed convolutional layers to invert these embeddings into phonated speech formants.

We have observed that most of the contours in the data do not present significant variations. This might introduce bias in our network towards a constant value in the formant and pitch contours. In order to maximise the similarity with the target contour, we choose a different loss function: the cosine distance between target and predicted contours, such that:

$$\mathscr{L}_{cos} = -\frac{1}{N}\sum_{i=0}^{N}\frac{\hat{y}_i \cdot y_i}{||\hat{y}_i||_2||y_i||_2} \quad (1)$$

where y_i and $\hat{y}_i$ stand for the target and denoised formant sequences of the i_{th} sample and $\cdot$ represents the dot product between the two vectors.

The correlation coefficients for F_1, F_2 and F_3 are presented in Table 1. We can observe that, while the correlation between phonated and denoised formant contours is improved for F_1, F_2, the network degrades the correlation for F_3 values. Therefore, we chose to apply denoised F_1 and F_2, while the contour of F_3 is used unmodified in the next step.

2.5. Implicit f_0 prediction

The second step in the method is a sequence prediction model that estimates the "implicit" f_0 contour based on the denoised contours of F_1, F_2 and (unmodified) F_3. This works under the assumption that the f_0 contour in the phonated data is related to the one "implied" in the whispered data. Admittedly, modelling the exact values of a non-existing pitch signal from formant contours is a challenging task. Therefore, this model focuses on estimating the relative changes in the pitch contour that represent intonation variability.

We use a recurrent neural network (RNN), a deep learning method that is especially efficient in uncovering temporal dependencies within sequences. Considering the temporal dependencies in the formant and f_0 contours, a sequence-to-sequence model [17] should be particularly useful in modelling dependencies within and between dynamic contours. The input formant contours are processed by two bi-directional LSTM layers [18] with 4 hidden units, resulting in a sequence with 8-dimensional features. The output sequence is then generated with a fully-connected layer that maps the corresponding 8 features to the target pitch values.

Similarly to the formant denoising, we require a function that will maximise the similarity to the target sequence. The cosine distance can be seen as a normalised correlation, thus providing a better approximation to the sequence shape than mean squared error (MSE). Additionally, in order to also approximate the actual pitch values of the target, we will also include MSE in the loss function. The resulting loss function is then a combination of MSE and cosine distance:

$$\mathscr{L}_{seq} = \frac{1}{N}\sum_{i=0}^{N}|\hat{y}_i - y_i|^2 - \frac{1}{N}\sum_{i=0}^{N}\frac{\hat{y}_i \cdot y_i}{||\hat{y}_i||_2||y_i||_2}. \quad (2)$$

Both the denoising and prediction models were trained over 300 epochs with an Adam optimiser and a learning rate of 10^{-4}. Additionally, a recurrent dropout of 0.4 was added in the training loop to the LSTM layers of the f_0 prediction network.

3. RESULTS

First of all, our results show that we were able to successfully leverage the close relationship between whispered and phonated speech to denoise whispered formants. Table 1 shows that denoised whispered formant contours exhibit a higher correlation with their phonated counterparts. The greatest improvement can be seen in F_1, known to exhibit the greatest difference between phonated and

Table 1: Pearson's correlation coefficient of whispered and denoised formant contours with respect to phonated ones.

Formant	Whispered	Denoised
F_1	0.19	0.34
F_2	0.44	0.47
F_3	0.26	0.16

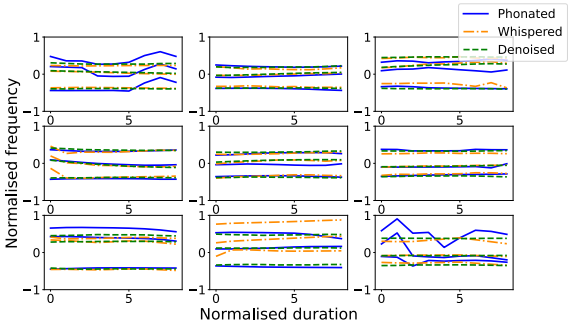


Figure 3: Denoised formants from whispered contours with respect to original values.

whispered speech [4, 5, 6]. In Figure 3, we present several examples of the denoised formant contours in which the denoising model acts as a smoothing function. It is evident that the cosine similarity loss function allows us to closely estimate formant frequency contours in stationary vowel intervals found in the CHAINS corpus. Where the network fails is in some extreme cases with high contour variability, which may be due to various factors such as formant tracking errors, residual influence of the flanking phones etc.

Second, the measured error and correlation values in the pitch prediction step are summarised in Table 2. The refined pitch contour (RPC) from [10] is used as baseline. Our model obtains a positive correlation (Mean $r = 0.56$, SD $= 0.3$) between predicted and target contours that allows to generally follow the target pitch. Some examples of the resulting mapping between predicted and phonated contours can be observed in Figure 4. We see a correct direction of the trend in most cases. We also compare the absolute difference between the mean f_0 frequencies. Regarding comparisons to the RPC baseline, while the LSTM shows improvement in several individual cases, the average performance is similar to the baseline in correlation and error values. This leads us to believe that a more complex system such as LSTM would benefit from using additional input features supplying information beyond the minimalistic formant set.

The absolute error from the predicted f_0 presents an average value of 38 Hz, with a standard deviation of 24 Hz. The number of different speakers

Table 2: Absolute error and correlation of implicit f_0 prediction from the refined pitch contour (RPC) estimation [10] and the proposed LSTM-based network with respect to phonated f_0.

Method	RPC [10]		LSTM	
	Mean	St. Dev.	Mean	St. Dev.
Error (Hz)	40	27	38	24
Correlation	0.55	0.3	0.56	0.3

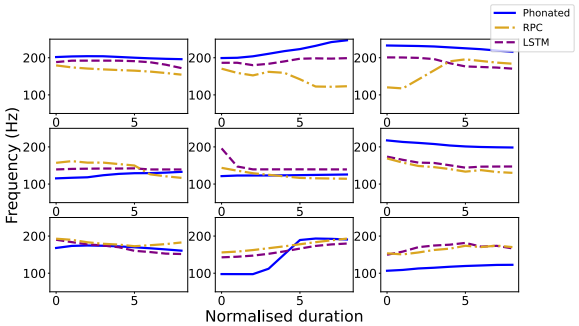


Figure 4: Predicted implicit f_0 contour values compared to the original contours.

(36) in this data was high and this error can be explained by the impact of speaker variability. This estimation could be improved by considering speaker-dependent features or tuning the model specifically for each speaker.

4. CONCLUSION

In this work, we have presented a machine learning method that estimates implicit f_0 contours from formant contours in whispered vowels. Our models have shown that it is possible to uncover an implicit f_0 trajectory in whisper, via its phonated equivalent, from F_1, F_2 and F_3 contours only.

In future lines of work, we will apply a speaker-dependent model. Our results show that the variability inherent in the 36 speakers modelled in this data has an influence on the f_0 prediction. Additionally, we would like to incorporate information on the analysed phones and their context into the formant analysis and pitch estimation. The present models showed promising results even before this variability was taken into account and we expect that adding the additional features will decidedly improve performance.

This result will allow us to observe variations in perceived pitch and provide an improved analysis of intonation in whispered speech. The implementation of modern methods to uncover implicit pitch in whisper will also lead to developments in voice conversion systems.

5. REFERENCES

[1] W. Heeren and V. J. van Heuven, "Perception and production of boundary tones in whispered dutch," *Proceedings of Interspeech, September 6-10, 2009*, pp. 2411–2414, 2009.

[2] J. N. Holmes and A. P. Stephens, "Acoustic correlates of intonation in whispered speech," *The Journal of the Acoustical Society of America*, vol. 73, no. S1, pp. S87–S87, 1983.

[3] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study," *Journal of Voice*, vol. 10, no. 2, pp. 155–158, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199796800427

[4] M. Żygis, D. Pape, L. L. Koenig, M. Jaskuła, and L. M. Jesus, "Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes," *Journal of Phonetics*, vol. 63, pp. 53–74, 2017.

[5] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a nam-captured whisper-to-speech system," *Speech communication*, vol. 52, no. 4, pp. 314–326, 2010.

[6] I. Eklund and H. Traunmüller, "Comparative study of male and female whispered and phonated versions of the long vowels of swedish," *Phonetica*, vol. 54, no. 1, pp. 1–21, 1997.

[7] J. Coleman, E. Grabe, and B. Braun, "Larynx movements and intonation in whispered speech," *Summary of research supported by British Academy*, 2002.

[8] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel mmse discogan for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop*. Google Office, 2018, pp. 1–3.

[9] H. Malaviya, J. Shah, M. Patel, J. Munshi, and H. A. Patil, "Mspec-net: Multi-domain speech conversion network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7764–7768.

[10] I. V. Mcloughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Trans. Access. Comput.*, vol. 6, no. 4, may 2015. [Online]. Available: https://doi.org/10.1145/2737724

[11] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *Proc of SPECOM*, vol. 6, no. 2006, 2006, pp. 431–435.

[12] J. Grosman, "Fine-tuned wav2vec2 large model for speech recognition in English," https://huggingface.co/jonatasgrosman/wav2vec2-large-english, 2021.

[13] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[14] P. Escudero, P. Boersma, A. S. Rauber, and R. A. Bion, "A cross-dialect acoustic description of vowels: Brazilian and european portuguese," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1379–1393, 2009.

[15] B. M. Lobanov, "Classification of russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, 1971.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[18] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673 – 2681, 12 1997.

NASALITY DETECTION FROM ACOUSTIC DATA WITH A CONVOLUTIONAL NEURAL NETWORK AND COMPARISON WITH AERODYNAMIC DATA

Amélie Elmerich[1], Lila Kim[1], Cédric Gendrot[1], Angélique Amelot[1], Lise Crevier-Buchman[1,2], Shinji Maeda[1]

[1] Laboratoire de Phonétique et Phonologie (UMR7018, CNRS - Sorbonne Nouvelle)
[2] Hôpital Foch: Service de Laryngologie Phoniatrie, Université Paris Saclay, Suresnes, France
{amelie.elmerich, lila.kim, cedric.gendrot, angelique.amelot}@sorbonne-nouvelle.fr

ABSTRACT

From a new acoustically transparent pneumotachograph mask, we simultaneously recorded aerodynamic (oral and nasal airflow) and acoustic data for 6 French male speakers, involving 3 oral and 3 nasal vowels out of logatoms (*i.e.* non words). A Convolutional Neural Network (CNN) trained on other acoustic corpora in French was tested on the data collected from the mask for the nasal/oral vowel distinction, with a 88% correct classification on average. We compared these CNN results with the nasal airflow extracted from all vowels of the corpora. Aerodynamic results showed a higher quantity of nasal airflow for the nasal vowels. However, for some speakers, distinction between nasal and oral vowels in terms of nasal airflow was less prominent, especially for /a/ *vs* /ɑ̃/, the 2 vowels for which the CNN have the least correct identifications. Finally, we discuss the discrepancies observed between aerodynamic data and CNN probabilities, and inter-speaker variations that can be approached by the CNN.

Keywords: Convolutional Neural Network (CNN), aerodynamic, nasality, speaker.

1. INTRODUCTION

Nasality is a distinctive feature in approximately a third of the world's languages [1]. The basic knowledge implies that the soft palate must be sufficiently lowered for the velopharyngeal port to be open allowing air to pass through the nose. The lowering of the soft palate as well as the passage of air through the nose will have an implication in the speech spectrum with acoustic zeroes that usually hinder the acoustic analysis for phoneticians [2].

There are temporal and spatial variations in the realisation of the [nasal] feature. It varies depending on the speaker's gender and anatomy [3 ; 4], speaker strategy [5 ; 6 ; 7], language [8], speaking style [9], speaking rate [10], the speech sound type, the phonetic and prosodic context [11], *etc*.

More specifically, the opening of the velopharyngeal port differs from one speaker to another [3, 4] and the morphology of the nasal cavities is highly variable between individuals. Nasal vowels and consonants are relevant for speaker identification as they contain more acoustic information relative to the speakers than the other sounds [12, 13].

Deep neural networks have recently shown an important development in the field of speech. Studies have been conducted in the clinical field with artificial neural networks to diagnose language pathologies, including hyper- or hypo-nasalisation [14, 15, 16]. Indeed it has been shown that CNNs have the ability to specialise on phonetic features such as place of articulation or articulatory mode [16, 18, 19].

From a new pneumotachograph developed at the Phonetics and Phonology Laboratory, composed of a paper-fiber mask which provides no acoustic distortion, it is possible to analyse acoustic and aerodynamic data in the same recording [20], which is not possible with most aerodynamic measurement systems. With this device, it is therefore possible to account for the aerodynamics in parallel with the resulting acoustics.

The purpose of the present study is to assess the detection of nasality from acoustic data with a CNN by comparing it with aerodynamic data collected during the acoustic recording. We take the vowel phonemic label 'nasal' or 'oral' as the reference and check whether CNN classification is correct from the acoustics. In a second time the nasal airflow provided will validate the CNN classification or help us understand the misclassifications. We will finally investigate to what extent the level of nasality for each speaker can be estimated with CNN.

2. MATERIALS AND METHODS

2.1. Corpora and data acquisition

The test data for this study was taken from 6 French male native speakers (mean age: 36 years) recorded

in a soundproof room. The speech samples consist of VCV sequences, where C=[p,b,t,d,v,s,z,m,n] and V=[i,a,y,u,o,e,ɑ̃,ɛ̃,ɔ̃]. The sequences were inserted in the frame sentence, for example: « Non, tu n'as pas dit apa quatre fois, tu as dit aba et ada quatre fois » ('No, you didn't say apa four times, you said aba et ada four times'). Finally, we have a total of 270 sequences with C= 270 et V= 540.

Aerodynamic and acoustic data were recorded simultaneously with a pneumotachograph mask. The advantages of this mask are that i) oral and nasal airflow can be recorded separately, ii) it is possible to adapt the size and position of the plate (Fig. 1a) to separate the nasal airflow (NAF) from the oral airflow (OAF) for each speaker iii) there is no acoustic distortion [20]. This mask is connected to 2 sensors and each one measures the differential pressure inside the mask with respect to the atmospheric pressure (see Fig. 1). There may be slight differences in flow measurements depending on the mask, the position of the sensor and the size and the position of the plate separating the nasal and oral airflow. Consequently, a calibration must be operated separately for each mask (individual mask for each speaker): one calibration for the oral compartment and another one for the nasal compartment. The calibration of the 2 pressure sensor modules allows to convert airflow values in the physical unit (liters/s, see Fig. 2). The mask provides a small resistance, as required to measure the airflow without affecting the sound propagation [21].

a b

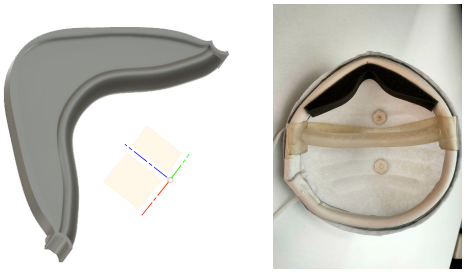


Figure 1: a. Flexible resin separation integrated into the mask to separate the nasal and oral airflow, b. Mask in fiber paper with plate and 2 adapters connected to the pressure sensors.

Acoustic data were captured with a microphone (AKG C520 L). All aerodynamic and acoustic sensors are linked to an acquisition card (DT9003). Audio and aerodynamic data were recorded at a sampling frequency of 20kHz.

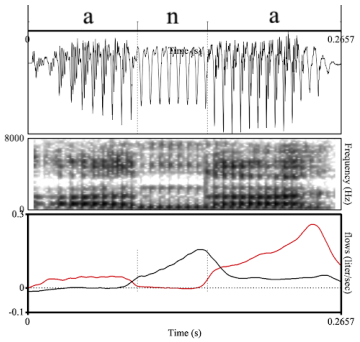


Figure 2: Example of acoustic and airflow recordings of [ana]. From top to bottom, (1) audio signal captured with a microphone, (2) spectrogram, (3) the nasal airflow (NAF in black) and oral airflow (OAF in red).

The data were manually segmented in Praat [22]. A Python script was used to automatically extract the mean of OAF and NAF for each vowel (l/sec). We used R [23] to perform graphics and statistical analysis.

2.2. Convolutional Neural Networks

For an acoustic nasal-non-nasal automatic classification task, we chose to work with Convolutional Neural Networks (CNN) and to focus only on 6 vowels /a,e,o,ɑ̃,ɛ̃,ɔ̃/, hence 3 vowel qualities /a,e,o/ and their nasal counterparts. We are aware that there is no exact articulatory match between our 3 oral and nasal vowels [24] and we decided to include all 6 vowels in the classification system (instead of comparing by pairs) so as to circumvent this asymmetry.

We opted for a CNN over other types of neural networks due to our intention of working with spectrograms of vowels, with the ultimate goal of applying a gradCam-like algorithm. This method is used to localise the specific parts of an image that have contributed to the final decision, in our case, the areas that contain information related to nasality. The training dataset is composed of the productions of these vowels extracted from 3 French corpora with various speech types: NCCFr [25], ESTER [26] and PTSVOX [27]. In all these corpora, automatic segmentations were provided at the phoneme level. Randomly chosen vowels were extracted at their boundaries (as determined by phonemic segmentation) in the form of a spectrogram, without any selection of the prosodic, lexical or phonemic context. For the first 2 corpora, the number of vowels of each type was checked. 10,887 productions of each type were taken from NCCFr and 9,186 from ESTER. For PTSVOX, we took all possible vowels respecting the natural frequency of phonemes as this brought better results.

The test dataset was derived from the acoustic data that was described in section 2.1. During this phase, we randomly selected vowel utterances and extracted their corresponding spectrograms. This set contains 66 productions of each vowel type (6 speakers * 11 occurrences), making 198 vowels for each category.

	Train & Val			Test
Source	NCCFR	ESTER	PTSVOX	Data recorded with the mask
nasal	32,661	27,558	65,669	198
non-nasal	32,661	27,558	135,119	198

Table 1: Number of vowels used for train and test sets according to the corpora.

All these spectrogram images with a frequency band from 0 to 8000 Hz were reduced in 48x48 pixel sizes and presented as input to our network. For the feature extraction part, the model is made of 2 pairs of convolution and pooling layers. The convolution layers were performed with a 5x5 kernel size and thus produced 32 and 64 filters respectively. After each convolution layer, a batch normalisation layer was inserted before applying an activation layer in order to allow the model to generalise [17] over different types of corpora and data. The max pooling layers were then used to reduce the size of the images with a 2x2 pool size. With the extracted features, 3 fully-connected layers performed the classification task with 1024 neurons. The ReLU activation function was applied after each batch normalisation layer and each fully-connected layer. Finally, a softmax activation function was used at the last fully-connected layer for nasal-oral classification. The labelling of "nasal" and "non-nasal" was based on the expected phonemic transcription in French. We performed a binary categorization to assign a probability to each class. The probability value for each class was also considered as a continuous variable for strength of nasality. During the model training, the model was improved by applying Adam as an optimisation technique and categorical cross entropy as a metric to measure model performance.

3. RESULTS

3.1 CNN results

For a given vowel, the classifier returns a value between 0 and 1, which we refer to as the probability of nasality obtained by the CNN model. When a vowel is identified as nasal by the classifier, the expected probability of nasality would be over 0.5 and conversely, a vowel classified as non-nasal would have a value close to 0 (or at least below 0.5).

Our image classifier could accurately identify 95% of non-nasal vowels and 82% of nasal vowels, reaching 88% overall accuracy and F1-score of 88% (k = 0.77) without fine-tuning between training and test.

The cross-speaker variability can be observed in Fig. 3. There are speakers for which the model makes many classification errors and others for which there are fewer errors from the model. For example, the model makes most misclassifications for speakers MT01 and MT04. Of the total misclassifications, 37% of the inaccurately classified vowels come from speaker MT04 (*i.e.* 17 out of 46 errors). Speaker MT01 gathers 13 incorrect occurrences (*i.e.* 28% of the total misclassifications) while speaker MT03 has only 1 error. In addition, errors on non-nasal vowels occur only for speakers MT01 and MT05. For the other speakers, the model accurately performed on non-nasal vowels, with errors occurring only for nasals.

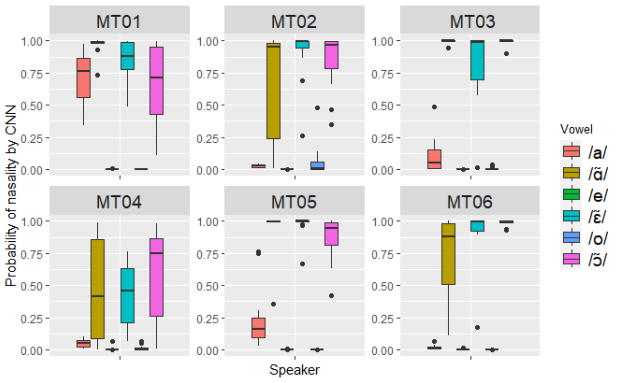


Figure 3: Nasality probabilities obtained by the CNN model for each speaker and each vowel type.

	Left	# Err	Right	# Err	Total
nasal	labial coronal	12 12	pause	9	36
non nasal	pause	5			10

Table 2: Contexts mostly found in misclassifications according to the nasal-oral categories with their number of occurrences in the errors.
('left' and 'right' for left and right context of vowels).

We observe that misclassifications mostly appear between /a/ and /ɑ̃/. Several phonetic contexts could be considered as factors in this confusion. On the one hand, as we can see in Table 2, when there is a pause in the left context, /a/ vowels tend to be classified as nasal by our model (5 out of 10 errors of /a/, i.e. 50%). On the other hand, the presence of a labial or coronal consonant before the vowels /ɑ̃/ can influence the decision of its class (respectively 5 and 6 out of 14 errors of /ɑ̃/). We notice the same influence when a pause is located after these vowels

(5 out of 14 misclassifications of /ɑ̃/). These 3 contexts for /ɑ̃/ vowels also appear in errors for other nasal vowels. Out of 36 incorrect nasal vowel classifications, 12 errors are caused by the labial and coronal consonant context preceding nasal vowels, and 9 errors occur with a pause in the right context.

3.2 Aerodynamic results

In Figure 4, we observe a higher quantity of nasal airflow for the 3 nasal vowels. Results of ANOVA were statistically significant with p<.001 for all speakers for the distinction of NAF between nasal and oral vowels.

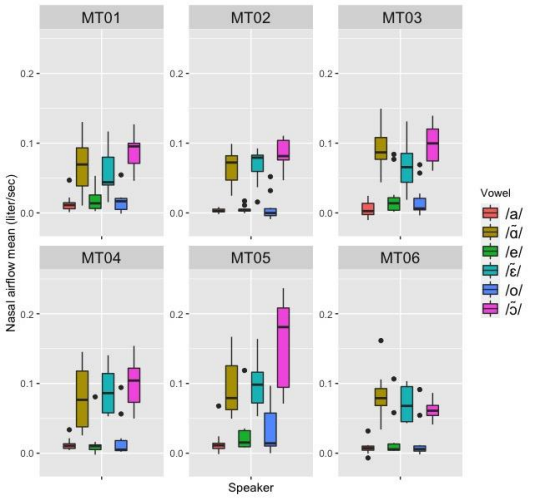


Figure 4: Mean of nasal airflow for each speaker and each vowel type.

	MT01	MT02	MT03	MT04	MT05	MT06	mean
/a/	0.010	0.003	0.005	0.012	0.015	0.008	0.010
s.d.	0,013	0.003	0.011	0.009	0.019	0.010	0.010
min	0.001	-0.002	-0.010	0.004	-0.002	-0.007	-0.002
max	0.047	0.008	0.025	0.033	0.068	0.032	0.035
/ɑ̃/	0.068	0.065	0.091	0.079	0.094	0.082	0.080
s.d.	0.039	0.026	0.030	0.044	0.041	0.036	0.040
min	0.010	0.025	0.044	0.026	0.050	0.034	0.031
max	0.130	0.01	0.149	0.145	0.167	0.161	0.142
/e/	0.020	0.005	0.023	0.015	0.027	0.019	0.020
s.d.	0.018	0.005	0.029	0.022	0.032	0.031	0.020
min	0.002	-0.0004	0.002	-0.002	0.008	0.003	0.002
max	0.053	0.017	0.084	0.081	0.119	0.106	0.076
/ɛ̃/	0.058	0.068	0.067	0.090	0.100	0.071	0.070
s.d.	0.031	0.025	0.033	0.031	0.036	0.029	0.030
min	0.015	0.016	0.019	0.053	0.053	0.043	0.033
max	0.117	0.092	0.131	0.140	0.164	0.103	0.124
/o/	0.016	0.007	0.018	0.020	0.034	0.017	0.020
s.d.	0.015	0.019	0.024	0.029	0.034	0.029	0.020
min	-0.001	-0.008	-0.003	0.00001	-0.002	0.054	-0.002
max	0.054	0.052	0.069	0.094	0.097	0.091	0.076
/ɔ̃/	0.087	0.086	0.099	0.100	0.154	0.061	0.100
s.d.	0.026	0.019	0.027	0.033	0.064	0.014	0.030
min	0.046	0.047	0.060	0.050	0.071	0.041	0.052
max	0.127	0.111	0.139	0.154	0.237	0.086	0.142

Table 3: Nasal airflow means (l/s) on vowels per speaker.

We can observe from Table 3 that each speaker has a minimum level of nasal airflow which differs between vowels. Also, for each pair of vowels (/a/ vs. /ɑ̃/, /e/ vs. /ɛ̃/, and /o/ vs. /ɔ̃/), the maximum nasal airflow for oral vowels can be larger than the minimum nasal airflow for nasal vowels. This obviously explains some of the misclassifications made by the CNN, in particular for speakers MT01, MT04 and MT05 who have a higher nasal airflow for oral vowels, or an overall lower nasal airflow for nasal vowels. For speaker MT02, the mean nasal airflow for /ɑ̃/ is 0.065 l/s, and all but 1 occurrence of /ɑ̃/ were misclassified below that threshold. As for speaker MT04, all but 2 occurrences of /ɛ̃/ were misclassified below a threshold of 0.059 l/s. Many examples of misclassifications were found according to these criteria. Overall, Pearson's correlation coefficient revealed that the prediction of nasality and non-nasality is correlated with the mean NAF measure (with r = 0.66).

4. DISCUSSION AND CONCLUSION

The main result of this work is the correct automatic classification of nasality for vowels up to 88% from a new acoustic corpus.

We showed that there is a significant connection between CNN probabilities and aerodynamic data. CNN misclassifications for speaker MT01 and MT04 or for the distinction between /a/ *vs.* /ɑ̃/ are found to be correlated with smaller differences in nasal airflow.

Our aim was also to correlate the CNN probabilities with the mean level of nasal airflow per speaker so as to evaluate the overall nasality per speaker and this point still has to be explored further. A first investigation indicated that the probability values given by the CNN were not related to the amount of nasal airflow, although misclassifications may give good hints for that matter.

It is unlikely that the values from the mask were erroneous due to the numerous calibration checks done during the recordings. However, we observed low values of nasal airflow that were unexpected (in particular preceding unvoiced obstruents) and the use of ratio [13] between nasal and oral airflow may provide useful answers.

In the near future, we will work on the activation functions so as to better relate the probability values with the level of nasal airflow. The analysis of spectral zones used by a CNN could be an important insight of our work since the modelling of the relationship between acoustics and articulation is still problematic for nasals. For example, it is known from articulatory studies that /a/ has a lowered velum compared to other vowels and this should

have an impact on the rate of classifications [28]. Overall, the implications from these results should help phoneticians in their analysis of nasal vowels, and it is planned to share that model and aerodynamic mask with the community.

5. REFERENCES

[1] Maddieson, I. 1984. Patterns of sounds. *Phonology*, *2*(1), 343-353.

[2] Styler, W. 2017. "On the acoustical features of vowel nasality in English and French," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 2469–2482, doi: 10.1121/1.5008854.

[3] Clarke, W. M. 1975. "The measurements of the oral and nasal sound pressure levels of speech," *J. Phon.*, vol. 3, pp. 257–262.

[4] Amelot, A. 2004. Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français. Université de la Sorbonne nouvelle - Paris III. Thèse de l'Université de la Sorbonne Nouvelle - Paris III.

[5] Croft, C. B., Shprintzen, R. J., & Rakoff, S. J. 1981. Patterns of velopharyngeal valving in normal and cleft palate subjects: A multiview videofluoroscopic and nasendoscopic study. *The Laryngoscope*, *91*(2), 265-271.

[6] Skolnick, M. L., McCall, G. N., & Barnes, M. 1973. The sphincteric mechanism of velopharyngeal closure. *The Cleft Palate Journal*, *10*(3), 286-305.

[7] Vaissière, J. 1988. "Prediction of velum movement from phonological specifications," *Phonetica*, vol. 54, pp. 122–139.

[8] Clumeck, H. 1976. "Patterns of soft palate movements in six languages," *J. Phon.*, vol. 4, no. 4, pp. 337–351, doi: 10.1016/S0095-4470(19)31260-4.

[9] Basset, P., A. Amelot, J. Vaissière, and B. Roubeau. 2001. Nasal airflow in French spontaneous speech. *J. Int. Phon. Assoc.*, vol. 31, no. 1, pp. 87–99, doi: 10.1017/S0025100301001074.

[10] Bell-Berti, F. and R. A. Krakow. 1991. Anticipatory velar lowering: A coproduction account. *J. Acoust. Soc. Am.*, vol. 90, pp. 112–123.

[11] Krakow, R. A. 1993. Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate. in *Phonetics and Phonology vol. 5: Nasals, Nasalization, and the Velum*, New York: Academic Press.: M.K. Huffman & R.A. Krakow, Eds., pp. 87–116.

[12] Ajili, M., Bonastre, J.-F., Rossetto, S., & Kahn, J. 2016. Inter-speaker variability in forensic voice comparison: a preliminary evaluation. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2114–2118.

[13] Kahn, J., Audibert, N., Bonastre, J.-F., & Rossato, S. 2011. Inter and Intra-speaker Variability in French: An Analysis of Oral Vowels and Its Implication for Automatic Speaker Verification. *ICPhS*, 1002–1005.

[14] Wang, X., Tang, M., Yang, S. *et al. 2019.* Automatic Hypernasality Detection in Cleft Palate Speech Using CNN. *Circuits Syst Signal Process* 38, 3521–3547.

[15] Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., & Al-Dhief, F. T. 2020. Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, *10*(11), 3723.

[16] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., & Woisard, V. 2022.. Validation of the Neuro-Concept Detector framework for the characterization of speech disorders: A comparative study including Dysarthria and Dysphonia. In *Interspeech 2022*.

[17] Ioffe, S., & Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning* (pp. 448-456). PMLR.

[18] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., & Woisard, V. 2020. Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders - step 1: Cnn model-based phone classification. In *Interspeech 2020* (pp. 2522-2526). ISCA; ISCA.

[19] Pellegrini, T., & Mouysset, S. 2016. Inferring phonemic classes from CNN activation maps using clustering techniques. In *Annual conference Interspeech (INTERSPEECH 2016)* (pp. pp-1290).

[20] Elmerich, A., Amelot, A., Maeda, S., Laprie, Y., Papon, J. F., & Crevier-Buchman, L. 2020. F1 and F2 measurements for French oral vowel with a new pneumotachograph mask. In *ISSP 2020-12th International Seminar on Speech Production*.

[21] Honda, K., & Maeda, S. 2008. Glottal opening and airflow pattern during production of voiceless fricatives: a new noninvasive instrumentation. *The Jour. of the Acous. Soc. of Am.*, *123*(5), 3738-3738.

[22] Boersma, P. Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.10, retrieved 17 March 2022 from http://www.praat.org/.

[23] R Core Team. 2022. R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.r-project.org/.

[24] Zerling, J. 1984. Phénomènes de nasalité et de nasalisation vocalique: étude cinéradiographique pour deux locuteurs. *Travaux de l'Institut de phonétique de Strasbourg* 16, 241-266.

[25] Torreira, F., Adda-Decker, M., & Ernestus, M. 2010. The Nijmegen Corpus of Casual French. *Speech Communication*, *52*(3), 201.

[26] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. 2004. The ESTER evaluation campaign of Rich Transcription of French Broadcast News, *Proc. Language Evaluation and Resources Conference*.

[27] Chanclu, A., Georgeton, L., Fredouille, C., & Bonastre, J.-F. 2020. PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire. *6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition)*, 73–81.

[28] Durand, M. 1953. De la formation des voyelles nasales, *Stud. Linguist.*, vol. 7, no. 1–2, pp. 33–53, doi: 10.1111/J.1467-9582.1953.TB00489.X.

Analysis-by-synthesis: phonetic-phonological variation in deep neural network-based text-to-speech synthesis

Christina Tånnander[1,2], David House[1] & Jens Edlund[1]

KTH Speech, Music and Hearing[1] & Swedish Agency for Accessible Media[2]
christina.tannander@mtm.se, davidh@kth.se, edlund@speech.kth.se

ABSTRACT

Text-to-speech synthesis based on deep neural networks can generate highly humanlike speech, which revitalizes the potential for analysis-by-synthesis in speech research. We propose that neural synthesis can provide evidence that a specific distinction in its transcription system represents a robust acoustic/phonetic distinction in the speech used to train the model.

We synthesized utterances with allophones in incorrect contexts and analyzed the results phonetically. Our assumption was that if we gained control over the allophonic variation in this way, it would provide strong evidence that the variation is governed robustly by the phonological context used to create the transcriptions.

Of three allophonic variations investigated, the first, which was believed to be quite robust, gave us robust control over the variation, while the other two, which are less categorical, did not afford us such control. These findings are consistent with our hypothesis and support the notion that neural TTS can be a valuable analysis-by-synthesis tool for speech research.

Keywords: analysis-by-synthesis, latent phonetic features, phonological variation, neural TTS

1. INTRODUCTION

The outstanding success of deep neural network-based text-to-speech synthesis (TTS) in recent years has not gone unnoticed (TTS usually refers to all types of text-to-speech synthesis; in the following we use the term to refer to deep neural network-based text-to-speech synthesis only). In some cases, it is indistinguishable from human speech. Even more impressive and perhaps surprising: it can be and is routinely trained on plain orthographic text. Although some findings show that phonetic or phonological transcriptions improve TTS quality (see e.g. [1, 2]), there are examples where grapheme and phoneme input yield similar results (e.g. [3]). In our own work, we have observed that a more detailed phonemic representation improves the TTS quality significantly.

What is not as widely known is how the transcription (or symbol sequence) used to represent speech impacts the synthesized speech in complex ways. Previous work has studied the impact of *suprasegmental* properties, in particular how duration can act as a proxy for prominence [4, 5]. In contrast, the present work focusses on symbols representing *segmental* speech properties. In essence, this work has dual goals:

Firstly, an improved understanding of the relationship between symbolic representation in training and synthesis can lead to better TTS, which is a key goal in speech technology and one which has societal benefits, perhaps most importantly in terms of increased accessibility – a central goal in many societal policy agendas [6].

Secondly, analyzing the largely opaque but almost unreasonably successful TTS models can provide new insights. We propose, as a novel application of analysis-by-synthesis [see e.g. 7], that post hoc interpretation [8] of TTS output as a function of its input can provide information on *whether a context-dependent phonetic-phonological distinction included in the transcription is reliably realized in the speech used to train the TTS*.

In this paper, we focus on the second of these goals, and conduct initial experiments to investigate if we can indeed learn from post-hoc interpretation of TTS output. We attempt to validate the proposed analysis-by-synthesis method using three well-known phonetic-phonological allophonic variations in Swedish. The first of these variations, the lowering of /ɛ/ and /ø/ in front of /r/, is known to be categorical and in complementary distribution in the dialect represented in our experiments. The second variation, the (de)aspiration of stop consonants following /s/, is described as less robustly governed by phonological context. The third variation, the allophonic variation of /r/ itself, is expected to be more difficult to capture due to its freer nature. We trained multiple TTS models systematically varying the symbol representations of these phonemes and analyzed their outputs in relation to these variations. Our results are consistent with our expectations and support the potential usefulness of TTS as a research tool for speech science.

2. BACKGROUND

2.1. Current TTS research

Most current TTS publications focus squarely on the machine learning (ML) aspect of TTS. They are primarily concerned with the algorithms and the mathematical fit of the model and use speech primarily as a case study for testing. For reasons of comparability of the algorithms used, that type of work exhibits a strong preference for pre-processed, well-known data sets. As an example, the LJ Speech dataset for English TTS [9], with fixed transcriptions and fixed preprocessing of the audio as well as the text, is used for the vast majority of all publications on English TTS. While these studies may be useful for speech technology, they have limited value for speech science. In contrast, we look in this paper to *analysis-by-synthesis* [7] as a method to investigate the nature of speech and provide insight into phonetics and phonology.

2.2. Latent phonetic features

Modern TTS systems are typically trained on a symbol sequence representing the orthographic text and a sequence of spectrograms representing the corresponding speech. This means that no explicit information is provided on the pronunciation of words in isolation or in context, and the representation of the acoustic signal is severely impoverished. Despite this, the speech generated by systems trained in this manner can rate on par with human speech [10, 11], at least within the limitations of current evaluation metrics [12–14].

The fact that TTS systems generate speech reminiscent of a native human speaker despite purely orthographic transcriptions suggests that the models capture pronunciation implicitly. In machine learning, characteristics that are encoded implicitly and opaquely in the models are often referred to as *latent features*, and in text processing, *latent semantic features* denote implicit semantics learned by language models [15]. By analogy, we discuss *latent prosodic features* in [4], and here we use *latent phonetic features* to refer to implicitly learned segmental speech characteristics which can provide insights into phonetics and phonology.

2.3. Effects of symbol inventory in neural TTS

TTS symbol sets that go beyond the graphemes often include representations for a range of speech phenomena. Most commonly, they hold phones or phonemes, stress, and boundary types. The symbol sets of four companies developing Swedish TTS [16–19] all include separate symbols for the lowered variants of short and long /ɛ/ and /ø/. One included separate symbols for aspirated /p, t, k/. None included more than one symbol for Swedish /r/, although some used separate symbols for non-Swedish /r/ sounds (e.g. English [ɹ]).

The selection of speech phenomena in the TTS symbol inventory impacts the TTS output. [20] trained several systems on podcast speech with different symbolic representations of filled pauses. Omitting filled pauses in the transcriptions resulted in a system that inserted filled pauses in reasonable places, but without the means to prevent these insertions. Using a single token for “uh” and “um” allowed for the removal of filled pauses and their insertion in specific places, but when the symbol was used in the text, the choice of "uh" or "um" was made implicitly by the system. Transcribing "uh" and "um" as separate tokens gave the user explicit control over both placement and type. We have observed the same effect with other phenomena, such as audible breath, in our own work. It may be worth noting in this context that explicit control is not always desirable in TTS. While the ability to control filled pauses allows the user to specify their placement, a TTS that inserts them in appropriate places on its own can solve this task without requiring user knowledge. In short, the two solutions will be good for different tasks.

We aim to use the insights from [21] to investigate segmental properties of a single Swedish speaker's speech. Specifically, we will examine three phonetic-phonological variations in Swedish.

2.4. Allophones of the vowels /ɛ/ and /ø/

In many Swedish dialects, and particularly in the eastern part of Sweden including the Stockholm area, the long vowels /ɛː/ and /øː/ are pronounced as the more open allophones [æː] and [œ̝ː] when preceding /r/ or a retroflex consonant [22–25]. This also applies to the short vowels /ɛ/ and /œ/ which are lowered to [æ] and [œ] in the same context. The lowering of short vowels is even more widespread among Swedish dialects than for long vowels [22]. There are exceptions to this pronunciation, such as when a morpheme boundary occurs after the vowel [23]. This variation is known as a robust complementary distribution and is clearly present in our training data.

2.5. Aspiration in the stop consonants /p/, /t/ and /k/

One of the most widely referred to phonetic variations used to illustrate complementary distribution in Swedish (as well as English) is the aspiration of the voiceless stops /p/, /t/ and /k/ and the absence of this aspiration when the stop is preceded by an /s/ in the same syllable [23, 24, 26, 27]. Following a morpheme boundary, such as the initial position of the second

Model	Context for additional symbols	Original symbols (N)	Additional symbols (N)
TTS$_{ÄÖ}$	Vowels before /r/ or retroflex	/ɛ:/ (2073) /ø:/ (2013) /ɛ/ (47996) /œ/ (1797)	[æ:] (2563) [œ:] (2763) [æ] (2515) [œ] (1981)
TTS$_{PTK}$	Voiceless stops after /s/ in same syllable	/p/ (9394) /t/ (36939) /k/ (17737)	[p] (723) [t] (4251) [k] (2424)
TTS$_{R}$	/r/ following voiceless stop in same syllable	/r/ (46403)	[r] (1841)

Table 1. Composition of the three models with separate symbols for pairs of allophones. In the base model **TTS$_{R}$** the original symbol is used for all allophones.

component of a compound (e.g. "*ås-topp"*, en. "ridge") the stop will generally retain its aspiration [22, 23]. The strength of aspiration can vary depending on many factors, and the strongest aspiration generally occurs in word-initial position in stressed syllables. Aspiration is often weaker and can even disappear in unstressed syllables [22, 24].

2.6. Free variation of /r/

The phoneme /r/ exhibits particularly rich phonetic variation and is often referred to as *the class of rhotics*. This variation occurs not only between languages, but also within languages representing both dialectal variation and individual speaker variation. Phonetic realization of /r/ includes trills, taps, flaps, fricatives and approximates as well as the influence on vowel quality by a following /r/. In addition to this variation in manner of articulation, place of articulation can range from dental to uvular [24, 28–30]. In Swedish, we find a pronounced dialectal variation in place of articulation. Southern Swedish dialects employ uvular place of articulation while central and northern Swedish dialects use alveolar place. In between, we find large areas where both places are found in complementary distribution [26, 31, 32].

The rhotics of the central Swedish dialect and the Stockholm area depends on individual preferences, the articulation force used and the degree of formality [24, 31]. However, both Elert and Malmberg point out that trills and taps are more common following a syllable-initial consonant (e.g. *prova, träna, kråka*). [31] includes syllable-initial /r/ before a stressed vowel as a position favoring a trill in certain situations but a fricative realization is also common. This variation is noted to be partially regulated by vocal effort and speaker preferences. According to [31], the fricative is the most common realization following a vowel (e.g. *bara*) or in word-final position (e.g. *pojkar*). [23] even lists /r/-deletion in unstressed positions as characteristic of the Stockholm dialect.

3. METHOD

Our proposed method starts with a TTS model that can produce a specific allophonic variation represented by a single symbol **S**. We then train a new model on the same speech data, but with an extended symbol inventory that separates allophones into separate symbols **S$_1$** and **S$_2$** based on phonological context rules. We test the model by synthesizing an utterance with **S$_1$** in a context where **S$_2$** is expected, or vice versa. If the mismatched allophone is noticeable to human listeners, we have gained explicit control over the allophones, otherwise we have not.

3.1 Assumptions and hypothesis

We assume that if we gain explicit control over the allophones in this procedure, it is because the phonological distinction expressed by the rule matches the distribution of acoustic/phonetic variation in the speech data, which is consistent with the variation being governed by the phonological distinction. Failure to control the allophones may occur for a wide range of reasons and is not informative.

We hypothesize that we will be able to control the famously robust and categorical /ɛ:, ø:, ɛ, œ/ allophones, and unable to control the notoriously free /r/ variation. As for aspiration of consonant stops, we defer prediction as this variation is described as robust in some contexts and much less so in others.

3.2 Stimuli design

We designed stimuli to test our hypothesis using four Swedish neural TTS voices trained on the same audio data of a professional female speaker from the Stockholm area. **TTS$_{BASE}$** used a minimal phonemic symbol inventory with extra symbols for pauses and

word boundaries. The other three models (**TTS$_{ÄÖ}$**, **TTS$_{PTK}$** and **TTS$_{R}$**) added explicit symbols for our selected allophonic pairs of /ɛː, øː, ɛ, œ/, /p, t, k/ and /r/ to the symbol inventories (see Table 1). Each model was trained for 600 epochs (iterations of the entire training data) in Nvidia's PyTorch implementation of Tacotron 2 [33] and synthesized with the WaveGlow vocoder [34] trained 650 epochs on the same voice as the Tacotron models.

We designed three sets of test sentences to showcase each of the three Swedish phonological variations. All sentences were synthesized with the **TTS$_{BASE}$** model. Next, each set of test sentences was synthesized with its respective TTS model in two ways: (1) with the specific symbols in their correct places according to the context and (2) with the opposite symbol of what the context would dictate. We synthesized each sentence five times in each condition to provide a sense of the variation in the models.

3.3 Analyses

To evaluate our hypotheses, we conducted the following analyses: (1) We evaluated the general capability of the **TTS$_{BASE}$** model by listening to and comparing the synthesized sentences to the original speaker's voice, dialect, and speaking style. (2) Three experienced Swedish phoneticians performed a qualitative analysis of the renditions with interchanged allophones for each series of test sentences to determine if our expectations were met, and our hypothesis supported. (3) For the /p, t, k/ test sets, we also conducted a quantitative analysis of voice onset time (VOT), a strong predictor of aspiration.

4. RESULTS

The base model performed well for all test sets. Even though 600 is a quite low number of epochs when training Tacotron 2, the TTS was similar to the original speaker in terms of voice, dialect, and speaking style. No major differences were observed between the test sets, and the allophones of interest were all produced in an appropriate manner.

For **TTS$_{ÄÖ}$**, the renditions with the allophone symbols correctly placed were indistinguishable from the base model. In contrast, interchanging the symbols consistently produced the allophones in a manner that sounded like proper speech, but highly unlikely for the specific voice due to the context. In pre-rhotic position, the more closed allophones gave a clear impression of a different dialect, and in non-pre-rhotic position, judgements varied between yet another dialect or, if semantics permitted, the same word with an infixed /r/, such that "*höna"* ([hœ̝ːna], en. "hen") was perceived as "*hörna"* ([hœ̝ːɳa], en. "corner").

For **TTS$_{PTK}$**, interchanging the symbols used to capture aspirated and non-aspirated voiceless stops did not produce any clear perceptual differences in the renditions. As a follow-up, we measured voice onset time (VOT) of stops after /s/ in the same syllable. We omit the details here since again, no differences were found.

Finally, for the **TTS$_{R}$**, models, the general quality was indistinguishable from the base model regardless of whether the symbols were interchanged. Two listeners perceived that a more pronounced, sometimes trilled /r/ was used more often when the symbol intended to capture trills was used, but we were unable to quantify this observation. The model produced a relatively wide range of /r/ realizations, all sounding appropriate, but these variations could not be predicted by phonological context.

5. DISCUSSION

Our primary goal was not to experiment with phonetic control for TTS production purposes, but rather to use TTS as a tool to gain insights into the powerful latent phonetic features in neural TTS systems. Our hypothesis that using separate symbols for /ɛː, øː, ɛ, œ/ and their lowered allophones, which are reliably governed by phonological context, would allow us to control the distribution of these allophones was borne out. This supports the notion that the ability to control a variation using different symbols that are assigned based on a principle (such as phonological context rules) indicates the accuracy of that principle in describing the variation. By utilizing the complex but inaccessible TTS models as a tool in conjunction with our perception, we can study speech beyond the limitations of easily measurable and accessible phonetic features. While this is currently a proof-of-concept, it has the potential to expand our understanding of speech.

6. ACKNOWLEDGEMENTS

This work is funded in part by the Vinnova funded project Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish (2018-02427). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

7. REFERENCES

[1] Fong, J., Taylor, J., Richmond, K., King, S. 2019. A comparison of letters and phones as input to sequence-to-sequence models for speech synthesis. *Proc. of SSW 10*, ISCA, 223–227.

[2] Taylor, J., Maguer, S. L., Richmond, K. 2021. Liaison and pronunciation learning in end-to-end text-to-speech in French. *Proc. of SSW 11*, ISCA, 195–199.

[3] Perquin, A., Cooper, E., Yamagishi, J. 2020. *Grapheme or phoneme? An analysis of Tacotron's embedded representations*.

[4] Tånnander, C., House, D., Edlund, J. 2022. Syllable duration as a proxy to latent prosodic features. *Speech Prosody 2022*, ISCA, 220–224.

[5] Tånnander, C., Edlund, J. 2021. Stress manipulation in text-to-speech synthesis using speaking rate categories. *Proc. of Fonetik 2021*, Lund University, 17–22.

[6] United Nations. 2015. *Transforming our world: the 2030 agenda for sustainable development*. Department of Economic and Social Affairs, United Nations.

[7] Bever, T. G., Poeppel, D. 2010. Analysis by synthesis: a (re-)emerging program of research for language and vision. *BIOLINGUISTICS* 4, 2–3: 174–200.

[8] Lipton, Z. C. 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3: 31–57.

[9] Ito, K., Johnson, L. 2017. *The LJ Speech dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[10] Shen, J., Pang, R., Weiss, R. J., et al. 2018. Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions. *Proc. of ICASSP 2018*, 4779–4783.

[11] Tan, X., Qin, T., Soong, F., Liu, T.-Y. 2021. *A survey* on neural speech synthesis. arXiv 2106.15561, Jun. 2021.

[12] Wagner, P., Beskow, J., Betz, S., et al. 2019. Speech synthesis evaluation - state-of-the-art assessment and suggestion for a novel research program. *Procs. of SSW10*, ISCA, 105–110.

[13] King, S. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1.

[14] Shirali-Shahreza, S., Penn, G. 2018. MOS naturalness and the quest for human-like speech. *In procs. of SLT 2018*, 346–352.

[15] Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., Miliani, M. 2022. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*.

[16] Acapela Group. 2005. Language manual, Swedish. [Online]. Available: http://www.acapela-vaas.com/Includes/language_manuals/Swedish.pdf

[17] Amazon. 2022. Amazon Polly: Developer guide. [Online]. Available: https://docs.aws.amazon.com/pdfs/polly/latest/dg/polly-dg.pdf

[18] CereProc. 2022. CereVoice phone sets. [Online]. Available: https://www.cereproc.com/files/CereVoicePhoneSets.pdf

[19] Microsoft. 2022. SSML phonetic alphabets. [Online]. Available: https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-ssml-phonetic-sets

[20] Székely, É., Henter, G. E., Beskow, J., Gustafson, J. 2019. Spontaneous conversational speech synthesis from found data. *Interspeech 2019*, ISCA, 4435–4439.

[21] Székely, É., Eje Henter, G., Beskow, J., Gustafson, J. 2019. How to train your fillers: uh and um in spontaneous speech synthesis. *Proc. of SSW 10*, ISCA, 245–250.

[22] Bruce, G. 2010. *Vår fonetiska geografi - Om svenskans accenter, melodi och uttal*. Studentlitteratur, Lund.

[23] Elert, C.-C. 1970. *Ljud och ord i svenskan*. Almqvist & Wiksell Förlag AB, Stockholm.

[24] Malmberg, B. 1968. *Svensk fonetik*. Gleerups, Lund.

[25] Schumacher, A. 2014. Hät, här, höt and hör - An articulatory and acoustic study of the Swedish vowels /ɛ:/ and /ø:/, Master Thesis, Lund University, Lund.

[26] Elert, C.-C. 1981. *Ljud och ord i svenskan 2*. Almqvist & Wiksell International, Stockholm.

[27] Ladefoged, P. 1982. *A course in phonetics*. Harcourt Brace Jovanovich, Inc., New York.

[28] Ladefoged, P., Maddieson, I. 1996. *The sounds of the world's languages*. Blackwell Publishers, Oxford.

[29] Laver, J. 1994. *Principles of phonetics*. Cambridge University Press, Cambridge.

[30] Lindau, M. 1980. *The story of /r/*. UCLA.

[31] Elert, C.-C. 1995. *Allmän och svensk fonetik*. Norstedts Förlag AB, Stockholm.

[32] Sjöstedt, G. 1936. Studier över r-ljuden i sydskandinaviska mål.

[33] Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. 2017. Tacotron: towards end-to-end speech synthesis. *Proc. of Interspeech 2017*, ISCA, 4006–4010.

[34] Prenger, R., Valle, R., Catanzaro, B. 2019. Waveglow: a flow-based generative network for speech synthesis. *Proc. of ICASSP 2019*, IEEE, 3617–3621.

NATURAL CHOICE: COMPARING PLACE CLASSIFICATION BETWEEN NATURAL AND TACOTRON FRICATIVES

Ayushi Pandey[1], Sébastien Le Maguer[1], Jens Edlund[2], Naomi Harte[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

[2]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
pandeya@tcd.ie, lemagues@tcd.ie, edlund@speech.kth.se, nharte@tcd.ie

ABSTRACT

As commercial prevalence of high-quality TTS voices continues to grow, little is known about their suitability for speech sciences. In this paper, we compare feature importance in obstruent place classification between natural voice, and two TTS systems: Tacotron WaveNet and Tacotron WaveGAN. Support Vector Machines (SVMs) are used for classification, using 20 acoustic-phonetic features. First, individual feature strength was evaluated by passing each feature individually to the SVMs. Then, classification was made using the full 20 featureset, to test the most high performing feature combination.

Top-ranked predictors in natural voice were compared with each of the Tacotron voices. We find that Tacotron voices greatly resemble natural in contrasts involving sibilant fricatives w.r.t similarity with high-performing features. But in non-sibilant fricatives, we find greater dissimilarity between natural and Tacotron voices.

Keywords: obstruents, Tacotron, WaveNet, WaveGAN, support-vector-machines

1. INTRODUCTION

In days of formant synthesis, speech science and synthesis technology enjoyed a reciprocal relationship. However, synthetic speech thus produced was more unintelligible than human speech, and its findings often did not extend to human language [1]. Although intelligibility improved with statistical parametric synthesis, it brought a roboticity or “unnaturalness" to the speech output [2]. But today, with neural Text-To-Speech (TTS), high-quality, natural-sounding synthetic speech has become quite accessible. Malisz et al. [3] show that data-driven TTS is now more realistic, highly intelligible, and perceptually closer to human speech. They argue that conclusions drawn on speech produced by modern TTS may generalize better to the human voice. It may thus be time, once again, for a closer exchange between speech science and synthesis technology.

Using synthetic speech as a research tool has attracted many phoneticians. Vector space embeddings in WaveNet synthesizers have been used to analyze prosodic patterns in Lombard speech, proposing new methodologies for prosodic research [4]. Controllable architectures like WaveBender [5] have reinstated the scope of feature-based phonetics within modern TTS. Additionally, using conversational synthetic speech has provided greater flexibility over controlled variation in determining paralinguistic characteristics [6]. Speech science has also increasingly contributed *to* TTS. Forensic investigation of speech spoofed through TTS provided important insights into human-likeness [7]. Segmental analysis of obstruents using acoustic-phonetic features in [8], have revealed system-specific weaknesses in WaveNet TTS. Furthermore, evaluation paradigms have borrowed actively from phonetics [9].

However, the majority of these papers involve prosodic explorations of synthetic speech. Phonemic contrast, which took centre stage in early formant synthesizers, has been overlooked in neural TTS. If contrastive trends in synthetic speech can generalize well to human speech, dependence on data collection can be immensely reduced. Phonemic contrasts could be synthesized in various positional, vocalic and cluster contexts, including speaking styles. Additionally, the success of multi-speaker, and accented TTS [10] can ensure the necessary diversity required to understand acoustic invariance.

To investigate whether phonemic contrast is maintained in neural TTS in the same way as in a human voice, this paper provides **place classification** of English fricatives as a targeted test case. Recently, voiceless fricatives have been shown to deviate from a natural voice in neural TTS [8]. However, despite deviation, if contrast

is encoded in the same parameters as in the natural voice, then TTS voices may be suitable tools for speech science research. Conversely, if phonemic contrast is indexed by divergent trends in TTS voices, then generalization may become more difficult. Additionally, unexpected acoustic detail may enforce a cognitive load condition and increase reliance on lexical cues [11]. This can cause greater problems for non-native listeners [12].

We compare **feature importance** for place classification between a natural voice, and two TTS systems: Tacotron WaveGAN and Tacotron WaveNet. The corpus for our analysis comes from the recently extended Blizzard Challenge 2013 (BC-2013) [13]. Features are analyzed both individually and as an ensemble, using Support Vector Machines (SVM)s for place classification between contrastive pairs (e.g, /s-ʃ/). Top-ranked features in each contrastive pair are compared. Through this analysis, we explore whether contrastive trends in Tacotron voices are similar to the natural voice in terms of top-ranked features. We find that sibilant fricatives in both the voices show very similar trends to the natural voice, while non-sibilant fricatives need further exploration.

2. EXPERIMENTAL SETUP

2.1. Dataset and features

2.1.1. Description of the dataset

The Blizzard Challenge 2013 [14] provides single-speaker read-speech data, collected from a female speaker of standard American English. Recently, the original 2013 challenge was extended [13] to include modernTTS voices. For the present analysis, we selected Tacotron WaveGAN (system R) , Tacotron WaveNet (system Z), and the natural voice. While both systems use Tacotron [15] as the acoustic model, waveform generation is handled by WaveGAN [16] and WaveNet [17]. With this dataset of a 100 sentences produced by each of the three voices (300 sentences total), we analyzed place classification in the following contrastive pairs of fricatives (*voiceless:*/f-θ/, /θ-s/, /s-ʃ/, *voiced:*/f-ð/, /ð-z/). Obstruents occuring in unstressed and consonant cluster positions were removed, hence the voiced sibilants /z-ʒ/ were excluded.

2.1.2. Feature extraction

A set of acoustic-phonetic features were extracted from these obstruent consonants in natural and each of the Tacotron voices. In English fricatives, place is determined largely by static spectral characteristics such as spectral tilt, shape, center frequency [18, 19]. Formant transitions have also frequently been discussed [20], although with inconsistent findings.

From the **vocalic** portion of the CV syllable, we extracted: *vowel duration*; *vocalic RMS amplitude*; *Formant values (F1-F3) at onset and midpoint*; *relative amplitude of F3*; *vocalic spectral tilt* and *Delta F1-F3*.

From the **consonantal** portion of the syllable, we extracted: *consonant duration*; *noise duration*; *RMS amplitude*; *spectral tilt*; *spectral shape*; *peak frequency*; *peak amplitude*; and *dynamic amplitude*. All these features have been compared across all English obstruents in [21] for various contrastive strength. They have also been used in analysing obstruents of WaveNet vocoders in [8].

2.2. Feature importance: individual and ensemble

Feature importance provides important links between speech stimuli and phonemic perception. Feature importance was analyzed in R using a two-pronged SVM based classification, and complemented with statistical analysis. This approach is greatly inspired by Styler's [22] work on nasality feature analysis. All features were pre-processed by median-normalization and scaling, and imbalances are removed by undersampling the abundant phonemic class. SVMs use a 5-fold cross validation, and a radial-basis kernel throughout.

2.2.1. Single-feature model (SF model)

In this step, place classification is performed using only 1 feature at a time. This reveals the independent predictive capacity of the predictor, before it is cast into a high-dimensional space.

Feature importance is calculated simply by checking the accuracies as returned by each feature. Next, in addition to accuracy, top-ranked features identified above were also checked for their statistical signficance in the dataset, using the Kruskal-Wallis rank sum test. If complementary information is provided by both the dataset and SVM, then it is crucial for the feature to be properly reproduced in Tacotron voices.

Similarity between natural and Tacotron voices is manually compared, using predictor accuracies, statistical significance scores, and median differences between phonemes.

2.2.2. Full-feature ensemble model (FF model)

In this step, we provide all 20 features to the SVM as an ensemble. This resembles real-world scenarios of speech perception, where all the contrastive information is available to the listener. This step investigates whether the same *feature combination* is shared between natural and Tacotron voices.

First, we calculate the accuracies and F1 score for each contrastive pair, for natural and both Tacotron voices. Then, ***feature importance*** is obtained using the DALEX [23] library, which ranks features based on the 1-AUC score. Thus, we obtain the top 7 (1/3rd of 20) most performant features for every contrastive pair in natural and Tacotron voices. Next, we compute a ***similarity score*** based on the Sørensen's-Dice coefficient. This calculates the magnitude of overlap among the top-7 features between natural voice and each Tacotron system. A high-score indicates greater overlap between high-performing feature combinations.

3. RESULTS

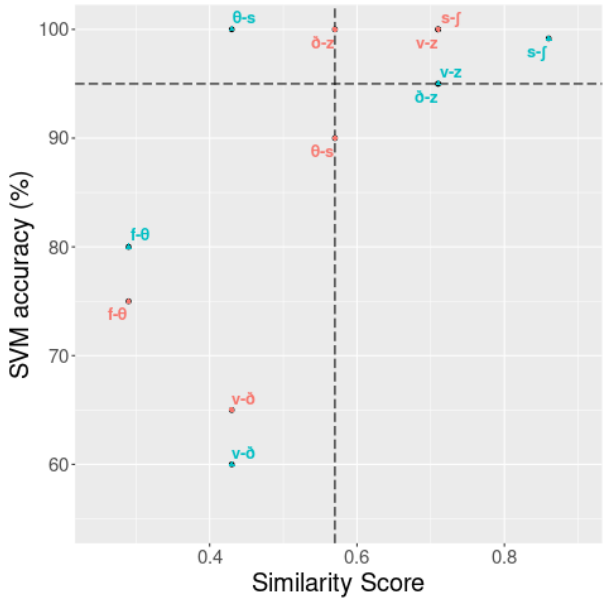


Figure 1: Relationship between accuracy and feature similarity with natural voice. Labels for contrastive pair presented for Tacotron WaveGAN (R) and Tacotron WaveNet (Z). Dashed lines show medians on every axis.

In this section, we present the results of place classification in sibilants, between sibilant-non-sibilants, and between non-sibilants. First, we report single-feature accuracies and statistical significance of top-ranked features in each voice using the SF model. These values are presented comparatively between natural and each Tacotron voice. Then, using the FF model, we report the accuracies of place classification, and similarity between top-7 features of natural and each Tacotron voice. Figure 1 displays the relationship similarity and accuracy obtained using the FF model.

3.1. Within-sibilants

***Voiceless* /s-ʃ/**: In sibilant place identification, both systems Z and R show remarkable similarity to natural voice. Through the single-feature analysis, we find **spectral shape** to be the most powerful predictor of sibilant place, with its standalone accuracy reaching 100% in natural voice, as well as in both the synthetic voices . In natural voice, the spectral shape of /ʃ/ is higher by +14.12 dB compared to /s/, [$\chi^2(1) = 86.26$, p-val < 0.0001]. The distinction of higher spectral shape in /ʃ/ is consistently maintained in system Z (median difference +13.29 dB), and in R by +12.83 dB respectively. This means that spectral shape is a robust indicator of within-sibilant place in natural voice, and gains a similar prominence in synthetic voices as well.

Using the FF model, we find near-perfect accuracy for place classification between sibilant fricatives. Figure 1 shows their clusters in the top-right quadrant, displaying high similarity with natural, as well as above-median accuracy. System Z scores 0.99, and R 0.71, on the Sørensen's distance metric. This indicates a very strong match in the features selected by the natural voice, and those by both the Tacotron systems.

Therefore, in addition to spectral shape, other high-performing features are also consistent in natural as well as synthetic voices for place identification in sibilants.

3.2. Sibilant vs non-sibilant

***Voiceless* /θ-s/**: Here, both Tacotron voices show high accuracy of classification, but share only a few features with natural voice. Through a single-feature analysis, we find a source of divergence between systems on the basis of **spectral tilt**. In Tacotron voices consonantal spectral tilt shows high standalone accuracy [Accuracy = 100%, $\chi^2(1) = 8.75$, p-val < 0.0001]. In natural voice, although spectral tilt for /s/ is significantly higher (p-val $<$ 0.001), it is a weak predictor for SVM, with only 55% individual accuracy. Using the FF model, system Z shows a 100% accuracy, with Sørensen's score of only 0.43. This means that the feature combination is not similar to natural, but a divergent

model gives comparable accuracy of classification.

Therefore, features in /θ-s/ place identification are more divergent from the natural voice, compared to within-sibilant place identification. However, significantly higher spectral tilts in /s/ means that the synthetic data, at least, preserves contrastive trends like the natural voice.

***Voiced* /ð-z/**: For the voiced sibilant-non-sibilant contrast /ð-z/, we see that Tacotron voices show greater similarity with natural in terms of features. Through a single feature analysis, we find that consonantal **spectral tilt**, **dynamic amplitude** and **peak amplitude** are robust across natural voice, as well as both the Tacotron voices. Here, dynamic amplitude between /ð/ and /z/ show statistically significant differences, [$\chi^2(1) = 12.79$, p-val < 0.001], with median dynamic amplitude of /z/ being +30.5 dB higher than θ. Comparable trends can be seen in Tacotron voices as well. Next, a high Sørensen's score of 0.71 and 0.57 in systems Z and R, indicate a majority of features in the full-model SVM are shared with natural.

Therefore, we can see that place classification between voiced sibilant-non-sibilant pair /ð-z/ shares great similarity with natural.

3.3. Within non-sibilants

***Voiceless* /f-θ/**: For this contrast, we can see very divergent trends. Figure 1 displays them in the bottom-left quadrant, indicating a low similarity with natural, as well as below-median accuracy. Next, we find that the contrast is very weak and confusable in natural voice, with only a few features showing modest statistical differences (p-val < 0.01). An SF model reveals consonantal RMS amplitude as the strongest predictor in natural voice [Accuracy 75%, p-val = ns]. On the other hand, both the Tacotron voices show high standalone accuracies for noise duration (System Z: [Accuracy 75%, $\chi^2(1) = 10.30$, p-val < 0.01], System R: [Accuracy 80%, $\chi^2(1) = 9.08$, p-val < 0.01]), which is conspicuously absent as a high-ranking feature in natural voice.

Using the FF model, we find that accuracy in natural voice is only 55%. This means that accuracy drops for natural when more features are added, and feature selection must be much more careful. Conversely, in Tacotron voices, accuracies are higher than chance in the FF model, even though it is lower than median.This means that features selected by SVM for /f-θ/ are different from natural.

***Voiced* /v-ð/**: Similarly in the voiced condition, both Tacotron voices show a low Sørensen's score, indicating only a minimal overlap with natural voice in the full-model SVM. As observed in the voiceless condition, noise duration is an important predictor of place in Tacotron voices [Accuracy 70%, $\chi^2(1) = 3.28$, p-val > 0.05] in system Z, and [Accuracy 70%, $\chi^2(1) = 3.28$, ns] in system R.

Therefore, in place identification between non-sibilants, i.e, /f-θ/, and /v-ð/, we see strongly divergent trends from natural voice in terms of feature importance, and classification results. So, a much deeper exploration is needed before their use can be recommended for speech sciences.

4. DISCUSSION & CONCLUSION

In this paper, we compared feature importance in place classification of obstruents between a natural voice and two Tacotron voices. Using a single-feature SVM model, we compared individual predictors that were most informative in each voice. Then, using a full-feature model in SVM, we compared feature importance between the voices using a similarity metric. We found that sibilant fricatives produced by both Tacotron vioces followed similar contrastive trends to the natural voice. However, those involving non-sibilant fricatives differ from the natural voice. A clear separation can be seen in Figure 1. Through these findings we can recommend that studies conducted on synthetically produced sibilant contrasts may be generalized to human speech.

Our findings are consistent with previous work on sibilants, w.r.t. the resilience of their features in impaired listening conditions [24]. Similarly, in non-sibilants, inadequate feature distinctions [25], and place confusions between /f-θ/ have been reported [26] for natural speech. Both Tacotron voices appear to overcome this inadequacy, basing classification primarily on longer noise durations in the labiodental /f/ and /v/. However, previous evidence suggests that spectral features are more informative of place in non-sibilants, instead of durations [19]. This indicates that insufficient contrastive information in the natural voice may cause cue trading in Tacotron voices to maintain contrast. This introduces an acoustic detail which may not be expected by listeners. Whether this impacts cognitive load, needs to be substantiated with carefully controlled perception experiments. Similarly, while this study shows featural importance through SVM and statistical tests, the perceptual relevance of these features should be established through subjective category perception tests, and supported further through cross-linguistic explorations.

5. ACKNOWLEDGEMENTS

This work has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, the ADAPT Centre (Grant 13/RC/2106), and a Google Faculty Award.

6. REFERENCES

[1] S. A. Duffy and D. B. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Language and Speech*, vol. 35, no. 4, pp. 351–389, 1992.

[2] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[3] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *International Congress of Phonetic Sciencesnces ICPhS 2019 5-9 August 2019, Melbourne, Australia Melbourne Convention and Exhibition Centre*, 2019.

[4] J. Šimko, M. Vainio, A. Suni *et al.*, "Analysis of speech prosody using wavenet embeddings: The lombard effect," in *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[5] G. T. D. Beck, U. Wennberg, Z. Malisz, and G. E. Henter, "Wavebender gan: An architecture for phonetically meaningful speech manipulation," *arXiv preprint arXiv:2202.10973*, 2022.

[6] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[7] C. Kirchhuebel and G. Brown, "Spoofed speech from the perspective of a forensic phonetician," *Interspeech 2022*, 2022.

[8] A. Pandey, S. L. Maguer, J. Carson-Berndsen, and N. Harte, "Production characteristics of obstruents in wavenet and older tts systems," in *INTERSPEECH*, 2022.

[9] E. Gutierrez, P. O. Gallegos, and C. Lai, "Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm," *ArXiv*, vol. abs/2107.02527, 2021.

[10] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, "Boffin tts: Few-shot speaker adaptation by bayesian optimization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7639–7643.

[11] S. L. Mattys and L. Wiget, "Effects of cognitive load on speech recognition," *Journal of memory and Language*, vol. 65, no. 2, pp. 145–160, 2011.

[12] S. L. Mattys, L. M. Carroll, C. K. Li, and S. L. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers," *Speech communication*, vol. 52, no. 11-12, pp. 887–899, 2010.

[13] S. L. Maguer, S. King, and N. Harte, "Back to the future: Extending the blizzard challenge 2013," in *Interspeech*, 2022.

[14] S. King and V. Karaiskos, "The blizzard challenge 2013," in *The Blizzard Challenge Workshop*, 2013, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.

[15] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *ArXiv*, vol. abs/1703.10135, 2017.

[16] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[17] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, 2016.

[18] A. Jongman, "Duration of frication noise required for identification of english fricatives," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1718–1725, 1989.

[19] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[20] A. Wagner, M. Ernestus, and A. Cutler, "Formant transitions in fricative identification: The role of native fricative inventory," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 2267–2277, 2006.

[21] C. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," *PhD dissertation, University of Kansas, Lawrence*, 2020.

[22] W. Styler, "On the acoustical and perceptual features of vowel nasality," Ph.D. dissertation, University of Colorado at Boulder, 2015.

[23] P. Biecek, "Dalex: explainers for complex predictive models in r," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3245–3249, 2018.

[24] J. Meyer, L. Dentel, and F. Meunier, "Speech recognition in natural background noise," *PloS one*, vol. 8, no. 11, p. e79279, 2013.

[25] S. J. Behrens and S. E. Blumstein, "Acoustic characteristics of english voiceless fricatives: A descriptive analysis," *Journal of Phonetics*, vol. 16, no. 3, pp. 295–298, 1988.

[26] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken english fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.

KNOWLEDGE-DRIVEN VS. DATA-DRIVEN METHODS FOR FILTERING ACOUSTIC MEASURES IN PHONETICS CORPORA

Lancien Mélanie, Adda-Decker Martine, Stuart-Smith Jane

FNSRS, LPP, LISN, GULP
melanie.lancien@unil.ch, madda@limsi.fr, Jane.Stuart-Smith@glasgow.ac.uk

ABSTRACT

This paper addresses the issue of vowel formant filtering for large scale phonetic analyses and evaluates an innovative data-driven method to remove spurious items. Automatic formant detection is error-prone due to formant jumps among other issues. A common solution is to adopt formant filters (FF) discarding tokens with measurements falling too far away from knowledge-based references. The proposed approach uses the Mahalanobis distance (MD) as a purely data-driven method. First, all vowel formant and duration values are used to compute vowel profiles. These learnt profiles supersede the need for reference values to carry out the filtering. We compare the two (knowledge-based vs data-driven) filtering approaches on the same dataset of French spontaneous speech. Results demonstrate the efficiency of MD filtering. The amount of filtered data is easily adjustable. Moreover, the data-driven status of the approach makes it well suited for less described languages.

Keywords: Vowels, Formants, Data Science, Automatic filtering, Corpus Phonetics

1. INTRODUCTION

Data derived from automatically annotated medium-size corpora are generally numerous and noisy. Different kinds of error arise from automated annotation and metrics extraction, such as formant-tracking jumps, transcription issues, or mismatches between the expected and the actual pronunciation of a word. These then lead to the presence of erroneous datapoints in data for analysis. The results from very large datasets may not be overly affected, but for medium-size datasets (e.g. with 4 or 5 tokens of each vowel type per speaker), more common in phonetic studies, erroneous values are more problematic, likely biasing subsequent results. Thus medium-size datasets need "cleaning" (filtering) as far as possible.

Unfortunately, the filtering method, and choices and outcomes linked to the process, are rarely reported. Among the few who do, [1] in a large-scale study on English /s/ showed a rejection rate of 10% with range filter methods, and [2] found 17% of her tokens were classified as non canonical with an ABX alignment tool [3] (see subsection 1.2). Thus data filter may result in rejecting up to 20%, whatever the corpus and method.

1.1. Post extraction methods

Several methods can be used to identify erroneous values during or after the value extraction process, some are knowledge-driven, others more data-driven. In phonetic studies, formant values are often extracted using Praat Burg algorithm [4]) and then checked via values range filters (e.g. [5]). This procedure is a knowledge-based "posthoc" procedure that consists of going through all acoustic values and identifying all likely erroneous values of e.g. F1 and F2 (e.g. F1 at 1500Hz for French /i/). However, "range filter" methods only account for expected values for the phoneme, and sometimes speaker gender; [5].

More recently, [6] tried another way of processing the extracted values, using Mahalanobis distance. This allowed the identification of erroneous datapoints based on a multi-parametric distribution and distance [7]. Mahalanobis distance is a multidimensional measure of the number of standard deviations between a point P and the mean of a distribution D; in our case P is a token of a vowel and D the distribution of the variable X for the group to which P belongs (e.g. P is an [i] and D is the distribution of the F2 of the /i/ type). The Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_p)^T$ to a set of mean value vectors $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$ and having a covariance matrix Σ is defined as follows:

$$(1) \quad Dx = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

The square root of $Dx = (x-\mu)^T\Sigma^{-1}(x-\mu)$ gives the number of standard deviations between the observation and the mean of the distribution. If $P = \mu D$, the distance is 0, and increases as P moves away from the mean in a determined space. Thus by choosing a threshold for the distance ([6] chose 3 standard deviations), it is easy to discard vowel tokens with features further than X standard deviations from the computed average profile.

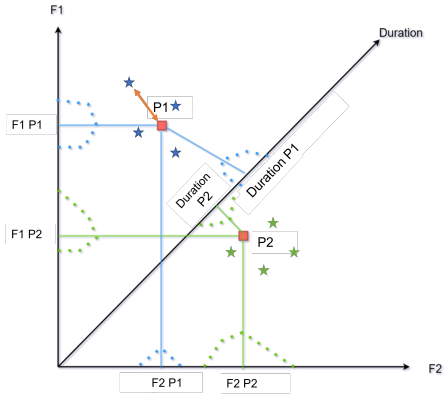


Figure 1: Schematic explanation of the Mahalanobis distance. Blue lines represent the distance of F1 (axis y) F2 (axis x) and duration (axis z) between the mean (orange squares) of two vowels (P1 and P2). Stars are tokens of P1 and P2 and the orange arrow indicate the distance measured.

1.2. Pre extraction methods

Other methods, cast in the NLP scope, can prevent the generation of some erroneous values directly during the annotation or extraction process. For instance, [2] uses sound prototypes and forced alignment to reclassify datapoints that do not match the sound x's profile directly during the labelling phase, by labelling them phone y instead of x. With this method, for example, occurrences of French /b/ (whose prototype will show negative VOT) which do not have a negative VOT will be annotated by the ABX system as its unvoiced counterpart /p/. Such methods can help identify the phonetic realization of sounds actually produced, but are restricted to the profiles of variation implemented in the system. Nor can they discard tokens that are too far away from the target or are noise resulting from alignment issues, without being given a typological description.

Finally, recent propositions like [8] with its new iteration of the FAVE automated formant measurement ([9, 10]), also uses Mahalanobis distance as part of the procedure. A matrix of several re-measurements of formants on the same phone is used to eradicate errors directly during the formant detection and measurement process instead of post-processing the extracted values with filters. Each vowel is measured several times with different numbers of LPC coefficients, and the resulting measures are compared to a prototype (mean formant frequencies/bandwidths) and a covariance matrix. The selected set of measures is that with the smallest Mahalanobis distance from the prototype. This leads to a smaller amount of erroneous values in the output, given that they are corrected directly during the extraction phase. However this method cannot be applied as a posthoc process, and must be run on the raw sound files, which is quite a limitation for studies on open-source datasets, relying on previously measured files (see SPADE [11] or ESTER [12]).

1.3. Conclusions on the state of the art

Even though they are largely preferred (maybe because easier to use), range filters are not optimal since the boundary values must be set by the user implying two main limitations : 1) there must be literature on those values for the given language; 2) ranges are generally set in a wide and arbitrary fashion (e.g. [5] use ranges from 1500 to 2500Hz for French /i/ F2). However it does efficiently identify the most obvious erroneous values.

In this paper we compare the two posthoc (post extraction) filtering methods mentioned above to provide more background on how they operate and which results can be expected. Thus we first replicate [5]'s work with the "knowledge-based" formant range filter (FF hereafter), and then uses a filter based on Mahalanobis' distance (MD), as in [6], on the same dataset.

2. KNOWLEDGE-BASED FORMANT FILTERS

In [5] the authors work on manually transcribed broadcast news corpora in French and German. They extract vowel formants and want to provide a clean dataset without manual checking. To do so, they use formant range filters to get rid of errors.

Their corpus consisted in 4h of journalistic speech: 2h in French (1h ♂, 1h ♀) and 2h in German (1h ♂, 1h ♀). The corpus was automatically segmented and labelled into phones using the LIMSI speech alignment system [3].

Oral vowel formant extraction was made thanks to the Burg algorithm implemented in Praat [4].For each formant, three values were computed on a given vowel segment (corresponding to locations at the first third, the middle and the last third of the segment). Then, these three measurements were averaged to provide a single formant value per vowel token. The filter was setup to take into account the vowel type and the speaker's sex. It excluded all the tokens with formant values not falling into preset frequency ranges. These ranges were chosen in a very tolerant way, based on Calliope [13] reference values. For instance, [i] was considered as erroneous if F1>750Hz (♂) or 900Hz (♀), and/or F2 values were not between 1500-2500Hz (♂) and 1600-3100Hz (♀).

About 4% of tokens were excluded. According to the authors, most of these vowel segments were very short in duration, or displayed devoicing, thus making formant detection more complicated. Compact high vowels (such as /u/) were also particularly prone to rejection as formant extraction sometimes couldn't distinguish F1 from F2.

3. REPLICATION OF THE G&A FF STUDY

3.1. ESTER subset

We selected a similar but different subset of 2h (1h ♀ and 1h ♂) from the ESTER corpus (namely France Inter radio files). This choice allows us to get an idea of the variation intrinsic to speech and speakers independently of a change in the filtering approach. Our subset gathers 26249 oral vowel tokens (against 24000 for [5]).

3.2. FF approach applied to the ESTER subset

Here we replicate the G&A study using their formant range filters. The formant values were also provided by the authors of[5].

Not surprisingly, we end up with rather comparable rejection rates. Among our 26k tokens, 3.2% were excluded by the filter against 4% in [5]. However, as can be seen in Table1, a lot more of /o/ vowels were excluded in our replication. This might at least partially be related to the very low number of /o/ occurrences in our subset (733 vs. 3824 for /i/, for instance).

Similarly, the mean formant values are quite comparable to [5], and even slightly closer to the reference values of [13]. This might be partially due to the choice of France Inter radio news, which is the most normative sample in ESTER.

Figure2 gives our (ESTER-FF - in light blue) and [5] (G&A - in coral) mean formant values, as well as Calliope's values [13] (in black) which were used as references for the filters. for /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/ to ease the comparison (only males data and a subset of vowels are presented for the sake of readability).

	i	**y**	**e**	**ɛ**	**a**	**oe**	**ø**	**ɔ**	**o**	u
G&A	5	15	1	0.3	0.6	4	0.4	1	4.9	25
replic.	2	5	0.3	.07	0.3	1	1	3	17	27

Table 1: Proportion of rejected segments for each vowel type (in %) for [5] and our replication.

4. DATA-DRIVEN FORMANT FILTERS

We now get back to our sample of ESTER and filter it again but with our data-driven method (MD).

We consider a three-dimensional space made of vowel tokens' duration (s), mean F1 (Hz), and mean F2 (Hz). In addition to the traditional mean F1 and F2, we take into account phone duration for it may be helpful to get more "discriminating" profiles.

The filter was coded as a R script [14], mainly using the *mahalanobis* function from *mvtoutlier* package [15].

The range filter (as set-up by [5]) considered the vowel type and the speaker's gender. However we tried different ways to set up MD calculation. The profile of vowels was computed in three fashions: accounting for the vowel type (T) only, the interaction between T and the speaker (S), or the interaction between T and the speakers' gender (G) - the later being the closest to [5]. For each three set-ups, the computed distance had a mean of ~ 2.9. Thus we choose $MD > 3$ as our threshold to tag values as erroneous. In other words, for each set-up, each vowel token that was further than 3 standard deviation away from its computed profile was tagged erroneous. Table 2 gives the proportion of rejection per vowel type, as a comparison with [5]'s results given in Table 1.

	i	**y**	**e**	**ɛ**	**ø**	**oe**	**a**	**o**	**ɔ**	**u**
T	19	22	28	29	23	25	30	28	29	27
T*S	25	31	29	34	32	30	31	34	35	33
T*G	18	23	26	29	25	25	29	29	29	27

Table 2: Rejection rate for each vowel type in each set-up used for MD calculation (in %). T = vowel type (e.g. /i, u, . . . /), S = speaker, G = gender (F/M).

The method allowing to keep the most data while excluding the most extreme values was the one using vowel type and gender (T*G, 26% rejection), closely followed by T (only the vowel type, 27%

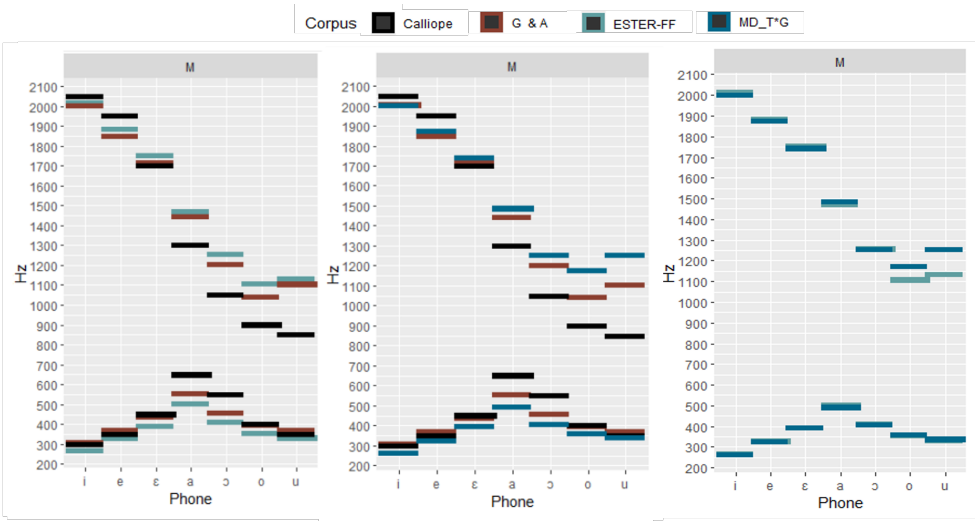


Figure 2: Comparison of formant structures for Calliope (Call) [13], [5] (G&A), our replication (ESTER-FF), and MD by Type and Gender (MD_T*G). Male speakers only.

rejection), T*S having higher rejection rate (31%).

For the three set-ups, we checked the resulting mean formant values and found pretty similar results both between our three set-ups, and between the MD filters and the formant range filter (see figure 2 where we only plot T*G - dark blue - for clarity).

Eventually, we ressorted to Linear Mixed Models (LMM)[16] to check the effects of the filters on the resulting formant values. To lower the number of levels for the factor âvowel typeâ, we grouped vowel types into 4 categories: Diffuse (/i, y, e, ɛ/), Compact (/o, u/), Mid (/ɔ, oe, ø/) and /a/ which was left in its own category for it is the only open and most variable vowel in French. For each of our LMMs, we compared different structures by adding fixed effects as we went along. These comparisons were made using likelihood ratio tests. When the LRT was significant ($\alpha < .05$), the most complete model was kept for further analysis. After model comparison the most parsimonious model was the one built as follow : $MeanFi \sim MethodOfFiltering * Gender * Phone + (1|Word) + (1|Speaker)$. The significance of each fixed factors was assessed using a Type III ANOVA.

Both models (on mean F1 and mean F2) showed a significant main effect of the method and its interactions with other IV as well as the 3 way interaction MethodOfFiltering*Gender*Phone ($p<.03$ for F1 and $p<.001$ for F2). However the analysis of the three way interaction and posthoc tests showed that the differences between the methods arised in very specific spots, such as males' compact vowels. The marginal estimated means for F2 differed by 152Hz max (compact ♂, FF having the lowest value), and 22Hz max for F1 (compact ♂, FF having the lowest value).

5. CONCLUSIONS

Mahalanobis distance was previously used by [8] as a way to choose the best measures during the formant computing phase but here we want to propose its use as a posthoc (post extraction) tool to correct datasets; as nowadays shared and open source data are available as spreadsheets, one might need a post processing tool. In this paper we wanted to put to the test the use of Mahalanobis distance, a data-driven method, as a way to filter erroneous values from acoustic phonetic datasets.

Our replication of [5] with a formant range filter and Mahalanobis distance showed that MD method have higher rejection rates then FF. The remaining tokens however have very similar formant values with both methods. One explanation is that MD method as we set it might remove more fine-grained variation that might be (socio)linguistically relevant. Thus we need to try several tightness degrees by moving the threshold for rejection (e.g. 4 or 5 standard deviations from the profile instead of 3) to have a less restrictive filter.

However MD is data-driven and gives a continuous variable (standard deviation from the profile computed on the data) as the index for rejection, meaning that the user can set it up according to the needs of the study in terms of data homogeneity. Therefor it is an efficient and adaptive tool for filtering extreme acoustic values.

6. ACKNOWLEDGEMENTS

The authors want to thank the Swiss National Science Fundation for the funds made available for this project, as well as the recorded speakers for the data we are using.

7. REFERENCES

[1] J. Stuart-Smith, M. Sonderegger, R. Macdonald, J. Mielke, M. McAuliffe, and E. Thomas, "Large-scale acoustic analysis of dialectal and social factors in english/s/-retraction," 2019.

[2] D. Amazouz, M.-A. Decker, and L. Lamel, "Variation du voisement des occlusives orales en code-switching: analyses par abx automatique et mesures acoustiques," in *Journées d'Études sur la Parole-JEP2022*, 2022.

[3] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[4] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[5] C. Gendrot and M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German." in *Interspeech 2005, Lisbon, Portugal*, 2005, pp. 2453–2456. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-00188096

[6] M. Lancien, "Le rôle de la réduction phonétique dans l'expression de la proximité sociale," Ph.D. dissertation, Université de Lausanne, 2021.

[7] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators," *Statistical Papers*, Nov 2019. [Online]. Available: http://dx.doi.org/10.1007/s00362-019-01148-1

[8] J. Mielke, E. R. Thomas, J. Fruehwald, M. McAuliffe, M. Sonderegger, J. Stuart-Smith, and R. Dodsworth, "Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several english speech corpora," 2019.

[9] K. Evanini, *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. University of Pennsylvania, 2009.

[10] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "Fave (forced alignment and vowel extraction) program suite," *URL http://fave. ling. upenn. edu*, 2011.

[11] M. Sonderegger, J. Stuart-Smith, M. McAuliffe, R. Macdonald, and T. Kendall, "Managing data for integrated speech corpus analysis in speech across dialects of english (spade)," 2022.

[12] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news." in *Interspeech*, 2005, pp. 1149–1152.

[13] Calliope (Firm), J. P. Tubach, and G. Fant, *La parole et son traitement automatique*. Paris ; Milan ; Barcelone : Masson, 1989.

[14] M. Lancien, "Filtering data for phonetics," 2021. [Online]. Available: https://github.com/M-Lancien/Filtering_data_for_phonetics

[15] P. Filzmoser and M. Gschwandtner, "mvoutlier: Multivariate outlier detection based on robust methods," 2018, r package version 2.0.9. [Online]. Available: https://CRAN.R-project.org/package=mvoutlier

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

THE EVERYDAY SPEECH ENVIRONMENTS OF PRESCHOOLERS WITH AND WITHOUT COCHLEAR IMPLANTS

Margaret Cychosz[1], Rochelle S. Newman[2], Benjamin Munson[3], Rachel R. Romeo[2], Jessica Kosie[4], and Jan R. Edwards[2]

[1]University of California, Los Angeles, [2]University of Maryland-College Park, [3]University of Minnesota, [4]Princeton University
mcychosz@umd.edu

ABSTRACT

Children learn speech sounds from the language used around them. Yet children with severe to profound deafness, who receive cochlear implants, have a different learning experience. They must learn speech on a different timescale, owing to the *absence* of auditory input pre-implantation, and the *degraded* speech signal transmitted by the cochlear implant post-implantation. These differences likely shape the everyday speech and language that children with cochlear implants are exposed to. This paper analyzes daylong audio recordings of 3- to 5-year-old children with cochlear implants (16 hrs/child), and their typical hearing peers, as they go about their daily lives to characterize differences in the *quantity*, *consistency*, and experience-related *growth* in their speech-language environments.

Keywords: speech, acoustics, first language acquisition, cochlear implants, language input

1. INTRODUCTION

Children learn the sounds and structure of their native language(s) from the input that they receive from caregivers around them. Even within North American children of the same age, exposed to the same language, there is large variability in the type and quantity of speech-language exposure such as the number of word types [1], ratio of male to female speech input [2], and phonetic realization of consonants and vowels [3, 4]. Crucially, these individual differences in exposure relate to some aspects of children's speech-language development. For example, 6- to 18-month-olds who engaged in more contingent vocal interactions with caregivers vocalized more and knew more words at 18 months [5]. Other work has demonstrated similar links between the phonetics of speech input and children's outcomes: expanded vowel spaces in speech directed to 18-month-olds predicts receptive and expressive vocabulary size at 24 months [6].

There are reasons to believe that the speech-language input directed to children may vary by hearing status, particularly for children with severe to profound hearing loss who receive cochlear implants (CIs). A CI is an auditory prosthesis consisting of an audio processor worn external to the ear and an electrode array inserted into the cochlea that directly stimulates the auditory nerve, partially restoring the sensation of hearing. There are two components of cochlear implantation that may systematically alter child implantees' speech-language environments. First, children often do not receive one or both of their CIs until their first or second birthdays, resulting in a period of AUDITORY ABSENCE pre-implantation. Second, CI electrodes stimulate the cochlea at discrete points, discretizing the speech envelope. This, in addition to issues inherent to the hardware such as electrode interaction and interaural mismatch, results in a DEGRADED SPEECH SIGNAL post-implantation.

Thus, both auditory absence pre-implantation and signal degradation post-implantation characterize the listening experience of children with CIs; these perceptual changes may lead to other changes in the input, shaping these children's everyday speech and language environments. For example, among children with aided hearing loss (hearing aids), those with lower better-ear pure tone averages tend to receive more language input from adult caregivers [7]. And a series of works now demonstrate how these individual differences in linguistic input impact speech-language outcomes among children with CIs in particular [8, 9, 10, 11]. Consequently, the goal of the current study is to evaluate how the listening experiences of children with CIs shapes their everyday the speech and language environments.

2. METHODS

2.1. Participants

Eighteen children with cochlear implants participated in this study. The children were matched by parent-reported gender, socioeconomic status (instantiated as number of years of maternal education), and age to two groups of children with typical hearing (TH): (1) by *chronological* age, to match for cognitive and articulatory development (N=18), and (2) by *hearing* age, to match for auditory experience (N=16 as 2 children with CIs had <1 year of hearing experience). All children were monolingual English speakers and age matching was made within 3 months whenever possible.

The children with CIs had profound deafness in both ears (N=14 bilateral CIs, N=2 unilateral, and N=2 bimodal CI+hearing aid). All children with TH had parent-reported typical speech-language development at the time of participation. See Table 1 for demographic information by hearing group. The average maternal education level was a college degree.

Table 1: Demographic and audiological information. Mean (SD), range. *Includes the 2 children with hearing ages < 12 mos.

	Chrono. age matches	Cochlear implant	Hearing age matches
Chrono. Age (mos)	46.28 (10.8), 32-66	47.72 (9.84), 31-65	35 (12.71), 17-52
Gender (F,M)	9,9	9,9	9,7
Hearing Age (mos)	NA	31.28 (14.3), 8-54*	NA
Implant Age (mos)	NA	16.44 (9.7), 7-45	NA

2.2. Recording procedure

Each child completed one daylong recording where they wore a small, lightweight Language ENvironment Analysis (LENA) recording device (2"x3"; 2 oz.) in a specialized vest for an entire day. Recordings were completed on a typical, non-school day in the child's life. Families were instructed to turn the device on in the morning when the child awoke and continue recording for the duration of the device battery (16 hrs.). One family instead completed a 12.83 hour recording.

2.3. Audio processing

Measures of the children's home speech-language environments were semi-automatically derived from each child's recording using LENA's diarization algorithm which assigns speaker tags, timestamps, and intensity/sound pressure levels to audio clips [12]. To minimize inter-recorder differences, LENA hardware captures intensity in dBC SPL (flat frequency response across the speech range). The algorithm reports intensity measures in dBFS (maximum of 0) so we offset each value by +97dB to facilitate interpretation. All speech clips tagged as "Target child near", "Male adult near", and "Female adult near" were extracted. Word token count estimates from the adult clips were likewise extracted. We filtered out target child clips that contained cries, adult male and female clips that contained any non-speech, and adult male and female clips > 10s (574 clips, 0.07%). Finally, algorithmic estimates of "conversational turns", defined as target child and adult utterances spoken within 5 seconds, were extracted (see Figure 1). This workflow allowed us to assess the quantity, consistency, and experience-related growth of the children's speech-language environments.

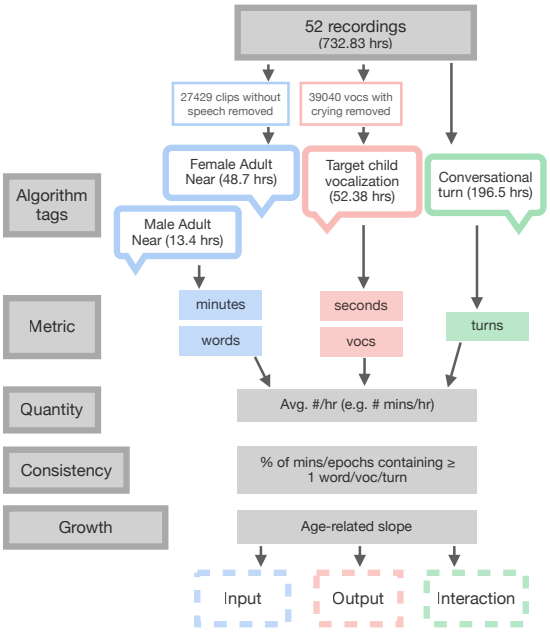


Figure 1: Daylong audio recording processing steps.

There has been substantial work evaluating the LENA system's algorithmic performance [13]. Crucially, our analyses relied on diarization and tags that have relatively high recall and precision for the language and age group studied (e.g., "Female adult near" > 60% for English-learning infants and preschoolers [14]) and not those categories, such as electronic speech, that have poorer reliability.

Nevertheless, as algorithmic performance is not exact for any of the analyzed categories, we interpret our results by comparing *across* hearing groups (as there should be no reason why algorithmic performance would be better or worse by hearing group), and stress that reports of exact amounts of e.g., words or vocalizations per hour, should be interpreted with caution.

3. RESULTS

We divide the results section into three components of each child's daily speech-language experience—caregiver input, target child output (production), and parent-child interaction—and evaluate the impact of hearing group upon each outcome. See tables 2 and 3 for summary statistics of the measures.

Table 2: Measures of the naturalistic speech environment, by hearing group. Mean(SD), range.

	Chrono. age matches	**CI**	**Hearing age matches**
Recording duration (hrs)	15.82(0.75), 12.83-16	16(0), 16-16	16(0), 16-16
Input			
Adult speech intensity	68.21(5.97), 48.98-83.7	68.87(5.81), 45.46-84.03	68.92(6.15), 48.08-87.94
Adult speech/hr (words)	1081.49(481.29), 285.63-2250.39	1217.23(508.87), 411.36-2127.7	1105.54(433.01), 170.88-1630.86
Adult speech/hr (s)	258.36(118.69), 72.79-533.67	288.52(121.52), 95.31-499.53	264.55(102.57), 43.67-386.22
Adult word consistency	0.52(0.13), 0.31-0.78	0.58(0.11), 0.25-0.7	0.51(0.12), 0.19-0.66
Output			
Voc. intensity	76.78(4.39), 47.51-84.77	76.95(4.35), 43.16-85.79	77.15(4.95), 44.79-90.31
Child voc. quantity	308.03(142.81), 90.12-575.81	271.75(69.23), 48.75-381.62	254.5(108.83), 42.5-424
Voc. duration (ms)	1004.46(662.3), 80-10940	937.93(569.76), 80-13270	966.59(627.6), 80-19730
Child voc. consistency	0.55(0.15), 0.34-0.84	0.58(0.13), 0.17-0.72	0.49(0.14), 0.22-0.69
Interaction			
Convo. turn quantity	61.71(32.78), 20.69-150.94	68.17(26.47), 8.5-116.75	65.13(25.47), 11.12-92.62
Convo. turn consistency	0.58(0.14), 0.38-0.84	0.64(0.13), 0.22-0.77	0.56(0.12), 0.36-0.74

Data were analyzed in the RStudio computing environment (R version 4.2.1; [15]). Visualizations were made using `ggplot2` [16] and modeling was conducted using `lme4` [17] and `lmerTest` packages [18]. All model fitting began with a baseline, random-effects only model. Model fit improvements were evaluated by comparing model log-likelihood values and AIC estimates. The predictor Hearing Group (3-levels: CI, chronological age matches, hearing age matches) was contrast-coded with 'CI' as the reference level so model coefficients for the chronological and hearing age match groups refer to deviance from the CI group unless noted otherwise.

3.1. Input

Children's speech-language input was quantified as the average number of minutes/hour containing speech from an adult female or male near the child, as well as the average number of words spoken by an adult near the child/hour. Throughout the results we will normalize by hour to account for time-of-day differences across recordings, as well as different recording lengths.

For repeated measures, such as speech intensity, we fit linear mixed effects models with random intercepts by child and a fixed effect of hearing group. Models of hourly measures (words, minutes) additionally included random intercepts by hour of recording. There were no reliable differences by hearing group for speech input intensity, or measures of input quantity (hourly words, hourly minutes of speech; log-likelihood tests all $p > .05$); thus, speech input was produced at a similar intensity and quantity (minutes and words) regardless of the child's hearing experience.

We next evaluated the consistency of speech input by hearing group, or the percentage of minutes in each recording containing ≥ 1 word from an adult. There were no differences in speech input consistency by hearing group ($p > .05$), although speech input did become more consistent with age (coded continuously, in months) across the sample, independent of hearing status (model fit: β=0.004, t=3.16, p=.003) meaning that as children aged, speech was more continuously present throughout their day.

Finally, we evaluated differences by hearing group in age-related growth of speech input. For this analysis, we modeled the effect of age (in mos) upon hourly adult word token count in the children's environments and compared the values across all three hearing groups (CI, chronological age matches, hearing age matches). For the children with CIs, we modeled both their growth by *chronological* age as well as *hearing* age (time since implantation): hourly word counts only increased in the TH groups (Table 3).

3.2. Output

Each child's speech production was quantified as the average number of vocalizations/hour from

Table 3: Relationship between age (mos) and measures of the naturalistic speech environment, by hearing group. β=model coefficient, p-value from model parameter (***$p \leq .001$, **$p \leq .01$, *$p \leq .05$), +$p \leq .1$, r=Pearson correlation coefficient. No p-value annotation indicates $p \geq .1$.

	Chrono. age matches	CI chrono. age	CI hearing age	Hearing age matches
Adult words	β=20.7+ r=0.46	β=3.81 r=0.07	β=0.77 r=0.02	β=16.49+ r=0.48
Child voc. quantity	β=3.98 r=0.34	β=2.71 r=0.33	β=1.06 r=0.17	β=8.03*** r=0.89
Child voc. duration (ms)	β=.59 r=0.02	β=3.16* r=0.05	β=1.84 r=0.03	β=6.59** r=0.11
Convo. turn	β=0.83 r=0.29	β=0.01 r=0	β=-0.33 r=-0.16	β=2.18*** r=0.93

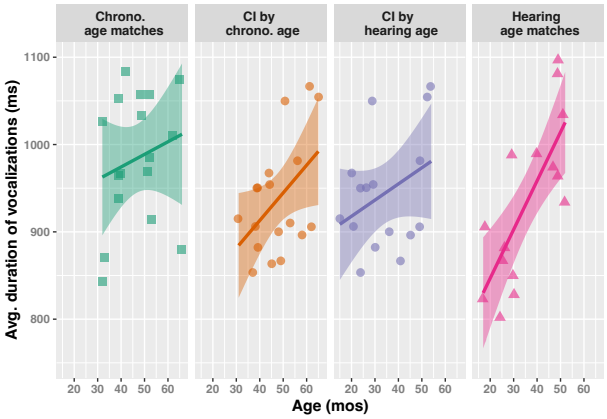


Figure 2: Age-related growth in child vocalization duration, by hearing group.

the target child and the average duration of each vocalization. There was no effect of hearing status on the number of vocalizations/hour or speech output intensity (both tests p >.05); so, hearing status did not dictate the amount or intensity of the children's speech. However, there was an effect of hearing status in the model predicting vocalization duration (χ^2=6.95, df=2, p=.03): the chronological age matches produced significantly longer vocalizations than the children with CIs (β=59.25) and hearing age matches (β=78.71).

We additionally measured consistency of speech output as the percentage of minutes in each recording containing at least one vocalization from the target child; there was no effect of hearing experience upon output consistency. And finally, we measured the age-related growth in vocalization quantity and duration: there was a significant, positive effect of age on vocalization duration among the children with CIs by chronological age and for the hearing age matches (Figure 2).

3.3. Interaction

Finally, we compared caregiver-target child interactions. We instantiated quantity of interaction as the average number of back-and-forth conversational turns per hour and the consistency of interaction as the percentage of 5-minute epochs containing at least 1 conversational turn. There was no effect of hearing group upon the quantity or consistency of turns (both p > .05). The age-related growth analysis showed increases in conversational turns for the hearing age matches.

4. DISCUSSION

This paper analyzed naturalistic daylong audio recordings from preschoolers to evaluate how the daily speech-language environment varied by hearing experience. Broadly, results show minimal differences by hearing status in caregiver-child interaction and speech input, even when we instantiated these measures in different manners (e.g., quantity, consistency, growth).

Differences by hearing group instead emerged in the children's own vocal productions: vocalization duration increases by chronological and hearing age among the children with CIs. Going forward it will be important to evaluate how different parameters of vocal production emerge in the speech of children with CIs (and how this compares to children with TH); for example, the rate of canonical babble, fluidity of consonant-vowel transitions, and number of syllables per utterance are all important indices of speech-language development. And these early comparative results suggest that they may exhibit substantial differences by hearing experience.

5. CONCLUSION

Children with CIs have a different early speech-language experience owing to the absence of auditory information pre-implantation and the degraded auditory signal conveyed by the implant post-implantation. Nevertheless, using hundreds of hours of audio capturing children's naturalistic speech-language environments, we were able to show that children with CIs and TH have fairly similar early speech-language experiences. The dynamics of the children's own vocal production, however, did vary by auditory experience, suggesting that future work should focus on the children's own speech dynamics.

6. ACKNOWLEDGEMENTS

The authors wish to thank the children and their families who participated in this research. Additional thanks to Julianna Gross for assistance with data collection and processing. All audio processing scripts and markdowns to replicate these results are included in the project's Github repository: https://github.com/megseekosh/everyday_CI.

7. REFERENCES

[1] M. L. Rowe, "A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development: Child-directed speech and vocabulary," *Child Development*, vol. 83, no. 5, pp. 1762–1774, Sep. 2012.

[2] E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni, "What do North American babies hear?: A large-scale cross-corpus analysis," *Developmental Science*, vol. 22, no. 1, p. e12724, 2019.

[3] L. C. Dilley, A. L. Millett, J. D. Mcauley, and T. R. Bergeson, "Phonetic variation in consonants in infant-directed and adult-directed speech: the case of regressive place assimilation in word-final alveolar stops," *Journal of Child Language*, vol. 41, no. 1, pp. 155–175, Jan. 2014.

[4] M. Kalashnikova and D. Burnham, "Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions," *Journal of Child Language*, vol. 45, no. 5, pp. 1035–1053, Sep. 2018.

[5] N. Ferjan Ramirez, S. R. Lytle, and P. K. Kuhl, "Parent coaching increases conversational turns and advances infant language development," *Proceedings of the National Academy of Sciences*, vol. 117, no. 7, pp. 3484–3491, Feb. 2020.

[6] K. M. Hartman, N. B. Ratner, and R. S. Newman, "Infant-directed speech (IDS) vowel clarity and child language outcomes," *Journal of Child Language*, vol. 44, no. 5, pp. 1140–1162, Sep. 2017.

[7] M. VanDam, S. E. Ambrose, and M. P. Moeller, "Quantity of parental language in the home environments of hard-of-hearing 2-year-olds," *Journal of Deaf Studies and Deaf Education*, vol. 17, no. 4, pp. 402–420, Oct. 2012.

[8] L. C. Dilley, E. A. Wieland, M. Lehet, M. K. Arjmandi, D. Houston, and T. R. Bergeson, "Quality and quantity of infant-directed speech by maternal caregivers predicts later speech-language outcomes in children with cochlear implants," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1822–1822, Mar. 2018.

[9] L. C. Dilley, M. Lehet, E. A. Wieland, M. K. Arjmandi, M. Kondaurova, Y. Wang, J. Reed, M. Svirsky, D. Houston, and T. R. Bergeson, "Individual differences in mothers' spontaneous infant-directed speech predict language attainment in children with cochlear implants," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 7, pp. 2453–2467, Jul. 2020.

[10] M. Arjmandi, D. Houston, Y. Wang, and L. C. Dilley, "Estimating the reduced benefit of infant-directed speech in cochlear implant-related speech processing," *Neuroscience Research*, p. S0168010221000213, Jan. 2021.

[11] J. L. DesJardin and L. S. Eisenberg, "Maternal contributions: Supporting language development in young children with cochlear implants," *Ear and Hearing*, vol. 28, no. 4, pp. 456–469, Aug. 2007.

[12] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA Language Environment Analysis System in young children's natural home environment," LENA Research Foundation, Boulder, CO, Technical Report lTR-05-2, 2009.

[13] M. Lehet, M. K. Arjmandi, L. C. Dilley, and D. Houston, "Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system." *Behavior Research Methods*, vol. 53, no. 1, pp. 113–138, 2020.

[14] A. Cristia, F. Bulgarelli, and E. Bergelson, "Accuracy of the Language Environment Analysis System segmentation and metrics: A systematic review," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 4, pp. 1093–1105, 2020.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[16] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag New York, 2016.

[17] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] A. Kuznetsova, P. Brockhoff, and R. Christensen, "lmerTest Package: Tests in linear mixed-effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

TURNING PODCASTS INTO A TRAINING CORPUS FOR CONVERSATIONAL TEXT-TO-SPEECH SYNTHESIS

Heete Sahkai, Liisi Piits, Liis Ermus, Indrek Hein, Meelis Mihkla, Indrek Kiissel, Kristjan Suluste, Egert Männisalu, Rene Altrov, Hille Pajupuu, Jaan Pajupuu

Institute of the Estonian Language
heete.sahkai@eki.ee

ABSTRACT

Text-to-speech synthesis (TTS) is increasingly used in applications that require a conversational speaking style, such as voice-enabled chatbots. This gives rise to the need to develop standard and affordable solutions for obtaining conversational training data. This paper describes the creation of a three-hour TTS training corpus for Estonian, using found podcast data, existing automatic speech recognition and transcription-editing software, and a simplified transcription-editing protocol. The corpus was evaluated in comparison with a corpus of read-aloud sentences recorded by the host of the podcast. The evaluation results showed that conversational utterances synthesised with the voices based on the podcast corpus were indeed perceived as representing a more spontaneous speaking style than the utterances synthesised with the voices based on the read-aloud data. Perceived spontaneity was increased by the presence of filled pauses and disfluencies in the stimulus utterances.

Keywords: TTS, spontaneous speech corpus, conversational speaking style, podcasts, Estonian

1. INTRODUCTION

Synthetic voices are increasingly used in applications that require a conversational speaking style, such as voice-enabled chatbots. This gives rise to the need for conversational training data. Obtaining phonetically representative conversational data from a single speaker with sufficient recording quality presents several challenges: what sources can provide such data; how to segment the data into utterances; how to treat the characteristic features of spontaneous speech like filled pauses, backchannels, disfluencies, reduced pronunciation, laughter, etc.

In previous attempts to train conversational synthetic voices, different types of training data have been used: scripted conversational lines or dialogues performed in a studio [1], [2], studio-recorded unscripted dialogues [2]–[4] and monologues [5], [6], existing podcasts [7], [8]. While [3] found that utterances synthesised with voices trained on spontaneous data were perceived as more conversational only when they contained discourse markers and filled pauses, [7] and [8] found that voices based on spontaneous data were generally perceived as more appropriate and spontaneous. Conversational data has also been used in combination with read-aloud data [3] or models based on such data [4], [7]–[9]. This has been found to improve the quality of synthetic speech while retaining its conversational character [3], [7].

The segmentation of unscripted conversational data into utterances has been done manually [3] or using a breath detection method [7], [9].

Regarding the treatment of disfluencies in the training data, the options are to exclude the utterances containing them, or to include them either without annotation or with a more or less detailed annotation [10]. When disfluent utterances are included, a selection must still be operated; for example, [3] retained filled pauses, backchannels, and discourse markers, but excluded utterances with imitations, word fragments, mispronunciations, heavily reduced pronunciations, mumbling and laughter. [7] and [10] found that a voice trained only on fluent utterances was perceived as equally conversational but had better quality than voices trained on disfluent data.

A further question is whether features like filled pauses should be added to input text at the time of synthesis. [3] found that utterances with and without fillers were perceived as equally conversational, while [10] found that filled pauses made synthetic speech sound more authentic.

The goal of the present study is to move towards developing an optimal and affordable workflow for obtaining spontaneous training data. A second goal is to test whether voices based on such data are perceived as having a more spontaneous speaking style than those based on read data, and whether the perception of spontaneity depends on the presence of disfluencies in the input sentences, as the previous evaluation results are mixed in these regards.

We describe the compilation of a spontaneous Estonian TTS corpus. We used found data (podcasts) as an affordable solution, as recording a corpus of conversational data in a studio is costly and time and effort consuming (Section 2). The data was

transcribed, edited, annotated, and segmented into utterances using existing automatic speech recognition (ASR) and transcription-editing tools (Section 3). The corpus was evaluated in comparison with a corpus of read-aloud sentences recorded by the host of the podcast (Section 4). We conclude that the resulting corpus does provide suitable training data for conversational style TTS (Section 5).

2. THE DATA

We used a podcast series as a source of conversational speech. The series was chosen from among a set of candidate podcasts and radio shows, based on the following criteria: most of the episodes were recorded by the same speakers, allowing to obtain more material per speaker; there was no background music or noise; the podcast was recorded by a man and a woman, allowing a more reliable speaker recognition, as the ASR system was not able to diarise several male or female speakers in the same recording; there were at most two speakers (ASR makes more diarisation mistakes when there are several speakers). The episodes were obtained through a free public portal for podcasts and radio shows (podcast.ee). We used 11 episodes with a duration of 8.68 hours.

The podcast series is titled "Intimately about private life"[1]. All the episodes are hosted by a married couple, who are not professional radio hosts. The podcast is unscripted and represents an informal spoken register. The topic of the podcast are relationship issues, which are discussed very openly.

An agreement was signed with both presenters, who gave the permission to use their voices for synthesis and for compiling a publicly available corpus.

3. PROCESSING THE MATERIAL

The sound files were transcribed using the Tallinn University of Technology Estonian ASR system[2] [11], [12]. The system performs speaker diarisation, speech-to-text, punctuation restoration to correspond to the written language, text normalisation (reduced forms are corrected). The system omits filled pauses and disfluencies, like interrupted words, from the output text. The system returns the transcription in several formats, from which trs and json were used.

The transcriptions were checked and corrected. The processing started by dividing the transcriptions into shorter chunks in one speaker's text according to the time codes in the json file, using a dedicated script. Then the transcription in the trs-files was checked manually, using the program Transcriber[3]. Recognition errors were corrected. Punctuation was matched with the structure of the speech (commas marked short pauses in an utterance, periods marked utterance boundaries).

A protocol was established for the annotation of two types of filled pauses (vowel- and consonant-based), backchannels, disfluencies, and laughter. The purpose of the annotation was to gain control over these phenomena (cf. [10]), and to test their influence on the perception of spontaneity.

Utterances containing overlapping speech, laughter, mumbling, and imitations were omitted.

It took about one hour to work through 10 minutes of the transcription.

After the correction and annotation, a second verification and the final selection of utterances was made by another annotator, using an internally developed web-interface. Finally, the trs and sound file were spliced into correspondent one-utterance files (txt and wav). The process resulted in two TTS corpora, one for each speaker. The corpus used for evaluation consisted of 2269 utterances by the female speaker with a duration of 3.69 hours. The corpus will be referred to as the SPON corpus.

4. EVALUATION

The SPON corpus was evaluated in order to answer the following questions: 1) Do listeners perceive a voice based on spontaneous training data as having a more spontaneous speaking style than a voice based on read-aloud data? 2) Does the presence of features that are unique to spontaneous speech affect the perception of spontaneity? 3) Are the evaluation results stable across different synthesis techniques? 4) Is the material sufficient and suitable for different synthesis techniques?

For the purpose of evaluation, a control corpus of read speech was first recorded by the same speaker (4.1). Next, three synthesis methods were used to train a synthetic voice on each corpus (4.2). Then input texts were selected and stimuli were synthesised with the different voices (4.3). The stimuli were then subjected to evaluation (4.4). The evaluation results are presented in Section 4.5.

4.1. Creation of the control corpus

In order to evaluate the SPON corpus against the baseline of read-aloud data a control corpus (READ corpus) was first recorded by the female host of the podcast. The corpus consisted of 1192 phonetically representative read-aloud sentences and had a total duration of 2.06 hours.

4.2. Training of the synthetic voices

Based on each of the two corpora, SPON and READ, three synthetic voices were trained, using different techniques (S1, S2, S3):

S1 uses HTS 2.0, a statistical-parametric TTS technique based on hidden Markov models [13]. The grapheme-based technique uses the language independent text processing libraries of the Ossian-TTS[4] as the front end of the HTS system [14];

S2 uses Merlin, a Neural Network based TTS [15]. The system relies on the Theano numerical computation library. To convert text into full-context labels, an internally developed front-end text processor was used [16];

S3 uses TransformerTTS[5], a TTS solution using a transformer-based neural network model. Spectrograms generated by the model were converted into waveforms using a HiFiGAN vocoder pre-trained on the LJSpeech dataset.

We failed to achieve a satisfying result with S3 on the SPON corpus, thus in total five voices were trained: S1READ, S1SPON, S2READ, S2SPON, and S3READ. S3READ was included in the comparison as a baseline as voices trained with S3 using read-aloud data have received the highest scores in previous evaluations [17].

4.3. The stimuli

Spontaneous speech synthesis should be evaluated using transcriptions of spontaneous speech (see [5]). We used the podcast that served to compile the SPON corpus and selected four 150–200 character passages from the episodes that were not included in the corpus. Each passage was used to create two input texts: 1) an exact transcription of the passage, with filled pauses, repetitions, self-repairs, filler words, etc., to be referred to as 'natural' text, and 2) an edited version of the passage, with disfluencies removed, referred to as 'clean' text (see Datasets[6]). Each of the eight input texts was synthesised with the five synthetic voices, giving 40 stimuli.

4.4. Testing

The listeners were asked to evaluate each stimulus for how spontaneous it sounds on a 7-point scale, where 1 = not spontaneous at all … 7 = spontaneous. The listeners were 10 women (aged 37–56, M = 43.0 years, SD = 6.0) and 8 men (aged 34–55, M = 43.3 years, SD = 6.8). All scores for each listener were normalised using the formula

(1) $y=(x-X)/s$,

where x is the score, X is the mean of the listener's scores, and s is the standard deviation of the listener's scores. We classified the performances with scores above zero as spontaneous, and those with scores below zero as not spontaneous.

To find out the degree of agreement among the listeners (inter-rater reliability), the intra-class correlation coefficient (ICC2k) was calculated using the 'psych' package in R [18]. A Welch Two Sample t-test was used to determine whether the average spontaneity scores differ significantly depending on the training corpus, the synthesis technique, and the type of input text [19].

4.5. Evaluation results

A good to excellent degree of agreement was found between the listeners' ratings. The average measure ICC2k was 0.90 with a 95% confidence interval from 0.85 to 0.94 $F(39, 663) = 9.8$, $p < 0.0001$).

The results of the t-test revealed that the voices trained on the SPON corpus were significantly more spontaneous than the voices trained on the READ corpus: $M_{S1SPON} = 0.01$ vs. $M_{S1READ} = -0.76$, $t(267) = -7.46$, $p < 0.0001$; $M_{S2SPON} = 0.30$ vs. $M_{S2READ} = -0.13$, $t(286) = -4.18$, $p < 0.0001$ (see Fig. 1).

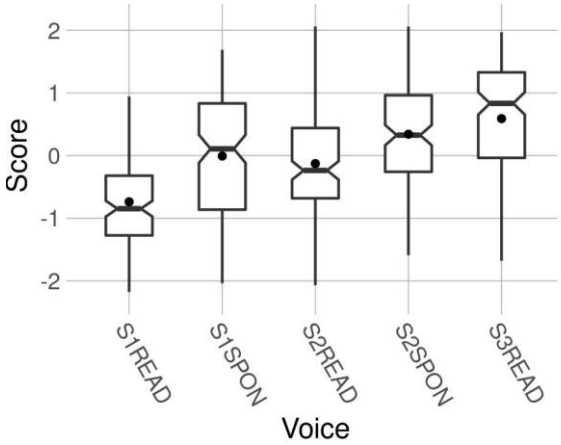


Figure 1: The spontaneity scores by synthetic voices (S1, S2, S3 – synthesis techniques; READ – trained on the read-aloud corpus, SPON – trained on the spontaneous corpus).

Taking into account the synthesis technique, the training corpus, and the type of input text, listeners gave the following spontaneity scores from the lowest to the highest: S1READ_natural -1.01; S2READ_natural -0.50; S1READ_clean -0.50; S1SPON_clean -0.12; S2SPON_clean 0.10; S1SPON_natural 0.14; S2READ_clean 0.24; S3READ_natural 0.51; S2SPON_natural 0.51; S3READ_clean 0.62 (see Fig. 2).

The three voices trained on the READ corpus received higher scores for the clean than for the natural input texts (significant difference for S1 and S2, no significant difference for S3): $M_{S1READ\_clean} = -0.50$ vs. $M_{S1READ\_natural} = -1.01$, $t(141) = 4.40$, $p < 0.0001$; $M_{S2READ\_clean} = 0.24$ vs. $M_{S2READ\_natural} = -0.50$,

$t(138) = 5.59$, $p < 0.0001$; $M_{S3READ\_clean} = 0.63$ vs. $M_{S3READ\_natural} = 0.51$, $t(142) = 0.73$, $p = 0.465$.

The two voices trained on the SPON corpus received higher scores for the natural input texts (significant difference in the case of S2): $M_{S1SPON\_clean} = -0.12$ vs. $M_{S1SPON\_natural} = 0.14$, $t(140) = -1.63$, $p = 0.106$; $M_{S2SPON\_clean} = 0.10$ vs. $M_{S2SPON\_natural} = 0.51$, $t(137) = -2.94$, $p = 0.004$.

The scores of S1SPON and S2SPON did not differ significantly in case of the clean input text: $M_{S1SPON\_clean} = -0.12$ vs. $M_{S2SPON\_clean} = 0.10$, $t(132) = -1.45$, $p = 0.151$. By contrast, S2SPON got a significantly higher spontaneity score for natural texts: $M_{S1SPON\_natural} = 0.14$ vs. $M_{S2SPON\_natural} = 0.51$, $t(142) = -2.42$, $p = 0.016$.

The performances of both natural and clean texts by S3READ received equal scores to the natural texts performed by S2SPON: $M_{S2SPON\_natural} = 0.51$ vs. $M_{S3READ\_natural} = 0.51$, $t(142) = 0.01$, $p = 0.995$; $M_{S2SPON\_natural} = 0.51$ vs. $M_{S3READ\_clean} = 0.63$, $t(142) = -0.71$, $p = 0.473$.

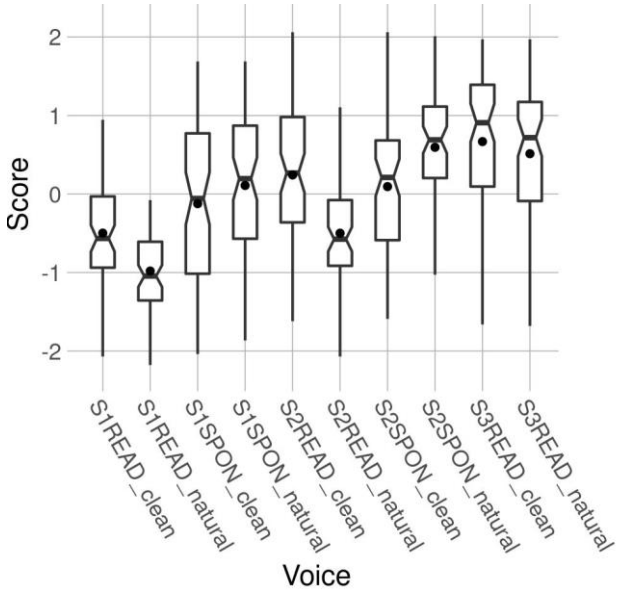


Figure 2: Spontaneity scores by the synthetic voices and the type of input text (S1, S2, S3 – synthesis techniques; READ – trained on the read-aloud corpus, SPON – trained on the spontaneous corpus; natural – input texts with fillers and disfluencies, clean – input texts without fillers and disfluencies).

5. DISCUSSION AND CONCLUSION

Both voices trained on the SPON corpus were perceived as significantly more spontaneous than the voices trained on the READ corpus using the same techniques (Fig. 1). The presence of fillers and disfluencies in the input text increased the perceived spontaneity of the voices trained on the SPON corpus, confirming the results of [10]. For the voices trained on the READ corpus, the effect was reversed: fillers and disfluencies in the input text decreased the perceived spontaneity (Fig. 2). Our results thus support the use of spontaneous training data as well as the insertion of fillers and disfluencies into synthetic speech by one of the methods in [10].

As to the performance of the different synthesis techniques, the SPON corpus was suitable for creating synthetic voices with two of the techniques, S1 and S2, while S3 did not yield results. Due to the characteristics of spontaneous speech (reduced articulation, fillers, disfluencies), this type of synthesis technique might require a larger training corpus (cf. [1]), or a combination of spontaneous and read data (see [3], [7]). From the two successful techniques, S2 received significantly higher scores for the natural input texts than S1. Equal scores were received by the baseline voice trained with S3 on the READ corpus (Fig. 2).

Regarding the methods for obtaining spontaneous training data, the results suggest that a relatively small amount of spontaneous speech (three hours) allows to train synthetic voices with certain methods. Previous attempts have been based on larger corpora, cf. [1], [7], [8].

The most time-consuming task was the manual correction and annotation of the data. This could be alleviated by developing an ASR tool for an accurate transcription of spontaneous speech, including filled pauses, disfluencies, etc. Another possible future avenue is the improvement of the annotation protocol based on the analysis of the synthesis results.

6. ACKNOWLEDGEMENTS

This work was financed by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies), by the National Programme for Estonian Language Technology 2018–2027, by the Estonian Research Council (project EAG144 Spontaneous speech synthesis), and by the basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research (project EKI-BAAS-2023/2 Speech studies at the service of speech technology). We wish to thank our annotators and informants.

7. REFERENCES

[1] Zandie, R., Mahoor, M. H., Madsen, J., Emamian, E. S. 2021. RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. *Interspeech 2021*, 2751–2755. https://doi.org/10.21437/Interspeech.2021-341

[2] Adigwe, A. O., Klabbers, E. 2022. Strategies for developing a Conversational Speech Dataset for Text-To-Speech Synthesis. *Interspeech 2022*, 2318–2322. https://doi.org/10.21437/Interspeech.2022-10802

[3] Andersson, S., Yamagishi, J., Clark, R. A. J. 2010. Utilising spontaneous conversational speech in HMM-based speech synthesis. *Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7)*, 173–178.

[4] Huang, Y.-C., Wu, C.-H., Chen, Y.-Y., Shie, M.-G., Wang, J.-F. 2017. Personalized Spontaneous Speech Synthesis Using a Small-Sized Unsegmented Semispontaneous Speech. *IEEE/ACM Trans. Audio Speech Lang. Process*, *25*(5), 1048–1060. https://doi.org/10.1109/TASLP.2017.2679603

[5] Székely, E., Edlund, J., Gustafson, J. 2020. Augmented Prompt Selection for Evaluation of Spontaneous Speech Synthesis. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6368–6374. https://aclanthology.org/2020.lrec-1.782

[6] Lameris, H., Mehta, S., Henter, G. E., Gustafson, J., Székely, É. 2022. Prosody-controllable spontaneous TTS with neural HMMs (arXiv:2211.13533). arXiv. http://arxiv.org/abs/2211.13533

[7] Székely, É., Henter, G. E., Beskow, J., Gustafson, J. 2019. Spontaneous Conversational Speech Synthesis from Found Data. *Interspeech 2019*, 4435–4439. https://doi.org/10.21437/Interspeech.2019-2836

[8] Yan, Y., Tan, X., Li, B., Zhang, G., Qin, T., Zhao, S., Shen, Y., Zhang, W.-Q., Liu, T.-Y. 2021. Adaptive Text to Speech for Spontaneous Style. *Interspeech 2021*, 4668–4672. https://doi.org/10.21437/Interspeech.2021-584

[9] Székely, É., Henter, G. E., Gustafson, J. 2019. Casting to Corpus: Segmenting and Selecting Spontaneous Dialogue for Tts with a Cnn-lstm Speaker-dependent Breath Detector. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6925–6929. https://doi.org/10.1109/ICASSP.2019.8683846

[10] Székely, É., Eje Henter, G., Beskow, J., Gustafson, J. 2019. How to train your fillers: Uh and um in spontaneous speech synthesis. *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 245–250. https://doi.org/10.21437/SSW.2019-44

[11] Alumäe, T., Tilk, O., Asadullah. 2018. Advanced Rich Transcription System for Estonian Speech. *Human Language Technologies – The Baltic Perspective*, 1–8. https://doi.org/10.3233/978-1-61499-912-6-1

[12] Olev, A., Alumäe, T. 2022. Estonian Speech Recognition and Transcription Editing Service. *Baltic Journal of Modern Computing*, *10*(3). https://doi.org/10.22364/bjmc.2022.10.3.14

[13] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K. 2007. The HMM-based Speech Synthesis System (HTS) Version 2.0. *SSW6-2007*, 294–299

[14] Vainio, M., Grönroos, S.-A., Smit, P., Suni, A., Watts, O. 2014. *Deliverable D2.2. Description of the final version of the new front-end*. https://simple4all.org/wp-content/uploads/2014/11/Simple4All_deliverable_D2.2.pdf

[15] Wu, Z., Watts, O., King, S. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 202–207. https://doi.org/10.21437/SSW.2016-33

[16] Kiissel, I. 2022. *Merlinil põhinev eesti keele kõnesüntesaator [Merlin based Estonian speech synthesizer]*. https://github.com/ikiissel/mrln_et

[17] Piits, L., Pajupuu, H., Sahkai, H., Altrov, R., Ermus, L., Tamuri, K., Hein, I., Mihkla, M., Kiissel, I., Männisalu, E., Suluste, K., Pajupuu, J. 2022. Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices. *Proceedings of the Language Resources and Evaluation Conference*, 1047–1053. https://aclanthology.org/2022.lrec-1.112

[18] Revelle, W. 2022. *psych: Procedures for Psychological, Psychometric, and Personality Research* (R package version 2.2.9). Northwestern University. https://CRAN.R-project.org/package=psych

[19] R Core team. 2022. *R: A Language and Environment for Statistical Computing* (4.2.2). R Foundation for Statistical Computing. https://www.R-project.org/

[1] https://podcast.ee/show/intiimselt-eraelust/

[2] http://bark.phon.ioc.ee/webtrans/

[3] https://sourceforge.net/projects/trans/files/transcriber/

[4] https://github.com/CSTR-Edinburgh/Ossian

[5] https://github.com/TartuNLP/TransformerTTS

[6] The datasets generated and analysed during the study: https://figshare.com/projects/Turning_podcasts_into_a_training_corpus_for_conversational_text-to-speech_synthesis/156248

USING MAHALANOBIS DISTANCE TO FILTER ERRONEOUS VOWEL FEATURES IN LESS-RESOURCED LANGUAGES : APPLICATION TO QUEBEC FRENCH

Lancien Mélanie, Stuart-Smith Jane, Adda-Decker Martine

FNSRS, LPP, LISN, GULP
melanie.lancien@unil.ch, Jane.Stuart-Smith@glasgow.ac.uk, madda@limsi.fr

ABSTRACT

A major challenge in building shareable datasets for phonetic studies consists of maximising data collections while minimising the errors involved, in particular when automatic processes come into play (e.g. labeling, formant detection...). Automatically extracting formants from large amounts of speech frequently produces artifacts, due to formant jumps, alignment problems, or noise. One solution is to use formant range filters. However this requires prior knowledge about the vowels, such as formant range information. Here we propose to use the Mahalanobis distance to remove erroneous values relying only on the labeled speech data. Our study is conducted on a Quebec French corpus including more than 170 k tokens of 16 vowel types. Results show that the proposed method can complement the threshold-based filter approach. Furthermore, it can be used autonomously for undocumented languages to eliminate erroneous values. The approach also makes it easy to adjust the degree of filtering.

Keywords: Vowels, Quebec French, Data Science, Automatic filtering, Corpus Phonetics

1. INTRODUCTION

In the context of automatically annotated medium to large size corpora, the data are generally numerous, heterogeneous, and noisy. Different kinds of errors arise from automated processes, such as formant-tracking jumps, transcription issues, or mismatches between the expected and the actual pronunciation of a word, among others. The said errors then lead to the presence of erroneous datapoints in the datasets. On very large datasets, erroneous values might not affect the results, but with medium-size datasets(e.g. with 4 or 5 tokens of each vowel type per speaker), more common in phonetic studies, erroneous values will likely bias the subsequent results. Thus medium-sized datasets need "cleaning" as far as possible.

Several methods can be used to identify erroneous values during or after the value extraction process (see [1] for methods applied during the formant extraction; see [2] for methods applied during the automatic alignment). Here we address the issue of "post processing" methods such as formant range filters (e.g.[3]). Formant range filters are the most common and work as follow :

1. the user sets up a series of value ranges between which formants must fall in order to correspond to the known/expected vowel profile
2. the filter go through and check the formant values extracted from the audio corpus
3. vowels with formant values oustside the ranges are discarded

The issue with those methods is that 1) there must be literature on those values for the given language; 2) ranges are generally set in a wide and arbitrary fashion (e.g. [3] use ranges from 1500 to 2500Hz for French /i/ F2). However it does efficiently identify the most obvious erroneous values.

Here we present a tool for filtering datasets from low resource language, with no value ranges or vowel profiles available. Quite counter-intuitively, Quebec French (QF) can be considered as low resource when it comes to phonetic description or NLP. But its closeness to France French allow us to compare our results to some expected vowel profiles. Thus we test our tool on the QF dataset from [4].

1.1. Quebec French vowels

Quebec French vowel system includ 16 phonological vowels [5, 6] : /i, y, e, ɛ, ɜ, ø, oe, a, ɑ, ɔ, o, u/ and the 4 nasals /ẽ, ɛ̃, õe, ɑ̃/. These vowels undergo a large amount of phonological processes, namely laxing, diphtonguization, devoicing, syncope, according to specific phonological rules (described by [5, 6, 7] among others).

Most vowel qualities resulting from these phenomena can be grouped according to the notion

of vowel class [8, 9]. A vowel class represents a phoneme and its position in a syllable [10], for instance iR is the class representing phoneme /i/ followed by a lengthening consonant. It is a very important concept in QF for vowel position in the word or syllable strongly affect the allophone used.

Table 1 gives the inventory of the classes relevant for our study. Vowels in class _C they are followed by a coda (any consonant), vowels in _K are followed by a non lengthening consonant coda, those in class _R are followed by a lengthening coda (/ʁ,v,z,ʒ/), finally in class _# vowels are in absolute final. Class _# usually don't trigger any process. In class _K high vowels are laxed (e.g. "bicycle" /bisik/ [bisɪk] *bicycle*). Class _R usually triggers lengthening and diphtonguization for non-high vowels (e.g. "père" /pɛχ/ [paɛχ] *father*). Class _C is only relevant for the /A/s and can trigger dipthonguization for /ɑ/ or posteriorisation for /a/ (e.g. "pâte" /pɑt/ [paɔt] *pasta* - "là" /la/ [lo] *here*).

Phoneme	Vowel Classes
i	i# iR iK
y	y# yR yK
u	u# uR uK
e	e#
ɛ	ɛ# ɛK ɛR
ø	ø#
oe	oeK oeR
o	o# oK oR
ɔ	ɔ# ɔK ɔR
a	a# aC
ɑ	ɑ# ɑC

Table 1: Vowel classes inventory, inspired from [9] and corresponding to [4]'s corpus.

1.2. Mahalanobis distance as a filtering method

In [11], authors showed that Mahalanobis distance can be used as a tool for filtering. The Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_p)^T$ to a set of mean value vectors $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$ and having a covariance matrix Σ is defined as follows:

$$(1) \quad Dx = (x-\mu)^T \Sigma^{-1} (x-\mu).$$

The square root of $Dx = (x-\mu)^T \Sigma^{-1} (x-\mu)$ gives the number of standard deviations between the observation and the mean of the distribution. In the case where $P = \mu D$, the distance is 0, this distance increases as P moves away from the mean in a determined space. Thus by choosing a threshold for the distance ([4, 11] chose 3 standard deviations), it is easy to discard any vowel token further then X standard deviation from the computed profile.

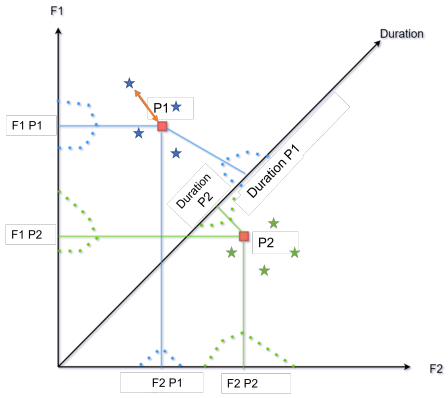


Figure 1: Schematic explanation of the Mahalanobis distance. Blue lines represent the distance of F1 (axis y) F2 (axis x) and duration (axis z) between the mean (orange squares) of two vowels (P1 and P2). Stars are tokens of P1 and P2 and the orange arrow indicate the distance measured.

2. METHODS

2.1. Quebec French data

The Quebec French dataset that we use here is part of [4]'s Ph.D work. They recorded 10 couples (10♂, 10♀), performing five tasks involving the pictures from the Diapix task [12, 13]: 1) reading, 2) image difference identification task alone, with a spouse, with an unknown investigator from the same region and finally with a stranger from France. A recording session lasted approximately 3h and the total corpus duration is 27.22h with a minimum of 1.19h of speech per speaker.

The recordings were all segmented in Speech Units and orthographically transcribed by hand. Author then used the SPPAS software (v. 2.7, adapted for QF phonetics and phonology; see [14]) for automatic segmentation and alignment (for more details see [4]). A total of 162 543 vowel tokens (belonging the 16 vowel classes introduced above) and their acoustic features were extracted from these recordings. Here we only use duration (s), mean F1 (Hz), and mean F2 (Hz).

2.2. Mahalanobis distances setup

In this work we consider a three-dimensional space made of the vowel tokens' duration (s), mean F1 (Hz), and mean F2 (Hz). In addition to the traditional mean F1 and F2, we take into account

phone duration for it may be helpful to get more "discriminating" profiles. The filter was coded as a R script [15], mainly using *mahalanobis* function from the *mvtoultier* package [16].

In the next section, we will investigate several settings for MD. As the data were build for a sociophonetic study we had more information than just the vowel class and the speakers' gender, and were able to try 4 different set-ups taking into account the vowel class, the speaker, their gender, or the speech condition. Thus for each 160k tokens of vowel we calculated a Mahalanobis distance between the phone (P) and:

- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class and speaker.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class and production condition.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class, speaker and production condition.

3. MAHALANOBIS DISTANCE AS A DATA-DRIVEN TOOL FOR PRUNING

The method allowing to keep the most data points while excluding the most extreme values was the one using only vowel class. The rejection rate was of 12% ($\sigma = 1.4$ depending on vowel classes), which is close the 10% reported by [17] on a similar work. The class ɛR shows the highest rejection rate (13.5%), which makes sense as these vowels can be both diphthongs or monophtongs, thus making the computation of a unique profile more complicated. The monologue condition lost just over 16% of its occurrences after filtering, compared to around 7.7% in Reading, 11.5% for spontaneous speech with a spouse, 13.6% with a stranger, and 13% with a foreign stranger. These style related results become very interesting in the light of phonostylistic findings and can be linked to work by [18, 19, 20, 21] (among others) on phonetic reduction and variation in speech, meaning that our method could be used efficiently for research in this field.

However, we have resorted to a more qualitative examination of the differences between the occurrences estimated as too extreme by our different filters in order to have a sharper vision of the different performances of these methods. Thus we examined the occurrences classified as valid or erroneous for three classes at the extremities of the system: i#, u#, and ɑ#. We have specifically chosen these three classes for the following reasons: i# are among the most stable in terms of phonological variation, and raise very few problems for formant detection (except for F1 detection when contextual devoicing happens); u# also have this stability, but represent a greater challenge for formants extraction due to compactness, very low F2, and the possibility of devoicing; as for ɑ#, they have a wide range of phonological variability, going from [a] to [ɔ], and can trigger detection problems when F1 and F2 are too close (e.g. $\sim$ 1000Hz).

Table 2 summarises the observations made for each filter and for the three classes. The table reads as follows: with the filter only using vowel class, out of the 3 171 tokens of i# tagged as erroneous 8% were actually valid datapoints (i.e. with mean F1s between 100Hz and 500Hz, mean F2s between 1800Hz and 2800Hz, and duration under 0.5s)[1], over the 570 u# tagged as erroneous, 21% were valid ($100 < F1 < 500$Hz, $600 < F2 < 1200$Hz, $dur < 0.5$s), and over the 1 699 tokens of ɑ# tagged as erroneous 17.5% were valid ($400 < F1 < 1$KHz, $800 < F2 < 1700$Hz, $dur. < 0.5$s). On the other hand, some tokens which should have been tagged as erroneous were not: for class ɑ# over 15 495 tokens classified as valid datapoints, about 23% should have been excluded by the filter ($F2 > 1800$Hz), for i# it drops to 8% ($F1 > 500$Hz & $F2 < 1600$Hz) of the 22,528 tokens, and about 27% of 3676 for u# ($F1 > 500$Hz & $F2 > 1500$Hz).

Filter	Diag.	i#	u#	ɑ#
class	WK	8%	27%	23%
	WD	8%	21%	17.5%
class*spkr	WK	6%	23%	21%
	WD	30%	28%	70%
class*sex* condition	WK	8%	27%	25%
	WD	30%	24%	57%
class* condition	WK	7%	28%	23%
	WD	34%	23%	52%

Table 2: Diagnostics of phones wrongly filtered/discarded (WD) or wrongly kept (WK) according to the different filters setups on the QF data.

Overall, MD filtering method did well with cases of devoicing (very common in QF), such as in the word "édifice" *building* /edifis/ where the first [i] is devoiced and drown in the affrication noise of /d/. However it missed some formants jumps in compact back vowels such as [ɔ] in [ʃprɔpoz] (reproduced in figure 2). In this example, F2

was wrongly detected around 1800Hz when it was actually around 1100Hz, merged with a high F1.

A few questionable events also arised. Figure 3 shows a case in which the vowel realization is very far from the average profile of the class: the /o/ is realized with an F2 closer to a [œ] (F1 at 370Hz and F2 at 1732Hz). This fronting movement was documented by [22, 10] as a rare but possible sociophonetic variation. Thus our filter might have removed some important tokens for the study of variation.

Finally table 3 gives the mean formants values that result from our filtered data. QF i#, u# and a# are often described as similar to France French /i, u, a/, and the formant values we report for these three vowels in QF are very close to what was found in France French reference values [23, 3], which is why we want to use them as a further evidence of the process's success.

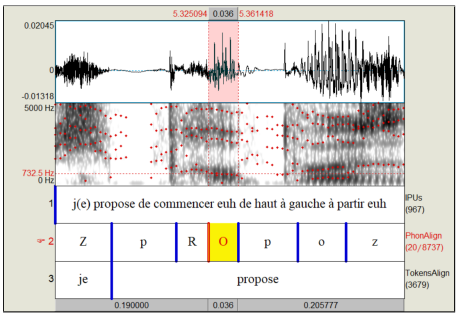


Figure 2: Spectrogram and oscillogram illustrating a case of poor detection of F2 in a [ɔ] from the word "propose" /prɔpoz/ *suggest*.

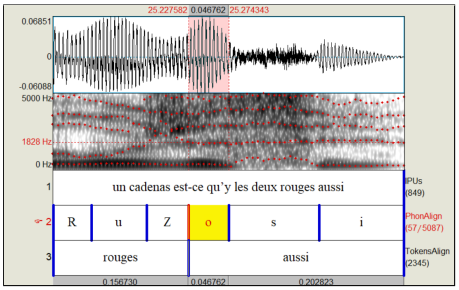


Figure 3: Spectrogram and oscillogram illustrating the non canonical realization of an /o/ in the word "aussi" /osi/ *also*.

.

4. CONCLUSIONS

The "by class" filtering process we used on the FQ corpus identified 12% of vowel tokens as erroneous. The number may seem very (too) substantial, however on a corpus of such magnitude, for which automatization was the main method, it is not unrealistic to consider that certain types of recurring errors have crept into segmentation and formant detection. In a way our rejection rate is pretty close to what [17] reported.

	mean F1		mean F2	
V. Class	F	M	F	M
i#	393	348	2184	1957
iK	435	368	2118	1896
iR	423	348	2191	2004
y#	425	396	1863	1717
yK	449	393	1788	1630
yR	474	425	1540	1473
e#	426	376	2139	1903
ɛ#	560	501	1858	1638
ɛK	539	474	1942	1692
ɛR	652	561	1748	1559
ø#	405	365	1690	1517
a#	609	532	1641	1516
aC	723	616	1627	1458
ɑ#	586	542	1325	1238
o#	442	407	1039	988
ɔ#	574	518	1436	1322
ɔK	562	513	1250	1180
u#	418	386	1299	1219
uR	443	398	1099	1012

Table 3: Mean formant values for each vowel class after the filtering by class only.

The classes experiencing the highest rejection rate were those with vowels that can be both diphthongs or monophtongs (e.g. ɛR), making the computation of a unique profile more complicated; while less variable classes had smaller rejection rates (e.g. 9.9% for ɑ#).

The differences of rejection rates between speech styles also demonstrate that MD filter can be used widely without previous information on style and will not erase style based differences. It also gives information on the variation rate of style, as the most standard style (Reading) rejection rate was of 7.7%, as opposed to 16% in self directed speech; which is in line with phonostylistic findings such as [18, 19, 20, 21].

However, more investigation is needed on the efficiency of Mahalanobis distances for filtering purpose. A comparison with [1]'s result would state whether MD is more efficient before or after features extraction. Different thresholds must be tested to evaluate which distance from the profile is the most suitable for phonetic studies and their different goals. Various type of data from high and low resources languages, different speech styles, and speaker profiles, should be tested. Further work also need to investigate on the effect of lexical categories and frequencies known to affect variation.

5. ACKNOWLEDGEMENTS

The authors want to thank the Swiss National Science Fundation as well as the recorded speakers.

6. REFERENCES

[1] J. Mielke, E. R. Thomas, J. Fruehwald, M. McAuliffe, M. Sonderegger, J. Stuart-Smith, and R. Dodsworth, "Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several english speech corpora," 2019.

[2] D. Amazouz, M. Adda-Decker, and L. Lamel, "Variation du voisement des occlusives orales en code-switching: analyses par abx automatique et mesures acoustiques," in *Journées d'Études sur la Parole-JEP2022*, 2022.

[3] C. Gendrot and M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German." in *Interspeech 2005, Lisbon, Portugal*, 2005, pp. 2453–2456. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-00188096

[4] M. Lancien, "Le rôle de la réduction phonétique dans lâexpression de la proximité sociale," Ph.D. dissertation, Université de Lausanne, 2021.

[5] D. C. Walker, *The pronunciation of Canadian French*. University of Ottawa press Ottawa, 1984.

[6] L. Santerre, "Voyelles et consonnes du français québécois populaire," in *Identité culturelle et francophonie dans les Amériques*. Québec : Presses Universitaires de Laval, 1976, vol. 1, pp. 21–36.

[7] M.-H. Côté, "Laurentian french (quebec) : extra vowels, missing schwas and surprising liaison consonants," in *Phonological variation in French: Illustrations from three continents*, R. Gess, C. Lyche, and T. Meisenburg, Eds. Amsterdam : John Benjamins, 2012, pp. 235–274.

[8] W. Labov, *Principles of Linguistic Change: Internal Factors*. Oxford, Blackwell, 1994, vol. 1.

[9] C. Paradis, "An acoustic study of the variation and change in the vowel system of chicoutimi and jonquière (québec)," Ph.D. dissertation, University of Pennsylvania, 1985.

[10] M. Yaeger, "Context-determined variation in montreal french vowels." Ph.D. dissertation, University of Pennsylvania, 1979.

[11] M. Lancien, J. Stuart-Smith, and M. Adda-Decker, "Knowledge-driven vs. data-driven methods for filtering acoustic measures in phonetics corpora," in *Proceedings of the 20th International Congress of Phonetic Sciences, ICPhS 2023, Prague, Czech Republic, 7-11 August 2023*, submitted.

[12] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and speech*, vol. 53, no. 4, pp. 510–540, 2010.

[13] R. Baker and V. Hazan, "Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior research methods*, vol. 43, no. 3, pp. 761–770, 2011.

[14] M. Lancien, M.-H. Cote, and B. Bigi, "Developing resources for automated speech processing of quebec french," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5325–5330. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.655

[15] M. Lancien, "Filtering data for phonetics," 2021. [Online]. Available: https://github.com/M-Lancien/Filtering_data_for_phonetics

[16] P. Filzmoser and M. Gschwandtner, *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*, 2018, r package version 2.0.9. [Online]. Available: https://CRAN.R-project.org/package=mvoutlier

[17] J. Stuart-Smith, M. Sonderegger, R. Macdonald, J. Mielke, M. McAuliffe, and E. Thomas, "Large-scale acoustic analysis of dialectal and social factors in english/s/-retraction," 2019.

[18] J.-L. Rouas, M. Beppu, and M. Adda-Decker, "Comparison of spectral properties of read, prepared and casual speech in french," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., 2010.

[19] N. Audibert, C. Fougeron, C. Gendrot, and M. Adda-Decker, "Duration-vs. style-dependent vowel variation: a multiparametric investigation," in *Proceedings of the 18th International Congress of Phonetic Sciences, ICPhS 2015, Glasgow, UK, 10-14 August 2015*, 2015.

[20] M. Lancien and M.-H. Côté, "Phonostyle et réduction vocalique en français laurentien," in *6e Congrès Mondial de Linguistique FranÃ§aise, 9-13 juillet 2018, Mons, Belgique*, vol. 46, 2018.

[21] M. Lancien, "Variation stylistique en français québécois : l'effet de l identité de l'interlocuteur," in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, and S. Schneider, Eds. Nancy, France: ATALA, 2020, pp. 308–316. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02812571

[22] J.-D. Gendron, *Tendances phonétiques du français parlé au Canada*. Paris : Klincksieck, 1966.

[23] Calliope (Firm), J. P. Tubach, and G. Fant, *La parole et son traitement automatique*. Paris ; Milan ; Barcelone : Masson, 1989.

[1] We willingly take a wide range of variation to take into account the different possible degrees of reduction.

STOP DEVOICING AND PLACE OF ARTICULATION: A CROSS-LINGUISTIC STUDY USING LARGE-SCALE CORPORA

Anisia Popescu[1], Mathilde Hutin[1], Ioana Vasilescu[1], Lori Lamel[1], Martine Adda-Decker[1,2]

Université Paris Saclay, LISN, CNRS; Université Sorbonne-Nouvelle, CNRS, LPP, Paris
anisia.popescu@universite-paris-saclay.fr

ABSTRACT

Voicing in stops is relatively hard to maintain from an articulatory and aerodynamic standpoint. Consonants with posterior constrictions (velars) are more prone to devoicing than consonants with more anterior constrictions (labials, coronals). Aerodynamic accounts have linked this pattern to the compliant tissue surface available for passive vocal tract expansion. This study investigates whether place of articulation plays a role in stop devoicing in five Romance languages using large corpora and automatic alignment with pronunciation variants. Results show that voicing maintenance patterns do not support an aerodynamic hypothesis: In Romance languages velars do not devoice at higher rates than labials and coronals. This suggests other elements than articulation naturalness are at play.

Keywords: large corpora, voicing maintenance, automatic alignment, pronunciation variants, Romance languages

1. INTRODUCTION

Maintaining voicing requires sufficient transglottal airflow. In stops, air accumulates in the oral cavity, reducing the oral and subglottal pressure gradient which may lead to insufficient transglottal airflow and result in voicing loss. This is known as the Aerodynamic Voicing Constraint (AVC) [1, 2]. Certain places of articulation are more compatible with voicing than others: It has been shown that voiced stops are more marked the more posterior the constriction location is [3] (i.e., voiced labials and coronals are more frequent than voiced velars, making /g/ a frequent phonological gap [4, 5]). Initially this pattern was attributed to oral cavity volume [6]: Pressure builds up faster in a posterior constriction location leading to faster voicing loss in velar stops.

Several studies [7, 8] have since shown that the difference in cavity volume alone has a negligible effect on pressure differences, and that the pattern is more likely related to the compliant surface available for passive enlargement of the vocal tract: Stops with more anterior places of articulation have more soft tissue available for increasing the vocal tract volume (i.e., lowering the larynx, elevating the soft palate, advancing the tongue root, depressing the tongue body [9, 10]).

However, most studies investigating voicing alternations as a function of place of articulation are done in controlled, small-scale laboratory studies (e.g., [11, 12]). Large scale cross-linguistic studies are scarce. A recent corpus study investigating the voicing of Danish stops in intervocalic position [13] did not find the predicted pattern supporting an aerodynamic account: The highest rate of voicing was found for coronals, not velars. The present study aims to test whether velar stops are more prone to devoicing than their labial and coronal counterparts in the European variants of five Romance languages (French, Italian, Portuguese, Romanian and Spanish), using large corpora and automatic alignment with pronunciation variants.

2. METHODOLOGY

2.1. Corpora and devoiced stop detection

The corpus consists of more than 1000 hours of French, Italian, Portuguese, Romanian and Spanish journalistic speech from TV and radio shows.These corpora were acquired from the LDC and ELRA, or developed during international projects: IST ALERT for part of the data in French and Portuguese [14, 15]; IRST for part of the data in Italian [16, 17] and OSEO Quaero [18, 19] for all languages. Only European shows were kept in the final corpus. Manual transcription was available for all data, except for Romanian, which only had 7h of manually transcribed data; The reminder of the Romanian data was automatically transcribed using a speech-to-text system [20].

An automatic speech recognition (ASR) system [21] was used to carry out the forced alignment, matching speech segments to their orthographic transcription using language specific acoustic

models and pronunciation dictionaries. A total of 486,366 bilabial, 1,557,644 coronal and 283,328 velar stops were labeled by the system across the five languages. Table 1 shows the counts of phonemically voiced stops in the corpora for each of the five languages.

	b	d	g	Total
Fre	25,274	197,543	24,731	247,548
Ita	34,289	183,571	26,877	244,737
Por	26,689	119,566	25,125	171,380
Rom	100,518	398,219	87,399	586,136
Spa	299,596	658,745	119,196	1,077,537

Table 1: Counts of phonemically voiced stops per language: French (Fre), Italian (Ita), Portuguese (Por), Romanian (Rom), Spanish (Spa).

Following [21, 22], the dictionaries were enriched with equiprobable pronunciation variants for stop voicing (e.g., the Fr. *beaucoup* /boku/ 'a lot' had four possible pronunciations: [boku], [poku], [bogu], [pogu]) allowing the system to choose the variant which best matches the language-specific acoustic models. This method has proven reliable and useful to the investigation of voicing and devoicing in several recent works (Romanian [23], French [24], Spanish [25, 26, 27], Italian [28] or several Romance languages [29, 30, 31, 32]). The rates of devoiced variants detected by the system are given in Tab. 2 and illustrated in Fig. 1.

Language	bilabial	coronal	velar
French	6.69%	6.7%	7.01%
Italian	7.58%	6.11%	8.25%
Portuguese	8.23%	15.12%	10.03%
Romanian	6.81%	8.28%	8.43%
Spanish	7.19%	8.79%	8%

Table 2: Percentage of bilabial, coronal and velar devoiced stops per language.

2.2. Statistical analysis

Various phonetic and phonological factors have been shown to correlate with loss of voicing. The factors we are considering in this study are place of articulation (PoA: bilabial, coronal, velar), stop duration, language (French, Italian, Portuguese, Romanian, Spanish), position in the word (initial, medial, final), position in the syllable (onset, coda), phonotactic complexity (singleton, cluster) and voicing status of the previous and following segment (pause, voiceless, voiced). The data was

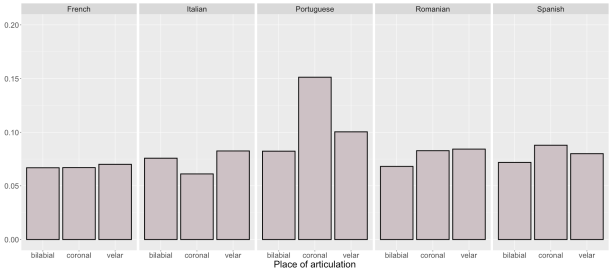


Figure 1: Percentages of devoiced stops detected in the corpora as a function of Language (from left to right: French, Italian, Portuguese, Romanian, Spanish) and Place of Articulation (from left to right: bilabial, coronal, velar).

modeled using a logistic regression and was largely inspired by the analysis in [13]. The duration variable was log-transformed in order to reach a near normal distribution. Categorical variables were contrast coded as follows: A sum contrast was coded for binary variables (phonotactic complexity, syllable position), theoretically motivated Helmert contrasts were coded for the three-level variables (see below) and deviation coding was used for the five-level Language variable (i.e., each language was compared with the overall mean of the Language variable). Each term and their interaction with place of articulation were added as factors to the model. Since speaker information (speaker ID) was not available for all data, random factors were not included for the present analysis. Below we describe the theoretically motivated Helmert contrast coding:

1. PLACE OF ARTICULATION
   - velar contrast ∼ -1/3 bilabials, -1/3 coronals, 2/3 velars
   - bilabials *vs* coronals: -1/2 bilabials, 1/2 coronals
2. POSITION IN WORD
   - final contrast ∼ -1/3 initial, -1/3 medial, 2/3 final
   - initial *vs* medial: -1/2 initial, 1/2 medial
3. VOICING ENVIRONMENT
   - voicing contrast ∼ -1/3 pause, -1/3 voiceless, 2/3 voiced
   - pause *vs* voiceless: -1/2 bilabials, 1/2 coronals

2.3. Predictions

Based on previous literature, we make the following predictions for each of our model variables:

- PLACE OF ARTICULATION: higher rates of devoicing in velar than in labial and coronal consonants (velar > coronal > bilabial) [7]
- STOP DURATION: higher rates of devoicing in

longer stops [32]
- LANGUAGE: no specific predictions
- POSITION IN WORD: higher rates of devoicing in final position followed by initial and medial positions (final > initial > medial) [29]
- SYLLABLE POSITION: higher rates of devoicing in coda position (coda > onset) [33]
- PHONOTACTIC COMPLEXITY: higher rates of devoicing in clusters than in singletons (clusters > singletons)
- VOICING ENVIRONMENT: higher rates of devoicing in pause or voiceless contexts (pause > voiceless > voiced) [32]

3. RESULTS

Results will be presented in two stages. First we report on overall tendencies for all variables (i.e., all five languages pooled together). We then present language-specific devoicing differences related to place of articulation.

3.1. Overall trends

- PLACE OF ARTICULATION
  - velar contrast ∼ velars do not devoice at higher rates than bilabials and coronals (opposite trend: Est.: -0.654994, t ∼ -10.08, p<0.0001).
  - labials *vs* coronals: coronals devoice at higher rates than labials (1.116872, t ∼ 18.923, p<0.0001).
- STOP DURATION: higher rates of devoicing in longer stops (Est. -0.371034, t ∼ -44.202, p<0.0001).
- POSITION IN WORD: higher rates of devoicing in final position compared to initial and medial positions (Est.: 0.503506, t ∼ 22.936, p<0.0001) and higher rates of devoicing in intial than in medial position (Est. -0.399143, t ∼ -49.466, p<0.0001).
- POSITION IN SYLLABLE: higher rates of devoicing in coda position (Est.: 1.271676, t ∼ 69.280, p<0.0001).
- PHONOTACTIC COMPLEXITY: higher rates of devoicing in clusters than in singletons (Est.: 0.147639, t ∼ 16.994, p<0.0001).
- VOICING ENVIRONMENT: lower rates of devoicing if the previous or following segment is voiced (previous: Est ∼ -1.44, t ∼ -144.634, p<0.0001; following: Est.: -0.993, t ∼ -40.937, p<0.0001) and lower rates of devoicing if the previous or following segment is a pause when contrasted with a voiceless stop (previous: -0.461502, t ∼ -28.185, p<0.0001; following: p>0.1).

The reported patterns are given in Tab. 3 and illustrated in Fig. 2, which show the rates of devoiced stops detected by the system for each of our categorical variables (word position, syllable position, phonotactic complexity, voicing status in the previous and following segments) for all five languages.

Variable	Levels	%
Word position	initial	9%
	medial	5.9%
	final	35.9%
Syllable position	onset	7.22%
	coda	28.9%
Complexity	singleton	7.9%
	cluster	10.7%
Previous segment	pause	21.6%
	voiceless	19.5%
	voiced	6.8%
Following segment	pause	40.8%
	voiceless	28.8%
	voiced	7.5%

Table 3: Rates of devoiced stops detected in the corpora for each categorical variable in the model.

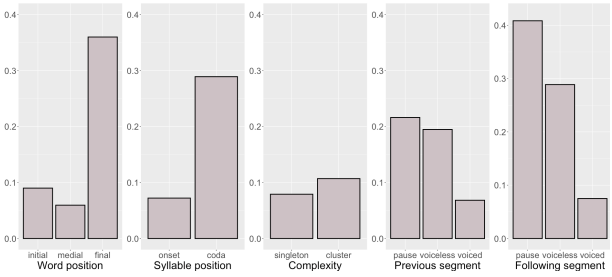


Figure 2: Rates of devoiced stops detected in the corpora for each categorical variable in the model.

3.2. Language-specific patterns

When modeling the velar and labial *vs* coronal contrasts per language (comparing each language with the overall mean of the LANGUAGE variable), we find that the velar contrast (velars compared to bilabials and coronals) is not significant for Italian, Spanish and French (p>0.1). Romanian and Portuguese pattern with the overall direction (velars tend to devoice less than bilabials and coronals), the effect being highest for Portuguese (Por: Est.: -0.106281, t ∼ -4.842, p<0.0001; Rom: Est.: -0.037202, t ∼ -2.204, p<0.05). When looking at contrasts between bilabials and coronals, we find

that there is no significant difference in devoicing rates for Romanian and French (p>0.1). Portuguese has significantly higher rates of coronal than bilabial devoicing (Est.: 0.540493, t ∼ 25.003, p<0.0001). Spanish and Italian both show patterns that support aerodynamic accounts of voicing maintenance: Bilabials tend to devoice more than coronals in both languages, when compared to the overall mean (Ita: Est.: -0.417066, t ∼ -20.304, p<0.0001; Spa: Est.: -0.025293 t ∼ -2.090, p<0.05).
To sum up, velars do not devoice at higher rates than bilabial and coronal consonants: No significant effect was found for Italian, Spanish and French and the opposite trends (i.e., velars devoice less than bilabials and coronals) was found for Portuguese and Romanian. The labial *vs* coronal contrast seems to be more language-specific: In Italian and Spanish, bilabials devoice at higher rates than coronals, Portuguese shows the opposite trend (coronals > bilabials) and Romanian and French show no difference between bialabial and coronal devoicing rates when compared to the overall mean.

4. DISCUSSION

The present study investigates whether there are any differences in stop voicing maintenance based on place of articulation in the European varieties of five Romance languages. More than 1000 hours of journalistic speech were automatically aligned using pronunciation variants, which allow the system to automatically detect devoiced stops based on language-specific automatic speech recognition (ASR) systems. By analyzing the effect of place of articulation along with other phonetic and phonological factors, we find that our predictions based on aerodynamic accounts of voicing maintenance are not confirmed: Stops with more posterior constriction locations (velars) are not more prone to voicing loss. Instead, in line with previous studies [13], we find that coronals tend to devoice at higher rates (a trend that is mainly driven by the high coronal devoicing in Portuguese). Several explanations can be put forward. First, coronals tend to be more frequent than bilabials and velars, and articulatory reduction phenomena (e.g., glottis spreading) tends to occur in more frequent environments. This is also the case for our data (there are twice as many coronal stops than velars and bilabials). Second, results might be influenced by the high rate of coronal devoicing found for Portuguese. Portuguese has been shown to behave differently than other Romance languages when it comes to voicing patterns [11, 34]. Another explanation could be that in languages which require actual vocal fold vibration to cue the voicing-contrast, such as Romance languages (with the possible exception of Portuguese), speakers might adapt their articulatory strategies to circumvent the AVC and maintain the voicing contrast, as suggested by [35]. In other words, the articulatory goal of vocal fold vibration for voiced stops [36] is achieved by compensating for the aerodynamic constraints.

Finally, there is also the issue of methodology. The present study relies on language dictionaries with pronunciation variants to detect devoiced stops. When selecting the appropriate variant the ASR system relies on acoustic features, which aren't accessible to us. Our analysis of the variant selection is mostly phonological. To thoroughly investigate the relationship between place of articulation and voicing maintenance, as well as language-specific differences within a single language family, large-scale acoustic studies of the data are planned. Acoustic analyses (e.g., voicing probability throughout the stop) would give us a clearer picture of the stop devoicing phenomenon. Furthermore, the present analysis should be extended to languages that, similarly to Romance languages, need vocal fold vibration during the stop's closure (Polish, Bulgarian, Catalan, Dutch) as well as to languages that do not rely on voicing during the stop's closure (German, English, Mandarin).

5. REFERENCES

[1] J. Ohala, "The origin of sound patterns in vocal tract constraints," in *The production of speech*, P. M. (Eds.), Ed. New York: Springer Verlag, 1983, pp. 189–216.
[2] ——, "Accommodation to the aerodynamic voicing constraint and its phonological relevance," in *Proc. of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 64–67.
[3] T. Gamkrelidze, "On the correlation of stops and fricatives in a phonological system," *Lingua*, vol. 35, 1975.
[4] D. Nikolaev, "A typology of consonant inventory gaps," *Linguistic Typology*, vol. 26, 2022.
[5] D. Sherman, "Stop and fircative systems: A discussion of paradigmatic gaps and the question of language sampling," *Working Papers on Language Universals*, vol. 17, pp. 1–31, 1975.
[6] A. Martinet, *Economie des changements phonétiques*. Berne: Ed. A. Francke, 1955.
[7] J. Ohala and C. Riordan, "Passive vocal tract enlargement during voiced stops," in *Speech Communication papers*, J. Wolf and D. Klatt, Eds. New York: ASA, 1979, pp. 82–92.

[8] H. Javkin, "Towards a phonetic explanation for universal preferences in implosives and ejectives," *Berk. Ling. Soc.*, vol. 3, pp. 559–565, 1977.
[9] J. Westburry, "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," *JASA*, vol. 73, pp. 1322–1336, 1983.
[10] M. Rhotenberg and J. Mahshle, *The breadth-stream dynamics of simple-released-plosive production*. Basel: Karger, 1968.
[11] D. Pape and L. Jesus, "Stop and Fricative Devoicing in European Portuguese, Italian and German," *Lang Speech*, pp. 224–246, 2015.
[12] P. Keating, "Physiological effects on stop consonant voicing," *UCLA Working Papers in Phonetics*, vol. 59, pp. 18–28, 1984.
[13] R. Pugaard-Rode, C. Horslund, and H. Jorgensen, "The rarity of intervocalic voicing of stops in danish spontaneous speech," *Laboratory Phonology*, vol. 13, 2022.
[14] T. F. da Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *Proc. of ICASSP*, 2011, pp. 4656–4659.
[15] ——, "Interpolation of acoustic models for speech recognition," in *Interspeech*, 2013, pp. 3347–3351.
[16] J. Després, L. Lamel, J.-L. Gauvain, B. Vieru-Dimulescu, C. Woehrling, V. B. Le, and I. Oparin, "The vocapia research asr systems for evalita 2011," in *Lecture Notes in Computer Science*, C. Bosco and A. Mazzei, Eds. Berlin Heidelberg: Springer, 2013, vol. 7689, pp. 286–294.
[17] F. Marcello, D. Giordani, and P. Coletti, "Development and Evaluation of an Italian Broadcast News Corpus," in *Proc. of LREC 2000*, 2000, pp. 921–924.
[18] L. Lamel, S. Courcinous, J. Despres, J.-L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, L. V. Bac, H. Ney, M. Nussbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. F. D. Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. T. Vu, A. Waibel, and C. Woehrling, "Speech Recognition for Machine Translation in Quaero," in *Proc. of IWSLT 2011, San Francisco, USA*, 2011.
[19] L. Lamel, "Multilingual speech processing activities in Quaero: application to multimedia search in unstructured data," in *International Conference on Human Language Technologies - The Baltic Perspective*. Tartu, Estonia: IOS Press, 2012.
[20] I. Vasilescu, B. Vieru, and L. Lamel, "Exploring pronunciation variants for Romanian speech-to-text transcription," in *Proc. 4th International Workshop on Spoken Language Technologies for Under-resourced Languages*, St. Petersburg, 2014, pp. 203–206.
[21] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary continuous speech recognition," in *Proc. 4th International Conference on Spoken Language Processing*, Philadelphia, 1996.
[22] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style," *Speech Comm.*, vol. 29, 1999.
[23] M. Hutin, A. Jatteau, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Ongoing Phonologization of Word-Final Voicing Alternations in Two Romance Languages: Romanian and French," in *Interspeech*, 2020, pp. 4138–4142.
[24] A. Jatteau, I. Vasilescu, L. Lamel, M. Adda-Decker, and N. Audibert, "'Gra[f]e!' Word-Final Devoicing of Obstruents in Standard French: An Acoustic Study Based on Large Corpora," in *Interspeech*, 2019, pp. 1726–1730.
[25] N. Ryant and M. Liberman, "Large-scale analysis of spanish /s/-lenition using audiobooks," in *Proceedings of the 22d International Congress on Acoustics*, 2016.
[26] I. Vasilescu, N. Hernandez, B. Vieru, and L. Lamel, "Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s," in *Interspeech*, 2018, pp. 2728–2732.
[27] J. Hualde and P. Prieto, "Lenition of intervocalic alveolar fricatives in catalan and spanish," *Phonetica*, vol. 71, pp. 109–127, 2014.
[28] J. Hualde and M. Nadeu, "Lenition and phonemic overlap in Rome Italian," *Phonetica*, vol. 68, pp. 215–242, 2011.
[29] I. Vasilescu, Y. Wu, A. Jatteau, M. Adda-Decker, and L. Lamel, "Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes," *Traitement Automatique des Langues*, vol. 61, no. 1, 2020.
[30] M. Hutin, Y. Wu, A. Jatteau, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Synchronic Fortition in Five Romance Languages? A Large Corpus-Based Study of Word-Initial Devoicing," in *Interspeech*, 2021, pp. 996–1000.
[31] M. Hutin, M. Adda-Decker, L. Lamel, and I. Vasilescu, "When phonetics meets morphology: Intervocalic voicing within and across words in romance languages," *Interspeech*, 2022.
[32] Y. Wu, M. Hutin, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Extracting Linguistic Knowledge from Speech: A Study of Stop Realization in 5 Romance Languages," in *LREC 2022*, Marseille, France, Jun. 2022.
[33] P. Ségéral and T. Scheer, "Positional factors in lenition and fortition," *Lenition & fortition, ed. J. Brandao de Carvalho, T. Scheer, & P. Ségéral*, pp. 131–172, 2008.
[34] M. Hutin, M. Adda-Decker, L. Lamel, and I. Vasilescu, "When phonetics meets morphology: Intervocalic voicing within and across words in romance languages," in *Interspeech*, 2022.
[35] M.-J. Solé, "Articulatory adjustments in initial voiced stops in spanish, french and english," *Journal of Phonetics*, vol. 66, pp. 217–241, 2018.
[36] C. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1989.

PROSODIC ANNOTATION OF LARGE SPEECH CORPORA

Philippe Martin

LLF, UFRL, Université Paris Cité
philippe.martin@u-paris.fr

ABSTRACT

Recent developments in linguistic analysis and automatic recognition made the availability of large spontaneous speech corpora pivotal. The ease of use of efficient research tools is becoming essential, given the usual large amount of data considered. The software described in this paper addresses these requirements, by offering fast access to recorded data in a variety of formats, allowing efficient acoustic analysis and prosodic annotation tools of selected speech segments of interest.

Pre-compilation of transcribed text for a very large number of recordings makes the extraction of selected transcribed material extremely fast, together with the display of spectrogram, wave, intensity, and fundamental frequency curves.

Prosodic annotation uses graphic drawing tools, operating automatically or manually either with the ToBI standard notation, or an automatic categorization of melodic contours based on their glissando levels.

Annotation of large speech corpora are presently available in French, English, Italian, Portuguese, Spanish, Dutch, German, Russian, Malay, Korean and Mandarin.

Keywords: Spontaneous speech, concordancer, prosodic annotation, spectrogram, fundamental frequency.

1. INTRODUCTION

Deep learning training for TTS (text to speech) and STT (speech to text) usually requires a very large number of transcribed speech samples. This leads both private and public research institutions to develop large, dedicated speech corpora, both read and spontaneous. Next to existing speech data designed by and for linguists (e.g., [2], [3], [5]), corpora designed for AI (e.g., [11], [14], 15]) could also be advantageously used for linguistic research, especially applied to spontaneous speech whose grammatical and prosodic characteristics may not yet be completely understood.

From a more general viewpoint, to acquire a better insight in a specific linguistic domain, automatic processing of data is not always advisable, as researchers may miss important experimental events buried in a large amount of data. Popular tools such as Praat [1], even associated with dedicated scripts, may not be always suitable, especially in the prosodic domain. More efficient software should then be made available to allow detailed examination by operators possibly with less efforts, in particular pertaining to the following points:

a) Handling of various text transcription formats, from character single byte [10] to UFT-8 coding [11].
b) Fundamental frequency analysis adapted to degraded recordings frequently found in spontaneous speech recordings (e.g. [2]).
c) Fast data mining, essential for users dealing with research topics that cannot or should not use automatic extraction processes.
d) Availability of user-friendly annotation tools, pivotal if extraction of data cannot be executed automatically due to degraded speech recordings.

This paper gives the main features of a dedicated spontaneous speech analysis software WinPitch [17], whose functions are tailored for prosodic analysis and annotation of large speech data sets.

2. CORPORA

2.1 Selected speech corpora

The software described in this paper can be used as is with any corpora using similar transcription encoding (see section 3).

2.1.1. dedicated to linguistic analysis, includes mostly spontaneous speech:

Orfeo, French [2].
OFROM, French [3].
C-ORAL-ROM: French, Italian, Spanish, European Portuguese [4].
C-ORAL-ROM Brazil: Brazilian Portuguese [5].
CGN Corpus Gesproken Nederlands, Dutch and Flemish varieties [6].
Santa Barbara corpus of American English [7].
Archive for Spoken German [8].

2.1.2. dedicated to TTS, STT and voice cloning, includes mostly read speech:

Decoda, Spontaneous French conversations [9].
The LJ Speech Dataset, American English [10].
SIWIS Speech Synthesis Database, French, Italian [11].

Beijing Magic Data Technology, French, Italian, Mandarin, Korean, Spanish, … (read and spontaneous speech) [12].
CSTR VCTK: English Multi-speaker Corpus [13].
LibriSpeech ASR corpus, English [14].
Russian read Speech [15].
Zeroth-Korean corpus [16].

3. TRANSCRIPTIONS

3.1 Text formats

All the above corpora come with text transcriptions associated with segments of recordings of various lengths, from isolated sentences (e.g., [11], [15]), to sequences aligned on prosodic breaks (e.g., [2]) or on syntactic boundaries (e.g., [8]) or even of random lengths (many recordings found in [2]).

There is no common standard for text coding transcription, varying from simple 1-byte ASCII per character to the more sophisticated JSON format (with meta information imbedded) or UFT-8 found in Magic Beijing [12]).

Automatic identification, conversion, and extraction of transcription text are implemented for the following formats:

XML: C-Oral-Rom [4],[5]
Trs (Transcriber): Orfeo [2], Decoda [9]
Tag (Transcriber): Orfeo [2]
TextGrid (Praat): Orfeo [2], OFROM [3], CGN (awd) [6]
CRF-ALG Text: Orfeo [2]
Necte XML text: Necte
RTF text: Beijing Magic Data [12]
JSON: Archive for Spoken German [8]
TXT (1 and 2 bytes): SIWIS [11]
UFT-8: Beijing Magic Data [12]

The corpora transcription file extensions do not necessarily give enough information on the text format used, so that automatic selection routines have been designed to handle the text transcription correctly, sometimes imbedded in other text information pertaining to speakers, recording conditions, etc. With the automatic selection of formats, files with any of the formats given above can be mixed in the same project and accessed with all available functions of the software.

3.2 Automatic IPA segmentation

Given that the transcriptions' grain varies from phone (e.g., Decoda [9]) to long sentences reaching more than 30 seconds (e.g., Orfeo [2]), an integrated segmentation algorithm has been implemented, based on dynamic Viterbi alignment with a TTS generated speech signal (available in more than 30 languages). It is then easy to segment accurately sentences transcribed deprived from any segmentation in SIWIS [11] or CSTR VCTK [13] into phones in IPA format.

Unexpected difficulties may arise from text coding variations sometimes found in transcription files. In SIWIS [11], some transcription texts are in plain ASCII (1 byte per character), whereas some others, in the same set of sentences, uses a 2-byte format. In other instances, as in Beijing Magic Data files [12], the transcribed text uses an expected UTF-8 format for Mandarin, but the RTF format for Spanish or French.

4. ACOUSTIC ANALYSIS

Acoustic speech analysis makes it possible to perform prosodic annotation on acoustically degraded recordings, such as those found in old entertainment or documentary movies, but also in many spontaneous speech recordings (e.g., Orfeo [2]). The software described here is specially designed for this purpose. It integrates various user-friendly functions to obtain reliable pitch curves and realize prosodic annotations in adverse acoustic conditions.

Besides the use of 7 user-selected different pitch tracking algorithms (spectral comb, 4 flavours autocorrelation, spectral brush, AMDF…), displaying of an overlay narrow band spectrogram whose frequency scale is automatically aligned on the pitch scale allows the annotator (after some training…) to visually ensure that the pitch curve obtained by the selected tracking algorithms is reliable, as it should correspond to the pitch variations indicated by the first (or the second) spectrogram harmonic.

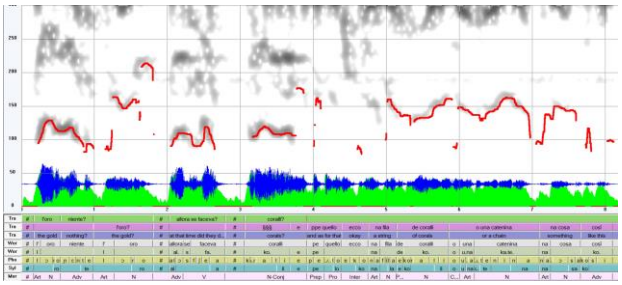


Figure 1. Superposition of the fundamental frequency curve and the first harmonic of the corresponding narrow-band spectrogram. This allows the user to visually check the validity of the fundamental frequency curve, and eventually select an alternate pitch tracking method, or draw on screen with the available drawing functions corrected pitch segments aligned on the first (or second) harmonic of the spectrogram.

Segments of pitch errors can then easily be located and corrected using another pitch tracking algorithm

or directly by the user with the available drawing functions to draw on screen pitch curve segments aligned on the first or second spectrogram harmonic.

This function allows the user to take care of the commonly found problems of spontaneous speech recordings, such as the use of unappropriated microphones (removing low pass frequencies), voice overlays, background noise or even creaky voice (relying on the second harmonic). An example is given Fig. 1.

5. DATA MINING

Dealing with a large amount of transcribed speech data requires an efficient data mining process, based on a concordancer whose input can be any sequence of words (in Unicode format, thus including Korean and Mandarin) to retrieve automatically the corresponding speech segment, together with its acoustic analysis (spectrogram, pitch and intensity curves).

To speed up the retrieval process, a precompiled file is generated for any selected group of recordings, mixing transcription formats if needed. This precompiled file consists of a database linking each word found in all the transcriptions of a user defined group of transcribed recordings to their corresponding speech segment, converted automatically in WAV format. While the compilation of all text transcriptions of a given recording set may take some time (from few seconds for corpora such as Orfeo [2] to few minutes for CSTR VCTK [13], which includes more than 88,000 sentences…). Once this file has been generated, it can be reloaded any time later for a very fast retrieval in less than 1 second of speech data from the concordancer entry. An example is given Fig. 2.

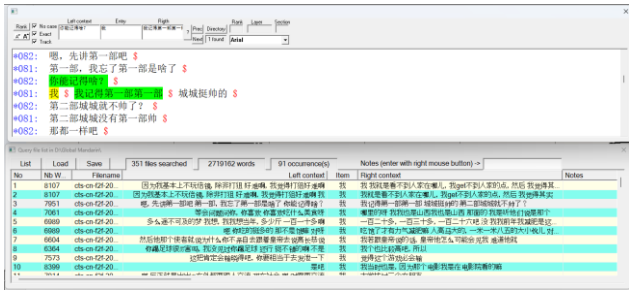


Figure 2. An example of speech segment retrieval in Mandarin from the sequence entered in the concordancer (here the word 我). The context is displayed and highlighted in a dedicated window (top figure). Any surrounding speech segment can be retrieved as well by selecting any part of the text with the mouse.

Fig. 3 and 4 display the pitch curve and the underlying narrow band spectrogram corresponding to the example of Fig 2.

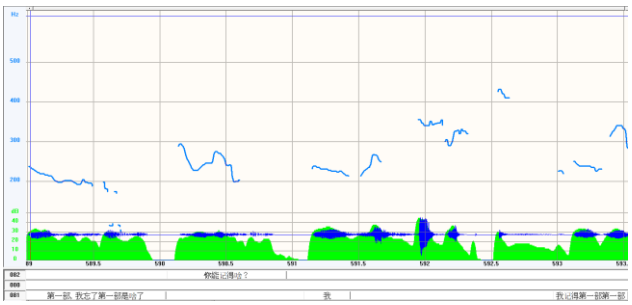


Figure 3. Display of the pitch curve corresponding to the text selected in Fig. 2, with the text transcription layers (bottom of the figure).

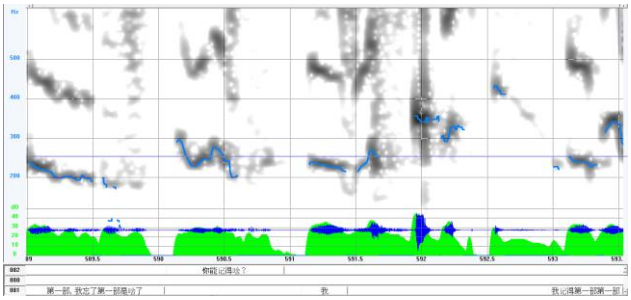


Figure 4. Display of the pitch curve with an overlaying narrow band spectrogram corresponding to the text selected in Fig. 2, with the text transcription layers (bottom of the figure). The validity of pitch tracking can then easily be checked visually.

6. PROSODIC ANNOTATION

Available prosodic annotation tools include text and graphic functions to allow the user to add any kind of information on screen, which could be saved later in text or Excel format. Segments of pitch or formant frequencies can be either automatically or manually highlighted and corrected, if necessary, with mouse-controlled graphic functions, with user-selected color and thickness.

Prosodic annotation presupposes an implied theoretical model (such as Autosegmental-Metrical) and the use of a specific annotation system (ex. ToBI). Fig. 5 shows an example of prosodic annotation by pitch contours of a French sentence. Stressed vowels, located either automatically of manually by the user, have their melodic contours automatically highlighted in various colors, according to their rising or falling pitch, above or below their pitch change perception level (glissando threshold). ToBI annotation can also be entered the same way, once melodic target properties and color coding have been predefined by the user and entered in the system.

With the implied prosodic model used in this example, prosodic annotation operates in three steps:

1) Manual or automatic localization of stressed syllables (excluding emphasis) defining accent

phrases “right” boundaries by their final position (in French);

2) Automatic classification and color coding of prosodic events occurring on stressed syllables. These events are instantiated by specific melodic contours assumed by hypothesis to indicate dependency relations between accent phrases. These user defined dependency relations define the prosodic structure.

3) The corresponding prosodic structure is then automatically displayed (Fig. 5).

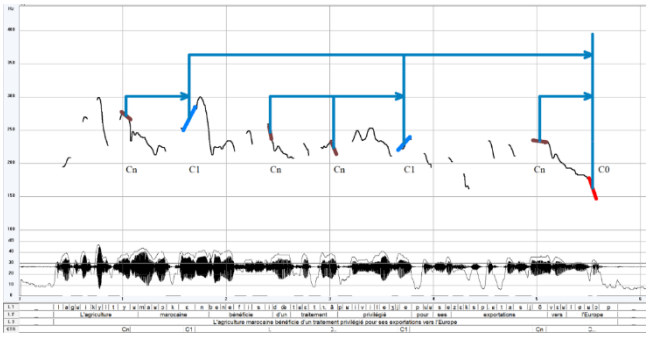


Figure 5. An example of prosodic annotation using a model relying on pitch contour on stressed syllables to indicate dependency relations between accent phrases, which in turn define the prosodic structure (in square brackets with oriented branches). (SIWIS corpus [11]).

Melodic contours, as well as ToBI melodic targets, can be dynamically edited with simple mouse commands, automatically updating the resulting prosodic structure, as well as the actual sentence prosody, modified by the imbedded Psola based melodic morphing routine.

7. EXAMPLES OF APPLICATIONS

7.1 Multi speaker realizations

The SIWIS French corpus [11] consists of a set of relatively short sentences read by 8 to 14 speakers, with a total of some 5340 isolated sentences. Analyzing the variations of prosodic realizations between speakers for a particular sentence is a breeze, requiring only a few minutes from the alphabetic ranking of text transcriptions. Clicking on text entries on the concordancer displays the corresponding speech samples, together with their pitch, intensity, and spectrographic data. One can then quickly in a few seconds observe how different speakers use different prosodic coding strategies even if they are reading the same text.

7.2 Phraseology

Analysis of prosodic realization of connectors, such as *enfin*, or *en fait* in French, requires with the commonly available tools many manipulations to extract the pertinent information from one of available corpora such as Orfeo [2] or OFROM [3]. However, with WinPitch [13], the software described in this paper, the 179 occurrences of *tu parles* found in the Orfeo corpus of more than 3,438,438 words are displayed in less than one second, and their pitch curves can be retrieved and displayed with a single mouse click (Fig. 6).

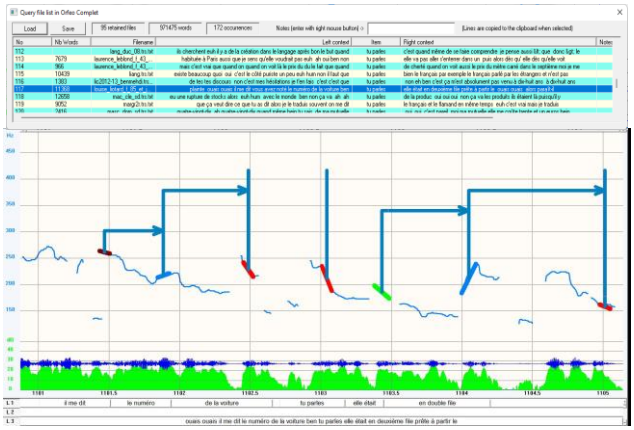


Figure 6. An example of analysis of the occurrence of the French locution *tu parles* "you speak" with its context *il me dit le numéro de la voiture tu parles elle était en double file* (Corpus ORFEO [2]). Pitch contours of stressed vowels are highlighted, and the corresponding prosodic structure (with square brackets) displayed automatically

8. CONCLUSION

The described user-friendly speech analysis software allows efficient prosodic annotations for prosodic research in general on a large variety of speech corpora, both read and spontaneous, making advantage of recent quasi-industrial corpora production designed for other purposes, such as speech recognition and synthesis based on deep learning algorithms, requiring a large amount of data.

The interoperability routines integrated in the software as well as the precompiled transcription data makes research of prosodic and other phonetic characteristics user selected sequences or words extremely fast and easy, making data collection that would normally take months executed in a few hours.

This interoperability allows the use of the program on any collection of transcribed recordings sharing one of the formats listed above, the precompiled transcription being generated by one mouse click operating on data simply collected in the same computer directory.

Developed by the author, the software can be downloaded from www.winpitch.com [17], and used freely for academic purposes.

REFERENCES

[1] Praat, doing phonetics by computers, https://www.fon.hum.uva.nl/praat/

[2] Orfeo, Le Corpus d'Etude du Français Contemporain - Le projet ORFEO www.projet-orfeo.fr/corpus/le-corpus-d-etude-du-francais-contemporain/14-orfeo

[3] Avanzi Mathieu, Béguelin Marie-José, Corminboeuf Gilles, †Diémoz Federica & Johnsen Laure Anne (2012-2023). Corpus OFROM – Corpus oral de français de Suisse romande. Université de Neuchâtel, http://ofrom.unine.ch

[4] C-ORAL-ROM French, Italian, Spanish, European Portuguese, http://www.elda.org/en/proj/coral/fr/coralrom.html

[5] C-ORAL-ROM Brazil, http://www.c-oral-brasil.org/english-site/index.html

[6] CGN Corpus Gesproken Nederlands, https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_English/topics/index.htm

[7] Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

[8] Archive for Spoken German, https://agd.ids-mannheim.de/index_en.shtml

[9] Bechet F. et al. (2012) DECODA: a call-centre human-human spoken conversation corpus, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 1343-1347.

[10] The LJ Speech Dataset, Version 1.0 July 5, 2017, https://keithito.com/LJ-Speech-Dataset

[11] The SIWIS French Speech Synthesis Database, https://datashare.ed.ac.uk/handle/10283/2353.

[12] Beijing Magic Data Technology Co., Ltd., High quality Training Datasets. https://www.magicdatatech.com/

[13] CSTR VCTK Corpus: English Multi-speaker Corpus, https://datashare.ed.ac.uk/handle/10283/3443.

[14] LibriSpeech ASR corpus, Large-scale corpus of read English speech, https://www.openslr.org/12/

[15] Pavel Skrelin et al., A Fully Annotated Corpus of Russian Speech, https://aclanthology.org/L10-1188/

[16] Zeroth-Korean corpus SLR40, https://www.openslr.org/40/

[17] WinPitch, Speech analysis software, https:// www.winpitch.com

HOW VARIABLE ARE ENGLISH SIBILANTS?

Morgan Sonderegger*, Jane Stuart-Smith†, Jeff Mielke‡, The SPADE Consortium

McGill University*; University of Glasgow†; North Carolina State University‡
morgan.sonderegger@mcgill.ca, Jane.Stuart-Smith@glasgow.ac.uk, jimielke@ncsu.edu

ABSTRACT

This paper takes a corpus-phonetic approach to consider variability in word-initial, prevocalic English sibilants in 5k speakers, 235k tokens, from 27 geo-social-ethnic regions of North America and the British Isles. We analyse ERB-transformed, spectral peak measures calculated from amplitude-normalised multitaper spectra using a 'distributional' Bayesian mixed-effects regression which explicitly models token, speaker, and region-level variability. Following previous phonetic and sociolinguistic research we expected English /s/ to be more variable than /ʃ/. The results, however, differ according to the level at which we consider variability. Across English regions, /s/ and /ʃ/ show a similar degree of variability. Across speakers within-region, /s/ is generally more variable than /ʃ/, and within speakers, /s/ is generally more variable by token than /ʃ/, both results likely reflecting linguistic and social-indexical sources of variation—such as gender, which has a greater effect on spectral peak for /s/ than for /ʃ/.[1]

Keywords: sibilants, corpus phonetics, variability, sociophonetics, spectral analysis

1. INTRODUCTION

The sibilant fricatives /s ʃ/, e.g. *seat, sheet*, contrast in North American and British Isles Englishes, with /s/ usually alveolar/denti-alveolar, and /ʃ/ palatoalveolar, also with lip-rounding. Bar subtle known differences in sibilant production from e.g. possible anatomical differences in vocal cavity size [1], English sibilants are assumed to be produced in fairly similar ways and to be acoustically stable across English dialects [2].

At the same time, experimental studies have shown that /s/ is more subject to coarticulatory pressures than /ʃ/, including lip-rounding [3] and articulatory and auditory retraction [4, 5]. This is also reflected in diachrony: /s/ is involved in contextually-induced sound change more often than /ʃ/ [6]. And numerous sociophonetic studies [7] have demonstrated the performative nature of /s/ productions linked to local social-indexical meanings, overriding anatomical differences [8], whilst /ʃ/ is generally assumed to be relatively less socially informative [9]. However, the very large body of phonetic research on English sibilants [2, 10, 11, 12] shows a dialect bias towards speech largely from younger educated adults speaking standard varieties of American and UK English [13]. Sociophonetic research also shows an observational bias towards /s/, with much less attention paid to /ʃ/. Here we take a corpus-phonetic approach, increasing sample size and dialect diversity, to ask: *Is /s/ more variable than /ʃ/ in English?*

2. METHOD

This study presents results from 32 public and private spoken corpora, phonetic and (socio)linguistic, comprising spontaneous and read speech, from the British Isles (England, Scotland, Wales, Northern Ireland and the Republic of Ireland), and from America and Canada. The 5042 speakers (2544 female and 2498 male) were subsequently grouped into 27 broad geo-social-ethnic dialect regions (here 'dialects'), following e.g. [14, 15].[2] All dialects contain spontaneous speech, and some also contain read speech. They may also have speakers of different ages, though we restricted our chronological window to recordings made from the 1990s to the 2010s.

We used [16] to import the force-aligned corpora, and then to extract and analyse all instances of word-initial, prevocalic /s ʃ/ in stressed syllables. We examined the empirical distributions of durations and excluded tokens with durations below 35ms (too short for spectral analysis) or above 400ms (likely alignment errors) (5.2k tokens/6% of the data). We also excluded 4.1k tokens (1.7% of the data) with unrealistically low spectral peak based on the empirical distribution for each corpus. The final dataset consists of 235,582 tokens from 27 regions, 5042 speakers, and 1429 words.

Measuring the main spectral peak in sibilants from female speakers in 16kHz recordings is problematic, as the peak is often above the Nyquist frequency (8kHz). We therefore analysed the main spectral peak only using recordings with sampling

rates of 22kHz or higher, which entailed excluding corpora from the full set [17] sampled at 16kHz. We also excluded recordings of 255 individual speakers after visual inspection of sibilant spectra. Often these were recordings which appeared to be high-sampling-rate digitized versions of low-quality original recordings, lacking energy at frequencies where sibilant noise is expected.

We created multitaper spectra from the middle 25ms of each sibilant interval using the `spectRum` package [18] for `R`, with 8 tapers, a bandwidth of 4, and no preemphasis. To compensate for varying recording conditions, we applied a spectrum normalization scheme. We divided all corpora into speech 'utterances' separated by segments of 150ms+ of non-speech in the forced-alignment, then sampled all utterance and non-utterance intervals at 1000 time points. Some corpora with many short sound recordings had no non-utterance intervals, so we measured utterance intervals instead. Within a given corpus, amplitudes were converted to decibels using the maximum frequency bin value for the utterance intervals as the 0 dB reference point. Spectra were then scaled, for each frequency bin, such that 1 equals the 90th percentile amplitude of utterance intervals and 0 equals the 10th percentile amplitude of non-utterance intervals. The main spectral peak was then measured as the frequency of the maximum amplitude between 1-11kHz.

Acoustic studies on sibilants vary in whether they report results in Hz or non-linear scales like Bark or ERB, and the results, especially for sibilant variation, are not always consistent across the two scales [12, 2, 19]; sociophonetic studies tend to report Hz. Given our interest in variability within and across the phonological categories (c.f. e.g. [20]), we transformed the peak frequency to ERB.

Our goals for modeling spectral `peak` were to assess three ways in which /s/ could be more variable than /ʃ/: **(1)** across dialects, **(2)** across speakers, within a given dialect, and **(3)** across tokens, within individual speakers. We modeled `peak` using a Bayesian regression model, fitted in Stan/brms [21, 22] using weakly-informative regularizing priors [23], which consists of two linear mixed-effects models (LMM): one for the mean of `peak`, as used in any corpus-phonetic analysis of spectral peak (e.g. [12]), and one for the *variance* of `peak` (specified as $\log(\sigma)$)—the amount of by-token variability, which is typically assumed to be constant, but in this kind of 'distributional regression' model [21] can itself vary, specifically to differ between /s/ and /ʃ/ [24].

The first LMM contains fixed effects for Onset (/s/ vs. /ʃ/), log-transformed sibilant Duration (expected: /s/ < /ʃ/), and their interaction, as well as speaker `gender` (expectation: F > M), and an interaction with `onset`, $\beta_{\text{gender:onset}}$, which partially captures comparison **(2)**, and allows for the F − M difference to differ between /s/ and /ʃ/. We include random effects to capture variability in `peak` for /s/ and /ʃ/ separately, by `dialect` and by `speaker` nested within `dialect`. These four random-effect variances, which we call $\sigma_{d,/s/}$, $\sigma_{d,/ʃ/}$, $\sigma_{s,/s/}$, and $\sigma_{s,/ʃ/}$, address **(1)** and **(2)** for an 'average speaker' and an 'average dialect', respectively. We also include near-maximal random slopes for `dialect` and `speaker`, as well as a by-`word` random intercept, which serves as a rough control for linguistic factors (e.g. following vowel height) beyond `onset` which could affect `peak`. The second LMM includes a fixed effect term of `onset` ($\beta_{\sigma,\text{onset}}$) which allows σ to differ between /s/ and /ʃ/, addressing **(3)**. Maximal by-`dialect` and by-`speaker` random effects are also included to accurately estimate this term.

3. RESULTS

We primarily report results relevant for our research questions **(1-3)**. We do so by summarizing the posterior distribution of a quantity of interest computed from model coefficients—e.g. the difference between $\sigma_{d,/s/}$ and $\sigma_{d,/ʃ/}$ for **(1)**—using the median, 95% credible interval (CredI), and probability of effect direction p_d, which are roughly like summarizing an effect for a frequentist model (with p_d = 1 minus p-value); see e.g. [23].

3.1. /s/ vs. /ʃ/ variability across dialects

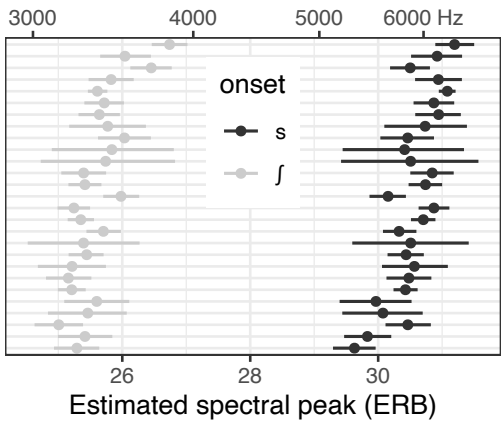


Figure 1: Estimated spectral peaks for English /s/ and /ʃ/ by dialect (one row per dialect). In all plots, dots/lines indicate posterior medians and 95% CredI's.

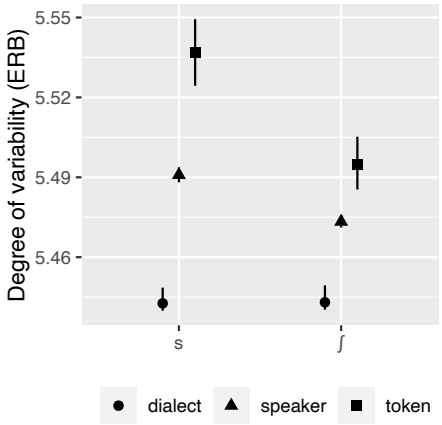


Figure 2: Estimated degree of variability for English /s/ and /ʃ/ by-dialect, by-speaker within-dialect ('speaker'), and by-token within-speaker.

Figure 1 shows spectral peak values estimated by the model for North American and British Isles English /s/ and /ʃ/. Note that dialects differ in certainty of estimates, corresponding to differing sample sizes, from $n < 250$ (3 rows with largest errorbars) to $n > 30$k (3 rows with smallest errorbars). Three findings are apparent. First, there is a gap between the highest /ʃ/ mean and the lowest /s/ mean, even accounting for uncertainty in model predictions. Second, there is substantial variability in the spectral peak location across dialects for both /s/ and /ʃ/. Third, the degree of variability does not clearly differ between /s/ and /ʃ/. This last point, which addresses **(1)**, can be shown by examining the estimated degree of by-dialect variability for each sibilant: the dots in Figure 2, $\sigma_{d,/s/}$ and $\sigma_{d,/ʃ/}$, are essentially the same, as confirmed by a hypothesis test ($\sigma_{d,/s/}$ - $\sigma_{d,/ʃ/}$: median = 0 ERB, 95% CredI = $[-0.005, 0.004]$, $p_d = 0.44$, BF = 0.77).

Two further observations emerge from Figure 1. Our finding that /s/ and /ʃ/ are equally variable cross-dialectally depends on using the ERB scale, closer to auditory processing, than to the linear Hz scale (top axis), where /s/ shows more dialectal variation than /ʃ/. Second, dialects with lower /s/ peaks seem to have lower /ʃ/ peaks. The model's estimated correlation is 0.29 (95% CredI = $[-0.03, 0.57]$, $p_d = 0.93$, BF = 14.0), which offers tentative support for 'contrast uniformity' of the sibilant place contrast, posited by [12], at the dialect level.

3.2. /s/ vs. /ʃ/ variability across speakers within dialects

Figure 3 shows the distribution of speaker /s/ and /ʃ/ peak as offsets from dialect means, i.e. how much every speaker differs in their realization of /s/ and /ʃ/ from what is expected given their dialect (see Figure 1). The widths of the distributions correspond to the estimated degrees of variability in sibilant spectral peak, $\sigma_{s,/s/}$ and $\sigma_{s,/ʃ/}$ (Figure 2: triangles). A hypothesis test addressing **(2)** confirms that the 'wider' distribution visible for /s/ (dark grey) is indeed more variable than /ʃ/ (light grey) ($\sigma_{s,/s/} - \sigma_{s,/ʃ/}$: median = 0.018 ERB, 95% CredI = $[0.015, 0.020]$, $p_d = 1$). We also find a difference in variability by speaker gender: the difference in spectral peak between male and female speakers is larger for /s/ than for /ʃ/ ($\beta_{\text{gender:onset}}$: 2.0 ERB, 95% CredI = $[1.9, 2.2]$, $p_d = 1$); see [25].

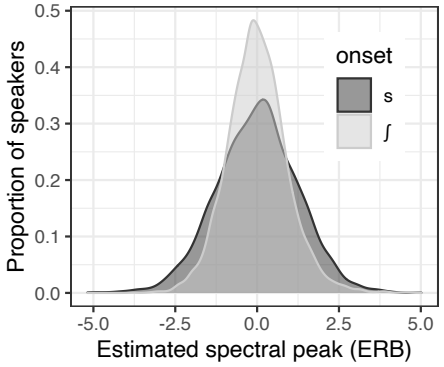


Figure 3: Distribution of estimated speaker offsets from dialect means (using posterior medians) for English /s/ and /ʃ/.

3.3. /s/ vs. /ʃ/ variability within speakers

Figure 2 (squares) addresses question **(3)**, showing the estimated amount of within-speaker variability for /s/ and /ʃ/, effectively capturing by-token variability for an 'average speaker' within an 'average dialect'. /s/ is generally more variable than /ʃ/, as confirmed by a hypothesis test ($\beta_{\sigma,\text{onset}}$: median = 0.08, 95% CredI = [0.04, 0.13], $p_d = 1$).

However, this result hides considerable variation across dialects and speakers. Figure 4 shows the estimated difference in within-speaker variability between /s/ and /ʃ/ for an 'average speaker' for *each* dialect (omitting the 3 dialects with $n < 250$). While most (16/27) dialects show greater variability for /s/ (95% CredI entirely positive), there is no clear difference in variability between /s/ and /ʃ/ in five, and in three, /s/ is *less* variable than /ʃ/ (95% CredI entirely negative). The five 'unclear' dialects may be inconclusive without more data; however, we can be sure that /s/ is not uniformly more variable than /ʃ/, at the token level in *all* dialects. As for variation within speakers, the actual predicted /s/-/ʃ/

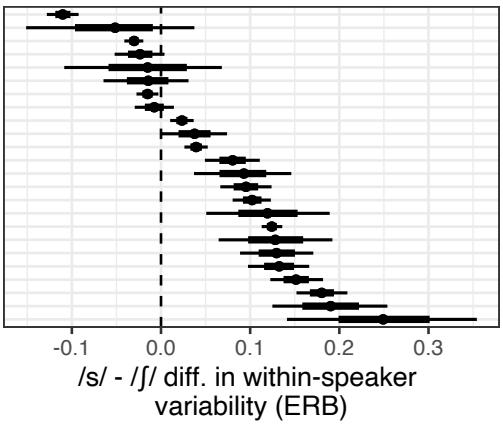


Figure 4: Estimated within-speaker variability between English /s/ and /ʃ/ by dialect.

variability differences are less informative because most speakers have sample size too small to say whether /s/ or /ʃ/ is more variable. Instead, we can examine the model's prediction for 'what proportion of speakers have /s/ variability > /ʃ/ variability?'. This fraction is 0.66 (95% CredI: [0.65, 0.68]), suggesting that /s/ is more variable at the token level for a majority of speakers, but not all.

4. DISCUSSION

This study considers acoustic invariance in English /s ʃ/ on a large scale. We find a robust psychoacoustic contrast between the sibilants across dialects, as consistently observed from previous lab studies [2, 20]. Whatever the precise articulations used across English, this sharp overall difference resonates with [26]'s assumption that the English phonological contrast exploits the abrupt acoustic consequences of a gradient shift in articulatory place and tongue shape, enhanced by lip-rounding.

Our main question is whether /s/ is more variable than /ʃ/ [7, 9], in a nationally, regionally, socially, and ethnically diverse sample of English dialect speakers. At the level of dialect, /s/ and /ʃ/ are equally variable. In contrast, /s/ is more variable than /ʃ/ across speakers within dialect, including gender differences being larger for /s/ [25]. Within speakers at the token level, /s/ is more variable than /ʃ/ on average, but not consistently across all dialects and speakers. At all levels we find greater variability in /ʃ/ than anticipated (c.f. [5]).

Thus, while we find that /s/ is 'more variable', the answer is more nuanced than expected. A major reason for this is our analytical choice of a non-linear, auditory scale (ERB, but we expect e.g. Bark would look similar) over Hz. In Hz, /s/ is more variable than /ʃ/ across dialects (Figure 1), and re-running our analysis in Hz also shows greater variability in /s/ across the board (as shown in the OSF project[1]). This is partly because differences at higher frequencies are exaggerated in Hz, an observational bias which is undone by using a non-linear scale, closer to auditory processing. Discrepancies between Hz and auditorily-scaled measures have been noted before [2, 12], but these discrepancies are especially large for research questions about *variance*, the focus here, rather than *means*, as in most previous work on sibilants. This has implications for sociophonetic research, which often uses Hz measures to infer social-indexical meanings, especially for gendered, social and ethnic identities e.g. [27, 28]. The broader point is that research questions about *variability* in any frequency measure will be greatly affected by the choice of scale, especially for higher-frequency (e.g. sibilant energy, stop bursts) versus lower-frequency spectral measures (e.g. vowel formants).

What are the sources of variation for greater across- and within-speaker variability in /s/? There are likely several factors, including greater susceptibility to coarticulatory, physiological and performative social-indexical variation [1, 7]. To this we can add additional factors in our sample, including speaker age, speech style, and further structured sociolectal variation which may be subsumed in the broad dialect groupings used here.

Finally, the relatively high variability in spectral peak for /ʃ/ compared to that of /s/ across dialects and speakers seems surprising. But perhaps our expectations about what English sibilants 'should' be like are skewed by theoretical attention on /s/ [29], and/or observational—and so theoretical—biases, unwittingly informed by previous smaller-scale studies on less heterogeneous dialects. Further corpus-phonetic work on these, and other sounds will help us to appreciate better the scope of variability in English sounds (cf. [30, 31, 32]).

5. REFERENCES

[1] S. Fuchs and M. Toda, "Do differences in male versus female /s/ reflect biological or sociophonetic factors?" in *Turbulent Sounds*, S. Fuchs, M. Toda, and M. Zygis, Eds. Berlin: de Gruyter, 2010, pp. 281–302.

[2] V. Evers, H. Reetz, and A. Lahiri, "Crosslinguistic acoustic categorization of sibilants independent of phonological status," *J. Phonetics*, vol. 26, no. 4, pp. 345–370, 1998.

[3] L. L. Koenig, C. H. Shadle, J. L. Preston, and C. R. Mooshammer, "Toward improved spectral measures of /s/: Results from adolescents,"

J. Sp. Lang. Hear. Res., vol. 56, no. 4, pp. 1175–1189, 2013.
[4] A. Baker, D. Archangeli, and J. Mielke, "Variability in American English s-retraction suggests a solution to the actuation problem," *Language Variation and Change*, vol. 23, no. 3, pp. 347–374, 2011.
[5] D. Recasens and C. Rodríguez, "A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound," *J. Phonetics*, vol. 59, pp. 58–75, 2016.
[6] M. Stevens and J. Harrington, "The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English," *J. Phonetics*, vol. 58, pp. 118–134, 2016.
[7] E. Levon, M. Maegaard, and N. Pharao, Eds., "The sociophonetics of /s/," *Linguistics*, vol. 55, no. 5, 2017.
[8] J. Stuart-Smith, "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Change in Phonology: Papers in Laboratory Phonology 9*, J. Cole and J. Hualde, Eds. Berlin: de Gruyter, 2007, pp. 65–86.
[9] P. Eckert and S. McConnell-Ginet, *Language and Gender*, 2nd ed. Cambridge: Cambridge University Press, 2013.
[10] P. Flipsen Jr., L. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic characteristics of /s/ in adolescents," *J. Sp. Lang. Hear. Res.*, vol. 42, no. 3, pp. 663–677, 1999.
[11] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *J. Acoust. Soc. Am.*, vol. 108, pp. 1252–1263, 2000.
[12] E. Chodroff and C. Wilson, "Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English," *Language*, vol. 98, no. 2, 2022.
[13] J. M. Scobbie and J. Stuart-Smith, "Socially-stratified sampling in laboratory-based phonological experimentation," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds. Oxford: Oxford University Press, 2012, pp. 607–621.
[14] A. Hughes, P. Trudgill, and D. Watt, *English Accents and Dialects*, 5th ed. London/New York: Routledge, 2013.
[15] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: de Gruyter, 2006.
[16] M. McAuliffe, A. Coles, M. Goodale, S. Mihuc, M. Wagner, J. Stuart-Smith, and M. Sonderegger, "ISCAN: A system for integrated phonetic analyses across speech corpora," in *Proc. 19th ICPhS*, Melbourne, 2019, pp. 1322–1326.
[17] "SPeech Across Dialects of English," 2018. [Online]. Available: spade.glasgow.ac.uk
[18] P. F. Reidy, "spectRum," 2013, R package.
[19] ——, "Spectral dynamics of sibilant fricatives are contrastive and language specific," *J. Acoust. Soc. Am.*, vol. 140, pp. 2518–2529, 2016.
[20] P. Boersma and S. Hamann, "The evolution of auditory dispersion in bidirectional constraint grammars," *Phonology*, vol. 25, no. 2, pp. 217–270, 2008.
[21] P.-C. Bürkner, "Advanced Bayesian multilevel modeling with the R package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.
[22] Stan Development Team *et al.*, "Stan modeling language user's guide and reference manual," 2019, version 2.29. [Online]. Available: https://mc-stan.org
[23] B. Nicenboim, D. J. Schad, and S. Vasishth. (2021) An introduction to Bayesian data analysis for cognitive science. [Online]. Available: https://vasishth.github.io/bayescogsci/book/
[24] L. A. Ciaccio and J. Veríssimo, "Investigating variability in morphological processing with Bayesian distributional models," *Psych. Bull. Rev.*, vol. 29, no. 6, pp. 2264–2274, 2022.
[25] J. J. Holliday, P. F. Reidy, M. E. Beckman, and J. Edwards, "Quantifying the robustness of the English sibilant fricative contrast in children," *J. Sp. Lang. Hear. Res.*, vol. 58, pp. 622–637, 2015.
[26] K. N. Stevens, "On the quantal nature of speech," *J. Phonetics*, vol. 17, no. 1, pp. 3–45, 1989.
[27] R. J. Podesva and S. Kajino, "Sociophonetics, gender, and sexuality," in *The Handbook of Language, Gender, and Sexuality*, S. Ehrlich, M. Meyerhoff, and J. Holmes, Eds. Wiley Online Library, 2014, pp. 103–122.
[28] N. Pharao, K. L. Appel, V. Wolter, and J. Thøgersen, "Raising of /a/ in Copenhagen Danish – Perceptual consequences over two generations," in *Proc. 18th ICPhS*, Glasgow, 2015.
[29] J. Stuart-Smith, "Changing perspectives on /s/ and gender over time in Glasgow," *Linguistics Vanguard*, vol. 6, no. s1, 2020.
[30] M. Y. Liberman, "Corpus phonetics," *Annual Review of Linguistics*, vol. 5, pp. 91–107, 2019.
[31] E. Chodroff, A. Golden, and C. Wilson, "Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization," *J. Acoust. Soc. Am.*, vol. 145, pp. EL109–EL115, 2019.
[32] J. Tanner, M. Sonderegger, J. Stuart-Smith, and J. Fruehwald, "Toward 'English' Phonetics: Variability in the Pre-consonantal Voicing Effect Across English Dialects and Speakers," *Frontiers in Artificial Intelligence*, vol. 3, 2020.

[1] We acknowledge funding from ESRC (ES/R003963/1) to Stuart-Smith; SSHRC (869-2016-0006, 435-2017-0925) and NSERC (RGPDD 501771-16) to Sonderegger; and NSF (SMA-1730479) to Mielke.
[2] Details of our data and analysis can be found in an OSF project at https://doi.org/10.17605/OSF.IO/XUBQM.

PhonD2: A DATABASE ON PHONOTACTIC STRUCTURES OF GERMAN DIALECTS

Alfred Lameli[1], Alexander Werth[2], Valeria Bunkov[2], Samantha Link[1]

Research Center Deutscher Sprachatlas, Philipps-Universität Marburg, Germany[1]
University of Passau, Germany[2]
lameli@uni-marburg,de, alexander.werth@uni-passau.de, valeria.bunkov@uni-passau.de, samantha.link@uni-marburg.de

ABSTRACT

We present the PhonD2 corpus, which is an open access online database on phonotactic structures of German dialects. The corpus is based on both the translation of sentences and free speech during interviews in 172 sites all over Germany. Data are annotated according to phonological and morphological criteria, e.g., sounds, CV-structure, morphemes. At present, the sub-corpus of translations is integrated, free speech will be added successively. The corpus focuses on syllable structures in dialects aiming at the description of geographical variation of syllable patterns at the phonology-phonetics interface and their typological classification. It shows, e.g., systematic regional differences in preferences for monosyllabic words, differences in how they arrange on the sonority scale as well as regional differences in the clustering of consonants in both onset and coda. The PhonD2 corpus thus opens up a systematic perspective on the clustering of sounds and on typological characteristics of German beyond standard language.

Keywords: Phonotactics, German dialects, syllable, sonority, CV phonology.

1. INTRODUCTION

Since the earliest days of dialectology, there has been an awareness of the fact that the realization of sounds may differ depending on their linguistic context. Such phonotactic conditions have always been taken into account in the relevant grammars of local dialects (e.g., [1]). However, they have not been systematically researched on a comparative regional scale. This is mainly due to the lack of data based on spoken dialects and suited for phonotactic questions. Although numerous corpora of spoken German language have been compiled in recent years (cf. [2]), they are either not regionally diversified or, if they are, they are not phonetically-phonologically specified.

This paper describes the planning and present implementation of the PhonD2 ("Phonotaktik der Dialekte in Deutschland") database, which is a database at phonotactic structures of German dialects. The focus of the PhonD2 database is on the regional documentation of syllable structures in Germany. For this purpose, data from both the translation of sentences and free speech during interviews documented in the *Marburg Phonetic Archive* (MRPhA) were selected from 172 sites all over Germany. First analyses based on this data indicate large-scale differences in the preference of syllable structures between dialect regions, which can be interpreted in terms of (micro)typological variation [3].

The aim of this paper is to introduce the PhonD2 database and to illustrate its linguistic potential. The remainder of the paper is as follows. Section 2 provides information on the PhonD2 data and its processing, section 3 takes a closer look on the data and section 4 concludes.

2. DATA AND DATA PROCESSING

2.1 Background

The data used for the PhonD2 database were selected from the *Marburg Phonetic Archive* (MRPhA), which is a largescale archive of dialectal sound recordings ranging from the first half of the 20th century to the present. The archive is maintained by the *Research Center Deutscher Sprachatlas* (DSA, Marburg/Germany).

Initially, the PhonD2 project made use of the MRPhA selection chosen by Göschel and colleagues in preparation of the *Phonetischer Atlas der Bundesrepublik Deutschland* (PAD; cf. [4, 5]). This selection documents the speech of so-called NORMs and NORFs (non-mobile, old, rural (fe)males), which were recorded between 1956 and 1996 by phonetically trained DSA explorers. The data collection was aimed at determining the oldest dialect forms known by these persons. For this purpose, translation tasks of Wenker sentences were carried out. The Wenker sentences are a standard instrument in German dialectology [6] consisting of 40 sentences plus additional words (e.g., numbers, weekdays).

Göschel selected a total of 184 recordings each of which was transcribed in parts during the 1980s and 1990s in a very narrow IPA notation supported by quantifying control measures to ensure the highest possible transcription quality (cf. [7]). Even though these transcriptions have been completed, the atlas remained unpublished. The data was released to the public for further analyses and have been used several times in dialectometric studies since then (e.g., [8, 9, 10, 11, 12]).

The PhonD2 corpus makes use of this PAD selection and its transcriptions of Wenker sentences. In order to be able to consider a further speech situation, additional recordings of the same persons from the MRPhA were integrated into the corpus. These are interviews on biographically relevant topics that were also recorded during the original data exploration. The monologues that developed from this, interrupted by short queries, form an excellent documentation of the free speech of these persons. However, since the documentation of free speech is not available for all PAD informants, the number of sites was reduced for the PhonD2 project. The PhonD2 corpus thus contains a total of ca. 20 hours of spoken language from 172 sites.

2.2 Workflow

In order to compute both a phonetic segmentation and labeling of the recordings WebMAUS [13] is used. We use WebMAUS' standard German language model, but enhance the orthographic notation required for it with dialectal phenomena and phenomena of spoken language such as clitization or sentence and word interruptions. This approach significantly improves the alignment of the dialect data. Further steps are the production of both a phonological and morphological representation as well as the implementation of language data and meta-data in the PhonD2 database.

2.2.1 Phonological representation

In order to prepare the phonotactic documentation, the PAD transcriptions must be modified. Take, for example, the notation [ˈmɔɪ̯ɪ̯də] 'tired' from the PAD corpus. Two essential work steps are involved here, namely the normalization of the transcript and the identification of syllables. The IPA transcriptions are thus assigned to phonological types that could be derived from the research literature. The identification of syllables is processed by a rule-based algorithm, the results of which are nevertheless checked by visual inspection. In the given case this procedure results in the representation [ˈmɔɪ̯.də].

The parts of the Wenker sentences, which have not been prepared during the PAD work, were directly transcribed. The same holds for the data from the situation of free speech context. Since some dialects are difficult to understand, we use a network of native dialect speakers to assist us in preparing the orthographic transcripts. We also do not transcribe the entire monologues, but only individual words, usually nouns, verbs, adjectives and adverbs.

The basis for the comparability of the transcriptions is, on the one hand, continuous training of the transcribers (including the comparison and control of individual transcriptions), and, on the other hand, a constantly growing catalog of regional features compiled in the course of the project.

Using these notations, a phonological representation that distinguishes plosives (P), affricates (A), fricatives (F), nasals (N), liquids (L), glides (G), short vowels (V), long vowels (V:) and diphthongs (VV) is automatically derived and implemented into the PhonD2 database (e.g., [ˈmɔɪ̯.də] ~ NVV.PV) as well as a less detailed CV representation ([ˈmɔɪ̯.də] ~ CVV.CV).[1] Furthermore, an indication of sonority with 1 = vowel, 2 = glide, 3 = liquid, 4 = nasal, 5 = fricative, 6 = plosive is implemented ([ˈmɔɪ̯.də] ~ 51.61) together with the information on strong and weak syllables ([ˈmɔɪ̯.də] ~ s.w). In addition, primary word accents and Middle-Franconian tone accents are annotated where necessary.

The classifications are sometimes difficult to implement, for example, with regard to the question as to whether a certain sound has to be defined as a glide or a fricative. From this point of view, the classification is to be understood as a suggestion (based on reference literature such as [14]), which can be modified in the context of individual data processing.

2.2.2 Morphological representation

In an earlier study on monosyllables [12], a dependency between PoS and phonotactic structure has been found for German dialects, in that some PoS do systematically prefer or disfavor certain syllable structures. Consequently, the PhonD2 data allows to systematically explore the interface between phonotactics and morphology. Based on STTS tagset [15], we automatically assigned PoS using TreeTagger [16]. To avoid inaccuracies, the tagged PoS are corrected manually. Tags not available in STTS are individually assigned based on the relevant literature on speech classifications for German (e.g., [17]). Usually, a word form is assigned to exactly one PoS; but some contexts require a word form assignment to more than one PoS (e.g., *mit* 'with' as a preposition vs. *mit* 'with' as a particle as in *mitgehen* 'to go along'). In such cases multiple individual tags

were provided (e.g., *mit* is assigned to "APPR" and "PTKVZ"). Apart from PoS, morphemes of the German dialects are segmented and identified manually. The corpus scheme for morphemes consists of (word-)stem, affixes, vowel mutation and vowel gradation, diminutives and suppletive forms.

2.3 Online access

All data will be available open access on the Internet.[2] At present (01/2023), the PhonD2 site provides an overview of several items of the Wenker sentences together with a map of the survey locations. Other data will be integrated successively. In addition, a mapping function illustrates the regional distribution of syllable types, for which Figure 2 and Figure 3 provide examples. These maps are intended only as a small glimpse into the geographical dimension of the data. For exhaustive analyses, the data must be processed individually. We do not achieve full coverage of the 172 sites in all cases; these instances are nevertheless included, provided they can be analyzed in phonotactic terms.

Additionally, the web site provides a first overview on sound frequencies regarding both individual sounds and sound combinations. In this regard, Figure 1 gives an impression of tri-grams with a liquid [l] core together with information on their frequency in the data.

trigrams	freq
blɐ	1
blɑ	1
ble	42
blə	2
blɛ	106
bli	62
blɪ	140
bln	6
blø	1
blœ	2

Figure 1: Extract of tri-grams with a liquid [l] core in the PhonD2 tri-gram table.

After completion of the project, the data can be downloaded in its entirety as a *.csv file. Until then, all data presented on the Internet must be considered preliminary. Nevertheless, they can be used for linguistic analyses.

3. DATA INSPECTION

3.1 Cartographic representations

Figure 2 illustrates the geographical distribution of syllable types of the standard German noun *Affe* 'ape' in the PhonD2 database. The map shows clear regional patterns in both the distribution of monosyllables vs. bisyllables (e.g., VC vs. V.CV ~ [af] vs. [ˈa.fə]) and the organization of syllables within those types (e.g., VC vs. V:C ~ [af] vs. [a:p]; glottal stops are not reported). It appears that monosyllables are preferred in the North and in the lower half of Germany, while bisyllabic realizations occur almost exclusively in a central strip stretching from West to East.

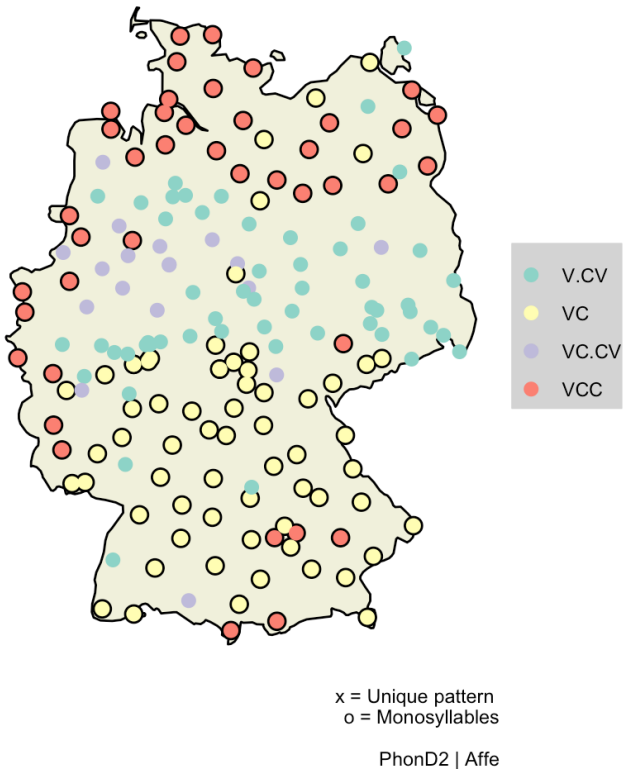


Figure 2: Regional distribution of syllable structure for the word *Affe* 'ape' from the PhonD2 corpus.

In Figure 2, several phonological processes come together, namely schwa apocopation ([af] vs. [ˈa.fə]), which indicates preference for monosyllabic or bisyllabic syllables and monosyllabic lengthening ([af] vs. [a:p]) [18, 19]. Against this background, the phonotactic map provides a linguistic systematization on a higher level; it classifies linguistic patterns with respect to more general structural characteristics.

Figure 3 represents the standard German adverb *genug* 'enough'. Again, there are preferences for monosyllabic forms in the North and in the South, even though the zone with bisyllabic forms between them is much more expanded. In this case, however, monosyllables are not attributed to schwa

apocopation, but schwa syncopation (e.g., [ˈknʊɐ̯] ~ CCVV or [ˈknʊːx] ~ CCV:C) or the loss of the *ge-*syllable ([ˈnɔx] ~ CVC or [ˈnaʊ̯x] ~ CVVC). The latter is interesting from the perspective that *ge-* is usually used as a prefix in German (e.g., past participle *gekannt* 'known'), which typically is lost in, for example, Bavarian dialects (*Ø-kent*). In the adverb *genug*, *ge-* is, however, no morphological marker thus indicating a phonological process driven by analogy. Regarding the distribution of syllable types, the pattern known from the *Affe* map is mirrored here to some extent.

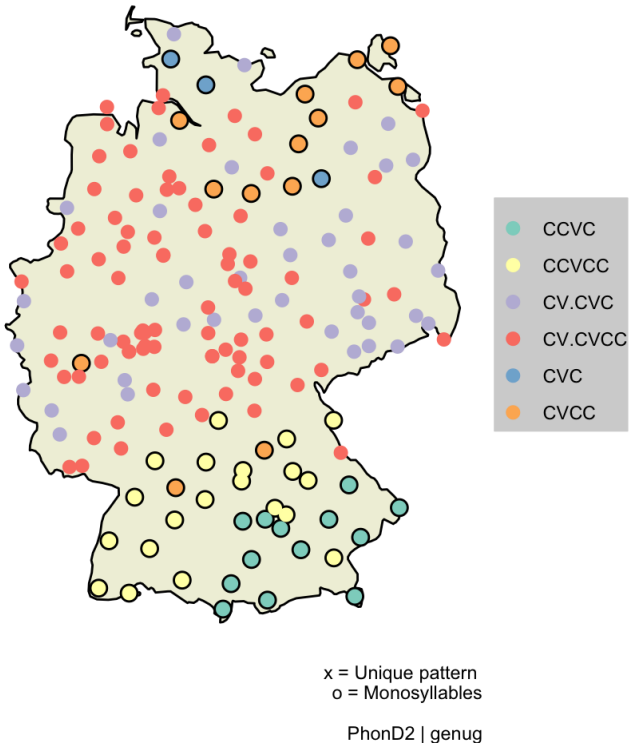


Figure 3: Regional distribution of syllable structure for the word *genug* 'enough' from the PhonD2 corpus.

At the same time, Figure 3 shows one of the rare examples, where onset and coda of a word are very closely arranged on the sonority scale as in [ˈgnʊŋk] ~ PNVNP or [ˈgnuːŋk] ~ PNV:NP [14]. With this pattern in mind, schwa syncopation can be understood as an optimization of syllable structure toward the most consistent sonority arrangement in the onset, supported by the integration of a velar nasal in the coda. From this point of view, we are not only dealing with the sonority optimization of the syllable, but the sonority optimization of the (phonological) word. As can be seen, this process is restricted to the Upper German dialect region (= yellow and green color). Whether this is a typological characteristic of Upper German dialects must be left to further analysis.

3.2 Data aggregation

In addition to single word representations, the aggregate evaluation of phonotactic structures is of interest. Figure 4 shows Finite State Automata (Markov Chains) for three subsamples of the PhonD2 corpus ([3]), representing the transition probabilities between consonants (C) and vowels (V) in monosyllables.

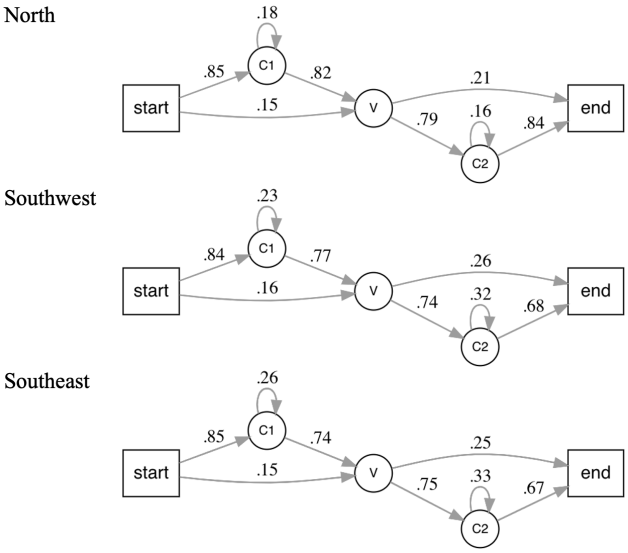


Figure 4: Finite State Automata for regional PhonD2 subsamples documenting CV transition probabilities in monosyllables

The Figure shows that in all regions, the probability that a monosyllable starts on C is about 85 %. However, then regional differences in the replication rate of the C position in Onset (C1) become apparent: 18 % in the North, 23%, 26 % in the South. It is similar for the replication rate in the coda (C2): 16 % in the North, 32 %, 33 % in the South. What the PhonD2 data here more generally shows is a regional difference in the syllable complexity of German dialects in the form of a South-North divide: more C clusters in the South, fewer C clusters in the North.

4. SUMMARY

We report on the PhonD2 database, which is a database on syllable structures of German Dialects. The phonotactic account enables a comprehensive view on more general structural characteristics of dialects, both with regard to individual phonetic characteristics, but also with regard to the typological structure of the dialects. The database is constantly being expanded. All data are freely accessible.

Acknowledgements
This research is funded by the German Research Foundation (DFG, grant 432304149). We are grateful

to two anonymous reviewers and to the discussants at project presentations in Marburg, Frankfurt, Vienna, and Salzburg. Thanks also go to both our student workers and our transcribers in the language regions.

5. REFERENCES

[1] Winteler, J. 1876. *Die Kerenzer Mundart des Kantons Glarus. In ihren Grundzügen dargestellt*. Winter.
[2] Kupietz, M., Schmidt, T. (eds) 2018. *Korpuslinguistik*. De Gruyter.
[3] Lameli, A. 2022. Syllable Structure Spatially Distributed: Patterns of Monosyllables in German Dialects. *Journal of Germanic Linguistics* 34, 241–287.
[4] Göschel, J. 1992. *Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas"*. Forschungsinstitut für deutsche Sprache.
[5] Göschel, J. 2000. Der Phonetische Atlas von Deutschland. *Јужнословенски Филолог* 56, 283–288.
[6] Chambers, J. K., Trudgill, P. 1998. *Dialectology. Second edition*. CUP.
[7] Almeida, A., Braun, A. 1986. "Richtig" und "Falsch" in phonetischer Transkription. Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik* 53, 158–172.
[8] Nerbonne, J. Siedle, C. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72, 129–147.
[9] Nerbonne, J. 2010. Mapping aggregate variation. In: Lameli, A., Kehrein, R., Rabanus, S. (eds), *Language and space. An international handbook of linguistic variation, vol. 2: Language mapping, part I*. De Gruyter Mouton, 476–495.
[10] Prokić, J, Çöltekin, C., Nerbonne, J. 2012. Detecting shibboleths. In: Butt, M., Carpendale, S., Penn, G., Prokić, J., Cysouw, M. (eds), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH: Visualization of linguistic patterns and uncovering language history from multilingual resources*. The Association for Computational Linguistics, 72–80.
[11] Prokić, J. 2017. Quantitative diachronic dialectology. In: Wieling, M., Kroon, M., van Noord, G., Bouma, G. (eds), *From semantics to dialectometry. Festschrift in honor of John Nerbonne*. College Publications, 293–301.
[12] Lameli, A., Werth, A. 2017. Komplexität und Indexikalität. Zum funktionalen Gehalt phonotaktischer Wortstrukturen im Deutschen. Hennig, M. (eds), *Linguistische Komplexität – ein Phantom?* Stauffenburg, 73–96.
[13] Kisler, T., Reichel U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
[14] Hall, T. A. 1992. *Syllable structure and syllable-related processes in German*. Niemeyer.
[15] Schiller, A., Teufel, S., Stöckert, C. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. University of Tuebingen.
[16] Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
[17] Hentschel, E., Weydt, H. 2021. *Handbuch der deutschen Grammatik*. de Gruyter. 5. Ed.
[18] Schirmunski, V. 2010 [1962]. *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Lang.
[19] Lameli, A. 2022. Remarks on the consistency of schwa apocope in the geography of German dialects. In: Nevaci, M., Floarea, I, Farcaş, J-M (eds), *Ex Oriente lux. In honorem Nicolae Saramandu*. Edizioni dell'Orso. 683–702.

[1] We differentiate between V: and VV in order to preserve the phonetic basis. Of course, unification can be done post hoc in the sense of CV representation (V:, VV = VC).
[2] http://www.dsa.info/phond2.

ARTICULATION RATE IN CONSONANTS AND VOWELS: RESULTS AND METHODOLOGICAL CHALLENGES FROM A CROSS-LINGUISTIC CORPUS STUDY

Roger Yu-Hsiang Lo, Márton Sóskuthy

Department of Linguistics, University of British Columbia
roger.lo@ubc.ca, marton.soskuthy@ubc.ca

ABSTRACT

We investigate how articulation rate is implemented by different segment types such as consonants and vowels using corpus data from eight unrelated spoken languages. Our study also sheds light on methodological issues that arise in the analysis of cross-linguistic corpus data. With forced-aligned data from 20-40 speakers for each language, we observe a robust effect of articulation rate on segment duration, such that vowels undergo a significantly greater extent of duration adjustment than consonants. However, the size of this effect varies across languages in ways that may be partially due to alignment quality and cross-linguistic differences in segment distributions. We propose methods for accounting for these factors and observe that bringing them into our model does indeed change some of the language-specific conclusions. We further discuss the cross-linguistic variation in our data in terms of language-specific phonetic implementation, and speculate on its implications for sound change.

Keywords: articulation rate, corpus, cross-linguistic, generalized additive mixed model

1. INTRODUCTION

Speech rate is an important and well-documented dimension of linguistic variation [1, 2, 3, 4, 5, 6]. We focus on *articulation rate*, based solely on segmental material to the exclusion of pauses, as opposed to *speaking rate*, which includes pauses as well. Articulation rate varies across age [7], gender [3], and dialect [8], and even within individual speakers and utterances [5]. However, much remains unknown about the implementation of changes in articulation rate. To this end, we ask the following question: how is variation in articulation rate reflected in consonant and vowel durations? It is reasonable to expect differences between consonants and vowels due to their articulatory and aerodynamic properties. An experimental study reports that the ratio of vocalic to consonantal material increases as articulation rate slows down [4] (though see [6]). In other words, they find that vowels stretch more than do consonants in slower speech. Here, we use speech corpora from multiple languages to address this research question. This allows us to test the universality of any observed effects, and also to look at variation across languages.

Corpus data come with significant challenges. The amount of speech data renders manual processing impractical and requires automated methods. The accuracy of these methods is often affected by recording quality. Therefore, when comparing corpora that were recorded under different conditions, it is crucial to explicitly account for quality differences. This is relevant for the current study, as we are interested in exploring cross-linguistic variation, but our corpora differ substantially in terms of their recording quality (see Section 2.1).

Cross-linguistic differences in linguistic structures and their use may also introduce confounds into statistical models. Of particular relevance to the current study is the fact that the precise make-up of broad categories such as consonants and vowels varies across languages: for instance, a given language may have a higher proportion of stops than another language. There may also be differences in the extent to which these finer-grained segment types respond to articulation rate (e.g., stops may be less stretchy than other consonants). Thus, any observed cross-linguistic differences at a broader level may simply be artifacts of different frequency distributions at a lower level (e.g., language A has more stops, which makes its consonants appear generally less stretchy). It may therefore be important to control for such differences across languages.

The current study attempts to answer the main research question while mitigating the aforementioned issues via statistical modeling. Thus, we look at the differential roles of consonant and vowel durations in carrying articulation rate, and explore variation across languages. At the same time, we also highlight the extent to which recording/alignment quality and varying segmental frequency distributions can

distort the results of cross-linguistic corpus studies. This allows us to strengthen our results, and also provides important pointers for future work on cross-linguistic corpus data.

2. METHODOLOGY

2.1. Corpora

Our data come from three databases that contain speech samples from eight genetically unrelated languages: Korean and Mandarin (the Origins of Patterns in Speech or OoPS-Lab corpus, created in-house); Amharic, Georgian, Swahili, Turkish, and Vietnamese (IARPA; [9]); and English (Buckeye Speech Corpus; [10]).

Our own OoPS-Lab corpus consists of read and spontaneous speech from 20 native speakers per language.[1] All speech data was collected remotely on speakers' computers or mobile devices via a dedicated website and transcribed by undergraduate research assistants. Each language contains 2-3 hours of speech data, and the recordings are of reasonably high quality.

The Intelligence Advanced Research Projects Activity (IARPA) corpus was created to develop speech recognition technology for noisy telephone conversations [9]. As a result, the recording quality for these corpora is substantially lower. To mitigate this issue, we only include telephone conversations from a home or office environment from a total of 40 speakers per language, but quality issues remain (more on this in Section 2.2). We analyse a total of 1-2 hours of data per language.

The Buckeye Speech Corpus [10] contains high-quality spontaneous speech recordings from 40 speakers (20 female and 20 male) from Columbus, Ohio. Importantly, given that the phonetic alignments were manually checked, we expect the alignments to be more accurate than the other corpora used in this study. We can therefore use it to rule out the possibility that our results from the other corpora are artifacts of the forced-alignment process.

For the recordings from the OoPS and the IARPA corpora, we used the Montreal Forced Aligner [11] to obtain alignments. All durations were extracted using the PolyglotDB [12] corpus management package. We analyse a total of 1.429 million segments (median of 89,547 per language) representing 35,483 utterances (median of 2,822 per language). Analyses and visualisations were generated using R [13]. All data and code are available at https://bit.ly/3GVjt0J.

2.2. Analyses

Articulation rate is operationalized as average segment duration in seconds in an utterance (a stretch of speech surrounded by at least 150 ms of pauses on each side). Note that average duration is the *inverse* of rate: a 1 s utterance with 5 segments has a segment rate of $5/1 = 5$ and an average segment duration of $1/5 = 0.2$. We define this measure at the level of segments (as opposed to, e.g., syllables) as our dependent variable—the duration of individual segments—is also at this level.

Our research question is operationalized as follows: do the durations of consonants and vowels change at different rates as average segment duration (i.e., inverse articulation rate) changes? For instance, when the average segment duration is short, consonants and vowels may have the same duration; but when the average segment duration is high, vowels may be 1.5 times longer than consonants. This should correspond to diverging consonant *versus* vowel durations when plotted over average segment duration in log-log space.[2]

This question is tested using generalized additive mixed models (GAMMs; [14]) with **segment duration** as the outcome and **C/V** (consonant vs. vowel), **average segment duration** and their interaction as fixed effects. We include random smooths over **average segment duration** by **speaker** $\times$ **C/V** (i.e. separate groups for each combination of **speaker** and **C/V**) and by **language** $\times$ **C/V**. This allows the **C/V** effect to vary across speakers and languages. The online materials include the full model structure. We fitted one model as described above (baseline model); one that also controls for alignment quality (alignment model); and one that controls for alignment quality and different segment distributions across languages (alignment + segment model).

To quantify the alignment quality of the recordings, we randomly sampled 20 aligned segments for every speaker. A research assistant judged each alignment to be "OK" or "misaligned", where misalignment meant that more than 50% of the aligned interval was placed incorrectly. This measure focuses on gross misalignment, not fine-grained errors, which are more subjective and harder to identify. The proportion of "OK" alignments for individual speakers are shown in Figure 1A. These proportions were passed through a logistic function with a midpoint of 0.6 and a growth rate of 15 (Figure 1B). The resulting weights were then used to run a weighted GAMM, where each data point contributes to the model in accordance with their weight. This means

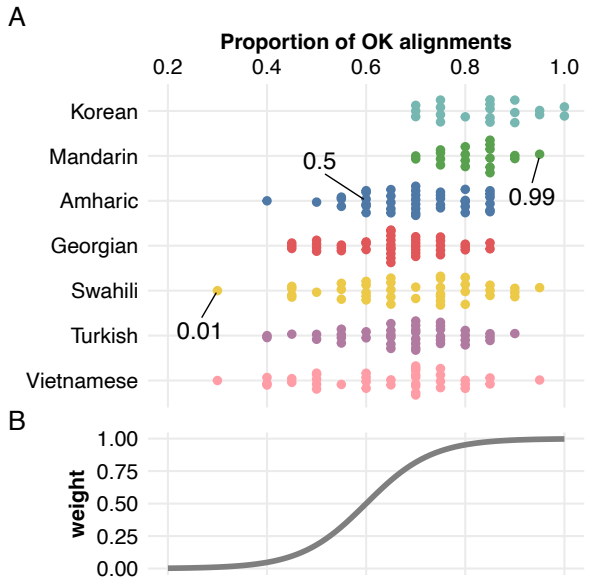


Figure 1: (A) Proportions of alignments judged "OK" across languages. Each point represents a speaker. (B) The mapping between proportions and model weights. The weights of three speakers are also annotated in the top panel.

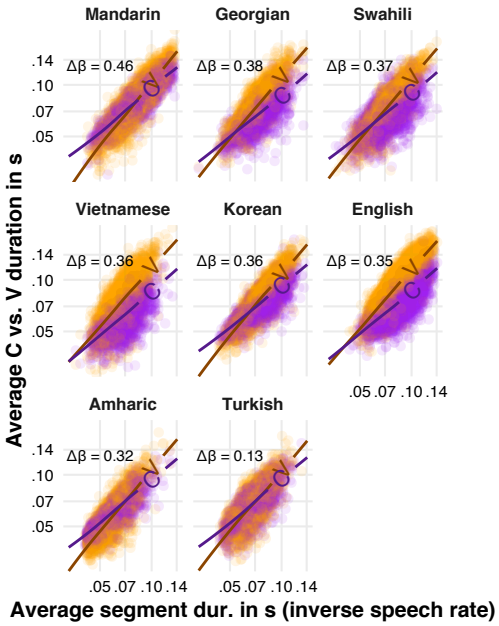


Figure 2: Average vowel and consonant durations as a function of average segment duration for each language. The dots are within-utterance averages. The lines are model predictions from the alignment + segment GAMM. Both axes are on the log scale.

that data points from speakers with lower alignment quality have less of an influence on model estimates.

Our third model controls for differing segment distributions across languages. While the proportion of vowels is approximately the same across all languages, the proportions of consonant types tend to vary more; for instance, Vietnamese has more nasals than all the other languages, while Amharic has more stop consonants. We bring this variation into the model by including a random smooth over **average segment duration** by **segment type**, where **segment type** has the levels STOP, AFFRICATE, FRICATIVE, LIQUID, NASAL, APPROXIMANT, HIGH VOWEL, and NON-HIGH VOWEL.

3. RESULTS

We first show the results from the alignment + segment model. Figure 2 shows model predictions plotted over data aggregated at the utterance level, with average consonant and vowel durations shown along the *y*-axis, and average overall segment duration along the *x*-axis. In all languages, vowels scale more readily with articulation rate than do consonants—that is, vowels are stretchier. These differences are significant across all languages. However, the effect sizes vary: for instance, Turkish shows only a small effect compared to the other languages. To compare the consonant and vowel slopes more holistically, we derive a metric called Δ effect size ($\Delta\beta$), defined as the difference in derivatives for the consonant and vowel prediction curves estimated at the median articulation rate. A bigger $\Delta\beta$ indicates a more pronounced difference in stretchiness. Turkish has the lowest $\Delta\beta$ (see Figure 4).

Figure 3 shows the results from the three different models for Amharic. For Amharic, bringing alignment quality into the model only makes a small difference, but controlling for segment distributions leads to a much higher effect size. Figure 4 shows the changes in $\Delta\beta$ across the three models for all the languages. Weighting by alignment quality tends to lead only to modest changes in the size of the effect (with the exception of Georgian). However, some effect sizes shift markedly once we bring variation in segment distribution under control. One noticeable difference between the baseline model and the alignment + segment one is that while the former shows a wide range of variation in effect sizes, the latter shows a tight cluster with only Turkish and Mandarin as outliers. Controlling for non-phonetic sources of variation in the data removes a great degree of apparent cross-linguistic variation.

4. DISCUSSION AND CONCLUSION

Our findings echo the work of [4], showing that vowels are stretchier than consonants: as articulation rate decreases, the proportion of an utterance occupied by vowels tends to increase. This pattern is stable

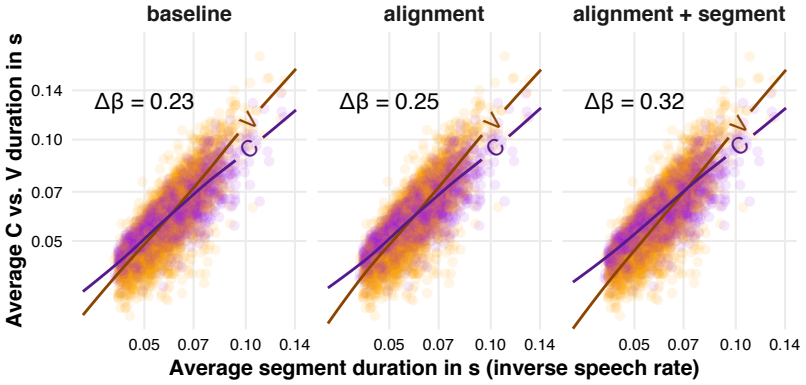


Figure 3: Differences in effect sizes between average consonant and vowel durations across models for Amharic. Each dot in the background represents an original data point. The solid lines show model predictions for the two segment categories. Note that both axes are on the log scale.

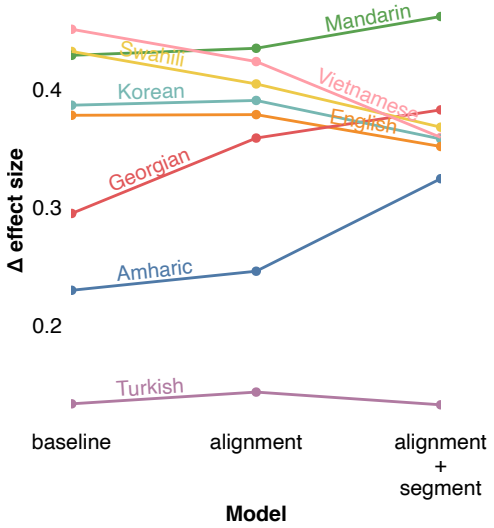


Figure 4: Changes in Δ effect size ($\Delta\beta$) across models for each language.

even after controlling for alignment quality and segment types. While all eight languages show a robust pattern, the size of the effect varies across languages. The relative ranking of different languages also changes across models with different degrees of control, and removing external sources of variation leads to more uniform estimates.

The differences in the way consonants and vowels carry articulation rate can be partly explained by reference to their articulation. Airflow is constricted for consonants, resulting in more aerodynamic and coordinatory complexity. To give an example, stops with closure voicing are impossible to sustain indefinitely due to the pressure build-up behind the constriction. It seems plausible that the added complexity of consonants makes them less responsive to changes in articulation rate.

Even after controlling for differences in segment distributions, languages do differ with respect to our key finding, suggesting that the same segment types (e.g., stops) can be more or less stretchy in different languages. This can potentially be attributed to subtle cross-linguistic differences in the phonetic implementation of the "same" sound (cf. [15]). For instance, [16] finds that English and Japanese /s/ exhibit differences in their spectral-temporal properties. Sundara [17] finds differences between the coronal stops of Canadian English and Canadian French in voice onset time, burst intensity and spectral shape. We speculate that at least some of the differences in our sample may be due to such fine-grained differences in phonetic implementation. This could be tested by taking a more detailed look at duration variation across different segment types in different languages, which, however, is outside the scope of this paper.

How articulation rate modulates the phonetic realization of segments also has implications for sound change. For instance, segments that are particularly inflexible in terms of their duration may be prone to 'catastrophic failure' under high articulation rate, that is, full deletion or reduction. To give another example, length contrasts may also be more susceptible to neutralization at high articulation rates in segments with limited durational maneuverability.

Finally, it is important to emphasise that the study conclusions do change as data quality and non-phonetic differences across languages are brought under statistical control. Cross-linguistic corpus phonetics [18, 19] has seen a marked surge in recent years. Our findings testify to the importance of accounting for systematic differences across data sources in cross-linguistic big data approaches. Statistical techniques such as the differential weighting of data points and random effects provide powerful tools for achieving this goal.

5. REFERENCES

[1] D. Byrd and C. C. Tan, "Saying consonant clusters quickly," *Journal of Phonetics*, vol. 24, no. 2, pp. 263–282, 1996.

[2] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 101–112, 1990.

[3] E. Jacewicz, R. A. Fox, and L. Wei, "Between-speaker and within-speaker variation in speech tempo of American English," *The Journal of the Acoustical Society of America*, vol. 128, no. 2, 2010.

[4] V. A. Kozhevnikov and L. A. Chistovich, *Speech: Articulation and perception*. Washington, D.C.: Joint Publications Research Service, 1965.

[5] J. L. Miller, F. Grosjean, and C. Lomanto, "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," *Phonetica*, vol. 41, no. 4, pp. 215–225, 1984.

[6] S. Wood, "What happens to vowels and consonants when we speak faster?" *Working Papers in Linguistics, Lund University*, vol. 9, pp. 8–39, 1973.

[7] H. Quené, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1104–1113, 2008.

[8] M. P. Robb, M. A. Maclagan, and Y. Chen, "Speaking rates of American and New Zealand varieties of English," *Clinical Linguistics & Phonetics*, vol. 18, no. 1, pp. 1–15, 2004.

[9] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, "Developing speech recognition systems for corpus indexing under the IARPA Babel program," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6753–6757.

[10] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. Buckeye Corpus of Conversational Speech (2nd release). Columbus, OH.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of INTERSPEECH 2017*, 2017, pp. 498–502.

[12] M. McAuliffe, E. Stengel-Eskin, M. Socolof, and M. Sonderegger, "Polyglot and Speech Corpus Tools: A system for representing, integrating, and querying speech corpora," in *Proceedings of INTERSPEECH 2017*, 2017, pp. 3887–3891.

[13] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[14] S. N. Wood, *Generalized additive models: An introduction with R*, 2nd ed. Boca Raton, FL: CRC Press, 2017.

[15] J. B. Pierrehumbert, "What people know about sounds of language," *Studies in the Linguistic Sciences*, vol. 29, no. 2, pp. 111–120, 1999.

[16] P. F. Reidy, "Spectral dynamics of sibilant fricatives are contrastive and language specific," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2518–2529, 2016.

[17] M. Sundara, "Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1026–1037, 2005.

[18] M. Y. Liberman, "Corpus phonetics," *Annual Review of Linguistics*, vol. 5, no. 1, pp. 91–107, 2019.

[19] E. P. Ahn and E. Chodroff, "VoxCommunis: A corpus for cross-linguistic phonetic analysis," in *Proceedings of the 13th Conference of Language Resources and Evaluation*, 2022, pp. 5286–5294.

[1] One female Taiwan Mandarin speaker was excluded from analysis due to poor recording quality.

[2] We look at the effects in log-log space as it is possible to observe diverging raw durations that maintain a constant ratio.

PEASYV: A PROCEDURE TO OBTAIN PHONETIC DATA FROM SUBTITLED VIDEOS

Adrien Méli[1], Nicolas Ballier[2]

[1]Université Paris Cité, CLILLAC-ARP, [2] Université Paris Cité, CLILLAC-ARP & LLF
adrienmeli@gmail.com, nicolas.ballier@u-paris.fr

ABSTRACT

This paper presents a procedure to easily obtain phonetic data from subtitled YouTube videos in English. Called PEASYV (Phonetic Extraction and Alignment of Subtitled YouTube Videos) and based on a UNIX shell script, this procedure makes it possible to obtain fine-grained phonetic and phonological data from any subtitled videos posted on YouTube. It uses two aligners, P2FA [1] and SPPAS [2], to generate Praat [3] TextGrids for each video. These TextGrids contain tiers specific to each aligner, with segmental, syllabic and lexical alignments. SPPAS tiers also contain prosodic information. After describing the procedure and the generated files, a case study of a YouTube channel on English pronunciation is presented, in order to assess the quality of the alignment: the formants of the monophthongs from the channel's 452 videos, amounting to more than 179 hours of spontaneous speech, are analysed. The potential alignment errors, along with their prevention, are then discussed.

Index Terms: English vowels, corpus phonology, formant extraction, automatic alignment

1. INTRODUCTION

Several forced aligners have been developed, among others MAUS (Munich AUtomatic Segmentation) [4] and WebMAUS [5], Prosodylab-aligner [6], LaBB-CAT [7] and the Montreal Forced Aligner (MFA) [8] (see [9] for a comparison of MAUS, FAVE, LaBB-CAT and MFA and [10] for a comparison of FAVE, MAUS and MFA and for a discussion of the acoustic models used in the forced aligners). Using manually aligned data from Trinidadian English [10] has compared Forced Alignment and Vowel Extraction (FAVE), Munich Automatic Segmentation (MAUS), and the Montreal Forced Aligner (MFA) and their performances in automatically segmenting Trinidadian speech. Forced aligners presuppose a text when deep neural network acoustic modeling is available for YouTube video transcription [11]. Several corpora have been made available as text transcriptions of YouTube videos [12]. We wish to explore the possibility of exploiting this automatic generation of captions for videos in English on Youtube to capture phonetic datasets online.

The emergence of tools such as YouGlish[1], that acts as a concordancer on a selection of the subtitled videos posted on YouTube, may be used for the observation of spontaneous spoken data in an ecological environment or at least as an alternative to the observer's paradox [13]. As time-stamped transcriptions of speech, subtitles have the potential to serve as data for phonetic and phonological purposes. The issue arises of the reliability of such a source. This study examines whether subtitle time stamps can be used for finer-grained segmental alignment. It focuses on English monophthongs because they can be easily represented in the F_1/F_2 space: the accuracy of the alignment can therefore be visually checked. The rest of the paper is organised as follows. Section 2 presents an outline of the procedure. Section 3 offers a case study investigating the reliability of PEASYV by visually inspecting the vocalic productions of a female speaker on a YouTube channel teaching British pronunciation. Section 4 discusses the results and explores future venues of research.

2. DESCRIPTION

This section outlines the workflow (Section 2.1) to align the subtitled video automatically, and then describes the resulting files (Section 2.2).

2.1. Workflow

PEASYV can be installed on any computer running a major Linux distribution. Execution on Windows or MacOS has not been tested. The following programmes are required: `yt-dlp` [14], `ffmpeg` [15], `R` [16], `Praat` [3], `SPPAS` and `P2FA`. The workflow is controlled by a UNIX-compatible shell script that takes a text file as argument. Each line of the text file contains a link to a subtitled YouTube video. Optional fields on the same line are the speaker's gender and a description of the variety of English they speak. The script then loops over each line of the

text file and executes a series of subscripts to perform the actions in the list below.

The script:

1. Downloads the video and its subtitles using `yt-dlp`;
2. Converts the video to a main 11kHz mono `.wav` file;
3. Creates a main TextGrid and a main PitchTier for that sound file;
4. Uses the time stamps of the downloaded subtitles to create matching intervals on a tier in the TextGrid;
5. Prints the subtitles in the corresponding intervals on the tier;
6. Splits the main files (*i.e.* the main sound file, the TextGrid and the PitchTier) into subfiles of the same types;
7. Inputs those subfiles into the two aligners, SPPAS and P2FA, successively;
8. Reintegrates the automatically aligned subfiles into the original main TextGrid

The main sound file is converted to a single-channel sound file sampled at 11,025Hz to comply with the requirements of the aligners. The PitchTier is generated following the procedure described in Hirst [17], which uses the interquartiles values obtained from Praat's default values. These interquartile values are obtained using a separate *R* script. Regardless of whether this extra step is carried out, the obtained prosodic tiers (*c.f.* next section) will only provide accurate information if the video features the same speaker. The reasoning behind step 6 in particular will be discussed in Section 4.

2.2. Resulting files

For each video, the procedure generates a TextGrid file which contains eleven tiers. The first tier is the transcription tier, with intervals corresponding to the time stamps in the original subtitle file. The next two tiers are specific to SPPAS, and deal with prosodic analysis: *(i)* the first prosodic tier is a tier with the target points calculated by the MOMEL algorithm (Modeling Melody, *c.f.* Hirst [18]); *(ii)* the second one is a tier with codes from the International Transcription System for Intonation (INTSINT). The remaining tiers include the phonemic and lexical alignments performed by each aligner, and four syllabic tiers (two for each aligner). The syllabic tiers were generated from the Longman Pronunciation Dictionary (*LPD*, [19]). The reason why there are four syllabic tiers is that although both aligners use the Carnegie Mellon University dictionary [20], they transcribe phonemes differently: SPPAS uses SAMPA, whereas P2FA uses ARPABet. Because the *LPD* offers transcriptions in the International Phonetic Alphabet (IPA), the final TextGrid contains one syllabic tier using the *LPD*'s IPA transcription for each aligner, and one syllabic tier for each aligner's prefered transcription system.

Another set of optional files is generated by PEASYV:

- a textual transcript of the video;
- two spreadsheets, one for each aligner.

The first file is a simple word list created to reference lexical usage and frequency. The two spreadsheets are structured in the same way and provide data for phonetic analysis. Each line corresponds to a vocalic nucleus. The following information is collected (the list is non-exhaustive): the video's metadata, the word and syllable in which the vowel appears, the preceding and succeeding phonemes, whether or not they belong to the same word as the vowel, the vowel's and syllable's durations, and formant readings for the first four formants at each centile of the vowel's duration.

This article contends that the relevance and accuracy of PEASYV can be verified by conducting phonetic analyses of vowels: the assumption is that if the vowels are correctly located on the F_1 /F_2 vocalic trapezoid, then the automatic alignment is also correct. Whether that is the case is explored in the next section.

3. A CASE STUDY

The purpose of this section is to assess the quality of the alignment performed by PEASYV. A cursory phonetic study of vowels is provided using raw, unnormalized F_1/F_2 values in Hertz: if these values are reasonable, then the data may be considered sound for future, more refined, research. The recordings come from a YouTube channel, EnglishLikeANative, which purports to offer lessons on British pronunciation: the channel was created by a female teacher offering advice on how to improve pronunciation in order to sound more British. Apart from a few exceptions, she is the sole speaker in her videos. The vowels she produces are therefore expected to be close to standard Southern British English (SBE) values, such as those presented in Deterding [21].

	SPPAS	P2FA
Vowels	973,780	821,125
Monopththongs	775,042	664,264
Diphthongs	198,738	156,861

Table 1: Per-aligner number of vowels extracted from YouTube channel EnglishLikeANative.

In total, 452 videos were aligned using both SPPAS and P2F, amounting to 172 hours and 39 minutes of recording. The combined storage space of all TextGrids and sound files is 3.5G and 15G respectively. 452 vidÃ©os Table 1 summarizes the number of vowels automatically obtained with the two aligners after running PEASYV on all the videos of the channel. Section 3.1 details the results obtained for SBE monophthongs, and Section 3.2, those for SBE diphthongs.

3.1. Monophthongs

One way to obtain an assessment of the accuracy of the alignment carried out by PEASYV is to plot the

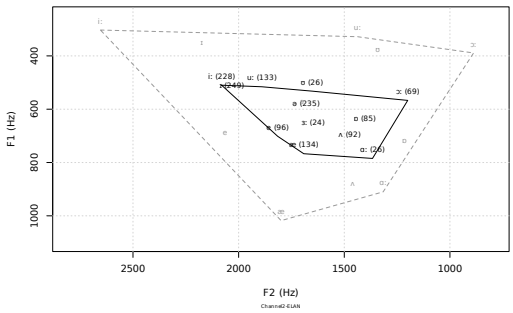


Figure 1: Convex hulls of F_1/F_2 values (in Hz) of all SBE monophthongs. Grey dotted line: Deterding; continuous black line: SPPAS. In brackets: numbers of occurrences (in thousands).

vowels in the F_1/F_2 space for each aligner, and to

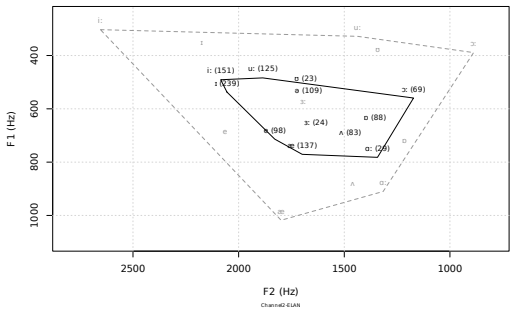


Figure 2: Convex hulls of F_1/F_2 values (in Hz) of all SBE monophthongs. Grey dotted line: Deterding; continuous black line: P2FA. In brackets: numbers of occurrences (in thousands).

compare the resulting values with those from controlled studies. Figure 1 and figure 2 show the mean mid-temporal values in Hertz of the first two formants of each monophthong in colours, along with the values reported in Deterding [21] in grey. The numbers between brackets indicate the number of occurrences of each monophthong in thousands. To facilitate visualization, the convex hulls of the plotted monophthongs have been drawn, in black for SPPAS- or P2FA-aligned values, and in a grey dotted line for Deterding's values. The x-axis (F_2) and

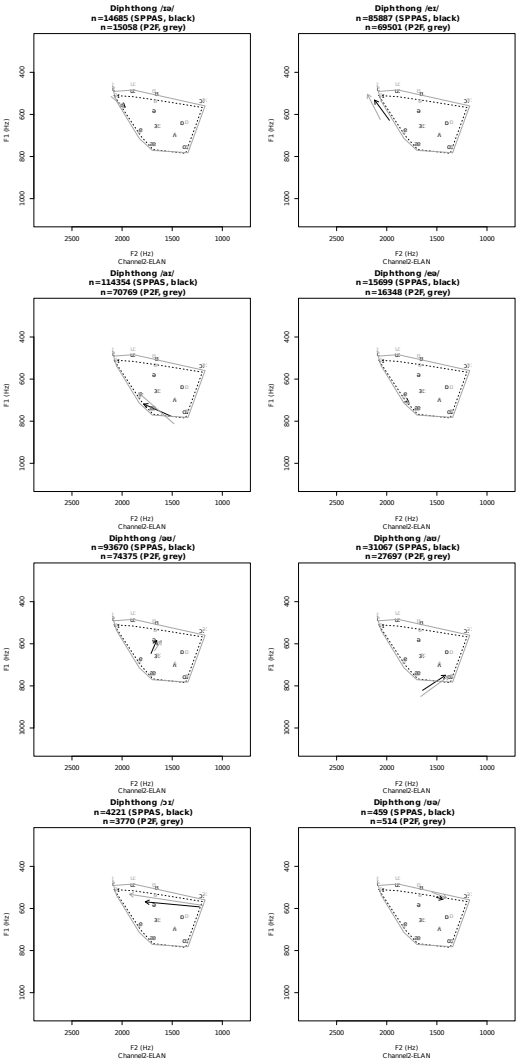


Figure 3: Diphthongs in the F_1/F_2 space compared to the trapezoids of monophthongs (SPPAS: black dotted line; P2FA: continuous grey line) as aligned by SPPAS (in black) and P2FA (in grey).

the y-axis (F_1) have been inverted to emulate the phonological vocalic trapezoids. As can be seen from these two figures, for both aligners the obtained values plot a consistent vocalic space, albeit more reduced than Deterding's (but *c.f.* section 4 for discussion).

3.2. Diphthongs

Figure 3 plots the eight diphthongs of SBE in the F_1/F_2 space. SPPAS-aligned formant values are represented in black, and P2FA-aligned values are shown in grey. The numbers of occurrences of each diphthong for each aligner is indicated on top of each diphthong's pane. The diphthongs are visualized with arrows: the starting point of the arrow is the mean F_1/F_2 values in Hertz at 20% of the diphthong's duration, while the pointed end of the arrow corresponds to the mean formant values at 80% of the diphthong's duration. The convex hulls obtained for monophthongs and presented in Section 3.1 were also plotted in order to assess the consistency of the results. The visual inspection of the glides confirms that the formant values extracted by Praat from the automatically generated TextGrids are in keeping with the female speaker's monophthongs, which themselves are located in a realistic vocalic space.

4. DISCUSSION

This section discusses the specifics and motivations of the procedure (Section 4.1) and then explores the validity of the obtained phonetic data (Section 4.2).

4.1. Procedure

One major obstacle to overcome when automatically aligning recordings of variable duration is containing the scope of misalignments and preventing a potential domino effect, where a misaligned segment might cause all succeeding segments to be misaligned too. The design of PEASYV makes it impossible for such cascading misalignments to occur: step 6 of the workflow presented in Section 2.1 entails that the two aligners are only fed extracts that are a few seconds long, possibly even shorter: in this study, the 404 videos were split into 301,511 short recordings. On average, these subfiles lasted 0.8 second. It is worth noting that such sectioning of the main file into smaller files has no consequences on intonational computations, *i.e.* MOMEL and INTSINT: the main PitchTier is created before, being generated following Hirst [18]. This also means that intonational data is only accurate for videos featuring one speaker, as PEASYV in its current state has no system of speaker detection: the interquartile values of the time-stamped pitch contours are therefore calculated across the entire recording.

4.2. Results

The results provided here are based on raw Hertz values: adding intermediary calculations such as vocalic normalization was deemed potentially counterproductive to assessing the reliability and accuracy of PEASYV. For monophthongs, the vocalic trapezoids feature accurately located vowels. The reduced size of the trapezoids compared to Deterding's most likely reveals hypoarticulation: this is to be expected not only because Deterding's reference values were elicited from potentially hyperarticulated /hVd/ templates, but also because the study's values are extracted from connected speech. However, the mean values of /uː/ admittedly prompt further research: the high F_2 may be indicative of an absence of lip-rounding, which might be the result of an abundance of hypoarticulated high-frequency words (such as "too"). For diphthongs, not only are the trajectories across the vocalic space consistent with phonological representations, the findings are also in keeping with more recent studies describing the progressive disappearance of centralizing diphthongs /eə/, /ɪə/ and /ʊə/ (*c.f.* Hannisdal [22] or Lindsey [23]), or advocating changes in transcriptions, for instance from /eə/ to /ɛː/ (*cf.* Collins [24]).

4.3. Availability and future research

PEASYV for the time being is only accessible on request to adrienmeli@gmail.com, along with the files and scripts used for this study. Deployment on a webpage, and access to the scripts on a github page, are planned for the near future.

One field of research to fine-tune the alignments is that of subtitle accuracy, as the quality of the alignment is highly dependent on that of the subtitles. Research is being carried out to assess the discrepancy between automatic and human-supervised subtitles. Another field is expansion of PEASYV to languages other than English, such as French, using SPPAS.

5. CONCLUSION

This study contends that PEASYV is a robust, low-resource procedure to align subtitled videos and collect phonetic data. The reliability of PEASYV was demonstrated with a study of all the vowels produced by a female speaker on a YouTube channel teaching British pronunciation. The mean formant values extracted from the files generated by PEASYV are consistent with established results and ongoing changes in SBE.

6. REFERENCES

[1] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus." *Journal of the Acoustical Society of America,*, vol. 123(5), pp. 5687-5690, 2008.

[2] B. Bigi, "Sppas: a tool for the phonetic segmentations of speech," in *The Eighth International Conference on Language Resources and Evaluation*, 2012, pp. 1748–1755.

[3] P. Boersma and D. Weenink. (2022) Praat: doing phonetics by computer [computer program]. version 6.3.03, retrieved 17 december 2022 from http://www.praat.org/.

[4] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. of the ICPhS 1999. San Francisco, August 1999*, 1999, pp. 607–610.

[5] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[6] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[7] R. Fromont and J. Hay, "Labb-cat: An annotation store," in *Proceedings of the Australasian Language Technology Association Workshop 2012*, 2012, pp. 113–117.

[8] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[9] S. Gonzalez, J. Grama, and C. E. Travis, "Comparing the performance of forced aligners used in sociophonetic research," *Linguistics Vanguard*, vol. 1, no. open-issue, 2020.

[10] P. Meer, "Automatic alignment for new englishes: Applying state-of-the-art aligners to trinidadian english," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2283–2294, 2020.

[11] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 368–373.

[12] S. Coats, "Dialect corpora from youtube," *Language and Linguistics in a Complex World*, vol. 32, pp. 79–102, 2023.

[13] W. Labov, "Some principles of linguistic methodology," *Language in Society*, pp. 97–120, 1972.

[14] D. yt dlp, "yt-dlp," https://github.com/yt-dlp/yt-dlp, 2022.

[15] F. Developers, "ffmpeg tool (version be1d324)[software]," *URL: http://ffmpeg. org*, 2021.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[17] D. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of Speech Sciences*, vol. 1, pp. 55–83, 01 2011.

[18] ——, "A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences*, vol. 12331236, 2007, pp. 1223–1236.

[19] J. Wells, *Longman pronunciation dictionary*. London: Pearson Longman, 2008.

[20] R. Weide, "CMU Pronouncing Dictionary," 1994. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[21] D. Deterding, "The formants of monophthong vowels in Standard Southern British English pronunciation," *Journal of the International Phonetic Association*, pp. 47–55, 1997.

[22] B. R. Hannisdal, *Variability and change in Received Pronunciation: A study of six phonological variables in the speech of television newsreaders*. [Doctoral dissertation, The University of Bergen], 2006.

[23] G. Lindsey, *English after RP: Standard British pronunciation today*. Palgrave, 2019.

[24] B. Collins and I. Mees, *Practical phonetics and phonology: A resource book for students*, ser. Routledge English language introductions. Routledge, 2013. [Online]. Available: https://books.google.de/books?id=faVJTQIw9eQC

[1] https://youglish.com/

IDEAR: A SPEECH DATABASE OF IDENTITY-MARKED, CLEAR, AND READ SPEECH

Valeriia Perepelytsia[1], Leah Bradshaw[1], and Volker Dellwo[1]

[1] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
valeriia.perepelytsia@uzh.ch, leah.bradshaw@uzh.ch, volker.dellwo@uzh.ch

ABSTRACT

Speakers adapt their speech in response to communicative context and listeners' needs. For example, when talking to hard of hearing or non-native listeners, or in the presence of background noise speakers make their speech more intelligible by speaking clearly. However, it is not clear which adaptations speakers will make when the goal is not intelligibility, but voice recognition. In this paper, we describe a speech database collected using a novel Wizard of Oz technique to investigate speakers' vocal adaptations when they are prompted to sound either more intelligible (clear speech) or more recognizable (identity marked speech). We recorded 39 speakers interacting with a mock speech recognizer, which repeatedly misunderstands speech, and a mock speaker recognizer, which misrecognizes the speakers' voice. We also collected read speech which served as a baseline. All recordings have been orthographically transcribed, forced aligned and manually corrected.

Keywords: speaker recognition, clear speech, speaker characteristics.

1. INTRODUCTION

Speakers dynamically adapt their speech depending on the communicative intent and/or perceptual difficulties on the listeners' side. For example, when talking to hearing-impaired individuals [1], children with and without learning disabilities [2], non-native listeners [3], when communicating under varying levels of background noise [4], or when talking to computers or voice-AI assistants [5] speakers adopt a mode of talk which can be broadly defined as *clear speech*, an intelligibility-enhancing speaking style. Generally, this clear speech is characterized by decreased speaking rate, longer segments, more frequent and longer pauses, increased intensity, and expanded vowel space compared to conversational speech [1], [6].

In addition to varying the degree of speech intelligibility, there is evidence to suggest that speakers can control the degree to which their voice is recognizable, for example, by means of voice disguise [7]–[10]. When disguising their voices, speakers tend to manipulate either the inborn speaker-specific information such as voice fundamental frequency [8] and/or acquired information such as regional accent. This suggests that speakers have some intuition about the speaker-specific information important for their and others' voice identity and are aware of some strategies for changing it.

Although it is evident that speakers can conceal their identity and make themselves less recognizable, it is not clear whether they can move in an opposite direction, namely, enhancing the recognizability of their voices. The ability to manipulate voice individuality and thus enhancing or reducing its recognizability is an important part of vocal communication in humans and many other animals. In human communication, the information about who is speaking is crucial for structuring and understanding speech content [11]. In animals, discriminating among individuals is vital for offspring and mate recognition, cooperative behaviors, and social hierarchies [12].

In this paper, we describe a database collected to investigate acoustic differences in speaking styles when speakers are presented with two types of communicative pressures: (1) being more intelligible (clear speech); and (2) being more recognizable (identity marked speech). Because clear speech is aiming at intelligibility, speakers tend to enhance the canonical properties of segments and suprasegmental features. Thus, speaker-specific properties should be mostly suppressed (see discussion in [11]). Revealing voice identity, on the other hand, implies that speaker-specific anatomical and acquired features should be enhanced. With our database, we want to investigate if speakers reveal more speaker-specific properties in identity-marked speech and whether this leads to their voices being better recognizable under such circumstances. Below we describe the speaker population, the experimental paradigm, the speech tasks, as well as the post-processing of audio files. The motivation for collecting this database is twofold: first, exploring to what extent speakers can enhance the speaker-specific information in their speech, and thus become more recognizable. Also, previous studies considering human-computer interactions (HCI) focus on speech recognition systems [13], [14], while voice-AI assistants are increasingly being used for alternative tasks, such as voice identification (for example, in banking). To our knowledge, this

experiment is a first attempt to explore voice-AI-directed speech on a wider scale.

2. DESIGN AND METHOD

2.1. General setup

To collect speech data in this experiment, we used the Wizard of Oz (WOz) paradigm, which has been widely used in HCI research [15]–[19]. In WOz experiments, participants interact with a computer system that they believe to be autonomous, but which in fact is being operated by an unseen experimenter (i.e., the wizard). WOz paradigm means participants receive seemingly relevant feedback from the system being tested but full experimental control can be maintained. It also allows eliciting different speech modes without explicit instructions for speakers to modify their speech in a certain way.

In our experiment, participants were asked to interact with two mock automatic systems, which we were in the process of developing and testing. The speakers completed three tasks. The first task was reading aloud 34 sentences. Participants were informed that the purpose of this task was to obtain a sample of their speech on which our systems would train. Speakers would then interact with our *speaker recognition* system, Verifico, modelled after the kinds of voice identification technology used in banking. Verifico was aiming to recognize/identify them from their voice sample. After this, speakers would interact with our *speech recognition* system, Vicky, which was similar to Amazon's Alexa or Apple's Siri. Vicky was attempting to correctly understand the speech of the participant.

To the participants knowledge, both Verifico and Vicky were genuine pieces of software which were developed due to a lack of existing systems for Swiss German speakers. However, for the purpose of greater experimental control, the feedback from both systems was fully designed and regulated by the experimenters.

2.2. Speakers

We recorded 39 speakers (mean age = 25.9 years, range = 19–34, SD = 3.36; 20 female) from the University of Zurich student population. All were native speakers of Swiss German. 67 % of speakers spoke Zurich dialect of Swiss German as their native dialect, while the rest of the speakers spoke other dialects of Swiss German. None of the speakers reported any history of speech, language, or hearing disorders. All gave their written consent and received a monetary compensation.

2.3. Recording conditions and procedure

Recordings were made in a sound-attenuated booth at the *Linguistic Research Infrastructure* (LiRI) laboratory at the University of Zurich using a Røde NT1 microphone. Recordings were made directly to disk in WAV format at 44.1 kHz sampling rate, 16 kbit/sec bitrate using Pro Tools software [20].

Each speaker participated in one recording session which lasted approximately 60 minutes. Prior to the recording, speakers were briefed about the procedure and tasks of the experiment. The experimenters informed the speakers that they will be communicating with two automatic systems: a speech recognizer and a speaker recognizer. The experimenters explained the basic principles of how the two systems work, as well as the key differences between the systems. Namely, that the speech recognizer is aiming for *speech recognition*, while the speaker recognizer is targeting *voice identification*.

Speakers were told that both systems are still in the development phase and that they have been already pretrained on a large number of other speakers, but that they still require calibration to improve their performance. They were also told that due to large variability between speakers the systems might make errors: the speaker recognizer might confuse the current speaker with the other speakers from the database on which it was pretrained, and the speech recognizer might misrecognize different words in a produced utterance.

After the briefing, the recording began. During the recording, the speaker sat in front of a computer screen on which prompts were presented via the Gorilla experiment builder [21]. During the speech tasks, the speakers received feedback from the mock automatic systems both visually on the screen and aurally via the loudspeakers inside the booth.

2.4. Speech tasks

The recording session consisted of three speech tasks, which the speakers were instructed to do in Swiss Standard German [22]: the read sentences task, followed by the speaker recognizer task for eliciting identity marked speech, and lastly the speech recognizer task for eliciting clear speech. We opted for Swiss Standard German since common voice-AI assistants, such as Siri or Alexa, do not yet exist for Swiss German, and therefore this was a plausible study motivation for our participants.

The order of tasks was fixed: all speakers started with the read sentences task, followed by the speaker recognizer task, and, finally, the speech recognizer task. We chose this fixed task order as speakers were far more familiar with clear speech strategy compared to identity-marked speech strategy, because of more

experience with speech recognition tools such as Apple's Siri or Amazon's Alexa. Therefore, we decided to present speakers with the speaker recognizer task before the speech recognizer task to avoid transfer effects from one task to another.

2.3.1. Read sentences task

In the first task, the speakers read 34 phonetically rich sentences presented in a random order. All sentences had SVO structure and were 5 words in length. The speakers were informed that both systems were already pretrained on the existing databases and their read speech would be used to enrol them into the datasets. This task was included as a baseline for future acoustic analyses.

2.3.2. Speaker recognizer task

In the second task, participants were interacting with the mock *speaker* recognizer. The task started with the brief interaction between the participant and the system. The system introduced itself as Verifico and asked the participant to select their gender and indicate their name. It then assigned an avatar to the speaker. This avatar was shown whenever the system "correctly recognized" the participant during the task. This introduction was included to give the participant the impression that the system was attempting to learn something about them and their identity.

We used text-to-speech (TTS) software to generate Verifico's utterances (available at https://ttsmp3.com). This software is part of Amazon Polly, a TTS cloud service supporting multiple languages and a variety of lifelike voices. We chose a male voice for our speaker recognizer system.

After the introduction, speakers did several practice trials to make sure they understood the goal of the task, which was to make the system recognize their *voice* correctly. The practice trials were excluded from subsequent analyses and are not available in the database. After the practice rounds, the real task began. Stimuli for this task consisted of the same 34 sentences that the participants had previously produced in the read sentences task.

In each trial (n=34), speakers would say the sentence prompted on the screen and wait for the system's feedback. Each trial had three feedback options. The first option was a correct recognition, whereby the speaker's avatar and their name appeared on the screen. The second option was a single misrecognition, whereby the avatar of another speaker and a different person's name appeared on the screen; in this case, the speaker would have to read the sentence again and would receive a correct recognition after that. The third option was a double misrecognition, whereby the speaker would read a sentence and receive two misrecognitions before being correctly recognized. All three feedback options were equally probable. In each trial, one feedback option was randomly selected by the Gorilla randomizing algorithm. The schematic overview of each trial is given in Figure 1.

During this task, speakers produced on average 67 utterances for all 34 trials in total (SD = 4.07 utterances, range = 60–75). On average, in 34.7% of the trials, speakers were correctly recognized after the first sentence production (SD = 7.5%, range = 17.6–50%), in 32.4% of the trials they received a single misrecognition (SD = 8.8%, range = 17.6–52.9%), and in the remaining number of trials – a double misrecognition (SD = 7.5%, range = 17.6–47%).

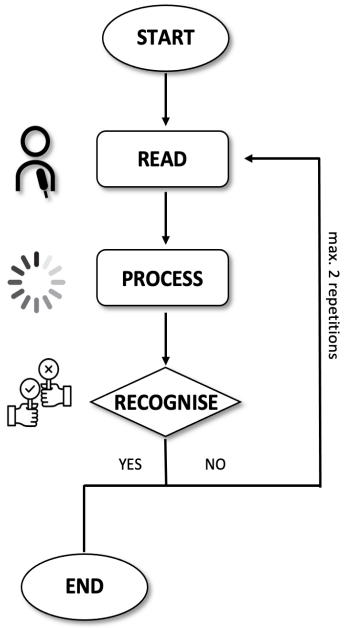


Figure 1: Flow chart showing trial structure in speaker and speech recognizer tasks. The speakers would read the prompted sentence and receive the system's feedback, which consisted of three equally probable options: (1) correct recognition; (2) a single misrecognition; and (3) a double misrecognition.

2.3.3. Speech recognizer task

In the third task, speakers were interacting with the mock *speech* recognizer named Vicky. We chose female voice for the speech recognizer to highlight the difference between the two systems in this experiment. We generated Vicky's utterances using the same TTS software as in the Verifico task described above.

This task also started with Vicky introducing herself to the participants. The introduction was followed by a similar round of practice trials as in the previous task. The practice trials were included to ensure the speakers understood the difference between the two systems and how to complete this new task, which now consisted of making the system

recognize their utterances correctly. In addition, the introduction to the new system and the practice trials served as a buffer to clearly separate the two systems and two tasks for the participants. The practice trials were again excluded from subsequent analyses. After this, the real task began.

As with the speaker recognizer task, the same 34 sentences were used as prompts for what the speakers should say to the system. Each trial (n=34) was structured similarly to the trials in the speaker recognizer task (see Figure 1), however, the feedback given to the speakers by the system differed. Speakers would read a sentence prompted on the screen and wait for Vicky's feedback. The feedback consisted of three equally probable options, which were randomly selected in each trial. The first option was a correct speech recognition, whereby the whole utterance was recognized correctly. The second option was a single misrecognition, whereby two words out of five in a produced utterance were "misheard" by the system. In this instance, speakers would be requested to produce the same sentence again and would be correctly understood by Vicky following this repetition. The final option was a double misrecognition, whereby the system would offer two misrecognitions before a correct recognition. In this instance, each misrecognition would contain two of five words in an utterance "misheard" by the system, and these "misheard" words would be different in each of the two misrecognitions. In the latter case, the speakers would repeat the sentence three times in total before a correct recognition. For instance, a possible double misrecognition scenario may be:

Original sentence:
German: *Die Richterin verliest das Urteil.*
English: 'The judge reads out the verdict.'
Misrecognition 1:
German: *Die Mieterin verliest den Hauptteil.*
English: 'The tenant reads out the main part.'
Misrecognition 2:
German: *Die Freundin vergisst das Urteil.*
English: 'The girlfriend forgets the verdict.'

The misrecognized words were selected based on the acoustic similarity with the target words to ensure that Vicky's misrecognitions were plausible and acceptable to the speakers. Misrecognition of single words might have led to local intelligibility adjustments (as shown by [13]), while we wanted speakers to produce intelligibility adjustments on the utterance level, and not on the word level. Therefore, we opted for misrecognizing two words out of five in each misrecognized utterance. This technique was also adopted to persuade speakers that word-level intelligibility adjustments would not assist the system with recognizing their speech.

During this task, speakers produced on average 67.5 utterances for all 34 trials in total (SD = 5.08 utterances, range = 51–80). In 33.3% of the trials, speakers were correctly recognized after the first sentence production (SD = 7.5%, range = 20.6–52.9%), in 33% of the trials they received a single misrecognition (SD = 7.4%, range = 17.6–52.9%), and in the remaining number of trials – a double misrecognition (SD = 7.5%, range = 17.6–55.9%).

For the preliminary results of acoustic analyses of the collected speech data, see [23].

2.5. Transcription and alignment

2.5.1. Orthographic transcription

All recordings were manually segmented and labelled, before individual sentence productions were extracted using a custom Praat [24] script. Sentence-level orthographic transcriptions were created for all utterances in the form of Praat TextGrids, whereby the first TextGrid tier contained the corresponding sentence, e.g., "Die Pflanze verliert ihre Blätter." (English: "The plant loses its leaves.")

2.5.2. Automatic forced alignment

All utterances were subjected to automatic forced alignment with the help of Montreal Forced Aligner (MFA) [25]. We chose MFA because it has previously shown better performance compared to other widely used forced aligners such as FAVE and MAUS [26]. For each utterance, MFA uses the wav file and the corresponding orthographic transcription in TextGrid format to generate word and phone level alignments. Although the current database contains speech in Swiss Standard German, the alignments were created using MFA Standard German dictionary and acoustic model, which were slightly adapted for the content of the dataset. MFA has been shown to perform well even when the dialect in the dataset to be aligned differs from the one on which the acoustic model was created [26].

All forced alignments were manually corrected by two trained phoneticians (the first and second authors): word boundaries were corrected, and vowel onsets and offsets were defined based in the presence of periodicity and higher formant structure.

6. ACKNOWLEDGEMENTS

This work has been supported by the Swiss National Science Foundation grant #185399. The recordings were carried out at the Linguistic Research Infrastructure laboratory of the University of Zurich.

8. REFERENCES

[1] M. A. Picheny, N. I. Durlach, and L. D. Braida, 'Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech', *J. Speech Lang. Hear. Res.*, vol. 29, no. 4, pp. 434–446, Dec. 1986, doi: 10.1044/jshr.2904.434.

[2] A. R. Bradlow, N. Kraus, and E. Hayes, 'Speaking clearly for children with learning disabilities: Sentence perception in noise', *J. Speech Lang. Hear. Res.*, vol. 46, no. 1, pp. 80–97, Feb. 2003, doi: 10.1044/1092-4388(2003/007).

[3] M. Uther, M. A. Knoll, and D. Burnham, 'Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech', *Speech Commun.*, vol. 49, no. 1, pp. 2–7, Jan. 2007, doi: 10.1016/j.specom.2006.10.003.

[4] K. L. Payton, R. M. Uchanski, and L. D. Braida, 'Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing', *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994, doi: 10.1121/1.408545.

[5] D. Burnham, S. Joeffry, and L. Rice, 'Computer- and human-directed speech before and after correction', presented at the Proceedings Of The 13Th Australasian International Conference On Speech Science And Technology, Melbourne, Australia, 2010, pp. 13–17.

[6] R. Smiljanić and A. R. Bradlow, 'Production and perception of clear speech in Croatian and English', *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1677–1688, Sep. 2005, doi: 10.1121/1.2000788.

[7] T. Kitamura, 'Acoustic analysis of imitated voice produced by a professional impersonator', presented at the INTERSPEECH 2008, 2008.

[8] I. Hove and V. Dellwo, 'The effect of voice disguise on f0 and on the formants', in *Proceedings of IAFPA*, 2014.

[9] A. Eriksson, 'The disguised voice: Imitating accents or speech styles and impersonating individuals', in *Language and Identities*, C. Llamas and D. Watt, Eds., Edinburgh University Press, 2010, pp. 86–96.

[10] A. Eriksson and P. Wretling, 'How flexible is the human voice? A case study of mimicry', presented at the Fifth European Conference on Speech Communication and Technology, 1997.

[11] V. Dellwo, E. Pellegrino, L. He, and T. Kathiresan, 'The dynamics of indexical information in speech: Can recognizability be controlled by the speaker?', *AUC Philol.*, vol. 2019, no. 2, pp. 57–75, Oct. 2019, doi: 10.14712/24646830.2019.18.

[12] E. A. Tibbetts and J. Dale, 'Individual recognition: It is good to be different', *Trends Ecol. Evol.*, vol. 22, no. 10, pp. 529–537, Oct. 2007, doi: 10.1016/j.tree.2007.09.001.

[13] M. Cohn and G. Zellou, 'Prosodic differences in human- and Alexa-directed speech, but similar local intelligibility adjustments', *Front. Commun.*, vol. 6, p. 675704, Jul. 2021, doi: 10.3389/fcomm.2021.675704.

[14] K. Maniwa, A. Jongman, and T. Wade, 'Acoustic characteristics of clearly spoken English fricatives', *J. Acoust. Soc. Am.*, vol. 125, no. 6, pp. 3962–3973, Jun. 2009, doi: 10.1121/1.2990715.

[15] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, 'Wizard of Oz studies — why and how', *Knowl.-Based Syst.*, vol. 6, no. 4, pp. 258–266, Dec. 1993, doi: 10.1016/0950-7051(93)90017-N.

[16] D. DeVault, J. Mell, and J. Gratch, 'Toward natural turn-taking in a virtual human negotiation agent', presented at the AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction., 2015.

[17] L. Gauder, M. Reartes, R. H. Gálvez, Š. Beňuš, and A. Gravano, 'Testing the effects of acoustic/prosodic entrainment on user behavior at the dialog-act level', in *Speech Prosody 2018*, ISCA, Jun. 2018, pp. 374–378. doi: 10.21437/SpeechProsody.2018-76.

[18] S. Oviatt, C. Darves, and R. Coulston, 'Toward adaptive conversational interfaces: modeling speech convergence with animated personas', *ACM Trans. Comput.-Hum. Interact.*, vol. 11, no. 3, pp. 300–328, Sep. 2004, doi: 10.1145/1017494.1017498.

[19] I. Gessinger, B. Möbius, N. Fakhar, E. Raveh, and I. Steiner, 'A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction', presented at the International Congress of Phonetic Sciences (ICPhS), Melbourne, 2019, pp. 1475–1479. [Online]. Available: https://www.assta.org/proceedings/ICPhS2019/papers/ICPhS_1524.pdf

[20] E. Brooks and P. Gotcher, 'Pro Tools'. Avid Audio. Accessed: Nov. 10, 2022. [Online]. Available: https://www.avid.com/pro-tools

[21] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder', *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[22] I. Hove, *Die Aussprache der Standardsprache in der deutschen Schweiz*. DE GRUYTER, 2002. doi: 10.1515/9783110919936.

[23] L. Bradshaw, V. Perepelytsia, and V. Dellwo, 'Vocal effort in human interactions with voice-AI', in *Proceedings of ICPhS*, 2023 [accepted].

[24] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2022. [Online]. Available: http://www.praat.org/

[25] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, 'Montreal Forced Aligner: Trainable text-speech alignment using Kaldi', in *Interspeech 2017*, ISCA, Aug. 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

[26] S. Gonzalez, J. Grama, and C. E. Travis, 'Comparing the performance of forced aligners used in sociophonetic research', *Linguist. Vanguard*, vol. 6, no. 1, p. 20190058, Jan. 2020, doi: 10.1515/lingvan-2019-0058.

PHONLAB PROJECT: COIMBRA – HARVARD. RETHINKING 20TH-CENTURY SCIENTIFIC CENTRES AND PERIPHERIES

Quintino Lopes – Institute of Contemporary History, University of Évora / IN2PAST – Associate Laboratory for Research and Innovation in Heritage, Arts, Sustainability and Territory
Gilberto Pereira – Science Museum of the University of Coimbra
Carla Ferreira – University of Minho, Documentation and Libraries Services
Helena Quaresma – Faculty of Letters of the University of Coimbra, Library and Documentation Services
Carolina Ferreira – Faculty of Letters of the University of Coimbra
Diogo Silva – Faculty of Letters of the University of Coimbra

ABSTRACT

The aim of PHONLAB is to conduct a case study of the Coimbra Experimental Phonetics Laboratory (1936-1979). Created by Portuguese phonetician Armando de Lacerda (1902-1984), the laboratory attracted scientists from leading universities such as Harvard, Cambridge and Bonn, who sought specialist training in the use of innovative laboratory research techniques. How is it that in a country regarded as being scientifically backward there was a laboratory with a reputation as the most advanced in Europe in the mid-20th century? We seek to produce a biography of the laboratory, documenting methods, instruments and providing recognition for its actors. This project will also contribute to the enhancement of a range of scientific sources that will later be made available at museums, and through publications and a digital experimental phonetics archive.

Keywords: Armando de Lacerda, Coimbra Phonetics Laboratory, Chromograph, PHONLAB.

1. INTRODUCTION

In this paper we present the PHONLAB project (2022.06811.PTDC). Funded by the [Portuguese] Foundation for Science and Technology, this research project takes place from January 2023 to December 2025. Its main objective is to reconstruct the historical and scientific importance of the Coimbra Phonetics Laboratory and its founder and head, Armando de Lacerda. This Portuguese phonetician coined the term 'coarticulation', along with Paul Menzerath. He was also an important member of the International Congresses of Phonetic Sciences. He actively participated at the 1st ICPhS (Amsterdam, 1932), where he presented the polychromograph, an innovative instrument that enabled the limitations of the kymograph, then predominantly used in the laboratories of experimental phonetics, overcome, and also the 3rd ICPhS (Ghent, 1938). At the 4th ICPhS (Helsinki, 1961) Lacerda became an honorary member of the Permanent Council for the organization of International Congresses of Phonetic Sciences. After the closure of the Coimbra Phonetics Laboratory in 1979, the memory of the laboratory and its founder and head was consigned to oblivion, and we now aim to recuperate it.

2. WORK PLAN

The importance of experimental phonetics stands in stark contrast with the ignorance of its history. Unlike linguistics, which has a long historiographical tradition, the evolution of the understanding of the sounds of human speech has been assigned little importance by phoneticians and science historians, the evolution of its theories, techniques, instruments and institutions being virtually ignored, and many actors in the field forgotten. Focusing on the Coimbra Phonetics Laboratory, the aim of PHONLAB is to promote understanding of the development of experimental phonetics, and help connect it with the mainstream of history of science research. How is it that the laboratory attracted researchers from the world's scientific centres? Who were the scientists who underwent specialist training in experimental phonetics at Coimbra? What research did they carry out there and how did it affect their careers? What impact did the work of the laboratory have on society? What influence did the innovative techniques and instruments used there have in the field? The aim of PHONLAB is to provide an understanding of the scientific and social impact of the laboratory at the global level.

The relevance of this project also derives from the fact that numerous unique historical sources that document the work of this laboratory and portray the evolution of experimental phonetics throughout the 20th century are in a poor condition and may be lost. A range of documentary and material sources have been found at the University of Coimbra and a number of archives in Europe, the USA, South America, and Australia. Phoneticians who trained at the laboratory have been identified, and it is vital that their testimony be preserved. The analysis of these historical sources will enable us to understand the

nature and impact of scientific practices at the laboratory, and could serve as the basis for a comparison with studies of other laboratories in this field of knowledge.

By opening up new perspectives in the history of experimental phonetics in the 20th century, PHONLAB research will enrich historical narratives and contribute to the reassessment of traditional dichotomies such as that of the 'creative centre' versus the 'passive periphery'. Rediscovering early developments in the field of experimental phonetics may have a positive impact on our current understanding of phonetics and potentially lead to new progress in the field.

2.1. Review of the literature

The few historical studies carried out in the field of experimental phonetics in recent years mainly consist of biographies of scientists and the description of historical scientific instruments [19], [13], [30], [2]. Also noteworthy are the papers published within the scope of the International Workshops on the History of Speech Communication Research [12], [22], [31], [32], [16]. Starting in 2015, these important meetings show that interest in this area of knowledge is recent. As a consequence, the important scientific standing of the Coimbra Phonetics Laboratory had been virtually forgotten until recently. The laboratory was important for the world community of phoneticians during the first half of the 20th century, which was discovered in exploratory studies conducted on the institution [15], [17]. Documenting the path followed by the laboratory enables the deepening of knowledge about the evolution of experimental phonetics in the 20th century and engagement with key current issues in the field of the history of science, such as the 'global history of knowledge', the circulation of knowledge, and scientific centres and peripheries [4], [23], [5], [26].

Spaces which have traditionally been understood as being peripheral have recently been afforded increased importance in the history of science – above all Latin America – while Portugal has remained almost completely absent from the focus of international historiographers who argue in favour of the circulation of knowledge [9], [29], [1], [10], [24], [3], [6]. Recent historiographical works on the Coimbra Phonetics Laboratory [17], [15], [20] provide an innovative contribution not only in the field of Portuguese historiography but also at the international level. The laboratory attracted scientists from major universities abroad to study Portuguese and undergo specialist training in the use of innovative research techniques used at the laboratory. Evidence of such an arrangement, for the 19th century and the first half of the 20th century, has hitherto only been found at traditional scientific centres such as the USA, Germany, France and the UK.

Therefore, the value of this project is twofold: first, the Coimbra Phonetics Laboratory, although located on the European periphery, attracted researchers from key institutions, such as the universities of Harvard, Cambridge, Bonn, Uppsala and the Sorbonne; and secondly, the aim is to gauge the impact of this training on their careers and its impact in terms of improvements made to research and teaching practices.

Following the discovery of a collection of some of the scientific instruments that belonged to the Coimbra Phonetics Laboratory in storage at the Coimbra Science Museum, it is also important that work be carried out to reconstruct their life-paths, which will provide an understanding if instruments such as the polychromograph and the chromographs developed by Lacerda from 1932 contributed to or anticipated current speech recognition and speech synthesis technology [11], [28], [21], [25], [18]. Taking into account that these instruments were used in linguistics studies, namely Portuguese and African dialectological studies, works associated with "linguistic variation" should also be considered besides mainstream literature on linguistics, enabling research on the Phonetics Laboratory to be contextualised within the framework of the scientific paradigms of linguistics [14], [7], [8].

2.2. Research plan and methods

This research project is subdivided into three smaller, interlinked projects:
1. History of experimental phonetics;
2. History of scientific institutions and the material culture of science;
3. Circulation of knowledge, scientific centres and peripheries.

These topics raise the following research questions:

- What is the role of the Coimbra Phonetics Laboratory's work in the development of experimental phonetics worldwide?
- What was the impact of the laboratory and its scientific production on society?
- What instruments were used at the laboratory and how were innovative techniques developed?
- How is it that in a country regarded as being scientifically backward there was a laboratory regarded as being the most advanced one of its kind in Europe in the mid-20th century?
- How did the laboratory manage to attract researchers from the world's scientific centres?

- Who were the scientists who sought specialist training in Experimental Phonetics at Coimbra?
- How was the internship experience of scientists from abroad at the Coimbra Phonetics Laboratory reflected in their academic careers and in subsequent improvements in teaching and research?
- What is the contribution of the laboratory to historiographical knowledge about scientific centres and peripheries?
- Will this case study enable traditional dichotomies such as that of creative centres versus passive peripheries to be reconceptualised and reassessed?
- Why was Lacerda, the laboratory director, a member of the "technical, ancillary and support staff" of the University of Coimbra (he was not a member of the teaching staff) and how this contributed to his historiographical invisibility? [27]

2.2.1. Methodology

This project will be based on an extensive review of the contemporary literature on the history of experimental phonetics, the history of scientific institutions and the material culture of science, and relations between scientific centres and peripheries, as well as the cross-referenced examination and analysis of a number of historical sources: journals, newspapers and periodicals from the era, correspondence, iconography, historical inventories of instruments and collections, reports of activities, trips and attendance of meetings, as well as oral testimony of phoneticians whose careers span the period under study. The use of this methodology will enable a solid scientific basis to be established for investigating the questions raised and delivering the outcomes expected.

2.2.2. Archives, libraries and testimony

The feasibility of PHONLAB is based on the fact that in 2017 the team members began to collect oral testimonies and items from the personal archives of phoneticians from the USA, Germany, Australia and Portugal. These oral testimonies, together with the many we seek to collect during the course of the project, will constitute historical sources providing the groundwork for the research to be conducted. The following archives have already been identified and the scanning of relevant documents is one of the tasks of this project: Coimbra Science Museum; the archives of the following universities: Coimbra, Lisbon, Harvard, Cambridge, Uppsala and São Salvador da Bahia; Bell Labs Archives (New Jersey); the family archives of Paulo de Lacerda (Porto), Sheila Ackerlind (NY), Sue Nogueira (Rio de Janeiro), António Almeida (Cologne), Ana Zamora (Madrid) and Marika Hammarström (Melbourne). We also seek to analyse and make available in the experimental phonetics digital archive monographs and periodicals produced on the basis of experimental work carried out at the Coimbra Phonetics Laboratory (held at the Library of the Faculty of Arts of the University of Coimbra), and document the instruments used at the Coimbra Phonetics Laboratory (some of these have already been transferred to the storerooms of the Coimbra Science Museum and the University of Lisbon Phonetics Laboratory).

2.2.3. Research team

The topics and methodology set out above will provide the basis for research, to be carried out by a transdisciplinary team composed of historians of science, phoneticians, linguists, heritage experts, librarians and audio-visual technicians. We plan to contract a junior doctoral researcher for a period of three years, supported by a research scholarship-holder with a master's degree (funded over 18 months). The scholarship-holder will organise and make available the information collected by the contracted researcher and other team members. Supervised by the PI, the scholarship-holder will be responsible for the producing of a catalogue and digital documents that will be made available in the experimental phonetics digital archive, one of the main outputs of this project. This catalogue includes collections of documents (such as letters, official documents and publications) and scientific instruments belonging to the Coimbra Phonetics Laboratory, as well as documentation from the Paulo de Lacerda family archive. The estate of Armando de Lacerda is held at the Paulo de Lacerda family archive in Porto. This catalogue will also contain documents from other scientific institutions and family archives (the aforementioned institutional and private archives). With the support of audiovisual technicians, an experimental phonetics digital archive will be set up, in which the catalogue, digital documents and interview recordings in connection with a number of figures associated with experimental phonetics during the period under analysis will be held and made available.

2.2.4. Outputs and dissemination

We aim to produce the following outputs:
1) Experimental phonetics digital archive;

2) Peer-reviewed articles published in international journals;
3) The submission of a thematic issue to an international journal in response to an open call;
4) The holding of one workshop and one conference;
5) The publication of an international peer-reviewed monograph as a result of the workshop and conference held.
In order to share the experience of PHONLAB with the scientific community and the general public, we also plan to hold an exhibition at the University of Coimbra. This exhibition will feature a short film about the work of the Coimbra Phonetics Laboratory.

2.3. List of milestones

- 01.07.2023 – PHONLAB website.
  The setting up of the website is one of the key elements of PHONLAB's dissemination strategy. This will be available at the end of the first semester.
- 01.07.2024 – Catalogue of collections: documents, scientific instruments.
  The catalogue is one of PHONLAB's sustainability indicators, which is of decisive importance for the management of collections, the implementation of dissemination actions, and research work. The catalogue will provide the basis for the digital archive.
- 30.12.2024 – Experimental phonetics digital archive.
  The Digital Archive is one of PHONLAB's most important tasks and outputs. It will enable research to be carried out during and after the project by team members and other researchers.
- 10.01.2025 – Workshop.
  The dissemination of PHONLAB research findings. The experimental phonetics digital archive will be presented to the public. The project team and consultants will take part in the workshop, which will be open to the public and other stakeholders.
- 01.12.2025 – International and national outputs.
  Publication of scientific articles, the holding of an exhibition and dissemination targeted at the general public, an internationally-published book, and the holding of the project-end international conference.

3. ACKNOWLEDGEMENTS

The Institute of Contemporary History, University of Évora Unit, receives national funding from the Fundação para a Ciência e a Tecnologia under projects UIDB/04209/2020 and UIDP/04209/2020. Research for this paper was also funded by the FCT under projects 2021.00906.CEECIND and 2022.06811. PTDC (Phonetics Laboratory: Coimbra – Harvard. Rethinking 20th-century scientific centres and peripheries).

4. REFERENCES

[1] Beigel, F., Gallardo, O., Bekerman, F. 2018. Institutional Expansion and Scientific Development in the Periphery: The Structural Heterogeneity of Argentina's Academic Field. *Minerva* 56, 305-331.
[2] Collins, B., Mees, I. 1998. *The real Professor Higgins: the life and career of Daniel Jones*. Mouton de Gruyter.
[3] Daston, L. 2015. At the Center and the Periphery: Joseph Pitton de Tournefort Botanizes in Crete. In: Arabatzis, T., Renn, J., Simões, A. (eds), *Relocating the History of Science: Essays in Honor of Kostas Gavroglu*. Springer, 85-98.
[4] Drayton, R., Motadel, D. 2018. Discussion: the futures of global history. *Journal of Global History* 13(1), 1-21.
[5] Gavroglu, K., Patiniotis, M., Papanelopoulou, F., Simões, A, Carneiro, A., Diogo, M. P., Bertomeu Sánchez, J. R., García Belmar, A., Nieto-Galan, A. 2008. Science and Technology in the European periphery: some historiographical reflections. *History of Science* 46, 153-175.
[6] Gemelli, G. 2002. A Central Periphery: The Naples Stazione Zoologica as an Attractor. In: Schneider, W. (ed), *Rockefeller Philanthropy and Modern Biomedicine. International Initiatives from World War I to the Cold War*. Indiana University Press, 184-207.
[7] Gonçalves, M. F. 2010. A Ortografia Nacional (1904) de Gonçalves Viana e as ideias ortográficas dos reformistas sul-americanos. *Eutomia. Revista de Literatura e Linguística* 3(2).
[8] Gonçalves, M. F. 2020. A «Arte de Enfermeiros» (1741): aspetos do léxico relativo a doenças e remédios no século XVIII. *Revista Panace@* XXI(52), 68-85.
[9] Gootenberg, P. 2007. A Forgotten Case of 'Scientific Excellence on the Periphery': The Nationalist Cocaine Science of Alfredo Bignon, 1884-1887. *Comparative Studies in Society and History* 49(1), 202-232.
[10] Hauswedell, T., Körner, A., Tiedau, U. (eds) 2019. *Re-Mapping Centre and Periphery. Asymmetrical Encounters in European and Global Contexts*. UCL.
[11] Hoffmann, R. 2019. Nothing but a Lung, a Glottis, and a Mouth – The Long Way of Speech Synthesis. In: Pucher, M., Trouvain, J., Lozo, C. (eds), *Proceedings of the Third International Workshop on the HSCR*. TUDpress, 9-28.
[12] Hoffmann, R., Trouvain, J. (eds) 2015. *Proceedings of the First International Workshop on the HSCR*. TUDpress.

[13] Kemp, J. 2001. The development of phonetics from the late 18th to the late 19th century. In: Auroux, S. (ed), *History of the language sciences*, vol. 2. Walter de Gruyter, 1468-1479.
[14] Kohler, K. 2017. *Communicative Functions and Linguistic Forms in Speech Interaction*. Cambridge University Press.
[15] Lopes, Q. 2023. *The global periphery: Armando de Lacerda and the Coimbra Experimental Phonetics Laboratory (1936-1979)*. Center for History of Science of the Royal Swedish Academy of Sciences.
[16] Lopes, Q., Ashby, M., Braun, A. (eds) 2022. *Lacerda 120. Proceedings of the Fifth International Workshop on the HSCR*. TUDpress.
[17] Lopes, Q., Pereira, E. 2019. Armando de Lacerda and Experimental Phonetics in the inter-war period: scientific innovation and circulation between Portugal, Germany and Harvard. In: Pucher, M., Trouvain, J., Lozo, C. (eds), *Proceedings of the Third International Workshop on the HSCR*. TUDpress, 95-104.
[18] Lourenço, M., Gessner, S. 2014. Documenting Collections: Cornerstones for More History of Science in Museums. *Sci. & Educ.* 23, 727-745.
[19] Mehnert, D. 2012. *Historische phonetische Geräte: Katalog der historischen akustisch-phonetischen Sammlung (HAPS) der Technischen Universität Dresden*. TUDpress.
[20] Nunes, M. F. 2020. Duas Culturas?! O caso do Laboratório de Fonética de Armando de Lacerda. In: Gonçalves, M. E., Brandão, T. (coords), *Diálogos Intergeracionais. À conversa sobre Ciência, Cultura, Política científica*. Centro Nacional de Cultura.
[21] Pisoni, D. 1986. A Brief Overview of Speech Synthesis and Recognition Technologies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1326-1330.
[22] Pucher, M., Trouvain, J., Lozo, C. (eds) 2019. *Proceedings of the Third International Workshop on the HSCR*. TUDpress.
[23] Raj, K. 2013. Beyond Postcolonialism … and Postpositivism. Circulation and the Global History of Science. *Isis* 104, 337-347.
[24] Roberts, L. 2009. Situating Science in Global History: Local Exchanges and Networks of Circulation. *Itinerario* 33(1), 9-30.
[25] Rogers, F. M. 1992. *Internationalism and the Three Portugals: the Memoirs of Francis Millet Rogers*. Peter Lang.
[26] Secord, J. 2004. Knowledge in Transit. *Isis* 95(4), 654-672.
[27] Shapin, S. 1989. The Invisible Technician. *American Scientist* 77, 554-563.
[28] Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.
[29] Thurner, M. 2015. Historical Theory Through a Peruvian Looking Glass. *History and Theory* 54(4), 27-45.
[30] Tillmann, H. G. 2006. Experimental and instrumental phonetics: History. In: Brown, K. (ed), *Encyclopedia of Language & Linguistics*. Elsevier, 374-389.
[31] Vainio, M., Šimko, J., Aulanko, R. (eds) 2017. *Proceedings of the Second International Workshop on the HSCR*. TUDpress.
[32] Volín, J., Šturm, P. (eds) 2021. *Proceedings of the Fourth International Workshop on the HSCR*. TUDpress.

ON THE HISTORY OF THE PHONETIC INSTITUTE OF THE UNIVERSITE DE PARIS FROM 1911 TO 1950

Cécile Fougeron, Angélique Amelot, Annie Rialand

Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/Sorbonne Nouvelle (France)
[cecile.fougeron, angelique.amelot, annie.rialand]@ sorbonne-nouvelle.fr

ABSTRACT

After the recent closure of the building hosting the Institut de Phonétique since 1928 at the 19 rue des Bernardins in Paris, we aim in this paper to retrace some of the key aspects of this historical location whose contribution to the development of Phonetic Sciences in France is no longer to be demonstrated. We pursue this endeavour with three main goals: to reconstruct the history of the development of the institute from its establishment in a fruitful context of emergence of experimental phonetics in France up to after WW2; to acknowledge its contribution in the training of a large number of students and future phoneticians; to acknowledge the large diversity of experimental research that was carried out. Finally, we aim to shed light on the contribution of a less renown female phonetician, Marguerite Durand, whose contribution to the history of the institute has been largely overlooked.

Keywords: Institut de Phonétique, Paris, experimental phonetics, history of phonetic sciences, Marguerite Durand.

1. INTRODUCTION

In June 2022, the moving of the current Laboratoire de Phonétique et Phonologie (CNRS/U. Sorbonne Nouvelle) from the 19 rue des Bernardins to its new location has led to the discovery of a large number of documents: registers, library cards, administrative reports, student theses. This article is based on these documents, and aims at retracing some of the key stages in the history of the Institut de Phonétique from its creation in 1911 to 1950.

2. AT THE ORIGIN OF THE INSTITUTE

At the end of the 19th century, research in experimental and instrumental phonetics in the Parisian landscape was vibrant. In 1881, Abbé Rousselot established at the Catholic Institute of Paris the first phonetics laboratory worldwide, obtained a chair in experimental phonetics in 1889, and developed another laboratory at Collège de France from 1897. Work in experimental phonetics in these years was stimulated by numerous contacts with international phoneticians visiting this laboratory, as well as by various collaborations with other institutions in Paris (Alliance Française, National Institute for the Deaf and Dumb, and several clinical institutions).

In 1911, the Institut de Phonétique of the Université de Paris was created at the Sorbonne [1]. Its foundation was made possible by Ferdinand Brunot, who, inspired by Abbé Rousselot's work and teaching, held progressist views as a fervent defensor of phonetics. Ferdinand Brunot had a strong influence at the academic and political levels: he was mayor of the 14th arrondissement from 1910 to 1919, dean of the Faculté des Lettres of Université de Paris from 1919-1928, and founder of the first school dedicated to the teachers of French as a foreign language in 1920. The Institut de Phonétique was part of a larger project between the Université de Paris, the city of Paris, and the Pathé firm. The intent was to build a Phonetics Institute along with Archives de la Parole, a speech archive, as was being done at the time in several other countries. The Institut de Phonétique and the Archives de la Parole were inaugurated at the Sorbonne on June 3rd 1911, with a speech of its future director - F. Brunot – whose vision for the newly created institution integrated both the preservation of the sound of living speech ('les sons de la parole vivante'), and 'the definitive creation of a new science: phonetics, which is related to acoustics, physiology, linguistics, but which is not to be confused with any of them and which must have its own methods, its own instruments and its own scientists.' (our translation, see [2] for the original text in French and [3] for a phonogram recording of some extracts).

The Archives de la Parole, available in part online [4] include several field trips of F. Brunot equipped by E. Pathé with special double horn phonograph (Ardennes in 1912, Berry and Limousin in 1913), field trips pursued after WW1 by Hubert Pernot (in Roumania in 1928 (Figure 2), in Czekoslovakia in 1929 or Greece in 1930), and by recordings made at the Institut de Phonétique. They included recordings of famous people (such as Guillaume Apollinaire, Emile Durheim, Alfred Dreyfus, Cécile Sorel), academics, speakers of French regional languages/dialects as well as recordings of pathological voices.

The Institut de Phonétique and the Archives de la Parole were directed by F. Brunot until 1920, Jean Poirot until 1924, Hubert Pernot until 1930. Pierre Fouché who took the direction in 1930, split the two institutions in 1932, taking the direction of the Institut till 1962, and leaving the direction of the Archives de la Parole to Roger Dévigne.

In the years 1925-26 the status of the Institute changed [1]. While it was still run by the Faculty of letters (Faculté des Lettres), it was officially attached to the Faculty of Sciences and Faculty of Medecine. In 1925 it incorporated the Laboratoire de la Parole of Hector-Victor Marichelle which originated from the National Institution for the Deaf and Dumb.

In 1926, the Institut de Phonétique and the Archives de la Parole benefited from the project of creating a Museum of Speech and Gesture in Paris on the initiative of André Gayot, and in 1928, and moved into the building located at 19 rue des Bernardins, allocated by the city of Paris. At the same time, the Archives de la Parole became the Musée du Geste et de la Parole, although the 'Gesture' part of the museum left no trace.

It was in 1966 that the Institut de Phonétique and the Institut de Linguistique of the Université de Paris merged to give place to the 'Institut de Linguistique et Phonétique Générales et Appliquées' (ILPGA), which, when the Université de Paris was divided after the 1968 riots, joined the Université Sorbonne-Nouvelle (Paris 3).



Figure 1: Plaque on the front of the building, 19 rue des Bernardins, Paris**.**

2. PHONETIC TRAINING AT THE INSTITUT DE PHONETIQUE DE PARIS

From 1920, the activity of the Institut de Phonétique, which, since its creation, was essentially geared towards the recording of the sound archives, moved to an activity of teaching experimental phonetics. For this purpose, F. Brunot pushed for the creation of an Assistant Professor position in phonetics at the Université de Paris. At that time, a chair of Experimental Phonetics had been created since 1890 at the Catholic Institute of Paris for Abbé Rousselot who, in 1922, two years before his death, obtained a chair in Phonetics at Collège de France. This chair disappeared with him. The Maître de Conférence position in phonetics created in 1920 at the Université de Paris was thus part of the plan envisioned by F. Brunot at the inauguration of the Institut de Phonétique to consolidate the creation of a new discipline. The position was entrusted to J. Poirot (from 1920 to his death in 1924), then to H. Pernot (up to 1930), and to P. Fouché who became Professor of Phonetics in 1937 until his retirement in 1962.

The phonetic courses and laboratory work were first given at the Sorbonne, and later at 19 rue des Bernardins starting from 1928. The classes were open to all the students registered in one of the Faculties of the Université de Paris. After 2 semesters, students could apply for one of two diplomas: the Certificat de Licence de Phonétique (created in 1921), and the Diplôme de Phonétique de l'Université de Paris (from 1923) [5]. The exams were mainly for students, French or not, intending to study linguistics or speech and language remediation. Both diplomas were defended in front of a jury composed of three professors, with a written part which was eliminatory and an oral part. During the oral part, the candidate had to defend an experimental work (for the Certificat) and a thesis of experimental phonetics (for the Diplôme), accompanied by two question-and-answer oral sessions (in general phonetics and experimental phonetics for the Certificat; in general phonetics and descriptive and historical phonetics for the Diplôme).

The curriculum of the training in phonetics consisted of 5 main courses [5]. The course of General Phonetics was very similar to a contemporary curriculum (articulation, phonation, audition. Phonemes, syllables, phonetic word. Combinatory processes and phonetic changes. Suprasegmentals). The course of Historical Phonetics, taught as a pro-seminar each year around a different question, aimed at applying the basic phonetic concepts to Romance languages evolution, with a focus on the detailed phonetic mechanisms and rigorous methods of analyses. The course of Experimental phonetics aimed at introducing the students to the use of instruments (see section 3). Practical training sessions took place every day of the week in the laboratory with the help of an assistant. Experimental works had to be presented every week during the first semester and during the second semester, the candidates worked on their own research topics for a thesis to be presented for the diploma (see below).

Starting from 1933, a course in Speech and Language Pathology was taught at the Institut by Dr André Ombredane. In parallel, was established an experimental section dedicated to remediation by

Mme Georges Lamarque, professor at the Institute of Deaf and Dumb.

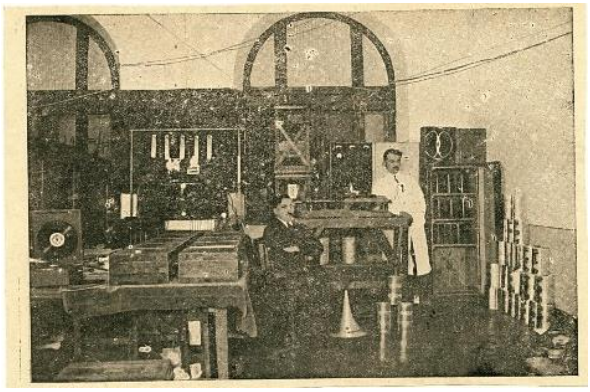


Figure 2. Laboratory set-up in Bucarest for the recording campaign in Romania by H. Pernot, 1928 (unknown source).

In parallel to the training in phonetic sciences, the Institut de Phonétique also provided a training in French pronunciation with 'special courses' of one semester leading to a "Certificat d'étude pratique de prononciation française" and summer courses during July or August leading to a an 'examen de fin de cours de vacances'. These French pronunciation courses had a great success, with 3609 students from around the world enrolled between 1931-39 and 1946-49 (these courses stopped during WW2).

The students trained in the phonetic curriculum at the Phonetic institute were less numerous, but also originated from a large number of countries. In his report of the year 1926-1927, H. Pernot [1] mentioned 53 students, from 18 nationalities, mostly European. An even larger diversity in the origins of the students was found in the 1931-1950 registers, with students from all around the world. Figure 3 illustrates the distribution of the 416 students listed in these registers.

Quite a few female students were trained and graduated, even though the vast majority of students were male. Among these students were such famous names as André Haudricourt, Georges Straka, Pierre Delattre, John Catford, Bertil Malmberg, Pierre Léon. An expected drop of students occurred during the years of WW2, but the training continued during the occupation of Paris, and we sadly found a few names of students that had been deported.

One point of interest is the large number of students who got training in this curriculum without registering to any diploma. Indeed, as illustrated in Figure 3, only a handful of diplomas (Certificat or Diplôme) were delivered each year. According to P. Fouché [5] this low number of candidates and of graduations was explained by the high standards of these exams.

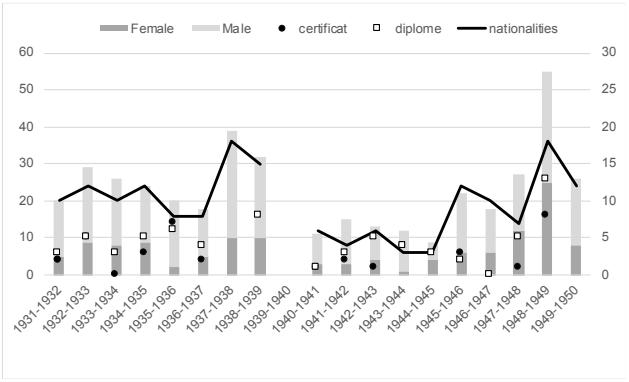


Figure 3: Left axis: Number of students (left axis), split by sex, registered at the Institut de Phonétique between 1931 and 1950. Right axis: Number of Certificats and Diplômes delivered and number of nationalities of the students per year.

3. RESEARCH IN EXPERIMENTAL PHONETICS ILLUSTRATED BY STUDENT WORKS AT THE INSTITUTE

The data presented in this section originates from our analysis of documents found: library cards, listings of the students' research titles in the registers, and original students research manuscripts defended at the Institute between 1923 to 1950. More that 800 theses covering 1923 to about 1990 had to be sorted. We identified more than 140 students' theses between 1923 and 1950. Figure 4 presents the distribution of these theses.

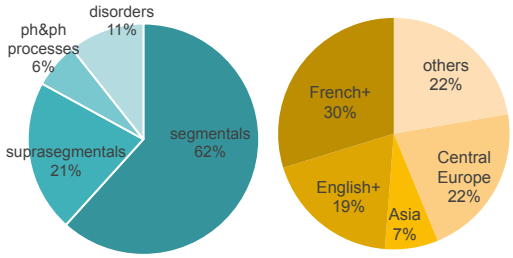


Figure 4: Distribution of the 1923-1950 students' theses in terms of research questions and languages studied.

Most of the theses addressed questions related to sound articulations (segmentals), some reported on tonal, rhythmic or intonative properties (suprasegmentals), a few on phonetic and phonological processes (assimilation, epenthesis, etc.), and still others on speech disorders (deaf speech, developmental disorders, cleft palate, aphasia, etc.). Over 24 languages and language varieties were studied thanks to the various nationalities of the students. Indeed, many of the theses reported on the students' own productions. One example of such theses for the Certificat was a

work on consonants and vowels of New York published in the Revue de Phonétique in 1928 [6].
Research in experimental phonetics was done on several types of instruments, as described by P. Fouché in his inventory of a fully equipped lab in 1935 [5]. Thanks to the funding provided by E. Pathé in 1911 to the Archives de la Parole, the Institute was equipped with several types of phonographs. Some equipments belonged to Rousselot and to Adolphe Zund-Bürget. J. Poirot came in 1920 with his labiograph, and he directed the doctoral research of Liu Fu, using a recording cylinder (built by Boulitte), equipped with an electric tuning fork and a manometric mask. The result was a pioneering thesis on the phonetics of Chinese tones [7]. From 1926, the merging with the speech laboratory of the Deaf and Dumb made accessible the Lioregraph. Palatography was common in students' work documenting their own pronunciation along with photographies. Cinematography was sporadically used from 1933. All the experimentations were done under the supervision of an assistant: M. Schramek till 1929, then Marguerite Durand, whose work is the focus of the next section.

4. SPOT LIGHT ON THE CONTRIBUTION OF MARGUERITE DURAND

Marguerite Durand (1904-1962) obtained her Diplôme in phonetics in 1928 and was hired as an assistant in Experimental Phonetics, and then as an 'Attaché de l'Institut' from 1929 until the end of her career. She never held an academic position despite her doctoral degree (in 1936), her publications, her teaching of the Experimental Phonetics course, and the supervision of many students' theses.
Besides articles ([8], [9], [10], [11], [12], [13], [14], [15]), M. Durand authored or co-authored four books before WW2, each of which was an important contribution to the field.
The first books of M. Durand [16], [17] which were published in 1936, concerned Parisian French viewed in its diversity within and around Paris. The first book included not only a phonetic study with several speakers but also a large-scale dialectological study of masculine and feminine adjectives (and some pronouns) with 400 speakers, showing that old forms, such as [kχў̆] cru (masc.) 'raw'/ [kχў̄ə] or [kχў̄j] crue (femin.) 'raw') were still present even in villages of the Paris region. The second book was a very detailed study of consonant durations, investigating up to 13 contexts (segmental contexts, position in clusters, in words, in sentences, etc.). The findings concerned the temporal organization of clusters and the influence of prosody and constituents on consonant durations.
In 1939, she co-authored a book on phonetic/ phonological errors in various forms of aphasia [18] with Théophile Alajouanine, a Professor of medicine at the Paris faculty and André Ombredane, Assistant director of the 'Laboratoire de psycho-biologie de l'enfant' at the EPHE, Paris. Similarly to R. Jakobson [19] at the same period, the authors viewed the 'phonetic disintegration' in pathologies as a mirror image of the child's acquisition. Phonetic investigations of patients' speech, with a noticeable attention to coarticulation, were performed by M. Durand. This book was pioneering in its integration of medical, psychological and phonetic approaches to improve the characterization and understanding of the disorders.
M. Durand's book on long and short vowels was published in 1946 [20], though it had been finished before the war. The book aimed at investigating the nature of the quantity contrast in a large diversity of languages, such as American English, Breton, Dutch, Swedish, Hungarian, Estonian, Lebanese Arabic, Thai, and several dialects of Chinese. Data were taken from published works, from her own investigations and extensively from those of students that she had supervised (such as Bertil Malmberg or Joe Larochette). The methodology to question the nature of the contrast between short and long vowels using a phonetic approach and the carefully worked out data for individual languages were the two main contributions of this book. Her attempt to capture the cross-linguistic nature of the contrast – which in her mind was not durational in essence – was controversial and is now outdated.

5. CONCLUSION

The analysis and archiving of the Institut de Phonétique documents are far from being complete. Next steps will be to process the content of the theses and relate them to the instruments and methodologies used at the Institut. While most of these instruments have already been donated to the Bibliothèque Nationale de France in 2000 (coll. Charles Cros [21], along with the Archives de la Parole), the opening of the attic has allowed the discovery of other instruments which are currently restored.

6. ACKNOWLEDGEMENTS

This work is dedicated to all past members of the 'Institut des Bernardins', in memory of this inspiring location.

7. REFERENCES

[1] Pernot, H. 1928. “L’institut de phonétique de l’Université de Paris.” *Revue de Phonétique*, 5, 28–46. https://gallica.bnf.fr/ark:/12148/bpt6k92981m
[2] Rousselot, P.-J. 1911. “Chronique.” *Revue de Phonétique,* 1, 197. https://gallica.bnf.fr/ark:/12148/bpt6k92977c
[3] Brunot, F. recording of the “Discours d’inauguration des Archives de la parole, June 1911” https://gallica.bnf.fr/ark:/12148/bpt6k1279113
[4] Cordereix, P. 2005. “Les fonds du département de l’Audiovisuel de la Bibliothèque nationale de France.” *Le temps des médias* 5 (2), 253–264.
[5] Fouché, P. 1935. “L’institut de Phonétique et le musée de la parole de l’Université de Paris.” *The French Review,* 9(1), 44–50.
[6] Strong, L. 1928. “Voyelles et consonnes de New-York.” *Revue de Phonétique,* 5, 70–107. https://gallica.bnf.fr/ark:/12148/bpt6k92981m.
[7] Fu, L. 1925.“Etude Expérimentale sur les tons du chinois,” Paris : Société des Belles Lettres, Pékin : Presses de l’Université Nationale de Pékin. https://gallica.bnf.fr/ark:/12148/bpt6k990887w
[8] Durand, M. 1929. “Etude sur les phonèmes postérieurs dans une articulation parisienne.” *Revue de Phonétique,* 6, 241-256. https://gallica.bnf.fr/ark:/12148/bpt6k92982z.r=revue%20de%20phonétique?rk=107296;4
[9] Durand, M. 1938 “Essai sur les vers de Boileau.” *Le français moderne,* 6, 331-336
[10] Durand, M. 1938. “Durée phonétique et durée phonologique.” *Proceedings of the 3rd International Congress of Phonetic Sciences (Gand),* 262-265
[11]Durand, M. 1939. “Essai sur la nature de la notion de durée vocalique ” *Etudes phonologiques dédiées à la mémoire de N.S. Trubetzkoy. Travaux du Cercle de Linguistique de Prague,* 8, 241-256.
[12] Durand, M. 1945. “Quelques observations sur un exemple de parisien rural” *Le français moderne,* 13, 83-91.
[13] Durand, M. 1946. “Perception de durées dans les phrases rythmées.” *Journal de psychologie normale et pathologique,* 305-321
[14] Durand, M. 1949. “De quelques éliminations d’homonymes chez un enfant.” *Journal de psychologie normale et pathologique,* 53-63
[15] Durand, M. 1950.“Ce que Racine penserait de notre prononciation.” *French Review, 24 (1). 47-52*
[16] Durand, M. 1936. *Le genre grammatical en français parlé à Paris et dans la région parisienne*. Paris: J. D’Artrey
[17] Durand, M. 1936. *Étude expérimentale sur la durée des consonnes parisiennes*. Paris: J. D’Artrey
[18] Alajouanine, T, Ombredane A & Durand, M. 1939. *Le syndrome de désintégration phonétique dans l’aphasie*. Paris : Masson et Cie.
[19] Jakobson, R. 1941.*Kindersprache, Aphasie und Allgemeine Lautgesetze*. Uppsala: Almqvist & Wiksell.
[20] Durand, M. 1946. *Voyelles longues et voyelles brèves. Essai sur la nature de la quantité vocalique*. Paris: C. Klincksieck.
[21] Collection Charles Cros, https://archivesetmanuscrits.bnf.fr/ark:/12148/cc97705f

THE 6th ICPhS IN PRAGUE AND THE ORGANIZED LIFE OF PHONETICS

Pavel Šturm

Institute of Phonetics, Charles University, Prague, Czech Republic
pavel.sturm@ff.cuni.cz

ABSTRACT

This paper examines the post-war efforts of phoneticians in Europe to facilitate the exchange of scientific knowledge in their field. It focuses on the organization of the 6th International Congress of Phonetic Sciences (ICPhS) in Prague in 1967, contextualizing it within the re-establishment of the tradition of phonetic congresses. The paper highlights the role of Bohuslav Hála, a prominent Czech phonetician of the time, in the planning and execution of the Prague congress, based on an analysis of his correspondence and other archival materials. Three areas of collaborative efforts are discussed: the development of the *Manual of Phonetics*, the organization of the phonetic congresses, and the hosting of smaller events (symposia) on specific phonetic topics.

Keywords: phonetic congresses, ICPhS, ISPhS, history of phonetics, Bohuslav Hála.

1. RE-ESTABLISHING THE SUSPENDED TRADITION OF PHONETIC CONGRESSES

The series of triennial meetings of phoneticians and related scientists, which had commenced in 1932, came to an abrupt halt with the outbreak of World War II. The last ICPhS before the war was held in Ghent in 1938. However, despite the challenges posed by the war, there was a persistent interest in continuing the congresses in the following decades. This paper will focus on the Czech efforts that culminated in the organization of the 6th ICPhS in Prague.

1.1. Hála–Kaiser correspondence (1948–1958)

Louise Kaiser is a renowned figure in the field of phonetics congresses. She pursued her studies in medicine at the University of Amsterdam and became the first female lecturer there. She played an instrumental role in establishing experimental phonetics in the Netherlands, particularly at her university [1]. Notably, she was one of the organizers of the inaugural ICPhS in 1932 in Amsterdam, and remained a member of the Permanent Council (PC) both before and after World War II. Additionally, she actively participated in the 1932, 1935, and 1938 congresses. Kaiser maintained extensive correspondence with Bohuslav Hála, a Czech phonetician who later became the director of the Institute of Phonetics in Prague after the passing of Josef Chlumský [2]. Hála also participated in the 1938 congress, but he and Kaiser had already become acquainted in 1923 during Kaiser's training in the Prague laboratory.

On 24/1/48, Hála informed Kaiser of his plans to tour phonetic laboratories in Holland, Belgium, and France, and expressed his intention to visit her in June. Kaiser responded positively, eager to hear Hála's critique of her workplace. She also informed him about an upcoming "small phonetic congress" that was disguised as a section of the Int. Congress of Anthropological and Ethnological Sciences, to be held in Brussels in August.

However, on 18/2/48, during the impending communist coup in Czechoslovakia, the Ministry of Education refused to fund Hála's journey to Western laboratories. On 22/6/48, Hála wrote to Kaiser, informing her that it was possible he would attend the congress in Brussels instead, with the hope of visiting her and checking some of her instruments. Unable to obtain a visa to Holland in time, their interaction was limited to the event in Brussels. Apparently, it was decided there (and was one of the aims of the event) that phoneticians should meet more frequently, continuing the interrupted tradition of phonetic reunions.

On 16/2/53, Hála complains to Kaiser that five years have passed and still nothing. Kaiser's answer is missing. At that time, she was busy preparing the *Manual of Phonetics* [3], a synthesis of various issues in phonetics, intended to be a continuation of the PC's former activities. "The spirit of the book was to be international" [3: ix]. On 5/5/53, Hála received the contents of the book, along with his name indicated under one section, which left him confused, as he had no prior knowledge of it. Kaiser responded with apologies, regretting that the contact between them had been less frequent. Hála then agreed to participate as an author and submitted his contribution on the Slavic languages at the end of the year. Kaiser was exited.

On 30/9/54, Hála wrote to her: "already a long time has passed since the last Phonetics Congress in 1948 (to tell the truth, the last independent and universal Congress was that of 1938). Wouldn't you find it [...] possible to think of organizing a new congress next year?" Kaiser replied on 5/10/54. She agreed the 1948 congress was "insufficient for lack of independence." She explained that she had been trying to organize a congress similar to those held in the 1930s since the

end of the war, but had encountered little response. This was why she had focused her efforts on producing the *Manual*. "Nevertheless, it would be worth to try anew to establish a congress." She promised she would discuss the matter with the PC if Hála came up with a plan of his own.

On 24/10/54, Hála responded with these propositions: determining the worldwide interest in the congress, considering moving the congress to 1956 if 1955 is too soon, and suggesting Milan as the possible location. Subsequent correspondence concerned the *Manual* again. However, a year later (23/9/55), Hála returned to the urgent question of the congress and proposed Milan or Paris as potential locations. Kaiser replied on 17/11/55, excusing the delay by not being able to help with the congress anyway. While she agreed with Hála that a new congress would be useful, she remained sceptical about its possibility based on her bad experience in Brussels. "Nevertheless, if someone succeeds in organizing a new congress, I would be very happy." Hála added a handwritten note to the letter: "Take care of it instead of Kaiser."

In the letter from 2/5/56, Hála reminds Kaiser of her last words and asks for her opinion on organizing the next congress in Prague. Apparently, there have been other proposals as well, as a circular sent to PC members on 14/5/56 informs them of three possibilities: Helsinki 1958 (Sovijärvi), Coimbra 1957/58 (de Lacerda) and Prague (Hála). The members of the council were asked to vote.

In a reply from 23/5/56, Hála offers to leave the organization to whoever contacted the PC first. He personally advises Coimbra 1957. On 30/7/56, Hála receives an official invitation from Kaiser to become a member of the PC, which also includes information that the site and time have not yet been settled. In October 1956, Hála responds to Kaiser's circular to PC members from 20/9/56, replying as follows: 1) he votes along with the majority for Coimbra in August 1958; 2) in exchange, he insists on excluding linguists from ICPhS, as they have their own congress, unless their contributions are phonetic in nature. On 30/10/56, Kaiser informs him that the decision about the congress has not yet been made. After several months, on 6/3/57, Kaiser informs Hála that she "might go to Coimbra in April this year to arrange the congress with M. de Lacerda." On 6/5/57, Hála reminds her of this and asks if it has been settled. Impatient, he also writes directly to de Lacerda (see Section 1.2). Kaiser's circular to PC members from 20/6/57 informs them that Coimbra is no longer an option because de Lacerda is mostly out of Portugal. The solution was to switch for Helsinki. However, on 29/6/57, Kaiser brings ill news: "difficulties are piling up around the 1958 Congress and I fear that the time left for the preparation will be too short."

On 20/12/57, Hála responds to a PC circular, voting for Milan as the site of the next Congress so that Gemelli's laboratory could be visited. However, as of 16/2/58, Hála has received no response or any news about the Italian congress and requires some information from Kaiser. There is no extant reply.

1.2. Hála–de Lacerda correspondence (1957)

On 26/5/57, Hála wrote to the Portuguese phonetician Armando de Lacerda: "For three years now I have been insisting with Madame Kaiser that a congress of phoneticians should be organized. At the beginning of this year, she informed me that she intended to go to Coimbra to discuss with you the conditions for a congress to be held in Coimbra. As Mrs. Kaiser has not until now answered my questions in relation to the congress and as, for certain reasons, I must know it to ensure my participation, from now on, I take the liberty of addressing you to ask for any details of your meeting with Madame Kaiser about the congress." There is no reply either in Prague or in the Lacerda archive (Quintino Lopes, personal communication).

1.3. Hála–Gemelli correspondence (1958)

In a letter to Agostino Gemelli from 11/4/58, Hála inquires whether the Congress that was to take place in August 1958 in Milan is postponed to the next year. Hála adds that Kaiser informed him in her last letter (which we do not possess) that she was resigning from the PC; as a result, Hála is unsure about the Congress' future. He believes that "a reunion of phoneticians, suspended since 1938, is of a great importance."

Gemelli replied curtly on 16/4/58 that the phonetic congress would not be occurring. Instead, he offered Hála a programme for the 5th congress of the Italian Society of Exper. Phonetics, Biological Phonetics, Phoniatrics and Audiology, scheduled for May 1958 in Milan. There is no evidence Hála attended the event.

1.4. Hála–Husson correspondence (1955–1959)

Another stock of correspondence is only indirectly relevant to our theme. Since at least 1951, Hála corresponded with the French physiologist and phoniatrist Raoul Husson. They admired each other's work highly, leading for instance to Hála being elected "Président général adjoint du Haut-Collège" for his outstanding work on Czech vowels. Hála was invited to Paris several times but due to a variety of reasons he was able to come only as late as 1957 (between April 5 and 20). He delivered a lecture in front of the Société de Linguistique on April 6. In a letter to Husson from 12/5/57, Hála fondly reminisces about the good times they had in Paris and the hours they spent discussing topics of mutual interest.

On 17/8/57, Hála informs Husson that, based on his report from the Parisian trip, the Czech ministry has promised to invite Husson to Prague the following year. On 16/12/57, Husson accepts the invitation and further informs Hála about a congress of the Italian Society of Experimental Phonetics (Gemelli) scheduled for September 1958 in Milan on the acoustic structure of consonants. Husson also mentions that the *Manual of Phonetics* has been published but criticizes it as a collection of "old fantasies," "outdated even before its publication." He wonders when a "serious handbook" will be published.

Hála responded to the letter on 19/1/58, saying nothing of the congress but commenting on the *Manual* and Husson's remark: "We drew up the plan together in Paris. I don't give it up, but for the moment there are so many pressing difficulties and other obstacles that this project must be postponed; we'll talk about it when the opportunity arises."

Eventually, Husson stayed in Prague in June 1959. He wrote to Hála on 11/5/59: "Important! I will certainly talk to you about our project, sketched out two years ago! Make a phonetics manual by our collaboration." Husson alleges that Panconcelli-Calzia has said to him that Kaiser's manual was already "born dead." Husson asks Hála to prepare a tentative list of contents for their new manual to be discussed in detail during his stay in Prague.

The visit was a success. Husson was pleasantly surprised by the progress of the Prague laboratories, writing on 16/6/59: "I remain stupefied and profoundly enthusiastic! [...] It's a model institution, the first of its kind in the world, that you have brought to life in Prague! [...] I have the impression, here, of returning to the past! I feel like in the plain Middle Ages."

Unfortunately, this is the last extant letter between the two from this period. Also, their collaboration on the manual was not continued, and nothing was published (the second edition of the *Manual of Phonetics* from 1968 [4] was an independent activity).

2. POST-WAR CONGRESSES

2.1. First post-war ICPhS: Helsinki, 1961

Almost a quarter of a century (1938–1961) passed between the third congress in Ghent and the fourth in Helsinki. In May 1960, Circular #1 was disseminated, announcing the event and explaining the intermission. "After the impossibility of having a Congress in 1958 had become obvious, the Finnish colleagues offered to stage the Fourth Congress at Helsinki in 1961. On January 30th, 1959, they were informed by Professor Roman Jakobson, the President of the [PC], that the Council had accepted the offer." Both Hála and de Lacerda were members of the PC. As a result, Hála finally participated in his second ICPhS, presenting an article on the phonetic nature of the syllable [5].

The PC underwent significant changes. Jakobson jokingly remarked that given the long gap, the "Council membership threatened to become permanent" as well. Dennis Fry was elected the new President, Hála Vice-President. In the final session, the PC accepted E. Zwirner's offer, and Bochum (later changed to Münster) was selected as the site for the 5th congress. Prague was proposed and tentatively selected for the 6th congress.

Hála also initiated here the restart of the International Society of Phonetic Sciences (ISPhS), with the aim to "constitute a permanent link between all phoneticians of the world" [6: xxviii].

2.2. Münster, 1964

Hála attended the congress, delivering a paper on the historical development of the Czech <ř> sound (/r̝/) [7]. Hála was Vice-President of the PC. The council decided at the congress that the next one would take place in three years in Prague, chaired by Hála. It's worth noting that Hála had initially suggested organizing a congress in Prague to Kaiser in 1956 (and to himself in 1955, if not earlier). More than ten years had passed between conceiving and holding the event. Hála was 73 years old at the time of the 6th ICPhS.

3. THE PRAGUE CONGRESS

3.1. Organizing Committee (OC)

On 24/3/65, the first meeting of the OC convened in Prague. Members included phoneticians but also phonologists, linguists and medical specialists, a total of 18 people. A narrower group of five individuals (four phoneticians and one administrative member), formed the Executive Committee, which met informally on a regular basis and held formal meetings 3-4 times a year. Although the organizer was the Czechoslovak Academy of Sciences, the primary responsibilities were carried out by the phoneticians from Charles University. Hála served as the President of the congress, while Milan Romportl assumed the roles of General Secretary and Chair of the Executive Committee and scientific programme committee. Přemysl Janota chaired the technical committee and Alena Skaličková was in charge of the social committee.

In December 1965, Circular #1 was distributed to all participants from Helsinki and Münster, as well as to selected universities, academies, and phonetic institutions worldwide. Additionally, the text was published in various journals, such as *Phonetica* and *Zeitschrift für Phonetik*. The circular provided information about the theme of the congress ("Acoustic Speech Signal and its Perception"), the date and

location (Sept. 7–13, 1967, Prague) and the official languages (English, German, French, Russian). Application forms were to be returned by March 15, 1966. A total of five circulars were issued.

Importantly, the OC coordinated the phonetic congress with the organizers of the 10th International Congress of Linguists in Bucharest, Romania, which took place a week before. This arrangement allowed overseas participants to attend both events. Furthermore, there was an agreement regarding the content, as phonetics and phonology papers were allocated exclusively to Prague and not to Bucharest.

The budget had to be repeatedly revised due to unforeseen circumstances. The original plans were based on the number of participants from the post-war congresses (cca 250 people). However, due to an unprecedented interest in the Prague event, the number of participants reached 606, necessitating adjustments in the budget. Additionally, constant changes in contracts and prices resulted in the reduction of some services, such as simultaneous interpretation being available only in the main auditorium instead of the parallel sessions as initially planned. The final budget was approved by the Czech Academy on 29/6/67.

The congress correspondence was extensive (1225 pages). The content is mostly organizational, related to travel, accommodation and registration/submission.

3.2. Scientific programme

There were five **plenary lectures**, related to the congress theme. The invited speakers talked for 30-45 min, followed by 60 min for discussion. During the sixth meeting of the OC (23/3/67), there was a discussion about whether a Czech representative should also have a plenary lecture and who it should be. However, this idea was eventually rejected and, instead, Hála was asked to deliver a historical summary of the development of Czechoslovak phonetics in his opening speech.

There were 278 talks in **parallel sessions** (five/six at once), of which 235 are included in the proceedings [8]. The main areas – Physiology and pathology of speech, Acoustics and perception of speech, Phonology and linguistic phonetics, and Biophonetics – were introduced with a longer paper (30 min) followed by a 15-min discussion. However, Husson's introduction was missing as he unexpectedly and angrily withdrew all participation in May 1967 in response to the 4th Circular. To avoid simultaneous occurrence of related topics in different sessions, the remaining contributions were organized into triads of papers. Each author had 10 min for presentation, and there was a joint 15-min discussion of the three contributions. This required precise timing and synchronization across the rooms. To great acclaim, a system of centralized optical signalization was designed and constructed, displaying the current presentation order (1 / 2 / 3 / D) in the rooms and in the foyer.

In addition to paper presentations, the congress also featured **other activities** such as scientific films, demonstrations of phonetic instruments and research equipment, and an exhibition of literature and apparatus. A **round-table discussion** involving 14 participants was also held to discuss the present-day tasks of the phonetic sciences, and a summary was included in the proceedings.

4. ISPhS SYMPOSIA

On 1/2/57, Kaiser wrote to the PC expressing uncertainty about whether ISPhS was still in existence, as it had been inactive until the first post-war congress. However, at the 4th ICPhS, the idea was born to have smaller events – symposia – dedicated to specific fields. Two such symposia were organized in 1970 in Czechoslovakia under the patronage of the ISPhS.

Milan Romportl held a **Symposium on Intonology** in Prague (October 6–8), marking the 50th anniversary of the Institute of Phonetics. The proceedings [9] was dedicated to the memory of Hála, who had passed away a few weeks before the event. The papers focused on various issues in intonation, including methods of pitch registration. In his closing speech, ISPhS Secretary General Martin Kloster-Jensen stressed the "need for the exchange of ideas especially in the form of personal contacts realized through symposiums [*sic*] in the way in which it has been practised here." He further entreated the 109 participants to "try and fill the vacuum between the international congresses [...] with meetings on a smaller scale." [9: 289].

Some of the participants moved on to Brno, where a **Symposium of Paedolinguistics** (child speech) was organized by Karel Ohnesorg, Hála's former student and assistant [2]. It took place on October 14–16 and attracted 56 researchers from 14 countries, who presented a total of 34 contributions [10].

5. CONCLUSION

Although many people were involved in organizing the 1967 ICPhS, it was B. Hála's merit that the congress took place in Prague at all. Hála had long dreamed of a phonetics congress and had been instrumental in pressuring various individuals, particularly L. Kaiser, towards organizing such an event. His acceptance into the Permanent Council in 1956 and subsequent election as its Vice-President in 1961 marked a significant step forward, underscoring the excellent reputation of Prague phonetics in the 1960s. In addition to congresses, the importance of ISPhS or IPA in the organized life of phonetics should also be acknowledged.

6. REFERENCES

[1] Meinsma, G. L. 1973. In memoriam Louise Kaiser: A remarkable woman. *Proc. Institute of Phonetic Sciences 3*, 122–149. Available online: https://www.fon.hum.uva.nl/archive/.

[2] Šturm, P. 2019. The birth of an institute: A centennial jubilee of Prague's Institute of Phonetics. *AUC Philologica* 2/2019, 9–26.

[3] Kaiser, L. (ed). 1957. *Manual of Phonetics*. North-Holland Publishing Company.

[4] Malmberg, B. (ed). 1968. *Manual of Phonetics [2nd ed]*. North-Holland Publishing Company.

[5] Hála, B. 1962. La nature phonétique de la syllabe. *Proc. 4th ICPhS* Helsinki, 418–427.

[6] Zwirner, E., Bethge, W. (eds). 1965. *Proceedings of the Fifth International Congress of Phonetic Sciences Held at the University of Münster 16–22 August 1964*. S. Karger.

[7] Hála, B. 1965. Sur la hiérarchie des éléments phoniques des sons du langage. *Proc. 5th ICPhS* Münster, 332–335.

[8] Hála, B., Romportl, M., Janota, P. (eds). 1970. *Proceedings of the Sixth International Congress of Phonetic Sciences Held at Prague 7–13 September 1967*. Academia.

[9] Romportl, M., Janota, P. (eds). 1972. *AUC Philologica 1, Phonetica Pragensia III: Symposium on Intonology, Prague, October 6–8, 1970*. Univerzita Karlova.

[10] Ohnesorg, K. (ed). 1972. *Colloquium Paedolinguisticum: Proceedings of the First International Symposium of Paedolinguistics Held at Brno, 14–16 October 1970*. Mouton.

Archival material used:

Archive of the Masaryk University, fond B91, box 14, inventory number 655:

- 6th ICPhS Organizing Committee reports

Unorganized archive of the Institute of Phonetics, Charles University:

- 6th ICPhS congress correspondence
- 6th ICPhS final report to the Czechoslovak Academy of Sciences
- 6th ICPhS materials (Circular #1, Programme)
- circulars of Louise Kaiser to members of the Permanent Council
- personal correspondence between Bohuslav Hála and Agostino Gemelli, Raoul Husson, Louise Kaiser, and Armando de Lacerda

ON THE EVOLUTION OF ICPhS BETWEEN 1932 AND 2019

Jürgen Trouvain

Language Science and Technology, Saarland University, Saarbrücken (Germany)
trouvain [at] lst.uni-saarland.de

ABSTRACT

A freely accessible online collection to all ICPhS proceedings (1932-2019) is used to study changes over eight decades in the community of phonetic sciences. ICPhS not only evolved from a meeting with 38 contributions to a congress with about 800 papers but also shows a gradual tendency from single to multiple authorship. "International" in its name, ICPhS starts as a European event in the 1930s, complemented by North America in the 1960s, with a constant rise of contributions from Asia in the last 20 years. English became the dominant congress language of writing in the 1970s, other languages used before 1995 were German, French, and occasionally Russian. Whereas female authors represent a minority in the 1930s (10%), the gender balance was reached in the 1990s, in contrast to technology-oriented speech research conferences. These facts are seen as first steps towards a more comprehensive, content-based analysis of the field reflected in these proceedings.

Keywords: history of phonetics, ICPhS, archiving

1. INTRODUCTION

The International Congress of Phonetic Sciences (ICPhS) is probably the most known, the most popular and one of the earliest conferences for all aspects of phonetic research in the world. Proceedings of meetings like ICPhS provide an optimal opportunity to record and learn more about the current topics and research themes at the time. Thus, congress proceedings represent an extremely useful resource to keep track of research development, besides journal publications. In a long-term view, congress proceedings over several decades provide the opportunity to study changes in the scientific field, be it on topics, theories and models, methods, measurement instruments, evaluation of findings, elicitation and assessment of data, but also on metadata such as gender, language, or country of research affiliation.

Since 2007 the proceedings of all ICPhS meetings are published online. To complement such effort, the International Phonetic Association (IPA) also made the proceedings of 2003 and 1999 available on their website. Recently, the Phonetics group at Saarland University has closed the gap between the proceedings which are accessible online (1999-2019) and those which are not online (1932-1995). The latter are now accessible at [1] as searchable pdf documents.

The following list includes all locations and years of past ICPhS meetings. Please note the rather long gap of 23 years between 1938 and 1961. This can be partially explained by the second world war (WWII), as well as other reasons [2]. No printed proceedings from ICPhS 1975 were available.

1. Amsterdam, Netherlands (1932)
2. London, United Kingdom (1935)
3. Ghent, Belgium (1938)
4. Helsinki, Finland (1961)
5. Münster, Western Germany (1964)
6. Prague, Czechoslovakia, now Czechia (1967)
7. Montreal, Canada (1971)
8. Leeds, United Kingdom (1975)
9. Copenhagen, Denmark (1979)
10. Utrecht, Netherlands (1983)
11. Tallinn, Soviet Union, now Estonia (1987)
12. Aix-en-Provence, France (1991)
13. Stockholm, Sweden (1995)
14. San Francisco, USA (1999)
15. Barcelona, Spain (2003)
16. Saarbrücken, Germany (2007)
17. Hong Kong (2011)
18. Glasgow, United Kingdom (2015)
19. Melbourne, Australia (2019)

Using the meta-data over the history of the congress, we concentrated on 1) the size of the meeting in terms of presented papers, 2) the development of multiple authorship, 3) the diversity of languages which the paper was written in, 4) the geographical location of research affiliation, and 5) the gender distribution of researchers. This information allows us to overview some sociological aspects of the research community in phonetic sciences over a period longer than typically inspected. The use of a longer period is important because it can reveal aspects of dominance and blind spots that possibly lead to biases in the research output.

2. NUMBER OF PAPERS

Counting the number of papers seems to be rather trivial. However, it requires a definition of what counts as a paper. In the first three congresses, papers were often transcripts of talks rather than manuscripts or written articles in a standard format [3]. It might also be the case that authors were unable or unwilling to submit their paper, so that it did not appear in the proceedings.

Fig. 1 shows the development of the size of each meeting in terms of number of published papers. We can see a dramatic expansion in 1987. From then onwards all congresses accepted between 500 and 800 papers. The drop in 2007 can be explained by the decision that full papers rather than abstracts had to be submitted and maybe with a more rigid acceptance policy along with a lesser known congress location.

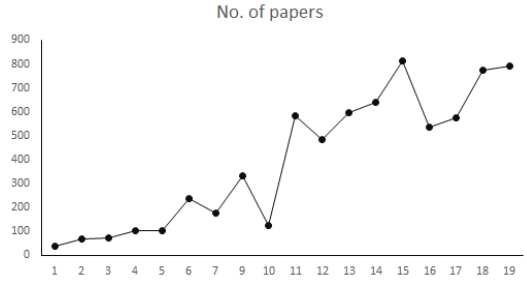


Figure 1: The absolute number of papers in the proceedings per year.

3. LANGUAGES IN PAPERS

Since 1995 English is the only language used in paper submission. Before that time, some degree of language diversity in published papers can be observed, see Fig. 2. Before WWII, German was the dominant language in 1932 and 1938, but not for the congress in 1935 in London. There were still 30% of all papers in German in 1967.

French was considerably used before WWII, but also thereafter, particularly when the congress was hosted in a French-speaking environment as it was the case 1971 in Montreal and 1991 in Aix-en-Provence. 1991 was the last time that papers were written in languages other than English.

Russian was also used at two meetings that were held in the *Eastern Bloc*: in 1967 (Prague, Czechoslovakia) with 34 papers, and in 1987 (Tallinn, Soviet Union, now Estonia) with 47 papers. It is more a side-remark that only one paper in Russian was given in 1971. The same is valid for Italian in 1932 and in 1983, respectively.

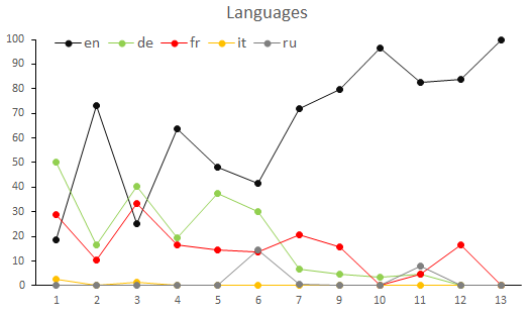


Figure 2: The relative number of papers written in different languages.

English has assumed a more dominant role as a language in research communication after WWII. Since 1971 constantly more than half of all papers per congress were written in English (leaving aside the peaks in 1935 in London, and in 1961, after WWII). From today's perspective, although English is the standard publishing language for most global or international conferences and journals, it is yet interesting to see that this dominance was present and established 50 years ago. The advantage of using just one language mastered to a certain degree by all congress delegates guarantees a common and constant means for sharing and exchanging ideas. The unavoidable disadvantage of monolingual exchange is the preferential treatment to native speakers of the dominant language being chosen, a group that usually represents a minority at such a conference.

4. NUMBER OF AUTHORS PER PAPER

From today's perspective, it is striking to see that exclusively all papers before WWII were written by only one author. In the beginning, single authorship was the rule, nowadays this is an exception (20%). This evolution is reflected in Fig. 3 where we can see a gradual change after WWII with a decreasing number of single-authored papers and an increasing number of papers with two authors. Papers with three authors increase with a slower pace but outnumber now single-authored papers. In the last congress in 2019, 14% of all contributions were written by four authors, a value reached in 1971 for two-authored papers and in 1983 for three-authored papers. Papers with five and more authors occur as well with the tendency to multiple-authorship, but they seem to be rather constant (below 10%).

The development from purely single to multiple authorship is clearly noticeable. The tendency to multiple authors is probably a consequence of the

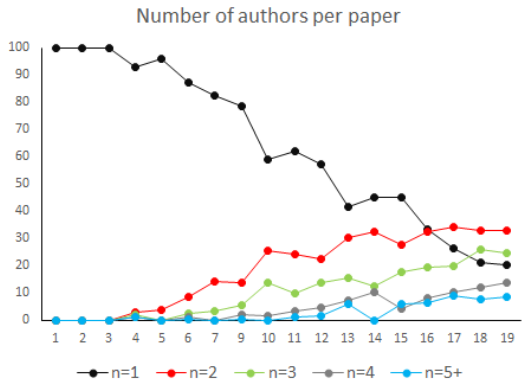


Figure 3: Percentage of papers pooled over the number of authors per paper.

fact that there are more researchers in general, and also more younger colleagues in early phases of their careers, which is likely to be linked to more employment opportunities and more labs.

5. GENDER

In relation to the paper counts reported in the prior sections, a pertinent aspect is to determine the gender of authors. It is not trivial and requires research on how to process or classify names. Typical challenges include that only initials are given, some names like *Kim* can be used for both sexes, and that for some names it was impossible to determine the sex (e.g. for romanized forms of Chinese names). In the end, nearly each congress contained a couple of unclear cases (ignored in this study). The absolute number of unclear cases increases with the number of authors, so the absolute numbers must be considered with some degree of uncertainty. However, the relative numbers do show a clear tendency.

In Fig. 4 we can see the gradual increase of female researchers in the phonetic community. However, this gradual increase does not start from the very beginning. Before WWII there were up to 18%, a value that was not unusual until 1983. Between 1987 and 2003, about a third of all authors were female. A real balance of nearly 50% has been reached in 2007 for the first time and has remained since then.

6. GEOGRAPHICAL AFFILIATION OF AUTHORS

Another important social factor which can be extracted from the metadata is the geographical affiliation of a paper's author/s presented at an international conference. For this parameter the country of the (first) research affiliation of the

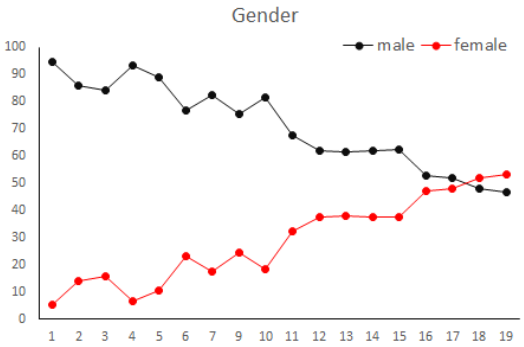


Figure 4: The relative number of authors pooled over gender of authors.

first author was determined. This also includes countries that do not exist nowadays like the Soviet Union, German Democratic Republic (GDR), or Czechoslovakia. In a number of cases, determining the country of affiliation required some further research when the country or the town was not indicated.

A pure list or table of countries would be rather complex. As a first approach we decided to allocate the countries to six continents: Europe (EUR), North America (NAM), Asia, South and Middle America (SAM), Africa (AFR), and Australia/Oceania (OCE). In Fig. 5 we can see the gradual shift from an exclusive European event in the 1930s to a real international conference where the authors with a European affiliation represent about a third in 2019. Contributions from North America started to increase after WWII. Papers from North America and Europe dominated with 90% of contributions. This dominant trend is gradually slowing in the millennium but still is rather strong with about two thirds in 2019. The trend was mostly contributed by authors from Asia.

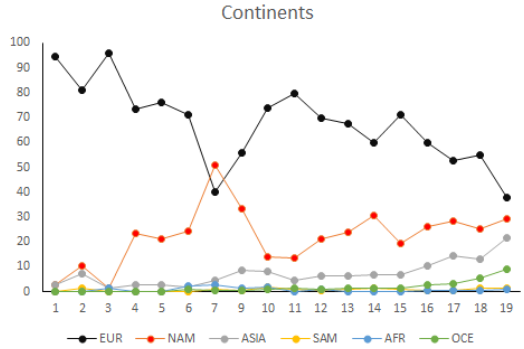


Figure 5: The relative number in % of first authors' first affiliations to countries binned by continents (see text for acronyms).

There is a clear association between congress location and country of delegates' affiliation. This link is also visible from the continent-wide perspective, e.g. the number of authors from North America at ICPhS in Montreal in 1971, or to a smaller extent, the number of authors from Oceania and perhaps also Asia at ICPhS in Melbourne in 2019.

A continent-wide view has its methodological weakness (e.g. both Saudia Arabia and Japan are currently classified under *Asia*, although geographically extremely distant). Nevertheless, it allows us to see the development from an intra-European congress to a conference with European and North American contributions to a meeting of the Northern hemisphere including Asia and Oceania as well. We can also see that South America and Africa are still underrepresented, although both are continents with the highest number of languages. Such imbalance in continental representation can have consequences for languages examined, recorded, taught, including phoneme inventory and supra-segmental characteristics.

7. SUMMARY AND CONCLUSIONS

Regarding the total number of papers ICPhS represents the largest congress in the phonetic sciences with an impressive growth over the years. It is obvious that the number of phonetic labs and also the number of researchers per lab have been growing. This is reflected by the increasing number of authors and the increasing number of countries different authors are affiliated with. Whereas female authors represented a minority in the 1930s (10%), the gender balance was reached in the 1990s. Regarding the number of authors per paper we can see a clear shift from single-authored to multiple- authored papers. Collaborative studies within a lab or between labs, often across countries, are the typical model. However, there seems to be a limit on the number of authors per paper. Regarding the countries of affiliation we can see that ICPhS evolved from a European to a much more international congress, still, with a great gap regarding countries of Africa, and to a lesser degree regarding countries of South America.

The freely available online access to the papers of all ICPhS proceedings allows us to conduct different types of studies with a historical perspective without the effort of procuring the primary literature resources. An example from the recent past is a study on the history of computational modelling and simulations in the phonetic sciences [4]. Other examples refer to historical views on explicitly this congress and this community as was performed by [5] and [2] on comparisons between two time points, or by [6] on longer periods.

It is also interesting to compare this to historical overviews of other related conference series. In a historical outline of 50 years of the speech research conference of the French-speaking community (JEP) we can see parallels regarding multiple authorship and gender balance [7]. The large-scaled analysis project for different conferences and journals in the field of Natural Language Processing by [8] also includes meetings like *Interspeech*. Here, we can see parallels as well regarding growing number of papers and multiple authorship but also a distinct difference regarding gender. There is also a balance in the number of male and female authorship, however, this convergence is extremely slower than at ICPhS and JEP. In the more technology-oriented conferences female authors represent only about 20% of all authors [8].

Next steps in the historical analysis of the ICPhS Proceedings should focus on topic- or content-specific issues. For instance, this could involve questions such as "When did acoustic measurements become popular?" or "How did *coarticulation* become an established technical term?". Concepts that are taken for granted nowadays such as *coarticulation* and *formant* were infrequently used in the congresses before WWII [3].

Other questions relate to the most-cited references over the years or also which languages were object of studies. Many more inspirations for content-based analysis in the history of a scientific discipline can be found in [8].

8. ACKNOWLEDGEMENTS

Thanks for the financial support of the Max Mangold estate which allows us to make the proceedings of earlier congresses available as searchable pdf documents [1]. The author is also grateful to Ivan Yuen and two anonymous reviewers for their feedback on earlier versions of the manuscript.

9. REFERENCES

[1] "ICPhS proccedings," https://www.coli.uni-saarland.de/groups/BM/phonetics/resources.html#icphs.

[2] J.-F. Bonnot and L.-J. Boë, "So where should the 4th ICPhS take place? And why did it never occur?" in *Proc. 14th ICPhS*, San Francisco, 1999, pp. 1325–1328.

[3] J. Trouvain, "On the early history of ICPhS," in *Proc. Fifth International Workshop on the History of Speech Communication Research (HSCR 2022)*, Porto, 2023.

[4] D. Duran, "On the history of computational modelling and simulations in the phonetic sciences," in *Proc. Second International Workshop on the History of Speech Communication Research (HSCR 2017)*, 2017, pp. 95–98.

[5] P. Šturm, "International phonetic congresses: the shift in research practices and areas of interest over 44 years," in *Proc. 18th ICPhS*, Glasgow, 2015.

[6] K. Kohler, "Two anniversaries: 75 years of international congresses of phonetic sciences and 50 years of Phonetica," *Phonetica*, vol. 64, pp. 73–79, 2007.

[7] V. Delvaux, G. Luxardo, and F. Hirsch, "Une histoire de JEP : 50 and d'études sur la parole," in *Proc. XXXIIe Journées d'Études sur la Parole*, Aix-en-Provence, 2018, pp. 365–373.

[8] J. Mariani, G. Francopoulo, and P. Paroubek, "The NLP4NLP Corpus (i): 50 years of publication, collaboration and citation in speech and language processing," *Frontiers in Research Metrics and Analytics*, vol. 3, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frma.2018.00036

PHONETIC DESCRIPTION OF FILLED PAUSES AS DISCOURSE MARKERS IN TOTOLI

Constantijn Kaland, Maria Bardají

Institute of Linguistics, University of Cologne, Germany
ckaland@uni-koeln.de, mbardaj1@uni-koeln.de

ABSTRACT

This study reports a phonetic analysis of filled pauses, similar to 'uh' and 'um' in English, in Totoli (Sulawesi, Indonesia). The focus of this study is on two types; 'a(h)' and 'e(h)'. The analysis was carried out on a corpus of spontaneous Pear Film retellings to investigate their frequency and position of occurrence, temporal characteristics and vowel quality. Results show that filled pauses function as markers of discourse structure, corroborating results for other languages. In particular, the Totoli filled pauses delimit two parts of the discourse that form different, semantically coherent units (i.e. 'paragraphs'). This function mainly applies to 'a'-like, rather than 'e'-like filled pauses. In addition, dialectal variation in the acoustic realisation of the filled pauses was found. The outcomes are interpreted as indications of the linguistic status of filled pauses in Totoli.

Keywords: filled pause, discourse marker, acoustic analysis, vowel quality, duration

1. INTRODUCTION

The endangered language Totoli is spoken in the Northern part of the Indonesian province Central Sulawesi by a decreasing number of speakers (approximately 7500 in [1]). The literature on this language is limited. Basic linguistic information and word lists are provided in [1], and a phonological and syntactic study of intonation units is reported in [2]. Verbal morpho-syntax has been studied to a larger extent ([3], [4], [5], [6], [7], [8], [9], [10], [11], [12]). Discourse structure has received limited attention, except for the recognition that multiple intonation units can form a paragraph. These often come with a specific sequence of intonation patterns, i.e. paragraph non-final intonation units end with a high boundary tone, whereas the final intonation unit in the paragraph ends with a low boundary tone [2].

Specific f0 movements in groups of subsequent intonation units (i.e. 'paratones') have been well-documented for English in read speech and spontaneous speech. They mark "a stretch of discourse presented by a speaker as forming a unit with a single topic" ([13], p.38), i.e. a paragraph. For Dutch, paragraphs showed an f0 downtrend over the course of the entire paragraph ('supradeclination', [14]). A follow-up experiment showed that listeners rate the paragraph as more natural when supradeclination is present than when it is absent. The perceptual relevance of paragraph intonation was furthermore shown for English listeners, who could hear whether a sentence was uttered in isolation or as part of a paragraph [15].

The literature has shown that there are other means to structure discourse at the level of the paragraph, such as pausing. For example, a study on Dutch showed that filled pauses (FPs) such as 'uh' and 'um' "carry information about larger-scale topical units" ([16], p.494). Filled pauses were more often present at strong breaks, which correspond to transitions between paragraphs, than at weak breaks, which correspond to within-paragraph boundaries. In addition, FPs were often preceded and followed by a silent pause at major breaks. If FPs occurred after a major break, they were most often found phrase-initially, i.e. still in the vicinity of the break. FPs were tentatively interpreted as linguistically relevant elements [16], which distinguishes them from 'placeholders' signaling whether the speaker has word-finding problems (e.g. [17]) and non-linguistic phenomena such as coughs and laughs.

Although traditional work studied FPs as disfluencies that are not part of the language (e.g. [18], [19]), more recent work has adopted linguistic interpretations. Thus, English 'uh' and 'um' were shown to be under control of the speaker just as any other word [20]. It was also shown that 'uh' occurs at minor breaks and 'um' at major ones and that their segmental makeup is very similar (central vowels) across languages spoken in Europe ([20], p.92) and also sound similar in Hebrew (e.g. 'eh', 'e-h', 'em', 'e-m', 'ah', 'a-m', e.g. [21]). Japanese FPs are segmentally different (e.g. 'eeto', 'etto', 'ano', 'anoo', 'uun', 'uunto', 'konoo', 'sonoo', 'jaa', e.g. [22]) and so do some Spanish ones ('este', e.g. [23]). It should be noted that

FPs in the latter two languages were derived from demonstratives [20]. Furthermore, studies reported dialectal variation in the production of FPs in English and Mandarin ([20], [24], [25]).

In the current study we explore the use of FPs in Totoli in order to advance our knowledge of their use in lesser documented languages. This is particularly valuable given the segmental similarity of FPs in well-studied European languages. There is no published documentation so far on Totoli FPs. The FPs in the available corpus [26] are mainly 'a(h)', 'e(h)' or 'mm(h)'. Given that the focus of the current study is on the potential discourse structuring function of Totoli FPs, 'mm(h)'-like FPs (backchannels) were excluded. We investigate 'a'- and 'e'-like FPs for their frequency of occurrence and their acoustic characteristics (temporal and vowel quality) in order to investigate the extent to which they contribute to discourse structure.

2. METHODOLOGY

This section describes the speech data, annotation and (acoustic) analyses. It is important to note that the notion of paragraph has been widely mentioned in the literature and proposed to be a grammatical unit at the level between discourse and phrase (e.g. [27]). However, to our knowledge, consensus on how to divide spontaneous spoken discourse into paragraphs is lacking. We will therefore outline the criteria used for Totoli in Section 2.2.

2.1. Participants and data collection procedure

Pear Film [28] monologue retellings [26] were elicited from 16 speakers (8F/8M; age range 31-82). All speakers were bilinguals (Totoli and Indonesian). Eight participants (four female, four male) were native speakers of the Totoli variety spoken in Tolitoli city and surrounding villages (henceforth 'southern dialect'), and the other eight participants (four female, four male) were native speakers of the variety of the Northern Tolitoli District (henceforth 'northern dialect').

The data was recorded with a Zoom Q8 camera and an AKG C520 external microphone. Participants were first shown the Pear Film, a 6-minute videoclip without speech widely used for linguistic elicitation [28], and were told to remember as many details as possible. Thereafter, participants were recorded whilst retelling the film to a Totoli interlocutor. The presence of the interlocutor ensured a more spontaneous nature of the recorded speech. Recordings in which there was a considerable amount of interaction between the interlocutors were excluded from analysis to ensure that all investigated speech was monologous.

2.2. Annotation

The selected recordings were transcribed in ELAN [29], translated into Indonesian by a Totoli native speaker and checked for consistency. Segmentation and annotation in Praat [30] consisted of two steps. First, segmentation into paragraphs and, second, segmentation and labeling of FPs with 'a' or 'e', depending on their perceived vowel quality. A paragraph is understood in this study as a part of discourse that is at least one intonation unit and is semantically coherent. Paragraph boundaries were determined based on prosodic, lexical, morpho-syntactic and semantic criteria. As for prosody, we followed the pattern of boundary tones described in Section 1 (i.e. [2], p.94). Final lengthening was also considered as a paragraph-final prosodic phenomenon. Lexical markers of discourse structure were words that (1) signal change of topic, e.g. *bali* 'so', *ingga (daan) noosa* 'not long after', *danna/daan* 'then', *tooka* 'finished', (2) introduce temporal clauses, e.g. *injan* 'after' or the Indonesian loan words *pas* 'exactly when' and *begitu* 'as soon as'. Morpho-syntactically, the repetition of the final part of the previous sentence as the initial part of the following sentence (tail-head linkage [31]) was taken as a possible paragraph transition. Semantically, paragraph boundaries were determined by dividing the discourse into coherent units, e.g. one event construal per paragraph. Not all prosodic, lexical or morpho-syntactic markers were consistently present, in which case we relied on semantic coherence to set the paragraph boundary.

2.3. Frequency of occurrence

The annotated FPs were categorized according to labeled vowel quality ('a' and 'e') and position relative to the paragraph (vowel quality measures in Section 2.5). Paragraph position was defined as occurring *between* two paragraphs, with the *start* of a paragraph, or *within* a paragraph. The items in each category were counted and a chi-square test was run to investigate whether their distribution was significantly different from chance level.

2.4. Temporal characteristics

The raw duration in ms was measured for all FPs. In addition, the distance to the start of the paragraph was measured for the *within* items. These distance measures were then analysed in a linear mixed model with distance measure (s) as response and vowel quality as factor. The model included a

random intercept for speaker.

2.5. Vowel quality

Acoustic measures of vowel quality were taken in order to confirm that the manual labels indeed correspond to different vowel targets acoustically. If the vowel qualities would overlap, it would be less plausible to assume a two-way distinction for the FPs. Vowel quality (F1 and F2) was measured following the procedure in [32]. This method takes formant measures from subintervals of the intervals corresponding to the FPs. The subinterval was set as the part of the FP for which the intensity dropped maximally 5% at either side of the intensity peak. Formant values are generally the most stable in this part. If the intensity level stayed within the 5% margin at the interval boundary, that boundary was taken as the boundary of the subinterval.

The formant measures were analysed in linear mixed models; one model for each formant. In both models the formant value in Bark was the response, with the interaction between labelled vowel quality (Vq: 'a','e') and dialect (Di: North, South) as factors, and with speaker and paragraph position (between, start, within) as random intercepts.

3. RESULTS

3.1. Frequency of occurrence

Table 1: Number of FPs (χ^2 residuals) split for vowel quality (Vq) and paragraph position.

Vq	Paragraph position		
	between	start	within
a	54 (0.59)	77 (0.90)	20 (-2.08)
e	12 (-1.04)	15 (-1.59)	22 (3.65)

The number of items occurring in each of the categories of FPs is given in Table 1. The results of the chi-square test show that their distribution was significantly different from chance level [$\chi^2(2, N = 200) = 22.42$, $p < 0.001$]. The chi-square Pearson residuals (observed - expected) furthermore indicate that 'a'-like FPs tend to occur in between paragraphs and at paragraph starts, whereas 'e'-like FPs tend to occur within paragraphs (Table 1).

3.2. Temporal characteristics

Table 2: Duration in ms (SD) of FPs split for vowel quality (Vq) and paragraph position.

Vq	Paragraph position		
	between	start	within
a	486.41 (138.67)	267.37 (101.96)	368.09 (199.41)
e	360.08 (148.87)	217.95 (80.18)	432.21 (212.12)

The raw duration measures (Table 2) show that the longest FPs (∼450 ms) are found between paragraphs for 'a' and within paragraphs for 'e'. Intermediate durations (∼360 ms) were found within paragraphs for 'a' and between paragraphs for 'e'. Both vowel categories showed the shortest FP durations (∼240 ms) at the start of paragraphs. The linear mixed model analysis showed that the distance to the start of the paragraph did not differ between 'a'-like items (μ = 4.69, sd = 3.89) and 'e'-like items (μ = 4.33, sd = 3.27): [β = -0.07, SE = 1.09, df = 39.91, t = -0.07, n.s.].

3.3. Vowel quality

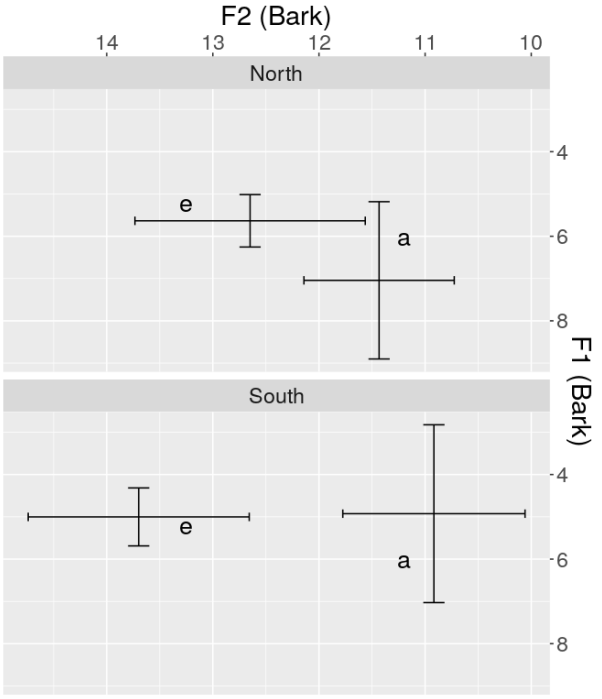


Figure 1: Formant measures for 'a'-like and 'e'-like FPs produced by speakers of the Northern (top) and Southern (bottom) dialect.

Table 3: Results of the linear mixed effects models for the formant measures (F1 and F2)

	Predictor	β	SE	df	t	p
F1	Intercept	7.26	0.47	13.05	15.57	< 0.001
	Vq	-1.22	0.32	186.68	-3.84	< 0.001
	Di	-2.38	0.61	13.29	-3.88	< 0.01
	Vq*Di	0.44	0.56	193.65	0.78	n.s.
F2	Intercept	11.51	0.21	12.43	56.06	< 0.001
	Vq	1.44	0.16	187.22	8.81	< 0.001
	Di	-0.44	0.31	13.17	-1.43	n.s.
	Vq*Di	0.70	0.29	194.09	2.40	< 0.05

The vowel quality results show that 'a' and 'e' were consistently labelled and differ acoustically in a significant way (Table 3). Dialect had a main

effect (F1) and an effect in interaction with vowel quality (F2). F1 values were lower for the Southern dialect than for the Northern dialect (Figure 1). The F2 interaction shows that the 'a' and 'e' values lie closer to each other in the Northen dialect than in the Southern dialect. In the Southern dialect, the F2 values are lower for 'a' and higher for 'e'.

4. DISCUSSION AND CONCLUSION

FPs in Totoli come in two vowel categories that show systematic differences according to paragraph position and speaker dialect. That is, the results revealed tendencies to use 'a'-like FPs near the boundaries of paragraphs (between or at their start), and 'e'-like FPs within paragraphs. Given that they were tendencies, there were still 'a'-like FPs used within pargraphs and 'e'-like FPs used near paragraph boundaries (Table 1). The tendencies were, however, further confirmed by the duration of the FPs; for both vowel categories the average FP was the longest at the position where it tended to occur more often (between and within paragraph). Note that 'a'-like FPs occurred often at the start of paragraphs but were the shortest at that position. This could be explained by the presence of following discourse material, that might have already been planned by the speaker.

The vowel quality measures furthermore confirmed that the vowel categories are acoustically different and that speakers from the northern Totoli dialect produce lower vowels which are more densely distributed within the acoustic space than the speakers of the southern dialect. The difference in vowel height (F1) and vowel place of articulation (F2) are interrelated. This can be explained according to the widely used visualisation of the acoustic space as a triangular shape (i.e. a vowel triangle). The lower the vowels (i.e. more toward the lower tip of the triangle), the closer their positions will be. In the current study, dialectel differences were the largest for 'a'; i.e. [a]-like targets in the southern dialect and [ɑ]-like in the northern dialect. Dialectal differences in FPs were also reported for other languages (e.g. [25]).

Taken together, the results of the current study point at a linguistic status of the FPs in Totoli. That is, they show clear distributional patterns with regard to their position in the paragraph. Their distribution suggests that they have a discourse marking function rather than being produced in an uncontrolled way. In particular the 'a'-like items seem to act as paragraph boundary markers. The possibility cannot be excluded, however, that among the 'e'-like items there might have been some hesitations, rather than discourse markers. This becomes particularly clear from the residuals of the FPs within paragraphs (Table 1). For those items the largest deviations from the expected values are observed, indicating the strongest tendency for 'a'-like items to *not* occur in that position as well as the strongest tendency for 'e'-like items to occur in that position. Thus, assuming that hesitations most often occur within paragraphs, where speakers might pause for planning reasons rather than for discourse structure reasons, the current data suggest that these would rather be 'e'-like. In addition, 'e'-like FPs have a wider spread in their F2 values (Figure 1), which indicates realisation toward the center of the acoustic space. Although central vowels have been reported to occur in linguistically relevant FPs crosslinguistically [20], they seem to be a minority in Totoli. Generally, FPs in Totoli occur as two vowels that are acoustically distinct. This could be taken as an additional indication of their linguistic nature. That is, their acoustic realisations map onto vowel targets and show dialectal variation. It could be speculated that 'a'-like FPs are derived from a demonstrative, i.e. *ia* (proximal), *ana* (medial) and (less likely) *itu* (distal), as listed in [1] (p.99). Also, it cannot be excluded that FPs originate from interjections such as *ya* or *ye* 'yes'. Totoli FPs thus show a clear correlation with linguistic units and are more articulate than central (schwa-like) vowels.

It remains to be seen whether there are more (discourse) structural differences in the use of 'a' or 'e' in Totoli FPs. One specific question pertains to the distribution of FPs at the start of paragraph-internal intonation units (the *within* items in this study). If 'e'-like items were indeed more often hesitations, it could be expected that they would not occur as often at the start of intonation units. That position could be occupied more often by 'a'-like items, just as they do at the level of paragraphs. This, as well as morpho-syntactic aspects of paragraphs, will be investigated in future studies.

5. ACKNOWLEDGEMENTS

Research for this paper was funded by the German Research Foundation (DFG) - Project-ID 281511265 - SFB 1252. The recordings were made by Christoph Bracks and the second author. The data (recordings, textgrids and dataframe) are made available at https://osf.io/yah89/. We gratefully acknowledge the Totoli speakers who participated in the recordings and contributed to the transcriptions.

6. REFERENCES

[1] N. P. Himmelmann, Ed., *Sourcebook on Tomini-Tolitoli languages: general information and word lists*, 1st ed., ser. Pacific linguistics. Canberra: Pac. Ling., Research School of Pacific and Asian Studies, Australian National U., 2001, no. 511.

[2] C. Bracks, "Intonation Units and grammatical units in Totoli," PhD Thesis, Universität zu Köln, 2020.

[3] S. Inghuong, A. Adnan, A. G. Hali, I. Halim, and N. Baso, *Morfologi dan sintaksis bahasa Totoli*. Palu: PPBSIDST, 1990.

[4] I. A. Sofyan, A. Adnan, Z. Mahmud, and M. Masyhuda, *Struktur bahasa Totoli*. Jakarta: Pusat Bahasa, 1991.

[5] N. P. Himmelmann and S. Riesberg, "Symmetrical Voice and Applicative Alternations: Evidence from Totoli," *Oceanic Linguistics*, vol. 52, no. 2, pp. 396–422, 2013.

[6] S. Riesberg, "A first take on information structure in Totoli - Reference management and its interrelation with voice selection," *Proceedings of the 2nd International Workshop on Information Structure of Austronesian Languages*, pp. 65–81, 2015.

[7] ——, "'Fern leaf, so you are challenging me?' Some observations on the Lelegesan, a form of verbal combat in Totoli," 2019.

[8] S. Riesberg, K. Malcher, and N. P. Himmelmann, "The many ways of transitivization in Totoli," in *Valency over Time*, S. Luraghi and E. Roma, Eds. De Gruyter, 2021, pp. 235–264.

[9] N. P. Himmelmann and S. Riesberg, "Expressions of directed caused accompanied motion events in Totoli, a western Austronesian language of Indonesia," in *Typological Studies in Language*, A. Margetts, S. Riesberg, and B. Hellwig, Eds. Amsterdam: John Benjamins Publishing Company, 2022, vol. 134, pp. 219–242.

[10] M. Bardají i Farré, S. Riesberg, and N. Himmelmann, "Limited-Control Predicates InWestern Austronesia: Stative, Dynamic, Or None Of The Above?" *Oceanic Linguistics*, 2021.

[11] S. Riesberg, M. Bardají i Farré, K. Malcher, and N. P. Himmelmann, "Predicting voice choice in symmetrical voice languages: All the things that do not work in Totoli," *Studies in Language*, vol. 46, no. 2, pp. 453–516, 2022.

[12] M. Bardají i Farré, "Nominalization in Totoli and other western Austronesian languages," PhD Thesis, Universität zu Köln, 2022.

[13] G. Yule, "Speakers' topics and major paratones," *Lingua*, vol. 52, no. 1-2, pp. 33–47, 1980.

[14] A. Sluijter and J. Terken, "Beyond Sentence Prosody: Paragraph Intonation in Dutch," *Phonetica*, vol. 50, no. 3, pp. 180–188, 1993.

[15] I. Lehiste, "The Phonetic Structure of Paragraphs," in *Structure and Process in Speech Perception*, K. S. Fu, W. D. Keidel, H. Wolter, A. Cohen, and S. G. Nooteboom, Eds. Berlin, Heidelberg: Springer, 1975, vol. 11, pp. 195–206.

[16] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[17] N. Amiridze, B. H. Davis, and M. Maclagan, Eds., *Fillers, pauses and placeholders*, ser. Typological studies in language. Amsterdam Philadelphia: Benjamins, 2010, no. 93.

[18] F. Goldman-Eisler, "A Comparative Study of two Hesitation Phenomena," *Language and Speech*, vol. 4, no. 1, pp. 18–26, 1961.

[19] M. Hayashi and K.-E. Yoon, "A cross-linguistic exploration of demonstratives in interaction: With particular reference to the context of word-formulation trouble," *Studies in Language*, vol. 30, no. 3, pp. 485–540, 2006.

[20] H. Clark and J. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[21] Y. Maschler, "Discourse markers at frame shifts in Israeli Hebrew talk-in-interaction," *Pragmatics*, vol. 7, no. 2, pp. 183–211, 1997.

[22] N. Iwasaki, "18 Japanese fillers as discourse markers: Meanings of "meaningless" elements," in *Handbook of Japanese Semantics and Pragmatics*, W. M. Jacobsen and Y. Takubo, Eds. De Gruyter, 2020, pp. 799–838.

[23] J. Brody, "Particles Borrowed from Spanish as Discourse Markers in Mayan Languages," *Anthropological Linguistics*, vol. 29, no. 4, pp. 507–521, 1987.

[24] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses," in *Disfluency in Spontaneous Speech*, 2005, pp. 179–182.

[25] Y. Tian, T. Maruyama, and J. Ginzburg, "Self Addressed Questions and Filled Pauses: A Cross-linguistic Investigation," *Journal of Psycholinguistic Research*, vol. 46, no. 4, pp. 905–922, 2017.

[26] M. Bardají i Farré and C. Kaland, "Phonetic description of filled pauses as discourse markers in Totoli," 2022, https://osf.io/yah89/.

[27] R. Longacre, "The Paragraph as a Grammatical Unit," in *Discourse and Syntax*, T. Givon, Ed. BRILL, 1979, pp. 113–134.

[28] W. L. Chafe, *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, N.J: Praeger, 1980.

[29] Max Planck Institute for Psycholinguistics, "ELAN," Nijmegen, 2022, https://archive.mpi.nl/tla/elan.

[30] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer," 2022, http://www.praat.org/.

[31] L. De Vries, "Towards a typology of tail–head linkage in Papuan languages," *Studies in Language*, vol. 29, no. 2, pp. 363–384, 2005.

[32] C. Kaland, "Acoustic correlates of word stress in Papuan Malay," *Journal of Phonetics*, vol. 74, pp. 55–74, 2019.

AERODYNAMICS AND ARTICULATION OF WORD-FINAL EJECTIVES IN EASTERN ARMENIAN

Tabita Toparlak, Hossep Dolatian

Yeditepe Üniversitesi; Stony Brook University
tabitatoparlak@gmail.com; hossep.dolatian@alumni.stonybrook.edu

ABSTRACT

Eastern Armenian is an Indo-European language with a three-way laryngeal contrast for stops: voiced, voiceless unaspirated, and voiceless aspirated, e.g., *b, p, pʰ*. Word-final voiceless unaspirated stops are variably ejectives: final *p* or *p'*. To analyze the phonetics of final ejectivization, we recorded three speakers reading from a word list of 51 words. The words had the three types of stops, at three places of articulation, at different syllable positions (initial, final, intervocalic). We measured the target stops in terms of their acoustics (VOT) and aerodynamics (oral airflow, intraoral pressure, and volume of exhaled air). We likewise recorded EGG signals for their articulation. We found that all three speakers almost always used ejectivization for word-final voiceless unaspirated stops. We suggest our speakers use ejectivization as a fortition process that creates a clearer phonetic contrast between final (ejectives) voiceless unaspirated stops vs. final voiceless unaspirated stops.

Keywords: Armenian, aspiration, ejectivization, aerodynamics, EGG

1. INTRODUCTION

Eastern Armenian is an Indo-European language with a 3-way laryngeal contrast. It is reported that word-final voiceless unaspirated stops can undergo ejectivization and become 'slightly ejective' [1]. We define such ejectivization as the use of laryngeal movement during the production of stops. This paper reports acoustic, aerodynamic, and articulatory data on final ejectivization in Eastern Armenian. We find that, for our sample of speakers, the voiceless unaspirated stops are pronounced as ejectives in the word-final position (V_#). We speculate that ejectivization acts as a type of phonetic enhancement. In the literature these consonants have been called either ejectives or glottalized. To avoid confusion, we adopted the term 'ejective' and 'ejectivization' for this paper.

2. BACKGROUND

Armenian is an Indo-European language. It forms a separate branch in the IE family, with two standardized varieties (Western and Eastern) that differ in the laryngeal contrasts of stops and affricates [2, 3, 4, 5, 6, 7]. Eastern has a three-way contrast between voiced, voiceless unaspirated, and voiceless aspirated stops [8, 9]: bilabial /b, p, pʰ/, coronal /d, t, tʰ/, and velar /ɡ, k, kʰ/.[1]

Table 1: Voicing contrasts in Armenian

	Eastern	Gloss	Orthography
/p/	[pɑɾ]	'dance'	պար
/pʰ/	[pʰɑk]	'closed'	փակ
/b/	[bɑr]	'word'	բառ

The three-way contrast has been confirmed in previous acoustic studies in terms of voice onset time (VOT) [10, 11, 12, 13, 14, 15]. Voiced stops have negative VOT, voiceless unaspirated have zero VOT or a short lag, and voiceless aspirated stops have long positive VOT. This paper replicates this finding.

Outside of Armenian, there is extensive work on the phonetics of ejectivization [16, 17, 18, 19, 20, 21, 22]. For Eastern Armenian, some phonological and phonetic studies report that some speakers variably use ejectivization for word-final voiceless unaspirated stops [23, 24, 25]. For example, the word կատակ 'joke' is typically pronounced with a final voiceless unaspirated stop [k]: *katak*. However, some speakers produce an ejective instead of a final voiceless pulmonic stop: *katak'*.[2]

It is unclear how widespread the use of final ejectivization is in Eastern Armenian. Some studies had participants who consistently used ejectivization [26], never used ejectivization [13], or a subset variably did [1, page 67]. One study [14] reports that 2 out of 8 speakers of Eastern Armenian (from Yerevan, Armenia) produced final ejectivization. Another study [27] reports that 4 out 225 tokens of voiceless unaspirates were ejectivized in their sample. The lack of final ejectivization is reported to be the prescriptive norm [28, page 17].

3. METHODOLOGY

Three speakers were recorded reading from a word list.[3] All three were native speakers of Eastern Armenian. Their age range was 25-38. Two speakers (M1 and M2) were male, born and raised in Armenia. One speaker (F3) was female, born and raised in Iran. The two regional dialects are not known to have significant differences for the laryngeal contrast [13, 29].

The stimuli was a word list of 51 words. Words were uttered in isolation without any carrier phrase or sentence. For each word, we asked the speaker to repeat the word three times. Some speakers would sometimes repeat 4 or 6 times by accident.

Each word had one or more of the following target stops: /p, t, k, p^h, t^h, k^h, b, d, g/. The stops were in different syllable positions. We report on a subset of these 51 words where where the target stops were found in one of the following positions: word-initial pre-vocalic (#_V), intervocalic (V_V), or word-final post-vocalic (V_#). In Table 2, we provide the recorded tokens per speaker and per context. Some stop-context combinations were sparsely populated, such as non-initial voiced stops.

Table 2: Tokens per speaker, context, and stop

		Bilabial			Coronal			Velar			Total
		b	p	p^h	d	t	t^h	g	k	k^h	
#_V	M1	9	6	6	9	3	9	12	3	6	63
	M2	9	6	6	9	3	9	12	3	6	63
	F3	9	12	6	9	3	9	12	3	6	69
	Total	27	24	18	27	9	27	36	9	18	195
V_V	M1		9	6	3	3	6	3	3	3	36
	M2		6	6	3	3	6	3	3	3	33
	F3		12	6	3	3	6	3	4	3	40
	Total		27	18	9	9	18	9	10	9	109
V_#	M1		3	3		6	3		12	3	30
	M2		3	3		6	3		12	3	30
	F3		6	3		6	3		12	3	33
	Total		12	9		18	9		36	9	93
Total		27	63	45	36	36	54	45	55	36	397

Participants were recorded by the first author in Paris 2017. Speaker F3 was recorded in a soundproof room of the Institut de Linguistique et Phonétique Générales et Appliquées at the Sorbonne Nouvelle university. Speakers M1-M2 were recorded in a quiet school-room. We recorded acoustic (VOT), aerodynamic (oral airflow, intraoral pressure, volume of exhaled air), and articulatory data. To get articulatory data, we used electroglottography (EGG) to visualize the contacts of the vocal folds [30]. To get aerodynamic data, we used a portable machine EVA2.

Our recording protocol was as follows. We placed an EGG device on the neck of the speaker. We asked the speaker to insert a 2mm diameter plastic tube inside the mouth. We attached a flexible rubber mask on the speaker's mouth. To avoid any air leakage, we used two types of masks, one for the female speaker and one for the male speakers. The speaker was asked to avoid blocking the tip of the tube as much as possible. We attached an inlet of the plastic tube to an EVA2. We ran a trial on the speaker to make sure that the equipment was stable. We recorded the stimuli where the target sound is a bilabial. We removed the tube that was in the mouth. We continued with the rest of the stimuli.

The signals were analyzed using the Phonedit Signaix software which is developed by the Aix-en-Provence Speech and Language Laboratory [31].[4] We used this software to record, visualize, and display multi-parameter data. Acoustic analysis was done with Praat [32]. Word-final segments were measured by taking the period between the first and last cycles of the signal.

4. RESULTS

All our word-final V_# tokens of /p, t, k/ were ejectives. This is likely because the stimuli were said in isolation, encouraging hyper-articulation. When discussing our measurements, we focus on the bilabial series for illustration.

4.1. Acoustic measurements: VOT

VOT was measured for the stops in all three syllable positions. Results are in Table 3, averaged across speakers. For measuring word-final VOT, we measured between the first and last cycles of the signal.

Table 3: Voice onset time per context and stop (in milliseconds)

		Bilabial			Coronal			Velar		
		b	p	p^h	d	t	t^h	g	k	k^h
#_V	Avg.	-75.67	19.18	62.12	-91.76	20.93	65.68	-79.63	31.74	66.58
	Std.	38.60	4.94	18.48	46.74	5.22	16.59	38.95	10.33	15.64
V_V	Avg.		20.72	40.73	-60.73	15.58	49.10	-49.18	34.25	45.19
	Std.		5.16	14.76	15.82	2.22	11.05	21.39	19.97	9.52
V_#	Avg.		33.71	81.47		42.30	85.74		36.00	52.24
	Std.		9.20	14.63		21.60	35.59		10.85	12.93

In all three syllable positions (initial, intervocalic, and final), the stops are distinguished in terms of negative, zero, and positive VOT. This is expected. Place has an effect on VOT values [10, 33, 19].

In the word-final position, we find clear VOT differences across 2 of the 3 laryngeal categories. Our stimuli lacked cases of word-final voiced stops. We impressionistically perceived that all three speakers pronounced word-final voiceless unaspirates as ejectivized or as ejectives.

4.2. Aerodynamic measurements

4.2.1. Intraoral pressure

Intraoral pressure was measured in hPa for the bilabial series. Results are in Table 4.

Table 4: Intraoral pressure for bilabials per context (in hPa)

		b	p	p^h
#_V	Avg (Std)	4.52 (1.31)	5.07 (2.1)	5.49 (1.36)
V_V	Avg (Std)		7.4 (1.73)	7.52 (1.32)
V_#	Avg (Std)		7.72 (3.77)	6.83 (1.66)

Within the bilabial series, first consider word-initial stops. the voiced stop /b/ has the lowest intra-oral pressure. The pressure is increased for the other two voiceless stops. For voiceless /p/, we see a rapid rise in pressure during closure (around 5 hPa), and then the pressure drops after the closure ends [34, 35]. For /p^h/, we also find that the pressure increases during the closure, and then reaches its maximum during the release (around 6hPa) [16]. Pressure is then dropped to around 1.4 dm^3/s.

Next consider intervocalic /p^h/. During closure, the intraoral pressure rises until release. At the release, the vocal folds are in contact, making the intraoral pressure have a zigzag shape, and making the EGG signal be not flat.

Finally consider final /p/ which surfaces as ejectivized [p'] as in [kɑrɑp'] 'swan'. After the vowel, the vocal folds spread to let the air pass. This causes an increase in intraoral pressure. Then the vocal folds close. During closure, the intraoral pressure increases until it reaches its maximum. At release, vocal fold vibration is stabilized.

4.2.2. Oral airflow

Oral airflow was measured in dm^3/s for all stops. Results are in Table 5.

Table 5: Oral airflow per context and stop (in dm^3/s)

		Bilabial			Coronal			Velar		
		b	p	p^h	d	t	t^h	g	k	k^h
#_V	Avg	0.33	0.34	0.91	0.26	0.44	1.16	0.24	0.39	1.17
	Std.	0.14	0.09	0.30	0.11	0.16	0.47	0.13	0.06	0.55
V_V	Avg		0.28	0.68	0.18	0.34	1.10	0.24	0.30	1.05
	Std.		0.09	0.32	0.06	0.07	0.20	0.07	0.07	0.43
V_#	Avg		0.38	0.93		0.36	0.76		0.26	0.66
	Std.		0.21	0.29		0.15	0.44		0.13	0.44

Consider word-initial bilabials. We see a rapid rise in airflow for /p/ at 0.56 dm^3/s. We find that the rise is strongest for the aspirated stops like /p^h/; this is expected [36].

For intervocalic /p^h/, the air flow reaches approximately 0.60 dm^3/s.

For the aspirate [p^h], the release is early, gradual, with significant airflow. But for the ejective [p'], the release is later; which implies that the delay of the release is due to the compress of the air between the closed larynx and the lips. Hence, when the air is released, the releasing is more sudden, with nearly no airflow.

Word-finally, the voiceless unaspirated stops are ejectivized. They differ from the aspirated stops because the latter have a higher air flow.

Across the three places, the coronal stop [t'] displays some differences. During the articulation of final [t'], we see that the glottis is still adducted and that the airflow is relatively low. The noise during release is smaller than that of the bilabial [p'] and velar [k']. For velar [k'], the noise upon release is more abrupt and sounds more intense.

4.2.3. Air volume

During the VOT, we calculated the air volume by taking the integral of the oral air flow. We treat the air volume as the amount of pulmonary air expelled from the mouth, per unit of time measured in dm^3/s (Table 6).

Table 6: Air volume per context and stop (in dm^3/s)

		Bilabial			Coronal			Velar		
		b	p	p^h	d	t	t^h	g	k	k^h
#_V	Avg	0.0022	0.0026	0.0384	0.0011	0.0036	0.0480	0.0024	0.0046	0.0461
	Std	0.0021	0.0015	0.0222	0.0010	0.0023	0.0287	0.0017	0.0022	0.0324
V_V	Avg		0.0022	0.0197	0.0009	0.0016	0.0297	0.0018	0.0049	0.0244
	Std		0.0017	0.0157	0.0003	0.0007	0.0127	0.0010	0.0039	0.0145
V_#	Avg		0.0056	0.0519		0.0096	0.0463		0.0050	0.0198
	Std		0.0042	0.0175		0.0112	0.0468		0.0037	0.0173

Consider word-initial bilabials. the largest amount of air is expelled for the voiceless aspirated stop /p^h/ than for its voiced and unaspirated counterparts /b, p/. Thus, aspirated stops have the highest air volume.

Next consider intervocalic /p^h/. Air volume is directly proportional to air pressure for 2/3 of the speakers. During VOT, the pressure reaches its maximum.

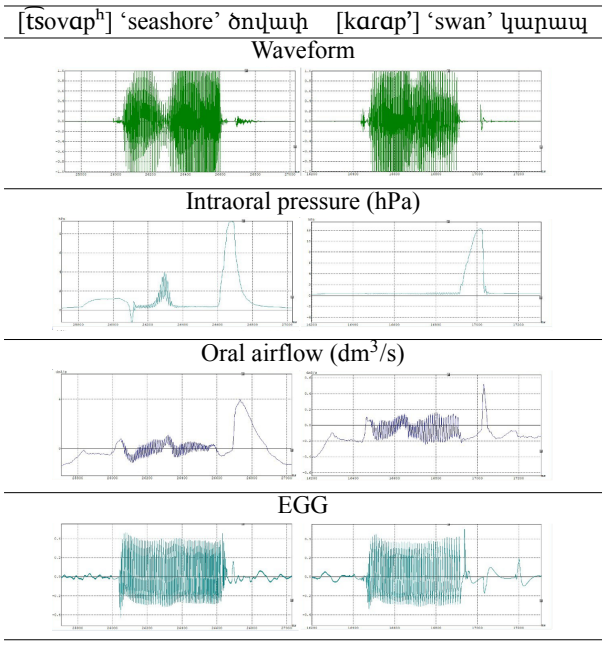
Word-finally, the voiceless unaspirated stops are ejectivized. They differ from the aspirated stops because the latter have a higher air volume.

4.3. Articulatory properties: EGG

We recorded the EGG signal as well. We provide illustrations of the EGG recordings in Figure 1 for the word [t͡sovɑp^h] with [p^h] and [kɑrɑp'] with an ejective [p'], produced by speaker M2. We provide the charts for the intraoral pressure and oral airflow. The audio is peak clipped, but that should not affect

measurement of VOT.

Figure 1: Contrasting final aspirated [pʰ] vs. ejective [p’]



The first screenshot of the EGG for the aspirated final [pʰ] implies simultaneous opening of the glottis and the release of the air. However, at the second, for the ejectivized final [p’], the glottis is held closed for a while, the pressure is raised during the closure of the glottis, and then there is a sudden release.

5. DISCUSSION

Our speakers produced final voiceless aspirated stops as ejectives. In contrast, for prescriptive Eastern Armenian speech, the norm is to lack ejectivization [28, page 17]. Acoustically, the final voiceless unaspirated stops have short-lag VOT. Thus they are unaspirated acoustically. But based on the aerodynamic data, they act as ejectives.

In our phonological interpretation, final ejectivization seems to be too speaker-variable to be allophonic. However, final ejectivization is not phonologically contrastive. It seems that final ejectivization is an optional fortition process in Eastern Armenian. Phonetically, we suggest that ejectivization acts as a way to phonetically enhance the contrast between final voiceless unaspirated stops and final voiceless aspirated stops. Glottal reinforcement of final voiceless stops is cross-linguistically attested [37]. Through ejectivization, the voiceless status of the final voiceless unaspirated stops is enhanced by the anticipation of glottal closure. During the stop closure, the glottal pressure occurs. The ejectivization does not replace the sounds [p, t, k]; it reinforces them.

This interpretation has been suggested in the past for Armenian. [1, page 67] demonstrate that the difference between voiceless unaspirated and aspirated stops is due to the force of the release. The unaspirated stops are released more “weakly”, or sometimes they are not released at all. In contrast, the aspirates have a relatively shorter closure duration, and then release, then noise and a flow of oral air. For voiceless unaspirated stops, they say that Armenian speakers “probably” have glottal closure during articulation and therefore sometimes slightly ejectivize.

6. CONCLUSIONS

Eastern Armenian has a 3-way laryngeal contrast between voiced, voiceless unaspirated, and voiceless aspirated stops. The language is reported to have word-final ejectivization of voiceless unaspirated stops, as a fortition process. Such a process is speaker-specific and highly variable across speakers and geographic regions.

Based on our data, we found three speakers who ejectivize their final voiceless unaspirated. Glottal closure exists during the articulation of these final voiceless unaspirates. The ejectivzed stops have noise during release, and the amount of air flow is not great. From an articulatory point of view, the glottis is closed during the production of these stops, thus triggering ejectivization.

From an articulatory point of view, the ejectivized consonants are extensions of the voiceless unaspirated consonants. And simultaneously from an acoustic angle, the ejectivized consonants [p’, t’, k’] are in contrast with their aspirated counterparts [pʰ, tʰ, kʰ] in the word-final position.

7. ACKNOWLEDGEMENTS

We thank Dr. Didier Demolin, Sipana Tchakerian, M. Rosario Signorello, and M. Nail Aras for help in setting up the experiment in Paris. We thank Scott Seyfarth and Canaan Breiss for helping us with guiding us though the literature, and we thank Jeremy Steffman for help in interpreting the data.

8. REFERENCES

[1] P. Ladefoged and I. Maddieson, *The Sounds of the World’s Language*. Malden, MA: Blackwell Publishing, 1996.

[2] H. Adjarian, *Classification Des Dialectes Arméniens*. Paris: Librairie Honoré Champion,

1909. [Online]. Available: https://archive.org/details/bibliothquedel173ecol/page/n7/mode/2up
[3] B. Vaux, *The Phonology of Armenian*. Oxford: Clarendon Press, 1998.
[4] L. Baronian, "Two problems in Armenian phonology," *Language and Linguistics Compass*, vol. 11, no. 8, p. e12247, Aug. 2017.
[5] N. E. Kelly and L. Keshishian, "The voicing contrast in stops and affricates in the Western Armenian of Lebanon," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1721–1725.
[6] ——, "Voicing patterns in stops among heritage speakers of Western Armenian in Lebanon and the US," *Nordic Journal of Linguistics*, vol. 44, no. 2, pp. 103–129, Oct. 2021.
[7] T. L. Tahtadjian, "Phonetic interference in the production of stops by Western Armenian bilinguals," Master's thesis, University of Toronto, 2021.
[8] E. W. Johnson, "Studies in East Armenian grammar," Ph.D. dissertation, University of California, Berkeley, 1954.
[9] Ա. Խաչատրյան, *Ժամանակակից Հայերենի Հնչույթաբանություն [Phonetics of Contemporary Armenian]*. Երևան: Հայկական ՍՍՀ Գիտությունների Ակադեմիա Հրատարակչություն, 1988.
[10] L. Lisker and A. S. Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," *Word*, vol. 20, no. 3, pp. 384–422, Jan. 1964.
[11] A. Braun, "An early case of "VOT"," in *Interspeech 2013*. ISCA, Aug. 2013, pp. 119–122.
[12] N. Hacopian, "A three-way VOT contrast in final position: Data from Armenian," *Journal of the International Phonetic Association*, vol. 33, no. 1, pp. 51–80, Jun. 2003.
[13] B. Amirian, "A study of acoustic features of stops and affricates in Eastern Armenian [in Persian]," Master's thesis, Department of Linguistics, Allameh Tabataba'i University, 2017.
[14] S. Seyfarth and M. Garellek, "Plosive voicing acoustics and voice quality in Yerevan Armenian," *Journal of Phonetics*, vol. 71, pp. 425–450, Nov. 2018.
[15] T. Toparlak, "Etudes phonétiques en arménien," Master's thesis, Université Paris 3 - Sorbonne Nouvelle, 2019.
[16] J. C. Catford, *Problems in Phonetics*. Midland Books, 1977.
[17] I. Maddieson and S. F. Disner, *Patterns of Sounds*, 1st ed. Cambridge University Press, Sep. 1984.
[18] J. C. Kingston, "The phonetics and phonology of the timing of oral and glottal events," Ph.D. dissertation, University of California, Berkeley, 1984.
[19] T. Cho and P. Ladefoged, "Variation and universals in VOT: Evidence from 18 languages," *Journal of Phonetics*, vol. 27, no. 2, pp. 207–229, Apr. 1999.
[20] P. D. Fallon, *The Synchronic and Diachronic Phonology of Ejectives*. Routledge, 2002.
[21] M. Stevens and J. Hajek, "A preliminary investigation of some acoustic characteristics of ejectives in Waima'a: VOT and closure duration," in *Proceedings of the 10th Australian International Conference on Speech Science & Technology*. Sydney: Macquarie University, 2004, pp. 277–282. [Online]. Available: https://epub.ub.uni-muenchen.de/id/eprint/14228
[22] G. Gallagher, "The perceptual basis of long-distance laryngeal restrictions," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.
[23] G. H. Fairbanks and E. W. Stevick, *Spoken East Armenian*. New York: American Council of Learned Societies, 1975.
[24] A. H. Khachatrian, "Some peculiarities of literary Armenian voiceless stops," *Annual of Armenian linguistics*, vol. 17, pp. 47–53, 1996.
[25] H. C. Fleming, "Glottalization in Eastern Armenian," *Journal of Indo-European Studies*, vol. 28, no. 1/2, pp. 155–196, 2000.
[26] W. S. Allen, "Notes on the phonetics of an Eastern Armenian speaker," *Transactions of the Philological Society*, vol. 49, no. 1, pp. 180–206, Nov. 1950.
[27] G. Schirru, "Laryngeal features of Armenian dialects," in *The Sound of Indo-European: Phonetics, Phonemics, and Morphophonemics*, B. N. Whitehead, T. Olander, B. Olsen, and J. E. Rasmussen, Eds. Copenhagen: Museum Tusculanum Press, 2012, pp. 435–457.
[28] J. Dum-Tragut, *Armenian: Modern Eastern Armenian*, ser. London Oriental and African Language Library. Amsterdam: John Benjamins Publishing Company, Dec. 2009, vol. 14.
[29] H. Dolatian, A. Sharifzadeh, and B. Vaux, "Grammar of Iranian Armenian: Parskahayeren or Iranahayeren," in review, unpublished manuscript.
[30] B. Gick, I. Wilson, and D. Derrick, *Articulatory Phonetics*. John Wiley & Sons, 2012.
[31] D. Demolin, "Aerodynamic techniques for phonetic fieldwork," in *Proceedings of the 17th International Congress of Phonetic Sciences. Hong Kong*, 2011, pp. 84–7.
[32] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.
[33] L.-m. Chen, K.-Y. Chao, and J.-F. Peng, "VOT productions of word-initial stops in Mandarin and English: A cross-language study," in *ROCLING 2007 Poster Papers*. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 2007, pp. 303–317.
[34] S. Hertegård and J. Gauffin, "Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 85–100, Feb. 1995.
[35] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time

in aphasia, apraxia of speech and dysarthria: A review," *Clinical Linguistics & Phonetics*, vol. 14, no. 2, pp. 131–150, Jan. 2000.
[36] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, CA: Singular, 2000.
[37] P. Ashby, B. Comrie, and G. Corbett, *Understanding Phonetics*. Routledge, Oct. 2013.

[1] The coronal series is typically dental, but some speakers use alveolar stops.
[2] Victoria Khurshudyan informs us that word-initial ejectivization is likewise attested in natural speech. Though we've been unable to find such cases in our data.
[3] We also recorded a fourth speaker. We had to set aside their recordings because we later learned they had ear surgery and their voice was altered.
[4] http://www2.lpl-aix.fr/ lpldev/phonedit/

FINAL LENGTHENING ACROSS VARIOUS SOUND CLASSES IN A LANGUAGE DOCUMENTATION CORPUS OF LOWER SORBIAN

Ludger Paschen

Leibniz-Zentrum Allgemeine Sprachwissenschaft
paschen@leibniz-zas.de

ABSTRACT

This paper investigates segmental lengthening before speech pauses in Lower Sorbian, an endangered Slavic language. The paper compares duration of vowels, sonorants, fricatives and other obstruents at various distances to speech pauses, based on a sample of 15,000 word tokens from a corpus of spontaneous speech. Results indicate lengthening of segments directly adjacent to boundaries for vowels, sonorants, and fricatives, but not for stops and affricates. The effect of lengthening was stronger for more sonorous sounds (vowels, sonorants) than for less sonorous sounds (fricatives). No evidence for lengthening of stressed vowels in (ante-)penultimate syllables was found, suggesting a narrow domain of lengthening that does not extend beyond the final syllable.

Keywords: Final lengthening, Lower Sorbian, corpus phonetics, elasticity.

1. INTRODUCTION

Lower Sorbian is an endangered West Slavic language spoken in the state of Brandenburg in Germany. While its sound inventory shares many common features with other Slavic languages, its phonetics and phonology are severely underdescribed, also compared to the closely related Upper Sorbian language [1, 2, 3, 4, 5]. The goal of this paper is to describe the durational properties of vowels and consonants before pauses in LS to test the scope and degree of *final lengthening* (FL), a common cross-linguistic process which lengthens segments before prosodic boundaries [6, 7, 8, 9]. More specifically, the paper tries to answer the following questions: Does FL in LS proceed progressively, i.e. are segments closer to a boundary more strongly affected than segments with greater distance to a boundary, as described in [10, 9]? Are all types of sounds (vowels, sonorants, fricatives, other obstruents) equally affected by FL or are there differences, as suggested by [11, 9]? Is the effect of FL stronger for stressed syllables compared to unstressed syllables [6, 12]?

2. DATA AND METHODS

2.1. The Lower Sorbian DoReCo dataset

DoReCo is a corpus containing time-aligned speech data from 51 endangered or minority languages from around the world [13]. Audio and annotations are freely available under CC-BY or similar licenses from the website: https://doreco.huma-num.fr/. The Lower Sorbian DoReCo dataset used in this study [14] derives from language documentation data within the DoBeS project archived at TLA [15]. The corpus includes more than 100 hours of transcribed interviews conducted between 2010 and 2015. The Lower Sorbian DoReCo dataset is a subset of the one archived at TLA consisting of 15 recordings of narrative texts from four speakers. It contains a total of 95 minutes of time-aligned transcriptions and 14,597 word tokens, or 41,923 individual phones.

2.2. Annotation and labeling

All recordings from the DoReCo core set feature time alignment on the level of phones, words, and inter-pausal units (IPUs). Time alignment was created using the MAUS forced aligner [16], with manual corrections of word and pause boundaries. DoReCo also includes manually annotated labels for special speech events such as hesitations, false starts, filled pauses, but also code-switching, singing, or unidentifiable stretches of speech (e.g. due to background noise). For the purpose of this study, all labeled tokens were excluded from the analysis.

2.3. Data preparation

A CSV file with phone start and end times from the Lower Sorbian DoReCo dataset was read in RStudio [17]. First, subsets were created for four major sound classes: vowels, sonorants, fricatives, and non-continuant obstruents, i.e. stops and affricates. Then, a number of segments were excluded:

- Labeled segments (see Sec. 2.2)
- Segments with a duration $\leq$ 30 ms
- Segments with a duration $>$ 3sd above the mean of the sound group
- Word-initial segments

The decision to set a global lower duration threshold of $>$ 30 ms was motivated by the fact that 30 ms was set as the minimum segment duration of MAUS, and segments of 30 ms are sometimes the result of last resort, and hence less accurate, alignments. Excluding segments with a duration $>$ 3sd above the mean in the respective sound group was done to remove outliers, again mostly corresponding to alignment errors. Word-initial phones were filtered in order to avoid effects of word-initial strengthening and/or weakening [18, 19]. This exclusion step also accounts for the fact that the onset of stop consonants following silent intervals cannot be determined with sufficient accuracy from the acoustic signal. After filtering, the remaining dataframe contained a total of 23,861 phones, of which there were 10,847 vowels, 6259 sonorants, 2666 fricatives, and 4089 plosives and affricates.

2.4. Statistical modeling

To test the hypothesis that segments are lengthened in the vicinity of speech pauses in Lower Sorbian, linear mixed models were run in R for each sound class (1). Models were run repeatedly with different values of the `position` variable to capture all possible combinations of positional effects, and the obtained significance levels were Bonferroni corrected (α*3).

(1)
```
log(duration) ~ position
+ (1|speaker) + (1|phone)
+ (1|left_context)
+ (1|speech_rate)
+ (1|pause_duration)
```

To account for speaker-specific properties, intrinsic segment duration, and coarticulatory effects, random intercepts for *speaker*, *phone*, and *left context* were included. Local *speech rate* was calculated as phones per second for each IPU. The factor *pause duration* was added to the model because pause duration may reflect the strength of prosodic boundaries, with shorter pauses correlating with weaker boundaries or greater textual cohesion [20].

To test the effect of lexical stress, another model was run on the vowel subset. This model had the same basic structure as in (1), but had the interaction between `position*stress` as the predictor variable, with `stress` being a binary variable distinguishing stressed and unstressed vowels.

3. RESULTS

3.1. An example of FL in Lower Sorbian

This subsection presents an example of context-dependent lengthening of the same word in two excerpts taken from the recording doreco_lowe1385_MEW-113-20121214, which is part of the Lower Sorbian DoReCo dataset [14]. The word *kutšu* 'carriage:ACC.F' appears twice in the examples, once in phrase-medial position (2) and once before a pause (3).

(2) doreco_lowe1385_MEW-113-20121214, 00:32–00:40
ten kněcht to (0.3) jo tu ***kutšu*** *pśigótował a farař se sednuł […] a pón su (.) z kónjom jěli (0.3) tam a cas (0.5)*
"The servant would prepare the carriage and the pastor would sit on it and then they went horseback riding from time to time."

(3) doreco_lowe1385_MEW-113-20121214, 01:48–02:04
ten jo (.) šeł flink domoj (0.5) jo holował (.) jogo (.) ***kutšu*** *(0.6)*
"He quickly went home, fetched his carriage […]."

Figure 1 illustrates the temporal properties of *kutšu* in different prosodic contexts. Image a. on the left corresponds to the token in (2), with a total duration of 520 ms. The duration of the word-final /u/ is 130 ms. In contrast, image b. on the right shows the *kutšu* token in (3), which displays significant lengthening of the final /u/ to 250 ms, nearly twice as much as the phrase-medial vowel in (2). The affricate /tʂ/ and the /u/ in the preceding syllable exhibit only minor traces of lengthening. Note that the onset of the plosive /k/ cannot be reliably determined in b. because it is preceded by a short silent pause.

We thus observe prosodically-driven differences in the acoustic realizations of the same word, with lengthening occurring mostly in the final vowel, and to a lesser extent also in the preceding two segments. The following section will provide a quantitative analysis of lengthening in the Lower Sorbian DoReCo dataset.

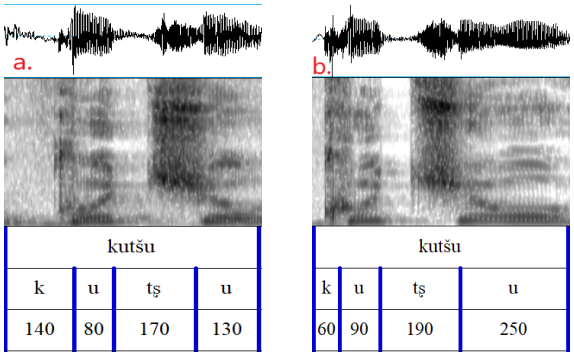


Figure 1: Praat editor windows showing time alignments for *kutšu* 'carriage' a. in phrase-medial position (2) and b. in pre-pausal position (3).

3.2. Quantitative results

Figure 2 shows the duration of vowels in the Lower Sorbian dataset, where "-n" refers to the n-th position before an IPU boundary and "Non-final" refers to all positions outside the prepausal 3-segment window. The boxplot reveals a gradual increase in segmental duration towards the prosodic boundary, and a notably longer duration of prepausal vowels (-1) compared to all other positions (-2, -3, Non-final).

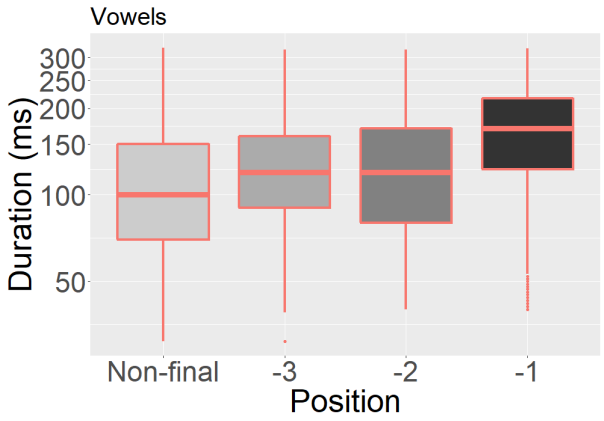


Figure 2: Vowel duration in various positions before IPU boundaries.

Differences between -1 and all other positions pause were highly significant ($p < 0.001$***). No significance was reached for differences between Non-final and -2 ($p = 0.180$); however, differences between Non-final and -3 were significant ($p < 0.01$**).

Sonorants show fairly similar durational patterns to vowels (Fig. 3). Again, a gradual increase in duration towards an IPU boundary can be observed, and segments in close proximity to a boundary (-1) were significantly longer than in any of the other tested positions ($p < 0.001$***). The interactions between other neighboring positions were also significant, albeit not as strongly (-2 vs. -3: $p < 0.05$*, Non-final vs. -3: $p < 0.05$*).

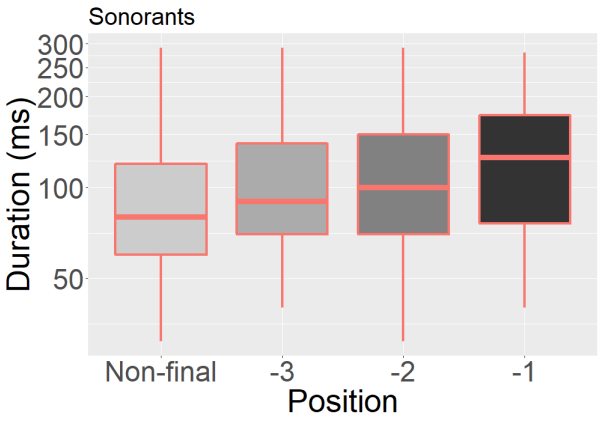


Figure 3: Sonorant duration in various positions before IPU boundaries.

Fricatives behave in a similar way to sonorants. Mean differences between positions were lower, however, suggesting a weaker effect of final lengthening (Fig. 4). Differences between positions reached significance in all but two cases: -2 vs. Non-final ($p = 0.160$) and -2 vs. -3 ($p = 0.864$).

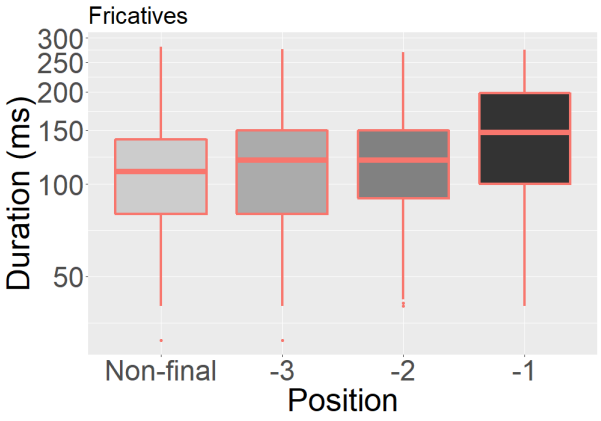


Figure 4: Fricative duration in various positions before IPU boundaries.

A divergent picture is presented by non-continuant obstruents (Fig. 5). No clear trend for durational differences is visible and durational differences did not reach significance, with one exception (-2 vs. -3, $p < 0.05$*). It is not clear why the -3 position shows a longer duration than the -2 position. One would expect that if any effect of FL exists at all for stops or affricates, it should be between Non-final and -1 rather than between -2 and -3. A potential explanation could be the phonotactics of Lower Sorbian, which heavily restricts the types of consonants that can appear word-finally. In the sample analyzed in this study, only three stops /p, t, k/ and two affricates /ts, tɕ/ occur in word-final position. The absence of voiced and palatalized sounds, which are generally associated with more

complex laryngeal and articulatory movements and thus greater intrinsic duration, may explain the lack of an effect for the -1 position in this group of sounds.

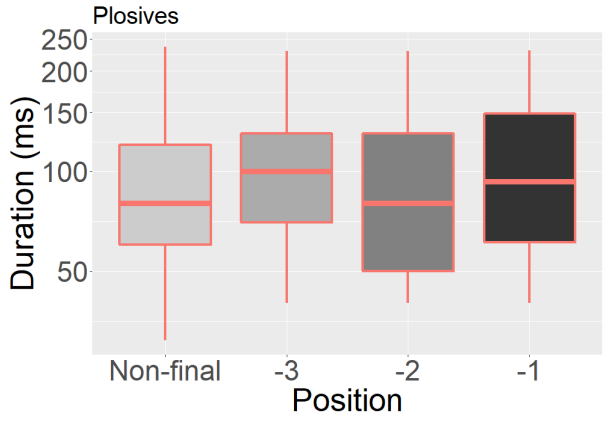


Figure 5: Stop and affricate duration in various positions before IPU boundaries.

Fig. 6 shows the duration of vowels in various syllable positions, split into unstressed and stressed syllables. Unsurprisingly, vowels in boundary-adjacent syllables (σ-1) were on average longer than in other syllables.Linear mixed models revealed several significant interactions both within and across position and stress groups. Depending on the base level, though, not all interactions involving stress reached significance, especially for the σ-2 and σ-3 positions. Estimates for σ-2 and σ-3 were shorter than for σ-1 and Non-final, suggesting that final lengthening in Lower Sorbian does not extend to the penultimate and antepenultimate syllables even when a stressed syllable occurs as a potential attractor in those positions.

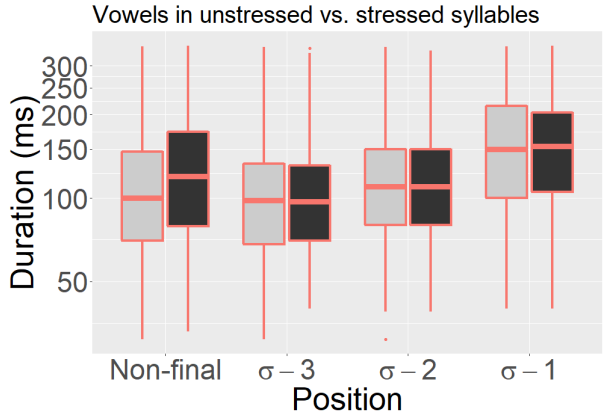


Figure 6: Vowel duration in unstressed (grey boxes) and stressed (black boxes) syllables in various positions before an IPU boundary.

4. DISCUSSION

Final lengthening in Lower Sorbian can be compared to other Slavic languages and to German, its major contact language. [21] reports longer syllable duration before prosodic boundaries in Polish. [22] describe lengthening of vowels and onset consonants in the two pre-boundary syllables in Czech, and an intricate interaction between final lengthening and phonological length. [23, 24] observe vocalic lengthening before prosodic boundaries and an increased effect before boundaries accompanied by pauses in Russian. They also report differences between stressed and unstressed vowels. [12] note systematic lengthening of IP-final vowels in Serbian.

For German, [9] present articulatory and acoustic evidence for lengthening of vowels and onset consonants, whereby lengthening is stronger in the ultimate syllable and more sonorous sounds are lengthened to a greater extent than less sonorous segments. This is well in line with the results obtained for Lower Sorbian, and confirms previous findings on general segmental elasticity [25, 26] as well as the relevance of those features for FL [11]. It should be noted, though, that the above mentioned studies on Slavic and German are experimental studies based on stimulus-based reading tasks, while the Lower Sorbian dataset consists of interviews and narrative texts. Ideally, as more experimental and corpus data become available in the future, these two types of data can complement each other by highlighting different aspects of the temporal organization of speech. Another promising point of departure for future research could be a closer look at various types of pauses such as silent, filled, and breath pauses in order to shed more light on the relation between speech planning, discourse structure, and the durational properties of segments.

5. CONCLUSION

This paper has offered an account of final lengthening in Lower Sorbian. Final Lengthening affects vowels, sonorants, and fricatives, and its effect is contingent on both the sonority of a segment and on the distance to an utterance boundary. The Lower Sorbian data show similarities with patterns reported for other Slavic languages and German. The study exemplifies the reusability of speech corpora originally created with the aim of language documentation.

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the German Research Foundation (DFG-PA2368/1-1).

7. REFERENCES

[1] P. J. Howson, "An acoustic examination of the three-way sibilant contrast in Lower Sorbian," in *Proc. Interspeech 2015*, 2015, pp. 2670–2674.

[2] ——, "Upper Sorbian," *Journal of the International Phonetic Association*, vol. 47, no. 3, pp. 359—367, 2017.

[3] ——, "Rhotics and palatalization: an acoustic examination of Upper and Lower Sorbian," *Phonetica*, vol. 75, pp. 132—150, 2018.

[4] P. J. Howson and A. Kochetov, "Lowered F2 observed in uvular rhotics involves a tongue root gesture: Evidence from Upper Sorbian," *The Journal of the Acoustical Society of America*, vol. 147, p. 2845, 2020.

[5] I. Kraljevski, M. P. Bissiri, F. Duckhorn, C. Tschoepe, and M. Wolff, "Glottal stops in Upper Sorbian: a data-driven approach," in *Proc. Interspeech 2021*, 2021, pp. 1001–1005.

[6] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[7] J. Fletcher, "The Prosody of Speech: Timing and Rhythm," in *The Handbook of Phonetic Sciences, Second Edition*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Chichester: Blackwell, 2010, pp. 521–602.

[8] S. Nakai, A. E. Turk, K. Suomi, S. Granlund, R. Ylitalo, and S. Kunnari, "Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish," *Journal of Phonetics*, vol. 40, no. 6, pp. 796–807, 2012.

[9] M. Belz, O. Rasskazova, J. Krivokapić, and C. Mooshammer, "Interaction between phrasal structure and vowel tenseness in German: An acoustic and articulatory study," *Language and Speech*, pp. 1–32, 2022.

[10] D. Byrd, J. Krivokapić, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1589–1599, 2006.

[11] J. P. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech & Language*, vol. 8, no. 2, pp. 95–128, 1994.

[12] B. Jakovljević and M. Marković, "Properties of I-boundary lengthening of vowels in English and Serbian," *Annual Review of the Faculty of Philosophy, Novi Sad*, vol. 45, no. 5, 2020.

[13] F. Seifart, L. Paschen, and M. Stave, Eds., *Language Documentation Reference Corpus (DoReCo) 1.1*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

[14] H. Bartels and M. Szczepański, "Lower sorbian doreco dataset," in *Language Documentation Reference Corpus (DoReCo) 1.1*, F. Seifart, L. Paschen, and M. Stave, Eds. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022. [Online]. Available: https://doreco.huma-num.fr/languages/lowe1385

[15] J. Měškank, M. Elikowska-Winklerowa, G. Wieczorek, K. Thorquindt-Stumpf, S. Krawcojc, L. Jocz, S. Elikowski-Winkler, J. Kaulfürstowa, M. Kóńcaŕ, F. Kaulfürst, and J.-F. Käthlitz, *Collection "Lower Sorbian"*. The Language Archive, 2011–2015. [Online]. Available: https://hdl.handle.net/1839/b7cb7b41-8362-4488-b8cb-86fbfea6c7c7

[16] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using the maus technique," in *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, 2011.

[17] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC, Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/

[18] L. White, S. Benavides-Varela, and K. Mády, "Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?" *Journal of Phonetics*, vol. 81, no. 100982, 2020.

[19] C. DiCanio and J. Sharp, "Initial weakening in Mixtecan languages," *The Journal of the Acoustical Society of America*, vol. 148, p. 2273, 2020.

[20] B. Andreeva, B. Möbius, and J. Whang, "Effects of surprisal and boundary strength on phrase-final lengthening," in *Proc. of Speech Prosody 10*, Tokyo, Japan, 2020, pp. 146–150.

[21] A. Wagner, "Description of polish speech rhythm using rhythm metrics and the time-delay approach: a comparative study," in *Proceedings of the 7th International Conference on Speech Prosody*, Dublin, Ireland, 2014, pp. 366–370.

[22] N. Spina and F. Schubö, "Vowel length affects pre-boundary lengthening in Czech," in *Proc. 1st International Conference on Tone and Intonation (TAI), Sonderborg, Denmark*, 2021.

[23] T. Kachkovskaia, N. B. Volskaya, and P. A. Skrelin, "Final lengthening in Russian: a corpus-based study," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*. Lyon, France: ISCA, 2013, pp. 1438–1442.

[24] T. Kachkovskaia, "Phrase-final lengthening in Russian: Pre-boundary or pre-pausal?" in *Speech and Computer*, A. Ronzhin, R. Potapova, and V. Delic, Eds., pp. 353–359.

[25] D. H. Klatt, "Synthesis by rule of segmental durations in English sentences," *Frontiers of Speech Communication*, pp. 290–297, 1979.

[26] N. Campbell, "Segmental elasticity and timing in Japanese speech," in *Speech perception, production, and linguistic structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds. Tokyo: Ohmsha, 1992, pp. 403–418.

COMPENSATORY LENGTHENING AND POLYSYLLABIC SHORTENING IN THE GHEG DIALECT OF ALBANIAN

Josiane Riverin-Coutlée[1], Enkeleida Kapia[1,2], Conceição Cunha[1] and Jonathan Harrington[1]

[1]Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität München, Munich, Germany
[2]Academy of Albanological Sciences, Tirana, Albania
josiane.riverin | enkeleida.kapia | cunha | jmh @phonetik.uni-muenchen.de

ABSTRACT

This study was concerned with understanding a case of diachronic vowel lengthening in the Gheg dialect of Albanian. Some of the long vowels of Gheg are thought to have emerged out of a historical process of compensatory lengthening following final schwa deletion. However, a confounding factor which may also explain why these vowels became longer is the effect of polysyllabic shortening being wiped out by schwa deletion. The main aim of this study was to disentangle the respective effects of compensatory lengthening and polysyllabic shortening on patterns of vowel duration in Gheg. An acoustic analysis of vowel duration in two word sets provided no evidence of an effect of polysyllabic shortening. Vowels were found to be longer, not shorter, in polysyllabic words, a fresh insight which suggests that vowel length may have instead been inherited from an open syllable and preserved when schwa was deleted.

Keywords: Gheg Albanian, vowels, syllable structure, compensatory lengthening, polysyllabic shortening

1. INTRODUCTION

This study probes into vowel length in the Gheg dialect of Albanian, with the aim of disentangling the respective effects of compensatory lengthening and polysyllabic shortening on patterns of vowel duration in a dialect for which empirical data is scarce.

Albanian is a language of the Indo-European family spoken by approximately 7 million people living mostly in Albania, Kosovo and North Macedonia [1]. In this study, we are concerned with Albanian spoken in Albania, which comprises two main dialects: Tosk, spoken in the southern part of the country; and Gheg, spoken in central and northern Albania, including in the capital Tirana [2], [3], [4]. We focus on a distinctive phonological feature of Gheg: contrastive vowel length.

The Gheg vowel inventory comprises 6 oral vowels which have short and long counterparts: /i y u e ɔ a/ and /iː yː uː eː ɔː aː/ [2], [5], [6], [7]. Long vowels occur frequently in indefinite singular nouns, whereas the corresponding definite singular nouns have short vowels, as illustrated in (1):

(1) /ˈkali/-/kaːl/: *kali* ‘the horse’ vs. *kalë* ‘a horse’

The presence of a long vowel in *kalë* /kaːl/ has been attributed to compensatory lengthening [6], [7], [8], [9], [10]. The leading hypothesis in the literature is that Gheg used to have an unstressed schwa at the end of the indefinite noun, /ˈkalə/, which was eventually dropped. The loss of this segment was compensated by an increase in duration of the remaining stressed vowel, with subsequent phonologization [11], [12], [13]. However, the historical passage from /ˈkalə/ to /kaːl/ also involved a reduction in the number of syllables in the word, which raises the question of whether so-called polysyllabic shortening could have played a role.

Polysyllabic shortening is a process through which the duration of vowels or syllables shortens as the number of syllables within the word increases. A classic example is that of *speed* > *speedy* > *speediness* [14] in which stressed /iː/ is shorter in disyllabic *speedy* than monosyllabic *speed*, and shorter in trisyllabic *speediness* than disyllabic *speedy*. Polysyllabic shortening has been observed mostly in stress-timed languages like English, Dutch, German or Swedish [15]. Since Albanian also is a stress-timed language [16], [17], the possibility has to be considered that the longer duration of the stressed vowel in monosyllabic *kalë* /kaːl/ compared to old disyllabic *kalë* /ˈkalə/ is due to the effect of polysyllabic shortening being wiped out when the final schwa (and thus one syllable) was deleted.

It is difficult to disentangle the effect of compensatory lengthening from that of polysyllabic shortening being lifted because they yield the same result: a longer vowel in the modern form. To overcome this issue, we propose to use the following proxies. First, we will verify whether polysyllabic shortening is a phenomenon that is found in Gheg using a corpus of definite-indefinite noun pairs as in (2). These have an analogical structure to the /ˈkali/-/kaːl/ pair in (1), but the indefinite noun never had a schwa, which was thus never dropped, not triggering compensatory lengthening:

(2) /ˈdaʃi/-/daʃ/: *dashi* ‘the ram’ vs. *dash* ‘a ram’

If polysyllabic shortening is found in Gheg, we expect the stressed vowel to be longer in monosyllabic /daʃ/ than in disyllabic /ˈdaʃi/.

Second, we will use a corpus of definite-indefinite pairs like disyllabic *kali* /ˈkali/ and monosyllabic *kalë* /ka:l/ complemented with a third word, in this case *fal* /fal/ 'to forgive'. This third word has a similar phonological structure to /ka:l/, but it is unrelated to the definite-indefinite pair and has not been subject to diachronic lengthening. If polysyllabic shortening plays any role in Gheg, we expect to replicate the results from the previous corpus, i.e. find the stressed vowel in /fal/ to be longer than that in /ˈkali/. In addition, if the wiping out of the effect of polysyllabic shortening alone can explain lengthening, then the vowel in monosyllabic /fal/ should be as long as that in /ka:l/.

2. METHODS

2.1. Speakers and recordings

Speech was collected from 14 native speakers of Gheg, 7 women and 7 men aged 38 to 79 years old (mean 58). The participants were born and raised in the village of Bërzhitë, near Tirana. While in this area some traditional features of Gheg tend to disappear due to dialect contact and influence of standard Albanian, contrastive vowel length has been found to be well preserved [18].

The participants were digitally recorded (44,1 kHz, 16 bits) in quiet rooms at the local school or at home, using a Beyerdynamic TG H54c head-mounted microphone, a Tascam US-2x2 and the Speech Recorder software [19]. They took part in a reading task featuring two repetitions of target words in carrier sentences, for example *kali thoni kali* 'the horse, say the horse' where the target word is *kali* /ˈkali/. Since the different prosodic positions in which the words were produced, i.e. initial and final positions of the carrier sentences, may have an impact on vowel duration, the variance induced by the prosodic position will be modeled in the statistical analyses by the inclusion of a fixed factor (see 2.3), but its effect will not be further analyzed because it is not our purpose to investigate this issue. Each sentence was read twice, for a total of four repetitions of each target word per speaker.

The speech signal was forced-aligned using WebMAUS General [20], [21], then hand corrected when needed. The material was structured into an EMU speech database [22] for further analysis.

2.2. Materials

Of the 241 target words produced four times by each speaker, 77 were used in this study while the rest was intended for other analyses. These 77 words were split into two corpora: the "analogy corpus" and the "triplet corpus". The analogy corpus comprises 26 words forming 13 definite-indefinite pairs on the model of /ˈdaʃi/-/daʃ/ in example (2). In each pair, the final unstressed vowel of the definite noun (/ˈdaʃi/) is absent in the indefinite noun (/daʃ/), leading to a reduction of one syllable in the indefinite noun (/daʃ/). As shown in Table 1, which provides examples of definite-indefinite pairs from the analogy corpus, the stressed syllable is always in penultimate position in definite nouns (/ˈdaʃi/), but in final position in indefinite nouns (/daʃ/). If polysyllabic shortening takes place in Gheg, the bolded vowels in the first row of Table 1 will be shorter than those in the second row.

definite	**da**shi 'the ram' /ˈ**da**ʃi/	**fi**ku 'the fig' /ˈ**fi**ku/	fish**e**ku 'the cartridge' /fiˈ**ʃe**ku/
indefinite	**da**sh 'a ram' /ˈ**da**ʃ/	**fi**k 'a fig' /ˈ**fi**k/	fish**e**k 'a cartridge' /fiˈ**ʃe**k/

Table 1: Examples of word pairs in the analogy corpus. Vowels in bold are the nuclei of the stressed syllable

The remaining 51 words make up the triplet corpus, that is, 17 triplets on the model of /ˈkali/-/ka:l/-/fal/. In each triplet, the definite noun (/ˈkali/) comprises a final unstressed vowel; the indefinite noun (/ka:l/) does not, therefore it has one less syllable, as well as a historically lengthened vowel; the third unrelated word (/fal/) has the same vowel quality and number of syllables as the indefinite noun (/ka:l/). As can be seen in Table 2, the stressed syllable is always in penultimate position in definite nouns (/ˈkali/), but in final position in definite nouns (/ka:l/) and unrelated words (/fal/). If polysyllabic shortening plays a role in Gheg, the bolded vowels in the third row of Table 2 will be longer than those in the first row. The bolded vowels in the second and third rows of Table 2 should also have similar durations if syllable count is the sole factor explaining why /ka:l/ got lengthened after schwa deletion.

definite	k**a**li 'the horse' /ˈk**a**li/	p**i**sha 'the pine' /ˈp**i**ʃa/	d**o**ra 'the hand' /ˈd**o**ɾa/
indefinite	k**a**lë 'a horse' /ˈk**a**:l/	p**i**shë 'a pine' /ˈp**i**:ʃ/	dorë 'a hand' /ˈdɔ:ɾ/
unrelated	f**a**l 'to forgive, INF' /ˈf**a**l/	p**i**sh 'to drink, 2SG SJV' /ˈp**i**ʃ/	m**o**r 'to take, 3SG SP PSS' /ˈm**o**ɾ/

Table 2: Examples of word triplets in the triplet corpus. Vowels in bold are the nuclei of the stressed syllable

2.3. Statistical analyses

The statistical analyses were carried out on the duration of the vowel in the stressed syllable in the 77 words forming the analogy and triplet corpora. Two separate mixed effect linear regression models were fitted [23], [24], one per corpus. For both analyses, the dependent variable was log-transformed, which helps reduce the skewness of residuals of non-

negative continuous values like durations.

A first model was fitted to the data from the analogy corpus using the R syntax in (3):

(3) lmer(log(duration) ~ definiteness + repetition + (definiteness|speaker) + (definiteness|wordpair)

where *definiteness* is the two-level fixed factor (definite, indefinite) of interest. *Repetition* is a two-level fixed factor modeling the variance induced by the different prosodic positions in which the words were produced, but not further analyzed (see 2.1). The interaction term *definiteness*repetition* was dropped because it was not significant. The model also includes random effects for *speaker* and *word pair* (e.g. *dashi* and *dash* correspond to the same word pair) and random slopes per *definiteness*.

A second model was fitted to the data from the triplet corpus using the same R syntax as in (3). However, the *definiteness* fixed factor had three levels (*definite*, *indefinite*, *unrelated*), and *word triplets* instead of *pairs* were set as random factor (e.g. *kali*, *kalë*, *fal* correspond to the same triplet). The interaction term *definiteness*repetition* was also dropped from this second model.

P-values were computed with *lmerTest* [25] and relevant pairwise comparisons with *emmeans* [26].

3. RESULTS

The top row of Figure 1 shows duration data for the stressed vowels in the 13 word pairs of the analogy corpus, while the bottom row shows data for the 17 word triplets of the triplet corpus. The words on the horizontal axis correspond to the definite nouns, i.e. those with one more syllable than their corresponding indefinite nouns and unrelated words.

For every word pair of the analogy corpus (top row), the duration of the stressed vowel is longer in the definite than indefinite nouns. That is to say, duration decreases as the number of syllables decreases (e.g. /ˈdaʃi/ > /daʃ/). This trend goes in the opposite direction from that expected from polysyllabic shortening, where the words represented by the black violins would have had the longest duration because they have one less syllable. The results of the statistical analysis indicate that the duration of the stressed vowels in the definite words is significantly longer than that in the indefinite words ($F[1, 20.36] = 80.12$, $p<0.001$).

The data from the triplet corpus shows that the vowels from the unrelated words tend to have the shortest duration, though this does not hold true for every triplet (e.g. *pisha-pishë-pish*). Overall, this replicates the trend observed in the analogy corpus: duration decreases as the number of syllables decreases, for instance from disyllabic definite /ˈkali/ to monosyllabic unrelated /fal/. This goes in the opposite direction from that expected from polysyllabic shortening, and certainly does not suggest that syllable count is the sole explanation for the increased length in indefinite nouns like /ka:l/, as the duration of their vowel does not resemble that of the unrelated words. For most triplets, vowel duration

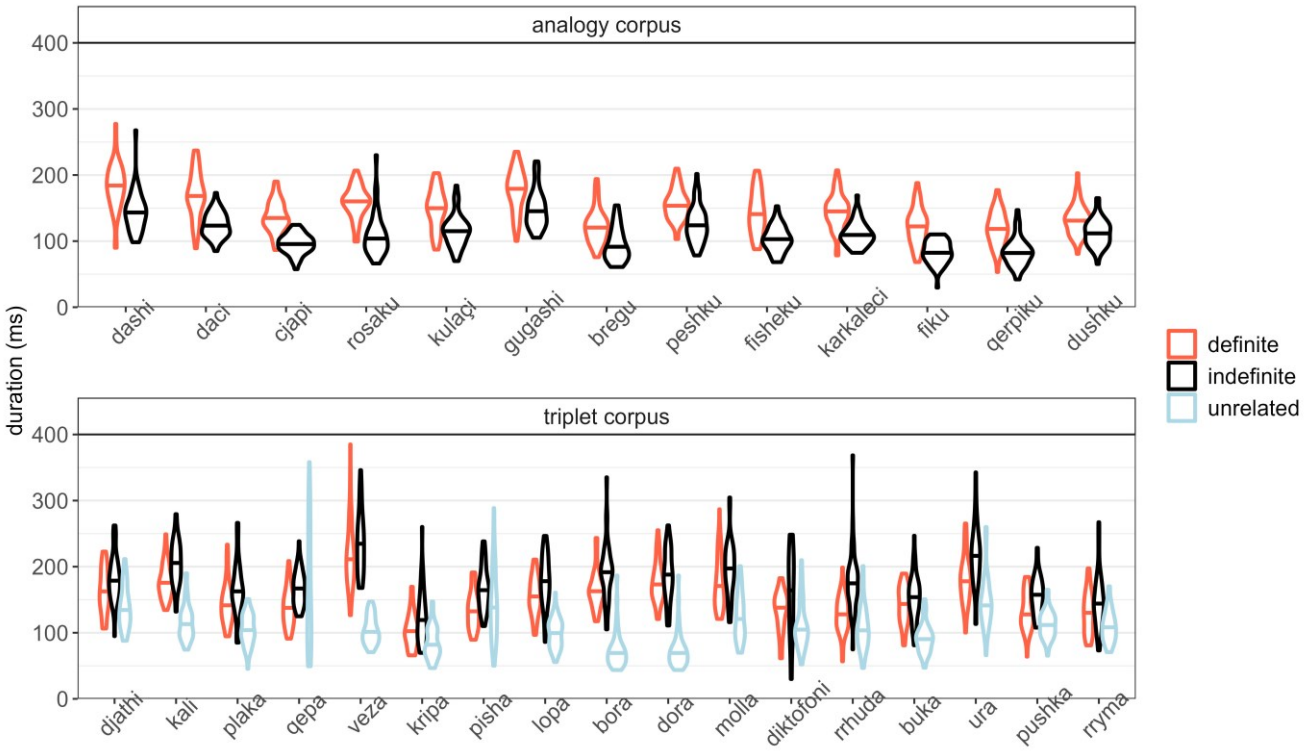


Figure 1: Violin plots of stressed vowel duration (ms) in word pairs from the analogy corpus (top) and word triplets from the triplet corpus (bottom). Definite nouns have one more syllable than the others, and are shown on the horizontal axes

is longest in the indefinite nouns, that is, those which have undergone diachronic lengthening. The results of the statistical analysis reveal a significant effect of the *definiteness* factor on vowel duration (F[2,16.78] =46.68, p<0.001). Post hoc pairwise comparisons indicate that vowel duration in every pair of words is different from each other. Indefinite nouns have a significantly longer vowel than both definite nouns (t[15.9] = 5.34, p<0.001) and unrelated words (t[20.2] = 7.49, p<0.001). Unrelated words also have a significantly shorter vowel than definite nouns (t[20.3] = 4.56, p<0.001).

4. DISCUSSION & CONCLUSION

In this study, we measured duration of stressed vowels in two word sets produced by Gheg speakers in order to better understand the respective effects of compensatory lengthening and polysyllabic shortening on length patterns in this dialect.

No evidence for an effect of polysyllabic shortening was found. Words with a larger number of syllables did not have shorter vowels than words with fewer syllables. This was found for both the definite-indefinite pairs from the analogy corpus and the definite-unrelated pairs from the triplet corpus. The main reason for supposing that polysyllabic shortening could be happening was that Albanian has a similar prosodic structure to that of languages where this phenomenon has been observed [16], [17]. However, there are other documented cases of languages with lexical stress which do not have polysyllabic shortening, e.g. Estonian and Finnish [15]. Therefore, perhaps more surprising than the absence of polysyllabic shortening is the consistent opposite effect found in this study. Words with fewer syllables, i.e. indefinite nouns in the analogy corpus and unrelated words in the triplet corpus, were found to have shorter stressed vowels than the definite nouns.

We relate this to a difference in syllabic structure: in definite nouns, the stressed syllable was open (*dashi* /ˈda.ʃi/, *kali* /ˈka.li/), while it was closed in indefinite nouns in the analogy corpus (*dash* /daʃ/) and unrelated words in the triplet corpus (*fal* /fal/). There is indeed a tendency across languages of the world for vowels in open syllables to have a longer duration than in closed syllables, all other things being equal [27]. The effect of syllable structure thus seems to outweigh that of syllable number in Gheg, at least under stress.

The finding that syllable structure seems to be the predominant factor does not allow us to rule out any effect of polysyllabic shortening in Gheg. An analysis of words with a closed penultimate stressed syllable would be necessary to shed more light on this issue. However, while principles of syllabification may vary across languages, it is usually agreed that Albanian favors open syllables unless phonotactic rules are violated [28], [29], [30]. Heterosyllabic intervocalic consonant clusters are thus rare, which may prove challenging for a targeted follow-up study of this factor.

The results also provided empirical confirmation, from a more controlled word set than ever before (c.f. [18]), of the longer duration of vowels thought to have undergone compensatory lengthening (/ka:l/). Given that the wiping out of the effect of polysyllabic shortening can be discarded as a potential explanation of the longer duration of these vowels, compensatory lengthening remains a plausible alternative, as hypothesized in earlier literature [6], [7], [8]. This being said, the marked difference in vowel duration depending on syllable structure found in this study tends to support the listener-oriented account of compensatory lengthening proposed by [13]. For sound changes of the CVCV > CV:C type, like old Gheg /ˈkalə/ > modern Gheg /ka:l/, the model in [13, p. 9] suggests the following: “Prior to the deletion of the final vowel, the longer vowel duration characteristic of open syllables is correctly parsed by listeners as a phonetic consequence of syllable structure in the first syllable of a CVCV sequence, and is discounted […]. Upon deletion of the final vowel, however, the duration of the vowel in the newly-closed syllable becomes inexplicable, since it is longer than is expected in the closed syllable”, setting the right conditions for this vowel to be reanalyzed as phonologically long (see also [31]). As argued in [11], this listener-based account departs from earlier models which mostly saw compensatory lengthening as a strategy of structure preservation, i.e. speaker-oriented, but it also puts into question the compensatory nature of the process. Length did not arise in order to compensate for the loss of schwa if it was already a property of the open syllable.

Even though our database still comprises only a small number of participants, which limits generalizability and statistical power, it includes rich material for detailed future work on vowel duration in Gheg, e.g. across positions within carrier sentences or vowel heights, as well as on syllable-level or word-level durations. In addition, of particular interest for future work on the topic is Tosk, the other main dialect of Albanian. Contrary to Gheg, Tosk has kept the final schwas in indefinite nouns and has not experienced diachronic lengthening [6], [8]. However, there have been reports some 50 years ago that Tosk speakers had started to occasionally drop schwas (e.g. [6]). Future work will thus examine length patterns in what can be seen as an earlier stage of the sound change that took place in Gheg.

5. ACKNOWLEDGMENTS

This research was funded by the project *InterAccent,* which has received funding from the European Research Council (grant agreement no. 742289). EK was also supported by a Humboldt fellowship.

6. REFERENCES

[1] A. Rusakov, "Albanian," in *The Indo-European Languages*, 2nd ed., M. Kapović, Ed. London: Routledge, 2017, pp. 552–608.

[2] B. Beci, *Të folmet veriperëndimore të shqipes dhe sistemi fonetik i të folmes së Shkodrës*. Tiranë: Mihal Duri, 1995.

[3] J. Gjinari, *Dialektologjia shqiptare*, Revised and Expanded. Tiranë: Akademia e Shkencave e Shqipërisë, 1988.

[4] J. Gjinari, B. Beci, G. Shkurtaj, X. Gosturani, and A. Dodi, *Atlasi dialektologjik i gjuhës shqipe*. Napoli: Università degli Studi di Napoli, 2007.

[5] B. Beci, "Vlera funksionale e gjatësisë së zanoreve të theksuara në shqipen e Veriut," *Studime Filologjike*, vol. 16, no. 3, pp. 89–134, 1979.

[6] M. Çeliku, "Kuantiteti i zanoreve të theksuara në të folmet e shqipes," *Studime Filologjike*, vol. 8, no. 4, pp. 65–100, 1971.

[7] E. Çabej, "Kuantiteti i zanoreve të theksuara të shqipes," *Buletin i Universitetit Shtetëror të Tiranës. Seria Shkencat Shoqërore*, vol. 2, 1957.

[8] M. Çeliku, *Gegërishtja jugperëndimore*. Tiranë: Akademia e Studimeve Albanologjike, 2020.

[9] G. Pekmezi, "Die Charakteristik der Unterschiedsmerkmale des Gegischen und Toskischen." Arkivi i Institutit të Gjuhësisë dhe të Letërsisë, 1901.

[10] G. Weigand, *Albanesische Grammatik in südgegischen Dialekt*. Leipzig: Barth, 1913.

[11] R. Gess, "Compensatory lengthening," in *The Blackwell Companion to Phonology*, vol. III, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds. Wiley Blackwell, 2011, pp. 1536–1557.

[12] L. M. Hyman, "Enlarging the scope of phonologization," in *Origins of Sound Change: Approaches to Phonologization*, A. C. L. Yu, Ed. Oxford: Oxford University Press, 2013, pp. 3–28.

[13] D. Kavitskaya, *Compensatory Lengthening: Phonetics, Phonology, Diachrony*. New York: Routledge, 2002.

[14] I. Lehiste, "The timing of utterances and linguistic boundaries," *Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2018–2024, 1972, doi: 10.1121/1.1913062.

[15] J. Fletcher, "The prosody of speech: Timing and rhythm," in *The Handbook of Phonetic Sciences*, 2nd ed., W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2010, pp. 523–602.

[16] G. Belluscio, A. Mendicino, and L. Romito, "L'albanese standard: vocalismo e strategie temporali," *Quaderni del Dipartimento di linguistica dell'Università degli studi di Firenze*, vol. 8, pp. 105–116, 1997.

[17] A. Jubani-Bengu and M. Çabej, "Izokronia, klasifikimi i shqipes mbështetur në veçoritë të organizimit kohor në shqipe," *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*, vol. 35, pp. 77–86, 2017.

[18] J. Riverin-Coutlée, E. Kapia, C. Cunha, and J. Harrington, "Vowels in urban and rural Albanian: The case of the Southern Gheg dialect," *Phonetica*, vol. 79, no. 5, pp. 459–512, 2022, doi: 10.1515/phon-2022-2025.

[19] C. Draxler and K. Jänsch, "SpeechRecorder – A universal platform independent multi-channel audio recording software," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. 559–562.

[20] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017, doi: 10.1016/j.csl.2017.01.005.

[21] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proceedings of ICPhS 14*, San Francisco, USA, 1999, pp. 607–610.

[22] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017, doi: 10.1016/j.csl.2017.01.002.

[23] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[24] R Core Team, "R: A language and environment for statistical computing." 2022. https://www.R-project.org

[25] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.

[26] R. Lenth, "emmeans: Estimated marginal means, aka least-squares means." 2022. https://CRAN.R-project.org/package=emmeans

[27] I. Maddieson, "Phonetic cues to syllabification," *UCLA Working Papers in Phonetics*, vol. 59, pp. 85–101, 1985.

[28] A. Dodi, *Fonetika dhe Fonologjia e gjuhës shqipe*. Tiranë: Akademia e Shkencave e Shqipërisë, 2004.

[29] R. Memushaj, *Fonetika E Shqipes Standarde*. Tiranë: Botimet Toena, 2009.

[30] A. Xhaferaj, "Syllabification of bi-consonantal clusters between vowels in Albanian," *International Journal of English Linguistics*, vol. 8, no. 5, pp. 230–237, 2018, doi: 10.5539/ijel.v8n5p230.

[31] J. J. Ohala, "The listener as a source of sound change," in *Papers from the Parasession on Language and Behavior*, C. S. Masek, R. A. Hendrick, and M. F. Miller, Eds. Chicago: Chicago Linguistic Society, 1981, pp. 178–203.

THE /t/ RELEASE IN JUTLAND DANISH: DECOMPOSING THE SPECTRUM WITH FUNCTIONAL PCA

Rasmus Puggaard-Rode

Institute for Phonetics and Speech Processing (IPS), LMU Munich
r.puggaard@phonetik.uni-muenchen.de

ABSTRACT

This study concerns variation in the realization of /t/ in Danish. /t/ is prominently affricated in Modern Standard Danish, but affrication is known to have been missing in some of the regional varieties that used to be spoken in Jutland. It remains unclear exactly which varieties lacked affrication. Due to relatively recent major changes in language policy and mobility, the traditional Danish dialect landscape has undergone rapid levelling, so we explore this problem using a large corpus of recordings of elderly speakers from the 1970s. We propose using the functional extension of principal component analysis to decompose the sources of spectral variance in stop release midpoints, and use the results as a proxy for determining affrication status. We find that /t/ affrication was limited to parts of eastern Jutland, perhaps due to early adaptation of the feature in some of the area's largest cities.

Keywords: Danish, functional PCA, regional variation, spectral decomposition, stop affrication

1. INTRODUCTION

The aims of this study are twofold: 1) to explore the geographical distribution of (missing) /t/ affrication in traditional varieties of Jutland Danish (see Fig. 1), and 2) to explore the use of functional principal component analysis (FPCA) [1] for compressing the information in speech spectra into a small number of easily interpretable discrete variables.

In Modern Standard Danish (MSD), /p t k/ are voiceless aspirated, and /t/ in particular is prominently affricated [2, 3, 4]. Until relatively recently, the majority of the Danish speech community spoke traditional dialects [5], but in the past century, rampant standardization has largely led to the levelling of these varieties [6]. An overt feature of some Jutlandic varieties was the use of a variant of /t/ without salient affrication, which is known colloquially as *tørt t* 'dry t' [7]. There is no consensus about which Jutlandic varieties lacked affrication. Sources have variously claimed that it was missing from all

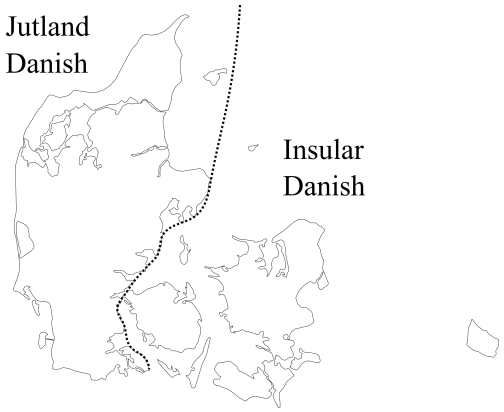


Figure 1: Map showing the areas associated with the two primary Danish dialect groups.

Jutlandic varieties [8], Northern Jutlandic [9], Western Jutlandic [10], or all varieties except Eastern Jutlandic [11]. We explore this mystery using a large legacy corpus of sociolinguistic interviews which was explicitly recorded to document the traditional language and culture of rural Denmark [12, 13]. These recordings (particularly the ones from Jutland) largely preserve a stage of language variation unaffected by the recent wave of standardization.

The acoustics of alveolar frication and aspiration are well-understood. In alveolar frication, a jet of air impinges on a hard surface (the upper front teeth) immediately in front of the coronal constriction, resulting in noise in a broad range of high frequencies (mainly above 4 kHz) [14]. In aspiration, lower frequency turbulence noise in a narrower frequency range (mainly below 1 kHz) is generated at and near the glottis [15]. This noise is filtered through the supraglottal cavity, where the resonance frequencies of the following vowel significantly affect the resulting noise. Analyzing how these spectral characteristics vary regionally is not at all straightforward.

Spectral moments are commonly used to decompose spectra generated from aperiodic portions of speech. The spectral mean, or *center of gravity* (COG) [16], is a particularly popular measure. Since spectra rarely resemble Gaussian distributions, the mean frequency alone is not very informative with respect to overall shape; furthermore, studies have

found conflicting evidence about which spectral moments (if any) can be used to consistently determine noise sources [17, 18]. Several other measures have been proposed for decomposing aperiodic spectra [19, 20, 21], but none have been as influential as COG.

For the reasons outlined above, we do not use spectral moments here. Instead, we adopt FPCA as an alternative method. FPCA is used to determine the main sources of variance in curves without imposing predetermined shapes. This makes it particularly suitable for exploratory work. Each resulting principal component (PC_i) corresponds to a source of variance relative to the average spectral shape, and each individual spectrum gets a score s_i for each PC indicating how well it corresponds to this source of variance. These scores can then be fitted to conventional statistical models; here, we fit PC scores to spatial generalized additive mixed models (GAMMs) [22] in order to test whether any of the variance in spectral shape is geographically determined. These models also include a range of phonetic contextual predictors. The influence of phonetic contextual predictors largely pattern as predicted, which we take as an indication that FPCA is indeed a suitable tool for decomposing spectra. FPCA has previously been used to analyze various time-varying phonetic measures, such as F_0 and formants [1, 23]; to our knowledge, it has not previously been used to analyze speech spectra.

2. METHODS AND MATERIALS

2.1. Recordings

Our materials consist of sociolinguistic interviews with elderly speakers of traditional regional varieties recorded between 1971–1976 [13, 24]. Interviews were conducted with informants in 230 parishes in Jutland, of which 213 were used in this study. 23% of the informants are women. Informants' median year of birth is 1896 (range 1871–1927). The geographical distribution of informants can be seen in Fig. 2 (which shows the location of Djursland for reference.) In spite of the large scope of these recordings, they have not been used for systematic research until very recently [7, 25].

2.2. Acoustic processing

5,169 /t/ releases were manually segmented. The release burst and the onset of periodicity were used as delimitation landmarks, as identified from the waveform. Using Praat [26], 5 ms snippets were extracted from the midpoint of each /t/ release, filtered to include only frequencies between 0.5–12.5 kHz. For each of these, multitaper spectra were generated in R [27] following the method described in [28]. Intensity values were log-transformed and standardized in order to keep all spectra on the same scale. Measurements at frequencies > 8 kHz were not included in the analysis, as they did not contribute much to the analysis other than noise.

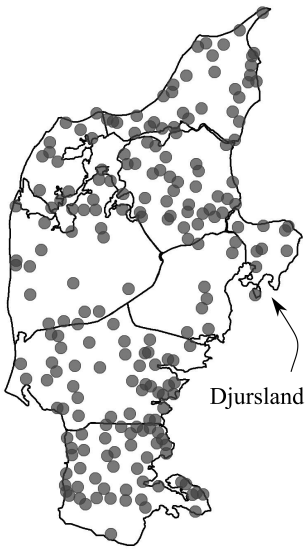


Figure 2: The geographical distribution of recordings used in this study.

2.3. Statistical analysis

All statistics were calculated in R [27]. The package `fdapace` [29] was used for FPCA. Spectra were smoothed using a local linear smoother with parameters set automatically using generalized cross-validation. Five PCs are needed to account for 95% of spectral variance. We focus on the first three $PC_{1–3}$, which account for 85.9% of the variance.

PC scores $s_{1–3}$ served as dependent variables in separate GAMMs with the following maximal structure: s_i ~ LON,LAT + HEIGHT + STRESS + SEX + BACK + ROUND + PALATALIZATION + VOT. (HEIGHT, ROUND, BACK, and STRESS refer to features of the following vowel; SEX refers to the sex of the speaker.) Most of these features have been shown to affect spectral shape in MSD stop releases [4]. There were various phonological palatalization rules in different Jutland Danish varieties [30]; we make no distinction between different sources of palatalization. All categorical variables are contrast coded [31]; the HEIGHT variable tests two Helmert contrasts, viz. 'high *vs.* non-high' and 'low *vs.* mid'. All other categorical variables are binary and coded with sum contrasts. The VOT variable is standard-

ized. By-speaker random slopes were included for each variable except SEX. The geographical variable LON,LAT is modelled with two-dimensional thin plate regression spline smooths [32]. GAMMs were fitted using the package `mgcv` [33]. The packages `itsadug` [34] and `mgcViz` [35] were used for health checks and visualizations of the models.

The source material is spontaneous speech, so the data are rather imbalanced: 41.9% of /t/ tokens were followed by non-high vowels, of which 15.4% were low. 33.7% of tokens were found in stressed syllables. 11.5% of tokens were followed by back vowels, and 22.7% were followed by rounded vowels. Only 2.5% of the tokens were palatalized.

We used a step-up model selection procedure, adding variables in the order given above (based on these variables' influence on MSD /t/ release spectra [4]). Variables were only kept if the cost of fitting a more complex model was sufficiently outweighed by an improved model fit, as determined with likelihood ratio tests [36]. Data and annotated analysis code are available online [37]. More details of the study are available in my dissertation [38].

3. RESULTS

PC_1 accounts for 58.4% of spectral variance, PC_2 accounts for 18.2%, and PC_3 accounts for 9.3%. The variance explained by $PC_{1–3}$ is visualized in Fig. 3. The plots all show the mean spectrum and the shape of the mean spectrum when weighted by the first and third quantile of all PC scores. This gives an indication of what spectra with a relatively high score s_i and a relatively low s_i look like.

The average spectrum has little energy below 1.5 kHz, relatively high amplitude at frequencies between 1.5–4.5 kHz, gradual energy loss above that, and peaks just below 2.5 kHz. The main source of variance, captured by PC_1, is the location and magnitude of the primary peak. Positive s_1 corresponds to a lower frequency peak, and generally more energy at lower frequencies. COG would likely have captured some of the same information as PC_1, but they are not equivalent: in addition to the location of the peak, PC_1 also captures information about variance in spectral slope, peak amplitude, and skew. Given what we know about aspiration and alveolar frication, strong negative s_1 seems to represent an alveolar noise source, and strong positive s_1 seems to represent a glottal noise source. PC_2 captures variance in the magnitude of the main peak. Positive s_2 corresponds to a more prominent peak, and less energy at higher frequencies, while negative s_2 corresponds to a less prominent peak, and more energy at higher frequencies. PC_3 mostly captures information about the peakedness of the energy distribution. Positive s_3 corresponds to a more restricted and more prominent peak in the same location as in the mean spectrum, and negative s_3 corresponds to a broader distribution of energy.

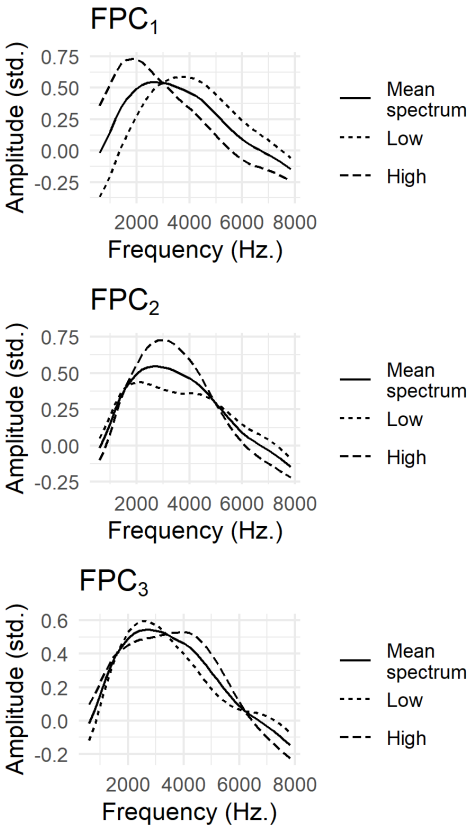


Figure 3: Variance in release spectra captured by $PC_{1–3}$ and the effect of weighting the mean by high and low scores $s_{1–3}$.

The GAMM modelling s_1 was fitted with all candidate independent variables except STRESS. This model has a medium strong effect size of $R^2 = .39$. A likelihood ratio test found that a model including the geographical variable LON,LAT performed significantly better than a nested model without this variable, with $\chi^2(3) = 4.1$, $p = 0.043$. The influence of the geographical variable is plotted on a map in Fig. 4; Fig. 4 shows that negative s_1 is common in Djursland and the surrounding area in mid-eastern Jutland (see Fig. 2); we return to this below. All other variables also significantly influence s_1 (see Table 1). The upshot is that there is less energy at high frequencies before non-high, particularly low, vowels; before rounded vowels; before back vowels; in non-palatalized releases; and in tokens from male speakers. Finally, there is an inverse relationship between VOT and s_1, such that shorter VOT

corresponds to higher s_1. These are all expected patterns, similar to what is found in MSD [4].

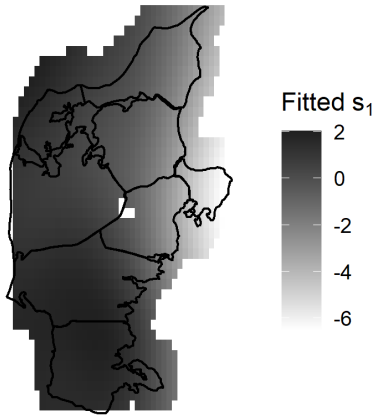


Figure 4: Fitted s_1 values attributable to area.

	est.	*SE*	*t*	*p*
intercept	-1.42	1.38	1.03	0.3
height				
- non-high, + high	-3.65	0.74	-4.96	<.001
- low, + mid	-4.38	0.91	-4.84	<.001
round				
- rd, + non-rd	3.47	0.97	3.58	<.001
back				
- back, + non-back	-3.91	1.21	-3.22	<.01
palatalization				
- non-pal., + pal.	-6.79	2.03	-3.35	<.001
sex				
- female, + male	6.26	1.87	3.35	<.001
VOT	-2.03	0.69	-2.95	<.01

Table 1: Parametric coefficients of GAMM modeling s_1.

Fewer variables contribute to the GAMM modeling s_2: HEIGHT, ROUND, STRESS, and BACKNESS. The geographical variable LON,LAT does not significantly improve the fit of this model. The final model has a medium effect size, with R^2 = .29. Only two variables significantly influence s_2, namely BACK with $\hat{\beta}$ = 1.97, *SE* = 0.7, *t* = 2.82, $p < .01$, and ROUND with $\hat{\beta}$ = 2.21, *SE* = 0.61, *t* = 3.6, $p < .001$. In other words, positive s_2, associated with an especially prominent energy peak around 2.5 kHz, is especially found before non-back and non-round vowels. This is also in line with previous findings for MSD [4].

All parametric candidate variables contribute to the GAMM modeling s_3; the geographical variable does not significantly improve the fit of this model. The resulting model has a medium small effect size of R^2 = .25. The only variable found to significantly influence s_3 is ROUND, with $\hat{\beta}$ = 2.62, SE = 0.45, *t* = -5.84, $p < .001$.

4. DISCUSSION AND CONCLUSIONS

Spectral shapes that suggest a coronal noise source are found in /t/ release midpoints in the varieties of eastern Jutland, particularly in Djursland, but not in other traditional varieties of Jutland Danish. This suggests that /t/ affrication was not historically a feature of most varieties of Jutland Danish.

Indeed, it is possible that /t/ affrication was not a feature of any traditional Jutlandic varieties. Southern Djursland coincides with the main water route between Jutland and the island of Zealand, which is where the capital (Copenhagen) and consequently the locus of Standard Danish is located. This is also where the largest city of Jutland (Aarhus) is located, and several of the major historical cities of Jutland are located immediately south of Aarhus [7]. Affrication may be a traditional feature of this area, but the feature may also have spread early from Standard Danish to Aarhus and from there to adjacent rural areas. This would be in line with the *cascade model* of interdialectal influence [39], which predicts that change spreads between population centers in a manner that is predictable from a combination of population size and geographical distance, thus also affecting rural areas close to major cities relatively early.

PC_1 and PC_2 capture the two most readily interpretable sources of variance in the release midpoint spectra: the location and relative magnitude of the energy peak. PC_1 seems to largely capture the difference between aspiration and affrication noise, which is modulated by both geography and phonetic context; lower PCs seem to capture effects of coarticulation, and these do not differ substantially across regional varieties. The influence of phonetic context on $PC_{1–2}$ is largely predictable, and is similar to what has been shown for MSD [4]; for example, the main peak is usually found at higher frequencies before high, non-back, and non-round vowels, in palatalized tokens, in stressed syllables, and in female speakers. The fact that these contextual phonetic effects all behave as expected are an indication that FPCA can indeed be fruitfully used to decompose variance in spectral characteristics. A possible venue to explore in future research would be using FPCA to decompose variance in the trajectories of spectral measures over the time course of stop bursts and releases.

5. ACKNOWLEDGMENTS

Many thanks to Janet Grijzenhout, Bert Botma, Johanna Cronenberg, Jiayin Gao, James Kirby, Michele Gubian, Kathleen Jepson, Vanessa Reichel, Josiane Riverin-Coutlée, and two anonymous reviewers for helpful comments on the presentation of this paper.

6. REFERENCES

[1] Gubian, M., Torreira, F., Boves, L. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *J. Phon.* 49, 16–40.

[2] Jespersen, O. 1897–1899. *Fonetik. En systematisk fremstilling af læren om sproglyd*. Schuboteske.

[3] Fischer-Jørgensen, E. 1954. Acoustic analysis of stop consonants. *Le Maître Phonetique* 32(69), 42–59.

[4] Puggaard-Rode, R. 2022. Analyzing time-varying spectral characteristics of speech with function-on-scalar regression. *J. Phon.* 95, art. 101191.

[5] Kristiansen, T. 2003. Danish. In: Deumart, A., Vandenbussche, W. (eds), *Germanic standardizations. Past to present*. John Benjamins, 69–91.

[6] Pedersen, I. L. 2003. Traditional dialects of Danish and the de-dialectalization 1900–2000. *Int. J. Soc. Lang.* 159, 9–28.

[7] Puggaard, R. 2021. Modeling regional variation in voice onset time of Jutlandic varieties of Danish. In: Van de Velde, H., Hilton, N. H., Knooihuizen, R. (eds), *Language variation – European perspectives VIII*. John Benjamins, 79–110.

[8] Brink, L., Lund, J. 1975. *Dansk rigsmål. Lydudviklingen siden 1840 med særligt henblik på sociolekterne i København*. Gyldendal.

[9] Petersen, N. R. 2009. Aspirater. In: *Den store danske*. Gyldendal.

[10] Petersen, J. H., Juul, H., Pharao, N., Maegaard, M. 2021. *Udtalt. En introduktionsbog til dansk fonetik*. Samfundslitteratur.

[11] Heger, S. 1981. *Sprog og lyd. Elementær dansk fonetik*. Akademisk Forlag.

[12] Andersen, T. A. 1981. Dialektbånd og databehandling. *Ord & Sag* 1, 11–18.

[13] Goldshtein, Y., Puggaard, R. 2019. Overblik over danske dialektoptagelser. *Ord & Sag* 39, 18–28.

[14] Stevens, K. N. 1993. Models for the production and acoustics of stop consonants. *Speech Comm.* 13(3), 367–375.

[15] Stevens, K. N. 1998. *Acoustic phonetics*. MIT Press.

[16] Forrest, K., Weismer, G., Milenkovic, P., Dougall, R. N. 1988. Statistical analysis of word-initial voiceless obstruents. Preliminary data. *J. Acoust. Soc. Am.* 84(1), 115–123.

[17] Stoel-Gammon, C., Williams, K. A., Buder, E. 1994. Cross-language differences in phonological acquisition. Swedish and American /t/. *Phonetica* 51(1), 146–158.

[18] Shadle, C. H., Mair, S. J. 1996. Quantifying spectral characteristics of fricatives. *Int. Conf. Spoken Lang. Proc.* 4, 1521–1524.

[19] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.* 108(3), 1252–1263.

[20] Bunnell, H. T., Polikoff, J., McNicholas, J. 2004. Spectral moment vs. bark cepstral analysis of children's word-initial voiceless stops. *Int. Conf. Spoken Lang. Proc.* 8.

[21] Koenig, L. L., Shadle, C. H., Preston, J. L., Mooshammer, C. R. 2013. Toward improved spectral measures of /s/. Results from adolescents. *JSLHR* 56(4), 1175–1189.

[22] Wieling, M., Nerbonne, J., Baayen, R. H. 2011. Quantitative social dialectology. Explaining linguistic variation geographically and socially. *Plos One* 6(9).

[23] Gubian, M., Boves, L., Cangemi, F. 2011. Joint analysis of F_0 and speech rate with functional data analysis. *IEEE ICASSP*, 4972–4975.

[24] *Dialektsamlingen*, 1971–1976. URL: https://dansklyd.statsbiblioteket.dk/samling/dialektsamlingen/.

[25] Goldshtein, Y., Ahlgren, L. M. 2021. Ideologies of language and place. Conflicting expectations to dialectal speech between informants and dialectologists. *J Postcolonial Ling* 5, 178–203.

[26] Boersma, P., Weenink, D. 2019. Praat. Doing phonetics by computer, v6.0.55.

[27] R Core Team, 2021. R. A language and environment for statistical computing, v4.1.2.

[28] Reidy, P. F. 2013. An introduction to random processes for the spectral analysis of speech data. *OSUWPL* 60, 67–116.

[29] Gajardo, A., Satarupa, B., Carroll, C., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Zhu, C., Müller, H.-G., Wang, J.-L. 2021. `fdapace`, R package v0.5.8.

[30] Bennike, V., Kristensen, M. 1912. *Kort over de danske folkemål med forklaringer*. Gyldendal.

[31] Schad, D. J., Vasishth, S., Hohenstein, S., Kliegl, R. 2020. How to capitalize on a priori contexts in linear (mixed) models. A tutorial. *J. Mem. Lang.* 110, art. 104038.

[32] Wood, S. N. 2003. Thin plate regression splines. *J. Roy. Stat. Soc. B* 65(1), 95–114.

[33] Wood, S. N. 2021. `mgcv`. R package v1.8-38.

[34] van Rij, J., Wieling, M., Baayen, R. H., van Rijn, H. 2020. `itsadug`. R package v2.4.

[35] Fasiolo, M., Nedellec, R., Goude, Y., Wood, S. N. 2021. `mgcViz`. R package v0.1.9.

[36] van Rij, J. 2016. Testing for significance. `itsadug` R package vignette.

[37] DataverseNL, DOI: 10.34894/FJKFD1.

[38] Puggaard-Rode, R. 2023. *Stop! Hey, what's that sound? The realization and representation of Danish stops*. PhD dissertation, Leiden University.

[39] Labov, W. 2003. Pursuing the cascade model. In: Britain, D., Cheshire, J. (eds), *Social dialectology*. John Benjamins, 9–22.

An MRI Examination of Lingual Fricatives in Upper Sorbian

Phil J. Howson[1] & Peter Birkholz[2]

[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft; [2]Technische Universität Dresden
howson@leibniz-zas.de; peter.birkholz@tu-dresden.de

ABSTRACT

Fricatives require precise gestural control to produce a narrow constriction in the vocal tract. This generates turbulent airflow that gives fricatives their distinctive acoustic properties. This study investigated the articulation of the lingual fricatives /s, ʃ, x/ of Upper Sorbian, which is an endangered language spoken in eastern Germany. We used MRI to examine the mid-sagittal plane for the fricatives. We also used the MRI tracings to simulate the fricatives' COG in VocalTractLab. 2 male and 2 female L1 Upper Sorbian speakers participated in the study. We performed 3D vocal tract scans using 1.2 x 1.2 x 1.8 mm^3 voxels with 44 sagittal slices.

The results revealed significant inter-speaker variation in constriction location and length, and tongue shape. The results thus implicate variation in gestural representations that deviate due to vocal tract size and shape to meet a specific acoustic target.

Keywords: Fricatives; Upper Sorbian; MRI; phonetics; endangered languages

1. INTRODUCTION

Fricative sounds involve articulations that produce turbulent airflow within the vocal tract. In sibilants, such as /s/, the airflow leaving the critical constriction is directed against the incisors for the generation of a specific noise pattern. Fricatives in general demand a great deal of precision to accurately produce the required spectral qualities. Even one millimetre difference in constriction location can impact the acoustics significantly [1, 5]. This makes them hard to produce and even harder to acquire.

The lingual fricatives of Upper Sorbian have not been studied in significant detail with MRI or acoustic measures, but the language shares a similar inventory as German, /s, ʃ, x/, and the impact from language contact is also likely to have caused similarities between the languages [7]. With respect to German fricatives, Jannedy & Weirich [9] found that for Hamburg and Jena dialects of German, there were high frequency spectral peaks for /s/, typically over 6,000 Hz, while /ʃ/ typically had a peak around 3,000 Hz, and /x/ had a peak often below 2,000 Hz. The overall distribution of noise differed significantly by segment and dialect, however, often spanning a range of 2,000 or 3,000 Hz. Gordon, Barthmaier, & Sands [6] examined seven languages that each had rich fricative inventories. Across the languages examined, /s/ had COG ranges from 4,400 Hz – 5,600 Hz, /ʃ/ had COG ranges from 3,900 Hz – 5,100 Hz, and /x/ had COG ranges from 3,900 Hz – 4,500 Hz. The wide and overlapping ranges of COG for the relevant fricatives suggests that there are a variety of distinct vocal tract shapes that distinguish each of the fricatives.

The purpose of this study was to examine the vocal tract shapes of lingual fricatives in Upper Sorbian as part of a documentation effort. Additionally, we compared and simulated COG to better understand how vocal tract geometry affects the acoustics for fricatives.

2. METHODS

2.1. Participants

Four L1 speakers of Upper Sorbian participated in this study (2 male and 2 female). Participants were all college-educated, ages between 20-24 years, and had no self-reported history of speech or hearing disorders.

2.2 Stimuli & Procedure

Participants were given a list of words with target segments prior to data collection and were instructed to practice producing them for an extended duration. The stimuli are presented in Table 1.

Segment	Stimuli	English
/s/	**s**adło	fat; grease
/ʃ/	**š**ach	chess
/x/	či**ch**awa	sneezing

Table 1. Stimuli with target segments in bold.

3D images of the vocal tract were collected with a Siemens 3T Trio, with a pixel size of 1.2 mm x 1.2 mm, and a sagittal slice thickness of 1.8 mm. 44 slices were taken in total and they were used to generate a 3D image of the vocal tract. In order to record 44 slices, participants had to produce a sustained articulation of a single segment for 14 seconds. To do this, participants were prompted with a power point presentation. A mirror was present in the MRI to reflect the power point screen. First, the target segment plus the example word were presented. Participants were asked to produce the word and the target segment through the intercom

system. Following confirmation that the correct segment was being articulated, participants were prompted with the target segment and example word and a bar on the screen that would indicate when to take a deep breath, when to begin articulation, and when to stop. After each articulation, MRI images were examined and in the case of blurriness or poor image quality, the participant repeated the trial until a clear, high-quality image was obtained.

Audio recordings were additionally performed in a quiet room in Leipzig, Germany. Data were collected using a Tascam Linear PCM recorder. Data were recorded at a sampling frequency of 48,000 Hz. Participants were sitting for audio recording, rather than laying as they had in the MRI. Participants read the target words used in the original MRI stimuli and then produced 5 second static production of each of the target segments /s, ʃ, x/. Participants were instructed to produce the segments as similarly to the MRI productions as they could.

2.2 Analysis

Dicom files were converted into 44 bitmap files for each segment produced by each of the participants. Bitmaps were loaded into Image3D [2]. Image3D uses Catmull-Rom Splines so users can trace the tongue, lips, palate, and pharyngeal wall. We generated 2 independent splines, one for the lower lip and tongue and one for the upper lip, palate and pharyngeal wall. Splines were then exported to a scalable vector graphics (SVG) file and imported to Inkscape [8]. Midsagittal contours were then overlayed on top of each other to facilitate comparisons.

Acoustic analysis was performed in Praat [4]. We took a 30-millisecond window centred on the midpoint of the static articulation of each fricative. A spectral slice was extracted and then the Center of Gravity (COG) was calculated. Then VocalTractLab (Special Version for 3D acoustic simulations) [3] was used to simulate tongue shapes and acoustic outputs. The 3D acoustic simulation suite was used. VocalTractLab [3] provides control points to manipulate the 2D and 3D tongue contours. It can then be used to compute transfer functions, modes, and the acoustic field in order to simulate the acoustics of the vocal tract shape. We manipulated the control points and tongue side elevation to match the MRI tracings.

3. RESULTS

3.1 MRI Results

The midsagittal traces for speaker US01 are presented in Figure 1. The midsagittal profile indicated a constriction along the alveolar ridge for both /s/ and /ʃ/, which was around the postalveolar region. However, there was a longer constriction for /s/, which extended further forward towards the lower incisors. Additionally, there was more tongue tip advancement compared to /ʃ/. The tongue body was more raised and had a downward slope for /ʃ/, indicating a slight degree of palatalization. /ʃ/ also had a sublingual cavity. The posterior tongue was also more advanced for /ʃ/, compared to /s/. /x/ had a constriction in the midpalate, accompanied by a retracted tongue tip and tongue dorsum.

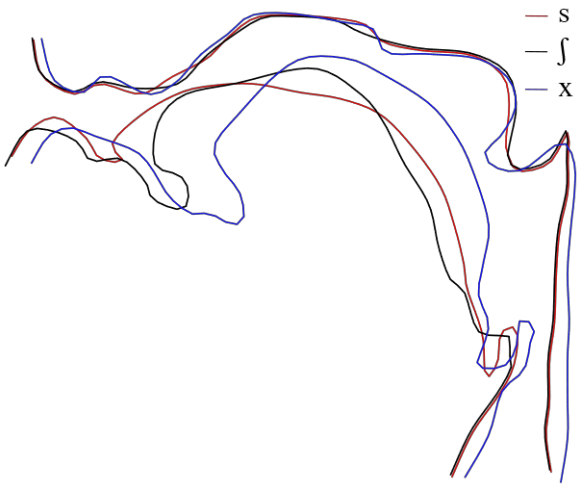


Figure 1: MRI traces of the midsagittal plane for lingual fricatives, /s/ (red), /ʃ/ (black), /x/ (blue) for US01.

The midsagittal traces for speaker US02 are presented in Figure 2.

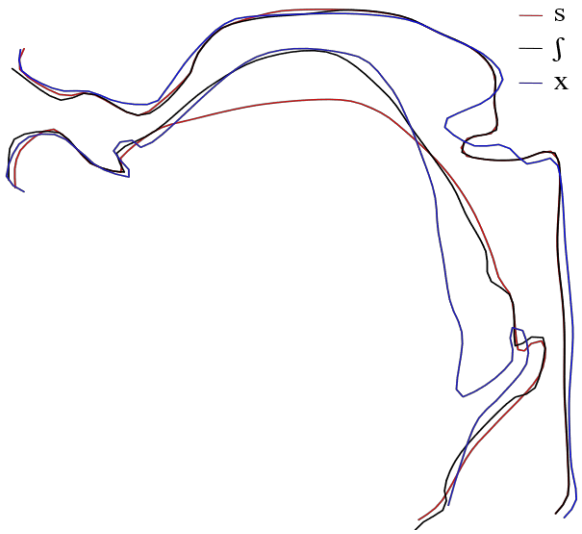


Figure 2: MRI traces of the midsagittal plane for lingual fricatives, /s/ (red), /ʃ/ (black), /x/ (blue) for US02

The constriction location for /s/ and /ʃ/ were similar. However, in the case of US02, the constriction for /ʃ/ was significantly longer than for /s/, spanning the entire post-alveolar ridge. /s/, on the other hand, had a short constriction and the most anterior portion of the alveolar ridge. /s/ had a low tongue body with a retracted tongue dorsum. /ʃ/ had a high degree of palatalization, exhibiting a much

narrower constriction in the midpalate. /x/ had a similar contour to /ʃ/. The major differences included a slight curling of the tongue tip and a much more advanced tongue dorsum. The posterior tongue body was also slightly more retracted and raised. This suggested a similar type of constriction for /x/ and /ʃ/, but with marginal differences that impacted the acoustics significantly.

Figure 3 presents the midsagittal tracings for US03.

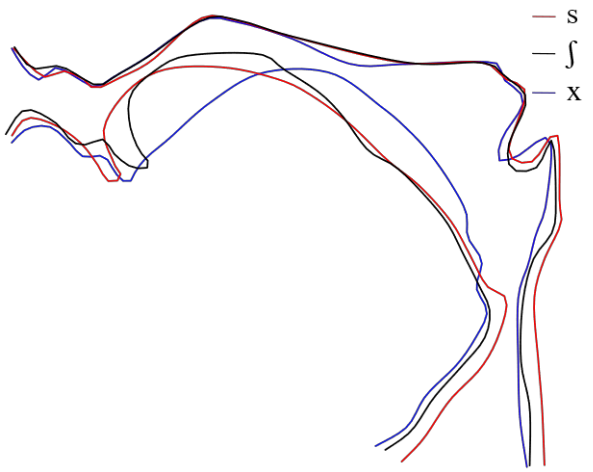


Figure 3: MRI traces of the midsagittal plane for lingual fricatives, /s/ (red), /ʃ/ (black), /x/ (blue) for US03.

Similar to US01 and US02, /s/ and /ʃ/ had overlapping constriction locations along the alveolar ridge, but /s/ had a more advanced constriction. /ʃ/ had a small, but noticeable sublingual cavity and an advanced tongue dorsum, while the tongue tip for /s/ was advanced towards the lower incisors. The tongue body for /ʃ/ was slightly raised along the alveolar ridge and hard palate, also indicating a small degree of palatalization. /x/ had a more retracted tongue dorsum and a constriction location along the soft palate. Additionally, for US03, /x/ did not have any observable sublingual cavity.

Figure 4 presents the MRI traces for participant US04. /s/ had a more advanced constriction right at the most anterior region of the alveolar ridge, compared to /ʃ/, which had a longer constriction along the posterior regions of the alveolar ridge. For both /s/ and /ʃ/, there was extensive tongue body raising, although for /s/ the raising was more in the mid-palate region and was achieved with the posterior tongue body. /ʃ/ had more anterior raising of the tongue body, congruent with the longer constriction along the posterior alveolar ridge. /x/ was produced with a long constriction along the mid-palate that extended towards the anterior portions of the hard palate. We also observed significant posterior tongue body and dorsum retraction into the velar and pharyngeal regions.

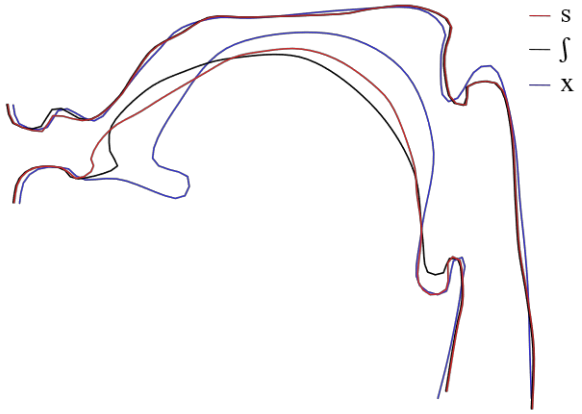


Figure 4: MRI traces of the midsagittal plane for lingual fricatives, /s/ (red), /ʃ/ (black), /x/ (blue) for US04.

3.2 Simulation Results

Table 2 presents the COG for each segment as produced by each speaker as extracted from the Praat analysis.

	/s/	/ʃ/	/x/
US1	8274	3711	1471
US2	9560	3328	4848
US3	7610	3725	3024
US4	8322	3589	2116

Table 2. COG (in Hz) for each of the segments, /s, ʃ, x/, for each of the participants US01, US02, US03, US04.

Table 3 presents the COG of the simulated data for each tongue shape for each segment using VocalTractLab. Figure 5 below presents an example simulation for US1 /ʃ/ with overlapped MRI tracing.

	/s/	/ʃ/	/x/
US1	6440	3749	1876
US2	6339	3456	4519
US3	6519	3966	3323
US4	5477	3282	2215

Table 3. Simulated COG (in Hz) for each of the segments, /s, ʃ, x/, based on MRI tongue contours for each of the participants US01, US02, US03, US04.

4. DISCUSSION

The MRI results revealed wide variation in the articulation of /s, ʃ, x/ across all speakers. Relative place of articulation across segments was consistent – that being that /s/ was most anterior, followed by /ʃ/, and finally /x/ – but that the precise location and length of the constriction along the palate varied by speaker. Significant overlap in constriction location was also observed between /s, ʃ/, implicating the shape of the constriction (i.e., grooved vs. domed) as an important predictor of acoustics [14]. The front cavity geometry also varied significantly with

respect to length, presence and size of sublingual cavity, and anterior tongue shape, which all may be relevant in the case of sibilants in order to direct airflow at the incisors [10, 11, 12]. Additionally, laminality or apicality was found to vary by speaker [11]. Differences in the anterior cavity for /x/ are likely related to retraction of the tongue body and dorsum to produce a constriction in the posterior palate. One limitation is that the teeth were not visible during the MRI, which play an important role in fricative acoustics [1, 5, 13].



Figure 5: Example vocal tract geometry for simulation of US1's articulation of /ʃ/ (black) with MRI of their vocal tract (red).

Simulations with VocalTractLab implicated specific regions of importance for differentiation between /s, ʃ/. Overlapping constrictions with different degrees of anterior tongue body raising played an important role in lowering COG to resemble the modelled /ʃ/. Doming or grooving of the tongue for either of /s, ʃ/ did impact COG simulations, but tongue posture and constriction location and length differences for /s, ʃ/ might implicate a biomechanical necessity rather than specific achievement of an acoustic target.

With respect to /x/, we observed that the posterior cavity played a significant role in the simulated COG, which may be due to the effects on preconstriction airflow. One limitation of the simulations was differences in the shape of the alveolar ridge and hard and soft palate for the model compared to each individual participant.

Taken together, we suggest that variation in articulation is not due simply to speaker specific vocal tract geometry. Rather, it seems that specific acoustic targets can be achieved with significant differences in place of articulation, and the front and back cavity geometry.

5. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101018840. We would also like to thank the Institut für Sorabistik at Universität Leipzig for their help scheduling participants.

6. REFERENCES

[1] Badin, P. (1989). Acoustics of voiceless fricatives: Production theory and data. Speech Technol. Lett, 45-52.

[2] Birkholz, P. (2023). *Image 3D*. https://www.vocaltractlab.de/index.php?page=image3d-download.

[3] Birkholz, P. (2013). Modeling consonant-vowel coarticulation for speech synthesis. *PLoS ONE*, 8(4): e60603.

[4] Boersma, P. & Weenink, D. (2022). *Praat: doing phonetics by computer* [Computer program]. Version 6.3.03.

[5] Fujiso, Y., Nozaki, K., & Van Hirtum, A. (2015). Estimation of minimum oral tract constriction area in sibilant fricatives in aerodynamic data. *Journal of the Acoustical society of America* EL20-EL25.

[6] Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, 142-174.

[7] Howson, P. (2017). Upper Sorbian. *Journal of the International Phonetic Association* 47, 1-9.

[8] Inkscape Project. (2020). *Inkscape*. Retrieved from https://inkscape.org.

[9] Jannedy, S. & Weirich, M. (2016). The acoustics of fricative contrasts in two German dialects. In Draxler C. & Kleber, F. (eds.), *Proceedings of Phonetics and Phonology in German Speaking Areas*, pp. 70-73. Ludwig-Maximilians-Universität München: Munich, Germany.

[10] McGowan, R. S. & Nittrouer, S. (1988). Differences in fricative production between children and adults: Evidence from an acoustic analysis of /ʃ/ and /s/. *Journal of the Acoustical Society of America* 229-236.

[11] Narayanan, S., Alwan, A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 1325-1437.

[12] Shadle, C. H. (1990). Articulatory-acoustic relationships in fricative consonants. In Hardcastle, W. & Marchal, A. (eds.), *Speech Production and Speech Modelling*, pp. 187-209. Kluwer Academic: New York.

[13] Shadle, C. H. (1991). The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics*, 19(3-4), 409-424.

[14] Shadle, C. H., Berezina, M., Proctor, M., & Iskarous, K. (2008). Mechanical Models of Fricatives Based on MRI-derived Vocal Tract Shapes. *8th International Seminar on Speech Production*, 417-420.

THE UTILITY OF TURNING-POINT ANALYSIS METHODS FOR WELSH AND IRISH INTONATION

Jack Pulman-Slater[1,a], Connor McCabe[2,b]

[1]Cardiff University, [2]Trinity College, Dublin
[a]pulman-slaterj1@cardiff.ac.uk, [b]cmccab1@tcd.ie

ABSTRACT

Two turning-point transcription systems (INTSINT and PoLaR) were used to examine intonation data from Southern Welsh and Munster Irish. Each study required a turning point approach to avoid issues of stress identification complicated by long-term contact with majority English. The particular choice of system was informed by study-specific aims. The Welsh investigation employed an INTSINT analysis of L2 learner data to focus on contour shape. The Irish investigation focussed on timing relationships by use of a PoLaR analysis of L1 speech. Welsh findings reveal significant differences in L2 intonation for learners of differing L1 backgrounds. Irish findings reveal regional variation and diachronic change in the relationship between high pitch and intensity prominence.

This summary comparison of the two studies illustrates the utility of turning point analyses as an intermediate step in addressing ambiguous intonation data while avoiding premature claims about phonological representation.

Keywords: Intonation, turning points, minority, Celtic

1. INTRODUCTION

Autosegmental-metrical (AM) approaches to the description of intonation such as Tones and Breaks Indices (ToBI; [23]) analyse pitch contours as strings of discrete tonal targets. The implementation of these strings is described with reference to metrical structure, e.g. H*+L referring to a falling pitch-accent in which the high target is associated with a metrically prominent syllable. Crucially, however, such analyses require the reliable identification of such metrical prominence, whether in terms of lexical stress (in, e.g., English) or predictable phrasal prominence (in, e.g., French).

The issue of lexical- and phrasal-prominence identification is far from trivial. There is growing cross-linguistic evidence of (i) diversity in the phonological relevance of lexical-level prominence marking [7, 12, 15], (ii) robust differences in phonetic implementation of phonological prominence [6, 27], and (iii) so-called 'stress deafness' in evaluating prominence as an L2 listener [8-10, 32-35].

This paper summarises two studies in which such identifications cannot be confidently or unambiguously made, dealing with two Celtic languages: Welsh and Irish. In each case, complicated contact with English and ambiguity surrounding lexical stress necessitated the use of a phonetic system of intonation analysis that did not force premature judgments about phonological structure, although this may be of interest for future work. Two systems based on turning points (TPs) in F0 were selected for this purpose: the International Transcription System for Intonation (INTSINT) [14] and Points, Levels and Ranges (PoLaR) [1]. The choice of each was determined by the respective focuses of the Welsh and Irish studies.

The case of Southern Welsh is first examined in Section 2, followed by Munster Irish in Section 3. The paper concludes with summary remarks in Section 4.

2. SOUTHERN WELSH

Examinations of sentence-level intonation in Welsh are almost exclusively based on perceptual studies employing several idiosyncratic and largely incomparable systems of analysis and notation [18, 19, 25, 31]. The only substantial work which draws on AM-style analysis is Cooper's 2015 investigation of Anglesey Welsh [5].

The study described here employs the first turning points analysis of Welsh by use of the International Transcription System for Intonation (INTSINT) [14]. The application of INTSINT described here facilitated an initial examination of L2 Welsh intonation amongst learners with L1 Welsh-substrate influenced English (Welsh English) and L1 Standard Southern British English (SSBE).

Conducting a cross-linguistic investigation of intonation in Welsh and English is complicated by the presence of differing acoustic realisations of lexical stress in the two languages. Welsh and English speakers have traditionally been held to produce lexical stress by drawing on different acoustic resources [26: 445, 34]. Previous investigations of Welsh stressed syllables find them to have relatively shorter duration, lower intensity and lower pitch compared to post-stress syllables [32-35]. The difference between

Welsh and English stressed syllables can lead speakers of L1 English to make erroneous stress-placement judgments on Welsh words by relying on English-style lexical stress pitch cues [33]. Such findings have led to a functional separation between stress and pitch in Welsh being posited [2: 90].

The situations of synchronic and post-shift (i.e. Welsh to English) language contact in which the two languages exist have resulted in the presence of a Welsh prosodic substrate in some varieties of Welsh English [28-30]. In communities where Welsh retains community-level functions, the presence of cross-linguistic influence has been found in the realisation of lexical stress by bilingual Welsh speakers and monolingual speakers of Welsh English [17].

Employing AM analyses of intonation which require pitch events to be related to stressed syllables is problematic given the differing acoustic realisations of lexical stress and the situations of language contact and acquisition in which this study is situated. Separate language- and variety-specific criteria for identifying stressed syllables against which pitch events could be anchored would need to be established. Generating such acoustically defined criteria, whilst perfectly possible, would not lead to systematically comparable data. This methodological problem is compounded by the lack of previous acoustically-informed studies of Welsh intonation.

2.1. Data

This study employed the first turning-points-style analysis of Welsh intonation, also providing one of the first examinations of L2 Welsh sentential intonation. Data were collected from 8 speakers of L1 SSBE and 8 Speakers of Welsh English (4 from North Pembrokeshire in south-west Wales and 4 from south-east Wales). The former group were designated 'non-native learners' and the second group 'native learners'. All were following formal online Welsh courses at CEFR level B1. Realisation of sentence-level pitch contours was compared with localised native speaker baselines.

Samples of controlled, laboratory-style speech were collected from participants in two remote recording sessions via Zoom (one in Welsh and one in English). Participants read aloud a total of 36 sentences (18 in Welsh and 18 in English). Using recording freeware, participants recorded themselves producing a declarative, a yes/no question and a declarative question of each sentence in both languages. An effort was made to ensure as few voiceless segments were included in the sentences as possible, whilst also ensuring that each sentence had a meaningfully equivalent version in each language. A total of 405 sentences were extracted from the recording process and annotated using INTSINT labels in Praat textgrids [4].

2.2. Analysis

INTSINT allows for the equivalent of a narrow phonetic transcription of intonation (14: 14). It allows for 2 broad types of turning points to be recorded. They can be defined in relative terms to the previously occurring pitch movement as Higher (↑), Lower (↓), the same, slightly Downstepped (>) or slightly Upstepped (<). The second option for the transcriber is to record more global pitch movements to the extreme Top (⇑) or Bottom (⇓) of the individual speaker's range within an utterance. Relative scaling is not marked as the analysis focuses instead on the relative height of each successive turning point. TPs in brackets denote utterance-final pitch movements. Unlike AM-style analyses, or the PoLaR analysis considered in Section 3 below, INTSINT does not anchor pitch events to phonological percepts such as stressed or prominent syllables. Figure 1 below shows an example of an annotated Welsh declarative (*Mae William yn gyrru i Wynedd* 'William is driving to Gwynedd') as produced by an L1 Welsh English speaker from the Rhondda in south-east Wales.

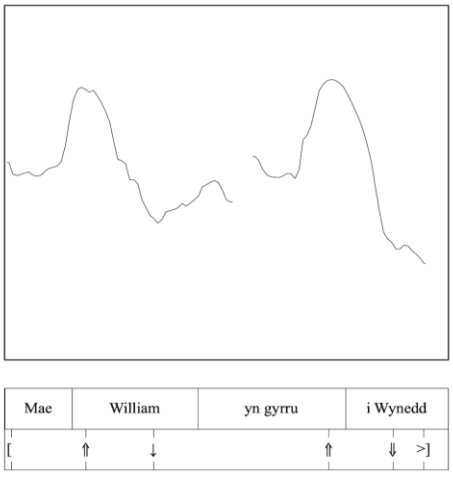


Figure 1: INTSINT annotation of the Welsh sentence *Mae William yn gyrru i Wynedd* ('William is driving to Gwynedd') spoken by a south-eastern Welsh learner (L1 Welsh English).

This study intended to provide an initial overview of utterance-level intonation in L2 Welsh and consider effects of transfer [11] from Welsh substrate-influenced L1 Welsh English. In order to make broad cross-linguistic comparisons, the INTSINT transcriptions for each utterance were converted into a series of 7 schematic pitch contours. The use of such schematic contours have been previously used in situations of synchronic and diachronic language contact between minority and majority languages [21].

Each schematic pattern was defined by the presence of a certain combination of TPs up to and including the pre-final TP. 'Sawtooth' patterns included an initial ↑/⇑ with a following ↓/⇓; a combination which repeated across the utterance in the case of double or triple sawtooths. Figure 1 shows an example of a double sawtooth pattern. Sequences of ↑/⇑-↓/⇓ have been described as a Welsh substrate-influenced feature of Welsh English sentential intonation [30].

2.2. Results

A binomial variable was established in which saw-tooth patterns were compared against all other schematic contours (i.e. those which didn't include ↑/⇑-↓/⇓ TPs). Results from linear mixed effects modelling found a significantly higher number of sawtooth patterns in Welsh compared to English in addition to varying degrees of the presence of a Welsh substrate between the two areas. Native learners (i.e. those with L1 Welsh English) in north Pembrokeshire, where Welsh retains community level functions, were found to use significantly more sawtooth patterns in Welsh than native speakers. However, contrary to the study's expectations, statistical modelling revealed that this significantly higher usage was not a transfer effect from the local L1 Welsh English variety. In the south-eastern region, on the other hand, the frequency of sawtooth patterns were unaffected by both the learners' L1 background. There were additionally no more general significant effect of speaker/learner on the use of sawtooth contours.

The analysis of utterance final TPs revealed that the use of <], ↑] and ⇑] was controlled by sentence type (Y/N question, declarative question and declarative) instead of by speaker or variety/language-based predictor variables. This finding runs contrary to previous work on Welsh English declaratives and interrogatives which has found high final pitch movements on both sentence types [28-29].

The INTSINT analysis employed presents an initial sketch of L2 Welsh intonation. Through facilitating inter-speaker/learner, intra-learner and inter-variety comparisons of sentential pitch shapes, this analysis also provides new intonational evidence of the potential weakening of the Welsh language substrate within Welsh varieties of English [24]. Through focussing on pitch movements without reference to stressed syllables, the study provides a starting out point for more-fine grained phonetic investigations of Welsh and Welsh English intonation.

3. MUNSTER IRISH

The Points, Levels, and Ranges (PoLaR) turning-point transcription system [1] was used to examine Munster Irish intonation, the first TP analysis of any Irish variety. This variety is said to exhibit a complex system of lexical stress assignment, in contrast to initial stress in other Irish varieties. Based on statistical findings for phonetic prominences at the lexical-level [16: 106-206], and suggestions in the literature that the rightward 'stress shift' in Munster Irish derived historically from high pitch occurring one or more syllables after lexical stress [3], it was decided to analyse the timing of high F0 relative to nearby heightened intensity as a starting point for the evaluation of such a misalignment hypothesis.

The distinctive character of Welsh lexical-stress marking noted in Section 2 was influential in encouraging scepticism of casual identification of stress in this related Celtic language. A TP analysis was therefore pursued to focus on the phonetic identification of regions of relatively high F0 in a contour without direct comment on the phonological status of the local metrically strong syllable (i.e. without explicitly diagnosing formal lexical 'stress' per se), and without categorising contour shapes.

3.1. Data

Naturalistic storytelling data from L1 MI speakers from two eras were used for this analysis.

The first dataset dates to 1928 – wax cylinder recordings, digitised by the Royal Irish Academy [22] and freely available online. Story (re)tellings from 20 male L1 MI speakers with age range 36-82 were selected. Modern (2020-21) data came from Zoom interviews in which story readings and retellings comparable to the 1928 data were collected from 14 L1 MI speakers, 11 female and 3 male, with age range 20-79.

All participants came from the Munster *Gaeltachtaí* (traditional Irish-speaking regions) of Counties Kerry, Cork, Waterford, Clare, and Tipperary. The latter two varieties went extinct by the end of the 20th century, and are only included in the 1928 data.

It was hypothesised that high pitch would frequently occur one or more syllables after metrical strength, especially in more conservative speech (i.e. less English-influenced) speech.

3.2. Analysis

PoLaR was selected for its modular structure, and for its flexible implementation. In contrast to INTSINT, used above with reference to Welsh, a PoLaR transcription includes the identification of 'strong' (if not

necessarily lexically 'stressed') syllables within utterances. This was crucial for the MI investigation's focus on timing. Analysis was carried out in Praat [4].

The system was adapted and operationalised for analysis of the selected storytelling recordings. Metrically strong syllables (PrStr/<*> in PoLaR notation) were identified primarily using maximum syllable intensity, supplemented by functional/lexical considerations. Content words were prioritised over adjacent function words for metrical strength. Strong syllables were labelled at the midpoint of the vowel, as suggested in the basic PoLaR conventions [1].

Intonation contours were then defined using a maximum of six TPs, with no more than one preceding the metrically strong syllable's vowel midpoint. Contours complex enough to initially require more than six TPs were scrutinised and either (i) divided into two contours, or (ii) filtered for superfluous TPs associated with, e.g., microperturbations in the F0 trace. Contours were identified as *nuclear* (phrase-final) or *prenuclear* (non-final) to allow for consideration of positional impact on high pitch timing. 8,487 contours were described in this way, using a total of 35,557 hand-placed TPs.

3.3. Results

A custom Praat script [20] was used to automatically extract described contours as strings of TPs with associated PoLaR levels, F0 value in Hertz (subsequently converted to semitones), and timing of the highest F0 point (or high plateau onset) relative to the local metrically strong syllable's vowel midpoint.

The latter revealed diversity in timing across eras and regions. An apparently conservative pattern of 100+ ms delays between strong-syllable vowel midpoints and the achievement of high pitch stood out in 1928 Cork (45% of prenuclear [PN] cases, 37% of nuclear [N] cases) and Kerry (45% PN, 38% N), and has been retained as a substantial minority pattern in 2020-21 (Cork: 35% PN, 34% N; Kerry: 21% PN, 16% N). This contrasts with a preference for closer alignment (+/-50 ms) between high pitch and metrical strength in 1928 Clare, Tipperary, and Waterford. This is retained in modern Waterford, and matches or outnumbers cases of large 100+ ms delays in modern Cork (36% PN, 34% N) and Kerry (56% PN, 51% N).

By divorcing the identification of metrical strength from the description of pitch-contour shape, the PoLaR analysis highlighted ambiguity in using F0 activity as a primary marker of lexical stress. In Figure 2, the light-heavy disyllable *troidín* 'fight.DIM' has intensity prominence on the initial syllable – the stressed syllable for this item in other Irish varieties – but excursion to a high pitch peak aligned with the end of the final heavy syllable predicted to receive lexical stress under typically-described MI conditions. This suggests that lexical stress in MI should not be uncritically diagnosed by F0 excursion, and further highlights problems that may arise in basing AM pitch-accent labels on presumed and/or prescribed stress location. Labelling for Figure 2's 1-1-5(-1) contour would likely vary (e.g. IViE L*+H versus H*+L [13]) if the initial or final syllable of *troidín* were identified as stressed.

These ambiguities would not be satisfactorily addressed by simply changing to impressionistic judgements of stress location, especially for an L1 English analyst. Results of the PoLaR approach to MI intonation data therefore demonstrate the utility of a decompositional phonetic approach to ambiguous intonation data.

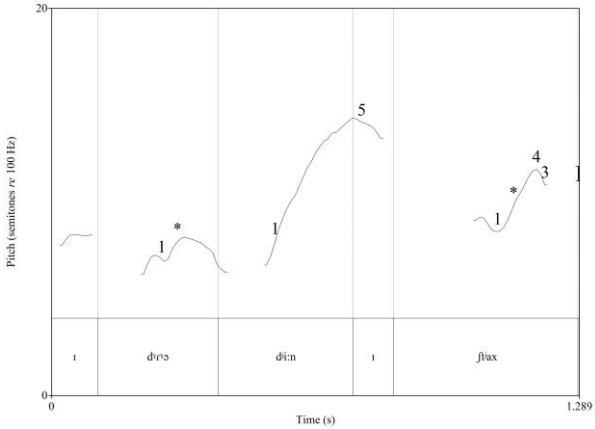


Figure 2: PoLaR annotation of 1928 Kerry utterance *i dtroidín isteach* from a monolingual MI speaker (b.1853). The initial light syllable /dʲɾʏɔ/ of *dtroidín* 'fight.DIM' is identified as 'strong' (<*>) based on intensity prominence, while high pitch (<5>) occurs just after the following heavy syllable /dʲiːnʲ/.

4. CONCLUSION

The use of PoLaR and INTSINT analyses sought to resolve the theoretical problems posed by the relationship between stress and pitch-contours. In the case of Southern Welsh, a shapes-based analysis was facilitated by INTSINT annotations which identified the presence of cross-linguistic influence in a language acquisition context. In the case of Munster Irish, a PoLaR analysis which adopted an ambiguous stance in the identification of lexical stress was nevertheless able to facilitate a timings-focussed analysis and shed light on the role of F0 excursion in this variety's realisation of lexical stress. Both studies illustrate the potential utility of turning-point analyses in initial phonetic investigations of the intonation of lesser studied languages in which AM pitch accent-labels cannot be easily applied.

5. REFERENCES

[1] Ahn. B., N. Veilleux, S. Shattuck-Hufnagel, Brugos, A. 2021. *PoLaR Annotation Guidelines*, Version 0.9.8. Accessed 18 October 2021, http://www.polarlabels.com/.

[2] Ball, M. J. 1989. The transcription of suprasegmentals in Welsh. *J. I.P.A. 19*(2), 89-96.

[3] Blankenhorn, V. 1981-82. Pitch, Quantity and Stress in Munster Irish. *Éigse 28*(2), 225-250.

[4] Boersma, P. & D. Weenink.

[5] Cooper, S. 2015. *Intonation in Anglesey Welsh*. Ph.D. dissertation. Bangor University.

[6] Cutler, A. 2011. Lexical Stress. In: Pardo, J.S., Nygaard, LC., Remez, R.E., Pisoni, D.B. (eds), *The Handbook of Speech Perception* (2nd ed.). Oxford: Blackwell Publishing Ltd, 264-289.

[7] Cutler, A., Jesse, A. 2021. Word Stress in Speech Perception. In: Pardo, J.S., Nygaard, LC., Remez, R.E., Pisoni, D.B. (eds), *The Handbook of Speech Perception* (2nd ed.). Oxford: Blackwell Publishing Ltd, 239-265.

[8] Dupoux, E. & S. Peperkamp. 2002. "Fossil markers of language development: phonological 'deafness' in adult speech processing". *Phonetics, phonology, and cognition 168*, 190-204.

[9] Dupoux, E., C. Pallier, N. Sebastián & J. Mehler. 1997. A destressing 'deafness' in French?. *J. of Memory and Language 36*(3), 406-421.

[10] Dupoux, E., S. Peperkamp & N. Sebastián-Gallés. 2001. A robust method to study stress "deafness". *J. A.S.A. 110*(3), 1606-1618.

[11] Ellis, N.C. 2008. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In: Robinson, P., Ellis, N.C. (eds), *Handbook of cognitive linguistics and second language acquisition*. Routledge, 382-415.

[12] Goedemans, R. & E. van Zanten. 2014. No stress typology. In: Caspers, J., Chen, Y., Heeren, W., Pacilly, J., Schiller, N.O., van Zanten, E. (eds), *Above and Beyond the Segments: Experimental linguistics and phonetics*. Amsterdam: John Benjamins, 83-95.

[13] Grabe, E., Nolan, F., Farrar, K. 1999. IViE – A Comparative Transcription System for Intonational Variation in English. *Proc. of the 5th Int. Conf. on Spoken Language Processing*.

[14] Hirst, D., Di Cristo, A. 1998. A survey of intonation systems. In: Hirst, D., Di Cristo, A. (eds.), *Intonation systems: A survey of twenty languages*. Cambridge University Press, 1-44.

[15] Hyman, L. 2014. Do all languages have word accent?. In: van der Hulst, H. (ed.) *Word Stress: Theoretical and Typological Issues*. Cambridge: CUP, 56-82.

[16] McCabe, C. 2023. *Prosodic Structure in Munster Irish*. Ph.D. dissertation. Trinity College, Dublin.

[17] Mennen, I., Kelly, N., Mayr, R., Morris, J. 2020. The Effects of Home Language and Bilingualism on the Realization of Lexical Stress in Welsh and Welsh English. *Frontiers in Psychology 10*, 3038.

[18] Pilch, H. 1975. Advanced Welsh phonemics. *Zeitschrift für celtische Philologie (ZcP) 34*(1), 60-102.

[19] Rhys, M. 1984. Intonation and the discourse. In: Ball, M.J., Jones, G.E. (eds), *Welsh phonology*. University of Wales Press.

[20] Rodgers, A. 2022. GetTimeAndF0Data [plugin]. Available at: https://github.com/AERodgers/plugin_getTimeAndF0Data

[21] Rodríguez Vázquez, R. 2019. The intonation of wh-questions in a language contact situation: the case of Galician and Galician Spanish bilingual speakers. *Estudios de Fonética Experimental* 28, 81-124.

[22] Royal Irish Academy. 2006. *The Doegen Records Web Project*. Available at: https://doegen.ie

[23] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.. 1992. TOBI: A standard for labelling English prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing*, 867-870.

[24] Thomas, A. 1984. Welsh English. In: P, Trudgill (ed.) *Language in the British Isles*. Cambridge University Press, 178-194.

[25] Thomas, C.H. 1967. Welsh intonation-a preliminary study. *Studia Celtica 2*, 8-28.

[26] van der Hulst, H., Hendriks, B., van der Veijer, J. (1999). A survey of word prosodic systems of European languages. In: H. Van der Hulst (ed.), *Word prosodic systems in the languages of Europe*. Mouton de Gruyter, 425-475.

[27] Vogel, I., A. Athanasopoulou & N. Pincus. 2016. Prominence, Contrast, and the Functional Load Hypothesis: An Acoustic Investigation. In: Heinz, J., Goedemans, R. & van der Hulst, H. (eds), *Dimensions of Phonological Stress*. Cambridge University Press, 123-167.

[28] Walters, J. R. 1999. *A study of the segmental and suprasegmental phonology of Rhondda Valleys English*. Ph.D. dissertation. University of Glamorgan.

[29] Walters, J. R. 2001. English in Wales and a 'Welsh Valleys' accent. *World Englishes 20*(3), 285-304.

[30] Walters, J. R. 2003. On the intonation of a South Wales Valleys accent of English. *J. I.P.A. 33*(2), 211-238.

[31] Watkins, T. A. 1961. *Ieithyddiaeth: Agweddau ar astudio iaith*. Gwasg Prifysgol Cymru.

[32] Williams, B. 1983. *Stress in modern Welsh*. Ph.D. dissertation. University of Cambridge.

[33] Williams, B. 1985. Pitch and duration in Welsh stress perception: the implications for intonation. *J. of Phonetics 13*, 381-406.

[34] Williams, B. 1986. An acoustic study of some features of Welsh prosody. In: John Lewis, C. (ed.) *Intonation in Discourse*. London: Croom Helm, 35-51.

[35] Williams, B. 1999. The phonetic manifestation of stress in Welsh. In: van der Hulst, H. (ed.), *Word Prosodic Systems in the Languages of Europe*, (pp. 311-335). Berlin: Mouton de Gruyter, 311-335.

Variable realization of the Arapaho glottal stop, despite its being distinctive and frequent

D. H. Whalen, Christian T. DiCanio and Wei-Rong Chen

City University of New York/Yale Child Study Center; University at Buffalo; Yale Child Study Center
douglas.whalen@yale.edu; cdicanio@buffalo.edu; wei-rong.chen@yale.edu

ABSTRACT

Complete glottal closure is typically treated as the canonical realization of a glottal stop, but it is quite unusual in running speech [1, 2]. Further, non-phonemic glottal stops are quite variable [3-5]. Even languages with contrastive glottal stops exhibit extensive variation [6-11]. Glottal stops are both common and contrastive in Arapaho (ISO 639 arp): Do they therefore have more canonical phonetic realizations? Glottal stops in an Arapaho corpus [12] were classified by phonetic realization and examined for duration and harmonics ratio (HR). Full glottal stops were rare (31.6%) and instead often realized as glottalization. Duration was longer for stops than for other realizations (but shorter than that of /t/). HR differed across all realizations. Arapaho reflects the pattern of other languages in using few stop realizations for the common, phonemic glottal stop, but may do so more often than reported for other languages.

Keywords: Glottal stop; Arapaho; variability; duration; harmonics ratio.

1. INTRODUCTION

Complete closure of the glottis is assumed to be the canonical realization of glottal stop [7]. Yet glottal stop realization has been found to be quite variable across many languages (e.g. [1, 10]). Glottal stops are found, at least allophonically, in 1131 of 3020 phonological inventories listed under PHOIBLE (37.5%) [13], but they are non-distinctive in other languages like English and Spanish. Does the presence of a phonemic contrast among oral stops and a glottal stop in a language predict the glottal stop's manner of realization? Some work assumes this to be true [14], but it has not been explicitly tested for many languages. The current study examines the variation in the realization of contrastive glottal stops in Arapaho as a test case.

While there is a substantial literature on variation in the production of non-modal phonation (for recent reviews, see [15, 16]), there is notably less instrumental phonetic work on the realization of glottal stops. In languages that lack contrastive glottal stops, such as English and German, glottal stops are often realized without complete glottal closure, but this variation is prosodically-mediated [3, 5, 17]. In languages with contrastive glottal stops, such as Itunyoso Triqui and Yucatec Maya, non-modal phonation is a frequent realization of the stop [6, 18]. A recent study on Hawaiian examined the realization of glottal stops across word positions within a corpus of radio program speech from eight speakers [2]. Unlike past studies, Davidson [2] categorized realizations of glottal stops into eight groupings related to the presence/absence of complete closure, the location of creaky phonation if complete closure was absent, and realizations involving modal phonation with or without local intensity perturbations (see [8, 21]). She found that glottal stops were realized with complete closure just 10-12% of the time in word-initial position and 2-6% of the time in word-medial position. Similar to studies on English and German, the realization of glottal stops in Hawaiian was prosodically-mediated. Therefore, overall, complete closure was rare in unscripted speech.

Glottal stop is sometimes treated as a glide in the phonology literature [19], but just as frequently as a stop [20]. We will be comparing it to /t/, and we can expect that the lack of an oral closure for glottal stop and/or its different phonological status may make it shorter in duration than /t/.

1.1. Arapaho

Arapaho (ISO 639-3: arp) is an Algonquian language currently spoken in Wyoming and Oklahoma. It has a relatively simple consonant inventory (see Figure 1), monophthongal vowels which contrast in length and quality /ɪ, iː, ɛ, æː, ʊ, uː, ɔ, ɔː/, and a few diphthongs /eɪ, ɪɬ, oʊ, aɪ/. For a phonetic description of the vowel system, see DiCanio and Whalen [21]. The phonetic properties of the stop inventory (excluding glottal stops) are investigated in Kakadelis [22].

	Bilabial	Interdental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Stop	b		t			k	ʔ
Affricate				tʃ			
Nasal			n				
Fricative		θ	s			x	h
Approximant	w				j		

Figure 1: Consonant inventory of Arapaho

Historical changes (from Proto-Algonquian (PA) to Arapaho-Gros Ventre (AGV)) resulted in many, distinctive glottal stops [23]. Goddard (1974) states

*"PA *x, *m/*n, *q, and *ʔ as first members of clusters all fall together to *ʔ."* [24]. Indeed, a rough estimate from Conathan's online Arapaho-English dictionary shows it to be the third most common consonant, accounting for 13% of all consonants. Do both its greater frequency in the lexicon and its distinctive status affect its realization?

1.2 Hypothesis

Given that Arapaho treats glottal stop as a stop and uses it with great frequency, we hypothesized that this would lead to greater use of the canonical, full glottal stop realization than in studies of other languages.

2. DATA

This question was studied by examining glottal stops in a portion of the corpus of Arapaho collected in Wyoming by Lisa Conathan (2004-2007) [12]. We selected two male and four female speakers for our analysis. Both narrative and elicited speech were analyzed. Overall, there were 493 tokens of glottal stop in elicited speech (98.6 per speaker) and 2117 in narratives (423.4 per speaker). Phonemic /t/ was measured for comparison, with 263 tokens in elicited speech and 1029 in narratives. A total of 284 minutes (4 hours, 44 minutes) of Arapaho speech was analyzed.

2.1. Alignment

Transcriptions were modified to use only ARPABET characters [25] so that an English-based aligner, P2FA [26], could be used. This required treating Arapaho glottal stop as a <T> since English lacks this phoneme. Those boundaries were hand corrected.

2.2 Delimiting glottal stops

We classified the realization of glottal stops into five groups (below):

Full glottal stop.
Creaky phonation.
Creaky phonation with local glottalization.
No glottalization.
Problematic stop.

These groupings were chosen organically by several raters of the Arapaho speech. The categories of *creaky phonation* and *creaky phonation with local glottalization* differed mainly in terms of the temporal extent of creaky voice. In the former, creaky phonation was found uniformly across the adjacent vowels, but in the latter, it was adjacent to a notably more aperiodic region corresponding to the lexically expected location for the glottal stop. The last category was used for those glottal stops whose realizations we could not categorize. Only 6.1% of all observed glottal stops fell into this category.

Each category was also measured for its duration. For full stops, this included the silence and adjacent noise (aspiration). The two creaky phonation types were challenging, but the portion that had the least clear formant structure was taken. Duration for tokens that lacked glottalization was taken as the stretch of low amplitude and/or low F0 associated with the glottal gesture; see [8, 27] for a discussion of glottalization-induced effects on amplitude and F0.

The underlying /t/s were analyzed for comparison with /ʔ/. As with full glottal stops, silence and associated aspiration were included as part of the durational measurement.

2.3. Calculating harmonics-to-noise ratio

One useful measure of voice quality is the harmonics-to-noise ratio (HNR). The less noise, the higher the ratio, and the more "modal" the phonation [28]. Though there are various implementations of HNR, it always relies on the presence of periodicity in the signal. Creaky phonation or glottalization, which we often find in the context of different realizations of a glottal stop, can strongly affect periodicity. HNR values may be indeterminate if there is substantial aperiodicity.

Instead of a typical measure of HNR, we used the "harmonic ratio" (HR) of Kim et al. [29] implemented in MATLAB, bounded between 0 (white noise) and 1 (sine wave). HR values were calculated continuously with 30 ms sliding windows stepped by 1 ms. We measured the median HRs during the glottal stop itself, and during /t/ as a comparison, weighted by the number of tokens per speaker. The advantage of this method is the ability to track HR changes in the speech signal regardless of the observed periodicity.

3. RESULTS

3.1 Category frequencies

The percentages of realizations for the glottal stops are shown in Table 1. These are averaged across talkers and speaking style (narrative or elicited speech). More glottal stops were realized without full closure than with full closure in the data (65.9% vs. 34.1%, respectively). This matches findings in other languages, such as Hawaiian [2]. However, the proportion is higher in Arapaho by a factor of 2 or 3, depending on the study. See Table 2 for realization percentages by speaker.

Realization	Average %	N
Full glottal stop	34.1	890
Creaky phonation	28.1	734
Creaky phonation with local glottalization	19.5	508
No glottalization	12.2	319
Problematic stop	6.1	159

Table 1: Glottal stop realization across speakers and styles.

Speaker	% stop	N
ZM	59.8	417
RMG	14.2	54
JA	33.3	23
MKU	24.6	170
ASE	25.1	87
LD	32.5	139
Mean/total	31.6	890

Table 2: Glottal stop realized as full stop, by speaker, across styles.

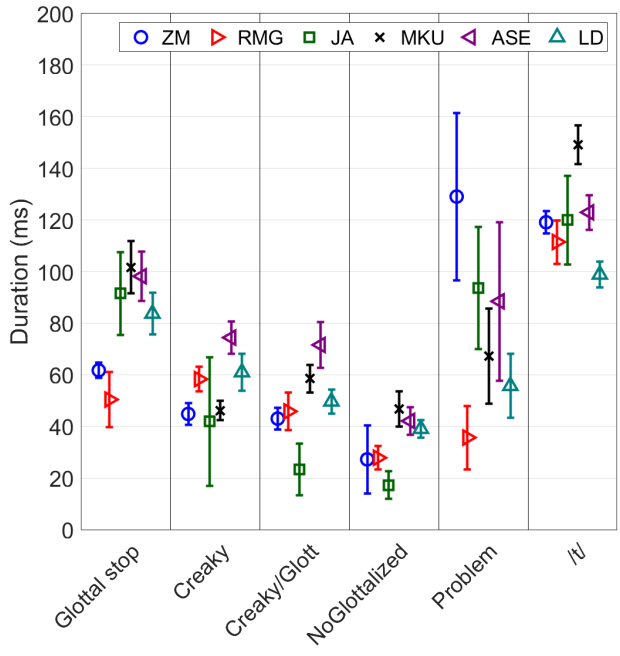


Figure 2: Duration of realizations of glottal stop and of /t/. Individual speakers' values are given, collapsed across speaking style. Error bars show 95% confidence intervals (CIs) of the mean.

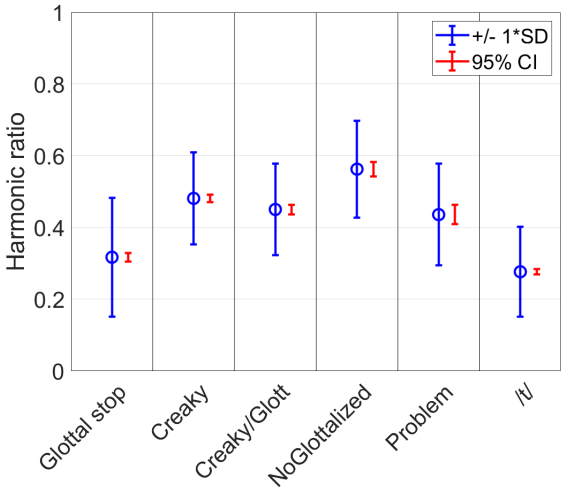


Figure 3: HR for realizations of glottal stop and of /t/. Averaged across speaker and style. Error bars show 95% Cis in red and +/- 1*SD in blue.

3.2 Duration

The average durations of the five realizations, and that of /t/, are shown in Figure 2. Complete closures are about 50% longer than the creaky phonation realizations (82.0 vs. 52.3 ms). Cases where no glottalization (i.e., aperiodicity) was observed had the shortest average duration (31.8 ms). These patterns suggest that there is a relationship between the duration of a glottal stop and its degree of lenition: full closure > creaky phonation > only intensity/f0 perturbations. These span the range from longer duration to shorter duration.

We also compared durations across the two speech styles, narrative and elicitation. Though we have recordings of both styles for most speakers, for speaker ZM, we had only narrative speech data, and for speaker JA, we had only elicited speech data. For some speakers, there were no instances of a glottal stop category for a particular style. The durations of both glottal stop and /t/ were quite similar in elicited speech and narratives. Because of these gaps, only about half of the comparisons were available for comparing realizations across styles. Either the speaker lacked any data for the style (two speakers), or the speaker had both styles but one of the styles did not have any instances of the realization. The average difference in duration across speech styles was 15.7 ms. The range was 77.3 ms. The slightly positive difference may simply be due to a slower speaking rate in elicited speech.

3.3 HR

Results for HR are shown in Figure 3. The non-glottalized realizations had the largest HR. The other glottalized versions were lower than that and similar

to each other. As can be expected, both the stops (glottal and underlying /t/) had the lowest values. Although not a full statistical test, CI overlap showed /t/ to have the lowest value, closely followed by glottal stop; the realization with no glottalization has the highest value, and the two realizations with creaky phonation had similar values.

4. DISCUSSION

Glottal stops were highly variable in the production of Arapaho speech, though as observed in past studies, complete glottal closure was significantly less common than realizations consisting of non-modal phonation. Complete glottal closure was rather more common in Arapaho (34.1%) than in Hawaiian (2-12% by context) [2]. One explanation for this may relate to differences in speech rate across languages. Though we do not have data on Hawaiian speech rate, previously reported data indicate that Arapaho speech is notably slower than either Sierra Norte de Puebla Nahuatl or Bardi [16]. Since the presence of complete glottal closure varies inversely with produced duration (both here and in Hawaiian), one possibility is that the longer realizations of underlying glottal stop in Arapaho result in a reduced potential for target undershoot [30] to occur with glottal closure.

Although both Arapaho and Hawaiian have distinctive glottal stops, their functional load is likely to be higher in Hawaiian. Most word-initial glottal stops in Arapaho occur within a small set of particles or interjections, whereas Hawaiian makes extensive use of it word-initially. Many final glottal stops in Arapaho are also epenthetic, while Hawaiian lacks coda consonants altogether.

Examining prosodic effects on the realization would be of extreme interest given the observations that the realization of glottal stops varies by prosodic position. However, we did not have the ability to do so here. Using unscripted speech corpora for phonetic data analysis increases the ecological validity of the observed phonetic patterns. However, a disadvantage is the difficulty of annotating prosodic detail given the great variability of the prosodic contexts. We could only reliably code for word position in the current corpus, and the limitations on initial glottal stop (only a few interjections and particles) would have made the comparison a weak one.

5. CONCLUSION

We have shown that Arapaho glottal stops are highly variable in their realization. This variation is strongly correlated with the duration of the glottal stop itself. Full glottal closure occurs in contexts of longer duration and is produced with low HR values; non-modal phonation is more common with shorter duration glottal stops and is produced with higher HR values; further, a lack of glottalization is more common with the shortest overall duration and is produced with the highest HR values. Full glottal closure is uncommon in the Arapaho corpus and these findings demonstrate a strong pattern of duration-induced reduction in spontaneous speech production [11, 30].

The Arapaho realization of distinctive glottal stop is similar to what has been observed in other languages, but the frequency of the “canonical” realization as a true stop is, in this corpus, approximately twice as high as that found for other languages. We theorize that this particular finding may relate to differences in speech rate across languages, though research on additional corpora of languages with contrastive glottal stops is required to test this hypothesis.

The lack of an effect of phonemic status both in Arapaho and other languages on the proportion of full glottal stop realizations would appear to relate to the ease with which listeners accommodate to the various realizations. The great variability in the realization of tones or nasalization suggests that canonical forms need not be the most frequent for a wide variety of linguistic signals.

Phonetic descriptions of endangered languages seldom include accounts of speech variation [31], even though such variation is important for language learners. The work here serves the additional goal of providing a description of this variation for Arapaho. The current results show that Arapaho is typical in its realization of distinctive glottal stops. However, it is atypical relative to Hawaiian and other languages in the frequency with which full glottal closure is produced by speakers. We hope that further research will reveal the precise mechanisms explaining these patterns.

6. ACKNOWLEDGMENTS

This work was supported by NSF Grant No. 0966411 and NIH grant DC-002717 to Haskins Laboratories.

7. REFERENCES

[1] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Oxford, UK; Cambridge, MA: Blackwell, 1996.

[2] L. Davidson, "Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian," *J. Phon.*, vol. 88, p. 101075, 2021/09/01/ 2021.

[3] L. C. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels

as a function of prosodic structure," *J. Phon.,* vol. 24, pp. 423-444, 10// 1996.
[4] J. B. Pierrehumbert and D. Talkin, "Lenition of /h/ and glottal stop," in *Papers in laboratory phonology II: Gesture, segment, prosody*, G. J. Docherty and D. R. Ladd Eds. Cambridge: Cambridge University Press, 1992, pp. 90-117.
[5] M. Garellek, "Voice quality strengthening and glottalization," *J. Phon.,* vol. 45, pp. 106-113, 7// 2014.
[6] C. T. DiCanio, "Coarticulation between tone and glottal consonants in Itunyoso Trique," *J. Phon.,* vol. 40, pp. 162-176, 2012.
[7] J. H. Esling, K. E. Fraser, and J. G. Harris, "Glottal stop, glottalized resonants, and pharyngeals: A reinterpretation with evidence from a laryngoscopic study of Nuuchahnulth (Nootka)," *J. Phon.,* vol. 33, pp. 383-410, 10// 2005.
[8] C. Gerfen and K. Baker, "The production and perception of laryngealized vowels in Coatzospan Mixtec," *J. Phon.,* vol. 33, pp. 311-334, 7// 2005,.
[9] C. Gerfen, "Topics in the phonology and phonetics of Coatzospan Mixtec," Ph.D. dissertation, University of Arizona, 1996.
[10] L. Storto and D. Demolin, "The phonetics and phonology of South American languages," in *The Indigenous languages of South America*, L. Campbell and V. Grondona Eds. Berlin: Mouton de Gruyter, 2012, pp. 331-390.
[11] C. T. DiCanio, W.-R. Chen, J. Benn, J. Amith, and R. C. García, "Extreme stop allophony in Mixtec spontaneous speech: data, prosody, and modelling," *J. Phon.,* vol. 92, no. 101147, pp. 1-18, 2022.
[12] L. Conathan, *Northern Arapaho linguistic corpus*: Endangered Language Archive, School of Oriental and African Studies, U. of London, 2004-2007.
[13] *PHOIBLE 2.0*. (2019). Max Planck Institute for the Science of Human History, Jena. Accessed: 2021-12-15. [Online]. Available: http://phoible.org
[14] D. Silverman, *Phasing and recoverability*. New York: Garland Publishing, 1997.
[15] L. Davidson, "The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages," *WIREs Cognitive Science,* vol. 12, no. 3, p. e1547, 2021/05/01 2021, doi: 10.1002/wcs.1547.
[16] M. Garellek, "The phonetics of voice," in *The Routledge handbook of phonetics*, W. F. Katz and P. F. Assmann Eds. London: Routledge, 2019, pp. 75-106.
[17] M. K. Huffman, "Segmental and prosodic effects on coda glottalization," *J. Phon.,* vol. 33, pp. 335-362, 7// 2005.
[18] M. Frazier, "The production and perception of pitch and glottalization in Yucatec Maya," Ph.D. dissertation, University of North Carolina, Chapel Hill, 2009.
[19] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT, 1951/1963.
[20] R. Jakobson, *Child language, aphasia and phonological universals*. The Hague: Mouton, 1968.
[21] C. T. DiCanio and D. H. Whalen, "The interaction of vowel length and speech style in an Arapaho speech corpus," in *Proceedings of the 18th International Congress of Phonetic Sciences*, vol. Paper number 513, The Scottish Consortium for ICPhS 2015 Ed. Glasgow: University of Glasgow, 2015, pp. 1-5.
[22] S. Kakadelis, "Phonetic properties of oral stops in three languages with no voicing distinction," Ph.D. dissertation, City University of New York, 2018.
[23] I. Goddard, "Recovering Arapaho etymologies by reconstructing forwards," in *Mír curad: Studies in honor of Calvert Watkins*, J. Jasanoff, H. C. Melchert, and L. Olivier Eds. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck, 1998, pp. 183-200.
[24] I. Goddard, "An outline of the historical phonology of Arapaho and Atsina," *International Journal of American Linguistics,* vol. 40, pp. 102-116, 1974..
[25] M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of occurrence of phonemes in conversational English," *Language and Speech,* vol. 21, pp. 221-241, July 1, 1978 1978, doi: 10.1177/002383097802100302.
[26] J. Yuan and M. Y. Liberman, "Speaker identification on the SCOTUS corpus," *The Journal of the Acoustical Society of America,* vol. 123, p. 3878, 2008.
[27] J. M. Hillenbrand and R. A. Houde, "Role of F0 and amplitude in the perception of intervocalic glottal stops," *J. Speech, Lg., Hear. Res.,* vol. 39, pp. 1182-1190, 1996.
[28] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Journal of the Acoustical Society of America,* vol. 71, pp. 1544–1550, 1982.
[29] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: Audio content indexing and retrieval*. Hoboken, NJ: Wiley Online Library, 2005.
[30] B. Parrell and S. Narayanan, "Explaining coronal reduction: Prosodic structure and articulatory posture," *Phonetica,* vol. 75, no. 2, pp. 151-181, 2018, doi: doi:10.1159/000481099.
[31] D. H. Whalen, C. T. DiCanio, and R. Dockum, "Phonetic documentation in three collections: Topics and evolution," *Journal of the International Phonetic Association,* vol. 52, no. 1, pp. 95-121, 2022.

THE REALIZATION OF VOICELESS NASAL CONSONANTS IN UPPER TANANA (DENE)

Olga Lovick[1], Kristy Reyes Herrera[1], and Jesse Stewart[1]

[1]University of Saskatchewan
Olga@lithophile.com, kbr458@mail.usask.ca, stewart.jesse@usask.ca

ABSTRACT

This study describes a rather rare nasal voicing contrast in Upper Tanana (Dene/Athabascan). The language has a rich inventory of nasal segments, distinguishing oral and nasal vowels and three phonemic consonants /m/, /n/, /n̥/. The voicing status of the coronal nasal (/n/ vs. /n̥/) is morphologically conditioned and surfaces as marking the contrast between light (unsuffixed) vs. heavy (suffixed) stems. We use the Earbuds Method [1] to analyze nasal events acoustically in order to better understand the processes involved in this contrast. Data come from two female speakers (both in their 80s). The results of this analysis suggest that, acoustically, the contrast between /n/ vs. /n̥/ may be more precisely described as a timing contrast rather than a voicing contrast. Additionally, there is evidence that the entire nasal gesture may be shifting leftward into the vowel in the light stem variant, mirroring other developments in the Upper Tanana stem syllable.

Keywords: Upper Tanana, Nasality, Earbuds, Dene/Athabascan, Timing Gestures

1. INTRODUCTION

Upper Tanana is a Dene (Athabascan) language spoken by fewer than 50, mostly elderly, individuals in eastern interior Alaska (USA) and the Yukon Territory (Canada). It has a large phoneme inventory distinguishing nasal and oral vowels as well as three nasal consonants /m/, /n/, /n̥/ [2, p. 43]. Nasal vowels can be reconstructed as sequences of vowels followed by nasal coda consonants [3]. Stem vowel nasalization has affected all long vowels (e.g., /nãː/ ‘mother’, /t͡sʰõː/ ‘grandmother’, /xĩː/ ‘melt (pfv.)’) but is less regularly found in stems with short vowels. Thus we find /ɲ̃ih/ ‘see (ipfv.)’ and /tʰãj/ ‘trail’, but also forms such as /tit͡ʃʰin̥/ ‘stick’ or /tonʔ/ ‘knock (pfv.)’.

The voicing contrast between /n/ and /n̥/ has a morphological origin between light (unsuffixed) and heavy (suffixed) stems. Voiceless /n̥/ occurs only word-finally in the absence of anysuffixes. Voiced /n/ occurs in syllables containing a historically vocalic suffix which has been absorbed into the stem. This morphological alternation has evolved into a surface contrast between word-final nasal phones, e.g., /nelt^{h}ʌn̥/ ‘it is frozen’ vs. /nelt^{h}ʌn/ ‘the frozen one’; /t͡θ’an̥/ ‘bone’ vs. /ʃt͡θ’anʔ/ ‘my bone’; or /tit͡ʃʰin̥/ ‘stick’ vs. /ʃtit͡ʃʰinʔ/ ‘my stick’. This contrast exists only for stems with short vowels. For long stem vowels, the contrast is between long nasal vowels and long oral vowels followed by voiced /n/, e.g., /tiht͡sʰĩː/ ‘I am hungry’ vs. /tiht͡sʰiːn/ ‘I, being hungry’. Verbal suffixes include the (productive) nominalizing suffix -v, the negative suffix -v́, and the imperative suffix -v. The only productive nominal suffix has the shape -v̀ʔ, but several non-productive suffixes with the shape -v are also attested ([2] chap. 11).

In this paper, we show that contemporary /n̥/ represents a transitional stage between a Vn sequence and suprasegmentalization of nasality on the preceding vowel Ṽ.

2. EARBUDS METHOD

While numerous techniques and methods exist for analyzing nasality in human speech nearly all share a common problem—they are not practical outside of a clinical setting, especially when working with Elders. Therefore, we opted for the Earbuds Method [1], which is specifically designed for collecting nasal data from the field using a non-invasive technique with minimal equipment. The Earbuds Method is based on the principle of nasalance, a nasal to oral amplitude output ratio ([4]), but instead of collecting amplitude (intensity) readings from a nasometer (see [5]), a pair of low impedance earbuds are used. This technique involves placing one earbud directly under a nostril with the silicon tip facing upward, channeling air from the nostril directly toward the earbud diaphragm ([1, p. 50]). The other earbud is held in place by the consultant at the corner of the mouth. After normalization of the intensity (dB) contours from each track, the resulting data can be quantified using the nasalance ratio (a proportion of nasal to oral energy within the signal) ([1, p. 59]). This information provides accurate details about the amplitude, timing, and duration of nasal gestures (e.g., pre & post-nasalization, nasal leakage, & coarticulation effects), which is what we seek to

better understand in the nasal voicing contrast in Upper Tanana.

When assessing figures made using the Earbuds Method, purely nasal sounds (e.g., [n]) are correlated with greater amplitude in both the wave form and intensity curves in the nasal tracks (displayed in red) compared to the oral tracks (displayed in blue). Purely oral sounds (e.g., [s]) show the opposite trend. Conversely, nasal vowels (e.g., [ã]) show similar degrees of amplitude in both the nasal and oral tracks given that air escapes from both tracts during the production of nasal vowels. Transitions between nasal and oral sounds are correlated with gradual increases in amplitude in the respective track as the velum lowers and raises.

3. METHOD

3.1. Participants

Given that Upper Tanana is a highly endangered language, we could only gather data from two participants, both women in their 80's.

3.2. Data

In total, 127 tokens were collected: 78 from one speaker and 49 from the other. An elicited wordlist was used, and tokens were produced in isolation. The data were then annotated into individual segments using Praat [6].

3.3 Procedure

Following the recommendations in [1], a pair of low impedance earbuds were connected to the microphone jack of a Marantz Professional Solid State Recorder PMD 671 and the device was set to record in stereo. The participants were asked to hold one earbud just beneath their nostril with the silicon tip facing upward while the second earbud was placed at the corner of the mouth. Participants were asked to hold the earbuds still and produce each word twice at a normal speech rate.

3.4 Acoustic analysis

Each token containing /n/ and /n̥/ in coda position was extracted using Praat [6]. Next, each channel from each token was independently extracted and converted to intensity. The wave forms from both channels were then drawn with the nasal track placed above the oral track. Below the wave forms, the intensity curve from the nasal track (red) was transposed on the intensity curve from the oral track (blue). Finally, a TextGrid with each phoneme segmented was drawn over the images. This allowed for the analysis of nasal events in and across phoneme boundaries using two amplitude metrics (Decibels for the intensity contours & Pascals for the wave forms) to identify timing gestures associated with /n/ and /n̥/.

4. RESULTS

4.1. Heavy stems

When a nasal coda appears in a heavy (suffixed) stem, the segment is traditionally described as voiced /n/ [2, p. 133]. With respect to timing, the nasal gesture frequently begins midway through the vowel and continues throughout the duration of the coda. Heavy stems with this archetype are observed in word-final position in Figures 1 and 2 where the amplitude in the nasal tracks substantially increases midway through the vowel and is maintained until the coda concludes. The coda in Figure 1 lasts 108 milliseconds (ms) while the entire nasal gesture, including the vowel, totals 185 ms. Likewise, the coda in Figure 2 totals 176 ms, while the entire nasal gesture, including the vowel, totals 247 ms.

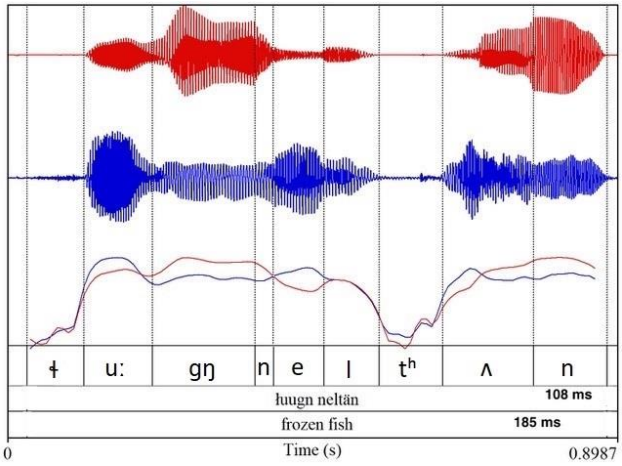


Figure 1: Word-final heavy stem with nasal coda in the word /ɬuːgŋ neltʰʌn/ *w*'frozen fish'

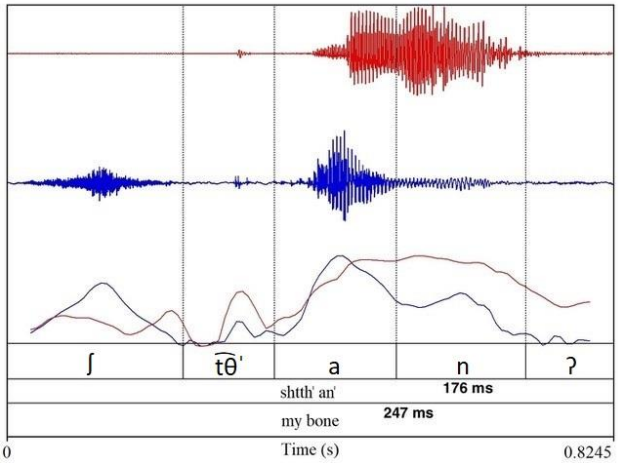


Figure 2: Word-final heavy stem with nasal coda in the word /ʃt͡θ'anʔ/ 'my bone'

4.2. Light stems

When a nasal coda appears in a light (unsuffixed) stem, the segment is traditionally described as voiceless /n̥/ [2, p. 133]. Yet the periodic wave cycles throughout the codas in Figures 3 and 4 suggest the segment may in fact be voiced. Instead, we observe that the contrast between heavy and light stem coda [n] might be better described as a difference in the timing of the nasal gesture. In light stems ending with coda [n] (as observed in Figures 3 & 4), the entire nasal gesture is shifted leftward into the vowel compared to the nasal gesture in heavy stems. This nasalizes a greater portion of the vowel yet shortens the duration of nasality in the coda. Consequentially, the nasal gesture concludes before the coda completes resulting in partial oralization of coda [n] ([n] → [ñ̥]).

In Figure 3, the nasalized portion of the coda only lasts 42 ms; substantially less than coda [n] in heavy stems. Yet two-thirds of the vowel in the light stem form is nasalized compared to roughly half the vowels in the heavy stem forms.

After the nasal gesture concludes, 24 ms of periodic wave cycles in the oral track wave form are observed throughout the remainder of the coda. Additionally, unlike in the heavy stem examples, both intensity curves drop sharply during the final phase, suggesting the latter half of the coda may be perceived as quieter. This could be the source of impressionistic observations of voiceless /n̥/ described in the literature.

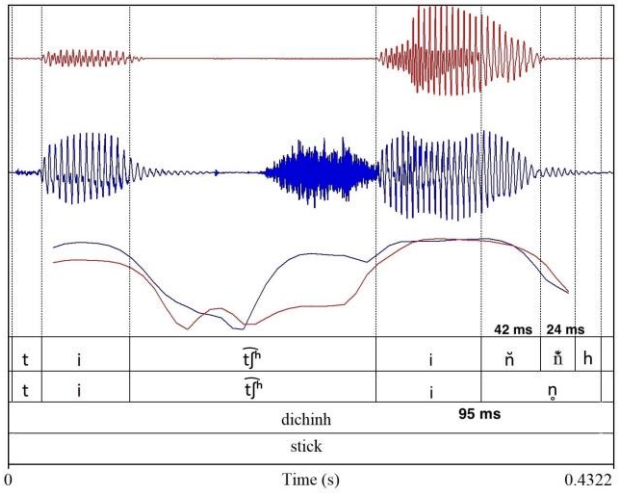


Figure 3: Word-final light stem with nasal coda in the word /titʃʰin̥/ 'stick'

In Figure 4, the nasal gesture nasalizes the entire vowel and concludes halfway through the coda. The nasalized portion of the coda lasts just 55 ms after which lower amplitude periodic cycles (seen in the blue wave form) continue for an additional 55 ms until the coda concludes.

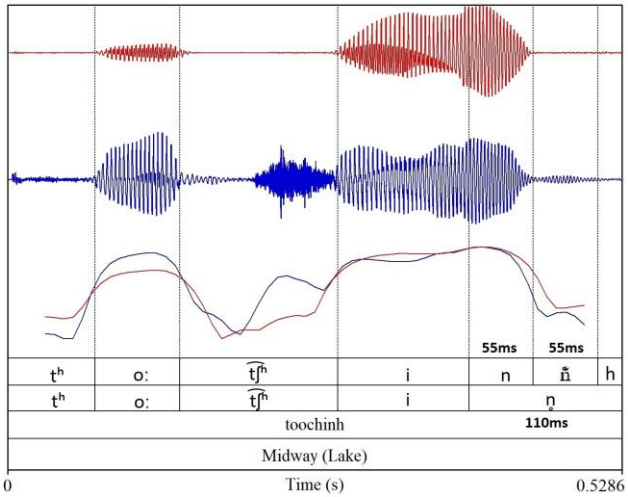


Figure 4: Word-final light stem with nasal coda in the word /tʰoːtʃʰin̥/ 'Midway Lake'

4.3. Quantifying Nasality in Upper Tanana Stems

To quantify these findings, we use the nasalance ratio, described in [1] as a proportion of nasal to oral energy throughout the signal. Nasalance is provided in Figures 5 & 6 as a line that fluctuates temporally between nasal (green) and oral (orange), which corresponds to the nasalance line crossing the 50% mark. Figures 5 and 6 with the nasalance lines were created in R (4.1.1) using the script provided in [1] and are formatted with the nasal and oral wave forms transposed to better illustrate the nasal gesture shifting leftward in the light stem example (Figure 6).

Figure 5 presents an example of a heavy stem form with a nasal coda in the word *dindeh etshyin* 'shaman'. As with Figures 1 and 2, the nasal gesture in the final syllable begins halfway through the vowel and concludes at the end of the coda. This observation is confirmed by the nasalance line crossing above the 50% mark 53% of the way through the vowel ((56 ms *crossing point* / 106 ms *total vowel duration*) *100) and dropping below the 50% mark upon conclusion of the coda. The entire nasal gesture is 139 ms with 56 ms positioned within the vowel and 83 ms positioned in the coda.

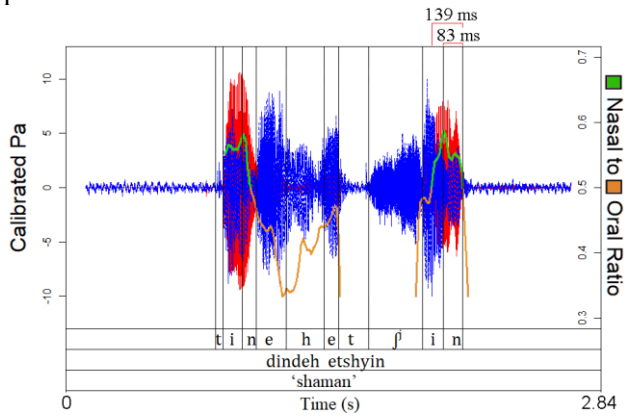


Figure 5: Word-final heavy stem with nasal coda in the word /tineh etʃʲin/ 'shaman'.

Figure 6 presents an example of a word-final light stem form with a nasal coda in the word *etshyinh* 'he's making magic'. As with Figures 3 and 4, the nasal gesture in the final syllable begins just after the onset of the vowel and concludes roughly halfway through the coda. This observation is confirmed by the nasalance line crossing above the 50% mark 18% of the way through the vowel ((21 ms *crossing point* / 95 ms *total vowel duration*) *100) and dropping below the 50% mark just 34% of the way through the coda leaving 66% of the coda with no nasality ((17 ms *until crossing point* / 50 ms *total coda duration*) *100).

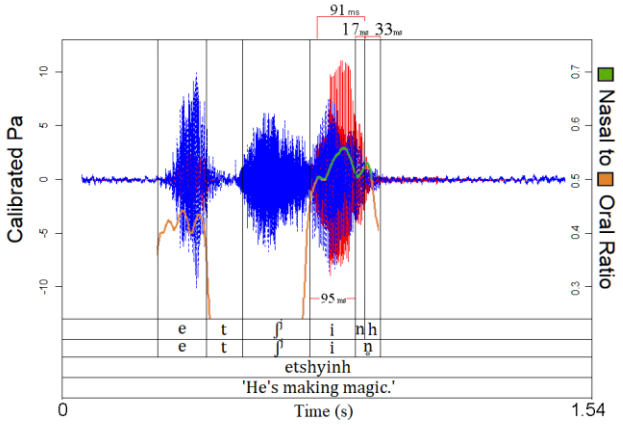


Figure 6: Word-final light stem with nasal coda in the word /etʃʲiɲ̥/ 'he's making magic'

As observed in Figures 1-6, the overall duration of the nasal gesture is variable, yet the proportions of the gesture that overlap the vowel and coda remain consistent. In heavy stem forms with a nasal coda, approximately half the vowel is nasalized and the entire coda is nasalized. In light stem forms with a nasal coda, the majority of the vowel is nasalized while the coda is only partially nasalized.

5. DISCUSSION

Results from this analysis suggest that the previously described nasal voicing contrast that alternates between heavy and light stems in coda position may be more accurately described as a timing contrast. Acoustic evidence presented in Figures 1–6 suggest that in the light stem forms, the gesture is shifting leftward into the vowel and out of the coda compared to heavy stem forms where the gesture occupies approximately half the vowel and the entire coda. These results are further supported by comparison with other Dene languages such as Ahtna [7] and Slave [8]. Ahtna has not developed phonemic nasal vowels. /n/ in codas is fully voiced, e.g., [tʰɛn] 'ice' (cf. Upper Tanana [tʰʌn̥]). This suggests that in Ahtna the nasal gesture minimally overlaps with the entire coda. Contrarily in Slave, which has lost most coda consonants, the nasal gesture has shifted completely leftward and initiates with the onset of the vowel leaving no overlap for a would-be coda (e.g., [tʰɛ̃] 'ice' [8, p. 42]). Upper Tanana light forms with nasal codas demonstrate a transitional system between languages like Ahtna and like Slave where the nasal gesture has shifted leftward but not to the point where it no longer overlaps with the coda. If the nasal gesture were to continue shifting leftward, we might predict a fully nasalized vowel along with an oral coda, similar to the development of /ʔīh/ 'see (ipfv.)' and /tʰāj/ 'trail' mentioned above, and to the loss of stem-final nasal consonants in syllables with long stem vowels. This trend, then, is yet another instance of coda consonant loss in Upper Tanana ([9], [10]). It is important to note that the timing of the nasal gesture in Upper Tanana light stems is not necessarily precise and might result in different sound combinations in the coda; especially during the transition phase between the nasal and oral segments while the tongue repositions to allow air to escape the mouth.

While the contrast between heavy and light stem [n] might be more accurately described as a timing contrast involving the nasal gesture, it may not be the best way to describe the contrast for language learners. Pedagogically, learners are already familiar with other voicing contrasts in Upper Tanana, specifically in the stop series. Therefore, applying this concept to nasal [n] is not new. However, explaining and understanding a timing gesture involving varying degrees of nasalization on both the verb and coda would be considerably more complicated for both instructors and learners alike. Additionally, the degree of endangerment of Upper Tanana makes it unlikely that the leftward shift of the nasal gesture will continue to the point where the entire vowel is nasalized and no nasal consonant remains. In that case, educators would have to make a decision as to whether they want to maintain a conservative orthography or make adjustments to the spelling system to reflect pronunciation.

6. CONCLUSION

Using the earbuds method, we have shown that what has been described as a voicing contrast between nasal consonants is acoustically a timing contrast, with the nasal gesture shifting leftwards. These findings provide yet another salient example of the multifaceted nature of phonetic cues that are described generically as 'voiceless nasals' (see e.g., [11], [12]). Future studies exploring the acoustic correlates of nasality in Ahtna and Slave will help to provide a more complete picture of these complexities in Dene languages.

7. ACKNOWLEDGMENTS

We want to thank Mrs. Ida Joe and the late Mrs. Sherry Demit-Barnes for being part of this study.

7. REFERENCES

[1] Stewart, J., Kohlberger M. 2017. Earbuds: A method for measuring nasality in the field, *Language Documentation and Conservation* 11, 49–80.

[2] Lovick, O. 2020. *A Grammar of Upper Tanana, vol. 1: Phonology, lexical classes, morphology*. Lincoln: University of Nebraska Press.

[3] Krauss, M. E., Leer, J. 1981. *Athabaskan, Eyak, and Tlingit sonorants*. in Native Language Center Research Papers No. 5. Fairbanks: Alaska Native Language Center.

[4] Samuel F.G., Iradj Sooudi D.D.S, Frost, S.D. 1974. Quantitative and graphic analysis of prosthetic treatment for 'nasalance' in speech. *The Journal of Prosthetic Dentistry* 32(3), 284–291.

[5] Ohala, M., Ohala, J. 1991.Nasal epenthesis in Hindi, *Phonetica*, 48, 207–220.

[6] Boersma, P., Weenink, D. 2016. Praat. Institute of Phonetic Sciences, University of Amsterdam, Amsterdam.

[7] Kari, J. 1990. *Ahtna Athabaskan Dictionary*. Fairbanks: Alaska Native Language Center

[8] Rice, K. 1989. *A grammar of Slave*, vol. 5. in Mouton Grammar Library, vol. 5. New York: Mouton de Gruyter.

[9] Leer, J. 1977. *Stem Syllable Development in Tanacross and Upper Tanana*, Fairbanks: Native Language Center (UT 974 L).

[10] Tuttle, S. G., Lovick, O. Núñez-Ortiz, I. 2011. Vowels of Upper Tanana Athabascan, *Journal of the International Phonetic Association* 41(3), 283–312.

[11] Chirkova, K., Basset, P., Amelot, A. 2018. Voiceless nasal sounds in three Tibeto-Burman languages, *Journal of the International Phonetic Association* 49(1), 1–32.

[12] Bhaskararao, P., Ladefoged, P. 1991. Two types of voiceless nasals, *Journal of the International Phonetic Association* 21(2), 80–88.

COTOPAXI MEDIA LENGUA VOWELS: A PRELIMINARY ANALYSIS

Jesse Stewart[1] and Sky Onosson[2]

[1]University of Saskatchewan, [2]University of Victoria
Stewart.Jesse@usask.ca, Sky@Onosson.com

ABSTRACT

Although Media Lengua was first documented by linguists in Ecuador's Cotopaxi province, this study represents the first phonetic account of Media Lengua vowel production in that region, as subsequent research on Media Lengua has focused exclusively on the variety spoken in Imbabura province. This preliminary case study reveals that Cotopaxi speakers differ from Imbabura speakers in how they accommodate the five vowels of Media Lengua's Spanish-origin vocabulary into Quichua's three-vowel system. Imbabura speakers have been previously shown to maintain significant but minimal acoustic differences between high and mid-vowels i.e., /i/ versus /e/, and /u/ versus /o/. In contrast, our results show that Cotopaxi speakers differ from this pattern in two ways: a greater distinction is maintained between front /i/ and /e/ as compared with Imbabura speakers, but there is merger of the back vowels /u/ and /o/, suggesting these two varieties may have other as yet unstudied differences.

Keywords: Media Lengua, Cotopaxi, vowels, acoustic phonetics, dialectal variation

1. INTRODUCTION

Media Lengua (ISO-693: mue) is an endangered 'mixed' language spoken in the Ecuadorian Andes by an estimated 2,904 people [1]. Media Lengua is generally described as having systematic divisions between lexicon and grammar where approximately 90% of the vocabulary is Spanish in origin and nearly the entire morphosyntactic system is Quichua in origin. Example (1) provides an example of a Media Lengua utterance where the bolded elements are of Spanish origin.

(1)

Orth	***Asi juga**shpaca **peligrosu**mi.*		
IPA	**asi**	**xuga**-ʃpa-ka	**peligrosu**-mi
Parse:	like.that	play-GEN-TOP	dangerous-VAL
Sp:	*Jugando así es peligroso.*		
Q:	*Shina pucllashpaca pilirusumi.*		
En:	Playing like that is dangerous.		

Media Lengua was originally documented in the province of Cotopaxi in the 1970's by Pieter Muysken [2]–[4] and later in the province of Imbabura in the mid 2000's [5]. Since its original documentation in the 1970's numerous recent attempts to find Media Lengua speakers in Cotopaxi have failed [6]–[8] leading researchers to believe that it had since gone dormant in the region. However, in 2022 new fieldwork revealed that Media Lengua is still spoken by an estimated 1,703 people in Cotopaxi [1].

The Media Lengua vowel system has been of particular interest to phoneticians due to the interplay between its two source vowel systems; Spanish with five vowels (/i, e, u, o, a/) and Quichua with three vowels (/i, u, a/). To date, all published phonetic research on the vowel system comes from the Imbabura dialect of Media Lengua [9]–[12]. Results reveal a stratified yet highly overlapping vowel system on the F1xF2 dimensions. Here, high and low vowels of both Spanish and Quichua origin are shown to be significantly different, although clusters only differ by 10 to 20 Hz, suggesting near-mergers [9]. Meanwhile, Spanish origin mid-vowels and high vowels, from both source languages, are significantly different with enough distance in acoustic space that they could potentially be contrastive aurally even though there is still a substantial degree of overlap [9]. The perception of mid- and high vowels was later tested using a 2-forced choice identification task experiment which revealed that Media Lengua listeners were indeed able to reliably identify differences different between /i/ versus /e/ and /u/ versus /o/ [10]. Figure 1 uses a bagplot to illustrate the F1xF2 dimensions of the Imbabura Media Lengua vowel system.

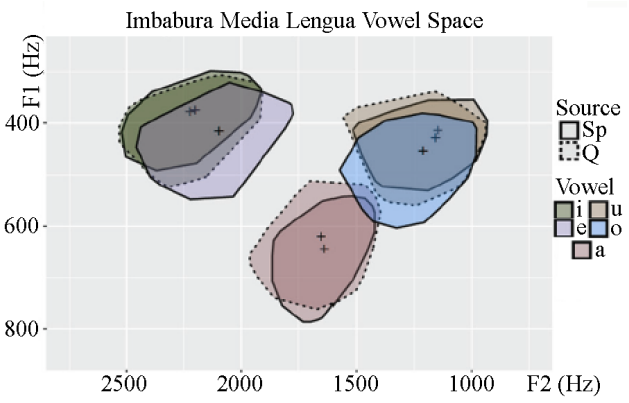


Figure 1: Imbabura Media Lengua vowel space based on [12].

Further analyses of Media Lengua's vowel system examined Spanish origin vowel sequences (diphthongs) [11]. Results revealed that Media Lengua speakers produce all Spanish origin vowel sequences but that these are adapted to Media Lengua's more compact vowel system (as compared to Spanish). Finally, it was shown that Media Lengua speakers make use of numerous cues beyond formant frequencies to identify the differences between mid and high vowels [12]. These primarily include speaker sex in addition to syllable structure and prosodic elements such as pitch.

Alongside these prior findings on the phonetic qualities of Imbabura Media Lengua vowels, this study provides a preliminary sketch of the Cotopaxi Media Lengua vowel system. F1xF2 measurements come by way of conversations from two women and is compared with previously analysed data from Imbabura Media Lengua. The goal of this study is twofold: (1) gain a general understanding of the shape of the Cotopaxi Media Lengua vowel system, especially with respect to the degree of overlap between mid- and high vowels and (2) infer how the vowel systems from these two speakers equate and/ or differ from the Imbabura Media Lengua vowel system.

2. METHOD

2.1. Participants

Given the recent and ongoing documentation of Cotopaxi Media Lengua, for the first time in approximately 40 years, we currently only have conversational data from two speakers. Both speakers are female, aged 29 and 37 respectively at the time of recording, and were born and raised in the community of Yacubamba, Cotopaxi, Ecuador.

For comparative purposes, previously analysed data from Imbabura Media Lengua was used [12]. These data come by way of five female speakers between the ages of 39 and 50 at the time of recording.

2.2. Data

F1 and F2 frequencies were extracted from the centre points of vowels. From the two Cotopaxi Media Lengua speakers, 516 tokens were extracted for analysis (Speaker 1 n = 257; Speaker 1 n = 259; /a/ n = 69, /e/ n = 136, /i/ n = 107, /o/ n = 141, /u/ n = 63). From the five Imbabura Media Lengua speakers, 420 tokens were extracted (Speaker 3 n = 59; Speaker 4 n = 42; Speaker 5 n = 96; Speaker 6 n = 36; Speaker 7 n = 191; /a/ n = 158, /e/ n = 56, /i/ n = 83, /o/ n = 90, /u/ n = 31).

2.3. Procedures

Data come from conversations recorded between Media Lengua speakers discussing daily life events. A Zoom H6 Handy Recorder was used to record the conversations in 16-bit WAV format with a sample rate of 44.1 kHz. Vowels were marked on a point tier in Praat [13] near the centre of the vowel where the F1 and F2 formants were at their most steady state. A Praat script was used to automatically extract the data points. The libraries *ggplot2* [14] and *LmerTest* [15] were used to analyse the data in R version 4.2.2 [16]. For each region, Cotopaxi and Imabura, linear mixed-effects regression (LMER) models were run on a per-formant basis with *Vowel* as fixed effect, and random intercepts for both *Speaker* and *Word*. For each model, the intercept was rotated between /i/ and /u/ to permit interpretation of high versus mid-vowel difference from the perspective of either the front or back regions of the vowel space. Additional cross-regional models were run for each formant by adding *Region* as a fixed effect, to examine cross-regional variation.

3. RESULTS

The LMER results for Cotopaxi F1 and F2 are shown in Table 1 (intercept = /i/) and Table 2 (intercept = /u/). Significant differences in one or both formants are sufficient to differentiate each vowel from all others in the system, although the formant differences between /u/ and /o/ (Table 2) are the least substantial, in particular for F1 where this difference is just above the significance threshold of p<0.05.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	418.77	**<0.001**	2428.39	**<0.001**
Vowel [a]	384.26	**<0.001**	-653.35	**<0.001**
Vowel [e]	107.74	**<0.001**	-105.28	**0.004**
Vowel [o]	174.29	**<0.001**	-948.40	**<0.001**
Vowel [u]	131.70	**<0.001**	-1052.62	**<0.001**

Table 1: Linear mixed-effects regression results for Cotopaxi Media Lengua F1 and F2, intercept = /i/.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	550.47	**<0.001**	1375.77	**<0.001**
Vowel [a]	252.56	**<0.001**	399.27	**<0.001**
Vowel [e]	-23.97	0.255	947.34	**<0.001**
Vowel [i]	-131.70	**<0.001**	1052.62	**<0.001**
Vowel [o]	42.58	0.050	104.22	**0.016**

Table 2: Linear mixed-effects regression results for Cotopaxi Media Lengua F1 and F2, intercept = /u/.

As in Cotopaxi, cross-vowel formant differences in Imbabura (Table 3 and Table 4) are significant for one or both formants, but are weaker in almost every case,

although notably the F2 difference between /u/ versus /o/ appears to be more robust than in Cotopaxi.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	423.06	**<0.001**	2357.12	**<0.001**
Vowel [a]	280.11	**<0.001**	-598.92	**<0.001**
Vowel [e]	34.88	**0.018**	-151.62	**<0.001**
Vowel [o]	79.83	**<0.001**	-891.46	**<0.001**
Vowel [u]	43.14	**0.015**	-1081.36	**<0.001**

Table 3: Linear mixed-effects regression results for Imbabura Media Lengua F1 and F2, intercept = /i/.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	466.20	**<0.001**	1275.76	**<0.001**
Vowel [a]	236.97	**<0.001**	482.43	**<0.001**
Vowel [e]	-8.26	0.666	929.74	**<0.001**
Vowel [i]	-43.14	**0.015**	1081.36	**<0.001**
Vowel [o]	36.69	**0.040**	189.90	**<0.001**

Table 4: Linear mixed-effects regression results for Imbabura Media Lengua F1 and F2, intercept = /u/.

The differences between regions are made clear via the cross-regional model (Table 5 and Table 6; to conserve space, only the interactions between *Vowel* and *Region* are displayed). From the perspective of Cotopaxi /i/ (Table 5), significant differences in F1 pertain to each Imbabura vowel, while from the perspective of /u/ (Table 6) this is only true relative to /i/; there are no significant cross-regional F2 differences at all.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
Vowel [a] × Region [Imbabura]	-102.87	**<0.001**	47.58	0.347
Vowel [e] × Region [Imbabura]	-73.08	**0.004**	-48.17	0.367
Vowel [o] × Region [Imbabura]	-93.60	**<0.001**	55.24	0.265
Vowel [u] × Region [Imbabura]	-88.44	**0.004**	-36.32	0.577

Table 5: Linear mixed-effects regression results for Media Lengua F1 and F2, intercept = Cotopaxi /i/.

	F1		F2	
Predictors	*Estimates*	*p*	*Estimates*	*p*
Vowel [a] × Region [Imbabura]	-14.43	0.637	83.91	0.195
Vowel [e] × Region [Imbabura]	15.35	0.626	-11.85	0.859
Vowel [i] × Region [Imbabura]	88.44	**0.004**	36.32	0.577
Vowel [o] × Region [Imbabura]	-5.17	0.865	91.56	0.154

Table 6: Linear mixed-effects regression results for Media Lengua F1 and F2, intercept = Cotopaxi /u/.

Visualization of the vowel distribution differences between regions is provided using two methods. Two-dimensional density estimates (Figure 2) portray the central region of each vowel's token distribution via opacity, with less dense regions being more transparent. Bag (or bag-and-loop) plots (Figure 3), being bivariate equivalents of box-and-whisker plots [17], show the central 50% of tokens in the opaque 'bag', with remaining tokens (excluding statistical outliers) contained in the larger, more transparent 'loop'.

Although these two methods are non-equivalent, they both illustrate the same pattern. Cotopaxi Media Lengua's front vowels are more distinct in the height dimension i.e., in F1, relative to Imbabura, while the back vowels are almost completely merged. In contrast, the Imbabura Media Lengua vowel space is more vertically compact with a large degree of overlap between high and mid-vowels, but with more front-to-back differentiation i.e., in F2.

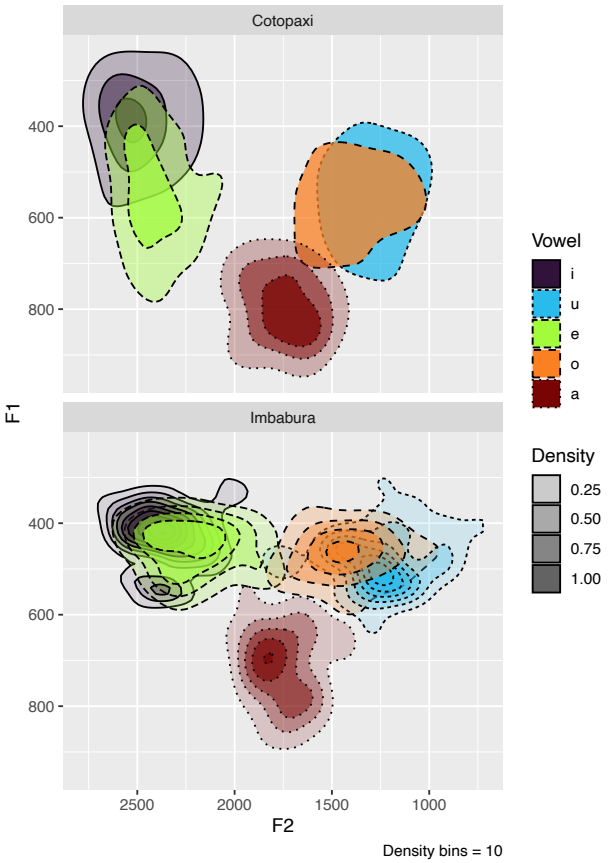


Figure 2: Density plots of Media Lengua vowels by region.

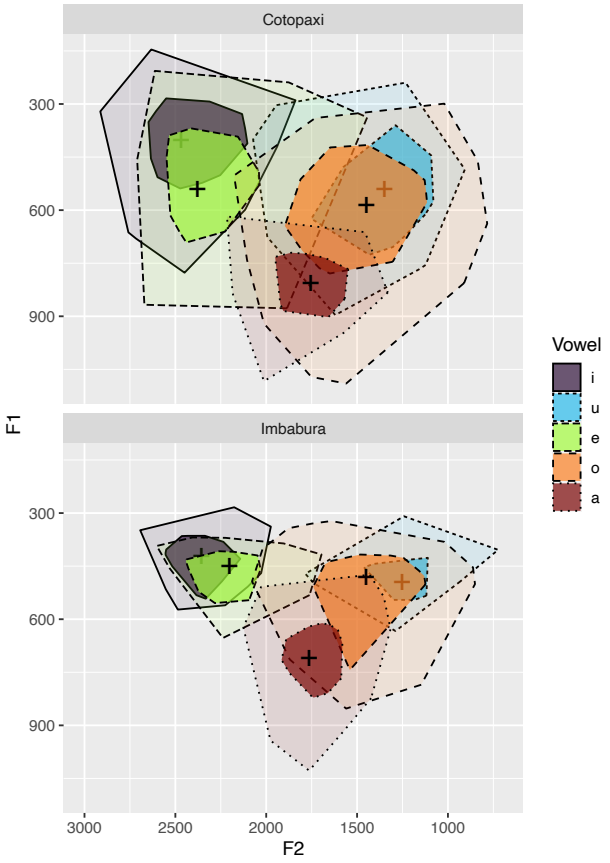


Figure 3: Bag plots of Media Lengua vowels by region.

4. DISCUSSION

Results from this study are still preliminary given the small sample size of speakers from Cotopaxi. Yet, they provide insights into how the Cotopaxi Media Lengua vowel system might differ from that of Imbabura.

Most noticeably, the back vowels in Cotopaxi show substantial overlap in F1xF2 acoustic space with no significant difference revealed in the statistical model (Table 2). Contrarily, Imbabura back vowels show significant, though not substantial, separation in F1 with /o/ being on average 37 Hz higher than /u/. Additionally, the back vowels from both Cotopaxi speakers are significantly lower in F1xF2 acoustic space compared to those from Imbabura suggesting that they have lowered and merged /u/ with /o/. This is substantiated by the statistical models which reveal that /u/ is significantly and substantially higher in F1 frequency compared to /i/ by, on average, 131 Hz and that /e/ and /o/ have similar average F1 frequencies, differing, non-significantly, by only 23 Hz (Table 2). Additionally, when comparing Cotopaxi /u/ to Imbabura /o/, the F1 averages are virtually identical only differing non-significantly, by 5 Hz.

Contrarily, both Cotopaxi speakers show significantly and substantially different F1 frequencies for /i/ and /e/, by on average, 107 Hz. While the average difference between /i/ and /e/ in Imbabura is also significant, the distance between the averages is only approximately a third of that in Cotopaxi /i/ and /e/ at 35 Hz.

These results suggest that these speakers of Cotopaxi Media Lengua are operating a four-vowel system with two front vowels (/i/ & /e/), a single mid-back vowel (/o/), and a low (/a/), while speakers of Imbabura are operating a five-vowel system, but with substantial overlap between mid and high vowel pairs (e.g., /i/ & /e/ and /u/ &/o/).

These results are interesting in that they may have implications for the origins of contact language phonology given the heterogeneity of the Cotopaxi and Imbabura vowel systems. Specifically, it appears that speakers of different dialects of the same mixed language have dealt with Spanish origin vowels in different ways. Both speakers from Cotopaxi seem to have placed substantially greater importance in differentiating /i/ from /e/ compared to Imbabura speakers, while at the same time the Cotopaxi speakers have almost completely ignored the differences between /u/ and /o/ while Imbabura speakers still maintain some degree of separation.

It is not yet known whether this is an effect of the underlying Quichua vowel systems from each region or innovations unique to Media Lengua. However, results from /o/ and /e/ in Spanish borrowings from Imbabura Quichua show a similar arrangement in acoustic space to that of Imbabura Media Lengua ([9]), therefore, it might stand to reason that /e/ and /o/ in Spanish borrowings in Cotopaxi Quichua might show clear separation between /i/ and /e/ while, /o/ and /u/ merge. Further investigation into the different phonetic realizations of the Quichua varieties spoken in each region may provide significant insight into how aspects of phonetics and phonology influence each other in the development of mixed languages.

Given the limitations inherent in a small sample size, we cannot at this stage make definitive statements regarding the differences between Media Lengua as spoken in Cotopaxi versus Imbabura, nor draw conclusions about how these differences arose. Nonetheless, the results described in this study suggest that regional differences do exist in Media Lengua, the extent and causes of which can only be guessed at presently.

5. REFERENCES

[1] Stewart, J., Gonza Inlago, L., Prado Ayala, G. Accepted. Cotopaxi Media Lengua is still very much alive. In: *Language Documentation and Conservation*.

[2] Muysken, P. 1997. Media Lengua. In: Thomason, S. (ed), *Contact languages: A Wider Perspective*. J. Benjamins Pub. Co., 365–426.

[3] Muysken, P. 1981. Halfway between Quechua and Spanish: The case for relexification. In: Highfield, R. (ed*), Historicity and variation in Creole studies*, vol. 57–78. Karoma Publishers.

[4] Muysken, P. 1979. La mezcla de quechua y castellano: El caso de la "media lengua" en el Ecuador. *Lexis* 3(1), 41–56.

[5] Gómez-Rendón, J. 2005. La Media Lengua de Imbabura. In: Muysken, P., Olbertz, H. (eds), *Encuentros y Conflictos: Bilingüismo y Contacto de Lenguas en el Mundo Andino*. Iberoamericana, 39–58.

[6] Stewart, J. 2011. *A Brief Descriptive Grammar of Pijal Media Lengua and an Acoustic Vowel Space Analysis of Pijal Media Lengua and Imbabura Quichua.* University of Manitoba MA Thesis.

[7] Lipski, J. 2019. Reconstructing the life-cycle of a mixed language: An exploration of Ecuadoran Media Lengua. *International Journal of Bilingualism*, 1–27.

[8] Müller, A. 2011. La media lengua en comunidades semi-rurales del Ecuador: uso y significado social de una lengua mixta bilingüe'. Universität Zürich PhD Dissertation.

[9] Stewart, J. 2014. A comparative analysis of Media Lengua and Quichua vowel production. *Phonetica* 7, 159–182.

[10] Stewart, J. 2018. Vowel perception by native Media Lengua, Quichua, and Spanish speakers. *Journal of Phonetics* 71, 177–193.

[11] Onosson, S., Stewart, J. 2021. The Effects of Language Contact on Non-Native Vowel Sequences in Lexical Borrowings: The Case of Media Lengua. *Language and Speech PaPE Special Issue*, 1–30.

[12] Onosson, S., Stewart, J. 2021. A multi-method approach to correlate identification in acoustic data: The case of Media Lengua. *Laboratory Phonology* 12(1), 1–30.

[13] Boersma, P. Weenink, D. 2016. *Praat*. Institute of Phonetic Sciences, University of Amsterdam, Amsterdam.

[14] Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

[15] Kuznetsova, A., Brockhoff, P. B., Bojesen, R. lmerTest. 2016. CRAN.

[16] R Development Core Team. 2017. *R*. R Foundation for Statistical Computing.

[17] Rousseeuw, P. J., Ruts, I., Tukey, J. W. 1999. The Bagplot: A Bivariate Boxplot. *The American Statistician* 53(4), 382–387.

THE PHONETIC BASIS FOR MODERN MONGOLIAN VOWEL HARMONY: AN ANALYIS BY MEANS OF REAL-TIME MRI

Yoshio Saitō[a], Kikuo Maekawa[b] & Yurong[c]

[a]Takushoku University; [b]National Institute for Japanese Language and Linguistics; [c]Inner Mongolia University
emailsaito@gmail.com; kikuo@ninjal.ac.jp; umyurong@yahoo.co.jp

ABSTRACT

Mongolian vowel harmony has been traditionally described by backness of the tongue as the main underlying principle, but it is difficult to apply it to the description of modern eastern Mongolian vowel harmony. Other features that have been proposed are the height of the tongue and the position of the tongue root. Recent studies suggest that the latter plays a more important role in grouping vowels into harmonic sets. However, further evidence is required to reveal the phonetic basis of modern eastern Mongolian vowel harmony.

We measured (1) the distance between the pharyngeal wall and the tongue root, (2) the larynx height, and (3) the tongue height of three native speakers using real-time MRI. The results show significant differences both in tongue height and in tongue root position between harmonizing sets of vowels in modern eastern Mongolian.

Keywords: Mongolian, vowel harmony, tongue root, ATR, tongue height, real-time MRI

1. INTRODUCTION

1.1. Vowel harmony

In languages with vowel harmony, the vowels are divided into groups, and a morphological unit can only have vowels belonging to the same group. Articulatory factors known to function as a basis for grouping vowels are palatality, labiality, degree of tongue height, and tongue root position.

1.2. Modern Mongolian vowel harmony

The modern Mongolian language we deal with here is the Bārin dialect, which is a variety of eastern Mongolian spoken in the south-eastern part of the Inner Mongolian Autonomous Region in China. The vowels investigated in this study are shown in Figure 1. (New vowels that emerged in a later period of history by assimilation, etc. were excluded. As the vowel *ï* has no long counterpart, the long *i* was also excluded.) These vowels are grouped into two categories in terms of vowel harmony, as shown in Figure 2. We employ the traditional Roman transcription system widely used in Western scholarship. The modern Bārin pronunciation is shown with IPA symbols in parentheses.

i [i] ***ü/ū*** [u(:)]
ï [ɪ] ***u/ū*** [ʊ(:)]
ö/ō [o(:)] ***e/ē*** [ɤ(:)]
o/ō [ɔ(:)]
a/ā [ɑ(:)]

Figure 1: Vowels investigated

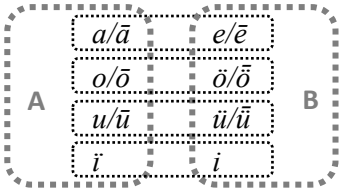


Figure 2: Four harmonic pairs and two harmonic sets

Mongolian vowel harmony has been traditionally described by backness of the tongue as the main underlying principle (cf. Poppe 1951), but as can be seen from the above two figures, it is difficult to apply the principle to the distinction of the two harmonic sets in modern eastern Mongolian from a phonetic point of view. Other articulatory phonetic features proposed for the distinction are the tongue height and tongue root position.

Chinggeltei & Sinedke (1959) first found that the difference in tongue root position plays an important role in the distinction of Mongolian vowels using X-ray radiography. Concerning Mongolia's Khalkha dialect, in which *a*, *o*, *ö*, *u*, *ü* and *i* have almost the same phonetic values as those in main dialects of Inner Mongolia, various views have been expressed based on acoustic data: Svantesson et al. (2005) describes that it has a pharyngeal harmony, and Jōo (2005) says both backness and degree of tongue height are involved in dividing vowels into harmonic sets.

Based on data of the vowels *a*, *e*, *o*, *ö*, *u* and *ü* obtained from a single speaker of the Bārin dialect using real-time MRI, Saitō et al. (2019) showed that the difference in the pharynx size is prominent between the two harmonic sets, and it could be the

major phonetic basis of modern Mongolian vowel harmony.

We have conducted a study of eight vowels differing in quality including *i* and *ï* of the same dialect using real-time MRI with three speakers including the one who participated in Saitō et al.'s (2019) work. See Figure 3 for MRI images of the vowels.

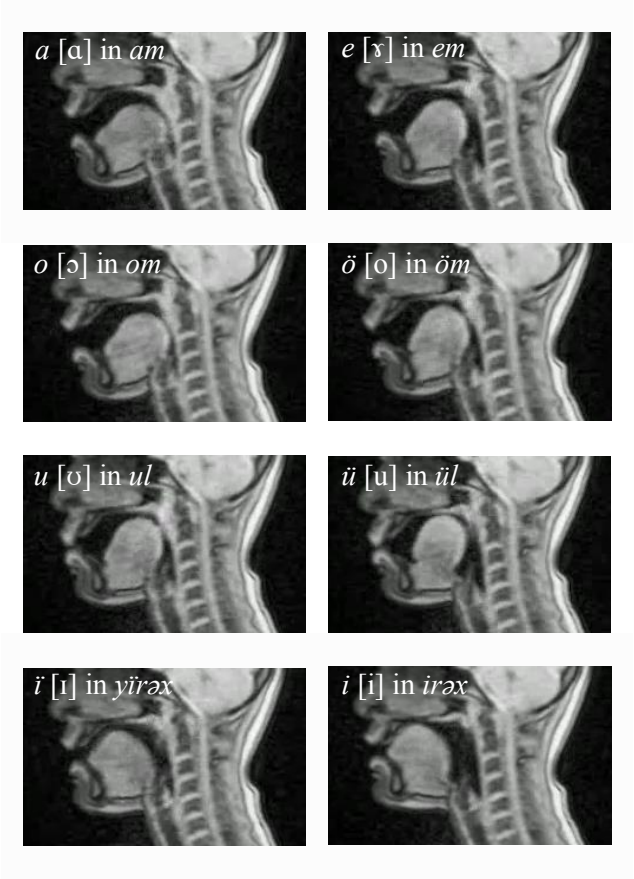


Figure 3: Sample MRI images of eight vowels differing in quality

2. METHOD

2.1. Data

MRI movies of 694 words each uttered by three female native speakers (S6, S22, and S23) of the Bārin dialect (194 words uttered three times each and the remaining 500 uttered once by each speaker) were recorded at BAIC (Brain Activity Imaging Centre) of ATR-Promotions, an affiliate company of ATR (Advanced Telecommunications Research Institute International) using the 3T MRI system (Siemens MAGNETOM Prisma fit 3T). Recordings of articulatory movements in the midsagittal plane were made in a FLASH sequence with acceleration factor 3. Spatial resolution was 256×256 pixels, and 1 pixel corresponds to 1 mm in length. Slice thickness was 10 mm, and temporal reconstruction rate was 14 frames per second.

The following vowels in minimal and quasi-minimal pairs of words were selected for the present study to achieve balanced data in the segmental environment:

a/e: ***a***m/***e***m, d***a***mjəx/d***e***mjəx, g***a***r/g***e***r, g***a***līg/g***e***rīg, x***a***ləx/x***e***ləx, x***a***ljē/x***e***ljē, x***a***rā/x***e***rē, x***a***rūl/x***e***rūl
ā/ē: ***ā***lj/***ē***lj, n***ā***x/n***ē***x, t***ā***/t***ē***, tūl***ā***/tūl***ē***, udl***ā***/üdl***ē***, xar***ā***/xer***ē***, yïr***ā***d/ir***ē***d
o/ö: ***o***m/***ö***m, t***o***səl/t***ö***səl, n***o***gō/n***ö***gō, ***o***rōd/***ö***rōd, x***o***r/x***ö***r
ō/ȭ: ***ō***/***ȭ***, ***ō***č/***ȭ***č, t***ō***/t***ȭ***, t***ō***l/t***ȭ***l, nog***ō***/nög***ȭ***, or***ō***d/ör***ȭ***d
u/ü: s***u***m/s***ü***m, s***u***rəg/s***ü***rəg, ***u***dlā/***ü***dlē, ***u***jūr/***ü***jūr, ***u***l/***ü***l, ***u***rəx/***ü***rəx, ***u***stai/***ü***stei
ū/ǖ: d***ū***/d***ǖ***, ***ū***r/***ǖ***r, n***ū***r/n***ǖ***r, t***ū***lā/t***ǖ***lē, ***ū***dəm/***ǖ***dəx, uj***ū***r/üj***ǖ***r, bïl***ū***/bil***ǖ***, xar***ū***l/xer***ǖ***l
ï/i: bï***l***ū̲/bi***l***ū̲, x***ï***r/x***i***r, y***ï***r/***i***r, y***ï***rād/***i***rēd, y***ï***rəx/***i***rəx

The vowels investigated were short vowels in initial syllables and long vowels in initial and second syllables. Short vowels in non-initial syllables, which are all reduced in modern Mongolian, were excluded. A frame from each movie at a point with vowel articulation least affected by adjacent consonants was chosen for measurements.

2.2. Measurements

Fig. 4 shows the measurement points and normalized xy-axes used in this study. Measurements of x- and y-coordinate values were made for each vowel at the black points outlined in white using an image processing programme, *ImageJ* (Ferreira & Rasband 2012). No measurements were made for points that were hard to identify due to unclear images.

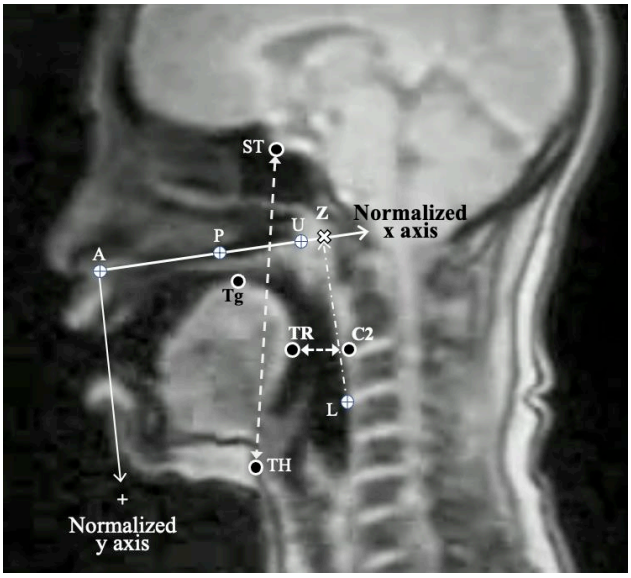


Figure 4: Measurement points and normalization method

We measured:

(a) The distance between the point on the inner lower left corner of the second cervical vertebra (C2) and its opposite point across the pharynx on the surface of the tongue root (TR) with the same y-axis value;

(b) The distance between the corner of the sella turcica (ST) and the top of the convex part of the sternohyoid muscle, which is made by being pressed forward by the thyroid cartilage (TH); and

(c) The xy-values of the closest point on the surface of the tongue to the palate (Tg).

A point on the pharyngeal wall was avoided because the surface of the wall is often blurred by saliva gathering on it.

The top of the convex part of the sternohyoid muscle was chosen to measure larynx height because the position of the vocal chords was less clear on our MRI images.

The "backness" and "height" of vowels in articulatory phonetics are not defined by physiological criteria alone, but to see the differences in shape of the oral cavity, we measured the closest point of the tongue to the palate, which we call the highest point of the tongue for brevity, for each vowel.

The obtained xy-values were normalized. A and P stand for the anterior and posterior nasal spines, respectively. U is the point where the line connecting A and P reaches the pharyngeal wall. L is the point on the pharyngeal wall corresponding to the boundary between the third and fourth cervical vertebrae. Z is the point where the line starting from point L intersects perpendicularly with the normalized x-axis.

The origin of the normalized coordinate is on point A. The normalized x-axis is the line connecting the points A, P, and U. The y-axis is perpendicular to the x-axis. The unit of measurement is different for x- and y-axes, and the units for the x- and y-axes are the distance between A and U, and between L and Z, respectively.

3. RESULTS

Table 1 summarizes the results of a series of Welch's two-sample t-tests, and Table 3 compares the means of A and B vowels. Note that we conducted only t-tests, assuming that there is no correlation among variables.

	Vowels		Distance		Highest point of the tongue	
		A vs. B (Number of tokens)	C2 - TR	ST - TH	Y-axis values	X-axis values
Pairs	Short	**a** (68)	**e** (68)	***	*	***	*
		o (38)	**ö** (32)	***	n.s.	***	n.s.
		u (56)	**ü** (51)	***	n.s.	***	***
		ï (44)	**i** (37)	***	*	***	***
	Long	**ā** (57)	**ē** (57)	***	n.s.	***	***
		ō (44)	**ō̈** (46)	***	n.s.	***	n.s.
		u (60)	**ṻ** (58)	***	n.s.	***	***
	Both	**a, ā** (125)	**e, ē** (125)	***	*	***	***
		o, ō (82)	**ö, ō̈** (78)	***	n.s.	***	n.s.
		u, ū (116)	**ü, ṻ** (109)	***	n.s.	***	***
Sets	All	A (367)	B (349)	***	***	***	n.s.
	All except *ï-i*	A (323)	B (312)	***	*	***	***

Significance levels * <.05, ** <.01, *** <.001

Table 1: Summary of the statistical tests

3.1. Tongue root position

The distance between the pharyngeal wall and the tongue root is significantly different at the significance level of .001 in all pairs and sets. The tongue root is retracted in Set A vowels, and is advanced in Set B vowels (see Figure 2 for the harmonic sets).

3.2. Larynx height

Because the results of the test on the larynx height (ST-TH distance) in Table 1 were hard to interpret, additional tests were conducted for each subject and vowel. The results are summarized in Tables 2 and 5. It is likely that the relevance of larynx height differs depending on subjects and vowels.

Speaker	Vowels A vs. B (Number of tokens)		ST-TH distance
s6	**a, ā** (42)	**e, ē** (43)	***
	o, ō (28)	**ö, ō̈** (29)	n.s.
	u, ū (39)	**ü, ṻ** (39)	***
	ï (15)	**i** (15)	***
s22	**a, ā** (43)	**e, ē** (43)	n.s.
	o, ō (25)	**ö, ō̈** (22)	n.s.
	u, ū (39)	**ü, ṻ** (38)	n.s.
	ï (14)	**i** (10)	n.s.
s23	**a, ā** (40)	**e, ē** (39)	n.s.
	o, ō (29)	**ö, ō̈** (27)	n.s.
	u, ū (38)	**ü, ṻ** (32)	n.s.
	ï (15)	**i** (12)	*

Significance levels * <.05, ** <.01, *** <.001

Table 2: Summary of the t-test on the larynx height

Vowels			Distance			
			C2-TR		ST-TH	
	A vs. B		Mean A	Mean B	Mean A	Mean B
Short	**a**	**e**	0.162	0.246	1.566	1.610
	o	**ö**	0.134	0.207	1.561	1.552
	u	**ü**	0.159	0.241	1.524	1.524
	ï	**i**	0.294	0.383	1.559	1.629
Long	**ā**	**ē**	0.154	0.210	1.589	1.629
	ō	**ō̈**	0.134	0.183	1.576	1.576
	ū	**ṻ**	0.172	0.253	1.534	1.567
Both	**a, ā**	**e, ē**	0.158	0.230	1.577	1.619
	o, ō	**ö, ō̈**	0.134	0.193	1.569	1.566
	u, ū	**ü, ṻ**	0.166	0.247	1.529	1.547
All	A	B	0.172	0.243	1.558	1.585
	A except *ï-i*	B	0.155	0.227	1.558	1.580

Table 3: Mean C2-TR and ST-TH distances

Vowels			Highest point of the tongue			
			Y-axis		X-axis	
	A vs. B		Mean A	Mean B	Mean A	Mean B
Short	**a**	**e**	0.205	0.146	0.632	0.648
	o	**ö**	0.213	0.156	0.704	0.701
	u	**ü**	0.175	0.118	0.712	0.658
	ï	**i**	0.118	0.094	0.421	0.394
Long	**ā**	**ē**	0.218	0.169	0.651	0.711
	ō	**ō̈**	0.214	0.158	0.729	0.733
	ū	**ṻ**	0.165	0.109	0.721	0.656
Both	**a, ā**	**e, ē**	0.211	0.157	0.641	0.677
	o, ō	**ö, ō̈**	0.213	0.157	0.718	0.720
	u, ū	**ü, ṻ**	0.170	0.113	0.717	0.657
All	A	B	0.187	0.137	0.656	0.650
	A except *ï-i*	B	0.197	0.142	0.688	0.681

Table 4: Mean vowel height

Speaker	Vowels		ST-TH distance	
	A vs. B		Mean A	Mean B
S6	**a, ā**	**e, ē**	1.413	1.505
	o, ō	**ö, ō̈**	1.447	1.422
	u, ū	**ü, ṻ**	1.369	1.422
	ï	**i**	1.438	1.564
S22	**a, ā**	**e, ē**	1.612	1.629
	o, ō	**ö, ō̈**	1.602	1.630
	u, ū	**ü, ṻ**	1.560	1.555
	ï	**i**	1.586	1.602
S23	**a, ā**	**e, ē**	1.708	1.729
	o, ō	**ö, ō̈**	1.657	1.663
	u, ū	**ü, ṻ**	1.655	1.678
	ï	**i**	1.654	1.735

Table 5: Mean ST-TH distance per subject

3.3. Tongue height

The members of each harmonic pair are significantly different in tongue height at the significance level of .001 in all pairs and sets in terms of y-axis value. Set B vowels generally show higher values than their counterparts in each pair (see Table 4).

As for the x-axis values, which concern the backness of the tongue, the difference between members of the pairs *o-ö*, *ō-ō̈*, *o/ō-ö/ō̈*, and A-B is not significant (see Tables 1 and 4).

4. DISCUSSION AND CONCLUSION

Our analysis confirmed the following: (1) Both the distance between the pharyngeal wall and the tongue root (TR-C2) and the degree of tongue height (Tg) are stable indices for grouping vowels into harmonic sets. (2) Backness of the tongue and the degree of larynx height are unstable as an index.

The above facts indicate that modern eastern Mongolian vowel harmony can be described with the features [ATR] and [high]. However, as the root and dorsum of the tongue may move in tandem to a certain degree, we must investigate if an inevitable cause-and-effect relationship exists between advancement of the tongue root and raising of the dorsum of the tongue. To solve this problem, we are starting to conduct perception tests using vowels synthesized from accurate 3D measurements of the vocal tract. Kaji et al. (2022) completed a small preliminary study using vowels generated by modifying vocal tract area functions. They have shown that modification of area functions of the pharynx or the oral cavity alone is insufficient for correct perception of vowels. However, there is a need to investigate the problem further with plenty of reliable data.

Because of some phonological changes, some additional vowels with different phonetic values have emerged in modern dialects. Their phonetic characteristics and phonological behaviours are different from the other vowels in terms of vowel harmony. We should also take such new vowels into consideration. Furthermore, we need to investigate characteristics of short vowels in non-initial syllables to complete the articulatory description of modern Mongolian vowel harmony.

ACKNOWLEDGEMENTS

This study was supported by a grant from the Japan Society for the Promotion of Science [Grant-in-Aid for Scientific Research (B), Project No. 20H01265, Principal investigator: Kikuo Maekawa)].

We thank Tatsuya Kitamura, Kiyoshi Honda, Yukiko Nōta, and Hironori Takemoto for their advice and help, Takuya Asai for development of an analysis and annotation system for real-time MRI movies, and anonymous reviewers for their comments and suggestions.

5. REFERENCES

[1] Chinggeltei & Sinedke. 1959. Monggol kelenü ündüsün egesigüdün tukai [On the basic vowels of the Mongolian language]. *Öbör monggolun yeke surgaguli erdem sinjilegenü sedkül* [*Journal of Inner Mongolia University*], 97–114. Reprinted in: *Chinggelteiyin jokiyal bütügelün tegübüri* 2 [*Collected works of Chinggeltei* Vol. 2]. 2010. Chìfēng: Öbör monggolun sinjilekü ukagan teknik mergejilün keblelün koriya, 125–146.

[2] Ferreira, T. & W. Rasband 2012. *ImageJ User Guide* https://imagej.nih.gov/ij/docs/guide/index.html

[3] Jōo, H. 2005. *Mongorugo boinchōwa no kenkyū* [*Studies on Mongolian vowel harmony*]. Tōkyō: Bensei-shuppan.

[4] Kaji, Y. et al. 2022. Seidō danmenseki kansū no sōsa ni yoru Mongorugo boin on'in tokuchō no kentō [Investigation of acoustic properties of Mongolian vowels by modification of vocal tract area functions]. *Proceedings of the Acoustical Society of Japan*, 997–998.

[5] Poppe, N. 1951. *Khalkha-mongolische Grammatik*. Wiesbaden: Franz Steiner Verlag.

[6] Saitō, Y. et al. 2019. An investigation into modern Mongolian vowel harmony using real-time magnetic resonance imaging. S. Calhoun et al. eds. *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne.

[7] Svantesson, J.-O. et al. 2005. *The phonology of Mongolian*. Oxford: Oxford University Press.

THE ACOUSTICS OF VOWEL HARMONY IN SAKHA

May Pik Yu Chan, Jianjing Kuang

University of Pennsylvania
pikyu@sas.upenn.edu, kuangj@sas.upenn.edu

ABSTRACT

Sakha (Yakut) is a Turkic language that is typologically rare to have both backness and roundness vowel harmony. This study aims to investigate the phonetic realizations of vowel harmony conditioned by vowel features (e.g., backness and roundness), monophthong or diphthong, as well as the distance of the target syllable from the trigger syllable. One native speaker of Sakha produced inflected or conjugated lexical items that were up to five syllables in length in a carrier phrase. Results show that F2 is the most important acoustic cue involved in both backness harmony and roundness harmony. And there is a salient effect of syllable position for vowel harmony: target vowels are gradually centralized when further away from the trigger; and this position-conditioned gradient vowel harmony effect is further modulated by vowel quality features: gradient backness harmony is observed for both monophthongs and diphthongs, but gradient roundness harmony is only observed for diphthongs.

Keywords: Yakut, Sakha, Vowel Harmony, Turkic language

1. INTRODUCTION

Sakha (or Yakut) is a North Siberian Turkic language spoken in Sakha/Yakutia. The eight monophthongs (/i, e, y, œ, ɯ, ɑ, u, o/) in Sakha differ in [+/-high], [+/-back] and [+/-round] dimensions respectively, with /i, ɯ, e, ɑ/ being unrounded counterparts to /y, u, œ, o/, and /i, ɯ, y, u/ being high counterparts to /e, ɑ, œ, o/. There are also four diphthongs (/ie, ɯɑ, yœ, uo/), which contrast in backness and roundess. As an important part of the morphosyntactic process, vowel harmony operates from left to right within an agglutinated word. Vowels consistently harmonize in backness with the word's initial vowel. Rounding harmony applies for all [-high, +round] triggers, but [+high, +round] vowels and diphthongs may only spread their roundness feature to [+high] vowels. There has been limited work looking at the phonetic realization of Sakha vowel harmony, except for [1] showing that Yakut-Russian bilinguals have robust vowel harmony adaptations in Russian word production.

One point of interest in studying vowel harmony is the comparison between alternating and non-alternating vowels. While vowel assimilation is found to be categorical and complete for some cases [2, 3, 4], gradient assimilatory effects between vowels are also fairly common [5, 6, 7]. For example, some other Turkic languages (e.g. Kazakh, Uyghur) have demonstrated centralization in vowel backness harmony patterns [5, 6]. This study focuses on describing gradient vowel harmony in Sakha as Sakha offers two extra dimensions of interest:

First, Sakha has both backness and roundness harmony. Unlike backness harmony, which is common across Turkic languages, systematic rounding harmony is much rarer typologically and often occurs with extremely limited contexts (see [5] for a detailed review). Subsequently, prior work on Turkic vowel harmony has focused on backness harmony only [5]. In comparison, Sakha's rounding harmony rules are less restrictive. In Sakha, low round vowels (/o/ and /œ/) may spread their roundness feature to both high and low vowels, but high round vowels (/y/ and /u/) may only spread roundness to following high vowels. The presence of both backness and relatively less restrictive roundness harmony allows us to explore whether gradient harmony effects are feature specific. The relatively freer rounding harmony rule also means that the entire vowel inventory could be examined, unlike languages like Uzbek, Kazakh, and Uyghur, where non-high round vowels may only occur in initial syllables [5].

Second, diphthongs (/ie, ɯɑ, yœ, uo/) also participate in harmony in Sakha. Specifically, diphthongs as triggers pattern like high vowels in Sakha. Words like /duoraɪ/ 'to make an echo' exemplifies this, since the second syllable in the word is unrounded. This is consistent with height restrictions of rounding vowel harmony in Sakha. Furthermore, verbal suffixes in Sakha also contain diphthongs that are underlying only specified in height but not backness and roundness (e.g. /bɑr/ 'to go'; /bɑrɯɑm/ 'go-1SG.FUT'). This allows us

to test whether diphthongs follow the same vowel harmony mechanisms.

The goal of this study is to offer a better understanding of the phonetic realizations of Sakha vowel harmony, as instrumental studies have not been done. We aim to capture the differences between alternating and non-alternating vowels in Sakha vowel harmony addressing the following: (i) whether there is gradient realization of harmony conditioned by syllable position in Sakha (ii) whether backness and roundness harmony follow the same harmony pattern and (iii) whether monophthongs and diphthongs pattern similarly.

2. METHODS

2.1. Procedure

One native speaker of Sakha in her mid-twenties participated in the study. Apart from Sakha, she is also fluent in English and Russian, and has beginner-level Italian and Japanese.

The speaker was given lists of lexical items containing nouns and verbs. The speaker learned to associate English translations of sentences with corresponding affixes. For instance, when given the English translation 'their leathers', she would produce [tirii-lere] 'leather-3pl.poss' instead of the full phrase [kiniler tiriilere] 'their leathers'. Similarly for verbs, when given the translation 'I will smile right now', she would produce only the conjugated target word [yœr-yœm] 'smile-1sg.fut', and when given the translation 'I will smile tomorrow', she would produce [yœr-yœʁym] 'smile-1sg.fut'. The speaker was asked to produce each target word in a carrier phrase [biligin _ dien tɯlɯ et] 'Say the word _ now', such that the target word was in a focused sentence medial position, avoiding phrase final prosody. This also avoids effects that may favor an interpolation-based account for harmony. The word domain edge would not align with the utterance edge, which may or may not be featurally specified in backness based on default articulatory settings. The carrier phrase and the root of the target word is presented in Cyrillic, and the full set of target word forms were presented in sets as English translations on the same slide as the target root word and carrier phrase. Each word was repeated three times.

Recording sessions were conducted online via Zoom, with the speaker recording the audio locally to her computer to avoid issues with data bandwidth filters. The data were recorded to the default microphone of a MacBook Air 2017 at a sampling rate of 32kHz and 32-bit float. Recordings were made over ten sessions; each lasted between 10 to 20 minutes.

2.2. Stimuli

The participant was presented with a set of target words containing all vowel quality contrasts in Sakha. This includes all eight monophthongs (/i, ɯ, e, ɑ, y, u, œ, o/) and four diphthongs (/ie, ɯɑ, yœ, uo/). Although Sakha does exhibit vowel length contrasts, words containing a short vowel as the initial syllable were chosen if possible. The comparative morpheme (-tAAʁA) also includes a long vowel. Two monosyllabic roots and two disyllabic roots per vowel were included for nouns, apart from the vowel /œ/, which had an extra disyllabic root token. There were 49 noun roots in total. For verbs, one monosyllabic and one disyllabic root were included per vowel, resulting in 24 verbal roots. This adds up to 73 roots in total. To the extent possible, monosyllabic roots and the first syllable of disyllabic roots ended in either a sibilant, liquid, or rhotic, with exceptions for coda-less lexical items, or those ending with a stop or /h/.

For nouns, seven forms were elicited: (i) singular, (ii) plural, (iii) first singular possessive, (iv) third plural possessive, (v) plural with a comparative, (vi) accusative, (vii) plural with an accusative marker. There is backness harmony that targets all non-initial vowels. This results in the following four alternations in backness: i∼ɯ, y∼u, e∼œ, ɑ∼o. With respect to roundness, since rounding harmony is parasitic on height, there are relatively fewer high-rounded vowels compared to unrounded vowels in the elicited forms.

		/kus/ 'duck'
a. SG	-Ø	kus
b. PL	-LAr	kustɑr
c. SG-1POSS	-(I)m	kuhum
d. PL-3POSS	-LAr-A	kustɑrɑ
e. PL-COMPAR	-LAr-TAAʁAr	kustɑrdɑɑʁɑr
f. ACC	-(n)I	kuhu
g. PL-ACC	-LAr-I	kustɑrɯ

Table 1: Examples of elicited nominal forms

Nine forms of the verbs were elicited: (i) the root in isolation (often corresponding to the imperative form), (ii) two forms of the first singular future, corresponding to an immediate or distant future, (iii) two forms of the third singular future, also corresponding to an immediate or distant future, (iv) first plural future, (v) third plural future, (vi) third singular progressive, (vii) third singular past.

As diphthongs consist of a high vowel followed by a low vowel of the same backness and roundness features, and they are also subject to backness harmony therefore, there are two alternating pairs in terms of backness (ie∼uɑ, yœ∼uo). We elicited fewer rounded than unrounded tokens for the same reason of parasitic rounding harmony.

		/bɑr/ 'to go'
a. root	-Ø	bɑr
b. 1SG.FUT	-IAm	bɑrɯɑm
c. 1SG.FUT	-IAʁIm	bɑrɯɑʁɯm
d. 3SG.FUT	-IA	bɑrɯɑ
e. 3SG.FUT	-IAʁA	bɑrɯɑʁɑ
f. 1PL.FUT	-IAxpIt	bɑrɯɑxpɯt
g. 3PL.FUT	-IAxtArA	bɑrɯɑxtɑrɑ
h. 3SG.PROG	-Ar	bɑrɑr
i. 3SG.PST	-PItA	bɑrbɯtɑ

Table 2: Examples of elicited verbal forms

Some complications to the word list include that when disyllabic verbs end in a coda /s/, the preceding vowel is deleted. For example, [umus] 'to dive' has a first singular future form of [umsuom] or [umsuoʁum] and not *[umusuom] nor *[umusuoʁum]. Two verbs have word final /s/. One verb [ɯarɯj] 'to become sick' has irregular conjugations and was excluded from the analysis. For monosyllabic roots, words are maximally four syllables in length, for disyllabic roots, words are maximally five syllables in length. This is the case for both nouns and verbs.

2.3. Segmentation and Measurement

Sound files were segmented manually in Praat [8]. The beginning and end of each vowel were segmented to the onset and offset of high-frequency energy. In cases where the vowel is adjacent to a sonorous consonant like laterals, rhotics and glides, the point of amplitude change was used to separate the vowel and the consonant. In cases where the token has noticeable background noise, the entire token (at the word level, not vowel level) was removed. After removing unwanted tokens, a total of 4948 vowel tokens entered the analysis. Among these tokens 3977 were monophthongs (short and long vowels included), 971 were diphthongs.

The first three formants as well as vowel duration were measured. Measurements were taken at 25, 50, and 75% of the vowel's entire duration using a Praat script. The "To Formant (burg)" function was used. The data were analyzed in R. For monophthongs, formants measured at 50% of the vowel duration were used for analyses. For diphthongs, formants measured at 25% and 75% of the vowel duration were analyzed. Since only one speaker's production was analyzed, no normalization was done.

3. RESULTS AND DISCUSSIONS

We first begin by exploring the overall vowel space in Sakha. The left facet of figure 1 shows the F1 and F2 of all monophthongs (short and long vowels included) measured at 50% of the vowel's total duration. It could be observed that /ɑ/ is located at the central portion of the vowel plot, suggesting that the [back] property of /ɑ/ is more phonological than phonetic. The facet on the right shows a corresponding F2-F3 plot. Generally, vowels differing in roundness do not differ much from the F3 dimension. Instead, vowel pairs contrasting in roundness differ mostly on the F2 dimension. Therefore, subsequent analysis will focus on F2.

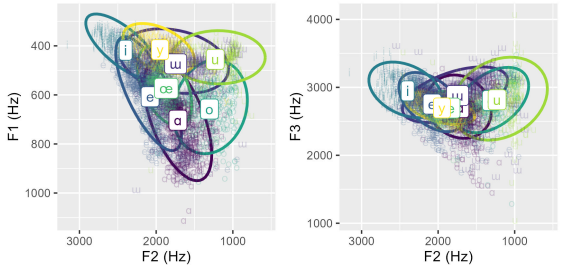


Figure 1: F1-F2 and F2-F3 vowel space plot.

We now turn to the effect of syllable position on vowel realization in vowel harmony. A linear mixed effects regression model (lme4 package [9]) was fitted to examine how vowel harmony is conditioned by syllable position, as well as the vowel features (backness and roundness) of the trigger and target vowels. Statistical significance is determined by Satterthwaite approximation via the lmerTest package [10]. Specifically, we predict F2 based on the fixed effects of initial vowel backness (which would be identical to target vowel backness), target vowel height, syllable position counting from the trigger, initial vowel roundness (which may differ from target vowel roundness), and whether the target vowel matches in roundness with the trigger (henceforth "roundness match"). The following two-way interactions are included: initial vowel backness and vowel height, initial vowel backness and syllable position, vowel height, and syllable position, initial vowel roundness and syllable position, and roundness match and syllable position. Three-way interaction effects of initial

vowel backness, vowel height, and syllable position are also included in the model. Random intercepts of the preceding consonant, the following consonant and the target vowel are included, and by-target vowel random slopes for syllable number are also included. Results show significant main effects of initial vowel backness (Reference = Front vowels; β = -1084.95, SE = 153.65, df = 3.110, p = 0.005), of initial vowel roundness (Reference = Unround vowels; β = -495.24, SE = 108.86, df = 3.134, p = 0.018), and of roundness match (Reference = Different; β = -369.56, SE = 112.12, df = 3.52, p = 0.036), confirming that both backness and roundness harmony modulates F2. Crucially, there is a significant interaction effect of initial vowel backness and syllable position (β = 183.24, SE = 37.85, df = 3.3, p = 0.013), suggesting that the later the syllable, the higher the F2 for back vowels. In other words, back vowels are fronted/centralized depending on syllable position. This effect can be clearly seen in Figure 2, which summarizes the F2 of vowels by position and backness harmony pairing. Both front and back vowels shift towards neutralization. However, there is no significant interaction between initial vowel roundness and syllable position, suggesting that the gradient neutralization effect is specific to backness for monophthongs.

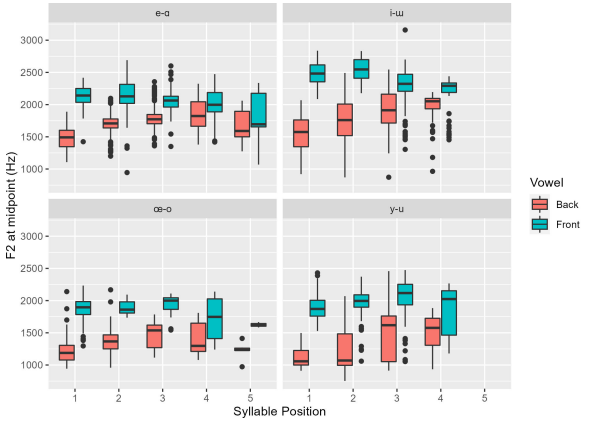


Figure 2: F2 at 50% by backness pairs (monophthongs).

Moving onto diphthongs, figure 3 visualizes F2 by syllable position and backness harmonic pairings in diphthongs. Facets correspond to F2 measurements taken at 25% and 75% of the vowel's total duration. Effects of F2 centralization appear to be generally present as well.

Two linear regression models are run to evaluate the effects of syllable position, initial vowel backness, and vowel roundness on F2 measured

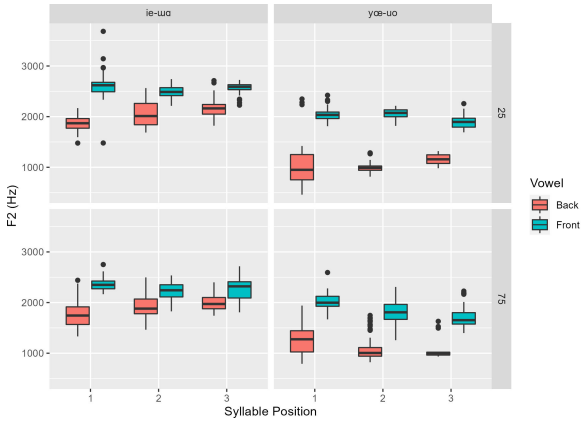


Figure 3: F2 at 25 and 75% of the vowel's duration by backness pairs (diphthongs).

at either 25% or 75% of the vowel's duration. For both models, the main effects of initial vowel backness, syllable position, and target vowel roundness are included. Two-way interaction effects between syllable position and initial vowel backness, and between syllable number and target vowel roundness are also included. Random intercepts by target-vowel are incorporated into the model. Due to a considerably lesser number of diphthongs in the dataset, other main and random effects are not included to ensure model convergence. Reference levels are identical to the monophthong models. Results find significant interactions of initial vowel backness and syllable position (25%: β = 140.36, SE = 16.10, df = 964.15, p < 2e-16; 75%: β = 111.254, SE = 17.74, df = 964.30, p = 5.37e-10) and initial vowel roundness and syllable number (25%: -63.12, SE = 16.33, df = 964.00, p = 0.0001; 75%: β = -162.1, SE = 17.99, df = 964.00, p < 2e-16) for both models. Therefore, for diphthongs, gradient centralization is conditioned by both backness and roundness.

4. CONCLUSIONS

This study aims to provide a better understanding of the phonetic realizations of Sakha vowel harmony. We show that F2 is the most important acoustic cue involved in both backness harmony and roundness harmony. And there is a salient effect of syllable position for vowel harmony: target vowels are gradually centralized when further away from the trigger; and this vowel harmony is further modulated by vowel quality features: gradient backness harmony is observed for both monophthongs and diphthongs, but gradient roundness harmony is only observed for diphthongs.

5. REFERENCES

[1] L. Vasilyeva, "Production and perception of vowel harmony: Phonological predictors of ratings and on-line adaptations of russian vowels in Yakut (Sakha)," PhD dissertation, University of Alberta, 2017.

[2] S. Hess, "Assimilatory effects in a vowel harmony system: an acoustic analysis of advanced tongue root in Akan," *Journal of Phonetics*, vol. 20, no. 4, pp. 475–492, 1992.

[3] B. Gick, D. Pulleyblank, F. Campbell, and N. Mutaka, "Low vowels and transparency in Kinande vowel harmony," *Phonology*, vol. 23, no. 1, pp. 1–20, 2006.

[4] S. G. Guion, M. W. Post, and D. L. Payne, "Phonetic correlates of tongue root vowel contrasts in Maa," *Journal of Phonetics*, vol. 32, no. 4, pp. 517–542, 2004.

[5] A. McCollum, "Gradience and locality in phonology: Case studies from Turkic vowel harmony," PhD dissertation, University of California, San Diego, 2019.

[6] ——, "Gradient morphophonology: Evidence from Uyghur vowel harmony," in *Proceedings of the Annual Meetings on Phonology*, vol. 6, 2019.

[7] ——, "Vowel harmony and positional variation in Kyrgyz," *Laboratory Phonology*, vol. 11, no. 1, 2020.

[8] P. Boersma, "Praat: doing phonetics by computer," http://www.praat.org/, 2006.

[9] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[10] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

PHONETIC VARIABILITY IN THE REALIZATION OF GLOTTALIZED STOPS IN USPANTEKO (MAYAN)

Ryan Bennett[a], Robert Henderson[b], Meg Harvey[b]

University of California, Santa Cruz[a] and University of Arizona[b]
rbennett@ucsc.edu, rhenderson@arizona.edu, mharvey3@arizona.edu

ABSTRACT

Glottalized *b' q'* often vary between ejective and implosive realizations in Mayan languages. Some of this variation has been attributed to phonetic differences between languages or dialects, but most descriptions are compatible with the alternative view that individual speakers of the same language/dialect differ in their inclinations to produce these sounds as ejective or implosive. We investigate the phonetic realization of *b' q'* in Uspanteko, a Guatemalan Mayan language. We find that labial *b'* is produced as an implosive more often than uvular *q'*; that word-final position favors ejective realizations; and that speakers of the same dialect differ in their tendencies to produce *b' q'* as implosive vs. ejective.

Keywords: ejectives, implosives, Mayan languages, allophony, inter-speaker variation

1. ALLOPHONIC REALIZATIONS OF GLOTTALIZED STOPS IN MAYAN

All Mayan languages contrast plain voiceless stops and affricates like /p t k t͡s t͡ʃ/ with glottalized counterparts at the same place of articulation [1].

(1) Glottalization contrasts in Uspanteko [2–4]
a. *ka'n* /kaʔn/ 'animal'
b. *k'a'n* /kʔaʔn/ 'angry'

The phonetic realization of these glottalized stops varies by language and place of articulation, among other factors. Bilabial *b'* is typically an implosive [ɓ] or [ɓ̥], though in some languages it may be preglottalized [ˀb]. The canonical realization of other glottalized plosives tends to be ejective (1).

However, glottalized stops in Mayan languages sometimes show extensive phonetic variability. Ejective [pʔ] allophones are commonly reported for glottalized *b'*, especially in word-final position [1, 4]. Variation between ejective and implosive realizations is often reported for uvular *q'* = [qʔ] ~ [ɢ̥], and to a lesser extent coronal *t'* = [tʔ] ~ [ɗ].[1]

Descriptive sources often report that this variation is conditioned by segmental context [1, 5, 6]. For example, [2] states that word-final *b'* in Uspanteko is realized as an unreleased implosive [ɓ̚], or an ejective [pʔ]; it is otherwise described as a voiceless implosive. [6] found that one speaker of Q'anjob'al produced *b'* as ejective [pʔ] at word edges, and as implosive [ɓ] word-medially, especially before [a].

Even so, in at least some Mayan languages, individual speakers may produce glottalized *b' q'* as either implosive or ejective in exactly the same environments, and even in the same words [1, 4–7]. Additionally, [1, 4] note that speakers seem to differ in their tendencies to produce ejective or implosive variants of these sounds. So while the segmental context may bias *b' q'* toward ejective vs. implosive realizations, it can also leave room for variability.

The phonetic variability observed for *b' q'* seems to involve a choice between qualitatively distinct allophones. Ejectives are produced with laryngeal raising, and implosives with laryngeal lowering (even if slight); these two vertical movements require the contraction of different muscles [5, 6, 8–12]. In that sense, the varied realization of *b' q'* as ejective vs. implosive implicates variation between categorically distinct phonetic outputs (cf. [6]).

In important early work addressing the phonetics of *b' q'* in Mayan with instrumental methods, [5] characterizes this variability in terms of differences between languages and/or dialects (based on data from the K'ichean languages Q'eqchi', Poqomchi', Kaqchikel, K'iche', and Tz'utujil). However, [5] presents results from just one speaker of each language/dialect, and no quantitative analysis. Consequently, it is unclear whether the phonetic variation observed by [5] reflects systematic differences between speech communities, or simply differences between individuals (as [5] herself notes; see also [7]). The same point can be made for [6, 13], who each report aerodynamic data from one speaker of Q'anjob'al. [7] is the only study of ejective vs. implosive variation in a Mayan language (Mam) which combines instrumental methods, quantitative analysis, and data from multiple speakers; still, [7] does not investigate contextual factors. Here, we provide a novel quantitative analysis of contextual

and inter-speaker variability between ejectives and implosives in another Mayan language, Uspanteko.

Data from Uspanteko could be useful for distinguishing the effects of dialect and language on the realization of *b' q'* from simple inter-speaker variation. Uspanteko is a fairly small language, with up to 6000 speakers; in comparison, there are over 70,000 speakers of Tz'utujil, over 150,000 speakers of Q'anjob'al, over 400,000 speakers of Kaqchikel, and over 1 million speakers of both K'iche' and Q'eqchi' (according to the most recent Guatemalan census; `https://www.censopoblacion.gt/`). Uspanteko is spoken in a compact region, and dialect variation does not occur on the same scale found for larger Mayan languages. Salient dialect differences do exist between the hamlet of Las Pacayas and other locations [2,4,14,15], but none of the speakers reported on here are from Las Pacayas.

In this paper we investigate variability in the realization of *b' k' q'* in a wordlist corpus [4]. We code each token of these sounds as either ejective or implosive based on their acoustic properties, and investigate what factors might influence this categorical variation in their phonetic form. We find that labial *b'* is implosive more often than uvular *q'*, and that speakers differ in their tendencies to produce ejectives vs. implosives. We also confirm that word-final position favors ejective realizations.

2. DATA COLLECTION

Audio recordings were collected from 9 native speakers of Uspanteko in 2018 (3 male; 23-50 years old, mean 35). Eight speakers were from the town of San Miguel Uspantán, and one from the nearby village of La Lagunita. The speakers translated target words written on index cards in Spanish to Uspanteko. The words were produced in the frame sentence [ˈjaχ ˈtekˀ ___ tiχ.ˈɓiχ] 'Diego says ___'. Most target items were 1-2 syllables long, but some were longer (3-5 syllables). Recordings were made in a quiet room with a headset microphone (Audio-Technica ATM73a) and solid-state recorder (Zoom H5) at a 48 kHz sampling rate. See [4] for details.

3. DATA ANNOTATION

Ejective vs. implosive realizations of *b' q'* and velar *k'* were annotated using TextGrids in Praat [16]. Implosives can be acoustically distinguished from ejectives and plain stops by the lack of a clear, salient release burst, particularly at CV transitions [4,5,7,9,17,18]. Though a release burst is sometimes present for implosives, it is typically very weak, and sometimes appears on the waveform as an initial negative pressure impulse at release, possibly reflecting ingressive airflow. Implosives, unlike ejectives, may also be produced with voicing during closure, especially just before or after the release of the oral constriction.

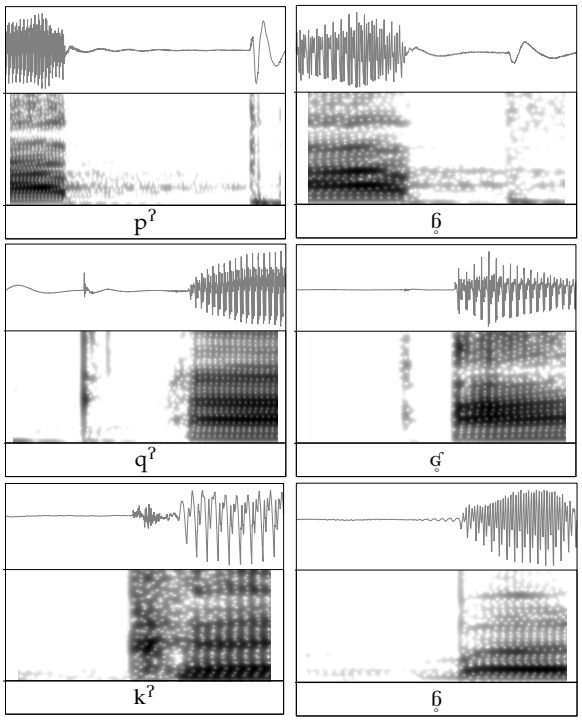


Figure 1: Some coding decisions for *b' k' q'*. Top row: /aC#/ context; bottom rows: /Ca/ context.

We thus annotated according to the following algorithm: (i) code as implosive if the release burst is very weak or absent, or if there is a negative impulse at release (Fig. 1, right); (ii) as ejective if there is a clear, strong release burst accompanied by a positive impulse on the waveform (Fig. 1, left); and (iii) if there is only a weak positive burst at release, look for other diagnostics, primarily voicing during closure or just prior to the oral release (Fig. 1, bottom right).

While most tokens were unambiguously ejective or implosive, some were difficult to annotate. Ejective tokens sometimes seemed to be 'slack', with weaker bursts [6, 7, 17, 19–21]. These could be difficult to distinguish from implosives, and so we had to rely on criteria other than the salience of the release burst to make coding decisions (e.g. voicing prior to release, direction of the initial impulse at release, etc.). Ambiguous tokens were most common with labial *b'*, likely because labials have weak release bursts in general, and so often lack the strong burst associated with ejectives [22].

We observed 14 tokens that appeared to be plain (non-glottalized) stops, and 28 unreleased stops in word-final position that could not be reliably

classified. We excluded these tokens from analysis.

Each annotation was checked multiple times, by at least two co-authors, to ensure that annotation guidelines were consistently followed. We did not compute any measures of inter-annotator agreement, but after joint review, it was clear our classification practices were quite consistent.

4. STATISTICAL ANALYSIS

Statistical analyses were carried out using R [23]. This dataset includes 312 ejective and 225 implosive tokens (Fig. 2). As expected, glottalized *b'* was mostly commonly produced as an implosive, and *q'* most commonly produced as an ejective, though ejective and implosive tokens did occur frequently for both sounds. Also as expected, velar *k'* was strictly realized as ejective, and never implosive except for one anomalous token.

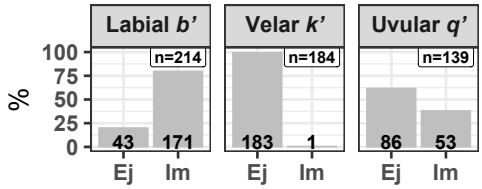
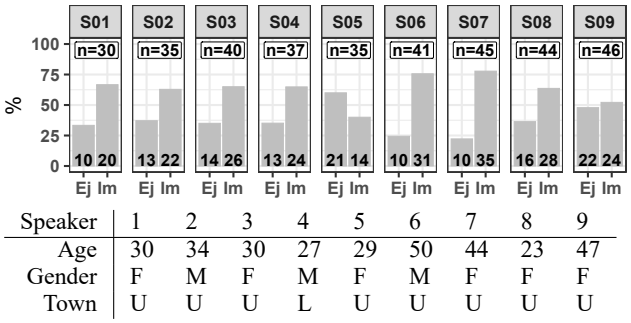


Figure 2: *b' k' q'* tokens by allophone. 'Ej' = ejective, 'Im' = implosive. Numbers in bars indicate token counts.

Since velar *k'* shows essentially no allophonic variability, it is not informative about the conditions favoring implosive vs. ejective variants of glottalized stops. Indeed, the effect of PLACE = VELAR cannot be estimated in the logistic regression presented below, because *k'* = [k^{ʔ}] is basically invariant. We thus exclude velar *k'* from further discussion. This leaves 129 ejective and 224 implosive tokens of *b' q'* (= 353 total) for analysis.

Speakers apparently differ in their propensities to produce ejectives vs. implosives for *b' q'* (Fig. 3). For example, speaker 7 seems to favor implosives, while speaker 5 seems to favor ejectives; both speakers are women from San Miguel Uspantán, though they differ in age.

We coded the segmental environment in terms of any neighboring consonants (= C), word-boundaries (= #), and vowels (= V), distinguishing stressed from unstressed vowels (= [±str], Fig. 4; stress may be final or penultimate in Uspanteko, and is predictable [4]). We treated lexical consonant clusters (e.g. *xq'aax*) and clusters produced by vowel deletion (e.g. *xáj(a)b'*) as equivalent (cf. [24]). With this coding, we observe more ejectives following stressed vowels and in word-final position, but no other obvious contextual effects.



Speaker	1	2	3	4	5	6	7	8	9
Age	30	34	30	27	29	50	44	23	47
Gender	F	M	F	M	F	M	F	F	F
Town	U	U	U	L	U	U	U	U	U

Figure 3: *b' q'* tokens by speaker.

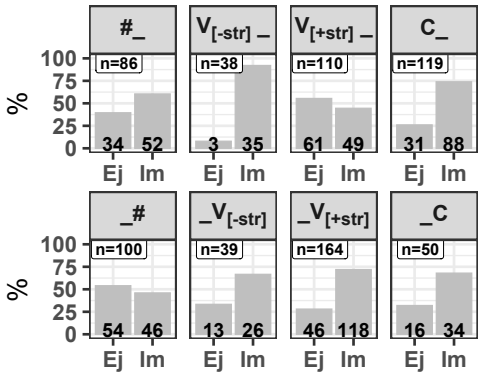


Figure 4: *b' q'* tokens by context. '_' = position.

The distribution of implosive and ejective variants was modeled with a mixed-effects binomial logistic regression (2) using the `glmer` function in `lme4` [25]. We included SPEAKER as a fixed effect rather than a random effect because we are interested in inter-speaker variation in ejective vs. implosive productions of glottalized stops, which is more straightforwardly assessed if SPEAKER is treated as a fixed effect. A random intercept for WORD was also included. All predictors were sum coded to reduce collinearity between simple predictors and interactions; this means that each level of a variable is compared to the grand mean across all levels of that variable. We chose speaker 4 as the reference level for SPEAKER because the estimated effect for speaker 4 was close to the overall sample mean. Fixed effects included SPEAKER, C PLACE, PRECEDING SEGMENT TYPE, and FOLLOWING SEGMENT TYPE, along with a two-way interaction between SPEAKER and C PLACE (2a). Models including other interactions involving C PLACE failed to converge, likely because some environments don't include enough tokens of both *b'* and *q'* to estimate an interaction (e.g. *b'* occurs in the context $V_{[-str]}\_$ 35 times, and *q'* just 3 times). The same is true for additional interactions involving SPEAKER. This model was then simplified using the `stepAIC` function in the `MASS` package [26]. The final model (2b) has reasonably low collinearity for a model

which includes interaction terms (VIF $\in [1.49, 2.92]$). To save space, we only report statistically significant levels of each predictor in the final model (Tab. 1).

(2) a. **Initial model**: PLOSIVE TYPE ~ C PLACE * SPEAKER + PRECEDING SEGMENT TYPE + FOLLOWING SEGMENT TYPE

b. **Final model**: PLOSIVE TYPE ~ C PLACE + SPEAKER + PRECEDING SEGMENT TYPE

	$\hat{\beta}$	Odds ratio ($e^{\hat{\beta}}$)	SE($\hat{\beta}$)	p-value
(Intercept)	0.95	2.60	0.25	$< .001$
LABIAL	1.30	3.68	0.19	$< .001$
SPEAKER 5	-1.51	0.22	0.46	$< .01$
SPEAKER 7	0.78	2.19	0.39	$< .05$
SPEAKER 9	-0.92	0.40	0.38	$< .05$
$V_{[-str]\_}$	1.09	2.96	0.55	$< .05$
$V_{[+str]\_}$	-1.54	0.21	0.29	$< .001$
SPEAKER 2 × LABIAL	1.25	1.39	3.49	$< .05$
SPEAKER 6 × LABIAL	-0.80	1.80	0.45	$< .05$
SPEAKER 7 × LABIAL	-0.80	2.19	0.45	$< .05$
SPEAKER 8 × LABIAL	-0.83	0.98	0.44	$< .05$

Table 1: Significant predictors in final logistic regression model. Odds ratios > 1 indicate higher likelihood of implosive, < 1 lower likelihood. $\hat{\beta}$ is relative to mean across all levels of the predictor.

Labial *b'* is realized as implosive more often than uvular *q'* is (Tab. 1). This finding parallels well-known typological tendencies, grounded in aerodynamic and articulatory factors [8, 22, 27]. In particular, labial stops have a larger oral cavity, and so it is more difficult to generate the increased intra-oral pressure needed to produce the high intensity 'pop' associated with ejective release. This may favor implosive realizations of *b'* across all contexts.

The segmental environment also plays a role: ejectives are more common after stressed vowels. This finding may actually indicate conditioning by word boundaries, which are highly correlated with stress in our data: most stops in $V_{[+str]\_}$ contexts are word-final (82/110 = 75%), while no word-final stops occur in the environment $V_{[-str]\_}$ (stress is often word-final [2, 4], and many target words were monosyllabic). In this dataset, *b'* is ejective in 42% of word-final tokens (26/62), and just 11% elsewhere (17/152); *q'* is ejective in 74% of word-final tokens (28/38), and 57% elsewhere (43/101) (see also Fig. 4). This is consistent with reports that word-final position favors ejective realizations of *b' q'* in K'ichean languages like Uspanteko [1, 2]. We speculate that word-final position favors ejectives because their releases are more salient, and so word-final stops at different places of articulation may be easier to distinguish when ejective [28]. (Word-final stops were also prosodically phrase-final in our study, due to our use of a frame sentence.)

Most of our significant effects concern inter-speaker variation. Compared to the overall sample of speakers, Speaker 2 is more likely to produce *b'* as implosive [ɓ̥]. Speakers 6 and 8 are instead more likely to produce *b'* as ejective [pˀ]. Speakers 5 and 9 favor ejectives for both *b'* and *q'*, when compared to the overall mean. Speaker 7 favors ejective realizations of *q'*, but not *b'*.

The source of this inter-speaker variation is not apparent. Some speakers who pattern together in their use of ejective vs. implosive variants of *b' q'* have different demographic characteristics (e.g. Speakers 6 and 8; Fig. 3). Notably, dialect variation is not a plausible explanation for these inter-speaker differences: our speakers were almost exclusively from one town, San Miguel Uspantán, and so geographically-based dialect variation cannot be the source of the observed variation [7, 14, 15]. We leave open the possibility that other sociolinguistic factors condition these patterns of phonetic variability.

5. CONCLUSION

While variation between implosive and ejective realizations of *b' q'* is sensitive to place of articulation and segmental context, a substantial component of this variation comes down to seemingly idiosyncratic differences between speakers. Our results are consistent with the view that glottalized consonants in Mayan languages are united by the presence of a laryngeal constriction of some kind, rather than by implosive or ejective articulations as such [1, 5]. Further, implosive *b'* was more common than implosive *q'*, in keeping with claims that the most canonical allophone of *b'* in K'ichean languages is implosive [1, 4]. If implosive and ejective productions of *b' q'* reflect qualitatively distinct articulatory maneuvers, variation between these sounds amounts to a variable allophonic process with categorical, rather than continuous, phonetic outcomes (cf. [6, 24]).

Coding glottalized stops as ejective or implosive on the basis of their acoustic properties can be challenging, particularly for *b'*. But the fact that it is sometimes difficult to distinguish ejective vs. implosive allophones of *b' q'* may help explain why we find so much intra- and inter-speaker variation in the production of these sounds in the first place: their acoustic similarity may faciliate their phonological equivalence.

6. REFERENCES

[1] R. Bennett, "Mayan phonology," *Language and Linguistics Compass*, vol. 10, no. 10, pp. 469–514, 2016.

[2] T. Can Pixabaj, *Gramática descriptiva Uspanteka*. Antigua, Guatemala: Oxlajuuj Keej Maya' Ajtz'iib' (OKMA), 2007.

[3] M. A. Vicente Méndez, *Diccionario bilingüe uspanteko-español*. Antigua, Guatemala: Oxlajuuj Keej Maya' Ajtz'iib' (OKMA), 2007.

[4] R. Bennett, R. Henderson, M. Harvey, and T. Méndez López, "The phonetics and phonology of Uspanteko (Mayan)," *Language and Linguistics Compass*, vol. e12467, 2022. [Online]. Available: https://dx.doi.org/10.1111/lnc3.12467

[5] S. Pinkerton, "Quichean (Mayan) glottalized and nonglottalized stops: a phonetic study with implications for phonological universals," in *Experimental phonology*, J. Ohala and J. Jaeger, Eds. Orlando: Academic Press, 1986, pp. 125–139.

[6] J. Kuang, "Contextual variation of glottalic stops in Q'anjob'al," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 1114–1118. [Online]. Available: https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1163.pdf

[7] S. Russell, "Some acoustic characteristics of word initial pulmonic and glottalic stops in Mam," Master's thesis, Simon Fraser University, 1997.

[8] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Malden, MA: Blackwell, 1996.

[9] G. Clements and S. Osu, "Explosives, implosives, and nonexplosives: the linguistic function of air pressure differences in stops," in *Papers in Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 299–350.

[10] J. Kingston, "The phonetics and phonology of the timing of oral and glottal events," Ph.D. dissertation, University of California, Berkeley, 1984.

[11] M. Oh, D. Byrd, L. Goldstein, and S. S. Narayanan, "Enriching the understanding of glottalic consonant production: Vertical larynx movement in Hausa ejectives and implosives," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1940–1941, 2018.

[12] C. Ludlow, "Central nervous system control of the laryngeal muscles in humans," *Respiratory physiology & neurobiology*, vol. 147, no. 2-3, pp. 205–222, 2005.

[13] R. Shosted, "Towards a glottalic theory of Mayan," in *New Perspectives in Mayan Linguistics*, H. Avelino, J. Coon, and E. Norcliffe, Eds. Cambridge, MA: MIT Working Papers in Linguistics, 2009, vol. 59, pp. 80–113.

[14] L. Campbell, *Quichean linguistic prehistory*, ser. University of California Publications in Linguistics. Berkeley, CA: University of California Press, 1977.

[15] J. A. Us Maldonado, *Monografía Uspanteka*. San Miguel Uspantán: Comunidad Lingüística Uspanteka, no date.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.05)," Computer program, 2020, retrieved from http://www.praat.org/.

[17] M. Lindau, "Phonetic differences in glottalic consonants," *Journal of Phonetics*, vol. 12, pp. 147–155, 1984.

[18] C. Henton, P. Ladefoged, and I. Maddieson, "Stops in the world's languages," *Phonetica*, vol. 49, no. 2, pp. 65–101, 1992.

[19] J. Kingston, "The phonetics of Athabaskan tonogenesis," in *Athabaskan Prosody*, S. Hargus and K. Rice, Eds. Amsterdam: John Benjamins, 2005, pp. 137–184.

[20] R. Wright, S. Hargus, and K. Davis, "On the categorization of ejectives: data from Witsuwit'en," *Journal of the International Phonetic Association*, vol. 32, no. 1, pp. 43–77, 2002.

[21] N. Warner, "Acoustic characteristics of ejectives in Ingush," in *4th International Conference on Spoken Language Processing*. New York: IEEE, 1996, pp. 1525–1528.

[22] H. Javkin, "Towards a phonetic explanation for universal preferences in implosives and ejectives," in *Berkeley Linguistic Society (BLS) 3*. Berkeley, CA: University of California, Berkeley Linguistics Society, 1977, pp. 559–565.

[23] R Development Core Team, *R*, Computer program, R Foundation for Statistical Computing, Vienna, Austria, 2020, version 4.0.2, retrieved from http://www.R-project.org/.

[24] R. Bennett, R. Henderson, and M. Harvey, "Vowel deletion as phonologically-controlled gestural overlap in Uspanteko," *Language*, to appear.

[25] D. Bates, M. Maechler, and B. Bolker, *lme4*, R package, 2020, version 1.1-23, retrieved from http://CRAN.R-project.org/package=lme4.

[26] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, 2002.

[27] L. Campbell, "On glottalic consonants," *International Journal of American Linguistics*, vol. 39, no. 1, pp. 44–46, 1973.

[28] J. Ohala, "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman, Eds. Cambridge, UK: Cambridge University Press, 1990, pp. 258–275.

[1] Some Mayan languages have a *p' b'* contrast, realized phonetically as a contrast between ejective [pʔ] and preglottalized or implosive [ʔb ɓ]. Uvular consonants are limited to Eastern Mayan and Q'anjob'alan languages.

Phonetics-prosody interplay is implicated in typologically rare and complex morphophonology in Tohono O'odham

Daejin Kim & Robert Cruz

Department of Linguistics, University of New Mexico
daejinkim@unm.edu, rsanta10@unm.edu

ABSTRACT

This paper explores the role of focus prominence in producing nouns with morphophonology in Tohono O'odham. Plural nouns ($C_1V_1\underline{C_2V_2}C$) are formed by reduplicating the first consonant and vowel of singular nouns ($\underline{C_1V_1}C$). Each CV receives recurring strong-weak stress, signaled with *f0* and duration. We examined how speakers produce words with the potential conflict between the C_2V_2's weak stress, indicating plurality, and focus prominence, correcting lexical and grammatical information. We find that C_1V_1 has longer duration and higher *f0* than C_2V_2. With contrastive focus, speakers increased *f0* more at C_1V_1 and decreased it more at C_2V_2 while maintaining duration. Our finding implies that the language operates both stress and mora as phonological timing units (i.e., mora-sensitive stress). We suggest that speakers' meticulous control of phonetic properties indicating prosody underlies typologically complex and rare morphophonology in Tohono O'odham.

Keywords: Tohono O'odham, CV reduplication, metrical stress, intonational prominence, mora-sensitive stress

1 INTRODUCTION

Tohono O'odham [1]–[5] is a Uto-Aztecan language spoken in South-Central Arizona (near Tucson, Arizona) in the United States and Northern Sonora in Mexico. The language is actively spoken in the community (15,000 English-bilingual speakers in 2007) but has been considered endangered [6]. The language's typologically rare CV reduplication and metrical stress phonology have been widely discussed in [7]–[15]. However, there have been no perspicuous accounts explaining how the interacting morpho-phonology is linguistically structured. We instead explored the morphophonology of the language differently by examining how speakers employ intonational (focus) prominence in producing words. This allowed us to examine how the morphophonology of the language is phonetically realized.

1.1 CV Reduplication

Previous studies [12]–[14] reported that Tohono O'odham marks plural agreements of nouns, adjectives, and verbs by copying the first CV sequence of a singular noun (e.g., *ban* 'a coyote', *gogs* 'a dog'), which is the base, and pasting the copied sequence after the base, which becomes the reduplicant of a plural noun (e.g., *baban* 'coyotes', *gogogs* 'dogs'). This process is called (full) CV reduplication [16]. Note that the language never permits the first syllable with the coda to be the base; only the first CV (or C) is permitted to be the base[i]. We argue that comparing the production of singular and plural nouns provides a good basis for understanding the morphophonology of the language. This is because the base and the reduplicant are equivalent morphophonological units with major grammatical functions, but these morphemes vary due to the metrical stress of the language.

1.2 Strong-weak metrical stress

In Tohono O'odham, strong-weak (SW) metrical stress occurs across words and phrases [7]–[11]. The stress pattern of the language has been considered trochaic with a left-headed foot structure [10]–[13]. The primary strong stress is always applied to the first vowel of a word (e.g., *kóji* 'a pig', *nówiyù* 'a cow'). The second vowel receives weak stress (or gets unstressed), and then the secondary strong stress is applied at the third vowel of words (e.g., *kókojì 'pigs'*, *nónowìyu* 'cows'). Word stress in this language is primarily realized with varying fundamental frequency (*f0*) and duration of segments (**Figure 1**) [17], as what other studies similarly found in various

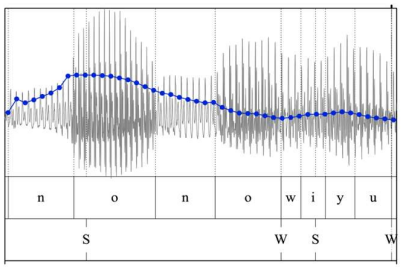


Figure 1: An *f0* contour of *nónowìyu* 'cows' with a strong-weak metrical stress in Tohono O'odham (from our recorded data).

languages. Because the language permits a CV to receive stress, the language has been claimed to have a *mora* [12], [13]. However, the metrical stress of the language has been regarded as phonologically complex because the stress pattern varies according to the status of a word, whether it is derived or not [11], and the interface with its morphology [10], [14]. In this study, we first introduce the phonetically analyzed data of Tohono O'odham, which consists of mini conversational dialogues with various words and focus prominence conditions, so we expect to gain better insight into certain aspects of Tohono O'odham's metrical stress phonology and morphology.

1.3 Intonational prominence in Tohono O'odham

Speakers actively employ prominence [18], [19] in their intonation to meet the communicative needs in conversation, such as presenting new meanings to listeners and implying speakers' intent. By comparing one element being prominent and distinctive to another, several studies cited in [18], [19] have found elements that contribute to organizing the phonetic and phonological structure of languages. Additionally, since word-level stress co-occurs with post-lexical prominence, it is necessary to distinguish stress behavior at two different levels to better understand the stress system of languages [20]. In Tohono O'odham, the realization of word-level stress should be different from that in languages previously reported because of its potential conflict with morphemes that serve a grammatical function, especially when a word needs to be contrasted lexically and grammatically in context.

As discussed in the previous section, the first CV, the base of singular and plural nouns, is subject to the assignment of strong stress. The second CV, the reduplicant of plural nouns, is subject to the assignment of weak stress. Speakers should produce the second CV prominently in their intonation when contrasting words lexically and grammatically by increasing (or decreasing) *f0* and duration, but it is under the influence of weak stress. It is unclear how language speakers produce a weakened CV sequence when it needs to be prominent due to focus.

- **Q1**: How do they produce the reduplicant (the second CV) with weak stress of plural nouns when contrasted with singular nouns?

To test this potential conflict, we compare the production of words with three types of focus prominence – broad focus (BF), narrow focus (NF), and contrastive focus (CF) [21] – which can also be implemented in Tohono O'odham. BF provides listeners with the whole phrase and sentence as new information in context, and NF narrows the range of attention to a word or a phrase as new information. CF, which has a stronger effect on articulating words and segments than BF and NF, provides a basis for correcting old information from questions (see **Table 2** for sentences designed for deriving each focus condition). In Tohono O'odham, CF can be implemented lexically (LCF) (e.g., *dog* vs. *cat*) as well as morphologically (MCF) (e.g., *dog* vs. *dogs*), which contrasts the plurality of the same lexical item. We, therefore, hypothesize that words with LCF and MCF are produced more prominently (*higher* in *f0* and *longer* in segmental duration) than those with BF and NF, and, therefore, words with MCF may have longer reduplicants in duration than those with LCF.

- **Q2**: How do speakers differentiate word prosody when lexically and morphologically contrasted in their intonation?

Taking these questions together, this study examined how speakers of Tohono O'odham produce singular and plural nouns in various focus prominence conditions, compared duration and *f0* of CVs, and explored how speakers produce a word with the potential conflict between metrical stress and focus prominence.

2 METHODS

We used eight common words (animals) as test words (**Table 1**), which consist of oral and nasal stop consonants (C = [p, t, k (fortis), b, d, g (lenis), m, n]) and vowels (V = [i, ɑ, o, u]) [1]-[4]. Test words have one to four CVs with or without a coda consonant. All plural nouns have more than two CVs. The dialogue (**Table 2**) was designed to make words be produced with various focus types. Words with LCF are contrasted with other lexical items in questions (e.g., *gogogs* vs. *totowa*) in Table 2. Words with MCF are contrasted with the same lexical items with and without CV reduplication (e.g., *gogs* vs. *gogogs*).

The dialogue was recorded in a sound-attenuated booth at the Linguistics Graduate Lab at University of New Mexico. We used a Zoom H6 recorder and a matched pair of two Rode M5 microphones. Two male native speakers of Tohono O'odham

Gloss	SG	Syll. Str.	PL	Syll. Str.
coyote	*ban*	CVC	*baban*	CV<u>CV</u>C
dog	*gogs*	CVCC	*gogogs*[ii]	CV<u>CV</u>CC
duck	*pado*	CVCV	*papado*	CV<u>CV</u>CV
turkey	*towa*	CVCV	*totowa*	CV<u>CV</u>CV
bull	*tolo*	CVCV	*totolo*	CV<u>CV</u>CV
pig	*koji*	CVCV	*kokoji*	CV<u>CV</u>CV
cat	*mistol*	CVCCVC	*mimistol*	CV<u>CV</u>CCVC
cow	*nowiyu*	CVCVCV	*nonowiyu*	CV<u>CV</u>CVCV

Table 1: Target words used for this study (SG (singular), PL (plural)). Underlined CVs are the reduplicants of plural words.

participated in this study. One speaker is Robert Cruz, a co-author and a native speaker of Tohono O'odham. During the recording, speakers were presented with conversational dialogues with questions and answers written in the language's orthography [3]–[5] and English translation on a computer monitor. Both speakers are literate in the language's orthography, and they read both questions and answers. In total, 512 word-tokens were collected from two speakers; 383 word-tokens were analyzed after excluding tokens with improper renditions.

Acoustic properties were measured using Praat [22]. Consonant duration was calculated based on closure duration and voice onset time for oral stop consonants and nasal murmur for nasal stop consonants. Vowel duration was measured based on the length of vocal fold vibration shown in the waveform and the spectrogram. *F0* values were taken from the highest values within V_1 and V_3 intervals and the lowest values within V_2 and V_4 intervals to gauge the stress pattern. We performed a series of linear mixed-effects regression analyses [23] in *R* [24] to estimate the statistical significance of independent variables (Focus Type [BF, NF, LCF, MCF], Number of CVs in words [1 – 4], and Word Type [SG and PL]) on dependent variables (CV duration, *f0*).

Prime and target sentences [Focus Type, Location]
Q: Hascu ia o'ohadag? (What is drawn here?) A: Heg añ ñeid mo g *gogogs* ha-huhu'id g mimistol. [*BF*, S] (What I see is that the *dogs* are chasing the cats.) Q: Hascu huhu'id g *mimistol*? (What chases the cats?) A: Heg añ ñeid mo *gogogs* ha-huhu'id g mimistol. [*NF*, S] (What I see is that the *dogs* are chasing the cats.) Q: No wuḍ g ***gogs*** (***totowa***) mo ha-huhu'id g mimistol? (Is the ***dog*** (are the ***turkeys***) chasing the cats?) A: Pi'a. ***Gogogs*** 'o ha-huhu'id g mimistol. [***MCF*** (***LCF***), S] (No. ***Dogs*** are chasing the cats.)
Q: Hascu ia o'ohadag? (What is drawn here?) A: Mimistol 'o ha-huhu'id g *gogogs* mo ap o'ohadag. [*BF*, O] (Cats are chasing the *dogs* is what is drawn.) Q: Hascu am ha-huhu'id g *mimistol*? (What are the *cats* chasing?) A: Mimistol 'o ha-huhu'id g *gogogs* 'o ap o'ohadag. [*NF*, O] (Cats are chasing the *dogs* is what is drawn.) Q: No wuḍ g mimistol mo (ha-)huhu'id g ***gogs*** (***totowa***)? (Are the cats chasing the ***dog*** (***turkeys***)?) [***MCF*** (***LCF***), O] A: Pi'a. Mimistol 'o ha-huhu'id g ***gogogs*** mo ap o'ohadag. (No. Cats are chasing the ***dogs*** is what is drawn.)

Table 2: Prime (Q) and target (A) sentences in Tohono O'odham and English Translation.

3 RESULTS

3.1 Focus prominence on CV duration of words

The model[iii] estimated that the recurring strong-weak stress pattern is signaled by CV duration (**Figure 2**). C_1V_1 (β (estimate) = 331, SE (standard error) = 13, p < .001) was longer than C_2V_2 (β = 210, SE = 6, p < .001). C_3V_3 (β = 226, SE = 7, p < .001) was longer than C_2V_2. C_4V_4 (β = 185, SE = 16, p < .001) was the shortest among other CV sequences (i.e., $\mathbf{C_1V_1 > C_2V_2; C_2V_2 < C_3V_3; C_3V_3 > C_4V_4}$). The model, however, found that focus prominence effects were not statistically significant for all singular and plural words (BF: β = 331, SE = 13, p > .1; NF: β = 340, SE = 6, p > .1; LCF: β = 340, SE = 7, p > .1; MCF: β = 337, SE = 7, p > .1) (i.e., **BF = NF = LCF = MCF** on *all* CV sequences)[iv]. Speakers signaled the metrical stress with varying duration of CVs, but focus prominence was *not* indicated by duration at all.

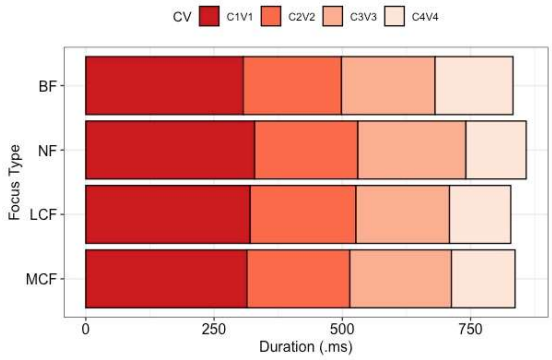


Figure 2: Mean duration of CV sequences with varying levels of focus prominence effect.

3.2 Focus prominence on *f0* of vowels

The model[v] estimated that *f0* was highest at V_1, and it gets lower towards the end of a word (V_1: β = 130, SE = 4, p < .001; V_2: β = 110, SE = 2, p < .01; V_3: β = 107, SE = 3, p < .01; V_4: β = 96, SE = 6, p < .01). Unlike the durational difference discussed in the previous sections, only V_1 and V_2 showed the statistically significant SW stress pattern. The rest of vowels did not show such a pattern (i.e., $\mathbf{V_1 > V_2 = V_3}$

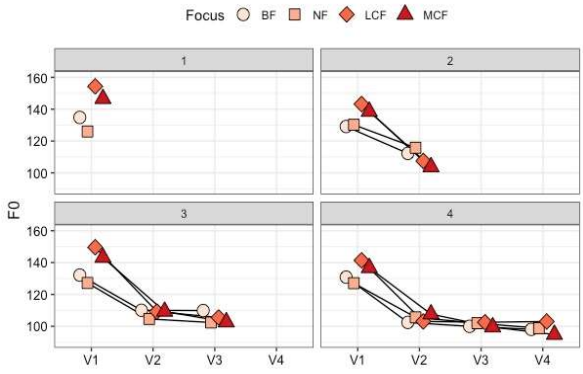


Figure 3: Mean *f0* values of vowels with varying levels of focus prominence effect. Numbers in each facet indicate the number of vowels (e.g., 1 = words with one vowel; 3 = with three vowels).

= V_4) (**Figure 3**). This may be due to the indication of a prosodic juncture signaling the end of a phrase. It can also be due to the assignment of an L tone indicating the end of a prosodic phrase [15], but we assume that the precise nature of the prosodic phrasing of the language is subject to further research.

Unlike the focus prominence on duration, the focus prominence on *f0* was realized differently across vowels. The model found that the LCF (β = 146, SE = 4, $p < .01$) and MCF (β = 139, SE = 4.5, $p < .001$) *increased f0* at V_1 as compared to BF (β = 130, SE = 4, $p < .001$) and NF (β = 128, SE = 4, $p > .01$) at V_1. LCF (β = 91, SE = 5, $p < .001$) and MCF (β = 95, SE = 5, $p < .001$) *decreased f0* at V_2 as compared to BF (β = 110, SE = 5.5, $p < .001$) and NF at V_2 (β = 112, SE = 5, $p > .1$). No contrastive focus effects (LCF and MCF) were statistically significant on V_3 and V_4. When contrasting a word, speakers enhanced word-level stress patterns by producing strong stress stronger and weak stress weaker using *f0* while maintaining its duration. Note that, as shown in **Table 3**, LCF and MCF varied depending on the number of vowels in each word. Focus prominence does not depend on grammatical status.

Word-level metrical stress is further distinguished with focus prominence, which is indicated solely with *f0*. Because speakers did not indicate focus prominence on duration, we argue that focus prominence does *not* necessarily enhance or strengthen all phonetic qualities of segments. Instead, it involves speakers' precise control of phonetic properties signaling the morphophonological structure of the language.

Words		Focus Type
1 CV (SG)	V_1	**BF = NF < LCF = MCF**
2 CVs (SG & PL)	V_1	**BF = NF < LCF = MCF**
	V_2	**BF = NF > LCF = MCF**
3 CVs (PL)	V_1	**BF = NF < LCF = MCF**
	V_2	BF = NF = LCF = MCF
	V_3	BF > NF = LCF = MCF
4 CVs (PL)	V_1	**BF = NF < LCF = MCF**
	V_2	BF = NF = LCF = MCF
	V_3	BF = NF = LCF = MCF
	V_4	BF = NF < LCF = MCF

Table 3: Post-hoc comparison of *f0* values with different focus types illustrated in **Figure 3.**

4 DISCUSSION AND CONCLUSION

We first aimed to understand how the metrical stress of the language is manifested with *f0* and duration by comparing the production of words conditioned with reduplication. Speakers employed both *f0* and duration to indicate the word-level stress pattern. Also, speakers heavily relied on acoustic cues of vowels rather than those of consonants to signal metrical stress (see endnote iv about this point).

Regarding focus prominence effects, our findings show that speakers precisely decreased *f0* of the second CV with LCF and MCF while maintaining the duration of CV sequences according to the stress pattern. In other words, when they need to contrast a word in context, they indicate more precise and distinctive metrical tones with *f0* to signal word stress. This implies that Tohono O'odham certainly treats the base and the reduplicant as equivalent morphophonological units but modifies these units according to strong-weak metrical stress using phonetic cues. We propose that Tohono O'odham uses both stress and mora as phonological timing units, and the recipient of stress is a mora (i.e., *mora-sensitive stress*), instead of a syllable, which has not been reported in languages previously studied. To better understand the co-existence of mora and stress and their interplay, further research should examine the production of words permitting C reduplication with acoustic and articulatory measurements.

We expected that speakers would produce words with LCF and MCF differently; however, MCF was statistically insignificant from LCF in both *f0* and durational measures. We assume speakers might have difficulty understanding orthographic stimuli because the language is primarily spoken rather than written. Picture-based tasks [25] may better test the effect of intonational prominence on word prosody in Tohono O'odham.

In conclusion, CV can be regarded as a basic morphophonological unit that explains lexical and post-lexical prominence variations in this language. The phonetic and prosodic modification of CV contributes to the formation of a certain aspect of morphophonology in Tohono O'odham. We look forward to further developing this study by involving more native speakers of the language; therefore, it will contribute to language documentation and revitalization of the Tohono O'odham language.

5 ACKNOWLEDGMENTS

We are highly grateful to the research participant (a native speaker of Tohono O'odham) for providing his voice and sharing in-depth knowledge of the language with us. We also thank Caroline L. Smith for making this study possible and helping us develop the study. The study is administratively reviewed and approved by the Institutional Review Board at University of New Mexico (Protocol #: 2250030063).

6 REFERENCES

[1] K. Hale, "Some Preliminary Observations on Papago Morphophonemics," *Int. J. Am. Linguist.*, vol. 31, no. 4, pp. 295–305, 1965.

[2] A. Alvarez and K. Hale, "Toward a Manual of Papago Grammar: Some Phonological Terms," *Int. J. Am. Linguist.*, vol. 36, no. 2, pp. 83–97, 1970.

[3] M. Mathiot, *A dictionary of Papago usage*, 2 vols. Bloomington, IN: Indiana University Press, 1973.

[4] O. Zepeda, *A Tohono O'odham Grammar*. Tucson, AZ: University of Arizona Press, 1983.

[5] D. Saxton, L. Saxton, and S. Enos, *Dictionary: Tohono O'odham / Pima to English, English to Tohono O'odham / Pima*. Tucson, AZ: University of Arizona Press, 1998.

[6] D. M. Eberhard, G. F. Simons, and Garcia, R. C., *Ethnologue: Languages of the world*. Dallas: SIL International, 2019.

[7] C. M. Fitzgerald, "Loanwords and Stress in Tohono O'odham," *Anthropol. Linguist.*, vol. 41, no. 2, pp. 193–208, 1999.

[8] C. M. Fitzgerald, "The morpheme-to-stress principle in Tohono O'odham," *Linguistics*, vol. 39, no. 5, pp. 941–972, 2001.

[9] C. M. Fitzgerald, "Tohono O'odham stress in a single ranking," *Phonology*, vol. 19, no. 2, pp. 253–271, 2002.

[10] C. M. Fitzgerald, "Prosodic Inconsistency in Tohono O'odham," *Int. J. Am. Linguist.*, vol. 78, no. 4, pp. 435–463, 2012.

[11] A. C. L. Yu, "Stress assignment in Tohono O'odham," *Phonology*, vol. 17, no. 1, pp. 117–135, 2000.

[12] J. H. Hill and O. Zepeda, "Derived Words in Tohono O'odham," *Int. J. Am. Linguist.*, vol. 58, no. 4, pp. 355–404, 1992.

[13] J. H. Hill and O. Zepeda, "Tohono O'odham (Papago) Plurals," *Anthropol. Linguist.*, vol. 40, no. 1, pp. 1–42, 1998.

[14] J. Riggle, "Infixing Reduplication in Pima and Its Theoretical Consequences," *Nat. Lang. Linguist. Theory*, vol. 24, no. 3, pp. 857–891, 2006.

[15] K. Hale and E. Selkirk, "Government and tonal phrasing in Papago," *Phonol. Yearb.*, vol. 4, no. 1, pp. 151–183, 1987.

[16] S. Inkelas and C. Zoll, *Reduplication: doubling in morphology*. in Cambridge studies in linguistics, no. v. 106. Cambridge: Cambridge University Press, 2005.

[17] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguist. Vanguard*, vol. 3, no. 1, p. 20170007, 2017.

[18] F. Cangemi and S. Baumann, "Integrating phonetics and phonology in the study of linguistic prominence," *J. Phon.*, vol. 81, p. 100993, 2020.

[19] M. Grice and F. Kügler, "Prosodic Prominence – A Cross-Linguistic Perspective," *Lang. Speech*, vol. 64, no. 2, pp. 253–260, 2021.

[20] T. Roettger and M. Gordon, "Methodological issues in the study of word stress correlates," *Linguist. Vanguard*, vol. 3, no. 1, p. 20170006. 2017.

[21] C. Gussenhoven, "Types of Focus in English," in *Topic and Focus*, C. Lee, M. Gordon, and D. Büring, Eds., in Studies in Linguistics and Philosophy, vol. 82. Dordrecht: Springer Netherlands, 2007, pp. 83–100.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. [Online]. Available: https://www.praat.org

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *J. Stat. Softw.*, vol. 67, no. 1, 2015.

[24] R Core Team, "R: A language and environment for statistical computing." Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[25] S.-A. Jun and J. Fletcher, "Methodology of studying intonation: from data collection to data analysis," in *Prosodic Typology II: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed., Oxford: Oxford University Press, 2014, pp. 493–519.

[i] The language also permits (partial) C reduplication, which copies the first consonant and pastes it after the base (e.g., *koton* 'a shirt', *kokton* 'shirts'). Some dialects permit the partial C reduplication to words differently; thus, the reduplication of Tohono O'odham varies across speakers.

[ii] The language allows complex onset and coda in the syllable [5]. [11] argued that no secondary stress is applied to odd-numbered syllables of polysyllabic words (e.g., *gogogs*, *mimistol*). Our data (or our knowledge of Tohono O'odham) indicates that the secondary stress is still applied to the odd-numbered syllables regardless of shape. Also, we see that strong-weak stress is applied to every CV of a word regardless of the derivational status of a word. Some test words are derived from Spanish (e.g., *towa* 'a turkey', *pado* 'a duck').

[iii] Duration ~ Focus Type * CV Numbers + Word Type + (1 | Speaker) + (1 | Word)

[iv] We did not include more detailed phonetic measures of CV sequences in this paper. This is because our models did not find any statistical significance of focus prominence effects on other durational measures, such as closure duration and voice onset time of consonants, and duration of vowels.

[v] F0 ~ Focus Type * CV Numbers (V) + Focus Type * Word Type + (1 | Speaker) + (1 | Word)

Focus prosody in Fijian: A pilot study

Albert Lee[1], Candide Simard[2], Apolonia Tamata[2], Jiaying Sun[1], Tsz Ching Mut[1]

[1] The Education University of Hong Kong, [2]University of the South Pacific
albertlee@eduhk.hk, candide.simard@usp.ac.fj, apolonia.tamata@usp.ac.fj, jsun@eduhk.hk, tcmut@eduhk.hk

ABSTRACT

This paper presents an acoustic analysis of focus prosody in Fijian – a verb-first language. Through analyzing pitch (f_0), intensity, and duration data, we compared (i) VP (i.e. initial) vs. neutral focus, (ii) subject vs. corrective-subject focus, and (iii) object vs. corrective-object focus. We found significant on-focus mean f_0 raising for VP focus, but no post-focus compression. Significant on-focus *shortening* of syllable duration was also observed. Compared with subject focus, corrective-subject is marked by significant on-focus *increase* in intensity and *shortening* of syllable duration. No significant difference between object and corrective-object focus was observed in the on-focus domain. Our findings pave the way for future systematic investigations of Fijian prosody, a currently understudied area in phonetics.

Keywords: Fijian, focus

1. INTRODUCTION

Studying prosodic focus often requires comparing the acoustic realisation of the same utterance in different focus conditions, elicited in a controlled manner (see overview in [1]). However, singling out prosodic cues can be less straightforward for languages where focus marking involves compulsory syntactic means (e.g. fronting in Samoan [2]), such as verb-initial languages (see review in [3]). One such language is Fijian, which to date has not been systematically studied in a controlled production experiment.

Fijian is an Austronesian language spoken by about 400,000 as a first language [4] in Fiji. It has positional stress. The basic word order of Fijian is verb-object-subject (but note alternative accounts such as [4]). Currently, there is no known experimental study reporting on the acoustic correlates of Fijian focus.

For a (reasonably) nearby verb-initial language Samoan, it was found that the initial phonological phrase was always the most prominent [2]. Thus when the focus was fronted, it was maximally prominent. In VSO sentences, the verb and agent were in the initial phrase. Speakers raised the pitch on the object in object focus, and lowered it in agent focus; although they did not do this consistently. There was no prosodic marking of focus on the agent.

As a first study of Fijian focus prosody, here we focused on identifying prosodic cues to focus and refrained from analysing conditions that would necessitate using syntactic focus markers like fronting. Therefore, in this paper we compared the following focus conditions: (i) VP (i.e. initial) vs. neutral focus, (ii) subject vs. corrective-subject focus, and (iii) object vs. corrective-object focus. As VP is sentence initial, analyzing VP vs. neutral focus allows us to see in situ prosodic marking of narrow focus in Fijian. In addition, we are also interested in whether in a corrective context, the same narrow focus (i.e. subject and object focus) would be marked differently.

2. METHODS

2.1. Participants

The study included ten native Fijian speakers (5 female and 5 male). They were born and raised in Fiji and spoke fluent Fijian. They were university students at the time of testing. No one reported any speech or hearing impairments.

2.2. Materials

Five different declarative sentences were designed (Table 1) to elicit prosodic focus. S1 and S2 are typical verb-object-subject (VOS) sentences in Fijian and S3-S5 are verb-subject (VS) sentences. S4 is in the passive voice. For S1 and S2, each base sentence has six possible focus conditions, namely, neutral focus, VP (verb phrase) focus, object focus, subject focus, corrective-subject focus (i.e., the subject is corrected by the speaker and under focus), and corrective-object focus. S3, S4, and S5 (VS sentences) do not have (corrective) object focus. When defining the on-focus domain [5], only content words were included (except for VP focus condition). All focus conditions were elicited by a precursor question asked by the interviewer (Table 2).

S1	Era vaqitora tiko na lali na gone. 3pl play-prep ASP ART drum ART child 'The children are playing with the drum.'
S2	E volia tiko na dalo na yalewa. 3sg buy-TR ASP ART taro ART woman 'The woman is buying taro.'

S3	Eratau qaqalo tiko na gone. 3pl swim ASP ART child 'The kids are swimming.'
S4	E waluvu tiko na vale. 3sg flood ASP ART house 'The house is being flooded.'
S5	E sa vuka vakababa na bebe. 3sg ASP fly sideways ART butterfly 'The butterfly is flying.'

Table 1: Base sentences used in this study.

Focus	Precursor questions	Target sentences
Neutral	Na cava e yaco tiko? ART what 3sg happen ASP 'What is happening?'	Era vaqitora tiko na lali na gone. 3pl play-prep ASP ART drum ART child 'The children are playing with the drum.'
VP	Na cava era cakava tiko na gone? ART what 3pl do ASP ART children 'What are the children doing?'	Era vaqitora tiko na lali na gone. 3pl play-prep ASP ART drum ART child 'The children are playing with the drum.'
Subject	O cei e vaqitora tiko na lali? ART who 3sg play-prep ASP ART drum 'Who is playing with the drum?'	O ira na gone, Era vaqitora tiko na lali. ART 3pl ART child 3pl play-prep ASP ART drum 'The children are playing with the drum.'
Object	Na cava era vaqitora tiko na gone? ART what 3pl play-prep ASP ART children 'What are the children playing with? '	Era vaqitora tiko na lali na gone. 3pl play-prep ASP ART drum ART child 'The children are playing with the drum.'
Corrective Subject	Eta vaqitora tiko beka na dramu? 1pl play-prep ASP perhaps ART drum 'Are we playing with the drum?'	Sega, o ira na gone era vaqitora tiko na lali. Not ART 3pl ART child 3pl play-prep ASP ART drum 'No, the children are playing with the drum.'
Corrective Object	Era qito rakavi tiko? ART play-TR rugby ASP 'Are the children playing rugby?'	Sega, era vaqitora tiko na lali na gone. Not 3pl play-prep ASP ART drum ART child 'No, the children are playing with the drum.'

Table 2: Examples of precursor questions and target sentences (the focused elements are underlined).

Speakers were asked to say each sentence three times. Altogether, we recorded 720 utterances ((2 sentences * 6 focus condition + 3 sentences * 4 focus condition) * 3 repetitions * 10 speakers). However, 168 (23.3%) utterances had to be discarded due to missing or redundant syllables, or unexpected word order. Consequently, a total of 552 sentences were retained for analysis. The seemingly high discard rate is considered to be due to the unscripted nature of the task (see §2.3).

2.3. Recording procedure

Recording took place in a room at the University of the South Pacific. We produced 21 cards, each with an image and a sentence describing it. The interviewer showed the participant the card and asked them to answer questions using the card's information. The presentation order of the cards was random. For the corrective focus condition, participants were required to correct the inaccurate information provided by the interviewer and give the complete and correct sentences. Before recording commenced, there were two tests in which the speakers were asked to describe the images, so that we could ensure that they understood the experimental procedure. The total recording time for each participant was about one hour.

2.4. Annotation and measurements

The raw sound data were first chunked into individual utterances, and then labelled by syllable in PRAAT [6]. As responses were unscripted, many deviated from the target list (see §2.2). Thus when labelling the recordings, we retained extra syllables that were semantically consistent with the expected answer. In cases where finding the syllable boundary was impossible (e.g. due to fast speech), it was placed in the dead centre of the two syllables concerned. Pauses, repeated syllables, and mispronounced syllables were not labelled.

After annotation, we checked and rectified vocal pulse markings manually for accurate f_0 tracking and obtained the time-normalised f_0, intensity, and duration with ProsodyPro [7].

3. RESULTS

3.1 F0

We fitted linear mixed effects models to pre-focus, on-focus, and post-focus mean f_0 (Hz) data using *lmerTest()* [8]. Model construction followed a bottom-up approach. Post-hoc comparisons were done using *emmeans()* [9]. For on-focus mean f_0, the best fitting model contained the fixed factor of focus condition (VP, neutral, subjective, Corrective-subject, object, Corrective-object), as well as by-subject and by-sentence random intercepts. Including random slopes prevented model convergence.

Intensity (§3.2) and duration (§3.3) data were analysed using the same approach.

The main effect of focus on on-focus f_0 mean was significant, $X^2(6) = 214.46$, $p < .001$. Post-hoc test shows that **VP focus** had significantly higher on-focus mean f_0 than corresponding neutral focus ($p < .0001$). No other significant contrast was found. However, SS ANOVA [10] plots for individual sentences (e.g. Figure 1) reveal that this difference was non-significant for 3 of 5 sentences.

The main effect of focus on post-focus mean f_0 was significant too, $X^2(6) = 159.71$, $p < .001$. The Post-hoc test shows that **Corrective-Subject focus** had significantly lower post-focus mean f_0 than corresponding **Subject** focus ($p = .0001$). However, inspection of SS ANOVA plots (e.g. Figure 2) revealed that this difference was significant for 2 of 5 sentences.

Finally, the main effect of focus on pre-focus mean f_0 was significant, $X^2(6) = 221.25$, $p < .001$. The Post-hoc test shows that **Corrective-Object focus** had significantly lower pre-focus mean f_0 than corresponding **Object** focus ($p < .0001$). SS ANOVA plots (e.g. Figure 3) show that this difference was observed in all 5 sentences.

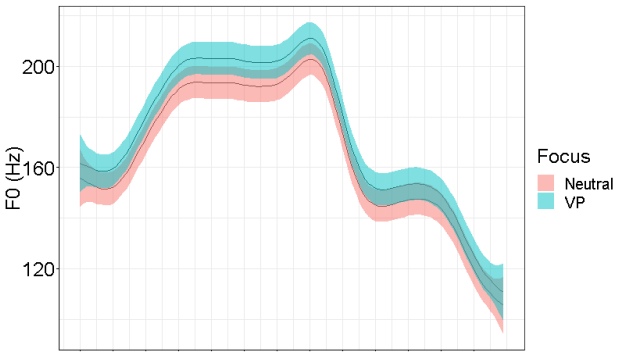


Figure 1: SS ANOVA comparing VP and neutral focus of S2 (see Table 1).

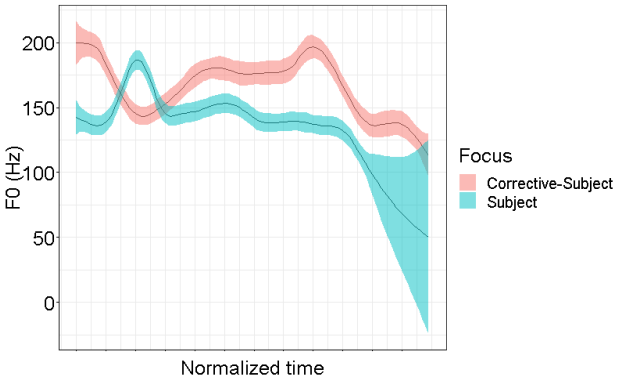


Figure 2: SS ANOVA comparing subject and Corrective-Subjective focus of S5.

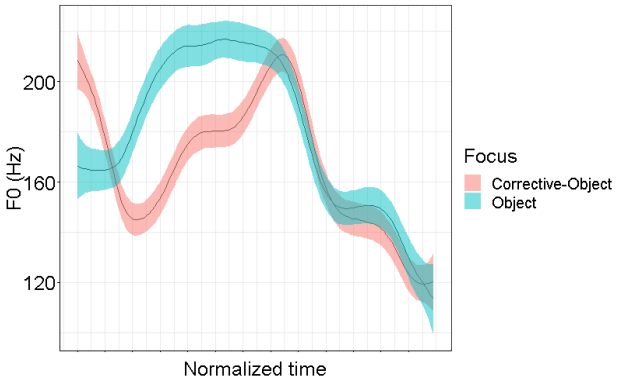


Figure 3: SS ANOVA comparing Object and Corrective-Object focus of S2.

3.2 Intensity

There was a main effect of focus condition on pre-focus ($X^2(6) = 70.24$, $p < .001$), on-focus ($X^2(6) = 115.19$, $p < .001$), and post-focus mean intensity ($X^2(6) = 41.34$, $p < .001$). Compared with neutral focus, **VP focus** had significantly lower on-focus intensity ($p = .0016$). Compared with **Subject focus**, **Corrective-Subject focus** had significantly higher on-focus intensity ($p < .0001$, see Figure 4), and lower pre-focus ($p < .0001$, see Figure 5) and post-focus intensity ($p = .0203$, see Figure 6). Compared with **Object focus**, **Corrective-Object focus** had significantly lower pre-focus ($p = .0241$) and post-focus intensity ($p = .0020$).

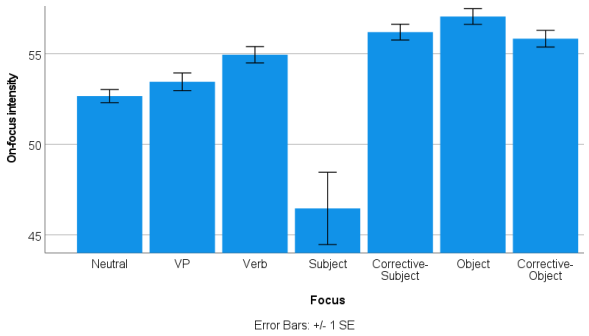


Figure 4: Barplot comparing on-focus mean intensity (dB) in different focus conditions.

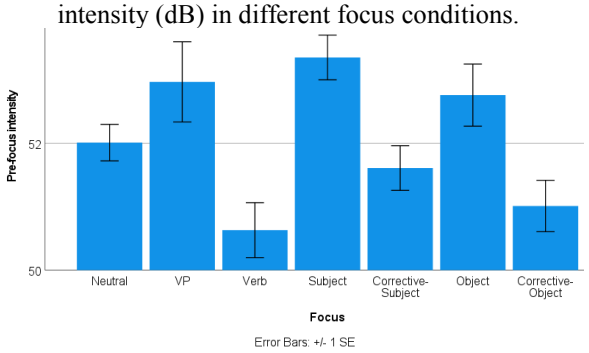


Figure 5: Barplot comparing pre-focus mean intensity (dB) in different focus conditions.

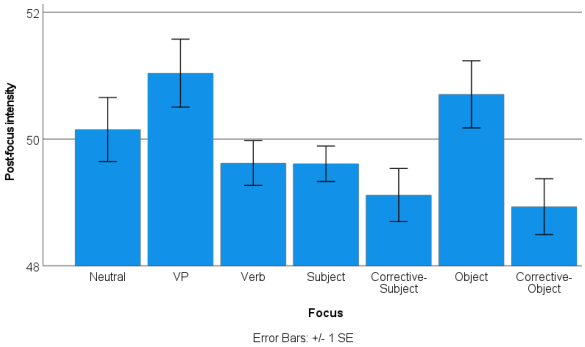


Figure 6: Barplot comparing post-focus mean intensity (dB) in different focus conditions.

3.3 Duration

There was a main effect of focus condition on pre-focus ($X^2(6) = 538.37$, $p < .001$), on-focus ($X^2(6) = 152.29$, $p < .001$), and post-focus mean syllable duration ($X^2(6) = 83.17$, $p < .001$). Compared with neutral focus, **VP focus** had significantly shorter on-focus ($p = .0001$) and pre-focus ($p < .0001$) duration. Compared with **Subject focus**, **Corrective-Subject focus** had significantly shorter on-focus ($p < .0001$, see Figure 7) but longer pre-focus ($p < .0001$, see Figure 8) and post-focus ($p = .0432$) duration. Compared with **Object focus**, **Corrective-Object focus** had significantly longer pre-focus ($p < .0001$) but shorter post-focus duration ($p = .0020$).

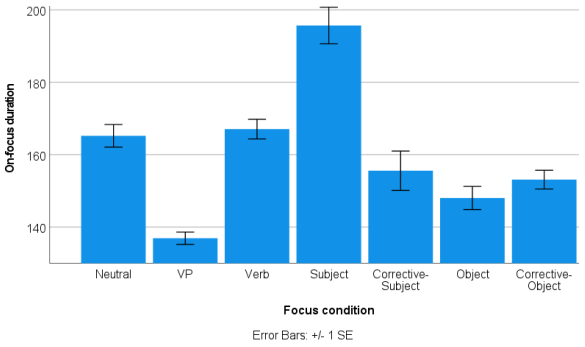


Figure 7: Barplot comparing on-focus mean duration (ms) in different focus conditions.

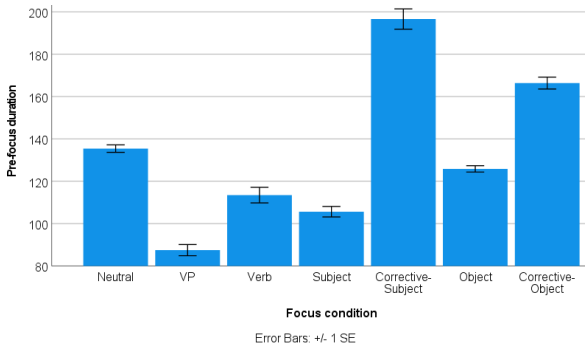


Figure 8: Barplot comparing pre-focus mean duration (ms) in different focus conditions.

4. DISCUSSION

This study set out to identify prosodic focus marking strategies in Fijian. We compared (i) VP (i.e. initial) vs. neutral focus, (ii) subject vs. corrective-subject focus, and (iii) object vs. corrective-object focus. We found significant on-focus mean f_0 raising for VP focus, but no post-focus compression. There was also considerable cross-sentence variability. Significant on-focus *shortening* of syllable duration was also observed. Compared with subject focus, corrective-subject is marked by significant on-focus *increase* in intensity and *shortening* of syllable duration. No significant difference between object and corrective-object focus was observed in the on-focus domain.

Through inferential statistics we have found that focus condition significantly affected prosodic focus markers in different ways. However, note also that there was considerable cross-speaker variability, in part due to the unscripted nature of participants' responses. Cross-sentence variability was also observed, showing that even highly significant effects may not necessarily be applicable to all sentences of the same focus condition.

As the first systematic production study of Fijian focus prosody, we have only been able to cover initial focus (i.e. VP focus), leaving out other positions. To bypass compulsory fronting, one could design production tasks such as nonce sentences or phone digits to elicit in situ prosodic marking strategies (see [11]). In addition, given Fijian focus is marked by multiple prosodic cues alongside syntactic strategies, one could use statistical tests such as linear discriminant analysis [12] to identify their respective predictive power of focus. Much more work is needed to get a better understanding of focus prosody in Fijian, a currently understudied Austronesian language.

6. REFERENCES

[1] Y. Xu, "Speech prosody: A methodological review," *J. Speech Sci.*, vol. 1, no. 1, pp. 85–115, 2011.

[2] S. Calhoun, "The interaction of prosody and syntax in Samoan focus marking," *Lingua*, vol. 165, pp. 205–229, 2015.

[3] F. Kügler and S. Calhoun, "Prosodic encoding of information structure: A typological perspective," in *The Oxford Handbook of Language Prosody*, C. H. M. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 454–467.

[4] P. A. Geraghty, "Fijian," in *Concise encyclopedia of languages of the world*, K. Brown and S. Ogilvie, Eds. Oxford, 2009, pp. 412–413.

[5] Y. Xu, C. X. Xu, and X. Sun, "On the temporal domain of focus," in *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)*, 2004, pp. 81–84.

[6] P. P. G. Boersma and D. J. M. Weenink, "Praat: Doing phonetics by computer." 2012.

[7] Y. Xu, "ProsodyPro: A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 2013, pp. 7–10.

[8] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017.

[9] R. V. Lenth, "Emmeans: Estimated marginal means, aka least-squares means." 2020, [Online]. Available: https://cran.r-project.org/package=emmeans.

[10] L. S. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 407–415, 2006.

[11] T. C. Mut, C. Simard, Apolonia Tamata, and A. Lee, "Focus prosody in Fijian: In-situ focus marking,", *Proc of ICPhS 2023*, 2023.

[12] M. E. Breen, E. Fedorenko, M. Wagner, and E. A. F. Gibson, "Acoustic correlates of information structure," *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 1044–1098, 2010.

6. ACKNOWLEDGEMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China awarded to AL (Project No. 18600621). We thank Ms. Jueyu Lu for helping with data annotation.

AN ACOUSTIC STUDY OF THE FRICATIVE VOWEL IN TWO ENDANGERED RYUKYUAN LANGUAGES

Shigeko Shinohara[1], Bowei Shao[2]

[1] Laboratoire de Phonétique et Phonologie, CNRS/Université Sorbonne Nouvelle, Paris, France
[2] Département d'études cognitives, École Normale Supérieure, Université PSL, Paris, France
shigeko.shinohara@sorbonne-nouvelle.fr, bowei.shao@ens.psl.eu

ABSTRACT

This study investigates the acoustic properties of the fricative vowel of two varieties of Ryukyuan languages: Aragusuku Yaeyama (AR) and Karimata Miyako (KR) Ryukyuans. Ryukyuans are endangered languages spoken in the Southwest Islands of Japan. The phonetic properties of the fricative vowel have been an issue of debate. We aim to characterize the frication in the acoustic signals of this vowel. We compared the zero-crossing rates (ZCR) between the fricative vowel and the regular vowel /i/. The results indicated that the fricative vowel has frication noise superimposed on its formant structure in the 5-8 kHz region. This noise distinguished the fricative vowel from /i/ in AR but not in KR based on ZCR. Nevertheless, the fricative vowels in both languages have a similar quantity of frication noise. These results suggest that the fricative vowels in AR and KR have a similar acoustic profile to those reported in other languages.

Keywords: Fricative vowel, Ryukyuan, Endangered language, Zero-crossing rate, GAMM

1. INTRODUCTION

Ryukyuan languages are spoken on the Southwest Islands of Japan. All six Ryukyuan languages were identified as endangered by UNESCO [15]. Some dialects of Miyako and Yaeyama Ryukyuans possess, cross-linguistically rare, fricative vowels (FVs) [6, 7, 13].

FVs are syllable nuclei composed of fricative sounds and are found in several unrelated languages [2, 13]. One well-known example is the apical vowels in Chinese languages, which is the most studied example of FVs [3, 4, 5, 11, 14, 23, 24, 25]. Acoustic analyses of FVs showed supralaryngeal frication generated at labial or alveolar place superimposed on voicing and that the frication evolves during the vowel interval. This frication could be observed following sibilant, plosive, and nasal onsets [23, 25]. Articulatory studies using ultrasound imaging showed a fricative-like tongue shape with medial grooving for the alveolar FV in Chinese languages [5, 25].

Data collection of the FV was planned in Miyako Islands for 2020 but has been suspended due to the pandemic. This situation drove us to explore the audio recordings of the Aragusuku dialect of Yaeyama Ryukyuan (AR) and the Karimata dialect of Miyako Ryukyuan (KR) made available by NII-SRC [17]. Ryukyuans are generally spoken by the elderly population. The number of speakers of each dialect is not censused. In 2022, the population over 65 years old in Aragusuku counts 8 and that of Karimata is 190. AR is the only dialect preserving an FV among Yaeyama Ryukyuan [7] whereas the FV is found in most of the dialects on the Miyako main Island but in distinct preconsonantal contexts depending on the dialect (see Sec. 2).



Figure 1: The Ryukyu Islands and the areas where the two languages are spoken: Karimata (KR) on Miyako Islands and Aragusuku (AR) Islets of Yaeyama Islands (Okinawa Travel Guide, link)

The quality of the FV in Ryukyuans has been an issue of debate in Japanese dialectology. It has been variably called an 'apical vowel' [8, 22] or a 'central vowel' [6, 20]. [20] investigated the formant structure of the vowels in the same database. The F2 values of the FVs were located midway between /i/ and /u/ in AR and KR. KR's FV was slightly more front, closer to /i/. Frication noise was observed at 3 kHz along with the presence of F1 and F2 in AR. As a complementary study to [20], we shall examine

frication noise in the FV in both regular and devoiced contexts (see below).

2. SOUND STRUCTURES OF AR AND KR

The phonological descriptions below are based on [12] for AR and [19] for KR and supplemented by [7] for AR and [10] for KR. AR contrasts six vowels /a i u e o I/ (/I/ denotes the FV); KR does five /a i u o I/. As we interpret the consonant segments in [12], AR and KR both have /p b t d k g ts f v s z h m n w r j/ as contrastive consonants. They have a similar basic syllable structure (N)(C)V(N). In addition, many of the segments have length contrast. /I/ appears in restricted consonantal environments in both languages. It can follow /p b k g s ts z r/ in AR and /p b k g s ts z/ in KR. In both languages, it can appear without an onset consonant (e.g., /I/ [I:] 'rice'). The high vowels (including /I/) are devoiced between two voiceless consonants in both AR and KR. They can be devoiced phrase-finally, and the other vowels can also be devoiced in similar environments. Vowels can be devoiced after a voiceless consonant, and a following voiced consonant can also be devoiced (e.g., [k̥I̥mu] 'liver').

3. STUDY QUESTIONS AND METHOD

3.1. Study questions

Our study aims to answer three research questions: 1) Do the FVs /I/ in AR and KR contain any frication noise in the acoustic signals?; 2) If they do, how does the frication noise evolve during the entire duration of /I/?; 3) Is there any difference in the degree of frication between the two languages?

3.2. Zero-crossing Rate and Generalized additive mixed model

To measure the degree of frication noise in the FV, we chose Zero-crossing Rate (ZCR) as an indicator. ZCR measures the number of upward and downward crossings of zero dB per second in the waveform, without involving the detection of voicing or pitch. Previous studies on the apical vowel in Chinese languages [3, 23, 25] demonstrated the robustness of this measurement in analyses of voiced segments containing frication noise.

In our study, ZCR was obtained as *PointProcess* objects in Praat [1] within a 20 ms sliding window on the acoustic signals. The ZCR was calculated as the number of zero-crossings divided by the length of the window. 22 data points (the onset and offset of all nucleic segments and 20 data points evenly spaced on each segment) were extracted. This method allowed us to visualize and analyze each segment's temporally dynamic ZCR patterns.

We chose Generalized Additive Mixed Models (GAMMs) [28, 29] to model the dynamic pattern of FVs, with *mgcv* package v1.8-40 in *R* v4.2.3 [21]. This method has been gaining ground in the investigation of dynamic speech patterns, including temporal and spatial aspects and their interactions [26, 27]. We chose GAMMs because ZCR was expected to change over time [3, 25], and we could also explore if the frication noise was influenced by other factors, notably by the type of onsets.

We made within-language and cross-language comparisons with GAMMs. Within-language comparisons would show the difference in ZCR between an FV /I/ and a regular vowel /i/ in a language. Cross-language comparisons were performed on the normalized ZCR scores, following [3]. The normalization took the ratio between the ZCR of /I/ and that of /s/ for each speaker (i.e., Relative zero-crossing rate (RZCR)). In the models, ZCR of the vowels was estimated over time, with factor smooths for onset and the voicing categories of the onset. Tweedie distributions were used in the model, as ZCR follows a left-skewed, long-tailed distribution. Results were visualized using *tidyverse* v1.3.2 and *tidymv* v3.3.2 [30, 31].

3.3. Speech materials

NII-SRC data contain recordings of a male Aragusuku speaker aged 83 in 1990 [12] and a female Karimata speaker aged 69 in 1996 [19]. The recording corpus included real words with five types of vowels: always-voiced long fricative vowel [I:], short fricative vowel [I] in voiced and devoiced environments, and regular vowel [i] in voiced and devoiced environments. The devoiced [I̥ i̥] tokens were treated separately from the voiced [I i]. The speakers repeated each word three times. The datasets turned out to be relatively small and uneven (AR: 37 [i], 20 [I], 4 [I:], 12 [I̥], 2 [i̥]; KR: 12 [i], 23 [I], 3 [I:], 8 [I̥], 2 [i̥]).

4. RESULTS

4.1. Aragukusu /I/

Fig. 2 presents three tokens of /I/: [I] preceded by a vowel, [maI] 'rice'; [I] preceded by a voiced affricate, [adzI] 'flavor'; and [I̥] between voiceless obstruents (i.e., a devoicing environment), [pI̥tu] 'human.' Frication noise is observed in the 4-6 kHz region in all three tokens. This suggests that the frication noise is not introduced by the preceding consonant. For the first two voiced tokens, the frication noise is superimposed on a clear formant structure, and the frication seems to diminish and disappear towards the

end of [I]. In the case of devoiced [I̥], the frication noise seems to be more intense in the same frequency range.

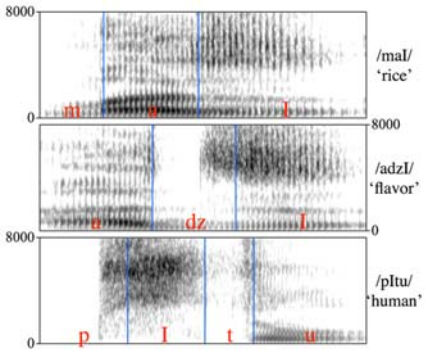


Figure 2: Spectrograms of three realizations of /I/: after a vowel, after a voiced consonant, and between voiceless consonants in Aragusuku.

The ZCR in FVs was further analyzed using GAMMs, compared with the high front vowel [i]. Fig. 3 shows the evolution of ZCR during [i], [I:], and devoiced [I̥] of AR. The devoiced segments' pattern is different from the voiced ones: devoiced segments have much higher ZCR. This is not surprising since when FVs are devoiced, their acoustic signal resembles that of a voiceless fricative consonant, having a much higher ZCR. For the voiced segments, overall, [i] has a lower ZCR and [I] has a higher ZCR, whereas [I:] has a descending pattern: it starts from a high point and gradually reaches the level of [i] and continues further down toward the end of the segment.

The [I:] and the devoiced [i̥] have larger confidence intervals compared to the other segments. This could be explained by the fact that there were fewer tokens of them in the dataset, thus, the fitted values of these two segments should be interpreted with caution.

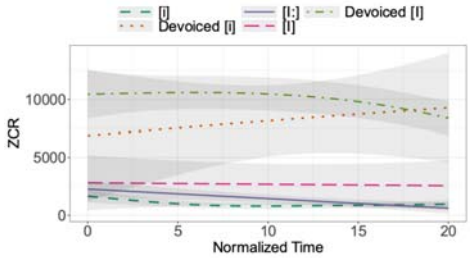


Figure 3: GAMM fitted values of Zero-crossing Rate during /i/ and /I/ in Aragusuku.

The differences among the three voiced segments [I I: i] are presented in Fig. 4. Compared to [i], [I:] has a higher ZCR in the first half of its interval; the second half is less fricated, having a similar ZCR to [i]. The ZCR of [I] is less reliable, however, it still has a higher ZCR overall than [i] does, especially in the middle portion, where the difference is significant.

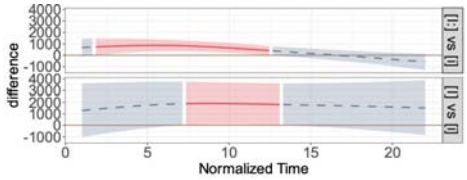


Figure 4: Difference between [i] and the fricative vowel [I, I:] in Aragusuku. The solid lines indicate portions with a significant difference.

4.2. Karimata /I/

Similar to AR, frication noise was observed in /I/ in all different contexts. Fig. 5 presents three tokens of /I/: [I] preceded by a vowel, [I] preceded by a voiced affricate, and [I̥] between voiceless obstruents. Frication noise is observed on the spectrogram of all three tokens but appears to be located at higher frequencies (specifically, in the 5-7 kHz region) than in AR. The frication noise is superimposed on a formant structure for the first two voiced tokens.

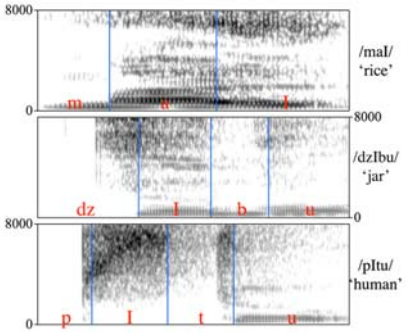


Figure 5: Spectrograms of three realizations of /I/ in Karimata. In /dzIbu/ 'jar,' the onset /dz/ is realized as [ts].

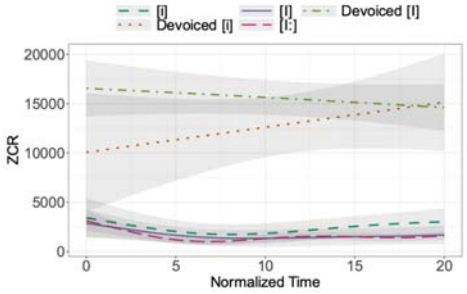


Figure 6: GAMM fitted values of Zero-crossing Rate during /i/ and /I/ in Karimata.

The ZCR of /i/ and /I/ were also analyzed using GAMMs. The first observation is, again, that the devoiced [I̥] and [i̥] pattern together differently from the voiced segments. For the voiced segments, contrary to AR [I], the KR [I] seems to pattern closely with [i], which can also be seen in Fig. 7. The overall difference between [I] and [i] is not significant, and the [I:] seems to have an even lower ZCR towards the end of its duration compared to [i]. However, the size

of the dataset calls for a cautious interpretation of the results.

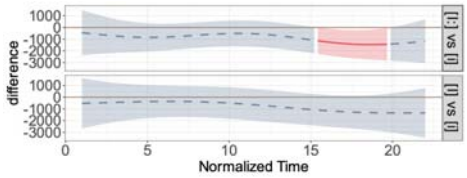


Figure 7: Difference between [i] and the fricative vowel [I, I:] in Karimata. The solid lines indicate a portion with a significant difference.

4.3. Comparison between AR and KR [I]

The results presented above may give the impression that the AR voiced [I] contains more frication noise than that in KR. The cross-language comparison (Fig. 8) indicates that although AR [I] has a slightly higher RZCR, there is no significant difference between the two languages. Fig. 9 arguably shows that the two languages' /I/s do not differ in terms of RZCR.

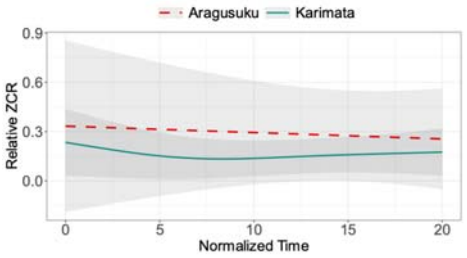


Figure 8. GAMM fitted values of Relative Zero-crossing Rate during /I/ in Aragusuku and Karimata.

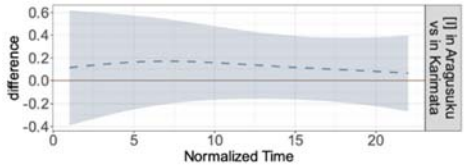


Figure 9. The difference in RZCR between the /I/s in Aragusuku and Karimata. The overall difference is not significant.

5. DISCUSSION AND CONCLUSION

This analysis is the first study conducted on the frication of the FV /I/ in two endangered languages, AR and KR. The results showed that there is frication noise on /I/ in the 5 kHz to 7 kHz region. This frication noise was observed when it was preceded by voiceless sibilants, vowels and voiced affricates. These observations show that the frication noise of /I/ could not be interpreted as being introduced by preceding sibilant consonants. When /I/ was devoiced ([I̥]) (i.e., between two voiceless consonants), the frication noise is more intense.

In AR, [I I:] contains significantly more frication than the regular vowel [i] at least in the first half of its duration. This pattern matches what was reported for apical vowels in Chinese languages [11, 25]. The pattern could be arguably explained by the aerodynamic adjustment [18]: the voicing and the frication are difficult to achieve simultaneously, and a trade-off has to happen during the fricative vowel, making the first half more fricative-like and the second half more vowel/approximant-like. In KR, [I I:] was not significantly more fricated than the regular vowel [i] as measured in ZCR. However, we note again that the limited number of tokens and speakers invites a careful interpretation. Recall that the AR speaker was a male and KR was a female. Female speakers might have a different physical structure and different socio-linguistic tendencies from male speakers influencing the results to some extent.

The interaction between voicing, frication, and duration among [I I: I̥] is complex. The long [I:] contains generally less frication compared to [I], and the devoiced [I̥] can be considered a fricative. Without knowledge of the articulatory characteristics of [I I:], it is difficult to determine how the frication noise is generated and superimposed on voicing. Similarly, understanding the acoustic shape of [I̥] requires an understanding of its supra-laryngeal articulation.

Finally, there was no significant difference in RZCR between [I] in the two languages, which might suggest that the difference in ZCR in each language was not solely caused by the frication noise present in [I I:], but also by the acoustic shape of [i].

Our study revealed that AR and KR have a strikingly similar type of fricative vowel, despite belonging to two different Ryukyuan languages, Yaeyama and Miyako. Historically, Ishigaki, the main island of Yaeyama (see Fig. 1) had contact with Okinawa, where the capital of the Ryukyu Kingdom was located in the past while Aragusuku islets were left out from the influence of the mainstream language changes [12]. However, historical language contact might not be the only reason for the similarity between AR and KR. Phonological environments may contribute to the preservation of the frication noise of /I/. /I/ in some of the Miyako Ryukyuan dialects (e.g., Hirara, Karimata, Bora) are reported to present frication noise in wider consonantal environments than in others (e.g., Ikema, Ogami) [16, 20]. And in the latter languages, the frication noise is reported declining. It would be desirable to conduct an articulatory study (e.g., ultrasound tongue imaging) across dialects of Miyako Ryukyuan along with acoustic and also phonological analyses to investigate the dialectal differences and the articulatory nature of the fricative vowel, which would contribute to a further understanding of the FV as a group of atypical sounds.

6. ACKNOWLEDGEMENTS

We would like to express our gratitude to the National Institute of Informatics' Speech Resources Consortium for providing us with access to the Aragusuku Dialect (Aragusuku) and Oogami Dialect (Oogami) speech databases. The study was supported by Labex EFL.

7. REFERENCES

[1] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.1.16.

[2] Dell, F. 1994. Consonnes à prolongement syllabique en Chine. *Cahiers de linguistique – Asie orientale* 23(1), 87–94.

[3] Faytak, M., Shao, B. 2022. Fricative vowels are noisier in crowded vowel spaces. *LabPhon18*. Online.

[4] Faytak, M. 2018. *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*, PhD dissertation. University of California, Berkeley.

[5] Faytak, M., Lin., S. 2015. Articulatory variability and fricative noise in apical vowels. *Proc. 18th ICPhS*, Glasgow.

[6] Hirayama, T (ed.). 1983. *Ryukyu Miyako shotou hougen kiso goi no sougouteki kenkyuu*. Tokyo: Ouhuu.

[7] Kajiku, S. 2016. Minami Ryukyu hougen gaisetsu, In: *Ryukyu no hougen, 40*, 107–200.

[8] Karimata, S. 1996. 'Miyako Hōgen no On'in Henkani Tsuite no Oboegaki: Kūki–Rikigakutekina Kanten kara Mite'. *Gengogakurin*, 1996–1997:709–22.

[9] Karlgren, B. 1915. *Études sur la phonologie chinoise*. Uppsala: K. W. Appelberg.

[10] Kinuhata, T., Hayashi, Y. 2014. Ryukyugo Miyako Karimata hougen no onin to bunpou, *Ryukyu no hougen*, 38, 17–49.

[11] Kong, H., Wu. S., Li, M. 2019. Héféi Huà Shéjiān Yuányīn de Mócā Xìngzhi jí Gǎnzhī Zēngqíang Lǐlùn Jiědú [The Frication Property of Apcial Vowels in Hefei Chinese and Its Perceptual Enhancement Account]. *Yuyan Yanjiu* 39(1), 23–33.

[12] Kuno, M., Kuno, M., Oono, M., Sugimura, T. 1993. *Zan o jounoushita hitotachi-Minami Ryukyu Aragusuku hougen no onsei to Jiraba*, Kokugakuin University, Institute for Japanese Culture and Classics.

[13] Ladefoged, P., Maddieson, I. 1996. *The sounds of the world's languages*. Oxford & Malden, MA: Blackwell.

[14] Lee-Kim, S-I. 2014. Revisiting Mandarin apical vowels: An articulatory and acoustic study. *J. Int. Phon. Assoc.* 44(3), 261–282.

[15] Moseley, C. 2010. *Atlas of the World's Languages in Danger* (3rd ed.). Paris: UNESCO Publishing.

[16] Nakahara, J, 2001. Okinawa Miyakojima Bora hougen no onin, *Ryukyu no hougen* 21, 105-123, Hosei University.

[17] NII-SRC (Speech Resources Consortium) The National Institute of Informatics (Last accessed 27 July 2022: http://research.nii.ac.jp/src/index.html).

[18] Ohala, J. J. 1983. The origin of sound pattern in vocal tract constraints. In MacNeilage, P. F. ed., The Production of Speech,189–216. Springer, New York.

[19] Oono, M., Kuno, M., Kuno, M., Sugimura, T. 1998. *Miyako Ogamijima no onsei – tango to bunpou (Appendix Karimata hogen)*, 1996-7 Grants in aid for Scientific Research Report.

[20] Oono, M., Kuno, M., Sugimura, T., Kuno, M. 2000. Phonetic substance of neutral vowels in the South Ryukyu Dialects, *Onsei Kenkyū* 4:1, 28–35.

[21] R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[22] Sakiyama, O. 1963. The Comparative Phonology of Miyako Dialects in the Ryūkyūs, *Kokugogaku*, 54, 6–21.

[23] Shao, B. 2020. *The apical vowel in Jixi-Hui Chinese: phonology and phonetics*, PhD dissertation. Université Sorbonne Nouvelle.

[24] Shao, B., Ridouane, R. 2021. 'Apical vowels' are not vowels: acoustic and ultrasound evidence from Jixi-Hui Chinese. *Proc. of the 12th International Seminar on Speech Production*, 110–113.

[25] Shao, B., Ridouane, R. accepted. On the nature of apical vowel in Jixi-Hui Chinese: Acoustic and articulatory data. *J. J. Int. Phon. Assoc.*

[26] Soskuthy, M. 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *J. Phon.*, *84*, 101017.

[27] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Pho.*, *70*, 86–116.

[28] Wood, S. 2006. *Generalized additive models: an introduction with R*. CRC Press.

[29] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3–36.

[30] Wickham H, Averick M, Bryan J, Chang W, McGowan L. D, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. "Welcome to the tidyverse." *Journal of Open Source Software*, 4(43), 1686.

[31] Coretta, S. 2022. tidymv: Tidy Model Visualisation for Generalised Additive Models, r package, version 3.3.2.

VOWEL SPACE AND ACOUSTIC CHARACTERISTICS OF STRESSED SYLLABLES IN TWO KIBUSHI DIALECTS SPOKEN IN MAYOTTE, FRANCE: A PILOT STUDY

Ahamada Kassime, Fabrice Hirsch, Miki Mori

Paul Valéry University Montpellier 3, Praxiling, CUFR de Mayotte
ahamadakassime5@gmail.com, fabrice.hirsch@univ-montp3.fr, miki.mori@univ-mayotte.fr

ABSTRACT

Kibushi is a principally oral language, which is spoken in Mayotte, a multilingual French island located between Madagascar and the Comoros Archipelago in the Mozambique Channel. The aim of this article is to study the two dialects of Kibushi: Kisakalava, used in a dozen villages with some twenty thousand speakers, and Kiantalautsi, spoken mostly in two villages, Ouangani and Poroani, by a few thousand speakers. 19 speakers in Kisakalava and 9 in Kiantalautsi were recorded producing 19 words. Results show a vowel space slightly different between the two dialects, as well as variation between villages.

Keywords: kibushi, variations, dialects, vowel space, stressed syllables.

1. INTRODUCTION

Mayotte is a multilingual French island located between Madagascar and the Comoros Archipelago in the Mozambique Channel. Its geographic location reflects its sociolinguistic landscape because there are two principal local languages: Shimaore, a Sabaki Bantu language used by the majority of the population, and Kibushi, an Austronesian language used only by 15% of Maore (people from Mayotte) [1]. In addition, the island was a French colony in the second half of the 18th century, before becoming a French department in 2011. Because of this, the French language is the only official language of administration and schooling.

Kibushi, mainly used orally, is understudied despite its apparent rich and complex variation. It is linguistically close to the Sakalava Malagasy dialect from the northwest of Madagascar [2]. There are two dialects with limited intercomprehensability: Kisakalava, used in a dozen villages with some twenty thousand speakers, and Kiantalautsi, spoken mostly in two villages, Ouangani and Poroani, by a few thousand speakers [1]. Intra-dialectal variation does occur, and it has been claimed that there is a "diversity" in the way Kibushi is spoken in Mayotte. [3]. To define these different ways to speak, Laroussi uses prosodic parameters: "intonation", "rhythm" and "accent" [3]. This prosodic variation depends on villages. In fact, Gueunier claims that each village is linguistically autonomous with its own accent and lexicon [4]. However, research demonstrating the existence of these prosodic and phonetic variations is lacking. Variation occurs mainly with regards to the lexicon and various phonetic, phonological, and prosodic levels. For example, the two dialects do not have the same word for "water," which is *ranu* in Kisakalava, but *mehetsaka* in Kiantalautsi.

This intriguing linguistic landscape is the focus of this pilot study which aims to explore variation in Kibushi based on two aspects: vowel space and stressed syllable characteristics. First, studying vowels is an interesting way to describe prosodic and phonetic variation of a language because unlike the consonants, their places of articulation are less stable [5]. It is unclear how vowel spaces vary by dialect and by village in Mayotte. Second, general observations of Kibushi suggest that there may be differences in word stress across villages and dialects. Specifically, there is one village, Chiconi, whose prosody has been described as "singing like" [3]. Research on Malagasy accent gives us some ideas to study stressed syllable characteristics in Kibushi. Indeed, to describe an "accented syllable" in Malagasy (Merina dialect), three acoustic parameters are used: intensity, duration, and pitch [6]. Duration (closely linked with pitch in Malagasy) and pitch have been shown to be by far the most reliable features for identifying a stressed syllable in Malagasy. That is, accent is associated with a syllable which contains higher pitch and longer duration [6].

Stress in both Kibushi dialects is penultimate, just as in the Malagasy language (Merina dialect). But it can also fall on the antepenultimate syllable with words that end in [ka], [t͡ra] and [na] [7]. About phones, Kibushi contains eight vowels (/a, ã, e, ẽ, i, o, õ, u/), two semi-vowels (/w, j/) and 30 consonants [1] [8]. Some of these phones are nonexistent in Malagasy Central dialects' phonology [9], such as the implosive consonants /ɓ/ and /ɗ/ and the semi-vowel /w/.

These observations led us to two research questions: 1. How do the two dialects and four villages differ in terms of vowel space? 2. How do the two dialects and four villages differ in terms of stressed syllable features, specifically duration, pitch, and intensity?

2. METHODS

Data from four villages were collected: Acoua and Chiconi, of Kisakalava dialect; and Ouangani and Poroani, of Kiantalautsi dialect. for Kisakalava, nine participants from Acoua of which four women, and 10 participants from Chiconi, of which nine women. For Kiantalautsi, nine participants were from Ouangani, of which 1 woman, and four were from Poroani, all men. Ages ranged from 12 to 24 years of age, with the average age of 16. Due to the voluntary nature of the project, age and gender are not representative. Participants were recorded in a quiet space using a Zoom H1N recorder with a sampling rate of 44,100 Hz. Audio files were cleaned for artifacts using Audacity before being processed via Praat [11]. Analyses and graphics were done using R [12]. In order to adjust for individual differences, including male versus female voices, frequency was converted from Hertz to semitones for each participant, using each participant's f0 averages (12*(log((zz)/100)) / log(2)). Formant normalization using Spectral Overlap Assessment Metric (SOAM) measurements was obtained using Wassink's VOIS3D [13].

The study elicited words using the 100-word Swadesh list [10]. 19 of them were analyzed in order to obtain the five vowels /a, i , u, e, o/, in stressed position: *'habu* (long) , *'hifi* (tooth), *'holu(ŋu)* (person), *'horu(ŋu)* (nose), *'lalaŋa* (path), *la'lahi* (man), *'lela(ka)* (tongue), *'liu* (blood), *'loku* (fish), *'luha* (head), *'maru* (a lot), *'meki* (dry), *'nonu* (breast), *'rohu* (heart), *'vatu* (stone), *'vava* (mouth), *'viti* (foot), *'voru(ŋu)* (bird), *'zahu* (I)

3. RESULTS

3.1. Vowel space variation

3.1.1. Vowel space by dialect

Figure 1 shows the vowel space by dialect. While vowels [a] and [o] are comparable in Kisakalava and Kiantalautsi, several differences are observed for [i], [e] and [u], specifically in terms of F2. This could be explained by prosodic differences between the two dialects, which can impact vowel pronunciation. As will be explained in Figure 2 of section 3.1.2, the anterior differences may stem from speakers in Chiconi rather than Acoua for Kisakalava. If differences are observed between the dialects, it is possible that contrasts are also notable between the villages where Kibushi is spoken.

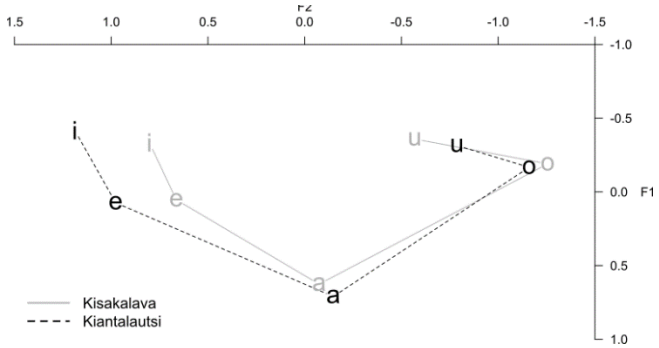


Figure 1: Vowel space by dialect (stressed and unstressed syllables), Lobanov normalization.

3.1.2. Vowel space by village

Variation by village can be seen in Figure 2. Indeed, differences are observed in the formant structure of Kibushi vowels among the four villages, again mainly in terms of F2. As can be seen, the Chiconi (Kisakalava dialect) vowel space is more constrained than the other villages, particularly for anterior vowels. In addition, while from two different dialects, the vowel spaces of Acoua (Kisakalava) and Ouangani (Kiantalautsi) are quite similar save for a slight difference with /u/.

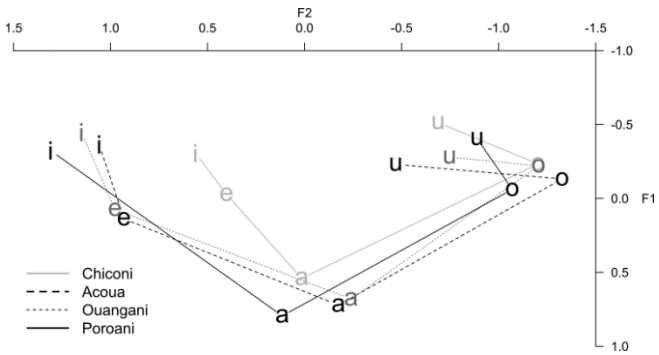


Figure 2: Vowel space by village (stressed and unstressed syllables), Lobanov normalization

3.2. Syllable stress variation

3.2.1. Stressed and unstressed syllable duration

Figure 3 shows the duration of stressed and unstressed syllables in Kibushi. They reveal that stressed syllables are systematically longer than the unstressed syllables. Indeed, the median duration of stressed syllables is 280 ms for Kibushi spoken in Acoua (*vs.* 182 ms for the unstressed syllables), 320 ms for Chiconi (*vs.* 185 ms), 287 ms for Ouangani (*vs.* 190 ms) and 270 ms for Poroani (*vs.* 179 ms).

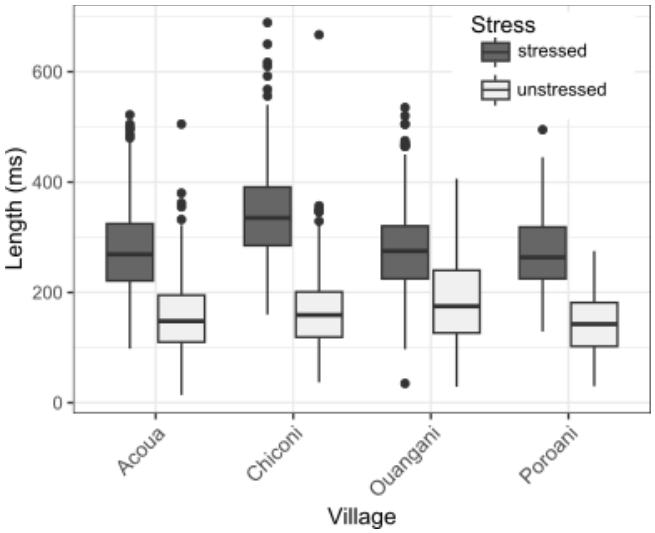


Figure 3: Syllable duration in ms by village.

The fact that the duration of the stressed syllables is longer can reveal prosodic phenomena, which is discussed in the sub-sections below which focus exclusively on stressed syllables.

3.2.2. f0 curves

Figure 4 shows SSANOVAs of the fundamental frequency throughout the stressed syllable. We observe a rise of the F0 on the second half of the syllable, and this for all villages. Variations between villages are also visible in Figure 5. While pitch for the Kiantalautsi villages (Poroani and Ouangani) are quite similar, those of the Kisakalava villages (Acoua and Chiconi) differ. In fact, pitch for Chiconi stressed syllables stand out from the rest of the villages throughout the syllable (consonant and vowel), in that the pitch rises more gradually. In addition, towards the end of the syllable, Acoua f0 begins to decrease compared to the other villages.

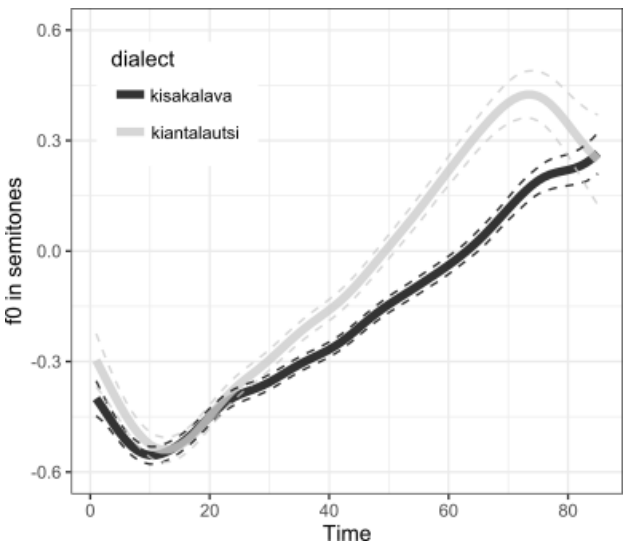


Figure 4: F0 semitones on stressed syllable by dialect

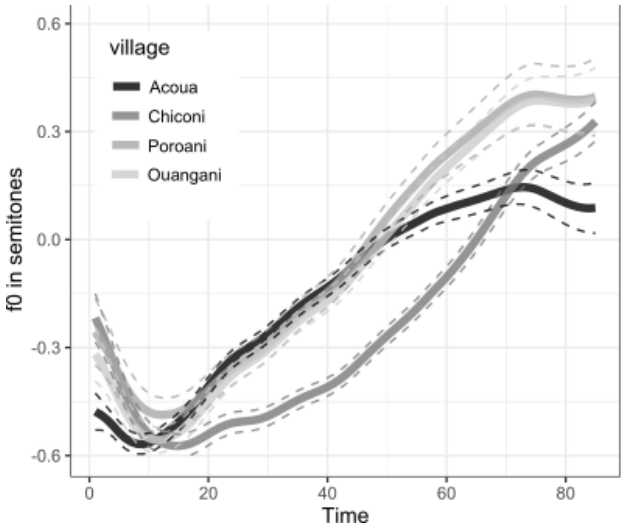


Figure 5: F0 semitones on stressed syllable by village

3.2.2. Intensity

A peak of intensity is noted on the accented syllables. As seen in Figure 6, this peak is located at about 40% of the syllable for Kiantalautsi, and slightly before that for Kisakalava. That is, intensity increases and decreases quicker in this dialect than for Kisakalava. As for village difference, as seen in Figure 7, villages vary quite a bit in terms of dB range and intensity curves, including location of peak intensity. Ouangani has the strongest stressed syllable intensity among the four villages, peaking around halfway through the syllable. Acoua's intensity peaks before that, and this intensity increases and decreases more sharply. Chiconi and Ouangani have similar curves, where intensity builds up steadily over the syllable. However, peak intensity of Ouangani occurs just after halfway, whereas that of Chiconi is near the end of the syllable, about ¾ of the way in.

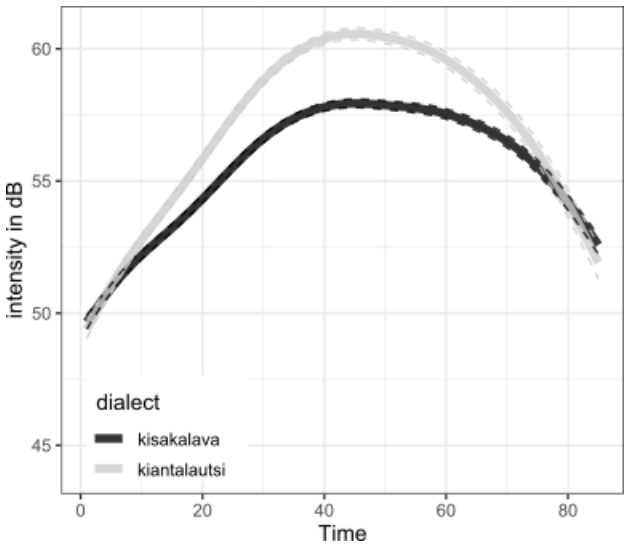


Figure 6: Intensity in dB stressed syllable by dialect

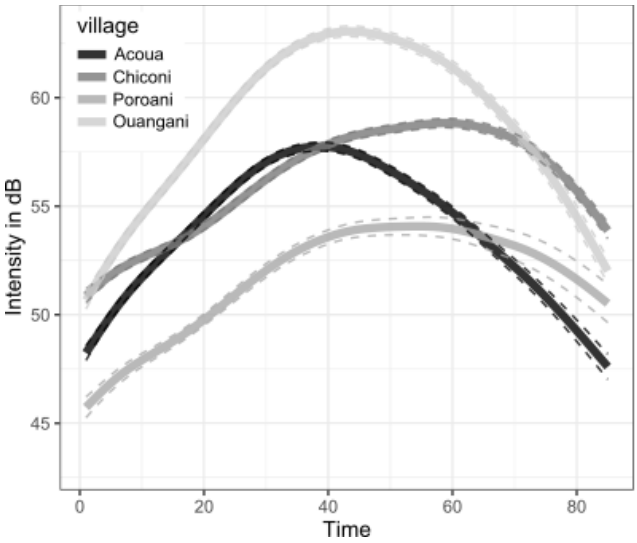


Figure 7: Intensity in dB on stressed syllable by village

4. DISCUSSION

The aim of this pilot study was to explore variation in Kibushi based on two aspects: vowel space and stressed syllable characteristics. Initial results show that Kibushi is subject to diatopic variation. Indeed, the formant structure of vowels is not totally identical from one dialect to another. Moreover, intra-dialectal variations are also noticeable. This indicates that the Kibushi dialects have evolved in slightly different ways. This observation could be explained by different migratory flows in each of the villages tested. Indeed, one can imagine that people who settled more recently on the island do not have the same linguistic origin, which is why vowel pronunciation varies from one village to another. Social networks and isolation between villages may also contribute to variation.

We also observed that the vowel space was larger for stressed syllables than for unstressed syllables. This result can be related to the length of these syllables which are longer than their unstressed counterparts. Indeed, we know that the formant structure of vowels depends on their duration: the longer the vowels are, the more the vowel space will be extended [14].

The other prosodic measures (F0 and intensity) also suggest marked variation. While looking at stressed syllable features by dialect is instructive, it appears to be just as important to look at differences by village. For example, Chiconi and Acoua, though part of the same dialect, differ in terms of f0 and intensity, with Chiconi standing out among the four villages. These findings support Guenier's [4] argument that village is the most important factor when understanding variation in Kibushi. That is, it does not appear to be enough to just address dialectal differences. Rather, when studying Kibushi, villages themselves must be considered. Further studies are needed to refine this observation.

5. REFERENCES

[1] Jamet, R. 2016. *Dictionnaire kibushi-français*, Dzoumogné, Association SHIME.

[2] Dominique B. 1976. *Le dialecte sakalava du nord-ouest de Madagascar : phonologie, grammaire, lexique*. Linguistic dissertation, Université de Paris V.

[3] Laroussi, F. 2010. *Langues, identités et insularité. Regards sur Mayotte*. Université de Rouen et du Havre.

[4] Gueunier, N. 2004. Le dialecte malgache de Mayotte (Comores) : une discussion dialectologique et sociolinguistique, *Faits de Langues,* 23-24, 397-420.

[5] Tanguy, S. 2017. Explorer la variation en breton grâce à la dialectométrie : la Basse-Bretagne considérée par la distance linguistique. *La Bretagne Linguistique, Centre de Recherche Bretonne et Celtique*, 21, 29-61.

[6] Raoniarisoa, N. 1990. *Accent and intonation in a Malagasy dialect*. Bangor University.

[7] Jake, A., Ileana, P. 2019. The Intonation of Malagasy: A Preliminary Look, *Proc. 19th ICPhS Melbourne,* 3792-3796.

[8] The Conseil Départemental de Mayotte. 2020.Shimaore and Kibushi alphabets. https://www.cg976.fr/ressources/dcp/?file=alphabets-langues-mahoraises

[9] Penelope, H. 2021. Central Malagasy. *Journal of the International Phonetic Association*, 51(1), 103-136.

[10] Swadesh, M. 1971. *The origin and diversification of language*. Aldine-Atherton. https://www.liquisearch.com/swadesh_list/original_final%20swadesh_list_1971283

[11] Boersma, P., Weenink, D. 2019. Praat: Doing phonetics by computer (version 6.1.01). http://www.praat.org

[12] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R project.org/

[13] Wassink, A. B. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three varieties, *Journal of the Acoustical Society of America,* 119(4), 2334-2350.

[14] Lindblom, B. 1963. Spectrographic study of vowel reduction. *The journal of the Acoustical society of America*, *35*(11), 1773-1781.

WORD POSITION AND STRESS COMBINE TO AFFECT VOICE ONSET TIME IN HAWAIIAN

Lisa Davidson[1] and ʻŌiwi Parker Jones[2]

New York University[1], University of Oxford[2]
lisa.davidson@nyu.edu, oiwi.parkerjones@eng.ox.ac.uk

ABSTRACT

Hawaiian, or ʻŌlelo Hawaiʻi, is an Eastern Polynesian language spoken on the islands of Hawaiʻi. This study examines voice onset time (VOT) in Hawaiian as produced by speakers on the 1970s-80s radio program Ka Leo Hawaiʻi. Hawaiian has only two primary oral stops /p k/ ([t] is a rare variant in these speakers' dialects). First, we show that /p/ and /k/ are unaspirated in this generation of Hawaiian speakers. Second, we examine whether syllable stress and word position affect VOT duration. Results show an interaction indicating that word initial position leads to longer VOT in unstressed and secondarily stressed syllables, but not in syllables with primary stress. In comparison to previous research on aspirated stops, which are often affected by prosodic prominence, and unaspirated stops, which are much less sensitive to prosody, stress and word position work together for a nuanced effect on the realization of VOT in Hawaiian.

Keywords: voice onset time, stops, Hawaiian, stress, word position

1. INTRODUCTION

Hawaiian, or ʻŌlelo Hawaiʻi, is an Eastern Polynesian language spoken on the islands of Hawaiʻi. After a precipitous decline in the intergenerational transmission and use of Hawaiian following the overthrow of the Hawaiian Kingdom and the banning of Hawaiian in schools in 1890s [1], a revitalization movement began in the 1970s. This movement has been successful in creating Hawaiian-medium schooling opportunities that have led to a new generation of native speakers of Hawaiian [2-4]. Around the same time that the Hawaiian-medium schooling began, a Hawaiian language radio program called Ka Leo Hawaiʻi (KLH) was broadcast that included on-air interviews with bilingual speakers from the last generation who had learned Hawaiian from birth from their families and communities [5]. This source of data has been used in previous studies of the phonetics of Hawaiian [6, 7], and is used in the current study.

The focus of this study is on the voice onset time (VOT) of Hawaiian, as it was produced by the KLH speakers in the 1970s-80s. Hawaiian contains 2 primary oral stops, /p/ and /k/ (as well a [t] variant of /k/ in a small number of words)[1], such as [papa] 'class', and [kokoke] 'near'. No instrumental study of VOT has previously been carried out for this generation of speakers, and previous impressionistic descriptions have not always agreed. Elbert and Pukui [8] comment that these stops are less aspirated than English, but Newbrand [9] observes that her speakers seem to produce aspiration at least word initially. A related finding is that Hawaiʻi Creole English (HCE) is reported to have VOT values that are significantly shorter than English aspirated stops [10], perhaps because HCE is influenced by Hawaiian.

The first goal of this study is to establish whether the Hawaiian spoken by the speakers who appeared on KLH contains aspirated or unaspirated stops. A related study on Māori found that greater use of English did correlate with longer VOT in Māori voiceless stops over time [11], so it may be the case that contact with English had already begun to lengthen VOT in Hawaiian among the generation of KLH speakers.

The second goal is to examine how the VOT of the single-series stops of Hawaiian interacts with stress and word boundaries. A typical finding for a language like English or (Jutland) Danish with (contrastive) aspirated stops is that VOT is longer in stressed position [12, 13], while [14] found an effect of phrase-initial position but not stress for stops that were unaccented in the phrase in English. For single series (non-contrastive) stops, the effect of prosodic variables on VOT seems to be minimal. For example, in Plains Cree, which has only voiceless unaspirated stops, syllable position (used as a proxy for either stress or pitch accent, which is still contested) is not significant [15]. Sierra Norte de Puebla Nahuatl, which has only one stop series that is consistently realized with positive VOT, word-initial vs. word-medial position has no consistent effect [16]. As for languages which contrast voiceless unaspirated (fortis) stops and voiced (lenis) stops, the role of prosodic prominence is less clear; prominence further shortens short lag VOT in Dutch [17], but has no effect in Spanish [18].

Expectations for the effect of stress and word position on Hawaiian VOT may depend on the degree of aspiration found for /p, k/, but the lack of contrast

with another stop series may also affect the findings. On one hand, if the KLH speakers produce stops in the aspirated range, then it may be expected that prosodic factors associated with prominence like syllable stress and word initial position would lead to longer VOT. If stops are instead unaspirated, then prominence could lead to shorter VOT, as in Dutch. Another possibility is that Hawaiian could show an intermediate level of aspiration, perhaps similar to HCE; previous research is agnostic about the prosodic effects on a "semi-aspirated" voiceless stop. On the other hand, since other languages without a contrast are generally insensitive to prosodic prominence for VOT length, Hawaiian may also follow this pattern too, regardless of the degree of VOT found for the stops.

2. METHOD

2.1. Participants

This analysis examines spontaneous speech collected from 7 Hawaiian elders (4M, 3F) who were interviewed on the radio program Ka Leo Hawai'i in the 1970s. These interviewees were bilingual speakers of Hawaiian and English, but the program was carried out entirely in Hawaiian.

2.2. Language background and materials

We report only on /p, k/, since use of [t] is very limited. Oral stops can appear in stressed and unstressed syllables in Hawaiian. The words in (1) illustrate the six possible combinations of heavy and light syllables that have been proposed for words that only have one stress in Hawaiian [19-21]. Syllables with diphthongs and long vowels are heavy, and all heavy syllables are stressed. If there are 2 or more light syllables at the end of the word, the penultimate syllable will contain stress [20, 22, 23]. Words containing more than 3 syllables or multiple heavy syllables will also have secondary stress, with primary stress always occurring on the rightmost stressed syllable. Examples of words with multiple stress are in (2) (this list of types is not exhaustive).

(1) a. ˈσL σL [ˈpa.ka] 'raindrops'
b. σLˈσL σL [pa.ˈke.ʔo] 'to project'
c. ˈσH [ˈkai] 'ocean', [ˈpo:] 'night'
d. σLˈσH [ku.ˈkui] 'candlenut tree'
e. ˈσH σL [ˈpai:.ka] 'fight'
f. σLˈσH σL [pa.ˈlao.a] 'bread'

(2) a. ˌσHˈσH [ˌpi:.ˈkoi] 'ache'
b. ˌσL σLˈσH [ˌpi.ha.ˈʔu:] 'jam-packed'
c. ˌσH ˈσL σL [ˌko:.ˈʔa.la] 'to broil'
d. ˌσL σLˈσL σL [ˌpa.ka.ˈla.na] 'Chinese violet'

2.3. Data analysis

Using episode transcripts and audio, each stop produced was segmented in Praat. Stops were not included if they were in a word containing an error or a hesitation. Each stop closure and burst was measured separately, as shown for [k, p] in Figure 1. The burst was marked as beginning where aperiodic noise started after the closure, and ended when F1 of the following vowel became visible.

The stops that were extracted from the KLH transcripts were annotated for word position and stress of the syllable using Parker Jones' [21] computational prosodic grammar of Hawaiian. For example, the grammar parses the word *pake'o* [pakeʔo] 'to project' as (WORD (PRWD (SYLL_L (CONSONANT p) (SHORT_VOWEL a)) (FOOT (SYLL_L (CONSONANT k) (SHORT_VOWEL e)) (SYLL_L (CONSONANT ʔ) (SHORT_VOWEL o))))). Regular expression matching was then used to detect whether /p/ or /k/ appeared in a primary stressed, secondary stressed, or unstressed syllable, or in a word-initial or word-medial syllable. For [pa.ˈke.ʔo], /p/ is word-initial, /k/ is word-medial, and /k/ is stressed (see [24] for details on the method). Word frequencies were also obtained from an independent Hawaiian text corpus of about 1.6M words [25].

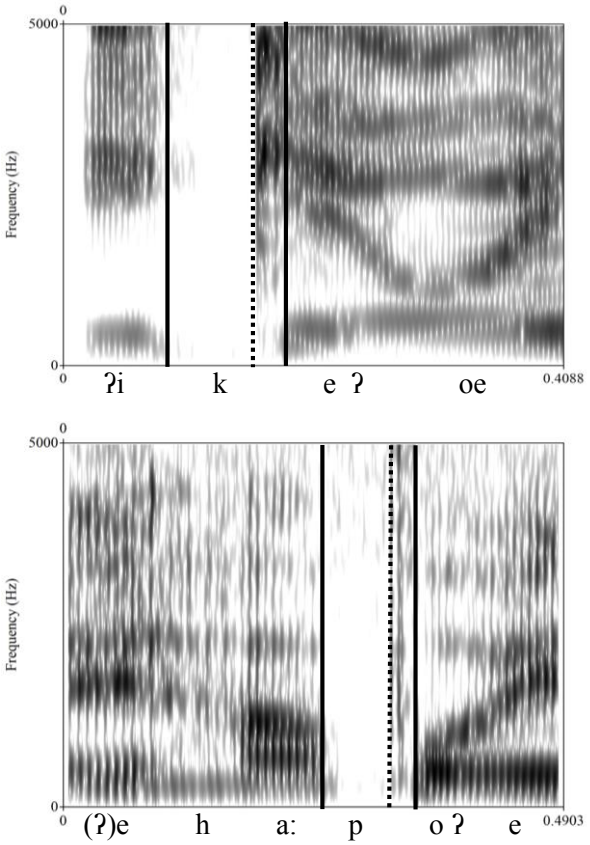


Figure 1: Multiple realizations of oral stops in [ʔike ʔoe] 'you see', 32ms, and [ʔeha: poʔe] 'four people', 23ms. Burst onset is demarcated with a dotted line.

In addition to the type of words in (1) and (2), Hawaiian also has 3 monosyllabic light syllable particles, [ka/ke] (allomorphs of the definite article) and [ko] (possessive marker), which are written as individual words in the orthography (e.g. *ka hale* 'the house'). These are typically unstressed, and are coded as such in this study. There are 5692 stops with a burst analysed in this study. In addition to the stops with closures and bursts, 1802 more are produced with frication, which will not be further examined here because of space limitations.

3. RESULTS

The histogram in Figure 2 indicates the range of values for VOT for /p/ and /k/ over all speakers and contexts in Hawaiian. The mean for /p/ is 24ms (s.d.=14ms, N=1547) and for /k/ is 39ms (s.d.=16ms, N=4145).

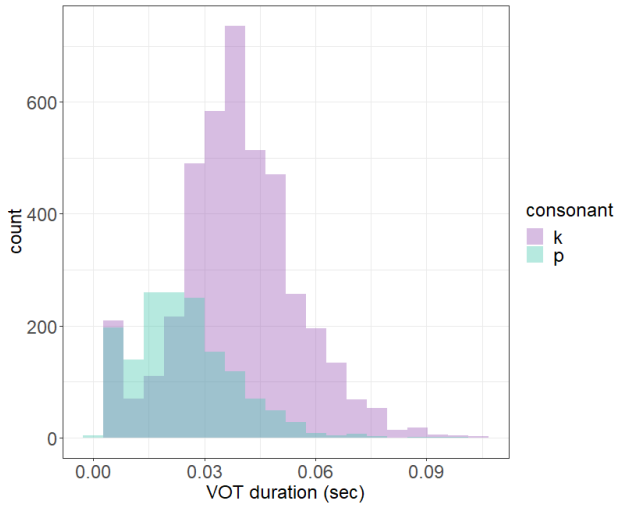


Figure 2: Histogram of VOT durations

The effects of consonant (/p, k/), word position (initial, medial), syllable stress (primary stressed, secondary stressed, unstressed), and lexical frequency (log transformed) on VOT duration are examined with a mixed effects linear regression in R. All of the categorical predictors were sum coded. A 3-way interaction between consonant, word position, and stress was included, as well as random intercepts for speaker and word. The results are in Table 1.

These results show that there is an effect of consonant: the 15ms difference between /k/ and /p/ is significant. Moreover, the histogram in Figure 2 indicates that some tokens do appear to have values that are in the range of aspirated stops. While individual differences are not the focus of this study, one concern is that these tokens may be attributable to one or two speakers. However, the plot of each of the talkers in Figure 3 demonstrates that despite some individual variability, all of the talkers are within the range of unaspirated stops, and that all of them produce some especially long tokens. This suggests that no individual speaker is producing most of the stops in the aspirated range.

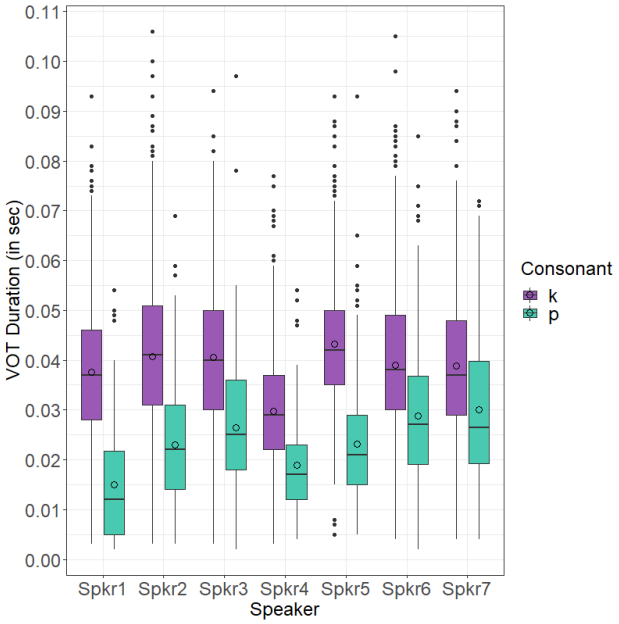


Figure 3: Average VOT durations for individual speakers

	Estim	Df	t val.	Pr(>|t|)
(Intercept)	0.0385	14.5	23.848	**<0.001**
consonant(p)	-0.0153	954.5	-14.05	**<0.001**
wd_pos(initial)	0.0017	1322	3.067	**0.002**
stress(primary)	0.0013	567.9	1.769	0.08
stress(secondary)	-0.0001	771.2	-0.103	0.92
log_word_freq	-0.0004	3501	-1.72	0.09
consonant(p): wd_pos(init)	-0.0003	1015	-0.256	0.79
consonant(p): stress(primary)	-0.0003	993.5	-0.282	0.78
consonant(p): stress(secondary)	0.0014	1515	0.898	0.37
wd_pos(init): stress(primary)	-0.0014	1435	-2.302	**0.021**
wd_pos(init): stress(secondary)	-0.0008	1654	-0.924	0.36
cons(p):wd_pos (init):str(prim)	0.0005	1058	0.413	0.68

Table 1: Linear mixed effects regression. Bold indicates a significant result.

There is no main effect for the levels of stress. Word position is significant, though the overall difference between the averages is small (word initial: 36ms, word medial: 33ms). There is no effect of word frequency. The negative interaction between word position and stress for primary stressed syllables occurs because there is a smaller

difference between word initial and word medial position than there is for secondary and unstressed syllables, both of which reflect the main effect of stress with longer VOT in word initial position. This effect is illustrated in Figure 4.

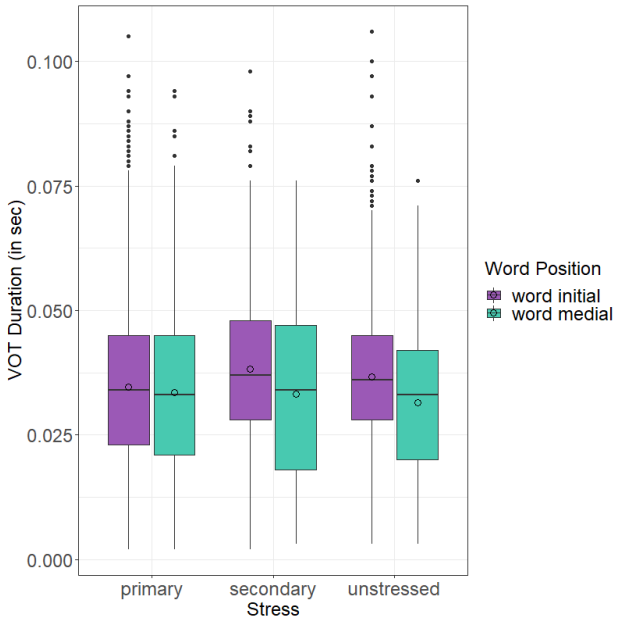


Figure 4: VOT durations by stress level (primary, secondary, unstressed) and word position (word-initial, word-medial)

4. DISCUSSION

Results indicate that for the generation of Hawaiian speakers who appeared as guests on Ka Leo Hawai‘i, the oral stops /p, k/ can be considered unaspirated. The values found here can be compared to previous cross-linguistic surveys [26, 27], which together report VOT values in more than 20 languages. The bilabial stops measured in [26, 27] range from 10-22ms, so while the Hawaiian average value of 24ms for [p] is on the high end of the range, it is consistent with other languages. As for velars, the Hawaiian value of 39ms for [k] fits squarely within the range of 27-56ms reported in these papers. Though we have not measured English speech for these same speakers, a recent study examining the VOT duration of aspirated stops in American English from 600 talkers reported a mean of 51ms for [p^h] and 54ms for [k^h] (across all contexts and averaged over all stress types) [28]. While only z-scored comparisons are reported in [29] for different contexts, VOT of stressed syllables is longer than unstressed syllables in English.

One limitation of this study is that we did not account for broader intonational patterns or prosodic elements like focus or accent, since this data comes from a corpus of spontaneous speech, and an intonational grammar of Hawaiian has not yet been developed. Impressionistic examination of the interviews suggests that especially long VOTs may be attributable to phenomena like focused utterances and animated speech, but a full accounting of these tokens awaits future research.

Previous research has shown two outcomes for the effect of prosodic variables on unaspirated stops. In Spanish, which has a voicing contrast, and in Plains Cree, which does not, stress had no effect on the duration of VOT [15, 18]. In Dutch, prosodically stronger positions including syllable stress, phrasal accent, and prosodic boundary strength all led unaspirated stops to have shorter VOTs than stops in weaker positions [17]. The authors concluded that in Dutch, VOT shortening is a fortition process that occurs to demarcate stronger prosodic positions.

The word position and stress results for Hawaiian show some significant effects, though their magnitude is small. Overall word initial VOT is 3ms longer, and the main effect of stress is not significant. The interaction of word position and stress indicates that longer word initial VOT is attributable to secondary and unstressed syllables (5ms difference), while the difference between word initial and medial positions for primary stress is nearly non-existent (1ms). Moreover, the numerical values for word-initial VOT in secondary and unstressed syllables are higher than for primary stress (Figure 4). This finding is not consistent with the literature for aspirated stops (e.g., [12, 29]), which usually reports that word initial position or primary stress lead to increased VOT duration (though in Scottish English spontaneous speech, for example, the word position difference for voiceless stops in stressed syllables is not significant [30]). If the effect for Hawaiian stops in secondary and unstressed syllables is perceptible, perhaps it could be an acoustic cue that a less prominent syllable belongs to the beginning of a word, instead of being the final unstressed syllable of a preceding word.

The picture that emerges for the interacting effects of stress and word position on Hawaiian VOT is more nuanced than most previous results for either aspirated or unaspirated stops. While the values are small, a 5ms difference is approximately the just noticeable difference for a 30ms burst [31]. Future work should examine whether this acoustic information is accessible for speech segmentation, as well as how it interacts with other cues to stress levels in Hawaiian.

5. ACKNOWLEDGMENTS

Thanks to Aidan Katson and JoJo Yang for their assistance with segmentation in Praat.

6. REFERENCES

[1] Wilson, W. 2013. Assessing Hawaiian. In: Kunnan, A. J. (ed.), *The Companion to Language Assessment* (pp. 1748-1758). Malden, MA: Wiley.

[2] Wilson, W., and Kamanā, K. 2001. Mai Loko Mai O ka ʻIʻini: Proceeding from a dream. In: Hinton, L. and Hale, K. (eds.), *The Green Book of Language Revitalization in Practice* (pp. 147-176). Leiden: Brill.

[3] Warner, S. L. N. E. 2001. The Green Book of Language Revitalization in Practice. In: Hinton, L. and Hale, K. (eds.), *The Movement to Revitalize Hawaiian Language and Culture* (pp. 133-146): Brill.

[4] Kawai'ae'a, K., Houseman, A. K., and Alencastre, M. 2007. Pūʻā i ka ʻŌlelo, Ola ka ʻOhana: Three generations of Hawaiian language revitalization, *Hūlili: Multidisciplinary Research on Hawaiian Well-Being*. 4: 183-237.

[5] Kimura, L. L. K. 2020. Kaniʻāina, ulukau.org Digital Repository of Ka Haka ʻUla O Keʻelikōlani College of Hawaiian Language, University of Hawaiʻi at Hilo, Ka Leo Hawaiʻi. 2020.

[6] Davidson, L. 2021. Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian, *J. Phon.* 88, 101075.

[7] Kettig, T. 2021. Haʻina ʻia mai ana ka puana: The vowels of ʻōlelo Hawaiʻi. PhD dissertation, University of Hawaii at Mānoa.

[8] Elbert, S., and Pukui, M. K., *Hawaiian Grammar*, Honolulu: University of Hawaii Press, 1979.

[9] Newbrand, H. L. 1951. A phonemic analysis of Hawaiian. MA thesis, University of Hawaiʻi at Mānoa.

[10] Drager, K., Grama, J., Gonzalez, L. *et al.* 2016. Voice onset time and closure duration of word-initial voiceless plosives in Pidgin conversation, . *J. Acoust. Soc. Am.* 140: 4, 3052-3052.

[11] Maclagan, M., and King, J. 2007. Aspiration of plosives in Māori: Change over time, *Australian Journal of Linguistics*. 27: 1, 81-96.

[12] Lisker, L., and Abramson, A. S. 1967. Some effects of context on voice onset time in English stops, *Lang. Speech*. 10: 1, 1-28.

[13] Puggaard, R. 2019. Modeling regional variation in voice onset time of Jutlandic varieties of Danish. In: Van de Velde, H., Haug Hilton, N. and Knooihuizen, R. (eds.), *Selected papers from the Tenth International Conference on Language Variation in Europe* (pp. 79-110). Amsterdam: John Benjamins.

[14] Cho, T., and Keating, P. 2009. Effects of initial position versus prominence in English, *J. Phon.* 37: 4, 466-485.

[15] Hodgson, M. 2021. Voice onset time in Plains Cree. MA Thesis, University of Calgary.

[16] Kakadelis, S. 2018. Phonetic Properties of Oral Stops in Three Languages with No Voicing Distinction. PhD Dissertation, CUNY Graduate Center.

[17] Cho, T., and McQueen, J. M. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress, *J. Phon.* 33: 2, 121-157.

[18] Simonet, M., Casillas, J., and Diaz, Y. 2014. The effects of stress/accent on VOT depend on language (English, Spanish), consonant (/d/, /t/) and linguistic experience (monolinguals, bilinguals). In: Campbell, F., Gibson, M. and Hirst, D. (eds.), *Proceedings of Speech Prosody 7* (pp. 202-206).

[19] Schütz, A. 2010. Measures and morphemes: A functional approach to Hawaiian accent. In: Bowden, J., Himmelmann, N. P. and Ross, M. (eds.), *A journey through Austronesian and Papuan linguistic and cultural space: Papers in honor of Andrew Pawley* (pp. 405-422). Canberra: Pacific Linguistics.

[20] Schütz, A. 1981. A reanalysis of the Hawaiian vowel system, *Oceanic Linguistics*. 20: 1, 1-43.

[21] Parker Jones, O. 2010. A computational phonology and morphology of Hawaiian. DPhil. thesis, University of Oxford

[22] Parker Jones, O. 2018. Hawaiian, *J. Int. Phon. Assoc.* 48: 1, 103-115.

[23] Schütz, A. 1978. Accent in two Oceanic languages, *Anthropological Linguistics*. 20: 4, 141–149.

[24] Anonymous. submitted. Prosodic effects on Hawaiian glottal stop realization, *Phonetica*.

[25] Parker Jones, O., and Shillingford, B., 2018. Composing RNNs and FSTs for small data: Recovering missing characters in old Hawaiian text, in 32nd Annual Conference on Neural Information Processing Systems IRASL Workshop, Montreal.

[26] Cho, T., and Ladefoged, P. 1999. Variation and universals in VOT: evidence from 18 languages, *J. Phon.* 27, 207-229.

[27] Cho, T., Whalen, D. H., and Docherty, G. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages, *J. Phon.* 72, 52-65.

[28] Chodroff, E., Godfrey, J., Khudanpur, S. *et al.* 2015. Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow.

[29] Sprinkle, M. 2022. In search of phonetic evidence for prosodically-motivated aspiration. BA Thesis, Linguistics, College of William and Mary.

[30] Stuart-Smith, J., Sonderegger, M., Rathcke, T. *et al.* 2015. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian, *Lab. Phon.* 6, 505-549.

[31] Stevens, K., *Acoustic Phonetics*, Cambridge, MA: MIT Press, 1998.

[1] While [t] is rare in the dialects of the KLH speakers, it does occur more commonly in the dialect of Niʻihau, a small island that maintained a continuously speaking community during the 20th century to today.

ACOUSTIC CORRELATES OF METRICAL FEET IN TWO YUPIK LANGUAGES

McKinley Alden, Anja Arnhold

University of Alberta
malden@ualberta.ca, arnhold@ualberta.ca

ABSTRACT

Yupik languages are known for their complex iambic stress systems. Yup'ik (Central Alaskan dialect) and Alutiiq (Chugach dialect) both generate feet iteratively from left to right and distinguish stressed long, stressed short, and unstressed short vowels. Yup'ik only expresses metrical structure via stress. Alutiiq, however, has several systems that complicate metrical expression, often associated with its ability to form ternary feet amid binary sequences. These include stress, tone, and two types of lengthening (affecting binary or ternary foot heads), which are complicated by length neutralization.

Measurements of 2,238 and 2,235 vowels, respectively, from narrative recordings showed that both languages use duration, intensity, and maximum f0 to differentiate stressed and unstressed syllables while preserving phonemic length contrasts. Alutiiq prioritizes these contrasts even when the syllable is obscured by overlapping metrical processes. The results confirm descriptions of these languages' complex metrical systems.

Keywords: Prosodic typology, stress, acoustics, North American languages, speech rhythm

1. INTRODUCTION

In this paper, we contrast the acoustics of metrical expression in the Central Alaskan dialect of Yup'ik (henceforth *Yup'ik*) and the Chugach dialect of Alutiiq (henceforth *Alutiiq*), focusing on the acoustic correlates of stress. Yup'ik and Alutiiq are spoken in southwestern Alaska. They are sisters in the Yupik branch of the Inuit-Yupik-Unangan family, and are highly lexically and grammatically similar with a high degree of mutual intelligibility.

Closely related as they are, Yup'ik and Alutiiq share similar metrical systems, though with some crucial differences (see [1]–[4] and [5]–[10] on Yup'ik and Alutiiq prosody, respectively, as well as [11], [12] for a comparison). Both build iambic feet iteratively from left to right, e.g. Yup'ik *nalluyagucaqunaku* 'don't forget it' is footed (na.ˈɬu).(ja.ˈgu).(ɮa.ˈqu).(na.ˈku). Both have been described as non-culminative, such that there are no primary-secondary distinctions among stresses within the same word. Both further feature phonemic vowel quantity and quantity-sensitive stress, though which syllables are considered heavy differs between the two languages. In Yup'ik, all syllables containing phonemically long vowels are heavy (and thereby stressed), and closed syllables are always heavy word-initially. Elsewhere, a (C)VC syllable may or may not behave as heavy, e.g. *ayagtuq* [a.ˈjax.tuq] 'he leaves' vs. *up'nerkaq* [ˈup.nəχ.ˈkaq] 'spring'. In Alutiiq, (C)VC syllables are only heavy word-initially.

The most notable prosodic difference, however, is that where Yup'ik builds only binary iambic feet, Alutiiq can form ternary feet amid sequences of binary feet. That is, in Alutiiq sequences of light syllables, some appear to be 'skipped' in the assignment of binary feet [5], [10], [13], [14]. Martinez-Paricio & Kager [9] model this with the construction of Internally Layered Ternary (ILT) feet, e.g. ((ma.ˈŋaʁ).*su*).(qu.ˈta).(qu.ˈni)) 'if he (refl) is going to hunt porpoise', where the third syllable is an adjunct to a nested minimal iambic foot, so that the first three syllables are all part of an ILT foot. ILT foot heads, like the second syllable in the preceding example, receive 'additional lengthening' [5], [9]. Furthermore, where Yup'ik only expresses its metrical structure via stress, Alutiiq additionally uses metrically conditioned tone and foot-initial onset consonant fortition to express metrical structure [5], [10], [15]. Alutiiq also has a compression rule that neutralizes length contrasts. In closed and word-final syllables, this compression neutralizes the difference between an underlyingly long or stress-lengthened vowel and an underlyingly short vowel according to Leer [5].

Despite these detailed descriptions of the two prosodic systems, to date there has been little acoustical evidence to support these detailed descriptions (the only published acoustical study being [16]). Here, we present first results of a study exploring and comparing the acoustics of metrical expression in Yup'ik and Alutiiq. For Yup'ik, this entailed a study on the acoustics of stress production. For Alutiiq, a series of analyses explored the effects of stress and compression.

2. MATERIALS AND METHODS

This project analyzed six recordings of Central Alaskan Yup'ik [17], [18] and seven recordings of

Chugach Alutiiq [19]. All the recordings are housed in the Alaskan Native Language Archive [20]. Each recording was transcribed in Praat [21] at the segment level, with tiers for the intonational phrase (IP), the word, the syllable, and the segment. A Praat script [22] took measurements of duration (in ms), maximum fundamental frequency (f0) (in semitones relative to a baseline of 100 Hz, henceforth ST), and intensity (in dB; all recordings were first scaled to 70 dB) for all vowels in the data corpus.

The Yup'ik recordings contained 2,282 vowels and the Alutiiq recordings 2,320 vowels in total. For statistical modelling, words ten syllables (in Yup'ik) and nine syllables (in Alutiiq) or longer, which were rare, were removed. As a result, there were 2,238 vowels in the final dataset for Yup'ik and 2,235 for Alutiiq. Among the Yup'ik vowels, 506 were long (and thereby obligatorily stressed), 785 were short and stressed, and 947 were short and unstressed. For Alutiiq, 286 vowels were long (and obligatorily stressed), 891 were short and stressed, and 1058 were short and unstressed. For f0, measurements were occasionally not possible due to excessive creak, noise, poor recording quality, et cetera, so that the datasets used for the f0 models contained 2,164 Yup'ik and 2,072 Alutiiq vowels.

Analysis was done separately for each acoustic measure via linear mixed-effect models using the package lme4 in R [23]–[25]. Model selection began with a model containing only the stress-length distinction (stressed long, stressed short, and unstressed short), as a fixed effect, with file, speaker, and phoneme as the random intercepts. Further fixed effects were added only if they significantly improved model fit [23]. While complex random effects were tested, they either did not improve the fit of any of the models or caused convergence issues. P-values for fixed effects were obtained with the package lmerTest [26]. For models containing significant interactions, pairwise comparisons were performed with the lsmeans function from the package emmeans [27], using the Tukey method for p-value adjustments and the (default) Kenward-Roger method for calculating degrees of freedom.

For the sake of brevity, below we list all fixed effects included in the best-fitting model of each dependent variable, but only describe the effects of the main factor of interest, the stress-length distinction (and, for Alutiiq, compression).

In addition to individual models for each language, we also computed a model of each measure in a combined dataset containing both Yup'ik and Alutiiq data. These models were constructed specifically for direct comparison of the effects of the stress-length distinction in the two languages and therefore only contained an interaction between the stress-length distinction and language as fixed effects, and file and phoneme as random effects.

3. RESULTS

3.1. Duration

For the Yup'ik data, the best statistical model contained the stress-length distinction, word length and distance from the left edge of the word (in number of syllables), whether the vowel was in an IP-final syllable, the presence of an onset, and the presence of a coda as fixed effects. The results showed that long (obligatorily stressed) vowels were significantly longer than stressed short vowels ($\beta = 86.6617$; $SE = 24.027$; $df = 3.0165$; $t = 3.607$; $p = 0.03626$) and that unstressed short vowels were significantly shorter than stressed short vowels ($\beta = -32.969$; $SE = 2.5207$; $df = 2147.1971$; $t = -1.079$; $p = <2.00e-16$).

The Alutiiq model contained distance from the left edge of the word, IP finality, the presence of an onset, and the presence of a coda as fixed effects in addition to the stress-length distinction. Stress and length were both distinguished by duration: stressed long vowels were significantly longer than stressed short vowels ($\beta = 32.7295$; $SE = 1.9358$; $df = 2172.705$; $t = 16.9077$; $p = <2e-16$), and unstressed short vowels were significantly shorter than stressed short vowels ($\beta = -12.7909$; $SE = 1.3765$; $df = 2172.1648$; $t = -9.292$; $p = <2.00e-16$).

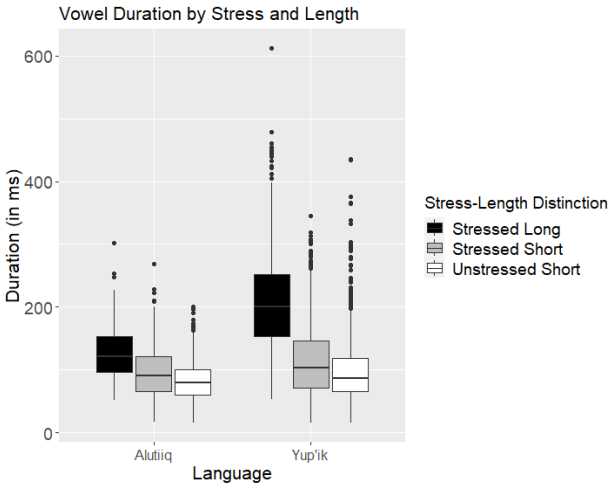


Figure 1: Boxplot of the duration of vowels by the stress-length distinction in Alutiiq and Yup'ik.

Thus, the duration linear mixed-effects models showed that, within each language, the stress-length categories were distinguished by their durations. However, the model directly comparing the two languages found the interaction between the stress-length distinction and the language to be significant. This means that the differences between the stress-length categories were not the same across the two

languages, as illustrated in Figure 1. The difference between the mean duration of stressed long and stressed short vowels in Yup'ik was 96.91 ms, while the difference in Alutiiq was notably shorter at 32.64 ms, with Yup'ik having considerably longer long vowels than Alutiiq. The difference between mean duration in stressed and unstressed short vowels was more similar across the languages (15.87 ms in Yup'ik, 13.19 ms in Alutiiq).

Follow-up pairwise comparisons showed that stressed long vowels were significantly shorter in Alutiiq than in Yupik (β = -85.4; *SE* = 3.68; *df* = 4468; *t* = -23.204; *p* = <.0001). This pattern also held true for stressed short (β = -20.8; *SE* = 2.51; *df* = 4073; *t* = -8.296; *p* = <.0001) and unstressed short vowels (β = -21.5; *SE* = 2.34; *df* = 3893; *t* = -9.176; *p* = <.0001).

Finally, to test the effect of compression in Alutiiq, which is described as neutralizing length and stress distinctions by shortening vowels in stressed closed and word-final syllables, we fit a linear mixed-effects model containing only a five-level factor coding distinctions in stress, length, and compression as a fixed effect (stressed long compressed, stressed long uncompressed, stressed short compressed, stressed short uncompressed, unstressed short). It showed the durational differences between all these categories to be significant. Firstly, stressed long uncompressed vowels were longer than stressed long compressed vowels (β = 26.239; *SE* = 3.587; *df* = 2193.156; *t* = 7.315; *p* = 3.60e-13) and stressed short uncompressed vowels were longer than stressed short compressed vowels (β = 26.717; *SE* = 2.291; *df* = 2194.104; *t* = 11.664; *p* = <2e-16). This indicates that compression did work to shorten long and lengthened vowels. Secondly, compressed long vowels were significantly longer than compressed short vowels (β = 20.896; *SE* = 2.889; *df* = 2193.671; *t* = 7.233; *p* = 6.49e-13), indicating that length distinctions were still preserved even when the vowels were compressed. Lastly, unstressed short vowels were significantly shorter than both compressed long vowels (β = -25.722; *SE* = 2.788; *df* = 2193.424; *t* = -9.227; *p* = <2e-16) and compressed short vowels (β = -4.826; *SE* = 1.500; *df* = 2193.456; *t* = -3.218; *p* = 0.00131). This shows that the effect was not fully neutralizing—that is, that compression did not reduce a long or lengthened vowel to the duration of an unstressed short vowel, and furthermore, that length distinctions are preserved even in compression.

3.2. Intensity

For Yup'ik, the best model of intensity contained the stress-length distinction, distance from the left edge of the word, IP finality, the presence of an onset, and the presence of a coda as fixed effects. It showed that stressed long vowels were significantly louder than stressed short vowels (β = 1.2004; *SE* = 0.221; *df* = 2170.4585; *t* = 5.431; *p* = 6.21e-08), and unstressed short vowels were significantly quieter than stressed short vowels (β = -1.3019; *SE* = 0.1949; *df* = 2171.8781; *t* = -6.68; *p* = 3.03e-11). Relevelling also confirmed that unstressed short vowels had significantly lower intensities than stressed long vowels (β = -1.2004; *SE* = 0.2210; *df* = 2170.46; *t* = -5.431; *p* = 2.61e-08). Like duration, then, intensity was associated with the production of both long vowels and stressed vowels.

Regarding Alutiiq, the model also contained the stress-length distinction, distance from the left edge of the word, IP finality, the presence of an onset, and the presence of a coda as fixed effects. Stressed long vowels were significantly louder than stressed short vowels (β = 1.9127; *SE* = 0.2572; *df* = 2187.9845; *t* = 7.4373; *p* = 1.47e-13), and unstressed short vowels were significantly quieter than stressed short vowels (β = -1.3545; *SE* = 0.1863; *df* = 2184.5074; *t* = -7.2519; *p* = 5.68e-13).

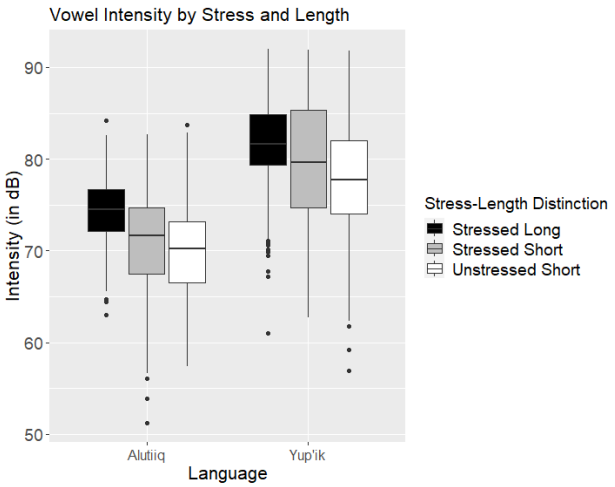


Figure 2: Boxplot of the intensity of vowels by the stress-length distinction in Alutiiq and Yup'ik.

In the model of data from both languages, the interaction between the stress-length distinction and language was significant, indicating that the differences between the categories were not the same across the two languages. Still, the difference in mean intensity between a stressed long and short vowel was 2.19 dB in Yup'ik and 3.15 dB in Alutiiq, while the difference between a stressed and unstressed short vowel was 1.78 dB in Yup'ik and 1.03 dB in Alutiiq. In short, the differences for intensity between the stress-length categories were more consistent across the two languages than the differences for duration. In Figure 2, however, Alutiiq intensity values appear overall lower than those of Yup'ik. Pairwise comparisons confirmed this for stressed long vowels (β = -9.1; *SE* = 0.332; *df* = 4502; *t* = -27.360; *p* =

<.0001), stressed short vowels (β = -10.1; *SE* = 0.227; *df* = 4505; *t* = -44.374; *p* = <.0001), and unstressed short vowels (β = -9.7; *SE* = 0.212; *df* = 4506; *t* = -45.830; *p* = <.0001).

3.3. F0

The maximum f0 model for Yup'ik contained only the stress-length distinction and IP finality as fixed effects. The difference between a stressed long vowel and a stressed short vowel was not significant (β = -0.65611; *SE*= 0.51476; *df* = 3.02931; *t* = -1.275; *p* = 0.291437). The difference between a stressed and unstressed short vowel was, with unstressed short vowels being lower (β = -1.51746; *SE* = 0.157; *df* = 2077.69558; *t* = -9.665; *p* = <2.00e-16). This indicates that f0, unlike duration and intensity, was only associated with stress and not with length.

For Alutiiq, the model contained the stress-length distinction, distance from the left edge and the presence of a coda as fixed effects. Stressed long vowels were significantly higher in f0 than stressed short vowels (β = 1.0538; *SE* = 0.1938; *df* = 2016.7382; *t* = 5.4376; *p* = 6.05e-08), and unstressed short vowels were significantly lower than stressed short vowels (β = -0.4755; *SE* = 0.1416; *df* = 2008.265; *t* = -3.3586; *p* = 8.00e-04).

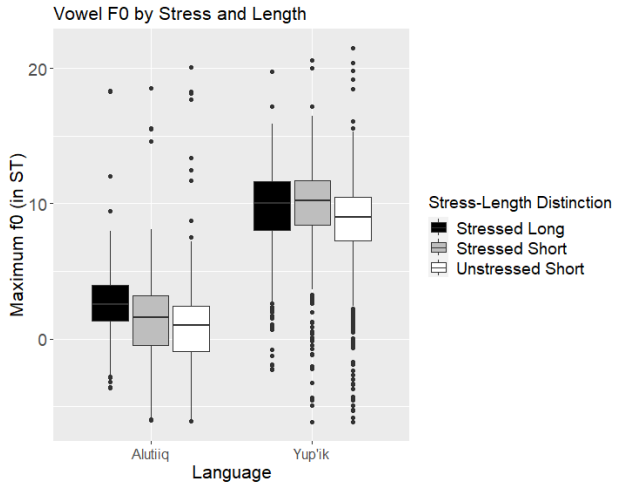


Figure 3: Boxplot of the maximum f0 of vowels by the stress-length distinction in Alutiiq and Yup'ik.

Figure 3 resembles the intensity figure more than the duration figure. However, in Yup'ik the non-significant difference in f0 between long and short stressed vowels was -0.04 ST, where the stressed short mean was slightly higher, while the difference between the same two categories in Alutiiq was larger, at 1.2 ST. Meanwhile, the difference between a Yup'ik stressed and unstressed short vowel was 1.47 ST, while the same difference in Alutiiq was only 0.52 ST. Modelling confirmed a significant interaction between the stress-length categories and language for f0. Pairwise comparisons showed that, like duration and intensity, Alutiiq vowels were significantly lower than their Yup'ik counterparts (stressed long: β = -6.52; *SE* = 0.231; *df* = 4271; *t* = -28.224; *p* = <.0001; stressed short: β = -7.70; *SE* = 0.161; *df* = 4253; *t* = -47.78; *p* = <.0001; unstressed short: β = -6.91; *SE* = 0.152; *df* = 4244; *t* = -45.606; *p* = <.0001).

4. DISCUSSION AND CONCLUSION

Table 1 summarizes the results of the acoustical analyses of Yup'ik and Alutiiq stress. Interestingly, despite the ways in which the two languages' metrical systems diverge, they largely express stress and phonemic length via the same acoustic correlates. There are, however, some differences. Notably, unlike Alutiiq, Yup'ik does not reliably mark length via f0. In turn, the durational difference between long and short stressed vowels was much larger in Yupik than in Alutiiq. This is likely because of duration's high functional load in Alutiiq: not only does it communicate phonemic length and stress information, but it is also affected by compression and, according to the literature, ILT head lengthening.

Table 1: Summary of results: acoustic correlates of stress and phonemic vowel length.

	Stress Correlates	Length Correlates
Yup'ik	Duration, Intensity, Maximum f0	Duration, Intensity
Alutiiq	Duration, Intensity, Maximum f0	Duration, Intensity, Maximum f0

Nevertheless, both languages consistently distinguish three types of vowel: long, obligatorily stressed vowels and short vowels that may be either stressed or unstressed. These distinctions are preserved even in cases of compression, contra previous descriptions of Alutiiq [5]. Our results are in line with a previous study of vowel duration in Central Siberian (St. Lawrence Island) Yupik [28], which found similar differences between long, stressed short and unstressed short vowels. Together, these studies show that the ternary stress-length distinction appears to be a stable factor in Yupik languages, despite the diversity of metrical systems.

Many questions remain surrounding the Yup'ik and Alutiiq metrical systems. This study represents the first acoustical examination and comparison of metrical expression in Yup'ik and Alutiiq, opening the door for further investigation into these intricate systems. Future work will determine acoustic correlates of Alutiiq additional lengthening and tone.

5. REFERENCES

[1] S. Jacobson, "Siberian Yupik and Central Yupik prosody," in *Yupik Eskimo prosodic systems: descriptive and comparative studies*, 1985, pp. 25–45.

[2] S. Jacobson and A. Jacobson, *A practical grammar of the Central Alaskan Yup'ik Eskimo language: with Yup'ik readings*. Fairbanks, AK: Alaska Native Language Center, 1995.

[3] A. C. Woodbury, "The postlexical prosody of Central Alaskan Yup'ik," Austin, TX, 1995.

[4] M. Mithun, "Prosody, grammar, and discourse in central Alaskan Yup'ik," p. 222, 1996.

[5] J. Leer, "Prosody in Alutiiq," in *Yupik Eskimo prosodic systems: descriptive and comparative studies*, 7th ed., M. Krauss, Ed. Fairbanks, AK: Alaskan Native Language Center, 1985, pp. 77–134.

[6] M. Hewitt and J. Ann, "Binarity and ternarity in Alutiiq," in *Arizona Phonology Conference Vol. 4*, 1991.

[7] M. Hewitt, "Deconstructing foot binarity in Koniag Alutiiq." University of British Columbia, Vancouver, BC, Feb. 1994.

[8] C. Rice, "Stress assignment in the Chugach dialect of Alutiiq," in *Proceedings of the Chicago Linguistic Society*, 1988, pp. 304–315.

[9] V. Martínez-Paricio and R. Kager, "Metrically conditioned pitch and layered feet in Chugach Alutiiq," *Loquens*, vol. 3, no. 2, 2017, doi: 10.3989/loquens.2016.030.

[10] J. Leer, "The Phonology of the Kenai Peninsula Dialect of Chugach Alutiiq." Alaska Native Language Center, Fairbanks, AK, 1994.

[11] A. C. Woodbury, "Eskimo and Aleut languages," in *Handbook of North American Indians*, vol. 5, D. Damas, Ed. Smithsonian Institution, 1984, pp. 49–63.

[12] B. Hayes, *Metrical stress theory: principles and case studies*. 1995.

[13] J. Leer, "Evolution of prosody in the Yupik languages," in *Yupik Eskimo prosodic systems: descriptive and comparative studies*, vol. 7, M. Krauss, Ed. Fairbanks, AK: Alaskan Native Language Center, 1985, pp. 135–158.

[14] J. Leer, *Yupik Eskimo Prosodic Systems: Descriptive and Comparative Studies*. 1985.

[15] J. Leer, "Toward a metrical interpretation of Yupik prosody," *Yupik Eskimo Prosodic Systems: Descriptive and Comparative Studies*, pp. 159–173, 1985.

[16] N. Gabas, "Phonetic correlates of stress in Yup'ik," *Prosody, Grammar, and Discourse in Central Alaskan Yup'ik, Santa Barbara Papers in Linguistics, vol. 7*, pp. 17–36, 1996.

[17] P. Afcan and E. Hofseth, *Napam cuyaa*. Fairbanks: Eskimo Language Workshop, University of Alaska, 1972.

[18] A. Blue, *Cungauyaraam Qulirai: Annie Blue's Stories*. Fairbanks, AK: Alaskan Native Language Center, 2007.

[19] D. Tabios and S. Meganack, *Sugcestun unigkuat: stories in the Sugcestun language from the villages of English Bay and Port Graham*. Fairbanks, AK: Division of Alaska Native Languages, Center for Northern Educational Research, 1972.

[20] "Alaska Native Language Archive." https://www.uaf.edu/anla/ (accessed Nov. 25, 2022).

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.1.10," *retrieved 23 March 2020*. 2020.

[22] A. Arnhold, "Duration, Intensity, F0 Measurement Script," *https://sites.ualberta.ca/~arnhold/scripts/OnlyMeasurements/MeasureIntensityDurationF0minF0maxF0contourpoints.praat*, 2018.

[23] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," 2006.

[24] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4.," *R package version*, 2014.

[25] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *J Stat Softw*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[26] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J Stat Softw*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.

[27] R. Lenth, "Package 'emmeans,'" *CRAN*, Sep. 2022.

[28] J. H. Koo and L. Badten, "Acoustic measurements of the 'fake' vowel length and degrees of vowel length in St. Lawrence Island Eskimo," *Phonetica*, vol. 40, no. 4, pp. 213–220, 1974.

VOWEL RAISING IN COTOPAXI QUICHUAN LANGUAGES

Sky Onosson[1] and Jesse Stewart[2]

[1]University of Victoria, [2]University of Saskatchewan
Sky@Onosson.com , Stewart.Jesse@usask.ca

ABSTRACT

This study investigates a phenomenon of morpheme-specific vowel raising that occurs in some Ecuadorian dialects of Quichua as well as the associated mixed language Media Lengua, where certain forms of /a/ are pronounced as [u]. Our data reveal that this process is region-specific, occurring in the Quichua dialect spoken in Cotopaxi province, but not in the relatively more well-known dialect of Imbabura province. Moreover, the dialects of Media Lengua in these respective regions also exhibit different realizations of the forms involved in this raising process which have both similarities to and differences from the patterns seen in their local Quichua dialects. These results may provide some insight into the relationships between these and other cross-regional dialects of Quichua, and how these may relate to similar dialects of Media Lengua, both contemporaneously and historically.

Keywords: Quichua, Media Lengua, vowels, acoustic phonetics, dialectal variation

1. INTRODUCTION

This paper offers a comparative analysis of vowels found in a set of morphemes in two dialects of Ecuadorian Quichua and two dialects of Media Lengua that are known to vary impressionistically based on the region where they are spoken. This analysis primarily focuses Cotopaxi Quichua (Glottocode: COTO1253) and Cotopaxi Media Lengua (Glottocode: MEDI1245). The former is a Quechuan language that forms part of the highland Quichua dialect continuum that extends throughout the Andean region of Ecuador. It is considered one of the central varieties and is spoken in Cotopaxi province by an estimated 50k speakers [1]. Like all Quechuan languages, Cotopaxi Quichua is agglutinating with SOV word order and contains three phonemic vowels (/i, u, a/). Cotopaxi Media Lengua is an endangered mixed language [2] which is spoken by an estimated 1703 people [3]. Media Lengua is essentially Quichua grammar, semantics, and phonology overlaid with Spanish vocabulary through the process of relexification [4].

To date, there are no phonetic studies of Cotopaxi Media Lengua and the only study that analyses phonetic data in Cotopaxi Quichua is a 2010 phonological description by Kohlberger ([5]). Included in his study is a brief overview of the vowel system which includes an F1xF2 plot with an estimated 50 vowels from six speakers (3 males & 3 females). This plot shows clear separation among /i/, /u/, and /a/ [5, p. 44]. Kohlberger's description also mentions that /a/ behaves differently in several morphemes when compared with other varieties in Ecuadorian Quichua. Specifically, the first /a/ in the suffixes, *-man* /man/ 'directional', *-manta* /manta/ 'ablative', and *-pac* /pak/ 'genitive' can be raised to [u] in spontaneous speech resulting in [muŋ], [munda], and [pux ~ puk], respectively while remaining as [a] in careful speech [5, p. 43]. Henceforth we refer to /a/ in these morphemes as /A/ to distinguish it from canonical /a/. /A/-raising appears to be a characteristic unique to central Quichua varieties (Chimborazo, Tungurahua, & Cotopaxi) as /A/ in Imbabura Quichua (north) and Loja Quichua (south) is documented as remaining canonically low ([a]) ([6, p. 18]). Additionally, multiple phonetic studies of Imbabura Quichua make no mention of /A/-raising in the previously mentioned suffixes ([7]–[10]).

The primary goal of this study is to revisit Kohlberger's analysis of the Cotopaxi vowel system and specifically focus on the morphemes involved in /A/-raising. In addition to *-man, -manta*, and *-pac*, we also investigate the verbal morpheme *-ngapac* [ngabu̯] 'same subject purpose marker' and the evidential *-mari* [mu̯ri] 'affirmative', both of which also give the impression of /A/-raising in our data.

The use of data from both Cotopaxi Quichua and Cotopaxi Media Lengua will further reveal whether the mixed language's vowel system patterns with those of Cotopaxi Quichua, makes use of its own innovations, or is more similar to other dialects of Media Lengua. To further the aims of this secondary goal and more fully understand the movement of /A/, this study compares Cotopaxi vowel data with previously analysed vowel data from both Imbabura Quichua and Imbabura Media Lengua (see [8], [9], [11]).

Coincidentally, both Cotopaxi and Imbabura provinces are home not only to different dialects of Quichua but also to different dialects of the mixed language, making them an ideal set of regions for comparison of the vowel raising process, among other

linguistic features. Example (1) provides a phrase from Cotopaxi Quichua (CQ) with parallel translations in Cotopaxi Media Lengua (CML), Imbabura Quichua (IQ), and Imbabura Media Lengua (IML) along with their respective glosses and Spanish and English translations. Bolded elements are of Spanish origin.

(1)

CQ	***Plaza**mun apangaboga na tarpanchic.*
Gloss	plaza-mụn apa-ngabụ-ga na tarpa-ntʃix
CML	***Plaza**mun **lleva**ngaboga **no sembra**nchi.*
Gloss	plaza-mụn ʒeβa-ngabụ-ga no sembra-ntʃi
IQ	***Plaza**man apangapaga na tarpanchic.*
Gloss	plaza-man apa-ngapa-ga na tarpa-ntʃik
IML	***Platsa**man **lleva**ngapaca **no sembra**nchi.*
Gloss	plaza-man ʒeβa-ngapa-ka no sembra-ntʃi
Sp:	Plaza-DIR llevar-PURP-TOP no sembrar-1P
En:	market-DIR bring-PURP-TOP no plant-1P
Trans	*No sembramos* [*comida*] *para llevar a la plaza.*
Trans	We don't plant [food] to bring to the market.

2. METHOD

2.1. Participants

Given the recent and ongoing documentation of Cotopaxi Media Lengua, for the first time in approximately 40 years [3], conversational data for this study was limited to two female speakers. Therefore, data from all languages and dialects in this study were also restricted to female speakers to avoid known differences in vowel production based on anatomical differences between males and females.

Cotopaxi Quichua data were gathered from three speakers in 2009 in the community of Quilotoa. A Spanish phrase list was used for elicitation. At the time of recording, the participants were 19, 28, and 38 years of age. For Cotopaxi Media Lengua conversational data comes from two speakers, aged 29 and 37 at the time of recording in 2022. Both were born and raised in the community of Yacubamba.

For comparative purposes, previously analysed data from Imbabura Quichua and Imbabura Media Lengua were also used [8], [11]. For Imbabura Quichua, data come from six female speakers from the community of Chirihuasi between the ages of 21 and 55 at the time of recording. For Imbabura Media Lengua, data come from five female speakers between the ages of 39 and 50 at the time of recording. Data from both Imbabura languages were recorded in 2011.

In terms of speech style or register, the Quichua data are primarily from elicitations and the Media Lengua data are primarily from conversations.

2.2. Data extraction and analysis

F1 and F2 frequencies were extracted from the centre points of vowels that were marked on a point tier in Praat [12] at their most steady state. A Praat script was used to automatically extract the data points. Table 1 provides a breakdown of the token counts by language and vowel. /A/ represents the vowel tokens extracted from the morphemes under analysis i.e., those potentially susceptible to the /a/-raising process, whereas the /a/, /o/, and /u/ columns reference vowels taken from full words and non-target morphemes. The libraries *ggplot2* [13] and *LmerTest* [14] were used to analyse the data in R version 4.2.2 [15].

	/A/	/a/	/o/	/u/	**Totals**
Cotopaxi Q	27	38		53	**118**
Cotopaxi ML	36	69	135	61	**301**
Imbabura Q	42	193		198	**433**
Imbabura ML	43	351	186	338	**918**
Totals	**148**	**651**	**321**	**650**	**1,770**

Table 1: Vowel token counts under analysis by language and region.

2.3. Procedures

Linear mixed-effects regression (LMER) models were run on a per-formant basis for F1 and F2, with *Vowel*, *Region* (Cotopaxi or Imbabura) and *Language* (Media Lengua or Quechua) as fixed effects including interactions between the fixed effects, and random intercepts for both *Speaker* and *Word*.

3. RESULTS

The LMER results for F1 and F2 are shown from the perspective of Cotopaxi Quichua /A/, which is set as the intercept in Table 2, and then from the perspective of its Imbabura Quichua counterpart which is set as the intercept in Table 3. To conserve space, Table 3 is simplified by removing rows involving *Region* as a single or interaction effect, as these are identical to the values in Table 2 except for the direction (sign) of the effect.

For Cotopaxi Quichua (Table 2), both formants of /A/ differ significantly from /a/ by a substantial amount, with a small (46 Hz) but significant difference from /u/ in terms of F1 only (there is also a non-significant difference from /o/, but this is irrelevant as the /o/ tokens are exclusively from Media Lengua). For Imbabura Quichua (Table 3), nearly the inverse is true; /A/ differs from /u/ in both formants, and from /a/ by a relatively large (308 Hz) difference in F2.

The effect of *Region* shows a significant difference for both formants. On a per-vowel basis, this shows up as a significant regional F1 difference

in /a/ (relative to /A/), and a regional difference in both formants of /o/ and /u/. *Language* turns out to have a relatively small role to play in the model, showing up in a two-way interaction with *Vowel* as a significant F1 difference between /A/ and /a/ in Cotopaxi (Table 2), and a corresponding F2 difference in Imbabura (Table 3). There is also a three-way interaction with *Region*, *Language* and *Vowel* involving F1.

	F1		F2	
Predictors	*Est.*	*p*	*Est.*	*p*
(Intercept)	507.78	**<0.001**	1306.06	**<0.001**
Vowel [a]	128.92	**<0.001**	439.33	**<0.001**
Vowel [o]	2.90	0.866	76.24	0.075
Vowel [u]	-45.73	0.038	-106.63	0.050
Region [Imbabura]	183.19	**<0.001**	209.04	**0.010**
Lang. [Media Lengua]	77.43	0.180	98.77	0.288
Vowel [a] * Region [Imbabura]	-120.85	**<0.001**	-131.76	0.061
Vowel [o] * Region [Imbabura]	-227.17	**<0.001**	-384.73	**<0.001**
Vowel [u] * Region [Imbabura]	-171.28	**<0.001**	-303.09	**<0.001**
Vowel [a] * Lang. [Media Lengua]	90.49	**0.003**	-67.92	0.360
Vowel [u] * Lang. [Media Lengua]	38.53	0.187	97.00	0.179
Region [Imbabura] * Lang. [Media Lengua]	-51.33	0.451	-22.05	0.846
(Vowel [a] * Region [Imbabura]) * Lang. [Media Lengua]	-112.76	**0.002**	-73.58	0.423
(Vowel [u] * Region [Imbabura]) * Lang. [Media Lengua]	-73.76	**0.043**	-44.18	0.625

Table 2: Linear mixed-effects regression results, intercept = Cotopaxi Quichua /A/.

	F1		F2	
Predictors	*Est.*	*p*	*Est.*	*p*
(Intercept)	690.96	**<0.001**	1515.10	**<0.001**
Vowel [a]	8.07	0.625	307.56	**<0.001**
Vowel [o]	-224.27	**<0.001**	-308.49	**<0.001**
Vowel [u]	-217.01	**<0.001**	-409.71	**<0.001**
Lang. [Media Lengua]	26.10	0.474	76.72	0.260
Vowel [a] * Lang. [Media Lengua]	-22.27	0.320	-141.51	**0.014**
Vowel [u] * Lang. [Media Lengua]	-35.22	0.117	52.82	0.358

Table 3: Linear mixed-effects regression results, intercept = Imbabura Quichua /A/; rows involving the effect of *Region* are excluded.

The regional differences in Quichua vowel production are visualized in Figure 1. In Imbabura (lower plot) while the F2 difference between /a/ and /A/ appears in the form of a slightly retracted distribution, the central regions of their respective distributions are nearly contiguous. In Cotopaxi (upper plot), contrastingly, there is virtually no overlap between /a/ and /A/. This is not the only difference, however, as /A/ also displays a clearly bimodal distribution, with one more retracted set of tokens which fully overlap the distribution of /u/, and another set which is more centralized, occurring in an intermediary position between /a/ and /u/. All /A/ realizations are substantially higher than /a/.

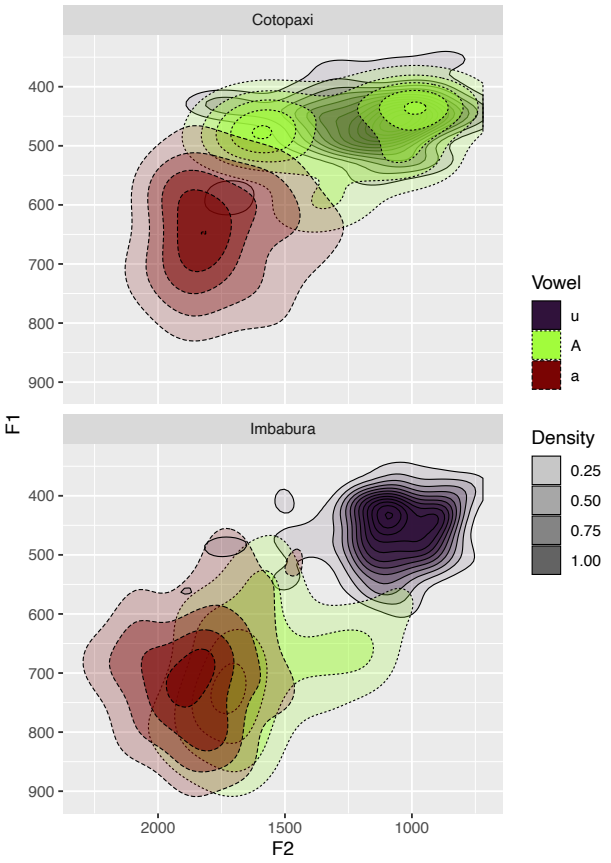


Figure 1: Quichua vowels by region.

For Media Lengua (Figure 2), the distributions of /A/ are similarly distinguished on a regional basis; however, the region-specific Media Lengua patterns do not precisely reflect those found for Quichua, and are moreover complicated by the presence of the additional back vowel /o/ (which itself overlaps substantially with /u/, a topic not focused on in this paper). Imbabura Media Lengua (lower plot) shows the most similarity with its regional Quichua variety: the distributions of /a/ and /A/ are largely overlapping, but the latter is slightly retracted. In Cotopaxi (upper plot), Media Lengua /A/, as in Quichua, appears to have a bimodal distribution, with one set of tokens largely overlapping /u/ (and, incidentally, /o/). The second set of /A/ tokens however, unlike Cotopaxi Quichua, are realized in a nearly identical manner as Imbabura's Media Lengua /A/, appearing as a slightly retracted /a/.

Note that the internal makeup of the distribution of Cotopaxi Media Lengua /A/ is not as readily discernible in Figure 2 compared with some of the

other plots. This is due to the combination of its bimodal pattern along with its relatively low token quantity, the fewest of any vowel token set (Table 1).

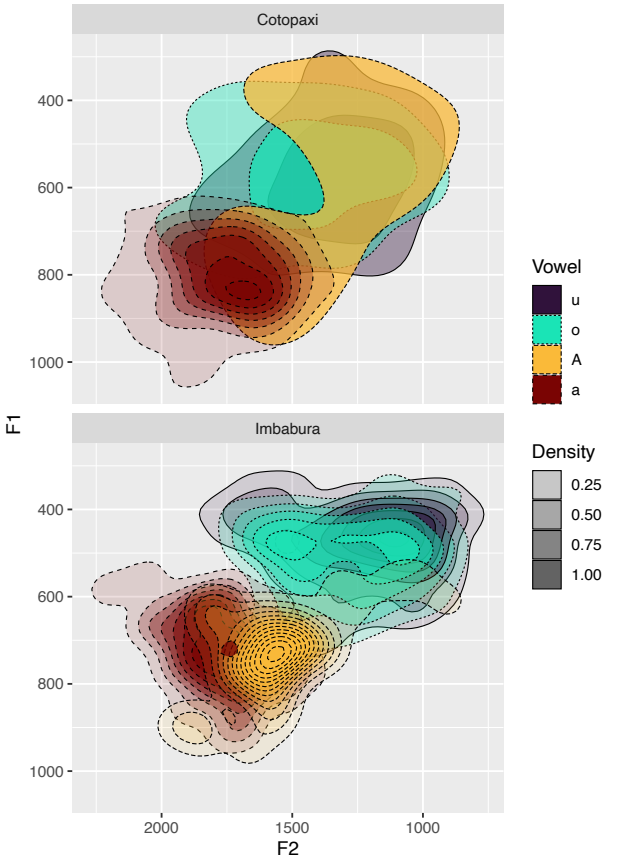


Figure 2: Media Lengua vowels by region.

4. DISCUSSION

This study supports Kohlberger's [5] observations of /A/-raising in Cotopaxi Quichua morphemes *-man* /man/ 'directional', *-manta* /manda/ 'ablative', and *-pac* /pak/ 'genitive', and further confirms that *-ngapac* 'same subject purpose marker' and the evidential *-mari* 'affirmative' also fit this paradigm. Kohlberger also observed some variation where /A/ remained [a] in careful speech but was raised to [u] in spontaneous speech. Although our Cotopaxi Quichua data were not set up to test differences in register, we found some variation in the production of /A/. While many of the /A/ tokens analyzed were both raised and retracted to the position of [u], a subset of /A/ tokens were only raised. These non-retracted vowels may give the perception of [a] as they are located directly above, and overlap in the front-back dimension with, the non-raising /a/ cluster, while remaining structurally distinct in that the average F1 frequency of /A/ remains far below that of /a/ (i.e., they remain raised). Based on this observation, an additional clustering analysis of each /A/-raising morpheme was conducted. While there was some variability between morphemes, the results were too inconsistent to suggest that any specific morpheme was responsible for the front-back variation. It may also be the case that the subset of non-retracted /A/s may be simply weakening to [ə] as this cluster is quite centralised in acoustic space. However, as our data are not currently set up to analyse other correlates of weakening (intensity, length, etc.), we must leave this question to future research.

Cotopaxi Media Lengua offers a somewhat different pattern. Like Cotopaxi Quichua, the majority of /A/s are raised. However, these raised realizations are generally not quite as retracted as their Quichua counterparts. A smaller cluster of /A/ tokens also occur in a position with both F1 and F2 averages close to that of canonical /a/. Once again, additional clustering analysis did not reveal that any specific morpheme was responsible for /A/ remaining unraised ([a]).

The analyses of both Cotopaxi languages, along with previous impressionistic observations ([6]), suggest that /A/-raising is a regional feature of central Quichuan. Viewing /A/-raising as an areal feature, it is plausible that register (careful vs. spontaneous) may not in fact be a robust predictor of /A/ production (low vs. high) as Kohlberger suggests, given that Cotopaxi Quichua data were elicited while Media Lengua data were conversational. Further investigation is warranted to validate this claim.

Comparing the Cotopaxi to Imbabura languages, it becomes clear that there is regional variation between the north (Imbabura) and central (Cotopaxi) varieties in terms of /A/ production. In both Imbabura Quichua and Imbabura Media Lengua, not a single token of /A/ is raised, and the production differences between /A/ and canonical /a/ are marginal. It is, however, noteworthy that both Imbabura Quichua and Imbabura Media Lengua /A/, are slightly retracted (but not raised) compared to canonical /a/ (307 Hz for Quichua & 166 Hz for Media Lengua). We do not expect that this entails an overall change in vowel quality for several reasons (1) /a/ and /A/ are still highly overlapping in both languages, (2) there is no other phonemic vowel occupying that region of acoustic space, and (3) both languages are known to occupy relatively large acoustic spaces making the precise articulation of low vowels inconsequential.

A final point of consideration is the similarity between the small subset of retracted-but-unraised Cotopaxi Media Lengua /A/ tokens, which positionally align with the distribution of Imbabura Media Lengua /A/. On this point, the cross-regional varieties of Media Lengua are in accordance, indicating that the areal differences are not wholesale, and that there is still some cross-regional influence between the Media Lengua dialects themselves.

5. REFERENCES

[1] Chisaguano, S. 2006. *La Población Indígena del Ecuador: Análisis de Estadísticas socio-demográficas*. Quito: Instituto Nacional de Estadísticas y Censos.

[2] Muysken, P. 1997. Media Lengua. In: Thomason, S. (ed) *Contact languages: A Wider Perspective*. J. Benjamins Pub. Co.

[3] Stewart, J., Gonza Inlago, L., Prado Ayala, G. Accepted. Cotopaxi Media Lengua is still very much alive. In: Language Documentation and Conservation.

[4] Muysken, P. 1981. Halfway between Quechua and Spanish: The case for relexification. In: Highfield, A. R. (ed) *Historicity and variation in Creole studies*, vol. 57–78. Karoma Publishers.

[5] Kohlberger, M. 2010. *Cotopaxi Quichua: A phonological description and an analysis of stops and affricates in Central Highland Ecuadorian Quichua*. University of Edinburgh.

[6] Orr, C. 1978. *Dialectos Quichuas del Ecuador*. Instituto Lingüístico de Verano.

[7] Guion, S. 2003. The vowel systems of Quichua-Spanish bilinguals: Age of acquisition effects on the mutual influence of the first and second languages. *Phonetica* 60, 98–128.

[8] Stewart, J. 2014. A comparative analysis of Media Lengua and Quichua vowel production. *Phonetica* 7, 159–182.

[9] Onosson, S., Stewart, J. 2021. The Effects of Language Contact on Non-Native Vowel Sequences in Lexical Borrowings: The Case of Media Lengua. *Language and Speech PaPE Special Issue*, 1–30.

[10] Lipski, J. 2015. Colliding vowel systems in Andean Spanish. *Linguistic approaches to bilingualism*, vol. 5, 91–121.

[11] Onosson, S., Stewart, J. 2021. A multi-method approach to correlate identification in acoustic data: The case of Media Lengua. *Laboratory Phonology* 12(1), 1–30.

[12] Boersma, P., Weenink, D. 2016. *Praat*. Institute of Phonetic Sciences, University of Amsterdam.

[13] Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

[14] Kuznetsova, A., Brockhoff, P. B., Bojesen, R. lmerTest. 2016. CRAN.

[15] R Development Core Team. 2017. *R*. R Foundation for Statistical Computing.

A STUDY ON TONES IN JIUHE BAI PRODUCED BY NAXI SPEAKERS

Meihao Wan, Feng Wang

The Chinese University of Hong Kong, Peking University
wanmeihao@link.cuhk.edu.hk, wfwf@pku.edu.cn

ABSTRACT

Bai is a language spoken in southwestern China which uses both pitch and non-modal phonation for tone contrast. In Jiuhe, there is a group of multilingual speakers of both Bai and Naxi languages. This paper aims to investigate the tone of Bai produced by local Naxi speakers with electroglottograph. The results show that the number of tones is reduced and some tones have merged in Naxi speakers' production. The high level tone is the least affected whereas the low tense tone is prone to merge with other tones. Some Naxi speakers also use non-modal phonation for tone contrast of Bai as native Bai speakers do, even though there is no phonation contrast in the tones of Naxi. Furthermore, this case implies that tone merger is not only due to the similarity of F0 contour, phonation can also be the condition in tone merge.

Keywords: Bai, tone, Naxi, phonation

1. INTRODUCTION

Bai is a tone language of Tibeto-Burman languages and also a register language, in which both F0 and voice quality function in tone contrast. Previous studies have shown that the tense tones in Bai are associated with non-modal phonation [1], but not accompanied with specific phonation types. The non-modal phonation types in the tense tones vary across different speakers and different dialects [2], [3]. For example, in Jianchuan Bai, one speaker may use harsh voice in the tense tone T7 while another speaker may use pressed voice. However, in Meiba Bai, the phonation of the same tone can be creaky voice. Previous research focused on the production of native Bai speakers and rarely on the Bai language produced by native speakers of other languages.

The Jiuhe township is located at the border of the Bai speaking region and the Naxi speaking region, where Bai and Naxi people live together in the villages. Because of their close contact, most people in the village are multilingual and can speak at least three languages: Bai, Naxi and Mandarin Chinese. Bai and Naxi are used in most situations in daily life. Jiuhe Bai has six lexical tones (see Table 1), which can be described with Chao's five-point scale [4]. Similar to other Bai languages, there is a tense vs. lax register in Jiuhe Bai: T1, T3, and T5 are tense tones, and T2, T4, and T6 are lax tones. In Jiuhe Bai, tense tones are not associated with specific phonation as well. For example, the phonation cue for T1 can be realized as creaky, breathy or harsh when produced by different speakers.

Register	Lexical tones		
Tense	T1 low-falling [tɐi21] 'flag'	T3 mid-falling [tɐi31] 'chase'	T5 mid-level [tɐi44] 'leech'
Lax	T2 low-falling [tɐi21] 'ground'	T4 mid-level [tɐi33] 'pull'	T6 high-level [tɐi55] 'many'

Table 1. Tones of Jiuhe Bai with examples

The phenomenon of tone merger has been studied for a long while. In previous studies such as Cantonese, another language with a complex tone system, some speakers merged several tone pairs which are similar in F0 contours [5]. Whether the similarity of F0 will lead to the merging of tones in Jiuhe Bai is an important issue of concern. The investigation of Naxi Bai can shed light on the study of mechanism of tone merger.

Moreover, unlike other well-studied tone languages, Bai is one of the register languages that use phonation for tone contrast. Another question that this study hopes to answer is whether Naxi speakers can use phonation for tone contrast. Jiuhe Bai is a good case to observe phonation as a contrastive cue in tone production of non-native Bai speakers.

2. METHOD

2.1. Data sampling

The data in this paper were all collected during a field trip in Jiuhe Township in August 2021. Four native Naxi speakers participated in the production experiment, including two males (one aged 45, one aged 46) and two females (one aged 72, one aged 73), in addition to two native Bai speakers (one aged 34, one aged 51) for control involving one male and one female. All of them were farmers who have lived in the village since birth and could speak Bai proficiently.

The recording software was Adobe Audition CC 2018, and the recording was taken in dual-channel recording, with the audio signal recorded in the left

channel by a condenser microphone (SONY ECM-44B), and the EGG signal recorded in the right channel by the Electroglottograph Model 7050A, and the two channels of signals were recorded through a mixer (XENYX 302 USB) and the sound card (SBX) into the computer (Huawei matebook 14 2019 model). The sampling frequency was 22050 Hz. The format of the audio file was WAV. The recording was performed in a quiet room in the villagers' house.

2.2. Bai materials

Since there is no data of Jiuhe Bai before, a corpus of Jiuhe Bai's vocabulary was created, in which eleven tonal monosyllabic minimal pairs were collected for the recording. All of the monosyllables were relatively balanced in frequency and were used frequently in daily life that Naxi speakers could respond to and pronounce instantly. For each speaker, these eleven minimal sets which contained 66 monosyllables (11 minimal sets × 6 tones) were pronounced twice. The minimal sets were identical in segmental structure but differed in tones as presented in Table 1.

2.3. Measures

In this study, three parameters were extracted from the acoustic signals and EGG signals: Fundamental Frequency (F0), Open Quotient (OQ) and Speed Quotient (SQ). F0 was extracted from the acoustic signals. OQ and SQ are extracted from EGG signals. F0 is defined as 1/period, OQ is defined as open phase/period, SQ is defined as opening phase/closing phase. These three parameters were extracted by the Matlab program 'VoiceLab', written by the Voice Laboratory of the Department of Chinese Language and Literature, Peking University.

The parameters extracted in this study were time normalized, that is, the F0, OQ, and SQ of all samples were composed of 20 points of data respectively. After the time-normalization, the first two data points and the last two points of each syllable were deleted, in that only the signals in the middle part are stable and reflect the phonetic features of the syllables. Ultimately, for each parameter, 16 stable data points were analyzed. Besides, in order to exclude the pitch range difference caused by the individual differences, we converted the F0 values into z-score for better comparison.[6]

3. RESULT

3.1. F0 contour

Figure 1 displays the F0 contours of the six lexical tones produced by native Bai speakers, which are well distinguished by pitch, in which T1 and T2 are low falling tones, T3 is a mid falling tone, T4 and T5 are mid level tones and T6 is a high level tone. The F0 contours of T1 and T2 are close but do not overlap, so are the F0 contours of T4 and T5. The F0 contour of T3 decreases more sharply with higher pitch compared to the other falling tones. The high level tone T6 is distinct from the other tones.

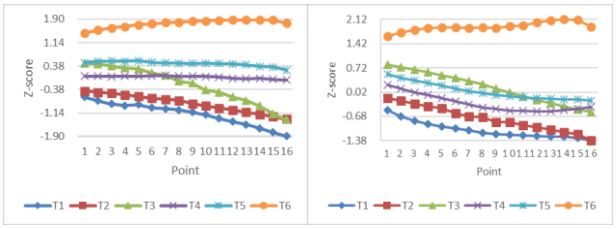


Figure 1: F0 contours of native Bai speakers (F1-B left, M1-B right)

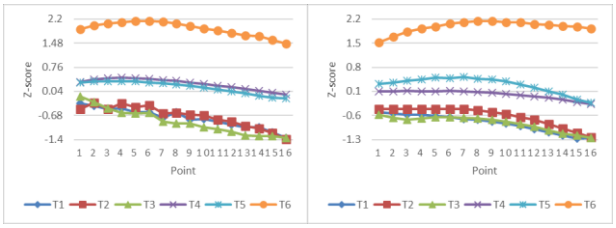


Figure 2: F0 contours of Naxi speakers (F1-N left, M1-N right)

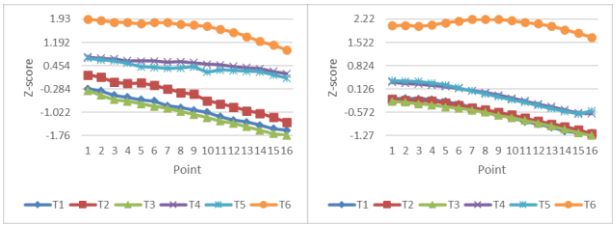


Figure 3: F0 contours of Naxi speakers (F2-N left, M2-N right)

However, some tones may merge in the production of Naxi speakers. The most common phenomena are the merging of T1 and T3 as well as the merging of T4 and T5 (see Figure 2 and Figure 3). There are also mergers of T1, T2, and T3. The affected degree varies among different tones. T6 is the least affected tone and the most affected tone is T1.

3.2. Phonation analysis

In Bai, phonation is also an important cue for tone contrast. However, there is no phonation contrast in the tones of Naxi. The phonation type can be identified by the OQ and SQ values [3], [7]. Table 2 shows the relationship between the vocalization type and the parameters based on previous studies. "+" is

marked if the value of tense tone is significantly greater than that of lax tone, and "-" is marked if the value of tense tone is significantly smaller. Based on the comparison of these parameters, we can identify the phonation type of the tense tone.

	Vocal fry	Breathy	Creaky	High-pitch	Harsh
F0	-	-	-	+	-
OQ	+	+	-	-	-
SQ	+	-	+	-	-

Table 2: Phonation types and their parameters

The figures below display the phonation parameters of Bai tones produced by Naxi speakers. Since the F0 contours of T1 vs. T3, T4 vs. T5 overlap in their production, we will compare the phonation by the merged tone pair. In addition, as both T1 and T3 are tense tones with non-modal phonation in Bai, we also display the parameter of T2 which is a lax tone with modal voice for comparison. All interpretations of the OQ and SQ values are summarized in Table 3.

3.2.1 T1, T2 vs. T3

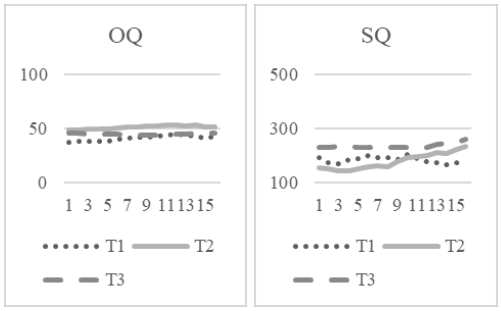


Figure 4: OQ and SQ contours of a Bai speaker (M1-B)

The OQ contours of Bai M1's T1, T2 and T3 are all relatively flat overall, with T2 slightly above T1 and T3 and the values are around 50%. The SQ contours of the three tones are also all relatively smooth, with T3 having the highest SQ at around 230%, T1 and T2 having SQ values between 150% and 230%.

Figure 5 displays the OQ and SQ contours of Naxi speaker F1. The OQ contours are flat, and the contours of the three tones are generally close to each other, with values around 50%. The SQ contours of F1 largely overlap and all show an upward trend, with values rising from about 200% to about 350%.

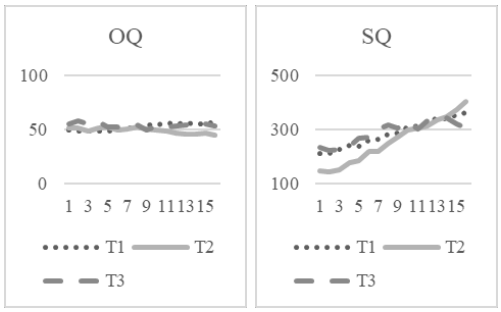


Figure 5: OQ and SQ contours of a Naxi speaker (F1-N)

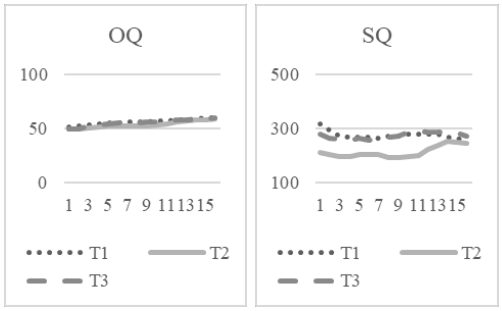


Figure 6: OQ and SQ contours of a Naxi speaker (F2-N)

The OQ and SQ contours of Naxi speaker F2 are displayed in Figure 6. The OQ contours show a slight upward trend, and the three contours are generally close, with values between 50% and 60%. In the SQ contours of F2, T1 and T3 largely overlap, with values between 250% and 300%; T2 is lower compared to T1 and T3, with lower values in the first two-thirds and slowly rising in the second third, close to the other two tones.

3.2.2 T4 vs. T5

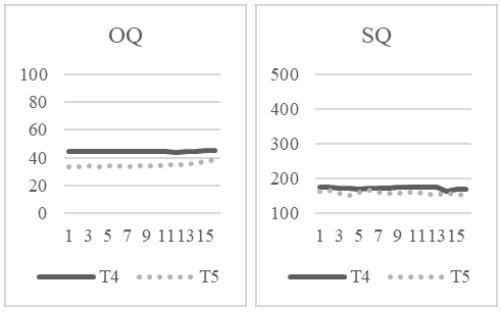


Figure 7: OQ and SQ contours of a Bai speaker (F1-B)

The OQ contour of T4 for Bai speaker F1 is relatively smooth, with values around 45%; the T5 OQ contour is also generally smooth and lower than that of T4 with values roughly around 34%. The SQ contours of the two tones are both relatively flat and largely overlap, with values around 165%.

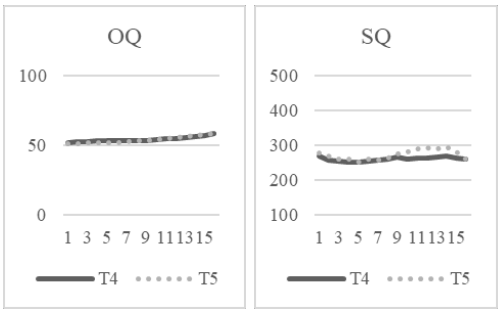


Figure 8: OQ and SQ contours of a Naxi speaker (F2-N)

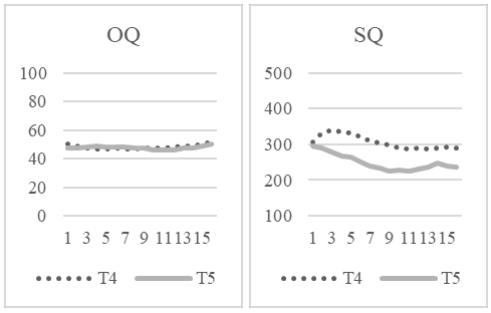


Figure 9: OQ and SQ contours of a Naxi speaker (M1-N)

The OQ contours of T4 and T5 for speaker F2 (see Figure 8) largely overlap, with an overall slight upward trend, with values ranging from about 50% to 60%. The SQ contours of T4 and T5 overlap roughly in the first half, and the curve of T5 is a little higher than that of T4 in the second half, reaching a maximum of about 290%, while T4 is around 265%.

The OQ contours of the two tones for M1 (see Figure 9) are smooth and largely overlap as well, with the value of around 50%. The SQ contours of M1's T4 is also smooth, with the SQ of T4 slightly larger than that of T5. The value of T4 is around 300%, and T5's SQ curve falls slightly, with the value dropping from 300% to around 250%.

To examine whether the differences are significant, independent sample *t*-tests are applied to the parameters point to point between the tone pairs. The results shows that there is almost no significant difference between the four Naxi speakers for each parameter of T1 vs. T3. Only the OQ values of M1's T1 and T3 have a significant difference. The results of all the speakers, including Bai and Naxi native speakers, are present in Table 3. "ns" is marked if there is no significant difference between the parameters of the two tones. The meaning of other marks is the same as in Table 2.

It can be seen from Table 3 that Bai M1 can distinguish the tone pairs by both F0 and phonation. Bai F1 can distinguish T4 and T5 with F0 and phonation but not T1 and T2, which may be due to her younger age.

Speakers	T1 vs. T2			T4 vs. T5		
	F0	OQ	SQ	F0	OQ	SQ
F1-B	-	ns	ns	+	-	ns
M1-B	-	-	ns	+	-	+
F1-N	ns	+	ns	ns	ns	ns
M1-N	-	-	-	+	-	-
F2-N	-	+	+	-	ns	+
M2-N	ns	ns	ns	ns	ns	ns

Table 3: Summary of parameter contrasts of the Bai tense vs. lax tones

For Naxi speakers, the tone merging of F1 and M2 is more radical compared to the other two speakers. They have poor performance in distinguishing the tone pairs in F0 and phonation. For the other two speakers, T2 is not only different from T1 and T3 in terms of F0, but also in terms of OQ and SQ. Based on the comparison between the parameters and Table 2, T1 of M1 is a typical harsh voice and T1 of F2 is a typical vocal fry voice. It can be seen that some Naxi speakers are also able to use phonation to distinguish different tones.

4. CONCLUSION

The results show that some Bai tone pairs have merged into one tone among the Naxi speakers. Based on the acoustic and EGG data, we can confirm that T1 vs. T3, T4 vs. T5 have merged in Naxi speakers' production, and some of the Naxi speakers have already merged T1, T2 and T3. The least affected tone is T6, which can be distinguished from other tones in all the Naxi speakers.

This case shows a different pattern of tone merging from previous studies. T4 and T5 are similar in F0 contours, which is consistent with the tone merging cases before. However, T1 and T3 are not similar in F0 contours in Jiuhe Bai. The similarity between the two tones is that they are both tense tones with non-modal phonation, which implies that tone merger is not only due to the similarity of F0 contour, phonation can also be the condition in tone merge. This also suggests that the phonation may affect the perception of tones, making them converge even if the F0 contours are not similar.

The investigation also reveals that there is a phonation contrast in the Bai tones produced by some Naxi speakers, which shows that phonation can be acquired to some extent in language contact as Jiuhe Naxi does not use phonation to distinguish different tones. Also, this may contribute to their failure in distinguishing Bai tones by phonation types.

5. ACKNOWLEDGEMENTS

This research was funded by Humanities and Social Sciences Foundation of the Ministry of Education (Grant 19JJD740001) and Major Project of Beijing Social Science Foundation (Grant 20ZDA20). We also thank Prof. Peggy Mok for helpful suggestions.

6. REFERENCES

[1] J. A. Edmondson and J. H. Esling, “The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies,” Phonology, vol. 23, no. 2, pp. 157–191, Aug. 2006, doi: 10.1017/S095267570600087X.

[2] X. Li and F. Wang, “Phonation variation and strategy of tone distinction: The case of Meiba Bai,” J. Chin. Linguist., vol. 48, no. 2, pp. 379–401, 2020, doi: 10.1353/jcl.2020.0009.

[3] 汪 鋒 (Wang, Feng), “VARIATIONS OF LARYNGEAL FEATURES IN JIANCHUAN BAI/劍川白語的嗓音變異,” J. Chin. Linguist., pp. 434–452, 2015.

[4] Y.-R. Chao, “A system of tone letters,” Maître Phon., 1930.

[5] P. P. K. Mok, D. Zuo, and P. W. Y. Wong, “Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese,” Lang. Var. Change, vol. 25, no. 3, pp. 341–370, Oct. 2013, doi: 10.1017/S0954394513000161.

[6] P. Rose, “Considerations in the normalisation of the fundamental frequency of linguistic tone,” Speech Commun., vol. 6, no. 4, pp. 343–352, Dec. 1987, doi: 10.1016/0167-6393(87)90009-4.

[7] 孔江平 (Kong, Jiangping), 论语言发声 (On Language Phonation). 中央民族大学出版社, 2001.

INTRINSIC FUNDAMENTAL FREQUENCY DIFFERENCES IN TWO TONAL AUSTRONESIAN LANGUAGES

Laura Arnold[1, a)], Jiayin Gao[2, 3, b)], James Kirby[3, 1, c)]

[1]University of Edinburgh, [2]LPP (UMR7018, CNRS/Sorbonne Nouvelle), [3]Institute for Phonetics and Speech Processing (IPS), LMU Munich

a) laura.arnold@ed.ac.uk, b) jiayin.gao@cnrs.fr, c) jkirby@phonetik.uni-muenchen.de

ABSTRACT

We report on intrinsic fundamental frequency (IF0) differences in Butlih Salawati and Biga, two closely related Austronesian languages. Both languages have small inventories of six underlying vowels; they both also have simple tone systems, with two specified tones. We show that IF0 differences between high and low vowels in the languages range from 1.5 to 2.8 ST, depending on tonal specification and utterance context. Overall, these results are larger than the mean crosslinguistic difference of 1.65 ST. Based on the previous literature, strong claims that IF0 is under the control of the speaker predict that IF0 differences are attenuated in tone languages, and not enhanced in languages with small vowel inventories. The results presented here do not support these claims. However, we suggest that these results may explain the comparatively frequent tonal developments conditioned by vowel height attested in Austronesian languages.

Keywords: Austronesian, intrinsic fundamental frequency, vowels, tone, production.

1. INTRODUCTION

1.1. Intrinsic fundamental frequency

Intrinsic fundamental frequency (IF0) is a near-universal phenomenon in which high vowels are produced with a higher F0 than low vowels. The mean crosslinguistic difference in F0 between the high vowels /i, u/ and the low vowel /a/ is 1.65 semitones (ST) [1]; however, IF0 differences vary between languages and between speakers [1, 2].

While IF0 has been documented in over 40 genealogically and typologically diverse languages, the mechanisms are the subject of much debate. Traditionally, there are two camps. In the first are advocates of the *biomechanical hypothesis*, who argue that IF0 is an automatic phonetic process caused by an interaction of the articulatory and phonatory systems [1, 3]. The main evidence supporting this hypothesis is the sheer number of languages in which IF0 is found—it is attested in nearly every language in which it has been investigated (e.g. [1, 4, 5])—as well as the presence of IF0 in the babbling of infants [3].

In the second camp are those who argue that IF0 is under speaker control, and that it is more-or-less consciously exploited—for example, to maximize distinctions elsewhere in the phonological system [6, 7]. We refer to this position as the *speaker control hypothesis*. Support for this hypothesis includes the presence of IF0 in esophageal speech, which can have no biomechanical explanation [8]; higher levels of cricothyroid activity during the articulation of high vowels [9], which is associated with active control of F0 [10]; and the existence of at least one language without IF0 differences (Mambila, [4]).

Previous literature suggests that IF0 may be enhanced or attenuated, depending on the phonological profile of a language. One proposed link relates to vowel inventory size. Because vowel height is better perceived when IF0 is enhanced [11], the following trend is predicted: a language with a larger vowel inventory enhances IF0 differences, in order to maximize phonological differentiation within a more clustered vowel space. While Dutch (with 12 vowels, mean IF0 = 21 Hz) and Arabic (with three vowels, mean IF0 = 8 Hz) [5] have been taken as evidence to illustrate such an effect of vowel inventory size, cross-linguistically the correlation between IF0 and number of vowels is not significant [1].

Whether a language has tone, and if so the nature of the tone system, may also have an effect on language-specific IF0 differences. As F0 has a contrastive function in tonal languages, it is reasonable to assume that IF0 differences might be attenuated so as not to jeopardize tonal distinctions. While IF0 differences are attested in most tone languages (e.g. [1, 4]), they tend to be smaller when compared with non-tone languages [2]; and the unusually crowded tone system of

Mambila is offered as the explanation for the lack of IF0 differences in this language [4]. This position predicts that IF0 will be attenuated in tone languages, with increased inhibition in languages with more complex tone systems. However, crosslinguistically, such a tendency is only weakly observed [1, 2]. Furthermore, these observations are largely based on data from a limited number of tone languages spoken in Africa and East and Southeast Asia. More typologically diverse data is required to develop stronger hypotheses on the relationship between tone and IF0.

While there is evidence on both sides of the debate, the emerging consensus is that IF0 has a mixed basis: there is a biomechanical foundation and IF0 is thus (near-)universal; but it is also under speaker control, and thus varies by language, dialect, or speaker [2, 5, 12, 13]. With this context in mind, we provide some background on the languages investigated in this study.

1.2. Target languages and phonological profiles

Biga and the Butlih variety of Salawati are spoken in the Raja Ampat archipelago, northwest New Guinea. Within Austronesian, they belong to the South Halmahera-West New Guinea (SHWNG) branch, which is the sister of Oceanic; and within SHWNG, the languages belong to the Raja Ampat-South Halmahera (RASH) subbranch [14, 15]. Both languages have 14 consonants (Butlih /p b t d k f s m n h r l w j/; Biga /p b t d k g f s m n r l w j/). Butlih contrasts five vowels /i e a o u/ plus a schwa in sesquisyllables; Biga contrasts six vowels /i ɪ e a o u/. Unusually for the Austronesian family—although not so unusually for the RASH languages of Raja Ampat [14, 16]—both languages have lexical tone. The tone systems are modest, and specifications are generally culminative and limited to word-final syllables. In Butlih, there is a contrast between High, Rise, and toneless syllables; in Biga, the contrast is between High, Extra-High, and toneless syllables.

The phonological profiles of the two languages lead us to predict that IF0 differences will be comparatively small in both Butlih and Biga. On the one hand, the vowel inventories of both languages are relatively small and uncrowded, so there is no need for speakers to disambiguate a crowded vowel space by enhancing IF0 differences. On the other, both languages have tone systems. While the systems are not complex in either language—so there is no need to neutralize IF0 differences to compensate for a crowded tonal space, as argued for Mambila [4]—we still expect to see some attenuation of the differences.

Based on the preceding discussion, we use novel data from these two underdocumented languages to address the following research questions: (1) Are there IF0 differences in Butlih Salawati and Biga? If so, what is the magnitude of the differences? (2) Are IF0 differences in Butlih Salawati and Biga in line with predictions based on the phonological profiles of the languages?

2. METHOD

2.1. Participants

Six speakers (3 male, 3 female) were recorded for each language during a fieldtrip to Raja Ampat in 2019. The mean age of the Butlih speakers was 47 (range: 36-65), and of the Biga speakers was 51 (range: 44-66). Informed consent was gained from participants, and all were remunerated for their time.

2.2. Speech material

In both languages, the target words were monosyllables of the shape CV, VC, and CVC. As far as the lexical inventories of the languages allowed, target words were selected that combined each vowel with each tonal specification. Each target word was recorded multiple times in two contexts: utterance medially and utterance finally. The carrier sentences used were as follows: 'I say X' (utterance-final; Butlih /ine jəwe (bo) X/, Biga /ɪn bɪtɪn X/) and 'I don't say X' (utterance-medial; Butlih /ine jəwe (bo) X po/, Biga /ɪn bɪtɪn X apo/). Prompts were provided orally by the first author, who asked for translations of equivalent sentences from Papuan Malay, the local lingua franca.

2.3. Data processing and analysis

This study reports a subset of the data, in which each syllable contains a monophthong high or low vowel /i (ɪ) u a/ and carries a specified tone (High or Rise in Butlih; Extra-High or High in Biga). Toneless words were excluded because the number of tokens was too small. In total, 525 tokens (Butlih) and 703 tokens (Biga) were retained for analysis. Recordings were stored as an EMU-SDMS database [17]. Annotations were performed either manually or using WebMAUS [18] with hand corrections. F0 values were extracted at 5 msec intervals using the `ksvF0` function of the `wrassp` package [19], and then converted to semitones with reference to the mean F0 of each speaker.

Statistical analyses were conducted using the

`lmerTest` [20] and `emmeans` [21] packages in R [22]. For both languages, F0 values were averaged over one portion of the vowel (see the VOWELREGION predictor below), and linear mixed-effect models were fitted to the averaged F0 data. The full models included Helmert-coded predictors VOWEL HEIGHT (VH) (highV /i (ɪ) u/ vs. lowV /a/), SEX, ONSETVOICING (zero, voiceless, or voiced), POSITION in the utterance (medial vs. final), VOWELREGION in the vowel (Vonset vs. Vmid), TONE (H vs. R or EH vs. H), and the interaction terms VH:ONSETVOICING, VH:SEX, and VH:POSITION:VOWELREGION:TONE. (Vonset and Vmid indicate the first third and the middle third of the vowel, respectively.) All these predictors have been found to affect F0 in previous studies (e.g., [1, 23, 24, 25]). Random intercepts were included for SPEAKER, ONSET, and CODA. By-speaker random slopes were included for VH, TONE (for both), and POSITION (for Butlih only). The final models for each language were selected based on backward elimination of non-significant effects and will be provided in the relevant result subsections.

3. RESULTS

3.1. Butlih Salawati

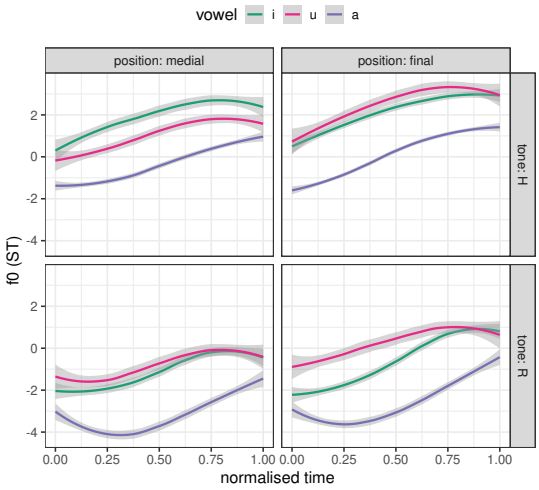


Figure 1: Aggregated F0 curves (loess-smoothed) by vowel, tone, and position from 6 speakers of Butlih Salawati. Shading: 95% CI.

As shown in Fig. 1, the two high vowels carry higher F0 than the low vowel almost throughout the entire trajectory. To quantify the magnitudes of IF0 in different contexts, pairwise comparisons were conducted on the finally selected model with the following predictors: VH, TONE, VOWELREGION, POSITION, ONSETVOICING, and the interaction terms VH:TONE:VOWELREGION and VH:ONSETVOICING.

In both utterance-medial and -final positions, IF0 is overall larger in H than R tones. When the syllable carries R tone, IF0 is larger at Vmid ($est.$ = 2.71, SE = 0.40, t = 6.81, $p < .0001$) than at Vonset ($est.$ = 1.96, SE = 0.40, t = 4.93, p = .0002). When the syllables carries H tone, IF0 is realized with a similar extent at Vonset ($est.$ = 2.70, SE = 0.36, t = 7.52, $p < .0001$) and Vmid ($est.$ = 2.60, SE = 0.36, t = 7.24, $p < .0001$).

3.2. Biga

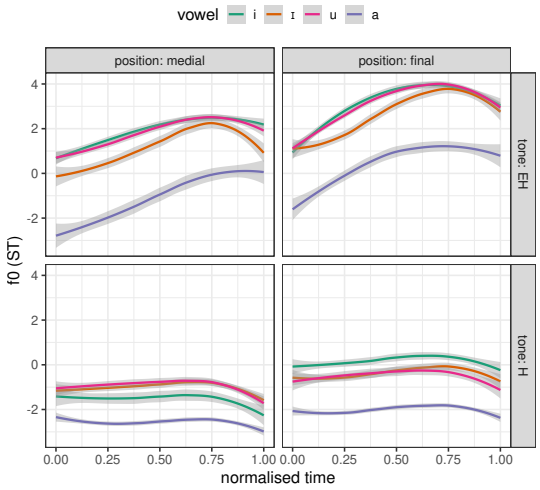


Figure 2: Aggregated F0 curves (loess-smoothed) by vowel, tone, and position from 6 speakers of Biga. Shading: 95% CI.

As shown in Fig. 2, here again, the three high vowels carry higher F0 than the low vowel almost throughout the entire trajectory. The effect of tonal context is more visible here than in Butlih. Pairwise comparisons were conducted on the finally selected model with the following predictors: VH, TONE, VOWELREGION, POSITION, ONSETVOICING, and the interaction terms VH:TONE:POSITION:VOWELREGION and VH:ONSETVOICING.

IF0 is larger in EH than H tones. When the syllable carries EH tone, IF0 is larger in utterance-medial (Vonset: $est.$ = 2.83, SE = 0.49, t = 5.79, $p < .0001$; Vmid: $est.$ = 2.44, SE = 0.49, t = 4.99, $p < .0001$) than -final position (Vonset: $est.$ = 2.40, SE = 0.42, t = 5.71, $p < .0001$; Vmid: $est.$ = 2.39, SE = 0.42, t = 5.67, $p < .0001$). When the syllable carries H tone, IF0 is smaller in utterance-medial (Vonset: $est.$ = 1.56, SE = 0.27, t = 5.75, $p < .0001$; Vmid: $est.$ = 1.70, SE = 0.27, t = 6.30, $p < .0001$) than -final position (Vonset: $est.$ =

1.86, SE = 0.32, $t = 5.87$, $p < .0001$; Vmid: *est.* = 1.94, SE = 0.32, $t = 6.14$, $p < .0001$).

4. DISCUSSION

1. Are there IF0 differences in Butlih Salawati and Biga? If so, what is the magnitude of the differences?

Yes, both languages exhibit IF0 differences to a similar extent, ranging from 1.5 to 2.8 ST depending on tonal specification and utterance context. Overall, this is larger than the mean crosslinguistic difference of 1.65 ST reported in careful speech [1], and the differences obtained for most languages from a study on large corpora of read speech [2]. While the nature of the speech material and the methods for measurements differ across studies, which would always remain an uncontrolled source of variability, we can conclude that IF0 in Butlih and Biga is probably larger than the crosslinguistic average.

IF0 also varies according to tone or the portion of the tonal contour. It is overall larger in syllables with a higher than a lower tone. For the Rise tone in Butlih, it is higher in the mid portion than the onset portion of the vowel, that is, it is higher in the portion of the vowel where overall F0 is higher. Previous studies report the disappearance of IF0 in the low F0 range [26], in low tones or the low portion of the tonal contour [1]. Although we do not observe an absence of IF0 even in the lowest portion of the tone, the increase of IF0 in higher-pitched tones or the higher portion is in line with previous findings. This relation between IF0 and the F0 range is widely attested, and thus has been taken as evidence for the biomechanical hypothesis, arguably because the activity of the strap muscles involved in larynx lowering counteracts the tongue-pull effect [1, 27] which is primarily responsible for the IF0 differences [28].

2. Are IF0 differences in Butlih Salawati and Biga in line with predictions based on the phonological profiles of the languages?

The results do not support a strong position linking speaker enhancement of IF0 differences to the phonological profiles of the languages. Following such a position, we would expect to see smaller IF0 differences in both Butlih and Biga, resulting from the combination of (1) a lack of enhancement due to relatively small vowel inventories and thus a lower need to maximize contrasts; and (2) attenuation due to the contrastive use of F0 in the tone systems. However, this is not what the results show: both Butlih and Biga have IF0 differences that are, overall, larger than average. These results indicate that, to the extent that IF0 is under speaker control, the effect of phonological profile may be limited, or perhaps overridden by some other factor—for example, the socio-cultural explanations offered by [29, 30]. The findings presented here emphasize the need to study IF0 in a cross-linguistically diverse sample of languages, in order to develop our theoretical understanding of this proposed universal.

This last point is highlighted by the diachronic relationships between tone and vowel height in several languages related to Butlih and Biga. Tonogenesis or tone changes conditioned by vowel height are typologically rare compared to the phonologization of co-intrinsic effects of consonants [31, 32]. Despite this rarity, a diachronic relationship between tone and vowel height is attested in several other eastern Austronesian languages: Ambel [33], Ma'ya [34] (both also RASH languages of Raja Ampat), Yerisiam [15] (a SHWNG language of northwest New Guinea), and Cèmuhî [35] (an Oceanic language of New Caledonia). One obvious phonetic source for the development of tone from vowel height is IF0 differences [31, 32]. The large IF0 differences of Butlih Salawati and Biga reported in this study may represent more general trends in the family. If so, this would indicate a longstanding and complex relationship between vowel height and F0 in Austronesian languages—the nature of which requires further data and research.

5. ACKNOWLEDGMENTS

This study was funded by a British Academy Postdoctoral Fellowship (PF19\100004) and a Moray Endowment Grant (University of Edinburgh) to L. Arnold, and an ERC Starting Grant (758605) to J. Kirby. Thanks to Dorlince Drimlol, Ahad Filis, Hasim Filis, Siti Hariani Klasin, Yunus Klasin, Penina Logof, Hodija Malakabu, Argenus Malbanat, Etresia Malbanat, Semuel Mandurun, Melianuis Samigita, and Aifa Warfandu for speech data, and Brisa Speier-Brito for help with data annotation.

6. REFERENCES

[1] D. H. Whalen and A. G. Levitt, "The universality of intrinsic F_0 of vowels," *J. Phon*, vol. 23, pp. 349–366, 1995.

[2] C. Ting, M. Clayards, M. Sonderegger, and M. McAuliffe, "The cross-linguistic distribution of vowel and consonant intrinsic F0 effects," PsyArXiv. March 24, 2023.

[3] D. H. Whalen, A. G. Levitt, P.-L. Hsiao, and I. Smorodinsky, "Intrinsic F0 of vowels in the babbling of 6-, 9-, and 12-month-old French- and English-learning infants," *JASA*, vol. 97, no. 4, pp. 2533–2359, 1995.
[4] B. Connell, "Tone languages and the universality of intrinsic F0: evidence from Africa," *J. Phon*, vol. 30, pp. 101–29, 2002.
[5] S. V. Hoof and J. Verhoeven, "Intrinsic vowel F0, the size of vowel inventories and second language acquisition," *J. Phon*, vol. 39, no. 2, pp. 168–177, 2011.
[6] R. L. Diehl and K. R. Kluender, "On the objects of speech perception," *Ecological Psychology*, vol. 1, no. 2, pp. 121–144, 1989.
[7] J. Kingston, "The phonetics and phonology of perceptually motivated articulatory covariation," *Lang Speech*, vol. 35, no. 1-2, pp. 99–113, 1992.
[8] J. Gandour and B. Weinberg, "On the relationship between vowel height and fundamental frequency: Evidence from esophageal speech," *Phonetica*, vol. 37, no. 5-6, pp. 344–354, 1981.
[9] N. Dyhr, "The activity of the cricothyroid muscle and the intrinsic fundamental frequency in Danish vowels," *Phonetica*, vol. 47, no. 3-4, pp. 141–154, 1990.
[10] A. Löfqvist, T. Baer, N. S. McGarr, and R. S. Story, "The cricothyroid muscle in voicing control," *JASA*, vol. 85, no. 3, 1989.
[11] H. Traunmüller, "Perceptual dimension of openness in vowels," *JASA*, vol. 69, no. 5, 1981.
[12] E. Fischer-Jørgensen, "Intrinsic F0 in tense and lax vowels wth special reference to German," *Phonetica*, vol. 47, no. 3-4, pp. 99–140, 1990.
[13] P. Hoole and K. Honda, "Automaticity vs. feature-enhancement in the control of segmental F0," in *Where do phonological features come from?*, N. G. Clements and R. Ridouane, Eds. John Benjamins, 2011, pp. 131–172.
[14] B. Remijsen, *Word-prosodic systems of Raja Ampat languages*. Utrecht: LOT, 2001.
[15] D. Kamholz, "Austronesians in Papua: Diversification and change in South Halmahera–West New Guinea," Ph.D. dissertation, University of California, Berkeley, 2014.
[16] L. Arnold, "Lexical tone in Metnyo Ambel," *Ocean. Linguistics*, vol. 57, no. 1, pp. 199–220, 2018.
[17] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech and Language*, pp. 392–410, 2017.
[18] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.
[19] L. Bombien, R. Winkelmann, and M. Scheffers, "wrassp: an R wrapper to the ASSP Library," *R package version 0.1*, vol. 8, 2018.
[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, 2017.
[21] R. Lenth, *emmeans: Estimated marginal means, aka Least-squares means*, 2022.
[22] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022.
[23] H. M. Hanson, "Effects of obstruent consonants on fundamental frequency at vowel onset in English," *JASA*, vol. 125, no. 1, pp. 425–441, 2009.
[24] J. Kirby and D. R. Ladd, "Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages," *JASA*, vol. 140, no. 4, pp. 2400–2411, 2016.
[25] Q. Luo, "Consonantal effects on F0 in tonal languages," Ph.D. dissertation, Michigan State University, 2018.
[26] D. R. Ladd and K. E. A. Silverman, "Vowel intrinsic pitch in connected speech," *Phonetica*, vol. 41, no. 1, pp. 31–40, 1984.
[27] M. Pettorino, "Intrinsic pitch of vowels: An Experimental study on Italian," *Procs. 11th ICPhS*, pp. 138–141, 1987.
[28] J. J. Ohala and B. W. Eukel, "Explaining the intrinsic pitch of vowels," in *In honor of Ilse Lehiste*, R. Channon and L. Shockey, Eds. Dordrecht: Foris, 1987, pp. 207–215.
[29] E. Jacewicz and R. A. Fox, "Intrinsic fundamental frequency of vowels is moderated by regional dialect," *JASA*, vol. 138, no. 4, pp. EL405–EL410, 2015.
[30] Y. F. Holt, E. Jacewicz, and R. A. Fox, "Sociocultural influences on fundamental frequency in vowels: The case of African American English," in *Procs. 19th ICPhS*, 2019, pp. 874–878.
[31] J. Kingston, "Tonogenesis," in *Blackwell Companion to Phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds. Oxford: Blackwell publishing, 2011, pp. 2304–2333.
[32] A. Michaud and B. Sands, "Tonogenesis," in *Oxford research encyclopedia of linguistics*, M. Aronoff, Ed. Oxford University Press, 2020.
[33] L. Arnold, "Highs and lows: A previously unattested tone split from vowel height in Metnyo Ambel," *Trans Philol Soc*, vol. 118, pp. 141–158, 2020.
[34] ——, "A preliminary archaeology of tone in Raja Ampat," in *Contact and substrate in the languages of Wallacea, part 2*, ser. NUSA: Linguistic Studies of Indonesian and Other Languages in Indonesia 64, A. Schapper, Ed., 2018, pp. 7–39.
[35] J.-C. Rivierre, "Tonogenesis and evolution of tonal systems in New Caledonia, the example of Cèmuhî," in *Procs. of the symposium Cross-linguistic studies of tonal phenomena*, S. Kaji, Ed. Tokyo: Institute for the Study of Languages/Cultures of Asia/Africa, 2001, pp. 23–42.

MAPPING PHONETIC VARIATION TO PHONOLOGICAL UNITS: FEATURE SHARING IN COMOX-SLIAMMON (SALISH) VOWELS

Gloria Mellesmoen and Amanda Cardoso

University of British Columbia
gloria.mellesmoen@ubc.ca; amanda.cardoso@ubc.ca

ABSTRACT

The way that sounds interact with one another is important for understanding how gradient phonetic variation maps to discrete phonological units, especially in cases where there is a significant overlap in phonetic variation across different sound categories, such as with vowels in Comox-Sliammon, an underrepresented Salish language spoken in Canada. We consider vowels in Comox-Sliammon in context (phonological environment as defined by the identity of adjacent consonants) using predicted vowel trajectories from Generalized Additive Mixed Models. Vowels are better dispersed (and differentiated) by derived feature specification (where features are shared from adjacent consonants) than by underlying feature specifications (where vowel identity alone is considered). Consequently, we gain a fuller understanding of the phonetic-phonological mapping of sound systems by building an understanding of the phonological grammar (context) into our phonetic analysis (trajectories). This ultimately allows us to posit and test data-driven hypotheses about the phonological units for users of the language.

Keywords: fieldwork, Salish, phonetic variation, phonological mapping, features

1. INTRODUCTION

Sounds are produced in context, and adjacent (and non-adjacent) sounds can influence each other [1], creating phonetic variation. Phonetic variation must be mapped back to phonological units to create meaning out of the phonetic variation.

When there is a clearer one-to-one mapping between an underlying phoneme and the phonetic realization, sounds are more evenly dispersed in acoustic space and form relatively discrete categories. In this case, phonetic variation can be differentiated by phonemic units with little information about phonological environment (i.e., context). However, when there is substantial overlap in the phonetic realizations or a less clear mapping between a sound and acoustic space, it may not be possible to see discrete categories without considering the environment.

If a phoneme corresponds to a broad distribution in acoustic space and overlaps substantially with other phonemes, we predict that dispersion will be clearest between phonologically discrete allophones (i.e., context-dependent units). These context-dependent units are an intermediate unit between a phoneme and its phonetic implementation (see, for example, [2]). Assuming a modular feed-forward architecture of phonology (e.g, [3]), the intermediate allophone is derived by the grammar, which may change or alter the phonological properties of a sound (e.g., feature sharing). Where the application of the grammar results in more substantial changes to a sound resulting in more phonetic variation (i.e., broader distribution in acoustic space), the phonetic implementation may result in variation that shows little differentiation between phoneme categories, but maintains clearer distinctions between allophones.

In languages with one-to-many mappings between phonemes and allophones, describing the acoustic variation by phonemic units is less informative, which suggests that it would be difficult for language users to differentiate the phonemic units. Here, we test whether considering the phonological grammar (which is language-specific knowledge available to a speaker) provides insight into acoustic overlap previously observed in the Comox-Sliammon vowel system [4]. We analyze F1 / F2 trajectories of Comox-Sliammon vowels to examine how well they are differentiated when classified by phoneme (underlying vowel) compared to allophone (derived vowel) to explore the mapping between phonetic variation and phonological units in an underrepresented language.

2. VOWELS IN COMOX-SLIAMMON

Comox-Sliammon, a Central Salish language spoken in Canada, has four vowels (/a/, /i/, /u/, and /ə/), which may be considered a minimal vowel

system (see [5]). The vowel /u/ is differentiated by position as a back (and rounded) vowel, while /a/, /i/, and /ə/ overlap considerably in F1 and F2 acoustic space corresponding to front vowels [4].[1] Adjacent consonants strongly influence the realization of /a/, /i/, and /ə/, and the inclusion of information about vowel environment greatly enhances contrast (i.e., dispersion of vowels) in the system [6].

We test the hypothesis that the unit of contrast in the Comox-Sliammon vowel system is not the underlying phoneme, but instead a derived allophone with a combination of phonological features inherent to the underlying vowel and those received in context (i.e., features shared from an adjacent consonant). Crucially, we focus on context rules (features received from an adjacent segment which affect the segment throughout its duration), rather than phonetic rules (co-articulation realized at the transition between two segments) [7].

We assume that consonant place features may be shared with adjacent vowels, consistent with earlier analyses of Comox-Sliammon (e.g., [8]). We adopt a feature geometry that assumes a unified set of features for consonants and vowels (e.g., [9]). We use [LOW] as a stand-in for either a [radical] feature or a [low] feature under [dorsal], and [FRONT] as a stand-in for a [-back] or [high] feature.

Table 1 provides the inherent phonological features associated with each of the three front vowels in Comox-Sliammon, alongside the relevant consonants.[2] The vowel /ə/ is underspecified for place prior to the phonological derivation.

Underlying Features		
Feature	Vowel	Consonants
[LOW]	/a/	/q/, /qʷ/, /qʼ/, /qʷʼ/, /χ/, /χʷ/
[FRONT]	/i/	/tʃ/, /tʃʼ/, /dʒ/, /ʃ/, /j/, /ˀj/
none	/ə/	
Derived Features		
Vowel Feature(s)	Environments	
[LOW]	$əC_{low}$, $aC_{elsewhere}$	
[FRONT]	$əC_{front}$, $iC_{elsewhere}$	
[FRONT, LOW][3]	iC_{low}, aC_{front}	
none	$əC_{elsewhere}$	

Table 1: Underlying (top) and derived (bottom) features for Comox-Sliammon vowels

We consider F1 / F2 over the duration of the vowel for each of the environments listed in Table 1 to test the following hypotheses (see [6, 10]):

1. **Underlying Phoneme Hypothesis:** Vowels are best differentiated by underlying feature specifications (i.e., vowel categories).
2. **Derived Allophone Hypothesis:** Vowels are best differentiated by derived feature specifications (i.e., vowels in context).

If the vowels are best categorized by phoneme categorization, we expect discrete F1 / F2 trajectories for each vowels. However, if the vowels are best categorized by allophones, we expect discrete F1 / F2 trajectories for each vowels *only* when plotted by allophone (not by phoneme).

3. METHODS

The data analyzed were produced by two speakers of Comox-Sliammon in their 80s: one male (labelled: f) and one female (labelled: e). The tokens were selected from a larger set of recordings produced for an e-dictionary [11]. The selected vowels all bear primary stress (as the first vowel in the word) and represent each of the target environments (see Table 1). The present analysis considers the front vowels, which show considerable overlap (/i/, /a/, and /ə/). As a vowel that is both back and round, /u/ overlaps less with the other vowels in F1 and F2 acoustic space.

Vowels were hand-aligned by two trained fieldwork linguists and then a praat script was used to extract 5 equidistant F1 and F2 measurements across the vowel (hereafter "vowel trajectories"). We use Generalized Additive Mixed Models (GAMMs) for an exploratory analysis of the F1 and F2 vowel trajectories by visualizing the predicted values for these trajectories. Separate models are run for each formant by speaker. GAMM models include non-normalized formant values (dependent variable) with parametric terms for vowel categories and derived features, smooth terms for trajectories by vowel categories and derived features, and random smooths for word. Table 2 shows token count across both speakers by vowel categories and derived features. While speaker e has more tokens (e=193; f=147), the number per vowel categories and context is similar in proportion.

	/i/	/a/	/ə/	**Feature Totals**
Front	74	0	45	119
Front.Low	37	45	0	82
Low	0	62	54	116
None	0	0	23	23
Vowel Totals	111	107	122	

Table 2: Total tokens counts by vowel category (columns) and derived features (rows)

This is an exploratory, not confirmatory, investigation, so statistical significance (e.g., p-values) is not discussed. Rather, we interpret

differences in trajectories through visualizations, which is recommended for exploratory investigation (e.g., [12]).

4. RESULTS

We see a large degree of overlap (i.e., little differentiation) in model predicted F1 and F2 trajectories across all vowel categories, which does not support the first hypothesis (F1 and F2 are best explained with reference to the underlying features associated with the vowel phonemes). In the left panels in Figure 1, F1 and F2 predicted vowel trajectories are plotted by underlying phoneme. These clearly demonstrate that the vowels are overlapped across the majority of the trajectory.

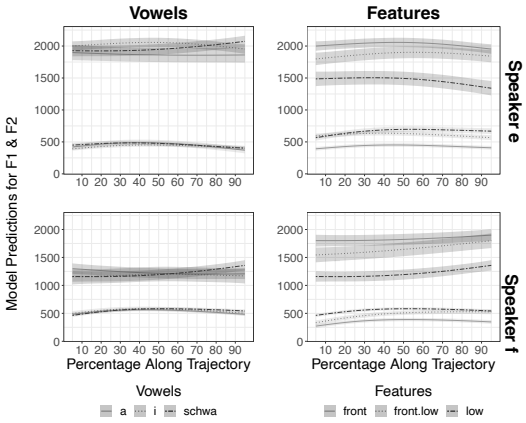


Figure 1: F1 and F2 model predicted trajectories with confidence intervals for each speaker (e - top row; f - bottom row) plotted by vowel category (left panels) and derived features (right panels)

The right panels in Figure 1 show the F1 and F2 predicted vowel trajectories plotted by categorically discrete allophones (i.e., derived features). For clarity, underspecified tokens (labelled: “none") are omitted in Figure 1. The panels on the right support hypothesis 2 (F1 and F2 are best explained with reference to derived feature specification). There is a clearer separation between the trajectories when plotted by derived feature specification than by underlying specification (phoneme).

The trajectories for derived feature groups occupy phonetic spaces consistent with the proposed features. The trajectories for the [FRONT] allophone have the highest predicted F2 (and lowest predicted F1), while the trajectories for the [LOW] allophone have the highest predicted F1 (and lowest predicted F2). Unsurprisingly, the trajectories for the [FRONT.LOW] allophone sits between the other two in both F1 and F2, moving in the expected direction (i.e., towards [FRONT] in the F2 space, and towards [LOW] in the F1 space).

While the overall findings hold for both speakers, we also find interspeaker differences. These differences are seen in Figure 2 where model predicted F1 (bottom row) and F2 (top row) trajectories are plotted by derived feature specifications in greater detail, including the underspecified (“none”) tokens. We compare the trajectory shapes and amount of overlap across derived feature groups in this figure, but as values are not normalised, we do not compare the Hz values themselves across speakers.

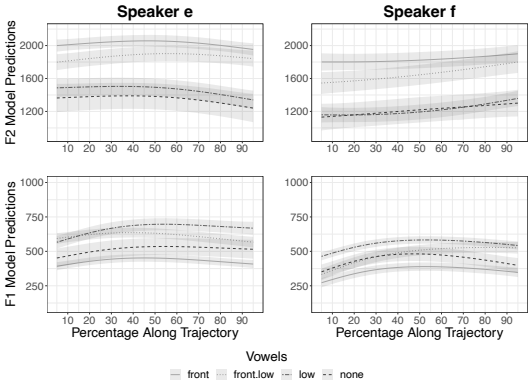


Figure 2: F1 (bottom row) and F2 (top row) model predicted trajectories with confidence intervals for each speaker (e - left panels; f - right panels) plotted by derived features

Generally, we see more movement over the trajectories for speaker f and more stable, flatter trajectories for speaker e. In the F1 space we see the [FRONT.LOW] trajectories are closer to the [LOW] trajectories and overlapping at points along the trajectory. However, there is more overlap at the beginning of the trajectory for speaker e, but more at the end for speaker f. For speaker f, the [FRONT.LOW] trajectory overlaps with the “none” trajectory, but not with the [FRONT] trajectory, and only the end of the [FRONT.LOW] trajectory overlaps with the [LOW] trajectory. Whereas, speaker e has more overlap between the [FRONT] and “none” trajectories, and between the [FRONT.LOW] and [LOW] trajectories. However, there is little overlap between the “none” and [FRONT.LOW] trajectories for speaker e.

Fewer interspeaker differences are observed for F2 trajectories. For both speakers, the [LOW] and “none” trajectories overlap (though speaker f’s trajectories are more similar to each other) and most of the [FRONT] and [FRONT.LOW] trajectories overlap, except at the beginning.

While specific patterns are found, there is evidence of stronger effects at the edges (onset and offset) of the vowel. The onset and offset of trajectories are often different from each other and from the midpoints. This is likely, in part due to phonetic coarticulatory effects, which are expected. The effect of the adjacent consonants is not *just* coarticulation though, as each of the derived allophones differs at the midpoint as well, supporting their status as categorically discrete phonological units. The difference between the midpoint and rest of the trajectory is information that would not be captured by midpoint measurements, supporting the descriptive and analytical benefit of plotting formant trajectories.

5. DISCUSSION

The current analysis supports the derived allophone hypothesis: there is less acoustic overlap when vowels are plotted by derived feature specification (i.e., allophones) than underlying vowel feature specification (i.e., phonemes) for these speakers. While it is often suggested that phonetic variation maps directly onto phonemic categories, this data suggests that phonetic variation is actually mapped to subphonemic categories (i.e., allophones).

For Comox-Sliammon, vowels cannot be clearly distinguished in acoustic space by underlying phoneme. Considering an intermediate phonologically derived allophone allows for a clearer mapping between phonological units and phonetic realization. This demonstrates that understanding the grammar of a language is integral to understanding the phonetic variation. In order to produce accurate descriptions and analyses of an underdocumented language, it is crucial to consider the language-specific knowledge (such as the phonological grammar) that is available to speakers in production and perception.

We also find that the proposed derived allophones (based on phonological features) are supported by the phonetic realizations. The most front (i.e., highest F2) trajectories are the [FRONT] trajectories and the lowest (i.e., highest F1) trajectories are the [LOW] trajectories. The [FRONT.LOW] trajectories sit between these, but closer to the [FRONT] trajectories for F2 and [LOW] trajectories for F1. The positions of each trajectory is consistent with the expected relationship between formants and vowel height and backness.

Vowel trajectories also show the difference between the phonetic coarticulatory effects of adjacent consonants at the edges and the phonological effects of feature sharing, that are evident throughout the trajectory (including midpoint). The edges of the vowel trajectories often show stronger effects, such as lowering and fronting in the [FRONT.LOW] trajectories, which are likely due to phonetic coarticulation with the adjacent consonants. On the other hand, wholesale changes to the trajectory, such as the entire [LOW] trajectory being lower than the other trajectories, likely demonstrate phonological distinctions based on the allophone as a discrete phonological unit.

Although there are differences between speakers, such as speaker f having a greater degree of movement across the vowel trajectories, these differences do not affect the overall conclusions. Both speakers show significant overlap in the F1 and F2 trajectories for different vowel phonemes, as well as separation between the derived feature group trajectories. Additionally, the general patterns of the derived feature group trajectories, such as the [FRONT.LOW] trajectories being positioned between the [FRONT] and [LOW] trajectories, are consistent across speakers. Therefore, the main findings are still valid even when taking individual speaker differences into account.

We find that the derived allophone is a more useful phonological unit to explore phonetic variation in Comox-Sliammon. This has broader implications for language description and analysis, which feed into language revitalization efforts. It is not sufficient to teach the vowels of the language without discussion of the context and interactions between vowels and adjacent consonants. This highlights the importance of considering the grammar of a language when conducting phonetic analysis (especially with little previous work to guide an exploration when considering underrepresented languages).

The derived allophone hypothesis has implications beyond the description and analysis of Salish vowels. Precision is not lost when describing vowel systems by categorically discrete allophones, rather than phonemes; if labelling by phoneme is sufficient for a language, the number of vowels and labels can remain the same due to a 1-to-1 mapping between phoneme and derived allophone. The derived allophone allows for more precision when describing vowel systems that are less well-documented and where the phonological grammar creates substantial overlap between phonemic categories in acoustic space. The derived allophone approach integrates language-specific patterns in sub-phonemic organization, which is useful in the cross-linguistic description of vowel systems.

6. ACKNOWLEDGEMENTS

We thank everyone involved with the ʔayʔaǰuθəm dictionary, which is the result of a project led by Henry Davis (SSHRC Insight Grant 435-2016-1694). In particular, this work would not be possible without the time and the patience of the speakers (Elsie Paul and Freddie Louie), as well as the work of the dictionary team. We also thank Marianne Huijsmans for help with preparing the data for analysis.

7. REFERENCES

[1] W. Hardcastle and N. Hewlett, *Coarticulation: Theory, Data and Techniques*, ser. Cambridge Studies in Speech Science and Communication. Cambridge University Press, 1999.

[2] A. J. Dinkin and R. Dodsworth, "Gradience, allophony, and chain shifts," *Language Variation and Change*, vol. 29, no. 1, pp. 101–127, 2017.

[3] J. Fruehwald, "The phonological influence on phonetic change," Ph.D. dissertation, University of Pennsylvania, 2013.

[4] G. Mellesmoen and M. Huijsmans, "The relationship between pronunciation and orthography: Using acoustic analysis as a practical illustration of ʔayʔaǰuθəm (Comox-Sliammon) vowel quality," in *Proceedings of the 19th International Congress of Phonetic Sciences*, M. T. Sasha Calhoun, Paola Escudero and P. Warren, Eds., 2019, pp. 979–983.

[5] C. Anderson, "Consonant colour and vocalism in the history of Irish," Ph.D. dissertation, Poznań: Adam Mickiewicz University in Poznań, 2016.

[6] G. Mellesmoen and A. Cardoso, "Contrast in the Comox-Sliammon vowel system," *Journal of Phonetics*, under review.

[7] P. A. Keating, "Underspecification in phonetics," *Phonology*, vol. 5, no. 2, pp. 275–292, 1988.

[8] S. J. Blake, "On the distribution and representation of schwa in Sliammon (Salish): Descriptive and theoretical perspectives," Ph.D. dissertation, University of British Columbia, 2000.

[9] G. N. Clements *et al.*, "Place of articulation in consonants and vowels: a unified theory," *Working papers of the Cornell phonetics laboratory*, vol. 5, no. 77, p. 123, 1991.

[10] G. Mellesmoen, "Modularity and the allophone in the Comox-Sliammon (Salish) vowel system (LabPhon17 poster)," University of British Columbia, 2020. [Online]. Available: https://www.mellesmoen.ca/research/talks/labphon17

[11] "ʔayʔaǰuθəm dictionary," https://mothertongues.org/ayajuthem/.

[12] M. Sóskuthy, "Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction," 2017. [Online]. Available: https://arxiv.org/abs/1703.05339

[1] We consider /a/, /i/, and /ə/ to be front vowels in Comox-Sliammon given previous phonetic work [4], as well as phonological analysis [8].

[2] The relevant consonants given in Table 1 are only a subset of consonants in Comox-Sliammon; we only list the consonants that share the identified feature with the vowel.

[3] A further prediction is that it should be possible to have a /ə/ map to a [FRONT, LOW] allophone as well if between a [FRONT] and a [LOW] consonant (e.g., $C_{front}əC_{low)}$. This is a question for future study.

EVALUATING FORCED ALIGNMENT FOR UNDER-RESOURCED LANGUAGES: A TEST ON SQULIQ ATAYAL DATA

Chi-Wei Wang[1], Bo-Wei Chen[1], Po-Hsuan Huang[1], Ching-Hung Lai[2], Chenhao Chiu[1,3]

[1]Graduate Institute of Linguistics, National Taiwan University; [2]School of Medicine, National Cheng Kung University; [3]Neurobiology and Cognitive Science Center, National Taiwan University

r09142007@ntu.edu.tw, r09142001@ntu.edu.tw, r09142003@ntu.edu.tw, i54051092@gs.ncku.edu.tw, chenhaochiu@ntu.edu.tw

ABSTRACT

Trainable forced alignment offers feasible solutions to document under-resourced languages. This study aims to assess the performances of a Montreal Forced Aligner (MFA) trained model using a small scale of phonetically transcribed field data in Squliq Atayal, an endangered Austronesian language spoken in Taiwan. Evaluations were implemented by comparing MFA outputs with manual annotations based on (1) the accuracy measurements on the interval boundaries of each segment, and (2) the acoustic measurements, by fitting the formant trajectories through the most common vowels [a, i, u] with generalized additive mixture models (GAMMs). The results showed that the general agreement reached 73.23% of accuracy when with a tolerance of 30 ms misalignment, and no statistical significance was found between the formant trajectories except for F1 trajectories of [a] and F2 trajectories of [u], revealing that MFA outcomes were fairly consistent with manual annotations when little but comprehensively labeled data were provided.

Keywords: forced alignment, phonetic fieldwork, language documentation, Squliq Atayal

1. INTRODUCTION

Recently, forced alignment (FA) systems have been providing feasible solutions to enhance the workflow of phonetic analyses. A prototypical pipeline of FA includes (1) integrating the audio files consisting of speech signals and the TextGrid files containing orthographic transcriptions at the utterance level to form a corpus. (2) Conducting the alignment with the help of the acoustic model, which calculates how likely a phone is given the acoustic features; and the pronunciation dictionary, which provides the reference for the grapheme-to-phoneme (G2P) mapping. (3) Generating a time-aligned TextGrid file with both word and phone tiers as the output.

Previous works on FA for under-resourced languages often adapted existing acoustic models made for well-documented languages [1, 2, 3, 4]. Nevertheless, the feasibility of such manipulation may be reduced due to the mismatch of sound inventories or orthographic systems between the two distinct languages. In order to solve the cross-linguistic issues, with the assistance of automatic speech recognition (ASR) toolkits such as Kaldi [5] or HTK [6], researchers can also consider training a new, language-specific acoustic model based on a relatively small corpus [2, 7, 8, 9]. Since all phonological and phonetic clues of the target language are embedded in a customized model, it is likely to outperform the pre-trained model in terms of the alignment results.

The overall performance of the FA can be assessed once the alignment is completed. In order to evaluate the FA outcomes, corresponding manually-aligned data are required and treated as the reference. There are several evaluation methods. One of the most common and essential ways is to measure the agreement (in percentage) at the onset and offset boundaries [1, 3, 4, 7, 8], which refers to the proportion of the FA and manual boundaries from the same segment that agree within a given threshold of tolerance (in ms). Intuitively, the larger the threshold, the better the agreement.

Alternatively, other studies proposed that such temporal-based accuracy measurement of the alignment can also be performed by calculating the overlap rate (in percentage) and the robustness (in ms) [9, 10, 11, 12, 13]. The former refers to the proportion of overlap between the intervals established by FA and human annotators. Similar to the analysis of agreement, the greater the overlap, the higher the accuracy; while the latter focuses on the actual displacement between the midpoints of the FA and manual intervals. Specifically, the mean

displacement of error tokens (i.e., the segments whose midpoints lie beyond a pre-determined threshold; e.g., 20 ms) was calculated; the less the displacement, the better the FA performances.

Furthermore, FA performances can also be evaluated by measuring acoustic features of the output [2, 14]. For instance, the location of pitch (F0) peak through the words, the vowel space constructed by the first and second formants (F1 and F2), and the consonant VOT. An FA model would be considered robust when statistical significance is absent among these measurements.

For the objectives of this study, an FA acoustic model is trained based on a small scale of phonetically transcribed field data in Squliq Atayal, an endangered Austronesian language spoken in Taiwan. The model performances are evaluated by the aforementioned measurements, including the agreement (AG), the overlap rate (OR), the midpoint displacement (MD, inspired by the robustness), and the formant measurements.

2. METHODOLOGY

The Montreal Forced Aligner (MFA) [15] was employed in this study. MFA is a Kaldi-based, open-source, cross-platform system that not only provides pre-trained pronunciation dictionaries and acoustic/G2P models in multiple languages but also offers a sophisticated API for training new dictionaries and models, operated in a user-friendly command-line interface.

2.1. Squliq Atayal dataset and model training

The dataset adopted as the source of the FA acoustic model training corpus was the complete recordings (in .wav format, mono at 16-bit/44.1kHz sampling rate) from a series of fieldwork sessions. All elicited utterances produced by one female Squliq Atayal native speaker (excluding the speech produced in the contact language) were manually labeled at both word and phone levels by two trained phoneticians using Praat [16].

To build the pronunciation dictionary, since every TextGrid file in the current dataset has a phonetically transcribed phone tier, the pronunciation dictionary was generated using Python [17] by concatenating the word and phone tiers in each manually-annotated transcription. Consequently, all words that occurred in the corpus were included and combined as a comprehensive pronunciation dictionary.

The acoustic model training and alignment procedures were operated simultaneously, with 31 pairs of audio/TextGrid files and the pronunciation dictionary as the inputs. Note that each TextGrid file could only include the word tier, as the MFA would recognize multiple tiers in a file as different speakers. The output files included a .zip file as a newly-trained acoustic model and 31 aligned TextGrid files. The overall workflow took approximately 40 minutes on a Linux server.

2.2. Evaluation of the MFA outputs

Once the aligned TextGrid file was retrieved, they were loaded using the `PraatIO` [18] package in Python along with their manually-annotated counterparts. Here, a customized algorithm was established to locate the identical segment in both manual and MFA phone tiers automatically. Nevertheless, an alignment issue was discovered: since the manual annotations of the segments were narrow transcriptions which reflect detailed surface representation of sounds, many allophonic variants of identical words were consequently appended into the pronunciation dictionary. Such variants could severely impact the workability of the algorithm, leading to segment mismatches between the manual and MFA phone tiers. In this case, only one pair of audio/TextGrid files was included in the current study (n = 2348; duration = 1065.75 seconds).

Regarding the evaluation procedures, the temporal-based measurements of the alignment accuracy (i.e., AG, OR, and MD) were calculated by comparing MFA outputs with manual annotations using Python. As for the acoustic measurements, the F1 and F2 frequencies among each phone interval from both MFA outputs and manual annotations were obtained using a customized Praat script [19] and were then imported to R [20] for statistical modeling. Particularly, the formant trajectories through the most common vowels [a, i, u] constructed by the 30 data points were fitted by six (2 formants * 3 vowels) generalized additive mixture models (GAMMs) [21]. GAMMs are particularly useful for analyzing non-linear relationships between variables. They also involve fitting smooth functions to the predictor variables (i.e., manual vs. MFA) and accommodate non-normal response variables (i.e., the acoustic data points), making it applicable to our study.

3. RESULTS

Figure 1 presents an example from the MFA output merged with its corresponding manual transcriptions. The first two tiers were annotated by a human annotator, and the lower two tiers were aligned by MFA. By eyeballing, the automatic

alignment results between human and MFA were fairly consistent, except for the unexpected empty intervals generated between words. This may be the case that MFA prefers utterance-level to word-level transcriptions, as suggested by its documentation.

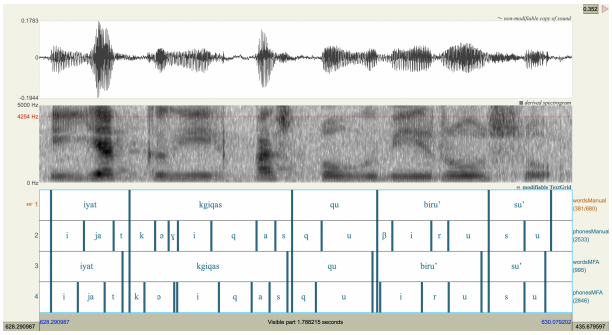


Figure 1: Annotated results by a human annotator (top two tiers) and MFA (bottom two tiers).

3.1. Measurements of alignment accuracy

Table 1 lists the agreements (AGs) at different thresholds, separated by vowel and consonant onset/offset; in addition to the mean AGs at both boundaries, as well as the overall AGs. As revealed by Table 1, better AGs were observed in vowel onsets than offsets. On the other hand, better AGs were observed in consonant offsets than onsets when with a higher tolerance. One possible reason is that the canonical syllable structure of Squliq Atayal is C(G)V(C), suggesting that both consonant clusters and hiatus are prohibited [22, 23]. As such, there should be mostly consonant-vowel sequences that occur throughout the entire phone tier. Moreover, the mean AGs of consonants outperform that of vowels, which is different from the previous studies [1, 2]. For the criteria of implementing the thresholds, according to the FA-related literature, 20 ms is considered the most robust threshold [24]; while the literature on under-resourced languages also reports at 30, or even 50 ms thresholds [1].

Table 2 lists the mean overlap rates (ORs) and midpoint displacements (MDs) of the following five different categories: vowels, consonants, overall performances, and the most common vowels [a, i, u]. Note that despite the algorithm suggested in [9, 13], the calculation of MDs here includes all segments instead of the error tokens pre-defined by a given threshold, in order to provide a relatively comprehensive view of the measurements. Overall, the results are in correspondence to those of AGs, where consonants have slightly higher ORs and narrower MDs than vowels. The results of ORs/MDs are also consistent among the three most common vowels. Specifically, [i] are usually better aligned; while [a] are less aligned.

Table 2: The mean overlap rates (ORs) and midpoint displacements (MDs) of different segment categories.

Category	OR	MD
vowel	53.26%	35.03 ms
consonant	55.20%	29.28 ms
overall	54.29%	31.99 ms
[a]	53.08%	39.63 ms
[i]	71.56%	19.43 ms
[u]	61.88%	28.05 ms

3.2. Measurements of acoustic features

Figure 2 renders the formant trajectories through vowels [a, i, u] fitted by GAMMs. The duration was normalized based on the 30 data points along with the shaded areas representing 95% CIs. The red dashed lines illustrate the formant trajectories of MFA-aligned vowels, and the blue solid lines depict those of manually-annotated vowels. Crucially, the overlap between the fits and confidence intervals suggest no difference between manual and MFA annotations.

Table 1: The agreements (AGs) of vowels and consonants at different thresholds.

Threshold	Vowel			Consonant			Overall
	Onset	Offset	Mean	Onset	Offset	Mean	
10 ms	30.62%	23.94%	27.28%	54.15%	22.64%	38.40%	32.84%
20 ms	56.82%	46.97%	51.90%	65.03%	49.15%	57.09%	54.49%
30 ms	79.67%	66.03%	72.85%	73.81%	73.41%	73.61%	73.23%
40 ms	89.97%	69.74%	79.86%	77.76%	82.59%	80.18%	80.02%
50 ms	91.15%	71.91%	81.53%	79.61%	84.21%	81.91%	81.72%

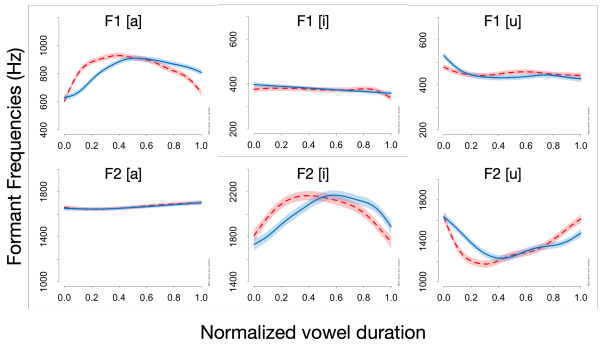


Figure 2: Formant trajectories over the normalized [a, i, u] fitted by GAMMs (MFA = red dashed lines; manual = blue solid lines).

As revealed in Figure 2, no statistical significance was found between MFA outputs and manual annotations except for F1 trajectories of [a] and F2 trajectories of [u]. Note that in the F2 [i] condition, despite that only limited overlap was found between two annotations, no statistical significance was reported. Overall, the acoustic results are not only consistent with the OR/MD results that the alignment of [i] outperforms that of [a] and [u], but also positively support the reliability of the current MFA model.

4. DISCUSSION

The discrepancy between current work and previous studies regarding the assertion that vowels were associated with better alignments than consonants may be accounted for by the effect of segment types. Table 3 further divides the vowel and consonant categories into the following six types: full vowels, weak vowel, complex vowels, plosives, nasals & liquids, and fricatives & affricates. In particular, full vowels include the canonical vowels [a, i, u, e, o]; weak vowels consist of pretonic reduced vowels [ə, ɨ]; while complex vowels are the combinations of vowels [a, i, u] and onglides/offglides [w, j].

Table 3: The mean ORs and MDs of different segment types.

Type	OR	MD
full vowels	60.21%	30.21 ms
weak vowels	39.07%	26.52 ms
complex vowels	38.54%	71.27 ms
plosives	60.86%	18.62 ms
nasals & liquids	46.55%	42.85 ms
fricatives & affricates	54.26%	33.04 ms

According to Table 3, it is obvious that complex vowels have the lowest mean OR and the largest mean MD. An additional linear regression analysis also reveals a statistical significance [$F(5, 2342) = 43.05$, $p < .001$ ***]. As a result, the relatively low performance of vowel alignments is likely to be affected by those segments classified as complex vowels. Previous studies such as [10, 25] also suggested that FA algorithms vary overtly in terms of their accuracy in different types of segments – specifically, semivowels had the greatest displacement among other segment types, which echoes our findings.

On the other hand, the asymmetry between the absence of statistical significance and the less overlapping fitted trajectories occurred in the GAMM results of F2 [i] condition in Figure 2 was likely to be accounted for by the normalization of the vowel duration. As shown in Figure 3, the two critically-misaligned vowels [a] in the word *kinlabang* "very wide" may be resulted from the intrinsic difference in duration. After normalization, formant trajectories might end up being stretched or compressed horizontally. Not to mention the displacement issue – the captured contours could be substantially off. These factors eventually influenced the fitting results of GAMMs.

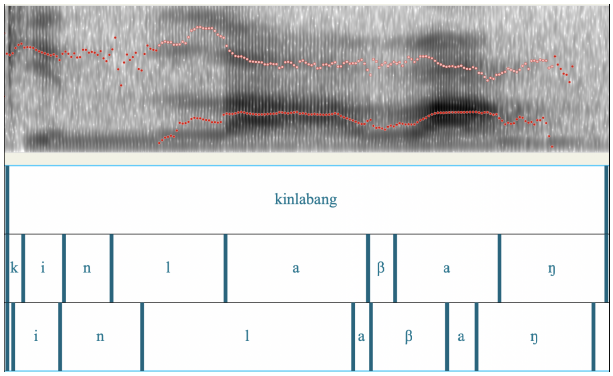


Figure 3: The F1 and F2 trajectories over the word *kinlabang* "very wide". Regarding the two phone tiers, upper tier: manual annotation; lower tier: MFA output.

Overall, the current results reveal that MFA outputs fairly correspond to manual annotations when little but comprehensively labeled data were provided. Furthermore, with the employment of the G2P model training, a function provided by the MFA, the minimum requirement of corpus size for constructing an acoustic model is also worth examining via the data recursion method proposed by [9]. This would call for future studies.

5. REFERENCES

[1] C. Jones, W. Li, A. Almeida, and A. German, "Evaluating cross-linguistic forced alignment of conversational data in north australian kriol, an under-resourced language," *Language Documentation and Conservation*, pp. 281–299, 2019.

[2] S. Babinski, R. Dockum, J. H. Craft, A. Fergus, D. Goldenberg, and C. Bowern, "A robin hood approach to forced alignment: English-trained algorithms and their use on australian languages," in *Proceedings of the Linguistic Society of America*, 2019, pp. 1–12.

[3] N. J. Young and M. McGarrah, "Forced alignment for nordic languages: Rapidly constructing a high-quality prototype," *Nordic Journal of Linguistics*, pp. 1–27, 2021.

[4] J. Leinonen, N. Partanen, S. Virpioja, and M. Kurimo, "Semiautomatic speech alignment for under-resourced languages," in *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, 2022, pp. 17–21.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011, pp. 1–4.

[6] S. J. Young, "The htk hidden markov model toolkit: Design and philosophy," Ph.D. dissertation, Department of Engineering, University of Cambridge, 1993.

[7] L. M. Johnson, M. Di Paolo, and A. Bell, "Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data," *Language Documentation & Conservation*, vol. 12, pp. 80–123, 2018.

[8] K. Tang and R. Bennett, "Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan)," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 2019, pp. 1719–1723.

[9] S. Gonzalez, C. Travis, J. Grama, D. Barth, and S. Ananthanarayan, "Recursive forced alignment: A test on a minority language," in *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, 2018, pp. 145–148.

[10] S. Paulo and L. C. Oliveira, "Automatic phonetic alignment and its confidence measures," in *International Conference on Natural Language Processing*, 2004, pp. 36–44.

[11] R. Fromont and K. Watson, "Factors influencing automatic segmental alignment of sociophonetic corpora," *Corpora*, vol. 11, no. 3, pp. 401–431, 2016.

[12] R. Coto-Solano and S. F. Solórzano, "Comparison of two forced alignment systems for aligning bribri speech." *CLEI Electron. J.*, vol. 20, no. 1, pp. 2–1, 2017.

[13] D. Barth, J. Grama, S. Gonzalez, and C. Travis, "Using forced alignment for sociophonetic research on a minority language," *University of Pennsylvania Working Papers in Linguistics*, vol. 25, no. 2, article 2, 2020.

[14] R. Billington, H. Stoakes, and N. Thieberger, "The Pacific Expansion: Optimizing Phonetic Transcription of Archival Corpora," in *Interspeech*, 2021, pp. 4029–4033.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2021. [Online]. Available: http://www.praat.org/.

[17] Python Core Team, *Python: A dynamic, open source programming language*, Python Software Foundation, 2022. [Online]. Available: https://www.python.org/.

[18] T. Mahrt, "Praatio," 2016. [Online]. Available: https://github.com/timmahrt/praatIO.

[19] W. Styler, "On the acoustical features of vowel nasality in english and french," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/.

[21] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[22] P. J. kuei Li, "The phonological rules of atayal dialects," *Bulletin of IHP*, vol. 51, no. 2, pp. 349–405, 1980.

[23] H.-c. J. Huang, "Squliq atayal syllable onset: Simple or complex," *Streams converging into an ocean: Festschrift in honor of Professor Paul Jen-kuei Li on his 70th birthday*, pp. 489–505, 2006.

[24] C. DiCanio, H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. García, "Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.

[25] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in *Proceedings of the 2nd European Conference on Speech Communication and Technology*, 1991, pp. 693–696.

MORPHOLOGY AND PHONETIC REALIZATION: LEXICAL VS MORPHOLOGICAL PRENASALIZATION IN SÀ'ÁN SÀVÍ ÑÀ ÑUÙ XNÚVÍKÓ (MIXTEC)

Guillem Belmar[1], Jeremías Salazar[1,2]

[1]University of California, Santa Barbara, [2]Mixteco Indígena Community Organizing Project
gbelmarviernes@ucsb.edu

ABSTRACT

This paper investigated the realization of the prenasalized stop /ⁿt/ and the prenasalized affricate /ⁿt͡ʃ/ in Sà'án Sàvĭ ñà ñuù Xnúvíkó (Mixtepec Mixtec, Otomanguean). These phonemes appear in native lexical items, and for many of these they can be traced back to proto-Mixtec. More recently, however, due to processes of segmental erosion triggered by grammaticalization, morphological prenasalization has become widespread in the language. Data from 5 speakers elicited in the field is analyzed in this study to investigate whether the nature of the prenasalization (lexical or morphological) has an effect on the realization of these two segments. Results show longer duration of the nasal closure when prenasalization is the result of morphological processes, which we suggest reflects the marked status of these forms.

Keywords: prenasalization; grammaticalization; Mixtec; voicing; morphology

1. INTRODUCTION[1]

Sà'án Sàvĭ ñà ñuù Xnúvíkó (Mixtepec Mixtec, Otomanguean, ISO693-3: mix) is spoken by approximately 9,170 speakers ([3]) in the municipality of San Juan Mixtepec, in the district of Juxtlahuaca, Oaxaca, Mexico. However, the number of speakers in diaspora community, mainly in Northern Mexico and the USA, is unknown ([11], [15]).

Just like other Otomanguean languages, Sà'án Sàvĭ ñà ñuù Xnúvíkó presents a contrast between plain and prenasalized stops (/p/ vs. /ⁿp/, /t/ vs. /ⁿt/, /k/ vs. /ⁿk/, and /kʷ/ vs. /ⁿkʷ/) and affricates (/t͡s/ vs. /ⁿt͡s/, and /t͡ʃ/ vs. /ⁿt͡ʃ/). These phones are commonly described as prenasalized, often voiced, segments ([3], [7], [13], [14]) or post-oralized nasal stops ([4], [9]) as opposed to consonant clusters ([10]) or a syllabic nasal stop, although there has not been any systematic study to support this. Evidence that these phones constitute the onset of syllables and cannot be separated was found in a related language, Zenzontepec Chatino ([2]). This study used play language syllable transposition as evidence for syllable structure and found that prenasalized segments indeed behave as single units, rather than clusters. We refer to these segments as prenasalized by convention, but whether these segments are to be analyzed as prenasalized oral stops or post-oralized nasal stops is outside of the scope of the present paper.

The present study compares the realization of prenasalized segments in Sà'án Sàvĭ ñà ñuù Xnúvíkó when they are lexical as opposed to when the result of morphological processes. In some lexical items, the prenasalization can be traced back to Proto-Mixtec ([6]) and is therefore part of the segmental composition of a lexical item (i.e., present-day [ⁿdu³t͡ʃi¹⁴] 'bean' from proto-Mixtec **ⁿduti?*). However, processes of segmental erosion ([4]) triggered by grammaticalization have also given rise to a different category of prenasalization that is morphological in nature. This newer prenasalization is currently widespread in Sà'án Sàvĭ ñà ñuù Xnúvíkó.

There are three main environments for morphological prenasalization: the prospective, the perfective verbal aspects, and irrealis negation (see Table 1).

C	English	nC	English
kítsáá [ki⁴t͡sa:⁴⁴]	IPFV.start	*kú nkìtsáá* [ku⁴ᵑgi¹t͡sa:⁴⁴]	PROSP.start
cháa [t͡ʃa:⁴³]	IPFV.write	*nchàa* [ⁿd͡ʒa:¹³]	PFV.write
katsí [ka³t͡si⁴]	POT.eat	*nkătsí* [ᵑga¹³t͡si⁴]	NEG.POT.eat

Table 1: Table showing prenasalization in the prospective aspect ('start'), the perfective aspect ('write') and the irrealis negation ('eat') in Sà'án Sàvĭ ñà ñuù Xnúvíkó.

The negative form is very common in fast speech, but speakers tend to produce the unpacked *mà* + V in careful speech (that is, *mà kătsí* instead of *nkătsí*, see Table 1). In addition, since the prospective forms of verbs include a prefix that may be variably realized as [ku⁴], [u⁴], or [ũ⁴], thus presenting potential difficulties for acoustic measurements, the perfective form of verbs is taken to represent morphological prenasalization in this study.

The perfective forms in the language may be expressed in different ways, all of which can be seen as different stages of grammaticalization, in which a prefix *nì-* has eroded and affected the tonal melody of the word to different degrees. In fact, the prefix surfaces in the negative forms of these verbs (see Table 2).

IPFV	PFV	NEG.PFV	English
kíxi [ki^{4}ʃi^{1}]	***nì****kìxì* [ni^{1}ki^{1}ʃi^{1}]	*kuě* ***ní****kìxì* [k^{w}e^{14}ni^{4}ki^{1}ʃi^{1}]	Sleep
tsíka [t͡si^{4}ka^{3}]	***n****tsìka* [^{n}d͡zi^{1}ka^{3}]	*kuě* ***ní****tsìka* [k^{w}e^{14}ni^{4}t͡si^{1}ka^{3}]	Walk
né'ě [ne^{4}ʔe^{14}]	*nè'ě* [ne^{1}ʔe^{14}]	*kuě* ***ní****nè'ě* [k^{w}e^{14}ni^{4}ne^{1}ʔe^{14}]	Get

Table 2: Imperfective, perfective, and negative perfective forms of the verbs 'sleep', 'walk', and 'get' in Sà'án Sàvĭ ñà ñuù Xnúvíkó illustrating the segmental erosion of the prefix *nì-* giving rise to these three different verbal paradigms.

This study aims to determine whether lexical and morphological prenasalization are phonetically realized differently by speakers of Sà'án Sàvĭ ñà ñuù Xnúvíkó. We hypothesize that there may be a difference in: a) longer relative duration of the nasal closure, and b) longer relative duration of voicing of the oral closure.

2. METHODOLOGY

2.1. Participants

Six participants, 3 females and 3 males, between the ages of 20 to 60 took part in this elicitation task. All participants were self-identified native speakers of Sà'án Sàvĭ ñà ñuù Xnúvíkó. Due to the presence of background noise, data from one speaker was dropped from the analysis, resulting in a sample of 5 speakers: 3 females and 2 males.

2.2. Materials and experimental procedure

As Mixtec is predominantly an oral language and most speakers are not used to reading it, nor is there a well-established writing system, the task consisted of watching and listening to a video presenting 40 stimuli sentences accompanied by pictures illustrating the meaning of a target word (see Fig. 1) and repeating the target word embedded in a carrier sentence twice, one repetition after the other.

The second author, also a speaker of Sà'án Sàvĭ ñà ñuù Xnúvíkó, recorded all stimuli sentences in one recording session using a Tascam DR-40X audio recorder and a Shure WH20XLR Dynamic Headset microphone. The first author used Audacity to process the audios and PowerPoint to create the video stimuli.



Figure 1: Screenshot of the video stimuli for the target word *ntà'vĭ* [^{n}ta^{1}ʔ^{1}βi^{14}] 'broke'.

All target words begin with either the prenasalized stop [^{n}t] (20 sentences) or the prenasalized affricate [^{n}t͡ʃ] (20 sentences), and the type of prenasalization in each target word is either lexical (10 words per consonant) or morphological (10 words per consonant). The carrier sentence (*vàtsi tù'un kávi-rà ___ sàtă iin líbrù* [βa^{1}t͡si^{3}tũ^{1}ʔũ^{3}ka^{1}βi^{3}ra^{1} | __ | sa^{1}ta^{14}ʔĩ:^{3}li^{4}βru^{1}], which translates to *X appears in the words he is reading in this presentation*) was designed to avoid any nasal or nasalized segment occurring adjacent to the initial prenasalized segments of the target words, as well as to place all target words in the same focus structure. The stimuli sentences presented the target word in context to make sure that speakers knew the exact word they needed to repeat, avoiding confusion with minimal pairs. The sentences were also accompanied by pictures that illustrated the meaning of the target word. All stimuli sentences and the carrier sentence were designed by the authors themselves.

Participants were recorded individually in one session in their houses by the second author using a Tascam DR-40X audio recorder and a Shure WH20XLR Dynamic Headset microphone. Participants were first explained what the task consisted of and given an example stimulus to make sure they understood the task. After that, the same video was played for all participants, containing 20-second-long pauses in between stimulus sentences so that they could repeat the carrier sentences with the target words.

2.3. Measurements

The recordings were then analyzed using Praat ([1]), measuring the duration of both the nasal and the oral closure of the initial prenasalized segments of the target words (see Figures 2 and 3), as well as the duration of voicing in the oral closure. In addition, each target word was coded for speaker, order (first or second time uttering the carrier sentence), as well as the vowel following the prenasalized segment and the number of syllables of the word.

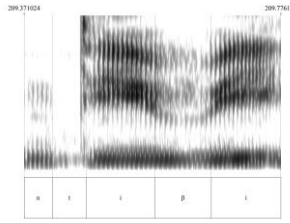


Figure 2: Segmented spectrogram for the word *ntivĭ* [ⁿdi³βi¹⁴] 'egg' produced by a male speaker.

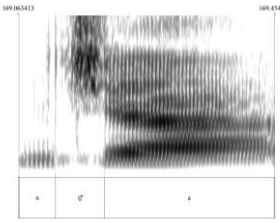


Figure 3: Segmented spectrogram for the word *nchǎá* [ⁿd͡ʒa⁴⁴] 'blue' produced by a male speaker.

The duration measurements were converted into percentages to facilitate comparison (i.e., how much of the total duration of the segment corresponds to either the nasal or the oral closure).

2.4. Statistical analysis

Using R Studio ([9]) we tested for significant differences in the relative duration of the nasal closure and the relative duration of voicing in the oral closure for lexical and morphological prenasalization. Since /ⁿt/ and /ⁿt͡ʃ/ behaved significantly differently, two separate analyses were conducted. Four linear models, two for /ⁿt/ and two for /ⁿt͡ʃ/, were constructed with the following independent variables: the nature of the prenasalization (lexical or morphological), the order of utterance (first or repetition), the vowel quality, the nasal quality of the vowel (oral or nasal), the number of syllables of the word (from one to four), and Speaker.

3. RESULTS

For /ⁿt/, we constructed a linear model to test if the nature of prenasalization (lexical or morphological) significantly predicted the relative duration of the nasal closure in the prenasalized segment. The fitted regression model was: Relative duration of the nasal closure for lexical prenasalization = 47.87 + 7.65** (for morphological prenasalization). The overall regression was statistically significant (R^2=.25, F(10, 189) = 7.746, p<0.001). It was found that the nature of the prenasalization significantly predicted the duration of the nasal closure of the prenasalized segment (β = 7.65, p<0.001). This trend is clear across speakers (see Fig. 4).

We also constructed a linear model to test if the nature of prenasalization (lexical or morphological) significantly predicted the relative duration of voicing in the oral closure of the prenasalized segment. The overall regression was statistically significant (R^2=.07, F(10, 189) = 2.498, p<0.05), with morphological prenasalization showing slightly longer duration of voicing in the oral closure (β = 8.78, p<0.05) (see Fig. 5).

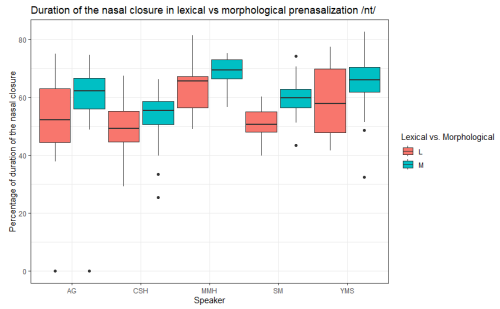


Figure 4: Relative duration of the nasal closure in /ⁿt/ by speaker and condition (lexical vs. morphological)

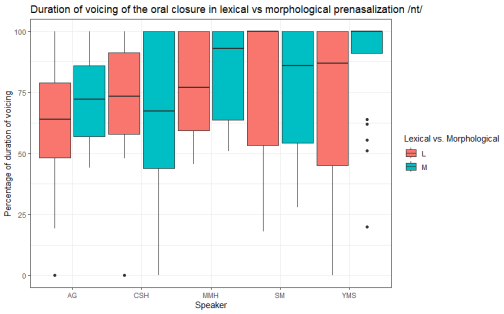


Figure 5: Relative duration of voicing in the oral closure in /ⁿt/ by speaker and condition (lexical vs. morphological)

For /ⁿt͡ʃ/, we constructed a linear model to see if the nature of prenasalization (lexical or morphological) significantly predicted the relative duration of the nasal closure in the prenasalized segment. The overall regression was statistically significant (R^2=.39, F(11, 188) = 12.52, p<0.001), but there was no significant difference in the relative duration of the nasal closure as a function of the nature of prenasalization. There were however significant differences across speakers (See Fig. 6).

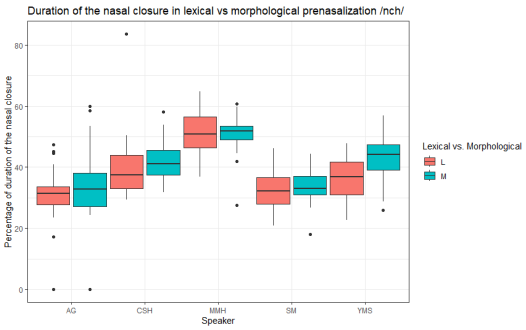


Figure 6: Relative duration of the nasal closure in /ⁿt͡ʃ/ by speaker and condition (lexical vs. morphological)

We also constructed a linear model of the relative duration of voicing in the oral closure as a function of nature of prenasalization (lexical or morphological) and speaker. This model was significant (R^2= .65, F(11, 188)=34.61, p<0.001). There was, however, no significant differences in the relative duration of

voicing in the oral closure as a function of nature of prenasalization. There were, however, significant differences across speakers (see Fig. 7), and a significant difference between the first and second utterance (repetition) of the words (see Fig. 8).

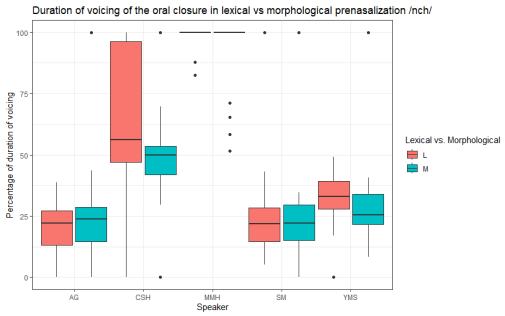


Figure 7: Relative duration of voicing in the oral closure in /ⁿt͡ʃ/ by speaker and condition (lexical vs. morphological)

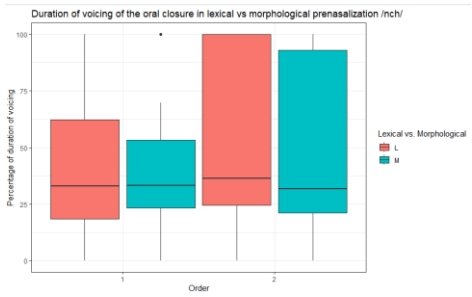


Figure 8: Relative duration of voicing in the oral closure in /ⁿt͡ʃ/ by order of utterance and condition (lexical vs. morphological)

4. SUMMARY AND DISCUSSION

The analysis of the prenasalized segments /ⁿt/ and /ⁿt͡ʃ/ in Sà'án Sàvĭ ñà ñuù Xnúvíkó revealed an interesting trend. While the number of speakers analyzed (5) and the number of tokens (N = 400) is too low to venture grand generalizations, the data show a clear trend for the nasal closure in /ⁿt/ to be significantly longer when the prenasalization is the result of morphological processes (with a mean duration of 61.2% of the closure for morphological prenasalization in comparison to a mean of 54.15% of the closure when the prenasalization is a feature of the lexical item itself). No difference was found in the duration of voicing in the oral segment.

This trend signals at the possibility that morphological status directly effects phonetic realization: although these are the same phones, undistinguished by speakers, their phonetic realization seems to be significantly different depending on their morphology (or lack thereof), with simplex lexical items displaying shorter duration of the nasal closure than those items in which prenasalization is morphological.

This may be due to morphological prenasalization in Sà'án Sàvĭ ñà ñuù Xnúvíkó historically deriving from a prefix *nì-* that can still be found in the language, and signals perhaps at an ongoing process of segmental erosion. It could be the case that the significantly longer realization of the nasal closure for morphological prenasalization is due to its importance in marking a relevant morphological distinction. That is, the forms that have undergone morphological processes all contrast directly with simplex non-prenasalized items. Prenasalization as a morphological process in Sà'án Sàvĭ ñà ñuù Xnúvíkó leads to the existence of several minimal and near-minimal pairs whose only differing segment is the prenasalized one. This may explain the need to emphasize the nasal closure in order to maximize difference with the unmarked, non-prenasalized form. However, prenasalization is not the only distinctive feature of these forms, as tone is also affected by the same morphological processes, and tone can even be the sole indicator of aspect change (see Table 2). A possible future direction to expand this analysis would be to include the prenasalized velar stop /ⁿk/ and its labialized counterpart /ⁿkʷ/, as well as other contexts in which prenasalization is morphological (such as the prospective aspect).

However, durational differences between lexical and morphological prenasalization were not observed for /ⁿt͡ʃ/. This is probably due to the already complex articulation of the affricate /t͡ʃ/ and the already crowded window of time for the gestures necessary to produce this segment. When prenasalized, therefore, there would be much less room for lengthening of the nasal closure.

Finally, we want to emphasize this study's combination of experimental and field methods. Experimental paradigms of phonetic research have traditionally placed an almost non-negotiable importance on tightly controlled settings and procedures, as well as heavy reliance on written representations of language. These present an unrealistic ideal when working with speakers of endangered minoritized languages residing in areas where access to higher education and laboratories is nearly non-existent. These were hurdles that we had to overcome during this study, leading to the mixed methodology presented in this paper: a fieldwork elicitation session using experimental stimuli. We hope that this study adds to the currently expanding body of literature working with languages underrepresented in scholarly research.

7. REFERENCES

[1] Boersma, P.; Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.10. Retrieved from http://www.praat.org

[2] Campbell, E.W. 2020. Probing phonological structure in play language: Speaking backwards in Zenzontepec Chatino. *Phonological Data and Analysis, 2*(1), 1-21.

[3] Cortés, F.; Mantenuto, I.; Steffman, J. 2023. San Sebastián del Monte Mixtec. *Journal of the International Phonetic Association.* doi:10.1017/S0025100322000226

[4] DiCanio, C.T.; Zhang, C.; Whalen, D.H.; Castillo García, R. 2019. Phonetic structure in Yoloxóchitl Mixtec consonants. *Journal of the International Phonetic Association, 50*(3), 333-365.

[5] Eberhard, D.M.; Simon, G.F.; Fennig, C.D. (eds.). 2022. *Ethnologue: Languages of the World.* Twenty-fifth edition. Dallas, Texas: SIL International.

[6] Heine, B.; Reh, M. 1984. *Grammaticalization and reanalysis in African languages.* Buske Helmet Verlag Gmbh.

[7] Iverson, G.K.; Salmons, J.C. 1996. Mixtec prenasalization as hypervoicing. *International Journal of American Linguistics, 62*(2), 165-175.

[8] Josserand, J.K. 1983. *Mixtec dialect history*. PhD Thesis. Tulane University.

[9] Marlett, S.A. 1992. Nasalization in Mixtec languages. *International Journal of American Linguistics, 58*(4), 425-435.

[10] Paster, M.; Beam de Azcona, R. 2004. A Phonological Sketch of the Yucunany Dialect of Mixtepec Mixtec. In: Harper, L.; Jany, C. (eds), *Proceedings of the 7th Annual Workshop on American Indigenous Languages,* 61-76. Santa Barbara: University of California.

[11] Reyes Basurto, G.; Hernández Martínez, C.; Campbell, E.W. 2021. What is community? Perspectives from the Mixtec diaspora in California. In Olko, J.; Sallabank, J. (eds), *Revitalizing endangered languages: A practical guide*, 100-102. Cambridge: Cambridge University Press.

[12] RStudio Team. 2020. *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA. http://www.rstudio.com

[13] Rueda Chavez, J.E. 2021. La caracterización fonológica del grupo mixteco: 70 años de descripciones segmentales. *Cuadernos de Lingüística de El Colegio de México*, 8. Ciudad de México.

[14] Salazar, J.; Alfaife, S., Belmar, G.; Campbell, E.W.; Mendoza, G.; Olguín Martínez, J.; Reyes Basurto, G.; Scanlon, C.; Troiani, G.; Vásquez-Aguilar, A. 2020. *Yucunani Mixtepec Mixtec: Collaborative Grammar Sketch.* Unpublished.

[15] Salazar, J.; Belmar, G.; Scanlon, C.; Troiani, G.; Campbell, E.W. 2021. Bridging diaspora: technology in the service of the revitalization of of Sà'án Sàvĭ ñà Yukúnanĭ. In Derhemi, E. (ed.), *Endangered Languages and Diaspora – XXVI Annual Conference Proceedings,* 178-185. Tirana, Albania: Foundation for Endangered Languages / QSPA.

[1] We want to thank Francisco Bautista Santiago, Alejandra García, Silverio García, Yuridia García Santiago, Miguel Martínez Hernández, and Claudia Salazar Hernández for sharing their language with us for this study. We also want to acknowledge everyone from the Phonetics Lab at UCSB for their feedback throughout the process: Jiyoung Jang, Argyro Katsika, Alonso Vásquez-Aguilar, James Yee, Chun-Jan Young, and specially Sherry Chien for her help with the statistical analysis. We also want to acknowledge everyone from the UCSB Field Methods Class 2019/20 for their contributions in the Documentation project for Yucunani Mixtepec Mixtec: Saleem Alfaife, Eric W. Campbell, Pingping Ge, Gabriel Mendoza, Jesús Olguín Martínez, Griselda Reyes Basurto, Catherine Scanlon, Giorgia Troiani, and Alonso Vásquez-Aguilar. The equipment used in this project was purchased with funds from the Jacobs Research Funds (Whatcom Museum Foundation, Bellingham, Washington USA). This project was supported partly by a UCSB Academic Senate Faculty Research Grant and is part of the collaboration between University of California, Santa Barbara (UCSB) and Mixteco Indígena Community Organizing Project (MICOP).

DEVELOPMENT OF SPEECH SYNTHESES FOR LOWER SORBIAN AND UPPER SORBIAN USING MARYTTS

Astrid Schmiedel, Ingmar Steiner

Serbski institut/Sorbian Institute, Bautzen, Germany
Fraunhofer IAIS, Sankt Augustin, Germany
astrid.schmiedel@serbski-institut.de, ingmar.steiner@iais.fraunhofer.de

ABSTRACT

Lower and Upper Sorbian are Slavic languages spoken in eastern Germany. As minority languages they are phonetically and phonologically less well documented than major languages and so far only benefit from limited digital resources. Thus, they are usually not considered for technologies such as text-to-speech (TTS) synthesis. This paper describes our approach, methods, and resources for creating two TTS synthesis systems for the Sorbian languages using the open-source MaryTTS platform. All data and resources for the unit-selection synthesis voices will be published under an open-source license and will be made freely available.

Keywords: Upper Sorbian, Lower Sorbian, text-to-speech synthesis, MaryTTS, unit selection speech synthesis

1. INTRODUCTION

The Sorbian people are one of four recognised autochthonous national minorities living in Germany. Their languages, Lower and Upper Sorbian, are protected in Germany under the European Charter for Regional or Minority Languages [1]. Not everyone who identifies as Sorbian is also proficient in the language; especially with regard to Lower Sorbian, there are only about 200 native speakers left (almost) all elderly. In addition, there are semi-speakers and, thanks to language revitalisation measures, also new speakers. Exact numbers of speakers are difficult to give; ten-year-old estimates speak of up to 5,000 speakers [2, p. 279] of Lower Sorbian and about 20,000 speakers [2, p. 291] for Upper Sorbian, both with a clear downward trend.

Still, as with many other minority and endangered languages, more and more Sorbian-language content and information in text form can be found on the internet. However, so far there are no technical aids such as auditory speech output or read aloud programmes, which would help achieve full accessibility for the Sorbian languages. The development of Sorbian text-to-speech (TTS) synthesis systems has both technical and linguistic prerequisites. The lack of a complete and scientifically supported standard pronunciation of Upper and Lower Sorbian is problematic. Though there are studies on Lower Sorbian and Upper Sorbian phonetics and phonology (e.g., [3, 4, 5, 6]), relevant questions, i.e., phoneme onset, number of phonemes and typical allophones, distribution, exact realisation individually and in phonetic compounds, relationship between spelling and pronunciation, need fundamental research.

The development of speech syntheses for Lower Sorbian and Upper Sorbian is a research project which will not be complete for several years, but the present aim is to fill as many of these gaps in the fundamental research as possible. Fundamental research on the phonetics of the Sorbian languages and work on the speech syntheses run in parallel, and the current focus of the fundamental research is on the connection between orthography and pronunciation in order to be able to work with preliminary orthoepic proposals.

The focus of this paper is the technical development using the open-source MaryTTS platform [7] and the voicebuilding process which it provides.

2. MATERIAL AND METHODS

Due to its modular architecture, the MaryTTS platform allows developers to access and modify the entire processing system from text input to speech output [8] and therefore enables the building of custom synthetic voices in new languages. It also supports unit-selection speech synthesis, which we choose for the benefit of its more natural synthetic sound. The process requires the creation of natural language processing (NLP) components for Lower and Upper Sorbian, and also requires appropriate speech data, i.e., natural language recordings of a speaker in the target languages in order to combine them for building the synthesised voices.

2.1. NLP components

Pronunciation prediction requires at least information on the pronunciation of the individual sounds as well as a word list as a lexical resource.

2.1.1. Allophone set

In lack of a pronunciation dictionary or an orthoepy of the Sorbian languages, a list of all the allophones of Lower Sorbian and Upper Sorbian was compiled on the basis of previous research [3, 4, 5, 6] and assembled in the required form. This allophone set contains IPA transcription as well as a customised SAMPA transcription plus the according standard phonological features. A customised SAMPA transcription was necessary because some standard SAMPA transcriptions – for example [ə̹] and [e̞] – include a backslash or an underscore, which are reserved characters in MaryTTS. The lists contain about 90 allophones for Lower Sorbian (36 vowels, 57 consonants/compounds) and 80 allophones for Upper Sorbian (33 vowels, 47 consonants/compounds). Since there are no standard variants yet, all currently accepted pronunciations have been collected. We decided to include also non-native sounds, mostly German, e.g., long vowels. The settlement area of the Sorbs within Germany has resulted in very close language contact with German, so the additional sounds are required to correctly pronounce foreign and loan words.

2.1.2. Lexical resource

The lexical resource consists of a word list containing relevant words in both orthographic representation and phonetic transcription. To select the words, morphological generators were used, which were developed at or with participation of the Sorbian Institute as the basis for programmes for automatic spell-checking for Lower and Upper Sorbian [9, 10, 11]. Based on a registered basic word form, all correct grammatical forms can be built applying the generators. The words were therefore not selected according to frequency, but rather – due to the availability of this resource – algorithmically. The algorithm was developed to output a list of words, taking into account as many language specific sound combinations as possible (based on preliminary analysis of the accompanying research on the connection between orthography and pronunciation).

For transcription we used the Grapheme-to-Phoneme (G2P) conversion tool [12, 13] available at BAS Web Services [14]. We make use of the option to create and upload customised mappings from orthographic to phonological form to use it for Lower and Upper Sorbian as "user-defined" languages. Although this sped up the transcription process a lot, it required and still requires constant checking and manual adjustment. Currently the word lists include approximately 12,500 items for Lower Sorbian and approximately 10,400 items for Upper Sorbian; we aim at about 40,000 entries for each. Letter-to-sound rule training, required for MaryTTS, takes place automatically based on the word lists.

2.1.3. Text normalisation

As Slavic languages, Lower and Upper Sorbian are highly inflected, so this is ongoing work.

In both NLP components, we have implemented a preprocessing module for text normalisation, which handles the conversion of symbols and numbers – which would be unpronounceable by the TTS engine in their raw form – into written words. Each token in the input text is checked against a symbol mapping table, and if applicable, replaced by a normalised string. Furthermore, following the approach of [15], we detect numeric tokens, and replace them by their spell-out pronunciation, using the Unicode Consortium's ICU4J library with custom rule-sets. In this way, an input string such as "100 €" will be normalized to the words, *sto eurow*.

Further improvements, including acronym handling, are planned.

2.2. Speech Data

In addition to the NLP components, natural language material is required to generate the synthesised voices using the unit-selection method.

2.2.1. Prompt material

Fortunately, we have access to two reference corpora representing the modern Sorbian written language. They are both based on Lower and Upper Sorbian publications from the period 1990 to 2010. The Lower Sorbian Reference Corpus consists of 68,058 randomly selected but evenly distributed sentences and sentence-like units of text, the Upper Sorbian one of 70,217. The selection of sentences should be as phonetically balanced as possible. We tried to achieve this by defining suitable letter combinations in connection with markings of word boundaries as well as sentence beginnings and endings, after the sentences of the corpora had been standardised.

The Upper Sorbian version of *The North Wind and the Sun* is taken from Howson [5] whereas the

material	duration
The North Wind and the Sun	52 s
ca. 360 sentences and sentence-like units of text	90 min

(a) Lower Sorbian

material	duration
The North Wind and the Sun	49 s
ca. 380 sentences and sentence-like units of text	70 min

(b) Upper Sorbian

Table 1: Prompt material and (approx.) durations.

Lower Sorbian version was translated by an expert. Details are given in Table 1.

2.2.2. Recordings

The main speaker criteria are native speakers with very good pronunciation, resilience and availability.

In the case of Upper Sorbian there was a wide range of potential native speakers. Test recordings were made with five eligible candidates, whose recordings were then evaluated by Sorbian linguists, including native speakers, for pronunciation clarity and possible problematic anomalies. The chosen male speaker was recorded in several sessions in 2022 in a studio environment in Bautzen, Germany. We used a Neumann TLM 103 large-diaphragm microphone mounted on a microphone arm approximately 30 cm from the speaker's mouth. The audio was sampled at 44.1 kHz (16 bit) using a Steinberg MR816 CSX audio interface along with Steinbergs Nuendo recording software. The material was presented in PDF format on a monitor and was read from screen by the speaker. At the moment, the material on Upper Sorbian comprises around 70 min of speech (see Table 1b for details).

For Lower Sorbian a male speaker with very good Lower Sorbian language skills and pronunciation was chosen. He did not learn Lower Sorbian from his parents, but through contact with his extended family, who still use it frequently. As presented in the introduction, we did have to take into account the difficult situation of Lower Sorbian. While younger speakers do learn Lower Sorbian at school or in language courses, the remaining true native speakers are not suitable for the recordings we are aiming at because of their old age and the problems resulting from it. So in the broader sense considering the context of an endangered language, the chosen speaker is to be understood as a native speaker. In several recording sessions in 2021 and 2022 the speaker was recorded in a studio environment in Cottbus, Germany. We used a AKG P420 large-diaphragm microphone mounted approximately 30 cm from the speaker's mouth. The audio was sampled at 44.1 kHz (16 bit) using a Motu 16A audio interface along with Apple Logic Studio recording software. The material was read from a printout by the speaker. At the moment, the material on Lower Sorbian comprises around 90 min of speech (see Table 1a for details).

2.3. Processing

The Lower Sorbian recordings were done in several takes and then each prompt was stored as a separate WAV file, whereas the Upper Sorbian material was stored individually directly after each prompt. In combination with the corresponding text files, the prompts were semi-automatically labelled using the MaryTTS forced-alignment plugin, a custom wrapper around the Montreal Forced Aligner [16]. The phonetic annotations were manually checked and corrected where necessary.

2.4. Voicebuilding

Another plugin handles voicebuilding, using the NLP components to assemble the processed Lower Sorbian and Upper Sorbian voice data into unit-selection voice components suitable for installation into MaryTTS. Summarizing from [8], this process comprises *(a)* extracting acoustic features from the audio, *(b)* extracting linguistic features from the text, *(c)* aligning features based on the phonetic labels, *(d)* building prosody models, and *(e)* packaging the data.

3. SUMMARY

The development of speech syntheses for Lower and Upper Sorbian is a research project which will not be complete for several years. The goal of creating a user-friendly service that enables users to have Sorbian-language texts on websites read aloud in a comprehensible way and – as far as technically feasible – with largely "good" pronunciation, requires fundamental research work on the phonetics and phonology of Lower Sorbian and Upper Sorbian. The project allows the linking of the service aspect with the closing of research desiderata in this area. It is ongoing work with constant improvements. The current state already allows an understandable synthesis of simple(st) texts in both Sorbian languages, live demo available at

https://tts-juro-matej.serbski-institut.de/. The quality of pronunciation should be improved with further data (lexical resource, audio data and especially further NLP components ideally considering different case endings). The lexical resources as well as the NLP components for Lower and Upper Sorbian are released under the Lesser GPL 3.0 license [17, 18, 19, 20]. First versions of both Sorbian synthesis voices are released free for non-commercial use [21, 22]. The Upper and Lower Sorbian Speech Databases are also publicly available under the same terms [23, 24].

4. ACKNOWLEDGEMENT

This project is supported with tax funds on the basis of the budget passed by the Saxon State Parliament (directive inclusion, no. 100664411).

The authors gratefully acknowledge one anonymous reviewer for detailed, helpful feedback.

5. REFERENCES

[1] Treaty Office of the Council of Europe, "ETS No. 148: European Charter for Regional or Minority Languages," 1998. URL: https://www.coe.int/en/web/european-charter-regional-or-minority-languages/text-of-the-charter

[2] F. Šěn, D. Scholze, and S. Hose, Eds., *Sorbisches Kulturlexikon*. Bautzen: Domowina-Verl., 2014.

[3] L. Jocz, *Wokalowy system hornjoserbskeje rěče přitomnosće [The vowel system of contemporary Upper Sorbian]*. Szczecin: volumina.pl Daniel Krzanowski, 2011.

[4] ——, *Studien zum obersorbischen Konsonantismus*. Szczecin: volumina.pl Daniel Krzanowski, 2015.

[5] P. Howson, "Upper Sorbian," *Journal of the International Phonetic Association*, vol. 47, no. 3, pp. 359–367, 2017, doi:10.1017/S0025100316000414.

[6] F. Kaulfürst, "Pśinosk k dolnoserbskej ortoepiji na zakładźe projekta awdijowych datajow za nimsko-dolnoserbski internetowy słownik [The creation and integration of sound data in the German - Lower Sorbian internet dictionary as a starting point for a discussion of Lower Sorbian orthoepy]," *Lětopis*, vol. 66, no. 1, pp. 3–41, 2019.

[7] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003, doi:10.1023/A:1025708916924.

[8] I. Steiner and S. Le Maguer, "Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform," in *Proc. 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, 2018, pp. 3171–3175. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1045.pdf

[9] "Lower Sorbian spell-checker," https://www.niedersorbisch.de/ortografija/kontrola.

[10] "Upper Sorbian spell-checker," https://soblex.de/download/download.html.

[11] "Soblex," https://www.soblex.de/?cmd=help.

[12] U. Reichel, "PermA and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, Portland, Oregon, 2012, pp. 1874–1877, doi:10.21437/Interspeech.2012-509.

[13] ——, "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," in *Proc. 25th Conference on Electronic Speech Signal Processing (ESSV)*, Dresden, Germany, 2014, pp. 42–49. URL: https://www.essv.de/pdf/2014_42_49.pdf

[14] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017, doi:10.1016/j.csl.2017.01.005.

[15] I. Steiner, S. Le Maguer, J. Manzoni, P. Gilles, and J. Trouvain, "Developing new language tools for MaryTTS: the case of Luxembourgish," in *Proc. 28th Conference on Electronic Speech Signal Processing (ESSV)*, Saarbrücken, Germany, 2017, pp. 186–192. URL: https://www.essv.de/pdf/2017_186_192.pdf

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502, doi:10.21437/Interspeech.2017-1386.

[17] "Upper Sorbian lexicon for MaryTTS | Hornjoserbski leksikon za MaryTTS," 2023. URL: https://github.com/marytts/marytts-lexicon-hsb

[18] "Lower Sorbian lexicon for MaryTTS | Dolnoserbski leksikon za MaryTTS," 2023. URL: https://github.com/marytts/marytts-lexicon-dsb

[19] "Upper Sorbian language component for MaryTTS | Hornjoserbska rěčna komponenta za MaryTTS," 2023. URL: https://github.com/marytts/marytts-lang-hsb

[20] "Lower Sorbian language component for MaryTTS | Dolnoserbska rěcna komponenta za MaryTTS," 2023. URL: https://github.com/marytts/marytts-lang-dsb

[21] "A male Upper Sorbian unit selection voice for MaryTTS," 2023. URL: https://github.com/marytts/voice-serbski-institut-hsb-matej

[22] "A male Lower Sorbian unit selection voice for MaryTTS," 2023. URL: https://github.com/marytts/voice-serbski-institut-dsb-juro

[23] "Upper Sorbian voice data for MaryTTS," 2023. URL: https://github.com/marytts/serbski-institut-hsb-data

[24] "Lower Sorbian voice data for MaryTTS," 2023. URL: https://github.com/marytts/serbski-institut-dsb-data

Resources and tools for pre-processing speech data in a lesser-known variety of English

Alexandra Vella[1], Sarah Grech[1], Ian Padovani[2], Maria-Christina Micallef[3]

[1]University of Malta, Malta; [2]Utrecht University, The Netherlands; [3]Uppsala University, Sweden
{alexandra.vella, sarah.grech}@um.edu.mt, i.padovani@students.uu.nl, maria-christina.micallef.8671@student.uu.se

ABSTRACT

Research on lesser-known language varieties can be hindered from the outset by the need for both data and tools for automating the required pre-processing work. For speech, whilst more ecologically valid data in the form of video and audio are sometimes available, these need to be accompanied by a machine-readable text, ideally segmented and labelled, both allowing for searchability. A significant initial commitment is needed even before the relevant phonetic and phonological research can begin. This paper demonstrates and evaluates the efficacy of already existing tools (YouTube captioning and WebMAUS forced alignment) in automating the pre-processing work required using Maltese English (MaltE), whilst also showcasing a sample analysis of the pronunciation of post-vocalic 'r' in the variety. As a low resource variety of English, MaltE presents a test case for showing how existing resources and tools can be utilised to work with language varieties which are digitally less well-supported.

Keywords: resources & tools, low resource, phonetics & phonology, Maltese English, variety of English

1. INTRODUCTION

A fair amount of manual effort is needed both to collect spoken data and to make such data ready for analysis, phonetic and/or phonological as well as otherwise [23]. Whilst this is the case for all languages, resources such as corpora and an increasingly wide range of automatic speech processing tools are available for so-called high resource languages such as English, Chinese and Spanish; Japanese and a few other European languages such as French can also boast an increasing wealth of resources and tools [12]. Low resource languages (see [6] for definition) trail behind, although the situation continues to improve. [19] report on their success in using WebMAUS to carry out forced alignment of data from a number of language documentation corpora whilst [14] survey research which shows that careful matching of a high to a low resource language can result in direct transfer of good performance to a new language using an existing model without any additional training. Similarly, [13] show that, whilst manual checking is still likely to be needed to correct errors, forced alignment software developed for American English performed well for six varieties of British English.

In this paper, we report on our application of two tools, YouTube's automatic captioning and the CLARIN-D web application WebMAUS, for pre-processing available data from Maltese English.

2. BACKGROUND

2.1. The Maltese English variety

Maltese English, MaltE, the variety of English of bilingual speakers of Maltese and English [22], can be said to be a lesser-known variety of English [17]. Whilst characterised by a great deal of both intra- and inter-speaker variability [4, 24], it is also readily identified by its speakers and others familiar with the variety [7, 18].

The accent of MaltE is particularly distinct in spite of the variability. Amongst the phonetic/phonological characteristics which have been noted [8, 24] are: (i) substitution of θ and ð; (ii) pronunciation of a velar plosive following ŋ; (iii) variability in the degree and nature of post-vocalic 'r' pronunciation; (iv) no clear/dark 'l' distinction; (v) aspiration and audible release of plosives; (vi) intervocalic voicing and final devoicing; (vii) vowel quality and durational differences (especially in the pronunciation of SSBE æ and ɜ:, as well as ə); (viii) lack of vowel weakening and of syllabic consonants; (ix) differences in lexical, compound and phrase stress; (x) distinct rhythm and intonation patterns.

In spite of its distinctness, empirical evidence for many of the phonetic/phonological characteristics of MaltE is scarce, partly due to the lack of pre-processed language corpora which, as [19] point out, provide an extremely good starting point for much general and applied linguistic research, sociophonetics included, with usefulness increasing the more fine-grained the temporal alignment of the transcription is. Moreover, although MaltE is spoken in daily life by most Maltese speakers, increasingly so as Malta becomes more "cosmopolitan" [24], collecting ecologically valid data from this variety is

never straightforward. Looking towards available data which can be pre-processed easily to give researchers access to data accompanied by a text, as well as by word and phone level time-aligned segmentations, would definitely be a step in the right direction. However, anecdotes of difficulties that technologies such as speech recognition systems have with different accents are frequent (see [15], but also [5]). It is therefore not clear whether and to what extent available tools can deal with data from MaltE, although reports of the sort in [13] are encouraging.

2.2. YouTube automatic captioning

Little information on the backend to YouTube's automatic captioning tool is available on the YouTube Help page [25]. It includes the caveat that "automatic captions may misrepresent the spoken content due to mispronunciations, accents, dialects or background noise". One might expect the variability in MaltE to exacerbate such difficulties resulting in a decrease in accuracy. On the whole, however, whilst the presence of all the usual elements of spontaneous speech including various discourse markers, normal disfluencies, and, in dialogue data, overlaps, make the captioning process less straightforward, the interface is generally user-friendly. Unlike other tools such as Google's Speech-to-Text API, YouTube's automatic captioning tool is accessible to people without prior programming knowledge, thus making it available to a wider pool of researchers (as well as other users).

2.3. WebMAUS

The ***M**unich **AU**tomatic **S**egmentation Service* **MAUS**, developed by the Bavarian Archive of Speech Signals and the corresponding CLARIN-D WebMAUS [9, 10, 16, 20] service uses acoustic models combined with pronunciation rules and, in some cases, a language model, as the basis for providing word and phone time-aligned segmentations given a) a speech signal; and b) an accompanying orthographic transcript [21]. Once again, it differs from other equivalent tools such as the Montreal Forced Aligner [26] in not requiring programming knowledge.

The WebMAUS service is available for more than 25 languages, including a number of dialects. It is currently available for five varieties of English: Australian (AU), American (US), British (GB), Scottish (SC) and New Zealand (NZ). The documentation on the different parameter sets suggests differences in the acoustic and pronunciation modelling and in the nature and amount of manually segmented and labelled data used for training. The results of using untrained forced alignment are however promising, see [19] and also [5], and, although [14] report on the importance of the need to select a high resource language or language variety to apply to a low resource one with care, they also note good performance if the match is well made.

At the start of this work, we tested WebMAUS's English (GB), (SC) and (US). Though Maltese speakers of English have come to be increasingly exposed to American English, British English continues to be favoured, particularly in education [24]. Australian and New Zealand English, particularly their vowel systems, present differently enough not to be in contention. After testing, it was found that the rhoticity of Scottish English served to minimise segmentation errors, and, to some extent also labelling errors. WebMAUS forced alignment was therefore carried out using English (SC).

3. METHODOLOGY

3.1. The data

For the purposes of this paper, we focus on a small amount of data taken from CoSME, a Corpus of Spoken Maltese English which is being collected from a variety of available sources, e.g. online platform or YouTube material, recordings of radio or television programmes, public, including University-based events, etc. Specifically, the data we report on here are: (1) a graduand monologue; and (2) a TimesTalk (dialogue) interview involving a MaltE interviewer and a MaltE singer, Ira Losco. We report on the application of tools available for use with more mainstream varieties of English to these MaltE data. The procedure used involved creating a text using the YouTube automatic captioning tool, followed by WebMAUS forced alignment segmentation and labelling of the relevant sound files.

The first requirement in the pre-processing of spoken data, whether in video or audio format, involves creating a text. This is because most automatic segmentation and labelling tools (specifically in this case WebMAUS) require a sound file together with an accompanying text. Our procedure, broadly based on [3], involves the following:

Step 1 – Use YouTube captioning to create a first instance of a text.
Step 2 – Use established conventions to "clean up" the text.
Step 3 – Make any boundary adjustments and add other details as necessary.
Step 4 – Run WebMAUS on the cleaned up version of the text created using YouTube to generate a Praat TextGrid containing the automatically generated segmentation/labelling.

In (the relatively rare) cases when a video or audio recording comes with a ready transcript, step 1 can be

skipped: steps 2 and 3 are nevertheless still necessary in order to ensure that the accompanying transcript is as "loyal" to what is in the actual audio as possible, thus minimising WebMAUS errors.

In the case of the dialogue data, we have not, so far, separated the speech signal into chunks on the basis of speaker (this is however something we intend to do as we continue with this work).

Overlap is marked in post-editing of the captions text and then removed from the original sound file and accompanying transcript using a simple Python script. In our pipeline, periods marked as overlap of any length are replaced by brief silence. This allows WebMAUS to annotate multiple voices in a single channel without the difficulties arising from overlap.

3.2. Using YouTube captioning

In order to use YouTube to automatically generate closed captions, audio must be uploaded to a YouTube account in video format. In the case of audio-only data, a still image was added to the sound file and the result saved in video (e.g. MP4) format.

YouTube automatically detects the language and generates captions accordingly. These can be manually edited in the available interface, and mistakes that occur, due to variety differences or for other reasons, can be corrected. Moreover, for the purpose of preparing the text for further processing, the more accurate the representation of an utterance is, the better. [19] report that the quality of the forced alignment is better when elements of spontaneous speech in the signal are included in the transcript. The closed captions were therefore manually edited by our annotators to include such elements. A Python script containing simple rules was also used to clean up other elements in the text such as the timestamps that appear in the subtitles, before passing it through WebMAUS. This was done to eliminate the risk that these would be interpreted as actual words in the audio signal.

3.3. Using WebMAUS

The edited YouTube close-captioned texts were saved as .txt and uploaded together with original audio files to WebMAUS. Forced alignment was carried out using English (SC) and a segmented and labelled Praat TextGrid [1] was generated. The ORT-MAU and MAU tiers were duplicated and edited manually by our annotators, see Fig. 1.

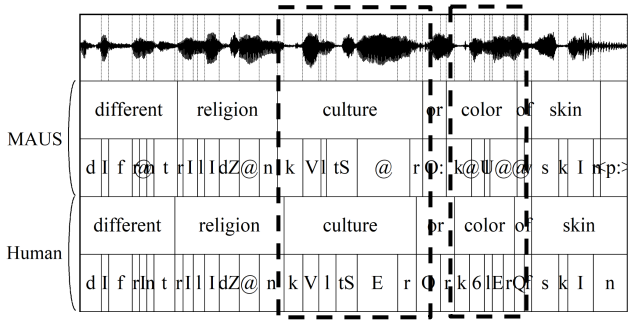


Figure 1: An example of an excerpt from a Praat TextGrid containing WebMAUS automatically generated word and phone level segmentation/labelling, together with the manual correction of the automatic output in the (bottom two) duplicated tiers.

4. EVALUATION OF TOOLS USED

4.1. Some general points

YouTube is quick at generating closed captions, typically taking from 30 minutes up to a few hours, depending on the file duration. The manual "cleaning up" required once the captions have been generated is also fairly quick, taking one of the annotators roughly an hour to cover 15 minutes of recording, although dialogue videos with high degrees of overlapping speech take more time. Still, creating a first instance text using this procedure takes far less time than full manual processing by human annotators.

[19] suggest that a "real-time factor up to 400" may be involved in manual segmentation and labelling of audio data at word and phone level. It took the annotator significantly less time to correct WebMAUS output.

4.2. YouTube captioning

Unlike the language documentation corpora data reported on in [19], recordings from available sources as outlined in 3.1 do not come with an accompanying text. In the case of MaltE, YouTube's closed captioning tool provides a surprisingly accurate starting solution.

A Word Error Rate (WER) analysis of our unedited (Orth) and edited (Clean) YouTube closed captioned texts is shown in Table 1.

Audio	Duration	Orth WER	Clean WER
Graduand	270	0.42%	1.55%
Interview	557	2.93%	6.99%

Table 1: WER scores for the data analysed. *Orth WER* compares to a plain orthographic transcription. *Clean WER* compares to a transcript including expanded acronyms, digits, added fillers etc.

Errors related to specific features of the variety do however occur. An example which arises from a feature of the pronunciation of MaltE, the substitution

of /ð/ with [d] (or [d̪], see feature i in 2.1 above), resulted in closed captioning of *though no* as *don't know*. Another example of a relatively common error type involves words such as *flags*, citation form /flægz/, being captioned as *flax* [flæks] a "mistake" likely due to final devoicing in Maltese English (see feature vii) or *flecks* [flɛks]: this is the result of substitution of /æ/ with [ɛ] (see feature viii).

As reported elsewhere, e.g. [14], proper nouns such as names and places often cause problems: the proper noun *Losco* was close-captioned as the (relatively) phonetically close approximation, *lost cause*. Errors related to detection of name and place entities in speech are not specific to the variety, see [14]. YouTube close-captioned *a new way* as *anyway* possibly for reasons related to speech rate. This is another issue not specific to MaltE, see [19].

4.3. WebMAUS

To evaluate the effectiveness of this step in our procedure, we ran WebMAUS on 270 seconds of data taken from the beginning, middle and end of a MaltE graduand monologue, (1) above, and an entire interview of 9 minutes and 17 seconds, involving two MaltE speakers, (2) above. Manual correction was carried out by our annotators. Boundaries at word beginnings and endings often needed adjusting (see Fig. 1) but segmentation was generally good. Discrepancies were more frequent at phone level, not just in the case of boundaries, but also with respect to phone labelling, a matter for further investigation.

Inter-annotator WebMAUS vs human annotator agreement was measured by means of a Cohen's Kappa score [11] based on comparison of the automatic and manual annotations. The results of the comparison for the graduand monologue and the interview data are shown in Table 2:

Sound File	Duration	Cohen's Kappa
Graduand	270	0.79
Interview	557	0.61

Table 2: Cohen's Kappa scores for the data analysed.

To utilise this metric, the audio and accompanying transcriptions were divided into non-overlapping frames 2 msecs in duration. Cohen's Kappa checks each frame to see if the segmentation and labelling of the phones in the two transcripts match or not, also taking into account the likelihood of them only matching by chance. As can be seen from Table 2, the agreement score for the two text-types is 0.79 and 0.61 respectively, indicating substantial agreement between WebMAUS and the human annotator. The score is higher for the graduand monologue data. This may be partly due to differences between scripted and unscripted speech. The former tends to display fewer normal disfluencies, and slower, more deliberate speech, making the latter more difficult to segment and label by comparison. Another possible reason for the difference is that different annotators were involved in annotating the data from the two sources, the latter receiving a more detail-oriented treatment.

4.4. WebMAUS vs human annotation of 'r'

Comparison of the WebMAUS vs human annotator data provides a wealth of information on the phonetic detail in the analyses. Phone substitutions made by the MaltE speakers clearly needed to be manually corrected. Characteristics of MaltE such as relatively heavy aspiration, including in word-final position, often resulted in the right-edge boundary being placed earlier by MAUS as compared to by the human annotator, the latter including the (often relatively long) period of aspiration as part of the relevant segment.

A characteristic we have started to examine in depth, but which we have minimal space to discuss here, is the MaltE realisation of 'r'. This, (see [2]) is shifting from non-rhotic to (more) rhotic. Examining the .csv file which served as the basis for the Cohen's Kappa analysis is interesting on several counts. It shows that the replacement rules employed in the English (SC) WebMAUS may not always generate truly Scottish English output. Fig. 1 (see the broken line box) shows an example in the labelling of *color* (YouTube spelling). WebMAUS does not include an 'r' following the vowel in spite of the rhoticity one would expect, also given the vowel following in *of skin*. Another example in Fig. 1 (see the heavy broken line box), *culture*, also includes a post-vocalic 'r'. Both cases involve an [ɛ] vowel similar in quality to Maltese /ɛ/ but shorter than what one would expect for SSBE, a sort of compensatory shortening. Further, in the interview data, 120 of the 215 instances of 'r' are post-vocalic. The human annotator labelled clear 'r' phones in all of these: WebMAUS gave only 15 'r' phones in this position. Two of the authors of this paper continue in the sociophonetic analysis of the MaltE pronunciation of 'r', including post-vocalically. Approximant, tap, trill and also affricated realisations have been noted, although the precise details of such realisations are still to be determined. Using data pre-processed as described above would provide a highly useful fast track entry point to the analysis.

5. CONCLUSION

In conclusion, we have shown that tools are available which allow researchers (and others), regardless of programming experience, to curate and pre-process

MaltE data, providing them with easy access to word and phone level time-aligned segmentations of ready data for analysis. As processing power, machine learning and natural language processing techniques continue to become more powerful, the time seems ripe to do all that is possible to increase the reach of already available tools to low resource languages and language varieties. User-friendly tools such as those we have described help bridge the gap allowing easy access to newcomers to these fields. We have shown that use of YouTube to create a first instance text to accompany a video/audio recording doubtlessly adds a welcome measure of automation at the early stages of transcription. Similarly, applying WebMAUS to MaltE gives good results, although manual correction remains necessary. Equally importantly, more in depth analysis of the accuracy of the output of both the closed captioning and forced alignment tools, and of the manual adjustments made by the human annotators, is likely to lead to insights of all sorts.

6. REFERENCES

[1] Boersma, P., Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 Dec 2022, http://www.praat.org/

[2] Bonnici, L. M. 2010. *Variation in Maltese English: The interplay of the local and the global in an emerging postcolonial variety*. PhD thesis, University of California.

[3] Cangemi, F., Fründt, J., Hanekamp, H., Grice, M. 2019. A semi-automatic workflow for orthographic transcription and syllabic segmentation. In: Piccardi, D., Ardolino, F., Calamai, S. (eds), *Gli Archivi Sonori al Crocevia tra Scienze Fonetiche, Informatica Umanistica e Patrimonio Digitale.* [*Audio Archives at the Digital Crossroads of Speech Sciences, Digital Humanities and Digital Heritage*.], Officinaventuno, Milano, 419-425.

[4] Caruana, S., Mori, L. 2021. Rethinking Maltese English as a continuum of sociolinguistic continua through evaluations of written and oral prompts. *English World-Wide. A Journal of Varieties of English* 42(3), 245-272.

[5] Choe, J., Chen, Y., Chan, M.P.Y., Li, A., Gao, X., Holliday, N. 2022. Language-specific effects on automatic speech recognition errors for World Englishes. *Proc. 29th International Conference on Computational Linguistics*, 7177-7186.

[6] Cieri, C., Maxwell, M., Strassel, S., Tracey, J. 2016. Selection criteria for low resource languages. *Proc. 10th International Conference on Language Resources and Evaluation (LREC '16)*, 4543-4549.

[7] Grech, S. 2015. *Variation in English: Perception and patterns in the identification of Maltese English*. PhD thesis, University of Malta.

[8] Grech, S., Vella, A. 2019. Rhythm in Maltese English. In: Paggio, P., Gatt, A. (eds), *The Languages of Malta*. Berlin: Language Science Press, 203-223.

[9] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech and Language* 45, 326-347.

[10] Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal processing via web services: The use case WebMAUS. *Proc. Digital Humanities*, Hamburg, 30-34.

[11] Kolesnyk, A. S., Khairova, N. F. 2022. Justification for the use of Cohen's Kappa statistic in experimental studies of NLP and text mining. *Cybernetics and Systems Analysis* 59(2), 280-288.

[12] Laumann, F. 2022. Low-resource language: what does it mean? https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

[13] MacKenzie, L., Turton, D. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6 s1.

[14] Magueresse, A., Carles, V., & Heetderks, E. 2020. Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264

[15] Reid, M., MacGregor, G. 2020. There can be no true Scottish speech recognition systém. https://jabde.com/2020/11/14/no-true-scottish-sls/

[16] Schiel, F. 1999. Automatic phonetic transcription of non-prompted speech. *Proc. 14th ICPhS* , San Francisco, 607-610.

[17] Schreier, D., Trudgill, P., Schneider, E. W., Williams, J. P. 2010. *The Lesser-Known Varieties of English: An introduction*. Cambridge: Cambridge University Press.

[18] Stilon, E. 2018. Perceptions of Maltese English: An experimental study. MA dissertation, University of Malta.

[19] Strunk, J., Schiel, F., Seifart, F. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. *Proc. 9th International Conference on Language Resources and Evaluation (LREC '14)*, 3940-3947.

[20] The **M**unich **AU**tomatic **S**egmentation System **MAUS**.https://www.bas.uni-muenchen.de/Bas/BasMAUS.html

[21] Tour de CLARIN: WebLicht and WebMAUS. https://www.clarin.eu/blog/tour-de-clarin-weblicht-and-webmaus

[22] Vella, A. 2012. Languages and language varieties in Malta. *International Journal of Bilingual Education and Bilingualism* 16, 532-552.

[23] Vella A., Grech, S. 2021. What can a corpus tell us about phonetic and phonological variation? In: O'Keeffe, A., McCarthy, M. J. (eds), *The Routledge Handbook of Corpus Linguistics*. Routledge, 281-295.

[24] Vella, A., Grech, S. Submitted. English in Malta.

[25] YouTube Help. https://support.google.com/youtube/answer/6373554?hl=en

[26] McAuliffe, Michael, Michaela Socolof, Elias Stengel-Eskin, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner [Computer program]. Version 1.0.

ACOUSTIC CORRELATES OF ALLOPHONIC PLOSIVE VOICING IN ROMEYKA GREEK

Merve Yazar, Metin Bağrıaçık, Stefano Canalis, Mehmet Can Dadan, Metehan Eryılmaz

Boğaziçi University, Turkey
merve.yazar@boun.edu.tr, metin.bagriacik@boun.edu.tr, stefano.canalis@boun.edu.tr, mehmet.dadan@boun.edu.tr, metehan.eryilmaz@boun.edu.tr

ABSTRACT

Romeyka is an endangered Greek variety spoken in north-eastern Turkey. Its phonological system differs in many ways from the other Greek varieties, and, to our knowledge, none of its properties have ever been investigated instrumentally ([1] documents Pontic Greek, which is spoken today in Greece by the descendants of refugees from Turkey who were relocated in the 1920s; however, Pontic differs substantially from Romeyka, which is still spoken in Turkey).

The aim of this paper is to present a preliminary investigation of the allophonic voicing of Romeyka plosives. At least in the native Greek lexicon, Romeyka does not seem to use voicing contrastively in plosives. We recorded word-initial and inter-sonorant plosives produced by three Romeyka speakers. The latter often show shorter/negative VOT, a substantially larger percentage of voicing into closure, and lower f_0 at the onset of the following vowel.

Keywords: acoustics of voicing, VOT, f_0, plosives, Romeyka

1. INTRODUCTION

Romeyka is an endangered language spoken in the Black Sea region of Turkey, mainly in the Trabzon province. It belongs to the Hellenic branch of the Indo-European language family. The number of speakers is supposed to be around five thousand according to the Turkish general census [2]; however, providing a precise figure may be difficult, since Romeyka speakers' attitude towards their language is influenced by sociological and political factors [3].

Only brief sketches of Romeyka phonetics and phonology exist. To the best of our knowledge, no aspect of them has ever been investigated instrumentally/quantitatively. [1] is a recent and detailed description of the phonetic structure of the Pontic Greek variety originating from Trabzon and nowadays spoken in Etoloakarnania, Greece by descendants of refugees from Turkey who were relocated there in the 1920s; however, this variety and Romeyka differ substantially, in all likeness because of the different language contact conditions (with Standard Greek for Pontic Greek as spoken in Etoloakarnania, and with Turkish for Romeyka).

1.1. Plosive voicing in Romeyka

In this paper we focus on some acoustic properties of voicing in Romeyka plosives. Previous descriptions say that voicing appears not to be contrastive in plosives, at least in the native Greek lexicon (loanwords from Turkish may preserve plosive voicing). This does not mean that plosives are uniformly voiceless or voiced. They are uniformly voiced after nasal consonants (e.g. [gonˈdo] *gondo* 'short'), and voiceless after voiceless consonants (e.g. [ndefˈtete] *ndeftete* 'how are you'). However, in other environments they show significant variability. Allophonic voicing of plosives has been impressionalistically reported in some descriptions of the Romeyka consonant system (cf. for example [4] p. 899: "The voice distinction is weakened for plosives, with underlying voiceless stops often realized as voiced"), but no previous sketch of Romeyka goes beyond this.

2. METHODOLOGY

To explore aspects of plosive voicing in Romeyka we recorded three speakers, with the goal of investigating three parameters related to plosive voicing/voicelessness, i.e., their VOT, proportion of the closure showing glottal pulses, and f_0 of the following vowel.

2.1. Subjects and material

Two female (aged 76 and 78) and one male (aged 78) Romeyka native speakers participated in this study. They are Turkish-Romeyka bilinguals who migrated to Istanbul from Trabzon thirty years ago. We observed them to dominantly use Romeyka at home and in their inner circle.

Given the preliminary nature of our study, we focussed on only two environments, the utterance-initial and the intervocalic one. Since most Romeyka words end in a vowel or a sonorant consonant, voicing in a word-initial but utterance-medial plosive would be comparable to that of a word-internal

intervocalic plosive; as said above, voicing before nasals appears to be quite systematic, as is voicelessness after voiceless consonants; since very few Romeyka words end in an obstruent other than /s/, we did not consider word-final plosives.

We created a list of Romeyka words with either a word-initial plosive or an intervocalic plosive at the onset of the second syllable (e.g., *târağma* 'mixing' and *atârağo* 'unmixable'). We strived to keep the rest of the environment (stress position, quality of the adjacent vowel(s)) as similar as possible. However, sometimes the ideal candidate was not available in the Romeyka lexicon and we decided not to use pseudowords, forcing us to occasionally relax these criteria. Romeyka has three contrastive places of articulation for plosives, i.e. bilabial, dental and velar. We tried to balance place evenly in our word list, which included a total of 40 items.

The speakers were recorded in a quiet room in Istanbul, where they produced the tokens in isolation. We asked them to repeat each word three times. A cardioid condenser microphone connected to a solid-state recorder was used to record their production at a sampling rate of 44.1 kHz onto a flash disc. The overall number of tokens we recorded was 326 (40 words by 3 repetitions by 3 speakers, minus some tokens that had to be discarded).

The digitized recordings were then imported into *Praat* [5] for analysis. We manually segmented and labelled the audio files. We annotated them for onset of plosive closure, offset of plosive closure and onset of voicing.

3. RESULTS

3.1. VOT

Voice Onset Time is one of the most widely adopted criteria to measure acoustic differences among different voicing categories within plosives [6, 7].

In Romeyka, utterance-initial plosives have on average a short positive VOT (Table 1).

Place	Mean VOT (ms)	SD
Bilabial	7	17
Dental	14	9
Velar	25	13

Table 1: Mean VOT of utterance-initial plosives

On the contrary, intervocalic plosives display negative VOT (Table 2).

Place	Mean VOT (ms)	SD
Bilabial	-10	15
Dental	-9	13
Velar	-12	24

Table 2: Mean VOT of intervocalic plosives

In fact, the moderately negative mean values of intervocalic /p t k/ result from combining what appear to be three distinct ways of realizing them. As Figure 1 shows, the statistical distribution of VOT values is anything but normal. Most consonants have a moderately negative VOT, but a fair proportion has moderately positive VOT, while a handful of plosives have substantially negative VOT.

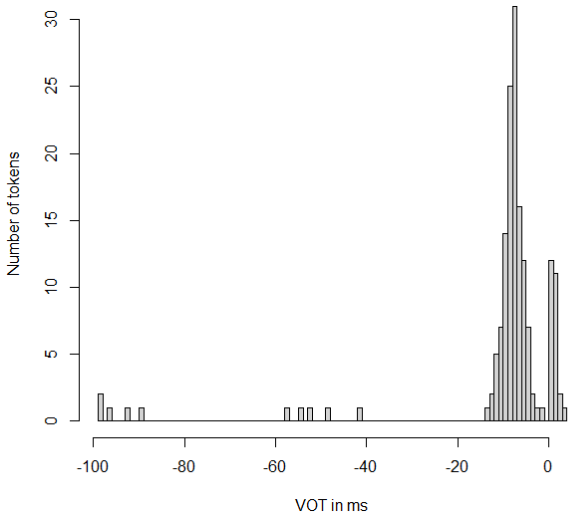


Figure 1: VOT intervocalically.

The statistical distribution of VOT in utterance-initial plosives is closer to a normal one, but sporadically VOT is amply negative (Figure 2).

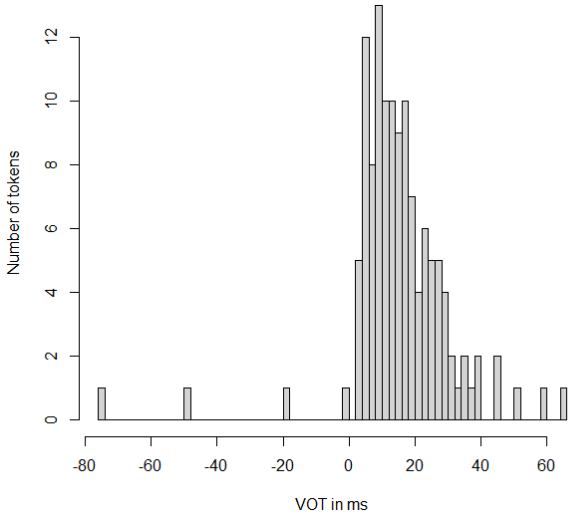


Figure 2: VOT utterance-initially.

3.2. Fraction of voicing into closure

The second parameter we used is the "fraction of locally unvoiced frames" measure in Praat's Voice Report. It is an automated measure of the percentage of the voiced part of a segment. The reliability of this measure has been questioned [8] in the past, mainly due to the widely variable values it gives when Praat's viewing window is scrolled and zoomed.

However, [9] has shown that, with the appropriate parameter setting, the results of Praat's Voice Report become comparable to those obtained with manual segmentation of voicing. Therefore, we adopted [9]'s suggestion to reduce the time step of cross-correlated pitch to 0.001 seconds and use gender-specific pitch ranges (70-250 Hz for male speakers and 100-300 Hz for female speakers). We extracted the "fraction of locally unvoiced frames" of each intervocalic plosive from the onset to the offset of its closure. As for the utterance-initial plosives, the onset of their closure cannot be determined acoustically if they are voiceless (and they nearly always are in Romeyka), as they will produce no sound until the closure is released. This means that their closure duration cannot be directly measured. We used an indirect estimation; we assumed that the closure duration of each utterance-initial token was the same as the mean duration of word-initial but intervocalic plosives that we measured for a separate experiment, i.e. 91, 79 and 70 ms for /p, /t/ and /k/ respectively. This admittedly only provides a rough estimation of their real duration, but we believe it can at least highlight broad tendencies.

As with VOT, most utterance-initial plosives cluster around a prototypical value, that is complete or almost complete absence of voicing (Figure 3), while most intervocalic plosives range from mild to complete voicing (Figure 4).

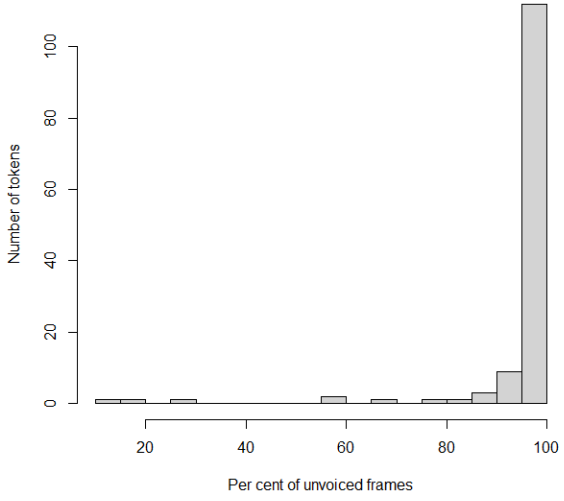


Figure 3: Fraction of locally unvoiced frames utterance-initially.

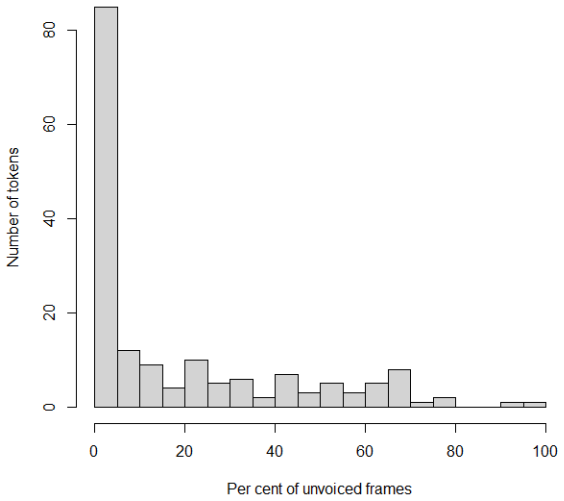


Figure 4: Fraction of locally unvoiced frames intervocalically.

3.3. F_0 of the following vowel

It has been often observed (see e.g. [10]) that voiced consonants tend to lower the fundamental frequency of following vowels. For each post-plosive vowel, we measured the interval between vowel onset and vowel midpoint, divided this interval by ten and extracted their f_0 values. The resulting f_0 contours are reported in Figures 5, 6 and 7.

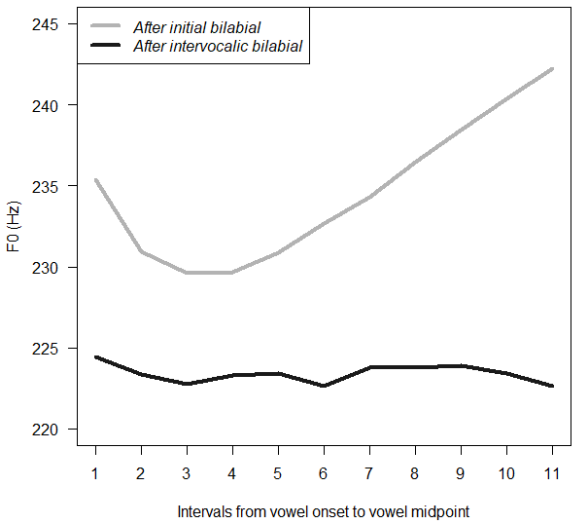


Figure 5: Mean f_0 after bilabial plosives

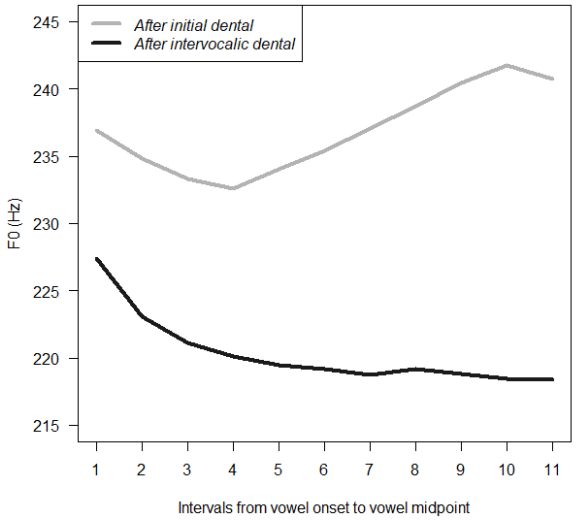


Figure 6: Mean f_0 after dental plosives.

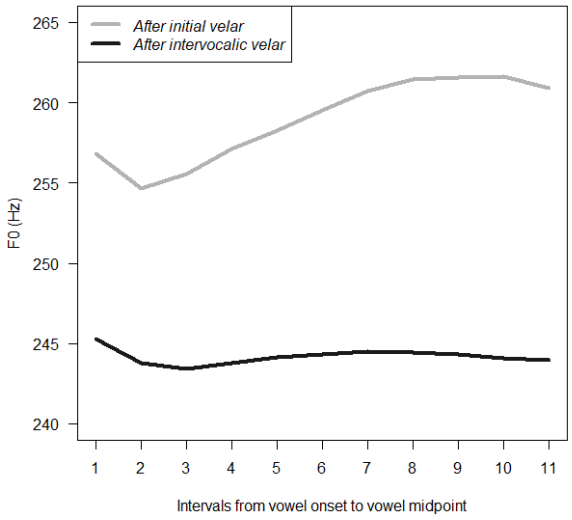


Figure 7: Mean f_0 after velar plosives.

4. DISCUSSION

The three indicators offer a broadly similar picture. As for plosives at the start of an utterance, in most cases they are unaspirated/weakly aspirated voiceless segments. Their VOT is almost always moderately positive, there is scarce if any acoustic trace of voicing during their closure, the fundamental frequency of the following vowel is not lowered. Nonetheless, occasionally a fully voiced initial plosive may occur (Figure 8).

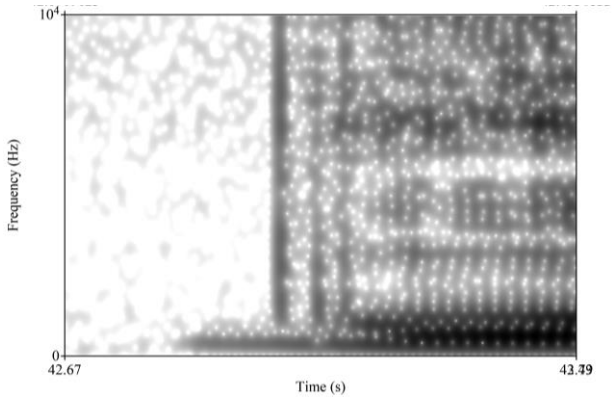


Figure 8: Initial portion [bo] of an utterance-initial token of the word *porta* /porta/ 'door', with voice bar clearly visible.

As for intervocalic plosives, they display more variability. The most common option is a segment with a moderate amount of voicing: moderately negative VOT, acoustic traces of voicing for most of the closure, some lowering of the following vowel's f_0. However, our data also show the less common but not rare occurrence of fully or nearly fully voiceless allophones, as well as of fully voiced allophones. We suspect that even the 'moderate voicing' tokens might in fact consist of two fairly distinct phonation types: some of them appear to have breathy voice throughout the consonant, while in others modal voicing is present at closure onset but then fades away, only to resume again shortly before the closure offset.

5. CONCLUSIONS

As our discussion suggested, several aspects of plosive voicing in Romeyka would deserve further investigation. Other aspects, such as the effect of stress on intervocalic voicing, have not been discussed at all. We plan to explore them in future research, as well as to examine other areas of Romeyka phonetics and phonology.

7. REFERENCES

[1] Armostis, S., Voniati, L., Drosos, K., Tafiadis, D. 2022. Trapezountian Pontic Greek in Etoloakarnania. *Journal of the International Phonetic Association*, 52(2), 328–340.

[2] Mackridge, P. 1987. Greek-speaking Moslems of north-east Turkey: Prolegomena to a study of the Ophitic sub-dialect of Pontic. *Byzantine and Modern Greek Studies* 11(1), 115–137.

[3] Schreiber, L., Sitaridou, I. 2017. Assessing the sociolinguistic vitality of Istanbulite Romeyka: an attitudinal study. *Journal of Multilingual and Multicultural Development*. 39(1), 1-16.

[4] Schreiber, L. 2018. Romeyka. In: Haig, G., Khan, G. (eds), *The Languages and Linguistics of Western Asia*. 6, 892-934.
[5] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program].
[6] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384–422.
[7] Abramson A. S., Whalen D. H. 2017. Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics* 63, 75–86.
[8] Kiss, Z. G. 2013. Measuring acoustic correlates of voicing in stops and fricatives. In Szigetvári, P. (ed.), *VLlxx. Papers in Linguistics. Presented to László Varga on his 70th Birthday*. Tinta Publishing House, 289–312.
[9] Eager, C. D. 2015. Automated voicing analysis in Praat: Statistically equivalent to manual segmentation. *18th ICPhS Glasgow*.
[10] Hombert, J. M., Ohala, J. J., Ewan, W. G. 1979. Phonetic explanations for the development of tones. *Language*. 55(1), 37–58.

WORD CLASS AND FREQUENCY EFFECTS IN HAWAIIAN STRESSED VOWEL CLUSTERS

Thomas Kettig

Department of Languages, Literatures & Linguistics, York University, Canada
tkettig@yorku.ca

ABSTRACT

Lexical factors like grammatical class and usage frequency condition phonetic variation. Previous research has found that sounds in more frequent words and function words tend to be shorter and hypo-articulated compared with less frequent words and content words. However, these patterns have been explored in a limited range of languages. This study investigates vowel reduction based on word class and frequency in Hawaiian. Durations and F1/F2 were measured for word-final, primary stressed clusters /ai, au, ei, ou/ (n=1046) in natural speech.

/ai/ is centralized in non-content words, with less diphthongal and shorter reflexes in higher-frequency words; /au/ is centralized, less diphthongal, and shorter in non-content words; and /ou/ is shorter in non-content and more frequent words. This finding that vowel clusters in non-content and higher-frequency words tend to be phonetically reduced in Hawaiian is relevant to phonological theories which remain to be tested in a wide typological range of languages.

Keywords: Hawaiian, vowels, reduction, word frequency, grammatical class

1. BACKGROUND ON HAWAIIAN

ʻŌlelo Hawaiʻi (Hawaiian) is the Indigenous Austronesian language of the Hawaiian Islands. After a steep decline, Hawaiian had only a few hundred speakers in the 1970s and 1980s, when bans on the language were lifted and community leaders spearheaded innovative revitalization programs. *Ka Leo Hawaiʻi* (KLH), a radio show which regularly broadcast interviews with native speakers beginning in 1972, was an early catalyst in the Hawaiian renaissance [1, 2]. The digitized KLH archives are widely used today in educational contexts.

Despite the availability of extensive audio archives, the phonetics and phonology of Hawaiian remain understudied. The present study investigates three aspects of variation – trajectory position, trajectory length, and duration – in a subset of primary stressed vowel clusters, probing the effects of wordform frequency and word class. The data analyzed here come from KLH interviews recorded in 1972-1974 with eight elderly native speakers. The results have implications not only for descriptions of Hawaiian, but also for language-independent theories of phonological representation and phonetic performance that are seldom tested on Austronesian languages.

2. FREQUENCY AND WORD CLASS

Reduction, the shortening and/or less full articulation of words or phones under certain conditions, is typically associated with unstressed environments [3, 4], but variation in hyper- and hypo-articulated outputs can also be observed within syllables bearing lexical stress. Higher **frequency** words have been found to have shorter acoustic durations than low frequency words in English [5], Dutch [6], and Chinese [7] and have been found to be centralized in the vowel space in English [8] and French [9] (though see [10] for conflicting observations in German).

Most psycholinguistic theories of speech assume separate processes for function and content word production (e.g. [11]). **Grammatical word class** thus also plays a role in vowel variation, with the phones in function words and pronouns typically reduced compared to content/lexical words. Such word class variation has been noted in English [12], Dutch [13], French [9], Icelandic [14], Scottish Gaelic [15], Vietnamese [16], and Japanese [17].

Observations of vowel reduction in higher-frequency and non-content words – words that tend to be most predictable [18] – are often taken as evidence for theories like the Smooth Signal Redundancy Hypothesis (SSRH): In order to balance articulatory ease with communicative success, speakers exhibit an inverse relationship between language redundancy and articulatory effort [19], resulting in undershoot of full durational and quality targets in reduced environments [20, 21]. When studies find an effect of word class even after accounting for prosodic effects predicted by the SSRH, their findings are taken to support a phonological model with independent lexical representations for homophonous words of different classes [22, 23]. Though these models of linguistic storage and performance are language-independent, the range of languages on which their predictions have been tested remains limited and typologically skewed toward Indo-European lects.

3. HAWAIIAN PHONOLOGY

In its native lexicon, Hawaiian has the vowel phonemes /i e a o u/ and /iː eː aː oː uː/, as well as the consonants /ʔ, h, l~r, m, n, p, k~t, v~w/ [24, 25]. The short diphthongs /ae ai ao au ei eu iu oi ou/ and the long diphthongs /aːe aːi aːo aːu eːi oːu/ are under most analyses not unitary phonemes, and are referred to herein as tautosyllabic clusters comprised of two separate vowel phonemes [25, 26, 27].

Syllable structure in Hawaiian is highly constrained, with no codas and no complex onsets in the native lexicon. Primary lexical stress is assigned to the syllable containing the penultimate mora of the word; stress is thus final in words ending in a tautosyllabic cluster or long vowel, and penultimate in words ending in a singleton short vowel [28, 29].

Kettig [27] detailed the acoustic profile of the Hawaiian vowel system based on F1/F2 measurements of the eight-speaker KLH dataset also used in the present study. He observed that the trajectories of /ai/ and /au/ in pronouns and directional particles exhibit trajectory reduction compared to their content word equivalents, though he did not investigate durations or report on word frequency. Previously, Parker Jones [25] reported qualitative reduction in Hawaiian /a/ in function words based on a small single-speaker sample.

4. RESEARCH QUESTIONS AND DATA

The present study probes two main research questions regarding whether Hawaiian fits reduction patterns previously described in unrelated languages:

- Does higher word frequency correlate with reduced vowel durations and trajectories?
- Do non-content words like pronouns and directionals have reduced vowel durations and trajectories compared with content words?

The data here comprise recordings of radio interviews from the 1970s, and thus represent an unscripted speech style in a studio environment. The eight speakers are one male and one female elder, each from the islands of Kauaʻi, Oʻahu, Maui, and Hawaiʻi [30]. Their interviews were transcribed and force-aligned using the Montreal Forced Aligner [31], with overlapping speech data discarded and all phone and word boundaries manually checked for accuracy. Durations and F1/F2 measurements at nine points in each token were extracted using the Fast Track plugin [32] in Praat [33], with re-measurements and outlier exclusion workflows outlined in [27]. Vowels were normalized using the ANAE log-diff method [34, 35].

Word frequency measures were obtained from Brockway [36] and are based on word counts in the first 40 transcripts of KLH episodes, which contained 4,826 word types and 315,785 tokens.

The clusters /ai, au, ei, ou/ were selected for analysis because the lexicon of Hawaiian contains both content and non-content words (specifically pronouns, pluralization/directional/verbal particles, and a question word) with each of these clusters bearing primary lexical stress. Table 1 provides examples of words analyzed as well as token counts after exclusions; word type counts are also provided in parentheses.

	Content word	Non-content word
/ai/	/ʔ**ai**/ eat n=141 (23)	/m**ai**/ hither n=237 (2)
/au/	/p**au**/ finished n=175 (25)	/v**au**/ 1SG n=159 (3)
/ei/	/l**ei**/ garland n=12 (6)	/n**ei**/ PROX n=61 (1)
/ou/	/h**ou**/ new n=17 (2)	/ˌkaːˈ**kou**/ 1PL.IN n=244 (5)

Table 1: Examples of words containing each of the clusters with word-final primary lexical stress; n=token counts (word type counts in parentheses).

Several exclusions were made to control for and anticipate some possible effects outside the scope of the present investigation: (1) Tokens immediately preceded or followed by a vowel without an intervening consonant (e.g. *...**lei** o ka...*) were excluded due to difficulty determining phone boundaries. (2) Only word-final primary stressed clusters are analyzed. Though primary stress can fall on either the final or penultimate syllable in Hawaiian, only one non-content word contains one of these primary stressed clusters in a penultimate syllable (/m**ai.**la/ 'hither'). In content words, token counts were also higher for clusters in final position. (3) Tokens immediately followed by a pause were excluded. The present dataset is not parsed for phrase boundary types, nor have the pauses been tagged for which indicate true phrase finality and which arise due to disfluencies or speaker overlap; it would therefore be difficult to disentangle the various types of utterance- and phrase-final lengthening observed cross-linguistically and noted previously in Hawaiian [25]. (4) Two tokens of greater than 500 ms were excluded, as inspection of these outliers showed that they were due to hesitation and listing prosody.

5. ANALYSIS AND RESULTS

5.1. Constructing Linear Mixed Effects Models

The following sections report the results of a series of linear mixed effects models using the R package *lmerTest* [37] which were constructed to measure the effects of word class and frequency on trajectory centralization, trajectory length, and duration. Each of these models contained the same structure: Fixed

effects were binary content word status (content/non-content) and word frequency (operationalized by the log normalized frequency), with a by-speaker random intercept in each model. Due to a high co-linearity between word class and frequency (with content words tending to have lower frequencies than non-content words), testing interactions in these models caused convergence issues, though future work on an expanded dataset should investigate this possible confound.

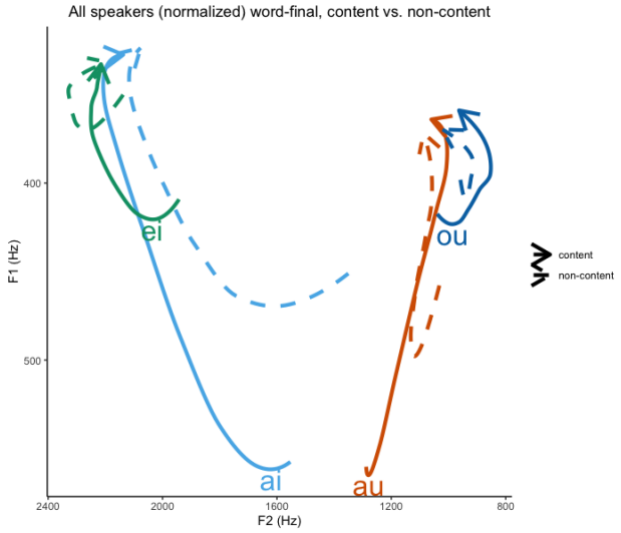


Figure 1: Mean trajectories of word-final, non-pre-pausal, primary stressed vowel clusters in content words (solid) and non-content words (dotted).

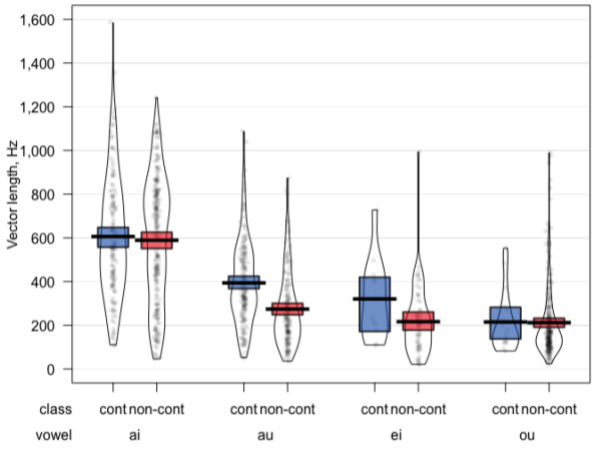


Figure 2: Trajectory length (vector length between F1 max and min) in content and non-content words.

5.2. Trajectory position

Fig. 1 plots the trajectories of /ai/, /au/, /ei/, and /ou/, split by whether they appear in tokens of content or non-content words. The lowest points of the /ai/ and /au/ trajectories appear to be centralized in non-content words compared to content words; /ei/ and /ou/ also appear to be more monophthongal in non-content words than content words.

To statistically test the centralization of the initial element /a/, two models were constructed, one for /ai/ and one for /au/; the predicted variable was the maximum F1 measurement within each token – that is, the lowest point in its trajectory. For /ai/, F1 was found to be significantly reduced (higher in the vowel space) in non-content tokens compared to content tokens (β=-95.8, SE=21.0, t=-4.6, p<.001). For /au/, F1 was significantly reduced in non-content tokens (β=-103.2, SE=15.2, t=-6.8, p<.001). In these models, word frequency did not significantly predict either /ai/ nor /au/ F1 maximum.

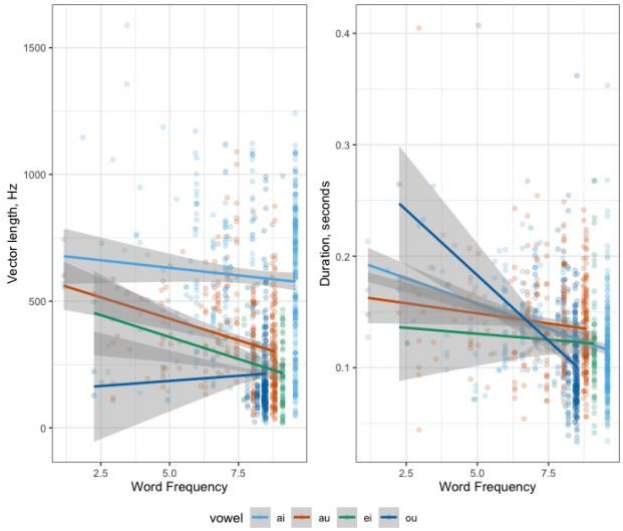


Figure 3: Trajectory length (left) and duration (right) by word frequency. Shaded standard error ranges.

5.3. Trajectory Length

To investigate whether word class or frequency predicted the length of trajectories – reflective of how monophthongal or diphthongal a given token was – four models were constructed, one for each cluster. Trajectory vector length was operationalized by calculating the F1/F2 Euclidean distance between the point of maximum F1 and the point of minimum F1 for each token (a modification of vector length calculations described by e.g. [38, 39]). A singular fit error was encountered for the /ei/ cluster model due the presence of only one non-content word type, so statistical test results are not reported for this cluster.

As illustrated in Fig. 2, a vector length difference between word classes was significant for /au/ (β=-132.9, SE=25.8, t=-5.2, p<.001), indicating that non-content words are pronounced with more monophthongal reflexes than content words. This was not a significant effect for /ai/ or /ou/. A vector length difference was predicted by word frequency for /ai/ (β=-31.9, SE=15.1, t=-2.1, p=.03), indicating that more frequent /ai/ words contain more monophthongal reflexes than less frequent ones. Though Fig. 3 (left) shows /au/ with a steeper

negative slope than /ai/, frequency was not a significant predictor for /au/ or /ou/ in these models.

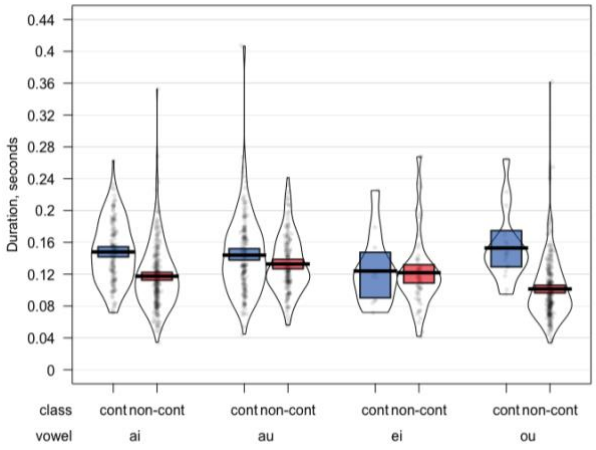


Figure 4: Cluster duration in content and non-content words.

5.4. Duration

Fig. 3 (right) shows token durations by word frequency and Fig. 4 illustrates duration differences between content and non-content words. Though it appears in Fig. 4 that /ai/ exhibits duration differences by word class, this effect was not significant; instead, frequency was found to significantly predict duration, with lower durations associated with higher frequencies (β=-0.008, SE=0.002, t=-3.9, p<.001). On the other hand, word class was a significant predictor for /au/ duration (β=-0.016, SE=0.006, t=-2.5, p=.012) while frequency was not. Neither class nor frequency were significant predictors for /ei/, while for /ou/, both word class (β=-0.036, SE=0.010, t=-3.6, p<.001) and frequency (β=-0.013, SE=0.002, t=-2.9, p=.004) were significant.

6. DISCUSSION

To summarize: /ai/ exhibits significant centralization in non-content words, as well as less diphthongal reflexes and reduced durations in higher-frequency words; /au/ has significant centralization, monophthongization, and durational reduction in non-content words; /ou/ has significant durational reduction based on word class and frequency; and /ei/ appears to be more monophthongal in non-content words, but statistical power in this sample was too low to obtain results. These findings of duration and trajectory variation in Hawaiian stressed vowel clusters are in line with Parker Jones' [25] previous observation for /a/ as well as cross-linguistic expectations, and may represent the first time these research questions have been tested on a substantial multi-speaker dataset in an Austronesian language.

Though a thorough discussion is outside the scope of the present paper, interspeaker variation was apparent in this sample. Inspection of individual speaker vowel plots showed that some of these elders exhibited large differences between content and non-content words, while others reduced very little across grammatical classes. Future work on this corpus should focus on factors that might condition variation in reduction rates, such as differing degrees of English/Hawaiian bilingual dominance or levels of comfort with the formal radio interview setting.

Future work should also address other factors that this study was not able to consider, particularly with regard to variation in vowel duration. For instance, accounting for within- and between-speaker speech rate differences may allow for clearer patterns in duration variation to emerge. While all tokens preceded or followed by vowels were excluded from consideration, the present study included tokens before or after /w/, /h/, and /ʔ/, for which segmentation was often difficult; excluding all such data would have massively reduced the number of words it was possible to analyze, but these tokens may have introduced inconsistencies (cf. [40]).

One challenge for researching phonetic variation in endangered and understudied languages like Hawaiian is that their phone-aligned datasets of spontaneous speech tend to be small, yielding a limited spread of words spoken per person and too few iterations per word to have statistical power. This investigation encountered such issues with statistical power in the /ei/ vowel, so future work should test the present observations on an expanded dataset. However, though the number of tokens analyzed here is several orders of magnitude smaller than those used to run similar tests in better-resourced languages, the results suggest that Hawaiian does exhibit the putative universal tendency for vowels in non-content and high-frequency words to be phonetically reduced in primary stressed position. The relative strength of the class and frequency effects may vary by vowel, and further research is necessary to disentangle the role of these two lexical factors.

As Hawaiian becomes better documented, with more thoroughly-tagged and larger phonetic datasets available, other predictors used to disentangle prosodic, grammatical, and frequency-based variation (such as lemma frequency, speech rate, position within word/utterance/phrase, etc.) will hopefully become available for improved statistical modeling. The continued documentation of such phonetic variation in not only Hawaiian, but also the world's thousands of other understudied languages, is crucial in order to test phonological theories against a set of empirical evidence reflective of the full range of human linguistic experience.

7. ACKNOWLEDGEMENTS

Mahalo to Rory Turnbull, ʻŌiwi Parker Jones, Lisa Davidson, and Nicole Holliday for their helpful comments and feedback. Mahalo hoʻi iā Larry Kimura a i nā kūpuna mānaleo o KLH.

8. REFERENCES

[1] Kimura, L. 2015. Ka Leo Hawaiʻi tapes and the regenesis of Hawaiian. *4th ICLDC*. http://hdl.handle.net/10125/25329.
[2] Wilson, W. H., Kamanā, K. 2019. ʻAha Pūnana Leo – Advancing From the Grassroots. *Linguapax* 7, 93–112.
[3] Crosswhite, K. 2004. Vowel reduction. In: Hayes, B., Kirchner, R., Steriade, D. (eds), *Phonetically Based Phonology*. Cambridge, 191–231.
[4] Ernestus, M., Warner, N. 2011. An introduction to reduced pronunciation variants. *J. Phon.* 39, 253–260.
[5] Gahl, S. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3), 474–496.
[6] Pluymaekers, M., Ernestus, M., Baayen, R. H. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *J. Acoust. Soc. Am*. 118(4), 2561–2569.
[7] Sherr-Ziarko, E. 2015. Word frequency effects on homophonous words in Mandarin Chinese. *Proc. ICPhS 2015*.
[8] Munson, B., Solomon, N. P. 2004. The Effect of Phonological Neighborhood Density on Vowel Articulation. *J. Speech, Lang. & Hearing Res.* 47(5), 1048.
[9] Meunier, C. and Espesser, R. 2011. Vowel reduction in conversational speech in French: The role of lexical factors. *J. Phonetics* 39(3), 271–278.
[10] Tomaschek, F., Wieling, M., Arnold, D., Baayen, R. H. 2017. Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. *INTERSPEECH 2013,* 1302–1306.
[11] Garrett, M. F. 1989. Processes in language production. In: Newmeyer, F. J. (ed), *Volume III: Language: Psychological and Biological Aspects*. Cambridge, 69–96.
[12] Shi, R., Gick, B., Kanwischer, D., Wilson, I. 2005. Frequency and category factors in the reduction and assimilation of function words: EPG and acoustic measures. *J. Psycholing. Research* 34(4), 341–364.
[13] Van Bergem, D. R. 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication* 12(1), 1–23.
[14] Schäfer, M. 2013. *Phonetic reduction of adverbs in Icelandic: On the role of frequency and other factors.* Doctoral diss., Universität Freiburg.
[15] Nance, C. 2015. ʻNewʼ Scottish Gaelic speakers in Glasgow: A phonetic study of language revitalisation. *Lang. in Soc.* 44(4), 553–579.
[16] Nguyễn, T. A. T. 2015. Vowel reduction in conversational speech in Vietnamese. *Intl. J. of Asian Lang. Processing* 23(2) 91–109.
[17] Shirai, S. 2005. *Lexical effects in Japanese vowel reduction*. Doctoral diss., University of Washington.
[18] Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee, J., Hopper, P. J. (eds), *Frequency and the Emergence of Linguistic Structure*. John Benjamins, 229–254.
[19] Aylett, M., Turk, A. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang. & Speech* 47(1), 31–56.
[20] Lindblom, B. 1963. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35(11), 1773–1781.
[21] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, W. J., Marchal, A. (eds), *Speech Production and Speech Modelling*. Kluwer, 403–439.
[22] Drager, K. 2011. Sociophonetic variation and the lemma. *J. Phon.* 39(4), 694–707.
[23] Lohmann, A. 2018. Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *J. Linguistics* 54(4), 753–777.
[24] Schütz, A. J. 1981. A reanalysis of the Hawaiian vowel system. *Oceanic Ling.* 20(1), 1–43.
[25] Parker Jones, ʻŌ. 2018. Hawaiian. *J. IPA* 48(1), 103.
[26] Rehg, K. L. 2007. Does Hawaiian have diphthongs? And how can you tell? In: Siegel, J., Lynch, J., Eades, D. (eds), *Language Description, History and Development*. John Benjamins, 119–131
[27] Kettig, T. 2021. *Haʻina ʻia mai ana ka puana: The vowels of ʻōlelo Hawaiʻi*. Doctoral diss., University of Hawaiʻi at Mānoa.
[28] Schütz, A. J. 2010. Measures and morphemes: A functional approach to Hawaiian accent. In: A. Pawley, J. Bowden, N. Himmelmann, M. Ross (eds.), *A journey through Austronesian and Papuan linguistic and cultural space: Papers in honor of Andrew Pawley*, Australian National University, 405–421.
[29] Parker Jones, ʻŌ. 2010. *A computational phonology and morphology of Hawaiian*. Doctoral diss., University of Oxford.
[30] Kimura, L. L. K. 1972–1974. (Producer, Presenter). Kaniʻāina, ulukau.org Digital Repository of Ka Haka ʻUla O Keʻelikōlani College of Hawaiian Language, University of Hawaiʻi at Hilo, *Ka Leo Hawaiʻi*. ulukau.org/kaniaina:
1972-10-31, HV24.013, Ida Kapuʻihilani Feary-Milton Nāone (guest).
1972-11-09, HV24.014, Rachel Mahuiki (guest).
1972-11-21, HV24.016, Joseph Makaʻai (guest).
1973-02-06, HV24.021, Henry Hanalē Machado (guest).
1973-04-24, HV24.032, Lilian Victor (guest).
1973-05-01, HV24.033, Sadie Kaluhiʻōpiopio Beebe (guest).
1974-03-03, HV24.057, Alfred Apaka, Sr. (guest).
1974-04-21, HV24.063, David Kaʻalakea (guest).
[31] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2018. *Montreal Forced Aligner*. Computer program, version 1.1.
[32] Barreda, S. 2021. Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Ling. Vanguard* 7(1).
[33] Boersma, P., Weenink, D. 2020. *Praat: Doing Phonetics by Computer.* Computer software. Version 6.1.13.
[34] Labov, W., Ash, S., Boberg, C. 2006. *The Atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.
[35] Barreda, S., Nearey, T. M. 2018. A regression approach to vowel normalization for missing and unbalanced data. *J. Acoust. Soc. Am.* 144(1), 500–520.
[36] Brockway, C.E.L., 2021. *Building high-frequency word lists for the semantic domain of ʻĀINA (ʻland') using a raw corpus of spoken ʻōlelo Hawaiʻi*. Doctoral diss., University of Hawaiʻi at Mānoa.
[37] Kuznetsova A., Brockhoff P. B., Christensen R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Software* 82(13), 1–26.
[38] Fox, R. A., Jacewicz, E. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *J. Acoust. Soc. Am.* 126(5), 2603–2618.
[39] Farrington, C., Kendall, T., Fridland, V. 2018. Vowel dynamics in the Southern vowel shift. *American Speech* 93(2), 186–222.
[40] Davidson, L. 2021. Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian. *J. Phon* 88, 101075.

Acoustic characterization of Najdi Arabic Vowel System

Ghazi Algethami

Taif University
g.gethami@tu.edu.sa

ABSTRACT

This study provides an acoustic description of the vowel system of Najdi Arabic, a major dialect spoken in Saudi Arabia. It aims to go beyond the phonological description provided by Ingham (1994) and offer an acoustic description of Najdi Arabic vowels. Fifteen native speakers of Najdi Arabic were recorded reading *CVC* words containing all the Najdi Arabic vowels described by Ingham (1994). As expected, Najdi Arabic vowels were found to occupy a narrow acoustic space. Temporally, the long vowels /i:/, /u:/ and /a:/ were almost double the length of their short counterparts /i/, /u/ and /a/. The long and short vowels also differed spectrally, with short vowels being significantly more centralized than their long counterparts. The classical Arabic diphthongs /ay/ and /aw/, as in many Arabic dialects, were realized as long monophthongs, /e:/ and /o:/, respectively. The results substantiate the impressionistic analysis of Najdi Arabic vowels system by Ingham (1994).

Keywords: Najdi, Arabic, Vowels, Phonetics, Acoustics.

1. INTRODUCTION

Modern Standard Arabic (MSA) is characterized by having a simple vowel system consisting of three long vowels /i:/, /u:/ and /a:/, their short counterparts /i/, /u/, and two diphthongs: /ay/ and /aw/ [1]. MSA is not used in day-to-day conversations and its use is limited to educational and formal contexts (e.g., news broadcasting) [2]. Arab speakers across the Arab world speak different regional dialects in their daily conversations. These dialects can vary widely in their phonological structure [1]. For instance, some variation exists between spoken Arabic dialects with regard to their vowel inventory and their phonetic realizations [3, 4, 5].

Najdi Arabic is one of the major dialects spoken in Saudi Arabia. It is spoken by a large population in central Saudi Arabia. Limited research has discussed the vowels system of Najdi Arabic [6,7], with no acoustic studies validating their phonological auditory description. The current study, thus, attempts to fill this gap by acoustically examining the vowel system of Najdi Arabic.

2. PREVIOUS STUDIES

The vowel system of Classical Arabic is traditionally described "as an example of the simplest kind of triangular vowel system" [8, p. 461]. It consists of three long vowels /i:/, /u:/ and /a:/ and their short counterparts /i/, /u/, and /a/. In addition, it has two diphthongs /ay/ and /aw/. The same can be said about MSA [2,3]. The long vowels are almost double the length of the short ones [9]. While all Arabic spoken dialects still preserve the three long vowels found in MSA, the phonetic realizations of the short vowels and diphthongs show considerable variation [3]. For example, in many colloquial spoken dialects, the two diphthongs /ay/ and /aw/ are realized as mid long vowels /e:/ and /o:/, respectively [3, 4]. A few recent studies have acoustically examined the phonetic realization of vowels in various regional dialects [4, 5]. For example, in Syrian Arabic, the short vowels /i/ and /u/ can be realized phonetically as /e/, /o/ or /ə/ [4]. Comparing Moroccan and Jordanian Arabic vowel systems, Barkat-Defradas et al [5] showed that Moroccan Arabic exhibited a more reduced short vowels than did Jordanian Arabic, leading to the emergence of /ə/ in Moroccan Arabic. It seems that a major source of variation among Arabic dialects is how they phonetically realize short vowels.

Ingham [7] listed the following vowel inventory for NA based on his auditory analysis of the dialect: /i:/, /u:/, /a:/, /i/, /u/, /a/, /e:/, and /o:/, which is illustrated in figure 1 below. On the phonetic realizations of these vowels, he mentioned that /a:/ in Najdi Arabic does not show fronting (aka *imalah*) as it does in Syrian and Egyptian Arabic. In addition, /i/ and /u/ can be centralized and have a "schwa-like quality in neutral environments" (p.15)

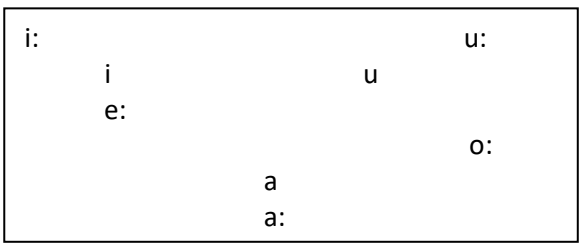


Figure 1: Najdi Arabic Vowel System; reproduced from Ingham [7], p. 15.

Given the scarcity of phonetic studies on Najdi Arabic vowels, the current study aims to acoustically corroborate the auditory description of Najdi Arabic vowels by Ingham (1994).

3. METHOD

Fifteen native speakers of Najdi Arabic were recruited to provide the data for the current study. They were all males and their age ranged from 19 to 37. None of them reported any speech or hearing impairment.

Each participant was met individually with the researcher in a sound-attenuated room. They were asked to read a list of CVC words (see Table 1 below), representing all the vowels described in Ingham [7], included in the carrier sentence *ʕaqra CVC wadxul lilqaʔah* (I read CVC and enter the hall). Some of the participants pointed out that they usually produce the word /kud/ as /kid/ in their local dialect, but they were asked to read it as /kud/ as illustrated to them by the use of the Arabic orthographic representation (كُد). None of them had any problem producing the word as such.

Table 1. The list of words used to elicit the Najdi Arabic vowels along with their IPA transcription and English translation.

Vowel	CVC word (translation)
i	/sid/ (block)
i:	/si:d/ (Master)
u	/kud/ (work hard)
u:	/ku:d/ (perhaps)
a	/bat/ (decided)
a:	/ba:t/ (slept)
e	/be:t/ (home)
o:	/so:t/ (so:t)

Although final-syllable consonants had no lengthening or shortening effect on preceding vowels in Saudi Arabic [10], an attempt was made to ensure each vowel pair of the short/long vowels is used by the same consonantal *CVC* environment to neutralize the effect of phonological context.

The participants were asked to read the sentences in their spoken dialect in normal speech pace. Since written Arabic is largely associated with MSA, I read all the sentences in dialectal Arabic first to set the required register for them and asked them to read the sentences as I did in their native dialect. The participants' renditions were digitized and saved into a computer drive for acoustic analysis using Praat software [11]. Spectral (midpoint F1/F2 formant values) and temporal data were obtained for each vowel per each speaker.

4. RESULTS AND DISCUSSION

Table 2 below presents the obtained average vowel formants (F1 and F2) and duration (in millisecond) for each vowel.

Table 2. Average F1 and F2 values and duration for all vowels.

Vowel	F1	F2	Duration
/a:/	653.743	1324.112	138.077
/a/	552.203	1445.433	70.26097
/i:/	371.1182	2091.597	102.3378
/i/	429.99	1626.148	54.22397
/u/	454.5769	1401.49	53.12449
/u:/	424.9873	1026.86	100.749
/e:/	497.603	1826.558	125.7461
/o:/	545.8368	1016.206	123.4459

Long vowels are almost double the duration of their short counterparts. The mean durational ratio of long to short vowels for all three vowel pairs is 1.9. This is similar to the result found in Alghamdi [9] for MSA produced by Saudi Arabians. The results generally corroborate the auditory description of vowel length in Najdi Arabic demonstrated by Ingham [7].

Figure 2 below plots the mean F1 and F2 values for all vowels to examine the phonetic realization of Najdi Arabic vowels. Short vowels are more centralized and reduced in quality than their long counterparts. Similar results were obtained for Jordanian Arabic in [4]. Different from the description provided by Ingham [7], /a/ is more centralized and slightly fronted.

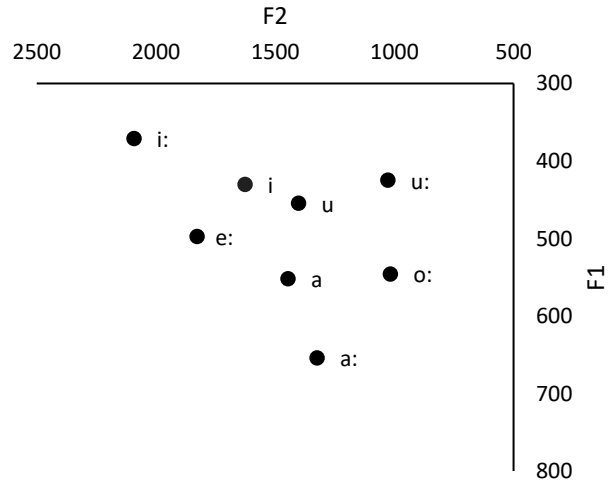


Figure 2. F1/F2 Plot for all Vowels

To capture individual speaker variation, the production of all speakers for each vowel is presented in Figure 3 below after normalization. Apart from the small spectral overlap between /i/ and /u/, the mean formant values show similar results to the individual and normalized formant values.

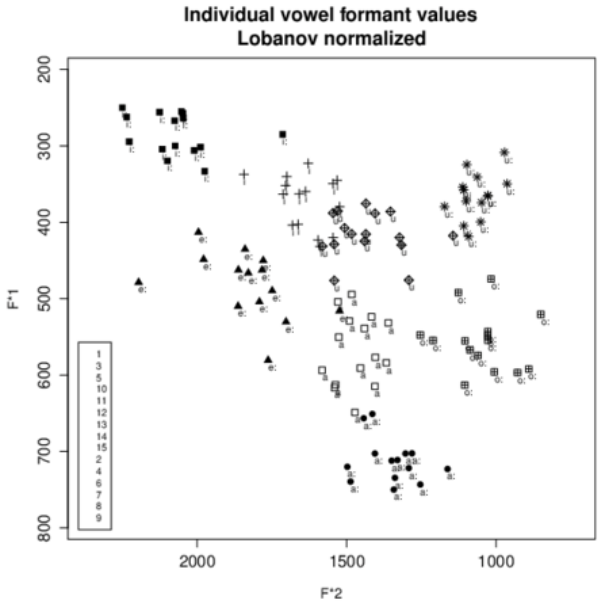


Figure 3. F1/F2 values of speakers' production of each vowel after normalization (Labov transformation).

To sum up, the vowel system of Najdi Arabic consists of three long vowels /i:/, /u:/ and /a:/ and three short counterparts /i/, /u/, and /a/, which differ in both quantity and quality. The short vowels are almost half the duration of the long ones and are significantly more centralized. Najdi Arabic also demonstrates, as many spoken Arabic dialects, the monophthongization of the MSA diphthongs /ay/ and /aw/, which resulted in the long vowels /e:/ and /o:/, respectively.

5. CONCLUSION

The findings of the current acoustic study corroborate the auditory description of Najdi Arabic vowels by Ingham [7]. Short vowels are almost half the length of the long vowels, and they are also more centralized or reduced in quality than long vowels. However, short vowels still form different spectral categories in their own. The only difference between the auditory description provided by Ingham [7] and the finding of the current study is that /a/ is more centralized and slightly fronted in the current study.

6. REFERENCES

[1] Ryding, K. 2005. *A reference grammar of modern standard Arabic*. Cambridge University Press.
[2] Holes, C. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.
[3] Watson, J. 2007. *The phonology and morphology of Arabic*. Oxford University Press.
[4] Almbark, R., Hellmuth, S. 2015. Acoustic analysis of the Syrian vowel system. *Proc. 18th ICPhS* Glasgow.
[5] Melissa, B., Al-Tamimi, J., Benkirane, T. 2003. Phonetic variation in production and perception of speech: a comparative study of two Arabic dialects. dialects. *Proc. 15th ICPhS* Barcelona.
[6] Prochazka, T. 2010. *Saudi Arabian Dialect*. Routledge.
[7] Ingham, B. 1994. Najdi Arabic: Central Arabian. John Benjamins Publishing.
[8] Ferguson, C. (1957) Two Problems in Arabic Phonology. *WORD* 13, 460-478.
[9] Alghamdi, M. 1998. A spectrographic analysis of Arabic vowels: A cross-dialect study. *Journal of King Saud University* 10, 3-24.
[10] Flege, J., Port, R. 1981. Cross-language phonetic interference: Arabic to English. *Language and speech*, 24, 125-146.
[11] Boersma, P., & Weenink, D. 2021. Praat: doing phonetics by computer [computer program].

ACOUSTIC AND AERODYNAMIC FEATURES OF HADZA CLICKS

Didier Demolin[1], Alain Ghio[2] & Andrew Harvey[3]

[1]LPP, UMR-7018, CNRS, Sorbonne Nouvelle, Paris
[2]Laboratoire Parole et Langage, CNRS-UMR 7309, Aix-Marseille Université
3 University of Bayreuth Faculty of Languages and Literatures

ABSTRACT

Hadza has 4 clicks in its sound inventory: bilabial, dental, alveolar and lateral [ʘ, |, !, ǁ] that can be aspirated, glottal and nasal [ʘ̬, |ˀ, |ʰ, ᴺ|, ᴺ|ˀ, !ˀ, !ʰ, ᴺ!, ᴺ!ˀ, ǁ, ǁʰ, ᴺǁʰ, ᴺǁˀ]. The bilabial click [ʘ] has a noisy release and is defined as [grave & noisy]. Acoustically the oral clicks [|, !, ǁ] are described with 2 features [grave vs. acute] and [abrupt vs. noisy]. The [grave] feature defining the alveolar click [!] has a peak in the FFT spectrum around 2 kHz. The [acute] feature defining the dental and lateral clicks [|, ǁ] has a peak around 6 kHz for the dental and an emphasis in frequencies between 4 & 5 kHz for the lateral click (all peaks in rounded values). The difference between the [abrupt] [!, ǁ] also has a duration component, i.e. short burst vs short noise band.

Keywords: Hadza, Clicks, Acoustic features, Aerodynamics.

1. INTRODUCTION

Hadza an isolate endangered language spoken in Tanzania has non-pulmonic consonants, clicks and ejectives, in its phonetic / phonological inventory. [1], [2] and [3] describe Hadza with 9 clicks while [4] suggests that there are 12. Recent data show that the language has 13: [ʘ̬, |ˀ, |ʰ, ᴺ|, ᴺ|ˀ, !ˀ, !ʰ, ᴺ!, ᴺ!ˀ, ǁ, ǁʰ, ᴺǁʰ, ᴺǁˀ]. Thus, there are 4 types of clicks with various accompaniments [k, g, q, ʰ, ˀ, ʱ, ᴺ, ᴺ̡] that can sometimes be combined.

The aim of this paper is to describe the acoustic features of Hadza clicks, evaluate the gestures involved with a measure of time and establish their distinctive features from acoustic and aerodynamic data. A precise articulatory and acoustic understanding of the mechanisms involved in their production and perception is necessary to characterize and formalize these consonants in terms of distinctive features. Another issue that will addressed is: are there aspirated clicks in Hadza? This implies to carefully establish the characteristics of the accompaniments for the different clicks. A more general issue involved in this study if to understand the mechanisms of the production of these segments and see how this type of data assesses, or extends, the limits of our knowledge on the diversity and functioning of speech production in languages.

2. MATERIAL AND METHOD

2.1 Speakers and corpus

Words including all Hadza consonants were recorded. Each word was repeated 3 times by 5 male and 4 female native speakers. For the present study (that only analyzes data from 2 females & 2 males) the following set of data has been used to compare the production of all the speakers. 1) [ʘ] Bilabial click [mʘ̬ʷa] ‘Thank God’; 2) [|] Dental click (glottal, aspirated, nasal, glottal nasal) : [|ˀu] ‘To end on an errand’, [|alʰa] ‘Threshing rock’, [ᴺ|̡ata] ‘Tongue’, [tuᴺ|̡ˀe] ‘To pluck feathers’; 3) [!] Alveolar click (simple, nasal, aspirated, glottal): [la!o] ‘To trip’, [kuᴺ!̡ana] ‘Kudu’, [!ʰaku] ‘To jump’, [!ˀoje] ‘Beeswax’, [!ˀakx̡itɬa] ‘The hard palate’; 4) [ǁ] Lateral click (simple, aspirated, glottal nasal, nasal): [ǁupʰi] ‘To sleep’, [ǁʰa] ‘To throw’, [beᴺǁ̡ˀe] ‘To shine’, [ᴺǁ̡a:] ‘For the whole camp to visit another one’. As the bilabial click has very limited occurrences, the main focus is set on the dental, alveolar and palatal clicks.

2.2 Method

Aerodynamic and acoustic data were recorded with an EVA2 Workstation [5] allowing simultaneous recordings of several aerodynamic parameters: intraoral pressure (Po) measured in hecto-Pascal (1hPa = 1.02cm H_2O), oral airflow and if necessary nasal airflow in dm^3/s. Audio recordings, synchronized with the aerodynamic data, are recorded with a microphone integrated to the EVA2 transducers. Oral Airflow (Oaf) was collected using a flexible silicone mask set on the subjects’ mouth. Intraoral pressure (Po) was recorded by inserting a small tube into the pharynx through the nasal cavity.

3. RESULTS

3.1 Introduction

Clicks have acoustically two components: an attack transient and an extinction transient. The attack transient is the explosion noise (sometimes called a burst) which is the impulse response to a change in the shape of the vocal tract. The extinction transient is the noise associated with the turbulent release. Figure 1 shows typical audio waveform shapes for the 4 types of clicks found in our Hadza data. The attack and extinction transients are indicated by arrows on the alveolar audio waveform. Acoustic measurements consist in duration for all clicks, duration of the glottal and aspirated parts following the attack transient and a FFT spectrum made in a 10 ms Hamming window around the peak of the initial transient noise.

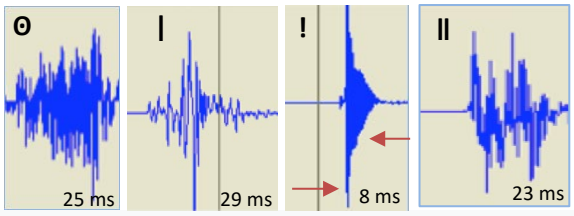


Figure 1. Audio waveform of the 4 clicks types.

3.2 Click duration

Figure 6 shows the comparison of duration, including the attack and extinction transients, between the dental, alveolar and lateral clicks made with 4 speakers. Results show that the alveolar is shorter compared to the dental and lateral clicks. The alveolar [!] is sharp compared to the dental [ǀ] that has a noisy or diffuse release and has a greater duration. This click has 2 phases: an initial rising until the peak of the release that is followed by a decreasing noise. The lateral click [ǁ] has a sharp release that lasts in intensity for an average of 23 ms before a sharp fall. This click is therefore produced as a noise band. These features are also displayed on the spectrograms of Figure 2, 3, 4 & 5.

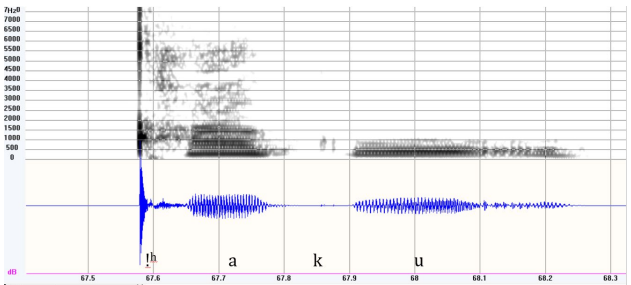


Figure 2. Audio waveform and spectrogram of the word [!ʰaku] ‘to jump’

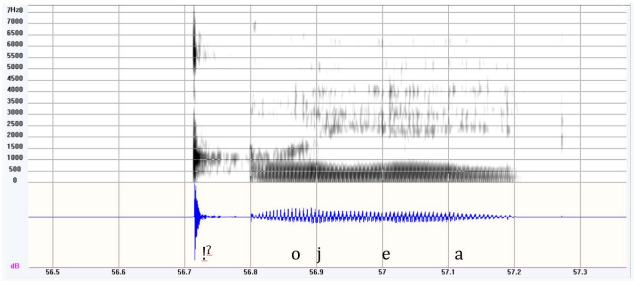


Figure 3. Audio waveform and spectrogram of the word [!ʔojea] ‘beeswax’

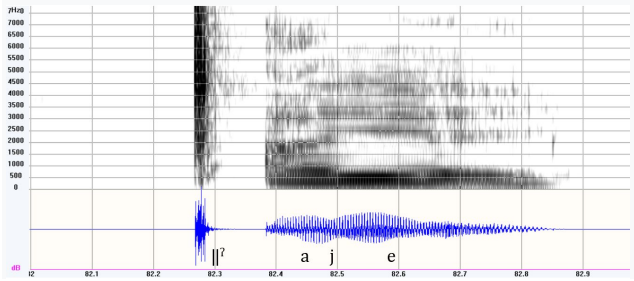


Figure 4. Audio waveform and spectrogram of the word [ǁʔaje] ‘to throw’

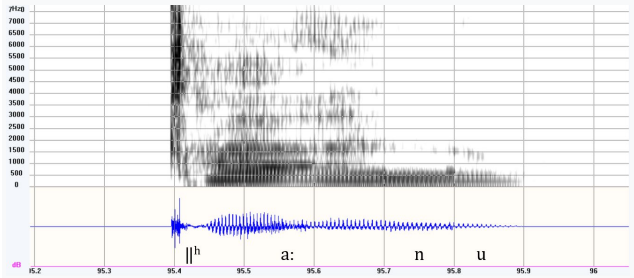


Figure 5. Audio waveform and spectrogram of the word [ǁʰanu] ‘dog’

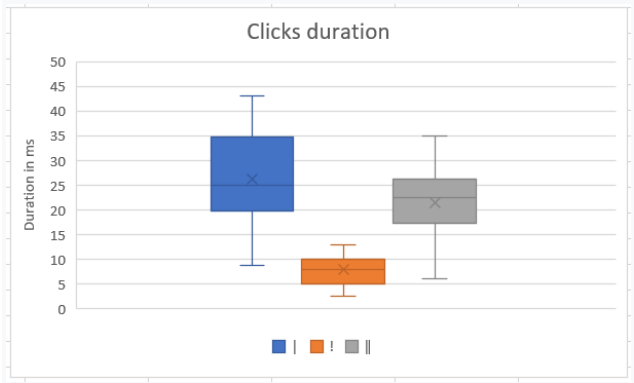


Figure 6. Mean duration of the dental, alveolar and lateral clicks measured produced by 4 Hadza speakers (n=70).

3.3 Click accompaniments

Spectrograms of Figures 2 and 5 show realizations of aspirated clicks. All clicks but the bilabial can be aspirated. The average values of the aspirated accompaniments are displayed in Figure 6. Aspiration noise following the transients has sorter duration when compared to the glottal lag accompanying all

clicks. Nasal clicks start by a nasal portion (sometimes devoiced) before the transient release.

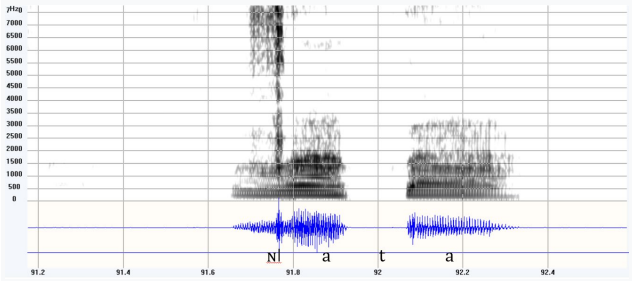


Figure 7. Audio waveform and spectrogram of the word [ᴺǀʰata] ‘tongue’

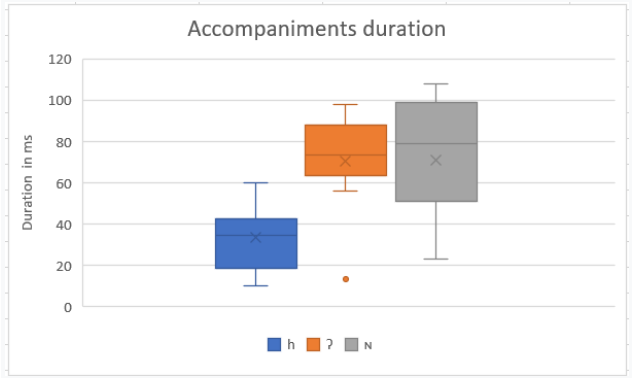


Figure 8. Duration of click accompaniments (aspirated, glottal and nasal) produced by 4 Hadza speakers (n=36).

3.4 Timing of the double closure

Figure 9 shows the audio waveform, intraoral (Po) and oral airflow (Oaf) of a dental click in the word [ǀaʔako]. There are 2 release transients that can be observed (1 & 2 in the Figure) on the waveform. The first is transducing the front release and the second the posterior (uvular) release. Aerodynamic data confirm this point. Indeed, Po curve shows that the maximum is reached after the 1st transient at the moment of the posterior release which means that there is, in this case, a short interval of time between the front and the posterior releases. Oaf become negative after the front release showing an ingressive airflow. The Oaf becomes positive after the posterior closure release.

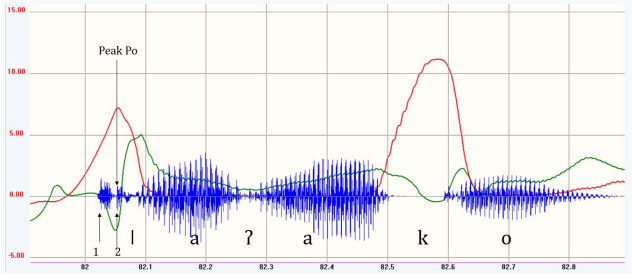


Figure 9. Audio waveform, Po in hPa (red plot) and Oaf in dcm³/s (green plot) of the word [ǀaʔako] ‘a large tree’. Peak Po is indicated by the arrow falling to the audio waveform.

1 and 2 indicate the moment of the front and back releases of the dental click [ǀ].

3.5 spectral features

Figure 8 shows average values taken at the peak of the transient releases for all clicks. The peaks of maximum energy were measured in a 10 ms Hamming window. This shows that the dental and alveolar clicks are clearly distinguished as acute and grave. The lateral click has a frequency band of similar intensity between 4 and 5 kHz on average.

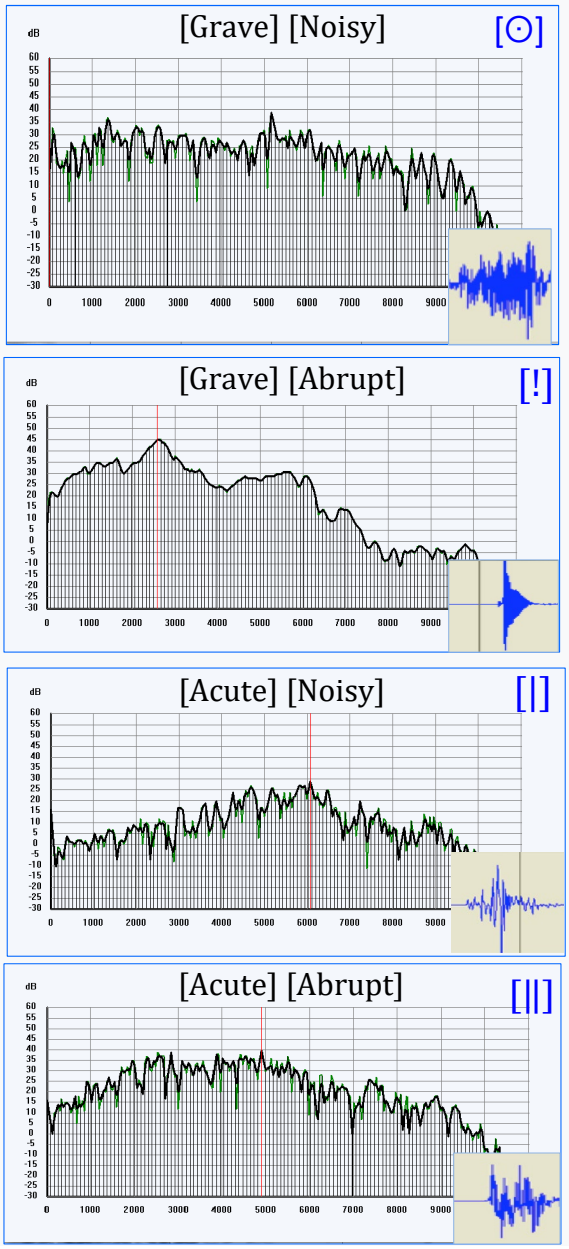


Figure 10. FFT spectra accounting the spectral characterization of the Hadza clicks.

4. DISCUSSION

The audio waveform observation of the 2 transients in clicks allow making inferences about the

movement and speed of the articulators involved. The anterior articulators (apex and tongue blade) move at different speed with the clicks. There is a difference between the apical clicks [|, !]. The alveolar click shows a much shorter initial transient (2 ms) when compared to the dental click which takes on average 12 ms before the peak of the noise. The lateral click [ǁ] where the tongue blade extends backwards during production has a sharp release and lasts for an average of 23 ms. The release of this click which is abrupt is also different from the other clicks that have a gradual extinction noise. This movement of the tongue dorsum can vary in different click realizations. Figure 9 gives a good example of the time difference that can occur between the font and the back closure of the articulatory click components. These sequences appear randomly with the dental click in our data. In a way the release of the lateral click is also made in a sequence as the initial lateral tongue release is followed about 20 ms later by a second sharp release accounting for the complete and abrupt opening of the tongue closure (See audio waveform at Figures 4 & 5).

The role of the posterior cavity is also important to consider in click releases. The short rising resonances - the partials - [between 1 and 2 kHz] following the posterior release (Figure 2 and 3), results from the lowering and retraction of the tongue dorsum reducing the volume of the pharyngeal cavity. These short resonances are visible after the 1st transient of the glottal and aspirated clicks of Figures 2 and 3. Although this has been observed with all accompaniments, but the bilabial, this occurs more frequently with the glottal.

Feature interpretation for clicks has to integrate a time factor. All the segmental cues discussed require to account very precisely for the relative coordination of articulatory movements. In addition to these tongue movements, the adjustments of the larynx and the velum movements which regulate and control nasal flow have to be taken into consideration. Regarding the nasal clicks [ᴺ|, ᴺ|ʔ, ᴺ!, ᴺ!ʔ, ᴺǁʰ, ᴺǁʔ], there is still a substantial amount of work to be done. First it is clear that nasalization spreads over the whole click duration. This makes that these clicks fully nasal and voiced most of the time. However, there are regular occurrences of clicks starting with a voiceless or very weak nasal component. These occurrences seem to occur randomly but this will have to be checked in future data analysis.

As mentioned by [6], the distinctions between clicks are unencoded in the following context and must reside in the noise burst only. The spectral information contained in the click noise burst is delivered to the auditory system with such saliency (temporal and/or intensity) that accurate perception is ensured. This makes that acoustic features are essential in making a correct description of click features.

From an acoustic point of view, Traill [6] proposed that clicks can be described with 2 features [grave vs. acute] and [abrupt vs. noisy]. This makes that the dental [|] and alveolar [!] Hadza clicks are both [grave] but [|] is [noisy] and [!] is [abrupt]. The lateral click [ǁ] is [grave and acute]. This description is deduced from an examination of the audio waveforms and acoustic spectra taken at the clicks release. Figure 11 includes the bilabial click [ʘ] which is noisy with a grave spectrum and some high frequency emphasis due to the labio-dental component of the articulation. The lateral click [ǁ] has a short band of intense noise between 4 & 5 kHz. This is higher than the alveolar [!] and lower than the dental [|].

A click can also be described as a series of features or a superposition of gestures, for example, nasal, lateral and aspirated [ᵑǁ͡ʰ]. Figure 11 displays articulatory and acoustic features for Hadza clicks.

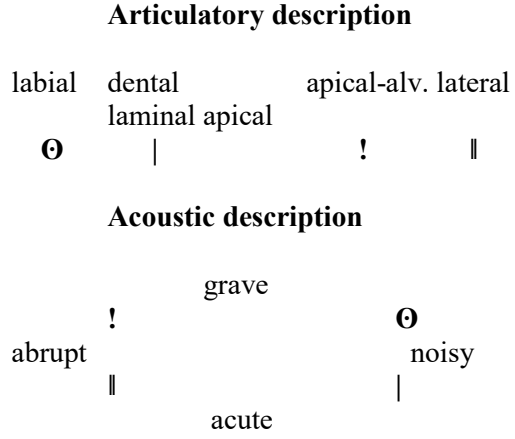


Figure 11. Articulatory and acoustic features for the characterization of Hadza clicks, following [6].

Finally, this contribution, coming from the partial analysis of a larger set of collected data, allows to make some observations about the relation between phonetics and phonology. The timing specifications and the time to produce some gestures are not directly related (one to one) with the phonological features. Hadza as most of the languages where clicks are found remain a challenge to give adequate answers to this question. This is why is so important to collect phonetic data in these languages.

REFERENCES

[1] Sands, B, Maddieson, I. & Ladefoged, P. 1996. The phonetic structures of Hadza. *Studies in African linguistics* (25), 2: 171-204.

[2] Sands, B. 2013. Phonetics and phonology of Hadza. In R. Vossen (ed). *The Khoesan languages*. London, Routledge: 38-42.

[3] Sands, B. 2020. Click consonants an introduction. In B. Sands (ed.) *Click consonants*. Leiden, Brill: 1-73.

[4] Miller, K. 2009. *Highlights of Hadza Fieldwork*. Ms. San Francisco.

[5] A. Ghio and B. Teston, "Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies," in *International Conference on Voice Physiology and Biomechanics,* 2004: 55–58.

[6] Traill, A. 1997. Linguistic phonetic features for clicks: Articulatory, acoustic and perceptual evidence. In R.K. Herbert (ed.) *African Linguistics at the Crossroads: Papers from Kwalusemi*. Köln, Rudiger Köppe: 97-117.

GEMINATION IN DOGRI AND PUNJABI: EVIDENCE OF DEVIATION AMONG 'LIKE' LANGUAGES

Pranav Badyal

Institute of Phonetics, Charles University
Pranav.Badyal@ff.cuni.cz

ABSTRACT

This study (i) investigates the timing properties of Dogri singleton and geminate voiceless stops in CVCV word form, and (ii) compares the findings with a related Indo-Aryan language, Western Punjabi.

LMER tests for part (i) confirm that the lengthened duration of the intervocalic consonant (C2) and contraction in the duration of the preceding vowel (V1) among geminates serve as the primary phonetic cues in differentiating Dogri geminate from a singleton.

The comparison of singleton – geminate distribution between Dogri and Punjabi in part (ii) reveals that the main cue of difference between the two languages lies in the flanking vowels. Whereas the contracted V1 duration appears promising in Dogri, Punjabi, instead, relies upon reducing the duration of V2. The study concludes that different treatment to maintain singleton – geminate contrast in the two genealogically related languages is due to the respective phonotactics that governs the distribution of duration across segments.

Keywords: gemination, double consonant, length, phonetics, speech production

1. INTRODUCTION

Dogri (ISO-639-3 dgo) and Punjabi are geographically and genealogically related languages that belong to the north-western sub-group of Modern Indo-Aryan (henceforth Indic) languages. In traditional Indian linguistics, a geographical belt comprising of related languages often leads to a situation, which is referred to as '*Kshetra'*, meaning, "linguistic area". Kshetra is a linguistic situation marked by plurality of languages in a given geographical area that show striking similarities [12].

The Kshetra distribution of languages in India has led to the problem of under-representation of minority languages, giving more proclivity and recognition to the non-minority languages. In case of Dogri, it is at times identified as a dialect of Punjabi, as pointed out by several authors [12, p. 271], [18], [14, p. 582].

In light of Kshetra, both Dogri and Punjabi have a similar syllable structure, and a similar phonological distribution of vowel length comprising of three short vowels /ɪ, ə, ʊ/ and seven long vowels /iː, eː, ɛː, ɑː, ɔː, oː, uː/ [8], [1], [4, p. 70], [11] and exhibit gemination.

In the context of gemination, earlier study on Dogri shows a lengthened duration of the word-medial consonant and contracted duration for the preceding vowel among geminates in CVCV singleton – geminate contrasting words [8]. Punjabi (Western, ISO-639-3 pnb), on the other hand, is reported to show no change in the duration of the preceding vowel but in the following vowel, the latter being contracted among geminates. The preceding vowel which is reported to show no durational change in a singleton – geminate pair places Western Punjabi in the category of languages like Estonian [6], Polish, Lebanese Arabic and Hungarian [9], and not with the Indic group.

Despite the fact that Dogri and Punjabi are genealogically as well as geographically related, yet both languages show different timing patterns for the pregeminate (preceding) vowel. Thus, it was deemed warranted to include Punjabi in contrast with Dogri.

2. METHODS

2.1 Participants

Participants included six speakers (all females) who reported Dogri as their L1. The language background of the participants was multilingual, in Dogri, Hindi and English. The educational background of all participants was at least Bachelor's or higher. All of them reported being residing in Jammu and Kashmir for at least 10 years. All participants reported having learned Dogri before Hindi and English. The age group ranged from 30 to 69 years. None of the participants reported a speech or hearing disorder.

2.2 Materials

A total of five minimal word pairs in singleton – geminate CVCV form were considered wherein,

- the word-initial consonant (C1) was either of the following: /t̪, b, s, k/,
- the preceding vowel (V1) was always /ə/,

- the word-medial consonant (C2) was either a singleton or a geminate: /p, t̪, ʈ, tʃ, k/, and
- the word-final vowel (V2) was always /ɑː/

These words were fitted in between a fixed carrier phrase (CP), ʊs _____ sʊneːjɑː, meaning *he/she* (third person) ____ *listen* (simple past).

2.3 Procedure

Recording sessions took place at Department of Dogri at University of Jammu, and All India Radio (AIR) studio, Jammu. All participants were shown the set of sentences in advance for a brief period.

A 20-inch monitor screen driven by a desktop PC computer was used to display sentences. An AudioTechnica ATR 1200 microphone was used for all recording purposes. The sentences were presented on a Microsoft PowerPoint slide, typed in Devanagari script in font size of 18. The sentences presented were manually controlled by the participants, by pressing the keyboard space-bar key in a self-paced fashion.

Each participant spoke each test-word five times except for VG who uttered 52 items. Thus, a total of 302 test-words were recorded. As far as the speaking rate is concerned, the participants were asked to read the material at a normal rate, in a relaxed and natural manner.

2.4 Acoustic segmentation

Broadband spectrograms and visual displays of their corresponding waveforms for each test-word embedded in a fixed phrase were generated and analysed using Praat software package [2]. After the manual inspection of waveforms and their corresponding spectrograms, the measurement criteria were set. An illustration of segmentation of a Dogri test-word is given in Fig.1.

C1: The onset for C1 was defined as the time interval from the appearance of frication noise of [s] and of start of the voicing in case of [b] in the spectrogram to the point where the amplitude for the following segment (the vowel) in the waveform would suddenly increase.

V1: The onset of V1 was defined at the point where higher formants of the vowel were observed. The end of V1 was defined as the point where the amplitude in the waveform would suddenly decrease or the F2 formant structures in the spectrogram would disappear.

C2: The closure duration of C2 was defined as the time interval from the offset of V1 until the point where a sudden release of energy would appear in the spectrogram.

V2: The onset for the word-final vowel was defined as the point from the offset of C2 until the point where the amplitude in the waveform would suddenly decrease, and the F2 formant structures in the spectrogram would disappear.

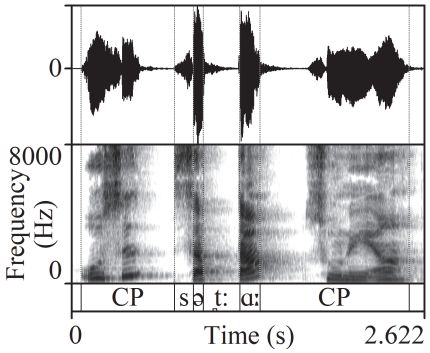


Figure 1: Acoustic waveform showing the segmentation of Dogri geminate test-word, sət̪ːɑː.

3. RESULTS

R [16] and lme4 [5] were used to perform a linear mixed effects analysis of the relationship between the phonemes (C1, V1, C2, V2) and consonant type (singleton/geminate). As fixed effects, consonant type (singleton/geminate) and place of articulation were entered (also with interaction term) into the model. As random effect, we had intercepts for subjects, as well as by-subject random slopes for the effect of consonant type. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The p value threshold was set at 0.05. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. The independent mean duration for four segments grouped by place of C2 articulation in singleton-geminate word pairs is presented in Table 1.

Concerning C1, it was anticipated that C1 duration would differ for singleton and geminates. It was found that consonant type did not affect C1 duration, ($\chi2(1) = 0.5136$, $p = 0.4736$). However, the place of articulation affected C1 duration significantly ($\chi2(2) = 107.31$, $p < 0.0001$) indicating that C1 duration varied across all five places of articulation. There was also no interaction effect to be found between consonant type and place, ($\chi2(2) = 0.4142$, $p = 0.8129$). Duration distribution of C1 among singleton and geminates across five places of C2 articulation is shown in Fig. 2.

Test-words having dental [t̪] and velar [k] as C1 were though segmented in Praat but these were not included in the analyses for estimating C1 duration because estimating onset boundaries of voiceless consonants accurately is not attainable.

Place	Consonant	Consonant type	C1	V1	C2	V2	Total word duration
Bilabial	pː	G	158 (41)	58 (11)	224 (32)	204 (42)	644
	p	S	155 (39)	73 (18)	125 (16)	211 (48)	564
Dental	t̪ː	G	150 (38)	67 (13)	238 (30)	213 (47)	668
	t̪	S	154 (35)	80 (22)	113 (18)	214 (45)	561
Palatal	tʃː	G	183 (56)	72 (10)	266 (40)	244 (49)	765
	tʃ	S	174 (61)	83 (15)	130 (28)	234 (56)	621
Retroflex	ʈː	G	97 (34)	71 (9)	213 (30)	216 (48)	597
	ʈ	S	103 (38)	97 (28)	115 (22)	218 (63)	533
Velar	kː	G	155 (32)	64 (14)	243 (24)	232 (56)	694
	k	S	156 (38)	85 (20)	117 (19)	214 (51)	572

Table 1: Mean durations (in ms) for C1, V1, C2, and V2 in singleton – geminate CVCV words across five places of articulation. The values in the parentheses represent standard deviation.

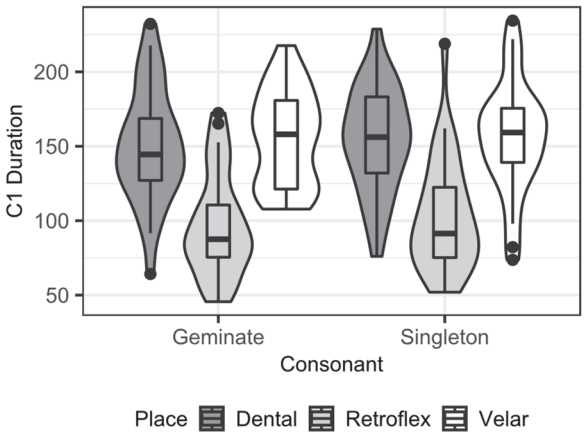


Figure 2: C1 duration (in ms) among Dogri singleton and geminates across all places of C2 articulation.

With regard to V1, there was significant effect of consonant type on the vowel, ($\chi2(1) = 4.9788$, $p = 0.02566$). Also, the place of articulation affected the V1 duration significantly ($\chi2(4) = 87.246$, $p < 0.0001$). However, no interaction effect could be established between place and consonant type on V1 duration, ($\chi2(4) = 9.4077$, $p = 0.05168$). Duration distribution of C1 among singleton and geminates across five places of C2 articulation is depicted in Fig. 3.

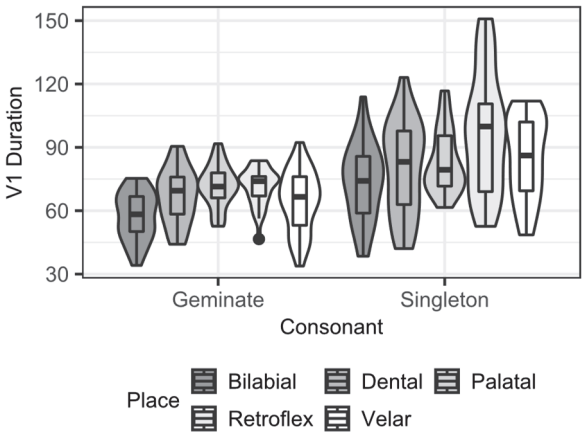


Figure 3: V1 duration (in ms) among Dogri singleton and geminates across all places of C2 articulation.

Regarding C2, consonant type affected C2 duration, ($\chi2(1) = 21.383$, $p < 0.0001$), lowering it by about 117.2 ms ± 8.6 (standard errors) at average in the case of singleton. The place of articulation affected C2 duration significantly, ($\chi2(4) = 78.903$, $p < 0.0001$), indicating that C2 duration varied at different places of articulation. There was also a significant effect between place and consonant type, ($\chi2(4)=23.348$, $p = 0.0001$). Duration distribution of C2 among singleton and geminates is shown in Fig. 4.

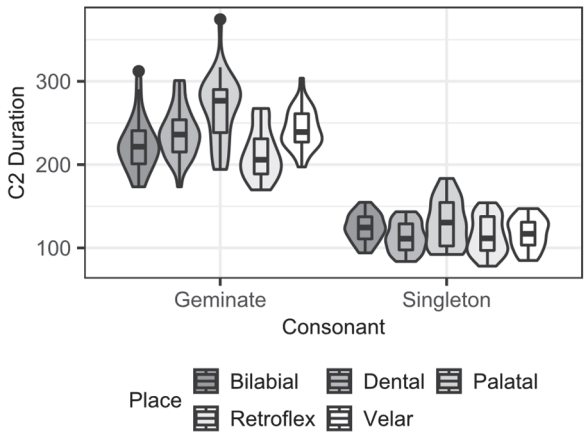


Figure 4: C2 duration (in ms) among Dogri singleton and geminates across all places of C2 articulation.

With respect to V2 duration, the results showed no significant effect of consonant type, ($\chi2(1) = 0.9101$, $p = 0.3401$). However, the place of articulation affected the V2 duration significantly, ($\chi2(4) = 66.372$, $p < 0.0001$), indicating that V2 duration differed across all places. As far as interaction effect between place and consonant type on V2 duration is contemplated, the results output remained insignificant, ($\chi2(4) = 6.4542$, $p = 0.1677$). Duration distribution of V2 is shown in Fig. 5.

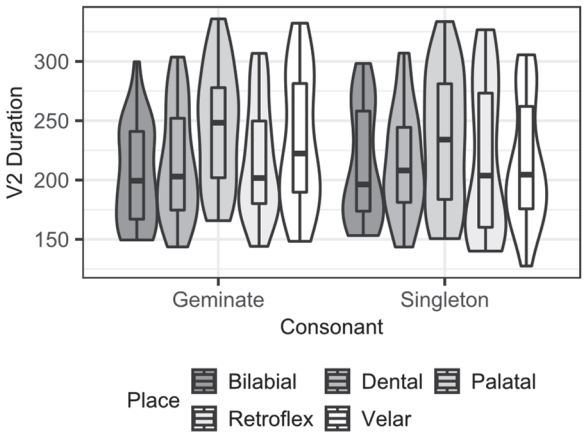


Figure 5: V2 duration (in ms) among Dogri singleton and geminates across all places of C2 articulation.

4. DISCUSSION AND CONCLUSIONS

This study aimed to appose linguistic differences between Dogri and Punjabi by gathering the congregated evidence in the existing literature, and by contributing new knowledge about the phonetics of Dogri and Punjabi from the view of gemination. It is found that the duration of the intervocalic consonant serves as the primary cue in distinguishing singleton and geminates in both Dogri and Punjabi (for all statistical values for Punjabi, cf. [10]).

The difference in C2 duration between singleton and geminates in Dogri at all five places of articulation, namely bilabial, dental, palatal, retroflex, and velar, are recorded to be 99, 125, 138, 98, and 126 ms respectively (cf. Table 1). On average, at five places of articulation, C2 is found to be 117 ms longer than their Dogri singleton counterparts. These mean durations suggest that Dogri voiceless geminate stops are almost double the duration in comparison to a singleton, thereby maintaining a ratio of 2:1. In contrast, Punjabi geminates are recorded to be on average 39 ms longer than their singleton counterparts at all five places of articulation, [$F(4, 16) = 13.359$; $p < .001$] [10].

The results presented for Dogri for V1 among singleton – geminate attest that V1 duration serves as the second primary cue in distinguishing voiceless geminate stops from their singleton counterpart at five places of articulation. On average, V1 is 17.2 ms shorter among geminates in comparison to singleton. These durational findings for the preceding vowel in Dogri geminates places Dogri in the category of languages that exhibit shortening of the vowel, as earlier attested in [8], also see [13, 19, 15] for a similar distribution among other Indic languages, and [7, 17] for other non-Indic languages. Interestingly, Punjabi, in versus, is not reported to show any durational difference in V1 by consonant type, [$F(1, 4) = 1.106$; $p = .352$], or by place of C2 articulation, [$F(4, 16) = 1.765$; $p = .185$].

Concerning the distribution of duration of the word-initial consonant in Dogri, no consonant lengthening is observed among geminates. Punjabi, in this respect, is also reported to show no significant difference in duration of the consonant in a singleton-geminate pair, [$F(1, 4) = 4.203$; $p = .110$]. This behaviour in speech is in opposition to the notion that the lengthening effects among geminates is due to mental anticipation, and to an articulatory effect that geminates are spoken with a "more energetic articulation" [20, p. 161].

The following vowel is found to be statistically similar in duration among Dogri singleton and geminates by consonant type (singleton/geminate) at all places of articulation. In contrast, Punjabi, interestingly, is reported to show shortened V2 duration among geminates in comparison to the singleton counterpart, [$F(1,4) = 170.844$; $p < .001$]. However, the effect of place on V2 duration was not found to be significant, [$F(4,16) = 1.873$; $p = .165$] .

From the discussion above, it is observed that besides a lengthened C2 among geminates in comparison to singletons, the V1 and the V2 serve as additional cues in the distinction of geminates from singleton in Dogri and Punjabi respectively. For geminates, Dogri maintains a shortened duration for V1 whereas Punjabi is reported to show a shortened duration for V2. It is however noted that though the distribution of duration among vowels in both languages is spread differently in a geminate, yet there is a similarity in their durational nature, i.e., there are contractions in the duration of the respective vowels (V1 for Dogri and V2 for Punjabi).

To speak briefly, there are though contractions observed in the respective vowels (V1 for Dogri, and V2 for Punjabi), but the location of the affected vowels is swapped for the two languages. The difference in the location of these contracted vowels for both languages could be attributed to separate phonotactics.

The study submits that geographically and genealogically related languages can be perceived very similar, yet phonetics provides evidence of deviation among the two.

In future research, it would be interesting to explore geminates in Eastern Punjabi (spoken in India) in parallel with Western Punjabi, to see if the former behaves akin to Western Punjabi or to Dogri. If the durational pattern for geminates in Eastern Punjabi does not correspond with that in Western Punjabi, it would suggest towards a possibility of language change on the part of Western Punjabi.

5. REFERENCES

[1] Bhatia, T. K. 1993. *Punjabi: A cognitive descriptive grammar*. London and New York: Routledge. ISBN-10: 0415003202. ISBN-13: 978-0415003209.

[2] Boersma, Paul., David, Weenink. 1992–2022. Praat: Doing phonetics by computer [computer program]. Version 6.1.26. retrieved 1st January 2023. www.praat.org.

[3] Dhillon, R. 2007. Stress in Punjabi. *Annual Meeting of the Berkeley Linguistics Society*, 33(1), 84–95. doi: http://dx.doi.org/10.3765/bls.v33i1.3519.

[4] Dhillon, R. 2010. *Stress and Tone in Indo-Aryan Languages*. Ph. D Dissertation. ProQuest LLC.

[5] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.*, 67, 1–48. doi: 10.18637/jss.v067.i01.

[6] Engstrand, O. & Krull, D. 1994. Durational Correlates of Quantity in Swedish, Finnish and Estonian: Cross-Language Evidence for a Theory of Adaptive Dispersion. *Phonetica*, 51(1-3), 80–91. doi: 10.1159/000261960.

[7] Esposito, A., Di Benedetto, M. G. (1999). Acoustical and perceptual study of gemination in Italian stops. *J. Acoust. Soc. Am.*, 106(4), 2051. doi: 10.1121/1.428056.

[8] Ghai, Ved Kumari. 1991. *Studies in Phonetics and Phonology: With special reference to Dogri*. Ariana Pub.

[9] Ham, William. 2001. *Phonetic and Phonological Aspects of Geminate Timing.* New York: Routledge. doi: 10.4324/9781315023755.

[10] Hussain, Q. 2015. Temporal characteristics of Punjabi word-medial singletons and geminates. *J. Acoust. Soc. Am.*, 138(4), EL388. doi: 10.1121/1.4931832.

[11] Hussain, Q. Proctor, M. Harvey, M., Demuth, K. 2020. Punjabi (Lyallpuri variety). *Journal of the International Phonetic Association*, 50(2), 282-297. doi:10.1017/S0025100319000021.

[12] Khubchandani, L. M. 1991. India as a sociolinguistic area. *Language Sciences*, 13(2), 265–288. doi: 10.1016/0388-0001(91)90018-V.

[13] Lahiri, A., Hankamer, G. 1988. The timing of geminate consonants. *Journal of Phonetics* (16). 327–338.

[14] Masica, C. P. 1993. *The Indo-Aryan Languages (Cambridge Language Surveys)*. Cambridge University Press. ISBN 10: 0521299446. ISBN 13: 9780521299442.

[15] Ohala, M. 2007. Experimental methods in the study of Hindi geminate consonants. In: Sole, M.J., Beddor, P., Ohala M. (eds.). *Experimental Approaches to Phonology*, Oxford University Press, 351–368.

[16] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

[17] Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. Journal of the International Phonetic Association, 37(2), 119–142.

[18] Shackle, Christopher. 2003. Panjabi. In: Cardona, G., Jain, D. (eds.), *The Indo-Aryan Languages*. New York: Routledge.

[19] Shrotriya, N. Verma, R. Gupta, S.K., Agrawal, S. 1996. Durational characteristics of Hindi consonant clusters. *Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP '96, 4, 2427-2430 vol.4. doi:10.1109/ICSLP.1996.607299.

[20] Trubetzkoy, N. S. 1969. *Grunzuge der Phonologie*. [principles of phonology]. Gottingen. Vandenhoeck & Ruprecht).

VARIATION AND PALATALISATION IN THE PRODUCTION OF THE PLURAL PREFIX IN MERU: A STUDY OF THREE DIALECTS GUE

Conceição Cunha[1], Fridah Kanana[1,2], Jonathan Harrington[1]

[1]Institute for Phonetics and Speech Processing (IPS), LMU Munich, Germany, [2]Kenyatta University, Kenya
cunha | fridah.kanana | jmh @phonetik.uni-muenchen.de

ABSTRACT

The proto-Bantu plural prefix marker *bi- for class 8 nouns/ has evolved into many forms across the Meru dialect continuum which belongs to the Central Kenya Group. However, the nature and extent of this variation is poorly understood. To shed light on this issue, five pairs of singular-plural class 7/8 nouns were recorded from 75 speakers across three dialects including the northern and southern dialects Tigania and Chuka as well as Imenti which is considered a standard variety of Meru bordering Tigania to the south. The prefixes were found to be palatalised labial in Imenti and varied between alveolar and alveolo-palatal in Chuka. For Tigania, there was a mixture of all of these both between and within speakers and whose choice was also influenced by the dialect of the experimenter. The observed variability nevertheless conforms to well-established patterns of sound change that connect palatalised labials, palatalised dorsals, and apical consonants.

Keywords: Bantu, Meru dialects, sound change, palatalisation, morphophonological variation.

1. INTRODUCTION

The present study forms part of a larger project to model morpho-phonological variation in dialects and dialect continua that happen to differ in one or more regular sound changes. The specific focus in this paper is on the variation in the plural prefix attachment to stems of class 8 nouns in the dialects of the Meru language which is of Bantu origin and spoken on the north eastern slope of Mount Kenya.

1.1 Background to Meru

For a very long time, the area where the Meru dialects are spoken belonged to one geo-political district called "Meru district" which was subdivided in 1992 into three districts namely, Meru, Nyambene and Tharaka Nithi. Later these districts were further sub-divided and amalgamated again in 2010 when Kenya introduced a devolved system of governance into 47 counties. The old "Meru district" currently belongs to two counties of Meru (County no.12) and Tharaka Nithi (County number 13). Chuka where data for this paper was collected from belongs to Tharaka Nithi county, and Tigania and Imenti belong to the Meru County. According to [1], before the colonial era, "Meru" comprised five (Igembe, Tigania, Imenti, Miutini and Igoji) of the nine sections of the so-called Meru dialects. The British administrators later included Tharaka in the adjoining eastern plains, as well as Mwimbi, Muthambi and Chuka that border Meru to the south (Fig. 1). The classification of these groups as belonging to the Meru dialects has been controversial.

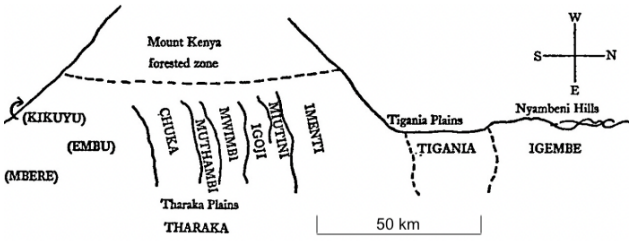


Figure. 1. The Meru district and its dialects/languages. Adapted from Fadiman (1973).

Some have suggested that Tharaka and Meru are two distinct languages [1-5]. Chuka has also been argued to belong to the separate Gikuyu language [4,6]. Despite these differences, there is some agreement that these varieties are mutually intelligible and form part of a dialect continuum (2,5,7,8-11]. Imenti is the most dominant dialect and sometimes considered to be the closest to a standard accent of Meru: it has the most developed literature and it is also used in the formative years of schooling (grades 1-3).

In common with many Bantu languages [12], the Meru dialects [8-11] have a seven vowel system and vowel length is contrastive. Voiced stops only occur in clusters following nasals (and may be prenasalised) and are lenited to voiced continuants intervocalically where they contrast with voiceless stops. There are nasal consonants at four places of articulation. Meru has a basic CV structure and may have lexical tone (although there are no studies to confirm this). As in most Bantu languages, singular and plural prefixes are paired in Meru according to the division of nouns into several noun classes [13, 14].

1.2 Morpho-phonological variation

A morpho-phonological and lexical analysis of six dialects of the Meru-Tharaka group was carried out

by Kanana ([9-11] who found one of the greatest areas of divergence between them was in the plural formation of class 8 nouns. Whereas the plural prefix originally derived from a proto-Bantu bilabial stop /*bi/ was typically voiced palatalised labial in Imenti, in Chuka it was found to be voiceless and had a primary non-labial, lingual constriction. These studies were however limited: there were no acoustic recordings and the data were based on typically just 1-2 participants per dialect. The present study extends this analysis to a larger pool of speakers and to the so-far unanalysed region of Tigania. The main reason for analysing Tigania was to determine whether plural formation is influenced to a greater extent by Imenti with which it shares its southern border compared with the geographically more distant Chuka. Another reason for analysing these dialects is that such differences in the plural prefix are relevant for understanding palatalisation as a sound change in the rare case of when palatalised labials become palatal, post-alveolar, or alveolar [15-17].

2. METHOD

The participants included 75 multilingual adult Meru speakers (median age of 36 years and age range 21-89 years) of three dialects: Chuka (n = 26, 14F), Imenti (n = 23, 6F), and Tigania (n = 26, 9F). The participants were screened as local dialect speakers of these regions. The participants provided informed written consent and were compensated for their participation. The participants spoke English and Swahili and none had any known language, speech or hearing problems.

The recordings took place in the participants' villages of residence. In each location, two participants were simultaneously recorded in two different rooms with limited background noise. The recordings were conducted by a native speaker of Imenti (the second author of this paper) in one of the rooms and by a native speaker of the local dialect as well as by a trained phonetician (the first author of this paper) in the other. The experiment was an isolated word production task consisting of a randomized order of 2-3 repetitions of 96 words, with one word at a time presented on a computer monitor in both English and in Swahili. The plural and singular were presented together on the monitor. The task was to produce the equivalents in the local dialect. Each presented word was formed from a morphological prefix and a stem. Words were repeated if the participants gave another equivalent (e.g., 'woman' or 'small girl' for the intended word 'girl'). The speech was recorded with a Beyerdynamic TG H54c head-mounted microphone at 44.1 kHz onto a Tascam US-2x2 interface connected to a laptop using SpeechRecorder 3.12.0 [18].

The present study is focused on a subset of these materials in order to analyse the variation in the place of articulation of a plural prefix combined with five different class 8 noun stems: /ara/ ('finger'), /eβa/ ('sorrow'), /imba/ ('corpse'), /oŋgo/ ('head') /ora/ ('frog'). A total of 979 words was analysed derived from these 5 plural prefix-stem combinations × 75 speakers with an average of 2.6 repetitions per word.

The speech signal was forced-aligned with WebMAUS [19]. The training model used for this purpose was German because it was found to give the most accurate segmentation. The output was structured into a speech database using EMU-SDMS [20]. Segment boundaries were hand-corrected independently by a trained research assistant and the first author of the paper. The analysis was based on an auditory transcription of the plural prefix supported when relevant by the visual information in the waveform and spectrogram. Cases of transcription difficulty were discussed amongst all three authors.

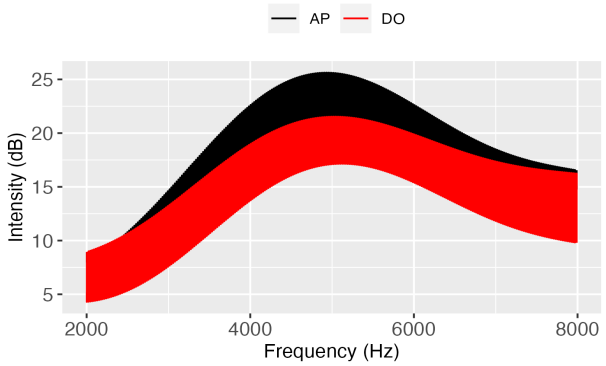


Figure. 2. Superimposed DCT-smoothed spectra extracted at the temporal midpoint of frication showing all 69 /tʃ, ʃ/ (AP: black) and all 275 /ʨ, ʨ/ (DO: red) spectra from Chuka and Tigania together.

The auditory distinction between the post-alveolar sibilants /tʃ, ʃ/ and alveolo-palatal /ʨ, ʥ/ was amongst the most difficult. Acoustically, /tʃ, ʃ/ fricatives can be expected to have a slightly more curved spectrum than those produced with a palatal constriction because they also have a steeper amplitude rise to a peak between 2-8 kHz [21, 22]: some acoustic evidence that the plural prefixes identified as /tʃ, ʃ/ spectra were more curved/had a higher amplitude peak that those identified as /ʨ, ʥ/ is shown in Fig. 2.

3. RESULTS

Since this was exploratory, non-hypothesis driven research, the results are presented without any statistical tests.

3.1 Dialect variation in the plural prefix

In Fig.3, the choice of place of articulation in the plural prefix was found to vary both within and between each dialect. As far as the latter is concerned, Fig. 3 shows that there is a division between Imenti whose plural prefix is a palatalised labial as opposed to the other two dialects that have a lingual prefix varying in place of articulation between dental and palato-alveolar.

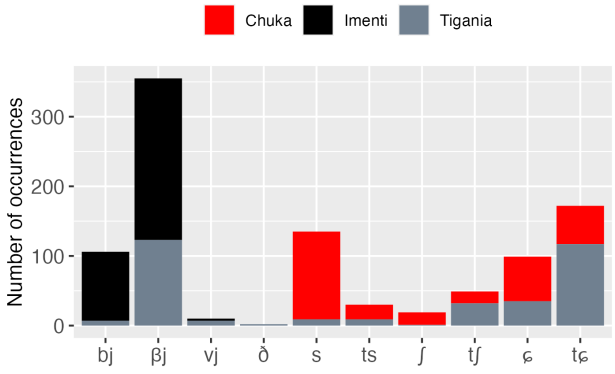


Figure. 3. The distribution of phonetic variants for marking plurals in class 8 nouns in three Meru dialects.

A further analysis was undertaken to determine whether there were any factors that might cause the variation between labial and non-labial prefixes in the Tigania plural suffixes. There was no influence of age nor speaker sex. However, as Fig. 4 shows, the choice of the plural suffix in Tigania was influenced by the interviewer: proportionally more speakers used a labial when the interviewer's dialect background was Imenti, but a (non-labial) lingual consonant when the interviewer was a local from the same region Tigania.

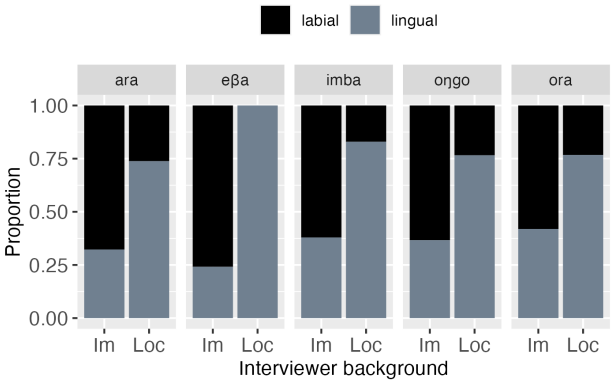


Figure. 4. The proportion of phonetic variants in Tigania by stem produced with a labial or lingual prefix when the experimenter was of an Imenti (Im) or Tigania (Loc) background.

Fig. 5 shows that this place difference in producing the plural prefix was not entirely caused by between-speaker variation: thus, although there were some speakers who always produced the prefix with a labial (black) and some who always produced it with a lingual (grey), there was also a minority (red) of speakers that for the same interviewer (and in the same session) produced the plural prefix sometimes with a labial and sometimes with a lingual consonant.

3.2 Differences between Chuka and Tigania

The focus here is on whether there were any place differences between Chuka and Tigania when plurals were produced with a lingual (non-labial) place of articulation. Fig. 3 had suggested a preference in Tigania for dorsal consonants /ɕ, tɕ/ as opposed to consonants produced with a primary apical/laminal stricture /s, ts, ʃ, tʃ/ in Chuka. Fig. 6 shows this by-dialect place preference for all five stems. Fig. 7 shows that while most speakers produced the prefix either only with an apical (black) or only with a dorsal (grey) there was also a minority of speakers (red) that variably produced the plural prefix either with an apical or with a dorsal consonant.

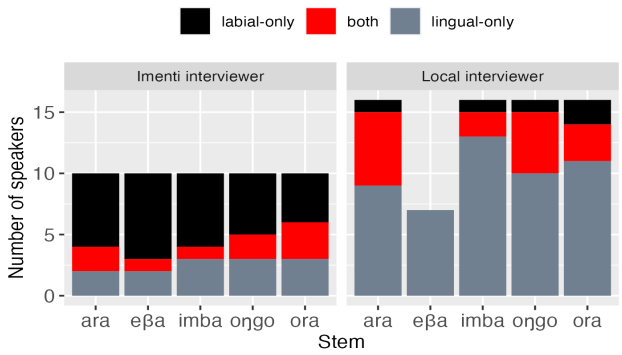


Figure. 5. The number of Tigania speakers by interviewer and stem who produced either only labial or only lingual or both labial and lingual plural prefixes.

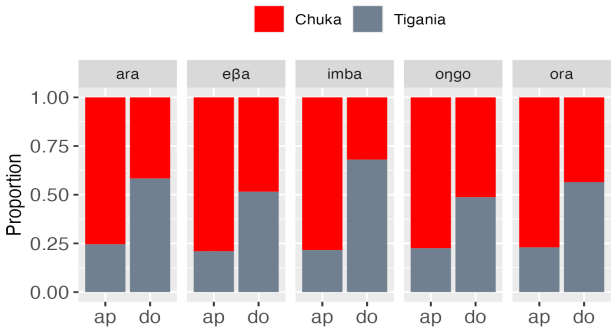


Figure. 6. The proportion of variants produced with an apical/laminal (ap) or dorsal (do) consonantal prefix in Chuka and Tigania.

Thus, together with Fig. 5, the data in Fig. 7 shows that while most of the variation in the place of articulation of plural prefixes is due to between-speaker differences, there is also a certain degree of

within-speaker variation in the selection of place of articulation in producing the plural suffix.

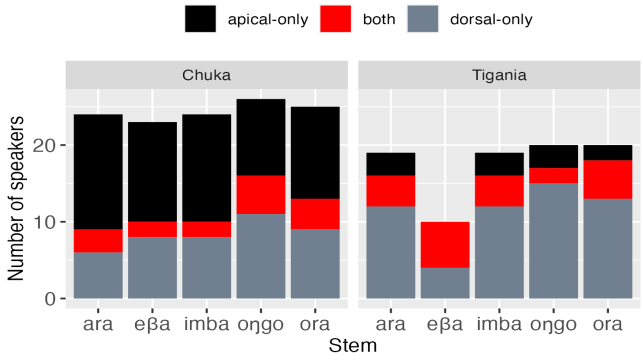


Figure. 7. The number of speakers in Chuka (left) and Tigania (right) by stem who produced either only an apical or only a dorsal or both apical and dorsal plural prefixes.

4. DISCUSSION

The study of palatalisation of the plural prefix has shown a wide range of place of articulation variation across three dialects of Meru that are located within some 100 km of each other. Consistently with earlier findings [8-11], whereas the plural prefix in Imenti was found to be a (palatalised) labial, in Chuka it was variably produced at alveolar, post-alveolar, and aveolo-palatal places of articulation. The new finding is that Tigania has characteristics of both regions: that is, the Tigania plural suffix can be labial as in Imenti, but also span the same range of lingual places of articulation as in Chuka. A further investigation showed that apical productions were more likely in Chuka than in Tigania.

The greater influence of Imenti on Tigania than on Chuka is likely to come about both because Imenti and Tigania are geographically closer and because Meru Town (which has the greatest population of the Imenti speaking region) is an administrative centre for Tigania. There is also frequent population movement between Imenti and Tigania due to trade in farming. Prior to the completion in the last 20 years of the tarmacked 100 km Mati road linking Meru and Chuka, travel between these regions was slow and sometimes (as in the rainy season) impossible [23]. Moreover, the Chuka region has its own administrative region in Kathwana to the east of Chuka town. For all these reasons, and consistently with much other research, communication density and population interaction [24-26] are likely to be major factors that explain the closer similarity in the phonetics of the Imenti and Tigania plural prefixes.

The variation in the place of articulation of the plural prefix was found to be both between and within speakers. This result suggests that this variation is cognitively represented especially so for the Tigania speakers for which the variation was skewed by the investigator's background. There is considerable evidence that speakers adapt their style to their interlocutors [27]. For the present study, we suggest that style-shifting may have formed a central part of the Tigania speakers' greater use of the labial plural prefix (that was found to typify the Imenti dialect) when the experiment was conducted by the Imenti investigator. When the same investigator conducted the experiment under the same conditions with the Chuka speakers, no such shift towards a labial place of articulation was observed. The lack of adaptation to the investigator in Chuka may be because Chuka only has lingual plural prefixes (Fig. 3) whereas some Tigania speakers produced a labial prefix even when the investigator was of a local Tiganian background (Fig. 5, right panel).

Based on the evidence that the plural prefixes are derived from reconstructed proto-Bantu /*bɪ/ [2], then Imenti with its palatalised labial prefixes is the most conservative of the dialects analysed in this study while Chuka and Tigania have introduced innovations that may have derived from the typologically rare sound change [28,12] by which labials undergo full palatalisation [29, 30]. Following Ohala [16], labial palatalisation typically involves a change from a labial to an alveolar place of articulation (e.g., standard Czech /pjɛt/, Bohemian /tɛt/, [31]) brought about by the acoustic similarity between palatalised labials and dentals/alveolars, and also because the labial burst is typically perceptually weaker than the high F2 transition that cues an alveolar place of articulation. Hock [28] questions the applicability of Ohala's acoustic idea to cases of labial palatalisation in Romance languages, while Bateman's analysis [32] of Moldavian suggests instead an articulatory explanation by which the glide hardens and then the labial is lost leading to the diachronic progression /bji → bʲji → bgji → gji/. Chuka may have developed palatalised dorsals (analysed as alveolo-palatals in this study) from palatalised labials following this type of articulatory progression suggested by Bateman [29, 32] and subsequently introduced a further innovation by which the palatalised dorsals have been fronted resulting in alveolar and post-alveolar prefixes following the well-attested sound change of velar palatalisation [33, 34]. In this scenario, Tigania would be diachronically less advanced than Chuka, given that its plural prefix is often produced with labials and fewer apicals. Whether such differences from Imenti to Tigania to Chuka mirror the diachronic progression of palatalised (labials → dorsals → apicals) requires further analysis also of the five or so dialects that intervene geographically between Imenti and Chuka.

5. ACKNOLEGMENTS

The research was funded by the project SoundAct which has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101053194).

6. REFERENCES

[1] Fadiman J. A. 1973. Early History of the Meru of Mt. Kenya. *Journal of African History*, XIV, I, 9–27.

[2] Guthrie, M. 1967-71. *Comparative Bantu: An introduction to the comparative linguistics and pre-history of Bantu languages*. 4 vols. Farnborough: Gregg.

[3] Whiteley, W.H. (eds) 1974. *Language in Kenya*. Nairobi: Oxford University Press

[4] Möhlig, W.J.G. 1980. Bantu languages. In: B. Heine and W.J.G. Möhlig (eds.), *Language and Dialect Atlas of Kenya*. Vol.1 Berlin: Dietrich Reimer, 11-58.

[5] Wamberia, K. 1993. *Kitharaka segmental morpho-phonology with special reference to the noun and the verb*. Unpublished Ph.D. thesis, University of Nairobi.

[6] Lambert, H.E. 1950. *The sytem of land tenure in the Kikuyu land unit*. Cape Town: School of African Studies, University of Cape Town.

[7] Heine, B., Möhlig, W.J.G. (1980). *Language and dialect atlas of Kenya*. Vol.1. Berlin: Dietrich Reimer.

[8] Kanana, F.E. 2010. *Lexical-phonological comparative analysis of selected dialects of the Meru-Tharaka Group,* unpublished Ph.D. thesis, Goethe University Frankfurt am Main.

[9] Kanana, F.E. 2014. *Lexical-phonological comparative analysis of selected dialects of the Meru-Tharaka Group*. Frankfurt: Peter Lang, edited by Rainer Vossen, doi: 10.3726/978-3-653-05161-2

[10] Kanana, F.E. 2011a. Dialect convergence and divergence: A case of Chuka and Imenti. *Selected Proceedings of the 40th Annual Conference on African Linguistics*, ed. Eyamba G. Bokamba et al., 190-205. Sommerville, MA: Cascadilla Proceedings Project.

[11] Kanana, F.E. 2011b. Meru dialects: The linguistic evidence. *Nordic Journal of African Studies* 20(4): 300–327.

[12] Maddieson, I. 2003. The sounds of the Bantu languages. In D. Nurse & G. Philippson (eds.), *The Bantu Languages,* 15–41. Routledge: London. 15–41.

[13] Guthrie, M. (1948). The Classification of Bantu Languages. Oxford University Press: Oxford.

[14] Herbert, R. (1991). Patterns in language change, acquisition and dissolultion: noun prefixes and concords in Bantu. Anthropological Ling., 33, 103-134.

[15] Bennett, W., Braver, A. 2015. The productivity of 'unnatural' labial palatalization in Xhosa. In M. Krämer & O. Urek (eds.), *Special Issue on Palatalization, Nordland, Vol. 42*. UIT The Arctic University of Norway.

[16] Ohala, J. 1978. Southern Bantu vs. the world: the case of palatalization of labials. *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*, 370–386.

[17] Recasens, D. 2020. *Phonetic Causes of Sound Change: the Palatalization and Assibilation of Obstruents*. Oxford University Press: Oxford.

[18] Draxler, C, Jänsch, K. 2004. SpeechRecorder – A universal platform independent multi-channel audio recording software, *Proc. of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 559–562.

[19] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services, *Computer Speech & Language*, vol. 45, pp. 326–347, doi: 10.1016/j.csl.2017.01.005.

[20] Winkelmann, R., Harrington, J., Jänsch, K. 2017 EMU-SDMS: Advanced speech database management and analysis in R, *Computer Speech & Language*, vol. 45, 392–410, doi: 10.1016/j.csl.2017.01.002.

[21] Jannedy S., Weirich, M. 2017. Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *Journal of the Acoustical Society of America,* 142, 395-405.

[22] Bukmaier, V., Harrington, J. 2016. The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *Journal of the International Phonetic Association*. 46, 311-329.

[23]https://www.businessdailyafrica.com/bd/corporate/shipping-logistics/new-road-cuts-transport-cost-in-meru-tharaka-nithi-counties-2286306

[24] Bloomfield, L.1933. *Language*. Holt: New York.

[25] Trudgill, P. 2008a. Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, 37, 241–254.

[26] Harrington, J., Kleber, F., Reubold, U., Schiel, F., Stevens, M. 2018. Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in Cognitive Science*, 1–21.

[27] Hay J, Jannedy S, Mendoza-Denton N. 1999. Oprah and /ay/: lexical frequency, referee design and style. *Proc. of the 14th international congress of phonetic sciences*, San Francisco

[28] Hock, H. 2006. *[+labial, +palatalized]. In R. Trask, J. Lakarra and J. Hualde (eds.) *Studies in Basque and Historical Linguistics in Memory of R. L. Trask.* Diputación Foral de Guipuzkoa: San Sebastian. 439–48.

[29] Bateman, N. On the typology of palatalization. *Language & Linguistic Compass*, 5, 588-602.

[30] Krämer, M., Urek, O. 2016. Perspectives on palatalization. *Glossa*, 1(1): 31. 1–17.

[31] Andersen, H. (1974). Abductive and deductive change. *Language*, 49, 765-793.

[32] Bateman, N. 2007). *A crosslinguistic investigation of palatalization*. Doctoral thesis, University of California, San Diego.

[33] Guion, S. (1998). The role of perception in the sound change of velar palatalization. *Phonetica*, 55, 18-52.

[34] Chang, S., Plauché, M., & Ohala, J. (2001). Markedness and consonant confusion asymmetries In E. Hume & K. Johnson (Eds.) *The Role of Speech Perception in Phonology*. Academic Press: San Diego, CA, p. 79–101.

THE QUANTITATIVE PHONETIC ANALYSIS OF LANGUAGES WITH SMALL ACCESSIBLE POPULATIONS

Márton Sóskuthy[1], Una Chow[2], Gloria Mellesmoen[3], Emily Sadlier-Brown[4]

[1–4]Department of Linguistics, University of British Columbia
[1]marton.soskuthy@ubc.ca, [2]una.chow@ubc.ca

ABSTRACT

Many of the world's languages – especially Indigenous languages – have small accessible populations. Such languages present challenges for quantitative phonetic analysis, which can in turn hamper the dissemination of results. This is due to phoneticians' overreliance on null hypothesis significance testing. Even data from a single speaker can provide valuable insights if a suitable approach is used. We present a case study that focuses on the VOT of ejectives in Gitksan based on a single speaker. We show how issues that stem from a small speaker pool can be remedied using a Bayesian approach that incorporates prior information from other languages. We also discuss further alternatives to null hypothesis significance testing such as foregrounding descriptive statistics and exploratory methods, and a flexible interpretation of what constitutes the statistical population for linguistic studies.

Keywords: Indigenous languages, small population, phonetic fieldwork, quantitative analysis

1. INTRODUCTION

According to the *Ethnologue* database [1], over 50% of the world's languages have fewer than 10,000 speakers, over 25% have fewer than 1,000 and over 12% have fewer than 100. In general, the smaller the population size of a language, the more difficult it is to access and recruit fluent speakers for phonetic study – though ease of access also depends on other factors such as geographical isolation. This paper uses the label *languages with small accessible populations* (LSAP) to refer to languages with a very small speaker pool and/or difficulties accessing speakers.

Indigenous languages are often LSAPs, including First Nations languages in British Columbia, Canada (the focus of some of our own work). In a 2018 survey representing 177 of 203 First Nations communities in BC (with a combined population of 137,653 individuals), only 3% of individuals spoke the language of their community fluently; among these speakers, 52% were aged 65+ and only 2.8% were under the age of 24 [2]. Government policies (including the residential school system) have disrupted language transmission in Indigenous communities, leading to low numbers of speakers. For example, there are fewer than 30 L1 speakers of Comox-Sliammon (Salish), and therefore a small population to recruit from. The representation of these languages in the literature is limited by the circumstances and total number of language users.

The phonetic study of Indigenous LSAPs is of utmost importance. It can feed into language teaching and revitalisation efforts, and it can enrich Indigenous peoples' experience with their language. These languages can also yield invaluable linguistic insights and broaden perspectives on patterns in speech, which are often heavily biased towards well-documented languages with large and easily accessible speaker populations [3,4].

However, the small size of the speaker samples that represent these languages constitutes a significant barrier to phonetic analysis. Researchers with restricted sample sizes cannot be sure how well their sample represents the language at large, which adds uncertainty to their conclusions. This is an issue regardless of the field of enquiry, but it is brought into especially stark focus in quantitative studies that rely on statistical analysis – a common approach in modern instrumental phonetics. As a result, phoneticians working on LSAPs often face significant challenges in data analysis, which in turn hampers dissemination of the results.

We argue that many of the difficulties surrounding work on LSAPs stem from a widespread overreliance on hypothesis testing using *p*-values (commonly referred to as *null hypothesis significance testing* or NHST). It has long been recognised that NHST is but a single tool in a much broader statistical toolkit [5]. In recent years, the literature on statistics in cognitive science has marked a move towards alternative approaches such as parameter estimation with uncertainty and exploratory data analysis [6,7]. We suggest that these tools allow phoneticians to better utilise the information in data from LSAPs. This paper serves two goals: (i) to provide an example of how such an approach can be implemented in practice and (ii) to convince analysts, reviewers and editors that quantitative studies of LSAPs are worthwhile.

In what follows, we first outline the main statistical problem in more detail. We then present a case study of voice onset time (VOT) in ejective stops in a single speaker of Gitksan, a Tsimshianic language spoken in

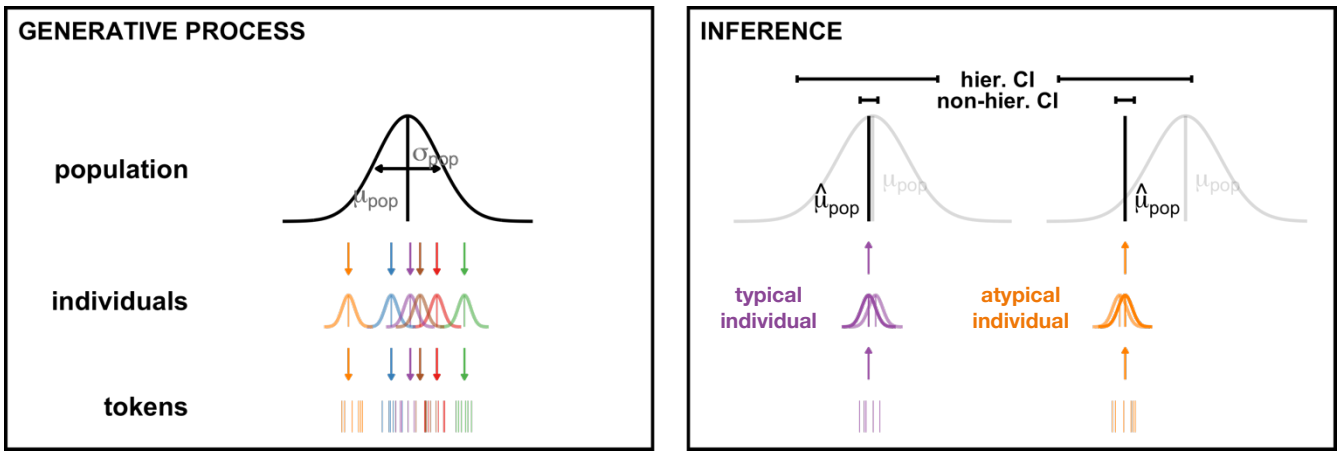


Figure 1: A graphical illustration of the main issue with statistical inference for LSAPs. The left-hand panel shows a schematic model of how phonetic data are generated by taking random individuals from a population and then sampling random tokens from those individuals. The right-hand panel shows what inferences can be made when phonetic data are only available from a single typical (left) or atypical (right) individual (hier. = "from a hierarchical model").

north-western British Columbia. We show that a Bayesian analysis utilising prior information from other languages can provide estimates of ejective VOT in Gitksan with a meaningful measure of uncertainty that allows for phonetic inferences. We also make suggestions as to what other tools could be used in the analysis of LSAPs and how reviewers and editors could approach evaluating work on these languages.

2. THE STATISTICAL CHALLENGE

Data from a limited number of speakers tend to strike analysts, reviewers and editors as insufficient. This occurs even when the small sample of speakers in a given study constitutes the entire extant speaker population. But why? Most of the time, we are interested in what is typical in a given language. In statistics, typicality is usually defined in terms of some central tendency (e.g. the mean) of the *statistical population* [8], which is often not the same as the accessible population. To give an example, in a medical study we might want to test a treatment that should work for all humans alive and yet to be born. This is our statistical population. However, it is theoretically impossible to test the treatment on humans who have not been born, and it is practically impossible to test it on all humans who are currently alive. Not all of the statistical population is accessible, and therefore the *statistical* and *accessible population* are not the same.

The same considerations arise in studies of LSAPs. The statistical population is often understood as a reasonably large set of speakers from the recent history of the language. Many of these speakers may not be accessible either because they are no longer alive or for practical reasons. Since typicality is defined in terms of the statistical population, the accessible population may or may not reveal typical patterns in the language.

Fig. 1 is a schematic illustration of the issue of typicality. The left-hand panel shows a common conceptualisation of how phonetic data are generated (and one that is implicitly assumed by anyone who uses hierarchical models, also known as mixed effects models [9]). Displayed horizontally is some phonetic quantity, such as VOT. The normal distribution at the top of the panel represents the statistical population, with the mean (μ_{pop}) and standard deviation (σ_{pop}) shown by the vertical line and horizontal arrows. Individuals sampled from this population are often fairly close to the population mean in terms of their phonetic behaviour (e.g. the purple and brown distributions in the middle), but sometimes they are quite far from it (e.g. the orange distribution on the left). What phoneticians observe, however, are not these individual-level distributions, but the actual tokens produced by the individuals, plotted at the bottom. These tokens show variation as well, and can sometimes deviate from the underlying individual-level patterns.

The right-hand panel shows two possible scenarios for statistical inference, one based on a single typical speaker and the other on a single atypical speaker. The analyst observes only the tokens plotted at the bottom. Based on these tokens, they estimate the individual-level distribution (darker shade), which may be different from the original distribution that generated the tokens (lighter shade). This is due to token-level variation and biases in sampling. Since only a single speaker is available in each scenario, the best estimate for the population-level mean is simply the mean of that individual. In the typical-speaker scenario on the

left, this estimate is close to the actual population-level mean, but in the atypical-speaker scenario, it is far from it. If there is only a single speaker or a handful of speakers in an LSAP study, the probability that the atypical scenario will arise is uncomfortably high. This is the reason for the intuition that a small speaker sample is not sufficient. Of course, the sample will necessarily be small in LSAPs due to the small size of the accessible population.

Quantitative researchers must accept some uncertainty in their estimates. In fact, a large part of statistical practice focuses on quantifying this uncertainty. LSAPs present two particular challenges in this context. First, the uncertainty of population-level values estimated from a small sample of speakers can be very large. This is illustrated by the wide confidence intervals (CI) at the top of the right-hand panel of Fig. 1, based on hierarchical models that account for both token-level and individual-level variation. As a result of this uncertainty, analyses of data from LSAPs using *p*-values as a binary criterion for robustness will rarely yield reliable effects. Second, when the speaker sample is extremely small (e.g. a single individual), it is impossible to reliably estimate variation across individuals and consequently impossible to come up with reliable estimates of uncertainty in population-level parameters. A common but mistaken reaction to this situation is to fall back on a non-hierarchical model, ignoring across-individual variation entirely. The estimates of uncertainty from such non-hierarchical models are unduly low, as illustrated by the second row of CIs in Fig. 1. While the CI from the hierarchical model is wide enough to include the true population mean even in the atypical-speaker scenario, the narrow non-hierarchical CI creates false confidence in a highly misleading estimate.

In section 3, we flesh out one possible alternative in the form of a Bayesian analysis. This approach avoids the issues above by (i) focusing the analyst's attention on graded measures of uncertainty instead of a simple binary decision criterion like the *p*-value; and (ii) approximating across-individual variation using external sources, which allows for the estimation of uncertainty around population-level parameters. In section 4, we suggest further solutions that question the assumptions made at the beginning of this section (e.g. what the statistical population is, or whether uncertainty must be quantified numerically).

3. CASE STUDY: GITKSAN EJECTIVE VOT

We analyse data from a single speaker of Gitksan to estimate the positive VOT associated with Gitksan ejective stops. The data and code for this study are available as part of the online supplementary materials at `osf.io/d7m34`. Isolated Gitksan words were elicited and recorded from a fluent L1 speaker (male, aged 65+) in a quiet room at the University of British Columbia. Word-initial prevocalic [t' c' k^w' q'] and their following vowels were labeled in Praat [10]. VOT values of 120 stop-tokens (4 places of articulation x 5 vowel contexts x 2 words x 3 repetitions) were calculated from the textgrids using a Praat script.

We aim to get a sense of what VOT values are plausible for this language, that is, to perform estimation with uncertainty. As noted above, a hierarchical model is needed to account for potential differences in typicality across speakers. However, there are difficulties in fitting such a model to this data, since it is not possible to estimate across-individual variation (corresponding to σ_{pop} in Fig. 1) from a single speaker. Our solution is to use a Bayesian hierarchical model (fitted with the `brms` package [11]) and exploit our ability to specify priors (see below). The model is essentially the same one that we would use for a larger speaker pool. The outcome variable is VOT, the only fixed effect is the intercept (corresponding to mean VOT) and there are random intercepts by speaker, as shown by the following `lme4`-style [12] formula:

(1) VOT ~ 1 + (1 | speaker).

Priors are distributions that embody prior knowledge or beliefs about parameters, and must be specified as part of a Bayesian analysis. The model returns a *posterior* distribution that results from updating the priors based on new evidence from the data. For the most part, we use *regularising* priors (details in online materials), which provide information about plausible values for parameters, but let the data determine the specific estimates. However, since across-speaker variation cannot be estimated from the data, this parameter is provided to our model as an *informative* prior obtained from three languages where by-speaker VOT figures for ejectives were available: Turkish Kabardian [13] (9 speakers), Cochabamba Quechua [14] (8 speakers) and Witsuwit'en [15] (11 speakers). It is a gamma distribution with a mean of 16.39 ms (shape = 24.53, scale = 0.67) estimated through a separate Bayesian model (details in online materials). This prior tells our model how much speakers tend to deviate from mean ejective VOT in other languages (but *not* what typical VOT values are). This then enables the construction of uncertainty estimates that allow for the possibility that our speaker is atypical.

The estimated population-level VOT for ejectives in Gitksan is summarised in the posterior distribution in Fig. 2. The mean of this distribution is 68 ms, which is the model's best guess for the population-level value, and is identical to the mean VOT for our single speaker (cf. Fig. 1 and the discussion in section 2). The rest of the distribution shows our uncertainty about

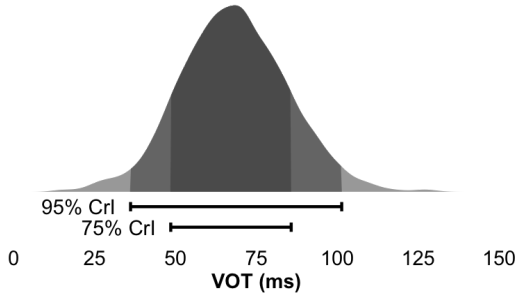


Figure 2: The posterior distribution of average VOT based on a Bayesian mixed model fitted to a speaker of Gitksan with across-speaker variation estimated from other languages. The whiskers show 95% and 75% credible intervals (corresponding to different levels of certainty).

this value. The 95% credible interval (CrI) is [36 ms, 101 ms]: we can be almost certain that the population-level value lies in this range. While null hypothesis significance testing usually limits itself to these two statistics, there is more information in the posterior. For instance, the 75% CrI is [49 ms, 86 ms]: we can be reasonably confident that ejective VOT in Gitksan is between these two values. Looking at the lower end of the distribution, it is extremely unlikely that the population mean is lower than 25 ms, placing it firmly in the long-lag VOT range. This contrasts with languages such as Witsuwit'en, where ejective VOT can be as low as 20 ms for some speakers [15]. The posterior can also be used as a prior for further studies of Gitksan ejective VOT with data from other speakers.

Using data from other languages to allow for hierarchical modelling is a compromise. However, it is better than (i) using a non-hierarchical model with unacceptably overconfident estimates (95% CrI of [62 ms, 75 ms]); or (ii) leaving the data unreported, which would lead to a loss of valuable information.

4. DISCUSSION

Sections 2 and 3 provided an exposition of the key statistical challenge of LSAPs and described a potential solution based on Bayesian modelling. Although estimating a single mean VOT value is simpler than most quantitative problems, the method generalises to more complex scenarios. The key is to be able to estimate across-speaker variation for relevant parameters from other languages. For instance, if the main research question is about differences in ejective VOT between two different places of articulation, one can look at how much this effect tends to vary in other languages.

As noted above, statistics provides a toolbox [5]. The method above is just a single tool. There are many other ways to approach the issue of small accessible populations. For instance, data description, exploration and visualisation are often under-utilised in phonetic studies. In some cases, the data are reduced to a single coefficient (or even just a *p*-value) from a regression model. We encourage analysts to put a heavier emphasis on communicating descriptive information in a format that is easy to digest (e.g. in the form of informative and well-designed graphs [16]). There are also many techniques for data exploration such as principal component analysis [17] and random forests [18] that provide alternative ways to summarise data. Providing a comprehensive but easily digestible summary can be more informative than running a hypothesis test that is doomed to fail or mislead due to a low sample size. While these descriptive methods do not offer measures of uncertainty, it is straightforward to verbally outline the challenges of parameter estimation with data from LSAPs.

Another tool that can be extremely useful in this context is the sharing of data and analysis code [19]. This facilitates both the evaluation of the analysis and helps future cumulative research efforts. We note that in our own attempt to find baseline figures on across-speaker variation in ejectives, we have not yet found a single publication providing raw data alongside the published results.

Finally, the statistical population for a given analysis is not predetermined by convention but depends on the goals and beneficiaries of the research. To give a concrete example, when descriptive work is carried out to facilitate language revitalisation, the community itself may prefer data from a small number of elders. In such a case, the statistical population may just be a single speaker, obviating the need for hierarchical modelling.

With these considerations in mind, how should a reviewer or editor approach the task of evaluating work on LSAPs? To be clear, we do not advocate for a blank cheque for this type of research. Quantitative work on LSAPs must be rigorous and must also show awareness of the dangers of a small speaker pool. The key to success for an analysis of LSAP data is to clearly identify research goals that can be attained with limited data; make the best use of the data through the judicious application of quantitative methods; and highlight any potential limitations of the conclusions. If these criteria are fulfilled, we see no reason to criticise work on LSAPs based solely on the number of speakers.

To summarise, NHST is overused in many scientific contexts, and is far too restrictive for work on LSAPs. Descriptive statistics, exploratory methods, visualisation and Bayesian modelling can all provide valuable alternatives. In addition, while all research requires careful thought about the scientific context and intended beneficiaries, this is even more important in work on LSAPs.

7. REFERENCES

[1] Eberhard, D. M., Simons, G. F., Fennig, C. D. 2022. *Ethnologue: Languages of the World (25th ed.)*. SIL International. http://www.ethnologue.com.

[2] Dunlop, B., Gessner, S., Herbert, T., Parker, A. 2018. *Report on the Status of B.C. First Nations Language (3rd ed.)*. First Peoples' Cultural Council.

[3] Evans, N., Levinson, S. C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32(5), 429–448.

[4] Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., Majid, A. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* 26(12), 1153–1170.

[5] Gigerenzer, G. 2004. Mindless statistics. *J. Socio.-Econ.* 33(5), 587–-606.

[6] Cumming, G. 2014. The new statistics why and how. *Psychol. Sci.* 25(1), 7–29.

[7] Kruschke, J. K., Liddell, T. M. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. B. Rev.* 25(1), 178–206.

[8] Yates, D. S., Moore, D. S., Starnes, D. S. 2003. *The Practice of Statistics (2nd ed.)*. Freeman.

[9] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59(4), 390–412.

[10] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/

[11] Bürkner, P. C. 2017. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01.

[12] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1), 1–48.

[13] Gordon, M., Applebaum, A. 2006. Phonetic structures of Turkish Kabardian. *J. Int. Phon. Assoc.* 36(2), 159–186.

[14] Gallagher, G., Whang, J. 2014. An acoustic study of trans-vocalic ejective pairs in Cochabamba Quechua. *J. Int. Phon. Assoc.* 44(2), 133–154.

[15] Hargus, S., 2011. *Witsuwit'en Grammar: Phonetics, Phonology, Morphology*. UBC Press.

[16] Healy, K. 2018. *Data Visualization: A Practical Introduction*. Princeton University Press.

[17] Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

[18] Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5–32.

[19] Foster, E. D., Deardorff, A. 2017. Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA* 105(2), 203.

EVALUATING FORMANT ESTIMATIONS AND DISCRETE COSINE TRANSFORM TO DIFFERENTIATE BETWEEN PHARYNGEAL FRICATIVES IN MEHWEB

Alexandre Arkhipov [a]*, Michael Daniel [a,b], Alexander Shiryaev [c], Ekaterina Shepel [d]

[a] Universität Hamburg, Germany; [b] Collegium de Lyon / Laboratoire Dynamique du Langage (Université de Lyon / CNRS), France; [c] Independent researcher, Singapore; [d] HSE University, Moscow
alexandre.arkhipov@uni-hamburg.de, misha.daniel@gmail.com, ashiryaev87@gmail.com, katerina11780@gmail.com

ABSTRACT

We investigate the laryngeal/pharyngeal part of the consonantal inventory of Mehweb Dargwa (East Caucasian), which has been shown to contrast plain laryngeals to pharyngeals, with additional contrasts due to pharyngealization feature.

We are using formant estimation on the fricative noise of laryngeal/pharyngeal fricatives to differentiate between contrasting segments and to approach their characterization in articulatory terms. Using formant analysis techniques is applicable here because the vocal tract configuration of pharyngeal and laryngeal segments is similar to that of vowels. While the formant structure of plain laryngeals accommodates to adjacent vowels, pharyngeals appear to have their own target formant values, which furthermore vary depending on the presence of pharyngealization.

We are comparing outcomes of our formant analysis to discrete cosine transform coefficients, which have been previously shown to differentiate well between sibilants.

Keywords: pharyngeal, pharyngealization, formants, DCT, East Caucasian.

1. MEHWEB LANGUAGE

1.1. Genealogical and sociolinguistic aspects

Mehweb (Glottocode: mege1234) is a one-village language of the Dargwa branch of the East Caucasian (Nakh-Daghestanian) family, sometimes considered to be a dialect of (Northern) Dargwa (Glottocode: darg1241, ISO 639-3: dar). It is spoken in the village of Mehweb (Russian: *Мегеб*) in central Daghestan at 1,800 m above sea level, geographically separated from other languages of the Dargwa branch and surrounded by speakers of Avar and Lak (belonging to other branches of East Caucasian).

As of 2019, the number of speakers is estimated to be between 800 and 900 [1], most of them living in Mehweb itself. “So far, there are no indications of language loss in Mehweb. All villagers speak Mehweb, and Mehweb is the first language acquired by children” [1: 2]. Mehweb has no written tradition; written languages used by Mehwebs are Russian and Avar. Most adults have a command of Russian, those born before the 1990s usually also speak Avar, and some of those born before the 1950s speak Lak.

Our starting point was the published description of Mehweb phonological system [2] based on fieldwork data from 2013–2016, briefly summarized in §1.2. A small sample of acoustic data from Mehweb, along with the data from several other East Caucasian languages, was used in [3], providing some phonetic observations presented in §1.3.

1.2. Mehweb phonology

The vowel system lists four vowel qualities: /i e a u/. All of them can be pharyngealized, whereby /uˤ/ is realized variably as [uˤ~oˤ].

Pharyngealization is claimed to be a ‘prosodic’ feature associated with a syllable (or root) rather than with a particular vowel or consonant; it may spread from lexically pharyngealized syllables (roots) to adjacent syllables. We follow [2] in marking the nucleus of the pharyngealized syllable with /ˤ/. Most pharyngealized syllables contain epiglottals [ʔ ʜ] or uvulars [q χ ʁ], but /aˤ/ is also attested in their absence (e.g. /laˤži/ ‘cheek’). Across lexicon, /aˤ/ is also the most frequent pharyngealized vowel, while front pharyngealized /iˤ eˤ/ are very rare.

The basic contrasting post-uvular segments are plain laryngeal (glottal) stop /ʔ/ and fricative /h/, and pharyngeal fricative /ħ/ (see examples in Table 1).

Sound	Phar.	Example	Gloss
ʔ	no	/muʔ/	‘back’ (noun)
h	no	/haruš/	‘fermented drink’
ħ	no	/ħaq’ur/	‘burdock’
[ʡ] /ʔ/	yes	/ʡaˤt’a/	‘frog’
[ʜ] /ħ/	yes	/ħaˤbal/	‘three’
ʡ	no	/ʡat’/	‘flour’ (Avar loan)
ʡ	no	/ʡarʁal/	‘long’

Table 1: Examples with laryngeals and pharyngeal/epiglottal sounds. “Phar.” marks the presence/absence of pharyngealization.

Two segments appearing in pharyngealized syllables, designated as epiglottal stop [ʔ] and fricative [ʜ], are considered to be pharyngealized allophones respectively of /ʔ/ and /ħ/ [2: 32–35]. However, in some Avar loans as well as in a few native roots a similar epiglottal stop appears without pharyngealization [2: 19, fn. 3]. While its properties require further study, this /ʔ/ will be provisionally treated here as a separate phoneme.

1.3. Phonetic properties of pharyngeals/epiglottals

Phonetic properties of pharyngeals/epiglottals are not discussed in [2]. Pharyngealization effect on vowels is described as 'centering' [2: 32] and symbolized as /iˤ eˤ aˤ uˤ/ yielding [eˤ ɛˤ æˤ uˤ~oˤ].

The study in [3: 1552] considers a limited sample of Mehweb data and characterizes [ʔ] in the context of pharyngealization as a 'weak epiglottal stop or approximant,' and the non-pharyngealized [ʔ] as 'a stronger epiglottal stop.' The difference between [ħ] and its pharyngealized counterpart [ʜ] is attributed to an 'increase in laryngeal constriction, possibly with larynx raising' in the latter. It is reported to have 'a flat spectrum between 1–3 kHz and peak at 3–3.5 kHz', while the spectrum of [ħ] has 'a deep valley above 1 kHz and peak at 2.5 kHz' (numbers for adult male speakers). Additionally, both fricatives or only the non-pharyngealized one, depending on the speaker, may optionally be strengthened through aryepiglottic trilling.

2. OBJECTIVES, METHODS AND DATA

2.1. Objectives

We follow [3] in adopting the Laryngeal Articulator Model approach [4, 5] and understand East Caucasian 'pharyngeal' and 'epiglottal' as essentially the same place of articulation, with further possible distinctions based on a combination of such parameters as the degree of aryepiglottic laryngeal constriction, larynx height (larynx raising being an expected synergetic accompaniment to laryngeal constriction) and aryepiglottic trilling. We thus use 'pharyngeal' as a cover term for both 'pharyngeal' and 'epiglottal' employed in [2].

Our main objective in the present paper is to obtain a more systematic acoustic description of the laryngeal and pharyngeal fricatives, which would allow to make informed hypotheses on the articulatory features involved (a proper articulatory study currently not being possible for extralinguistic reasons). To achieve this, we perform spectral analysis of the fricatives across different vocalic contexts and in different positions within a word.

2.2. Methods

Acoustic metrics used in previous work to differentiate between fricative categories in various languages include spectral moments [6] and, more recently, discrete cosine transformation, or DCT coefficients [7]. The application of spectral moments is problematic for various reasons, highlighted in [8]. Therefore, we only recur to them for comparison with previous work and as to a baseline for other metrics.

DCT has been shown to be efficient in discriminating sibilants in rich sibilant systems such as Polish [9] and German [10]. However, while DCT analysis is good at capturing spectral differences between categories, the coefficient values may not always be directly interpretable in articulatory terms. This is why we supplement DCT with the formant analysis of the fricative spectra, more informative in this respect.

Formants are characteristics of resonances of the vocal tract which determine the acoustic structure of vowels. Fricatives are produced with a noise source, hence their spectra are aperiodic. However, as shown in [11], spectral peaks in sibilants may show seamless transition into vowel formants. Spectral peaks in fricatives in the range of vowel formants F2–F5 are discussed in [12]. The length of the oral cavity in front of the main constriction has a strong influence on the spectral shape of fricatives. Spectra of anterior non-sibilants [f θ] are generally weak and have no sharp peaks since the front cavity is negligible and there is no significant obstacle downstream. Sibilants do present sharp peaks, but the lower frequency region corresponding to F1 has typically very little energy [11]. For more posterior consonants, the length of the resonance cavity increases; "[t]he more back fricatives, x, χ, ħ, have a spectral peak that decreases in frequency as the place of articulation approaches the glottis, and additional peaks in the higher part of the spectrum" [13]. Finally, it is (almost) the entire supraglottal tract which contributes to the production of laryngeals and pharyngeals. Consequently, their spectrum is most similar in structure to that of vowels.

Due to the aperiodicity of the spectrum, direct formant measurements in fricative noise are less reliable than in vowels. We use FastTrack [14], a plugin for Praat [15], to perform a set of automated measurements with different settings and select the optimal analysis for each individual sound based on the smoothness of the formant trajectories. Beyond calculations of spectra and formants, most other analyses and visualizations were done in R [16].

2.3. Data

The recordings were made in summer 2022 by the second author from four male and four female

speakers born between 1955 and 1971 and one male speaker born in 1994. All the older speakers formed married couples, and more generally were part of a tight social network within the speech community.

A list of target words was compiled to include laryngeal and pharyngeal stops and fricatives in word-initial, word-medial and word-final position in different vocalic contexts. Here, we limit our discussion to fricatives. A total of 134 stimuli were recorded, with the list of stimuli slightly different across speakers. The interviewer usually only uttered the Russian translation equivalent. Stimuli were presented in random order for each speaker, who were asked to produce the stimulus four times in isolation and once in a carrier phrase; the actual number of tokens varied across speakers and stimuli. Only isolated productions are considered in this paper.

The data were recorded with a Sennheiser HSP4-EW headset and Olympus DM-901 voice recorder as Linear PCM at 48 kHz/16 bit, manually segmented and labelled in Praat. Their annotation and analysis are currently in progress. In this paper, data from one female (*Mn*, 63 y.o.) and five male speakers (*Ab* (51), *An* (63 y.o.), *Kz* (67), *Mh* (28), *Mz* (59)) are used.

3. ACOUSTIC ANALYSIS

3.1. Preview

Generally, laryngeal [h] is markedly different from the two pharyngeals. It is shorter and less intensive; its mean duration in word-medial position was 44% to 73% of the mean duration of [ħ ʜ] across speakers. Therefore, the main focus was on distinguishing between the two pharyngeals.

Ensemble-averaged pre-emphasized spectra for the three target categories, plain laryngeal [h], pharyngeal [ħ] (labelled "ħ–") and its pharyngealized counterpart [ʜ] ("ħ+") in contexts before [a/aˤ] for speakers *An*, *Kz*, *Mn* are given in Fig. 2. Sharper peaks and valleys in [ħ] can be noticed compared to [ʜ].

3.2. Discrete cosine transformation

For DCT analysis, fricatives were resampled at 24 kHz and pre-emphasized by 6 dB/octave from 50 Hz. Series of 3 spectra at 10 ms step were extracted with 21.3 ms Hamming window around the temporal midpoint of the fricative, averaged and transformed to mel scale. Frequencies below 500 mel (414 Hz) were discarded to avoid possible effects of voicing. DCT coefficients (DCT0–DCT4) were computed with *emuR* package [17] for each averaged spectrum.

DCT coefficients behaved variably across speakers and contexts. DCT4 was the most stable and efficient in distinguishing between pharyngeals; it tended to be positive for [ħ] and negative for the pharyngealized [ʜ]. DCT3 tended to be positive for the laryngeal [h] and negative for both pharyngeals, so DCT4 and DCT3 combined performed well in distinguishing between the three fricative categories (see Fig. 3).

3.3. Spectral moments

The four spectral moments (center of gravity, standard deviation, skewness and kurtosis) were obtained on the spectra (in Hertz scale). Although laryngeal [h] was clearly separated from pharyngeals in many speaker x moment combinations, the results for the two pharyngeals were inconsistent and often contradictory across speakers. Specifically in the case of post-velars it is not surprising since the spectral moments do not capture the complex spectral shapes.

Speaker	Ab	An	Kz	Mh	Mn	Mz	Total
h	44	50	24	39	40	38	235
ħ ('ħ–')	82	108	81	80	91	87	529
ʜ ('ħ+')	70	77	55	59	62	81	404
Total	196	235	160	178	193	206	1168

Table 2: Number of tokens per speaker for three fricative categories. The pharyngeals [ħ] and [ʜ] are labelled in the data resp. as 'ħ–' and 'ħ+'.

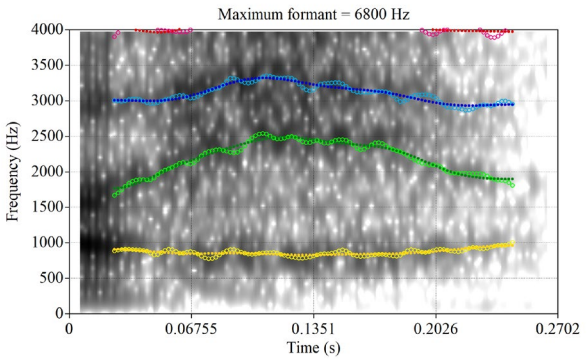


Figure 1: Sample image of formant analysis in noise by FastTrack. Speaker *Mn*, [ħ] in *tamaħ* 'consciousness'.

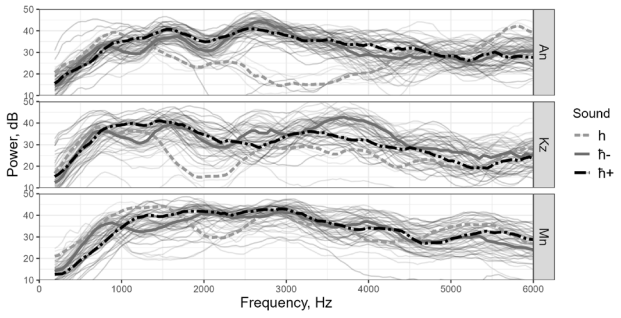


Figure 2: Averaged spectra for [h] (light grey), [ħ] (dark grey), [ʜ] (black) in contexts before [a/aˤ].

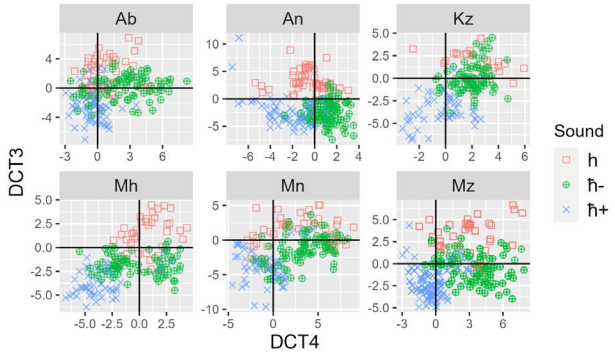


Figure 3: Fricatives in DCT3xDCT4 space.

3.4. Formant analysis

For this analysis, fricatives were submitted to the formant extraction algorithm of FastTrack. Most tokens were processed with the maximum frequency setting within 4000–6000 Hz for male speakers and 5200–6800 Hz for the female speaker. In some cases manual correction of the maximum frequency range was performed to force a better analysis.

The results confirm the high variability of the spectrum of [h]. Very often, formants are steady between [h] and the vowel, i.e. the formant structure of the laryngeal adapts completely to the vowel. In contrast, the two pharyngeals exhibit a far more stable formant configuration across vocalic contexts, its own for each sound and distinct from the adjacent vowels. The non-pharyngealized [ħ] exhibits has a high F1 and a relatively high F3 (up to ca. 3 kHz in male speakers). Its pharyngealized counterpart [ʜ] has much more tightly spaced formants, with an extremely high F1 and a low F3. It is worth noting that no formant configuration in pharyngeals is apparently continued into the vowel. Formant transitions thus deserve further study.

3.5. Comparison

We used interpretative machine learning (iML) to compare the efficiency of different types of metrics to distinguish between the two pharyngeal categories, [ħ] and [ʜ]. Three datasets were created, each including the 933 pharyngeal tokens, with Speaker and Gender as predictors. The competing dataset-specific predictors were spectral Moments (1 to 4), DCT coefficients (0 to 4), and Formants (F1 to F3, measured around the fricative midpoint). We also tested Formants datasets augmented with formant bandwidths (FormBW), formant frequencies at 0.3, 0.5 and 0.7 relative timepoints (FormDyn), and both (FormDynBW).

Three different ML algorithms were run on each dataset: logistic regression (LR), random forest (RF) and gradient boosting (GB). Each algorithm was trained 1,000 times on random 30% of the data and tested on the other 70%. F1-score was calculated each time as a quality metric (once for each of [ħ, ʜ]). The pipeline of the model was realized in Google Colabs using Scikit-learn package [18]. The median of the quality metric is reported in Table 3.

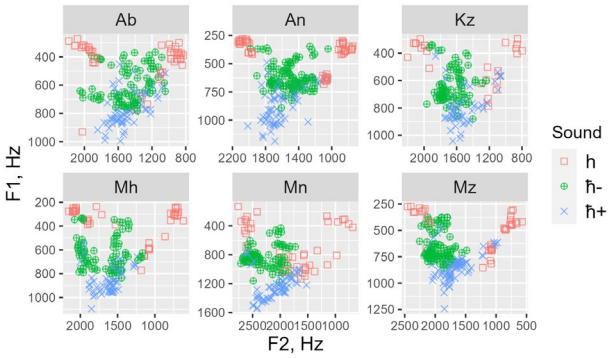


Figure 4: Fricatives in F2xF1 space.

	LR1	GB1	RF1	LR2	GB2	RF2
Moments	50.8	63.8	66.0	70.1	74.4	75.3
DCT	82.5	85.9	87.2	86.8	89.2	90.2
Formants	86.2	86.0	88.0	90.0	90.0	91.2
FormBW	87.6	89.0	90.3	91.0	92.1	93.0
FormDyn	89.8	91.6	92.8	92.7	93.8	94.6
FormDynBW	90.1	91.9	92.7	92.7	94.0	94.5

Table 3: Median F1-score (%) of models predicting the pharyngeal categories. LR1, GB1, RF1 is the score for [ʜ], LR2, GB2, RF2 the score for [ħ].

In all settings, RF achieved better results than GB and LR. The score for [ħ] was always better than for [ʜ], presumably because of the unbalanced token quantity. Results on Moments were by far the worst, starting at 50.8% for LR1. Formants yielded slightly better F1-score than DCT, while introducing formant dynamics gave a noticeable improvement (a stronger one than adding bandwidths).

4. DISCUSSION

The two contrasting pharyngeal categories have demonstrated consistent formant configurations distinct from each other and from that of vowels. Formant values allow for an articulatory interpretation; the high F1 values for [ħ ʜ] are consistent with a narrowing in the lower pharyngeal domain, and low F3 values have also been reported previously for pharyngealized vowels in East Caucasian [13; 19].

Directions for future work include the analysis of formant transitions, of voice quality in adjacent vowels, and of the pharyngeal and laryngeal stops.

5. ACKNOWLEDGEMENTS

*Contribution by this author was made in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

Segmentation of audio files was done by Anna Rozzorenova-Helle, Ekaterina Shepel and Georgiy Bulgakov. We are indebted to Yuliy Daniel for implementing the interpretative machine learning pipeline.

We are grateful to two anonymous reviewers for their valuable comments.

6. REFERENCES

[1] Dobrushina, N. 2019. The language and people of Mehweb. In: Daniel, M., Dobrushina, N. & Ganenkov, D. (eds.), *The Mehweb language: Essays on phonology, morphology and syntax*. Language Science Press, 1–15. DOI:10.5281/zenodo.3402054

[2] Moroz, G. 2019. Phonology of Mehweb. In: Daniel, M., Dobrushina, N. & Ganenkov, D. (eds.), *The Mehweb language: Essays on phonology, morphology and syntax*. Language Science Press, 17–37. DOI:10.5281/zenodo.3402056

[3] Arkhipov, A., Daniel, M., Belyaev, O., Moroz, G., Esling, J. H. 2019. A reinterpretation of lower-vocal-tract articulations in Caucasian languages. *Proc. 19th ICPhS* Melbourne, 1550–1554.

[4] Esling, J. H. 2005. There are no back vowels: The laryngeal articulator model. *Cdn. J. Ling.* 50, 13–44.

[5] Esling, J. H., Moisik, S. R., Benner, A., Crevier-Buchman, L. 2019. *Voice Quality: The Laryngeal Articulator Model*. Cambridge: CUP.

[6] Forrest, K., Weismer, G., Milenkovic, P., Dougall, R. N. 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.* 84(1), 115–123.

[7] Harrington, J. 2010. Acoustic phonetics. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds.), *The Handbook of Phonetic Sciences*. Wiley-Blackwell, 81–129.

[8] Shadle, C. 2023. Alternatives to moments for characterizing fricatives: Reconsidering Forrest et al. (1988). *J. Acoust. Soc. Am.* 153, 1412–1426.

[9] Bukmaier, V., Harrington, J. 2016. The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *J. Int. Phonetic Assoc.* 46, 311–329.

[10] Jannedy, S., Weirich, M. 2017. Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *J. Acoust. Soc. Am.* 142(1), 395–405.

[11] Soli, S. D. 1981. Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *J. Acoust. Soc. Am.* 70(4), 976–984.

[12] Stevens, K. N. 1998. *Acoustic Phonetics*. MIT Press.

[13] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Blackwell.

[14] Barreda, S. 2021. Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, 7(1). https://doi.org/10.1515/lingvan-2020-0051.

[15] Boersma, P., Weenink, D. 2022. *Praat: Doing phonetics by computer* [Computer program]. Version 6.2.23. http://www.praat.org/ (accessed 08.10.2022).

[16] R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/.

[17] Harrington, J. 2010. *The Phonetic Analysis of Speech Corpora*. Blackwell.

[18] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. 2011. Scikit-learn: Machine Learning in Python. *J. Machine Learning Research* 12, 2825–2830.

[19] Arkhipov, A. 2015. The acoustic correlates of vowel pharyngealization in Archi (East Caucasian). *Proc. 18th ICPhS* Glasgow, paper 1014.

RATE-SENSITIVE DIFFERENCES BETWEEN MODAL AND NON-MODAL VOWELS IN SAN MARTIN PERAS MIXTEC

Ben Eischens

University of California, Los Angeles
beischens@ucla.edu

ABSTRACT

The difference between laryngeal consonants and non-modal phonation is often difficult to determine based solely on phonetic data. This pilot study demonstrates a novel source of phonetic evidence for an analysis of [V?] sequences in San Martín Peras Mixtec as non-modal vowels that are phased as modal-then-glottalized. In slow speech, the non-modal portion of these vowels lengthens significantly more than the portion made up of modal voicing. This pattern differs from sequences of a modal vowel and a consonant, supporting an analysis of [V?] sequences as glottalized vowels. The study also investigates the behavior of [h], a sound whose (supra-)segmental status is not as clear, finding inter-speaker variation that merits further study.

Keywords: Mixtec, speech rate, non-modal phonation

1. INTRODUCTION

Cross-linguistically, the representational distinction between laryngeal consonants and non-modal phonation is not always apparent from investigations of their acoustic correlates [1]. For example, Davidson [2] found that the phonemic glottal stop consonant in Hawaiian is typically realized as a period of creaky voice, especially intervocalically, similarly to the phonetic realization of contrastively creaky vowels in Jalapa Mazatec [3]. Additionally, laryngealized vowels in many languages are often associated with glottal closure [4, 5, 6]. Because of this, analyses of whether laryngeal gestures are best analyzed as consonants or instantiations of phonation types are usually made based on phonological evidence. This is the case in Mixtec languages, which are commonly analyzed as contrasting modal and 'glottalized' vowels [7, 8, 9, 10, 11, 12, 13], though see also [14, 15, 16]. In these languages, glottalized vowels are usually realized as a sequence of modal voicing followed by creaky voicing or glottal closure. This contrast is found in San Martín Peras Mixtec (SMP Mixtec; ISO: jmx; glottocode: west2643), where the motivation for an analysis of [V?] sequences as glottalized vowels comes from phonotactic evidence [17]. The present study analyzes the characteristics of modal and glottalized vowels across speech rates and provides additional support for this analysis.

In SMP Mixtec, the proportion of a glottalized vowel that is made up of modal voicing decreases in slow speech. This happens regardless of whether the vowel is followed by a consonant or by another vowel. For example, in both words in Figure 1, the period of glottalization increases in duration disproportionately to the preceding modal voicing.

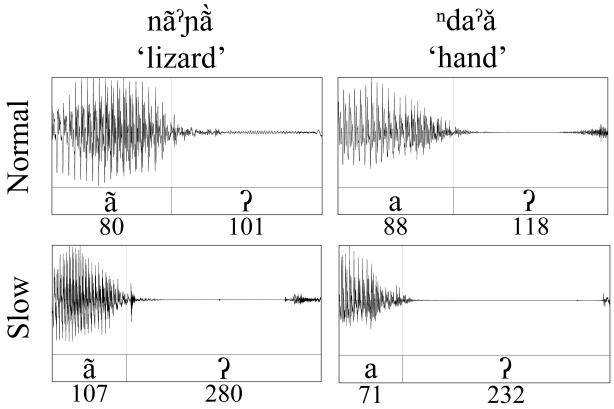


Figure 1: Waveforms and duration (ms) of the glottalized [aʔ] vowel in [nãʔɲǹ] ('lizard') and [ⁿdaʔă] ('hand') at normal and slow speech rates

Crucially, it is not the case that modal vowels never lengthen much in slow speech. The first vowel in the words in Figure 2 lengthens proportionately to the following consonant in slow speech.

The lengthening pattern in Figure 1 has implications for the decision to analyze phonetic [V?] sequences as glottalized vowels or as sequences of a vowel and a consonant. If they constitute glottalized vowels, the disproportionate reduction in duration of the modal vowel in slow speech in Figure 1 can be said to result from a production strategy in which the glottalized portion of a vowel is disproportionately lengthened in slow speech. However, an analysis of [?] as a consonant has trouble accounting for the pattern

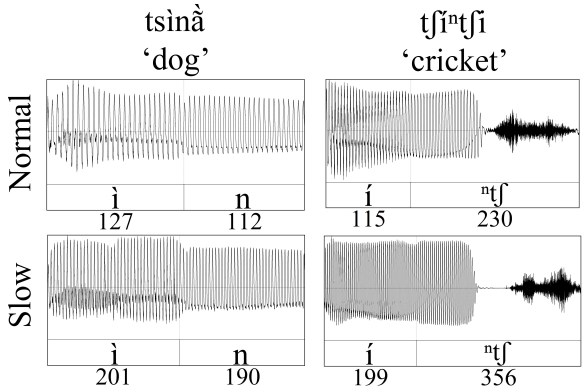


Figure 2: Waveforms and duration (ms) of the modal vowel [i] and the following consonant from [tsìnǎ] (‘dog’) and [tʃĭⁿtʃi] (‘cricket’) at normal and slow speech rates

because of how [ʔ] would be syllabified: It should be syllabified as a coda in [nãʔɲǎ] (‘lizard’) and as an onset in [ⁿdaʔă] (‘hand’). Importantly, there is no evidence of exceptional coda syllabification in CVCV words, and [ʔC] is not a possible onset.

When syllabified as a coda in [nãʔɲǎ] (‘lizard’), the disproportionate lengthening of [ʔ] can be understood as lengthening of the initial rime via selective lengthening of the coda (see [18] for similar asymmetries in Cantonese, and [19, 20] for onset-coda lengthening asymmetries in English and Dutch). However, when syllabified as an onset in [ⁿdaʔă] (‘hand’), the [ʔ] no longer forms a rime with the preceding vowel. The preceding vowel nucleus should lengthen more than in [nãʔɲǎ] (‘lizard’) because the coda is no longer available to be lengthened. However, this prediction is not borne out—the modal vowel decreases disproportionately to the following [ʔ] in slow speech in both cases. The fact that [ʔ] lengthens uniformly relative to the preceding vowel regardless of whether it would be a coda or an onset points toward an analysis of [Vʔ] sequences in SMP Mixtec as glottalized vowels, and away from an analysis of [ʔ] as a consonant.

This paper reports the results of a pilot study demonstrating this lengthening pattern for glottalized vowels. It also investigates the lengthening patterns of words with contrastive and non-contrastive [h] (referred to as breathiness and pre-aspiration throughout, respectively), which share many phonotactic properties with [ʔ] but cannot be as conclusively analyzed as non-modal phonation on the basis of phonotactic evidence. Ultimately, the findings of this study provide convergent evidence in favor of an analysis of [Vʔ] sequences as glottalized vowels in SMP Mixtec, and they do so through a novel method of speech rate manipulation. Though the results must still be replicated with a larger participant pool, they present a novel method for probing representational distinctions between laryngeal consonants and non-modal phonation that may also prove useful outside of SMP Mixtec.

2. LANGUAGE BACKGROUND

SMP Mixtec is an Otomanguean language spoken by about 10,000 people in western Oaxaca, Mexico [21] and by diaspora communities throughout Mexico and the US. As is the case across Mixtec [13, 22], it has a complex tonal system [17, 23], and roots are minimally (and usually maximally) bi-moraic. There are no coda consonants. The first vowel in a root may be modal (V), glottalized ($V^{ʔ}$), or followed by [h] (V^{h}). [h] predictably precedes all root-medial, voiceless consonants, but is contrastive before vowels and voiced consonants. Examples of roots of these types can be seen in Table 1.

3. METHODS

3.1. Participants

Two middle-aged, female speakers of SMP Mixtec who live and work in Watsonville, CA participated separately in this production task. Both grew up in the municipality of San Martín Peras, are bilingual in SMP Mixtec and Spanish, and speak Mixtec on a daily basis. The task was administered in Spanish.

3.2. Materials and procedure

Each participant (henceforth P1 and P2) produced target words in the carrier sentence in (1). They were asked to produce the first utterance at a normal rate, the second more slowly, and the third very slowly.

(1) kâ̂ʔ=ì̀ __ βiʰtsĩ
POT.say=1SG __ now
‘I will say __ now.’

To avoid priming effects, and because the language currently has no standardized orthography, the prompt sentences were read out loud in Spanish by the author. If a participant produced a target word other than the desired one, the author then specified the desired word in SMP Mixtec, and the participant produced the three utterances using that word. Target words varied in whether the initial vowel was glottalized, breathy, modal, or was followed by a pre-nasalized or pre-aspirated consonant (Table 1). Glottalized and breathy words varied in whether or not they included a root-medial consonant, with a roughly equal number of items in each condition.

				V1+P-V1		
				V1	**P-V1**	
Glottal	C2	nã^{ʔ}ñà 'lizard'	n	ã	ʔ	ñà
	No C2	^{n}da^{ʔ}ǎ 'hand'	^{n}d	a	ʔ	ǎ
Breathy	C2	nù̀^{h}nǐ 'corn'	n	ù̀	h	nǐ
	No C2	nì^{h}ǐ 'blood'	n	ì̀	h	ǐ
Preasp		tʃú^{h}tu 'cat'	tʃ	ú	h	tu
Prenas		tʃí^{n}tʃi 'cricket'	tʃ	í	n	tʃi
Modal		tsìnǎ 'dog'	ts	ì	n	ǎ

Table 1: Relevant portions of target words by condition. 'P-V1' stands for 'Post-V1.'

Words with a medial, pre-nasalized consonant were included for purposes of comparison because pre-nasalization occurs in the same local context as glottalization and breathiness but is unambiguously a feature of the medial consonant [17].

P1 produced the sentence in (1) at three different speech rates for 151 target words (41 glottalized + 40 breathy + 21 modal + 28 pre-aspirated + 21 pre-nasalized) for a total of 453 productions. P2 produced this three-sentence sequence for 153 target words (42 glottalized + 43 breathy + 22 modal + 22 pre-aspirated + 22 pre-nasalized) for a total of 459 productions. Few target words were prompted more than once, except in the pre-nasalized condition because pre-nasalization of root-medial consonants is rather uncommon in the language.

3.3. Data analysis

Each token was spliced and annotated by hand in Praat [24]. A Praat script extracted the total duration of each carrier sentence and each segment in the target words. Sentence duration spanned from the utterance-initial stop burst to the end of periodic voicing on the utterance-final vowel. In the target word, vowel duration was measured from the offset of the preceding consonant until either (1) the onset of the following consonant, or (2) the point at which periodic voicing ceased and silence or aperiodic noise began. When present, consonant off-glides were included in the duration of the following vowel because many such vowels never reached a steady state. [ʔ] duration was measured from the end of periodic vocal fold vibration on the preceding vowel until the onset of periodic vocal fold vibration in the following consonant or vowel. [h] duration was measured from the onset of frication noise, excluding any periodic, breathy voicing, (c.f. [25]) to the onset of the following consonant or vowel. The relevant portions of the tokens whose duration was measured are given in Table 1.

The dependent variable was the ratio of V1 to V1 + Post-V1, which is the ratio of the modal vowel's duration to the duration of the modal vowel plus whatever immediately follows it in the same word. Raw duration was not used because the phenomenon of interest is how much the modal vowel lengthens in slow speech relative to what follows, whether that is [ʔ], [h], the [n] of pre-nasalization, or a full consonant in the case of the modal words. To satisfy the linear regression's assumption of unbounded variables, each token's ratio of V1 to V1 + Post-V1, which was necessarily bounded at [0, 1], was centered and scaled. To obtain a continuous measure of speech rate, rate was quantified as moras per second (μ/sec) for each token's carrier sentence, all of which had 6 moras. P1's average μ/sec values for the three prompted rates from normal to slowest were 4.19 (SE = .04), 2.77 (SE = .03), and 1.79 (SE = .02). P2's rate change was smaller, with average μ/sec values for the three prompted rates being 4.66 (SE = .01), 3.7 (SE = .03), and 3.41 (SE = .03). Because μ/sec is bounded at 0, each token's value was centered and scaled for the linear regression.

3.4. Statistical models

All statistical tests were carried out in R [26]. The dependent variable was the centered and scaled value for each token's ratio of V1 to V1 + Post-V1, and the independent variables were speech rate (centered and scaled μ/sec), phonation type, and their interaction. A random effect of item was also included. To establish a baseline for comparison, Pearson's correlation coefficient was calculated to check for correlation between speech rate (μ/sec) and the ratio of V1 to V1 + Post-V1 for roots with a medial, pre-nasalized consonant. The test did not reach significance for either participant (P1: t = -.55, p = .58; P2: t = -1.32, p = .19). Because this ratio did not reliably correlate to speech rate, the pre-nasalized condition was used as the baseline for each participant. A linear mixed effects model was run using the lme4() package [27]. Because there were only two participants in this task, separate statistical models were run for each participant, with the results subjected to Bonferroni correction for multiple comparisons (α=0.05/2).

4. RESULTS

The models' results for P1 and P2 are given in Table 2. The residuals for each model were normally distributed (R = .991 for P1, .973 for P2). Model criticism was carried out using the drop1() function in the lmerTest package [28]. The full models were compared with simpler models omitting the phonation by μ/sec interaction. These comparisons were both significant ($p < .001$) using Satterthwaite's method, indicating that this interaction should not be excluded from the models. No further model simplification was possible.

Both models found negative main effects of glottalized and breathy phonation type, meaning the ratio of V1 to V1 + Post-V1 was lower across the board in these conditions than in pre-nasalized words. There was no main effect of pre-aspiration or modal phonation for either participant. For P1, there were significant positive interactions between rate (μ/sec) and the ratio of V1 to V1 + Post-V1 for glottalized and breathy words. This interaction held regardless of whether or not there was a medial consonant, and also for pre-aspirated words, which always have a medial consonant. This means that the modal portion of the vowel decreased disproportionately relative to [ʔ] or [h] as rate slowed down. This interaction was not seen for modal words, suggesting that the ratio of V1 to V1 + Post-V1 is consistent across speech rates for modal words for P1.

For P2, the positive interaction between rate and pre-aspirated roots was significant, and the interaction of rate with glottalized phonation was significant regardless of whether or not there was a medial consonant. However, the interaction between rate and breathy phonation was significant only for words with a medial consonant. That is, modal vowel duration decreased disproportionately relative to [ʔ] as rate slowed down whether or not there was a following consonant, but the modal vowel only decreased disproportionately relative to [h] when there was a following consonant. There was no interaction between rate and modal phonation.

5. DISCUSSION

For both participants, modal vowels are shorter before [ʔ] than in the baseline case, and their proportion relative to [ʔ] shrinks as speech rate slows regardless of whether or not there is a consonant following [ʔ]. As discussed in §1, this finding is straightforwardly explained by an analysis of [Vʔ] sequences in SMP Mixtec as glottalized vowels, but not by an analysis of [ʔ] as a consonant.

Predictor	β (SE)		|t|		p-value	
	P1	P2	P1	P2	P1	P2
Intercept	.9 (.18)	.93 (.15)	5.1	6.1	***	
μ/sec	-.04 (.05)	-.07 (.04)	-.84	-1.7	.4	.09
Glottal (w/ C2)	-1.28 (.2)	-1.62 (.19)	-6.3	-8.7	***	
Glottal (no C2)	-1.99 (.2)	-2.05 (.19)	-9.8	-11.06	***	
Breathy (w/ C2)	-1.31 (.21)	-1.25 (.2)	-6.2	-6.26	***	
Breathy (no C2)	-1.91 (.2)	-1.58 (.19)	-9.41	-8.45	***	
Preasp	-.09 (.2)	.11 (.15)	-.49	.72	.63	.47
Modal	-.17 (.2)	.003 (.18)	-.82	.02	.41	.98
μ/sec x Glottal (w/ C2)	.56 (.07)	.33 (.06)	7.54	5.54	***	
μ/sec x Glottal (no C2)	.45 (.08)	.26 (.06)	5.92	3.96	***	
μ/sec x Breathy (w/ C2)	.21 (.08)	.29 (.06)	2.73	4.95	*	***
μ/sec x Breathy (no C2)	.3 (.08)	.09 (.06)	3.89	1.58	***	.12
μ/sec x Pre-asp	.36 (.07)	.28 (.06)	5.19	4.46	***	
μ/sec x Modal	.06 (.07)	.03 (.06)	.77	.55	.44	.59

Table 2: Results of the mixed effects models with Bonferroni-corrected α codes: * = < .025, ** < .005, *** = < .0005

The data regarding [h] are less clear. For P1, [h] behaves the same as [ʔ]. However, as speech rate slows for P2, [h] lengthens disproportionately only when it would be syllabified as a coda, not an onset. This is consistent with the consonantal analysis. Interestingly, the phonotactic evidence for [h] as the realization of a phonation type in SMP Mixtec is not quite as strong as it is for [ʔ], despite their phonotactic similarity: [ʔ] is never root-initial, but one lexeme, the demonstrative [hàáˊ] ('that'), begins with [h]. With only two participants, it is not possible to tell if this inter-speaker variation reflects a principled distinction between groups of speakers. That said, the behavior of [h] does clear up one confound: The cases in which modal vowels decrease disproportionately as rate slows down involve the voiceless consonants [ʔ] and [h] (c.f. [1]).[1] The other cases involve sonorants and pre-nasalized obstruents, which are voiced. However, P2's modal vowels do not always decrease disproportionately to [h], obviating this confound.

6. ACKNOWLEDGEMENTS

I am very grateful to Natalia Gracida Cruz, Roselia Durán Cruz, Juan Gracida Ortiz, and other members of the San Martín Peras community in Oaxaca and in California for graciously sharing their language with me.

7. REFERENCES

[1] M. Garellek, Y. Chai, Y. Huang, and M. Van Doren, "Voicing of glottal consonants and non-modal vowels," *Journal of the International Phonetic Association*, 2021.

[2] L. Davidson, "Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian," *Journal of Phonetics*, vol. 88, 2021.

[3] D. Silverman, B. Blankenship, P. Kirk, and P. Ladefoged, "Phonetic structures in Jalapa Mazatec," *Anthropological Linguistics*, pp. 70–88, 1995.

[4] D. Silverman, "Laryngeal complexity in Otomanguean vowels," *Phonology*, vol. 14, no. 2, pp. 235–261, 1997.

[5] C. Gerfen and K. Baker, "The production and perception of laryngealized vowels in Coatzospan Mixtec," *Journal of Phonetics*, vol. 33, no. 3, pp. 311–334, 2005.

[6] H. Avelino, "Acoustic and electroglottographic analyses of nonpathological, nonmodal phonation," *Journal of Voice*, vol. 24, no. 3, pp. 270–280, 2010.

[7] M. Macaulay and J. C. Salmons, "The phonology of glottalization in Mixtec," *International Journal of American Linguistics*, vol. 61, no. 1, pp. 38–61, 1995.

[8] C. Gerfen, *Phonology and phonetics in Coatzospan Mixtec*. Springer Science & Business Media, 2013, vol. 48.

[9] I. McKendry, "Tonal association, prominence and prosodic structure in South-Eastern Nochixtlán Mixtec," 2013.

[10] B. Becerra Roldán, "Un estudio fonológico del mixteco de Santo Domingo Huendio, Oaxaca," Master's Thesis, 2015.

[11] J. Mendoza Ruiz, "Fonología segmental y patrones tonales del Tu'un Savi de Alcozauca de Guerrero," Master's Thesis, 2016.

[12] O. León Vásquez, "Sandhi tonal en el mixteco de Yucuquimi de Ocampo," *MA thesis, Centro de Investigaciones y Estudios en Antropología Social*, 2017.

[13] K. Penner, "Prosodic structure in Ixtayutla Mixtec: Evidence for the foot," Ph.D. dissertation, 2019.

[14] C. Mak, "The tonal system of a third Mixtec dialect," *International Journal of American Linguistics*, vol. 24, no. 1, pp. 61–70, 1958.

[15] K. L. Pike, "Analysis of a Mixteco text," *International Journal of American Linguistics*, vol. 10, no. 4, pp. 113–138, 1944.

[16] L. Pankratz and E. V. Pike, "Phonology and morphotonemics of Ayutla Mixtec," *International Journal of American Linguistics*, vol. 33, no. 4, pp. 287–299, 1967.

[17] B. Eischens, "Tone, phonation, and the phonology-phonetics interface in San Martín Peras Mixtec," Ph.D. dissertation, University of California, Santa Cruz, 2022.

[18] E. Zee, "The effect of speech rate on the temporal organization of syllable production in Cantonese," in *Speech Prosody 2002, International Conference*, 2002.

[19] A. E. Turk and J. R. Sawusch, "The domain of accentual lengthening in American English," *Journal of Phonetics*, vol. 25, no. 1, pp. 25–41, 1997.

[20] T. Cambier-Langeveld and A. E. Turk, "A cross-linguistic study of accentual lengthening: Dutch vs. English," *Journal of Phonetics*, vol. 27, no. 3, pp. 255–280, 1999.

[21] Instituto Nacional de Estadística y Geografía, "Censo de población y vivienda," 2010.

[22] C. DiCanio and R. Bennett, "Prosody in Mesoamerican languages," *The Oxford Handbook of Language Prosody*, 2018.

[23] S. L. Peters, "The inventory and distribution of tone in Tù'un Ndá'vi, the Mixtec of Piedra Azul (San Martín Peras), Oaxaca," Master's Thesis, 2018.

[24] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 6.1.27)," 2020.

[25] J. Kingston, "The coordination of glottal with oral articulations in Icelandic and Hindi," *The Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S114–S114, 1987.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[1] The phonotactic properties of SMP Mixtec, in which all root-medial voiceless consonants are preceded by [h], do not allow for direct comparison between [Vh/Vʔ] sequences and [VC] sequences where the C is voiceless.

THE ACOUSTIC AND ARTICULATORY PROPERTIES OF THE PRENASALISED CORONAL TRILL IN TWO OCEANIC LANGUAGES

Tihomir Rangelov

Max Planck Institute for Evolutionary Anthropology, Leipzig
tihomir_rangelov@eva.mpg.de

ABSTRACT

A few Austronesian languages have a prenasalised coronal trill with a plosive-like release (hereinafter NDR), whose exact place and manner of articulation are rarely discussed in detail; the existing impressionistic descriptions vary. This paper adds to our knowledge about this rare complex segment by offering a detailed acoustic and articulatory description of NDR in two Oceanic languages spoken in Vanuatu – Ahamb and Farsaf. The description includes measurements of the nasal portion, the frequency of vibration, and the exact place and manner of contact between the articulators, and is based on audio recordings of citation forms, linguograms and palatograms.

Both languages contrast NDR with a plain coronal trill /r/. Perceptually, both prenasalisation and the plosive-like release in NDR appear to be salient cues for contrast. The results suggest that trilling is the main phonetic characteristic, and the plosive-like release is an effect of trilling being initiated following an oral closure.

Keywords: trills, prenasalisation, plosive release, Oceanic, Austronesian

1. INTRODUCTION

A voiced prenasalised coronal trill has been documented in a number of languages, most notably languages of the Austronesian family. There are different reports about the exact place and manner of articulation of such sounds. The more general designation "coronal" can refer to what have been claimed to be alveolar/apical [1, 2, 3, 4] or post-alveolar/retroflex [5, 6, 7, 8] articulations. The designation "prenasalised trill" is debatable as well – such sounds are also sometimes called "plosives with a trilled release" [9] since they normally feature a perceptually salient plosive-like release, followed by the period in trilling.

Phonological evidence from the languages where prenasalised coronal trills have been described, shows that we have one complex segment at hand here, rather than a sequence of more than one segment, e.g. [ndr]. Because of the perceptual salience of all three phases – prenasalisation, plosive-like release and trilling – this complex segment is often denoted in orthographies with the trigraph <ndr> or the digraph <dr> (as prenasalisation is often not marked in the orthographies of languages where it is distinctive; languages with a prenasalised coronal trill often also have other prenasalised sounds, mostly plosives, but also affricates or bilabial trills). Hereinafter I refer to the prenasalised coronal trill as NDR (to refer to the three portions of this complex segment) for the sake of clarity and brevity. In phonemic transcriptions, I use /D/.

To our knowledge, no detailed articulatory and acoustic studies of NDR exist. This study aims to partially fill this gap by offering an articulatory and acoustic analysis of NDR in two languages of Vanuatu – Ahamb (ISO 639-3 code: ahb, Glottocode: axam1237) and Farsaf (ISO 639-3 code: nrg, Glottocode: nara1263). Ahamb is spoken on the small Ahamb Island off the south coast of Malekula, the second largest island of Vanuatu. Farsaf (also known as Narango) is spoken in a few villages (the main one being Narango) near the south coast of Espiritu Santo, Vanuatu's largest island. Both languages fall within the North-Central Vanuatu (NCV) group of Oceanic languages, but are members of different subgroups of NCV, see Glottolog [10] for a genealogical tree.

A number of other NCV languages have NDR; these are some of the other 80 or so languages of Malekula and Espiritu Santo [11], as well as Nafsan, another NCV language spoken on Efate Island further south [3, 11]. Outside of Vanuatu, within the Oceanic group, NDR has been reported in Fijian [4, 12], a few languages of the Admiralties group, (Manus Island, Papua New Guinea) [1], Sa'a (Solomon Islands) [1], as well as in the non-Oceanic Austronesian languages Nias (Sumatra, Indonesia) [5, 13] and Malagasy [8]. NDR has been reconstructed for Proto Oceanic [14]. Maddieson [15] suggests that a similar sound can be reconstructed for Proto Austronesian. Outside of Austronesian, prenasalised coronal trills have been reported in some languages of Africa [16, 17, 18].

NDR has two main phases: a closure period and a period in trilling. The closure period can be divided into *nasal closure* (nasal airflow is present, oral airflow is absent), followed by a *total closure* (absence of both nasal and oral airflow). This is then normally followed by the plosive-like release and the trilling phase.

In Ahamb, NDR contrasts with the plain alveolar trill /r/, the prenasalised alveolar plosive /ⁿd/ (a marginal phoneme found in only two words), and the

prenasalised and plain bilabial trills /ᵐʙ, ʙ̥/ [7, 19]. Prenasalisation is the main contrast in Ahamb's plosives and bilabial trills [7, 19]. Farsaf lacks proper phonological analysis and description. However, newly collected data suggest that in this language, NDR contrasts with a plain trill /r/. Farsaf appears to lack a coronal plosive, but there is a prenasalisation contrast in the bilabial and velar plosives.

A morphophonological process in Ahamb, which is relevant to this study, is that /nr/ clusters are produced by younger speakers similarly to NDR, with a perceptually salient plosive-like release. /nr/ clusters in Ahamb are almost exclusively found in word-initial position in nouns, across a morpheme boundary, where /n/ is a common noun marker, a reflex of a former article *na, and the noun stem begins with /r/, e.g. /n-ran/ 'ground', /n-ras/ 'sea', and borrowings such as /n-rum/ 'room' [7]. Older speakers of Ahamb lack such "intrusive" plosives in /nr/ clusters, and in some cases appear to retain the historical vowel as an extra short schwa.

This paper adds to our knowledge about NDR, whose phonetic properties have received little in-depth attention. More specifically, it aims to establish the acoustic and articulatory characteristics of NDR in Ahamb and Farsaf, including measurements of the duration of the nasal and oral closures and the frequency rate of trilling. Linguography and palatography data give evidence of the exact place and manner of articulation of NDR. Furthermore, I attempt to rank the importance of the different phonetic characteristics of NDR, in terms of their role in distinguishing it from other sounds in the consonant inventories of Ahamb and Farsaf.

2. DATA AND METHODOLOGY

2.1. Recording procedures

The analysis presented in this paper is based on controlled wordlist audio data, as well as linguograms and palatograms. All data were collected in late 2022 in Vanuatu by the author.

One speaker of each language was recorded, a 30-year-old male speaker of Ahamb and a 32-year-old male speaker of Farsaf. Both speakers were born, raised and currently reside in their respective communities. Ahamb and Farsaf were their respective first languages, which they use at home, when socialising with other members of their communities, and in other domains. The Ahamb wordlist was compiled in collaboration with Ahamb speakers, based on previously available Ahamb data [7, 19]. The Farsaf wordlist was compiled in Narango, in collaboration with Farsaf speakers, by extracting items from a list of over 500 words previously collected by the first author. In both cases, the wordlists contain items in which NDR appears in different positions, both in syllable onset and coda position, word-initially, word-finally, between vowels and preceding another consonant.

The Ahamb list contains 19 items with NDR, five items with word-initial /nr/ clusters (the recorded Ahamb speaker normally produces these with the "intrusive" plosive), four items with /r/ in other environments, and two items with /ⁿd/. The Farsaf list includes 19 items with NDR and eight items with /r/ in different environments.

The audio data for each speaker were recorded in two sessions. In a first session, each word was recorded at 24-bit/48kHz using an Audix HT5 headset microphone and a ZOOM H5 audio recorder. In a second session, a recording was made for the purpose of analysing nasality, using a method described by Bruil & Stewart [20] (a modified version of Stewart & Kohlberger's [21] method), which allows for easy and affordable recordings for nasality analysis in remote settings. With this method, earbud headphones are plugged into the microphone input of the recording device. The speakers insert the two earbuds in their nostrils, with the signal from both earbuds being recorded in one channel, while another microphone is placed in front of the mouth and feeds the other channel. After normalisation, the nasal channel of the resulting stereo recording clearly shows the cut-off points and duration of nasal gestures, see [20, 21] for more details. I used unbranded earbuds and the Audix HT5 headset microphone, and the ZOOM H5 device at the 24-bit/48kHz setting. In all cases, speakers were prompted in Bislama, the national language of Vanuatu, and were asked to repeat each word three times. The second repetition was used for analysis, unless it was judged unnatural or spoilt by noise, in which case the first or third repetition was used.

It was only logistically possible to perform linguograms and palatograms for the Ahamb speaker. These were made on Ahamb Island with the same speaker who provided the audio data.

Speakers provided informed consent to their participation. Strict hygiene procedures were followed.

2.2. Data processing and analysis

Waveforms and spectrograms were inspected in Praat [22]. The recordings from the first (non-earbud) sessions were used to measure the duration of the period in trilling and the number of oscillations, in examples with NDR and /r/.

The earbud recordings were used to measure closure durations. Nasal closure was measured as the

period from the onset of high amplitude in the nasal channel when NDR is in initial position, or from a change in the nasal waveform with a simultaneous drop in energy in the oral channel, when NDR is after a vowel, until a significant drop in energy in both channels, more clearly visible on the spectrogram. This drop in energy signals the onset of the total closure period, which ends with a sudden burst of energy, corresponding to the release. Figure 1 shows an example of a spectrogram and waveform of a Farsaf word recorded with this method.

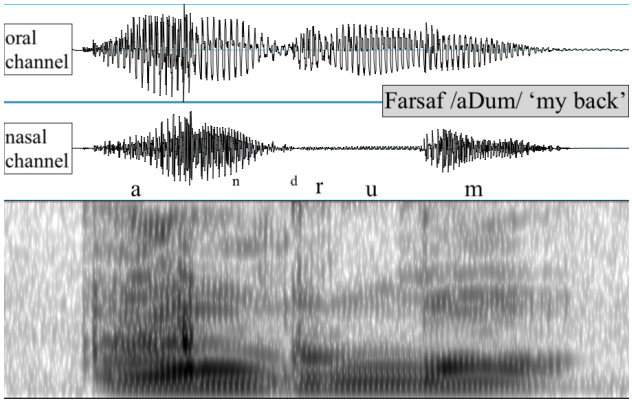


Figure 1: A spectrogram and waveform for Farsaf /aDum/ 'my back', recorded using the earbud method

3. RESULTS

3.1. The articulators

The linguogram and palatogram for the Ahamb speaker (target word /Daᵑg/ 'strong') in Figure 2 show that the articulation involves the apex of the tongue, which comes into contact with the alveolar ridge at the base of the upper teeth and immediately behind them. This means that, at least in Ahamb, NDR is an apical alveolar consonant without any major change of the place or manner of articulation throughout its different stages. Palatograms and linguograms for the same speaker for /n, r/ (the targets were /a**n**aᵑg/ 'my mother' and /**r**aŋ/ 'to cry'; the figures are not included here in the interest of space), show a very similar place of contact between the articulators. The Farsaf speaker reported a similar place and manner of articulation when asked.

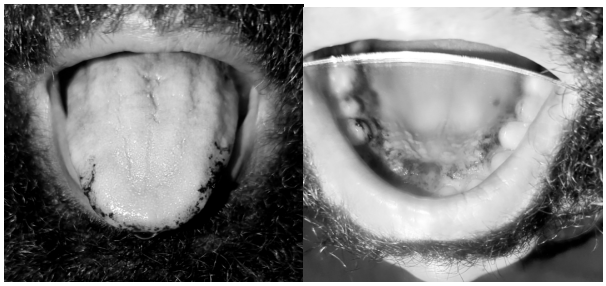


Figure 2: Linguogram and palatogram for Ahamb /Daᵑg/ 'strong'

3.2. Prenasalisation

The mean duration of the entire closure phase for NDR was 120 ms (s.d. = 29) for Ahamb and 102 ms (s.d. = 28) for Farsaf. Of this, the nasal closure (prenasalisation) was on average 94 ms for Ahamb and 85 ms for Farsaf, corresponding to 78% (Ahamb) and 82% (Farsaf) of the entire closure phase. The differences in closure duration between the two speakers are small enough to be attributed to speech rate effects or individual characteristics, especially considering that the proportion of prenasalisation is less significantly different.

The substantial proportion of the nasal closure suggests that prenasalisation is a major cue for contrast between NDR and /r/ in both languages. This is not surprising, considering that prenasalisation is the main contrast in plosives in both languages and also between the two bilabial trills in Ahamb [19].

3.3. Trilling

Between one and four trill periods were observed during spectrogram and waveform analysis, most commonly one or two. Figure 3 demonstrates a realisation of NDR with two trill periods.

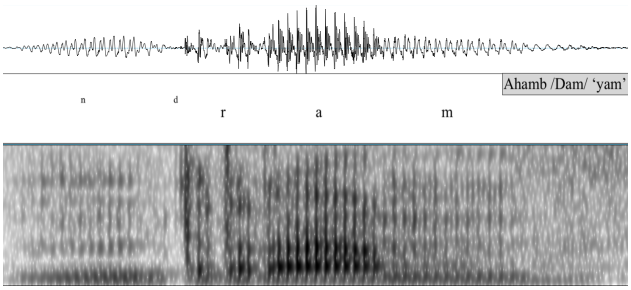


Figure 3: A spectrogram and waveform for Ahamb /Dam/ 'yam', showing two complete oscillations

The duration of the period in trilling varied substantially based on the number of complete oscillations recorded. The mean duration of each oscillation was 41 ms for Ahamb and 39 ms for Farsaf. This corresponds to a frequency rate of trilling of 24 Hz and 26 Hz respectively. These numbers are within the normal rates observed in trills cross-linguistically [9]. The frequency rates of trilling in items with /r/ for the two languages were very similar – 25 Hz for Ahamb and 28 Hz for Farsaf. This means that there are no substantial differences in terms of trilling between NDR and /r/.

Trills are generally known to easily fail to be realised properly, since they are "very sensitive to small variations in the articulatory and aerodynamic conditions" [9]. For example, the bilabial trills of Ahamb are often realised as plosives or fricatives [19]. In our data, in some cases where NDR was

realised with only one oscillation of the tongue, trilling was perceptually available but was not as clearly distinguishable on the waveform and spectrogram as, for example, in the clear oscillations in Figure 3 above. (This allophonic realisation of the trill should be more precisely referred as a tap, rather than a trill.) However, in all such cases there was a clear drop in amplitude/airflow, following the initial release (see Figure 4 below), which cannot be observed in instances of a regular plosive release with /ⁿd/, as in Figure 5. This suggests that this drop in amplitude/airflow is an important articulatory and perceptual aspect of NDR.

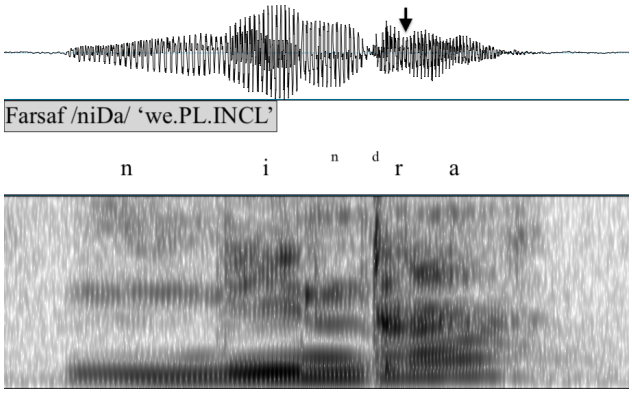


Figure 4: A spectrogram and waveform for Farsaf /niDa/ 'we.PL.INCL', showing a drop in amplitude after the initial release (indicated by an arrow) in a realisation of NDR with one oscillation of the tongue

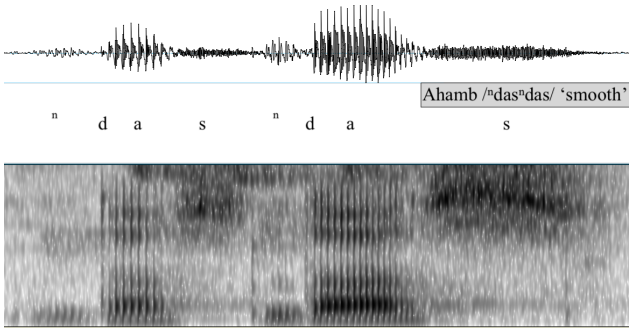


Figure 5: A spectrogram and waveform for Ahamb /ⁿdasⁿdas/ 'smooth', showing a constant rise in amplitude following the initial release in both instances of /ⁿd/

3.4. /nr/ clusters in Ahamb

For the examples of /nr/ clusters in Ahamb, which the recorded speaker produced with the plosive-like release, the frequency rate of trilling was 25 Hz, which is similar to the values for NDR and /r/. The duration of the entire closure was 132 ms and the nasal closure duration was 111 ms (84% if the entire closure duration). These numbers are somewhat higher than those for NDR, albeit not substantially so. A possible explanation is the presumably recent origin of this prenasalisation as a full /n/ segment.

4. DISCUSSION AND CONCLUSIONS

This study of NDR has allowed to establish more precisely the phonetic properties of this rare sound in two small indigenous languages of Vanuatu. The results show that prenasalisation is the main strategy for contrast between NDR and plain alveolar trills. Trilling rates are similar to those for /r/. In all cases, a drop in airflow following the initial release of the articulators, appears to be an important characteristic.

In Farsaf, a prenasalised plosive is absent and in Ahamb it is only a marginal sound, but we do not see NDR changing to /ⁿd/ in order to occupy this gap in these two languages' phonological systems. In fact, our data suggest that trilling (or at least a partial closure following the initial release) is more important than the plosive-like release in the production and perception of NDR. In fact, from an aerodynamic point of view, the plosive-like release can be viewed as a product of the build-up of intraoral pressure during the total closure period. Even though this pressure likely does not have the chance to rise significantly (since air is escaping through the nose during most of the closure phase, as in bilabial trills [15]), it likely rises enough to cause a sudden and energetic first release, similar to that of plosives. The same process has been observed with prenasalised bilabial trills (which have a very similar manner of articulation as NDR [23]), where the first release is usually more sudden than the following releases during the trilling phase [19]. Furthermore, Ahamb's /nr/ clusters' surfacing as NDR also suggests that the plosive-like release is a bi-product of the articulation.

Based on these results, we can now establish that NDR is a *prenasalised apico-alveolar trill* – [ⁿr], or more narrowly transcribed as [ⁿᵈr] to account for the plosive-like release – at least in Ahamb, for which we have the palatography and linguography data. This is most likely also the case for Farsaf.

Future work on Ahamb, Farsaf and other languages with NDR can shed more light on the properties of this rare complex segment and the importance of its different aspects with regard to articulation, acoustics and perception.

5. ACKNOWLEDGEMENTS

I am grateful to the speakers recorded for this study for their time and patience during data collection and to the Ahamb Island and Narango communities for their generosity and hospitality and for helping overcome all logistical issues as much as possible. This work has been supported financially by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany.

6. REFERENCES

[1] Blust, R. 2007. The prenasalised trills of Manus. In: Siegel, J., Lynch, J., Eades, D. (eds), *Language description, history, and development: Linguistic indulgence in memory of Terry Crowley*, 297–311. Amsterdam, Netherlands: John Benjamins

[2] Crowley, T. 2006. *The Avava language of central Malakula (Vanuatu)*. Canberra: Pacific Linguistics.

[3] Thieberger, N. 2006. *A grammar of South Efate. An Oceanic language of Vanuatu*. Honolulu: University of Hawai'i Press.

[4] Dixon, R.M.W. 1988. *A grammar of Boumaa Fijian*. Chicago: University of Chicago Press.

[5] Catford, J.C. 1988. Notes on the phonetics of Nias. In McGinn, R. (ed), *Studies in Austronesian linguistics,* 151-172. Athens, OH: Center for Southeast Asian Studies.

[6] Dimock, L. 2009. *A grammar of Nahavaq (Malakula, Vanuatu)*. Victoria University of Wellington. (PhD thesis.)

[7] Rangelov, T. 2020. *A grammar of the Ahamb language (Vanuatu)*. Hamilton, New Zealand: University of Waikato. (PhD thesis.)

[8] Howe, P. 2021. Central Malagasy. *Journal of the International Phonetic Association* 51(1). 103-136.

[9] Ladefoged, P., Maddieson, I. 1996. *The sounds of the world's languages*. Oxford: Blackwell.

[10] Hammarström, H., Forkel, R., Haspelmath, M., Bank, S. 2022. *Glottolog 4.7*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

[11] Tryon, D. 1976. *New Hebrides languages: An internal classification*. Canberra: Australian National University.

[12] Schütz, A.J. 1985. *The Fijian language*. Honolulu: University of Hawaii Press.

[13] Yoder, B. 2010. Prenasalization and trilled release of two consonants in Nias. In Baart, J. (ed), *Work papers of the Summer Institute of Linguistics, University of North Dakota session*. University of North Dakota

[14] Ross, M., Pawley A., Osmond, M. (eds). 2016. *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 5: People: body and mind*. Canberra: Australian National University.

[15] Maddieson, I. 1989. Aerodynamic constraints on sound change: The case of bilabial trills. *UCLA Working Papers in Phonetics* 72. 91-115

[16] Demolin, Didier. 1988. Some problems of phonological reconstruction in Central Sudanic. *Belgian Journal of Linguistics* 3. 53-95.

[17] Demolin, D. 1992. *Le mangbetu: Étude phonétique et phonologique.* Brussels: Université Libre de Bruxelles. (PhD thesis.)

[18] McKee, R.G. 2001. Concerning Meegye and Mangbetu's bilabial trills. In Payne, D.L., Reh, M. (eds), *Advances in Nilo-Saharan Linguistics. Proceedings of the 8th Nilo-Saharan Linguistics Colloquium, University of Hamburg, August 22–25 2001 (Nilo-Saharan Linguistic Analyses and Documentation 22)*, 181-189. Cologne, Germany: Rüdiger Köppe.

[19] Rangelov, T. 2019. The bilabial trills of Ahamb (Vanuatu): Acoustic and articulatory properties. In: Calhoun, S., Escudero, P., Tabain M., Warren, P. (eds), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, 1292-1296. Canberra: Australasian Speech Science and Technology Association Inc.

[20] Bruil, M., Stewart, J. 2022. Phonetics and phonology of nasality in Ecuadorian Siona. *Phonological Data and Analysis* 4(3). 1-34

[21] Stewart, J., Kohlberger, M. 2017. Earbuds: A New Method for Measuring Nasality in the Field. *Journal of Language Documentation and Conservation* 11. University of Hawai'i, 49-80.

[22] Boersma, P., Weenink, D. 2022. *Praat: Doing Phonetics by Computer* [Computer program]. Version 6.3.03.

[23] Rangelov, T., Walworth, M., Barbour, J. 2023. A multifaceted approach to understanding unexpected sound change: The bilabial trills of Vanuatu's Malekula Island. *Diachronica*. https://doi.org/10.1075/dia.21051.ran

VOWEL DURATION IN ENENLHET

Paige Erin Wheeler

The University of Texas at Austin
pewheeler@utexas.edu

ABSTRACT

This study describes vowel duration in Enenlhet (ISO 639-3 tmf; Paraguay), based on a corpus of 3.5 hours of naturalistic speech. Enenlhet is both under-documented and endangered, having only around 1200 speakers. The Enenlhet corpus was aligned using untrained forced alignment, and vowel boundaries were hand corrected. Duration data were analysed with a linear mixed-effects model which examined factors shown by cross-linguistic research to regularly affect vowel duration. Results show that vowels are longer in pre-pausal syllables, in open syllables, and before voiced consonants in closed syllables; these three factors interact. The mid vowel /e/ is shorter than both /o/ and /a/.

Keywords: vowel duration, Enenlhet (Enlhet-Enenlhet) language description, naturalistic speech

1. INTRODUCTION

Enenlhet is one of six Enlhet-Enenlhet languages, all spoken in the Gran Chaco region of Paraguay. All six languages are endangered, and the family is under-described [1]. There are no detailed linguistic studies of Enenlhet, and only two phonetic studies of related languages (Enxet and Sanapaná) [2, 3]. The Enenlhet vowel inventory contains three vowels: /a, e, o/. Though sister languages have been argued to have phonemic vowel length, Enenlhet has not been described to have it [4, 5].

This study examines Enenlhet vowel duration through a corpus of naturalistic speech data [6]. In addition to contributing to the description of Enenlhet, it also provides information about vowel duration in spontaneous, as opposed to elicited speech, which may differ [7, 8]. Factors often found to increase vowel duration are considered: utterance-final position, stress, voiced following consonants, open syllables, and low vowels.

Utterance-final lengthening is widely attested [9, 10, 11, 12], though the domain may vary (e.g., final syllable, final foot, final word [13]). Though much of the literature on stress focuses on European languages, greater duration has been found to be a robust acoustic correlate of stress [14, 15, 16, 17]. Vowels in many languages are also longer preceding a voiced consonant [18, 19] (though cf. [20]). Open-syllable lengthening is also present in many languages [21, 22], though some studies do not find longer vowels in open syllables [23]. Low vowels have also been shown to have a longer inherent duration than mid or high vowels, perhaps due to articulatory constraints [24, 25, 26].

3. METHODS

3.1. Corpus

This study uses naturalistic speech collected as a part of a project to create an Enenlhet dictionary and available in the Archive of the Indigenous Languages of Latin America [6]. Recordings (24bit/96kHz) were made with a Zoom h4n recorder and an XLR lavalier microphone. Transcription, translation, and utterance-level segmentation were done by the corpus compilers. Utterance-level segmentation was done based on syntactic criteria; most utterances were also bounded by silence.

Interviews or narratives were selected for this study. The selected recordings amounted to 3.5 hours of speech from 8 speakers (4 men, 4 women), with a total of 15,031 vowels from 2013 unique lexical items. Most words contained between one and three syllables; the longest words in the corpus had eight.

3.2. Alignment and annotation

To speed data processing, the corpus was force-aligned using the Praat [27] plugin Easyalign [28]. Since there are no trained acoustic models of Enenlhet, this force-alignment used a model for Spanish (with *seseo*). Transcriptions were adjusted to account for deletion and speech disfluencies, both of which were frequent.

Utterance boundaries were marked at all utterance boundaries in the original transcription, as well as after any vowel ending >250ms before the start of the next vowel (i.e., followed by silence). Each word was marked as either pre-pausal or non-pre-pausal. Vowels were marked for their position within a word, numbered from right word edge and then again from the left. Enclitics [29] were counted as part of the host word. Vowels were marked for the voicing of the following consonant. Syllables were marked as either open or closed using a Python [30] script which segmented consonants as onsets whenever possible (i.e., CV.CV not CVC.V).

3.3. Boundary correction and measurement

Vowel boundaries were manually adjusted based on the following criteria: following a closure, release bursts were included as part of the vowel; preceding a closure, vowel offset was placed at the end of F2; adjacent to nasals, boundaries were placed at the beginning of a sharp intensity drop/rise or, if not visible, the onset and offset of antiformants; adjacent to frication, onset/offset were marked at the end/beginning of aperiodic noise; after approximants, onset was marked at the start of the amplitude increase or, if not visible, the beginning of the F1 rise; vowel offset before approximants was placed at the start of the amplitude decrease marking the start of the constriction.

The duration of each vowel segment was extracted using a Praat script, which also extracted the word label associated with each vowel. This analysis considers only tokens transcribed as a single vowel. The original transcriptions also included some adjacent vowels not separated by glides and /VʔV/ sequences with identical vowels. In related languages, these cases have been analyzed as sequences of segments rather than single units [3, 5], and therefore they warrent a separate analysis.

3.4. Hypotheses

Because utterance-final lengthening is so widely attested, Enenlhet vowels were expected to be longer before a pause (utterance boundary) than utterance-medially. If Enenlhet has fixed stress, it was expected to be identifiable as a lengthening effect associated with a vowel's position within a word. Lexically idiosyncratic stress would be indistinguishable from phonemic length. Vowels were expected to be longer before the voiced consonants – nasals (/m, n, ŋ/), glides (/j, w/), and the lateral approximant (/l/) – than the voiceless ones (/p, t, k, q, ʔ, ɬ, s, h/). Consonant clusters are rare in Enenlhet, with most syllables being either CV or CVC. Vowels were anticipated to be longer in open (C(C)V) syllables than in closed ((C)CVC(C)) ones. Enenlhet has no high vowels, and studies of vowel quality in sister languages have found more variation in the F2 dimension compared to the F1 [2, 3]. Nevertheless, the low vowel /a/ was predicted to be longer than the mid vowels /e/ and /o/. The mid vowels were not expected to differ.

3.5. Statistical analysis

A linear mixed effects model was calculated in R [31] using the *lme4* package [32]. The model treated duration as the dependent variable and included vowel quality (/a, e, o/), following consonant voicing (voiced or voiceless), syllable structure (open or closed), word position, and utterance position of the word (pre-pausal vs. medial) as fixed effects. Lexical item and speaker were random intercepts. It also included all possible two-way interactions between the fixed effects. The baselines for each variable were: vowel quality /e/, medial word, final syllable, voiceless following consonant, closed syllable.

This model showed significant effects of word position, utterance position, and their interaction. These two effects were investigated separately by examining two overlapping subsets of the corpus: all word-final vowels, in both medial and pre-pausal words; and all medial words, with vowels in any position. The model including only medial words did not show any significant word position effects, regardless of whether word position was anchored to the left or the right word-edge, so the word position and utterance position variables were subsequently recoded and combined. Each syllable was marked as either immediately pre-pausal or non-pre-pausal, and a third model was run using this variable in place of utterance and word position separately. This model also included all possible two-way interactions between consonant voicing, syllable structure, and pre-pausal position; it used the same baseline variables as the previous one.

4. RESULTS

Table 1 shows the significant effects and their estimated intercepts from the model with word position and utterance position pooled. Effects which were not significant are not shown in the tables. All results are significant at $p < 0.0001$.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	1.04	9.92	10.28
vowel quality: /a/	0.61	17.82	10.88
pre-pausal syllable	1.04	15.61	16.28
vcd following C	0.81	10.31	8.35
open syllable	0.87	6.53	5.70
vcd C * pre-pausal syll	1.40	-7.44	-10.39
open syll * pre-pausal	1.36	8.10	11.00
vcd C * open syll	1.11	-8.17	-9.12

Table 1: Significant effects and interactions from model combining word and utterance position variable (n=14,810)

Table 1 shows a significant effect of vowel quality, with /o/ and /a/ both being longer than /e/. Two-way comparisons with *emmeans* [33] indicate a significant difference between /a/ and /e/, and between /e/ and /o/, but not between /o/ and /a/.

Table 1 also shows a negative interaction between consonant voicing and pre-pausal syllable. To investigate these factors separately, an additional model containing only utterance-medial vowels in closed syllables was run; results appear in Table 2.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	1.56	4.74	7.37
vowel quality: /a/	0.78	8.88	6.96
vcd following C	0.75	9.83	7.34

Table 2: Significant effects in model with medial vowels in closed syllables (n=6238)

Table 3 shows the results of a model containing only medial vowels followed by voiceless consonants, which examined the effect of syllable structure.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	1.47	7.74	11.35
vowel quality: /a/	0.87	14.31	12.48
open syllable	0.86	6.15	5.32

Table 3: Significant effects in model with only medial vowels followed by voiceless consonants (n=5494)

Finally, Table 4 shows the results of a model containing only vowels in closed syllables followed by voiceless consonants, both medial and pre-pausal.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	2.54	5.62	14.30
vowel quality: /a/	1.75	8.23	14.42
pre-pausal syllable	1.44	19.82	28.44

Table 4: Significant effects in model containing closed syllables with voiceless consonant codas (n=3522)

Tables 2, 3, and 4 show significant effects of voiced following consonants, open syllables, and pre-pausal position, respectively. To investigate the interaction between these three factors, another pair of models were run, one including only utterance-medial syllables, and one including only voiceless consonants. Table 5 shows the results for the model containing only utterance-medial vowels.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	1.01	10.66	10.80
vowel quality: /a/	0.58	17.41	10.14
vcd following C	0.75	10.26	7.69
open syllable	0.81	6.11	4.97
vcd C * open syll	1.09	-6.77	-7.36

Table 5: Significant effects and interactions from model containing utterance-medial vowels (n=11,661)

The significant negative interaction between consonant voicing and syllable structure shown in Table 5 suggests that the lengthening effect due to voiced following consonants is smaller in open syllables than in closed ones. Table 1 also includes a negative interaction between consonant voicing and utterance position, which suggests that the lengthening associated with voiced following consonants is limited in pre-pausal position.

Table 6 shows the results from the model containing only voiceless following consonants, which investigated the effect of syllable structure and its interaction with utterance position.

Effect	Std. error	t-value	Intercept (ms)
vowel quality: /o/	1.51	5.87	8.84
vowel quality: /a/	0.94	12.93	12.20
pre-pausal syllable	1.23	12.50	14.14
open syllable	1.01	4.48	4.52
pre-pausal * open syll	1.78	8.17	14.52

Table 6: Significant effects and interactions from model containing voiceless following consonants (n=7717)

Table 6 shows a significant positive interaction between pre-pausal position and syllable structure, suggesting more lengthening in pre-pausal open syllables than medial ones.

5. DISCUSSION AND CONCLUSION

5.1. Pre-pausal lengthening

The largest effect in Table 1 is that of pre-pausal position. Pre-pausal vowels are, as expected, longer than vowels not adjacent to the pause. This effect is the only positional effect observed in the corpus; there is not an independent effect of a vowel's position within a word.

These results are congruent with previous work on the effect of utterance position on duration, though the effect in Enenlhet is smaller than has been reported for other languages (e.g., [9, 10]). Unlike studies which find word-final lengthening as well as utterance-final lengthening [13] or progressive lengthening over the last several syllables in an utterance [12], utterance-final lengthening in Enenlhet is restricted to the final vowel in the utterance-final word.

5.2. Stress

There is no independent word position effect in the corpus, and therefore no evidence for fixed stress. Future research is needed to determine if Enenlhet lacks word-level stress entirely. Alternatively, stress may be lexically specified, variable based on word class, or some morphemes may fall outside the stress domain and therefore obscure stress in this analysis.

5.3. Consonant voicing

The effect of consonant voicing is consistent with results from other languages. Enenlhet vowels are significantly longer before voiced consonants. The negative interaction between syllable structure and consonant voicing suggests that this effect is primarily limited to closed syllables, when the following consonant forms a coda.

Consistent with studies that find that other duration effects are not strictly additive [25, 35], the effect of coda consonant voicing is restricted in pre-pausal syllables. This finding suggests a potential global limit to the amount of utterance-final lengthening which is permitted.

Because Enenlhet does not make a voicing distinction in the obstruent class, voicing also splits the obstruents from the sonorants. However, there may be differences in consonant behaviour dependent on manner of articulation (e.g., nasals vs. approximants or stops vs. fricatives) akin to findings from other languages [25, 34].

5.4. Syllable structure

As shown in Table 3, there is a significant effect of syllable structure. As expected based on other languages, vowels in open syllables are longer than those in closed syllables. In addition to the relationship between syllable structure and consonant voicing, Table 1 shows an interaction between syllable structure and utterance position. In contrast to the effect of the voicing of the following consonant, the interaction between syllable structure and utterance position in Table 6 shows that open syllable lengthening is greater in final syllables than medial ones.

5.5. Vowel quality

Vowel quality has an unexpected effect on vowel duration. Pairwise comparison confirms that /e/ is significantly shorter than /a/, as expected based on the shorter inherent durations of mid vowels than low vowels. However, /e/ and /o/ are also significantly different, and /o/ and /a/ are not.

The estimated differences between /e/ and /o, a/ are much smaller than has been reported previously [24, 25], amounting to only about 8% of the mean duration of /e/. Even so, this finding suggests that /e/ is higher than /o/ and that /o/ and /a/ have about the same height. However, the average formant values for each speaker, shown in Figure 1, suggest that each vowel occupies a different area of the acoustic space.

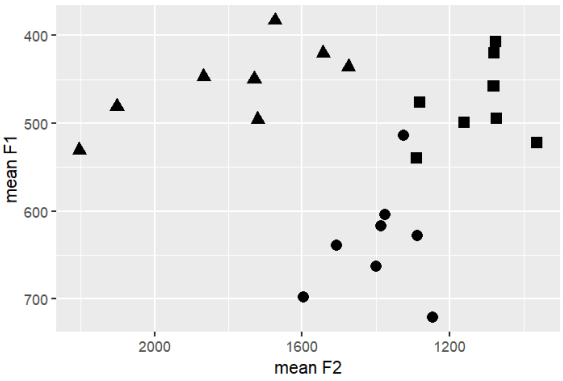


Figure 1: Mean F1 and F2 values for /a, e, o/ for each speaker (triangles: /e/, squares: /o/, circles: /a/)

For most speakers, the F1 values for /e/ and /o/ fall in the range of 400–500 Hz, and for all but one speaker, the F1 values for /a/ (600–700Hz) are higher than the means for the mid vowels. These mean F1 and F2 frequencies suggest that the vowel height cannot fully explain the duration difference between /e/ and /o/ and the similarity between /o/ and /a/.

5.5. Conclusion

This study describes vowel duration in Enenlhet, finding that vowels are longer utterance-finally, in open syllables, and before voiced consonants. The syllable structure effect is enhanced utterance-finally, while the effect of voiced consonants is limited. Lengthening due to voiced following consonants is greater in closed syllables than in open ones. Compared to other languages, the effects in Enenlhet are quite small, though still significant. Utterance-final lengthening, open syllable lengthening, and lengthening preceding voiced consonants are also found in unrelated languages, suggesting that they are relatively universal phenomena, albeit with language-specific implementations.

Vowel quality has an unexpected relationship to duration: /e/ is shorter than /a/ and /o/. Further study of the relationship between quality and duration is needed, as vowel height alone does not appear to explain this difference.

This study used a corpus of naturalistic speech, which is underutilized in phonetic research. Future work comparing these results with those from experimental work will be fruitful in gaining a greater understanding of how running speech differs from careful production.

6. ACKNOWLEDGEMENTS

I would like to thank Scott Myers, Pattie Epps, Elizabeth Wood, Raina Heaton, and the members of the UT Austin research group in phonetics and phonology; their valuable comments have substantially improved this analysis. This research was supported by the National Science Foundation (Doctoral Dissertation Research Improvement Grant No. 2024000). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect views of the National Science Foundation. All remaining errors are my own.

7. REFERENCES

[1] Enenlhet. 2023. *Catalogue of Endangered Languages*. University of Hawai'i at Manoa. https://endangeredlanguages.com/lang/2141

[2] Elliott, J. 2016. For bilinguals, Enxet vowel spaces smaller than Spanish. *J. Acoust. Soc. Am*. 140. 3107.

[3] van Gysel, J. 2022. The influence of language shift on Sanapaná vowels: An exemplar based perspective. *Lang. Var. and Change* 34. 1–29.

[4] Unruh, E. & Kalisch, H. 2003. Enlhet-Enenlhet: Una familia lingüística chaqueña. *Thule* 14/15, 207–231.

[5] Elliott, J. 2021. *A grammar of Enxet Sur*, Manoa.

[6] Heaton, R. 2019. Enenlhet Documentation. *The Archive of the Indigenous Languages of Latin America.* https://ailla.utexas.org. PID: ailla:266555.

[7] DiCanio, C., Whalen, D. 2015. The interaction of vowel length and speech style in an Arapaho speech corpous. *Proc. 18th ICPhS* Glasgow.

[8] DiCanio, C., Nam, H., Amith, J., Castillo García, R., Whalen, D. 2015. Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. *J. Phon*. 48, 45–59.

[9] Oller, D. 1973. The effect of position in utterance on speech segment duration in English. *J. Acoust. Soc. Am.* 54. 1235–1247.

[10] Berkovits, R. 1994. Durational effects in final lengthening, gapping, and contrastive stress. *Lang. and Speech* 37. 237–250.

[11] Hockey, B. Faygal, Z. 1999. Phonemic length and pre-boundary lengthening: An experimental investigation of the use of durational cues in Hungarian. *Proc. 14th ICPhS* San Francisco. 313–316.

[12] Nakai, S., Kunnari, S., Turk, A., Suomi, K., Ylitalo, R. 2009. Utterance-final lengthening and quantity in Northern Finnish. *J. Phon.* 37. 29–45.

[13] Beckman, M., Edwards, J. 1990. Lengthenings and shortenings and the nature of prosodic constituency. *Between the grammar and physics of speech.* Cambridge.

[14] Fry, D. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am*. 27. 765–768.

[15] Lindblom, B., Rapp, K. 1973. *Some temporal regularities of spoken Swedish*, Stockholm

[16] Ortega-Llebaria, M., Prieto, P. 2011. Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Lang. and Speech* 54. 73–97.

[17] Gordon, M., Roettger, T. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Lin. Vanguard.* https://doi.org/10.1515/lingvan-2017-0007

[18] Fintoft, K. 1961. The duration of some Norwegian speech sounds. *Phonetica* 7. 19–39.

[19] Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22. 129–159.

[20] Mitleb, F. 1984. Voicing effect on vowel duration is not an absolute universal. *J. Phon.* 12. 23–7.

[21] Benguerel, A. 1971. Duration of French vowels in unemphatic stress. *Lang. and Speech* 14. 383–391.

[22] Maddieson, I. 1985. Phonetic cues to syllabification. In: Fromkin, V. (ed.) *Phonetic linguistics: Essays in honor of Peter Ladefoged*, Orlando. 203–221.

[23] Lippus, P., Asu, E., Teras, P., Tuisk, T. 2013. Quantity-related variation of duration, pitch, and vowel quality in spontaneous Estonian. *J. Phon.* 41. 17–28.

[24] Klatt, D. 1975. Vowel duration is syntactically determined in connected discourse. *J. Phon.* 3. 129–40.

[25] O'Shaughnessy, D. 1981. A study of French vowel and consonant durations. *J. Phon*. 9. 385–406.

[26] Esposito, A. 2001. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica.* 59. 197–231.

[27] Boersma, P., Weenink, D. 2019. *Praat: Doing phonetics by computer* [computer program]. http://praat.org. Version 6.04.49.

[28] Goldman, J. 2011. Easyalign: An automatic phonetic alignment tool under Praat [Praat plugin]. *Proc. Interspeech* Firenze. http://latcui.unige.ch/phonetique/easyalign.php. Version 04.2012.

[29] van Gysel, J. 2017. *Temporal predicative particles in Sanapaná and the Enlhet-Enenlhet language family (Paraguay): A descriptive and comparative study,* Leiden.

[30] van Rossum, G., Drake, F. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

[31] R core team. 2019. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

[32] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Statistical Software* 67. 1–48.

[33] Lenth, R. V. 2022. *Estimated marginal means, aka least-squares means.* R package version 1.8.2. https://CRAN.R-project.org/package=emmeans.

[34] Begus, G. 2017. Effects of ejective stops on preceding vowel duration. *J. Acoust. Soc. Am*. 142. 2168–2184.

[35] Klatt, D. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am.* 59.5. 1208–21. https://doi.org/10.1121/1.380986.

A NASALANCE-BASED STUDY OF THE /h/ VS. /h̃/ OPPOSITION IN ZUBEROAN BASQUE

Ander Egurtzegi,[1,2] Andrea García-Covelo[2,3,4] & Iñigo Urrestarazu-Porta[1,2,4,5]

[1]CNRS, [2]IKER-UMR5478, [3]IPS-LMU, [4]UPPA & [5]UPV/EHU
ander.egurtzegi@iker.cnrs.fr, andrea.garcia@phonetik.uni-muenchen.de & inigo.urrestarazu-porta@iker.cnrs.fr

ABSTRACT

Eastern varieties of Basque have been described with an extremely rare opposition between an oral /h/ and a nasalized aspirate /h̃/ (1), which some researchers considered impossible (see (2)). This paper presents the first nasalance-based study of the /h/ vs. /h̃/ contrast in Basque, with data from the endangered Zuberoan variety.

We report the production of a reading task by 5 participants from the village of Larraine (4 male, 1 female; age range 60-70) including items with nasalized and oral aspirates (e.g. *ih̃ue* 'no one' vs. *aihai* 'dinner'), with oral aspirates comprising both words with only oral segments and words with a nasal stop that might trigger nasal assimilation (3).

Our results suggest that the /h/ vs. /h̃/ opposition is still present in Larraine Basque, although some lexical items sporadically lost nasality, and some speakers have completely merged both segments by consistently producing /h/.

Keywords: Nasal aspiration; nasalance; rare sound patterns; endangered languages; Zuberoan Basque

1. INTRODUCTION

Two endangered varieties of Basque (basq1248), Zuberoan (1) and Mixean (4; 5), show evidence of a scarcely attested phonological opposition between an oral aspirate /h/ and a nasalized /h̃/ (1; 3; 6). This opposition has been proposed as a potential analysis in only a handful of languages, and some authors deemed it theoretically impossible (2). In the Basque varieties that show it, some aspirates are produced with audible nasalization that spans a whole [VCV] sequence, resulting in sequences phonetically realized as [Ṽh̃Ṽ] (7; 8), where nasality is phonologically analyzed as originating in /h̃/ and then spreading to the surrounding vowels (1; 3). However, due to the difficulty of obtaining recordings of these endangered varieties, there are very few studies that present phonetic evidence of the /h/ vs. /h̃/ opposition (9). This study is the first to present nasalance data of this contrast, with recordings of speakers from Larraine (Larrau in French), a small village in the South of Zuberoa.

1.1. Zuberoan Basque

The historical province of Zuberoa (Soule in French) is located in the South-West of France, in the Pyrénées-Atlantiques department, and the north-eastern corner of the Basque Country. Its population is in clear recession, from 15350 in 2006 to 12716 as of 2019 (10). Zuberoa is formed by 43 villages, the main town being Maule-Lextarre (Mauléon-Licharre, 2947 inhabitants in 2019). The village of Larraine (195 inhabitants in 2019), where we recorded, is in the southern end of the province, bordering with Spain.

Zuberoan is usually considered the most deviant variety of Basque. It has historically been in contact with Bearnese Gascon to the North and Spanish to the South, but the importance of French has increased exponentially in the last century, severely threatening this variety, which is currently highly endangered. As a result, the number of speakers of Zuberoan Basque is much lower than the population of the area; the passive knowledge of the language being estimated at 22.8%, and the effective usage at 5.8%. In addition, its speakers are among the oldest population, making the study of this variety very pressing.

1.2. Typological background

The oral glottal approximant /h/ can be found in most of the world's languages, being part of 88% (279/317) of the inventories in the UPSID data-base (11) and 56% (1703/3020) of the inventories in the PHOIBLE data-base (12). On the other hand, the nasalized glottal approximant /h̃/, its nasal counterpart, while not unheard of, is by no means equally common, being attested in just a handful of languages including Krim (Niger-Congo, bomk1234), Lisu (Tibeto-Burman, lisu1250) and Pirahã (isolated, Amazon, pira1253). This rareness

is reflected in the fact that its cell is shaded in gray on the IPA chart, a feature used to signal impossible segments.

The impossibility attributed to /h̃/ is rooted in the aerodynamic definition of nasality; whether or not enough air-stream can go through the nasal cavity once it has produced glottal friction and after it has been divided between the nasal and the oral tract. However, nasalized aspirates are not as problematic from an articulatory perspective, where any sound produced with a lowered velum can be considered nasal (2).

From the fact that /h̃/ is an uncommon segment follows that the phonological opposition between /h/ and /h̃/ is exceedingly rare, the set of the languages that include /h/ and /h̃/ being a small subset of the languages that include /h̃/. In fact, a phonological opposition between /h/ and /h̃/ has only been proposed for 3-4 languages: kwangali (kwan1273) and thimbukushu (a.k.a. mbukushu, mbuk1240) both Bantu languages from northern Namibia (13), Seimat (seim1238) an Oceanic language from the Admiralty Islands (14), and Basque. Thanks to recent research, Basque is now arguably the clearest example of this opposition (15).

1.3. Previous research

Early work on Zuberoan Basque by Larrasquet (7) described a number of vowel-aspirate-vowel sequences that were nasalized throughout (now transcribed as [Ṽh̃Ṽ], see (16)), in contrast to other comparable [VhV] sequences which were not nasalized. The initial phonological analysis for this sequence of sounds assumed that both vowels were underlyingly nasal, while /h/ was similar to other oral /h/s in the language (17). This analysis was likely conditioned by the typological rarity of the /h/ vs. /h̃/ opposition. Later on, Hualde proposed a different analysis (1), namely that /h̃/ was phonologically nasal, and the surrounding vowels were only phonetically nasalized, given that in Basque nasality spreads to all vowels in contact with a nasal consonant (18). Recent research has found strong phonological evidence in support of this proposal (3). In parallel, historical research has established intervocalic *n as the segment that gave rise to /h̃/ (17; 19). However, phonetic research on this opposition is still scarce. There is only one acoustic analysis of the aspirate opposition in Mixean Basque (9), and only a handful of illustrative spectrograms from Zuberoan speakers can be found in the published literature thus far (16; 6). This paper aims to start filling this gap by means of a quantitative study of nasalance data recorded from speakers from the Zuberoan village of Larraine.

2. EXPERIMENT

We set up a field-lab at a house that was rented in Larraine. The local participants performed an elicitation task mostly including words containing an aspirate and we recorded their productions using a nasalance device or nasometer.

2.1. Participants

We recorded the utterances of 5 volunteer Larraine-Basque speakers (4 male, 1 female; mean age 64.6, range 60-70). Participant profiles were self-reported: they were L1 Larraine-Basque speakers; they were born and grew up in Larraine; they started to learn French (L2) at school at age 5; some left the Basque Country for work, but returned to town later on. To our knowledge, there were no direct family bounds between them.

2.2. Stimuli

Stimuli consisted of isolated words taken from Larrasquet's lexicon (8). We selected 12 words with initial /h/ (e.g. *hiru* 'three'), and around 90 with medial aspirates: 17 with etymologically oral /h/ (e.g. *soho* 'cropland' from Lat. *solu*(*m*)), 43 with etymologically nasalized /h̃/ (e.g. *ah̃ate* 'duck' from Lat. *anate*(*m*)), 7 with an aspirate adjacent to a nasal stop (e.g. *lanho* 'fog'), and 23 including an aspirate and a nasal stop in the same word, which may cause long distance nasalization (e.g. *nahi* 'will' and *ihun* 'nowhere'). Crucially, aspirates were all prompted written with <h>, irrespective of their etymology and alleged phonological status. Thus, any degree of nasalization in the uttered aspirates should be attributed to the participants' intended pronunciation. Finally, we selected another 56 words with no aspirates as fillers. Difficulties came from the fact that some participants were not used to reading. In addition, there were slight divergences in some words between our source (8) and the current local pronunciation, so we included a French translation in the prompts. The translation helped participants understand the target words.

2.3. Procedure

The recordings were conducted using a Glottal Enterprises nasalance with a separator handle (NAS-1 SEP)[1], which consists of two microphones separated by a wooden plate. This plate allows the separation of the acoustic signal coming from

the mouth and the nose. The nasometer stood in a tripod, which was adjusted to a suitable height for each participant, so that they could sit comfortably while the nasometer was pressed against their upper lip. *SpeechRecorder* software was used in order to present the stimuli in a 32 inch TV screen and record the productions. The order of the stimuli was randomized for each participant. The recording started automatically 500 ms after each stimulus was presented and was manually stopped once the participant had read out loud the target word or stated it was unknown to them.

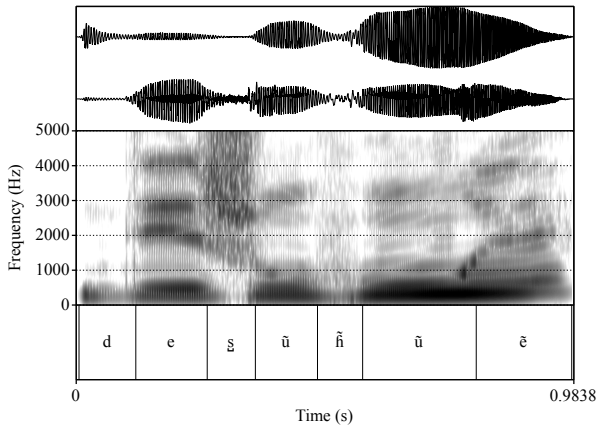


Figure 1: Waveforms and spectrogram of *desuhue* 'dishonor' as produced by speaker 2

2.4. Analysis

The stereo nasalance data was analyzed using Praat. Fig. 1 shows an example of the production of /h̃/. The top waveform corresponds to the nasal channel, while the bottom waveform corresponds to the oral one. For the acoustic analysis, both the nasal and oral channels were band-pass filtered (80 Hz-10000 Hz) and two measurements describing the degree of nasalization were computed: 1. the mean-corrected nasal amplitude (nasal amplitude divided by the mean intensity of the token); and 2. the nasalance (ratio of the nasal amplitude to the sum of the oral and nasal amplitudes, i.e. $A_n/(A_o + A_n) \times 100$) (20; 21). As the description of nasalized aspirates using nasometry is an unexplored topic, the suitability of both metrics for the available data was tested. All measurements of both metrics were then z-transformed (grouped by speaker), and median values of the z-scores of all observations were computed for each aspirate. All recorded words were coded by the etymological category of the aspirate included in them, given that Basque only allows for one aspirated segment in each word (17). A first coding was made following Larrasquet's transcriptions (8), but that needed to be revised, as the variety described by Larrasquet is from the North of Zuberoa and it was spoken 85 years ago.

Two different Bayesian Gaussian models were fitted in *R* using the *brms* interface to *Stan*. Model 1 included nasalance as a response, while Model 2 included mean-corrected nasal amplitude instead. In both cases, etymological category and trial (which was previously min-max normalized) were used as predictors. Speaker and word were added as random effects (including correlated varying intercept and slope) to both models. Both models were fitted in eight chains of 10000 iterations each (including 5000 warm-up iterations). We used *Student's t* distributed priors (df = 5, μ = 0, σ = 2) for all predictors and the intercept, with default priors from *brms* for the random effects and correlation parameter. In both cases, we had to adapt the delta to 0.999 and the maximum tree depth to 12 to ensure correct convergence (no issues, all $\hat{R}$ = 1.00).

2.5. Results

Fig. 2 shows the posterior predictions for Model 1, with nasalance as the response. The posterior distribution plot shows that the median values and overall distributions of the etymologically nasalized /h̃/s and the /h/s nasalized after assimilation are very similar in our data, while the distribution of the oral /h/s falls to lower nasalance values. The main results of the model are in the table in Fig. 2.

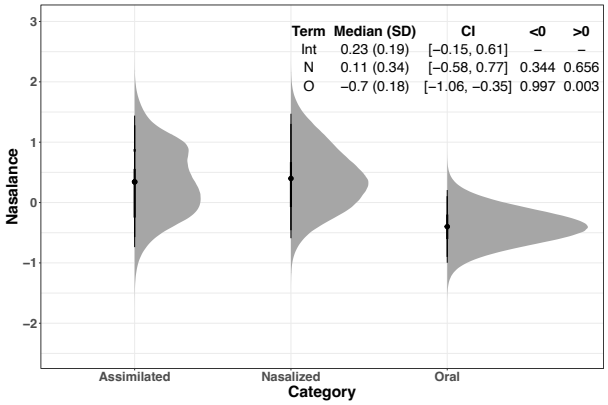


Figure 2: Posterior probability distribution of nasalance by aspirate category

Taking the factor level with the category "assimilated" as the intercept, the "etymologically nasalized" category had a lower 95% credible interval (CI) of -0.58 and an upper 95% CI of

0.77, with a comparable span over both sides of zero, which can be interpreted as the categories "etymologically nasalized" and "nasalized due to assimilation" being quite similar in our data. On the other hand, the oral category had a lower 95% CI of -1.06 and an upper 95% CI of -0.35, showing little overlap with the two nasalized categories and not crossing over 0, which suggests that the "oral" category is less nasalized than the two nasalized categories in our data. The results from model 2 (μ-corrected A_n) were very similar.

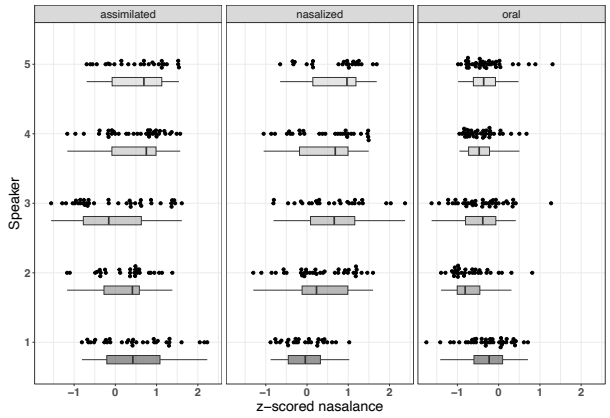


Figure 3: Nasalance by aspirate category

3. DISCUSSION

Our models shows a difference between etymologically nasalized and etymologically oral aspirates in the Zuberoan variety of Larraine. Nonetheless, there are multiple factors that need to be taken into account in the interpretation of these results. First, not all speakers show the same degree of nasalization: while some speakers consistently nasalize the aspirates in the relevant lexical items, others do not produce /h̃/ consistently, if at all. This can be seen in Fig. 3, which presents the nasalance values of the aspirate segments by aspirate category. In parallel to the model results in Fig. 2, the values for assimilated and etymologically nasalized tokens are quite similar, whereas there is a substantial difference between etymologically nasalized and etymologically oral aspirate tokens. Fig. 3 points to big inter-speaker differences: while speaker 1 produced aspirates in all categories with a low nasalance level, the two aspirate groups (nasalized and assimilated vs. oral) are clearly differentiated in speakers 2 or 5.

Second, not all etymologically nasalized aspirates in Larrasquet's lexicon (8) are in fact nasalized in modern-day Larraine. Fig. 4 shows the mean

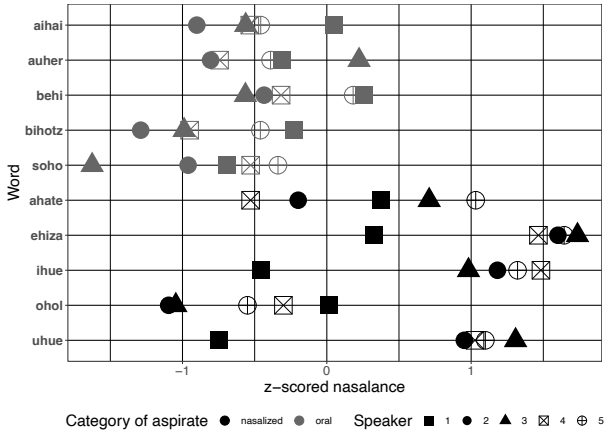


Figure 4: Mean speaker nasalance in example words including an aspirate of oral vs. nasal origin

nasalance value of the aspirate included in ten different lexical roots (i.e. words with the same root but potentially with different declension). The first five roots include etymologically oral /h/s, while the bottom five include aspirates that Larrasquet transcribed as nasal. Each symbol represents the production of a given speaker. When a speaker produced more than one token of the same lexical root, the nasalance values were averaged for this plot. As can be observed in Fig. 4, some nasalized aspirates are indeed produced with a high nasalance value by most speakers (with the recurring exception of speaker 1), such as *eh̃iza* 'hunt', *ih̃ue* 'no one' or *uh̃ue* 'honor', but other tokens are less often nasalized, such as the classic example of /h̃/ *ah̃ate* 'duck', or not at all, as is the case of *ohol* 'board'. In contrast, all aspirates characterized as oral have consistently low nasalance values.

4. CONCLUSION

This research presents evidence that the typologically rare opposition between nasalized and oral aspirates is still present in the Zuberoan Basque variety of Larraine. However, it seems that this contrast is being gradually neutralized as lexical items expected to be nasalized have lost their nasality and some speakers show loss of aspirate nasalization, even in the older generations.

The Zuberoan variety is severely endangered, so further data collection is more difficult by the day. This research reminds us that it is paramount to continue collecting data from endangered languages, given that they might be our only opportunity to witness the rarest phenomena of human language.

REFERENCES

[1] J. I. Hualde, "Topics in Souletin phonology," in *Generative Studies in Basque Linguistics*, J. I. Hualde and J. Ortiz de Urbina, Eds. Amsterdam: John Benjamins, 1993, pp. 289–327.

[2] R. Walker and G. K. Pullum, "Possible and impossible segments," *Language*, vol. 75, pp. 764–780, 1999.

[3] A. Egurtzegi, "On the phonemic status of nasalized /h̃/ in Modern Zuberoan Basque," *Linguistics*, vol. 56, pp. 1353–1367, 2018.

[4] I. Camino, *Amiküze eskualdeko heskuara [The Basque of the region of Amiküze (Mixe)] (Mendaur 11)*. Bilbao: Euskaltzaindia, 2016.

[5] A. Egurtzegi and C. Carignan, "An acoustic description of Mixean Basque," *The Journal of the Acoustical Society of America*, vol. 147, pp. 2791–2802, 2020.

[6] A. Egurtzegi, "/h̃/ hasperen sudurkarituaren inguruan [On the nasalized aspiration /h̃/]," *International Journal of Basque Linguistics and Philology (ASJU)*, vol. 57, 2023.

[7] J. Larrasquet, "Phonétique du basque de Larrajá (Quartier de barcus)," *Revista Internacional de Estudios Vascos*, vol. 23, pp. 153–191, 1932.

[8] ——, *Le basque de la Basse-Soule orientale*. Paris: C. Klincksieck, 1939.

[9] A. Egurtzegi and C. Carignan, "A typological rarity: the /h/ vs. /h̃/ contrast of Mixean Basque," 2020, paper presented at LabPhon 17.

[10] "Populations légales en vigueur á compter du 1er janvier 2022," 2021, Available at: https://www.insee.fr/fr/statistiques/6011060, Accessed: 2022-12-26.

[11] I. Maddieson, *Patterns of sounds*. Cambridge: Cambridge University Press, 1984.

[12] S. Moran and D. McCloy, "Phoible 2.0," http://phoible.org, 2019.

[13] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Oxford: Blackwell, 1996.

[14] R. Blust, *The Austronesian languages (rev. ed.)*. Canberra: Asia-Pacific Linguistics Open Access, 2013.

[15] J. Blevins and A. Egurtzegi, "Refining explanation in evolutionary phonology: Macro-typologies and targeted typologies in action," *Linguistic Typology*, vol. 27, 2023.

[16] A. Egurtzegi, "Diferentes tipos de aspiración en vasco (con análisis espectrales del dialecto suletino actual)," in *Iberia e Sardegna. Legami linguistici, archeologici e genetici dal Mesolitico all'Età del Bronzo*, E. Blasco, P. Francalacci, A. Nocentini, and G. Tanda, Eds. Firenze: Le Monnier, 2013, pp. 151–169.

[17] L. Michelena, "Fonética histórica vasca," in *Obras completas 6*, J. A. Lakarra and I. Ruiz Arzalluz, Eds. Donostia-San Sebastián & Vitoria-Gasteiz: Diputación Foral de Guipúzcoa & University of the Basque Country, 2011.

[18] J. I. Hualde, "Segmental phonology," in *A grammar of Basque*, J. I. Hualde and J. Ortiz de Urbina, Eds. Berlin: Mouton de Gruyter, 2003, pp. 16–65.

[19] I. Igartua, "Diachronic effects of rhinoglottophilia, symmetries in sound change, and the curious case of Basque," *Studies in Language*, vol. 39, pp. 635–663, 2015.

[20] C. Carignan, "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 143, pp. 2588–2601, 2018.

[21] ——, "A practical method of estimating the time-varying degree of vowel nasalization from acoustic features," *The Journal of the Acoustical Society of America*, vol. 149, pp. 911–922, 2021.

[1] We thank the town of Larraine and the first *Basque Summer Tutorial in Language Documentation* for providing us with the setting for this research, Marianne Pouplier for lending us her nasalance device and for her invaluable insights on Praat scripting, Roy Alderton for his help with *SpeechRecorder*, and Shelece Easterday, Aitor Egurtzegi and Francesco Rodriquez for their readiness to answer our questions. All errors are ours. This research has been partially funded by the ANR project MADPAB (ANR-20-CE27-0007).

Pre-Pausal Glottalization in Chicontepec Nahuatl

Andrés Ehecatl Aguilar

Department of Chicana and Chicano Studies, San Diego State University, San Diego, California, USA
aaguilar26@sdsu.edu

ABSTRACT

This paper analyzes the distribution of glottalization associated with pauses in Chicontepec Nahuatl (ChN), a variety of Eastern Huasteca Nahuatl, which is a Uto-Aztecan language spoken in Veracruz Mexico. This paper contributes to the literature on glottal effects at prosodic boundaries by describing a pattern in ChN that is distinct from other documented phenomena, such as phrase-final creak. This glottalization is analyzed as a [+constricted glottis] suprasegmental feature that marks the end of phrases. Data gathered in the field suggest that the realization of utterance-final glottalization is sensitive to the phonological status of the final consonants. In addition, for a subset of speakers, there is a pattern of phrase-final devoicing, such that glottalization moves leftward to the nearest voiced syllable, suggesting that it is a suprasegmental feature associated with prosodic boundaries.

Keywords: Nahuatl, utterance-final glottalization, suprasegment, glottal stop.

1. INTRODUCTION

Glottal effects at prosodic boundaries, including phenomena such as prosodic strengthening, phrase-initial glottal stops, phrase-final creak, phrase-final boundary tone truncation by devoicing, etc., have been of interest in the literature. In this paper I describe a distinct glottal effect found in Chicontepec Nahuatl (ChN): glottalization (realized as both glottal stops and voiced glottalization) that is associated with pauses. I analyze these glottal stops as a [+constricted glottis] suprasegmental feature that marks the end of a prosodic unit. While pausal glottalization is found both utterance-medially and finally, this paper focuses on the utterance-final context. I first establish the canonical realization of utterance-final pausal glottalization as a glottal stop after vowel-final words. Then, I show how pausal glottalization is realized when cooccurring with utterance-final consonants, and show that the realization of this [+constricted glottis] feature is sensitive to phonological status. Finally, I describe a pattern in which utterance-final pausal glottalization moves left to the nearest voiced syllable, suggesting that this type of glottalization is a suprasegmental feature associated with a prosodic unit.

Nahuatl is the Southernmost Uto-Aztecan language with varieties spoken from west and central Mexico, as far south as El Salvador, and in diaspora communities across the United States. It is spoken by roughly 1,651,958 speakers in Mexico [1]. ChN is a variety of Eastern Huasteca Nahuatl, spoken in Chicontepec de Tejeda, Veracruz and surrounding communities by about 32,428 speakers [2].

This description is based on data gathered through fieldwork conducted with ten speakers (six female and four male) in Zacatecas, Mexico. They are native Nahuatl speakers who grew up in the municipality of Chicontepec. All speakers were between the ages of 25-45. The sentences that were analyzed in this study were drawn from a combination of data gathered through elicitation and directed tasks using carrier sentences, as well as natural conversations and narrative texts. Data were recorded in the field in mono with a Marantz PMD660 solid-state recorder and an Audio-Technica AT897 shotgun microphone.

1.1. Laryngeal Sounds in Nahuatl

Across Nahuatl varieties, there is typically one contrastive laryngeal sound. In Classical Nahuatl and other central varieties of Nahuatl, the contrastive sound is a glottal stop /ʔ/. In many other varieties of Nahuatl, including ChN, the corresponding laryngeal sound is the glottal fricative /h/, such that a word like 'medicine' will be something like [paʔt͡ɬi] in some varieties and [paht͡ɬi] in others. ChN is a laryngeally rich language in that it has contrastive /h/, as well as a number of phonologically derived surface glottal fricatives from underlying coda /w/, word-final nasals, and /k/ in a number of contexts [3]. In ChN, glottal stop is not a contrastive sound; however, glottalization is widely distributed in a number of contexts: prefix-stem boundary, stem-stem boundaries, hiatus contexts and in pre-pausal contexts. In this paper, I focus on the pre-pausal context for glottalization.

2. PRE-PAUSAL GLOTTALIZATION

There is a regular pattern in which glottalization aligns with pauses in ChN. A pause is defined as a period of silence, a perceptible and physical break in the speech stream, that may or may not align with other prosodic boundaries. Utterance-medially, this glottalization is typically realized as a canonical

phonetic glottal stop in which glottalization can be seen both leading into the pause and out of the pause. The glottal stop is thus the period of silence. Utterance-finally, this pausal glottalization is realized as a canonical glottal stop after vowel-final words, with some amount of noticeable irregular pulsing leading to a period of silence, followed by some degree of breathy noise in the release of the glottal stop. Because neither a detailed intonational model for, nor a syntactic analysis of ChN is yet available, I focus on utterance-final pauses which necessarily align with two prosodic boundaries: a *phrase* and the *utterance*. The *utterance* is defined as a breath group that is a phonetic string that is bounded by breaths.

Utterance-final glottalization is a robust pattern. In particular, if the utterance-final word is vowel-final, all utterances trigger a glottal stop. The waveforms and spectrograms in Figures 1–2 show two equivalent sentences: one in which there is noun incorporation and the utterance ends verb-finally (Figure 1), and one in which the utterance ends in an unincorporated noun (Figure 2). Glottalization occurs after both final vowels, demonstrating that glottalization is not a feature of a particular lexical item, but rather an articulation associated with the end of an utterance.

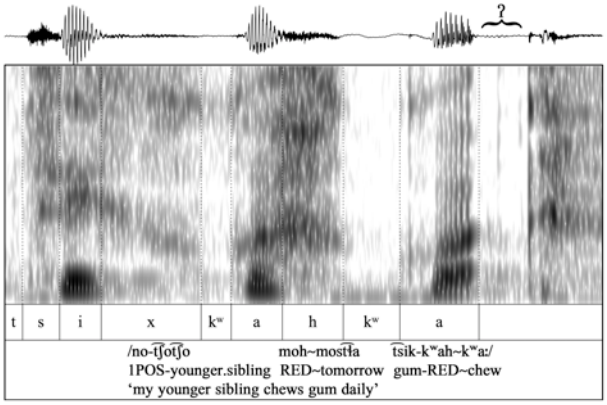


Figure 1: Utterance-final glottalization, as marked by "?" over the waveform, on final verb[1]

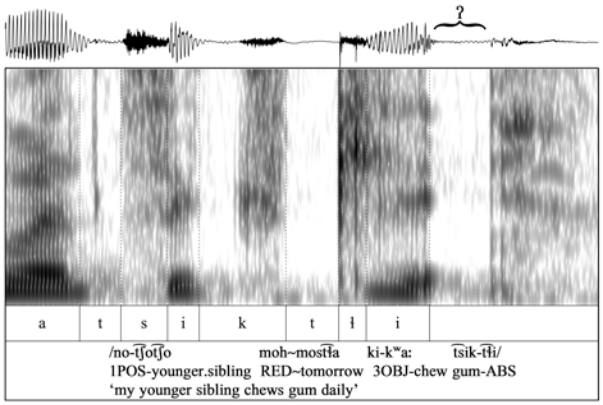


Figure 2: Utterance-final glottalization, as marked by "?" over the waveform, on final noun

2.1. Utterance-final consonants

Glottalization has been shown to have variable realization across and within languages [4, 5, 6]. As expected, the phonetic realization of glottalization in ChN ranges from canonical glottal stops to varying degrees of creaky phonation on adjacent voiced sounds. In the vowel-final context, glottalization is realized as a canonical glottal stop. In utterances that are consonant-final, glottalization is realized as creaky phonation. In this study, creaky phonation is identified visually in the spectrogram [7] as irregular pulsing, which is often accompanied by other features such as pitch-halving, or interruption of pitch tracking altogether. Figure 3 shows how glottalization is realized when the utterance-final word is consonant-final: creaky phonation during the final vowel.

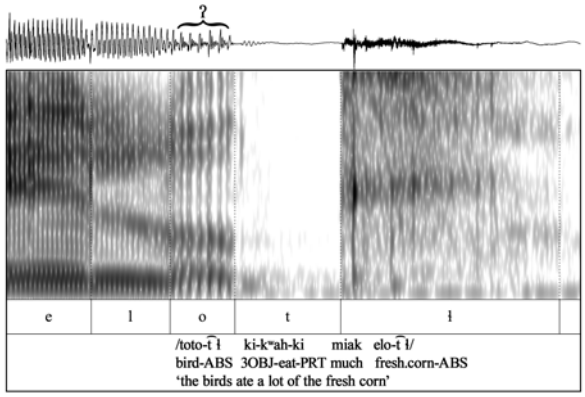


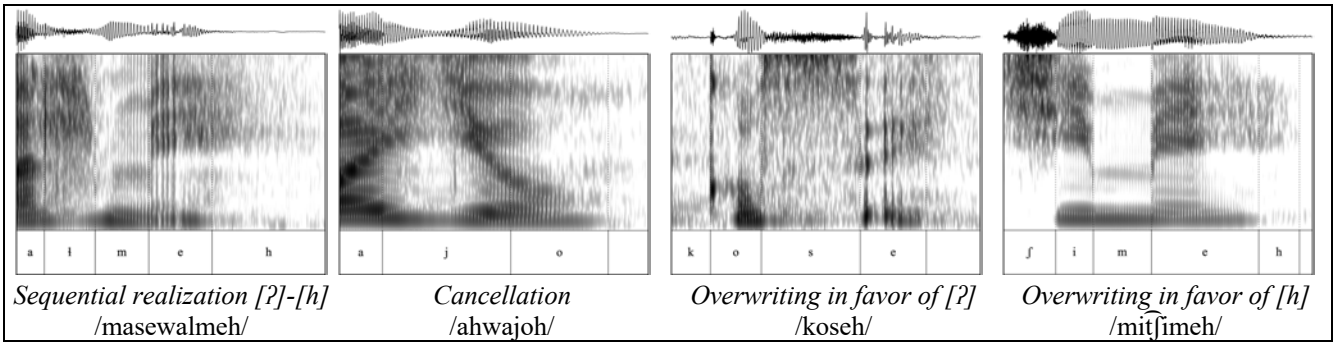
Figure 3: Utterance-final glottalization realized as creaky phonation on C-final word

While it is articulatorily possible to realize a glottal stop after or during the final consonant, for utterances ending in a consonant, glottalization instead is realized on the final vowel of the utterance, perhaps because here it would be more acoustically salient. This supports an analysis that glottalization is a floating feature associated with the end of a phrase or utterance, and that it must anchor to a final vowel.

The occurrence of glottalization in this context is much more variable than vowel-finally. I attribute this to perhaps differences in intonational structure or as an artifact of the methodologies employed. While glottalization seems to be absent in this context in many tokens gathered, it is realized overtly much more regularly in the naturalistic data from conversations.

2.2. Utterance-final [h] and phonological sensitivity

The distribution of /h/ in ChN is such that it almost exclusively occurs in the coda position. This is an interesting distribution since it has been shown that the glottal fricative /h/ is more perceptually salient in the onset position and is not expected to have such a



Sequential realization [ʔ]-[h] /masewalmeh/ | *Cancellation* /ahwajoh/ | *Overwriting in favor of [ʔ]* /koseh/ | *Overwriting in favor of [h]* /mit͡ʃimeh/

Figure 4: Realizations of utterance-final glottalization with final [h]

distribution [8]. In addition, word-final [h] not only corresponds to the contrastive underlying /h/ in the inventory, but also to nasals (/m/ and /n/) and the labiovelar glide /w/. This is because phonological processes of neutralization target word-final nasals and coda /w/s.

Utterance-final glottal fricatives provide an interesting context for the realization of utterance-final glottalization since, following Ladefoged's Continuum model of phonation [9, 10, 11], [h] and [ʔ] are articulatorily opposites. However, in ChN, both occur utterance-finally, in a "stacked" prosodic position (glottalization at the utterance-level, [h] word-level). Possible outcomes for the realization of both the word-final [h] and utterance-final glottalization (schematized here as [ʔ]) include: 1) sequential realization [ʔ]-[h]; 2) sequential realization in the reverse order [h]-[ʔ]); 3) cancellation such that neither is realized; 4) overwriting in favor of [ʔ] (/h/→Ø); and 5) overwriting in favor of [h] (/ʔ/→Ø). A survey of 554 tokens with final [h]s from underlying /h/, /n, m/, and /w/ showed that surface realization was variable. Note that nasal-final words are less frequent than /h/-final words, and /w/-final words are even less frequent, which is reflected in the number of tokens available to survey. Table 1 summarizes the distribution of realizations by underlying representation (UR).

	/h/	**/N/**	**/w/**
Sequential [ʔ]-[h]	58 [.20]	27 [.15]	17 [.19]
Sequential [h]-[ʔ]	0	0	0
Cancellation	45 [.16]	15 [.09]	3 [.03]
Overwriting [ʔ] (/h/→Ø)	20 [.07]	6 [.03]	3 [.03]
Overwriting [h] (/ʔ/→Ø)	165 [.57]	128 [.73]	66 [.73]
Total	288	176	90

Table 1 Realization of [h] and [ʔ] by UR. Proportion of total tokens in brackets.

The distributions of realizations for overt glottalization and glottal fricatives by underlying representation all follow the same general trends. Glottalization is more often deleted than realized overtly in the context of [h] (overwriting /ʔ/→Ø). If glottalization surfaces, the most likely realization is for it to occur *sequentially* in a [ʔ]-[h] sequence as seen for other utterance-final consonants (e.g., Figure 3), though there is a less frequent pattern of *overwriting* of [h] in favor of [ʔ]. Finally, *cancellation* occurs less than sequential realization. Note that the contrast between /Vh/ and final /V/ is maintained, as the latter are always realized with glottalization as shown in Figures 1-2. Examples of each type of attested surface realization are given in Figure 4, all with words with underlying final /h/. There were no cases of glottalization occurring after the glottal fricative.

Looking at overt glottalization only, there appears to be a subtle difference in the realization of glottalization by UR, suggesting that the phonetic implementation of the stacked final [h] and utterance-final glottalization is sensitive to phonological status. It can be argued that [ʔ] is realized more strongly when co-occurring with /h/ than with /N/ and /w/: overwriting in favor of [ʔ] occurs twice as often for tokens with underlying /h/ than for /N/ and /w/; and, overwriting in favor of [h] occurs more often for /N/ and /w/ than for /h/. Cancellation also occurs at a higher rate for /h/ than for /N/ and /w/. Taken together, the distribution of [ʔ] and [h] realizations suggests that [h] that is underlyingly /h/ is more likely to be reduced compared to [h] that is underlyingly /N/ or /w/. In other words, [h] that is underlyingly nasal or /w/ is more persistent. The persistence of [h] that is derived from an underlying nasal or /w/, in light of a pressure to reduce in the context of [ʔ], in effect ensures that the derivational phonological information is not lost.

2.3. Utterance-final glottalization as a suprasegmental

That utterance-final glottalization can move inward from the edge supports the analysis that it is a floating [+constricted glottis] feature specified to occur near the end of an utterance. For a subset of speakers

(4/12), there is a marginal pattern of devoicing utterance-final syllables. This may be partially driven by coarticulation with voiceless consonants, declination, and/or perhaps the anticipation of the open-glottis gesture associated with an intake of breath [12, 13, 14]. When the final syllable is devoiced, glottalization is realized on the penultimate syllable rather than the final syllable. Figure 5 shows glottalization realized on the utterance-penultimate syllable when the final syllable is devoiced.

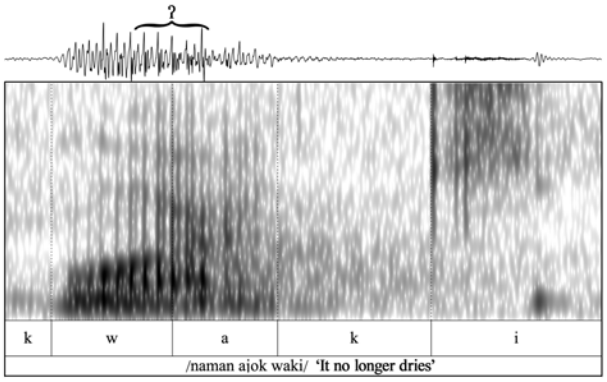


Figure 6: Glottalization when final syllable is devoiced

What is of interest here is that glottalization that marks the end of a phrase can be realized before the end of the segmental string. This suggests that this laryngeal target moves leftward to the nearest possible anchor point: a sonorant. This supports the analysis that this is a suprasegmental [+constricted glottis] associated with the utterance/phrasal boundary, since it can move in from the edge depending on the context.

This glottalization pattern in ChN parallels other suprasegmental phenomena. Devoicing has been shown to cause pitch-accent to shift in Japanese [15, 16]. Similarly, edge or boundary tones, which mark the ends of the prosodic phrases, have been shown to move and appear before the phrase edge or move to an epenthesized vowel when no voiced nucleus is available in Tashlhiyt Berber [17, 18]. Similar in position and behavior to Tashlhiyt Berber boundary tones, ChN glottalization behaves like a boundary target analogous to tone, but involving vocal fold constriction rather than stretching. The more conservative analysis taken here is of a floating [+constricted glottis] feature associated with utterance (or loosely-defined phrasal) boundaries. Figure 6 schematizes the movement of glottalization anchor points.

3. DISCUSSION AND CONCLUSION

In this paper I show how glottalization, realized as both canonical glottal stops and voiced glottalization, is associated with pre-pausal contexts, utterance-

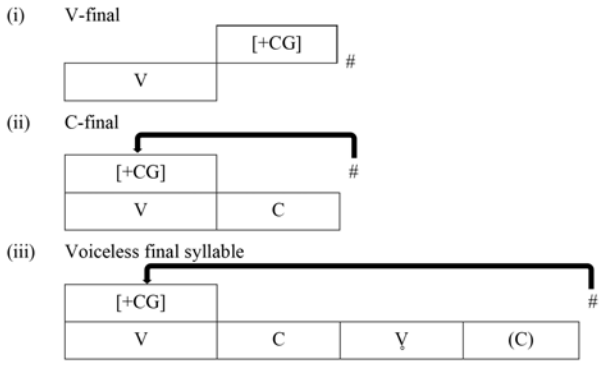


Figure 5: Anchoring points for suprasegmental [+CG]

medially and finally. I analyze utterance-final glottalization as a [+constricted glottis] suprasegmental feature that marks the end of a phrase given that it can move leftward from the edge to the nearest possible anchor point. The realization of this suprasegment is sensitive to the phonological status of the U-final consonant such that it is more strongly realized with utterance-final [h] that is underlyingly /h/ compared to [h] derived from underlying /N/ or /w/.

Laryngeal articulations associated with utterance boundaries, and presumably phrase boundaries, are not unprecedented. In Choguita Rarámuri, utterance-final vowels with falling tones are rearticulated, such that a word like /naʰpô/ 'prickly pear' surfaces as [naʰpóʔò] [19, 20, 21]. In Warekena, a glottal fricative with a copy vowel /-hṼ/ is reported in pre-pausal and utterance-final vowels [22]. In Cahuilla, words with "moveable" glottal stops are reported to occur most reliably in specific phrasal contexts such as phrase-finally and in isolation [23]. In Dagbani, glottal stops are reported to mark phrasal boundaries utterance-medially (pauses) and utterance finally [24]. Similarly in Lahu (Tibeto-Burman), a phrase-final glottal stop is associated with the imperative construction [25].

These patterns from diverse language families suggest that laryngeal articulations associated with prosodic boundaries, and specifically that are not physiological or aerodynamic in origin, may be more common than previously documented in the world's languages. More locally, this raises the question for future investigations of how representative this pattern is of other Nahuatl varieties.

This work is important for a number of reasons. First, it has implications for prosodic typology in that it has the potential to expand our definitions of "intonation" to include other laryngeal articulations beyond f0. Second, it is a first step towards an intonational analysis of ChN. Finally, it contributes to the limited literature on the phonetics of Nahuatl and other Uto-Aztecan and Indigenous American languages in general.

4. ACKNOWLEDGEMENTS

Miac tlazcamati, special thanks to my teachers, notlamachtianiwah: Catalina Cruz de la Cruz, Faustino Cruz de la Cruz, Sabina Cruz de la Cruz, Ofelia Cruz Morales, Victorino Francisco de la Cruz, Eduardo de la Cruz Cruz, Marcelina de la Cruz Cruz, Rosa de la Cruz Cruz, Delfina de la Cruz de la Cruz, my other Nahuatl teachers, and the Instituto de Docencia e Investigación Etnológica de Zacatecas. I am also grateful to Marc Garellek, Gabriela Caballero, and September Cowley for your helpful comments.

5. REFERENCES

[1] Hablantes de lengua indígena en México, Instituto Nacional de Estadística, Geografía e Informática (INEGI) https://cuentame.inegi.org.mx/poblacion/lindigena.aspx

[2] Sistema de Información Estadística y Geográfica del Estado de Veracruz de Ignacio de la Llave (SIEGVER). 2021. *Cuadernillos municipales 2021: Chicontepec. Veracruz, Mexico*.

[3] Aguilar, A.E., 2020. *Phonology and phonetics of laryngeal sounds in Chicontepec Nahuatl*. University of California, San Diego.

[4] Pierrehumbert, J., Talkin, D. 1992. Lenition of /h/and glottal stop. *Papers in laboratory phonology II*, 90–117.

[5] M. Garellek, Y. Chai, Y. Huang, and M. Van Doren, "Voicing of glottal consonants and non-modal vowels," Journal of the International Phonetic Association, pp. 1–28, 2021. doi: 10.1017/S0025100321000116

[6] Garellek, M., 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 101155. doi: 10.1016/j.wocn.2022.101155

[7] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org

[8] Silverman, D. 1997. *Phasing and recoverability*. Taylor & Francis.

[9] Ladefoged, P. 1971. *Preliminaries to linguistic phonetics. Chicago*: University of Chicago Press.

[10] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of phonetics* 29, 4, 383–406.

[11] Garellek, M. 2019. The phonetics of voice. In: Katz, W. F., Assmann, P. F. (eds), *The Routledge Handbook of Phonetics*. Routledge, 75–106.

[12] Ohala, J. J. 1983. The origin of sound patterns in vocal tract constraints. In: MacNeilage, P.F. (ed), *The production of speech*. Springer, New York, NY, 189–216.

[13] Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87, 2, 820–857.

[14] Slifka, J. .2006. Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of voice* 20, 2, 171–186.

[15] Vance, T. J. (1987). *An introduction to Japanese phonology*. SUNY Press.

[16] Maekawa, K. 1990. Production and perception of the accent in the consecutively devoiced syllables in Tokyo Japanese. *Proc. First International Conference on Spoken Language Processing*, 517–520.

[17] Grice, M., Roettger, T. B., Ridouane, R., Fougeron, C. 2011. The association of tones in Tashlhiyt Berber. *Proc. 17th ICPhS* Hong Kong, 775–778.

[18] Gordon, M., Nafi, L. 2012. Acoustic correlates of stress and pitch accent in Tashlhiyt Berber. *Journal of Phonetics* 40, 5, 706–724.

[19] Garellek, M., Aguilar, A., Caballero, G., Carroll, L. 2015. Lexical and Post-Lexical Tone in Choguita Rarámuri. *Proc. 18th ICPhS,* Glasgow, 10–14.

[20] Caballero, Gabriela. 2022. *A grammar of Choguita Rarámuri: In collaboration with Luz Elena León Ramírez, Sebastián Fuentes Holguín, Bertha Fuentes Loya and other Choguita Rarámuri language experts*. Berlin: Language Science Press. doi: 10.5281/zenodo.7189161

[21] G. Caballero, Y. Chai, and M. Garellek. 2022. Stress, tone, and intonation in Choguita Rarámuri. In Kubozono, H., Ito, J., and Mester, A. (eds), *Prosody and Prosodic Interfaces*. Oxford University Press, 227–248. doi: 10.1093/oso/9780198869740.003.0008.

[22] Aikhenvald, A. Y. 1998. Warekena. In: Derbyshire, D.C., Pullman, G.K., (eds), *Handbook of Amazonian languages vol. 4*, Berlin & New York: Mouton de Gruyter, 225–439.

[23] Seiler, H. 1965. Accent and morphophonemics in Cahuilla and in Uto-Aztecan. *International Journal of American Linguistics* 31, 1, 50–59.

[24] Hyman, L. M. 1988. The phonology of final glottal stops. *Proc. of WECOL*, 113–130.

[25] Matisoff, J. A. 1973. *The grammar of Lahu*. University of California Press.

[1] All spectrograms show up to 5000 Hz.

FRICATION AND FORMANT FREQUENCIES IN THE MUNDABLI HIGH VOWELS

Matthew Faytak[1], Bowei Shao[2], Angèle Douanla Taffre[3], and Nelson C. Tschonghongei[1]

[1] University at Buffalo, [2] École Normale Supérieure, Université PSL, [3] Université Yaoundé I
faytak@buffalo.edu, bowei.shao@ens.psl.eu, angeletaf@gmail.com, nelsonts@buffalo.edu

ABSTRACT

Mundabli (Yemne-Kimbi, Cameroon) is reported to contrast two sets of high vowels: extra-high /i u/ and high /ɪ ʊ/, by way of frication intrinsic to /i u/. In this study, we assess the role of aperiodicity (zero crossing rate, ZCR) and formant frequencies (midpoint F1-F2) in these contrasts. Analysis of the dynamics of ZCR in the vowels of interest using generalized additive mixed models shows elevated aperiodic energy early in the duration of /i u/ compared to /ɪ ʊ/, modulated by onset consonant type. Small, inconsistent differences in F1 and F2 are observed, and /i/ tends to exhibit lower F2 than /ɪ/, suggesting that this contrast is not simply one of height. These findings contribute to our understanding of the fricative vowels and their development from plain high vowels; and add to the literature on vowel contrasts involving frication.

Keywords: frication, fricative vowels, zero crossing rate, formant analysis, Mundabli

1. INTRODUCTION

1.1. Mundabli vowels in context

Mundabli (ISO 639-3: boe; autonym [ɲɔ̄ ⁿdʒân]) is a Yemne-Kimbi language spoken by 350-450 people from a single village on rugged terrain in the Lower Fungom area of northwestern Cameroon [30, 11]. Lower Fungom is notable for its linguistic-genetic diversity and intense multilingualism [11, 26], as well as the general degree of complexity of its languages' vowel systems [30, 11, 22, 24].

Mundabli is notable for exhibiting unusually close contrasts among its high vowels, which have been reported as cued entirely or in part by fricative noise. Voll [30, 39-41] describes the contrasts between the vowels /i, u/ and /ɪ, ʊ/ as extremely close in height, but quite reliably differing in frication: /i u/ are said to exhibit considerable fricative noise, and often cause delayed release, affrication, or trilling (in the case of /b/ > [bʁ]) of onset consonants (see Fig. 1). These extra-constricted vowels are also known to occur in neighboring languages of Lower Fungom such as Ajumbu, Fang, Koshin, and Mungbam [11, 22, 24]; the Grassfields area to the south [9, 27]; and languages in contact with both groups [3, 19].

1.2. Fricative vowels

Mundabli /i, u/ share some features with *fricative vowels*, vowels produced with an overlay of frication attributable to a coronal or labial constriction [3, 16]. Fricative vowels are mainly described for Chinese languages, where coronal fricative vowels are also known as apical vowels [17, 29]. Most often, fricative vowels have evident supralaryngeal frication for at least the first half of their duration [6, 21, 29]. They are also known to trigger affrication or trilling of preceding onset consonants [33, 10], not unlike the extra-constricted vowels found in Mundabli, and may also preferentially occur with affricate and fricative onsets [17, 10].

Unlike the /i-ɪ/ and /u-ʊ/ contrasts described for Mundabli, fricative and apical vowels tend to differ in formant frequencies from non-fricated high vowels. Fricative vowels in Chinese languages are known to have F1-F2 values similar to high central vowels, with lower F2 for coronal fricative/apical vowels compared to [i], and higher F2 for labial fricative vowels compared to [u] [21, 12, 29, 5]. These differences are generally thought to arise as enhancements to frication production, specifically the modification of tongue-palate contact to generate strident frication in coronal fricative vowels [21, 13, 8], or lowering or 'troughing' of the tongue during labial vowels [10, 28].

1.3. Research goals

Fricative vowels are generally thought to develop diachronically from phonologization of fricative noise occurring passively on very constricted high vowels [7, 13]. The Mundabli high vowels merit examination because they appear to contrast less robustly in terms of formant frequencies, and may thus represent a diachronic precursor to fricative vowels. As such, here we evaluate (1)

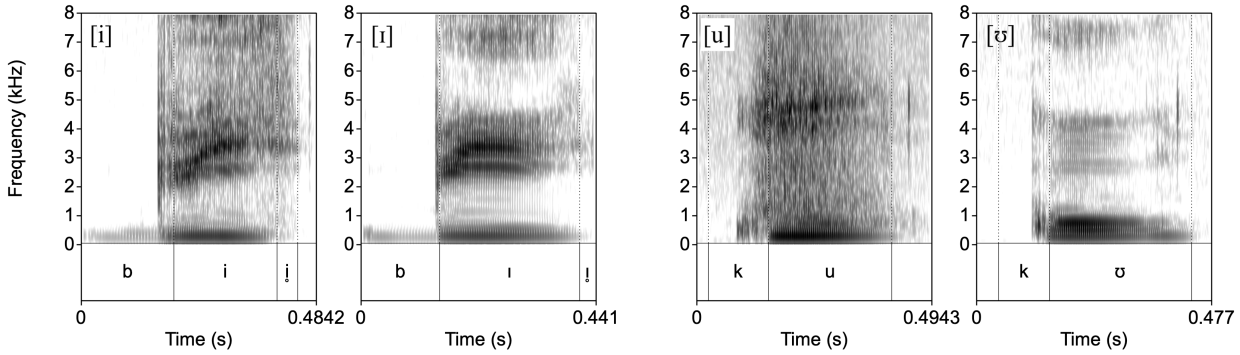


Figure 1: Sample tokens (speaker 1F) of /i ɪ u ʊ/: [bī] 'fish', [bí] 'go out-imp', [kū] 'rat-mole', [kʊ̄] 'bone'.

the contributions of frication noise and formant frequencies in the contrast between the pairs /i/-/ɪ/ and /u/-/ʊ/. We aim to identify (2) the time-dynamic pattern of frication in fricative vowels, and investigate (3) if the frication is modulated by consonants or by lexical tones. Using these data, we aim to answer the research question: are these pairs distinguishable by frication alone, or by a combination of frication and formant structure?

2. METHODOLOGY

2.1. Stimuli and data collection

The data set analyzed here was collected from four Mundabli speakers (2F, 2M) in Douala, Cameroon, in July 2022. Speakers were recorded in a quiet room using Shure SM10A head-mounted cardioid dynamic microphones and a Zoom H4n recorder (44.1 kHz sampling rate). From the resulting eight-hour corpus of elicited lexical items, we selected lexical items of the shape CV, where C is a non-nasal consonant and V is in /i ɪ u ʊ/. We excluded vowels occurring in pronouns, demonstratives, and in the first syllable of multisyllabic items (e.g. /kpʊ̱̋.kpóˤ/ 'woodpecker'; /dɪ̱̀.də̄m/ 'chest') since these appear to occur in prosodically weak positions. This yielded 1748 tokens in total (547 /u/, 498 /i/, 324 /ɪ/, 379 /ʊ/).

2.2. Data processing and analysis

Data were segmented in Praat v6.1.39 [2]. The first and second formants (F1, F2) were estimated at vowel midpoint. Praat's default settings for LPC formant estimation were used for the front vowels (ceiling 5.5 kHz, five formants estimated) for speakers 1F, 2F, and 2M; a lower 5 kHz ceiling was used for speaker 1M. For back vowels, the ceiling was lowered to preclude formant misidentification, and only two formants were estimated: a ceiling of 1.5 kHz was applied for speakers 1F, 2F, and 2M, and a ceiling of 1.4 kHz for speaker 1M. Tokens more than three standard deviations away from vowel-speaker means for F1 or F2 were removed (n=39). Measures were not normalized for speaker anatomical differences, due to the small portion of the vowel space analyzed (four of the 16 Mundabli monophthongs) and the similar height and (inferred) vocal tract length of three of the speakers (1M being roughly 0.3m taller than the others).

Formant measures were submitted to linear mixed-effects models in R v4.2.2 using *lme4* v1.1-31 [1], with *p*-values estimated using *lmerTest* v3.1-3 [15]. Separate F1 and F2 models were constructed for the front and back vowels. Models included fixed effects of vowel (/ʊ/ vs. /u/ or /ɪ/ vs. /i/), speaker, and their interaction, with random intercepts for onset and word. In front vowel models, onset was omitted as a random effect as it did not improve model fit. The less-constricted vowels /ɪ ʊ/ and speaker 1F are taken as reference levels. Post-hoc comparisons (Tukey's HSD tests) were carried out in R for F1 and F2 on estimated marginal means for each vowel pair within speaker using *emmeans* v1.8.5 [18].

To measure the timecourse of frication, we chose zero-crossing rate (ZCR) to measure the number of crossings of zero dB per second in the waveform, as in [25, 29]. To model the dynamics of ZCR, generalized additive mixed models (GAMMs) were constructed using *mgcv* v1.8-40 [32]. Separate models were constructed for comparison of the pairs /i/-/ɪ/ and /u/-/ʊ/. In the models, ZCR of the vowels was estimated over time, with factor smooths for speaker and onset. Tweedie distributions were used in the model, as ZCR follows a left skewed, long-tailed distribution. Results were visualized using

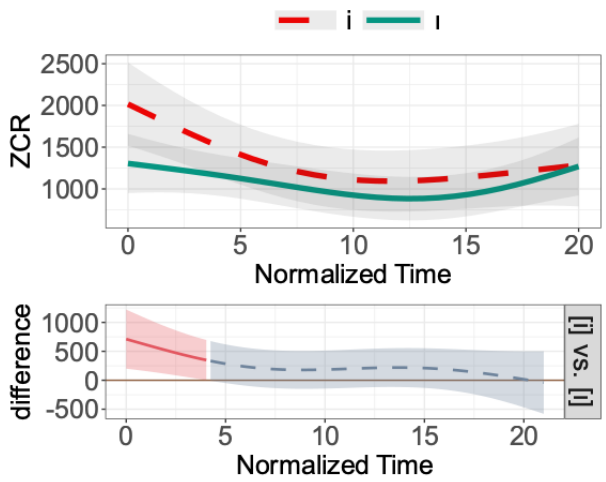


Figure 2: Top: GAMM fits for ZCR, /i/ and /ɪ/. Bottom: difference smooth; red indicates significant difference.

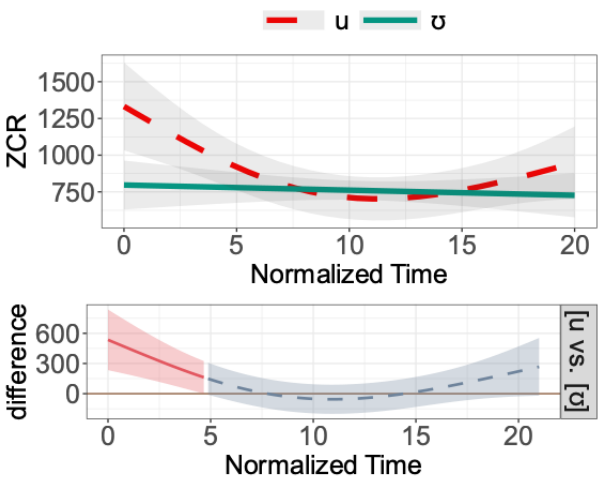


Figure 3: Top: GAMM fits for ZCR, /u/ and /ʊ/. Bottom: difference smooth; red indicates significant difference.

tidyverse v1.3.2 and *tidymv* v3.3.2 [31, 4].

3. RESULTS

3.1. Zero-crossing rate

The ZCR GAMM fits for /i, ɪ, u, ʊ/ are presented in Figs. 2–3. The /i/-/ɪ/ and /u/-/ʊ/ comparisons show clearly that more constricted /i u/ have significantly elevated ZCR at onset compared to /ɪ ʊ/. ZCR of /i u/ follows a high descending pattern, indicating that frication noise gradually attenuates over the timecourse of the vowel. After onset, neither vowel pair differs significantly in ZCR.

The level of ZCR in the vowel onset is substantially modulated by the type of onset consonant which precedes the vowel. Fricative and affricate onsets introduce visibly more frication into the following vowel, compared to plosives (Fig. 4). Because the models take by-onset variability into account, the significant differences in ZCR observed in Figs. 2-3 suggest that /i u/ contain somewhat more aperiodic energy than /ɪ ʊ/ even when occurring in "less favorable" contexts for producing frication.

3.2. Formant frequencies

Summary F1-F2 data are shown in Fig. 5. For the sake of brevity, we report only main effects of vowel and interactions of speaker and vowel, as the size of the acoustic contrast between the vowels for each speaker is mainly reflected in the interactions.

Moving first to the front vowel models, the main effect of the vowel /i/ on F1 reaches significance (β=-24.84, t=-3.84, p<0.0001), indicating that /i/ tends

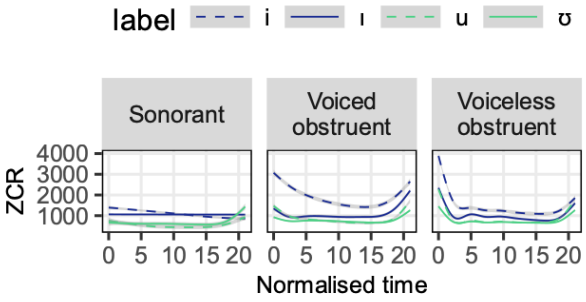


Figure 4: GAMM fits for ZCR by onset type and vowel.

to exhibit a lower F1 than /ɪ/. This tendency varies by speaker: speaker 2M's interaction with vowel reaches significance (β=27.271, t=4.253, p<0.0001), suggesting a slightly higher F1 for /i/ compared to /ɪ/. The main effect of /i/ on F2 fails to reach significance (β=-21.542, t=-0.95, p=0.35). However, speakers 2F and 2M exhibit significant interactions with vowel /i/ (2F: β=-56.95, t=-3.03, p=0.0026; 2M: β=-184.10, t=-8.78, p<0.0001), suggesting they exhibit a lowered F2 for /ɪ/ relative to /i/.

Turning to the back vowel models, the main effect of the vowel /u/ on F1 reaches significance (β=-81.64, t=-14.51, p<0.0001), indicating a tendency for /u/ to exhibit a lower F1 than /ʊ/. However, all interactions of speaker and vowel /u/ reach significance (1F: β=36.96, t=5.40; 1M: β=62.62, t=8.42; 2M: β=66.78, t=8.91; all p<0.0001): non-reference speakers exhibit a smaller F1 difference for the /u-ʊ/ pair. The main effect of the vowel /u/ on F2 is weakly significant (β=33.86, t=2.017, p=0.045),

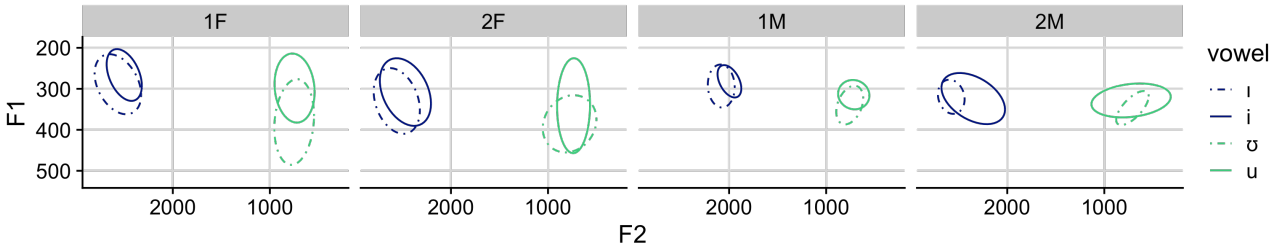


Figure 5: Confidence ellipses (95%) for F1 and F2 frequencies by vowel and speaker.

suggesting a slight tendency for /u/ to exhibit a higher F2 than /ʊ/. Again, however, interactions of speaker and the vowel /u/ reach significance for 1M (β=-100.33, t=-6.12, p<0.0001) and 2F (β=-44.67, t=-3.080, p=0.0021).

The above models suggest inter-speaker variation in implementation of the /i/-/ɪ/ and /u/-/ʊ/ contrasts in terms of formant frequencies. To investigate the degree to which speakers contrast each pair on each formant, we turn to by-speaker estimated marginal means (EMMs) for each vowel's F1 and F2. Differences for EMMs of front vowels (/ɪ/ minus /i/) and back vowels (/ʊ/ minus /u/) are shown in Table 1. F1 differs for both pairs in the predicted direction (i.e. /ɪ, ʊ/ have higher F1) for all speakers except for 2M. Unexpectedly, /i/ has lower estimated F2 compared to /ɪ/ for three speakers, with speaker 2M exhibiting a particularly large difference.

A. /ɪ/-/i/ difference

	F1 Est.	p	F2 Est.	p
1F	**24.84**	**<0.0001**	21.5	0.35
2F	**13.80**	**0.0472**	**78.5**	**0.0013**
1M	**18.58**	**0.0059**	**52.6**	**0.026**
2M	-2.43	0.7312	**205.6**	**<0.0001**

B. /ʊ/-/u/ difference

	F1 Est.	p	F2 Est.	p
1F	**81.6**	**<0.0001**	**-33.9**	**0.0477**
2F	**44.7**	**<0.0001**	10.8	0.5459
1M	**19.0**	**0.0034**	**66.5**	**0.0002**
2M	**14.9**	**0.0238**	-31.4	0.0776

Table 1: Difference in estimated marginal means for front vowels (/ɪ/ - /i/; A) and back vowels (/ʊ/ - /u/; B) by speaker. Significant differences in bold.

4. DISCUSSION

ZCR modeling reveals that both Mundabli vowel pairs /i/-/ɪ/ and /u/-/ʊ/ are distinguished by frication at vowel onset, modulated by onset consonant: more frication occurs following obstruents than sonorants. This overall difference in aperiodic energy and effect of onset consonant type have been reported for fricative vowels in Chinese [21, 8, 29], but not for Mundabli or other Lower Fungom or Grassfields languages with similar vowel contrasts.

The relationship between each vowel pair in terms of formant frequencies shows inter-speaker variation, and suggests a difference not reducible to height. While extra-high /i, u/ exhibit lower F1 than less-high /ɪ, ʊ/ as might be expected, F2 is generally lower for /i/ compared to /ɪ/, a reversal of the typical pattern. Speaker 2M lacks an F1 distinction for /i/-/ɪ/ and has an estimated F2 difference of more than 200 Hz for the same pair. This lowered F2 is typical of apical vowels in Chinese, due to fronting of the tongue-palate constriction [20]. Speaker 2M may exhibit a similar distinction for the /i/-/ɪ/ pair; whether this is idiosyncratic or typical of more Mundabli speakers is not yet clear.

Observed F1 and F2 differences are near each measure's just-noticeable difference [23, 14] and may not reliably cue the /i/-/ɪ/ or /u/-/ʊ/ contrasts. Frication may be more reliable, such that Mundabli (and languages with similar high vowel contrasts) may be on a hypothesized path for fricative vowel development in which tighter constrictions inadvertently produce frication[13, 7]. Increased airflow at obstruent release, as suggested by the Mundabli ZCR data, plausibly encourages the phonologization and enhancement of this frication.

5. ACKNOWLEDGEMENTS

Zɔ̀ɔŋ kɛ'ɛ kə̀ ne kō Ikom Christopher, Pierpaolo DiCarlo, Jeff Good, and jʊ̀ŋ kə́n mɔ̀ˢmɔ̀ˢ Ntambong Irene, Dwo Ivette, Ntemfang Ignatius, and Dwo Clifort for your contributions as language experts.

6. REFERENCES

[1] D. M. Bates, "lme4: Mixed-effects modeling with R," 2010.
[2] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2022, computer program, version 6.3.03.
[3] B. Connell, "Mambila fricative vowels and Bantu spirantisation," *Africana Linguistica*, vol. 13, pp. 7–31, 2007.
[4] S. Coretta, "tidymv: Tidy model visualisation for generalised additive models," 2022, r package, version 3.3.2.
[5] M. Faytak, "Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese," PhD dissertation, University of California, Berkeley, 2018.
[6] M. Faytak and S. Lin, "Articulatory variability and fricative noise in apical vowels," in *Proc ICPhS XVIII*, Glasgow, 2015.
[7] M. Faytak, "High vowel fricativization and chain shift," *UC Berkeley PhonLab Annual Report*, vol. 10, 2014.
[8] ——, "Place uniformity and drift in the suzhounese fricative and apical vowels," *Linguistics Vanguard*, 2022.
[9] M. Faytak and P. W. Akumbu, "Kejom (Babanki)," *Journal of the International Phonetic Association*, vol. 51, no. 2, pp. 333–354, 2021.
[10] M. Faytak, J. Kuo, and S. Wang, "Lingual articulation of the Sūzhōu Chinese labial fricative vowels," in *Proc ICPhS 19*, Melbourne, 2019, pp. 1440–1444.
[11] J. Good, J. Lovegren, J. P. Mve, C. Nganguep Tchiemouo, R. Voll, and P. Dicarlo, "The languages of the Lower Fungom region of Cameroon: Grammatical overview," *Africana Linguistica*, vol. 17, no. 1, pp. 101–164, 2011.
[12] F. Hu and Y. He, "Frication as a Vowel Feature? Evidence from the Rui'an Wu Chinese Dialect," in *Proc INTERSPEECH 2019*, 2019, pp. 3955–3959.
[13] F. Hu and F. Ling, "Fricative vowels as an intermediate stage of vowel apicalization," *Language and Linguistics*, vol. 20, no. 1, pp. 1–14, 2019.
[14] D. Kewley-Port and Y. Zheng, "Vowel formant discrimination: Towards more ordinary listening conditions," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2945–2958, 1999.
[15] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, pp. 1–26, 2017.
[16] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Blackwell Oxford, 1996.
[17] S.-I. Lee-Kim, "Revisiting Mandarin 'apical vowels': An articulatory and acoustic study," *Journal of the International Phonetic Association*, vol. 44, no. 3, pp. 261–282, 2014.
[18] R. V. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means," 2022, r package, version 1.8.3.
[19] D. Lewis and S. Shittu, "Phonemic Status of Len Fricative-Vowels," *Ibadan Journal of Humanities Studies*, vol. 24, pp. 27–45, 2014.
[20] F. Ling, "The articulatory and acoustic study of fricative vowels in Suzhou Chinese," in *Proceedings of ICPhS 16*, 2007.
[21] ——, "A phonetic study of the vowel system in Suzhou Chinese," Ph.D. dissertation, City University of Hong Kong, 2009.
[22] J. S. J. Lovegren, "Mungbam grammar," Ph.D. dissertation, State University of New York at Buffalo, 2013.
[23] P. Mermelstein, "Difference limens for formant frequencies of steady-state and consonant-bound vowels," *The Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 572–580, 1978.
[24] J. P. Mve, "Aspects of the phonology of Fáŋ," Master's thesis, Université de Yaoundé I, 2013.
[25] C. Nance and J. Stuart-Smith, "Pre-aspiration and post-aspiration in Scottish Gaelic stop consonants," *Journal of the International Phonetic Association*, vol. 43, no. 2, pp. 129–152, 2013.
[26] R. A. Ojong Diba, "The sociolinguistic dynamics of rural multilingualism in Africa: The case of Lower Fungom," Ph.D. dissertation, University of Buea, 2019.
[27] K. Olson and Y. Meynadier, "On Medumba bilabial trills and vowels," in *Proc ICPhS 18*, Glasgow, 2015.
[28] C. H. Shadle, H. Nam, A. Katsika, M. Tiede, and D. Whalen, "Aeroacoustic consequences of tongue troughs in labiodentals," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3579–3579, 2017.
[29] B. Shao and R. Ridouane, "On the nature of apical vowel in Jixi-Hui Chinese: Acoustic and articulatory data," *Journal of the International Phonetic Association*, pp. 1–26, 2023.
[30] R. M. Voll, *A grammar of Mundabli, a Southern Bantoid (Yemne-Kimbi) language of Cameroon*. Netherlands Graduate School of Linguistics, 2017.
[31] H. Wickham, "tidyverse: Easily Install and Load the 'Tidyverse'," 2022, r package, version 1.3.2.
[32] S. N. Wood, *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006.
[33] D. Yuan, F. Ling, R. Shen, and M. Shi, "Bilabial Trill Induced by Fricative High Rounded Vowel: The Emerging of "Тв" in the Wu Dialect of Su-Wan Boundary," *Journal of Chinese Linguistics*, vol. 47, no. 1, pp. 172–192, 2019.

PRE-ASPIRATED SONORANTS IN SHEHRET, A MODERN SOUTH ARABIAN LANGUAGE

Janet C.E. Watson,[1] Barry Heselwood,[1] Gisela Tomé Lourido,[1] Amer al-Kathiri[2]
[1]University of Leeds, UK, [2]University of Technology and Applied Sciences, Dhofar, Oman
j.c.e.watson@leeds.ac.uk, G.TomeLourido@leeds.ac.uk, b.c.heselwood@leeds.ac.uk, aa.alkathiri@sct.edu.om

ABSTRACT

This paper examines pre-aspirated sonorants in the Central and Eastern varieties of Shehret, an endangered Modern South Arabian language (MSAL) spoken by c. 50,000 speakers in Dhofar, southern Oman. We assume pre-aspirated sonorants fall in the class of breathy sonorants, acknowledged to be typologically rare [1], [2], and phonotactically tightly restricted [3]. Shehret pre-aspirated sonorants are restricted to the offset of stressed word-final syllables in a closed set of words [4]; they contrast both with non-pre-aspirated sonorants and with strings of /h/ followed by a sonorant (/hS/ realised as [həS] with epenthesis), giving a /ʰS/ versus /hS/ phonological sequence contrast; pre-aspiration also characterises voiceless obstruents in the language. Western Shehret apparently lacks pre-aspirated sonorants [6]. While related Soqotri exhibits a post-aspirated palatal glide [7], pre-aspirated sonorants appear not to be attested elsewhere within the Semitic language family.

Keywords: pre-aspirated/breathed sonorants; Shehret; Modern South Arabian; endangered languages; unbreathed sonorants.

1. INTRODUCTION

Shehret (aka Jibbāli) is one of six endangered Modern South Arabian languages (henceforth MSAL) spoken along the southern edge of the Arabian Peninsula in an area covering Oman, eastern Yemen, the island of Soqotra and the southern fringe of Saudi Arabia, an area where Arabic is the dominant and majority language. With approximately 50,000 speakers, Shehret itself is spoken on the mountain ranges parallel to the coast of Dhofar, southern Oman, and along the Dhofari coastline.

Segments in Shehret, as in other MSAL, fall into two laryngeal classes: 'breathed' (or 'open glottis') for voiceless obstruents, and 'unbreathed' (or 'constricted glottis') for vowels, for all canonically voiced consonants and for consonants typically described in the literature on Semitic languages as 'emphatic' [8]. Henceforth pre-aspirated sonorants are described as 'breathed' and non-pre-aspirated sonorants as 'unbreathed'.

It is known that sonorants in Shehret are usually realised without voice in utterance-final position [9]. Dufour [10] transcribes breathed sonorants with a devoicing symbol, as *n̥*; however, both breathed and unbreathed sonorants are realised without voice and are typically inaudible in utterance-final position [4]. The significance of breathed sonorants in Shehret is fivefold: a) the long duration of breathy voice and aperiodic noise ~100ms; b) native speakers' awareness of breathed sonorants, transcribing them with preceding 'h' in Arabic-based orthography; c) the typical lack of any acoustic trace of utterance-final sonorants, although EPG data show articulation in place; d) the loss of breathiness utterance-medially before an unbreathed segment; e) (near-)minimal contrasts with both unbreathed sonorants, e.g. *be-ḏóʰr* 'with blood' v. *bóḏór* 'to sow' and /hS/ strings, e.g. *ḏáʰn* 'this' v. *ḏáh[ə]n* 'mind'. We assume epenthesis applies here to prevent a surface *ʰS* v. *hS* contrast, predicted not to occur cross-linguistically [5].

2. METHODS

This paper draws on acoustic (Ac), electrolaryngographic (ELG) and electropalatographic (EPG) data. ELG and acoustic recordings were produced in the field in Dhofar; EPG recordings were produced at the University of Leeds (UoL).

2.1. Speaker information

For this study, 7 speakers (1 woman, 6 men) provided ELG data, 3 of whom also provided EPG data; a second female provided ELG data (J116a), these were excluded due to poorly fitting electrodes. Her acoustic data were extracted from the ELG files by taking channel 1 on PRAAT [11] and analysed with other acoustic data from ELG and EPG files. 2 speakers are from Central Dhofar, 2 from East Dhofar and 3 from Central-West Dhofar. Eastern speakers are predicted to exhibit pre-aspirated sonorants in more words than Central/Central-Western speakers.

	Sex	Age	Region	Ac	ELG	EPG
J001	M	39	C	√	√	√
J002	M	34	C	√	√	
J028	M	34	E	√	√	
J043	M	40	E	√	√	√
J116	M	23	C-W	√	√	
J116a	F	36	E	√		
J117	F	46	C-W	√	√	
M026	M	48	C	√	√	√

Table 1: Speaker information

2.2. Acoustic analysis

Acoustic data were extracted by taking channel 1 of the ELG file on PRAAT and from EPG data through taking the corresponding .wav file. A wordlist of 83 different words elicited 594 sonorants: 350 (59%) unbreathed and 244 (41%) breathed: /l, ʰl/ = 127, /m, ʰm/ = 158, /n, ʰn/ = 236, /r, ʰr/ = 73. Extracted acoustic data were segmented in PRAAT TextGrids to note the presence and duration of breathy voice (BV) and aperiodic noise (N), as in Figure 1:

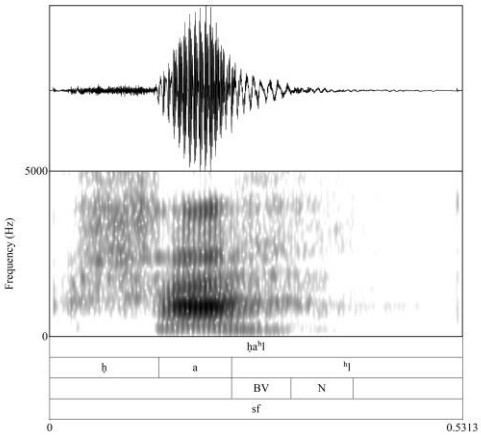


Figure 1: J028 *ḥaʰl* 'pressed oil; time'

The >100ms duration of pre-aspiration in breathed sonorants (Figure 1 BV + N = 156ms) contrasts with impressionistically typical 20~65ms pre-aspiration in breathed plosives (Figure 2 BV + N = 62ms):

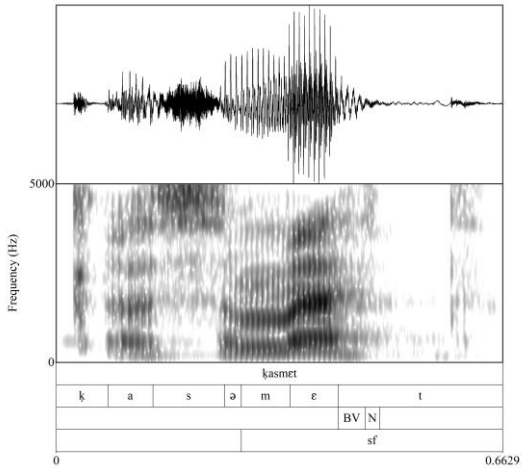


Figure 2: J028 *ḳasmét* 'present'

2.3. Electrolaryngographic analysis

ELG data were recorded on a Laryngograph EGG-D200 microprocessor with an ECM 500L/SK lapel microphone. 226 utterance-final breathed sonorants were analysed alongside 226 utterance-final unbreathed sonorants. A selection of utterance-medial breathed sonorants was examined for maintenance or lack of breath in this position.

Data were analysed through Speech Studio for Connected Speech Analysis by measuring the closed quotient (CQ) value at the mid-point and at the offset of the vowel preceding utterance-final sonorants. For utterance-medial breathed sonorants, the CQ value was measured at the mid-point and offset of the preceding vowel. The CQ of a glottal voicing cycle is the proportion of the cycle during which the vocal folds are in contact, expressed as a percentage of the cycle's duration [12]. The lower the CQ value of a cycle, the more air is free to flow through the glottis, giving an inverse relation between CQ and transglottal airflow, with values <40% sounding increasingly breathy [13].

Table 2 gives the number of ELG tokens of breathed and unbreathed utterance-final sonorants by speaker:

	Unbreathed	Breathed
J001	24	30
J002	12	0
J116	27	48
J117	24	19
J028	27	48
J043	95	63
M026	17	18
TOTAL	226	226

Table 2: Number of utterance-final sonorant tokens by speaker for ELG analysis

The CQ of vowel offset preceding breathed /ʰl/ in *[eg]miʰl* 'the camels' with a value of 20% is indicated on the laryngogram in Figure 3 by the vertical line. Above the trace are, from top down, the spectrogram, speech waveform, and larynx waveform.

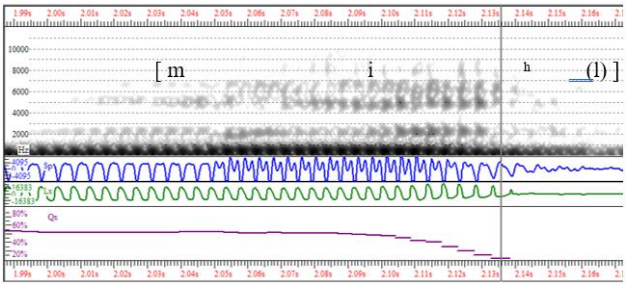


Figure 3: J001 *[eg]miʰl* 'the camels' with aligned phonetic transcription: (l) = silently articulated /l/

This compares with the CQ of vowel offset preceding unbreathed /l/ in *źīyɛl* 'camel owners' in Figure 4 with a value of 60.67% (CQ of vowel midpoint 53.56% (arrowed). ({V̩}) = creaky phonation, (l) = silent /l/.)

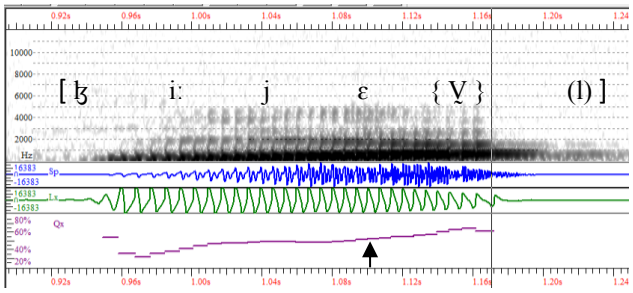


Figure 4: J043 *źīyɛl* 'camel herders' with aligned phonetic transcription; position of silent /l/ approximate

In utterance-medial position, the breathiness of breathed sonorants is lost before unbreathed consonants and vowels, shown by large CQ differences between utterance-medial and utterance-final positions. For J028, CQ values of between 52–54% at vowel offset before [l] in utterance-medial *ḥa^{h}l* in *ḥa^{h}l mɛken* 'lots of pressed oil' contrast with values of between 25–27% at vowel offset before /^{h}l/ in *ḥa^{h}l* in utterance-final position; CQ values of between 48–55% at vowel offset before [n] in utterance-medial *sɛ^{h}n* 'they f.' in *sɛ^{h}n bū^{h}n* 'they f. are here' contrast with values of between 20–29% at vowel offset before /^{h}n/ in *sɛ^{h}n* in utterance-final position; CQ values of between 48–53% at vowel offset before [r] in utterance-medial *śḥɛ^{h}r* in *śḥɛ^{h}r əd lɔ̄d rḥim* 'the mountains became good' contrast with values of between 25–33% at vowel offset before /^{h}r/ in *śḥɛ^{h}r* in utterance-final position. For J043, CQ values of between 51–55% at vowel offset before [m] in utterance-medial *šu^{h}m* in *šu^{h}m bū^{h}n* 'they m. are here' contrast with values between 23–28% at vowel offset before /^{h}m/ in *šu^{h}m* in utterance-final position.

Where a breathed sonorant precedes a breathed obstruent, however, the CQ at vowel offset drops below 40% for some speakers, continuing to drop through the sonorant towards the breathed obstruent; the sonorant is sounded throughout, but typically becomes increasingly breathy; in *ri^{h}m fɛjˈat* 'extremely tall' in Figure 5 the CQ at vowel offset (B) is 38% and drops to 27% at the boundary between breathed /^{h}m/ and breathed /f/ (vertical line).

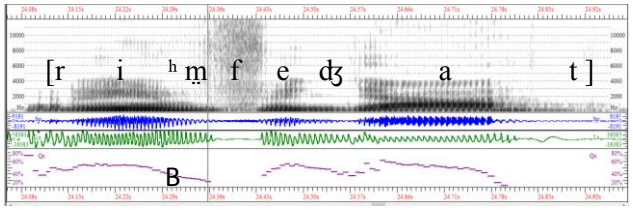


Figure 5: J028 *ri^{h}m fɛjˈat* 'extremely tall' with aligned phonetic transcription

For other speakers, the breathed sonorant begins non-breathy before a breathed obstruent, but becomes increasingly breathy, as in J043's production of *šu^{h}m tōlɛn* 'they m. are with us' in Figure 6. The CQ at vowel offset before utterance-medial /^{h}m/ is around 48%. The start of breathy voice with a CQ value of 35% is marked by the vertical line.

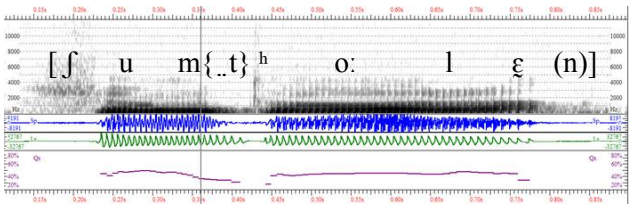


Figure 6: J043 *šu^{h}m tōlɛn* 'they m. are with us' with aligned phonetic transcription; curly braces delimit the breathy voice portion of /^{h}m/ which leaks into the closure for /t/

2.4. Electropalatographic analysis

EPG data were collected to identify tongue–palate contact patterns and any acoustic/articulatory gaps between vowel offset and onset of articulation. EPG data were collected in the UoL Phonetics Laboratory through Articulate Assistant version 1.18; EPG palates for all speakers were Articulate style produced by Bristol Dental Hospital. EPG data was sampled at 100Hz. The number of EPG tokens of each sonorant by speaker is given in Table 3 below:

Unbreathed	**Total**	J043	M026	J001
/l/	**41**	20	12	9
/n/	**51**	25	18	8
/r/	**42**	27	12	5
TOTAL	**134**	**72**	**42**	**20**
Breathed				
/^{h}l/	**15**	12	3	0
/^{h}n/	**51**	38	14	0
/^{h}r/	**14**	12	6	0
TOTAL	**85**	**62**	**23**	**0**

Table 2: Number of utterance-final sonorant tokens by speaker for EPG analysis

The following EPG measures were taken:

1. Time in milliseconds (ms) to nearest 10ms (as 100Hz sampling rate generates 1 frame every 10ms) of any delay between final voicing pulse of preceding vowel and ONSET of the sonorant articulation taken as first palate frame showing more contact than for the preceding vowel; this duration is referred to as 'ONSET DELAY'.
2. Time to nearest 10ms between ONSET and first palate frame of maximum articulatory contact ('MAX'), referred to as 'ONSET-MAX'.
3. Time to nearest 10ms between MAX and first frame showing breaking of the articulatory closure ('OFFSET'), referred to as 'MAX-OFFSET'.

Landmarks for measuring ONSET DELAY, ONSET-MAX, and MAX-OFFSET are shown in Figure 7.

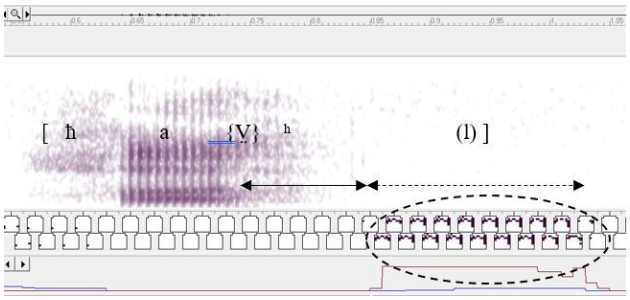


Figure 7: J043 *ḥaʰl* 'pressed oil' with aligned phonetic transcription; silent articulation and tongue-palate contact profile circled; ONSET DELAY (110ms) = double-headed arrow; ONSET-OFFSET (180ms) = dashed double-headed arrow

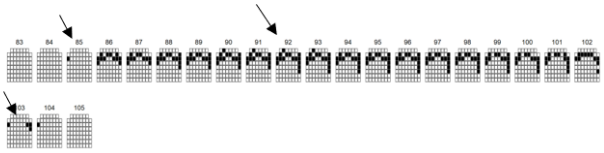


Figure 8: Enlarged image of palates circled in Figure 7; measurement landmarks ONSET (frame 85), MAX (frame 92), OFFSET (frame 103) arrowed

3. RESULTS

3.1. Acoustic

Utterance-final breathed sonorants show a pre-aspiration noise duration of ~100ms, with /ʰr/ and /ʰl/ exhibiting longer durations than /ʰm/ and /ʰn/, as in Table 3. This compares with the typical 20–50ms pre-aspiration duration of Shehret breathed obstruents.

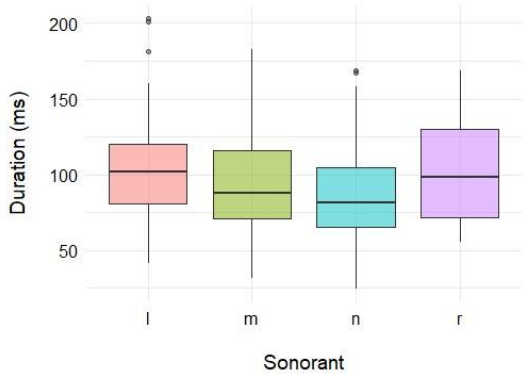


Table 3: Duration of preaspiration

3.2. Electrolaryngographic results

The distribution of CQ values for vowel offset before breathed sonorants gives 5 tokens (2.21%) with CQ of >40%, and 179 (79.2%) with CQ values <30%, well within the range for breathy voice phonation. Comparing the CQ of vowel midpoint with that of vowel offset, the CQ slope of a vowel drops by an average of 22.61% before a breathed sonorant.

Only 27 tokens (11.95%) of unbreathed sonorants exhibited CQ values <40%, with 176 (77.88%) exhibiting CQs of >45%. Comparing the CQ of vowel mid with vowel offset, the CQ slope remains almost flat in these, rising from an average of 49.33% to 50.52%.

3.3. Electropalatographic results

Mean ONSET DELAY, ONSET-MAX, MAX-OFFSET and ONSET-OFFSET duration values with ranges and standard deviations are provided in Table 4 for breathed, and Table 5 for unbreathed sonorants. All breathed tokens show an onset delay which typically reached 50ms or more: 66.7% for /ʰl/, 96.1% for /ʰn/, 71.4% for /ʰr/. Delays of >100ms are very common: 40% for /ʰl/, 72.5% for /ʰn/ and 57.1% for /ʰr/. Of the 134 unbreathed tokens, 3 tokens of /l/ and 2 of /n/ show no onset delay. The remainder show delays of >50ms in the majority of cases: 63.2% /l/, 78.4% /n/, 84.1% /r/; delays of >100ms are not uncommon: 26.3% /l/, 39.2% /n/, 22.7% /r/.

	ONSET DELAY			ONSET-MAX			MAX-OFFSET		
	M	SD	R	M	SD	R	M	SD	R
/ʰl/	81.6	56.5	12-210	21.33	17.27	0-60	51.47	27.8	150-240
/ʰn/	141.31	76.47	21-470	51.37	55.32	0-320	110.0	66.16	40-280
/ʰr/	92.07	45.1	37-140	12.86	7.26	0-30	38.57	21.07	10-70

Table 4: Duration values for ONSET DELAY, ONSET-MAX and MAX-OFFSET for /ʰl, ʰn, ʰr/; M=mean, SD=standard deviation, R=range

	ONSET DELAY			ONSET-MAX			MAX-OFFSET		
	M	SD	R	M	SD	R	M	SD	R
/l/	74.13	46.95	0-170	35.53	27.28	0-90	154.21	70.47	30-320
/n/	90.14	57.68	0-261	45.49	36.84	0-180	158.04	68.03	40-330
/r/	75.7	29.23	17-142	20.91	23.21	0-110	37.27	25.09	10-100

Table 5: Duration values for ONSET DELAY, ONSET-MAX and MAX-OFFSET for /l, n, r/

As well as exhibiting an onset delay, all tokens of breathed and unbreathed utterance-final sonorants show the articulation in place but no auditory trace of the sonorant.

4. SUMMARY

Impressionistically, the duration of pre-aspiration in utterance-final breathed sonorants is considerably longer than pre-aspiration in breathed obstruents. Articulatory onset delays are greater before breathed than before unbreathed utterance-final sonorants. The restriction of breathed sonorants to utterance-final or pre-breathed position supports earlier observations for the cross-linguistic phonotactic restriction of breathy sonorants [3].

EPG data show the sonorant articulation in both breathed and unbreathed sonorants in place utterance-finally, while acoustically completely inaudible [4].

5. ACKNOWLEDGEMENTS

This work has been supported through a Leverhulme Trust Major Research Fellowship to Watson (MRF-2018-121). We are grateful to Míša Hejná for encouraging our research on pre-aspirated sonorants and commenting on a draft abstract, to Chris Norton for producing a script to create the PRAAT drawings, and to all our Shehret speakers for freely providing the data analysed.

6. REFERENCES

[1] Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press.

[2] Berkson, K. H. 2016. Production, perception, and distribution of breathy sonorants in Marathi. In: Menon, M., Syed, S. (eds.), *Proceedings of Formal Approaches to South Asian Languages*. Konstanz, 4–14.

[3] Berkson, K. H. 2019. Acoustic correlates of breathy sonorants in Marathi. *Journal of Phonetics* 73, 70–90.

[4] Watson, J. C. E., Heselwood, B. Tomé Lourido, G., al-Kathiri, A., al-Mahri, A. In press. Silent sonorant articulations in Mehri and Shehret. In: Anderson, C., Kuznetsova, N., Easterday, S. (eds), *Rarities in Phonetics and Phonology: Evolutionary, Structural, Typological and Social Dimensions*. Language Science Press.

[5] Kehrein, W., Golston, C. 2004. A prosodic theory of laryngeal contrasts. *Phonology* 21, 1–33.

[6] Al-Maʿšanī, A. 2014. *Muʿjam lisān Ẓufār*. Muscat.

[7] Naumkin, V., Kogan, L., al-Daʿrhi, I. G., al-Daʿrhi, A. I., Cherkashin, D., Bulakh, M., Vizirova, E. 2015. *Corpus of Soqotri Oral Literature: Volume 1*. Brill.

[8] Watson, J. C. E., Heselwood, B. C. 2016. Phonation and glottal states in Modern South Arabian and San'ani Arabic. In: Haddad, Y. A., Potsdam, E. (eds), *Perspectives on Arabic Linguistics XXVIII*. John Benjamins, 3–36.

[9] Johnstone, T. M. 1981. *Jibbāli lexicon*. Oxford University Press.

[10] Dufour, J. 2016. *Recherches sur le verbe subarabique modern*. Habilitation. École pratique des hautes études.

[11] Boersma, P., Weenink, D. 2017. *Praat: Doing phonetics by computer* (uva.nl).

[12] Abberton, E. R. M., Howard, D. M., Fourcin, A. 1989. Laryngographic assessment of normal voice: A tutorial. *Clinical Linguistics & Phonetics* 3, 281–296.

[13] Heselwood, B., R. Maghrabi 2015. An instrumental-phonetic justification for Sībawayh's classification of *ṭā, qāf* and *hamza* as *majhūr* consonants. *J. Semitic Studies* 60, 131–175.

Evaluating cluster analysis on f0 contours: an information theoretic approach on three languages

Constantijn Kaland, T. Mark Ellison

Institute of Linguistics, University of Cologne, Germany

ckaland@uni-koeln.de, tellison@uni-koeln.de

Abstract

This paper proposes a method for evaluating cluster analyses of f0 contours. Contour clustering was recently introduced as a theory-neutral approach to grouping f0 contours. This approach is applicable to any type of speech data (spontaneous to scripted) from any language. It lets the user explore how f0 shape differences relate to differences in linguistic content. Many cluster analysis methods leave the number of clusters as a parameter. The current paper proposes a method for identifying the ideal value for this parameter for a given dataset. In particular, we use a the Bayes-equivalent method of finding the minimum description length (MDL). This method is illustrated here using f0 data from three typologically different languages, obtained using different elicitation methods. The results show that the MDL measure selects cluster counts corresponding natural classes identified by humans.

Index Terms: cluster analysis, f0 contour, evaluation metric, information theory.

1. Introduction

A well known question in prosody research is how form and function are related in languages (e.g. [1]). For example, how does the shape of an f0 contour relate to its communicative function in a certain language? Although research has provided different approaches to this issue, much work still focusses on a small number of well-known languages [2]. In addition, autosegmental metrical approaches are frequently applied to identify an inventory of pitch accents and boundary tones [3], [4]. Often, the resulting description of intonation in a given language is limited to a small number of fixed functions, commonly based on scripted speech from a small number of speakers. These descriptions rarely represent natural, spontaneously produced speech, which is in part the result of methodological limitations. In order to lower the workload threshold for including less-known languages in prosody research, and in order to incorporate more diverse speech data, new methods need to be developed. Most importantly, these new methods need to work with all languages and improve the analysis of their prosody and intonation. This paper proposes a novel way of evaluating cluster analyses of f0 contours.

Cluster analysis (CA) has recently been proposed as a means of identifying prototypical f0 contours [5]. This method starts with N f0 contours (individual observations) each in a singleton cluster. Larger clusters are formed by merging smaller clusters (agglomerative hierarchical clustering). Each contour is numerically represented by time-series measurements (vectors) of the fundamental frequency (either Hz, ST or a speaker-normalized conversion). Merging clusters into larger ones halts only when all observations form a single cluster. Clusters are merged deterministically based on 1) the distance between all possible pairs of observations as expressed in a distance matrix, here calculated using Euclidean distances [6] and 2) the linkage criterion used when clusters with multiple observations need to be merged, here *complete linkage*. Complete linkage "iteratively merges two clusters from the current CA that have the smallest diameter when merged into a single cluster" ([7], p.1131). This linkage criterion maintains maximal dissimilarity between the clusters, in order to obtain the largest differences between contours from unmerged clusters. The output of the hierarchical clustering is a dendrogram; a tree structure showing the entire merging process from separate clusters for each individual observation (bottom) to all observations in a single cluster (top). The height at which one analyses the dendrogram corresponds to the number of clusters in the analysis.

Hierarchical clustering, unlike K-means clustering, performs the analysis and allows for the resulting dendrogram to be inspected *before* deciding on the number of clusters. However, determining the number of clusters often remains a key issue in CA. There are numerous statistical methods available to determine the number of clusters based on several aspects of the CA [8]. However, many of these evaluation methods are generic, i.e. not tailored to the type of data under analysis. A coarse division has been made between evaluation of a 'goodness of fit' given a within and between cluster variation (global measures) and the evaluation of whether or not a cluster should have been merged with another one (i.e. not clustered) or subdivided into more clusters (local measures) [9]. A common way to find the number of clusters is by finding the point at which the variance explained by the CA no longer increases such that it is worth assuming more clusters (so called 'elbow' or 'knee' method). Such a stopping-criterion is entirely based on the fact that with more clusters, more variance is explained. In the approach outlined in this paper, we account for multiple factors that play a role in evaluation the CA, specifically when applied to f0 contours. The choice of the number of clusters is particularly crucial when the analysis seeks to reveal an inventory of phonologically distinct contours ([3] [4]) from data in which there is (potentially) a high degree of 'allophonic' contour variation. Thus, the two main competing factors to take into account here are explaining contour variation (seeks a higher number of clusters) whilst finding a compact set of contour prototypes (seeks a low number of clusters). Allowing many clusters may reveal smaller and smaller (potentially uninformative) yet real differences between the clusters (overfitting). At the same time, assuming that there are contour prototypes (as done in most theories of intonational phonology, e.g. [1]) implies to look for contours that can be subsumed under the same type. The approach outlined in this paper assumes that the ideal number of contour clusters defines an optimum between those competing factors. It is important to note that choosing the right number of clusters remains a challenge, particularly when exploring the prosody of an understudied language. There might be lit-

tle to no background knowledge about which f0 movements are meaningful, a dendrogram might not provide sufficient directions, and it is important not to rely on distinctions perceived by the researcher which will reflect their language background. Thus, it is crucial to evaluate the CA based on the f0 variation in the dataset separate from interpreting the outcome based on other knowledge about the language (e.g. previous literature).

This paper proposes to evaluate the number of clusters using the Minimum Description Length (MDL) principle of [10]. MDL is based Shannon's (e.g. [11]) theory of information. Shannon related the level of information in a random variable X, with the probabilities of its $(x_i)_i$ possible outcomes (see 1).

$$\mathrm{H}(X) = -\sum_{i=1}^{\mathrm{n}} \mathrm{p}(x_i)\log(\mathrm{p}(x_i)) \quad (1)$$

The Shannon entropy (or information cost) for rolling a dice (6 possible outcomes) is -(6*1/6)*log(1/6) = 1.79, where entropy is expressed in nats (natural bits, each around 1.44 binary bits). In contrast, the entropy of flipping a coin is much smaller: $-2 * \frac{1}{2} * \log(\frac{1}{2})$ = 0.69. Events with low probability contain more information. The entropy measure can be used to find the least information needed to describe a certain dataset. By Ockham's Razor (e.g. [12]), this will be the best account. Hence we propose the model-selection criterion of Minimum Description Length (MDL, [10]).

In order to find the number of clusters that best fits a speech dataset according to MDL, all information needed to describe the data given the clustering model must be evaluated. This includes 1) the cost of specifying a mean contour for each cluster. This cost is expected to increase with more clusters. The measure also must take into account 2) the cost of classifying each contour in the data into the corresponding cluster. This quantity is likely to increase with more clusters. Finally, 3) the cost of specifying each contour, given that we know the mean contour for that cluster. This quantity will decrease with more clusters, if the CAs are sensible, as in smaller clusters the variance from the mean will be less. Given that (1) and (3) most clearly differ in direction with higher number of clusters, the sum of the three information cost measures is expected to approximate a U-shape, with the lowest point indicating the least information cost, corresponding to the MDL of the dataset. In short, the total information cost (IC) per round of CA is the sum of:

1. IC of expressing the mean contour per cluster relative to the contour average across the whole dataset,
2. IC of expressing for each contour what cluster it is in,
3. IC of expressing each contour relative to the mean contour of the cluster it belongs to.

In the remainder of the paper we illustrate the application of the above evaluation metric using data from three typologically different languages. We have chosen datasets with elicited data in order to obtain a high level of control over the number of contours produced. Thus the accuracy of the evaluation metric can be assessed by comparing the MDL with the knowledge of the context in which the contours were elicited. The evaluation method is coded in R [13], built-in and shared along with the tool available for contour clustering ([5], ver. 2022-04).

2. Datasets

The following subsections describe the three datasets to which we have applied contour clustering and the proposed evaluation metric. The datasets come from typologically different languages (German, Papuan Malay and Zhagawa) and from different constituents (intonation phrase, noun phrase, syllable). The investigation in this paper compares the number of contours as hypothesized in the respective studies with the number of contours suggested by the evaluation metric *within* each language. The following subsections list relevant aspects of the experiments used to collect the speech data. Additional methodological details (e.g., speakers, number of items) are listed in the respective papers and not repeated here for space reasons.

2.1. German

The German data came from the standard variety spoken in Germany (ISO: deu). The elicitation took the form of a discourse completion task in which participants responded to a pre-recorded speaker by reading out loud written stimuli. The stimuli were presented so that the final sentence (target utterance) occurred in different speech acts (exclamative [E] or question [Q]) and in different focus conditions (given [G] or contrast [C]). For example, the pre-recorded speaker would say: "Anna hat sich in ihrer Dissertation jetzt auf Germanen spezialisiert" (*Anna is doing a PhD in history and has specialized in Germanic peoples*). The participant would then respond with:

[E-G] "Ja, das hat sie mir neulich erzählt. Sie ist irre viel unterwegs, um an Originalquellen von Germanen heranzukommen. Wo die schon überall Germanen erforscht hat!"
Yes, I heard that, too. Anna has been traveling a lot for this. The places where she's done research on the Germanic peoples!

[Q-G] "Wirklich? Da ist sie bestimmt viel unterwegs, um an Originalquellen von Germanen heranzukommen. Weißt du zufällig, wo die schon überall Germanen erforscht hat?"
Really? I'm sure Anna has been traveling a lot for this. Do you know more? Where has she done research on the Germanic peoples?

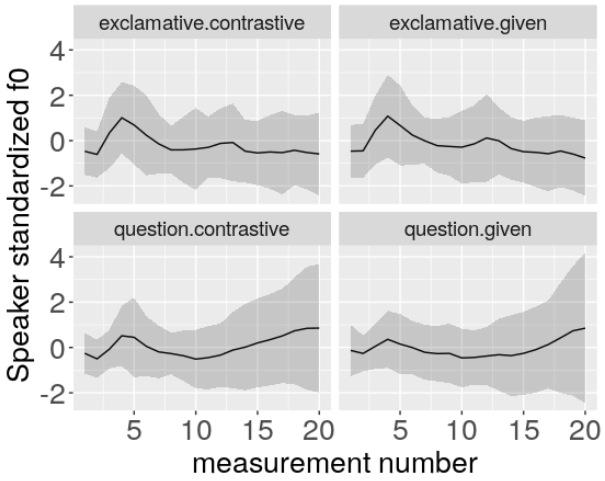


Figure 1: *DEU: mean f0 contours in each condition.*

In the contrastive context, the pre-recorded sentence has *Paul* as the subject of the first sentence and *Anna* is introduced as a contrastive referent in the pre-final sentence read by the participant. See detailed descriptions of the stimulus materials and procedure in [14] and [15]. The f0 contours of the recorded target utterances were measured using 20 points per contour, which were equally distributed over the contour. The mean f0 in each experimental condition shows that exclamatives end low,

whereas questions end high (Figure 1). Furthermore, there appears to be more subtle differences between the focus conditions in that the height and the width of the initial f0 peaks are larger for contrastive contexts than for given contexts. Thus, at least four different clusters are expected when all the (subtle) differences can be successfully clustered.

2.2. Papuan Malay

Papuan Malay (ISO: pmy) is spoken in the Eastern-Indonesian provinces Papua and West-Papua [16]. The data comes from elicitations of contrastively focused noun phrases (noun-adjective) that refer to pictures differing in shape and color (details in [17]). The descriptions all have the same structure (matrix sentences), referring to a picture displayed on the left side of a screen (antecedent A) and a picture displayed on the right side of the screen (target T). The noun phrases occur in either medial or final phrase position, for example:

"Saya liat [A] di sebla kiri, dang saya liat [T] di sebla kanang."
I see [A] on the left side, but I see [T] on the right side.

"Di sebla kiri saya liat [A], dang di sebla kanang saya liat [T]."
On the left side I see [A], but on the right side I see [T].

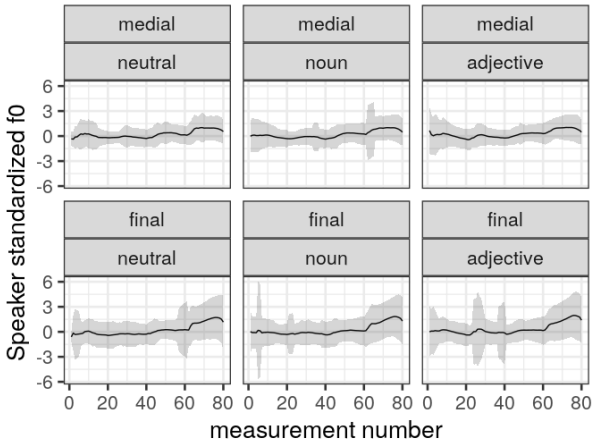


Figure 2: *PMY: mean antecedent f0 contours in each condition.*

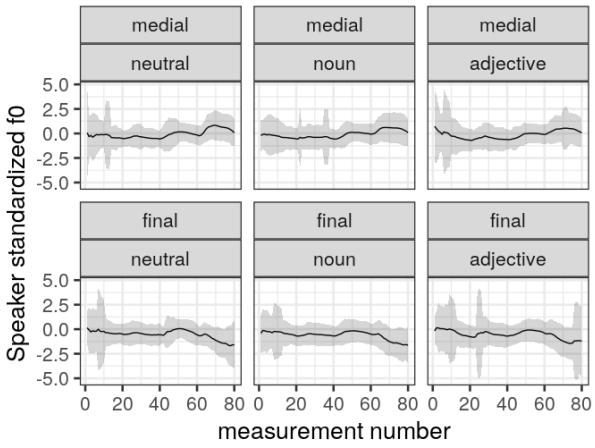


Figure 3: *PMY: mean target f0 contours in each condition.*

The noun phrases referring to the pictures expressed a shape contrast; eliciting focused nouns (e.g., A: babi hitam, T: pisang hitam, *black banana - black pig*), a color contrast; eliciting focused adjectives (e.g., babi hitam - babi mera, *black pig - red pig*), or no contrast at all (e.g. A: babi hitam, pisang mera, *black pig, red banana*). All nouns and adjectives were bisyllabic. The f0 contour of each of the four syllables in the target noun phrase was measured using 20 points per syllable. The mean contours of the entire noun phrase (80 points) in each of the experimental conditions (phrase position: medial, final; focus: neutral, noun, adjective) are shown in Figure 2 (A) and Figure 3 (T). The contours do not appear to differ between the focus conditions. For both A and T there appears to be a difference between the phrase positions: in medial position, the final syllable shows a mid level contour (invariable between A and T), whereas in final position antecedents end high and targets end low. Thus, three clusters is hypothesized to be a sufficient number to describe the differences between the Papuan Malay contours in this dataset.

2.3. Zhagawa

Zhagawa (ISO: zag) is a tone language spoken in Darfur, Africa. The data was obtained from one speaker (no f0 speaker normalisation) and consists of elicitations of body parts, kinship terms, animals and colours for the purpose of investigating number marking by tone ([18], see [5] for details). F0 contours (20 measurements) were taken from the final syllable of 212 elicited words. The final syllable is the reported location for tonal number marking ([19],[20],[21]). These analyses differ in the number of tones are used to mark number. A preliminary CA indicated the highest success rate of distinguishing singulars from plurals with six clusters [5], corroborating the analysis in [19], in which three register tones precede a modulated tone (LH, LM, MH, ML, HM, HL). The state of the research for Zhagawa represents an ideal test case for the evaluation metric; there are hypotheses about the use of tone for number marking, but the exact inventory of tones still remains uncertain.

3. Cluster analyses evaluation

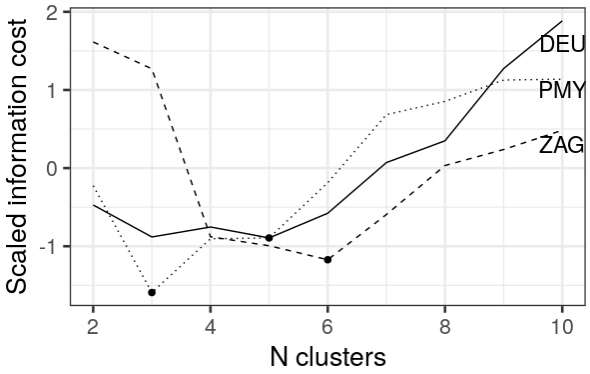


Figure 4: *Scaled information cost per number of clusters for each of the languages. Dots indicate MDL.*

The f0 measures from the DEU, PMY and ZAG datasets were run through several rounds of cluster analyses (2-10 clusters). The output CAs were subsequently measured for their information costs and then scaled to compare them in a line graph (Figure 4). The MDLs indicated 5, 3 and 6 as optimal cluster numbers for the respective data sets. In what follows, the mean contours per cluster are plotted and discussed per language.

For German, analysis with 5 clusters (Figure 5) indicates that the majority of contours in cluster 2 (60/68) and 5 (21/22) are produced when the elicited speech act concerned exclamatives. In these clusters a final rise can be observed. The other clusters also show a majority of exclamatives, although with relatively lower percentages; cluster 1: 67/78, cluster 3: 53/94, cluster 4: 25/35. No (contrast/given) cluster shows a

clear majority of subject focus for either category. Cluster 2 (37/68) and cluster 4 (16/35) show the highest percentage of contours elicited in the contrastive focus condition. Accent ratings in a perception task revealed that some participants placed a pitch-accent on the object *Germanen* when the subject was contrastively focused, which led to an increased f0 towards the end of the phrase. This is exactly what can be observed for the contours in cluster 2 and 4. In this way, the object-accented contours have their unaccented object counterparts in cluster 1 (exclamatives) and 2 (questions) respectively. The contours in cluster 3 show smaller excursions overall, and most often occur in exclamative, given contexts.

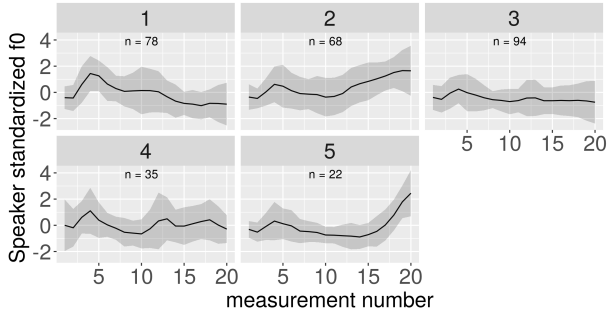


Figure 5: *DEU: mean f0 contour in each cluster (N = 5).*

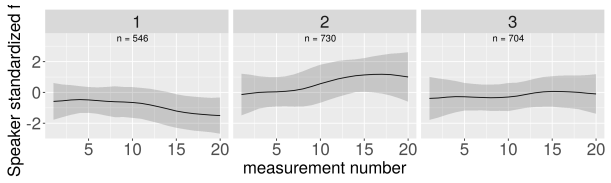


Figure 6: *PMY: mean f0 contour in each cluster (N = 3).*

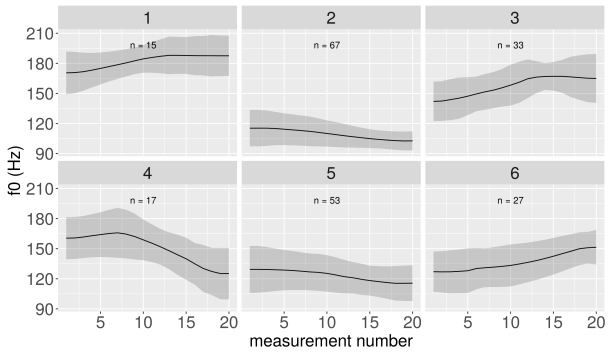


Figure 7: *ZAG: mean f0 contour in each cluster (N = 6).*

The Papuan Malay analysis with 3 clusters reveals overall shallower contours than German. This may be a result of the contour length: an entire phrase is used in German and a noun phrase in Papuan Malay. The differences between the contours can best be described by looking at the final syllable in the noun phrase (measurement points 60-80). This part shows a final fall (cluster 1), final rise (cluster 2) or a sustained mid-level f0 (cluster 3). Note that these three prototypes match the variation observed in Figure 2 and 3, with falls mainly produced in target final position, rises in antecedent final position and sustained f0 in medial positions of both antecedent and target.

The analysis with 6 clusters for Zhagawa provides a close match with one of the analyses in the literature [19]. That is, cluster 1: MH, cluster 2: ML, cluster 3: LH, cluster 4: HL, cluster 5: HM, cluster 6: LM. It should be observed that the absolute levels do not match the registers perfectly (c.f. cluster 2 and 5). In addition, the CA performed in [5] indicated the most accurate distinction between singulars and plurals in this dataset using 6 clusters (83.33%). This accuracy level did not improve when more clusters were permitted.

4. Discussion

The evaluation metric applied to these three data sets can be used widely to identify the number of clusters, and corresponding prototypes, in a set of f0 contours. Note that in addition to the typological differences, the datasets concerned different (prosodic) constituents (Sec. 2) and different prosodic phenomena (DEU: speech act/focus, PMY: phrase position/focus, ZAG: number marking). Still, MDL selected a number of categories for each dataset that could be interpreted as a plausible outcome for each of the investigated languages. The evaluation metric was able to discover relevant contour differences that were not manipulated in the experimental setup (DEU) as well as to reduce the contour differences to the most essential ones, even lower than the number of experimental conditions (PMY).

It should be noted that the number of clusters chosen by MDL remains sensitive to a number of factors related to the dataset. Most importantly, fine-grained differences that are potentially linguistically relevant do not necessarily show if there are larger-scale differences. For example, the f0 range of boundary tones generally comes out in different clusters before scaling or alignment differences of pitch accents are revealed (see also discussion in [5]). This is exemplified by the German data in this paper. Although some unexpected differences were found, the potential differences between contrastive and non-contrastive pitch accents on the subject were not entirely revealed with 5 clusters. It is likely that the unexpectedly accented object masked some of these differences, however, more research is needed to confirm this hypothesis.

Evaluating hierarchical cluster analyses using entropy-based metrics of information cost are a crucial step towards a principled choice of the number of clusters for a particular dataset. More general critical notes should be made, however. The outcome of the analyses presented here do not constitute a basic set of phonologically relevant f0 movements that compare to an autosegmental-metrical analysis. This is in part due to the fact that none of the datasets used here consisted of all possible contours of a language. In this respect it is important to note that in the typological overviews ([3],[4]) the inventories of phonologically relevant f0 movements had a weaker empirical basis. Contour clustering is able to bridge this gap by obtaining reproducible and entirely data-driven results showing which contours are (linguistically) relevant. It should also be noted that the outcome in terms of prototypical contours obtained by CA does heavily rely on the type of data under investigation. None of the datasets used in this paper are fully representative of *the prosody* of a certain language. This is a natural result of the data being collected in specific contexts. It is therefore crucial that multiple carefully-obtained datasets from one language are combined in order to provide a representative sample. In addition, perception research should be incorporated to further verify hypotheses created on the basis of any contour clustering.

5. Acknowledgements

Research for this paper was funded by the German Research Foundation (DFG) Project-ID 281511265 SFB 1252. The authors thank Heiko Seeliger for providing the German data.

6. References

[1] D. R. Ladd, *Intonational phonology*, 2nd ed., ser. Cambridge studies in linguistics. Cambridge ; New York: Cambridge University Press, 2008.

[2] N. P. Himmelmann and D. R. Ladd, "Prosodic Description: An Introduction for Fieldworkers," *Language Documentation & Conservation*, vol. 2, no. 2, pp. 244–274, Dec. 2008. [Online]. Available: http://scholarspace.manoa.hawaii.edu/handle/10125/4345

[3] S.-A. Jun, Ed., *Prosodic typology: the phonology of intonation and phrasing*, ser. Oxford linguistics. Oxford ; New York: Oxford University Press, 2005.

[4] ——, *Prosodic typology II: the phonology of intonation and phrasing*, ser. Oxford linguistics. Oxford: Oxford University Press, 2014.

[5] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, pp. 1–30, 2021. [Online]. Available: https://doi.org/10.1017/S0025100321000049

[6] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering A decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733

[7] A. Growendt and H. Rglin, "Improved Analysis of Complete-Linkage Clustering," *Algorithmica*, vol. 78, no. 4, pp. 1131–1150, Aug. 2017. [Online]. Available: http://link.springer.com/10.1007/s00453-017-0284-6

[8] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software*, vol. 61, no. 6, 2014. [Online]. Available: http://www.jstatsoft.org/v61/i06/

[9] A. D. Gordon, *Classification*. Boca Raton: Chapman & Hall/CRC, 1999, oCLC: 152674146. [Online]. Available: http://www.crcnetbase.com/isbn/9781584880134

[10] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978. [Online]. Available: http://doi.org/10.1016/0005-1098(78)90005-5

[11] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948. [Online]. Available: http://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[12] E. Sober, *Ockham's razors: a user's manual*. Cambridge: Cambridge University Press, 2015.

[13] R Core Team, "R: the R project for statistical computing," 2020. [Online]. Available: https://www.r-project.org/

[14] S. Repp, "The Prosody of *Wh* -exclamatives and *Wh* -questions in German: Speech Act Differences, Information Structure, and Sex of Speaker," *Language and Speech*, vol. 63, no. 2, pp. 306–361, Jun. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0023830919846147

[15] H. Seeliger and C. Kaland, "Boundary tones in German wh-questions and wh-exclamatives - a cluster-based approach," in *Proceedings of the 11th International Conference on Speech Prosody 2022*, S. Frota and M. Vigrio, Eds., Lisbon, Portugal, 2022, pp. 27–31.

[16] A. Kluge, *A grammar of Papuan Malay*. Berlin, Germany: Language Science Press, 2017. [Online]. Available: http://langsci-press.org/catalog/book/78

[17] C. Kaland and N. P. Himmelmann, "Time-series analysis of F0 in Papuan Malay contrastive focus," in *Proceedings of the 10th International Conference on Speech Prosody 2020*, 2020, pp. 230–234. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2020-47

[18] Language Archive Cologne, "Beria," 2018, publisher: re3data.org - Registry of Research Data Repositories. [Online]. Available: http://doi.org/10.17616/R3JV4W

[19] Tourneux, H., "Inventaires phonologiques et formation du pluriel en zaghawa (Tchad)," *Afrika und bersee*, vol. 75, no. 2, pp. p.267–277, 1992, place: Hamburg. [Online]. Available: https://db.degruyter.com/view/IABO/iab19933512

[20] A. M. Wolfe, "Towards a Generative Phonology and Morphology of the Dialects of Beria," Master's thesis, Harvard University, 2001, pages: 110.

[21] S. N. Osman, "Phonology of the Zaghawa Language in Sudan," in *Proceedings of the 9th Nilo-Saharan Linguistics Colloquium*, ser. Nilo-Saharan, A.-A. Abu-Manga, L. G. Gilley, and A. Storch, Eds. University of Khartoum: Kln: Kppe, 2006, meeting Name: Nilo-Saharan Linguistics Colloquium OCLC: ocm77077755.

DATA FROM ONLINE PRODUCTION EXPERIMENTS AND CHALLENGES FOR COLLECTING GOOD-QUALITY RECORDINGS FOR PROSODIC ANALYSES

Dorotea Bevivino, Barbara Hemforth, Giuseppina Turco

Université Paris Cité, LLF, CNRS
dorotea.bevivino@u-paris.fr, barbara-edith.hemforth@cnrs.fr, giuseppina.turco@cnrs.fr

ABSTRACT

As an effect of pandemic-related restrictions, experimental research has been challenged to find alternative methods for data collection and analyses. When it comes to laboratory phonology and production studies, few experiments were run online so far, mainly measuring accuracy and latencies. The present study used speech data collected online via a crowd-sourcing platform to perform prosodic analyses on sentence productions. We first ran a pilot study with six participants to test audio quality and feasibility. After preliminary analyses, we conducted a large-scale experiment with remote speech production data collection. The results suggest that we can collect online good-quality recordings for prosodic analyses, particularly duration measurements. However, reduced audio quality and increased accuracy demand for large samples result in increased time and effort to collect and process the desired amount of usable data.

Keywords: online experiments, remote recordings, language production, prosodic analyses, Prolific

1. INTRODUCTION

In recent years, Covid-19 lockdowns and less restricted pandemic-related measures later on have significantly affected the way research is conducted. Experimenters have been challenged to find alternative methods for data collection and analyses not to stop research entirely. Despite recent concerns about low participant validity [1, 2], crowd-sourcing platforms have become increasingly common. Sophisticated experimental features have been implemented on web platforms, and multiple paradigms have been successfully adapted and validated for online data collection in various domains [3, 4, 5]. When it comes to phonetic sciences, remote speech data collection is not new. Some applications had already been developed to investigate production in less-documented languages [6], and remote recordings had been tested for clinical purposes [7]. More recently, new functionalities have been added to common experimental web platforms, and in the last year, some specialized tools have been developed for remote collection of production data [8, 9].

The feasibility of online platforms, web browsers, and everyday devices as alternative equipment for speech studies has been tested and validated in several meta-studies. The intelligibility of the audio stimuli has been confirmed in perception [10]. The time accuracy of the recording devices has been investigated in production [11, 12]. Moreover, several studies have simultaneously recorded and thoroughly compared the sound quality of remote alternatives and high-quality recorders, to test the reliability of acoustic parameters extractions. These extra-controlled multi-tests suggest a distortion effect on all measurements to some extent [13], with F0 being more robust than other formants [14, 15] or other acoustic parameters [16], but still giving pitch visualization problems if HNR levels are poor [17].

Despite this growing body of tools and methodological literature on remote speech data collection, to the best of our knowledge, only a handful of production experiments were indeed run online in the past years. Remarkably, these experiments mainly measured latencies and accuracy [18, 19, 20, 21] without performing more fine-grained acoustic measurements.

It is worth noting at this point that one online production experiment investigating acoustic and prosodic measurements [22] appeared while we were implementing our study. Although both [22] and the present study investigate prosodic measurements from speech data collected online, the two studies differ with respect to some key elements, and we think they independently contribute to the field. Specifically, in [22], participants were recruited from the lab pool (and not a crowd-sourcing platform); the experiment was implemented on a web-platform specifically designed for the lab (and not a widely accessible experimental platform); experimenters collected

multi-word naming productions (and not full sentence productions in a priming paradigm); duration measurements were taken on the entire multi-word utterance and in-between word pauses (and not in specific critical regions of the sentence in the continuous speech stream).

In this paper, we report our experience conducting a speech production experiment online to perform prosodic analyses. We first ran an online pilot study to test audio quality and feasibility. After preliminary analyses and necessary technical adjustments, we ran a large-scale study. Below, we present procedures and challenges we encountered while collecting and processing speech production data online, using a crowd-sourcing platform and web-based experimental software. The data were collected for a large-scale experiment aiming at performing prosodic analyses on specific regions of sentence productions in British English. We believe that both the successful parts of our procedures and the challenges we faced will be helpful for the community to run online production experiments more smoothly.

2. METHODS

Participants' eligibility criteria, experimental materials and design, data collection and data processing procedures were identical for the pilot and the full-scale experiment, unless otherwise specified. The study design and analysis plan were preregistered [23], and all procedures received ethical approval.

2.1. Participants

Young adult native English speakers from the same UK area were recruited online via Prolific, controlled for gender, age, and education level. Six participants took part in the pilot study; 60 in the full-scale experiment. All participants had normal or corrected-to-normal vision, no hearing impairments, and no known neurological, speech, or communication disorders at the time of testing. All participants provided informed consent.

2.2. Study design

Participants were tested on a prosodic priming task. The task consists in repeating a series of sentences auditorily (primes) or visually (targets) presented (see [24] for details on materials and design). The experimental session lasted around 30 minutes, including setup, questionnaires, and practice.

2.3. Technical setup

The task was coded using the Penn Controller for Ibex [25], taking advantage of its MediaRecorder element. The script was implemented and run on the Ibex farm server installed and maintained by CNRS technicians at Université Paris Cité. The server was properly set up to host the recordings. Participants were directly redirected to the experiment website from Prolific. All recordings were made from participants' personal computers, using either the built-in speakers and microphones or external ones.

Device restriction labels (computer, audio, microphone) were included in the study preview on Prolific, to inform participants beforehand and restrict participation. Also, all technical requirements were clearly stated in the study description on the website. Specifically, participants were explicitly advised not to use mobile phones or tablets to join the study, as submissions from these devices would be rejected. We restricted participation to computer-only users primarily because of the nature of the main priming experiment, which required stimuli to be displayed in full-screen size. At the same time, collecting recordings only from computers helped minimize noise due to different devices. To ensure participants only used computers, we checked for screen-width at the beginning of the script and denied access to the experiment if needed.

Similarly, participants were explicitly informed that the task was solely compatible with desktops and laptops using the Chrome browser. Participants were advised against using a different web browser, to avoid the experiment crashing and their submission being rejected. We made the experiment accessible only via Chrome to ensure the most common combination of operating system and web browser [12] and guarantee the stability of the testing platform. Once again, restricting participation to a single web-browser helped minimize potential extra noise in the data.

Along with device and web browser restrictions, participants were explicitly asked to wear a headset and be in a quiet space prior to accepting participation in the study. It was clearly stated that no data analyses would be possible on unclear audios, resulting in a rejected submission. Following [17, 26] suggestions, after being redirected to the experimental page, participants were provided with detailed guidelines to recreate a lab-like environment and ensure good-quality recordings (checking on the audio settings, cancelling excessive ambient noise, avoiding

electronic interference, etc.). No further checks were implemented to ensure that participants actually recorded alone in a silent room and used headsets, due to ethical and practical restrictions on video-recording the experimental sessions.

2.4. Recording procedure

Before starting the practice trials, participants were invited to perform an audio quality check. The recorder interface was displayed on the screen, and participants were asked to record and replay a short precise sentence, while paying attention to the sound quality. Suggestions for improving audio quality were also displayed again. Participants were invited to freely use the interface to take multiple tests until their audio was clear. To encourage engagement with the audio test, the button to proceed to the next trial was displayed only after a short delay.

In the regular experimental trials, the recording interface was no longer displayed on the screen, and participants were instructed to simply repeat sentences using the spacebar to advance. Participants were asked to repeat each sentence, even if not accurate, and to say *Pass* or similar if they could not recall anything. This would enable us to avoid empty recordings and immediately detect potential technical issues or bad-faith participants.

Each recorded sentence was immediately uploaded to the server in the background. At the end of all trials, completion screens remained on the screen until all recordings were uploaded to the server. In the full-scale experiment, participants were encouraged to download the recording zip file locally, to be able to send it to experimenters in case of connection failure or server problems before the upload was complete.

2.5. Data processing and analysis

Each recorded sentence was automatically saved as a separate WebM file and stored in zip files with unique names by PCIbex. WebM files were converted offline to .wav files for analysis. The audio-recorded data were transcribed automatically using Praat [27] scripts. Each file was then manually checked for sound problems, accuracy errors, and to correct any discrepancies between the automatically transcribed expected productions and the speakers' actual productions. The data were hence forced-aligned using the Montreal Forced Aligner [28].

For each repeated sentence, we measured the duration at two specific critical regions, to address the research question of our main experiment. In the full-scale experiment, we also extracted F0 values from a subset of data using ProsodyPro [29], to run exploratory analysis of pitch modulation.

3. PILOT RESULTS

Due to issues with the server and how the original Ibex farm hosts multimedia files, 4 participants experienced difficulties loading the audio stimuli. As a result, their experimental session was slowed down or interrupted. Hence, we listened to and overall evaluated all the collected recordings. More fine-grained analyses were performed only on a subset of data specifically selected to investigate the research question of the main experiment.

After the overall evaluation, none of the 6 participants was excluded for large number of blank recordings or general poor audio quality. In the subset selected for prosodic analyses (N = 158 recordings), no recordings were discarded due to background noise or poor sound quality, few recordings were discarded due to blank (N = 1) or *Pass* (N = 2) productions, several recordings were excluded because the sentence ROI was missing (N = 9) or presented disfluencies (N = 1) (see Fig. 1). No specific problems were encountered when conducting the measurements and running scripts at any stage of the data processing.

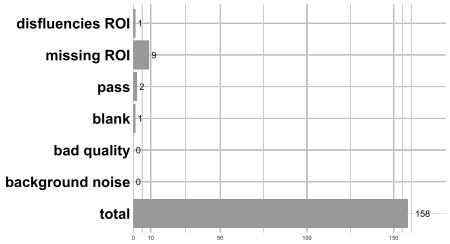


Figure 1: Raw counts of discarded recordings by error type in the subset of pilot data.

These preliminary results attested the general good quality of the recordings to perform prosodic analyses. Thus, online collection of speech production data seemed not only to be feasible and sufficiently reliable for accuracy and latencies but also for prosodic analyses. The results of the pilot study allowed us to proceed safely with the pre-registered experiment on a greater scale and including extended analyses.

4. FULL-SCALE EXPERIMENT RESULTS

Once the resource loading problem was fixed, no participants in the full-scale experiment reported any issues preventing them from completing the task.

Two participants reported being slowed down due to misreading the instructions (they did not press the spacebar and instead let the experiment dictate the pace). With an original target sample size of approximately 80 participants, we began analyzing the data after collecting recordings from the first 40.

After overall preliminary analyses to assess the audio quality of the recordings, we excluded 5 participants due to constant electronic interference and/or severe background noise (e.g., people around, street noise, dog barking, mouse clicking, etc.). During the transcription process, we further excluded 12 participants who produced unusable recordings for more than 20% of their productions. This resulted in a sample of 23 participants out of the 40 whose data were analyzed. Similarly, we did not include in the final analyses bad recordings (e.g., blank or *Pass* productions, noises, fragments, disfluencies at/near the critical regions) from good participants, resulting in a final sample of 1625 productions out of the total 1830 sentences produced by the 23 good participants (see Fig. 2). Due to the large number of excluded participants and recordings, we decided not to proceed further with data processing for the other 20 participants, but rather to replace excluded participants to ensure a balanced sample.

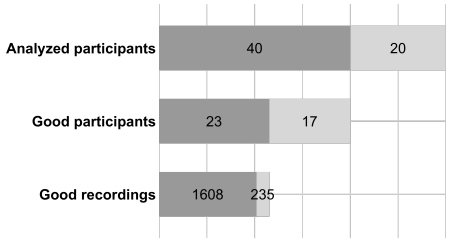


Figure 2: Raw counts of included vs. excluded participants and recordings in the full-scale experiment.

Alike the pilot data, no particular issues arose during file conversion, or measurements, or script execution for speech-to-text alignment or duration extraction. All extracted duration data were successfully and reliably modeled to address the research question of the main experiment [24].

Regarding the exploratory analysis of pitch modulation conducted on a subset of data, we again did not encounter any technical issues while running the scripts to extract normalized F0 curves across speakers. However, an initial examination of the plotted curves revealed a high number of sudden jumps and considerable variability in the curves, rendering pitch visualization unreliable.

5. GENERAL DISCUSSION

Our results suggest that we can conduct production experiments online, via crowd-sourcing platforms and web-based experimental softwares, and still collect good-quality recordings. These recordings not only seem to be reliable for accuracy and latencies, but also for prosodic analyses on sentence productions, particularly duration measurements.

The observations on F0 are in line with previous reports of pitch visualization problems [17]. Also, it is worth noting that for this exploratory analysis, we did not perform any landmarking procedure to normalize timing differences in the individual productions. The nature and structure of our stimuli did not allow for a straightforward isolation of the ROIs. The lack of landmarking may account for some variability and alignment issues in the data. However, it still remains unclear to what extent the variability resulting from different devices, HNR levels, recording distances, and noisy environments might still affect intonational measurements in online data collection.

Notwithstanding the reliability of data for duration measurements, in our full-scale study, we had to discard a large number of recordings (up to roughly half of the analyzed participants) due to technical and behavioral issues that could have been avoided in a lab setting. This is despite the emphasis on the critical importance of audio quality and all reminders provided to participants on the environment setup. While reduced audio quality and increased accuracy demand are physiological for large samples, our results seem rather to denote a general lack of precision and attention during online tasks, in line with previous reports on online participants, even when recruited from lab pools [19]. These findings also point out how our data suffered from the lack of monitoring of participants and their productions during the task, a standard practice in lab data collection. All of this results in increased time and effort required to collect and process the desired amount of usable data.

In conclusion, our study suggests that remote speech data collection is a feasible and reliable option not only for accuracy and latencies but also for prosodic analyses, particularly duration. This approach is a potentially valuable strategy for extending research on less-documented languages or for collecting naturalistic interactional data. However, it appears more suitable for small samples and simple tasks, where noise and attention-related issues can be minimized.

6. REFERENCES

[1] M. A. Webb and J. P. Tangney, "Too good to be true: Bots and bad data from Mechanical Turk," *Perspect. Psychol. Sci.*, 2022.

[2] A. J. Moss, C. Rosenzweig, S. N. Jaffe, R. Gautam, J. Robinson, and L. Litman, "Bots or inattentive humans? Identifying sources of low-quality data in online platforms," *PsyArXiv*, 2021.

[3] M. Vos, S. Minor, and G. C. Ramchand, "Comparing infrared and webcam eye tracking in the visual world paradigm," *Glossa Psycholinguistics*, vol. 1, 2022.

[4] M. S. Slim and R. J. Hartsuiker, "Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js," *Behav. Res. Meth.*, 2022.

[5] A. Chuey, M. Asaba, S. Bridgers, B. Carrillo, G. Dietz, T. Garcia, J. A. Leonard, S. Liu, M. Merrick, S. Radwan, J. Stegall, N. Velez, B. Woo, Y. Wu, X. J. Zhou, M. C. Frank, and H. Gweon, "Moderated online data-collection for developmental research: Methods and replications," *Front. Psychol.*, vol. 12, 2021.

[6] "Lessons learned after development and use of a data collection app for language documentation (lig-aikuma)," in *Proc. 19th ICPhS*, Melbourne, 2019.

[7] S. Jannetts, F. Schaeffler, J. Beck, and S. Cowen, "Assessing voice health using smartphones," *Int. J. Lang. Commun. Disord.*, vol. 54, no. 2, pp. 292–305, 2019.

[8] A. L. Thomas and P. F. Assmann, "SpeechCollectR: An R package for web-based speech data collection," *J. Acoust. Soc. Am.*, vol. 150, no. 4, pp. A356–A357, 2021.

[9] T. T. Schnoor and B. V. Tucker, "SpeakerPool: A remote speech data collection platform," *J. Acoust. Soc. Am.*, vol. 152, no. 4, pp. A60–A60, 2022.

[10] M. Cooke and M. L. Garcia Lecumberri, "How reliable are online speech intelligibility studies with known listener cohorts?" *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1390–1401, 2021.

[11] A. Anwyl-Irvine, E. S. Dalmaijer, N. Hodges, and J. K. Evershed, "Realistic precision and accuracy of online experiment platforms, web browsers, and devices," *Behav. Res. Meth.*, vol. 53, no. 4, pp. 1407–1425, 2021.

[12] D. Bridges, A. Pitiot, M. R. MacAskill, and J. W. Peirce, "The timing mega-study: Comparing a range of experiment generators, both lab-based and online," *PeerJ*, vol. 8, 2020.

[13] C. Sanker, S. Babinski, R. Burns, M. Evans, J. Johns, J. Kim, S. Smith, N. Weber, and C. Bowern, "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," *Language*, vol. 97, no. 4, 2021.

[14] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 3910–3916, 2021.

[15] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements," in *Interspeech 2021*, Brno, 2021.

[16] J. Penney, A. Gibson, F. Cox, M. Proctor, and A. Szakay, "A comparison of acoustic correlates of voice quality across different recording devices: A cautionary tale," in *Interspeech 2021*, Brno, 2021.

[17] G. Magistro, "Speech prosody and remote experiments: Atechnical report," *PsyArXiv*, 2021.

[18] M. Kandel, C. R. Wyatt, and C. Phillips, "Agreement attraction error and timing profiles in continuous speech," *Glossa Psycholinguistics*, vol. 1, no. 1, 2022.

[19] A. Fairs and K. Strijkers, "Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors," *PLOS ONE*, vol. 16, no. 10, p. e0258908, 2021.

[20] A. Vogt, R. Hauber, A. K. Kuhlen, and R. A. Rahman, "Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice," *Behav. Res. Meth.*, vol. 54, no. 4, pp. 1954–1975, 2022.

[21] K. Stark, C. van Scherpenberg, H. Obrig, and R. Abdel Rahman, "Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities," *Behav. Res. Meth.*, 2022.

[22] J. He, A. S. Meyer, A. Creemers, and L. Brehm, "Conducting language production research online: A web-based study of semantic context and name agreement effects in multi-word production," *Collabra: Psychology*, vol. 7, p. 29935, 2021.

[23] D. Bevivino, B. Hemforth, and G. Turco, "Prosodic Priming (Replication of Tooley et al. 2018)," 2022. [Online]. Available: osf.io/z24ke

[24] D. Bevivino, G. Turco, and B. Hemforth, "Priming prosodic boundaries across constructions," in *Linguistic Evidence 2022*, Paris, 2022.

[25] J. Zehr and F. Schwarz, "PennController for Internet Based Experiments (IBEX)," 2018.

[26] E. James, "Tools and tips for collecting spoken responses online," 2020.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," [Computer program], 2021. Version 6.1.56.

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*, Stockholm, 2017.

[29] L. Xu, "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis," in *TRASP 2013*, 2013, pp. 7–10.

7. ACKNOWLEDGEMENTS

This work was financed by Labex EFL (ANR-10-LABX-0083). It contributes to the IDEX Université Paris Cité (18-IDEX-0001).

GETTING FROM A TO B: COMPLEXITIES OF TURN CHANGE AND RETENTION IN CONVERSATION

Emer Gilmartin[1,2], Marcin Włodarczak[3]

[1]ADAPT Centre, Trinity College Dublin, Ireland; [2]INRIA Paris, France

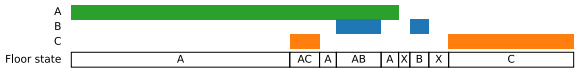
[3]Department of Linguistics, Stockholm University, Sweden
gilmare@tcd.ie, wlodarczak@ling.su.se

ABSTRACT

Spoken interaction proceeds as a series of contributions or utterances from participants. The question of who 'has the turn' and how turns are managed is fundamental to understanding of all genres of spoken interaction, and to conversation in particular, as turntaking is managed locally by participants. We use floor state, descriptions of who is speaking or not at any time, to study between and within floor state transitions (BST and WST), sequences of activity from one substantial stretch of single party speech to the next, analagous to turn change or retention. We explore the patterns of speech and silence in the Switchboard corpus of over 2000 casual two-party telephone conversations using an annotation and analysis paradigm that has previously exposed interesting features of multiparty talk. We find that over half of the transitions observed in Switchboard involve more than one interval of silence, a phenomenon first noted in multiparty talk. We also find that although the transition distribution in Switchboard's dyadic conversations broadly follows patterns found in multiparty talk, there are fewer complex transitions observed. **Index Terms**: dialogue, corpus studies, turntaking, human-computer interaction

1. INTRODUCTION

The arrangement of participant contributions to spoken interaction, particularly during turn change or retention, is of fundamental interest to studies of speech and dialog, and has become a topic of interest in speech and language technology with increasing uptake of natural language interfaces. Consequently, patterns of speech and silence in interaction have been widely used in large scale analyses of dialogue, often involving machine learning, to infer a wide range of information about spoken interaction – examples include predicting speaker activity from conversation history [1, 2, 3], inferring speaker and context characteristics [2, 4], or extracting personality traits of speakers in dyadic and multiparty interaction [5, 6, 7]. However, in much of this work, the focus is on optimising models' predictive power over entire conversations rather than on elucidating the specifics of how conversation works locally.



A . . . cameras. You're just not normally as aware of it.
C Yeah
B It's true.
B Yeah.
C There's two on every bus.

Figure 1: An excerpt from a casual conversation in English corresponding to a between-speaker transition, **A_AC_A_AB_A_X_B_X_C**, from speaker A to C with seven intervening intervals (**AC**, **A**, **AB**, **A**, **X**, **B**, **X**). *Top*: A diagram showing temporal organization of individual speakers' contributions (represented as color bars) and the resulting floor states. *Bottom*: A simplified transcript. Speakers' contributions are color-coded for consistency.

In this paper we examine how interaction progresses in the Switchboard corpus [8], a large collection of dyadic phone conversations, and compare the results to patterns previously observed in multiparty conversation. To do this we use an annotation and analysis paradigm developed and previously employed to analyse a range of smaller corpora of multiparty dialogue. The analysis is based on the concept of *floor state* - who is speaking or silent at any moment during interaction. By annotating *floor state intervals*, stretches of time during which a particular floor state holds, we can analyse *floor state transitions* or sequences of contiguous floor states. While these transitions can record any sequence of conversational activity, we are particularly interested in sequences of activity between longer stretches of single party speech, as these transitions can throw light on how turn-taking is accomplished. In order to track changes in floor possession, we further categorize the sequences as

either *between speaker transitions* (BST) or *within speaker transitions* (WST). In WST, the speaker on either side of the transition is the same, analogous to a turn retention, while in BSTs, the single party speech bounding the transition is produced by different speakers, analogous to a turn change. Figure 1 shows an example of a short exchange from a three-party conversation to illustrate the labelling scheme and concepts. The example involves nine floor states – solo speech (**A**, **B**, **C**), overlaps (**AC**, **AB**) and general silence / nobody speaking (**X**). Existing data-driven approaches to turn-taking could treat this stretch as a series of four transitions: **A_AC_A** from A to A, **A_AB_A** from A to A, **A_X_B** from A to B, and **B_X_C** from B to C. However, looking at the transcript and the speech patterns, it seems more likely that the *longer* stretches of solo speech (**A**, **C**) delimit a single more complex transfer of floor possession from speaker A to C. In the analysis described below, we use floor state transitions to discover these larger conversational structures which can be otherwise overlooked by large-scale corpus studies.

2. BACKGROUND

Previous work on floor state transitions in several collections of multiparty dialogue has identified interesting similarities – for all multiparty corpora studied less than half of between and within speaker transitions are accomplished with a single interval of silence or overlap, indicating that turn change and retention usually involve more complex sequences of speech (either in the clear or involving overlap) and silence. It has also found high levels of uniformity in the most common WSTs and BSTs found in different languages and settings [9, 10, 11, 12, 13]. In a study of speaker transitions in multiparty three-, four- and five-party casual conversation in English, one-interval transitions were the largest category, with a higher proportion of one-interval transitions in WST, perhaps reflecting breath pauses or single backchannels during monologic stretches [10]. Nevertheless, the majority of transitions involved more than one intervening interval to complete. In work on English, Estonian and Swedish three-party casual face-to-face conversations [14], similar patterns were observed, with over 95% of transitions completed in 15 or fewer intervening intervals, and over 65% involving more than one interval, with a higher likelihood of more complex transitions in the BST category. Additionally, the minimum duration threshold imposed on single-party speech affected the transition type label and number of intervening intervals assigned, with transition type label (WST or BST) changing for 28.29% of transitions if the right duration hand threshold of one second was removed. These results imply that turn change and retention cannot be comprehensively modelled without consideration of both left and right hand context - if the right hand minimum duration threshold is relaxed, most transitions revert to one-interval and the complexity of what actually happens in real conversation is lost. Subsequent work investigated the more common transitions in three-party casual and task-oriented data [12], finding considerable complexity and growing incidence of participation by more speakers with transition length. Even though very many different transition labels were present in the data, a small subset of 17 labels (4 one-interval and 13 three-interval) accounted for most of floor state transitions occurring.

Having replicated these results across a number of multiparty corpora, below we (1) analyse the Switchboard corpus of dyadic phone conversations to investigate whether these findings on multiparty talk also hold for the Switchboard conversations, and (2) compare the characteristics of within and between speaker transitions in Switchboard with those of other corpora analysed previously.

3. DATA AND ANNOTATION

The analysis was based on the Switchboard-1 Telephone Speech Corpus (Release 2) [15], comprising 2438 dyadic phone conversations between 542 speakers of American English (302M/241F). Speech and silence labels for each participant were derived from word level transcription [16], with all non-speech sounds suppressed to silence, resulting in a set of 520135 talkspurts (interpausal units comprising speech from a participant bounded by silence from the same participant) from 259 hours of conversation. The speech/silence labels were used to generate labels for each floor state interval in the corpus, which were combined to form transition labels, stretches of conversation starting with an interval of single party speech in the clear of at least one second in duration and continuing through the next one-second interval of single party speech in the clear encountered in the data. Each transition was classified as either between-speaker (BST) or within-speaker (WST), depending on whether it was bounded by speech from the same or two different speakers.

We used three-party dialogue data from

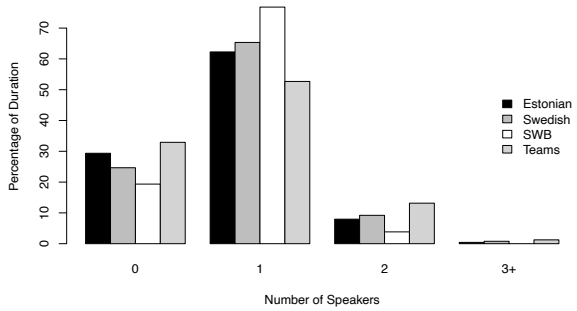


Figure 2: Distribution of silence (0 speakers), single-party speech (1 speaker) and overlap (2 or more speakers) as percentage of total conversation time for Switchboard (SWB) and for the English, Estonian and Swedish three-party corpora.

previous studies to compare with the Switchboard results. These data are three-party spontaneous conversations in Estonian [17] and Swedish [18], and collaborative conversational games in English [19]. All the data had been segmented manually. The three-party data set contained 22106 talkspurts in 9 hours and 51 minutes hours of conversation.

4. RESULTS

Results are first presented for the Switchboard corpus, and then contrasted with results from previous work on multiparty corpora.

Distribution of Speech, Silence and Overlap There were 1,042,516 floor state intervals identified in Switchboard – 415,552 silent (0 speaker) intervals accounted for 19.35% of conversation time, 520161 single-party speech for 76.82%, and 106,993 two-party overlap for 3.83%. The proportions of single-party speech, silence and overlap in Switchboard are shown along with the proportions in the three-party data for comparison in Figure 2. It can be seen that Switchboard has lower incidence of silence and overlap than the other data sets, and higher incidence of single-party speech in the clear.

Distribution of speaker transitions The Switchboard data set contained 256,655 speaker transitions in 259 hours of talk, an average of one every 3.7 seconds. In the three-party data, there was an average of one transition every 4.7 seconds.

Figure 3 shows relative frequencies for odd number interval transitions, further split between BST and WST classes, in the four data sets. Similar to the three-party corpora, the vast bulk (over 99%) of transitions in Switchboard comprised less than 16 intervening intervals. There are vanishingly few transitions involving even numbers of intervening intervals (47 out of 256,655). This low number is unsurprising as such transitions would involve smooth switches or simultaneous onset or offset of speech, which are very rare in conversation.

One-interval transitions are the largest class in Switchboard, and the frequency of transitions decreases with increasing numbers of intervening intervals: 47.72% of all transitions (41.65% of BSTs and 50.03% of WSTs) were accomplished with one intervening interval, 27.14% (24.77% of BSTs and 28.15% of WSTs) with two intervening intervals, and 12.86% (15.98% of BSTs and 12.16% of WSTs) with 3 intervening intervals.

Overall, 78.28% of transitions are WST, greatly outnumbering BST transitions. WST are particularly frequent in short transitions, accounting for 81% of one-interval transitions, 80% of three-interval, and falling with increasing numbers of intervals to 60% of 15 interval transitions.

Most Common Transition Sequences In the Switchboard data, we found all 4 possible one-interval transition sequences, two BSTs (A_X_B,A_AB_B) and 2 WSTs (A_X_A,A_AB_A). There were 16 three-interval transition sequences (8 BSTs, 8 WSTs), and 64 five-interval transition sequences (32 BSTs, 32 WSTs).

The most frequent transition types in Switchboard overall was A_X_A at 36.69% of all transitions, followed by A_X_A_X_A (13.1%) and A_X_B (6.55%). The most common WST was A_X_A, followed by A_X_A_X_A and A_X_B_X_A - all of which involve only single party speech and silence. The most frequent WST involving overlap was a three-interval transition, A_X_B_AB_A, with the one-interval overlap WST A_AB_A appearing as the 6th most common WST. For BSTs, the most common was A_X_B followed by A_AB_B. Transitions involving overlap were more frequent in BSTs than in WSTs.

5. DISCUSSION

The distribution of speaker transitions in Switchboard largely reflects the patterns found in the three-party data used for comparison (and also in four- and five- party data analysed in [11]. The largest category are one-interval transitions, even-number interval transitions are extremely rare, and the number of transitions drops off with increasing numbers of intervals. The proportion of one-interval transitions in Switchboard is greater than that seen in the three-party corpora, but still only accounts for 47.7% of all transitions, highlighting the fact that most transitions involve

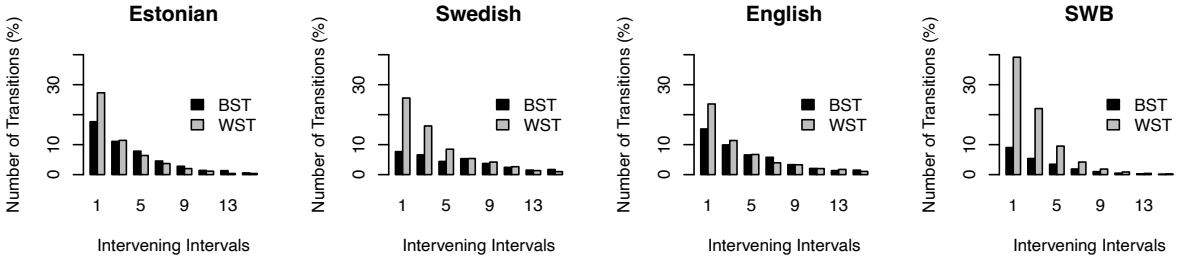


Figure 3: Distribution of between- and within-speaker transitions in Switchboard (SWB) and the English, Estonian and Swedish three-party corpora.

more than a simple single silence or overlap, even in dyadic phone conversations.

The split between within- and between- speaker transitions in Switchboard reflects that found in the three-party data in that there are more WSTs present. WSTs in Switchboard dramatically outnumber BSTs, more than in the three-party data. This could reflect long turns being taken in Switchboard, perhaps as a consequence of the fact that participants were strangers, or indeed, be a feature of telephone conversation.

The results on the distribution of specific transitions in Switchboard are very interesting to compare to those found on the three-party data used in this paper. The most frequent transition types in Switchboard was A_X_A at 36.69% of all transitions, followed by A_X_A_X_A (13.1%) and A_X_B (6.55%). In the 3 party data, the most common sequences overall were A_X_A (within-speaker silence) and A_X_B (between-speaker silence). The prominence of A_X_A_X_A in Switchboard reflects the high proportion of within-speaker transitions seen in Switchboard compared to the three-party data. For WSTs, the ranking in three-party data was similar to that in Switchboard, with both A_X_A_X_A and A_X_B_X_A (three-interval WSTs involving only single party speech and silence) appearing more frequently than one-interval WSTs involving overlap (A_AB_A). As with Switchboard, the top two three-party BSTs were one-interval silence and overlap (A_X_B and A_AB_B). Transitions involving overlap were more frequent in BSTs than in WSTs in both Switchboard and the three-party data.

Previous work on multiparty talk [12] has found that participation by all speakers becomes more likely in transitions involving more intervals, leading to quite complex and varied possibilities. However, in the dyadic Switchboard data, the number of transitions drops more rapidly as the number of intervals increases (see Figure 2).

Switchboard is also characterized by less silence and overlap and more speech in the clear than the three-party conversations. This may be due to the modality [20, 21] – on the phone, speakers may wait for their interlocutor to finish before commencing to speak, and may not give as much verbal feedback in overlap. It could also reflect differences between dyadic and multi-party talk.

In summary, our analysis of the 2400 Switchboard conversations has shown that more than half of all between and within speaker floor state transitions in these dyadic conversations involve more than one intervening interval of speech, silence or overlap between longer stretches of single party speech. This reflects previous results on multiparty spoken interaction, implying that turn change and retention even in dyadic phone conversations exhibit a level of complexity that is not covered by modelling them as a simple gap or overlap. This has implications for design of spoken dialog technology - while 'ping-pong' style turntaking may suffice for simple question-answer transactions, creating more convincing human-like interaction will require more nuanced modelling of turn taking.

6. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106. The work was also funded by Swedish Research Council project 2019-02932 *Prosodic functions of voice quality dynamics* to Marcin Włodarczak.

7. REFERENCES

[1] J. Jaffe, L. Cassotta, and S. Feldstein, "Markovian model of time patterns of speech," *Science*, vol. 144, no. 3620, pp. 884–886, 1964.

[2] B. Beebe, D. Alson, J. Jaffe, S. Feldstein, and C. Crown, "Vocal congruence in mother-infant play," *Journal of Psycholinguistic Research*, vol. 17, pp. 245–259, 1988.

[3] B. Beebe, J. Jaffe, F. Lachmann, S. Feldstein, C. Crown, and M. Jasnow, "Systems models in development and psychoanalysis: The case of vocal rhythm coordination and attachment," *Infant Mental Health Journal*, vol. 21, no. 1-2, pp. 99–122, 2000.

[4] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *Proceedings of SIGdial 2008*, Columbus, OH, 2008, pp. 148–155.

[5] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[6] L. H. Gilpin, D. M. Olson, and T. Alrashed, "Perception of speaker personality traits using speech signals," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. LBW514.

[7] M. Yu, E. Gilmartin, and D. Litman, "Identifying personality traits using overlap dynamics in multiparty dialogue," in *Proceedings of Interspeech 2019*, 2019, pp. 1921–1925.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1992, pp. 517–520.

[9] E. Gilmartin, K. Aare, M. O'Reilly, and M. Wlodarczak, "Between and within speaker transitions in multiparty conversation," in *Proceedings of Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 799–803.

[10] E. Gilmartin, M. Yu, and D. Litman, "Comparing speech, silence and overlap dynamics in a task-based game and casual conversation," in *Proceedings of ICPhS 2019*, 2019, pp. 3408–3412.

[11] E. Gilmartin, "Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis," Ph.D. dissertation, Trinity College Dublin, 2021.

[12] M. Włodarczak and E. Gilmartin, "Speaker transition patterns in three-party conversation: Evidence from English, Estonian and Swedish," in *Proceedings of Interspeech 2021*, 2021, pp. 801–805.

[13] E. Gilmartin, C. Saam, C. Vogel, N. Campbell, and V. Wade, "Just talking - modelling casual conversation," in *Proceedings SIGdial 2018*, Melbourne, Australia, 2018, pp. 51–59.

[14] E. Gilmartin, K. Aare, M. O'Reilly, and M. Wlodarczak, "Between and within speaker transitions in multiparty conversation," in *Speech Prosody 2020*, 2020, pp. 799–803.

[15] J. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, CA, 1992, pp. 517–520.

[16] N. Deshmukh, A. Ganafileiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *Proceedings of ICSLP*, Sidney, Australia, 1998, pp. 1543–1546.

[17] P. Lippus, T. Tuisk, N. Salvestre, and P. Tiras, "Phonetic corpus of Estonian spontaneous speech." [Online]. Available: https://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech

[18] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Frontiers in Psychology*, vol. 8, p. 708, 2017.

[19] D. Litman, S. Paletz, Z. Rahimi, S. Allegretti, and C. Rice, "The teams corpus and entrainment in multi-party spoken dialogues," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1421–1431.

[20] L. ten Bosch, N. Oostdijk, and J. P. de Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *Proceedings of 7th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, 2004.

[21] F. Hammer, P. Reichl, and A. Raake, "Elements of interactivity in telephone conversations," in *Eighth International Conference on Spoken Language Processing*, 2004.

CHARACTERISTICS AND DISTRIBUTION OF SILENT PAUSES IN CONVERSATIONS BETWEEN AUTISTIC AND NON-AUTISTIC DYADS

Simon Wehrle[1], Kai Vogeley[2], Martine Grice[1]

[1]I*f*L-Phonetik, University of Cologne, Germany [2]Dept. of Psychiatry, University Hospital Cologne, Germany
{simon.wehrle, martine.grice}@uni-koeln.de; kai.vogeley@uk-koeln.de

ABSTRACT

Silent pauses are a pervasive feature of speech, but one that has not received much attention in the context of atypical populations. We investigated the use of silent pauses in conversations between dyads of autistic as compared to non-autistic adults. Previous research on silent pauses in Autism Spectrum Disorder (ASD) is limited to three studies, with contradictory results. We found many similarities in silent pause use between groups, but also a robust difference, as autistic speakers produced more long silent pauses. We also investigated the effect of filled pauses on the duration of subsequent silent pauses. While we found no group-level difference, we were able to replicate, but also qualify, the previous finding that *uhm* precedes longer silent pauses than *uh*. This study hence 1) is a substantial contribution to our understanding of conversational behaviour in ASD and 2) extends important previous findings on silent pauses in the general population.

Keywords: autism spectrum disorder; silent pause; filled pause; conversation; Bayesian modelling

1. INTRODUCTION

Silent pauses feature in the majority of spoken utterances, and they are particularly prevalent in conversational speech. While there is a solid amount of general research on the topic, and in the context of second-language speech in particular [1, 2], very little is known about the use of silent pauses in atypical populations, such as in the speech of persons diagnosed with Autism Spectrum Disorder (ASD).

Previous work on silent pause use in ASD seems to be limited to three studies comparing autistic speakers with matched controls (CTR), with contradictory results. In [3], fewer silent pauses in picture book narrations by English-speaking autistic children are reported. In contrast, [4] report a higher rate of silent pauses in interview-style conversations between experimenters and English-speaking autistic adults. Finally, [5] found equivalent silent pause rates for autistic and matched non-autistic adults in a sentence repetition task.

Crucial differences in the age of participants and/or speech material make comparison with our own data difficult in the cases of [3, 5]. As these confounding factors are less of a concern regarding the work by [4] (age range and speech data being similar to our own), we will focus on this latter study for comparison with our own findings.

The authors of [4] report a higher rate of silent pauses in autistic adults compared to non-autistic controls, but did not examine any silent pauses with a duration of under 2 seconds. The authors provide no specific reasons for using this extremely high cut-off point, only stating that "this was done in order to ensure that we excluded normal prosodic pauses" (p. 138). We can further call into question the sheer utility of such a threshold from a pragmatic–analytic point of view, as employing it entails excluding almost all silent pauses in a given data set; speakers from the control group in [4] in fact did not produce *any* silent pauses longer than 2 seconds.

For this study, we separately analysed 1) silent pauses of any duration, 2) a subset of silent pauses over 2 seconds in duration (for comparison with [4]) and 3) silent pauses of 700 ms or longer (a subset of 1) and a superset of 2); see Section 3.3 for rationale).

In addition to simply comparing the rate and duration of silent pauses, we decided to also probe for other potential group differences in the form of the distributional characteristics of silent pauses. Specifically, we investigated which effect the lexical form of a preceding filled pause ("uh" or "uhm") had on the duration of the following silent pause.

This is chiefly inspired by the highly influential work in [6] comparing the use of *uh* und *uhm* in spontaneous speech. The authors claim that there is a considerable difference in the average duration of silences following *uh* as compared to *uhm*, with *uhm* preceding silences of at least twice the duration of silences following *uh*. In a comparison of autistic and non-autistic children, [7] confirmed this effect for their CTR, but not their ASD group.

While [6] also showed, in a binary distinction, that silences following "lengthened" productions of both *uh* and *uhm* were considerably longer overall, the duration of the *uh* vs. *uhm* tokens themselves was not controlled for. In fact, to our knowledge, none of the subsequent papers examining this phenomenon involved an analysis that systematically controlled for the inherent average duration of *uh* and *uhm*.

This is a serious concern since, in our dataset at least, *uhm* is considerably longer (521 ms) than *uh* (329 ms) on average . Thus, it is important to establish whether and to what extent the effect ascribed to a difference in filled pause type (nasal vs. non-nasal) is in fact simply due to filled pause duration, independent of whether a final nasal was present (which for simple reasons of physiology and aerodynamics increases the likelihood of longer durations). We thus attempted to replicate the relevant effect while controlling for the confound of filled pause duration, all in the context of investigating differences between the ASD and CTR groups in our data set.

2. METHODS

2.1. Material and subjects

We analysed data from a corpus of approximately 5 hours of semi-spontaneous (Map Task) speech produced by 28 German adults [8–10]. Dialogues had a mean duration of 20 min. overall (SD = 12'; range: 9'–49'), and they tended to be shorter in the ASD (mean = 15'; SD = 7'; range: 9'–30') compared to the CTR group (mean = 26'; SD = 15'; range 13'–49').

Half of the participants (ASD group) had been diagnosed with ASD, the other half (CTR group) were matched for age, gender and verbal IQ. Dialogues took place in disposition-matched dyads, that is, both interlocutors in a dyad either did or did not have a diagnosis of ASD. We collected Autism Spectrum Quotient (AQ) scores for all participants and found that all ASD subjects scored above the suggested threshold of 32 points and all CTR subjects scored below the same threshold [11, 12]. Differences between groups were confirmed using Bayesian modelling (see below).

2.2. Data and methods

All dialogues were manually transcribed using *Praat* [13], and annotations were divided into either silences or inter-pausal units (IPU) of speech with a minimum duration of 200 milliseconds [2, 14, 15]. The corpus contains a total of 3473 silent pauses.

Portions of dialogue that contained only audible breathing, clicks, and similar noises were counted as being part of silent intervals. In contrast, all other speech sounds, including filled pauses, were counted as being part of IPUs. While we acknowledge that most such "silent" pauses are not completely silent from a strictly acoustic perspective [16], we chose to adhere to the conventional definition outlined above, since the main aim of this study is to enable comparison with the (sparse) previous literature on silent pauses in ASD as well as with the more general literature on the topic.

For the analysis of silences following filled pauses, we investigated all 1027 filled pause tokens in the data set, as well as their surrounding linguistic context (see [9] for a full analysis of filled pause use in this corpus). If filled pauses were followed not by any period of silence, but instead directly by another utterance (by either of the interlocutors), we assigned the following silence a duration of 0.

We used Bayesian linear modelling for statistical analysis [17–19]. All code, scripts, model specifications and data frames are available in the *OSF* repository at https://osf.io/bph2t/.

3. RESULTS

3.1. All silent pause tokens

The mean duration of silent pauses was close to identical across groups, with means of 677 ms (SD = 563) for the ASD and 628 ms (SD = 527) for the CTR group. The mean rate of silent pauses was exactly identical across groups, with a value of 12.2 silent pauses per minute. A dyad-specific analysis of silent-pause rates also gives no indications of ASD-specific behaviour; see Fig. 1. Note that there was a considerable degree of by-dyad variability and overlap between groups, not only for this analysis, but also all of the ones described below.

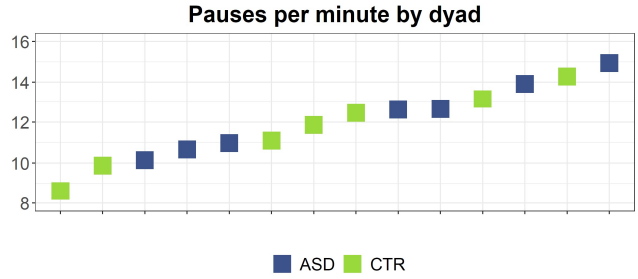


Figure 1: Rate of silent pauses by dyad and group (ASD in blue, CTR in green).

3.2. Silent pauses >2 seconds

To allow for a direct comparison with [4], we separately analysed a subset of all silent pauses with a duration of over 2 seconds. Silent pauses of this kind were very rare in our corpus (73 tokens, or 2%, of the total 3473). The number of such pauses produced by each dyad ranged from 0 to 13.

The ASD group produced a higher mean rate of long silent pauses (>2 s) per minute (0.33; n = 34) than the CTR group (0.21; n = 39); see Fig. 2. Given the low overall number of instances, a more intuitive way of stating the same observation is that a 20-minute dialogue (average duration) would

typically contain 7 long pauses (>2 s) in the ASD group and 4 long pauses (>2 s) in the CTR group.

Bayesian Poisson regression suggests that this was a robust difference between groups (mean δ = -0.12; 95% CI [-0.23, -0.01]; P (δ > 0) = 0.97). However, the proximity of the higher end of the credible interval to zero and the very low overall number of observations are reasons for exercising some caution in the interpretation of these data.

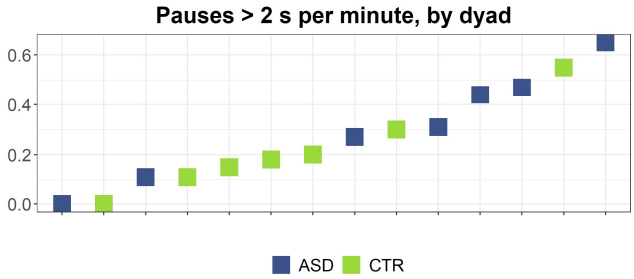


Figure 2: Rate of silent pauses >2 s in duration, by dyad and group (ASD in blue, CTR in green).

3.3. Silent pauses ≥700 milliseconds

For a more reliable and representative metric of "long pauses" in dialogue, we proceeded to use a lower cut-off value, at 700 milliseconds. We chose this particular threshold mainly because it clearly exceeds mean pause durations in the data set used here (646 ms across groups) as well as in previous work [2, 14, 20]. Additionally, the same value has been used for categorising long silent gaps between speakers [10], based on the findings that gaps of 700 ms or longer are perceived as unusual by listeners and often cue repair initiations or non-affiliating responses [21–23]. As the difference between within-speaker pauses and between-speaker gaps structurally lies only in who takes the following turn, these findings further support the use of a 700-millisecond threshold for silent pauses.

Using this cut-off point leaves far more observations for analysis (n = 1052) and will therefore also yield more robust and reliable results.

The group rate of silent pauses ≥700 ms was higher for the ASD group (4.02) than for the CTR group (3.52); see Fig. 3. Although this group difference is not very large, Bayesian negative binomial regression shows the effect to be robust, confirming that the CTR group produced a lower rate of long silent pauses (≥700 ms) than the ASD group (mean δ = -0.5; 95% CI [-0.9, -0.1]; P (δ > 0) = 0.98).

Thus, we can conclude that autistic dyads produced more long silent pauses than non-autistic dyads, independently of the exact cut-off point used to define a "long" pause, although no difference between groups was found when tokens of any duration were taken into account.

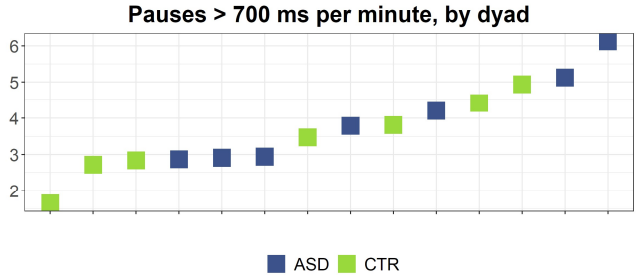


Figure 3: Rate of silent pauses ≥700 ms in duration, by dyad and group (ASD in blue, CTR in green).

3.4. Silent pauses following *uh* vs. *uhm*

The effect of *uh* and *uhm* on subsequent silent pause duration was equivalent for the ASD and the CTR group overall, as *uhm* was followed by longer silences in both groups and for all analyses. We therefore report results across groups below.

When disregarding filled pause duration (as in previous studies), we found a clear difference in the mean duration of following silence according to filled pause type: silences were on average 355 ms longer following *uhm* (mean = 541; SD = 1056) than following *uh* (mean = 186; SD = 517). We further calculated the proportion of filled pauses followed by a period of silence with a duration > 0 (i.e. not followed directly by speech). This was the case more frequently for *uhm* (69.4%) than for *uh* (45%).

As a sanity check, we next ran a Bayesian linear regression model with hurdle log-normal distribution to check whether filled pause duration, independent of filled pause type, could actually be shown to be correlated with the duration of the following silence at all. Model output unambiguously confirms this to be the case (mean δ = 0.29; 95% CI [0.16, 0.43]; P (δ > 0) = 1): longer filled pauses clearly tended to be followed by longer intervals of silence.

To conclusively establish whether differences between filled pause types were independent of the fact that *uhm* tokens in themselves were typically considerably longer than *uh* tokens, we fitted a model with log-normal distribution to the duration of the following silence, with speaker and, crucially, duration of filled pause, as random factors.

We found that silences following *uhm* were indeed longer than those following *uh*, regardless of the duration of filled pause tokens, even though the difference was quite small (150 ms). More details on statistical modelling are reported below.

In our main model, we only included observations where filled pauses were followed by at least 200 ms of silence (i.e. followed by a new, separate IPU). The difference between types is presented with *uhm* as the reference level. Model output shows the difference to be robust, even though the upper bound of the credible interval is close to zero (mean δ = -0.15; 95%

CI [-0.28, -0.02]; $P(\delta > 0) = 0.97$). A second model, including also all cases where the following silence was 0 (using a hurdle log-normal model), confirms the finding in also showing a robust effect for the difference between filled pause types (mean δ = -0.12; 95% CI [-0.19, -0.07]; $P(\delta > 0) = 1$) .

4. DISCUSSION

The analyses presented in this paper show that there were more long silent pauses in conversations between autistic dyads as compared to non-autistic control dyads. This is broadly in line with results from one of the three previously published studies on the same topic [4], but stands in contradiction to an earlier account [3]. We did not find group differences when considering all silent pauses regardless of duration, nor for mean pause duration (similarly to results in [5]). There was also no between-group difference regarding the effect of preceding filled pause type on subsequent silent pause duration. This stands in contrast to results in [7], where longer silences following *uhm* were found for non-autistic, but not autistic children.

While differences were thus rather subtle overall, the higher rate of long silent pauses in the ASD group is still likely to have a discernible effect on spoken interaction [2, 15], and might thus contribute to perceptions of a difference in communication styles. This is all the more true when considering that we found evidence for idiosyncratic behaviour by the same autistic speakers in the related domain of turn-timing, where they produced more long silent gaps *between* speakers compared to non-autistic dyads (but only in the early stages of conversations) [10].

Besides uncovering group differences in silent pause use, we were able to replicate the finding that silent pauses tend to be longer following *uhm* compared to *uh* [6]. Our study does not merely provide a replication, however, but rather extends and qualifies the original finding, as we added duration of filled pause as a random factor in a Bayesian regression model. This allowed us to explicitly show that the effect described is independent of intrinsic filled pause length (importantly, as *uhm* tends to be longer than *uh)*.

Moreover, our results suggest that the effect of longer silent pauses following *uhm* compared to *uh* is more subtle than previously described. While a two-fold difference in silence duration according to filled pause type is reported in [6], we found a difference of only 150 ms (with an average silent pause duration of 646 ms) when factoring in filled pause duration. It is not obvious how relevant such a difference might be in real-life spoken interactions. We will leave this question for future perception experiments.

There are a number of important limitations to the present study. These include the facts that we tested subjects from one extreme end of the autism spectrum (verbal, socially relatively skilled and motivated individuals with average or above-average IQ) and that conversations were task-based and limited to the spoken modality—although it could be argued that both of these limitations add to the specificity and interpretability of our findings. Additionally, the comparability between our own and previous studies is limited, obviously in the case of [3, 5], but also in the case of the more closely related work in [4]. In this study, we analysed semi-structured conversations between disposition-matched dyads of autistic speakers (ASD–ASD), whereas [4] analysed structured interactions between autistic adults and (presumably) non-autistic interviewers (ASD–CTR). While this is representative of most previous research on interactive communication in ASD, we submit that such research cannot yield any reliable insights into "autistic communication" per se. It instead only allows us to observe the particular patterns of behaviour arising in interactions between individuals with different cognitive styles [24–26].

5. CONCLUSION

We have shown evidence for a robust tendency towards a higher rate of long silent pauses in conversations between autistic compared to non-autistic dyads, but also found many similarities in the silent pause use of both groups. For instance, we replicated, extended and qualified the finding that longer pauses follow the filled pause type *uhm* compared to *uh,* across groups, showing further that the effect is subtle and independent of filled pause duration. Thus, besides extending important findings on the distribution of silent pauses in general, the current study specifically advances our understanding of silent pause use in ASD, which is especially important given that previous research has not provided any unambiguous indications in this regard. We hope that these insights might, in conjunction with related findings, ultimately serve as an important component in a multi-dimensional characterisation of conversational behaviour in ASD [8–10, 27].

6. ACKNOWLEDGEMENTS

Thanks to Martina Krüger and Alicia Janz for help with recording, annotation and processing. Funding was provided by the *Deutsche Forschungsgemeinschaft* (Project-ID: 281511265; SFB 1252) and the *Studienstiftung des deutschen Volkes* (Promotionsstipendium).

7. REFERENCES

[1] A. R. Bradlow, M. Kim, and M. Blasingame, 'Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate', *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 886–899, 2017.

[2] N. H. De Jong and H. R. Bosker, 'Choosing a threshold for silent pauses to measure second language fluency', in *The 6th workshop on disfluency in spontaneous speech (diss)*, 2013, pp. 17–20.

[3] C. Thurber and H. Tager-Flusberg, 'Pauses in the narratives produced by autistic, mentally retarded, and normal children as an index of cognitive demand', *Journal of Autism and Developmental disorders*, vol. 23, no. 2, pp. 309–322, 1993.

[4] J. K. Lake, K. R. Humphreys, and S. Cardy, 'Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders', *Psychonomic bulletin & review*, vol. 18, no. 1, pp. 135–140, 2011.

[5] P. E. Engelhardt, O. Alfridijanta, M. E. G. McMullon, and M. Corley, 'Speaker-Versus Listener-Oriented Disfluency: A Re-examination of Arguments and Assumptions from Autism Spectrum Disorder', *J Autism Dev Disord*, vol. 47, no. 9, pp. 2885–2898, Sep. 2017, doi: 10.1007/s10803-017-3215-0.

[6] H. H. Clark and J. E. Fox Tree, 'Using uh and um in spontaneous speaking', *Cognition*, vol. 84, no. 1, pp. 73–111, May 2002, doi: 10.1016/S0010-0277(02)00017-3.

[7] R. Lunsford, P. A. Heeman, L. Black, and J. van Santen, 'Autism and the use of fillers: differences between 'um'and "uh"', in *DiSS-LPSS Joint Workshop 2010*, 2010.

[8] S. Wehrle, 'A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder', PhD Thesis, University of Cologne, Cologne, Germany, 2021.

[9] S. Wehrle, M. Grice, and K. Vogeley, 'Filled Pauses Produced by Autistic Adults Differ in Prosodic Realisation, but not Rate or Lexical Type', *Journal of Autism and Developmental Disorders*, accepted.

[10] S. Wehrle, F. Cangemi, A. Janz, K. Vogeley, and M. Grice, 'Turn-timing in conversations between autistic adults: Typical short-gap transitions are preferred, but not achieved instantly', *PLOS ONE*, vol. 18, no. 4, p. e0284029, Apr. 2023, doi: 10.1371/journal.pone.0284029.

[11] K. L. Ashwood *et al.*, 'Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire', *Psychological medicine*, vol. 46, no. 12, pp. 2595–2604, 2016.

[12] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, 'The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians', *Journal of autism and developmental disorders*, vol. 31, no. 1, pp. 5–17, 2001.

[13] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer [Computer program]'. 2020. [Online]. Available: http://www.praat.org/

[14] H. Cho and D. Hirst, 'The contribution of silent pauses to the perception of prosodic boundaries in Korean read speech', *Proceedings of Speech Prosody 2006*, 2006.

[15] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press, 1968.

[16] M. Belz and J. Trouvain, 'Are 'Silent'Pauses Always Silent?', in *19. International Congress of Phonetic Sciences (ICPhS)*, Melbourne, 2019, pp. 2744–2748.

[17] P.-C. Bürkner, 'brms: An R package for Bayesian multilevel models using Stan', *Journal of statistical software*, vol. 80, no. 1, pp. 1–28, 2017.

[18] R Core Team, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[19] RStudio Team, 'RStudio: Integrated Development Environment for R'. RStudio, PBC, Boston, MA, 2021. [Online]. Available: http://www.rstudio.com/

[20] B. Megyesi and S. Gustafson-ˇCapková, 'Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish', in *Seventh International Conference on Spoken Language Processing*, 2002.

[21] K. H. Kendrick, 'The intersection of turn-taking and repair: the timing of other-initiations of repair in conversation', *Frontiers in psychology*, vol. 6, no. 250, pp. 10–3389, 2015.

[22] K. H. Kendrick and F. Torreira, 'The timing and construction of preference: A quantitative study', *Discourse Processes*, vol. 52, no. 4, pp. 255–289, 2015.

[23] F. Roberts and A. L. Francis, 'Identifying a temporal threshold of tolerance for silent gaps after requests', *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. EL471–EL477, 2013.

[24] D. Milton, 'On the ontological status of autism: the "double empathy problem"', *Disability & Society*, vol. 27, no. 6, pp. 883–887, 2012.

[25] G. L. Williams, 'Theory of autistic mind: A renewed relevance theoretic perspective on so-called autistic pragmatic "impairment"', *Journal of Pragmatics*, vol. 180, pp. 121–130, Jul. 2021, doi: 10.1016/j.pragma.2021.04.032.

[26] G. L. Williams, T. Wharton, and C. Jagoe, 'Mutual (Mis)understanding: Reframing Autistic Pragmatic "Impairments" Using Relevance Theory', *Front. Psychol.*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.616664.

[27] S. Wehrle, F. Cangemi, K. Vogeley, and M. Grice, 'New evidence for melodic speech in Autism Spectrum Disorder', *Proc. Speech Prosody 2022*, pp. 37–41, 2022.

POINTS OF MAXIMUM GRAMMATICAL CONTROL – THE PROSODY OF A TURN-HOLDING PRACTICE

{Anneliese Kelterer, Saskia Wepner}*, Julian Linke, Barbara Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology
anneliese.kelterer@edu.uni-graz.at, saskia.wepner@student.tugraz.at, {linke, b.schuppler}@tugraz.at

ABSTRACT

In conversation, speakers may pause in the middle of a turn construction unit to hold their turn, at a point of *maximum grammatical control*. We call such turn holds 'incomplete holds'. Based on 837 inter-pausal units from a corpus of Austrian German spontaneous conversations, we extract 11 prosodic features (F0 and RMS, one durational feature and one categorical feature) to analyse the prosody of incomplete holds. We do so in two steps: First, we use a Random Forest classifier to find the most important features that contribute to distinguishing incomplete holds from other kinds of pre-pausal turn-holds as well as from turn-changes. Second, we use regression analysis to examine how these features contribute to the different turn-taking categories. We find that incomplete holds are characterised by continuing intonation, high intensity, low articulation rate and a flatter F0 than grammatically complete holds.

Keywords: Turn-taking, prosody, conversational speech, Austrian German, incomplete hold

1. INTRODUCTION

Speakers often produce pauses at the end of turn construction units (TCU) [1]. However, speakers may also pause within a TCU to hold their turn, at a point of *maximum grammatical control* [2] (e.g., "I understand it, but ... no one else does"), called incomplete holds in this paper. This phenomenon has rarely been studied. One notable contribution is the *qualitative* investigation of turn-holding and turn-yielding practices in a North-Western variety of German by Selting [1]. We continue this line of research with a *quantitative* analysis of the prosody of incomplete holds in comparison to complete turn holds and turn changes in Austrian German.

Sacks [3] and Selting [1] argued that if speakers want to hold their turn at the end of a TCU, they have to employ specific turn holding strategies, such as rush-through or a specific prosody that is distinct from the prosody of a turn change. In incomplete holds, turn-holding is already signalled by syntactic projection. Therefore, we expect a prosodic realisation of incomplete holds that is distinct from both complete holds and turn changes. We draw the following hypotheses from the literature:

Final intonation. Ford & Thompson [4] reported a considerable number of syntactically complete turn holds that are also prosodically complete instead of projecting more talk by means of prosody. Incomplete holds, on the other hand, are expected to display prosody that does project more talk. Therefore, we expect more continuing intonation contours in incomplete holds, more terminal contours in change, and terminal as well as continuing contours in complete holds.

Final F0. Turn changes have been found to be associated with rising or falling pitch at the end [5, 6, 7] while turn holds displayed a flat pitch in the middle of a speaker's range [1, 6, 7, 8, 9]. A study of Swedish, on the other hand, found that while level pitch was a turn-holding cue, rising pitch was not clearly associated with either hold or change [10]. However, most of these studies did not distinguish syntactically complete from incomplete holds. Incomplete holds have been described to display final level or (slightly) rising pitch [1, 11]. Thus, we expect to find flat F0 in incomplete holds.

Intensity. We expect turn changes to have lower intensity than turn holds [9]. Within holds, we expect higher intensity in complete holds than in incomplete holds due to the need for specific prosodic turn-holding devices [1, 3], which might not be necessary when more talk is projected already by the incomplete syntax.

Articulation rate. Findings for lengthening are not consistent. Studies found more final lengthening in change than in hold [5], the reverse pattern [6, 7] or no consistent pattern at all [12]. For incomplete holds, Local & Kelly [11] found no noticeable slowing of tempo at the end. In the data analysed for this study, however, we had the auditory impression that speakers frequently do slow down in incomplete holds. Thus, we expect a lower articulation rate in incomplete hold, but we do not have clear expectations about the articulation rate in complete hold

*equal contribution

vs. change.

To address these hypotheses, we analyse the data two steps. First, we perform a Random Forest (RF) classification. The purpose of the RF is not classification per se, but to tell us *which* features are important for distinguishing the three turn-taking categories. Second, we perform regression analyses to find out *how* exactly the features are related to incomplete hold, complete hold and change.

2. MATERIALS

This study is based on 70 minutes of conversation (5 minutes in 14 conversations: 14f & 14m speakers) from the *Graz Corpus of Read and Spontaneous Speech* (GRASS) [13]. The corpus contains spontaneous face-to-face conversations between pairs of native Austrian German speakers. The data was annotated on three levels: turn-taking, prosodic phrases, and a phonetic segmentation via forced alignment, which was manually corrected [14].

For both prosodic phrasing and turn-taking, annotations were created in two stages: One annotator created a first version, which was then corrected by one of the other annotators. Prosodic phrases were labelled as *termination*, *continuation* or *high-rise* [15]. In Schuppler et al. [15], *continuation* was called *rise*, but other intonation contours indicating continuation were also annotated with this label.

Turn-taking was labelled in Inter-Pausal-Units (IPUs) separated by pauses longer than 150 ms [16]. IPUs were labelled as `change` before a turn change (questions are not included); as `complete hold` for a syntactically complete turn hold; and as `incomplete hold` if the same speaker continued talking after producing a pause at a point of *maximum grammatical control* [2]. The labelling process was based on sequential criteria in the tradition of Conversation Analysis (i.e., on interlocutors' behaviours and not on a speaker's intentions).

From a set of $N = 1011$ tokens, we excluded tokens containing laughter, uncertain labels and turn-changes with overlapping speech at the end of the IPU, resulting in $N = 837$ tokens (`in-hold`: 222, `com-hold`: 368, `change`: 247) for this study.

3. METHODS

3.1. Features

For each target IPU, we extracted 10 acoustic features (5 F0, 4 intensity, 1 duration) and the categorical feature *phrase-final intonation*. Since we are interested in turn-holding and turn-yielding practices, acoustic features were calculated for a window at the end of each IPU. When comparing the classification performance (cf., sec. 3.2) of three different window lengths (0.6, 0.8 and 1.0 seconds), a window of 0.6 s yielded the best performance (mean F_1-scores over all turn-taking categories: $F_{1,0.6\,\mathrm{s}} = 0.61$, $F_{1,0.8\,\mathrm{s}} = 0.60$, $F_{1,1.0\,\mathrm{s}} = 0.60$). If an IPU was shorter than 0.6 s, features were calculated for this shorter window.

F0 was tracked with the Python [17] package Parselmouth [18], and corrected for octave jumps with [19]. F0 was speaker-normalised by converting Hertz to semitones based on each speaker's overall median F0; then we calculated the maximum, minimum ($\mathrm{F0_{max}}$, $\mathrm{F0_{min}}$), median and range ($\mathrm{F0_{med}}$, $\mathrm{F0_{range}}$) and the $\mathrm{F0_{slope}} = (\Delta \mathrm{F0})/(\Delta t_{(\mathrm{F0_{min}},\,\mathrm{F0_{max}})})$.

We calculated the articulation rate (*ArtR*), and four intensity features ($\mathrm{I_{max}}$, $\mathrm{I_{med}}$, $\mathrm{I_{std}}$ and the position in time of $\mathrm{I_{max}}$ ($t(\mathrm{I_{max}})$)) as z-score speaker-normalised root mean square (RMS) values.

finIntonation indicates the annotated (functional) phrase-final intonation contour (levels: terminal, continuing, high-rising contour). It is important to note that these annotations are not the same as the final F0 slope; for instance, continuing intonation subsumes various final F0 movements (cf., comma intonation in [20]).

3.2. Random forest classifier

We used a random forest (RF) classifier from *scikit-learn* [21] (with 100 estimators, the square root as maximum number of features and Gini impurity) that we trained on the 11 features described in section 3.1. We trained the classifier for three different comparisons: `in-hold` vs `com-hold`, `in-hold` vs. `change` and `com-hold` vs. `change`. For each comparison, we split the data into 80 % training and 20 % test data and cross-validated with ten randomly chosen splits. Cross-validation is crucial since our data shows much variation both between and within speakers. With these pairwise comparisons, we obtained a better classification and thus a more reliable feature ranking than with a three-way classification.

3.3. Regression models

We built linear mixed effects regression models with R's [22] `lme4` package [23], where the dependent variable was one of the 10 continuous acoustic features. Models included the independent variables *Category* (`in-hold`, `com-hold`, `change`), and *Speaker* ($N = 18$) as a random variable. We then performed pairwise comparisons of the three values of the variable *Category* with `emmeans` [24]. For the categorical variable *finIntonation*, we built a logistic

in-hold vs. com-hold	in-hold vs. change	com-hold vs. change
$F_{1,\texttt{in-hold}}$=.60	$F_{1,\texttt{in-hold}}$=.78	$F_{1,\texttt{com-hold}}$=.68
$F_{1,\texttt{com-hold}}$=.79	$F_{1,\texttt{change}}$=.81	$F_{1,\texttt{change}}$=.40

Table 1: F_1 scores of the three RF classifications.

regression model with `multinom` from `nnet` [25], with *finIntonation* as dependent variable and *Category* as independent variable. We rotated the reference levels of *finIntonation* and *Category* to examine all comparisons. Here, we present significant predictors only. Features and model outputs can be found at[1].

4. RANDOM FOREST RESULTS

Table 1 presents F1 scores of the pairwise classifications. Two of the three pairwise comparisons yielded good results in the RF classification: `in-hold` vs. `com-hold` and `in-hold` vs. `change`. The classification of `com-hold` vs. `change` showed a bad performance. `in-holds` could be distinguished correctly from `changes` in 75.45% of cases (true positives), but from `com-holds` only in 56.36% of cases. `com-holds` were correctly distinguished from `in-holds` in 82.03% of cases, and from `changes` in 74.73% of cases. `changes` were classified correctly in 83.60% of the cases when compared to `in-holds` but only in 34.49% when compared to `com-holds`.

Table 2 shows the four highest ranked features for the pairwise comparisons with their corresponding importance values. We do not present further ranks because of their low importance values. *finIntonation* was the most important feature in distinguishing `in-hold` from the other two categories. Due to the low importance values of the other features and the overall poor classification performance of `com-hold` vs. `change`, the ranking of these features should be taken with caution.

5. REGRESSION RESULTS & DISCUSSION

5.1. Final intonation

Figure 1 presents the distribution of the annotated phrase-final intonation contours over the three turn-taking categories. The logistic regression showed that in `change`, terminal intonation was significantly more likely than continuing ($p < .0001$) and high-rising intonation ($p < .05$), compared to `com-hold`. In `change`, terminal intonation was significantly more likely than continuing ($p < .0001$) and high-rising intonation ($p < .01$), compared to `in-hold`. In `in-hold`, continuing intonation was more likely

in-hold vs. com-hold		in-hold vs. change		com-hold vs. change	
finInt	.123	finInt	.222	I_{med}	.021
ArtR	.012	*ArtR*	.024	*ArtR*	.015
I_{med}	.009	I_{med}	.008	I_{max}	.014
$t(I_{max})$	.007	I_{std}	.008	$F0_{slope}$	.014

Table 2: Highest ranked features with respective importance values of the three RF classifications – a higher value indicates a higher importance.

than high-rise ($p < .05$), compared to `change`. In `in-hold`, continuing intonation was more likely than high-rising ($p < .001$) and terminal intonation ($p < .0001$), compared to `com-hold`. Though the majority of `com-holds` had a terminal contour, the models do not indicate that this relationship is significant. The relatively high percentage of terminal intonation in `com-holds` (49% of all terminal contours were produced in `com-holds`) fits with the finding that only about half of all transition relevance places actually involve a turn-shift [4].

As expected, `in-holds` are clearly associated with continuing intonation and `changes` with terminal intonation, and *finIntonation* was the most important feature in the RF feature ranking (except for `com-hold` vs. `change`). High-rises occurred most frequently in `com-holds`, even though they were generally rare, but `com-holds` were not consistently associated with either terminal or continuing intonation, which also matches our expectations.

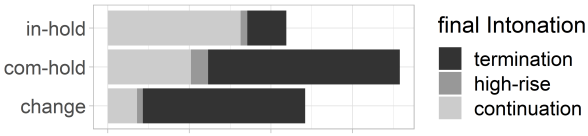


Figure 1: Distribution of phrase-final intonation labels over the three turn-taking categories.

5.2. Final F0

$F0_{max}$ was higher for `com-hold` than for `change` ($p < .05$) as well as `in-hold` ($p < .05$), but `in-hold` vs. `change` was not significant. $F0_{range}$ was higher for `com-hold` than `change` ($p < .05$) as well as `in-hold` ($p < .01$), but `in-hold` vs. `change` was not significant. $F0_{med}$ was higher for `com-hold` than for `change` ($p < .05$). For $F0_{min}$ and $F0_{slope}$, none of the comparisons showed a significant difference between the turn-taking categories. Figure 2 shows the distributions of $F0_{slope}$ values over the three categories. The two modes in the bimodal distribution indicate falling and rising contours. Even though distributions overlap considerably, some tendencies are visible: In `in-hold`, the peak at negative val-

ues is considerably higher than the one at positive values, indicating more falls than rises. The peak at negative values is close to zero, and the density around zero is clearly higher than in the other categories. This indicates that falls are not as steep and there are more flat F0 curves in `in-holds`. In `change`, the peak at negative values is much higher than the one at positive values, and the area under the curve at negative values extends more towards lower values, indicating that `changes` are more often falling than rising and that they display steeper falls than the other categories. In `com-hold`, the peak at positive values is only slightly higher than the one at negative values, but the area under the curve extends more towards higher values, indicating that, even though there are also falls in `com-holds`, there are overall more rises, and the rises are steeper than in other categories. Thus, we found more flat slopes as well as narrower F0 ranges in `in-holds`, which meets our expectations, more and steeper falls in `changes`, and more and steeper rises as well as higher F0 and a larger F0 range in `com-holds`. The higher F0 values in `com-holds` could be related to the higher percentage of high-rises in this category. However, even though some of the F0 features were significant, they were generally of low importance in the RF, except for the F0 slope in `change` vs. `com-hold`.

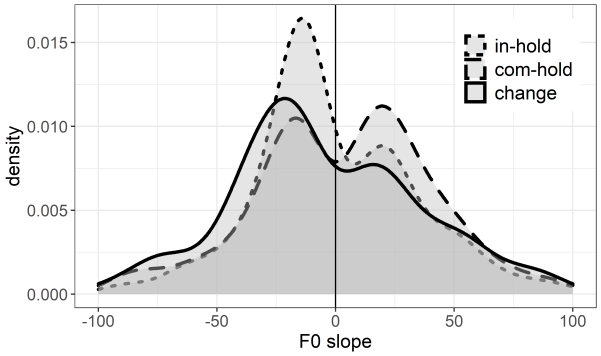


Figure 2: Density plot of $F0_{slope}$ for `in-hold` (dotted), `com-hold` (dashed) and `change` (solid). Positive values indicate a rise, negative values a fall, and values around zero indicate a flat F0.

5.3. Intensity

I_{med} was higher for `in-hold` than for `change` ($p < .001$) and `com-hold` ($p < .01$). I_{max} was marginally significantly higher for `in-hold` than for `change` ($p = .092$), but the other comparisons were not significant. I_{std} and the position of I_{max} were not significantly affected by the turn-taking categories. Thus, we did not find the expected higher intensity in `com-hold`. Instead, we found higher intensity in `in-hold`. This higher intensity at the end of `in-holds` might be due to higher sub-glottal pressure in the middle than at the end of a TCU, or it might be a strategy to contrast them to trail-offs [26].

5.4. Articulation rate

ArtR was highest for `change`, intermediary for `com-hold` and lowest for `in-hold` (`com-hold` vs. `change`: $p < .01$; `in-hold` vs. `change`: $p < .0001$; `com-hold` vs. `in-hold`: $p < .0001$). Thus, we found the fastest articulation rate in `change`, and, as expected, the slowest articulation rate in `in-holds`. This result does not confirm Local & Kelly's [11] observation of no noticeable slowing, but could be interpreted in the light of hesitations or disfluency. While turn-taking was not explicitly investigated by Betz et al. [27], they found disfluent lengthening most often in cases which could be characterised as points of *maximum grammatical control* (e.g., in articles, conjunctions and prepositions).

6. CONCLUSION

In this study, we analysed the prosody of syntactically incomplete turn-holds, that is, when a pause is produced at a point of *maximum grammatical control*, and compared them to syntactically complete turn-holds as well as to turn-changes. Our analyses showed that syntactically complete holds stand out through a higher maximum and median F0, which is probably due to a higher percentage of high-rises in this category. However, contrary to the expectation of more prosodic effort to distinguish syntactically complete holds from syntactically complete changes [1, 3], we did not find a higher intensity in complete holds, a specific articulation rate or a clear association with continuing intonation. Turn-changes, on the other hand, are characterised by terminal intonation and a higher articulation rate. The poor RF classification of syntactically complete turn-holds vs changes, suggests an insufficient prosodic distinction between these two categories. Overall, we found that the prosody at points of *maximum grammatical control*, that is, at the end of syntactically incomplete holds, is characterised by continuing intonation, high intensity, low articulation rate, and a flatter F0 than grammatically complete holds.

7. ACKNOWLEDGEMENTS

The work by A. Kelterer and S. Wepner was funded by grant P-32700-NB from FWF (Austrian Science Fund).

8. REFERENCES

[1] Selting, M. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* 6(3), 371–388.

[2] Schegloff, E. A. 1996. Turn organization: One intersection of grammar and interaction. In: Ochs, E., Schegloff, E. A., Thompson, S. (eds), *Interaction and grammar*. Cambridge: Cambridge University Press, 52–133.

[3] Sacks, H., Schegloff, E., Jefferson, G. 1974. A simplest systematics for the organisation of turn-taking for conversation. *Language* 50, 696–735.

[4] Ford, C., Thompson, S. 1996. *Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns*, 134–184. Cambridge University Press.

[5] Local, J., Kelly, J., Wells, W. 1986. Towards a phonology of conversation: turn-taking in Tyneside English. *J. Linguist.* 22(2), 411–437.

[6] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech* 41, 295–321.

[7] Gravano, A., Hirschberg, J. 2011. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang* 25(3), 601–634.

[8] Duncan, S. 1972. Some signals and rules for taking speaking turns in conversations. *J.Pers.Soc.Psychol.* 23(2), 283–292.

[9] Skantze, G. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67, 1–26.

[10] Edlund, J., Heldner, M. 2005. Exploring prosody in interaction control. *Phonetica* 62(2-4), 215–226.

[11] Local, J., Kelly, J. 1986. Projection and 'Silences': Notes on Phonetic and Conversational Structure. *Human Studies* 9(2/3), 185–204.

[12] Hjalmarsson, A. 2011. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication* 53(1), 23–35.

[13] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: The Graz corpus of Read And Spontaneous Speech. *Proceedings of LREC*, 1465–1470.

[14] Ludusan, B., Schuppler, B. 2022. An analysis of prosodic boundaries across speaking styles in two varieties of german. *Speech Communication*.

[15] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94, 62–74.

[16] Schuppler, B., Kelterer, A. 2021. Developing an Annotation System for Communicative Functions for a Cross-Layer ASR System. *Proceedings of Workshop "Integrating Perspectives on Discourse Annotation"*.

[17] Python Software Foundation, 2009. Python Language Reference. Version 3.9.

[18] Jadoul, Y., Thompson, B., de Boer, B. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71, 1–15.

[19] Paierl, M. 2021. Detektion und Visualisierung von F0-Berechnungsfehlern.

[20] Chafe, W. 1988. Linking intonation units in spoken english. In: Haiman, J., Thompson, S. A. (eds), *Clause combining in grammar and discourse*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 1–27.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

[22] R Core Team, 2020. R: A Language and Environment for Statistical Computing. Version 4.2.1.

[23] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[24] Lenth, R. 2020. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1.

[25] Venables, W. N., Ripley, B. D. 2002. *Modern Applied Statistics with S*. New York: Springer fourth edition. ISBN 0-387-95457-0.

[26] Walker, G. 2012. Coordination and interpretation of vocal and visible resources: 'Trail-off' conjunctions. *Language and Speech* 55(1), 141–163.

[27] Betz, S., Wagner, P., Voße, J. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. *Tagungsband der 12. Tagung Phonetik und Phonologie im Deutschsprachigen Raum*.

[1] Feature table and model outputs can be found at: https://cloud.tugraz.at/index.php/s/mAH4LC2NadQcXSi

TIMING OF LAUGHTER IN CONVERSATION: BETTER LATE THAN NEVER?

Tamara Rathcke[1], Eleni Kapogianni[2], Lucy Page[2]

[1]University of Konstanz, Germany; [2]University of Kent, UK
Tamara.Rathcke@uni-konstanz.de, E.Kapogianni@kent.ac.uk, L.Page@kent.ac.uk

ABSTRACT

Studies in Conversation Analysis show that timing is a cue for preference organisation of pragmatic meanings. Responses that occur outside of an expected temporal window tend to be associated with dispreferred options and/or negatively charged meanings. The present perception study applied the predictions of conversation analysis to laughed (as opposed to spoken) responses, hypothesising that timing of genuine laughter obeys similar time constraints. We investigated perceptions of voiced and unvoiced laughs occurring at three time points and in three discourse contexts. The results show that delayed laughs were rated as significantly less genuine and more sarcastic than laughs that occurred within the expected time window or earlier. The timing effect was moderated by laughter type and discourse context. The results confirm that timing of a response contributes to conversational meanings, though the effect is not as general as previously suggested.

Keywords: laugh, laughter types, conversation, timing, humour, sarcasm

1. INTRODUCTION

Laughter permeates communicative interaction and is ubiquitous in our daily lives, e.g., [1], [2], [3], [4]. Being a form of non-verbal vocal behaviour, laughter occurs not only in humorous contexts but is generally known to promote positive and cooperative relationships between interlocutors during a conversation [5]. In fact, the prevalent function of laughter in a conversation seems to be purely interactional since most laughs tend to occur in the absence of humour [1], [2]. The full range of meanings that laughter can help to express during conversations is still rather poorly captured. In particular, the role of different laughter types and their timing in conversation have not been fully addressed in previous research.

Communicative exchanges take place rapidly, with interlocutors' turns concatenating with each other smoothly [6], [7]. Temporal gaps between turns observed in natural conversations average around 200 ms, with a vast majority transitions happening within a window of 100 to 500 ms [6]. Very few turns tend to overlap or occur outside of the specified time window derived from large speech corpora [6]. It has been suggested that spoken interactions obey certain temporal organisation principles, such as minimal overlap or delay of interlocutors' contributions to the ongoing discourse [7]. There are also proposals suggesting that conversational turn-taking follows a regular 'beat' or an internal 'clock' allowing participants to orchestrate their verbal contributions in time [8], [9].

Moreover, the time window of a conversational response may indicate the meaning of that response. Conversation analysis of large databases has established that preferred responses (such as acceptances of invitations, inter-speaker agreements and encouragements) tend to occur early and smoothly, on completion of a turn or in a slight overlap with it [10], [11]. In contrast, dispreferred responses (such as rejections of invitations or disagreements) often occur after a noticeable delay [10], [11]. Accordingly, the timing of a response can be meaningful by itself. A response delivered after 700 ms or more is very likely to break bad news to the speaker waiting for their interlocutor's response. Early responses overlapping in time with the previous turn also seem to deliver dispreferred information, though the effect is not as consistent as the effect of a delayed response [11].

Little research exists on the timing of laughter in turn-taking and its conversational meaning. The interactional concept of 'invitation-acceptance' has been applied to laughs produced in a small corpus of spoken dialogues in German [12]. Accordingly, turn transitions involving laughter could be classified as either invited or self-initiated. The prevalent pattern of the dialogues involves the speaker's invitation to a joint laugh which is then followed by their interlocutor while self-initiated laughter of the interlocutor was observed slightly less often. Speed dating dialogues seem to follow a similar pattern whereby the incoming interlocutor tends to join an ongoing laugh some 300 ms after its start [13]. These findings suggest that temporal dynamics are at play when laughs are produced in conversation. Given that laughter prevails at the end of turns and utterances [14], it constitutes an ecologically relevant turn-taking phenomenon to study.

The present paper reports a perception study that was created to examine the question whether or not laughter, like speech, is expected to occur within a certain time window after a completed turn in order to be interpreted as genuine. Based on the existing works of Conversation Analysis [10], [11], we hypothesised that laughter deviating from a baseline time window would be perceived as disaffiliative, i.e., expressing distance and/or disapproval towards the previous utterance [15]. The perception can be expected to map onto laughter evaluations as insincere ("fake") and critical/negatively disposed towards the addressee's previous utterance ("sarcastic"). The testing of this hypothesis also takes into consideration different laughter types and laughter-eliciting contexts.

2. METHODS

2.1. Speech materials

Six conversational topics of daily relevance were chosen. Test items followed a 4-turn structure that varied according to the type of context in the second turn (i.e., the turn immediately preceding laughter). Three contexts were included in the present study: humorous, sarcastic, and literal. Table 1 gives an example of a topic and the pragmatic manipulation. Humour appeared as jokes or wordplays. Sarcasm indicated a non-ironic, insulting or mocking contribution. Literal evaluation was a genuine response without intended humour and was used as a control condition. This resulted in a total of 18 scripted conversations.

The materials were recorded by two female speakers of Standard Southern British English, with distinct voices that could be easily perceived as interlocutors A and B. The speakers were friends in real life, making the recordings of the conversations and laughs more authentic. The speakers were recorded in the linguistics laboratory of the University of Kent.

2.2. Laughter types

Different approaches to the categorisation of laughter tend to agree on at least one phonetic feature distinguishing between laughter types, namely the presence of voicing [16], [17], [18]. Harmonically rich, vocalised laughs that may contain vowel-like portions are sometimes considered stereotypical [16]. However, voiceless laughter that consists of bouts of nasal or oral air turbulence is a more frequently observed laughter behaviour [16]. Moreover, research suggests that the presence or absence of vocal fold vibration during a laugh may result in different perceptual evaluations of its function and authenticity: Voiceless laughter tends to receive less positive ratings and may be perceived as less genuine but rather fake or sarcastic [16], [17], [19], [20].

The present study used a voiceless and a voiced laugh produced by one of the study speakers (A, Table 1). The spontaneously produced laughs differed slightly in duration. The duration difference was normalised using Praat, by slightly shortening both laughs to 570 ms. The intensity of the two laughs was also normalised by setting the maximal amplitude of each sound to a pre-defined value.

Speaker	Discourse context		
	Humorous	Sarcastic	Literal
A	*I've been running 10 miles every day preparing for the marathon*		
B	*It'll be worth it …in the long run.*	*Can you even run a mile?*	*Sounds like hard work*
A	LAUGH		
B	*Good luck anyway*		

Table 1: Example of a conversation topic manipulating the discourse context of laughter occurrence.

2.3. Timing manipulations

The timing of laughter occurrence was manipulated to create three versions: early, late and baseline. The baseline was set to 300 ms, following previous research on the timing of turn-transitions in natural conversation [6], [11], [21]. The timing of a late response followed the threshold of 700 ms that has been previously shown to increase the amount of dispreferred or disaffiliative evaluations in conversations [11]. The early timing was set to 300 ms of overlap time between the previous turn and the incoming laughter.

2.4. Procedures

The conversation recordings were paired with the two laughs under three timing conditions, resulting in a total of 108 experimental stimuli (18 conversations x 2 laughs x 3 timings). The stimuli were divided into two sets, containing either voiced or voiceless laughs. The participants listened and responded to one set of the stimuli only, rating a total of 54 conversations. The ratings were collected on a 7-point Likert scale, with the participants responding to the question *'What is your impression of speaker A's laughter?'*. Only

the endpoints of the scale were labelled, 1 as 'genuine and friendly' and 7 as 'fake and sarcastic'. Since meta-pragmatic intuitions on sarcasm are expected to vary [22], [23], the scale combining elements of "speaker sincerity" (genuine-fake) and "speaker disposition" (friendly-sarcastic) was deemed representative of the distinction between affiliative and disaffiliative functions of laughter that the present study is concerned with.

The responses were collected online, using the Qualtrics software. All stimuli were presented in a randomised order. Listener responses were not timed but experimental instructions encouraged the participants to respond spontaneously and not to overthink their ratings. A short practice was given prior to the experiment. At the end of the experiment, participants were presented with 14 different laughs by the same female speaker (including the two experiment laughs) and asked to rate these laughs on a scale from 1 (controlled, not spontaneous at all) to 7 (extremely spontaneous, not controlled at all). The additional ratings established the baseline evaluation of the two experiment laughs outside of the selected conversational contexts.

2.5. Listeners

Sixty-seven listeners aged between 18 and 73 (mean age: 40), without any known speech or hearing disorders volunteered to participate in the study. Responses from five listeners were excluded from the analyses as the listeners were non-native speakers of English. This left a total of 62 complete datasets for the remaining analyses.

The study was approved by the Research Ethics Advisory Board of the University of Kent in June 2019. All participants gave an informed consent prior to the recordings and the perception experiment taking place.

3. RESULTS

Given the ordinal scaling of the dependent variable, we fitted Poisson mixed models to these data (estimated using ML and BOBYQA optimizers), to predict the rating of a laugh by laughter type (voiced, voiceless). For conversations, the predictors of interest further included laughter timing (baseline, early, late) and discourse context (literal, humorous, sarcastic) and their interactions with each other and with laughter type. Only results of the best-fit models are reported below.

The best-fit model of ratings obtained for the conversation stimuli included two significant interactions, between laughter type and timing (χ^2 = 7.60, $p < 0.05$; plotted in Figure 1) and between timing and discourse context (χ^2 = 13.36, $p < 0.01$, plotted in Figure 2).

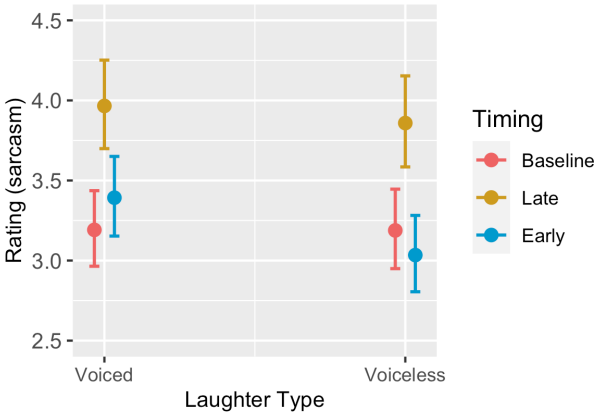


Figure 1: Estimated effects for the interaction between laughter type and timing.

Overall, there was no effect of laughter type on the perceptual ratings. Laughter type only mattered if its timing was early, i.e., overlapping with the preceding speaker turn. In these (but not any other) cases, the voiced laugh was perceived as significantly more sarcastic than the voiceless laugh ($z = 3.42$, $p < 0.001$). In addition, both types of laughter were rated as more sarcastic than the baseline (voiced: $z = 5.02$, $p < 0.001$, voiceless: $z = 3.80$, $p < 0.001$) if they were coming in as a late response. In contrast, both early-timed laughs were perceived similar to the baseline.

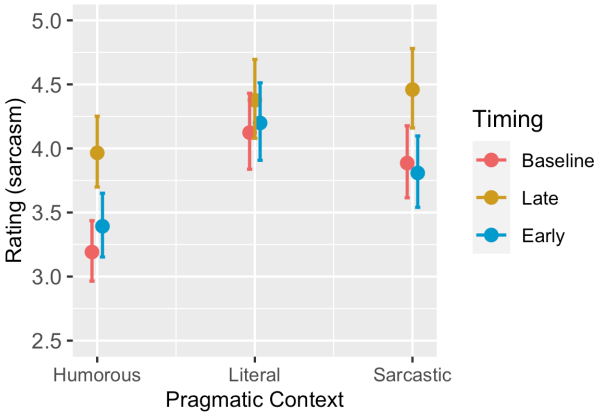


Figure 2: Estimated effects for the interaction between laughter timing and discourse context of conversation.

The interaction of laughter timing and discourse context indicates that timing does not matter for the perception of laughter in literal discourse contexts (see Figure 2). Instead, late timing leads to a sarcastic interpretation of the laugh in those contexts that carry the meaning of humour (late/baseline comparison: $z = 5.56$, $p < 0.001$; late/early comparison: $z = 5.42$, $p < 0.001$) or sarcasm (late/baseline comparison: $z = 3.66$, $p < 0.001$; late/early comparison: $z = 5.75$, $p <$

0.001). Early-timed laughs are perceived similar to baseline in both humorous and sarcastic contexts while late-timed laughs are perceived as more sarcastic in sarcastic than in humorous contexts ($z = 3.46$, $p = 0.0015$).

Finally, we examined the perception of isolated laughter as genuine and spontaneous, comparing only the two experimental stimuli. The two laughs indeed differed in their perceived spontaneity, with voiceless laughter being rated as slightly but significantly less genuine and spontaneous than voiced laughter ($z = 2.28$, $p < 0.05$).

4. DISCUSSION

The aim of the present study was to investigate the perception of laughter when its timing in the conversation is altered. Following existing findings of conversation analysis [10], [11], our main hypothesis proposed that laughter would be expected to follow the typical conversational transition of 300 ms, with deviations from this temporal window eliciting dispreferred or disaffiliative interpretations. We tested this idea by investigating perceptual responses to laughter in three types of mini-conversations (with literal, humorous, and sarcastic pragmatics) in combination with three conversational timings (baseline, early, late) and two laughter types (voiced, voiceless).

Overall, the results of the perception study with 62 listeners of different ages indicate that timing of laughter plays a major role in creating meaning in conversation, with delayed laughs being more likely interpreted as disaffiliative (sarcastic) [7], [24]. The latter finding supports the hypothesis based on previous research, though only partially [10], [11]. When it comes to a spoken response, any deviance from the baseline time window of 100-500 ms at the end of a turn tends to evoke a dispreferred interpretation. Yet when it comes to a laughed response, early timing does not seem to lead to a dispreferred interpretation. Perceptions of early laughs that show a substantial temporal overlap with the ongoing turn did not differ from the baseline; only delayed laughs did.

This finding may be related to general observation that laughter, unlike speech, often overlaps with turns of other interlocutors [13], [25]. Consequently, the conversational timing hypothesis ought to take the type of verbal responses into account. While late timings may tend to evoke dispreferred interpretations across a range of verbal responses, pragmatic interpretations of early timings depend rather heavily on the response being spoken or laughed.

Moreover, the perception of laughter timing in conversation varied with laughter type. Previous studies comparing voiced and voiceless laughter [16], [19] identified that voiced laughs tended to be perceived as more genuine and spontaneous than voiceless laughs. The present study confirmed this tendency for the two types of laughs used here, though only when these were judged in isolation. Embedded in a conversation and overlapping with the preceding speaker turn, it was specifically the voiced laugh that was perceived as more fake and sarcastic and not the voiceless laugh. The ratings from the two parts of the experiment are somewhat at odds with each other and highlight the need to study laughter under more naturalistic conditions (e.g., in pragmatically meaningful contexts).

As previously suggested, voiced laughter (due to the prevalence of high amplitudes, high mean F0 and strong F0 modulations during its production) has an arousing effect on listeners [16], [26]. Such high arousal of a voiced laugh may evoke the perception of sarcasm, given that an early-timed laugh of the present study does not follow the prevalent pattern of "invitation-acceptance" found in natural dialogues with laughed turn-transitions [25]. Even though voiced laughter is often considered a prototypical laugh, it is rather infrequent in comparison to unvoiced laughter [26]. Being unsolicited by context and otherwise infrequent, a voiced laugh may attract perceptual salience [27] that then, in combination with high arousal, is attributed to speaker sarcasm. This interpretation of the present finding requires further investigation.

Finally, the discourse context also affects the interpretation of laughter timing. The most natural occurrence of laughter is immediately following a joke, i.e. in a non-sarcastic wordplay [28]. It is the intended (and preferred) response after a joke's punchline [29]. Our data confirms that the humorous context was indeed conducive to the perception of the laughter response as less sarcastic and more genuine. On the other hand, sarcasm is typically a disaffiliative act (despite the fact that it may not necessarily have a negative interpersonal effect, [30]) and is therefore more likely to be responded to by disaffiliative laughter [24], [31], a pattern which is also present in our data.

Taken together, the results of the present study indicate that turn timing actively contributes to the creation of pragmatic meaning in conversation [7], [22-24]. While the pragmatics of early timing may depend on the type of response, the pragmatics of late timing appears to generally indicate dispreferred, disaffiliative meanings across a range of verbal and vocal responses.

REFERENCES

[1] Jefferson, G. (2004). A note on laughter in 'male–female' Interaction. *Discourse Studies,* 6(1), 117–133.

[2] Vettin, J., Todt, D. (2004). Laughter in Conversation; Features of Occurrence and Acoustic Structure. *Journal of Nonverbal Behaviour*, 28, 93–115.

[3] Tanaka, H., Campbell, N. (2014). Classification of social laughter in natural conversational speech. *Computer, Speech and Language*, 28, 314–325.

[4] Truong, K. P., Trouvain, J. (2014). Investigating prosodic relations between initiating and responding laughs. *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, 1811–1815.

[5] Bryant G. A. *et al.* (2016). Detecting affiliation in co-laughter across 24 societies. *Proceedings of the National Academy of Sciences*, 113(17), 4682–4687.

[6] Levinson, S. C., Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6.

[7] Sacks, H. (1974). An analysis of the course of a joke's telling in conversation. In: R. Baumann, J. Sherzer (eds.) *Explorations in the Ethnography of Speaking*, Cambridge: CUP.

[8] Wilson, T. P., Zimmerman, D. H. (1986). The structure of silence between turns in two-party conversation. *Discourse Processes*, 9, 375–390.

[9] Wilson, M., Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12(6), 957–968.

[10] Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In: J Maxwell Atkinson, J. Herit (eds.) *Structure of social action*. Cambridge: CUP.

[11] Kendrick, K. H., Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, 52, 255–289.

[12] Trouvain, J., Truong, K. P. (2013). Exploring sequences of speech and laughter activity using visualisations of conversations. *Workshop on Affective Social Speech Signals.*

[13] Rathcke, T., Fuchs, S. (2022). Laugh is in the air: An exploratory analysis of laughter during speed dating. *Frontiers in Communication*, 7.

[14] Provine, R. R., Emmorey, K. (2006). Laughter among deaf signers. *Journal of Deaf Studies and Deaf Education*, 11(4), 403–409.

[15] Clift, R. (2016). Don't make me laugh: Responsive laughter in (dis)affiliation. *Journal of Pragmatics,* 100, 73-88.

[16] Bachorowski, J.-A., Owren, M., J. (2001). Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, 12(3), 252–257.

[17] Kohler, K. J. (2008). 'Speech-smile', 'speech-laugh', 'laughter' and their sequencing in dialogic interaction. *Phonetica*, 65(1–2), 1–18.

[18] Trouvain, J. (2003). Segmenting phonetic units in laughter. *Proceedings of the 15th International Congress of Phonetic Sciences*.

[19] Grammer, K., Eibl-Eibesfeldt, I. (1990). The ritualisation of laughter. *Natürlichkeit der Sprache und der Kultur*, *18*, 192–214.

[20] Ruch, W., Ekman, P. (2001). The expressive pattern of laughter. In: A. Kaszniak (ed.) *Emotions, Qualia, and Consciousness*, 10, World Scientific.

[21] Levinson, S. C. (2013). Recursion in Pragmatics. *Language*, 89(1), 149–162,

[22] Dynel, M. (2017). Academics vs. American scriptwriters vs. academics: A battle over the etic and emic 'sarcasm' and 'irony' labels. *Language Communication*, 55, 69–87.

[23] Kapogianni, E. (2021). Types and Definitions of Irony. In: P. Stalmaszczyk (ed.) *The Cambridge Handbook of the Philosophy of Language*, Cambridge: CUP, pp. 622–638.

[24] Truong, K. P., Trouvain, J. (2012). On the acoustics of overlapping laughter in conversational speech. *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, 851–854.

[26] Bachorowski, J.-A., Smoski, M. J., Owren, M. J. (2001). The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110(3), 1581–1597.

[27] Boswijk, V., Coler, M. (2020). What is Salience? *Open Linguistics*, 6,(1), 713–722.

[28] Gervais, M., Wilson, D. S. (2005). The evolution and functions of laughter and humor: a synthetic approach. *Quarterly Review in Biology*, 80(4), 395–430.

[29] Norrick, N. R. (2003). Issues in conversational joking. *Journal of Pragmatics*, 35, 1333–1359.

[30] Partington, A. (2006) *The Linguistics of Laughter*. Routledge.

[31] Hanks, E., Egbert, J. (2022). The Interplay of Laughter and Communicative Purpose in Conversational Discourse: A Corpus-Based Study of British English. *Corpus Pragmatics,* 6(4), 261–290.

VOICE QUALITY DYNAMICS OF TURN-TAKING EVENTS IN SWEDISH AND GERMAN

Marcin Włodarczak*, Mattias Heldner*, Anna Bruggeman†, Petra Wagner†

*Stockholm University, †Bielefeld Univeristy
{wlodarczak,heldner}@ling.su.se, {anna.bruggeman,petra.wagner}@uni-bielefeld.de

ABSTRACT

The study investigates variation of two voice quality features (smoothed cepstral peak prominence and the relative strength of the fundamental) preceding speaker changes and turn holds in spontaneous conversations in Swedish and German. We demonstrate that the overall pattern is to a large extent linked to the syllabic structure in speech. While speech before turn holds is characterised by more periodic phonation than before speaker changes, the effect sizes are very small and are unlikely to be of practical importance for speech communication. We observe no evidence in favour of language specificity of the reported contrasts.

Keywords: phonation type, turn-taking, neck-surface acceleration, Generalized Additive Mixed Models

1. INTRODUCTION

The speech signal is rich in linguistic and paralinguistic information [1], including prosodic patterns related to coordination of turn-taking in conversation [2, 3]. However, studies of prosodic turn-taking cues rarely include properties related to voice quality (i.e. phonation type), in spite of a growing body of evidence for the importance of voice control to prosodic expression [4, 5, 6]. This state of affairs can to a large extent be explained by the difficulty of studying the voice source in spontaneous speech, where inverse filtering of supraglottal resonances is either prohibitively time consuming or error-prone [7].

Existing work has thus relied primarily on manual annotations of specific voice quality types (e.g. creak [8]) or acoustic analysis of the raw audio signal [9], often using features which are difficult to interpret when calculated from connected speech, such as jitter or shimmer [3, 10]. The methodological differences notwithstanding, this line of research has found that speech preceding speaker changes is less periodic and more creaky than in turn-holds.

More recently, [11] revisited the question of phonatory turn-taking cues using a signal collected with miniature accelerometers attached to speakers' tracheal wall below the larynx. The method is simple to use, noninvasive and can be used to estimate voice source parameters [12, 13, 14]. It has also showed promise for classification of voice quality [15], although it might offer little advantage over the microphone signal in highly controlled speech [16].

[11] used the method to analyze several voice quality parameters (harmonicity, strength of the fundamental and spectral balance) extracted from the last voiced interval preceding turn-taking events. They found that, in line with previous research, speaker changes accompanied by silence were characterised by less modal phonation than turn holds. In addition, they demonstrated that while some voice quality features did help to distinguish between the categories, their contributions were limited when compared to that of intensity level and, to a lesser extent, fundamental frequency (f_0).

An obvious limitation of the analysis in [11] is that it was based on averages calculated over the last voiced interval in a talkspurt. Thus, it failed to account for variability of the parameters in question over time. In this paper, we remedy this shortcoming by analysing the dynamical properties of voice quality using Generalized Additive Mixed Models (GAMM). By including two languages (German and Swedish), we address the question of language-specificity of voice quality variation.

2. METHOD

2.1. Material

The material consisted of seven dyadic conversations in Swedish and four in German. The participants (8 males and 14 females) were instructed to talk with each other on a topic of their choice for about 30 minutes (mean durations: 28 and 30 minutes for the German and Swedish conversations, respectively). All participants knew

each other before the recording. The recordings were made in sound treated rooms at Stockholm University and Bielefeld University. In addition to audio recordings using regular microphones, participants' vocal activity was also recorded with miniature accelerometers (Knowles BU-27135) attached to the neck below the level of the glottis.

2.2. Feature extraction

For Swedish, talkspurts were identified from the audio recordings using the voice activity segmentation method proposed in [17]. For German, due to strong crosstalk between the channels, talkspurts were labelled manually by a student assistant. Each silence was then automatically classified as either within- or between-speaker (henceforth, WSS and BSS), depending on whether it was associated with a speaker change (BSS) or a turn-hold (WSS). Given that very short utterances correspond largely to backchannels [18], WSS and BSS intervals in which the talkspurt preceding the silence was shorter than one second were excluded from the analysis.

The following voice quality features were extracted from voiced frames (50 ms in length with 2 ms step, Gaussian window) in the accelerometer signal:

Smoothed cepstral peak prominence (CPPS): amplitude of the first rahmonic relative to the regression line over the power cepstrum, in dB [19] characterising the periodicity of the signal.

***L_1L_2*:** characterising the strength of the fundamental relative to the second harmonic and/or spectral slope in the low frequency part of the spectrum: $L_1 - L_2$, in dB.

Given that voice source parameters vary significantly as a function of subglottal pressure and sound pressure level [20], intensity level (in dB) was also computed. Feature extraction was performed with Praat [21], identical to that in [11], except for the fact the same (default) voicing threshold was applied to f_0-dependent (L_1L_2) and and f_0-independent (intensity level, CPPS) features.

Voiced intervals of at least 20 ms were identified based on the output of Praat's pitch tracking. Next, the last voiced interval in the final second of each talkspurt was found, the CPPS, L_1L_2 and intensity level trajectories were interpolated with cubic splines (without smoothing) and sampled at 15 evenly spaced points. Frames that included pitch-halving errors were removed. If this procedure resulted in a loss of more than 10% of frames in a voiced interval, the entire interval was excluded.

Only intervals longer than 78 ms (corresponding to the minimum of 15 50-ms frames with a 2-ms step) in which all the feature values were within three standard deviations of speaker's mean were included in the final analysis. Additionally, data from one Swedish speaker were excluded due to evidence of strong pitch halving. The final data set consisted of: 234 BSS and 649 WSS intervals in Swedish, and 384 BSS and 618 WSS intervals in German.

2.3. Modeling

The data were modeled using Generalized Additive Mixed Models (GAMM, [22]) using the *mgcv* package in R [23]. Separate models were fitted for each response variable (intensity level, CPPS, L_1L_2). The following model structure was used, using the `mgcv::formula.gam` syntax (`s`: smooth term, `te`: tensor product smooth, `bs='fs'`: factor smooth):

```
Y ~ s(time) +
  s(time, by=is.swe) +
  s(time, by=is.wss) +
  s(time, by=is.swe.wss) +
  te(time, voiced.dur) +
  s(time, speaker, bs='fs', m=1) +
  s(time, speaker, by=is.wss.ord,
    bs='fs', m=1)
```

The models included effects for language (`is.swe`) and interval type (`is.wss`) as well as for the difference between the BSS/WSS contrast across the two languages (`is.swe.wss`). Since we have no specific hypotheses about contour shapes as opposed to the overall level, we use binary difference smooths [24]. They combine both of these differences and do not result in inflated type-I error rates associated with separate tests for the parametric and the non-linear components [25]. In addition, the models included a tensor product interaction between time and the duration of the voiced interval (`voiced.dur`, in $log_2(\text{s})$), as well as random smooths for speakers (`speaker`), grouped by interval type (`is.wss.ord`, modelled using a reference and random difference smooths, following the recommendation in [25]).

All the models were fitted with the `mgcv::bam` function, using fast restricted maximum likelihood estimation (fREML) and discretization of covariate values (i.e. with the `discrete` parameter set to true). A scaled t model (`family = 'scat'`) was used to deal with non-linearity of the residuals. Autocorrelation in the data was controlled for with a

first-order auto-regressive model (AR1), with ρ set to autocorrelation at lag of one in a model without the AR1 component.

Secondary data and the scripts used in the analysis are available at: https://doi.org/10.5281/zenodo.7870015.

3. RESULTS

The predicted curves for intensity level, CPPS and L_1L_2 are shown in Figure 1.

The observed intensity effects reflect the expected patterns associated with syllabic organisation in speech: intensity level increases at the beginning of the voiced interval and falls again towards its end. Indeed, median duration of the voiced interval was equal to 0.2 seconds, which corresponds roughly to average syllable duration in speech. The end point is somewhat lower than the start point, most likely corresponding to energy declination across prosodic phrases. The difference between German and Swedish was significant for the reference BSS category ($p < 0.001$). Visual inspection of Figure 1 as well as of the difference smooth indicates that the curve for Swedish is somewhat lower (by about 2 dB) and shows a more gradual decrease of intensity throughout its duration, as opposed to the German curve, which decreases more sharply towards its end. Neither the BSS/WSS contrast in German ($p = 0.09$) nor the difference in marking of the contrast across the languages ($p = 0.6$) were significant.

CPPS mirrors the inverted-U pattern present in the intensity curves, reflecting the fact that centers of syllabic nuclei are characterised by more periodic phonation than the peripheries. We found no evidence for a difference in the BSS category across the languages ($p = 0.7$). The BSS and WSS intervals differed significantly for German ($p < 0.001$): within-speaker silences were characterised by higher CPPS values in the middle portion of the voiced interval. However, the difference curves reveal that this difference is very small (less than 1 dB). There was no evidence that the BSS/WSS contrast is realised differently in Swedish ($p = 0.4$).

L_1L_2 shows the opposite pattern to intensity level and CPPS with low values in the middle of the voiced interval and high values towards the edges. This is again expected given that a strong fundamental (and, consequently, high L_1L_2) is indicative of more breathy phonation. The language effect for the BSS category was significant ($p < 0.001$): the curve for Swedish is somewhat higher and is characterised by a less steep increase towards the end. The BSS/WSS contrast in German was also significant ($p = 0.02$), with lower values for WSS intervals but, similar to CPPS, the difference was smaller than 1 dB on average. We found no evidence for language-specific difference with respect to the BSS/WSS effect ($p = 0.7$).

4. DISCUSSION

This paper is part of project where we are exploring voice quality as a dimension of prosodic expression, following the voice prosody hypothesis [4, 5, 6]. Our approach has been to examine short-term variations in acoustic voice quality features as correlates of various prosodic functions (e.g. turn-yielding, turn-holding, and word and utterance level prominences). To avoid the difficulties in estimating voice quality features from running speech, we have based our features on signals from miniature accelerometers attached to the speaker's neck below the level of the glottis where the influence of the vocal tract is minimal.

The results presented in this work indicate that voice quality varies over syllable-sized units roughly in the same way as intensity level. The middle part of the last voiced interval was associated with higher intensity level and, consequently, more periodic phonation, indicated by higher CPPS and lower L_1L_2. This effect is present in both languages and regardless of the different segmentation methods employed in the two data sets.

The other effects investigated here, although present, are very small and did not vary across languages. In line with previous research [8, 9, 11], between-speaker intervals were characterised by less periodic phonation (lower CPPS, higher L_1L_2) but the differences were below 1 dB, on average. However, it is possible that the feature extraction methods employed, and in particular the removal of frames with pitch-halving, reduced the influence of creak on the results. Thus, the role of creak as a turn-taking cue might be greater than indicated by our results. It is also conceivable that the the BSS/WSS contrast is realised differently as a function of the duration of the last voiced interval, or depends on the duration of the talkspurts, with greater aperiodicity being more likely towards the end of longer turns [26]. Finally, voice quality cues might be employed over larger temporal domains than the last voiced interval of a talkspurt. We leave these questions open for future research.

Of all the features investigated here, intensity level showed the least amount of separation across the between and within-speaker categories. This stands in stark contrast to the results in [11],

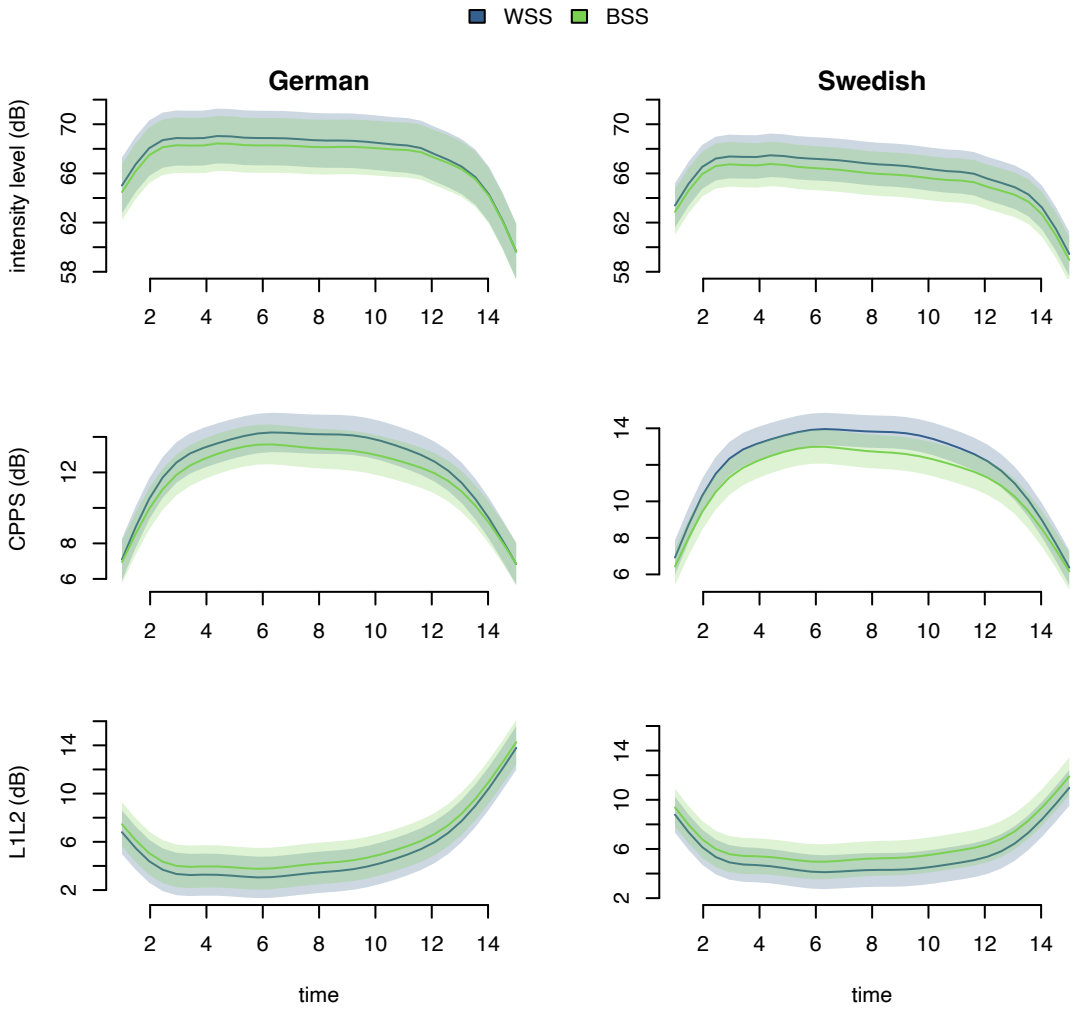


Figure 1: Model predictions for intensity level (top), CPPS (middle) and L_1L_2 in Swedish (left) and German (right) for between-speaker and within-speaker silences (BSS and WSS, respectively).

where intensity made the biggest contribution to discriminating between turn-taking categories. This could be partly explained by the fact that intensity level (as well as other f_o-independent features) in [11] included values extracted from intervals detected with a lower voicing threshold in order to capture even less periodic phonation types (this procedure was not used in the present study to produce time-aligned trajectories for f_o-dependent and f_o-independent features). Such intervals, which involve very breathy or creaky voice quality, are likely to be much quieter than the voiced intervals investigated here. Furthermore, the analysis in [11] did not control for covariates such as interval duration and did not model the grouping structure in the data.

The fact that the CPPS and L_1L_2 curves in Figure 1 exhibit greater separation across the BSS/WSS classes might suggest that voice quality does in fact carry non-redundant information related to turn-taking. Similarly, while the shape of the L_1L_2 contour is essentially a mirror image of the intensity level curve, the CPPS shows a more distinct, symmetrical pattern, possibly indicating that CPPS is to a less extent determined by sound pressure level (and subglottal pressure) variation. Nonetheless, given the small effect sizes, the effects are most likely of little practical importance in speech communication.

5. ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council project *Prosodic functions of voice quality dynamics* (VR 2019-02932) to Marcin Włodarczak.

6. REFERENCES

[1] J. Local, "Variable domains and variable relevance: Interpreting phonetic exponents," *Journal of Phonetics*, vol. 31, no. 3, pp. 321–339, 2003.

[2] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica*, vol. 62, no. 2–4, pp. 215–226, 2005.

[3] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, 2011.

[4] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2012, pp. 378–423.

[5] A. N. Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The voice prominence hypothesis: The interplay of f0 and voice source features in accentuation." in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 3527–3531.

[6] A. N. Chasaide, I. Yanushevskaya, and C. Gobl, "Prosody of voice: Declination, sentence mode and interaction with prominence," in *Proceedings of the XVIIIth International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK, 2015.

[7] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[8] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[9] M. Heldner, M. Włodarczak, Š. Beňuš, and A. Gravano, "Voice quality as a turn-taking cue," in *Proceedings of Interspeech 2019*, Graz, Austria, 2019, pp. 4165–4169.

[10] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," in *Proc. Interspeech 2017*, 2017, pp. 2351–2355.

[11] M. Włodarczak and M. Heldner, "Contribution of voice quality to prediction of turn-taking events," in *Proceedings of Speech Prosody 2022*, Lisbon, Portugal, 2022, pp. 485–489.

[12] A. S. Fryd, J. H. Van Stan, R. E. Hillman, and D. D. Mehta, "Estimating subglottal pressure from neck-surface acceleration during normal voice production," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. 1335–1345, 2016.

[13] V. S. McKenna, A. F. Llico, D. D. Mehta, J. S. Perkell, and C. E. Stepp, "Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, pp. 3404–3416, 2017.

[14] D. D. Mehta, V. M. Espinoza, J. H. Van Stan, M. Zañartu, and R. E. Hillman, "The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. EL386–EL392, 2019.

[15] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5060–5064.

[16] M. Włodarczak, B. Ludusan, J. Sundberg, and M. Heldner, "Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound," *Journal of Voice*, In press.

[17] K. Laskowski, "Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2011.

[18] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011, pp. 2837–2840.

[19] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, 1996.

[20] J. Sundberg, "Flow glottogram and subglottal pressure relationship in singers and untrained voices," *Journal of Voice*, vol. 32, no. 1, pp. 23–31, 2018.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Computer program, 2022.

[22] S. N. Wood, *Generalized Additive Models. An Introduction with R*. Boca Raton: CRC Press, 2017.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[25] M. Sóskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, pp. 1–22, 2021.

[26] K. Aare, P. Lippus, M. Włodarczak, and M. Heldner, "Creak in the respiratory cycle," in *Proceedings of Interspeech 2018*, 2018, pp. 1408–1412.

PEOPLE WITH PARKINSON'S DISEASE EXHIBIT PHONETIC FLEXIBILITY

Véronique Delvaux[1,2], Eva Goeseels[1], Kathy Huet[1], Myriam Piccaluga[1] Virginie Roland[1], Clémence Verhaegen[1], Bernard Harmegnies[1]

[1]Research Institute for Language Science and Technology, UMONS, Belgium
[2] Fund for Scientific Research (F.R.S.–FNRS), Belgium
veronique.delvaux@umons.ac.be

ABSTRACT

Phonetic flexibility is defined as the ability of the speaker-listener to adapt his speech behavior to the internal and external constraints weighing on the communication situation. This exploratory study deals with phonetic flexibility in elderly people with Parkinson's disease (PD). Twelve participants aged 57 to 79, half of whom had PD, produced the 25 sentences of the corpus under 6 successive experimental conditions: (i) reading; (ii) repetition of an oral model; (iii) interactive game; (iv)=(ii); (v) repetition with instruction to imitate the oral model; (vi) repetition with instruction to inhibit imitation. The results show that the participants, regardless of whether or not they belonged to the PD group, exhibited substantial phonetic flexibility. Indeed, the average pitch and sentence durations systematically deviated from those measured in reading to approach those of the target, and this according to the experimental condition: moderately in repetition, greatly in imitation, almost not in inhibition.

Keywords: Parkinson's disease, phonetic flexibility, phonetic convergence, repetition, imitation.

1. INTRODUCTION

Phonetic flexibility is defined as the ability of the speaker-listener to adapt his/her speech behavior to the internal and external constraints of the communication situation. Speech is thus considered as a complex behavior supported by multiple, complex and interactive cognitive processes, under the control of the executive functions.

In the linguistic tradition, it is generally considered that the sound form of messages is determined by linguistic constraints and rules, but also depends on a large number of so-called "paralinguistic" factors (attitudes, emotions, dialectal or sociocultural origin of the speaker, etc.), or even "extra-linguistic" factors (e.g. sex, age, physical attributes of the speaker). The second half of the 20th century has documented many cases of phonetic variation related to the dialectal or sociolinguistic origin of the speakers. The last twenty years have seen many advances in our understanding of the effect on speech and voice of many other factors, which had been previously little explored, e.g., acoustic correlates of body size and mass [1], vocal expression of emotion [2], prosodic characterization of attitudinal factors such as irony [3], effects of stress and cognitive load [4], etc.

In traditional linguistics, the focus is on the form of the message, which is considered to be the resultant – unique? – of the combined action of the linguistic, paralinguistic and extralinguistic factors mentioned above. We propose to reverse the perspective and to consider speech from the point of view of cognitive sciences centered on the human subject, the speaker-listener. Each speech production is then the unique and irreducible product of an interaction between human subjects in a specific communication situation. The different factors determine a framework of constraints - the field of possibilities - but do not allow the prediction of a unique sound form. The individual shows flexibility, adaptability and phonetic control. According to her needs, expectations, preferred strategies, which can themselves evolve in relation to the situation, the speaker constantly adapts, in real time, the phonetic form of her speech productions.

Recent work has demonstrated the existence of important resources in speakers in terms of what we call "phonetic flexibility". This is evidenced by the numerous studies on imitation and vocal disguise abilities [5]. Moreover, it is now accepted that this flexibility potential is used daily by speakers during language interactions (cf. the phonetic convergence mechanism typically occurring between two interlocutors: [6]), as well as all the phenomena of adaptation to the interlocutor: 'child-directed speech', 'foreigner-directed' speech, speech addressed to the hearing impaired, etc.; for a review, see [7]). More generally, a large literature concerns adjustments to speech productions in response to constraints on the communication situation, both 'external' constraints (e.g., speech in noise [8]), and constraints internalized by the speaker (the traditional hypo- vs. hyper-articulation continuum: [9,10]). Finally, recent experiments using various speech perturbation paradigms have shown the extent to which speakers are able to adjust their motor programs in real time in order to compensate, totally or partially, for the effects of an articulatory (lip tube, bite block) or auditory (altered feedback) perturbation (e.g. [11]).

These adjustment/adaptation mechanisms of the healthy speaker constitute in fact as many resources accessible to the individual confronted with a set of

disturbances due to the development of a pathology affecting speech. Beyond the effects of the pathology itself, of the drug treatment, as well as their interactions, it is commonly admitted that patients set up specific compensatory strategies in order to maintain an adapted production of their speech sounds. These compensatory strategies, to be effective, must be based on a solid foundation of phonetic flexibility, as well as on other elements such as preserved access to cognitive resources (elderly and/or neurodegenerative patients) and motivation (personal, professional) to maintain rich and nuanced spoken communication with the environment.

Previous work in our laboratory has investigated phonetic flexibility abilities via a phonetic learning paradigm of aspirated plosives in French-speaking speakers over 65 years of age [12], as well as in individuals with speech aphasia [13]. In the study reported here, we explore phonetic flexibility ability in elderly individuals with and without Parkinson's disease via a repetition/imitation/inhibition paradigm [14]. Parkinson's disease (PD) is a neurodegenerative disease responsible for progressive disorders of the motor sphere. Among the associated disorders is dysarhtria, a generalized disorder of motor execution of speech, involving difficulty in initiation, lack of coordination or slowness in the execution of articulatory gestures, which typically results in dysphonia, articulatory imprecision and prosodic insufficiency (monopitch, monoloudness, timing abnormalities), associated or not with respiratory problems [15]. At an early stage of the disease, the atypicalities observed in the speech may be consecutive to the disorder itself (or even to the drug treatment), but also to compensatory strategies put in place by the patient to cope with it (e.g. maintaining speech rate at the expense of articulatory precision).

The principle of the experimental paradigm used here is to induce in the participants controlled variations in speech rate and pitch register, linked to instructions associated with a succession of experimental tasks or situations involving spoken communication. In addition to the clinical perspectives, this research aims at better understanding the cognitive and linguistic mechanisms that support the production of speech, conceived as a strategic behavior.

2. MATERIAL AND METHODS

2.1. Participants and speech material

Twelve participants, six men and six women between the ages of 59 and 79, took part in the study. Half of them had Parkinson's disease. Relevant biographical characteristics are collected in Table 1.

The corpus consisted of 25 sentences, 5 initial sentences (produced by a 22-year-old native French speaker) each having undergone 5 types of modifications. These modifications were performed via the "Manipulate" procedure in Praat and concern either the total duration (75%, 100%, 125% of the initial duration) of the sentence, or its average pitch (75%, 100% or 125% of the initial average pitch). The initial phrase is thus coded D100P100 and the four other versions: D75P100, D125P100, D100P75, D100P125. The sentences were: "Lalie vend du lilas (*Lalie sells lilacs*), "Le loup vit dans les bois" (*The wolf lives in the woods*), "C'est un papa ou un papy?" (*Is he a daddy or a grandpa*?), "Il a une barbe ou une moustache?" (*Does he have a beard or a moustache*?), "Il a une grande bouche ou une petite bouche?" (*Does he have a big mouth or a small mouth*?). The sentences in the corpus vary in length (6 to 11 syllables), syllabic complexity and phonological content (e.g., the first two sentences are totally voiced), as well as in prosodic structure.

	Healthy speakers					
Speaker	S01	S02	S03	S04	S05	S11
Sex	F	F	F	M	M	M
Age	64	64	75	75	67	79
Education	Ba	Ba	Ma	Ba	Ma	Ma
	PD patients					
Speaker	S06	S07	S08	S09	S10	S12
Sex	F	F	F	M	M	M
Age	59	71	66	64	57	75
Education	Ba	Ma	Ba	Ma	Ba	Ma
MMSE	30	29	30	30		29
Medication	Azi, Pro	Pro	Pro	Sta, Azi	Mo, Azi	Pro
Year of diagnosis	2013		2018	2013	2014	2016
Prior SLT	No	Sw	Sw	No	No	No
Complaint	N/A	Tr	N/A	A.R.	N/A	Tr

Education: Ba/Ma = bachelor's/master's degree
Medication: Azi: Azilect; Pro: Prolopa; Sta: Stalevo; Mo: Modapar
Prior Speech-Language Therapy: Sw: swallowing
Complaint: Tr: Tremor; A.R.: Arm rigidity

Table 1: Participants' characteristics

2.2. Experimental paradigm and measurements

The experimental paradigm consisted of six successive steps, completed in a single experimental session. At each stage of the paradigm (except (iii)), the participant produced the 25 sentences of the corpus according to a specific task: (i) reading; (ii) repetition of an oral model (the "target" voice); (iii) interactive game; (iv) repetition of an oral model (same as (ii)); (v) repetition with explicit instruction to imitate the oral model; (vi) repetition with explicit instruction to inhibit as much as possible the imitation of the oral model. The 25 sentences were presented in the same pseudo-random order at each stage. The interactive game was a slightly adapted version of the board game "Who am I?", so only certain sentences from the corpus were likely to be produced, without guarantee. The games (at least four games per participant) were all played between the participant and one and the same interlocutor, the one to whom the target voice belonged. The speech productions during the game are not analyzed here.

Acoustic data were recorded using a cell phone (iPhone 8) placed 30 cm from the participant's mouth on a separate stand. The recordings were converted to .wav

format (44100Hz, mono) and processed in Praat. A dedicated script allowed the alignment between the transcriptions of the produced sentences and the speech signal, as well as the taking of measurements under the supervision of the user. The measures are: the total duration of the sentence, then converted into speech rate (number of syllables per second), as well as the fundamental frequency (Hz) calculated every 5 ms, from which the median value was extracted and expressed in z-score. For each utterance of each participant, we thus have two indicators, which form the main dependent variables of the statistical analysis: the median pitch (z-score) and the speech rate (syll/s).

Moreover, for tasks (ii), (iv), (v), (vi), these two indicators were transformed into a "convergence rate". The principle is to express (in %) the proportion of the path covered in relation to the distance to be covered to reach the target. The absolute value (e.g., the average measured pitch for a given repetition of a given sentence) is expressed relatively to the target:

$$\frac{\text{Measured\_Pitch} - \text{Baseline\_Pitch}}{\text{Target\_Pitch} - \text{Baseline\_Pitch}} * 100$$

The baseline selected was the average of the values measured for the respective participant during the reading task. The convergence rate could indicate either a partial convergence (e.g. 25%) or almost total convergence (90-100%), or even an overshoot of the target (above 100%) or a divergence (negative value).

3. RESULTS

Two repeated measures analyses of variance (ANOVA) were first performed, one with speech rate (syll/s) as the dependent variable, the other with pitch (z-score). The within-subjects factor was the task (pregame repetition; postgame repetition; imitation; inhibition), the between-subjects factor concerned the characteristics of the stimuli (respectively, for rate: D75, D100, D125; for pitch: P75, P100 and P125). In these analyses, the participant was considered as a random factor and the group (control vs. PM) was not taken into account. Thus, this was an initial exploration of the data, all subjects combined.

Analyses revealed a significant effect of task and stimulus characteristics on rate and pitch, as well as a significant interaction of these two factors illustrated in Fig.1 (Rate: $F(4,40.003)=30.255$, $p<.001$; Pitch: $F(4,40.081)=10.272$, $p<.001$). In repetition tasks, speakers tended to follow the oral model, producing sentences with higher/slower speech rates in response to higher/slower speech rates (similarly for pitch). In the imitation task, the observed differences in production as a function of stimulus characteristics were exacerbated (slopes are steeper in Fig.1). In the inhibition task, the line is almost horizontal, which means that speakers were globally successful in maintaining a given level of rate or pitch, regardless of those characterizing the voice that pronounced the sentences to be repeated.

In a second step, the effect of group (participants with and without Parkinson's disease) was studied. A multivariate analysis of variance was performed with convergence rates calculated for flow and pitch as dependent variables and task (repetition: includes pre- and post-game repetition tasks; imitation; inhibition) and group (control vs. PD) as independent variables.

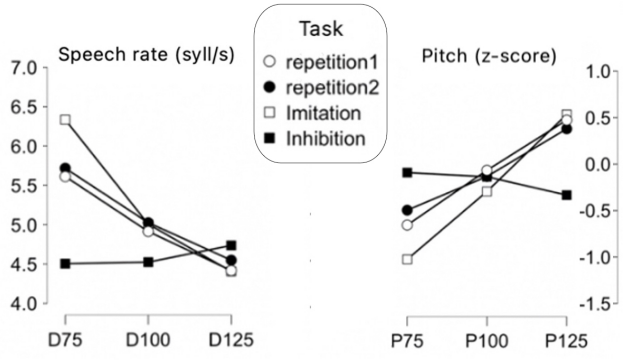


Figure 1: Speech rate (syll/s) (left) and pitch (z-score) (right) as a function of task and stimulus properties. All participants combined.

Regarding the convergence rate on speech rate, the analysis revealed only a significant effect of task ($F(2,1045)=6.713$, $p= .001$). There was no group effect, i.e. subjects with PD behaved like control subjects for speech rate (Fig.2, left): there was low convergence in repetition, high convergence (about 50% on average) in the imitation task and high divergence rates in inhibition.

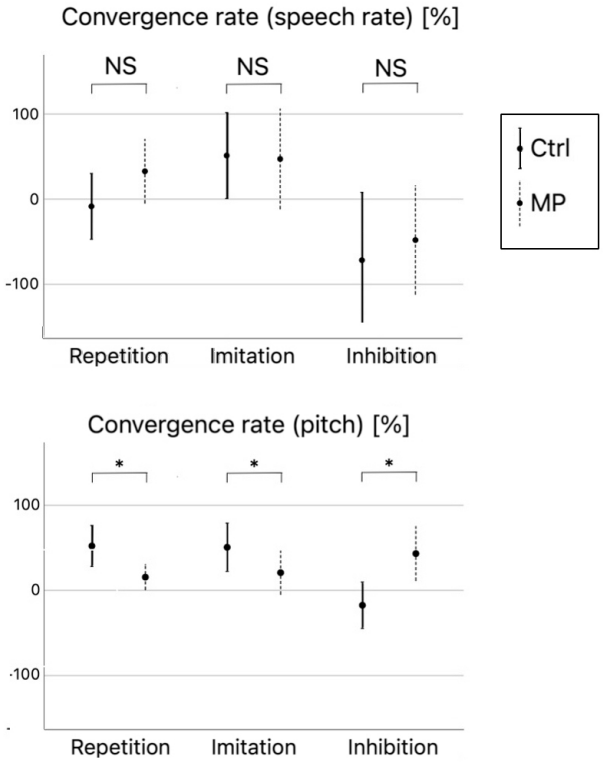


Figure 2. Convergence rate (%) as a function of task and participant group. Left: Speech rate, right: Pitch.

Regarding the convergence rate on pitch, only the Group*Task interaction was significant ($F(2,1045)=7.939$, $p< .001$). Regarding the convergence rate on pitch, only the Group*Task interaction was significant ($F(2,1045)=7.939$, $p< .001$). Fig.2 (right) illustrates this interaction: speakers in the control group converge to the pitch of the target voice, in both repetition and imitation, and manage to counteract this tendency in inhibition. In contrast, speakers with Parkinson's disease are unable to follow the given instructions, especially in the inhibition task: they approach (slightly) the pitch of the target voice in all tasks.

Note that when we added Sex to the independent variables in the statistical analysis, we observed a significant triple interaction Sex*Group*Task for the convergence rate on pitch ($F(2,1039)=7.915$, $p< .001$). This interaction reflects the fact that the trend shown in Fig.2 (right) concerned exclusively female subjects. Men, with or without PD, did not converge on pitch, regardless of the experimental task.

4. DISCUSSION

The aim of this study was to explore phonetic flexibility in French-speaking participants over 60 years of age, half of whom have Parkinson's disease. The paradigm used allowed us to study the fluctuations of the phonetic parameters of speech productions in relation to the production conditions. On the one hand, the instructions varied according to the task: simple repetition, explicit imitation, repetition without imitation, of the model voice. On the other hand, the sentences to be produced were manipulated in such a way as to obtain 5 versions of the same sentence, differing only in the speech rate or the pitch register.

First, the results showed that the subjects in the study were overall capable of phonetic flexibility. Their productions varied in the expected direction. When asked to simply repeat the sentences produced by a target voice, participants tended to imitate the speech rate as well as the fundamental frequency of the target voice, in a moderate way. This result, typical of the shadowing task, can be interpreted as a latent phenomenon of phonetic convergence, which occurs even in non-interactive communication situations [16]. When the instruction was explicitly to imitate the target voice, the convergence was more marked (the average convergence rate was close to 50% in imitation for the control subjects, for both speech rate and pitch). In contrast, convergence was close to zero, or even turned into divergence (speech rate), during the voluntary inhibition of imitation task, confirming the ability of the participants to actively adapt their speech productions in order to follow a given instruction. Comparative analysis of convergence rates as a function of conditions (shown in Fig.2) suggests that speech rate was "easier" for participants to manipulate at will than the mean register of f0.

Overall, this study confirms previous results obtained on a population of elderly French-speaking subjects via a phonetic learning paradigm [12]. The two experimental paradigms are different, as well as the subjects who participated in the two studies, but the finding at the end is similar: even elderly speaker-listeners are able to adapt in real time their speech productions in order to accomodate the internal and external constraints that frame the communication situation. Theoretical implications are numerous, especially for the study of the effects of aging on speech motor control (see e.g. [17]).

The study presented here also provides an exploratory approach to the question of the potential preservation of phonetic flexibility resources in people with PD. The exploratory nature is due to the small number of participants studied to date, with additional subjects currently being recruited to complete the numbers. Initial results suggest that PD participants exhibit phonetic flexibility, as they do not behave significantly differently from controls, at least for speech rate. On the other hand, PD subjects were found to be unable to converge to (or diverge from) the fundamental frequency of the target voice when instructed. This result could be interpreted in light of the prosodic insufficiency typically observed in these patients [18]. Recall, however, the interaction found with Sex: men, with or without PD, did not converge on the pitch, regardless of the experimental task. It is indeed worth noting that the target voice was a female voice (mean pitch at P100: 209 Hz), which makes the imitation/inhibition task radically different for male participants compared to female participants. The small number of participants does not allow at this stage to generalize the effects highlighted by the statistical analysis on our sample: is the lesser convergence on the pitch an effect of gender, an effect of the disease, or more globally an expression of individual variation? Future analyses on a larger sample should answer this question.

In any case, the phonetic flexibility demonstrated by PD patients with respect to speech rate (whereas deficits related to the temporal properties of speech are typical of Parkinsonian dysarthria [15,18]) opens interesting clinical perspectives. The speech therapy management of these patients – unfortunately not very systematic in the typical care pathway – could be based on these resources in order to develop compensatory strategies best adapted to their motor speech disorders.

In conclusion, the exploratory study presented here confirms previous results which suggest that phonetic flexibility capacities are maintained after 60 years of age. These abilities would also be preserved, at least in part, in people with Parkinson's disease. Thus, the production-perception link would remain effective, mobilizable in order to respond to the constraints of the communication situation, and this despite the presence of a motor execution disorder more or less extensive in these patients.

5. ACKNOWLEDGEMENTS

This study was partially funded by the Action de Recherche Concertée ARC "EvalDy" Grant (Wallonia-Brussels Federation, UMONS), Belgium. Thanks to Yousra El Omari for her help in data collection.

6. REFERENCES

[1] Pisanski K., Fraccaro P., Tigue C., O'Connor J., Röder S., Andrews P., Fink B., Debruine L., Jones B., Feinberg D. 2014. Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour* 95, 89–99.

[2] Grandjean D., Baenziger T. 2019. Chapitre 4. Expression vocale des émotions. In : Sander D. (ed.), *Traité de psychologie des émotions* (pp. 109-155). Paris:Dunod.

[3] Rouas J.-L., Shochi, T., Guerry M., Rilliard A. 2019. Categorisation of spoken social affects in Japanese: human vs. machine. *Actes de l'International Congress of Phonetic Sciences ICPhS*, Aug 2019, Melbourne, Australia. ffhal-02317743

[4] Huttunen K., Keränen H., Väyrynen E., Pääkkönen R., Leino T. 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied ergonomics* 42, 348-57.

[5] Revis J., De Looze C., Giovanni A. 2013. Vocal flexibility and prosodic strategies in a professional impersonator, *Journal of Voice* 27, 524.e23- 524.e31.

[6] Pardo J. S., Urmanche A., Wilman S., Wiener J., Mason N., Francis K. & Ward M. 2018. A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics* 69, 1-11.

[7] Cooke M., King S., Gargnier M., & Aubanel V. 2014. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language* 28, 2, 543-571.

[8] Garnier M., Henrich Bernardoni N. 2014. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*. Doi: 10.1016/j.csl.2013.07.005.

[9] Moon S.J., Lindblom B. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.* 96, 40–55.

[10] Smiljanic R., Bradlow A.R. 2008. Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics* 36, 91-113.

[11] Ménard L., Perrier P., Aubin J. 2016. Compensation for a lip-tube perturbation in 4-year-olds: Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *J. Acoust. Soc. Am.* 139(5), 2514.

[12] Delvaux V., Cano-Chervel J., Huet K., Leclef C., Verhaegen C., Harmegnies B. 2015. Capacités d'apprentissage phonétique et vieillissement, in : A.-M. Etienne (ed). *Evolutions sociales, Innovations et Politiques: nouvelles questions pour la psychologie de la santé*, Mardaga

[13] Verhaegen C., Delvaux V., Huet K., Fagniart, S., Piccaluga M. Et Al. 2020. Capacités d'apprentissage phonétique chez des patients aphasiques francophones: étude de cas. *Actes des 33e Journées d'Études sur la Parole (JEP),* 2020, Nancy, France, 606-616.

[14] Delvaux, V., Harmegnies, B., Huet, K., Piccaluga, M., Roland, V., Verhaegen, C. (2022) Les personnes atteintes de la maladie de Parkinson font preuve de flexibilité phonétique. *Proc. XXXIVe Journées d'Études sur la Parole*, JEP 2022, Noirmoutier, 289-297.

[15] Rusz J., Cmejla R., Ruzickova H., Ruzicka E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J. Acoust. Soc. Am*. 129, 350–367.

[16] Delvaux V., Soquet A. 2007. The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica* 64, 145-173.

[17] Fougeron C., Guitard-Ivent F., Delvaux V. 2021. Multi-Dimensional Variation in Adult Speech as a Function of Age. *Languages* 6(4), 176.

[18] Skodda S., Visser W., Schlegel U. 2011. Gender-Related Patterns of Dysprosody in Parkinson Disease and Correlation Between Speech Variables and Motor Symptoms. *Journal of Voice* 25(1), 76–82.

TEMPORAL ASPECTS OF TURN-TAKING IN ZOOM CONVERSATIONS

Qiang Xia

Humboldt-Universität zu Berlin
qiang.xia.1@hu-berlin.de

ABSTRACT

The proper timing is crucial for turn transitions. Turn-taking in conversations usually takes place without noticeable gaps or overlaps longer than 200 ms [1, 2]. With the widespread adoption of video-mediated communication during the COVID-19 pandemic, there is a growing need to understand how turn-taking behavior differs in remote situations. The current study compares several temporal aspects of turn-taking in face-to-face and Zoom interactions. Spontaneous dialogues in the Berlin Dialogue Corpus [3] were investigated. Twenty native German speakers conversed in pairs to complete two spot-the-difference tasks in the respective situations. Interlocutors tend to speak more slowly and produce longer gaps and longer overlaps between turns over Zoom. In face-to-face interactions, speaker change happens more frequently, leading to slightly more overlaps. This implies that dialogues over Zoom follow a distinct temporal pattern from those in co-present situations. A threshold higher than 200 ms is necessary to describe the temporal pattern of turn-taking in Zoom conversations.

Keywords: Zoom-conversation, turn-taking, speech corpus, spontaneous speech.

1. INTRODUCTION

Since Sacks, Schegloff and Jefferson [1] first focused on the "speech exchange systems" and concluded that conversations overwhelmingly follow the "one-speaker-at-a-time" pattern where turn transitions with no gap and no overlap are quite common, turn-taking behaviour has been extensively studied for almost fifty years. Empirical studies confirmed that the timing of turn-taking is highly precise, but not as seamless as "zero-gap-zero-overlap" [4, 5]. Transitions with acoustic silences of about 200 ms compose the majority of turn switches. Actually, gap durations within a range of 250 ms are examined to be cross-linguistically valid [6, 7]. Wilson and Wilson [2] explained that smooth transitions are realised by the synchronisation of neural oscillators based on speaker's syllable rate, about 200 ms in informal English dialogues. Thus, listeners' readiness to speak usually counterphases with that of the speaker.

However, it is noteworthy that face-to-face interaction has been the default conversational setting in the body of literature. Still relatively little is known about how turn-taking behaviour is organised in video-mediated conversations, though the COVID-19 pandemic has immensely increased our use of videoconferencing programs such as Zoom in the last few years.

Zoom conversations differ from face-to-face interactions in several ways. Some important para-linguistic cues, for instance, gaze direction, that can facilitate anticipating turn-ends [8], are hardly available in remote conversations [9]. In fact, it has been demonstrated that conversational rhythm is disrupted in Zoom interactions [10], partly because electronic transmission takes time and unavoidable delays of about 30–70 ms are long enough to disturb neutral oscillators that synchronise during conversation [2, 10]. They reported massive differences of turn transition time in face-to-face (135 ms) and Zoom interactions (487 ms). Given the differences between the two conversational situations, Zoom interactions are assumed to include longer gaps. The 200 ms threshold commonly used to describe the turn-taking rhythm in face-to-face conversations may not be suitable for Zoom conversations.

This paper aims to investigate how turn-taking behaviour differs between face-to-face and Zoom interactions in terms of temporal aspects, specifically identifying the threshold that describes the majority of turn transition in Zoom interactions.

2. METHODS

The Berlin Dialogue Corpus [3] was investigated in the study. Twenty native German speakers (mean age = 25.7, SD = 3.8, 10 females, 10 males) who knew each other well prior to the experiment were asked to finish four spot-the-difference tasks (two over Zoom and two in face-to-face situation) in pairs. In the so-called Diapix tasks [11, 12, 13],

participants needed to cooperate and identify 13 differences between two similar pictures within 10 minutes. In the co-present situation, participants conversed in a phonetic laboratory. For the Zoom conversation, participants were sitting in different rooms in the same university building and connected via Zoom. To avoid differences arising from recording methods, conversations were recorded by separate microphones in both situations. Zoom acknowledges that there can be latency from a few milliseconds to a few seconds during Zoom calls. Yet, the delay durations vary with various factors and can hardly be specified [14]. Using separate recordings can include the latancies perceived by both sides. Recordings from the separate channels were merged together to reconstruct the dialogue. The collected data and its orthographic transcription were first aligned by WebMaus [15]. Turn-taking relevant elements like speech turns were annotated in Praat [16]. Interval boundaries have been manually corrected. After annotation, the automatic detection tool [17] was used to identify empty and overlapping annotation chunks. In this way, gaps and overlaps between speech turns were created automatically based on the annotations aligned with speech signals. In the end, the data were transformed into an emuR-database [18] for temporal analysis.

The study focused only on between-speaker silences, specifically gaps between speech turns, while silences within a speech turn were not considered. Gaps and overlaps are often treated as two ends of a continuum: when measuring the interval between the end of the first speaker's speech and the starting point of the second speaker's speech, gaps have positive values and overlaps negative [6, 5, 19].

To compare speech rate in the two situations, syllable durations were calculated by subtracting pauses within a turn from the length of the turn before dividing it by the number of syllables produced within it. Syllable were counted by using the R-package SYLLY.DE [20].

In order to investigate the correlation between the position of gaps or overlaps and the task-oriented conversation timeline, the starting point of a gap or overlap was normalised by dividing it by the entire conversation length and rounded to two decimal place. The study excluded the preparing and ending phase of a conversation, where speakers were trying to be heard or finish the recording, corresponding to the time point beyond the range from 0.05 to 0.95. For each normalised time slot, gap and overlap durations were averaged.

3. RESULTS

Extra long gaps are observed at the beginning and near the end of conversations, because participants were still trying to adjust themselves to the new conversational settings or they had difficulties finding more differences. To eliminate the influence of these extra-linguistic factors, gaps that exceeded the upper bound of the data (Q3+1.5IQR, calculated as the upper quartile of the data plus 1.5 times the interquartile range) were removed. Similarly, outliers of overlaps (longer than the lower bound, Q1-1.5IQR) were not taken into account. Previous studies [5, 21] have shown that geometric means provide a more realistic estimate of central tendency than arithmetic means. Therefore, geometric means are presented in the figures.

3.1. Syllable durations

On average, it takes about 217 ms for speakers to produce a syllable over Zoom, significantly longer than in face-to-face interactions, about 206 ms per syllable (t = -5.97, DF = 39, $p < .001$). In other words, speakers tend to speak more slowly when they talk over Zoom; as illustrated in Fig. 1.

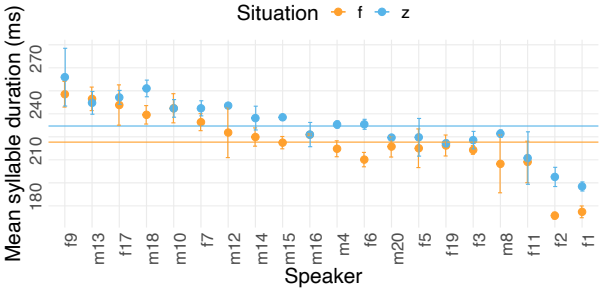


Figure 1: Mean syllable durations in face-to-face (f) and Zoom (f) situations by speaker with error bars. The horizontal lines show the respective mean syllable durations in the two situations.

3.2. Gap durations

As can be seen from Fig. 2, gaps in face-to-face conversations (shown in yellow) are generally shorter than those in Zoom situations. In the co-present scenario, around 50% of gap durations were within the 270 ms range, whereas the median value in Zoom conversations was considerably longer, at approximately 438 ms. To examine the differences further, a linear mixed effect model was computed: using situation and the normalised time point of a gap in the dialogue as the mixed effect, gap duration

as the dependent variable and conversation as the random intercept. The results revealed that both situation and normalised time point significantly impact gap durations. Specifically, the Zoom situation increased the duration of gaps. Gaps are expected to be longer nearer to the end of a conversation. The correlations are additionally summarised in Fig. 3.

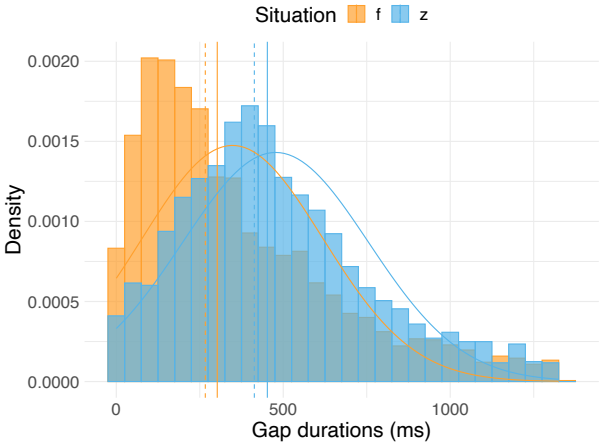


Figure 2: Histogram of gap durations in face-to-face (f) and Zoom (z) situations with the estimated distributions. Bin size 50 ms. Outliers excluded. Dashed lines show the geometric means, solid lines the medians.

Situation	Face-to-face	Zoom
Arithmetic mean	348.04	477.08
Geometric mean	231.60	365.76
Median	269.56	437.49
Mode	152.38	402.38
Standard deviation	270.52	278.80
Without outliers	3088	2722
Total N	3317	2936
Percentage	93.10 %	92.71%

Table 1: Descriptive statistics of gap durations in face-to-face and Zoom situations (in ms).

3.3. Overlap durations and occurrences

Fig. 4 presents that Zoom conversations exhibit longer overlaps compared to face-to-face situation. The median value of overlap durations in Zoom interactions is 416 ms, about 100 ms longer than the 319 ms observed in face-to-face situation; see Tab. 2. A mixed effect linear model was used to determine the effects of situation and normalised time point (mixed effect) on mean overlap durations (dependent variable) with conversation as a random effect. The results showed that only situation

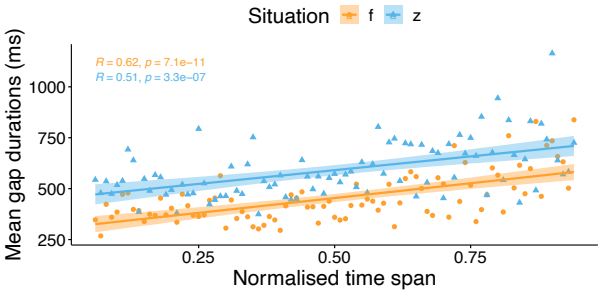


Figure 3: Linear regression between mean gap durations and normalised time point (range: 0.05-0.95) in conversations.

significantly affects overlap durations, and there is no correlation between the normalised time point of an overlap and its duration (see Fig. 5).

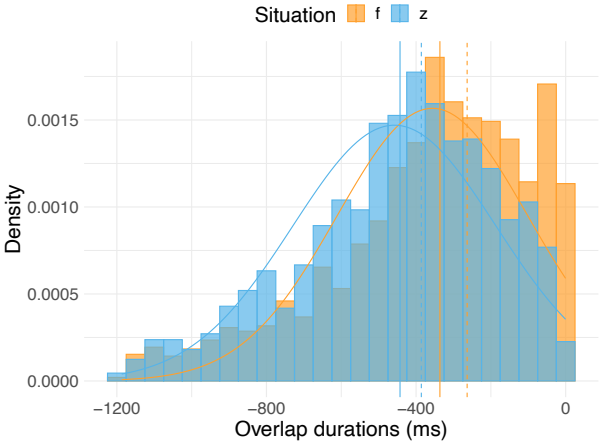


Figure 4: Histogram of overlap durations in face-to-face (f) and Zoom (z) situations with the estimated distributions. Bin size 50 ms. Outliers excluded. Dashed lines show the geometric means, solid lines the medians.

Situation	Face-to-face	Zoom
Arithmetic mean	-356.19	-457.95
Geometric mean	-238.27	-359.05
Median	-319.41	-416.35
Mode	-247.62	-397.62
Standard deviation	254.56	271.35
Without outliers	1935	1790
Total N	2047	1880
Percentage	94.53%	95.21%

Table 2: Descriptive statistics of overlap durations in face-to-face and Zoom situations (in ms).

Although more overlaps are observed in Zoom interactions, the difference is not significant at the 5% level (p = 0.23). Interlocutors switch more

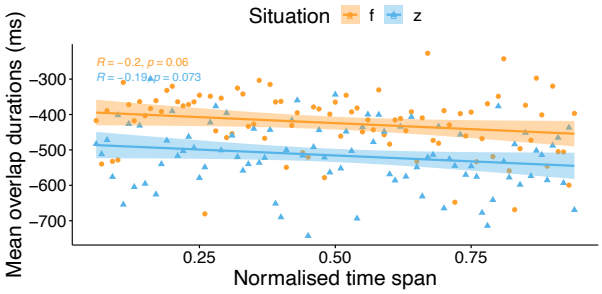


Figure 5: Linear regression between mean overlap durations and normalised time point (range: 0.05-0.95) in conversations.

frequently the speakership when talking face-to-face, approximately 18 times per minute; while in Zoom interactions, the turn transition frequency is about 15 times per minute ($t = 4.25, p < .05$). Yet, the overlap-to-transition ratio in co-present situation is smaller (0.63) than that in Zoom interactions (0.66). However, the difference was not significant.

4. DISCUSSION

It is common to have slight gaps and overlaps during turn-taking in conversations [1, 2, 19], but previous studies have suggested different interval durations for the *zero-gap-zero-overlap* pattern is not realistic [4], such as 120 ms in [5], 200 ms in [2, 22, 19]. The current study found that for face-to-face interactions, a threshold of up to 300 ms could describe the common cases of turn transition intervals, while for Zoom conversations, the threshold was 450 ms.

The longer gaps in face-to-face interaction might be due to the high cognitive loads resulting from the experimental settings. Participants were asked to cooperate to spot all differences between two pictures, which might be cause high cognitive loads while conversing [19]. However, studies suggesting 200 ms as the frequent threshold are primarily based on polar questions [2, 6]. Considering the differences in cognitive loads resulting from experiment settings, the longer gaps found here should come as no wonder.

In Zoom conversations, both gaps and overlaps are longer than those in the co-present situation. Unavoidable transmission delays can explain this phenomenon to some extend. To record conversations from a holistic view, recordings were made separately from both sides of a dialogue. By this means, actual latencies perceived by the two sides were included in the recording. Speakers would stop talking if overlap occurs, so that the *one-speaker-at-a-time* pattern can soon be restored [1]. If speech signal is received with latency, reaction on the signal will also be delayed. In this case, speakers would not be able to relinquish their speech turn *on time*. As a result, overlap continues a bit till it is perceived, causing longer durations. Unfortunately, statistics on electronic transmission delays over Zoom are not available to the public because of commercial interests and legal limitations. For this reason, the true proportion of technical latency in gap and overlap durations can hardly be removed from the current analysis. Even if the assumed 30-70 ms transmission delays [23] were subtracted from the gap and overlap durations in the Zoom data, the thresholds will still be longer than those in face-to-face situations.

Given transmission delays over Zoom and their interference in conversational rhythm, more overlaps had been expected in remote communication. Also [24] confirmed that latencies up to 800 ms increase the number of unintended interruptions. On the contrary, slightly more overlaps were found in face-to-face conversations, which can be traced back to the higher frequency of speaker changes in this situation. Edelsky [25] demonstrated that overlaps can reflect high interactivity and high engagement in a conversation. From this point of view, it can be assumed that participants were more engaged when talking face-to-face, hence overlapped more.

In addition, interlocutors spoke with a lower syllable rate when talking over Zoom. The reason for this difference might be the uncertainty caused by the omnipresent transmission delay. Though no obvious latency was reported during the experiment, speakers could probably still sense the brief delay, which made them unsure if their conversational partner could perceive their speech punctually and completely. It could also explain the the longer gaps in Zoom conversations since the speech rhythm of speakers will synchronise during a conversation [2].

5. CONCLUSION

This study aimed to compare temporal aspects of turn-taking behavior in face-to-face and Zoom conversations. Zoom interactions are found to have longer gaps and overlaps. However, face-to-face interactions had more frequent speaker changes and slightly more overlaps. The impact of transmission delay on turn-taking timing in video-based conversations needs to be explored further, including measuring the exact duration of latency in signal transmission and its interaction with the temporal pattern of turn-takings.

6. ACKNOWLEDGEMENTS

The author would like to thank Christine Mooshammer for valuable input and discussions. And a lot of thanks to Malte Belz for guidance in corpus annotation and helpful comments. Also thanks to reviewers for great suggestions. The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

7. REFERENCES

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic Bulletin & Review*, vol. 12, no. 6, pp. 957–968, 2005.

[3] M. Belz, A. Zöllner, M. Terada, R. Lange, L. S. Adam, and B. Sell, "Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo," 2021.

[4] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[5] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[6] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[7] K. Weilhammer and S. Rabold, "Durational aspects in turn taking," in *Proc. 15th ICPhS*, Barcelona, 2003, pp. 2145–2148.

[8] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[9] J. N. Bailenson, "Nonverbal overload: A theoretical argument for the causes of Zoom fatigue," *Technology, Mind, and Behavior*, vol. 2, no. 1, 2021.

[10] J. E. Boland, P. Fonseca, I. Mermelstein, and M. Williamson, "Zoom disrupts the rhythm of conversation," *Journal of Experimental Psychology: General*, pp. 1272–1282, 2021.

[11] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and Speech*, vol. 53, no. 4, pp. 510–540, 2010.

[12] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.

[13] O. M. Bullock and B. Sell, "PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels," 2022.

[14] Zoom, "Video conferencing, web conferencing, webinars, screen sharing," https://zoom.us/, 2023.

[15] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2022.

[17] T. Mahrt, "A python library for working with praat, textgrids, time aligned audio transcripts, and audio files: PraatIO," https://github.com/timmahrt/praatIO, 2016.

[18] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[19] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, 2015.

[20] M. Michalke, "Sylly: Hyphenation and syllable counting for text analysis," https://reaktanz.de/R/pckg/sylly/citation.html, 2018.

[21] J. P. de Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[22] M. B. Walker and C. Trimboli, "Smooth transitions in conversational interactions," *The Journal of Social Psychology*, vol. 117, pp. 305–306, 1982.

[23] J. E. Boland, "Conversation Transition Times: Working Memory & Conversational Alignment," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, A. K. Goel, C. M. Seifert, and C. Freksa, Eds. Cognitive Science Society, 2019, pp. 159–165.

[24] S. Egger-Lampl, R. Schatz, and S. Scherer, "It takes two to tango - Assessing the impact of delay on conversational interactivity on perceived speech quality," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, Interspeech 2010*, 2010, pp. 1321–1324.

[25] C. Edelsky, "Who's Got the Floor?" *Language in Society*, vol. 10, no. 3, pp. 383–421, 1981.

INVESTIGATING THE RELATIONSHIP BETWEEN PROSODIC ENTRAINMENT AND INTERACTION STYLE

Joanna Kruyt[1 2*], Lena-Marie Huttner [3*], Johannah O'Mahony [4*]

[1] Institute of Informatics, SAS, Bratislava, [2] Faculty of Informatics and Information Technologies, Slovak Technical University, [3] LPL, CNRS, Aix-Marseille University, [4] CSTR, University of Edinburgh, *All authors contributed equally to this paper

joanna.kruyt@savba.sk, lena-marie.huttner@univ-amu.fr, johannah.o'mahony@ed.ac.uk

ABSTRACT

Entrainment is a phenomenon whereby speakers imitate each others' speech patterns during a conversation. It is influenced by social and individual factors. This study aims to investigate whether interaction style predicts entrainment. Following existing studies, we operationalise interaction style using features related to expressive paralinguistics, turn-taking behaviour, and topic persistence. We extracted these features from conversations in the Switchboard I corpus before reducing them into a single *interaction style* variable. We measured the overall prosodic entrainment of the conversation as well as each speaker's contribution using a geometric approach. To investigate whether speakers' involvement scores and entrainment contributions are correlated, we used an actor-partner interdependence model. Results suggest that interaction style and prosodic entrainment are not associated. Possible explanations for these findings, as well as suggestions for future research are discussed.

Keywords: entrainment, convergence, involvement, interaction style, conversation style

1. INTRODUCTION

During interaction, interlocutors' speech patterns sometimes become more similar [1, 2]. This phenomenon is referred to as *entrainment*. Why entrainment occurs is debated: while some argue it is an automatic process resulting from priming [3, 4], others hypothesise that social factors play a central role. Communication Accommodation Theory (CAT) [5] states that speakers (dis-)entrain to minimise or emphasise social differences between themselves and their interlocutor. Entrainment may be more likely to occur if it yields a communicative benefit [6]. Studies have also found that individual differences, such as openness and attention switching abilities [7], can affect entrainment. Additionally, [8] found that individuals who speak with greater variability in their acoustic and articulatory features tend to entrain to others more than individuals who speak with less variability. This suggests that the way someone speaks may relate to entrainment at the prosodic level. In other words, entrainment behaviour seems to be affected both by social and individual factors.

A factor that pertains both to individual differences and to social factors is *interaction style*. Interaction style describes how a person behaves in a conversation in terms of turn-taking behaviour, discussed topics, and paralinguistic expressiveness [9, 10]. Interaction style and entrainment are both associated with rapport [9, 1] and both influence the way people coordinate conversations. We aim to investigate whether an individual's interaction style is associated with their prosodic entrainment behaviour.

Interaction style is an abstract concept that can be operationalised in different ways, but one of the most popular frameworks is Tannen's [9, 10]. Central to her theory is that interaction styles can be considered either *high involvement* or *high considerateness*. High involvement is characterised by a preference for personal topics, quick shifts in topic, topic persistence or repetition of previously brought up topics, telling of stories and rapid turn-taking with short or even negative pauses and overlapping speech. High involvement speakers have a relatively fast speech rate and large pitch and intensity ranges [9, 10]. High considerateness, on the other hand, is associated with little overlapping speech, longer, more frequent pauses both between and within turns, relatively slow speech rate, little variation in pitch, and a preference for less personal topics [9, 10].

Measuring interaction style relies on manual annotations of conversation style, which can be expensive and time-consuming. In this paper

we adopt a method by [11], which relies on automatically extracted features to quantify interaction style. While [9, 10] describes high involvement and high considerateness as two opposing styles, in [11]'s operationalisation, which we follow here, interaction style is a spectrum. To our knowledge, no research has directly investigated the association between interaction style and prosodic entrainment. However, [12] suggested that entrainment behaviour is part of linguistic style, and that speakers tend to match each other's linguistic style. [13] found that more prosodic entrainment was observed in conversations where speakers were more engaged or involved with one another. Both [12] and [13] measured their variables at the conversation level, rather than for the individual speakers. In this study, we measured entrainment and involvement per speaker to characterise the relationship between the two.

We aim to investigate whether a difference in interaction style affects entrainment in conversation: according to Tannen, having similar conversation styles facilitates conversation [9, 10], and may thus decrease the perceived social distance between speakers. Following CAT [5], this perceived social distance based on difference in interaction style may influence entrainment.

2. METHOD

2.1. Data

We analysed Switchboard I (SWB) [14] using the Mississippi State University word alignments. We excluded any recordings where a speaker spoke to someone in the background. For our analysis we used 1295 conversations in which both speakers were the same gender (so they could be considered "indistinguishable dyads" in our statistical analyses).

2.2. Measure of involvement

To quantify *interaction style*, we follow the method used in [11]. [11] selected and extracted 11 variables that reflect the characteristics of involvement – topic, pace, and paralinguistic expressiveness – according to Tannen [9, 11]. We extract the same variables as [11], but have operationalised them slightly differently to allow for cross-corpus replication. Unlike [11], who used a line of transcript from a speech-to-text system as their main conversational unit, we used inter-pausal units (IPUs) as our baseline unit. Following [15], the data were split into IPUs using a silence threshold of 180ms and a minimum IPU duration threshold of 100ms [16]. We treated non-speech markers in the SWB dataset as silences. We excluded any word transcriptions with a duration of 0s. We implemented the silence and overlap classification (SOC) of [15] to find pauses (segment of silence within a turn of a single speaker), gaps (silence between turns of two different speakers) and overlaps.

2.2.1. Acoustic Features

[11] extracted features to reflect a speaker's expressive paralinguistics, based on [9]. Pitch variance (pv) was calculated by extracting pitch for each IPU of a speaker using the autocorrelation function in Praat [17] via Parselmouth [18]. We used the two-pass method in [19] to determine the speaker-specific pitch floor and ceiling and removed pitch points which were more than two standard deviations from the speaker's mean. Pitch values were then transformed into semitones using the speaker's median, and the pitch range was calculated. Intensity variance (iv) was calculated by extracting the intensity for each IPU using Praat via Parselmouth [18] and calculating range across all IPUs. Finally, speech rate (wps) was calculated in words per second across the entire recording.

2.2.2. Turn Features

[11] selected a number of features to quantify turn-taking behaviours. We calculated mean length of between-own pauses (boplen) by taking the mean duration of all *pauses* per speaker in the SOC classification. The mean length of post-other pauses (poplen) was calculated per speaker as the mean length of all gaps. Words per utterance (wpu) was calculated as the mean number of words per IPU.

2.2.3. Textual Features

[11] also measured conversation topics and topic persistence. The rate of personal pronoun use (ppron) was calculated by dividing the number of words in a speaker's transcript by the total number of first and second personal pronouns. Following [11], we removed stopwords and filler words and used the Stanford Stanza NLP toolkit [20] to lemmatize remaining tokens. To quantify topic repetition, we calculated the mean number of terms in each IPU that were repeated from the previous IPU (rept). We also calculated the percentage of IPUs which contained a repeated item (repu) by dividing the number of IPUs with a repetition by the total number of IPUs.

2.2.4. *Involvement variables*

Following [11], interaction style was calculated using a principal component analysis (PCA) on the extracted variables, which were scaled to zero mean and unit standard deviation.The initial scree plot suggested that the variance in the data can be explained by 3 principal components with an eigenvalue >1. The first component explained 28% of the variance, while two less important components explained 16% and 9%, respectively. On a theoretical level, the extracted components account for a single underlying theoretical construct: involvement [9, 10, 11]. In accordance with [11] we interpreted involvement to be represented by the first component, explaining 28% of the variance. A speaker's involvement was quantified as the individual coordinates on the first factor [11]. Table 1 shows the loadings for the individual variables on the involvement component.

Table 1: Loadings of variables on the involvement component

rept	repu	wpp	wpu	olap	wps	iv	pv	ppron	boplen
0.46	0.45	0.38	0.33	0.30	0.20	0.07	0.06	-0.18	-0.29

2.3. Measure of entrainment

To measure prosodic entrainment, we calculated the mean f0 for each speaker (mean f0 per IPU, weighted by IPU duration) for the first and last third of the conversation (based on IPU number). Following [21, 22], we took a geometric approach to quantify entrainment: with speaker A's f0 on the x-axis and speaker B's f0 on the y-axis, a "matching line" is drawn on the x=y line to represent a situation in which both speakers have the same mean f0. Two points were drawn to represent both speakers' mean f0 during the first and last third of the conversation. The minimum distances between these points and the matching line were calculated. The difference between these minimum distances represents the overall entrainment. A speaker's contribution to the overall entrainment was calculated by determining the proportion of change that occurred along each axis. This relative indication of contribution to entrainment was converted to an absolute value by multiplying it with the overall entrainment.

2.4. Statistical analysis

To assess the relationship between involvement and entrainment, we used actor-partner interdependent models (APIM), implemented using the Shiny web-app by [23]. APIMs can be used to quantify actor- and partner-effects, where actor effect refers to how one individual's independent variable predicts their own dependent variable, and partner effect refers to how an individual's independent variable predicts their partner's dependent variable. Both members of the dyad have actor and partner effects. Typically, APIM is used to calculate k, the ratio between the actor and partner effects, which can be used to estimate dyadic patterns. For more information on APIM, see [23]. Here, APIM is used to investigate whether a speaker's involvement score correlates with their own absolute entrainment contribution, their interlocutor's absolute entrainment contribution, and their interlocutor's involvement score. To investigate whether a mismatch in interaction style influenced the observed entrainment, a Pearson correlation was run on the overall entrainment and the absolute difference between both speaker's involvement scores. To investigate whether speakers' combined involvement influenced entrainment, a Pearson correlation between the sum of both speaker's involvement scores and the overall entrainment was conducted.

3. RESULTS

Results of the entrainment analyses show that approximately half of the 1295 dyads (664) showed entrainment and in this group, the mean overall observed entrainment was 6.81Hz and the the average absolute entrainment contribution of both speakers was 3.40Hz. In the other 631 conversations, disentrainment was measured, with the mean overall disentrainment being -6.93Hz and the mean absolute contribution of both speakers in these conversations being -3.46Hz. Table 2 shows a summary of the APIM. The variance of the errors is 31.72 and the R squared is 0.001. The partial intraclass correlation for entrainment contribution while controlling for the other predictor variables is 0.47 and is statistically significant ($p<0.001$, 95% CI [0.41, 0.54]). The predicted entrainment contribution when involvement score equals zero (i.e. the intercept)is equal to 0.06 and is not statistically significant ($p=0.68$, 95% CI [-0.21, 0.33]). The actor effect is -0.07 ($p=0.237$, 95% CI [-0.19, 0.04]). The standardised actor effect is -0.02 (partial r=-0.02). The partner effect is -0.03, which is not statistically significant ($p=0.555$, 95% CI [-0.15,0.08]), and its overall standardised effect is -0.01 (partial r=-0.01). Since neither actor nor partner effects are significant, k cannot be estimated adequately, meaning that calculating and

interpreting it will provide no insight regarding the relationship between partner and actor effects.

Table 2: Summary of APIM.

effect	estimate	lower bound CI	upper bound CI	p	beta	r
intercept	0.06	-0.21	0.33	0.68		
actor	-0.07	-0.19	0.04	0.33	-0.02	-0.02
partner	-0.03	-0.15	0.08	0.56	-0.01	-0.01

The Pearson correlation between the speakers' difference in involvement scores and the overall entrainment is not significant (r(1293)=-0.008, p= 0.764), nor is the Pearson correlation between the overall entrainment and the sum of the speakers' involvement scores (r(1293)=-0.032, p=0.256).

4. DISCUSSION

The results of the APIM suggest that there are no significant partner or actor effects, i.e. that an individual's absolute entrainment contribution is not predicted by their own or their conversation partner's involvement score. The two conducted Pearson's correlations suggest no significant correlations between the sum or difference between two speakers' involvement scores and the overall entrainment in their conversation. Taken together, these results suggest that there is no clear link between prosodic entrainment as measured using [21]'s method and interaction style as defined by Tannen [9, 10] and operationalised by [11].

Entrainment was observed in 51.2% of the conversations, whereas disentrainment was found in the other 48.7% of the conversations. The only significant correlation was the intraclass correlation between the two speakers' absolute contribution to the overall entrainment: if one speaker showed a lot of entrainment, the other speaker tended to entrain, too. In conversations where overall entrainment was high, both speakers' contributions were more likely to be high than if the overall entrainment was low.

The proportion of explained variance in the PCA is similar to the analysis reported by [11], though our loadings of variables on the first principal component differ. The differences in the loadings between our analysis and that of [11] could be because [11] analysed involvement in human-machine interaction, while SWB [14] consists of free telephone conversation between people. Different features may play different roles in these different contexts: for example, both pitch and intensity variation both play a large role in [9, 10]'s definition of involvement but may play a smaller role in phone conversations compared to face-to-face interactions, where people can rely on non-verbal cues. In addition, SWB consists of conversations between strangers on the phone. Our findings may not translate to face-to-face interactions or conversations between familiar individuals.

Our results suggest that prosodic entrainment is not associated with interaction style. There are several possible explanations for this. The most straight-forward one is that individual interaction style and prosodic entrainment are simply unrelated. Alternatively, measurement level may have affected the results: while previous studies measured both entrainment and involvement at a conversation level [12, 24], we measured both at an individual level, following [21, 22] and [11]. Possibly, the two phenomena interact differently at different measurement levels.

While we did not find any effect of interaction style on entrainment at the prosodic level, there may be a correlation between interaction style and coordinative behavior at other levels of language such as syntax or lexical choice. It is also possible that "interaction style" is not a stable individual trait, but is context-dependent and can change dynamically. A recent study suggests this: [24] calculated a measure of interaction style based on features extracted from 30 second intervals, and found that individuals exhibited great variation in their interaction styles. Our methods could not capture such dynamic changes.

5. CONCLUSIONS

Our findings suggest that interaction style as as defined by Tannen [9, 10] and operationalised by [11] is not associated with prosodic entrainment as measured with methods by [21, 22]. This could be because the two are simply unrelated, but it could also be that interaction style is not a stable individual trait but changes dynamically throughout a conversation, which our analysis did not account for. Nonetheless, Tannen's [10, 9] definition of involvement as operationalised by [11], is a robust measure of involvement style, which is now replicable over different corpora. Future studies may investigate if involvement relates to entrainment on linguistic levels other than prosody, whether entrainment and involvement dynamically vary together during a conversation, or whether our findings translate to face-to-face interactions or conversations between friends or acquaintances.

6. ACKNOWLEDGEMENTS

This work has been conducted in the framework of the Conversational Brains (COBRA) Marie Sklodowska-Curie Innovative Training Network and

has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 859588.

7. REFERENCES

[1] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.

[2] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[4] ——, "An integrated theory of language production and comprehension," *Behavioral and brain sciences*, vol. 36, no. 4, pp. 329–347, 2013.

[5] H. Giles, N. Coupland, and I. Coupland, "1. accommodation theory: Communication, context, and," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[6] R. Ostrand and E. Chodroff, "Itâs alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue," *Journal of phonetics*, vol. 88, p. 101074, 2021.

[7] A. C. Yu, C. Abrego-Collier, and M. Sonderegger, "Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and âautisticâ traits," *PloS one*, vol. 8, no. 9, p. e74746, 2013.

[8] Y. Lee, L. Goldstein, B. Parrell, and D. Byrd, "Who converges? variation reveals individual speaker adaptability," *Speech Communication*, vol. 131, pp. 23–34, 2021.

[9] D. Tannen *et al.*, *Conversational style: Analyzing talk among friends*. Oxford University Press, 2005.

[10] D. Tannen, "Toward a theory of conversational style: The machine-gun question." 1980.

[11] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark, "Style and alignment in information-seeking conversation," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 2018, pp. 42–51.

[12] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002. [Online]. Available: https://doi.org/10.1177/026192702237953

[13] C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring dynamics of mimicry by means of prosodic cues in conversational speech," in *ICPhS 2011*, Hong-Kong, China, 2011. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01705525

[14] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517â520.

[15] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[16] B. Bigi and B. Priego-Valverde, "Search for Inter-Pausal Units: application to Cheese! corpus," in *9th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, May 2019, pp. 289–293. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02428485

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[18] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[19] C. D. Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *INTERSPEECH*, 2009.

[20] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," 2020.

[21] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic entrainment in conversations of verbal children and teens on the autism spectrum," *Frontiers in Psychology*, vol. 11, p. 2718, 2020.

[22] H. Lehnert-LeHouillier, S. Terrazas, S. Sandoval, and R. Boren, "The relationship between prosodic ability and conversational prosodic entrainment," *Speech prosody (Urbana, Ill.)*, vol. 2020, p. 769, 2020.

[23] L. Stas, D. A. Kenny, A. Mayer, and T. Loeys, "Giving dyadic data analysis away: A user-friendly app for actor–partner interdependence models," *Personal relationships*, vol. 25, no. 1, pp. 103–119, 2018.

[24] N. Ward, "Individual interaction styles: Evidence from a spoken chat corpus," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 27–31.

EXPLORING GLOBAL AND LOCAL ARTICULATION RATE ENTRAINMENT IN TYPICAL AND ATYPICAL SPEAKERS

Lotte Eijk[a)], Stefany Stankova[b)], Sophie Meekings[c)]
Department of Psychology, University of York, United Kingdom
[a)]lotte.eijk@york.ac.uk [b)]ss3345@york.ac.uk [c)]sophie.meekings@york.ac.uk

ABSTRACT

Entrainment, in which talkers adapt their speech to become more similar to their interlocutor's, seems to be a persistent phenomenon among typical speakers. However, it is not clear how this behaviour is affected when one speaker has difficulty controlling their speech production, for example because they stutter. We explore whether typical speakers entrain their Articulation Rate differently to people who stutter than to fluent speakers. Twenty participant pairs (10 typical-typical, 10 typical-atypical) participated over Zoom, completing a solo picture description, and three rounds of a spot-the-differences Diapix task in interaction. Local entrainment was defined as turn-by-turn entrainment. Global entrainment was investigated by comparing the Articulation Rate of speakers of a pair between Diapix rounds, and the solo picture description. Results indicate that speakers entrained locally. No evidence for global entrainment nor differences between the two groups were found. Overall, this suggests speakers entrain locally to both typical and atypical speakers.

Keywords: entrainment; articulation rate; stuttering; interaction.

1. INTRODUCTION

Speakers in conversation tend to change their speech to become more like that of their interlocutor. This phenomenon is, among other things, referred to as entrainment, alignment, or accommodation. Here we will use the term *entrainment* to refer to speakers changing their speech to match their interlocutor's speech more closely. Entrainment has been thoroughly investigated on different linguistic levels in typical speakers (e.g., phonetic entrainment [1], syntactic entrainment [2], lexical entrainment [3]), and can occur over the course of the whole interaction (global entrainment) or turn-by-turn (local entrainment). However, studies on entrainment are less prevalent in atypical populations. It is thus not clear whether speakers entrain to atypical speech to the same extent as they do to typical speech.

Entrainment has been found to relate to a range of positive aspects, such as greater task success [4] or higher likeability [5]. Therefore, it is important to understand whether disabilities that affect speech production also influence conversational entrainment, as this may affect how speakers are perceived, with implications for their general quality of life. As hypothesised by Borrie and Liss [6], impaired entrainment abilities may partially contribute to intelligibility difficulties in conversation for this group of speakers.

This study explores Articulation Rate (AR) entrainment between typical speakers and people with persistent developmental stuttering. Stuttering is a neurological speech condition that causes involuntary syllable repetitions, prolongations, and 'blocks' where speakers are unable to produce sounds. This affects speaking and articulation rate [7, 8]. In addition, people who stutter demonstrate difficulty with speech motor skills, which may affect their ability to adapt their speech to different situations [9]. Stuttering is also associated with negative listener perceptions and influences listener's recall of spoken information [10], meaning that the condition presents several challenges to successful spoken interaction.

In this study, we aimed to investigate whether these difficulties affect Articulation Rate entrainment in conversation. Specifically, we asked whether typical speakers, when partnered with a person who stutters, would adapt their own speech more to compensate for their interlocutor's reduced speech motor control. We also aimed to investigate whether entrainment in these interactions occurs at the local or global level, both, or neither.

Studies on entrainment usually investigate local and/or global entrainment in different manners. Even though definitions of what a turn is can differ between studies, local entrainment is commonly investigated as turn-by-turn entrainment [e.g., 2]. There are a larger range of possible ways to measure global entrainment. Most definitions of global entrainment are, however, similar, where global entrainment is defined as speakers becoming more similar over a conversation. Within this definition there is variety in how this measure is employed. Examples of different measures are comparing one half of a conversation to a different half [11] or comparing real pairs to pairs who did not interact [4]. Entrainment is thus investigated in a variety of ways.

Articulation Rate (AR) entrainment has been investigated in many studies in typical speakers [e.g., 5, 11, 12]. These studies show mixed results. Where

Eijk and colleagues [12] find no evidence for local entrainment of AR (number of syllables divided by phonation time – speaking time excluding pauses), and indications of global entrainment, Levitan and Hirschberg [11] find both local and global entrainment in Speaking Rate (the same as AR, but including pauses), and Schweitzer and Lewandowski [5] find overall divergence in AR (but reversed or weakened depending on mutual likeability). Another study, exploring entrainment in an atypical population, investigated Speaking Rate to atypical populations of speakers with dysarthria [6]. The authors found global Speaking Rate entrainment to both typical and atypical speakers in a sentence reading paradigm. However, speakers entrained more to typical speakers.

To get further insight into entrainment to different atypical populations in a semi-naturalistic setting, the research question was studied in a spot-the-differences task with typical speakers in interaction with speakers who stutter over Zoom, an online meeting application. AR is defined as the number of syllables in an utterance divided by the phonation time (time excluding pauses). An utterance is defined as an inter-pausal unit, separated by at least 0.5s of silence [e.g., 11]. We define local entrainment as turn-by-turn entrainment and global entrainment as similarity between speakers of a pair in their AR over different pictures.

2. METHODS

2.1. Participants

Forty native English participants participated in this study. The participants were paired into two different groups, resulting in ten pairs of typical-typical speakers (age: $M = 31.3$, $SD = 12.6$) and ten pairs of typical-stuttering speakers (age: $M = 32.8$, $SD = 12.2$). Participants who stuttered were self-identified stutterers. The typical-typical pairs consisted of 3 female-male, 6 female-female, and 1 male-male pairs, and the typical-stuttering pairs were 7 female-male, 2 female-female, and 1 male-male pairs.

2.2. Tasks

Speakers completed multiple experimental parts over Zoom. The first day, they completed different tasks by themselves, including a solo picture description. In this task, participants were asked to describe a picture from a spot-the-differences task - the Diapix task [13] - by themselves, starting from the top left corner and continuing clockwise.

On a different day, participants joined over Zoom again to participate in the Diapix task. In this task, participants were asked to turn off their camera to only provide the interlocutor with information via speech. They were furthermore asked to use a separate microphone and headphones. The interactional task had a mean duration of 29.84 min ($SD = 2.90$ min) over all participants.

2.2. Procedure

After doing the solo picture description on the first day, on a different day participants completed three rounds of the Diapix task, and thus described three different pictures to each other (Beach, Farm and Street). Pairs were divided into a participant *A* and a participant *B*. Participant A was the leader in Round 1 and 3, and participant B was the leader in Round 2. In the pairs with a participant who stuttered, this participant was always participant A. Being the leader entailed starting the picture description. The other participant was the follower in that round and was instructed to notify the other participant when something seemed different in their picture and together find the differences. There was a time limit of 10 minutes per picture and each picture contained 12 differences to be identified.

2.3. Pre-processing

2.3.1. Audio pre-processing

Since data was collected online and varied quite drastically in quality, pre-processing was necessary to conduct analyses. This process consisted of six steps. The first step was to use dynamic time warping with a custom Python (version 3.9.12) script to align the separate mono wav files from the participants to be time aligned to a wav file containing both participants' speech. These files were then combined to a stereo file with one channel for each speaker and cut up into separate files per described picture. These were split into mono again to reduce the noise per participant with a custom script in Praat [14] using the "Reduce noise…" function. Next, we used the "To TextGrid (silences)…" function in Praat with individual optimal settings per participant to separate the utterances. Minimum pause duration was set to 0.5s for all participants to divide the audio into inter-pausal units (IPUs; commonly divided by a minimum pause of 0.5s [e.g., 11]). The resulting TextGrid files were manually checked by two trained researchers. Stutters on a word were annotated to be included in the same IPU as the word itself. The last step was to extract all separate IPUs, taking into account the starting time within the file.

2.3.2. Articulation Rate extraction

Articulation Rate (AR) was calculated per IPU with a script [15] in Praat. The standard settings for the script were used (Silence threshold (dB): -25, Minimum dip near peak (dB): 2, and Minimum pause duration (s): 0.3). After extraction of the AR, IPUs with less than 5 syllables were removed to avoid including utterances too short for a reliable AR (due to e.g., filled pauses or phrase final lengthening) in the analyses. This resulted in 6971 IPUs for all participants together (about 58 percent of the total).

2.4. Analyses

2.4.1. Global entrainment

To investigate global entrainment, we calculated a single AR per picture description. This score was calculated by dividing all syllables from the different utterances used to describe a picture by the total phonation time per picture. This resulted in AR values per participant per picture. Within pairs, these scores were then subtracted from each other, after which we took the absolute value to calculate a difference score per participant pair (see (1) below). These scores represent how similar a participant A's AR was to that of participant B in the solo picture description and in the different Diapix rounds. A score closer to 0 means that participants' ARs in a round are more similar.

$$\text{Difference score} = |\text{AR}_\text{A} - \text{AR}_\text{B}|. \quad (1)$$

These difference scores were then used to perform two linear mixed-effects regression (LMER) models in R version 4.2.1 [16], using the lme4 version 1.1.31 [17] and car package version 3.1.1 [18]. Ggplot2 version 3.4.0 [19] was used for visualisation. The first LMER model tested for differences between groups (typical-typical pairs versus typical-stuttering pairs) between the solo picture description and the different Diapix rounds and included a random intercept of participant. The second model was the same, but excluding the interaction with group to test for overall differences between the rounds in the whole group.

2.4.1. Local entrainment

After deletion of utterances with less than 5 syllables, utterances of participant B that were preceded by an utterance of participant A were identified. This was counted as consecutive turns and used in the statistical models. This led to a total of 1297 data points for the statistical analyses of local entrainment.

We tested two LMER models in R version 4.2.1 [16]. In the first model, we tried to predict the AR of participant B with the AR of participant A, in interaction with group (typical-typical versus typical-stuttering). This model included a random intercept of participant. The second model was the same model, excluding the interaction with group to test for an overall effect within both groups together.

3. RESULTS

Figure 1 shows the difference scores per participant pair per round of the experiment. Descriptively, the rounds seem to differ from each other. Figure 2 shows the same data points, but now connected with a line for each pair. This shows that the entrainment patterns considerably differ between the different pairs.

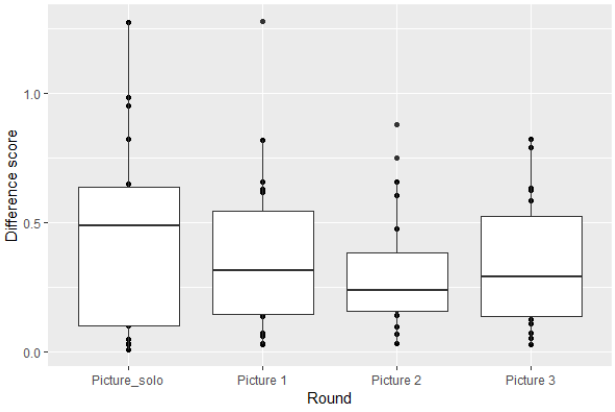


Figure 1: Difference scores between participants' AR in different parts of the experiment.

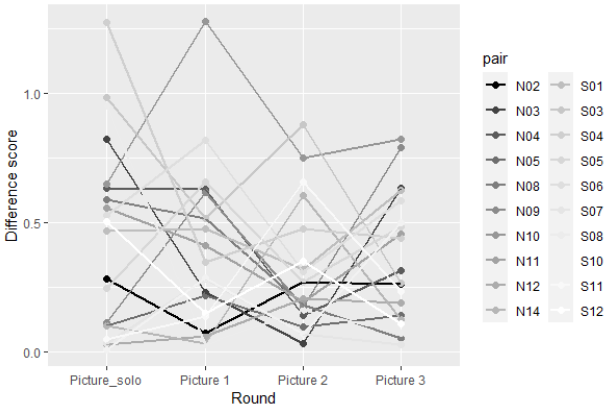


Figure 2: Difference scores between participants' AR in different parts of the experiment. Lines connect each pair. N = typical-typical pair, S = typical-stuttering pair.

3.1. Global entrainment

We performed an LMER testing for differences between the groups between the different rounds (Solo Picture, Diapix Round 1, Diapix Round 2, and Diapix Round 3). These analyses showed no significant effects, as indicated by running an ANOVA from the car package on this model. We then tested the same model, but taking out the interaction with group to test for global entrainment

in the whole group. This showed no significant difference either. This is not too surprising given the diversity of patterns shown in Figure 2.

3.2. Local entrainment

Results of the first LMER predicting the Articulation Rate (AR) of participant B with that of participant A and checking for group effects showed no significant effects. We thus do not find evidence for differences in local entrainment between the two groups.

We then tested a similar model, without the interaction. This model showed a significant effect of the AR of participant A on the AR of participant B (see Table 1). This indicates that over the total group of twenty pairs, we find evidence of local entrainment.

Variable	β	SE	T Value
Intercept	3.55	0.12	29.50
AR_A	0.07	0.03	2.46

Table 1: Output from the local entrainment model.

4. DISCUSSION AND CONCLUSION

This study investigated whether typical speakers entrained more to atypical speakers than to typical speakers, and whether entrainment in Articulation Rate (AR) was reflected in local or global measures. We studied AR in a solo picture description task and in speakers in interaction over Zoom performing three rounds of a spot-the-differences task.

No evidence was found to suggest that speakers entrain their AR differently to speakers who stutter as compared to typical speakers. No entrainment was found globally. However, over the whole group, we found evidence for local entrainment, indicating that speakers changed their AR to more closely match the AR of their interlocutor in the turn prior to theirs.

These findings are in line with Levitan and Hirschberg [11] who found local entrainment of speaking rate. However, the authors also found global entrainment (like [6, 12]). One may assume that speakers must also entrain globally when local entrainment takes place. However, the measure used in this study may have not been optimal to detect global entrainment as found in other studies. Here, we used one AR over a whole picture to get an indication of possible global entrainment. Other ways of measuring global entrainment such as the measurement by Levitan and Hirschberg, who measured the degree to which Speaking Rate became more similar over time between two speakers, may shed more light on global entrainment. Global entrainment will thus be further explored in future studies using the same dataset. Nevertheless, since there is no standard way of measuring global entrainment, this study shows a first insight into this measure in a population of speakers who stutter.

Another reason why our results may differ partly from the literature, and why findings in the literature may be rather divergent, is the possible effect of the task. Since entrainment has been shown to relate to e.g., task success [4] or social factors [11], it is likely that the kind of task influences the outcome.

Moreover, different studies may investigate slightly different measures. As mentioned before, some of the studies cited here investigate Speaking Rate [e.g., 11] instead of Articulation Rate. Another difference in the measurement of both Articulation Rate and Speaking Rate is that one can either use the actually realised syllables found in the speech signal (excluding reduced syllables) or base the rate on the number of syllables that the canonical words contain. The possible differences and the possible effect of these different measures of speech tempo on entrainment patterns should be further investigated to gain insight into what speakers actually entrain to.

Different pairs of speakers showed various entrainment patterns. Stuttering is a heterogeneous disorder [20], and the amount of stuttering in an affected person's speech can vary widely depending on the situation [21, 22]. Differences in overall stuttering severity as well as variation in the amount of stuttering may have affected a pair's ability to entrain; additionally, the online nature of the experiment may have introduced additional challenges to entrainment such as environmental noise or connection problems.

In conclusion, this study advances our understanding of how speech patterns align during communication, particularly for individuals who stutter. A key finding is that local entrainment of articulation rate occurs in both typical-typical and typical-stuttering pairs. This suggests that stuttering may not affect rhythmic conversational entrainment when people who stutter converse with typical talkers. Local entrainment was observed despite the relatively poor audio quality of Zoom compared to in-person conversation. This study thus shows the potential of using Zoom data to test populations that may be difficult to assess in the lab, even though these data pose audio quality challenges. Although no significant global entrainment was found, the individual variation in entrainment patterns suggests that factors such as stuttering severity may play a role in how pairs align to each other. These findings could have implications for how clinicians and researchers assess and treat individuals with stuttering.

6. REFERENCES

[1] Pardo, J.S., 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), pp.2382-2393.

[2] Branigan, H.P., Pickering, M.J. and Cleland, A.A., 2000. Syntactic co-ordination in dialogue. *Cognition, 75*(2), pp.B13-B25.

[3] Brennan, S.E. and Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition, 22*(6), p.1482.

[4] Reitter, D., and Moore, J. D., 2007. Predicting success in dialogue. In A. Zaenen & A. van den Bosch (Eds.), *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 808–815). Prague, Czech Republic: Association for Computational Linguistics.

[5] Schweitzer, A. and Lewandowski, N., 2013. Convergence of articulation rate in spontaneous speech. In *INTERSPEECH* (pp. 525-529).

[6] Borrie, S.A. and Liss, J.M., 2014. Rhythm as a coordinating device: Entrainment with disordered speech. *Journal of Speech, Language, and Hearing Research, 57*(3), pp.815-824.

[7] Hall, K., Amir, O., and Yairi, E. 1999. A longitudinal investigation of speaking rate in preschool children who stutter. *Journal of Speech, Language, and Hearing Research, 42*, 1367–1378.

[8] de Andrade C.R., Cervone L.M., Sassi F.C., 2003. Relationship between the stuttering severity index and speech rate. *Sao Paulo Med J. 121*(2):81-4. doi: 10.1590/s1516-31802003000200010.

[9] Namasivayam, A.K. and van Lieshout, P., 2011. Speech Motor Skill and Stuttering. *Journal of Motor Behavior*, *43*(6), 477-489.

[10] Healey, E., 2010. What the Literature Tells Us about Listeners' Reactions to Stuttering: Implications for the Clinical Management of Stuttering. *Seminars in Speech and Language, 31*(04), 227–235.

[11] Levitan, R. and Hirschberg, J.B., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR.

[12] Eijk, L., Ernestus, M., Schriefers, H., 2019. Alignment of pitch and articulation rate. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia (2019), pp. 2690-2694.

[13] Baker, R. and Hazan, V., 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods, 43*(3), pp.761-770.

[14] Boersma, P. and Weenink, D., 2022. Praat: Doing phonetics by computer (Version 6.2.23).

[15] De Jong, N.H., Pacilly, J., & Heeren, W., 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically, *Assessment in Education: Principles, Policy & Practice,* 28:4, 456-476, DOI: 10.1080/0969594X.2021.1951162

[16] R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

[18] Fox, J. and Weisberg, S., 2019. An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[19] Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

[20] Stager, S. V., 1990. Heterogeneity in stuttering. *Journal of Fluency Disorders, 15*(1), 9–19. doi:10.1016/0094-730x(90)90028-q

[21] Schwartz, H. D., & Conture, E. G. 1988. Subgrouping Young Stutterers. *Journal of Speech Language and Hearing Research, 31*(1), 62.

[22] Bloodstein, O., 1950. Hypothetical conditions under which stuttering is reduced or absent. *Journal of Speech & Hearing Disorders, 15*, 142–153.

TSCR: A comprehensive coding system for task-oriented vocal interaction analysis

Brandon Copping[1], Elizabeth Holzmeyer[1], Santosh Kumar[1], Deniz S. Ones[2], Eugene H. Buder[1]

[1]University of Memphis, Memphis, TN, United States
[2]University of Minnesota, Minneapolis, MN, United States
bcopping@memphis.edu

ABSTRACT

A coding system is developed for the assessment of vocal coordination during task-oriented discourse to systematically assess synchronization and alignment phenomena at multiple levels concurrently. The TSCR implementation includes auditory-spectrographic inspection for Talk unit identifications, Syllable segmentations, Conversational event labels, and Respiratory kinematic signal analysis. TSCR is named from these levels of analysis: talk, syllable, conversation, and respiration. The work-based discourse goals for this corpus required interdependency, and the study incorporated confederates as members of the same-sex dyads who either facilitated or impeded productivity. Participants met each confederate once. Each 30-minute interaction was coded for 3-minute intervals at the beginning, middle, and end. The highest interactivity segments were identified by the product of participants' talk onset frequencies. Alignment results derived from coding of these segments may be interpreted in terms of task efficiency. More broadly, the coding scheme is designed to identify and evaluate physical entrainment dynamics in communicative interactions.

Keywords: Conversation, Turn Exchange, Syllable, Respiration, Entrainment

1. INTRODUCTION

Conversations require every participant to actively listen, predict, and respond to one another in real time. This complicated maneuver requires ongoing coordination between participants, most clearly seen in the turn exchange. The concept of turn-holding can be ambiguous [8] but relatively easy to operationalize dyadically as involving transitions between partners' talk episodes. Numerous past studies have shown that people have remarkable ability to predict when it is they may begin speaking so that they will not overlap too much as to seem rude or interruptive but also not be after too much of a gap as to create an awkward silence [10], [11], [16]. However, how this is performed and what the basis for this synchronization ability is in speech remains a matter of ongoing debate.

There have been many different approaches to analyzing and understanding conversational interaction focusing on turn-exchange dynamics. Sacks et al.'s [18] seminal framework described rules initiated at 'projected transition-relevance places' specifying whom may continue or defer to take a turn, implying physical units defining such places and their projection, but also the units of silence that a turn-holder may use to defer. Similarly, Couper-Kuhlen's [6] groundbreaking discourse analyses distinguished between turn-exchange pauses that were of standard brevity, relative to surrounding speech rhythms, from those that were markedly longer for conspicuous discourse purposes. Meanwhile, Stivers et al. [20] identified turn-exchange pauses cross-culturally as averaging on the order of syllable durations but most typically at 0 ms. This suggests that projection by such units facilitates turn-exchanges, though there is also evidence that other longer prosodic units may be involved [3]–[5] along with respiratory dynamics [12]. Wilson & Wilson's coupled oscillator model for turn exchange [21] provides a contemporary synchronization approach that is useful for modeling such phenomena at multiple levels, indicating that studies might assess whether respiratory versus syllable-level alignments are more significant. For progress to be made on such questions, multiple level descriptions and analyses need to be conducted closely parallel to one another.

The present work introduces a novel, multi-layer coding system for human coding of conversational interactions at three levels, talk, syllable, and conversation, in a corpus that also includes respiratory kinematics. The talk level is the most inclusive, marking any vocal sounds that could have communicative effect and determining the units on which coders at other levels operate. Syllable coders designate the boundaries of all sufficiently articulated vocalizations, accommodating the reductions and hesitations typical of spontaneous interaction to arrive whenever possible at a face-valid count with duration patterning that matches perceived speech rhythms. At the conversational level, which also builds on the scaffolding of talk units, vocalizations are coded for concrete aspects of participants' interactivity. This is done primarily in terms of vocal onsets and offsets relative to the perceived turn-holder, including classification of turn-exchanges as

with or without gaps, or with overlapping between yielder and taker, and noting of simultaneities and interruptive behaviors.

2. METHODS

2.1. Participants and data collection procedure

Recordings included came from 16 sessions: eight same-sex pairs, four male and four female between the ages of 20 and 60, consisting of one participant ("P1") and one or the other trained confederate ("P2"). All were native speakers of American English with a regional, urban dialect reflective of their professional status. Pairs, who had never met, completed a Merit Bonus task: working together on performance evaluations of a set of simulated employee profiles to evaluate and determine an "overall merit score" affecting the employees' year-end salary bonuses [17]. They were told they could accomplish this task in any way they desired and that they would be receiving a compensation bonus depending on how many evaluations were completed across their two sessions combined. These 16 sessions come from a previously recorded corpus of 50 sessions encompassing 25 different participants interacting with the different confederates with these 16 being chosen due to lack of clear recording anomalies or other technical issues with any of their audio or respiratory signals.

In the first session, the facilitating confederate was trained to be *constructive* and helped participants try to complete as many as possible in the 30 minutes allotted. In the second session, their confederate was trained to be *obstructive* and impede productivity. In total there were four confederates, but due to the sessions being sex-matched participants only ever interacted with two of them. Productivity was always higher in the first session, even though being first it required more task orientation. From the point of view of 'complex interdependence,' these circumstances simulate very demanding levels of workplace communication proficiency. After completing the second session, participants were debriefed regarding use of confederates in the design, and they received full compensation regardless of actual productivity in the study.

Conversation partners were seated in a comfortably furnished recording suite with materials on a worktable in between and without any acoustic separation between them. Channels of current interest were acquired by over-the-ear headset microphones (Countryman Associates E6) for full audio, contact microphones adhered to the throat wall for speaker-specific audio (incorporating an accelerometer assembly obtained from PentaxMedical), and respiratory inductance plethysmographic bands [9] for thoracic and abdominal kinesiology (Ambulatory Monitoring Inductotrace systems). Audio from both participants was picked up by the over-the-ear microphones while the contact microphones only picked up signals from the person to whom they were adhered. Audio signals were coded in the AACT coding environment [7], which implements features of the TF32 acoustic analysis program [13].

2.2. Coding procedures

Before any coding was initiated, sessions were split into three segments: the first, middle, and last three minutes of the recording. All segments underwent three rounds of coding. Two of these rounds involved coding only one person at a time, *talk* and *syllable*, while *conversation* coding involved judgments of the interactions within the pairs. *Talk* coding was always performed first as both *syllable* and *conversation* coding used *talk* codes as scaffolding for where work needed to be done. Coding review sessions were held consistently during the process where coders could bring up questions they had during coding to the group and consensus developed as necessary. Statistical reliability assessments were conducted at the conclusion of coding at which point one main coder each remained for the different coding types. Intercoder reliabilities were calculated between these main coders and the same work coded previously by different coders while intracoder reliability was assessed between these remaining main coders and their own previous work. For each type of coding, five one-minute segments were used from the entire corpus to assess reliability.

2.2.1. Talk Coding

The main goal of *talk* coding was to label all sounds within each session partners used communicatively. What could be considered communicative was left purposefully broad to include sounds such as tongue clicks, audible inhales and exhales, and lip smacks, but only if they were deemed by the coder to be intentional, directed, and able to be perceived by the opposing conversational partner. These gestures are being analyzed using the respiratory alignment procedures outlined later in this paper. If such gestures were deemed too minute as to be detected by the other person, they were not coded. Later *syllable* coding was then more restrictive regarding articulated speech.

For segmenting talk units, consistent duration criteria were applied to unarticulated silences: if the silence was less than 200 ms, no gap between codes was allowed and if over 300 ms, a gap between codes was required, but if the silence was between 200 and 300 ms, coders were allowed to consider the salience

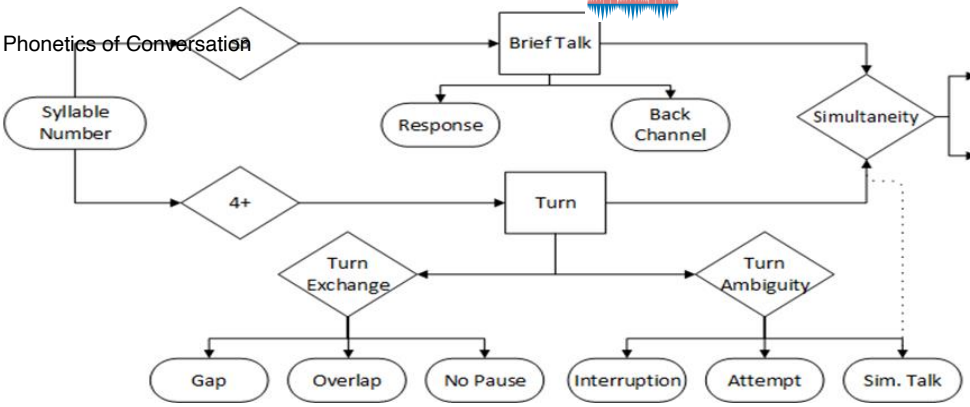


Figure 1: A flowchart of potential *conversation* coding decisions starting with syllable number.

of that gap in their decision, based on such considerations as surrounding speech rate and linguistic discontinuities.

To incorporate laughter, *laugh* and *laughed* codes were available. A talk code could include no more than one syllable of laughter. Ongoing full laughter was coded as *laugh*, and talk that was spoken on laughter was coded as *laughed*; occasionally single talk units were segmented into several such codes.

Reliability was assessed using relative agreement between code onsets with a tolerance of +/- 50 ms. Across the segments used, intercoder reliability ranged between a kappa of κ=0.97 and 0.99 (mean of 0.98) while intracoder ranged between 0.9 and 0.99 (mean of 0.95).

2.2.2. Syllable coding

Similarly to *talk* coding, systematic guidance for *syllable* coding, e.g., lexical or phonological definitions, don't accommodate articulatory rhythms or the variety of spontaneous productions. The criteria applied here were more broadly based on maximal onset and sonority principles, and allowed for extensive reductions, hesitation phenomena (e.g. glottal stopping or fry) or paralinguistic utterances (e.g. laughter, or non-lexical floor-holding 'tunes'): syllable boundaries were coded whenever a distinct rhythmic unit could be both heard and identified spectrographically—the latter was especially important for identifying evidence for reduced articulations, sometimes leading to rejection of a percept that fulfilled lexical expectations but appeared illusory on closer inspection. No specific restrictions were placed on syllable durations; syllables on the order of 50 ms or less occurred, though rarely less (e.g., a syllabic /n/), and drawn-out articulations in the corpus sometimes resulted in syllable durations exceeding 1 s.

Anticipating acoustic modeling of the signal our approach also incorporated "landmark" criteria, for example formant frequency and amplitude inflections and obstruent energy onsets [19]. Training assessments reveal very high agreement rates (over 90%).

Reliability was assessed using relative agreement between code onsets with a tolerance of +/- 50 ms. Across the segments used, intercoder reliability ranged between a kappa of κ=0.92 and 0.99 (mean of 0.95) while intracoder ranged between 0.97 and 0.99 (mean of 0.99).

2.2.3. Conversation coding

Focusing on both individual and dyadic patterns of vocalization, this coding involved the highest number of decisions and used previously coded talk codes as its scaffolding. The primary focus is on turn exchanges and the labeling of brief non-turn utterances, defined as utterances with 3 or fewer heard syllables. Fig. 1 provides a listing of all codes displayed as a decision flow-chart.

Turn exchange codes are based on the timing of the exchange. No code is needed when there is a salient gap between turns, a simple *overlap* code is placed when their turns overlap (and no one has shown interruption), and a *no pause* code is placed when it sounds like there is neither a gap nor an overlap between the two speakers. Signal criteria are consulted, e.g., a tolerance of +/- 300 ms constrains 'no pause' decisions.

Overlapping of speech does not indicate interruption, and to avoid construals of intent, interruption codes were only used when a speaker's turn-holding articulations were cut short along with lexical discontinuity after the partner began speaking concurrently; the speaker who stopped is coded as *interrupted*. If overlapping speech begins concurrently with the partner's onset with a gap having occurred and then is cut short, this is assumed to have been a turn *attempt*. Overlapping speech may also be coded as *simultaneous talking* when a partner joins in during the other's speech, both contributing turns concurrently without either showing articulatory evidence of being stopped by the other.

Utterances with three or fewer syllables spoken by the partner while the other is taking their turn are coded as either *backchannels* or *responses*. A *backchannel* affirms that the partner has the turn, while a *response*, following some question posed by the partner, merely provides information without taking a full turn.

For synchronization study purposes, any audibly *simultaneous onsets* or *offsets* of partners' speech were coded, and two more codes were developed to

reflect the study's task orientation. An *extended silence* was marked whenever interactivity ceased for more than 3 seconds (partners frequently identified independent tasks), and *self-talk* was marked when both manner of speaking and lack of interactivity clearly indicated that the partner was simply 'thinking aloud' during an independent task.

Since conversation coding used talk coding as a scaffold for timing purposes, reliability was instead assessed using agreement on how the interactions were labeled using the various possible codes. Across the segments used, intercoder agreement ranged from 82-96% while intracoder agreement ranged from 87-99%.

2.2.4. Respiratory signals

Thoracic and abdominal wall signals were summed with 2/3 weighting on the thorax [1] and then analyzed in MATLAB for cross-correlations, along the lines pursued by McFarland [12] to assess phase alignments relative to conversational interactions. These cross-correlations were performed using a 10 s sliding window every 5 s.

3. RESULTS

Primary analyses planned for the current corpus of codes and respiratory signals will focus on 25-30 s extracts from each of the three segments of each of the 16 conversations to yield 48 sets of observations, with the goal of comparing partner alignments in constructive versus obstructive sessions. These extracts are selected by a simple algorithm counting the number of turn onsets in each 10 s window, taking the product across participants, and finding the maximally scored interval within each segment. Alignments will be represented primarily in terms of phase-relations between syllable- and respiratory-related units.

To assess the scheme's utility, each of the 48 samples is under examination for interpersonal synchronization phenomena focusing on turn-exchanges, brief talk timing, synchronous behaviors, turn ambiguity, etc. At the most global level, the nature of the turn-exchange flow can be visualized through the respiratory cross-correlation functions. For example, during a task where participants are working simultaneously on a task but not actively exchanging much novel information, respiratory alignment tends to be more in-phase, as can be seen in Fig. 2a This synchrony can rapidly switch to anti-phase on the order of a few seconds when both interlocutors return to exchanging novel information and must therefore alternate with each other, as can be seen in Fig. 2b Since these are conversation samples,

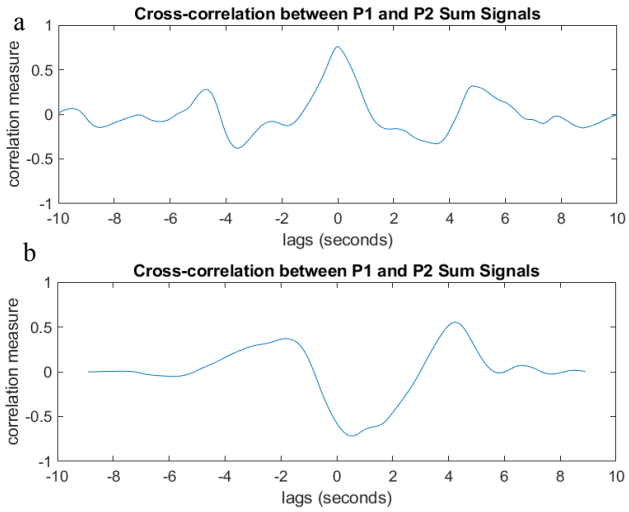


Figure 2a. Respiratory cross-correlation of the first 10 s of a 25 s interactive sample between two female participants showing in-phase alignment.
b. Last 10 s of the same sample as in a, showing that a shift occurred to anti-phase alignment.

this phase switching tends to line up with similar alignments in *talk* codes as well as, during the interactive segments used for analysis, most of their expiratory time is spent speaking. Syllable durations also tend to align during these handoffs with both mono- and di-syllable timing units appearing to be involved. For example, during the final 10 s of the interactive segment referenced in Fig. 2, there is a point where P1 produces 2 ~150 ms syllables simultaneously with 1 ~300 ms syllable from P2, followed by a ~300 ms syllable from P1, and then a similarly-sized syllable from P2 with a "no pause exchange" coded between the final two by conversation coders. This transition was not labeled as an interruption or attempted interruption by any conversation coder and so was judged to be smooth. These multiple levels taken together are what are able to provide more comprehensive insight into conversation dynamics and its relationship to phase dynamics and synchronicity between conversational partners.

4. DISCUSSION

The approach developed here identifies alignments at multiple levels of analysis in task-related discourse. With the present corpus, phase alignments will be examined in relation to task productivity. Furthermore, this framework potentially integrates numerous prior approaches to the phonetics of conversation [15], [22], [23] and accords with coupled-oscillator models [14], [21]. Respiration mobile-sensing models could also be implemented based on these lab data to detect performance-related conversational dynamics remotely [2].

5. ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.
The authors also thank students and staff at the MD2K Center and the Social Interaction Lab at the University of Memphis for data collection.

6. REFERENCES

[1] R. B. Banzett, S. T. Mahan, D. M. Garner, A. Brughera, and S. H. Loring, “A simple and reliable method to calibrate respiratory magnetometers and Respitrace.”

[2] R. Bari, R. J. Adams, Md. M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, “rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–27, Mar. 2018.

[3] E. H. Buder and J. L. Edrington, “Conversational prosodic interactivity when one partner has aphasia,” in *Speech Prosody 2008 Conference*, P. A. Barbosa, S. Madureira, and C. Reis, Eds. Campinas, Brazil: Editora RG/CNPq, 2008, pp. 501–504.

[4] E. H. Buder and A. Eriksson, “Time-series analysis of conversational prosody for the identification of rhythmic units,” in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 2, A. C. Baile, Ed. San Francisco, CA: University of California, Berkley, 1999, pp. 1071–1074.

[5] E. H. Buder and A. Eriksson, “Prosodic cycles and interpersonal synchrony in American English and Swedish,” in *Eurospeech 1997 Proceedings*, vol. 1, E. Dermatas, Ed. Grenoble, France: European Speech Communication Association, 1997, pp. 235–238.

[6] E. Couper-Kuhlen, *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins, 1993.

[7] R. E. Delgado, “AACT - Action Analysis Coding and Training Software.” Intelligent Hearing Systems Corp., Miami, FL, 2020.

[8] C. Edelsky, “Who’s got the floor?,” *Lang. Soc.*, vol. 10, pp. 383–421, 1981.

[9] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al’Absi, and S. Shah, “AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field,” in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, Seattle Washington, 2011, pp. 274–287.

[10] S. Garrod and M. J. Pickering, “The use of content and timing to predict turn transitions,” *Front. Psychol.*, vol. 6, pp. 27–38, 2015.

[11] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *J. Phon.*, vol. 38, pp. 555–568, 2010.

[12] D. H. McFarland, “Respiratory Markers of Conversational Interaction,” *J. Speech Lang. Hear. Res.*, vol. 44, no. 1, pp. 128–143, Feb. 2001.

[13] P. Milenkovic, “TF32.” University of Wisconsin-Madison, Madison, WI, 2018.

[14] M. O’Dell, M. Lennes, and T. Nieminen, “Hierarchical levels of rhythm in conversational speech,” in *Proceedings of the Speech Prosody 2008 Conference*, Campinas, Brazil, 2008, pp. 355–358.

[15] R. Ogden and S. Hawkins, “Entrainment as a basis for co-ordinated actions in speech,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, 2015, vol. Paper number 0599, pp. 1–5.

[16] J. P. de Ruiter, Holger. Mitterer, and N. J. Enfield, “Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation,” *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[17] R. Saavedra, P. C. Barley, and L. V. Dyne, “Complex Interdependence in Task-Performing Groups,” *J. Appl. Psychol.*, vol. 78, no. 1, pp. 61–72, 1993.

[18] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, pp. 696–735, 1974.

[19] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distincitve features,” *J. Acoust. Soc. Am.*, vol. 111, pp. 1872–1891, 2002.

[20] T. Stivers *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 26, pp. 10587–10592, Jun. 2009.

[21] M. Wilson and T. P. Wilson, “An oscillator model of the timing of turn-taking,” *Psychon. Bull. Rev.*, vol. 12, no. 6, pp. 957–968, 2005.

[22] M. Włodarczak and M. Heldner, “Respiratory Turn-Taking Cues,” in *Interspeech 2016*, 2016, pp. 1275–1279.

[23] M. Włodarczak, J. Sˇimko, and P. Wagner, “Syllable boundary effect: temporal entrainment in overlapped speech,” in *Proceedings of Speech Prosody 2012*, Shanghai, China, 2012, pp. 611–614.

DO HUMANS CONVERGE PHONETICALLY WHEN TALKING TO A ROBOT?

Tom Offrede[1], Chinmaya Mishra[2], Gabriel Skantze[2,3], Susanne Fuchs[4], Christine Mooshammer[1]

[1]Humboldt-Universität zu Berlin, [2]Furhat Robotics,
[3]KTH Royal Institute of Technology, [4]Leibniz-Centre General Linguistics (ZAS)
offredet@hu-berlin.de, chinmaya@furhatrobotics.com, skantze@kth.se, fuchs@leibniz-zas.de, mooshamc@hu-berlin.de

ABSTRACT

Phonetic convergence—i.e., adapting one's speech towards that of an interlocutor—has been shown to occur in human-human conversations as well as human-machine interactions. Here, we investigate the hypothesis that human-to-robot convergence is influenced by the human's perception of the robot and by the conversation's topic. We conducted a within-subjects experiment in which 33 participants interacted with two robots differing in their eye gaze behavior—one looked constantly at the participant; the other produced gaze aversions, similarly to a human's behavior. Additionally, the robot asked questions with increasing intimacy levels.

We observed that the speakers tended to converge on F0 to the robots. However, this convergence to the robots was not modulated by how the speakers perceived them or by the topic's intimacy. Interestingly, speakers produced lower F0 means when talking about more intimate topics. We discuss these findings in terms of current theories of conversational convergence.

Keywords: phonetic convergence; fundamental frequency; human-robot interaction; intimacy

1. INTRODUCTION

Humans having a conversation are known to change their speech depending on how their partner is speaking. Among other terms, this phenomenon is called convergence [1]. On a phonetic level, for example, it has been shown that speakers alter their speech rate, intensity, and shimmer [2] in the direction of their interlocutor's speech acoustics. The current study is concerned with convergence on the fundamental frequency (F0) level, as an initial feature to be investigated. Phonetic convergence has been observed not only in interactions between two humans, but also between a human and a voice AI system [3, 4], and, preliminarily, between a human and a robot [5]. Importantly, phonetic convergence tends to be a variable and often subtle effect [1].

Various theories propose different explanations for this phenomenon. For instance, the Interactive Alignment Model argues that a priming mechanism causes speakers to entrain to each other [6]. On the other hand, [7]'s Communication Accommodation Theory (CAT) argues that convergence (or accommodation) depends on interindividual and intergroup factors, and that it can be used to manage the social distance between two interlocutors. For example, the speakers' evaluations of each other's personal history and group-belonging dynamics may affect their (non-)accommodating behavior.

In addition to social factors, situational elements may also influence speech overall and phonetic convergence. [8, 9] have suggested that the topic being discussed may influence convergence. One example could be the emotional content of the conversation: more intimate themes, especially in interaction with the interlocutors' perceptions of each other, might influence their speech behavior. Further, [10] have demonstrated that speakers tended to converge to their interlocutor on a few acoustic features when they felt engaged in the conversational task, but not when the task was not engaging. Interestingly, these speakers still adapted their speech according to the task, independently of convergence to their partner.

One important communicative cue during conversation is eye gaze—specifically, gaze aversion (GA), i.e., the breaking of eye contact between interlocutors. Different functions have been attributed to GA, such as management of cognitive load [11], regulation of turn-taking [12], and modulation of intimacy expressions [13]. Since this behavior is ubiquitous in human-human interaction, we hypothesize that gaze behavior may also have an impact on human-to-robot convergence. Specifically, we expect that a robot that produces GA more or less similarly to a human, in comparison to a robot that looks continuously at its human interlocutor, would be perceived more positively. This, in turn, might affect the human's phonetic (F0) convergence behavior.

Hence, in this study, we investigated (1) whether humans converged on F0 to a robotic interlocutor, and whether (2) social factors—the participants' perception of the robot—and (3) the conversation topic—intimacy of the questions asked by the robots—would modulate such convergence.

2. METHOD

Each participant interacted with two Furhat robots separately under two conditions in a within-subjects design. One robot looked constantly at the participant (*Fixed Gaze* condition). The other one averted its gaze away from its interlocutor every few seconds [14], similarly to what a human would do (*Gaze Aversion* condition). This GA behavior respected known gaze cues related to conversational floor management: for example, at the end of the robot's turn, it always made eye contact with the participant, as humans usually do. We recorded the participants' speech, eye movement behavior (data discussed elsewhere) and ratings of their perception of the robots and conversations.

2.1. Participants

Thirty-three participants assigned male at birth participated in the study. Their ages ranged from 21 to 56 (M = 30.55; SD = 8.07). Five participants spoke English as a first language (L1)—the language in which the study took place—while the others' L1s were one or two of 16 different languages. According to their LexTALE scores and Lemhöfer & Broersma's classification [15], 16 of these participants would be categorized at the C1 to C2 level, 15 at B2, and two at B1 [16]. We obtained similar results on F0 behavior regardless of whether the dataset included the participants with lower English proficiency.

All participants provided written consent and received a 100 SEK-voucher as compensation.

2.2. Procedure and Materials

Before each conversation with the robot, the eye-tracker was calibrated. Then, the first interaction began. The robot started out introducing itself and explaining that they would have a conversation together. It then proceeded to ask the participant six questions, which the participant answered freely. After the participant's answer, the robot would also answer its own question before moving on to the next one. Each participant interacted with two robots (each differing in eye gaze behavior; counterbalanced in order), and after each interaction, the participant responded to a questionnaire that measured their perception of the robot and the conversation. Between the two conversations, they also filled in [17]'s version of the Big Five personality inventory and the LexTALE, a test that measures lexical knowledge in English and that correlates with general proficiency [15].

The questionnaire the participants filled in after each interaction contained items about their perception of the robot and of the conversation itself, which they rated on a 1–9 Likert scale. Since we observed moderate to strong correlation in the answers to multiple questions, we conducted a Principal Components Analysis using the R package *parameters* [18] to reduce dimensionality and identify patterns in the data. Two principal components were determined. *Conversational Quality* included items such as "My conversation with the robot flowed well" and "I was able to understand when the robot wanted me to speak." The component *Evaluation of Robot* included items such as "The robot responded to me at the appropriate time" and "The robot's behavior was very human-like."

The questions asked by the robots during the interaction always went from less (e.g., *What did you have for breakfast this morning?*) to more intimate (e.g., *For what in your life do you feel most grateful?*). To determine their intimacy level, 28 questions were taken from [19, 20] and rated on a 1–9 Likert scale by residents of the city where the data were collected (N = 130). The researchers then selected 12 questions with increasing intimacy ratings and divided them into two sets with similar intimacy distributions. Each robot asked one set of questions in a counterbalanced fashion across conditions.

2.3. Feature Extraction and Analysis

Using Praat [21], we automatically extracted voiced and voiceless portions during speech using autocorrelation. In a next step, we defined pauses as unvoiced portions with at least 350 ms duration. This threshold was determined through visual inspection. Unvoiced portions of 100, 150, 200, and 300 ms had also been considered; however, these periods often corresponded to unvoiced segments or consonant clusters within the speech stream. The speech streams between pauses are called Interpausal Units (IPU). From these, we extracted F0 information of the participant's and robot's audio streams, obtaining the mean F0 in each IPU.

For all statistical analyses, we used the R package *lmerTest* [22] to fit linear mixed-effects models.

All the models contained random intercepts for participants. All the continuous predictors were centered around their mean value in each given conversation. We consider p values below 0.05 as indicating a statistically significant effect. F0 data are given in Hertz, intimacy data are the mean ratings of the questions (1 to 9), and intensity data (see below) are given in decibels.

3. RESULTS

3.1. Do humans converge on F0 to a robotic interlocutor?

To investigate whether the participants displayed overall convergence to the robot, we fit a mixed-effects model with F0 mean as the dependent variable and the robot's previous F0 mean (i.e., the robot's F0 values in the IPUs of the preceding turn) as a predictor (*human's F0 ~ robot's previous F0*). We considered that convergence happened if the robot's F0 was a significant positive predictor of the human's F0. This method of convergence measure was based on [23]. The model revealed that the robot's F0 mean in the preceding turn was a significant predictor of the human's current F0 mean (see Figure 1 and the model's coefficients on Table 1). This suggests that the participants' F0 tended to converge to their robotic interlocutors. Importantly, there was great interindividual variability, meaning that not all speakers converged to the robot. This drove the overall convergence effect to be small: marginal $R^2 = 0.002$ and conditional $R^2 = 0.58$ [24], as calculated with the R package *MuMIn* [25].

Variables	Estimate	St. Err.	*t* value	*p* value
F0 mean				
Intercept	115.86	3.05	38.02	< 0.001
Robot's F0	0.06	0.02	3.09	0.002

Table 1: Regression coefficients indicating the effect of robot's F0 mean on human's F0 mean.

3.2. Does the participants' perception of the robot modulate convergence?

Next, we investigated whether the participants converged to the robots differently depending on the gaze condition (Fixed Gaze and Gaze Aversion). First, we evaluated whether the participants had different perceptions of the robots and conversations according to the condition. In linear regression models, we included the participants' ratings of the *Conversational Quality* and *Evaluation of Robot* components as the dependent variables and gaze condition as a predictor. Contrary to our predictions, neither of the models revealed a significant difference between conditions in the participants' ratings (Conversational Quality: $p = 0.91$; Evaluation of Robot: $p = 0.19$).

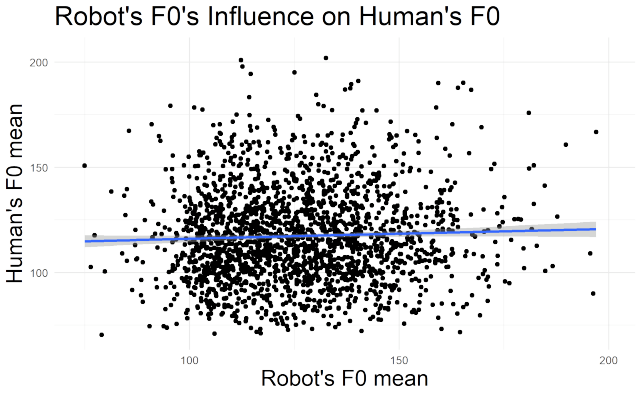


Figure 1: The influence of the robot's F0 mean on the human's F0 mean.

Despite this null result, we fit a regression model to test whether gaze condition influenced F0 convergence, since differences in perception or attitude might not be captured by our previous analysis. We added gaze condition as an interacting predictor to the aforementioned model measuring convergence (*human's F0 mean ~ robot's F0 mean : gaze condition*). This model suggested a difference between the conditions. However, a comparison between the AICs of the models with and without condition as a predictor indicated that the one without it was preferred. We thus conclude that there was no reliable difference in convergence between the conditions. Neither the participants' perception of the robot nor its gaze behavior affected their convergence on F0.

3.3. Does the conversation topic (intimacy of the questions asked by the robots) modulate convergence?

To evaluate whether intimacy of the topic influenced the convergence effect, we also included intimacy value in the model described above (*human's F0 ~ robot's F0 : intimacy*). This model did not output any significant values that might indicate convergence ($p = 0.77$). Thus, the intimacy of the topics being talked about did not influence how much people converged to the robots.

Interestingly, however, increases in the topic's intimacy level were related with a decrease in the F0 mean of the participants' speech. The robot's

gaze condition did not interact significantly with the effect of intimacy on F0.

It could be hypothesized that these decreases in F0 were related to lower intensity: participants might speak less loudly when discussing more personal topics, which would then lead to lower F0 values. Hence, for an additional post hoc analysis, using Praat's default settings we obtained the mean intensity of the IPUs for which we had calculated F0. After a mixed-effects model confirmed that intimacy also negatively predicted intensity, we ran a mediation analysis using the R package *mediation* [26]. F0 mean was the dependent variable, intimacy the independent variable, and speech intensity the mediator. Intimacy's direct effect on F0 when accounting for the mediator was -0.77 ($p < 0.001$, with F0's intercept being 115.14 and the 95% confidence interval of the effect between -1.07 and -0.49). This analysis corroborated intimacy's direct effect on F0.

4. DISCUSSION

Our data suggest that human speakers tend to converge on F0 mean to a robotic interlocutor—albeit to a small degree, as has been shown elsewhere [1]. To the extent of our knowledge, this had been shown to happen to spoken dialogue systems [3, 4], but not systematically to a physical robot and in a conversational setting—i.e., where the participant could produce unscripted speech. Although we have only analyzed F0, we will investigate convergence on other acoustic features in future work.

We attempted to modulate the humans' perception of their robotic interlocutor by having them interact with robots that differed in their eye gaze behavior. However, their ratings of the robots and conversations were not significantly different across conditions. In addition, gaze condition did not reliably influence the participants' convergence to the robot. That is, in our data, factors related to interpersonal dynamics did not seem to modulate convergence, as theories such as CAT [7] would predict. Rather, it could be that the mechanisms behind our participants' convergence behavior are of (nonsocial) cognitive nature. For instance, [6] would argue that this happens through priming: speakers produce speech that matches the acoustics they have been exposed to in the conversation. This entrainment process would then facilitate alignment on mental representations and, in turn, the conversational flow and information sharing in general.

The level of intimacy of the topics discussed also did not interact with the robot's F0 to influence the humans' speech. Interestingly, however, intimacy had a direct impact on the participants' F0 mean. This finding is in line with [10]'s observation that the circumstances of the interaction (in their case, the speakers' engagement with the conversational task) modified individuals' speech.

The impact of intimacy on F0 cannot be fully accounted for in terms of the valence of the emotions evoked. Some of the more intimate questions evoked positive emotions (*For what in your life do you feel most grateful?*) while others evoked negative memories (*What is one of the more embarrassing moments in your life?*).

One possible explanation for the effect of intimacy on F0 is [9]'s Audience Design model. Speakers produce different speech depending on their addressees, accounting for relevant social factors such as level of formality. According to [9], speakers also associate certain topics to specific addressees, causing them to vary their speech when talking about said topics even in the absence of those addressees. This might explain why our participants produced lower F0 when talking about more intimate topics.

It is important to acknowledge that the questions' intimacy is confounded with their order (every conversation went from less to more intimate), which could be the reason for the F0 effect. We argue that this is not related to such a habituation effect to the robots because the F0 reduction was similar for both conversations the participants had. When they started the second interaction, they did not have to get used to the robot as they did to the first. Still, this effect was herein analyzed in an exploratory manner and should be further investigated experimentally.

5. CONCLUSION

This paper has demonstrated that human speakers may converge on F0 to robotic interlocutors, as they do to other humans. Their perception of these interlocutors, however, does not modulate the convergence behavior, as has been shown to happen in human-human interactions. We have also observed that the intimacy level of the topics being discussed had a direct influence on the speakers' F0.

6. ACKNOWLEDGMENTS

We are thankful to Melina Pfundstein for her work on data annotation. COBRA is a European project funded by the European Union's Horizon

2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement n° 859588. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–SFB 1412, 416591334.

7. REFERENCES

[1] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, Perception, & Psychophysics*, vol. 79, pp. 637–659, 2017.

[2] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.

[3] M. Cohn, K. Predeck, M. Sarian, and G. Zellou, "Prosodic alignment toward emotionally expressive speech: Comparing human and alexa model talkers," *Speech Communication*, vol. 135, pp. 66–75, 2021.

[4] I. Gessinger, B. Möbius, S. Le Maguer, E. Raveh, and I. Steiner, "Phonetic accommodation in interaction with a virtual language learning tutor: A wizard-of-oz study," *Journal of Phonetics*, vol. 86, p. 101029, 2021.

[5] O. Ibrahim, G. Skantze, S. Stoll, and V. Dellwo, "Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing." Interspeech, 2019.

[6] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[7] C. Gallois, T. Ogay, and H. Giles, "Communication accommodation theory: A look back and a look ahead," in *Theorizing about intercultural communication*. Thousand Oaks: Sage, 2005, pp. 121–148.

[8] U. C. Priva and C. Sanker, "Distinct behaviors in convergence across measures." in *CogSci*, 2018.

[9] A. Bell, "Language style as audience design," *Language in society*, vol. 13, no. 2, pp. 145–204, 1984.

[10] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Communication*, 2022.

[11] G. Doherty-Sneddon and F. G. Phelps, "Gaze aversion: A response to cognitive or social difficulty?" *Memory & cognition*, vol. 33, no. 4, pp. 727–733, 2005.

[12] G. Brône, B. Oben, A. Jehoul, J. Vranjes, and K. Feyaerts, "Eye gaze and viewpoint in multimodal interaction management," *Cognitive Linguistics*, vol. 28, no. 3, pp. 449–483, 2017.

[13] A. Abele, "Functions of gaze in social interaction: Communication and monitoring," *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 83–101, 1986.

[14] C. Mishra and G. Skantze, "Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1201–1208.

[15] K. Lemhöfer and M. Broersma, "Introducing lextale: A quick and valid lexical test for advanced learners of english," *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[16] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.

[17] E. R. Thompson, "Development and validation of an international English big-five mini-markers," *Personality and individual differences*, vol. 45, no. 6, pp. 542–548, 2008.

[18] D. Lüdecke, M. S. Ben-Shachar, I. Patil, and D. Makowski, "Extracting, computing and exploring the parameters of statistical models using R." *Journal of Open Source Software*, vol. 5, no. 53, p. 2445, 2020.

[19] M. Kardas, A. Kumar, and N. Epley, "Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation." *Journal of Personality and Social Psychology*, 2021.

[20] E. Hart, E. M. VanEpps, and M. E. Schweitzer, "The (better than expected) consequences of asking sensitive questions," *Organizational Behavior and Human Decision Processes*, vol. 162, pp. 136–154, 2021.

[21] Boersma, Paul and Weenink, David, "Praat: doing phonetics by computer [computer program]," http://www.praat.org/, 2022, version 6.2, retrieved August 23, 2022.

[22] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[23] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech." in *INTERSPEECH*, 2013, pp. 525–529.

[24] S. Nakagawa, P. C. Johnson, and H. Schielzeth, "The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of the Royal Society Interface*, vol. 14, no. 134, p. 20170213, 2017.

[25] K. Bartoń, *MuMIn: Multi-Model Inference*, 2023, r package version 1.47.5. [Online]. Available: https://CRAN.R-project.org/package=MuMIn

[26] D. Tingley, T. Yamamoto, K. Hirose, L. Keele, and K. Imai, "Mediation: R package for causal mediation analysis," 2014.

WHAT MAKES A GOOD PAUSE? INVESTIGATING THE TURN-HOLDING EFFECTS OF FILLERS

Bing'er Jiang, Erik Ekstedt, Gabriel Skantze

KTH Speech, Music and Hearing, Stockholm, Sweden
{binger, erikekst, skantze}@kth.se

ABSTRACT

Filled pauses (or *fillers*), such as *uh* and *um*, are frequent in spontaneous speech and can serve as a turn-holding cue for the listener, indicating that the current speaker is not done yet. In this paper, we use the recently proposed Voice Activity Projection (VAP) model, which is a deep learning model trained to predict the dynamics of conversation, to analyse the effects of filled pauses on the expected turn-hold probability. The results show that, while filled pauses do indeed have a turn-holding effect, it is perhaps not as strong as could be expected, probably due to the redundancy of other cues. We also find that the prosodic properties and position of the filler has a significant effect on the turn-hold probability. However, contrary to what has been suggested in previous work, there is no difference between *uh* and *um* in this regard.

Keywords: Hesitation, fillers, turn-taking, spoken dialog, computational modelling

1. INTRODUCTION

In spontaneous speech, filled pauses (or *fillers*), such as *uh* and *um*, are frequent and are typically associated with a hesitation on part of the speaker [1]. Some studies have found an association between the use of fillers and higher cognitive load in both monolog and dialog [2]. From the perspective of coordinating turn-taking in conversation [3, 4, 5], it has been suggested that fillers work as a *turn-holding cue* for the listener, indicating that the current speaker is not done yet [6, 7]. To what extent such fillers are produced intentionally for this purpose, and how much information they carry, is not yet clear [8].

According to [7], *uh* and *um* are to be considered as words in the speaker's vocabulary. While these words do not have any propositional content, it is argued that fillers conform to the "phonology, prosody, syntax, semantics, and pragmatics of English words". In this view, they should be seen as interjections (similar to *ah* and *oh*), indicating that the speaker wants to keep the floor. In addition, [7] claim that *uh* and *um* have different meanings, where the latter signals a longer upcoming delay. This claim is based on an analysis of the pause length following these fillers in several spontaneous speech corpora. Later studies of other corpora have also found such differences [2], while others have not [9]. Another possibility is that it is not the lexical form of the filler that is important for their function in signalling delay, but rather their prosodic characteristics. If so, it would perhaps be wrong to regard them as carrying distinct symbolic meaning. However, we are not aware of any studies that have investigated this.

A general problem with corpus-based analyses is that we can only find correlations, which do not necessarily imply causation. Thus, if we compare places with filled pauses with places without them, we do not know whether it is the filled pauses that actually cause the observed differences, or whether they were due to other factors that happened to correlate with the usage of fillers. One alternative is to perform perception experiments where stimuli are systematically manipulated, while all other factors are kept constant (e.g. [10, 6, 11, 12]). However, such experiments are costly to perform on a larger scale and they are often done with a small set of made-up examples that do not necessarily reflect general distributions. In this paper, we take a third route: To train a model that makes predictions in conversation and then use that model to investigate the effects of large-scale systematic stimuli manipulation.

Using the recently proposed **Voice Activity Projection** model [13] (**VAP**), which models the dynamics of conversation and can be used to predict the **turn-hold probability** (**THP**) (i.e., the probability that the turn will not shift to the other speaker) at any given point, we systematically remove and insert fillers in order to investigate their effects. We want to answer the following questions:

1. How is the THP affected when fillers are removed from their natural occurrences?
2. How is the THP affected when fillers are inserted at places which should be clearly turn-

yielding?

3. How do the properties of the filler (position, lexical form, pitch, intensity, duration) affect the THP?

2. VOICE ACTIVITY PROJECTION

Voice Activity Projection [13], VAP, is a predictive model of conversational dynamics developed for turn-taking research. The model objective is to continuously predict the upcoming voice activity of both speakers in a dialog. The voice activity is defined in binary terms (i.e., whether they will be speaking or not at a certain point in time), and the two speakers' future activities are jointly encoded into a discrete label that represents the next 2s of dialog.

Similar to a language model (text-to-text), the VAP model (speech-to-activity) is optimized through MLE (cross-entropy). The model processes raw spoken dialog audio and outputs a likelihood distribution over each of the discrete activity labels. This likelihood can then be used for controlling the turn-taking in a spoken dialog system, or as a tool to analyze the likelihood of turn-taking events in recorded dialogs. In previous work, it has been shown that the model is sensitive to prosodic cues [14]. In this work, we utilize the VAP model to investigate the effect of a filler has on how long the current turn holds until the other speaker takes the turn.

The activity projection is defined by eight time window bins of a 2s window, four for each speaker (0.2s, 0.4s, 0.6s, 0.8s respectively), which indicates if the speaker is active or not in that time window. The various combinations of these bins result in 2^8 = 256 discrete activity labels. While these labels represent various turn-taking-related events (such as turn-shifts and backchannels), we are here primarily interested in calculating the probability of who will be the next speaker. To do this, we take a weighted average over the probability assigned to each label, focus on the bins that cover our region of interest (0-600ms), and compare the contributions for each speaker. The result is a value between 0 and 1 that represents the probability that the current speaker will continue, defined as **turn-hold probability (THP)**. In our study, we apply the model on dialogs and get incremental predictions of THP.

The model used in this experiment is a stereo version of the original VAP model that operates on two separate waveforms (one for each speaker) and is trained on subsets of the Fisher part 1 and the Switchboard corpus [15, 16]. We define a held-out test set containing 577 sessions from the dialog act annotated subset [17], swda, of the switchboard corpus to be used in the subsequent analysis.

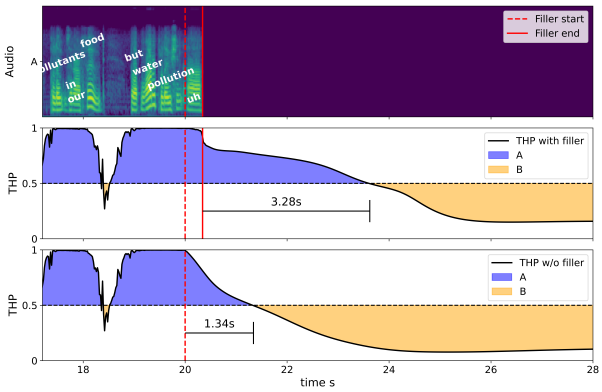


Figure 1: The THP following an utterance ending with *uh*. Top: the spectrogram of the spoken utterance; Middle: the THP with the filler; Bottom: the THP without the filler. The blue/yellow areas corresponds to A/B being the most likely next speaker. The black number bars represent the duration of silence before a turn-shift prediction (THP falls under 0.5).

3. EXPERIMENT 1: EXCLUDING FILLERS

We evaluate the effect of fillers through examining how 'soon' the VAP model predicts a turn shift after the end of a speech segment. In Experiment 1, we investigate the effect of excluding a valid filler present in the data. As illustrated in Figure 1, we apply the model to 20s of dialog (a pair of data samples including and excluding the filler), followed by 10s of artificial silence, and track the THP over the silence segment. While a pause of 10s is very unlikely in real life conversations, we use a longer length anyways to fully examine the prediction of the turn shift. The time t where the THP drops below 50% is considered the position of turn shift.

We select filler candidates *uh* and *um*, based on word-level transcriptions, and define three criteria in order to consider them valid fillers: 1) the duration of the candidates must be longer than 0.2s; 2) the fillers should be followed by a pause of at least 0.2s of the same speaker and 'isolated' from activity of the listener of at least 1s, before and after the filler; 3) there should be at least 20s of past dialog context (for the model to make its predictions). This results in 5316 valid fillers from the 577 test sessions.

We use *Survival Analysis* to examine the results, as some cases may not reach the turn shift predicted by the model even at the end of the 10s silence. The *Survival Analysis* includes a *censoring* status to

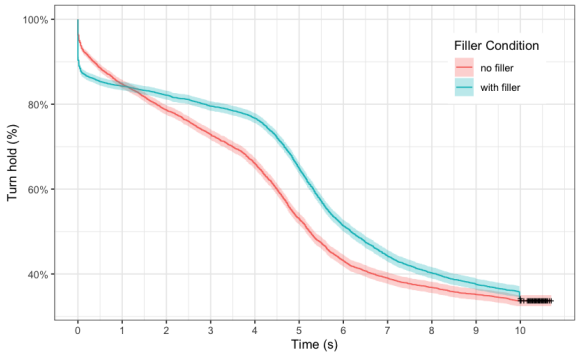


Figure 2: Kaplan-Meier curve of turn hold predictions through 10s of silence for Experiment 1 (excluding fillers), with or without fillers.

identify whether the expected event has happened or not (*uncensored* vs. *censored*). In our case, dialogs that are not predicted to reach turn shift by the end of the silence are marked as *censored*.

In order to examine whether the filler makes a difference, we plot the Kaplan-Meier curve of the two conditions. Figure 2 shows the proportion of data samples that keep the hold (y-axis) as a function of time (x-axis). Note that the y-axis does *not* reflect the THP; instead, it refers to percentages of cases where the THP are above 50%, i.e., percentages of data that are likely to keep the hold at time t.

Figure 2 shows that the filler indeed makes a difference in holding the turn, confirmed by the log-rank test ($p < 0.001$). Moreover, the effect changes throughout time. In particular, the filler does not help holding the turn at the very beginning of the silence, as shown by the drastic drop of the green line. At the later stage of the silence (after 1.2s), the effect of keeping the hold by the filler becomes more prominent, reaching the greatest between 4s and 5s, and then decreases until the end of the silence.

While there is a clear effect of the filler, it is interesting to note that the effect is perhaps not as big as expected; the turn hold effect is quite strong even with the filler removed. This is likely due to other redundant turn-holding cues present before the filler (prosody, syntax, etc).

4. EXPERIMENT 2: INSERTING FILLERS

While Experiment 1 indeed shows a turn-holding effect of existing fillers, it is unclear how strong the effect is, given the confounded turn-holding context. In Experiment 2, we further investigate the effects of fillers in contexts that are clearly turn-yielding. Specifically, we insert fillers at the end of yes/no questions and examine the difference in THP between the original data samples and the ones with

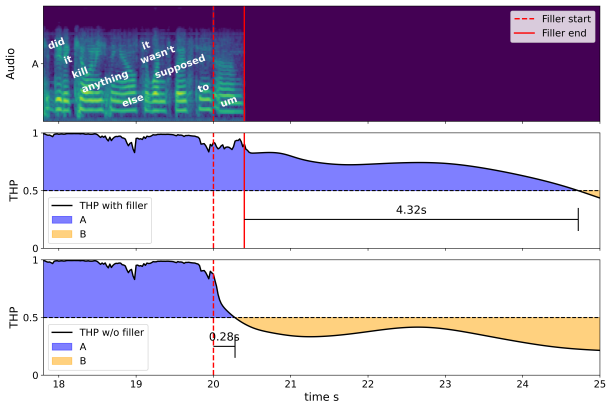


Figure 3: The THP following a yes/no question. Top: the spectrogram of the spoken utterance; Middle: the THP with the insertion of the filler *um*; Bottom: the THP of the original utterance.

inserted filler at the end.

We define valid utterances by the following conditions: we first select all utterances that are labeled as a yes/no-question dialog act according to the *swda* annotation [17]. The original utterance annotations may include multiple dialog act segments in a single utterance and we omit those that do not end with a yes/no-question dialog act. Valid utterances must be followed by a pause longer than 0.5s (from the same speaker) and 'isolated' from activity of the listener of at least 0.5s, before and after the end of the utterance. The candidates with a context duration of less than 20s are omitted. The final condition is that the model assigns a turn-shift in the first 5s of silence when 20s of context up and including the utterances are processed. This approach result in 245 valid turn-shift utterances.

For each valid utterance we select all the valid fillers, from the same speaker and session, as candidates to add to the utterance (1436 total filler candidates). Similarly to Experiment 1, we create paired data samples constituting 20s of dialog context, up to and including the utterances (with or without the added filler), followed by 10s of silence. This is illustrated in Figure 3.

Following the same preprocessing as described in Experiment 1, we plot the Kaplan-Meier curve, as shown in Figure 4. The figure shows that inserting a filler consistently lengthens the hold, and the two curves differ significantly from the log-rank test ($p < 0.001$). As can be seen, for the original yes/no questions, the model clearly predicts a fairly quick turn shift in most cases. While adding a filler does not flip all of them to turn-holds, the fillers seem to have a stronger effect here than in Experiment 1, and so the model clearly is sensitive to them.

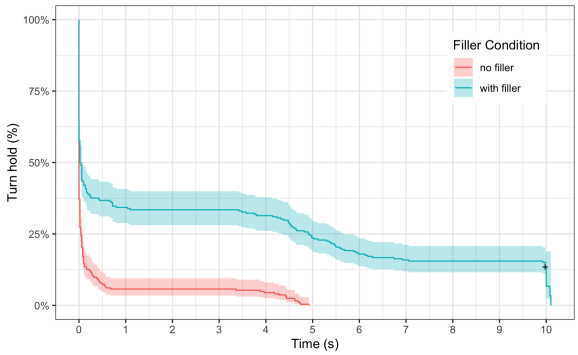


Figure 4: Kaplan-Meier curve of turn hold predictions through 10s of silence for Experiment 2 (inserting fillers after yes/no questions), with or without fillers.

5. EXPERIMENT 3: PROSODIC AND LEXICAL EFFECTS

Experiment 3 examines how prosodic and lexical properties of the filler affect the turn hold. For capturing prosodic information, we include by-speaker standardized mean **pitch**, by-speaker standardized mean **intensity** and filler **duration** on the log scale. We also examine whether the **lexical form** (*uh* or *um*), as well as the **position** of the filler in the current dialogue act (*mid* or *start*) has any influence on the THP. All continuous variables (prosody) are standardized by subtracting the mean and divided by two standard deviations. Note that the pitch and intensity variables are normalized with regard to each speaker, while duration is not.

We fit a Cox proportional hazards model on data with fillers from Experiment 1 to examine the joint effects from the five factors. The independent variables are the five factors and an interaction term between F0 and the lexical form, and the dependent variables are time reaching the turn shift derived from the VAP model and the censor status.

Table 1, summarizing the Cox model, shows that except for the filler type, all four other variables are statistically significant. The coefficients reflect the change of hazards function (here, the probability of the turn shift): a negative coefficient estimate corresponds to less 'hazard' of turn shift (higher probability of turn hold), while a positive coefficient estimate indicates an increased probability of turn shift. The exponentiated coefficients are termed as *hazards ratio*, which represents the difference in the hazard of an event at any given time.

The results from Table 1 reveals that the following factors help keep the hold (decrease the 'hazard'): higher pitch, start of the dialog act, higher intensity and longer duration. Moreover, while the lexical form of the filler itself does not significantly affect the hold, it has an indirect influence through pitch: *uh* has higher pitch and thus results in longer holds.

	coef	coef(exp)	SE	Pr(>|z|)
F0	-0.725	0.484	0.246	**0.003**
Intensity	-0.127	0.879	0.035	**0.0003**
Lex_{um}	-0.077	0.925	0.050	0.12
Duration	-0.118	0.888	0.037	**0.001**
Pos_{mid}	-0.305	0.736	0.065	**<0.0001**
F0:Lex_{um}	-1.237	0.237	0.290	**0.007**

Table 1: Model summary of the Cox regression model. Bold *p* values are significant.

Specifically, if a speaker raises their voice by 0.5 standard deviation (SD) of their pitch range, the probability of holding the turn determined by the model will increase by 51.6% (here, the change of hazards of turn shift from *hazards ratio* is 0.484 - 1 = -0.516). Increasing 0.5 SD of the intensity, however, only raises the probability by 12.1%. The duration also affects the turn hold: if the filler is lengthened by 0.5 SD, the probability is raised by 11.2% . In terms of the position, if a filler is at the beginning of the dialog act, it tends to raise the probability by 26.4% compared to one in the middle.

6. DISCUSSION AND CONCLUSIONS

In this study, we use computational modelling to evaluate the effects of fillers on holding a turn, and examine a variety of prosodic, positional and lexical properties of the fillers. Experiment 1 shows that excluding existing fillers indeed results in shorter holds, while the effect is perhaps not as strong as expected. This is likely due to the redundancy of other turn-holding cues. Experiment 2 further reveals that fillers have a strong turn-holding effect when added to clearly turn-yielding contexts (yes/no questions). Experiment 3 shows that a number of properties of the filler contribute to the turn-hold effect: higher pitch, higher intensity, and longer duration of the filler are associated with longer turn hold. Moreover, utterance-initial fillers have a stronger turn-hold effect than mid-utterance fillers. However, there is no significant difference between the lexical forms *uh* and *um*. Nevertheless, the two types of fillers still indirectly affects the THP through increasing the pitch of the filler. Thus, instead of the lexical forms, it is their phonetic realization that makes a difference, as *um* includes a nasal which associates with lower pitch compared to the oral *uh*. This speaks against the interpretation given by [7], where *uh* and *um* are seen as different words with different meanings.

7. ACKNOWLEDGEMENTS

This work was supported by Riksbankens Jubileumsfond (RJ), through the project *Understanding predictive models of turn-taking in spoken interaction* (P20-0484), as well as the Swedish Research Council, through the project *Prediction and Coordination for Conversational AI* (2020-03812).

8. REFERENCES

[1] J. E. Fox Tree, "Interpreting Pauses and Ums at Turn Exchanges," *Discourse Processes*, vol. 34, no. 1, pp. 37–55, Jul. 2002.

[2] R. L. Rose, "Um and Uh as Differential Delay Markers: The Role of Contextual Factors," in *Disfluency in Spontaneous Speech (DiSS)*, 2015.

[3] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 12 1974.

[4] S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[5] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction : A Review," *Computer Speech & Language*, vol. 67, 2021.

[6] P. Ball, "Listeners' responses to filled pauses in relation to floor apportionment," *British Journal of Social and Clinical Psychology*, vol. 14, no. 4, pp. 423–424, 1975.

[7] H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[8] M. Corley and O. W. Stewart, "Hesitation Disfluencies in Spontaneous Speech: The Meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[9] D. C. O'Connell and S. Kowal, "Uh and um revisited: Are they interjections for signaling delay?" *Journal of Psycholinguistic Research*, vol. 34, no. 6, pp. 555–576, 2005.

[10] M. Cook and M. G. Lalljee, "The interpretation of pauses by the listener," *British Journal of Social and Clinical Psychology*, vol. 9, pp. 375–376, 1970.

[11] J.-P. de Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[12] S. Bögels and F. Torreira, "Listeners use intonational phrase boundaries to project turn ends in spoken interaction," *Journal of Phonetics*, vol. 52, pp. 46–57, Sep. 2015.

[13] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," in *Proc. Interspeech 2022*, 2022, pp. 5190–5194.

[14] ——, "How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Edinburgh, UK: Association for Computational Linguistics, Sep. 2022, pp. 541–551.

[15] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*. USA: IEEE Computer Society, 1992, p. 517â520.

[17] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, Tech. Rep. 97-02, 1997.

Changes in parents' phonetic parameters during the first year of a child's life. Comparative study of German and Swedish

Melanie Weirich* & Adrian P. Simpson*

*Friedrich Schiller University Jena, Germany
melanie.weirich@uni-jena.de, adrian.simpson@uni-jena.de

ABSTRACT

This paper deals with inter- and intra-speaker variation over time in adult-directed speech of first-time parents. The study investigates the potential effects of sociopsychological parameters such as parental involvement and gender identity in German and Swedish mothers and fathers.

Picture descriptions of 71 participants were recorded at four time points (one before the birth of the child, three more during the baby's first year) and acoustic analyses were conducted regarding mean fundamental frequency and acoustic vowel space.

Inter-speaker variation was found between the parental groups reflecting (culture-specific) gender roles (expressed through parental involvement and femininity). Intra-speaker variation over time in mean f0 only found in the German mothers, points to the role of sociopsychological factors besides hormonal influences due to pregnancy.

Keywords: fine phonetic detail, f0, AVS, parents, intra-speaker variation

1. INTRODUCTION

Until recently, sex- and gender-specific fine phonetic detail has been investigated by highlighting differences between male and female speakers. Alongside biological factors responsible for differences between the sexes [1-4], social factors have been claimed to explain gender-specific variation in adults [5-9]. By looking at variation within genders and incorporating the impact of sexual orientation, gender research has become increasingly multi-facetted [10-12].

Most recently, gender identity in terms of self-rated femininity, masculinity or gender role orientation has become a further intriguing factor being considered to explain gender-specific variation in fine phonetic detail within females and males independently of or in addition to sexual orientation. Heterosexual German males have been found to indicate self-ascribed femininity through fine phonetic detail [13]. The study found that the more feminine a speaker rated himself, the higher his speaking fundamental frequency (SFF) and the larger his acoustic vowel space (AVS) – i.e. these males exhibited phonetic parameters shifted towards female values. In addition, [14] showed that gender identity differs between German and Swedish participants with Swedish females showing lower femininity scores than German females, in turn resulting in similar scores between the genders in Sweden, but significant differences in Germany.

The present study also looks at the relationship between gender identity and phonetic cues in German and Swedish speakers, but additionally investigates the gender role, more specifically differences in parental role of individual speakers. Moreover, intra-individual variation is examined in self-ascribed gender identity in young parents during the first year of their first child's life – looking at possible changes in fine phonetic detail. [16] found lower SFF and less variation in SFF in English females one year after giving birth than during pregnancy and 5 years later. The authors name several potential reasons for these intra-individual changes over time. The main factor the authors suggest are hormonal changes after pregnancy (decrease in estrogen). However, they also point to the potential influence of behavioral factors such as mental and physical fatigue. A similar decrease in SFF and AVS has also been found in German mothers during the first year after the birth of their first child [15], accompanied by changes in self-ascribed femininity. The present study adds a later time point for the German data set as well as comparing the German speakers with Swedish mothers and fathers from speech samples recorded at the same four time points before and after the birth of the first child. Sociopsychological data (self-ascribed femininity and parental involvement) are collected and investigated as possible interacting factors.

2. METHOD

Written consent from all subjects taking part in the study was obtained prior to the recording. The project of which this study is part (see [15]) was reviewed and approved by the ethics committee of the University of Jena. Acoustic recordings and sociopsychological data (regarding femininity and parental involvement) were collected from Swedish and German mothers and fathers at four time points before the baby's birth and within its first year.

2.1. Participants and sociopsychological measures

35 Swedish speakers (15 females and 20 males) and 36 German speakers (15 females, 21 males) participated in the study. The German participants were on average 29.6 years (SD = 3.7, mothers) and 30.8 years (SD = 6.0, fathers), respectively. Swedish participants were on average 31.5 years (SD=3.9, mothers) and 32.8 years (SD=4.5, fathers), respectively. The fathers differed in the time they stayed at home with the child as primary caregiver and were separated into 2 groups, accordingly. One group (Fa) consisted of fathers who took parental leave for at least 3 months when their child was between three and ten months old, this being the crucial period when recordings two and three were made. The other (control) group comprised fathers (CF) who, on average, only took 12 days (Sweden) or 10 days (Germany) leave. In Sweden 6 males form the Fa group, while in Germany 9 Fa males took part. The CF group is represented by 11 German and 15 Swedish participants.

Participants were asked at each time point about the time they spend with their child and the activities they do with them (eight items, asking about the frequency of doing things like play, comforting, nappy changing in comparison to their partner). Figure 1 shows the values averaged over the three time points within the first year separated by group and country. A split between the groups, with Fa participants lying between CF and Mo, is apparent in both countries. However, it is more distinct in the German group. The LMM with involvement as dependent variable and speaker as random effect reflects this by showing a significant interaction of country and speaker group ($\chi^2(3) = 23.5$, $p < .001$). Post hoc comparisons reveal significant differences between all three groups in the German participants ($p < .05$), but only between Mo and CF in the Swedish participants ($p < .001$). In the Swedish sample, the less clear separation between the father groups and the genders in general points to a more balanced participation in childcare.

Regarding gender identity, data was gathered in terms of self-ascribed femininity using the GEPAQ_F+ questionnaire (positive Femininity scale of the German Extended Personality Attributes Questionnaire [17]). The questionnaire consists of 8 bipolar adjective pairs describing positively evaluated personality traits traditionally attributed to women (such as kind, empathetic, sensitive, gentle). The German version was translated into Swedish. Participants were asked to rate themselves on these adjective pairs on a scale from 1 to 7. Mean values were calculated for each participant with high values reflecting a high femininity score.

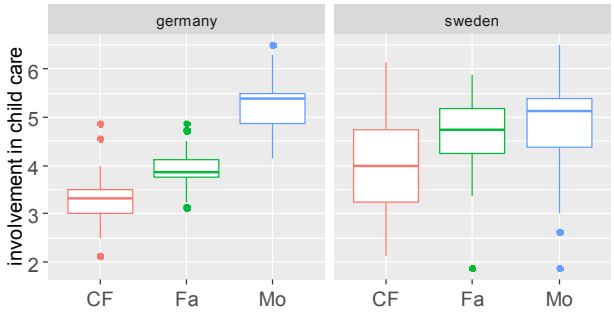


Figure 1: Involvement in childcare separated by country and parent group. High values reflect a strong involvement in childcare.

2.2. Speech data collection

Acoustic recordings were conducted using a headset-microphone (Sennheiser ew 100 G3 – SK100) and a ZOOM – H6 Handy Recorder. Each speaker was recorded at four different time points at their homes by a German or Swedish female experimenter, respectively. The first recording took place 4 weeks before the expected due date. The second recording was conducted approximately 5 months, the third one 8 months and the last one 12 months after the birth.

Semi-spontaneous speech material was elicited using a picture description task. Participants were asked to describe the same fifteen pictures (in a randomized order) at each of the four recording sessions to the experimenter. The pictures showed various objects and animals that were chosen as carrier words for the selected target vowels, with each carrier word appearing on at least three pictures (see [14]). Vowels slightly differ between the languages due to language specific vowel inventories. The respective vowels (for German: /i: ɛ a u:/, for Swedish: /i: ɛ a ɑ u:/) were chosen to estimate the vowel space area (VSA) by calculating the area of the polygon formed by joining the points defined by the first two formants of these vowels. In total, 5446 vowel tokens were analysed for the Swedish data set and 4839 vowel tokens for the German data.

2.3 Acoustic analysis

Acoustic analyses were carried out in PRAAT [18]. Modified versions of publicly available scripts - distributed under the GNU General Public License by Mietta Lennes and Jonas Lindh - were used to calculate the respective phonetic parameters. A psychoacoustic transformation of formant and fundamental frequency values from Hz to Bark was carried out [19]. The calculation of the AVS was done using the phonR package [20] in R [21].

2.3.1 Speaking fundamental frequency (SFF)

SFF was measured for each participant and time point over the whole sequence of the spontaneous data recorded during the picture description task. Recordings lasted on average 223 seconds (SD =167 s) in the German data, and 287 seconds (SD:178 s) in the Swedish data. The analysis parameters in PRAAT consisted of a minimum pitch of 75 Hz, a maximum pitch of 600 Hz and a time step of 0.01 s.

2.3.2 Formants and acoustic vowel space (AVS)

The first two formants (F1, F2) in the target vowels of the carrier words were measured at the vowel midpoint. The analysis parameters consisted of a time step of 0.01s, the maximum number of formants = 5, a window length of 0.025 s and a pre-emphasis from 50 Hz. The maximum formant value was set to 5000 Hz for males and 5500 Hz for females. Measurements were checked manually if necessary. Mean F1 and F2 values of each vowel category were calculated for each speaker and time point and used to estimate the size of a speaker's AVS for each recording.

3. RESULTS

Prior to statistical analysis we begin by providing a description of certain general patterns that are apparent from Fig. 2. This allows us to consider similar developments in the different parameters (femininity, SFF, AVS) over time, as well as possible similarities in the differences between the parental groups (CF, Fa, Mo) across the different parameters. In the statistical analysis, on the other hand, we will be concentrating on individual parameters separately as they differ between the groups and across the time points.

3.1 Variation patterns in femininity and phonetic parameters over time – descriptive analysis

Figure 2 shows the variation in self-ascribed femininity (top), SFF (middle) and AVS (bottom) over the child's first year for the different parent groups.

Independent of time, the figure shows interesting differences between parent groups and countries. As already shown in [14], regarding femininity, the genders only differ in Germany but not in Sweden, which is reflected in smaller phonetic gender differences in Sweden than in Germany. In addition, we can see that while the German CF and Fa groups vary with higher values for Fa and thereby more towards the femininity scores of the mothers, the Swedish Fa group lies below the CFs – except for the time following the baby's birth (2, green in Figure 2). Regarding the phonetic parameters (middle and bottom panels), patterns similar to the femininity scores arise: while the German Fa parents lie in between CF and Mo (analogue to their femininity score), the Swedish Fa parents pattern with the CF males (for AVS) or show values below them (for SFF) – reflecting their somewhat lower femininity score.

Looking at the variation over time, the clearest pattern can be seen in the German Mos, with decreasing values in all three parameters (femininity, SFF and AVS) from time 1 to 3 and a slight increase again to time 4 (c.f.[15]). However, for the other groups the patterns are not as apparent. The Swedish Fa group also shows time-dependent variation in femininity with highest values at time 2 after the birth of the child (as mentioned above). However, this is not reflected in the phonetic parameters. In the Swedish mothers a similar pattern appears in femininity and AVS but not SFF.

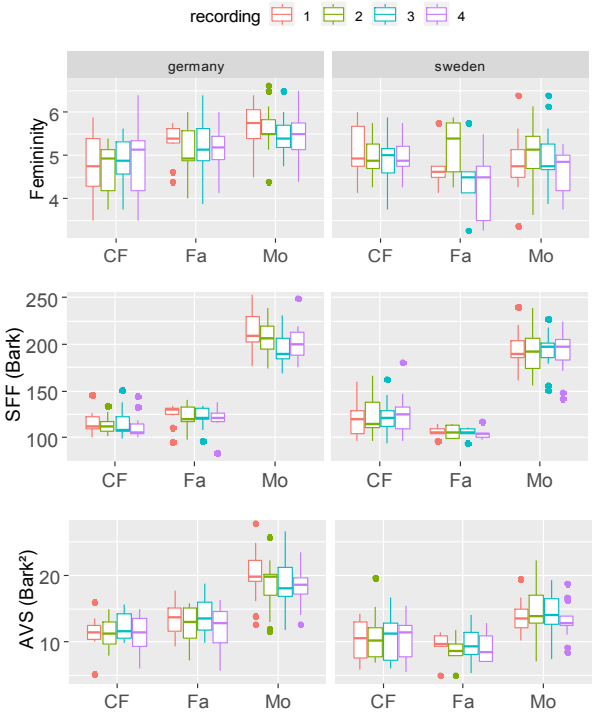


Figure 2: Femininity (top), SFF (middle) and AVS (bottom) separated by time of recording, country and parent group.

3.2 Statistical analysis

Statistical analyses were conducted in R [20]. Linear mixed models were run using *lme4* [22]. P-values were obtained using Likelihood ratio tests comparing models with and without factors in question. Post hoc Tukey tests were calculated using *lsmeans* [23]. As fixed effects we entered parent group, time and country, and speaker as random effect.

Since in [13] significant positive correlations between femininity and SFF and AVS were found for German male participants at the first time point, here, we investigated this relationship for this time point in the

Swedish participants. In contrast to the German data, no significant correlations were found, either for males or females.

Regarding femininity, a significant interaction of group and country ($\chi^2(4) = 19.46$, $p < .01$) was found. Post hoc tests reveal only the difference between the German CF and Mo speakers to be significant (estimate: 0.79, $p < .001$). The three-way interaction of group, country and time fails to reach significance ($p = .09$), which is probably due to the high intra-group variation for each of the time points.

For SFF, the three-way interaction of group, time and country is significant ($\chi^2(15) = 32.1$, $p < .01$). Post hoc Tukey tests reveal significant differences between time points for the German mothers only, with decreasing values starting from time 1 (1 vs. 2,3,4 and 2 vs.3, $p < .05$). No significant differences in the father groups and also in any of the Swedish speakers are found. Comparing the countries, significant differences between Swedish and German mothers appear at times 1 and 2, due to the decrease in SFF in the German mothers at the later time points.

For AVS, a significant interaction of country and group was found ($\chi^2(2) = 6.2$, $p < .05$) with significant differences between Mo and both father groups in both languages ($p < .05$), but higher estimates in the German speakers than in the Swedish speakers (7.5 and 5.6 in German vs. 3.8 and 3.6. in Swedish). When language groups are compared, significant differences were found for mothers (estimate: 5.02, $p < .001$) and Fa males (estimate: 3.04, $p < .05$), but not CFs. In addition, a significant main effect of recording ($\chi^2(3) = 11.69$, $p < .01$) was found, with a decrease from time 1 to 4 (estimate: 0.75, $p < .05$) and from time 3 to 4 (estimate: 0.8, $p < .05$). Even though group specific variation over time appeared, e.g., with a clearer decrease in German mothers compared to Swedish mothers, the three-way interaction of group, corpus and time failed to show significance ($p = .10$).

4. SUMMARY AND DISCUSSION

Although the countries analysed, Germany and Sweden, are similar in terms of political system, economy, religion and gender equality, even here, a clear difference in self-ascribed femininity was found. In Sweden, which has been at the forefront of providing financial support for both parents following childbirth since the 1970s, females do not differ from males in self-ascribed attributes such as emotional, empathetic or helpful, and this is not due to higher femininity ratings in males, but to lower scores in females. While in Germany some men also rate themselves high on these traditionally female attributes, in Sweden, a lower score in terms of conservative femininity is the accepted standard in females. The difference between the countries in the construction of gender is reflected in phonetic variation. In both parameters, females revealed lower values in the Swedish than in the German participants, thereby leading to a less clear phonetic gender-specificity. In addition, the German father group which showed heightened femininity ratings (Fa) also showed acoustic characteristics more towards the female values (in SFF and AVS), and thus lies between those of females and CF males. In the Swedish participants the Fa group shows lower femininity and phonetically patterns with the CF group or lies below them. Note though, that conclusions drawn from the Swedish Fa group are limited due to the small participant number.

Thus, results indicate that gender-specific variation in speech is influenced by a culturally affected understanding of gender. In addition, while it has been found that femininity is indexed and perceived through fine phonetic detail in German [13], Swedish speakers do not seem to signal femininity in these phonetic cues. From that we have to ask whether the sociopsychological score we have used does not measure femininity in Sweden correctly or whether Swedish speakers express their femininity in speech through other parameters not measured here.

One rationale for the investigation over time were previous findings showing that British English mothers' SFF decreased during the baby's first year [16]. A potential reason suggested for this variation is the hormonal change after pregnancy with a decrease in estrogen, progesterone, and cortisol levels. The results of the German mothers corroborate this finding, while the Swedish mothers do not exhibit a similar decrease. Although we did not measure hormonal levels, we suggest that these differences between German and Swedish mothers do cast doubt on a mainly hormone-based explanation of the phonetic changes. In addition to the behavioural shifts in self-ascribed femininity that we found, changes in physical and mental fatigue within the baby's first year might lead to lower SFF [24-25] and potentially also hypoarticulated vowels and a smaller AVS [26]. While we did not ask for self-ratings of fatigue or depression, we assessed the involvement in childcare and clear differences between the genders in Germany and between the countries appeared: While in the German group the main part of childcare is done by mothers (irrespective of time and parental leave of the partner), in the Swedish group the participation of both parents (also working CFs) is much more balanced.

The effect of culture-specific gender role concepts and intra-individual variation over time in self-rated gender identity point to a complex interplay of situational changes in life (becoming a parent) and (culture-specific) socio-phonetic variables.

5. ACKNOWLDEGMENTS

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grant WE 5757). Many thanks to Fabienne Neder, Heidi Springel, Franziska Rauch and Sandra Schäder (Friedrich Schiller University Jena) for their help in data acquisition and annotation in Germany and Jasmin Öjbro and Christine Ericsdotter Nordgren (University of Stockholm) for collaboration and data acquisition in Sweden. We are very grateful to all participating subjects.

6. REFERENCES

[1] Goldstein, U. (1980). An articulatory model for the vocal tracts of growing children (PhD thesis). Massachusetts Institute of Technology.

[2] Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. Journal of the Acoustical Society of America, 85, 1699–1707.

[3] Stevens, K. N. (1998). Acoustic phonetics. Cambridge, MA: MIT Press

[4] Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America, 106(3), 1511– 1522.

[5] Boe, L.-J., Contini, M., and Rakotofiringa, H. (1975). Etude statistique de la frequence laryngienne ("Empirical study on laryngeal frequency"), *Phonetica*, 32, 1–23.

[6] Takefuta, Y., Jancosek, E. G. & Brunt, M. (1972). A statistical analysis of melody curves in the intonation of American English, *Proceedings of the VIIth International Congress of Phonetic Sciences*, Montreal, pp. 1035-1039

[7] van Bezooijen R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Lang Speech*. 38: 253 265.

[8] Mennen, I., Schaeffler, F. & Docherty, G. (2012) Cross-language differences in fundamental frequency range: A comparison of English and German. *J Acoust Soc Am*. 131(3): 2249–2260.

[9] Simpson, A. & Weirich, M. (2020) Phonetic Correlates of Sex, Gender and Sexual Orientation. In: Mark Aronoff (Ed*.) Oxford Research Encyclopedia of Linguistics*, Oxford University Press, New York.

[10] Munson, Benjamin, Elizabeth C. McDonald, Nancy L. DeBoe, & Aubrey R. White (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics* 34, 202-240.

[11] Sulpizio, S., Fasoli, F., Maass, A., Paladino, M.P., Vespignani, F., Eyssel, F. & Bentler, D. (2015) The sound of voice: Voice-based categorization of speakers' sexual orientation within and across languages. *PLOS ONE* 10: e0128882. https://doi.org/10.1371/journal.pone.0128882 PMID: 26132820

[12] Kachel, S., Simpson, A.P. & M.C. Steffens (2018). Do I sound straight?: Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech. *J Speech Lang Hear Res*.; 61: 1560–1578. https://doi.org/10.1044/2018_JSLHR-S-17-0125 PMID: 29955829

[13] Weirich M. & Simpson, A. (2018) Gender identity is indexed and perceived in speech. *PLoS ONE* 13(12):e0209226. doi:10.1371/journal.pone.0209226

[14] Weirich, M., Simpson, A. P., Öjbro, J., & Ericsdotter Nordgren, C. (2019) The phonetics of gender in Swedish and German. In P*roceedings of FONETIK* (pp. 49–53)

[15] Weirich M. & Simpson, A. (2019) Effects of Gender, Parental Role, and Time on Infant- and Adult-Directed Read and Spontaneous Speech. *Journal of Speech, Language and Hearing Research.62(12):1-14.*

[16] Pisanski, K., Bhardwaj, K., & Reby, D. (2018). Women's voice pitch lowers after pregnancy. *Evolution and Human Behavior*, 39, 457–463. https://doi.org/10.1016/j.evolhumbehav.2018.04.002

[17] Runge, T.E., Frey, D., Gollwitzer, P.E., Helmreich, R.L. & Spence J.T. (1981). Masculine (instrumental) and feminine (expressive) traits: A comparison between students in the United States and West Germany. J Cross Cult Psychol. 12(2): 142–162.

[18] Boersma, P., Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0.40) [Computer program]. Retrieved from http://www.praat.org/

[19] Traunmüller H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. doi:10.1121/1.399849

[20] McCloy DR (2016). *phonR: tools for phoneticians and phonologists*. R package version 1.0-7.

[21] R Core Team (2016). *R: A language and environment for statistical computing* [Computer software manual]. Vienna: R Statistics Foundation

[22] Bates D, Maechler M, Bolker B, Walker S. (2015) Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw.,* 67(1): 1–48.

[23] Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *J. Stat. Sofw*. 69, 1–33.

[24] Cho, S.-W., Yin, C. S., Park, Y.-B., & Park, Y.-J. (2011). Differences in self-rated, perceived, and acoustic voice qualities between high-and low-fatigue groups. *Journal of Voice*, 25(5), 544–552.

[25] Whitmore, J., & Fisher, S. (1996). Speech during sustained operations. *Speech Communication*, 20(1-2), 55-70.

[26] Lindblom B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In: Hardcastle W.J., Marchal A. (eds) *Speech Production and Speech Modelling*. NATO ASI Series (Series D: Behavioural and Social Sciences), vol 55. Springer, Dordrecht

Variation in FACE and FLEECE trajectories in Australian English adolescents according to community language diversity

Joshua Penney, Felicity Cox, Andy Gibson

Department of Linguistics, Macquarie University, Sydney
joshua.penney@mq.edu.au; felicity.cox@mq.edu.au; andy.gibson@mq.edu.au

ABSTRACT

Australia is an increasingly multicultural society with high levels of linguistic diversity found within its larger cities. Nevertheless, research on Australian English is often based on an Anglo-centric, monolingual model. In this paper, we explore variation in the production of FACE and FLEECE vowels produced by adolescent speakers from areas of Sydney that differ in their level of language diversity and the dominant non-English languages spoken within them. A dynamic vowel formant analysis demonstrates that speakers from more linguistically diverse areas produce a raised first element of FACE and less onglide of FLEECE, compared with speakers from a less linguistically diverse area, whose production of these vowels more closely resembles previous descriptions of 'mainstream' Australian English. Productions from the linguistically diverse areas pattern similarly, indicating the changes are not direct transfer effects.

Keywords: Australian English; diphthongs; sound change; language contact; sociophonetics

1. INTRODUCTION

Australian English (AusE) exhibits a number of phonetic characteristics that distinguish it from other varieties of English. The majority of the described differences occur within the vowel system [8]. In this paper, we focus on two vowels that have long been considered markers of an Australian accent: FACE and FLEECE [10]. FACE, generally transcribed as /æɪ/ in AusE, can be described as a front closing diphthong comprising an open first element in the region of TRAP and a gliding component pointing towards KIT [17]. There is evidence that after a period of historical lowering and fronting described as anticlockwise diphthong shift by [25] and 'drift' by [23], the first element of this vowel may now be in the process of raising towards the vicinity of DRESS [7, 12, 21] in a clockwise reversal of the previously described diphthong shift.

Although FLEECE is generally described as a long monophthong, it shows a delayed target with a long onglide in AusE in which F1 decreases and F2 increases over the course of the vowel from a more centralised location towards its position at the high front extremity of the space, giving this vowel a diphthongal quality [17, 11]. This onglided FLEECE, which may be transcribed as [ᵊi:], is considered characteristic of AusE [9, 21]. The diphthongal nature of AusE FLEECE is made possible by the lowered first element of FACE ensuring separation between the two vowels.

Contemporary Australian society is increasingly multicultural, particularly in larger cities such as Sydney. According to the 2016 national census, more than two thirds of Sydney's residents have at least one overseas-born parent, and nearly 40% of households speak a language other than English at home [2]. Despite such high levels of linguistic diversity, and evidence that this has driven changes to AusE in the recent past [18, 15], until relatively recently much of the research on AusE has been based on what is often referred to as 'mainstream' AusE (MAusE), based on an Anglo-centric, monolingual/monocultural model [6, 13] (though see [5] for a review of studies examining ethnocultural variation in AusE). As part of the Multicultural Australian English project, we are currently building a corpus of audio recordings and survey data from adolescent AusE speakers from areas of Sydney that vary in their levels of community language diversity and the dominant non-English languages spoken within them, including very heterogeneous areas with high proportions of linguistic diversity through to a relatively homogeneous, English-dominant area. This project will enable us to create a more inclusive model of variation in AusE and to examine the effects cultural and linguistic diversity have on sound change.

Previous research has demonstrated that speakers from linguistically diverse communities may propel changes in vowel realisations. For example, [4] found that second-generation AusE speakers with Italian and Greek heritage advanced change away from overtly local (broad) forms that had been adopted by their first-generation parents. More recently in London English, speakers from linguistically diverse communities have been shown to advance change to the diphthong system (among a raft of other changes). This includes the raising and fronting–and in some cases monophthongisation–of FACE [3, 4, 14]. Such changes are likely due to effects of language contact in diverse communities that are then spread through

social interaction and accommodation to innovative forms [4, 16].

We here present an analysis of FACE and FLEECE vowels produced by AusE speakers from areas of differing linguistic diversity. We consider the entire shape of the trajectory to account more fully for dynamicity. We may expect to observe that speakers from areas with higher levels of linguistic diversity will produce vowels that differ from those produced by speakers from a more homogeneous, English-dominant area, who we would expect to produce vowels that represent traditional AusE descriptions.

2. METHODS

2.1. Participants and locations

Speech data from 59 adolescent male speakers of AusE aged 15-18 (mean age: 15.8) are included in this analysis. All participants were either born in Australia or migrated to Australia at a very young age. Participants were recruited from high schools throughout Sydney. The schools were selected due to their locations in areas of Sydney that differ in the level of community diversity and the dominant non-English languages spoken within these areas. Five areas of Sydney were selected: the Northern Beaches (NB, n=20) was selected as it has a high level of residents who speak English only and low levels of community language diversity. The other areas were selected as they have high proportions of residents who speak the four most common non-English languages: Arabic in Bankstown (BK, n=9); Chinese (Mandarin and Cantonese) in the Inner West (IW, n=14); Hindi (and other Indo-Aryan languages) in Parramatta (PM, n=5); Vietnamese in Cabramatta (CB, n=11). Note that participants were recruited based on their residence/schooling in these areas, not their language background. Therefore, speakers within each area do not necessarily speak these languages but nonetheless live in the areas in which the dominant non-English languages are spoken. For each of the linguistically diverse areas the majority of participants had a non-English speaking background; for the low diversity area the majority of participants had monolingual English-speaking backgrounds.

2.2. Data collection

41 of the speakers were recorded in a face-to-face setting in a quiet room of their school (n=36), a local library (n=4), or their home (n=1); 18 speakers were recorded remotely via supervised video call while at school during periods of restriction on face to face contact due to COVID-19. Face-to-face participants were recorded to a Zoom H6 recorder through a RODE HS2 headset microphone with a sampling rate of 44.1kHz and 16 Bit quantisation. Remote participants were recorded through an online recorder [https://mmig.github.io/speech-to-flac/] with a sampling rate of either 44.1kHz or 48kHz and 16 Bit quantisation. All files were subsequently resampled to 44.1kHz.

The data for this analysis were collected as part of a picture naming task, in which participants were prompted to produce 225 words and short phrases through the presentation of images on a computer screen. Participants were also recorded while taking part in a spontaneous conversation with a peer and a research assistant; however, these conversational data are not included in this analysis.

For this analysis, we focus on items with FACE and FLEECE vowels produced in single words with /(C(C))VC/ syllable structure, where the coda /C/ was either a voiced or voiceless obstruent (e.g. *spade*, *seat*, etc.). Each word was produced once by each speaker with the exception of one of the words containing FLEECE (*seed*), which was elicited three times for the purposes of formant normalisation. This resulted in 764 items for analysis (FACE: 353; FLEECE: 411).

2.3. Data processing and acoustic analysis

Audio files were segmented with [19] utilising an AusE model. Segment boundaries were subsequently hand corrected by trained annotators. The data were imported into an emu database using [26]. Formant frequencies were estimated with [1] utilising default settings for male speakers (5 formants; 5000Hz ceiling; 0.025s window length) and time normalised. Values in the first and last 15% of the vowel were excluded to minimise effects of neighbouring segments. F1/F2 trajectories were visually inspected and obvious outliers hand corrected. F1/F2 values were *z*-score normalised within speakers [20].

2.4. Statistical analysis

The data were analysed with generalised additive mixed models (GAMMs) using the mgcv [27] and itsadug [24] packages in R [22]. Separate models were fitted for F1 and F2 for each of the two vowels. Each model included a parametric difference term for the five areas (to model constant difference between areas), a smooth over normalised vowel duration (with NB as reference), a smooth over normalised vowel duration by area (an ordered factor modelling non-linear differences from the NB reference for each of the other areas), and a random smooth over normalised vowel duration by participant. For each model basis functions were set to ten (i.e. k=11) and an AR1 error term was included to account for autocorrelation in the models.

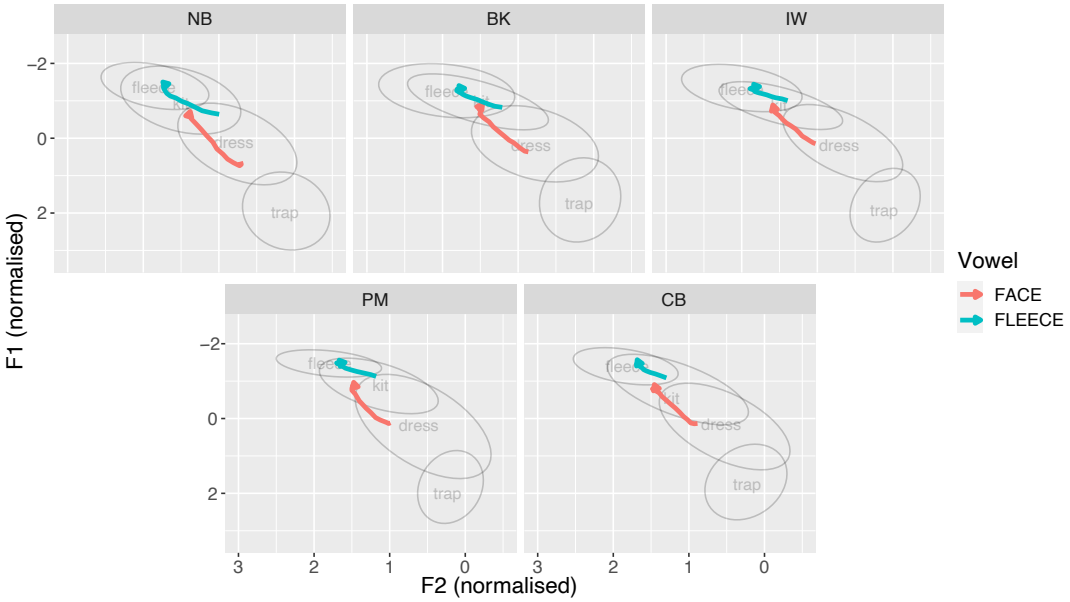


Figure 1. Mean normalised F1/F2 formant trajectories for FACE (red lines) and FLEECE (blue lines) vowels with reference to mean monophthongal vowel space for speakers from five areas of Sydney.

3. RESULTS

Figure 1 illustrates the mean F1/F2 trajectories of FACE and FLEECE for each of the five areas with reference to vowel spaces based on mean target values of monophthongal vowels for each area. (Note that the monophthongs were elicited in the same task as the FACE and FLEECE vowels outlined above; however, their inclusion here is for illustrative purposes only and a complete analysis of the monophthongs is not included in this paper.) From the figure it can be seen that the NB speakers exhibit a lowered and retracted onset of FACE compared to other speaker groups beginning from below the mean of DRESS. This is consistent with recent reports of MAusE FACE decoupling from TRAP that has recently been described for this vowel [12, 21]. In contrast, the speakers from each of the other areas show an even more raised and fronted onset. Additionally, the NB speakers exhibit a clear onglide in FLEECE, as expected according to past descriptions of AusE [17, 9]. The speakers from all of the other areas appear to show a reduced onglide relative to that observed in the NB group, although the BK group appears to produce an onglide that is intermediate between the extremes shown by the NB group and the other three groups.

The results showed significant parametric (i.e. constant) and non-linear effects for the GAMM analyses for both F1 and F2 of FACE for each of the areas relative to the reference NB area (see Table 1 for a summary). This indicates that speakers from all four of the areas showed lower F1 values (i.e. higher realisations) and higher F2 values (i.e. fronter realisations) in general compared to the NB group, but that they also showed differences in the shape of the vowel trajectory. Figure 2 shows the predicted values of the GAMMs for F1 and F2 of FACE and illustrates these differences between the areas.

Area/Vowel	BK	IW	PM	CB
FACE F1 (parametric)	*	*	*	*
FACE F1 (non-linear)	*	*	*	*
FACE F2 (parametric)	*	*	*	*
FACE F2 (non-linear)	*	*	*	*
FLEECE F1 (parametric)	-	*	*	*
FLEECE F1 (non-linear)	*	*	*	*
FLEECE F2 (parametric)	*	*	-	-
FLEECE F2 (non-linear)	-	*	*	*

Table 1: Summary of significant differences from the reference EN group for GAMMs for F1 and F2 in FACE and FLEECE. Significant effects (at $\alpha = 0.05$) are marked by asterisks.

For FLEECE, the F1 model showed significant parametric differences for all of the groups except BK, and significant non-linear differences for all areas relative to the reference NB group. For F2, there were significant parametric differences for BK and IW relative to the NB reference group, and significant non-linear differences for all areas apart from BK. This indicates that all of the groups, apart from BK, showed overall lower F1 values (i.e. higher realisations) in addition to differences in trajectory shape of F1. For F2, the BK and IW groups showed overall higher values (i.e. fronter realisations) than the NB group; however, the PM and CB groups did not show overall fronter realisations. With the

exception of the BK group, all of the groups produced F2 trajectories that differed in shape from the NB group. Figure 3 shows the predicted values of the GAMMs for F1 and F2 of FLEECE and illustrates the differences between the areas.

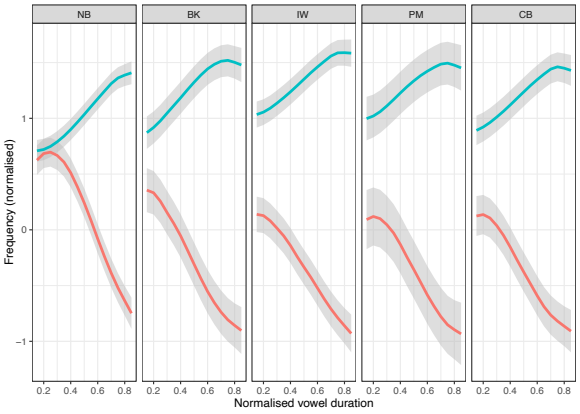


Figure 2: Model predictions of normalised F1 (red lines) and F2 (blue lines) of FACE by normalised vowel duration for speakers from five areas of Sydney. Error ribbons show 95% CIs.

4. DISCUSSION

The results described above demonstrate that there is variation in the production of these two vowels in AusE according to the area a speaker comes from. We assume that the trajectories of the NB area group can be taken to represent the productions of AusE speakers with reduced influence from speakers with language backgrounds other than English compared to the other groups. Correspondingly, as seen in Figure 1, their productions conform to previous descriptions of 'mainstream' AusE: for FACE, we observe a lower first element compared to the other groups with evidence of the recently reported raising change to this vowel, and for FLEECE, we observe the characteristic onglide.

On the other hand, compared to descriptions in the AusE literature, the four other groups appear to produce innovative realisations of these vowels: FACE and FLEECE are both produced with markedly shorter trajectories by all of these groups (see Figure 1) – and, for the most part, with higher and fronter onsets – and for FACE, a higher offset is also evident. As these areas represent communities with high levels of linguistic diversity, we interpret these results as indicating change to these vowels due to language/dialect contact within these communities.

Interestingly, speakers from the linguistically diverse areas seem to pattern together and produce similar realisations of these vowels, although the BK group appear to produce realisations that lie somewhere between the NB group and the other groups. This would seem to indicate that the changes are not due to direct transfer effects from one of the non-English languages spoken in these areas, or that, if they originally were due to such transfer effects, these have now spread through the linguistically diverse areas of Sydney beyond those with a particular language background [3, 4]. Additionally, at least with regard to FACE, the innovative realisations appear to be an advancement of the raising change that has been noted for MAusE speakers. While we cannot speculate as to the initiation of this change, it would appear that it is now being driven by speakers in the linguistically diverse areas, with the speakers from the homogeneous area lagging behind. Taken together, these findings may suggest a shift in speakers from linguistically diverse areas away from overtly local, or mainstream, AusE forms, similar to the observations made by [18].

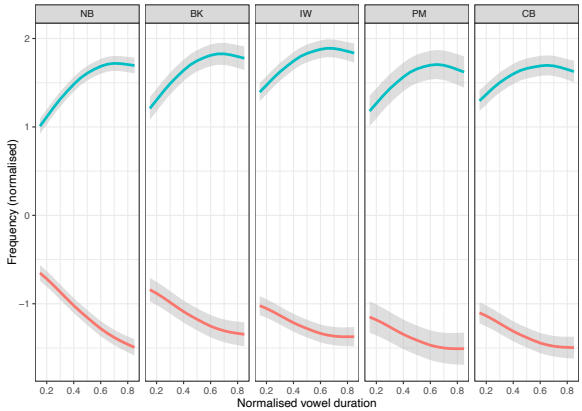


Figure 3: Model predictions of normalised F1 (red lines) and F2 (blue lines) of FLEECE by normalised vowel duration for speakers from five areas of Sydney. Error ribbons show 95% CIs.

Another important aspect of this analysis is that the vowels in FACE and FLEECE appear to pattern together in that greater onglide for FLEECE (non-linear F1 effect for NB compared to the other groups) occurs with a lowered first element of FACE in the NB speakers (both parametric and non-linear differences between NB and other groups in F1). Conversely, for the speakers from areas of high diversity, reduced FLEECE onglide patterns with a more raised first element of FACE. The relationship between these two vowels in chain shifting of the type described by [25] and [23] is a fruitful area for future research.

Finally, we note that this analysis is preliminary, and that a limitation is that some of the groups have fewer speakers, and hence fewer items, than others. Data collection for this project is ongoing, and in our future research we aim to broaden our scope to include both female and male speakers, and to extend the analysis to the entire AusE diphthong system.

5. ACKNOWLEDGEMENTS

This research was supported by Australian Research Council Future Fellowship Grant FT180100462 to the second author.

6. REFERENCES

[1] Albin, A. 2014. PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language. *J. Acoust. Soc. Am 135*, 2198.
[2] Australian Bureau of Statistics. 2016. *Census 2016*.
[3] Cheshire, J., Fox, S., Kerswill, P., Torgersen E. 2008. Ethnicity, friendship network and social practices as the motor of dialect change: Linguistic innovation in London. *Sociolinguistica 22*, 1–23.
[4] Cheshire, J., Kerswill, P., Fox, S., Torgersen E. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *J. Sociolinguistics 15*, 151–196
[5] Clothier, J. 2020. Ethnolectal variability in Australian Englishes. In: Willoughby, L, Manns, H., (eds.), *Australian English reimagined*. Routledge, 155–171.
[6] Clyne, M., Eisikovits, E., Tollfree, L. 2001. Ethnic varieties of Australian English. In: Blair, D., Collins, P. (eds.), *English in Australia*. Benjamins., 223–238.
[7] Cox, F. 1999. Vowel change in Australian English. *Phonetica 56*, 1–27.
[8] Cox, F. 2020 Phonetics and phonology of Australian English. In: Willoughby, L, Manns, H., (eds.), *Australian English reimagined*. Routledge, 15–33.
[9] Cox, F., Palethorpe, S. 2007. Illustrations of the IPA: Australian English. *J. Int. Phon. Assoc. 37*, 341–350.
[10] Cox, F., Palethorpe, S. 2012. Standard Australian English: The sociostylistic broadness continuum. In: Hickey, R. (ed.), *Standards of English: Codified Varieties around the World*. Cambridge, 294–317.
[11] Cox, F., Palethorpe, S., Bentink, S. 2014. Phonetic archaeology and fifty years of change to Australian English /i:/. *Aust. J. Ling. 34*, 50–75.
[12] Cox, F., Palethorpe, S., Harrington, J. 2015. *Diphthong shift in Australian English,* presented at Australian Linguistic Society Conference, Sydney.
[13] Cox, F., Penney, J., Palethorpe, S. 2022. Fifty years of change to prevocalic definite article allomorphy in Australian English. *J. Int. Phon. Assoc, 1–31. doi:10.1017/S002510032200007X*
[14] Gates, S. M. 2018. Why the long FACE? Ethnic stratification and variation in the London diphthong system. *U. Penn. Working Papers in Linguistics 24*, 39–48.
[15] Grama, J., Travis, C. E., Gonzalez, S. 2020. Ethnolectal and community change ov(er) time: Word-final (er) in Australian English. Aust. J. Ling. 40, 346–368.
[16] Harrington, J. Gubian, M., Stevens, M., Schiel, F. 2019. Phonetic change in an Antarctic winter. *J. Acoust. Soc. Am. 146*, 3327–3332.
[17] Harrington, J., Cox, F., Evans. Z. 1997. An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Aust. J. Ling. 17*, 155–184.
[18] Horvath, B. 1985. Variation in Australian English: The sociolects of Sydney. Cambridge.
[19] Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language 45*, 326–347.
[20] Lobanov, B. 1971. Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am. 49*, 606–608.
[21] Purser, B., Grama, J., Travis, C. 2020. Australian English over time: Using sociolinguistic analysis to inform dialect coaching. *Voice and Speech Review 14,* 269–291.
[22] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
[23] Trudgill, P. 2004. *New-dialect formation: The inevitability of colonial Englishes*. Edinburgh.
[24] van Rij, J., Wieling, M., Baayen, R. H., van Rinj, H. 2022. Itsadug: Interpreting time series and autocorrelated data using GAMMs. https://CRAN.R-project.org/package=itsadug.
[25] Wells, J. C. 1982. *Accents of English*. Cambridge.
[26] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language 45*. 392-410.
[27] Wood, S N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. Royal Stat. Soc. 73,* 3–36.

IMPLICIT VS. EXPLICIT PERCEPTION OF FRENCH OPTIONAL LIAISON AS A MARKER OF FORMALITY

Megan M. Dailey, Sharon Peperkamp

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS-PSL, CNRS, EHESS)
megan.dailey@ens.psl.eu, sharon.peperkamp@ens.psl.eu

ABSTRACT

This study examines the effect of explicit awareness (or, *noticing*) of socially-indexed variation on the generation of social inferences during speech perception. In a matched-guise task, French listeners associated sentences that contained optional liaison—a sociolinguistic variable at the level of *stereotype* in French—more frequently with 'formal language' than with 'everyday language'. This effect was larger for participants who reported noticing the optional liaisons than for participants who did not.

These results demonstrate that while noticing variation in context is not a necessary condition for social inferences to be generated, language users' subjective experience of socially-conditioned variation influences the frequency with which social inferences arise, which may ultimately lead to individual differences in the representation of the same socially-indexed variation.

Keywords: sociophonetics, speech perception, awareness, individual differences, French liaison.

1. INTRODUCTION

Social inferences about a speaker or context arise rapidly, automatically, and implicitly during speech perception [1]. At the same time, language users have explicit knowledge or awareness of some, but not all socially-indexed variation in the speech signal [2], and the extent of this knowledge varies by sociolinguistic variable (i.e., an association of linguistic form and social category). Labov [2] defines a scale of awareness of sociolinguistic variables that distinguishes *indicators*, of which language users have no awareness, *markers,* of which they have implicit awareness, in that they generate social inferences upon encountering a marker without explicit awareness, and *stereotypes*, of which they have explicit awareness and which are often the subject of metalinguistic commentary.

There is a large body of work dedicated to describing language users' knowledge of sociolinguistic variation [3]; additionally, evidence from sociophonetic perception studies has demonstrated that social information given to listeners about a speaker or context can bias speech perception [4-5]. Yet little is known about the effect of explicit awareness of socially-indexed variation on social inferences made during language processing (described as *noticing* by Schmidt [6]).

In a study on noticing using a self-paced reading task, Squires [7] compared listener awareness of standard, dialectal and unattested subject-verb agreement forms. Results showed that the different-dialect form [singular-NP + *don't*], a variable at the level of *stereotype* in American English, impeded reaction times of participants who reported noticing the form in a follow-up questionnaire to a greater extent than those who did not report noticing the form, though both groups exhibited slower reaction times to non-standard forms than to standard forms. Therefore, while dialectal and unattested variation affected the performance of both groups, it was noticing that determined the speed at which variants were processed.

In this paper, we aim to quantify the effect of noticing on the generation of social inferences related to it. Our case study involves French liaison, a complex external sandhi phenomenon in which certain word-final consonants are produced only when the following word starts with a vowel. In French, liaison is obligatory in some contexts (e.g., between an article and a following noun, as in *les‿idées* [le**z**ide] 'the ideas'), and optional in others (e.g., between a plural noun and a following adjective, as in *idées‿intéressantes* [ide**z**ɛ̃teʁesɑ̃t] or [ideɛ̃teʁesɑ̃t] 'interesting ideas'). The variation induced by optional liaison is socially structured: corpus studies of French have found higher rates of optional liaison in formal speech (radio news broadcast) than in casual speech (conversations between friends) [8-9]. These empirical observations back a theoretical hierarchy of discourse contexts in which different optional liaisons would most likely be deployed [10].

Optional liaison is a sociolinguistic variable at the level of *stereotype* in Metropolitan French. It is associated with a hyper-standard, prestige variety of French learned in school and linked to political, legal, academic and business language [11]. Although optional liaison is a robustly studied phenomenon, there is to our knowledge no experimental data showing that French listeners systematically link the

presence of optional liaison with formality[1]. The goal of this study, then, is two-fold: we examine, first, whether French listeners reliably associate optional liaison to formality during speech perception, and, second, whether their noticing of optional liaison mediates the strength of this association. To do so, we employ a matched-guise paradigm [13] and a post-experiment questionnaire probing explicit awareness.

2. METHODS

2.1. Materials

Stimuli were 17 short, descriptive French sentences, each with a single optional liaison environment. They had simple syntactic structure and contained neutral observations or well-known facts. Sentences were constructed such that obvious lexical and other phonetic cues that signal discourse (in)formality in French were absent. In particular, they contained no environments for schwa-deletion or post-obstruent liquid deletion (both of which signal casual speech in French [14]), no singular second-person pronouns (*tu* or *vous*, depending on relational proximity), no first-person plural pronouns (*on* or *nous,* depending on level of discourse formality), and – with one exception[2] – no negations (e.g., *pas* or *ne…pas* 'not', depending on level of discourse formality). An example, with the optional liaison environment underlined and bolded, is shown in (1). A full list of sentences can be found in the Appendix.

(1) *Les plat**s**‿**i**taliens sont très populaires*.
Italian dishes are very popular.

Sentences were recorded in a soundproof booth by a male native speaker of French with state-of-the-art recording equipment. The speaker was instructed to produce each sentence both with and without liaison, at a normal speaking rate and by keeping the productions as similar as possible for all aspects except liaison. Both guises were recorded 2-3 times, and the best takes were selected for subsequent manipulation in Praat [15].

Sentences were first scaled to match in average intensity, and then manipulated in the following manner: For each sentence, the two-word sequence that constituted the optional liaison environment was spliced out of both guises. Pitch contours and duration were then matched across guises, and the resulting sequences were both spliced back into the sentence originally produced in the guise without liaison. As a result, the two guises of a sentence differed only in the presence or absence of liaison in the crucial two-word sequence.

Two counterbalanced stimulus lists were created, varying which sentences were presented with and without liaison in each, such that no sentence appeared in the same list twice. The number of sentences with and without liaison was also balanced across lists, such that in the first list, there were 8 sentences with liaison and 9 sentences without, and vice versa for the second list.

2.2. Participants

Participants were 109 native speakers of Metropolitan French (age range: 19-60, median: 26), located in France and recruited on Prolific (www.prolific.co). Data from three additional participants who were speakers of other varieties of French were collected but not analyzed.

2.3. Procedure

The experiment was run on the experiment-hosting platform Labvanced [16]. Participants were instructed to sit in a quiet room and use headphones to complete the task. They were randomly assigned to one of the two lists. Within the relevant list, sentences were presented in a random order.

In each trial, participants listened to a sentence and subsequently provided a binary judgment on its formality: *langage soutenu* – "formal language" or *langage courant* – "everyday language". In order to minimize the effect of semantic content on formality judgments, participants were instructed to pay more attention to what each sentence sounded like than to its meaning.

At the end of the experiment, participants completed a questionnaire on their experience with the task as well as their language background and demographic characteristics. Two questions, shown in (2), probed participants' explicit awareness of optional liaison in the sentence stimuli.

(2) *a. Quelle stratégie avez-vous utilisée pour faire vos choix ?*
What strategy did you use to make your choices?
b. Y a-t-il un aspect des phrases qui a particulièrement attiré votre attention ?
Was there an aspect of the sentences that attracted your attention in particular?

3. RESULTS

Participants were first assigned to one of two groups based on their mention of optional liaison in responses to the two relevant questions in the post-experiment questionnaire. Those that included the word *liaison* in at least one of their responses to the

two relevant questions above or gave an example of an optional liaison they heard in the experiment were coded as Mentioners (n = 32), and those that did not were coded as Non-Mentioners (n = 77).

Formality ratings were coded as 1 for *langage soutenu* 'formal language' and as 0 for *langage courant* 'everyday language'. The overall mean formality rating was 0.31; hence, sentences were rated as 'everyday language' more than as 'formal language'. Table 1 reports mean formality ratings by Liaison and Mention status, and Figure 1 shows individual difference scores of formality ratings for sentences with minus sentences without optional liaison.

Table 1. Mean formality ratings by Liaison and Mention status. Standard errors are shown in parentheses.

Mention	Liaison	No Liaison
Yes	0.55 (0.09)	0.24 (0.05)
No	0.32 (0.08)	0.23 (0.05)

We performed a logistic mixed effects regression with fixed effects Liaison, Mention, and their interaction, random intercepts for Participant and Sentence pair, and a random by-sentence slope of Mention. The analysis was carried out with the lme4 package [17] in the R environment [18]; significance was established by means of model comparison and a likelihood ratio test, and the random structure was the maximal one that allowed for convergence while avoiding overfitting. Results revealed main effects of Liaison ($\beta = 0.60$, $SE = 0.07$, $z = 8.83$, $Chi^2(1) = 83.5$, $p < 0.001$) and Mention ($\beta = 0.37$, $SE = 0.17$, $z = 2.17$, $Chi^2(1) = 4.58$, $p < 0.04$), and a significant interaction between the two ($\beta = 0.29$, $SE = 0.07$, $z = 4.19$, $Chi^2(1) = 17.8$, $p < 0.001$).

Sentences with liaison were rated as more formal than sentences without liaison, and Mentioners rated sentences as more formal overall than Non-Mentioners. Further, the difference in rating for sentences with and without liaison was bigger for Mentioners than for Non-Mentioners.

4. DISCUSSION

In a matched-guise task, French adult listeners were sensitive to the presence of optional liaison as a sociolinguistic cue, linking it to formal language. In other words, the presence of optional liaison biased interpretations of sentence formality even in the absence of noticing, indicating that French listeners have implicit knowledge of optional liaison as a marker of formality. Furthermore, listeners who reported noticing optional liaison in the sentences they listened to made this association more frequently than those who did not. Based on our finding that noticing socially-conditioned variation leads to more frequent generation of social inferences related to it, we hypothesize that listeners who tend to notice socially-indexed variation during speech perception

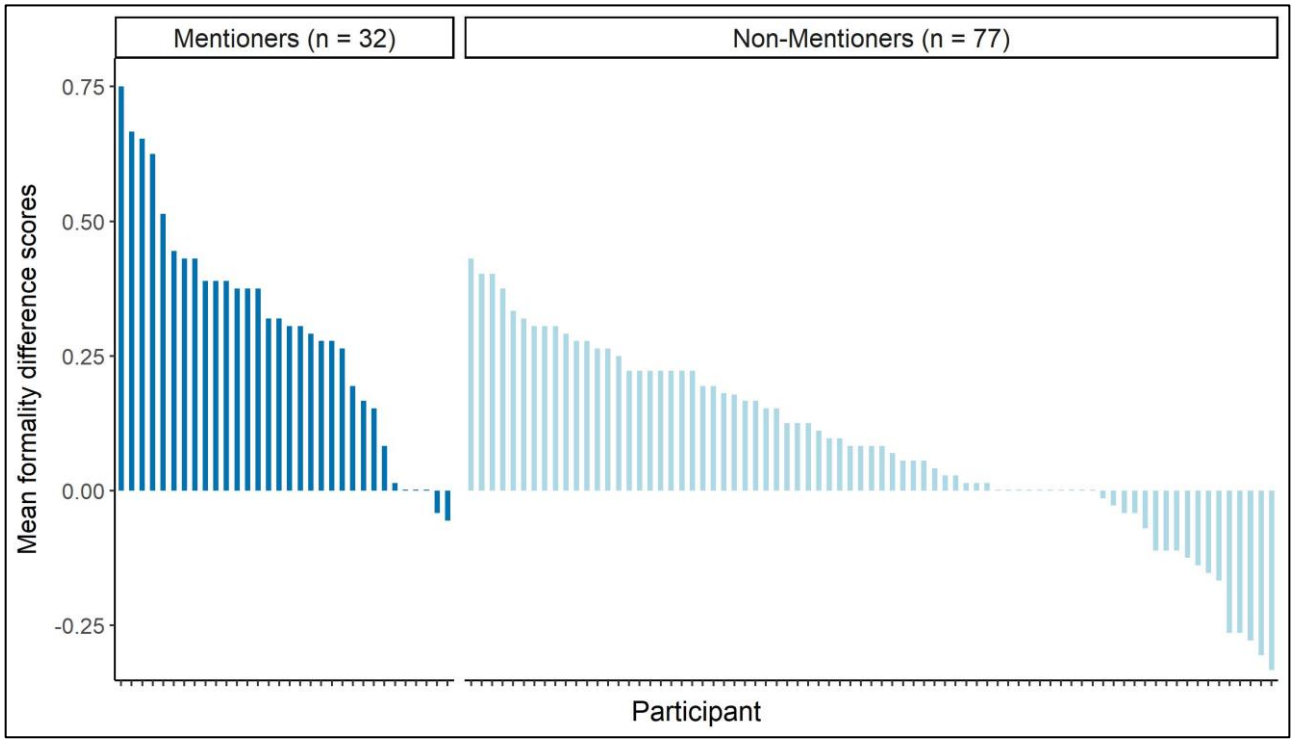


Figure 1. Mean individual difference scores: formality ratings for sentences with minus for sentences without optional liaison for Mentioners (left) and Non-Mentioners (right).

have stronger associations between acoustic cues and social categories.

The hypothesis that explicit awareness of, or attention to, socially-conditioned variation could amplify the robustness of associations between acoustic cues and social categories is not a new one. Specifically, it has been proposed in so-called dual route models of speech perception [19]; these models view interaction between social and linguistic information as an integral part of language processing. Though intuitive, it has as far as we know not been tested directly. This hypothesis could be tested in future research, especially with a wider range of methods for assessing listener awareness, including EEG [20].

Finally, the results of the current study highlight the importance of probing individual awareness of socially-indexed variation in language processing experiments. Going a step further, it would be interesting to investigate which factors account for individual differences observed in noticing socially-indexed variation during speech perception. For the case at hand, given that optional liaison is linked with political, legal, academic and business language, one may wonder if level of education plays a role. Our sample was skewed in that two thirds of participants had at least a college degree, and several of the remaining participants were college students. For what it's worth, the percentage of Mentioners did not differ between those who had a college degree and those who did not ($t < 1$). More generally, cognitive factors such as working memory and social factors such as autistic-like traits have been shown to influence sociolinguistic perception to some degree [21-22]. Whether factors such as these correlate with noticing is a promising avenue for future inquiry.

APPENDIX

A full list of sentence stimuli, each with one optional liaison environment (bolded).

1. Les népalais sont‿**h**abitués au froid.
2. Les légumes sont vraiment‿**e**xcellents pour la santé.
3. Les bars ont‿été fermés pendant longtemps.
4. Les livres‿**a**nciens sont vendus en ligne.
5. La viande n'est pas‿**a**u menu tous les jours.
6. Les canicules seront‿**a**ssez fréquentes.
7. Les enfants peuvent‿**a**pprendre plusieurs langues.
8. Les américains sont vraiment‿**a**llés sur la lune.
9. La statue vivante était parfaitement‿**i**mmobile.
10. Plusieurs pays veulent‿**o**rganiser la prochaine Coupe du Monde.
11. Les couteaux suisses sont très souvent‿**u**tilisés au camping.
12. La victoire était totalement‿**i**nattendue.
13. Le Canada a deux langues‿**o**fficielles.
14. Les koalas et les kangourous sont des animaux‿**a**ustraliens.
15. Les jeunes bébés boivent‿**u**niquement du lait.
16. Le portugais est parlé au Portugal mais‿**a**ussi au Brésil.
17. Les plats‿**i**taliens sont très populaires.

6. REFERENCES

[1] Drager, K. & Kirtley, M. J. 2016. Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In: Babel, A. (ed), *Awareness and Control in Sociolinguistic Research.* Cambridge University Press, 1-10.

[2] Labov, W. *Sociolinguistic Patterns.* Philadelphia: Philadelphia University Press, 1972.

[3] Campbell-Kibler, K. 2010. Sociolinguistics and perception. *Language and Linguistics Compass* 4/6 377-389.

[4] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1), 62-85.

[5] Drager, K. 2010. Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4(7), 473-480.

[6] Schmidt, R. W. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11, 129-158.

[7] Squires, L. 2016. Processing grammatical differences: Perceiving versus noticing. In: Babel, A. (ed), *Awareness and Control in Sociolinguistic Research.* Cambridge University Press, 80-103.

[8] Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E. & Lamel, L. 2012. La liaison dans la parole spontanée familière : Une étude sur grand corpus. *Revue Française de Linguistique Appliquée* 17, 113-128.

[9] Hutin, M., Weng, C., Adda-Decker, M. & Lamel, L. 2022. La liaison facultative en français : étude de grands corpus combinant approche automatique relâchée et jugement perceptif. SHS Web of Conferences 138, *Congrès Mondial de Linguistique Française – CMLF 2022.*

[10] Delattre, P. 1947. La liaison en français, tendances et classification. *The French Review* 21(2), 148-157.

[11] D. Hornsby. *Norm and Ideology in Spoken French: A Sociolinguistic History of Liaison.* Cham, Switzerland: Springer, 2020.

[12] Dautricourt, R. G. French liaison: Linguistic and sociolinguistic influences on speech perception. Ph.D. dissertation, Dept. of Linguistics, The Ohio State Univ., Columbus, OH. 2010.

[13] Lambert, W. E. 1967. The social psychology of bilingualism. *Journal of Social Issues* 23, 91-109.
[14] N. Armstrong. *Social and Stylistic Variation in Spoken French.* Amsterdam: John Benjamins, 2001.
[15] Boersma, P. & Weenink, D. 2019. *Praat: Doing phonetics by computer* [computer program]. Available at: http://www.praat.org (accessed April 18, 2019).
[16] Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. 2017. LabVanced: a unified JavaScript framework for online studies. *International Conference on Computational Social Science,* Cologne.
[17] Bates, D., Maechler, M., Bolker, B., & Walker, S. 2015. Fitting linear mixed effects models using lme4. *Journal of Statistical Software* 67, 1–48.
[18] R Core Team (2020) *R: A language and environment for statistical computing* [software]. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org (accessed January 2021).
[19] Sumner M., Kim, S. K., King, E. & McGowan, K. B. 2014. The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology* 2014(4), Article 1015.
[20] Hanulikova, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. 2012. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience* 24(4), 878-887.
[21] Lev-Ari, S., Dodsworth, R., Mielke, J. & Peperkamp, S. 2019. The different roles of expectations in phonetic and lexical processing. *Proceedings of INTERSPEECH,* Graz, Austria, 2305-2309.
[22] Hesson, A. & Shellgren, M. 2015. Discourse marker *like* in real time: Characterizing the time-course of sociolinguistic impression formation. *American Speech* 90(2), 154-186.

[1] A little differently, Dautricourt [12] shows that Metropolitan French listeners associate optional liaison with high socio-economic status (SES) speakers.

[2] In this sentence (Sentence 5, Appendix), deletion of *ne* is highly unlikely due to the surrounding phonological context.

PRONUNCIATION OF GENDER-NEUTRAL GERMAN WITH RESPECT TO SPEAKER ATTITUDE

Malte Belz, Miriam Müller, Christine Mooshammer

Humboldt-Universität zu Berlin, Germany
malte.belz@hu-berlin.de

ABSTRACT

In recent years the pronunciation of gender-neutral forms has been a matter of highly controversial discussions in German media. In this reading study we investigate the pronunciation variants of orthographically marked gender-neutral forms of four groups of native speakers of German (age: younger and older, gender: women and men) depending on their attitude towards the usage of gender-neutral forms. Younger and older women mark about 76% of the tokens phonetically or syntactically, whereas the group of younger men marked 72% and the older men only 44%. The most frequent phonetic markers are glottalisations and glottal stops, followed by combinations of these and pauses, but a wide range of other variants was also found. There is a tendency that speakers with a positive attitude towards gender-neutral forms use them more frequently, but there are also speakers whose attitude and usage are inconsistent.

Keywords: sociophonetic variation, gender-neutral language, read speech, attitude

1. INTRODUCTION

German is a language with grammatical gender, in which the so-called generic masculine has historically been used to refer to persons of any gender. In recent years generic masculine forms have come under criticism for having a male bias [1, 2]. In order to make women and other genders not only socially but also linguistically visible, an early effort was to substitute generic masculine nouns in favour of pair forms as in *Köchinnen und Köche* 'chefs.PL.MASC and chefs.PL.FEM' In recent years, as an alternative strategy that may be interpreted to be more inclusive, gender-neutral nouns are marked in writing with glyphs such as the asterisk (e.g., *Köch*innen* 'chefs.PL.NEUT') or the colon (*Köch:innen* 'chefs.PL.NEUT') [3]). How these glyphs (e.g., *Köch*innen*) are pronounced in spoken language was unclear until, in recent years, the insertion of a pause and/or a glottal stop has become more and more established as a strategy to mark gender-neutral forms in oral language. To the best of our knowledge, up to now only Slavik et al. [4] investigated the realisation strategies and phonetic features of gender-neutral forms. They found that the most frequent marker for gender-neutrality was a lengthening of the /ɪ/, often accompanied by a stress shift on the suffix. Pauses and glottal stops were far less frequent.

In this paper we explore the various strategies for the pronunciation of gender-neutral forms and how these are phonetically realised. Furthermore, we investigate whether the attitude of the speaker towards gender-neutral language affects the choice of the pronunciation strategy.

RQ 1: What pronunciation strategies (feminine, masculine, pair form, neutral, other) do speakers of German use when reading gender-neutral forms?

RQ 2: What are the most frequent phonetic markers for gender-neutrality?

RQ 3: Does speakers' attitude towards gender-neutrality affect gender-neutral pronunciation?

2. METHOD

In the corpus of gender-neutral language, KGGS [5], the pronunciation strategies were investigated in a reading study. Participants were instructed to read sentences aloud that included nouns marked as gender-neutral. The attitude of the participants towards gender-neutral language was collected in a post-experimental questionnaire.

2.1. Participants

40 native German speakers from across Germany participated in the study, identifying themselves as male (n=21) and female (n=19). The younger group consisted of 26 speakers, ranging between 21 and 34 years of age (13 male, 13 female), and 14 speakers belonged to the older group, ranging between 47 and 83 years of age (8 male, 6 female). Except for three speakers, all qualified for university entrance. Many also graduated from institutions of higher education.

2.2. Material

The stimuli consisted of twelve test words (see Table 1) embedded in sentences in three conditions (masculine, feminine and gender-neutral), each in singular or plural form. Only the gender-neutral forms will be considered here. The gender-neutral forms were marked with one of six possible glyphs `I * : _ () /` and varied between subjects. As in the current corpus a maximum number of two participants per gender and age group produced each glyph list, the different glyphs are not considered as a condition. A total of 960 items were included in the analysis (12 test words x 2 numbers x 40 speakers), of which two instances were excluded due to technical errors. Thus, 958 instances of gender-neutral forms were included in this study.

2.3. Procedure

The stimulus sentences were prompted one by one on a screen for the participants to read aloud. After the reading experiment, they filled out a questionnaire about their personal attitude towards gender-neutral language. As a parameter for the participants' attitude, we used a Likert scale that assigned values of 1 (= 'disagree' to 5 (= 'agree') to the participants' responses to the following assertion: 'You find it important and necessary to use gender-neutral forms, e.g., students.PL.NEUT'.

2.4. Data annotation

The data were annotated on several interval tiers in Praat [6] by one of the authors, based on acoustic and auditory cues. Unclear cases were discussed by all authors. For the realisation of the gender marker (GM) we defined GM as the interval after the offset of the preceding segment and before the /n/ of the suffix *in(nen)*, including the vowel /ɪ/ as well as pauses or glottal stops. For gender-neutral realisations the subsegmental parts preceding the segment /ɪ/ were annotated for pauses (#), single glottal stops (?), partial glottalisation (q), and glottalisation on the whole vowel (qI). The segment /ɪ/ was annotated for modal voice vowel (I). Unexpectedly, in some cases the vowel was realised with breathy voice which was marked by t_I. Silences before glottal stops are only annotated if they were longer than 40ms, accounting for the closure of the stop [7]. Otherwise, no silence threshold was used.

Stops preceding the GM which were perceived as longer than usual were marked with an additional (#), e.g., #d. On a separate tier, we annotated the perceived strategy employed by the speaker for expressing gender-neutral forms. The annotation values consisted of *n* if the target word was judged as gender-neutral by the annotators, of *f* if the feminine form was perceived, and of *s* for syntactic expansions, including coordinating conjunctions (e.g., *Kolleginnen und Kollegen* 'colleagues (female and male)', and asyndetic conjunctions (e.g., *Kollegen Kolleginnen*).

test words (feminine)	test words (neutral)	translation (feminine)
Königin	*König* □ *in*	queen
Köchin	*Köch* □ *in*	chef
Freundin	*Freund* □ *in*	friend
Probandin	*Proband* □ *in*	subject
Studentin	*Student* □ *in*	student
Referentin	*Referent* □ *in*	speaker
Zeugin	*Zeug* □ *in*	witness
Kollegin	*Kolleg* □ *in*	colleague
Bäuerin	*Bäuer* □ *in*	farmer
Fördererin	*Förderer* □ *in*	patron
Doktorin	*Doktor* □ *in*	doctor
Autorin	*Autor* □ *in*	author

Table 1: Overview of the test words in their feminine and gender-neutral forms. Only the singular forms are shown here. For the plural forms the suffix *-en* [ən] is attached, e.g., *König*innen*. The box □ represents the six gender-neutral glyphs `I * : _ () /`.

2.5. Analysis

For the phonetic analysis, the annotation values of the gender marker were grouped into four overarching categories, which are *glott* (glottal activity such as glottal stops, glottalisation or breathy voice before or during /ɪ/), *sil* (silence), *combi* (a combination of glottal activity and pauses), and *none* (no additional marker in the acoustic signal). For a qualitative analysis of attitude towards marking gender-neutrality, we assigned the values 1 and 2 to a negative attitude, 3 to an indifferent attitude, and 4 and 5 to a positive attitude.

3. RESULTS

3.1. Pronunciation strategies

Regarding RQ 1 (what pronunciation strategies do speakers of German use when reading gender-neutral forms?), women mainly choose the strategy of using gender-neutral forms, followed by the strategy of pronouncing the gender-neutral forms the way feminine forms are pronounced (s. Table 2).

Men prefer the neutral and the feminine strategy as well, but also tend to include other strategies such as pair forms. Some replace the gender-neutral form with the masculine form (s. Table 2). Age alone does not affect the chosen strategy ($\chi = 2.1, p = 0.15$). However, there is a noticeable difference between older men and the other speaker groups. Only about half of the older men chose a strategy other than the feminine strategy, while all other speaker groups mainly realise neutral or pair forms in approx. 75% of the gender-neutral stimuli. Singular or plural number does not seem to exert a strong influence on the chosen strategy (s. Figure 1).

Table 2: Proportions of pronunciation strategy per gender and age.

	neutral	pair form	feminine	masculine
fem.younger	75.9	0.3	23.8	0.0
fem.older	75.7	0.0	24.3	0.0
male.younger	61.7	10.6	27.3	0.3
male.older	42.2	2.1	53.1	2.6

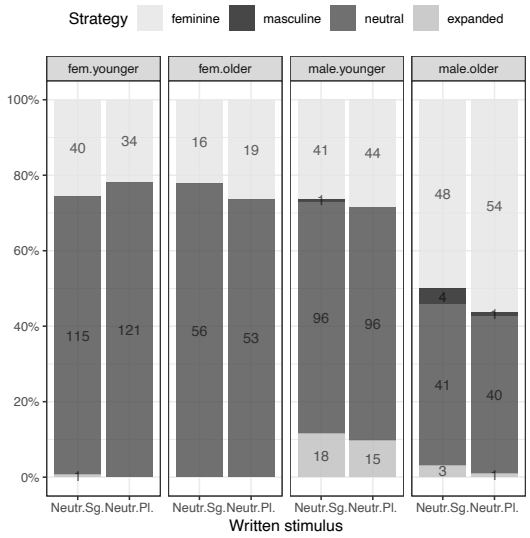


Figure 1: Pronunciation strategy of gender-neutral singular/plural target words by speaker gender and age.

3.2. Phonetic markers

The most frequent phonetic markers for gender-neutrality (n = 617) are glottalisations (53.6%), followed by combinations (29.8%), no marker (12.8%), and silences (3.7%). Breathy voice was observed for 3% of the data and only once occurred together with glottalisation. Table 3 shows the proportions per gender and age groups.

Table 3: Proportion of phonetic markers per age and gender group. Participants use glottalisation and glottal stops (glott), combination of glottalisation and glottal stops with silence (combi), silence (sil), or no markers (none).

	glott	combi	sil	none
fem.younger	80.5	14.0	2.1	3.4
fem.older	18.3	69.7	2.8	9.2
male.younger	53.9	27.2	6.3	12.6
male.older	22.2	28.4	3.7	45.7

3.3. Influence of the speaker's attitude

Table 4 shows the results from the questionnaire regarding the attitude of the participants towards the usage of gender-neutral forms. Except for the group of older men the majority of participants states that they have a positive attitude. In the older generation the attitude is more often negative for men and women compared to the younger generation where only 7.7% disagree with the usage of gender-neutral forms. Regarding gender, male participants are more often indifferent than female participants.

Table 4: Proportion of positive, indifferent, and negative attitude per age and gender group.

	negative	indifferent	positive
fem.younger	7.7	7.7	84.6
fem.older	33.3	0.0	66.7
male.younger	7.7	23.1	69.2
male.older	25.1	37.3	37.6

Figure 2 shows the proportion of gender-neutral forms relative to all realized forms per speaker depending on attitude, shown as emojis. No clear trend can be observed, which is corroborated by a generalised linear mixed-effects model with strategy as the dependent variable, attitude as the independent variable, and speakers and words as random intercepts. Attitude as speaker-specific random slope did not improve the model significantly. No significant effect was found for different test words. Attitude is not affecting the gender-neutral strategy significantly ($\beta = -0.1$, std. error = 0.1, p = 0.5). Accordingly, the mean proportion of neutral and extended forms amounts to 74.5% for speakers indicating a positive attitude towards gender-neutral forms, to 63.9% for speakers with a negative attitude, and to 49.5% for speakers with an indifferent attitude.

4. DISCUSSION

In summary, German speakers in this study show a clear effect of gender and age when reading written gender-neutral stimuli. Younger female speakers use the neutral forms most frequently, followed by older females, younger males, and older males. This

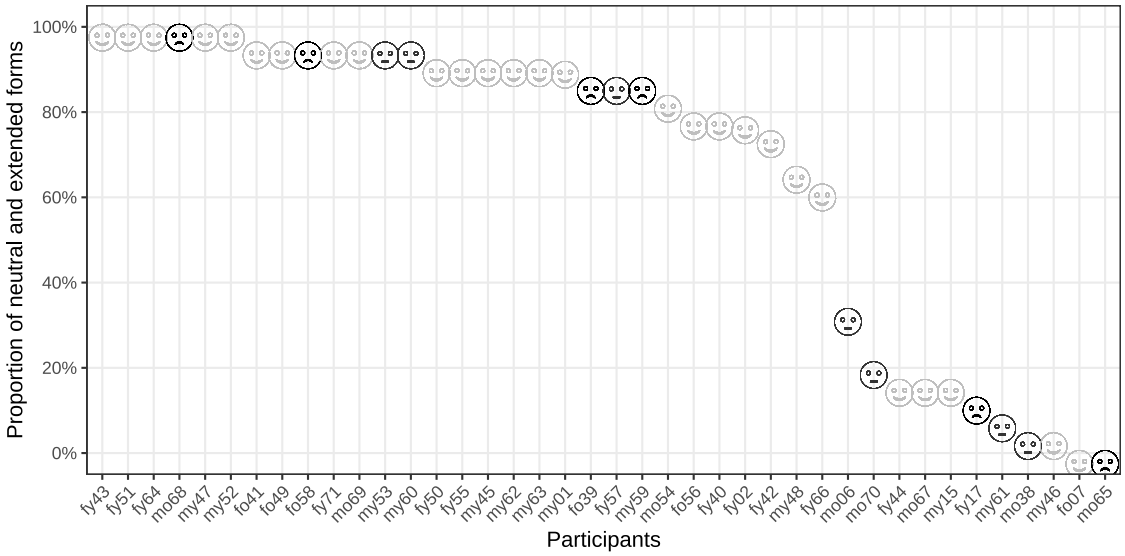


Figure 2: Proportion of gender-neutral and pair forms for attitude (1 and 2 = frowning face, 3 = neutral face, 4 and 5 = smiling face) per participant. The first two letters of the participant ID refer to their gender group (fy = female younger, fo = female older, my = male younger, mo = male older).

trend is in line with sociophonetic findings on the innovative potential of younger women in language change [8, 9].

Phonetically, glottalisations (including the singleton glottal stop as a subgroup) are most common when realising gender-neutral words, followed by combinations consisting of glottalisation and silences. Compared to Slavik et al. [4], a more detailed picture emerges, as inserted glottalisations and silences might be the driving factor for the gender marker at the position of /ɪ/.

The strategy of pronouncing written gender-neutral forms neutrally or as a pair form have been already discussed by [4] as possibilities to mark gender-neutrality. The feminine strategy (i.e. realising the neutral form without a perceivable gender marker on /ɪ/) probably arises because the written gender-neutral form looks quite similar to the feminine form, but with a glyph added. Additionally, the gender-neutral forms perceived as feminine may come about as speakers either wanted to mark it phonetically and failed, or that they applied other means, such as word accent or loudness. The decision to use a masculine strategy, in contrast, seems to be either a deliberate decision, as a part of the written word is deliberately omitted, or an epiphenomenon of being insecure about the neutral pronunciation, or even a signal of attitude towards gender-neutral language.

However, contrary to our expectations, attitude does not affect the realisation of gender-neutral language significantly. The data show both participants with a positive attitude realising few neutral forms, and, vice versa, participants with a negative attitude realising many neutral forms. Nevertheless, the data show a general tendency that speakers with a positive attitude towards gender-neutral forms use them more frequently, followed by speakers with a negative attitude. Speakers with indifferent attitudes use gender-neutral forms the least. Since mostly men are indifferent towards the usage of gender-neutral forms this could also be a confound with gender.

In conclusion, the majority of speakers investigated in this study are capable of and aiming to pronounce the glyphs for the gender-neutral suffix as expected, namely with a glottal stop, glottalisation and/or a pause in-between. Except for the group of older men, these phonetic features seem to be well-established markers for gender-neutrality in read German. In future studies we plan to include that word stress and the perception of the neutral strategy.

5. ACKNOWLEDGMENTS

We thank Alina Zöllner for help with the recordings and annotation. We are thankful to the Department of German Studies and Linguistics for financial support.

6. REFERENCES

[1] Braun, F., Gottburgsen, A., Sczesny, S., Stahlberg, D. 1998. Können Geophysiker Frauen sein? Generische Personenbezeichnungen im Deutschen. *Zeitschrift für germanistische Linguistik* 26, 265–283.

[2] Schunack, S., Binanzer, A. 2022. Revisiting gender-fair language and stereotypes – a comparison of word pairs, capital i forms and the asterisk. *Zeitschrift für Sprachwissenschaft* 41(2), 309–337.

[3] Kotthoff, H. 2020. Gender-Sternchen, Binnen-I oder generisches Maskulinum,...(Akademische) Textstile der Personenreferenz als Registrierungen? *Linguistik Online* 103(3), 105–127.

[4] Slavik, K., Cronenberg, J., Draxler, C. 2018. A study on the pronunciation of gender-neutral nouns in German. Belz, M., Mooshammer, C., Fuchs, S., Jannedy, S., Rasskazova, O., Żygis, M. (eds), *Proceedings of the Conference on Phonetics & Phonology in German-speaking countries (P&P 13)*, 185–188.

[5] Mooshammer, C., Belz, M. 2021. Korpus gelesener geschlechtergerechter Sprache (KGGS), Version 2 [= Corpus of read gender-fair language]. https://rs.cms.hu-berlin.de/phon.

[6] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [computer program].

[7] Garellek, M. 2013. *Production and perception of glottal stops*. PhD thesis UCLA.

[8] Eckert, P. 1989. The whole woman: Sex and gender differences in variation. *Language variation and change* 1(3), 245–267.

[9] Eckert, P. 1997. Gender and sociolinguistic variation. In: Coates, J. (ed), *Language and gender: A reader*. Oxford: Blackwell, 64–75.

BEING NORTH KOREAN IN KOREATOWN: THE ROLE OF IDENTITY & EXPERIENCE IN ACCOMMODATION IN THE LONDON NORTH KOREAN COMMUNITY

Bronwen G. Evans, Rakgyun Shin & Justin Lo
Department of Speech, Hearing & Phonetic Sciences, University College London, London, U.K.

bronwen.evans@ucl.ac.uk, rakgyun.shin.21@ucl.ac.uk, justin.lo@ucl.ac.uk

ABSTRACT

North Korean refugees living in South Korea frequently adopt a Standard Seoul accent to avoid stigmatization, but in New Malden (London, UK), they are an active part of the large Korean community. In this study, we investigate accommodation in North Koreans in New Malden, focussing on the vowel /ʌ/, which is produced differently in North and South Korean accents.

Results demonstrate that whilst age of arrival in South Korea affects production of /ʌ/, identity drives the outcome of accommodation. In read speech, those who left North Korea before age 16yrs (early leavers) patterned more closely with South Koreans than did late leavers. However, in a conversational task, both groups diverged from rather than converged to their South Korean interlocutor. Interview responses showed a clear sense of solidarity within the New Malden North Korean community and we argue that these speakers used this variable to reclaim and index their North Korean identity.

Keywords: accommodation, D2 acquisition, sociophonetics, speech production, Korean.

1. INTRODUCTION

It is well-known that speakers of different regional accents modify their speech when living in a community where that accent is not commonly spoken (e.g., [1]). Such behaviour typically manifests as convergence, where speakers modify their accent to sound more like that of speakers in their new community. This may be to make themselves more easily understood – speakers might avoid variants that are markedly regional or unusual to facilitate communication (e.g., [2]) – but may also be for social reasons (see [3] for a review). Speakers who have moved to a new community may adopt features of the accent of their new community in order to better fit in and perhaps even to show belonging to their new community (e.g., [4]). Indeed, such short-term accommodation has been hypothesized to be a basis for second dialect (D2) acquisition (e.g., [1]).

Changing one's accent may be particularly important in a community where speakers are both easily identifiable from their accent and discriminated against as a result of their background. In South Korea, North Koreans who have defected to the South are typically stigmatized as a result of their socio-political background [5]. To hide their identity and pass as non-North Korean, they often report trying to lose their distinctive northern accent and adopt Standard South Korean (SK), the prestige accent spoken in Seoul (e.g., [5]). Contemporary SK is typically described as having a 7-vowel system /i ɐ̟ a ʌ o u ɯ/, where /ɐ̟/ is the result of an /e/ and /ɛ/ merger that took place over the past century [6, 7]. North Korean (NK) accents, in contrast, have not undergone an /e/-/ɛ/ merger [8]. Further, in the northeast, where our participants came from, /ʌ/ is raised (i.e., produced with a much lower F1) such that it overlaps with /o/ [9].

The present study focuses on a group of NK speakers in New Malden, London, who initially defected to South Korea but who then sought asylum in the UK. Due to the stigma they experience and the difficulties they face in building relationships in South Korea, some North Koreans emigrate to other countries in pursuit of a better life [10]. New Malden is attractive because it has the largest concentration of Koreans in Europe (8,000–10,000; [11]) and, although dominated by South Koreans, is also home to the largest population of North Korean refugees in the world (approx. 1,000; [12]). These refugees typically work in the many Korean-run businesses and are relied upon by many South Koreans to staff their businesses [12], despite ongoing tensions between North and South Koreans. North Koreans have also become increasingly visible and active within the community, e.g., establishing their own community schools. In a community where the two Koreans sometimes make strategic alliances in order to survive as minority groups in the UK, New Malden presents a unique sociolinguistic environment for the Korean language that contrasts with that in South Korea.

Of interest in our study, is (1) whether NK speakers living in New Malden have acquired and continue to use SK-like vowels and (2) given the different relationship between North and South Koreans in New Malden, if they continue to

accommodate towards SK speakers in a conversational speech task (Diapix: [13]). We focus on /ʌ/, a vowel that differs markedly in NK and SK accents. Given that previous research has suggested that earlier is better in terms of native-like acquisition of a D2 [14], we were also interested in whether those who had left North Korea during their childhood used more native-like SK vowels than those who had left later in life, i.e., in early adulthood.

2. METHOD

2.1. Participants

Ten female NK speakers, aged 27–45yrs, were recruited through a Korean school in New Malden, which runs Saturday classes in Korean language and culture for the children of North and South Koreans. All NK participants were refugees who had sought political asylum in the UK: 5 had left North Korea before age 16 (early NK leavers; ENK) and 5 after age 16 (late NK leavers; LNK). The participants were known to the second author, a native SK speaker who had been living in New Malden for 1 year. This was important for recruitment; many NK speakers are concerned that taking part in research could result in reprisals for family who remain in North Korea and so great care was taken to reassure them about the aims of the project.

A male SK confederate was recruited for the Diapix task. This speaker was known to all NK participants. He was aged 53yrs, was born in Busan, South Korea and had spent his childhood in both Seoul and Busan before moving to Seoul for university and work. He moved to Bristol, UK for further study 12 years ago, and at the time of testing, had been living in New Malden for 6 years.

In addition, 4 SK-speaking female teachers from the school, aged 30–45yrs, completed the read speech tasks for comparison.

2.2. Stimuli & Materials

All recordings and interviews took place in a quiet room in a Korean community café in New Malden. Recordings were made using a Roland R-09 recorder at a sampling rate of 44.1 kHz, 16-bit resolution and were then transferred directly to a laptop for analysis.

Wordlist & Read passage. The wordlist consisted of 24 bisyllabic words; 8 monophthongs (/i e ɛ a o ʌ u ɯ/) covering the Korean vowel space in 3 phonetic contexts, /hVdV(C)/, /sVgV(C)/ and /kVbV(C)/. Vowels were embedded in the 1st syllable of each word. To enable the use of real words, the 2nd syllable of a small number of words (N=4) differed through addition of an extra consonant in the onset of the 2nd syllable. The read passage was the Korean translation of the Aesop fable, “The North Wind and the Sun”, which contains all vowels in multiple phonetic contexts.

Diapix. Three Diapix scenes, a Beach, Farm and Street scene, were selected from the DiapixUK materials [13] and adapted for use in this study. These scenes were chosen because it was possible to adapt them to include multiple instances of the target vowel, /ʌ/. Words in the pictures were changed into Korean script and object layers modified, e.g., by changing colours, so that all pictures included 3-5 items that were represented by words that included the target variable. All pictures contained 10 differences (reduced from 12 in DiapixUK [13]).

2.3 Procedure

Recordings. SK speakers, including the confederate, recorded the wordlist and read passage. NK speakers first recorded the wordlist and read passage. They were then introduced to the confederate before completing the Diapix task.

The wordlist was presented in a randomized order via PowerPoint. To avoid list effects, each word was presented on a separate slide. Participants were recorded producing 1 clean repetition of each word, i.e., good voice quality, no hesitation. The read passage was presented on a laminated sheet of paper. Participants were given time to read through the passage before being instructed to read the passage aloud at their normal reading speed.

The 3 Diapix scenes were completed in a randomized order. The experimenter explained that they had 10 minutes to spot the differences for each scene, that they were not to show each other their pictures but instead were to communicate verbally to describe the pictures. Participants were told to start in the top left corner of the picture and to work clockwise. Pictures were presented on laminated pieces of paper; participants were instructed to circle the differences using the marker pen provided and to shout “all found” when they had finished spotting the differences. As soon as they had spotted all the differences they were instructed to move to the next scene. If they did not finish within the allotted time, they were stopped by the experimenter and told to move on to the next scene (N=10). Before completing the experiment, the NK participants did not know that the SK speaker was acting as a confederate. The confederate himself was instructed to behave in a way that did not indicate that he knew the answers (cf. [15]).

After a short break, NK participants completed a questionnaire which included key personal information, factors affecting language use (length of residence in North and South Korea and the UK; use

of Korean; attitudes towards NK and SK). This formed the basis of a short, informal interview. Interviews were not recorded due to the sensitive nature of the community, but any important information was noted.

All testing was carried out by the second author.

Analysis. For the wordlist, a total of 360 tokens were analyzed. Tokens were manually segmented in Praat [16]. F1 and F2 values were measured at the vowel midpoint for each target word. The read passage was segmented using a Korean forced aligner [17] and F1 and F2 values for /ʌ/ were extracted (N=180). For Diapix, recordings were transcribed using the transcription function in Office 365 and then hand-checked for errors. F1 and F2 values for /ʌ/ that were not in a nasal or liquid context were manually extracted (N=601). For all data, formant measures that were over 2 standard deviations outside the F1 or F2 mean per vowel for either NK or SK speakers were checked and hand-corrected if necessary. Data were normalized using Nearey2 [18]: vowels from the passage and Diapix were normalized using the log-means calculated from the wordlist.

Statistical analysis was carried out in R [19] using the *lmerTest* package [20] using Satterthwaite's approximation to obtain *p* values. Any post-hoc comparisons were conducted using the *emmeans* package [21]. *ggplot2* was used for data visualization [22].

3. RESULTS

3.1 Wordlist & Read passage

As displayed in Fig. 1, the majority of vowels in the wordlist were produced similarly by all NK and SK speakers. The only exceptions were for /ʌ/ and /e/-/ɛ/. As expected, /e/-/ɛ/ were fully merged in SK but not NK speakers. LNK speakers had a clear split between /e/-/ɛ/, whilst ENK speakers produced /e/ with a lower vowel, intermediate between the NK and SK vowel. SK speakers produced /ʌ/ as a back, mid-central vowel but LNK speakers produced this as a high back vowel, whilst ENK speakers produced it with a quality intermediate between LNK and SK speakers.

To investigate potential differences in production of /ʌ/, a linear mixed effects model was built for F1 and F2 separately, with group (SK, ENK, LNK) and phonetic context as fixed factors, and participant as a random factor. For F2, there were no significant effects of group, context or their interaction. Normalized F1 produced by LNK speakers was significantly lower than that produced by ENK speakers ($p = .0190$), which was in turn significantly lower than F1 for SK speakers ($p < .0001$). There were also significant interactions between group and phonetic context. Post-hoc Tukey-adjusted comparisons showed that F1 did not differ by context among ENK or LNK speakers, but was higher in /hVdV(C)/ than in /kVbV(C)/ ($p = .0010$) and /sVgV(C)/ ($p = .0068$) for SK speakers.

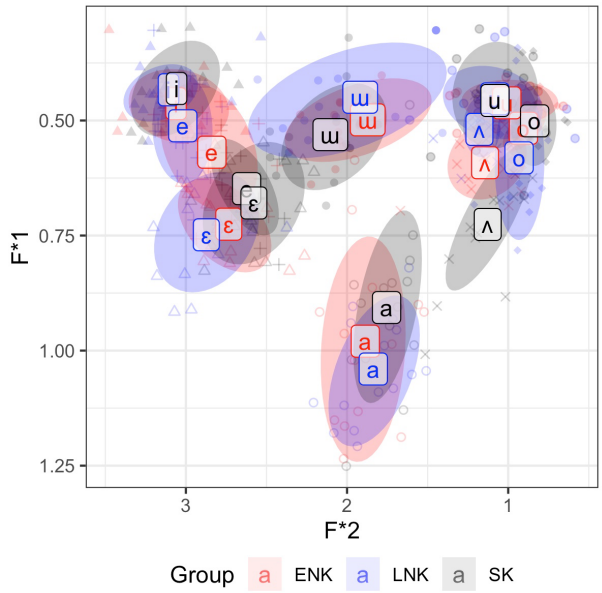


Figure 1: Scatterplot to show vowel production in the wordlist, averaged across 3 vowel contexts and split by accent background: ENK (Early North Korean), LNK (Late North Korean), SK (South Korean).

Fig. 2 shows production of /ʌ/ for all speakers in the wordlist and read passage, and for Diapix (NK speakers only). As expected, there is a clear centring effect from the wordlist to the read passage; in the read passage, all speakers show vowel reduction with /ʌ/ becoming more centralized. However, in the read passage itself, there are few differences between speaker groups. This was confirmed in linear mixed effects models for F1 and F2 separately with group

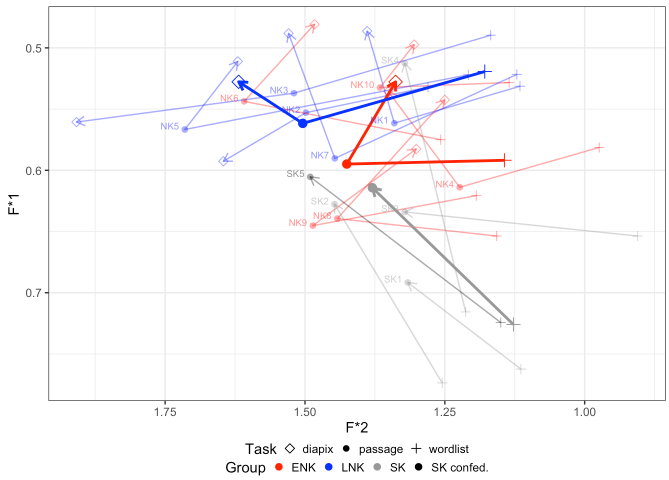


Figure 2: Scatterplot to show average production of /ʌ/ in the wordlist, passage and Diapix by SK and NK speakers, split by accent background: ENK (red), LNK (blue) and SK (grey) speakers, split by group (darker shade) and by individual (lighter shades). The male confederate is SK5.

included as a fixed effect. There were no significant effects for F1 or F2 ($p > .05$).

3.2 Investigating Accommodation: comparison of read speech & Diapix (NK speakers only)

To assess potential phonetic accommodation (i.e., convergence or divergence), we compared production of /ʌ/ in the read passage and Diapix tasks. In this set of models, task, group (ENK, LNK; sum-coded) and their interaction were included as fixed factors and participant as a random factor. For F2, there were no main effects of task or group but there was a significant interaction of group and task ($p = .0007$). As displayed in Fig. 2, F2 moves in different directions for the two groups: it decreases (i.e., backs) for ENK and increases for LNK (i.e., fronts), with the overall fronting effect for LNK driven by NK2 and NK3. For F1, there was a significant effect of task ($p < .0001$): F1 in Diapix was lower than that in the passage for all participants. There was no significant main effect of group or interaction between group and task ($p > .05$).

4. GENERAL DISCUSSION

The current study investigated phonetic accommodation in the New Malden Korean community where North and South Koreans live and work alongside each other. This is very different from the situation in South Korea, where North Koreans struggle to integrate into society and report actively trying to hide their NK identity (e.g., [5, 23]). Of interest was whether, given this very different social environment, North Koreans would continue to use SK-like vowels, particularly when interacting with an SK speaker.

Although not native-like, our results show that some NK speakers do acquire SK-like vowels. In the wordlist, ENKs who had more experience with SK produced /ʌ/ with a higher F1 and had a smaller /e/-/ɛ/ split such that /e/ was intermediate between that of LNK and SK speakers (cf. [24]). LNK speakers, who had less experience with SK, used their native NK vowels (cf. [14, 25]): they produced /ʌ/ with a lower F1 and maintained an /e/-/ɛ/ split typical of NK varieties.

However, regardless of experience, all of our NK speakers diverged from their SK interlocutor in production of /ʌ/ in the Diapix task. This was somewhat surprising. Previous studies have emphasized North Koreans' desire to fit in with their new South Korean community (e.g., [5]) and so, based on this, we had expected that in this task in particular, all our speakers would produce /ʌ/ as a more central or lower vowel, to better match their SK interlocutor.

Why did these NK speakers diverge? One possibility is that although at least some of the speakers had SK-like variants available in their repertoire (i.e., the ENKs), using these later-learned variants requires a higher degree of cognitive control, and that when attention is diverted (e.g., by a complex task or emotive topic) first-learned variants emerge instead (cf. [26]). Although we cannot rule out a role for cognitive factors, Diapix is a straightforward task and was a format familiar to all participants. There is little time pressure and the pictures are not designed to elicit an emotional response.

Perhaps a more likely explanation is that our speakers diverged as a result of social factors. As previously mentioned, North Korean refugees living in South Korea experience stigmatization as a result of their background and some, like those in our study, emigrate for a better life [10]. New Malden is attractive because of the large Korean community, where North Koreans are an important part of the local workforce. They have created their own community, distinct from that of South Koreans, and characterized by close ties (e.g., [27, 28]) and active participation in the community (cf. [29]). The majority of our participants worked as waitresses and cleaners within South Korean businesses and had formed close relationships with others in the NK community, e.g., through involvement in the North Korean school and workplace. In the interviews, they reported that they no longer felt the need to hide their identity and saw themselves instead as members of a London-based Korean community (cf. [5]). We argue then that these speakers, settled in a new community outside Korea, were able to reclaim and index their NK identity through use of their NK /ʌ/ vowel. It is possible that this effect was magnified because the confederate was known to them. Further, different variants, e.g., the /e/-/ɛ/ split, may behave differently: NK speakers may develop a hybrid accent in order to signal their new identity as North Koreans in a community where North and South Korea can coexist (cf. [24]). These questions require further analysis and experimentation (i.e., completing Diapix with an unfamiliar interlocutor), both of which we plan to do in the future.

In conclusion, NK speakers who had left North Korea before age 16yrs (i.e., ENKs) had acquired SK-like vowel targets and used these in more formal speech tasks. However, identity rather than experience determined the outcome of phonetic accommodation: in New Malden, where NK speakers are settled, active members of the community, they used their native NK /ʌ/ vowel to reclaim and index their NK identity.

5. REFERENCES

[1] Chambers, J.K. 2002. Dynamics of Dialect Convergence. *J. Sociolinguistics* 6(1), 117-130.

[2] Trudgill, P. 1986. *Dialects in Contact.* Blackwell: Oxford.

[3] Foulkes, P. & Docherty, G. 2006. The social life of phonetics and phonology. *J of Phonetics* 34: 409-438.

[4] Evans, B.G. & Iverson, P. 2007. Plasticity in vowel perception and production: A study of accent change in young adults. *J. Acoust. Soc. Am*. 121(6), 3814–3826

[5] Park, M-Y. 2021 'I don't want to be distinguished by others': language ideologies and identity construction among North Korean refugees in South Korea. *Lang. Awareness* 31(3), 271-287

[6] Eychenne, J. & Jang, T-Y. 2005. On the Merger of Korean Mid Front Vowels: Phonetic and Phonological Evidence. *J Korean Soc of Sp Sci,* 7(2), 119-129.

[7] Lee, K.-M. & Ramsey, R.S. 2011. *A history of the Korean language*. Cambridge: Cambridge University Press.

[8] Kwak, C.-G. 2003. The vowel system of contemporary Korean and direction of change. *J. of Korean Linguistics* 41, 59–91.

[9] So, S. A. 2021. Hamnam pangŏnŭi moŭm ch'egyewa kŭ pyŏnhwa - hamhŭng chiyŏgŏrŭl chungshimŭro [The Vowel System of the South Hamgyeong dialect and its Change -Focusing on Hamheung subdialect]. Kugŏhak 97, 177-214.

[10] Lee W. J and Baek. N. S. 2021. Puk'anit'alchumin t'allam shilt'aebunsŏng min taeŭngbangan yŏn'gu [Analysis and countermeasures for North Korean defectors who settled in Korea to leave Korea]. Han'gukkyŏngch'arhak'oebo, 23(3), 49-76.

[11] Fisher, P. 2015. The Korean Republic of New Malden: How Surrey became home to the 70 year-old conflict. *The Independent.* https://www.independent.co.uk/news/uk/home-news/the-korean-republic-of-new-malden-how-surrey-became-home-to-the-70-yearold-conflict-10063055.html

[12] Shin, H. 2018. The territoriality of ethnic enclaves: dynamics of transnational practices and geopolitical relations within and beyond a Korean transnational enclave in New Malden, London. *Annals of the American Association of Geographers* 108(3), 756-772.

[13] Baker, R. & Hazan, V. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Beh. Res. Methods* 43, 761–770.

[14] Chambers, J.K. 1992. Dialect Acquisition. *Language* 68(4), 673-705.

[15] Stamp. R., Schembri, A., Evans, B.G. & Cormier, K. 2016. Regional Sign Language Varieties in Contact: Investigating Patterns of Accommodation. *J. of Deaf Studies and Education,* 70-82.

[16] Boersma, P. & Weenink, D. 2021. Praat: Doing phonetics by computer (Version 6.2.14).

[17] Yoon. T. J & Kang. Y. (2013). The Korean Phonetic Aligner Program Suite. https://korean.utsc.utoronto.ca/kpa/

[18] Nearey, T. 1989. Static, dynamic and relational properties in vowel perception. *J. Acoust. Soc. Am*. 85(5): 2088-2113.

[19] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[20] Kuznetsova A., Brockhoff, P.B., & Christensen R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J. of Statistical Software* 82(13), 1–26.

[21] Lenth. R., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H. & Singmann, H. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://github.com/rvlenth/emmeans

[22] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[23] Kim, H., & Hocking, C. 2018. The stories of North Korean refugees settling in South Korea: Implications for South Korean educators. *Multicultural Education Review*, 10(3), 203–223.

[24] Tomé Lourido, G. & Evans, B.G. 2019. The effects of language dominance switch in bilingual: Galician new speakers' speech production and perception. *Bilingualism: Language and Cognition* 22(3), 637-654.

[25] Tagliamonte, S. & Molfenter, S. 2006. How'd you get that accent? Acquiring a second dialect of the same language. *Language in Society* 36, 649–675.

[26] Sharma, D. 2018. Style dominance: Attention, audience, and the 'real me'. *Lang. in Soc.* 47, 1–31.

[27] Milroy. L. 1982. Language and group identity. *Journal of Multilingual and Multicultural Development* 3(3), 207-216,

[28] Milroy, L., & Milroy, J. 1992. Social Network and Social Class: Toward an Integrated Sociolinguistic Model. *Language in Society* 21(1), 1–26.

[29] O'Rourke, B., & Ramallo, F. (2015). Neofalantes as an active minority: Understanding language practices and motivations for change amongst new speakers of Galician. *International Journal of the Sociology of Language* 231, 147–165.

WHAT YOU SEE IS (NOT) WHAT YOU HEAR: THE ROLE OF SOCIAL EXPECTATIONS ON SPEECH INTELLIGIBILITY IN A BRITISH CONTEXT

Ruohan Guo, Fengjie Cheng, Bronwen G. Evans

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
ucjurgu@ucl.ac.uk, fengjie.cheng.20@ucl.ac.uk, bronwen.evans@ucl.ac.uk

ABSTRACT

Several studies have found that giving listeners information about the talker affects speech comprehension. This study investigated whether exposing listeners to regionally-meaningful cues has similar effects on the intelligibility of native accents in a British context. Two groups of participants, who differed in their familiarity with Glaswegian (GE) and Standard Southern British English (SSBE), transcribed GE-accented and SSBE-accented sentences at three SNRs (+3 dB, 0 dB, -3 dB) while they were exposed to visual cues indicating (1) a congruent accent region, (2) an incongruent accent region or (3) silhouetted cues representing no specific region. As expected, noise level and accent familiarity affected performance. However, contrary to previous findings, congruence between regional cues and spoken accent did not affect speech intelligibility in either listener group and did not interact with noise level or familiarity. This null result raises questions about the generalizability of the effects of social expectations on speech perception.

Keywords: sociophonetics, speech perception, exemplar-models, speech intelligibility in noise.

1. INTRODUCTION

A major area of interest in sociophonetics is how non-linguistic information is integrated into speech processing. There is evidence that perceived talker characteristics can influence listeners behaviour in perceptual tasks. For example, Niedzielski [1] found that listeners from Detroit categorised /aʊ/ differently depending on the description they were given of the speaker's regional origin. Niedzielski suggested that listeners' phoneme categorization decisions reflected awareness of phonetic features that they stereotypically attributed to certain regional accents. Subsequent studies have also found that listeners use social information to guide speech processing and that even implicit cues such as stuffed toys [2] or newspapers [3] can affect perceptual performance.

The observed effects have led to an interest in usage-based and exemplar approaches to speech perception [4, 5]. These argue that, based on observations or stereotypes, listeners form and store in memory associations between linguistic instances and socioindexical information (about speakers or contexts) and exploit these when processing newly-encountered speech.

In recent years, a growing body of research has found that exposure to socially-meaningful cues can also affect the processing of utterances beyond the phoneme level. In a transcription task, Babel and Russell [6] reported that listeners who were shown a photo of a Chinese speaker transcribed L1-English sentences less accurately than listeners who were shown a photo of a white Caucasian. McGowan [7] complemented this finding, demonstrating that Chinese-accented sentences in noise were more intelligible to American listeners when paired with an Asian face than with a Caucasian face. In contrast, misleading social expectations (i.e., incongruent pairing of face and voice) led to activation of inappropriate associations and hindered word recognition [6]. These findings are consistent with an exemplar-based account in which congruent speaker information activates appropriate associations in listeners' memory and facilitates recognition of utterances (see also [8]).

The reported effects of the role of social expectations in speech intelligibility have not only contributed to theories of speech perception but have also helped us better understand how stereotypes can disadvantage a speaker in communication (e.g., [9, 10]). However, most existing studies have been conducted in either New Zealand (e.g., [2]) or North America, with those in North America focussing primarily on the effect of perceived talker ethnicity on the intelligibility of standard native and/or foreign accents (e.g., [6, 7, 8]). It is unclear if this effect generalizes to different geographical contexts and to the perception of regional identity.

This study contributes to existing knowledge by testing how exposure to regionally-meaningful cues affects speech intelligibility in a British context. In two online experiments, we measured listeners' sentence transcription accuracy in two British English accents, Glaswegian English (GE) and Standard Southern British English (SSBE), whilst they were exposed to visual cues indicating; (1) a congruent accent region, (2) an incongruent accent region or (3) silhouetted cues representing no specific region (control condition). The first experiment involved

listeners from Glasgow who were familiar with both accents, while the second recruited listeners from the south of England, highly familiar with SSBE but not GE. Based on previous findings [7], we hypothesised that listeners would transcribe sentences in noise more accurately in the congruent condition than in the control condition and least accurately in the incongruent condition, but that this would be modulated by familiarity such that greater familiarity would lead to a larger congruence effect. In addition, stimuli were presented at three signal-to-noise ratios (SNR) (+3 dB, 0 dB, -3 dB). We expected that transcription accuracy would decrease with noise [11, 12] but of interest here, was whether effects of regional cues would be modulated by listening conditions.

2. EXPERIMENT ONE

2.1. Methods

2.1.1 Participants

Twenty-nine monolingual native English speakers aged 18-50 yrs were recruited via Prolific (www.prolific.co) [13]. All participants had lived in Glasgow for at least 3 months and reported being very familiar with both GE and SSBE. Twenty-six self-identified as having a GE accent. The participants reported no history of any speech, hearing, or language disorders.

2.1.2 Auditory stimuli

The stimuli were 108 sentences from the Institute of Electrical and Electronics Engineers (IEEE) sentence list [14], taken from a previous study [15]. Sentences were recorded at 44.1 kHz, 16-bit resolution, then segmented in Praat [16] and downsampled to 22,050 kHz. Amplitude was normalized to 70 dB. The sentences were produced by two GE and two SSBE speakers, all male, monolingual English speakers, aged 26-46 yrs, who identified as middle class.

Half of the sentences (N=54) were randomly assigned to the GE accent condition, and the other half (N=54) to the SSBE accent condition. For each accent, a third of the sentences (N=18) were paired with images representing a congruent accent region; the second set (N=18) was paired with images representing an accent region incongruent with the sentence accent; the final set (N=18) was paired with neutral images representing no particular place.

2.1.3 Visual stimuli

Visual stimuli were chosen to represent Glasgow (GE) and London (SSBE). The images included iconic cartoon-style images of the city skylines, football club mascots and a culturally salient figure representing each city (London – Beefeater; Glasgow – piper in traditional Scottish dress). For the neutral background, a silhouetted cityscape and two stick figures were selected. The icons were arranged in the centre of the screen and around the transcription input box to ensure that participants would notice them.

A naïve assessor confirmed that the images represented each city or, for the neutral background, no specific place. Post-experiment questionnaires showed that participants also linked the visual cues with each city or a neutral place.

2.1.4 Procedure

The experiment was designed and conducted online in Gorilla Experiment Builder (www.gorilla.sc) [17].

Before beginning the experiment, participants were told that their performance would be used to assess whether the experimental setup was suitable for children and that, as such, they would see images as well as hear a sentence. This was to encourage them to pay attention to the images. However, consistent with some previous studies [2], these regional cues were not attributed to the speaker.

For the transcription task, participants were instructed that they would hear each sentence only once and were asked to type in anything they heard. Participants completed 108 trials in a randomised order. Sentences were played at a self-adjusted comfortable listening level. The experiment was self-paced. Participants clicked a button to start each trial: the screen then simultaneously displayed the images and a play button which participants clicked to hear the sentence.

Finally, participants completed a post-experiment questionnaire. They rated the images and gave basic information about their language background and experience with GE and SSBE.

2.1.5 Data analysis

Data was pre-processed to remove punctuation. A custom-built Excel script was used to automatically calculate the proportion of words correctly transcribed in each sentence. Manual spot checks were completed afterwards, e.g., to correct typos ("squireel" for "squirrel").

Data was then analysed using R [18]. A binomial generalized linear mixed model was fitted. The proportion of correctly transcribed words (bounded between 0 and 1) was used as the dependent variable. Based on the predictions, the model included *Congruence* (Congruent, Control, Incongruent), *Noise Level* (-3 dB, 0 dB, 3 dB) and *Sentence Accent*

(GE, SSBE) as fixed effects, as well as the interactions between them. Random intercepts for participant and sentence were also included. The *mixed*() function in the *afex* package was used to compute the p-values of all fixed effects using likelihood ratio tests [19]. Follow-up tests were conducted with the *emmeans* package [20] and corrected pairwise comparisons with Holm correction.

2.2 Results

As expected, the likelihood ratio tests for the full model demonstrated a significant main effect of *Noise Level*, $\chi^2(2) = 50.70$, $p<0.001$, but a non-significant effect of *Sentence Accent*, χ^2 (1) = 1.13, p =0.28. Overall, GE listeners performed more poorly at higher noise levels but performed equally well with GE and SSBE (see Fig. 1). Follow-up tests showed that, on average, listeners transcribed sentences more accurately at +3 dB than at 0 dB (z=2.16, p =0.03) and -3 dB (z=7.757, p <0.001). Additionally, transcription accuracy at 0 dB was significantly higher than at -3 dB (z=5.624, p <0.001).

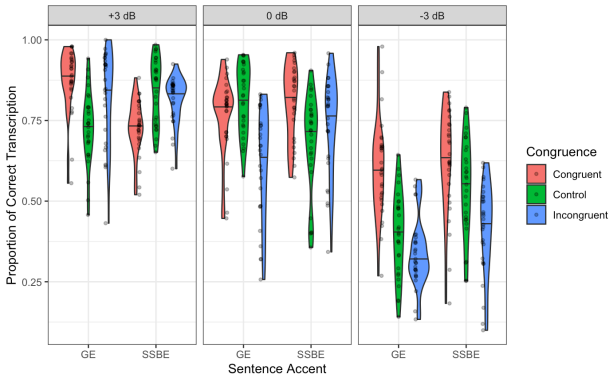


Figure 1: Proportion of correctly transcribed words for GE listeners in each accent (GE, SSBE) split by noise level and congruence condition.

However, there was a non-significant effect of *Congruence*, χ^2 (2) = 5.96, p = 0.051, and no significant interactions. Although the data, as displayed in Fig. 1, showed a pattern matching our prediction (i.e., *Congruent* > *Control* > *Incongruent*) at the highest SNR, -3 dB, this was not statistically significant.

2.3 Summary

Contrary to our expectations, we failed to find clear evidence for the effect of regional cues on speech intelligibility under adverse listening conditions.

Consistent with previous findings, GE listeners transcribed sentences less accurately at higher noise levels [11] and performed equally well with both GE and SSBE [12]. This likely reflects GE listeners' high level of familiarity with both these accents. However, the fact that GE listeners had high levels of experience with and, therefore, presumably detailed phonetic representations of both accents did not lead to any reliable congruence effects. Although there was some indication that sentences were identified more accurately when presented with congruent visual cues, particularly at higher noise levels, this was not reliable. This contrasts with previous studies [7, 8] which have found that listeners are able to benefit from matching social information.

One possibility is that these GE listeners, highly familiar with both GE and SSBE, have detailed enough phonetic representations and so either do not recruit non-linguistic information to support word recognition or are better able to ignore incongruent visual cues. In Experiment 2, we investigate whether listeners from the South of England who are highly familiar with SSBE but not GE perform similarly.

3. EXPERIMENT TWO

3.1 Methods

3.1.1 Participants

Twenty-four monolingual native English speakers aged 18-50 yrs were recruited via Prolific [13]. They had been born and raised in London and South East England and described their accent as SSBE. In the post-experiment questionnaire, all reported that they were highly familiar with SSBE but not GE, that they had not spent any significant amount of time in Scotland and did not have regular contact with GE speakers (e.g., family members, partner). Participants reported no history of any speech, hearing, or language disorders.

3.1.2 Stimuli and procedure

The design and procedure were the same as in Expt 1.

3.2 Results

We performed a binomial generalized linear mixed model on the SSBE group data, with *Congruence*, *Noise Level* and *Sentence Accent* as fixed effects and interaction terms between them. The model included random intercepts for both participant and sentence.

The results of likelihood ratio tests for the mixed effects model showed that, as predicted, there was a significant main effect of *Noise Level*, χ^2 (2) = 52.56, p <0.001. As in Expt 1, intelligibility decreased as the noise level increased (see Fig 2). Follow-up tests showed that listeners recognized utterances more accurately at +3 dB than at 0 dB (z=2.168, p =0.03) and -3 dB (z=7.93, p <0.001). Transcription accuracy

was also significantly higher at 0 dB than at -3 dB (z=5.78, p <0.001) where participants performed particularly poorly for GE in all conditions. There was also a significant main effect of *Sentence Accent*, χ^2 (1) = 28.66, p <0.001: SSBE listeners performed better with SSBE than GE (Fig 2). Follow-up tests showed that GE-accent sentences were less accurately transcribed than SSBE-accent sentences in noise (z=-5.72, p <0.001).

However, there was no significant main effect of *Congruence*, χ^2 (2) = 4.27, p =0.118, and no interaction terms were statistically significant. Although there was a trend for listeners to perform better with GE sentences in the congruent than incongruent condition, particularly at higher noise levels (Fig 2), SSBE listeners' transcription accuracy in noise was not reliably affected by the match or mismatch between the visual regional cues and spoken accent.

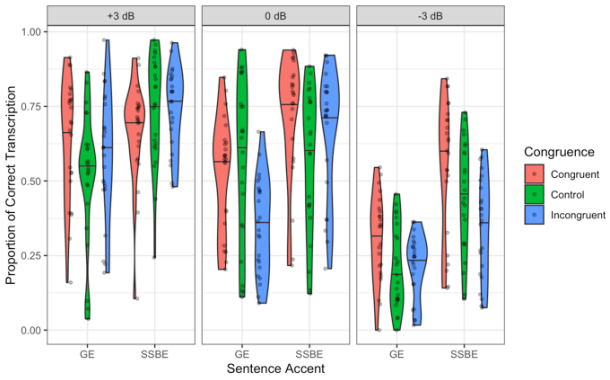


Figure 2: Proportion of correctly transcribed words for SSBE listeners in each accent (GE, SSBE) split by noise level and congruence condition.

3.3 Summary

The results of this experiment are similar to those reported in Expt 1. Listeners performed more poorly as the noise level increased. As predicted, SSBE listeners, highly familiar with SSBE but not GE, performed more poorly with GE than SSBE (cf. [12]), but there was no evidence for a congruence effect: there was no reliable difference in performance in the congruent, control or incongruent conditions for either accent at any noise level. This indicates that the effect of regional cues on sentence transcription was not modulated by familiarity with a given accent.

4. GENERAL DISCUSSION

This study explored how exposure to regionally-meaningful visual cues affects speech intelligibility in a British context. Although both experiments showed expected effects of noise level and familiarity, there was no evidence that transcription accuracy was affected by visual cues to regional accents. This null result is inconsistent with earlier studies showing that alignment between social cues and spoken accents increases sentence intelligibility in noise (e.g., [7]).

The null effect may be attributed to the subtle visual prompts used in this study. Earlier studies have primarily examined how perceived ethnicity affects speech intelligibility of standard L1- or foreign (L2-) accents in North America (e.g., [6, 7, 8]). In these experiments, listeners are typically led to believe that the speech they hear comes from the talker, depicted in a photo [6, 7] or explicitly described [8]. This may lead to a closer association of social cues and auditory information than is possible in an experiment like the present one, where visual stimuli were cartoon-like icons, and may in turn lead to a congruence effect.

Another explanation is that although regional cues may have guided social expectations, at least some of the listeners did not have sufficient experience with the accents and instead used stereotypical or simplified structures that did not match the speech they actually heard (cf. [7]). This could explain why SSBE listeners showed no effect of congruence for GE, with which they were unfamiliar. Nevertheless, this cannot explain why listeners did not show the expected priming effect for accents with which they were highly familiar. Although it could be the case that listeners discard mismatching visual information when they can reliably access the auditory signal (i.e., at easier SNRs), it is still unclear why visual cues appeared to have no effect on speech intelligibility at the higher noise levels (0dB, -3dB), similar to those used in previous studies (e.g., [7]).

Some studies have found a significant effect of regional cues on vowel categorization, mainly in New Zealand (e.g., [2] and more recently, [21]). Researchers in Britain have attempted to replicate these effects but have failed to find convincing evidence that exposure to regional cues affects phoneme matching [22, 23]. The current study is not directly comparable to these previous experiments due to methodological differences (sentence transcription *vs* phoneme matching). However, the lack of a significant effect in these different studies may indicate either that the effect of social cues on speech processing is more limited than expected or that listeners do use this information but these paradigms are not always able to detect this [cf. 22, 23]. Further studies using a variety of approaches are needed to investigate the role of these and other factors, in particular sociocultural context, on the role of social expectations in speech processing.

5. REFERENCES

[1] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1), 62–85.

[2] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48(4), 865–892.

[3] Portes, C., German, J. S. 2019. Implicit effects of regional cues on the interpretation of intonation by Corsican French listeners. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10, 1-26.

[4] Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34, 485–499.

[5] Docherty, G. J., Foulkes, P. 2014. An evaluation of usage-based approaches to the modelling of sociophonetic variability. *Lingua* 142, 42-56.

[6] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *The Journal of the Acoustical Society of America 137*(5), 2823-2833.

[7] McGowan, K. B. 2015. Social expectation improves speech perception in noise. *Language and Speech 58*(4), 502-521.

[8] Vaughn, C. R. 2019. Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America* 145(5), 3218-3232.

[9] Kang, O., Rubin, D. L. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28, 441–456.

[10] Kutlu, E., Tiv, M., Wulff, S., Titone, D. 2022. Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications* 7(1), 1-16.

[11] Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., Abrams, H. B. 2006. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics* 27(3), 465-485.

[12] Adank, P., Evans, B.G., Stuart-Smith, J., Scott, S.K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance* 35 (2), 520-529.

[13] Palan, S., Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17, 22-27.

[14] Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust* 17, 227–246.

[15] Evans, B.G., Adank, P. 2008. Differences in the time-course of accent adaptation: a comparison of adaptation to foreign-accented and unfamiliar regionally-accented speech. *Journal of the Acoustical Society of America* 123(5), 3703-3703.

[16] Boersma, P. & Weenink, D. 2021. Praat: Doing phonetics by computer (Version 6.2.14)

[17] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1), 388-407.

[18] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[19] Singmann, H., Bolker, B., Westfall, J. 2015. afex: Analysis of factorial experiments. http://CRAN.R-project.org/package=afex.

[20] Lenth. R., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H. & Singmann, H. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* https://github.com/rvlenth/emmeans.

[21] Hurring, G., Hay, J., Drager, K., Podlubny, R., Manhire, L., Ellis, A. 2022. Social priming in speech perception: revisiting Kangaroo/Kiwi priming in New Zealand English. *Brain Sciences* 12(6), 684.

[22] Lawrence, D. 2015. Limited evidence for social priming in the perception of the BATH and STRUT vowels. *Proc. 18th ICPhS* Glasgow.

[23] Juskan, M. 2018. *Sound Change, Priming, Salience: Producing and Perceiving Variation in Liverpool English*. Language Science Press.

[24] Hanulíková, A. 2021. Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PLoS ONE* 16(10), e0259230.

CHANGES IN FUNDAMENTAL FREQUENCY AND F1 ACROSS CHRONOLOGICAL AGE

Seren Parkman

Lancaster University
s.parkman@lancaster.ac.uk

ABSTRACT

Chronological age is used widely in academic research and yet it is poorly understood how accurately it measures ageing across the lifespan. The current study considers 17 speakers of English between 22-79 years old (M=51.65, SD=18.09) focusing on whether their f_o and F_1 correlate with chronological age in three different vowels (/ə, iː, ɛ/). I hope to determine whether there is a linear correlation between f_o and chronological age, and F_1 and chronological age, or perhaps the changes we see in speech production are as a consequence of other factors. Results indicate that there is a very weak effect of chronological age on f_o and F_1. Therefore, the evidence suggests more research into other age-grading measures, such as the social and biological should be considered in the future [1, 2], though I argue that chronological age should not be dismissed entirely from research as proven in previous studies [3, 4].

Keywords: chronological age, f_o, speech production, formants, lifespan

1. INTRODUCTION

The present study aims to determine whether f_o and F_1 correlate linearly with chronological age (CA), in hopes of continuing current ageing research in linguistics [3, 4], and consider how we should be evaluating age across the lifespan in future research [2]. When we speak of CA we speak of the numerical value assigned at birth, which is used widely across research areas including linguistics, and yet we know little about how well this measure fully encapsulates all aspects of ageing [5]. Previous research determines that the voice can be impacted by lots of different factors as we age. Social factors known to influence speech production include: class, race, gender, language attitudes, politics and economic experience [6]. In addition, previous research notes how speech production can change depending on our relationship to the individual we are talking to, determining whether we accommodate or not [7] to them, resulting in us subconsciously moving our phonetic targets to align with this other individual [8]. On a more biological level, puberty can cause male speakers to experience vocal fold mass increases, a descending larynx from an increased neck length and width, and a larger nasal cavity [2]. Likewise, hormonal changes such as the menopause have been evidenced to influence voice quality in female speakers, generally causing a 'decrease in lung power, atrophy of laryngeal muscles, stiffening of laryngeal cartilages, vocal fold thickening, and a loss of elastic and collagenous fibres' [9], resulting in a lowering of f_o and F_1 over the lifespan. This lowering of f_o and F_1 is evidenced in previous studies [3, 4] on CA across the lifespan. Reubold et al [4] speculate the reasons for this could be due to: vowel harmonics, with f_o tracking the formants across the lifespan, due to the intrinsic relationship between them [10, 11], vocal tract lengthening over time, or a more perceptual reason to maintain perceived phonetic height over time [3]. Harrington et al [3] conclude that whilst f_o and F_1 both appear to correlate with CA over time, there are still unanswered questions as to why this occurs. As a result, my own study aims to be a precursor to future studies analysing how language may change as a consequence of age by tracking f_o and F_1 across CA. This should help us to understand how speech production may be influenced by increasing age, and how perhaps in future we can incorporate other ageing models into linguistic research. For example, in more recent years, Hejná and Jespersen [2] suggest that perhaps biological and social age are better measures for capturing and understanding linguistic behaviour over CA. As a result, by understanding how much CA can explain variance in f_o and F_1, we can begin to understand why other models of age may be useful to consider alongside CA in future research.

1.1. Summary and predictions

In this study, I compare f_o and F_1 across chronological age (CA), in order to understand the relationship between f_o and CA, and F_1 and CA in

speech production. In light of the existing research I suggest the following predictions:

Any correlations seen between CA and f_o could result from hormonal changes such as lower testosterone levels in males and lower oestrogen and progesterone after menopause in females [12]. Additionally, correlations seen between F_1 and CA suggests possible evidence for vowel centralisation as older speakers' vocal tracts centralise with age [12]. If we see correlations between both f_o and CA, and F_1 and CA we would assume that perhaps CA does have an influence on our speech production as we age. Any differences we see between f_o and F_1 may result from differences in voice quality, [13] and the size and strength of the vocal folds causing changes in f_o [9] or changes to the size and shape of the vocal tract causing changes in F_1 [14, 15]. Nonetheless, as mentioned f_o and F_1 are intrinsically linked, so we would expect them to follow a similar trajectory across CA [10, 11]. If correlations between CA and f_o, and CA and F_1 cannot fully be explained or do not occur then future research should consider other biological and social factors in order to observe age-related changes in speech production.

2. METHODS

2.1. Participants

Data from 17 speakers (9 female and 8 male) were collected for the purposes of this study ranging between 22-79 years old (M=51.65, SD=18.09). Male ages ranged from 22-79 years old (M=57, SD=18.21) and females from 22-67 years old (M=46.89, SD=17.61). All speakers within the study were born or raised in Kendal, a small town in the county of Cumbria, UK, bordering Lancashire to the South, Scotland to the North and Yorkshire to the East. Social information revealed that the participants all fell into a lower to upper middle-class socioeconomic status.

2.2. Data collection

Speech data was collected using a Zoom H1 recording device which produced .wav files at a sample rate of 44,100 Hz. Recordings took no more than 10 minutes. Participants also completed a questionnaire to elicit some social information and their own personal thoughts about ageing.

2.3. Materials

The word list was made up of 40 words. Within this, there were 10 words aiming to elicit a /ə/ vowel, 10 for the /iː/ vowel, and 10 for the /ɛ/ vowel. The vowels were chosen due to their relatively stable nature across the lifespan in the North West of England, being less affected by social change. The /ə/ vowel was also chosen to help replicate findings in previous studies [3, 4], where both f_o and F_1 declined with CA. In addition, there were an extra 10 distractor words. In total this produced 680 tokens, when the distractor words were removed, and the data was analysed and filtered, 469 tokens were eventually used in the final data.

An additional questionnaire was presented to participants, based on existing metrics presented in previous studies [16, 2] which aim to find out how the voice can be affected by speech and how we might measure age, including occupation and social background, how an individual feels about age, and hormonal and biological effects which may impact on an individual's age. In future studies it would be useful to use this data to understand how social factors may be influencing on an individual's age.

/ə/ Words	/ɛ/ Words	/iː/ Words
liar	dress	fleece
mother	protest	bleed
question	effort	sheep
vessel	confess	grief
again	attest	leaf
caution	net	evil
denim	head	believe
melon	regrets	beach
across	press	breeze
extra	excess	freed

Table 1: Word List used to extract /ə/, /ɛ/ and /iː/ vowels from participants.

2.4. Data analysis

Speech data was collated and analysed using Praat [17]. Before analysis, recordings were downsampled to 22,050 Hz, low-pass filtered to 11,025 Hz and high-pass filtered to 50 Hz. From here individual vowels were annotated by word, vowel and speaker in order to extract the formant values. I took the mean of f_o and F_1 across the vowel duration. This was in order to provide a more robust estimate across the vowel, rather than relying on a single timepoint. I present inferential statistics of the results extracting Pearson's r and p values for

this paper. This was in order to understand whether or not we could see correlations between CA and f_o, and CA and F_1, and to understand whether there is a relationship between them [18].

3. RESULTS

Figures 1, 2 and 3 present the results of this study showing f_o and F_1 against CA for male and female speakers in the three separate vowels (/ə, ɛ/, iː). Results indicate that f_o and F_1 are both not correlating strongly with CA for all three vowels. In the next few sections I look into each vowel in detail.

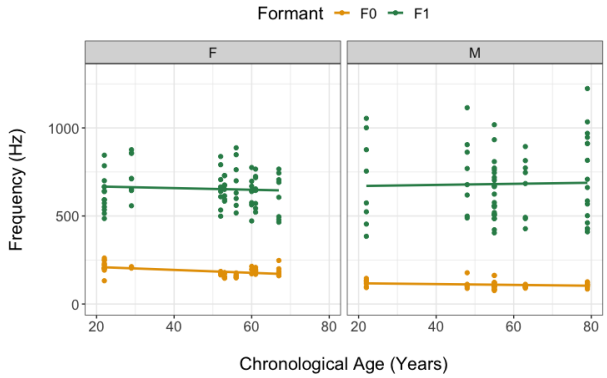


Figure 1: f_o and F_1 against chronological age results for the /ə/ vowel.

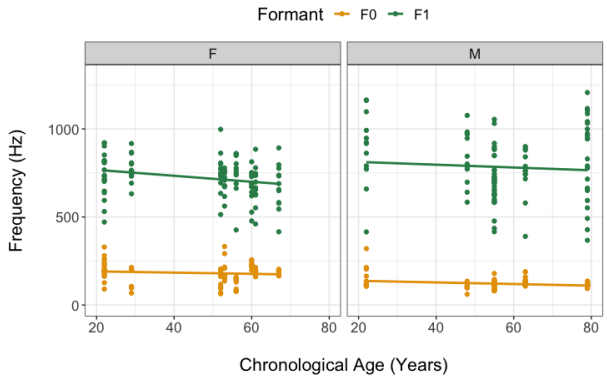


Figure 2: f_o and F_1 against chronological age results for the /ɛ/ vowel.

3.1. /ə/ vowel

Overall, there does not appear to be a strong correlation between CA and either f_o or F_1 for the /ə/ vowel in both male and female speakers. Using statistical testing, we notice that for male speakers the correlation is r=−0.2, p=.2 for f_o and r=0.03, p=.84 for F_1. This shows no significant correlation

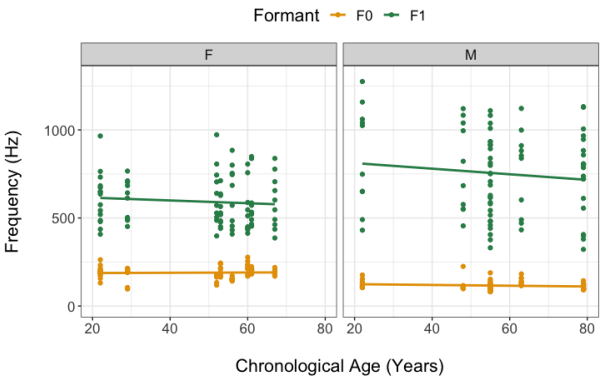


Figure 3: f_o and F_1 against chronological age results for the /iː/ vowel.

between CA and F_o, and CA and F_1, for male speakers. For female speakers the correlation is r=−0.48, p=<.001 for f_o and r=−0.08, p=.55 for F_1. This shows no significant correlation between CA and F_1 and a weak significant correlation between CA and f_o. Looking at the graphs, we can note that for female speakers there is a slight decline in f_o and F_1 overtime, but not to a great extent and with considerable participant variation across CA. Similar results can be observed for male speakers, however this time with a slight rise in f_o and decline in F_o, but with variation and 'troughs' forming at around 55 years old with a steep decline, and then rise again in F_1 across CA. Nonetheless, these correlations are non-significant, suggesting the /ə/ vowel does not appear to vary much across CA in terms of f_o and F_1.

3.2. /ɛ/ vowel

Both male and female speakers produce similar results in their f_o and F_1 for the /ɛ/ vowel across CA. Statistical measures indicate that for male speakers the correlation is r=−0.24, p=.02 for F_o, and r=−0.071, p=.49 for F_1. This shows no significant correlation between f_o and CA, and F_1 and CA. For female speakers the correlation is r=−0.11, p=.26 for f_o, and r=−0.24, p=.014 for F_1. Much like the men this shows no significant correlation between f_o and CA, and a weak-significant correlation between F_1 and CA. The graphs demonstrate that there is substantial variation within participants, especially for F_1, but also f_o in female speakers, but not all these patterns were significant. Results therefore indicate that there is not a decline over CA for either f_o or F_1, going against previous literature in this area [3, 4].

3.3. /iː/ vowel

There were non-significant negative correlations found for the majority of data for f_o and F_1 for the /iː/ vowel across CA. Male speaker correlation shows that r=−0.15, p=.2 for f_o and r=−0.11, p=.32 for F_1. This shows no significant correlation between f_o and CA, and F_1 and CA. Female results follow a similar pattern with the correlation as r=0.039, p=.72 for f_o and r=−0.089, p=.42 for F_1. This shows no significant correlation between f_o and CA, and F_1 and CA. As a consequence, the /iː/ vowel demonstrates little correlation between F_1 and CA, and f_o and CA. The results indicate that instead, f_o and F_1 remain relatively stable across CA, especially for f_o, suggesting that perhaps there is little change in f_o and F_1 across CA.

4. DISCUSSION

The results of this study indicate that there is very weak non-significant correlations between f_o and CA, and F_1 and CA. There appears to be little difference between the vowels presented in this study suggesting the weak relationship between CA and f_o, and CA and F_1 is consistent across the vowels. The slight linear decrease we do see could stem from changes in f_o and F_1 over CA due to the centralising of the vowel space due to ease of articulation as we get older. Though previous results on this are mixed and vary between the genders [19, 20, 21]. In addition, particularly for the female participants, slight declines after around 50 years old could stem from hormonal influences such as the menopause [22]. Nevertheless, changes in F_1 are not linear, instead we see some fluctuation. These troughs, most notably seen in male speakers, have been acknowledged in previous work [23, 24] and occur around the age of 55 years as reflected in the results of this study, though some studies argue for much later, around 80 years old [4]. The conclusion presented in the past has been that these troughs are speaker dependent [4]. Perhaps these individual speaker discrepancies are actually what is being observed in this study. Despite this, male speakers' F_1 is far more variable than female F_1, perhaps because of anatomical differences such as men having a longer pharynx [25] or perhaps a more social reasoning such as Smyth and Roger's [26] suggestion that women have more careful speech articulations than men. Nonetheless, the evident similarity in trajectory between f_o and F_1 across CA could stem from the intrinsic relationship between f_o and F_1 [10, 11], with F_1 simply following f_o across CA in order to maintain perceptual distance [4]. Overall, however there is little change across CA in f_o, and in F_1, suggesting that there is a weak relationship between f_o and CA, and F_1 and CA. These results could also have been influenced by low token numbers, using an apparent-time methodological approach and significant gaps in age groups such as between 30-40 years for both sexes. Future research should explore further the implications of social and biological ageing, and explore a wider range of ages in order to fully understand how speech production changes across the lifespan.

5. CONCLUSIONS

The aim of this initial study was to determine whether chronological age correlates with f_o and F_1, in order to understand whether there is a linear relationship between them, or whether we need to consider other forms of ageing, such as social and biological, in order to accurately measure ageing as a whole. In light of this research there is a very weak effect of chronological age on f_o and F_1 in the vowels /ə, iː, ɛ/, suggesting that other factors may better explain variation in the data, such as biological or social age. Future research should hope to determine how much of an effect these different forms of ageing are having and influencing each other across the lifespan.

6. REFERENCES

[1] P. Eckert, "Age as a sociolinguistic variable," in *The Handbook of Sociolinguistics*, F. Coulmas, Ed. Blackwell Publishing Ltd, 2017, pp. 151–167.

[2] M. Hejná and A. Jespersen, "Ageing well: Social but also biological reasons for age grading," vol. 16, no. 5, 2022.

[3] J. Harrington, S. Palethorpe, and C. I. Watson, "Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers," in *Interspeech*. ISCA, 2007, pp. 2753–2756.

[4] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers," vol. 52, no. 7, pp. 638–651, 2010.

[5] D. Kotter-Grühn, A. E. Kornadt, and Y. Stephan, "Looking beyond chronological age: Current knowledge and future directions in the study of subjective age," vol. 62, no. 1, pp. 86–93, 2016.

[6] A. C. Zentella, "Lexical leveling in four new york city spanish dialects: Linguistic and social factors," vol. 73, no. 4, p. 1094, 1990.

[7] H. Giles, "Communication accommodation theory: negotiating personal relationships and social identities across contexts," 2016.

[8] A. Schweitzer and N. Lewandowski, "Social

factors in convergence of f1 and f2 in spontaneous speech," 2014, publisher: Unpublished.
[9] R. T. Sataloff, D. Caputo Rosen, M. Hawkshaw, and J. R. Spiegel, "The aging adult voice," vol. 11, no. 2, pp. 156–160, 1997.
[10] H. Traunmüller, "Perceptual dimension of openness in vowels," vol. 69, no. 5, pp. 1465–1475, 1981.
[11] U. Reubold and J. Harrington, "Disassociating the effects of age from phonetic change: A longitudinal study of formant frequencies*," in *IMPACT: Studies in Language and Society*, A. Gerstenberg and A. Voeste, Eds. John Benjamins Publishing Company, 2015, vol. 37.
[12] S. Schötz, "Acoustic analysis of adult speaker age," *Speaker classification I: Fundamentals, features, and methods*, pp. 88–107, 2007.
[13] J. Kreiman, B. R. Gerratt, K. Precoda, and G. S. Berke, "Individual differences in voice quality perception," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 3, pp. 512–520, 1992.
[14] B. H. Story, H. K. Vorperian, K. Bunton, and R. B. Durtschi, "An age-dependent vocal tract model for males and females based on anatomic measurements," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 3079–3102, 2018.
[15] S. E. Linville and J. Rens, "Vocal tract resonance analysis of aging voice using long-term average spectra," *Journal of Voice*, vol. 15, no. 3, pp. 323–330, 2001.
[16] C. L. Rose, "The measurement of social age," vol. 3, no. 2, pp. 153–168, 1972.
[17] P. Boersma and D. Weenink, "Praat software," *Amsterdam: University of Amsterdam*, 2013.
[18] W. Lowie and B. Seton, *Essential statistics for applied linguistics*. Palgrave Macmillan, 2013.
[19] P. Torre III and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of communication disorders*, vol. 42, no. 5, pp. 324–333, 2009.
[20] M. P. Rastatter, R. A. McGuire, J. Kalinowski, and A. Stuart, "Formant frequency characteristics of elderly speakers in contextual speech," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 1, pp. 1–8, 1997.
[21] S. A. Xue and G. J. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production," 2003.
[22] G. Prelevic, "The effects of sex hormones on the female voice," in *Physical and Emotional Hazards of a Performing Career: A special issue of the journal Musical Performance*. Routledge, 2013, pp. 93–103.
[23] S. E. Linville, "The sound of senescence," vol. 10, no. 2, pp. 190–200, 1996.
[24] R. Baken, "The aged voice: A new hypothesis," vol. 19, no. 3, pp. 317–325, 2005.
[25] R. L. Diehl, B. Lindblom, K. A. Hoemeke, and R. P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," *Journal of phonetics*, vol. 24, no. 2, pp. 187–208, 1996.
[26] R. Smyth and H. Rogers, "Phonetics, gender, and sexual orientation," in *Proceedings of the annual meeting of the Canadian Linguistics Association*. lâUniversity du Quebec au Montreal Montreal, 2002, pp. 299–301.

Reconstructing the perception of gender identity, sexual orientation, and gender expression in American English

Benjamin Lang

University of California, San Diego
blang@ucsd.edu

ABSTRACT

Previous research has identified associations between acoustic cues, such as mean f0 and center of gravity (COG) of /s/, and a speaker's gender identity (GI), sexual orientation (SO), and gender expression (GE). This study investigates how acoustic features combine to influence the perception of GI, SO, and GE from speech. 197 listeners rated the speech of 66 speakers of American English along scales of GI, SO, and GE. Measurements of acoustic features such as f0, vowel formants, fricative spectra, and creaky voice were correlated with listener judgments using random forest and effect size analyses. Results indicate that well-known features such as f0 and COG of /s/, and less-studied ones like creaky voice and diphthong formants, matter for perception of GI, SO, and GE. Results further suggest that listeners reconstruct a speaker's identity along traditional binaries such as masculine and feminine, and outside those binaries for queer-sounding voices.

Keywords: sociophonetics, speech perception, gender, sexual orientation

1. INTRODUCTION

The current work seeks to investigate the ways in which multiple acoustic features combine to influence perception of speakers' gender identity (GI), sexual orientation (SO), and gender expression (GE). Sociophonetic studies of speech production have identified many correlates of GI, SO, and GE. For American English, these include vowel formants and dispersion [1–2], fundamental frequency (f0) [3–8], acoustic characteristics of voiceless alveolar fricative /s/ [4, 5, 9], and creaky voice [9–10]. This body of work suggests a strong association between certain acoustic cues and a speaker's identity.

During speech perception, studies have further shown that listeners can readily infer the sexual orientation of a speaker from acoustic cues [11–15], again presenting a similar association between acoustic cues and a speaker's (perceived and veridical) identity. Mack and Munson [14] report on one specific feature and its acoustic characteristics, /s/, which was associated with the perception of more homosexual-sounding speech, following a common cultural stereotype about "the gay lisp." Importantly, the effect was strong enough that the actual sexual orientation of the speakers did not matter; when heterosexual speakers' speech was modified to include /s/ with a higher acoustic center of gravity (COG) or as [θ], their speech was rated as more homosexual-sounding. Additionally, f0 has been shown to cue listeners' perceptions of gender identity, with higher f0 values associated with the perception of woman-like individuals and lower f0 values associated with the perception of man-like individuals [7–8]. Furthermore, the perception of gender broadly includes the perception of gender expression as well as gender identity. Similar to the perception of sexual orientation, a number of studies [3, 16–17] show that acoustic characteristics of f0 and /s/ influence the perception of a speaker's voice as more or less masculine-sounding or feminine-sounding, again regardless of the speaker's veridical gender identity.

Presently, there is a need to characterize the interaction of multiple acoustic features with the perception of a speaker's multidimensional social identity. Prior work has often focused on "one-to-one" associations between a single feature and a single dimension (for example, the association between /s/ and sexual orientation in homosexual men) [14]. These one-to-one associations can also be shared across multiple identities such that /s/ cues the perception of sexual orientation in both men [14] and women [5]. However, many-to-one associations are common in speech perception, and the combinations of cues may influence reconstruction of multifaceted identities. The current study seeks to investigate how a listener reconstructs a speaker's gender identity, sexual orientation, and gender expression from speech.

2. METHOD

2.1 Stimuli

Stimuli consisted of short excerpts drawn from interview-format podcasts from American media. Speakers were selected primarily according to their gender identity and sexual orientation, as drawn from as many public sources (articles, videos, profiles, etc.) of a speaker's identity as possible. Excerpts were

controlled for socially-meaningful lexical content. For example, the excerpt for one speaker was "learn more and subscribe."

2.2 Participants

Data was collected from 210 undergraduate students at UCSD via an online survey designed on Qualtrics and recruited from the UCSD undergraduate experimental subjects pool. Each participant self-identified as a native English speaker. All participants were required to use headphones in order to proceed with the survey. Data from 197 participants are reported in the Results section (13 participants were removed due to blank responses in the survey).

2.3 Procedure

Participants were instructed to listen to short speech excerpts of 66 speakers of American English and provide Likert scale judgments of the gender identity (GI), sexual orientation (SO), and gender expression (GE) of each speaker, followed by a demographic survey. Participants were given a practice phase with one speaker not included in analysis, and then completed the main block. All 3 Likert scale dimensions were presented simultaneously for each speaker on a 7-point continuous scale. For GI, the question was "I think this person identifies as…" and responses could be recorded as Male (1), non-binary, gender-non-conforming, genderfluid, etc. (4), and Female (7). For SO, the question was "Would this person seek relationships with people of only the same gender identity, any gender identity, or only an opposite gender identity?" and responses could be recorded as Only same gender (1), Any gender identity (4), and Only opposite gender identity (7). Finally, for GE, the question was "Does this person's voice sound more traditionally masculine or feminine?" and responses could be recorded as Masculine (1), Neither/Both (4), and Feminine (7). Scales were continuous such that participants could provide responses anywhere between the landmarks, e.g. 2.58 or 6.43.

2.4 Feature Extraction

Each speaker excerpt was resampled to 16kHz and automatically segmented via textless forced alignment using Charsiu [18], and segments were hand-corrected by the author. The following acoustic features were then extracted from each excerpt:

1. f0: mean fundamental frequency (f0), f0 standard deviation (std), f0 90th percentile, f0 10th percentile, f0 range, from across the entire utterance
2. Monophthongal Formants: mean, maximum, and minimum F1-F4 of /ɑ æ ʌ ɔ ə ɛ ɪ i ʊ u/
3. Diphthongal Formants: mean, maximum, and minimum F1-F4 for the first and last third of segments of /a͡ʊ a͡ɪ e͡ɪ o͡ʊ o͡ɪ/
4. Speaker Vowel Duration: mean vowel duration across utterance
5. Individual Consonant and Vowel Durations: mean individual segment duration of all segments
6. Vowel Dispersion: distance of mean formant measure of /ɑ æ ʌ ɔ ə ɛ ɪ i ʊ u/ by speaker from group mean formant values
7. Fricative Spectral Features: spectral center of gravity (COG), standard deviation (SD), skew, intensity, duration, and kurtosis of /s z f v ʃ ʒ/
8. Creaky Voice: percentage of creaky voice across utterance

For feature groups 1-6, measurements were derived and calculated from VoiceSauce [19]. For feature group 7, measurements were derived from Praat scripting [20–21], and for feature group 8, measurements were derived via creaky voice detection from COVAREP [22]. In total, 370 individual acoustic features were identified. However, not all excerpts contained a given feature since they were extracted from spontaneous speech in podcasts. For example, only 38 out of 66 speaker excerpts contained an /s/ token. Any missing acoustic value was imputed from the mean of the distribution of that acoustic value across excerpts.

3. RESULTS

3.1 *K*-Means Clustering

Participant ratings of speakers were clustered using *k*-means [23], such that each cluster represents speakers whose GI, SO, and GE vary along similar dimensions in perception. Plotting distortions and using the 'elbow' method indicated that 5 clusters were the best fit for the model. Figure 1 shows each of the 5 clusters and their relative position along the three perceived dimensions: GI, SO, and GE. Cluster labels were assigned post-hoc, based on their relative position along the 3 dimensions, GI, SO, and GE. For example, in Figure 1, the cluster with the most man-like rating for GI, the most same gender rating for SO, and the most masculine-sounding, was categorized as SM, loosely representing the commonly-perceived perceptual category of straight men. The clusters are SM (13 speakers), QM (14 speakers), QE (9 speakers), QW (10 speakers), and SW (20 speakers).

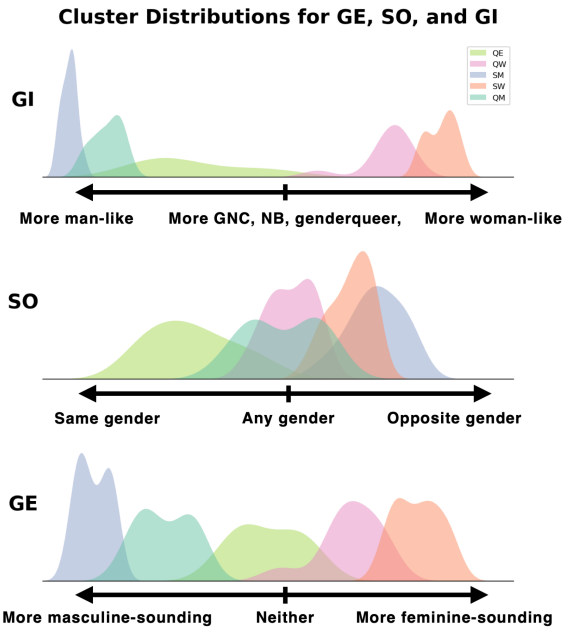


Figure 1: Perceived clusters of speakers according to GI, SO, and GE. Cluster labels are indicated for each of the individual distributions. The clusters are SM (13 speakers), QM (14 speakers), QE (9 speakers), QW (10 speakers), and QE (20 speakers).

Results of the perceptual rating task shown in Figure 1 indicate the emergence of 5 commonly-perceived categories: queer men (QM), queer women (QW), generally queer (non-binary) people (QE), straight men (SM), and straight women (SW).

3.2 Random Forest Analysis

Following clustering, a random forest classifier [23] was trained and tested on the 370 acoustic features with SM, QM, QE, QW, and SW as classes. The random forest classifier establishes the relative importance of the acoustic features for a given cluster, and thus provides a list of features that contribute to the perception of a certain class. The random forest classifier was accurate up to 100% for each cluster on train and test data, and performed at chance (20%) when clusters were randomly-assigned (rather than using the *k*-means clustering results). Top features for each cluster of the 370 features in the model are shown in Table 1. The reported features are truncated for brevity and due to the fact that relative importance values for features begin to plateau within the top 20 features for each cluster.

In an effort to reduce dimensionality further and investigate the apparent similarity in the GI and GE dimensions (see GI and GE in Figure 1), a Pearson correlation analysis [24] was conducted between the participant ratings along the two dimensions and revealed a significant strong correlation ($r = 0.85$, $p < 0.05$). Subsequent analyses are focused on GI and SO only.

Cluster	Features
SM	f0, Monophthongal Formants, Diphthongal Formants, /s/: COG
QM	Diphthongal Formants, Monophthongal Formants, Dispersion
QE	Creaky Voice, Vowel Duration, Monophthongal Formants
QW	f0, Dispersion, Monophthongal Formants, /s/: Duration
SW	f0, Monophthongal Formants, Diphthongal Formants

Table 1: Top features for each cluster via random forest classification.

3.3 Effect Size

In order to investigate the size of the effect of a given acoustic feature on the overall perceptual dimensions GI and SO, the effect sizes of each feature were measured for GI and SO. Linear regressions built with statsmodels [24] using the OLS method were conducted for each feature of interest and the two perceptual dimensions, GI and SO. For example, GI ~ mean f0 and SO ~ mean f0. The coefficient and t-values for select models are reported in Table 2 below.

Results for the effect size analysis show that an ensemble of acoustic features, some expected and others unexpected, contribute significantly to the perception of the GI and SO dimensions. These significant acoustic features were then compared to the 5 clusters to establish what feature or combination of features predict the perception of one cluster identity over another.

4. DISCUSSION

Summarized in Figure 2, certain features remain important to the perception of a single dimension. Replicating some of the results from studies like Mack and Munson [14], voices perceived as belonging to queer men and queer non-binary people can be distinguished from those of perceived straight men, supported by the acoustic characteristics of /s/. Similarly, perceived straight men and straight women's voices can be distinguished from each other on the basis of f0, replicating the effect of f0 on the perception of gender identity (GI) [6–8].

Unexpected features that emerged as important to the perception of identity include creaky voice and diphthongal formants. Drawing attention to the importance of creaky voice in perception as well as production [9–10], queer men, women, and non-

binary people are more easily distinguished from straight men and women via an increase in creaky voice across an entire utterance. Reflecting previous investigations into the idea that queer men share similar perceived speech styles with women [2–4, 12, 14], queer men are distinguished from straight men via an increase in diphthongal formant values, where increases in diphthongal formant values contribute to the perception of more woman-like characteristics along the perceived gender identity (GI) dimension; meanwhile, increases in diphthongal formant values contribute to distinguishing straight women from queer women.

Feature	GI coeff (t-value)	SO coeff (t-value)
f0	0.018 (108.896)	0.002 (12.875)
/ɪ/: F1	0.005 (40.092)	-
/ɪ/: F2	0.001 (48.109)	-
/ɪ/: F3	0.002 (65.992)	0.000 (8.803)
/ɪ/: F4	0.002 (65.78)	0.000 (6.492)
/a͡ɪ/: F1, 1st seg.	0.002 (36.764)	0.001 (10.597)
/a͡ɪ/: F2, 1st seg.	0.001 (24.569)	-
/a͡ɪ/: F3, 1st seg.	0.000 (3.647)	-
/a͡ɪ/: F4, 1st seg.	0.002 (43.149)	0.000 (3.294)
/a͡ɪ/: F1, 3rd seg.	0.002 (36.142)	0.001 (10.69)
/a͡ɪ/: F2, 3rd seg.	0.001 (25.318)	-
/a͡ɪ/: F3, 3rd seg.	0.000 (2.818)	-
/a͡ɪ/: F4, 3rd seg.	0.002 (43.192)	0.000 (3.283)
/ɪ/: F1 distance	0.002 (11.753)	0.000 (-2.715)
/ɪ/: F2 distance	0.001 (22.453)	0.000 (4.043)
/ɪ/: F3 distance	0.000 (5.683)	0.000 (5.672)
/ɪ/: F4 distance	0.000 (-2.867)	0.000 (11.289)
Avg. vowel dur.	-0.001 (-3.297)	-0.004 (-12.34)
/s/: COG	0.000 (34.096)	0.000 (-11.691)
/s/: skew	-0.494 (-32.476)	0.122 (10.006)
/s/: duration	-11.359 (-34.7)	-
Creak (%)	-0.008 (-6.577)	-0.017 (-18.95)

Table 2: Top features with significant effects on GI and SO. Mean feature values were used to compute models. Each cell contains the model coefficient and t-value. "-" indicates non-significance.

Regarding the perceptual dimensions, an asymmetry arises where the variation along perceived sexual orientation (SO) best explains the perception of the man-like individuals; smaller variation along perceived gender identity (GI) explains variation in woman-like individuals. Listeners are overall better at reconstructing different sexual orientations for men

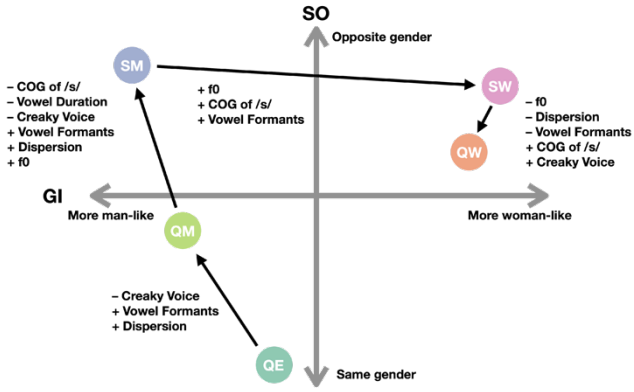


Figure 2: Acoustic features listed alongside arrows show the listener's reconstruction of GI and SO as perception travels between clusters. "+/-" indicate the direction of the effect on acoustic feature values. Vowel formants include both monophthongal and diphthongal formants.

than for women. Hazenburg [5] and Zimman [9] both report relatively constrained ranges of acoustic values produced by straight, man-like voices in comparison to queer men and women, and straight women. Any deviation from those ranges suggested the man-like individual is not straight, which aligns with the observed wider variation in perceived sexual orientation (SO) for man-like individuals. Additionally, Hazenburg [5] and Willis [15] show that any variation in acoustic values that would normally cue the perception of sexual orientation can be attenuated due to an interaction with gender identity. Thus, the more minute differences between the perception of straight and queer women are explained by the perception of their status as woman-like individuals along perceived gender identity (GI).

5. CONCLUSION

The purpose of this study was to investigate the multitude of acoustic features that contribute to a nuanced perception of gender identity, sexual orientation, and gender expression in speakers of American English. The reported associations between speaker identity and expected acoustic features provide further support for existing evidence on the perception of identity, while unexpected acoustic features such as creaky voice and diphthongal formants suggest that the perception of speaker GI, SO, and GE involves more than one acoustic feature. These results illustrate a multidimensional perceptual space in which listeners use many acoustic cues to construct multiple possible speaker identities along traditional binaries, but also create space for speaker identities outside of those binaries, particularly for queer-sounding voices.

6. REFERENCES

[1] J. B. Pierrehumbert, T. Bent, B. Munson, A. R. Bradlow, and J. M. Bailey, "The influence of sexual orientation on vowel production (L)," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 1905–1908, Oct. 2004, doi: 10.1121/1.1788729.

[2] R. Smyth and H. Rogers, "Do gay-sounding men speak like women?," *Toronto Working Papers in Linguistics*, vol. 27, p. 16, 2008.

[3] R. P. Gaudio, "Sounding Gay: Pitch Properties in the Speech of Gay and Straight Men," *American Speech*, vol. 69, no. 1, pp. 30–57, 1994, doi: 10.2307/455948.

[4] S. E. Linville, "Acoustic Correlates of Perceived versus Actual Sexual Orientation in Men's Speech," *Folia Phoniatr Logop*, vol. 50, no. 1, pp. 35–48, 1998, doi: 10.1159/000021447.

[5] E. Hazenberg, "Walking the straight and narrow: linguistic choice and gendered presentation," *GENL*, vol. 10, no. 2, pp. 270–294, Feb. 2015, doi: 10.1558/genl.v10i2.19812.

[6] N. Houle and S. V. Levi, "Acoustic differences between voiced and whispered speech in gender diverse speakers," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 4002–4013, Dec. 2020, doi: 10.1121/10.0002952.

[7] N. Houle and S. V. Levi, "Effect of Phonation on Perception of Femininity/Masculinity in Transgender and Cisgender Speakers," *Journal of Voice*, vol. 35, no. 3, p. 497.e23-497.e37, May 2021, doi: 10.1016/j.jvoice.2019.10.011.

[8] B. Merritt and T. Bent, "Revisiting the acoustics of speaker gender perception: A gender expansive perspective," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 484–499, Jan. 2022, doi: 10.1121/10.0009282.

[9] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men," *JLS*, vol. 2, no. 1, pp. 1–39, Feb. 2013, doi: 10.1075/jls.2.1.01zim.

[10] R. J. Podesva, "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," *Journal of Sociolinguistics*, vol. 11, no. 4, pp. 478–504, 2007, doi: 10.1111/j.1467-9841.2007.00334.x.

[11] R. Smyth, G. Jacobs, and H. Rogers, "Male voices and perceived sexual orientation: An experimental and theoretical approach," *Lang. Soc.*, vol. 32, no. 3, pp. 329–350, Jun. 2003, doi: 10.1017/S0047404503323024.

[12] E. Levon, "Hearing 'Gay': Prosody, Interpretation, and the Affective Judgments of Men's Speech," *American Speech*, vol. 81, no. 1, pp. 56–78, Feb. 2006, doi: 10.1215/00031283-2006-003.

[13] B. Munson, S. V. Jefferson, and E. C. McDonald, "The influence of perceived sexual orientation on fricative identification," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2427–2437, Apr. 2006, doi: 10.1121/1.2173521.

[14] S. Mack and B. Munson, "The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype," *Journal of Phonetics*, vol. 40, no. 1, pp. 198–212, Jan. 2012, doi: 10.1016/j.wocn.2011.10.002.

[15] C. Willis, "Bisexuality and /s/ production," *Proc Ling Soc Amer*, vol. 6, no. 1, p. 69, Mar. 2021, doi: 10.3765/plsa.v6i1.4942.

[16] J. D. Avery and J. M. Liss, "Acoustic characteristics of less-masculine-sounding male speech," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3738–3748, Jun. 1996, doi: 10.1121/1.414970.

[17] B. Munson, "The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation," *Lang Speech*, vol. 50, no. 1, pp. 125–142, Mar. 2007, doi: 10.1177/00238309070500010601.

[18] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A Semi-supervised Approach," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[19] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A Program for Voice Analysis," *Hong Kong*, p. 5, 2011.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. [Online]. Available: www.praat.org

[21] C. DiCanio, "Spectral Moments of fricative spectra script in Praat." 2021.

[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

ATTITUDES IN THE GERMAN-POLISH CONTEXT BASED ON PERCEPTUAL EVIDENCE

Marzena Żygis[1,2] , †Joanna Blaszczak[3], Aleksandra Ćwiek[1], Markus Saint-Pettersen[3], Sarah Wesołek[1,2], Piotr Gulgowski[3]

[1]Leibniz-Centre General Linguistics (ZAS), [2]Humboldt University Berlin, [3]University of Wrocław

zygis@leibniz-zas.de, markus.stpettersen@gmail.com, cwiek@leibniz-zas.de, wesolek@leibniz-zas.de, piotr.gulgowski@uwr.edu.pl

ABSTRACT

The present study investigates how national attitudes/stereotypes influence listeners' perceptions of accented speakers. We studied judgments of Polish listeners towards German and American accented Polish and expected the participants to prefer American over German accent, as implied by the results of a relatively recent national survey of attitudes conducted by a Public Opinion Research Centre (CBOS) [1]. In addition, we hypothesized that the more positive (or negative) judgements would be strengthened when the accented speech is accompanied by a corresponding symbol of the country, i.e., its flag.

Our results revealed no preference of one of the accents over another, nor did the presence of primes affect participants' judgements. They rated the speakers individually and independently from their perceived nationality. The results suggest that preconceived notions of other nationalities are not necessarily evoked when an opinion on an individual speaker is expressed.

Keywords: stereotypes, German, Polish, perception

1. INTRODUCTION

Non-native speakers of a language often encounter difficulties understanding speech and being understood. They also may find themselves the target of prejudice from both native speakers and other non-native speakers alike. These prejudices may be directed at foreigners in general, but an accent can specifically indicate their origins/nationality and evoke particular negative or positive attitudes.

Negative attitudes towards non-native speakers might have different reasons. On the one hand, the cognitive load and the need to adopt new strategies may subconsciously lead to a more negative judgement of the accented speaker. On the other hand, listeners may carry prejudices and stereotypes either about speakers using accented speech in general or about nationalities associated with specific accents.

Previous research has shown that speech perception is not only based on the acoustic signal but also integrates non-linguistic information.

Niedzielski [2] demonstrated that phonemes are perceived differently when accompanied by (true or false) information about the speaker's identity. Hay & Drager [3] showed that perceptions can be influenced by exposure to a symbol of a geographical region. New Zealander participants exposed to a kangaroo were more likely to perceive vowels on a synthesised continuum as spoken by an Australian English speaker. Conversely, participants who saw a kiwi were more likely to perceive the same vowels as spoken by a New Zealand English speaker. Jannedy et al. [3] examined German attitudes towards speakers of Hood German, a sociolect associated with the immigrant communities of Berlin. The evaluations of the speakers were more favourable when they were presented with a French name than when the same speakers were presented with a Turkish name. In summary, the studies show how listeners' judgment can be affected by non-linguistic factors.

The present study continues this line of research by investigating how attitudes towards speakers that may otherwise remain undisclosed can be revealed by means of perceptual evidence. In particular, we scrutinize how Polish listeners judge German-accented Polish versus American-accented Polish. The choice of the accents was motivated, among others, by considerably different attitudes Polish speakers expressed towards German as opposed to American people in a relatively recent national survey of attitudes conducted by a Public Opinion Research Centre (CBOS [1], 2020, N = 958 participants): 50% Poles have positive attitudes towards American people and 12% negative (30% remained indifferent and 8% did not know). In the same survey 36% Poles were positive towards Germans but 29% negative (29% indifferent, and 6% did not know). Taking the results into account the present study examines the attitudes of Polish speakers towards American and German speakers by means of indirect evidence gained from a perceptual test.

2. EXPERIMENT

2.1. Hypotheses

Polish participants are expected to prefer American accented speech over German accented speech. The more positive judgements should also be strengthened when the accented speech is accompanied by a corresponding symbol of the country (a flag).

2.2. Experimental design

To test our hypotheses two on-line perceptual experiments were conducted through SoSci Survey (Leiner [5]), an online questionnaire platform.

In both experiments, we probed how Polish listeners perceive American and German accents in Polish with the only difference being in the presence (Experiment 1) or absence of a flag (Experiment 2). The corresponding flag was displayed in the top right corner of the screen mirrored the intended accent of the recording.

Recordings of four female German speakers speaking German and American English (AE) accented Polish were included in each experiment. The speakers were asked to imitate German and American accent while speaking Polish. They were selected from a pool of 6 female speakers by two phoneticians speaking all three languages (German, Polish, and English) who paid attention to both an overall acoustic impression of the accent and their most prominent phonetic cues, e.g., the retroflex /ɽ/ in American English, absent in both German and Polish. Each speaker produced several different sentences during conversations with another speaker (two speakers pronounced stimuli with American accent and two with German accent), and forty sentences of similar length were selected and presented in each experiment. The stimuli were automatically randomized for each participant. Table 1 provides an overview of both experiments.

	Listeners	Flag	Speakers	Accent
Exp1	Polish (43)	YES	German (4)	American/ German
Exp2	Polish (47)	NO	German (4)	American/ German

Table 1: An overview of Experiment 1 and 2

The participants' task was to listen to a recording and rate the speaker on a 7-point Likert scale. Opposing adjectives were presented at the extreme ends of the scale, e.g., *serious* (left extreme point) and *not serious* (right extreme point). No numbers were displayed on the scale.

There were eight adjectives determining the competence of the speaker (*strong, certain, serious, competent, extravert, independent, ambitious, dominant*) and seven adjectives describing their benevolence (*warm, likable, sensible, helpful, sociable, emotional, honest*; cf. Brown et al. [6] and Brown [7] for the concept of competence and benevolence). Please note that the listeners were most probably thinking that they were listening to German and American speakers while in fact they were only exposed to German speakers.

2.3. Participants

Experiment 1 yielded data from 59 listeners, 43 of whom were included in the further analysis. The average participant age was 21.3 (1.88 sd). In the second experiment data from 47 participants out of 59 were evaluated. Their average age was 21.6 (2.52 sd). Overall, the ratings from 28 participants could not be included because they either did not finish the experiment or they were not native speakers of Polish.

2.4. Statistics

All statistical analyses were conducted with R (version: 4.0.0 [8]). We used ordinal regression analysis (package "ordinal" by Christiansen [9]) to investigate the influence of LANGUAGE ACCENT, i.e., German vs. American accent in Experiment 1 (section 3.1) and Experiment 2 (section 3.2) on ANSWER [1:7]. We also included INFORMANT as a random intercept and LANGUAGE ACCENT as its slope. In addition, in a separate model (section 3.3.) we linked the results from Experiment 1 and Experiment 2 and included FLAG as a fixed factor with two levels [flag; no flag] and added the interaction LANGUAGE ACCENT * FLAG to the model. The random structure was identical as in the previous models.

7697 measurement points from Experiment 1 and 7755 from Experiment 2 were submitted to the statistical analysis.

3. RESULTS

3.1. Experiment 1

Our results show that Polish with German accent was in general rated more "positively" than American accent. This general pattern, however, slightly differed when the attitude groups were examined in detail. While Benevolence showed a clearly parallel pattern, for Competence, the ratings varied depending on the adjective, see Figure 1.

The adjectives, for which the ratings (almost) converged, were *serious, extravert, independent, and dominant*. Thus, it appears that Polish speakers perceived German accented participants as more benevolent (mean POL_GER 3.01 vs. mean POL_AME 3.87; t = -7.91, p < .001) and more competent than American accented participants (mean POL_GER 3.46 vs. mean POL_AME 3.78; t = -3.77, p<.001).

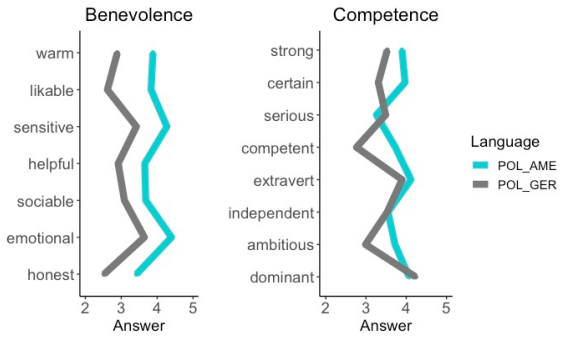


Figure 1: Results obtained for Polish stimuli with American and German accent split for Benevolence and Competence (presented with flags)

If we split the results according to ratings obtained for individual speakers, a slightly different picture emerges. Regarding Competence it becomes evident that one speaker with American accent was judged more negatively (mean POL_AME1 = 2.93 vs. mean POL_AME2 = 4.64; t = -11.41, p<.001); see Figure 2 (right). The two speakers with German accent were rated differently as well, although the difference was slightly smaller (mean POL_GER1 = 2.93 vs. mean POL_GER2 = 4.64; t =-9.02, p < .001).

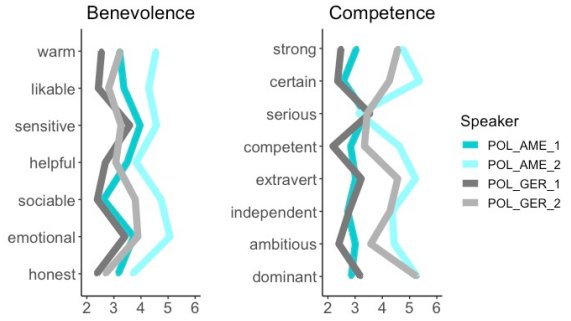


Figure 2: Results obtained for Polish stimuli with American and German accent split for Benevolence and Competence and individual speakers (with flags)

Differences between individual speakers were also found in Benevolence ratings, see Figure 2 (left). The first speaker with German accent was rated as more benevolent than the second speaker (mean POL_GER1 = 2.77 vs. mean POL_GER2 = 3.24; t = -3.75, p <.001). The difference between speakers with American accent was significant, too, whereby the first speaker was rated as more benevolent than the second speaker (mean POL_AME1 = 3.36 vs. mean POL_AME2 = 4.39, t=-5.47, p < .001).

3.2. Experiment 2

In the second experiment, the same stimuli as in Experiment 1 were used, yet no flag was displayed on the screen. In this condition, there are again no clear patterns discernible if the results are concatenated for all adjectives. However, splitting the results according to Benevolence and Competence unveils that German accent was rated as more benevolent (mean POL_GER = 3.06 vs. mean POL_AME = 4.06, t = -9.47, p < .001) than the American accent. Figure 3 (left) shows that for Benevolence, the ratings of German and American accent do not overlap. In the case of Competence, German accent was judged more competent only with respect to five out of eight adjectives. Taken together the difference between German and American accents was still significant leading to a conclusion that German accent was perceived as more competent (mean POL_GER = 3.58 vs. mean POL_AME = 3.81, t = -2.72, p < .01; see Figure 3 (right).



Figure 3: Results obtained for Polish stimuli with American and German accent split for Benevolence and Competence (without flags)

If we consider rankings assigned to individual speakers, considerable differences become evident. First, as far as Benevolence is concerned, both speakers with German accent were generally judged more benevolent than speakers with American accent, see

Figure 4 (left). The difference between the speakers within an accent was significant: speaker 2 with German accent was rated as less benevolent than speaker 1 (mean POL_GER1 = 2.62 vs. mean POL_GER2 = 3.42, t = -7.53, p < .001). Similarly, speaker 2 with American accent was perceived as less benevolent than speaker 1 (mean POL_AME1 = 3.58 vs. mean POL_AME2 = 4.56, t = -7.99, p< .001).

As for Competence, the speakers within the same accent were rated significantly different as well. For Competence, the differences were larger than in the

Benevolence context for German (mean POL_GER1 = 2.63 vs. mean POL_GER2 = 4.42, t = -14.79 p < .001) and American accent (mean POL_AME1 = 3.14 vs. mean POL_AME2 = 4.48, t = -8.53, p < .001); see Figure **4** (right).

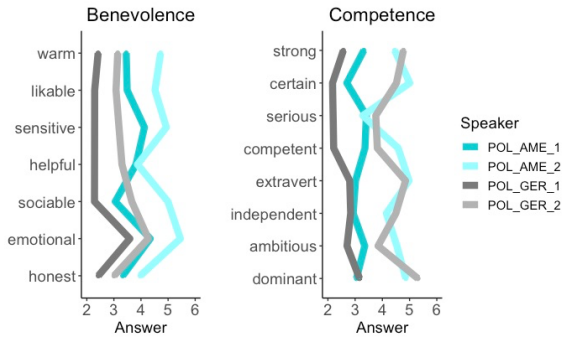


Figure 4: Results obtained for Polish stimuli with American and German accent split for Benevolence and Competence and individual speakers (without flags)

3.3. Comparison of Experiment 1 and 2

Finally, if we compare results from Experiment 1 (with flag) and 2 (without flag) there are only very small differences for both Benevolence (see Figure 5) and Competence (see Figure 6).

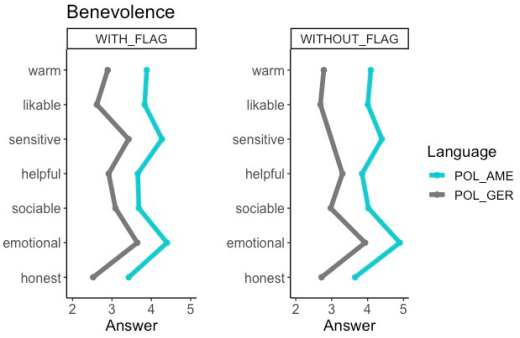


Figure 5: Comparison of results obtained for Polish stimuli with American and German accent split for flag and without-flag condition for Benevolence

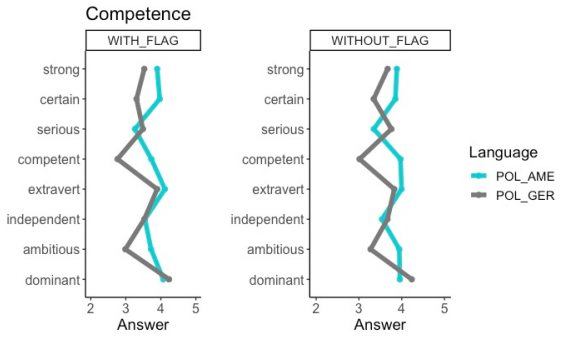


Figure 6: Comparison of results obtained for Polish stimuli with American and German accent split for flag and without-flag condition for Competence

With regard to Benevolence, the American accent was rated as more benevolent when the flag was shown but the presence of the German flag did not influence the ratings of the German accent (Flag*Language: t= -3.52, p<.001). Still, the German accent was perceived as more benevolent than the American one (t=-10.46, p<.001).

Furthermore, in the case of Competence, the influence of the flag was not significant, neither was the interaction of Flag and Language. The only significant effect was that of Language where German was rated as more competent (t =-4.50, p<.001); see Figure 6.

4. CONCLUSIONS

The aim of this study was to test attitudes towards German and American people by means of an indirect evidence, i.e. a perceptual test of the respective accented speech.

Two hypotheses have been put forward. Firstly, Polish participants prefer the American-accented version of their language to the German-accented version. This hypothesis was based on a CBOS study [1]. Secondly, we hypothesized that participants' judgements are strengthened when the accented speech is accompanied by a corresponding national symbol, that is, its flag. However, neither of these hypotheses were supported by the results.

Although Polish participants rated German-accented speakers as generally more benevolent and competent than American-accented speakers, a closer inspection of rating of individual speakers revealed a considerable variance between individual speakers within the two accent groups. Responses evoked by inter-speaker differences in voice quality may outweigh the attitudes that listeners hold towards specific accents or nationalities. In other words, while people may have preconceived notions of other nationalities, they do not rely exclusively on them when forming an opinion on an individual speaker. This finding was consistent even when participants saw the national flags.

However, there may be another explanation as to why the hypotheses were not supported. Firstly, the idea that Polish participants would rate German accented speakers more negatively originated mainly from the CBOS [1] survey. In that survey, the sample was representative of the adult population of Poland whereas the Polish participants in the present study were on average young. Therefore, it is possible that the attitudes found in the CBOS "representative" population are not as prevalent as those found among young people which in turn suggests a change of attitudes of Poles towards Germans.

Acknowledgments
Sarah Wesolek and Piotr Gulgowski were supported by the Deutsch-Polnische Wissenschaftsstiftung and Marzena Zygis by the Leibniz Society.

5. REFERENCES

[1] CBOS. Centrum Badania Opinii Społecznej 2020. Attitude to other nationalities. CBOS [Public Opinion Research Centre]. Warsaw, Poland: CBOS.

[2] Niedzielski, N. 1999. The effect of social information on the phonetic perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1), 62–85.

[3] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48, 865-892.

[4] Jannedy, S., Mentoza-Denton, N., Weirich, M. 2019. Social capital and the production and perception of fine phonetic detail in Berlin. In: Heyd, T., von Mengden, F., Schneider, B. (eds.) Language and Social Life [LSL] 17. Berlin: DeGruyter, 25-144.

[5] Leiner, D. J. 2019. SoSci Survey (Version 3.1.06) [Computer software]. Available at https://www.soscisurvey.de

[6] Brown, B.L.; Strong, W.J.; Rencher, A.C. 1975. Acoustic determinants of perceptions of personality from speech. *International Journal of the Sociology of Language* 6, 11-32.

[7] Brown, Bruce, L. 1982. Experimentelle Untersuchungen zur Personenwahrnehmung aufgrund vokaler Hinweisreize. In: Scherer, K. (ed.): Vokale Kommunikation. Basel: Beltz Verlag, 211-227.

[8] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 4.0.0. URL https://www.R-project.org/.

[9] Christensen, R.H.B. 2018. Ordinal. R package. https://cran.rproject.org/web/packages/ordinal/ordinal.pdf.

Indexing femininity in vowel acoustics: A comparison between speakers in eastern and western parts of Germany

Melanie Weirich[1] & Katharina Zahner-Ritter[2]

[1]Friedrich Schiller University, Jena, Germany, [2]University of Trier, Germany
melanie.weirich@uni-jena.de, k.zahner-ritter@uni-trier.de

ABSTRACT

Previous research has identified cross-linguistic differences in the relation between gender identity and acoustics. This study tests whether such differences extend to different cities in the same country (in western and eastern parts of Germany, formerly BRD and GDR, with distinct historical developments). Results show that gender role concepts vary between regions, with males in the eastern city showing more egalitarian views and higher femininity than males in the west. On the group level, the difference in gender identity was not evidenced in the acoustic vowel space. On an individual level, variation in vowel productions could however be explained by femininity, especially with respect to the second formant (front-back dimension). The effect of femininity was modulated by city, with speakers in the east indexing femininity in vowel productions while speakers in the west did not. Implications for the analysis of inter-speaker variation in fine-phonetic detail is discussed.

Keywords: socio-phonetics, vowel spaces, formants, gender identity, femininity

1. INTRODUCTION

Research on phonetic differences between male and female voices has shown that perceptual, biological, and social factors explain gender-specific variation in speech [1-3]. At the same time, the conception of *gender* in western cultures is moving away from a strictly dichotomous term. Instead, gender is expressed and perceived on a more fluid scale; people may even choose to reject this categorization altogether.

Phonetic studies highlighting the significance of a speaker's *gender identity* in explaining acoustic-phonetic parameters (independent of sex) are still scarce. Several studies have investigated the impact of a speaker's sexual orientation on fine-phonetic detail: While some found differences in mean fundamental frequency (F0) between straight and gay speakers [4,5], others did not [6-8]. For German, [9] found lower mean first vowel formants (F1) in straight than in gay men. Fewer studies have investigated female voices and results are similarly inconclusive with regard to the influence of sexual orientation ([10,11] vs. [12,13]). One reason for these mixed results might be the binary distinction of sexual orientation (cf. [14]). Specifically, [14] emphasize the importance of a fine-grained analysis of psychological characteristics including gender identity in addition to sexual orientation: While there was no acoustic difference between straight and non-straight women, intra-group variability in F0 and vowel formants was related to the exclusivity of sexual orientation (intermediate score between exclusively gay/hetero) and gender-role concept. Independent of sexual orientation, [15] found that males indexed self-reported femininity scores in speech through F0 and vowel space sizes, and listeners reliably detected this information on one-word stimuli [15]. Similar results were found for females with correlations between self-ascribed and perceived femininity and hints towards the role of F2 [16]. Cross-linguistic studies have revealed that gender-specific variation is affected by cultural norms and expectations: [17] highlighted cultural aspects of differences in F0 between Dutch and Japanese women. [18] found larger differences between genders in F0 and vowel space size in German speakers than in Swedish speakers, due to lower values for Swedish females than for German females, accompanied by lower self-ascribed femininity in Swedish females.

This study tests whether such cross-linguistic differences in the relation between gender identity and acoustics also pertain to different cities within one country. We focus on two cities in Germany that differ in their historical/societal developments: Jena (eastern Germany, former GDR) and Trier (western Germany, former BRD) – both university cities of about 110.000 inhabitants. The historical difference between the two cities dates back to the division of Germany into a capitalist democracy in the West and a socialist state in the East after World War II. These differences in political systems were accompanied by varying gender roles with women working full time shortly after having children in the East while staying at home as housewives in the West. Even well after reunification, there are still economic and societal differences between the two regions: Crucially, they differ in the amount of females in leading positions (more in eastern G), gender-based salary differences (in favor of females in eastern G [19]), and in labor division (more equal in eastern G [20]). Regarding dialectal differences, the two cities are comparable in their region-specific vowel space sizes [21]. Against this background, we address three research questions:

RQ1: Do females and males differ in self-ascribed gender identity (femininity) and views on gender roles across cities (higher fem. and more egalitarian gender-role concept in males in Jena than in Trier)?

RQ2: Are these potential regional differences reflected in gender-specific variation in vowel spaces (larger diff. between sexes in Trier than in Jena)?

RQ3: Does the self-ascribed gender identity explain inter-speaker variability **within the sexes** in vowel formants (F1, F2) and if yes, does this differ between regions?

2. METHOD

The study was approved by the Ethics Committee of the Friedrich Schiller University, Jena (2019-1389-BO); all speakers gave their written consent.

2.1. Participants

We recruited speakers between 30 and 50 years who were born and raised in Jena (eastern Germany) or Trier (western Germany). In total, 37 speakers from Jena and 25 from Trier participated for a small payment, see Table 1. The groups (per gender) did not differ in age, height, schooling, sexual orientation, as well as how much they liked their city (all p>0.12).

	Jena		Trier	
	f	m	f	m
N	15	22	15	10
Mean age in years (SD)	36.53 (7.19)	38.59 (4.44)	38.10 (6.64)	38.40 (5.38)

Table 1: *Number of participants per city and sex.*

2.1.1. Socio-psychological parameters

GEPAQ_Fem: Self-ascribed gender identity was assessed via the positive femininity scale of GEPAQ (*German Extended Personality Attributes Questionnaire* [22]) which was constructed to determine the personality-based gender-role self-concept. The scale consists of eight adjective pairs describing positively evaluated personality traits traditionally attributed to women (e.g., kind, gentle). Participants were asked to rate themselves on these adjective pairs on a scale from 1 to 7 (e.g., 1=very unkind, 7=very kind). Mean values were calculated for each speaker, with higher scores reflecting higher femininity.

NGRO: The Normative Gender Role Orientation Questionnaire [23] measures explicit stereotypes towards women and men. Here, we selected nine items to measure participants' view on gender roles. Participants indicated the extent to which they agree with these statements on a 7-point scale (1=I strongly disagree; 7=I strongly agree). Higher means indicate a more egalitarian view on gender roles.

SO: Sexual orientation was measured using a Kinsey-like scale ranging from 1 (exclusively heterosexual) to 7 (exclusively gay/lesbian).

2.2. Speech material

A picture-description task was used to elicit semi-spontaneous speech data (cf. [15], [18], [24]). Participants were asked to describe 15 different pictures that showed several objects and animals, see Table 2. Pictures were designed to contain several carrier words with the target vowels in stressed position, e.g., *Tiger* (tiger), *Biene* (bee) for /i:/; *Äpfel* (apples) for /ɛ/; *Katze* (cat), *Tasse* (cup) for /a/; and *Kuh* (cow), *Buch* (book) for /u:/. Table 2 shows a summary of the number of vowels in the dataset, split by *city* and *sex*.

	Jena		Trier	
	f	m	f	m
i:	217	314	246	132
ɛ	61	94	85	46
a	202	320	246	139
u:	185	293	214	126



Table 2: *Observations included per group (left) and example picture showing target words (right).*

2.3. Phonetic analyses: Vowel acoustics

The first two formants (F1, F2) in the vowels (/i: a ɛ u:/) of the carrier words were measured at the vowel midpoint using Praat's [25] linear predictive coding formant measurement algorithm. Analysis parameters included a time step of 0.01s, a maximum number of 5 formants, a window length of 0.025s, and a pre-emphasis from 50 Hz. The maximum formant value was set to 5500 Hz for females and 5000 Hz for males. For /u:/, the maximum number of formants was set to 2 and the maximum formant value to 1200 for males and 1500 for females. Erroneous measurements were checked and corrected manually. To get an estimation of vowel dispersion, the polygon size spanned by F1 and F2 of the corner vowels (/i: a ɛ u:/) was calculated for each speaker.

3. RESULTS

3.1. Femininity and gender role in Jena and Trier

Figure 1 shows the self-ascribed femininity scores (GEPAQ_Fem; Cronbach's α=0.82) and gender-role orientation (NGRO, Cronbach's α=0.53), separated by city and gender. Fig. 1 suggests that the difference in the scales between sexes is smaller in Jena (similar values across sex) than in Trier (lower values in males than in females). Compared to males in Jena, males in Trier showed a less egalitarian view on gender roles (NGRO, t=2.6, df=14.5, p=0.02) and a tendenc

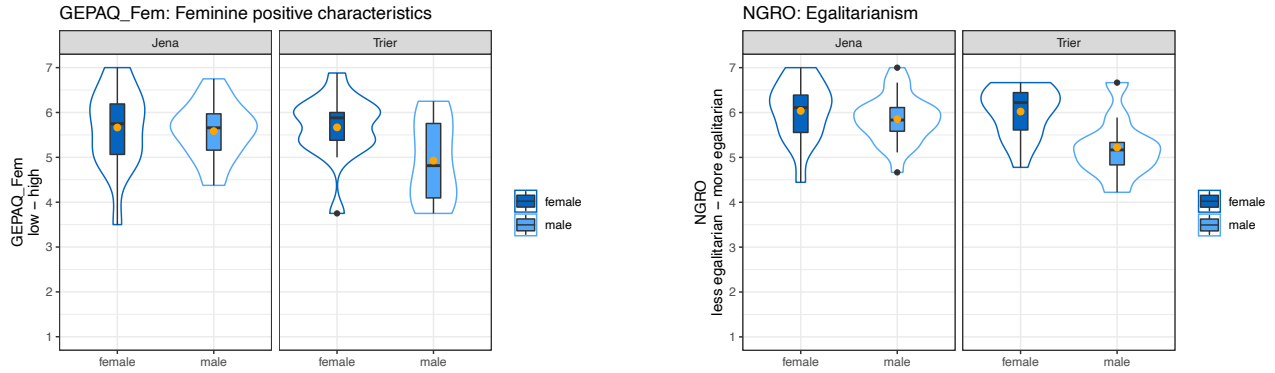


Figure 1: Self-ascribed gender identity in Jena (east) and Trier (west); GEPAQ_Fem left and NGRO (egalitarianism) right.

for lower femininity (GEPAQ_Fem, t=2.05, df=12.96, p=0.06, **RQ1).** Females in Trier vs. Jena did not differ in femininity or egalitarianism (both p>0.94). Responses for both scales were significantly correlated (r=0.4, p=0.001). To analyse the relation between gender identity and vowel acoustics, we use GEPAQ_Fem, which had a higher scale reliability.

3.2. Variability in vowel acoustics and femininity

Figure 2 shows the acoustic vowel spaces in Jena and Trier **(RQ2).** The female polygon spanned by F1 and F2 of the four corner vowels (/i: a ɛ u:/)) was 0.45 kHz^2 (SD=0.09) in Jena and 0.42 kHz^2 (SD=0.07) in Trier. The male polygon was 0.25 kHz^2 (SD=0.05) in Jena, and 0.26 kHz^2 (SD=0.05) in Trier. Not surprisingly, females had a significantly larger vowel space than males (ß=0.18, SE=0.02, t=10.86, p<0.001). There was, however, no interaction between *sex* and *city* (p=0.35), indicating that the difference in vowel space between males and females was not different in the two cities. Hence, the difference in self-ascribed gender identity between Jena and Trier (Figure 1) was not indexed in the acoustic vowel space on the group level **(RQ2)**, see Figure 2.

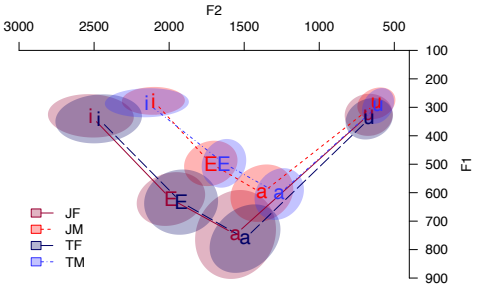


Figure 2: Acoustic vowel space for speakers in Jena (red) and Trier (blue); females in dark, males in light.

We will now turn to the third research question **(RQ3)** on whether inter-speaker variation within the sexes can be accounted for by individual gender identity. We report on a potential relationship between vowel acoustics (in terms of F1 or F2) and GEPAQ_Fem (femininity). We will investigate the first two formants instead of vowel space to base our analysis on more data points, and to be able to test effects of gender identity for individual vowels. In addition to *vowel*, we include *city* and *sexual orientation* as potentially interacting factors with the effect of *gender identity*. In response to **RQ3**, we run separate linear mixed effects regression models [26] for each sex for both F1 and F2, testing for an effect of GEPAQ_Fem and its interaction with the factors *vowel*, *city*, and *sexual orientation*. Categorical fixed factors were dummy coded (*vowel*=/a/ and *city*=Jena as reference levels). For convergence reasons, *sexual orientation* was grouped into straight (Kinsey-Scale=1: n=36 speakers) and non-straight persons (Kinsey-Scale=2-7, n=26 speakers). *Speaker* and *word* were included as random effects (by intercept adjustments). Model comparisons based on LogLikelihood (using anova()) indicated the final model.

F1. The model for **females** revealed an interaction beween GEPAQ_Fem**vowel** ($\chi^2(6)=131.2$, p<0.001) such that for the vowels /a/ and /ɛ/ the relation was slightly negative (F1 decreases with higher femininity), while /u:/ and /i:/ showed no relation. The effect of femininity is, however, small (ß =-17.3 for the reference /a/) and in an unexpected direction. There was no interaction with city or sexual orientation. The model for F1 for **males** did not reveal an effect of femininity (neither as main effect nor in interaction with the other factors), only the expected main effect of vowel type ($\chi^2(3)=134.7$, p<0.001).

F2. The final model for **females** revealed a significant interaction between GEPAQ_Fem**vowel** ($\chi^2(6)=126.2$, p<0.001) and a marginally significant interaction between GEPAQ_Fem**city** ($\chi^2(2)=4.9$, p=0.08), see Figure 3 (left panel) for predicted values for F2. In particular, for the vowel /a/ in females from Jena (reference level), F2 significantly increased with higher femininity (ß=85.4), i.e., the vowel is more fronted. This relation was also observed for the three other vowels, but to a smaller extent (for /i:/: ß=-26.11; for /ɛ/: ß=-47.48, for /u:/: ß=-62.12).

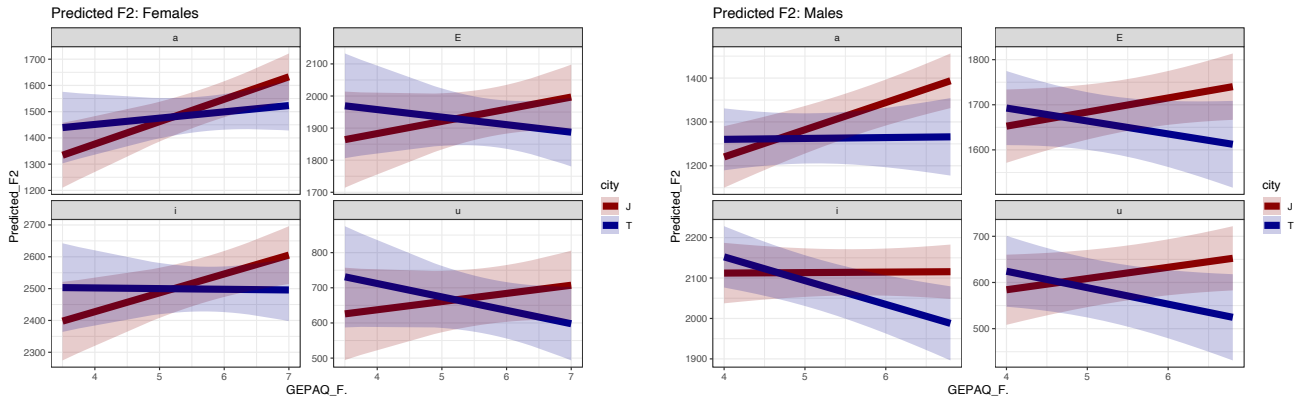


Figure 3: Predicted values for F2 for female (left) and male speakers (right) for individual vowels; speakers from Jena are coded in red, speakers from Trier in blue. Note that y-axes scaling is different for individual vowels.

The effect of femininity on F2 was smaller in females in Trier than in females in Jena (ß=-61.41 with respect to the reference /a/). For other vowels, there is no or a (slight) negative relation in females in Trier.

The final model for **F2 in males** revealed significant interactions between GEPAQ_Fem*_vowel_ ($\chi^2(6)=150.9$, $p<0.001$) and GEPAQ_Fem*city ($\chi^2(2)=9.2$, $p<0.01$), see Figure 3 (right panel) for predicted values for F2. Similar to females in Jena, F2 for male speakers from Jena significantly increased with higher femininity for the vowel /a/ (reference level, ß=62.0; i.e., more fronted); this relation was present but smaller for /ɛ/ and /u:/ (ß=-30.7 and -7.7), but lacking for /i:/ (ß=-60.8). Also, the effect of femininity on F2 was different in male speakers from Trier than in males from Jena (ß=-60.00, with respect to the reference level), see also Figure 3 (no other or even negative relationships). Sexual orientation did not play a role (as main effect or interaction) in the models for F1 or F2 (all $p>0.5$).

4. DISCUSSION

Extending previous socio-phonetic studies on cross-linguistic differences in gender-related phonetic variability [18], the present study assessed the self-ascribed gender identity in two cities in Germany, which differ in their historical background and societal development. As expected, gender identity – captured via the Femininity Scale of the GEPAQ [22] and the NGRO [23] – differed between the sexes to varying degrees in the two cities Jena (eastern Germany) and Trier (western Germany). While female speakers' gender role orientations did not differ across cities, male speakers in Jena showed a more egalitarian view on gender roles and ascribed themselves higher femininity (RQ1). However, on a group level, these differences in ascribed gender identity were not directly evidenced in the vowel space sizes (RQ2).

Importantly, inter-speaker variation in vowel productions could be explained by individual scores for self-ascribed femininity, especially with respect to the second formant (F2, RQ3); sexual orientation did not play a role. The effect of femininity was shaped by vowel category and city. While higher femininity values clearly go along with higher F2 (more fronted vowels) in male and female speakers from Jena, this relation is absent or even negative in speakers in Trier (especially in men, but given the small sample size in this group, interpretation needs to remain cautious). Hence, while speakers who consider themselves more feminine front vowels in Jena, speakers in Trier do not. With regards to vowel quality, our results point to a strong effect of femininity on /a/, followed by /ɛ/ for males and /i:/ for females.

Earlier studies have shown the relevance of F2 regarding the indexicality of gender identity in fine phonetic detail ([14, 16]). However, in light of the differences between the cities found in the present study, generalization on the social meaning of F2 regarding femininity should be made with caution (see [14] for similar regional differences). Our results clearly point to a regionally influenced phonetic encoding of gender identity in speech – within one language, potentially affected by differences in historical developments – which, in turn, might have influenced people's normative gender-role concepts.

Along with other socio-phonetic studies that point to locally constructed social meaning of phonetic variation [27], our study emphasizes that the analysis and explanation of inter-speaker or inter-group variation in fine phonetic detail needs to incorporate a number of social factors, and to consider ideologies, norms and expectations of both speakers and listeners, which are shaped by the local environment, society, and history. The reasons underlying the regional differences in indexing femininity in speech yet remain to be detected. A possible reasoning is that it might be more important for speakers from Jena to index their femininity in speech since differences between the genders in many aspects of life have become smaller. Using fine phonetic detail is hence one source of variation to express one's female identity.

5. ACKNOWLEDGEMENTS

We thank D. Kanngießer, A. Maljavin, L. Hilliger (Jena), and V. Meyer, M. Loes, M. Hildebrand, N. Elsässer (Trier) for help with testing and annotation, and all participating subjects. The project was funded by the DFG (WE5757/3-1).

6. REFERENCES

[1] G. Fant, 'A note on vocal tract size factors and non-uniform F-pattern scaling', *STL QPSR*, vol. 4, pp. 22–30, 1966.

[2] M. Weirich, S. Fuchs, A. P. Simpson, R. Winkler, and P. Perrier, 'Mumbling: macho or morphology?', *Journal of Speech, Language and Hearing Research*, vol. 59, no. 1587–1595, 2016.

[3] M. Weirich and A. P. Simpson, 'Investigating the relationship between average speaker fundamental frequency and acoustic vowel space size', *JASA*, vol. 134, no. 4, p. 2965-74, 2013.

[4] H. Baeck, P. Corthals, and J. V. Borsel, 'Pitch Characteristics of Homosexual Males', *Journal of Voice*, vol. 25, no. 5, pp. e211–e214, 2011.

[5] A. Suire, A. Tognetti, V. Durand, M. Raymond, and M. Barkat-Defradas, 'Speech acoustic features: A comparison of gay men, heterosexual men, and heterosexual women', *Arch Sex Behav*, vol. 49, no. 7, pp. 2575–2583, 2020.

[6] J. W. Lerman and P. H. Damsté, 'Voice pitch of homosexuals', *Folia Phoniatr Logop*, vol. 21, no. 5, pp. 340–346, 1969.

[7] D. Rendall, P. L. Vasey, and J. McKenzie, 'The Queen's English: An alternative, biosocial hypothesis for the distinctive features of "Gay Speech"', *Arch Sex Behav*, vol. 37, no. 1, pp. 188–204. 2008.

[8] S. Sulpizio *et al.*, 'The sound of voice: Voice-based categorization of speakers' sexual orientation within and across languages', *PLoS ONE*, vol. 10, no. 7, p. e0128882, 2015.

[9] S. Kachel, A. P. Simpson, and M. C. Steffens, '"Do I sound straight?": Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech', *J Speech Lang Hear Res*, vol. 61, no. 7, pp. 1560–1578, 2018.

[10] M. Camp, 'Japanese lesbian speech: Sexuality, gender identity, and language', Dissertation, Department of East Asian Studies, The University of Arizona, Columbus, OH, 2009.

[11] B. Moonwoman-Baird, 'Toward the study of lesbian speech', in *Queerly phrased,* A. Livia & K. Hall, Oxford: Oxford University Press, 1997, pp. 202–213.

[12] B. Munson, E. C. McDonald, N. L. DeBoe, and A. R. White, 'The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech', *Journal of Phonetics*, vol. 34, no. 2, pp. 202–240, 2006.

[13] S. Sulpizio, F. Fasoli, R. Antonio, F. Eyssel, M. P. Paladino, and C. Diehl, 'Auditory *Gaydar*: Perception of sexual orientation based on female voice', *Lang Speech*, vol. 63, no. 1, pp. 184–206, 2020.

[14] S. Kachel, A. P. Simpson, and M. C. Steffens, 'Acoustic correlates of sexual orientation and gender-role self-concept in women's speech', *JASA*, vol. 141, no. 6, pp. 4793–4809, 2017.

[15] M. Weirich and A. P. Simpson, 'Gender identity is indexed and perceived in speech', *PlosOne*, vol. 13, no. 12, p. e0209226, 2018.

[16] M. Weirich and A. P. Simpson, 'Do you hear what I say? The expression and perception of femininity in female voices', Paper presented at the Workshop on Social Meaning, Leibniz-ZAS Berlin, Leibniz-ZAS Berlin, 2019.

[17] R. van Bezooijen, 'Sociocultural aspects of pitch differences between Japanese and Dutch Women', *Lang Speech*, vol. 38, no. 3, pp. 253–265, 1995.

[18] M. Weirich, A. P. Simpson, J. Öjbro, and C. Ericsdotter Nordgren, 'The phonetics of gender in Swedish and German.', *Proceedings of FONETIK 2019*, pp. 49–53, 2019.

[19] Agentur-für-Arbeit Deutschland, 'Day of German Unity: In East Germany, women continue to earn more than men'. Accessed: Dec. 21, 2022. [Online]. https://www.arbeitsagentur.de/presse/2022-43-ostdeutsche-frauen-verdienen-mehr-als-meanner

[20] Bundesministerium für Familie Frauen und Jugend (Hrsg.), '25 Jahre Deutsche Einheit. Gleichstellung und Geschlechtergerechtigkeit in Ostdeutschland und Westdeutschland', 2015.

[21] B. Siebenhaar and M. Hahn, 'Vowel space, speech rate and language space', in *Proceedings of the International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 879-883, 2019.

[22] T. E. Runge, D. Frey, P. E. Gollwitzer, R. L. Helmreich, and J. T. Spence, 'Masculine (instrumental) and feminine (expressive) traits: A comparison between students in the United States and West Germany', *J Cross Cult Psychol*, vol. 12, no. 2, pp. 142–162, 1981.

[23] U. Athenstaedt, 'Normative Geschlechtsrollenorientierung: Entwicklung und Validierung eines Fragebogens', *Zeitschrift für Differentielle und Diagnostische Psychologie*, vol. 21, pp. 91–104, 2000.

[24] M. Weirich and A. P. Simpson, 'Effects of gender, parental role, and time on infant- and adult-directed read and spontaneous speech', *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 11, pp. 4001–4014, 2019.

[25] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer, version 6.2.23 [Computer program]. http://www.praat.org/', 2022.

[26] R. H. Baayen, *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.

[27] R. J. Podesva, 'Phonation type as a stylistic variable: The use of falsetto in constructing a persona', *J Sociolinguistics*, vol. 11, no. 4, pp. 478–504, 2007.

THE SOCIOLINGUISTICS OF /R/-VOCALIZATION IN SPOKEN SWISS STANDARD GERMAN

Selma Vonlanthen[1], Adrian Leemann[1,2], Carina Steiner[1,2], Péter Jeszenszky[3]

[1]University of Bern, Switzerland; [2]University of Zurich, Switzerland; [3]Independent researcher
selma.vonlanthen@unibe.ch, adrian.leemann@unibe.ch, carina.steiner@unibe.ch, pjeszenszky@gmail.com

ABSTRACT

In German-speaking Switzerland, two varieties are used in parallel: dialect (informal, everyday use) and Swiss Standard German (formal, official documents). Depending on speakers' attitudes, they may articulate certain sounds as more "Swiss-like" or more "German-like" when talking in Swiss Standard German. The present study examines this by focusing on /r/-vocalization, e.g., [d̥ɐ̯] as opposed to [d̥er] in the German masculine article *der*. We analyzed /r/-vocalization of nearly 500 speakers across 125 localities in two contexts (word-final and preconsonantal /r/, both following a vowel). Overall, very few speakers (23/500) vocalized in both contexts. The results further revealed effects of region, age, attitudes towards Standard German, and an influence on how /r/ is realized in dialect. We discuss these findings against the backdrop of diglossia and the status attributed to the Swiss and German Standard varieties.

Keywords: Swiss Standard German, diglossia, language attitudes, accommodation

1. INTRODUCTION

German is a prototypical example of a pluricentric language with multiple codified standards, including Swiss Standard German (CHStG) and German Standard German (DStG) [1]. If we zoom into German-speaking Switzerland, we find a special diglossic situation with untypically high prestige attributed to dialects [2-5]. Against this backdrop, discussions about the use of dialects or standard, and if standard, which one, are often emotionally loaded. Some Swiss speakers feel particularly proud of their dialects and distance themselves from sounding like a German altogether [6]. Turning to the standard language, studies on the perception of both standard varieties have shown that many Swiss people do not regard CHStG as "correct German" compared to DStG [7] and some Swiss even feel inferior when speaking the standard language [6]. This situation, then, has people choose between phonetic forms that sound more dialectal or more German (meaning DStG) when speaking the standard language.

DStG and CHStG show a variety of differences in the realization of phonemes, one of which is that of /r/. In DStG, /r/ is either realized as a consonant or vocalized depending on the phonological context, while it is most commonly realized as a consonant in CHStG [8]. Hove [9] describes this typically missing vocalization of /r/ as one of the most distinct features of CHStG. If, however, a Swiss person vocalizes their /r/s in CHStG, these vocalizations are constrained by a number of linguistic and sociolinguistic factors. Hove [9], for example, found that /r/ in a word-final position is vocalized more often (in 20% of cases) than in a preconsonantal position (15% of cases). As for sociolinguistic constraints, women have been found to vocalize /r/ slightly more often than men. Christen et al. [10], examining the CHStG pronunciation in a corpus of police phone calls, noticed accommodation effects: Swiss speakers vocalized more when talking to German or Austrian speakers. Siebenhaar [11] further showed regional effects: St. Gallen – a Northeastern canton close to the German and Austrian border – featured more speakers vocalizing than for example those in Zurich and Bern, further away from the Northern border.

We identified a number of research gaps in the study of /r/-vocalization in CHStG: there are no studies contextualized with language attitudes (e.g., testing if a positive association with Standard German leads to more vocalization); further, there are also no studies that have investigated the link between age and the place of articulation (alveolar or uvular) in dialect with a potential effect on vocalization in CHStG. In the present study, we addressed these research gaps by analyzing the pronunciation of /r/ in nearly 500 Swiss German speakers talking in CHStG with a specific focus on the following potential sociodemographic and attitudinal effects: age, region, gender, education, and language attitudes. We hypothesized that younger speakers, women, and those from regions closer to Germany or Austria would vocalize more frequently. Furthermore, we expected speakers with more favorable attitudes towards Standard German as well as speakers with higher education to vocalize more frequently.

2. METHODS

2.1. Speakers

For this study, we analyzed a total of 500 speakers from the SDATS database (Swiss German Dialects Across Time and Space) [12] – an extensive database of Swiss German, representing 125 localities across German-speaking Switzerland. The sample of the current study consisted of two women and two men per locality, evenly balanced between two age cohorts: two 20–35-year-olds and two 60+ year-olds.



Figure 1: The 125 localities included in the SDATS study [12].

2.2. Material

For the current study, data from a read text in Standard German were used. The text started with the following phrase: *Die direkte Demokratie in der Schweiz fordert viel Kompromissbereitschaft*. (Eng.: 'Direct democracy in Switzerland demands a great deal of willingness to compromise'). /r/ was examined in word-final position (as in *der* 'the') and preconsonantal position (as in *fordert* 'demands'). To study whether the dialectal realization of /r/ has any effect on the CHStG vocalization of /r/, we further integrated dialectal realization as alveolar and uvular in a sentence completion task where speakers pronounced *Bernerin*, 'Bernese woman', in their local dialect.

2.3. Procedures

Data collection took place during the COVID-19 pandemic in 2020–2021 [13]. This read speech task was embedded in an oral interview principally geared to eliciting dialectal speech (~300 items, i.e, words and sentences, were elicited in participants' dialectal speech). We asked participants to spontaneously read the text in Standard German which was displayed on their smartphones, see Figure 2.

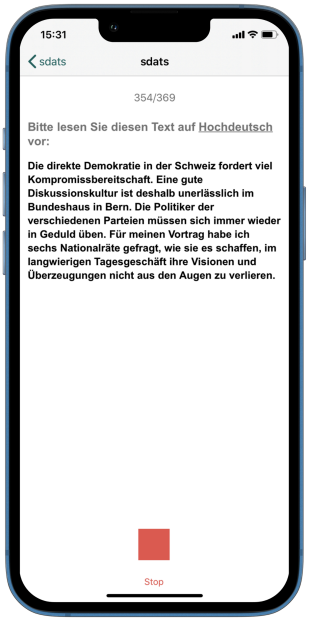


Figure 2: Elicitation prompt for the short reading passage in Standard German.

Metadata on sociodemographic variables, language use, and attitudes were obtained via an online questionnaire after the interview. Both variables – word-final and preconsonantal /r/ – were coded auditorily for the 500 speakers. Three speakers were excluded due to poor recording quality, resulting in a total of 497 analyzed cases. Vocalization was coded as 1, while consonantal realization was coded as 0. In case of doubt, the coding was checked by a second coder (N=42, 8.5%). Only a-schwa sounds ([ɐ]) were regarded as vocalized [8]. Some dialects of Swiss German may realize /r/ in a vocalic-like way, such as [ˈfɔɻdət] for *fordert* ('demands'). For this realization, we introduced a third coding category "2". An analysis in PRAAT [14] with randomly selected samples showed that these dialectal transfers were not specifically of [ɐ]-vocalized sound quality and were also difficult to attribute to a clear consonantal quality. Given this ambiguity, we decided to exclude these tokens from further analysis, which occurred 57 times in word-final and 55 times in preconsonantal position (further studies will need to be conducted on just this phenomenon).

Logistic regressions were run in JMP [15] and R [16] using region, age, dialectal /r/, and attitudes to Standard German as predictor variables. The latter variable was operationalized as a factor resulting from three statements that the participants were asked to agree or disagree with (using a seven-point Likert-scale, where 1 = strongly disagree and 7 = strongly agree; e.g., "I like speaking Standard German" etc.). Since variation in the degree of /r/-

vocalization in both contexts was small and different factors appeared to be at play for each of the two variables, the analysis was split into two separate models – one for each context.

3. RESULTS

3.1 Quantitative results

Table 1 shows the frequency of vocalization in both contexts.

Context	word-final /r/	preconsonantal /r/
Frequency	7% (N=34)	5% (N=23)

Table 1: Frequency of vocalization in the two models.

Overall, /r/ was rarely vocalized: by 34/497 speakers in word-final context and by 23 speakers in preconsonantal context. As for the word-final context, despite the small proportion of vocalization, the logistic regressions revealed some interesting effects: first, the regional origin of the participants affected vocalization of word-final /r/ (X^2(7, N=440)=25, p=.0007)): speakers from Northwest and Northeast Switzerland vocalized word-final /r/ significantly more frequently than those from other regions. Second, language attitudes played a role (X^2(1, N=440)=11, p=.0009)): the more positive the language attitudes, the more vocalizations occurred (Figure 2). Third, this effect of language attitudes was only true for the younger cohort (age*language attitudes, X^2(1, N=440)=8 p=.0056)). Figure 3 shows the effect of attitudes to Standard German on vocalization in word-final position by age cohort.

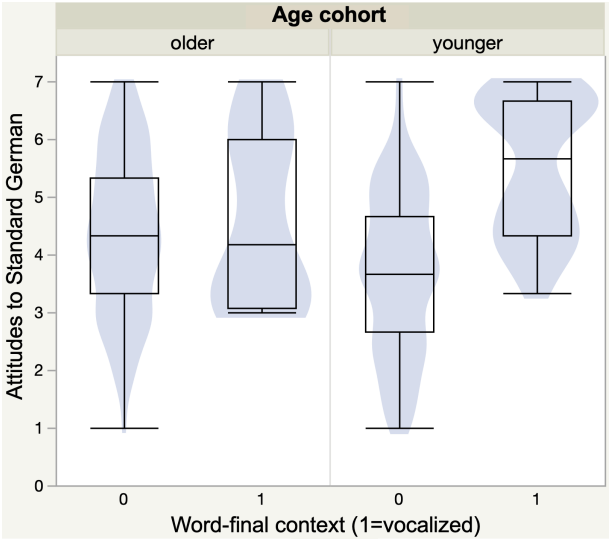


Figure 3: Attitudes towards Standard German (from 1 rather negative to 7 very positive) (y-axis) versus (non)-vocalization in word-final context by age cohort (y-axis).

Preconsonantal /r/ was vocalized 5% of the time. The same regional effect as for word-final /r/ was observed (X^2(7, N=442)19, p=.0091)): Northwestern and Northeastern dialects vocalized more frequently. Moreover, the age cohorts showed statistically significant differences (X^2(1, N=442)=34, p<.0001)): none of the older 250 speakers vocalized /r/ in that position; those who did vocalize, all came from the younger cohort. Finally, the realization of dialectal /r/ factors into whether speakers vocalize preconsonantal /r/ (X^2(1, N=440)=12, p=.0004)): speakers who used uvular /r/ in the prompt *Bernerin* vocalized in the preconsonantal position 31% of the time; those using alveolar /r/ in *Bernerin* vocalized only slightly more than 3% of the time. There is incongruity in the realization of /r/ in these two contexts: vocalization in preconsonantal position corresponded to vocalization in word-final position but not vice versa (i.e., those who vocalized in preconsonantal position also vocalized in word-final position). In what follows, we focus in a qualitative fashion on these 23 participants who vocalized in both contexts.

3.2. Qualitative analysis

The 23 speakers who vocalized /r/ have the following characteristics: 14 of them are male; 14 do not have a tertiary degree but are currently enrolled in a degree program at an institute of higher education. Interestingly, those 23 who vocalized in both instances tend to have jobs that have a communicative function, e.g., journalist, attorney, flight attendant, customer service or musical performer, or they hold a management position. Regarding language attitudes, they had much more positive attitudes towards Standard German (M=5.6/7) compared to those who did not vocalize (N=474) in either context (M=4/7). Further, there is a clear geographic pattern in the origins of these 23 speakers: almost all of them were socialized in the (Northern) Midlands (Figure 4).

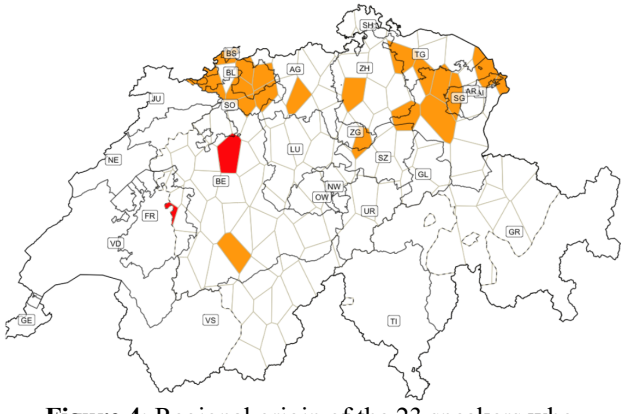


Figure 4: Regional origin of the 23 speakers who vocalized /r/ in both contexts. Orange polygons refer to one such speaker, red polygons refer to two speakers.

4. DISCUSSION

We will first discuss the general trends, followed by the effects of region, attitudes, and age, before moving on to interpret the qualitative analysis. The overall picture is clear: vocalizing /r/ in CHStG is still very rare in German-speaking Switzerland. Our findings further confirm that word-final /r/ is more often vocalized than preconsonantal /r/. We can speculate as to potential reasons for this effect: perhaps vocalizing /r/ in word-final position may be easier from an articulatory standpoint than it is to do before a consonant [17]. Another reason why word-final /r/ was vocalized more may have to do with the order in which the words with the contexts studied appeared in the text: the word-final variable (*der*) appeared before the preconsonantal one (*fordert*); we noticed how some speakers started by vocalizing the first item but switched back to the more Swiss consonantal realization by the second item. Perhaps, while reading, some speakers realized they had vocalized at first, noticed that this may sound inadequate in an interview principally geared towards elicited dialect items, and – consequently – switched back to the more expected non-vocalized realization.

As for regional and attitudinal effects, we argue as follows: the region effect for both variables can be explained to a certain degree by the geographical proximity to Germany. Speakers in Northeastern and Northwestern Switzerland experience more linguistic and personal exchange with DStG than speakers from other regions [18]. Regarding attitudes, positive attitudes towards StG seem to encourage the vocalization of /r/ among younger speakers. This finding corroborates a growing body of research which shows the importance of language attitudes and their effect on the way someone speaks [6, 19].

Regarding age, the results confirmed our expectations that younger speakers may vocalize more frequently than older speakers. In preconsonantal position, /r/ was vocalized exclusively by young speakers and, moreover, all these speakers also vocalized in word-final position. This generation grew up using social media [20], which might exert an influence on their articulation (watching – and maybe also creating – videos in DStG on TikTok, Facebook, Instagram, YouTube etc.). Muhr [21], for example, showed a similar influence of DStG on Austrian German pronunciation through television in the early 2000s. A future study exploring the influence of social media on the CHStG pronunciation should therefore be conducted.

Moving on to the qualitative findings of our study regarding detailed analysis of the 23 speakers who vocalized /r/ in both contexts, we can offer the following explanations. About two-thirds of them are enrolled in a higher education degree. In the Swiss context, you would expect this cohort to vocalize the most [22], given, for example, their exposure to DStG in a university context (many of the lecturers are of German origin). We further found that these participants are predominantly in communication-oriented jobs, such as working as a journalist or musical performer. A musical performer, for instance, typically undergoes training in DStG for musical productions. This, in turn, may have an effect on their articulation of CHStG. Regarding attitudes towards DStG, we found that these 23 speakers, on average, had a mean of 5.6 on a seven-point Likert scale indexing attitudes towards the standard language – compared to an average of 4 for the other 474 participants who did not vocalize both items. This replicates the findings of the quantitative analyses on a micro scale. Further, the regional effects found also reproduce the empirical analyses for word-final and preconsonantal /r/, in that most of these 23 speakers come from the Northern Midlands, close to the German border (where uvular /r/s are predominant – hence the effect reported in 3.1).

While the findings reported are highly plausible, it needs to be borne in mind that the effects reported are based on a small number of total occurrences of vocalization. In the future, we hope to expand the analyses to the full 1000-speaker dataset and, potentially, collect additional data for enhanced statistical explanatory power. Further, this study should be repeated with other, typical CHStG phonetic variables which we assume have sociolinguistic constraints: for example, affrication of DStG [kʰ] as in *Politiker* ('politician'), backed articulation of DStG [ç] as in *ich* ('I'), or lowered articulation of DStG [ɛ], as in *Nationalräte* ('members of parliament'). An analysis of additional variables will give us a more complete picture of the sociolinguistic factors affecting pronunciation in CHStG.

5. ACKNOWLEDGMENTS

This research was funded by the Swiss National Science Foundation (Grant no. PCEFP1_181090).

5. REFERENCES

[1] Ammon, U., Bickel, H., Lenz, A. N. (eds) 2004. *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. de Gruyter.

[2] Ferguson, C. 1959. Diglossia. *Word* 15, 325–340.

[3] Haas, W. 2004. Die Sprachsituation der deutschen Schweiz und das Konzept der Diglossie. In: Christen, H. (ed.), *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum. Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen*. Edition Praesens, 81–110.

[4] Berthele, R. 2004. Vor lauter Linguisten die Sprache nicht mehr sehen – Diglossie und Ideologie in der deutschsprachigen Schweiz. In Christen, H. (ed.), *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum. Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen*. Edition Praesens, 111–136.

[5] Petkova, M. 2012. Die Deutschschweizer Diglossie: eine Kategorie mit fuzzy boundaries. *Zeitschrift für Literaturwissenschaft und Linguistik* 42(4), 126–154.

[6] Bickel, H., Hofer, L. 2013. Gutes und angemessenes Standarddeutsch in der Schweiz. In: Schneider-Wiejowski, K., Kellermeier-Rehbein, B., Haselhuber, J. (eds), *Vielfalt, Variation und Stellung der deutschen Sprache*. De Gruyter, 79–100.

[7] Scharloth, J. 2006. Schweizer Hochdeutsch – schlechtes Hochdeutsch? In: Dürscheid, C., Businger, M. (eds.), *Schweizer Standarddeutsch: Beiträge zur Varietätenlinguistik*. Narr Francke Attempto, 81–96.

[8] Krech, E-M., Stock, E., Hirschfeld, U., Anders, L. C. (eds) 2009. *Deutsches Aussprachewörterbuch*. De Gruyter.

[9] Hove, I. 2002, *Die Aussprache der Standardsprache in der deutschen Schweiz (= Phonai Bd. 47)*. Niemeyer.

[10] Christen, H., Guntern, M., Hove, I., Petkova, M. 2010. *Hochdeutsch in aller Munde: Eine empirische Untersuchung zur gesprochenen Standardsprache in der Deutschschweiz*. Steiner.

[11] Siebenhaar, B. 1994. Regionale Varianten des Schweizerhochdeutschen: Zur Aussprache des Schweizerhochdeutschen in Bern, Zürich und St. Gallen. *Zeitschrift für Dialektologie und Linguistik* 61(1), 31–65.

[12] Leemann, A., Jeszenszky, P., Steiner, C., Messerli, J., Studerus, M. 2020. SDATS Corpus – Swiss German dialects across time and space. [Online]. Available: osf.io/s9z4q.

[13] Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard* (6.3).

[14] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.02, retrieved 29 November 2022 from http://www.praat.org/.

[15] JMP®, Version <15>. SAS Institute Inc., Cary, NC, 1989–2021.

[16] R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[17] Martinet, A. 1981. *Sprachökonomie und Lautwandel. Eine Abwandlung über die diachronische Phonologie*. Klett-Cotta.

[18] Bundesamt für Statistik BFS. 2021. Grenzgängerinnen und Grenzgänger in der Schweiz 1996–2020. https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/publikationen.assetdetail.17205597.html

[19] Beaman, K. V., Tomaschek, F. 2021. Loss of historical phonetic contrast across the lifespan: Articulatory, lexical, and social effects on sound change in Swabian 1. In *Language Variation and Language Change Across the Lifespan*. Routledge, 209–234.

[20] Morning Consult. 2022. Gen Z is extremely online. https://morningconsult.com/2022/12/12/gen-z-social-media-usage/

[21] Muhr, R. 2003. Language change via satellite: The influence of German television broadcasting on Austrian German. *J. Hist. Pragmat.* 4(1), 103–127.

[22] Guntern, M. 2012. Dialekt und gesprochene Standardsprache: Wie Laien gesprochenes Schweizerhochdeutsch beurteilen. *Sociolinguistica* 26.

ACOUSTIC AND ARTICULATORY CHARACTERISTICS OF RHOTICITY IN THE NORTH-WEST OF ENGLAND

Claire Nance, Maya Dewhurst, Lois Fairclough, Pamela Forster, Sam Kirkham, Takayuki Nagamine, Danielle Turton, Di Wang

Lancaster University
c.nance@lancaster.ac.uk

ABSTRACT

We present a sociophonetic, acoustic, and articulatory analysis of coda rhoticity in East Lancashire, North-West England. We analysed data from 24 participants aged 8–73 recorded at a public engagement event in Blackburn Market (598 tokens). Auditory analysis shows coda rhoticity is declining across generations, with speakers born after 1990 being mostly non-rhotic. Audible rhoticity is realised by lowered F3 and raised F2. GAMMs fitted across the vowel(+rhoticity) interval show that audibly rhotic tokens have a significantly smaller distance between F3 and F2 than audibly non-rhotic tokens in all vowel contexts. Our ultrasound analysis compares minimal pairs e.g. 'core' and 'caw'. Principal Component Analysis of tongue splines shows that speakers use different tongue shapes in auditorily rhotic tokens.

Keywords: rhoticity, ultrasound, acoustics, sociophonetics, language variation and change

1. INTRODUCTION

In this paper, we consider the distribution and realisation of residual coda rhoticity in East Lancashire, North-West England, for example in words like 'farm' and 'core'. Coda rhoticity is one of the classic and most widely studied sociophonetic variables across the English-speaking world [1], see [2] for a recent overview. In England, rhoticity appears to have been in decline since early modern times [3]. Dialect survey work and analysis of oral history data from the 20th century indicates pockets of coda rhoticity remaining in Central/East Lancashire, (as well as rural West Yorkshire, and south-West England) [4, 5], though 21st century surveys and sociophonetic analysis suggests that East Lancashire rhoticity is now in decline [6, 7].

Acoustic approaches to English rhoticity have largely focussed on lowered F3 as a cue to rhoticity e.g. [5]. However, other work suggests that formants below F3 also contribute to perceptions of rhoticity [8, 9]. In East Lancashire in particular, previous work indicates that audible rhoticity is realised with lowered F3 and also raised F2 [10, 11]. We therefore chose F3−F2 as the acoustic measure best representing rhoticity. We present the raw values of F3−F2 as comparing the distance between formants should normalise to some extent for differing vocal tract lengths, and further normalisation may over-normalise the data [11].

Previous articulatory work on rhotics in English suggests that speakers broadly use a retroflex or bunched tongue gesture to realise rhoticity [12, 13]. Tongue shape is usually investigated in the midsagittal dimension. Due to the importance of the midsagittal perspective, ultrasound tongue imaging has been widely used to study rhotic articulation cross-linguistically e.g. [14, 15]. In English in the UK, ultrasound has especially been used to study variable rhoticity in Scottish English e.g. [16, 17]. Ultrasound analysis in Scotland has shown considerable sociolinguistic variation in rhoticity realisation [16]. Our analysis represents the first articulatory study of rhoticity in England.

In this paper we investigate the following research questions:

1. How is rhoticity distributed according to different demographic factors?
2. How is audible rhoticity realised in acoustics?
3. How is audible rhoticity realised in articulation?

2. METHODS

2.1. Participants and setup

Our data collection was carried out as part of a public engagement event in the Blackburn Market, East Lancashire, inspired by [18]. We recorded a total of 35 participants. In this paper, we present data from 24 participants (14f, 10m) who were long-term residents of East Lancashire, mostly Blackburn. They were aged 8–73 ($\bar{x} = 39$; $\sigma = 17.5$). Blackburn has a large British Asian population and we recorded one participant who identified as British Pakistani, and two who were born in India but now work in Blackburn. Our speakers represented a range of occupations, as well as children and non-working

people. Due to this variety, we have not analysed social class further but note that this will be an important dimension in future work.

2.2. Materials and recording

Participants were asked to read a list of words which were presented in the Articulate Assistant Advanced (AAA) software [19]. Words were displayed orthographically and with pictures. Participants were asked to read the words twice, but we only collected one repetition from some speakers due to time constraints. Our word list included nine words potentially containing coda rhoticity, and four distractors/minimal pair counterparts (Table 1). Note that in Lancashire the NURSE and SQUARE lexical sets are often merged [20].

Word	Lexical set	Word	Lexical set
beard	NEAR	goat	GOAT
caw	THOUGHT	paw	THOUGHT
core	NORTH/FORCE	pour	NORTH/FORCE
cake	FACE	stair	SQUARE/NURSE
farm	START	stir	NURSE
fair	SQUARE/NURSE	worm	NURSE
fur	NURSE		

Table 1: Word list included in the study.

We used the Telemed MicrUs system to record our ultrasound data into AAA, with a 64-element probe, 20mm radius. The probe frequency was 2MHz, depth was 80mm, and field of view was 90–100% resulting in a frame rate of 80–90Hz. The probe was held in place using a plastic helmet [21]. The acoustic data were recorded with a headset microphone attached to the helmet. Acoustic data were recorded onto a laptop in AAA with a Sound Devices USBPre2 audio interface at 22,050Hz.

2.3. Analysis

Data were exported from AAA into Praat for auditory and acoustic analysis. Each token was first coded auditorily for the presence/absence of coda rhoticity. The interval containing the vowel plus any following rhoticity was then labelled, and extracted for analysis. The first three formants were estimated in Fasttrack [22]. The optimal analyses from Fasttrack were then analysed further in R. Acoustic data from one participant were excluded due to very poor audio quality. Ultrasound data were excluded for two participants due to poor image quality. TextGrids were then imported into AAA and the labels serve as landmarks for the articulatory analysis. Tongue splines were generated using DeepLabCut [23] in AAA at 11 key points on the tongue, and then rotated to the occlusal plane for each participant using bite plate traces [24] (except for one participant). Further analysis was done in R.

The analyses in this paper are structured around our research questions. Code and data for these analyses are available at https://osf.io/wb9m5/.

Analysis 1: Auditory coding results from words containing orthographic 'r' are presented according to speaker and vowel context. We then consider variation in the data according to age, gender and ethnicity. These results are analysed with a logistic mixed effects regression model: rhoticity $\sim$ year of birth + gender + (1|word) + (1|speaker). Ethnicity and vowel context are explored qualitatively due to small token counts. 598 tokens from 24 speakers.

Analysis 2: Acoustic data from words containing orthographic 'r' were divided by their auditory coding of 'rhotic' or 'non-rhotic'. We then used GAMMs [25] to compare the F3−F2 values across 11 timepoints in the vowel(+rhoticity) interval in different vowel contexts. GAMMs were fitted following [26]. Predictor variables included a parametric term of auditory coding for rhoticity and smooth terms of normalised time and a normalised time-by-rhoticity interaction. We also fitted random smooths of time-by-speaker and time-by-word. For significance testing, we compared a full autoregressive model to nested models excluding all rhoticity predictors to investigate for significant differences in trajectory height, and then the full model to a model excluding rhoticity-by-time to investigate for significant differences in trajectory shape [27]. 394 tokens from 23 speakers.

Analysis 3: Ultrasound analysis focusses on two rhoticity minimal pairs: *core/caw* and *pour/paw* at 80% duration of the vowel(+rhoticity) interval. We compare the tongue shapes used by speakers visually, and then via Principal Component Analysis for statistical comparison of auditorily rhotic vs. non-rhotic tongue shapes [28, 15]. Values of PC1 accounted for 95% of the variation in the dataset. We therefore tested z-scored PC1 values comparing words containing 'r' (*core, pour*) against words not containing 'r' (*caw, paw*) for auditorily rhotic and non-rhotic speakers. This was carried out via linear mixed effects model of the formula PC1z $\sim$ r-ful word*auditory perception + (1|speaker) + (1|word). 120 tokens from 16 speakers.

3. RESULTS

3.1. Distribution of rhoticity according to demographic factors

Auditory perceptions of rhoticity are shown in Figure 1 for individual speakers and words.

Statistical testing considered age and gender of participants. Due to the correlation between vowel

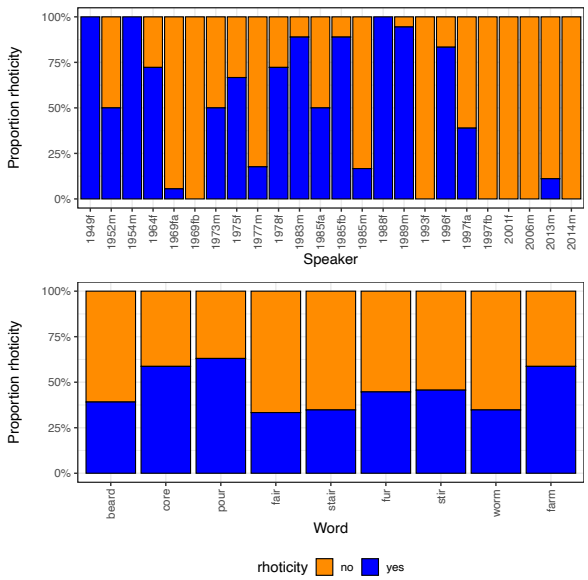


Figure 1: Top panel: Auditory coding of rhoticity for individual speakers. Speakers are ordered by year of birth oldest to youngest left to right. Speaker number shows year of birth and gender. Bottom panel: Auditory coding of rhoticity for individual words.

and word, we have modelled the vowel effect as a random intercept of word only. Visual inspection of Figure 1 indicates that rhoticity is most present in words from the NORTH/FORCE and START lexical sets, though there are not large differences between vowels. In terms of ethnicity, our British Pakistani speaker was completely non-rhotic. The two Indian speakers were almost 100% rhotic and used taps or trills, while the White British speakers used pharyngealisation or approximant rhotics. Model comparison shows a significant effect of age, where older speakers are more likely to be rhotic ($\chi^2(1) = 8.19$; $p(\chi^2) = .004$). Gender was not significant.

3.2. Acoustics of audible rhoticity

Here, we compare the F3−F2 values of auditorily rhotic and non-rhotic productions in the vowel(+rhoticity) interval via GAMMs. Results are shown in Figure 2, and the results of model comparisons testing for significant differences in trajectory height and shape are in Table 2. For significance testing via model comparison, we used an AR1 model with rho estimated as the autocorrelation at lag 1. See code for information (https://osf.io/wb9m5/). Results show significant differences for height in every vowel context, and significant differences for shape in NORTH/FORCE, SQUARE, and START words. In each case, F3−F2 is lower in auditorily rhotic tokens.

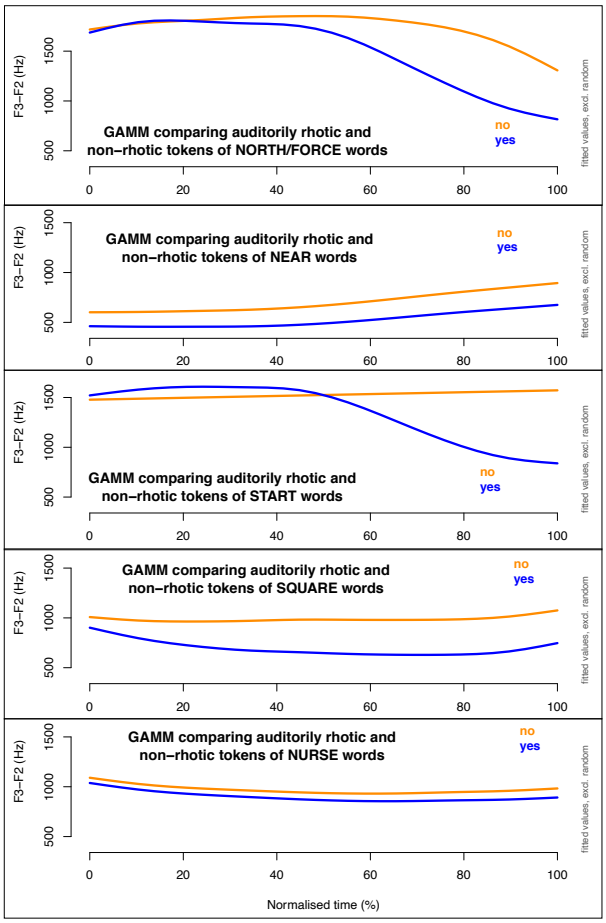


Figure 2: GAMMs comparing auditorily rhotic (blue) and non-rhotic (orange) tokens. Top panel NORTH/FORCE words; second panel NEAR words; third panel START words; fourth panel SQUARE words; fifth panel NURSE words.

Lexical set	Trajectory height	Trajectory shape
NORTH/FORCE	$\chi^2(3) = 26.7; p < .001$	$\chi^2(2) = 24.1; p < .001$
NEAR	$\chi^2(3) = 14.8; p < .001$	$\chi^2(2) = 0.32; p = .72$
START	$\chi^2(3) = 16.5; p < .001$	$\chi^2(2) = 16.5; p < .001$
SQUARE	$\chi^2(3) = 40.3; p < .001$	$\chi^2(2) = 8.50; p < .001$
NURSE	$\chi^2(3) = 7.52; p = .002$	$\chi^2(2) = 0.33; p = .71$

Table 2: Acoustic GAMM model comparison showing differences between auditorily rhotic and non-rhotic tokens.

3.3. Articulation of audible rhoticity

Here, we compare the tongue shapes used for *core/caw* and *pour/paw*. We have plotted the splines for the sixteen speakers analysed so far in Figure 3. The DeepLabCut output was smoothed using a 10 Hz first-order Butterworth low-pass filter. In the

eight speakers who contrast for rhoticity, *core/pour* are produced with a raised and fronted tongue compared to the back vowel in *caw/paw*.

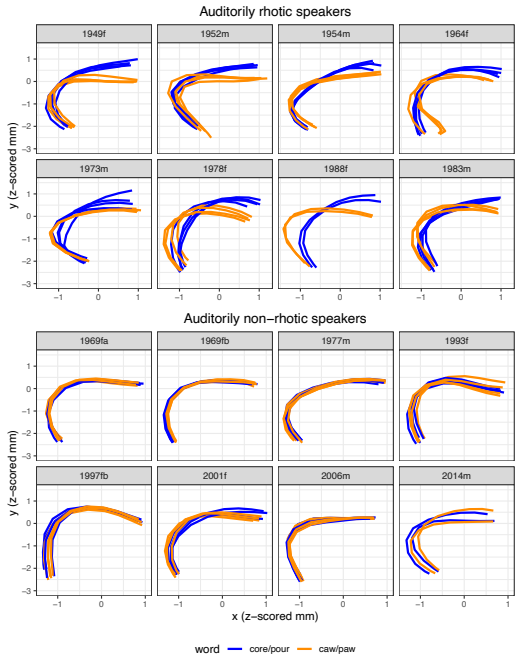


Figure 3: Tongues splines comparing rhoticity minimal pairs in auditorily rhotic and non-rhotic speakers. Code for this: https://osf.io/wb9m5/

The differences in tongue shape in Figure 3 were compared with a PCA run on the spline values. We report values of PC1 only, as this accounted for 95% of the variation in the dataset. (PC2 accounted for 2% and higher PCs even lower values). The variation in PC1 is plotted in Figure 4, along with a comparison of PC1 values in auditorily rhotic and non-rhotic speakers. PC1 appears to capture variation in tongue fronting and height.

The values of PC1 were modelled to test for an interaction between perceived rhoticity and minimal pair. Model comparison indicates that this interaction is significant ($\chi^2(2) = 23.6; p < .001$), meaning that for speakers who produce audible rhoticity, tongue shapes are significantly different for *core/pour* vs. *caw/paw*.

4. DISCUSSION AND CONCLUSIONS

Production of coda rhoticity appears to vary across our sample. However, we found a general pattern towards decline in rhoticity according to age, supporting previous work in [7, 6, 11]. In our data, young people born after 1990 are mostly non-rhotic. We were not able to fully investigate variation across vowel contexts due to the small number of words in

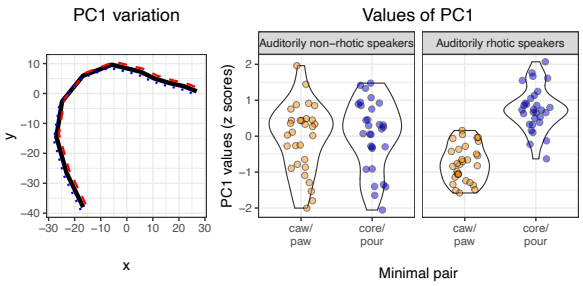


Figure 4: Left panel: variation explained by PC1. Black line shows tongue mean, red dashed line shows maximum values, blue dotted line shows minimum values.
Right panel: Values of PC1 according to perceived rhoticity and whether the word contains 'r'.

our dataset, but it is interesting to note that NURSE words were not necessarily the most rhotic as has been suggested in other work on rhoticity in England [2]. The three British Asian speakers in our dataset show some interesting patterns based on the extent to which they have acquired a South Asian English variety (rhotic), or grew up in Lancashire (one young person, non-rhotic). This variation would be interesting to investigate in the future in more depth. Our acoustic results indicate that F3−F2 is a good measure of audible rhoticity in East Lancashire, especially for non-central vowels. It was less good for the NURSE vowel, though still showed significant differences in trajectory height. Our preliminary analysis of the ultrasound data shows differences in the tongue shapes used in minimal pairs where speakers produce audible rhoticity.

To conclude: rhoticity is still relatively widespread in East Lancashire, even though it is now very rare in the north of England generally [7]. This appears to be changing in apparent-time though future larger-scale work is now needed. We presented the first articulatory analysis of rhoticity in England and demonstrate that different tongue shapes are used for auditorily rhotic and non-rhotic tokens. Future work will look in more detail at the timing of rhotic gestures, as well as a larger set of lexical items. We are also interested in the possibility of differing voice qualities and/or articulatory settings across generations in East Lancashire (similar to the description of Glasgow in [29]) and would like to investigate this further with acoustic and articulatory methods.

5. ACKNOWLEDGEMENTS

Thank you to our participants, Blackburn Council, and Blackburn Market. Thank you to the ESRC Festival of Social Sciences for funding.

6. REFERENCES

[1] W. Labov, *The social stratification of English in New York City*. Washington D. C.: Center for applied linguistics, 1966.

[2] T. Blaxter, K. Beeching, R. Coates, J. Murphy, and E. Robinson, "Each person does it their way: Rhoticity variation and the community grammar," *Language Variation and Change*, vol. 31, no. 1, pp. 91–117, 2019.

[3] E. Gordon, L. Campbell, J. Hay, M. Maclagan, A. Sudbury, and P. Trudgill, *New Zealand English: its origins and evolution*. Cambridge: Cambridge University Press, 2004.

[4] H. Orton and E. Dieth, *Survey of English Dialects: The Six Northern Counties and the Isle of Man*, H. Orton and W. J. Halliday, Eds. Leeds: University of Leeds Press, 1962.

[5] S. Durkacz Ryan, H. Dann, and R. Drummond, "'Really this girl ought to be going to something better': Rhoticity and social meaning in oral history data," *Language in Society*, p. First view, 2022.

[6] W. Barras, "The sociophonology of rhoticity and r-sandhi in East Lancashire English," Ph.D. dissertation, University of Edinburgh, Edinburgh, 2010.

[7] A. Leemann, M.-J. Kolly, and D. Britain, "The English dialects app: The creation of a crowdsourced dialect corpus," *Ampersand*, vol. 5, pp. 1–17, 2018.

[8] B. Heselwood and L. Plug, "The role of f2 and f3 in the perception of rhoticity: Evidence from listening experiments," in *Proceedings of the 17th International Congress of the Phonetic Sciences*. Hong Kong: City University Hong Kong, 2011.

[9] D. Villarreal, L. Clark, J. Hay, and K. Watson, "From categories to gradience: Auto-coding sociophonetic variation with random forests," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, pp. 1–31, 2020.

[10] B. Heselwood, "Rhoticity without F3: Lowpass filtering, F1-F2 relations and the perception of rhoticity in NORTH-FORCE, START and NURSE words," *Leeds Working Papers in Linguistics and Phonetics*, vol. 14, pp. 49–64, 2009.

[11] D. Turton and R. Lennon, "An acoustic analysis of rhoticity in Lancashire, England," *Journal of Phonetics*, Under review.

[12] P. Delattre and D. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 6, no. 29-68, 1968.

[13] H. King and E. Ferragne, "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics*, vol. 80, pp. 1–19, 2020.

[14] M. Tabain and R. Beare, "An ultrasound study of coronal places of articulation in Central Arrernte: Apicals, laminals and rhotics," *Journal of Phonetics*, vol. 66, pp. 63–81, 2018.

[15] C. Nance and S. Kirkham, "Phonetic typology and articulatory constraints: The realisation of secondary articulations in Scottish Gaelic rhotics," *Language*, vol. 98, no. 3, pp. 419–460, 2022.

[16] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "The social stratification of tongue shape in postvocalic /r/ in Scottish English," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 256–268, 2011.

[17] J. Stuart-Smith, E. Lawson, and J. M. Scobbie, "Derhoticisation in Scottish English: a sociophonetic journey," in *Advances in Sociophonetics*, C. Celata and S. Calamai, Eds. Amsterdam: John Benjamins, 2014, pp. 59–96.

[18] M. Heyne, X. Wang, D. Derrick, K. Dorreen, and K. Watson, "The articulation of /r/ in New Zealand English," *Journal of the International Phonetic Association*, vol. 50, no. 3, pp. 366–388, 2020.

[19] Articulate Instruments, *Articulate Assistant Advanced version 2.20.2*. Edinburgh: Articulate Instruments, 2022.

[20] K. Watson and L. Clark, "How salient is the Nurse-Square merger?" *English Language and Linguistics*, vol. 17, no. 2, pp. 297–323, 2013.

[21] L. Spreafico, M. Pucher, and A. Matosova, "UltraFit: A speaker-friendly headset for ultrasound recordings in speech science," in *Interspeech 2018*, 2018.

[22] S. Barreda, "Fast track: fast (nearly) automatic formant-tracking using praat," *Linguistics Vanguard*, vol. 7, no. 1, 2021.

[23] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[24] J. M. Scobbie, E. Lawson, S. Cowen, J. Cleland, and A. Wrench, "A common co-ordinate system for mid-sagittal articulatory measurement," *QMU CASL Working Papers*, vol. 20, 2011.

[25] S. N. Wood, *Generalized Additive Models*. Chapman and Hall/CRC, 2017.

[26] M. Sóskuthy, *Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction*, 2017. [Online]. Available: https://arxiv.org/abs/1703.05339

[27] S. Kirkham, C. Nance, B. Littlewood, K. Lightfoot, and E. Groarke, "Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English," *Journal of the Acoustical Society of America*, vol. 145, pp. 784–794, 2019.

[28] D. Turton, "Categorical or gradient? an ultrasound investigation of /l/-darkening and vocalization in varieties of English," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 1, pp. 1–31, 2017.

[29] M. Sóskuthy and J. Stuart-Smith, "Voice quality and coda /r/ in Glasgow English in the early 20th century," *Language Variation and Change*, vol. 32, pp. 133–157, 2020.

EVALUATING THE ROLE OF SELF-DESCRIPTION IN DEMARCATING ACCENTS

Amanda Cole[1] and Patrycja Strycharczuk[2]

[1]University of Essex, [2]University of Manchester
amanda.cole@essex.ac.uk, patrycja.strycharczuk@manchester.ac.uk

ABSTRACT

This study tests if speakers' self-description of accent, using both pre-selected labels and free-classification, is predictive of their linguistic productions. Based on the production of diphthongs in a sample of 186 speakers from Southeast England, we identify three clusters of speakers, representing Standard Southern British English, Estuary English and Multicultural London English. We explore participants' self-description of accent guided by pre-selected accent labels, such as 'London' or 'Queen's English', and their unguided, self-description of accent in their own terms. The former corresponded to the clusters of vowel features in 62% of cases. Unguided descriptions revealed some patterns, but typically speakers favoured geographical terms such as 'British' or 'London'. We conclude that self-description is a relatively poor predictor of accent, especially in a complex sociolinguistic context.

Keywords: accent variation; Southern British English; self-description; linguistic identity; vowels

1. INTRODUCTION

A central tenet of sociophonetics is that the rates of production of phonetic features vary systematically by social factors (e.g., class, age, gender) and/or regional factors (where a person is from) for many speech communities. However, sociophonetics is not only interested in the parameters of linguistic variation in speech production, but also how this variation is perceived, categorised and evaluated by speakers.

We know that the perceptual linguistic categories held by non-linguists do not always align with objective patterns of linguistic variation, for instance, meaning that the regional dialect of some speaker groups can be identified by listeners more frequently than others [1]. This becomes all the more evident when listeners are asked to identify speakers' regional dialect in a free classification task rather than choosing between fixed choice categories [2]. However, we do not currently know how reliably speakers self-define their own accent. Reliability is interpreted here as speakers of the same accent - in terms of similarities in linguistic content – demonstrating a shared understandings in how they label and describe their own accent. How reliably speakers self-define their own accent has important implications for linguistic research which uses self-description as a tool for identifying authentic speakers of a variety

This study builds on our previous work in which we delineated three accents spoken in Southeast England based on patterns of linguistic co-variation using a cluster analysis [3]. This new study investigates how speakers' definition of their own accent maps onto their linguistic productions i.e., their cluster membership. As a baseline for comparison, we also test whether a combination of the speakers' social and regional information is predictive of cluster membership. Our research questions are:

1. How reliably do speakers of the same accent - based on similarities in linguistic production - use the same labels to describe their accent?
2. Are speakers of some accents more likely to converge in how they define their accent compared to speakers of other accents?
3. To what extent are demographic factors, comprising social and regional information, predictive of cluster membership?

1.1. THE ACCENTS OF SOUTHEAST ENGLAND

Many different accents have been documented in Southeast England, most notably: Received Pronunciation (RP), Standard Southern British English (SSBE), Estuary English (EE), Cockney and Multicultural London English (MLE). RP is a standard accent spoken by the higher classes, linked to the fee-paying school system [4]. Over recent decades, RP has experienced language change towards vernacular productions, provoking debates about whether it is still spoken in Britain [5, 6]. Some linguists have instead begun to use the term "SSBE" to describe the more contemporaneous, upper-middle class accent evolved from RP [7].

Similar to SSBE, EE is an accent often described imprecisely as falling somewhere between RP and Cockney and being spoken across Southeast England [8]. Cockney is a working-class dialect of East London which includes a shifted diphthong system: /ʊ/- and /ɪ/-diphthongs are rotated clockwise and anti-clockwise respectively. Cockney is no longer prevalent among young East Londoners, particularly those from ethnic minority backgrounds, who instead speak MLE. This new variety includes innovative features, particularly in the diphthong system, influenced by other dialects of English and other languages [9].

In sum, there are not clear, widely shared or empirically determined boundaries between the accents spoken in Southeast England in terms of linguistic content or group membership, and there is little consensus on whether these accents are contemporaneous. Our previous research has addressed this issue by analysing the production of diphthong vowels in a sample of 193 speakers from Southeast England, using a bottom-up exploratory analysis, a combination of functional Principal Component Analysis and Gaussian Mixture Modelling. This unsupervised approach identified three clusters of speakers. Diphthongs were chosen as they are the most notable locus of variation in the Southeast. Mean formant values for diphthong vowels in each cluster are illustrated in Figure 1. These vowel sub-systems align closely with previous descriptions of variants attested in the southeast of England, specifically Multicultural London English (MLE, cluster 1), Standard Southern British English (SSBE, cluster 2) and Estuary English (EE, cluster 3).

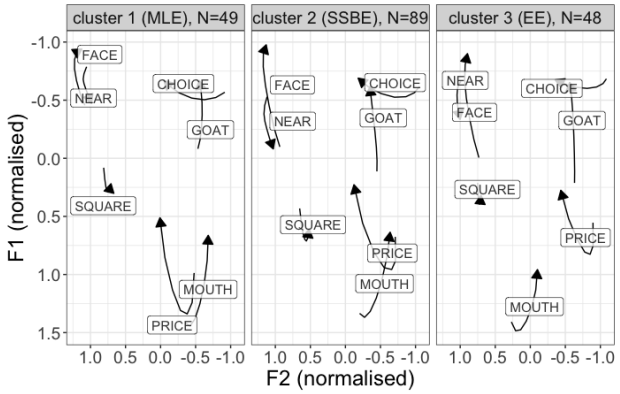


Figure 1: The three clusters based on diphthong productions for 193 speakers from Southeast England from [3].

The clusters emerged from the data through a bottom-up analysis, corresponding to previous descriptions of accent features in the Southeast. This confirms that there is structured phonetic variation on the data. The current study investigates whether this structured variation corresponds to how speakers describe their own accent.

2. METHOD

We analysed responses from 186 speakers. Seven speakers were excluded from the original sample of 193 because they did not answer all the questions. The speakers were all aged 18-33 years and they had lived in Southeast England for at least half of the years between the ages of three and 18. Out of the 186, based on the previous cluster analysis, 49 were classed as speakers of MLE, 89 were classed as speakers of SSBE, and 48 were speakers of EE.

We tested to what extent speakers' demographic information and self-description of accent predicted cluster membership (MLE, SSBE or EE) in a series of random forest models, fitted as described below.

In order to obtain a baseline level of accuracy, we built a model predicting cluster membership using only phonetic predictors, which were principal component scores capturing the main aspects of dynamic variation in the F1 and F2 of the following vowels: FACE, PRICE, CHOICE, MOUTH, GOAT, NEAR and SQUARE. This was done in order to quantify the separability of the three clusters from the phonetic point of view, given that some aspects of vowel variation were continuous. The model was trained on 75% of the data and tested on the remaining 25%. We repeated this procedure on 1000 bootstrapped samples with a 75% training set – 25% test set split. At each iteration, we extracted the overall accuracy, defined as the percentage of correct classification relative to the size of the test set. We also extracted the two most important predictor variables in each model.

We applied the same procedure to explore the role of demographic predictors (regional and social ones) in predicting cluster membership. Understanding the structure of social and regional variation within and between our clusters broadens our understanding of why speakers within each cluster may use certain terms to define their accent. The demographic predictors that we included were ethnicity, gender, age, social class, whether or not the participant attended a boarding school, and the participant's geographic county. A choice of response categories was not given, and participants provided the demographic information in their own terms. We collapsed the responses to the following response categories: ethnicity: Asian British, Black British or White British; gender: female or male; social class: middle-class or working class; whether or not the participant attended a boarding school: yes or no; speaker's county: London vs. other locations.

Further, we tested how well speakers' self-description of accent predicted cluster membership. Speakers indicated on numerical, sliding scales (from 0-100) whether they felt their accent could be described as 'typical of where they live', 'Southeast', 'Queen's English', 'Estuary', 'Cockney' or 'London'. These labels were judged to be potentially familiar to lay participants, as opposed to terms such as SSBE, or MLE, which are linguistic terms, but which do not often feature in public discourse. We used the speakers' responses to the pre-selected accent labels to predict cluster membership (MLE, SSBE or EE) in a series of random forest models, following the same approach as previously used with phonetic and demographic predictors. In addition, speakers described their accent in their own words in a free-classification task in response to the question "How would you define your accent?". We extracted the most frequent words from the speaker's free-classification, which we then analysed in relation to cluster membership to explore which labels were used most for each cluster.

3. RESULTS

3.1. Phonetic predictors

The mean accuracy of models predicting cluster membership using only phonetic predictors was 86%. Our interpretation is that the phonetically informed models performed well, but not perfectly because there is some overlap between different accents. In essence, we find a continuum of variation between MLE and SSBE on the one hand, and between SSBE and EE on the other hand. The prototypical pronunciations in each cluster, as illustrated in Figure 1, are indisputably different, but the more peripheral productions may span a boundary with a neighbouring cluster. The two features consistently ranked as highest ranking were the principal components relating to the MOUTH and PRICE vowels. This finding is consistent with varying degrees of PRICE-MOUTH cross-over we find when comparing the three clusters. MOUTH is fronted and lowered in EE, retracted in MLE, and somewhat raised in SSBE. PRICE is retracted and raised in MLE, fronted and lowered in MLE, while the SBE realisation is intermediate between the two.

3.2. Demographic predictors

The models informed by only demographic predictors performed considerably worse, with a mean success rate of 54%. Given three possible response categories, this is better than random, however, the error rate is considerable. In terms of variable importance, ethnicity and gender were consistently selected as the most important. Broadly, Asian British and Black British speakers were more likely to be classed as MLE. White British males were more likely to have EE features, whereas the prototypical SSBE speakers were White British females. However, these were only tendencies, and we found representatives of each gender and ethnicity in each cluster.

3.3. Guided accent ratings for pre-selected labels

The models based on guided accent ratings had a mean accuracy of 64%. This is more accurate compared to the models based on demographic predictors, however, the overall accuracy is still relatively low. The two most highly ranked variables in terms of importance were "London" and "Queen's English". MLE speakers had a higher mean rating for their own accent as London: 72 as opposed to 38.7 in SSBE speakers, and 48.9 in EE speakers. SSBE speakers had a higher mean rating in the Queen's English category, compared to the two other groups (45.7 for SSBE, 30 for MLE, 39 for EE). However, given the 100-point scale, speakers typically did not see their own accent as 'Queen's English', regardless of their cluster membership.

3.4. Unguided, free classification of own accent

Table 1 lists the words that appeared most frequently in the descriptions produced by the participants, depending on accent cluster.

Cluster	Term	Frequency
MLE	*London*	17
	British	10
	Nigerian	3
	English	3
SSBE	*London*	17
	posh	13
	Essex	13
	South	11
	British	11
EE	*London*	13
	Essex	8
	Southern	5
	common	5
	British	5

Table 1: The terms used most frequently in the descriptions of own accent by members of each cluster

As we can see from the table, most participants across all three clusters, used the term 'London'. The participants also tended to use geographical terms such as 'British' or 'Essex'. The only terms that

suggest an attitudinal distinction were 'posh', used frequently by SSBE speakers, and 'common', which appears in the descriptions of EE speakers. These terms are evocative of social class stereotypes associated with SSBE (middle-class) and EE (working class). Note, however, that while these terms signal a difference between the aggregated responses, most individuals did not use them.

4. DISCUSSION

This study tested if speakers' self-description of accent, using both pre-selected labels and unguided, free classification, is predictive of their accent. The relationship between demographic predictors and cluster membership is broadly as expected. The MLE cluster is most common in Asian British and Black British speakers, the SSBE cluster is most common in White British females, and most White British males are in the EE cluster. However, the mean accuracy of the demographic predictors was relatively low at 54%, and the predictors do not fully separate the data. This result is perhaps to be expected as the relationship between rates of linguistic features and demographic factors in any speech community is only ever a trend and not categorical.

What is interesting, speakers' ratings of their own accent according to pre-selected labels, predict cluster membership better than demographic factors (64 % vs. 54%). The two accent labels consistently ranked as most important were "London" and "Queen's English". Though speakers in all clusters identified with the "London" label to some extent, MLE speakers had the highest mean rating for this label, followed by EE speakers and then SSBE speakers. Further, as expected, SSBE speakers were those who most identified with the "Queen's English" label. However, even the SSBE speakers did not identify strongly with the "Queen's English" label, suggesting that this term does not capture the accent identity for most south-eastern speakers.

In general, most labels received relatively low mean ratings. It may be that some other labels would have been a better expression of how speakers perceive their own accent. However, it is not clear what those labels might be when we consider the outcome of the unguided self-descriptions. Two possibilities are the terms "posh", which SSBE speakers used in the free-classification task, and "common" which was used by some EE speakers. However, these are morally-loaded labels which may reinforce stereotypes about the value of different speech patterns. In addition, these terms occurred alongside many other responses which overlapped between groups or were not helpful delimiters of the accents. For instance, speakers in both the SSBE and EE clusters, but not the MLE cluster, define their accent as "southern", and speakers from all clusters use the term "British". The level of precision is also an important consideration in how reliably speakers' label accents – whether categorising their own accent or that of others. "British" is too broad to meaningfully delineate any of our clusters.

Results of the unguided, free-classification task also indicated some expected trends between groups, but again these did not reliably separate the clusters. Much like for the guided task, "London" was the most prolific term used by speakers in all three clusters. This is likely because all three clusters include speakers from London. Furthermore, as a large, and culturally and linguistically heterogeneous city, "London" likely holds ambiguous designations. This result coincides with previous research which has found much variability on language attitude scales when participants are presented conceptually with a "London" accent [10, 11].

A further geographic label which SSBE and EE speakers commonly used in the free-classification task was "Essex", a county bordering East London where accents have been heavily influenced by Cockney [12]. "Essex" occurred frequently but no other county was explicitly mentioned. A person in the EE cluster might define their accent as "London" or "Essex" if they are from these places, but if they are not, they are more likely use more generic terms like "Southern" or "common". "Essex", like "London", has become a salient and socially-meaningful label in how linguistic variation is categorised in the Southeast. Our findings highlight that the importance of place in defining accent varies considerably depending on the sociolinguistic salience of particular varieties and the cultural prominence [13] of the location.

5. CONCLUSIONS

Though much research has looked at how speakers categorise and label the speech of others, this paper tested how reliably speakers self-define their own accent. In both the guided task in response to pre-selected labels and the unguided, free-classification task, patterns have emerged which broadly align with expectations for the SSBE, EE and MLE speakers. However, self-description is a relatively poor predictor of actual linguistic production in the sense that it does not entirely separate the speakers with different patterns of vowel production. These results have important implications for linguistic research that uses self-description as a tool for identifying authentic speakers of a variety.

6. REFERENCES

[1] Leach, H., Watson, K., Gnevsheva, K. 2016. Perceptual dialectology in northern England: Accent recognition, geographical proximity and cultural prominence. *Journal of Sociolinguistics* 20(2), 192-211.

[2] McKenzie, R. M. 2015. The sociolinguistics of variety identification and categorisation: Free classification of varieties of spoken English amongst non-linguist listeners. *Language Awareness* 24(2), 150-168.

[3] Cole, A., Strycharczuk, P. Accepted. The search for linguistically coherent accents: Unsupervised clustering of diphthong variation in Southeast England. *English World-Wide*

[4] Agha, A. 2003. The social life of cultural value. *Language & Communication* 23(3-4), 231-273.

[5] Milroy, J. 2001. Received pronunciation: Who "receives" it and how long will it be "received"? *Studia Anglica Posnaniensia* 36, 15–33.

[6] Fabricius, A. 2002. Ongoing change in modern RP: Evidence for the disappearing stigma of t-glottalling. *English World-Wide* 23(1), 115-136.

[7] Lindsey, G. 2019. *English after RP: Standard British pronunciation today.* Palgrave Macmillan.

[8] Altendorf, U. 2003. Estuary English: levelling at the interface of RP and South-Eastern British English. Gunter Narr Verlag Tübingen

[9] Cheshire, J., Kerswill, P., Fox, S., Torgersen, E. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *Journal of Sociolinguistics* 15(2), 151-196.

[10] Bishop, H., Coupland, N., Garrett, P. 2005. Conceptual accent evaluation: Thirty years of accent prejudice in the UK. *Acta Linguistica Hafniensia: International Journal of Linguistics* 37(1), 131–54.

[11] Cole, A. 2021. Disambiguating language attitudes held towards sociodemographic groups and geographic areas in South East England. *Journal of Linguistic Geography* 9(1), 13–27.

[12] Cole, A. 2022. Cockney moved East: the dialect of the first generation of East Londoners raised in Essex. *Dialectologia et Geolinguistica* 30(1), 91-114.

[13] Montgomery, C. 2012. The effect of proximity in perceptual dialectology. *Journal of Sociolinguistics* 16(5), 638-668.

VOWEL RAISING ACROSS SYRIA AND JORDAN IN THE DIVAL CORPUS

Sam Hellmuth[1], Rana Almbark[1], Chris Lucas[2] and Georgina Brown[3]

[1]University of York, [2]SOAS University of London, [3]Lancaster University
sam.hellmuth@york.ac.uk, rana.alhusseinalmbark@york.ac.uk, cl39@soas.ac.uk, g.brown5@lancaster.ac.uk

ABSTRACT

Levantine Arabic dialects display a pervasive pattern of front vowel raising (*imala*), whose distribution is subject to variation and change across the region. There are few acoustic descriptions of *imala*, but prior auditory transcriptions point to variation in the degree of raising, as well as in the categorical presence or absence of *imala*, between contexts and dialects. We report values of F1/F2 and F2-2F1 in two potential *imala* contexts (word-internal /a:/ and feminine suffix /-a/), in scripted speech produced by 123 speakers of dialects from across Syria and Jordan, from the Dialect Variation in the Levant [DiVaL] corpus, with comparison to earlier dialect descriptions. We find limited evidence of present day dialectal differences in the conditioning environments of *imala* of word-internal /a:/, but clear evidence of both gradient and categorical variation in the degree of movement in the front vowel diagonal affecting feminine suffix /-a/.

Keywords: dialectal variation, Arabic, vowel raising, speech corpora.

1. INTRODUCTION

Levantine Arabic dialects display varying degrees of a vowel raising/fronting pattern known in Arabic as *imala*. This name dates back to the very earliest descriptions of Arabic phonetics by mediaeval grammarians [1], and the typical phonological conditioning and regional distribution of the pattern is well-known [2-4]. However, even though *imala* has been described as involving lower F1 and higher F2 [3], there are very few studies which actually use or report results of acoustic analysis of *imala* [5-7]. This study seeks to help remedy the lack of acoustic descriptions of *imala* by exploiting a new corpus of speech data in dialects from across Jordan and Syria: the Dialect Variation in the Levant (DiVaL) [8].

Our aim is: i) to introduce the DiVaL corpus, ii) to demonstrate its potential through acoustic analysis of F1 and F2 in two *imala* conditioning contexts (word-medial /a:/ and word-final feminine suffix /-a/, and iii) to compare the results with prior dialectological descriptions based on auditory impression [9].

2. BACKGROUND TO THE STUDY

2.1. Prior descriptions of *imala*

The traditional term *imala* ([ʔima:la] lit. 'inclination') covered a range of phenomena involving some degree of raising (and fronting) of /a/. Here we focus on general *imala* of word-medial long /a:/ and *imala* in the morphological context of word-final feminine suffix /-a/. One type of *imala* does not entail the other, but in all Arabic varieties that exhibit one or both types, *imala* is blocked by emphatic consonants (with secondary post-velar articulation) which trigger backing of /a/ ([tafxi:m] lit. 'intensification'). Syrian dialects vary in both the presence and degree of *imala*: for word-medial /a:/ [9] gives Damascus *lābis* (no *imala*) vs. Aleppo *lēbis* vs. Soukhne (Homs region) *lībis* 'wearing'; but for suffix /-a/: Dēr iz-Zōr (NE Syria) *wazza* (no *imala*) vs. Damascus *wazze* vs. Soukhne *wazzi* 'goose'; *imala* of word-medial /a:/ is present in all non-*tafxi:m* environments in some dialects (e.g. Tartus) but in others only in non-*tafxi:m* environments historically adjacent to a high front vowel or palatal consonant (most of NW Syria).

Recent sociolinguistic studies show that both types of *imala* are subject to variation and change in Levantine dialects and beyond [7, 10, 11], so the incidence of *imala* in Syria may have changed since the 1980s data collection which informed [9]. In one of the best sociophonetic descriptions of *imala* to date, [5] measured F1 and F2 at five points in tokens of suffix /-a/ in an age-stratified sample of Palestinian Arabic speakers from Gaza City (using a within-speaker linear-normalised 'N-score' [(F2-F1) / F1]), and found significant differences between the youngest and oldest speakers at all measuring points. Crucially, [5] revealed variation in apparent time in the phonetic realisation of /-a/, which had not been detected in auditory analysis of the larger corpus from which the analysed data subset was drawn [12].

2.2. Dialect Variation in the Levant [DiVaL] corpus

A number of very large speech databases of Arabic dialects exist [13], but their size and data collection practices lead authors such as [5] to concede to frequent errors in the metadata. In addition, the level

of granularity in the metadata typically cannot support analysis at the sub-regional level. The DiVaL corpus was created to facilitate reliable and detailed analysis tasks to be carried out at a fine-grained level.

Data was collected in 2021 using four production tasks: *st*: reading a scripted folk *story* twice; *rs*: ten scripted *read sentences* which target phonological and morphological variables such as phoneme /q/ and feminine suffix /-a/; *gs*: free translation of three *grammar sentences* from Standard Arabic to dialect, which target morpho-syntactic and lexical variables such as negation; *pd*: three *picture description* tasks, with visual prompts which target lexical items known to vary within and across dialects.

Text for the *st* and *rs* tasks were presented in pdf format in Arabic script using dialectal spelling conventions (see Table 1). Participants were asked to read in their own dialect not Standard Arabic. Text for the *gs* task was presented in Standard Arabic.

Participants were recruited to represent key major dialect regions in Syria and Jordan. All Jordanian data and most Syrian data was collected in Jordan; a few Syrian-origin speakers were recorded in the UK. Since most Syrians are displaced, their data is coded according to place of origin and not where recorded. In Jordan we worked with local fieldwork assistants to ensure the accuracy of the place of origin metadata. Full metadata will be published with the corpus [8].

Due to Covid restrictions, data was self-recorded remotely using the Awesome Voice Recorder (AVR) smartphone app [14]. We used AVR to obtain wav audio files containing full spectral information [cf. 15]. Other platforms were explored but these did not reliably yield wav files with full spectral information.

3. METHODS

This study uses data from the story (*st*) and read sentences (*rs*); sample scripted items are in Table 1.

The full corpus comprises 133 speakers: 52 from Jordan; 81 from Syria. We also identified subsets of speakers who can be grouped into linguistically meaningful dialect subgroups in Jordan or Syria with at least 14 speakers per subgroup. Table 2 shows the split of speakers in these subgroups by gender, and Figure 1 visualises speakers' place of origin as a map. The speakers were aged between 18-65 at time of recording with median age in the 26-35 age range. Level of education is included in the corpus metadata.

A romanised orthographic transcription was force aligned to the data using Prosody Lab Aligner [16]. Tokens of word-medial long /a:/, feminine suffix short /-a/ and word-medial short /-a-/ were identified using a Praat script, which also extracted the first and second formants (F1 and F2) in each identified vowel at the vowel midpoint. To remove erroneous values potentially arising from tracking or alignment errors, outliers were removed using the Modified Mahalanobis Distance method implemented in [17] using the joeyr package [18]. The remaining formant values were Lobanov normalised [19] using phonR [20]. We calculated linear F2-2F1 [F2-(2*F1)] which has been shown in prior dialect contact studies [21] to effectively capture direction and extent of movement in the front diagonal of the vowel space in a single metric. Values of F2-2F1 were explored in linear mixed-effects models (LMM) using lme4 [22].

Code	Target
rs08	الكل ظروفو صعبة كثير هال أيام
	/l-kull ðˤuru:f-o-h sˤaʕb-a hal-ʔajja:m/
	<lkull DHuru:fuh Sa3ba hal2ayya:m>
	Everyone's circumstances are hard these days.
st13	قالّو جحا ثلاثة ريال ومافي غير ها
	/qallu ʒuħa θala:θa rja:l w ma:fi ɣayrha/
	<qallu juHa thala:tha rya:l w ma:fi ghayrha>
	Juha told him: "Three riyals and nothing more"

Table 1: Arabic orthography, IPA, romanised transcription and translation of sample stimuli.

Code	Dialect area	F	M	Total
sy-neast	Al-Raqqa/Al-Hasakah	10	4	14
sy-nwest	Aleppo/Idlib	8	6	14
sy-homha	Homs/Hamah	8	8	16
sy-damas	Damascus + environs	12	6	18
sy-south	Daraa	8	6	14
jo-rural	Irbid/Al-Ramtha/Ajloun	9	7	16
jo-urban	Amman/Al-Salt	11	4	15
jo-south	Karak/Al-Tafilah/Ma'an	9	7	16
Total		76	48	123

Table 2: DiVaL subgroups included in this study.

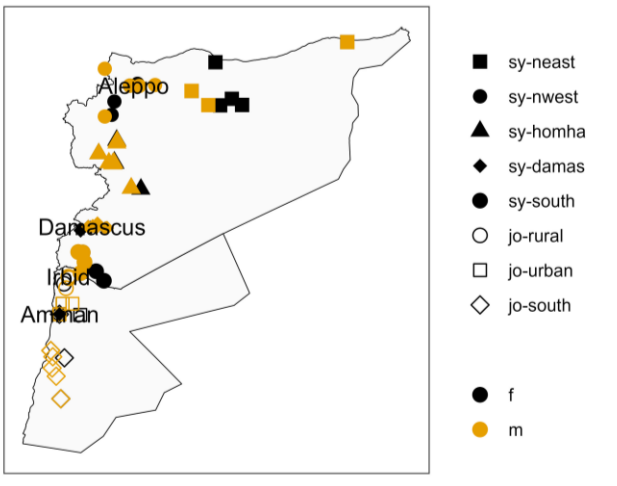


Figure 1: Place of origin of DiVaL speakers in this study.

4. RESULTS

4.1. Word-medial /a:/ in different environments

We examined F1 and F2 at the midpoint of four types of target words produced in the *rs* task: words

expected to contain plain /a:/ such as <thala:tha> /θala:θa/ ‘three’; words where /a:/ is preceded by a palatal such as <2ayya:m> /ʔajja:m/ ‘days’; words followed by a palatal or front vowel such as <na:zil> /na:zil/ ‘going down’; and words containing a trigger for tafxi:m (e.g. an emphatic coronal consonant) such as <Ta:la3> /tˤa:liʕ/ ‘going up’. Target /a:/ in [tˤa:laʕ] is expected to resist *imala* in all dialects due to the preceding emphatic [tˤ] and following post-velar [ʕ].

Our research question in this section is thus: to what extent do any patterns of (non-)overlap in the phonetic realisation of /a:/ in plain versus palatal versus emphatic contexts in the DiVaL corpus (collected in 2021) align with patterns reported in the earlier dialectological description by [9]. The corpus subset yields 2617 tokens for analysis (Table 3).

dialect	plain	prepal	follpal	emphatic	Total
sy-neast	102	38	90	56	286
sy-nwest	106	33	88	66	293
sy-homha	118	46	106	77	347
sy-damas	132	50	118	85	385
sy-south	98	39	93	67	297
jo-rural	127	41	102	71	341
jo-urban	107	39	102	72	320
jo-south	121	45	106	76	348
Total	911	331	805	570	2617

Table 3: Count of /a:/ tokens by dialect~wordtype.

Figure 2 visualises the F1/F2 vowel space of the different dialects, with varying degrees of overlap of formant values in /a:/ in plain versus pre-/post-palatal environments, and variation also in the degree of relative lowering/backing in emphatic environments.

The data were explored using a series of LMMs to predict values of F2-2F1 in the /a:/ data subset, with *wordtype* and dialect *subgroup* plus the interaction between them as fixed effects; the best fit model included fixed effects for *gender* and *task* (but not for *age* or *education*), with random intercepts for *speaker* and *word* and a random slope for *speaker* by *wordtype*. The model indicates a main effect of *task* (with larger F2-2F1 difference in *rs* across the board), but no effect of gender. There is a significant dialect *subgroup* by *wordtype* interaction with smaller F2-F1 values in emphatic environments in all dialects, as expected, being lower/backed in the vowel space.

Figure 3 visualises model predictions: the main differences are in the degree of difference between the emphatic versus all other environments, with a smaller degree of raising in Damascus and Daraa than in northern Syrian dialects, and which is similar to the degree of raising in northern Jordanian dialects. There is no indication of a different degree of vowel raising in plain versus palatal contexts, contra Behnstedt’s observation in [9] that *imala* applied only in palatal contexts in northern Syria (equivalent to our sy-neast and sy-nwest).

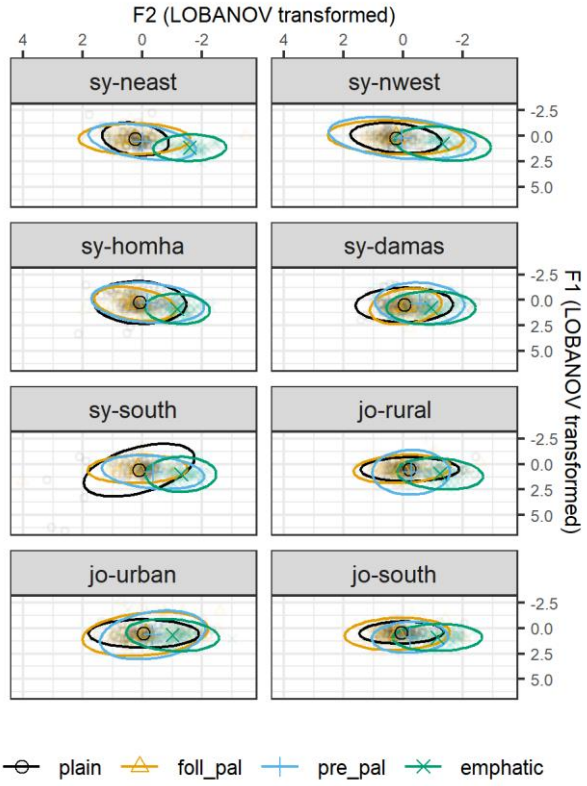


Figure 2: Mean normalised F1/F2 at the vowel midpoint in word-medial /a:/ by environment and by dialect group.

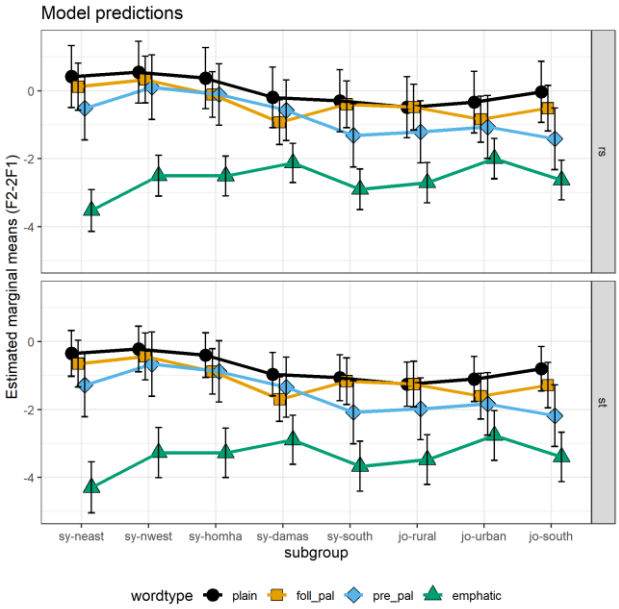


Figure 3: Estimated marginal mean + 95%CI of normalised F2-2F1 at the vowel midpoint in word-medial /a:/ by *wordtype*, dialect *subgroup* and *task*.

4.2. Feminine suffix /-a/ versus word-medial /-a-/

We examined F1/F2 at the midpoint of suffix /-a/ and word-medial /-a-/ in three target words produced at multiple points in the *rs* and st tasks: <madrasa> /madrasa/ 'school', <thala:tha> /θala:θa/ 'three', and <madi:na> /madi:na/ 'city'.

Our research question here is: to what extent do patterns of (non-)overlap in phonetic realisation of suffix /-a/ versus word-medial /-a-/ in DiVaL (collected in 2021) align with patterns reported in earlier dialectological descriptions by [9]. The corpus subset yields 3156 tokens for analysis (Table 4).

dialect	/-a/ suffix	medial /-a-/	Total
sy-neast	176	192	368
sy-nwest	172	180	352
sy-homha	204	213	417
sy-damas	225	234	459
sy-south	175	190	365
jo-rural	199	213	412
jo-urban	184	194	378
jo-south	195	210	405
Total	1530	1626	3156

Table 4: Short /a/ tokens by dialect~environment.

Figure 4 visualises the F1/F2 vowel space of each dialect, with varying degrees of overlap of formant values in suffix /-a/ versus word-medial /-a-/ between dialect sub-groups. The data were explored using a series of LMMs to predict values of F2-2F1 in the short [a] data subset, with *voweltype* and dialect *subgroup* plus the interaction between them as fixed effects; the best fit model included a fixed effect for *age* only (but not for *gender*, *task* or *education*), with a random slope for *speaker* by *voweltype*. There is a main effect of *age* (less F2-2F1 difference for the one male speaker in the oldest 56-65 age group, consistent with natural effects of aging on formant values). There is a significant dialect *subgroup* by *voweltype* interaction such that the degree of F2-2F1 difference between suffix /-a/ versus word-medial /-a-/ varies.

Figure 5 visualises the model predictions. There is no *imala* in the Al-Raqqa (sy-neast) data, and clear raising in suffix /-a/ in northern Syrian and southern Jordanian, but a reduced (intermediate) degree of raising in southern Syrian dialects (Damascus and Daraa). These patterns align closely with Behnstedt's observations about *imala* of suffix /-a/ in Syrian varieties, both as a categorical (absent in Al-Raqqa) and gradient (raised to [i] in the Homs region but only to [e] in Damascus) phenomenon. The patterns also align with raising patterns in Jordan reported in [11], that is, with raising in all equivalent dialect groups to the three investigated here, but noting a higher degree of raising in southern Jordan (e.g. Karak) [4 fn4].

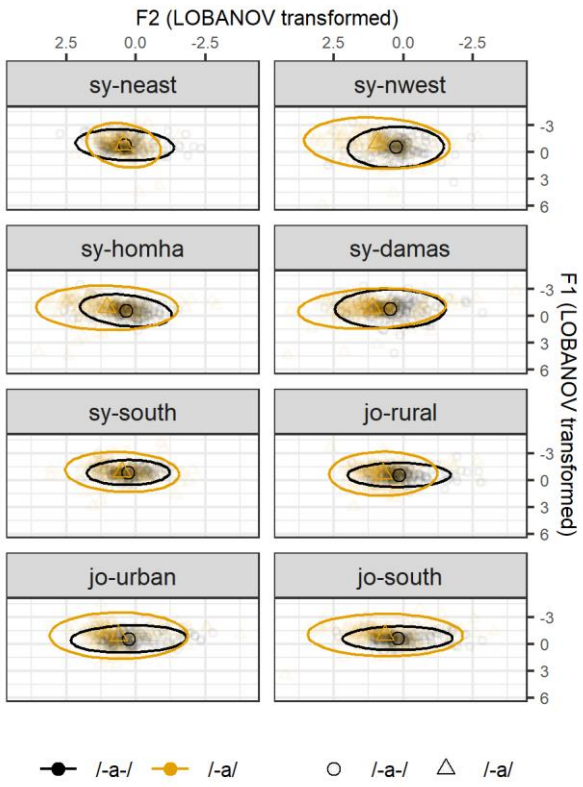


Figure 4: Mean normalised F1/F2 at the vowel midpoint in suffix /-a/ and word-medial /-a-/ by dialect group.

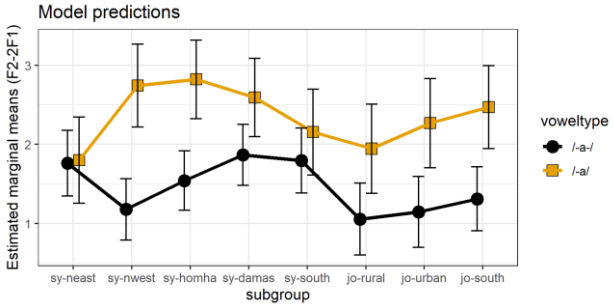


Figure 5: Estimated mean + 95%CI of norm. F2-2F1 at midpoint in suffix /-a/~word-medial /-a-/ by dialect.

5. CONCLUSION

We report quantitative investigation of F1/F2 in tokens of long and short /a/ in two contexts expected to display the vowel raising pattern known as *imala* (word-internal /a:/ and feminine suffix /-a/), from data newly collected in 2021 with speakers of a range of Syrian and Jordanian dialect areas. Observed patterns in the 2021 data, of categorical and/or gradient variation in the degree of raising in each context, mostly align with the earlier descriptions in [9], except for the apparent lack of sensitivity of *imala* to plain versus palatal contexts in NW Syria. This initial indication based on group-level acoustic analysis, can be further explored in future through auditory analysis at speaker level, thanks to the availability of the Dialect Variation in the Levant [DiVaL] corpus.

6. REFERENCES

[1] Owens, J. 2006. *A linguistic history of Arabic*. Oxford: Oxford University Press.

[2] Levin, A. 2011. ʾImāla, in *Encyclopedia of Arabic Language and Linguistics*, K. Versteegh, et al., Editors. Brill: Leiden. p. 311-315.

[3] Barkat-Defradas, M. 2008. Vowel raising, in *Encyclopedia of Arabic Language and Linguistics*, K. Versteegh, et al., Editors. Brill: Leiden. p. 678–82.

[4] Mitchell, T.F. 1993. *Pronouncing Arabic 2*. Oxford Clarendon Press.

[5] Cotter, W.M. 2016. A sociophonetic account of morphophonemic variation in Palestinian Arabic. *Proceedings of Meetings on Acoustics*, **26**(1): p. 060001.

[6] Kelly, N.E. 2017. A phonetic case study of a bidialectal speaker of Lebanese and Palestinian Arabic, in *Paper presented at Structural and Developmental Aspect of Bidialectalism, The Arctic University of Norway, Tromsø, Norway*.

[7] Abou Taha, Y. 2022. Contact-Induced Change in the Levantine: Evidence from Lebanese and Palestinian Arabic. PhD Dissertation, University of Ottawa.

[8] Almbark, R., S. Hellmuth, and G. Brown. (forthcoming). *Dialect Variation in the Levant*. UKDS. https://reshare.ukdataservice.ac.uk/856484/

[9] Behnstedt, P. 1997. *Sprachatlas von Syrien: Kartenband*. Otto Harrassowitz Verlag.

[10] Hennessey, A. 2011. The linguistic integration of the Palestinian refugees in Beirut: a model for analysis. MA Dissertation, American University of Beirut.

[11] Horesh, U., et al. 2022. Dialect contact and change in the Arabic feminine ending morpheme, in *Perspectives on Arabic Linguistics XXXIII. Papers selected from the Annual Symposium on Arabic Linguistics, Toronto, Canada, 2019*, A.-K. Ali and A. Hachimi, Editors. John Benjamins Publishing Company. p. 27-49.

[12] Cotter, W.M. 2013. Dialect contact and change in Gaza City. *Unpublished MA thesis. Colchester, UK: University of Essex*.

[13] Ali, A., et al. *The MGB-5 challenge: Recognition and dialect identification of dialectal arabic speech*. in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019. IEEE.

[14] Newkline. 2020. Awesome Voice Recorder [Android/iOS smartphone application]. Newkline Ltd.

[15] Zhao, L. and E. Chodroff. *The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects*. in *Proceedings of the Language Resources and Evaluation Conference*. 2022. ELRA.

[16] Gorman, K., J. Howell, and M. Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Journal of the Canadian Acoustical Association*, **39**(3): p. 192-193.

[17] Stanley, J.A. 2020. 6. The Absence of a Religiolect Among Latter-Day Saints in Southwest Washington. *Publication of the American Dialect Society*, **105**(1): p. 95-122.

[18] Stanley, J.A. 2021. joeyr: Functions for Vowel Data (R package version 0.6.2).

[19] Adank, P., R. Smits, and R. Van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, **116**(5): p. 3099-3107.

[20] McCloy, D.R. 2012. Vowel normalization and plotting with the phonR package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, **1**: p. 1-8.

[21] Dinkin, A.J. 2013. What's really happening to short-a before L in Philadelphia? *American Speech*, **88**(1): p. 7-31.

[22] Bates, D., et al. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1): p. 1-48.

Developmental changes of fundamental frequency and temporal characteristics in Estonian adolescent speech

Einar Meister, Lya Meister

Tallinn University of Technology, Estonia
einar.meister@taltech.ee, lya.meister@taltech.ee

ABSTRACT

Age- and gender-related anatomical changes and the development of the speech-motor system during childhood and adolescence affect speech acoustics. Development trends of F0 and temporal features in the speech of Estonian adolescents are investigated as a function of age and gender, using speech samples read by 309 Estonian-speaking subjects (175 girls and 134 boys) aged 9–18 years. The results show the expected age-dependent patterns: (1) a decline in F0 with a sharp drop of about 100 Hz in boys aged 12–15 years and a gradual decline in girls, (2) an increase in speech and articulation rates up to 15 years of age. Gender differences in temporal characteristics are not significant, although boys tend to speak slightly faster than girls. The results are in line with the findings reported for several other languages and can be considered reference data for Estonian-speaking subjects aged 9–18 with typical language development.

Keywords: adolescent speech, age- and gender-related variations, acoustic analysis, fundamental frequency, speech tempo

1. INTRODUCTION

Age- and gender-related anatomic changes of the vocal apparatus during childhood and adolescence have a direct impact on speech acoustics manifested in a sudden drop of fundamental frequency (F0) during puberty in boys and in a gradual decrease of the acoustic vowel space and formant frequencies in both genders [13, 15, 26, 27]. In parallel with anatomical changes, the development of speech-motor control as well as cognitive and linguistic processing takes place, which manifests itself in the increase of speech and articulation rates [14, 24].

In infants (under 1 year old) the length of the vocal folds is 4–5 mm, at the age of 20 the vocal folds have reached a length of 11–15 mm in women and 17–25 mm in men [7, 20]. Consistent with vocal fold growth, changes in F0 have been documented, e.g. mean F0 in boys at the age of 7 is 266 Hz, at the age of 12 years 226 Hz, and lowers to 127 Hz by the age of 15 with marginal changes thereafter. The average F0 of 7-year-old girls is 275 Hz and gradually decreases to 231 Hz until the age of 12, after which there is no significant change [13]. However, another study reports a further decline of girls' mean F0 from 230 Hz to 218 Hz between the ages of 13 and 19 [5].

Studies of speech-motor development have shown that children's articulatory movements are slower and more variable than that of adults. A study [24] examined native English-speaking children and adults aged 4–22 years by recording upper lip, lower lip, and jaw movements while reading different sentences. The results showed that as age increases, the variability of articulatory movements decreases and the time required to form a sentence shortens, and speech-motor control in both boys and girls becomes similar to adults only after the age of 14. The study concludes that the age-related increase in speaking rate is due to improvements in cognitive and linguistic processing and motor control of speech.

According to various studies, the articulation rate in the spontaneous speech of English-speaking children aged 3–6 years varies from 2.9 to 4.3, and the speech rate from 2.3 to 2.6 syllables per second; for 7- to 12-year-olds, 4.5 to 5.6 and 2.4 to 2.9 syllables per second, respectively [14]. The speaking rates increase equally with age in boys and girls, and the differences within age groups are not statistically significant (however, males tend to speak slightly faster) e.g. [10, 11, 12, 19]. Articulation rates of 3.3–5.8 and the speaking rates of 3.7–4.7 syllables per second have been documented in adults [12].

When reading written texts aloud, the speaking rate is affected by the text length as confirmed by several studies e.g. [1, 4, 22]. This is known as anticipatory shortening, whereby in the speech planning process, the speaker adjusts his average syllable duration to the expected length of the phrase [2]. The study [22] explored the relationship between text length and speaking rate in a comparison of children (aged 5–16) and adults (aged 20–23) and found that the duration of a test phrase read in isolation is longer than when read in a frame sentence both in adults and 9-16-year-olds, but not in 5-7-year-old children. The authors hypothesize that children and adults use different motor planning strategies when reading more complex sentences: younger children plan their speech in smaller units (words or syllables), while older children and adults plan speech in longer units (phrases).

The present study aims to document the acoustic characteristics of Estonian adolescent speech related to the speaker's age and gender and gain a better understanding of these variations from a cross-linguistic perspective. In particular, we explore (1) the acoustic variations of the fundamental frequency F0, and (2) the measures of speaking rate.

2. MATERIALS AND METHOD

2.1. Speech corpus

The Estonian adolescent speech corpus [16] consists of speech samples from 309 native-speaking subjects (175 girls and 134 boys) aged 9–18 years. The corpus contains cross-sectional speech data of different age groups recorded in ten schools across Estonia with high-quality recording equipment. 60 read and 10 spontaneous items were recorded from each speaker, averaging about 15 minutes of speech per speaker. In total, the corpus contains approximately 70 hours of speech.

In the present study, acoustic analyses were performed on a sub-corpus of read speech samples consisting of 20-25 utterances (average duration of utterance 7.8 seconds) per subject, totalling 6964 utterances. All utterances were manually segmented at the word and phoneme levels using Praat [3]. Syllable boundaries and types were added using a custom Praat script.

2.2. Acoustic analysis

In the study, fundamental frequency (F0) and speech and articulation rates were explored depending on gender and age.

2.2.1. F0

For the F0 analysis, a custom Praat script was compiled using the two-step procedure recommended in [8]. First, the F0 values of each utterance were found in the frequency range 75–600 Hz, then the range was narrowed according to the speaker's F0 variation as follows: F0max = 1.5 x 3rd quartile value, F0min = 0.75 x 1st quartile value. This approach takes into account the subject's individual F0 range and thus provides more reliable results. However, about 20% of the utterances required manual F0 validation in cases where the automatically found F0 values seemed unlikely, e.g. when boys had maximum F0 values above 400 Hz or girls had minimum F0 values below 150 Hz. There were also more F0 detection errors in the speech of boys with a pubertal voice mutation period, where F0 variations were larger and sometimes reached the falsetto register.

2.2.2. Speaking rate

To calculate the speaking rate measures, the number of syllables in each utterance, and utterance duration with intra-sentence pauses and without pauses (pauses longer than 250 ms were excluded) were found. Two speaking rate measures for each utterance were calculated:

- the speaking rate = the number of syllables in an utterance / the duration of an utterance including pauses,
- the articulation rate = the number of syllables in an utterance / the duration of an utterance without pauses.

2.3. Statistical analysis

For statistical data processing RStudio [21] was used with the *mgcv* package [29] for Generalized Additive Mixed Models (GAMM) and the *itsadug* package [25] for model validation and visualization of the results).

3. RESULTS

3.1. F0

The histograms (Figure 1) show the distribution of mean F0 in boys and girls. In the case of boys, the F0 distribution is binomial with clearly distinguished peaks at 110 Hz and 215 Hz. The first peak represents the most frequent mean F0 value of boys who have undergone pubertal voice change (boys aged 15–18), and the second peak corresponds to the most frequent mean F0 value of younger boys (aged 9–12) before pubertal voice change. As expected, there is only one peak in the histogram for girls, representing the median F0 value for all girls (226 Hz).

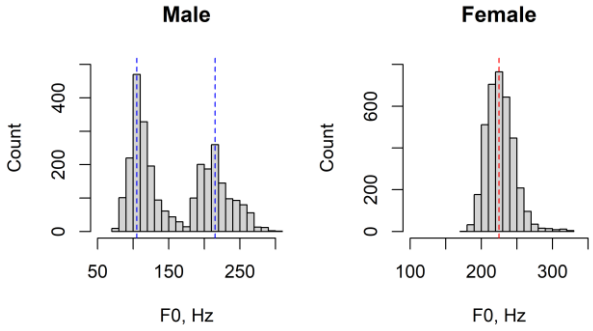


Figure 1: Histograms of F0 mean values, males on the left, females on the right.

Figures 2 and 3 represent the boxplots of the measured F0 means for each age group in boys and girls, respectively. In boys, the general developmental pattern of F0 mean values shows a gradual decrease between the ages 9 and 12 by 17 Hz, a more

prominent decline (by 28 Hz) occurs between the ages 12–13 followed by the largest drop (by 61 Hz) between the ages 13–14, a further decrease (by 19 Hz) continues till age 15. During the ages of 15–18, the mean F0 stabilizes around 110 Hz (in pairwise comparison, the differences between the age groups are still statistically significant). However, individual F0 developmental paths can be different from the general pattern, as the outliers in Figure 3 manifest. There are two 13-years old (F0 means 101 and 113 Hz), three 14-years old (F0 means 228, 194, and 164 Hz), two 15-years old (F0 means 149 and 141 Hz), and one 16-years old (F0 mean 201 Hz) boys whose F0 mean values deviate significantly ($p<0.001$) from the other speakers in the respective age group (F0 means 199, 114, 105, and 111 Hz, respectively). We suggest that these deviating F0 mean values reveal an early (in the case of 13-year-olds) and late (in the case of 14–16-year-olds) onset of the pubertal voice mutation period.

In girls, F0 development shows an almost linear decline pattern between the ages 9 and 18, with the decrease of mean F0 from 245 Hz to 212 Hz. The outliers occurring in several female age groups could be attributed to the individual characteristics of the subjects' voice source.

Gender difference is significant ($p<0.001$) already from age 9 (F0 means 235 and 245 Hz for boys and girls, respectively).

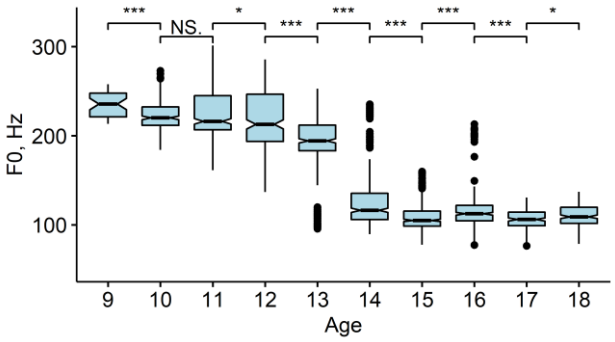


Figure 2: Male speakers' boxplots of F0 means by age with significance levels of pairwise t-test (NS. $p > 0.05$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

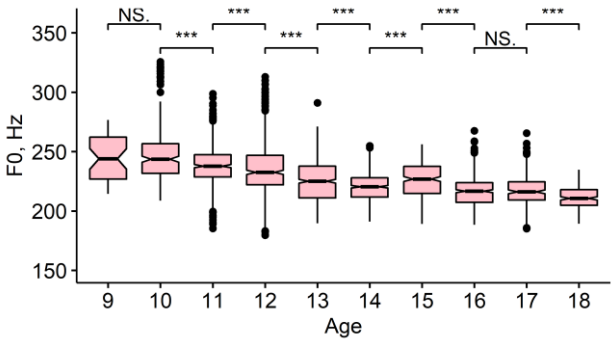


Figure 3: Female speakers' boxplots of F0 means by age with significance levels of pairwise t-test.

The measured F0 mean values for each read utterance were used to fit the GAMM with the smooth term *age,* the explanatory factor *gender,* and *subject* as the random effect. The model-predicted F0 values for both genders are presented in Figure 4.

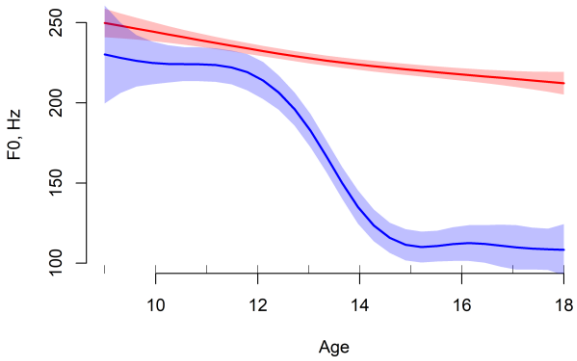


Figure 4: The age-dependent change of F0 mean with 95% confidence bands for girls (top) and boys (bottom) as predicted by the GAMM.

According to the GAMM's prediction, in girls between the ages of 9 and 18 years, the F0 mean gradually decreases from 250 Hz to 212 Hz. In boys between the ages 9 – 12, the mean F0 decreases by 17 Hz, followed by a rapid drop of F0 (by 105 Hz) between the ages 12–15, followed by minor F0 variations in further ages.

Table 1. The GAMM-predicted mean F0 values and the standard errors (SE) for each age group by gender (in Hz).

		9	10	11	12	13	14	15	16	17	18
F	F0	250	244	238	233	228	224	221	218	215	212
	SE	4,4	2,7	1,6	1,2	1,1	1,1	1,3	1,6	2,2	3,4
M	F0	230	225	224	216	184	134	111	112	110	108
	SE	15,6	6,2	4,9	5,5	4,8	4,8	4,4	5,1	6,7	7,8

3.1. Speaking rate characteristics

GAMMs were fitted for the speech and articulation rates with the smooth terms *age* and *text length*, and the explanatory factor *gender*; *subject* was added as an independent random variable. The GAMM-predicted speech and articulation rates are presented in Table 2 and the plots of the smooth terms are shown in Figure 5.

Table 2. The GAMM-predicted articulation and speech rates and the standard errors (SE) for each age group (in syllables per second).

	9	10	11	12	13	14	15	16	17	18
Art. rate	4.0	4.3	4.6	4.9	5.0	5.2	5.3	5.3	5.2	5.2
SE	0.13	0.07	0.04	0.04	0.04	0.04	0.04	0.05	0.06	0.09
Sp. rate	3.8	4.2	4.5	4.8	5.0	5.1	5.2	5.2	5.2	5.1
SE	0.14	0.07	0.05	0.04	0.04	0.04	0.04	0.05	0.06	0.09

The data in Table 2 confirm similar age-related development patterns of both speaking rate characteristics (Figure 5, left): at the age 9–15 both rates increase significantly (pairwise comparison of age groups is statistically significant, $p<0.001$) and reach a maximum at 15–16 years of age, in ages 17–18 both rates slightly decrease ($p<0.05$). The effect of text length on the speaking rates has similar patterns (Figure 5, right): in sentences with 8–12 syllables speaking rates increase with the text length (differences are significant at least at the $p < 0.05$ level), in sentences with 12–18 syllables the speaking rates are stable, followed by a decreasing trend (differences are insignificant) in longer sentences (19–20 syllables). There is no gender difference in both rates ($p>0.5$).

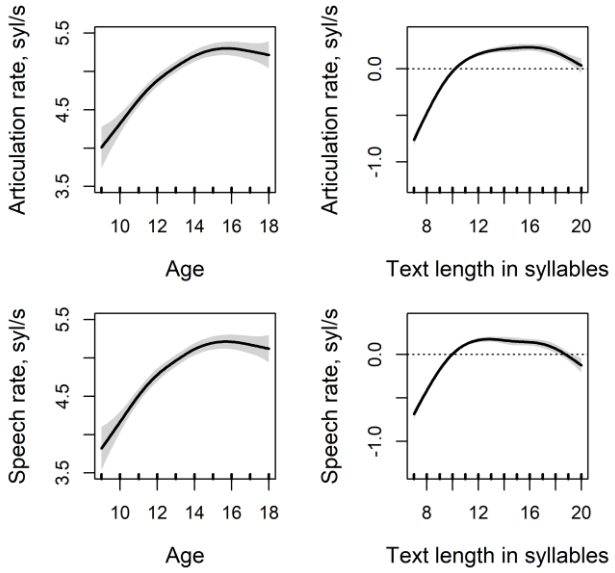


Figure 5. The partial effect of features fitted with the GAMM models for articulation rate (top) and speech rate (bottom) with 95% confidence bands.

4. DISCUSSION

In the study, we explored the developmental characteristics of F0 and two speaking rate measures in Estonian adolescent speech. Due to the cross-sectional nature of the corpus, the findings reveal general patterns of the age- and gender-related development of the studied acoustic characteristics.

According to GAMM prediction, the largest drop in boys' F0 (by 105 Hz) occurs between age 12 (F0 = 216 Hz) and age 15 (F0 = 111 Hz). The finding is in line with [13], which suggests that pubertal voice change in male speakers starts between ages 12 and 13, and ends around age 15. Similar to previous studies [9, 28], individual differences in 13–16-year-old males have been found in our corpus, with mean F0 values that significantly differ from those of age-matched peers suggesting that the onset of puberty varies among speakers.

Both articulation and speech rates increase between the ages of 9 and 15 and become stable at further ages. The difference between articulation rate and speech rate in the analyzed speech material is rather small as there are considerably fewer pauses in read speech compared to spontaneous speech. The effect of text length on speech rate characteristics was significant for sentences up to 12 syllables long. In texts longer than this, the articulation rate shows a further rising trend up to 18-syllable sentences followed by a slight decrease in 19–20-syllable sentences, while the speech rate decreases in 13–20-syllable sentences because speakers make more pauses when reading longer texts. Such patterns of speaking rate variations may be related to motor planning in text reading.

According to previous research, the temporal characteristics of children's speech are acoustically more variable up to the age of 12 years [13] and articulatory up to the age of 14 years [24]. These results have been interpreted as evidence of the achievement of adult-like speech-motor skills at ages 12–14 [18]. On the other hand, it has been argued that the development of motor patterns continues into late adolescence [23]. In line with the latter, it has been reported that the speech rate of 13–14-year-old British children has not yet reached adult levels [6], and the speech rate of 13–17-year-old Hebrew-speaking children continues to increase [1]. The average speech rate of 14–18-year-old Estonian speakers (5.1 syllables per second) compared well to the read speech rate of Estonian young adults of 4.9–5.3 syllables per second [17].

Based on the discussion above and our results, we suggest that the speech-motor control of Estonian adolescents achieve adult-like skills between the ages of 14 and 15 with further improvement of proficiency in ages 15–18.

5. SUMMARY

The study explored the changes in F0 and two speaking rate measures in Estonian adolescent speech as a function of age and gender. The discovered developmental patterns – (1) a decline in F0 with a sharp drop of about 100 Hz in boys aged 12–15 years and a gradual decline in girls, (2) an increase in speech and articulation rates up to 15 years of age without gender-specific differences – are similar to the findings reported for several other languages.

The results of the study will further the knowledge about the age- and gender-related variability of the acoustic properties of adolescent speech and can be considered reference data that are typical for Estonian-speaking individuals aged 9–18 years with normal language development.

6. ACKNOWLEDGEMENTS

This study was supported by the ERDF (the project "Centre of Excellence in Estonian Studies") and by the national programme “Estonian Language Technology 2018-2027” (the project “Speech Recognition”).

7. REFERENCES

[1] Amir, O., Grinfeld, D. 2011. Articulation Rate in Childhood and Adolescence: Hebrew Speakers. *Language and Speech*, 54(2), 225–240.

[2] Bishop, J., Kim, B. 2018. Anticipatory shortening: Articulation rate, phrase length, and lookahead in speech production. *Proceedings of 9th International Conference on Speech Prosody 2018*, 235–239.

[3] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3., retrieved 15 November 2022 from http://www.praat.org//

[4] Darling-White, M., Banks, S. W. 2021. Speech Rate Varies with Sentence Length in Typically Developing Children. *Journal of Speech, Language, and Hearing Research*, 64 (6S), 2385–2391.

[5] Draxler, C., Schiel, F., Ellbogen, T. 2008. F0 of adolescent speakers: First results for the German Ph@ttSessionz Database. *Proceedings of the 6th International Conference on Language Resources and Evaluation*. European Language Resources Assoc. http://www.lrec-conf.org/proceedings/lrec2008/

[6] Hazan, V.; Pettinato, M. 2014. The emergence of rhythmic strategies for clarifying speech: Variation of syllable rate and pausing in adults, children and teenagers. *10th International Speech Production Seminar*, Köln, Germany.

[7] Hirano, M., Kurita, S., Toh, Y. 1981. Growth, Development and Aging of the Vocal Fold, *Practica Oto-Rhino-Laryngologica*, 74(8), 1791–180.

[8] Hirst, D. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the XVIth International Conference of Phonetic Sciences*, 1233–1236.

[9] Hollien, H., Green, R., and Massey, K. 1994. Longitudinal research on adolescent voice change in males. *J. Acoust. Soc. Am*. 96, 2646–2654.

[10] Jacewicz, E., Fox, R. A., O’Neill, C. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2), 233–256.

[11] Jacewicz, E., Fox, R. A, Wei, L. 2010. Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America*, 128(2), 839–850.

[12] Lee, A., Doherty, R. 2017. Speaking rate and articulation rate of native speakers of Irish English. *Speech, Language and Hearing*, 20(4), 206–211.

[13] Lee, S., Potamianos, A., Narayanan, S. 1999. Acoustics of children’s speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am*. vol. 105, no. 3, 1455–1468.

[14] Logan, K. J., Byrd, Courtney T., Mazzocchi, E. M., Gillam, R. B. 2011. Speaking rate characteristics of elementary-school-aged children who do and do not stutter. *Journal of Communication Disorders*, 44(1), 130–147.

[15] Meister, E., Meister, L. 2021. Developmental changes of vowel acoustics in adolescents. *Interspeech 2021: Brno, Czech Republic.* International Speech Communication Association, 4014−4018.

[16] Meister, L., Meister, E. 2014. Development of the corpus of Estonian Adolescent Speech. *Human Language Technologies – the Baltic Perspective: Proceedings of the Sixth International Baltic Conference, Baltic HLT 2014.* Amsterdam: IOS Press, 206−209.

[17] Meister, E., Meister, L. 2022. Estonian adolescent speech III: analysis of speech tempo and syllable durations [in Estonian]. Keel ja Kirjandus, 3, 226−244.

[18] Redford, M. A.; Oh, G. E. 2017. The representation and execution of articulatory timing in first and second language acquisition. *Journal of Phonetics*, 63(1) 127–138.

[19] Robb, M. P., Maclagan, M. A., Chen, Y. 2004. Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics*, 18(1), 1–15.

[20] Rogers, D. J., Setlur, J., Raol, N., Maurer, R., Hartnick, C.J. 2014. Evaluation of true vocal fold growth as a function of age. *Otolaryngol Head Neck Surg*. 151(4), 681–687. doi: 10.1177/0194599814547489.

[21] RStudio Team. 2020. RStudio: Integrated Development for R, RStudio, PBC, Boston, MA. http://www.rstudio.com/.

[22] Sadagopan, N., Smith, A. 2008. Developmental Changes in the Effects of Utterance Length and Complexity on Speech Movement Variability. *Journal of Speech, Language, and Hearing Research*, 51, 1138–1151.

[23] Smith, A. 2006. Speech motor development: Integrating muscles, movements, and linguistic units. *Journal of Communication Disorders*, 39(5), 331–349.

[24] Smith, A., Zelaznik, H. N. 2004. Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, 45(1), 22–33.

[25] van Rij, J., Wieling, M., Baayen, R., van Rijn, H. 2022. itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4.1.

[26] Vorperian, H. K., Kent, R. D. 2007. Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, 1510–1545.

[27] Whiteside, S.P. 2001. Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *J. Acoust. Soc. Am*., vol. 110, no. 1, 464–478.

[28] Whiteside, S. P, Hodgson, C., Tapster, C. 2002. Vocal characteristics in pre-adolescent and adolescent children: a longitudinal study. *Logopedics Phoniatrics Vocology,* 27(1), 12–20.

[29] Wood, S. N. 2017. Generalized Additive Models: An Introduction with R (2nd ed.), CRC Press.

EFFECTS OF DIALECT PRIMING ON PHONETIC CONVERGENCE

Cynthia G. Clopper, Ellen Dossey, Roberto Gonzalez

Ohio State University
clopper.1@osu.edu, dossey.1@osu.edu, gonzalez.594@buckeyemail.osu.edu

ABSTRACT

Non-linguistic social information affects speech perception. The goal of the current study was to examine the effect on phonetic convergence of providing non-linguistic regional information about the model talker in a word shadowing task. Shadowers produced a set of target words in a baseline reading task and then repeated those words after a model talker from the Northern dialect of American English. Half of the shadowers were asked simply to repeat the model talker. The other half were told that the model talker is from the Northern United States and then asked to repeat her. Phonetic convergence was assessed perceptually using an AXB perceptual similarity task. The results revealed significant phonetic convergence overall, with more convergence by the shadowers who were told where the model talker is from than by those who were not. These results suggest that non-linguistic social information can prime phonetic convergence in a word shadowing task.

Keywords: phonetic convergence, word shadowing, dialect priming, Northern Cities Shift

1. INTRODUCTION

Priming social information about a talker affects how that talker's speech is perceived [1, 2, 3, 4, 5, 6, 7]. Effective primes range from movies of talking faces synchronized to the speech signal [7] to still visual images of faces presented synchronously with the speech signal [1, 3, 5, 6] to regional labels at the top of the response sheet [4] to casually-referenced stuffed animals associated with different regions [2]. The effects of these primes on speech perception include changes in perceptual category boundaries between phonemes [7], identification of best-matching synthetic vowels [2, 4], word and sentence recognition [1, 3, 5, 6], and accent ratings [1, 5].

Previous research examining social priming has further demonstrated that primes that accurately reflect the listeners' expectations about the talker's social background can facilitate speech processing, relative to primes that mismatch the listeners' expectations about the talker's social background [1, 3, 5]. For example, McGowan [3] found that speech intelligibility was higher for a Chinese-accented talker paired with a Chinese face than for the same talker paired with a white face. The goal of current study was to explore the effects of this kind of congruent social priming on phonetic convergence in a word shadowing task.

Word shadowing tasks have been shown to elicit phonetic convergence, even in the absence of explicit instructions to imitate the model talker [8, 9]. Although some aspects of the phonetic convergence observed in word shadowing tasks may therefore reflect automatic processes at the perception-production interface [8], more recent research has revealed considerable social selectivity in which features exhibit convergence. For example, the baseline phonetic distance between the model talker and the shadowers affects the magnitude of convergence, such that larger baseline differences allow for greater convergence [10]. As a result, phonetic convergence across dialects can be greater than phonetic convergence within dialects [11].

At the same time, however, social stereotypes appear to affect the magnitude of phonetic convergence to some dialects and linguistic variants. For example, more positive attitudes towards the talker [10, 12] and the talker's dialect region [13] lead to greater convergence. In addition, negative stereotypes associated with particular linguistic variants appear to block convergence to those forms [13, 14, 15]. Thus, even for a single model talker dialect, shadowers can exhibit varying degrees of phonetic convergence across variants, with greater convergence to non-stereotyped forms than to stereotyped forms [13, 14, 15]. For example, Clopper and Dossey [14] observed phonetic convergence in a word shadowing task to non-stereotyped Southern American English /u oʊ/ fronting, but not to stereotyped Southern American English /aɪ/ monophthongization.

The current study examined phonetic convergence to the Northern dialect of American English in a word shadowing task, in which social priming was manipulated across shadowers. The Northern dialect is characterized by the Northern Cities Shift, including the raising and fronting of /æ/, lowering and backing of /ɪ ɛ/, and fronting and lowering of /ɑ ɔ/ [16]. Although the Northern Cities Shift results in a phonetic vowel space that differs considerably from the ideological standard variety of American English, the Northern dialect is not socially stereotyped as a

unique dialect of American English (i.e., it is not enregistered), nor are any of the Northern Cities Shifted variants negatively stereotyped [17]. We therefore expected to observe robust phonetic convergence to the Northern Cities Shifted vowels of the model talker in the word shadowing task in the current study, consistent with previous work suggesting automatic phonetic convergence to non-salient forms [8, 13, 14, 15].

We manipulated social priming in the current study by either providing or not explicit information to the shadowers about where the model talker was from. Given the positive effects of congruent social information on speech processing [1, 3, 5], we expected that providing explicit, authentic information about the model talker's region of origin would facilitate phonetic convergence to the Northern Cities Shift.

2. METHOD

Phonetic convergence was assessed perceptually in an AXB perceptual similarity task [8, 11, 14, 15, 18, 19]. This approach to assessing phonetic convergence was selected because it provides a holistic measure that is difficult to capture using individual acoustic distance measures [15, 18, 19]. In American English, vowel variation is observed across dialects in multiple acoustic dimensions, including midpoint F1 and F2, duration, and formant trajectories [16, 20]. The AXB perceptual similarity task in the current study therefore allowed for a robust test of phonetic convergence because this multidimensional vowel variation could be assessed simultaneously and holistically by the listeners, rather than separately in a series of acoustic analyses.

2.1. Participants

Forty-nine adult listeners (30 female, 19 male) were recruited from a local science museum in Columbus, OH, to participate in the AXB perceptual similarity task. The listeners were all native speakers of American English and ranged in age from 18-67 years old (M = 31 years).

2.2. Stimulus materials

The stimulus materials were produced in a word shadowing task and comprised the model talker utterances from the word shadowing task, as well as baseline read utterances and shadowed (i.e., repeated) utterances from 10 female shadowers. The model talker was a 22-year-old female lifetime resident of the Northern dialect region of American English, selected from the Indiana Speech Project corpus [21]. The 10 female shadowers were all lifetime residents of the Midwestern United States, including both the Midland and Northern dialect regions of American English. They ranged in age from 19-39 years old (M = 28 years).

The word shadowing task involved two blocks. In the first block, the shadowers read a set of 48 multisyllabic English words aloud one at a time from the computer screen. In the second block, the shadowers repeated the same set of 48 words after the model talker. Both the read and shadowed utterances were recorded. The set of 48 words included 24 target words containing a stressed vowel implicated in the Northern Cities Shift (/ɪ ɛ æ ɑ/), as well as 24 filler words containing other stressed vowels (/i oʊ u aɪ/). The 24 target words were used in the AXB perceptual similarity task in this study and are shown in Table 1.

Vowel	Target Words
ɪ	amphibian, aristocrat, conspicuous, imposition, liberator, precipitate
ɛ	clarinet, embezzle, epilepsy, legendary, obsession, silhouette
æ	appetizer, caterpillar, deactivate, evaporate, procrastinate, spatula
ɑ	hypnotic, octopus, rhinoceros, robin, roster, slobber

Table 1: Target words in the AXB perceptual similarity task.

Five of the shadowers were asked in the shadowing block to simply repeat the words aloud after the model talker. The other five shadowers were asked to repeat the words aloud after the model talker from Chicago, IL, a major, well-known city within the Northern dialect region of American English. Thus, the latter group of shadowers ("primed") were primed with the model talker's region of origin, whereas the former group of shadowers ("unprimed") were not.

An acoustic analysis of the stimulus materials was conducted to confirm the presence of the Northern Cities Shift in the model talker utterances and to assess the baseline phonetic distance between the model talker and the shadowers. Given the mixed residential histories of the shadowers, including both the Midland and Northern dialect regions, this acoustic analysis allowed us to confirm sufficient baseline distance between the model talker and the shadowers to ensure that we could observe potential phonetic convergence in the acoustic space [10]. First and second formant frequencies were estimated at the temporal midpoint of the stressed vowel in each target word for the model talker and for each shadower in the baseline block. A summary of these vowel spaces

is shown in Fig. 1, separately for the two shadower priming conditions.

The model talker's vowel space (black symbols) is shifted down and to the left overall relative to the shadowers' vowel spaces (gray symbols), making direct acoustic comparisons between the model talker and the shadowers' baseline vowels difficult. However, the phonetic features of the Northern Cities Shift are evident in the acoustic approximation of /ɛ æ/ and in the relative position of /ɛ æ/ in the F2 dimension for the model talker, in comparison to the unshifted shadowers' baseline utterances. The shadowers' baseline utterances in the two priming conditions are similar overall, except that /ɑ/ is lower and more similar to the model talker's /ɑ/ for the shadowers in the unprimed condition (left panel) than in the primed condition (right panel).

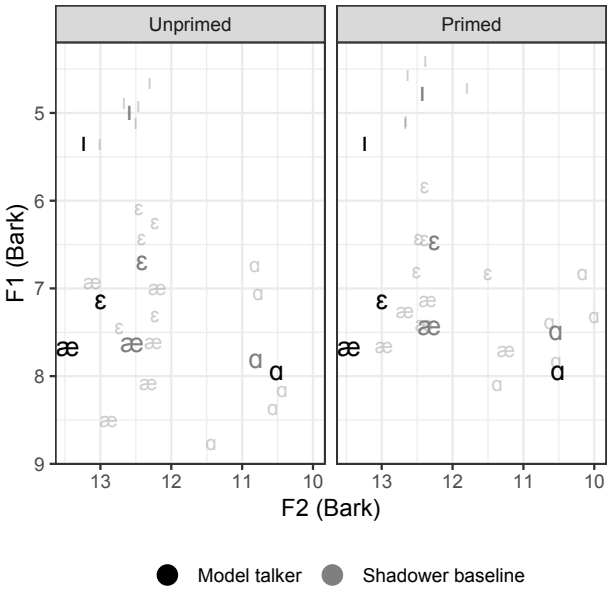


Figure 1: Mean vowel formant frequencies for the model talker (black) and for the shadower baseline utterances (gray), separately by priming condition. Large gray symbols show shadower grand means. Small gray symbols show individual shadower means.

Together, the limited selection of four target vowels and the considerable variation in the relative positions of the selected vowels across the model talker and shadowers make normalization of the acoustic vowel measures inappropriate [20]. This lack of acoustic normalization further motivates the use of a perceptual assessment of phonetic convergence. In particular, we assumed that the listeners in the AXB perceptual similarity task would normalize for overall talker differences and attend to the phonetic properties of the Northern Cities Shift in making their similarity judgments.

2.3. Procedure

The listeners in the AXB perceptual similarity task were seated at personal computers equipped with headphones and button boxes in a quiet lab space in the science museum. On each trial, they were presented with three tokens of the same word: the baseline read utterance from one of the shadowers, the model talker utterance, and the shadowed utterance from the same shadower. The model talker's utterance was always presented second (X). The order of the shadower's baseline and shadowed utterances (A/B) were counterbalanced within and across experimental lists. The three stimulus tokens were separated by 100 ms inter-stimulus intervals. Listeners were asked to press the leftmost button on the button box if the first utterance was more similar to the second utterance and to press the rightmost button on the button box if the third utterance was more similar to the second utterance.

Each listener heard all 24 target words for each of five shadowers, either from the unprimed condition or the primed condition, for a total of 120 trials. Prime condition was therefore a between-listener variable. For each prime condition, two lists were constructed so that across lists, the baseline and shadowed utterances of each word from each shadower appeared in both A/B orders. Trial order was randomized separately for each listener. Twenty-five listeners assessed phonetic convergence by the unprimed shadowers and 24 listeners assessed phonetic convergence by the primed shadowers.

3. RESULTS

Phonetic convergence is assumed to be perceptible when listeners in an AXB perceptual similarity task identify the shadowed utterance as more similar to the model talker utterance than the baseline utterance [18, 19]. The mean proportion of shadowed utterances identified as more similar to the model talker utterance in each of the two priming conditions is shown in Fig. 2. As is typical for perceptual assessments of phonetic convergence [8, 11, 14, 15, 18, 19], the proportions are near 0.50, which corresponds to chance performance in a two-alternative forced-choice task. Despite this low performance overall, performance was higher in the primed condition than in the unprimed condition, as expected.

A logistic mixed-effects model predicting shadowed responses from priming condition, target word stressed vowel, and their interaction was constructed. Both predictor variables were sum-contrast coded. The maximal random effect structure that converged included only a random by-listener

intercept. The model revealed a significant positive intercept (Est. = 0.10, $z = 3.64$, $p < 0.001$), confirming significant perception of convergence overall. The main effect of priming condition was also significant (Est. = -0.07, $z = -2.48$, $p = 0.013$), confirming greater perception of phonetic convergence in the primed condition than the unprimed condition, as expected. Although the model returned a significant slope estimate for /ɑ/ (Est. = -0.10, z = -2.13, p = -0.034), suggesting less convergence to /ɑ/ than to the overall mean, no by-vowel pairwise comparisons using estimated marginal means were significant. The interaction was also not significant.

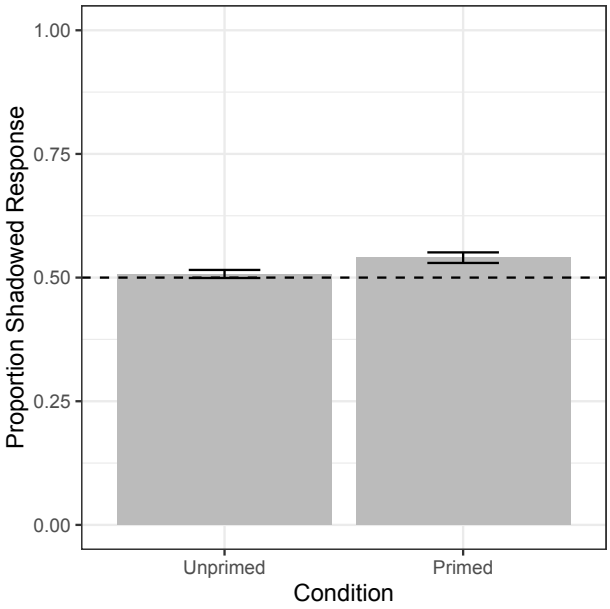


Figure 2: Mean proportion shadowed responses in the AXB task for the two priming conditions (unprimed, primed). Error bars are standard error of listener means.

4. DISCUSSION

The AXB perceptual similarity task revealed significant convergence to the Northern model talker by the Midwestern shadowers, as well as more robust convergence in the primed condition than in the unprimed condition. Consistent with previous work [11, 14, 15, 19], the effect sizes for both the intercept (i.e., overall convergence) and priming condition were small (see Fig. 2), reflecting the difficulty of the AXB perceptual similarity task.

The significant overall convergence is consistent with previous research demonstrating cross-dialect convergence in word shadowing tasks [10, 11, 13, 14, 15]. As shown in Fig. 1, the model talker produced phonetic features of the Northern Cities Shift and the shadowers' baseline exhibited sufficient acoustic distance from the model talker for phonetic convergence to be observed. In addition, the lack of social stereotyping of the Northern dialect and its vowel variants led to convergence for all four vowels, as expected based on previous work [10, 11, 13, 14, 15]. These results therefore provide further evidence for cross-dialect phonetic convergence when social stereotypes are not present to block it.

The stronger perception of phonetic convergence in the primed condition than in the unprimed condition suggests that providing shadowers with accurate information about the model talker's region of origin can facilitate phonetic convergence. That is, simply being told where the model talker is from, without any further instruction to imitate her, led to more robust phonetic convergence than when no information about the model talker was provided. This result parallels previous findings that accurate social information can facilitate speech processing [1, 3, 5] and extends these findings to include phonetic convergence.

However, this result critically differs from a parallel study involving a Southern American English model talker, for whom phonetic convergence did not differ in unprimed and primed conditions [14]. Given that Southern American English is negatively stereotyped in the U.S. [22], this difference across studies suggests that the blocking of phonetic convergence due to negative social stereotypes may outweigh any priming effects arising from providing social information about the model talker. That is, the negative effects on phonetic convergence of social stereotypes related to Southern American English in the previous study appear to have been stronger than the positive effects of social priming that we observed for the Northern American English model talker in the current study. As a result, no positive effects of social priming were observed in the earlier study with a Southern American English model talker.

Taken together, the results of this study suggest robust cross-dialect phonetic convergence to the non-stereotyped Northern dialect of American English, consistent with previous work on cross-dialect phonetic convergence [10, 11, 13, 14, 15]. In addition, the results provide novel evidence for enhancement of convergence when the model talker's region of origin is explicitly provided. This effect of social priming on phonetic convergence advances our understanding of the role of non-linguistic social information in speech processing [1, 2, 3, 4, 5, 6, 7]. In particular, the results of the current study suggest that social priming extends beyond speech perception to include speech production, consistent with a very tight connection between speech perception and production processes [8, 23].

5. ACKNOWLEDGMENTS

This research was conducted at the Buckeye Language Network Language Sciences Research Lab at the Center of Science Industry in Columbus, Ohio, USA. We would like to thank Laura Beebe, Emily Behm, Sarah Chilson, Ben Elsbrock, Chance Goodwin, Conor Higgins, Rachel Monnin, and McKenna Reeher for assistance with data collection and analysis.

6. REFERENCES

[1] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137, 2823–2833.

[2] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48, 865–892.

[3] McGowan, K. B. 2015. Social expectation improves speech perception in noise. *Lang. Speech* 58, 502–521.

[4] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *J. Lang. Social Psychol.* 18, 62–85.

[5] Rubin, D. L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Res. High. Educ*. 33, 511–531.

[6] Staum Casasanto, L. 2008. Does social information influence sentence processing? In: *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 799–804.

[7] Strand, E. A. 1999. Uncovering the role of gender stereotypes in speech perception. *J. Lang. Social Psychol.* 18, 86–99.

[8] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251–279.

[9] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Percept. Psychophys.* 66, 422–429.

[10] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon.* 40, 177–189.

[11] Walker, A., Campbell-Kibler, K. 2015. Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Front. Psychol.* 6(546), 1–18.

[12] Yu, A. C. L., Abrego-Collier, C., Sonderegger, M. 2013. Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and 'autistic' traits. *PLoS One* 8(e74746), 1–13.

[13] Babel, M. 2010. Dialect divergence and convergence in New Zealand. *Lang. Soc*. 39, 437–456.

[14] Clopper, C. G., Dossey, E. 2020. Phonetic convergence to Southern American English: Acoustics and perception. *J. Acoust. Soc. Am.* 147, 671–683.

[15] Ross, J. P., Lilley, K. D., Clopper, C. G., Pardo, J. S., Levi, S. V. 2021. Effects of dialect-specific features and familiarity on cross-dialect phonetic convergence. *J. Phon.* 86(101041), 1–23.

[16] Labov, W., Ash, S., Boberg, C. 2006. *Atlas of North American English*. Mouton de Gruyter.

[17] Campbell-Kibler, K. 2012. Contestation and enregisterment in Ohio's imagined dialects. *J. Eng. Linguist.* 40, 281–305.

[18] Pardo, J. S. 2013. Measuring phonetic convergence in speech production. *Front. Psychol*. 4(559), 1–5.

[19] Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E. 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *J. Mem. Lang.* 69, 183–195.

[20] Clopper, C. G., Pisoni, D. B., de Jong, K. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. *J. Acoust. Soc. Am*. 118, 1661–1676.

[21] Clopper, C. G., Carter, A. K., Dillon, C. M., Hernandez, L. R., Pisoni, D. B., Clarke, C. M., Harnsberger, J. D., Herman, R. 2002. The Indiana Speech Project: An overview of the development of a multi-talker multi-dialect speech corpus. In: *Research on Spoken Language Processing Progress Report No. 25*. Speech Research Laboratory, Indiana University, 367–380.

[22] Niedzielski, N., Preston, D. R. 2003. *Folk Linguistics*. Mouton de Gruyter.

[23] Pickering, M. J., Garrod, S. 2013. An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347.

Quantifying nasalance in Korean *aegyo*

Drew Crosby[1] & Amanda Dalola[2]

[1]University of South Carolina, [2]University of Minnesota
dmcrosby@email.sc.edu, adalola@umn.edu

ABSTRACT

Nasality has historically been a difficult quantity to measure in linguistic studies because it has required expensive and cumbersome equipment. A recent study in language documentation [18] proposed using earbuds to separately measure nasal and oral air intensity for descriptive purposes.

This study extends [18]'s earbud methodology to the field of sociophonetics by using it to measure and operationalize nasalance in the Korean baby talk register, *aegyo*, as a continuous variable. Previous research reports aegyo as associated with the speech of young women, IP-final nasality and following epenthetic nasal consonants. The present study corroborates both structural and social findings: surrounding nasal consonants resulted in a higher degree of nasalance, high vowels had more nasalance than low vowels, and nasalance was greater in the *aegyo* condition particularly for young women. This suggests that both [18]'s earbud methodology and a continuous operationalization of nasalance are meaningful metrics in sociophonetic inquiry.

Keywords: Aegyo, Earbuds, Nasalance, Nasality, Sociophonetics

1. INTRODUCTION

Measuring nasality has traditionally required cumbersome and expensive equipment such as nasometers. Recently, [18] developed an inexpensive, field-friendly and non-invasive method that uses a pair of earbuds to measure the intensity of the airflow separately in the nasal and oral passageways and allows for the calculation of a nasalance score derived from these measures, as shown in (1).

(1) $100 * \frac{A_{nasal}}{A_{nasal} + A_{oral}}$

A = intensity in decibels

[18, p. 59]

This method has so far been applied in descriptive linguistics and linguistic documentation in order to determine the presence of phonemically nasal segments in phonological inventories. However, nasality as a subphonemic parameter also represents an important dimension in sociolinguistics, and [18]'s methodology has the potential to inform our understanding of the role it plays in linguistic variation, particularly in the creation of linguistic styles. One area where nasality plays a clear role is in the babytalk registers of Korean, Mandarin, and Japanese [4, 6, 8, 13, 15, 17]. However, due to the various difficulties of measuring nasality in the field, this dimension of these styles has thus far been neglected.

In order to explore the applicability of this method to sociophonetics, specifically to a continuous operationalization of nasality in linguistic styles, we tested [18]'s methodology by applying it to the Korean baby talk register of *aegyo*. *Aegyo* is a particularly apt style to test this on, given that its nasal component also has a name, *kossori* (lit. 'nose sounds') [13]. *Kossori* can manifest as either increased nasalization of the intonation phrase (IP)-final vowel or by the addition of an epenthetic nasal consonant to said vowel. It can also be represented orthographically by the addition of an ***ieung, ㅇ,*** the Korean symbol for a /ŋ/, as in (2), with the standard version given in (3) [6, 8, 15]. The Korean writing system, Hangul, is an alphabet that is organized into syllable blocks. Thus the relevant character here changes from 자, /ʨa/, to 장, /ʨaŋ/, with the character representing /ŋ/, ㅇ, appearing below the onset vowel grouping.

(2) *Aegyo*-style Script
자기야 잘 장~~~
/ʨakija ʨal **ʨaŋ**/

(3) Standard Script
자기야 잘 자.
/ʨakija ʨal **ʨa**/
'Honey good night.'

[15, p.13]

2. METHODS

2.1. Participants

A total of 114 Korean native speakers were interviewed in their heterosexual romantic couple. Birth year ranged from 1945 to 1999. The age distribution by number of participants per decade is

as follows: 1940s: 3; 1950s: 20; 1960s: 32; 1970s: 17; 1980s: 25; 1990s: 17 (MEAN = 1972.1). Region of origin varied but most participants were from the Seoul Capital Area (Capital Area: 68, *Gyeongsang-do*: 27; *Jeolla-do*; *Chungcheong-do*: 6, *Gangwon-do*: 3).

2.2. Stimuli

The stimuli consisted of three short dialogues (dinner invitation, complaining about work, vacation planning), and three open-ended scenarios in which couples were asked to roleplay with one another (requesting, comforting, expressing love).

2.3. Procedure

Recording took place in a private room (a rented room in a study café, a local church, the participants' home, or the researcher's parents-in-law's home) with the researcher present. Recording was done on a pair of Zoom H4n Pros at 44.1 kHz (one recorder per participant) and a pair of low-resistance (32 Ω) earbuds with silicone attachments (Joha JE01) used as microphones. One earbud was placed in the corner of the mouth and one was placed under a nostril to separately record the oral and nasal tracts. A JLab TALK GO Plug & Play USB Microphone was also attached to the laptop used to present the data and used to record the full interaction for use in other research.

The participants were first requested to read the dialogues and then switch roles and do it again. They were then asked to do the roleplays with each partner performing each role once. They were asked to do both of these while speaking "as they usually do." They were then asked to repeat the procedure again while performing *aegyo*.

3. ANALYSIS

3.1. Measurements

The purported location of nasality in *aegyo* is the intonation-phrase (IP)-final syllable (Moon 2013). Accordingly, the participants' speech signals were divided into IPs following K-ToBI [7]. The IPs were then extracted with 100ms on either side to facilitate the calculation of intensity at the edge of the IP, and the amplitude of both the oral and the nasal tracts were normalized to 0db using Adobe Audition (Build 14.4.0.38), following [18]. Vowel duration and intensity values for each tract were then extracted across the IP-final vowel (and any epenthetic nasal consonant following it) using a Praat script and nasalance calculated following the above equation.

3.2. Statistics

Statistical analyses of nasalance were conducted in R [16]. A mixed effects linear regression was calculated using the *lmer()* function of the lmerTest package [9], with nasalance as the dependent variable and the interaction of *preceding segment type* (oral consonant, vocoid, nasal consonant), *vowel duration* (continuous in seconds), *presence of following epenthetic nasal consonant* (yes, no), and *vowel height* (low, mid, high) as independent variables. Additionally, it was suspected that sounds in the 100ms surrounding the IP might influence the amplitude normalization if their amplitude was higher than any sound in the IP itself, so two additional independent variables addressing noise on the edge of each track (i.e., noise from the vocal tract that occurred in the 100ms on the edge of each sound file that was outside of the IP and was larger in amplitude than any sound in the IP itself) were added: *presence of oral-edge noise* (yes, no) and *presence of nasal-edge noise* (yes, no). Additionally, to examine the role of social parameters, we included the three-way interaction between (*performance of) aegyo* (yes, no), continuous *age* (calculated using the Korean system, 2022 - year of birth +1), and *gender (women/men)*. Random intercepts for *participant* and *lemma* were also included. Graphics were generated using the ggplot2 [19] and sjplot [12] packages of R.

4. RESULTS AND DISCUSSION

Table1 displays the computed statistical model.

Factor	Estimate	*p*-value
Intercept	4.833	<0.0001****
Preceding vocoid	1.277	<0.0001****
Preceding nasal cons.	3.075	<0.0001****
Vowel duration	1.188	<0.0001****
Mid vowel	0.585	<0.0001****
High vowel	1.540	<0.0001****
Epenthetic Nasal	2.709	<0.0001****
Aegyo	0.512	<0.0001****
Oral-track Noise	-0.819	<0.0001****
Nasal-track Noise	0.978	<0.0001****
Age	-0.000884	0.933
Men	0.538	0.481
Preceding vocoid * Vowel duration	-0.696	0.022*
Preceding nasal cons. * Vowel duration	-2.015	<0.0001****
Aegyo * Age	-0.0150	<0.0001****
Aegyo * Men	-1.165	<0.0001****
Age * Men	-0.00460	0.752
Aegyo * Age * Men	-0.0197	0.000128***

Table 1: Nasalance linear regression model (nasalance ~ *preceding segment type* * *vowel duration* + *epenthetic nasal consonant* + *vowel height* + *Oral-track noise* + *Nasal track noise* + *aegyo* * *age* * *gender* + (1|*participant*) + (1|*lemma*).

The model revealed a main effect for *preceding segment type* such that IP-final vowels preceded by a vocoid had a nasalance measure that was 1.277% higher than when preceded by an oral consonant (p<0.0001****). Additionally, IP-final vowels preceded by a nasal consonant had a nasalance measure that was 3.075% higher than when preceded by an oral consonant (p<0.0001****). There was also a main effect for *vowel duration* such that for every one second of increase in vowel duration there was a 1.188% increase in nasalance (p<0.0001****).

Figure 1 displays the interaction between *preceding segment type* and *vowel duration*. Here we see the longer a vowel is, the greater its nasalance when preceded by an oral consonant or vocoid. However, when preceded by a nasal consonant the opposite is true: the longer a vowel is, the lower its nasalance.

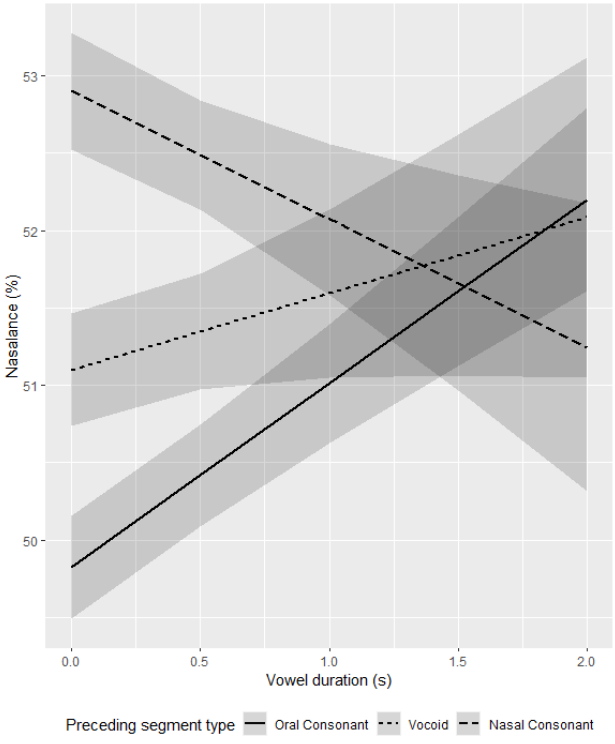


Figure 1: Predicted nasalance by preceding segment type and vowel duration. Shaded areas = 95% confidence interval.

Additionally, the presence of an epenthetic nasal consonant resulted in a 2.709% greater degree of nasalance across the IP-final vowel, as shown in Figure 2 (p<0.0001****).

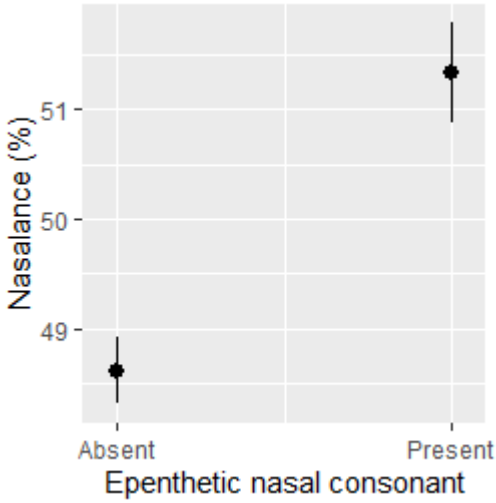


Figure 2: Predicted nasalance by presence of an epenthetic nasal consonant. Bars = 95% confidence interval.

The model also revealed a main effect for vowel height such that mid vowels had a 0.585% greater nasalance (p<0.0001****) and high vowels a 1.540% greater nasalance (p<0.0001****) than the reference low vowels, as illustrated in Figure 3.

Oral-track noise and *nasal-track noise* were also significant. When high amplitude noise was present on the edge of the oral track, the IP-final vowel was predicted to have a 0.512% greater degree of nasalance than when noise was not present (p<0.0001****). Meanwhile, when high amplitude noise was present on the edge of the nasal track, the IP-final vowel was predicted to have a 0.819% lower degree of nasalance than when noise was not present (p<0.0001****).

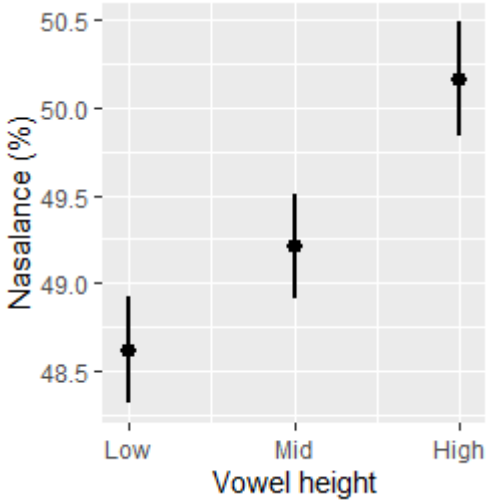


Figure 3: Predicted nasalance by vowel height. Bars = 95% confidence interval.

The findings from the structural factors suggest that the use of [18]'s earbud methodology and the continuous operationalization of nasalance are meaningful metrics in sociophonetic inquiry. The

results for *preceding segment type* and *presence of an epenthetic nasal consonant* show that the coarticulation of a nasal consonant with a vowel on either side is reflected in measures of nasalance; the fact that nasalance decreases as a vowel gets longer when preceded by a nasal consonant also shows that this coarticulation gets smaller the further you get from the nasal-consonant vowel boundary. Additionally, low vowels have also been shown to have lower degrees of nasalance than high vowels [3, 11]. However, /u/ has been shown to have lower nasalance than /a/ in Korean [5, 10], and contrary to [11], [2]'s data shows /u/ having similar nasalance to /a/ in English. The authors suggest this is due to the tongue's proximity to the velum in /u/ which dampens velar vibrations and thus the transfer of energy between the oral and nasal tracts. However this should not affect the present data much: /i/ comprised 76.042% of high vowels; /u/ 22.541%, and /ɯ/ 1.416%. This finding also makes phonetic sense, given that low vowels are more sonorous than mid and high vowels, due to their larger aperture for airflow, while the nasal aperture remains relatively constant regardless of vowel height. However, given that the sounds surrounding the IP do have an effect on nasalance values, these should probably not be directly compared to values measured from a nasometer which span a much larger domain than a single IP.

The model also revealed a main effect for *performance of aegyo* such that nasalance was 0.978% greater when *aegyo* was being performed than when it was not (p<0.0001****).

Figure 4 displays the three-way interaction between *performance of aegyo*, *age*, and *gender* (p=0.000128***). Here we see that nasalance decreases sharply the older a woman is when she is performing *aegyo*, but there is little effect for age when she is not. Men also show a slight inverse correlation between age and nasalance but it is actually stronger when they are not performing *aegyo*.

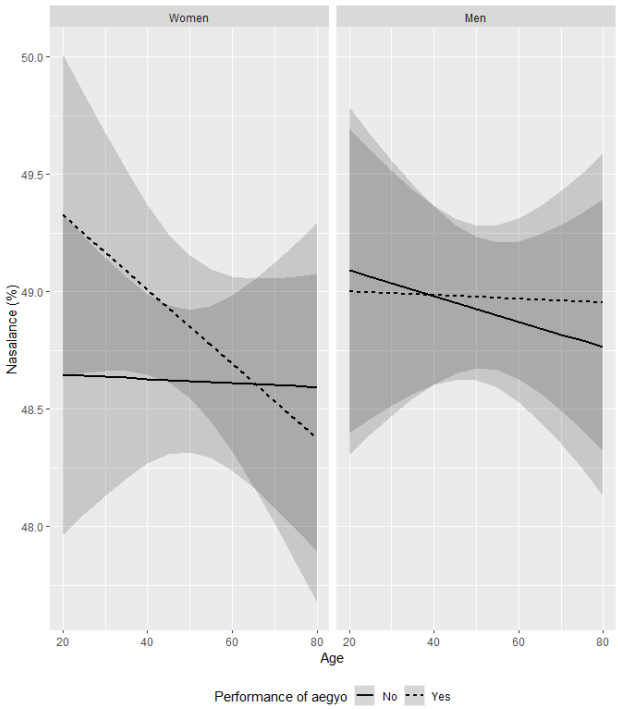


Figure 4: Predicted nasalance by *performance of aegyo*, *age*, and *gender*. Shaded areas = 95% confidence interval.

Like the structural findings, the results from the social factors also argue for the validity of [18]'s earbud methodology and the continuous operationalization of nasalance in variationist studies. *Aegyo*'s association with nasality and with young women is clearly reflected in the nasalance values. We can also see *aegyo*'s effect in the interaction between vowel duration and *preceding segment type*. IP-final vowel lengthening is also associated with *aegyo* [14], so we would expect nasalance and vowel length to increase in tandem when *aegyo* is also a factor, particularly when there is no nasal consonant to influence nasalance values, as suggested by Figure 1.

5. CONCLUSION

The results of this study have shown that continuous nasalance measures obtained using [18]'s earbud methodology follow expected directions in terms of coarticulation with nasal consonants, vowel height, vowel length, style, age and gender. This strongly suggests that this methodology and operationalization can be meaningfully applied to variationist research, opening up the possibility of the exploration of a new dimension of sociophonetic phenomena.

Because most of the nasalance values hovered around 50%, future work should explore alternative normalization methods which may yield more robust results, particularly when measuring more subtle more nasal phenomena.

6. REFERENCES

[1] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.1.53, http://www.praat.org/.
[2] Gildersleeve-Neumann, C. E., & Dalston, R. M. 2001. Nasalance scores in noncleft individuals: why not zero?. T*he Cleft palate-craniofacial journal* 38(2), 106-111.
[3] Ha, S., & Cho, S. H. 2015. Nasalance scores for normal Korean-speaking adults and children: effects of age, vowel context, and stimulus length. *International journal of pediatric otorhinolaryngology* 79(8), 1235-1239.
[4] Hardeman, K. 2013. Gender and second language style: American learner perceptions and use of Mandarin sajiao. *University of Hawaii at Mānoa dissertation*.
[5] Hwang, Y. J., Kim, H. K., Jeong, O. R., Lee, J. H. 2007. The average nasalance as a function of Korean and Chinese in vowels. *Speech Sciences* 14(3), 139-146.
[6] Jang, H. 2021. How cute do I sound to you?: Gender and age effects in the use and evaluation of Korean baby-talk register, Aegyo. *Language Sciences*, 83, 1-18.
[7] Jun, S. A. 2000. K-tobi (korean tobi) labelling conventions. *Speech Sciences* 7(1), 143-170.
[8] Kim, J. 2020. Phonetics and Phonology of Speech Styles in Korean. *Seoul National University dissertation.*
[9] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2015. Package 'lmertest'. *R package version* 2(0), 734.
[10] Lee, S. E., Hong, J. H., Choi, H. S. 2002. The Effects of Vocal Loudness on Nasalance Scores in Korean Vowels-A Preliminary Study Using NasalView System. *Speech Sciences* 9(4), 49-58.
[11] Lewis, K. E., Watterson, T., Quint, T. 2000. The effect of vowels on nasalance scores. *The Cleft palate-craniofacial journal* 37(6), 584-589.
[12] Lüdecke D (2022). sjPlot: *Data Visualization for Statistics in Social Science*. R package version 2.8.12, https://CRAN.R-project.org/package=sjPlot.
[13] Miller, L.2004. You are doing burikko!: Censoring/scrutinizing artificers of cute femininity in Japanese. In: Okamoto S., Shibamoto Smith J. (eds.) *Japanese language, gender, and ideology: Cultural models and real people*. Oxford University Press, 148-165.
[14] Moon, K. 2017. Phrase final position as a site of social meaning: Phonetic variation among young Seoul women. *Stanford University dissertation*.
[15] Moon, K. 2013. Phrase final position as a site of social meaning: Phonetic variation among young Seoul women *Stanford University unpublished master's thesis*.
[16] R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
[17] Starr, R. L., Wang, T., Go, C. (2020). Sexuality vs. sensuality: The multimodal construction of affective stance in Chinese ASMR performances. *Journal of Sociolinguistics* 24(4), 492-513.
[18] Stewart, J., & Kohlberger, M. 2017. Earbuds: A method for analyzing nasality in the field. L*anguage Documentation & Conservation* 11, 49-80.
[19] Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

APPARENT TIME CHANGE IN THE ARTICULATION OF ONSET RHOTICS IN SOUTHERN BRITISH ENGLISH

Patrycja Strycharczuk[1], Susan Lloyd[2] & James M. Scobbie[2]

[1]University of Manchester; [2]Queen Margaret University, Edinburgh
patrycja.strycharczuk@manchester.ac.uk; SLloyd@qmu.ac.uk; JScobbie@qmu.ac.uk

ABSTRACT

The /r/ phoneme in Anglo-English is known to correspond to a number of relatively distinct articulatory variants. However, little is known about the social structure of this variation. In this study, we investigate the effect of two social factors, age and gender, on the production of /r/, in a sample of 36 speakers from the South of England. We analysed ultrasound images of pre-vocalic /r/ tokens. We measured the distances between the short tendon and 11 points on the tongue surface. We compared these distances across speakers in representative /r/ frames. We find an apparent time difference whereby the distance between the tongue tip and the short tendon reduced in apparent time, potentially signalling an ongoing sound change from a tip-up to tip-down /r/.

Keywords: rhotics; variation; Southern British English; ultrasound; change

1. INTRODUCTION

Southern British English (SBE) rhotics are well-studied, but mostly from the point of view of whether they occur or not in specific contexts. In contrast, the quality of the /r/ in this variety has received less empirical attention. It is commonly described as a post-alveolar approximant [1, 2], although other variants, notably labio-dental, have also been documented [3, 4, 5].

An early articulatory study of /r/ in Anglo-English shows pronunciation by two speakers to be a retroflex approximant [ɻ] in onset position, and a vowel otherwise [6]. A recent study combining ultrasound and a video of the lips documents the articulation of /r/ in a geographically diverse sample of 24 speakers from England [7]. This study captures considerable articulatory variation, spanning three variants of tip-up /r/, two of which are types of voiced post-alveolar approximant [ɹ], while the other is a retroflex approximant [ɻ], as well as two types of tip-down bunched /r/. These variants resemble those found in other varieties of English, including Scottish English [8, 9], American English [10], and New Zealand English [11]. Bunched /r/ does not appear in most earlier descriptions of Anglo-English, even as an coarticulatory variant in supportive environments. As far as constraints on rhotic variation are concerned, tip-up /r/ is generally less likely to occur preceding front vowels than preceding back vowels. Otherwise, the variation is not known to be systematically structured.

In this study, we explore whether rhotic variation in SBE, a dialect of Anglo-Eglish, is systematically conditioned by social factors, specifically age and gender. Investigating these factors from a variationist point of view requires a relatively large sample, which can be challenging in the context of an articulatory study. However, articulatory data can be crucial in case of rhotics, given acoustic similarities between multiple articulatory /r/ variants. In this project, we used data collected at a scientific outreach event hosted by the British Academy in London in June 2019, featuring a live ultrasound demonstration (see [11] for a similar approach). This event was attended by many visitors local to the area, making the sample geographically homogeneous, and there was a relatively good balance of age and gender, which allowed us to explore the role of these factors in conditioning /r/-variation.

2. METHOD

We recorded the data while delivering a live ultrasound demonstration at a scientific outreach event. Visitors had the opportunity to try ultrasound tongue imaging (UTI), and they were invited to participate in the study. 55 visitors agreed to participate, 36 of whom were included in the current study. The inclusion criteria were having being born in and grown up in South of England. 18 participants described their gender as male (mean age = 38.4, SD =14.0), and 18 described it as female (mean age = 36.1, SD =17.7).

The participants read the following sentences: *That Mary Rivers! She's so overbearing. I'd feel better if I never saw her again. If you're not feeling well, a nice hot cup of tea can help you bear it. Try*

it out and see how you feel.

The sentences were designed such that most (though not all) /r/ tokens occurred in a stable segmental environment, following a mid vowel and preceding a high vowel. The prosodic and morphological context was varied, including a word-initial onset (*Rivers*), a word-medial onset (*Mary*), morpheme-final /r/ (*overbearing*) and linking /r/ (*bear it*). We also aimed to elicit prosodically natural speech with these stimuli.

One of the experimenters held the ultrasound probe under the participant's chin along the midsagittal line. No head stabilisation was used. A lapel condenser microphone by AudioTechnica was used to capture the accompanying audio. Audio signal and the ultrasound image were recorded and synchronised using Articulate Assistant Advanced (AAA) v.2.18 [12]. The ultrasound system was EchoB, with a 2-4MHz probe. The frame rate was 60 frames per second, the depth was 75 mm and the number of scanlines was 127.

We first classified a selection of tongue images (at least one from each speaker) qualitatively, in order to identify the main types of rhotics in the data. We followed the classification system by [7]. The classification was done collectively by the three authors, who discussed the shapes and arrived at a consensus classification.

We then traced the tongue contour in the ultrasound image using Deep Lab Cut (DLC) [13]. This algorithm is trained on manually selected labels intended to pick out consistent flesh points in a midsagittal sequence, either with direct reference to articulatory features evidenced in the image (not limited to surface edge features), or through interpolation. The DLC process is trained therefore to assign the same labels to novel data. These tongue surface labelled points are located by DLC as part of a surface contour (a 2D spline) based on all the data in the image in reference to the training set provided by AAA (v.2.20). This surface contour can be exported as a series of Cartesian or polar points relative to an origin (e.g. the location of the short tendon).

Using a method based on [13], we measured the distance between the short tendon and the tongue surface along 11 points on the surface of the tongue. Approximate location of the points is illustrated in Figure 1. The radials connecting the points on the tongue surface and the short tendon form a fan, similar to a standard approach in ultrasound methodology. The distances between the short tendon and the tongue surface form a normalised tongue shape that can be compared across speakers.

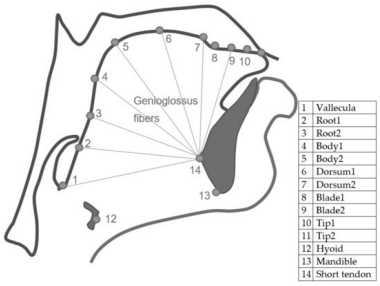


Figure 1: The reference points used in the analysis. Figure reproduced from [13] with permission.

We extracted the normalised tongue shapes for a single ultrasound frame in each /r/-token, at the time point when the tongue was judged to reach the maximum /r/-constriction. These representative UTI frames were selected by the first author, based on visual inspection of the ultrasound videos.

We fitted a series of GAMMs (Generalised Additive Mixed Models) modelling normalised tongue shapes, by predicting the distance from the tongue surface to the short tendon, depending on a by-fanline smooth. A stepwise model comparison procedure was used to evaluate the following predictors:

- main effect of gender;
- a smooth term for fanline by gender;
- a smooth term for participant age;
- tensor product interaction between the fanline and participant age;
- tensor product interaction between the fanline and participant age by gender;
- main effect of *r*-token;
- a smooth term for fanline by *r*-token.

The number of knots for each smooth was set to 10. All models included a tp by-participant random smooth for fanline with 10 basis functions [14]. Significance of the individual predictors was established through Maximum Likelihood comparisons between nested models. We selected the best model using a backward comparison procedure. We then corrected for autocorrelation by fitting an AR1 version of the selected model [15].

3. RESULTS

We found instances of the following /r/-types reported by [7]: front-up, tip-up and curled-up, as well as front-bunched /r/. We also found some tokens tentatively classified as mid-bunched, although a more appropriate label would perhaps

be 'weakly bunched'. Representative examples of all types are illustrated in Figure 2. The left column of the figure shows tongue outlines in Cartesian coordinates. The right column shows the corresponding normalised tongue shape, i.e. the distance from the short tendon to the tongue surface, depending on fanline. All these examples are from word-initial /r/ (*Rivers*). 13 out of 36 participants (36%) had tip-down /r/ in this position.

The distance measure used to produce normalised tongue shapes highlights that front-up, tip-up and mid-bunched /r/ are characterised by various degrees of convexity in the tongue surface, but they differ by the location of the constriction, which is more anterior for mid-bunched /r/. Additionally, tongue tip is down in mid-bunched /r/, which is reflected in the reduced distance at fanline 11. Curled-up /r/ and front-bunched /r/ both have a degree of tongue body concavity, which results from a simultaneous tongue root and dorsal constriction. The tip is notably lower in front-bunched /r/, compared to curled-up /r/.

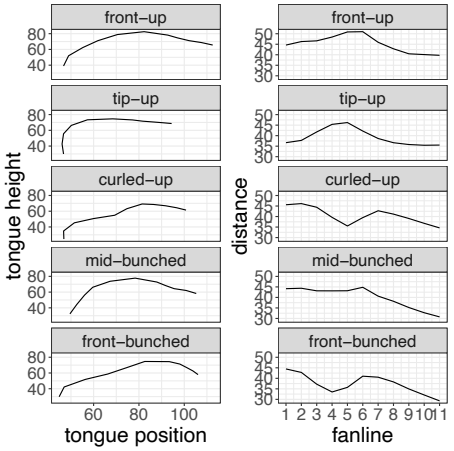


Figure 2: Left column: tracings of the tongue contour for word-initial /r/, as pronounced by three selected participants. Tongue tip is on the right. Right column: distance between the short tendon and the 11 reference points on the tongue surface for the same /r/-tokens.

Based on model comparison, normalised tongue shape was significantly affected by speaker age ($p<.05$) and by the /r/ token ($p<.001$), but not by speaker gender ($p= .07$). Gender had a significant main effect on distance ($p<.001$), whereby male speakers had overall greater distances compared to females, consistent with males having on average larger vocal tracts.

Figure 3 illustrates the model predictions for normalised tongue shape in word-initial /r/ (*Rivers*), depending on the fanline and speaker age. The figure shows predictions for female speakers (recall that males are predicted to have greater distance values but not significantly different shape). The mean smooths for various age groups separated by decade would appear to show a continuous progression from a more convex to a more concave tongue body, associated with a secondary dorsal constriction in speakers of ca. 50 years of age and younger. However, the only area of significant difference is fanline 11 (tongue tip), which suggests that variation in the tongue body is somewhat idiosyncratic. The tongue tip, on the other hand, shows significant lowering in apparent time, with youngest speakers being more likely to have a tip-down /r/.

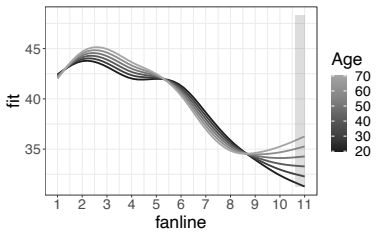


Figure 3: GAMM predictions for mean normalised tongue shape, depending on speaker age. Predictions are for a word-initial /r/ (*Rivers*) produced by a female speaker. Shading indicates areas of significant difference.

In terms of token effects, we find systematic differences in the tongue root and degree of dorsal constriction. Word-initial /r/ (*Rivers*) had more advanced tongue root compared to morpheme-medial /r/ (*Mary*), which in turn had more advanced tongue root and increased dorsal constriction, compared to morpheme-final one (*overbearing*) or linking /r/ (*bear it*). Morpheme final /r/ had a higher tongue blade, compared to linking /r/. While some systematic trends seem to emerge, they might be partially due to long-distance co-articulation. The rhotics in *her again* and *try* were markedly different, presumably due to the segmental context, and especially affrication in the latter. Otherwise, token-conditioned differences we find are relatively subtle. While some speakers showed forms of /r/-allophony, only few used clearly different /r/-shapes in different contexts. Typically, speakers had a single dominating /r/-shape, although we acknowledge that we tested a limited set of possible /r/ environments.

4. DISCUSSION

In some aspects, our results are consistent with previous studies on /r/-variation, crucially [7]. We

find a mixture of various /r/-shapes, and an overall preference for tip-up /r/. A third of the speakers in our sample (36%) had a tip-down /r/ in their repertoire, compared to 42% in [7]. However, our results show an apparent time difference, suggesting an ongoing sound change from a tip-up to tip-down /r/ in the South of England. While, there is variation in tongue shape across all ages, younger speakers (under 25) are more likely to have a tip-down /r/ than a tip-up one.

From a certain point of view, this might seem surprising, as tip-up /r/ is generally the preferred variant in non-rhotic varieties of English [7, 11]. In general, tip-down /r/ is more likely in postvocalic positions [6, 10], so speakers who do not produce coda rhotics, are thought to be less likely to acquire bunching as a rhotic articulation [11]. We might therefore ask what could be driving a potential sound change towards tip-down /r/ in a non-rhotic variety.

We believe that a factor potentially contributing to this change is lip posture. Anglo-English /r/ has a prominent labial component, which has been shown to co-vary with tongue shape: tip-down /r/ is typically produced with more lip protrusion, compared to tip-up /r/, as shown by [7], who propose that lip protrusion is a compensatory strategy that allows speakers to lower the F3 despite the relatively small sublingual cavity produced by bunching the tongue (compared to retroflex). This explanation is consistent with their finding that various /r/ shapes are not associated with significant differences in the third formant, a key acoustic correlate of rhoticity. A further study has shown that Anglo-English listeners are highly sensitive to the visual cues associated with labialisation /r/ [16]. Visual information alone was sufficient for participants to disambiguate /r/ and /w/ minimal pairs, whereas an auditory classification task showed some confusability, with a general bias for /r/ perception. This was attributed to the presence of labio-dental [ʋ] as a frequent variant of /r/ in Anglo-English, which is an /r/-allophone with a relatively high F3.

Given that labialisation is a distinct feature of tip-down /r/, and also a prominent perceptual cue of rhoticity in Anglo-English, the labial gesture likely contributes to the propagation of tip-down /r/ in England [16]. A question that arises in this context is whether the shift towards tip-down /r/ we observed is a part of ongoing increase of labio-dental [ʋ]. This interpretation would be consistent with the presence of some weakly bunched /r/ tokens in our data, and also with the fact that we do not find a significant apparent time difference affecting tongue body. While we see an apparent time increase in tip-down /r/, we do not find an accompanying shift to a concave tongue body, which suggests that the emerging tip-down /r/ variants are not uniformly front-bunched. However, this proposal is only tentative, since we are not in a position to support it with acoustic or auditory evidence. We collected the data in a public space with significant background noise present, and as a result the audio data we have do not lend themselves to either spectral or auditory analysis.

Nevertheless, our results provide a compelling case for future studies of articulatory /r/-variation in South of England, also highlighting the need for stratifying the speaker sample for age. The current findings are among few that show systematic social structure in the variation between tip-up and tip-down /r/ in an accent of English. This type of variation is known to be conditioned by linguistic factors such as syllable position and segmental environment [6, 7, 11, 10], as well as by the variety of English itself, but otherwise, it can be idiosyncratic: some speakers have a bunched /r/ and some do not. The somewhat free nature of this variation is sometimes attributed to the acoustic similarity of retroflex and bunched /r/, especially at the frequencies typically salient for phonological contrasts [17]. A notable exception, however, is Scottish English, where /r/-variation is systematically conditioned by social class [8, 9].

Future articulatory studies may benefit from the approach we used to normalise tongue shape, originally proposed by [13]. A key aspect of the methodology is the use of a fan that uses an anatomically informed original point. Thanks to this, the distance measures can be interpreted in terms of articulatory constrictions, and they can be compared systematically across different speakers. Furthermore, by being fitted separately to each frame based on identifiable landmarks, the fan partially corrects for probe rotation and probe movement. While this approach is highly promising, it also raises questions concerning the level of systematicity one can expect when aggregating articulatory data. For example, we note some differences in the exact location of constriction for different types of /r/-variants (Figure 2). It is not clear whether such differences systematically distinguish /r/-variants, which would make them valuable for automatic classification of /r/-shape, or whether they are partially dictated by anatomical differences. Future models may improve on this approach by predicting tongue shape based on functional muscle bundle control.

5. ACKNOWLEDGMENTS

We thank our participants for contributing to our study. We also thank the British Academy for inviting us to participate in the Summer Showcase 2019, and for their support for our data collection during the exhibit.

6. REFERENCES

[1] P. Carley, I. M. Mees, and B. Collins, *English phonetics and pronunciation practice*. Abingdon, UK/New York, NY: Routledge., 2017.

[2] R. Ogden, *An introduction to English phonetics*. Edinburgh University Press, 2009.

[3] J. Wells, *Accents of English. 3 vols*. Cambridge: Cambridge University Press, 1982.

[4] P. Foulkes and G. J. Docherty, "Another chapter in the story of /r/: 'Labiodental' variants in British English," *Journal of Sociolinguistics*, vol. 4, pp. 30–59, 2000.

[5] E. Lawson, J. Stuart-Smith, J. M. Scobbie, and S. Nakai, "Dynamic Dialects: An articulatory web resource for the study of accents." [Online]. Available: https://www.dynamicdialects.ac.uk/

[6] P. Delattre and D. C. Freeman, "A dialect study of American *R*'s by X-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[7] H. King and E. Ferragne, "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics*, vol. 80, p. 100978, 2020.

[8] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation," *Journal of Phonetics*, vol. 41, pp. 198–210, 2013.

[9] E. Lawson, J. Stuart-Smith, and J. M. Scobbie, "The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study," *The Journal of the Acoustical Society of America*, vol. 143, pp. 1646–1657, 2018.

[10] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/," *Language*, pp. 101–140, 2016.

[11] M. Heyne, X. Wang, D. Derrick, K. Dorreen, and K. Watson, "The articulation of /ɹ/ in New Zealand English," *Journal of the International Phonetic Association*, pp. 1–23, 2018.

[12] Articulate Instruments Ltd. (2019) Articulate Assistant Advanced ultrasound module user manual, revision 2.18.

[13] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut," *Sensors*, vol. 22, p. 1133, 2022.

[14] M. Soskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, p. 101017, 2021.

[15] R. H. Baayen, J. van Rij, C. de Cat, and S. Wood, "Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models," in *Mixed-effects regression models in linguistics*, D. Speelman, K. Heylen, and D. Geeraerts, Eds. Springer, 2018, pp. 49–69.

[16] H. King and I. Chitoran, "Difficult to hear but easy to see: Audio-visual perception of the /r/-/w/ contrast in Anglo-English," *The Journal of the Acoustical Society of America*, vol. 152, pp. 368–379, 2022.

[17] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/," *The Journal of the Acoustical Society of America*, vol. 123, pp. 4466–4481, 2008.

Sounds Heimish: A sociophonetic study of London's orthodox Jewish community

Keren Rubner[1], Kathleen M. McCarthy[1]
[1]Queen Mary University of London
k.rubner@hss20.qmul.ac.uk, k.mccarthy@qmul.ac.uk

ABSTRACT

This paper presents English, or *Judeo-English,* vowel production patterns by members of the Haredi Jewish community in Stamford Hill, north London, UK. The aim of the paper is twofold: to provide the first description of Judeo-English vowels as spoken in London's community; and secondly, to explore how community-specific gender differences influence vowel production patterns. Members of the community who have lived, studied and worked in Stamford Hill for either their whole lives, or from early childhood were recorded producing /i/, /ɪ/, /ɛ/, /æ/, /ɔ/, ʊ/, /u/ within words. Overall, the vowel production patterns showed little influence from surrounding London English varieties. An acoustic analysis of the first and second formant frequencies found a difference between men and women in their production of some vowels, reflecting their distinct language experiences. These findings highlight speech patterns in an understudied close-knit community, and illustrate factors that can influence phonetic variation in diaspora communities.

Keywords: Judeo-English, phonetic variation, gender, monophthongs

1. INTRODUCTION

The Jewish community of Stamford Hill, north London is the largest *Haredi*, or strictly orthodox Jewish community in Europe [12]. The community grew rapidly after the Second World War, with immigration from Central and Eastern Europe. The community is self-sufficient and very close-knit, with its own schools, businesses and charitable organisations. This has resulted in members having little need to interact with people beyond the metaphysical borders of Stamford Hill. This intra-dependence has also meant that access to the community is difficult, especially so for academic research [2]. While there has been some phonetic research on Judeo-English in the United States [4], to our knowledge there is no published sociophonetic study on the Judeo-English spoken in the Stamford Hill community, London, UK. The aim of the current study was therefore to provide the first descriptive study of monophthong production by adults in the community, and an exploratory analysis of gender differences in vowel production.

Previous research has shown that tight-knit diaspora communities with a dominant heritage language (L1), can give rise to English speech production patterns that reflect the L1 [13, 14, 18]. These production patterns tend be modulated by multiple factors such as home language environment (e.g., intergenerational living), educational experience, and more commonly explored factors such as age of arrival, and similarities and differences between the L1 and L2.

Speakers in the Stamford Hill Haredi community are mostly bilingual. Yiddish is spoken at home and in the community, however, gender plays a unique role in determining the language environment. Specifically, for men, Yiddish is typically their dominant language, whereas for women Judeo-English is the language of study and work [2]. During childhood, girls and boys only socialise within the nuclear family unit and are educated separately. Boys are mainly taught in Yiddish and read Hebrew and Aramaic texts, completing formal secular education at the age of 14. Girls are taught secular subjects e.g., mathematics, in English and sit their GCSE's and sometimes A-Levels. Women are also traditionally the breadwinners. This means that women in the community are predominately exposed to and use English and Judeo-English from the start of schooling and through their working lives, whereas men are typically primarily exposed to and use Yiddish, in their daily lives from childhood onwards [2].

In the current paper we explore how these community-specific gender differences in language environment influence the Judeo-English vowel production patterns. Men and women in the community were recorded producing seven monophthongs: /i/, /ɪ/, /ɛ/, /æ/, /ɔ/, ʊ/, /u/ within English words. All vowels are present in English and Yiddish, apart from /æ/ which is not present in Yiddish [15]. Based on the distinct language environment differences between men and women in the community, we predicted that men and women would display different English vowel production patterns. Specifically, men might reflect their Yiddish dominant environment, whereas women, who have comparatively more contact outside of the community, will show Judeo-English production patterns that present greater alignment with English

vowels, possibly as spoken in London. A second aim of this study was to compare, descriptively, the community production patterns to previous research on neighbouring London English varieties spoken in the North London Borough of Hackney [6, 9], to see to what extent, if any, the community has been influenced by local production patterns.

2. MATERIALS AND METHODS

This study had to contend with a number of unique factors when recording community members. The research took place during the Covid pandemic. UK Government restrictions at the time allowed the mixing of two households. A further consideration was the community's cultural and religious restrictions, which had to be accounted for when designing the study. Haredi societies generally shun mass media and the internet. Members of the Stamford Hill community do not own smart phones or televisions, nor do they have internet access. This meant that interviews could not be conducted over virtual meeting platforms, such as Zoom. Due to these restrictions, recording devices were delivered to participants' homes with instructions on how to operate the devices. The interviews with the researcher were then carried out over the landline telephone.

2.1. Participants

Nineteen adults (18 – 73 years old, median = 35 years, mean = 37.4 years, 10 women, 19 men) from the North London Stamford Hill Haredi Jewish community were recorded. All speakers lived in the community from birth (n =18) or from a very early age (n=1). Participants were bilingual, speaking Judeo-English and Yiddish. The primary home language of most speakers in this study was Judeo-English. All male participants used Yiddish on a daily basis. Subjects were recruited through the first author's personal contacts and acquaintances.

2.2. Materials and procedure

All recordings took place in the participants' homes using a Zoom H2n recorder with a sampling frequency of 44.1 kHz. The seven target vowels /i/ (FLEECE), /ɪ/ (KIT), /ɛ/ (DRESS), /æ/ (TRAP), /ɔ/ (THOUGHT), ʊ/ (FOOT), /u/ (GOOSE) were elicited within words in CVC stressed position, avoiding words with coda lateral or approximant. Participants read a total of 49 target words in the carrier phrase *say* _____ *again*. Speakers were asked to repeat the list in case of background noise or unclear diction. Recordings from 4 participants could not be used because of excessive background noise, leaving the final analysis with data from 9 women and 6 men.

2.3. Acoustic analysis

A total of 736 tokens were analysed. The first (F1) and second (F2) formants were measured using Praat [5] with F1 and F2 measurements taken from the midpoint of the steady-state part of each vowel. Frequency data was normalised using NORM [8] using the Lobanov transformation [11]. This allowed data to be compared across men and women.

3. RESULTS

We first describe the overall vowel production patterns and the gender comparison. Separate mixed-effect linear regression models were run for F1 and F2. Fixed effects were gender (binary: men, women) and vowel, while participant was included as a random effect. Post-hoc group Tukey adjusted group comparisons were conducted, along with Cohen's *d* effect size. All analyses were conducted in R [17] using the packages lme4 [1], CAR [7], and emmeans [10]. The second section of the results provides a short descriptive comparison with previous research on the neighbouring London English varieties. All vowels are described using Well's Lexical Set [20]

3.1. Gender comparison

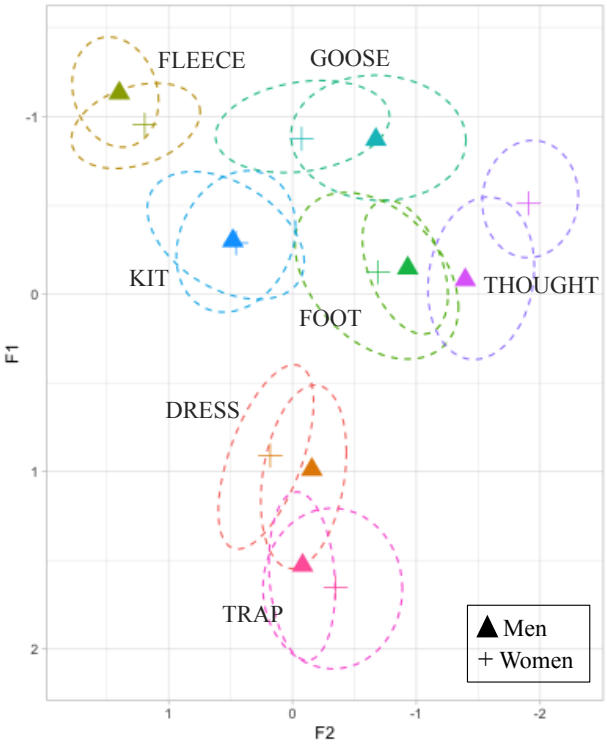


Figure 1. Vowel plot of the mean normalised F1 and F2 frequencies. The ellipses represent 95% confidence intervals.

Observation of Figure 1 shows differences and similarities between the gender groups. The women display more GOOSE-fronting than the men. The DRESS vowel is slightly more fronted for the women than the men, which is more central in the vowel space. The women's THOUGHT vowel is slightly further back in comparison to the men's THOUGHT. Additionally, then men's THOUGHT vowel is also slightly lower in the vowel space than the women. The ellipses of DRESS and TRAP overlap for the men, whereas as for the women these vowels are distinct. The remaining vowels, FLEECE, KIT and TRAP, did not display any marked difference in production between men and women.

The mixed model analysis revealed a significant main effect of vowel for F1 ($\chi^2(6) = 5404.18$, $p < .0001$) and F2 ($\chi^2(6) = 2681.22$, $p < .0001$), and a significant interaction between gender and vowel (F1: $\chi^2(6) = 43.20$, $p < .0001$, F2: $\chi^2(6) = 90.79$, $p < .0001$). Post-hoc group comparisons revealed a significant gender difference for THOUGHT in the F1 ($t(536) = 5.28$, $p < .001$, $d = 1.26$) and F2 domain ($t(536) = 4.72$, $p < .0001$, $d = 1.12$), with the women's production having a higher F1 (closer) and higher F2 (further back) than the men. There was also a significant gender difference in F2 for DRESS ($t(399) = -3.66$, $p < .01$, $d = -0.74$), and GOOSE ($t(536) = -5.53$, $p < .001$, $d = -1.32$). For DRESS, women had a higher F1 (more front) than men. For GOOSE, women displayed a lower F2 (more fronted) than the men. The men also displayed a significant overlap in F2 for TRAP and DRESS ($t(736) = -0.77$, $p > .05$, $d = -0.16$), but this was not the case for women who showed a distinct difference in F2 ($t(736) = 6.66$, $p < .0001$, $d = 1.16$). All other group comparisons were not significant.

3.2. Descriptive comparison with neighbouring London English varieties

Given the close-knit nature of the Stamford Hill community, a secondary aim of the project was to compare our data with previous descriptions of monophthong vowels as produced in the neighbouring London communities [6, 14].

One of the central characteristics of the neighbouring London English monophthongs, as well as many British English dialects, is GOOSE-fronting [6]. As can be seen in the Stamford Hill vowel plot (Figure 1), the GOOSE vowel in the Judeo-English as produced by the men in the community appears to be further back than the GOOSE vowel recorded in other contemporary London dialects [6]. To some extent, this is also true of the women. While they display a more fronted GOOSE vowel than the men, their production is still further back than has been found in other contemporary London varieties [6]. This shows a that the GOOSE vowel in Stamford Hill's Haredi community is moving forward, but at a slower pace than in neighbouring communities. Although to a lesser extent, the FOOT vowel has also seen a front movement in London and other British English dialects. This fronting is not apparent in our data, which shows FOOT as backed in the Judeo-English of Stamford Hill.

The data shows an overlap between DRESS and TRAP vowels in the formant values of some speakers from Stamford Hill. This was especially noticeable in the oldest speaker in our study, suggesting potential age-graded differences that require further investigation. The DRESS vowel in the Stamford Hill data is also considerably lower than in other London varieties and shows closer alignment with RP production patterns than London English varieties [6].

4. DISCUSSION

This study examined a linguistic community that has had limited exposure. Previous research on this community has focused on Yiddish dialects [3]; this study is the first to investigate the phonetic production of Judeo-English in the UK, in Europe's largest Haredi Jewish community [12].

The first objective of this study was to ascertain whether there is a gender difference in the vowel production of Judeo-English in Stamford Hill's Haredi community. The community's pedagogical approach to the genders means that Yiddish is the language of instruction and socialising for most boys and men, while girls and women mostly speak Judeo-English. The separation of the genders from a very young age and the disparate educational experiences of boys and girls, led to the prediction that there would be a significant difference in their vowel production. Our analyses found some significant differences in vowel production between the men and women in our study. The women displayed a more fronted GOOSE vowel in comparison to the men, with the women showing a closer alignment with neighbouring English varieties. We also found slightly more overlap between the TRAP and DRESS vowels produced by the men. A possible explanation for this pattern is there is no TRAP vowel in the Hasidic Yiddish vowel inventory and the closest approximation would be DRESS [15]. More generally, the gender disparity in the vowel production of the Judeo-English spoken in Stamford Hill may stem from the community's approach to employment. Women in traditional Jewish societies were expected to be the breadwinners so their husbands and sons could focus their attention on the study of Torah [16]. This has translated into the contemporary acceptance

of girls learning in English, so they are better prepared for the job market. Although most women work within the community, employment has resulted in slightly more interaction of women with people outside of the community. It is perhaps this point of contact that has prompted the developments that we find in the women's vowel data.

The second aim of this study was to provide a descriptive comparison with the neighbouring London English varieties spoken in Hackney. Our data showed closer alignment with more traditional vowel production patterns. This was particularly noticeable for the close-back vowels FOOT and GOOSE, which, for current SSBE and London dialects are nowadays more fronted [6, 14]. Despite the likely fronting in neighbouring dialects, the data from this study shows a further back GOOSE and to some extent the FOOT vowel in the Stamford Hill community. This degree of fronting is similar to what has previously been observed for older Cockney speakers in Hackney [19]. We also found that the DRESS vowel is lower in the Stamford Hill data and there is some overlap between the DRESS and TRAP vowels for the men. These data suggest that there is minimal influence from surrounding dialects. In many ways, these production patterns are not surprising given the close-knit nature of the community. There is little opportunity or need for speakers to interact with neighbouring communities, whether through work or education. Younger generations in Stamford Hill are taught by members of the same community, who themselves will have learnt English from second or third generation members of the same community. This suggests that the English spoken in Stamford Hill is possibly changing in approximate isolation; Stamford Hill's Jewish community is in some ways a linguistic island.

Research on immigrant minority language groups has shown that a move towards local varieties is typical by the second or third generation [13, 18]. Stamford Hill's Haredi community is now in its fourth and fifth generation from the point of immigration. The community's self-reliance and close-knit society means that the move towards dominant speech patterns outside of the community is progressing at a far slower pace than the norm. While not investigated in the current study, this adherence to a phonetic characteristic may also be a conscious choice on the part of the community to create a unique identity and sense of belonging among its speakers.

We have been privileged to record members of a community that is normally closed off to research. This inaccessibility has led to very little exposure and understanding of Judeo-English in the UK and we have just started to scratch the surface. Stamford Hill is in the position for researchers to examine both extrinsic and intrinsic phonetic change in a traditional bilingual immigrant community.

5. REFERENCES

[1] Bates, D., Machler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67, 1-48.

[2] Baumel, S. D. 2003. BLACK HATS AND HOLY TONGUES: Language and Culture among British "Haredim." European Judaism: *A Journal for the New Europe*. 36(2), 91–109.

[3] Belk, Z., Kahn, L., & Szendroi, K. E. 2020a. Loss of case and gender within a generation: Evidence from Stamford Hill Hasidic Yiddish. *J. Comp. Germ. Ling*. 23(3), 271–326.

[4] Benor, S. B. 2009. Do American Jews Speak a "Jewish Language"? : A Model of Jewish Linguistic Distinctiveness. *JQR*. 99(2), 230–269.

[5] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

[6] Cheshire, J., Fox, S., Kerswill, P. Torgersen, E. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *J. Socioling*. 15(2):151–196.

[7] Fox, John., Weisberg, S. 2011. *An R Companion to Applied Regression,* Second Edition. Sage Publications, Thousand Oaks California.

[8] Kendall, Tyler and Erik R. Thomas. 2010. *Vowels: Vowel Manipulation, Normalization, and Plotting in R*. R package, version 1.1.

[9] Kerswill, P., Torgersen, E. 2017, London's Cockney in the twentieth century: Stability or cycles of contact-driven change? *Listening to the Past: Audio Records of Accents of English.* Hickey, R. (ed.). Cambridge: Cambridge University Press, p. 85-113

[10] Lenth, R. V. 2016. Least-Squares Means: The R Package lsmeans. *J. Stat Software*. 69(1), 1–33

[11] Lobanov, Boris. 1971. Classification of Russian Vowels Spoken by Different Speakers. *J. Acoust. Soc. Am*. 49. 10.1121/1.1912396.

[12] Local Government Association. 2020. Retrieved from :https://www.local.gov.uk/case-studies/london-borough-hackney-charedi-orthodox-jewish-community-engagement

[13] McCarthy, K. M., Rosen, S., Mahon M., Evans, B. G. 2014. Speech perception and production by sequential bilingual children: a longitudinal study of VOT acquisition. *Child Dev.* 85(5), 1965-1980.

[14] McCarthy, K. M., Evans, B.G., & Mahon, M. 2013. Acquiring a second language in an immigrant community: the production of Sylheti and English stops and vowels by London Bengali speakers. *J. Phon*. 41, 344-358.

[15] Nove, C.R, 2021. Phonetic Contrast in New York *Hasidic Yiddish Vowels: Language Contact, Variation, and Change*. PhD. Dissertation. The City University of New York.

[16] Proverbs. Old Testament. Ch.31

[17] R Core Team 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/.

[18] Sharma, D., Sankaran, L. 2011. Cognitive and social forces in dialect shift: Gradual change in London Asian speech. Language Variation and Change. 23(3), 399-428.

[19] Torgersen, E., Kerswill, P Fox, S. 2006. Ethnicity as a source of changes in the London vowel system. In F. Hinskens (ed.). Language Variation - European Perspectives. Selected Papers from the Third International Conference on Language Variation in Europe (ICLaVE3), Amsterdam, June 2005. Amsterdam: Benjamins. 249-263

[20] Wells, J.C. 1982. *Accents of English*. Cambridge University Press

WHAT'S THAT *PHTHONG*? AUTOMATED CLASSIFICATION OF DIALECTAL MONO- AND STANDARD DIPHTHONGS

Simon Oppermann & Beat Siebenhaar

Universität Leipzig
simon.oppermann@uni-leipzig.de, siebenhaar@uni-leipzig.de

ABSTRACT

In East Central German, there's an opposition in the reflexes of Middle High German *ei*, *ou*, *öu*: In Standard German, these diphthongs are shifted to [aɪ̯, aʊ̯, ɔɪ̯] whereas in the dialects they've been monophthongised to [eː, oː, eː] respectively. However, dialectal monophthongs are increasingly replaced by standard diphthongs. This paper uses publicly available data from non-professional speakers recorded over the past 20 years to examine (supposed) lifespan variation in *phthong* use. Combatting this plethora of data by auditory categorisation alone is not feasible. Thus, a procedure to automate the distinction between mono- and diphthongs for larger datasets is proposed: Relevant segments are force-aligned and formant tracks are calculated automatically. Their DCT coefficients and temporal parameters are then used to train multiple random forests. With a classification accuracy of over 95% this scalable model promises sufficiently accurate results for the analysis of lifespan change in *phthong* realisations.

Keywords: DCT, random forest, diphthong, East Central German, lifespan change.

1. INTRODUCTION

In our research project, we are interested in examining linguistic (in-)stability across the lifespan in the use of multiple East Central German (ECG) variables. In this paper, the focus will only be on the realisation of Middle High German (MHG) *ei* and *ou*. As the available dataset currently consists of more than 420 hours of footage, this paper proposes a procedure for automating the distinction between monophthongs and diphthongs using random forest models.

1.1. Automatic distinction of mono- and diphthongs

In sociophonetics, the standard practice to analyse vowels acoustically is the calculation of the first and second formants. These static measurements of F1 and F2 at the temporal midpoints of each vowel, commonly used for distinguishing monophthongs, lack the ability to account for the diphthongs' dynamic nature, which is manifest as a temporal systematic change in formant values. In recent years, several different methods have been proposed to quantify the formant paths of diphthongs. Instead of measuring just the vowel centre, the measurement is extended to multiple time points. Here, time points can either be chosen relative to the vowel length, e.g. at discrete points spaced equally around the midpoint of each vowel [1, 2, 3], or in absolute steps, e.g. every 2.5 milliseconds [4]. Various methods have been proposed to utilise these discrete values to represent the dynamic formant trajectories: polynomial functions, additive models, target-locus scaling, vector-based measurements, and discrete cosine transformations [2]. Polynomial functions [5] and additive models [4, 6, 7] fit "curves to the sampled formant frequencies over time" [3]. Although "they have the advantage of not forcing a parameterisation on trajectory shapes" [3], they are difficult to "generalize across different datasets" [3]. Vector-based approaches [2, 8] measure the Euclidian distance between two data points and/or calculate the angle of these vectors. Despite relying on little spectral information, they have proven to be quite effective [3]. Nonetheless, classification results can be even further improved by representing trajectory shapes with discrete cosine transformations [2].

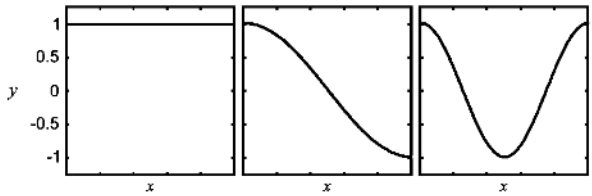


Figure 1: Zeroth, first and second DCT basis functions. Taken from [8], edited by the first author.

Discrete cosine transformations (DCT) [1, 3, 10, 11, 12] work by expressing an input signal sequence as a sum of weighted cosine functions with different frequencies. "The curve parameterization is based on a trajectory's mean plus ½ cosine multiples, each with amplitudes representing deviations from this mean" [3]. The amplitudes of these cosine functions are referred to as DCT coefficients. Figure 1 shows that the zeroth DCT coefficient DCT-0 correlates with the mean of the input signal, DCT-1 (half of a cosine),

with the slope's magnitude and direction and DCT-2 (a full cosine) with the overall curvature [3]. Several studies use these first three DCT coefficients to successfully describe vowel-inherent spectral change [3, 10, 11, 12], while DCT-0 and DCT-1 are the most important [1, 3, 9].

1.2. Lifespan variation and change

In the past two decades, an ever-growing number of studies have been published that more closely examine lifespan stability, variation, and change in several languages (see e.g. the edited volumes [13, 14]). For German, there are only a few studies to date [4, 15, 16, 17, 18]. All these studies show a great amount of inter-individual variation: While the greater part of individuals show linguistic stability, a few others show considerable change over time. E.g. in [4], the majority of speakers change their language over the course of their lifespan, following the community-wide trend toward variants closer to the standard. Only two individuals reveal a retrograde change which is explained by retirement and a pronounced sense of dialect identity. The result that most speakers change their language over their lifespan contradicts those findings, that stability is the predominant individual pattern [19]. However, in a situation with a strong dialect levelling as in Swabia [4], it is likely that most community members conform to the trend.

The majority of these studies only work with two different recording dates [4, 15, 16, 17, 18], and the differences they find at these two stages are interpreted as lifespan change. Studies that look at more time stamps [21, 22, 23] reveal more complex structures, as the realisations of different variants often seem to oscillate, which can be interpreted as a somewhat stable variation. If only two points in time are considered within such oscillating patterns, these differences are overinterpreted or misinterpreted as change. Against this methodological background, a denser temporal grid for the investigation of lifespan change (or variation) would be preferable [21].

1.3. The linguistic situation in East Central German

The data for this paper was recorded in the ECG area, where on the one hand the traditional dialects have already been largely lost and their function is now filled by a regional vernacular – in contrast to many other German areas. On the other hand, the regional realisation of the standard contains many of these vernacular features so that standard and vernacular tend to be quite close to each other [24, 25]. One of the features considered typical of ECG dialects is the realisation of the MHG diphthongs *ei*, *ou*, and *öu* as monophthongs [e:], [o:] and [e:] respectively, thus preserving a phonological contrast that has been lost in New High German (NHG). At present, an abandoning of the monophthongal realisations in the regiolect [24] and thus a trend towards the standard diphthongs can be observed.

MHG	*î*	*ei*	*û*	*ou*	*iu*	*öu*
ECG	aɪ̯	e:	aʊ̯	o:	ɔɪ̯	e:
NHG	aɪ̯	aɪ̯	aʊ̯	aʊ̯	ɔɪ̯	ɔɪ̯

Table 1: Merger of six MHG mono- and diphthongs to three NHG diphthongs. In ECG dialects, the phonological contrast is preserved.

2. DATA

The data basis for this study is the German zoo docusoap "Elefant, Tiger & Co." which follows the everyday life of the zoo animals and the work of Leipzig Zoo employees and has been broadcast weekly since 2003. The employees are accompanied during their work, they speak directly to the TV audience, to the animals, and to each other, while acting entirely without a script. Most speakers' educational, occupational, and socio-professional statuses remain consistent throughout the series. As of the submission of this paper, over 1000 episodes have been broadcast, each approximately 25 minutes long. For the current study, 133 of these episodes are divided into four evenly spaced periods: 2003/2004, 2008/2009, 2013/2014, and 2018/2019. In total, more than 20 regular interviewees have made appearances in at least three of these periods. Three of the most prominent speakers are selected for this paper.

3. METHOD

Each episode is transcribed orthographically in Praat [27]. These transcripts are then automatically force-aligned using WebMAUS [28] and other BAS web services [29, 30]. For the three selected speakers, 258,515 segments were obtained, 8,786 of these being either /aɪ̯/, /aʊ̯/, or /ɔɪ̯/. F1 and F2 values are calculated at 31 equally spaced time points (20%, 22% … 80% of total vowel duration) in Praat and z-normalised in R (v4.2.2) [31] using the package vowels [32]. Additionally, the duration of every segment is extracted. The NHG diphthong in each lexeme is then matched to the corresponding MHG vowel. 4,683 tokens (= 53.3%) correspond to MHG *î*, *û*, *iu*, which are realised as the same diphthongs in standard and dialect (see Table 1). Those that go back to MHG *ei*, *ou*, *öu* add up to 3,662 tokens (= 41.68%). 262 tokens (= 2.98%) correspond to neither of those historical categories. MHG *öu* is by far the rarest with only 83 tokens overall.

In the present, data only diphthongal realisations [ɔɪ̯] can be found. Hence all tokens for MHG *öu* will be excluded from further analysis.

Due to the large size of the entire corpus, neither the force alignment nor the formant paths can be corrected manually. To tidy the data automatically and mitigate the influence of potential noise in the process, tokens with a median F1 less than 200 Hz or larger than 1000 Hz, with a median F2 less than 800 Hz or larger than 2500 Hz, or those longer than 400 ms are excluded [4]. In Table 2 all remaining tokens for the actual analysis are listed, separated by speaker and period.

	MHG vowel	Period 03/04	08/09	13/14	18/19
Speaker 1	*ei*	323	380	253	316
	ou	242	256	162	195
Speaker 2	*ei*	58	29	24	227
	ou	49	12	13	94
Speaker 3	*ei*	307	70	55	173
	ou	138	28	39	84

Table 2: MHG vowel tokens per speaker per period.

For each token, the first three DCT coefficients for F1 and F2 are calculated in R using the package emuR [33]. Together with the log10-transformed duration values these DCT coefficients are used as predictors for a random forest model, calculated with the R-package randomForest [34]. Random forest models have proven to be more accurate than regression models in separating multivariate linguistic data [35], despite being somewhat hard-to-interpret black boxes. They work by fitting multiple decision trees to data, but for each tree not only randomising the specific samples using bootstrap aggregation (bagging) but also changing the subsets of the predictor variables [35]. The final prediction is the average of all individual trees' predictions.

The specific random forest used to categorise automatically all tokens into either mono- or diphthongs is trained on a subset of tokens (n = 324), that previously have been auditorily categorised by the two authors and two student assistants. Tokens assigned being neither mono- nor diphthongs, mostly due to wrong alignment or phonetic reduction, have been excluded from this subset.

4. RESULTS

To achieve the most accurate classification, the random forests were calculated in three different ways: A) Should the model distinguish between MHG vowels? After all, different phonemes correspond to different formant tracks. B) Should the formant data be normalised? Z-normalisation is common when comparing formant data of different speakers. However, previous test runs with smaller samples achieved better results with non-normalised data. C) Should the model distinguish between different speakers? Here, only three speakers are analysed, but ideally the model should work for all the other additional speakers to be analysed later.

± MHG	± Norm.	± Speaker	Accuracy
+	+	+	.909
+	+	−	.896
+	−	+	**.976**
+	−	−	.969
−	+	+	.877
−	+	−	.883
−	−	+	.944
−	−	−	.938

Table 3: Highest classification accuracy scores of different random forest models, each calculated with unique parameters explained above.

Each of the eight configurations in Table 3 was tested on 40 different random forest model settings, each with different numbers of decision trees (ntree = 50, 100 … 2000) and repeated 100 times. The data was split 50/50 into training and test datasets. As there were eight or nine independent variables, depending on whether speaker information was included, the number of variables tried at each split (mtry) was limited to three, as recommended [35]. Where MHG vowels were differentiated, the best classification accuracy scores for MHG *ei* and MHG *ou* respectively were averaged. Table 3 shows that the highest classification scores were achieved with random forest models that were separated by MHG vowel, used non-normalised data, and differentiated between different speakers. The best-performing models achieved an accuracy score of .964 for MHG *ei* and .988 for MHG *ou*, resulting in an average accuracy of .976. As those models were trained at a very specific subset of the original data, they were then used to classify the *phthongs* with different data subsets. After 100 iterations each, the model for MHG *ei* achieved accuracies between .964 and 1 with a median of .976, and the model for MHG *ou* between .988 and 1 with a median of .988.

According to the variable importance scores (see Table 4) the first two DCT coefficients for F1 and F2 generally were the most important variables in both random forest models, contributing best to the homogeneity at the nodes and leaves (= Gini) and leading to the biggest drop in accuracy if they were excluded (= Acc.).

Variable	Acc. *ei*	Gini *ei*	Acc. *ou*	Gini *ou*
k0_f1	8.351	6.724	46.042	8.205
k0_f2	7.859	6.896	4.649	2.424
k1_f1	5.347	6.207	5.283	3.359
k1_f2	5.278	5.698	1.427	2.116
k2_f1	3.171	3.684	2.461	2.326
k2_f2	-0.197	2.710	-6.218	1.646
length	7.013	6.537	-0.091	1.993
speaker	1.184	0.828	4.873	0.718

Table 4: Variable importance scores (mean decrease in accuracy and mean decrease of Gini coefficient) of the best performing random forest models for MHG *ei* and *ou* respectively.

Length is the third most important variable for classifying the realisations of MHG *ei*, but one of the least important ones for MHG *ou*. Information about the speaker also turned out not to be too essential for classification, an observation supported by the fact that those random forests excluding speaker information altogether achieved only minimally worse accuracy scores than those that did not (see Table 3). In a final step, the two best-performing random forests – one for MHG *ei* and one for MHG *ou* – were used to analyse the usage of dialectal monophthongs across the three individual speakers' lifespans.

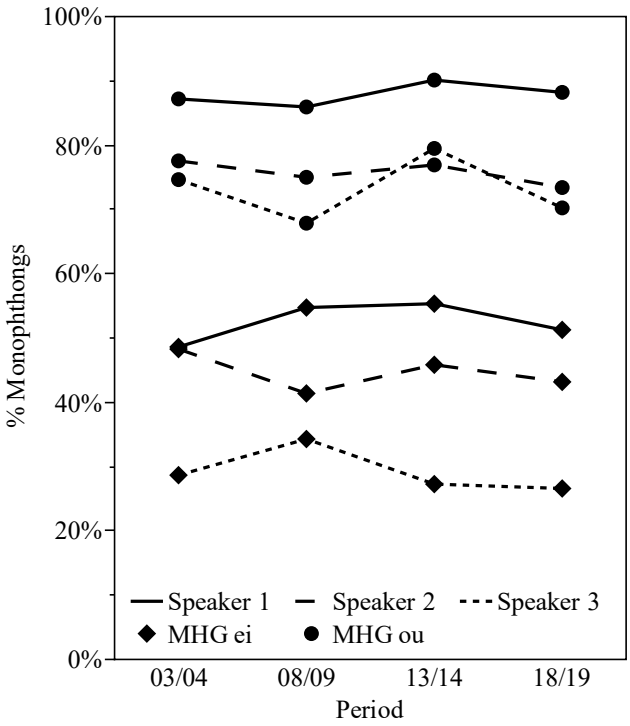


Figure 2: Percentage of monophthongs per speaker per period. Phonemes were classified automatically using the best-performing random forest models for MHG *ei* and *ou* respectively.

Overall, all speakers seem to show stability in their use of *phthongs* but tend to use a higher percentage of monophthongal realisations of MHG *ou* than of MHG *ei* (see Figure 2). Where variation does occur, it is of an oscillatory nature, increasing and decreasing alternately. The envelope of variation across the lifespan is quite low, remaining comfortably in single digits, except for MHG *ou* of Speaker 3. These small variations are reflected in the results of subsequent Pearson's chi-square tests at $\alpha = .05$, which show no statistically significant differences for each speaker and each phoneme ($p > .25$ in all cases). In contrast to this low level of intra-speaker variation, the overall inter-speaker differences are significantly distinct for MHG *ei*, χ^2 (2, N = 2,215) = 94.36, $p < .001$ and MHG *ou*, χ^2 (2, N = 1,312) = 39.21, $p < .001$.

The slight oscillations of the lifespan trajectories, not exhibiting statistically significant differences, could very well be influenced by comparatively low token numbers. Other factors possibly contributing to the varying use of mono- or diphthongs include situation-specific performance effects, as the speakers are recorded in different communicative situations, as well as the frequency of the respective words since in ECG the monophthongisation tends to be highly lexicalised.

5. DISCUSSION

Methodologically, random forest models with the three first DCT coefficients of F1 and F2 and duration of force-aligned segmentations can reach high accuracies in distinguishing monophthongs and diphthongs – at least in the ECG area. Hence this procedure has proven to be valuable for analysing the use of these *phthongs* in larger datasets.

Using this method, the analysis of lifespan variation of three adult ECG speakers reveals relative stability, sometimes slightly oscillating around their individual means. However, the speakers are different in their individual realisation of mono- and diphthongs, resulting in a considerable inter-speaker variation. Their quasi-stable use of monophthongisation mirrors their stable educational, occupational, and socio-professional circumstances.

The general change within the community as reported in [24] is not reflected in an individual change, supporting findings of general stability as one of several possible patterns of individual linguistic variation across the lifespan [19]. Nonetheless, the corpus allows for further individuals to be analysed over their lifetime and for apparent time analyses at different time points with a dataset of more than 100 speakers. Thus, a deeper insight into the correlation between language change and individual variation in a longitudinal study can be expected.

6. REFERENCES

[1] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am* 140(1), 576–581.

[2] Tanner, J., Sonderegger, M., Stuart-Smith, J. 2022. Multidimensional acoustic variation in vowels across English dialects. In: Nicolai, G., Chodroff, E. (eds) *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. ACL, 72–82.

[3] Williams, D., Elvin, J., Escudero, P., Gafos, A. I. 2019. Multidimensional variation in English diphthongs. *Proc. 19th ICPhS* Melbourne, 587–591.

[4] Beaman, K. V. 2021. *Coherence in Real- and Apparent-Time: A sociolinguistic variationist investigation of language change in Swabia*. PhD Thesis. Queen Mary University of London.

[5] Themistocleous, C. 2017. Dialect classification using vowel acoustic parameters. *Speech Communication* 92. 13–22.

[6] Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., Groake, E. 2019. Dialect variation in formant dynamics: the acoustics of lateral and vowel sequences in Manchester and Liverpool English. *J. Acoust. Soc. Am* 145(2), 784–794.

[7] Renwick, M. E. L., Stanley, J. A,. 2020. Modeling dynamic trajectories of front vowels in the American South. *J. Acoust. Soc. Am* 147(1), 579–595 .

[8] Olsen, R. M., Olsen, M., and Renwick, M. E. L. 2018. The impact of sub-region on /ai/ weakening in the U.S. South. *Proc. Mtgs. Acoust.* 31(1), 060005

[9] Morrison, G. S. 2013. Theories of vowel inherent spectral change. In: Morrison, G. S., Assmann P. F. (eds), *Vowel Inherent Spectral Change*. Springer, 31–48.

[10] Watson, C. I., Harrington, J. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *J. Acoust. Soc. Am* 106(1), 458–468.

[11] Williams, D., Escudero, P. 2014. A crossdialectal acoustic comparison of vowels in Northern and Southern British English. *J. Acoust. Soc. Am* 136(5) 2751–2761.

[12] Wolfswinkler, K., Harrington, J. 2021. The influence of Standard German on the vowels and diphthongs of West Central Bavarian. *J. Int. Phon. Assoc*. 1–33

[13] Beaman, K. V., Buchstaller, I. (eds) 2021. *Language Variation and Language Change Across the Lifespan. Theoretical and Empirical Perspectives from Panel Studies*. Routledge.

[14] Wagner, S. E., Buchstaller, I. (eds) 2017. *Panel Studies in Language Variation and Change*. Routledge.

[15] Siebenhaar, B. 2002. Sprachwandel von Sprachgemeinschaften und Individuen. In: Häcki Buhofer, A. (ed), *Spracherwerb und Lebensalter*. Francke, 313–325.

[16] Bausch, K.-H. 2000. Dialektologie und interpretative Soziolinguistik am Beispiel des Sprachwandels im Rhein-Neckar-Raum. In: Stellmacher, D. (ed), *Dialektologie zwischen Tradition und Neuansätzen*. Steiner, 78–98.

[17] Bülow, L., Vergeiner, Ph. C. 2021. Intra-individual variation across the lifespan: Results from an Austrian panel study. In: *Linguist. Vanguard* 7(s2), 1–11.

[18] Ruge, J. 2011. Veränderungen im Dialektgebrauch derselben Sprecher innerhalb von drei Jahrzehnten. In: Glaser, E., Schmidt, J. E., Frey, N. (eds), *Dynamik des Dialekts — Wandel und Variation.* Steiner, 287–300.

[19] Sankoff, G. 2019. Language Change Across the Lifespan: Three Trajectory Types. *Language* 95(2), 1–36.

[21] Bowie, D. 2019. Individual variation in the development of the Western Vowel System of Utah. *Linguist. Vanguard* 5(2), 20180020.

[22] Reubold, U., Harrington, J., Kleber, F. 2010. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Commun.* 52, 638–651.

[23] Reubold, U., Harrington, J. 2017. The Influence of Age on Estimating Sound Change Acoustically From Longitudinal Data. In: Wagner, S. E., Buchstaller, I. (eds), *Panel Studies in Language Variation and Change*. Routledge, 129–151.

[24] Rocholl, M. J. 2015. *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Olms.

[25] Siebenhaar, B. 2019. Ostmitteldeutsch: Thüringisch und Obersächsisch. In: Herrgen, J., Schmidt, J. E. (eds): *Deutsch: Sprache und Raum – Ein Internationales Handbuch der Sprachvariation*. De Gruyter, 407–435.

[27] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer. Computer program. Version 6.3.02. http://www.praat.org/ (accessed 16.11.2022)

[28] Schiel, F. 2015. A Statistical Model for Predicting Pronunciation. *Proc. 18th ICPhS* Glasgow. Paper number 195.

[29] Poerner, N., Schiel, F. 2018: A Web Service for Pre-segmenting Very Long Transcribed Speech Recordings. *Proc. LREC* Miyazaki, 2860–2865.

[30] Reichel, U. D. 2012. PermA and Balloon: Tools for string alignment and text processing. *Proc. Interspeech*. Portland, OR. Paper number 346.

[31] R Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/ (accessed 16.12.2022).

[32] Kendall T., Thomas, E. R. 2018. vowels: Vowel Manipulation, Normalization, and Plotting. R package version 1.2-2. https://CRAN.R-project.org/package=vowels (accessed 14.12.2022).

[33] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. In: *Comput. Speech Lang.* 45, 392–410.

[34] Liaw A., Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/ (accessed 14.12.2022).

[35] Tomaschek, F., Hendrix, P., Baayen, R. H. 2018. Strategies for addressing collinearity in multivariate linguistic data. *J. Phon.* 71, 249–267

TOWARDS A MODEL OF THE LISTENER IN PERCEPTION OF QUEER SPEECH

Cooper Bedin

University of California, Santa Barbara
cbedin@ucsb.edu

ABSTRACT

A fast-growing body of research is examining the relationship between the acoustic properties of a speaker's voice and sexual orientation ascribed to that speaker by listeners, however, these studies by and large only examine variation between speakers. In this paper I ask: what considerations are necessary to develop a model of how listeners integrate acoustic information to make sexuality judgements? This is done by conducting a perception experiment, and modeling sexuality judgements from two listeners of different sociodemographic backgrounds. I focus on modeling architecture, comparing exemplar models against logistic regression and random forest models. I find that the cognitively-motivated exemplar model best captures the behavior of each listener, indicating that future work on queer speech would benefit from engaging with cognitive computational modeling. Ultimately, a model of how listeners integrate acoustic information to make perceived sexuality judgements will contribute towards a more complete understanding of how queer speech is constructed.

Keywords: Sociophonetics, computational modeling, queer speech, cognitive modeling

1. BACKGROUND

A wide collection of experimental studies have drawn connections between acoustic variation in speakers' voices and a speaker's perceived or actual sexual orientation. Early work finds that listeners can be relatively "accurate" in distinguishing gay vs. heterosexual cisgender male speakers from audio alone [1], even with prompts as short as a single word [2]. However, the picture of how phonetic variation contributes to (perceived) sexual orientation remains murky. Take the phoneme /s/, which is one for which effects have been found in many studies. The way /s/ variation is operationalized is inconsistent, where any of center of gravity [3, 4, 5, 6], peak frequency [7], skewness [8, 4], duration [9, 10, 4], or a combination thereof may be analyzed. Additionally, for any one measure, one study may find a relationship between skewness of /s/ and sexual orientation [8], while another finds none [4]. Studies of other acoustic variables such as pitch have been even less definitive [4, 5].

Furthermore, research on sexuality and the voice often focuses on contrasting cisgender gay and heterosexual men. When researchers have examined other LGBTQ+ groups, such as bisexual [4] or transmasculine [5, 6] individuals, individual speakers have patterned with respect to /s/ in ways that are different than in literature on gay men, indicating that there are ways to sound queer beyond the dominant stereotype of the "gay lisp" [3, 11].

Little such attention has yet been paid to variation in listeners. Given that socially-grounded expectations of listeners inform phonetic perception [12, 13] and the interpretation of sociolinguistic variables [14], it would be reasonable to expect that listeners with different experiences perceive sexuality in the voice differently. Indeed, one study suggests that gay and straight speakers may respond to experimental sexuality judgement tasks differently [9], and another finds that listener variability significantly accounts for variability in elicited sexuality judgements [2].

So, a complete model of queer speech needs to account for complex sociophonetic patterns, a diversity of queer speech styles, and differential expectations of listeners. This paper focuses on what such a model—from the perspective of the listener—would look like.

2. EXEMPLAR MODELING

I argue that missing from current literature on sexuality and the voice is cognitively-motivated model of how listeners go from phonetic input to sexuality judgements. Thus, the central focus of this paper is to evaluate the effectiveness of different modeling strategies' abilities to account for the results of an experiment devised based on prior work [2, 7]. In particular, this paper evaluates exemplar models, which are foundational to a class of modern theories of speech perception and have

been applied to other domains in phonetics and phonology [15, 16, 17]. Exemplar theory posits that, when classifying, humans will compare a new stimulus with experiences, or "exemplars," already stored in memory. The more similar the new stimulus is to the exemplars of a category, the more likely a person will judge that the new stimulus is of that category. An exemplar model operationalizes this theory in computational form. For this paper, exemplar modeling is compared against logistic regression and random forests as baselines.

Exemplar modeling makes a different set of assumptions about its data than the other two models considered. In logistic regression, a change in an input variable linearly affects model output (e.g. higher center of gravity directly predicts a more gay-sounding voice). Random forest models are non-linear, but still require imposing the structure of a decision tree onto variables. Exemplar models, by contrast, directly compare new data to data stored in memory, which make them both non-linear and motivated by a class of theories about how humans do categorization.

Given the complex objective of this line of investigation, I take a highly zoomed-in approach to my data for this preliminary approach. I model the behavior of two individual listener subjects chosen for their vastly different sociodemographic backgrounds, and focus on acoustic variation in sibilant phonemes.

3. EXPERIMENTAL METHODOLOGY

Data was collected in an experiment [2, 7] consisting of two tasks, each of which involved a separate group of subjects: a speaker task, which generated audio to be used in a listener task.

3.1. Speaker task

14 speakers were recruited for this study, all graduate or undergraduate students 18–27 years old. Speakers self-identified as "male, masculine, masc, male-presenting, or male-coded" as a prerequisite to participating in the study. After completing the experimental task, speakers described their sexual orientation and gender in open-responses boxes—categorized gender and sexual orientation responses are given in Table 1. Speakers were intentionally recruited to represent a range of sexual orientations and gender identities, to include a diversity of queer voices; it is worth noting that only 2 speakers identify as straight or heterosexual.

Speakers were administered a list of 25 sentences selected from the MOCHA-TIMIT [18] corpus, and recorded reading each sentence aloud. Due to the COVID-19 pandemic, subjects recorded themselves in a quiet part of their home rather than in a lab. Speakers read through the list of sentences three times to ensure that there would be at least one quality recording per sentence per speaker, yielding 350 recordings (14 speakers $\times$ 25 sentences).

Gender	**Count**
Male	11
Non-binary or gender-nonconforming	3
Sexual orientation	**Count**
Straight or heterosexual	2
Gay	5
Bisexual or bi/pansexual	4
Queer	2
Demisexual/asexual panromantic	1

Table 1: Categorized gender and sexual orientation responses for speakers.

3.2. Listener task

Listeners were recruited and administered their task through Prolific [19]. After completing the experimental task, subjects filled out a multiple-choice survey on demographics and connections with the LGBTQ+ community. This paper focuses on modeling the behavior of two listeners—which will be referred to throughout as listeners A and B—who were chosen because they provided maximally different answers to many survey questions, especially those on exposure to LGBTQ+ communities. Answers to relevant questions on the survey are provided in Table 2.

For the listener task, the 350 speaker recordings were shuffled and divided into three lists of 116–117 recordings each. Each listener subject who participated heard one of these lists—listeners A and B both heard the same list (116 recordings total), and so provided judgements for the same recordings presented in the same order. For each recording a listener heard, they would provide an ordinal 1–7 rating for how "queer/gay" the reading sounded to them (1="not at all queer/gay," 7="very queer/gay"). Spearman's rho indicates that listener A and B's ratings in the generated dataset (described below) are significantly—although not strongly—correlated ($\rho = 0.323$, $p < 0.0001$), so there is variation in how they responded to the task, but their ratings are not entirely independently motivated.

Question	Listener A	Listener B
How old are you?	18–20	61–70
How do you describe yourself with respect to gender?	Male	Female
How do you describe yourself with respect to race and ethnicity?	White or European	White or European
How do you describe yourself with respect to sexual orientation?	Bisexual or Pansexual	Heterosexual or Straight
Do you self-identify as LGBTQ+?	Yes	No
How many of your close friends are LGBTQ+?	Most, more than half	Some, less than half
How frequently do you go to spaces that are explicitly LGBTQ+?	About once a month	Never or almost never
How accurately do you believe you are able to determine if someone is queer or gay, upon meeting them?	Very accurately	Not at all

Table 2: Selected survey responses for listeners A and B.

4. DATASET FORMATION

Existing research suggests that acoustic information from multiple segments and types of segments is simultaneously integrated to form judgements about perceived sexual orientation [2, 6, 14, 20]—this study focuses on the sibilants /s/, /z/, and /ʃ/. Sibilants were chosen because they are a relatively small class of phonemes within which acoustic variation is easily measured, and because there already exists extensive literature connecting /s/ to queer speech. Given that prior work has considered several different acoustic measures of variation, center of gravity, skewness, and duration were all measured and given as features to the models. Typically studies consider specifically /s/, however, /ʃ/ and /z/ were also considered for this study, which allowed for the ability to evaluate models' abilities to handle multiple phonemes, and to implicitly test whether future work should continue to focus on /s/.

TextGrids were generated for all recordings using the Montreal Forced Aligner [21], and alignments for all sibilants were then corrected manually. Acoustic measurements were collected via the Python library Parselmouth [22]. Any sibilant of duration $\leq$ 50ms was excluded from the dataset, in addition to any sibilant which occurred in a function word (in this data, the words 'is' and 'was'). This was done to exclude any sibilant whose articulation was highly reduced. This yielded a dataset of 176 sibilant tokens, and three features for every token (center of gravity, skewness, and duration). All three features in the dataset were normalized to a 0–1 range as a preprocessing step for the models.

Ratings given by listeners were simplified by converting them to binary category labels. To convert ratings to labels, all ratings above the listener's mean rating (listener A: 4.2, listener B: 3.5) were converted to YES, and all below to NO. Every sibilant token was matched with the label the listener assigned to the recording in which the token occurred. This yielded two alternative labelings of the same set of sibilant tokens.[1]

5. EVALUATING MODEL PERFORMANCE

All models were evaluated using the Scikit-learn package in Python [23]. Implementations of logistic regression and random forests were also taken from the Scikit-learn library, and exemplar models were implemented based on formulations used by Johnson [17]. Hyper-parameter tuning was done by five-fold cross validation,[2] and model performance was valued by average accuracy[3] across ten-fold cross validation. Models were trained, tuned, and tested separately on the labelings given by listeners A and B. Additionally, architectures were evaluated over four subsets of the data: the entirety of the data, and subsets composed of each individual phoneme (/s/, /z/, and /ʃ/). This yields 7 different scores for each architecture: 2 speakers $\times$ 4 different subsets of the data, minus the the /ʃ/ subset for listener B, because it only contained three YES labels, making it unsuitable for evaluation.

6. RESULTS

All scores are given in Table 3. The logistic architecture performs at or near chance for nearly all of the data subsets. The forest model fared worse, consistently performing well below chance. The exemplar architecture almost entirely outperformed the others, although confidence intervals indicate that there is considerable variance in its performance.

The results suggest that a purely linear treatment of this data is insufficient to capture the full story of listener behavior. With this in mind, consideration was also given to improving the linear model by preprocessing of the data. Model architecture was kept exactly the same, however, each feature was transformed by generating many univariate B-spline

Listener	Subset	# tokens	Chance	Logistic (Raw)	Logistic (Splined)	Forest	Exemplar
A	All	176	0.52	0.52 ± 0.01	0.55 ± 0.13	0.44 ± 0.10	$\mathbf{0.58 \pm 0.10}$
	/s/	81	0.53	0.53 ± 0.03	$\mathbf{0.64 \pm 0.09}$	0.42 ± 0.09	$\mathbf{0.64 \pm 0.16}$
	/z/	73	0.53	0.54 ± 0.09	0.53 ± 0.03	0.41 ± 0.07	$\mathbf{0.61 \pm 0.20}$
	/ʃ/	22	0.55	0.58 ± 0.13	0.59 ± 0.10	0.50 ± 0.07	$\mathbf{0.77 \pm 0.31}$
B	All	176	0.55	0.55 ± 0.01	0.58 ± 0.13	0.56 ± 0.12	$\mathbf{0.69 \pm 0.08}$
	/s/	81	0.53	0.53 ± 0.03	0.59 ± 0.14	0.48 ± 0.23	$\mathbf{0.68 \pm 0.10}$
	/z/	73	0.55	0.62 ± 0.07	0.63 ± 0.13	0.64 ± 0.13	$\mathbf{0.74 \pm 0.05}$

Table 3: Model performance across all data subsets, listeners, and architectures, with 95% confidence intervals. Also given are the number of tokens in each data subset, and expected performance for each subset if a model were to perform at chance. The score of the best performing model for each subset is bolded. "Logistic (Raw)" is the logistic architecture's performance on the data without additional preprocessing, and "Logistic (Splined)" is logistic architecture performance after projecting raw features onto spline bases.

bases for the feature, and projecting the raw features onto these bases. 51 cubic B-spline bases—50 knots—were generated for each of the 3 features, resulting in 53 basis features.[4] This gives the logistic architecture the ability to learn non-linear fits to the data, and have many more degrees of freedom. Average cross-validation scores are given in the "Logistic (Splined)" column in Table 3. For most data subsets an improvement in performance is seen with the transformed data. However, performance does not quite reach that of the exemplar model, and improvements are not seen on every data subset. This indicates that underlying patterns in the data are broadly non-linear, but also that scaffolding a linear architecture to perform as well as the exemplar model would require considerably more machinery.

7. DISCUSSION

In this study, several modeling architectures were compared for their ability to capture patterns in sexuality judgements given by listeners, based on acoustic information from sibilants. Given the central goals of this paper, discussion will focus on identifying relevant considerations for future iterations of modeling work.

Most importantly, treating the data non-linearly improved performance, and exemplar modeling performed best. This indicates that future work should continue to engage with cognitive science research and cognitive modeling architectures. Follow-up work should consider phonemes beyond sibilants—especially simultaneous information from multiple phonemes, and should consider the ordinal ratings given by listeners (instead of simplified binary labels) to create a more detailed model of listener responses. Additionally, while this paper did not investigate the parameters learned by models, such analysis will be crucial in future work.

Results from modeling for this study do not indicate that /s/ is the only sibilant phoneme with an effect on sexuality judgements—in fact, the best, most consistent performance across models is seen on the $\{$/z/, listener B$\}$ data subset. A similar jump is also seen over the $\{$/ʃ/, listener A$\}$ subset, however, the wide confidence interval make it difficult to draw as strong of conclusions. This suggests that effects previously observed on /s/ may extend at least as far as these two other phonemes, and encourages future work to consider how to parametrize information from many different phonemes and types of phonemes.

Lastly, it was observed that models were, across data subsets, able to fit the labelings given by listener B more accurately than listener A. In fact, only for listener B does the entire confidence interval of all exemplar model scores lie above chance performance. This suggests that, in fact, the two listeners did bring different expectations about queer speech to the given task. Table 2 indicates that listener B did not self-identify as LGBTQ+, where listener A did, and listener B also indicated less connection with the LGBTQ+ community than listener A. One possible explanation for why B's ratings were able to be better-described by the model could come from this (lack of) experience—if the ability of a model to fit data is used as a proxy for how "systematic" the data is, then listener A may be basing his judgements on a larger and more diverse set of experiences with queer and trans individuals, where listener B is referencing a more limited set of experiences that may be better-characterized by dominant stereotypes such as the "gay lisp" [3, 11]. This indicates the explanatory ability of modeling to explore differences between listeners, and invites expanding models to compare an arbitrary number of listeners.

8. ACKNOWLEDGEMENTS

I would like to thank Keith Johnson for his support as I collected the data for this project, and Simon Todd for his support on the writing of this paper, as well as the CPLS lab at UCSB for their feedback.

9. REFERENCES

[1] R. P. Gaudio, "Sounding gay: Pitch properties in the speech of gay and straight men," *American Speech*, vol. 69, no. 1, pp. 30–57, 1994.

[2] E. C. Tracy, S. A. Bainter, and N. P. Satariano, "Judgments of self-identified gay and heterosexual male speakers: Which phonemes are most salient in determining sexual orientation?" *Journal of Phonetics*, vol. 52, pp. 13–25, Sep. 2015.

[3] J. Calder, "From sissy to sickening: The indexical landscape of /s/ in SoMa, San Francisco," *Journal of Linguistic Anthropology*, vol. 29, no. 3, pp. 332–358, 2019.

[4] C. Willis, "Bisexuality and /s/ production," in *Proceedings of the Linguistic Society of America*, vol. 6, no. 1, 2021, pp. 69–81.

[5] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech," *Journal of language and sexuality.*, vol. 2, no. 1, 2013-2-27.

[6] ——, "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/," *Language in Society*, vol. 46, no. 3, pp. 339–370, Jun. 2017.

[7] K. Johnson and E. Tracy, "Acoustic and perceptual characteristics of vowels produced by self-identified gay and heterosexual male speakers," *The Journal of the Acoustical Society of America*, vol. 136, pp. 185–214, 2014, 10.1121/1.4899862.

[8] B. Munson, E. C. McDonald, N. L. DeBoe, and A. R. White, "The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech," *Journal of Phonetics*, vol. 34, no. 2, pp. 202–240, Apr. 2006.

[9] G. Jacobs, R. Smyth, and H. Rogers, "Language and sexuality: Searching for the phonetic correlates of gay- and straight-sounding male voices," *Toronto Working Papers in Linguistics*, vol. 18, Jan. 2000.

[10] E. Levon, "Sexuality in context: Variation and the sociolinguistic perception of identity," *Language in Society*, vol. 36, no. 04, p. 533, Oct. 2007.

[11] S. Mack and B. Munson, "The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype," *Journal of Phonetics*, vol. 40, no. 1, pp. 198–212, Jan. 2012.

[12] B. Munson, S. V. Jefferson, and E. C. McDonald, "The influence of perceived sexual orientation on fricative identification," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2427–2437, Apr. 2006, publisher: Acoustical Society of America.

[13] E. A. Strand and K. Johnson, "Gradient and visual speaker normalization in the perception of fricatives." in *KONVENS*, 1996, pp. 14–26.

[14] K. Campbell-Kibler, "Accent, (ING), and the social logic of listener perceptions," *American Speech*, vol. 82, no. 1, pp. 32–64, Feb. 2007.

[15] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency," *Frequency and the emergence of linguistic structure*, vol. 45, no. 137, pp. 10–1075, 2001.

[16] K. Johnson, "Speech perception without speaker normalisation," *Talker variability in speech processing*, pp. 145–65, 1997.

[17] ——, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of Phonetics*, vol. 34, no. 4, pp. 485–499, Oct. 2006.

[18] A. Wrench, "The MOCHA-TIMIT articulatory database," https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, 1999.

[19] "Prolific," https://www.prolific.co, 2014.

[20] K. Campbell-Kibler, "Intersecting variables and perceived sexual orientation in men," *American Speech*, vol. 86, no. 1, pp. 52–68, Feb. 2011.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[22] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[1] Ratings were bisected along the listener average, although one could instead have converted all ratings ≥ 5 to YES, ≤ 3 to NO, and dropped all 4's. Splitting along listener average was done instead because it preserves *all* tokens for analysis, without having to shrink the dataset; and it generates labelings that are balanced between YES and NO labels, reducing the risk that the balance of the dataset affects model performance.

[2] Cross-validation is a standard method for evaluating machine-learning models. In five-fold cross-validation, the data is divided into five equal parts, and for each fifth the model is trained on the other four-fifths of the data, and then tested for its ability to predict the labels of the remaining fifth. This allows one to test not just how well a model learned patterns in its training data, but also how well it can generalize to novel data.

[3] Simple accuracy was used, as all datasets were relatively balanced between YES and NO labels.

[4] Several bases were tested, with 2, 5, 10, 50, and 100 knots, and it was found that 50 knots struck the best balance against over- and under-fitting.

The effects of gender and social class on the deaffrication of /ʤ/ in Jordanian Arabic

Amenah Salman & Amanda Dalola

University of South Carolina, University of Minnesota
ajsalman@mailbox.sc.edu, adalola@umn.edu

ABSTRACT

This study investigates the phenomenon of the deaffrication (DAF) of the voiced palato-alveolar affricate /ʤ/ to the voiced postalveolar fricative [ʒ] in Jordanian Arabic (JA) among indigenous Ammani Jordanian speakers. Previous work has found DAF to be more common in the speech of women than in men in other large cities in Jordan (ex., Irbid, Al-Karak). The goal of this study is to assess if the same is true in Amman, a more culturally diverse and capital city in Jordan, via an examination of not only gender, but also social class and word position. Eight L1-JA speakers participated in a self-paced reading task targeting informal, colloquial JA speech. Results suggest that men were less likely to produce the DAF variant than women and that JA speakers from lower-middle social classes were less likely to produce it than those from high social classes. There was no effect for word position.

Keywords: deaffrication, Jordanian Arabic, gender, social class, sociophonetics

1. INTRODUCTION

Deaffrication (DAF) in Jordanian Arabic (JA) is a phenomenon in which the voiced palato-alveolar affricate /ʤ/ weakens into the voiced postalveolar fricative [ʒ]. Although DAF has received a great deal of attention in the literature on phonology and language acquisition [1,6,9,16], only a few studies have examined its social predictors [3,4,8]. The present work aims to contribute to this body of research through a sociophonetic examination of both social and structural predictors of the variable.

The purpose of this study is to investigate DAF as it occurs in JA among indigenous Ammani Jordanians (from Amman, Jordan) of different social classes. The DAF variable was chosen because it is a high-frequency phonetic variant of the /ʤ/ phoneme in JA; it is important to point out that although /ʒ/ is a phoneme in other varieties of Arabic, it is not in JA. In Modern Standard Arabic (MSA), the phoneme that is orthographically represented as / ج/ is realized as [ʤ] in some Arabic varieties and as [ʒ] in others. While there is an "overlap regarding the two variants as far as the geographical distribution is concerned" [5], the JA phonemic inventory includes the phoneme /ʤ/ to the exclusion of /ʒ/. Given the way assimilatory processes behave across affricates in JA [14], it is expected that there will be word positional differences in how the DAF variant is conditioned. Considering the widespread alternation between [ʤ]~[ʒ] in JA, a rigorous examination of its distribution should provide insight into the roles of gender and social class in Ammani society [8,10].

Affricates are stops followed by fricative releases [7]. Hence, the deaffrication of an affricate simply involves the dropping of the stop segment and the retention (and/or lengthening) of the fricative portion. Figure 1 illustrates the difference between the Arabic affricate /ʤ/ and its deaffricated fricative variant [ʒ] on a spectrogram.

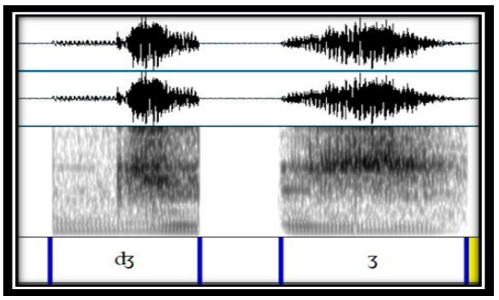


Figure 1: /ʤ/ and [ʒ] on a spectrogram

Ethnicity and rurality have been found to condition DAF usage: Palestinian Jordanians of urban Palestinian origin have been found to use higher rates of the [ʒ] variant, while indigenous Jordanians in provincial towns show higher rates of the standard [ʤ] [2,3,8]. Others have reported interactions between age and education level [3], in which more educated, younger women were found to use the DAF variant more often than less educated, older ones. Previous studies have identified urban women (i.e. in urban centers like Irbid) to be the initiators of language change in Jordanian society [8]; [4] later found that both gender and social class were the strongest predictors, with higher class speakers using more innovative variants than lower class speakers. Since younger speakers were found to use the DAF variant more often than older speakers, [4] argues

for the possibility of a change in progress with regard to DAF in JA.

2. METHODS

2.1 Research Questions

The present study seeks to answer the following research questions (RQs):

- **RQ1:** Which gender group (males or females) is more likely to use the DAF variant in a reading task?
- **RQ2:** Which class group (low-middle vs. high) is more likely to use the DAF variant in a reading task?
- **RQ3:** Does word position (word-initial, word-medial, and word-final) influence the appearance of DAF in a reading task?

2.2. Hypotheses

Based on previous literature, the present research makes the following hypotheses (H):

- **H1:** Females are more likely to use the DAF variant than their male counterparts.
- **H2:** High-class individuals are more likely to use DAF than those from lower classes.
- **H3:** It is expected that there will be positional differences in the conditioning of the DAF variant. However, due to gaps in the pre-existing literature, the directionality of these differences cannot be hypothesized.

2.3 Participants

16 participants (aged 22-24) balanced for gender, were recruited through word-of-mouth and WhatsApp. They were placed into two social class groups on the collective basis of the following criteria, following [10]: university type, residential area, and parental yearly income. Each social class group comprised 8 participants, balanced for gender (4 males, 4 females). One group consisted of upper-class individuals born and raised in West Amman, “a prestigious residential area" [10] who attended private universities and had parents of high socioeconomic status. The second group comprised lower-middle-class individuals born and raised in East Amman, “a non-prestigious residential area” [10], who attended public universities and came from families of lower-middle socioeconomic status. Each of the four groups ultimately contained four participants: upper-class women, lower-middle class women, upper-class men, lower-middle class men (16 total). In an attempt to eliminate the effects of education, urbanization, and age, this study collected data from participants who attained the same level of education (4 years of college), were in the same age group (22-24), and who lived in the same city (Amman). Thus, the only two differences between the participants were gender and social class. To avoid the contact effect of other Arabic dialects [12] where the DAF variable is a phoneme and not a phonetic variant, this study examined only the speech of Indigenous Jordanians, to the exclusion of Jordanians of Palestinian or Syrian origins. Amman was chosen as the site of this investigation due to its status as the capital and largest city of the country. This choice was made because smaller cities with more homogenous populations may not exhibit the same patterns of speech indexing as seen in culturally diverse areas with higher rates of both indigenous and nonindigenous JA speakers.

2.4 Stimuli and Procedure

A reading passage was designed to mimic an informal JA style using colloquial words and phrases popular among L1-JA speakers. This method was chosen to facilitate data collection during the global pandemic, and because previous work has reported that young speakers who use the DAF variant in speech also use it in reading tasks [8]. The reading passage contained fifteen instances of /ʤ/ distributed equally in word-initial, word-medial, and word-final positions. The target words used in the study are shown in Table 1.

IPA	Gloss	Position
ʤədhæ	grandpa	Initial
ʤədethæ	gramma	
ʤæbæl	mountain	
ʤɑlilæh	Jaleeleh	
ʤænub	south	
ældʒæʊ	weather	Medial
jidʒmæʕ	collect	
dædʒædʒ	chicken	
θælædʒeh	fridge	
estaʕdʒelhum	hurry them up	
eldæræʤ	stairs	Final
beħtæʤ	needs	
tawwadʒ	crowned	
zæwædʒ	marriage	
bejbhedʒ	to charm	

Table 1: Target Words

Participants were sent the passage via email and asked to read and record it as a .wav file using a Knox TM gear microphone and the recording application "Voice Recorder" on a Windows operating system, when there was no one present with them in the room. The recordings were then fed into the software Praat [11] for phonetic analysis.

3. ANALYSIS

3.1. Measurements

Tokens of /ʤ/ were assessed and measured using two different spectrographic cues: 1) presence of stop closure in affricates, and 2) the presence of friction in fricatives. Given that fricatives are formed by forcing airflow through a narrow constriction between passive and active articulators, they were assessed on the spectrogram via the presence of aperiodic vibrations (turbulence) at high frequencies. Conversely, affricates were assessed via the presence of a stop closure followed by the high-frequency turbulence of a fricative [12]. Figures 2 and 3 show the visualizations of standard [ʤ] and the DAF variant [ʒ] on the spectrogram in the word /daraʤ/ 'stairs.' As illustrated in Figure 2, the affricate [ʤ] has two parts: a short stop closure followed by short high-frequency turbulence, characteristic of a fricative. Figure 3, on the other hand, shows the DAF variant, which features a longer period of turbulent energy at a high frequency.

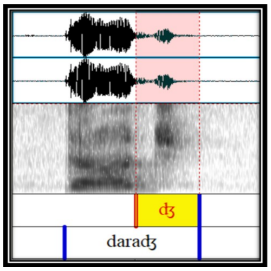


Figure 2: /ʤ/

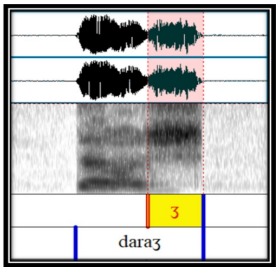


Figure 3: DAF variant

3.2. Statistical Analyses

Measures of binary DAF occurrence were submitted to a stepwise logistic regression with gender (male, female), social class group (high, mid-low), word position (initial, medial, final), and all possible interactions thereof as independent variables, using the glm() function in the *glm2* statistical package in R [13]. All visuals were created using Excel and the *interaction.plot* function in R [13].

4. RESULTS

The stepwise model determined the bestfit model to contain: gender + social class group + gender*social class group. Word position did not figure in the bestfit model. The output of the optimal model is visualized in Table 2:

	Odds Ratio	Std Error	P-value
Intercept	18.99926	0.5923	<0.001 ***
genderMale	0.03271571	0.6491	<0.001 ***
groupmid-low	0.234453	0.6798	<0.005 *
genderMale:groupmid-low	1.4414	1.0247	0.159

Table 2: Bestfit model

Results of the logistic regression found a main effect for gender ($p<0.001$) such that male speakers were 96.72% less likely to produce the DAF variant than female speakers. This effect is visualized in Figure 4:

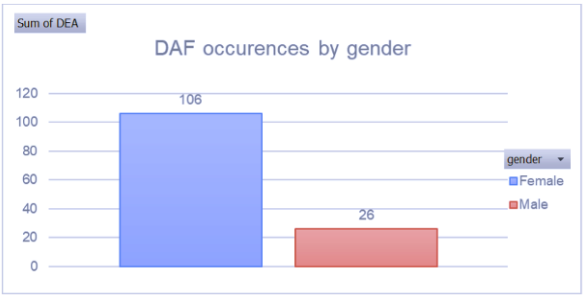


Figure 4: DAF occurrences by gender

There was also a main effect for social class group ($p<0.001$), such that the mid-low class speakers were 76.34% less likely to use the DAF variant than high-class speakers. This effect is visualized in Figure 5:

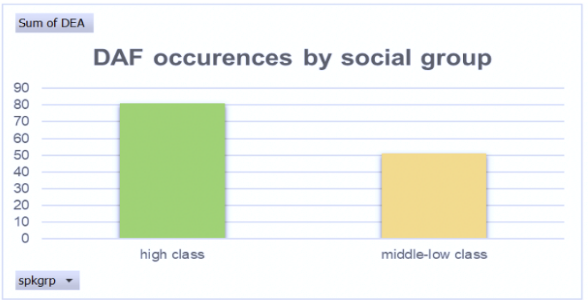


Figure 5: DAF occurrences by social class group

No effect was found for the two-way interaction between group and gender ($p=0.159$). High-class females and males used DAF more often than mid-low class females and males, but the effect was not significant, as visualized in Figure 6:

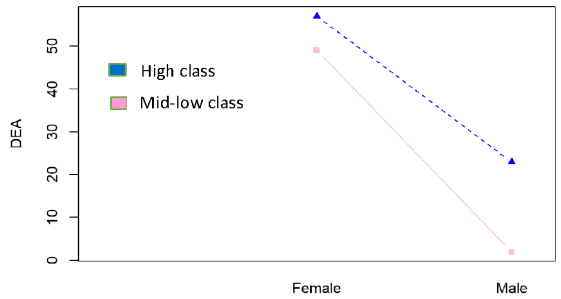


Figure 6: DAF by gender and group

5. DISCUSSION

The main findings of this study suggest that both gender and social class affect how often people produce the DAF variant in Amman, Jordan. In regards to RQ1, it was found that women produce the DAF variant significantly more often than men, in support of our hypothesis. This provides additional support for [3] who targeted uniquely female speakers in their investigation of other social predictors of language variation. Additionally, the finding supports [3,4,8,14] by suggesting that (young Ammani educated) women are more prolific users of the DAF variant than men, indirectly signaling that they may be leading this linguistic change in Amman, as has already been reported more generally for other linguistic variables in cities like Irbid [8] and Al-Karak [14].

For RQ2, it was found that high-class individuals produce the DAF variant more frequently than lower class individuals, also in support of our hypothesis. This finding lends additional evidence to [4,8] who reported speakers in higher classes in Jordan to use more innovative variants in speech, including the DAF variant, all around. This finding also develops on previous work in sociolinguistics that has reported higher classes to be less sensitive to mechanisms of prestige and linguistic accommodation overall, due to their inherently privileged status [15].

For RQ3, it was found that word position did not significantly or meaningfully predict the occurrence of the DAF variant, contra our hypothesis. In fact, the word position factor did not figure into the bestfit model reported here (or any competing models proposed by the stepwise analysis). This finding suggests several possibilities: 1) that phonological environment is not a meaningful predictor of the DAF variant, or 2) that our operationalization of DAF (word-initial, word-medial, word-final) was not adequate to capture any phonological conditioning that may be present in the variable. [8] suggests that the alternation between the two variants is, in fact, phonetically conditioned, but that its appearance is predicted not by where it occurs in the word, but by the place features of certain adjacent consonants, i.e. it is more common in front of dentals and bilabials than in front of vowels. A third possibility is that phonological environment may, in fact, play some role in predicting the appearance of DAF, but that its effect is much weaker than that of the social factors of gender and social class under investigation here.

Future research should seek to more robustly examine the interaction of age with gender and social class. The participants in the present study were all between the ages of 22 and 24, and therefore represented only a sliver of the Ammani population. This fact also calls for a more robust operationalization of social class, in particular one that is based on the participants' own income rather than that of their parents.

In terms of structural variables, future research should seek to examine the conditioning behaviors of adjacent segments, as well as overall syllable count in the target word, as the variable may be preferred in internal position, primarily when it occurs in bisyllabic words [8]. The role of lexical frequency should also be considered, given the oft-reported tendency of higher-frequency words to more readily exhibit sociophonetic variation. From a phonetic standpoint, future work may choose to include durational and spectral measures of the DAF variant that seek to qualify and compare its energy distribution relative to whole segments. In spite of its social distribution, it has been reported that the [ʤ]~[ʒ] alternation is not a salient and, therefore, not a stigmatized linguistic marker in Jordan [4,8]. However, a post-conversation with one of the participants indicated otherwise. When asked why she deaffricated her /ʤ/, she claimed that the DAF variant made her speech "sound more feminine." As such, it is possible that the deaffricated variant [ʒ] is, therefore, becoming a more salient marker in JA.

6. CONCLUSION

This study investigated the role of gender, social class, and word position on the appearance of DAF in JA. Results revealed that women and individuals from high social classes were more likely to use the DAF variant than men and individuals from lower social classes. While the interaction did not reach statistical significance, it can be observed descriptively that women from higher social classes were the most likely to use DAF than any other group. A variety of elicitation methods and more engagement in ethnographic fieldwork are needed to determine the socio-indexical properties of DAF and listener attitudes towards it in JA.

7. REFERENCES

[1] A. Alqarni, *The realization for the English voiceless postalveolar affricate /t[esh]/ in Najdi Saudi ESL learners production*, Southern Illinois University: Carbondale ProQuestDissertations Publishing, 2013.

[2] E. Al-Wer, The formation of the dialect of Amman: From chaos to order, *Arabic in the City* Routledge, 2007.

[3] E. Al-Wer, E. Essa. *Phonological variation in the speech of women from three urban areas in Jordan*. Diss. University of Essex, 1991.

[4] F. AI-Tamimi, *Phonetic and Phonological Variation in the Speech of Rural Migrants in a Jordanian City*, Ph.D. diss: The University of Leeds Department of Linguistics and Phonetics , 2001.

[5] H. Monassar, Modern Standard Arabicج: /ʒ/ And /ʤ/.*International Journal Of Linguistics*, Literature, And Culture, 2014.

[6] J. Taqia, Consonant Lenition among Speakers of Arabic, College of Education for Women, Iraqia University: *International Journal of Innovation*, Creativity and Change, 2020.

[7] L. Katz, *Temporal resolution and voice -onset -time perception in sensorineural hearing loss*, University of Wisconsin--Madison, 1985.

[8] M. Al-Khatib, *Sociolinguistic change in an expanding urban context: A case study of Irbid city, Jordan*., Doctoral Dissertation: University of Durham, 1988.

[9] Marta Szreder, Derrick Donald, and Ben-Ammar Chahla, Affricate variation in Emirati Arabic, *Experimental Arabic Linguistics*, 2021.

[10] O. Omari and A. Jaber, Variation in the acoustic correlates of emphasis in Jordanian Arabic: Gender and social class, *Folia Linguistica*, 2019.

[11] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.2.03, from *http://www.praat.org/*, 2021.

[12] R. Aronow, Phonetics: The Basics About Acoustic Features Of Consonants In Standard English, *Linguisticsnetwork.Com,* 2021.

[13] RStudio Team, RStudio: Integrated Development for R. RStudio PBC, Boston, MA URL *http://www.rstudio.com/*., 2020.

[14] S. Albdairat, *Sociolinguistic Study of Phonological Variation in Jordanian Arabic in Al-Karak*, Ph.D. diss., Indiana University, 2021.

[15] W. Labov, The social stratification of (r) in New York City department stores. In *Dialect and language variation.*, Academic Press, 1986.

[16] Y. al-Rojaie, Regional dialect leveling in Najdi Arabic: The case of the deaffrication of [k] in the Qaṣīmī dialect, *Language Variation and Change*: Cambridge University Press, 2013.

The urge to unmerge: a case of structural change across the lifespan

Marc Barnard, Robert James Hellyer & Sophie Holmes-Elliott

Queen Mary University of London
m.e.barnard@qmul.ac.uk, r.j.hellyer@qmul.ac.uk, s.holmes-elliott@qmul.ac.uk

ABSTRACT

This paper considers change across the lifespan by investigating an adult speaker's development from a one-part low back vowel system to a two-part system. Specifically, we use podcast data to track the real-time development of a LOT-THOUGHT split in the speech of an American English speaker who previously exhibited a merged system in production. Using automatic forced alignment, we extracted measures of the first and second formants. Degree of overlap was analysed using lmers and Pillai scores. Analysis revealed that the speaker's one-part system diverged into two statistically significantly distinct categories, coming to resemble the dominant pattern of the region in which he resides. As previous work indicates that grammatical plasticity is constrained in adulthood, the current finding contributes a new insight into the nature of post-critical period change. These findings are discussed within the broader context of underlying phonological representations and motivations behind language change across the adult lifespan.

Keywords: lifespan change, sociophonetics, low back vowels, merger, plasticity

1. INTRODUCTION

1.1. Change across the lifespan

Research into language change across the lifespan has presented evidence that adult language systems tend to be relatively stable [24]. Relatedly, some aspects of grammatical plasticity appear to be significantly constrained after the end of the critical period [24], hypothesised to occur in late adolescence [29]. Despite this, a growing body of work shows that lifespan change is possible to some degree [7, 9, 12, 14, 15, 20, 21, 24–26, 28]. The full extent to which post-adolescent acquisition is limited to certain types of language change is uncertain [12]. It is still unclear if adult language change involving structural phonological categories is possible, or whether it is restricted to more superficial adjustments [24].

The current paper contributes to this area by presenting a case study focussed on production data from one adult speaker whose speech evidences the development of a phonemic split where one vowel category becomes two.

1.2. The limits of change across the lifespan

Previous findings regarding the existence and nature of limits to post-adolescent language change are inconclusive. However, evidence of complex structural change is rarely, if ever, attested.

The most commonly reported type of lifespan change is that which occurs among already-variable elements of the grammar. Adult speakers may alter their rate of one variable linguistic feature due to sociolinguistic pressures, while maintaining the other variant [25]. For example, Sankoff & Blondeau [26] found that Montreal French speakers who already varied between dorsal and apical /r/, shifted their preference towards the dorsal variant in line with a community-wide change in progress. Likewise, in her analysis of Sir David Attenborough's speech, MacKenzie [12] finds that he removes a previous existing constraint on variable tapped /r/ use. Attenborough initially uses a greater rate of /r/-tapping in intervocalic, word-internal positions (e.g., *forest*) and fewer in situations where /r/ links two words (e.g., *for a*). However, he begins to show similar tapping rates in both contexts, due to the reanalysis of certain frequently-uttered collocations as single words. In this way, MacKenzie [12] argues that while his surface pronunciation shifts, the grammatical rules governing his use of this variable remain fixed.

Further, evidence shows that it is easier for adults to simplify their phonological system by merging categories than complexifying it through adding a new phoneme or allophonic rule [2]. In her study of Noam Chomsky, Kwon [9] uses historical recordings to track his vowel system across time. She finds that he approximates the local merged low back vowel system after moving to Boston, having previously produced separate LOT and THOUGHT categories. However, the merger does not reach categoricity, suggesting that the underlying phonemes may remain distinct and no phonological restructuring takes place. Similarly, Riverin-Coutlée & Harrington [21] show how speakers may adopt merged categories that can be used stylistically alongside their initial system. Michaëlle Jean, a Quebec French public speaker, appears to lose a typically Quebecois allophonic distinction between tense and lax vowels when

regularly addressing an international audience, though reverts to her earlier split system when these pressures relax. That adult acquisition of simple changes such as mergers may be easily reversible [20, 21, 27] may indicate a lack of underlying grammatical change. Instead, speakers may be altering the surface realisation of their speech for stylistic purposes.

More complex patterns such as splits are relatively rare in the literature [4, 9, 14] and where speakers do acquire these, they often do not achieve native-like categoricity [26]. For example, Kwon [9] finds that Chomsky begins to reorganise his short-a system towards the more complex nasal system, but he fails to completely master the allophonic split. Likewise, Evans & Iverson [7] find that northern English speakers begin to change their STRUT and FOOT realisations towards southern-like vowels, but this is not accompanied by changes in perceptual processing. This may indicate that while phonetic realisations are available for change, adult speakers' phonological representations remain constant. In both of these cases, speakers who change do not achieve a native-like split between the two categories.

1.3. An example of complex change?

Taken together, the above findings suggest that adult language change appears to be largely limited to more simple changes. Acquiring new structural categories, such as phonemes, appears to be a much harder task. This supports Chambers' [2] third principle of second-dialect acquisition, whereby simpler phonological rules are easier to acquire than complex ones.

The current research presents preliminary evidence that complex structural change, in this case the adoption of a split low back vowel system, may be possible among adult speakers. The implications of this for our understanding of lifespan change are also discussed.

2. METHODOLOGY

2.1. Sample

This study exploits readily available data in the form of a podcast, *Overdue*. The book-discussion podcast is recorded remotely by two native US-English speakers, Andrew Cunningham and Craig Getting. Andrew was brought up in Central Ohio, an area with a merged LOT-THOUGHT category [10]. At the time of recording, he was living in New Jersey, where these categories tend to be split [10]. Craig was also living in a split area, Philadelphia, where he was born and grew up. The study examines speaker accommodation to the ambient dialectal pattern in their respective communities. If this occurred, we would predict that Craig's system would remain stable, whereas Andrew's one-part system would shift towards a two-part, in line with the ambient Philadelphian dialect.

2.2. Data processing

Data were drawn from four episodes spanning roughly three years (Ep. 22, July 2013; Ep. 65, Jun 2014; Ep. 119, Jun 2015; Ep 183, Jun 2016). Episodes were accessed on Spotify, recorded using Audacity and orthographically transcribed using Transcriber. The data were forced aligned using DARLA [19]. F1 and F2 values were extracted at one-third of the way between onset and offset, following Labov, Rosenfelder & Fruehwald [11], and Lobanov normalised. The output was manually checked by the researchers, with unstressed and non-native words removed. A total of 825 tokens of LOT and THOUGHT were analysed, from 159 minutes of podcast recordings (see Table 1 below).

Table 1: Table showing the minutes of recorded data and the number of tokens analysed.

Time point	Minutes	LOT tokens	THOUGHT tokens
July 2013	35	*Andrew*, 19 *Craig*, 95	*Andrew*, 17 *Craig*, 55
June 2014	35	*Andrew*, 26 *Craig*, 31	*Andrew*, 14 *Craig*, 19
June 2015	56	*Andrew*, 42 *Craig*, 164	*Andrew*, 30 *Craig*, 56
June 2016	33	*Andrew*, 87 *Craig*, 93	*Andrew*, 34 *Craig*, 42

2.3. Analysis

Acoustic analysis was performed in three stages. First, the Lobanov normalised F1 and F2 of each token was plotted (see Figure 1 below). Following this, the overlap of the distributions was quantified using Pillai scores. Pillai scores are the output of a MANOVA and consist of a score between 0-1 showing degree of overlap, with 0 representing no distinction between the categories and 1 showing no overlap [16]. Following this, data were analysed using Linear Mixed Effects Regression (lmer) modelling in R [18], to statistically confirm any significant differences in Lobanov normalised F2 values over the time points. Fixed effects included a three-way interaction of Speaker, Vowel and Time, as well as Following Segment. Word was included as a random intercept.

3. RESULTS

Analysis of Andrew's LOT/THOUGHT categories suggest that over the course of the three years his initially merged system comes to approximate Craig's split system.

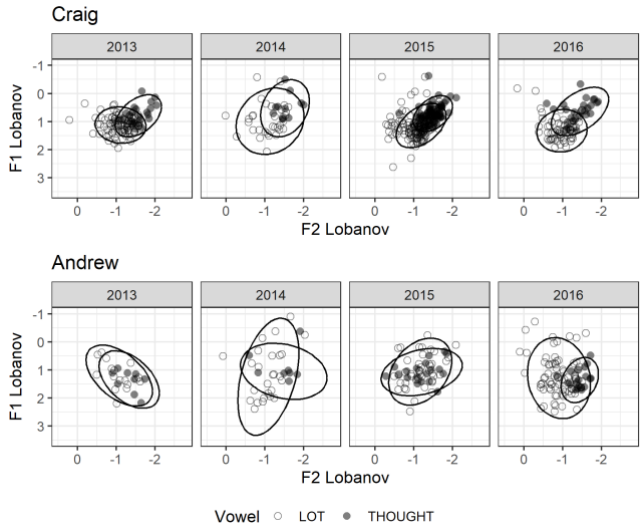


Figure 1: Distribution plots showing Lobanov-normalised F1 and F2 values for Craig and Andrew.

Figure 1 shows the Lobanov-normalised F1 and F2 distributions of LOT and THOUGHT categories for both speakers. Each point represents a single token, LOT in outlined points, THOUGHT in solid, with ellipses generated to represent the spread of each speakers' vowel distribution and relative overlap. Craig's results serve as an example of a typical split system, as used by a speaker with a similar demographic profile to Andrew. His token distributions suggest that he has two low back vowel categories, as the distribution ellipses coalesce around two separate centre points, indicating different mean F1 and F2 values for both categories. The ellipses remain distinct throughout all time points, suggesting that Craig's vowel system shows little change over time.

In contrast, Andrew's vowels do not show substantially separate distributions until 2016. From 2013 to 2015, the distribution of his LOT and THOUGHT categories closely resemble each other, with high degrees of overlap. By 2016, the means and distributions of the two vowels have begun to separate into two distinct categories. During this episode, Andrew's vowel system appears to resemble Craig's patterning, with a fronter LOT vowel and a backer THOUGHT category.

Table 2: Table showing Pillai scores and p-values from the three-way interaction in the lmer model (Speaker*Vowel*Time), with vowel F2 as the dependent variable.

	Craig		Andrew	
Year	**Pillai Score**	**lmer 'vowel' *p*-value**	**Pillai Score**	**lmer 'vowel' *p*-value**
2013	0.401	<.0001	0.146	0.56
2014	0.296	<.0001	0.128	0.32
2015	0.251	<.0001	0.005	0.78
2016	0.491	<.0001	0.278	<.0001

The Pillai scores in Table 2 further support this analysis. From 2013-2015, Craig's scores are consistently higher than Andrew's, indicating a greater distinction between the two categories. By 2016, Andrew's Pillai score is within the range of Craig's scores, suggesting the distribution of the two categories has become more distinct.

The lmer model indicates that the observed pattern reaches statistical significance. The difference in F2 between Craig's LOT and THOUGHT categories remains significant at every time point (at $p<0.0001$; see Table 2). However, Andrew's categories are not significantly distinct until 2016.

4. DISCUSSION

4.1. General discussion

The current findings show that Andrew's previously merged low back vowel system splits into two separate categories, a fronter LOT and backer THOUGHT. In this way, he begins to approximate a speaker with a native unmerged system. This result is surprising in light of previous work that has emphasised the difficulty with which adult speakers acquire complex phonological patterns, such as splits. The development of such a pattern may indicate that the post-adolescent language faculty is more malleable than previously hypothesised. Further, the fact that Andrew's split resembles Craig's native two-part system stands in contrast to previous examples of complex change [7, 9], where the patterns acquired are not native-like. This may provide further indication that adult speakers are better at acquiring new and complex language features than earlier work suggests.

While any complex production shift is unexpected, it is difficult to ascertain from the current data whether this reflects changes to Andrew's underlying representations. Labov, Ash & Boberg

[10] show that for many Central Ohioan speakers, the merger is in a transitional stage, or present in perception but not production. Likewise, Durian [6] argues that while many speakers' realisations of the two phonemes are very close, not all speakers in Central Ohio have a complete merger. It may be the case that Andrew's original phonological system contained two low back vowel categories, but that he realised these categories identically, in line with the increasingly dominant pattern in that part of the Midwest [6]. Andrew could be reverting to an earlier pattern that he had already stored, rather than learning a new system. His case is perhaps similar to that of Michaëlle Jean, who exhibited a simpler system in adulthood, but retained her more complex native system for stylistic use [21]. However, from the available production data, these interpretations are necessarily speculative. Psycholinguistic methods would be required to more fully understand the status of Andrew's underlying grammar.

4.2. Mechanisms for lifespan change

The mechanisms by which Andrew's language shifts are also unclear. One relevant factor in adult language acquisition is degree of exposure to new features [23, 28]. Andrew is surrounded by the two-category system in his adopted home state of New Jersey, and it is common for speakers who have moved region to acquire new dialect features through ambient exposure [14, 23, 28]. Regular interaction with his unmerged co-host may also reinforce use of the split system. Interspeaker phonetic convergence has been shown to occur within individual conversations [1, 8], as well as longitudinally [17]. Recent scholarship has emphasised that relatively automatic acquisition from the surrounding environment and socially-determined convergence to individual speakers may play complementary roles in processes of phonetic change [1, 3, 22]. For example, Ross et al. [22] find that familiarity from exposure to a second dialect facilitates phonetic convergence in dialogic situations. Similarly, Andrew's shift towards the split system may be driven by both immersion in the unmerged dialect of New Jersey, and a desire to accommodate towards Craig's speech. The fact that both speakers currently live in states with a split low back vowel system could indicate why it is Andrew, rather than Craig, whose speech changes.

4.3. Motivations for lifespan change

This account, however, still fails to explain why Andrew split his categories at that particular moment. By 2016, Andrew had been living in New Jersey for five years, and co-hosting the podcast with Craig for three. One possibility is that these five years represent a threshold at which he had acquired enough input to produce the unmerged system. However, this hypothesis finds little support in previous work, where length of residence in a second dialect region has shown to be an insufficient predictor of adult language change [28]. Other social motivations must be at play. One possible factor is his upcoming move to Philadelphia, another split region and the hometown of his co-host. This occurs around a month after the 2016 episode analysed here and is extensively discussed in one of the following episodes (Ep. 186). Orientations towards place have been shown to correlate with adoption of new dialect features [13, 15], and may influence production before any actual move takes place [5, 13]. For example, De Decker [5] finds that Ontarians who wish to move to the city show distinct vowel patterns compared to those who intend to remain in a rural area. This suggests that the timing of Andrew's vowel shift could have been determined by his anticipation of the upcoming move, which increased the salience of Philadelphia and his degree of orientation towards the city. While Andrew may have acquired the split categories through ambient exposure, sociolinguistic factors such as place orientation and communication accommodation may have determined why the split system appears in his own speech in 2016.

5. CONCLUSION

The current research provides evidence of an adult speaker who begins to produce split LOT and THOUGHT categories after previously exhibiting a merged low back vowel system. This finding runs counter to previous claims that adult speakers are unable to reliably acquire new complex patterns. Accordingly, the podcast hosts offer a valuable case study that contributes to our understanding of the types of language change that are possible following the critical period. The present research is limited to a focus on production; the question of whether the change constitutes phonological restructuring would require a longitudinal study of perception. Further perceptual work would also elucidate the extent to which Andrew's vowels perceptually resemble Craig's for Philadelphians. The mechanisms and motivations behind Andrew's shift also remain unclear, but a combination of ambient exposure, sustained interaction with Craig, and a stronger orientation towards a dialectal region which exhibits the split are all likely contributing factors.

6. REFERENCES

[1] M. Babel, 'Dialect divergence and convergence in New Zealand English', *Lang. Soc.*, vol. 39, no. 4, pp. 437–456, 2010.

[2] J. K. Chambers, 'Dialect acquisition', *Language*, vol. 68, no. 4, pp. 673–705, 1992.
[3] E. H. Coles-Harris, 'Perspectives on the motivations for phonetic convergence', *Lang. Linguist. Compass*, vol. 11, no. 12, p. e12268, Dec. 2017.
[4] J. Conn and U. Horesh, 'Assessing the acquisition of dialect variables by migrant adults in Philadelphia: A case study.', *Univ. Pa. Work. Pap. Linguist.*, vol. 8, no. 3, pp. 47–57, 2002.
[5] P. De Decker, 'A real-time investigation of social and phonetic changes in post-adolescence', *Univ. Pa. Work. Pap. Linguist.*, vol. 12, no. 2, pp. 65–76, 2006.
[6] D. Durian, 'A New Perspective on Vowel Variation across the 19th and 20th Centuries in Columbus, OH', The Ohio State University, Columbus, OH, 2012.
[7] B. G. Evans and P. Iverson, 'Plasticity in vowel perception and production: A study of accent change in young adults', *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3814–3826, 2007.
[8] M. Kim, W. S. Horton, and A. R. Bradlow, 'Phonetic convergence in spontaneous conversations as a function of interlocutor language distance', vol. 2, no. 1, pp. 125–156, 2011.
[9] S. Kwon, 'Phonetic and phonological changes of Noam Chomsky: A case study of dialect shift', *Am. Speech Q. Linguist. Usage*, vol. 93, no. 2, pp. 270–297, 2018.
[10] W. Labov, S. Ash, and C. Boberg, *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2008.
[11] W. Labov, I. Rosenfelder, and J. Fruehwald, 'One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis', *Language*, vol. 89, no. 1, pp. 30–65, 2013.
[12] L. MacKenzie, 'Frequency effects over the lifespan: a case study of Attenborough's r's', *Linguist. Vanguard*, vol. 3, no. 1, pp. 1–12, 2017.
[13] M. Monka, P. Quist, and A. R. Skovse, 'Place attachment and linguistic variation: A quantitative analysis of language and local attachment in a rural village and an urban social housing area', *Lang. Soc.*, vol. 49, no. 2, pp. 173–205, 2020.
[14] J. Nycz, 'Second dialect acquisition: Implications for theories of phonological representation', New York University, New York, 2011.
[15] J. Nycz, 'Stylistic variation among mobile speakers: Using old and new regional variables to construct complex place identity', *Lang. Var. Change*, vol. 30, no. 2, pp. 175–202, 2018.
[16] J. Nycz and L. Hall-Lew, 'Best practices in measuring vowel merger', *Proc. Meet. Acoust.*, vol. 20, no. 1, p. 060008, Dec. 2013.
[17] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, 'Phonetic convergence in college roommates', *J. Phon.*, vol. 40, no. 1, pp. 190–197, 2012.
[18] R Core Team, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022.
[19] S. Reddy and J. Stanford, 'A Web Application for Automated Dialect Analysis', *Proc. NAACL-HLT 2015*, 2015.
[20] U. Reubold and J. Harrington, 'The influence of age on estimating sound change acoustically from longitudinal data', in *Panel studies of variation and change*, Routledge, 2017, pp. 129–151.
[21] J. Riverin-Coutlée and J. Harrington, 'Phonetic change over the career: a case study', *Linguist. Vanguard*, pp. 1–12, 2022.
[22] J. P. Ross, K. D. Lilley, C. G. Clopper, J. S. Pardo, and S. V. Levi, 'Effects of dialect-specific features and familiarity on cross-dialect phonetic convergence', *J. Phon.*, vol. 86, p. 101041, 2021.
[23] M. L. Sancier and C. A. Fowler, 'Gestural drift in a bilingual speaker of Brazilian Portuguese and English', *J. Phon.*, vol. 25, no. 4, pp. 421–436, 1997.
[24] G. Sankoff, 'Language change across the lifespan', *Annu. Rev. Linguist.*, vol. 4, pp. 297–316, 2018.
[25] G. Sankoff, 'Language change across the lifespan: Three trajectory types', *Language*, vol. 95, no. 2, pp. 197–229, 2019.
[26] G. Sankoff and H. Blondeau, 'Language change across the lifespan:/r/in Montreal French', *Language*, pp. 560–588, 2007.
[27] A. Shapp, N. LaFave, and J. V. Singler, 'Ginsburg v. Ginsburg: A longitudinal study of regional features in a Supreme Court Justice's speech', *Univ. Pa. Work. Pap. Linguist.*, vol. 20, no. 2, pp. 149–158, 2014.
[28] J. Siegel, *Second dialect acquisition*. Cambridge University Press, 2010.
[29] F. van der Slik, J. Schepens, T. Bongaerts, and R. van Hout, 'Critical Period Claim Revisited: Reanalysis of Hartshorne, Tenenbaum, and Pinker (2018) Suggests Steady Decline and Learner-Type Differences', *Lang. Learn.*, vol. 72, no. 1, pp. 87–112, 2022.

VOICE TYPES AND VOICE QUALITY IN JAPANESE ANIME

Carlos T. Ishi[1,2], Akira Utsugi[3], Ichiro Ota[4]

[1] Guardian Robot Project, RIKEN, [2] ATR, [3] Nagoya University, [4] Kagoshima University
carlos.ishi@riken.jp, utsugi@nagoya-u.jp, iota@leh.kagoshima-u.ac.jp

ABSTRACT

Japanese anime is noteworthy for being dubbed by professional voice actors and is unique in its vocal characteristics. In this study, recordings were taken of students of voice actor course at a music college reading short texts and anime lines in various voice types. These recordings were analyzed by focusing on the acoustic parameters related to voice quality. The results showed that for a repeating task, when students were asked to repeat an actual animated voice, acoustic parameters such as F0 and sub-band harmonics-to-noise ratio (HNR) clearly captured the differences between several anime characters. When analyzing recordings of short texts in various voice types, a similar trend was found, although not as distinct as the repeating task. These results suggest that multiple anime voice types have been established in Japanese anime and voice quality helps to distinguish them.

Keywords: anime, voice quality, fundamental frequency, sub-band HNR, register

1. INTRODUCTION

Japanese anime voices are unique from the perspective of the voice culture. Unlike Western animation, which is often performed by actors, a majority of Japanese anime is dubbed by professional voice actors called "seiyu." Further, anime voices of children are performed by adult female seiyus [1]; this indicates that Japanese anime voices have been unique from its early days. In contemporary Japan, seiyu is a popular profession among young people [2]. This background and contemporary trend are speculated to have influenced the characteristics of the anime voices. It has been suggested that multiple anime voice types are becoming established in contemporary anime, and can be viewed sociolinguistically as registers [3]. One of the aims of this study is to determine whether voices performed by different voice actors have similar vocal characteristics if the types of characters they perform are similar.

The study of anime voices has several approaches, that can be categorized by their descriptive and data-collection methods.

Impressionistic and quantitative methods form the descriptive categorization. Anime voices are often described impressionistically when discussed from the perspective of media studies. In contrast, sociophonetic studies have attempted to acoustically analyze voices via characteristics such as F0 [4-5] and voice quality [3,6-8].

Further, the data collection has two approaches: analyzing actual animated speech [5-8], and obtaining multiple types of speech from the same speaker through researcher intervention [3-4].

One of the goals of this study is to examine how the voices of different types of anime characters exhibit different characteristics in terms of acoustic features related to voice quality. Although previous studies have analyzed the acoustics of anime voices, few studies have been conducted on character-specific analysis of contemporary Japanese anime. Another goal of this study is to determine whether character voices have similar characteristics, regardless of the individual seiyus.

This study compared multiple types of speech recorded from the same speaker under the same conditions. As previous studies using a similar approach were ambiguous in terms of whether the listener was receiving the speech as intended by the speaker, this study conducted acoustic analysis only on recorded speech that met certain impression ratings.

2. METHOD

2.1. Data collection

Voice performances from 11 young females in their early 20s, who were students of a seiyu course at a music college in Japan, was collected. Their speech utterances were digitally recorded using an audio interface (PerSonus AudioBox iTwo) and a condenser microphone (PreSonus M7). The sampling and quantization bit rates were 44100 Hz and 24 bits, respectively.

Data collection comprised four different tasks structured in the following order: sentence reading (Part I), passage reading, dubbing, repeating, and sentence reading (Part II). Next, 15-minute informal interviews were conducted to collect tokens in a natural setting. This study only reports the results for the passage reading and repeating tasks.

For passage reading, the speakers read "The North Wind and the Sun" by performing four different characters (plus with the speaker's own voices). The characters were of the same types as the anime characters they were to perform in the dubbing and repeat tasks. However, at this stage, the characters were simply designated as a gentle adult woman or older sister ("ordinary"), a strong, bold type like a woman warrior ("brave"), a sexually neutral or immature girl ("boyish"), and a kawaii girl ("cute"), without mentioning the title of the anime and the names of the characters.

In the repeating task, the speakers imitated the performance of professional seiyus in several scenes from an anime immediately after hearing them. The anime for this task was *Kono Subarashii Sekai-ni Shukufuku-o!* ("God's Blessing on this Wonderful World!"), released in 2016. Six scenes from this anime were selected, and the statements spoken by the four characters in these scenes were analyzed. The targeted characters corresponded to the four character types defined in the passage reading task: Aqua (hereafter, "AQ," ordinary), Darkness ("DN," brave), Megumin ("MG," boyish), and Yunyun ("YY," cute).

2.2. Impression ratings

Three listeners (research assistants who did not know the anime characters) annotated the perceived character impressions, from the voices of all 11 speakers. The character impressions were graded over three levels (1 = slightly, 2 = clearly, 3 = strongly) for each of the four character types. Data where two or more annotators assigned level 2 or above to the target character, were used for acoustic analysis.

In the passage task, the number of speakers analyzed based on impression ratings was as follows: ordinary: 9 speakers, brave: 3 speakers, boyish: 8 speakers, cute: 9 speakers.

In the repeating task, the data of the utterances with very strong emotional expressions, when the impressions of the characters changed, were omitted from acoustic analysis.

2.3. Acoustic parameters

F0 and voice-quality-related parameters were extracted. Whereas a previous study analyzed acoustic parameters related to formants and spectral slopes [3], this study focused on harmonicity analysis.

F0 was extracted by filtering the maximum peak of the autocorrelation function (ACF) of the speech signal filtered through a low frequency band of 100–1200Hz. This band guaranteed the presence of one or more harmonics in this range, even for very high pitch voices. The ACF of the speech signal without traditional LPC inverse filtering is more robust, as anime voices include very-high-pitch voices. F0 was extracted at 32-ms frame lengths, with a 10-ms frame step. Subsequently, it was converted to the log scale with semitone intervals, according to:

$$F0[semitone] = 12 * log_2(F0[Hz]) + c \qquad (1)$$

where c = -36.376. Hence, $F0[semitone]$ = 57 corresponds to 220Hz (A3 on the musical scale) [9].

For voice quality, we tested several parameters related to harmonicity. By considering that spectral harmonicity vary at different frequency bands depending on vocal tension (stiffness), we analyzed sub-band harmonic-to-noise ratio (HNR).

The sub-bands were divided as follows: sub-band 1 (100–1200 Hz), 2 (1000–2200 Hz), 3 (1800–4000 Hz), 4 (4000–6000 Hz), 5 (5500–7500 Hz).

Sub-band HNR was computed in dB using an autocorrelation-based method [10], with adaptation to sub-band computation, according to:

$$HNR[k] = 10 * log_{10}\left(\frac{r[k][\tau_{max1}]}{1-r[k][\tau_{max1}]}\right) \qquad (2)$$

where k is the sub-band number (1 to 5), τ_{max1} is the lag of the maximum peak in the ACF of sub-band 1, and $r[k][\tau_{max1}]$ is the autocorrelation value of sub-band k at lag τ_{max1}, normalized by $(T - \tau_{max1})/T$, where T is the window length.

Note that τ_{max1} of sub-band 1 is used to compute the HNR of all the sub-bands assuming that harmonic components in all frequency bands are compatible with the harmonicity of the fundamental frequency obtained from the sub-band 1 peak. However, a small range around τ_{max1} was searched to account for slight deviations between the ACF peaks in different bands. Accordingly, HNR[k] will have positive values if harmonic components are predominant, and negative values if noise components are predominant in each sub-band k. Thus, high HNR values in high-frequency bands are expected for voices with high vocal tension/stiffness.

The mean values of each parameter were extracted as representative values for each utterance, and the mean values of the utterances for each character type performed by each speaker were used as representative values for each speaker.

3. RESULTS

Fig. 1 shows the scatter plots of the average F0 and HNR values of each speaker for each target character in the repeating task. The upper panel shows the distributions for HNR[3] (1800–4000Hz), whereas the lower panel shows the distributions for HNR[5]

(5500–7500Hz). The small red circles indicate the data of the professional seiyus (extracted from the original DVDs). The results for HNR[2] and HNR[4] were omitted because of redundancies.

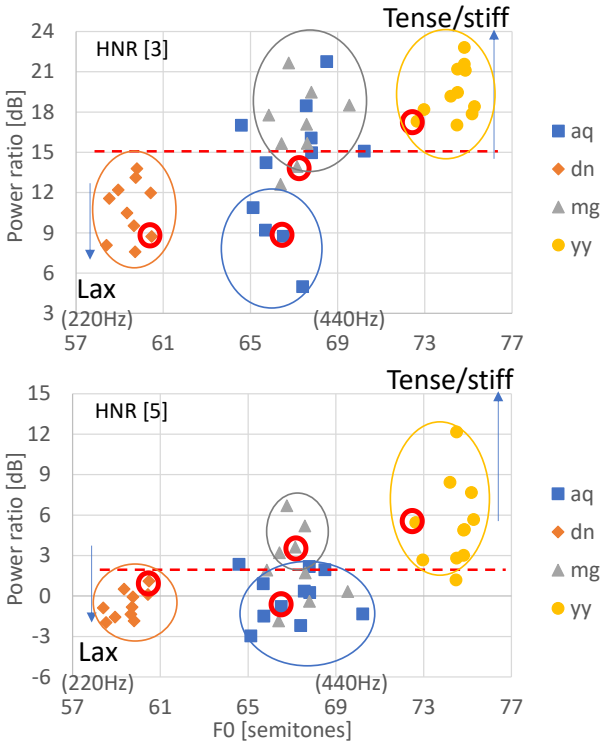


Figure 1: Distributions of the average F0 and HNR values (HNR[3] and HNR[5]) for each target character in the repeating task.

For each acoustic parameter, between-subject ANOVA was conducted with the speakers as random variables and character types as independent variables. Although all speakers performed all character types, a between-subject design was adopted as some samples were omitted through the impression ratings in Section 2.2. Multiple comparisons were performed using the Bonferroni method. Significant main effects were found for all parameters: $F(3,33) = 273.75$, $p < 0.01$, for F0; $F(3,33) = 13.42$, $p < 0.01$, for HNR[3]; and $F(3,33) = 12.6$, $p < 0.01$, for HNR[5].

From Fig. 1, DN and YY can be observed to show distinct distributions in F0, HNR[3] and HNR[5] (DN < YY; $p < 0.05$). The distributions of AQ and MG have some overlap, but MG tends to have a tenser voice than AQ, as observed in HNR[3] (AQ < MG; $p < 0.05$). However, the distributions in HNR[5] show that YY has a tenser voice quality than the other characters (YY > MG,DN,AQ; $p < 0.01$).

The results for the professional seiyus (small red circles in Fig. 1) indicate that the voices imitated by the students were close to the mean values of the seiyus.

Fig. 2 shows the distributions of passage reading task. The target characters were “brave”, “boyish”, “ordinary”, and “cute”. The distributions for “self” indicate the reading style in the speaker’s natural voice.

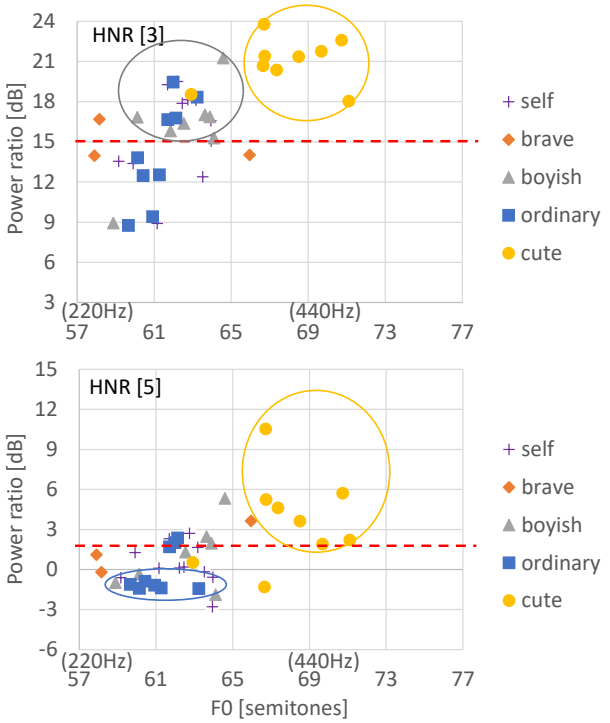


Figure 2: Distributions of the average F0 and HNR values (HNR[3] and HNR[5]) for each target character in the passage reading task.

Between-subject ANOVA and multiple comparisons were conducted similar to that done for the repeating task. Significant main effects were found for all parameters: $F(4,35) = 14.5$, $p < 0.01$, for F0; $F(4,35) = 6.18$, $p < 0.01$, for HNR[3]; and $F(4,35) = 3.75$, $p < 0.05$, for HNR[5].

The distributions in Fig. 2 differ from the results of the repeating task. Overall, the F0s in the passage reading task (Fig. 2) are lower than those in the repeating task (Fig. 1). This is because utterances in the repeating task were mostly accompanied by emotional expressions (such as surprise and excitement), whereas sentences in the passage reading task were mostly emotionless.

However, some similarities can be observed in the HNR[3] and HNR[5] distributions between some voice types. For example, the distributions of “cute” shows similar trends with YY, with a high-pitched tense voice (YY > MG,AQ,DN; $p < 0.05$). The distributions of “boyish” are also similar with MG character, i.e., mostly higher HNR[3] values and mostly lower HNR[5] values. However, the distributions between “ordinary” and AQ are less similar. A possible reason is that “ordinary” may not

be an appropriate attribution for AQ. In fact, in the preliminary analysis, utterances in the repeating task were mostly perceived as “cute girl” rather than “adult female.” Regarding “brave”, data from only three of the speakers could be used for the acoustic analysis, of which two had very low F0s as in DN, while one of the speakers was perceived as “brave+boyish,” with higher F0 and tenser voice quality. Finally, regarding the distributions of “self”, most of the speakers have voices closer to “ordinary” or “boyish”. This is also in agreement with subjective impressions.

4. DISCUSSION AND CONCLUSION

4.1. Character types and acoustic parameters

A previous study on anime character types identified significant differences for F0 but not for HNR [3]. In contrast, the present analysis successfully detected between-character differences for HNR as well. This is because we conducted sub-band HNRs instead of full-band HNR in this study. Another reason would be that we conducted the acoustic analysis after impression ratings.

Character classification in this study was comparable to that in Kawahara's study [4]. He compared the F0s of two characters, "moe" and "tsun," and their normal voices. According to him, "moe" can be characterized by a set of adjectives like "pretty, cute, … and cheerful", whereas "tsun" was characterized as "beautiful, tall, … and inaccessible". Given that "moe" coincides with "cute" and "tsun" overlaps with "brave" in this study, the high and low F0 values for "cute" and "brave", respectively, are consistent with the results of Kawahara's study.

A variety of acoustic parameters related to voice quality were analyzed by Starr [7]. However, directly comparing Starr's character classification with that used in this study is difficult. She compares “sweet” and “nonsweet” voices in Japanese anime, where sweet voice characters include mothers, older women, etc. Although her classification is different from ours, it has a similar implication to this study in that it shows that pitch and voice quality are related to distinction in characters.

Regarding the interpretation of the results for HNR[3] and HNR[5], which reflect the levels of harmonicity components in the mid (1800–4000Hz) and high (5500–7500Hz) frequency bands, it is important to clarify that they can be affected by different factors. First, tense voices tend to have stronger harmonicity components in higher frequency bands than lax voices. Second, vocal tract length may also affect harmonicity in different frequency bands. Longer vocal tracts (male speakers) tend to have lower formant frequencies, and consequently weaker harmonicity components in higher frequency bands compared with shorter vocal tracts (females and children). The HNR[5] differences between the distributions of “cute” and “boyish” could be because both are tense qualities but “cute” is generated by raising the larynx (i.e. shorter vocal tract length), in comparison to “boyish”. Finally, HNR is also strongly affected by breathy or whispery voices, which contain higher levels of turbulent noise in mid- and high-frequency components. In the analyzed dataset, breathy and whispery voices were locally observed primarily at the end of phrases. However, they were not strongly reflected in the analysis in this present study, since mean values at utterance level were used. Such local breathiness is an important factor in emotional expression; therefore, a more detailed analysis is required in future investigations.

4.2. Consistency of distribution across speakers

Notably, voice characteristics of the participants were distributed close to each character. Such a distribution in the repeating task may not be surprising as the participants were asked to listen to actual anime voices and replicate them. However, this trend was also observed to some extent in the passage reading task, wherein participants were not asked to imitate a specific character. This suggests that anime character types are socially shared.

However, the results for the passage reading task were not as clear as those of the repeating task. In addition, many impressions were not judged as intended during impression ratings and were excluded from analysis. This may be either because the passage reading task was difficult for the participants, who were still in the process of training to become seiyus, or because some of the characters are difficult to be expressed through a reading task. In the passage reading task, the participants had to pronounce the passage based only on the instruction without professional model voices.

Finally, the balance of the emotions included in the utterances differed depending on the anime character type. Therefore, distinguishing character differences from differences in emotion in this study is difficult, and separating these two factors will be the subject of a future study.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP21K18116 and JP22H04875. We would like to thank KADOKAWA and Konosuba Partners for their cooperation.

6. REFERENCES

[1] Ishida, M. 2020. *Anime to Seiyu no Mediashi: Naze Josei ga Shonen o Enjirunoka*. Tokyo: Seikyusha.

[2] Iwata, M. 2017. Seiyu-do. Tokyo: Chuokoron-shinsha.

[3] Ota, I., Ota, Y., Utsugi, A. 2021. The voice quality of anime as a sociocultural register. A paper presented at the 16th International Conference of the European Association of Japanese Studies.

[4] Kawahara, S. 2016. The prosodic features of the "moe" and "tsun" voices. *Journal of the Phonetic Society of Japan* 20 (2), 102–110.

[5] Marushima, A. 2020. Josei seiyu ni yoru yakugara no seibetsu no kotonaru onsei no onkyoteki tokucho: Kihonshuhasu ni chakumoku shite. *Osaka-keizaihokadaigaku Ronshu* 115, 23-33.

[6] Teshigawara, M. 2003. Voices in Japanese animation. Doctoral dissertation, University of Victoria.

[7] Starr, R. L. 2015. Sweet voice: The role of voice quality in a Japanese female style. *Language in Society* 44, 1–34.

[8] Utsugi, A., Wang, H., Ota, I. 2019. A voice quality analysis of Japanese anime. *Proc. 19th ICPhS* Melbourne, 1853–1857.

[9] Ishi, C.T., Ishiguro, H., Hagita, N. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531-543.

[10] Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sample sound. *IFA Proceeding* 17, 97-110.

Stop voicing in two settler communities on the Canadian Prairies

Alexandra Pfiffner, Nicole Rosen

University of California, Berkeley, University of Manitoba
apfiffner@berkeley.edu, nicole.rosen@umanitoba.ca

ABSTRACT

This study investigates stop voicing in two rural Canadian communities representing differing ethnoreligions: Manitobans of Ukrainian descent from the INTERLAKE region and of Low German-speaking Mennonite descent from the SOUTHERN region. Results show stop voicing differences between the two communities and a degree of prevoicing not normally found in North America English, typically described as contrasting short-lag and long-lag VOT [1]. Furthermore, Interlake speakers show more word-final voicing of their voiceless stops, and Southern speakers do not devoice, suggesting that German substrate effects seen in analogous communities in the US [2,3,4] do not continue across the border into Canada. Overall, Canadian Prairies speakers show a different stop voicing pattern than currently described in North American varieties. This is explained in part by the block settlement patterns of the region, where European immigrants of differing descent were settled into ethnically-determined villages during the main period of settlement, 1890-1940.

Keywords: Stop voicing, VOT, Canadian English, sociophonetics.

1. INTRODUCTION

1.1 The Canadian Prairies

While Canadian English has been an area of sociophonetic interest for several decades, very little work has focused on the speech of rural areas of the country, and even less in the rural areas of the Canadian Prairies. These regions, however, are of particular interest due to the settlement patterns that differed from those in the East. While most of Ontario and the Maritime provinces were settled by Anglophones, the Prairies saw a large wave of other Europeans such as Ukrainians, Poles, Icelanders, and Mennonites immigrating beginning in the 1880s, as Anglophone immigrants were not numerous enough for the Canadian government to populate the West. At the time, particular social groups were considered to have 'superior potential,' and special privileges were granted to them as they immigrated to the Prairies in large numbers, attracted by the possibility of free land to work. Groups were settled in ethnic blocks scattered around the Prairies, and these blocks remain relevant today. [5]

The Interlake region under discussion here are what were marginal agricultural lands west of Lake Winnipeg, settled by some Poles, Métis or French, but primarily Ukrainians over a few decades starting in the 1890s. Thanks to the block settlement, Ukrainian language, food and customs were maintained into the middle of the 20th century, at which point the English language had won out.

The Southern region was settled at a similar time, by Low German-speaking Mennonites seeking religious freedom. Low German in this region seems to have persisted somewhat longer than Ukrainian did in the Interlake, but in both regions, young people are learning English as a first language and no longer speak Low German, although it is often still heard in the community. Even today, however, older residents in these areas mostly speak Ukrainian or German as a first language, even if they rarely use it anymore.

1.2 Stop voicing patterns

In North American English, initial voiced stops are typically described as having no glottal pulsing and a short-lag voice onset time (VOT), while voiceless stops have long-lag VOT [1]. Recent sociophonetic studies have found that some regions show prevoicing in word-initial stops, such as in regions of Mississippi and Alabama in the US [6,7] and in Winnipeg, Canada [8]. If there are substrate effects in the rural communities in this study, we may expect to see differences in VOT, as Ukrainian has prevoicing [9] while Low German does not [10].

Word-final voicing in North American English varies by dialect, with high amounts of glottal pulsing in voiced stops reported in Southern and Western dialects [11,12], and low amounts of glottal pulsing reported in Midwestern dialects [2,3,4]. Furthermore, Purnell et al. [2,3] and Pfiffner [4] discuss German substrate effects contributing to the final stop neutralization found in Wisconsin and Minnesota English, as German has final stop neutralization [2]. This shows up not just in glottal pulsing, but also in subsegmental measurements, as shorter closure and

burst durations are cues to voicing, and longer closure and burst durations are cues to voicelessness. Given the block settlement pattern in Canada and the recent switch to English as a lingua franca, we may expect to find similar substrate features in the speech of Southern residents of Mennonite descent. This may contrast with Interlake residents of Ukrainian descent, as Ukrainian does not have final stop neutralization [9]. The interest of these groups is then twofold: to first document stop voicing patterns in two rural areas of the Prairies to see whether they follow similar patterns as elsewhere in Canada, and to investigate whether substrate influences of Ukrainian or Low German are evident in the local speech patterns.

2. METHODS

2.1. Speakers

21 residents of the Southern region and 15 residents of the Interlake region participated in the study. Age and gender distribution are given in the tables below:

	Interlake	**Southern**	**Total**
Youngest (b 86-00)	7	5	**12**
Middle (b 63-76)	6	6	**12**
Oldest (b 25-56)	8	4	**12**
Total	**21**	**15**	**21**

Table 1: Participant demographics

Speakers were also coded for (binary) gender and socioeconomic status. Details of the demographics are outlined in Rosen [13].

2.2. Materials

A list of 210 words based on Boberg [14] and Wassink [15] was elicited as part of the Languages in the Prairies Project. All words with stops in word-initial and word-final position in this word list were analyzed in this study.

2.3. Procedure

Participants were recruited and recorded between 2015-17 by two research assistants, each local to the area where they conducted the recordings. Recordings were made in participants' homes using an H4N Zoom recorder with external lavalier microphones. These recordings included a word list, two reading passages and a sociolinguistic interview that was transcribed in Elan [16]. Only the word list was used in this analysis.

2.4. Analysis

Word lists were force-aligned using the Montreal Forced Aligner [17] and stop boundaries were hand corrected in Praat [18]. For word-initial stops, the VOT was measured. Word-finally, the duration of the closure and burst were measured as well as the glottal pulsing to closure duration ratio ('voicing ratio'). After excluding tokens that were unclear, lenited, or had background noise, we had 2,503 word-initial stops and 2,951 word-final stops.

Linear mixed-effects models in R [19] were fitted in a step-up-step-down procedure to predict VOT, closure duration, burst duration, and voicing ratio. Fixed effects included underlying voicing, plosive, ethnicity, gender, age group and first language. Random effects were word and speaker.

3. FINDINGS

We present the findings of word-initial stops first in 3.1, then examine word-final stops in 3.2.

3.1. Word-initial stops

We found a substantial amount of negative VOT in initial voiced stops, with speakers in the Interlake region in particular prevoicing significantly more than those in the Southern region ($p=0.03$), as seen in Figure 1. Approximately 51% of voiced stops were produced with prevoicing in the Interlake region compared to 31% of stops in the Southern region. There were no significant differences in the voiceless initial stops (Figure 2).

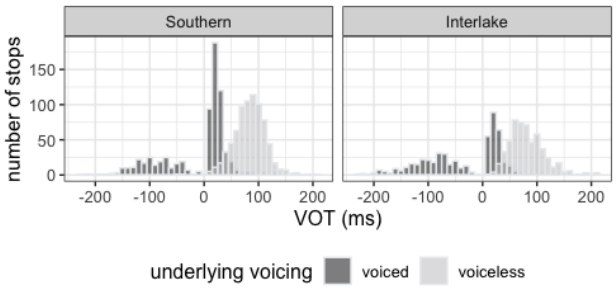


Figure 1: VOT measurements from both communities

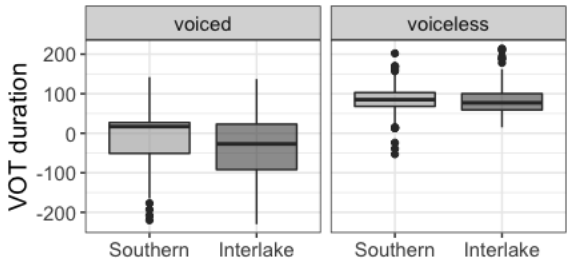


Figure 2: Distribution of VOT measurements by voicing

No other demographic factors tested were significantly predictive of VOT of initial stops. Additionally, there were no significant factors in predicting the bimodal distribution of negative and positive VOT values within the underlyingly voiced stops.

3.2. Word-final stops

Word-finally, underlyingly voiced stops showed no significant differences between communities, while underlyingly voiceless stops did pattern differently. Closure and burst durations, as well as the voicing ratios, are given in Table 2.

Voiced stops	**Closure**	**Burst**	**Voicing ratio**
Interlake	65	72	90%
Southern	75	96	79%
Voiceless stops	**Closure**	**Burst**	**Voicing ratio**
Interlake	99	97	38%
Southern	121	122	18%

Table 2: Word-final measurements

In underlyingly voiceless stops, speakers in the Interlake region had significantly shorter closure (p=0.007) and burst (p=0.024) durations in comparison to speakers in the Southern region. Additionally, Interlake speakers showed significant differences in voicing ratio based on first language; L1 English speakers had significantly higher voicing ratios than L1 Ukrainian speakers (p=0.004). L1 English speakers on average voiced approximately 47% of the closure duration in underlyingly voiceless stops (Figure 3).

Interlake speakers in particular had high amounts of 'voicing bleed,' a term used to describe a pattern of glottal pulsing (possibly, but not necessarily, from a preceding sonorant) into a stop closure [20]. The voicing ratio in Table 2 shows us that this 'bleed' is very high in comparison to those reported in Wisconsin [2,3] and Minnesota [4].

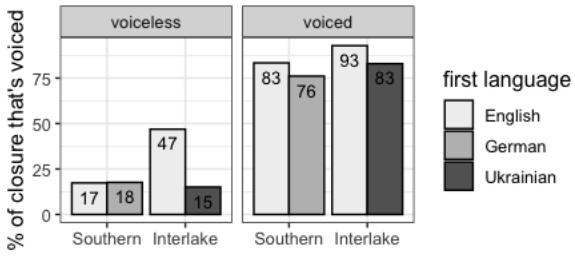


Figure 3: Glottal pulsing during closure by L1

We now turn to discussing the implications of these findings.

4. DISCUSSION

There are three main generalizations to be made from our findings. First, both communities display high levels of prevoicing and voicing bleed in comparison with other reported varieties of English. Second, within these highly voiced varieties, there were significant differences in stop voicing between the Manitoba Interlake and Southern regions. Lastly, the locus of this variation was different: word-initially, voiced stops were variably prevoiced. Word-finally, voiceless stops were produced with more voicing than expected.

We suspect that the tendency towards more voicing may be a substrate effect from Ukrainian, where /b, d, g/ are produced with voicing in all positions [9]. Note however though that there was no evidence of a German substrate effect of word-final stop voicing neutralization in either community. Given that nearby German communities in Wisconsin and Minnesota do display word-final voicing neutralization [3,4,5], we had hypothesized that the Southern community would likewise exhibit devoicing, but that was not the case. This means that there are cross-border differences in how German substrate is realized in English.

5. CONCLUSIONS

Overall, stops are strongly voiced in the Canadian Prairies. Within the Prairies, however, the two regions under investigation are consistently statistically different from each other. Both the Interlake and Southern regions under study are rural, agricultural, and settled within similar time periods, and so the primary difference between the two is the settlers' origins. These findings tell us that there are important differences between rural areas and that a closer look at these areas is important for our understanding of linguistic patterning.

6. REFERENCES

[1] Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

[2] Purnell, T., Salmons, J., & Tepeli, D. (2005). German substrate effects in Wisconsin English: Evidence for final fortition. *American Speech, 80*(2), 135-164.

[3] Purnell, T., Salmons, J., Tepeli, D., & Merced, J. (2005). Structured heterogeneity and change in laryngeal phonetics. *Journal of English Linguistics, 33*(4): 307-338.

[4] Pfiffner, A. (2021) Cue-Based Features: Modeling Change and Variation in the Voicing Contrasts of Minnesotan English, Afrikaans, and Dutch. Doctoral dissertation. Georgetown University.

[5] Gagnon, Erica (2022). Settling the West: Immigration to the Prairies 1867 to 1914. Available at: https://pier21.ca/research/immigration-history/settling-west-immigration-to-prairies.

[6] Hunnicutt, L. & Morris, P. A. (2016). Prevoicing and aspiration in Southern American English. *Proceedings of the 39th Annual Penn Linguistics Conference, 22*:215-224.

[7] Herd, W. (2020). Sociophonetic voice onset time variation in Mississippi English. *JASA, 147*(1): 596-605.

[8] Li, L. (2021). A sociophonetic study of Filipino English in Winnipeg, Canada. Doctoral dissertation. University of Manitoba.

[9] Pompino-Marschall, B. Steriopolo, E. & Żygis, M. (2017). Ukrainian. *Journal of the International Phonetic Association, 47*(3): 349-357.

[10] Jessen, M. & C. Ringen. (2002) Laryngeal features in German. Phonology 19: 189-218.

[11] Jacewicz, E., Fox, R. A., & Lyle., S. (2009). Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association, 39*(3):313-334.

[12] Podesva, R., Eckert, P., Fine, J., Hilton, K., & Jeong. S. (2015). Social influences on the degree of stop voicing in inland California. *University of Pennsylvania Working Papers in Linguistics, 21*(2):167-176.

[13] Rosen, N. (2019). Applying the Nunaliit framework to the visualization of complex linguistic variation 3rd edition, in *Further Developments in the Theory and Practice of Cybercartography: International Dimensions and Language Mapping*, ser. Modern Cartography, D. R. F. Taylor, E. Anonby, and K. Murasugi, Eds.San Diego, CA. Elsevier Press, 2019, vol. 9, pp. 461–474.

[14] Boberg, C. (2008). Regional phonetic differentiation in standard Canadian English. *Journal of English Linguistic, 36*(2): 129–154

[15] Wassink, A. (2023). The Pacific Northwest English study," [Online]. Available at: https://zeos.ling.washington.edu/PNWEnglish/

[16] Wittenburg, P., H. Brugman, A. Russel, A. Klassmann & H. Sloetjes (2006). "ELAN: A professional framework for multimodality research [software application]." In *Proceedings of LREC 2006*, Fifth International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, The Language Archive

[17] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*, pp. 498-502.

[18] Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer (Version 6.3.02). [Computer program]. http://www.praat.org

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[20] Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics, 54*: 35-50.

Acknowledgements: This work draws on research supported by the Social Sciences and Humanities Research Council and the Canada Research Chairs Programs.

FRONT VOWEL PATTERNING IN THE INTERLAKE REGION OF MANITOBA, CANADA

Nicole Rosen & Lisa Sullivan

University of Manitoba
nicole.rosen@umanitoba.ca, lisa.sullivan@umanitoba.ca

ABSTRACT

This study investigates vowel patterning in the Interlake region of Manitoba, a rural community in the Canadian Prairies. Results suggest the Low Back Merger Shift is in progress: middle age speakers retract /æ/ and /ɛ/ while young speakers lower /ɪ/. However, contrary to the shift, the young speakers are fronting /æ/. Young and non-professional speakers also had significant overlap between /e/ and /i/. Furthermore, /æg/-raising is present, particularly in middle-and-young speakers and non-professionals. The recurring pattern of younger and non-professional speakers as drivers of change in rural communities suggests that more research in rural areas is important for our understanding of language change across different communities.

Keywords: Canadian English, Low Back Merger Shift, phonetics, sociophonetics, Pre-velar raising

1. INTRODUCTION

While Canadian English has been an area of sociophonetic interest for several decades, very little work has focused on the speech of rural areas of the country, and even less in the rural areas of the Canadian Prairies. These regions, however, are of particular interest due to the settlement patterns that differed from those in the East. The Interlake region discussed here is what were marginal agricultural lands west of Lake Winnipeg, settled primarily by Ukrainians and Icelanders over several decades starting in the 1890s.

The goal of this paper is to document the phonetics of the speech of the region. Anecdotally, rural Prairies speech is distinctive, especially as compared to urban speech, but no phonetic studies have been conducted to quantify these differences. This work is situated within a recent sociolinguistic literature (e.g. [1, 2]) conducted in rural areas, which find social meanings and identities to be developing differently than in the urban areas most often studied.

In this paper, we focus on the three features which emerge from the data: low back merger shift, overlap between /e/ and /i/, and pre-velar /æ/ raising. The Low-Back-Merger Shift (LBMS) [3] is a widespread pull-chain shift in North America said to have occurred due to instability in the English short vowel system whereby the front vowels retract and lower. /æ/-lowering is the first step in the pull-chain, with /ɛ/ and /ɪ/ normally following in turn. Pre-velar /æ/-raising (/æg/-raising) is a process where /æ/ is raised before /g/. It has been acoustically documented to occur across Canada (e.g. [4]), as well as in the Pacific Northwest (e.g. [5], Upper Midwest (e.g. [6]) and California [7] regions of the US. Pre-velar raising of /ɛ/ also occurs before /g/ (e.g. [8]), and some have proposed that these vowels are merging with each other and /e/ in this context in some regions [5]. In addition to these well-known phonetic features, we found substantial overlap between /i/ and /e/, particularly among the youngest speakers in the corpus.

2. METHOD

2.1. Speakers

25 residents of the Interlake region participated in the study. Age and gender distribution are given in Table 1. Speakers were also coded for socio-economic status (professional vs non-professional; hereafter SES). For more details, see Rosen [9].

Age Group	Year of Birth	Male	Female
Younger	1990-2000	1	5
Middle	1963-1972	4	4
Older	1925-1956	4	7
Total		**9**	**16**

Table 1: Participant Age and Gender

2.2. Stimuli

A word list based on Boberg [4] and Wassink [10] which included vowel tokens comprising the full vowel system of Canadian English was used.[1]

2.3. Procedure

Participants were recruited and recorded in 2019 by a student research assistant local to the area. Recordings were sociolinguistic interviews made in participants' homes using an H4N Zoom recorder with external lavalier microphones. The interviews included one repetition of the word list.

2.4. Analysis

We limit the current analysis to the word list data (172 to 296 tokens per model). Recordings were force aligned in FAVE [11] and manually corrected in Praat [12]. F1 and F2 measurements were taken at the midpoint of each vowel and z-score normalized. /e/-/i/ overlap was measured using Pillai scores [13].

Linear regression models were constructed in R [14] using either the *lm*() function (/e/-/i/ overlap) or the *lmer*() function (low back merger shift, /æg/-raising) from the lme4() package [15]. Mixed effects regression models were constructed with F1 or F2 as response variables for the low back merger shift vowels (/ɪ ɛ æ/), and with F1 as the response variable for /æg/-raising. A multivariate model with Pillai score as the response variable was constructed for /e/-/i/ overlap. All models included age, gender, and socioeconomic status as predictor variables. The models for /æg/-raising also included coda consonant (/g/ vs non-/g/ obstuents) and its interaction with the other three variables. Mixed effects models included random intercepts for participant and item.

Gender, coda consonant and socioeconomic status were simple coded with female, non-/g/ obstruents and non-professional as -0.5 and male, /g/ and professional as 0.5, respectively. Age was coded for comparisons between older (+) and middle (-) and between middle (+) and younger (-).

3. RESULTS

In this section we examine our findings for front vowel patterning among Interlake speakers.[1] A summary of significant effects from the regression models can be found in Table 2.

3.1. Low Back Merger Shift

The Interlake speakers appear to participate in the LBMS, as the vowel plots in Figure 1 show. Looking at each of the vowels in the LBMS individually, we see that for /æ/, the middle age group is statistically more retracted than the older group, while the younger group retracts less than the middle group.

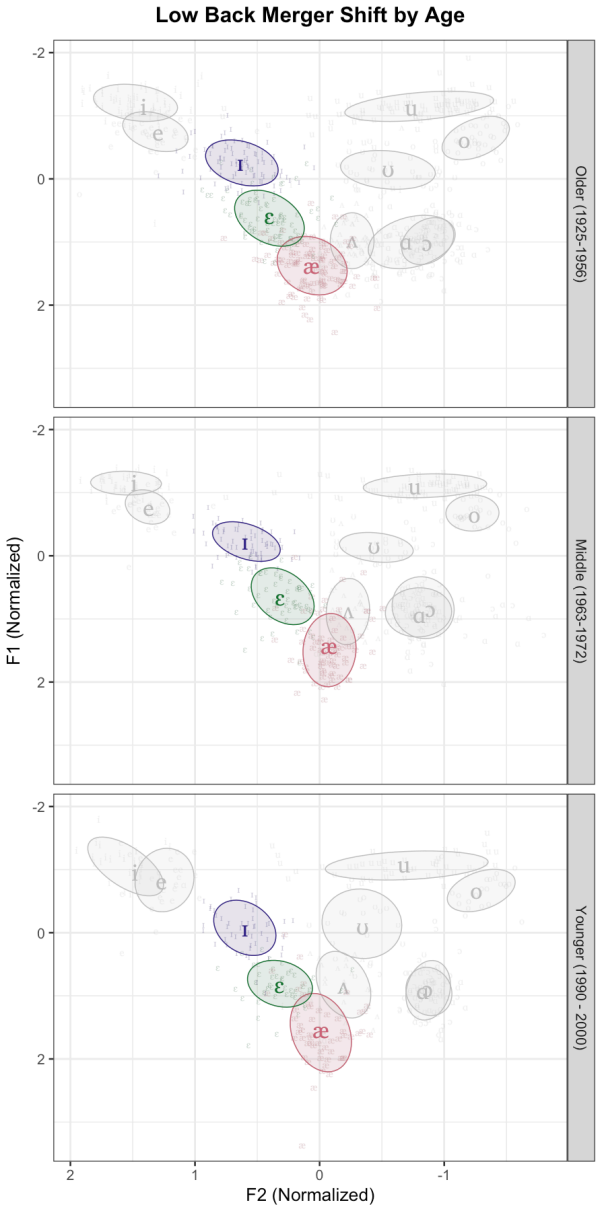


Figure 1: Low Back Merger Shift by Age

Women also retract more than men. For /ɛ/ and /ɪ/, we find non-significant trends in that the middle age group retracts /ɛ/ more than the older group while the younger group retracts /ɪ/ more than middle age group. Socioeconomic status did not have a significant effect on participation in the LBMS.

3.2. Overlap of /e/ and /i/

Interestingly, this analysis found a large degree of overlap between /e/ and /i/ for some Interlake speakers. This seems to be predictable by age, where the youngest speakers have the highest degree of overlap as seen in Figure 2, though this difference is only significant at the 0.1 level. Non-professional speakers also appear to have more overlap than

Variable		Predictor/ Interaction	Est.	*SE*	t	p	
LBM Shift	ɪ F1	*Age* (*Y*(-) *v M*(+))	-0.13	0.08	-1.73	0.1	
	ɛ F2	*Age* (*M*(-) *v O*(+))	0.11	0.05	1.96	0.07	.
	æ F2	*Gender* (*M* (*0.5*) *vs F* (*-0.5*)	0.15	0.04	3.51	0.002	**
		Age (*Y v M*)	-0.13	0.056	-2.27	0.03	*
		Age (*M v O*)	0.17	0.047	3.64	0.001	**
BAG-raising (F1)		*Coda* (*g*(*.5*) *v Obs*(*-.5*))	-0.40	0.22	-1.80	0.08	.
		*Coda * Age* (*M v O*)	0.32	0.12	2.77	0.006	**
		*Coda * SES* (*NP*(*-.5*) *v P*(*.5*))	0.23	0.11	2.11	0.04	*
e-i (Pillai)		*Age* (*Y v M*)	0.25	0.13	1.86	0.08	.

Table 2: Significant and Trending Effects

professional speakers (Figure 3), though, unlike age, SES was not a significant predictor in the regression model.

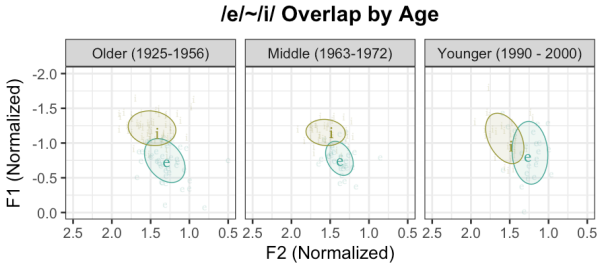


Figure 2: /e/ - /i/ Overlap by Age

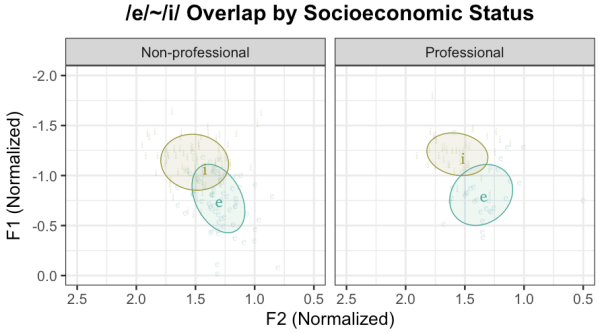


Figure 3: /e/ - /i/ Overlap by SES

3.3. Pre-velar /æ/-raising

Figures 4 and 5 show the Interlake speakers' production of /æg/, /ɛg/, /æ/ and /ɛ/ broken down by participant age and socioeconomic status, respectively. Both /ɛg/ and /æg/ appear to be raised relative to /ɛ/ and /æ/ for all groups of speakers, but they remain distinct from each other. The statistical analysis shows two patterns for /æg/-raising: the middle age group raises more than the older group, as seen in Figure 4, and non-professionals raise more than professionals, as in Figure 5.

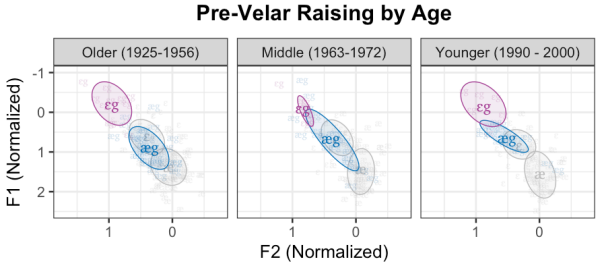


Figure 4: /æg/-Raising by Age

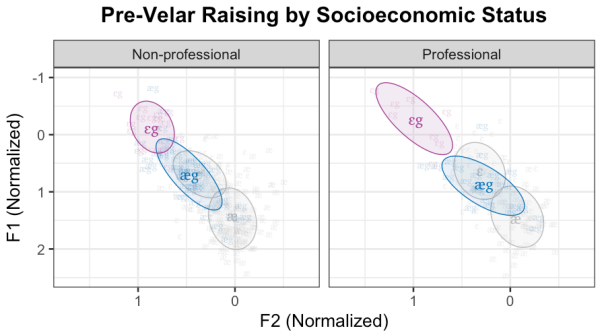


Figure 5: /æg/-Raising by SES

4. DISCUSSION

4.1. Low back merger shift (LBMS)

Apparent-time analysis suggests that Interlake speakers participate in the LBMS mostly as would be expected. Middle-aged speakers retract /æ/ more than the older speakers, and women retract more than men; both differences suggesting a change in progress towards more retraction and reflect studies elsewhere (e.g. [16, 17, 18]). Furthermore, the next two vowels in the pull-chain trend similarly, with middle age retracting /ɛ/ more than the older speakers, and the younger speakers retracting /ɪ/ more than the middle generation. We interpret this to mean that /ɪ/-shifting has just begun in this youngest Interlake generation.

Of particular interest, however, is the fact that the youngest speakers appear to be initiating a reversal of the LBMS, with /æ/ in a more front position than for older speakers. This effect could be due simply to the low number of younger speakers in the sample. However, given the gender imbalance within the younger speakers skewing female, and the tendency for women to drive change, we expect that

the fronting found within our data to be reflective of change presently occurring within the community. This reversal of the retraction found elsewhere could be interpreted as moving away from 'city' pronunciation. This interpretation becomes more interesting as we see similar effects in /e/-/i/ overlap (4.2) and /æg/-raising (4.3).

4.2. /e/-/i/ overlap

The degree of overlap found between /e/ and /i/, to our knowledge, has not been reported elsewhere, but, there is some preliminary evidence supporting /i/-/e/ overlap as a locus of variation in the area. Onosson [19], for example, found that Manitobans of Filipino descent had a more raised and front /e/ than those of Mennonite descent. Similar overlap of high and mid vowels was also found in the Metis French spoken in the area [20], attributed to influence of the nearby Algonquian languages.

It is interesting to note that it is the youngest speakers with the most overlap, and so this may be evidence of a change in progress. Such an overlap has not been observed in 18 to 35 year olds in Ontario or Colorado [18], whose degree of overlap is similar to that of the older and middle-age Interlake speakers. Although we cannot speak definitively on /e/-/i/ overlap as a change in progress, this is an area worth developing for future research. It is furthermore interesting to note that this overlap is predictable not only by age but by SES, where younger and non-professional speakers are overlapping more than older and professional. This pattern is one we will return to in the next section.

4.3. Pre-velar /æ/-raising

Overall, there is substantial /æg/-raising among the speakers in the region, consistent with findings elsewhere in the Canadian West [21, 22, 23, 24, 19, 25, 26]. On the other hand, we do not find the BEG-BAG merger or overlap documented by Freeman [5], and there is less overlap than in Onosson [19]'s findings among the urban Filipino Manitobans.

Recall that while both /ɛg/ and /æg/ are raised relative to /ɛ/ and /æ/ for all groups of speakers, the statistical analysis shows two /æg/-raising patterns: the middle and younger speakers raise more than the older group, as in Figure 4, and non-professionals raise more than professionals, as in Figure 5. This is the same patterning together of younger and non-professional speakers as we just saw with the /e/-/i/ overlap. This change towards raising is consistent with Swan [26]'s work in Vancouver, and Rosen & Skriver [25]'s in Southern Alberta, both of which showed more raising among younger speakers. However while the Interlake pattern is consistent with a change occurring in apparent-time, beginning with the middle age group, the socio-economic differences are particularly intriguing.

Linguistic changes in progress are often associated not only with younger speakers in apparent-time, but also women, and professional women in particular, with the explanation that young upwardly mobile women have particular incentive to adopt changes as a way to gain status in the linguistic marketplace. However in this case, there is a) no gender effect, and b) younger speakers and non-professionals are /æg/-raising more. Although more research is needed, we suggest that this may be a rural effect as found in Stanley [2]'s study of /æg/-raising in rural Cowlitz County, Washington, in the Pacific Northwest. Stanley found that 'higher tokens of BAG are correlated with degree of connectivity and positive feelings about the Pacific Northwest while lower tokens are a result of time away from the community and more education' ([2]:142). It may be that /æg/-raising is doing similar social work in the Interlake region, an area where it is not a social advantage to talk like 'city people.' In the Interlake case, we find that non-professionals are raising more than professionals, suggesting that it may be developing into a local rural marker, differentiating from the city. More research is necessary to determine whether this is the case, however.

5. CONCLUSION

Overall, we found that while Interlake speakers participate actively in the LBMS, there appears to be a reversal of this shift at the beginning stage among the youngest speakers in the sample. In addition, we found a significant degree of overlap of / e/ and / i/ not previously reported for Canadian English, and lastly, that there is a pattern of æg-raising in the community.

Of particular interest is the pattern of movement away from upwardly-mobile urban features found across the three features. Specifically, the non-professional and younger age groups are patterning together for both æg-raising and /e/-/i/ overlap. When added to the potential movement away from the LBMS among the younger speakers, this recurring pattern is worthy of further investigation to find out whether the drivers of change in rural communities are different than in the urban communities more commonly studied in sociolinguistics.

6. REFERENCES

[1] R. J. Podesva, A. D'onofrio, J. Van Hofwegen, and S. K. Kim, "Country ideology and the california vowel shift," *Language Variation and Change*, vol. 27, no. 2, pp. 157–186, 2015.

[2] J. A. Stanley, "Changes in the timber industry as a catastrophic event: BAG-raising in Cowlitz County, Washington," *University of Pennsylvania Working Papers in Linguistics*, vol. 24, no. 2, pp. 137–146, 2018.

[3] K. Becker, *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and Short Front Vowel Shifts across North America*. American Dialect Society, 2019.

[4] C. Boberg, "Regional phonetic differentiation in standard Canadian English," *Journal of English Linguistics*, vol. 36, no. 2, pp. 129–154, 2008.

[5] V. Freeman, "Vague eggs and tags: Prevelar merger in Seattle," *Language Variation and Change*, vol. 33, no. 1, pp. 57–80, 2021.

[6] E. Koffi, "The acoustic vowel space of central Minnesota English: Focus on female vowels," *Linguistic Portfolios*, vol. 2, no. 1, pp. 2–16, 2013.

[7] L. Esposito and E. Lake, "Complicating prevelar raising in the West," *American Speech*, pp. 1–45, 07 2021.

[8] J. A. Stanley, "Regional patterns in prevelar raising," *American Speech*, vol. 97, no. 3, pp. 374–411, 2022.

[9] N. Rosen, "Applying the nunaliit framework to the visualization of complex linguistic variation 3rd edition," in *Further Developments in the Theory and Practice of Cybercartography: International Dimensions and Language Mapping*, ser. Modern Cartography, D. R. F. Taylor, E. Anonby, and K. Murasugi, Eds. San Diego, CA: Elsevier Press, 2019, vol. 9, pp. 461–474.

[10] A. Wassink, "The Pacific Northwest English study," 2023. [Online]. Available: https://zeos.ling.washington.edu/PNWEnglish/

[11] I. Rosenfelder, J. Fruehwald, K. Evanini, S. Seyfarth, K. Gorman, H. Prichard, and J. Yuan, "Fave (forced alignment and vowel extraction," software application suite, 2014.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.1.38," 2021. [Online]. Available: http://www.praat.org/

[13] J. A. Stanley and B. Sneller, "Sample size matters in calculating pillai scores," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 54–67, 2023.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://R-project.org/

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] C. Boberg, *The English language in Canada: Status, history and comparative analysis*. Cambridge University Press, 2010.

[17] C. Holland and T. Brandenburg, "Beyond the front range: The Coloradan vowel space," *Publication of the American Dialect Society*, vol. 102, no. 1, pp. 9–30, 2017.

[18] L. Sullivan, "Pre-velar /æ/-raising in Ontario and Colorado English: Production, perception and metalinguistic awareness," Ph.D. dissertation, Univesrity of Toronto, Toronto, ON, 2022.

[19] S. Onosson, "Prevelar vowel raising and merger in Manitoba English," *Journal of English Linguistics*, vol. 50, no. 3, pp. 247–280, 2022.

[20] N. Rosen and E. Lacasse, "Une comparaison des voyelles postérieures du français mitchif et du français franco-manitobain," in *À l'ouest des Grands Lacs: communautés francophones et variétés de français dans les Prairies et en Colombie-Britannique*, R. Papen and S. H. Bres, Eds. Presse de l'Université Laval, 2014.

[21] J. H. Esling and H. J. Warkentyne, "Retracting of/æ/in Vancouver English," in *Focus on Canada*, S. Clarke, Ed. John Benjamins, 1993, vol. 11, pp. 229–246.

[22] J. Jones, "I bag your pardon: The Albertan æ/ɛ vowel shift as a window into community grammars," Master's thesis, University of Calgary, 2015.

[23] G. Mellesmoen, "Front vowel phonology of British Columbia English," Master's thesis, University of Toronto, Toronto, ON, 2016.

[24] ——, "A remedial path to merger: Merger by phonological transfer in British Columbia English," *Toronto Working Papers in Linguistics*, vol. 40, no. 1, 2018.

[25] N. Rosen and C. Skriver, "Vowel patterning of mormons in southern Alberta, Canada," *Language & Communication*, vol. 42, pp. 104–115, 2015.

[26] J. T. Swan, "Canadian English in the Pacific Northwest: A phonetic comparison of Vancouver, BC and Seattle, WA," in *Proceedings of the 2016 annual conference of the Canadian Linguistic Association*, 2016.

[1] For a complete wordlist and full size figures, see lisasullivan.ca/lipp/icphs2023

Acknowledgements: This work draws on research supported by the Social Sciences and Humanities Research Council of Canada and the Canada Research Chairs Program. We wish to thank Jessica Padre for assistance with the acoustic analysis.

How gender affects immigrants' accent variation—An acoustic study on Hunan immigrants in Beijing

ZOU Ziyan[1], YOU Shuxiang[2]
[1]Chinese University of Hong Kong, [2]Macau University of Science and Technology
ziyanzou@link.cuhk.edu.hk, sxyou@must.edu.mo

ABSTRACT

This study examines accent variation among Hunan immigrants in Beijing with a focus on gender. Using the picture elicitation method, we compared the quantitative and acoustic differences of *er* suffixed words produced by immigrants of different genders in two production tasks. Our results show that female immigrants produce significantly more *er* suffixed words than male immigrants. Acoustically, female immigrants also differ from female Beijing locals in their retroflexion, as evidenced by lower F2 and F3 mean values of the same *er* suffix. We suggest that this accent shift may be due to the fact that female immigrants who are subject to social pressure from both their feminine and immigrant social status seek to assert themselves in the current patriarchal society by adopting the language of the place where they have immigrated. This study sheds light on the complex interplay of gender, social status, and language use in immigrant communities.

Keywords: *er* suffixation; gender; sociolinguistic; accent variation

1. INTRODUCTION

The inquiry into gender-related linguistic behaviours constitutes a salient and extensively debated topic within the field of sociolinguistics. Previous studies in sociolinguistics have posited that women exhibit a propensity for employing the prestigious variant of language, albeit lacking a cogent definition of what precisely constitutes "*prestigious form*". In order to comprehensively investigate the veracity of such claims and reveal the underlying drivers of potential accent shifts, this sociolinguistic study undertakes a focused production task of Hunan immigrants in Beijing. By analysing accent variation in words suffixed with "*er*" produced by immigrants of two genders through quantitative and acoustic analyses, this study endeavours to offer insights into the relationship between gender, language use, and sociocultural dynamics in the context of immigration and language contact.

2.LITERATURE REVIEW

2.1. Gender and the prestige form

Early studies on phonological gender differences have affirmed the tendency of women to be sensitive to the prestige form, and they shared the same conclusion that women are more conservative in their use of language and their usages are closer to the standard form which is used by the upper class in society.

Trudgill et al.[1] studied the use of *ng* in different classes in Norwich. The variable *ng* had two different phonologically free variants which are the prestige form [ŋ] and lower-class form [n]. The results showed that in formal contexts, women used the prestige form more than men, regardless of class. Also, when asked what form of speech they were using, women were more likely to report that they were using a form with more prestige attributes. Similar results were found in Macaulay [2] on different phonologically free variants of /i/ in Glasgow English. He found that people of lower social status had a more backward and lower tongue position when pronouncing the /i/ sound than people of higher social status. However, within the same social class, women pronounced it higher and more forward than men.

Both of the above studies provide evidence from the phonological perspective. Eisikovits' [3] studies about the speech of adolescents in Sydney provided us with evidence from grammar aspects. Eisikovits found that there were three ungrammatical usages in teenager speech, which were (1) Non-standard past tense form; (2) Multiple negations; (3) Invariable *don't*. Her results showed in each non-standard grammatical usage, a higher percentage was found in boys' speech in their daily lives than girls.

One can argue based on these results that it is the desire for being prestige that drives women to change their accents. However, some opposite linguistics strategies taken by females have been found in the literature.

Labov's [4] Philadelphia data showed that women lag in the use of variants that were widely used in the larger community. In order to explain this phenomenon, Labov proposed a higher level of the concept of prestige. He believed that two prestige forms should exist, a global prestige form on a larger scale, and a local prestige form confined to the local area. As the local culture fluctuated, some of the accents that were once popular and representing the standard may have become stigmatized in today's

society. And because the world's cultural mobility did not keep pace with local change, phonological forms that were considered backward at the level of local prestige were still compliant with global prestige. Nevertheless, the motivation where came from Philadelphia women used stigmatised variants also can be interpreted as not pursing the prestige form.

Other scholars have also offered their views by explaining that women's usage of stigmatised variants was motivated to gain more social benefits. For example, Eckert [5] suggested that not the prestige but the power is the most appropriate underlying sociological concept for the analysis of gender-based linguistic variation. Later in her long-term studies [6] she did show us a new way of thinking.

At a high school in Detroit, Eckert studied the students' accents. The majority accent was still American English. However, Eckert noticed that two free phonological variants were developing. One was backing of front vowel, and the other was raising of low nuclei. Students in this high school could also be separated into two groups generally, 'jocks' and 'burnouts'. The jocks engaged in school activities passionately, aiming at getting into a good university, whereas the burnouts focused more on outer-school affairs and aimed at finding jobs after finishing high school education. It turned out that the usages of these two variants were polarised in two groups. There was a higher percentage of girls in jocks than boys using the prestige form while female burnouts tended to have a heavier non-standard accent than male burnouts.

Based on this, Eckert argued that woman may not always choose the prestige form of language. Then she tried to account for the polarised usage by arguing that females concerned more with category membership. Therefore, girls must rely more on symbolic manifestations for social membership than boys. This motivated girls to use more symbolic linguistics forms. This interpretation also can fit the situation in the study of Clader et al [7]. This study examined /s/ variation in two African-American (AA) communities: one was AAs as the majority, the other was not. The results indicated that there was no gender difference in /s/ articulation among AA majority area, but a gender pattern existed among AA minority community, with women being significantly more fronted than men.

As to why women are more concerned with the social category to which they belong. Eckert provided an explanation by introducing the concept Bourdieu's Linguistic Market [8]. Eckert speculated that it is the result of women's powerlessness within society that plays a role. She argued that because society denied women the opportunity to advance their power or status through the marketplace, they were forced to rely on their symbolic resources, including language, to enhance their social position. Linguistic market was defined as a symbolic market. It was the market that can legitimate all kinds of linguistic features. Diverse linguistic features were unequally contributed to this market. It was the social culture that determined each feature's value. The most valued features were at the central position of the market. And those who held legitimate linguistic features can benefit from the market.

So far as reviewed, women did not simply pursue for the prestige form. The choice of different accents was a self-empowerment strategy adopted by women on their path to power. We can see that the studies reviewed above on women's linguistic strategies mainly focused on natives', but how about the situation in an immigrant society? Would females take more actions to adopt the most beneficial linguistic form than males?

Mendoza-Denton [9] argued that speakers were able to negotiate an identity that transcends ethnicity and interacts with gender and place-based notions of identity. In an immigrant society, the language of the immigrated place can play a role of benefit form for all immigrants. Through mastering the language of the immigrated place, one can be treated like a native who comes with a number of advantages in life and respect for social status. In order to better observe how the accent of the immigrant is changed, Hunan immigrants in Beijing were recruited, as the native language of the immigrants, namely the Xiang dialect, differs significantly from the Beijing dialect. Moreover, a crucial difference between these two dialects laied in the use of *er* suffixation—while *er* suffixation seldom appeared in Xiang dialect, it was an outstanding characteristic in Beijing Mandarin.

Therefore, we assumed that compared to immigrant males, immigrant females are more likely to use words suffixed with "*er*" in their daily conversations.

3. METHODOLOGY

To elicit the participants' *er*-suffixation production in natural spontaneous speech, two "storytelling with pictures" tasks were designed. As word frequency affects the rate of speech, and the more familiar a word is, the faster it can be extracted from the mind Oldfield & Wingfield [10], we randomly selected 12 familiar words with *er*-suffixation from the 60 "required *er* suffixed words" mentioned in Peng [11] as our target words. The participants were expected to produce these target words when they described the pictures and completed the storytelling tasks.

3.1. Participants

Thirty participants (15 males and 15 females) were recruited and divided into three groups. Group A

consisted of ten subjects (5 males and 5 females) from Miluo, Hunan Province; Group B consisted of ten subjects (5 males and 5 females) from Miluo, Hunan Province, who immigrated to Beijing and have been living in Beijing for 6-7 years; Group C consisted of ten local Beijing locals (5 males and 5 females).

3.2. Production test

3.2.1. Task 1(See in Figure 1,Appendix)

Scenario: Two little cats are describing "Specialties in Beijing". There are five specialties listed in the pictures and each of the specialty contained an *er* suffixation word.

Target words: 猫儿（catty) [mɑɚ]; 胡同儿 (HuTong) [tʰũɚ]; 吹糖人儿(suger-made people) [ɹəɚ]; 豆汁儿(soybean juice) [tʂəɚ]; 捏面人儿(rice cracker-made people)[ɹəɚ]; 四合院儿(SiHeyuan) [yəɚ]

3.2.2. Task 2（See in Figue 2 , Appendix)

Scenario: A little girl was crying under a tree because her schoolbag was hanging on a branch when a little boy came with a long pole to help her get her schoolbag down from the tree.

Target words: 孩儿(kid) [xaɚ]; 枝儿(branch) [tʂəɚ]; 包儿(bag) [pauɚ]; 杆儿(stick)[kaɚ]; 曲儿 (song) [tɕʰyɚ]; 鸟儿(bird) [njɑɚ]

3.3. Occurrence Er suffixation counting and acoustically analysis

To count the *er*-suffixed word, each occurrence of an *er*-suffixed word was assigned a score of one point. In cases where an *er*-suffixed word occurred multiple times within a subject's production, only one score was assigned. It should be noted that words containing the same *er*-suffixation root with different preceding modifiers were considered as different tokens for the purposes of scoring.

For the acoustic analysis of *er*-suffixed words, we chose the top 3 most frequently occurring words in the production of female speakers from Group B, ensuring a sufficient number of tokens for robust analysis. These preeminent *er*-suffixed words were [xaɚ] (little kid) (10 times), [ɹəɚ] (people) (6 times), and [tʂəɚ] (juice) (5 times). We annotated each token of these three words produced by the female speakers in Group B in Praat. Subsequently, for each token, the F2 and F3 formants at 11 equidistant points across their temporal span were extracted. The mean F2 and F3 values for each *er*-suffixed word were obtained by calculating the mean of the group data, and their values were utilised to graphically plot the final formants of females in Group B for subsequent intergroup contrasts. Similar procedures were replicated for females production in Group C where the three most frequently produced words were [xaɚ] 10 times, [ɹəɚ] 10 times, and [tʂəɚ] 5 times. Furthermore, we conducted T-tests to see if statistically significant differences between the groups could be found, using mean F2 and F3 values from Group B and Group C.

4. RESULTS

4.1. Statistical analysis of *er* suffixation

Group A achieved a total score of 15, indicating that the participants produced 15 valid tokens of er-suffixation words. Among Group A members, males scored a total of 11 with an average of 2.2, while females scored a total of 4 with an average of 0.8. In Group B, the total score was 43. The male participants in Group B scored a total of 4 with an average of 0.8, whereas females scored a total of 39 with an average of 7.8. Group C had a total score of 122. In Group C, men scored a total of 58 with an average of 11.6, while women scored a total of 64 with an average of 12.8. (*See Figure 3 & 4, Appendix*)

To investigate whether significant differences existed between various groups, we conducted a one-way ANOVA, which yielded a p-value of <0.01 (F=45.637). Based on this, we can confirm that there were notable disparities in scores across the groups. Subsequently, we performed a post hoc test using multiple comparisons, and the analysis revealed that while the difference in scores between group A and group B was significant (p=0.011), it was not as dramatic as the differences observed between Group C and A (p<0.01), as well as between Group C and B (p<0.01). These findings imply that Hunan immigrants in Beijing tend to employ more *er*-suffixed words in their daily speech than those who have never been to Beijing, likely due to the influence of the Beijing dialect on their accent. However, the number of *er*-suffixed words that Hunan immigrants in Beijing use still falls well short of the figure among the locals, indicating that other factors, such as length of residency in Beijing, might play a role. Overall, we can infer that the accent of Hunan immigrants in Beijing is influenced by the Beijing dialect, but this effect appears to be subject to certain constraints, leading to a nonnative-like performance in *er* suffixation production among this immigrant group. These findings will be further discussed in the subsequent sections.

In terms of within-group differences by gender, only Group B yielded a very significance (F=55.682, p<0.01). It indicated that although the accent of both male and female Hunan immigrants in Beijing is influenced by Beijing dialect, female immigrants are more affected by Beijing accent. Moreover, there is no significant difference in the number of *er* suffixed

words produced by male participants between Group A and Group B (F=1.2, p=0.305>0.05), which revealed the truth that the accent of male Hunan immigrants in Beijing is less influenced by Beijing dialect as compared to the accent of males who have never been to Beijing.

4.2. Acoustic analysis

For [tʂəɚ], there was no significant difference between Group B's F2 and Group C's F2 (t= -0.883, p=0.3>0.05), nor was there a statistically significant difference between Group B's F3 and Group C's F3 (t=0.245, p>0.8>0.05). For [ɹəɚ], both F2 and F3 in Group B were statistically significantly different from F2 and F3 in Group C (t=6.16, p<0.01; t=3.12, p<0.01). For [xaɚ] the results were similar to those for [ɹəɚ], with both F2 and F3 in Group B being statistically significantly different from those in group C (t=2.91, p<0.01; t=2.33, p<0.01).

The reason for the lack of significant difference between Groups B and C for [tʂəɚ] may be because the "soybean juice (豆汁儿)" is a local Beijing word, and for immigrants who have acquired Beijing as a second language, the *er* suffixation is part of the word and not learnt as an affix but an impartible part of it. Therefore, when it comes to the production of [tʂəɚ], immigrants would pronounce it with *er* suffixation unconsciously. Moreover, the better production performance of it may due to the consonant-vowel harmony. Since /tʂ/ itself is a retroflex consonant. (*See Figure 5 in Appendix*)

But why does Group B perform so differently from Group C in [ɹəɚ] but not so different in [tʂəɚ] given that they both contain a retroflex consonant initial? Here we offer two possible reasons for this contrast. Firstly, the absence of /ɹ/ in the consonants inventory of the Miluo dialect itself has a certain impact on the pronunciation for Group B, which was why they pronounced the consonant part of /ɹ/ not as smoothly and natural as /tʂ/. Secondly, in [ɹəɚ], the values of F2 and F3 are consistently lower in Group C than in Group B. This is probably due to the higher degree of retroflexion in Group C. This difference in the degree of retroflexion is hard to perceive, so it created more obstacles for Group B to truly acquire the pronunciation if they simply imitate what they hear.(*See Figure 6 in Appendix*)

For [xaɚ], the spectrum (Figure 7) shows that a sharp drop in F3 in Group C begins earlier than in group B, and the values for both F2 and F3 are lower than in Group B. This indicates that *er* suffixation process begins earlier in the native Beijing locals, and that the retroflexion begins almost as soon as the pronunciation of the initial consonant ends. Moreover, there is no significant difference between the two groups in the pronunciation of the initial consonant, and the significant difference between Groups B and C begins at the moment of vocalisation of the rhymes. The reason for this difference in the pronunciation of the *er* suffixation of the rhymes would be consistent with the results about the acquisition of the *er* suffixation of [ɹəɚ] mentioned earlier.

5. DISCUSSION

The results showed that, quantitatively, there is a large difference between male and female migrants in terms of the number of *er* suffixed words produced. Female immigrants produced more *er* suffixed words (up to one-third of the number of *er* suffixed words produced by native Beijingers), while male immigrants produced almost no *er* suffixed words. This suggested that female immigrants are more likely to take the immigrated place's accents than male immigrants. Acoustically, there is also a large difference in terms of the same *er* suffixed words produced by female immigrants and that produced by native female Beijing locals. Beijing females have a higher degree of retroflex by showing lower F2 and F3 mean values. This also indicated that more efforts were needed for female immigrants to acquire the same manner of articulation about *er* suffixation.

Findings in this study made it clear that the gender differences in immigrants' accent variation do exist. Female immigrants were more likely to use migrated area's dialect as an approach to empowering themselves. This is in line with Eckert's [4] suggestion that women are in an insecure social position in a patriarchal society and therefore use all means possible to enhance their social position. Females in immigrant societies are more motivated to change their accent than males in general societies because they are under pressure from both their immigrant status and famine status. We can use the linguistic market to explain this phenomena: the more native linguistic features and the more masculine linguistic features all participate simultaneously in the linguistic market. A distinction between these two is that the more native linguistic identity has a higher value in this market. As a result, female migrants change their accents in order to gain more capital to retain and enhance their social status. After all, benefits brought up by changing accents come more obviously than talking in the way which males do. Most importantly, the case of *er*-suffixed words we have observed in an immigrant society substantiates the assertion that women, deprived of access to real power, must claim their social status through the use of symbolic features.

6. REFERENCES

[1]Trudgill, P., Trudgill, S., Anderson, S. R., & Huddleston, R. (1974). The social differentiation of English in Norwich (Vol. 13). CUP archive.

[2]Macaulay, R. K. (1978). Variation and consistency in Glaswegian English. Sociolinguistic Patterns in British English, 132-143.
[3]Eisikovits, E. (1987). Variation in the lexical verb in inner-Sydney English. Australian journal of linguistics, 7(1), 1-24.
[4]Labov, W. (2001). Principles of Linguistic Change. Volume II: Social Factors, chapter 10. Social Networks, 325-365.
[5]Eckert, P. (1998). Gender and sociolinguistic variation. Language and gender: A reader, 64-75.
[6]Eckert, P., & McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. Language in society, 28(2), 185-201.
[7]Calder, Jeremy and King, Sharese(2020) "Intersections between Race, Place, and Gender in the Production of /s/," University of Pennsylvania Working Papers in Linguistics: Vol. 26: Iss. 2, Article 5
[8]Bourdieu, P. (1975). Le langage autorisé. Actes de la recherche en sciences sociales, 1(5), 183-190. Mrs. Thatcher interrupted so often? *Nature* 300, 744–747.
[9]Mendoza-Denton, N. (2002). 19 Language and Identity. The handbook of language variation and change, 11, 475.
[10]Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. Quarterly Journal of Experimental Psychology, 17(4), 273-281.
[11]Peng, Z.P. (2004). Beijing Erhuaci Yanjiu. Beijing Language University.

7. APPENDIX

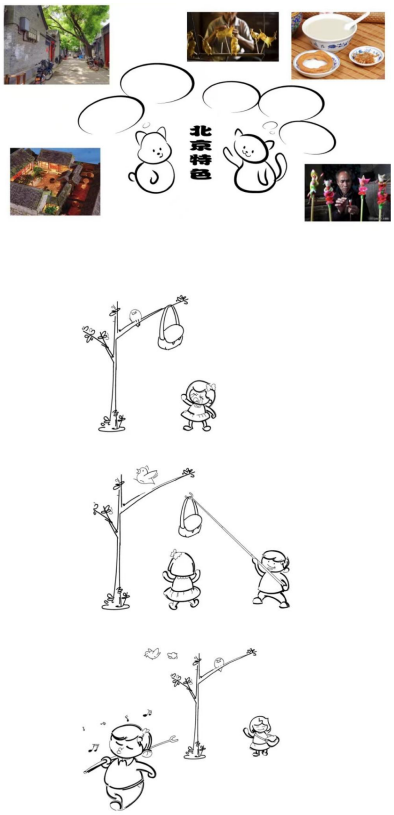


Figure 1&2 : Task pictures

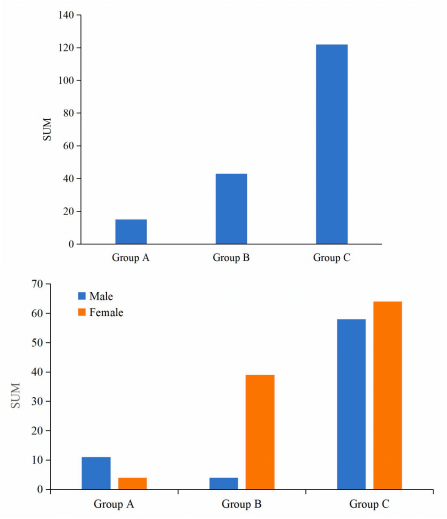
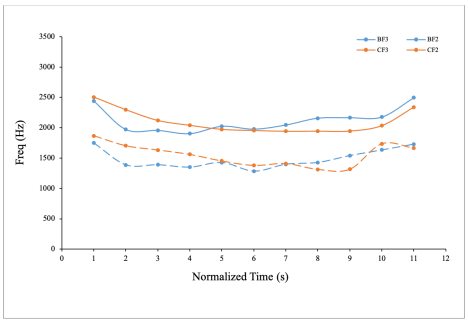
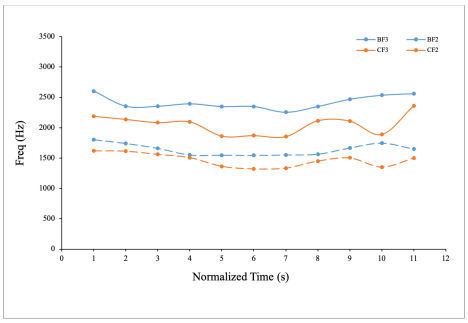


Figure 4&5: Group differences of results





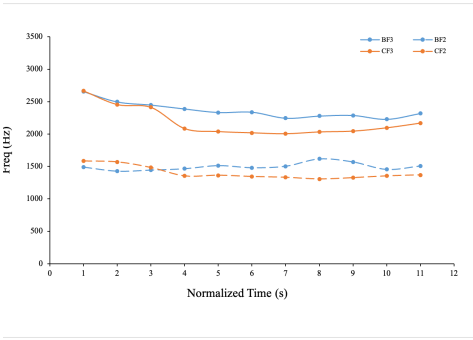


Figure 5, 6 &7: The F2 and F3 of [tʂəɚ]; [ɹəɚ]; [xaɚ]

CONTEXT-INDUCED PHONETIC SHIFT IN SILESIAN CZECH SPEAKERS

Šárka Šimáčková and Václav Jonáš Podlipský

Palacký University Olomouc
sarka.simackova@upol.cz, vaclav.j.podlipsky@upol.cz

ABSTRACT

We assessed the regional accent of young educated urban-dwellers from Czech Silesia and measured their phonetic accommodation in (in)formal situations. Silesian Czech reduces the vowel length contrast existing in Common Czech. We compared the duration of /a:/ produced by 12 Silesians and 6 Common Czech controls, as well as accent ratings of these speakers by 102 Czech listeners, distinguishing between regionally mobile and region-bound Silesians and, for the mobile Silesians, also between their reading to a native-accent peer and to a standard-Czech senior teacher. Both the /a:/ durations and the accent ratings confirmed that Silesian and Common Czech speakers differ. For the mobile Silesians, the results of /a:/ duration revealed short-term accommodation but neither the acoustic nor the accent-rating data provided evidence of long-term change as compared to the region-bound Silesians.

Keywords: Silesian Czech, vowel duration, accent recognition, accommodation

1. INTRODUCTION

Instrumental (socio)phonetic studies of contemporary Czech speech in Moravia and Silesia (the east of the Czech Republic) are sparse [1]. As noted already half a century ago [2], traditional dialectal areas in these regions have been undergoing dialect levelling. The reduction of distinct local features has been shown more recently in a survey of Ostrava teenagers' language use [3]. At the same time, Moravian and Silesian teenagers may prefer (broadly) regional features to Standard Czech and to Common Czech, the prestigious, regionally unmarked inter-dialect widely used in Bohemia and west Moravia [4].

Our study focuses on Czech Silesia (north-eastern Czech Republic), specifically on the Ostrava region. Early industrialization and urbanisation of the area promoted in-migration. Contacts between speakers of different dialects from within and outside Silesia resulted in a reduced differentiation of local dialects and the formation of a Silesian inter-dialect that continues to undergo further levelling [5]. Phonologically, Silesia remains a distinct dialectal area [6], its regional accent being easily recognizable by outside listeners [7]. The features distinguishing Silesians from Czech speakers in other regions include penultimate lexical stress, the absence of contrasting vowel length, the presence of the /i/-/ɪ/ contrast, and progressive voicing assimilation in /tv, sv, kv/ clusters [2, 3, 8, 9]. However, even these features are subject to variation, their instability in contemporary Silesian occasionally commented on (e.g. [10] for penultimate stress). Inter-speaker variation in the Ostrava region seems to be driven, above all, by socioeconomic status and education [5], while an important intra-speaker factor is register: in formal contexts, Silesians tend to switch to Standard Czech [3, 5]. It is the impact of register we are exploring in this study.

In face-to-face verbal exchanges, the interlocutors may adopt phonetic characteristics of each other's speech [11]. According to Communication Accommodation Theory, such adjustments facilitate coherent interaction and allow interlocutors to negotiate social distance [12, 13]. Empirical studies provide evidence of phonetic accommodation during a cooperative goal-oriented exchange [14-16]. However, cooperation may not be a prerequisite for inducing a phonetic shift towards one's interlocutor; e.g. the listeners in [17] needed as little as a visual hint at the speaker's regional background to adjust their perception of vowel sounds.

In our study, we used a non-interactive reading task to elicit speech from young, regionally mobile speakers from Silesia on two occasions: once in the presence of a peer with the same dialect background (Ostrava Mode) and once in the presence of a senior instructor using Standard Czech (Standard Mode). We tested if the speakers would adapt phonetically to the sociolinguistic identity of a passive interlocutor. The study further asked whether the regionally mobile Silesians' accent in the Ostrava Mode would differ from the accent of Silesians who did not leave the region regularly. Since both groups consisted of young university-educated urban dwellers, it was possible that their speech would not contain salient local features. This is why comparison data were collected from a parallel population of speakers of the widely spread Common Czech [18].

We used ratings by listeners as a global measure and vowel duration as a specific acoustic measure of the speakers' accent. The acoustic analysis focused on the vowels /a/ and /a:/. In both Standard and Common Czech phonology, vowel length cuts across

all vowel qualities, producing a vowel system of 5 short and 5 long vowels, including the two low vowels with minimal pairs such as *dal* /dal/ 'he gave' and *dál* /da:l/ 'further', for which duration is the main cue [19]. Typical for Silesian speech is the absence of the vowel-length contrast, with *dal* and *dál* pronounced as homophones.

Our research questions concern (1) the duration of /a:/, described in Standard Czech as a long low central unrounded vowel, and (2) Czech listeners' responses to speech of mobile and non-mobile Silesians and Common Czech speakers. With respect to each point, we ask if there is a difference between (i) Silesians and Common Czech speakers, (ii) regionally mobile and non-mobile Silesians, and (iii) mobile Silesians' speech in the Ostrava Mode and the Standard Mode.

2. METHODOLOGY

2.1. Speakers

The speakers were 18 young adults: 2 groups of 6 Silesians and one of 6 speakers from central or western Bohemia or western Moravia. One Silesian group, the "mobile" Silesian Czech (mSCz) speakers, included 4 women and 2 men (mean age 21.8 years) from Ostrava who pursue academic degrees at Palacký University Olomouc in central Moravia but maintain a frequent contact with their home region, returning to Silesia for the weekends (1 participant even commuting daily). Another 3 men and 3 women from Silesia (4 from Ostrava, 1 from Frýdek-Místek, 1 from Havířov; mean age 23.5 years) could be described as "non-mobile", conducting their day-to-day life within the Silesian region; they were university students or recent graduates (University of Ostrava, Ostrava Technical University). We take them to represent the accent of young educated adult speakers of Silesian Czech (SCz). The third group included 3 men and 3 women (mean age 22.5), also students at Palacký University Olomouc; they were speakers of Common Czech (CCz).

2.2. Production: vowel duration

These speakers read the Czech version of *The North wind and the Sun* [20]. The mSCz speakers were recorded twice in order to elicit performance in different speaking modes. First, they met with a peer with the same native dialect background (a 22-year-old man from Ostrava, Collector 1), who introduced the session by a 5-minute chat about everyday life at home. At least a week later, they were recorded by a university instructor (a 55-year-old woman speaking in Standard Czech, Collector 2), who inquired about their academic work before the recording. Both sessions took place in a recording studio at Palacký University. This venue was also used to record the CCz speakers, interacting with Collector 2. The SCz speakers were recorded in a quiet room at home by a native Silesian (either Collector 1 or a 17-year-old high school student). A Zoom H4n recorder at 16-bit and 44.1 kHz without compression was used for all the recordings. The participants practised the passage and then read it twice.

For the analysis we used 22 words from the first reading. When a token could not be analysed, the same word from the second reading was used. Two types of intervals were manually annotated in Praat [21]: the target vowel, /a(:)/, and a CVC interval comprising /a(:)/ and the immediately preceding and following consonants. No word-initial or -final /a(:)/'s were analysed. The words *pak* 'then' and *tak* 'so', typically following a pause, were excluded as the beginning of the CVC sequence could not be identified. In total, 12 long /a:/'s and 12 short /a/'s were measured per reading. We calculated relative V duration as the V-to-CVC duration ratio to attenuate variability due to reading tempo differences. Although there are other variables affecting V duration in Czech, including word length (4 syllable-words were excluded), a vowel's position in the word, and the word's position within an intonational phrase [22], these were kept constant between participants since the same words and text were used.

2.3. Accent recognition

The sentence that was delivered most fluently by the largest number of participants (*Ujednali tedy, že ten se má považovat za silnějšího, kdo první dokáže, aby si pocestný svlékl plášť.* "They agreed the one who first made the traveller take his cloak off should be considered stronger.") was extracted from the 24 recordings, yielding 6 CCz clips, 6 SCz clips, and 12 mSCz clips (6 per mode). The clips were assigned in a pseudorandomized manner into 6 sets, each comprising 8 items – 2 from SCz and 2 from CCz speakers, and 2 from mSCz speakers performing in the native mode and 2 from other mSCz speakers in the standard mode. Each clip occurred in two sets but only once in each set. Each set was heard by 17 listeners, i.e. each clip was judged by 34 listeners.

In a Praat MFC experiment, listeners heard a clip and responded to the question 'Do you think this person comes from the Ostrava region?' on a 6-point Likert scale (1 = definitely not, 6 = definitely yes). They could replay each sentence twice.

The listeners were 102 young adults, all students at Palacký University (mean age 21.3 years, 80 women, 22 men), out of which 18 came from the Ostrava region. The data was collected in a computer room, and the listeners used circumaural headphones.

3. RESULTS

3.1. Production of /a(:)/ duration

Fig. 1 shows the distributions of the relative durations of the target /a(:)/'s for the CCz and SCz speakers, as well as the mobile SCz speakers performing in the Ostrava Mode (SCz.o). The figure suggests shorter durations of long /a:/ for the SCz speakers than for CCz, while the measured /a:/ durations for the SCz.o speakers seem to be intermediate. Listed in Tab. 1, and plotted in Fig. 2, are the coefficients estimated by a linear mixed model with relative V duration as the response variable, Speaker Variety and Vowel Length as the fixed effects, and Item (varying intercepts) and Speaker (varying intercepts and slopes for Vowel Length) as the random effects (analyses in R [23-26]).

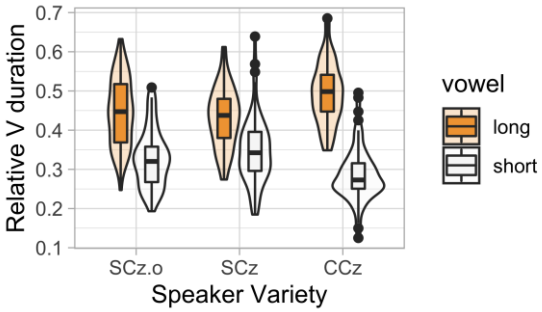


Figure 1: Violin and boxplots of relative durations of the target long and short /a(:)/ split by the speaker's native variety.

	Estimate	SE	df	t	Pr(>|t|)
(Intercept)	0.440	0.020	31.40	22.62	**<0.0001**
varietySCz	-0.015	0.015	17.38	-0.96	0.3500
varietyCCz	0.048	0.015	17.38	3.21	**0.0050**
vowelShort	-0.113	0.024	26.80	-4.68	**<0.0001**
varietySCz:vowelShort	0.034	0.015	17.16	2.25	0.0380
varietyCCz:vowelShort	-0.091	0.015	17.26	-5.99	**<0.0001**

Table 1: Coefficients estimated by a linear mixed model fitted to the relative V durations produced by the speakers in their native varieties.

The analysis showed that the CCz speakers' /a:/ durations were reliably longer (est. +0.048, *SE* = 0.015, $t = 3.21$, $p = 0.005$) than those produced by the mobile SCz speakers in the native Ostrava Mode (SCz.o, the Intercept in Tab. 1), which in turn did not differ reliably from those produced by the non-mobile Silesians (SCz).

We plotted (in Fig. 3) and modelled (with Bonferroni alpha correction) the mobile Silesians' productions in the Ostrava Mode and in the Standard Czech Mode separately, as this is a within-speaker comparison. Fig. 3 shows the distributions of the relative durations of the target /a(:)/ for these speakers. While short /a/ durations did not seem to be affected by Mode, long /a:/ seemed lengthened when produced in the Standard Mode as compared with the Ostrava Mode.

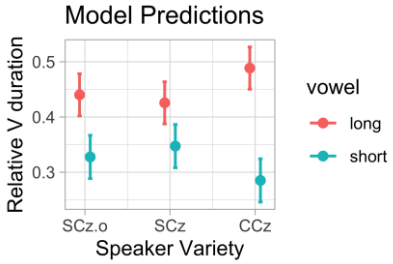


Figure 2: Predicted relative durations of /a(:)/'s produced by the speakers in their native varieties.

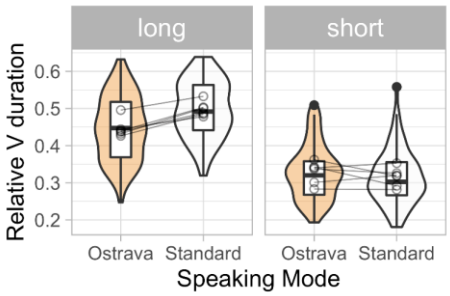


Figure 3: Violin and boxplots of relative durations of the target long and short /a(:)/ for the mobile Silesian speakers split by Mode. Unfilled circles connected by lines show each speaker's means.

Listed in Tab. 2 are the coefficients estimated by a linear mixed model with relative V duration as the response variable and Speaking Mode and Vowel Length as the fixed effects, and Item (varying intercepts) and Speaker (varying intercepts and slopes for Vowel Length) as the random effects. The model confirmed that the bidialectal speakers' lengthening of /a:/ in the Standard Mode was statistically reliable (est. +0.052, *SE* = 0.008, $t = 6.12$, $p < 0.0001$).

	Estimate	SE	df	t	Pr(>|t|)
(Intercept)	0.436	0.021	25.68	21.30	**<0.0001**
modeStandard	0.052	0.008	256.86	6.12	**<0.0001**
vowelShort	-0.109	0.026	25.15	-4.11	**0.0004**
modeStandard: vowelShort	-0.065	0.012	256.86	-5.45	**<0.0001**

Table 2: Coefficients estimated by a linear mixed model fitted to the relative V durations produced by the bidialectal speakers in the two Modes.

3.2. Accent recognition

For each step on the accent recognition scale (where 1 = the speaker is definitely not from the Ostrava region, and 6 = definitely yes), Fig. 4 plots the counts

of responses per Speaker Variety, or in the case of the bidialectal Silesians, Speaking Mode. It can be seen that for CCz the frequency of responses decreases rapidly with the growing scale step, while for SCz it increases (though more gradually), both as expected. For the mobile Silesians, the counts are rather evenly spread across the scale and do not differ much between the Ostrava (SCz.o) and Standard (SCz.s) modes.

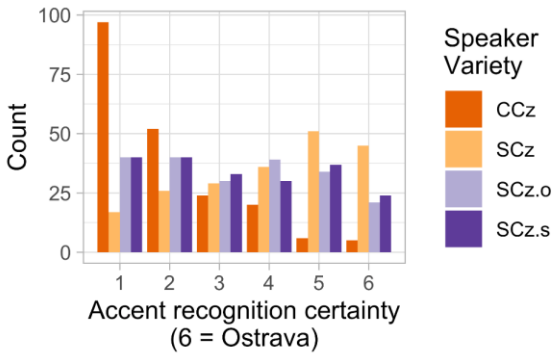


Figure 4: Listeners' accent recognition certainty. Counts of responses per scale step (1 = definitely not from the Ostrava region, 6 = definitely yes) and Speaker Variety / Speaking Mode.

	Estimate	SE	z	Pr(>|z|)
(Intercept)	1.155	0.085	13.57	**<0.0001**
var/modeSCz.s	0.010	0.055	0.17	0.8684
var/modeSCz	0.220	0.119	1.86	0.0635
var/modeCCz	-0.463	0.124	-3.75	**0.0002**

Table 3: Coefficients estimated by a generalized linear mixed model (Poisson family) fitted to the listeners accent recognition certainty scores for the stimuli produced by speakers with the different native varieties and in the different speaking modes.

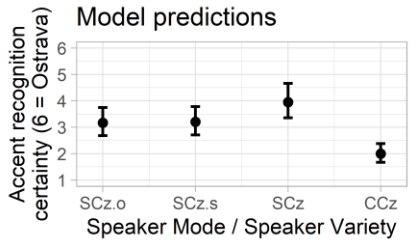


Figure 5: Predicted accent recognition certainty ratings of speakers with different native varieties, and for the mobile Silesians, speaking in the different modes.

Given in Tab. 3 and plotted Fig. 5, are the coefficients estimated by a generalized linear mixed model of the Poisson family with Speaker Variety / Mode as the fixed effect, and Listener and Speaker as the random effects (both with varying intercepts). This analysis revealed a reliable difference of the mobile Silesian speakers' ratings in the native Ostrava Mode (the Intercept in Tab. 3) from the CCz ratings (est. -0.463, *SE* = 0.124, *z* = -3.75, *p* = 0.0002), but not from the mobile Silesians' Standard Mode ratings, or from the non-mobile Silesians' ratings.

4. DISCUSSION

The speech of our (young, educated) Silesian participants shows the hallmark of Silesian Czech, i.e. long vowel shortening. Their "long" /a:/ was shorter than the Common Czech long /a:/. Interestingly, the Silesian "short" /a/'s were longer compared to Common Czech /a/'s. Together, this indicates a reduction of length contrasts in Silesian Czech, vowel duration possibly being more variable due to prosodic factors (word-level stress, intonation phrasing). Yet, the mobile Silesians did differentiate between /a:/ and /a/ (Tab. 1), with the non-mobile Silesians also displaying some, albeit smaller, difference (see Fig. 1). Thus, a systematic investigation of vowel duration in Silesian Czech, accounting for prosody-conditioned variation, needs to follow if the phonological role of vowel length in contemporary Silesian Czech is to be clarified.

Our study did not find evidence of long-term accommodating. The Silesians who regularly spend time away from their home region did not produce reliably longer /a:/ compared to their region-bound counterparts. From Fig. 2 it appears though that the magnitude of the /a:/-/a/ difference, rather than just the duration of /a:/, should be explored next before it is concluded that young mobile Silesian speakers do not experience phonetic drift [27]. However, the mobile Silesians did adjust /a(:)/ durations due to the social setting, producing longer /a:/'s and shorter /a/'s (Tab. 2) in the presence of a Standard Czech speaker. The /a:/ shifts were consistent across speakers, suggesting that V duration is changed dynamically.

In the accent recognition task, the listeners made a global judgement of the phonetic characteristics of the target utterances, reliably distinguishing between the Silesian and Common Czech speakers, which confirms a wide recognisability of the Silesian accent in the Czech Republic. The spread of responses to the Silesian speakers indicates a degree of uncertainty and is possibly due to variation among the Silesians. The counts of endpoint responses on the accent scale were different for the mobile and non-mobile Silesians (Fig. 4), however the model did not predict a reliable difference (Tab. 3). The listeners did not differentiate between the Ostrava and the Standard Mode utterances of the mobile Silesians, suggesting that phonetic adjustments on the part of the Silesian speakers were not sufficient to affect accent perception.

5. ACKNOWLEDGMENTS

This research was supported by Czech Science Foundation grant number 21-09797S.

6. REFERENCES

[1] Uličný, O. 2018. Mluvená čeština na Moravě a ve Slezsku dnes: otázka sociolingvistická [Spoken Czech in Moravia and Silesia: A Soliolinguistic Issue]. In: Bláha, O., Svobodová, J. (eds), *Současná jazyková situace na Moravě a ve Slezsku*. UP v Olomouci, 49–55.

[2] Bělič, J. 1972. *Nástin české dialektologie [A sketch of Czech Dialectology]*. SPN.

[3] Havlásková, P. 2007. *Působení tradičního dialektu na mluvu mladé generace v Ostravě*. PhD dis., Masaryk university Brno.

[4] Krčmová, M. 1977. Jazyk mládeže na Moravě [Language of young people in Moravia]. *Naše řeč*. 59, 225. http://nase-rec.ujc.cas.cz/archiv.php?art=5980.

[5] Šrámek, R. 1997. Jak se proměňují slezská nářečí [How Silesian dialects change]. In: Daneš, F., Bachmannová, J., Čmejrková, S., Krčmová, M. (eds), *Český jazyk na přelomu tisíciletí*. Academia, 231–238.

[6] Palková, Z. 2018. *Variabilita krátkých vokálů z hlediska kvality* [Variability of short vowels wrt quality]. In: Bláha, O., Svobodová, J. (eds), *Současná jazyková situace na Moravě a ve Slezsku*. UP v Olomouci, 153–166.

[7] Drkošová, S. 2013. *Jazyková nerovnost a jazyková diskriminace v české společnosti: Ostrava vs. Praha [Linguistic Inequality and language discrimination in the Czech Society]*. BA dis., Charles University Prague.

[8] Blažková, J. 2008. *Výslovnost dvojího" i" ve východolašském dialektu [Pronunciation of two i's in the Eastern Lachian dialect]*. MA dis., Charles University Prague.

[9] Kuldanová, P. 2018. Ke zvukové stránce současné češtiny na Ostravsku [On the Sound of Contemporary Czech in the Ostrava Region]. In: Bláha, O., Svobodová, J. (eds), *Současná jazyková situace na Moravě a ve Slezsku*. UP v Olomouci, 144–152.[10] Janáková, B. 2015. *Mluva mladší generace obyvatel ve Frýdku-Místku [The speech of younger generation in Frýdek-Místek]*. BA dis., Charles University Prague.

[10] Janáková, B. 2015. Mluva mladší generace obyvatel ve Frýdku-Místku [The speech of younger generation in Frýdek-Místek]. BA dis., Charles University Prague.

[11] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* 119(4), 2382-2393.

[12] Giles, H., Smith, P. M. 1979. Accommodation theory: Optimal levels of convergence. In: Giles, H., St. Clair, R., (eds), *Language and Social Psychology*. Oxford: Blackwell 45–65.

[13] Dragojevic, M., Gasiorek, J., & Giles, H. 2016. Accommodative Strategies as Core of the Theory. In: Giles, H. (ed), *Communication Accommodation Theory*. CUP, 36–59.

[14] Aubanel, V., Nguyen, N. 2020. Speaking to a common tune: *Between-speaker convergence in voice fundamental frequency in a joint speech production task*. *PloS one*, 15(5), e0232209.

[15] Schweitzer, A., Wokurek, W., Pützer, M. 2019. Convergence of harmonic voice quality parameters in spontaneous dialogues. In *Proceedings 19th ICPhS* 363–367.

[16] Šturm, P., Skarnitzl, R., & Nechanský, T. 2021. Prosodic Accommodation in Face-to-face and Telephone Dialogues. *Proc. Interspeech 2021*, 1444-1448.

[17] Hay, JB & K Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48(4). 865-892.

[18] Krčmová, M. 2017. Obecná čeština. [Common Czech]. In: Karlík, P. Nekula, M. Pleskalová J. (eds), *CzechEncy – Nový encyklopedický slovník češtiny.* https://www.czechency.org/slovnik/OBECNÁ ČEŠTINA

[19] Podlipský, V. J., Chládková, K., Šimáčková, Š. 2019. Spectrum as a perceptual cue to vowel length in Czech, a quantity language. *J. Acoust. Soc. Am.* 146(4), EL352-EL357.

[20] Šimáčková, Š., Podlipský, V. J., Chládková, K. 2012. Czech spoken in Bohemia and Moravia. *J. Int. Phon. Assoc.* 42(2), 225-232.

[21] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.16. http://www.praat.org/

[22] Spina, N., Schubö, F. 2021. Vowel length affects pre-boundary lengthening in Czech. *1st International Conference on Tone and Intonation (TAI) 2021, Sonderborg, Denmark, December 6-9, 2021, Proceedings,* 166–170.

[23] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL https://www.R-project.org/.

[24] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67, 1-48.

[25] Singmann, H., et al. 2022. Afex: Analysis of Factorial Experiments. R package version 1.1-1. https://CRAN.R-project.org/package=afex

[26] Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

[27] Chang, C.B. 2019. Phonetic drift. In Schmid, M. S, Köpke, B. (eds), *The Oxford handbook of language attrition*, 191–203. OUP.

Connected Speech Strategies at Word Boundaries: Taps and Glottal Stops in RP

Ana M. Sánchez-Leira

University of Murcia
anamaria.sanchez17@um.es

ABSTRACT

This study examines the different connected speech strategies at word boundaries in potential t-tapping cases in a non-rhotic variety of English. A number of interviews from a talk show broadcast on BBC One were analysed auditorily and acoustically for the occurrence of potential t-tapping contexts at word boundaries. The results indicate that t-tapping is not very frequent in spontaneous speech at word junctures. The results also show that t-tapping is favoured over the production of glottal stops, whilst word-final /t, d/ are more frequent than tap productions.

Keywords: t-tapping, glottal stops, word boundaries, connected speech

1. INTRODUCTION

Connected speech phenomena refer to modifications in the pronunciation of phonological units during the speech chain [1]. Phonological modifications may be subject to the linguistic context, speech rate or register. Speech perception concerns the interaction between incoming and stored information [4]. [12] suggested that pronunciation is strongly determined by either habit or mental storage of words rather than speed rate or formality.

T-tapping is one of the many reduction processes that has been considered to be a naturally and phonetically motivated sound change [7]. From an articulatory perspective, t-tapping may be considered a natural type of sound change since it might be derived from the common tendency of articulatory ease [7]. T-tapping could also be considered a phonetically motivated process given the principle of lexical diffusion [13] since t-tapping affects high-frequency words first [7].

The connected speech phenomenon of t-tapping implies the pronunciation of an intervocalic /t, d/ as a voiced tap realisation instead of a plosive consonant. It can occur in three different phonetic contexts: within words, word-internally at morpheme boundaries and externally across word boundaries (e.g. *pretty, pudding, I don't*).

Glottal stops refer to a type of laryngeal strategy whereby a complete or partial obstruction of the airflow occurs at the glottis [2, 10]. When a plosive is replaced with a glottal stop, this type of articulation is usually referred to as t-glottaling or glottal replacement (e.g. *get it*) [2].

2. WORD BOUNDARIES STRATEGIES IN CONNECTED SPEECH: THE CASE OF T-TAPPING

While t-tapping and t-glottaling have long been observed and described in the specialised literature, t-tapping remains scarcely investigated in non-rhotic accents of English, such as RP, where it may function as a competing variant to avoid the negative connotations associated to the glottal stop [2].

As a case in point, most studies in the field have investigated and described t-tapping in rhotic varieties of English [3, 5, 6, 11, 15] as well as in non-rhotic varieties in the southern hemisphere [8, 14]. However, few empirical studies have focused on the production of taps across accents in the British Isles remains scarcely investigated [2, 7].

The findings obtained from these studies indicated that taps were highly frequent and were the preferred allophonic variant of medial /t/ over glottal stops in American English [11, 14, 17].

Concerning t-tapping and t-glottaling, some of the studies found that glottalised variants seemed to be competing with tap productions in intervocalic final position in Australian English [14].

Some empirical studies addressed the acoustics of taps, which were characterised for being acoustically short [17], whilst other studies accounted for the distribution of allophonic variants of /t/ [5].

The production t-tapping has generally been correlated to linguistic factors such as morpheme boundaries, as it has been claimed to be more frequent in morphologically simple words, usage-based factors such as highly frequent lexical representations and as lexical frequency, since its rate of occurrence was higher in high-frequency words and high-frequency phrases [3, 11, 15]. However, phonetic factors such as stress and vowel quality may also condition its variability [6, 15].

Sociolinguistic factors were also considered to play a crucial role in the ratio of production of t-tapping in some varieties of English, such as Australian English, where younger speakers exhibited the highest production of t-tapping over t-glottaling [14].

In New Zealand English, t-tapping production followed similar trends to the factors affecting its variation in other varieties of English, such as American or Australian English. For instance, empirical studies revealed that t-tapping use has increased over time and it affected high-frequency words first [8].

Empirical studies on RP found that young speakers exhibited an increasing use of taps because of the introduction of phonological innovations from non-standard variants in younger generations of RP speakers. In [2], 317 out of 5,248 tokens that have been analysed corresponded to tap realisations in the following phonological environments: word-final prevocalically, word-medial intervocalically and word-medial before syllabic /l/ contexts, which were found to occur principally in informal speech while they rarely occur in formal speech. This study on RP revealed that the highest influence of production for taps are medium-frequency words, and high-frequency items containing /t/ in word-medial position. Lexical frequency did not have any impact on word-final position /t/ in potential t-tapping sites [2]. In terms of sociolinguistic factors, age, gender and socioeconomic factors played a role in affecting t-tapping variation. Adults' utterances contained more instances of taps (i.e. 8.1%) in comparison to teenagers' utterances (i.e. 4.2%), whereas men tended to produce more taps (i.e. 7.1%) than women (i.e. 4.8%). The production of taps in [2] was also analysed concerning the type of school of the participants. The highest rates of production corresponded to boarding school speakers (i.e. 9.6%), followed by comprehensive school speakers (i.e. 5.1%) while the lowest rates corresponded to private non-boarding school speakers (i.e. 3.5%).

T-tapping was analysed both within words and across word boundaries in the speech of RP television newsreaders in [7]. The findings revealed that the voiceless plosive was the most frequent realisation of /t/ in intervocalic environments (i.e. 64.8%) and t-tapping occurred more frequently (i.e. 35.2%), which suggested that t-tapping may be considered a well-established variant in RP.

In [7], variation was mostly linked to gender since males used taps more than females with an average percentage frequency of 42.5% and 26.4%, respectively. Linguistic factors, such as word position, affected the production of t-tapping considerably. 66.9% of the tokens were produced word-finally whereas 17.5% corresponded to word-medial [7]. Usage-based factors found in this study were in line with previous research since t-tapping was found to affect mainly monosyllabic function words, frequent phrases and common content words. The analysis of phonetic factors in [7] revealed that t-tapping was equally produced across word boundaries when followed by stressed and unstressed syllables; however, it was less frequent after a stressed syllable than after an unstressed syllable. Regarding t-tapping within words, it was frequently found between a stressed syllable and an unstressed syllable.

The findings obtained in the studies mentioned earlier indicate, in general terms, that taps were frequent at word boundaries in all the varieties of English under investigation, even in RP; however, little is known about speakers' preferred strategies in connected speech when t-tapping is not applied.

Given that little empirical research on t-tapping in connected speech in RP has been conducted, and that variability on the frequency of its competing variant (i.e. t-glottaling) when taps are not used remains scarcely investigated in RP, the present study addresses the variability of potential t-tapping contexts at word boundaries in connected speech in RP.

The principal objective of this paper addresses the following research questions: a) what is the status of t-tapping in connected speech in RP?; and b) is t-tapping favoured over laryngeal strategies (i.e. glottal stops)?

Based on previous studies, it is hypothesised that a) t-tapping will be frequent in connected speech in RP; and b) t-tapping will be favoured over t-glottaling.

3. METHODOLOGY

3.1. Data

The interviews analysed for this study have been extracted from The Andrew Marr Show, a talk show broadcast in the United Kingdom from 2005 to 2021. The guests who normally appear are one senior member of the UK Cabinet, a representative member of the Opposition, and sometimes, famous non-political guests as well (e.g. well-known journalists). The interviews were always conducted in the same programme set and following the same turn-taking method, which enables maintaining consistency for data collection and analysis.

The analysis focuses on the non-rhotic variety of English known as RP, as represented in UK political figures' speech.

Even though in traditional studies of linguistic variation sociolinguistic interviews are the most common practice for collecting data, using speech from a TV programme seems to be a more practical and better-suited method for observing contemporary pronunciation [9; 7]. However, significant drawbacks are found. One of the major disadvantages is the

irregularity of the speakers' appearances and the differences in interview duration (i.e. neither all the speakers appear the same number of times nor all the interviews last the same amount of time).

The speech of six UK political figures, three males and three females, was analysed from 29 interviews corresponding to the years 2010-2013, due to the transcriptions of the interviews are not available until 2010. The duration of the interviews is from 07:17 for the shortest one to 29:06 for the longest one. The text analysed amounts to ca. 74,600 words (excluding interview titles, headings, and other written-down specifications that are not part of the dialogue).

3.2. Procedure

The procedure involved two stages: 1) identification of the potential orthographic contexts where t-tapping may occur at word boundaries; and 2) auditory and acoustic analysis of the potential contexts identified.

To conduct the orthographic search for potential cases of t-tapping, specific orthographic contexts were explored through the word search tool available in Adobe Acrobat Reader. The searches for potential spelling contexts across word boundaries comprised the combination of the letters <t>, <tt>, <d> and <dd> preceded and followed by vowels, with a space in-between to check word boundaries.

All of the spelling searches were done taking into consideration orthographic pauses (e.g. , and .), paying attention to whether they were real phonetic pauses or not.

Despite decisions were carefully made regarding the presence or absence of instances of the connected speech phenomena described earlier, auditory analysis of the data was occasionally considered to be unsatisfactory and spectrographic software was used for acoustic clarification.

To achieve a significant amount of data, a threshold of 300 tokens per speaker was established, resulting in 1,800 tokens being analysed.

3. RESULTS AND DISCUSSION

The results obtained from the analysis show that in a compiled corpus of approximately 74,600 words, 526 cases of t-tapping have been identified out of 1,800 potential cases. This, therefore, corresponds to 29.22% of corpus tokens, whereas 1,274 of the potential cases where t-tapping was not produced correspond to 70.78% of the corpus (see Figure 1 below).

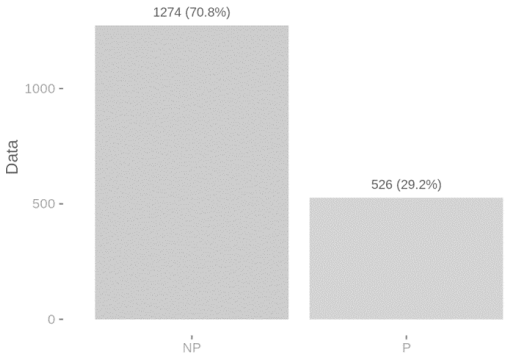


Figure 1: Rates of produced and non-produced cases for t-tapping.

The ratio of production of glottal stops when t-tapping is not applied has also been explored to determine how frequently a glottal stop production is produced in potential t-tapping contexts.

As Figure 2 shows, the quantified results indicate that out of the 1,800 potential cases, 410 are produced as taps in <t> contexts (22.78%), whilst 116 tokens are produced as taps in <d> contexts (6.44%), which yields 29.22% of the overall tap production. The results also indicate that t-tapping is favoured over glottal stop productions in /t/ contexts, which occur in 14.7% of the potential cases when t-tapping is not produced. The ratio of production of word-final /t, d/ corresponds to 63.3% of the overall potential cases where t-tapping or glottal stop are not applied.

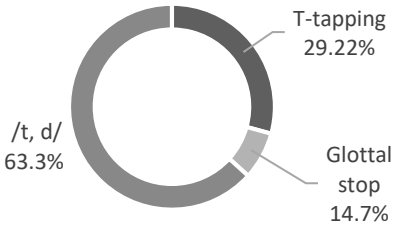


Figure 2: T-tapping strategies at word boundaries.

Since t-tapping is not categorical, two strategies are used in contexts in which the production of t-tapping s avoided: glottal stop [ʔ] and word-final /t, d/.

As Figure 2 above indicates, the analysis reveals that word-final /t, d/ are invariably more favoured over the glottal stop production. Furthermore, the findings also indicate that tap rates vary between /d/ and /t/ contexts. Tap rates for /t/ contexts are higher than for /d/ contexts. One possible explanation might be that glottal stops are an alternate variant of /t/ that RP speakers tend to avoid.

As to whether tap rate varies in terms of gender, the quantified results indicate that the tokens produced by females are slightly lower in comparison to the cases produced by males. As Table 1 shows, the

difference in the production of t-tapping between females and males amounts to 6%.

Gender	Males	%	Females	%
n	290	32.20	236	26.20

Table 1: Overall t-tapping cases by gender.

Concerning the production for t-tapping divided by gender, the tap rate by speaker reveals some patterns. As Table 2 shows, the three female speakers and only one male speaker have a consistent production of t-tapping between 20% and 30%, whilst the other two male speakers have a production of t-tapping above 35%.

Speaker	n	%
Male 1	108	36.00
Female 1	67	22.33
Male 2	67	22.33
Female 2	80	26.67
Male 3	115	38.33
Female 3	89	29.67

Table 2: Overall t-tapping cases by speaker.

These findings support previous research in which male speakers produced more tap variants than female speakers, which might be in relation to female speakers using more standard variants.

The findings obtained in the analysis revealed that the production of t-tapping in spontaneous speech in RP is not very frequent, since the percentage of production is low. The quantified data indicate that t-tapping is not fully established in contemporary pronunciations of RP speakers, as it is produced in less than 50% of the cases. However, a rate of production of 29.22% is high considering the relative formality of the TV programme, and it may mainly be linked to the social status represented by the interviewees, together with the fact that t-tapping is not generally considered a common RP feature.

This low percentage of production seems to refute the statement that its usage could be a representative feature of a casual style of speech [16] and it is in line with previous studies on RP [7] and on AusEng [14] but in contrast to studies on young RP speakers [2].

The findings on the production of t-glottaling also reveal that the traditional alveolar plosives in word-final position are favoured over the production of t-tapping and glottal stops across word boundaries in RP.

The low percentage of production for t-glottaling is related to the negative social connotations associated to the glottal stop variant.

Even though [7] pointed out that several interpretations of the phenomenon suggested its spreading in British English accents, the analysis indicates that t-tapping is a case of a recent development rather than a case of stable variation in contemporary RP. T-tapping has been long associated to sociolinguistic factors such as male speakers, working-class speakers, and with younger speakers [7], which explains t-tapping's low ratio of production in the data. In line with previous studies on RP [7], these findings, therefore, reveal that t-tapping can be considered to be well-established in contemporary RP, and thus, not limited to non-standard accents of English.

4. CONCLUSION

This study has focused on the rate of production of t-tapping in spontaneous speech across word boundaries in contemporary RP and the strategies that speakers use when t-tapping is not applied. Given the data obtained, it is clear that t-tapping still represents an ongoing development in the phonological inventory of RP speakers. Despite the overall production is low, several factors such as the relative formality of the speech situation and the social factors associated to the phenomenon may have conditioned its variability. The findings also reveal differences between the strategies speakers favour in potential t-tapping contexts. In most cases, RP speakers favour word-final /t, d/ pronunciations over both t-tapping production and glottal stops, which represent an extremely low percentage of production in the data.

The present findings represent another step to the understanding of the phenomenon of t-tapping across word boundaries in RP; however, the study has certain limitations that should be acknowledged. Some clear limitations are the number of speakers, tokens and a wider time span. Future research should include more speakers, more potential cases and more years of study, which would result in more potential cases per speaker, with more inter and intra speaker variability in a wider diachronic perspective.

Despite the limitations, this study has tried to contribute to the growing empirical body of data. It shows that t-tapping occurs across word boundaries in spontaneous speech in contemporary RP speakers, although traditional word-final /t, d/ pronunciation is preferred. It also shows that RP speakers favour t-tapping production, which can be described as spreading into RP phonological inventory, over glottal stops. The findings also reveal that alveolar plosives are the common pronunciation in potential t-tapping contexts and that t-tapping is more frequently used than in glottal stops in the same contexts.

5. REFERENCES

[1] Alameen, G., Levis, J. M. 2015. Connected speech. In: Reed, M., Levis, J. M. (eds), *The Handbook of English Pronunciation*. Blackwell, 159-174.

[2] Badia Barrea, B. 2015. *A Sociolinguistic Study of T-glottaling in Young RP: Accent, Class and Education.* Unpublished Doctoral Dissertation. University of Essex.

[3] Connine, C. M. 2004. It's not what you hear but how often you hear it: on the neglected role of phonological variant frequency in auditory word recognition. *Psychomic Bulletin & Review* 11(6), 1084-1089.

[4] Denes, P. B., Pinson, E. N. 1993. *The Speech Chain: The Physics and Biology of Spoken Language* (2nd ed.). W. H. Freeman and Company.

[5] Eddington, D. 2007. Flaps and other variants of /t/ in American English: allophonic distribution without constraints, rules or abstractions. *Cognitive Linguistics* 18(1), 23-46.

[6] Eddington, D., Elzinga, K. 2008. The phonetic context of American English flapping: quantitative evidence. *Language and Speech* 51(3), 245-266.

[7] Hannisdal, B. R. 2006. *Variability and Change in Received Pronunciation: A Study of Six Phonological Variables in the Speech of Television Newsreaders.* Doctoral Dissertation. Department of English, University of Bergen.

[8] Hay, J., Foulkes, P. 2016. The evolution of medial /t/ over real and remembered time. *Language* 92, 298-330.

[9] Labov, W. 1984. Field methods of the project of linguistic change and variation. In: Baugh, J., Sherzer, J. (eds), *Language in Use: Readings in Sociolinguistics.* Prentice-Hall, 28-53.

[10] Mompeán, J. A., Gómez, A. 2011. Hiatus-resolution strategies in non-rhotic English: the case of /r/-liaison. *Proc. 17th ICPhS Hong Kong,* 1414-1417.

[11] Patterson, D., Connine, C. M. 2001. Variant frequency in flap production. A corpus analysis of variant frequency in American English flap production. *Phonetica* 58(4), 254-275.

[12] Shockey, L. 2003. *Sound Patterns of Spoken English.* Blackwell.

[13] Tollfree, L. 1999. South east London English: discrete versus continuous modelling of consonantal reduction. In: Foulkes, P., Docherty, G. (eds), *Urban Voices: Accent Studies in the British Isles.* Arnold, 163-184.

[14] Tollfree, L. 2001. Variation and change in Australian English consonants: reduction of /t/. In: Blair, D., Collins, P. (eds), *English in Australia.* John Benjamins, 45-67.

[15] Warner, N., Tucker, B. V. 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *J. Acous, Soc. Am.* 130, 1006-1617.

[16] Wells, J. C. 1982. *Accents of English,* 3 vols. Cambridge University Press.

[17] Zue, V. W., Laferriere, M. 1979. Acoustic study of medial /t, d/ in American English. *J. Acous. Soc. Am.* 66, 1038-1050.

Acoustic merger between /e/ and /æ/ in Singapore English: insights into stylistic variation and sub-varietal difference

Canaan Zengyu Lan, Olga Maxwell, and Chloé Diskin-Holdaway

The University of Melbourne
zengyul@student.unimelb.edu.au, omaxwell@unimelb.edu.au, chloe.diskinholdaway@unimelb.edu.au

ABSTRACT

This study investigates the extent of the /e/-/æ/ acoustic merger in Singapore English produced by young Chinese Singaporeans in their twenties. We examine the effects of speech style on the acoustic properties of these vowels and the amount of spectral overlap. Participants were recorded reading a wordlist and sentences written in Singapore Standard English (SSE) and Singapore Colloquial English (SCE). Results show some evidence of merger between /e/ and /æ/ on the group level, with the SSE and the SCE sentence styles showing almost complete overlap, but the wordlist showing the least overlap. However, these overlapping patterns are highly gradient when examined individually, with a merger present in the wordlist for some of the speakers. These findings raise questions about the status of Singapore English as a stable or emergent variety and emphasise the need for further fine-grained phonetic research on the vowel systems in New Englishes.

Keywords: Acoustic merger, Singapore English, front vowels, speech production.

1. INTRODUCTION

This study focuses on variability in vowel production in Singapore English (henceforth SgE), which is an umbrella term used to describe English spoken in Singapore. We investigate the extent of acoustic overlap between the DRESS and TRAP vowels (henceforth /e/ and /æ/) across different speech styles in SgE, of which there are at least two syntactically, morphologically and phonologically distinct sub-varieties. Singapore Standard English (henceforth SSE) is generally defined as the standard variety spoken in formal situations and Singapore Colloquial English (henceforth SCE), otherwise known as Singlish, is an informal variety used among friends and family [1]–[3].

Speakers of SgE have been found to demonstrate linguistic unity, homogeneity and confidence in the local standard. This development of "pan-Singaporean" features [4] has been attributed to the rising status of English as the first language and most spoken home language amongst young Singaporeans [5]. Meanwhile, there is an ongoing debate about "standardness", more specifically, the use of SSE in co-existence with SCE or Singlish [6] and the perceived fluidity of boundaries between the two sub-varieties among Singaporeans [3]. Consistent phonetic and phonological inter- and intra-speaker variation [3] further raises questions about whether this is a result of stylistic variation (e.g., [7]) or code-switching between SSE and SCE (e.g., [1]). The front vowels are of particular interest in SgE, given the reported loss of acoustic distinction in SCE [8], and a large degree of variability in vowel realisations [1], [7]–[11].

Previous acoustic investigations of SgE have examined /æ/ alongside /e/ (e.g., [8], [9]). They focus on the /e/-/æ/ acoustic merger, a tendency among most SgE speakers to merge the two vowels, resulting in their phonetic properties being closer to [ɛ]. While a body of work (e.g., [7]–[9]) has explored the realisation of /e/, less is known about the spectral variability of /æ/, addressed in the present paper precisely for this reason.

Studies that have examined the extent of /e/-/æ/ merger in formal speech styles in SgE, such as the citation form or reading tasks, reveal that spectral overlap can be more partial due to the unpredictability in /e/ realisations, where some /e/ tokens merge with /æ/ and the other tokens show raising towards the diphthongal [eɪ] position [12], [13]. As first noted in [13], /e/-raising occurs in the speech of some educated young SgE speakers and only in certain words, such as *bed* and *head*. Further studies that find raising in *next* but not *text* suggest an emerging NEXT-TEXT split in SgE (e.g., [7], [9]). This is in line with theoretical concerns raised by [14] that in some cases of lexically specific phonetic variation, often in contexts of new Englishes, a merger exists more as a synchronic rather than a diachronic phenomenon.

Unlike phonologically conditioned /e/-raising before a voiced velar found across many parts of North America [15], work on SgE (e.g., [9]) indicates that such raising seems rather inconsistent, despite the observation of more frequent /e/-raising before voiced stops (e.g., [7]). A recent examination by [7] comparing older and younger age groups corroborates previous observations that not all young speakers exhibit an /e/-split but that the /e/-/æ/ merger is a strong motivator for /e/-raising among younger speakers. Further, [7]'s observation about stylistic stratification among young Chinese Singaporean participants emphasises the importance of both

speech style and age group as key parameters in the investigation of the acoustics of /e, æ/. Moreover, [7] suggests a negative relationship between /e/-raising and speech formality, with more formal speech styles (i.e., a wordlist versus a reading passage) resulting in less /e/-raising.

Most previous linguistic observations of SgE have assumed correlations between speech styles and sub-varieties, where formal style/reading tasks means the production of SSE, and casual style/interviews means the production of SCE (e.g., [16], [17]). The present preliminary study attempts to disambiguate sub-varieties and styles in SgE. We examine read speech elicited with the help of a wordlist and a set of sentences, further subdivided into sentences written in SSE and SCE (see §2.2 for study design). The production data is based on three sub-sets: **wordlist** (SSE, most formal), **SSE sentence** (SSE, formal but less controlled), and **SCE sentence** (SCE, least formal).

The research questions are as follows:

1. How does speech style (most formal to least formal) influence the acoustic characteristics of /e/ and /æ/ in SgE?
2. How do sub-varietal differences (SSE vs. SCE) contribute to variation in SgE?

2. METHODS

2.1. Participants

The study participants were six Chinese Singaporeans (3M, 3F, gender self-identified; age $\bar{x}$=25, s=2.61) who were either enrolled in or had completed their tertiary degree, with no prior residence in an English-speaking country other than Singapore prior to data collection. All participants self-identified as bilingual speakers with L1 English and varying degrees of fluency in Mandarin. This is except for speaker 005_M whose L1 was Mandarin and L2 English.

2.2. Materials and procedure

The data reported here forms part of a larger project on SgE, investigating 14 monophthongal vowels (including the monophthongised diphthongs GOAT and FACE) across a variety of speech styles, including sociolinguistic interviews, which also elicited participants' demographic information. The analyses presented in this paper are based on a wordlist and a set of sentences. The data collection took place remotely: the participants followed a set of guidelines (how to approach each reading task) shared via WhatsApp and recorded themselves using their smartphones. Participants were instructed to place their smartphones in front of them on a table, complete their recording in one sitting, and read as naturally as possible, allowing for a short pause after each sentence. Additionally, participants were instructed to repeat the entire sentence if they stumbled over a word, without stopping the recording to ensure consistency in volume and quality.

For the wordlist, eight target words were embedded in the carrier phrase "Please say ____ again". For the sentence set, two versions of a semantically similar sentence were provided for participants to read aloud, one written in SCE and the other in SSE, with grammar and lexis modified based on the differences mentioned in the literature (e.g., [3]) and the first author's native speaker intuition. There were eight sentence pairs in total, and participants were briefed about the difference between the two versions before reading the randomised sentences. The words "Standard English" or "Singlish" were written before each sentence to further prompt the intended response (e.g., Standard English: *Have you finished eating? Why are you not having your soup?*; Singlish: *Finish eating? Why you not drinking your soup hah?*). The terms Standard English and Singlish (instead of SSE and SCE) were used as they are more common in Singapore. Both the wordlist and the sentence set involved seven repetitions, and only accentually prominent words were considered for the analysis of the sentence set (Table 1).

	/e/	/æ/
wordlist	bet, bed, het, head	bat, bad, hat, had
SSE sentence	anything, else, seven, ten, very	and, apple, can, has, have, having, Japanese, mad
SCE sentence	better, seven, ten, very	and, apple, can, Japanese, mad
Total	529	781

Table 1: Target words and token numbers for /e/ and /æ/, presented by speech style/sub-variety.

2.2. Data processing and analysis

All data were manually annotated in *Praat* using the default setting [18] and analysed via the *emuR* [19] package in *R* [20]. The F1/F2 estimates were automatically extracted at vowel midpoints using *forest* with the parameters set according to speaker gender. All output data were manually inspected in *Praat*, with any datapoints subject to formant tracker errors corrected or removed. Lobanov 2.0 normalisation [21] was applied, combining the data for 14 monophthongs across all speech styles/sub-varieties and speakers, following the assumptions of a vowel-extrinsic formula.

The Pillai-Bartlett trace, or Pillai score, was calculated with the *tidyverse* [22] package. Pillai scores are one of the most conventionally applied approaches to measuring acoustic overlap [23], [24]

and range from 0 (greater overlap) to 1 (greater distinction). A series of Linear Mixed Modelling (LMM) analyses were performed to determine the relationships between the predictors (e.g., speech style, vowel) and the response variables (i.e., normalised F1 and F2) with the inclusion of random effects using *lme4* [25] and *lmerTest* [26]. The final model included vowel, speech style/sub-variety, and the interaction between vowel and speech style/sub-variety as fixed factors. The random intercepts were speaker and word; factor significance was calculated using the Satterthwaite likelihood t-tests. Tukey post-hoc tests were performed to locate the source of differences using the *emmeans* package [27].

3. RESULTS

3.1. Spectral results across speakers

Fig. 3.1 presents normalised F1/F2 ellipse plots for the vowels /e/ and /æ/. For all three speech styles/sub-varieties, there is a degree of overlap for the two target vowels, with the SSE sentence showing the most variation and overlap in the F1/F2 vowel space. The wordlist has the least amount of overlap due to considerable variability in /e/, whereby half of the /e/ tokens are produced in the same vowel space as /æ/, with /æ/ showing the least variation in the wordlist. The other half of the /e/ tokens in the wordlist show raising and some fronting.

Both /e/ and /æ/ show more variation in the SSE sentence sub-set, with /e/ exhibiting fronting as it raises, compared to tokens in the SCE sentence sub-set. The SCE sub-set indicates greater separation for the two target vowels as /æ/ raises, with a reversal of vowel position in which the centroid for /æ/ is somewhat higher and more forward as compared to /e/.

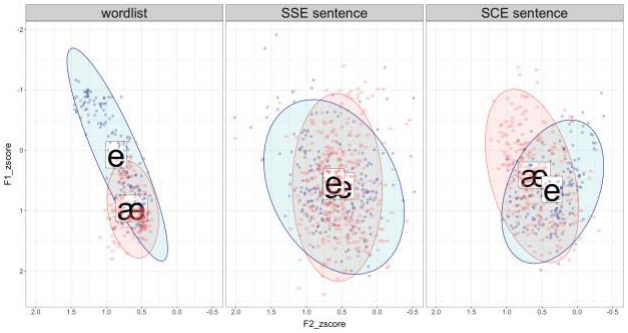


Figure 3.1: Normalised F1/F2 at midpoints of /e/ (blue ellipses), /æ/ (red ellipses) according to speech style/sub-variety (wordlist – left, SSE – middle, SCE – right), with mean F1/F2 estimates of /e/ and /æ/.

The LMM analysis shows no significant effects of vowel or speech style/sub-variety on F1/F2 values. However, the interaction between vowel and speech style/sub-variety is significant for F1 ($p<0.05$), indicating that there are differences between /e/ and /æ/ for the F1 parameter in the wordlist. Further analysis using post-hoc tests reveals a significant difference in F1 ($p<0.05$) and F2 ($p<0.05$) for /e/ between the wordlist and SSE sentence, and in F2 between the wordlist and SCE sentence ($p<0.05$).

In addition, the results revealed potential conditioning effects of the following nasal on vowel formants. The interaction between speech style/sub-variety and the vowel indicates that the F1 of the pre-nasal /æ/ is significantly different from that of the tokens in non-nasal environments in SSE speech style/sub-variety ($p <0.05$). However, caution must be exercised when interpreting these results, given the unbalanced data set (327 out of 1310 tokens where the vowel is followed by a nasal stop). Further investigation using a larger dataset is warranted.

3.2. Individual speaker behaviour

Fig. 3.2 presents the mean F1 (top) and F2 (bottom) estimates for each speech style/sub-variety by speaker. The spectral distributions are highly gradient both individually and stylistically.

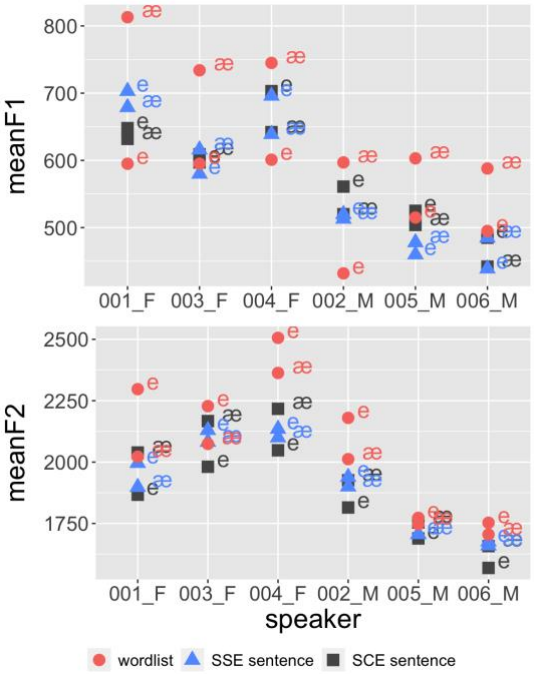


Figure 3.2: Mean F1/F2 estimates (Hertz) (y-axis, F1 – top, F2 – bottom) presented by speaker code in order of gender (x-axis) and speech style/sub-variety (wordlist – red, SSE – blue, SCE – grey)

As shown in the top panel, all participants produce /æ/ lower in the vowel space (i.e., high mean F1 values) in the wordlist (red) in comparison to the sentences (blue, grey). Three speakers (003_F, 005_M, 006_M) raise /e/ in the SSE sentence (blue), i.e., have low F1 values, whereas the other three raise

/e/ in the wordlist (red), at the same time showing a reversal of the /e/ and /æ/ position in the F1/F2 space for the two sentence sets. Among speakers exhibiting this reversal, 001_F produces the two target vowels with varied degrees of distinction in F1, both within and across speech styles/sub-varieties, whereas 004_F, although having an /e/-/æ/ distinction, merges SSE-/æ/ (blue) with SCE-/æ/ (grey), and SSE-/e/ with SCE-/e/. The results for speaker 002_M exhibit difference in F1 in the SCE sub-set, with the mean values for /æ/ clustered with those for SSE-/e, æ/.

As illustrated in the bottom panel, all speakers produce a more fronted /e/ in the wordlist (i.e., the highest mean F2, relative to the speakers' vowel spaces) and a more back target in the SCE sentence (i.e., the lowest F2 mean). For the F2 values, the distribution of /e/ and /æ/ seems to be more consistent for three of the speakers (i.e., 001_F, 003_F, 004_F, all female) showing signs of /æ/ fronting in the SCE sentence. In contrast, male speakers 002_M, 005_M, 006_M produce /e/ and /æ/ with little differences in F2 both within and across speech styles/sub-varieties, except for 002_M whose F2 mean for /e/ is lower in the wordlist.

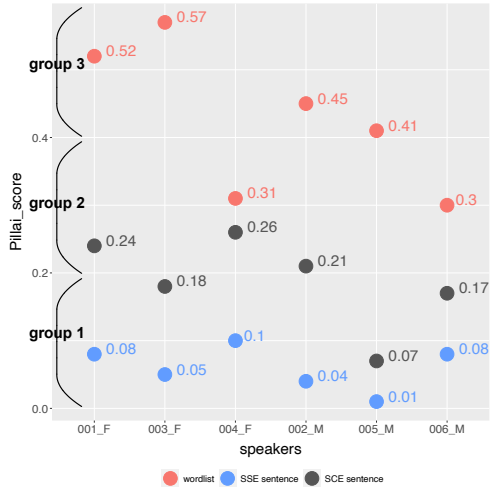


Figure 3.3: Pillai score results with Pillai score on the y-axis and speakers in order of gender on the x-axis, presented by speech style/sub-variety (wordlist – red, SSE – blue, SCE – grey).

Fig. 3.3 illustrates individual Pillai scores. The degree of overlap between /e/-/æ/ across individuals can be grouped according to three main types of overlapping behaviour. Group 1 (scores of 0-0.20, indicating a substantial amount of overlap) reveals the SSE sentence as the most overlapped style/sub-variety for all speakers, with three speakers' (003_F, 005_M, 006_M) SCE sentence also showing a similar degree of overlap. Group 2 (scores of 0.2-0.4, indicating a considerable amount of overlap) includes three speakers (001_F, 004_F, 002_M) whose Pillai scores for the SCE sentence fall within this range, with 004_F and 006_M also having their wordlist showing a similar degree of overlap. Group 3 (scores of 0.4-0.6) includes four speakers (001_F, 003_F, 002_M, 005_M) whose Pillai scores suggest more modest amount of overlap between the two vowels in the wordlist only.

4. DISCUSSION

Returning to the research questions, our analysis has shown that the acoustic results on the group level reveal the least amount of overlap between /e/ and /æ/ for the most formal speech style (i.e., wordlist) and the most overlap for the *less* formal style (SSE sentence), but not the *least* formal style (SCE sentence).

The study confirms [8] and [10]'s findings on /e/-raising in wordlist style in SgE, but not [7]'s findings who reported less frequent /e/-raising in the wordlist. The patterns in our data also exhibit the greatest amount of variation in /e/ and the least variation in /æ/ in the wordlist. We offer two possible explanations. First, hyperarticulation in careful speech could have led to a greater distinction between /e/ from /æ/ in the vowel space. Second, previous research notes that the production of words such as *bed* and *head* is characterised by a higher tendency to be raised by some speakers [13]. This lexical effect was also found in the current study, with different phonetic variants of the close-mid vowel across the target words and individuals, resulting in higher F1 for some of the tokens in the wordlist.

Our analysis also reveals a reversal of /e/ and /æ/ in the SCE sub-set. It is likely that /e/ lowering and backing in the SCE sub-set represent sub-varietal difference between SCE and SSE. Future work will expand the analysis of /e/-/æ/ raising to a much larger dataset and will extend the investigation by applying a dynamic approach, i.e. vowel formant trajectories, including the monophthongised /eɪ/ vowel.

5. CONCLUSION

Our analysis has shown greater /e/-/æ/ overlap in the formal but more controlled style (SSE) as compared to the least formal style (SCE). This pattern potentially reveals complex patterns of variation in SgE, raising questions about acoustic boundaries between speech styles and sub-varieties. The observed /e/-split adds to our knowledge of mergers in (new) Englishes (e.g., [8], [28]), whereby in SgE, a merger may be more accurately conceptualised as a synchronic rather than a diachronic state [14], [29]. This may call for a change of theoretical perspectives and methodological approaches in future studies of mergers and splits in new Englishes.

6. REFERENCES

[1] L. Lim, Ed., *Singapore English: A grammatical description*, vol. G33. Amsterdam: John Benjamins Publishing Company, 2004. doi: 10.1075/veaw.g33.

[2] Z. Bao, *The Making of Vernacular Singapore English: System, Transfer, and Filter*. Cambridge: Cambridge University Press, 2015. doi: 10.1017/CBO9781139135375.

[3] J. R. E. Leimgruber, *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press, 2013. doi: 10.1017/CBO9781139225755.

[4] J. H. Sim, "Negotiating social meanings in a plural society: social perceptions of variants of /l/ in Singapore English," Open Science Framework, preprint, Aug. 2021. doi: 10.31219/osf.io/4jbdm.

[5] Census of 2020, "Census of Population 2020 Statistical Release 1: Demographic Characteristics, Education, Language and Religion," Depart of Statistics Singapore, Singapore, 2020. [Online].

[6] D. Y. S. Chye, "Standard English and Singlish: The Clash of Language Values in Contemporary Singapore," presented at the 2009 Conference of the Australian Linguistic Society Conference, 2010, pp. 1–33.

[7] R. L. Starr and A. Choo, "The NEXT TEXT split in Singapore English: Comparing self report and speech production," presented at the 2019 New Ways of Analyzing Variation, University of Oregon, 2019.

[8] D. Deterding, "Emergent patterns in the vowels of Singapore English," *English World-Wide*, vol. 26, no. 2, pp. 179–197, Jun. 2005, doi: 10.1075/eww.26.2.04det.

[9] D. Deterding, "The vowels of the different ethnic groups in Singapore," in *English in Southeast Asia: varieties, literacies and literatures*, D. Prescott, A. Kirkpatrick, H. Azirah, and I. Martin, Eds. Newcastle, UK: Cambridge Scholars Publishing, 2007, pp. 2–29.

[10] B. H. Suzanna and A. Brown, "The [e] and [æ] vowels in Singapore English," in *The English language in Singapore: research on pronunciation*, 2. pr., A. Brown, D. Deterding, and E.-L. Low, Eds. Singapore: Singapore Association for Applied Linguistics, 2002, pp. 84–92.

[11] E. L. Low and A. Brown, *English in Singapore: an introduction*. Singapore ; New York: McGraw Hill, 2005.

[12] D. Deterding, "Phonetics and Phonology," in *Singapore English*, Edinburgh University Press, 2007, pp. 12–39.

[13] M. W. J. Tay, "The Phonology of Educated Singapore English," *English World-Wide*, vol. 3, no. 2, pp. 135–145, Jan. 1982, doi: 10.1075/eww.3.2.02tay.

[14] W. Maguire, L. Clark, and K. Watson, "Introduction: what are mergers and can they be reversed?," *English Language and Linguistics*, vol. 17, no. 2, pp. 229–239, Jul. 2013, doi: 10.1017/S1360674313000014.

[15] J. A. Stanley, "Regional Patterns in Prevelar Raising," *American Speech*, pp. 1–50, Jul. 2021, doi: 10.1215/00031283-9308384.

[16] K. Kalaivanan, F. Sumartono, and Y.-Y. Tan, "The Homogenization of Ethnic Differences in Singapore English? A Consonantal Production Study," *Lang Speech*, vol. 64, no. 1, pp. 123–140, Mar. 2021, doi: 10.1177/0023830920925510.

[17] A. J. Chong, "Towards a model of Singaporean English intonational phonology," presented at the ICA 2013 Montreal, Montreal, Canada, 2013, pp. 060213–060213. doi: 10.1121/1.4800916.

[18] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." 2022. [Online]. Available: http://www.praat.org

[19] R. Winkelmann, K. Jänsch, S. Cassidy, and J. Harrington, "emuR: Main package of the EMU Speech Database Management System." 2022. [Online]. Available: https://cran.r-project.org

[20] R Core Team, "R: A language and environment for statistical computing." 2022. [Online]. Available: https://www.r-project.org

[21] J. Brand, J. Hay, L. Clark, K. Watson, and M. Sóskuthy, "Systematic co-variation of monophthongs across speakers of New Zealand English," *Journal of Phonetics*, vol. 88, p. 101096, Sep. 2021, doi: 10.1016/j.wocn.2021.101096.

[22] H. Wickham *et al.*, "Welcome to the Tidyverse," *JOSS*, vol. 4, no. 43, p. 1686, Nov. 2019, doi: 10.21105/joss.01686.

[23] J. Nycz and L. Hall-Lew, "Best practices in measuring vowel merger," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 4198–4198, Nov. 2013, doi: 10.1121/1.4831400.

[24] W. Heeringa and H. Van de Velde, "A New Vowel Normalization for Sociophonetics," in *Interspeech 2021*, Aug. 2021, pp. 4024–4028. doi: 10.21437/Interspeech.2021-1846.

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Soft.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[26] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Soft.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[27] R. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2022. [Online]. Available: CRAN.R-project.org/package=emmeans

[28] E. W. Schneider, *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press, 2007.

[29] J. Nycz, "New contrast acquisition: methodological issues and theoretical implications," *English Language and Linguistics*, vol. 17, no. 2, pp. 325–357, Jul. 2013, doi: 10.1017/S1360674313000051.

Front vowel lowering in New Zealand English: investigating a regional sound change

Brooke Ross, Elaine Ballard, and Catherine Watson

The University of Auckland
bros138@aucklanduni.ac.nz, e.ballard@auckland.ac.nz, c.watson@auckland.ac.nz

ABSTRACT

New Zealand English is often noted for its lack of regional variation. Recent research, however, has suggested young speakers in New Zealand's largest city Auckland might have lowered **DRESS**, **TRAP**, and **NURSE** vowels. In the following paper we report findings from a study investigating regional sound change. Auckland speakers were compared with New Zealand English speakers from a regional city, Nelson. Sociolinguistic interviews were recorded with 85 participants stratified by age (16-25 and 40+), gender, and location. Hand-corrected formants from over 25,000 monophthong tokens in stressed syllables, were analyzed. The results confirm earlier findings where young Aucklanders have lowered and retracted **DRESS**, **TRAP** and **NURSE** vowels. There is also some indication, however, that young Nelsonians might also be starting to participate in vowel lowering. The implication of these results is discussed.

Keywords: New Zealand English, Sound Change, Acoustic Analysis, Monophthongs, Vowels.

1. BACKGROUND

1.1 New Zealand English

In this analysis we focus on two ongoing vowel shifts that have been at the forefront of New Zealand English (NZE) phonetic research. The first is the short front vowel shift involving **DRESS**, **TRAP**, and **KIT**. New Zealand English research over the past 40 years has tracked the raising and fronting of **TRAP** and **DRESS** vowels [1, 2, 3]. This sound change has also resulted in the **KIT** vowel retracting and lowering. Studies such as [4] also show that the raising of **DRESS** has resulted in the diphthongization of the long vowel **FLEECE**. The second shift of interest is the raising of the **NURSE** vowel towards GOOSE [5]. Until recently NZE research has documented the continuation of these shifts [2, 3, 4, 6].

Recent phonetic analysis of young speakers in Auckland [7, 8], however, found **TRAP** and **DRESS** lowering, and reduced **FLEECE** diphthongization. This is significant as New Zealand English is generally considered to have little regional variation [9, 10]. If vowel raising is continuing outside Auckland then this could indicate the emergence of an Auckland based variety of NZE. The findings from this study were limited to read speech from young speakers, however. This meant the conclusions reached from this analysis are preliminary in nature.

1.2 Auckland (Urban) vs. Nelson (Regional)

In this paper we report findings from a study comparing monophthongs of speakers from two New Zealand regions. The first, Auckland, is New Zealand's largest city/urban area. It has a population of 1.5 million people, and since the early 1990s has been undergoing rapid demographic change, resulting in increased linguistic diversity compared to elsewhere in NZ. Although Statistics New Zealand does not collect data specifically about the number of New Zealand English speakers, we can show linguistic diversity through other metrics. For example, 41.6% of Aucklanders were born overseas (27.4% nationwide). In addition, 67.4% of Aucklanders speak only one language (77.7% nationwide) [11]. This is significant as studies done in the UK and Europe have suggested that linguistic diversity in large cities can be a catalyst for language change [12].

Our second region, Nelson-Tasman is located in the far north of New Zealand's South Island. It has seen demographic change at a slower pace than Auckland. It has a population of around 100,000, with the largest urban centre having a population of about 65,000 residents. In relation to Auckland, Nelson's population is more monolingual, with a higher proportion of New Zealand English speakers. 78.6% of Nelson-Tasman residents were born in New Zealand compared to 58% of Aucklanders. 87% of residents speak only one language (77.7% nationwide) [13, 14]. It is also important to note, about half of Nelson's migrant population are from the United Kingdom and Ireland. This starkly contrasts Auckland's immigrant population who are from a diverse range of countries across Asia, Europe, and the Pacific Islands. This has resulted in a population which is more like the demographics of New Zealand pre-1990.

2. METHODOLOGY

2.1 Speakers

This study uses data from two sources. Most of the data was collected for the Auckland Voices Project (for details see [7]). This database consists of 67 New Zealand English speakers from Auckland. For this analysis they are stratified by age (Under 25, and Over 40), and gender. An additional 18 recordings with New Zealand English speakers from Nelson were collected for this study, also stratified by age and gender. A breakdown of speakers is given in *Table 1.* Participants were all either New Zealand born or had arrived in New Zealand under the age of seven. Younger speakers had to have lived in their respective location for their entire time in New Zealand. Older Auckland speakers had to have lived in Auckland for 20+ years. Nelson speakers had to have lived in Nelson for 20+ years, and additionally could not have lived in Auckland for any extended period.

Table 1. *Speaker distribution by age, gender, and location.*

		Under-25	Over 40
Auckland	Women	20	17
	Men	18	12
Nelson	Women	5	5
	Men	3	5

Participants were recorded in sociolinguistic style interviews for 1-2 hours in a quiet location of their choice. Interviews were recorded on two devices. A Zoom H5 and a Marantz PMD 661, both using a TDK lavalier clip-on microphone. In instances where the main microphone failed, backup recording from the main recording devices were used. The speech signal was sampled at 44.1 kHz and quantized to 24 bit.

2.2 Data Preparation

In this analysis we took vowel tokens from a ten minute section of the larger interview. This section was taken at the 30-minute mark. This time point was selected as far enough into the interview for the speaker to be comfortable with the interviewer and recording device, but before the speaker was fatigued. Recordings were transcribed using ELAN [15] and passed through the forced aligner WebMAUS (NZ English service) [16]. Further processing was then done using the EMU-webApp [17]. Phonetic boundaries were hand checked and corrected where necessary. Formant tracks were calculated using forest within the EMUR [17] package in R [18]. These formants were then hand checked and corrected where necessary in the EMU-webApp. Additionally, stressed monophthongs were labelled at the vowel target based on the criteria given in [19 & 20]. The F1 and F2 values were then extracted at these vowel targets using EMUR in R. In total this analysis looks at 25,000 hand-labelled monophthongs tokens.

Preliminary investigations into this dataset showed that the impact of gender and aging on the vocal tract made it difficult to meaningfully compare group means. In most instances it would be beneficial to use speaker normalization such as Lobanov or Hearney [21] to minimize speaker differences. Our testing, however, found that, because the changes we are interested in largely involve high front vowels, normalization warped the vowel space towards the high front vowel space, and did a poor job of minimizing vocal tract differences while also retaining sociolinguistic differences. Instead, we use a modified version of the linear transformation performed in [8]. We perform two linear transformations on the F1 and F2 values of the dataset, using the anchor vowels **FLEECE**, **THOUGHT**, and **START** (the peripheral vowels of New Zealand English). The first transforms the male formant values towards the female speaker values. The second takes the transformed male to female formant values, stratifies them by age, and transforms the younger speakers towards the older speakers. Details and examples of this transformation process can be found in [8].

3. RESULTS

3.1 Visual Analysis

Figure 1. shows the results of the formant analysis. Our statistical analysis suggests no significant differences between men and women, therefore, in this instance for space reasons, we combine male and female groups. Centroid means are transformed formant values as described section 2.2. This allows us to combine male and female formant values in the visual analysis. All formant values are in bark. All four plots retain the characteristic triangular vowel space of New Zealand English, with centralized **START** and **STRUT** vowels. All four groups also have a fronted GOOSE, and a mid-back **LOT** vowel. Considering our four vowels of interest, it is worth noting that all four groups have the lowered and retracted **KIT** vowel that distinguishes NZE from Australian English. In addition, the two groups of speakers aged 40+ have raised **DRESS**, **TRAP** and **NURSE** vowels. The Auckland Under 25 group, however, have lowered and retracted **TRAP** and

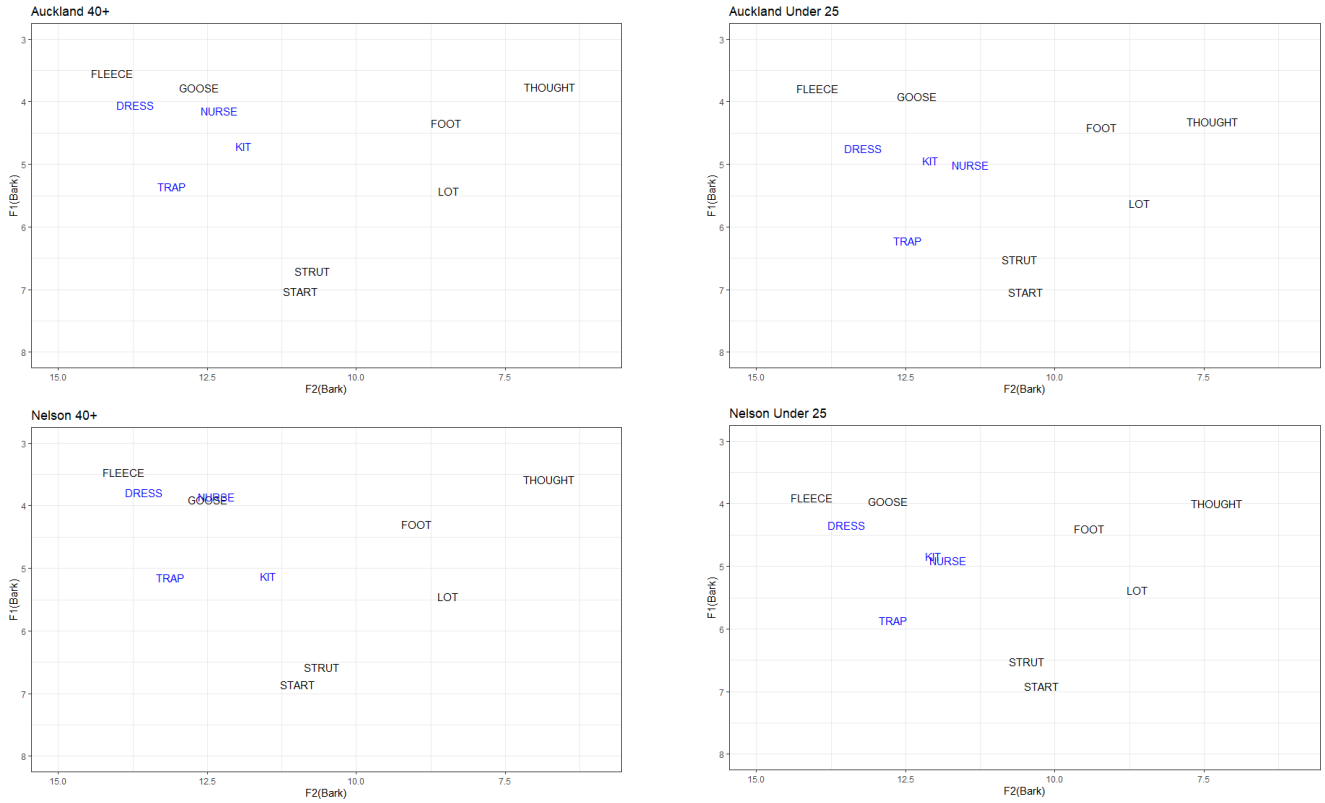


Figure 1: F1/F2 plots of all speaker groups monophthongs. Centroids are transformed means.

DRESS vowels. In addition, **NURSE** is lowered and retracted patterning with **KIT** rather than GOOSE. Interestingly, the visual analysis suggests the younger Nelson speakers retain a raised **DRESS** vowel, but have a lowered **NURSE** vowel and a somewhat lowered **TRAP** vowel compared to the Nelson 40+ group.

3.2 Statistical Analysis

We performed four statistical analyses on our vowels of interest (**DRESS**, **KIT**, **TRAP**, and **NURSE**), outlined in *Table 2,* to investigate whether the visual changes observed in *Figure 1* were significant. All analyses use linear mixed models performed in R, following the methodology in [8] and [22]. The models were calculated using `lme()` function in the `nlme` package. For each vowel three linear mixed models were built for observations of the first and second formants. Comparison between models was done with the `nlme` package using the `anova()` function. When each of the null models were compared by gender (Type vs. Sex) there were no significant differences at a significance level of 0.01. This suggests that gender differences in this dataset are negligible, and for space reasons these results will not be reported here.

Our first analysis (*Table 3*) looks at Auckland speakers to establish whether the vowel lowering identified in the visual analysis is significant. When

Table 2. *The four statistical analyses presented in this investigation.*

Model	Fixed Effects	Random Effects	Observation
Analysis1: Auckland 40+ vs. Auckland Under 25			
Null	Type	Speaker	Formant value
g1	Type*Age	Speaker	Formant value
g2	Type*Sex	Speaker	Formant value
Analysis 2: Auckland 40+ vs. Nelson 40+			
Null	Type	Speaker	Formant value
g1	Type*Place	Speaker	Formant value
g2	Type*Sex	Speaker	Formant value
Analysis 3: Nelson 40+ vs. Nelson Under 25			
Null	Type	Speaker	Formant value
g1	Type*Age	Speaker	Formant value
g2	Type*Sex	Speaker	Formant value
Analysis 4: Auckland Under 25 vs. Nelson Under 25			
Null	Type	Speaker	Formant value
g1	Type*Place	Speaker	Formant value
g2	Type*Sex	Speaker	Formant value

the null model was compared with g1 (Type*Age) **TRAP**, **DRESS**, and **NURSE** differed at a significance level of 0.01. Further analysis from post-hoc t-tests were completed with the older group set as the reference. These show that Auckland Under 25, compared to Auckland 40+ speakers, have lowered and retracted **TRAP** (AgeY:t (66)=7.4, p<.0001; typeF2:AgeY: t(4555)= -6.5, p<.0001), **DRESS** (AgeY:t (66)=4.88, p<.0001; typeF2:AgeY: t(5445)= -4.75, p<.0001), and **NURSE** (AgeY:t (66)=6.35,

p<.0001; typeF2:AgeY: t(1951)= -8.7, p<.0001) vowels. Interestingly, however, the **KIT** vowel is not significantly different between the two groups.

Table 3. Analysis 1: Auckland 40+ vs. Auckland Under 25 (*Null model vs g1 model (Age)).*

	Degrees of Freedom	AIC Difference	Log Likelihood Ratio	P-value
TRAP	10	-38.14	42.14	<.0001
NURSE	10	-48.76	52.76	<.0001
DRESS	10	-18.86	22.86	<.0001
KIT	10	-0.8	4.81	0.0905

Analysis number 2 (*Table 4)* compares the two older groups and finds one significant difference at 0.01. The post-hoc t-tests with Auckland as the reference shows that the Nelson speakers have a lowered and retracted **KIT** vowel compared to the Auckland speakers (PlaceN:t (38)=3.95,p0.0003; typeF2:PlaceN: t(3071)= -4.38,p<.0001). Previous NZE research has identified **KIT** lowering/retraction is a stigmatized feature of broad NZE which may explain this difference [23].

Table 4. Analysis 2: Auckland 40+ vs. Nelson 40+ (*Null model vs g1 model (Place)).*

	Degrees of Freedom	AIC Difference	Log Likelihood Ratio	P-value
TRAP	10	2.247	1.75	0.42
NURSE	10	1.5	2.5	0.29
DRESS	10	2.65	1.34	0.51
KIT	10	-14.19	18.19	0.0001

In the third analysis we look at differences between the two Nelson groups (*Table 5)*. We find two significant differences, for **NURSE** and **KIT**. The post-hoc t-tests with the Nelson 40+ speakers as the reference are as follows. Nelson under 25 speakers have lowered and retracted **NURSE** (AgeY:t (17)=3.56, p0.002; typeF2:AgeY: t(468)= -4.64, p<.0001) vowels, while **KIT** (AgeY:t (17)=-3.62, p0.002; typeF2:AgeY: t(1436)= 8.83, p<.0001) vowels are raised and fronted for the Under 25 group.

Table 5. Analysis 3: Nelson 40+ vs. Nelson Under+ (*Null model vs g1 model (Age) – significant differences (significance level 0.01)).*

	Degrees of Freedom	AIC Difference	Log Likelihood Ratio	P-value
TRAP	10	-3.127	7.13	0.02
NURSE	10	-11.48	15.47	0.0004
DRESS	10	0.112	3.89	0.14
KIT	10	-26.77	30.77	<.0001

Table 6. Analysis 4: Auckland Under 25 vs. (*Null model vs g1 model (Place) – significant differences (significance level 0.01)).*

	Degrees of Freedom	AIC Difference	Log Likelihood Ratio	P-value
TRAP	10	-0.54	4.53	0.1
NURSE	10	0.43	3.57	0.16
DRESS	10	-3.92	7.92	0.019
KIT	10	3.36	0.72	0.72

Our final analysis compares the mean formant values of the two younger groups (*Table 6)*. Surprisingly this analysis finds no significant differences between the Auckland and Nelson speakers Under 25.

4. DISCUSSION AND CONCLUSION

In this study we compared the monophthongs of NZE speakers from urban and regional New Zealand areas. Both the visual analysis and the statistical analysis support earlier research suggesting the **TRAP**, **DRESS**, and **NURSE** vowels are lowering and retracting for young Auckland speakers compared to older Auckland speakers. For these speakers, however, the short front vowel **KIT** does not appear to be raising or fronting in response to this shift. This may be because of its status as an identity marker for NZE speakers as one of the major features that differentiates NZE and Australian English [20]. In addition, we have shown that for both Auckland speakers Nelson speakers aged 40+ **DRESS**, **TRAP** and **NURSE** remain raised and fronted. This suggests front vowel lowering in NZE is a relatively recent phenomenon.

One purposes of this analysis was to investigate whether vowel lowering is a regional change for young Aucklanders. It appears this may not be the case. For Nelson speakers Under 25 the **NURSE** vowel, is lowered and retracted to a similar height and backness as the Auckland speakers. While their **TRAP** and **DRESS** vowels are not significantly different to either Auckland under-25 or Nelson 40+ speakers. These results lead us to conclude that while Auckland speakers under 25 appear to be most advanced in this sound change, these sound changes might not indicate a regional change. These results may instead indicate sound change emerging from the linguistically diverse Auckland which are then being adopted by young people elsewhere in New Zealand. This is an interesting finding considering NZE research from outside Auckland has maintained that vowel raising for **DRESS**, **TRAP** and **NURSE** is ongoing [6, 24]. Given the small size of the regional dataset in this study, however, further data from other New Zealand regions is needed.

7. REFERENCES

[1] Bauer, Laurie. "The second great vowel shift?." Journal of the International Phonetic association 9(2): 57-66, 1979.

[2] Maclagan, M., & Hay, J., "The rise and rise of New Zealand English DRESS.", The Proceedings of the Australian International Conference on Speech Science and Technology, 183-188, 2004.

[3] Watson, C., Maclagan, M., and Harrington, J., "Acoustic evidence for vowel change in New Zealand English", Language variation and change 12., 12(1): 51-68, 2000.

[4] Maclagan, M., & Hay, J., "Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift." Language variation and change, 19(1):1-25, 2007.

[5] Maclagan, M., Watson, C., Harlow, R., King, J., and Keegan, P., "Investigating the sound change in the New Zealand English nurse vowel /ɜ/". Australian Journal of Linguistics, 37(4):465-485, 2017.

[6] Warren, P., Quality and quantity in New Zealand English vowel contrasts. Journal of the International Phonetic Association, 48(3):305-330, 2018.

[7] Ross, B., An acoustic analysis of New Zealand English vowels in Auckland. Master's thesis, Victoria University of Wellington, 2018.

[8] Watson, C., Ross, B., Ballard, E., Charters, H., Arnold, R., & Meyerhoff, M., "Preliminary investigation into sound change in Auckland." In Proceedings of the 17th Australasian International Conference on Speech Science and Technology, 17-20, 2018.

[9] Gordon, E., Maclagan, M., & Kortmann, B. (2008). Regional and social differences in New Zealand phonology. Varieties of English, 3, 64-76.

[10] Bauer, L. (1994). "English in New Zealand." In R. Burchfield (Ed.), The Cambridge History of the English Language. Vol5, English in Britain and overseas: origins and development, 382-429.

[11] Stats NZ. "Place Summaries | Auckland Region | Stats NZ." https://www.stats.govt.nz/tools/2018-census-place-summaries/auckland-region, accessed on 10 Dec 2022.

[12] Cheshire, J., Fox, S., Kerswill, P. & Torgersen, E. "Language contact and language change in the multicultural metropolis." Revue Française de Linguistique Appliquée, 17(2):63-76, 2013.

[13] Stats NZ. "Place Summaries | Nelson Region | Stats NZ." https://www.stats.govt.nz/tools/2018-census-place-summaries/nelson-region, accessed on 10 Dec 2022.

[14] Stats NZ. "Place Summaries | Tasman Region | Stats NZ." https://www.stats.govt.nz/tools/2018-census-place-summaries/tasman-region, accessed on 10 Dec 2022.

[15] Sloetjes, H., & Wittenburg, P., "Annotation by category – ELAN and ISO DCR." Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[16] Kisler, T., Reichel, U., & Schiel, F., "Multilingual processing of speech via web services." Computer Speech and Language, 45:326–347, 2017.

[17] Winkelmann, R., Harrington, J., & Jänsch, K., "EMU-SDMS: Advanced speech database management and analysis in R." Computer Speech and Language, 45:392–410, 2017.

[18] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, accessed on 24 July 2022.

[19] Harrington, J., Phonetic analysis of speech corpora, Wiley Blackwell, 2010.

[20] Watson, C., Harrington, J., & Evans, Z., "An acoustic comparison between New Zealand and Australian English vowels." Australian journal of linguistics 18(2): 185-207, 1998.

[21] Disner, Sandra Ferrari. "Evaluation of vowel normalization procedures." *The Journal of the Acoustical Society of America* 67(1):253-261, 1980.

[22] Watson, C. I., Ballard, E., Ross, B., & Charters, H., "Divergence of FACE and TRAP in Auckland English: A Potential Regional Sound Change in New Zealand English." The 19th International Congress of Phonetic Sciences, Melbourne, Australia, 2019.

[23] Maclagan, Margaret A., Elizabeth Gordon, and Gillian Lewis. "Women and sound change: Conservative and innovative behavior by the same speakers." Language variation and change 11(1): 19-41.

[24] Watson, C., Maclagan, M., King, J., Harlow, R., and Keegan, P., "Sound Change in Māori and the influence of New Zealand English." Journal of the International Phonetic Association, 46(2):185-218, 2016.

Effects of Sociolinguistic Factors on Second Dialect Acquisition of North Korean Refugees in Seoul

Jungah Lee[1], Kaori Idemaru[2], Charlotte Vaughn[2]
[1]Hanyang Institute for Phonetics & Cognitive Sciences of Language (HIPCS), Hanyang University;
[2]University of Oregon
jlee14@hanyang.ac.kr; idemaru@uoregon.edu; cvaughn@uoregon.edu

ABSTRACT

This study investigates to what extent various factors affect North Korean (NK) refugee speakers' stop production, from the perspective of second dialect acquisition. The VOT and F0 associated with 7945 stops were measured from 22 NK speakers (F:16, M:6). The factors examined were age of arrival (AoA), length of residence (LoR), and three different aspects of social adaptation (Identification, Assimilation, and Language attitudes). Results showed that the NK speakers with stronger South Korean (SK) identity and longer LoR produced more SK-like VOT patterns. Moreover, the speakers with positive language attitude toward SK and longer LoR produced more SK-like F0 patterns. Our results also suggest gender may moderate these factors and play a role in NK speakers' acquisition of SK stops.

Keywords: Second dialect acquisition, Sociolinguistics, Sociophonetics, Korean Phonetics

1. INTRODUCTION

Considerable linguistic divergence between North Korean (NK) and South Korean (SK) varieties exists, arising due to the seventy-three years of physical and political separation, as well as different language policies [16]. One of the differences has been observed in the three-way stop contrast (fortis vs. lenis vs. aspirated stops). While acoustic cues of fortis stops have not changed, lenis and aspirated stops in have gone through changes in VOT and F0 patterns in utterance initial (IP-initial) position [6, 17, 21]. Specifically, in SK, VOT is no longer a salient cue to distinguish lenis and aspirated stops, showing VOT merger (e.g., 60 ~ 80 ms) in this position. Instead of VOT, now F0 of the following vowels is used as a cue; the lenis stops are produced with lower F0 than aspirated stops on the following vowel.

Recent studies examining the stop production of NK refugees living in SK showed that NK speakers' stop production was different from that of SK speakers [13, 17]. Unlike the SK speakers, the NK speakers used VOT to distinguish lenis and aspirated stops (30-60 ms VOT for lenis and 80-120 ms for aspirated stops). In addition, the NK speakers distinguished F0 between lenis and aspirated stops less robustly than the SK speakers [16]. Thus, the NK speakers' stop production appeared to align with SK speakers' previous stop patterns of the early 1990s [11, 17].

However, the pattern of NK stop production was not necessarily uniform across contexts. Instead, [16] observed more SK-like stops in some speech styles (conversational) than others (read speech). Such variable production in the process of second dialect acquisition is possible and even likely, potentially triggered by such factors as the formality of the context [15] and interlocutor [4]. Prior research also has suggested that extralinguistic and sociolinguistic factors may exert influence on individual differences in second dialect production and acquisition, including (i) age of arrival (AoA), (ii) length of residence (LoR), (iii) degree of identification with the first dialect (D1) and second dialect (D2), (iv) degree of assimilation in the D2 region, and (v) language attitudes towards D1 and D2 [5, 20]. Such factors may matter since studies have shown that geographically mobile speakers who identified as members of the D2 community acquired D2 features better than those with weaker D2 identity [5, 20]. Speakers who report being more satisfied with living in D2 region and socialized with D2 locals produced more D2 variants [18].

These prior findings suggest that such extralinguistic and sociolinguistic factors may influence individual differences in NK speakers' stop productions and their acquisition of SK stop patterns. Given this, we examined the effects of the five factors listed above on the pattern of stop production (i.e., NK-like vs. SK-like) focusing on the use of VOT and F0 as cues to distinguish lenis and aspirated stops.

2. METHODOLOGY

2.1. Participants

Twenty-two NK speakers now living in Seoul (F: 16, M: 6) in their early 20s (Mean: 22.4, SD: 4.1) participated for a small amount of payment. The speakers were from towns near the capital city of Pyongyang, North Korea, in the Pyongan province, whose regional variety is considered to be standard

North Korean [22]. Their mean AoA in SK was 18.9 years old (SD: 4.9, range 9 – 31) and their mean LoR in Seoul was 3.7 years (SD: 3.4, range 1 - 10).

2.2. Speech tasks and procedure

All participants completed two reading tasks, a syllable elicitation task and a phrase elicitation task, and then a sociolinguistic interview. The syllable elicitation task included 18 CV syllables, each with one of the three-way stop contrasts (e.g., [k]/[k*]/[k^h]) followed by cardinal vowels [ʌ] or [o] (e.g., [kʌ]), across 3 places of articulation. The phrase elicitation task included 9 phrases, adopted from [10]. Each phrase had four syllables, including one of the stops in the first two (e.g., /p^{h}an.p^{h}an.ha.ta/ 'To be even/flat'). Typically, the initial syllable was repeated (e.g., /p^{h}an.p^{h}an/, except for the case of /p*an.c*ak/ and /pan.c*ak) and followed by /ha.ta/ ('to be'). The vowel was always [a]. The test syllables and phrases were elicited in the same carrier sentence "__ (*la)ko malha-yss-ta* (I said__)" [10]. Our analyses in this study focused on the lenis and aspirated stops in the initial syllable.

Each speaker sat in front of a laptop computer, wearing a lavalier microphone Audio-Technica AT 899, which was connected to a Marantz PMD 670 flash drive recorder in a quiet room. Speakers completed the syllable elicitation task first, followed by the phrase elicitation task. The computer screen presented each syllable or phrase in the carrier sentence in three randomized orders. This task took about ten minutes to finish.

Immediately after the reading tasks, the participant and the first author engaged in a sociolinguistic interview over approximately 45 minutes. The participant and the interviewer sat face to face, and the same recording device described above was used for this task. The interview questions were modified from sociolinguistic interview questions intended for an immigrant population developed by [1]. The demographic section of the interview asked participants questions about their hometown, AoA, and LoR in SK. The SK and NK topics sections asked questions about their lives, neighborhoods, and the people and language in each region. All participants answered all questions in the interview protocol.

2.3. Stops and measurements

The duration of the VOT in the IP initial stops was measured from the left edge of the burst release to the onset of the following vowel, defined as the left zero crossing of the first complete periodic cycle [9]. F0 of the following vowels was measured at the mid-point of the vowel following the stop [10]. The F0 was centered between participants to remove gender-related effects. The articulation rate was calculated (number of syllables per second) because it is known to affect the duration of VOT [9]. Each stop was coded for the word containing the stop, and the following vowel.

2.4. Identification, Assimilation, and Language attitude scores

The NK speakers' Identification, Assimilation, and Language attitude scores were calculated based on their sociolinguistic interview responses, combining and modifying the methodologies in [5, 20, 25].

In our study, *Identification* score represents to what extent the NK speakers identify more with D1 (NK) or with D2 (SK). *Assimilation* score indicates to what extent the NK speakers are motivated and willing to live like SK people do. The degree of the speaker's adaptation in the D2 community has been identified as a critical factor in previous studies [20, 25]. The *Language attitude* score, based on the Accent category in [25], reflects the extent to which NK speakers were interested in and willing to acquire the SK variety.

To code speaker responses vis-à-vis these categories, the first author entered 1 point every time one of their responses reflected a sense of identification with or assimilation with the SK society and a positive attitude toward SK variety. One negative point was entered every time a speaker response indicated the same sentiment about NK society and the NK variety. For instance, for *Assimilation* scores, if a speaker responded that she liked living in South Korea, this was considered an expression of assimilation to South Korea. And if she mentions this sentiment four times during the interview, we counted this as four points for the category of Assimilation.

2.5. Analysis

We first examined collinearity between scores of Identification, Assimilation, and Language attitude. Since the degree of correlation among these scores was not very strong, albeit reliable (r ranging from -0.16 to 0.43), we took this to mean that the categories of Identification, Assimilation, and Language attitude captured some different aspects of NK speakers' stance and adaptation to SK, while overlapping slightly with each other. We thus retained the categories in the subsequent analysis.

We used mixed-effect linear regression models to examine the factors that influence variation in VOT and F0 [3, 19]. The predictors were Stop (lenis vs. aspirated), AoA, LoR, Assimilation, Identification, and Language Attitudes and the interaction between

Manner and each of the other factors. Articulation rate was included in the VOT model only. Random intercepts for Participants and Word, and by-Participant random slope for Manner and by-Word random slopes for AOA, LOR, Identification, Assimilation, and Language Attitude were also included.

The dredge() function from the MuMIn package was used to validate the contribution of predictors [2, 8] and it confirmed that the best-fit model should retain all factors. The function model.avg () was used to evaluate predictors by averaging coefficients across the model's fit. Finally, the sw() function was used to calculate predictor importance in a model, i.e., the probability that each predictor contributes to the VOT and F0 production.

3. RESULTS

3.1. VOT

Results for VOT showed that the main effects of Articulation rate and Stop were significant as expected (β = -8.5, SE =0.2, z = 33.9, $p < 0.001$; β = 35.8, SE =1.4, z = 25, $p < 0.001$, respectively). More importantly, the interaction between Manner and LoR was significant (β =-5.0, SE = 1.9, z = 2.6, p = 0.008) and the interaction between Stop and Identification score was also significant (β = -3.2, SE= 1.8, z= 1.7, p = 0.083). As the model averaging analysis provides more conservative and standardized results [8], we interpret that Stop*LoR is significant. The interactions indicate that the VOT difference between lenis and aspirated, which indexes SK-likeness vs. NK-likeness, was modulated as a function of LoR and Identification score. Longer LoR and higher Identification scores (with SK) predicted more SK-like VOT patterns (Fig 1 and 2). The two factors were fairly close in their strength of prediction. LoR was the most effective factor among the factors (98% of predictability for the SK-like VOT production) and Identification was the second most influential factor (92% of predictability for the SK-like VOT pattern) as indicated by the sw() results. This suggests that the NK speakers who have lived in SK only for a few years still might be able to acquire and produce SK-like VOT if they have strong alignment with SK identity. Unlike findings in previous studies [12, 24], AoA was not a significant factor in here.

3.2. F0

Results for F0 showed that the main effects of Stop and Language attitude were significant (β = 0.1948, SE = 0.0051, z = 28.2, $p < 0.001$; β =-0.00085, SE = 0.003, z = 2.7, p = 0.005, respectively). Also, there were reliable interactions between Manner and Language attitude (β = 0.0464 SE = 0.0061, z = 7.6, $p < 0.001$) and between Manner and LoR (β = 0.027, SE = 0.0097, p = 0.004). Again, the two factors were close in the strength of prediction. Language attitude (Fig 3) can predict the most SK-like F0 production (100%) among the factors and LoR (Fig 4) can be the second most factor that predicts 97% of SK-like F0 production. It is notable that Language attitude score outperformed LoR here. This suggests that NK speakers with positive language attitudes toward SK can acquire and produce SK-like F0 patterns, even when they have not been in SK for a long time. Unlike the previous literature [12, 18], AoA and Assimilation were not an important predictor in F0 production here. Interestingly, whereas Identification predicted SK-like VOT patterns, it did not predict SK-like F0 patterns. Likewise, Language attitude predicted SK-like F0 but not VOT.

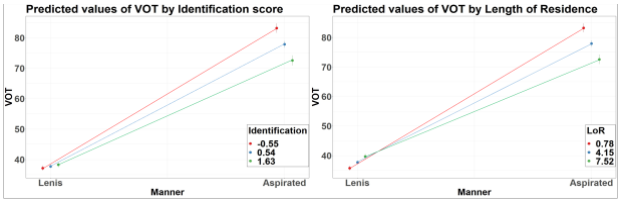


Figure 1 (left) and 2 (right)

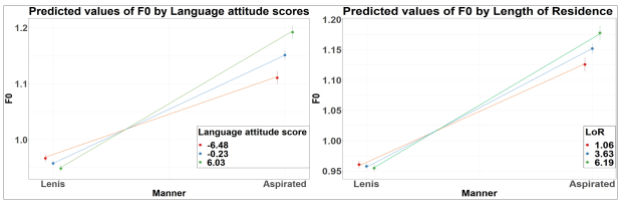


Figure 3 (left) and 4 (right)

3.3. Individual analysis

We noticed that some participants showed strong SK-like stop patterns even though they had only been in SK for one year or less; while others showed strong NK-like stop patterns even when they have lived in Seoul for 10 years. These cases prompted us to identify SK assimilators and SK dissimilators and examine their adaptation scores post hoc.

The mean difference in VOT between lenis and aspirated was calculated for each participant. We then separated assimilator and dissimilator groups based on the group mean (40 ms for VOT and 42 Hz for F0).

We noted that most of VOT assimilators were female (7 females and 1 male). Three of them had very short LoR (1-2 yr) who nonetheless were VOT assimilators. Examination of these 3 participants' adaptation scores showed that their Identification to SK, Assimilation in SK and Language attitude toward SK were stronger than others. VOT dissimilators

included both female and male, and most (10 out of 11) had short LoR (1-3 yrs). One exception was a male participant who showed a dissimilar VOT pattern despite living in SK for 10 years. His adaptation scores showed zero Identification (neutral between SK and NK), high Assimilation to SK, and strong NK linguistic bias (i.e., positive language attitude to NK and negative language attitude to SK).

Similar to VOT, all nine F0 assimilators were female. While some participants had longer LoR (4-10 yrs), 3 participants had only 1 yr LoR. These 3 participants showed strong Assimilation to SK and positive Language attitude toward SK. On the other hand, F0 dissimilators included both female and male, and many had short LoR (1-2). Interestingly, four participants in this category who had longer LoR (4-10) were all male. These four participants' scores indicated zero Identification (neutral), high Assimilation to SK, and strong positive NK linguistic bias. Their SK assimilation scores were higher than the general mean, indicating that they were satisfied with their lives in SK. However, in their interview responses, they indicated that SK variety sounded too feminine and they did not wish to acquire SK dialect as, to them, NK dialect was cooler, stronger, and more masculine.

4. DISCUSSION

The current study investigated to what extent the factors AoA, LoR, Identification, Assimilation, and Language attitude influenced NK refugee speakers' stop production. We found that among those factors, *LoR*, *Identification*, and *Language attitude* predicted variation in VOT and F0. The more experience NK speakers had with SK (as estimated by longer LoR), the more their stops look like those of SK, with shorter VOT and higher F0 for aspirated stops. Finding the effect of LoR but not AoA was rather unexpected, given the opposite patterns or the strong effect of AoA reported in other studies [13]. The LoR effect we found might be because of the LoR of our participants were generally not very long (M = 3.7, range 1-10) and it is reported that the effect of LoR may exert a stronger influence for those speakers whose learning has not plateaued [7].

Additionally, NK speakers who more identified as members of SK society (high Identification scores) tended to produce more SK-like VOTs, and those who had more positive attitudes toward the SK variety (Language attitude scores) tended to produce more SK-like F0. The flip-side of these results is that NK speakers who maintained strong NK identity and affinity toward the NK variety retained more NK-like stops. It is noteworthy that these social stances and attitudes toward the dialect community and regions affected the use and acquisition of the second dialect. We note that the effect of Language attitude was even stronger than that of LoR in predicting the F0 pattern.

Unlike Identity, Assimilation scores did not influence NK speakers' VOT or F0. In other words, while identifying as a member of SK society was related to more SK-like stop production, adapting well in SK was not linked to more SK-like stop production. On the surface, this may seem puzzling as identification as a person of that dialect area and assimilating to that area may seem indistinguishable. However, we suggest that the NK speakers' unstable status as refugees in South Korea is relevant here. These speakers chose to escape from North Korea, and to settle down in South Korea to find a better life [14]. As some of them reported in the interview, they preferred to live in South Korea because of the freedom, capitalistic ideology, and better infrastructure and environment in South Korea. This attitude toward their environment might lead them to more easily assimilate in South Korean society. However, enjoying and appreciating a free life with material convenience (which would lead to high assimilation scores) might not lead NK speakers to necessarily acquire SK-like stops, in the same way that identifying as a member of SK society might.

Post-hoc individual analyses indicated that the same NK speakers did not necessarily show SK-like patterns in both VOT and F0. Some showed only SK-like VOT and some showed only SK-like F0. In fact, only one speaker showed an SK-like pattern in both VOT and F0. Further, four out of six male speakers showed NK-like F0 usage, i.e., lower F0 following the aspirated stops. The content of their responses in the sociolinguistic interview indicated that while they were content with living in SK (Assimilation), they showed strong NK identity and negative attitudes toward the SK dialect, commenting, in particular, that the NK variety is more masculine and cool. The responses suggest that some male NK speakers view the SK variety as having social meanings of 'femininity' and 'softness' as opposed to the NK variety's 'masculinity.' This, in turn, may have resulted in these speakers' resisting the acquisition of higher F0 of aspirated stops in SK. Further study should investigate relations between gender (masculine) identity and acquisition of standard D2, in more depth.

5. ACKNOWLEDEMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884) awarded to TC (PI) and SK (Co-PI).

6. REFERENCES

[1]Anastassiadis. T., Ralli, A., Gekas, S., Pappas, P., Siotou, A., Tsolakidis, S., Karachaliou, R., Rsimpouris. C., & Papanagiotou, C. (2017). Immigration and Language in Canada: Greeks and Greek-Canadians Questionnaire "The questionnaire created and used for the interviews which are part of the Immigrec Database"

[2] Bartoń, Kamil. (2013). MuMIn: Multi-model inference. R package version 1.10.0.

[3] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005.

[4] Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

[5] Carmichael, Katie. (2017). Displacement and local linguistic practices: R-lessness in post-Katrina Greater New Orleans. Journal of Sociolinguistics. 21. 696-719. https://doi.org/10.1111/josl.12253.

[6] Choi, J., Kim, S., & Cho, T. (2020). An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. *Plos one*, *15*(10), e0240682.

[7] Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3-83.

[8] Franklin, A. B., Anderson, D. R., & Burnham, K. P. (2002). Estimation of long-term trends and variation in avian survival probabilities using random effects models. *Journal of Applied Statistics*, *29*(1-4), 267-287.

[9] Holliday, J. J. (2015). A longitudinal study of the second language acquisition of a three-way stop contrast. *Journal of Phonetics*, *50*, 1-14. https://doi.org/10.1016/j.wocn.2015.01.004.

[10] Idemaru, K., & Guion, S. G. (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the International Phonetic Association*, *38*(2), 167–186.

[11] Jang, H. J. (2017). 20 seyki cenki kwuke etwu phyeyswayumuy sengtaycintong sicaksikan (VOT) yenkwu:'pothonghakkyo cosenetokpon','kimpokcinuy tonghwakwuyen'calyolul pathangulo [Study of VOT in stops in early 20th:'Based on Kim Bok-Jin's Joseon language reading]. *Hankwukehak [Korean linguistics], 77*, 311-329.

[12] Kang, K. H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *The Journal of the Acoustical Society of America*, *124*(6), 3909-3917.

[13] Kang, Y. J., & Yun, S. Y. (2018, June). *Acquisition of second dialect features by North Koreans in Seoul.* Poster session presentation at the meeting of *LabPhon16*, University of Lisbon, Lisbon, Portugal.

[14] Kim, S. Y. (2006). Saythemin cengchakuy silthaywa mwunceycem mich kaysenpangan [An analysis of the current state of North Korean refugee resettlement, along with identified issues and proposed solutions for improvement]. *Hankwuktongpwukanonchong [Journal of Korea and East Asia],* (38), 249-269.

[15] Labov, W. (2006). *The social stratification of English in New York city*. Cambridge University Press.

[16] Lee, H. B (1991). Pwukhanuy wulimal umsenghak yenkwu hyenhwang. [Phonetic study in North Korea]. *Hankul [Korean]* (213), 7-28.

[17] Lee, J. (2022). *Second Dialect Acquisition of North Korean Refugees in Seoul* (Order No. 29252949.). [Doctoral dissertation, University of Oregon]. ProQuest Dissertations & Theses Global.

[18] Nycz, J. (2018). Stylistic variation among mobile speakers: Using old and new regional variables to construct complex place identity. *Language Variation and Change*, *30*(2), 175-202.

[19] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[20] Siegel, J. (2010). *Second dialect acquisition*. Cambridge University Press.

[21] Silva, D. J. (2006). Variation in voice onset time for Korean stops: A case for recent sound change. *Korean Linguistics*, *13*(1), 1-16.

[22] Sohn, H. M. (2001). *The Korean language.* Cambridge University Press.

[23] Stanford, J. (2007). *Dialect contact and identity: A case study of exogamous Sui clans* ProQuest Dissertations & Theses Global.

[24] Stanford, J. (2008). A sociotonetic analysis of Sui dialect contact. *Language Variation and Change, 20*(3), 409-450.

[25] Walker, A. (2014). *Crossing oceans with voices and ears: Second dialect acquisition and topic-based shifting in production and perception*

SEGMENTAL AND AUTOSEGMENTAL MODIFICATION IN RAKUGO STYLE JAPANESE

Shigeto Kamano[1)], Chikau Sakamoto[2)], Rin Tsujita[3)], Seunghun J. Lee[4)]

[1,2,3,4)]International Christian University, [4)]IIT Guwahati
shig.kamano@gmail.com[1)], sakachi.jp@gmail.com[2)], c231421r@icu.ac.jp[3)], seunghun@icu.ac.jp[4)]

ABSTRACT

Rakugo is a Japanese performance art form where a solo performer delivers a story with comedic parts, often altering their voice when they produce a speech style unique to Rakugo. This paper compared the phonetic patterns of Rakugo-style speech and natural speech produced by 8 Rakugo performers (called *Hanashika*). An analysis of 32 bi- or tri-syllabic words varying in onset type, accent type, and vowel in word-initial syllables show differences in the Rakugo speech style. Results show that pitch range is larger and syllable duration is longer in the Rakugo style. Vowels were more dispersed but high vowel devoicing occurs with much less frequency in the Rakugo style. The expanded use of acoustic cues is suggested to be the phonetic features of Rakugo.

Keywords: Rakugo speech, Japanese, pitch range, syllable duration, vowel dispersion, vowel devoicing

1. INTRODUCTION

Rakugo is a storytelling form that originates from performances conducted since the 17th century. Rakugo stories are told by a *Hanashika* (the narrator), who personifies all the characters in a story. Both the personification part and the narration part result in the modification of voice by a *Hanashika*. Speech styles used in performance arts are often understudied. Some exceptions exist, however: [1], [2] on maid voice and [3] on anime voice.

Modification of speech in performance arts is well recognized by the audience, but identification of such phonetic aspects is less known. [1] reports that maid voice can be produced with raised F0, more dynamic F0 movement, or exaggerated vowels. Results in [2] show that anime voice is produced with stronger harmonics than normal speech.

In this paper, we focus on the speech production in non-Rakugo vs. Rakugo speech style and investigate the similarities and differences between the two, such as vowel quality, durational differences, vowel devoicing, and pitch range.

The rest of the paper is organized as follows. In section 2, methods for data collection are described. The results in section 3 demonstrate that Rakugo and non-Rakugo speech styles are not identical. We conclude the paper with a discussion section.

2. METHODS

2.1. Stimuli

32 randomized sentences containing target words in Table 1 were produced 3 times each by 8 Kanto Japanese speakers (in two different styles: Rakugo and non-Rakugo. The frame sentence was *kono ___-ttenaa, donna imidai* for the Rakugo style and *kono ___-to yuu tangono imiwa nandesuka* for the non-Rakugo style (both means "What is the meaning of this word _____ ?"). The final number of tokens is 1536 (32 target words x 8 participants x 3 repetitions x 2 styles).

target	accented	unaccented
na	*naka* 'middle' *nana* 'seven'	*naki* 'crying' *nami* 'moderate'
ni	*niki* '2 terms' *niji* '2 o'clock'	*nishi* 'west' *niji* 'rainbow'
nu	*nuka* 'nuka' *nugu* 'taking off'	*nuke* 'missing' *nuno* 'cloth'
ne	*negi* 'green onion' *neko* 'cat'	*neta* 'content' *nekki* 'heat'
no	*nodo* 'throat' *noshi* 'noshi'	*noki* 'eave' *nozomi* 'hope'
ta	*tane* 'seed' *take* 'length'	*tana* 'shelf' *take* 'bamboo'
te	*teko* 'lever' *teba* 'wings'	*teki* 'enemy' *tetsu* 'iron'
to	*toki* 'Japanese Ibis' *tomo* 'friend'	*toku* 'benefit' *tomoni* 'together'

Table 1: A stimuli table varying in onset nasality and vowel types.

2.2. Recording sessions

We recruited four male and four female participants who have performed Rakugo storytelling as a member of a Rakugo club (mean years of experience = 3.5). The age of participants was 18 to 23 at the

time of recording (mean = 21.2). All participants utilize Edo Rakugo, a style that uses the Edo dialect (old Tokyo Japanese).

The participants produced the Rakugo style speech while seated in a *seiza* style, where the participant sits on a cushion following the mode of Rakugo performance. The non-Rakugo speech was elicited while sitting on a chair. After reading a snippet from a Rakugo story for approximately one minute to immerse themselves into Rakugo speech style, participants produced the target items in the Rakugo style. The non-Rakugo sentences were elicited after participants changed their sitting position. The recording was conducted in a quiet room using the audio recording software Audacity.

2.3. Annotation

Target words in the elicited data set were first processed using Praat [4]. Segment-level annotation was conducted manually by examining the periodicity of vowels as well as information from the spectrogram.

A single interval was marked for vowels and nasals. Voiceless plosives were annotated for their closure duration and voice onset time (vot), while voiced plosives were annotated for the closure duration (cvo) and vot. In this paper, boundaries were selected in a conservative manner. An example of an annotated target word is shown in Figure 1.

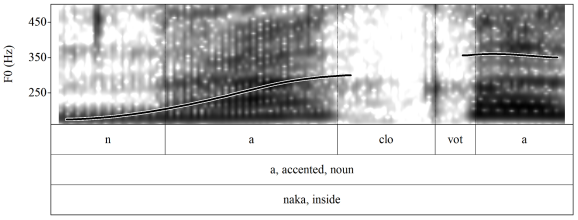


Figure 1: Example of the annotation.

Annotations were conducted by three Japanese-speaking assistants, two of whom were trained in linguistics. All items were double-checked for their annotation accuracy.

3. RESULTS

3.1 Vowel Quality

R [5] was used to create vowel charts to examine the distribution of vowels in the Rakugo and non-Rakugo styles, using the extracted values of F1 and F2 from the mid-point of an annotated vowel interval. Aberrant formant values (74 out of 1515 values) were excluded from analysis by calculating Cook's distance for each pair of gender and vowel. Gender differences discussed in a previous study [6] are also observed in our data, as shown in Figures 2 and 3.

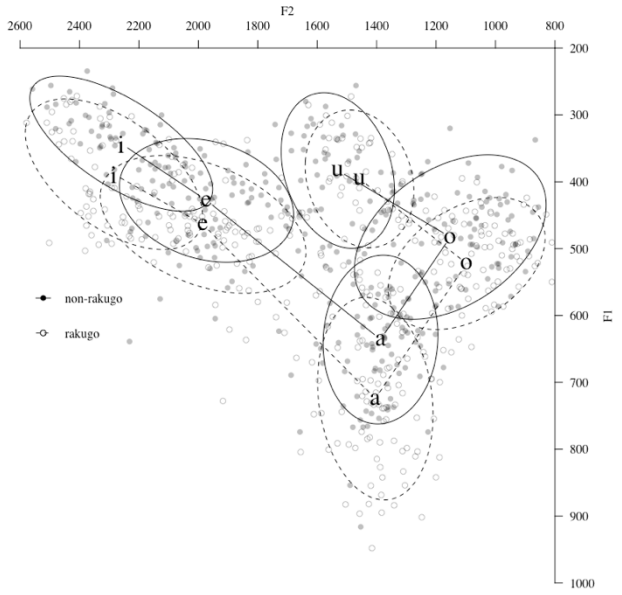


Figure 2: Vowel plot of male speakers.
Non-Rakugo in solid lines and Rakugo in dotted lines.

As Figure 2 shows, male speakers tend to produce vowels in a lower position in the Rakugo style than in the non-Rakugo style. Such lowering tendency has been reported for non-high vowels [o], [e], and [a] in the previous study where syllables without any context are used as stimuli [6]. However, in the current study in which speakers are asked to speak words in different styles, the high vowel [i] is also subject to the lowering tendency in the Rakugo style.

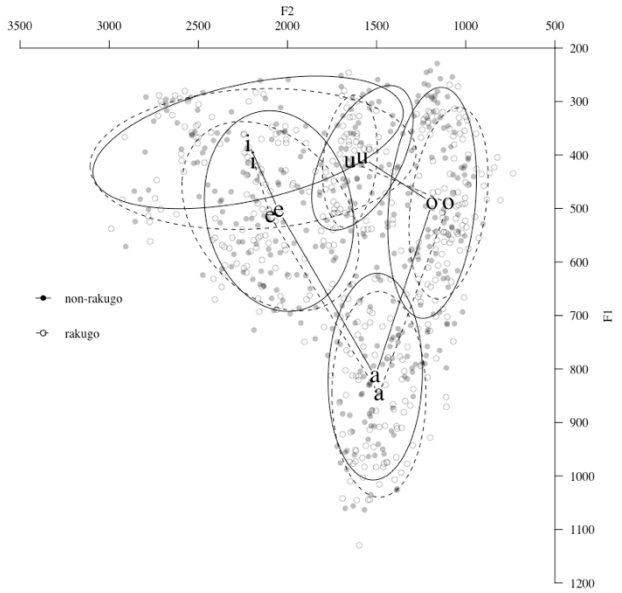


Figure 3: Vowel plot of female speakers.
Non-Rakugo in solid lines and Rakugo in dotted lines.

Although [6] describe that female speakers produce vowels in the Rakugo style in less varied styles, their results do not extend to the vowel pattern in Figure 3 because female speakers show variation to a lesser degree, instead of less variation.

3.2 Duration

An examination of the first syllable in the same target words read in both styles revealed that speakers tend to produce the target syllable longer in the Rakugo style than in the non-Rakugo style. To show this relationship, a correlation plot is created.

Figure 4 displays durational differences between Rakugo and non-Rakugo styles of the same target words. The duration of the first syllable in the non-Rakugo style is plotted on the x-axis, and the duration in the Rakugo style on the y-axis. As shown in the regression line plotted by R, a positive correlation was found (R^2 = 0.35). Round-shaped dots above the line represent cases when a speaker produces the first syllable of a word as longer in the Rakugo style than in the non-Rakugo style. On the other hand, tokens in which a speaker produces the first syllable as longer in the non-Rakugo style are triangle-shaped. The distribution of the dots indicates that Rakugo speakers tend to produce each segment longer in the Rakugo style.

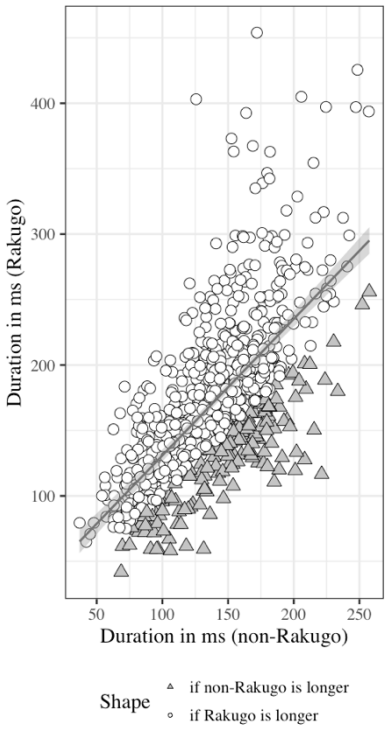


Figure 4: Correlation between non-Rakugo and Rakugo (Duration).

Durational differences between the two styles in individual tokens by speakers are shown in Figure 5.

Six out of eight speakers used durational cues by producing longer vowels (more than 70% of the tokens) in the Rakugo-style speech.

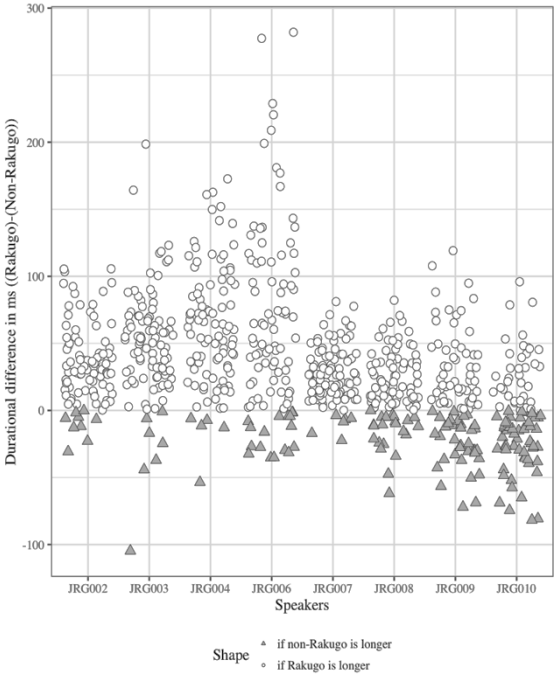


Figure 5: Durational difference between the non-Rakugo and Rakugo styles by Speakers.

		non-Rakugo	Rakugo
naki	unaccented	16 / 24	1 / 24
niki	unaccented	16 / 24	0 / 24
nishi	unaccented	17 / 24	1 / 24
teki	unaccented	18 / 24	0 / 24
tetsu	unaccented	22 / 24	0 / 24
toki	accented	16 / 24	0 / 24
toku	unaccented	18 / 24	0 / 24
	Total	123 / 168	2 / 168
		73.2%	1.1%

Table 2: A table showing tokens with the devoiced vowel in the final syllable.

3.3 High vowel devoicing

High vowel devoicing is a feature of Japanese that is well-reported (cf. [7] and reference therein). The number of words produced with a devoiced vowel in the final syllable is summarized in Table 2. While

target words in non-Rakugo speech undergo high vowel devoicing as expected, the same vowel produced in words with the Rakugo style is almost never produced with a devoiced vowel.

3.4 Pitch range

Speech in performance arts often exhibits widened speech range to increase the transmission of information. To check whether the Rakugo style displays this feature, we calculated the pitch difference between the maximum F0 and minimum F0 of the first and second vowels of each word, considering differing accent patterns. The results show that a larger pitch range is used in the Rakugo style than in the non-Rakugo style.

The data we analyzed consists of words in two accent types: falling pitch in the penultimate syllable (the -2 accent condition) and no falling pitch (the unaccented condition). Words with the -1 accent condition were present in the stimuli but excluded from the analysis due to overlapping effects in pitch accent patterns with unaccented words when the -1 accent words are produced in isolation.

As the words in the -2 accent condition contain a pitch drop between the first and second syllables, we subtracted the value of the minimum pitch of the second syllable from the value of the maximum pitch of the first syllable. Tokens were excluded if the first mora was produced with a lower pitch than the second mora. The distribution of the dots indicates that speakers use a larger pitch range in the Rakugo style than in the non-Rakugo style, as shown below in Figure 6.

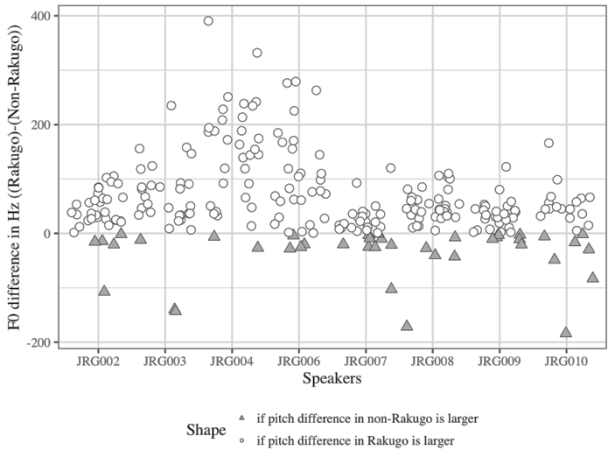


Figure 6: F0 difference in the -2 accent condition between the non-Rakugo and Rakugo styles by Speakers

As unaccented words in Japanese have a pitch rise from a low initial F0, the difference between the minimum pitch of the word-initial mora and the value of the maximum pitch of the second mora was calculated. Most words are plotted above the y = 0 line in Figure 7, which means that the pitch range in the first two morae of unaccented words also becomes larger in the Rakugo style than in the non-Rakugo style.

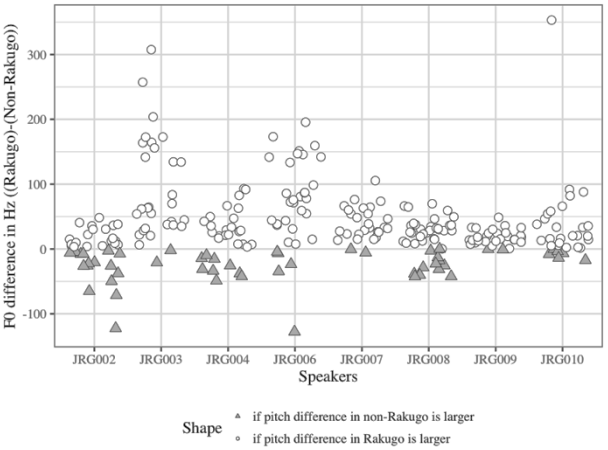


Figure 7: F0 difference in the unaccented accent condition between the non-Rakugo and Rakugo styles by Speakers

4. DISCUSSION AND CONCLUSION

Our research reveals that Rakugo-style speech exhibits dispersed vowels, longer syllable duration, lower rate of high vowel devoicing, and larger pitch range. Although in their normal voice participants devoiced their high vowels, due to the nature of rakugo, this common phenomenon of Tokyo Japanese is not observed in Rakugo style. However, since Rakugo speakers do not acquire but learn Rakugo-style speech, these segmental or suprasegmental modifications in the Rakugo style can be subsumed in an optional effort to convey Rakugo stories correctly and dramatically. The inter-speaker variations support the possibility that the phonetic characteristics in the Rakugo style are not obligatory.

While verbal performance or acting is well-studied in terms of its aspects of history and culture, the speech patterns in the performance often remain understudied. In this study, we focused on the speech patterns displayed in Rakugo, one of such verbal performances. More research on the phonetic characteristics of Rakugo-style speech such as the use of non-modal phonation or the implementation of pitch excursion will further elucidate what makes a Rakugo-style Rakugo.

5. ACKNOWLEDGEMENTS

We thank the Rakugo performers who participated in the data collection. This work was partially supported by JSPS KAKENHI Grant #20K00578

(POP), and the ILCAA joint research project "Phonetic typology from cross-linguistic perspectives (PhonTyp)" (2021-2023).

6. REFERENCES

[1] Kawahara, S. 2013. The phonetics of Japanese maid voice I: A preliminary study. *On-in Kenkyuu [Phonological Studies]* 16, 19–28.

[2] Kawahara, S. 2016. The prosodic features of the 'moe' and 'tsun' voices. *Journal of the Phonetic Society of Japan* 20(2), 102-110.

[3] Utsugi, A., Wang, H., Ota I. 2019. A voice quality analysis of Japanese anime. Proceedings of the *International Congress of Phonetic Sciences* 19, 1853-1857.

[4] Boersma P., Weenink, D., 2019. Praat: doing phonetics by computer. Accessed: Dec. 17, 2019. [Online].

[5] R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[6] Kamano, S., Tsujita, R., Sakamoto, C., Lee, S. J. 2022. Vowel production in Rakugo Speech. *Proceedings of the 36th General Meeting of the PSJ*, Kobe, Japan. 151–155.

[7] Fujimoto, M. 2015. Vowel devoicing. In: Kubozono, H. (ed) *Handbook of Japanese Phonetics and Phonology*, De Gruyter. 167–214.

Variation in the Scottish BIT vowel: Comparing two corpora

Lauren Hall-Lew,[1] Zuzana Elliott, [1] Jessica Göbel,[2] Claire Cowie, [1] and Nina Markl[1]

[1]The University of Edinburgh and [2]New York University
lhlew@ed.ac.uk, zuzana.elliott@ed.ac.uk, jessica.goebel@nyu.edu, claire.cowie@ed.ac.uk, nina.markl@ed.ac.uk

ABSTRACT

We analyse variation in the Scottish BIT vowel as produced by men from different birth years and social classes. An analysis of two Edinburgh-based speech corpora allows for a comparison of white Scottish men born between 1899 and 1996. The results indicate that BIT may be raising over time among New Middle Class and Working Class speakers. However, the conclusion is complicated by differences between the two corpora, including recording technology, addressee, register, and topic. These are just as likely to account for the patterns in BIT F1 found with respect to corpora and social class group. The paper calls for further work on the social and stylistic differences between corpora used in real time studies.

Keywords: sociophonetics, social class, style, vowels

1. INTRODUCTION

Examples of stable phonetic variables are rarer than examples of sound changes [10], with famous English examples including *-ing* velar/alveolar alternation and consonant cluster reduction. One question is how to best account for stylistic and other social changes over time which co-occur with the passage of time, itself. In this paper, we consider variation in the height and anteriority of vowels, in particular the vocalic variable known as English KIT [29] or Scots BIT [27]. Both Scottish English and Scots, and mixed styles between them, are spoken in Edinburgh. By examining speech from two Edinburgh corpora, with respondents born between 1899 and 1996, we find variation in KIT/BIT over time among Working Class speakers. We consider whether this indicates a change in progress or if stylistic differences between corpora may better account for the difference.

The Scottish BIT vowel is lower and backer than the standard English KIT vowel, so much so that it is often merged with STRUT; this is the case in general and in Edinburgh, in particular [22]. The variation correlates with socioeconomic class, with middle class speakers producing a higher and fronter (i.e., more 'English') variant than working class speakers [25]. This pattern has also been documented in Glasgow, in the 1970s [18] and the 2000s [7, 26].

Despite this consistent stratification across the decades, previous work has suggested that the BIT vowel might be undergoing a change in progress. Data from Glasgow in the early 2000s found a lower vowel among middle class boys than middle class men, and a backer vowel among all boys than among all men, suggesting a lowering and/or backing change in progress [7, 8], or possibly an age graded pattern. Here, we consider data from Edinburgh which may suggest a change in progress. However, the conclusion is complicated by social and stylistic differences between corpora.

2. THE CORPORA

The present paper compares data from two corpora of Edinburgh speech, which we refer to as the Esling Corpus and the Lothian Diary Corpus. The corpora do both contain uses of Scots, but most of the speech can be best described Scottish English, albeit ranging the spectrum from very non-standard to very standard.

The Esling Corpus [9] was collected by John Esling in 1975. It consists of sociolinguistic interviews in the tradition of Labov [15] and adapted for the Scottish context. The participants were white Scottish men, stratified by socioeconomic class. Data collection took place in participants' homes, over cups of tea, with family, friends, and pets coming and going during the recordings. About this atmosphere, Esling [9] writes, "it is extremely helpful in eliciting casual speech from the informant." The interview topics included childhood games, school memories, and memories about the local area.

At the time, Esling, an American, was an PhD student at the University of Edinburgh, and the interviews were framed as contributions to a student project. Esling [9] writes, "[a]ll informants appeared favourably disposed towards the interviewer's American accent, and this is thought to have been an advantage in the Scottish situation in encouraging informal speech." Impressionistically, this seems to be true; there is no indication that speakers converged phonetically towards Esling's American accent, nor towards a more educated or middle-class accent that might have been primed by his university affiliation. Most of the speakers were older than Esling was, at the time of recording, and Esling positions them as local experts and in the position of authority throughout the interaction. Speech acts such as joking and teasing, and channel cues [15] such as laughter, indicate that Esling was highly successful in evoking

a casual speaking style from the participants. The read-speech portions of the interview only occurred at the very end of the interaction, and participants seem unaware of attention to their speech beforehand. John Esling generously contributed the recordings to the University of Edinburgh's 'Edinburgh Speaks' archive in 2017, for further analysis.

The Lothian Diary Corpus [13, 14] was collected by a team of researchers from the University of Edinburgh between 2020-2021, and consists of audio or video self-recordings about the effects of the COVID-19 lockdowns on daily life. Self-recordings are increasingly common but remains understudied in sociophonetics [14]. In contrast to the Esling Corpus, the Lothian Diary participants represent a wide range of genders, socioeconomic classes, ages, and other social factors, but all were living in the greater Edinburgh area at the time of recording. Data collection took place wherever the participant chose, often in their home, but not always, with no direct communication between the research team and the participant prior to recording. Many participants spoke in response to prompts on the project website (e.g., "How has your life changed in lockdown?"), but many others did not refer to these questions.

The recordings were framed on the project website as contributions to (a) research on individual experiences with the Scottish lockdown measures, (b) a report for Scottish Parliament, and (c) a COVID-19 oral history collection for the City of Edinburgh. As a result, the speaker's potential *imagined audience* [2] was quite varied and diverse; it might have included the research team members, members of the Scottish Parliament, the general public, or all three. Speakers might have oriented to a more standard accent, due to the researchers' affiliation with the university, or they may have instead oriented to their role as authority and expert (on the topic of their personal experience). While some speakers produced the occasional laugh, most contributions are impressionistically sombre and serious, often reflective and philosophical. Others are delivered with a flat affect, listing the changes between life before and during the lockdown. Read speech was never directly elicited, but some participants read the website prompts aloud interspersed with their more spontaneous answers.

3. METHODS

Following Esling [9], social class was coded as ternary, which better reflects the reality of social class experience in Edinburgh than the binary coding often seen in other studies. In Edinburgh, social class corresponds, among other things, to a speaker's level and location of education, their type of work and income level, their parents' type of work and income level, their level of wealth, and the level of social deprivation characterizing their neighbourhood. Esling's [9] three levels included 'I' (middle class, from the Morningside neighbourhood), 'III' (working class, from the Pilton neighbourhood), and 'II' (in between, e.g., working class origins but living in Morningside). These levels, and the Lothian Diary participants, were (re)coded as in [6]: the Established Middle Class ('EMC', speakers whose parents were also middle class; Esling's class 'I'), the Working Class ('WC', speakers whose parents were also working class; Esling's class 'III'), and the New Middle Class ('NMC', upwardly mobile speakers who were born to working class parents, but middle class at the time of data collection; Esling's class 'II'). The NMC primarily captures upward social mobility, but also includes stable members of the Lower Middle Class ('LMC', see [6, 16]).

For the current paper, we analysed a subset of 13 Esling Corpus speakers born between 1899-1953, aged 22-76 at the time of recording. Four were NMC and four were WC. We compare these speakers to a subset of 8 Lothian Diary Corpus speakers, also white Scottish men, born between 1964-1996, ages 24-66 at the time of recording. Three were NMC and two were WC (the only available speakers of that demographic in the corpus). The Lothian Diary Corpus speakers are all cis-gender and straight; this information was not collected for the Esling Corpus speakers but all of the 13 were in visibly heterosexual marriages at the time.

Tokens of the Scottish BIT vowel in the Esling Corpus were hand-annotated by the third author, and midpoint vowel formants and durations were hand-measured in Praat [3]. Since analysis of the Lothian Diary Corpus speakers is part of a much larger project, tokens of the same variable in the were annotated automatically using the Montreal Forced Aligner [19], and midpoint vowel formants and durations were automatically extracted using a script in Praat. Ten percent of automated annotations were checked for accuracy, though very few annotations in the sample required manual adjustment.

Valid tokens of BIT were those carrying primary lexical stress, not immediately following vowels, and not immediately preceding /l/ or /r/ (total N=945). Models were built in *R* [23], using the *lme4* package [1]. Initial model factors included vowel DURATION (log normalized), lexical FREQUENCY (based on the spoken British National Corpus), PRECEDING phonological place (*labial, apical, dorsal, pause*), FOLLOWING phonological place (*labial, apical, dorsal*), speaker AGE (continuous, 22-76), and social CLASS (*EMC, NMC, WC*), and CORPUS (*Esling, LothianDiaries*). Speaker year-of-birth is not included as it is nested within corpus. WORD and SPEAKER were entered as random intercepts; models

with random slopes did not converge. Models were fit to F1 and F2 separately. Model fitting used sum-coding and was based on drop-one comparisons of an initially maximal model with only logically predicted interaction factors included.

4. RESULTS

The best fits models for each formant of BIT were:

F1 ~ LOGDURATION + CORPUS * CLASS
F2 ~ LOGDURATION + PRECEDING + FOLLOWING

Thus, the preceding and following phonological environment significantly predicted vowel anteriority (F2), and the interaction between the corpus and social class significantly predicted vowel height (F1). Vowel duration is a significant predictor for both variables. Vowels are both lower and backer when produced with longer durations [28].

The non-significant predictors for both variables are lexical frequency and speaker age. Neither phonological environment factor predicted variation in F1, and neither corpus nor socioeconomic class were significant predictors for F2. The reasons for the effects of duration on both formants, and the effect of phonological environment on F2, are phonetically straightforward. We therefore concentrate for the rest of this paper on analysing the significant interaction effect for F1 (Figure 1).

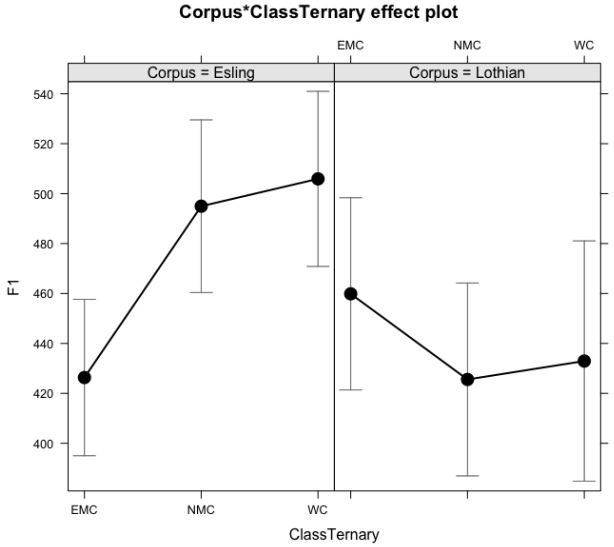


Figure 1: Second significant predictor (after vowel duration) for variation in the F1 of the Scottish BIT vowel among Edinburgh speakers of two corpora. Social class predicts vowel height, but in opposite ways by corpora.

Corpus, alone, is a significant predictor, with BIT being lower overall in the Esling Corpus than the Lothian Diary corpus. Speaker social class, alone, is not a significant predictor. This is not surprising when class is viewed in interaction with corpus (Figure 1). While there is no significant difference between corpora among Established Middle Class speakers, New Middle Class and Working Class speakers both produce significantly higher BIT vowels in the Lothian Diary Corpus than in the Esling Corpus. In the Esling Corpus, NMC and WC BIT vowels are lower (with a larger F1) than EMC BIT vowels, in keeping with all other studies of BIT in the Scottish Central Belt. However, in the Lothian Diary Corpus, the reverse is true. So, while a low BIT vowel has been seen in previous work to be an index of working-class speech, the opposite class pattern obtains for the self-recorded COVID-19 diaries. For this small speaker sample, both the New Middle Class and Working Class produce a higher BIT vowel, closer to the standard English KIT variant, than the variant produced by the Established Middle Class.

5. DISCUSSION

There are several ways to interpret these results. The first is that there has been a change in progress in the height of the KIT/BIT vowel among white Scottish men in Edinburgh, such that the NMC and WC are shifting towards a higher variant over time. This is an entirely plausible analysis, and one that can only be addressed through a longitudinal analysis that controls for differences between corpora.

A second interpretation is that the differences between these two corpora are so great that it is impossible to determine whether or not there is a change in progress here. For this to be likely, the difference would specifically affect the NMC and WC speakers. In this section, we will consider this possibility in more depth, setting the stage for a future analysis that can tease these issues apart.

An issue for any longitudinal analysis is changes over time in recording technology, such as the type and quality of the microphone (see, e.g., [24]). This was held constant in the Esling Corpus. Every recording in the Lothian Diary Corpus was made with a different recording setup, but this information (device and microphone) was collected for each participant, and there does not appear to be any stratification in recording quality according to social class. Future analysis, with more speakers, could include recording device as a factor.

The more striking differences between the Esling Corpus from the Lothian Diary Corpus are stylistic. At least three stand out: addressee, register, and topic. Addressee refers to the person or group of people to whom participants' speech is directed [2]. In the Esling corpus, the addressee is clear: John Esling conducted casual sociolinguistic interviews with

participants and framed himself as a student and the speakers as local experts. There is no evidence that speakers converged to Esling's American accent or to a more middle-class standard, which suggests that speakers maintained a casual speaking style. For NMC and WC speakers, this included a lower BIT vowel. In comparison, the addressee in the Lothian Diary Corpus is less clear. There was no direct communication between the research team and the speakers during the recordings. The recordings themselves were framed as contributions to a COVID-19 oral history collection and a report for Scottish Parliament. Thus, the speakers' imagined audience was vaguer and more formal than in Esling's project. For NMC and WC speakers, the imagined audience is also in a position of relatively more power, while EMC speakers are at roughly the same status position (e.g., compared to members of the university and Parliament). If the former are hypercorrecting towards an EMC norm, then this would account for the production of a much higher BIT vowel than we see for speakers of the same social classes in Esling's data. From a Labovian perspective, this would suggest that the COVID-19 recordings are stylistically comparable to minimal-pair word lists.

The register difference between the two corpora is the difference between an interview and a self-recorded audio/video diary. Previous work on self-recorded speech argues that self-recordings represent an important resource for sociophonetic research due to their propensity to elicit a wider array of stylistic variation than is seen in traditional sociolinguistic interviews [411, 12]. In particular, self-recordings have been seen to elicit more advanced sound change variants than those produced in interview speech [4]. At the same time, self-recordings have also evidenced more hyperarticulated variants than interview speech, namely more fronted /s/ variants produced in Skype calls [11]. In addition, there may be an additional register effect resulting from speech directed to a physically co-present interlocutor and speech directed to a screen (whether self-recorded or not). Both of the working-class men from the Lothian Diary Corpus who were analyzed here submitted video diaries, as opposed to audio diaries, and were looking at the cameras on their recording devices for the duration of the recordings. Other Lothian Diary contributors, not analyzed here, submitted only audio recordings, and future work could compare those submissions to video made by similar participants.

Lastly, the two corpora differ in terms of topic. Esling obtained narratives about childhood games, family memories, and stories of local history. Some speakers mentioned their time in World War II, but spoke more about work and travel than about suffering and death. In contrast, the main topics of Lothian Diary narratives were the COVID-19 pandemic, the stay-at-home order, and the effects these had on people's lives. Although the discussion prompts did ask about 'bright spots' in lockdown, most of the contributions discussed restrictions of movement and increased feelings of fear, anxiety, loneliness, and loss. These topics resulted in a much more sombre mood than in the Esling interviews. It is possible that, for NMC and WC speakers, this affect is indexically linked to a more formal speaking style, and thus a higher BIT vowel (see, e.g., [21, 22]). EMC speakers wouldn't draw on this variation for affective work because they already produce a high, front KIT/BIT vowel as their typical variant.

The potential role of affect in sociophonetics is interesting in light of classic variationist theories of style and crisis narratives. COVID-19 narratives could be considered akin to answers to Labov's [17] famous 'danger of death' question, the answers to which were purported to produce maximally informal speech. The results here suggest that there raise the questions to whether there might be stylistic differences between 'single event' crises (e.g., [5]), versus a more abstract, prolonged, and multi-faceted crises like COVID-19. In the former case, speakers are expected to be highly animated, while in the latter case, we might predict speakers to be withdrawn and subdued. Perhaps there are two types of 'danger of death' question that result in two very different, even opposite, styles. In any case, topic represents a distinct stylistic factor likely affecting results between the two corpora, given how the focus of discussions in each corpus had little overlap.

In short, while there might be a change in progress in the height of BIT, there are too many differences between the corpora to claim this conclusively.

6. CONCLUSION

Based on an analysis of two speech corpora of white Scottish men from Edinburgh, spanning 98 years of birth, there may be a change in progress in the height of the Scottish BIT vowel. If so, this change appears to be a raising of the vowel among the New Middle Class and Working Class, towards the location of the standard English KIT vowel, with no significant change in the speech of the Established Middle Class. However, differences between the two corpora with respect to addressee, register, and topic are just as likely account for the wide variation seen between the speakers of each corpus. Future work should consider other linguistic variables (e.g., *-ing*, /t/-glottal replacement; rhoticity), and other kinds of speakers, in order to understand the effect of social and stylistic differences between corpora used in real time studies.

7. REFERENCES

[1] Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4x. *Journal of Statistical Software* 67, 1–48.

[2] Bell, A. 1984. Language style as audience design. *Language in Society* 13(2): 145-204

[3] Boersma, P., Weenink, D. 2013. Praat: Doing phonetics by computer [Computer program]. Version 5.3.56. Retrieved August 2014 from http://www.praat.org/.

[4] Boyd, Z., Elliott, Z., Fruehwald, J., Hall-Lew, L., Lawrence, D. 2015. An Evaluation of Sociolinguistic Elicitation Methods. In: The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow.

[5] Clark, L., MacGougan, H., Hay, J., Walsh, L. 2016. "Kia ora. This is my earthquake story" Multiple applications of a sociolinguistic corpus. *Ampersand* 3, 13–20.

[6] Dickson, V., Hall-Lew, L. 2017. Class, Gender and Rhoticity: The Social Stratification of Non-prevocalic /r/ in Edinburgh Speech. *Journal of English Linguistics* 45(3), 229-259.

[7] Eremeeva, V. 2002. *A Sociophonetic Investigation of the Vowels OUT and BIT in Glaswegian*. PhD dissertation, University of Glasgow.

[8] Eremeeva, V., Stuart-Smith, J. 2003. Sociophonetic Investigation of the Vowels OUT and BIT in Glaswegian. In: Solé, M. J., Recasens, D., Romero, J. (eds.), *15th International Congress of Phonetic Sciences*. Barcelona, Spain. ISBN 1-876346-48-5.

[9] Esling, J. H. 1978. *Voice Quality in Edinburgh: A Sociolinguistic and Phonetic Study*. PhD dissertation, University of Edinburgh.

[10] Gardiner, S., Nagy, N. 2017. Stable variation vs. language change and the factors that constrain them. *University of Pennsylvania Working Papers in Linguistics* 23(2), 10.

[11] Hall-Lew, L., Boyd, Z. 2017. Phonetic Variation and Self-Recorded Data. *University of Pennsylvania Working Papers in Linguistics* 23(2), 86–95.

[12] Hall-Lew, L., Boyd, Z. 2020. Sociophonetic perspectives on stylistic diversity in speech research. *Linguistics Vanguard* 6, s1.

[13] Hall-Lew, L., Cowie, C., McNulty, S. J., Markl, N., Lai, C., Llewellyn, C., Alex, B., Fang, N., Elliott, Z., Klinger, A. 2021. The Lothian Diary Project: Investigating the Impact of the COVID-19 Pandemic on Edinburgh and Lothian Residents. *Journal of Open Humanities Data* 7(4), 1–5.

[14] Hall-Lew, L., Cowie, C., Lai, C., Markl, N., McNulty, S. J., Liu, S. S., Llewellyn, C., Alex, B., Elliott, Z., Klingler, A. 2022. The Lothian Diary Project: Sociolinguistic Methods during the COVID-19 Lockdown. *Linguistics Vanguard* 8(s3), 321-330.

[15] Labov, W. 1966a. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

[16] Labov, W. 1966b. The effect of social mobility on linguistic behavior. *Sociological Inquiry* 36(2), 186-203.

[17] Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

[18] Macaulay, R. K. S. 1977. *Language, social class and education: A Glasgow study*. Edinburgh: Edinburgh University Press.

[19] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 498–502.

[20] Millar, R. M. 2018. *Modern Scots: An Analytical Survey*. Edinburgh: Edinburgh University Press.

[21] Podesva, R., Callier, P., Voigt, R., & Jurafsky, D. 2015. The connection between smiling and GOAT fronting: Embodied affect in sociophonetic variation. In: The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow.

[22] Pratt, T. 2023. Affect in sociolinguistic style. *Language in Society*, 52(1), 1-26.

[23] R Core Team. 2020. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Available at www.R-project.org

[24] Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., and Bowern, C. 2021. (Don't) try this at home! The effects of recording devices and software on phonetic analysis. *Language* 97(4): e360-e382.

[25] Speitel, H. H., Johnston, P. 1983. *A sociolinguistic investigation of Edinburgh speech*. Social Science Research Council End of Grant Report. C/00/23/0023/1.

[26] Stuart-Smith, J. 2003. The phonology of modern urban Scots. In: Corbett, J., McClure, J. D., Stuart-Smith, J. (eds.), *The Edinburgh companion to Scots*, 110-137. Edinburgh: Edinburgh University Press.

[27] Stuart-Smith, J. 2004. Scottish English: Phonology. In: Kortmann, B., Burridge, K., Schneider, E. W., Mesthrie, R., Upton, C. (eds.), *A handbook of varieties of English, vol. 1, Phonology*, 47-67. Berlin: Mouton de Gruyter.

[28] Toivonen, I., Blumenfeld, L., Gormley, A., Hoiting, L., Logan, J., Ramlakhan, N., Stone, A. 2015. Vowel height and duration. In: *Proceedings of the 32nd west coast conference on formal linguistics* 32, 64-71. Somerville, MA: Cascadilla Proceedings Project.

[29] Wells, J. C. (1982). *Accents of English beyond the British Isles*. Cambridge: CUP.

THE VOICELESS VELAR STOP /k/ IN RIJAL ALMA ARABIC: REVISITED

Leila Yaqoub[1], Sam Hellmuth[2], George Bailey[3]

[1]King Khalid University, [1, 2, 3]University of York
lyacob@kku.edu.sa, sam.hellmuth@york.ac.uk, george.bailey@york.ac.uk

ABSTRACT

An earlier documentation study showed that the voiceless velar stop /k/ is articulated in Rijal Alma Arabic (spoken in southwest Saudi Arabia) as a post-palatal fricative /x̟/, a phoneme absent from most other Arabic varieties. However, a recent sociolinguistic study in Rijal Alma revealed patterns of variation and change in use of post-palatal /x̟/ in favour of the supra-local voiceless velar stop /k/.

To better understand variation and change in the speech patterns of speakers of Rijal Alma, this paper presents an auditory and acoustic analysis of the realisation of /k/ in three positions (word-initial, intervocalic, and word-final). The data are from wordlist fieldwork recordings of 33 target items produced by 43 Rijal Alma speakers (24M/19F aged 18-60). Results indicate that target /k/ displays more complex phonetic properties than previously reported, pointing to a range of different realisations, which appear to be both auditorily and acoustically distinct: [k x kx xk].

Keywords: Rijal Alma dialect, velar stop /k/, auditory realisation, acoustics measurements.

1. INTRODUCTION

Rijal Alma (RA) is a variety of Tihamat Asir Arabic spoken in the RA province of Saudi Arabia in the south-west with a population of approximately 65406 [1].

The phonology of RA has been reported to differ from Modern Standard Arabic (MSA) and most Arabic dialects in a number of respects. A notable difference is the absence in RA of the voiceless velar stop /k/. In most Arabic varieties, the voiceless velar stop /k/ is a common sound. However, in some Arabic varieties, the voiceless velar stop /k/ can also be realised as either affricate [t͡ʃ], as in some rural dialects of the Levant, parts of Jordan and Iraq dialects and Gulf dialects, or as affricate [t͡s], as in Najdi Arabic [2], where the occurrence of [t͡ʃ] or [t͡s] is conditioned mainly in front vowel environments [3].

In contrast, in RA, the voiceless velar stop /k/ has been proposed to manifest instead as a fricative sound [4, 5, 6, 7, 8]. Prochazka [4] described the realisation of voiceless velar stop /k/ in RA as a voiceless uvular /χ/. A another description by Hamzah cited in [6] suggested that target /k/ in RA is perceived as a mixture of two fricatives /ʃ/ and /χ/.

An acoustic analysis conducted by Asiri [5, 6] describes the production of target /k/ by native RA speakers as a post-palatal fricative [x̟], occurring in all phonological environments. Use of the local [x̟] was, however, observed to be reducing in favour of the voiceless velar stop /k/ particularly among educated speakers [7]. There is, however, a lack of evidence regarding how the observed variants of local [x̟] are phonetically characterised. In the previous studies [5, 6], the focus was only on the phonological behaviour of the local [x̟], with only limited acoustic investigation.

This paper presents the first preliminary results of a wider study into linguistic variation in Rijal Alma. Our focus here is the evidence for variation in the phonetic realisation of the local variant [x̟] in the speech of native speakers of Rijal Alma, by examining the auditory and acoustic characteristics of their speech.

2. METHODS

2.1. Participants, Materials and Procedure

The data for this study stems from recordings of 43 RA speakers, comprising two genders and are balanced by age (24 men and 19 women aged 18 to 60). The participants were asked to read 33 words as naturally as possible. The target words are placed in a carrier phrase, “ʔana ʔagool [target word] marratti:n” ~‘I say [target word] twice’.

Each participant was recorded on location in RA using a Marantz model PMD660 with an external microphone (Shure SM10 professional unidirectional head-worn dynamic microphone). The recording was saved in WAV format, with a sampling rate of 44.1 KHz 16 bit.

2.2. Data Preparation and Auditory Analysis

The data was phonetically transcribed using Al-Tamimi Romanization system (ATR) [9] in ELAN [10] then forced aligned at the phone level using WebMaus [11] for further acoustic analysis. Manual auditory coding was performed by the first author. Auditory judgements were made using Praat [12] with close reference to the waveform and

spectrogram. The local variant [x̟] was found to display a number of complex phonetic structures that can be grouped into three variants.

A token was coded as a plosive [k] if it has three clear phases of articulation; closure, burst, aspiration (as in Figure 1(a)). Tokens with aperiodic energy typical of a fricative, and perceived as a fricative, were coded as a fricative [x] [i](Figure 1(b)). In some cases, however, potential tokens of [k] had heavy aspiration or frication noise after the closure phase and were perceived auditorily as a combination of stop and fricative [kx] (Figure 1(c)). In addition, there was a further variant classified as [xk], in which there is a period of aperiodic frication before the transient burst begins, and in some cases, an incomplete closure phase is present (Figure 1(d)). The two variants [kx] and [xk] are labels used during coding for ambiguous or intermediate variants which exhibit some characteristics of both plosive and fricative sounds; they comprise 11% of the sample.

Most variants often involved a glottal presence in which it appeared that two constrictions were occurring, one at the glottis and the other at one or more phases of articulation (e.g., before or after closure, before or after frication, or after aspiration), though this rarely developed into a non-pulmonary stop such as clicks.

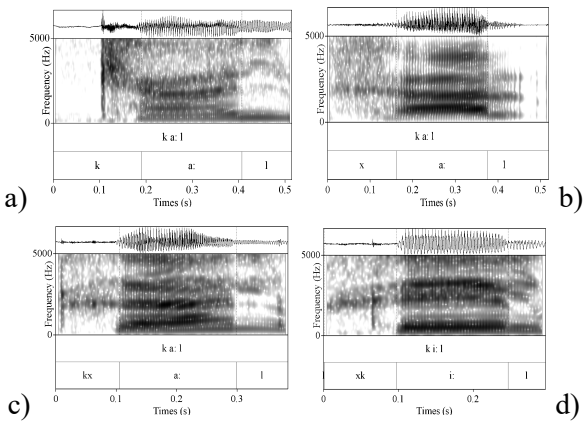


Figure 1: Waveforms and spectrograms of [ka:l] 'weight': a) a plosive variant [k]; b) a fricative variant [x], and two with intermediate variants c) [kx] and d) [xk].

2.3. Acoustic Measurements

Acoustic measurements were also taken to corroborate the auditory coding and provide a better basis for distinguishing tokens heard as plosive [k], fricative [x] or intermediate [kx]/[xk]. Following prior literature on measuring the duration of obstruent consonants [13, 14, 15, 16, 17, 18] each variant [k], [x], [kx], or [xk] was measured according to the various acoustic events associated with each variant (closure, burst, aspiration/frication). This allows comparison of the duration of a certain phase between variants and provides more comparable results to prior literature regarding the distinction between stops, affricates and fricatives. In this report, non-pulmonic variants were excluded due to their small number (0.96% of the sample).

The whole segment duration of variants coded as plosive [k] was defined based on its acoustic events (closure, burst, aspiration), in which the onset and offset of each phase was marked (see Figure 2). In cases where multiple release bursts were observed, the whole segment duration was taken from the first to final transient at the release.

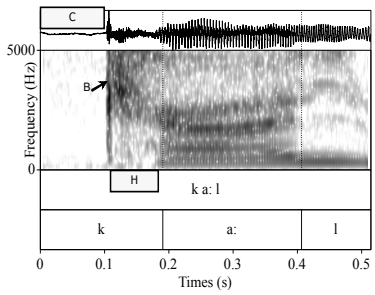


Figure 2: Example of segmentation for [k], showing hold phase (C), burst release (B) and aspiration (H).

As for the variant coded as [x], there was no closure or burst, and the whole segment duration was defined as the interval between the onset and offset of the frication noise. The whole segment duration for variants coded as [kx] or [xk] were defined in a similar manner to the durations defined for the variant coded as [k], where the time between onset and offset of each phase of articulation was defined. However, in cases where no closure phase occurred, only burst and aspiration duration measurements were recorded.

The second acoustic measurements used in this study is rise time to gain a further insight in distinguishing different possibilities among the sound realised as a stop and other obstruent that were identified auditorily. It has previously been suggested that the rise time of the noise energy of obstruent is an important cue to the manner of articulation [17]. In this document, a rise time is measured by obtaining the duration from the time of onset of aspiration/frication noise to the time at which peak amplitude reaches its maximum.

The acoustic measurements (durations and rise time) were analysed statistically using mixed-effects linear regression model in the lme4 package [19] in R to assess whether there were any overall differences among the variant types. The data was fitted with the following model (DV (acoustic measurements) ~ auditory label * word_position + (1 | word) + (1 | speakerID). Acoustic measurements are the dependent variable, which is expressed as a

continuous variable. The fixed effects are auditory labels, and word position. Words and speakers were fitted as random intercepts in order to account for individual differences.

3. RESULTS

The data examined in this document were derived from the wordlist task outlined above. The sample included 24 males and 19 females, balanced by age. Specifically, the younger participants were aged between 18 and 35 years, while the older participants were 40 years or older. The total number of auditorily labelled tokens from the wordlist data included 2888 tokens. Among the 2888 tokens, there were 2152 (74%) tokens were realised as [k], of which 48% were produced by younger participants and 26% by older participants, 447 (15%) tokens were realised as [x] , 7 % are produced by younger participants and 8% by older participants, 157 (6%) tokens were realised as [kx], 2% of them are produced by younger participants and 4% by older participants, and 132 (5%) tokens were realised as [xk], 2% of which are produced by younger participants and 3% older participants.

3.1 Duration

The distribution of durations for the whole segment of all variant types are presented in Figure 3. The mean duration of tokens labelled [k] (annotated on the plot) appears to be longer than that of other variants.

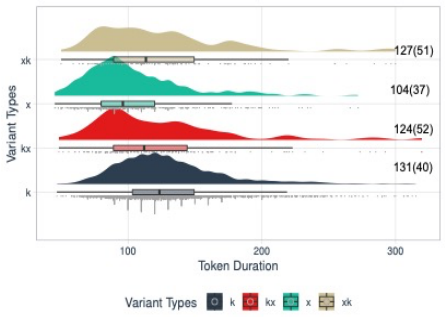


Figure 3: Mean and standard deviation (ms) given in brackets of Total segment duration for all variants.

A closer look at the durations of each component phase is provided in Table 1. As predicted, tokens with intermediate variants [kx, xk] appear to have a shorter mean duration of closure and burst than tokens coded as a stop [k]. Further, the mean duration of aspiration for the intermediate variants [kx, xk], is longer than the mean duration of aspiration for the variant coded as a stop [k]. In a comparison of tokens realised auditorily as fricatives [x] and intermediate variants [kx, xk], it was found that the mean duration of aspiration for intermediate variants was shorter than that for tokens realised auditorily as fricatives [x].

Variant:	k	x	kx	xk
Closure	69.67 (27.8)	- -	28.9 (31)	22 (28.9)
Burst	12.1 (9.09)	- -	12.7 (12)	9.57 (9.56)
Aspiration	47.7 (23)	- -	82.3 (44. 1)	94.9 (50.6)
Frication	- -	104 (35.9)	- -	- -

Table 1: Mean and standard deviation (ms) (in brackets) of closure, burst, aspiration and frication durations by variant.

A statistical analysis performed by using linear mixed effects models confirms the observations from the raw data; closure duration was significantly shorter for intermediate variants [kx] (β= -7.005, SE = 3.418, t = -2.049, p < 0.04), and [xk] (β= -32.424, SE = 3.009, t = -10.777, p < 0.001) than for the stop [k], across all different word positions (see Figure 4). There was no statistically significant difference between the burst durations of the stop and intermediate variants.

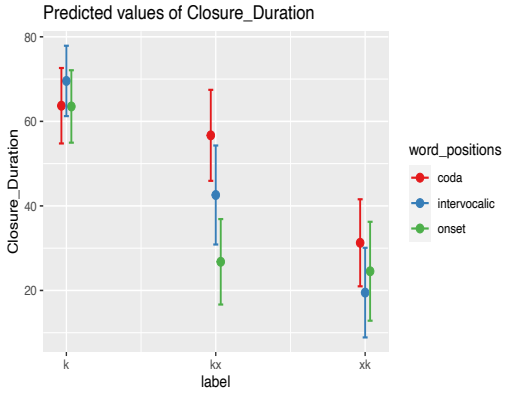


Figure 4: Predicted mean and 95% CI of closure duration for variants [k, kx, xk] by word position.

As for aspiration duration, the statistical results show a significant effect with a longer duration for the intermediate variants [kx] (β= 28.325, SE = 3.519, t = 8.048, p < 0. 001), and [xk] (41.583, SE = 3.097, t = 13.428, p < 0.001) compared to the variant [k], across all different word positions. It appears that these results are in agreement with the literature concerning durational differences between stops and affricates. According to [14] when affricates are compared with stops of the same voicing category and manner of articulation, the noise stage appears to be lengthened while the silence stage is shortened.

Furthermore, the statistical results confirm the longer duration for the fricative form [x] in comparisons with the intermediate variants (β= 27.913, SE = 5.226, t = 5.186, p < 0. 001) compared to [kx], and [xk] (-13.252, SE = 4.897, t = -2.706, p < 0.007) is significantly shorter than [x]. In light of this, the duration means of the fricative variant [x] is longer than that of intermediate variants, which further indicates that intermediate variants might be classified as affricates. Considering the literature that examines the durational differences between affricates and fricatives, this is an expected result in which it suggests that the affricate sounds are found to be distinguished by a shorter duration for the fricative portion than for the corresponding fricative part [14, 15].

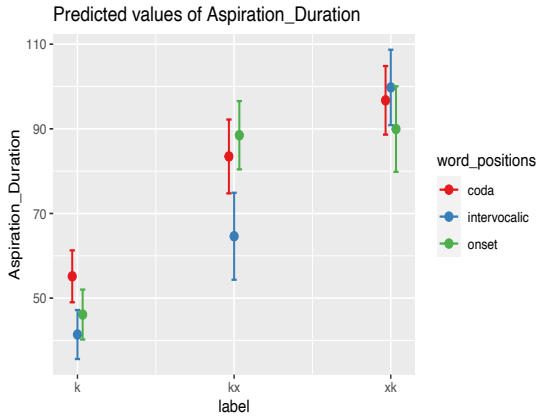


Figure 5: Predicted mean and 95% CI of aspiration duration for variants [k, kx, xk] by word position.

3.2. Rise Time of Stop - Intermediate - Fricative

The values of rise time by variant are presented in Figure 6 and Table 2. The table contains the mean values and (SD) of rise time in aspiration/frication phases across 2152 [k], 447 [x], 157 [kx], and 132 [xk] tokens from 43speakers. In general, the rise time variation between intermediate and fricative variants does not appear as large as the variation between intermediate, stop variants. However, there are still noticeable differences that may be better accounted for by a linear regression model.

Variant:	k	x	kx	xk
Rise time	21.2 (21.3)	48.6 (55.5)	52.7 (45.5)	42.1 (34. 4)

Table 2: Mean and standard deviation (in brackets) of rise time (ms) by variant.

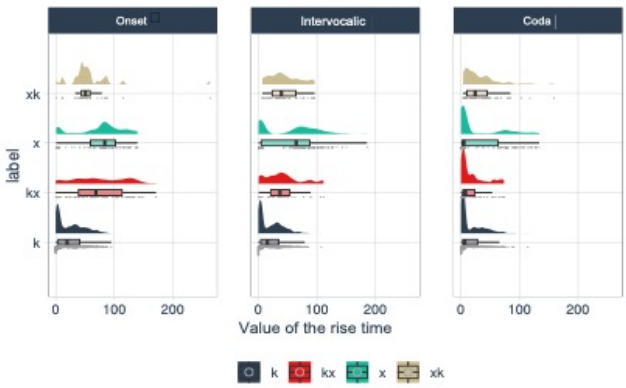


Figure 6: Distribution of rise time values (ms) for each variant type.

A mixed-effects linear regression revealed a significant difference with a longer rise time for the [x] (β= 9.432., SE = 4.346, t = 2.170, p < 0.03) relative to intermediate variant [kx], but no statistically significant difference between [x] and [xk]. Additionally, a significant difference is found between the two intermediate variants with a longer rise time for the variant [kx] (β= 22.769, SE = 2.229, t = 10.21, p < 0.001) compared to the plosive variant [k], and a longer rise time for the intermediate variant [xk] (β= 19.554, SE = 2.130, t = 9.18, P < 0.001) relative to the plosive variant. Accordingly, the plosive variant [k] exhibits the most rapid build-up of acoustic energy, followed by intermediate and then fricative variants. According to the wider literature examining how the obstruent can be phonetically distinguished, this seems to be an expected result, which indicates that noise energy is gradually built up in fricative consonants; it increases faster in affricates than in fricatives, but not as rapidly as in stop consonants [14, 15].

4. CONCLUSIONS

In this document, we have attempted to reconsider the evidence for variations in the phonetic realisation of the local variant [x̱] in the speech of native speakers of Rijal Alma, by examining the auditory and acoustic characteristics of their speech. Although these results are in their preliminary status, they indicate quite clearly that the local variant previously described as (x̱) in Rijal Alma possesses complex phonetic patterns. Based on the acoustic measurements presented in this document, it appears that there are four variants of the local variable (x̱) which are acoustically distinct from each other in Rijal Alma. However, a future sociolinguistic study is still needed to investigate possible associations with specific social categories.

5. ACKNOWLEDGEMENTS

Financial support is gratefully acknowledged from King Khalid University, Saudi Arabia.

6. REFERENCES

[1] General Authority for Statistics Kingdom of Saudi Arabia. (2022). *Population and Housing census.* Available: https://www.stats.gov.sa/en/13

[2] Ingham, B., & Henry Sweet Society for the History of Linguistic Ideas, provenance. 1994. *Najdi Arabic: central Arabian / Bruce Ingham.* Amsterdam; Philadelphia: J. Benjamins Pub. Co.

[3] Johnstone, T. M. 1967. *Eastern Arabian dialect studies / by T. M. Johnstone.* London; New York [etc.]: Oxford U.P.

[4] Prochazka, T. 1988. *Saudi Arabian dialects / Theodore Procházka Jr.* London; New York: Kegan Paul International: Distributed by Routledge, Chapman, and Hall.

[5] Asiri, Y. 2009. "Aspects of the Phonology and Morphology of Rijal Alma'Dialect (south-west Saudi Arabia)", Doctoral dissertation, University of Salford. Manchester

[6] Asiri, Y. 2009. "Remarks on the dialect of Rijal Alma'(South-west Saudi Arabia)". *Wiener Zeitschrift für die Kunde des Morgenlandes*, 9-21.

[7] Assiri, A. 2014. "Sociolinguistic variation in Rijaal Almaʕ, Saudi Arabia: a dialectological study". *King Khalid University Journal for Humanities*, *23*(2), 73–119.

[8] Watson, Janet. 2018. "South Arabian and Arabic dialects". In C. D. Holes (Ed.), *The Historical Dialectology of Arabic.* Oxford University Press. 316–334.

[9] Al-Tamimi, J., Schiel, F., Khattab, G., Sokhey, N., Amazouz, D., Dallak, A., & Moussa, H. 2022. "A Romanization System and WebMAUS Aligner for Arabic Varieties". *Proc.13th Conference on Language Resources and Evaluation (LREC 2022)* Marseille, 7269–7276.

[10] Sloetjes, & Wittenburg, L. 2008. "Annotation by Category: ELAN and ISO DCR". *Proc.6th International Conference on Language Resources and Evaluation*. Retrieved from https://archive.mpi.nl/tla/elan".

[11] Kisler, T., Reichel, U., & Schiel, F. 2017. "Multilingual processing of speech via web services". *Computer Speech & Language*, *45*, 326-347.

[12] Boersma, & Weenink. 2020. *Praat: doing phonetics by computer*.

[13] Johnson, K. 1997. *Acoustic and auditory phonetics / Keith Johnson.* Cambridge, Mass.: Blackwell Publishers.

[14] Hayward, K. 2000. *Experimental phonetics/ Katrina Hayward.* Harlow, Eng.: Longman.

[15] Kent, R. D., & Read, Charles. 2002. *The acoustic analysis of speech / Ray D. Kent and Charles Read.* (2nd ed.). Albany, NY: Singular Press.

[16] Ladefoged, P. 2003. *Phonetic data analysis: an introduction to fieldwork and instrumental techniques / Peter Ladefoged.* Oxford: Blackwell Pub.

[17] Mitani, S., Kitama, T., & Sato, Y. 2006. "Voiceless affricate/fricative distinction by frication duration and amplitude rise slope. *The Journal of the Acoustical Society of America*, *120*(3),1600–1607.

[18] Foulkes, P., Docherty, G., & Jones, M. 2011. Analyzing stops. In D. P. Marianna & M. Yaeger-Dror (Eds.), *Sociophonetics: A Student's Guide*. Routledge, 58–71.

[19] Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. 2015. "Fitting linear mixed-effects models using lme4". *Journal of Statistical Software*, *67*(1), 1-48.

[i] The transcription [x] is used here to indicate a fricative variant of the voiceless velar stop /k/, but this does not imply that it is identical to the target realisations of the phoneme /x/ used in other Arabic dialects, and a detailed analysis of this area is beyond the scope of this paper. A further work focused on the fricative variant in RA is required so that the exact place of articulation can be classified.

A PRELIMINARY INVESTIGATION OF VOCALIC REALIZATIONS OF HISPANIC MICRO-WORKERS FROM LATIN AMERICA

Ioana Vasilescu[1], Yaru Wu[2], Lori Lamel[1], Juana Torres Cierpe[3], Paola Tubaro [4]

[1]LISN, Univ. Paris-Saclay, Orsay, France,

[2]CRISCO/EA4255, Université de Caen Normandie, Caen, France,

[3]Télécom Paris, Institut Polytechnique de Paris, France,

[4]CNRS & CREST, UMR 9194, Palaiseau, France

{ioana.vasilescu, lori.lamel}@lisn.fr, yaru.wu@unicaen.fr, juana.torrescierpe@telecom-paris.fr, paola.tubaro@cnrs.fr

ABSTRACT

This paper focuses on socio-phonetic features of micro-workers supporting the data and annotation needs for artificial intelligence, from several Latin American countries. It represents a first attempt to document the linguistic identity of this emerging professional community. We raise the issue of a group of speakers that might not fall under classical sociolinguistic variables as they have the same activity without sharing a common workplace. Instead, they are spread throughout a large geographical area and likely to cover several regional varieties of Spanish. We make use of a corpus of interviews recorded as part of a larger sociological survey and focus on the acoustic properties of vowels as function of socio-demographic variables. We perform principal component analysis and hierarchical clustering to correlate formant values with socio-demographic variables. The preliminary results suggest vocalic realizations and reduction patterns correlate with level of education and social status.

Keywords: socio-phonetics, LA Spanish, vocalic quality, socio-demographic variables.

1. INTRODUCTION

The present work is a first attempt to document the linguistic characteristics of the emerging professional community of micro-workers supporting the data and annotation needs for artificial intelligence (AI). We exploit a corpus of recorded interviews as part of sociological surveys, to investigate patterns of spoken variation of AI micro-workers from several Latin America (henceforth LA) countries. "Microwork" encompasses a range of small, fragmented tasks performed remotely online by large numbers of providers [1]. It is a by-product of the "datafication" of our societies [2], whereby more data and increasing computational capacity have brought AI research to the fore and in turn, further raised the demand for ever-larger, ever-better data assets. AI micro-workers operate via online platforms and perform tasks as varied as annotating images, recording short sentences or correcting a speech recognition systems' output. Since Labov's founding work [3, 4], the socio-linguistic analysis described patterns of language variation of many languages and from many kinds of data. Therefore, today we know that different demographic and socio-economic groups deploy different linguistic behaviors and characteristics, as do the inhabitants of different geographic areas and particular populations over time. On this basis, it can be assumed that the larger the territory where a language is spoken, the greater the number of linguistic varieties.

This paper addresses the challenge of phonetic particularities of a group of speakers that may not fall under classical socio-linguistic variables. The speakers in our corpus are micro-workers recruited through an international platform open to Spanish speakers and they all declare Spanish as mother tongue. They perform similar tasks without sharing an identified workplace. According to the sociological survey our speakers are coming from different social backgrounds [5]. If some of them are in virtual contact *via* forums related to work on platforms, they are spread through a large geographical area as they are coming from 6 LA countries (Argentina, Colombia, Dominican Republic, Ecuador, Peru and Venezuela). They are thus likely to cover several regional varieties of LA Spanish. We focus here on the acoustic properties of vowels as function of socio-demographic variables.
The remainder of this paper is organized as follows. Section 2 provides a brief survey of patterns of segmental variation in Spanish and motivates the focus on vowels. Section 3 describes the corpus and methodology, followed by an analysis and preliminary results in Section 4. We conclude in Section 5.

2. SPANISH VOWELS VARIATION: A STATE OF THE ART REVIEW

Today Spanish is one of the top 5 most spoken languages in the world and mother tongue for nearly 500 million speakers[1]. With such a worldwide coverage, it benefits from an extensive body of work dedicated to patterns of variation at many linguistic levels and with respect to various socio-linguistic factors [6]. The reason we focus on segmental variation and, in particular, on the acoustic properties of vowels, is twofold. First, we want to contribute to the socio-linguistic study of variation in Spanish of an occupational group hitherto ignored by linguistic explorations but increasingly present in social science studies. State of the art of socio-linguistic perspective points out a range of factors responsible of vocalic variation in Spanish. Among them, the role of bilingualism and language contact has been often mentioned. Speech style is another factor, as well as gender, sexual orientation, social class, and social networks [7, 8]. Our data offer the opportunity to take into account other variables such as education level and social status based on the presence vs. absence of a main occupation, and the reported income.

Second, we aim at contributing to the growing work dedicated to vocalic variation in Spanish. Spanish vocalic system has been described as a particularly stable five-vowel system inherited from Latin. The vowels show more stability compared to consonants that seem to present much more variation especially between Peninsular and LA Spanish [9, 10, 11]. Despite this stability, increasing work is dedicated to patterns of dialectal and socio-linguistic variation of Spanish vowels which seem to be less stable than previously thought [12, 13, 14, 15]. Among the patterns of variation that affect Spanish vowels, reduction is often mentioned, even if other phenomena are also discussed such as vowel harmony [6]. Vocalic reduction leads to weakened vowels (eg. shortened, unstable and even deleted realizations) and is responsible for changes in vocalic quality (formants and voicing). Both lexical stress and word position play a role in vocalic reduction, at least as far as LA varieties are concerned [16, 6], whereas studies show that stress does not have a conventionalized effect on vowel quality in Iberic Spanish [17].

3. CORPUS AND METHODOLOGY

This section presents the data and methodology, and describes the main features of the corpus.

3.1. Corpus acquisition

The dialogues were acquired in the framework of the project TRIA ("The Real voices of Artificial Intelligence"), an interdisciplinary project involving sociologists, economists and linguists from Paris-Saclay University and Institut Polytechnique de Paris. The aim of the project was to document the lived experiences and identity of platform workers from different countries. In particular it focused on the digital labour in Spanish speaking countries, increasingly known as one of the most important reservoirs of micro-work in Europe and beyond[2]. TRIA is a two-part corpus:

1. Written: online questionnaires to document the socio-economic profiles of micro-workers. The surveys consist of over 100 questions covering a wide range of socio-demographic information from gender, age and level of education to patterns of their online earning activities (eg. tasks, platforms used etc.) with the aim of defining their socio-economic profile as comprehensively as possible.
2. Spoken: in-depth interviews with a sub-sample of the same micro-workers, with the objective of corroborating questionnaires. The interviews are scripted to cover the following topics: career path, arrival on the micro-work market, the micro-working tasks and their place in the life of the micro-worker (i.e. the balance between micro-working and everyday life). The wording of the questions as well as the time allocated to each question varies from one interview to another.

The available data correspond to interviews with 57 micro-workers, resulting in approximately 40 hours of speech. For this preliminary work we have processed and we report on 14 hours of speech, corresponding to 18 interviews conducted with 13 male and 5 female micro-workers. They cover the 6 countries mentioned in Section 1: Argentina (4), Colombia (3), Ecuador (2), Peru (1), Dominican Rep. (1), Venezuela (6). Among them 3 micro-workers have a migration history (Spain-Ecuador, Venezuela-Argentina and Venezuela-Colombia). The interviewers are 3 female sociologists (2 Spanish natives, 1 bilingual Italian-Spanish). All the interviews were conducted over videocalls due to the pandemic during the year 2020-2021. Volunteers received compensation for their participation.

3.2. Methodology

Interviews were manually transcribed by a native speaker of Spanish (the fourth author). Then

Corpus	Counts
#duration	14h
#duration speech	11h (80%)
Mean dur./interview	48min.
Mean dur./speaker turn	23min.
#words	90632
#tokens	6441
#tokens micro-w.	5553

Table 1: General description of the TRIA corpus.

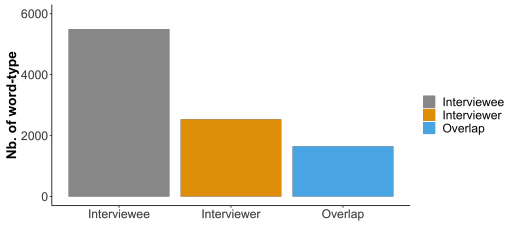


Figure 1: Word-types used by interviewee (gray), interviewer (yellow) and overlapped speech (blue).

the language-specific speech recognition system described in [18] was used to carry out a forced alignment of the speech with the manual transcription. The acoustic models were trained on data from both Peninsular and LA Spanish.

The novelty of our approach is that we take advantage of the socio-demographic variables available through the written survey to enrich the speech with meta-data corresponding to the profile of the micro-workers. Thereafter a mapping is realized between each speaker turn and the corresponding variables. The selected labels, which will later serve as factors are the following: Gender (male, female), Country of residence (Argentina, Colombia, Dominican Republic, Ecuador, Peru, Venezuela), Education level (Secondary (high-school), Post-secondary, Short Higher Education (up to 2 years), Bachelor's degree (3 years), Master's or higher (5 years), working situation (unemployed, family assistance, fixed-term contract, open-term contract, employer, independent, student), Primary income (present, absent).

We focus here on the acoustic quality of the vowels and used PRAAT [19] to extract formants values for the first four formants (F1, F2, F3 and F4). First, we used the formants values to perform a principal component analysis (PCA) in R [20]. Thereafter we applied hierarchical clustering on principal components (HCPC) on the result of the PCA. In order to illustrate individual entries of HCPC result, we calculated the mean value of F1, F2, F3, F4 for each factor (each level of the meta-data as mentioned above). Second, we used the measurements of the first and second formants (F1, F2) to generate the vocalic space according to selected factors (see Section 4). Three values are extracted for each vowel: one at the beginning, one in the middle and one at the end of the segment.

Vowel	a	e	i	o	u
occ.	26450	31851	12551	22917	6163

Table 2: Nbr. of lexical vowels used in the study.

3.3. Corpus features

The general characteristics of the data used in the study are summarized in Table 1. Figure 1 shows that interviewees' (micro-workers) speaking time is significantly longer compared to the interviewers (sociologists), suggesting a rather spontaneous character of the exchanges. The presence of overlapping speech as well as a ratio of verbal vs non-verbal speech events (here empty and filled pauses) similar to other Spanish spontaneous corpora previously analyzed [18] suggest that TRIA has the features of a spontaneous corpus (Figure 1). From there we hypothesise that phenomena specific to spontaneous speech may occur such as reductions and/or non-canonical realizations of segments. With respect to the vocalic system, previous investigations dedicated to acoustic quality of the Spanish vowels as function of the speech style underlined universal (i.e. language-independent) trends such as the tendency to centralization of the vocalic system [21]. In addition, studies also pointed out the propensity to schwa-like realizations of Spanish vowels modulated by lexical stress, the unstressed position triggering more centralized realizations as they correspond to shorter realizations compared to stressed vowels [22].

4. PRELIMINARY RESULTS

We focus on the acoustic quality of monophthong vowels regardless of the lexical stress: near 100k occurences are used for the analysis (Table 2). The acoustic quality of the Spanish vowels is examined with respect to socio-demographic meta-data used as factors in the two following procedures. First, mean values of F1, F2, F3 and F4 of the focal vowels /i-a-u/ are used to perform Principal Component Analysis (PCA), followed by HCPC (hierarchical clustering on principal components), in order to have an estimate of the vowels co-variation according to socio-demographic labels (i.e. Gender, Country of Residence, Diploma, Professional situation, +/- Income). Figures 2, 3 and 4 show the co-

variation for each vowel. The two axes represented correspond to the first two dimensions of the PCA models. Note that the dimensions are not directly formants, but the two dimensions that model the best the dataset. Dispersion show that Gender is predictably responsible for clustering and for all vowels (Y-axis). However, they suggest that other variables play a role in the clusters' composition, such as level of education and socio-economic situation in a broad sense (e.g. the main occupation, the reported income). In particular, X-axis separates values corresponding to lower vs. higher level of education in particular for the vowel /a/.

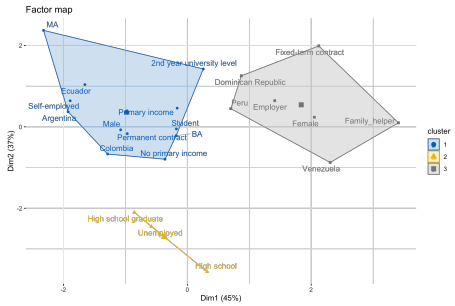


Figure 2: Visualization of categories of each factor based on the PCA map and HCPC : [a].

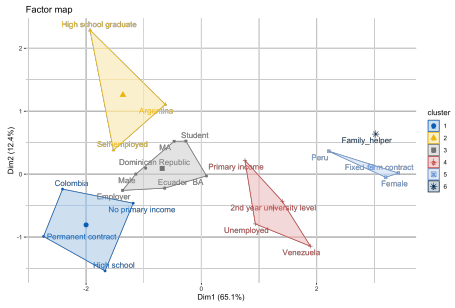


Figure 3: Visualization of categories of each factor based on the PCA map and HCPC : [i].



Figure 4: Visualization of categories of each factor based on the PCA map and HCPC : [u].

Second, we used the more represented variables in the dataset, that is Gender and reported Income, to represent the 5 Spanish vowels in the F1/F2 space. Table 3 groups the speakers according to the source of income and, as a result, the main or complementary character of the micro-work (+/-Income), and underlines the correlation with Gender. Figure 5 suggests the tendency to centralization potentially correlated to selected variables, e.g. male and female speakers for which micro-work is the main source of income produce the more centralized and thus potentially reduced vowels. Such profiles also correlate with lower level of education suggesting diaphasic variation as function of the social status.

Income	Male	Female
+Income	7	5
-Income	7	0

Table 3: Data distribution as function of Gender and reported Income.

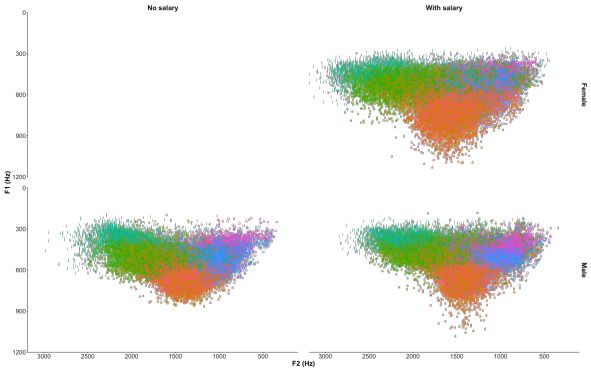


Figure 5: Vocalic dispersion according to Gender and reported Income.

5. CONCLUSION

We proposed a preliminary analysis of the acoustic quality of Spanish vowels from a corpus of interviews made by sociologists with micro-workers from Latin America. Measurements of the first four formants of the focal vowels were used to conduct principal component analysis and hierarchical clustering with respect to the socio-demographic labels. The most represented labels (Gender, Income) were used as factors to compare several vocalic dispersions as a function of F1/F2. The results suggest that educational profile and the socio-economic situation influence in a measurable manner the vocalic variation. This multidisciplinary research is in progress: we will both enrich the corpus with more data and add other parameters of variation to draw an in-depth picture of the micro-workers from Latin America and beyond.

6. ACKNOWLEDGEMENTS

This research was supported by the MSH Paris-Saclay grant "TRIA (Le TRavail de l'Intelligence Artificielle)" awarded to Paola Tubaro.

7. REFERENCES

[1] P. Tubaro and A. A. Casilli, "Micro-work, artificial intelligence and the automotive industry," *Journal of Industrial and Business Economics = Economia e politica industriale*, vol. 46, p. 333–345, Jun. 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02148979

[2] G. Bastin and P. Tubaro, "Le moment big data des sciences sociales," *Revue française de sociologie*, vol. 59, no. 3, pp. 375–394, Sep. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01885416

[3] W. Labov, "The study of language in its social context," in *Volume 1 Basic concepts, theories and problems: alternative approaches*, J. A. Fishman, Ed. Berlin, Boston: De Gruyter Mouton, 1971, pp. 152–216.

[4] W. Labov, I. Rosenfelder, and J. Fruehwald, "One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis," *Language*, vol. 89, pp. 30–65, 2013.

[5] P. Tubaro, "Learners in the loop: hidden human skills in machine intelligence," *Sociologia del Lavoro*, vol. 163, pp. 110–129, 2022. [Online]. Available: https://hal.inria.fr/hal-03787017

[6] J. M. Lipski, *Geographical and Social Varieties of Spanish: An Overview*. John Wiley Sons, Ltd, 2012, ch. 1, pp. 1–26. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118228098.ch1

[7] M. de la Fuente Iglesias and S. Pérez Castillejo, "Spanish mid vowels as sociolinguistic variables in galicia," *Spanish in Context*, vol. 17, no. 3, pp. 464–487, 2020. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/sic.18027.igl

[8] R. Ronquest, *Vowels*, ser. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2018, p. 145–164.

[9] J. Hualde, *The Sounds of Spanish*. Cambridge University Press, 2005. [Online]. Available: https://books.google.be/books?id=vI-wswEACAAJ

[10] P. Delattre, "An acoustic and articulatory study of vowel reduction in four languages." *Iral-international Review of Applied Linguistics in Language Teaching*, vol. 7, pp. 295–326, 1969.

[11] T. Navarro Tomás, *Manual de pronunciación española /*, 13th ed. Madrid:: Monteverde,, 1967.

[12] J. I. Hualde, M. Simonet, and F. Torreira, "Postlexical contraction of nonhigh vowels in spanish," *Lingua*, vol. 118, no. 12, pp. 1906–1925, 2008, studies on the Phonetics and Phonology of Glides. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0024384107001787

[13] M. F. Dabkowski, "Acoustic properties of mexico city spanish vowel weakening," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3578–3578, 2017. [Online]. Available: https://doi.org/10.1121/1.4987616

[14] S. M. Alvord and B. Rogers, "Miami-cuban spanish vowels in contact," *Sociolinguistic Studies*, vol. 8, no. 1, p. 139–170, Jul. 2014. [Online]. Available: https://journal.equinoxpub.com/SS/article/view/7300

[15] R. Ronquest, "Stylistic variation in heritage spanish vowel production," *Heritage Language Journal*, vol. 13, no. 2, pp. 275 – 298, 2016. [Online]. Available: https://brill.com/view/journals/hlj/13/2/article-p275_9.xml

[16] J. Hualde and I. Chitoran, "Surface sound and underlying structure: The phonetics-phonology interface in romance languages," in *Manual of grammatical interfaces in Romance*, S. Fisher and C. Gabriel, Eds. De Moution Gruyter, 2016, vol. 10, pp. 23–44.

[17] M. Nadeu, "Stress- and speech rate-induced vowel quality variation in catalan and spanish," *Journal of Phonetics*, vol. 46, pp. 1–22, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447014000540

[18] I. Vasilescu, N. Hernandez, B. Vieru, and L. Lamel, "Exploring temporal reduction in dialectal spanish: A large-scale study of lenition of voiced stops and coda-s," in *Proc. Interspeech 2018*, 2018, pp. 2728–2732. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1256

[19] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[21] B. Harmegnies and D. Poch-Olivé, "A study of style-induced vowel variability: Laboratory versus spontaneous speech in spanish," *Speech Communication*, vol. 11, no. 4, pp. 429–437, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939290048C

[22] F. Santiago and P. Mairano, "The role of lexical stress on vowel space and duration in two varieties of Spanish," Proceedings of the 9th International Conference on Speech Prosody, pp. 453–457, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01717914

[1] https://www.exteriores.gob.es/en/PoliticaExterior/Paginas/ElEspanolEnElMundo.aspx

[2] The study protocol was approved by the Data Protection Officer of CNRS under GDPR

REDEFINING THE VARIABLE CONTEXT(S) FOR ENGLISH (T/D)-DELETION

James A. Walker

La Trobe University
j.walker2@latrobe.edu.au

ABSTRACT

The variable deletion of /t,d/ in word-final consonant clusters (T/D) is a ubiquitous feature of spoken English that is conditioned by a number of social and linguistic factors. Phoneticians dispute characterising the process as deletion and prefer gestural overlap or masking with a following consonant. This concern raises the question of whether (T/D) is a single variable or multiple variables. Using spontaneous English data collected in two locales, this paper examines the linguistic conditioning of (T/D) with each following context (consonant, vowel, pause). Multivariate analysis reveals some commonalities across the three contexts, suggesting an underlying shared basis for the variable. Differences across contexts can be explained by the nature of following consonants and effects of frequent lexical items in some contexts.

.

Keywords: English, deletion, consonants, variation.

1. INTRODUCTION

The variable deletion of underlying /t/ and /d/ in word-final consonant clusters, (t/d-deletion) or (T/D), is considered a 'showcase' variable of variationist sociolinguistics [1]. This variable occurs in all varieties of spoken English, has been studied in a number of locales and has been shown to be conditioned by an array of social and linguistic factors [1,2,3,4,5,6,7,8,9].

Phoneticians have objected to using the term 'deletion' for this variable, as it implies a complete absence of articulatory gesture [10]. Instead, they suggest that it is better characterised as an overlap or masking of one articulatory gesture by another, especially when the following sound is a consonant. Even without a following consonant, gestural traces may remain, although they are not easy to detect without laboratory methods [11,12].

Until now, (T/D) has been considered as a unitary variable process, but the abovementioned concerns raised by phoneticians suggest that we consider the possibility that it represents different articulatory processes with a similar perceptual output. This paper takes this possibility into consideration by re-analysing data collected in two widely separated cities (Toronto and Melbourne) to test for the effect of locale, dividing the data into three contexts: before consonants, before vowels and before a pause.

2. DATA AND METHODS

The data on which this study is based are taken from two corpora of sociolinguistic interviews conducted in different cities: Toronto, Canada, and Melbourne, Australia. From 27 recordings of conversational interaction with speakers of British/Irish background (15 in Toronto, 12 in Melbourne, evenly balanced for sex), 2,494 tokens of (T/D) were extracted and coded impressionistically as deleted or not deleted (on the basis of any audible occlusion) by two coders in each location.

Each token was additionally coded for social factors (individual speaker, location and speaker sex) and for a series of linguistic factors shown to be relevant in previous studies:

- whether the /t,d/ shared voicing with the preceding segment (*sand*, *kept*) or not (*went*) [6];
- whether the /t,d/ shared its place of articulation (coronal) and manner of articulation (non-sonorant) with the preceding sound [6];
- whether the following sound was a consonant, a vowel or a pause;
- if the following sound was a consonant, whether the /t,d/ shared its voicing, place of articulation and manner of articulation with that consonant [6];
- whether the /t,d/ was part of the root morpheme (monomorphemic: *mist*, *past*), indicated past tense (*missed*, *passed*), indicated past tense along with a root vowel change (semiweak: *kept*, *left*, told) or whether it was ambiguous as to its morphological function (*went*, *built*) [2,3];
- the individual lexical item in which the /t,d/ occurred; and,
- the frequency of that lexical item in each dataset [7,9].

3. RESULTS

A mixed-effects logistic regression analysis of the effects of the above factors on deletion (with speaker

and lexical item as random effects) showed that the effect of the following segment outranked those of all other linguistic and social effects, which provided further justification for conducting separate analyses for each following context.

Table 1 shows the results of three mixed-effects logistic regression analyses for each of the following contexts (logodds values were calculated using sum contrasts, with values centred on 0).

	Following Context					
	Consonant		Vowel		Pause	
	Logodds	N	Logodds	N	logodds	N
Location						
Melbourne	.403	424	---		---	
Toronto	-.403	633	---		---	
Voicing						
Same	-.346	810	-.95	822	-.342	235
Different	.346	247	.95	259	.342	121
Preceding Place						
+coronal	.509	881	.902	820	---	
-coronal	-.509	176	-.902	261	---	
Preceding Manner						
+sonorant	---		---		.529	247
-sonorant	---		---		-.529	109
Following Place						
+coronal	-.372	355				
-coronal	.372	702				
Following Manner						
+sonorant	.373	220				
-sonorant	-.373	837				
Morphological Structure						
Monomorph	---		1.211	561	---	
Past	---		-.127	404	---	
Semiweak	---		-.509	34	---	
Ambiguous	---		-.576	82	---	
Frequency						
+1	---		.041		---	

Table 1: Factors selected as significant to (t/d)-deletion by following segment.

These analyses reveal a number of differences in the conditioning of (T/D) across following context. Location is selected as significant only with following consonants, with deletion favoured more in Melbourne than in Toronto. The manner of the preceding segment is selected as significant only for following pause, with preceding sonorants favouring deletion. The manner and place of the following consonant are selected as significant, with non-coronal and sonorant consonants favouring deletion. Morphological status is selected as significant only for following vowels, with monomorphemic forms favouring deletion over all past-tense forms (including semiweak and ambiguous forms). Frequency is selected as significant only for following vowels.

There are some consistent effects across contexts. Preceding place is significant across both following consonants and following vowels, with a consistent favouring effect of coronals, and shared voicing is significant across all contexts, with a favouring effect of different voicing.

Taken together, these results provide some justification for separating the three following contexts. Despite the similarity in the effects of the preceding segment across most or all contexts, variability in deletion is conditioned by different linguistic and social factors. However, these differences do not have ready explanations from a purely phonetic perspective. While the effects of the preceding and following segments may reflect gestural overlap (or differentiation), the morphological and frequency effects require further exploration.

Although the mixed-effects regression models take into consideration the overall rate of deletion of individual lexical items, the underlying assumption is that all lexical items have an equal opportunity of occurring in each context. Figures 1-3, which plot the occurrence of lexical items by their frequency in the dataset and their overall rate of deletion, show that the lexical composition of each context is different.

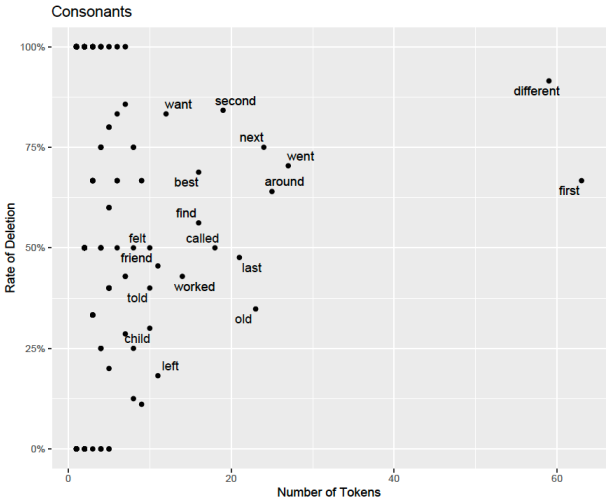


Figure 1: Distribution of lexical items by frequency and rate of deletion before consonants.

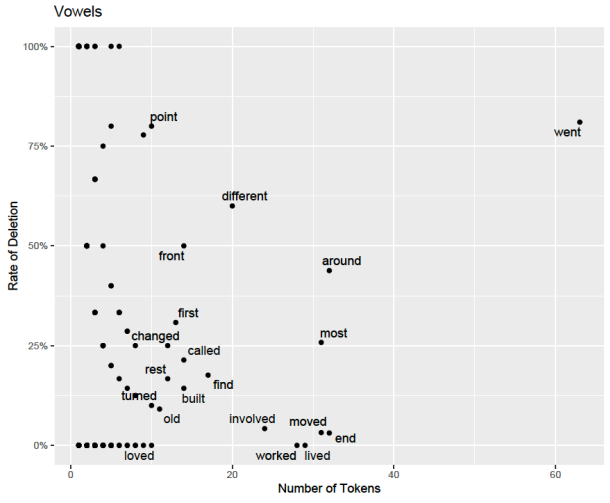


Figure 2: Distribution of lexical items by frequency and rate of deletion before vowels.

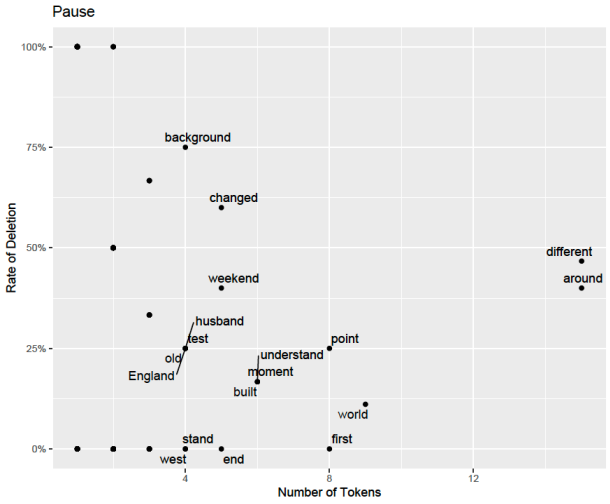


Figure 3: Distribution of lexical items by frequency and rate of deletion before pause.

In particular, the most frequent lexical items for following consonant and pause are *different* and *first* (consonant) or *around* (pause), which are monomorphemic forms, whereas the most frequent lexical item before vowels is *went*, an ambiguous form. The frequency of each item within each context is not correlated with the rate of deletion: while *went* (before a vowel) shows a relatively high rate of deletion, the rates for *different* differ depending on whether the following context is a consonant or pause.

The interaction between following context and morphological structure becomes even more apparent when the rate of deletion is plotted separately for each location and combination of linguistic factors, as shown in Figure 4.

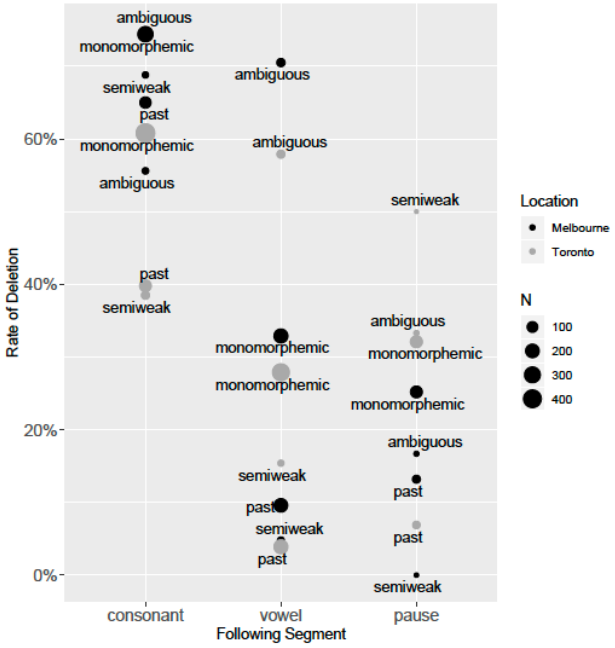


Figure 4: Rate of deletion by location, following context and morphological structure.

The results shown in Figure 4 are largely consistent for location (with higher rates of deletion in Melbourne) and for some aspects of morphological structure (with monomorphemic forms showing higher rates of deletion). However, the other morphological categories are not consistent across location or following context. Ambiguous forms before a consonant or before a vowel show high rates of deletion but much lower rates of deletion with following pause. Semiweak forms show completely different effects by location and following context: higher deletion before a pause in Toronto but before a consonant in Melbourne. However, a re-examination of the distribution of lexical items (Figures 1-3) suggests that these effects have less to do with phonological and morphological considerations and more to do with the effect of the individual lexical items that make up the ambiguous (*went*) and semiweak (*left*, *kept*, *told*) categories.

4. CONCLUSION

Concerns about the phonetic characterisation of (T/D) prompted conducting separate analyses of the variation by following context. These analyses revealed differences in the morphological conditioning of the variation in different locations, but these differences can largely be traced to the effects of individual lexical items, which show different distributions and rates of deletion in each context. Phonological conditioning by the preceding segment is shared across the three contexts, suggesting that deletion is a unitary phonological process in which differences can be attributed to non-phonological considerations. The effects of following consonants on deletion merits further consideration, although the inability of acoustic methods to detect gestural overlap in conversational data would require innovative methods that combine sociolinguistic data collection with more detailed laboratory analysis [8,9].

5. REFERENCES

[1] Guy, G. 1980. Variation in the group and the individual: The case of final stop deletion. In: Labov, W. (ed), *Locating Language in Time and Space*. New York: Academic Press, 1-36.

[2] Guy, G., Boyd, S. 1990. The development of a morphological class. *Language Variation and Change* 2, 1-18.

[3] Guy, G. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3, 1-22.

[4] Patrick, P. 1991. Creoles at the intersection of variable processes: t, d deletion and past-marking in the

Jamaican mesolect. *Language Variation and Change* 3, 171-189.

[5] Baranowski, M., Turton, D. 2020. TD-deletion in British English: New evidence for the long-lost morphological effect. *Language Variation and Change* 32, 1-23.

[6] Guy, G., Boberg, C. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9, 149-164.

[7] Guy, G., Hay, J., Walker, A. 2008. Phonological, lexical, and frequency factors in coronal stop deletion in early New Zealand English. In *Laboratory Phonology 11*. Wellingon, New Zealand.

[8] Tagliamonte, S., Temple, R. 2005. New perspectives on an ol' variable: (t,d) in British English. *Language Variation and Change* 17, 281-302.

[9] Walker, J. 2012. Form, function, and frequency in phonological variation. *Language Variation and Change* 24, 397-415.

[10] Browman, C., Goldstein, L. 1990. Tiers in articulatory phonology, with some implications for casual speech. In: Kingston, J., Beckman, M, (eds), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge: Cambridge University Press, 341-397.

[11] Temple, R. 2017. Phonetic detail and variationist phonology: the case of (t,d). *Dialectologia* 18, 129-155.

[12] Purse, R. 2019. The articulatory reality of coronal stop 'deletion'. *ICPhS19 Proceedings*. Melbourne, Australia.

Aspiration of fortis plosives in multiethnolectal Zurich German

Marie-Anne Morand[1], Camille Watter[2] & Stephan Schmid[2]

[1]University of Fribourg, Switzerland, [2]University of Zurich, Switzerland
marie-anne.morand@unifr.ch, {camille.watter, stephan.schmid}@uzh.ch

ABSTRACT

In the last decades, new ways of speaking have emerged in urban areas of German-speaking Switzerland (so-called multiethnolects), which are distinct from more traditional Swiss German dialects (i.e., Zurich German) on various linguistic levels.

In this study, we analyze aspiration patterns of word-initial fortis plosives in read and spontaneous speech of Zurich German speaking adolescents. Speakers who were perceived as more multiethnolectal show a smaller difference of normalized VOT values between phonologically unaspirated and aspirated plosives in read speech. This difference even disappears completely for bilabial plosives. For alveolar plosives, the difference is bigger for speakers who were perceived as more traditional. These results are mostly confirmed by the spontaneous speech data.

Future research may show whether this VOT-merging of traditionally unaspirated and aspirated fortis plosives is a sign of an emerging sound change and to what extent these findings affect the plosive system of Zurich German in general.

Keywords: Sociophonetics, multiethnolects, Zurich German, sound change, aspirated plosives

1. INTRODUCTION

Since the turn of the millennium, the emergence of so-called multiethnolects [3, 16] has been observed in multiple European countries with relatively high linguistic and cultural diversity [7]. Multiethnolects are used by several minority groups collectively and express a new kind of group identity [3, p. 87]. In this study, we investigate a phonetic feature that might distinguish multiethnolectal Zurich German from a more traditional variety of the same dialect.

The consonant system of Zurich German is characterized by a binary opposition between 'fortis' and 'lenis' plosives [6, 19], which is basically implemented in terms of closure duration (CD) and not in terms of voice onset time (VOT). In principle, both fortis and lenis plosives are unaspirated, but in several lexical items (mainly loanwords) aspirated fortis plosives occur as well [10]. However, it appears that in multiethnolectal Zurich German word-initial fortis plosives are often aspirated in words where they would be unaspirated in traditional Zurich German.

Therefore, the aim of this study is to examine aspiration patterns in multiethnolectal Zurich German more systematically; to this end, read and spontaneous speech data was recorded from adolescents in the city of Zurich.

2. PLOSIVES IN TRADITIONAL ZURICH GERMAN

The phonological contrast between fortis and lenis plosives is quite pervasive in the Zurich German lexicon, as shown by numerous minimal pairs such as /ˈhuːpə/ 'to hoot' vs. /ˈhuːb̥e/ 'bonnet', /ˈlɒtə/ 'lath' vs. /ˈlɒd̥ə/ 'store', /ˈhɒːke/ 'hook' vs. /ˈhɒːg̊e/ 'to fence in' [6, p. 244]. The functional load of this contrast is relatively high for bilabial and alveolar plosives in intervocalic word-internal position, but less so for word-initial plosives and for velar plosives in general. Due to diachronic sound changes, the equivalent of word-initial /k/ in Standard German appears in traditional Zurich German either as a fricative /x̊/ or as an affricate /k͡x/ (e.g., Zurich German /x̊ɒt͡s/ 'cat' and /ˈk͡xɒvi/ 'coffee' vs. Standard German /ˈkat͡sə/ and /ˈkafe/). For this reason, the current study only examines bilabial and alveolar plosives.

Regarding the phonetic implementation of the fortis/lenis contrast in Zurich German, studies have indicated closure duration as the most important acoustic correlate, while VOT (i.e., voicing and/or aspiration) plays no role at all. An additional cue seems to lie in F0 effects in the following vowel [10].

Besides the basic opposition between fortis and lenis plosives, the consonant system of traditional Zurich German displays a third plosive type. For instance, word-initial aspirated plosives do appear in proper names and in loanwords from Standard German and English, such as /ˈpʰeːtər/ 'Peter' and /tʰom/ 'Tom', /pʰɒk͡x/ 'parcel' and /tʰeː/ 'tea', or /ˈpʰɒrti/ 'party' and /t͡ʰiːm/ 'team' [10, p. 231].

Now, it has been observed that among younger speakers there is an increasing number of lexical items (e.g., *Panda* 'panda' or *Puuder* 'powder'), where the word-initial consonant seems to shift from an unaspirated fortis plosive towards an aspirated one. In addition to this lexical diffusion, there is evidence that younger speakers tend to produce fortis plosives with longer VOT values than older speakers, which would point to a regular sound change [20].

Furthermore, there appears to be a certain amount of sociophonetic variation even among younger speakers with regards to the pronunciation of word-initial plosives. More precisely, it has been found that the aspiration of traditionally unaspirated fortis plosives (e.g., [ˈtʰoxtər] 'daughter' instead of [ˈtoxtər]) is typical for adolescents who were perceived as speaking more multiethnolectal Zurich German [15]. It therefore remains to be seen whether a general increase in VOT (in addition to the lexical diffusion type of sound change) constitutes a sociophonetic feature of multiethnolectal Zurich German as well.

3. MULTIETHNOLECTAL ZURICH GERMAN

Recent research has shown that multiethnolectal Zurich German differs from more traditional forms of the dialect with respect to several segmental and suprasegmental variables.

On the segmental level, two sociophonetic features of multiethnolectal Zurich German are related to the fortis/lenis contrast. First, multiethnolectal speakers tend to produce fully voiced lenis plosives which normally do not occur in the traditional dialect [12]. Moreover, multiethnolectal speakers tend to use word-initial fortis fricatives which are excluded by a phonotactic constraint in the traditional dialect [17].

On the suprasegmental level, multiethnolectal Zurich German displays a more 'syllable-timed' speech rhythm, deducible from a lower durational variability of subsequent vocalic intervals [13]. This was observed both in read and in spontaneous speech, which suggests that speakers are not aware of it and unable to control it actively. Therefore, a rather 'syllable-timed' speech rhythm seems to be an 'indicator' rather than a 'marker' [9] as evident by the lack of style shifting between more and less formal settings. It remains to be seen whether this holds for the aspiration of fortis plosives as well or if we detect any differences due to speaking style.

4. DATA AND METHODS

To analyze the aspiration of word-initial fortis plosives in multiethnolectal Zurich German, a corpus consisting of read and spontaneous speech data was compiled. This section presents the speakers, the material, and the data analysis of the two sub-corpora.

4.1 Speakers

For the sub-corpus of read speech, 50 speakers (22 female; mean age = 14.9; *SD* = 0.51) were recorded. Thirteen indicated that they had spoken only Swiss German before they entered kindergarten. The other 37 spoke additional or other languages, which include amongst others (in descending order): Standard German (n = 10), Italian (n = 9), Albanian (n = 5), Portuguese (n = 5), Spanish (n = 5), and English (n = 4).

For the sub-corpus of spontaneous speech, 28 speakers (17 female; mean age = 14.2; *SD* = 0.72) were analyzed. Six adolescents indicated that they had spoken only Swiss German before they entered kindergarten. The other 22 spoke additional or other languages, which include amongst others (in descending order): Albanian (n = 4), Standard German (n = 4), Spanish (n = 3), Bosnian (n = 2), Portuguese (n = 2), and Serbian (n = 2).

In addition, short speech samples of all speakers were rated by peers, who indicated on a 7-point Likert scale how multiethnolectal the speakers sounded; the ratings yielded a continuum rather than two groups of traditional and multiethnolectal speakers [14]. Calculating Spearman correlations between the obtained rating scores and the acoustic measurements of the production data (VOT) will allow (1) to determine to which extent aspiration of traditionally unaspirated plosives constitutes a phonetic feature of multiethnolectal Zurich German, and (2) to test whether there are additional linear relationships between the two variables for the three analyzed categories (see §4.2).

4.2 Material

Read speech data was collected in a Zurich secondary school using *SpeechRecorder* [5]. The sentences were carefully designed and chosen to elicit bilabial and alveolar word-initial fortis plosives. There are 30 sentences in total, containing target words with word-initial plosives: in traditional Zurich German, 10 of these words have a phonologically unaspirated plosive (e.g., (1) and (2)), 10 words have a phonologically aspirated plosive (e.g., (3) and (4)), and in 10 cases the word-initial plosive can be either unaspirated or aspirated (e.g., (5) and (6)). Each of the three categories (-asp, +asp, ±asp) contains 5 bilabial and 5 alveolar fortis plosives.

(1) Am Flughafe bruucht er de **Pass**.
(2) Ich nimm en grosse **Täller**.
(3) Mir händ e langi **Pause**.
(4) Ich hett gern en **Tee**.
(5) Die bäide sind es **Paar**.
(6) Du muesch en **Tescht** mache.

Spontaneous speech was recorded at two different secondary schools in Zurich. Adolescents played a game of 'spot the difference' in pairs using so-called *Diapix* [1, 18]; games lasted approx. 15 minutes.

Recordings were transcribed using so-called 'Dieth orthography' [4] in Praat [2].

4.3 Data Analysis

The sound files for read speech were automatically annotated and segmented using *WebMAUS* [8]. Annotation of VOT (the acoustical correlate of aspiration) and manual correction of segmentation was done in Praat according to the principles stated in [11]. VOT values were normalized for speech rate (nVOT). The analyzed corpus consists of 1357 tokens.

For spontaneous speech, the transcriptions of 48 speakers were searched to find all word-initial bilabial and alveolar fortis plosives characters (i.e., <p t>) followed by a vowel with a Praat script using regular expressions. Twenty-eight adolescents were included in the analysis who produced at least 3 target words with word-initial plosives that are traditionally unaspirated and 3 target words with word-initial plosives that are traditionally aspirated. The data was preprocessed in the steps described above. The analyzed corpus consists of 336 tokens. In our data, there is a bias towards alveolar plosives that are traditionally unaspirated.

5. RESULTS

5.1 Overall aspiration in read speech

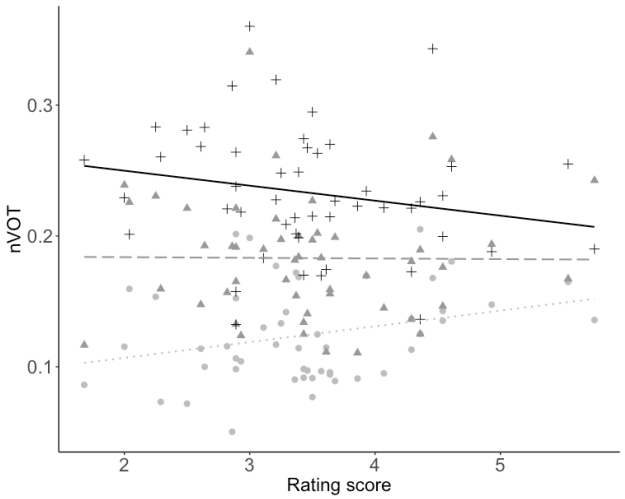


Figure 1: Normalized VOT (read speech) as a function of rating score (-asp: light grey points, dotted; +asp: black crosses, solid; ±asp: dark grey triangles, dashed).

As shown in Figure 1, the speakers' rating score correlates in different ways with VOT for phonologically aspirated and unaspirated fortis plosives. Adolescents who were rated as speaking rather multiethnolectal Zurich German (i.e., higher rating scores) exhibit lower normalized VOT values for word-initial fortis plosives that are traditionally aspirated than adolescents who were rated as speaking rather traditional Zurich German (i.e., lower rating scores); however, this trend does not reach a conventional significance level (r_s = -.246, p = .084).

VOT values for word-initial plosives that can be both unaspirated and aspirated by speakers of traditional Zurich German do not significantly correlate with rating score (r_s = -.090, p = .536). Word-initial plosives that are traditionally unaspirated (e.g., in words such as *Pass* (1) or *Täller* (2)) tend to be aspirated more by adolescents who were rated as speaking rather multiethnolectal Zurich German than by rather traditional Zurich German speaking adolescents (r_s = .219, p = .127).

Finally, the difference between traditionally aspirated and traditionally unaspirated fortis plosives becomes significantly smaller as the rating score increases (r_s = -.291, p = .040).

5.2. Aspiration in bilabial plosives (read speech)

Traditionally unaspirated and traditionally aspirated bilabial fortis plosives are completely merged regarding normalized VOT values for adolescents who were rated as speaking rather multiethnolectal Zurich German, as is shown by the crossing dotted and solid lines in Figure 2. There is a significant correlation between rating score and VOT values for traditionally aspirated bilabial plosives (r_s = -.291, p = .040): Adolescents who were rated as speaking rather multiethnolectal Zurich German aspirate them less than adolescents who were rated as speaking rather traditional Zurich German.

VOT values for traditionally unaspirated bilabial plosives (r_s = .074, p = .611) and for bilabial plosives that are either unaspirated or aspirated by speakers of traditional Zurich German (r_s = -.015, p = .917) do not significantly correlate with rating score.

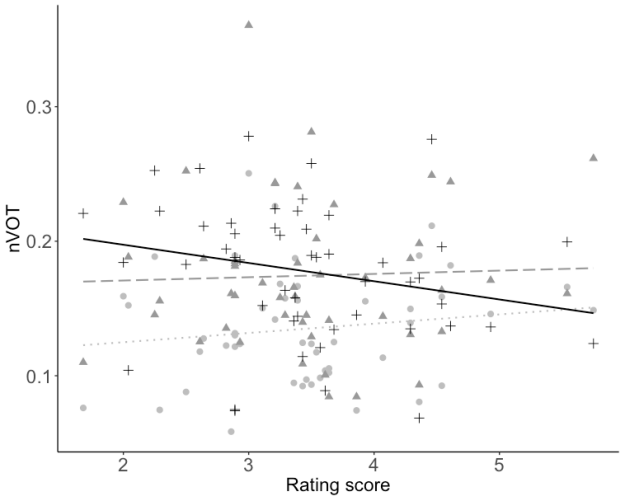


Figure 2: Normalized VOT for bilabial plosives (read speech) as a function of rating score (-asp: light grey points, dotted; +asp: black crosses, solid; ±asp: dark grey triangles, dashed).

5.3. Aspiration in alveolar plosives (read speech)

Traditionally unaspirated and traditionally aspirated alveolar fortis plosives are still very different regarding normalized VOT values for adolescents who were rated as speaking rather multiethnolectal Zurich German (see Figure 3). There is a significant correlation between rating score and VOT values for traditionally unaspirated alveolar plosives (r_s = .322, p = .023), which are more aspirated by multiethnolectal speakers. However, VOT values for traditionally aspirated alveolar plosives (r_s = -.119, p = .409) and VOT values for alveolar plosives that can be unaspirated or aspirated by speakers of traditional Zurich German (r_s = -.119, p = .411) do not correlate with rating score.

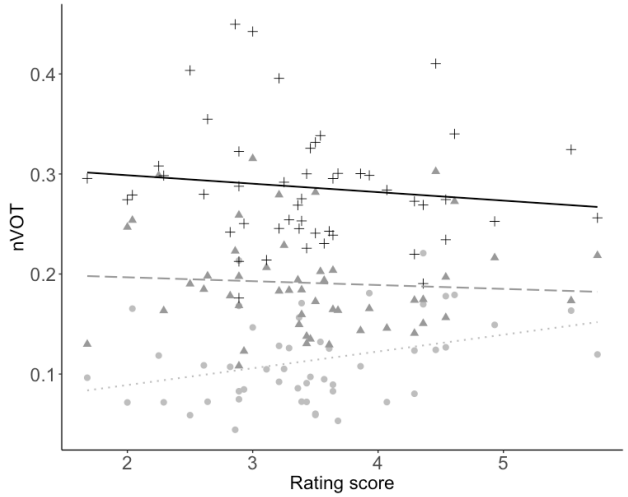


Figure 3: Normalized VOT for alveolar plosives (read speech) as a function of rating score (-asp: light grey points, dotted; +asp: black crosses, solid; ±asp: dark grey triangles, dashed).

5.4. Overall aspiration in spontaneous speech

In spontaneous speech, the difference between VOT values for traditionally unaspirated and traditionally aspirated fortis plosives merges around rating scores of a little bit more than 4 (r_s = -.541, p = .003); afterwards, VOT values for traditionally unaspirated plosives are even higher than VOT values for traditionally aspirated plosives.

No correlation between VOT values for traditionally unaspirated plosives and rating score was found (r_s = .211 p = .282). However, there is a significant negative correlation between VOT values for traditionally aspirated plosives and rating score (r_s = -.486, p = .009). Thus, adolescents who were rated as speaking rather multiethnolectal Zurich German aspirate these plosives less than adolescents who were rated as speaking rather traditional Zurich German.

6. DISCUSSION

In read speech, adolescents who were rated as speaking rather multiethnolectal Zurich German tend to aspirate traditionally unaspirated fortis plosives in word-initial position more than adolescents who were rated as speaking rather traditional Zurich German; the opposite holds for traditionally aspirated fortis plosives. However, both correlations fail to reach conventional levels of statistical significance. Moreover, there is a significant negative correlation between rating score and the difference of the two phonological categories (-asp vs. +asp). Although there is much interspeaker variation for VOT values for plosives that can be both unaspirated and aspirated by speakers of traditional Zurich German, VOT values do not correlate with rating score.

The relation between rating score and VOT values seems to be influenced by place of articulation. While traditionally aspirated bilabial plosives are less aspirated by speakers of multiethnolectal Zurich German than by speakers of traditional Zurich German, traditionally unaspirated alveolar plosives are aspirated more by the former.

In spontaneous speech, multiethnolectal Zurich German speaking adolescents show lower VOT values in traditionally aspirated fortis plosives in word-initial position than adolescents who were rated as speaking rather traditional Zurich German. Therefore, the results of our study suggest that multiethnolectal Zurich German is evolving towards a merger between the two categories of (unaspirated) fortis and aspirated plosives.

Indeed, there is additional evidence for a sound change in progress regarding (unaspirated) fortis and aspirated plosives in Zurich German. A recent sociophonetic study focusing on generational differences revealed that younger speakers produce fortis plosives with longer VOT values than older speakers [20]. Future research may show if multiethnolectal adolescent speakers represent the most advanced stage in this sound change in progress and how a potential merger of (unaspirated) fortis and aspirated plosives will affect the entire plosive system of Zurich German (including lenis plosives).

7. ACKNOWLEDGMENTS

This research was supported by the Swiss National Science foundation (Grant Nos. 165798 and 200616). We would like to thank all the speakers, raters, and their teachers for the participation in this study. Furthermore, we are grateful to Zora Naef who helped with recording read speech as well as with controlling and correcting segment boundaries of the audio files.

8. REFERENCES

[1] Baker, R., Hazan, A. 2011. DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3), 761–770.

[2] Boersma, P. Weenink, D. 2018. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.40, http://www.praat.org/

[3] Clyne, M. 2000. Lingua Franca and ethnolects in Europe and beyond. *Sociolinguistica* 14, 83–89.

[4] Dieth, E. 1986. *Schwyzertütschi Dialäktschrift: Dieth-Schreibung* (2nd rev. ed.). Sauerländer.

[5] Draxler, C., Jänsch, K. 2004. SpeechRecorder: A universal platform independent multichannel audio recording software. *Proc. 4th LREC* Lisbon, 559–562.

[6] Fleischer, J., Schmid, S. 2006. Zurich German. *Journal of the International Phonetic Association* 36, 243–253.

[7] Kerswill, P., Wiese, H. (eds.) 2022. *Urban Contact Dialects and Language Change: Insights from the Global North and South.* Routledge.

[8] Kisler, T., Reichel, U., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

[9] Labov, W. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

[10] Ladd, D.R., Schmid, S. 2018. Obstruents voicing effects on F0, but without voicing: Phonetics correlates of Swiss German lenis, fortis, and aspirated stops. *Journal of Phonetics* 71, 229–248.

[11] Machač, P., Skarnitzl, R. 2009. *Principles of Phonetic Segmentation.* Epocha Pubishing House.

[12] Morand, M.-A., Bruno, M., Julmi, N., Schwab, S., Schmid, S. 2019. The voicing of lenis plosives in Zurich German: A sociophonetic marker of (multi-)ethnolectal speech. *Proc. 19th ICPhS* Melbourne, 1600–1604.

[13] Morand, M.-A., Bruno, M., Schwab, S., Schmid, S. 2022. Syllable rate and speech rhythm in multi-ethnolectal Zurich German: A comparison of speaking styles. *Proc. 11th Speech Prosody* Lisbon, 336–341.

[14] Morand, M.-A., Schwab, S., Schmid, S. 2020. The perception of multiethnolectal Zurich German: A continuum rather than clear-cut categories. *Loquens* 7(2), e072.

[15] Morand, M.-A., Schwab, S., Schmid, S. 2021. Standarddeutsche Interferenzen im Dialektwortschatz Schweizer Jugendlicher: Lexikalische und lautliche Entlehnungen. *Bulletin suisse de linguistique appliquée* 113, 137–154.

[16] Quist, P. 2000. Ny københavnsk 'multietnolekt': Om sprogbrug blandt unge i sprogligt og kulturelt heterogene miljøer. *Danske Talesprog* 1, 143–211.

[17] Schmid, S., Morand, M.-A., Schwab, S. In press. Multiethnolektales Zürichdeutsch: Dialekttransformation und soziolinguistische Wahrnehmung. *Journal LIPP.*

[18] van Engen, K.J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M., Bradlow, A.R. 2010. The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53, 510–540.

[19] Winteler, J. 1876. *Die Kerenzer Mundart des Kantons Glarus in ihren Grundzügen dargestellt*. Winter.

[20] Zebe, F., Watter, C., Schmid, S. In revision. Increasing aspiration of Swiss German plosives: A sound change in progress? *Proc. 20th ICPhS* Prague.

ACOUSTIC CORRELATES OF SELF-ASCRIBED MASCULINITY/FEMININITY IN AN AMERICAN ENGLISH SAMPLE

Chen Zhou[1], Jason Bishop[1,2]

[1]The Graduate Center, CUNY; [2] College of Staten Island, CUNY
czhou1@gradcenter.cuny.edu, jason.bishop@gc.cuny.edu

ABSTRACT

Recently it has been argued that some aspects of the signal usually associated with a speaker's biological sex may also vary systematically in relation to measures of the speaker's masculinity/femininity, suggesting gender identity, gradiently defined, may be a source of inter-speaker variation in speech. The present study attempted to extend the findings from a recent production study reported in [1] (which tested a German sample) to English. Read speech from a sample of 74 American English speakers (38 women, 36 men) was collected and analyzed with respect to two variables known to vary, in part, as a function of speaker sex and/or gender: vowel space area and the spectral center of gravity for sibilant /s/. These acoustic phonetic variables were then considered in relation to gradient measures of the speakers' self-ascribed femininity. Results were partially and only weakly consistent with the previously-reported German findings.

Keywords: masculinity/femininity, socio-phonetics, gender identity, individual variability in speech.

1. INTRODUCTION

A number of acoustic-phonetic factors are known to distinguish, to varying degrees, the speech of men and women. Some of the most well-studied aspects of the signal believed to distinguish the sexes, biologically defined, include measures of spoken fundamental frequency (f0) – especially mean f0 ([2], [3]) and f0 range ([4]) – as well as differences related to voice quality ([5]), acoustic vowel space size ([6], [7]), and the spectral characteristics of sibilant fricatives ([8], [9]). For example, in comparison to men's speech, women have generally been shown to produce higher mean f0, a wider f0 range, a larger acoustic vowel space area and higher spectral mean (i.e., center of gravity), for sibilant /s/. Importantly, however, such phonetic differences have been described as reflecting not only biophysical differences between the male and female sexes, but also learned differences that distinguish men and woman as gender groups, where 'gender' is understood as a cultural/socio-cognitive construct (e.g., [10], [11]). In other words, one's speech patterns encode information about both sex (biological dimorphisms distinguishing males and females) and gender (learned/abstract social divisions between men and woman).

However, both in the language sciences and in many societies more generally, there is growing recognition that understanding the role that sex and gender play in speech will likely require recognizing more than just these binary, categorical distinctions. On the one hand, there is well documented within-sex variation tied systematically to distinctions in sexual orientation, such as straight versus non-straight speakers ([12], [13]). However, it is also known that gender characterizations not directly related to sexual orientation, such as a coarse distinction between cisgender and gender expansive speakers, have correlates in the speech signal as well ([14], [15]). Most pertinent to our purposes here, there is reason to believe that gender – again a socio-cognitive construct – can be understood as not only non-binary, but perhaps not even categorical. If this is true, then we would expect to see gradient measures of gender or 'gender identity' predict properties of speech outputs in correspondingly gradient ways.

In fact, there is some evidence for this kind of intra-group variability as well. For example, two studies testing German speakers ([16], [17]) found that some acoustic cues associated with sex-specific differences varied as a function of speakers' scores on continuous self-report scales of masculine/feminine social/personality characteristics, which have been claimed to capture fundamental aspects of a person's sense of gender identity ([18]). Moreover, this was true within groups of lesbian women (in terms of median f0) as well as within groups of straight and gay men (in terms of mean second formants for /a:/, /i:/ and /u:/). Such findings suggest the possibility that previous work may have overestimated the role of sex in determining some phonetic properties of speech. Instead, a significant amount of variation we see might actually be due to individual differences along some socio-cognitive dimension(s) related to gender identity.

Few studies have explored the effect of such individual variability in gender identity on sex- and gender-specific speech patterns. One exception was [1]'s investigation of German speakers. In that study, gender identity was operationalized in terms of self-reported ratings on a scale of masculinity/femininity and the authors examined the relationship between

these ratings and several acoustic parameters known to encode sex and gender. Two questionnaires were used to elicit speakers' femininity ratings: the F+ scale of the German version of the Extended Personal Attributes Questionnaire (GEPAQ-F+) ([19]) and the Traditional Masculinity-Femininity scale (TMF) ([20]). The authors found that both vowel space area (VSA) and mean f0 varied systematically in relation to self-ascribed femininity ratings, though only for male speakers. More specifically, the more feminine/less masculine that men rated themselves to be, the larger their VSA and the higher their f0 were (i.e., in the direction of more female-like values). The other exception was [21]'s study on the moderating effect of social meaning on pitch range differences of female and male Japanese-British English sequential bilinguals. Both female and male speakers' English pitch level (measured with mean f0) were found to vary as a function of individual gender identity, also operationalized as scores on continuous scales of masculinity and femininity (in their case, using the short version of the Bem Sex Role Inventory (BSRI-short). The findings from both studies invite further examination of subtle-but-systematic acoustic cues to constructs such as gender identity, which can arguably be usefully operationalized in terms of self-ascribed masculinity/femininity.

To this end, the present study attempted to partially replicate the basic findings reported by [1], who found gradient measures of femininity to be systematically related to subtle differences in speech usually associated with sex. Our goal, however, was to extend the discussion to another language, namely American English, and to use a larger sample size, as [1]'s data set was rather modest for evaluating interindividual variation tied to what is, again, best regarded as a socio-cognitive variable. Moreover, we also sought to control for prosodic position and segmental context in the speech samples, which led us to utilize read speech rather than spontaneous or unscripted speech (as in [1]'s study). In line with the investigation in [1], the current study treated gender identity as a continuous variable, based on speakers' self-reported scores on established scales of femininity. With these basic considerations in mind, the present study asked two research questions:

1.) Do speakers with higher self-ratings of femininity produce more female-like values for phonetic variables like VSA or spectral center of gravity (CoG) for /s/? These two variables have been shown previously to vary significantly based on speaker sex (e.g., [22] for VSA; [8], [9] for CoG).

2.) If so, are these gender-dependent patterns to any extent also sex-specific?

Based primarily on the findings in [1] and [21], we made the following predictions. First, we predicted that the more feminine a speaker rates themselves, the more female-like their productions of the acoustic variables will be; in the present case, that would mean larger VSA and higher spectral CoG for /s/. Second, we predicted that the patterns may be specific to a particular sex group. That is, the systematic relationship between the acoustic variables examined and the self-reported femininity scores of the speakers, if any such relationships exist, might be limited to just one sex group, as this was what was found previously by [1] for VSA and mean f0, which were affected by femininity scores for male speakers only. (See also [21], who found gender ratings to have the opposite relationship to mean f0 for male and female speakers).

2. EXPERIMENT

2.1. Methods

2.1.1. Participants

A total of 74 monolingual American English speakers (38 female speakers, 36 male speakers, mean age 21.8 years, SD 3.7) participated in the current study, a previous analysis of which was reported in [23]. In order to limit dialectal variation in the speech sample, all participants recruited in the present study originated and lived all or most of their lives in New York City, and more specifically, Staten Island. None of the participants reported any history of a speech or hearing disorder. When asked to choose options about their self-identified (categorical and binary) gender, only one participant (a biological female who identified as a man in gender) chose an option that was different from their biological sex.

2.1.2 Stimuli

Two sets of English CVC wordforms, consisting of sixteen target words, were used in the study. One set contained eight words eliciting the four vowels /i/, /ɛ/, /u/, and /ɑ/ in two consonantal contexts: h_d and b/p_t. (While the /æ/ vowel is in fact more peripheral than /ɛ/ in American English, it was not used to construct vowel spaces here due to dialectal variation associated with 'a-tensing' in NYC English; [30]). The other set of CVCs consisted of four words intended to elicit /s/ in onset and coda positions. Target words were to be presented in carrier sentences in which the target word was contrasted with another CVC (filler) word in order to induce

contrastive/emphatic prosodic prominence. The phrase positions of target words and their relative order with the filler words were also varied to balance the prosodic effects associated with sentence position, yielding four positions for the target words in four carrier sentences:

(1, 2) *Target/filler* is the first word, and *target/filler* is the second word. (Phrase-initial)

(3, 4) The first word is *target/filler*, and the second word is *target/filler*. (Phrase-final)

Stimuli were pseudo-randomized and presented in two blocks whose order was counterbalanced across participants; one block contained phrase-initial targets and fillers and the other block contained phrase-final targets and fillers. One production of each item in each sentence was collected, resulting in 96 samples per speaker and thus 7,104 tokens for acoustic analysis in total.

2.1.3. Measurement of gender identity

Similar to the approach adopted in [1], gender identity was operationalized as scores on two continuous measures of self-reported femininity, also used in [1]: the Femininity scale of the Extended Personal Attributes Questionnaire (EPAQ-F+) ([25]) and the Traditional Masculinity-Femininity Scale (TMF) ([20]), both of which are intended to assess feminine traits and behaviours in English-speaking adults. For each of the questionnaires, the higher the score, the higher the self-rated femininity.

2.1.4. Procedure

All the recordings were made in a sound-attenuated booth using a head-mounted Shure SM-10A microphone and a 44.1kHz sampling rate. Participants were seated in the sound booth in front of a computer screen, on which the target and filler words were presented in their carrier sentences, one carrier sentence at a time on the screen. To create an interactive speaking context for the task, the participants were told to read the carrier sentences for the benefit of a confederate, who sat with the participant in the sound booth (but behind a curtain) so as to write down the target and filler words in each sentence produced by the participant. The confederate would say "got it" to give verbal confirmation after each sentence was produced before moving on to the next. Recordings were saved as *wav* files for later analysis in Praat (version 6.1.15, [26]).

2.2. Acoustic and statistical analyses

As described above, the two acoustic measures in the present study included: (1) the spectral CoG of sibilant /s/, and (2) VSA, the geometric area of the quadrilateral enclosed by the points in an F1 × F2 space whose coordinates were defined by the values (transformed to a Bark scale) of the first two formants for each of the four vowels (/i/, /ɛ/, /u/, and /ɑ/). For each of the acoustic measures, Linear Mixed-Effects (LME) models were used to identify any effects related to self-rated femininity scores in predicting CoG for /s/ and VSA in speakers' productions.

3. RESULTS

3.1. Self-rated femininity ratings

We first consider the distributions of scores on the two gender-related measures, their relation to each other and how well they each distinguish the two sex groups. To that end, the scatterplot in Fig1 shows the relationship between TMF and EPAQ-F+ for male and female speakers. Simple Pearson correlations showed a marginally significant positive relationship between TMF and EPAQ-F+, but for female speakers only. The box and whisker plots in Fig1 illustrate the amount of overlap in EPAQ-F+ and TMF scores exhibited by the two sex groups; Welch Two sample t-tests for the inter-sex differences on EPAQ-F+ and TMF scores showed a significant sex group difference for TMF scores ($t = 15.716$, $df = 65.751$, $p < 0.0001$), but not for EPAQ-F+ scores ($t = 1.6404$, $df = 68.73$, $p = 0.1055$). Therefore, only the TMF gender scale clearly differentiated the two sex groups in this sample.

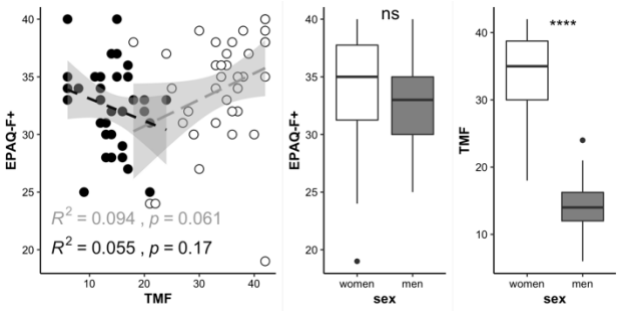


Figure 1. Plots for EPAQ-F+ and TMF scores (male speakers: black; female speakers: white). Left plot: the relationship between the two measures for the two sexes. Right plots: boxplots show the distribution of EPAQ-F+ and TMF scores for each sex.

3.2. Effects of femininity ratings on CoG

Fig2 (left two plots) shows the effect of each measure of femininity on CoG values for /s/. For TMF, though the relationship was nominally in the predicted

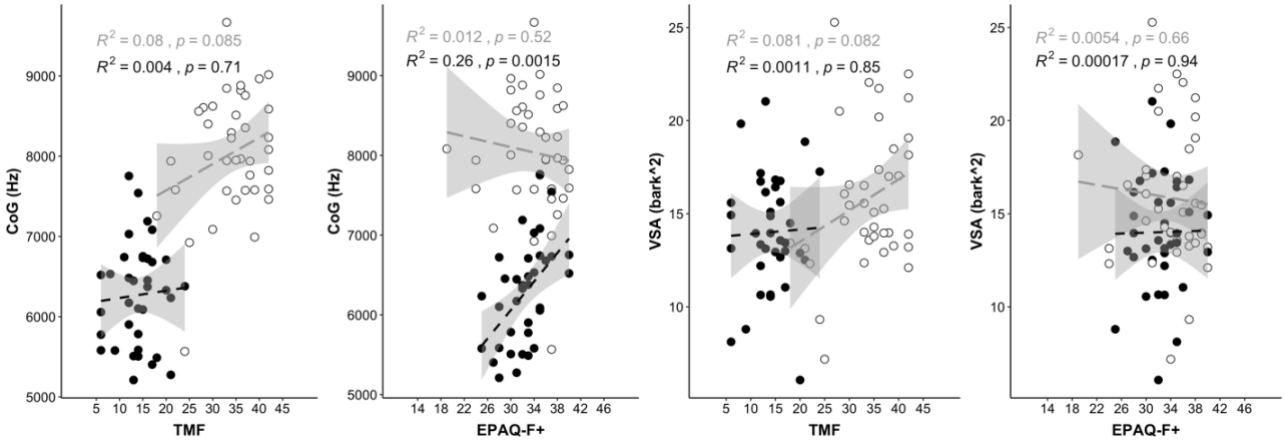


Figure 2. Scatter plots that show the relationships between the femininity scores and the acoustic variables. (Male speakers: black; female speakers: white/grey). Left two plots: the femininity scores in relation to CoG; right two plots: the femininity scores in relation to VSA.

direction for both sexes, a simple Pearson correlation suggested this relationship was weak – only marginally significant, and only in the case of female speakers. EPAQ-F+, on the other hand, was found to have a significant correlation, and in the predicted direction, with CoG values – but for male speakers only. LME analysis further confirmed these patterns; models showed a significant interaction between EPAQ-F+ and sex (χ^2 (1) = 12.01, p < 0.001), as well as a significant main effect for TMF (χ^2 (1) = 7.20, p = 0.007). Separate sex-specific models showed that TMF was only marginally significant in predicting female speakers' CoG values (χ^2 (1) = 3.16, p = 0.075) and that EPAQ-F+ significantly predicts male speakers' CoG values (χ^2 (1) = 12.49, p < 0.001).

3.3. Effects of femininity ratings on VSA

Fig2 (right two plots) illustrate the relationships between the measures of femininity and VSA. There was a weak positive correlation between TMF and VSA, though for female speakers only. This relationship was shown to hold in a more rigorous LME modelling; TMF was found to be a marginally significant predictor of VSA in a sex-specific model of female speakers (intercept = 0.17, SE = 0.09, t = 1.805, p = 0.08), but not for male speakers' VSA (intercept = 0.03, SE = 0.12, t = 0.25, p = 0.80).

4. DISCUSSION/CONCLUSION

The present study examined the relationship between gradiently defined gender identity, operationalized as scores on two continuous scales of self-reported femininity, and two acoustic phonetic variables usually associated with differences in speaker sex. Statistical analyses indicated that, at least for male speakers, higher scores on EPAQ-F+ (indicating higher levels of femininity) were significantly associated with productions of /s/ with higher, more female-like CoG values. This is, in fact, the predicted relationship if we assume that (a) gender can be defined gradiently between two extremes, and (b) speakers encode this subtle information about gender in their speech. Interestingly – and as in [1] – this relationship only held for male speakers, although in the present case, for a different acoustic variable, and only when EPAQ-F+ was the measure of femininity.

However, the findings related to EPAQ-F+ in the present study suggest that that measure may not be tapping into the same gender-related construct as TMF, at least not for our American English sample. That is, unlike GEPAQ-F+ scores in the German sample in [1], EPAQ-F+ scores in the present study did not clearly differentiate male and female speakers, and were also not well correlated with TMF within-sex (see Fig1, above). One possibility, though speculative, is suggested by [20], who note that some well-established femininity/masculinity scales tap into socio-cognitive constructs more closely related to instrumentality or expressivity of gender traits rather than gender identity per se. Further research is therefore likely needed to evaluate this measure. For the time being, we think that the results related to TMF in the present study are likely the most reliable, and therefore that the role of gradiently-defined gender identity was detectable in the speech of this American English sample (for both acoustic variables) – but only marginally so, and only for female speakers.

7. REFERENCES

[1] M. Weirich and A. P. Simpson, "Gender identity is indexed and perceived in speech," *PLOS ONE*, vol. 13, no. 12, p. e0209226, Dec. 2018, doi: 10.1371/journal.pone.0209226.

[2] K. N. Stevens, *Acoustic phonetics*, vol. 30. MIT press, 2000.

[3] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1699–1707, Apr. 1989, doi: 10.1121/1.397959.

[4] J. Haan and V. J. van Heuven, "Male vs. female pitch range in Dutch questions," presented at the Proceedings of the 14th International Congress of Phonetic Sciences, NN, 1999, pp. 1581–1584.

[5] C. G. Henton and R. A. W. Bladon, "Breathiness in normal female speech: Inefficiency versus desirability," *Language & Communication*, vol. 5, no. 3, pp. 221–227, Jan. 1985, doi: 10.1016/0271-5309(85)90012-6.

[6] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, May 1995, doi: 10.1121/1.411872.

[7] M. Weirich and A. P. Simpson, "Differences in acoustic vowel space and the perception of speech tempo," *Journal of Phonetics*, vol. 43, pp. 1–10, 2014.

[8] K. Heffernan, "Evidence from HNR that/s/is a social marker of gender," *Toronto Working Papers in Linguistics*, vol. 23, 2004.

[9] A. P. Simpson and M. Weirich, "Gender-specific differences in sibilant contrast realizations in English and German.," in *ICPhS*, 2015.

[10] B. Munson and M. Babel, "The phonetics of sex and gender," in *The Routledge Handbook of Phonetics*, Taylor and Francis, 2019, pp. 499–525.

[11] A. P. Simpson and M. Weirich, "Phonetic Correlates of Sex, Gender and Sexual Orientation," *Oxford Research Encyclopedia of Linguistics*, Oct. 27, 2020. http://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-749 (accessed May 12, 2021).

[12] B. Munson, E. C. McDonald, N. L. DeBoe, and A. R. White, "The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech," *Journal of Phonetics*, vol. 34, no. 2, pp. 202–240, Apr. 2006, doi: 10.1016/j.wocn.2005.05.003.

[13] J. B. Pierrehumbert, T. Bent, B. Munson, A. R. Bradlow, and J. M. Bailey, "The influence of sexual orientation on vowel production (L)," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 1905–1908, Oct. 2004, doi: 10.1121/1.1788729.

[14] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men," *Journal of Language and Sexuality*, vol. 2, no. 1, pp. 1–39, 2013.

[15] L. Zimman, "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/," *Language in Society*, vol. 46, no. 3, pp. 339–370, Jun. 2017, doi: 10.1017/S0047404517000070.

[16] S. Kachel, A. P. Simpson, and M. C. Steffens, "Acoustic correlates of sexual orientation and gender-role self-concept in women's speech," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4793–4809, Jun. 2017, doi: 10.1121/1.4988684.

[17] S. Kachel, A. P. Simpson, and M. C. Steffens, "'Do I Sound Straight?': Acoustic Correlates of Actual and Perceived Sexual Orientation and Masculinity/Femininity in Men's Speech," *Journal of Speech, Language and Hearing Research (Online)*, vol. 61, no. 7, pp. 1560–1578, Jul. 2018, doi: http://dx.doi.org.ezproxy.gc.cuny.edu/10.1044/2018_JSLHR-S-17-0125.

[18] J. T. Spence and C. E. Buckner, "Instrumental and Expressive Traits, Trait Stereotypes, and Sexist Attitudes: What Do They Signify?," *Psychology of Women Quarterly*, vol. 24, no. 1, pp. 44–53, Mar. 2000, doi: 10.1111/j.1471-6402.2000.tb01021.x.

[19] T. E. Runge, D. Frey, P. Gollwitzer, R. L. Helmreich, and J. T. Spence, "Masculine (Instrumental) and Feminine (Expressive) Traits: A Comparison between Students in the United States and West Germany," *Journal of Cross-Cultural Psychology*, vol. 12, no. 2, pp. 142–162, 1981, doi: 10.1177/0022022181122002.

[20] S. Kachel, M. C. Steffens, and C. Niedlich, "Traditional Masculinity and Femininity: Validation of a New Scale Assessing Gender Roles," *Front. Psychol.*, vol. 7, 2016, doi: 10.3389/fpsyg.2016.00956.

[21] E. Passoni, E. de Leeuw, and E. Levon, "Bilinguals Produce Pitch Range Differently in Their Two Languages to Convey Social Meaning," *Language and speech*, vol. 65, no. 4, pp. 1071–1095, 2022.

[22] C. Henton, "Cross-language variation in the vowels of female and male speakers," in *Proc. XIIIth ICPhS*, 1995, pp. 420–423.

[23] J. Bishop *et al.*, "Brief Report: Autistic Traits Predict Spectral Correlates of Vowel Intelligibility for Female Speakers," *J Autism Dev Disord*, May 2021, doi: 10.1007/s10803-021-05087-5.

[24] K. Becker and A. W. Wong, "The Short-a System of New York City English: An Update," *University of Pennsylvania Working Papers in Linguistics*, vol. 15, no. 2, Mar. 2010, [Online]. Available: https://repository.upenn.edu/pwpl/vol15/iss2/3

[25] J. T. Spence and R. L. Helmreich, *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents*. University of Texas Press, 1979.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.1. 15," 2020.

THE SOCIOPHONETIC VARIATION OF /j/ AND /ɻ/ IN MALAPPURAM MALAYALAM

Naseera Thesni K, Reenu Punnoose

Indian Institute of Technology Palakkad
232004002@smail.iitpkd.ac.in

ABSTRACT

Malappuram Malayalam is a variety known for its use of /j/ as a variant of the post-alveolar approximant /ɻ/ found in most varieties of Malayalam and unique for its proto-Dravidian ancestry and intriguing phonetic-phonological features. This study investigates the sociophonetic factors that condition the /j/and /ɻ/ variation in Malappuram Malayalam amidst recent reports of dialect levelling. 615 tokens were collected from seven female and seven male participants in their early twenties via a reading task and picture-promoted conversations. Early results show the presence of three variants- /j/, /ɻ/ and a null variant, distributions of which reveal the effects of education, gender and surrounding vowel environments. Acoustically, /j/ shows a higher F2 and F3 compared to /ɻ/. The shared palatalisation feature of /j/ and /ɻ/, and the presence of a null variant surrounding /i/, also strengthen the case for the significance of secondary articulation in the phonetics-phonology interface of Malayalam.

Keywords: Sociophonetics, Variation, Approximants, Malappuram Malayalam.

1. INTRODUCTION

The variety of Malayalam spoken mainly by the Muslims in the Northern district of Malappuram is interesting for two reasons: Socially, it is a minority variety spoken by a historically oppressed and socioeconomically disadvantaged community. Linguistically, it is one of the handful of varieties of Malayalam that has an alternate-/j/- for the otherwise stable postalveolar approximant /ɻ/ the latter tracing its origin to the Proto-Dravidian sound inventory and existing presently only in Malayalam and a few Tamil varieties. For example, [maɻa] 'rain' in other varieties of Malayalam including the standard becomes [maja] in Malappuram Malayalam. The use of /j/ where other varieties have a /ɻ/ is one of the most salient features of this variety. Phonotactically, /ɻ/ occurs only intervocalically in Malayalam; /j/ occurs both initially and medially; and the two sounds are not lexically contrastive in Malayalam. Interestingly, in all the other Dravidian languages, /ɻ/ has merged with a range of different sounds (/ɭ,ɳ,ɖ,ʈ/) [3] and its alternation with /j/ appears to be unique to the Malappuram Malayalam variety and a handful of others spoken in Northern Kerala.

A recent study claims that the variety is undergoing dialect levelling which is characterised by the decline in the use of /j/ by younger speakers in favour of /ɻ/, among other features [2]. In [2], however, the researcher is an outsider to the community, a speaker of the standard variety and the only social variables of interest considered are age and gender. The general role of formal education in the adoption of an urban, mainstream, standard variety and its consequent effects on the use of a regional /non-standard one is well attested in the sociolinguistic literature. In the case of the Malappuram Muslim community, however, it is important to note that it has historically been and at present continues to be one of the most educationally deprived communities in Kerala. Nevertheless, increased outward migration over the last thirty years and greater exposure to the standard and other non-standard varieties via media and internet could all have consequences for the local speech variety.

2. PRESENT STUDY

The present study has the following objectives:

i) Examine the phonetic characteristics of the /j/-/ɻ/ variation among the young Malappuram Malayalam speakers and
ii) Investigate the effects of gender and education levels on the varying productions of /j/ and /ɻ/.

In contrast to previous work [2], the positionality of the main researcher in this study is that of a part insider-part outsider: She is a L1 speaker of the local variety spoken in Malappuram which is the subject of this study which results in the part-insider perception. She is also a highly educated female in a community in which the educational attainment level of females is low and visits to the field revealed that this contributes to a part-outsider perception.

3. METHODS

Seven male and seven female participants between 18-22 years of age having two different educational attainment levels [see table 1 below] were recorded

using ZOOM H1n recorder and atr3350xis Audio Technica microphone. A total of six hundred and fifteen tokens were collected via two types of tasks- reading task and picture-prompted conversations. The data was analysed auditorily and also acoustically using PRAAT [1]. Formant frequencies, F1, F2 and F3 of the /j/, /ɻ/ tokens as well as of the surrounding vowels were extracted at mid-point. The data set consisted of five different vowel contexts- -/a_a/, /o_i/, /a_i/, /u_u/ and /e_a/.

As the present study is interested in assessing the effects of gender and education on the /j/, /ɻ/ productions by participants as manifested in their formant patterns, a Multivariate Analysis of Variance (MANOVA) was conducted. Results are reported in 4.2.

Participants	Male	Female
School Dropouts	3	3
College-going	4	4
Total	7	7

Table 1: Social features of the participants

4. RESULTS

4.1. Auditory Analysis

Overall, the male speakers, irrespective of their education levels, produced most of the total tokens as either [j] or a [ø], the latter indicating a null variant or a deletion of the approximant altogether. Female participants' productions exhibited a higher degree of variation. Both education levels among the female participants used more [ɻ] than their corresponding male counterparts. A small proportion of tokens were realised as a tap-approximant by college- going participants. Tap-approximants are tap-like sounds characterised by a continuous formant structure (similar to approximants) and a reduced waveform, i.e., a decrease in waveform amplitude. In the present study, this variant is restricted to only one lexical item, /uɻuɲɲə/, in the u_u context, which is an uncommon vowel context for /ɻ/ in this variety and therefore was not further analysed for this paper.

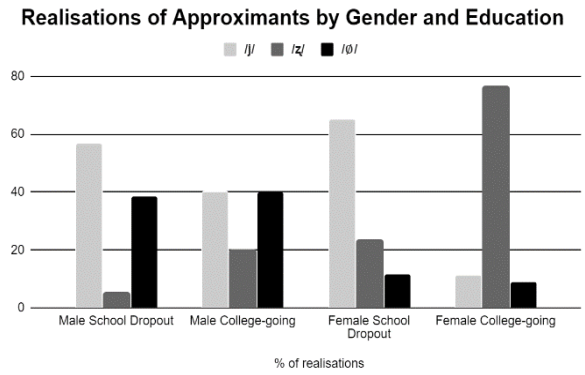


Figure 1: % of different variants produced by male and female school dropout and college-going participants

Among the two education levels, college-going female participants realised the majority of the target tokens as [ɻ] compared to their male and female school dropouts and college-going males.

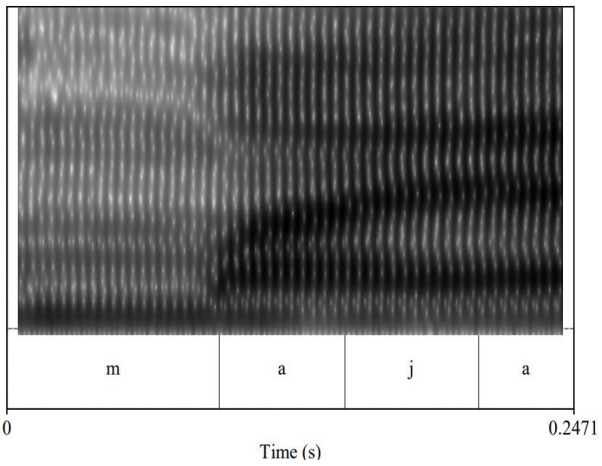


Figure 2: Spectrogram of /maɻa/ produced as [maja]

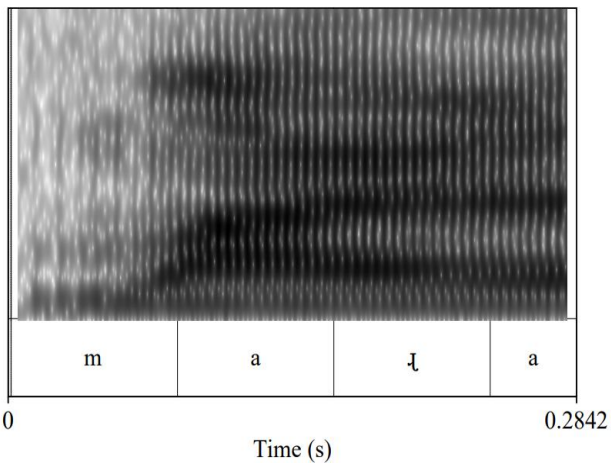


Figure 3: Spectrogram of /maɻa/ produced as [maɻa]

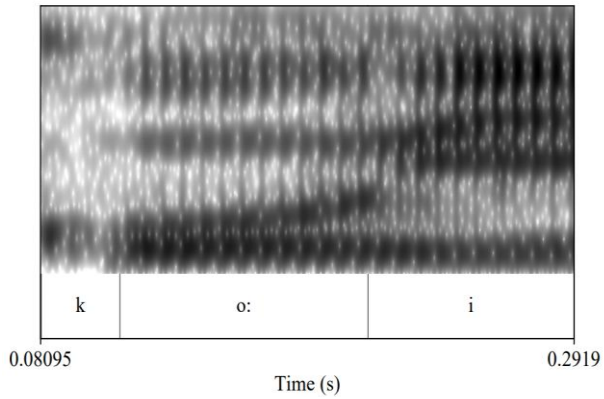


Figure 4: Spectrogram of /ko:ɻi / produced as [ko:i]

[j] and its surrounding vowels sound fronter and more raised than [ɻ] and its surrounding vowels. In the case of the male and female participants, fronter vowel environments were found to promote a [j] realisation whereas backer vowel environments were found to promote a [ɻ] realisation. A surrounding /i/ vowel was also found to promote a null variant ([ø]), especially in male participants, i.e., the deletion of the approximant altogether in 10% of such tokens among female participants and 39.2% of such tokens among male participants.

4.2. Acoustic Analysis

As shown in fig. 6, the F2-F3 frequencies of [ɻ] realisations among female college-going speakers and those of [j] realisations among female school drop-outs show a wide distribution.

For male speakers, the F2-F3 frequencies of [j] realisations among college-going speakers have the widest distribution followed by those of [j] realisations among school dropouts (fig. 5). Some degree of overlap notwithstanding, in terms of the formant characteristics of [ɻ] and [j], the latter shows overall higher F2 and higher F3 values compared to the former among female speakers. The [ɻ] and [j] realisations of college-going female speakers show considerable degree of overlap in their F2-F3 values while those of the school dropouts are more clearly distinguished. Among the male group, interestingly, the college-going speakers produce their [j] realisations with overall higher F3 values than their school dropout counterparts.

The statistical analysis (MANOVA) was carried out using the statistics software jamovi (version 2.3) [6]. The dependent variables are the F1, F2, and F3 values of different realizations of the approximant. The independent variables are education and gender of the participants and vowel context.

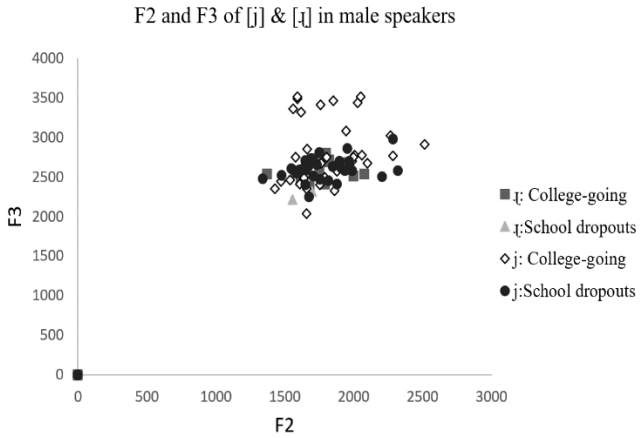


Figure 5: F2-F3 distribution of [ɻ]&[j] in college-going vs school dropout male speakers

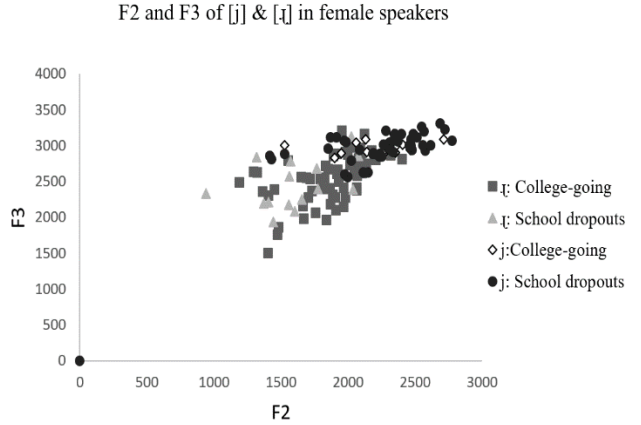


Figure 6: F2-F3 distribution of [ɻ]&[j] in college-going vs school dropout female speakers

The findings revealed that there are statistically significant interactions between gender and vowel context, F (12,127) =5.967, p<.001, Wilk's Lambda=0.307. Interaction between gender and education, gender, education and vowel contexts were not found to be statistically significant.

The segmental effect is conditioned by gender such that the /a_i/ vowel context had a significant effect on female versus male productions of the approximant. Within each of the two gender groups, /i/ in the following vowel context was found to have a significant effect on productions compared to a following /a/ context.

5. DISCUSSION

The auditory and acoustic analysis both show the significant effects of gender on the realisations of approximants in Malappuram Malayalam. Previous work [2] reported a widespread dialect levelling in the variety with a majority of the younger age group participants, irrespective of gender, producing the standard variant [ɻ] instead of the local variant [j]. In the present study, we find that young participants,

except for the college-going female group, produced around 80% of their total approximant tokens as [j]. The female college-going participants have the opposite pattern to the other three groups- male and female school drop outs and male college-going participants, in that they produce the majority of their tokens as [ɻ].

It is a well-established pattern in the sociolinguistic literature that young women lead non-stigmatised linguistic change. It appears to be the same for Malappuram Malayalam since it is the college-going female speakers who use the [ɻ] variant the most. Nevertheless, the increased use of [ɻ] among the college-going women do not appear to be causing a linguistic change among the other female and male groups in favour of the non-stigmatized variety. The focus group discussions and participant observations made during the field visits also suggest that the participants (including the educated female group) were overtly critical of people who shifted their speech towards the standard variety.

Moreover, other reasons why young women may not lead linguistic change in this community could also be to attributed to the low educational attainment levels of Muslim women in Malappuram. This is due to several factors internal and external to the community such as early marriage, religious beliefs, gender bias, poor infrastructure, fewer number of higher educational institutions, etc. Also, the nature of the tasks reported in the present study were reading task and the picture-prompted conversations, both of which were relatively on the formal side of the style continuum, the former more than the latter. Therefore, it must also be taken into account that even in tasks that are not the most spontaneous, only the female college-going group produces most tokens as standard variant, [ɻ].

As part of a larger study in progress, qualitative ethnographic methods like focus group discussions and participant observations were also conducted with a different group of male and female participants and although not directly comparable with the results of the data presented here, it is worth mentioning that qualitative research methods could significantly contribute to a more comprehensive understanding of the relationship between variation, both intra- and inter-speaker, and social meaning.

The statistical analysis of the formant measurements, did not show a significant effect of education. We speculate that this could be due to a relatively small data set and possibly also that formant frequency measures alone do not capture the all the significant aspects of the [j]-[ɻ] phonetic distinction.

Acoustically, [j] and [ɻ] are characterised by differences in F2 and F3, higher in the former compared to the latter. In the present study, vowel context was also found to have a significant effect on the nature of approximant realisation in the variety. Of the various vowel contexts, the following /i/ context was found to result in several null variants across speakers. This combined with the acoustic characteristics of /j/ and /ɻ/ are interesting in the context of Malayalam. Previous work on the phonetic characteristics of several contrasts in Malayalam have shown that clear vs dark differences as perceptual correlates of secondary articulation differences are a common distinguishing feature underlying several contrasts in the language- the singleton-geminate contrast, the dental-alveolar contrast and the five member liquid contrasts [for more on this point, see [4]]. This raises important questions regarding the presence of [j] as a variant in Malappuram Malayalam, unlike any of the variants in other languages that has led to the loss of [ɻ] in those languages. The shared palatalisation feature between [j] and [ɻ] makes the former an ideal candidate for a variant without risking the use of relative clear-dark distinction between [ɻ] and the other four liquids in Malayalam as a contrast maintenance strategy. As suggested in [4], future work must examine the role of resonance as an integral aspect of the phonology of Malayalam.

6. CONCLUSION

Auditory and acoustic analyses have shown that there are significant effects of gender and surrounding vowel context on the realisations of [j] and [ɻ] in Malappuram Malayalam, a minority variety characterised by the [j] variant as salient marker and which has recently been reported to be undergoing widespread dialect levelling. Although a larger data set would yield more robust results, the patterns observed in the present study suggest that the phonetic differences between [j] and [ɻ] and the shared palatalisation feature between the two mirror several phonemic contrasts in Malayalam in which relative resonance differences serve to maintain a crowded phonological sound system.

7. REFERENCES

[1] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer (version 6.2.06). http://www.praat.org/

[2] Jayaraj,Thapasya.2017 The Realisation of the Phonological Variation of /ɻ/ in Malabar Mappila dialect of Malayalam. *Jadavpur Journal of Languages and Linguistics.*

[3] Krishnamurti, Bh. 2003. *The Dravidian Languages.* Cambridge: Cambridge University Press

[4] Punnoose, Reenu. 2011. An auditory and acoustic study of liquids in Malayalam. Thesis (PhD). Newcastle University.

[5] Punnoose, Reenu & Khattab, Ghada & Al-Tamimi, Jalal. 2013. The Contested Fifth Liquid in Malayalam: A Window into the Lateral-Rhotic Relationship in Dravidian Languages. Phonetica. 70. 274-97. 10.1159/000356359.

[6] The jamovi project.2022. jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

THE IMPACT OF SOCIAL INFORMATION ON VOT SHADOWING BY NONBINARY SPEAKERS

Jack Rechsteiner & Betsy Sneller

Michigan State University
rechste4@msu.edu, sneller7@msu.edu

ABSTRACT

Social information can impact the degree to which one speaker phonetically converges with another speaker. There is also evidence that nonbinary speakers alter their speech due to their social environment, specifically in environments where there is a threat of being misgendered. In this paper, we investigate whether nonbinary speakers' convergence toward extended voice onset time (VOT) in word-initial English /p, t, k/ is impacted by whether they believe they are listening to another nonbinary speaker or to a cis speaker. We tested 15 speakers in an online VOT shadowing input-driven elicitation task, and we found that nonbinary speakers show statistically significant greater divergence away from the cis-labeled voice than in other conditions. These results suggest that the threat of being misgendered is a primary motivation for nonbinary speakers shifting their linguistic productions in differing social contexts.

Keywords: sociophonetics, nonbinary, phonetic imitation, social information, voice onset time

1. INTRODUCTION

Existing research on individuals who are not cisgender — or 'cis', referring to those whose gender identity matches their sex-assigned-at-birth — has largely focused on the experiences of trans people with binary trans identities [1]. However, the amount of research on speakers with gender identities that are outside of the gender binary is significantly smaller [2, 3, 4]. Nonbinary individuals present direct opposition to gender binaries, in that they may identify as being somewhere along the spectrum between male and female or outside of the male-female dichotomy entirely. The experiences of nonbinary individuals have been shown to be different from the experiences of binary-gendered individuals, such as marginalization in both cis and LGBT communities in addition to the unique challenges of being gender non-conforming in spaces that tend to understand gender as a binary framework [2, 1].

Gratton [5, 6] observed nonbinary participants varying their linguistic patterns in queer contexts compared to cisgender contexts, and argued this was motivated by the desire to counteract the possibility of their gender being assumed incorrectly. The present study aims to build upon these findings by analyzing the results of 15 nonbinary American English speakers who participated in an online phonetic imitation shadowing task. This experimental paradigm was chosen in order to investigate the degree to which nonbinary speakers are influenced by socially salient identities, even in minimally interactive conditions. We hypothesize that participants who are exposed to a model speaker stated to share a queer identity with them will show higher rates of convergence than participants who are exposed to a model speaker who does not share a queer identity with them.

1.1. Phonetic convergence

Phonetic convergence is a process whereby a speaker takes on acoustic-phonetic traits that are present in the speech of a person they are interacting with. Convergence falls within the broader category of linguistic accommodation, which has been argued to be "motivated by a desire to affiliate with or decrease social distance to a fellow interactant" as well as "underscoring common social identities" [7]. While phonetic convergence is at least partly socially facilitated [8, 9] speakers also exhibit phonetic convergence in minimally social laboratory settings [10, 11, 12] as well as in cooperative and conversational situations [13, 14].

1.2. Voice onset time

The specific phonetic variable of interest in the current study is voice onset time (VOT), which is defined as the length of time between the release of a stop consonant and the onset of voicing for the following vowel. For the purpose of this study, VOT is equivalent to a practical measurement of the aspiration of voiceless stops. VOT has been well documented as a phonetic object that

produces convergence [12, 15, 16]. Nielsen [15] found speakers extending their VOT productions for stress-initial words beginning with a voiceless stop after exposure to recordings of a model speaker with artificially extended VOT in a non-shadowing elicitation task. Extended VOT was chosen as our variable of interest because it is not an explicit stereotype of gender (in the sense of Labov 1972's *indicators, markers*, and *stereotypes* [17]), extended VOT stimuli are easy to artificially create through acoustic manipulation, and extended VOT has no phonological perception consequences for voiceless stops in English [15].

2. METHODS

2.1. Participants

The participants in this study included 15 American English speakers (ages 18-35, mean age 27) who identified as nonbinary and reported that they were born and currently live in the United States. Participants were recruited through the researchers' social media networks. Participating in the study took roughly 15 minutes from beginning to end, and participants were compensated for their participation.

2.2. Stimuli

The stimuli consisted of 54 words — 40 target words and 14 filler words. All of the words in the stimuli are bisyllabic, stress-initial words. with a frequency between 1 and 25 per million based on frequency scores provided by the SUBTLEXUS database [18]. Low frequency words were chosen because previous research has shown phonetic properties in low frequency words to show a higher degree of convergence than high frequency words [10]. For the target words, 16 have word-initial /p/, 16 have word-initial /k/, and 8 have word-initial /t/; no target words have initial onset clusters. All 14 filler words begin with vowels. This stimuli set is consistent with the stimuli used in previous studies on extended VOT convergence [12, 15, 16]. Mean frequencies for the words in each category of word-initial stop can be seen in Table 1.

The model speaker was an American English speaker who was determined to sound appropriately gender-ambiguous to listeners via a pre-experiment norming study. The model speaker provided recordings of the 54 stimuli words, as well as the audio instruction portion of the experiment. The original VOT of initial consonants was measured and then extended using the Duration Tier in Praat's manipulation features [19] to create VOTs that were, on average, 102% longer than the original VOT. This method was chosen in order to avoid auditory aberrations, such as aperiodic bursts, that can occur when manipulating VOT through other means.

Initial stop	Mean FPM	Example word
/p/	9.87	pollen
/t/	13.21	timer
/k/	10.78	cabin

Table 1: Mean frequency per million (FPM) of target words by word-initial stop with examples of words used.

2.3. Procedure

The experiment consists of a shadowing input-driven elicitation task where each participant is assigned to 1 of 3 conditions for a between-subject experiment design. In this experimental paradigm, words are presented to participants before, during, and after exposure to a model speaker and participants record themselves speaking the word aloud. The experiment was built with and administered online using PsychoPy [20]. After giving informed consent, participants completed a demographic survey to collect information on their age, gender identity, residential history, race/ethnicity, and education. Participants then took part in 3 phases of the shadowing task where they recorded themselves saying the given word within the carrier phrase, "The word is ____."

Phase 1 (Baseline Phase) elicited participants' baseline productions by presenting written instructions and words on the screen without any auditory exposure. Phase 2 (Exposure Phase) presented participants with audio instructions and words read aloud by the model speaker. Phase 3 (Post-exposure Phase) again presented participants with written words with no accompanying audio. The order in which the words were presented in each phase was randomised for all participants.

In the Exposure Phase, participants were given auditory instructions from 1 of 3 conditions. In the Nonbinary Condition, the model speaker begins by explicitly identifying themself as nonbinary ("My name is Sam. I am nonbinary and my pronouns are they/them"). In the Neutral Condition, the model speaker does not give any information about their gender ("My name is Sam"). In the Cis Condition, the speaker explicitly identifies themself as cis ("My name is Grant and my pronouns are he/him").

Aside from this introductory gender identity information, the recordings for the model speaker

were identical in each condition. The recordings were from the same model speaker in each condition, and the pre-experiment norming study on gender ambiguity of the model speaker aimed to mitigate effects that would cause participants to assume the gender of the speaker in the Neutral Condition. The structure of these conditions was motivated by the hypothesis that nonbinary speakers are more likely to converge with a model speaker they perceive as nonbinary, as suggested by the results from Gratton [5, 6] which showed nonbinary participants were more likely to pattern together in queer spaces than in non-queer spaces. Participants were distributed evenly across the 3 conditions, resulting in 5 participants for each condition.

3. RESULTS

Following Nielsen [15], the VOT of participant responses was measured in Praat [19]. Recordings of the shadowing task were transcribed orthographically and force aligned using the Montreal Forced Aligner [21] via DARLA [22]. Measuring the VOT of target words in Praat was done manually, assisted by the get_vot Praat script [23].

Unexpectedly, all conditions saw a decrease in participant VOT values during the Exposure Phase compared to their Baseline Phase, suggesting *divergence* from the model talker (Fig. 1). This may be the result of participants using a hyperarticulated "citation style" in the Baseline phase, and becoming more familiar with the task in subsequent phases, or it could be the result of social divergence. Either way, our focus here is not on the direction of effect, but rather on the differences in degree of divergence across the three social conditions. These results were analyzed using a linear mixed-effects model in RStudio [24], with VOT as the dependent variable and an interaction term between the fixed effects of Experiment Phase (Baseline, Exposure, and Post-Exposure) and Condition (Neutral, Cis, and Nonbinary), fixed effects for the initial stop (p, t, k) and the height of the following vowel (low, mid, high). Random intercepts were included for speaker and word. Table 2 shows an overview of the statistics of this model. The model formula used was:

$$(1)\quad lmer(VOT \sim Phase * Condition + Initial\,stop + Vowel\,height + (1|Speaker) + (1|Word))$$

Here, we discuss the significant results. The reference level shows that the average VOT value for participants in the Baseline Phase of the Neutral

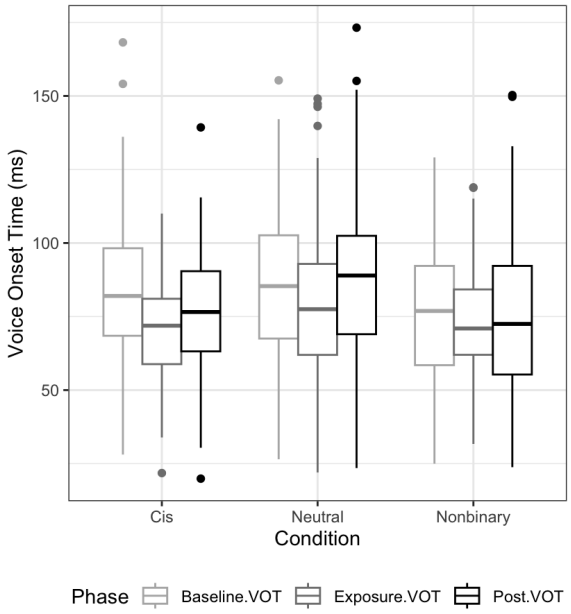


Figure 1: VOT values across the 3 conditions.

Fixed Effect	Estimate	P value
Neutral Baseline VOT	**77.47**	**<.001*** **
(Neutral) Exposure	**-8.07**	**<.001*** **
(Neutral) Post	1.94	.21
Cis (Baseline)	-2.81	.68
Cis : Exposure	**-4.94**	**.03***
Cis : Post	**-9.52**	**<.001*** **
Nonbinary (Baseline)	-11.12	.12
Nonbinary : Exposure	**5.86**	**.008** **
Nonbinary : Post	-2.94	.18
Initial Stop /p/	**-9.32**	**.01***
Initial Stop /t/	5.53	.19
Vowel Height Low	**18.91**	**<.001*** **
Vowel Height Mid	7.11	.08

Table 2: Results of the mixed effects model.

condition was 77.47 ms. We see a significant main effect ($p < 0.001$) of Exposure phase for the Neutral condition, indicating that participant VOTs diverged from the model talker by *decreasing* by 8.07 ms. The interaction between Cis condition and Exposure phase shows a marginally significant effect ($p = 0.03$), showing participants diverging even more in the Cis condition Exposure phase than in the Neutral condition Exposure phase (an additional 4.94 ms shorter, on top of the main effect of Exposure phase). The interaction between Cis condition and Post phase shows participants in the Cis condition maintaining their divergence ($p <$

0.001) in the Post exposure phase, meaning that their divergence from the model talker persisted even beyond immediate exposure. The interaction between Nonbinary condition and Exposure phase ($p = 0.008$) shows participants in the Nonbinary condition still diverging (-8.07 main effect + 5.86 interaction effect = -2.21 ms), but significantly *less* than participants in the Neutral or Cis conditions. Finally, we find expected significant main effects of initial stop, with /p/ showing significantly shorter VOT ($p = 0.1$, 9.32 ms), and low vowels showing significantly longer VOT ($p < 0.001$, 18.91 ms).

4. DISCUSSION

This study examined VOT imitation effects in 15 American English speakers across 3 different experimental conditions with the prediction that nonbinary speakers would show stronger convergence when they were told the model speaker is also nonbinary (Nonbinary Condition) as compared to the other conditions. This prediction was based on previous observations that suggest the threat of being misgendered is a primary motivation for nonbinary speakers shifting their linguistic productions in differing social contexts [5, 6].

The results showed a surprising tendency for participants in all 3 conditions to diverge from, rather than converge with, the model speaker's VOT. Patterns of consistent divergence away from a model talker, like those seen in this study, highlight that phonetic imitation is not simply an automatic process, but instead mediated by social factors [8, 13, 14]. For example, Babel [8] found that male participants who rated a model talker as attractive were more likely to diverge from that talker's production. Babel posits that these participants "were, perhaps, *socially threatened* and distanced themselves in response to the threat" (emphasis ours). In our case, the difference in divergence across conditions also shows the strong influence of social factors. We found that nonbinary participants diverged the *most* in the Cis Condition (-9.52 ms, $p < 0.001$). We posit that nonbinary participants interpreted a social threat associated with a cis model talker, such as the threat of being misgendered [5, 25], which was strong enough to motivate participants to linguistically distance themselves from a cis-identified talker.

Additionally, VOT values from the Exposure Phase diverged the *least* in the Nonbinary Condition (5.86 ms, $p = 0.008$), suggesting that nonbinary participants align their speech most closely to a model talker when they are explicitly identified as sharing a nonbinary identity. We interpret this that participants who are in an explicitly queer virtual setting, even a very low-interaction one, converge towards a shared nonbinary speech norm.

These findings furthermore align with previous work which argued that in conversational speech in queer contexts, nonbinary speakers pattern more like each other regardless of sex assigned at birth, effectively creating a distinct nonbinary speech community [5, 26].

5. CONCLUSION

This study aimed to investigate the impact that model talker gender identity has on the direction and degree to which nonbinary speakers converge in VOT. We found that compared to a model talker who is unlabeled for gender identity, a nonbinary model talker resulted in significantly less divergence for nonbinary participants. We additionally found that a cis-labeled model talker resulted in significantly more divergence for nonbinary participants. These results suggest that even in low-interaction virtual settings, being in an explicitly queer context enables nonbinary speakers to pattern more like another nonbinary speaker than like a cis-identified speaker.

Previous phonetic imitation studies have shown that speaker gender does not have a consistent, significant effect on imitation [13, 27, 28]. This is not to say, however, that gender does not matter for imitation or convergence. Pardo [13] noted that phonetic imitation "is subject to situational constraints that influence the direction and magnitude of phonetic convergence", and this is precisely what our findings show. Different situational contexts — in this case, whether nonbinary participants have entered an explicitly queer virtual environment or an explicitly heteronormative one — impact phonetic imitation.

6. REFERENCES

[1] A. E. Goldberg and K. A. Kuvalanka, "Navigating identity development and community belonging when "there are only two boxes to check": An exploratory study of nonbinary trans college students," *Journal of LGBT Youth*, vol. 15, no. 2, pp. 106–131, Apr. 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/19361653.2018.1429979

[2] N. J. Bradford, G. N. Rider, J. M. Catalpa, Q. J. Morrow, D. R. Berg, K. G. Spencer, and J. K. McGuire, "Creating gender: A thematic analysis of genderqueer narratives," *International Journal of Transgenderism*, vol. 20, no. 2-3, pp. 155–168, Jul. 2019. [Online].

Available: https://www.tandfonline.com/doi/full/10.1080/15532739.2018.1474516
[3] A. Garmpi, "The Discursive Construction and Performance of Non-Binary Identity," Master's thesis, The University of Edinburgh, 2020.
[4] J. Jones, "Authentic Self, Incongruent Acoustics: A Corpus-Based Sociophonetic Analysis of Nonbinary Speech," Ph.D. dissertation, University of Canterbury, 2022.
[5] C. Gratton, "Resisting the Gender Binary: The Use of (ING) in the Construction of Non-binary Transgender Identities," *University of Pennsylvania Working Papers in Linguistics*, vol. 22, no. 2, 2016. [Online]. Available: https://repository.upenn.edu/pwpl/vol22/iss2/7
[6] ——, "Non-binary identity construction and intraspeaker variation," in *The 91st Annual Meeting of the Linguistic Society of America, University of Texas at Austin*, 2017. [Online]. Available: https://academia.edu/30897570/Non_binary_Identity_Construction_and_Intraspeaker_Variation
[7] J. Gasiorek, H. Giles, and J. Soliz, "Accommodating new vistas," *Language & Communication*, vol. 41, pp. 1–5, 2015.
[8] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *Journal of Phonetics*, vol. 40, no. 1, pp. 177–189, 2012.
[9] N. Coupland, "Accommodation at work: Some phonological data and their implications," *International Journal of the Sociology of Language*, vol. 39, no. 2, pp. 49–70, 1984.
[10] S. D. Goldinger, "Echoes of echoes? an episodic theory of lexical access," *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998.
[11] S. D. Goldinger and T. Azuma, "Episodic memory reflected in printed word naming," *Psychonomic Bulletin & Review*, vol. 11, no. 4, pp. 716–722, 2004.
[12] K. Shockley, L. Sabadini, and C. A. Fowler, "Imitation in shadowing words," *Perception & Psychophysics*, vol. 66, no. 3, pp. 422–429, 2004.
[13] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.
[14] J. S. Pardo, A. Urmanche, S. Wilman, J. Wiener, N. Mason, K. Francis, and M. Ward, "A comparison of phonetic convergence in conversational interaction and speech shadowing," *Journal of Phonetics*, vol. 69, pp. 1–11, Jul. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0095447017300992
[15] K. Nielsen, "Specificity and abstractness of VOT imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, Apr. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0095447011000027
[16] J. Schertz, E. K. Johnson, and M. Paquette-Smith, "The independent contribution of voice onset time to perceptual metrics of convergence," *JASA Express Letters*, vol. 1, no. 4, p. 045205, Apr. 2021. [Online]. Available: https://asa.scitation.org/doi/10.1121/10.0004373
[17] W. Labov, *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972.
[18] M. Brysbaert and B. New, "Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior Research Methods*, vol. 41, no. 4, pp. 977–990, Nov. 2009. [Online]. Available: http://link.springer.com/10.3758/BRM.41.4.977
[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org/
[20] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019. [Online]. Available: https://doi.org/10.3758/s13428-018-01193-y
[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," *Interspeech 2017*, 2017.
[22] S. Reddy and J. Stanford, "A web application for automated dialect analysis," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 71–75. [Online]. Available: https://aclanthology.org/N15-3015
[23] J. Kang, "get_vot Praat Script," 2017. [Online]. Available: https://github.com/HaskinsLabs/get_vot
[24] Posit team, *RStudio: Integrated Development Environment for R*, Posit Software, PBC, Boston, MA, 2022. [Online]. Available: http://www.posit.co/
[25] L. Konnelly, "Nuance and normativity in trans linguistic research," *Journal of Language and Sexuality*, vol. 10, no. 1, pp. 71–82, Feb. 2021. [Online]. Available: http://www.jbe-platform.com/content/journals/10.1075/jls.00016.kon
[26] J. Rechsteiner and B. Sneller, "Nonbinary speakers' use of (ING) across gender-related topics," 2021, NWAV 49. [Online]. Available: https://vimeo.com/627644620
[27] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender Differences in Vocal Accommodation:: The Role of Perception," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 422–432, Dec. 2002. [Online]. Available: http://journals.sagepub.com/doi/10.1177/026192702237958
[28] J. S. Pardo, K. Jordan, R. Mallari, C. Scanlon, and E. Lewandowski, "Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures," *Journal of Memory and Language*, vol. 69, no. 3, pp. 183–195, Oct. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0749596X13000545

MODELING SOCIAL MEANINGS OF PHONETIC VARIATION AMID VARIABLE CO-OCCURRENCE: A MACHINE-LEARNING APPROACH

Dan Villarreal & James Grama

University of Pittsburgh; Sociolinguistics Lab, University of Duisburg-Essen
d.vill@pitt.edu, james.grama@uni-due.de

ABSTRACT

The social meaning of phonetic variation is central to sociophonetics. Despite the co-occurrence of multiple potentially meaningful variables in speech, previous experimental approaches generally focus on single features. This analysis investigated whether machine-learning methods can uncover vowel features that impact sociolinguistic perceptions. Ninety-seven listeners rated short, spontaneous California English stimuli on 12 affective scales; /æ/ and /u/ tokens were manipulated to create "California-shifted" and "unshifted" guises, with other vowels left to vary naturally. Mirroring "bag-of-words" approaches to text corpora, we treated stimuli as "bags of features" based on 11 vowel phonemes. We used the Boruta feature-selection algorithm to assess the importance of these features, plus guise, on affective scale ratings. The most frequently selected variables included both predictable (/æ/) and less well-attested variables (/ʊ/). However, guise was never selected, suggesting it was an overly-coarse axis of relevant variation. We argue that "bottom-up" approaches can model social meanings amid variable co-occurrence.

Keywords: vowel variation, computational sociolinguistics, social meaning, methods, California English

1. INTRODUCTION

Social meanings are the glue that connects the social and linguistic dimensions of language variation and change.[1] More formally, a social meaning is an ideologically mediated relationship between a linguistic structure and some stance, identity, or personal characteristic; these meanings emerge (and may be observed) in both production and perception. For example, a New York City English speaker may delete postvocalic /r/ to emphasize that they are an "authentic" New Yorker [1]; listeners may likewise perceive a New York City English speaker who deletes postvocalic /r/ as better suited for working-class than middle-class jobs [2]. These meanings are a central concern of the "third wave" of language variation research, which views social meaning as an essential (rather than incidental) part of language [3].

One strand of research attempts to investigate social meanings in perception, most often via matched-guise tasks (MGTs), which compare listeners' reactions to pairs of stimuli (i.e., guises) that differ only in their use of a particular variable (e.g., [ɪn] vs. [ɪŋ] variants of English *-ing* [4]). This method necessarily entails hypothesizing that a particular variable is socially meaningful, but this researcher choice means we may fail to account for other meaningful variables that are present in the stimulus. In other words, variables do not occur in isolation, including in chain shifts where multiple changes are occurring at once. This phenomenon of *variable co-occurrence* thus creates a dilemma for operationalizing and analyzing social meanings.

Given this quandary, the present analysis investigated whether machine-learning methods can uncover vowel variables that impact sociolinguistic perceptions, as these methods are particularly good at sorting through complex data. To test this methodology, we undertake a re-analysis of Villarreal [5], which investigated the social meanings of California English vowels (TRAP backing and GOOSE fronting) using an MGT. (In this paper, we use Wells' [6] notation for English vowel phonemes.) We use the Boruta algorithm [7], which selects all predictors that are important to some dependent variable (in this case, attributes like *familiar* and *feminine*). Our findings suggest that "top-down" approaches to sociolinguistic perception can miss crucial details about which co-occurring variables influence social meanings.

1.1. California English

The last 30 years has seen a groundswell of work on the vowel system of California English as spoken by white, coastal speakers (but see [8–11]). The bulk of this work has focused on two phenomena [10–15]:

1. The *Low-Back-Merger Shift* (LBMS) [16], consisting of:
   a. the merger (or near-merger) of LOT and THOUGHT in the low-back corner of the vowel space; this is hypothesized to trigger
   b. lowering and/or retraction of the short front vowels TRAP, DRESS, and KIT;
2. *Back Vowel Fronting* (BVF), involving long back vowels GOOSE and GOAT, particularly following coronal consonants.

While neither phenomenon is restricted to California or the western US states (e.g., [17–20]), perceptual research has suggested that Californians and non-Californians alike associate TRAP-backing with a Californian identity [5, 21, 22]. Proportionally less attention has been paid to the putative fronting of short back vowels FOOT and STRUT despite structural parallels to BVF [23]. The work that directly investigates these vowels rarely identifies compelling evidence of STRUT fronting over time [11], and identifies FOOT centralization as having reached completion [24].

2. METHODS

This study represents a re-analysis of Villarreal's matched-guise perception task, which used an MGT to investigate which social meanings Californian listeners perceive in California English vowels [5].

In this study, TRAP and GOOSE were chosen to represent California English vowels more broadly; each is implicated in a major vocalic subsystem of California English (TRAP in LBMS, GOOSE in BVF), and there is evidence linking each to Californian social meanings [8, 22, 25, 26]. Listeners (n=97), all from California, heard stimuli created from 12 voices (also from California) producing spontaneous speech retelling a short cartoon. In each trial (n=580), the listener identified the speaker's region and rated the speaker on 12 attribute scales. Scales were chosen based on a pretesting task and previous research on California English social meanings. Each stimulus contained either a "California-shifted" guise (where TRAP and GOOSE were acoustically manipulated to be backer and fronter, respectively), or a "conservative" guise (where TRAP and GOOSE were fronter and backer, respectively). Other vowel phonemes were left to vary naturally. For further details on these methods, including acoustic manipulation, see [5] and https://github.com/djvill/Vowel-Manipulation.

Villarreal [5] found that guise significantly affected perceived attributes for *Californian*, *sounds like a Valley girl*, and (for male speakers) *confident*; however, there was substantial variance in attribute ratings that guise failed to capture. In other words, while TRAP and GOOSE have a clear impact on listener assessments of Californian-ness, listeners were clearly also attending to cues beyond these two vowels in the signal. Furthermore, because TRAP and GOOSE were shifted in tandem in the stimuli, it was not possible to disentangle the effects of either vowel individually. Thus, we turn here to machine-learning methods to see how multiple vowel variables may have influenced listener behavior.

2.1. Feature set

A conceptual challenge to modeling the effects of multiple cues is sparsity. English has a large number of vowel phonemes unevenly distributed through the lexicon, meaning short stretches of speech are unlikely to include all possible vowels. Because stimuli were spontaneously produced (albeit all on the same topic), they all contain slightly different content and thus different vowel variables. For example, 13 of the 24 stimuli in the perception task include zero FLEECE tokens. To deal with sparsity, we treated each stimulus as a "bag of features" (based on the computational approach to text corpora as "bags of words" [27]).

These features sorted vowel variation into discrete bins, in order to assess the effect of (e.g.) high FOOT or front TRAP on listeners' perceptions. Our procedure for creating features is schematized in Fig. 1. Vowels' midpoint F1 and F2 were measured via Praat [28], hand-checked, and normalized using the *Atlas of North American English* (ANAE) procedure (based on 466–712 tokens per speaker) [13]. In order to contextualize vowel variation in our stimuli against the broader range of North American vowels, we used ANAE's "natural breaks", which define F1 and F2 bins for North American English vowel variation. This procedure yielded features across 11 vowel phonemes: FLEECE, KIT, FACE, DRESS, TRAP, PRICE, LOT, STRUT, GOAT, FOOT, and GOOSE. Four vowels (CHOICE, MOUTH, THOUGHT, and pre-nasal TRAP-N) were discarded from the final analysis as they were not adequately represented across all stimuli. The final dataset comprised 379 tokens across 24 stimuli.

2.2. Analysis

To determine which vowel variables were important for sociolinguistic perception, our analysis used the Boruta algorithm [7]. This algorithm's primary use case is feature selection, answering the question "which predictor(s) appear to impact the dependent variable?" For example, Dickson & Durantin used Boruta to determine important predictors of reflexive pronoun choice in Australian Kriol [29]. Briefly, Boruta identifies whether features increase the prediction accuracy of the dependent variable, by comparing each feature's prediction accuracy to that of randomly shuffled "shadow" features.

In this study, separate Boruta models were run with each attribute as the dependent variable, via the `Boruta` package in R [30, 31]. This procedure returned a list of selected features for each attribute scale. We interpret a feature's selection as indicating that the feature impacted listeners' perceptions vis-à-vis the attribute scale in question.

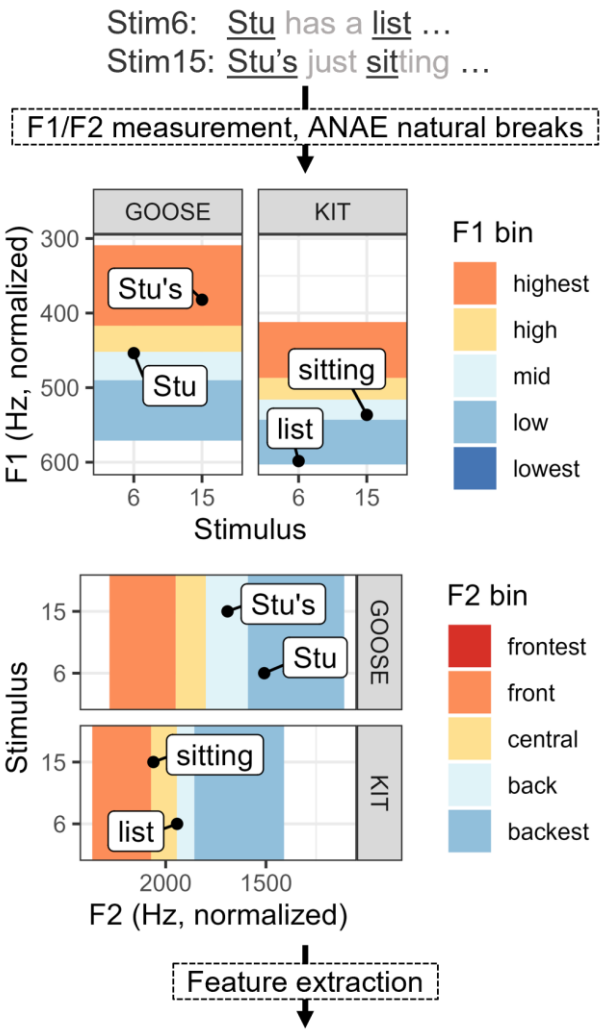


Figure 1: Schematic for creation of feature set from stimuli, based on ANAE [13] natural breaks.

3. RESULTS

Here, we primarily focus on which variables were most frequently selected by the Boruta algorithm. These results are presented in Table 1. We identify three key findings:

1. Vowels that are changing (or have changed) in California English are not necessarily more impactful to listeners' perceptions.
2. The most impactful vowel variables were FOOT, TRAP, and GOAT.
3. Guise was never selected as important to attribute scale ratings.

Variable	F1	F2
FOOT	10	8
TRAP	8	6
GOAT	4	10
FLEECE	7	6
KIT	6	6
FACE	4	6

Variable	F1	F2
LOT	5	5
GOOSE	5	5
PRICE	5	4
STRUT	5	3
DRESS	3	4

Table 1: For each variable and formant, the number of times at least one feature was selected for an attribute scale (out of 12); bold = change in progress; italics = marginal evidence of change in progress.

As Table 1 indicates, different vowel variables did not all influence attribute scale ratings to the same degree. For example, at least one FOOT F1 feature was selected as important for 10 of 12 scales, but only 3 of 12 scales for DRESS F1. Pooling these results across F1 and F2 presents a picture of each variable's impact on listeners' perceptions overall, and which variables were more *impactful* than others.

These results indicate that oft-studied variables in California English are not necessarily more impactful to listeners' perceptions, at least with regard to this particular set of attributes. The most impactful variable identified here, FOOT, is rarely investigated in California despite being structurally related to well-attested GOOSE and GOAT fronting; FLEECE, which is almost completely unattested as undergoing change (but see [11]), outranks several well-attested California English variables. Moreover, the historical ordering of LBMS sound changes (TRAP then DRESS then KIT) is not reflected in these variables' selection frequencies relative to one another. This evidence bolsters Eckert and Labov's claim that only individual variables, not the structural phenomena that they comprise (e.g., chain shifts), are available for social meaning [32].

Underneath the results in Table 1 were different distributions of features selected as important. This is exemplified by the F2 features from the highest-ranking variables: FOOT, TRAP, and GOAT (see Fig. 2). In some cases, a single attribute scale was influenced by multiple features from the same variable (e.g., both front GOAT and central GOAT influenced *friendly* perceptions); in others, a single feature acted alone. Pooling across scales, it is clear that some features stood out to listeners more than others. For FOOT, fronter features had a greater impact on listeners' perceptions than backer features; for GOAT, however, impact was assigned relatively evenly across features.

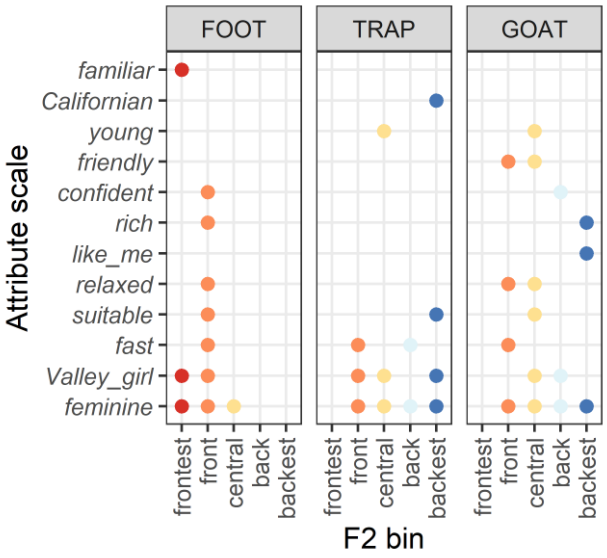


Figure 2: F2 features selected by Boruta models for FOOT, TRAP, and GOAT.

None of the Boruta models selected guise as an important predictor, despite the frequent selection of TRAP and GOOSE, the variables implicated in the guise manipulation. This result is surprising given Villarreal's [5] original finding that guise affected *Californian*, *sounds like a Valley girl*, and *confident*. In addition, our finding that TRAP was more impactful than GOOSE (Table 1) suggests that this original guise finding was driven by more by TRAP than GOOSE. This re-analysis suggests that as operationalized in [5], guise was an overly-coarse method of modeling socially meaningful vowel variation.

4. DISCUSSION

Social meanings are crucial to understanding language variation and change, and sociolinguists have an array of methodological tools at their disposal for investigating them, both in production and perception. Our findings raise critical questions about what gets overlooked by existing methods such as MGTs. However, the feature-selection method we have presented here is not without its limitations. We thus suggest that, rather than undermining the established methodological toolkit for investigating social meanings, the feature-selection method we have presented complements this toolkit.

The status of FOOT as the most frequently selected variable suggests that it may be a more prominent indicator of Californian-ness than previous research on California English identifies. Despite being part of early descriptions of California vowels [23], FOOT is often left out of discussions of California English vowel variation (but see [24]), and almost completely absent from research on California vowels' social meanings (but see [25] for an example of fronted FOOT in Kristen Wiig's parodic performance and [33] for a putative KIT–FOOT merger among South Asian and Korean-identifying Californians). This overlooking of FOOT is likely a matter of methodological expedience—compared to better-studied long back vowels GOOSE and GOAT, FOOT tokens are shorter (and thus more prone to measurement error), and less frequent in content words. Still, our findings imply that FOOT may be more relevant to Californian social meanings than previously assumed (see also [34]).

A further strength of this feature-selection method is that it accounts for variable co-occurrence when measuring social meanings in perception. Variable co-occurrence is a "known problem" in sociolinguistics (e.g., [35]), but actually accounting for co-occurrence in perception is challenging. Production research on co-occurrence typically appeals to analytic constructs such as stances and styles [36], indexical fields [37], and sociolects [38]. Studies of real-time listener reactions [39] represent a significant step forward, but can only measure one perceived attribute at a time.

This method is not without its limitations. While it mitigates the potential bias MGTs introduce by choosing particular variables to manipulate, there is still an element of researcher choice with respect to which features are in the "bag". Our feature set included a broad range of vowel categories, but we lack features (e.g., consonants, prosody) that may shape social meanings. Next, the particular selection of attribute scales in any perception task should not be taken to represent all possible social meanings. We would not expect the selection frequencies in Table 1 to be exactly replicated under a different set of attribute scales. Finally, this study benefited from benchmarks for discretizing vowel variation based on a large reference corpus, which is unavailable in under-resourced language contexts.

Perhaps most importantly, this method (and indeed large-scale methods more broadly) can only give a "bird's-eye" view of social meanings. Feature-selection algorithms like Boruta cannot, on their own, identify the direction or magnitude of any feature's effect, only that it has been assessed as important. MGTs offer unparalleled experimental control for zeroing in on particular variables. Like MGTs themselves, this method does not account for listener expectations that are known to affect sociolinguistic perceptions (e.g., [40]). The true value of this feature-selection method is guiding researchers' choice of variables for further study using tried-and-true methods for investigating social meaning.

5. REFERENCES

[1] Becker, K. 2009. /r/ and the construction of place identity on New York City's Lower East Side. *J. Sociolinguistics.* 13, 634–658.

[2] Labov, W. 2006. *The social stratification of English in New York City*, 2nd ed. Cambridge U. Press.

[3] Eckert, P. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Ann. Rev. Anthro.* 41, 87–100.

[4] Campbell-Kibler, K. 2007. Accent, (ING), and the social logic of listener perceptions. *Am. Speech* 82, 32–64.

[5] Villarreal, D. 2018. The construction of social meaning: A matched-guise investigation of the California Vowel Shift. *J. Eng. Ling.* 46, 52–78.

[6] Wells, J. C. 1982. *Accents of English*. Cambridge U. Press.

[7] Kursa, M. B., Jankowski, A., Rudnicki, W. R. 2010. Boruta: A system for feature selection. *Fundamenta Informaticae* 101, 271–285.

[8] Eckert, P. 2008. Where do ethnolects stop? *Intl. J. Biling.* 12, 25–42.

[9] Podesva, R. J., D'Onofrio, A., Van Hofwegen, J., Kim, S. K. 2015. Country ideology and the California Vowel Shift. *Lang. Var. and Change* 27, 157–186.

[10] D'Onofrio, A., Eckert, P., Podesva, R. J., Pratt, T., Van Hofwegen, J. 2016. The low vowels in California's Central Valley. In: Fridland, V., Evans, B., Kendall, T., Wassink, A. (eds), *Speech in the Western states, volume 1*. Duke U. Press, 11–32.

[11] D'Onofrio, A., Pratt, T., Hofwegen, J. V. 2019. Compression in the California Vowel Shift: Tracking generational sound change in California's Central Valley. *Lang. Var. and Change* 31, 193–217.

[12] Hagiwara, R. 1997. Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America* 102, 655–658.

[13] Labov, W., Ash, S., Boberg, C. 2006. *The atlas of North American English*. Mouton.

[14] Kennedy, R., Grama, J. 2012. Chain shifting and centralization in California vowels: An acoustic analysis. *Am. Speech* 87, 39–56.

[15] Hall-Lew, L., Cardoso, A., Kemenchedjieva, Y., Wilson, K., Purse, R., Saigusa, J. 2015. San Francisco English and the California Vowel Shift, in *Proc. 18th ICPhS*.

[16] Becker, K. 2019. *The Low-Back-Merger Shift*. Duke U. Press.

[17] Boberg, C. 2019. A closer look at the short front vowel shift in Canada. *J. Eng. Ling.* 47, 91–119.

[18] Fridland, V., Kendall, T., Evans, B., Wassink, A. B. 2016. *Speech in the Western states, volume 1*. Duke U. Press.

[19] Fridland, V., Kendall, T., Evans, B., Wassink, A. B. 2017. *Speech in the Western states, volume 2*. Duke U. Press.

[20] Nesbitt, M., Stanford, J. N. 2021. Structure, chronology, and local social meaning of a supra-local vowel shift: Emergence of the Low-Back-Merger Shift in New England. *Lang. Var. and Change* 33, 269–295.

[21] Villarreal, D., Kohn, M. 2021. Local meanings for supra-local change: Perceptions of TRAP backing in Kansas. *Am. Speech* 96, 45–77.

[22] D'Onofrio, A. 2015. Persona-based information shapes linguistic perception: Valley Girls and California vowels. *J. Sociolinguistics.* 19, 241–256.

[23] Hinton, L. et al. 1987. It's not just the Valley Girls: A study of California English. *Proc. 13th Ann. Mtg. Berkeley Ling. Soc.* 117–128.

[24] Holland, C. L. 2014. Shifting or shifted? The state of California vowels. PhD dissertation, University of California, Davis.

[25] Pratt, T., D'Onofrio, A. 2017. Jaw setting and the California Vowel Shift in parodic performance. *Lang. in Soc.* 46, 283–312.

[26] Podesva, R. J. 2011. The California Vowel Shift and gay identity. *American Speech* 86, 32–51.

[27] Jurafsky, D., Martin, J. H. 2022. *Speech and language processing*, draft of 3rd ed.

[28] Boersma, P., Weenink, D. 2022. Praat. Version 6.2.23

[29] Dickson, G., Durantin, G. 2019. Variation in the reflexive in Australian Kriol. *Asia-Pacific Lang. Var.* 5, 171–207.

[30] Kursa, M. B., Rudnicki, W. R. 2010. Feature selection with the Boruta package. *J. Stat. Soft.* 36, 1–13.

[31] R Core Team 2022. R: A language and environment for statistical computing. Version 4.2.1

[32] Eckert, P., Labov, W. 2017. Phonetics, phonology and social meaning. *J. Sociolinguistics.* 21, 467–496.

[33] Cheng, A. 2016. A Survey of English vowel spaces of Asian American Californians. *UC Berkeley PhonLab Annual Report* 12, 348–384.

[34] Gardner, M. H., Roeder, R. V. 2022. Phonological mergers have systemic phonetic consequences: PALM, trees, and the Low Back Merger Shift. *Lang. Var. and Change* 1–24.

[35] Gumperz, J. J. 1982. *Discourse strategies*. Cambridge U. Press.

[36] Kiesling, S. F. 2009. Style as stance. In: Jaffe, A. (ed), *Stance: Sociolinguistic perspectives*. Oxford U. Press, 171–194.

[37] Eckert, P. 2008. Variation and the indexical field. *J. Sociolinguistics.* 12, 453–476.

[38] Guy, G. R. 2013. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *J. Prag.* 52, 63–71.

[39] Montgomery, C., Moore, E. 2018. Evaluating S(c)illy voices: The effects of salience, stereotypes, and co-present language variables on real-time reactions to regional speech. *Language* 94, 629–661.

[40] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48, 865–892.

[1] We thank Nicole Holliday, Carina Ahrens, and Anne-Marie Mölders, as well as two anonymous reviewers, for valuable feedback on an earlier version of this paper. This research was funded in part by the UC Davis Department of Linguistics.

BIASES AND SPEECH-TO-TEXT EFFICACY FOR BRITISH ENGLISH VARIETIES

Kyra Hung[a], Amanda Cardoso[b], Devyani Sharma[c], Erez Levon[d]

[a,b]University of British Columbia, [c]Queen Mary University of London, [d]University of Bern
[b]amanda.cardoso@ubc.ca, [c]d.sharma@qmul.ac.uk, [d]erez.levon@unibe.ch

ABSTRACT

Multiple factors impact speech-to-text (STT) performance. Previous work demonstrated poorer performance for North American marginalised English varieties [1] and faster speech rates [2]. We examine STT efficacy of British English varieties, considering: i) the varieties' social statuses; ii) individual repertoires; and iii) speech rate to better understand how potential biases relate to STT efficacy. Speech rate differences impact STT efficacy most with faster (compared to slower) speech rates performing worse. These effects interact with varieties' status and individual repertoires. STT performs better for non-marginalised compared to marginalised varieties with similar speech rates. Marginalised variety speakers who use more accent features have worse STT performance compared to speakers who use fewer accent features. For these speakers, slower speech rates also do not consistently improve performance, which is not found for non-marginalised speakers. While speech rate impacts STT efficacy, we also find accent bias through the differential performance for varieties and individuals.

Keywords: accent bias, speech rate, google speech-to-text, non-standard varieties, accentedness

1. INTRODUCTION

Investigations of differences in automatic speech recognition and speech-to-text performance, which is termed "efficacy" in industry discussions, have found evidence of social biases (e.g., [1, 3, 4]). These investigations suggest that performance differences relating to a number of social variables, such as gender, race, and social status of North American varieties, demonstrate social biases that are upheld in the datasets used in the back-end of the speech recognition systems (e.g., in training) through inclusion and/or exclusion of particular speakers. For example, STT performs worse with speakers of marginalised varieties (often also referred to as non-standard) compared with speakers of non-marginalised ones (often also referred to as standard).

However, some inconsistencies are reported in the results across investigations, either in the magnitude or presence of STT efficacy differences relating to these social variables (e.g., [4]). These inconsistencies across studies may, in part, be due to different speakers being used in the investigations when assessing STT efficacy, as individuals who speak the "same" variety will produce different features or have different levels of use of features that are generally associated with those varieties. In other words, individual repertoires will mean that every speaker of a variety sounds different, to some extent, from every other speaker of that variety. These differences could relate to STT efficacy, in particular, for speakers that use many features that are stigmatised, associated with lower status, or associated with marginalised groups. These speakers are more likely to be excluded in the training data due to social biases and access to speakers.

STT efficacy is also affected by a range of other factors, such as speech rate [2] (see [5] for a review). It has been found that faster speech impacts speech recognition efficacy. Speech rate differences for individuals and varieties have so far not been considered when assessing performance differences in speech recognition systems and social biases. However, there may be differences in the typical speech rate across varieties and across individuals, which could contribute to the STT efficacy differences that are found comparing varieties.

Our investigation considers these three factors (variety social status, individual repertoires, and speech rate) by exploring STT efficacy differences for a content-controlled (i.e., same lexical content) set of recordings by speakers of British English varieties. We examine whether the type of accent bias found for North American varieties hold in this sample (i.e., worse performance for marginalised varieties), what the effect of individual repertoires is

on STT performance (i.e., differential use of accent features) and what the effect of speech rate is on its own and when considering variety social status and individual repertoires. Based on previous work, we would predict that STT efficacy will be better for: i) non-marginalised varieties compared with marginalised ones; ii) speakers whose individual repertoires include fewer marginalised features compared with those that include more marginalised features; iii) slower speech rates compared with faster ones.

2. METHODS

STT outputs were generated using Google API STT (used by, e.g., YouTube's auto-captions) for scripted responses to 15 interview questions performed by 10 male speakers of 5 British English varieties (2 speakers per variety) which are high quality recordings created for the Accent Bias Britain project (https://accentbiasbritain.org/; for more details about the data see [6]). These scripted responses are the same for all speakers, which limits content or lexical differences and allows for a more controlled comparison of STT performance differences. The five varieties represent intersections of different socially-relevant contrasts in the UK, such as region and social status. Estuary English (EE), Multicultural London English (MLE) and Received Pronunciation (RP; aka Queen's English) are all southern varieties with the first two being marginalised. General Northern English (GNE) and Urban West Yorkshire English (UWYE) are northern varieties with the first being non-marginalised. MLE is also considered a multiethnic urban variety [7], while all the others are associated with white ethnicity.

The STT outputs are compared to the spoken transcripts to calculate word error rate (WER) for each speaker and for each question by speaker. WER is used in academic and industry assessments of STT efficacy (see, e.g., [1]) and is calculated by summing the number of deletions (i.e., a word appears in the spoken transcript but not in the STT output), insertions (i.e., a word appears in the SST output but not in the spoken transcript), and substitutions (i.e., a different word appears in the spoken transcript compared with the STT output) divided by the number of words in the full spoken transcript. Higher WERs indicate worse performance.

A dialect density measure (DDM) is used to quantify differences in individual repertoires by calculating the extent of use of accent features. Accent features here refer to ways of producing sounds or combinations of sounds that differ from careful speech and are associated with, at least, one British English variety. Two question responses (Q6 and Q13) for each speaker were coded by two trained phoneticians for presence of 61 accent features, which encompassed a wide variety of features ranging from very localisable (i.e., associated with one variety) to completely non-marginalised (i.e., occurs regularly in all British English varieties). These features were chosen as they are reported in previous literature to occur in one or more of the varieties (see [8] for more details about accent features chosen). For example, we coded for /k/-backing which occurs only in MLE, merged *and* split FOOT-STRUT vowels which differentiate northern and southern varieties, /ð/-fronting which occurs in the northern and southern marginalised varieties, and intrusive-r which occurs regularly in all British English varieties (see [8] for more information about DDMs, where we also discuss relative stigmatisation of features and their effect on perceived accentedness). The DDM value is the sum of all accent features used by the speaker divided by the number of times that an accent feature could have been used based on the known patterns for these accent features. This measure correlates well with differences in perceived accentedness [8] and quantifies differences in individual repertoires within and across varieties. Using a proportional measure accounts for differences in the likelihood of accent features occurring across the varieties, as all features do not occur equally likely in all varieties. Furthermore, DDMs are scaled to allow us to better compare across varieties as well as across speakers within a variety.

Finally, for a measure of speech rate, the average phone duration for each utterance for the two interview responses used in the DDM calculations (Q6 and Q13) was calculated. Speech was divided into utterances using pause length. Pauses greater than 300 ms were taken to indicate a new utterance. Then average phone duration is taken across utterances by speaker and by question for each speaker. Average phone duration across utterances as a proxy for speech rate has been used in previous work and is termed inverse speaking rate (ISR). Higher numbers indicate slower speech and lower numbers indicate faster speech, which is different from many other speech rate measures. ISR compared with other speech rate measures have also been used in discussions of text-to-speech modeling and their improvement (e.g., [9]).

3. RESULTS

Many speakers vary considerably in the STT WERs and DDMs across different questions, while ISR is more consistent across speakers. RP1 is the only speaker that has a difference of more than 0.003 (3 ms) in ISR across the questions (0.066, 0.076). Table 1 indicates the word error rate (WER) and inverse speaking rate (ISR) for each speaker, and dialect density measures (DDM) for questions by speaker and provides the social contrasts (labeled "SC"; i.e., region and social status) associated with the variety for each speaker. Standard deviations are provided for WER and are calculated from the WER values for each question by the speaker. The highest value in each of the different measures is indicated through bold font and the lowest value is indicated through underlining. Recall that high values indicate poorer performance for WER, slower speech for ISR, and more accent feature use for DDMs. Therefore, according to our predictions based on previous work, the best SST performance (i.e., lowest WER) should occur with higher ISR, lower DDMs, and non-marginalised varieties. Note that marginalised speakers are listed at the top of table and non-marginalised are at the bottom, and within each group northern speakers are listed first.

ID	SC	WER(sd)	DDM	ISR
UWYE1	NM	0.09(0.03)	0.88,**1.40**	**0.072**
UWYE2	NM	0.10(0.05)	0.23, 1.36	**0.072**
EE1	SoM	0.22(0.10)	-0.14,0.50	0.067
EE2	SoM	0.12(0.05)	-0.37,-0.28	0.069
MLE1	SoM	0.15(0.07)	-0.81,-0.15	0.064
MLE2	SoM	**0.31**(0.11)	0.94,1.17	0.062
GNE1	NSt	0.11(0.04)	-0.87,0.33	0.071
GNE2	NSt	0.18(0.08)	0.75,1.0	0.063
RP1	SoSt	0.16(0.06)	-1.71,-0.17	0.070
RP2	SoSt	0.18(0.12)	-1.71	0.068

Table 1: Summary of word error rates, scaled dialect density measure, and inverse speaking rate for each speaker ("ID"). "SC" indicates main social contrasts for each variety and is coded as So=southern, N=northern, St=non-marginalised, M=marginalised. Col. 3 has speaker word error rate (WER) with standard deviation in brackets (higher=worse & lower=better performance). Dialect density measures (DDM) for each question (higher = more & lower = fewer accent features used) is in col. 4. Average speaker inverse speaking rate (ISR) is provided in col. 5 (higher = slower & lower = faster speech rate). Bold font = highest and underlined = lowest WER, DDM and ISR.

Speech rate has the strongest effect of all the variables under consideration (see Figure 1). UWYE1, UWYE2, GNE1 are the speakers with consistently slower speech rates and they also have the lowest WERs. Therefore, they obtain the best STT performance. On the other hand, MLE2, the speaker with the fastest speech rate, obtains the worst STT performance.

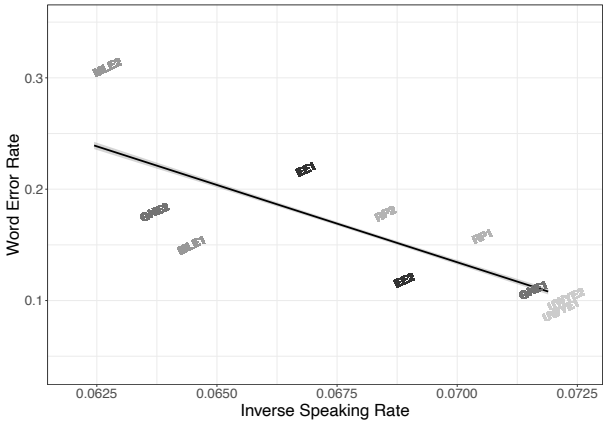


Figure 1: Speech-to-text word error rate plotted against average inverse speaking rate (ISR) for each speaker using the ggplot lm smooth function. Speaker labels represent exact values.

Variety does not appear to have a clear effect on STT efficacy (see Figure 2). The non-marginalised varieties do not have better STT performance than marginalised varieties overall. UWYE speakers, who speak with a northern marginalised variety, have the lowest WER (i.e., best performance) and RP speakers, who speak with a southern non-marginalised variety have the 3rd and 4th highest WERs (i.e., relatively poor performance). In the case of the UWYE speakers, these are also the speakers with the slowest speech rates, which may partially explain this result. However, the RP speakers have slower speech rates as well and have worse STT performance than speakers with similar speech rates, so it cannot be entirely due to speech rate effects. For example, RP2 has a similar ISR to both EE speakers, but has a much higher WER than EE2 and lower one than EE1.

Differences across varieties only seem to appear when individual repertoires are considered. This can be seen through the inconsistent STT performance for speakers of the same variety, who have large differences when comparing their DDMs to each other (see Table 1). For example, EE1 and EE2 both speak EE, which is a southern marginalised variety, and have similar speech rates. However, EE1 has a relatively high WER (i.e., poorer performance) and

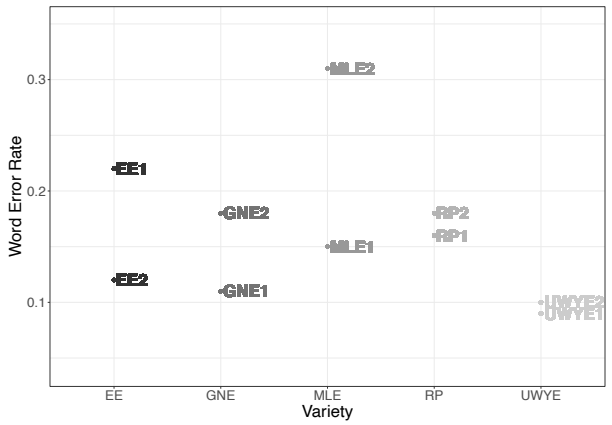


Figure 2: Speech-to-text word error rates (WER) for each speaker plotted by variety. Speaker labels indicate WER for that speaker within the variety.

EE2 does not. If we take the individual repertoires into account then we see that EE1 uses more accent features than EE2. A similar finding occurs when comparing the MLE speakers, who have similar speech rates, but differ in their accent feature use and have substantially different WERs. Again the speaker who uses fewer accent features (MLE1) has better STT performance.

Considering the non-marginalised varieties (GNE and RP), it is difficult to tease apart the effects of individual repertoire compared with speech rate, as these often point in the same direction. The RP speakers have similar speaking rates and use of accent features and also have similar WERs. On the other hand, STT performance differs for the GNE speakers, but this aligns with both the DDM and ISR results. GNE1 who has better performance also has a much slower speech rate and uses fewer accent features than GNE2.

4. DISCUSSION

Previous work has indicated that non-marginalised North American varieties have better STT performance than marginalised ones, but this does not appear to be straightforwardly the case when considering these British English varieties. We do find the worst STT performance for a speaker of a marginalised variety (MLE2), but the best STT performance also occurred for speakers of a marginalised variety (UWYE). This result seems to be, at least, partly due to the difference in speech rate, as the UWYE speakers also have the slowest speech rate. Furthermore, it seems that at the slowest speech rates differences in accent feature use does not equate to differences in STT performance. The UWYE speakers also have the highest DDMs. Non-marginalised varieties, such as RP, do not have very low WERs despite both speakers having relatively slow speech rates and the lowest accent feature use.

Differences in the individual repertoires of speakers within an accent are able to account for some of the differences in STT performance that we find and this occurs most obviously for EE and MLE, which are southern marginalised varieties. Speakers who use more accent features (EE1, MLE2) also have worse STT performance. These performance differences do not seem to be related to speech rate differences as both of the MLE speakers and both of the EE speakers have similar speaking rates compared to each other. Therefore, individual repertoires rather than variety has a more substantial effect on the STT performance in the direction that was predicted.

Overall speech rate is found to strongly affect WER, but again this is not found consistently across all speakers and varieties. The three speakers with the slowest speech rates (UWYE1, UWYE2, and GNE1) also have the best STT performance and the speaker with the fastest speech rate (MLE2) has the worst STT performance. Other than these speakers, we need to consider more than just speech rate to understand the differences in STT performance.

While more work on STT performance for a wider range of speakers of different varieties needs to be done, our findings suggest a complicated type of accent bias might occur in training and language models for STT systems. Even when marginalised varieties may be included and those varieties or speakers use slower speech rates, there may continue to be unequal outcomes for speakers of marginalised varieties who use more accent features. Building on the suggestions from other work, such as [3, 1], more diverse datasets are required when training these speech recognition systems. These datasets need to be more inclusive and should represent multiple varieties (marginalised and non-marginalised) *and* speakers with a wide range of individual repertoires within these varieties. Only then it may be possible to obtain more equal STT performance for everyone and limit the effects of social biases.

5. REFERENCES

[1] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/

W17-1606
[2] S. Oviatt, M. MacEachern, and G.-A. Levow, "Predicting hyperarticulate speech during human-computer error resolution," *Speech Communication*, vol. 24, no. 2, pp. 87–110, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639398000053
[3] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117
[4] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, 2017, pp. 934–938.
[5] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007, intrinsic Speech Variations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000404
[6] *Journal of English Linguistics*, 2021.
[7] J. Cheshire, P. Kerswill, S. Fox, and E. Torgersen, "Contact, the feature pool and the speech community: The emergence of multicultural london english," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 151–196, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9841.2011.00478.x
[8] *Language and Speech*, in preparation.
[9] G.-T. Liou, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, "Estimation of hidden speaking rate," in *9th International Conference on Speech Prosody*, 2018, pp. 592–596.

Quantifying Prosodic Variation in AAE as a Function of Gender and Interlocutor Using the PoLaR Framework

Nicole R. Holliday

Department of Linguistics and Cognitive Science, Pomona College
nicole.holliday@pomona.edu

ABSTRACT

While African American English (AAE) is well-documented, descriptions of its prosody remain scarce. This study is the first phonetic paper to describe the prosody of AAE speakers using the PoLaR framework, which represent a flexible framework for describing prosodic variation. This study examines data from 29 AAE-speaking students with attention to variation by speaker gender and interlocutor. Data was annotated and extracted using PoLaR and a series of regression models then tested differences between speaker groups by number of points, level and range changes. Results indicate female speakers use wider pitch ranges and more dramatic local F0 changes than male speakers. Additionally, speakers are sensitive to interlocutor; those who interacted with an AAE-speaking interlocutor used greater level changes, more points, and wider pitch ranges. These results demonstrate the utility of PoLaR for describing prosodic variation in English, and document the role of gender and interlocutor in prosodic variation in AAE.

Index Terms: prosody, sociophonetics, African American English, ethnolinguistic variation, PoLaR

1. INTRODUCTION

Though African American English is extremely well-documented in the sociolinguistic literature, descriptions of its prosody remain scarce. Previous works have explored the variety's pitch accent inventory and distribution as well as boundary tones using MAE-ToBI annotations based on the AM framework, without attention to the fact that these conventions were not designed to accommodate variation in American English [1, 2, 3] (see section 1.1). The current study is the first phonetic paper to describe the prosody of AAE speakers using the PoLaR conventions developed by Ahn et al. 2021 [1], which may represent a more flexible framework for detailing such variation. This study is an examination of reading data from 29 AAE-speaking college students with special attention to variation conditioned by speaker gender and interlocutor. The aims of the paper are to test a new methodology for describing the prosodic system of a non-mainstream variety of American English, taking into account phonetic variables, as well as to better describe sociolinguistic factors that may contribute to intra-variety differences within AAE. With respect to testing the utility of PoLaR for describing prosodic differences within and across varieties, the current study represents a step forward in using phonetic methods that are more independent from phonological assumptions in order to describe intonational systems. Such work will be increasingly necessary for linguists hoping to expand our knowledge of how prosody works in underdescribed languages and dialects across the world, without having to rely on intonational phonological descriptions that either may not exist,or were not intended to capture the rich variation present in naturalistic speech.

This work is also especially of interest due to the fact that not only is AAE prosody poorly understood, but also because it appears to be especially perceptually salient for listeners [2, 4], thus making it a key element for dialect identification. As a result, understanding the prosody of AAE has important utility for better addressing linguistic profiling and discrimination, developing more fair speech technologies, and understanding more broadly the realm of possible prosodic differences between American English varieties.

1.1. Prosodic Variation in MAE and AAE

Previous studies on prosodic variation in AAE have observed a number of general properties that may differentiate it from MAE. In general, studies have claimed that AAE may use a greater degree of macro-rhythm; that is, more local F0 alternations [2, 3]. This may be quantified as more (bitonal) pitch accents in the MAE-ToBI framework, as claimed by [3, 5, 6]. A number of previous works have also observed that AAE-speakers may employ a wider pitch range and/or greater use of falsetto phonation [7, 8]. While many of these previous works have been purely descriptive, they provide motivation for testing such claims quantitatively. The current study does not directly compare AAE and MAE speakers, but rather aims to provide a description of these aspects of AAE prosody as well as test for

potential effects of gender and interlocutor that may help to document such differences in ways that will lay the groundwork for future projects.

2. METHODS

One challenge for studying prosodic variation in English is that studies are still limited by existing methods, nearly all of which were designed for Mainstream American English [2]. Recent advances in both phonetics and intonational phonology, however, have prompted the development of new methods for studying prosodic variation. While it is still the case that the majority of studies in the U.S. have coded data using the MAE-ToBI conventions developed by Beckman and Ayers-Elam [9], these conventions are specifically designed for MAE, without attention to the potential for systematic variation. This is a particular limitation because ToBI systems are intentionally phonetically underspecified and intended to capture phonological categories of MAE, which may not apply for speakers of other varieties.

The Points, Levels and Ranges (PoLaR) annotation system designed by Ahn et al. [1] may be more appropriate for describing the prosody of a broader range of linguistic varieties. PoLaR uses four ties to capture phonetic detail, which is an important feature that distinguishes it from ToBi. The tiers used in PoLaR are as follows:

1. *Prosodic Structure*: Prominences and Boundaries
2. *Points:* Moments where F0 turns occur
3. *Levels*: Locally-specified F0 movements, within a given IP. Corresponds to marks on the Points Tier, and is derived from Point and Range Info
4. *Ranges*: How much does F0 change within a given IP?

Figure 1 shows an example of PoLaR coding from the current dataset.

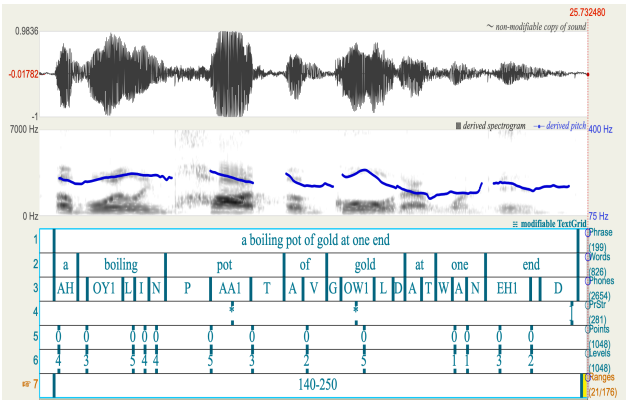


Figure 1: Tier 1 (Phrase) is the phrase's orthography. Tier 2 (Word) shows individual words. Tier 3 (Phone) shows individual phones. Tier 4 (PrStr) is the Prosodic Structure tier, roughly corresponding to pitch accents and boundaries in MAE-ToBI. Tier 5 (Points) shows points, or each instance of a phonologically meaningful F0 turn. Tier 6 (Levels) takes the points from Tier 5 and gives them a locally-determined rating of pitch level, based on the global F0 range of the phrase. Finally, Tier 7 (Range) shows the F0 range (max-min, measured in Hz) of the phrase.

2.1. Data Collection and Coding

The dataset consists of 29 recordings of self-identified AAE speakers (19 female, 10 male) reading the Rainbow Passage aloud [10]. Participants were recorded as a part of a larger study of the use of AAE on college campuses [11]. That study captured thorough demographic information about participants, but also utilized a unique design such that half of the participants were interviewed by a white, non-AAE speaking interlocutor and half were interviewed by a Black, AAE-speaking interlocutor. This design allows for the current study to test not only the effects of personal demographic categories such as gender, but also possible interlocutor effects.

Individual Intonational Phrases (IPs) were identified perceptually and annotated manually, and IPs with disfluencies or background noise were excluded, ultimately yielding 3764 IPs for analysis. IPs were then force-aligned and subsequently annotated using the PoLaR conventions developed by Ahn et al. [1] and data was extracted using a series of scripts that accompany the Praat PoLaR plugin. The relevant data is number of Points, local F0 changes within the phrase (Levels), and F0 range differences (Range) between phrases for each speaker.

2.2. Statistical Analysis

A series of logistic and linear mixed effects regression models in R using the lme4 package [12] tested differences between speaker groups related to Points, Levels, and Ranges. For comparisons related to number of points, a linear mixed effects regression (LMER) model tested for differences by gender, interlocutor, and the interaction with a random effect of speaker. Using the same fixed effects of gender and interlocutor with a random speaker effect, the model for Levels tested differences between adjacent level labels within the same IP in order to examine variation in how speakers use larger or smaller local jumps within their individual F0 range. Finally, again using fixed effects of gender and interlocutor, the model for Ranges tested differences in F0 max-min across IPs for speakers.

3. RESULTS

3.1. Points

Results for the model for number of points per second of speech by gender and interlocutor indicate that female speakers use a greater number of points than male speakers overall (Est=0.186, p<.05). The model also reveals that speakers of both genders use a greater number of points when interacting with the AAE-speaking interlocutor, though this difference is not significant. We do not observe an interaction effect between the gender and interlocutor. While Figure 2 shows the mean results for each speaker group, the lack of an interaction effect may obscure the main results.

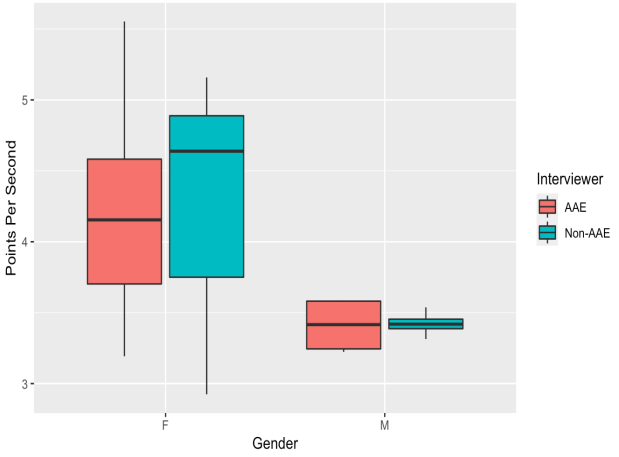


Figure 2: Boxplot for number of points by gender and interlocutor.

3.2. Ranges

Overall, female speakers use wider pitch ranges across IPs than male speakers (Est=31.878, p<.001), as do speakers interacting with the AAE-speaking interlocutor, though the effect is somewhat smaller (Est=3.423, p<.01). Again, we observe no interaction between gender and interlocutor.

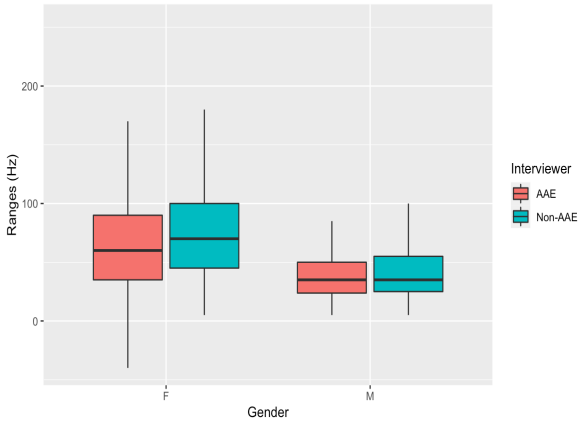


Figure 3: Boxplot for range by gender and interlocutor.

3.3. Levels

As this paper is among the first to conduct a quantitative analysis of the Levels differences coded in the PoLaR system, it will report on speakers' proportion of use of each Level, which may indicate how much time they spend in the middle versus towards the extreme ends of their F0 ranges for any given IP. In this framework, we observe that female speakers also use more dramatic local pitch differences between points, employing a greater number of 1-5 and 5-1 level jumps than male speakers (Est=0.228, p<.05). However, in contrast to the results for Points and Ranges, we do not observe interlocutor or interaction effects for this variable. Figure 4 shows the distribution of the use of each Level by speaker and interlocutor in order to show the general distribution of Levels. While this paper focuses only on the most dramatic level jumps as an illustration, future work should also examine whether there are systematic differences between speakers or varieties for other types of level jumps.

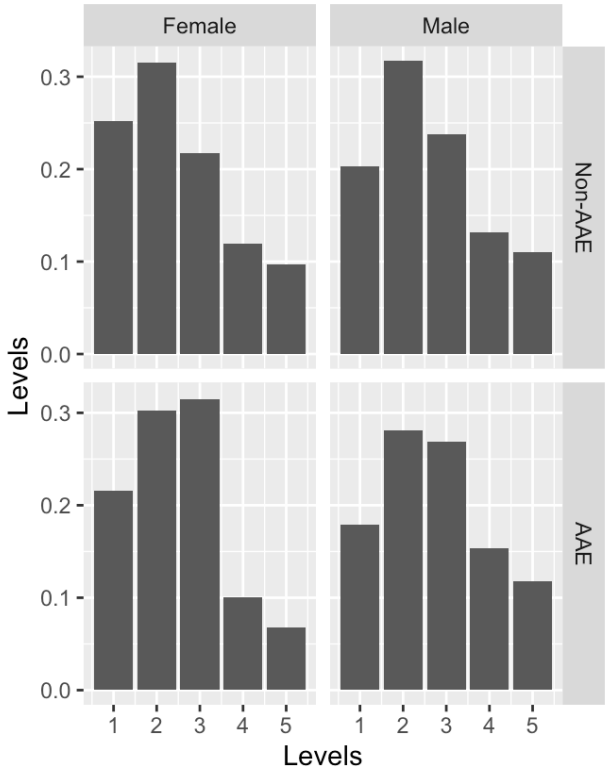


Figure 4: Proportion of use of each Level by Gender and Interviewer.

4. DISCUSSION AND CONCLUSION

In this study of the use of prosodic variation using Points, Levels, and Ranges among AAE-speaking college students, female speakers use more points and wider pitch ranges per phrase, and show more instances of dramatic level changes than male speakers. Additionally, speakers appear to be sensitive to interlocutor-based differences; speakers who interacted with an AAE-speaking interlocutor generally used more points and wider pitch ranges than those that interacted with a non-AAE speaking interlocutor. Interestingly, however, we see limited interaction effects between gender and interlocutor for any of the variables tested, indicating that perhaps effects on the use of these variables operate somewhat independently of one another. These results support the claims of previous works that assert that AAE generally has a strong macrorhythm, or more local F0 changes that manifest as simple changes, jumps between adjacent points, and differences in the use of F0 range between phrases in the same passage. The strong effects for both gender and interlocutor across phonetic variables also shed light on some social factors that may affect systematic prosodic variation in AAE. Previous work such as Li et al. [13] has documented gender differences in the use of F0 range between female and male speakers in both MAE and AAE, so the results obtained here strengthen claims that female AAE speakers generally employ a wider pitch range than their male counterparts. The findings for Levels also support this claim; greater local pitch jumps may also facilitate the use of a wider F0 range. Finally, the larger number of points used by female speakers and those interacting with an AAE-speaking interlocutor support findings such as those by McLarty [3] who observed that AAE speakers generally employed a larger number of pitch accents, though in this case, the larger number of points may be indicative of speaker accommodation to the AAE-speaking interlocutor.

Finally, these results also demonstrate the utility of the PoLaR conventions for underdescribed varieties of English. Earlier works on AAE were limited by the fact that they had to describe the variety's prosody based on assumptions that were designed for MAE speakers. This study has shown that the PoLaR annotation conventions can be successfully used in sociophonetic analysis and descriptions of prosodic variation across prosodic systems that may or may not conform to the assumptions inherent in the MAE-ToBI conventions. Future work can build on these results to arrive at a better description not only of AAE prosody, but also of other marginalized varieties of English, as well as other languages.

5. REFERENCES

[1] Ahn, Byron, Nanette Veilleux, Stefanie Shattuck-Hufnagel, and Alejna Brugos. 2021. PoLaR Annotation Guidelines (version 1.0). doi: 10.17605/OSF.IO/USBX5.

[2] Thomas, E. R. Prosodic features of African American English. In S. Lanehart (Ed.), *The Oxford handbook of African American Language*, 2015, pp. 420–438. Oxford: Oxford University Press.

[3] McLarty, J.. African American Language and European American English Intonation Variation Over Time in The American South. *American Speech: A Quarterly of Linguistic Usage*, vol. 93, no. 1, pp. 32–78, 2018.

[4] Purnell, Thomas, William Idsardi, and John Baugh. "Perceptual and Phonetic Experiments on American English Dialect Identification." *Journal of Language and Social Psychology* 18: 10–30, 1999.

[5] Holliday, N.. *Intonational variation, linguistic style, and the black/biracial experience*. PhD dissertation, New York University, 2016.

[6] Burdin, R.S., Holliday, N.R. and Reed, P.E.. American English pitch accents in variation: Pushing the

boundaries of mainstream American English-ToBI conventions. *Journal of Phonetics*, *94*, p.101163, 2022.
[7] Tarone, E. E.. "Aspects of intonation in Black English". *American Speech*, vol. 48, 29–36, 1973.
[8] Nielsen, Rasmus. *Reassembling Ethnicity: Stylistic Variation in African American English Prosody*. Ph.D. dissertation, Georgetown University, Washington, D.C., 2012.
[9] Beckman M.E. and G. Ayers Elam, "Guidelines for ToBI labelling (Ver. 3)," unpublished mw., The Ohio State University, 1997.
[10] Fairbanks, G.. The Rainbow Passage. From Voice and articulation drillbook, 2nd Edn. New York: Harper & Row, 1960.
[11] Holliday, N.R. and Squires, L.. Sociolinguistic labor, linguistic climate, and race (ism) on campus: Black college students' experiences with language at predominantly white institutions. *Journal of Sociolinguistics*, *25*(3), pp.418–437, 2021.
[12] Bates D., Mächler M., Bolker B., Walker S.. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1), 1–48, 2015.
[13] Li, A., Purse, R. and Holliday, N.. Variation in global and intonational pitch settings among black and white speakers of Southern American English. *The Journal of the Acoustical Society of America*, *152*(5), pp.2617–2628, 2022.

REGIONAL AND INDIVIDUAL VARIATION IN ACOUSTIC TARGETS OF /AI/ and /AU/ IN AMERICAN ENGLISH

Irina Shport[1], Marie Bissell[2], Kelly Berkson[3], Katie Carmichael[4]

[1]Department of English, Louisiana State University; [2]Department of Linguistics, The Ohio State University; [3]Department of Linguistics, Indiana University; [4]Department of English, Virginia Tech, United States
ishport@lsu.edu, bissell.43@osu.edu, kberkson@iu.edu, katcarm@vt.edu

ABSTRACT

English diphthongs are represented as bisegmental (e.g., /aɪ/) and assumed to have two acoustic targets (onset, offset). These phones are standardly represented with variable symbology (e.g., /ai/, /aj/), and previous work indeed reports variability in diphthong offsets. To investigate whether variation can be explained by dialectal and/or individual differences, we examined diphthongs (/aɪ, aʊ/) produced by 41 speakers of American English from Ohio (Midland, Northern regions) and Louisiana (Southern region).

Comparison of offset spectral estimates to nearby monophthongs indicate that the /aɪ/ offset was relatively consistently acoustically close to [ɪ], but the /aʊ/ offset was highly variable for both groups. While some of these findings, such as the degree of dynamic spectral change in diphthongs, can be explained by dialectal differences, it is also possible that the diphthongs have different underlying structure (e.g., more clearly biphasic onset-offset for /aɪ/ than for /aʊ/).

Keywords: diphthongs, nearby monophthongs, positional relationship, cross-dialectal differences.

1. INTRODUCTION

This study examined spectral trajectories of two American English diphthongs, /aɪ/ and /aʊ/. In speech sciences, three sets of notations are used to represent these phones. The offglide targets are variably represented as glides, lax high vowels, or tense high vowels [1]. These three sets of notations – /aj/-/aw/, /aɪ/-/aʊ/, /ai/-/au/ – bear inherent theoretical assumptions about the underlying targets of diphthong offglides, which are rarely critically examined in context of variation in vowels systems across English dialects.

Herein, we investigate whether phonological representation of the offsets as high vowels is (a) accurate for both diphthongs and (b) supported by robust acoustic data in word productions by English native speakers from two dialectal regions. Diphthongs are often assumed to have two steady-state targets as their onset and offset [2]. Yet the single study (to our knowledge) that has investigated what acoustic properties of English diphthongs are most relevant to their classification tested three models – onset plus offset, onset plus slope, onset plus direction – and yielded correct classifications of a diphthong corpus exceeding 90% accuracy [3]. Further, a study comparing the English and German equivalents of these two diphthongs suggested that they may have different onsets in English, concluding that /a/ accurately represents the onset of /aɪ/ but not /aʊ/ [4]. Both these works, though very different, raise questions about how best to represent the diphthongs.

The use of the IPA phonetic symbols for lax vowels in English diphthongs may be traced back to Daniel Jones (1881-1967) [1, footnote 1]. The /aɪ/ and /aʊ/ notations took deep root in the linguistic community and are commonly used by students in phonetics classes and by researchers alike (also, in this paper). Research findings, however, have showed that the acoustic targets of the diphthongs can approximate a variety of phones for both onset (/æ/, /a/, /ɑ/, or /ʌ/) and offset (/i/, /ɪ/, /ɛ/, /u/, /ɔ/, or /ʊ/), although they are more consistent for the former than the latter [2-3, 5-7]. For this reason, we chose to focus on offsets and, specifically, on the positional relationship between the estimates of offset targets and nearby monophthongs.

In sociolinguistic tradition, ARPABET phonetic symbols AY and AW are often used for vowels in *bite* and *bout* [8], representing the offset targets as underlying glides. Recent work suggests that while high vowels /i, u/ and their glide counterparts /j, w/ may belong to distinct phonemic categories, categorical distinction in their phonetic properties is not so much in the degree of constriction (as reflected in formant values), but in their temporal organization (as reflected in relative timing of articulatory gestures) [9]. Based on these findings and our focus, we excluded the /aj/ and /aw/ representations from consideration here.

In research so far, data were often collected in Midwestern states and might have been taken to represent General American English [3, 7]. Following [6], we contribute cross-dialectal data collected in a variety of phonetic contexts to this line of research. Because the phonological voicing status of the

following coda was shown to significantly alter the duration and spectral properties of vowels (e.g., American English in [10-11]; Australian English in [11]; Canadian raising in [13-15]) we examined formant values of diphthongs when followed by voiced and voiceless segments. Because speech in Southern U.S. communities is often characterized by /aɪ/-monophthongization and resistance to back vowel fronting which would affect diphthong offglides, participants were recruited in the state of Louisiana to represent the Southern region and in the state of Ohio to represent the Midland and Northern regions [16]. In what follows, we examine whether offglides approximate tense or lax high vowels (or other vowels), and whether speaker dialect or voicing context can explain offset variability in the target diphthongs.

2. METHODS

2.1. Data collection

Data was collected as a part of a larger study in which the two diphthongs were elicited in pre-voiced and pre-voiceless contexts (Table 1). Words with /aɪ/ totalled 52 and words with /aʊ/ totalled 35. In addition, thirteen monophthong vowels /iː, ɪ, eː, ɛ, æ, ə, ɜː, ɑ, ɔ, ʌ, oː, ʊ, u/ were elicited in b_t and h_d English word contexts (e.g., *beat/heed, bat/had*).

Target	Pre-Voiced Context	Pre-Voiceless Context
/aɪ/	*prize, cyber, rider, siding, vibration*	*price, viper, writer, biting, citation*
/aʊ/	*browse, cows, cloudy, however*	*house, couches, pouting, outlandish*

Table 1: Example stimulus words.

Participants were recruited in Columbus, Ohio (18 speakers, 14 females) and Baton Rouge, Louisiana (23 speakers, 15 females). They recorded themselves reading stimulus words one at a time in an online Qualtrics survey (words were randomized and presented three times each) and emailed the recordings to the investigators.

2.2. Analyses

All data was force-aligned using the MFA aligner [17]; then, segmentation was manually hand-checked in Praat [18]. A Praat script was used to extract first and second formant (F1, F2) values at 21 equidistant point from the onset of each vowel token, at 5%-increment into vowel duration. Next, we describe a set of analyses conducted for each speaker's data.

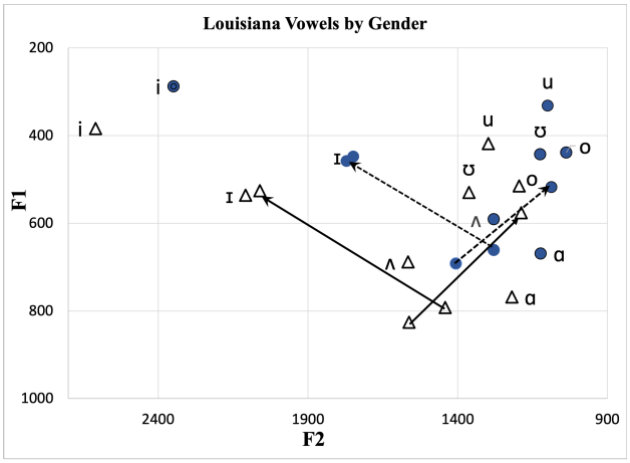
To provide a single estimate for each monophthong vowel (Mono), means of median values across 17 time points (i.e., 90% of vowel duration to exclude coarticulatory effects) were calculated [12]. To provide an estimate for the onset and offset of each diphthong (Di.on and Di.off), means of median values across the 3-5th and 17-19th time points (i.e., the 10-20% and 80-90% of vowel duration) were calculated [3]. After initial observations of individual F1/F2 plots, ten Euclidean distances (ED) between Mono and Di.off were calculated per speaker to estimate the distance between the /aɪ/'s offset and /i, ɪ, ʌ, æ, ɑ/ and between the /aʊ/'s offset and /u, o, ʌ, ʊ, ɑ/ (ED types). The formula in (1) was used to calculate EDs and served as a speaker-intrinsic normalization technique.

(1) $$ED_{/\text{Mono}/-/\text{Di.off}/} = \sqrt{(F1_{/\text{Mono}/} - F1_{/\text{Di.off}/})^2 + (F2_{/\text{Mono}/} - F2_{/\text{Di.off}/})^2}.$$

Mixed modeling was used for statistical analyses of these ED values, five ED types per diphthong.[1]

3. RESULTS

3.1. Variability in positional relationships



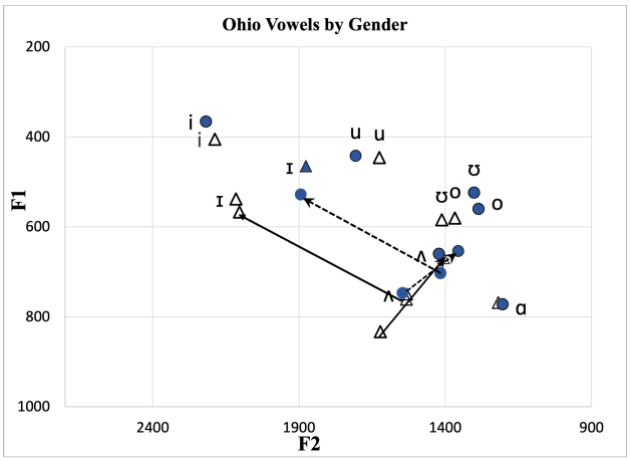


Figure 1: Mean vowel characteristics for two regions (females: open triangles, solid lines; males: filled circles, dashed lines). Monophthongs are represented by singular F1/F2 datapoints; diphthongs are represented by lines connecting the Di.on and Di.off estimates.

Figure 1 illustrates some regional differences in the vowel systems, such as [u] and [o] fronting (a relatively high F2) and [ʊ] and [o] lowering (a relatively high F1) in Ohio as compared to Louisiana. The degree of spectral change in diphthongs appears to be smaller in Ohio than Louisiana, especially for /aʊ/. In both plots, the offset of /aɪ/ is approximating [ɪ] the most. The offset target of /aʊ/ is less apparent: likely, [o ʊ] for Louisiana speakers, and [o ʊ ʌ] for Ohio speakers.

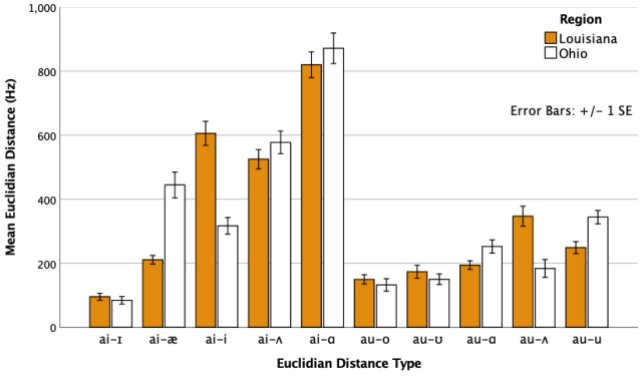


Figure 2: Summary of offglide positional relationship to nearby monophthongs (ten ED types) by region. Small ED values indicate relative proximity of diphthong offsets to monophthongs.

Generalized linear mixed modeling (GLMM) on ED values with region (2) and nearby monophthongs (5) as fixed factors showed a significant interaction between region and nearby monophthong for /aɪ/, $F(4,195) = 25.11$, $p < .001$, and for /aʊ/, $F(4,195) = 13.32$, $p < .001$. Figure 2 illustrates this interaction. For /aɪ/, the largest regional difference is in the offset proximity to [i] and [æ] ($t = 5.60$, $p < 0.001$, $t = 6.05$, $p < 0.001$). For /aʊ/, the largest regional difference is in the offset proximity to [u] and [ʌ] ($t = 3.39$, $p = 0.002$, $t = 3.79$, $p < 0.001$).

To examine individual variability in offset category membership, the numbers of speakers exhibiting specific primary and secondary proximal positions (i.e., the smallest and the second smallest EDs per diphthong) are summarized in Table 2. This summary confirms that for /aɪ/, the primary offset target is robust – the lax /ɪ/, with secondary proximal positions varied by region. For /aʊ/, the primary and secondary proximal positions largely varied.

Trends	LA	OH
aɪ - ɪ (i)	-	14 speakers
aɪ - ɪ (æ)	22 speakers	4 speakers
aɪ - æ (ɪ)	1 speaker	-
aɪ - ɪ (ʌ)	1 speaker	-
aʊ - o	12 speakers	7 speakers
aʊ - ʊ	7 speakers	2 speakers
aʊ - ʌ	-	8 speakers
aʊ - ɑ	4 speakers	-
aʊ - u	-	1 speaker

Table 2: A summary of speakers by absolute proximity of their diphthongs' offsets to selected monophthongs (i.e., ED types). The smallest ED was taken as the closest match. The second closest matches are in parentheses for the /aɪ/ offset but not for /aʊ/ because their number exceeded ten.

3.2. Voicing context

We also examined whether voicing of the following segment influences the acoustic realization of diphthong offglides, specifically, their F2 values at onset and offset.

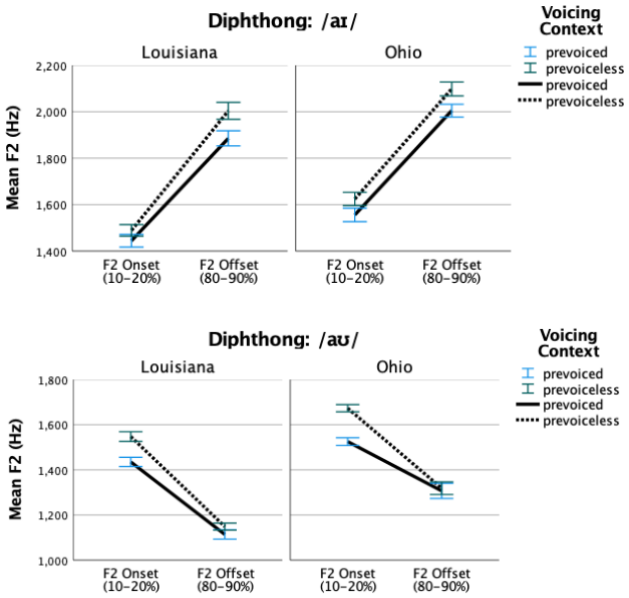


Figure 3: F2 estimates at diphthong onsets and offsets by region and the following voicing context.

GLLM analyses of F2 estimates with target (onset, offset), region (LA, OH), and voicing (pre-voiced, pre-voiceless) as fixed factors yielded a significant effect of voicing (/aɪ/: $F(1,322) = 32.83$, $p < .001$; /aʊ/: $F(1,322) = 46.07$, $p < .001$) and region (/aɪ/: $F(1,322) = 5.70$, $p = .018$; /aʊ/: $F(1,322) = 19.38$, $p < .001$). The interactions between voicing and region were not significant in either diphthong model. The interaction between voicing and target was significant for /aʊ/ only, $F(1,322) = 22.38$, $p < .001$. Figure 3 illustrates these results: F2 was higher in pre-

voiceless than pre-voiced contexts except for /aʊ/ offsets, and in productions of Ohioans than Louisianans.

4. DISCUSSION

This study addressed the question of whether acoustic data allow us to infer robust phonological offglide targets in productions of two diphthongs, and if so, what those targets are. We found that across two dialectal regions of North American English, the offset of /aɪ/ consistently approximated the monophthongal /ɪ/, although the secondary trend was for this offset to approximate [i] for Ohioans and [æ] for Louisianans. The latter secondary trend may be explained by /aɪ/ monophthongization, a still frequently observed characteristic of Southern speech where /aɪ/ does not have much of an offglide, with relatively small dynamic spectral change across the diphthong, as illustrated in Figure 4. The productions of /aɪ/ were overall more fronted by speakers from Ohio than speakers from Louisiana, and more in pre-voiceless than pre-voiced contexts. It is possible that if /aɪ/ offsets were examined only in a dialect such as Midland / Northern, in words where it is followed by a voiceless segment, a researcher may conclude that the offset target is closer to the tense /i/ rather than the lax /ɪ/. Based on our data, it is reasonable to infer that the underlying form of the diphthong is /aɪ/.

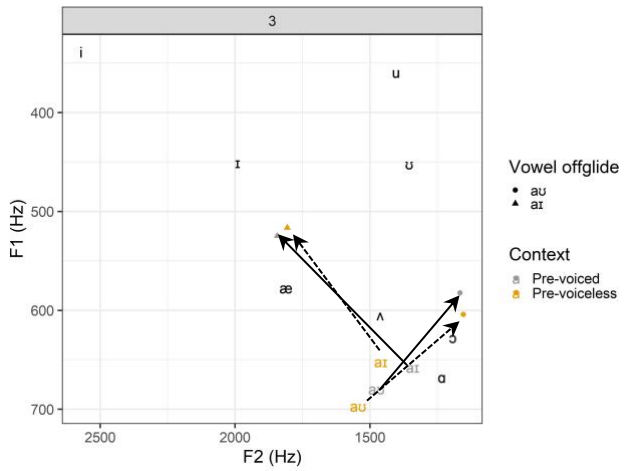


Figure 4: F1/F2 estimates of diphthong onsets and offsets in productions of a male speaker from Louisiana showing relative monophthongization (prevoiced contexts: solid lines; prevoiceless contexts: dashed lines).

It is difficult to make a robust inference for the /aʊ/ offset, however, because of the rampant variability in its acoustic values within individual speakers as well as across regions. Based on group averages (Figure 2) and individual absolute values (Table 2) and under the assumption of phonetics-phonology continuum, /ao/ (or /aɔ/ for varieties without /ɑ/-/ɔ/ merger) could have been proposed as a likely underlying form, but a close competition with /aʊ/, /au/, and in some dialects /aʌ/ casts a doubt that the assumption of two monophthong-like targets [1] would serve us well here. Furthermore, the lack of the effect of voicing in the following segment on /aʊ/'s F2 offset estimates, which is observed elsewhere in our data, strengthens the inference that the offset of this diphthong may not be a phonological target in the same sense as the offset of /aɪ/.

It is possible that these two American English diphthongs have different types of phonological representations, such as "onset plus offset" for /aɪ/ versus "onset plus direction" for /aʊ/ [3]. Such a proposal has, for example, been put forth for Ningbo Chinese, where rising diphthongs were argued to have two static targets while falling diphthongs have just one dynamic target [19]. Perhaps, at least in some varieties of English, front-gliding diphthongs like /aɪ/ might have two phonetic targets whereas back-gliding diphthongs like /aʊ/ may have a dynamic target. This proposal does not solve the issue of the most accurate IPA-style notations for diphthongs that a) do not have their offsets and onsets closely aligning to nearby monophthongs (e.g., significant differences in their positional relationship in the acoustic space) and that b) have dynamic targets that are not easily conveyed by IPA symbols.

One limitation of this study is that diphthongs' offset values are estimated as averages of the median F1/F2 values in the 80-90% portion of a diphthong, which is not an absolute endpoint of a vowel. However, F1/F2 values close to vowels' absolute endpoints are affected by co-articulation and phrasal position [11]. We assumed that the 80-90% portion of a diphthong is representative enough of the terminal acoustic target. After all, in practice, the 20-50-80% timepoints are often used for estimation of the vowel trajectory [20-21]. To check our conclusions, diphthongs in open syllables can be examined in future research, or other statistical methods can be used for the projection of the endpoints of diphthong trajectories (e.g., by estimating the amount and speed of spectral change in the last 10% of vowel duration).

Another limitation is that we took the acoustic-phonetic data to be representative of phonological targets. However, we cannot dismiss a possibility that phonological targets may be obscured by processes such as an speakers' articulatory undershot as in the hypoarticulation theory [22]. Other types of data (e.g., duration and duration ratios, kinematic data, response times in categorization tasks, rhyming trends in poetry) may provide evidence for mental representations that speakers' hold for diphthongs.

5. REFERENCES

[1] Thomas, E. R. 2020. Sociophonetic trends in studies of Southern U.S. English. *J. Acoust. Soc. Am*. 147, 529-540.

[2] Lehiste, I., & Peterson, G. E. 1961. Transitions, glides, and diphthongs. *J. Acoust. Soc. Am*. 33, 268–277.

[3] Gottfried, M., Miller, J. D., Meyer, D. J. 1993. Three approaches to the classification of American English diphthongs. *J. of Phonetics* 21, 205-229.

[4] Raffelsiefen, R., Geumann, A. 2016. AI vs. AU in American English compared to German. In: Draxler, C., Kleber, F. (eds), *Tagungsband: 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*. Ludwig-Maximilians-Universität München, 155–157.

[5] Holbrook, A., Fairbanks, G. 1962. Diphthong formants and their movements. *J. Speech Hear. Res*. 5, 38–58.

[6] Thomas, E. R. 2001. *An acoustic analysis of vowel variation in New World English*. NC Duke UP.

[7] Lee, S., Potamianos, A., Narayanan, S. 2014. Developmental acoustic study of American English diphthongs. *J. Acoust. Soc. Am*. 136, 1880–1894.

[8] Shoup, J.E. 1980. Phonological aspects of speech recognition. In: Lea, W. (ed), *Trends in Speech Recognition*. Prentice Hall, 125-138.

[9] Burgdorf, D. C., Tilsen, S. 2021. Temporal differences between high vowels and glides are more robust than spatial differences. *J. of Phonetics,* 88, 1-47.

[10] Moreton, E. 2004. Realization of the English postvocalic [voice] contrast in F1and F2. *J. of Phonetics* 32, 1–33.

[11] Pycha, A., Dahan, D. 2016. Differences in coda voicing trigger changes in gestural timing: A test case from the American English diphthong /aɪ/. *J. of Phonetics* 56, 15-37.

[12] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am*. 140, 576–581.

[13] Chambers, J. 1973. Canadian Raising. *Canadian Journal of Linguistics* 18, 113-135.

[14] Davis, S., Berkson, K., Eds. 2021. *American Raising.* Publication of the American Dialect Society.

[15] Moreton, E., Thomas, E. 2007. Origins of Canadian Raising in voiceless coda effects: A case study in phonologization. In: Cole, J., Hualde, J. (eds.), *Papers in Laboratory Phonology 9*. Cambridge UP, 37-64.

[16] Labov, W., Ash, S., Boberg, C. (2006). *Atlas of North American English: Phonetics, phonology, and sound change*. New York: Mouton de Gruyter.

[17] McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., Sonderegger, M. 2017-2022. *Montreal Forced Aligner* [computer program].

[18] Boersma, P., Weenink, D. 1992-2022. *Praat: Doing phonetics by computer* [computer program].

[19] Hu, F. 2013. Falling diphthongs have a dynamic target while rising diphthongs have two targets: Acoustics and articulation of the diphthong production in Ningbo Chinese. *J. Acoust. Soc. Am*. 134, 4199.

[20] Fox, R. A., Jacewicz, E. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *J. Acoust. Soc. Am*. 126, 2603–2618.

[21] Farrington, C., Kendall, T., Fridland, V. 2018. Vowel dynamics in the Southern Vowel Shift. *Am. Speech*, 93, 186–222.

[22] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W.J, Marchal, A. (Eds.), *Speech production and speech modelling*. Kluwer, 1–35.

[1] We concur with an anonymous reviewer that GAMMs may be a useful analytical tool for exploring alternative questions related to this topic. For our specific research questions, we prioritized capturing multiple formant dimensions at once.

The emergence of rhotic vowels in Quebec French: a change from below?

Massimo Lipari

McGill University, Canada
massimo.lipari@mail.mcgill.ca

ABSTRACT

Quebec French is undergoing a rarely observed kind of sound change whereby the front mid rounded vowels become rhotic (produced with a bunched tongue or retroflex gesture resulting in low F3, like English /ɚ/). Previous work has suggested that this is a change from below—i.e., one which speakers are not conscious of and which is thus not socially marked. These previous studies, however, have not had the breadth to robustly demonstrate this finding throughout the entire speech community, and the extent of interspeaker variation (especially according to age, gender, and geographic origin) remains largely unknown. This study, which employs a sizeable corpus of parliamentary speech, adds to the documentation of the change, while directly examining the effects of the aforementioned predictors. The findings are consistent with the hypothesis of change from below (with some caveats), although additional data from new speakers is likely needed to draw any strong conclusions.

Keywords: rhotic vowels, change from below, Quebec French, sound change, corpus phonetics.

1. INTRODUCTION

Rhotic vowels, like English /ɚ/, are a rare class of sounds, occurring in less than 1 percent of the world's languages [1]. Their retroflex or bunched tongue articulation creates their signature low third formant [2]. In Quebec French, rhotic-sounding realizations of the mid front rounded vowels /ø/ and /œ/ were first noted by Dumas [3] more than 50 years ago in the speech of a handful of young working-class male Montrealers—particularly in final open syllables and before rhotic codas. This discovery, however, received little attention until a pair of studies by Mielke [4, 5], which found a change in progress also affecting the /œ̃/ phoneme.

The first of these is a corpus study of 75 speakers from the Gatineau-Ottawa region (~170 km west of Montreal, straddling the Quebec-Ontario border) [4]. A significant fall in F3 was found (in apparent time) for each of the three vowels, beginning with speakers born in the 1960s—the drop occurs slightly earlier and steeper for /ø/ than for the others. The effects of gender and socioeconomic class were not significant.

A follow-up production study with 23 speakers from the same region (save for a few) [4, 5] then showed, via ultrasound imaging, that there exists a high degree of variability in how the mid front rounded vowels are realized articulatorily, both between and within speakers. That is, varying degrees of tongue bunching (the much more commonly used gesture compared to retroflexion) were observed, with tokens lying on a continuum from not bunched (fully non-rhotic) to English /ɚ/-like (fully rhotic).

The non-significance of the effects of the social predictors leads Mielke to reject Dumas' (somewhat speculative) suggestion that rhotacization results from deliberate borrowing from English for stylistic or prestige reasons.[1] Instead, he suggests it is a change from below: a kind of change which speakers are not conscious of during its progression, and which is accordingly not socially marked (cf. [6]). He also cites in favour of his view anecdotal evidence of native speakers generally failing to perceive a salient difference between rhotic and non-rhotic productions of the vowels when asked directly. Mielke proposes a perceptual motivation for the change, noting that rhotacization may serve to enhance the acoustic cues produced by lip rounding.[2] Rhotacization would thus serve to increase perceptual distance between the front mid rounded vowels and their unrounded counterparts /e, ɛ, ɛ̃/ in a crowded vowel space.[3]

While the theory of change from below is certainly consistent with Mielke's results, the studies have some limitations. For one, the model reporting is sparse. In particular, nothing is said about the random effects: those results, however, could be illuminating, given the finding of much inter-speaker variability. For another, the corpus used is geographically circumscribed. Since we know from Dumas that rhotacization is present in other regions of Quebec, data from these other areas would allow us to test how robustly the theory of change from below holds up across the speech community.

The research questions of the present study are thus twofold. The first set is descriptive: does the state of rhotacization look similar throughout Quebec to the situation in the Gatineau-Ottawa region? The second aims to determine whether the evidence from other dialects is consistent with a change from below. Were this the case, we may expect to find no effect of gender, and either no effect of dialect or perhaps some evidence of gradual, wave-like diffusion throughout

the territory (since a change from below would not be consciously associated with any particular group).

2. DATA & METHODS

The data used in the present study are from a combination of two corpora of proceedings of the National Assembly of Quebec (ANQ), Quebec's national parliament. The first corpus was developed by Milne [9]: it consists of approximately 9 hours of recordings from 61 parliamentarians,[4] which was previously force-aligned with purpose-built software. The other is a new corpus in the same style, developed for this study, containing almost 7 hours of additional data from 26 speakers (including some of the same speakers as Milne's) and aligned using the *Montreal Forced Aligner* [10]. Neither corpus originally contained biographical data (other than speaker names): gender, year of birth and place of birth were thus obtained manually from the ANQ website [11]. The first two predictors were added as-is, whereas place of birth was used to construct a ternary DIALECT predictor. The traditional assumption that Quebec is roughly divided into two dialect regions, West and East, by a line perpendicular to the Saint Lawrence River somewhere between Montreal and Quebec City (cf. the discussion in [12] on the historic apical vs. dorsal rhotic isogloss) was followed, and speakers in the ambiguous region between those cities were coded as speaking a "Central" dialect.[5] Non-native speakers and speakers with missing biographical data were excluded. The remaining 70 speakers are distributed by gender and dialect per Table 1.

	West	Central	East	**Total**
Female	11	4	8	**23**
Male	22	5	20	**47**
Total	**33**	**9**	**28**	**70**

Table 1: Breakdown of speakers by gender/dialect.

Formants were measured at regular intervals using the "refined" formant analysis function of the Python package *PolyglotDB* [13]. This algorithm, described in [14], iteratively searches for the optimal parameters (number of LPC coefficients and frequency ceiling) to use with the *Praat* [15] formant measurement function on a per-phoneme and per-speaker basis by comparing the results at each step to a set of language-specific input vowel prototypes.[6] The measured F3 values were then vowel-extrinsically normalized using the Neary2 method [16], with the modifications suggested in [17].[6] Phones less than 50 ms in length and/or from stop words were discarded, leaving a total of 5,438 tokens: 1,885 of /ø/, 1,812 of /œ/, and 1,741 of /œ̃/. The first and last 20 percent of duration of each token were then discarded to avoid coarticulation effects from preceding or following segments, and finally only the minimum F3 value was retained for modelling (for the sake of comparability, as this closely mimics Mielke's methodology in [4]).

A mixed-effects linear regression model of minimum F3 was fitted using the *lmerTest* package in *R* [18, 19]. In addition to predictors for GENDER, YEAR OF BIRTH (YOB), and DIALECT, as well as a ternary PHONEME predictor distinguishing between the three vowels, a pair of controls were added. First, SYLLABLE POSITION (σ POS) marks out word-final syllables from others, since these bear accent in the language. Second is a control for following CONTEXT: this is needed because four consonants (/ʁ, v, z, ʒ/) are known to trigger an allophonic lengthening and diphthongization process which may interact with rhotacization (cf. [20]). Since the rhotic may by its very nature have a particular effect on F3, the factor has three levels: ʁ, other lengthening C, and other. Rounding out the fixed effects are a series of interactions terms—in particular, GENDER:YOB, DIALECT:YOB, and GENDER:DIALECT, as well as all binary interactions between PHONE and the social predictors. As for random effects, by-speaker intercepts and slopes for PHONE and CONTEXT, as well as by-word intercepts and slopes for YOB (controlling for lexical effects), were included.

All binary predictors were centred around the mean and scaled by 2σ. Ternary predictors used weighted Helmert contrasts (due to the unbalanced nature of corpus data). Accordingly, the contrasts of PHONE represent /ø/ vs. /œ/ and nasal vs. non-nasal; those of DIALECT represent West vs. Central and non-East vs. East; those of CONTEXT represent /ʁ/ vs. /v, z, ʒ/ and lengthening C vs. other.

3. RESULTS

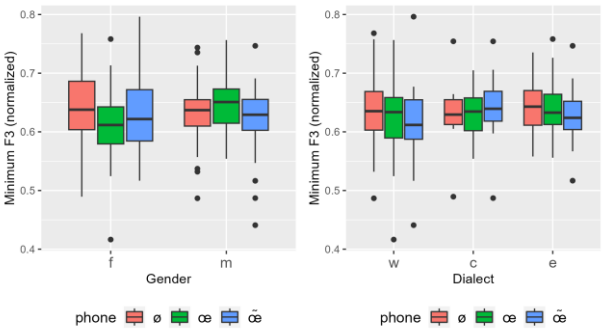


Figure 1: Empirical plots of GENDER (left) and DIALECT (right) by PHONEME, using speaker means.

Before examining the model predictions, it is useful to consider the empirical distribution of the data along the dimensions of interest.[8] In the following

discussion and figures, by-speaker averages were first computed in order to prevent statistics from being skewed by a few prolific speakers.

The average minimum F3 values across the dataset for each phoneme are quite similar, at 0.64 for /ø/, 0.63 for /œ/ and 0.62 for /œ̃/, with equally similar standard deviations of 0.056, 0.055, and 0.058, respectively—indicative of a sizeable amount of inter-speaker variation. As seen in Figure 1, however, these figures obscure some differences between the GENDER and DIALECT groups. While the mean of /œ̃/ is not clearly different for women and men ($\mu = 0.63 \pm 0.13$ vs. $\mu = 0.62 \pm 0.10$), there do appear to be differences in the other two vowels: for women, /ø/ appears to be slightly higher than /œ/ (0.64 ± 0.13 vs. 0.61 ± 0.14), while this pattern is reversed for men (0.63 ± 0.10 vs. 0.64 ± 0.08)—although note the large confidence intervals. As for DIALECT, West and East look quite similar, both exhibiting the pattern /ø/ = /œ/ > /œ̃ (0.63 ± 0.12, 0.63 ± 0.12, 0.61 ± 0.12 vs 0.64 ± 0.08, 0.64 ± 0.09, 0.63 ± 0.09), albeit with there being somewhat less variability in the East. The Central dialect, conversely, shows very little difference between the phonemes (0.64 ± 0.14, 0.63 ± 0.09, 0.64 ± 0.09 for /ø/, /œ/ and /œ̃/, respectively).



Figure 2: Empirical plot of YOB vs. speaker-averaged minimum F3, faceted by PHONEME.

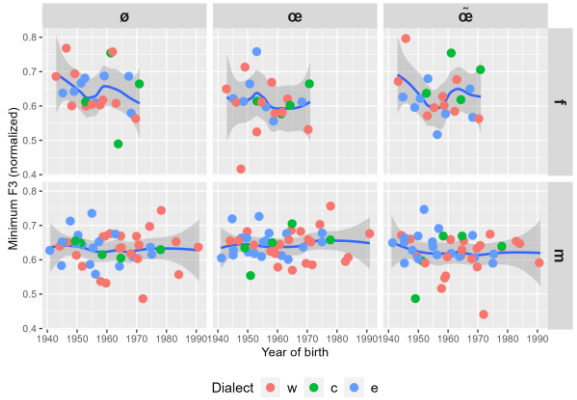


Figure 3: Empirical plot of YOB vs. speaker-averaged minimum F3, faceted by GENDER and PHONEME.

Figures 2 and 3 show individual speaker means for each phoneme (with points labelled by both gender and dialect), as well as their evolution through apparent time. Here, we see that all dialect and gender groups have both very rhotic and very non-rhotic speakers. Across the whole dataset, the YOB smooths by phoneme (Figure 2) do not reveal much in the way of change: there may be a general downward trend (at least for /ø/ and /œ/), although this is very uncertain (largely due to the lack of speakers towards the end of the range, after about 1975). When broken down by GENDER (Figure 3), the pattern becomes a bit clearer: the results are suggestive of a decrease across all three phonemes for women (although the confidence intervals still include 0), whereas no change is manifest for the men. No pattern over time across dialects emerges.

Fixed effects				
Coefficient	Est.	SE	*t*	*p*
(Intercept)	0.633	0.0072	88.08	< 0.001
PHONEME1	0.026	0.0099	2.65	0.009
PHONEME2	-0.013	0.0137	-0.91	0.362
YOB	-0.006	0.0123	-0.48	0.633
GENDER	0.005	0.0126	0.43	0.670
DIALECT1	0.014	0.0172	0.82	0.418
DIALECT2	0.001	0.0129	0.05	0.959
GENDER:DIALECT1	-0.024	0.0341	-0.71	0.478
GENDER:DIALECT2	0.021	0.0258	0.82	0.414
YOB:GENDER	0.039	0.0303	1.28	0.204
YOB:DIALECT1	0.058	0.0366	1.59	0.120
YOB:DIALECT2	-0.033	0.0271	-1.21	0.231
PHONEME1:YOB	0.009	0.0140	0.66	0.514
PHONEME2:YOB	-0.006	0.0159	-0.38	0.710
PHONEME1:GENDER	0.037	0.0141	2.60	0.012
PHONEME2:GENDER	-0.017	0.0141	-1.17	0.245
PHONEME1:DIALECT1	-0.004	0.0179	-0.23	0.817
PHONEME1:DIALECT2	0.004	0.0143	0.31	0.761
PHONEME2:DIALECT1	0.010	0.0183	0.55	0.588
PHONEME2:DIALECT2	-0.009	0.0144	-0.61	0.543
Random effects				
	Variance	σ		
SPKR (INTERCEPT)	0.0015	0.039		
SPKR (PHONEME1)	0.0011	0.033		
SPKR (PHONEME2)	0.0014	0.038		

Table 2: Abridged model table.

Table 2 lists the key model estimates.[9] Only two predictors are found to be significant:[10] the difference between /ø/ and /œ/ ($p = 0.009$), and the interaction between that predictor and GENDER ($p = 0.012$). The combined effects of these terms are visible in Figure 4. Together, they mean that while there is an overall significant /ø/–/œ/ difference, this is due entirely to the men (0.62 ± 0.02 vs. 0.65 ± 0.02), whereas for women there is no clear difference (0.63 ± 0.03 vs 0.63 ± 0.02). The relationship between /œ̃/ and either

of its oral counterparts is unclear both overall and for each gender group due to the greater variability.

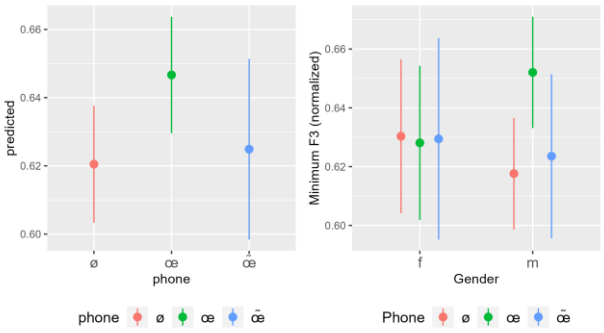


Figure 4: Left: Predicted minimum F3 by phoneme. Right: Interaction between PHONEME and GENDER.

Finally, the overall effect of YOB, while trending downward, is not significant. Figure 5 shows the predicted effects of the interaction between YOB on the one hand and GENDER (left) and DIALECT (right) on the other. A hint of a GENDER difference is visible here: while there is no clear effect for men (the very slight positive slope may be due to a few high-leverage speakers), a decrease in F3 over time can indeed be seen for women. This effect, however, is not significantly different from 0, due to the sparsity of speakers (especially female ones) towards the right end of the YOB range. Similarly, for DIALECT, a similarly sized negative effect is predicted for both the West and East dialects over time, although neither is significant. (Although a positive effect is predicted for the Central dialect, it is very uncertain: it is likely an artefact of the small sample size of this group.) This is likely due to the large amount of inter-speaker variability: the standard deviations of the by-speaker random intercepts ($\sigma = 0.04$), as well as those of the by-speaker random slopes for the PHONEME effects (0.03 for /ø/–/œ/ and 0.04 for oral–nasal), are larger than most of the predicted fixed effect coefficients. There is thus a high level of variability that is not conditioned by the social predictors examined here.

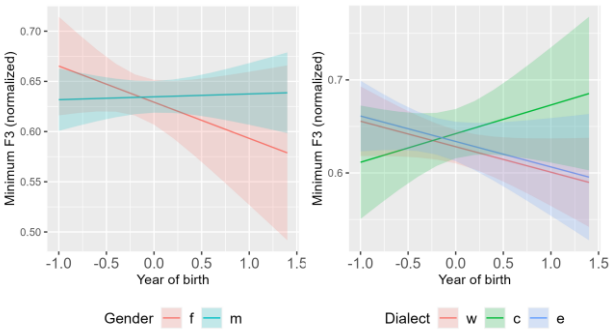


Figure 5: Predicted effects of GENDER (left) and DIALECT (right) by YOB

4. DISCUSSION

The present study confirms one of the major findings of Mielke's studies and extends it to all regions of Quebec: the existence of a large range of inter-speaker variability with respect to rhotacization of the front mid rounded vowels. The evidence for change in progress, however, is not as strong here as it was there, since the effect of YOB is not significant—not even in any of the subgroups. This is perhaps not surprising, as the number of speakers in the corpus used here thins out just as Mielke finds rhotacization begins in earnest in Gatineau-Ottawa. It is possible, then, that rhotacization—despite existing elsewhere, at least in a minority of speakers—only began to progress at an appreciable rate anywhere in Quebec with the cohort of speakers born after 1965. (In fact, it has been argued that a high degree of inter-speaker variability is a prerequisite for sound change [23], which could mean this study has correctly gauged the state of affairs just before the rhotacization took off.) However, it is also possible that the negative effects of YOB observed in 2 of the 3 dialects could become clearer with more data. In any case, the present study unfortunately does not bring clarity to the question of where rhotacization originated, nor to whether its spread is consistent with change from below.

There is also here a hint of a gender effect in two places. One of these is the negative YOB effect found in women but not in men, *contra* Mielke's findings. To be sure, this difference is not significant. Even if it were to be borne out, however, despite contradicting one of the initial hypotheses, it does not necessarily pose a problem for the account of change from below: women in particular have been found to lead changes from below by 1-2 generations in several different studies (cf. [24]). The results here (however tentative) therefore do not contradict Mielke's findings, although more data must be brought to bear on this question.

A further expansion of the corpus with 20-30 additional younger speakers (especially women and speakers from the Central dialect) is currently in progress, and may be sufficient to answer the research questions in much more depth than was possible here. A further goal for future work is to address the issue of the source of rhotacization as a change, which was not taken up here. Finally, there may be additional signal to be found in modelling formant trajectories: this is worthy of investigation in any follow-up study.

Acknowledgments This work was funded by FRQSC B1Z award #321056, as well as Morgan Sonderegger's SSHRC grant #435-2017-0925. Many thanks to MS for his guidance and insight, and to Irene Smith for her comments and support.

5. REFERENCES

[1] Ladefoged, P., Maddieson, I. 1996. *The sounds of the world's languages.* In: *Phonological Theory*. Oxford, UK: Blackwell.

[2] Delattre, P., Freeman, C. 1968. A dialect study of American r's by x-ray motion picture. *Linguistics* 44, 29–68.

[3] Dumas, D. 1972. Le français populaire de Montréal : description phonologique. *Université de Montréal* M.A. thesis.

[4] Mielke, J. 2013. Ultrasound and corpus study of a change from below: Vowel rhoticity in Canadian French. *University of Pennsylvania Working Papers in Linguistics* 19(2), 141–149.

[5] Mielke, J. 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *Journal of the Acoustical Society of America* 137(5), 2858–2869.

[6] Labov, W. 1994. *Principles of Linguistic Change,* vol. 1: Internal Factors. Chichester, UK: Blackwell.

[7] Stevens, K. 1998. *Acoustic phonetics*. Cambridge, MA, USA: MIT Press.

[8] Tousignant, C. 1987. Les variantes du /R/ montréalais: Contextes phonologiques favorisant leur apparition. *Revue québécoise de linguistique théorique et appliquée* 6(3), 73–113.

[9] Milne, P. 2014. The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French. *University of Ottawa* Ph. D. Thesis.

[10] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proc. 18th Interspeech*, 498–502.

[11] Assemblée nationale du Québec. Dictionnaire des parlementaires du Québec de 1764 à nos jours. *https://www.assnat.qc.ca/fr/membres/notices/index.html*

[12] Saint-Amant Lamy, H. 2016. Kilomètre zéro : /R/ et le statut épicentral de Montréal dans le domaine linguistique laurentien. *Cahiers internationaux de sociolinguistique* 10, 101–134.

[13] McAuliffe, M., Stengel-Eskin, E., Socolof, M., Coles, A., Mihuc, S., Goodale, M., Sonderegger, M. 2019. PolyglotDB [Computer program]. Version 0.1.0, https://github.com/MontrealCorpusTools/PolyglotDB.

[14] Mielke, J., Thomas, E.R., Fruehwald, J. McAuliffe, M., Sonderegger, M., Stuart-Smith, J., Dodsworth, R. 2019. Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several English speech corpora. *Proc. 19th ICPhS* Melbourne, 1258–1262.

[15] Boersma, P., Weenink, D. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.09, http://www.praat.org/.

[16] Neary, T.M. 1978. *Phonetic feature systems for vowels*. Bloomington: Indiana University Linguistics Club.

[17] Voeten, C.C., Heeringa, W., Van de Velde, H. 2022. Normalization of nonlinearly time-dynamic vowels. *Journal of the Acoustical Society of America* 152(5), 2692–2710.

[18] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26.

[19] R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

[20] Dumas, D. 1974. Durée vocalique et diphthongaison en français québécois. In: *Cahiers de linguistique de l'Université du Québec* 4. 13–55.

[21] Lüdecke, D., Aust, F., Crawley, S., Ben-Shachar, M.S. 2022. ggeffects: Create tidy data frames of marginal effects for 'ggplot' from model outputs [R package]. Version 1.1.4, https://strengejacke.github.io/ggeffects/.

[22] Wickham, H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York. https://ggplot2.tidyverse.org.

[23] Baker, A., Archangeli, D., Mielke, J. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change* 23, 347–374.

[24] D'Arcy, A. 2015. Variation, transmission, incrementation. In: *The Oxford Handbook of Historical Phonology*, Honeybone, P., Salmons, J. (eds.), 583–602.

---

[1] Cf. [3] : « […] l'usage de la rétroflexion prend en montréalais une valeur stylistique propre, celle de l'affirmation de soi par référence à l'anglais, qui est sans doute perçu comme 'argument d'autorité' ». (p. 105)

[2] In particular, the lowered F3 which results from creating a constriction at a displacement antinode [7].

[3] Alternatively, at least for /ø/ and /œ/, phonologization of the coarticulatory effects of a following rhotic consonant (which also weakens and deletes), a context where these vowels occur before disproportionately often, may be at cause (cf. [8]).

[4] Milne reports there being 62 speakers, but in the course of this study it was discovered that one was originally double-counted.

[5] More specifically, this Central dialect includes speakers born in the Mauricie, Centre-du-Québec and Estrie administrative regions.

[6] The prototypes are composed of by-phone (but across-speaker) mean formant frequencies, bandwidths and amplitudes, as well as the covariance matrix of these measures. The algorithm then attempts to minimize the Mahalonbis distance between the measured formants and the prototype distributions.

[7] This method was chosen both for its strong performance in [17]'s comparative testing and for comparability purposes with [4], since it is also used there.

[8] All empirical plots are made in R with the *ggplot2* package [21], and all model prediction plots are made with the *ggeffects* package [22].

[9] The full model is available at this project's OSF page, available online at https://osf.io/p3gr7/.

[10] While this might be thought to suggest the model is overparametrized, subsets of this model had essentially the same results with respect to significance.

LINGUISTIC STABILITY & MALLEABILITY IN LONG-TERM DIALECT CONTACT: THE COT/CAUGHT CONTRAST AMONG MOBILE SPEAKERS

Jennifer Nycz

Georgetown University
jennifer.nycz@georgetown.edu

ABSTRACT

This paper examines the low back vowel systems of two mobile English-speaking groups: 40 people from Toronto, Ontario now living in New York City and 30 native New Yorkers living in Toronto. Vowel measurements of words in the COT and CAUGHT lexical sets were drawn from minimal pair tasks and conversational speech to determine whether speakers in either sample show evidence of acquiring the system of their new ambient dialect. Results indicate that natively COT/CAUGHT-merging Torontonians exhibit small yet significant phonetic distinctions between word classes after years living in New York, though this seems to be the result of changes in individual lexical items, not a category-level split resulting in new phonemic structure. Similarly, natively COT/CAUGHT-distinguishing New Yorkers in Toronto show evidence of phonetically approximating these vowels, but their underlying contrast remains robust. That is, both groups exhibit structural stability yet phonetic malleability as a result of mobility-induced contact.

Keywords: dialect acquisition, accent change and mobility, sociophonetics, vowel merger and split

1. INTRODUCTION

People often change the way they speak after moving to and living in a new region. This phenomenon is typically referred to as *second dialect acquisition* (SDA), but speakers do not change every (or even most) of their dialect features to match those of their new community. Studies of dialect change among mobile individuals indicate that specific features are altered depending on a range of linguistic, social, and developmental factors (see [1, 2] for overviews). Focusing on the linguistic domain, several scholars have argued that the likelihood of adopting particular features depends on the linguistic complexity of the change involved [3, 4, 5]: changes which seem to require learning a complex set of phonological, grammatical, or lexical conditioning factors are less frequently attested than those which are less constrained. The most "complex" (and thus most difficult-to-acquire) type of feature in this view is a non-native phonemic contrast. To acquire a new phonemic distinction, a learner would have to not only create new contrastive categories but also assign every relevant lexical item to the correct new category. Mobile speakers faced with a non-native merger, however, need only implement a relatively simple change (a single process of neutralization), to accommodate to new dialect forms [6].

Previous studies have investigated whether mobile speakers exposed to a new contrast show signs of acquiring that contrast. In a study of six Canadian children living in England, Chambers [4] found that most showed little evidence of distinguishing words like COT and CAUGHT (which have been merged for several generations in most varieties of Canadian English [7]) in word list production. Through acoustic analysis of conversational speech, however, Nycz [8] found that a majority of seventeen Canadian adults who had been living in the New York City area showed gradient shift towards the local COT/CAUGHT contrast, producing a significant (though small) phonetic distinction between these word classes. Neither study, however, can address the relative difficulty or likelihood of acquiring a merger versus a distinction, as they do not examine the reciprocal case, in which individuals from a low-back-vowel-distinguishing region move to a place characterized by a low back merger. This paper fills this empirical gap, by examining production and judgments in both types of contact scenarios: Torontonians in New York (TO-in-NYC) encountering a local dialect in which two vowel categories map onto their single low back vowel, and New Yorkers in Toronto (NYC-in-TO) facing the opposite situation.

2. METHODS

2.1. Participants

The data analyzed here are drawn from interactions with 71 participants: 40 people born-and-bred in Toronto, Ontario, Canada who had moved as adults to New York City, New York US, and 31 natives of New York City who had been living in Toronto.[1] All participants had been living in their current city for at least 4 years at time of interview, though time in that city as well as the age at which speakers moved there varied within and across the samples (Figure 1). The two speaker groups have a similar mean "Age of Arrival" (mid-20s), though on average the Toronto sample is 12 years older than the New York sample, and thus (again, on average), has spent longer in their current region. Both samples vary internally with respect to self-reported ethnicity and social-attitudinal factors which will not be discussed here.

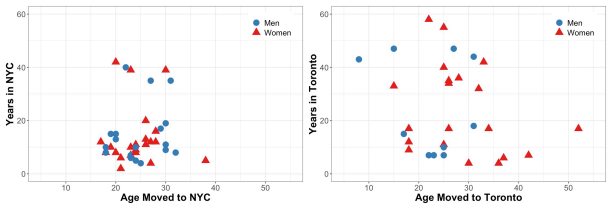


Figure 1: Study participants

2.2. Data collection and tasks

This analysis draws on data collected in New York City and Toronto in 2018-2020. Participants took part in conversational interviews focusing on growing up in Canada or the United States (as appropriate), moving to their new country and city, and impressions of the people and culture there.[2] After the conversational portion of the interview, participants completed word list readings and minimal pair tasks. In the minimal pair task, word pairs distinguished by the target low back vowels (e.g. *cot* and *caught*; *dawn* and *don*) were presented to participants on paper or via an iPad, with other potential pairs (e.g. *Mary/merry, pin/pen*) interspersed to pull focus away from the low back contrast of interest. Interviewers instructed participants to read each pair out loud and then say whether the two words sound the same or different to them. The minimal pair task thus yielded two types of data: production data as well as judgments regarding whether words in pairs contrast.

All interviews were recorded to 44.1kHz/16 bit wav files using a Zoom H4N solid state recorder and an Audio-Technica AT831b lavalier microphone. Interviews took place in the participant's home or a quiet public space (such as in a meeting room in the New York Public Library, for NYC participants). Consent was obtained and participants received an Amazon gift card worth $20 in local currency upon completion of study activities.

Data from the conversational portion of the interview and the minimal pair lists will be the focus of the analysis presented here.

2.3. Data processing and acoustic analysis

Each interview was transcribed in ELAN [9] by student research assistants. Transcripts and wav files were submitted to FAVE-Align python scripts [10] to generate time-aligned and segmented Praat textgrid files [11]. FAVE-Extract scripts were then used to extract formant measurements from all stressed vowels longer than 50 msec. Default configuration settings for FAVE-Extract were used, with maximum formant set automatically according to gender of the speaker. Vowel (i.e., word class) codes assigned by the CMU dictionary were hand-checked and corrected as needed. Raw formant measurements were Lobanov-normalized using the scale() function in R [12].

2.4. Statistical analysis

For each speaker (and in each style), a MANOVA model was fit with normalized F1 and F2 as the outcome variables and relevant phonological factors (following place, following manner, and vowel duration) and word class (COT or CAUGHT) as predictors. The Pillai statistic associated with word class was used as a measure of distinction [13]; because Pillai scores are sensitive to sample size, assessment of whether each score represented a significant phonetic distinction between COT and CAUGHT was carried out using the formula suggested in [14].

3. RESULTS

3.1. Minimal pair tasks

The TO-in-NYC and NYC-in-TO samples differed substantially in their minimal pair judgments. Of the 400 judgments produced by the 40 TO-in-NYC speakers, 335 were coded as *same*, 18 as *unsure*, and 47 as *different*, with 30 participants confidently judging at least 8 of the 10 pairs to sound the same. The 31 NYC-in-TO participants gave 309

judgments: 263 *different*, 10 *unsure*, and 36 *same*; 24 of these participants judged at least 8 pairs to sound different. The productions of these two groups are consistent with their judgments: the TO-in-NYC speakers individually produce COT and CAUGHT with very little distance between their means, and as a group show almost complete overlap between these categories, while the NYC-in-TO sample show a clear separation between word classes at the individual level and very little overlap at the group level (Fig. 2).

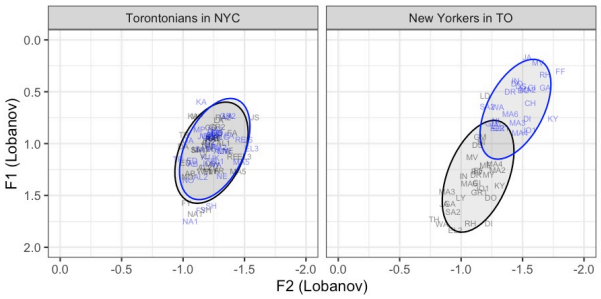


Figure 2: Productions of COT words (black) and CAUGHT words (blue) in the minimal pair task. Each label plots a speaker category mean; ellipses depict one standard deviation around the mean of the speaker means

The Pillai statistics associated with the MANOVA models of each speaker's minimal pair productions corroborate the overall difference suggested in the previous paragraph (see the righthand portion of Fig. 3). 35 of the TO-in-NYC participants have Pillai scores in this style that are low enough to indicate no significant difference between these vowels. For the NYC-in-TO group, models for all but two participants yielded a Pillai score high enough to indicate a significant phonetic distinction between COT and CAUGHT. To summarize, in the minimal pair task, both groups as a whole provide judgments and productions which reflect the dialect of the region in which they were born and raised: the TO-in-NYC speakers, whose first dialect does not distinguish two low back vowels, largely do not produce or report a difference between relevant words in the minimal pair task, while the NYC-in-TO speakers, whose first dialect does distinguish these word classes, mostly do produce this difference in their minimal pairs and recognize these words as different.

3.2. Conversational speech

Minimal pair productions of COT and CAUGHT across the two participant groups are quite different,

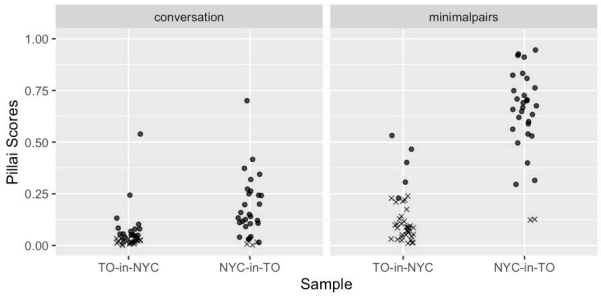


Figure 3: Pillai scores across samples and styles. Dots indicate Pillai scores high enough to suggest a distinction given token count; crosses indicate Pillais suggesting merger

reflecting a qualitative difference in the dialect that each group first acquired in their home region. The groups pattern more similarly, however, in their conversational speech: while TO-in-NYC as a group show signs of a split between these word classes, NYC-in-TO bring them closer together (Fig 4). This difference is also reflected in the Pillai scores for conversational style (see the lefthand portion of Fig. 3). For TO-in-NYC, Pillai scores still tend to be low, indicating much overlap between categories, but 17 have a Pillai that is sufficiently high given their token count to indicate a significant phonetic distinction between these categories. For NYC-in-Toronto, the majority (27 people) still have a Pillai that suggests a distinction between word classes, though these Pillais tend to be lower in conversational speech as compared to minimal pair productions.

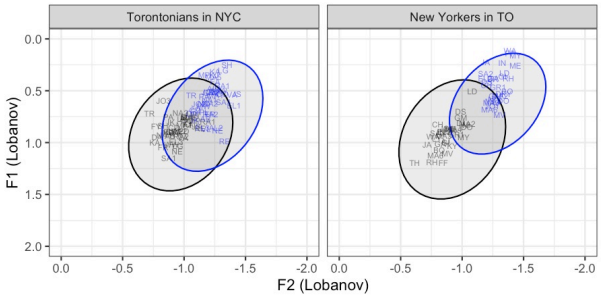


Figure 4: Productions of COT words (black) and CAUGHT words (blue) in conversational speech. Each label plots a speaker category mean; ellipses depict one sd around the mean of the speaker means

4. GENERAL DISCUSSION

In this data we observe a style difference in both samples, though in opposite directions. Native Torontonians in NYC mostly exhibit a merger in

minimal pair productions and judgments, but are more distinct in conversational productions. Native New Yorkers in TO are qualitatively more stable as a whole, showing a COT/CAUGHT distinction in both styles, though for them the magnitude of this distinction is smaller in conversational speech. The NYC-in-TO style difference could be explained in terms of a general tendency for conversational speech to be less dispersed than citation speech; the TO-in-NYC data, however, is not explainable in these terms. For both groups, the minimal pair task seems to elicit linguistic behavior which "shifts radically backward toward an older, corrected value" [15] - in these cases, productions reflecting the dialect acquired in one's region of origin. Conversational speech, meanwhile, shows more influence of the ambient second dialect: natively merged Torontonians exhibit a small but significant split which is more consistent with the New York City input they now receive, while New Yorkers whose native dialect contrasts the COT and CAUGHT sets reduce this contrast in conversation, consistent with *their* new Toronto English input. These patterns suggest that long-term convergence towards (or adoption of) a non-native *phonetic* distinction is not more difficult or unlikely than long-term convergence towards a non-native phonetic merger, contrary to claims that an acquisition asymmetry is to be expected: both groups show gradient, incomplete convergence towards their new dialect in spontaneous speech.

But what about long-term acquisition of a non-native *phonemic* contrast or merger? Here I would also argue that no asymmetry is present, because both samples show stability with respect to their underlying low back vowel system. New Yorkers may reduce the magnitude of their phonetic distinction in conversational speech, but their judgments that these words differ (and in which direction) remain robust. Similarly, while 43% of native Torontonians produce a significant measurable difference between COT and CAUGHT in conversation, their citation productions and judgments suggest that they do not recognize that this is a "real", in the sense of categorical, difference. Thus while we can say that new phonetic distinctions (or lack thereof) have been partially acquired by these mobile speakers in adulthood, there is as yet no evidence for changes at the phonemic level.

These patterns may be accounted for in a hybrid theory of lexical representation (such as current versions of Exemplar Theory [16]), which contains levels of representation corresponding to both phonemic categories and individual lexical items. The data presented here suggests that while the representations corresponding to individual lexical items accrue new exemplars over time, gradiently shifting the production targets of COT and CAUGHT words away from each other (for TO-in-NYC) or towards each other (for NYC-in-TO), the phoneme-level representations remain stable for both groups. So, while phonetic differences between word classes may emerge in certain styles (for example, conversational talk about life in one's new region, which may be more likely to activate exemplars associated with the new dialect), this does not imply that any change has occurred (or will occur) at the category level. In cases of mobility-induced dialect contact, structural stability and phonetic malleability can co-exist.

5. REFERENCES

[1] J. Siegel, *Second Dialect Acquisition*. Cambridge: Cambridge University Press, 2010.

[2] J. Nycz, "Second dialect acquisition: A sociophonetic perspective," *Lang. & Linguistics Compass*, vol. 9, pp. 469–482, 2015.

[3] A. Payne, "The acquisition of the phonological system of a second dialect," Ph.D. dissertation, University of Pennsylvania, 1976.

[4] J. K. Chambers, "Dialect acquisition," *Language*, vol. 68, no. 4, pp. 673–705, 1992.

[5] P. Kerswill, "Children, adolescents, and language change," *Lang. Variation & Change*, vol. 8, pp. 177–202, 1996.

[6] W. Labov, *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell, 1994.

[7] C. Boberg, *The English Language in Canada*. Cambridge: Cambridge University Press, 2010.

[8] J. Nycz, "New contrast acquisition: Methodological issues and theoretical implications," *English Lang. & Linguistics*, vol. 17, no. 2, pp. 325–357, 2013.

[9] ELAN, "Elan [computer software]. version 4.94." Available at: https://tla.mpi.nl/tools/tla-tools/elan/, 2018.

[10] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "FAVE (Forced Alignment and Vowel Extraction)," http://fave.ling.upenn.edu, 2011.

[11] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," Retrieved from http://www.praat.org/.

[12] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[13] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *J. Phon.*, vol. 34, pp. 458–484, 2006.

[14] J. Stanley and B. Sneller, "Sample size matters in calculating pillai scores," *Journal of the Acoustical Society of America*, Forthcoming.

[15] W. Labov, M. Karen, and C. Miller, "Near-mergers and the suspension of phonemic contrast," *Language Variation and Change*, vol. 3, pp. 33–74, 1991.
[16] J. Pierrehumbert, "The next toolkit," *J. Phon.*, vol. 34, pp. 516–530, 2006.

[1] This work is supported by National Science Foundation (NSF) Grant BCS-1651108.
[2] Participants in New York were interviewed by a native of Long Island, NY; participants in Toronto were interviewed by one of three graduate students, two of whom are Canadian and one of whom is a German national. All interviewers are white women in their 20s with sociolinguistics training.

Acquiring allophony: GOOSE and SCHOOL vowels in the speech of Australian children

Andy Gibson[a,b], Felicity Cox[a] & Joshua Penney[a]

[a]Macquarie University, Sydney; [b]Queen Mary University of London
gibsonism@gmail.com; felicity.cox@mq.edu.au; joshua.penney@mq.edu.au

ABSTRACT

The fronting of GOOSE has occurred in many varieties of English. In Australian English (AusE), prelateral tokens of GOOSE (referred to here with a lexical set label, SCHOOL) are the exception, and have become backed in recent generations. This study investigates the speech of 149 Australian children sampled through an online picture naming task over three timepoints, as they transition from preschool to school. Many children in the corpus speak a language other than English (LOTE) in the home. Higher levels of LOTE-usage are associated with less fronted (non-prelateral) GOOSE vowels. With respect to age, we find that while GOOSE does not appear to change across real-time, SCHOOL backs with increasing age. Additionally, children who have an older sibling show more retracted SCHOOL vowels. These findings suggest that SCHOOL-backing is an ongoing sound change, that may be subject to incrementation as children begin vernacular reorganisation.

Keywords: Australian English, child speech, GOOSE-fronting, sound change, diversity.

1. INTRODUCTION

In mainstream Australian English (AusE), as in many other varieties of English, the high /ʉː/ vowel of the GOOSE lexical set varies in its realisation from central–fronted [ʉː] in most phonetic contexts to retracted [uː] in prelateral contexts [1, 2]. This paper focuses on this allophonic split. We refer to the non-prelateral allophone of the /ʉː/ phoneme (in words such as *two*, *shoes*, *boot* and *juice*) as GOOSE, while we will refer to the prelateral allophone (in words such as *pool*, *stool*, *school*, and *cool*) as SCHOOL.

In the 1980s, [3] found that SCHOOL was retracted for speakers from Adelaide (and to a certain extent Melbourne) but not for speakers from Sydney or Brisbane, suggesting that the feature was strongly regional. Even as recently as 2008, [4:114] described retracted prelateral GOOSE as a feature of South Australian English, stating that 'In most regional varieties, similar vowel qualities occur for [GOOSE] vowels with or without a following lateral'. More recent research [1, 2, 5] suggests that the retraction of SCHOOL is a rapidly spreading change in progress for speakers from New South Wales (NSW). This SCHOOL-backing change, along with GOOSE-fronting, a change now nearing completion, leads to a strong split between the prelateral and non-prelateral allophones.

For preschoolers who speak languages other than English at home, the situation may be different. If a child's primary caregiver produces high back [uː] vowels, either in their L1 or in their (L2) English, this may impact upon a child's realisation of /ʉː/ in two ways: a backer GOOSE and possibly a less retracted SCHOOL. It should be noted, however, that even preschoolers do not always follow the model set forth by their caregivers: [6] found that, for 4–5 year old non-Anglo children growing up in London with parents born outside the U.K., there was no correlation between the children's and their principal caregivers' degree of GOOSE-fronting. They suggested that even at this age, children might reject variants used by their parents that are 'saliently non-local' [6:168]. Rather than focusing on ethnolinguistic background, in this study we look at the extent to which a given child uses a language other than English (LOTE) in the home, and consider whether this predicts degree of GOOSE-fronting.

1.1. Predictions for analysis

We assume based on recent analyses of AusE [1, 2, 5] that SCHOOL will be backer overall than GOOSE, but that the degree of this allophonic split will vary according to social factors. We have three specific predictions:

1. The more time a child spends speaking a language other than English in the home, the less fronted their GOOSE will be. This difference should not apply to SCHOOL which, if anything, may be more retracted for monolingual AusE speakers than for those who frequently use a LOTE at home;
2. If the backing of SCHOOL is an ongoing sound change, as shown in [5], we expect to see signs of incrementation [7], with retraction of SCHOOL over real-time;
3. If the backing of SCHOOL is an ongoing sound change, we also expect that children who have an older sibling will have a more

retracted SCHOOL vowel, having received early exposure to incremented variants.

The first prediction rests on the assumption that most of the LOTEs represented in the sample have a more retracted high-back vowel in their phoneme inventory than AusE. The children in the sample speak a wide range of languages, with the most frequently occurring being Mandarin, Urdu and Arabic. All of these languages include a high back vowel in their phonologies which may result in backer GOOSE vowels in the speech of caregivers who moved to Australia in adulthood [8-10]. In the analysis below, we examine the potential influence of speaking a LOTE at home, as well as exploring SCHOOL-backing as a change in progress.

2. METHODS

2.1. Participants and data collection methods

Children engaged in a self-recorded picture naming task delivered via the Gorilla online platform [11], framed as a game where the child helped a cartoon alien find its friends and its spaceship. 150 single words and short phrases were elicited but incidental items were also recorded as the children engaged with the task. Self-recorded in their own homes, the recordings come from a wide range of devices of varying quality. The project was approved by the Macquarie University Human Research Ethics Committee. Participants received compensation for their involvement.

This analysis includes recordings from 149 children (143 born in Australia and six who were born overseas but moved to Australia by 18 months of age). Almost all children are from New South Wales, with a large proportion from Sydney, and a cluster from the regional area of the Mid North Coast of NSW. Our original intention was to recruit children at three timepoints: before starting school, soon after starting school, and towards the end of the first year of schooling. However, due to limitations caused by the COVID-19 pandemic, including moving the task to open recruitment online, there was a high level of variation in the ages of the children. We present here mean age values at each timepoint to give an idea of ages and attrition rates. Our analysis in this paper however will use age in months, rather than timepoint, as the key independent variable for modelling real-time change. At their first recording timepoint (T1), the mean age was 59.1 months (n=148), at T2 (n=108) it was 66.5, and at T3 (n=97) it was 76.3.

Parents were invited to complete an in-depth questionnaire to provide details of the child's language and ethnic background. This revealed that 92 participants speak only English at home while 57 speak at least one language other than English (LOTE) at home. The survey also asked for the percentage of time each language is used in the home. This measure is used as a continuous predictor in the current analysis. Where this value was missing, proxy values were used. For those who had a LOTE listed as either their L1, or as their 'main language spoken at home' (or both), a value of 53.4% was assigned. For those with a LOTE listed as an 'other language spoken at home', a value of 24.7% was assigned. These values were chosen as being half a standard deviation either side of the mean % LOTE-usage for those who speak a LOTE at home (39.1).

2.2. Data processing and formant estimation

Data were orthographically transcribed by research assistants, and then processed through webMAUS [12] to provide phoneme level annotation. Boundaries were then hand-checked by phonetically trained research assistants who were instructed to use the automatic boundary unless it was obviously incorrect, in which case to place the boundary at the midpoint of the vowel-lateral sequence.

From an initial dataset of 7540 tokens of /ʉː/, we removed all prevocalic tokens, which can involve a linking glide and consequent coarticulation issues (n=1447), all instances of the lexical items *you* and *to*, which are often subject to reduction (n=864), and tokens that were found to be mislabelled (n=34).

Formants were estimated for the remaining tokens using FastTrack [13], choosing the best fit across a range of settings for the maximum formant value, ranging from 6000Hz–8000Hz. FastTrack estimated formants at seven time bins across each vowel, and the best fit was determined by smoothness of the formant trajectories.

Our analysis in this paper is entirely based on the midpoint measurement for each vowel. While an examination of the dynamics of these vowels would enrich this analysis, it was deemed to be outside the scope of the present paper. Given the automated nature of formant estimation, a stringent series of data pruning steps were then taken. The following 900 tokens were removed:

- Formant estimates missing (n=40).
- Midpoint F1 value > 800Hz (n=243).
- Midpoint measurement of F2 differed from either adjacent measurement by > 400Hz (n=441).
- Vowels shorter than 50ms (n=59).
- F2 measures > 2.5 standard deviations above, or < 2.5 standard deviations below

the mean, calculated separately for GOOSE (n=94) and SCHOOL (n=23).

For each of the above criteria, some tokens were listened to (by the first author) to determine a reasonable compromise between rejecting correctly tracked tokens and retaining mis-tracked tokens. After these exclusions, all tokens produced by five children who moved to Australia after the age of 18 months were also removed from the dataset (n=132). The final dataset used for the analyses below consisted of 4163 tokens (GOOSE n=3716; SCHOOL n=447) produced by 149 children.

3. RESULTS

3.1. Raw results

The mean F2 for children who speak only English at home was 2251Hz for GOOSE and 1028Hz for SCHOOL, compared to 2050Hz and 1077Hz, relpectively for children who speak a LOTE at home at least some of the time. Amongst these LOTE-speaking children, the mean F2 of GOOSE for speakers of Mandarin, Urdu and Arabic was 2019Hz, 1955Hz and 2074Hz, respectively.

3.2. Methods for statistical modelling

Linear mixed effects regression models were fit to predict the midpoint F2 of the /ʉː/ vowels, using *lme4* [14], considering the following four predictors:

- 'preL', that is, whether the following environment was /l/ or not, (a binary factor, non-prelateral GOOSE vs. prelateral SCHOOL);
- percentage of the time a LOTE is spoken at home (as a continuous predictor, centred and scaled), referred to as LOTE-usage;
- age in months (as a continuous predictor, centred and scaled);
- whether or not the child has an older sibling (a binary factor).

Random intercepts were included for participant and word. A slope for age on participant was also included since most children are represented at more than one age. This allows the model to predict the effect of age above and beyond idiosyncratic changes over timepoints for individual children. The slope for following environment on participant was also attempted but was abandoned as it led to singular fit errors. We followed a backward model selection process, beginning with the four-way interaction and sequentially removing least significant terms, using model comparison to confirm that each step in model simplification was justified.

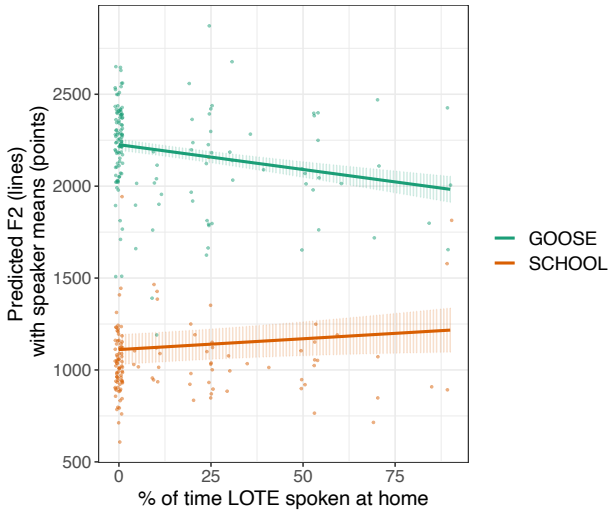


Figure 1: Predicted F2 (in Hertz) of GOOSE and SCHOOL according to percentage of time a given child speaks a language other than English (LOTE) at home.

3.3. Final statistical model

No four-way or three-way interactions were significant at an alpha of 0.05. The final model included three significant two-way interactions (all $p<0.01$), and the following syntax:

```
F2 ~ preL*OlderSibling + preL*LOTE-usage +
preL*Age + (1 + Age | Speaker) + (1 | Word)
```

Effect sizes from the model for each of these interactions are represented in Figures 1–3. Overall, a following /l/ robustly decreases the F2 of /ʉː/, that is, SCHOOL is consistently backer than GOOSE. The F2 of these two allophones varies as a function of LOTE-usage, age and having an older sibling.

Figure 1 shows that high levels of LOTE-usage are associated with backer tokens of GOOSE, and fronter tokens of SCHOOL. The figure shows the model predictions for this interaction with lines, along with points for the speaker means according to LOTE-usage, calculated separately for the prelateral and non-prelateral tokens.

Figure 2 shows the interaction of following environment with age. GOOSE does not change in any robust way with increasing age, but there is a strong tendency for SCHOOL to be more and more backed as the children get older.

Whether or not the child has an older sibling was found to significantly interact with following environment, as shown in Figure 3. Children with an older sibling have fronter GOOSE and more retracted SCHOOL. Though not included in the final model, there was a trend ($p=0.077$) for a three-way interaction between following environment, older sibling and age. This possible three-way interaction suggested that while SCHOOL retracts over real-time for all children, the slope of this change was steeper for children that have an older sibling.

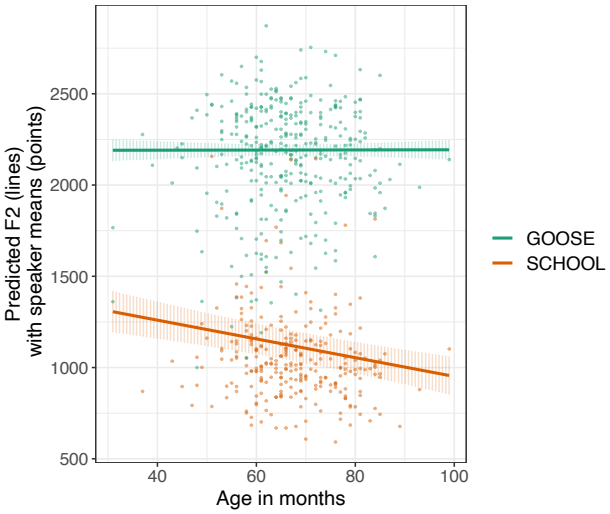


Figure 2: Predicted F2 (Hz) of GOOSE and SCHOOL according to the child's age in months.

4. DISCUSSION

4.1. Effect of LOTE-usage on GOOSE-fronting

Across the sample as a whole, we found a significant effect of the amount of LOTE-usage at home on the F2 of /ʉː/. Those who use a LOTE at home more often tend to have backer GOOSE vowels, supporting our first prediction. While this result may seem to depart from those of [6] discussed earlier, they are not directly comparable since we have focused on the extent to which children actually speak a language other than English in the home. Our results may thus capture transfer effects happening within the speech of these bilingual/multilingual children.

We were surprised at the non-significance of the three-way interaction of following environment, LOTE-usage, and age. We might expect that children who begin with a retracted goose would have begun to undergo rapid GOOSE-fronting to converge to their peers in the time scope of our recordings. The continuous LOTE measure used here may not have been the right measure to explore any fronting that may have occurred over real-time for the subset of LOTE-speaking children that had the most backed GOOSE vowels at the time of their first recording.

4.2. SCHOOL-backing as a change in progress

The second and third interactions reported revealed signs of incrementation of a change in progress, supporting our second and third predictions. While GOOSE showed no signs of change with increasing age, there was a significant backing of SCHOOL as children got older in real-time.

SCHOOL backs with increasing age, suggesting the real-time incrementation of an ongoing change in progress. These young children are acquiring and then incrementing the allophonic split between GOOSE and SCHOOL.

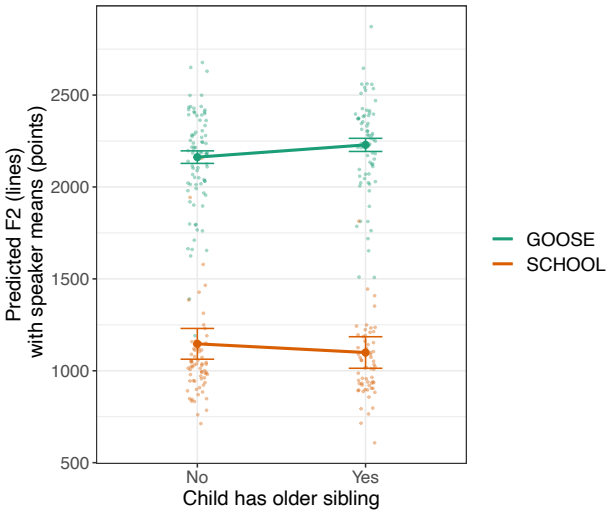


Figure 3: Predicted F2 (Hz) of GOOSE and SCHOOL according to whether the child has an older sibling or not.

Children with an older sibling have both fronter GOOSE and backer SCHOOL than children who lack a slightly older role model in the home. The presence of an older sibling may give these children a 'head-start' in extracting the direction of changes in progress. While we did not see evidence of GOOSE fronting in real-time, the older sibling effect for GOOSE suggests there may still be residual levels of GOOSE-fronting, as the change nears completion.

4.3. Limitations of automated methods

This study has used automated formant estimation methods, and it should be kept in mind that some tokens with formant tracking errors are almost certainly included in the results. While the gold-standard would be to check the formant tracking of every token, FastTrack is a useful and efficient tool for automatically achieving reasonable formant estimates from a large number of tokens. Additionally, the boundary between /ʉː/ and /l/ was largely determined by automated methods (though manually checked). Post-hoc analysis ruled out possible biases related to word-length: neither child age nor F2 of /ʉː/ correlated with word duration.

5. CONCLUSION

The allophonic split between GOOSE and SCHOOL appears to be an ongoing sound change for young speakers in NSW. Additionally, we see an effect of LOTE-usage on these vowels. A full understanding of this phenomenon needs to take into consideration the rich ethnolinguistic diversity in urban centres like multicultural Sydney.

6. ACKNOWLEDGEMENTS

This research was supported by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the second author.

7. REFERENCES

[1] T. Szalay, T. Benders, F. Cox, S. Palethorpe, and M. Proctor, "Spectral contrast reduction in Australian English /l/-final rimes," *J Acoust Soc Am,* vol. 149, no. 2, p. 1183, Feb 2021.

[2] B. Purser, J. Grama, and C. E. Travis, "Australian English over Time: Using Sociolinguistic Analysis to Inform Dialect Coaching," *Voice and Speech Review,* vol. 14, no. 3, pp. 269-291, 2020.

[3] H. Oasa, "Phonology of current Adelaide English," in *Australian English: The language of a new society*, P. Collins and D. Blair Eds. St Lucia: University of Queensland Press, 1989, pp. 271–287.

[4] D. Bradley, "Regional characteristics of Australian English: phonology," in *Varieties of English. Volume 3: The Pacific and Australasia*, K. Burridge and B. Kortmann Eds. Berlin: Mouton de Gruyter, 2008.

[5] T. Szalay, T. Benders, F. Cox, and M. Proctor, "Prelateral vowel change in Australian English POOL–PULL," in *Proceedings of the International Congress of Phonetic Sciences 2023*, Prague, 2023.

[6] J. Cheshire, P. Kerswill, S. Fox, and E. Torgersen, "Contact, the feature pool and the speech community: The emergence of Multicultural London English," *Journal of Sociolinguistics,* vol. 15, no. 2, pp. 151–196, 2011.

[7] W. Labov, *Principles of Linguistic Change, vol. 2: Social Factors*. Oxford: Wiley, 2001.

[8] N. Amir, O. Amir, and J. Rosenhouse, "Colloquial Arabic vowels in Israel: A comparative acoustic study of two dialects," *The Journal of the Acoustical Society of America,* vol. 136, no. 4, pp. 1895-1907, 2014.

[9] I. Rehman, "Urdu vowel system and perception of English vowels by Punjabi-Urdu speakers," University of Kent, 2019.

[10] B. D. Barkana and A. Patel, "Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems," *Applied Acoustics,* vol. 162, p. 107203, 2020.

[11] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods,* vol. 52, no. 1, pp. 388-407, 2020/02/01 2020.

[12] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services.," *Computer Speech & Language,* vol. 45, pp. 326–347, 2017.

[13] S. Barreda, "Fast Track: fast (nearly) automatic formant-tracking using Praat," *Linguistics Vanguard,* vol. 7, no. 1, 2020.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, no. 1, 2015.

BIAS AND CONSISTENCY OF INDIVIDUAL LINGUAL ARTICULATORY BEHAVIOR AND ITS RELATIONSHIP TO THE FIRST AND SECOND FORMANTS

Carolina Lins Machado[1], Lei He[1 2]

[1] Dept. Computational Linguistics, University of Zurich, Zurich, Switzerland
[2] Dept. Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland
cmachado@ifi.uzh.ch, lei.he@uzh.ch

ABSTRACT

Individual variation in articulatory behavior can be characterized by bias and consistency in movement outcome. Consistency can be indicated by *variable error* (VE) representing precision of individual performance and bias by *constant error* (CE) representing tendency in movement outcome. The present study employs CE and VE to characterize individual articulatory behavior, and assesses the relationship between consistency and bias in the articulatory and acoustic domains. We computed CE and VE of tongue blade and dorsum kinematic trajectories and the first two formants' curves in the production of /æ/ and /ɑ/ by 20 native U.S. English speakers. The relationship between acoustic and kinematic VE and CE were revealed using gradient boosting machines. Results indicate that individual CE and VE vary over the time course of a vowel and that movement outcome is affected by linguistic constraints.

Keywords: articulation, consistency, bias, individual variability, speech production

1. INTRODUCTION

Speakers differ in their vocal tract morphology and in the articulatory strategies they adopt to produce a highly similar acoustic output [2]. Besides phonemic information, an acoustic signal carries speaker-specific information. For instance, in vowels, formant contours have been shown to contain substantial individual information related to idiosyncratic articulatory behavior [23, 26, 16, 24, 18]. This behavior is far from constant, since speakers can produce the same sound employing different articulatory strategies [19].

In this study we seek to characterize speakers' articulatory behaviors using two measures employed in the investigation of human motor performance, the error scores *constant error* (CE) and *variable error* (VE) [22, 10, 9]. CE indicates the tendency, or bias, in individual motor performance. It represents the amount and direction of error, or deviation, from a target or criterion in a subject's movement [10]. In the context of this paper, *target* is defined as the mean of all speakers in a cohort, because in the motor control literature individual differences in movement outcome are described in relation to a predefined target or to other performers in the same environmental condition [22]. In speech, CE may indicate how much and in which direction a speaker deviates from a target. For instance, compared to others, a speaker could have the tendency to overshoot (direction) their tongue movement, fronting this articulator by 8 mm (amount) when producing an open vowel. Moreover, this speaker may be quite variable in their advancement. This inconsistency in movement outcome is indicated by the VE. Together, CE and VE scores can characterize the articulatory performance of individual speakers, highlighting their consistency and bias in movement outcome.

Articulatory movements modulate the acoustic signal; therefore, computing error scores of features in the acoustic domain (even though error scores are primarily used to describe movement performance) becomes a reasonable step to assess the relationship between the acoustic and articulatory dimensions. As previously stated, formant contours carry significant speaker-specific information. Thus, in this study, formant dynamics of the English vowels /æ/ and /ɑ/ were the acoustic feature selected for analysis. These vowels were chosen since they differ in inherent formant changes. While /æ/ requires formant movement to convey phonemic information [12], therefore requiring more precision in motor command, /ɑ/ is believed to have a relative stable acoustic outcome, less sensitive to small variations in articulatory movements [15]. Consequently, /ɑ/ may be more variable across speakers, since articulatory movements may be less stifled by linguistic constraints.

The present study is exploratory in nature, seeking to understand individual variation in the articulatory domain using error scores. More specifically, this study seeks to (i) investigate individual tendency and consistency of speech production in the articulatory dimension, and to (ii) examine the relationship between acoustic and articulatory error scores of the English vowels /æ/ and /ɑ/. We expect speaker's error scores to be less variable in the production of /æ/,

since linguistic constraints may affect the amount of individual information in formant contours.

2. METHOD

2.1. Material

Productions of the vowels /æ/ and /ɑ/ in single-word citation form by twenty native speakers of U.S. English (10 M, 10 F) with an upper Midwest American English dialect background were selected from the EMA-MAE corpus [1]. We excluded vowel tokens produced in the context of rhotic, lateral, nasal, and approximant syllable onset or codas, due to the coarticulatory effects on vowel formants related to these consonants [25].

Kinematic data (sampling rate = 400 Hz) were collected using the electromagnetic articulograph (NDI Wave) along with time-synchronized acoustic recordings (sampling rate = 22 kHz). Sensor data related to the position of the tongue dorsum (TD) and the tongue blade (TB) were analyzed in both anteroposterior (x) and vertical (y) directions. For each speaker the kinematic data were head-corrected and calibrated using a bite-plate. The tongue kinematic data in this analysis contains tongue-jaw compound movements corresponding to the tract variables of *tongue body constriction location/degree* and *tongue tip constriction location/degree* in Articulatory Phonology [7].

Acoustic and kinematic data were processed in Praat [20] following the same steps as [6]. The resulting first and second formant curves and kinematic trajectories, both comprising 5 analysis points, were subsequently processed in R [21]. After outlier removal (vowels with long silences in the middle or excessive creak leading to incorrect measures), subsets of the data were created per vowel containing two acoustic variables, F1 and F2, and four kinematic variables: TDx, TDy, TBx, TBy. Overall the datasets consisted of 1240 data points for /æ/ and 990 for /ɑ/.

2.2. Data analysis

2.2.1. Error scores

Speaker bias and consistency of articulatory and acoustic outcomes were determined by calculating error scores for every variable at each analysis point. CE, indicating bias, was calculated as shown in equation (1), where x_i is the value for repetition i, T is the target value (the average over all speakers), and k is the number of repetitions the participant performed. A CE score of zero indicates no bias in outcome. A value greater than zero indicates overshoot, and a score less than zero expresses movement undershoot.

$$\text{CE} = \Sigma(x_i - T)/k. \quad (1)$$

$$\text{VE} = \sqrt{\Sigma(x_i - M)^2/k}\,. \quad (2)$$

Equation (2) shows the calculation of VE, indicating consistency, where x_i and k are defined as in (1), and M is the speaker's mean. A VE score of 0 indicates that the speaker's productions were always exactly the same, and a value greater than zero indicates the extent of outcome inconsistency. The extent to which inconsistency is considered high or low is an empirical question not addressed in the present study. We believe the level of error to be a function of the linguistic environment, where some phonemes may allow more error with little consequences while others may not, thus affecting what is considered high or low inconsistency. Here, we used the mean VE over all speakers for each vowel as a threshold; a VE > 1.5 for /æ/ or a VE > 2 for /ɑ/ indicates high inconsistency.

The above computations yield CE and VE scores for the first two formants (CE_F1, CE_F2, VE_F1 and VE_F2) and the four articulatory variables (CE_TBx, CE_TBy, CE_TDx, CE_TDy, VE_TBx, VE_TBy, VE_TDx, and VE_TDy). Subsequently, to determine a possible effect of vowel on the error scores related to lingual movement, we employed the Levene's test for equality of variances [13]. CE or VE were the response variable grouped by vowel, articulatory variable and analysis point.

2.2.2. Gradient Boosting

To better understand the acoustic-articulatory relationship in each vowel, we used generalized boosted regression modeling from the R package gbm [4]. Our models may reveal an acoustic-kinematic relationship not despite the speaker effect, but by actually taking speaker characteristics into consideration, since the CE and VE reflect non-normalized speaker-specific characteristics.

For each vowel, gradient boosted models with gaussian loss function were built for CE and VE, with the formants error score as the predicted variables and the articulatory error scores as explanatory variables. Important hyper-parameters [3] were tuned over 5000 iterations using grid search with 10-fold cross-validation. For the hyper-parameter optimization and model performance metric, the root mean square error (RMSE) was used. Models were trained on 80% of the data and performance was evaluated on the remaining 20%.

3. RESULTS

3.1. Error Scores

Error scores indicated the consistency and bias of each speaker's motor performance. Here we only present the main differences between two randomly

selected speakers, due to the page constraint. Error scores of all speakers are available in the supplementary figures S1 and S2 via https://osf.io/xud7h.

Figure 1 shows the tendency (CE) and consistency (VE) in vowel productions of two male speakers. Regarding their articulatory tendencies for the vowel /ɑ/, speaker 15ENM's anteroposterior movements (TBx and TDx) tend to be directionally biased to fall short of the target, while speaker 35ENM tends to overshoot TBx. Further, speaker 35ENM tends to undershoot TBy and speaker 15ENM tends to overshoot this movement. In their production of /æ/, CE indicates that speaker 35ENM tends to undershoot his vertical tongue movements (TBy, TDy) as opposed to speaker 15ENM, whose vertical movements tend to be overshot.

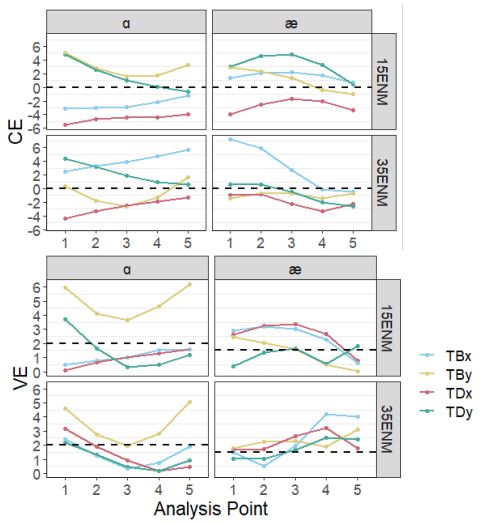


Figure 1: Constant Error (CE, top plot) and Variable Error (VE, bottom plot) scores of two male speakers calculated for four articulatory variables at five time points. In VE plots dashed lines correspond to the average inconsistency level.

In terms of consistency, VEs indicate that both speakers are most inconsistent in TBy of /ɑ/ production. In the anteroposterior movement (TBx and TDx), speaker 15ENM seems more consistent in the beginning of this vowel, while speaker 35ENM becomes more consistent towards the end. In the production of /æ/, speaker 15ENM becomes more consistent with time in his anteroposterior movement (TBx and TDx) and in TBy. Conversely, speaker 35ENM becomes more inconsistent with time. Together, CE and VE reveal that in the production of /ɑ/ 15ENM inconsistently overshoots TBy and 35ENM inconsistently undershoots it. Moreover, during the production of /æ/, although speaker 15ENM overshoots TBy, he tends to become more consistent with time. Contrariwise, speaker 35ENM tends to become less consistent with time undershooting TBy.

Due to differences in the magnitude and spread of VE and CE scores between /æ/ and /ɑ/ over all speakers, we decided to look into the effect of vowels in these scores. The results of the Levene's test indicated that the variance between vowels was not equal. Overall /ɑ/ had a significantly higher dispersion ($p < 0.001$) of CE and VE scores for all variables. The differences of variance between vowels for TBx were 1.54 (CE) and 0.75 (VE); for TBy 2.04 (CE) and 2 (VE); for TDx 4.28 (CE) and 1.12 (VE); and for TDy 0.99 (CE) and 3.06 (VE).

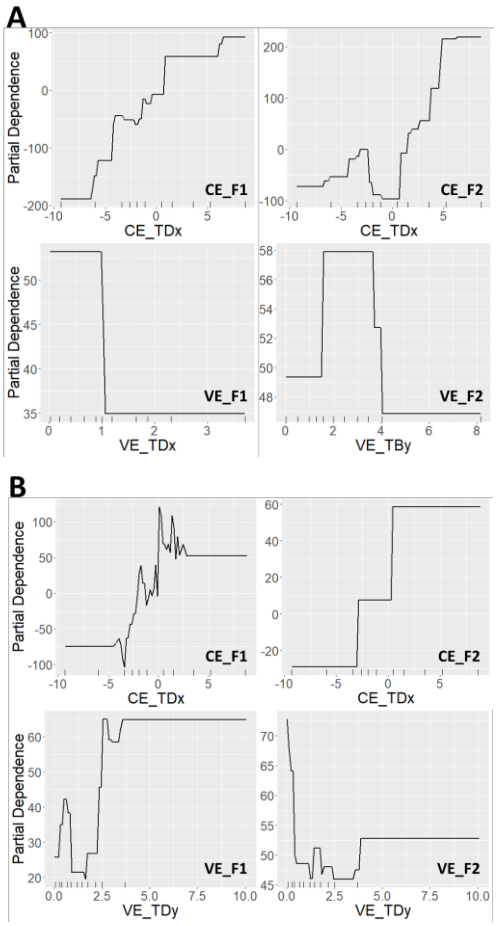


Figure 2: Partial dependence plots of the four models built for /æ/ (A), and /ɑ/ (B). The y-axis represents the marginal impact of the articulatory error scores to the acoustic error scores. Hash marks at the base of plots show distribution of error scores of each variable, in millimeters.

3.2. Gradient boosting

The normalized RMSE of all GBMs was equal or below 0.26, indicating good model fitting. The individual influence of the predictor variables on the response variable is indicated by their *relative importance*, a predictor's ranking based on their contributions to each model. For /æ/ the most important variables in predicting CE_F1, CE_F2 and VE_F1 were the median error scores of TDx, with a relative contribution of 45% (CE_F1), 59% (CE_F2) and 44% (VE_F1). In predicting VE_F2, TDx (30%)

was second to TBy (37%). For /ɑ/ the most important predictor of both formants CEs was TDx, with a contribution of 27% for F1 and 68% for F2. For the formants' VEs, TDy was the most important predictor to VE_F1 (39%) and VE_F2 (32%).

The partial dependence plots (Figure 2) illustrates the relationship between the error scores of the most relevant articulatory and acoustic variables. For both vowels, the partial dependence of CE_TDx on CE_F1 and CE_F2 is increasing over the main body of the data. Generally, a large CE_TDx means higher CEs for both formants. As for VE, in /æ/ VE_F1 seems higher with low VE_TDx, stepping downwards after a VE_TDx of 1, conversely, VE_F2 starts low, steeping upward after a VE_TBy of ca. 2. In /ɑ/, relationships between VE scores of F1–TDy and F2–TDy appear less straightforward. However, a general increase in the VE_TDy increases VE_F1 whereas a general increase in VE_TDy decreases VE_F2.

4. DISCUSSION

In this paper error scores were used to characterize individual articulatory performance. CE indicated bias, whether speakers in this dataset tend to undershoot or overshoot their tongue movements in the production of /æ/ and /ɑ/, and VE indicated how consistent their movements were.

Analyzing vowel dynamics revealed that bias in movement outcome is not a stationary characteristic. For instance, in speaker's 15ENM production of /ɑ/, although his TBx and TDx movements seem to show a quasi-constant CE score, other movements (e.g. TDy) seem to start with a great directional bias and approach the target over time. Similarly, changes in movement bias were observable in this vowel for speaker 35ENM, where TBx increased in bias with time, and TDy, showing the opposite progression, decreased in bias.

At an individual level, differences in the directional bias of the tongue articulatory variables could be an indication of quantal regions [17], i.e., a region of stability in the articulation-to-acoustics mapping. In this context, error scores may reveal which regions along the movement trajectory are more stable than others, therefore exhibiting a greater amount of movement bias. Our results seem to be in line with previous research, hypothesizing that formant trajectories may be more sensitive to continuous tongue movement at certain time points [5]. In addition, at a group level, differences in speakers' tendencies could be attributed to motor equivalence [19], which states that a similar acoustic outcome can be reached by employing different articulatory movements. Furthermore, we believe that aspects related to a speaker's physiology and motor preferences (e.g. [2]), along with linguistic constraints may determine the amount and direction of bias in tongue movement outcome.

Similarly, consistency in movement outcome seems to be restricted by linguistic constraints. Our results reflected the assumption that the production of /ɑ/ would be more variable than /æ/, because the latter relies on formant trajectories to denote phoneme identity, consequently requiring more controlled articulatory movements during its production, and thus constraining speaker-specific behavior. Moreover, our findings support previous studies demonstrating that some speech features are indeed less variable between speakers than others [24, 11, 8].

On the questions of the relationship between the error scores of articulatory and acoustic variables, we believe that interactions between the directions of tongue movement [14] may be expressed in the acoustic outcome. Regarding F2 bias, its dependence to CE_TDx for both vowels was unsurprising, since the relationship between this formant and the anteroposterior tongue movement follows a widely held view that F2 increases as the tongue advances. However, F2 consistency dependence to vertical tongue movements could be the result of an interaction between the vertical and horizontal directions of tongue movement. The current interpretation remains speculative since we did not test this assumption.

As to F1, we expected changes in tongue height to affect this formant, since vertical tongue movement alters the pharyngeal cavity where F1 resonates. Yet, with the exception of /ɑ/'s VE_F1, our results revealed a relationship between both vowels' F1 bias and /æ/'s consistency and the anteroposterior tongue dorsum movement (TDx). Ours is not the first study to find a relationship between F1 and the tongue anteroposterior direction (see [5] for an account on diphthongs). In the present study, we believe that tongue retraction may be the primary variable affecting the volume of the pharyngeal cavity [14] and consequently F1 CE and VE.

Ultimately, this study has shown that consistency and bias are non-stationary individual characteristics affected by linguistic constraints and that the acoustic-articulatory relationship still requires further investigation to understand its complexity.

6. ACKNOWLEDGEMENTS

This work was supported by the Forschungskredit of the University of Zurich (Grant No. FK-20-078), and the Swiss National Science Foundation (Grant #PZ00P1_193328) to LH.

7. REFERENCES

[1] A. Ji, J. J. Berry, and M. T. Johnson, "The Electromagnetic Articulography Mandarin Accented

English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7719–7723. doi: 10.1109/ICASSP.2014.6855102.

[2] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker Variability in Hard Palate Morphology and Vowel Production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, Dec. 2013, doi: 10.1044/1092-4388(2013/12-0211).

[3] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[4] B. Greenwell, B. Boehmke, J. Cunningham, and G. B. M. Developers, "gbm: Generalized Boosted Regression Models." 2020. [Online]. Available: https://CRAN.R-project.org/package=gbm

[5] C. Dromey, G.-O. Jang, and K. Hollis, "Assessing correlations between lingual movements and formants," *Speech Communication*, vol. 55, no. 2, pp. 315–328, 2013, doi: https://doi.org/10.1016/j.specom.2012.09.001.

[6] C. Lins Machado, V. Dellwo, and L. He, "Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis," in *Proc. Interspeech 2022*, 2022, pp. 754–758. doi: 10.21437/Interspeech.2022-10397.

[7] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989, doi: 10.1017/S0952675700001019.

[8] C. Schindler and C. Draxler, "Using spectral moments as a speaker specific feature in nasals and fricatives," in *Interspeech 2013*, Aug. 2013, pp. 2793–2796. doi: 10.21437/Interspeech.2013-639.

[9] D. Guth, "Space saving statistics: An introduction to constant error, variable error, and absolute error.," *Peabody Journal of Education*, vol. 67, no. 2, pp. 110–120, Jan. 1990, doi: 10.1080/01619569009538684.

[10] F. M. Henry, "Variable and Constant Performance Errors Within a Group of Individuals," *Journal of Motor Behavior*, vol. 6, no. 3, pp. 149–154, Sep. 1974, doi: 10.1080/00222895.1974.10734991.

[11] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, "A forensic phonetic study of dynamic sources of variability in speech: The dyvis project," in *Proceedings of the 11th Australasian international conference on speech science and technology*, 2006, pp. 13–18.

[12] G. S. Morrison and P. F. Assmann, Eds., *Vowel Inherent Spectral Change*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-14209-3.

[13] H. Levene, "Robust Tests for Equality of Variances," I. Olkin, Ed. Palo Alto, CA: Stanford University Press, 1960.

[14] J. H. Esling, "There Are No Back Vowels: The Laryngeal Articulator Model," *Canadian Journal of Linguistics/Revue canadienne de linguistiqu*e, vol. 50, no. 1–4, pp. 13–44, 2005, doi: 10.1017/S0008413100003650.

[15] J. S. Perkell *et al.*, "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *Journal of Phonetics*, vol. 28, no. 3, pp. 233–272, 2000, doi: https://doi.org/10.1006/jpho.2000.0116.

[16] K. McDougall, "Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies," *International Journal of Speech, Language and the Law*, vol. 13, no. 1, pp. 89–126, Jun. 2006, doi: 10.1558/sll.2006.13.1.89.

[17] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 17, no. 1–2, pp. 3–45, Jan. 1989, doi: 10.1016/S0095-4470(19)31520-7.

[18] L. He, Y. Zhang, and V. Dellwo, "Between-speaker variability and temporal organization of the first formant," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL209–EL214, 2019, doi: 10.1121/1.5093450.

[19] O. M. Hughes and J. H. Abbs, "Labial-Mandibular Coordination in the Production of Speech: Implications for the Operation of Motor Equivalence," *Phonetica*, vol. 33, no. 3, pp. 199–221, 1976, doi: doi:10.1159/000259722.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Jul. 22, 2021. [Online]. Available: http://www.praat.org/

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.Rproject.org/

[22] R. A. Schmidt, T. D. Lee, C. Winstein, G. Wulf, and H. N. Zelaznik, *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics, 2018.

[23] T. Kitamura and M. Akagi, "Speaker Individualities in Speech Spectral Envelopes and Fundamental Frequency Contours," in *Speaker Classification II*, vol. 4441, C. Müller, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 157–176. doi: 10.1007/978-3-540-74122-0_14.

[24] W. Heeren, "The contribution of dynamic versus static formant information in conversational speech," *International Journal of Speech Language and the Law*, vol. 27, no. 1, pp. 75–98, Aug. 2020, doi: 10.1558/ijsll.41058.

[25] W. Labov, S. Ash, and C. Boberg, *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter, 2006.

[26] X. Yang, J. B. Millar, and I. Macleod, "On the sources of inter- and intra- speaker variability in the acoustic dynamics of speech," *Proceedings of Fourth International Conference on Spoken Language Processing*. ICSLP '96, Oct. 1996, vol. 3, pp. 1792–1795 vol.3. doi: 10.1109/ICSLP.1996.607977.

Investigating the effects of talker age and listener age on trait perception from adult voices

Authors: Carolyn McGettigan[1] & Nadine Lavan[2]

Affiliations: 1: Speech, Hearing and Phonetic Sciences, UCL, London UK; 2: Department of Psychology, Queen Mary University of London, London, UK
Email: 1: c.mcgettigan@ucl.ac.uk; n.lavan@ucl.ac.uk

ABSTRACT

We tested whether "first impression" judgements of talkers are affected by the age of the individuals being evaluated, the age of the listeners, and the interaction of these two factors.

We presented listeners in 3 age groups (20-34, 35-49, 50-65 years old) with the vowel [a:] produced by 80 male and female talkers aged from 20 to 65. Talkers were rated on 1-9 Likert scales for perceived trustworthiness, attractiveness, and dominance.

The results showed significant main effects of both talker age (older talkers sounded less trustworthy and less attractive) and listener age (older listeners found talkers more attractive overall), but there were no significant interactions. Within-age-group agreement and cross-age-group correlations in ratings were high. Overall, our study indicates some talker-age effects but minimal listener-age effects on trait perception across the lifespan.

Keywords: Voice, identity, age, traits, first impressions.

1. INTRODUCTION

There is considerable evidence that human listeners can make judgements of person characteristics from voices, including physical attributes (e.g. sex, age, height), traits (e.g. attractiveness, dominance), and social characteristics (e.g. educatedness, professionalism). These "first impression" judgements can be made rapidly and with high agreement [1, 2], on stimuli including vowels, words, phrases, and longer passages [3,4]. When considering trait perception, it has been found that first impressions of voice stimuli can influence not only explicit trait judgements but also decision-making and behaviours toward other individuals – for example, electoral voting choices [5].

For this study, we were interested in the extent to which voices of different ages may yield differing first impressions, as evidence of stereotyping might have important social and economic implications. Several previous studies have focused on how accurately listeners can perceive voice age itself, finding that age perception is to some degree possible [6,7]. However, age estimates are not always precise: for example, listeners tend to overestimate the ages of younger adult voices, while underestimating the ages of older adult voices. There is also some suggestion of effects of *listener age* on judgements – a number of studies have reported an "own-age" bias, where listeners are more accurate at estimating the age of voices from their own age group, versus groups older or younger than themselves (although this effect is not always replicated). See [6] for a detailed review.

Aside from studies directly examining the perception of age, few studies of other aspects of talker first impressions have measured lifespan effects on perception of other talker characteristics. One study found that young listeners associate fewer positive stereotypes with elderly voices [8], while another found that perceived older age in talkers was associated with negative stereotyping, such as being ill-natured, frail, and subdued [9]. However, we are not aware of any previous studies that have comprehensively examined the relationships between listener age and talker age in the formation of trait impressions from voices.

We therefore designed a study to test for effects of talker age, listener age, and their interaction on the formation of trait impressions. Groups of young, middle-aged, and older adults listened to vowels produced by talkers of matched age ranges and provided ratings of their perceived attractiveness, trustworthiness, and dominance (previously identified in a 3-dimensional model underlying first impressions of faces [10]). We predicted that we would find main effects of talker age group (e.g. stereotyped responses where older voices are perceived as less dominant), main effects of listener age group (e.g. positivity bias in older listeners; [11]), and significant interactions of talker age and listener age (e.g. young and older listeners trait ratings showing different trends in relation to talker age; own-age biases [12]).

2. METHOD

2.1. Participants

Listeners were recruited in 3 age groups: 20-34 years old, 35-49 years old, and 50-65 years old. After

exclusions (see section 2.4.3) a final sample of 116 participants included 39 participants aged 20-34 years old (Mean age = 26.7, s.d. = 4.3; 20 female), 39 participants aged 35-49 years old (Mean age = 41.4, s.d. = 3.5; 23 female), and 38 participants aged 50-65 years old (Mean age = 53.8, s.d. = 4.3; 27 female). Participants were recruited from Prolific (prolific.co) and paid at a rate of £7.50 per hour. All participants were native speakers of English, had normal or corrected-to-normal vision, and had no self-reported hearing difficulties. The study was approved by the local Ethics Committee.

2.2. Materials

Audio clips of the sustained vowel [aː] were retrieved from the Saarbrücken Voice Database (http://stimmdb.coli.uni-saarland.de/). We chose to study vowel perception as a first step, and to avoid presenting German connected speech samples to English-speaking listeners. Eighty clips (40 female, 40 male) were selected, covering an age range from 20 to 65. This included 24 voices (12 female) in each of the age ranges 20-34 and 35-49, and 32 voices (16 female) in the age group 50-65.

In Audacity [13], the experimental stimuli were each trimmed to 500ms with a ramp on and ramp off of 25ms. Stimuli were then root-mean-square amplitude normalised in Praat [14], and converted to mp3 format using the MediaHuman Audio Converter.

A further 5 mp3 files featured a synthetic voice saying one of the digits 1-5, for use in vigilance trials. These were root-mean-squared amplitude normed to the experimental stimuli as above.

2.3. Design and Procedure

Participants completed the experiment in the Gorilla Experiment Builder [15]. After providing informed consent, participants completed a headphone volume calibration before proceeding to the main experiment.

The experiment was divided into 3 rating blocks: Attractiveness, Trustworthiness, and Dominance. The order of the blocks was counterbalanced across participants. There were two trial types within each block. The main experimental trials involved a single presentation of a vowel sound followed by onscreen presentation of a visual rating scale from 1-9. Participants used this scale to give their trait rating for that voice, where 1 = "not at all Attractive/Trustworthy/Dominant" and 9 = "very Attractive/Trustworthy/Dominant". Although the task was self-timed, participants were instructed to give their first impression or "gut feeling" when providing a rating. The second trial type was the vigilance trial: When hearing a spoken digit, participants were instructed to respond with the matching number on the rating scale. There were 80 experimental trials (1 per voice) and 5 vigilance trials (1 per digit 1-5) per block; trial order was fully randomised within blocks.

After completing the three rating blocks, participants finally answered a debrief questionnaire, in which they could report technical issues and other observations.

2.4. Statistical Analysis

2.4.3 Data cleaning

Data were prcoessed and analysed using R in RStudio (Version 2022.07.2).

Data were inspected for quality prior to performing statistical analyses. Four participants who failed more than 20% of the vigilance trials were removed, as was 1 participant who used the same rating for >80% of trials for any rating scale.

2.4.2 Inter-rater reliability

The intra-class correlation coefficient (ICC) was used as a measure of inter-rater reliability. After [16], the ICC2k was calculated on raw ratings per trait scale and listener group, using *ICC* function within the *psych* package in RStudio.

2.4.3 Inter-group correlations

The mean trait ratings per talker were calculated by trait scale and listener age group. These values were entered into three Pearson correlations per scale to exhaustively calculate correlations across the 3 listener age groups (i.e. 20-34 vs 35-49; 35-49 vs 50-65; 20-34 vs 50-65).

2.4.4 Effects of talker age and listener age on mean trait impressions

We used generalised linear mixed models (GLMMs) to investigate the effects of listener age, talker age, and their interaction on trait ratings, using the *lme4* package in R. Listener age group (20-34, 35-49, 50-65) and talker age group (binned into categories: 20-34, 35-49, 50-65) were modelled as fixed factors, with stimulus and voice gender as random intercepts. Pairwise comparisons were computed in *emmeans*.

3. RESULTS

3.1. Inter-rater reliability

Table 1 shows the mean intra-class coefficient (ICC) values by trait scale and listener age group. ICCs

above 0.75 indicate good reliability, and values above 0.9 signal excellent reliability [17].

Table 1: Intra-class coefficients for trait ratings, by scale and listener age group.

Trait	Listener Age	ICC	
		Mean	CIs
Attractiveness	20-34	0.857	0.817 - 0.893
	35-49	0.915	0.891 - 0.936
	50-65	0.846	0.801 - 0.884
Trustworthiness	20-34	0.792	0.735 - 0.843
	35-49	0.763	0.698 - 0.821
	50-65	0.712	0.634 - 0.782
Dominance	20-34	0.864	0.826 - 0.896
	35-49	0.861	0.821 - 0.896
	50-65	0.881	0.847 - 0.910

3.2. Inter-group correlations

Table 2 shows Pearson correlation coefficients and significance values across age groups for the 3 evaluated traits. Ratings were significantly correlated across all three age groups, with little evidence that listener groups who more similar in age (20-34 vs 35-49; 35-49 vs 50-65) show more strongly correlated ratings than groups more distant in age (20-34 vs 50-65). Notably, correlations for trustworthiness are overall lower than correlations for the other traits.

Table 2: Inter-group correlations of mean trait ratings. ** indicates p < .001

Trait	Inter-Group Correlations		
	20-34 vs. 35-49	35-49 vs. 50-65	20-34 vs. 50-65
Attractiveness	0.935**	0.922**	0.894**
Trustworthiness	0.769**	0.577**	0.646**
Dominance	0.888**	0.900**	0.842**

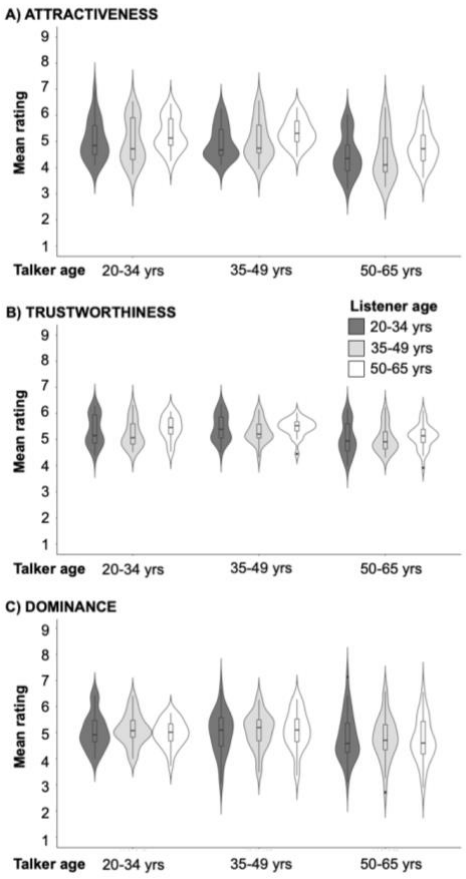
3.3. Effects of talker age and listener age on mean trait impressions

Figure 1 shows plots of mean Attractiveness, Trustworthiness, and Dominance ratings, by talker age and listener age.

3.2.1 Attractiveness

There was a significant main effect of listener age ($\chi2[2] = 71.55$, $p < .001$). The oldest listener group (50-65) gave significantly higher attractiveness ratings than the two younger groups (*p*s < .001). There was also a significant main effect of talker age ($\chi2[2] = 17.64$, $p < .001$). The oldest voices (50-65) were rated as significantly less attractive than the younger voices (*p*s < .001). There was no significant interaction between listener age group and talker age group ($\chi2[4] = 2.94$, $p = .569$).

Figure 1: Mean ratings by trait, talker age and listener age



3.2.2 Trustworthiness

There was a significant effect of listener age group ($\chi2[2] = 10.81$, $p = .005$), but within this none of the pairwise comparisons was significant (ps > .099). There was also a significant main effect of talker age ($\chi2[2]= 71.55$, $p < .001$). The oldest talkers (50-65) were rated as significantly less trustworthy than the other talker groups (ps < .003). There was no significant interaction between listener age group and talker age group ($\chi2[4] = 2.52$, $p = .641$).

3.2.3 Dominance

There were no significant effects of listener group or talker group, nor any significant interaction of the two factors (all $\chi2$s < 3.36, ps > .187).

4. DISCUSSION

We conducted a study investigating the effects of talker age, listener age, and their interaction in the formation of first impressions from voices.

Effects of talker age varied by trait category. For Attractiveness, we replicated findings from the face perception literature, which found that more attractive faces tend to have a youthful appearance [10].

For Trustworthiness, we found a significant effect of talker age, where the oldest talkers were perceived as *less* trustworthy. This trend opposes previous reports in the first impressions literature, where for example more trustworthy faces are associated with older appearance [10]. However, Sutherland's [10] face morphing approach found that more trustworthy faces were both older *and more feminine* in appearance – thus, a positive relationship between age and trust may be dependent on additional characteristics. Alternatively, the cues to ageing in voices may truly yield less positive impressions than apparent ageing in faces. Future work might resolve this issue by comparing explicit trait perceptions across faces and voices using a lifespan sample.

For Dominance, we found no effect of talker age on trait perceptions. The ICC values suggested good inter-rater agreement on dominance judgements, and exploratory inspection of the ratings broken down by talker sex also indicated that listeners reliably rated dominance higher for male than female talkers. Therefore, our interim conclusion is that dominance is perceptible in our talker sample, but impressions of this trait are unaffected by age-related stereotypes.

Attractiveness impressions may be more closely linked to physical characteristics than trustworthiness or dominance. While trustworthiness is conceptually complex and potentially context-dependent [18], attractiveness percepts from voices have been shown to be correlated with actual reproductive success [19]. The relationship between vocal age (as an index of mate fitness) and perceived attractiveness may therefore be more direct than that seen for other traits.

We found a significant effect of listener age group on ratings of talker attractiveness and trustworthiness. Specifically, listeners in the oldest age group rated the voices higher in attractiveness than the listeners in the two other age groups. This finding echoes the "positivity bias" previously reported for older adults [11]. However, this "positivity bias" did not extend to impressions of dominance or trustworthiness.

Our GLMMs revealed no significant interactions of talker and listener age, suggesting that the influence of talker age on trait perception operated similarly regardless of listener age. A set of correlation analyses further supported this conclusion: Listeners across the three listener age groups showed highly correlated talker trait ratings, on all three scales, again suggesting negligible differences in how listeners of different ages ranked the voices. However, there was some evidence of lower correlations between different age groups for trustworthiness. Trustworthiness was also the scale yielding the lowest ICC values, potentially reflecting the conceptual complexity of this trait compared with attractiveness and dominance.

These findings add valuable new evidence in relation to previous claims of listener-talker interactions in the perception of voices. As mentioned in the Introduction, the "own-age bias" reported for age perception is not always replicated. As human social groups are typically multi-generational, it is presumably adaptive for listeners of all ages to be similarly sensitive to socially-relevant information across the lifespan.

There are several possible limitations to the current study: The perception of age from voices is imperfect, thus an alternative way to examine our research questions would be to define talker age based on *perceived* age of the current voice sample. Future work should therefore explore the effects of both chronological and perceived talker age on trait perception, and in relation to listener age.

We used a single sustained vowel as the stimulus material representing each talker. Vowels have the advantage of being short and relatively steady-state samples of the talkers' vocal physiology. However, these stimuli minimise other potentially useful cues, such as speech rate and quality of articulation, which could more reliably communicate cues to physical age. While the current study showed only weak talker-age effects on dominance perception from vowels, a replication using spoken phrases may yield stronger cues to age and thus be more sensitive to detect age-related stereotyping of voices. That said, recent evidence for consistency of trait impressions across stimulus types suggests our results should be generalisable to more realistic speech [e.g. 4].

5. CONCLUSIONS

Our study is one of the first to directly examine how the ages of talkers and listeners might interact in the formation of first impressions of traits from voices.

Our findings have potential societal implications. For example, in call-centre occupations, age-related stereotyping of a voice may lead to less positive evaluations of older workers - if recognised by employers, our findings could help address inequities in the workplace. Future work measuring first impressions of age, and including more naturalistic spoken samples, will help to establish the magnitude and generalisability of these effects.

6. ACKNOWLEDGEMENTS

CM is funded by a Research Leadership Award from The Leverhulme Trust (Grant No. RL-2016-013). NL

is funded by a Sir Henry Wellcome Fellowship from Wellcome (Grant No 220448/Z/20/Z).

7. REFERENCES

[1] McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS ONE 9*(3), e90779.

[2] Mileva, M., & Lavan, N. (2022). Trait Impressions from Voices Are Formed Rapidly Within 400ms of Exposure. *PsyArXiv*. doi:10.31234/osf.io/zd4un

[3] Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS ONE* 13(10), e0204991.

[4] Groyecka-Bernard, A., *et al.* (2022). Do voice-based judgments of socially relevant speaker traits differ across speech types? *Journal of Speech, Language, and Hearing Research* 65(10), 3674-3694.

[5] Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PLoS ONE* 10(8), e0133779.

[6] Moyse, E. (2014). Age estimation from faces and voices: A review. *Psychologica Belgica* 54(3), 255-266.

[7] Lavan, N. (2022). The effect of familiarity on within-person age judgements from voices. *British Journal of Psychology* 113(1), 287-299.

[8] Lee Hummert, M., Mazloff, D., & Henry, C. (1999). Vocal characteristics of older adults and stereotyping. *Journal of Nonverbal Behavior* 23(2), 111-132.

[9] Mulac, A., & Giles, H. (1996). 'You're Only As Old As You Sound': Perceived Vocal Age and Social Meanings. *Health Communication* 8(3), 199-215.

[10] Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* 127(1), 105-118.

[11] Zebrowitz, L. A., Franklin Jr, R. G., Hillman, S., & Boc, H. (2013). Older and younger adults' first impressions from faces: similar in agreement but different in positivity. *Psychology and Aging* 28(1), 202-212.

[12] Zebrowitz, L. A., & Franklin Jr, R. G. (2014). The attractiveness halo effect and the babyface stereotype in older and younger adults: Similarities, own-age accentuation, and older adult positivity effects. *Experimental Aging Research* 40(3), 375-393.

[13] Audacity Team (2022): Audacity [Computer program]. Version 2.4.2, retrieved Dec 19, 2022, from http://audacityteam.org/

[14] Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved Dec 19, 2022, from http://www.praat.org/

[15] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1), 388-407.

[16] Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 86(2), 420-428.

[17] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2), 155-163.

[18] Knight, S., Lavan, N., Kanber, E., & McGettigan, C. (2018). The social code of speech prosody must be specific and generalizable. *Proceedings of the National Academy of Sciences* 115(27), E6103-E6103.

[19] Pisanski, K., & Feinberg, D.R. (2018) Vocal Attractiveness. In S. Frühholz, and P. Belin (eds), *The Oxford Handbook of Voice Perception* (pp.607-626). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.27, accessed 20 Dec. 2022.

WHAT CAN VOICE LINE-UPS TELL US ABOUT VOICE SIMILARITY?

Renata R. Passetti[1], Sandra Madureira[2], Plínio A. Barbosa[3]
[1]Postgraduate Program in Linguistics, UFSCar, Brazil, [2] Graduate Program in Applied Linguistics and Language Studies, PUCSP, Brazil, [3]Dep. of Linguistics, UNICAMP, Brazil
re.passetti@gmail.com, madusali@pucsp.br, pabarbosa.unicampbr@gmail.com

ABSTRACT

This study considers the outcomes of a voice line-up experiment to shed light on the acoustic-prosodic features related to voice similarity. Data analysis comprises a similarity function based on acoustic parameters, and a multidimensional technique based on perceptual evaluation of voice quality. Target and foil stimuli are compared in terms of a set of 18 prosodic-acoustic parameters. The voice quality settings of the target, the most chosen foil and the non-chosen foil are also analysed according to the Voice Profile Analysis (VPA) and contrasted by a multidimensional scaling technique. The timing interval between stimuli presentations had great impact on voice identification correctness. The f0 parameters regarding melody were influential for establishing correct voice identifications. Voice quality settings were relevant regarding to false alarms, and voice dynamics settings for correct stimuli rejections. Results also highlight the importance of segmental and speaking tempo parameters to perceived voice similarity.

Keywords: voice line-up, voice similarity, VPA, similarity function

1. INTRODUCTION

Voice similarity is a pivotal topic in forensic phonetics. Knowing the phonetic features which can make voices sound similar is of special interest for forensic tasks concerning earwitness evidence. In such situations, voice similarity can be exploited both to improve procedures for preparing voice line-ups as well as to ensure fair identification when conducting them [1].

Regarding voice line-up construction, the analysis of voice similarity is crucial for selecting foil speakers, i.e., volunteers whose speech samples will be added to the voice line-ups. According to a set of guidelines on voice line-ups developed for application in the Dutch forensic context [2], the recording of foils should meet the description of the suspect's and perpetrator's voices in terms of physiological, prosodic, and sociolinguistic features.

Another key factor for voice similarity is related to the kind of parameters which may play a role on its perceptual assessment. This question was addressed by Lindh [3] in a study that compared the outcomes of a perceived voice similarity test to those of a line-up experiment. The results showed voice similarity judgments were partially explained by similarities in speaking tempo parameters, such as articulation rate and pausing measurements.

In a follow-up study, Lindh and Eriksson [4] compared the outcomes of voice similarity judgments made by listeners to those of automatic systems to explore convergence and divergence between them. The results pointed out to the influence of linguistic factors, such as speaking style, on listeners' judgments, which do not seem to play a role in the automatic systems analysis. Furthermore, the authors suggest that when performing line-ups tasks, listeners may also pay attention to other linguistic factors, such as pronunciation and voice quality.

The acoustic correlates of perceived voice similarity were also topic of a study by Nolan et al. [5]. To avoid interference of linguistic factors on listeners' perception, speech samples from the same dialect were used in a paired comparison test. Correlations between the perceptual dimensions of similarity judgments and the set of acoustic variables have shown the major importance of fundamental frequency in perceived voice similarity. Other acoustic parameters in order of importance from higher to lower are F3, F2 and F1.

The present study considers issues raised in the literature on voice similarity and uses the outcomes of a voice line-up experiment to shed light on the prosodic features related to perceived voice similarity. Some of our research questions are: What prosodic cues may play a role in listeners performance on voice line-ups? In misleading identifications, are there prosodic similarities between the target and the most chosen foil voices? Why were some foils not chosen?

The methodological approach comprises perceptual and acoustic analysis of speech samples, which were used in perceptual tasks performed by attendants in a workshop on voice similarity held by the authors.

2. METHODOLOGY

2.1. Voice database

The speech samples used in the perceptual experiment were selected from the Corpus Forense do Português Brasileiro (henceforth CFPB). This corpus

has been compiled by the Audiovisual and Electronics Section of the Criminalistics Institute of the Brazilian Federal Police and consists of text-reading and semi-spontaneous recordings of officials from Brazilian law enforcement agencies.

At the time of the study the corpus contained 280 recordings from male speakers and 70 from female speakers, aged between 19 and 64 and from five regions in Brazil (North, Northeast, Central-West, Southeast and South). All the CFPB recordings are recorded at a sampling rate of 44.1 kHz.

2.2. Voice line-up experiment

2.2.1. Line-ups

The perceptual experiment comprised four voice line-ups. The line-ups were constructed following the guidelines proposed by [2] and consisted of the target voice randomly arranged with five other distractor voices.

The selection criteria of the 24 voices (six for each line-up) were mainly based on the availability of the samples in the CFPB. Therefore, we selected semi-spontaneous recordings of male speakers originally from Brazilian states with more than 10 samples at the corpus.

The chosen states were Ceará (CE) in the Northeast, São Paulo (SP) and Rio de Janeiro (RJ) in the Southeast, and Paraná (PR) in the South. For each state, the target and distractor voices were selected based on their similarity regarding the mean of f0, F1 and F2. This analysis was performed automatically by a Praat script developed by the third author.

The line-up stimuli lasted twenty seconds, while the target stimuli lasted forty seconds. We ensured that the content of the target speakers' speech in their first presentation was different from the content of their speech within the line-up since similar content could influence listeners' choices.

Finally, another Praat script was developed to automatically arrange the stimuli along the line-ups and normalizing them to the same level of intensity. Moreover, this script added a five-second-long silent pause between the stimuli.

The position of the target stimulus in each line-up is shown in Table 1.

Line-up	Target position
Ceará (CE)	#1
São Paulo (SP)	#3
Rio de Janeiro (RJ)	#5
Paraná (PR)	#5

Table 1: Position of the target stimuli in the line-ups.

2.2.2. Experimental procedure

The workshop practices were held over two days separated by a one-week interval. Two voice line-ups were performed in the first day, and two in the second day. Therefore, the experimental procedure consisted of running two voice line-ups (from CE and SP) immediately after the target voice exposure, one voice line-up (from RJ) one week after the target voice exposure, and one voice line-up (from PR) 30 minutes after the target voice exposure. The target and line-up voices were played only once, and the attendees registered their choices on an online form.

2.2.3. Listeners

From the 21 participants who attended the workshop, the sixteen who attended it on both days were included in the present study.

They were 11 men and 5 women, aged between 20 and 64 years (mean of 34 years). All of them were higher educated (completed or ongoing) in speech-related courses, and three of them were forensic experts. None of them reported cognitive or hearing impairments.

2.3. Voice similarity analysis

2.3.1. Acoustic analysis

In order to verify whether prosodic-acoustic similarity reflects the listeners' choice, the third author implemented a function in R that computes this similarity according to three different criteria. Before that, the audio files were segmented into chunks separated by silent pauses.

Acoustic measures were computed with the Praat script "Prosody Descriptor Extractor" [6] on all chunks. This script extracts twelve fundamental frequency (f0) descriptors, two intensity descriptors and four voice quality descriptors. Among them, the following were significant in this study: f0 median, f0 semi-amplitude between quartiles, f0 minimum, f0 maximum, standard deviations of f0 local peak values, mean f0 peak rate, mean f0 peak bandwidth, mean f0 rates of rises and falls, spectral emphasis [7], Harmonic-to-Noise ratio (HNR) and jitter.

The similarity function gives the order of foils similar to the target starting from the most similar to the less similar. Three criteria of computing the distance between the target and a foil are used: (1) the first one computes the sum of absolute errors between target and a foil for each parameter where the error is computed as the difference between the medians of the corresponding parameters divided by the median of that parameter in the target voice; (2) the second one computes the sum of absolute errors between

target and a foil for each parameter where the error is computed as the difference between the medians of the corresponding parameters divided by the pooled standard-deviation of target and foil for that parameter; and (3) the third criterion arranges the foils in descending order according to the number of parameters with minimum error as established by criterion 1.

2.3.2. Perceptual analysis

In order to verify the role of voice quality (henceforth VQ) and voice dynamics (henceforth VD) settings on listeners' performance, the first and second authors assessed the target, the most chosen foil and the non-chosen foil stimuli for the SP, RJ and PR line-ups by means of the Vocal Profile Analysis (VPA) [8]. The CE line-up stimuli were discarded from this analysis since all listeners correctly chose the target's stimulus.

After having evaluated the stimuli separately, the first and second authors came up with an agreed version of the VPA for each of those stimuli. Then, the agreed versions of the VPA were subjected to the Multidimensional Scaling technique (MDS) [9, 10], which was performed by means of the R *smacof* package [11, 12]. This analysis was carried out in two steps: firstly, the VQ settings (i.e. vocal tract, muscular tension, and phonation features), and secondly, the VD settings (i.e. pitch, loudness, and speech rate features) as input parameters.

3. RESULTS AND DISCUSSION

3.1. Voice line-up experiment

The number of choices for each stimulus in the line-ups is shown in Figure 1. In all the line-ups, the target voices were the most correctly identified ones.

The better performance of the CE line-up can be explained by a combination of the latency time after the target voice exposure (immediately after it) and its position in the line-up (first position). The RJ line-up, on the other hand, had the worst performance, since only six listeners correctly chose it. This score can also be explained by the latency time of one week between the voice target exposure and the line-up running.

Some stimuli were not chosen by anyone. This was the case for all stimuli except the first (i.e. the target) in the CE line-up, the fifth stimulus in the SP line-up, the sixth in the RJ line-up and the second in the PR line-up.

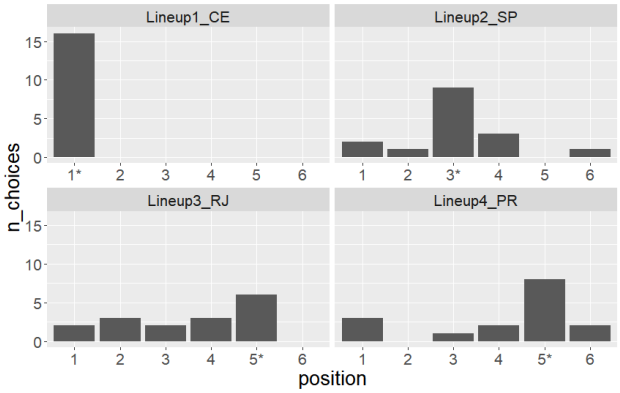


Figure 1: Number of choices for each stimulus position in the line-ups. Target positions are identified with an asterisk.

3.2. Acoustic analysis

The use of criterion 3 for the similarity function was the best predictor of the listeners' choices. For the purposes of analysis, in each line-up the two most similar foils to the target have been considered.

For the SP and RJ line-ups the function output predicted the target as the first choice. For the PR line-up, the target was the second function choice (foil 2, PR_2, being the first choice). The parameters with a minimum error according to criterion 3 which were common to the two correct choices were f0 semi-amplitude between quartiles, f0 maximum, mean f0 rates of rises and falls, and voice quality parameters (spectral emphasis for the SP line-up, and HNR and jitter for the RJ line-up). For the PR line-up, the parameters with a minimum error were f0 median, f0 minimum, standard deviations of values of f0 local peaks, mean of f0 peak rate, and jitter, and those associated with the second place (i.e. the target) were mean f0 peak bandwidth, mean f0 rates of rises, spectral emphasis, and HNR.

For all the line-ups, the melodic patterns appear as a relevant parameter to being used for prosodic-acoustic similarity purposes.

3.3. Perceptual analysis

The bi-dimensional plots for VQ and VD settings are shown in Figure 2. Dimensions D1 and D2 account for the transformation of the VPA scores into fitted distances in the MDS models. Data from RJ has two most chosen foil stimuli because both were equally chosen.

The normalized stress for the VQ analysis is 0.024, whereas for the VD analysis is 0.039, both considered acceptable values [12].

The analysis of the stimuli spatial arrangement sheds some light on listeners' false alarms, as well as on the possible reasons why some foils were not chosen.

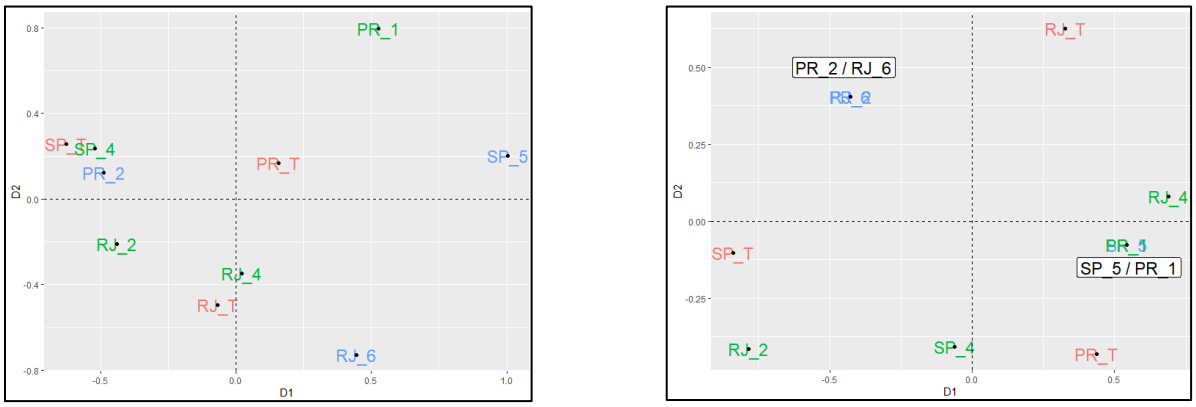


Figure 2: Bi-dimensional plots of the MDS analysis for the VQ (on the left side) and VD (on the right side) settings. Same line-up stimuli are identified by the acronyms of their states. The colours red, green, and blue stand for the target, the most chosen foil, and the non-chosen foil stimuli, respectively.

The comparison of the stimuli arrangement suggests that VQ may account for most of the false alarms in the SP ($d_{(SP\_T,SP\_4)} = 0.11$) and RJ ($d_{mean(RJ\_T, RJ\_2, RJ\_4)} = 0.47$) line-ups, whereas VD may account for most of the false alarms in the PR line-up ($d_{(PR\_T,PR\_1)} = 0.37$), since the target and the most chosen foil stimuli are closer to each other in the respective plots.

Regarding the non-chosen foil stimuli, it can be observed, in both plots, that the target and non-chosen foil stimuli pairs were placed on opposite sides of D1. The mean distances of these stimuli pairs suggest that VQ may also have been a relevant cue to perceived voice dissimilarity judgments in the SP line-up ($d_{(SP\_T,SP\_5)} = 1.63$), whereas VD may have been relevant to judging voice dissimilarity in the RJ ($d_{(RJ\_T,RJ\_6)} = 1.30$) and PR ($d_{(PR\_T,PR\_2)} = 1.20$) line-ups.

It is also worth mentioning that some stimuli were overlapped in the VD plot, which suggests that these speakers share similarities regarding VD settings. The "PR_2" and "RJ_6" stimuli, for example, share neutral settings for pitch and loudness variabilities, and for speech rate. The "SP_5" and "PR_1" stimuli, on the other hand, share non-neutral scalar degrees for minimized pitch range, low mean loudness, and slow rate.

4. CONCLUSION

Although it is very difficult (to not say almost impossible) to predict listeners' strategies when judging voice (dis)similarity, exploring prosodic parameters can give us some guidance. Therefore, the combination of acoustic and perceptual procedures proves to be complementary and necessary, as they delve into perception-production links.

In this study, the acoustic analysis showed that intonational, especially those regarding melody (e.g. mean f0 rates of rises and falls), and voice quality features are relevant for an overall voice similarity judgment. The perceptual analysis, on the other hand, shed some light on false alarms as well as on the non-chosen voices by analyzing voice quality and voice dynamics features separately. Moreover, it also added to the acoustic analysis by considering the analysis of speech rate features.

Special attention can be given to the PR_2 foil, which has not been chosen by any listener but was the first choice of the automatic analysis (i.e., the similarity function) for the PR line-up. These results highlight the importance of two aspects not considered in the acoustic analysis carried out here. The first one is the analysis of speaking tempo parameters, whose relevance to voice similarity had already been addressed by [3]. Speech rate probably also played a role among the voice dynamics settings in distinguishing this foil from the target, as observed in the VD plot. The other one is the analysis of segmental parameters, as mentioned by [4], which may signal sociolinguistic as well as idiosyncratic cues. In this specific case, the PR_2 foil produced an allophone of /t/ different from the target, and as this is a salient feature in some Brazilian dialects, it may have been one of the cues on which listeners relied to discard him. Therefore, speaking tempo as well as segmental parameters should be addressed in future studies.

Although voice line-ups are not part of the forensic casework in some countries, we believe they are useful to assess voice similarity in the forensic context. Furthermore, voice line-up analysis opens up possibilities to discuss why people confuse voices.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -

Brasil (CAPES) - Finance Code 001. Therefore, the authors acknowledge a grant from CAPES/PROCAD/Segurança Pública e Ciências Forenses #88887.516306/2020-00. The first author acknowledges a grant from CAPES #88887.804443/2023-00. The second author acknowledges a grant from PIPEq #21672, PUCSP.

6. REFERENCES

[1] McDougall, K. 2013. Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, *20*(2).
[2] Broeders, A. P. A., Van Amelsvoort, A. G. 1999. Lineup construction for forensic earwitness identification: A practical approach. In *Proceedings of the 14th International Congress of Phonetic Sciences* 2, 1373-1376).
[3] Lindh, J. 2009. Perception of voice similarity and the results of a voice line-up. *FONETIK 2009*, 186.
[4] Lindh, J., Eriksson, A. 2010. Voice similarity-a comparison between judgements by human listeners and automatic voice comparison. *Working papers/Lund University, Department of Linguistics and Phonetics*, 54, 63-68.
[5] Nolan, F., McDougall, K., Hudson, T. 2011. Some Acoustic Correlates of Perceived (Dis) Similarity between Same-accent Voices. In *ICPhS* , 17, 1506-1509.
[6] Barbosa, P. 2021. *Prosody Descriptor Extractor* [Praat script]. Available: https://github.com/pabarbosa/prosodyscripts/tree/master/ProsodyDescriptorExtractor
[7] Traunmüller, H., Eriksson, A. 2000. Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, 107(6), 3438-3451.
[8] Mackenzie-Beck, J. 2007. Vocal profile analysis scheme: A user's manual. *Queen Margaret University College–QMUC, Speech Science Research Centre*, Edinburgh.
[9] Young, F. W. 1985. Multidimensional scaling. *Encyclopedia of Statistical Sciences*, University of North Carolina.
[10] Wickelmaier, F. 2003. An introduction to MDS. *Sound Quality Research Unit*, Aalborg University, Denmark 46.5, 1-26.
[11] de Leeuw, J., Mair, P. 2009. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 1–30. Available: https://www.jstatsoft.org/v31/i03/
[12] Ashkiani, S. 2017. *Dimensionality Reduction and Data Visualization Using MDS-SMACOF package in R*. Available: https://rstudio-pubs-static.s3.amazonaws.com/246348_b31bca1e4be04bb395825dc6a00de364.html

"YOUTUBE SPACE"—A PRELIMINARY INVESTIGATION OF VOWEL SPACES AND CHARISMA PERCEPTION ON YOUTUBE

Stephanie Berger[1], Margaret Zellers[1], Oliver Niebuhr[2]

[1]ISFAS, Kiel University, Germany [2]CIE, University of Southern Denmark, Denmark
sberger@isfas.uni-kiel.de, mzellers@isfas.uni-kiel.de, olni@sdu.dk

ABSTRACT

This study investigates correlations between vowel space size (measured as the mean formant ranges of the first two formants F1 and F2) and listener ratings of charisma and charisma-adjacent features (*authentic*, *enthusiastic*, *likeable*, and *persuasive*). The results show that for stimuli from a perception experiment, speakers with larger F1 ranges are perceived as more charismatic, but at the same time less authentic. The opposite is the case for a full 50-minute data set: For this data set, speakers with a smaller F1 range were perceived as more charismatic. The results differ from previous research suggesting that charisma on YouTube and other genres like business speeches may work differently.

Keywords: YouTube, charisma, perception, formants, vowel space

1. INTRODUCTION

Charisma (see [1] for an overview of the concept of charisma), i.e. the ability to draw in an audience, is incredibly important for online content creators like YouTubers. In a digital world where billions of videos are only one click away, audiences need to be kept engaged to stay and come back to a channel.

Phonetic charisma research has identified a large number of prosodic features that increase perceived charisma: a wider pitch range, faster speech rate, frequent use of emphatic accents, or a frequent use of high pitch accents, among others (e.g. [2, 3, 4, 5, 6, 7, 8, 1, 9]). News outlets that popularized the term "YouTube Voice" also include "overstressed vowels" in their assessment of the particular speaking style in YouTube vlogs [10, 11]. (Vlogs are video blogs in which a YouTuber speaks to the audience. This is semi-spontaneous speech, as it is not scripted, but planned and edited.)

The prosodic features mentioned above (as well as others, see [8] for an overview of features investigated for charismatic speech) can be connected to increased vocal effort, which in turn tends to be correlated with increased charisma in business and political speakers [1]. Increased vocal effort and increased charisma go hand in hand with clear, minimally reduced speech, and "clear speech includes 'an expansion of the vowel space'" [12, p.346]. [13] also found that speakers in psychological distress (e.g. depression) speak with significantly reduced vowel spaces compared to healthy speakers, which they correlate with decreased expressivity. That suggests the opposite—expanded vowel spaces—goes in hand with increased expressivity.

Furthermore, [14] studied differences in vowel space size and mean ranges of the first and second formants (F1 and F2) in L2 English produced by L1 German speakers. Listeners rated the speakers on five attributes: *charismatic*, *passionate*, *decided*, *trustworthy*, and *captivating*. The study found significant positive correlations between the listener ratings of all attributes and speakers' F1 and F2 ranges. This means that a larger formant range on either vowel space dimension was rated higher [14]. Large vowel spaces were also perceived as most charismatic, small vowel spaces as least charismatic by listeners. Additionally, when the vowel space was compressed in the F2 (back–front) dimension, speakers were perceived as less trustworthy and less decided, while they were perceived as less passionate and captivating if the F1 (closed–open) dimension was small [14]. Thus, in [14], the F1 range was more relevant for attributes related to emotions, while the F2 range was more closely related to attributes connected to cognition.

2. QUESTIONS AND HYPOTHESES

The current study investigates correlations between charisma ratings from a perception experiment and vowel space size, which is operationalized as the mean ranges of F1 and F2 (see Section 3.3). This study, similar to [14], includes a direct charisma rating (*charismatic*), as well as emotion-related (*likeable* and *enthusiastic*) and cognition-related attributes (*persuasive* and *authentic*). While [14] investigated business-oriented L2 English speakers, the current study investigates L1 English speaking

YouTubers who entertain their audience.

This study addresses three research questions (RQ). Do the F1 and F2 ranges (as measurements of vowel space size) correlate with the ratings of charisma and charisma-adjacent attributes (RQ1)?

Are correlations between ratings and vowel spaces similar for YouTubers and the speakers from a business context from previous literature [12, 14] (RQ2)? Differences between YouTubers and other speaker groups would indicate that charisma may be encoded depending on the medium. First indications of differences between acoustic features of charismatic speech in business and on YouTube exist: e.g., pitch range is important for charisma in business [1], but it did not correlate with the video view count (considered an approximation of charisma) in [15].

In addition, we compare the formant ranges from the entirety of annotated speech material (50 minutes) to those from the stimuli used to collect the perception ratings: Do similar correlations arise between charisma ratings and vowel space (F1 and F2 ranges) differ between the full data set and the perception experiment stimuli (RQ3)?

Four hypotheses are connected to these research questions. In line with [14], larger vowel spaces (in terms of larger F1 and F2 ranges) are expected to be perceived as more charismatic (H1). Larger F1 ranges are expected to be positively correlated with higher ratings for emotion-related attributes (*likeable* and *enthusiastic*, H2). Larger F2 ranges are expected to correlate positively with higher ratings of cognition-related attributes (*persuasive* and *authentic*, H3). We also predict that the direction of correlations matches between the two data sets (H4).

3. METHODS

3.1. Speech material and measurements

The first data set (*full data set*, *FData*) consists of 50 minutes of speech material from ten English-speaking YouTubers[1] (5 male, 5 female, 5 from North America, 5 from England, age at time of video publication: 24 to 32, mean=29.1, SD=2.64). All videos are vlogs where the speaker talks to the audience about different topics like mental health, the YouTube business, or their community; only monolog passages were used for the current study.

F1 and F2 were extracted at vowel midpoints using a Praat [16] script based on [17] using Praat's default settings for its `To Formant (robust)` function with the maximum formant at 5000 Hz for male and 5500 Hz for female speakers. The segment boundaries and their SAMPA annotations were created with WebMAUS [18] and manually corrected. Four corner vowels of the vowel space (IPA: [iː], [uː], [æ] and [ɒ]) are included in the study. Formants were normalized using the `normLobanov()` function in the `phonR`-package [19] to account for the different origins and genders of the speakers and the different recording set-ups of each YouTuber.

Mean F1 and F2 of each of the four vowels were calculated per speaker. The F1 means of [iː] were subtracted from those of [æ] for the front dimension, and [uː] was subtracted from [ɒ] for the back dimension. Afterwards, the mean of these differences was calculated to obtain a mean F1 range for the center height of the vowel space. For F2, [uː] was subtracted from [iː] for the closed dimension, and [ɒ] was subtracted from [æ] for the open dimension to get the center width of the vowel space.

The same calculations were made for a second data set (a subset of *FData*: *stimuli data set*, *SData*). It comprises four interpausal units for each speaker which were stimuli used in perception experiments (see Section 3.2). The stimuli were not initially chosen to study vowel spaces, so some substitutions were necessary: For speaker ZS, the [iː] was exchanged for /ɪ/ to obtain a complete vowel space with four corner points. Speaker PL was excluded for the *SData*, as only two of four corner points were available in the experiment stimuli and no substitutions were possible.

3.2. Perception experiment

Stimuli (both manipulated in pause duration and breathing as well as unmanipulated) were presented to participants in perception experiments and were rated on 5-point Likert scales. 20 participants (male: N=8, female: N=11, prefer not to answer: N=1, age: mean=27.5, SD=5.07) rated the stimuli in terms of charisma directly. 20 other participants (male: N=8, female: N=11, non-binary: N=1, age: mean=29.95, SD=5.23) rated the stimuli on charisma-adjacent attributes (*authentic*, *likeable*, *enthusiastic*, *persuasive*). All participants were L1 English speakers, originally from the British Isles. None reported issues with their hearing. The mean ratings of the unmanipulated stimuli per speaker and scale are used for the correlations, both with the vowel spaces of *FData* and *SData*.

3.3. Statistical analyses

Kendall's τ rank correlation coefficients were calculated between formant ranges and mean ratings. The significance level is set at α=.05. All statistical analyses were run in RStudio [20, R version 4.1.3].

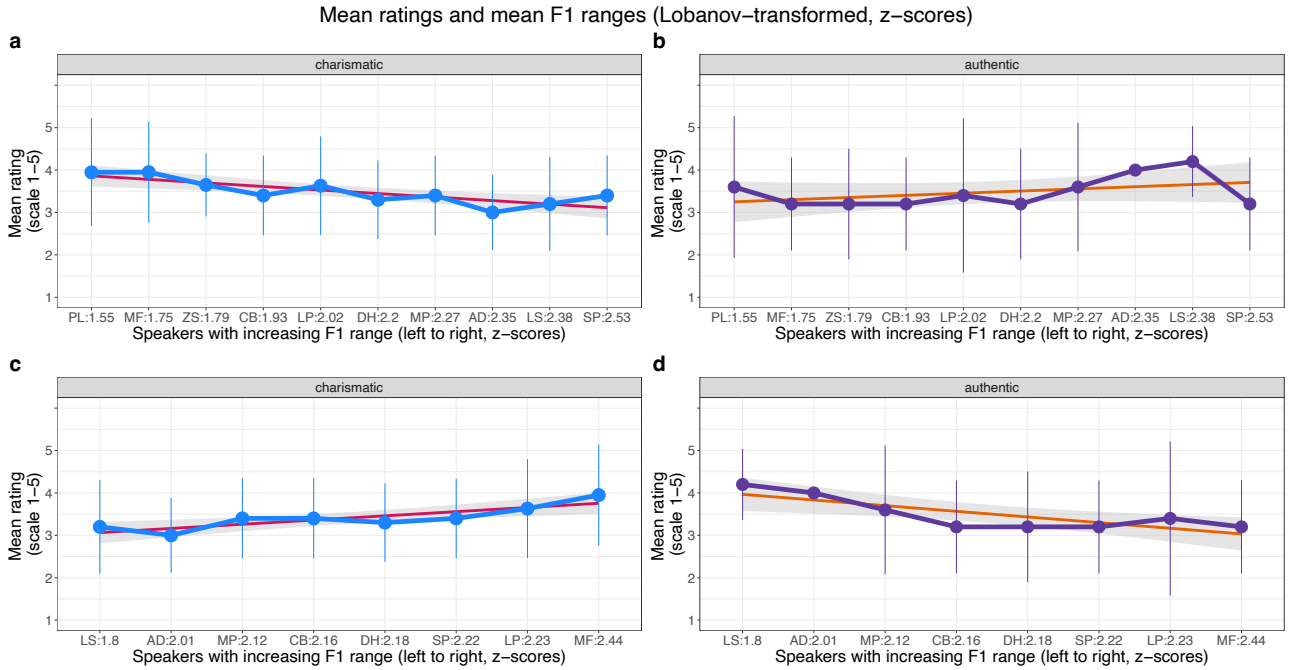


Figure 1: Data points and correlations between a) charismatic and b) authentic ratings and mean F1 range for the full data set, and c) charismatic and d) authentic ratings and mean F1 range for the stimuli. The speakers and their respective F1 ranges on the X-axis are arranged in ascending order. Whiskers represent standard deviation.

4. RESULTS

4.1. Full data set (*FData*)

There is a significant negative correlation between the *charismatic* ratings and the mean F1 range (τ = -0.64, p = .01). Speakers that have the largest F1 ranges in their videos overall are rated as less charismatic in the experiment (see Figure 1a). There are no other significant correlations for mean F1 range. Visually, there is also a tendency for a positive correlation for the *authentic* ratings (Figure 1b), and while this suggests the opposite direction than for the *charismatic* rating, this was not significant (p = .25). There are no significant correlations between the ratings on any scale and F2 range (all p-values >.1; see Table 1). There is a non-significant trend in that larger F2 ranges tend to be perceived as less likeable (τ=-0.4, p=.12).

4.2. Stimuli (*SData*)

Two significant correlations arise for the F1 ranges. The *charismatic* rating is positively correlated with the F1 range (τ = 0.72, p = .02); thus speakers with a larger F1 range are perceived as significantly more charismatic than speakers with a smaller F1 range (see Figure 1c). At the same time, there is a significant negative correlation between the *authentic* rating and the F1 range (τ = -0.64, p = .03): the larger the F1 range, the less authentic a speaker is perceived (see Figure 1d). These two correlations are opposite, but—taking into account only the visual tendency for the *authentic* ratings in the *FData*—the correlation directions between the *FData* and the *SData* are also opposite. There are no significant correlations between the ratings on any scales and the F2 ranges of the vowel spaces of the stimuli (all p-values >.1; see Table 1). However, the correlation tests reveal some trends, suggesting that speakers with larger F2 ranges tend to be perceived as more charismatic (τ = 0.49, p = .1) and persuasive (τ = 0.46, p = .12), but less likeable (τ = -0.44, p = .13).

5. DISCUSSION

This study investigated correlations between formant ranges and charisma ratings for YouTubers. RQ1 and RQ2 are only partially supported for the *SData*, not the *FData*. There are differences between the *SData* and the *FData*, so RQ3 is confirmed for the current sample.

The correlation between *charismatic* and F1 range for the *SData* in the current study was significant and positive (as in previous literature, and partially supporting H1), while the negative correlation with

Type	Attribute	F1 range (*F*)	F1 range (*S*)	F2 range (*F*)	F2 range (*S*)
charisma	*charismatic*	τ=-0.64, p =.01*	τ=0.72, p =.02*	τ=-0.05, p =.86	τ=0.49, p =.1^{+}
emotion	*enthusiastic*	τ=-0.26, p =.31	τ=0.4, p =.18	τ=-0.17 p =.52	τ=0.08, p =.79
emotion	*likeable*	τ=-0.26, p =.32	τ=-0.15, p =.62	τ=-0.4, p =.12^{+}	τ=-0.44, p =.13^{+}
cognition	*authentic*	τ=0.31, p =.25	τ=-0.64, p =.03*	τ=0.1, p =.7	τ=-0.16, p =.6
cognition	*persuasive*	τ=0.07, p =.78	τ=0.39, p =.2	τ=0.22, p =.4	τ=0.46, p =.12^{+}

Table 1: The results of the Kendall's correlation tests for each scale attribute for the mean F1 and F2 ranges (*F* = *FData*, *S* = *SData*). Significant correlations are marked with an asterisk, non-significant trends with a plus.

the *authentic* rating does not match previous findings [14]. In the *FData*, the correlations were opposite to those in the *SData*, and also opposite to those in the literature, and not supporting H1. Unlike in [14], there were no other significant correlations for F1 ranges and attributes, and no significant correlations between attributes and F2 ranges.

Neither H2 nor H3 were supported by the data. There is no evidence that larger mean F1 ranges are positively correlated with higher ratings of emotion-related attributes, as there were no significant correlations with the attributes *likeable* and *enthusiastic*. Similarly, there is also no evidence for positive correlations between F2 range and cognition-related attributes, as there were no significant differences in the data at all, though there were tendencies for the F2 ranges that seem to align with previous research. There is a tendency for speakers with larger F2 ranges to be perceived as more charismatic and persuasive, which mirrors the results in [14], and may partially align with H3. The more emotion-related rating *likeable* has a tendency to be rated lower the larger the F2 range is which does not seem to be in line with [14]. It might also be that the *authentic* rating in this current study is more emotion-related than cognition-related, though. In [14], the term *trustworthy* was explained to participants to mean "capable of living up to [...] promises" (p.270), while the *authentic* rating (taken as similar) in the current study was explained as the speaker not putting on an act. If this is the case, there was a positive correlation between F1 range and *authentic* rating in the *SData* that may partially support H2.

We further predicted (H4) that the direction of the correlations would match between the two data sets. This was not the case. Where the correlation with the *charismatic* rating was negative in the full data, it was positive in the *SData*. The (visual, not significant) correlation with the *authentic* rating in the *FData* was positive, while the (significant) correlation was negative in the *SData*. This suggests that for charisma and authenticity, formants may work well as predictors for perception of the stimuli the ratings were made on, but that there is more going on in larger data sets that cannot be extrapolated by using only the perception of a small subset. In particular, the overall shape and area of the vowel spaces differs and should be investigated in future studies.

The difference in the direction of correlation for *charismatic* and *authentic* in the *SData* might be explained by a balancing act on YouTube: Speakers with larger F1 ranges (i.e. exhibiting increased vocal effort) were perceived as more charismatic, and at the same time less authentic. YouTubers tend to try to appear approachable, friendly and genuine [21], while still having to keep an audience engaged and entertained. In the context of YouTube, speech is then likely perceived as less authentic when it is produced with more effort. This is similar to results from [22] which found that moderately reduced German speech was perceived as most sincere, compared to unreduced or reduced speech.

Charisma may be a different dimension that, on YouTube, is less tied to authenticity, as the goal is entertainment. It is also likely that YouTubers intentionally or subconsciously adjust their speech to be more colloquial, like a friendly conversation. This may coincide with what Labov calls the "principle of attention" in that speaking styles "can be ordered along a single dimension, measured by the amount of attention paid to speech. [...] Casual and intimate styles can be stationed at one end of this continuum, and frozen, ritualistic styles at the other" [23, p.112]. Vlogs may be positioned somewhere towards the casual end of the continuum, but still not at the extreme, as they are planned and edited (see also [24]).

In addition to investigating the vowel space area and shape instead of F1/F2 ranges, future studies should look at a larger sample of YouTubers. Preliminary studies have also shown that the jaw movement is connected to formants and a more consistent predictor for speaker charisma [25]. While physical measurements from YouTube videos are impossible, future studies may also include distance measurements of mouth and jaw facial landmarks calculated from videos with programs like OpenFace [26].

6. REFERENCES

[1] O. Niebuhr, A. Brem, J. Michalsky, and J. Neitsch, "What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and Mark Zuckerberg," *Cadernos de Linguistica e Teoria da Literatura*, vol. 1, no. 1, pp. 1–40, 2020.

[2] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert, "A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech," in *Speech Prosody*, vol. 37, 2008.

[3] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[4] R. Signorello, F. D'Errico, I. Poggi, D. Demolin, and P. Mairano, "Charisma perception in political speech: A case study," in *International Conference on Speech and Corpora (GSCP 2012)*, 2012, pp. 343–348.

[5] F. D'Errico, R. Signorello, D. Demolin, and I. Poggi, "The perception of charisma from voice: A cross-cultural study," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 552–557.

[6] S. Berger, O. Niebuhr, and B. Peters, "Winning Over an Audience–A Perception-based Analysis of Prosodic Features of Charismatic Speech," in *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, 2017, pp. 1454–1457.

[7] O. Jokisch, V. Iaroshenko, M. Maruschke, and H. Ding, "Influence of age, gender and sample duration on the charisma assessment of German speakers," *Proc. 29. Konferenz Elektronische Sprachsignalverarbeitung (ESSV2018)*, 2018.

[8] S. Berger, O. Niebuhr, and A. Brem, "Of voices and votes: Phonetic charisma and the myth of Nixon's radio victory in his first 1960 TV debate with Kennedy," in *An den Rändern der Sprache*, ser. Kieler Forschungen zur Sprachwissenschaft, M. Elmentaler and O. Niebuhr, Eds. Peter Lang, 2020, pp. 109–145.

[9] S. Berger and M. Zellers, "Pitch accent position, peak height, and prominence level relative to accented vowel onset on YouTube," in *1st International Conference on Tone and Intonation (TAI)*, 2021, pp. 137–141.

[10] J. Beck, "The Linguistics of 'YouTube Voice'," *The Atlantic, December 7, 2015*, 2015.

[11] S. Hagi, "The rise of the 'YouTube Voice' and why vloggers want it to stop," *Vice.com, March 28, 2017*, 2017.

[12] O. Niebuhr and S. Gonzalez, "Do sound segments contribute to sounding charismatic? Evidence from a case study of Steve Jobs' and Mark Zuckerberg's vowel spaces," *International Journal of Acoustics and Vibration*, vol. 24, no. 2, pp. 343–355, 2019.

[13] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4789–4793.

[14] O. Niebuhr, "Space fighters on stage—How the F1 and F2 vowel-space dimensions contribute to perceived speaker charisma," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 265–277, 2020.

[15] S. Berger, O. Niebuhr, and M. Zellers, "A Preliminary Study of Charismatic Speech on YouTube: Correlating Prosodic Variation with Counts of Subscribers, Views and Likes," in *INTERSPEECH*, 2019, pp. 1761–1765.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2018. [Online]. Available: http://www.praat.org/

[17] Y. Xu and H. Gao, "FormantPro as a tool for speech analysis and segmentation," *Revista de Estudos da Linguagem*, vol. 26, no. 4, pp. 1435–1454, 2018.

[18] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[19] D. R. McCloy, "phonR: Tools for phoneticians and phonologists," *R package version*, vol. 1, 2016.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[21] R. Kyncl and M. Peyvan, *Streampunks: How YouTube and the new creators are transforming our lives*. Random House, 2017.

[22] O. Niebuhr, "Clear Speech-Mere Speech? How Segmental and Prosodic Speech Reduction Shape the Impression That Speakers Create on Listeners," in *INTERSPEECH*, 2017, pp. 894–898.

[23] W. Labov, "Some principles of linguistic methodology," *Language in Society*, vol. 1, no. 1, pp. 97–120, 1972.

[24] S. Lee, "Style-shifting in vlogging: An acoustic analysis of "YouTube voice"," *Lifespans & Styles: Undergraduate Working Papers on Intraspeaker Variation*, vol. 3, no. 1, pp. 28–39, 2017.

[25] O. Niebuhr and A. Gutnyk, "Pronunciation engineering: Investigating the link between jaw-movement patterns and perceived speaker charisma using the MARRYS cap," in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2021, pp. 1–6.

[26] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[1] The nine videos used in this investigation are collected in a playlist on YouTube: https://bit.ly/3N1TnfC

Just one word: An analysis of *just* as a speaker discriminant using various acoustic measures

Ben Gibb-Reid

University of York
ben.gibb-reid@york.ac.uk

ABSTRACT

The task of forensic voice comparison (FVC) requires phonetic variables that are frequent in speech and easy to measure. The aim of this paper is to investigate the value of a frequent word in British English, *just*, as a speaker discriminant by comparing its performance across simple and complex combinations of measurements. Various likelihood ratio (LR)-based FVC systems are evaluated to assess the performance of vowel formants, segment and word durations and centre of gravity (CoG) measurements individually and in combination.

Promisingly, *just* performs better than STRUT and the vowel in *um*. Overall, including vowel durations and /s/ CoG measurements in a model offers slight improvements, but *just* F1~F3 vowel measurements alone perform well as speaker discriminants. In the case of *just*, an increased complexity of analysis does not offer a useful improvement in performance.

Keywords: forensic phonetics, acoustic phonetics, forensic voice comparison

1. INTRODUCTION

An ongoing goal of forensic speech science is to find suitable features of the voice which are highly variable between speakers but have low variation within an individual's speech [1,2]. To ensure comparison across samples of speech in forensic voice comparison (FVC), it is also necessary that these features are frequent, easy to measure, and ideally resistant to disguise and robust in transmission [2]. Looking at the most frequently spoken words could therefore provide a good source of potential FVC features. This study aims to assess the suitability of one such word: *just*. This is the 17th most frequent word in the spoken 2014 British National Corpus (BNC) [3], and it is increasing in usage in English when compared with the 2007 BNC (where it was the 42nd most used word). There are studies which also show increased usage in Tyneside [4] and Toronto English [5].

Although *just* has potential for the task of FVC, it is not always easy to measure. Frequent words are more likely to reduce phonetically [6], and *just* is no exception. For example, reliable formant estimates cannot be extracted from a token of *just* where the vowel is elided. As *just* is a word canonically made up of four segments /d͡ʒʌst/, there is the potential to combine various segmental and long-term acoustic measurements, with the aim of producing a more robust FVC analysis than one which relies on formants alone.

The measurements selected for analysis here are vowel formants, fricative centre of gravity (CoG), segment and word durations, and Mel-Frequency Cepstral Coefficients (MFCCs). Most of these are *single parameter* measurements, involving one segment only – for example vowel formants are only estimates of the vowel quality and do not pay attention to the whole token. However, word durations and MFCCs are comparatively *long-term* measurements and capture information from a broader range of the token. It is predicted that long-term measurements will perform better than single-parameter ones as they offer information from the entirety of a token. Due to their holistic nature, long-term measurements are also arguable easier to measure than single parameter ones as they do not rely on segmental variation to the same degree. Understanding how these measurements (and their combinations) affect the discriminatory power of *just* assesses the potential of word-based phonetics as new 'features' for the FVC toolkit. First, the performance of *just* vowel formant combinations is evaluated, and contrasted with the performance of vowel formants from *um* and STRUT. This is followed by an analysis of *just* durations and CoG measurements and a brief description of the MFCC results. The combinations of all features is also assessed and finally there is a discussion of the implications of these results for FVC.

2. METHODS

2.1. Data and token extraction

Data is taken from Task 1 in the DyViS corpus [7] – 100 male speakers of standard southern British English recorded in mock police interviews. Tokens were annotated manually and measured using Praat [8] with formant settings adjusted to look for five formants with a formant ceiling of 55kHz and a window length of 25ms. 1,019 tokens of *just* were extracted as suitable for vowel formant analysis, from

76 speakers. 92 vowel tokens were excluded due to fricative coarticulation, overlapping talk or short durations which made formant readings unreliable. In addition, 496 tokens of the vowel in *um* and 584 tokens of STRUT were extracted from the same 76 speakers for comparative analysis. STRUT is selected as the canonical vowel in *just*, and *um* is selected as it has performed well in FVC testing in the past [12]. 596 tokens of *just* from 55 speakers were also extracted as suitable for segment duration, CoG and MFCC measurements. 12 MFCCs were extracted from 20ms frames within 10ms shift across each token of *just* using a MATLAB script [9]. These results were then processed within the likelihood-ratio (LR) framework using various scripts in R, tidyverse packages were utilised to plot the results of these analyses [11,12].

2.2. Likelihood Ratios

To analyse potential FVC features, likelihood ratio (LR)-based testing is undertaken. LRs are a logical way to numerically express a conclusion comparing features of the voice and they are a broadly accepted way of presenting voice comparison evidence [13]. LR testing involves assessing the potential of a system to separate same-speaker (SS) and different-speaker (DS) pairs. In this study there are three stages of analysis which were achieved using the *fvclrr* package in R [14]. The first two (*feature-to-score* and *score-to-LR*) are achieved following previous research on potential FVC features [12]. The third stage (*replication*) follows the methods of [15]. In testing, the data is split into three subsets (test, training, and reference) before comparisons are run. Firstly, LR-like scores are calculated for the training data set against a set of reference data (*feature-to-score* stage) which assesses typicality using Multivariate Kernel Density (MKVD) [16]. Then, these test scores are calibrated using scores made from a training dataset (*score-to LR* stage). This calibration stage uses separate training data to make a logistic-regression model, which converts the LR scores to log likelihood ratios (LLRs) [17]. After this, the results were validated (*replication* stage). The first two stages were run with 25 replications – varying the arrangement of speakers across the training, test, and reference datasets. For example, one replication may have had speakers 1–25 in the training subset, speakers 26–50 in the test subset and speakers 51–76 in the background subset. Further replications would arrange the speakers in different permutations.

Two metrics are used to measure the validity of LRs in the study, the equal error rate (EER) – a measure of how many errors a system makes – and the log LR cost function (C_{llr}) [18], a measure of the severity of errors. The lower these are, the better a system is performing. For all results, the mean EER and C_{llr} across all replications is presented to ensure that variability in performance caused by sampling is accounted for. As speakers are compared against each other, a lower limit of six tokens is set to ensure no fewer than three same-speaker comparisons occurred. Any speakers who had less than six tokens which contained the required segments for analysis were excluded from that system. This excluded 24 speakers for vowel measurements and 45 speakers for segment duration, CoG and MFCC measurements.

3. RESULTS

The following section outlines the performance of each acoustic feature in various combinations. The EER and C_{llr} values for each system are used to assess them against each other in the form of averages across replications and the number of replications in which there is an improvement between systems.

3.1 Formants

Firstly, the performance of combinations of vowel formant is assessed and displayed in Figure 1.

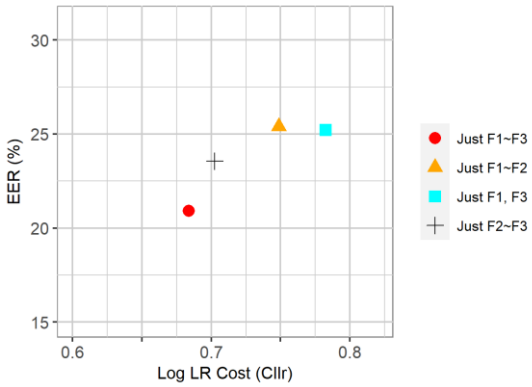


Figure 1: Mean C_{llr} plotted against mean EER comparing F1, F2 and F3 *just* vowel formant combinations across replications.

Just F1~F3 performs best overall. It has a lower mean EER (21.2%) than other systems, and a slightly lower mean C_{llr} (0.66, compared to 0.70 for F2~F3), meaning that a system relying on F1~F3 produces fewer errors of a smaller magnitude than a system based on any other combination of formants. F1~F3 has a lower C_{llr} than F2~F3 across 22 of the 25 replications and a lower EER for 21 replications, showing robustness across variable samples.

Figure 2 shows a comparison of *just* FVC performance against STRUT and *um*. Surprisingly *um* F1~F3 performs worse than *just* with a C_{llr} of 0.75 and an EER of 22.1%–more errors and of a greater magnitude. One reason for this difference in

performance could be that there were fewer tokens of *um* in the model than *just*. *Um* has a higher C_{llr} than *just* in 19 out of the 25 replications and a higher EER in 13 replications. However, these are smaller differences compared with the low performance of STRUT (0.80 C_{llr}, 30.6% EER)–a higher C_{llr} than *just* in 22 out of 25 of the replications and a higher EER in all 25 replications.

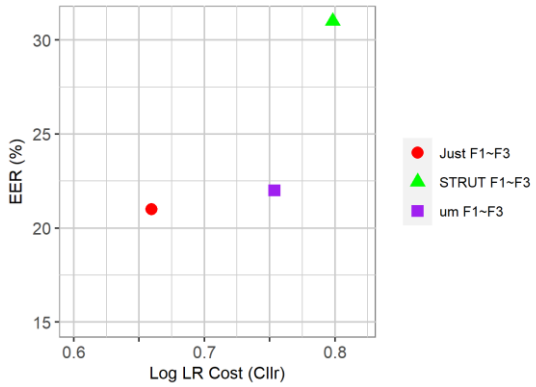


Figure 2: Mean C_{llr} plotted against mean EER comparing F1~F3 formant estimates from *just,* STRUT and the vowel in *um* across replications.

3.2 Durations

Figure 3 displays the performance of duration measures. The models are compared with a system which contains only F1~F3 vowel estimates (as there are fewer speakers, the performance of this F1~F3 system is slightly reduced against the one displayed in Figures 1 and 2).

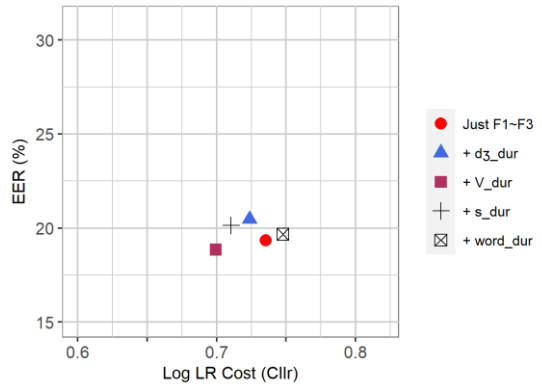


Figure 3: Mean C_{llr} plotted against mean EER comparing just segment duration measurements across replications.

There are very slight differences in performance when including segment or word durations compared with a system containing only F1~F3 (19.35% EER, 0.74 C_{llr}). Including word duration reduces model performance (a change of 0.37% EER and 0.01 C_{llr}). Including /d͡ʒ/ and /s/ duration measurements in a model leads to a very slight improvement in mean C_{llr} (by -0.02 and -0.03 respectively), with an improvement seen across 14 out of the 25 replications for /d͡ʒ/ and 17 replications for /s/. However, /d͡ʒ/ and /s/ also slightly increase EER scores (by 1.12% and 0.08% respectively). This is seen across 17 replications for /d͡ʒ/ and 13 replications for /s/. The biggest effect, though, is the improvement in performance seen when vowel duration is included. A system which contains F1~F3 and vowel duration has a C_{llr} 0.04 lower and an EER 0.5% lower than a system containing F1~F3 (this occurs in 20 out of the 25 replications for C_{llr} and 17 for EER scores). Adding vowel durations to a model therefore improves its FVC performance and it also reduces C_{llr} and EER standard deviation.

3.3 Centre of gravity

Centre of gravity (CoG) measurements were extracted from /dʒ/ and /s/. Mean performance across systems containing CoG measurements is shown in Figure 4. Three systems were tested: one containing tokens with F1~F3 vowel measurements and both CoG measurements, and two containing the vowel formant measurements and one CoG measurement.

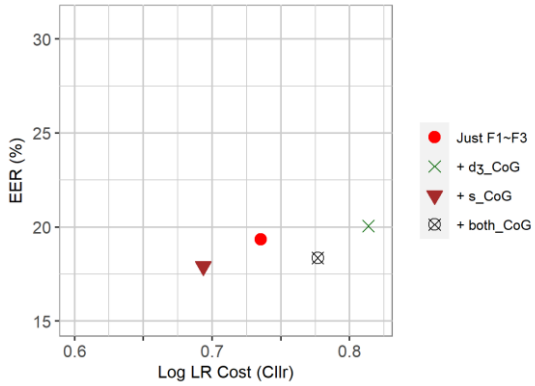


Figure 4: Mean C_{llr} plotted against mean EER comparing just centre of gravity measurements across replications.

Adding /s/ CoG measurements to a voice comparison system improved its performance. This is true for C_{llr} (a decrease of 0.04) and EER (a decrease of 1.43%) scores. Improvement is seen across 23 out of the 25 replications for C_{llr} scores and 20 replications for EER scores. Adding /dʒ/ CoG to a system reduced model performance (an increase in mean C_{llr} of 0.08, and in mean EER of 0.71%). However, an increase in C_{llr} and EER for /dʒ/ CoG is only seen across 12 out of the 25 replications. With a standard deviation of 5.6% for EER scores, /dʒ/ CoG seems to be less stable than vowel formants or /s/ CoG in terms of *just* voice comparison performance. A system containing both CoG measurements performed in-between systems containing only /s/ or /dʒ/ CoG – a slight improvement for mean EER, but an increase In mean C_{llr}. The addition of /dʒ/ CoG measurements to a system which contains /s/ CoG does not improve voice comparison performance.

3.4 MFCC measures of *just*

MFCC measurements were extracted from 726 tokens of *just* across 55 speakers. Calibrated LR voice comparison tests were run for 25 replications. The results of these suggest that MFCCs performed poorly, with a mean C_{llr} of 1.58 and a mean EER of 43.3%. C_{llr} and EER scores increase across 24 out of the 25 replications. However, this result is not a reliable reflection of the potential of MFCCs but rather a well-known statistical modelling issue, as is discussed in section 4.

3.5 Combining measures

The analysis of measurements of *just* has shown the following: for vowel formant estimates, a system containing F1~F3 performs best, and including vowel durations and /s/ CoG measurements improves model performance. A combined system made up of the best-performing single-parameter measures was also tested. This contains vowel F1~F3 and durations, and /s/ CoG measurements from each *just* token. Figure 5 displays a comparison between the best performing systems across segmental measurements, and this combined system.

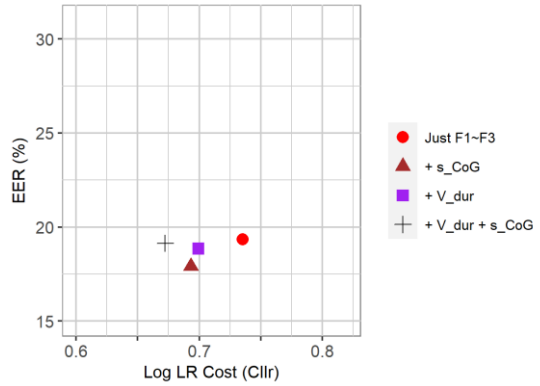


Figure 5: Mean C_{llr} plotted against mean EER comparing the best performing combinations of *just* acoustic measurements.

Overall, the combined single-parameter system performs better than F1~F3, with a mean C_{llr} of 0.67 and a mean EER of 19.1% – the lowest mean C_{llr} score across all systems tested. There are, however, very slight improvements in model performance against a system containing only F1~F3 measurements. A reduction in C_{llr} measures is seen across 19 out of the 25 replications, whereas a reduction in EER is only seen across 9 replications. The system with the lowest mean EER is the one containing F1~F3 and /s/ CoG measurements (17.9%).

4. DISCUSSION

Combining *just* single-parameter measurements provides the best system for FVC in this study. Looking at the word as a forensic feature of the voice, it is possible to combine many components to make a more robust speaker discriminant. As [2] argues, ideal speaker discriminant features of the voice should be easy to measure, but display low within-speaker and high between-speaker variation. There is a fine balance between these two requirements. On one hand, features should be as robust as possible in their discriminatory power, but often, as in the case of *just*, a higher performance is found with a complex combination of features. Vowel formant estimates are widely used in FVC analysis and casework, and that is why the baseline for much of this analysis has been F1~F3 vowel midpoint estimates. By adding /s/ CoG and vowel duration measurements, this increases model complexity but also produces the best performing system. Even when only considering vowel formants, however, *just* outperforms *um* and STRUT as a speaker discriminant in this study indicating that a word token contains more speaker-specific information than the lexical vowel it canonically relates to.

It was predicted that long-term measurements would perform better as speaker discriminants than the single-parameter (segmental) measurements. There is, in fact, a reduction in performance when including word durations and MFCCs.). It is possible that word duration is sensitive to the segmental variation of *just* in that the presence/absence of segments influenced how long or short a token would be. However, the performance of MFCCs in this study is problematic. C_{llr} scores higher than 1 are not useful for FVC, so it is surprising that MFCC measurements of *just* yielded a C_{llr} of 1.58. The problem may be to do with MVKD, which assumes that each data point is correlated even though each MFCC is not correlated. Previous research on LR-based FVC has highlighted an issue with MVKD when dealing with a high number of parameters [19]. Future research may find a better performance for word-based MFCCs in FVC systems using a less-problematic method of analysis such as Gaussian Mixture Model-Universal Background Models (GMM-UBMs).

Just shows good potential as a speaker discriminant. More work may reveal that other long-term measurements such as long-term formants and fundamental frequency contribute to its effectiveness. As it is, the potential of word tokens in FVC analysis broadens what is thought of as an individual's speech pattern beyond segmental or suprasegmental variation.

5. REFERENCES

[1] P. Rose, 'Forensic speaker identification', 2003.

[2] F. Nolan, *The phonetic bases of speaker recognition / Francis Nolan*. Cambridge: Cambridge University Press, 1983.
[3] R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery, 'The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations', *Int. J. Corpus Linguist.*, vol. 22, no. 3, pp. 319–344, 2017.
[4] K. Woolford, 'Just in Tyneside English', *World Englishes*, Jan. 3AD, doi: 10.1111/weng.12542.
[5] S. Tagliamonte, *Teen talk: The language of adolescents*. Cambridge University Press, 2016.
[6] J. Bybee, 'Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change', *Lang. Var. Change*, vol. 14, no. 3, pp. 261–290, 2002.
[7] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, 'The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *Int. J. Speech Lang. Law*, vol. 16, pp. 31–57, 2009.
[8] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. p. [Computer program], 2022. Accessed: May 18, 2022. [Online]. Available: https://www.praat.org
[9] V. Hughes, 'Script for extracting segmental MFCCs from a series of tokens saved as individual sound files'. 2018.
[10] R Core Team, 'R: A Language and Environment for Statistical Computing.' R Foundation for Statistical Computing, Vienna. URL: https://www.r-project.org, 2020.
[11] H. Wickham *et al.*, 'Welcome to the Tidyverse', *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.
[12] V. Hughes, S. Wood, and P. Foulkes, 'Strength of forensic voice comparison evidence from the acoustics of filled pauses', *Int. J. Speech Lang. Law*, vol. 23, no. 1, pp. 99–132, 2016.
[13] B. Robertson and G. A. Vignaux, 'DNA evidence: Wrong answers or wrong questions?', *Genetica*, vol. 96, no. 1, pp. 145–152, Jun. 1995, doi: 10.1007/BF01441160.
[14] J. Lo, 'fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison'. 2022. Accessed: Oct. 04, 2022. [Online]. Available: https://github.com/justinjhlo/fvclrr#readme
[15] B. X. Wang, V. Hughes, and P. Foulkes, 'The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison', *Speech Commun.*, vol. 138, pp. 38–49, Mar. 2022, doi: 10.1016/j.specom.2022.01.009.
[16] C. G. Aitken and D. Lucy, 'Evaluation of trace evidence in the form of multivariate data', *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 53, no. 1, pp. 109–122, 2004.
[17] G. S. Morrison, 'Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio', *Aust. J. Forensic Sci.*, vol. 45, no. 2, pp. 173–197, 2013, doi: 10.1080/00450618.2012.733025.
[18] N. Brümmer and J. Du Preez, 'Application-independent evaluation of speaker detection', *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 230–275, 2006.
[19] B. Nair, E. Alzqhoul, and B. J. Guillemin, 'Determination of Likelihood Ratios for Forensic Voice Comparison Using Principal Component Analysis', *Int. J. Speech Lang. Law*, vol. 21, no. 1, pp. 83–112, Jun. 2014, doi: 10.1558/ijsll.v21i1.83.

Sensitivity of x-vectors and automatic speaker recognition scores to vocal variation

Jessica Wormald[1], Paul Foulkes[1], Philip Harrison[1], Vincent Hughes[1], Finnian Kelly[2], David van der Vloed[3], Poppy Welch[1], Chenzi Xu[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]Oxford Wave Research, Oxford, UK
[3]Netherlands Forensic Institute, The Hague, The Netherlands
{jessica.wormald|paul.foulkes|philip.harrison|vincent.hughes|poppy.welch|chenzi.xu}@york.ac.uk,
finnian@oxfordwaveresearch.com, d.van.der.vloed@nfi.nl

ABSTRACT

Automatic speaker recognition (ASR) systems rely on a complex processing chain in order to compare speech signals and produce likelihood ratios. The complexity of this chain, and of the speech signals themselves, mean that there is still limited understanding about what makes a certain voice easy or difficult for a system to recognise. This gap in understanding is holding back the use of ASR in forensic casework.

This study considers two specific parts of the ASR chain: x-vectors (speaker models) and within-speaker comparison scores. Using heavily-controlled data from two phoneticians, we demonstrate that variability in vocal setting results in phonetically-predictable shifts in x-vectors and scores. Shifts in supralaryngeal voice quality produce the biggest deviations from modal voice. The results provide a basis for exploring how properties of the voice affect ASR performance, which in turn can contribute to helping courts and practitioners take advantage of ASR systems in forensic casework.

Keywords: speaker recognition; forensic phonetics; vocal setting

1. INTRODUCTION

1.1 The context

Automatic speaker recognition (ASR) systems are increasingly used around the world in forensic speaker comparison cases [1,5]. State-of-the-art systems utilise deep neural networks (DNNs) to convert acoustic features (typically mel frequency cepstral coefficients, MFCCs) to a compact, fixed-length speaker representation, known as an x-vector [8]. Such systems perform very well, even in forensically realistic conditions, when optimised using case-specific data [4,6].

While much research in ASR has focused on tackling technical challenges arising from different recording types (e.g. channel effects such as telephone transmission, background noise), still very little is known about why certain voices are easy or difficult for systems to recognise. This, in part, explains why some courts are still cautious about ASR as a form of expert evidence [7]. A particular concern is being able to explain what information about the voice is being captured by complex models like DNNs. This study is the first step in a wider project which will provide a more comprehensive understanding of the linguistic and phonetic bases of ASR system behaviour, which in turn will help courts and practitioners make best use of ASR systems in forensic casework.

1.2 Some challenges

ASR systems involve a complex set of processes to analyse and compare two speech signals (e.g., one of a known suspect and one of an unknown offender) in order to compute a numerical value which represents the strength of evidence. Features are initially extracted from the 'voice-active' portion of each speech signal (i.e. all parts of the signal containing speech). In state-of-the-art systems the features are then passed through a DNN to generate x-vectors. The x-vectors from two signals are compared to produce a score. Scores are numerical representations of the similarity and typicality of the x-vectors derived using pre-trained models within the system. However, scores are not directly interpretable, i.e. it is not always possible to judge whether a score is 'high' or 'low' in isolation, or given a single score, whether the system has made an error. In part this is due to differences between the conditions of the evidential comparison and the data used to train the system. Therefore, the final stage in the ASR processing chain is *calibration*, which is a means of interpreting the evidential score in light of same- and different-speaker scores from comparisons that are representative of the evidential comparison, but where the ground truth is known. This process of calibration converts the score to a likelihood ratio (LR); it is only at this point that the strength of evidence can be evaluated directly.

Speech recordings from forensic casework are also complex. Within a case, there is likely to be considerable variability across samples in terms of

speaking style, interlocutor, and speaking level, as well as technical factors caused by e.g. channel differences, sample duration and background noise. Each forensic case is also unique, meaning that analysis could in principle involve speakers of any language, regional accent and socio-economic background, in any situational context.

1.3 A solution

In this study, we take an initial step towards understanding the sensitivity of ASR systems to different types of voices and the extent to which systems capture different phonetic information. We do this using highly controlled, and therefore forensically unrealistic, recordings of phoneticians in different vocal conditions. We focus on specific parts of the ASR chain in order to localise where speaker effects emerge and how they contribute towards system output.

Specifically, we focus on (i) x-vector speaker representations (extracted from a DNN using MFCCs) for different vocal conditions within-speakers, and (ii) the subsequent within-speaker score distributions which are generated by running a comparison in the ASR, but before any system calibration. All other variables are held constant (see description of data below) so that variation in the distributions can be interpreted relative to changes within a speaker's voice. While the focus on x-vectors and scores means that we cannot directly interpret whether differences are 'important' in terms of their ultimate impact on system performance, it does enable us to identify conditions of interest for further work.

2. DATA

This study uses a heavily controlled corpus which includes variation in speaker, vocal condition, time, and technical condition. In this paper, we report on a subset of the available material.

2.1 Participants

We report on data from two phoneticians: Paul Foulkes (PF - P1) and Francis Nolan (FN - P4). By using data from experienced phoneticians we could ensure there was minimal variability between participants when varying vocal conditions.

2.2 Vocal conditions

Each participant read the first two paragraphs of *The Rainbow Passage* in twenty four vocal conditions. These were selected to reflect changes in segmental and suprasegmental vocal parameters and fell into four groups (individual conditions are listed in Figures 1 and 2 below):

- *Modal voice.* This was used as a baseline for a speaker's 'normal' way of speaking.
- *Accent guises (5).* A range of linguistic changes, no significant shift in any one feature.
- *Miscellaneous (7).* Other forensically relevant variation.
- *Voice quality (7 supralaryngeal, 3 laryngeal).* One vocal setting at extreme, all others held as constant as possible.

Each vocal condition was repeated three times within a session (non-consecutive repetitions). Both participants took part in three sessions which were at least a week apart. In each session the conditions were repeated in the same order.

2.3 Technical conditions

Sessions were recorded in an anechoic chamber. Participants were seated at one end of the chamber throughout. Repetitions were simultaneously recorded in four technical conditions: headband microphone (DPA 4066 omnidirectional headset), near microphone (1m from participant), far microphone (2m from participant), and landline-to-VOIP call. Recordings were made in PCM WAV format with a 48kHz sample rate at 24 bits. For the purposes of the present study, only the headband microphone recordings were analysed

Individual repetitions of each vocal condition were extracted from within each session and are referred to throughout as a sample (i.e. 1 sample = 1 repetition of 1 condition in 1 session). There were 218 samples for PF, and 216 for FN. (There was one additional repetition of the high pitch and whisper conditions for PF.)

3. METHODS

3.1 The system

Analysis was carried out in the ASR software VOCALISE 2021 (version 3.0.0.1746) [2]. We report on outputs from the system at different stages in the comparison process.

3.2 x-vectors

For each participant, we first generated x-vectors for each sample using the default VOCALISE x-vector model.

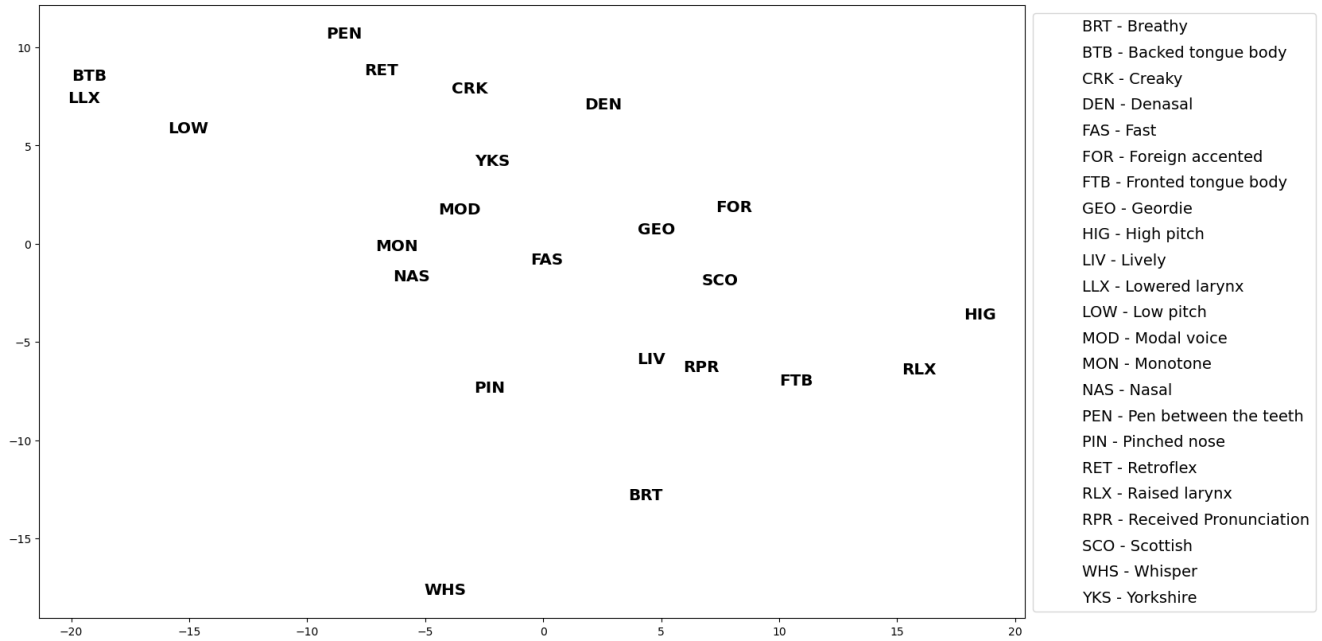


Figure 1: A t-SNE visualisation showing the average x-vector position for FN in each condition in 2-D space.

These were generated from MFCCs, which were extracted on a frame-by-frame basis across the sample and then passed through a DNN [2,8]. The x-vector is a representation of the speaker's voice in a given sample and contains 512 values. t-SNE (t-Distributed Stochastic Neighbour Embedding) plots were generated using the x-vectors to assess within-speaker patterns. t-SNE plots provide a way of visualising high-dimensional data in two or three dimensional space [9]; similar points appear closer together, and dissimilar points further apart [3].

3.3 Score distributions

Within-speaker comparisons were also carried out in VOCALISE using the default x-vector PLDA model. For each participant, each sample was compared to every other sample they produced. Each comparison results in a score. As noted earlier, scores are numerical representations of the similarity and typicality of the x-vectors derived using pre-trained models within the system, and interpretable only relative to other sets of similar scores. In this study we discuss only the within-speaker modal-to-other scores.

4. RESULTS

4.1 Within-speaker x-vector space

The t-SNE plot in Figure 1 visualises the x-vector space for FN (more detailed plots for both speakers are available online).

For FN, all within-condition repetitions clustered around the same area, so for ease of visualisation we have included only the average of the coordinates. The distribution of conditions in the x-vector space generally reflects the phonetic distance between the conditions. Low pitch, lowered larynx and backed tongue body all cluster in a similar area in the top left of the plot. At the bottom, breathy and whisper are separated from other conditions. Although his plot is not included here, PF's foreign accented repetitions did not cluster together, but instead three samples appeared away from the main cluster. These three were characterised by more marked larynx lowering, harsh phonation and some lingual retraction; this was in contrast to the other repetitions which were more breathy and perceptually higher in pitch. Taken together, it is clear that the x-vector distributions are capturing phonetic variability.

4.2 Within-speaker score distributions

The plot in Figure 2 shows the within-speaker (i.e. same speaker) modal-to-other score distributions. The dark grey distributions are PF, and the light grey are FN. The dashed line represents the median modal-modal score across PF and FN and is used here as a baseline for assessing which conditions demonstrate the biggest effects on scores. Individual medians for each distribution are shown as a solid line within the distribution. Vocal conditions are grouped, and ranked within each group based on the distance from modal (i.e. those which are more different appear at the bottom of the display).

The within-speaker modal-modal distributions for PF and FN overlap almost completely; both have similar degrees of variability across their repetitions in this condition and the scores are very similar.

The modal-accent guise scores are generally the closest to the modal-modal distributions. Both PF and FN varied their speech within these conditions on a range of different phonetic dimensions but the scores are generally in the same range as the modal-modal distributions. The exception to this is PF's foreign accented voice; as discussed earlier, these samples showed within-condition variability, and were also more different from his modal voice in a range of features.

Overall, the modal-supralaryngeal voice quality scores are the most different from the modal-modal scores. However, there is also considerable variability within this group, and between FN and PF. Modal-lowered larynx and modal-backed tongue body comparisons have distributions furthest from modal-modal, and FN and PF have different distributions, both in shape and degree of difference from modal.

For the miscellaneous conditions, those which involve supralaryngeal changes have the biggest impact on the score distributions. Modal-pinched nose, and modal-high pitch are most different from modal-modal. The difference between FN and PF in the modal-high pitch condition likely reflects the different strategies employed to produce high pitch; PF's high pitch involved some laryngeal raising and falsetto, thus involving supralaryngeal as well as laryngeal changes compared with PF's modal. The score distributions appear to reflect this difference.

When considering laryngeal voice quality, modal-whisper scores are the most different from those for modal-modal.

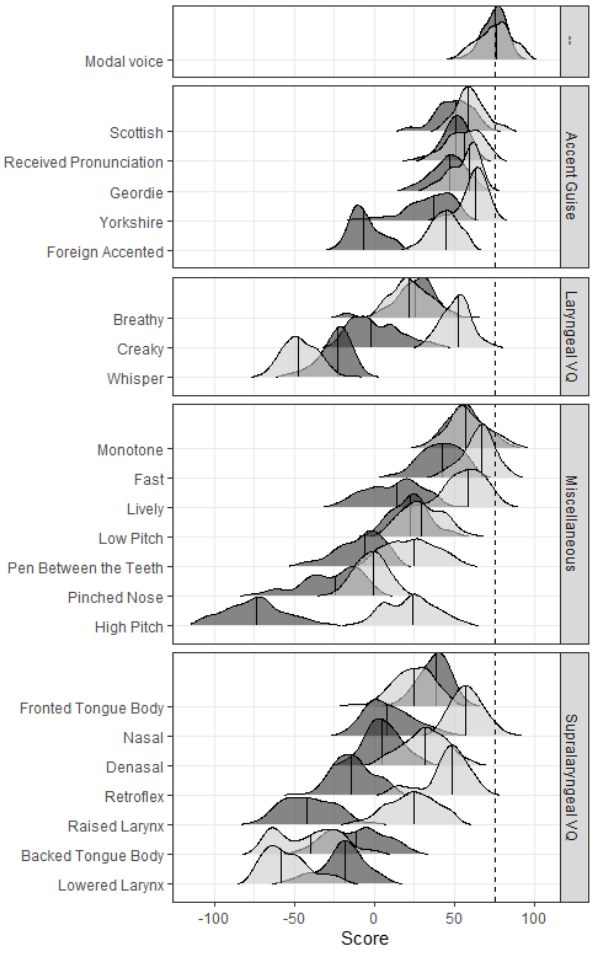


Figure 2: Within-speaker modal-to-other score distributions. Dark grey = PF; light grey = FN.

5. DISCUSSION

5.1 Within-speaker variability

The relationship between vocal conditions in the x-vector and score distributions reflects the degree of phonetic variability from modal. Supralaryngeal changes and whisper have the most divergent score distributions compared with the modal condition; the outcomes of these impact on a wide range of phonetic patterns (e.g. backed tongue body results in more retracted realisations of all vowels). In contrast, making intermittent, more targeted shifts to a range of features (e.g. the modal-accent guises) does not markedly impact the score distributions when compared to the modal-modal distributions.

5.2 Forensic applications

The findings allow us to further our understanding of potential mismatch conditions which might be relevant in casework situations.

Although we do not know at this stage whether such factors would affect system performance. However, the fact that the system is sensitive to phonetic variation in a linguistically-predictable way (i.e. large supralaryngeal changes result in the most difference) at the two stages considered here is useful when it comes to addressing issues of explainability, for example, in a courtroom.

5.3 Future plans

In future work, we will consider further stages in the chain such as calibration and validation, and include different-speaker comparisons. We will also explore how the findings from this small controlled corpus can be extrapolated to other, less controlled data and a larger number of speakers.

6. ACKNOWLEDGEMENTS

This work was funded by ESRC grant ES/W001241/1.

7. REFERENCES

[1] Gold, E., French, P. 2019. International practices in

forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law* 26, 1–20.

[2] Kelly F, Forth O, Kent S, Gerlach L, Alexander A. 2019. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *AES International Conference on Audio Forensics*. Available: http://www.aes.org/e-lib/download.cfm?ID=20477

[3] Kelly, F., and Hansen, J. 2021. Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition. *IEEE/ACM Trans Audio Speech Lang Process* 29, 927–942.

[4] Morrison, G., Enzinger, E. 2016. Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. *Speech Communication* 85, 119–126.

[5] Morrison, G., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., et al. 2016. INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263, 92–100.

[6] Morrison, G., Enzinger, E. 2019. Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01)–Conclusion. *Speech Communication*. 112, 37-39

[7] R v Slade & Ors. 2015. Lord Justice Davis, Mr Justice Wilkie, Mr Justice Holroyde. Available: https://www.casemine.com/judgement/uk/5a8ff7a560d03e7f57eb0c31

[8] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[9] van der Maaten, L., Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605.

Automatic assessment of voice similarity within and across speaker groups with different accents

Linda Gerlach[1,2], Kirsty McDougall[1], Finnian Kelly[2], Anil Alexander[2]

[1]University of Cambridge, United Kingdom, [2]Oxford Wave Research, United Kingdom
{lg589|kem37}@cam.ac.uk, {finnian|anil}@oxfordwaveresearch.com

ABSTRACT

As larger speaker databases materialise and speech technology becomes more efficient, an opportunity arises to automate assessments of perceived voice similarity ('PVS') among selections of voices for applications such as voice synthesis and forensic voice parades. Expanding on previous research [1], the present study addresses whether the correlation observed between listener ratings and automatic estimates of PVS can be found within different speaker groups with the same accent and across speaker groups with a different accent of English. Further, the impact on this correlation of variations in the automatic approach for PVS, based on a pre-trained automatic speaker recognition system, is explored; specifically, combinations of different feature extraction methods, speaker modelling approaches, and distance measures are considered. Results are positive and statistically highly significant within and across speaker groups particularly when relying on automatically extracted perceptually relevant phonetic features, demonstrating the generalisability of the method.

Keywords: voice similarity, voice parades, forensic speech science, automatic speaker recognition

1. INTRODUCTION

The assessment of perceived voice similarity ('PVS') between speech samples is a costly and time-consuming task when conducted manually. Subjective ratings from a number of listeners are required to approximate a 'ground-truth' value of PVS for a speaker pair, e.g. [2].

For forensic voice parades, where a witness is asked whether they can identify a perpetrator's voice among a number of foil voices that sound similar with regard to their pitch, speed, accent and more, assessing PVS is crucial for ensuring a fair selection of foils [3]. Under the scrutiny of an expert phonetician, partial automation of the selection process could allow for a larger pool of potential foils to be considered while facilitating a more objective selection at greater speed.

Voice banking applications aim to provide a person who has lost their voice, e.g., due to a medical condition, with a personalised speech synthesis device [4]. If audio recordings from the patient are not sufficient or unavailable, a 'donor voice' may be used for synthesis. Automatic assessment of PVS between the patient's voice (recordings) and a pool of donor voices may help swiftly narrow down suitable candidates for the patient to choose from and find their ideal donor voice.

Recent developments in voice synthesis show a need for fast assessments of similarity for large numbers of target speakers and their synthesised samples. Das et al. [5] explored the use of an automatic speaker recognition (ASR) system based on x-vectors for similarity assessments with encouraging results. In a similar vein, Deja et al. [6], trained a model based on spectral features to automatically evaluate PVS and yielded Pearson correlations of up to 0.78 between their model's output and perceptual ratings.

Past research also considered the use of phonetic features in automatic assessments of PVS [1, 7], but used only a small number of speakers. Building on Gerlach et al. [1], the present paper proposes assessing PVS using an existing pre-trained ASR system. This paper will address 1) whether there is a correlation between listener ratings and automatic estimates of PVS within different groups of speakers with the same accent and 2) whether this correlation can also be found across groups of speakers for different accents of English. It will further be explored 3) how the correlations observed are impacted by different feature extraction and speaker modelling approaches, as well as distance measures.

2. METHOD

2.1. Speaker databases and groups

Six speaker groups were chosen from DyViS [6, 7], YorViS [10], and WYRED [11]. DyViS contains 100 male speakers of Standard Southern British English (SSBE) aged 18-25. Three sets of 15 DyViS speakers (henceforth D1, D2, and D3) were employed to enable observation of variability between speaker groups of the same accent. Further, 15 male speakers of York English from YorViS (referred to as Y), also 18-25, were used. Finally, two groups of 15 speakers (hereafter W1 (Bradford) and W2 (Wakefield)) out of

180 male speakers (18-30 y.o.) of West Yorkshire Englishes were selected from WYRED. This selection enables the analysis of variation between three Yorkshire accent groups as well as comparison with the three same-accent SSBE groups. The recordings from all three databases contained spontaneous speech in studio quality from a mock police interview and a telephone call with an 'accomplice'.

2.2. Listener experiment

Listener ratings of PVS for all six speaker groups, collected as part of the VoiceSim [12], YorViS [10], and IVIP [13] projects, and presented in [14], were made available for the present study. For the listener experiments, the researchers had created two short voice samples of about 3s from each speaker's telephone call recording (studio quality). The samples were paired to form 120 speaker pairs (including same-speaker) for each speaker group. Listeners were then asked to rate the voice similarity of each speaker pair on a 9-point scale (1 – very similar, 9 – very different). Each speaker group was rated by a different set of 20 listeners (aged 18-40, English L1, born and raised mostly in England until aged 18), reporting no hearing impairments, and approximately balanced for sex.

For the present study, the given listener ratings were inverted (1 – low similarity, 9 – high similarity) to ensure that the scales used for the automatic estimation of voice similarity and the listener ratings were in the same direction.

2.3. Automatic comparisons

VOCALISE forensic ASR software [13, 14] was used in this experiment to obtain similarity scores automatically. For speaker modelling, the software provides pre-trained i-vector and x-vector frameworks that can use spectral Mel-frequency cepstral coefficients (MFCCs) or so-called auto-phonetic (AP, automatically extracted phonetic) features, which are regarded as perceptually relevant [7]. Independent of the feature extraction and speaker modelling approaches, in the automatic speaker comparison two speaker models are compared and a comparison score is calculated. In this study, the comparison scores are interpreted as indicators of voice similarity (low score – low similarity, high score – high similarity).

VOCALISE provides combinations of pre-trained modelling approaches and parameters as 'sessions' [15]. Table 1 shows the VOCALISE sessions tested in this study. I-vector session (a) with its combination of auto-phonetic features, including F0 among others, was included as this had previously been applied in similar experiments by [7] and [1] and is considered to be the most relevant session with regard to PVS (see also [16] regarding the importance of F0 and formant frequencies in PVS, and [12]). I-vector session (b) was included to evaluate the relevance of formant frequencies and F0 for the assessment of voice similarity. The i-vector session (c) based on MFCCs was included as similar approaches had been used previously, for instance in [18], and to test the hypothesis that approaches incorporating perceptually relevant phonetic features are superior to those that do not. Due to the development of new speaker modelling approaches, this experiment also includes x-vector sessions based on Deep Neural Networks. X-vector session (d) using spectral features outperforms i-vectors in speaker recognition tasks [16] and it is hypothesised that this speaker modelling approach may also have an impact on the representation of PVS. An x-vector session relying on auto-phonetic features (e) is explored as the inclusion of perceptually relevant features yielded promising results in previous experiments assessing voice similarity [1, 15]. Finally, an x-vector session using auto-phonetic features including F0 (f) is explored, and results compared to its i-vector equivalent (a).

	Feature extraction	Speaker model
a	Auto-phonetic (AP), ie. F0, semi-tones of F0, derivatives, LTF1 to LTF4	i-vector
b	AP, ie. LTF1 to LTF4	i-vector
c	spectral/MFCC	i-vector
d	spectral/MFCC	x-vector
e	AP, ie. LTF1 to LTF4	x-vector
f	AP, ie. F0, semitones of F0, derivatives, LTF1 to LTF4	x-vector

Table 1: Feature extraction and speaker modelling combinations used in VOCALISE.

Comparison scores were initially calculated based on the widely used PLDA (probabilistic linear discriminant analysis) which has great discriminatory power to tell speakers apart [16]. For comparison, cosine distance was also evaluated for calculating similarity scores to explore whether a less powerful discriminator may be more useful for assessing voice similarity.

Comparisons were conducted between all speakers within each of the six speaker groups. For each speaker, two samples (~4min) were used, one each from the telephone call and the interview task. The resulting comparison scores were then interpreted as voice similarity estimates. The scores were calibrated using Bio-Metrics [19] in order to normalise their numerical range and ensure comparability across speaker groups. The two

comparison scores for each speaker pair (spk1_1 vs spk2_2, spk2_1 vs spk1_2) were averaged to obtain one score per speaker pair [1].

2.4. Evaluation and analysis

To adjust for bias, the 5% trimmed mean of the listener ratings was calculated. Spearman rank correlation was performed to explore the degree and direction of the relationship between listener ratings of PVS and automatically obtained similarity estimates. To assess the potential for establishing thresholds for PVS in the future, the linearity of the relationship was measured using Pearson's correlation coefficient.

3. RESULTS

In sections 3.1. and 3.2. below, automatic comparison results are based on the i-vector session using auto-phonetic features including F0 as described in section 2.3. (a), and PLDA. Section 3.3. gives an overview of the performance of all VOCALISE sessions using PLDA to calculate scores, while section 3.4. considers the impact of PLDA versus cosine distance.

3.1. Same-accent background

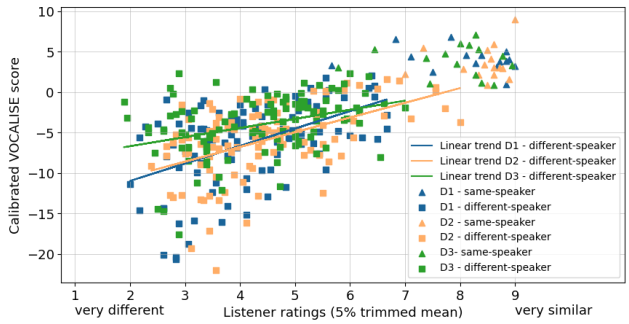


Figure 1: Scatter plot of listener ratings versus automatically obtained scores based on session (a) as described in 2.3. for DyViS D1 (blue), D2 (yellow), and D3 (green). DS comparisons are displayed as squares, SS comparisons as triangles.

Figure 1 illustrates the results for the three DyViS speaker groups, with listener ratings on the x-axis, and automatically obtained similarity estimates on the y-axis. Linear trendlines are shown for the different-speaker (DS) comparisons only. The graph shows that listeners seem to treat same-speaker (SS) comparisons in a similar way to DS comparisons, as listeners used a range of scores and did not rate SS comparisons consistently as 'very similar' across the three speaker groups. The automatic system showed overall good discrimination between SS and DS comparisons. The degree of the linear relationship between automatic scores and listener ratings varies across the three SSBE speaker groups, nevertheless, the arising trends all show a positive, largely linear relationship. In the statistical analyses which follow only DS comparisons are included.

3.2. Different-accent background

Speaker group	Spearman's rho	Pearson's r
D1	0.484**	0.503**
D2	0.519**	0.502**
D3	0.357**	0.366**
Y	0.463**	0.704**
W1	0.380**	0.398**
W2	0.552**	0.631**

Table 2: Correlation coefficients for each speaker group using the AP i-vector session with F0 and PLDA to approximate listener PVS ratings. Level of statistical significance highlighted as **<0.001, 2-tailed.

Table 2 shows that across all six speaker groups, the correlations between listener ratings and automatically obtained scores are statistically highly significant. For Spearman analyses, the lowest correlation is observed for D3, whereas W2 yielded the highest correlation, followed by D2. These results show that there is some variation between different speaker groups regardless of accent. This is also true for variation in linearity as indicated by the Pearson results; note that for Y, a very distinctive speaker had a substantial influence on the correlation.

3.3. Session performance across speaker groups

	Session	Spearman's rho	Pearson's r
a	i-vector AP inc. F0	0.457**	0.472**
b	i-vector AP	0.276**	0.269**
c	i-vector MFCC	0.349**	0.339**
d	x-vector MFCC	0.196**	0.201**
e	x-vector AP	0.417**	0.399**
f	x-vector AP inc. F0	0.244**	0.272**

Table 3: Correlation coefficients for each VOCALISE session based on automatic similarity estimates using PLDA and listener ratings of five speaker groups (D1, D2, D3, W1, W2). Level of statistical significance highlighted as **<0.001, 2-tailed.

To evaluate the performance of individual VOCALISE sessions, DS scores from all speaker groups except YorViS were combined and correlated with the aggregate of the corresponding 5% trimmed mean listener ratings. The Y speaker group was excluded as it was not possible to obtain calibrated scores for this set.

Table 3 shows that correlations for all sessions tested were statistically highly significant and broadly linear. X-vector session (d) relying on spectral features yielded the lowest Spearman correlation coefficient, followed by the i-vector session using auto-phonetic features without F0 (b). The next highest correlation coefficient was found using the i-vector session with spectral features (c). The sessions performing best in approximating ratings of PVS were the x-vector session using auto-phonetic features (e) and, ultimately, the i-vector session relying on auto-phonetic features and F0 (a). Of further note is that for the two best performing sessions, Spearman correlations for all individual speaker groups were positive and statistically highly significant, while this was not the case for the other sessions.

Based on these results, it was expected that a session using x-vectors and auto-phonetic features including also F0, semitones of F0 and derivatives (f) would have the potential to outperform the i-vector session using these features. Hence, a session combining these features for an x-vector model was also evaluated. In fact, results showed a statistically highly significant but low correlation between listener ratings of PVS and the new calculated scores compared to the majority of the previously tested sessions.

3.4. Impact of distance measure on correlation

Since PLDA as well as cosine distance are commonly encountered distance measures in speaker recognition, the correlation experiment using the five speaker groups (excl. Y) was repeated using cosine distance instead of PLDA.

Overall, again all correlation coefficients were positive and statistically highly significant. On average, Spearman correlation coefficients were higher using PLDA. On an individual session level, using cosine distance had an adverse effect on the ranking of the i-vector session using auto-phonetic features including F0, leaving it in third place (rho=0.309, $p<0.001$, 2-tailed), while the i-vector session using spectral features came up second best (rho=0.318, $p<0.001$, 2-tailed). Results showed a slight increase of 0.023 for the correlation coefficient for the x-vector session using auto-phonetic features (rho=0.440, $p<0.001$, 2-tailed), suggesting it as the best choice when using cosine distance.

4. DISCUSSION

This study expanded on [1] to show the generalisability of the proposed voice similarity assessment across more diverse speaker groups. Overall, the relationship between listener ratings and automatic estimates of PVS was found to be positive and statistically significant within the speaker groups of the same accents as well as across speaker groups of a different accent of English. Variability was present in the strength and linearity of the correlations, but there were no apparent trends related to the speakers' accents, likely due to the variation of individual speakers or listeners exceeding any variation related to the different accents and impact of sample choice.

The best performance in automatically approximating listener ratings of PVS using a pre-trained ASR system was yielded by applying the i-vector session incorporating LTF1 to LTF4, F0 and semitones, as well as derivatives in combination with PLDA as a distance measure, closely followed by the x-vector session drawing solely on LTF1 to LTF4 in combination with PLDA. The latter session's correlation with the listener ratings improved further when using cosine distance. Further conclusions can be drawn if interrater agreements can be better assessed.

While the similarity rating scales used by Deja et al. [6] (0-100) and in the current experiment (1-9) may not be directly comparable, the correlations in this study are not as high as those reported by Deja et al.. Additional experiments considering synthetic speech are needed to draw further conclusions.

Regarding the different applications mentioned in section 1., using an existing ASR system with features that are relevant for PVS may offer an opportunity to select similar-sounding natural voices for voice parades and voice banking. It should be noted that the search space for similar-sounding voices may need to be restricted in terms of demographic data to be appropriate for the task. Further, a suitable degree of similarity between voices acceptable for a particular application must be quantified, e.g., by using calibrated scores from the automatic system.

5. CONCLUSION

This paper showed that using a pre-trained ASR system to assess PVS provides positive, statistically highly significant results within and across speaker groups of different accents of English, particularly using automatically extracted, perceptually relevant phonetic features. Future work will investigate score thresholds for PVS for different applications and explore large and diverse speaker databases using similarity scores and clustering methods.

6. ACKNOWLEDGEMENTS

The authors thank Oscar Forth for his assistance with VOCALISE modifications for the paper. This work was partially funded by the Selwyn Cambridge – Oxford Wave Research PhD Studentship in Forensic Phonetics and Automatic Speaker Recognition and the IVIP project 'Improving Voice Identification Procedures' which is funded by the UK Economic and Social Research Council, reference ES/S015965/1.

7. REFERENCES

[1] Gerlach, L., McDougall, K., Kelly, F., Alexander, A., Nolan, F., 2020. Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Commun.*, 124, 85–95. doi: 10.1016/j.specom.2020.08.003.

[2] McDougall, K., 2013. Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *Int. J. Speech Lang. Law*, 20(2), 163–172. doi: 10.1558/ijsll.v20i2.163.

[3] Home Office, 2003. Advice on the use of voice identification parades. [Online]. Available: https://tinyurl.com/26r4ucn6 (28.04.2023).

[4] Yamagishi, J., Veaux, C., King, S., Renals, S., 2012. Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction. *Acoust. Sci. Technol.*, 33(1), 1–5. doi: 10.1250/ast.33.1.

[5] Das, R. K., Kinnunen, T., Huang, W., Ling, Z., Yamagishi, J., Zhao, Y., Tian, X., Toda, T., 2020. Predictions of subjective ratings and spoofing assessments of Voice Conversion Challenge 2020 submissions. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 99–120. doi: 10.21437/VCC_BC.2020-15.

[6] Deja, K., Sanchez, A., Roth, J., Cotescu, M., Jul. 2022. Automatic evaluation of speaker similarity. In *Proc. of INTERSPEECH 2022* Incheon, Paper 75. doi: 10.48550/arxiv.2207.00344.

[7] Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J., Åkesson, J., 2016. Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. In *Proc. of INTERSPEECH 2016* San Francisco, Paper 2018, 1567–1568. Available: https://tinyurl.com/mr3nyhhr (28.04.2023).

[8] Nolan, F., Dynamic Variability in Speech: a forensic phonetic study of British English, 2006-2007. 2011, UK Data Service. SN: 6790. doi: http://doi.org/10.5255/UKDA-SN-6790-1.

[9] Nolan, F., McDougall, K., de Jong, G., Hudson, T., 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int. J. Speech, Lang. Law*, 16(1), 31–57. doi: 10.1558/ijsll.v16i1.31.

[10] McDougall, K., Duckworth, M., Hudson, T., 2015. Individual and group variation in disfluency features: a cross-accent investigation. *Proc. 18th ICPhS*, Paper 0308. Available: https://tinyurl.com/jw2v3uu9 (28.04.2023)

[11] Gold, E., Ross, S., Earnshaw, K., 2018. The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. In *Proc. of INTERSPEECH 2018* Hyderabad, Paper 0065, 2748–2752. doi: 10.21437/Interspeech.2018-65.

[12] Nolan, F., McDougall, K., Hudson, T., 2011. Some acoustic correlates of perceived (dis)similarity between same-accent voices. In *Proc. 17th ICPhS* Hong Kong, 1506–1509. Available: https://tinyurl.com/yf9skafm (28.04.2023).

[13] Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A., Smith, H. M. J., 2023. Identifying unfamiliar voices: examining the system variables of sample duration and parade size. *Q. J. Exp. Psychol.*, 1–19. doi: 10.1177/17470218231155738.

[14] McDougall, K., 2021. Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. In *Proc. of XVII AISV (Associazione Italiana Scienze della Voce) conference: "Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications"*, 33–56. doi: 10.17469/O2108AISV000002.

[15] Alexander, A., Forth, O., Atreya, A. A., Kelly, F., 2016. VOCALISE : a forensic automatic speaker recognition system supporting spectral , phonetic , and user-provided features. In *Odyssey 2016 Show & Tell* Bilbao, Paper 88. Available: https://odyssey2016.org/papers/Show_tell/88.pdf (28.04.2023)

[16] Kelly, F., Forth, O., Kent, S., Gerlach, L., Alexander, A., 2019. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In *Proc. of the AES International Conference* Porto, Paper 27. Available: http://www.aes.org/e-lib/browse.cfm?elib=20477 (28.04.2023).

[17] McDougall, K., 2013. Earwitness evidence and the question of voice similarity. *Br. Acad. Rev.*, 21, 18–21. Available: https://tinyurl.com/47umjb3u (28.04.2023).

[18] Lindh, J., Eriksson, A., 2010. Voice similarity - a comparison between judgements by human listeners and automatic voice comparison. In *Proc. of FONETIK 2010* Lund, 63–68. Available: https://journals.lub.lu.se/LWPL/article/view/19647/17775 (28.04.2023)

[19] Bio-Metrics 1.8 Performance Metrics Software. 2019, Oxford Wave Research. Available: https://oxfordwaveresearch.com/products/bio-metrics/ (28.04.2023).

VOICE DISTINCTIVENESS: AN INVESTIGATION OF THE ROLE OF SPEAKERS' POSITION IN A POPULATION WITH RESPECT TO f0

Kirsty McDougall, Alice Paver and Francis Nolan

University of Cambridge
kem37|aep58|fjn1@cam.ac.uk

ABSTRACT

The role played by f0 in listeners' assessment of voice distinctiveness is investigated. In Experiment 1, listeners judged the (dis)similarity of low-, medium-, and high-pitched voices selected from an accent- and demographic-matched population. Listeners' judgements tended to cluster speakers together according to their position in a population distribution for f0, yet the tightness of these clusters varied, with members of the high-pitched group being consistently judged as most different from each other. This suggests that other phonetic dimensions are likely to be relevant as well as pitch. Experiment 2 collected similarity judgements of the same stimuli resynthesised with pitches shifted between the same low, medium and high positions in the population distribution. Results show that shifting the pitch of stimuli did not lead to a significant change in judgements, again indicating that more than pitch alone is driving the judgements phonetically. Implications for earwitness identification are discussed.

Keywords: voice distinctiveness, voice similarity, fundamental frequency, earwitness, voice parades

1. INTRODUCTION

In criminal cases where a witness has heard but not seen a perpetrator, a voice parade may be conducted to see whether the witness is able to pick out the voice of a suspect from a line-up of voice samples. Voice parade studies such as [4, 11] for familiar listeners and [7, 14] for unfamiliar listeners show certain target speakers being identified more readily than others: voices which sound more distinctive to a listener are presumably more likely to be recognised, yet the phonetic underpinnings of voice distinctiveness are not well understood.

[6] shows a correlation between listeners' judgements of voice distinctiveness and an acoustic metric combining f0, formants and harmonics-to-noise ratio. However, this study uses isolated utterances of *had* only, does not appear to account for demographic variation other than sex, and does not unpack the detail of the combined acoustic measure.

A study by Sørensen [13] focusses on the role of mean f0 in Danish listeners' ability to recognise unfamiliar male speakers within a group homogeneous for accent. Using a population distribution of mean f0, her experiment compared recognition of target speakers with centrally positioned mean f0 values with those with mean f0 values from the upper or lower tails of the distribution. Mid f0 voices were recognised correctly in 56% of trials, improving to 74% for high or low f0 voices. Sørensen interprets this as showing that mid f0 voices are harder to remember and recognise than those with a more extreme mean f0. This may be the case, but further evidence is needed; it is not clear how many different target voices were used and it is possible that further individual factors were at play. A crucial question arising from Sørensen's work is whether the relevant population for a voice parade is the population of foil voices used in the parade or the whole population.

The present study investigates whether, for a population homogeneous for accent/demographic characteristics, the distinctiveness of a voice is linked with its position in the population's distribution of mean f0 values. It tests whether speakers at the extreme ends of the f0 distribution are judged as sounding more distinctive from one another than speakers in the centre of the distribution through two experiments. In Experiment 1, listeners judge the similarity of pairs of speakers from within and between the lowest, highest and central parts of a population distribution for f0. In Experiment 2, listeners judge the similarity of the same group of speakers whose speech has been resynthesized across low, medium and high f0 positions.

2. EXPERIMENT 1

2.1. Method

2.1.1. Speakers

The *DyViS* database [9] provides a population of 100 matched-demographic speakers: male, aged 18-25 years, Standard Southern British English accent. Using 3-5 minutes net speech per speaker from a semi-spontaneous telephone call task (*DyViS* Task 2), the mean f0 was calculated for each speaker [5] using *Praat* [3]. f0 values were converted from Hertz to semitones (st, base 50Hz) to reflect the perceptual

scaling of pitch [8]. 12 speakers were selected, four from each of the lowest (speaker L1: mean f0 8.56st, L2: 8.77st, L3: 9.18st, L4: 9.38st), highest (H1: 16.7st, H2: 17.2st, H3: 17.3st, H4: 17.5st), and central (M1: 12.7st, M2: 13.00st, M3: 13.2st, M4: 13.5st) parts of the mean f0 distribution of the *DyViS* population. Each voice was reviewed auditorily to confirm that it did not sound particularly unusual (other than its pitch) before selecting it for the group.

2.1.2. Materials and experimental design

For each speaker, two three-second samples (U1 and U2) were created using short sections from the Task 2 recordings. Samples were controlled for content between speakers, taken from discussion of the same events prompted by the Task 2 scenario.

A similarity judgement task was built and hosted on Gorilla [1]. Each speaker was paired with himself and all other speakers, creating 78 pairings. A stimulus was prepared for each pairing, containing U1 and U2 separated by a silence of one second. Participants were randomly and evenly assigned to one of two groups, whereby in group 1 voice A appeared first, then voice B, and vice versa in group 2. Listeners rated the (dis)similarity of the two voices in each stimulus pairing on a Likert scale from 1 (very similar) to 9 (very dissimilar), having been asked to take into account voice quality and accent, and to ignore any meaningful content of the utterances, i.e. to focus on the sound of the voice. After six practice judgements, where listeners heard all 12 voices and became familiar with the task, they heard all pairings in a randomised order, with five evenly spaced rest breaks amongst the stimuli. The experiment took on average 17 minutes to complete.

2.1.3. Listener participants

35 participants aged 18-40 years (18 female, 16 male, 1 undisclosed) were recruited using the online recruitment platform Prolific.co. They lived and had spent most of their lives in the UK. They were native British English speakers with no reported hearing difficulties. Participants undertook a test [15] to ensure they were wearing headphones.

2.2. Results

Multi-dimensional scaling [12] was applied to the Likert scale judgements to reduce the many pairwise distances in a smaller number of dimensions. The relative positions of the 12 speakers according to the first two (perceptually most important) MDS dimensions are shown in Figure 1. Speakers in the same pitch group are clustered close together, particularly along dimension 1, with some overlap for

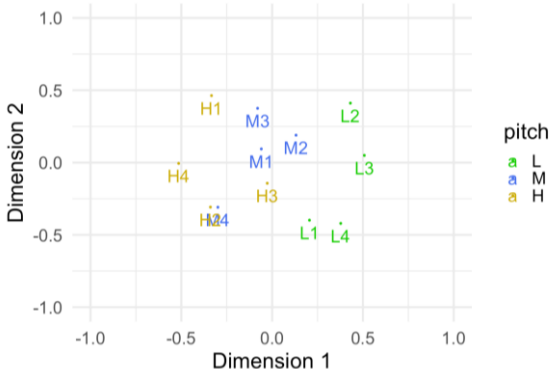


Figure 1: Scatterplot of the 12 speakers' locations on the first two dimensions (of five) produced by MDS using listeners' judgements of the speaker pairings [S-stress = 0.08299, Dispersion Accounted For (D.A.F.) = 0.97851]

the high- and medium-pitched speakers. Medium group speakers fall centrally in the display, with the high group overlapping and to the left of the overall collection of speakers, and the low group on the right side of the collection. In other words, along Dimension 1, the speakers are distributed approximately according to their pitch relationships, but neither the high nor low groups are positioned in an extreme location.

Regarding within-pitch-group comparisons, the individual members of the high group were judged to be more different from one another than the members within the low and medium groups, with H1 to H4 being spread further apart than L1 to L4 and M1 to M4. The boxplots in Figure 2 illustrate this; the median Likert scale score for the high group was 7, while the low and medium groups each had a median Likert scale score of 3. A Kruskal-Wallis test shows pitch group has a significant effect on Likert scores ($\chi^2(2) = 127.71$, $p < 0.0001$). A post-hoc Dunn test shows that the high group is significantly different from both the low group ($p < 0.0001$) and the medium group ($p < 0.0001$), whilst medium and low are also significantly different from one another ($p = 0.0186$).

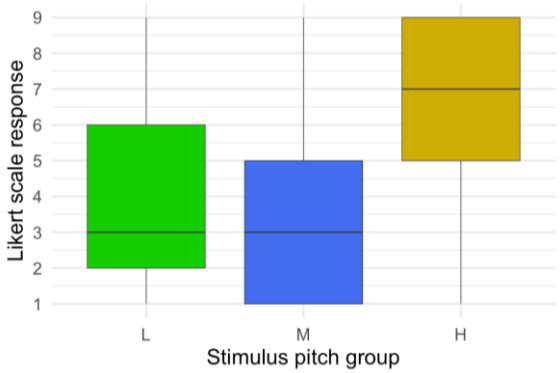


Figure 2: Boxplots of the Likert scale listener responses for within-group comparisons for each pitch group.

2.3. Interim summary

Voice (dis)similarity judgements show some clustering among speakers according to their pitch groups. However, the extent to which speakers are spread apart within each pitch group differs depending on the group. The clustering adds evidence that pitch plays a role in similarity judgements, while the spreading highlights the possibility that other speech dimensions are contributing. In particular, the speaker pairings within the high group are spread further apart in their Likert scale judgements than pairings in the other two pitch groups: could this be due to the high speakers differing more widely on speech dimensions other than pitch? To investigate this further, a second experiment was conducted in which similarity judgements were collected for the same speech stimuli which had been resynthesised after pitch-shifting. If speech dimensions other than pitch are driving the judgements, the high group speakers should always be the most different from one another even if they are resynthesised to have a low or medium pitch.

3. EXPERIMENT 2

3.1. Method

3.1.1. Materials and experimental design

The speech stimuli were all resynthesised so that there were three versions of each stimulus in low, medium and high pitch. For each stimulus, two of these resynthesized versions were a pitch-shifted version of the stimulus. The third was a resynthesised version of the stimulus at the same pitch as the original, undertaken to make sure that all stimuli sound equally 'unnatural' as a result of their resynthesis. The resynthesis process included all possible options as in Table 1. Stimuli under-went pitch-shifting in *Praat* using the 'Manipulate' and 'Shift pitch frequencies…' tool to shift the pitch contour by the appropriate number of semitones for the new median pitch value. (Note while speakers were chosen using *mean* f0 values, *Praat* uses *median* f0 for pitch-shifting). Speakers occupied the same relative position within a new pitch group, i.e., speaker L1 became speaker LM1 when shifted to medium pitch and LH1 when shifted to high pitch. Pitch values for the transformation were determined by the corresponding speaker in the new pitch group; for example, when being shifted to the new medium pitch group, speaker L1 was manipulated from his original median f0 of 8.56st to 12.7st, the original f0 of speaker M1. Stimuli were then evaluated to ensure naturalness after resynthesis. The same experimental set-up as for Experiment 1 was used, but in Experiment 2 listeners heard one of three different combinations of resynthesised stimuli. All listeners listened to the original 12 speakers whose stimuli had undergone resynthesis.

Resynthesis	Code
Low → Low	LL
Low → Medium	LM
Low → High	LH
Medium → Low	ML
Medium → Medium	MM
Medium → High	MH
High → Low	HL
High → Medium	HM
High → High	HH

Table 1: Pitch shift patterns used in Experiment 2.

The 12 speakers were heard in three pitch groups, Low, Medium and High, with four speakers in each. The combinations of resynthesised stimuli played to listeners are shown in Table 2. 20 listeners were

	Stimuli Group		
Stimuli Combination	Low-pitched	Medium-pitched	High-pitched
Combination 1	LL	HM	MH
Combination 2	HL	MM	LH
Combination 3	ML	LM	HH

Table 2: The three combinations of resynthesised stimuli, each heard by a separate group of listeners.

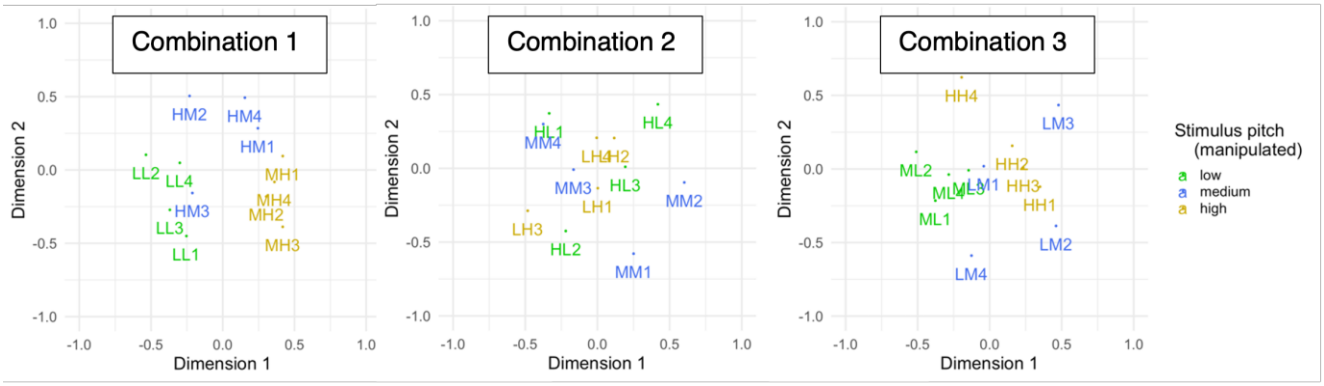


Figure 3: Scatterplots of the 12 speakers' locations on the first two dimensions (of five) yielded by MDS using judgements of the resynthesised stimuli for Combinations 1 (S-stress 0.08188, D.A.F. = 0.97865), 2 (S-stress .08789, D.A.F. = 0.97618) and 3 (S-stress .08477, D.A.F. = 0.97819).

randomly allocated to each stimulus combination.

3.1.2. Listener participants

60 participants (30 male, 29 female, 1 undisclosed), different individuals from Experiment 1 with the same characteristics, were recruited using Prolific.co.

3.2. Results

Scatterplots of the first two MDS dimensions for each stimuli combination are shown in Figure 3. In Combinations 1 and 3, speakers form relatively clear clusters according to their new pitch groups. In Combination 2, there is much more overlap among pitch groups, perhaps due to the fact that for the shifted speakers the pitch shift was the most extreme, i.e. low to high, and high to low.

Considering the relative positions of speakers, the shifted-to-medium group speakers (speakers labelled '…M') fall relatively centrally in each of the three graphs, but only for Combination 1 is there a sense of a low-medium-high continuum along Dimension 1, similar to that seen in Figure 1 for the natural stimuli. For Combination 3, the shifted-to-high pitch cluster largely falls within the datapoints of the shifted-to-medium cluster, while the shifted-to-low cluster is to the left of the overall collection of speakers. Meanwhile the display for Combination 2 does not show clear pitch-group-based clusters.

Figure 4 shows boxplots of the Likert scale judgements for the pitch groups in each stimuli combination. Reading each of the top, middle, and bottom panels of Figure 4 from left to right, it is clear that the originally H speakers (at right) still yielded the highest Likert scores regardless of f0 manipulation. It can be inferred that while pitch is clearly important, dimensions in addition to pitch are driving the similarity/difference judgements of these speakers. Reviewing the stimuli auditorily shows that

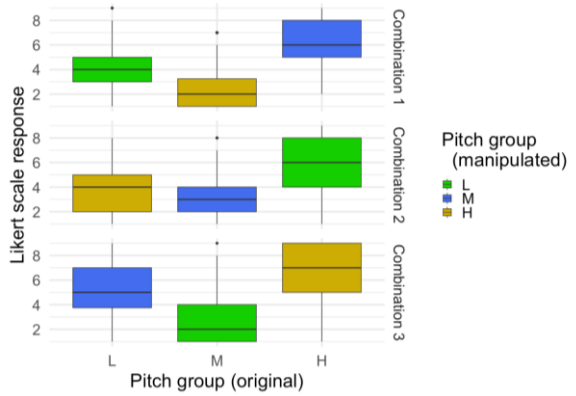


Figure 4: Box plot of the Likert scale listener responses for within-group comparisons for the low, medium and high pitch groups in Combinations 1 (top panel), 2 (middle panel) and 3 (bottom panel)

the originally high-pitched speakers have a greater range of voice qualities when compared with the originally low- and medium-pitched groups.

A Kruskal-Wallis test showed that pitch-shift status had no significant effect on Likert scores for all the within-pitch-group comparisons combined ($\chi^2(2) = 0.187$, $p = 0.665$). This test was repeated for the individual pitch groups, confirming that pitch-shift status had no significant effect on Likert scores for the originally low-pitched ($\chi^2(1) = 2.421$, $p = 0.1197$), medium-pitched ($\chi^2(1) = 3.741$, $p = 0.0531$), or high-pitched ($\chi^2(1) = 0.983$, $p = 0.3215$) groups.

4. DISCUSSION AND CONCLUSION

When listeners judged the (dis)similarity of all pairings among a group containing speakers of extreme low, extreme high, and mid mean f0 values, pitch clearly played a role, with speakers from the same pitch group clustering together, and an apparent 'low-mid-high pitch continuum' forming in correspondence with the most important MDS dimension [cf. 2, 10]. Neither the high or low groups were positioned remotely from other speakers on the MDS scatterplot, so Sørensen's contention that speakers with an extreme mean f0 value are perceived as more distinctive (and therefore more memorable) is not directly refuted, yet it is not unqualifiedly confirmed. The MDS scatterplots also exhibited varied extents of spreading of the speakers' locations within each pitch group, indicating that other phonetic dimensions are likely to be relevant.

Collecting (dis)similarity judgements for stimuli which had been resynthesised across the three pitch groups confirmed that while pitch is relevant, a more complex phonetic picture is at play. The within-pitch-group spread patterning observed for the original stimuli (larger spread of the H group speakers) was once again present and echoed in an auditory review, particularly with respect to voice quality.

The immediate implications of this study for voice parade construction are that a consideration of f0 is essential when selecting foils for comparison with a suspect, and that particular care should be taken in cases where the suspect's mean f0 is at the low or high extreme of the distribution for the relevant population, due to listeners' sensitivity to this feature. However, judgements of (dis)similarity are influenced by more than f0 alone [cf. 6], and further research is needed to uncover the complexity of the relationship between f0 and other phonetic features and speaker distinctiveness in order to better understand both what makes certain voices sound distinctive and the implications for earwitness procedures of the degree of perceived distinctiveness of a suspect's voice.

5. ACKNOWLEDGEMENTS

This research was supported by the UK Economic and Social Research Council as part of the project Improving Voice Identification Procedures (IVIP), reference ES/S015965/1. Additional funding was provided by the Isaac Newton Trust.

5. REFERENCES

[1] Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 31(1), 388-407.

[2] Baumann, O., Belin, P. 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research* 74, 110-120.

[3] Boersma, P., Weenink, D. 1992-2023. PRAAT: Doing phonetics by computer. [Computer program]. http://www.praat.org/.

[4] Foulkes, P., Barron, A. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics* 7(2), 180-198.

[5] Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. 2007. F0 statistics for 100 young male speakers of Standard Southern British English. In: Trouvain, J., Barry, W.J. (eds), *16th International Congress of Phonetic Sciences, Saarbrücken, August 6-10, 2007, Proceedings.* Universitaet des Saarlandes, 1809-1812

[6] Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P. 2013. Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology* 23(12), 1075-1080.

[7] McDougall, K., Nolan, F., Hudson, T. 2015. Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica* 72, 257-272.

[8] Nolan, F. 2003. Intonational equivalence: an experimental evaluation of pitch scales. In: Solé, M. J., Recasens, D., Romero, J. (eds), *15th International Congress of Phonetic Sciences, Barcelona, August 3-9, 2003, Proceedings.* Causal, 771-774.

[8] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The *DyViS* database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1), 31-57.

[9] Nolan, F., McDougall, K., Hudson, T. 2011. Some acoustic correlates of perceived (dis)similarity between same-accent voices. In: Lee, W.-S., Zee, E. (eds), *17th International Congress of Phonetic Sciences, Hong Kong, August 17-21, 2011, Proceedings.* City University of Hong Kong, 1506-1509.

[10] Rose, P., Duncan, S. 1995. Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics*, 2(1), 1-17.

[11] Schiffman, S.S., Lance Reynolds, M., Young, F.W. 1981. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. Academic Press.

[12] Sørensen, M.H. 2012. Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law* 19(2), 145-158.

[13] Thompson, C.P. 1985. A language effect in voice identification. *Human Learning* 4, 19-27.

[14] Woods, K.J.P., Siegel, M., Traer, J., McDermott, J.H. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception and Psychophysics* 79(7), 2064-2072.

EVALUATING DISCRETE FORENSIC VOICE EVIDENCE: A PRELIMINARY INVESTIGATION BASED ON FILLED PAUSE OCCURRENCE

Michael Carne

Speech and Language Laboratory, The Australian National University
michael.carne@anu.edu.au

ABSTRACT

There are a variety of statistical methods for generating likelihood ratios (LR) from continuous acoustic data for the purposes of forensic voice comparison (FVC). Few methods exist for evaluating voice evidence where the data are discrete and those that do, remain largely untested. The pattern of filled pause (FP) occurrence in speech is a behavioural characteristic that is potentially individuating and is an example of discrete voice evidence. This paper evaluates the evidential value of FP occurrence (*um* /ɐːm/, *uh* /ɐː/) in Australian English (AE) using a Poisson-Gamma model for LR estimation. Results indicate that while speakers do exhibit idiosyncratic patterns of FP occurrence, the discriminatory potential of the feature set is limited. The paper suggests that including a larger set of speech disfluency-based features may yield improvement, as well as fusion with acoustic measurements. Limitations associated with the statistical model are also canvassed.

Keywords: forensic voice comparison, likelihood ratios, filled pauses, speech disfluency.

1. INTRODUCTION

In a likelihood ratio-based (LR) approach to evidence evaluation [1] [2], forensic material (e.g DNA, fingerprints, voice recordings) from a known source and unknown source are compared with the goal of determining how much more likely the two are from the same-origin (H_p) *vs.* different origins (H_d). Given evidence (E) in the form of forensic sample(s), H_p and H_d are evaluated as ratio of conditional probabilities: $p(E|H_p)/p(E|H_d)$) (i.e. the LR term in Bayes Theorem). In FVC, E is usually quantified by set of measurements from continuously-valued acoustic features. Typically these are either measurements based on phonetically informed features (e.g. vowel formants) or automatic acoustic features derived from signal processing techniques (e.g. MFCC). Acoustic features capture speaker-specific characteristics associated with anatomical differences between vocal tracts. A variety of methods for LR estimation exist for continuous acoustic variables, including MVKD [3], GMM-UBM, PLDA *i*-vector and *x*-vector LR estimation [4].

Speaker-specific characteristics are also reflected in longer term patterns of language use and are quantifiable discretely; either in terms of their presence or absence in speech, or by the frequency of their occurrence. These characteristics include things like a speakers' habitual lexical and syntactic choices, discourse patterns, conversational style, patterns of speech disfluency and so forth. Auditory phonetic features that specify pronunciation differences between speakers (e.g. /sɪti/ *vs.* /sɪɾi/ in AE), delineate regional accents and non-obligatory phonological processes are other examples. Little attention has been given to evaluation of discrete voice data within the LR framework. In [5] the evidential value of lexical choice in AE was evaluated. [6] evaluates the occurrence of clicks (paralinguistic/discourse level usage) in British English. In this paper a Poisson-Gamma LR model described in [6] is used to evaluate the evidential value of AE FP occurrence for the first time. The aim is to determine how well can we distinguish one speaker from another based on the occurrence of *ums* and *uhs* in speech applying an LR-based approach to evidential evaluation. In doing so, it builds previous work by evaluating a new feature type for FVC, as well as on [7] by demonstrating a feature-based procedure for LR estimation for discrete data.

1.1. Speech disfluency and forensics

Disfluency is a natural part of spontaneous speech. Speaker's hesitate, repeat words, or parts of words, and use various other strategies to delay or prolong speech. FPs (e.g *ums* and *uhs*) are a category of this phenomenon for which various explanations exist. There is evidence FPs serve pragmatic and discourse functions [8][9]. For instance, FPs are

often used to indicate when one is about to start or continue speaking. Cognitive studies suggest FPs are a strategy that allow time for a speaker to synchronize cognitive and speaking processes [9], as well as potentially play a role in listener's language processing [10]. The occurrence of FPs also appear to be linked to indexical factors such as age, sex and education [11]. Forensic phonetician's interest in FPs stems from evidence that individual speakers vary in the choice of filler words, patterns of occurrences [12][13] as well as acoustically [11]. While there are several studies evaluating the acoustic properties of FPs within the LR framework [11][14][15], there is only one evaluating FP occurrence [7]. A shortcoming of [7] though is that it employs a score-based LR estimation method, which only accounts for the similarity, not the typicality of forensic samples [16][17][18]. An alternative is a feature-based approach which estimates LRs directly from the feature values and incorporates typicality, such as the Poisson-Gamma model evaluated here.

2. METHODS

2.1. Data, pre-processing & feature extraction

The data consists of 208 speech transcriptions from 104 normally fluent AE male speakers (2 x non-contemporaneous recording sessions per speaker). Transcriptions were made from unscripted conversational speech obtained from [19]. The recordings varying between 3-5 minutes for each conversation-side. For these experiments transcriptions of the first 2 minutes are used. Counts of *um*, *uh* and *total FPs* (i.e. the total count of both) were vectorised separately for each speaker and each transcription. Summary statistics of tokens elicited from each session are given in Table 1.

recording	feature	$\bar{x}$	sd	range
1	uh	4.95	4.45	0 – 22
1	um	6.63	4.85	0 – 26
2	uh	4.56	4.19	0 – 20
2	um	6.82	5.53	0 – 24

Table 1: FP occurrence statistics (N speakers = 104)

2.2. Poisson-Gamma LR model

The following description is adapted from [6]. Let $x = (x_1, ..., x_{k_x})$ be the count of *ums* or *uhs* from the suspect (or known) recording and $y = (y_1, ..., y_{k_y})$ those from the offender (or unknown) recording. Counts of each are taken from consecutive time periods (in minutes) from the recordings, with k_x and k_y representing the number of periods sampled in each. To estimate the LR we are concerned with two quantities from the voice evidence t_x and t_y, which are the total number of occurrences from the recordings and given by $t_x = \sum_{i=1}^{k_x} x_i$ and $t_y = \sum_{i=1}^{k_y} y_i$. Assuming the counts in consecutive periods are independent and follow Poisson distribution (the implications of this assumption are discussed in Section 4), evidential value can be evaluated using the Poisson-Gamma LR model given in 1.

(1)
$$LR = \frac{\Gamma(\alpha + t_x + t_y)\Gamma(\alpha)}{\Gamma(\alpha + t_x)\Gamma(\alpha + t_y)} * \frac{(\beta + k_x)^{\alpha + t_x}(\beta + k_y)^{\alpha + t_y}}{\beta^{\alpha}(\beta + k_x + k_y)^{\alpha + t_x + t_y}}$$

where, Γ is a gamma distribution defined by α (shape parameter) and β (rate parameter). The Γ distribution assesses the typically of samples under comparison and is estimated using reference data from the relevant speaker population. In these experiments α and β were estimated via maximum likelihood estimation.

2.3. Testing and validation

The following features are evaluated: (1) counts of *um*; (2) counts of *uh* (3) total FP count (*um* + *uh*) and; (4) fused *um* and *uh*. The scores estimated from the Poisson-Gamma model were calibrated using a logistic-regression procedure [20]. In the case of (4) the procedure simultaneously calibrates and fuses LRs. The data were randomly split into test (20%), training (40%) and reference data (40%) and 5-fold cross-validation applied. The test data was used to simulate suspect and offender comparisons, the reference data to assess the typicality of the samples being compared and the training data to calculate weights for the calibration-fusion procedure. All features were tested using 60 and 120 seconds of netspeech.

2.4. Evaluation

The log-likelihood ratio cost function (C_{llr}) [20] and the Equal Error Rate (EER) are used to evaluate performance. C_{llr} is an information-theoretic measure of the validity (accuracy) of voice comparison systems. C_{llr} can be decomposed into two additional measures C_{llr}^{min} and C_{llr}^{cal} ($C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$) which tell us the contribution of discrimination and calibration loss to the overall performance. The

C_{llr} value is a gradient metric which is small when LRs support the correct hypothesis (e.g. same-speaker comparisons are assigned an LR greater than 1) and increases as the magnitude of counter factual LRs do. A $C_{llr} = 0$ indicates perfect accuracy. Accuracy is worse as C_{llr} approaches and exceeds 1. The EER is a statistic commonly used in the evaluation biometric systems and is the point at which the false acceptance (or false positive) and false rejection (false negative) rates are equal; the lower the EER, the better the accuracy.

3. RESULTS

Table 2 shows the C_{llr} and $\%EER$ values for 60 and 120 seconds of net speech for each of the feature parametrisations described in Section 2.3. Results are mean values following a 5-fold cross-validation. The C_{llr} values are rather high, ranging between 0.94 and 0.98 bits. C_{llr} values are less than one though, indicating that some speaker-specific is present in FP occurrence albeit very little given values are close to 1. Commensurate with this result, the high $\%EERs$ reflect rather poor discrimination accuracy; the average $\%EER$ is 36.76%. The best performance is achieved via fusing FP features (*um* and *uh*) (C_{llr} = 0.94). The same result is obtained for both 60 sec. and 120 sec. of net speech. However, a slightly lower $\%EER$ is achieved (33.98% *vs.* 34.16%) and discrimination performance is marginally better with 120 sec. ($C_{llr}^{min} = 0.8$ *vs.* 0.83).

Table 2: Performance metrics for FP features for 60 and 120 seconds of net speech.

	feature	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	$\%EER$
60*sec.*	fused	0.94	0.83	0.11	34.16
	uh	0.96	0.85	0.11	38.79
	um	0.98	0.89	0.09	38.93
	total	0.96	0.85	0.10	36.13
120*sec.*	fused	0.94	0.80	0.14	33.98
	uh	0.95	0.84	0.11	35.81
	um	0.98	0.87	0.12	37.49
	total	0.98	0.84	0.15	38.86

The amount of net speech used does not result in meaningful performance gains. The additional minute of speech data only yields improvement in C_{llr} for *uh* and it is very slight - only 0.01 bits (C_{llr} = 0.96 *vs.* 0.95). For *um* C_{llr} is the same (0.98) and for *total FPs* it is worse (0.96 *vs.* 0.98). We do though see an improvement in discrimination performance in the 120 sec. condition, where C_{llr}^{min} values overall are lower. However, this improvement comes at the cost of poorer calibration which is evident from the higher C_{llr}^{cal} values obtained. Overall, it appears the strength of evidence obtained from FP features is rather weak. This can be appreciated in more detail via the Tippett plot in figure 1 which shows the cumulative proportion of same- and different speaker trials for 120 secs. of speech data for the best performing settings (i.e. fused, 120 secs.).

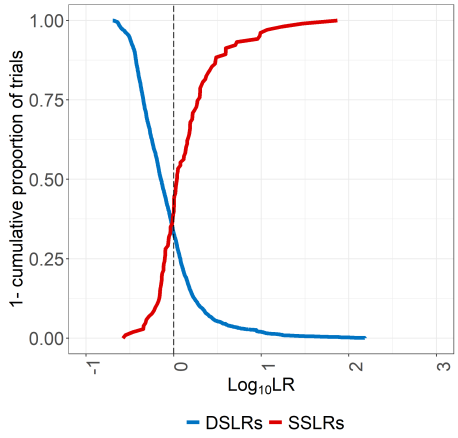


Figure 1: Tippett plot showing $Log_{10}LRs$ from SS (red line) and DS (blue line) comparison trials based on fused FP features (*uh* + *um*) from 120 seconds of net speech.

In figure 1 blue lines converging from the left are LRs from different-speaker trials (DSLRS) and red lines converging from the right are same-speaker trials (SSLRs). The *y*-axis shows the cumulative proportion of trials plotted as a function of $Log_{10}LR$ on the *x*-axis. The minimum $Log_{10}LR$ achieved for DS comparisons is -0.69 and overall a large proportion of $Log_{10}LR$ from the DS comparisons are close to the decision threshold (i.e. 0). This indicates most comparisons yield very weak support for the DS hypothesis. The magnitude of $Log_{10}LRs$ supporting the incorrect hypothesis (i.e. $Log_{10}DSLRs$ greater than zero) are also relatively large (the maximum $Log_{10}DSLR = 2.18$). The situation is slightly less severe for the SS comparisons in that the magnitude of counter factual SSLRs is less (minimum SSLR = -0.58). However, the strength of evidence for correctly assessed SS pairs is weak. The maximum $Log_{10}SSLR$ achieved is 1.87). Figure 2 shows Tippett plots from the voice comparisons for individual FP features. Figure 2A and figure 2B (*um*) are components of the fused system shown in figure 1 (*uh*). The results for *um* (figure 2A.) show a similar configuration to what we have just seen for the fused result. That is, $Log_{10}DSLRs$ yield relatively weak strength of

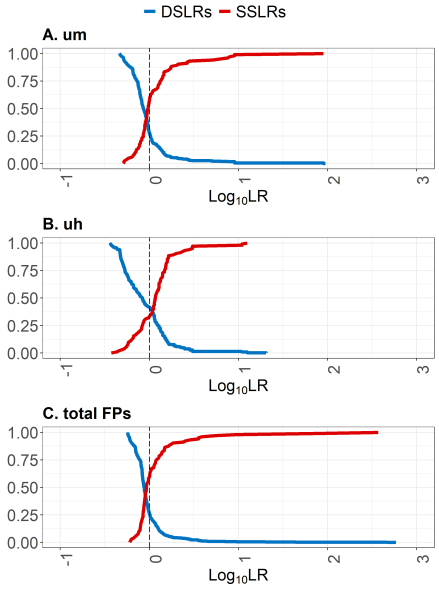


Figure 2: Tippett plots showing $Log_{10}LRs$ from SS (red line) and DS (blue line) comparison trials based for individual FP features (*uh* , *um* and *total FPs*) from 120 seconds of net speech, *y*-axis = 1 - cumulative proportion of trials.

evidence with a large proportion of trials supporting the contrary-to-fact hypothesis. The minimum LR is slightly weaker ($Log_{10}DSLRs$ = -0.33) than in the fused system, but this is to be expected. SS comparisons fair better and produce almost identical results to SS comparisons in the fused system in terms of LR magnitude. For *uh* (figure 2B), LRs are less biased towards the SS comparisons evident from the more symmetric appearance of the Tippett lines for SS and DSLRs. Here the magnitude of LRs for SS comparisons is slightly weaker than in figure 2A, but this is balanced by a reduction in the magnitude of counter-factual DSLRS. The magnitude of counter-factual DSLRs for *total FPs* (figure 2C) by contrast is substantial relative to the tippetts in 2A and B.

4. DISCUSSION & CONCLUSIONS

The results indicate AE FP occurrence is of relatively limited forensic use given the poor strength of evidence that is obtainable. This is reflected in high C_{llr} values (close to 1) and $\%EER$ observed. That said, these experiments have only considered a narrow set of speech disfluencies. Discriminant analysis in [13] demonstrated that speaker classification rates are improved substantially when several types of disfluencies are combined (e.g. word repetitions, silent and FPs, prolongations etc.). Similarly, in a previous LR-based study of FP occurrence in Japanese [7] substantial improvements are seen in C_{llr} values as the number of FPs used in LR estimation is increased. In fact, evidence of this can been seen in the results presented in this paper where logistic regression fusion of *um* and *uh* is shown to yield better performance. Also worth exploring is fusion of disfluency-based features with acoustic and temporal information. It was shown in [11] that formant and nasal duration information from *um* yields low C_{llr} and $\%EER$ values. It has also been shown that F_0 associated with FPs [12] varies between speakers. Incorporating this information into the voice comparison would also likely be beneficial. While acoustic features require continuous methods for LR estimation, LRs evaluating different kinds of evidence (from different models) can be combined to give a measure of the overall strength of evidence; either directly through multiplication of LRs or, where evidence is correlated, via a fusion procedure such as the logistic-regression procedure used here. In this way, FP occurrence and other disfluency features may prove useful in providing complementary information to acoustic-based systems; particularly in instances where data is limited.

Another set of limitations relate to the statistical model used to generate LRs. First, the Poisson-Gamma model assumes FPs occurrence in consecutive periods are independent. In reality, FP occurrence is very likely dependant on factors such as discourse structure. For example, where FP are used to signal turn-taking, the occurrence of FPs will be dependant the discourse structure between interlocutors - this is likely to vary minute-to-minute. Second, Poisson-Gamma is a univariate model and therefore LRs were calculated separately for *um* and *uh*. This ignores the potential for interactions between individual speakers' choice of FP. For example, it may be that speakers exhibit a preference for *uh* over *um* and do so in different contexts. A multivariate discrete LR model, such as a multinomial LR, may be more appropriate to better capture these kinds of interactions.

This paper has evaluated the evidential value of FP occurrence in AE using a discrete model for LR estimation. The feature set was shown to be a limited forensic value, but may complement acoustic-based systems where data is limited.

5. ACKNOWLEDGMENTS

The author's research is supported by an Australian Government Research Training Scholarship and

ANU Supplementary Scholarship. Dr. Shunichi Ishihara provided valuable comments on a draft version of this paper. I would also like to thank the reviewers for their constructive comments.

6. REFERENCES

[1] I. W. Evett, "Towards a uniform framework for reporting opinions in forensic science casework," *Science and Justice*, vol. 38, no. 3, pp. 198–202, 1998.

[2] I. W. Evett, J. A. Lambert, and J. S. Buckleton, "A bayesian approach to interpreting footwear marks in forensic casework," *Science and Justice*, vol. 38, no. 4, pp. 241–247, 1998.

[3] C. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Journal of the Royal Statistical Society: Series C (Applied Statistics))*, vol. 53, pp. 109–122, 2004.

[4] G. Morrison, E. Enzinger, D. Ramos, J. GonzÃ¡lez-RodrÃguez, and A. Lozano-DÃez, *Statistical models in forensic voice comparison*. Boca Raton, FL: CRC, 2020, book section 21.

[5] M. Carne, Y. Kinoshita, and S. Ishihara, "High level feature fusion in forensic voice comparison," in *Proc. Interspeech 2022*, 2022, pp. 5293–5297.

[6] C. Aitken and E. Gold, "Evidence evaluation for discrete data," *Forensic Science International*, vol. 230, no. 1-3, pp. 147–155, 2013.

[7] S. Ishihara and Y. Kinoshita, "Filler words as a speaker classification feature," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, M. Tabain, J. Fletcher, D. Grayden, J. Hajek, and A. Butcher, Eds., 2010, pp. 34–37.

[8] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[9] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[10] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[11] V. Hughes, S. Wood, and P. Foulkes, "Strength of forensic voice comparison evidence from the acoustics of filled pauses," *International Journal of Speech Language and the Law*, vol. 23, no. 1, pp. 99–132, 2016.

[12] A. Braun and A. Rosin, "On the speaker-specificity of hesitation markers," in *The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow., 2015, pp. 731–736.

[13] K. McDougall and M. Duckworth, "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16–27, 2017.

[14] X. Wang, V. Hughes, and P. Foulkes, "The effect of speaker sampling in likelihood ratio based forensic voice comparison," *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 97–120, 2019.

[15] B. X. Wang, V. Hughes, and P. Foulkes, "The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison," *Speech Communication*, vol. 138, pp. 38–49, 2022.

[16] G. S. Morrison and E. Enzinger, "Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality," *Sci Justice*, vol. 58, no. 1, pp. 47–58, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29332694

[17] A. Bolck, H. Ni, and M. Lopatka, "Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic mdma comparison," *Law, Probability and Risk*, vol. 14, no. 3, pp. 243–266, 2015.

[18] S. Ishihara and M. Carne, "Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods," *Forensic Science International*, vol. 334, pp. 111–268, 2022.

[19] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, D. Chow, and A. Szczekulska, "Forensic database of voice recordings of 500+ australian english speakers (auseng 500+)," 2021. [Online]. Available: Available:http://databases.forensic-voice-comparison.net

[20] N. Brummer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2, pp. 230–275, 2006.

Speaker discriminatory power of voice quality acoustics under forensic conditions

Ricky K. W. Chan

Speech, Language and Cognition Laboratory, School of English, University of Hong Kong
rickykwc@hku.hk

ABSTRACT

Voice quality has been regarded as one of the most popular and useful features for forensic speaker comparison, but relevant empirical validation is limited, let alone the acoustic aspects of voice quality. This study assesses the evidential strength of spectral tilt and additive noise parameters under the likelihood-ratio framework. Speech data of 75 male speakers aged 18-45 were obtained from a forensically-oriented database of Australian English speakers in Sydney/New South Wales. Results show that, contrary to previous findings, spectral tilt and additive noise parameters generally carry limited speaker-discriminatory power, especially when speech style mismatch and/or non-contemporaneous recordings are involved. Implications for forensic speaker comparison are discussed.

Keywords: Forensic voice comparison, voice quality, likelihood-ratio, speech style mismatch, non-contemporaneous recordings

1. INTRODUCTION

The task of forensic voice comparison (FVC) mostly involves comparison of voices on disputed and known samples. The disputed sample typically contains an unknown voice of an offender (e.g. hoax call, ransom demand, threatening message), and the known sample typically involves the voice of a suspect captured during a police interview [1]. The goal of FVC is to assist the investigating authorities (e.g., police) or trier-of-fact (e.g., jury or judge) in deciding whether the known and unknown voices are from the same speaker or different speakers. One of the main goals in FVC research is to empirically test the speaker-discriminatory power of speech features under conditions that are typically found in forensic casework. This paper focuses on voice quality.

'Voice quality' generally refers to the quasi-permanent characteristics of one's speech running through all the sounds from the speaker as a result of various laryngeal and supralaryngeal settings [2]. Voice quality has been regarded as one of the most useful and popular features for FVC casework [1,3]. However, empirical tests of the evidential strength of voice quality features are surprisingly limited, not to mention the acoustic aspects of voice quality. A recent study [4] tested the evidential strength of harmonic and inharmonic components of laryngeal voice quality (VQ hereafter) acoustic parameters using contemporaneous speech data of 97 Southern British English male speakers. They found that the combination of these parameters could perform very well in distinguishing speakers, and telephone or mobile phone transmission only led to a small decline in performance. Also, the addition of voice quality information was found to improve MFCCs-based system performance, especially when transmission quality degraded. The present study extends these promising findings and explores how speech style mismatch and the use of non-contemporaneous recordings may affect the performance these parameters using Australian English speech data.

2. METHODS

2.1. Corpus

A forensically-oriented database of 552 Australian English speakers (332 females and 231 males at the time of writing) [5] was selected. Each speaker was recorded on one to three or more occasions based on the protocol proposed in [6]. In each recording session, each speaker completed three speaking tasks: casual telephone conversation with a friend/colleague (CNV), fax information exchange over the telephone, and pseudo-police interview (INT). For speakers recorded on more than one occasion, the time interval between the recording sessions was about two weeks.

75 speakers were chosen from the database based on the following considerations: age, gender, regional background, and the availability of non-contemporaneous recordings. Besides, only speakers who were recorded on more than one occasion (separated by at least a one-week interval) were chosen given the importance of using non-contemporaneous recordings in the evaluation of FVC systems [6]. Then we strived to control for speakers' age and regional background as far as possible. To this end, 75 male speakers aged between 18 and 45 were selected and most of them were from Sydney and other areas within the state of New South Wales. For each speaker, four recordings—the CNV and INT tasks recorded in two separate sessions (i.e. CNV1, CNV2, INT1, INT2)—were analyzed. We focused on these two tasks because the speaking

styles involved are typically found in forensic casework [6].

2.2. Feature extraction and parameterization

Vowel-only portions of the recordings were manually segmented and labelled in Praat. Approximately 33 seconds of net vocalic material was extracted per speaker per recording. VQ parameters reported in [4] were selected so that the findings will be comparable. They were extracted using VoiceSauce [7] with a 20ms window length and 10ms window shift.

- *H1-H2* and *H2-H4*. These are the amplitude differences between the first and second and second and fourth harmonics respectively.
- *H1-A1*, *H1-A2*, and *H1-A3*: these are the amplitude differences between the first harmonic and the spectral magnitude at the first, second and third formant respectively.

The harmonic/spectral amplitudes were corrected for formant frequencies and bandwidths. In general, the higher values of these measure, the greater the spectral tilt which suggests a higher degree of glottal spreading due to breathiness. Vice versa for glottal constriction due to creakiness. These five measures were combined as 'spectral tilt' for further analysis.

- *Cepstral peak prominence (CPP)*. a measure of cepstral peak amplitude normalized for overall amplitude. In principle, modal phonation has well-defined periodic waves that result in larger cepstral peaks, while breathy phonation is likely to have less well-defined ones and lower cepstral peaks. A larger CPP value indicates a more modal voice, whereas a smaller CPP value a breathier voice.
- *Harmonic-to-noise ratio (HNR)*. HNR captures the spectral noise level and is positively correlated to the degree of perceived breathiness. The HNR was extracted over 0-500Hz, 0-1500Hz, 0-2500Hz, and 0-3500Hz respectively, resulting in four separate measures.

These five measures were combined as the 'additive noise' for further analysis. Outliers were removed as they were deemed unrepresentative of the speakers' typical long-term voice quality characteristics. They were defined as data points that are three median absolute deviations away from the overall median, as opposed to the commonly used "the mean plus or minus three standard deviations" approach because the mean and standard deviation are strongly influenced by outliers [8].

2.3. Statistical analysis

To assess the evidential strength of voice quality parameters, the multivariate kernel-density (MVKD) formula [9] was used for same-speaker and different-speaker comparisons. Calibrations were conducted using logistic regression. The 75 speakers were randomly assigned to one of the three datasets: training, test, or reference set (25 speakers in each set). The procedure above was replicated 100 times with different speakers in the training, test, and reference sets, as it has been demonstrated that the reliability of system performance hinges on the speaker samples involved [10]. To test the effects of speech style mismatch and non-contemporaneous recordings, the evidential strength of VQ features were tested using three different sets of speech data: 1) CNV1 vs. CNV2 (same speech style, non-contemporaneous recordings); 2) CNV1 vs. INT1 (different speech styles, contemporaneous recordings); and 3) CNV1 vs. INT2 (different speech styles, non-contemporaneous recordings). System validity was evaluated based on log-LR cost (C_{llr}). The lower the C_{llr} value, the better the system performance. C_{llr} values close to or greater than 1 imply limited evidential value of the input parameter(s). Due to space constraints, equal error rate values will be presented in the conference.

3. RESULTS & DISCUSSION

CNV1 vs. CNV2: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.61	1.07	0.77	0.08
Additive noise	0.52	0.91	0.67	0.08
Spectral tilt + Additive noise	0.85	1.08	0.93	0.04

CNV1 vs. INT1: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.91	1.04	0.96	0.02
Additive noise	0.84	1.05	0.92	0.04
Spectral tilt + Additive noise	0.85	1.02	0.93	0.04

CNV1 vs. INT2: C_{llr}				
	Min	**Max**	**Mean**	**SD**
Spectral tilt	0.91	1.05	0.97	0.03
Additive noise	0.76	1.01	0.88	0.05
Spectral tilt + Additive noise	0.87	1.02	0.93	0.03

Tables 1 to 3: statistics of C_{llr} values across 100 replications with VQ parameters as input in CNV1 vs. CNV2, CNV1 vs. INT1, and CNV1 vs. INT2 respectively.

Tables 1 to 3 show the descriptive statistics of C_{llr} values based on the combined spectral tilt measures (5 parameters), the combined additive noise measures (5 parameters), and spectral tilt + additive noise (10 parameters) for the three comparisons. In general, all the input parameters yielded a rather small standard deviation in C_{llr} value (less than 0.1) across the 100 replications, suggesting that system performance using these parameters as input were generally stable (i.e. high system reliability). Specific results from the three sets of recordings are summarized below.

CNV1 vs. CNV2. With non-contemporaneous recordings and the same speech style, both spectral tilt and additive noise measures yielded good results, with the best replications returning C_{llr} values of 0.61 and 0.52 correspondingly. Surprisingly, using spectral tilt + additive noise as input led to worse system performance, with the lowest C_{llr} being 0.85.

CNV1 vs. CNV2. With contemporaneous recordings and mismatch in speech style, the spectral tilt and the additive noise measures only performed slightly worse, with the best replications producing C_{llr} values of 0.91 and 0.84 correspondingly. Using both spectral tile and additive noise as input did not seem to improve system performance (best performance: C_{llr} = 0.85).

CNV1 vs. INT2. The comparison of these two datasets most closely reflects real-life forensic situation where both speech style mismatch and non-contemporaneous recordings are involved. The spectral tilt measures did not appear to provide much speaker-discriminator information, with the best replication having a C_{llr} of 0.91. On the other hand, the additive noise measures performed better and yielded a C_{llr} of 0.76 in the best replication. However, adding spectral tilt measures to additive-noise-based system led to poor performance— a C_{llr} of 0.87 in the best replication.

4. DISCUSSION

The present study sought to determine the evidential strength of VQ acoustic parameters under the LR framework, and tested the effects of speech style mismatch and the use of non-contemporaneous recordings on system performance. In general, the additive noise measures performed slightly better than the spectral tilt measures. This suggests that additive noise measures carry more speaker-specific information and higher evidential strength for FVC. However, combining spectral tilt and additive noise measures led to worse system performances, suggesting that they may provide overlapping or even conflicting information for discriminating speakers. While it has been found that the harmonic components may interact with the inharmonic components (noise) in various frequency bands in the expression of VQ-based phonological contrasts in some languages [11]), VQ is not known to be used for phonation contrasts in Australian English. The interaction between the harmonic and inharmonic components of VQ in speaker characterization under forensically-relevant conditions deserves further investigation.

[4] used the same parameters but only with contemporaneous speech data that involved speech style mismatch (conversation vs. interview; similar to CNV1 vs. INT1 in this study). They found that the spectral tilt and the additive noise measure carry considerable speaker-discriminatory information, and system validity based on these input features was only slightly affected by channel mismatch and deterioration of recording quality. However, our findings are less promising, and this may be attributable to the methodological differences between the two studies such as the number of speakers (75 vs. 97), length of speech material per recording (approximately 33s vs. 60s), the mathematical models used for speaker modelling (MVKD vs. the Gaussian Mixture Model-Universal Background Model (GMM-UBM)), and the variety of English (Australian English vs. standard southern British English) involved. The last of these suggests that the evidential strength of VQ parameters may be language/variety-specific and cautions should be exercised when generalising the results to other languages.

In addition, the present study involves two more sets of comparison to determine the effects of speech style mismatch and non-contemporaneous recordings on the speaker-discriminatory power of VQ acoustic parameters. VQ parameters performed relatively well in CNV1 vs. CNV2 (same speech style, non-contemporaneous recordings). However, results are much less promising when speech style mismatch is involved (CNV1 vs. INT1 or INT2). The results of CNV1 vs. INT1 and CNV1 vs. INT2 are similar. These findings suggest that speech style mismatch have more detrimental effects than non-contemporaneous recordings in this study, and that the involvement of non-contemporaneous recordings does not have clear effects on system performance. This highlights the fact that aspects of voice quality effects are not only ingrained by habits and bound by physiological limits, but may also be voluntarily manipulated in different speaking styles. The within-speaker variation involved may render two recordings of the same speaker sound very different. On the other hand, the negative effects of non-contemporaneous

recordings suggest that a speaker voice quality may change from occasion to occasion which can be random or conditioned [6], but these affect system performances to a lesser extent. Future research may investigate how the performance of VQ parameters may be affected by other speech styles typically found in forensic casework (e.g. speech under various emotional states), and with recordings separate by different time periods. With both speech style mismatch and non-contemporaneous recordings (i.e. CNV1 vs. INT2) as in typical forensic casework, all but the additive noise measure yielded high C_{llr} values, suggesting that the VQ parameters examined in this study may bear little speaker-discriminatory value in actual forensic cases, as least when Australian English is concerned.

At a broader level, this seems to be at odds with the claim among some forensic experts that voice quality is one of the most useful features for FVC casework [1,3]. Nonetheless, the present study only analysed laryngeal voice quality, with a focus on a number of spectral tilt and additive noise parameters. These parameters correspond to auditory VQ labels such as creaky voice and breathy voice, but they only constitute a portion of the voice quality characteristics normally analysed in actual forensic casework. [1] note that in typical voice quality analysis using a version of the Laver VPA scheme, around 38 speech features and vocal tract settings may be analysed auditorily. A comprehensive analysis of the acoustic correlates of all these features/settings is necessary in order to fully evaluate the value of voice quality analysis in forensic casework, and such endeavour will help transform the categorical auditory-based labels in VPA into continuous acoustic variables whose evidential strength can be easily assessed using the LR framework for forensic casework.

In the current study, the evidential strength of VQ parameters were assessed using multivariate kernel-density (MVKD) formula. Reanalysis of the data based on GMM-UBM is in progress and the results will be reported in the future.

5. CONCLUSION

Given the importance of empirically validating speech features to be used in FVC casework, the present study investigated the evidential strength of spectral tilt and additive noise parameters under the LR framework. It was found that these laryngeal voice quality parameters generally offer limited speaker-discriminatory value, particularly when speech style mismatch and non-contemporaneous recordings were involved. Forensic analysts should be cautious when using spectral tilt measures and/or additive noise measures as speaker discriminants.

6. REFERENCES

[1] French, P., & Stevens, L. (2013). Forensic speech science. In M. Jones & R.-A. Knight (Eds.), *The Bloomsbury Companion to Phonetics*. London: Bloomsbury, 183-197.

[2] Laver, J. (1980). *The Phonetic Description of Voice Quality*. New York: Cambridge University Press.

[3] Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, *18*(2), 293-307.

[4] Hughes, V., Cardoso, A., Foulkes, P., French, J. P., Harrison, P. & Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*. Melbourne, Australia.

[5] Morrison, G. S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B. K., De Souza, S., Cummins, N., & Chow, D. (2015). Forensic database of voice recordings of 500+ Australian English speakers.

[6] Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), 155–167.

[7] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

[8] Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

[9] Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series c (Applied Statistics)*, *53*(1), 109–122.

[10] Wang, B. X., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech Language and the Law*, *26*(1), 97–120.

[11] Garellek, M. (2019). The phonetics of voice. In W. Katz & P. Assmann (Eds.), *The Routledge handbook of phonetics*. Abingdon-on-Thames, UK: Routledge.

BETWEEN-SPEAKER SYLLABLE INTENSITY VARIABILITY IN PERSIAN

Homa Asadi[1], Batool Alinezhad

Department of Linguistics
University of Isfahan, Iran
{h.asadi; b.alinezhad}@fgn.ui.ac.ir

ABSTRACT

This study examines speaker-specific temporal features with a particular focus on the variability of syllable intensity in a Persian-speaking population. Two types of intensity variability measures (mean and peak) between syllables were examined as a function of speaker in two Persian databases with different sources of within-speaker acoustic variability. Results showed that the intensity measures that take into account the difference between all intervals in an utterance were better at revealing between-speaker variability than the measures based on the difference between successive intervals. Peak intensity measures were also found to be more speaker-specific. We also found that the degree of between-speaker rhythmic variability in terms of syllable intensity was not affected by the language-specific features of Persian. However, the discriminatory power of the intensity measures was reduced in the database with extreme rate variability. We discuss how results could be applied in forensic speaker comparison.

Keywords: speaker discrimination, intensity measures, vocal variability.

I. INTRODUCTION

Acoustic characteristics of voice with high variability between speakers and low variability within speakers are crucial to forensic speaker comparison (FSC) [1, 2, 3, 4]. In FSC, two samples of voices, a known and an unknown (disputed) sample, are compared to estimate the probability that the same speaker has produced the speech samples (same-speaker hypothesis) versus the probability that the speech samples have come from two different speakers (different-speaker hypothesis) [3]. This goal can be achieved by learning about acoustic parameters that discriminate between speakers well and show high stability within each speaker.

Previous studies provided strong evidence that acoustic measures of speech rhythm based on the durational characteristics of consonantal and vocalic intervals as well as on the syllable intensity characteristics can help distinguish different speakers [5, 6, 7, 8]. It has been demonstrated that anatomical differences in the structural dimensions of speakers' articulators, as well as acquired idiosyncratic ways of operating their speech articulators to produce sounds, will significantly alter the acoustic characteristics of speech and could, therefore, result in a high degree of between-speaker variability in temporal aspects of speech [5, 6, 7, 9].

Rhythmic variability between speakers is largely influenced by intensity in acoustic signals. Empirical studies based on audiovisual experiments have shown that mouth-opening size and vocal intensity are highly correlated and that articulatory movements of the lips and mouth opening are reflected in the temporal structure of the amplitude envelope [10, 11, 12, 13].

Acoustically, between-speaker variability in syllable intensity was first investigated by [7, 14]. They argued that individual articulatory movements could lead to a high degree of speaker specificity, which can be reflected in the intensity aspects of speech signals. They applied different intensity measures to the speech corpora with speakers from Zürich German and North German and found enormous variability between speakers in the two databases.

In addition to anatomical and learned differences that affect temporal characteristics of speech, cross-linguistic differences may also play a role in syllable intensity variability. It has been demonstrated that a language with a more phonotactically complex structure typically has a higher degree of intensity variability than languages with simpler structures [15]. Additionally, languages that allow vowel reduction have higher variability of syllable intensity because reduced vowels have lower amplitude envelopes, resulting in lower intensity levels in terms of mean intensity or peak intensity [16]. Another important factor affecting intensity is related to how lexical stress is signaled in a language. Languages that use intensity as an acoustic cue for signaling lexical stress have higher levels of intensity variability than those that do not utilize intensity for the production of stress [17].

Persian has a simple syllable structure of CV(C)(C), and it does not have a vowel reduction

pattern [18, 19, 20, 21, 22, 23]. In terms of stress pattern, duration is the most reliable correlate of stress in Persian, while intensity is a poor marker of stress position [24]. We thus hypothesize that the degree of between-speaker rhythmic variability in terms of syllable intensity is prone to be decreased in Persian because of less complex syllable structure, little to no vowel reduction, and different phonetic realization of stress.

In line with [7] and [14], we have selected two types of intensity measures and analyzed between- and within-speaker variability, posing the following questions:

1) Do intensity measures vary between speakers in Persian?

2) Do intensity measures remain stable across different sources of within-speaker variability?

3) Which intensity measures explain possible between-speaker variability best?

2. METHOD

2.1. Participants and task

Two speech corpora with different sources of within-speaker acoustic variability were collected for this study. In the first corpus (hereafter non-contemporaneous corpus), 20 male native speakers of Persian (age mean=31.2, SD = 4.3) were recorded in two sessions separated by a two-month time-lapse. Speakers were equipped with a fixed microphone and asked to read 40 Persian sentences at a normal rate with a three-second pause between sentences. Recording sessions were conducted in a soundproof booth with a sampling rate of 44.1 kHz and a quantization of 16 bits. For the second corpus (hereafter tempo corpus), following the procedure used in the collection of the BonnTempo corpus in German [25], 10 male native speakers of Persian (age mean=34.3, SD = 3.6) were instructed to read *The North Wind and the Sun* in Persian at five different speaking speeds (normal, slow, slower, fast, and fastest possible). Before each recording session, participants were asked to read the text several times to familiarize themselves with the passage. Speakers were then asked to slow their pace in two steps and then to read the text faster and as fast as possible. This resulted in strong syllable rate variability across the five different reading passages. The recording location and setup were the same as in the non-contemporaneous corpus.

2.2 Acoustic parameters

Speech materials were acoustically analyzed using Praat [26]. Speech tokens were annotated in segments, syllables, and peak tiers. We developed two sets of intensity variability metrics following the procedure used in [7, 14]. From the syllable tier, we calculated the stdevM, varcoM, nPVIm and rPVIm, while stdevP, varcoP, nPVIp and rPVIp were calculated from the peak syllable tier. To quantify the mean syllable intensity, we divided the sum of the intensity values between the onset and offset of a syllable by its duration. Peak intensity was calculated at the syllable peak point interpolated with the cubic function. We calculated the global quantification of intensity measurements based on the standard deviations of the mean and peak intensity of the syllables in an utterance. Local intensity variations were calculated by measuring syllable-to-syllable intensity differences. A detailed description of the intensity measurements is given in the appendix.

2.3. Statistical analyses

All statistical analyses were performed using R (R core Team, 2022) version 4.2.2 [27]. First, we conducted a principal component analysis (PCA) to see whether the different intensity measures formed independent categories (eigenvalues $\geqq 1$; rotation method = varimax). Second, we constructed a multinomial logistic regression (MLR) model on the collected speech data to address the question of which intensity measures could better explain the variability between speakers. We modeled the acoustic parameters as predictor variables and the speaker as a nominal response variable. The proportion of between-speaker variability explained by intensity measures was calculated using the likelihood ratio χ^2 of each acoustic parameter divided by the sum of the likelihood ratio χ^2s of all parameters. Furthermore, linear mixed-effects models were run to analyze the significance of within-speaker acoustic variability, i.e., repetition and tempo on intensity measures in our datasets. In the non-contemporaneous corpus, repetition was entered into the model as a fixed effect, and the speaker and sentence were treated as random factors, whereas in the tempo corpus, the tempo was considered as a fixed effect and the speaker and sentence as random factors.

3. RESULTS

3.1. Principal component analysis (PCA)

The PCA results from Table 1 show that three components were extracted for the non-contemporaneous corpus, with component 1 including all measures of mean intensity, while components 2 and 3 were primarily based on peak intensity measures. For the tempo corpus, two components were extracted. The first component contains measures computed over the entire

utterance, while local measures have emerged in the second component. This suggests that the different types of intensity measurements contain complementary information.

Table 1: PCA table showing a correlation matrix for variables loaded on the PCs extracted from the data analysis.

	Non-contemporaneous corpus		
	Comp1	Comp2	Comp3
rPVIm	0.952		
nPVIm	0.950		
varoM	0.786		
stdevM	0.772		
stdevP		0.921	
varcoP		0.920	
nPVIp			0.973
rPVIp			0.965
variance	**0.39**	**0.29**	**0.25**
	Tempo corpus		
	Comp1	Comp2	
varcoP	0.952		
stdevP	0.950		
stdevM	0.786		
varcoM	0.772		
nPVIm		0.892	
rPVIm		0.884	
nPVIp		0.885	
rPVIp		0.846	
variance	**0.43**	**0.38**	

3.2. Multinomial logistic regression

MLR analysis results show that the speaker's effect was significant in all investigated intensity measures. Nonetheless, these measures were not balanced in explaining between-speaker variability. As shown in Tables 2 and 3, the strongest effects were found for varcoP in both speech corpora. Based on the results, most acoustic variation is explained by peak and global measures.

Table 2: Summary of the results of MLR on intensity measures for non-contemporaneous corpus.

Intensity measures	-2 Log Likelihood of Reduced Model	χ2 [df]	P	Variability explained
stdevM	11476.703	193.510 [19]	<0.0001	12%
varcoM	11465.566	182.373 [19]	<0.0001	11.3%
rPVIm	11381.953	98.760 [19]	<0.0001	6.1%
nPVIm	11381.067	97.874 [19]	<0.0001	6.1%
stdevP	11638.987	355.795 [19]	<0.0001	22.1%
varcoP	11649.108	365.916 [19]	<0.0001	22.8%
rPVIp	11433.896	150.704 [19]	<0.0001	9.3%
nPVIp	11441.893	158.701 [19]	<0.0001	9.8%

Table 3: Summary of the results of MLR on intensity measures for tempo corpus

Intensity measures	-2 Log Likelihood of Reduced Model	χ2 [df]	P	Variability explained
stdevM	1187.803	130.399 [9]	<0.0001	13.9%
varcoM	1190.133	132.729 [9]	<0.0001	14.1%
rPVIm	1129.827	66.544 [9]	<0.0001	7.1%
nPVIm	1123.948	72.423 [9]	<0.0001	7.7%
stdevP	1298.924	241.52 [9]	<0.0001	25.8%
varcoP	1299.406	242.002 [9]	<0.0001	25.8%
rPVIp	1083.017	25.613 [9]	<0.0001	2.7%
nPVIp	1081.619	24.215 [9]	<0.0001	2.6%

3.2. Linear mixed-effects models

The results of within-speaker occasion-to-occasion variability analysis showed that the variability of the tested parameters as a function of repetition is not significant ($p>0.05$). Results also revealed that the effect of speaker on the intensity measures was significant across different speech rates. Further, we analyzed the effect of speaker at the five different speech rates separately. The result showed that speakers are differentiated well when speaking at normal, slow, fast and fastest rates, but they behave similarly when speaking at the slowest rate. Post hoc analyses using Bonferroni adjusted pairwise t-test were also applied to quantify the differences between speakers in both corpora. It has been shown that the number of significant paired comparisons is higher in the non-contemporaneous corpus. For example, for nPVIp, 215 of the 361 (59%) comparisons were significant ($p< 0.0001$) in the non-contemporaneous corpus, while 19 out of the 81 (23%) possible paired comparisons in the tempo corpus were significant ($p<0.05$). Comparing the results of the tempo corpus with those obtained in the non-contemporaneous corpus shows that between-speaker variability is drastically reduced in the dataset with high prosodic variability. Figure 1 shows the boxplot of between- and within-speaker variability for varcoP in both investigated corpora.

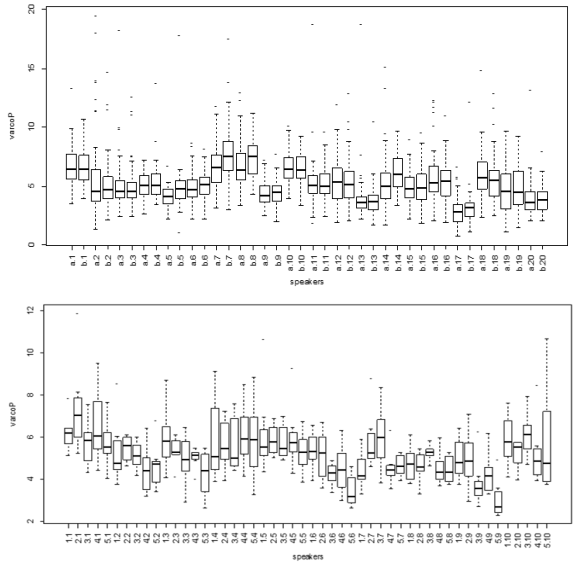


Figure 1: Boxplots of between- and within-speaker variability for varcoP in non-contemporaneous (top) and tempo corpus (bottom) (Rerecording sessions are shown with a and b. Speech rates are shown as follows: slow=1, slowest=2, normal=3, fast=4, fastest=5).

4. DISCUSSION AND CONCLUSION

In this study, we investigated between-syllable intensity variability as a source of between-speaker rhythmic variability in two Persian speech corpora with various sources of within-speaker variability. Our findings showed that intensity measures vary considerably and consistently among Persian speakers in both speech corpora. Results of our study based on the non-contemporaneous corpus replicate the previous studies on German and Swiss German databases [7]. As such, we hypothesized that language-dependent features, such as syllable structure complexity, vowel reductions, and the use of different acoustic cues to signal lexical stress, may affect speaker discrimination strength. However, our findings proved that such features do not affect the discriminatory power of intensity measures. This implies that intensity measures can discriminate between speakers regardless of their language of communication. We also found that speakers' performances were similar across their two recording sessions which is further indicative that intensity measures are robust to within-speaker acoustic variability caused by time-lapsing. Additionally, the results suggested that despite high prosodic variability in tempo corpus, speakers exhibited differences in intensity measures. Nevertheless, the obtained results are not totally in line with the previous studies [5, 7]. In Persian, intensity measures performed well in discriminating speakers when they were speaking at normal, fast, very fast and slow rates. Still, those measures were unable to identify speakers when they reduced their speed to the slowest possible rate. This is not the case for German speakers. Intensity measures could discriminate among German speakers in the BonnTempo corpus, a dataset of the same speech material with high within-speaker speech rate variability. A possible reason for the different performance of intensity measures at the slowest rate in German and Persian may be that Persian speakers uttered the passage with varying strategies of speed and loudness, which subsequently led to some changes in the intensity scores. On the one hand, speakers usually utilize less energy at a slow rate in moving their articulatory organs of speech, especially lips and the blade of the tongue. On the other hand, they have less control over the idiosyncratic muscular movement of speech organs. These will result in less speaker-specific information in the intensity variability across syllables in the utterances at the slowest rate.

In line with [7, 14], we also found that the measures that consider the difference between all the intervals in an utterance or sentence perform better in revealing between-speaker variability than those based on the difference between consecutive intervals. Our results also showed that the peak measures collectively contain more between-speaker variability than the mean measures in both databases. Intensity peaks are strongly correlated with articulatory correlates of jaw and tongue tip maximum displacements [13, 28, 29]. Thus, speaker-specific articulatory behavior may be more influential on the intensity peak that occurs when the mandible or tongue tip reaches its maximum displacement [14]. VarcoP yielded the best result for speaker discrimination in both Persian databases. This accords with the findings of [7], who also reported varcoP as a powerful speaker-specific variable with robustness against high prosodic variability. According to previous and current studies, we can conclude that varcoP is a powerful universal speaker-specific feature that goes beyond typological differences across languages.

Some implications of our results are relevant to the field of forensic phonetics. This study supports the notion that intensity measures can be used to determine speaker individuality. According to previous studies demonstrating that intensity measures can discriminate speakers in German and Swiss German, the results of this study in Persian suggest that speech rhythm intensity measures can be characterized as language-independent measures capable of being used when speaker-specific rhythms are not known.

6. APPENDIX

– The global intensity measures:
- stdevM: the standard deviation of average syllable intensity levels;
- stdevP: the standard deviation of syllable peak intensity levels;
- varcoM: the variation coefficient of average syllable intensity levels
- varcoP: the variation coefficient of syllable peak intensity levels

– The local intensity measures:
- rPVIm: the raw pairwise variability of adjacent mean syllable intensity levels;
- rPVIp: the raw pairwise variability of adjacent syllable peak intensity levels;
- nPVIm: the normalized pairwise variability of adjacent mean syllable intensity levels;
- nPVIp: the normalized pairwise variability of adjacent syllable peak intensity levels.

The formulae for calculating intensity variability can be found in [7, 14].

7. REFERENCES

[1] Wolf, J. 1972. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*. 51(6B), 2044–2056.

[2] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

[3] Rose, P. 2003. *Forensic Speaker Identification*. New York: Taylor & Francis.

[4] Morrison, G. S. 2010. Forensic Voice Comparison, In I. Freckelton & H. Selby, *Expert Evidence*. Sydney: Thomson Reuters.

[5] Dellwo, V., Leemann, A., Kolly, M. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*. 137, 1513–1528.

[6] Leemann, A., Kolly, M., Dellwo, V. 2014. Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Science International*, 238, 59–67.

[7] He, L., Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. *The International Journal of Speech, Language and the Law*. 23, 243–273.

[8] Asadi, H., Nourbakhsh, M., He, L., Pellegrino, E., Dellwo, V. 2018. Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals. *International Journal of Speech, Language and the Law*. 25(2), 151–174.

[9] Dellwo, V., Huckvale, M. and Ashby, M. 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification". In C. Müller (Ed), *Speaker Identification 1*, 1-20, Berlin: Springer Verlag.

[10] Summerfield Q. 1992. Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London Biological Sciences*. 335(1273):71–8.

[11] Grant KW, Seitz PF. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*. 108(3),1197–208.

[12] Garnier, M., Wolfe, J., Henrich, N., Smith, J. 2008. Interrelationship between vocal effort and vocal tract acoustics: a pilot study. *Proceedings of INTERSPEECH* Brisbane, 2302–2305.

[13] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. 2009. The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5(7), e1000436.

[14] He, L., Dellwo, V. 2014. Speaker idiosyncratic variability of intensity across syllables. *Proceedings of INTERSPEECH* Singapore.

[15] Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., Post, B. 2012. Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English and Spanish. *Speech Communication*. 54, 681–702.

[16] He, L. 2017. *Speaker Idiosyncratic Intensity Variability in the Speech Signal*. Ph.D. Dissertation, University of Zurich.

[17] Wang, Q. 2008. L2 stress perception: The reliance on different acoustic cues. *Speech Prosody* Campinas, 635–638.

[18] Windfuhr, G. L. 1979. *Persian grammar: History and state of its study*. New York: Mouton de Gruyter.

[19] Lazard, G. 1992. *Grammar of contemporary Persian*. Mazda Publishers.

[20] Sheikh Sangtajan, Sh., Bijankhan, M. 2010. The study of vowel reduction in Persian spontaneous speech. *Journal of Research in Linguistics*, 2(1): 35–48.

[21] Bijankhan, M. 2018. Phonology. In A. Sadeghi and P. Shabani-Jadidi (eds) *The Oxford Handbook of Persian Linguistics*; 111–141. Oxford: Oxford University Press.

[22] Yavaş, M. 2011. *Applied English phonology*, United Kingdom: Wiley-Blackwell.

[23] Sadeghi, V. 2015. A phonetic study of vowel reduction in Persian, *Language Related Research.* 30, 165–187.

[24] Sadeghi, V. 2011. Acoustic correlates of lexical stress in Persian. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)* Hong Kong 1738–1741.

[25] Dellwo, V. 2010. *Influences of speech rate on the acoustic correlates of speech rhythm: an experimental phonetic study based on acoustic and perceptual evidence*. Ph.D. Dissertation, Bonn University.

[26] Boersma, P., & Weenink, D. (2022) Praat: Doing Phonetics by Computer. Version 6.2.14, retrieved 24 March, 2021 from http://www.praat.org/.

[27] R Core Team R, A Language and Environment for Statistical Computing (version 4.2.2). 2022. R Foundation for Statistical Computing. http://www.Rproject.org.

[28] Birkholz, P., Kröger, B. J., Neuschaefer-Rube, C. 2011. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*. 9(5), 1422–1433.

[29] Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C., Suemitsu, A., Moore, J. 2015. Bridging articulation and perception: The C/D model and contrastive emphasis. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow.

SPEAKER IDIOSYNCRATIC INTENSITY AND MOUTH OPENING-CLOSING VARIATIONS: THE CASE OF ENGLISH

Yu Zhang, Lei He, Volker Dellwo

Department of Computational Linguistics, University of Zurich, Switzerland
{yu.zhang | lei.he | volker.dellwo}@uzh.ch

ABSTRACT

This study investigated speaker idiosyncrasy in intensity and mouth opening-closing variations using an English corpus containing both acoustic and articulatory data (19 speakers × 59 read sentences). The speeds of intensity as well as mouth opening-closing movements were calculated and summarized in terms of the mean, standard deviation, and pairwise variability index per sentence. Multinomial logistic regressions were used to test the speaker effect and evaluate the amount of between-speaker variability explained by each measure. It was found that all measures showed significant speaker effect. Moreover, the measures pertaining to the speeds of intensity and mouth opening-closing movements explained more between-speaker variability in English.

Keywords: speaker idiosyncrasy, intensity variability, mouth opening-closing variability, English.

1. INTRODUCTION

People differ in how they speak and what they sound like. The major processes in speech production – phonatory and articulatory activities – all contain speaker idiosyncratic information; such speaker idiosyncrasies leave traces in the speech signal and are thus measurable acoustically (see [1] for a general review). Salient between-speaker variabilities in various temporal acoustic features have been explained in terms of speaker idiosyncratic articulatory movements (e.g., [2, 3] for duration variabilities; [4, 5] for formant variabilities; [6, 7] for intensity variabilities,). This paper focuses on between-speaker variability in intensity variability and corroborates it from the articulatory movements of the mouth.

We followed the method for examining between-speaker intensity variability developed by He and Dellwo [7], who partitioned the intensity curve into rising and falling segments between peaks and troughs (peaks were identified between syllable boundaries; troughs were identified between peaks). The speeds of intensity increases and decreases for all rising and falling segments[1] were calculated using the Zurich German TEVOID corpus [2, 3]. They found that the speeds of intensity decreases explained more between-speaker variability than the speeds of intensity increases across an utterance [7]. Similar results have been obtained from a Thai corpus [8]. The authors suggested that speakers might differ more in mouth-closing movements [7, 8], as a high degree of covariation between intensity and mouth aperture size was observed [9]. Similar results were obtained from the same Zurich German corpus by calculating the first formant variability, another acoustic feature that covaries with mouth opening and closing movements [5].

For the current study, we aim to (1) investigate whether the speeds of intensity decreases also explain more between-speaker variability in English, and (2) whether the findings from intensity variations can be supported by the speeds of mouth opening-closing movements.

2. METHOD

2.1. Corpus

The sentence-reading part of the EMA-MAE corpus[2] [10] was used. Twenty native speakers of American English were enrolled; each participant read 59 sentences. Both articulatory and acoustic data were recorded simultaneously (The NDI Wave electromagnetic articulograph; $Fs_{\_NDI}$ = 400 Hz; $Fs_{\_acoust}$ = 22.05 kHz). After data cleaning, 19 speakers (9F, 10M) were kept with 1108 analyzable sentences in total.

2.2. Data Preprocessing

The intensity curve of each sentence was extracted using the Praat [11] function "Sound: To Intensity…" with default settings[3]. The intensity curve was then normalized such that the average intensity equated to 65 dB (SPL).

The mouth opening-closing movements were characterized using three NDI sensors at the upper lip (UL), lower lip (LL) and lip corner (LC), as indicated in Fig. 1. The area of the triangle formed

by these three sensors was used to estimate the magnitude of mouth opening (abbreviated as MO henceforth) at each instant in time based on the x, y and z coordinates of these sensors [15][4]. The resultant MO curve was smoothed by low-pass filtering at 10 Hz.

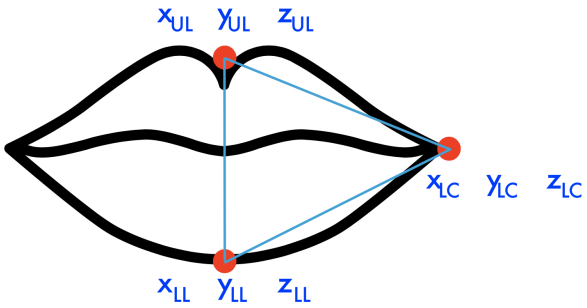


Figure 1: An illustration of estimating the mouth opening-closing movements from three NDI sensors at the upper lip, lower lip, and the lip corner. Endnote 4 shows the formula.

Syllable boundaries were placed using the BAS Web Services[5]. First, the WebMAUS Basic [12, 13] was used for phonemic transcriptions based on the acoustic signals and orthographic transcriptions. The annotations were saved in the BAS Partitur Format (*.par). The *.par files were then submitted to Pho2Syl [14] for the syllabification processing; syllable boundaries were saved in Praat TextGrids. The boundaries were cross-checked by the authors.

2.3. Speeds of change in intensity and mouth movements

For each sentence, the speeds of increase and decrease in intensity and mouth movement magnitude were calculated (i.e., the steepness of the arrows in red, purple and caramel in Fig. 2). For intensity changes, the peaks were pinpointed between syllable boundaries and the troughs were identified between consecutive peak points (see Fig. 2(a), the peak points were indicated by blue down arrows, and the trough point was indicated by a blue up arrow). The speeds of intensity increases from a trough point to its consecutive peak point (i.e., the steepness of the upward arrows in Fig. 2(a)) is notated as **v_I[+]**; the speeds of intensity decreases from a peak point to its neighboring trough point (i.e., the steepness of downward arrows in Fig. 2(a)) is notated as **v_I[–]**.

For the speeds in MO increases and decreases, two methods were followed: (1) the syllable-based method and (2) the derivative-based method.

Re (1) the syllable-based method: Similar to **v_I[+]** and **v_I[–]**, MO peaks were identified between syllable boundaries, and MO troughs were detected between peaks (see Fig. 2(b), the MO peaks were indicated by blue down arrows, and the trough point was indicated by an blue up arrow). The speeds of increases from an MO trough to its following peak (i.e., the steepness of the upward arrows in Fig. 2(b)) is notated as $\mathbf{v\_MO_{syl}[+]}$; the speed of MO decrease from a peak to the next trough (i.e., the steepness of the downward arrows in Fig. 2(b)) is notated as $\mathbf{v\_MO_{syl}[–]}$.

Re (2) the derivative-based method: The first derivative of the MO curve (abbreviated as derMO hereafter) was approximated in terms of the difference between the $(n+1)^{th}$ and the nth samples. An MO peak was pinpointed where the derMO crossed zero with a downward trend; an MO trough was pinpointed where the derMO hit zero with an upward trend [16] (Fig. 2(d) indicates the peaks by blue down arrows, and a trough by an blue up arrow). This captured all local flexions in MO. The speed of local trough-to-peak increase (i.e., the steepness of the upward arrows in Fig. 2(d)) is notated as $\mathbf{v\_MO_{der}[+]}$; The speed of local peak-to-trough decrease (i.e., the steepness of the downward arrows in Fig. 2(d)) is notated as $\mathbf{v\_MO_{der}[–]}$.

MO speeds calculated using the first method focus on macroscopic MO movements within the frame of syllables, while MO speeds calculated using the second method focus on a finer scale of MO movements.

2.4. Variables calculated from the speeds of change in intensity and mouth movements

(a) Variables based on **v_I[+]** *and* **v_I[–]** Each sentence contains a number of **v_I[+]**s and **v_I[–]**s. Their central tendencies, dispersions and sequential variabilities were calculated in terms of the means, standard deviations, and pairwise variability indices[6] ad modum He and Dellwo [7]: mean_vI[+], stdev_vI[+] and pvi_vI[+], as well as mean_vI[–], stdev_vI[–] and pvi_vI[–].

(b) Variables based on $\mathbf{v\_MO_{syl}[+]}$ *and* $\mathbf{v\_MO_{syl}[–]}$ Similarly, each sentence contains a number of $\mathbf{v_MO_{syl}[+]}$s and $\mathbf{v_MO_{syl}[–]}$s. Their means, standard deviations, and pairwise variability indices were calculated: mean_vMO_{syl}[+], stdev_vMO_{syl}[+] and pvi_vMO_{syl}[+], as well as mean_vMO_{syl}[–], stdev_vMO_{syl}[–] and pvi_vMO_{syl}[–].

(c) Variables based on $\mathbf{v\_MO_{der}[+]}$ *and* $\mathbf{v\_MO_{der}[–]}$ In the same manner, the means, standard deviations and pairwise variability indices were calculated based on $\mathbf{v\_MO_{der}[+]}$ and $\mathbf{v\_MO_{der}[–]}$ for each sentence: mean_vMO_{der}[+], stdev_vMO_{der}[+] and pvi_vMO_{der}[+], as well as

mean_vMO$_{der}$[–], stdev_vMO$_{der}$[–] and pvi_vMO$_{der}$[–].

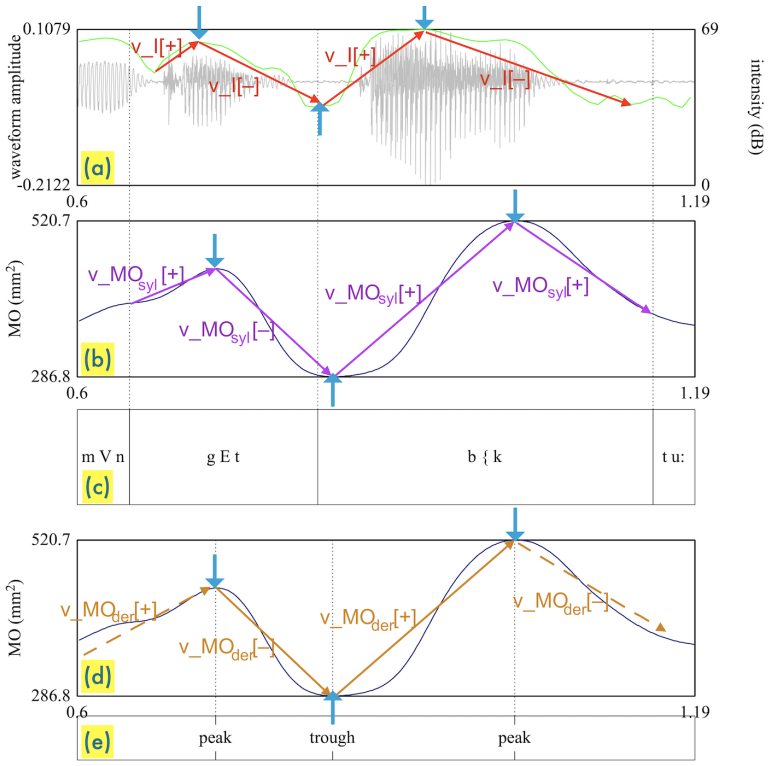


Figure 2: (a) An illustration of speeds of intensity increases and decreases; (b) an illustration of speeds of MO increases and decreases using the syllable-based method; (c) syllable boundaries with the SAMPA transcription ("get back"); (d) an illustration of speeds of MO increases and decreases using the derivative-based method; (e) MO peaks and troughs pinpointed from derMO.

2.5. Statistical analysis

JMP was used for statistical analysis. To quantify how between-speaker variability was differentially explained by the positive and negative variables in §2.4, the multinomial logistic regressions (MLRs) were fitted with *speaker* as the nominal response variable and the variables in §2.4 as the numeric predictor variables. *Sentence* effect was mitigated by z-score normalization. To bypass the collinearity issues, nine MLR models were separately fitted (M1 – M9 below).

For variables based on speeds of intensity increases and decreases: (M1) *speaker* ~ mean_vI[+] + mean_vI[–], (M2) *speaker* ~ stdev_vI[+] + stdev_vI[–], (M3) *speaker* ~ pvi_vI[+] + pvi_vI[–].

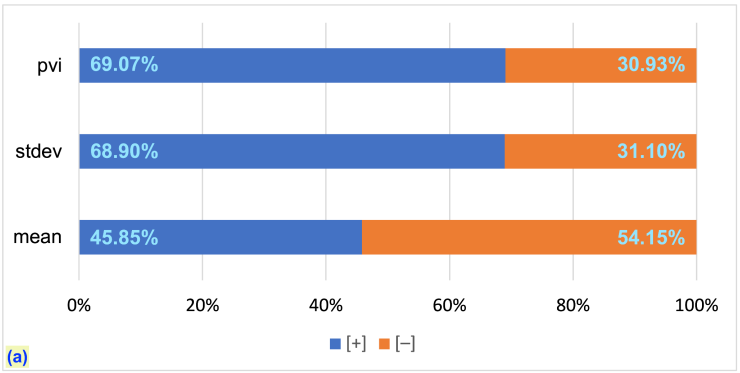
For variables based on speeds of MO increases and decreases calculated across syllable boundaries: (M4) *speaker* ~ mean_vMO$_{syl}$[+] + mean_vMO$_{syl}$[–], (M5) *speaker* ~ stdev_vMO$_{syl}$[+] + stdev_vMO$_{syl}$[–], (M6) *speaker* ~ pvi_vMO$_{syl}$[+] + pvi_vMO$_{syl}$[–].

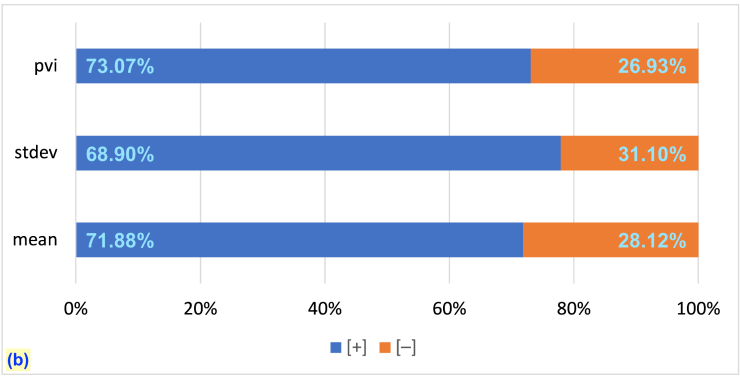
For variables based on speeds of MO increases and decreases calculated based on derivatives: (M7) *speaker* ~ mean_vMO$_{der}$[+] + mean_vMO$_{der}$[–], (M8) *speaker* ~ stdev_vMO$_{der}$[+] + stdev_vMO$_{der}$[–], (M9) *speaker* ~ pvi_vMO$_{der}$[+] + pvi_vMO$_{der}$[–].

For each model, the amount of between-speaker variability was calculated as $(\chi^2/\Sigma\chi^2)\times100\%$, where χ^2 refers to the likelihood ratio χ^2 of a particular variable, and $\Sigma\chi^2$ refers to the sum of likelihood ratio χ^2s of both variables in a model.

3. RESULTS

Due to the page constraint, the model fitting details and statistical results of the nine models are not tabulated in this paper; they are documented as a supplementary material accessible via https://osf.io/n7jcu. For each model, the variable based on the positive speed explained more between-speaker variability than its negative counterpart (Fig. 3 indicates the percentage of between-speaker variability explained by each variable per model).





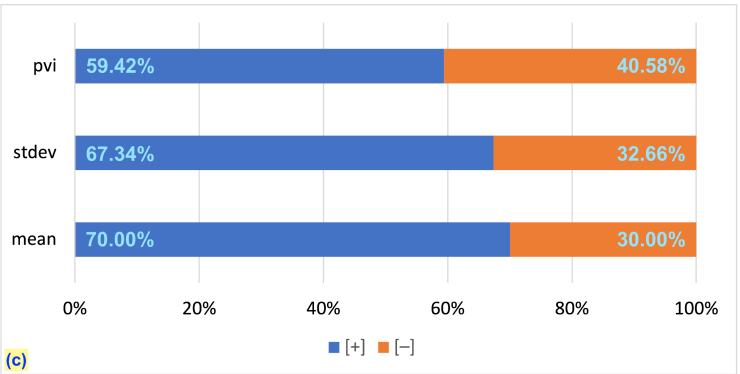


Figure 3: Between-speaker variability calculated from nine MLR models. (a) variables based on vI[+] and vI[−]; (b) variables based on vMO_{syl}[+] and vMO_{syl}[−]; (c) variables based on vMO_{der}[+] and vMO_{der}[−].

4. DISCUSSION

For native English speakers, measures based on speeds of intensity increases explained more between-speaker variability in general (except for the mean measure, although the difference is small-scale). The speeds of mouth opening and closing also differentially explain between-speaker variability in a similar manner: the measures based on the speeds of MO increases are more powerful in explaining speaker variability than the ones based on the speeds of MO decreases.

Congruency in the results obtained from both acoustic and articulatory data supports the attempt to investigate speaker idiosyncrasy in mouth opening-closing articulation via variabilities in intensity increases and decreases [7, 8]. This suggests that speaker idiosyncratic features can be shared across interconnected systems. In linguistics and speech communications, features shared between domains (visuo-articulatory, acoustic, neurological and gestual) are helpful, and sometimes crucial for comprehension [20, 21, 22].

In terms of the disparities in the speeds of increases and decreases for intensity and mouth opening-closing movements in encoding speaker-specific information, it is likely that during speech articulation, English speakers exhibited more freedom in the mouth opening phases, while in the mouth closing phases they coordinate their articulatory movements in a more controlled way. In a study on the reproduction of articulatory trajectories using a model-based method, it has already been demonstrated that the motor programs may be different in the opening and closing gestures [23].

It is also interesting to see that the English speakers exhibited the opposite pattern compared to Zurich German and Thai speakers [7, 8]. It is likely that the role of mouth opening-closing cycles is not universal across languages, and the way speaker-specificity is encoded in this dynamic process and its acoustic outcomes are also different in different languages. Using spontaneous speech, a number of studies have shown that the distribution of speaker-specific information in speeds related to the increases and decreases of intensity and F1 are more or less balanced in Dutch, English and German [24, 25, 26]. The differences might be attributed to the use of spontaneous speech, which is believed to be less controlled and thus displaying larger variations in articulatory patterns in general [27, 28].

Although different studies involving different languages and speaking styles exhibited different results in terms of the amount of speaker specificity explained by measures that are attributed by the speeds of mouth opening and closing movements, *speaker* effect remained significant regardless of the speed directions (see the model fitting results in the supplementary material https://osf.io/n7jcu) and the languages involved [5, 7, 8]. These findings have implications for forensic phonetics, ASR and other research fields that require the characterization of individual speakers. In forensic voice comparison closer attention should be paid to parts of the speech signals which are more informative about speaker identity in a particular language. In terms of ASR, it may be possible to achieve higher system performance if different parts of the acoustic signal are modeled differently. Of course more work needs to be done to gain better insights in our general understanding of the acoustic-articulatory dynamics.

5. ACKNOWLEDGEMENTS

We acknowledge support from the University of Zurich (Forschungskredit, Grant Nos. FK-19-069 (Y.Z.) and FK-20-078 (L.H.)), and the Swiss National Science Foundation (Grant No. PZ00P1_193328 (L.H.)).

6. REFERENCES

[1] Dellwo, V., French, P., He, L. 2018. Voice biometrics for speaker recognition applications. In: Frühholz, S., Belin, P. (eds), *The Oxford Handbook of Voice Perception*. Oxford University Press, 777–795.

[2] Leemann, A., Kolly, M.-J., Dellwo, V. 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Sci. Int'l* 238, 59–67.

[3] Dellwo, V., Leemann, A., Kolly, M.-J. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *J. Acoust. Soc. Am.* 137, 1513–1528.

[4] McDougall, K. 2006. Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *Int'l J. Speech, Lang. Law* 13, 89–126.

[5] He, L., Zhang, Y., Dellwo, V. 2019. Between-speaker variability and temporal organization of the first formant. *J. Acoust. Soc. Am.* 145, EL209–EL214.

[6] He, L., Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. *Int'l J. Speech, Lang. Law* 23, 243–273. .

[7] He, L., Dellwo, V. 2017. Between-speaker variability in temporal organizations of intensity contours. *J. Acoust. Soc. Am.* 141, EL488–EL494.
[8] Zhang, Y., He, L., Kerdpol, K., Dellwo, V. 2021. Between-speaker variability in intensity slopes: The case of Thai. Abstract presented at the XVIIth AISV Conference (Zürich, 4–5 February 2021).
[9] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A. A. 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, e1000436.
[10] Ji, A., Berry, J. J., Johnson, M. T. 2014. The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In: *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing* (Florence, 4–9 May, 2014), 7769–7773.
[11] Boersma, P., Weenink, D. 2020. Praat: Doing phonetics by computer (version 6.1.16) www.praat.org
[12] Schiel, F. 1999. Automatic phonetic transcription of non-prompted Speech. In *Proc. of the 14th ICPhS* (San Francisco, 1–7 August 1999), 607–610.
[13] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347.
[14] Reichel, U. D., Kisler, T. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In: Hoffmann, R. (ed), *Elektronische Sprachverarbeitung* (Studientexte zur Sprachkommunikation 71). TUDpress, 42–49.
[15] He, L. 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am.* 152, 567–579.
[16] Hass, J., Heil, C., Weir, M. 2019. *Thomas' Calculus* 14th edition, (SI units, international edition) Pearson Education.
[17] Low, E. L., Grabe, E., Nolan, F. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech* 43, 377–401.
[18] MacNeilage, P. F. 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499--546.
[19] Morrill, R. J., Paukner, A., Ferrari, P. F., Ghazanfar, A. A. 2012. Monkey lipsmacking develops like the human speech rhythm, *Develop. Sci.* 15, 557--568
[20] Strauss, A., and Schwartz, J.-L. 2017. The syllable in the light of motor skills and neural oscillations, *Lang. Cogn. Neurosci*. 32, 562--569.
[21] Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J, Kello, C., Schaefer, R.S., Wiggins, G.A. 2021. Multilevel rhythms in multimodal communication, *Phil. Trans. R. Soc*.B376: 20200334.
[22] Park, H., Kayser, C., Thu, G., Gross, J. 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility, *eLife* 5, e1452.
[23] Birkholz P., Kroger, B. J., and Neuschaefer-Rube, C. 2011. Model-based reproduction of articulatory trajectories for consonant-vowel sequences, *IEEE Trans. Audio, Speech Lang. Process*. 19, 1422--1433.
[24] Lins Machado, C. 2021. A cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech. In *Proc. of the* XVII AISV Conference (Zurich, 4-5 February 2021), 79-80
[25] He, L., Heeren, W. 2021. Between-speaker variability in dynamic formant characteristics in spontaneous speech. In Proc. of the XVII AISV Conference (Zurich, 4-5 February 2021), 71-72.
[26] Heeren, W., He, L. 202., Between-speaker variability in segmental F1 dynamics in spontaneous speech. In: Proc. of the 30th Annual Conference of the International Association for Forensic Phonetics and Acoustics (Marburg, 10-13 July 2021).
[27] De Nil, L.F., Abbs, J.H. 1991. Influence of speaking rate on the upper lip, lower lip and jaw peak velocity sequencing during bilabial closing movements. *J. Acoust. Soc. Am.* 89(2), 845-849.
[28] Illa, A., Ghosh, P.K. 2020. The impact of speaking rate on acoustic-to-articulatory inversion. In *Computer Speech & Language*, 59, 75-90.

---

[1] He and Dellwo [7] referred to the speeds of intensity increases and decreases as "positive and negative dynamics" of intensity. In the context of [7], "dynamic(s)" was used in its most general sense to mean "non-static" properties of the intensity curve. Here, we refrain from using the word "dynamics" to avoid confusion with its more technical sense in mechanics, where "dynamics" entails elucidating the forces underpinning movements. We did not characterize these forces.

[2] The EMA-MAE corpus is based on work supported by the National Science Foundation of the United States under Grant #IIS-1142826 to Marquette University, which support does not constitute an endorsement.

[3] The underlying signal processing details are described in He and Dellwo [7].

[4] The formula for calculating the triangle area based on 3D coordinates is $\frac{1}{2}\sqrt{\begin{vmatrix} y_{UL} & z_{UL} & 1 \\ y_{LL} & z_{LL} & 1 \\ y_{LC} & z_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} z_{UL} & x_{UL} & 1 \\ z_{LL} & x_{LL} & 1 \\ z_{LC} & x_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} x_{UL} & y_{UL} & 1 \\ x_{LL} & y_{LL} & 1 \\ x_{LC} & y_{LC} & 1 \end{vmatrix}^2}$. The MATLAB script (mouth_opening_area.mlx) for the computation is available via https://osf.io/qjh6p [15].

[5] URL: clarin.phonetik.uni-muenchen.de/BASWebServices

[6] For a series of numbers (x_1, x_2, x_3, …, x_n), the pairwise variability index (pvi) was calculated as $(|x_2 - x_1| + |x_3 - x_2| + \cdots + |x_n - x_{n-1}|) \div (n - 1)$. The pvi was initially developed to measure speech rhythm based on the durations of vocalic or intervocalic intervals [17].

The Acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech

Grace Wenling Cao, Peggy Mok

The Chinese University of Hong Kong
gracecao@cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

This paper explored filled pauses *uh*, *um* and *m* in Cantonese-Mandarin-English trilingual speech by 20 Hong Kong trilingual speakers. Duration, frequency and vowels of 1493 filled pauses were examined. Auditory analysis suggested that [ɛ, a, ə] vowels were commonly used for *uh* and *um* in the three languages. A preference for *uh* was found for all three languages, but the distributions of *uh*, *um* and *m* were more similar in the two Chinese languages than in English. For the duration, *uh* was similar across the three languages. For formants of *uh*, the main difference between Cantonese and the other two languages was found on F2, partially supporting the language-specific view in filled pause research. The speakers had similar vowel qualities for *uh* in English and Mandarin. Their parents' L1, primary school background, age and length of learning English and Mandarin significantly predicted their use of filled pauses.

Keywords: forensic phonetics, filled pause, cross-language speaker comparison

1. INTRODUCTION

Previous studies suggested that filled pauses such as *uh* and *um* can be useful for speaker identification in a monolingual context because they appear to be language-specific [1] and consistent for individual speakers [2]. However, whether filled pauses are language-specific in the context of multilingual speech remains unclear. This paper aims to tackle this issue and explores filled pauses in Cantonese-Mandarin-English trilingual speech. In this case, speakers' L2 (e.g. English) is typologically different from L1 Cantonese, whereas L3 (e.g. Mandarin) is typologically similar to their L1.

1.1. Filled pauses in multilingual speech

It has been a debate about whether filled pauses are language-dependent in the context of bilingualism. On the one hand, Wong and Papp [3] argued that filled pauses in English-Maori bilingual speech had a high transferability across languages, suggesting an L1-transfer view. Gosy, Gyarmathy and Beke [4] shared a similar stance, showing that Hungarian-English bilinguals in their study had no perceptual difference in the vowel production of English and Hungarian filled pauses. On the other hand, a few recent studies suggested a language-specific view [5]–[7]. For instance, de Boer and Heeren [5] found consistency in duration and F0 for Dutch-English bilingual speakers, but vowel realization showed some adaptation towards L2. Lo [6] also found that German-French simultaneous bilingual speakers did not use filled pauses with the same distributions and acoustic profiles in the two languages. Spreafico [7] explored filled pauses produced by an Italian-German-English trilingual speaker, and his results supported the language-specific view, where F2 was significantly different across the three languages. Studies which support the language-specific view tend to use bilingual speakers with high proficiency, whereas studies supporting the L1-transfer view had bilingual speakers with various L2 proficiency. None of these studies, however, examined the role of bilingual speakers' linguistic background in their use of filled pauses.

1.2. Present study

The present study explored filled pauses in Cantonese-English-Mandarin trilingual speech. Studies on filled pauses in Chinese are very limited. Among the handful of studies [8]–[10], not all of them conducted an acoustic-phonetic analysis. Common filled pauses in Mandarin include *uh*, *um* and *mm* [8]-[9]. Wu [10] examined L2 Chinese and suggested *uh* with two different representations (呃, 哦) and *um* (嗯). Zhao and Jurafsky [9] suggested that *uh* was more common than *mm* and speakers in the south used more filled pauses than speakers in the north. The lack of phonetic research on filled pauses in Chinese languages and trilingual speech calls for more in-depth investigations. To fill the gap, this study asked two research questions:

- Is vowel quality of filled pauses language-dependent across the three languages?
- Do trilingual speakers' linguistic background influence their use of filled pauses?

Our hypotheses were first, the speakers would use language-specific vowels in three languages; and

second, factors of speakers' linguistic background such as parents' L1, age and length of learning English and Mandarin would influence how they use filled pauses cross-linguistically.

2. METHOD

2.1 Participants

Twenty (8F) Cantonese-English-Mandarin trilingual speakers participated in three mock police interviews. They were all students (age range: 18-27 years old) from a university in Hong Kong and spoke Cantonese as their first language. Seven participants reported that at least one of their parents does not speak Cantonese as their L1, among which six speak Mandarin and one speaks Teochew. Nine participants reported that they attended primary or secondary schools which used Mandarin as the medium of instruction for the Chinese subject. Their English proficiency was IELTS 6-8, the age of learning English was 1-6 years old, and the length of English learning was 15-21 years. Regarding their Mandarin proficiency, none of the participants took the National Putonghua Proficiency Test. Therefore, their proficiencies were estimated by a separate rating task. 100 native Mandarin speakers listened to a 10-second recording of each participant and judged the standardness of the speaker's Mandarin accent. Results suggested that the mean was 5.42 (out of 10), ranging from 2.81 to 8.13. Their age of learning Mandarin was 1-10 years old, and the length of learning Mandarin was 9-27 years. Participants also reported their daily use of the three languages. On average, they used 88% Cantonese, 3% Mandarin and 9% English.

2.2 Experiment

Participants attended three mock police interviews as part of a larger forensic phonetic project. Participants were told that they were involved in an international investment fraud case. A police officer from Hong Kong, Shanghai and London would interrogate them separately in three interviews. The three police officers were role-played by the first author, who is a Cantonese-Mandarin balanced bilingual and speaks English with a near-native British accent. The participants had an interview in Cantonese first, followed by Mandarin and English interviews. To help participants to adjust their language mode for the coming interviews, they had a 5-minute break after each interview and watched a 3-minute video in Mandarin or English before they started the next task.

Interviews were conducted in a sound-proof recording booth. Participants and the 'police officer' sat face-to-face by a desk. A desktop computer was placed facing the participants to show information about the case. Participants were told that they could look at the screen if they could not remember the case details. An H4N recorder was placed about 30 cm away from the participant and was used to record the interviews. Each interview was about 6 to 8 minutes long. Participants also completed a questionnaire about their linguistic backgrounds.

2.3 Data annotation and analysis

First, filled pauses in the three interviews were coded by two research assistants and two student helpers using Praat [11]. The Cantonese interviews were coded by the research assistant who is a native speaker of Cantonese, whereas the Mandarin and English interviews were marked by the research assistant/student helpers who are Mandarin-English bilinguals. Although code-switching is common for trilingual speakers in Hong Kong, only in the Cantonese interviews intra-sentential code-switching was found, which should not affect the coding for Cantonese interviews. In total, 1493 filled pauses including 413 Cantonese tokens, 471 Mandarin tokens and 609 English tokens were coded. To measure the frequency of filled pauses, participants' interviews were transcribed orthographically and the word/character counts were calculated.

As there is limited research on filled pauses in Chinese languages and trilingual speech, both auditory and acoustic analyses were conducted. The auditory analysis aimed to explore the vowels of filled pauses used in the three languages. First, the first author listened to every single token of the filled pauses in the interviews and annotated the vowels. Another trained forensic phonetician who is also a Cantonese-Mandarin-English trilingual crosschecked all the vowel annotations in three interviews. For any disputed annotations, the first author and the phonetician would discuss to reach an agreement. For the acoustic analysis, a Praat script was used to extract F1, F2 and F3 values at the midpoint of the vowels and calculate the duration of the filled pauses. Data were imported to R along with participants' background information for further analysis.

3. RESULTS

3.1 Counts and frequency

The frequency of *uh*, *um* and *m* over the total word/character counts are shown in Table 1. Note that for Cantonese and Mandarin interviews the number of Chinese characters was calculated, whereas for English the number of words was counted. The comparisons between the Chinese languages and English shall be interpreted cautiously.

	Count	*uh*	*um*	*m*	all fps
Cantonese	15048	2.53%	0.07%	0.14%	2.74%
Mandarin	9804	4.09%	0.42%	0.30%	4.80%
English	17524	2.21%	0.91%	0.35%	3.48%

Table 1. Percentage of filled pauses in the interviews.

The distributions of *uh*, *um* and *m* over all filled pauses are shown in Figure 1. The most frequently used filled pause in all three languages was *uh*. The distributions of filled pauses in Mandarin and Cantonese were similar, whereas the distribution in English suggested a more frequent use of *um* and *m*.

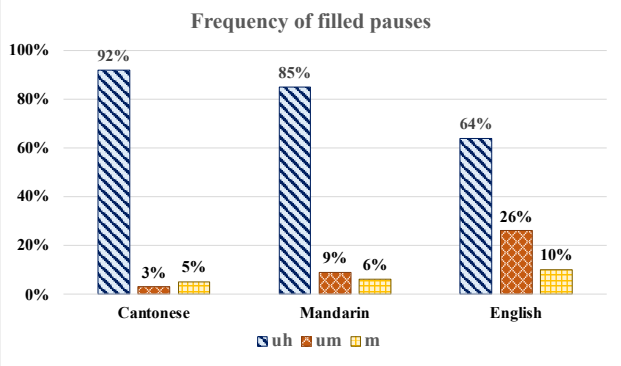


Figure 1. Distribution of *uh*, *um* and *m* in Cantonese, Mandarin and English.

Following Lo [6], the dataset was re-coded for mixed-effects logistic regression models. Tokens of *m* were removed and the use of *uh* and *um* was coded as a binary variable. The glmer() function in lmer package [12] was used to run the models. The basic model included language and sex as fixed effects, and included random intercept by speaker and random slopes by speaker over language. Factors of speakers' linguistic backgrounds such as whether the L1 of participant's parents is Cantonese or not (Parents' L1), age and length of learning Mandarin and English, attending primary schools in which the medium of instruction is Mandarin for the Chinese subject (MAN as MOI), proficiency of English, the experience of learning other languages and the standardness of Mandarin were examined one by one using a forward stepwise approach. Models were then compared with the basic model using the anova function in R. Only factors that significantly improved the model fit would add to the model.

The best model contained the fixed factors shown in Table 2. Results suggest that the distribution of *um* over *uh* was significantly different between Cantonese and English, but not between Cantonese and Mandarin (um/uh ratio for Cantonese: 0.028; for English: 0.409; for Mandarin: 0.102). For participants whose parents' L1 is *not* Cantonese, their *um:uh* ratio (0.29) was higher than those whose parents' L1 is Cantonese (0.11). For participants who had Mandarin as the Medium of Instruction in their primary schools, their *um:uh* ratio (0.28) was higher than those who did not (0.15). Factors which were not included in Table 2 did not improve model fit.

	Est	SE	z	*p*
(Intercept)	-5.3	1.7	-3.1	.002
Language (ENG)	1.9	0.9	2.2	**.03***
Language (MAN)	0.5	1.0	0.5	.60
Parent's L1 is Cantonese (yes)	-2.9	1.3	-2.3	**.02***
MAN as MOI (yes)	2.4	0.8	2.9	**.003***
Sex (male)	-1.3	0.7	-1.7	.09
Standardness of MAN	0.6	0.4	1.4	0.17
Language (ENG) × Parents' L1 is Cantonese (yes)	1.4	1.3	1.1	0.26
Language (MAN) × Parent's L1 is Cantonese (yes)	-1.5	1.7	-0.9	0.38

Table 2. Summary of fixed effects in the mixed-effects logistic regression model for the distribution of *uh* over *um*.

3.2 Acoustic and auditory analysis of vowels in *uh*

Due to the limited space, only acoustic and auditory analysis of vowels in *uh* will be reported here. For F1, F2 and F3, data of males and females were separated into two subsets and linear mixed-effects models were run separately. The basic model contained language as a fixed effect and included random intercept by speaker and random slopes by speaker over language. Factors of participants' linguistic backgrounds were added to the model one by one to test its significance. Results are shown in Table 3.

The most significant cross-linguistic difference was found on F2, where *uh* was articulated with much fronter vowels in Cantonese than in Mandarin and English for both females (about 400 Hz) and males (about 250 Hz). F1 was relatively similar across the three languages, except that female speakers produced vowels with a significantly lower F1 in their Cantonese interviews compared to English ones. For F3, females and males showed different cross-linguistic patterns. Female speakers had no significant change in their F3 across the three languages, whereas male speakers produced vowels with a significantly higher F3 in Mandarin and English than in Cantonese. The durations (in ms) of *uh* were similar across the three languages for female speakers (CAN: 374ms, MAN: 350ms, ENG: 370ms). No significant comparisons were found for male speakers too: English (360ms), Cantonese (337ms) and MAN (311ms).

Auditory analysis suggested that the participants used [ɛ, a, ɐ, ə, ʌ] vowels for most of the *uh* in the interviews. The filled pause *uh*-[ɛ] is commonly found in colloquial Cantonese, the quality of this vowel is similar to the Cantonese /ɛ/ as in 賒 'lend'. The *uh*-[ɛ] is not commonly found in colloquial Mandarin and English. The *uh*-[a] is close

to the Cantonese /ɐ/ as in 濕 'wet'. For simplification, two categories *uh*-[ɛ] and *uh*-[a] were distinguished based on auditory analysis. The *uh*-[a] contained representations vary from [ə] to [ʌ].

F1-females	**Est**	**SE**	**t**	***p***
(Intercept)	546	26	21	**<.001***
Language (MAN)	42	22	2	.109
Language (ENG)	61	21	3	**.019***
MAN as MOI (yes)	-113	51	-2	**.033***
Age of learning MAN	41	12	3	**.002***
F1-males	**Est**	**SE**	**t**	***p***
(Intercept)	318	82	4	**.003***
Language (MAN)	10	23	0.4	.672
Language (ENG)	-33	26	-1.3	.227
Length of learning ENG	15	4.5	3	**.009***
F2-females	**Est**	**SE**	**t**	***p***
(Intercept)	1775	78	23	**<.001***
Language (MAN)	-446	66	-7	**<.001***
Language (ENG)	-408	57	-7	**<.001***
Other languages (yes)	129	47	3	**.012***
F2-males	**Est**	**SE**	**t**	***p***
(Intercept)	1735	68	26	**<.001***
Language (MAN)	-253	53	-5	**<.001***
Language (ENG)	-240	53	-5	**<.001***
MAN as MOI (yes)	-242	102	-2	**.035***
F3-females	**Est**	**SE**	**t**	***p***
(Intercept)	3256	154	21	**<.001***
Language (MAN)	-23	81	-0.3	.787
Language (ENG)	84	64	1.3	.223
Length of learning MAN	-30	8	-4	**.003***
Other languages (yes)	310	40	8	**<.001***
F3-males	**Est**	**SE**	**t**	***p***
(Intercept)	2418	52	46	**<.001***
Language (MAN)	105	43	2.5	**.028***
Language (ENG)	124	31	4	**.005***

Table 3. Summary of fixed effects in the mixed-effects regression models for F1 – F3 of *uh*.

For Cantonese interviews, the frequency of [ɛ] was (82%) much higher than [a] (18%). In contrast, [ɛ] was less common for MAN (19%) and ENG (15%). Figure 2 shows the vowel plots for all these vowels in three languages. The vowel [a] was similar across the three languages, whereas the CAN-[ɛ] vowel was distinguished from the MAN-[ɛ] and ENG-[ɛ] for both male and female participants.

4. DISCUSSION

This study investigated Cantonese-Mandarin-English trilingual speakers' filled pauses in three languages. Different from Lo [6], distributional difference was found between Cantonese and English, but not between Cantonese and Mandarin. This could be because Cantonese and Mandarin are typologically similar, and the speakers might directly transfer the distribution of Cantonese filled pauses to Mandarin. Both Cantonese and Mandarin belong to the Sino-Tibetan language family. Although the two languages differ in phonology and lexicons, they have the same SVO word order and they both are tonal languages. As factors of language proficiency were not significant for the models (Table 2), the difference is unlikely due to speakers' proficiency in English and Mandarin.

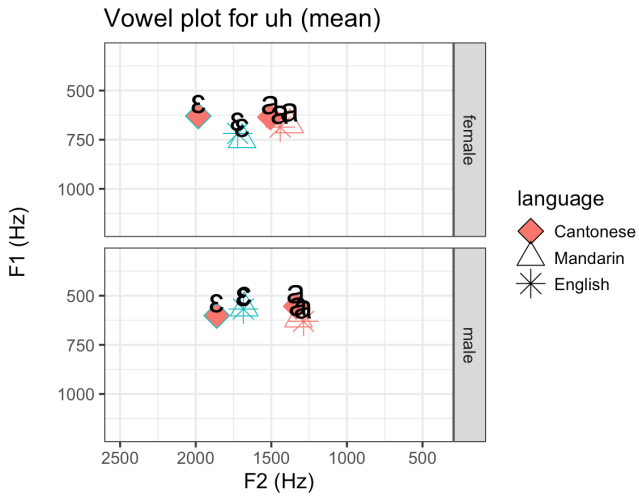


Figure 2. Vowel plot for *uh* by sex and language.

The cross-linguistic difference was mainly found in F2, which is in line with Lo [6] and Spreafico [7]. The significant frontness of Cantonese-*uh* is likely driven by the Cantonese-specific filled pause *uh*-[ɛ]. As shown in Figure 2, Cantonese *uh*-[ɛ] was much fronter than the other two languages, but *uh*-[a] was similar across the three languages. Although F2 was significantly different between Cantonese and the other two languages, no significant difference was found for F1, F2 and F3 between English and Mandarin (except for F1-male), suggesting that the speakers used similar vowels for *uh* in English and Mandarin. In other words, only part of the results supports the language-specific view.

Several factors of linguistic background had a significant impact on the speakers' use of filled pauses, including parents' L1, MAN as MOI, the length and age of learning Mandarin and English, and other languages. This reveals the necessity of taking these factors into consideration for future studies.

5. CONCLUSION

To conclude, this paper partially confirms the language-specific view with trilingual speech data: Cantonese-Mandarin-English trilingual speakers used a fronter CAN-[ɛ] than ENG-[ɛ] and MAN-[ɛ]. Speakers' linguistic backgrounds such as parents' L1 and age and length of learning L2 and L3 are found significant to predict their use of filled pauses in three languages.

6. ACKNOWLEDGEMENTS

This research is supported by Hong Kong RGC Postdoctoral Fellowship. We are grateful to So Ka Ki, Liu Xinyue, Wang Wenhui and Zhou Henan for their assistance with data coding and to Fung Hei for annotation crosschecking.

7. REFERENCES

[1] E. de Leeuw, "Hesitation Markers in English, German, and Dutch," *J. Ger. Linguist.*, vol. 19, no. 2, pp. 85–114, 2007.

[2] H. J. Künzel, "Some general phonetic and forensic aspects of speaking tempo," *Int. J. Speech Lang. Law*, vol. 4, no. 1, pp. 1350–1771, 1997.

[3] S. G. Wong and V. Papp, "Transferability of non-lexical hesitation markers across languages: Evidence from te reo Maori-English bilinguals," *Proc. 26th IAFPA*, pp. 35–66, 2018.

[4] M. Gósy, D. Gyarmathy, and A. Beke, "Phonetic analysis of filled pauses based on a Hungarian-English learner corpus," *Int. J. Learn. Corpus Res.*, vol. 3, no. 2, pp. 149–174, 2017.

[5] M. M. de Boer and W. F. L. Heeren, " Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English ," *J. Acoust. Soc. Am.*, vol. 148, no. 6, pp. 3612–3622, 2020.

[6] J. J. H. Lo, "Between Äh(m) and Euh(m): The Distribution and Realization of Filled Pauses in the Speech of German-French Simultaneous Bilinguals," *Lang. Speech*, vol. 63, no. 4, pp. 746–768, 2020.

[7] L. Spreafico, "Filled pauses in multilingual speech: an acoustic analysis," *Linguist. e Filol.*, vol. 36, no. 2016, pp. 99–116, 2016.

[8] X. Wang and X. Jin, "汉语二语学习者口语非流利填充型停顿研究," 東北師大學報, vol. 3, pp. 84–92, 2020.

[9] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses," *Proc. Disfluency Spontaneous Speech, DiSS 2005*, no. September, pp. 179–182, 2005.

[10] C. Wu, "Filled Pauses in L2 Chinese: a comparison of native and non-native speakers," *Proc. 20th North Am. Conf. Chinese Linguist.*, vol. 1, pp. 213–227, 2008.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

ASR-BASED DEVELOPMENT OF CHALLENGING SPEAKER DISCRIMINATION TESTS

Andrea Fröhlich[1,2,3], Volker Dellwo[2], Peter French[2], Meike Ramon[3]

[1]Zurich Forensic Science Institute, Switzerland
[2]Department of Computational Linguistics (UZH), University of Zurich, Zurich, Switzerland
[3]Applied Face Cognition Lab, University of Lausanne, Switzerland

ABSTRACT

We report our ongoing efforts towards the development of challenging speaker discrimination tests. Our goal is to establish highly sensitive tests that enable characterization of individual differences in voice processing skills. Such tools are required but lacking for the identification of so-called "voice super-recognizers" – individuals with superior voice processing ability – who could aid criminal investigations involving audio material. To find such individuals we manipulated test difficulty using an ASR and delta F0-based stimuli selection method. Independent groups of participants performed 2-alternative forced choice voice discrimination tests of stimulus pairs selected either randomly, or systematically based on inter-item similarity. On average, performance was significantly higher for the prior (80.0%) as compared to the latter (68.8%) test. Thus, we have established a method to manipulate task difficulty in speaker discrimination tests. We will further validate and extend our method to include a wider range of audio material implemented as a within-subject design.

Keywords: voice super-recognizers (VSR), speaker discrimination, forensic phonetics.

1. INTRODUCTION

In criminal proceedings with audio files as evidence, police and prosecution authorities can seek the assistance of forensic phoneticians. Recently, criminal cases with voice evidence have increased where forensic phoneticians are expected to analyse mounting volumes of data in ongoing police investigations. Scenarios such as potentially finding a voice in an extensive collection of case-relevant audio files are impossible to solve by applying traditional auditory-phonetic and acoustic methods within a reasonable time frame. Therefore, new techniques or approaches are required to deal with criminal investigations involving large numbers of audio files. One potential approach to this big data and time-pressure dilemma is the employment of automatic speaker recognition (ASR) systems. They can rapidly calculate similarity scores or (log) likelihood ratios between large numbers of files. However, the performance of these systems is highly affected by the quality and the duration of the audio recordings [1] and humans have also been shown to outperform them in certain scenarios [2]. Therefore, ASR systems are typically used only alongside traditional methods in forensic casework [3]. If, after using ASR systems, forensic phoneticians must still analytically review all the ASR-generated results by applying auditory and acoustic methods, big data and time pressure of ongoing investigations still pose a significant challenge. We are therefore seeking additional solutions.

Since the considerations above are not exclusive to auditory material, ongoing efforts from the domain of vision sciences may offer a solution. For the past decade, researchers have been investigating so-called "Super-Recognizers" (SR), individuals with untrained exceptional face identity processing abilities [4],[5]. As interest from international law enforcement agencies steadily increases [5], [6], so do reports of formal collaborations between researchers and police agencies who seek to actively deploy individuals due to an empirically documented unique ability [5].

Notwithstanding that SR are a very recent phenomenon in the visual domain, we have adopted an interdisciplinary approach to determine whether an analogous superior ability also exists for the processing of voice identity. Our current definition of "voice super-recognizers" (VSR) describes untrained individuals with *superior ability to efficiently process speaker identity consistently across input variations*. This could manifest as highly proficient speaker discrimination and/or recognition across objectively challenging conditions associated with increased task difficulty. Such conditions could include, e.g., low-quality audio material or high similarity of speakers' voices. Following this definition, we currently envision the employment of VSR in post-processing of ASR-generated results in police investigations or intelligence gathering with big data cases (Figure 1).



Figure 1: Potential employment scenario of VSR.

Given the nature of to-be-processed material, we recommend identifying VSR among police employees.

One approach to identify VSR proposed by Jenkins *et* al. [7] is to test if the abilities of individuals who identified as visual SR, also extend to voice processing.

Their participants completed the Bangor Voice Matching Test [8] (discrimination test), the Glasgow Voice Memory Test (recognition test) [9] and a bespoke Famous Voice Recognition Test [7]. The authors reported positive correlations between voice- and face processing and that some but not all individuals with superior face ability showed superior performance for voice processing.

We sought to expand on this work and further explore the concept of VSR by developing tests incorporating a larger quantity of speaker identities while systematically varying task difficulty using automatic approaches.

The current study proposes a method to control task difficulty by selecting voice stimuli based on within- and between-speaker similarity. We created a speaker discrimination test based on ASR-generated similarity scores, analogous to how we envision the employment of VSR to post-process ASR results. This provides the basis for creating a challenging test of speaker discrimination, which is the focus of the present work.

1.1 Previous work

To the best of our knowledge, the subject of *superior* voice processing ability, an umbrella term for various voice and speaker-related tasks, has been addressed in only a small number of studies so far [8]–[10]. Mühl *et al.* developed the aforementioned Bangor Voice Matching Test to assess voice discrimination abilities in a short and standardised screening test. While their research on speaker discrimination focused on assessing listeners' abilities, other groups focused on factors influencing task difficulty [11]. Alongside 'natural' variations, such as speaking style that render speaker discrimination tasks more challenging, acoustic properties of voices and their correlation to listeners' judgments in speaker discrimination tests have been investigated [12]. While speaker discrimination tests force listeners to make binary decisions, similarity ratings give more granular insights into the perceived similarity scale of voices [13]. As to the correlation between automatic and human similarity ratings, so far only phonetic features were investigated [14]. As far as we know, there is currently no research indicating that MFCC-based ASR systems correlate with human speaker discrimination performance. Still, we suspect that stimulus pairs rated similarly by ASR systems will be harder to discriminate by humans.

In the domain of human assisted speaker recognition (HASR), it was found that fusing machine-generated and human results showed improvements [2], [15], [16]. These findings provided inspiration for how VSR could be employed in investigative case work for post-processing ASR system-generated results. Contrary to some of the analysed schemes, we envision the employment of lay people only in the investigatory stage of cases.

2. METHOD

Different factors influence the difficulty of a speaker discrimination task. For this test, we only investigated and tested "voice-inherent factors", such as two voices sharing similar frequency properties [17], by using ASR-derived similarity scores and delta F0 values to find challenging stimulus pairs with low within-speaker, and high between-speaker similarity. We have refrained from making the experiment more difficult by addressing "technical factors" (such as recording devices) other than re-sampling the stimuli down to 8 kHz. For future experiments, we also plan to implement "acoustic environment factors", such as background noise, reverberation etc. We only tested short stimuli of around 1.2 seconds as Bricker and Pruzansky have shown that participants' discrimination accuracy increased with longer speech sample duration [18].

2.1 Corpus and participants

We used stimuli from the TEVOID corpus (Temporal Voice Idiosyncrasy, [19]), initially created to investigate speakers' temporal features in short sentences. We further processed the studio quality recordings that consist of read sentences by native Zurich German speakers (male and female) as described below. The study was run with participants from different classes of police cadets in Zurich, Switzerland.

2.2 Stimulus selection

As in a possible scenario where VSR would assist in a real police case, we combined an ASR system to pre-process the data for the speaker experiment. The goal of using an ASR system was to have a fast method to select similar stimuli from large corpora.

Yet, most ASR systems show high equal error rates with short speech sequences since there is insufficient information to build an adequate speaker model [2], [20]. To mitigate this challenge, we worked with a state-of-the-art speaker comparison model for short stimuli, which won the "Short-duration Speaker Verification Challenge 2020" [21],[22]. This ECAPA-TDNN algorithm (release 05-03-21) is open-source from Speechbrain [23] (version 0.5.10) and was included in our stimuli processing pipeline implemented in Python (version 3.8.12).

To select the stimulus pairs for the discrimination experiment, we first loaded all stimuli from the corpus using the `AudioNormalizer` and created embeddings using the `EncoderClassifier` classes of Speechbrain [23]. We then calculated pairwise cosine distances between the embedding vectors to obtain a full similarity matrix for all stimuli. We use the pairs with the lowest within-speaker and the highest between-speaker similarity scores. No likelihood ratios were calculated, the system was solely used on a score basis.

It is noteworthy that in the MFCC-feature extraction process, F0 information is neglected [21],[22]. However, it has been found that mean pitch influences listeners' judgements of voice dissimilarity [17]. We therefore measured and averaged the fundamental frequency (F0) of all audio samples for each speaker and calculated "delta F0" values for all speaker pairs. After investigating the F0 distribution of male and female voice identities in TEVOID, we only considered stimulus pairs for speakers with a delta F0 of 10 Hz or less for female speakers and 15 Hz for male speakers.

We then sorted all stimulus pairs by ascending cosine similarity and started picking pairs from the top of this list. A maximum number of four stimuli per speaker were included for different-speaker and same-speaker scenarios each. Two speakers only appeared in a pair with each other up to two times. The two stimuli of a pair never originated from the same sentence, and each selected audio stimulus only appeared once in the whole experiment. We performed all the above processing steps separately for male and female speakers and used ten speakers per gender. Overall, the experiment included 20 same-speaker (SS) and 20 different-speaker (DS) trials per gender.

2.3 Stimulus processing

To harmonise the audio file durations, which varied inter-individually with speakers' speech tempo, we normalised the selected stimuli by cutting the files to a duration of around 1.2 seconds, chosen from the centre of the audio file. We normalised the amplitude to 65 dB (RMS), but not the number of syllables contained in one snippet. For post-processing the selected files, we applied an amplitude smoothing function with the software Praat [24] to make the starting and ending of the stimuli less abrupt. In addition, we removed non-speech sections before and after the utterances.

Furthermore, we down-sampled the resulting stimuli to 8 kHz to make them more similar to a frequently observed setting in forensic casework (i.e. telephone transmitted speech). Finally, we concatenated each stimulus pair in random order with a pause of one second between the stimuli and saved them as MP3 files.

To validate our ASR and delta F0-based stimuli selection method, we implemented a second discrimination experiment where participants were presented with randomly selected stimulus pairs of the same audio quality, length, and out of the same TEVOID-corpus. We followed the same rules concerning the number of files per speaker and the spoken sentences. This random discrimination test was assessed with two unique groups of police cadets.

2.4 Experimental design

We used the Gorilla Experiment Builder [25] to create and host the speaker discrimination experiment. Pairs of audio stimuli (number of pairs = 80) were presented sequentially to the participants in random order, who had to indicate whether both were uttered by the same speaker or by two different speakers. We ran the two experiments on different classes of police cadets on different days. Participants completed the experiment onsite. They all used the same headphones.

3. RESULTS

We initially assessed the discrimination performance of the ECAPA-TDNN model on the original stimuli of the complete TEVOID corpus. The model showed a ROCCH EER [26] of only 0.92% for males and 2.13% for female stimuli. We compared this EER to the one of a x-vector-based system validated for forensic case work that showed a ROCCH EER of 16.1% for the males and a ROCCH EER of 16.5% for the female stimuli.

The results of the speaker discrimination tests are as follows: In the randomly combined stimulus pairs, participants, on average, correctly classified 80.0% of the pairs (38 participants with 80 stimulus pairs each), whereas in the challenging discrimination task, participant groups only classified 68.8% of the pairs correctly (52 participants with 80 stimulus pairs each). This observed difference is significant ($p<0.01$, two-sided t-test). The distributions of correct

answers per experiment are visualised in Figure 2. The challenging task was run with three groups of police cadets, of which the first scored 68.1%, the second 69.1% and the third 70.3% of the pairs correctly.

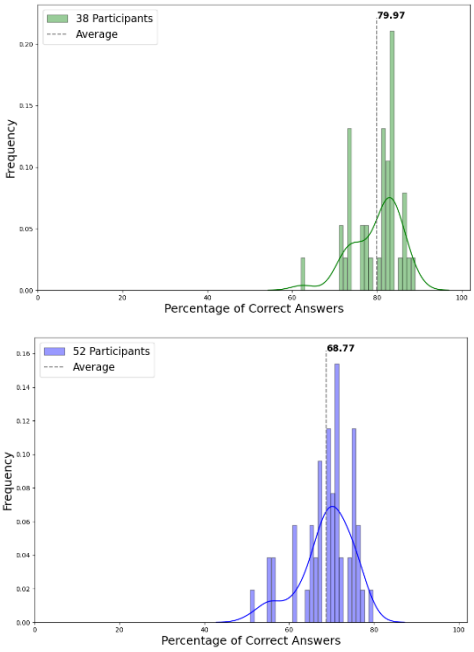


Figure 2: Test result of random (top) and score-based (bottom) stimuli selection method.

4. DISCUSSION

The presented study aimed to implement and test a new stimulus selection method to control voice-inherent speaker similarity in a discrimination task. We did so by employing an ASR model combined with delta F0 values for objective stimulus selection. The used ECAPA-TDNN model was not yet tested on forensic data sets or in a forensic casework scenario as suggested by Morrison and Enzinger [27]. Yet, it performed best in short speech stimuli discrimination and showed very low EER on our stimuli. Also, we did not employ the model to perform any forensic speaker discrimination but only to select difficult stimulus pairs to be used in our test. In future experiments, we will calibrate the ASR-system and try to make the stimuli as similar as possible to a forensic scenario.

So far, we ran our experiment on three groups of police cadets that scored similarly regardless of group size or composition. To further validate our method, we implemented and tested a second experiment with randomly chosen stimulus pairs which was tested on two additional independent participant groups. When comparing the results between random stimulus selections and stimulus selections based on our new method, the test designed using our method turned out to be significantly more challenging. This finding supports our hypothesis that discrimination task difficulty can be controlled by selecting stimulus pairs based on ASR system scores and delta F0 values.

The current test did not yet show any distinct superior group of participants in the distribution of scores. However, one could argue that the participants at the high end of a distribution of voice discrimination and/or recognition abilities could possibly be VSR. VSR capabilities could manifest as highly proficient speaker discrimination and/or recognition across objectively challenging conditions associated with increased task difficulty. Whether the best performers of our test are in fact VSR, will have to be evaluated after they have completed additional challenging and forensically relevant tests. However, we consider a challenging discrimination test like ours a necessary precondition for identifying VSR if they do, in fact, exist.

VSR could pave the way for human-centred approaches in biometrics and forensics with implications for law enforcement. Despite currently non-existent research regarding VSR, their discovery and employment in actual cases receive interest in the community. Currently, we can envision VSR deployment for postprocessing of automatically generated results in investigatory stages of a case involving large amounts of audio files. Identification of VSR will ultimately provide the needed basis for further research, including the benefits of their employment for criminal investigation, and the neural basis of their abilities. Finally, fusing the abilities of VSR and state-of-the art ASR systems could further improve voice processing performance.

We plan to repeat our experiments with more participants and as a within-subject design. Alongside voice-inherent factors, technical and acoustic-environmental factors will further be included in future tests, where we plan to extend our method beyond speaker discrimination to further test scenarios, such as one-to-many speaker recognition, sorting, or clustering tasks. Finally, our upcoming tests will include forensically relevant scenarios and stimuli to potentially find VSR best equipped for the employment in investigatory case work.

7. REFERENCES

[1] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, 'Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)', Speech Commun., vol. 112, pp. 30–36, Sep. 2019, doi: 10.1016/j.specom.2019.06.005.

[2] S. J. Park, A. Afshan, J. Kreiman, G. Yeung, and A. Alwan, 'Target and Non-target Speaker Discrimination by Humans and Machines', in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 6326–6330. doi: 10.1109/ICASSP.2019.8683362.

[3] E. Gold and P. French, 'International Practices in Forensic Speaker Comparisons: Second Survey', Int. J. Speech Lang. Law, vol. 26, no. 1, pp. 1–20, Jun. 2019, doi: 10.1558/ijsll.38028.

[4] R. Russell, B. Duchaine, and K. Nakayama, 'Super-recognizers: People with extraordinary face recognition ability', Psychon. Bull. Rev., vol. 16, no. 2, pp. 252–257, Apr. 2009, doi: 10.3758/PBR.16.2.252.

[5] M. Ramon, 'Super-Recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work', Neuropsychologia, vol. 158, p. 107809, Jul. 2021, doi: 10.1016/j.neuropsychologia.2021.107809.

[6] M. Ramon and S. Rjosk, beSure? – Berlin Test for Super-Recognizer Identification: Part I: Development. Verlag für Polizeiwissenschaft, 2022.

[7] R. Jenkins et al., 'Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks', 2020, doi: 10.31234/osf.io/7xdp3.

[8] C. Mühl, O. Sheil, L. Jarutytė, and P. E. G. Bestelmeyer, 'The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability', Behav. Res. Methods, vol. 50, no. 6, pp. 2184–2192, Dec. 2018, doi: 10.3758/s13428-0.

[9] V. Aglieri, R. Watson, C. Pernet, M. Latinus, L. Garrido, and P. Belin, 'The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices', Behav. Res. Methods, vol. 49, no. 1, pp. 97–110, Feb. 2017, doi: 10.3758/s13428-015-0689-6.

[10] D. Humble, S. R. Schweinberger, A. Mayer, T. L. Jesgarzewsky, C. Dobel, and R. Zäske, 'The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices', Behav. Res. Methods, Jun. 2022, doi: 10.3758/s13428-022-01818-3.

[11] A. Afshan, J. Kreiman, and A. Alwan, 'Speaker discrimination performance for "easy" versus "hard" voices in style-matched and-mismatched speech', J. Acoust. Soc. Am., vol. 151, no. 2, pp. 1393–1403, 2022.

[12] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, 'The relationship between acoustic and perceived intraspeaker variability in voice quality', in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[13] K. McDougall, 'Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades', Int. J. Speech Lang. Law, vol. 20, no. 2, pp. 163–172, Dec. 2013, doi: 10.1558/ijsll.v20i2.163.

[14] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, and F. Nolan, 'Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features', Speech Commun., vol. 124, pp. 85–95, Nov. 2020, doi: 10.1016/j.specom.2020.08.003.

[15] R. G. Hautamäki, V. Hautamäki, P. Rajan, and T. Kinnunen, 'Merging human and automatic system decisions to improve speaker recognition performance', in Interspeech 2013, ISCA, Aug. 2013, pp. 2519–2523. doi: 10.21437/Interspeech.2013-422.

[16] R. Schwartz et al., 'USSS-MITLL 2010 human assisted speaker recognition', in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 5904–5907. doi: 10.1109/ICASSP.2011.5947705.

[17] T. K. Perrachione, K. T. Furbeck, and E. J. Thurston, 'Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices', J. Acoust. Soc. Am., vol. 146, no. 5, pp. 3384–3399, Nov. 2019, doi: 10.1121/1.5126697.

[18] P. D. Bricker and S. Pruzansky, 'Effects of stimulus content and duration on talker identification', J. Acoust. Soc. Am., vol. 40, no. 6, pp. 1441–1449, 1966.

[19] V. Dellwo, A. Leemann, and M.-J. Kolly, 'Speaker idiosyncratic rhythmic features in the speech signal', Interspeech Conference Proceedings, 2012.

[20] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, 'Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles', J. Acoust. Soc. Am., vol. 144, no. 1, pp. 375–386, 2018.

[21] J. Thienpondt, B. Desplanques, and K. Demuynck, 'The IDLab Short-duration Speaker Verification Challenge 2020 System Description', Tech. Rep., 2020.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, 'Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification', ArXiv Prepr. ArXiv200507143, 2020.

[23] M. Ravanelli et al., 'SpeechBrain: A general-purpose speech toolkit', ArXiv Prepr. ArXiv210604624, 2021.

[24] P. Boersma, 'Praat, a system for doing phonetics by computer', Glot Int, vol. 5, no. 9, pp. 341–345, 2001.

[25] A. Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder.', 2020.

[26] N. Brümmer and E. de Villiers, 'The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF'. arXiv, Apr. 10, 2013. Accessed: Dec. 31, 2022. [Online]. Available: http://arxiv.org/abs/1304.2865

[27] G. S. Morrison and E. Enzinger, 'Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction', Speech Commun., vol. 85, pp. 119–126, 2016, doi: 10.1016/j.specom.2016.07.006.

THE EFFECT OF TARGETED VOICE MANIPULATIONS ON LONG-TERM ACOUSTIC CHARACTERISTICS

Alžběta Houzar, Tomáš Nechanský & Radek Skarnitzl

Institute of Phonetics, Charles University, Czech Republic
alzbeta.houzar@ff.cuni.cz; tomas.nechansky@seznam.cz; radek.skarnitzl@ff.cuni.cz

ABSTRACT

The human voice is characterized by enormous flexibility and plasticity, which is highly relevant for speaker identification. Even though naturalistic voice disguise in the forensic context tends to manifest little sophistication, some perpetrators have been shown to change atypical or more speech parameters. This paper aims to find how and whether fifteen targeted manipulations in the articulatory and phonation domains affect the characteristics of the fundamental frequency, long-term formant values, harmonicity, and the spectral slope when compared to the ten Czech male speakers' habitual voice. Whilst we anticipated that formant values would be influenced by articulatory settings rather than phonatory modifications; and that, on the other hand, the spectral slope and the harmonics-to-noise ratio would mostly reflect voice quality changes, our results turned out to be less straightforward, and it would be misleading to draw such general conclusions. Possible reasons for these findings are discussed.

Keywords: voice manipulation, fundamental frequency, formants, harmonicity, spectral slope

1. INTRODUCTION

In everyday life, a speaker's voice characteristics change due to a wide range of factors. These include, first, behavioural factors stemming from speakers themselves, such as using different speech styles [1, 2], being under the effect of various affective states and especially emotions [3, 4], speaking in a loud voice [5] or using whispered [6] speech. The sound of our voice also changes when we are speaking in different languages [7, 8] or even accents [9]. The person to whom we are talking may also affect our speech production, in what is referred to as phonetic accommodation, or vocal convergence toward one's conversation partner [10]. The second group of factors may be described as physiological; these include the effects of alcohol consumption [11], vocal fatigue [12], or the time within the female menstrual cycle [13] on speech; all these are also known to affect the sound of a speaker's voice. Naturally, the voice changes with time as well; various studies have examined the effect of different time spans on the voice, from changes within one day [14] to changes across a number of years [15].

All these shifts in our voices, all this within-speaker variability is made possible by the tremendous plasticity of the speech production mechanism [16, 17]: while our physiology imposes some limits on the sound of our voice, it allows for great variability, with countless degrees of freedom, that we use on a daily basis to express the various components of communicative intent [18]. This intra-speaker variability is, of course, of vital importance in the forensic phonetic context: an analyst comparing two voices and deciding on or against their identity must be aware of possible within-speaker differences.

One behavioural factor, which has not been listed above and which is also crucial in forensic voice comparison, is intentional voice disguise. Voice disguise refers to a speaker's deliberate attempt to modify the sound of their voice and thus to conceal their vocal identity. Forensic speech examiners encounter voice disguise particularly in crimes where the perpetrator suspects that they may be recognized (for example, in an anonymous blackmail telephone) or that the telephone call may be recorded (when calling an emergency line, for instance) [19].

When disguising their voice, perpetrators typically rely on shifting their speaking fundamental frequency (typically upwards for male speakers and downwards for females [19, 20]) placing a foreign object in front of or into their mouth (such as holding a tin can in front of their mouth as a resonator, covering their mouth with a handkerchief or cloth, or holding a pen between their teeth [21]), on imitating a regional dialect or foreign accent, changing the temporal patterning of their speech (e.g., monotonous syllable durations, unnatural pausing), or changing their voice by modifying some articulatory or phonatory settings [21, 22]. Fortunately, sophisticated complex shifts, involving more of the above-mentioned voice disguise strategies, are quite rare.

Outside of the forensic context, studies instructing speakers to change their voice as much as they can seem to confirm the rather simple strategies adopted by most speakers. However, a paper on speakers of Czech [23] identified several speakers who performed multiple, sophisticated modifications and still managed to sound natural; these changes resulted in significantly lower recognizability of the speakers. In his seminal study, Nolan examined the effect of targeted voice manipulations on the speech signal,

using his own voice [18]. In this study, we focus on targeted voice disguise in speakers of Czech and examine the effect of fifteen modifications on long-term formant, spectral, harmonic and f_0 characteristics.

We hypothesize that (1) formant values will mostly be affected by articulatory modifications, as they are considered an acoustic correlate of articulation setting, with F3 being relatively most stable [24]; (2) spectral and harmonic characteristics will depend mainly on phonatory modifications; and (3) f_0 will remain relatively stable across most settings, as speakers were instructed to only perform the targeted manipulation and not change other aspects of their speech like pitch.

2. METHOD

2.1. Material

Ten male speakers of Common Czech were analyzed in this study. All of them are experienced voice users, trained in phonetics, and they were generally able to perform the targeted voice manipulations. They were instructed to read a short text (translation of the Rainbow Passage) in their habitual voice, and fifteen more times, always performing a different modification.

The modifications were chosen mostly based on the SVPA scheme [25]:

- lip-spreading and lip-rounding
- closed jaw and open jaw
- palatalization and pharyngealization
- nasalization and denasalization
- pressed, breathy, whispery, and creaky voice

In addition, we included three combinations of one phonatory and one articulatory modification:

- spread lips and breathy voice
- labialized (rounded lips) and whispery voice
- open jaw and creaky voice

The recordings were obtained at the sound-treated recording studio of the Charles University's Institute of Phonetics, using 48kHz sampling frequency and 16-bit quantization. The speakers were told to repeat a passage in case they failed to maintain the targeted voice manipulation; the recordings were later edited so as to contain the best realization of each sentence vis-à-vis the intended modifications. As two speakers were not able to perform all the modifications, we analyzed 157 recordings in total (10 speakers * 16 versions – 3 not performed).

2.2. Analyses

Having extracted vocalic streams in Praat Vocal Toolkit [26], we dealt with the following acoustic signal domains: the fundamental frequency (f_0), long-term formants (LTF), harmonics-to-noise ratio (HNR), and long-term average spectrum (LTAS). All values were obtained automatically in Praat [27].

As for f_0 values, the median and alternative baseline (the level below which 7.64% of f_0 values fall [28]), were extracted in the 60–350 Hz range. Regarding LTF [29], first to third formant values were measured every 10 ms with the default 'robust' settings; then, all values below the 5th and above the 95th percentile were removed. For HNR [30], the floor of 60 Hz and standard cross-correlation settings for harmonicity extraction were used. Finally, as concerns LTAS, we focused on three spectral slope parameters: the Hammarberg index [31], which reports the difference between spectral amplitude (dB) maxima in the 0–2 and 2–5 kHz ranges; the alpha ratio [32], which describes the energy (Pa^2·s) ratio between the 0–1 and 1–5 kHz ranges; and the BgNoF0 index [33], which corresponds to the energy (Pa^2·s) ratio between the 350–1100 and 2300–5500 Hz ranges, thus without (most of) the fundamental frequency and F2 ranges.

To assess whether a modification had influenced the speech parameters in a consistent way across all speakers, we conducted the Wilcoxon paired test and applied the Bonferroni correction.

For visualization purposes, the data were normalized, so that the manipulated voice values show the difference from the habitual voice values.

3. RESULTS

3.1. The fundamental frequency

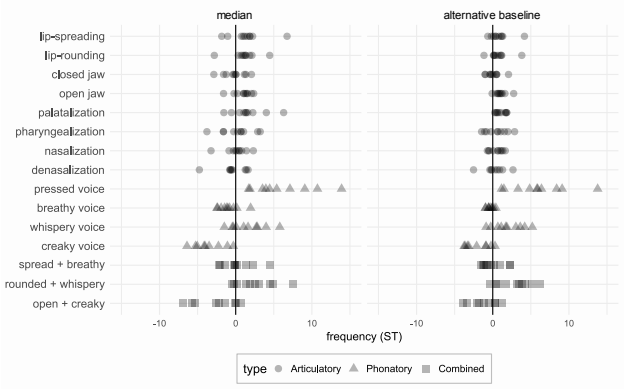


Figure 1: f_0 median and alternative baseline shifts (in ST) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

As is shown in Figure 1 and also in Table 1 below, there is a consistent shift of both the median and the alternative baseline in pressed phonation, which has also been confirmed statistically. As expected, a certain tendency could be observed in creaky voice as well; however, only a marginally significant result was returned regarding the f_0 median. It is noteworthy that the combined modifications copy the single

phonatory trends. As for the articulatory settings, only palatalized speech resulted in a homogeneous and statistically significant rise of the baseline value.

3.2. Long-term formants

Contrary to our assumptions, we witness shifts in long-term first to third formant values (LTF1–3) in relation to not only articulatory, but also phonatory manipulations. Even though LTF3 seems relatively stable in Figure 2, there are 5 significant comparisons (see Table 1). A significant increase/decrease in LTF1 has been revealed in 10 modifications; and 6 settings regularly influenced LTF2. Interestingly, lip-rounding caused a significant drop in all LTF values; and breathy phonation (both on its own and in combination with lip-spreading) lowered LTF1 but raised LTF2 and LTF3 medians.

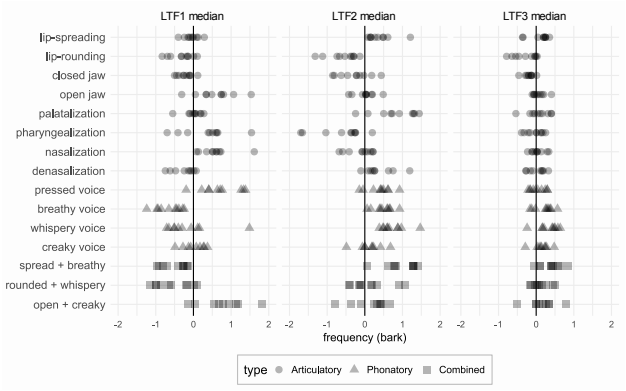


Figure 2: Long-term first to third formant shifts (in barks) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

3.3. Harmonics-to-noise ratio

Unsurprisingly, Figure 3 shows that speakers appear to have randomly oscillated around the HNR mean when performing articulatory manipulations. Regarding voice quality, which is more striking, the only significant exception is the breathy phonation alone and in combination; although creaky voice (again both single and combined) exhibits marginally significant results as well. As anticipated, creaky quality decreased HNR; nevertheless, breathy proved to have done the opposite.

3.4. The spectral slope

Finally, we were interested whether three spectral slope metrics would reflect differences between any of the manipulated and normal setting. As for BgNoF0, we cannot confirm this to be the case since no comparison turned out to be even marginally significant. Concerning the Hammarberg index and the alpha ratio, once again it was the breathy phonation that consistently shifted both indexes.

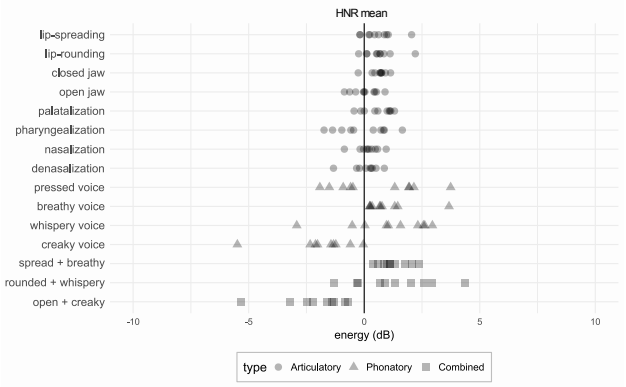


Figure 3: Harmonics-to-noise ratio shifts (in dB) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

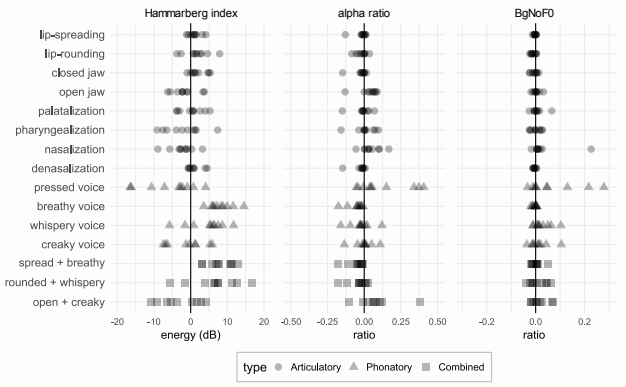


Figure 4: Spectral slope shifts (in Pa2·s and dB) in individual modifications. The vertical line represents habitual voice; individual points represent speakers.

To find out how all the voice manipulations shifted the studied acoustic parameters in individual speakers, see an interactive visualisation at http://tinyurl.com/voice-manipulations.

4. DISCUSSION

This study focused on targeted voice manipulations by experienced Czech voice users. The ability to perform (at least the majority of) these manipulations was crucial, and that is why the speaker sample is relatively limited: since performing these modifications is highly demanding, it would be unrealistic to obtain significantly more than ten speakers.

It should be pointed out that our results may have been affected by the different durations of extracted vocalic streams: due to the lack of periodic signal especially in phonatory settings, the durations ranged from 8.2 to 44.7 seconds (mean 29.5 s, SD 4.9 s).

As for the results, it seems that most manipulations brought along changes of parameters that were not predicted by our hypotheses. Some of

modification	median f_0	alternative baseline f_0	LTF1	LTF2	LTF3	HNR	Hammarberg index	alpha ratio	BgNoF0
lip-spreading	0.064	0.037	0.188	0.004	0.454	0.010	0.084	0.124	0.152
lip-rounding	0.064	0.049	**0.001**	**<0.001**	**<0.001**	0.010	0.193	0.124	0.027
closed jaw	0.722	0.922	**0.001**	0.007	**<0.001**	0.004	0.020	0.059	0.260
open jaw	0.084	0.004	**<0.001**	0.188	0.679	0.799	0.193	0.105	0.721
palatalization	0.049	**0.002**	0.934	**0.001**	0.599	0.024	0.625	0.906	0.770
pharyngealization	0.846	0.375	0.121	**<0.001**	0.421	0.846	0.275	0.415	1.000
nasalization	0.557	0.105	**<0.001**	0.041	0.934	0.193	0.064	0.037	0.037
denasalization	0.910	0.426	**0.002**	0.020	0.855	0.496	0.203	0.012	0.058
pressed	**0.002**	**0.002**	**<0.001**	0.004	0.978	0.322	0.037	0.084	0.124
breathy	0.064	0.049	**<0.001**	**<0.001**	**0.002**	**0.002**	**0.002**	**0.002**	0.286
whispery	0.084	0.014	0.064	**<0.001**	**0.001**	0.105	0.037	0.160	0.041
creaky	0.004	0.012	0.952	0.761	0.542	0.004	0.570	0.910	0.570
spread + breathy	1.000	0.846	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.002**	**0.002**	0.919
rounded + whispery	0.020	0.004	**0.001**	0.762	0.073	0.049	0.014	0.037	0.557
open + creaky	0.030	0.074	**0.001**	0.426	0.715	0.004	0.203	0.098	0.359

Table 1: Significant ($p \leq 0.0033$, in black) and marginally significant ($p \leq 0.0066$, in darker grey) p-values returned by the Wilcoxon paired test after Bonferroni correction ($\alpha = 0.05/15$).

these divergences should not be surprising, however, given the multiple interconnections of the phonatory and articulatory apparatus via the palatopharyngeus muscles, or of tongue position with nasality via the palatoglossus muscle [34]. For instance, most articulatory manipulations also triggered a change in fundamental frequency. This seems understandable with palatalization or pharyngealization, modifications which pull the larynx upwards or downwards, respectively, but is much less straightforward, for example, with lip-spreading.

One of our hypotheses stated that long-term formant values, being dependent on the shape of the vocal tract, would be mostly affected by the articulatory settings. Whereas we were able to confirm the widely known influence of lip-rounding, palatalization and pharyngealization, or jaw opening and closing on the drop or rise of either one or both F1 and F2, other findings were less common.

Firstly, formant values were significantly shifted in all settings but creaky phonation, which implies an articulation change above the larynx. Secondly, our data show an LTF1 increase in nasalization and a decrease in denasalization, which is probably caused by movements of the larynx when anchoring the velum to an opened and closed position, respectively. Thirdly, even though F3 is regarded as somewhat speaker-specific, it partially does reflect vowel quality. Based on our data, it is possible to confirm a stable F3 drop for lip-rounding and the closed jaw modification; with open articulation, the opposite trend is visible, though not significant (see Table 1).

Concerning phonatory modifications, it should be mentioned that even though HNR is usually measured in the middle third of sustained vowels, some researchers [35] do make use of long-term vocalic streams similarly to the presented paper. Creaky, whispery, and breathy phonation were anticipated to result in lower HNR since such voice qualities are expected to be less periodic than the modal voice. Whereas such a tendency was observed with creaky voice in our data (marginal significance), breathy phonation yielded the opposite result.

Spectral slope is considered another acoustic correlate of voice quality; hence, we assumed that phonatory settings would consistently shift the spectral slope values. Steeper spectral slope, with less energy in higher frequencies, has been reported for breathy phonation [33], which is also supported by our results. Conversely, creaky [31] and pressed [16] voice have been associated with flatter spectral slope, but neither has been confirmed in the present study.

To conclude, voice disguise is a phenomenon with which forensic voice comparison practitioners have to calculate [19, 22]. It would be beneficial, upon detection that a given case involves voice disguise, to be able to extrapolate acoustic values between speakers' disguised voice used in unknown recordings and their habitual voice obtained, for example, during interrogation. However, the multiple differences observed in our data beyond those that were predicted, as well as some acoustic manifestations which countered our hypotheses, do not warrant such a procedure as realistic. Nevertheless, our study provides another evidence of the fascinating plasticity of our voices.

5. ACKNOWLEDGEMENTS

This study was supported by the Grant Schemes at CU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935; the last author was supported by the ERDF-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

6. REFERENCES

[1] M. Jessen, "Forensic phonetics and the influence of speaking style on global measures of fundamental frequency," in *Formal linguistics and law*, G. Grewendorf and M. Rathert, Eds. Mouton de Gruyter, 2009, pp. 115–139.

[2] K. McDougall, and M. Duckworth, "Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English," *Int. J. of Speech, Lang. and Law*, vol. 25, pp. 205–230, 2018.

[3] E. J. Eriksson, R. D. Rodman, and R. C. Hubal, "Emotions in speech: Juristic implications," in *Speaker classification I*, C. Müller, Ed. Springer-Verlag, 2007, pp. 152–173.

[4] K. R. Scherer, "Acoustic patterning of emotion vocalization," in *Oxford handbook of voice perception*, S. Frühholz and P. Belin, Eds. Oxford University Press, pp. 61–91.

[5] H. Traunmüller, and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, vol. 107, pp. 3438–3451, 2000.

[6] Y. Swerdlin, J. Sith, and J. Wolfe, "The effect of whisper and creak on vocal tract resonances," *J. Acoust. Soc. Am.*, vol. 127, pp. 2590–2598, 2010.

[7] S. X. Chen, and M. H. Bond, "Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context." *Pers. Soc. Psychol. B.*, vol. 36, pp. 1514–1528, 2010.

[8] B. Lee, and D. Van Lancker Sidtis, "The bilingual voice: Vocal characteristics when speaking two languages across speech tasks," *Speech Lang. Hearing*, vol. 20, pp. 174–185, 2017.

[9] B. G. Evans, and P. Iverson, "Plasticity in vowel perception and production: A study of accent change in young adults, *J. Acoust. Soc. Am.*, vol. 121, pp. 3814–3826, 2007.

[10] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, pp. 2382–2393, 2006.

[11] B. Baumeister, C. Heinrich, and F. Schiel, "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers," *J. Acoust. Soc. Am.*, vol. 132, pp. 442–451, 2012.

[12] V. J. Boucher, and T. Ayad, "Physiological attributes of vocal fatigue and their acoustic effects: A synthesis of findings for a criterion-based prevention of acquired voice disorders," *J. Voice*, vol. 24, pp. 324–336, 2010.

[13] M. Hejná, "A case study of menstrual cycle effects: global phonation or also local phonatory phenomena?", *Proc. 19th ICPhS*, Melbourne, 2019, paper 13.

[14] M. Artkoski, J. Tommila, and A.-M. Laukkanen, "Changes in voice during a day in normal voices without vocal loading," *Log. Phon. Vocol.*, vol. 27, pp. 118–123, 2002.

[15] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *Int. J. of Speech, Lang. and Law*, vol. 24, pp. 177–199, 2017.

[16] J. Laver, *The phonetic description of voice quality*. Cambridge University Press, 1980.

[17] F. Nolan, "Degrees of freedom in speech production: An argument for native speakers in LADO," *Int. J. of Speech, Lang. and Law*, vol. 19, pp. 263–289, 2012.

[18] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge University Press,1983.

[19] A. Braun, "Stimmverstellung und Stimmenimitation in der forensischen Sprechererkennung," in *Das Phänomen Stimme: Imitation und Identität*, T. Kopfermann, Ed. St. Ingbert, 2006, pp. 177–181.

[20] H. J. Künzel, "Effects of voice disguise on speaking fundamental frequency," *Forensic Linguist.*, vol. 7, pp. 149–179, 2000.

[21] R. M. Figueiredo, and H. S. Britto, "A report on the acoustic effects of one type of disguise," *Forensic Linguist.*, vol. 3, pp. 168–175, 1996.

[22] H. Masthoff, "A report on a voice disguise experiment," *Forensic Linguist.*, vol. 3, pp. 160–167, 1996.

[23] A. Růžičková, and R. Skarnitzl, "Voice disguise strategies in Czech male speakers," *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, pp. 19–34, 2017.

[24] R. Skarnitzl, and T. Nechanský, "Segmental cues," in *Oxford handbook of forensic phonetics*, F. Nolan, K. McDougall, and T. Hudson, Eds. Oxford University Press, 2023 (in print).

[25] E. San Segundo, and J. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *J. Voice*, vol. 31, pp. 644.e11–644.e27, 2017.

[26] R. Corretge, *Praat Vocal Toolkit*, retrieved from https://www.praatvocaltoolkit.com, 2022.

[27] P. Boersma, and D. Weenink, *Praat*, retrieved from http://www.praat.org, 2020.

[28] J. Lindh, and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Proc. Interspeech*, Antwerpen, 2007, pp. 2025–2028.

[29] F. Nolan, and C. Grigoras, "A case for formant analysis in forensic speaker identification," *Int. J. of Speech, Lang. and Law*, vol. 12, pp. 143–173, 2005.

[30] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, pp. 1544–1550.

[31] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngol.*, vol. 90, pp. 441–451, 1980.

[32] J. Sundberg, and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 453–457, 2006.

[33] J. Volín, and J. Zimmermann, "Spectral slope parameters and detection of word stress," *Technical Computing Prague*, pp. 125–130, 2011.

[34] A. Marchal, *From speech physiology to linguistic phonetics*. John Wiley & Sons, 2009.

[35] V. Hughes, A. Cardoso, P. Foulkes, P. French, A. Gully, and P. Harrison, "Forensic voice comparison using long-term acoustic measures of laryngeal voice quality," *Proc. 19th ICPhS*, Melbourne, 2019, paper 639.

ANALYSIS OF SPECTRAL FEATURES' CHANGE FOR SPEAKER COMPARISON

Tekla Etelka Grászi[1], Valéria Krepsz[1, 2], Anna Huszár[1], Bettina Száraz[1], Andrea Deme[3], Kornélia Juhász[1, 3], Alexandra Markó[1, 4]

[1]Hungarian Research Centre for Linguistics, [2]Humboldt-Universität zu Berlin, [3]Eötvös Loránd University, [4]SSNS Institute for Expert Services

graczi.tekla.etelka@nytud.hu, krepsz.valeria@nytud.hu, huszar.anna@nytud.hu, szaraz.bettina@nytud.hu, deme.andrea@btk.elte.hu, juhasz.kornelia@nytud.hu, marko.alexandra.phd@gmail.com

ABSTRACT

Medium-term changes of speech parameters are understudied, while forensic tasks often require the comparison of non-contemporary speech samples. The present study compares segmental features of the same speakers' speech recorded a decade apart.

F_1–F_4 values of five vowels and four spectral moments of five voiceless obstruents were measured in Hungarian reading aloud tasks. The ratios of the mean values were compared between-sessions, and the possible group-level change was tested by LMM. The relevance of the spectral features in speaker verification was analysed using LDA within the first and between the two recording sessions.

Most spectral features showed high intraspeaker variability in terms of their change between the sessions. The LDA-results showed that the sibilants may have similar relevance in within- and between sessions speaker verification, while vowels though having the highest relevance in within-session verification, lose on this in the case of between-session comparisons. Stops had the lowest relevance.

Keywords: spectral changes, speaker comparison, vowels, voiceless obstruents

1. INTRODUCTION

The most common task in forensic speaker identification involves acoustic and auditory comparison of several voice samples. Traditionally, analyses primarily focus on voice parameters (mainly f_0 and voice quality) and vowel formants [1, 2]. However, in studies of English spectral features of consonants have also been involved in forensic context recently [3]. Study of spectral parameters of /m, n, ŋ, l, s/ [4] found high speaker-specificity effect of /m/ and /s/. Earnshaw [5] studied voiceless plosives (/t/, /k/ and /t/ & /k/combined) in terms of VOT, closure duration, ratio, duration & VOT, VOT & ratio, and all three combined. The best Likelihood Ratio results were achieved using the combined /t/ & /k/ data in the VOT & ratio system. A study of three sibilants /s, z, ʃ/ focusing on both static features (intensity, CoG, SD, skewness, and kurtosis) and dynamic ones (CoG depending on F_2 vowel onset and offset) [3] revealed high speaker-specificity of CoG, SD, and intensity (while the latter highly depends on the recording circumstances).

The comparison of temporally distant speech samples, which is the topic of interest of the present study, is relatively understudied, although this is the typical setting in the context of forensic speaker comparisons. The extent of the temporal gap between recording sessions is different case by case, which might vary from a couple of months to several years (the longest documented interval is 27 years [6]).

The relatively long time lag between the first and the second recordings of the speech can be interpreted from the aspect of age-related speech changes (disregarding here other factors, like speech mode, health, sociophonetic variation, context, etc., that may affect the variation of speech parameters). However, age-related speech changes are not linear, they might be abrupt and are influenced by several factors, e.g., the age of the speaker at the time of recording.

With respect to the comparison of non-contemporary samples, Rhodes [2] studied 8 speakers (6 men and 2 women) between the ages of 21 and 49, in 7-year intervals. He reported large interspeaker variation in all parameters. F_1 changed between the analysed samples and decreased from 21 to 49 years of age (for all but one vowel). However, the change showed large individual differences. F_2 and F_3 showed similar but less generalizable tendencies within the 28 years but no clear tendencies were found in the 7-year interval comparisons. Consonants were compared both in contemporaneous and non-contemporaneous pairs of speech samples in [5]; however, the exact temporal distance between the latter ones is not known.

In the present study, we analyse speech samples from two age groups, young and middle aged speakers, i.e., the typical population of forensic speaker analysis, and compare samples that were recorded from them with a 10-year-long time lag.

Within the age interval under analysis, i.e. in 10 years, no large biological changes are expected due to aging, although minor differences resulting from lifestyle, work environment, smoking or hormonal changes (especially in women) might appear.

The realization of segmental parameters in part are language-dependent, which should be considered in the methodology of forensic analysis [1]. In our study, we analyse realizations of vowels and consonants of Hungarian language which differs from English in many respects [7].

The present explorative study raised the following questions. Is there in group-level difference in vowel or consonant realizations between speech samples recorded with a time lag of 10 years? How reliably can formant and spectral moment values be used in non-contemporary speaker comparison?

2. METHODS

The speech of eight male and eight female speakers was analysed. All of them participated in collection of speech samples for 'BEA' database [8], and approximately ten years later in another for "Longitudinal" database [9] (henceforward, 'LD'). LD was established in 2017, in order to investigate non-contemporary speech variability and its effect on speaker verification. All the recording circumstances of the two databases were kept identical. The speakers' age varied between 20 and 34 years at the first recording and between 30 and 44 years at the second one.

Two reading tasks (15 separate sentences, and a 13-sentence long text) were used from the two databases. 16 occurrences of five (interconsonantal) vowels /ɒ, aː, i, ɛ, eː/ and 8 (intervocalic) occurrences of five consonants /s, ʃ, p, t, k/ were selected. Target sounds were manually labelled and automatically measured in Praat [10]. In vowels, the first four formants' frequencies were measured at the midpoint of the duration with the Formant (burg) algorithm of Praat with the basic settings (5 formants in the range of 0–5000 Hz for men, and 0–5500 for women). For the consonants, the four spectral moments were measured in the sound files resampled at 22 kHz in a 0.01 s window at the midpoint of the time interval of the fricatives, and from the start of the release burst of the stops. In the case of alternative realizations of stops (spirantisation), the mid 0.01 s was selected for measurement. Outliers were eliminated by speaker, by vowel, and by recording session. The ratios of the median spectral values in the LD to the BEA were calculated ('LD to BEA ratio'). The inter-database, i.e. between-session differences were tested by linear mixed models (LMM) [11, 12, 13, 14, 15]. The dependent variable was one of the attested spectral features. After model selection (anova()), the fixed effects were RECORDING ('BEA' vs. 'LD', GENDER, and SPEECH SOUND allowing for their interaction, and the model included random slopes only on RECORDING by SPEAKERS. A simpler model was selected for kurtosis, where the factors CONSONANT and RECORDING were the fixed effects, and the model only included random intercepts by SPEAKER. *P*-values were obtained with Satterthwaite-approximation by anova().

We opted for LDA analysis [16], as in the context of speaker comparison, this method is used "to assess the speaker-specificity of a given variable and how useful it might be as a parameter in forensic casework" ([4]: 68). LDAs were run separately for the speech sounds and genders using all four spectral features measured for the specific speech sound. Spectral features were Z-transformed within the speech sound categories. A set of LDAs (2 speech sound groups * 5 speech sounds * 2 genders) was run within the first ('BEA') recording dividing the datasets into training and testing sets (.6 : .4 ratio for vowels, and .5 : .5 ratio for consonants for each speaker randomly). Another set of LDAs was trained on the dataset of the 'BEA' recordings, and tested on the dataset of the 'LD' recordings. The true positive and the true negative ratios were calculated over speakers. True positive ratio is the ratio of the items which were correctly classified to originate from the specific speaker to all the items that originate from the same speaker. True negative ratio is the ratio of the items correctly classified as not originating from a specific speaker to all the items that originate from another speaker.

3. RESULTS

3.1. Between-database comparison of spectral features

Median formant frequency values compared between the first ('BEA') and the second ('LD') recording showed large variability across vowels, speakers, and formants, and suggested no group-level changes over time in the studied time frame (Fig.1). F_1-data in women's pronunciation and the first two formants in men's pronunciation would suggest some possible change; however, these group-level differences did not appear on the speaker level, but in different speakers and vowels. No significant main effect of RECORDING for F_1 was found (LMM: Tab.1.) Although the interaction of the factors RECORDING and VOWEL was significant, post hoc tests revealed no significant F_1-shift tendency between recording sessions. RECORDING played a role neither as main effect, nor in any of the interactions on F_3. A general, small, significant increase was found in F_2 (estimated

means: 1329 Hz, 1338 Hz, which is a negligible ratio of +0.006), and F_4 (estimated means: 3787 Hz, 3821 Hz, which is a negligible ratio of +0.009). However, post hoc tests for the significant threefold interaction did not reveal any vowel- and/or gender-specific tendency for any of these formants. This means that, no group-level change was detected between the two recording sessions for any of the formants in any of the vowels.

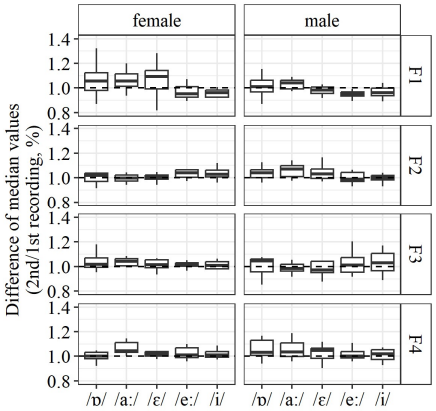


Figure 1: Change (%) of the formant values from the first ('BEA') to the second ('LD') recording (in 10 years) of the analysed vowels by gender.

factor	F	p	F	p
	F_1		F_2	
R	F(1,14.0)=1.8,	n.s.	F(1,14.3)=7.6	*
V	F(4,2251.3)= 1692.7	***	F(4,2143.0)= 1688.1	***
G	F(1,14.0)=22.6	***	F(1,14.03)=36.8	***
R:V	F(4,2251.3)=9.6	***	F(4,2143.3)=0.9	n.s.
R:G	F(1,14.0)=0.02	n.s.	F(1,14.3)=0.04	n.s.
V:G	F(4,2251.3)=14.8	***	F(4,2143.0)=9.7	***
R:V:G	F(4,2251.3)=1.3	n.s.	F(4,2143.3)=3.5	**.
	R_m^2=.693, R_c^2=.789		R_m^2=.756, R_c^2=.802	
	F_3		F_4	
R	F(1,14.0)=2.938,	n.s.	F(1,14.0)=15.7	*
V	F(4,2309.1)= 1136.3	***	F(4,2289.4)=14.3	***
G	F(1,14)=33.3	***	F(1,14.0)=5.7	*
R:V	F(4,2309.2)=0.5	n.s.	F(4,2289.4)=1.7	n.s.
R:G	F(1,14.0)=0.03	n.s.	F(1,14.0)=10.0	n.s.
V:G	F(4,2309.2)=0.9	n.s.	F(4,2289.4)=6.2	***
R:V:G	F(4,2309.2)=1.5	n.s.	F(4,2289.4)=3.5	**
	R_m^2=.284, R_c^2=.416		R_m^2=.104, R_c^2=.296	

Table 1: The results of the linear mixed models for the vowel formants. (R = recording, G = gender, V = vowel. Grey cells = significant differences.) (Sign. levels: 'n.s.' = not significant, '*' < .05, '**' < .01, '***' < .001).

Alternative realizations were found in some stops (non-burst release, but a fricative-like opening, i.e., with lower intensity, or spirantization/constant leak along the entire duration): in 10% of /p/, 2% of /t/, and 18% of /k/ sounds, altogether in the two databases. The 'LD to BEA' ratio of CoG showed larger interspeaker variability in the stops than in the sibilants (Fig.2). SD for women showed similar tendencies as CoG, but not for men. Between-session differences in skewness and kurtosis showed larger variability across speakers within the fricatives than CoG and SD. LMMs (Tab. 2) found a significant effect of RECORDING as main effect for skewness (estimated mean and change: +3.77, +2.53, which means a ratio of +1.67) and kurtosis (estimated mean and change: +28.21, +45.98, which means a ratio of +2.63). The interaction of the factors RECORDING and CONSONANT also significantly affected skewness and kurtosis. The post hoc test of these interactions found a significant difference in skewness of /p, k/ and kurtosis of /p, t, k/ between the two recording sessions. RECORDING as main factor did not have a significant effect on CoG and SD, while the interaction of RECORDING and CONSONANT was significant for these spectral moments, and the threefold interaction for CoG. However, the post hoc test of these interactions revealed no group-level tendencies in the change of CoG and SD between the two recording sessions.

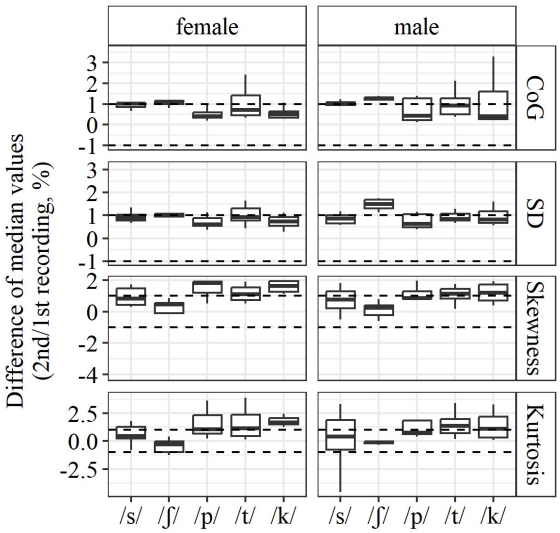


Figure 2: Median change of the spectral moments of the analysed consonants from 'BEA' to 'LD' (ratio).

3.2. Results of LDAs

The true negative ratios (TNR) did not considerably change between the two LDA scenarios (contemporary: 0.87–0.94, non-contemporary: 0.83–0.91). The true positive ratios (TPR) for the vowels

decreased approximately to their halves (contemporary: 0.33–0.61, non-contemporary: 0.10–0.30), especially for men's production. The TPRs for the consonants were considerably lower than for the vowels (contemporary: 0.11–0.3, non-contemporary: 0.11–0.32). In the contemporary tests /ʃ/ for women (0.29), and /s/ for men (0.38), in the non-contemporary tests both sibilants for women (/ʃ/: 0.32, /s/: 0.24) and /ʃ/ for men (0.29) reached considerably higher TPRs than the further consonants which's TPRs stayed below 0.2.

factor	F	p	F	p
	CoG		SD	
R	F(1,14.1)=0.1	n.s.	F(1,14.0)=0.3	n.s.
G	F(1,13.9)=15.3	**	F(1,14.0)=0.8	n.s.
C	F(4,1186.2)= 1812.9	***	F(4,1209.1)=19.8	***
R:G	F(1,14.07)=2.6	n.s.	F(1,14.0)=0.8	n.s.
R:C	F(4,1186.0)=5.0	***	F(4,1209.2)=10.1	***
G:C	F(4,1186.2)=16.8	***	F(4,1209.1)=2.8	*
R:G:C	F(4,1186.0)=2.6	*	F(4,1209.2)=0.9	n.s.
	R_m^2=.844, R_c^2=.859		R_m^2=.089, R_c^2=.318	
	Skewness		Kurtosis	
R	F(1,14.3)=9.4	**	F(1,1228.2)=52.4	***
G	F(1,14.0)=0.73	n.s.		
C	F(4,1226.0)= 258.7	***	F(4,1228.0)=40.1	***
R:G	F(1,14.3)=0.1	n.s.		
R:C	F(4,1226.0)= 17.2	***	F(4,1228.1)=12.2	***
G:C	F(4,1226.0)= 2.0,	n.s.		
R:G:C	F(4,1226.0)= 0.9	n.s.		
	R_m^2=.444, R_c^2=.522		R_m^2=.163, R_c^2=.221	

Table 2: LMM results for the non-contemporary comparison of the spectral features (Sign. levels: 'n.s.' = not significant, '*' < .05, '**' < .01, '***' < .001).

4. DISCUSSION AND CONCLUSIONS

The present paper discussed the relevance of speaker-specific variability of eight spectral features', namely F_1–F_4 of five vowels and the four spectral moments of five voiceless obstruents between speech samples of Hungarian from a decade apart.

F_1 of the low, and mid-low vowels in women's pronunciation, and F_3, and F_4 in men's pronunciation showed larger interspeaker variability in the change between the two recordings. Only the F_2 and F_4 differences between the two recordings were found to be significant (LMM); however, their ratios of the estimated means were below 10%. As for the consonants, the CoG of the stops in men's pronunciation, and the skewness and kurtosis for most consonants showed large interspeaker variability in the comparison of the two recording sessions. Group-level change was found for the skewness of /p, k/, and the kurtosis of /p, t, k/. The low number and random occurrence of alternative realizations of stops do not explain the results of kurtosis and skewness.

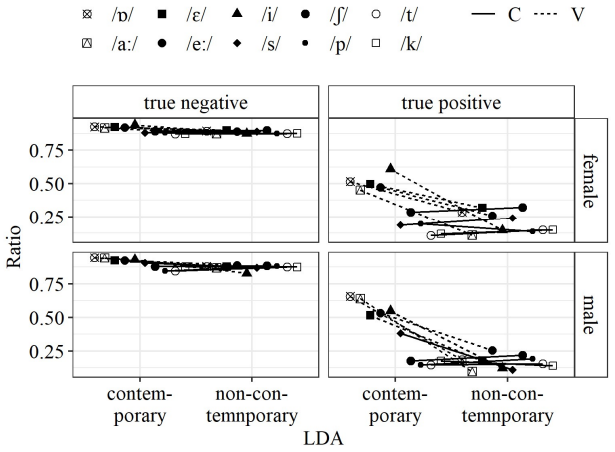


Figure 3: The true negative and the true positive values of LDAs within 'BEA' and between the two databases.

The true negative rates of LDA results did not vary considerably between the contemporary and non-contemporary testing. Its values ranged quite high, above 0.8 due to the relatively high number of speakers (at least for LDA). The TPRs for LDA tests using the spectral moments of the consonants were low, but they were higher for sibilants than for stops. These values did not change considerably between the two LDA scenarios. The TPRs for vowels were twice-trice higher in the contemporary than in the non-contemporary LDA tests. In the non-contemporary comparison, the TPRs for vowels and sibilants were similar.

Based on these results, we conclude that our hypotheses are not supported, as there is no considerable difference among the LDA tests using different speech sounds.

The limitations of our study are the relatively low number of speakers, and the low number of obstruents analysed. The results, however, suggest that the interspeaker differences in the between-recording variance in non-contemporary speech samples raise difficulties in forensic tasks.

5. ACKNOWLEDGEMENT

The research was funded by the National Research, Development, and Innovation Office (grant number: FK128814).

6. REFERENCES

[1] Rose, P. 2002. *Forensic Speaker Identification.* Taylor & Francis.

[2] Rhodes, R. W. 2012. *Assessing the strength of non-contemporaneous forensic speech evidence.* PhD thesis. The University of York.

[3] Fernandez, S. C. 2022. Speaker Specific Information in the Acoustic Characteristics of English Fricatives. *IJFL* 3 , 105–115.

[4] Kavanagh, C. M. 2012. *New consonantal acoustic parameters for forensic speaker comparison.* PhD thesis. The University of York.

[5] Earnshaw, K. 2016. Assessing the Discriminatory Power of /t/ and /k/ for Forensic Speaker Comparison using a Likelihood Ratio Approach. *IAFPA 25.* York, 92–93.

[6] French, J. P. F., Harrison, P., Windsor-Lewis, J. 2006. R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. *IJSLL* 13 , 256–273.

[7] Siptár, P., Törkenczy, M. 2007. *The Phonology of Hungarian.* Oxford University Press.

[8] Gósy, M. 2013. BEA – A multifunctional Hungarian spoken language database. *Phonetician* 105, 50–61.

[9] Gráczi, T. E., Huszár, A., Krepsz, V., Száraz, B., Damásdi, N., Markó, A. 2020. Longitudinális korpusz magyar felnőtt adatközlőkről. In Berend, G., Gosztolya, G., Vincze, V. (eds), *XVI. Magyar Számítógépes nyelvészeti konferencia. Proceedings.* University of Szeged. 103–114.

[10] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer.* http://www.praat.org/

[11] R Core Team 2022. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

[12] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *JSS* 67, 1–48.

[13] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J SS* 82, 1–26.

[14] Lenth, R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.3, https://CRAN.R-project.org/package=emmeans.

[15] Bartoń, K. 2022. MuMIn: Multi-Model Inference. R package version 1.47.1, https://CRAN.R-project.org/package=MuMIn.

[16] Venables, W. N., Ripley, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer.

Reviewing the performance of formants for Forensic Voice Comparison: a meta-analysis of forensic speech science research

Lois Fairclough[1], Georgina Brown[1,2] and Christin Kirchhübel[2]

[1]Department of Linguistics and English Language, Lancaster University, UK
[2]Soundscape Voice Evidence, Lancaster, UK
{l.fairclough | g.brown5}@lancaster.ac.uk, ck@soundscapevoice.com

ABSTRACT

It is widely accepted in forensic speech science that formants have speaker discriminatory power, and therefore formant measurements commonly feature in forensic speech casework. Courts in Northern Ireland even go as far as to insist that formant analysis has to be carried out as part of forensic voice comparison analysis. However, work in the broader phonetics field has started to challenge the weight that has been traditionally attached to formants in a range of related subdisciplines – this justifies a review of the use and performance of formants in forensic speech science research studies. This paper therefore presents part of a meta-analysis of forensic speech science studies which test formants as a parameter for speaker discrimination. The results are highly variable across the 277 results from the 37 papers included. Some performance trends that might be expected are not evident from this meta-analysis.

Keywords: forensic phonetics, vowel formants, speaker discrimination

1. INTRODUCTION

The job of a forensic speech analyst typically involves comparing questioned and known speech samples in order to address the question of whether it is the same speaker or different speakers featuring in the recordings. The auditory phonetic and acoustic approach is a well-established method applied to the forensic voice comparison (FVC) task, as shown in a survey of forensic speech practitioner practices [1]. Within this approach, analysts consider multiple parameters, one being vowel formants. In fact, in UK practice, there is a general expectation that practitioners will include an analysis of formants in their voice comparison analyses. This is because formants are said to reflect the resonances of the vocal tract, thereby tapping into the physical characteristics of the speaker [2].

Despite the weight that is given to formants on theoretical grounds, there is not a clear account of how well formants perform in practice at speaker discrimination. In recognition of this, the overarching objective of this paper is to carry out a systematic review of the relevant research literature and to offer a description of the performance of formants for the purpose of discriminating speakers in the forensic setting.

Section 2 lays out further background information and motivation for this work. Section 3 explains the methods used in this meta-analysis, while section 4 reveals the initial findings of the analysis. Section 5 discusses the findings and some implications around the use of formants in forensic speech science.

2. BACKGROUND

This section first describes what formants are, and how they can be measured in forensic speech science research and practice. The section then moves on to outline the status given to formants in FVC.

2.1. What are formants?

Formants represent the resonant frequencies of the vocal tract [3], and as such reflect a speaker's physical and articulatory characteristics. While F1 and F2, to a great extent, reflect the phonetic quality of a vowel, the higher formants (F3, F4 and F5) are said to represent more speaker-specific characteristics [3]–[5]. Generally, F4 and F5 are not available for analysis in forensic casework due to the frequency bandwidth reduction typical of the types of recordings involved.

Formants can be analysed in different ways. Three key ways of measuring formants include midpoint measurements, dynamic measurements and long-term formant analysis (LTF). All three of these appear in the forensic speech science literature. Midpoint measurements capture formant frequencies from the centre of the vocalic segment in question [6], while dynamic measurements capture multiple time points across the segment [7]. LTF analysis captures formants across all vocalic portions of speech in a sample [8]. Within each of the three techniques there are different strategies by which the formant analysis can be carried out. For example, the intervals from which LTFs can be calculated might differ – some might take measurements every 5ms, others every 10ms.

Irrespective of the type of formant measurement used, recent discussion in acoustic phonetics has questioned the reliability of formants as a parameter in speech analysis. This emphasises that the methods used to obtain formants need more attention and caution [9], and reinforces the idea that formant measurements are actually only "estimates", not true values [10]. Together, this puts into question the representation of formants in the forensic literature and beyond.

2.2. Formants in FVC casework

Formants are widely used by forensic speech practitioners. In the survey by [1], it was found that of 36 experts, 97% of them carried out some form of formant analysis. Of the practitioners using formants, all measure F2, 87% measure F1 and F3, while 17% measure F4. Importantly, 94% measured the centre frequencies of monophthongs, 71% reported measuring the trajectories of diphthongs, and 45% measured vowel-consonant or consonant-vowel formant transitions [1]. In addition, practitioners also make use of LTF analysis [11, 12]. Detailed results of formant measurement practice in [13] revealed diversity in carrying out formant measurements, highlighting the flexibility in the field when it comes to integrating formants into FVC analyses.

In the legal context, there is a court ruling in the jurisdiction of Northern Ireland that places an expectation on practitioners to include formants in their FVC analyses. It states that "*no prosecution should be brought in Northern Ireland in which one of the planks is voice identification given by an expert which is solely confined to auditory analysis… there should also be expert evidence of acoustic analysis… which includes formant analysis*" [14]. This reinforces the significance placed on using formant analysis for FVC.

Given the theory that underpins formants, and the significance that has been placed on formants within and outside the forensic speech science community, it seems fitting to carry out a review of the forensic speech science research literature to interrogate their speaker discrimination performance. This can be achieved through a meta-analysis of the existing research literature.

3. METHODOLOGY

In selecting papers for inclusion in the meta-analysis, an initial screening of each paper was undertaken to ensure that formants were used as a parameter for investigation in isolation of other parameters. A basic sampling approach was adopted by collecting papers from Google Scholar, university library online resources, and journals which have particularly relevant contributions to forensic speech science. From this initial set of papers, the references were examined for more papers, increasing the sample size.

The above exercise resulted in over 100 forensic speech science studies being considered for inclusion in the meta-analysis. Reasons for exclusion comprised, among other things, the fact that some studies were only represented by a conference abstract, or they combined formant measurements with other parameters (e.g., MFCCs) to generate results. Of the papers that were included, two key types of study motivation emerged: 1) analysis of formants for the purpose of speaker discrimination and 2) analysis of formants to test their 'robustness' (i.e., within-speaker variation owing to speaker-internal or speaker-external factors). This paper focuses solely on the analysis of speaker discrimination-based results which led to the inclusion of 277 results in this meta-analysis.

Each paper was interrogated according to 23 different factors which cover the nature of the dataset that was used, the formant measurement technique, the analysis, and the results. The present paper pays particular attention to the following factors: measurement type (e.g., midpoint, dynamic, LTF), the number of formants (e.g., F1, F2, F3), and the performance result.

Papers generally reported speaker discrimination performance through Equal Error Rates (EERs) or Classification Rates (CRs). However, there were instances of alternative quantitative measures and qualitative comments which did not fall into the EER or CR categories. EERs reflect the threshold for a system's false acceptance and false rejection rates. The lower the EER, the higher the performance of the measurement parameters for speaker discrimination [15]. CRs most often present the results of a discriminant analysis, whereby linear combinations of features are identified to characterise and group speakers. This multivariate approach can be used to determine whether a set of predictors can be combined to predict group membership, or in this case speaker membership [16]. Importantly, EERs are produced as a result of studying speaker discrimination as an "open-set" type of problem (i.e., does the analysis support the same-speaker or different-speaker view?), whereas the studies that use CRs to reflect speaker discrimination performance assume a "closed-set" type of problem (i.e., which speaker is it out of the "closed-set" of X speakers?). As such, the CR results are heavily dependent on the number of speakers that were included in the "closed set" of speakers.

This paper focuses on EERs and CRs as they account for the majority of results in the speaker discrimination papers.

4. FINDINGS

The meta-analysis results are derived from 37 different papers from 25 different first authors. Some authors feature frequently, with the most frequent appearing as first author on 6 publications and many more as a co-author. The papers included are dated from 1996-2021, therefore spanning 25 years. The number of speakers in each study range from 5 to 171. The findings cover 8 languages including Cantonese, Czech, Dutch, English, German, Mandarin, Shanghainese, and Swedish. The research predominantly included English (21 of the 37 papers). Within English, three varieties were investigated: Australian English, North American English and Standard Southern British English (SSBE). Of the 37 papers, 12 analyse spontaneous speech, 18 read speech, and 7 semi-spontaneous. The majority of the papers are based on laboratory recorded speech (both direct and using telephone transmission), and only two papers are based on casework recordings.

Two mixed-effects linear regression analyses were run for each type of result (i.e., EER and CR). For these analyses, only factors that featured in all of the studies were included. Number of speakers, linguistic variety, and measurement type were included as fixed effects, while the paper ID was included as a random effect. The analysis of EER results brought about one significant effect which corresponded to the number of speakers ($p = 0.04$).

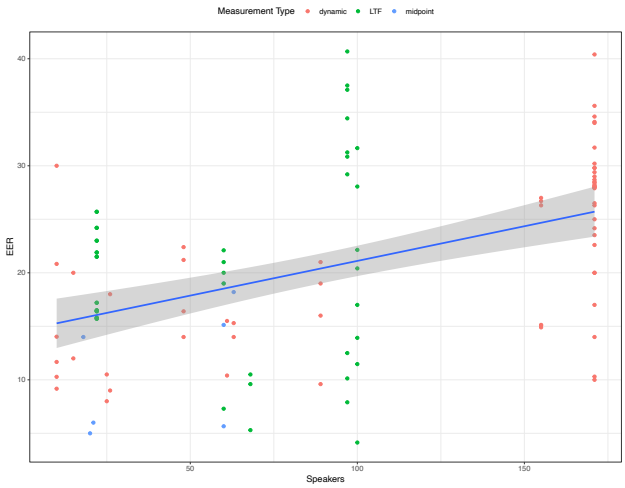


Figure 1: Number of speakers by EER results

Looking more closely, there is a weak tendency for studies which include a larger number of speakers to display higher EERs. This could be connected to the idea that a larger pool of speakers creates more room for different speakers to sound more similar to one another – therefore there is opportunity for more discrimination errors. As we would expect, the analysis of the CR results uncovered a significant effect for number of speakers ($p = 0.01$). In addition, however, linguistic variety also resulted in a significant effect ($p = 0.01$), and measurement type resulted in a near-significant effect ($p = 0.07$).

Figure 2 shows the results of the EER dataset. Studies which have measured the midpoint have the lowest median EERs and they all fall below 20% EER. Studies which have measured dynamic formant trajectories and those which have included LTF measurements display similar median EERs and similarly wide ranges (i.e., EERs ranging from below 10% to above 30%). Importantly, there is only a very low number of studies which have analysed midpoints and reported EERs. This might be the reason for the lower degree of variation in midpoint results compared to dynamic and LTF results.

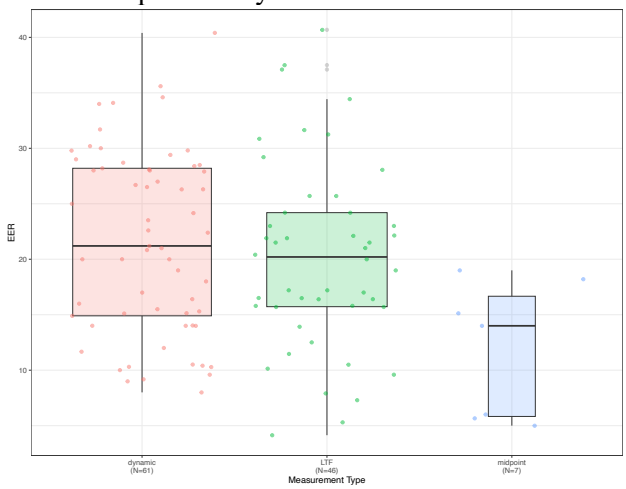


Figure 2: Measurement type by EER results

To further investigate these patterns, the results were visualised according to the specific formant frequencies analysed. This was in order to discover whether certain formants have greater speaker distinguishing power than others (see Figure 3).

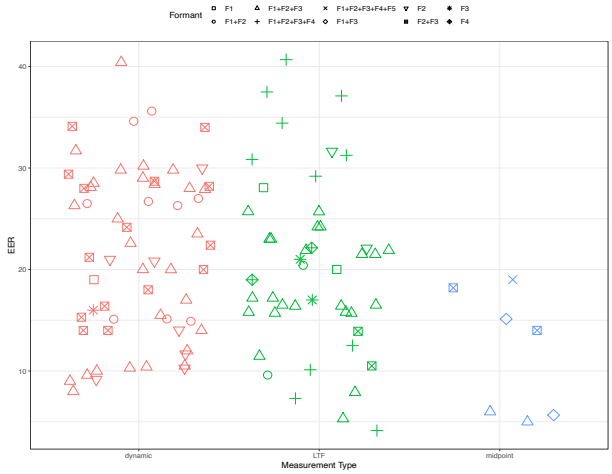


Figure 3: Formant frequencies and measurement type by EERs

There is no clear pattern in terms of specific formant frequencies and low EERs. However, it appears that adding more formants into the analysis does not show

a clear improvement. To illustrate, taking the largest number of formants in a measurement combination for the dynamic category, i.e., F1+F2+F3, this shows values ranging from the lowest EERs to the highest. Likewise, when focusing on the F1+F2+F3+F4 combination for the LTF category, we can see a similarly broad range of performance results. The midpoint category contained the largest measurement combination that occurred across all the studies (F1+F2+F3+F4+F5) - this resulted in the highest EER for the midpoint category. Overall, relatively little research has focused on the higher formants, i.e., F3, F4 and F5, in isolation. It is anticipated that some of the variability within the dynamic and midpoint measurement types may be connected to the specific vocalic segments that were in focus. In addition, differences in recording context are also likely to have contributed to the variability in performance across all studies. Further work would be needed to unpack these interactions.

Turning to the CR dataset, Figure 4 presents a plot of the number of speakers relative to the CR achieved. Obviously, the more speakers included in the discrimination task, the worse the CR result. More interesting, though, is the incidental relationship between measurement type and the number of speakers included in the "closed-set". This may explain the overall better classification rates of dynamic over midpoint measurements (and the near-significance of measurement type). To illustrate, a cluster of midpoint values at the 50-speaker mark has poor classification rates compared to the high CR rates of between 50 and 90% using dynamic measurements with <10 speakers.

Figure 4 also shows that even when the number of speakers remains stable – for instance all the midpoint results – there are high levels of variation in classification rates. This highlights the importance of understanding the influences on CRs other than the number of speakers, such as the specific formant frequencies measured, the vowel segments included and the recording contexts.

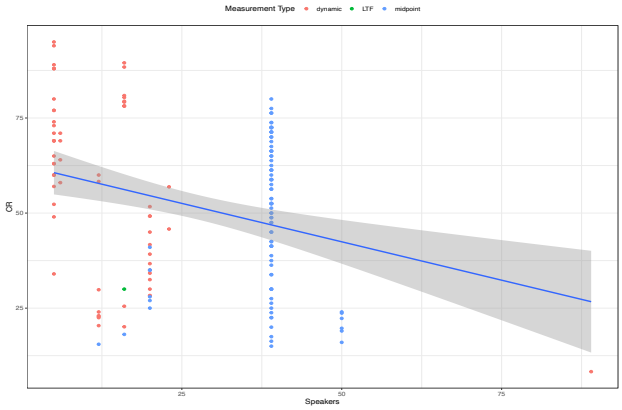


Figure 4: Number of speakers by classification rates

Having further investigated the significant effect of linguistic variety, it transpires that SSBE shows overall poorer classification rates compared to other varieties. However, this finding should not be overinterpreted in view of the low number of studies.

5. DISCUSSION

It might come as a surprise to learn that relatively few studies use midpoint formant measurements in the speaker discrimination research literature. Midpoint measurements seem to exhibit better performance overall than dynamic and LTF measurements, potentially as a result of the small number of studies that use midpoint measurements and report EERs. This overall low usage of midpoint measurements in the forensic speech science research literature seems to be at odds with the 97% of practitioners who reported to use midpoint measurements in forensic casework [1]. This discrepancy between research and practice further extends to the data itself, with the majority of the forensic speech science research using lab-based speech which is not representative of casework material. This meta-analysis reveals other trends that go against some of our expectations including the notion that adding to the number of formants is necessarily better for speaker discrimination, dynamic measurements are more fruitful than midpoint measurements, and that some formant analysis processes might be "better" than others. Additionally, there is little research to support the notion that higher formants are necessarily better speaker discriminators.

Overall, it seems to be difficult to characterise the performance of formants in speaker discrimination tasks, given the wide range of results presented in the forensic speech science research literature. It is acknowledged that as this is a meta-analysis, all of the studies had varying factors, some of which have been explored and commented on in this paper. Others that have not been commented on in this initial phase of the meta-analysis include the role played by specific vocalic segments. This naturally hinders the ability to directly compare results from all of the included papers. Even so, in view of the wide-ranging results in this meta-analysis and ongoing discussion regarding the obscurities around formants in phonetic research [9], it may be appropriate to review the importance placed on formants in FVC casework. This is not to say that formants should not be analysed in FVC casework at all. Rather, a measured approach to their inclusion should be taken.

6. ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council (ESRC) [grant number 2385933].

7. REFERENCES

[1] E. Gold and P. French, 'International Practices in Forensic Speaker Comparison', *International Journal of Speech, Language and the Law*, vol. 18, no. 2, pp. 293–307, Nov. 2011, doi: 10.1558/ijsll.v18i2.293.

[2] F. Nolan, 'The "telephone effect" on formants: a response', *International Journal of Speech, Language and the Law*, vol. 9, no. 1, pp. 74–82, Mar. 2002, doi: 10.1558/ijsll.v9i1.74.

[3] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. Wadsworth, Cengage Learning, 2011.

[4] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, 'Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material', *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1596–1607, Dec. 1968, doi: 10.1121/1.1911302.

[5] P. Rose, *Forensic Speaker Identification*, 1st ed. Taylor Francis, 2002.

[6] C. Byrne and P. Foulkes, 'The "Mobile Phone Effect" on vowel formants', *International Journal of Speech, Language and the Law*, vol. 11, no. 1, pp. 83–102, Mar. 2004, doi: 10.1558/ijsll.v11i1.83.

[7] K. McDougall, 'Speaker-characterising properties of formant dynamics: A case study', *Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne, December 2 to 5, 2002.*, pp. 403–408, 2002.

[8] F. Nolan and C. Grigoras, 'A case for formant analysis in forensic speaker identification', *International Journal of Speech, Language and the Law*, vol. 12, no. 2, pp. 143–173, Aug. 2005, doi: 10.1558/sll.2005.12.2.143.

[9] D. H. Whalen, W.R. Chen, C. H. Shadle, and S. A. Fulop, 'Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986)', *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 933–941, Aug. 2022, doi: 10.1121/10.0013410.

[10] T. Kendall and V. Fridland, *Sociophonetics*. Cambridge, UK: Cambridge University Press, 2021. doi: 10.1017/9781316809709.

[11] M. Jessen, 'Speaker profiling and forensic voice comparison: The auditory-acoustic approach', in *The Routledge Handbook of Forensic Linguistics*, 2nd ed., M. Coulthard, A. May, and R. Sousa-Silva, Eds. London: Routledge, 2020, pp. 382–399.

[12] M. Jessen, *MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis*. 2021, p. 415. doi: 10.21437/Interspeech.2021-1697.

[13] T. Cambier-Langeveld, 'Current methods in forensic speaker identification: Results of a collaborative exercise', *International Journal of Speech, Language and the Law*, vol. 14, no. 2, pp. 223–243, Mar. 2008, doi: 10.1558/ijsll.2007.14.2.223.

[14] R v O'Doherty [2002] NICA 20.

[15] M. Jessen, 'Forensic voice comparison', in *Handbook of Communication in the Legal Sphere*, J. Visconti, Ed. De Gruyter, 2018, pp. 169–200. doi: 10.1515/9781614514664-010.

[16] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, 'A Forensic Phonetic Study of "Dynamic" Sources of Variability in Speech: The DyViS Project', *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, Dec. 2006, [Online].

Some measures of phonetic similarity for use in legal trademark disputes

Sandra Ferrari Disner[1] & Vincent J. van Heuven[2, 3, 4]

[1] University of Southern California, Los Angeles, USA
[2] Leiden University Centre for Linguistics, Leiden, The Netherlands
[3] University of Pannonia, Veszprém, Hungary
[4] Fryske Akademy, Leeuwarden, The Netherlands
s.disner@usc.edu; v.j.j.p.van.heuven@hum.leidenuniv.nl

ABSTRACT

A trademark or brand name may be legally disputed if it sounds too similar to another trademark. US courts would benefit from simple and understandable measures of sound-alikeness that judges and lay jurors would be able to understand. We compared the usefulness of transitional probabilities (bigrams and trigrams in phonetic transcriptions) and various string edit measures (Levenshtein distance, length-normalised or not, feature-weighted or not) in two studies. In the first, we computed similarity and distance measures for all US trademark litigation cases we could find that were settled predominantly on the basis of sound similarity. We aimed to find optimal cut-off values to separate the names that were deemed too similar from those deemed sufficiently distinct and thus allowed to compete in the market. Our second study examined 120 pairs of generic drug names used in the USA that were deemed confusable by pharmaceutical experts. Again, length-normalized and feature-weighted Levenshtein distance proved the best predictor of confusability.

Keywords: trademark sound-alikes, shared n-grams, Levenshtein string edit distance, phonetic feature weighting, string length normalization.

1. INTRODUCTION

A trademark is a type of intellectual property consisting of a recognizable sign (e.g., logo), design (characteristic shape, size, color), and/or linguistic expression, which identifies and distinguishes products or services of a manufacturer or provider from those of potential competitors. It happens with some regularity that a new brand, e.g., *Americrest* (a mortgage lender) on the market semi-copies the trademark of an earlier, well-established product of the same type, e.g., *Ameriquest*, in an attempt to derive unearned benefits from the reputation of the 'senior' mark ('reputation parasitism'). Although such stealing of intellectual property is illegal, no *a priori* checks are mandatory for an entrepreneur who wants to penetrate the market with a new trademark. Any initiative for starting litigation, and the burden of proof, is on the owner of the senior mark, who must show to the satisfaction of a court of law that the junior mark bears a confusing similarity to the senior mark, and may be held accountable for (potential) loss of product turnover. In such cases, forensic phoneticians may be called upon as experts to argue that the sound shapes in a trademark dispute are (or are not) similar enough to cause confusion among consumers [1, 2, 3].

Generally, in such cases, the court does not allow litigants to run (field) experiments with human participants to determine the degree of auditory confusability between trademarks. The reason for this is that the confusability of competing brand names is strongly affected by the articulation and voice characteristics of the speaker, the auditory acuity and motivation of the listener, and on the noisiness of the communication channel. Rather, the court wants theoretically grounded, generally applicable reasoning to establish the degree of confusability that can be explained to (and understood by) a lay jury. In current judicial practice, the court's decision whether two contested trademarks are sufficiently distinct or sound so similar that auditory confusion may arise, are made on intuitive grounds, and are typically not based on systematic phonetic reasoning. But here we explore the possibility of more objective methods to determine auditory confusability in legal trademark disputes. We will do this in two background studies. However, before we present summaries of these studies, we will first explain techniques used in the literature, and by ourselves, to quantify the degree of similarity or difference between two sound shapes, be they existing words or nonce trademarks.

2. MEASURES OF SIMILARITY

Characteristic of all the measures we will survey in this paper is that they are not computed directly on some acoustic signal recorded from a human speaker. We abstract away from a specific human speaker by converting the sound shapes to a phonetic transcription, which may be broad (one symbol per phoneme) or somewhat narrow, marking predictable allophones with dedicated transcription symbols or diacritics. In our studies on American English trademarks so far,

we have used standardized broad transcriptions only, including the primary stress, which we treat on a par with the segmental phonemes.

Measures of similarity between strings of transcription symbols are typically based on the number of symbols shared by two words or names [1, 2, 3]. To preserve information on sequential order, measures of similarity are based on the number of n-grams shared by two sound shapes, where $n > 1$ [4]. In most studies the n-grams are either bigrams (sequences of two adjacent symbols) or trigrams (sequences of three adjacent symbols). A word boundary symbol '#' is inserted at the beginning and end of the transcription of a name/word, and is counted on a par with the other symbols, to preserve information about the edge status of the first and last sound of a word. The bigram and trigram measures are generally correlated but provide partly independent information about similarity between two strings of symbols. A general concern in expressing the degree of similarity (and difference) between two sound shapes is the length of the strings.

The longer the strings are, the better the chances of the same n-gram occurring in both strings. Rather than counting the absolute number of shared n-grams, therefore, we compute the percentage of shared n-grams. Any n-gram that occurs in both string A and string B, increments the count by 2. The sum of the shared n-grams is then divided by the total number of n-grams in A and B added together. Table 1 shows how to compute the percentage of shared bigrams and trigrams (Pbi, Ptri) for the brand names *Phexxi-Imvexxy* /f'ɛksii/-/ɪmv'ɛksii/, i.e., 67 and 63, respectively.

Table 1: Computation of shared n-grams.

	Bigrams			Trigrams		
	f 'ɛksii	ɪmv'ɛksii	*N*	f 'ɛksii	ɪmv'ɛksii	*N*
1.		#ɪ	0		#ɪm	0
2.		ɪm	0		ɪmv	0
3.	#f	mv	0	# f '	mv'	0
4.	f '	v'	0	f 'ɛ	v'ɛ	0
5.	'ɛ	'ɛ	2	'ɛk	'ɛk	2
6.	ɛk	ɛk	2	ɛks	ɛks	2
7.	ks	ks	2	ksi	ksi	2
8.	si	si	2	sii	sii	2
9.	ii	ii	2	ii#	ii#	2
10.	i#	i#	2			
	Σ shared bigrams		12	Σ shared trigrams		10
	Σ bigrams		18	Σ trigrams		16
	Pbi = (12/18) × 100		67	Ptri = (10/16) × 100		63

In our broad transcription, long/tense vowels are represented as geminates, including the vowels in *bad* /ææ/ and *hot* /ɑɑ/. Diphthongs in *find* /aɪ/, *found* /aʊ/ and *coin* /ɔɪ/, as well as the diphthongized tense vowels in *fame* /eɪ/ and *foam* /ou/ are transcribed and computationally treated as two vowels in sequence.

3. MEASURES OF DIFFERENCE

Differences between strings of symbols are based on string edit counts. The Levenshtein distance (LD) measure was first proposed by the eponymous Levenshtein [5], and later adopted in the field of computational dialectology as a convenient and valid measure of the distance between related varieties (dialects, accents) of a language [6, 7]. LD counts the number of string edit operations needed to convert a string of symbols A to its counterpart B. Possible string operations are: insertion, deletion and substitution of a symbol [8]. LD software is available on the internet [9, 10, 11]. Before strings can be compared, they have to be optimally aligned through dynamic programming. Vowel symbols are aligned with vowel symbols, consonant symbols with consonant symbols. Semivowels /j, w/ can be aligned with vowels as well as consonants. Similarly, the neutral vowel schwa can be aligned with any other vowel or with /r/. The LD algorithm minimizes the cost of the conversion by finding the optimal alignment of symbols and the least number of edit operations. In the binary (or 'plain') application of the algorithm, each edit incurs a penalty of 1 point. Some implementations of the algorithm allot .5 penalty to either an insertion or a deletion (also called 'indel') and 1 penalty point to a substitution. Operation on a diacritic in a narrow transcription incurs a cost of .5. Optional length normalization is achieved by expressing the percentage of the raw LD relative to the maximal cost that could be incurred by the two strings under comparison, where the max cost of a substitution equals 1, and of an indel .5. Table 2 exemplifies the computation of LD for the string pair /f'ɛksii/-/ɪmv'ɛksii/, assuming a broad transcription as input.

Table 2: Computation of raw and length-normalized Levenshtein Distance (LD).

			Binary (plain)		F-weighted	
	f 'ɛksii	ɪmv'ɛksii	cost	max	cost	max
1.		ɪ	.5	.5	.39	.5
2.		m	.5	.5	.50	.5
3.	f	v	1.0	1.0	.09	1.0
4.	'	'	0	1.0	0	1.0
5.	ɛ	ɛ	0	1.0	0	1.0
6.	k	k	0	1.0	0	1.0
7.	s	s	0	1.0	0	1.0
8.	i	i	0	1.0	0	1.0
9.	i	i	0	1.0	0	1.0
	Raw LD		2.0	8.0	.98	8.0
	Norm. LD (%)			25		12

It has been objected that the binary LD is too crude a measure, as the substitution of /f/ for /v/ raises the cost no higher than the substitution of, e.g., /t/ for /w/. Though no improvement could be demonstrated [6], judge and jury will be more easily convinced of the validity of the LD measure when the costs are weighted for the number of distinctive features that differ between the two sounds involved in the substitution. The LD algorithm we use optionally determines the difference between two vowels, or two consonants, according to the number of phonetic features that have to be changed to convert one segment to the other. It uses the feature system proposed by [12], and adapted for use in the LD computation by [13]. Changing /f/ to /v/ affects one consonant feature, [voice], which incurs a small cost of only .09 (see Table 2 under F-weighted, i.e., feature weighted). For reasons of space, we do not include the weighted cost matrices for vowels and consonants but will make these available in the supplementary materials. We also refer to the help files included in the LED-A software [11, 14].

4. STUDY 1: PREDICTING COURT DECISIONS

In our first study we examined the first 200 trademark cases in the USA listed in the Lexis-Nexis legal database.[1] These are the most frequently cited cases involving likelihood of confusion in the past 25 years. We kept only those that involved a dominant role of auditory confusion of the product names in the court's decision. In a second round, we searched the entire database with the conjoint terms: 'lapp test' & 'sound' & 'likelihood of confusion', which narrowed the cases down to those in which the court utilized a well-known test for the existence of likelihood of confusion based on the degree of sound similarity. The search yielded a set of 51 pairs of contested trademarks for which a court decision was known. Thirty-four contested pairs were judged too similar to compete on the market, while the remaining 17 pairs were judged sufficiently different to avoid consumer confusion.

We computed four similarity/distance measures between each of the trademark pairs as explained above. Table 3 shows the (Pearson) correlation matrix for the four measures.

Table 3: Correlation matrix for four similarity/distance measures ($p < .001$ in all cases).

	Pbi	Ptri	pLD
Shared bigrams (%)			
Shared trigrams (%)	.942		
Plain Levenshtein Distance	−.900	−.908	
Feature-weighted wLD	−.871	−.858	.920

Given the high intercorrelations, we deleted one similarity and one distance variable, i.e., those that were least successful in distinguishing pairs that had been judged too similar from those that had been judged sufficiently different. The two remaining variables, Pbi and wLD, were then entered as predictors in a Linear Discriminant Analysis (LDA [15]) distinguishing the two types of pairs. We ran the LDA three times, i.e., once with Pbi as a single predictor, once with wLD as a single predictor, and a third time with both predictors combined. Results are shown in Table 4.

Table 4: Results of LDA. *N* correct classifications in green cells.

Predictor(s)	Court's decision (down)	Predicted		Total
		TS	SD	
Pbi	Too similar (TS)	24	10	34
	Sufficiently different (SD)	3	14	17
	76.5% correct, TS ≥ 61.0% Pbi			
wLD	Too similar (TS)	28	6	34
	Sufficiently different (SD)	4	13	17
	80.4% correct, TS ≤ 14.5% wLD			
Pbi+wLD	Too similar (TS)	29	5	34
	Sufficiently different (SD)	4	13	17
	82.4% correctly classified boundary: 1.251×z(wLD) + .322×z(Pbi) = 0			

In Figure 1, we illustrate the contribution of wLD and Pbi to the correct separation between 'too similar' vs 'sufficiently different'.

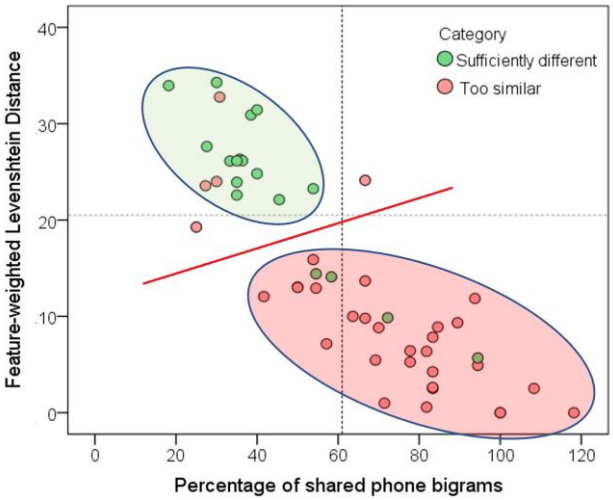


Figure 1: Separation of contested pairs of trademarks by wLD and Pbi together. Ellipses and boundary drawn by hand.

Some of the error decisions (red dots in green area, or green dots in red area) seem to be the result of the court's decision having been based on non-phonetic grounds after all (see [16: slide 30] for examples and discussion). Generally, the critical values for Pbi and

wLD work very well to distinguish 'too similar' and 'sufficiently different'. We ran an external validation on these parameters by computing the distribution of Pbi, Ptri, pLD and wLD for all possible pairs of English mono-morphemic words taken from the 3000 most frequent lexemes according to the British National Corpus but using the American English phonemic transcriptions in the CMU digital pronouncing dictionary [17]. In the 4,305,656 unique word pairs, a Pbi ≤ 61% is seen in fewer than .03% of the pairs. A wLD < 14.5% is found in 1.2% of the pairs. The conclusion follows that trademarks have to be exceptionally similar before they will be banned by a US court decision.

5. STUDY 2: PREDICTING CONFUSABLE DRUG NAMES

In our second study we did not attempt to predict (or rather postdict) the court's decision on competing trademarks; rather, we addressed the issue of consumer confusion more directly, by looking at confusable (generic) pharmaceutical drug names.[2] About 25% of all medication errors can be ascribed to some confusion of drug names [18], which may cause the issuing of a wrong prescription because of incorrect recognition of a drug name (either by eye or by ear). We were given a list of 1,250 pairs of drug names that can be obtained in the USA, which were deemed confusingly similar by a panel of pharmacologists.[3] We intended to use these pairs to see how confusable pairs could be automatically identified, using the parameters we found in the previous study.

The first problem we had to solve was to establish the official pronunciation of the drug names concerned. Here we should distinguish between branded drug names and generic drug names. The pronunciation of branded drug names is up to the manufacturer, and cannot easily be determined. The pronunciation of generic drug names, however, is laid down by the United States Adopted Name (USAN) Council, and can be obtained online from the USP dictionary of USAN and International Drug Names.[4] The pronunciation is specified in a laymen's phonetic transcription, which includes both primary and secondary stresses (sometimes debatable) and uses orthographic letter combinations that can only be pronounced in the way the USAN council wants the names to sound. We selected only the generic drug names from the list of 1,250 pairs, and within this subset we eliminated all compound names, i.e., drug names with elements separated by spaces, hyphens, numbers or other non-letter characters. Next, we checked in the USP dictionary if a pronunciation was listed (in which case we were certain that the name was indeed generic), copied the transcription and converted it (automatically) to the IPA transcription required by the LD software. This selection yielded 120 LASA (Look-Alike, Sound-Alike) pairs of generic drug names. The same name may occur in multiple pairs, e.g., as in *Oxazepam-Quazipam* and *Oxazepam-Oxaprozin*.

Apart from the test set of 120 LASA pairs we created a control set by generating all non-implicated pairs of these 120 drug names, i.e., a set of 2,428 non-confusable pairs. The total dataset then comprised 120 + 2,428 pairs of drug names, for which we computed 14 similarity/distance measures. Eight addressed phonetic similarity/distance:

Nbi	(raw)	Pbi	(length normalized)
Ntri	(raw)	Ptri	(length normalized)
pLD_r	(raw)	pLD_n	(length normalized)
wLD_r	(raw)	wLD_n	(length normalized)

The other measures were the same as the above but computed on the orthographic forms (printed names); however, no orthographic measures for weighted letter shapes were defined.

The 14 measures were used to classify the 2,548 pairs of drug names into those that were held to be confusable, and those that were non-confusable controls. Each measure was tested once separately, while the most successful predictors were also tested in selected combinations. Single predictors successfully separated the two categories between 83.8 and 86.4% correct for phonetic parameters, and even better for orthography-based parameters, i.e., between 82.5 and 92.3% correct. The best combination of phonetic predictors was (as before) length-normalized feature-weighted LD plus the percentage of shared bigrams Pbi (87.5% correct classification). Length-normalized predictors were always more successful than their raw (non-normalized) counterparts. Feature-weighted LD (raw as well as length normalized) proved a better predictor of confusability than the plain (binary) counterpart measures.

6. CONCLUDING REMARKS

In this paper we have tried to demonstrate that intuitive (court) decisions on similarity and confusability of consumer products can be predicted with high accuracy by relatively simple and straightforward linguistic-phonetic procedures that a layman (judge or member of a jury) would understand. Our measures and boundary values can be profitably used to signal potential consumer confusion, so that manufacturers may pre-empt legal procedures and choose names that are on the safe side of the boundaries.

More research is needed to separate look-alike from sound-alike confusability. This will require selection of critical name pairs in which sound-alikeness does not correlate with look-alikeness.

7. REFERENCES

[1] Butters, R. R. 2010. Trademark linguistics. Trademarks: Language that one owns. In: Coulthard, M., Johnson, A. (eds.), *Routledge handbook of Forensic Linguistics*. Routledge, 351–364. Doi: 10.4324/9780203855607

[2] Shuy, R. 2002. *Linguistic battles in trademark disputes*. Palgrave Macmillan. Doi: 10.1057/9780230554757

[3] Shuy, R. 2012. Using Linguistics in trademark cases. In: Solan, L., Tiersma, P. (eds.), *Oxford handbook on Language and Law*. Oxford University Press, 449–462. Doi: 10.1093/oxfordhb/9780199572120.013.0033

[4] Stephen, G. A. 1994. *String searching algorithms*. World Scientific.

[5] Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Sovjet Physics Doklady* 10, 707–710.

[6] Heeringa, W. J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Center for Language and Cognition Groningen. https://pure.rug.nl/ws/portalfiles/portal/9800656/thesis.pdf.

[7] Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, 60–67

[8] Kruskal, J. B. 1999. An overview of sequence comparison. In: Sankoff, D., Kruskal, J. (eds.), *Time warps, string edits, and macromolecules. The theory and practice of sequence comparison*. Stanford Center for the Study of Language and Information, 1–44.

[9] Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., Leinonen, T. 2011. Gabmap – a web application for dialectology. *Dialectologia: revista electrónica*, 65-89. https://www.clarin.nl/sites/default/files/restore/Gabmap-long-2011-jan-07-rev-mei.pdf

[10] Leinonen, T., Çöltekin, Ç., Nerbonne, J. 2016. Using Gabmap. *Lingua*, 178, 71–83. Doi: 10.1016/j.lingua.2015.02.004

[11] Heeringa, W. J. 2021. Levenshtein Edit Distance App (LED-A). Computer program. https://fryske-akademy.nl/fa-apps/led-a/#run

[12] Almeida, A., Braun A. 1986. "Richtig" und "falsch" in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschr. Dialektol. Ling.* 53, 158–172. https://www.jstor.org/stable/40502947

[13] Heeringa, W. J., Braun, A. 2003. The use of the Almeida-Braun system in the measurement of Dutch dialect distances. *Comp. Hum.* 37, 257–271. http://wjheeringa.nl/papers/cath02a.pdf

[14] Heeringa, W. J., van Heuven, V. J., Van de Velde, H. 2022. LED-A: a web app for measuring distances in the sound components among local dialects. Poster presented at the 17th New Methods in Dialectology Conference, Mainz. https://methodsxvii.uni-mainz.de/files/2022/06/

[15] Klecka, W. R. 1980. *Discriminant analysis*. Sage.

[16] van Heuven, V. J., Disner, S. F., Heeringa, W. J. 2021. What's in a name? On the phonetics of trademark infringement. Plenary talk presented at the 29th Annual Conference of the IAFPA, Marburg. Doi: 10.13140/RG.2.2.17529.95846

[17] The Carnegie Mellon University Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[18] Lambert, B. L. 1997. Predicting look-alike and sound-alike medication errors. *Am. J. Health-System Pharm.* 54, 1161–1171. Doi: 10.1093/ajhp/ 54.10.1161

[19] van Heuven, V. J., Disner, S. F. 2022. Utility of length-normalization for predicting confusion of generic drug names from Levenshtein string edit distance. Plenary talk presented at the 30th Annual Conference of the IAFPA, Prague. Doi: 10.13140/RG.2.2.30951.73125

[1] This section is based on a plenary talk [16] at the 29th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) in Marburg. The PowerPoint slides we presented there can be downloaded (see reference section).

[2] This section is based on a plenary talk [19] we presented at the 30th IAFPA Annual Conference 2022 in Prague. The slides can be downloaded (see reference section).

[3] We are most grateful to Professor Bruce Lambert, director of the Center for Communication and Health at Northwestern University (Evanston, IL), for making his (updated) list of 1,250 confused medication brand names available to us.

[4] https://www.usp.org/products/usp-dictionary

VOICE COMPARISON ANALYSIS OF FORENSIC RECORDINGS USING THE VOICESAUCE PROGRAM

Michael Jessen, Canan Konrat and Jonathan Horn

Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{michael.jessen | canan.konrat | jonathan.horn}@bka.bund.de

ABSTRACT

The computer program VoiceSauce was applied to a set of authentic telephone-based forensic recordings in order to examine the speaker-discriminatory potential of the acoustic voice quality information extracted with the program. Several conditions were tested, including variation on the input audio (all-vocalic vs. /a/-based) and the option of uploading hand-corrected formant and f0 streams. Among the voice quality parameters investigated, CPP (cepstral peak prominence) turned out to be the most speaker-distinguishing one, yet it stayed behind formant frequencies when used as baseline parameter. Judged from the performance indices EER and Cllr, speaker discrimination for the forensic data analysed with VoiceSauce remained quite limited. Multivariate models with several voice quality parameters, as well as logistic regression fusion, did not improve performance over the use of single parameters.

Keywords: forensic voice comparison, laryngeal voice quality, VoiceSauce

1. INTRODUCTION

Voice quality is a feature complex widely used in forensic phonetics [1]. Most commonly, forensic voice quality is judged auditorily [2, 3]. The focus of this paper is on laryngeal voice quality, also known as phonation type, as opposed to supralaryngeal voice quality also referred to as articulatory settings; henceforth "voice quality" will be used in the narrow sense of phonation type. Due to limitations in the intra- and interrater reliability of auditory voice quality ratings [4], being able to measure voice quality acoustically would be an advantage. There has been scepticism, however, that most acoustic voice quality measurements might not very well survive the telephone-passband or other forms of signal degradation typical in forensics [2]. Confirming such reservations, Enzinger et al. [5], using software called Glottex, found a speaker discrimination reduction in telephone-transmitted speech as compared to high-quality speech. However, Hughes et al. [6], using VoiceSauce [7], presented more mixed results, and especially with harmonic amplitude parameters there was little difference between studio and telephone quality, even when signal quality was further reduced within the latter category.

The main goal of this study is to explore further to which extent the program VoiceSauce can provide speaker-discriminatory information when applied to forensically realistic recordings. Though the focus of this study is on this program, there is an interest in forensic-phonetic acoustic voice quality research more generally.

2. METHOD

2.1. Forensic recordings

This study is based on a speech corpus called GFS 2.0 (German Forensic Speech). The corpus contains two recordings each of 23 male adult speakers of German. These recordings derive from anonymized authentic forensic cases involving telephone conversations (converted into PCM wav files sampled at 16 bit, 8 kHz). Single recordings from additional 25 speakers were used for the UBM (Universal Background Model), addressed in section 2.4. This corpus was also used in [8, 9]; the specific version applied is the one of [9], in which one of the 23 speakers was not included due to difficulties in measuring formant structure. Signal quality is on average lower than in non-forensic telephone recordings. In terms speech style, sections of speech with increased vocal effort or emotional involvement are included, but severe levels of signal distortion or non-neutral speech style were avoided in the compilation of the corpus. Net duration of the recordings is between about 20 and 60 seconds.

2.2. Audio data preparation

The measurements with VoiceSauce are based on vowel-only speech material. In some of the voice comparison tests (test condition design summarized in Table 1) the material used was the same as in [9], where for the purpose of long-term formant analysis (LTF) each recording was cut down manually to a 10-seconds long stream that contained only vowels with visible F1, F2 and F3. This condition will be called "LTF".

In other tests only the vowel /a/ was included in the stream. Depending on the number and lengths of

a-tokens in the recordings, the duration of the a-stream differs between recordings from between about 1.5 to 4 seconds. Tokens of /a/, regardless of phonological context, were identified based on listening from roughly sentence-level context, but when listening to the selected tokens in isolation, they had a broader range of mid-to-low vowel qualities. This condition is referred to as "a-tokens".

In both the LTF- and the a-token-data, cutting passages out of the recordings resulted in abrupt transitions between the remaining segments. In the spectrogram, these cuts may be visible as very short transients across the frequency range. During the brief moments of these transients, spectral tilt is radically reduced. Since spectral tilt is among the parameters measured with VoiceSauce, results could be affected negatively by the existence of these transients. To investigate that possibility, a subset of the a-token data was processed in a way that hard cuts were eliminated. This was performed by writing a Praat script that carried out the following steps for each of the a-tokens audio files: The a-tokens had been labelled in Praat. For each of the labelled tokens per audio file, the first zero crossing was identified and the signal from the beginning of the labelled segment to that zero crossing was muted. The same procedure was applied in the right-to-left direction of the segment. Subsequently all the segments per file were concatenated. The condition using data processed in this manner is called "a-tokens-zero".

2.3. VoiceSauce analysis

The program VoiceSauce (accessed August 8, 2022 from www.phonetics.ucla.edu/voicesauce) with compiled Matlab executables for Windows was used to analyse the audio data described in 2.2. Out of the voice source parameters offered by the program, the following set was used for analysis (see the program manual for their definition): H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-H2K*, H2K*-H5K, CPP, HNR05, HNR15, HNR25, HNR35. For reasons explained below, the parameter H2K* was added as well. VoiceSauce requires good estimation of formant frequencies and fundamental frequency in order to find the spectral events necessary for measurement of the voice source parameters. The Snack algorithm was selected for formant analysis and Praat for f0-analysis. Snack was selected due to compatibility with Wavesurfer [10], used in the LTF study in [9]. Praat was preferred over other f0-algorithms based on relative success (frequency and severity of f0 outlier values as criterion) on a subset of the data. Parameter defaults were used, except for frame shift, which was set to ten milliseconds instead of one, and Max F0 in Praat, which was set to 250 Hz.

Instead of automatic f0 and formant estimation, VoiceSauce also offers the option of manual data input. This option was used for some of the voice comparison test conditions. It allows the user to upload files containing hand-measured or manually corrected formant- or f0-value streams into the software, and in that case the uploaded values will be used for subsequent voice quality parameter extraction instead of the automatically measured f0 and formant values. In some test conditions only formants were uploaded. The formant values were taken from the formant measurements of [9] performed with Wavesurfer. Alignment across VoiceSauce parameters was ensured by taking the automatically calculated formant values and temporally aligning them with the uploaded formant values, after which only the uploaded values were used in the analysis. Maximum alignment was achieved by using a value of 2 for the data offset parameter in VoiceSauce. F1 was rarely in need of correction, therefore formant upload was limited to F2 and F3. In another test condition both hand-corrected formants and f0-values were uploaded. Uploaded f0-values were measured and corrected in Wavesurfer. Alignment was again achieved with a data offset value of 2. Pre-analysis showed that VoiceSauce-measured formants (F2 and F3) were more error prone with the data used here than f0. This is the reason why only in one test condition f0 data were uploaded in addition to formant data.

Table 1 shows the test conditions of this study. The term test means that questioned-speaker recordings from 22 speakers are compared against suspect recordings from the same 22 speakers, yielding 462 different-speaker comparisons and 22 same-speaker comparisons. Within a test condition the test is repeated several times while varying the voice quality parameter and the use or non-use of MAP, as well as two values for the number of Gaussians, all explained in 2.4.

Cond.	Audio	Upload (manual data input)
1	LTF	-
2	LTF	F2, F3
3	LTF	F2, F3, f0
4	a-tokens	-
5	a-tokens	F2, F3
6	a-tokens-zero	-

Table 1: Test conditions used in this study.

2.4. Forensic-phonetic analysis

For each test as defined in 2.3, each of the 484 comparisons resulted in a score by using the GMM-

UBM approach. This approach was originally developed for automatic speaker recognition [11] and was adapted to the analysis of acoustic-phonetic data by Becker et al. [12]. In the current study the software VOCALISE version 1.6 [13] was used. The software accepts the frame-by-frame output generated by VoiceSauce for each of the voice quality parameters. This stream of values is modelled as a Gaussian Mixture Model (GMM) for the suspect recording and the 25-speaker set representing the UBM. The questioned-speaker values are not modelled in that manner, but instead they are evaluated frame-by-frame against the GMM of the suspect and the UBM, technically resulting in a Likelihood Ratio. These results are averaged across the frames. Further details about GMM-UBM analysis as applied here are the same as in [9].

In the tests, one of the parameters varied is the number of Gaussians for the GMM. In [9] the value of 17 was used. In pre-analyses it was tested whether this is a viable value for the current data. This value was accepted and a lower value of 3 was selected as well, because analysis of acoustic-phonetic data, formants in particular, tend to saturate at lower values of Gaussians as far as speaker-discrimination performance is concerned [12, 14]. The other parameter varied is the inclusion or exclusion of MAP (Maximum a posteriori) adaptation [9, 11]. One of the goals of MAP adaptation is to provide a more stable suspect model in case of insufficient suspect data. This might be relevant here with the a-tokens in particular.

For each individual voice quality parameter speaker discrimination performance was measured with EER (Equal Error Rate, using the convex hull method) and Cllr (log-likelihood-ratio cost) [15]. For the comparisons reported with Cllr, prior calibration was performed using cross validation logistic regression calibration [16].

Further analyses were made by grouping several parameters. In combination tests, the parameters of the group were included in a multivariate model, which is a regular option with the GMM-UBM approach. In fusion tests, the parameters were fused using logistic regression fusion [16]. Three sets of parameters were grouped for this purpose. The first group consists of all parameters using relative harmonic amplitudes (except, H2K*-H5K, which, as will be explained, is replaced by H2K*), the second one consists of those targeting spectral noise (CPP and HNR), and the third group consists of all parameters contained in the previous two groups. For the combination results Cllr is reported after calibration, for the fusion results Cllr is reported directly because fusion has a calibrating effect. For the functions described in this paragraph and the preceding one, the software BIO-METRICS was used [17].

3. RESULTS

Results are reported in two ways. First, numerical detail is presented for Condition 2. Secondly, in a more descriptive manner, patterns are pointed out that are similar or different between Condition 2 and the other five ones. A full report of all numerical results across conditions is made available at www.researchgate.net/profile/Michael-Jessen-2.

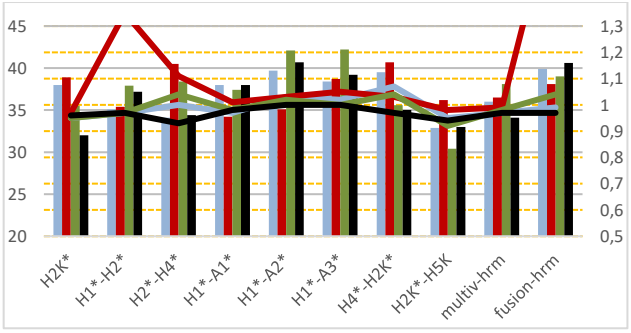


Figure 1: EER (bars; left vertical axis) and Cllr (lines; right) in Condition 2 for harmonic amplitude parameters.

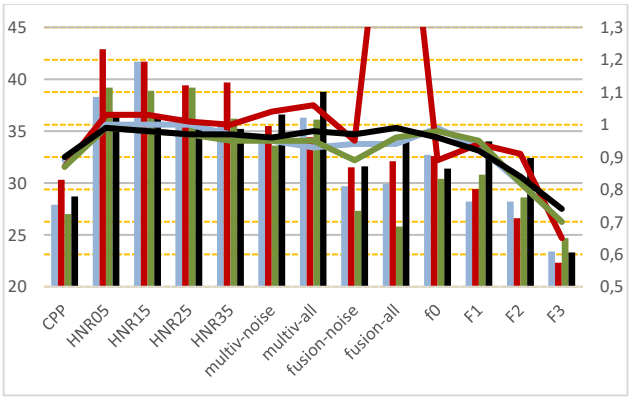


Figure 2: EER and Cllr in Condition 2 for spectral noise parameters, combination and fusion across all voice quality parameters, f0 and formant frequencies F1, F2 and F3.

For test condition 2 (see Table 1), Figs. 1 and 2 show the EER results (bars) and the Cllr results (lines) in the following four test variations (bars from left to right; same colours/greyscales for lines): no MAP, 3 Gaussians; no MAP, 17 Gaussians; MAP, 3 Gaussians; MAP, 17 Gaussians. Lower values indicate better performance. EER (in percent) of 50 and Cllr of 1 are at chance level. Looking at EER, the four test variations have an impact on the results, but there are no systematic performance differences. There are some differences between the voice quality parameters however. Among the harmonic amplitude parameters (Fig. 1), the best performance of single

parameters is with H2K*-H5K. Paradoxically, this is a parameter that should bear little meaning since the sampling rate of the recordings was 8 kHz and therefore 5 kHz is not represented. Could this result be driven by the component H2K*? In order to investigate this question, this parameter was included. Fig. 2 indeed shows that H2K*, although it is not a relative amplitude measurement (subtraction), is among the better-performing parameters, but it is not as good as H2K*-H5K (although in Conds. 4 and 6 it is slightly better). This has to remain an open issue. Applying multivariate processing or fusion to all harmonic amplitude parameters (excluding H2K*-H5K due to its questioned validity) leads to no particular performance advantage over single parameters. Among the spectral noise parameters (included in Fig. 2), a striking result is that CPP has the best performance. It is surpassed only by F3, which along with f0, F1 and F2 was included for comparison purposes with more familiar forensic-phonetic parameters as baseline. Fusing spectral noise parameters or applying a multivariate model did not improve upon CPP. The same holds for fusion and multivariate models across all parameters (but H2K*-H5K), except for fusion in one of the four variations. That particular fusion advantage, however, is not paralleled by Cllr. There was also an exception with Cond. 3, where noise-based or overall fusion slightly improved upon CPP.

Turning to the Cllr results, they roughly mirror the EER results. For example, Cllr is relatively low with CPP and still lower with F3, just like EER. There are some differences in detail however. In a few cases Cllr is massively out of scale. This is due to outliers in the scores, which Cllr is sensitive to but EER is not. The outliers occur in one of the two variations that make no use of MAP. MAP has the effect of preventing such extreme scores. Outlier values that occur without MAP were further negatively enhanced by cross validation calibration.

When comparing the voice quality results from Cond. 2 to the results from the other conditions mentioned in Table 1 (not shown, but provided through link), no major differences occur. It is not the case that single methodological steps such as uploading formants as opposed to fully automatic processing or removal of transients left any clear marks on the results. Most remarkable is the lack of any major difference between the conditions using 10-seconds LTF data and those using much shorter /a/-vowel data. Although there is a tendency for results to be better in the LTF data, this effect is quite small. Absence of a clear difference could mean that the extraction of the relevant information can be achieved with very short vocalic material, or it could mean that there is a trading relation in which the advantage for voice quality analysis of limiting the analysed material to the mid-to-low range of the vowel space [cf. 18, 19] can be traded against the advantage of increased duration of the vocalic input. Interestingly, formants (F2 and F3) suffer more in performance when switching from LTF to a-vowels than voice quality (shown in the link-provided data).

4. DISCUSSION

Speaker discrimination studies with VoiceSauce or related methods have been conducted predominantly on the basis of microphone-quality speech. Examples of such studies, with promising results, include [20-23]. There is less work on telephone-based or otherwise signal-degraded speech. A previous study that is of particular relevance to the present research is Hughes et al. [6].

VoiceSauce measurements on the forensic speech corpus GFS 2.0 have resulted in speaker discrimination performance that is quite limited. EER most of the time turned out to be in the range of 30 to 40% and only the parameter CPP showed better performance, reaching down into the 25-30% range. This is close to the performance of some formants based on the same material, though F3 has better performance still. Cllr was never below 0.8 even with CPP, and sometimes it was around chance level. Hughes et al. [6] in their telephone conditions differing in landline/mobile status and quality reported better values. For their multivariate models that correspond closely to “multiv-noise” and “multiv-hrm” in Figs. 1 and 2, they on average reported EER of about 20% and Cllr of between 0.7 and 0.9 for the former and for the latter EER of about 15% and Cllr slightly below 0.6. This is better performance than found here, especially for the harmonic amplitude parameters, which is contrary to the present results, where additive noise was better-performing than harmonic amplitude (probably CPP dominated in the multivariate model, so it really narrows down to that single additive noise parameter). Possible reasons for the performance difference between [6] and this study, respectively, include: telephone-transmission mostly simulated from microphone data vs. original telephone-recordings; fixed vs. varying signal quality across recordings; 60 seconds vs. 10 seconds vowels or less (a-tokens); contemporaneous speech vs. non-contemporaneous speech.

With more in-depth scrutiny of the extraction process it could be investigated further whether the limitations observed in this study are intrinsic to the forensic audio material or whether they result from difficulties of the algorithms to detect the relevant spectral events.

5. REFERENCES

[1] Gold, E., French, P. 2011. International practices in forensic speaker comparison. *Int. J. Speech, Lang. Law* 18, 293–307.

[2] Nolan, F. 2005. Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle, W. J., Mackenzie Beck, J. (eds), *A Figure of Speech. A Festschrift for John Laver*. Lawrence Erlbaum Associates, 385–411.

[3] San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., Kavanagh, C. 2019. The use of the vocal profile analysis for speaker characterization: Methodological proposals. *J. Int. Phon. Assoc.* 49, 353– 380.

[4] Kreiman, J., Gerratt, B. R., Ito, M. 2007. When and why listeners disagree in voice quality assessment tasks. *J. Acoust. Soc. Am.* 122, 2354–2364.

[5] Enzinger, E., Zhang, C., Morrison, G. S. 2012. Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proc. Odyssey* Singapore, 78–85.

[6] Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., Gully, A. J. 2019. Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. *Proc. 19th ICPhS* Melbourne, 1455–1459.

[7] VOICESAUCE software. http://www.phonetics.ucla.edu/voicesauce/

[8] Solewicz, Y.A., Jessen, M., van der Vloed, D. 2017. Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proc. Interspeech* Stockholm, 2849–2853.

[9] Jessen, M. 2021. MAP adaptation characteristics in forensic long-term formant analysis. *Proc. Interspeech* Brno, 411–415.

[10] WAVESURFER software. https://www.speech.kth.se/wavesurfer/index2.html

[11] Reynolds, D., Quatieri, T., Dunn, R. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19–41.

[12] Becker, T., Jessen, M., Grigoras, C. 2008. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech* Brisbane, 1505–1508.

[13] VOCALISE software. https://oxfordwaveresearch.com/products/vocalise/

[14] Jessen, M., Alexander, A., Forth, O. 2014. Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proc. Audio Engineering Soc. 54th Int. Conf. Audio Forensics* London, 28–35.

[15] Van Leeuwen, D. A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed), *Speaker Classification I: Fundamentals, Features, and Methods*. Springer, 330–353.

[16] Morrison, G. S. 2013. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Aust. J. Forensic Sci.* 45, 173–197.

[17] BIO-METRICS softw. https://oxfordwaveresearch.com/products/bio-metrics/

[18] Hanson, H. M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.* 101, 466–481.

[19] Hanson, H. M., Chuang, E. S. 1999. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* 106, 1064–1077.

[20] Jessen, M. 1997. Speaker-specific information in voice quality parameters. *Forensic Linguistics* 4, 84–103.

[21] Vaňková, J., Skarnitzl, R. 2014. Within- and between-speaker variability of parameters expressing short-term voice quality. *Proc. Speech Prosody* 7, 1081–1085.

[22] Kreiman, J., Park, S. J., Keating P. A., Alwan, A. 2015. The relationship between acoustic and perceived intraspeaker variability in voice quality. *Proc. Interspeech* Dresden, 2357–2360.

[23] Lee, Y., Keating, P. A., Kreiman, J. 2019. Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am.* 146, 1568–1579.

EXPLORATION OF SIMPLE METHODS OF LIKELIHOOD RATIO CALCULATION IN FORENSIC VOICE COMPARISON

Michael Jessen, Almut Braun and Sonja Menges

Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{michael.jessen | almut.braun | sonja.menges}@bka.bund.de

ABSTRACT

Two multi-speaker datasets were used, one of them from a laboratory corpus, one from forensic casework. Long-term formants were analysed, both with the maximally performing GMM-UBM approach and with simpler methods for likelihood ratio (LR) calculation. Results show that for the casework data, though not the laboratory data, GMM-UBM performance could be approximated with the simple methods. Reasons for the difference probably lie in signal quality and speech style differences which are expected to affect the complexity of the long-term formant distributions. Results support the view that simple methods of LR calculation can improve the process of estimating LRs, which is a widely practiced method in forensic science, they can be a point of departure for the development of more complex methods, or an educational device.

Keywords: likelihood ratio, forensic voice comparison, long-term formants

1. INTRODUCTION

The view that conclusions of forensic science casework should be expressed in terms of likelihood ratios (LR) has been gradually gaining acceptance internationally [1, 2]. Methods of calculating LRs are found across the forensic sciences with more progress in some than others so far. Forensic voice comparison is actually one of the very earliest forensic sciences having promoted and developed useful concepts and tools in that regard [3, 4, 5]. Methods of LR calculation are usually complex. In voice comparison, the most complex ones are from the domain of automatic speaker recognition [6]. Fairly complex also are methods used in semiautomatic speaker recognition, i.e. the application of manually performed or supervised methods of (most commonly) acoustic phonetics to voice comparison [7]. The two best known of such methods are the Multivariate Kernel Density (MVKD) approach and the Gaussian Mixture Model-Universal Background Model (GMM-UBM) approach. The former was developed for the purpose of various forensic sciences, most specifically glass evidence [8], and initially adapted to semiautomatic speaker recognition by Rose [9]. The latter was developed in automatic speaker recognition [10] and was initially adapted to semiautomatic speaker recognition by Becker et al. [11]. The MVKD approach is best suited for token-based data (e.g. twelve tokens of /a/ per recording with one set of formant measurements each) and the GMM-UBM is best suited for stream-based data (e.g. measurements of formants every 10 milliseconds scanning through the vocalic parts of a recording). Since the data analysed in this study are semiautomatic and stream-based, the GMM-UBM approach will be used to establish the top-line indicating optimum performance.

The goal of this study is to compare the performance of one of these high-end methods to the performance characteristics of simpler methods of LR calculation. There are several reasons why we think this is a worthwhile endeavour. Firstly, estimation of LRs in situations where insufficient data are available to calculate them is an option that is explicitly included in current guidelines for forensic analysis [12]. The simpler the methods the better the analyst can understand them and use them as conceptual tools in the estimation process, thereby making the estimations more concept-driven and less intuition-driven. Secondly, simple methods can form the beginning of the development of more complex ones in new areas of the voice comparison feature space or in other forensic sciences with little exposure so far to the LR framework. Thirdly, looking at simple methods can have an educational aspect. One can observe how performance can gradually change by adding more complexity to the methods.

2. METHODS

Two datasets were used as the basis for the experiments in this paper. The first one is called GFS 2.0 (German Forensic Speech). It contains anonymized speech from forensic casework involving telephone interception data. There are data from 22 male adult speakers of German, each with one questioned speaker recording and one suspect recording. There are additional single recordings from 25 adult male speakers recorded under the same conditions. That latter collection will be referred to as the UBM (Universal Background model). More detail

on GFS is found in [13, 14]. The second dataset is a subset of Pool 2010 [15]. This subset contains 21 male adult speakers with two recordings each and a UBM consisting of 22 further such speakers. The same dataset was used in [16], where further description is provided. Long-term formant analysis (LTF) [17] was performed on these datasets, and the data streams containing F1 to F3 from the vowels in each recording form the input to the methods applied.

Among the LR calculation methods referred to as simple in the title of this paper, three different ones were used, M(ethod) 1 to 3. M2 was performed with two variations, called M2a and M2b.

Method 1: M1 generates pure similarity scores. In a two-dimensional matrix of 22 questioned-speaker recordings compared against 22 suspected-speaker recordings each of the 484 comparison cells contains the result of the following calculation (for the GFS dataset; for Pool 2020 it is 21 x 21 test speakers). Separately for each formant F1, F2, F3, the mean formant frequency value from the second recording is subtracted from the mean value of the first. The absolute value of the result is taken because it does not matter which of the two compared mean values is higher. Since scores represent degree of similarity, whereas subtraction results represent dissimilarity, a minus-sign is added to each of the results. [1]

Method 2: M2 generates scores that take into account both similarity and typicality, but no probabilities are involved. In addition to the step of similarity calculation in M1, each questioned-speaker value is also compared against the UBM. The result from the similarity calculation step is divided by the result from the typicality calculation step involving the UBM. Taking the absolute value and adding a minus-sign to the resulting ratio concludes the calculation. There are two versions of dealing with the UBM. In **M2a** the UBM value is the mean of the means (of F1 etc.) of all the 25 UBM-speakers (for GFS; 22 for Pool 2010). This is the value subtracted from each questioned-speaker recording in order to obtain the denominator of each of the ratios. In **M2b**, a given questioned-speaker value is compared against each UBM speaker and the mean of these 25 (or 22) comparisons is taken as the denominator of the ratio involving that questioned speaker.

Method 3: M3 generates scores that take into account both similarity and typicality, and it does so by producing probabilities and arranging them in the structure of a LR. The formula for this was taken from Morrison [18, see p. 175]. In addition to the mean value for each questioned and suspected speaker, the formula also requires standard deviations for the suspect and the UBM. In order to obtain mean and standard deviation for the UBM, all individual formant values (extracted every 10 ms) across all UBM speakers were collated into one large data column, from which mean and SD were calculated. The formula assumes one single data point from the questioned speaker (or many points with averaging after the entire procedure), but here the mean value was taken for the questioned speaker instead – a methodological option mentioned on [18, p. 177]. Although the formula contains some higher-level math, the method which it captures can be visualised and understood quite easily (see figure on p. 175). Morrison (pers. comm.) presented this formula for introductory purposes, hence in a mainly educational intent. For these reasons M3 is classified with the simple LR methods.

As the top-line method, expected to represent maximum performance, GMM-UBM was used. The same method was used in [14], where further technical detail is found (except that the option "symmetric testing" was disabled here). The software used for that purpose is VOCALISE version 1.6 [19]. The method was applied in two versions. In the one called **GMM1**, suspect model and UBM were modelled with a single Gaussian, in **GMM3** with three Gaussians. MAP (maximum a posteriori) adaptation was not applied. In contrast to the GMM method, M1 to M3 were all limited to univariate data, i.e. there is a calculation for each formant separately. There was also no modelling of more complicated than single Gaussian distributions: M1 and 2 have no distribution model at all and M3 works with single Gaussians. Both of this differs from GMM, which can work with multivariate data and several Gaussians.

Each of the methods generates scores. These are subsequently transformed into LRs with the use of logistic regression cross validation calibration or with fusion across all three formants ([20] on cross validation, [18] on calibration and fusion). Some results are shown visually as Tippett plots [7] and all reported numerically as Equal Error Rate (EER, convex hull method) and log-likelihood-ratio cost (Cllr) [21]. For calibration, fusion, plots and performance calculation the software BIO-METRICS version 1.8 [22] was used.

3. RESULTS

Figure 1 shows Tippett plots for the score results (not the calibrated LRs) of Methods 1 to 3 in the GFS dataset, applied to the parameter F3, which turns out to be the most successful among the single formants. In the Tippett plots, same-speaker scores are shown in the cumulative distributions rising from left to right and different-speaker scores are the ones rising from right to left.

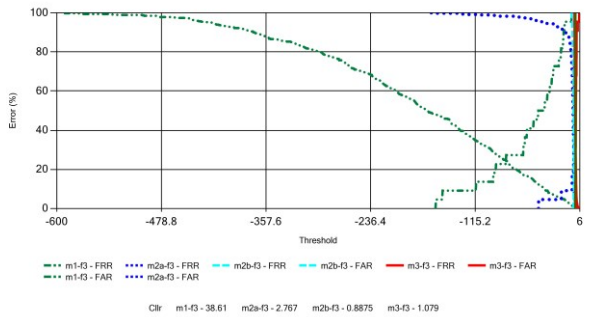


Figure 1: Tippett plots of scores for parameter F3 analysed with Methods 1, 2a, 2b and 3 on dataset GFS. (Cllr values shown are prior to calibration.)

Fig. 1 shows that M1 generates a wide range of scores, which due to the added minus-sign have the value zero as upper limit (that limit also holds for M2ab). All the other methods result in scores that are limited to a much narrower range of values. This range is expanded in Fig. 2.

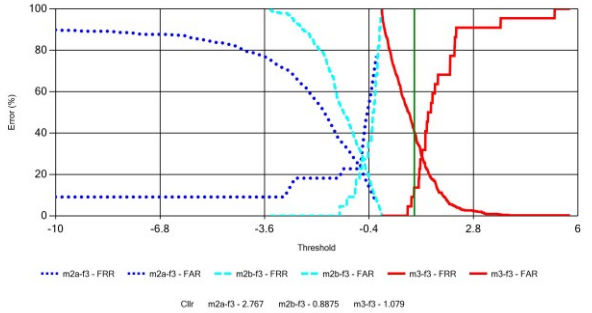


Figure 2: Tippett plots of scores for parameter F3 analysed with Methods 2a, 2b and 3 on dataset GFS. Green vertical line indicates the value 1.

Fig. 2 shows that M2a and M2b have similar patterns but there are more values extending into the lower range of the scores in M2a. As will be discussed below, this can be at least partially the result of outliers that are due to a limitation of the method. M3 generates scores that in their intersection point are quite close to the value one. This proximity to unity indicates that the purpose of the formula behind M3 to create LRs has been successful, i.e. the scores are quite well calibrated even before applying logistic regression calibration.

Speaker discrimination performance measured as EER and Cllr for GFS is shown in Table 1 (lower values indicate better performance). Addressing EER first, when the different methods are compared there are no clear systematic differences except that the results for F1 and F2 are better with GMM3 (but not GMM1) than with the simple methods. F1 and F2 capture phonological vowel distinctions, F3 only marginally so. It could therefore be expected that the distributions of F1 and F2 values across a recording are more complex than the distribution of F3, because they capture differences between different vowel categories. In previous LTF studies this effect however is stronger for F2 than F1 [16, 23 for illustration], whereas here the degree of improvement is similar, though indeed proportionally slightly stronger for F2. Three Gaussians allows for a better modelling of complex distributions than one (in GMM1 as well as M3) or none at all (other methods), which would explain the improvement.

Method	F1	F2	F3	Fusion F123	Multiv F123
M1	36.3 0.93	34.9 0.91	21.8 0.69	20.0 0.60	-
M2a	37.2 1.09	36.5 1.00	23.2 1.31	35.4 1.41	-
M2b	36.2 0.92	34.6 0.94	23.5 0.73	20.8 0.66	-
M3	38.0 0.97	39.0 0.95	26.5 0.84	26.1 0.82	-
GMM1	36.1 0.91	33.2 0.89	25.4 0.76	19.2 0.68	20.9 0.68
GMM3	32.1 0.89	29.2 0.90	25.0 0.79	20.0 0.70	21.5 0.71

Table 1: EER (upper value per cell) and Cllr (lower value) in GFS 2.0 for different methods applied to single formants, fusion of all three formants and multivariate combination of the formants, the latter only possible for GMM.

When comparing EER in different formants and their fusion/combination, F3 turns out to be stronger than the other formants across methods. Fusion of all formants compared to single formants is of benefit for most methods. In GMM, where multivariate processing is possible, multivariate combination of the formants produces very similar results as fusion of the formants.

Turning to the Cllr results now, Cllr most of the time has relative patterns similar to EER, which is expected because discrimination is part of what Cllr measures (and calibration loss, which can affect Cllr, is minimised due to cross validation calibration or fusion). There is strong discrepancy however in M2a, where values of Cllr above 1 occur (1 is chance level). These are likely due to outlier scores that originate from the circumstance that when the questioned speaker value is close to the UBM value, the difference approximates zero and this can strongly increase the value of the ratio calculated in M2a (hence reduce the final score when the minus-sign is added; see Figs. 1 and 2). A few of such problems also occurred in M2b, but less frequently or strongly so.

Method	F1	F2	F3	Fusion F123	Multiv F123
M1	44.9 1.01	30.3 0.88	26.0 0.71	23.0 0.66	-
M2a	43.5 1.00	35.8 1.00	31.3 1.91	40.8 1.81	-
M2b	43.0 1.02	28.9 0.89	28.3 0.73	24.5 0.69	-
M3	45.7 1.01	39.9 0.99	27.7 0.82	27.6 0.81	-
GMM1	34.4 0.92	30.4 0.85	28.1 0.74	22.8 0.64	20.6 0.65
GMM3	27.5 0.83	19.2 0.66	25.8 0.72	11.8 0.42	11.2 0.49

Table 2: EER (upper values) and Cllr (lower values) in Pool 2010.

Table 2 shows EER and Cllr for the Pool 2010 dataset. Pool 2010 is based on a laboratory collection and is expected to yield better performance than the real-case data in GFS. A comparison between the different methods again shows an improvement for F1 and F2 of GMM3 over the simple methods (for F1 also of GMM1). Proportionally this improvement is stronger now than it was with the GFS data. Percentage of improvement is again slightly higher for F2 than for F1. It is likely that the formant distributions are more complex in the Pool 2010 data than the GFS case data because the formants can be measured better on average and the speech style in Pool 2010 is farther to the right on a scale between casual and clear speech. This could explain why the improvement of GMM3 over the simple methods is stronger with the Pool 2010 data. When all these formants are fused or combined, EER is considerably better with Pool 2010 than GFS.

The Cllr results in Table 2 generally follow the pattern of EER, with the same kind of problems for M2a as before.

4. CONCLUSIONS

Based on two datasets involving long-term formant analysis, comparisons were made between the use of a method that is state-of-the-art for semiautomatic voice comparison and simpler methods of LR calculation. Justification for the term LR is given by the use of cross validation calibration, which turns even the results of the simplest method of pure similarity scoring (M1) into LRs. When the methods are applied to real case data (GFS 2.0) the simple methods show nearly the same speaker discrimination performance as the top-line method. When applied to a laboratory collection (Pool 2020), the top-line method clearly outperforms the simple methods, which is probably due to more complex formant value distributions caused by more defined formant structure and a clearer speech style.

The demonstrated relative success of the simple methods with case data does not imply that they should be preferred in casework over state-of-the-art methods if these are available, nor that pure similarity scores should be used in more than preliminary ways [24]. Moreover, it is hard to predict whether even within casework, conditions occur that give the more complex methods a clear advantage over simpler ones. One crucial factor seems to be whether the forensic feature analysed can be approximated by a normal distribution or whether the distribution of feature values is more complex. Features might also have a multivariate structure, which is the case with formants. As could be shown however, performance of multivariate processing (which the simple methods do not perform) is equivalent to univariate processing followed by fusion (which can be done with the simple methods).

Though it is clear that simple methods should not replace complex ones if available, results have shown that simple methods of LR calculation in casework conditions can approximate the performance of complex ones (with the provision that this study is limited to just one type of forensic-phonetic feature). This is interesting since results could have turned out to show a greater performance difference. This outcome suggests there is some substance to the reasons for this study, stated in the introduction, that simple methods of LR calculation can serve as conceptual tools for LR estimation, a starting point for the development of more complex methods or an educational device.

It should be mentioned that the method of cross validation calibration that was used to arrive at LRs from the scores, as well as fusion, are not entirely simple methods, but very common in forensic phonetics and increasingly so in other forensic sciences [1, 18]. The performance index EER however is not affected by calibration, but it can be by fusion. When thinking of LR estimation, the equivalent of numerical calibration would be to either have a good estimate of similarity and also typicality, e.g. based on experience and some research results [12 for examples], or, especially in the case of pure similarity scores (as in M1), to develop knowledge about where the turning point is between differences that support the same-origin hypothesis and those that support the different-origin hypothesis. Fusion would be about developing an idea about degree of correlation between features.

5. REFERENCES

[1] Robertson, B., Vignaux, G. A., Berger, C.E.H. 2016. *Interpreting Evidence*, 2nd ed. Wiley.

[2] Gold, E., French, P. 2019. International practices in forensic speaker comparisons: second survey. *Int. J. Speech, Lang. Law* 26, 1–20.

[3] Meuwly, D., El-Maliki, M., Drygajlo, A. 1998. Forensic speaker recognition using Gaussian Mixture Models and a Bayesian framework. *COST-250 Workshop on Speaker Recognition by Man and by Machine*, Ankara, 52–55.

[4] Rose, P. 2002. *Forensic Speaker Identification*. Taylor & Francis.

[5] Morrison, G. S. 2022. Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Sci. Int.: Synergy* 5, 100270.

[6] Morrison, G. S., Enzinger, E., Ramos, D., González-Rodríguez, J., Lozano-Díez, A. 2020. Statistical models in forensic voice comparison. In: Banks, D., Kafadar, K., Kaye, D. H., Tackett, M. (eds.), *Handbook of Forensic Statistics*. CRC, 451–497.

[7] Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen J., Niemi, T. 2015. *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Verlag für Polizeiwissenschaft. http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf.

[8] Aitken, C. G. G., Lucy, D. 2004. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53, 109–122.

[9] Rose, P. (2006): The intrinsic forensic discriminatory power of diphthongs. *Proc. 11th Australian Int. Conf. Speech Science and Technology*, 64–69.

[10] Reynolds, D., Quatieri, T., Dunn, R. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19–41.

[11] Becker, T., Jessen, M., Grigoras, C. 2008. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech, Brisbane*, 1505–1508.

[12] European Network of Forensic Science Institutes (ENFSI), 2015. *ENFSI Guideline for Evaluative Reporting in Forensic Science*. http://enfsi.eu/wp-content/uploads/2016/09/m1_ guideline.pdf.

[13] Solewicz, Y.A., Jessen, M., van der Vloed, D. 2017. Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proc. Interspeech*, Stockholm, 2849–2853.

[14] Jessen, M. 2021. MAP adaptation characteristics in forensic long-term formant analysis. *Proc. Interspeech*, Brno, 411–415.

[15] Jessen, M., Köster, O., Gfroerer, S. 2005. Influence of vocal effort on average and variability of fundamental frequency. *Int. J. Speech, Lang. Law* 12, 174–213.

[16] Jessen, M. Alexander, A., Forth, O. 2014. Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proc. Audio Engineering Soc. 54th Int. Conf. Audio Forensics*, London, 28–35.

[17] Nolan F., Grigoras, C. 2005. A case for formant analysis in forensic speaker identification. *Int. J. Speech, Lang. Law* 12, 143–173.

[18] Morrison, G. S. 2013. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Aust. J. Forensic Sci.* 45, 173–197.

[19] VOCALISE software. https://oxfordwaveresearch.com/products/vocalise/

[20] Morrison, G. S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM). *Speech Commun.* 53, 242–256.

[21] Van Leeuwen, D. A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*. Springer, 330–353.

[22] BIO-METRICS softw. https://oxfordwaveresearch.com/products/bio-metrics/

[23] Jessen, M., Becker, T. 2010. Long-term formant distribution as a forensic-phonetic feature. *J. Acoust. Soc. Am.*, 128 (4), 2378. (For illustr. see p. 6 of https://www.researchgate.net/publication/343268341_Long-Term_Formant_Distribution_as_a_forensic-_phonetic_feature.)

[24] Morrison, G. S., Enzinger, E. 2018. Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Sci. Justice* 58, 47–58.

[1]Mathematical expression of the proposed methods. Abbreviations not explained so far: s_x = score according to method x; i, j = the two recordings to be compared; f = formant results; Q = questioned speaker; S = suspected speaker; topline = mean.

Method 1

$$s_1(i,j) = -\left|\overline{f_{Q,i}} - \overline{f_{S,j}}\right|$$

Method 2a

$$s_{2a}(i,j) = -\left|\frac{\overline{f_{Q,i}} - \overline{f_{S,j}}}{\overline{f_{Q,i}} - \overline{f_{UBM}}}\right|$$

Method 2b

$$s_{2b}(i,j) = -\frac{\left|\overline{f_{Q,i}} - \overline{f_{S,j}}\right|}{\frac{1}{K}\sum_{k=1}^{K}\left|\overline{f_{Q,i}} - \overline{f_{UBM,k}}\right|}$$

Method 3

$$s_3(i,j) = \frac{\frac{1}{\sigma_S\sqrt{2\pi}}\, e^{-\frac{\left(\overline{f_{Q,i}} - \overline{f_{S,j}}\right)^2}{2\sigma_S^2}}}{\frac{1}{\sigma_{UBM}\sqrt{2\pi}}\, e^{-\frac{\left(\overline{f_{Q,i}} - \overline{f_{UBM}}\right)^2}{2\sigma_{UBM}^2}}}$$

ARE THERE INDIVIDUAL DISFLUENCY PATTERNS?

Angelika Braun & Nathalie Elsässer
University of Trier, Germany
brauna @uni-trier.de and s2naelsa@uni-trier.de

ABSTRACT

This contribution takes a look at disfluencies[1] from the forensic practitioner's point of view. It focuses on individuality in the use of disfluency markers. The materials analyzed consist of several minutes of spontaneous speech by eight speakers on three different points in time. Analyses cover fillers including two elements which have not received much attention in previous research: the nasal filler and verbal fillers. Within- and between-speaker differences are assessed. Statistical analysis shows that disfluency markers will distinguish speakers at a level well above chance. At the same time, results show that it is impossible to pin down a single measure which will characterize the disfluency behavior of individual speakers. Rather, a combination of parameters is needed. The forensic implications of these findings are discussed.

Keywords: forensic phonetics, disfluencies, fillers, hesitation

1. INTRODUCTION

Whenever speakers engage in spontaneous conversation, disfluencies, i.e., disruptions of the speech flow are bound to occur. Since about the 1960s, fillers in particular have been looked at primarily as indications of verbal planning and self-monitoring on the part of the speaker (cf. e.g. [18] or [25] [16]) and thus as examples of the symptom function of speech.

The lexical status of fillers has been subject to debate. In this context, Clark and Fox Tree [11] argue vehemently that fillers are normal words (interjections), which would make them signals in the sense of Bühler [9], Corley and Stewart [12], based on an in-depth literature search, summarize that "[t]here is no conclusive evidence that fillers are words" (p. 600).

A key issue in this discussion is to what extent speakers have control over the use of fillers. Clark and Fox Tree [11] find evidence that speakers consciously distinguish between *uh* and *um* in a systematical manner: This is a point which bears relevance to using this parameter in the forensic domain: anything which is subject to active influence on the part of the speaker is prone to voice disguise. If speakers actively control the use of fillers, the value of the parameter for forensic analysis will decrease.

2. INDIVIDUALITY IN DISFLUENCIES

There are observations in various studies pointing to the fact that patterns in the use of disfluency markers may be individual, [19], [16], [2], [18], [10], [13], [21], [25], [1], [11:97-98], [14:6], [15]. This is consistent with the notion that disfluency behavior reflects the cognitive planning process of a specific individual. In the early literature on disfluencies, individuality is stressed much more than in more recent publications. However, none of these studies report data for individual speakers.

A number of issues have been studied in conjunction with disfluencies. Many studies focus on pausing. They generally distinguish between filled and unfilled or silent pauses and among the latter between breath pauses and still pauses [29]. The fillers studied include *uh* and u*m* and their cognates in various languages (e.g. German *äh* and *ähm*, French *euh*; see [1], [11], and [17] for an overview). This, however, does not describe the full range of fillers in actual speech [6].

The present contribution seeks to establish a more comprehensive concept of disfluency than has been done in previous studies. It makes use of the "classical" fillers, which have been studied for decades, but it also proposes new elements which have so far hardly, if ever been considered, such as the nasal filler and verbal fillers. Verbal fillers are multifunctional lexical items [27] which may either carry propositional meaning or serve as fillers. Examples from German are *und* or *ja*, but also phrases which make the search for the appropriate word explicit, such as *wie sagt man* ('how do you put it') or *mir fällt gerade das Wort nicht ein* ('I can't think of the word right now'). In order to shed light on speaker individuality, results are reported separately for each speaker and also by session.

A dedicated forensic approach to disfluencies was developed by McDougall and colleagues [24], [25]. They list a number of parameters which describe the behavioral profile of a given speaker and can be used in forensic casework. This is more comprehensive than any other framework, but it still falls short of being exhaustive, and the intraspeaker consistency of the features is assumed but not tested. The study by Hughes et al. [20] also has a forensic focus. Their approach is confined to studying various aspects of the *um* formant dynamics in a likelihood-ratio based format. They find that to provide relevant information about voice identity, which underlines the importance to include the spectral characteristics

of fillers in the analysis. The key questions to be explored by this research are thus

(a) Are there speaker characteristic features in the disfluency behavior which have so far not been exploited?

(b) Are speakers at all consistent in their disfluency behavior?

(c) Are there features which are suitable for distinguishing between speakers?

3. MATERIALS AND METHODS

The materials used in this study have been analyzed to some degree earlier on [8]. Additional analyses are presented here. Materials consist of recordings from 8 middle-aged female speakers. Subjects' age varied between 45 and 65 years. They were all from the same part of the country. This eliminates gender, age, and dialect as influencing factors. In the present data set, no significant correlation was found between the frequency of disfluencies and speaker age (r = .286; Spearman rank correlation.) Speakers were recorded talking spontaneously about a fixed set of topics (their recent vacation, books they had read, soccer, their opinion on the ban on smoking in restaurants and pubs, what they would tell Angela Merkel if they had a chance to meet her etc.) They were prompted by the investigator once they had nothing more to say on a given topic, but other than that the speech material was monological. Thus, certain factors which have been found to influence disfluency rates, such as relationship, topic, and syntactic complexity [5] could be ruled out for this data set. There were three recording sessions per speaker, which took place about a week apart. This was done in order to be able to look at within-speaker variability vs. between-speaker variability. Sessions lasted between seven and 15 minutes, depending on the frequency of disfluencies and the speaking tempo.

Recordings were analyzed by creating Praat [4] textgrids and annotating the type of pause (breath, silent, filled), the type of filler, the way fillers were worked into the text (preceded and/or followed by a pause), glottal constrictions, verbal fillers, as well as restarts and repairs. The steady phase of the "classical" fillers was marked. It was subsequently used for formant measurements. All annotations were done manually. The following parameters were analyzed by way of Praat scripts: Number of filled and unfilled pauses, number of breath pauses, number and type of fillers, prolongations, connection of fillers with surrounding text, restarts and repairs, formant frequencies of "classical" fillers[2], voice quality of fillers. Only a fraction of results can be presented here.

4. RESULTS

4.1 How to report results

The question of how to report the results is by no means trivial. It has not received much, if any attention in previous studies. Disfluency rates may be expressed in terms of occurrence per lexical item or per time unit. In the English-speaking world, it is common to report disfluency rates per 100 words, cf. e.g. [5], [11], [21]. While this may work for English, it is certainly not a good idea for languages like German which have an almost unlimited potential for compounding, as the word *Bundesausbildungs-förderungsgesetz* ('Federal Education Promotion Act') shows.

The second option is reporting disfluency rates as N per time unit, e.g., per minute. This solves the problem of different word lengths in different languages or speaking styles, but not that of different speech rates. In the forensic context, in which individuality is aimed at, it is crucial to find the measure which best characterizes individual disfluency behavior. In our materials, the correlation between the number of words and the duration of the individual recordings is quite high (r = .63; p = .00093). This means that the results of disfluencies per minute and per 100 words should be similar. This, is indeed the case most of the time, but not if there are large differences in speaking tempo.

4.2 Numbers and proportions

The presentation of results is largely descriptive because the focus is on individual behavior rather than averaged findings [6,7]. Figure 1 shows the proportion of the various disfluency markers per speaker and session.

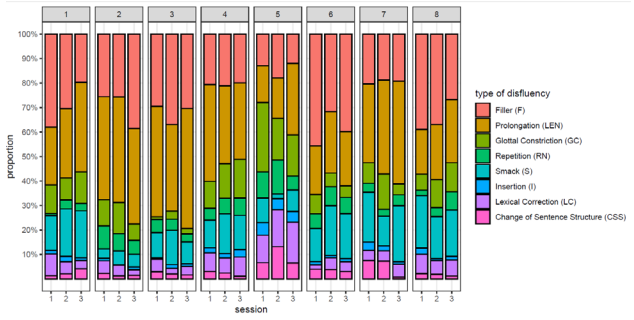


Figure 1: Proportion of disfluency types per speaker and session.

While distributions in the three recording sessions per speaker look strikingly similar, there are notable between-speaker differences. These apply for instance to the relation of fillers (F) to prolongations (LEN). Speakers #3 and #4 show n inverse relationship between the two, whereas #6 stands out by her rare use of prolongations. Other speakers, e.g.,

#7 and #8, are more difficult to distinguish by this parameter.

Table 1: Proportion of vocalic and consonantal prolongations (per 100 words)

prolongations speaker no.	V per 100 words	V %	C per 100 words	C %
#1.1	3.35	54.7	2.77	45.3
#1.2	3.82	61.5	2.45	39.5
#1.3	3.88	45.2	4.71	54.8
#2.1	.58	43.0	.77	57.0
#2.2	.43	19.9	1.73	80.1
#2.3	.78	27.1	2.10	72.9
#3.1	1.78	30.8	4.0	69.2
#3.2	1.79	28.9	4.41	71.1
#3.3	1.29	24.9	3.89	75.1
#4.1	6.88	48.1	7.41	51.9
#4.2	4.67	46.1	5.44	53.9
#4.3	5.54	50.6	5.41	49.4
#5.1	2.72	44.7	3.36	55.3
#5.2	.94	21.1	3,52	78.9
#5.3	1.14	30.3	2,62	69.7
#6.1	.38	15.0	2.15	85.0
#6.2	.53	22.8	1.79	77.2
#6.3	.92	23.1	3.07	76.9
#7.1	.5	16.9	2.45	83.1
#7.2	.95	22.2	3.32	77.8
#7.3	.72	15.4	3.96	84.6
#8.1	1.42	31.2	3.15	68.8
#8.2	.96	20.4	3.75	79.6
#8.3	2.52	39.4	3.88	60.6
Total	**2.09**	**35.07**	**3.87**	**64.93**

Somewhat unexpectedly, prolongations make up the bulk of disfluency markers. For four speakers, their number exceeds that of fillers, three have more fillers than prolongations, and for speaker #1 they are about equal in number. The large proportion of prolongations can possibly be explained by a very meticulous annotation process on our part. Since this was carried out for all speakers alike, the results are comparable. The distribution of vocalic and consonantal prolongations varies between speakers. Table 1 shows the results.

The only study which these results can be compared to is [3]. They report a proportion of vowels of 44.3. The mean proportion of lengthened vowels in our data is 35.1% and thus looks compatible with their results. However, considerable between-speaker differences are evident from the range (15 – 61%). The proportion varies among our speakers but is very consistent over sessions in four of them, whereas two are quite variable. Seven out of eight speakers lengthen consonants more frequently than vowels.

Figure 1 also shows the frequency of occurrence of clicking sounds (S) (smacks). It is evident from the distribution that there are speakers who use this disfluency marker quite frequently. In fact, speaker #7 uses them almost as often as fillers, whereas speakers #3 and #6 hardly exhibit clicks at all.

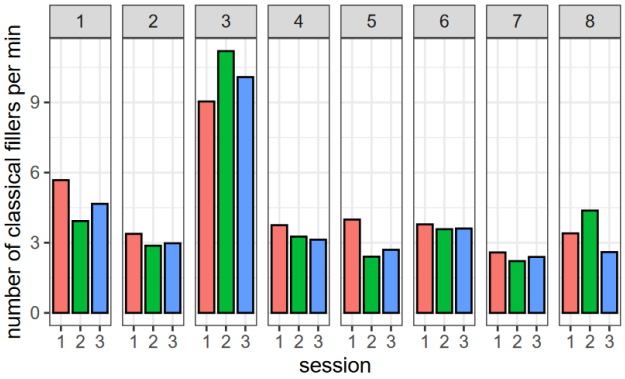


Figure 2: Number of "classical" fillers including *mh* per minute per speaker and session

Figure 2 depicts the number of the "classical" fillers including *mh* per minute. Again, there iss high within-speaker consistency, but this parameter is not suitable to distinguish between speakers very well. Six out of eight speakers are very similar, the remaining two stand out.

The number of fillers per 100 words ranges from 1.3 to 7.3[3]. This considerably exceeds the numbers given by Clark and Fox Tree [11], who find a median of 1.73. One explanation for this difference may be that German, due to its more complex lexical and probably also syntactical structure (the finite verb being at the very end of a subordinate clause), may be more prone to disfluencies than English. If disfluencies per minute are counted, our numbers (2.2 to 11.2) correspond well to those reported by Belz [1] for German in that format, i.e., 1.9 to 11.3.

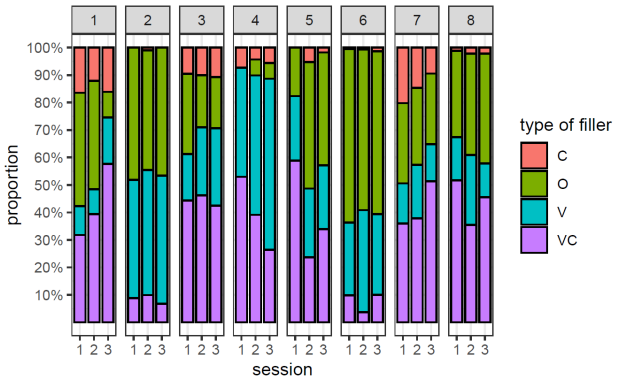


Figure 3: Proportion of types of fillers by speaker and session

Figure 3 depicts the relative proportion of the fillers, i.e,, *äh* (V), *ähm* (VK) and *mh* (K) as well as verbal fillers (termed "O" in the caption) only. As pointed out earlier, the distribution between *äh* and *ähm* is of special interest. Speaker #1, for instance, uses practically no *äh*s, whereas speakers #2 and #6 very rarely exhibit *ähm*. All in all, the majority of subjects (N = 5) show

a clear preference for *ähm,* three for *äh*, and one speaker actually uses verbal fillers more frequently than all others. The preference is clearly speaker specific in our materials. T-tests (unpaired, two-tailed) show a significant preference for *ähm* over *äh* in four speakers (#1: t = -3.82; df = 2.39; p = .046; #3: t = -6.01; df = 2.403; p = .016; #7: t = -4.974; df = 2.566; p = .022; #8: t = -4.314; df = 3.873; p = .013). Two speakers prefer *äh* significantly (#2: t = 26.173; df = 3.916; p = .000; #6: t = 6.09, df = 3.428; p = .005). For the remaining two speakers, results fall short of significance (#4: t = 1.134; df = 3.899; p = .321: #6: t = -1.425; df = 2.011; p = .289). Our data thus do not confirm the observation by Belz [1] and Wieling et al. [30] that speakers, women in particular, generally produce significantly more CV-fillers than V-fillers.

Additional support for considering the nasal filler and verbal fillers to be hesitation markers as opposed to interjections comes from the F0 data. There is no significant difference in this respect between *äh* and *ähm* on the one hand and *mh* and verbal fillers on the other (unpaired T-test, two-tailed; t = 1.064, *p* =.293). This can be taken as yet another indication that these two constitute fillers in their own right. Once again, though, there is one speaker (#7) who does not follow this pattern.

The statistical challenge in this study consists in the fact that in the present context one is trying to prove what is normally considered a confounding factor, i.e., that individual speakers differ. We attempted to do this via a Random Forest model. For training the model, the single takes[4] at all three points in time were treated as separate events. All measurement results in 70% of the takes were used for training. The remaining 30% were used in the classification task. The overall accuracy achieved was 78.5%. This is relatively high considering that the chance level is at 12.5%. The single parameters that were most important for the classification task were expectedly F1, F3, and F4 of the vowel in *äh* and *ähm* as well as, rather unexpectedly, the duration of the inspiration noise.

6. DISCUSSION

Looking at individual disfluency behavior opens a new perspective [7]. In some respects, previous results are confirmed, but it is evident that they can hardly ever be confirmed for all speakers even in this relatively small group. This applies e.g. to filler fundamental frequency. Incidentally, Belz [1] mentions the same thing for his cohort. This is where the forensic interest kicks in. Some findings which have been considered to be well-established can no longer be confirmed if individual speakers are looked at.

As far as the general debate about the lexical status of fillers is concerned, our data do not support Clark and Fox Tree's theory about their being words. One counterargument is the high between-speaker variability paired with a low within-speaker variability of those fillers. In fact, we find the distribution of *äh* vs. *ähm* to be highly individual. If a speaker has a clear preference for one of the two, it can hardly be argued that speakers make a conscious choice in each individual situation.

7. CONCLUSIONS

With the forensic perspective in mind, the following conclusions can be drawn:

(1) There are quite a few parameters in disfluency behavior which would merit being added to the list contained in the TOFFA framework proposed by McDougall et al. [25] along the lines suggested here.

(2) Most results show a striking intra-subject similarity across the three sessions. This applies for instance to the number and type of filler per session, but also the formants. This demonstrates that disfluency behavior is by no means random but instead follows individual patterns, given that the setting is kept constant. That said, while there are speakers who show a large degree of intra-speaker consistency, others are more variable.

(3) In this small group, between-speaker differences exceed within-speaker differences most of the time. Similarity is not a sufficient criterion on which to draw conclusions about speaker identity, though. The typicality criterion also needs to be taken into account [26]. In order to do this, however, much more material needs to be collected. Typicality will have to be assessed with respect to a speaker pattern as opposed to singular disfluency markers.

It is by no means the intention of this research to argue that speaker identification should rely on disfluency behavior alone. But the present results do indicate that while no single parameter is likely to suffice to distinguish between speakers, looking at the complete disfluency pattern has the potential of doing just that.

It might be argued that this approach is unrealistic in the sense that recordings of sufficient duration will hardly be available in a forensic setting. That is only partly true, though. Particularly in jurisdictions which allow telephone intercepts, there are often many minutes of both questioned and reference speech material available from the same telephone surveillance measure. A detailed analysis of the disfluency behavior is therefore possible.

ACKNOWLEDGMENTS

A warm word of thanks goes to Melissa Hildebrand and Vivien Meyer for doing much of the tagging.

7. REFERENCES

[1] Belz, M. 2021. Die Phonetik von *äh* und *ähm*. Akustische Variation von Füllpartikeln im Deutschen. In: Eklund, R. (ed.) *DiSS. The 7th Workshop on Disfluency in Spontaneous Speech*.. Berlin, 1–4.

[2] Belz, M. and J. Trouvain. 2019. Are 'silent' pauses always silent? *Proc. 19th ICPhS,* Melbourne.

[3] Betz, S., Eklund, R., and Wagner, P. 2017. Prolongation in German. In *DiSS 2017 The 8th Workshop on Disfluency in Spontaneous Speech, KTH, Royal Institute of Technology, Stockholm, Sweden,* pp.13-16.

[4] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer* [Computer program]. Version 6.3.03, retrieved 17 December 2022.

[5] Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F., Brennan, S.E. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44(2), 123-147.

[6] Braun, A. 2021 Forensische Sprach- und Signalverarbeitung. In J. Bockemühl (ed.) *Handbuch des Fachanwalts Strafrecht*. 8th ed. Köln: Carl Heymanns Verlag, 1890 – 1914.

[7] Braun, A. 2021: Nonverbal vocalizations – A forensic phonetic perspective. *Proceedings Workshop on Laughter and other Nonverbal Vocalisations,* Bielefeld 2020.

[8] Braun, A., Rosin, A. 2015. On the speaker specificity of hesitation markers – a pilot study. *Proceedings of ICPhS Glasgow,* 5p.

[9] Bühler, K. 1934. *Sprachtheorie. Die Darstellungs-funktion der Sprache*. Jena: G. Fischer.

[10] Butcher, A. 1973. *Aspects of the perception and production of pauses in speech.* Arbeitsberichte of the Phonetics Institute of the University of Kiel, Nr. 1.

[11] Clark, H. H., Fox Tree, J.E. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73-111.

[12] Corley, M., Stewart, O. M. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. *Language and Linguistics Compass* 2/4, 589

[13] Duez, D. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech* 25, 11-28.

[14] Eklund, R. 2001. Prolongations: A dark hoarse in the disfluency stable. In *Proceedings of DISS '01*, Edinburgh, 5-8.

[15] Fant, G., Kruckenberg, A. and Barbosa Ferreira, J. 2003. Individual variation in pausing. A study in read speech. *PHONUM* 9, 193-196.

[16] Finlayson, I.R, Corley, M. 2012. Disfluency in Dialogue: An Intentional Signal from the Speaker? *Psychon Bull Rev* 19, 921-928.

[17] Fuchs, S., König, L. L., Gerstenberg, A. 2021. A Longitudinal Study of Speech Acoustics in older French Females: Analysis of the Filler Particle *euh* across Utterance Positions. *Languages*, *6*(4): 211; https://doi.org/10.3390/languages6040211 .

[18] Goldman-Eisler, F. 1961. A comparative study of two hesitation phenomena. *Language and Speech* 4: 18-26.

[19] Goldman Eisler, F. 1968. *Psycholinguistics. Experiments in Spontaneous Speech*. London and New York: Academic Press.

[20] Hughes, V., Foulkes, P., Wood, S. 2016. Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, 23(1): 99-132.

[21] Kjellmer, G. 2003. Hesitation. In defence of ER and ERM. E*nglish Studies* 84(2): 170–198. Open yale courses. https://oyc.yale.edu/. Accessed: June 6th, 2022.

[22] Levelt, W J.M. 1983.Monitoring and self-repair in speech. *Cognition* 14: 41-104.

[23] Maclay, H., Osgood, Ch.E. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15: 19-44.

[24] McDougall, K., Duckworth, M. 2018. Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law* 25(2): 205–230. https://doi.org/10.1558/IJSLL.37241

[25] McDougall, K., Rhodes, R., Duckworth, M., French, P., Kirchhübel, Chr. 2019. Application of the 'Toffa' Framework to the Analysis of Disfluencies in Forensic Phonetic Casework. In: Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren (eds), *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia. Canberra, Australia: Australasian Speech Science and Technology Association Inc, pp. 731-735.

[26] Rose, Ph. *Forensic Speaker Identification*. London/New York: Taylor and Francis.

[27] Stenström, A-B. 2012. Pauses and hesitations. In G. Andersen and K.Aijmer (eds.), *Pragmatics of Society.* Berlin/Boston: De Gruyter Mouton, 537-567.

[28] Trouvain, J., Fauth, C., Möbius, B. 2016. Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. *Proceedings of Speech Prosody (SP8)*, 31-35.

[29] Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., Liberman, M. 2016.: Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change* 6 (2): 199–234. doi:10.1163/22105832-00602001.

[1] *Disfluency* is used as a cover term for hesitations (covert repairs) including fillers and (overt) repairs [22].

[2] Four formants were measured.

[3] The corresponding graph is not given here for reasons of space. Both had to be calculated since reference units differ in the literature.

[4] One „take" is defined as a part of a recording dealing with one topic.

Acoustic characteristics of stop consonants in Mandarin-speaking children with cochlear implants

Jing Yang[1], Li Xu[2]
[1]Communication Sciences and Disorders, University of Wisconsin-Milwaukee, USA
[2]Hearing, Speech and Language Sciences, Ohio University, USA

ABSTRACT

The present study examined the acoustic features of stop consonants in Mandarin-speaking children with cochlear implants (CIs). The speakers included 22 children with normal hearing (NH) and 35 children with CIs regrouped into chronological age-matched (CA) and hearing age-matched (HA) subgroups. The speech material was Mandarin disyllabic words including six Mandarin stops (/p, t, k, p^h, t^h, k^h/) followed by different vowels. The VOT, ΔVOT, normalized amplitude, and F2 onset were measured and compared between the NH children and each of the CI subgroups. The results revealed that the children with CIs, regardless of CA or HA, showed no difference from the NH controls on the VOT and ΔVOT. However, both CI subgroups produced lower amplitude for aspirated stops than the NH controls. Both CI subgroups followed the NH pattern of F2 onset varying as a function of place of articulation. In sum, Mandarin-speaking children with CIs have developed NH-like temporal and spectral features for stops but their respiratory control for stop production has not fully matured.

Keywords: children, cochlear implant, stop consonants, acoustic features, Mandarin Chinese.

1. INTRODUCTION

Phonological studies on normally developing children suggest that stops emerged very early and most children produce acceptable stops by three years of age [1-3]. However, detailed investigation of stop acoustics in normally developing children reveals that the refinement of articulatory and laryngeal sequences for stop consonants is a gradual process, which shows a continuing development after three years of age [4-6]. For the hearing-impaired population especially children with severe-to-profound hearing loss since birth, accurately producing speech segments is challenging. Regarding the production of stops, deaf children show deviation from the phonological system rather than individual, isolated speech errors in stop productions [7]. Congenital deaf speakers show highly variable and less contrastable voicing features for English stops [8, 9]. Meanwhile, they tend to use alternative features to produce phonological/phonetic contrasts [10, 11].

With the facilitation of auditory prosthesis, children with CIs show an acquisition order for consonant and vowel inventories comparable to NH children [12]. Among different types of consonants, stops are one of the most frequent and predominant manner of articulation in children with CIs [13, 14]. Acoustically, the change in hearing status introduced by CIs exerts an observable influence on the acoustic-phonetic characteristics of stops in CI users. Some researchers reported enlarged voiced-voiceless contrast represented by the VOT change post-implantation [15], adjusted VOT values in accordance with the change of syllable duration [16], increased voicing lead for voiced stops [17], and the NH-like production of VOT in most pediatric CI users [18]. Some researchers found that although early emerged and frequently used, the speech behaviors including articulatory and phonatory gestures of stop consonants in children with CIs still deviate from NH children after a certain period of device use [19, 20].

Extending from the previous studies that focused on the detailed speech characteristics of speech productions in English-speaking children with CIs [18, 21], The present study aims to characterize the acoustic profile of stop consonants in Mandarin-speaking children with CIs. Mandarin stops are produced at bilabial, alveolar, and velar places, similar to English stops. But Mandarin stops are characterized by aspirated-unaspirated distinction, different from the voiced-voiceless contrast in English. Despite the distinctive phonological labelling of the voicing feature, the phonetic representations of word-initial stops in these two languages are both reflected in short-lag vs. long-lag contrast along the VOT continuum. Given the phonological differences and phonetic similarities between Mandarin and English stop systems, we wondered whether the prelingually deafened Mandarin-speaking children with CIs, after a certain period of device use, develop NH-like acoustic features in stop production.

2. METHODS

2.1. Participants

The speakers included 57 Mandarin-speaking children (22 children with NH and 35 children with

CIs) recruited from the Beijing area, China. The children with NH (13 girls and 9 boys) were all aged between 3.25 and 10 years old (M = 6.17 yrs, SD = 1.69 yrs). None of them was reported as having any speech, language, or developmental issues. The children with CIs (16 girls and 19 boys) aged between 3.77 and 15 years old. They were prelingually deafened and all received unilateral implantation at ages between 1.08 and 9.42 years with the length of device use ranging between 0.08 and 9.49 years. To match with the children with NH, the children with CIs were assigned to chronological age-matched (CA) and hearing age-matched (HA) groups. The CA subgroup included 26 children aged between 3.77 and 10.0 years old (M = 7.08 yrs, SD = 1.67 yrs). In the HA subgroup, there were 26 children with the hearing age ranging between 3.33 and 9.49 years (M = 6.12 yrs, SD = 1.88 yrs). Note that the children with CIs were all prelingually-deafened, the influence of pre-surgery acoustic experience from hearing aid, if any, should be very limited. Therefore, we regarded the length of CI use as their hearing age. Some children with CIs were assigned to both subgroups because their chronological age and hearing age were both within the age range of the NH controls. Independent sample t-test revealed no significant group difference in age (both $p > 0.05$) between NH and CA and between NH and HA children.

2.2. Speech materials

All participants were recorded producing a list of 18 Mandarin disyllabic words containing six Mandarin stop consonants /p, t, k, pʰ, tʰ, kʰ/ each followed by three vowels /a, i, u/, respectively. Note that the vowel /i/ was substituted with /ɤ/ for velar stops /k, kʰ/ due to the phonotactic constraints in Mandarin. The target stop sounds all occurred at the word-initial position in the first syllable. The selection of these Mandarin words considered the vocabulary size of young children and the picturability of the target words. By virtue of the same consideration, the tone environment was not strictly controlled.

2.3. Recording

Speech recording was conducted through a visual-auditory paradigm in a quiet room or sound-treated auditory booth. Each participant was seated in front of a laptop computer with which the pictures representing the target words were presented on the screen in random order. An audio prime was then played and the participants were asked to repeat the target word. The speech samples were recorded through a digital recorder with a 16-bit quantization rate and 44.1 kHz sampling rate. All recorded speech samples were transferred to a desktop hard disk with the same quantization rate and sampling rate. The speech samples were then segmented into individual words using a spectrographic analysis program (Adobe Audition 3.0) and saved as separate .wav files.

2.4. Acoustic measurements

The productions that were identified as stop sounds were subject to acoustic analysis. The acoustic measures included VOT, ΔVOT, normalized amplitudes, and F2 onset. The landmark locations of each token including the burst onset, stop offset/vowel onset, and vowel offset were determined by the observation of spectrograms with the assistance of waveform using the digital audio editing program (Adobe Audition 3.0). VOT was measured as the time interval from the release of oral occlusion to the onset of voicing [22]. Because Mandarin stops are all voiceless stops, the onset of voicing was defined as the start of the periodicity of the following vowel. ΔVOT was calculated for the aspirated-unaspirated pair in each place of articulation followed by each vowel. This measure indexes the magnitude of aspirated-unaspirated contrast. Normalized amplitude was calculated to examine the airflow control. To calculate normalized amplitude, the root-mean-square (RMS) amplitude of the stop and the RMS of the following vowel were calculated, then the difference in dB between these two segments was computed. To measure F2 onset, a vowel analysis program TF32 [23] was used to extract formant tracks. A manual correction was implemented when errors occurred in formant extraction. Then, the frequency of F2 onset was measured at the start point of the vowel periodicity. Considering that the participants in the present study varied in chronological age, the F2 values were converted to z scores for each individual speaker [24] to minimize the influence of vocal tract size on the formant values. To ensure ease of interpretation, the z scores were rescaled to Hz-like values following the methods in Thomas and Kendall (2007) [25]. Note that only /a/ and /u/ vowels were used for the analysis of F2 onset because they are the common vowel context used for all six stops.

3. RESULTS

Fig. 1 presents the VOT of NH children and two CI subgroups for each stop. All three groups of children produced longer VOTs for the aspirated stops than the unaspirated stops. Of the three places, the NH children produced longer VOT for both unaspirated

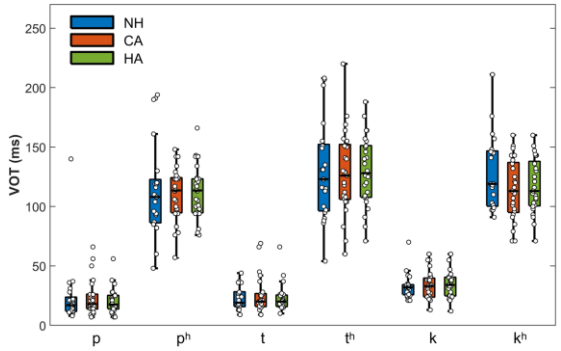


Fig. 1. Box plot showing the VOT of each Mandarin stop in the NH, CA, and HA children.

and aspirated velar stops. The two CI subgroups showed similar place patterns only for the unaspirated stops. A Linear Mixed-Effects Model (LMM) was used to examine VOTs between the NH children and each CI subgroup, respectively. The factors including children's group (NH vs. CA or NH vs. HA), place of articulation, and aspiration were set as fixed effects and the subject was set as a random effect with a random intercept for the subject included. For the NH-CA comparison, the results revealed a significant place difference ($F(2, 223.9) = 4.77$, $p = 0.009$) and aspiration difference ($F(1,223.7) = 1187.3$, $p < 0.0001$) and a place by aspiration interaction ($F(2, 223.4) = 3.78$, $p = 0.024$). For the NH-HA comparison, the results revealed a significant place difference ($F(2, 224.6) = 6.67$, $p = 0.002$) and aspiration difference ($F(1,224.4) = 1523.2$, $p < 0.0001$) and a place by aspiration interaction ($F(2, 224.1) = 5.07$, $p = 0.007$). No significant group difference or group-related interactions were found for both comparisons.

The ΔVOTs across vowel context are shown in Fig. 2. The data were fitted with an LMM in which the children's group (NH vs CA, NH vs HA) and place of articulation were set as the fixed effects, and the subject was set as a random effect with a random intercept for subject included. For the comparison between NH and CA, the results revealed no significant group or place difference. For the comparison between NH and HA, there was a

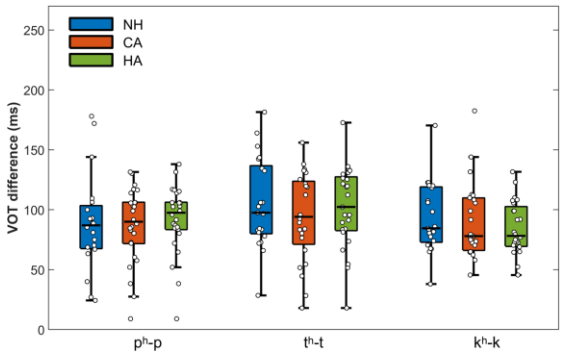


Fig. 2. Box plot showing the ΔVOT for the bilabial, alveolar, and velar stops in the NH, CA, and HA children.

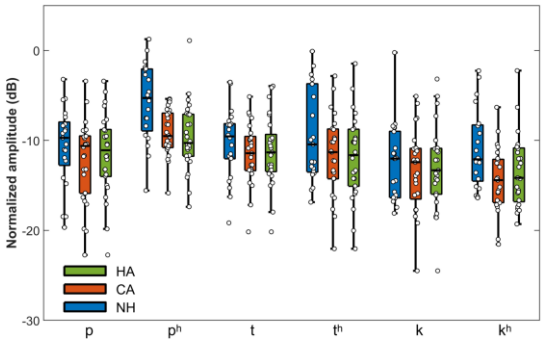


Fig. 3. Box plot showing the normalized amplitude of each Mandarin stop in the NH, CA, and HA children.

significant place difference ($F(2, 86.8) = 4.3$, $p = 0.016$) but no significant group difference.

Fig. 3 presents the normalized amplitudes of the six Mandarin stops collapsed across vowel contexts in the NH and two CI subgroups. The children with NH showed a higher amplitude for aspirated stops than unaspirated stops and the velar stops were produced with lower amplitudes than the other two places. Both CA and HA subgroups followed the NH pattern varying with the place of articulation, but they produced lower normalized amplitude than the NH controls, especially for the aspirated stops. An LMM was implemented to compare the normalized amplitude between the NH and each CI subgroup, respectively. For the NH-CA comparison, the results revealed a significant group effect ($F(1, 46.6) = 6.83$, $p = 0.012$), place effect ($F(2, 222.8) = 23.65$, $p < 0.001$), and aspiration effect ($F(1, 222.6) = 14.18$, $p < 0.001$). Meanwhile, the results revealed a significant group by aspiration interaction ($F(1,222.6) = 5.42$, $p = 0.021$) and place by aspiration interaction ($F(2, 222.4) = 9.12$, $p < 0.001$). For the NH-HA comparison, the results revealed a significant group effect ($F(1, 46.5) = 4.10$, $p = 0.049$), place effect ($F(2, 223.4) = 20.83$, $p < 0.001$), and aspiration effect ($F(1, 223.2) = 11.51$, $p < 0.001$). Meanwhile, the results revealed a significant group by aspiration interaction ($F(1,223.2) = 8.29$, $p = 0.004$) and place by aspiration interaction ($F(2, 223.0) = 6.82$, $p = 0.001$).

Fig. 4 shows the F2 onset of the six stops followed by the vowels /a/ and /u/ in the NH group and two CI subgroups. All children, NH or CI alike, produced higher F2 for both /a/ and /u/ following the unaspirated alveolar stop /t/. An LMM was used to compare the group difference in F2 onset in /a/ and /u/ contexts, respectively. In the /a/ context, the NH-CA comparison yielded a significant place ($F(2, 203.5) = 14.130$, $p < 0.001$) and group effect ($F(1, 47.0) = 6.76$, $p = 0.012$) and a significant place by aspiration interaction ($F(2, 202.8) = 10.79$, $p < 0.001$). The NH-HA comparison yielded a significant place effect ($F(2, 206.9) = 15.18$, $p < 0.001$) and place by aspiration interaction ($F(2, 206.04) = 8.50$, $p < 0.001$). In the /u/ context, the NH-CA comparison yielded a

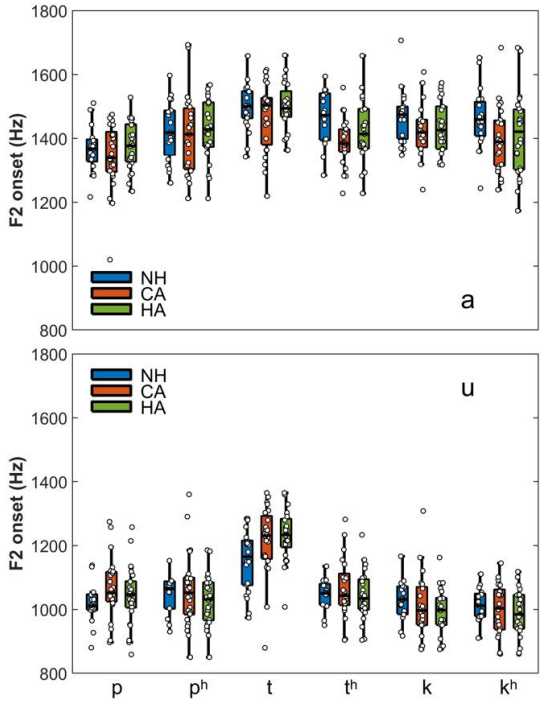


Fig. 4. Box plot showing F2 onset for each Mandarin stop followed by /a/ (top) and /u/ (bottom) in the NH, CA, and HA children.

significant place ($F(2,188.4) = 36.04$, $p < 0.0001$) and aspiration effect ($F(1,190.7) = 19.23$, $p < 0.0001$) and a significant place by aspiration interaction ($F(2,187.7) = 16.66$, $p < 0.001$) but no significant group effect or group related interactions. The NH-HA comparison yielded significant place ($F(2, 194.7) = 58.91$, $p < 0.0001$) and aspiration effects ($F(1,196.4) = 35.74$, $p < 0.0001$) but no significant group effect. All two-way interactions were significant (all $p < 0.05$).

4. DISCUSSION

VOT is the "acoustic result of laryngeal-oral coordination reflected in the temporal domain" (p.1336, [26]). Lane and colleagues (1994) [16] reported that postlingually deafened CI users modified the VOT values for English voiced and voiceless stops in accordance with the change in their hearing status caused by CIs. Uchanski and Geers (2003) [18] found that although the CI children demonstrated greater inter-subject variations in the VOT for English /t/ and /d/, over 60% of CI children who used either total communication or oral communication mode had VOTs and ΔVOT values within the normal-hearing range. In the present study, our data revealed that the CI children, regardless of CA or HA, showed no significant difference from the NH children on the VOT for unaspirated or aspirated stops or the ΔVOT for the aspirated-unaspirated contrasts. These results suggested that the children with CIs approximated the NH children and could use the temporal features to effectively distinguish aspiration contrasts in their stop production.

Among the tested acoustic measures, the normalized amplitude is associated with respiratory and airflow control during the process of occlusion and release. Lane and Perkell (2005) [26] proposed that the reduced aerodynamic distinction between voiced and voiceless stops in prelingually deafened speakers might be reflected as having higher than normal peak flow for voiced stops while lower than normal peak flow for voiceless stops. Other researchers also found that deaf children produced lower expiratory airflow than NH peers did [27]. In this study, we observed that both CI subgroups produced significantly lower normalized amplitude than the NH control for the aspirated stops but showed no significant difference from the NH controls for the unaspirated stops. The production of aspirated stops involves a strong airflow immediately following the release burst, which therefore has a higher amplitude than the unaspirated stops. Our data suggested that the aerodynamic distinction of amplitude between unaspirated and aspirated stops in the children with CIs was reduced than the NH controls.

F2 onset contains information regarding the placement of articulation [28, 29]. Of the three places, the alveolar stops have spectral prominence at a higher frequency than the bilabial stops [28]. The F2 onset following the alveolar stops is normally higher than that following bilabial stops. Our data showed significant place difference, in particular, higher F2 onset following the alveolar stops in both /a/ and /u/ contexts in the NH group, which was consistent with the pattern reported in previous studies. The two CI subgroups generally followed the NH pattern of F2 onset varying as a function of place of articulation. This finding suggested that children with CIs could differentiate the place of articulation in their stop production. While F2 onset reflects the articulatory configuration of the preceding consonant, it is also determined by the vowel target. The group difference between the CA and NH children might be partially accounted for by less definable vowel productions in children with CIs [30].

In summary, our acoustic data indicated that Mandarin-speaking children with CIs had not fully developed NH-like respiratory control for stop production, but they could produce NH-like temporal features for the aspirated-unaspirated contrast and produce distinguishable place features in their stops. All three groups of children showed considerable variabilities in the tested measures, which indicated that the acoustic development of stop consonants is a long-term process.

7. REFERENCES

[1] Dodd, B., Holm, A., Hua, Z., Crosbie, S. 2003. Phonological development: a normative study of British English-speaking children. *Clin Linguist. Phon.* 17(8), 617-643.

[2] Smit, A., Hand, L., Freilinger, J., Bernthal, J. Bird, A. 1990. The Iowa articulation norms project and its Nebraska replication. *J. Speech. Hear. Disord.*, 55, 779–798.

[3] Vihman, M., 1996. Phonological Development (Oxford: Blackwell).

[4] Imbrie, A. K. K. 2005. *Acoustical study of the development of stop consonants in children* (Doctoral dissertation, Massachusetts Institute of Technology).

[5] Yang, J. 2018. Development of stop consonants in 3-to-6-year-old Mandarin-speaking children. *J. Child Lang. 45,* 1091-1115.

[6] Liu, C. T. 2021. A first step toward the clinical application of landmark-based acoustic analysis in child Mandarin. *Children*, *8*(2), 159.

[7] Monsen, R. B. 1976. Normal and reduced phonological space: The production of English vowels by deaf adolescents. *J. Phon. 4*(3), 189-198.

[8] Harris, K. S., Rubin-Spitz, J., McGarr, N. S. 1985. The role of production variability in normal and deviant developing speech. In J. L. Lauter (Ed.), *Proceedings of the Conference on the Planning and Production of Speech in Normal and Hearing-Impaired Individuals: A Seminar in Honor of S. Richard Silverman* (pp. 50-57). Rockville, MD: American Speech-Language-Hearing Association.

[9] Metz, D. E., Samar, V. J., Schiavetti, N., Sitler, R. W. 1990. Acoustic dimensions of hearing-impaired speakers' intelligibility: Segmental and suprasegmental characteristics. *J. Speech Lang. Hear. Res*. *33*(3), 476-487.

[10] Slis, I. H., Cohen, A. 1969. On the complex regulating the voiced-voiceless distinction I. *Lang. Speech.* 12(2), 80-102.

[11] Higgins, M. B., Carney, A. E., McCleary, E., Rogers, S. 1996. Negative intraoral air pressures of deaf children with cochlear implants: Physiology, phonology, and treatment. *J. Speech. Lang. Hear. Res*. *39*(5), 957-967.

[12] Serry, T. A., Blamey, P. J. 1999. A 4-year investigation into phonetic inventory development in young cochlear implant users. *J. Speech. Lang. Hear. Res*. *42*(1), 141-154.

[13] Tobey, E. A., Pancamo, S., Staller, S. J., Brimacombe, J. A., Beiter, A. L. 1991. Consonant production in children receiving a multichannel cochlear implant. *Ear. Hear*. 12, 23–31.

[14] Grogan, M. L., Barker, E. J., Dettman, S. J., Blamey, P. J. 1995. Phonetic and phonological changes in the connected speech of children using a cochlear implant. *Ann. Otol. Rhinol. Laryngol. Suppl. 166, 390–393.*

[15] Economou, A., Tartter, V. C., Chute, P. M., Hellman, S. A. 1992. Speech changes following reimplantation from a single-channel to a multichannel cochlear implant. *J. Acoust. Soc. Am*. *92*(3), 1310-1323.

[16] Lane, H., Wozniak, J., Perkell, J. 1994. Changes in voice-onset time in speakers with cochlear implants. *J. Acoust. Soc. Am*. *96*(1), 56-64.

[17] Kishon-Rabin, L., Taitelbaum, R., Tobin, Y., Hildesheimer, M. 1999. The effect of partially restored hearing on speech production of postlingually deafened adults with multichannel cochlear implants. *J. Acoust. Soc. Am*. *106*(5), 2843-2857.

[18] Uchanski, R. M., Geers, A. E. 2003. Acoustic characteristics of the speech of young cochlear implant users: A comparison with normal-hearing age-mates. *Ear. Hear*. *24*(1), 90S-105S.

[19] Higgins, M. B., McCleary, E. A., Carney, A. E., Schulte, L. 2003. Longitudinal changes in children's speech and voice physiology after cochlear implantation. *Ear. Hear*. *24*(1), 48-70.

[20] Yu. J., Xia, X. 2019. Production of Mandarin stop consonants in prelingually deaf children with cochlear implants. In S. Calhoun, P.Escudero, M. Tabain & P. Warren (eds.) *Proc. 19th ICPhS Melbourne, Australia*

[21] Joy, D. A., Sreedevi, N. 2019. Temporal characteristics of stop consonants in pediatric cochlear implant users. *Cochlear. Implants. Int. 20*(5), 242-249.

[22] Lisker L. Abramson A.S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384-422.

[23] Milenkovic, P. (2003). TF32 software program. Madison: University of Wisconsin

[24] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different listeners. *J. Acoust. Soc. Am*. 49, 606-08.

[25] Thomas, Erik R. Tyler Kendall. 2007. NORM: The vowel normalization and plotting suite. Retrieved November 14, 2022, from http://lingtools.uoregon.edu/norm/about_norm1.php#scaling.

[26] Lane, H., Perkell, J. S. 2005. Control of voice-onset time in the absence of hearing. *J. Speech Lang. Hear. Res*. 48(6), 1334-1343.

[27] Żebrowska, A., Zwierzchowska, A., Manowska, B., Przybyła, K., Krużyńska, A., Jastrzębski, D. 2016. Respiratory function and language abilities of profoundly deaf adolescents with and without cochlear implants. In *Prospect in Pediatric Diseases Medicine* (pp. 73-81). Springer, Cham.

[28] Delattre, P. C., Liberman, A. M., Cooper, F. S. 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am*. *27*(4), 769-773.

[29] Cassidy, S. Harrington, J. M. 1995. The place of articulation distinction in voiced oral stops: evidence from burst spectra and formant transitions. *Phonetica*. 52, 263-284.

[30] Yang, J., Xu, L. 2021. Vowel production in prelingually-deafened Mandarin-speaking children with cochlear implants. *J. Speech Lang. Hear. Res*. 64(2), 664-682.

ARTIFICIAL GRAMMAR LEARNING IN INDIVIDUALS WITH DOWN SYNDROME: THE ROLE OF PROSODIC CUES

Vesna Stojanovik[1], Vitor Zimmerer[2] and Jane Setter[1]

University of Reading, UK[1], University College London, UK[2]
v.stojanovik@reading.ac.uk, v.zimmerer@ucl.ac.uk, j.e.setter@reading.ac.uk

ABSTRACT

The artificial grammar learning (AGL) paradigm allows investigation of cognitive processing and learning under controlled conditions. We compared AGL between 17 individuals with Down syndrome (DS) aged 3-12 years and 60 typically developing (TD) individuals aged 5-18 years and examined age effects, whether prosodic cues affected learning, and the effect of familiarity of stimulus sequences.

In general, the TD group strongly outperformed the DS group. In both groups, older children demonstrated better learning. Prosody did not affect learning in the TD group but had a positive effect in the DS group.

While limited by the small group size, our results identify factors that affect or support language learning in DS individuals. This may have clinical and educational implications, as DS individuals may need stimuli augmented with prosodic cues to make them more accessible (e.g., through song).

Keywords: Down syndrome, prosody, artificial grammar, language impairment, learning.

1. INTRODUCTION

Down syndrome (DS) is a developmental disorder which affects approximately 1 in 700 live births [1] resulting from an extra chromosome 21. Individuals with DS are characterized with a range of distinct physiological features including narrow auditory canals, a small oral cavity, short stature, hypotonia, ophthalmological disorders, hearing issues, psychiatric and behavioral difficulties [2]. Individuals with DS have a different cognitive profile from typically developing (TD) individuals, with IQ ranging between 30 and 70, and significant delays with language acquisition.

Infants with DS tend to acquire their first words later than TD children [3]. Development of language beyond the level of single word acquisition has been studied less often and the few existing longitudinal studies on the acquisition of language reached very different results. English speakers with DS were reported to start producing two-word combinations in English around 36 months of age, while Hebrew speaking children with DS started combining words at 55 months [5]. Grammatical acquisition is a challenge and children with DS often score lower than expected for non-verbal mental age on expressive and receptive grammar tasks and on standardized and non-standardized assessments of morphology and syntax [5, 6].

The language input children hear is thought to contain strings of words which are not random but follow the grammatical patterns, or rules, of the target language [7]. While listening to and processing language input, children need to parse the word strings they hear, extract patterns or regularities, and generalize this knowledge to be able to produce and understand new utterances. This language input also contains prosodic cues, which accentuate language structure and are particularly important in the early stages of child language acquisition [8]. In the first few months of life, infants use prosodic cues to help them segment the incoming speech stream into words [9,10]. Opinions are divided as to whether TD 5-year-old children use prosodic cues when disambiguating ambiguous utterances [11,12] because as children grow older, other cues become available, such as segmental and lexical cues [13]. This may be different for children with developmental conditions.

2. ARTIFICIAL GRAMMAR LEARNING

Artificial Grammar Learning (AGL) is an empirical paradigm that allows the investigation of the basic principles underlying the ability to parse combinations of stimuli, extract patterns about their structure and generalize these to novel stimuli [14]. AGL tasks typically involve a familiarization phase where participants are exposed to sentences generated by a (target) grammar, and a test phase, where they judge the correctness of novel sequences. The advantage of AGL is that artificial systems can be

designed to be easy to learn through short exposure, and that language features can be controlled.

AGL has been used extensively over the past few decades to study and identify how representations are acquired during learning and what processing biases may be underlying different developmental profiles, and in different developmental and acquired disorders. For example, a study of auditory artificial grammar learning in children and adolescents with Williams syndrome (WS) [15] found that they made decisions on the familiarity of a particular stimulus combination rather than on whether the sequences followed the rules of the target language (grammaticality); this familiarity-based processing was only present in young TD children, while older switched from familiarity- to grammaticality-based judgments. Importantly, the individuals with WS showed evidence of learning only when the stimuli were presented with prosodic cues, unlike the TD individuals.

A reasonable body of literature documents the relative strengths and weaknesses in the language profile of individuals with DS, including difficulties with grammar (e.g., [6,16]). No studies to our knowledge have investigated AGL in individuals with DS using verbal or auditory stimuli in order to determine which factors contribute to learning.

The aim of the current study is to investigate the performance of a group of individuals with DS on a verbal AGL task, how their performance compares to that of TD individuals and determine factors that relate to task performance (age, the presence of prosody in the stimulus set, and processing biases). The research questions were:

- How do individuals with DS perform in an AGL task compared to TD individuals?
- What is the effect of age?
- Does either group benefit from prosodic cues?
- Do participants extract abstract grammatical knowledge, or do they make judgments based on familiarity with stimuli?

3. METHODOLOGY

3.1. Participants

23 participants with DS were recruited through local charities. Six participants were excluded either because they did not complete the AGL, or they showed a strong response bias (i.e., only pressed one of the two keys). The analysis is therefore based on 17 individuals with DS (6 male), with a mean age of 9 years, 7 months (9;7); range 5;10 – 18;4. They had British English as their main language. Children with DS and additional diagnoses were included.

63 TD children, speakers of British English as their main language, were recruited through a participant database at the University of Reading and from local schools. Three children were excluded from the analyses due to missing data, strong response bias, or failure to complete the task, hence analyses are based on data for 60 children (Mean age = 7 years 5 month; range 3;4 – 12;1; 31 males, 29 females). These participants are the same set as in [15].

3.2. Materials

3.2.1. Baseline assessments

There were two baseline assessments. The non-verbal Raven's Coloured Progressive Matrices (RCPM [17]) assesses fluid intelligence. As individuals with DS are known to have non-verbal cognitive delays, we administered this task to account for non-verbal mental ability. As individuals with DS also present language delays, we measured verbal ability. Depending on a participant's age and their general developmental level, the *Word Structures* subtest of the Clinical Evaluation of Language Fundamentals (CELF-4) or the Pre-School version, CELF-2, were used [18] to assess expressive language ability. We report percentage correct to make the results of the standardised language assessments comparable and also because different versions of the *Word Structures* task have different numbers of items (the pre-school version has 24 and the school version 32).

Table 1 shows participant group demographic information regarding chronological age and their raw scores for the RCPM and CELF percentage correct scores. The groups are not matched on any variables; the TD group are younger on average than the DS group, however the verbal and non-abilities of the TD group are higher than those of the DS group.

	Age in months	**CELF Raw %**	**RCPM Raw**
	Mean (sd) Min-Max	Mean % correct (sd) Min-Max	Mean score (sd) Min-Max
DS (n = 17)	118 (42) 70-221	33 (18) 6-33	13 (4) 3-23
TD (n = 60)	90 (30) 38-145	81 (16) 29-100 (n = 56)	25 (8) 9-36 (n = 59)

Table 1: Participants' age, non-verbal ability and language ability.

3.2.2. The AGL task

Traditional AGL tasks can be too long and complex for children. We designed a novel task which is child-friendly and suitable for individuals with learning

difficulties. It has previously been used with individuals with William syndrome [15]. It involves a narrative about a magician trying to learn his spells. Unlike traditional AGL tasks, ours has a prosody and a no-prosody condition.

Children tend to be more attentive if characters appear and move on a screen than if they are only presented auditorily [19]. Auditory stimuli were shown simultaneously with events on a computer screen. A grammatical string in this task had the structure A(B)C, consisting of three-word classes: A, B and C. Class A referred to the object type appearing on screen (e.g., bird, rabbit). Class B, which was optional in a grammatical string, referred to the size or color of the object (big/small, red/blue). Class C referred to an action happening to the object (e.g., it spins or zooms in). Each word class corresponded to one aspect of an event which was presented on the screen. Each word class was distinguished from the others by having a distinct phonological onset, and all words were bisyllabic. Violations were generated by changing the order of word classes, repeating the same word class in a string, or by having both a word order change and a repetition. For a full list of items, see [15].

There were two conditions: *Prosody* and *No Prosody*. These are described in section 3.3.

3.3. Procedure

The task contained a familiarisation and a judgment phase. Familiarisation trials were created and presented through Microsoft PowerPoint using adapted clipart images and animations of a magician and the objects from the artificial grammar scheme. Ten grammatical strings were generated based on the artificial grammar described above using Cepstral's [20] male British English synthesised voice ('Lawrence'). The speaking rate was 136 words per minute. Phrases were sampled at a rate of 44.1kHz, 16 bit stereo with intensity scaled to 70dB.

In the *Prosody* condition, fundamental frequency (F0) fell across the phrase: the F0 of A class words was 125Hz, for B words it was 100Hz, and for C class words it was 75Hz. There was a falling tone on the last word. In the *No Prosody* condition, the F0 of the phrases was kept constant at 100Hz using Praat [21]. Participants were randomly assigned to either the *Prosody* or *No Prosody* condition. 33 TD participants completed the *Prosody* condition and 27 the *No Prosody* condition. In the DS group, 11 participants completed the *Prosody* and 6 participants the *No Prosody* condition.

The experimental trials were created and presented on a laptop computer using Eprime 2.0 Professional (Psychology Software Tools, Inc). There were 20 strings: 10 ungrammatical and 10 grammatical phrases. Of the grammatical phrases, six were identical to phrases from the familiarisation trials, and four were novel, unfamiliar phrases. The ungrammatical phrases consisted of repetitions of word classes or violations of dependency rules (such as a B class word followed by an A class word, which is the opposite of the rule). In the *Prosody* condition, the F0 was the same as the familiarisation trials. Hence, the ungrammatical phrases were distinct from the grammatical phrases acoustically due to an unfamiliar F0, in addition to violations at the grammar level. *No Prosody* condition phrases were presented at 100Hz.

Participants were instructed that they were about to see a magician practising his spells and all they had to do was to watch and listen carefully. They were told the spells would sound funny because the magician comes from another planet. The familiarisation phase lasted approximately 8 minutes. Phrases were repeated on average 10 times (range 9-11) and were presented simultaneously with corresponding animations in the PowerPoint presentation. A simple game was used between each block to maintain participants' attention and motivation. The test phase followed immediately afterwards. Participants were told that the magician was teaching another magician some spells, and that sometimes these spells would be right and sometimes wrong. The participants had to judge whether spells would work or not by pressing a green smiling face on the keyboard (if they judged a spell to be correct) or a red sad face (if incorrect). Participants received no feedback. When they had completed the task, participants were thanked for listening and asked if they knew which spells would work. Participants commonly responded with: "I was guessing" or "I don't know" suggesting that they had most likely not explicitly extracted a pattern.

Ethical approval was granted by the Research Ethics Committee at the University of Reading. All participants' parents/carers provided written consent and the participants themselves provided child assent (if children).

4. RESULTS

Accuracy means for all groups AGL task are presented in Table 2.

	Prosody	**No Prosody**
	Mean % correct (sd)	Mean % correct (sd)
DS	48.2 (8.4)	42.5 (2.2)
TD	69.4 (18.8)	69.6 (13.2)

Table 2: Accuracy in the AGL across groups and conditions

A One-Way ANOVA showed an effect of group $F(3,73) = 9.398$, $p < .001$. Tukey's HSD found that for each stimulus condition, the TD group performed significantly better than the DS group was significant ($p < .003$).

We used hierarchical generalized linear models for analysis, with each trial entered as a separate data point. The decision to accept the string ("Accepted") as the binary outcome. We started with simpler models and adding variables if they resulted in significant improvements. For each group separately, we started with Grammaticality, then added chronological age (CA), then Prosody.

Model 1: Accepted ~ Grammaticality
Model 2: Accepted ~ Grammaticality*CA
Model 3 Accepted ~ Grammaticality*CA*Prosody

Model 1 investigated the effect of Grammaticality on responses. In the TD group, the model strength was $\chi^2(1) = 187.48$, $p < .001$ compared to the null model. The effect of Grammaticality was significant, $\beta = -1.65$ (SE = .12), $z = -9.51$, $p < .001$. The odds ratio was 5.19, i.e. TD participants were five times more likely to accept a sequence if grammatical. Model 2, which added CA, was a significant improvement to Model 1, $\chi^2(2) = 54.91$, $p < .001$. Model 3, which added Prosody, did not make a significant contribution compared to Model 2, $\chi^2(4) = 6.19$, $p = .19$, suggesting that the Model 2 is most appropriate. There was a significant interaction between Grammaticality and CA, meaning older TD children accepted more grammatical and fewer ungrammatical strings than younger children.

In the DS group, Model 1 was no significant improvement over the null model, $\chi^2(1) = 1.99$, $p = .15$, with no significant effect of Grammaticality. Model 2, was a significant improvement, $\chi^2(2) = 11.55$, $p = .003$. After entering CA, the Grammaticality effect became significant, $\beta = -1.53$ (SE = .71), $z = -2.14$, $p = .03$. However, there was no significant effect of Age, and no interaction between Age and Grammaticality. Entering Prosody in Model 3 was a significant improvement over the previous model, $\chi^2(4) = 11.7$, $p = .01$.The interaction between Grammaticality and Prosody was significant, $\beta = 4.88$ (SE = 1.81), $z = 2.69$, $p = .007$, meaning that in the Prosody condition, participants with DS more consistently accepted grammatical and rejected ungrammatical strings. A three-way-interaction between Grammaticality, CA, and Prosody was also significant, $\beta = -.03$ (SE = .01), $z = -2.208$, $p = .003$, showing a greater effect of Prosody on grammaticality judgments in older participants.

To test the effect of familiarity, we calculated Edit Distance (ED) for each string, a measure of how similar one test string is to the most similar familiarization string. Replacing Grammaticality with ED resulted in weaker models for both groups.

5. DISCUSSION

Individuals with DS were less successful in the AGL task than the TD group, despite having a higher average chronological age. The best fitting statistical model implies that participants with DS were more successful if they were older and the stimuli contained prosody. The positive effect of prosody on learning was stronger in older children with DS.

The TD group, however, reliably based decisions on the grammaticality of strings, with no effect of prosody. However, older children performed better than younger children. Age effects are in line with existing results [22,23].

We found that both groups were more likely to make judgment based on grammaticality than on familiarity of strings. This suggests that language in children with DS although delayed, may not be qualitatively different from neuro-typical children.

Prosody had a facilitatory effect for the individuals with DS only. This is similar to Stojanovik et al. [15], which also showed that prosody also had a facilitatory effect on the AGL for children and adolescents with WS. These findings can be interpreted within the speech segmentation hierarchy proposed by Mattys et al. [13] according to which prosodic cues are the lowest in the hierarchy (lexical cues and segmental cues being higher) and, as such, they are thought to be the earliest and most critical in the early stages of language acquisition. A large body of evidence shows that prosodic cues play a crucial role during young children's speech processing in helping them identify between phrase boundaries and structural relations between phrases (e.g., [9,10]). Given that individuals with DS present with significant delays with language acquisition, it seems likely that they need prosodic cues to help them with a language learning task. This has implications for language interventions for individuals with DS, as it suggests that adding prosody, or making language stimuli more prominent, may aid learning. This is unsurprising, as prosody is a fundamental part of human language.

This study has limitations. The DS group is small, partially because 25% of the recruited participants with DS we recruited were unable to complete the task. Future studies interested in this research should seek replication.

6. REFERENCES

[1] Martin, G. E., Klusek, J., Estigarribia, B., Roberts, J. E. 2009. Language characteristics of individuals with Down syndrome. *Topics in Language Disorders, 29,* 112-132

[2] Baum, R.A., Nash, P.L., Foster, J.E., Spader, M., Ratliff-Schaub, K., Coury, D.L., 2008. Primary care of children and adolescents with down syndrome: an update. *Current Problems in Pediatric and Adolescent Health Care*, 38(8), 241-261.

[3] Stoel-Gammon, C. (2001). Down syndrome phonology: Developmental patterns and intervention strategies. *Down Syndrome Research and Practice,* 7, 93-100.

[4] Oliver, B., Buckley, S. 1994. The language development of children with Down syndrome: First words to two-word phrases. *Down Syndrome Research and Practice*, 2(2), 71-75

[5] Levy, Y., Eilam, A. 2013. Pathways to language: a naturalistic study of children with Williams syndrome and children with Down syndrome. *Journal of Child Language*, 40(1), 106-138.

[6] Joffe, V., Varlocosta, S. 2007. 'Patterns of syntactic development in children with Williams syndrome and Down's syndrome: evidence from passives and wh-questions', *Clinical Linguistics & Phonetics,* 21 (9), 705-27

[7] Pinker, S., 1991. Rules of language. *Science*, *253*(5019), 530-535.

[8] Morgan, J. L., Meier, R. P., Newport, E. L. 1987. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology* 19, 498–550.

[9] Jusczyk, P. W., Friederici, A. D. Wessels, J., Svenkerud, V. Y., Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* 32*,* 402–420.

[10] Jusczyk, P. W., Hirsch-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., Piwoz, J. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology* 24(2), 252–293.

[11] Beach, C. M., Katz, W. F., Skowronski, A. 1996. Children's processing of prosodic cues for phrasal interpretation. *The Journal of the Acoustical Society of America* 99(2), 1148-1160.

[12] Snedeker, J., Trueswell, J. C. 2001. Unheeded cues: Prosody and syntactic ambiguity in mother-child communication. Paper presented at the *26th Boston University Conference on Language Development*, Boston, MA.

[13] Mattys, S. L., White, L., Melhorn, J. F. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General* 134, 477–500.

[14] Reber, A. S. 1967. Implicit learning of artificial grammars. *Verbal Learning and Verbal Behavior* 5(6), 855–863.

[15] Stojanovik, V., Zimmerer, V., Setter, J., Hudson, K., Poyraz-Bilgin, I., Saddy, D., 2018. Artificial grammar learning in Williams syndrome and in typical development: The role of rules, familiarity, and prosodic cues, *Applied Psycholinguistics* 39(2), 327–353.

[16] Næss, K.A.B., Lyster, S.A.H., Hulme, C. Melby-Lervåg, M., 2011. Language and verbal short-term memory skills in children with Down syndrome: A meta-analytic review. *Research in developmental disabilities* 32(6), 2225-2234.

[17] Raven, J. 2007. *Coloured Progressive Matrices*. London: Harcourt Publishers.

[18] Semel, E., Wiig, E. H., Secord, W. A. 2004. *Clinical Evaluation of Language Fundamentals - 4 (CELF-4).* Toronto, Canada: The Psychological Corporation/A Harcourt Assessment Company.

[19]Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science*, *14*(5), 427–432. doi: 10.1111/1467-9280.01458.

[20] Cepstral https://www.cepstral.com/

[21] Boersma, P., Weenink, D. 2005. *Praat: doing phonetics by computer* (Version 4.3.01) [Computer program]. http://www.praat.org/.

[22] Fletcher, J., Maybery, M.T., Bennett, S. 2000. Implicit learning differences: A question of developmental level? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(1), 246–252.

[23] Maybery, M. T., Taylor, M. O'Brien-Malone, A. 1995. Implicit learning: Sensitive to age but not IQ. *Australian Journal of Psychology* 47(1), 8–17.

RHYTHM DEFICITS IN DEVELOPMENTAL DYSLEXIA: EVIDENCE FROM ITALIAN

Marina Rossi, Eline Smit, Tamara Rathcke

University of Konstanz
marina.rossi@uni-konstanz.de, eline.smit@uni-konstanz.de, tamara.rathcke@uni-konstanz.de

ABSTRACT

Rhythm deficit in developmental dyslexia has recently received much attention. Three major accounts consider it to be the key issue in dyslexia, though supporting evidence is somewhat conflicting. The present experiment investigated rhythm processing in Italian – a language with a comparably high number of dyslexia diagnoses. Adolescents with and without developmental dyslexia were asked to tap in synchrony with metronome and monosyllable sequences at a fast and a slow pace. The results revealed subtle differences in rhythmic performance at the group level, though limited to the slow-paced condition only. Synchronisation performance of dyslexic participants showed longer absolute asynchronies indicative of a larger synchronisation error as compared to asynchronies measured for typically developing adolescents. No differences were found for sensorimotor stability, signed asynchronies, and none could be uncovered by more fine-grained time-series analyses of rhythmic performance. We conclude that there may be a rhythmic processing deficit in developmental dyslexia, though it is rather subtle and cannot predict the range of issues associated with developmental dyslexia.

Keywords: rhythm, timing, developmental dyslexia, sensorimotor synchronisation, temporal processing, time-series analyses

1. INTRODUCTION

1.1. Timing and rhythm in developmental dyslexia

Developmental dyslexia (DD) refers to specific difficulties in learning to read, write and spell, which are not accompanied by intellectual disabilities [28]. Despite many years of extensive study, the underlying causes of DD are not fully understood [17]. Sometimes considered "a heterogeneous disorder" [28], a growing number of proposals has been put forward to explain DD as a temporal processing deficit [25], a rhythm impairment [11, 14], or a rhythmic prediction deficit [4, 5, 16].

The hypothesis of a rapid temporal auditory deficit [25] proposes that the key issue in DD is a deficit in processing short or fast-paced acoustic signals. Given that phoneme identification often relies on the ability to perceive formant transitions and other spectral changes on very short timescales [9], a rapid temporal auditory deficit would lead to issues in establishing adequate phonological representations and result in instable phoneme-to-grapheme mappings that are essential for reading.

An alternative account is presented by the temporal sampling framework [11]. This suggests that DD arises as a result of an auditory sensory deficit, where the underlying cause of the deficit is seen in altered brain activity involving neural oscillations specifically within the delta range (specifically 2-2.5 Hz, [12]). These brain oscillations are assumed to lock on to acoustic modulations of speech signals on similar timescales representative of syllable rates in speech signals [10]. According to this framework, DD is the result of a temporal misalignment between neural activity and syllabic rise-times during speech perception. Such misalignment impairs the extraction of linguistic information and leads to a poor encoding of syllables that then gives rise to a poor grapheme-to-phoneme mapping.

Most recently, an inefficient anticipation hypothesis [16, 24] has advanced the idea that DD may reflect a rhythm deficit that stems from an impaired ability to make profitable use of temporal structure in speech that guides anticipation of upcoming linguistic events [4]. Such predictive processes are known to facilitate speech and language processing [29]. Accordingly, DD is suggested to be a consequence of impaired predictive processes in auditory and visual modalities affecting word recognition in both speech [4] and reading [13].

1.2 Sensorimotor synchronisation in DD

Given that several accounts of DD suggest a rhythm deficit [14], empirical methods examining temporal aspects of auditory processing are key to capturing the underlying causes of DD. One of the frequently deployed methods is sensorimotor synchronisation (SMS) that studies rhythm perception and predictive processes often by means of finger tapping in time with a metronome [26]. The paradigm asks

participants to synchronize taps of their dominant index finger with an auditory rhythm sequence, e.g. a metronome beat, and performance analysis shows that participants frequently tap ahead – i.e. anticipate – a metronome beat [22]. SMS thus offers a simple way of testing the predictions of all three accounts of DD outlined above.

Current evidence documenting clear SMS deficits in DD is somewhat conflicting. On the one hand, SMS with a metronome may show deviant patterns in DD, e.g. larger asynchronies or higher variability than in typically developing participants [26]. On the other hand, the timescales at which the group-level deficits are measurable vary across studies. For example, some research provides evidence for an impairment at timescales of 2-2.5 Hz (or 500-400 ms IOI) [26] whereas others at longer timescales of 1.3 Hz (or 750 ms IOI) [16]. However, other studies fail to establish any deviant metronome synchronisation performance in DD, at any timescale [19]. Overall, existing findings do not straightforwardly support the three accounts outlined above, though the conducted analyses of SMS behaviours in DD have also been quite limited in nature.

So far, only few studies investigated rhythmic processing in DD by means of SMS, focusing primarily on tapping variability and mean accuracy [5, 16]. Predictive processes as indicated by negative mean asynchronies as well as the time-course of synchronisation have also received little attention, though individuals are known vary in the ease with which they establish a rhythmic pattern and resume synchronisation [21]. The present study aimed to fill this gap by investigating potential group-level differences in (1) signed asynchronies as an indication of predictive processes [5, 21] and (2) the time-course of synchronisation (as an indication of the ease in establishing the temporal frame of synchronisation). Moreover, there is little evidence if an increased spectral complexity of the auditory prompt (such as spoken monosyllabic words as opposed to a metronome, cf. [20]) may pose more difficulties to individuals with DD.

The language of the present study is Italian which has a large number of school-age children being annually diagnosed with DD (3,5%, [2]), even though Italian has a more consistent phoneme-to-grapheme mapping than English [27].

2. METHOD

2.1. Participants

Seventy adolescent volunteers were recruited at a local high school in Greater Milan area. Forty adolescents (29 F) aged between 14 and 21 years (M= 16.6, SD =1.8) had a formal diagnosis of DD defined according the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM–V, [1]). Thirty typically developing adolescents (23 F) aged between 14 and 20 years (M 16.2, SD=1.5) took part as a control group. Attention was paid to (proportionally) matching age and gender across the two groups of participants. All participants were Italian monolinguals without any hearing impairments, mental health issues, cognitive or developmental comorbidities (though dyscalculia occurred in 60% of the adolescents with DD and was not considered an exclusion criterion).

2.2. Materials

The study's materials consisted of faster- vs. slower-paced metronome and syllable stimuli. The metronome stimuli contained short (50 ms) beats of a metronome, repeated at two IOI. The faster-paced metronome had IOI of 300 ms and the beat repeating 40 times while the slower-paced metronome had IOI of 800 ms and the beat repeating 20 times. The syllable stimuli contained an Italian monosyllable noun [re] (meaning: *king*), produced by a female speaker of Northern Italian. The monosyllabic word had a total duration of 250 ms and was repeated 25 times, with a shorter pause of 300 ms (resulting in a faster-paced sequence with 550 ms IOI) or a longer pause of 800 ms (resulting in a slower-paced sequence with 1050 ms IOI) between repetitions.

2.3. Procedure

Participants were first presented with the metronome and then with the verbal stimuli, slow-paced sequences preceded the faced-paced ones. The task was to listen to each sequence and to start synchronizing with the pacing sounds (i.e., the beats or the syllables) as soon as possible, by tapping with the index finger of the dominant hand on the KAT KTMP1 drum pad placed in front of the participants. At the start of a session, the experimenter (the first author) explained what tapping pressure was needed and which drum-pad area had to be used during the experiment to ensure that taps were recorded correctly.

The stimuli were played back via good-quality headphones (Sennheiser HD 380). Taps were collected using LogicPro [7] running on the MacBook Air 13 laptop. The set-up specific time delay consisted of 34 ms that were subtracted from the tapping data collected, prior to calculating asynchronies.

Participants were tested individually in a quiet room allocated to the experiment by the local school authorities. At the beginning of a session, participants

were asked to self-report their previous musical training by filling in a questionnaire (cf.Rathcke and Lin [19]). The questionnaire collected information if participants ever had any musical training (coded 0 for "no", 1 for "yes"), if they were still actively practicing their hobby (again, coded 0 for "no", 1 for "yes"), at what age they started their training (coded 2 for under 10 y.o., 1 for over 10 y.o.), the number of years they engaged with the hobby and how many instruments they played. Musical training included not only playing an instrument, but also singing or dancing. An individual musicality score was a numerical composite of all questionnaire answers provided. Higher scores were indicative of a higher level of musical training and experience [19].

All participants and their guardians consented to their participation. The study received ethical approval from the Ethics Board of the University of Konstanz (IRB statement 05/2021).

2.4 Data preparation and analyses

Temporal onsets of the metronome and the syllable targets were extracted using Praat [6]. Subsequently, time delays between each metronome beat or syllable onset and the nearby tap were calculated[15], resulting in the following set of measures:

- *mean coefficient of variation (CV)* as an individual measure of SMS consistency, calculated as standard deviation of inter-tap intervals divided by the mean interval duration;
- *mean absolute asynchronies (AA)* as an individual measure of accuracy, reflecting the temporal distance between a target and a tap;
- *mean signed asynchronies (SA)* as an individual measure of predictive processes, or the distance from a tap to the preceding (negative values) or the following (positive values) target.

These measures are widely used in studies of clinical and non-clinical SMS-profiles[8, 20]. In addition, we examined signed asynchronies as time-series data for each participant and stimulus, focusing on potential group-level differences in the ease of establishing and maintaining the temporal frame of synchronization. These analyses provide additional evidence on potential anticipation issues that may be key to DD [14].

3. RESULTS

3.1 Mean-based measures of SMS performance

The best-fit models reported in this section were established using a backward-fitting procedure and estimated using REML and *nloptwrap* optimizer [3]. They included *participant* as random effect.

The best-fit linear mixed-effects regression with *CV* as the dependent variable did not include *group* either as a main effect or as an interaction, for either type of stimuli. *IOI* was the only significant predictor showing that tap consistency was higher for *the faster IOI* compared to the *slower IOI* across both groups and stimulus types (metronome: ß = 0.25, SE = 0.08, z = -2.86, p <.01; verbal: ß = 0.12, SE = 0.033, z = 3.64, p <.001).

The best-fit linear mixed-effects regression with *AA* as the dependent variable included *group* in interaction with *IOI* as a significant predictor, for both stimulus types. The dyslexic group had higher *AA* (indicative of lower accuracy) than the control group, but only in the *slower IOI* compared to the *faster IOI* (metronome: ß = 63.28, SE = 9.12, z = 6.93, p <.001; verbal: ß = 24.74, SE = 9.05, z = 2.73, p <.01, see Figure 1). There was no significant effect of musical training.

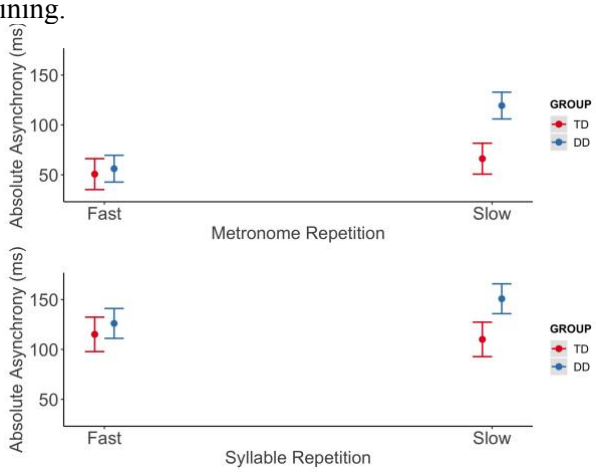


Figure 1: Group-level estimates of absolute asynchronies from best-fit models for synchronization with metronome (top panel) and syllable (bottom panel) sequencies.

The best-fit linear mixed-effects regression with *SA* as the dependent variable did not include *group* either as a main effect or as an interaction, for either type of stimuli. Only *IOI* was a significant predictor for verbal (but not metronome) stimuli, showing that all participants anticipated the upcoming target more in the *faster IOI* than the *slower IOI* condition (ß = 53.52, SE = 7.48, z = 7.15, p <.001).

3.2 Time-series measures of SMS performance

Generalised additive mixed models (GAMMs) were fit to the time-series data measuring *SA* with each synchronisation target. The models were estimated using ML and *outer newton* optimizer and included *participant* as random effect [23]. Best-fit GAMMs did not included *group* as the significant smoothing term at either fast pace (*300 ms IOI*: edf = 4.9, Ref. df = 5.6, F = 1.8, p = 0.2; *550 ms IOI*: edf = 1.06, Ref. df = 1.01, F = 0.24, p = 0.6) or slow pace (*800 ms IOI*: edf = 1, Ref. df = 1, F = 0.05, p = 0.8; *1050 ms IOI*:

edf = 1.1, Ref. df = 1.5, F = 0.02, p = 0.9; see Figures 2 and 3).

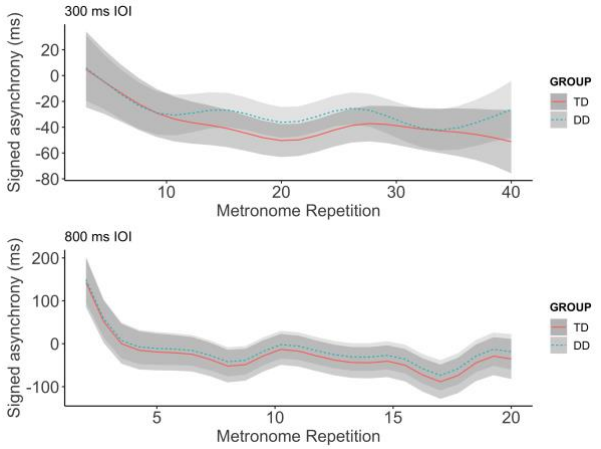


Figure 2: Best-fit GAMMs estimating SA in the *metronome* condition as a function of *group*.

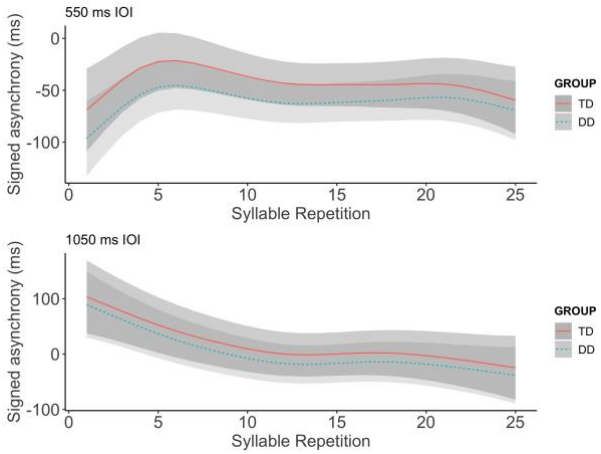


Figure 3: Best-fit GAMMs estimating SA in the *syllable* condition as a function of *group*.

Visual comparisons of the time-series data indicate that all participants experienced more difficulties with maintaining stable tapping when synchronizing with metronome than syllable sequencies and when synchronizing at slower than faster pace. Across all experimental conditions, dyslexic participants showed anticipatory tapping responses to a similar degree control participants did.

4. DISCUSSION

The purpose of the current study was to investigate rhythmic processing in Italian adolescents with and without developmental dyslexia. The method of the study involved a SMS paradigm used in previous research [19, 26]. Participants had to synchronize with slow and fast-paces stimuli containing either metronome beats or syllables [19]. Overall, our results support the idea of a rhythm impairment in DD [11, 24, 25], even though reliable group-level differences are very subtle and limited to slow-paced rhythmic prompts.

We observed significant group-level differences in synchronisation with both metronome and syllable sequences, but these differences were specific to the slow-paced stimuli. Time-series analyses did not help to capture these group-level differences, though here, we focused primarily on signed asynchronies to test the hypothesised anticipation issues in the dyslexic group[18]. Overall, sensorimotor performance of the two groups differed exclusively in their absolute mean asynchronies. No other measure of rhythmic synchronization showed a significant effect of group.

Current findings do not straightforwardly support existing accounts of DD. The rapid temporal auditory deficit hypothesis [25] predicts an impaired rhythmic performance primarily for fast-paced stimuli. However, we found the opposite pattern. The assumptions of the temporal sampling framework [11] are not supported for two reasons. First, the framework predicts main issues at 2-2.5 Hz (i.e., around 500-400 ms IOI), but not at slower tempi. However, the present study shows key differences between the groups at longer IOIs. Second, the framework suggests that rhythmic processing issues stem from an auditory issue with encoding complex amplitude modulations. Accordingly, tapping long with acoustically complex syllables should cause more difficulties for dyslexic participants than tapping with a metronome beat. The prediction was not borne out by the present study. Given that we tested a relatively large sample, it is rather unlikely that the lack of statistical evidence to support group-level differences is due to low power.

Finally, the inefficient anticipation hypothesis [16] would predict lack of anticipation of upcoming rhythmic events, with taps lagging behind targets during SMS [4]. The results of the present time-series analyses clearly show that Italian adolescents with DD can indeed tap *ahead* of rhythmic targets similar to typically developing control group. When synchronizing with slow-paced prompts, both groups lagged behind the first few targets but then started anticipating. When synchronizing with fast-paced prompts, all participants tended to anticipate from the onset of their tapping.

The study of developmental dyslexia has seen a number of controversial findings. A previous review of the literature concluded that disordered production and/or perception of prosodic phenomena in dyslexia has only been consistently attested in tasks involving metalinguistic judgements or other factors that increased processing difficulty (e.g., the presence of background noise, high short-term memory load, or time pressure on decision making) [18]. The results of the present study speak in support of the suggestion that a rhythmic impairment is unlikely to be the primary cause of the issues observed in dyslexia.

7. REFERENCES

[1] American, Psychiatric, and Association, 2013, Diagnostic and Statistical Manual of Mental Disorders. DSM-5, ed: American Psychiatric Publishing.

[2] Barbiero, C. *et al.*, 2019, The lost children: The underdiagnosis of dyslexia in Italy. A cross-sectional national study, *PLoS One,* vol. 14 no. 1 e0210448

[3] Bates, D., Mächler, M., Bolker, B., and Walker, S., 2015, Fitting Linear Mixed-Effects Models Using lme4, *Journal of Statistical Software,* vol. 67 no. 1 1 - 48

[4] Beach, S. *et al.*, 2022, The Neural Representation of a Repeated Standard Stimulus in Dyslexia, *Frontiers in Human Neuroscience,* vol. 16

[5] Bégel, V., Dalla Bella, S., Devignes, Q., Vandenbergue, M., Lemaître, M. P., and Dellacherie, D., 2022, Rhythm as an independent determinant of developmental dyslexia, *Dev Psychol,* vol. 58 no. 2 339-358

[6] *Boersma, P. and Weenink, D., 2007, PRAAT: Doing phonetics by computer (Version 5.3.51)*.

[7] *Cousins, M. and Hepworth-Sawyer, R., 2014, Logic Pro X*. Oxon: Focal Press, Abingdon.

[8] Dalla Bella, S. *et al.*, 2017, BAASTA: Battery for the Assessment of Auditory Sensorimotor and Timing Abilities, *Behav Res Methods,* vol. 49 no. 3 1128-1145

[9] Dorman, M., Cutting, J., and Raphael, L., 1975, Perception of temporal order in vowel sequences with and without formant transitions, *J. of exp psychology. Human perc and performance,* vol. 104 2 147-153

[10] Flanagan, S. and Goswami, U., 2018, The role of phase synchronisation between low frequency amplitude modulations in child phonology and morphology speech tasks, *J of the Acoustic Soc of America,* vol. 143 no. 3 1366-1366

[11] Goswami, U., 2011, Temporal sampling framework for developmental dyslexia, *Trends in cognitive sciences,* vol. 15 3-10

[12] Keshavarzi, M. *et al.*, 2022, Atypical delta-band phase consistency and atypical preferred phase in children with dyslexia during neural entrainment to rhythmic audio-visual speech, *Neuroimage Clin,* vol. 35 103054

[13] Kirkby, J. A., Barrington, R. S., Drieghe, D., and Liversedge, S. P., 2022, Parafoveal processing and transposed-letter effects in dyslexic reading, *Dyslexia,* vol. 28 no. 3 359-374

[14] Ladányi, E., Persici, V., Fiveash, A., Tillmann, B., and Gordon, R. L., 2020, Is atypical rhythm a risk factor for developmental speech and language disorders?, *Wiley Interdiscip Rev Cogn Sci,* vol. 11 no. 5 e1528

[15] *Natick, 2022a, MATLAB*. Massachusetts: The MathWorks Inc.

[16] Pagliarini, E., Scocchia, L., Granocchio, E., Sarti, D., Stucchi, N., and Guasti, M. T., 2020, Timing anticipation in adults and children with Developmental Dyslexia: evidence of an inefficient mechanism, *Scientific Reports,* vol. 10 no. 1 17519

[17] Protopapas, A., 2014, From temporal processing to developmental language disorders: mind the gap, *Philos Trans R Soc Lond B Biol Sci,* vol. 369 no. 1634 20130090

[18] Ramus, F. and Ahissar, M., 2012, Developmental dyslexia: the difficulties of interpreting poor performance, and the importance of normal performance, *Cogn Neuropsychol,* vol. 29 no. 1-2 104-122

[19] Rathcke, T. and Lin, C.-Y., 2021, Towards a Comprehensive Account of Rhythm Processing Issues in Developmental Dyslexia, *Brain Sci,* vol. 11 no. 10 1303

[20] Rathcke, T., Lin, C.-Y., Falk, S., and Dalla Bella, S., 2019, When language hits the beat: Synchronising movement to simple tonal and verbal stimuli, presented at the International Congress of Phonetic Sciences, Melbourne, Australia.

[21] Rathcke, T., Lin, C.-Y., Falk, S., and Dalla Bella, S., 2021, Tapping into linguistic rhythm, *Laboratory Phonology: Journal of the Association for Laboratory Phonology,* vol. 12(1):11 1-32

[22] Repp, B. H. and Su, Y.-H., 2013, Sensorimotor synchronization: A review of recent research (2006–2012), *Psychonomic Bulletin & Review,* vol. 20 no. 3 403-452

[23] Sóskuthy, M., 2017, Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction,

[24] Taha, J., Stucchi, N., Pagliarini, E., and Guasti, M. T., 2022, *What is wrong with rhythm in Developmental Dyslexia? The inefficient anticipation hypothesis*.

[25] Tallal, P., 1980, Auditory temporal perception, phonics, and reading disabilities in children, *Brain Lang,* vol. 9 no. 2 182-198

[26] Thomson, J. M. and Goswami, U., 2008, Rhythmic processing in children with developmental dyslexia: auditory and motor rhythms link to reading and spelling, *J Physiol Paris,* vol. 102 no. 1-3 120-129

[27] Traficante, D. and Burani, C., 2014, List context effects in languages with opaque and transparent orthographies: a challenge for models of reading, *Front Psychol,* vol. 5 1023

[28] van Bergen, E., van der Leij, A., and de Jong, P. F., 2014, The intergenerational multiple deficit model and the case of dyslexia, *Front Hum Neurosci,* vol. 8 346

[29] Zettersten, M., 2019, Learning by predicting: How predictive processing informs language development, in *Patterns in Language and Linguistics*, Busse and Moehlig-Falke Eds. Berlin, Boston: De Gruyter Mouton, 255-288

A PILOT FIELDWORK ULTRASOUND STUDY OF TONGUE SHAPE VARIABILITY IN CHILDREN WITH AND WITHOUT SPEECH SOUND DISORDER

Amy Smith, Maria Dokovova, Eleanor Lawson, Anja Kuschmann, Joanne Cleland

University of Strathclyde
amy.smith@strath.ac.uk, maria.dokovova@strath.ac.uk, eleanor.lawson@strath.ac.uk, anja.kuschmann@strath.ac.uk, joanne.cleland@strath.ac.uk

ABSTRACT

Children with Speech Sound Disorders (SSD) have been described as having increased tongue shape variability during speech. However, most studies do not compare this variability to typically developing children (TD) using ultrasound tongue imaging (UTI). Open access corpora suitable for answering this question are scarce. This pilot study addresses both the theoretical question of whether variability differs between TD children and children with SSDs; and the feasibility of a potential solution for acquiring ultrasound data quickly at a public science exhibition. We compare tongue shapes during multiple repetitions of various consonants in TD children across different ages, and children with SSD, using mean Nearest Neighbour Distance (NND). Results suggest no significant effect of age in the TD group. Children with SSD had significantly higher tongue shape variability than TD children. Field data collection was a viable method for collecting UTI data, despite some limitations.

Keywords: speech sound disorders, ultrasound, variability, articulation, field methods

1. INTRODUCTION

Speech Sound Disorder (SSD) is an umbrella term encompassing many subtypes that affect a child's ability to acquire clear or intelligible speech. With a prevalence of between 2.3% and 24.6% [1], SSDs make up a large proportion of Speech and Language therapy caseloads. Disagreement in how these disorders are classified creates challenges for Speech and Language Therapists (SLTs) to diagnose and treat SSD. However, SSDs can be broadly divided into those of a phonological nature (e.g., phonological disorders) and those of an articulatory or phonetic nature (articulation disorders, childhood apraxia of speech). Without understanding the true nature of the difficulty and due to limitations of current investigation practices, crucial information can be missed. Instrumental techniques such as Ultrasound Tongue Imaging (UTI) can negate this problem by providing detailed information about articulation, instead of relying on perceptual tools such as phonetic transcription, which have reliability issues [2].

In UTI, a probe is placed under the chin to image the midsagittal, or coronal tongue surface, displaying movement in real time. UTI is relatively inexpensive and portable, and the use of UTI in SSD research is growing [3]. Use of UTI is no longer limited to speech labs and hospital rooms and, as such, it is possible to collect data in a variety of settings. Despite the challenges field collection can present, it opens the possibility for larger N studies.

UTI has been shown to aid in the detection of covert errors in children with SSDs [4]. One of these is increased variability, meaning repetitions of the same sound in the same phonetic context are produced differently. Here we define variability as sub-phonemic, that is, repetitions are typically phonetically transcribed using the same symbol. This variability is therefore subtle and can only be detected using a modality such as UTI, in addition to e.g. auditory analysis [4]. Increased variability is suggestive of subtle motor difficulties [5]. This has strong implications for SLTs and could affect differential diagnosis and decisions surrounding therapy approaches. However, for SLTs to fully understand the articulatory variability associated with SSDs, it is essential to quantify articulatory variability in typically developing children of different ages.

In contrast to previous evidence, which suggested little to no difference in articulatory gestures made by children when compared to adults [6], recent literature suggests protracted development of speech motor control well into adolescence [7]. Younger typically developing children show greater articulatory variability than their older counterparts. What is not yet known is what this developmental trajectory looks like and what the typical range is. UTI has previously been used to measure lingual variability by comparing sets of tongue curves and supports the idea of protracted development [8]. However, as [9] points out, "there are still crucial gaps in our knowledge about the developmental paths taken by children to adult-like speech motor control".

Crucially for SLTs, we do not yet know how this compares to children with SSDs.

2. RESEARCH QUESTIONS

This is an exploratory pilot study, aiming to describe articulatory variability, as measured using mean Nearest Neighbour Distance (NND) [10] in typically developing (TD) children, and children with SSD across four key consonants: /t/, /k/, /ɹ/, /s/. Higher mean NND are suggestive of increased variability.

This study explores the following questions:

- Is it feasible to collect ultrasound data from children at a public event in a museum?
- Do older TD children have lower articulatory variability than younger TD children?
- Do TD children have lower articulatory variability than age-matched children with a Speech Sound Disorder?

3. METHODS

3.1. Participants and recruitment

TD children were recruited at a public engagement event (ethics approval from the University of Strathclyde). This free event was held at the Glasgow Riverside Museum on a Saturday afternoon. Recordings were made at a public stall, in a large exhibition room. Inclusion criteria were: spoken English, age 5-15 years, and no developmental conditions. Any child who approached the exhibit with their family was invited to be recorded and parents gave consent via an online survey on their mobile phone or a tablet, which also collected demographic information. They were recorded by one of five researchers present at the exhibit. SSD data was selected from the UltraSuite corpus [11]. These are one-off recordings from children with unresolved SSD attending SLT services. After the science event, the TD children were age-matched with 10 children with SSD (2 female). The children with SSD had a diagnosis of any SSD of unknown origin, given by their SLT. The exclusion criteria were: current or repaired cleft lip and/or palate, no spoken English, severe or profound hearing loss, severe or profound learning disability. They received the same instructions and produced the same set of consonants.

3.2. Data collection feasibility

To address the feasibility of data collection at a public event, the following topics were considered:

- How many parents and children would consent to recordings in a public venue?
- Would children tolerate the headset and duration of the recording?
- Would the recordings yield usable data?

3.3. Stimuli

Each child was instructed to produce 10 repetitions of all English consonants in an /aCa/ environment. Some children produced fewer than 10 repetitions. The consonants /s/ /t/ /k/ /ɹ/ were used for analysis in this pilot, as they represent high frequency segments in English, and therefore have a crucial impact on intelligibility [12]. Participants were given an orthographic representation and asked to read or were given a verbal demonstration and asked to imitate.

3.4. Ultrasound Set-Up

UTI data of the TD children were acquired using a Micro machine controlled via a Windows laptop running AAA v. 2.20.01 [13]. The echo return data were recorded at ~100 fps over a field of view of 162°. This field of view allowed the greatest view of the tongue, including both the hyoid and mandible shadows. A 5–8 MHz 10 mm radius microconvex ultrasound probe was stabilized with an Ultrafit lightweight plastic headset [14]. Audio data was recorded using an Audio Technica3350 microphone attached to the headset. The same set up and materials were used for both groups, however the SSD data were collected with Sonospeech [15] in a quiet clinical setting by SLTs trained in UTI data collection, rather than at a public event.

3.5. Analysis

3.5.1. Spline fitting

The tongue surface was identified by fitting splines at the point of maximal lingual gesture of each consonant, using AAA. The frame of interest, containing the maximal lingual gesture was identified manually, using the audio, waveform, and spectrogram. Regions of interest were observed frame by frame (stop consonant closure phases and whole durations of continuants). Splines were added using semi-automatic edge detection in AAA. Regions of low confidence were manually trimmed and the automatic AAA correction “snap-to-fit” applied.

3.5.2. Mean Nearest Neighbour Distance

The variability of tongue shapes was analysed using the mean Nearest Neighbour Distance (NND) procedure [10], implemented in AAA (version 220.04.01) at 80% minimum confidence. In this procedure splines are first analysed in pairs. The

distance from each point on one spline to the nearest point on the other spline is measured then all distances between the two splines are averaged. Using this procedure, we compared each spline from one half of the repetitions of one consonant from one child to each of the splines in the other half of repetitions of that consonant from that child, resulting in a mean and standard deviation of NND per consonant per child.

3.5.3. Statistical analyses

Data processing and visualisation were performed with "tidyverse" [16] in RStudio (version 2022.02.1+461 for Windows) [17]. Statistical analyses were carried out using the lme4 package [18] (data and code available at: https://osf.io/3byvp/).

4. RESULTS

4.1. Feasibility results and descriptive statistics

Age group	Consonant	Mean NND	SD NND	n participants
[7-8]	/k/	0.10	0.04	4
	/ɹ/	0.10	0.05	4
	/s/	0.07	0.02	4
	/t/	0.08	0.03	5
[9-11]	/k/	0.07	0.02	5
	/ɹ/	0.07	0.03	5
	/s/	0.05	0.03	5
	/t/	0.07	0.02	5
[14]	/k/	0.07	0.04	1
	/ɹ/	0.07	0.02	1
	/s/	0.07	0.02	1
	/t/	0.05	0.02	1

Table 1: Mean and SD of Nearest Neighbour Distance (cm) per consonant per age group in TD children.

Eighteen children and their parents consented to participate at the public event, over a four-hour period. Of those, 7 datasets were excluded due to: corrupt files (n=1), missing audio (n=5), and 1 who had autism (identified after the event from the survey), leaving 11 TD participants (5 female). There was no age match for a 14-year-old TD participant in the SSD sample, so the analyses comparing both groups were run with and without that participant. Anecdotally, we consider this a higher rate of data wastage than previous lab-based recordings of children. Their mean age was 9 years (range 7-14). All children who consented to the recording completed the tasks, were enthusiastic, and tolerated the headset and duration of the recording. Duration of recordings ranged from 3-9 minutes (Mean=6 minutes). Recording varied in length as not all participants completed 10 repetitions of each stimulus. Table 1 provides mean and SD NND (cm) per consonant per age for the TD children.

4.2. Age

A linear mixed effects model was run with mean NND as an outcome variable for TD children and age (in years, centred around the mean) as a continuous predictor. The model included random intercepts per participant. Age was not a significant predictor for NND (Table 1).

Predictor	β	*t*- value	*p*- value
Intercept	0.07	12.10	< 0.001
Age	-0.002	-0.61	0.56

Table 1: Model results from the statistical analysis of age and mean NND in TD children.

4.2. Diagnosis

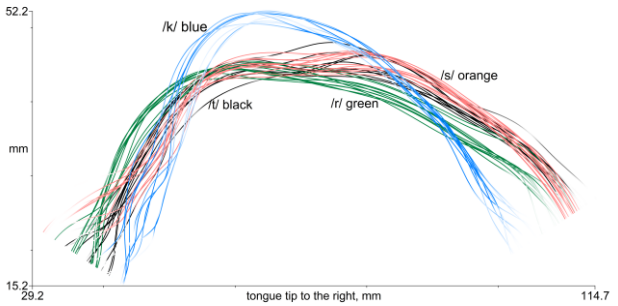


Figure 1: Tongue splines for /k/, /t/, /ɹ/, /s/ of a 10-year-old male child with SSD.

Lastly, we explored whether TD children had lower articulatory variability than age-matched children with SSD. We ran a linear mixed effects model with mean NND as an outcome variable and Diagnosis (TD as reference) as a predictor. The model included random intercepts per participant. Results suggest that SSD diagnosis is associated with increased mean NND per consonant (Table 1). Excluding the 14-year-old TD participant did not change direction or significance of the results. Fig. 1 and 2 show all tongue splines per consonant from a TD child and a child with SSD. Fig. 3 shows the differences in variability by group.

Predictor	β	*t*- value	*p*- value
Intercept	0.14	8.79	< 0.001
Diagnosis (TD vs SSD)	0.10	4.41	< 0.001

Table 2: Model results from the statistical analysis of diagnosis and mean NND.

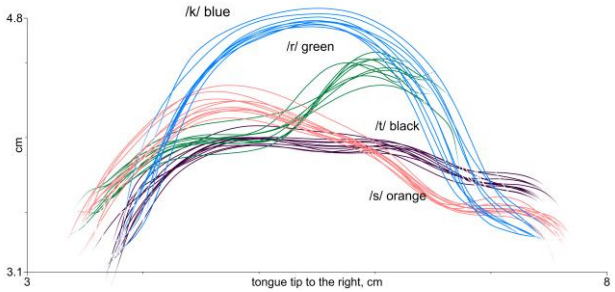


Figure 2: Tongue splines for /k/, /t/, /ɹ/, /s/ of a 10-year-old male TD child.

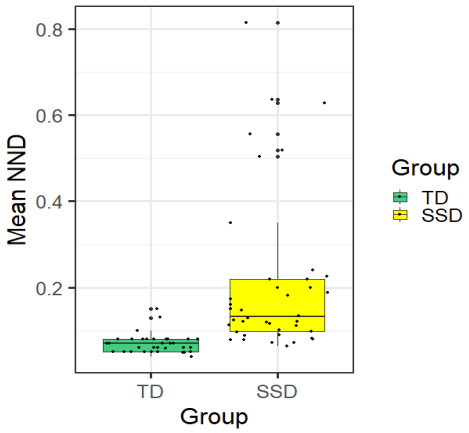


Figure 3: Differences in mean NND between TD children and children with SSD.

5. DISCUSSION

This paper provides preliminary normative data on tongue shape variability in TD children; compares this to children with SSD and describes the feasibility of collecting data at a public engagement event. Despite greater losses of data, it was feasible to collect recordings that can be used for analysis at these kinds of events. For future research we suggest a recording protocol be established to avoid discrepancies between researchers undertaking recordings and minimise experimenter error. In addition, future field research should consider recording the age of the children in years and months, for increased precision in investigating age-effects.

Tongue shape variability was measured using mean Nearest Neighbour Distance (NND) per consonant, across /s/, /t/, /k/, /ɹ/. In addition, the data was explored to investigate the relationship between age and tongue shape variability, and the relationship between having an SSD diagnosis and variability in consonant production. The results were inconclusive about the relationship between age and mean NND in TD children. However, there was a significant effect of SSD diagnosis. Children with SSD had greater tongue shape variability than children without a diagnosis.

These results differ from some previous findings. For example, Zharkova, et al. [19] report greater within-speaker variability in children than in adults, opposite to their expectation of greater variability associated with larger oral cavities. Namasivayam, et al. [20] discuss evidence demonstrating that between-articulator coordination tends to complete its development in child speech before within-articulator coordination, specifically the coordination of multiple tongue gestures. The present study has a limited number of participants (potentially contributing to the null result for the effect of age on consonant variability), which prevents a more detailed by-consonant analysis. Future research could explore differences in age-related variability of tongue gestures with only a primary lingual gesture e.g. /t/, compared with those with a primary and secondary lingual gestures such as /r/ or /l/.

The greater tongue shape variability in the SSD group is consistent with an account of variability being linked to motor skill immaturity or disorder. While earlier research such as [6] did not observe increased variability during repetitions in most children with SSD, this might also be related to differences in instrumentation. Electropalatography, which was used by Hardcastle, et al. [6] only shows which part of the palate the tongue is in contact with. The advantage of using UTI is that it allows measurements of the whole midsagittal surface of the tongue, from the root to near the tip. Future research could explore whether any specific area of the tongue contributes disproportionately to increased variability in the SSD group.

In conclusion, it was feasible to collect data from children at a public event. The likelihood of larger than usual amounts of data wastage can be mitigated by developing a recording protocol before the event. Results of this pilot suggest increased variability in children with SSD compared to TD children. Larger studies are needed to examine the effects of age.

6. ACKNOWLEDGMENTS

This research was funded by grants from the Engineering & Physical Sciences Research Council [EP/P02338X/1] and the Economic and Social Research Council [ES/W003244/1]. For Open Access, the author has applied a CC licence to any Author Accepted Manuscript (AAM) version arising from this submission. We would like to thank the Glasgow Riverside Museum and the children, for participating.

7. REFERENCES

[1] J. Law, J. Boyle, F. Harris, A. Harkness, and C. Nye, "Prevalence and natural history of primary speech and language delay: findings from a systematic review of the literature," *International Journal of Language & Communication Disorders,* vol. 35, no. 2, pp. 165-88, 2000, doi: 10.1080/136828200247133.

[2] D. Sell, "Issues in perceptual speech analysis in cleft palate and related disorders: a review," *International Journal of Language & Communication Disorders,* vol. 40, no. 2, pp. 103-121, 2005, doi: https://doi.org/10.1080/13682820400016522.

[3] E. Sugden, S. Lloyd, J. Lam, and J. Cleland, "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *International Journal of Language & Communication Disorders,* vol. 54, no. 5, pp. 705-728, 2019, doi: 10.1111/1460-6984.12478.

[4] E. Sugden and J. Cleland, "Using ultrasound tongue imaging to support the phonetic transcription of childhood speech sound disorders," *Clinical Linguistics & Phonetics,* pp. 1-20, 2021, doi: 10.1080/02699206.2021.1966101.

[5] J. C. Vick *et al.*, "Data-Driven Subclassification of Speech Sound Disorders in Preschool Children," *Journal of Speech, Language, and Hearing Research,* vol. 57, no. 6, pp. 2033-2050, 2014, doi:10.1044/2014_JSLHR-S-12-0193.

[6] W. J. Hardcastle, R. A. M. Barry, and C. J. Clark, "An Instrumental Phonetic Study of Lingual Activity in Articulation-Disordered Children," *Journal of Speech, Language, and Hearing Research,* vol. 30, no. 2, pp. 171-184, 1987, doi: doi:10.1044/jshr.3002.171.

[7] B. Walsh and A. Smith, "Articulatory Movements in Adolescents," *Journal of Speech, Language, and Hearing Research,* vol. 45, no. 6, pp. 1119-1133, 2002, doi:10.1044/1092-4388(2002/090).

[8] D. Abakarova, S. Fuchs, and A. Noiray, "Developmental Changes in Coarticulation Degree Relate to Differences in Articulatory Patterns: An Empirically Grounded Modeling Approach," *Journal of Speech, Language, and Hearing Research,* vol. 65, no. 9, pp. 3276-3299, 2022, doi:10.1044/2022_JSLHR-21-00212.

[9] N. Zharkova, N. Hewlett, and W. J. Hardcastle, "Coarticulation as an Indicator of Speech Motor Control Development in Children: An Ultrasound Study," *Motor Control,* vol. 15, no. 1, pp. 118-140, 2011.

[10] N. Zharkova and N. Hewlett, "Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /ɑ," *Journal of Phonetics,* vol. 37, no. 2, pp. 248-256, 2009/04/01/ 2009, doi: https://doi.org/10.1016/j.wocn.2008.10.005.

[11] A. Eshky *et al.*, "UltraSuite : a repository of ultrasound and acoustic data from child speech therapy sessions," presented at the Interspeech 2018, 2018. [Online]. Available: https://strathprints.strath.ac.uk/64824/.

[12] L. Gilner and F. Morales, "Functional load: Transcription and analysis of the 10,000 most frequent words in spoken English," *The Buckingham Journal of Language and Linguistics,* vol. 3, pp. 135-162, 2010.

[13] *Articulate Assistant Advanced Ultrasound Module user manual, revision 2.16.,* Articulate Instruments, 2014.

[14] M. Pucher, N. Klingler, J. Luttenberger, and L. Spreafico, "Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings," *Speech Communication,* vol. 123, pp. 83-97, 2020/10/01/ 2020, doi: https://doi.org/10.1016/j.specom.2020.07.001.

[15] *Sonospeech user manual*, Articulate Instruments, 2023.

[16] H. Wickham, Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H., "Welcome to the Tidyverse," *Journal of Open Source Software,* vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.

[17] *R: A Language and environment for statistical computing*. (2010). R Foundations for Statistical Computing, Vienna, Austria.

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1 - 48, 10/07 2015, doi: 10.18637/jss.v067.i01.

[19] N. Zharkova, N. Hewlett, and W. J. Hardcastle, "An ultrasound study of lingual coarticulation in /sV/ syllables produced by adults and typically developing children," *Journal of the International Phonetic Association,* vol. 42, no. 2, pp. 193-208, 2012, doi: 10.1017/S0025100312000060.

[20] A. K. Namasivayam, D. Coleman, A. O'Dwyer, and P. van Lieshout, "Speech Sound Disorders in Children: An Articulatory Phonology Perspective," (in English), *Frontiers in Psychology,* Review vol. 10, 2020-January-28 2020, doi: 10.3389/fpsyg.2019.02998.

A Speech Therapy Animation and imaging Resource (STAR)

Eleanor Lawson†, Joanne Cleland†, Jane Stuart-Smith‡, Brian Aitken‡, Janet Beck •

† University of Strathclyde,
‡ University of Glasgow,
• Queen Margaret University, Edinburgh

ABSTRACT

The Speech Therapy Animation and imaging Resource (STAR) is a Web-based anglophone Speech and Language Therapy resource, that hosts over 1,000 videos showing the visible and hidden movements of speech articulators, using (i) ultrasound tongue imaging (UTI) with lip camera video, (ii) magnetic resonance imaging (MRI), and MRI-based animation. The *STAR* resource focuses on children's speech therapy and comprises of (1) a teaching and training website, which will host UTI videos of disordered and nondisordered child speech, nondisordered adult speech, and MRI videos of modelled speech (2) an in-clinic website hosting vocal-tract animations to aid speech therapy and speech practise at home.

STAR will be the first resource of its kind, providing Speech and Language Therapy students and teachers with real examples of imaged and animated vocal tract movement in disordered and nondisordered speech.

Keywords: Speech and Language Therapy, Ultrasound-Tongue-Imaging.

1. VISUAL APPROACHES TO SPEECH THERAPY

SLT training and practice has been dominated by auditory and descriptive approaches. However, there is increasing evidence that vocal-tract visualisation can provide breakthroughs in speech error remediation, see §§1.1-1.2. Recent research into visually-aided speech therapy has focussed on two main remediative tools that reveal hidden vocal organ movement (i) Visual Biofeedback (VB) (ii) Visual animated articulatory models (VAAMs).

1.1 Visual Biofeedback

VB provides the speech therapist and client with real-time visual information about the position and movement of the client's tongue during speech production, through e.g. UTI, using standard medical ultrasound machines**.** UTI VB therapy helps with both diagnosis and remediation of speech-sound errors. Research involving VB therapy shows breakthroughs for clients with persisting speech sound disorders - errors that show no improvement, despite years of conventional speech therapy, see [1], [2], [3]; including those caused by congenital sensorineural hearing impairment [4], and childhood apraxia of speech (CAS) [5]. However, it has been noted that viewing visual UTI video examples of target articulations plays a significant role in speech remediation, even without the feedback element of the client seeing their own tongue movement [3]. This finding may be due to the fact that vocal-tract visualisation is a more intuitive method of understanding correct articulations, avoiding complex descriptions that clients, especially children, may struggle to understand.

1.2 Visual animated articulatory models (VAAMs)

VAAMs show midsagittal cut-aways of the head and vocal tract, or partially-transparent 3-D heads, revealing the movements of hidden articulators. They provide the speech and language therapist (SLT) and their client with a visual representation of a speech target. VAAMs have proved to be effective in therapy for children with hearing loss [6], and for remediation of persisting speech-sound disorder [7]. While VB requires investment in technology and training, VAAM-based therapy is comparatively cheap. Roxburgh [9] compared the effects of UTI VB therapy versus VAAM-based therapy (the mobile/i-pad app "Speech Trainer 3D" [8], for children with submucous cleft palate. Results showed no advantage of VB over the VAAM [9].

Research using UTI VB tools show huge potential in Speech Therapy, but UTI is used extremely rarely by SLTs, especially in the U.K. Studies involving VAAMs in clinical settings also show that the provision of accurate visual models of the hidden speech articulators can have a breakthrough role in remediation of speech-sound disorders, but quality and accuracy of commercial VAAMs is a potential issue, i.e. it is not always clear how animated articulatory movements have been generated and whether they are based on real vocal-tract movement. The addition of normative

articulatory visual models (imaged and animated) to the SLTs arsenal, has the potential to make speech therapy easier, reduce caseloads, and, crucially, to improve client outcomes. Additionally, animated and imaged speech materials of nondisordered and disordered speech has the potential to greatly improve SLT training; providing teachers and students with accurate information about how the vocal organs move inside the vocal tract in both nondisordered and disordered speech.

2. FROM SEEING SPEECH TO STAR

2.1 Clinical use of Seeing Speech

Seeing Speech www.seeingspeech.ac.uk [10] was developed as a Phonetics teaching and learning resource, hosting UTI and MRI videos of speech, and animation (based on MRI vocal-tract recordings frame-by-frame). However, *Seeing Speech* is not specifically designed for SLT needs; it does not have an optimal interface for use in clinic and is missing many important resources for SLTs. A user survey conducted between 2019 – 2020 revealed that 18% (N=140) of *Seeing Speech* user-respondents were either SLTs, SLT teachers, researchers, or students, using the site for: continuing professional development (CPD); revision, in preparation for sessions with patients; in clinic and to recommend to clients and their carers. Specific feedback from SLTs on the usefulness of *Seeing Speech* stated that viewing speech articulations helped with diagnosis; feedback to clients/carers; provided a visual support for children to better explain tongue placement, and improved understanding of non-English speech sounds. An informal survey of local Scottish SLTs, asked how *Seeing Speech* had changed practice. SLTs reported greater understanding of articulatory movement for themselves, and improved description and feedback to clients and their carers. SLTs reported that priorities for a clinical resource, similar to *Seeing Speech*, would be: a more child/carer-friendly interface; addition of missing items such as affricates; animated examples of speech sounds for clients and their carers to view; materials to help clients practise at home; addition of the *extensions to the International Phonetic Alphabet for disordered speech* (extIPA) chart [11]; imaged examples of different disorders – e.g. cleft palate, speech sound disorders, cerebral palsy etc.; imaged examples of speech from different age groups; electropalatographic images; different ways of producing some sounds e.g. /s/ made with the tongue tip down; more animations showing the difference between /ʃ/ and /s/; more speech imaging examples in the coronal plane, especially for sibilants and /r/.

2.2 The STAR website

Funding was obtained from the *Economic and Social Research Council*'s *Secondary Data Analysis Initiative* to create a vocal-tract imaging resource, after the model of *Seeing Speech*, but with a children's Speech and Language Therapy focus: the *Speech Therapy Animation and Imaging Resource* (STAR). *STAR* will be launched in November 2023 and will comprise of two websites:

(1) www.seeingspeech.ac.uk/speechstar/ - a site for Speech and Language Therapy teaching and training that will provide users with access to imaged and animated video examples of midsagittal vocal tract movements in disordered and nondisordered speech.

(2) www.speechstar.ac.uk - an in-clinic site for child Speech and Language Therapy use that will provide users with quick and easy access to midsagittal vocal-tract animations of English speech samples to: help explain vocal-organ movement to clients and their carers; identify speech-production targets, and to aid practise at home.

An additional aim of the *STAR* project is to make curated UTI recordings, collected during research projects for over fifteen years, available to key user groups who might benefit from having access to them (see table 1).

Table 1: Projects from whose corpora UTI/MRI video were selected for the STAR databases. Dark grey cells indicate clinical-focussed child-speech projects.

Project title, date, imaging modality	Speaker age	No. of speakers
Dynamic Dialects (2014-15) (UTI+lip video/MRI)	Adult (18-65)	54
Changes in shape, space and time (2016-19) (UTI+lip video)	Adult (18-50)	32
Ultrax (2011-14) (UTI+lip video)	Child (5-13)	68
Ultraphonix (2015-16) (UTI)	Child (6-13)	20
Visualising Speech (2017-18) (UTI)	Child (3-15)	39
Ultrax2020 (2017-20) (UTI)	Child (5-16)	50

2.2.1 STAR content

The *STAR* teaching and training site will contain:

i. Clinically-focussed introductory materials explaining how the resource can be used; describing and showing how UTI, the MRI recordings were collected, and how the speech animations were created.
ii. An interactive clickable extIPA chart showing midsagittal vocal tract movements in modelled disordered speech sounds, imaged using UTI, MRI and in animated form.
iv. The Edinburgh MRI modelled speech corpus – a new database of modelled speech recorded with MRI (see §2.2.2 below).
iii. A set of filterable UTI (and lip camera) speech databases: (1) a non-disordered speech database of key words, (2) a disordered child-speech sentences database, and (3) a child speech-error database (containing words and phrases).

STAR databases are underpinned by the MySQL database management system [12]. The UTI databases are filterable by some or all of the following filters: *speaker code*; *age*; *sex*; *speech-disorder diagnosis*; *speech-error type*; *speech sound*; *place of articulation*; *syllable position*. Users will be able to select up to four videos to play side-by-side for comparison. All videos are accompanied by metadata: demographic and stimulus/target. The child speech-error database also contains a phonemic transcription of the target word and phonetic transcription of the utterance produced.

The *STAR* in-clinic site contains a reduced set of resources, more suited to use in-clinic and with child clients. This site will contain speech animations, which offer a more easily interpretable stimulus for children and their carers:

i. Animations showing the speech organs and how they move.
ii. Animations of a set of key English consonant sounds in syllables.
iii. Speech animations to support practise at home.

2.2.2 The Edinburgh MRI modelled speech corpus

Funding was obtained from the *Royal Society of Edinburgh* (RSE) and also financial support was given by *NHS Research Scotland* (NRS), through the *Edinburgh Clinical Research Facility*, to create a new MRI-based speech corpus for *STAR*. The Edinburgh MRI modelled speech corpus (EMMSC) consists of dynamic midsagittal vocal-tract MRI recordings of English speech produced by model talker Prof. Janet Beck. Materials consist of English minimal sets (recordings of consonants that are common English therapy targets: [s], [ʃ], [tʃ], [t] and [k] in onset and coda position with a variety of high, low, front and back vowel nuclei), and polysyllabic words containing these key consonant sounds.

MRI recordings were made at the *Edinburgh Imaging Facility*, *Royal Infirmary Edinburgh* (EIF RIE), using a 3 Tesla Siemens Prisma MRI scanner (Figure 1), with a head coil receiver.



Figure 1: Siemens Prisma MRI scanner at EIF RIE.

The recording protocol used for the EMMSC corpus maintains a similar number of midsagittal scans per second (6.25) to the MRI scans currently available on *Seeing Speech* (6.5), but increases image resolution by 70% to 320x320 pixels (Figure 2).

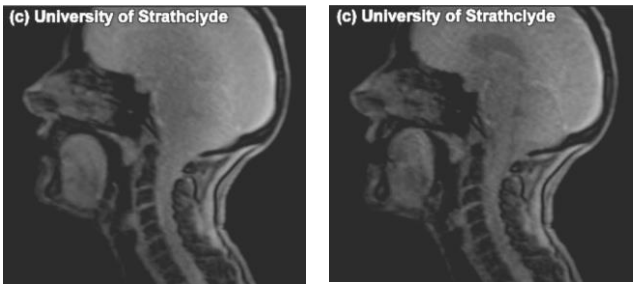


Figure 1: (left) an EMMSC midsagittal scan of a [t] constriction in "Betty", (right) a [k] constriction in "Becky".

The DICOM files obtained were turned into video using ImageJ software [13], then further manipulated (increased brightness and contrast, noise-cancelled audio and copyright marks added) using VirtualDub [14] with frameserver editor AviSynth [15].

Speech audio was recorded inside the MRI scanner using an OptoAcoustics FOMRI III dual-channel, fibre-optic microphone system. The microphone was fixed to the MRI head coil using

velcro straps and the microphone was positioned as close as possible to the speaker's lips using the microphone's goose neck. Further noise-cancelling was carried out using Audacity [16]; however, it was not possible to completely remove extraneous noise relating to the MRI machine without removing too much of the speech signal. For the purposes of exemplifing speech sounds to SLT students, teachers and child Speech and Language Therapy clients, we plan to add MRI videos with "clean" audio at a later date, i.e. having our model talker match her productions in the MRI machine, while being recorded in a sound studio. We will use the denoised MRI audio recordings as stimuli for our model talker to copy, while recording her simultaneously with UTI. We then examined the lingual articulation in the MRI and the UTI recording to make sure that they were comparable.

3. EVALUATION

STAR has been co-designed from the outset with input from a Clinical Advisory panel of SLTs, and SLT trainers. Users will also be invited to evaluate the resources in a beta version of *STAR*, followed by a period of revision and improvement before the main launch with an online survey attached to the sites for ongoing feedback.

4. SUMMARY

Speech and Language Therapy research over the past decade has highlighted that adding vocal-tract imaging, or animation of the movements of the visible and hidden speech articulators, to the SLTs arsenal can be game changing for remediation of persistent speech sound disorders. Building on the success of *Seeing Speech, STAR* will be a web-based resource providing SLTs, SLT teachers, students, clients and carers access to curated databases of imaged and animated vocal-tract movement to improve understanding of how the vocal organs move in disordered and nondisordered speech.

7. REFERENCES

[1] Adler-Bock, M., Bernhardt, B., Gick, B. and Bacsfalvi, P., 2007. The use of ultrasound in remediation of North American English /r/ in 2 adolescents. *AJSLP*, 16(2), 128-139.

[2] Byun, T., Hitchcock, E. and Swartz, M., 2014. Retroflex Versus Bunched in Treatment for Rhotic Misarticulation: Evidence From Ultrasound Biofeedback Intervention. *JSLHR*, 57(6), 2116-2130.

[3] Cleland, J., Scobbie, J.M. and Wrench, A., 2015. Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *CLP*, 29(8-10), 575-597.

[4] Bernhardt, B., Gick, B., Bacsfalvi, P. and Ashdown, J., 2003. Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *CLP*, 17(3), 199-216.

[5] Preston, J., Brick, N. and Landi, N., 2013. Ultrasound Biofeedback Treatment for Persisting Childhood Apraxia of Speech. *AJSLP*, 22(4), 627-643.

[6] Massaro, D.W. and Light, J., 2004. Using visible speech to train perception and production of speech for individuals with hearing loss. *JSLHR*, 47(2), 304-320.

[7] Fagel, S and Madany, K., 2008. A 3D virtual head as a tool for speech therapy for children. Proceedings of the ninth annual conference of the international speech communication association (Interspeech 2008) 22-26 Sept. 2008, Brisbane, Australia, 2643-2646.

[8]Smartyears, 2017. Speech Trainer 3D. https://www.smartyearsapps.com

[9] Roxburgh, Z., 2018. Visualising articulation: real-time ultrasound visual biofeedback and visual articulatory models and their use in treating speech sound disorders associated with submucous cleft palate. (unpublished PhD thesis). Edinburgh: Queen Margaret University.

[10] Nakai, S., Beavan, D., Lawson, E., Leplatre, G., Scobbie, J.M. and Stuart-Smith, J., 2018. Viewing speech in action: speech articulation videos in the public domain that demonstrate the sounds of the International Phonetic Alphabet (IPA). *ILLT*, 12(3), 212-220.

[11] ICPLA (2008) Extensions to the International Phonetic Alphabet for disordered speech. https://www.internationalphoneticassociation.org/sites/default/files/extIPAChart2008.pdf

[12]Widenius, M. 2009. MariaDB v10.6.5 https://mariadb.org/

[13]Rasband, W.S. 1997. ImageJ v1.50i https://imagej.nih.gov/ij/

[14]Lee, A. (1998) VirtualDub v1.10.4. https://www.virtualdub.org/

[15] Rudiak-Gould, B., van Eggelen, E., Post, K., Berg, R., Brabham, I. (2000). AviSynth, v2.5. http://avisynth.nl

[16]Audacity Project. (2005). Audacity. v2.0.5 https://www.audacityteam.org/

Probing production-perception relationships in children with myotonic dystrophy

Lucie Ménard[1], Marie Bellavance-Courtemanche[2], Paméla Trudeau-Fisette[1], Camille Vidou[1], and Cristina Uribe[1]

[1]Université du Québec à Montréal, Département de linguistique, Montréal (CANADA)
[2]*Le Code Morse, Granby (CANADA)*

menard.lucie@uqam.ca, marie@lecodemorse.com, ptrudeaufisette@gmail.com, vidou.camille@gmail.com, uribe.cristina@gmail.com

ABSTRACT

Myotonic dystrophy, a neurodegenerative disease that causes muscle weakness and difficulties with muscle relaxation after contraction, frequently affects orofacial articulatory dynamics, leading to decreased speech intelligibility, particularly in children. We investigated the effects of myotonic dystrophy on sensorimotor relationships in children's speech by studying compensations for a lip-tube perturbation. Nine French-speaking children diagnosed with myotonic dystrophy and nine age-matched typically developing children were recruited. They were asked to produce repetitions of the vowel /u/ with and without a 17-mm-diameter tube inserted between the lips. A synchronized ultrasound system, Optotrak motion tracking system, and audio recording system were used to track lip and jaw displacement, as well as tongue shape and position. Combined with articulatory and acoustic results, perceptual ratings of the produced vowels suggest that children with myotonic dystrophy rely more on auditory feedback than their typically developing peers.

Keywords: clinical speech, production-perception relationships, lip-tube perturbation, speech development.

1. INTRODUCTION

It is well established that speech production and perception processes interact with each other. From a theoretical point of view (for a review see [1]), during speech production, the central nervous system generates an efferent signal, corresponding to the estimated auditory (or somatosensory) consequences of the articulatory movements. This efferent signal is then compared to the actual auditory (or somatosensory) feedback signal produced by the speaker (reafferent sensory input). When a discrepancy is observed, an error signal occurs and the motor system adapts to minimize the discrepancy between the expected and actual auditory or somatosensory feedback. In the course of speech development, children first learn to control phonemic goals through feedback-based mechanisms (auditory and somatosensory). Although several studies have used the artificial perturbation paradigm to probe the contribution of each of those mechanisms (feedback-based and feedforward) to speech production in typically developing speakers, much less is known about speakers with a motor control impairment. In this paper, we report on a study of speech production and perception using a lip-tube perturbation with children diagnosed with myotonic dystrophy.

1.1. Myotonic dystrophy

Myotonic dystrophy (MD) is a genetic neuromuscular disorder characterized by muscle weakness (myopathy) and difficulty relaxing muscles (myotonia). Myotonic dystrophy type 1, also known as Steinert's disease, is the most common form of muscular dystrophy. Its global prevalence is estimated at 1 in 20,000 inhabitants, but its proportions are greater in certain regions of Quebec, particularly the Estrie region, where the approximate prevalence is 1/1,500 ([2]), and in the Charlevoix and Saguenay-Lac-St-Jean regions, where the prevalence is approximately 1/600 ([2, 3]). One of the most common features in people with MD is bilateral weakness of the facial and jaw muscles. Children with MD often display reduced labial and lingual displacement in non-speech tasks, and produce speech sounds perceived as being less intelligible than those produced by typically developing peers ([4, 5]). To our knowledge, no data on French-speaking children are available. To gain a deeper understanding of the speech motor abilities of children diagnosed with this medical condition, a group of speech language pathologists from CHU Ste-Justine in Montreal reached out to us to gather quantitative data on their patient cohort. Consequently, the study was collaboratively developed by both clinicians and researchers.

1.2. The lip-tube perturbation paradigm

Perturbation experiments have been widely used to study reliance on sensory feedback ([6]). A lip tube is an appropriate perturbation to affect the geometry of the vocal tract, since it increases or decreases the lip area compared with the typical configuration. In the French vowel /u/, for example, the lips are rounded so that the lip area is small (around 0.3 cm^2 for adults; [7]); the associated tongue position is elevated towards the velopalatal region. It has been found that, when a 2.5-cm-diameter lip tube keeps them from rounding their lips, French adult speakers who are asked to produce /u/ cannot immediately compensate for the perturbation. After a learning session, during which a gradual improvement in the quality of /u/ was observed, some speakers proved better able to compensate than others; their compensation strategies involved a high and backward movement of the tongue towards the velo-pharyngeal region as a mean to correct for the auditory feedback error ([7, 8]).

A more recent study using a lip tube with children ([9]) showed that 4-year-olds could successfully compensate for a lip-tube perturbation (as assessed by the intelligibility of their produced vowels), but the compensation was based on trial and error and did not seem to be guided by a systematic error correction mechanism. We would expect a systematic correction mechanism to be associated, first, with a continuous evolution towards a better production of the vowel and, second, with the end of the correction process once the production is considered to be appropriate. It is important to note that, in both children and adults, compensatory responses were never complete: acoustic and articulatory values in the lip-tube condition remained significantly different from values in the normal pre-perturbed condition. This difference could be ascribed to the competing demands of somatosensory feedback, which limits articulatory displacement (which is required to compensate for auditory feedback error).

The objective of this study is to investigate the compensatory response of children with MD and typically developing children to a lip-tube perturbation during production of the French vowel /u/. The effects of the compensation will be assessed with acoustic, articulatory, and perceptual analyses.

2. METHODS

A speech production task and a perceptual task were administered to different groups of participants.

2.1. Speech production

2.1.1. Participants

Nine children ranging in age from 6 years old to 9 years old (mean age: 7.6; 5 males) were recruited at Ste-Justine paediatric hospital in Montreal (the “MD group”). These children had received a diagnosis of MD several years ago. Nine age-matched and gender-matched typically developing children were also recruited. They reported no speech or language disorder. All children were native speakers of Canadian French and were screened for auditory thresholds. The study received approval from both UQAM and the hospital's Institutional Review Boards. Additionally, all parents of the participating children provided written consent for the experiment.

2.1.2. Experimental procedure

Children had to produce 30 tokens of isolated /u/, in three conditions: first, without the lip tube, in a normal condition (referred to as the normal pre-perturbed condition). In the second condition, children had to produce the vowel /u/ 10 times with a 1.7-cm-diameter lip tube between their lips (referred to as the lip-tube condition). This diameter was chosen based on the results of a previous study with typically developing children ([9]). In the third condition, children were asked to produce another set of 10 /u/ vowels without the lip tube (referred to as the normal post-perturbation condition).

2.1.3. Data analysis

Articulatory and acoustic data were recorded using a combined ultrasound system (Sonosite 180Plus) and Optotrak system (NDI Certus). The synchronized acoustic signal was also recorded using a SHURE unidirectional microphone, with a sampling frequency of 22050 Hz. Ultrasound images corresponding to each vowel’s midpoint were selected. Tongue contours were semi-automatically tracked using the GetContour program (described in [10]) and corrected for head movements using the Optotrak measures (HOCUS; [11]). Measures of tongue maximal curvature index (MCI) and tongue height, as described in [10], were also extracted. Acoustic data were analysed using Praat ([12]). The first (F1) and second (F2) formant frequencies were also extracted at the vowel midpoint, using the Burg algorithm embedded in Praat. In order to account for between speaker variability in vocal tract size, all data were normalized by dividing the value of a given trial by the mean value measured in the normal pre-pre-perturbation phase.

Linear mixed-effects models ([13]) were built using R. The dependent variables were articulatory and acoustic measures (one model for each variable), and the independent variables corresponded to the experimental phase and the speaker group. The participant was included in the models as a random effect (slope).

2.2. Perceptual assessment

For each participant, the first two and last two trials of each experimental phase (normal pre-perturbation, lip-tube perturbation, normal post-perturbation) were selected and used as stimuli in a speech perception test administered to 20 French-speaking adults. Listeners did not report any speech or language disorder or any hearing deficits. They ranged in age from 20 to 29 years old. Stimuli were presented in randomized orders. Listeners' first task was to indicate, by selecting the corresponding button on a computer screen, whether the vowel they had heard was /u/ or not (open choice). Then they had to rate the quality of the vowel on a scale ranging from 1 (poor) to 7 (excellent). The percentage of /u/ responses and average quality ratings were calculated for each trial. Linear mixed effects models were built for each of the dependent variables (percent correct and quality rating), with experimental phase and speaker group as the independent variables and participant as the random effect.

3. RESULTS

3.1. Speech production results

The mean F2 values for each trial, averaged across speakers within each group, are presented in Figure 1. As the graph shows, control and MD groups produced higher values of F2 in the lip-tube condition than in the normal pre-perturbation condition ($\chi^2(2)=38.62$; $p<0.001$). This pattern can be interpreted as reflecting incomplete compensation. However, unlike the control group, the MD group had significantly reduced F2 values during the 10 trials with the lip tube. This effect was significant ($\chi^2(5)=15.58$; $p<0.01$). In the normal post-perturbation condition, no significant difference was observed between groups, nor between this condition and the normal pre-perturbed condition.

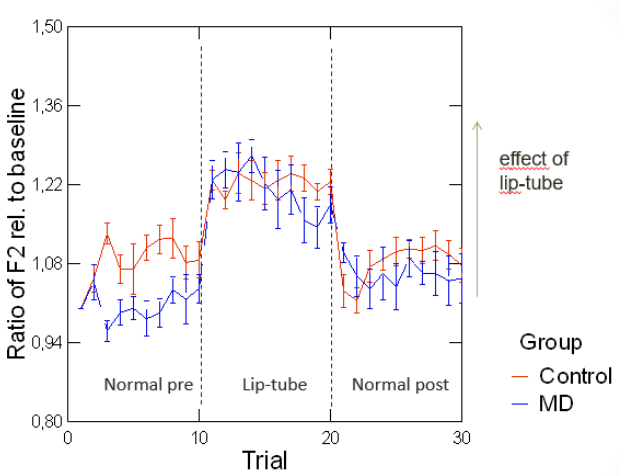


Figure 1: Mean normalized values of F2, averaged across speakers within each group.

The mean F1 values, depicted in Figure 2, show a similar pattern. Children with MD produced significantly lower F1 values than typically developing children in the lip-tube condition ($\chi^2(2)=37.91$; $p<0.001$). Furthermore, as was the case for the F2 values (Figure 1), F1 values decreased significantly over the course of the perturbed condition for speakers of the MD group only ($p<0.01$).

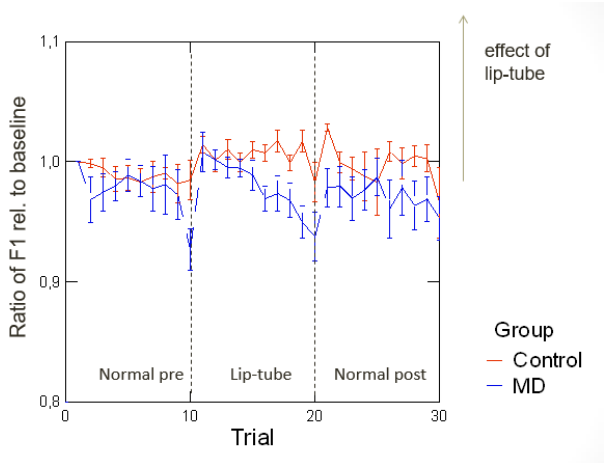


Figure 2: Mean normalized values of F1, averaged across speakers within each group.

In the articulatory domain, tongue displacement showed a similar trend (Figure 3). The expected compensatory manoeuvre –elevating the tongue– was produced in the lip-tube condition by children with MD, but not control children ($\chi^2(5)=21.42$; $p<0.001$). Again, a significant relationship between trial number and tongue height for the 10 trials of the perturbed condition ($p<0.01$) was observed only for the MD group.

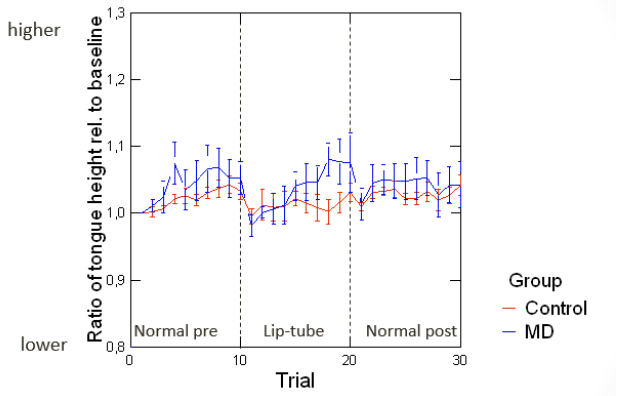


Figure 3: Mean normalized values of tongue height, averaged across speakers within each group.

3.2. Perceptual results

The perceptual test did not reveal any significant differences between the percentages of correct responses for /u/ produced by the children with MD and the typically developing children. However, mean quality-rating values exhibit different patterns, as shown in Figure 4. A significant effect of condition is noted, with /u/ produced in the lip-tube condition being perceived significantly better than /u/ produced in either of the two normal conditions ($\chi^2(2)=24.15$; $p<0.001$). Furthermore, children in the MD group produced /u/ vowels that corresponded overall to lower quality than their typically developing peers ($\chi^2(3)=25.42$; $p<0.001$). Importantly, a significant effect of the interaction between trial and group was noted in the lip-tube condition ($\chi^2(5)=28.42$; $p<0.001$): for the MD group only, the last two trials received higher scores than the first two trials, confirming the tendencies described above in the articulatory and acoustic domains.

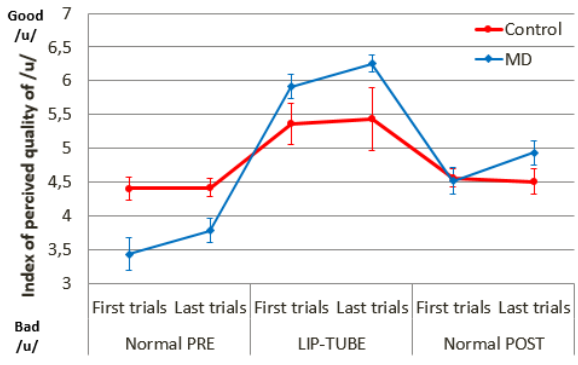


Figure 4: Mean quality rating values of /u/, averaged across speakers within each group.

4. DISCUSSION

In this study, we used a lip-tube perturbation to examine production-perception relationships in children with motor deficits (myotonic dystrophy) compared to typically developing children. Our results showed that children with MD exhibited better abilities than age-matched typically developing children in using their auditory feedback to compensate for the lip-tube perturbation. Indeed, despite the reduced motility of the tongue reported in the literature, children in this group were able to modify their tongue position and reduce formant values in the expected direction in the perturbed condition. This pattern emerged gradually from the first to the last trials, suggesting a sensory-based error correction mechanism. Importantly, typically developing children did not display this compensatory response, in line with previous studies ([9]).

Our results also provide information about the reliance on auditory and somatosensory feedback in children with MD. In line with Lametti et al., 2012 ([14]), we can consider the magnitude of the compensatory response as an index of sensory reliance in both speaker groups. Since children with MD produced a greater compensation in the acoustic dimension, they seem to rely more heavily on auditory feedback than typically developing children when producing French /u/. However, since the MD group produced larger tongue displacements compared to the normal pre-perturbed condition than the control group, they induced a larger discrepancy between the actual somatosensory feedback and the somatosensory feedback predicted by the central nervous system (see section 1). Children with MD thus attached less weight to somatosensory feedback than typically developing children.

The results reported here could have important consequences for rehabilitation strategies clinicians use for children with myotonic dystrophy. Indeed, since an artificial perturbation such as a lip tube elicits larger displacement of the tongue, more exercises of this type, taking advantage of the strong reliance of children with MD on auditory feedback, could be administered to ensure these children explore the entire lingual articulatory space and map it to their sensory experiences. Further experiments to assess the clinical value of such perturbed conditions are currently under way.

Acknowledgements
This work was supported by an SSHRC grant and an NSERC grant. Thanks to Zofia Laubitz for copy editing the paper.

5. REFERENCES

[1] Parrell, B., Lammert, A. C., Ciccarelli, G., & Quatieri, T. F. 2019. Current models of speech motor control: A control-theoretic overview of architectures and

properties. *The Journal of the Acoustical Society of America*, *145*(3), 1456-1481.
[2] Perron, M., Veillette, S., Desbiens, F., and Mathieu, J., 1986. « Comportements sociodémographiques des individus atteints de dystrophie myotonique ». *Cahiers québécois de démographie*, 15, 75-110.
[3] Desjardins, P. 2012. *Évaluation de la capacité fonctionnelle chez des patients atteints de la dystrophie myotmûque de type 1.* (Maîtrise en médecine expérimentale Mémoire), Université Laval et Université du Québec à Chicoutimi.
[4] Sjögreen, L. 2010. *Orofacial dysfonction in children and adolescents wilh myotonic dystrophy type 1: evaluation and intervention.* (Ph. D.), The Sahlgrenska Academy at University of Gotbenburg, Gotbenburg, Sweden.
[5] Sjogreen, L., Engvall, M., Ekstrom, A B., Lohmander, A, Kiliaridis, S., & Tulinius, M. 2006. Orofacial dysfunction in children and adolescents with myotonic dystrophy. *Deve!opmental Medicine & Child Neurology, 49,* 18-22.
[6] Lindblom, B. and Sundberg, J. 1971. "Acoustical consequences of lip, tongue, jaw and larynx movement," *J. Acoust. Soc. Am. 50,* 1166-1179.
[7] Savariaux, C., Perrier, P., and Orliaguet, J.-P. 1995. "Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production," *J. Acoust. Soc. Am. 98,* 2428-2442.
[8] Savariaux, C., Perrier, P., Orliaguet, J.-P., and Schwartz, J.-L. 1999. "Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis," *J. Acoust. Soc. Am. 106,* 381-393.
[9] Ménard, Lucie, PERRIER, Pascal, SAVARIAUX, Christophe, AUBIN, Jérôme et THIBEAULT, Mélanie 2008: "Compensation strategies for a lip-tube perturbation of French [u] : an acoustic and perceptual study of 4-year-old children ", Journal of the Acoustical Society of America, 124, 1192-1206.
[10] Dawson, K. M., Tiede, M. K. & Whalen, D. H. 2016. Methods for quantifying tongue shape and complexity using ultrasound imaging, Clinical Linguistics & Phonetics, 30:3-5, 328-344, DOI: 10.3109/02699206.2015.1099164
[11] Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., & Vatikiotis-Bateson, E. 2005. The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research, 48,* 543–553.
[12] Boersma, P. and Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/.
[13] Bates, D., Maechler, M., Bolker, B., & Walker, S. 2013. *Lme4: Linear mixed-effects models using Eigen & S4*. R package version 1.0-5.
[14] Lametti, D. R., Nasir, S. M., & Ostry, D. J. 2012. Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience*, *32*(27), 9351-9358.

EFFECT OF AUDITORY FEEDBACK ON HIGH LEXICAL TONES IN DEAF ADULTS WITH COCHLEAR IMPLANTS

Tsung-Lun Alan Wan

National Yang Ming Chiao Tung University, Taiwan
tsunglun.wan@hotmail.com

ABSTRACT

The study investigated how auditory feedback can affect fundamental frequencies (F0) of high lexical tones (high-level tone and high-falling tone) produced by deaf speakers of Taiwan Mandarin. Observations made in previous research on non-tonal languages indicated that auditory enhancement in deaf people leads to F0-lowering. Results of the present study showed that high lexical tones in Taiwan Mandarin did not always invite F0-lowering when there is auditory feedback, given that the standard phonetic realization of these tones invites F0-raising.

Keywords: clinical phonetics, cochlear implants, lexical tones, gender, clinical sociolinguistics

1. INTRODUCTION

Speech production relies on both somatosensory and auditory feedback to calibrate it [1]. Auditory feedback can affect both segmental and suprasegmental features. Previous research on deaf speech shows that after cochlear implantation, voice fundamental frequency (F0) drops [2, 3, 4]. It is however noticed that other studies have reported heterogeneous findings: there exists individual variation among deaf users of cochlear implants (CIs) in F0 variation post-implantation [5].

Tonal languages rely on F0 variation to convey semantic meanings. An effect of auditory feedback on F0 may interact with the lexical tone quality and undermine phonemic distinction. Previous research has explored how short-term auditory deprivation may affect F0 in non-tonal languages: a 24-hour deprivation leads to F0 raising, and returning to auditory feedback results in F0 dropping [6]. However, it is not clear how a short-term auditory deprivation would affect lexical tone qualities. Particularly, for lexical tones which occupy higher F0 register to be differentiated from other lexical tones in lower F0 register, the relation between auditory feedback and F0-lowering may be challenged. The present paper works with deaf speakers of Taiwan Mandarin, where four lexical tones exist – Tone 1 (high-level), Tone 2 (low-rising/low-level), Tone 3 (low-falling)[7], and Tone 4 (high-falling). This study investigates how the two high lexical tones – Tone 1 and Tone 4 [8]– are affected by auditory feedback among deaf speakers.

2. RESEARCH DESIGN

Two sentence lists written in traditional Chinese characters were prepared. Participants read aloud the sentence lists before recordings to make sure there was no unfamiliar Chinese character.

Each sentence list had fourteen sentences. The first word in each sentence was bracketed to invite focus. To control for prosody, only the first word in each sentence is examined in the present study.

In the first experiment (On-Off), participants read aloud the first sentence list firstly with the CI and then without the CI. The second reading occurred right after the CI was turned off.

In the second experiment (Off-On), the same participants read aloud the second sentence list without the CI and then with the CI. The second reading occurred right after the CI was turned on.

3. PARTICIPANTS

A total of thirteen deaf participants (six women and seven men) were recruited from Taipei, Taiwan. All spoke Mandarin as the dominant language and received oral education. Five used both single-sided CI and a hearing aid, i.e., bimodal users; eight used only single-sided CI. Participants had profound (sensorineural) deafness. Age and duration of cochlear implant use (in years) are shown in Figure 1.

4. STATISTICS

F0 values are normalized into semitones by the following formula:

$$(1) \quad semitone = \frac{(12ln(F0/Reference))}{ln(2)}$$

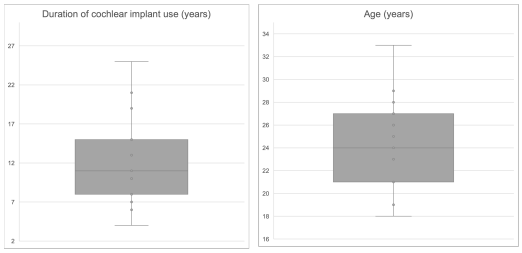


Figure 1: Demographical information

As this study involves inter-speaker comparisons, a speaker-specific reference is adopted [9]. Each speaker's mean Tone 1 is used as the reference value, with auditory feedback in the first On-Off experiment.

Taking F0 trajectories into consideration, this study does not look at single-point F0 values. Instead, F0 values are automatically extracted at each 10% of the segment interval by a Praat script. To exclude F0 tracking errors, only values from 20% to 80% are included in data analysis. Generalized additive models (GAMs) are fit to the data [10]. Smooth terms are included for scaled time, with speaker and word as random effects.

In each experiment, participants read aloud the same sentence-list twice. Thus, linguistic factors are not included. To control for prosody, vowel duration is included as a factor, and it can reveal how speakers engage with F0 when tokens are longer (more articulatory planning). Previous research shows that bimodal users who use both CI and hearing aid perform lexical tones in a more standard way, compared to unimodal users using single-sided CI [11]. In this study, bimodal users only turned off the CI during auditory deprivation, with the hearing aid on. We predict that bimodal users would be less affected by auditory deprivation. Unimodal/bimodal use is considered as a factor, so is duration of CI use. To test whether there is any sociolinguistic engagement with the experiment [12], gender is considered.

5. FINDINGS

5.1. Auditory deprivation

5.1.1. F0 height of high-level tone

Table 1 shows significant terms that are relevant to the present study. We focus our analysis on the two three-way interaction terms among auditory feedback, vowel duration, and CI use or gender.

The durational effect is significantly larger when there is auditory feedback, in male speakers and bimodal users. This durational effect is significantly reduced in female speakers and single-sided CI users, failing to achieve significance in female speakers (Figure 2 and 3).

	Estimate	Std.Error	*t* value	*p* value
(Intercept)		n.s.		
Gender = woman x **Vowel duration**	1.92	0.58	3.30	<0.001
Auditory feedback = on x **Vowel duration**	6.74	0.47	14.31	<0.001
Device = Single-sided CI x **Auditory feedback** = on x **Vowel duration**	-3.65	0.77	-4.69	<0.001
Gender = woman x **Auditory feedback** = on x **Vowel duration**	-6.08	0.84	-7.19	<0.001

Table 1: Summary of the best-fitting GAM predicting normalized F0 trajectory of Tone 1 in the On-Off experiment

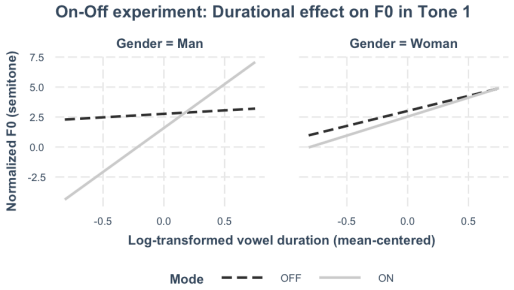


Figure 2: Durational effect by gender and auditory feedback in the On-Off experiment (model prediction)

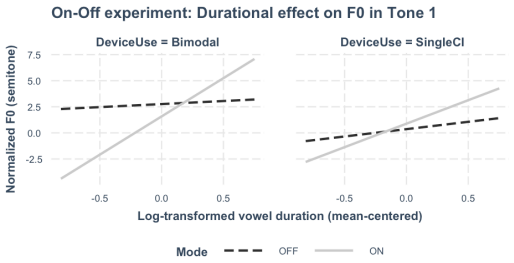


Figure 3: Durational effect by device use and auditory feedback in the On-Off experiment (model prediction)

Longer tokens receive more articulatory planning. When there is auditory feedback, bimodal users and single-sided CI users (to a less extent) treat a higher F0 in tone 1 as a standard speech style; shorter tokens invite less attention from speakers and exhibit the mechanistic effect of auditory feedback on F0.

Men exhibit a vowel durational effect with auditory feedback; no effect of vowel duration is observed among men without auditory feedback. That is, there is a mechanistic effect of auditory feedback on F0 in male speakers. In contrast, regardless of auditory feedback, women realize longer tokens with higher F0 values.

5.1.2. F0 height of high-falling tone

	Estimate	Std.Error	*t* value	*p* value
(Intercept)		n.s.		
Vowel duration	-2.98	0.59	-5.03	<0.001
Vowel duration x **Gender** = woman	2.45	0.34	7.12	<0.001
Vowel duration x **Duration of CI use**	-0.15	0.02	-6.82	<0.001
Vowel duration x **Device** = Single-sided CI	1.82	0.24	7.40	<0.001
Gender = woman x **Auditory feedback** = on x Vowel duration	1.64	0.43	3.76	0.001

Table 2: Summary of the best-fitting GAM predicting normalized F0 trajectory of Tone 4 in the On-Off experiment

Tone 4 is a high-falling tone. Different from Tone 3, a mid-falling tone, phonemically, Tone 4 occupies a higher F0 register. However, current findings show that longer tokens tend to receive lower F0 in male speakers.

As shown in Table 2, when a person uses CI for more years, their shorter tokens are realized with higher F0. Previous research argues that CI is a source of F0 lowering; yet, the current finding reveals that high F0 in short Tone 4 tokens is not a direct mechanistic product of auditory feedback. Previous research shows that longer duration of CI use contributes to more standard tone production [13]. The current finding may be a case where speakers with longer CI duration have learned a standard speech style of Tone 4 production.

There is no effect of auditory feedback on F0 in Tone 4 in men. Shorter tokens invite higher F0. The effect of vowel duration is significantly reduced among female speakers (Figure 4), but the direction is retained. There does not seem a clear effect of short-term auditory deprivation on F0 in Tone 4 tokens.

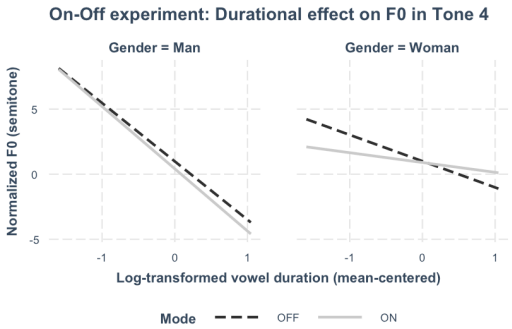


Figure 4: Durational effect on normalized F0 by gender and auditory feedback in the On-Off experiment

5.2. Auditory enhancement

5.2.1. F0 of high-level tone

Without auditory feedback, longer tokens invite lower F0 (Table 3). This durational effect under auditory deprivation is significantly reduced in single-sided CI users, failing to achieve significance. When the speakers return to auditory enhancement, an inverse relation between vowel duration and F0 emerges: longer tokens invite higher F0. Likewise, this interaction is significantly reduced in single-sided CI users.

	Estimate	Std.Error	*t* value	*p* value
(Intercept)		n.s.		
Vowel duration	-3.74	0.24	-15.33	<0.001
Vowel duration x **Auditory feedback** = on	5.71	0.31	18.02	<0.001
Gender = woman x **Vowel duration**	1.48	0.23	6.21	<0.001
Device = Single-sided CI x **Vowel duration**	3.01	0.25	11.75	<0.001
Device = Single-sided CI x **Auditory feedback** = on x Vowel duration	-4.99	0.39	-12.55	<0.001
Gender = woman x **Auditory feedback** = on x Vowel duration	-1.14	0.36	-3.51	0.001

Table 3: Summary of the best-fitting GAM predicting normalized F0 trajectory of Tone 1 in the Off-On experiment

5.2.2. F0 of high-falling tone

The model (Table 4) reveals a significant effect of auditory feedback in general: when speakers return to auditory enhancement, F0 is raised. This effect is significantly larger in longer tokens, and the interaction between vowel duration and auditory

	Estimate	Std.Error	*t* value	*p* value
(Intercept)		n.s.		
Auditory feedback = on	3.21	1.36	2.35	0.01
Vowel duration	-4.52	0.45	-9.98	<0.001
Vowel duration x **Auditory feedback** = on	4.15	0.40	10.32	<0.001
Vowel duration x **Gender** = woman	0.82	0.34	2.39	0.01
Vowel duration x **Device** = Single-sided CI	3.31	0.37	8.72	<0.001
Gender = woman x **Auditory feedback** = on x **Vowel duration**	-1.78	0.47	-3.76	0.001
Device = Single-sided CI x **Auditory feedback** = on x **Vowel duration**	-3.84	0.51	-7.43	0.001

Table 4: Summary of the best-fitting GAM predicting normalized F0 trajectory of Tone 4 in the Off-On experiment

feedback is significantly reduced in female speakers, while remaining in the same direction.

The interaction between vowel duration and auditory feedback is significantly reduced in single-sided CI users – no clear trend of vowel durational effect during auditory deprivation (Figure 5).

The F0-raising during auditory enhancement is inconsistent with the prediction that CI leads to F0-lowering. A Tone 4 token with high F0 is aligned with standard Mandarin speech, showing that the effect of auditory feedback does not universally lead to a lower F0; instead, it is constrained by Mandarin phonology.

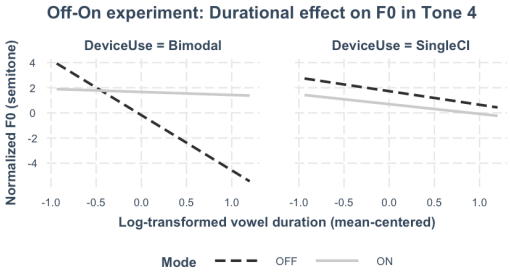


Figure 5: Durational effect by CI use and auditory feedback in the Off-On experiment

6. DISCUSSION

In the On-Off experiment, F0 in Tone 1 is less affected by auditory feedback in single-sided CI users; in the Off-On experiment, F0 in both tones is less affected by auditory feedback in single-sided CI users. It is inconsistent with the prediction that bimodal users, with the hearing aid on during auditory deprivation, would be less affected by auditory deprivation. Future research is needed to investigate why bimodal users, compared to unimodal users, are more affected by the removal of auditory feedback generated by CI alone.

Gender plays an important role in the present study. Tone 1 in women is nearly unaffected by auditory deprivation in the On-Off experiment, but Tone 1 in men shifts toward higher F0 in short tokens, as predicted by the theory that CI results in F0-lowering. In the Off-On experiment, both women and men shift to higher F0 in long Tone 1 tokens and lower F0 in short Tone 1 tokens. Short Tone 1 tokens in men follow the mechanism that auditory feedback drives F0 lowering and auditory deprivation drives F0 raising. The F0-raising in longer tokens of Tone 1, instead, may be driven by attention paid to speech made possible by auditory feedback. While auditory feedback leads to F0-raising in longer Tone 1 tokens, the effect is modulated by attention paid to speech, driving a style-shift to standard speech [14]: high-level tone goes higher in F0.

Auditory feedback does not seem to have much effect on F0 in Tone 4 in the On-Off experiment. In the Off-On experiment, however, longer Tone 4 tokens invite higher F0 when the auditory feedback is recovered. That is, F0-raising is modulated by attention paid to speech during auditory enhancement. The lack of shift in F0 of Tone 4 in the On-Off experiment indicates that Tone 4 may be a linguistic variable whose phonetic realization invites more awareness from speakers; in contrast, Tone 1 is directly affected by auditory deprivation among men but less so among women, meaning that women in general seem to be more sensitive to shift in F0 and apply more effort to maintain F0 when the auditory feedback is just off.

The present research shows that the effect of auditory feedback on F0 is largely conditioned by the phonology of lexical tones in a tonal language like Mandarin. In standard realization of high tones, including high-level and high-falling tone, F0 should move upwards rather than downwards; this directly challenges the mechanistic effect of auditory enhancement, which is supposed to lower F0. The interaction between vowel duration and auditory feedback reveals different effects of auditory feedback on short tokens and long tokens: the former usually follows the relation between F0-raising/lowering and auditory deprivation/enhancement, and the latter instead is under more speaker control and therefore invites F0 movements that are associated with standard speech.

7. REFERENCES

[1] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, Aug. 2011.

[2] S. B. Leder, J. B. Spitzer, and J. C. Kirchner, "Immediate effects of cochlear implantation on voice quality," *Archives of Oto-Rhino-Laryngology*, vol. 244, no. 2, pp. 93–95, Jul. 1987.

[3] J. Hamzavi, W. Deutsche, W. D. Baumgartner, W. Bigenzahn, and W. Gstoettner, "Short-Term Effect of Auditory Feedback on Fundamental Frequency after Cochlear Implantation," *Audiology*, vol. 39, no. 2, pp. 102–105, Jan. 2000.

[4] M. K. Evans and D. D. Deliyski, "Acoustic Voice Analysis of Prelingually Deaf Adults Before and After Cochlear Implantation," *Journal of Voice*, vol. 21, no. 6, pp. 669–682, Nov. 2007.

[5] A. Gautam, J. G. Naples, and S. J. Eliades, "Control of speech and voice in cochlear implant patients," *The Laryngoscope*, vol. 129, no. 9, pp. 2158–2163, 2019.

[6] M. A. Svirsky, H. Lane, J. S. Perkell, and J. Wozniak, "Effects of short-term auditory deprivation on speech production in adult cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 92, no. 3, pp. 1284–1300, Sep. 1992.

[7] J. Kuo, "A large-scale smartphone-based sociophonetic study of Taiwan Mandarin," *Asia-Pacific Language Variation*, vol. 4, no. 2, pp. 197–230, 2018.

[8] K. Huang, "From pitch contour variation to tone change: An analysis of the phonological representations of the tones in Taiwan Mandarin," *International Journal of Chinese Linguistics*, vol. 4, no. 2, pp. 273–307, Jan. 2017.

[9] J. N. Stanford, "Sociotonetics using connected speech: A study of Sui tone variation in free-speech style," *Asia-Pacific Language Variation*, vol. 2, no. 1, pp. 48–81, Sep. 2016.

[10] Y.-Y. Chuang, J. Fon, I. Papakyritsis, and H. Baayen, "Analyzing phonetic data with generalized additive mixed models," in *Manual of Clinical Phonetics*, M. Ball, Ed. London: Routledge, 2021.

[11] D.-D. Tao, J.-S. Liu, and N. Zhou, "Acoustic analysis of tone production in Mandarin-speaking bimodal cochlear implant users," *JASA Express Letters*, vol. 2, no. 5, p. 055201, May 2022.

[12] T.-L. A. Wan, L. Hall-Lew, and C. Cowie, "Feeling disabled: Vowel quality and assistive hearing devices in embodying affect," *Language in Society*, vol. Online First, pp. 1–27, 2022.

[13] P. Tang, I. Yuen, N. Xu Rattanasone, L. Gao, and K. Demuth, "Longer Cochlear Implant Experience Leads to Better Production of Mandarin Tones for Early Implanted Children," *Ear and Hearing*, vol. 42, no. 5, p. 1405, September/October 2021.

[14] W. Labov, *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1973.

Effect of Hearing Loss on Prosodic Intervention in Parkinson's Speech

Caroline Menezes, Lori Pakulski

University of Toledo
Caroline.menezes@utoledo.edu, lori.pakulski@utoledo.edu

ABSTRACT

This study tests the effect of hearing loss on prosody in people diagnosed with Parkinson's disease (PD). Speech therapy for hypophonia in this population involves cueing speakers to increase their speech loudness. A complete audiological evaluation was completed on all participants that included pure tone hearing range, speech discrimination, comfortable loudness levels and dynamic range. Audiological testing was used to group the participants into two groups: mild to moderate and severe hearing loss. They were then asked to read a phonetically balanced sentence at four levels of self-perceived loudness (50%, 100% 200% and 400%). Changes in intensity and pitch range were compared for the two groups. The results indicate a negative impact of low dynamic hearing range on the participants' perceptual scale for loudness which resulted in lower intensity and pitch range. Impact of these findings on current speech therapy is also discussed.

Keywords: Parkinson's disorder, hearing loss, speech discrimination threshold, loudness, and pitch.

1. INTRODUCTION

Parkinson's disease (PD) is a neurological disorder of the basal ganglia, which according to Canter [1] is characterized by disturbances in muscular control including tremors, slowed movements, and muscle rigidity. Typical speech characteristics of PD include inappropriate silences, monopitch, monoloudness, short rushes of speech, slightly fast rate of speech, harsh or breathy voice, and imprecise consonants, due to the reduced range of movement of the speech articulators (Darly et al., 1969) [2]. Hypophonia or reduced speech loudness is a common indication of speech involvement in individuals with PD (Dykstra, 2012) [3] and also the major symptom that is remediated in speech therapy [4, 5]. Along with disruption to the speech system, researchers have established that auditory system dysfunction is a common nonmotor feature of patients with PD (Jafari, Kolb, & Mohajerani, 2020) [6]. Yet, there has been little attention in clinical practice to the impact of hearing impairment on speech perception, social communication, and quality of life [6].

Ho et al. (2000) [7] reported that individuals with hypophonia can improve their loudness when provided with external cues. Clark *et al.* (2014) [8] examined the loudness perception of individuals with hypophonia compared to a healthy control group by administering three loudness perception tasks. One task involved participants producing a sentence at normal level, then two times and four times quieter and louder. The participants with PD demonstrated a more restrictive loudness range and lower speech intensity compared to the control group. These results suggest that hypophonia may be due to a deficit in speech loudness perception. Yet, this study did not incorporate audiometric evaluation, so it is challenging to say if the restricted loudness range is due to speech perception deficits or hearing performance. De Keyser *et al*., (2016) [9] studying the relationship between speech production and speech loudness perception in individuals with mild PD, found no significant difference between PD and healthy control, however, their PD participants were better able to meet their target intensity when they were provided with cues on intensity levels.

In this study, we test the effect of hearing loss on participants' self-perception of loudness and its effect on speech prosody. In this preliminary publication we test only the aspect of pitch range when participants produce speech at different levels of loudness with the inclusion of hearing evaluation to better understand the influence of hearing.

2. METHOD

2.1. Participants

The participants analyzed in this study include four female individuals with PD. Participants were recruited from the University of Toledo Speech-Language Clinic where they received speech diagnosis and PD therapy.

2.2. Audiological evaluation

Audiometric evaluations were completed in a sound treated room with a calibrated diagnostic audiometer. Testing revealed the presence of sensorineural hearing loss in all subjects. Only one subject utilized hearing aids and those instruments provided inadequate access to the full range of speech sounds.

Table 2: Participant demographics.

Participant	Age (years)	Months since diagnosis
2PD	70	12
5PD	92	96
6PD	88	132
7PD	64	144

2.3. Speech paradigm

Speech data was extracted from the speech diagnosis protocol. Here the participants were asked to read the utterance "Kick the ball straight and follow through," at different levels of self-perceived loudness, similar to the protocol employed by Clark [5]. Participants were told to read the utterance at their most comfortable loudness and habitual pitch. They were then instructed to read it four more times. Their most comfortable loudness was referred to as 100%. They then had to read the same sentence at 200%, 400% and 50% loudness. All participants produced the utterances in the same order. It is expected that the reading at 50% would be influenced by the 200% and 400% productions, so they were asked to produce the 100% again one time before reading the utterances at half loudness. Each of these levels were produced five times. At no time were the participants provided the decibel rating of their loudness.

2.4. Analyses

For this publication, participants were grouped into mild-moderate hearing impairment (MHI) for hearing thresholds between 25-55 dBHL and severe hearing impairment (SHI) for thresholds >56 dBHL. Mean speech intensity values were calculated over the entire utterance separately for all individuals and levels of loudness. Similarly, pitch range values were calculated separately for all individuals and levels of loudness. We believe that pitch range gives us a more comprehensive understanding of the participants' pitch profile.

3. RESULTS

3.1. Audiological report

Audiometric findings are summarized in Table 2. Two subjects had MHI with normal word discrimination ability in optimal listening conditions. Two subjects had SHI with fair word discrimination ability in optimal listening conditions. All demonstrated a reduced dynamic listening range: the greater the loss, the more diminished the range for these subjects.

Table 2: Audiometric report for all speakers.

Subject	Ear	Hearing Range	SRT	Speech Discrimination	MCL	UCL	Dynamic Range (UCL-SRT)
2PD	Right	Normal to mild loss	10	92%@50 dBHL	55	85	75
	Left		10	100%@50 dBHL	60	85	75
7PD	Right	Normal to moderate loss	20	100%@50 dBHL	55	80	60
	Left		15	100%@50 dBHL	50	80	65
5PD	Right	Mild to profound loss	60	62%@75 dBHL	70	95	35
	Left		60	84%@75 dBHL	75	95	35
	Aided	Aided	35	86%@50 dBHL	50	70	35
6PD	Right	Moderate to profound loss	75	68%@85dBHL	80	100	25
	Left		70	64%@85 dBHL	80	100	25

3.2. Intensity

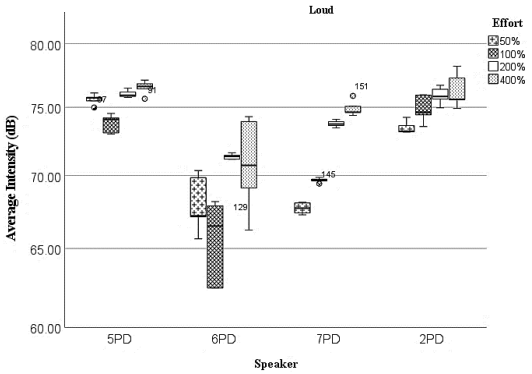


Figure 1: Boxplots of average intensity for speakers separated by loudness level. Speakers 5PD, 6PD SHI, speakers 7PD, 2PD MHI

Figure 1 displays the average intensity values calculated over the entire utterance separated by each speaker for the different levels of self-perceived loudness. Speaker 6PD had the lowest overall intensity and the greatest difference between the different levels of loudness; however, she also demonstrated a lot of variation across the individual repetitions. Looking at Table 3, we see that this participant had a 14dB increase from 100% to 200% but approximately 2dB decrease from 200% to 400%. In contrast 5PD, who is grouped together with 6PD on hearing response had limited change in intensity across the loudness levels, but she also had lesser sentence to sentence variations. Referring to Table 3 again, we see that 5PD has intensity values congruent with the MHI participants but the difference between the levels is most evident only between 100% and 200% which was a difference of 2dB. She increased her intensity by only 0.5 dB when she doubled her loudness from 200%. Speakers with mild to moderate hearing impairment were able to clearly distinguish each level of loudness with lesser variation. For every level of

loudness, they exhibited a difference of 1 or 2dB amplitude, depending on the speaker. Generally, we note that PD participants perception of doubling loudness was on average 1.5 to 2dB.

Table 3: Average intensity and pitch range values for all speakers for different loudness levels and HI.

Hearing	Speaker	50% Loudness		100 % Loudness		200% Loudness		400% Loudness	
		Intensity (SE)	Pitch Range (SE)	Intensity (SE)	Pitch Range (SE)	Intensity (SE)	Pitch Range (SE)	Intensity (SE)	Pitch Range (SE)
Severe	5PD	76 (.18)	167 (11)	73.8 (.31)	135.7 (15.5)	76 (.1)	168 (11)	76.5 (.24)	167 (8.25)
	6PD	68 (.89)	81.8 (9.2)	63.9 (2.5)	119 (14)	78.6 (.85)	145.9 (14.5)	76.9 (1.5)	154 (20.7)
Moderate -Mild	7PD	67.7 (.17)	84.5 (5.8)	69.7 (.07)	130 (4.5)	73.8 (.11)	178 (6.6)	74.9 (.26)	206.8 (1.8)
	2PD	73.5 (.21)	170.9 (10.8)	74.9 (.46)	267 (6.5)	75 (.30)	317.8 (14.3)	76 (.61)	303.6 (12.8)

3.3. Prosody

Pitch range was calculated for all speakers separately for the different levels of loudness due to individual variation in symptoms of PD and hearing loss. For all speakers pitch increased as loudness increased with the exception of 5PD at 50% level of loudness. However, it should be noted that this level was produced with higher intensity than her 100%. The correlation of loudness and pitch range was positive and significant for all speakers. However, the correlation was lesser in the SHI group ($r(77) = .67$, $p < .001$) when compared to the MHI group ($r(77) = .83$, $p < .001$).

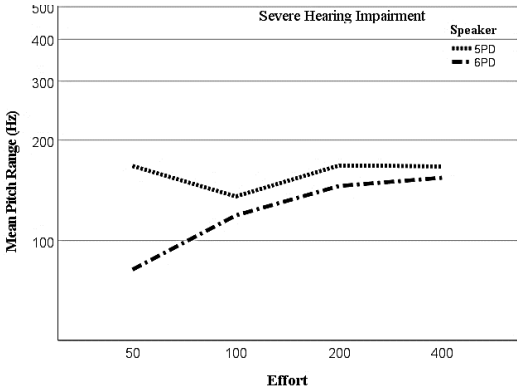


Figure 2: Line graphs of average pitch range for SHI speakers separated by loudness level.

Figure 2 plots line graphs for the two SHI participants separately for the different levels of loudness. We see that pitch range for these two speakers are relatively narrow falling between 68 to 167 Hz. Note that 50% for 5PD is higher but her intensity measure for this level is also higher than her habitual production (100%). Her pitch range decrease from 100% to 50% is approximately 30 Hz, however, between 50%, 200% and 400% her pitch range does not change similar to her intensity values. Speaker 6PD exhibited on average a change in pitch range equivalent to 30Hz through the levels of 50%, 100% and 200%. From 200% to 400% she had decreased her intensity by approximately 1.5 dB, but she increased her pitch range by approximately 10 Hz. In general, this speaker was not able to accurately scale her intensity, but she was better with her pitch modulation.

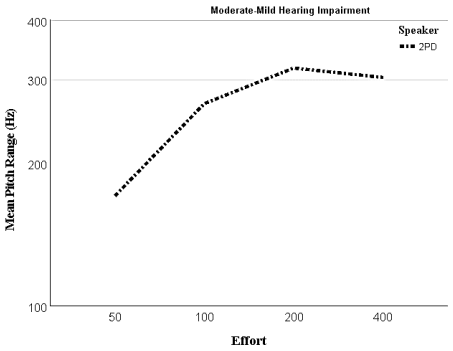


Figure 4a: Line graphs of average pitch range for MHI speaker 2PD separated by loudness level.

Figures 4a plots the pitch range for 2PD. The pitch range for this speaker was approximately 148 Hz with approximately 45 Hz increase for every perceptual doubling of loudness. We see a similar large pitch range (122.5 Hz) in Speaker 7PD as seen in Figure 4b. For every perceptual increase in loudness this speaker also had approximately 45 Hz increase in pitch range.

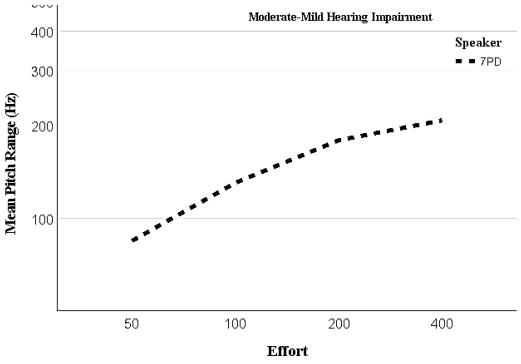


Figure 4b: Line graphs of average pitch range for MHI speaker 7PD separated by loudness level.

4. CONCLUSION

Our results indicate negative interaction between hearing loss and pitch range depending on the speech dynamic range in people diagnosed with Parkinson's disease. Pitch range was reduced in persons classified as SHI when compared to those classified as MHI. However, pitch range significantly increased with greater loudness.

Except for Speaker 6PD, all speakers had a gradual increase in intensity as the perceived levels of loudness were doubled. 6PD's intensity change was erratic increasing by 14dB difference between 100% to 200% but a decreasing from 200% to 400%. This indicates that she was not able to achieve an accurate scaling of her perceptual loudness. Interestingly, she has the most impaired loudness perception, tolerating only 25 dBHL (dynamic range), which may certainly impact self-monitoring and control. Speaker 5PD, also classified as SHI, on the other hand was not able to perceive a halving of loudness from 100% to 50%. Similarly, she was only able to increase loudness by half a decibel when asked to double her intensity from 200%. She is the only subject with hearing aids, but due to their inadequate support, may negatively influence her responses. Further, the devices themselves may also impact her loudness perception in the way they process sound input.

On the other hand, the speakers classified as MHI were able to demonstrate a gradual increase in their perceptual scale of loudness though it varied from one speaker to the other. On average, a 1.5 to 2dB difference in amplitude was observed for every doubling on the perceptual scale. This intensity change over the scale also showed a steadying increase in pitch range. Both also demonstrated better loudness tolerance with wider dynamic listening ranges.

5. DISCUSSION

Within our small subset of speakers, we see several individual variations in intensity and pitch range related to their difference in Parkinson's disease symptoms and hearing loss. However, we see a compromise in their perceptual scale of loudness based on their speech dynamic range. Speakers with better hearing and dynamic range were classified as mild to moderate speech while speakers with more severe hearing loss and smaller speech dynamic range were classified as severe in this study. This classification for this preliminary study held true when comparing their performance on their self-perceived loudness scale.

All participants were able to perform the task to some degree of accuracy, but it is clear that speech therapy for the remediation of hypophonia needs to be more granular to address the inherent prosodic changes and the interaction of prosody. Further, these results support findings reported by Jafari et al. (2020) with regards to the need for far more attention in clinical work to a wide range of aspects of the impaired auditory system [6]. More comprehensive audiological evaluation is warranted. For without it, it is unclear how patients with hearing impairments that are largely untreated can benefit from intervention. Further, without audiological information, it is challenging to determine if benefits from intervention are simply due to adequate auditory access and dynamic listening range more so than the treatment itself.

7. REFERENCES

[1] Canter, G. J. (1963). Speech characteristics of patients with parkinson's disease: I. Intensity, pitch, and duration. *Journal of Speech and Hearing Disorders*, 28(3), 221–229. https://doi.org/10.1044/jshd.2803.221

[2] Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnosis patterns of dysarthria. *Journal of Speech and Hearing Research*, 12, 246–249.

[3] Dykstra, A., Adams, S., & Jog, M. (2012). The Effect of Background Noise on the Speech Intensity of Individuals with Hypophonia Associated with Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *20*(3), 19–30.

[4] Behrman, A., Cody, J., Elandary, S., Flom, P., & Chitnis, S. (2020). The effect of speak out! and the loud crowd on dysarthria due to Parkinson's disease. *American Journal of Speech-Language Pathology*, *29*(3), 1448–1465. https://doi.org/10.1044/2020_ajslp-19-00024

[5] Ramig, L. O., Sapir, S., Fox, C., & Countryman, S. (2001). Changes in vocal loudness following intensive voice treatment (LSVT®) in individuals with parkinson's disease: A comparison with untreated patients and normal age-matched controls. *Movement Disorders*, *16*(1), 79–83. https://doi.org/10.1002/1531-8257(200101)16:1<79::aid-mds1013>3.0.co;2-h

[6] Jafari, Z., Kolb, B.E. and Mohajerani, M.H. (2020), Auditory Dysfunction in Parkinson's Disease. Mov Disord, 35: 537-550. https://doi.org/10.1002/mds.28000

[7] Ho, A. K., Bradshaw, J. L., & Iansek, R. (2000). Volume perception in Parkinsonian speech. *Movement Disorders*, 15, 1125–1131.

[8] Clark, J. P., Adams, S. G., Dykstra, A. D., Moodie, S., & Jog, M. (2014). Loudness perception and speech intensity control in Parkinson's disease. *Journal of Communication Disorders*, *51*, 1–12.

[9] De Keyser, K., Santens, P., Bockstael, A., Botteldooren, D., Talsma, D., De Vos, S., Van Cauwenberghe, M., Verheugen, F., Corthals, P., & De Letter, M. (2016). The relationship between speech production and speech perception deficits in parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 59(5), 915–931.

EXPLORING THE USE OF PHONOLOGICAL FEATURES FOR PARKINSON'S DISEASE DETECTION

Nina Hosseini-Kivanani [1], Juan Camilo Vásquez-Correa[2], Christoph Schommer[1], Elmar Nöth[3]

[1]Faculty of Science, Technology, and Medicine, University of Luxembourg, [2]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA) Donostia-San Sebastian, Spain,[3]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nüremberg, Germanyy
nina.hosseinikivanani@uni.lu, jcvasquez@vicomtech.org, christoph.schommer@uni.lu, elmar.noeth@fau.de

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that causes motor and non-motor symptoms. Speech impairments are one of the early symptoms of PD, but they are not always fully exploited by clinicians. In this study, the use of phonological features extracted from speech data collected from Spanish-speaking patients was explored to predict PD patients from healthy subjects using phonet, which was trained on Spanish data, and PhonVoc, which was trained on English data. These features were then used to train and test several machine learning models. The XGBoost model achieved the best performance in classifying patients from HCs, with an accuracy of over 0.76. However, the model performed better when using a phonological model trained on Spanish data rather than English data.

Keywords: Parkinson's disease, machine learning models, classification, Phonet, PhonVoc.

1. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's that affects movement and can cause symptoms such as tremor, slowness of movement, muscle rigidity, and different writing and speech deficits. PD is typically diagnosed through a combination of a patient's medical history, symptoms, and a neurological and physical exam. However, PD is a slow-progressive disease and may have a prodromal period of 3 to 15 years during which the main motor symptoms may not be clearly perceivable [1]. This highlights the importance of developing robust and automatic methods for detecting the disease in its early stages to improve the patient's quality of life.

Speech impairments are often one of the earliest motor symptoms of PD, and they can worsen as the disease progresses. Recent studies have indicated that over 90% of PD patients express some kind of speech impairment [2], with disorders mostly related to phonation and articulation, including alterations in speech rate and pitch variations, imprecise articulation of vowels and consonants, and monotonous speech, leading to decreased speech intelligibility [3]. Speech impairments in individuals with PD often result in imprecise articulation, particularly affecting the production of stop, affricate, and fricative sounds [4]. Analysis of speech materials such as sustained vowels and diadochoknetic (DDK) exercises (i.e., rapid syllable repetition) can aid in monitoring the disease severity of patients [5]. which have been widely used for detecting early symptoms of the disease [6].

Various machine learning (ML) methods have been widely used to classify PD patients vs. HCs using speech signals ([7, 8]). In [7], feature subset selection ranging from 8 to 20 was used to represent dimensionality reduction in complexity, and the accuracy of ML models was improved using 10-fold cross-validation on a small dataset to implement a voice-based detection methodology In [9], the importance and effectiveness of ML models were investigated using utterances of the sustained vowel /ah/ from 188 PD patients and 64 HCs. The results showed that repeating voice patterns are identified in PD speech, proving that ML models can support the diagnosis of PD with similar accuracy as movement disorder therapists.

Phonological posterior features have been found to be helpful in gathering information about the presence and severity of PD in patients [10]. Among the common toolkits for extracting phonological features, PhonVoc [11] has been used to evaluate the progress of apraxia based on speech. [12] used PhonVoc to extract phonological features of impaired speech to better distinguish the disease. In addition, the Phonet toolkit [13] is also being used for speech processing, specifically for predicting the

posterior probability of speech files, which refers to the likelihood of a speech file based on the data it contains. [14] recently used Phonet to extract phonological features from Dutch patients with dysarthria to develop a computer-based therapy approach. This study aims to address the pertinent problem of identifying and monitoring PD using speech signals as a measurement, with the intention of creating useful tools for this purpose.

It is uncertain how neurodegenerative diseases impact the production of different groups of phonemes, including those related to the manner and place of articulation. Additionally, the majority of available language resources are in English. We are conducting a study to compare the performance of two phonological feature extraction models, one trained on English data and one trained on Spanish data. This is the first cross-language study of these models, and the goal is to determine whether these models are language independent or if one model performs better on the Spanish data due to being trained on data in the same language.

2. MATERIAL & METHODS

2.1. PD Dataset

The study used data from the PC-GITA database [15], which consists of speech utterances from 68 individuals with PD and 48 healthy controls (HC), all of whom were Spanish-speaking and balanced in terms of age and gender. Patients' neurological state was evaluated using the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS-III scale) [16]. Recording were taken while patients were on anti-Parkinsonian medication. This was done because of ethical reasons. The statistical tests for validating gender and age balance are detailed in Table 1([17]).

2.1.1. Speech material

The Speech material includes six DDK tasks, which consist of rapid repetition of syllables such as /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, and /ka/. These tasks are designed to evaluate the speakers' ability to move their vocal tract articulators (e.g, lips, tongue, and soft palate) as quickly and consistently as possible.

2.2. Toolkits

We used **PhonVoc** [11] and **Phonet** [13] toolkits to extract phonological features from the speech signal, including features related to phonation. **PhonVoc** is a toolkit that uses a deep neural network (DNN) to estimate the probability that a sound belongs to a specific phonological class in English, and it has an accuracy of over 96% for detecting these classes. **Phonet** toolkit is based on Gated Recurrent Neural Networks to extract phonological posteriors from speech. The toolkit was trained on 17 hours of clean FM podcasts in Mexican Spanish, which is able to detect the phonological classes with high accuracy (90%). For each toolkit, we have calculated values for the mean, standard deviation, skewness, and kurtosis. The list of phonological classes that correspond to both stop and voiced segments of speech for evaluating consonant imprecision includes "Back," "Voice," "Vocalic," "Consonantal," "Continuant," "Coronal," & "Silence". "Silence" is considered a separate phonological class, because it represents the absence of sound. These classes can be used to analyze the accuracy of the toolkits in detecting specific phonological features in speech signals.

2.3. Classical ML models

In this project, ML techniques were used to classify speech based on its features. ML allows computers to recognize patterns in data and make decisions based on those patterns. We used several commonly used ML models such as Naïve Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), *k*-nearest neighbors (*k*NN), and Ensemble methods (Voting & eXtreme Gradient Boosting (XGBoost)) for the binary classification problem. These models were implemented using the *scikit-learn* library in python (3.9).

2.3.1. Data sampling methods

The number of speech samples for HCs and PD patients was not equal: 72% of the samples related to tasks performed by PD patients, while 28% related to tasks performed by HCs. Imbalanced distribution of the classes could lead to a higher likelihood of misclassification for the minority class. A balancing technique was applied to create equal samples for each class. We used a number of commonly employed techniques for handling imbalanced datasets, including:

Oversampling: It is a technique for balancing imbalanced datasets by increasing the number of minority class samples in the training dataset. One commonly used oversampling technique is Synthetic Minority Over-sampling (SMOTE) [18], which generates synthetic data by selecting the nearest data points (typically k = 5) based on the

	PD patients	HC subjects	Patients vs. controls
Gender [F/M]	31/37	24/24	$*p = 0.99$
Age [F/M]	60.9(11.2)/64.7(9.4)	59.9(8.7)/63.5(10.4)	$**p = 0.08$
Time since diagnosis [F/M]	15.5(14.5)/8.1(5.9)	–	
MDS–UPDRS–III [F/M]	36.2(18.1)/36.3(18.9)	–	

Time since diagnosis and age are given in years. [F/M]: Female/Male. Average (Standard deviation).
$*p$–value calculated through Chi–square test. $**p$–value calculated through Mann-Whitney U-test.

Table 1: Clinical & demographic information of the subjects from the PC-GITA database [15].

Euclidean distance between them in the feature space and duplicating these points in the dataset.

Class weight: This technique involves computing the frequency of each class and then inverting it such that the underrepresented class has a much larger error when multiplied by the class loss compared to the majority class.

Threshold: It is a technique for adjusting the predicted values by setting a threshold value. If the predicted value is greater than this threshold, it is set to 1; otherwise, it is set to 0.

2.4. Evaluation

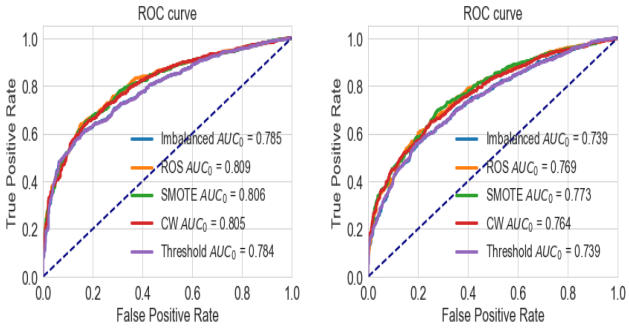
The dataset was split into a training set (75% of data, including 84 out of 112 participants) and a testing set (25% of data) based on the participants' IDs. To optimize performance, k-fold cross validation was used, specifically **group k-fold** cross validation (n_splits=5), to ensure the same group of recordings from one participant does not appear in both the training and testing sets, avoiding overfitting and ensuring subject-independent cross-validation. To evaluate the performance of the ML models, we used metrics such as accuracy, precision, recall, and receiver operating characteristic (ROC) to obtain the Area Under Curve (AUC). These metrics are based on four outcomes: true negative (TN), false negative (FN), false positive (FP), and true positive (TP), which are shown in a confusion matrix (Figure 3).

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- AUC $= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$

3. RESULTS & DISCUSSION

The outputs of using Phonet to extract Spanish phonological features (SpNet) and English phonological features (EnVoc) were presented to evaluate the performance of the models with these features. By comparing the results with different types of features, it is possible to determine which models work best with specific types of data. To evaluate the effectiveness of balancing techniques, we used SVM for binary classification tasks which yields the best accuracy score with relatively low training times. We calculated the accuracy scores of the models with and without a balanced dataset for the baseline model to determine which balancing techniques were most effective.

The baseline classifier model had an accuracy of 0.75 for the SpNet and 0.73 for the EnVoc, but it was biased towards the majority class (PD). To address this issue, various balancing approaches were employed (as described in Section 2.3.1). The results of the baseline model on different sampling techniques (Figure 1) showed that oversampling performed better than SMOTE, class weight, or threshold techniques, with AUC values of 0.81 and 0.77 for SpNet and EnVoc, respectively. AUC values between 0.7 and 0.8 are considered fair [19] , and the values obtained for SpNet and EnVoc fall within this range. This suggests that the model is relatively accurate in its classification tasks.



(a) SpNet **(b)** EnVoc

Figure 1: ROC curves & AUC values: imbalanced, Oversampling, SMOTE, Class weight & threshold.

To further improve the performance of the ML models, we implemented a grid search procedure to optimize the hyperparameters of the models. This entailed specifying a range of values for each hyperparameter and evaluating the performance of all possible combinations to identify the optimal configuration. Although this did not significantly improve as a result of this process, with the exception of the SVM model, the results of the other models were still relatively comparable to those of the SVM model.

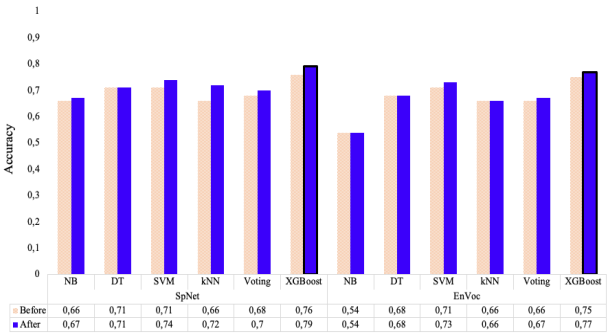


Figure 2: Result of ML models before & after hyperparameter tuning.

In addition to the ML models previously mentioned, we also used ensemble methods such as the Voting method and the XGBoosting technique to try to improve the performance of the ML models. The Voting method involves combining the results of multiple base classifiers based on weighting, while XGBoosting is a combination of bagging and boosting techniques. After applying hyperparameter tuning and oversampling to the datasets, we found that the XGBoost model was more effective at classifying PD vs. HC than the other models, achieving a mean accuracy of 0.79 for SpNet and 0.77 for EnVoc 2. The confusion matrix for this model also supported its effectiveness (Figure 3).

The confusion matrix in Figure 3 illustrates the model's performance and the types of errors it is making by comparing the actual labels (vertical levels) to predicted labels (horizontal levels) by XGBoost model for each phonological feature to evaluate the model's performance in the binary classification problem. These values were used to calculate accuracy, recall, and precision. The results of the XGBoost model are presented in Figure 3, which shows the correct (i.e., TP and TN) and incorrect (i.e., FP and FN) predictions made by the model. These results demonstrate high recall and precision values: recall for SpNet was 0.85 and for EnVoc was 0.84, while precision for SpNet was 0.84 and for EnVoc was 0.83.

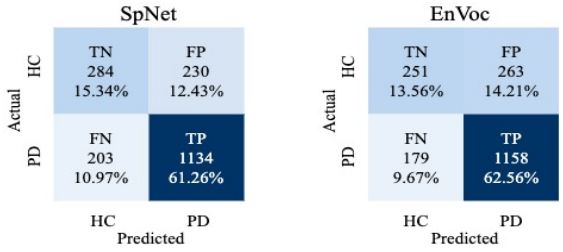


Figure 3: Confusion matrix of XGBoost model.

The XGBoost model was used to conduct a feature importance analysis in PD speech by ranking the input features based on their phonological features on SpNet and EnVoc. This analysis identified the most important and influential features for classifying PD speech from HCs in the dataset. The top team important features are shown in Figure 4, with higher scores indicating a greater impact on the model's predictions. The mean of continuant in SpNet and the kurtosis of voice consonant in EnVoc received the highest importance scores among all the features (Figure 4). These listed features have the potential to improve the model's performance. The underlying aspects of speech production warrant further investigation.

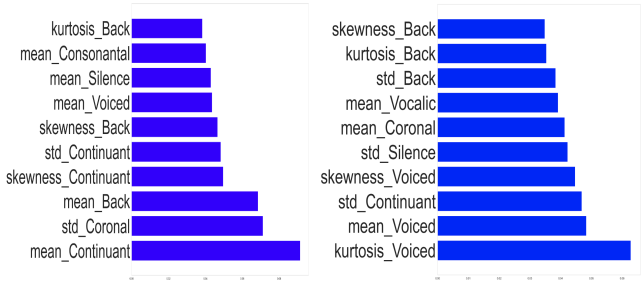


Figure 4: Top 10 important features of SpNet (left) & EnVoc (right) for XGBoost model.

4. CONCLUSION

We conducted a study to assess the ability of ML models to classify PD patients and HCs based on speech features. We compared the performance of phonVoc and phonet toolkits: one trained on a large corpus of Librispeech consisting of 1000 hours of clear speech in English, and another trained on the Spanish corpus of radio podcasts in Spanish. Our results showed that SpNet had slightly better performance than English (0.79 vs 0.77). The difference in the quality and size of the training data may have contributed to the slightly lower performance of phonVoc model. But this difference could be larger if we had access to a higher quality corpus of Spanish speech. Our study is the first to compare the performance of ML models trained on English and Spanish for our task, and this opens up new research avenues, such as considering the use of the UPDRS-III for multiclass classification problems. These results support the usefulness of the Phonet toolkit, which was specifically designed for the Spanish language and optimized for classifying speech data.

5. ACKNOWLEDGEMENT

This publication is based upon work from COST Action GoodBrother, COST Action

19121—Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living, supported by COST (European Cooperation in Science and Technology).

6. REFERENCES

[1] R. Savica, J. M. Carlin, B. R. Grossardt, J. H. Bower, J. E. Ahlskog, D. M. Maraganore, A. E. Bharucha, and W. A. Rocca, "Medical records documentation of constipation preceding Parkinson disease," *Neurology*, vol. 73, no. 21, pp. 1752–1758, 2009.

[2] G. Solana-Lavalle, J. C. Galán-Hernández, and R. Rosas-Romero, "Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 505–516, 1 2020.

[3] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific Reports*, vol. 7, no. 1, p. 12, 2017.

[4] J. A. Logemann and H. B. Fisher, "Vocal Tract Control in Parkinson's Disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 348–352, 1981.

[5] J. C. Vásquez-Correa, C. D. Rios-Urrego, A. Rueda, J. R. Orozco-Arroyave, S. Krishnan, and E. Nöth, "Articulation and Empirical Mode Decomposition Features in Diadochokinetic Exercises for the Speech Assessment of Parkinson's Disease Patients," in *Iberoamerican Congress on Pattern Recognition*, vol. 11896 LNCS. Springer, 2019, pp. 688–696.

[6] F. Karlsson, E. Schalling, K. Laakso, K. Johansson, and L. Hartelius, "Assessment of speech impairment in patients with Parkinson's disease from acoustic quantifications of oral diadochokinetic sequences," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, p. 839, 2 2020.

[7] G. Solana-Lavalle and R. Rosas-Romero, "Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation," *Biomedical Signal Processing and Control*, vol. 66, p. 102415, 2021.

[8] J. R. Orozco-Arroyave, J. C. Vádsquez-Correa, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, no. March, pp. 6490–6494, 2016.

[9] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.

[10] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech and Language*, vol. 46, no. June, pp. 196–208, 2017.

[11] M. Cernak and P. N. Garner, "PhonVoc: A phonetic and phonological vocoding toolkit," in *INTERSPEECH*, 2016, pp. 988–992.

[12] A. Asaei, M. Cernak, and M. Laganaro, "PAoS Markers: Trajectory Analysis of Selective Phonological Posteriors for Assessment of Progressive Apraxia of Speech," in *Proceeding on the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016.

[13] J. C. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: a Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech." in *INTERSPEECH*, 2019, pp. 549–553.

[14] V. M. Ramos, J. C. Vasquez-Correa, R. Cremers, L. Van, D. Steen, E. Nöth, M. De Bodt, and G. V. Nuffelen, "Automatic boost articulation therapy in adults with dysarthria: Acceptability, usability and user interaction," *International Journal of Language & Communication Disorders*, vol. 56, no. 5, pp. 892–906, 2021.

[15] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 342–347.

[16] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, and others, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[17] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[19] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638–1645, 2016.

TOWARDS DISFLUENCY FEATURES FOR SPEECH TECHNOLOGY BASED AUTOMATIC DEMENTIA CLASSIFICATION

Megan Thomas[1*], Samuel Hollands[1*], Daniel Blackburn[2], Heidi Christensen[1]

[1]Department of Computer Science, University of Sheffield, UK

[2]Sheffield Institute for Translational Neuroscience, University of Sheffield, UK
[*]*Authors Contributed Equally*
{mcthomas2, shollands1, d.blackburn, heidi.christensen}@sheffield.ac.uk

ABSTRACT

Automatically detecting cognitive decline using speech and language technologies is a field of growing research. This domain deploys AI-based classification technologies into clinical environments to assist traditional diagnostic methods. This paper explores a new series of disfluency features to assist the widely adopted *acoustic-linguistic* elements of contemporary automatic dementia diagnosis.

Utilising automatic speech recognition, phoneme recognition, and voice activity detection technologies on the Sheffield IVA data [1], we show improved accuracy from a baseline of 78.4% to 83.2% using automatically calculated disfluency features. This paper contributes disfluency features as a method of enhancing current feature sets used in the automatic detection of cognitive decline, and highlights potential approaches towards reliably extracting these disfluency features automatically. Features that are less dependent on raw acoustic information are potentially more generalisable and robust for dementia classification tasks in varied environments, a key obstacle in this domain.

Keywords: Dementia, Disfluency, Speech Analytics, Classification

1. INTRODUCTION

Cognitive decline (CD) and diseases such as Alzheimer's Dementia (AD) pose a significant challenge to health services worldwide, with an estimated 139 million people worldwide living with dementia by 2030 [2]. As the need for early and accurate CD detection increases, a growing body of work within the speech and language technologies (SLTs) domain is looking at assisting doctors and clinicians in the diagnosis of CD through the use of automatic cognitive decline detection (ACDD) systems, specifically through the automatic analysis of speech [3]. Speech is an effective biomarker for this task as it is readily available, easy to collect, and contains features that correlate well to different levels of CD [4, 5, 6, 7].

Numerous studies have highlighted the effectiveness of different temporal measures of speech for discriminating between people with CD and healthy controls. Some temporal measures such as speech rate are even salient enough to differentiate between prodromal dementia and early stages of AD [8]. Pauses are another common feature frequently investigated for the purpose of automatically detecting CD, with correlations found between the number and duration of filled and unfilled pauses and levels of CD (for example in [9] and [10]). Despite this, little research has been conducted into the feasibility of using other disfluencies to aid in the detection of CD. This is due, in part, to the difficulties of including disfluencies in automatically produced transcripts.

This study, building on work from [11], explores the efficacy of using manually annotated disfluency features to aid in the automatic detection of cognitive decline. The capability of these features to enhance the performance of traditional *acoustic-linguistic* based systems is demonstrated, and initial experiments to automatically calculate these features are presented. Automation of these features is imperative in order to apply them to speech analytics based classification systems.

2. BACKGROUND

Recent successful ACDD systems have consisted of the following components. A virtual interface asks a patient some pre-written prompts and records the response [12]. An automatic speech recognition (ASR) system is used to transcribe their utterance into textual information. Two sets of features are typically generated; acoustic (non-verbal) and linguistic (verbal) [13]. Acoustic features include

Table 1: Disfluency features used within this study and their descriptions with constituent elements

Disfluency Feature	Description
Unfilled Pauses	Silence lasting >200ms
Filled Pauses	Sustained sound, typically a vowel or a sound that is not part of a word, lasting >200ms
Repetitions	Part word, whole word, or phrase repetitions
Prolongations	Prolongation of a phone lasting >200ms
Repairs	When a speaker alters what they had been saying
Speech Errors	Phonetic additions, substitutions, or deletions. selection errors, retrieval errors, and blends

metrics extracted directly from the audio recordings, and linguistic features include parameters extracted from a transcript of the speech to be analysed. These features are then fed into a classification system which is trained to discriminate between classes (or conditions).

Recent work from [11] suggests that disfluency information could be valuable in a CD detection task. A manual disfluency analysis found that features such as number of word repetitions and prolongations were significantly different between people at different stages of cognitive decline. The present study aims to use this disfluency information to inform an ACDD system alongside *acoustic-linguistic* features. However, automating this process is no simple feat. Typical ASR systems are designed to ignore disfluencies in speech and therefore disfluencies are not typically present in automatically created transcripts. Unlike "traditional" features used for similar tasks, disfluency features have the added benefit of being easily explainable. The issue of explainability and interpretability is especially relevant when working in the medical domain, as discussed in [14].

3. DATA

A subset of the Sheffield IVA dataset [1] was used for this project. This data is collected via a virtual agent as part of an ACDD system that poses a number of language tasks to patients and records their answers. The task used for this study was the Cookie Theft picture description task [15]. Participants belonged to one of three different diagnosis groups; healthy controls (HC), mild cognitive impairment (MCI), and neurodegenerative dementia (ND). The ND group contains participants with various types of dementia, which could also include AD. Each group contained 18 recordings. Recordings were selected for this study based on the intelligibility of the speech from the participants. Classification tasks within this study use 15 HC and 10 ND participants. In the larger IVA dataset there may be multiple recordings of the same speaker. Our sample contains only one recording for each participant.

4. METHODOLOGY

4.1. Feature Selection & Extraction

4.1.1. Manual Disfluency Features

A disfluency schema was created to determine exactly which disfluencies should be investigated. This was based largely on the work of [16]. After the work conducted in [11], minor changes were made to the disfluency schema resulting in the schema outlined in Table 1 above.

Praat [17] was used to manually annotate all instances of the above disfluencies within the recordings. Diagnosis labels were hidden from the annotator, a trained phonetician, to avoid any potential bias.

4.1.2. Automatic Disfluency Features

A Jupyter Notebook was created to calculate the disfluency features automatically. Voice activity detection (VAD) [18, 19] with a threshold of 50ms onset/offset and a 50ms minimum speech threshold (informed by the average length of voiced and unvoiced British English syllables [20]) was used to identify and parse the speech components of each file. A convolutional recurrent deep neural network (CRDNN) based Commonvoice [21] ASR system was used; this was preferred over a transformer based system given the potential for improved performance on non Southern British English accent groups [22]. For the syllabic (and particularly non-lexical) aspects of the task both phoneme recognition (PR) and grapheme to phoneme (G2P) technologies were explored. PR (using CMUSphinx) transcribes a phone given an audio signal input. G2P takes the output of an ASR system and transcribes back into phones. PR was found to outperform G2P and was therefore used for syllable parsing and all non-lexical feature

Table 2: Linguistic features used within this study [12], the number of features (n) within each category, and the automated elements needed for each feature to function

Tok: Tokeniser, POS: Part-of-Speech Tagger, W_List: Word List, CoRef: Co-reference Tagger, Sem_Tag: Semantic Tagger, POS_Pat: POS Pattern Matcher, Tree: Syntactic Tree Parser

Feature(s)	n=#	Automated Components
Content Density	1	Tok, POS, W_List
Part-of-Speech Rate	45	Tok, POS
Reference Rate to Reality	1	Tok, POS
Personal, Spatial and Temporal Deixis Rate	3	Tok, POS, W_List, CoRef
Relative Pronouns and Negative Adverbs Rate	2	Tok, POS, W_List
Lexical Richness	3	Tok
Action Verbs Rate	1	Tok, POS, Sem_Tag
Frequency-of-Use Tagging	1	Tok, W_List
Propositional Idea Density	1	Tok, POS, POS_Pat
Mean Number of Words in Utterance	1	Tok
Number of Dependent Elements Linked to the Noun	2	Tok, POS, Tree
Global Dependency Distance	2	Tok, POS, Tree
Syntactic Complexity	1	Tok, POS, Tree
Syntactic Embeddedness	2	Tok, POS, Tree
Utterance Length	2	Tok

calculations. The disjoint between the ASR and PR provided some inaccuracy; a robust ASR system for Northern British English with a phone level output would have resulted in better accuracy in this area.

4.1.3. Automatic Acoustic Features

The openSMILE toolkit [23] was used in this study for acoustic feature extraction using the Python API. eGeMAPSv02, Emobase, and ComParE were used as feature sets that have been well explored within the literature. All three feature sets were calculated for each of the participant recordings. 42 features containing SMB contour smoothing within ComParE were redacted due to a strong negative impact on support vector machine (SVM) accuracy due to the values being calculated as near binary measurements (distance away from min|max $< 10^{-5}$ when normalised).

4.1.4. Automatic Linguistic Features

Table 2 highlights the linguistic features extracted for this research [12]. These features were extracted using a variety of different tools highlighted within the *Automated Components* column. The libraries used were spaCy [24] (word tokenisation, POS tagging, coreference resolution, and semantic tagging), NLTK [25] (sentence tokenisation and tree parsing), and Re(gex) (POS pattern matching). A Jupyter Notebook was developed using Python 3.10 containing custom written functions to extract each feature.

4.2. Baseline System

The baseline classifier in this study was developed using exclusively interpretable metrics; we did not use features such as word embeddings as it was counter-productive to the ambition of finding interpretable solutions. A baseline two way classifier was built to discriminate HC-MCI/ND. An SVM was trained using acoustic feature sets; eGeMAPSv02, Emobase, and ComParE using the following hyperparameters {“C”: 100, “gamma”: 0.001, “kernel”: “rbf”}. All SVMs were developed using 5-fold cross validation. Baseline performance was 78.4% for HC-MCI/ND. The baseline system used exclusively acoustic features for reasons discussed in Section 5.

For our disfluency features, normalisation was explored using three options; relative to speech duration, instances per 100 syllables, and relative to word frequency. Word frequency was the most

underperformant normalisation metric hindering performance of the disfluency features to below the baseline. This may be due to the syllabic nature of disfluencies and that an ASR system does not provide a reliable output for features such as speech errors and non-speech events, limiting the utility of word frequency as a way of measuring the length of discourse. Speaker duration worked well, achieving performance slightly above the baseline at 79.2% (HC-MCI/ND). Syllabic relative frequency, a common method for counting disfluencies, was the most effective outperforming the baseline with 83.2% (HC-MCI/ND). Syllabic relative frequency was not found to be beneficial for the normalisation of textual features, therefore the final SVMs utilised conventional normalisation as detailed in [12] for these features.

5. RESULTS & DISCUSSION

The linguistic features used in this study were developed for what is arguably the most popular corpus of ND speech; the DementiaBank Pitt Corpus [26]. Tested on Pitt Cookie Theft, an SVM was built with an accuracy of 78.8% on a 2-class HC-MCI/ND discrimination task; however these features perform with an accuracy of 49.6% on the Sheffield IVA data. There is a large limitation in this area of research caused by data scarcity and mono-corpus challenge based research that results in difficulties in determining the generalisability of feature sets across different corpora. These results highlight the need to test ACDD systems on unseen corpora to ensure reliable efficacy outside of a single data set. For this reason, linguistic features were not included in our baseline. A preliminary series of experiments on the Pitt Corpus demonstrated enhanced performance with linguistic features and disfluency features combined. Therefore these features were retained for the final Sheffield IVA classification system to acknowledge the potential of these features to increase generalisability.

Table 3: HC-MCI/ND SVM Accuracy Results (5-fold cross validation)
{M-#: Manual, A-#: Automatic}

Model	Accuracy
Acoustic Only	77.6%
Linguistic Only	57.6%
Acoustic + Linguistic	68.8%
Manual Disfluency	**88.8%**
Automatic Disfluency	78.4%
Ac + Ling + Auto Disfluency	83.2%

An SVM trained exclusively on the manually annotated disfluency features yielded a baseline performance of 88.8%. This strongly suggests human annotated disfluency features have the capacity to enhance the performance of existing models. In order to automate disfluency feature extraction certain features were omitted. Blends and substitutions (categorised as speech errors in the above schema) were removed from the feature bank due to falling below a threshold of 10 instances within our subset of the Sheffield IVA dataset.

Table 3 demonstrates improved performance of 5.6% from the acoustic only SVM baseline trained on the Sheffield IVA data when using an SVM enhanced with automatically calculated disfluency features. The performance of the manually annotated features remains substantially higher at 88.8%. Overall these results demonstrate that disfluency features provide meaningful performance improvements to ACDD SVMs in the context of HC-MCI/ND discrimination when using conventional and interpretable linguistic/acoustic feature sets.

6. CONCLUSION

This study has demonstrated the potential for enhancing ACDD systems through the automatic extraction of disfluency features. The difference between manual and automatic feature performance remains large; whilst this paper has demonstrated the potential for boosting system efficacy more research is needed to optimise approaches for extracting each feature. Future research may wish to focus on a larger scale analysis of disfluency feature generalisability. Our future research will focus on the behaviour of linguistic features to supplement this work on disfluency features and aim to inform greater intuition about the behaviour of different modes of automated features in real world environments.

7. ACKNOWLEDGEMENTS

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]

8. REFERENCES

[1] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent

virtual agent for the detection of early signs of dementia," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2732–2736.
[2] W. H. Organization *et al.*, "Global status report on the public health response to dementia," 2021.
[3] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: a review," *Expert systems with applications*, vol. 150, p. 113213, 2020.
[4] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, "Phonological and articulatory impairment in alzheimer's disease: a case series," *Brain and language*, vol. 75, no. 2, pp. 277–309, 2000.
[5] P. Gómez-Vilda, V. Rodellar-Biarge, V. Nieto-Lluis, K. L. de Ipiña, A. Álvarez-Marquina, R. Martínez-Olalla, M. Ecay-Torres, and P. Martínez-Lage, "Phonation biomechanic analysis of alzheimer's disease cases," *Neurocomputing*, vol. 167, pp. 83–93, 2015.
[6] N. Biassou, M. Grossman, K. Onishi, J. Mickanin, E. Hughes, K. Robinson, and M. D'Esposito, "Phonologic processing deficits in alzheimer's disease," *Neurology*, vol. 45, no. 12, pp. 2165–2169, 1995.
[7] A. J. Astell and T. A. Harley, "Tip-of-the-tongue states and lexical access in dementia," *Brain and language*, vol. 54, no. 2, pp. 196–215, 1996.
[8] V. Vincze, G. Szatlóczki, L. Tóth, G. Gosztolya, M. Pákáski, I. Hoffmann, and J. Kálmán, "Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild alzheimer's disease," *Clinical Linguistics & Phonetics*, vol. 35, no. 8, pp. 727–742, 2021.
[9] V. Vincze, G. Gosztolya, L. Tóth, I. Hoffmann, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "Detecting mild cognitive impairment by exploiting linguistic information from transcripts," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 181–187.
[10] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.
[11] M. Thomas, N. Pevy, and T. Walker, "Disfluencies and cognitive decline: An investigative study," in *ICPLA 2023- Conference of the International Clinical Phonetics and Linguistics Association*, 2023.
[12] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.
[13] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.
[14] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.
[15] H. Goodglass and E. Kaplan, *The assessment of aphasia and related disorders*. Lea & Febiger, 1972.
[16] K. McDougall and M. Duckworth, "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16–27, 2017.
[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." [Online]. Available: http://www.praat.org/
[18] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.
[19] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021.
[20] U. Gut, "Analysing phonetic and phonological variation on the suprasegmental level," *Research Methods in Language Variation and Change*, pp. 244–259, 2013.
[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
[22] S. Hollands, D. Blackburn, and H. Christensen, "Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation," in *Proc. Interspeech 2022*, 2022, pp. 3958–3962.
[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
[24] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.
[25] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
[26] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

CHARACTERIZING RHYTHM IN DYSARTHRIC SPEECH USING THE TEMPORAL ENVELOPE

Eugenia San Segundo[1], Jonathan Delgado[2] and Lei He[3]

[1]Dept. of Spanish Language & General Linguistics, UNED (Spain), [2]Dept. of Developmental & Educational Psychology, La Laguna University (Spain), [3]Dept. of Computational Linguistics, University of Zurich (Switzerland)

[1]esansegundo@flog.uned.es, [2]extjdelgado@ull.edu.es, [3]lei.he@uzh.ch

ABSTRACT

This study investigated the rhythmic characteristics of 15 dysarthric speakers and 15 non-dysarthric speakers. Speech rhythm was viewed from MacNeilage's frame/content theory. Specifically, the temporal envelope was analysed. The subjects, native speakers of Canarian Spanish, read four phonetically-balanced sentences. Five common spectral measures (centroid, spread, rolloff, flatness, and entropy) were computed from the temporal envelope of each sentence. A binomial logistic regression model was built to evaluate how well the five spectral measures can characterize speech rhythm in dysarthric speakers versus the control group. The results show that the dysarthric group present a significantly lower centroid and lower spread. Possible explanations for these results are discussed in relation to previous phonetic studies into the rhythmic speech patterns of dysarthric speakers.

Keywords: dysarthria, rhythm, temporal envelope, Spanish, binomial logistic regression.

1. INTRODUCTION

This paper is an exploratory investigation into the rhythmic characteristics of dysarthria. Speech rhythm is viewed here from MacNeilage's frame/content perspective, where the mouth opening-closing cycles (and thence the temporal modulations) constitute the rhythmic frame in speech [1]. In particular, the method followed in this article for speech rhythmic characterization consists in the extraction of five spectral measures common in audio engineering (centroid, spread, rolloff, flatness, and entropy), computed from the temporal envelope of sentences in read speech.

Dysarthria is a speech disorder with neurological origin that causes difficulties in speech motor programming and execution [2]. Its evaluation is complex due to the heterogeneity of its symptoms [3] and requires the combination of a battery of objective and subjective tests [4].

In the upcoming section, literature is reviewed in terms of: (1) previous phonetic studies on dysarthric speech and (2) the frame/content approach to the study of rhythm.

2. PREVIOUS STUDIES

2.1. Phonetic studies on dysarthria

Different acoustic analyses of dysarthric speech have been conducted in order to obtain objective data that distinguish individuals with this type of speech disorder, or to determine the severity of dysarthria by looking for measures that correlate with the level of speech intelligibility. Acoustic analyses are also useful to distinguish among different dysarthria subtypes. Commonly analysed acoustic measurements include: the area of the vowel space [3, 4, 5], the range and slope of the second formant [3, 6, 7], VOT [5, 8], temporal measures, such as speech rate [3, 6, 9], or a combination of temporal, spectral, and cepstral parameters [3, 10].

Summarizing the findings of these investigations is not a simple task, as some of them focus on one particular type of dysarthria or analyse patients of a particular language. Hence results cannot be extrapolated to other dysarthria subtypes or to different languages. While some processes affecting motor speech planning and execution in dysarthria could be largely language-independent, others could be prone to cross-language differences [11].

In terms of rhythmic analyses, since Holmes's historical paper [12], dysarthria has been described as presenting reduced articulation rate, irregular duration contrasts between stressed and unstressed syllables and a scanning rhythm (*staccato*) [10]. However, this description refers to one particular type of dysarthria: ataxic dysarthria. Recent investigations tend to study a wide range of dysarthria subtypes, although results are sometimes disparate. As a case in point, while Liss et al. [9] concluded that rhythm metrics (acoustic measures of vocalic and consonantal segment durations) allow to distinguish control speech from dysarthria and to discriminate dysarthria subtypes, another study [13] concluded that none of the rhythm metrics based on segmental durations could differentiate disordered from healthy speakers, "despite clear perceptual differences,

suggesting that factors beyond segment duration impacted on rhythm perception" [13, p.1].

Investigations into the perceptual characterization of dysarthria are rarer, although perceptual assessment in clinical studies is still considered the gold standard against which acoustic measures are compared [14]. In [15] a simplified version of the Vocal Profile Analysis (VPA) protocol [16] is used to analyse the voice quality of dysarthric speakers perceptually. The authors found that the perceptual settings 'vocal tract tension' and 'laryngeal tension' were the most useful to characterize this disorder. In particular, trained raters were more likely to score a voice with high vocal tract tension if it belonged to the dysarthric group than if it belonged to the control group of neurologically healthy speakers. The authors suggest that acoustic-perceptual assessment through this protocol could be an important complement to other types of evaluations, especially because several settings of the VPA refer to supraglottic structures commonly affected in dysarthria, a speech disorder in which the muscles used to produce speech are damaged, paralyzed, or weakened. However, more research is necessary into the acoustic correlates of perceptual settings such as 'vocal tract tension'. San Segundo et al. [17] suggested that low inter-rater agreement for this perceptual dimension could be due to the different salience of prosodic aspects in each rater. They showed that the perceptual ratings of one evaluator correlated with the mean intensity variability across syllables. In contrast, the ratings of the other evaluator correlated positive and significantly with two rhythmic measures related to mean consonant duration.

Rhythm is a key element of prosody, together with intonation and tempo. Since dysprosody, or prosody degradation, is considered a hallmark of dysarthria [11], the current investigation has focused on the rhythmic characterization of this speech disorder. This will complement previous investigations, such as [15], where 13 acoustic features (mainly spectral and cepstral measures) were analysed, together with voice quality perceptual settings, with the aim of describing dysarthria phonetically.

Due to the fact that traditional acoustic approaches to rhythm (i.e. durational variability in different phonetic intervals) in dysarthric speakers have given rise to disparate results [9, 13], a different acoustic method is proposed here, which analyses rhythm from the speech temporal envelope.

2.2. Speech rhythm: The frame/content perspective

The key idea which lies behind the concept of rhythm is cyclicity or regularity. The first phonetic studies of speech rhythm distinguish two major types of regularity: isochronous syllables and isochronous feet [18, 19]. However, it has not been possible to corroborate strict syllable- or foot-isochrony in traditionally considered syllable-timed and stress-timed languages [20-21]. Different rhythmic metrics have since been developed by calculating durational variabilities in different phonetic intervals, usually consonantal and vocalic intervals [22]. These metrics have been used to characterize the rhythm of languages or speaking styles, but also to evaluate speech disorders [9, 13]. Other approaches to speech rhythm can be roughly grouped into 'prominence', 'modulation' and 'coupling strengths' [23], which basically depend on which aspect gives rise to different rhythmicities: syllable intensity, peak frequencies in the amplitude envelope spectrum, or coupling oscillations.

A recent investigation [23] suggests that MacNeilage's frame/content theory of speech evolution unifies the different approaches to the study of rhythm: "MacNeilage held that speech rhythm evolved from pre-existing cyclical mandibular movements in ancestral primates in the form of lip-smacking [1]. It is also believed that the coupling between mouth opening-closing cycles and vocalization emerged en route to human evolution: the sonority of speech typically waxes and wanes with mouth opening and closing gestures [24, 25]. Such opening/closing alternations are organized into syllable-sized units corresponding to the temporal modulations, which constitute the rhythmic frames; the open and closed phases are filled with vocalic and consonantal contents. More recently, Fitch [26] endorsed the frame/content mechanism and argued that it affords the evolution of the language system at large" [23, p. 568].

As explained in [23], the rhythm metrics quantifying the durational variability of vocalic and consonantal intervals typically focus on the *content* perspective, whereas the modulation-based approaches target the *frame* perspective with different emphases: recurring frequencies in the temporal envelope [27, 28] and coordination between envelopes at slower and faster rates [29]. This is how the different approaches to speech rhythm are related to the frame/content theory. Furthermore, Erickson and colleagues [30] demonstrated that the jaw displacement well explained the metrical structure and subjective prominence ratings in a number of languages in which a more prominent syllable was typically associated with a lower jaw position. To integrate both temporal modulation and the opening and closing cycles of the mouth for the characterization of the rhythmic frame, spectral coherence has been used [23, 31]. The coherence is calculated from the spectra of two signals to show the

connectivity between them in terms of their correlation in the frequency domain. For instance, He [23] uses spectral coherence between the temporal envelope and the mouth opening and closing kinematics to characterize speech rhythm in L2 English speakers of Mandarin. He found that the native group was significantly higher than the non-native group in terms of spectral centroid and spread; two of the five variables extracted from the spectral coherence, which can also be simply extracted from the temporal envelope (see section 3.3).

3. METHOD

3.1. Participants

30 subjects voluntarily participated in this study: 15 with dysarthria (mean age 42.93, SD 10.31) and 15 neurologically healthy (mean age 41.86, SD 13.62). The two experimental groups (dysarthria and control) were sex matched. They were all speakers of Canarian Spanish. Within the dysarthric group, 10 participants presented ataxic dysarthria, 2 spastic dysarthria and 3 mixed dysarthria; with different medical origins. After a preliminary analysis, one dysarthric participant presenting mixed dysarthria was discarded. Her audio samples presented signal saturation and stammering. These aspects were deemed unfit for this type of acoustic analyses.

3.2. Recording setup and speech samples

All recordings were conducted in a soundproof booth with an AKG C544L head-mounted condenser microphone. They were digitized at a sampling rate of 44.1 kHz and 16 bits of resolution using the audio interface Alesis io2 express. The signal-to-noise ratio (SNR) was measured post hoc to check the level of environmental noise of the voice recordings. All samples were consistent with the recommended threshold proposed by [32]. The speech material consisted in reading aloud four phonetically balanced sentences of the Spanish Matrix Sentences Test [33].

3.3. Acoustic analysis

First, the acoustic signal per sentence was bandpass filtered between 700 and 1300 Hz (100 Hz smoothing) to keep the vocalic energy while removing the glottal energy and obstruent noise. This filter has been used to detect the P-centers or "beats" in the speech signal [35]. Then, the filtered signal was full-wave rectified and downsampled to the Nyquist frequency of 20 Hz, yielding the temporal envelope. Five spectral measures (CENTROID, SPREAD, ROLLOFF, FLATNESS, and ENTROPY) [34] were calculated from the temporal envelope of each sentence. The CENTROID calculates the "balancing point" in the coherence and serves as a point estimate of the coherence. The SPREAD calculates to what extent the coherence disperses around the centroid. The ROLLOFF indicates the degree of skewness in the coherence. The FLATNESS and ENTROPY quantify the amount of unpredictability or disorder in the spectrum. For details please look at the supplementary material (https://osf.io/3z9uq).

3.4. Statistical analysis

All statistical procedures were performed using R [35] and associated packages (*lme4*, *car*, *effects*, and *ggplot2*). The generalized linear mixed model (binomial logit) was used for data analysis. The five spectral variables extracted from the temporal envelope were modeled as the fixed-effect numeric predictors; GROUP (control and dysarthric speakers) was modeled as the dichotomous response variable. Both groups produced the same set of sentences, and thus SENTENCE was modeled as a random factor.

4. RESULTS

Table 1 summarizes the model-fitting results of all predictors. The reference of the response variable group is C (control speakers). The model explains between 66.54% and 88.76% of the dependent variable variation (Cox & Snell's R^2 = 0.6654; Nagelkerke's R^2 = 0.8876).

Predictor	Estimate	Std. Error	z
CENTROID	–4.51	1.72	–2.63 **
SPREAD	–5.95	2.84	–2.09 *
FLATNESS	10.87	11.34	0.96
ROLLOFF	0.22	0.73	0.30
ENTROPY	3.11	1.88	1.66 .

Table 1: Results of the fixed-effect predictors in the generalized linear mixed model (binomial logit). **p<0.01; *p<0.05; . p<1.

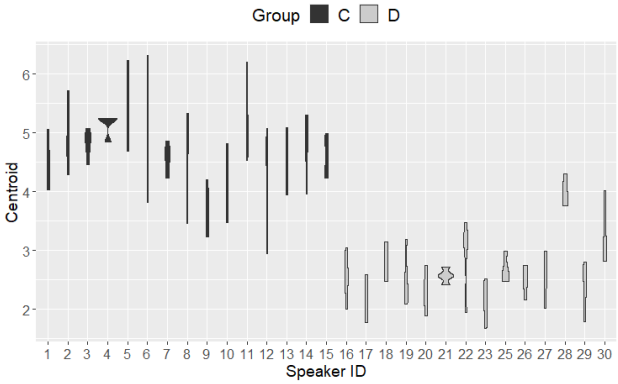


Figure 1: Boxplots showing the distribution of CENTROID values per speaker. C = Control; D = Dysarthria.

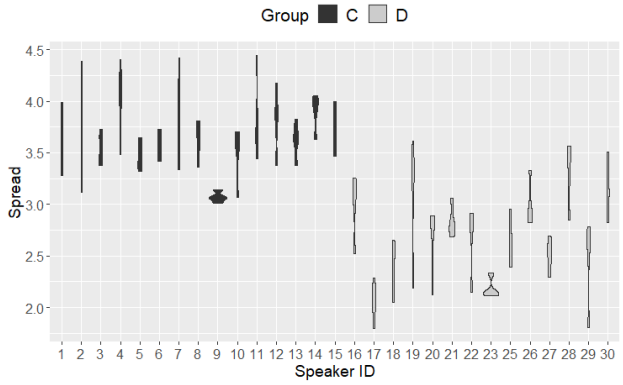


Figure 2: Boxplots showing the distribution of SPREAD values per speaker. C = Control; D = Dysarthria.

Fig. 1 and Fig. 2 show the distribution of CENTROID and SPREAD values, respectively, per speaker and group (black for the control group and grey for the dysarthric speakers).

5. DISCUSSION

This paper has investigated the rhythmic characteristics of read sentences in a group of Spanish speakers, distinguishing dysarthric speakers and non-dysarthric speakers. For this purpose, five variables capturing the temporal envelope of the sentences were computed and analysed as numeric predictors in a binomial logistic regression model.

Dysarthric speakers in general showed significantly lower CENTROID frequencies in the temporal envelope (estimate = –4.51, with $p<0.01$). Although this study was mainly exploratory, this finding agrees with our expectations. Dysarthric speech typically exhibits a slow speech rate due to slow articulatory movements. As a result, a stretched rhythmic *frame* was expected, together with a centroid shifted towards lower frequencies, in comparison with those of the control speakers.

A more restricted spread could also be anticipated in the dysarthric group. Our results corroborate that indeed dysarthric speakers in general showed a significantly narrower spectral SPREAD in the temporal envelope (estimate: –5.95, with $p<0.05$) in comparison with the non-dysarthric group. This suggests that the oscillations in the rhythmic frame were more regular around the centroid. A possible explanation for this could be that jaw displacements in the non-dysarthric population are variable while jaw oscillations in dysarthria do not exhibit large variation. Instead, in terms of openness, both the jaw and the mouth remain stationary in these speakers throughout utterance production. These results can be explained by the fact that dysarthria is a motor speech disorder in which the muscles used to produce speech are damaged, paralyzed, or weakened.

Our acoustic findings agree with impressionistic perceptual descriptions of dysarthria. For instance, Ziegler [10] highlights reduced articulation rate and 'drawling' speech as clinical symptoms of ataxic dysarthria. In future studies we will consider analysing ataxic, spastic and mixed dysarthric patients separately. Upon observation of the boxplots (Figs 1-2), there seems to be some differences among the speakers belonging to different subtypes of dysarthria. For instance, speakers 16, 18, 21, 22, 26, 27, 29 and 30 are ataxic speakers while the rest of dysarthric speakers present either mixed or spastic dysarthria. Possible explanations for these differences could be related to what Ziegler [10, p.14] states about ataxic dysarthria, namely that voice impairment in this disorder can be more irregular than in other dysarthrias: "vocal pitch or loudness may suddenly change and thereby interrupt the natural intonation pattern, and voice quality may change from strained-strangled to breathy or rough (…). Likewise, articulation may change between lenis and fortis consonant production, with staccato transitions between syllables or with syllable lengthening".

6. CONCLUSIONS

This study have shown that the dysarthric speakers analysed present a significantly lower centroid and lower spread in the temporal envelope of their read sentences, in comparison with a control group of non-dysarthric speakers. These variables can discriminate dysarthric and non-dysarthric speakers, although more studies, preferably with a larger number of subjects and in different languages, are necessary to ensure replicability and validation of these findings.

The literature has shown that *content*-based approaches to rhythm (durational variability in different phonetic intervals) have resulted in disparate findings in dysarthric studies. The method proposed here analyses rhythm from the speech temporal envelope and presents the advantage of not requiring audio transcriptions or phonetic interval segmentation. Previous L2 studies have shown that rhythmic characteristics may not be sufficiently explained from the traditional *content*-based perspective alone [23], highlighting the importance of the *frame* aspect in rhythm acquisition. In a similar way, investigations on speech disorders can benefit from this combined perspective. Since the sonorant/obstruent alternations typically follow the opening-closing mouth movements, according to the frame-content theory [1], future dysarthric studies could analyse the spectral coherence between the temporal envelope and mouth opening-closing articulatory data, as well as correlations with jaw openness ratings in the VPA perceptual protocol.

7. REFERENCES

[1] MacNeilage, P.F. 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499–511.

[2] Melle, N. 2007. *Guía de intervención logopédica en la disartria*. Madrid: Editorial Síntesis.

[3] Kim, Y., Kent, R. D., Weismer, G. 2011. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *J Speech Lang Hear Res* 54(2), 417–429.

[4] Sapir, S., Ramig, L., Spielman, J., Fox, C. 2010. Formant centralization ratio (FCR) as an acoustic index of dysarthric vowel articulation: comparison with vowel space area in Parkinson disease and healthy aging. *J Speech Lang Hear Res.* 53, 114–125.

[5] Delgado Hernández, J. 2016. Medida de la severidad de la disartria atáxica a través del análisis acústico, *Estudios de fonética experimental*, 25, 149–166.

[6] Yunusova, Y., Green, J. R., Greenwood, L., Wang, J., Pattee, G. L., Zinman L. 2012. Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica*, 64(2), 94–102.

[7] Delgado Hernández, J., Izquierdo Arteaga, L. 2016. Relación entre las pendientes del segundo formante y las alteraciones motoras del habla en la disartria. *Revista de logopedia, foniatría y audiología* 36(2), 71–76.

[8] Auzou, P., Ozsancak, C., Morris, R. J., Jan, M., Eustache, F., & Hannequin, D. 2000. Voice onset time in aphasia, apraxia of speech and dysarthria: a review, *Clinical linguistics & phonetics*, 14(2), 131–150.

[9] Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., Caviness, J. 2009. Quantifying speech rhythm abnormalities in the dysarthrias, *J. Speech Lang. Hear. Res*. 52(5), 1334–1352.

[10] Ziegler, W. 2016. The phonetic cerebellum: cerebellar involvement in speech sound production. In: Marien, P., Manto, M. (eds), *The linguistic cerebellum*. Academic Press, 1–32.

[11] Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., Sadat, J. 2017. A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. *J Phonetics* 64, 156–167.

[12] Holmes, G. 1917. The symptoms of acute cerebellar injuries due to gunshot injuries. *Brain*, 40(4), 461–535.

[13] Lowit, A. 2014. Quantification of rhythm problems in disordered speech: A re-evaluation, *Philos. Trans. R. Soc. B.,* 369, 20130404.

[14] Ma, E.P.M., Yu, E.M.L. 2005. Multiparametric evaluation of dysphonic severity. *J. Voice* 20, 380–390.

[15] San Segundo, E., Delgado, J. 2021. A preliminary approach to the acoustic-perceptual characterization of dysarthria. *Proc. ISAPh 2021*, Universitat Rovira i Virgili, Tarragona (pp. 6-8).

[16] Laver, J. 1980. *The phonetic description of voice quality*. London: Cambridge Studies in Linguistics.

[17] San Segundo, E., Schwab, S., Dellwo, V., He, L., González, J. A. M. 2017. Perception of vocal tract tension: Exploring possible prosodic correlates. In *Tendencias actuales en fonética experimental: Cruce de disciplinas en el centenario del Manual de Pronunciación Española* (pp. 79-82). UNED.

[18] Jones, D. 1922. *An Outline of English Phonetics*. G. E. Stechert & Co.

[19]Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[20] Roach, P. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In: Crystal, D. (ed), *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F. R. Palmer*, Edward Arnold, 73–79.

[21] Dauer, M. 1983. Stress-timing and syllable-timing reanalyzed, *J. Phon*. 11, 51–62.

[22] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal, *Cognition* 73, 265–292.

[23] He, L. 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am*. 152(1), 567–579.

[24] Ghazanfar, A. A., Chandrasekaran, C., Morrill, R.J. 2010. Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: Implications for the evolution of audiovisual speech, *Eur. J. Neurosci*. 31, 1807–1817.

[25] Morrill, R. J., Paukner, A., Ferrari, P.F., Ghazanfar, A.A. 2012. Monkey lipsmacking develops like the human speech rhythm, *Dev. Sci.* 15, 557–568.

[26] Fitch, W. T. 2019. Sequence and hierarchy in vocal rhythms and phonology, *Ann. N.Y. Acad. Sci.* 1453, 29–46.

[27] Tilsen, S., Johnson, K. 2008. Low-frequency Fourier analysis of speech rhythm. *J. Acoust. Soc. Am.* 124 (2), EL34–EL39.

[28] Tilsen, S., Arvaniti, A. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages, *J. Acoust. Soc. Am*. 134, 628–639.

[29] Leong, V., Stone, M.A., Turner, R.E., Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception, *J. Acoust. Soc. Am.* 136, 366–381.

[30] Erickson, D., Kawahara, S. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns, *Linguist. Vanguard* 2, 20150025.

[31] Chandrasekaran, C., Trubanova, C., Stillittano, S., Caplier, A., Ghazanfar, A.A. 2009. The natural statistics of audiovisual speech, *PLoS Comput. Biol.* 5, e1000436.

[32] Deliyski, D.D., Shaw, H.S., Evans, M.K. 2005. Adverse effects of environmental noise on acoustic voice quality measurements. *J Voice* 19, 15–28.

[33] Hochmuth, S., Brand, T., Zokoll, M.A., Zenker, F., Wardenga, M., Kollmeier, B. 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol.* 51, 536–544.

[34] Giannakopoulos, T., Pikrakis, A. 2014. *Introduction to Audio Analysis*, Academic Press, 78–86.

[35] R Core Team. 2023. *R: A language and environment for statistical computing (R4.1.0)* [computer program].

Accommodation to Vocal Pitch in Children with Autism

Anqi Hu[1], Zhenghan Qi[1, 2], Kathryn Franich[1, 3]

[1]University of Delaware, [2]Northeastern University, [3]Harvard University
anqihu@udel.edu

ABSTRACT

Speech accommodation is a common phenomenon in typically developing (TD) adults. However, little is known regarding whether TD children and autistic children, who have atypical prosodic profiles and social communication skills, also spontaneously accommodate to others' speech. Thirty 5-to-10-year-old autistic children and 30 age-matched TD children repeated sentences after two model talkers either speaking with their original pitch or with artificially raised pitch. Acoustic analysis revealed that only the autistic children displayed evidence of pitch accommodation. In a follow-up experiment, 25 TD children were explicitly asked to copy the voice of the talkers; here, they showed a similar degree of pitch accommodation to the autistic children. We discuss results in the context of developmental patterns of speech accommodation and suggest TD children were less likely to accommodate to pitch changes due to their communicative function, and to the slightly less typical-sounding voice of the higher-pitch model talker.

Keywords: speech accommodation, f0, pitch, convergence, imitation, autism, speech development

1. INTRODUCTION

Speech accommodation (also known as speech *entrainment, imitation,* or *convergence*) is a process of adapting to the speech characteristics of another talker. Robust evidence has demonstrated that adults can accommodate to different dimensions of speech including pitch, speech rate, vowel spectra, and voice onset time of model talkers or conversation partners in both experimental contexts and naturalistic speech settings [1].

Although speech accommodation has been under-investigated in children, existing studies in children found evidence of phonetic accommodation using measures of fundamental frequency (F0), voice onset time (VOT), and lexical stress [2]–[4]. For example, Ko and colleagues [2] examined pitch convergence between toddlers and their mothers over the course of one to two months and found greater similarities in pitch between mother-and-child dyads in their last conversational turn compared to the first one. Nielsen [3] examined accommodation to VOT in 4-year-old and 8-year-old children and adults. All age groups showed effect of accommodation and produced /p/ with lengthened VOT after exposure, but the effect was stronger in children than in adults, with comparable effects between younger and older children. These studies hinted at the spontaneous nature of phonetic accommodation in children and a potential reduced speech accommodation effect across development.

Children's ability to accommodate to a model has been studied as a broad cognitive phenomenon for decades under the umbrella term "imitation". Robust evidence has demonstrated that children can imitate gestures, actions, and facial expressions since early infancy [5]. Speech accommodation can be considered as an example of spontaneous imitation that occurs automatically without explicit instructions [6]. It has been proposed that social affiliations and liking towards a model talker, as well as the talker's proficiency, social status, and resemblance to oneself can all influence likelihood of imitation [7]. In the speech domain, children are found to imitate more when models are more friendly [8], speak with a native accent as opposed to a foreign accent [9], or are adults as opposed to children [10]. It has also been argued that the development of these social biases during imitation are contingent upon children's age and development of different cognitive skills [7]. Although there is still debate regarding what kind of contexts and type of biases can elicit different profiles of imitation, existing evidence demonstrates that typically developing (TD) children show spontaneous imitation across domains from an early age, and the fidelity of their imitation is guided by the development of different social biases towards the models.

Children with autism spectrum disorder (ASD) often display atypical prosodic profiles [11]. Yet, little is known regarding the processes of speech accommodation in autism, and whether speech accommodation is impacted by social biases in this population with social communication impairment. Studies of speech accommodation in autistic children could provide important information about the source of their prosodic differences.

Around 1 in 40 to 59 children in the US are diagnosed with ASD [12]. Among individuals with autism that do acquire spoken language, atypical speech prosody has been reported and included as a

crucial diagnostic criterion [13]. In the domain of pitch, the focus of the present study, several studies have reported similar mean pitch or pitch variability in sentences between individuals with autism and controls [14], [15], though others have found a higher mean pitch and a wider pitch range in ASD [16], [17].

In one study with 9-to-15-year-old autistic children, Lehnert-LeHouillier and colleagues examined speech accommodation during an interactive picture description task led by a female experimenter [18]. Only 25% of autistic children showed increasing similarity in their mean F0 values to the experimenter over the course of the task, compared to 66% in the TD group. While the TD group converged to the pitch of the experimenter, the similarity in pitch found in the ASD group was mainly driven by speech accommodation of the experimenter.

Given the role that social dynamics can play in accommodation, it isn't yet clear from these results whether autistic children cannot accommodate to adults as well as TD children, or whether they may have failed to accommodate due to some aspect of the social dynamic in the study. Autistic children are known to display greater speech and language deficits in contexts involving higher cognitive and social demand [19]; therefore, a socially-demanding interactive task may have made speech processing—and, hence, accommodation—more challenging for the ASD group. In the present work, we explore speech accommodation among children with ASD in a less socially demanding context, that of a sentence repetition task. This type of task—where a participant listens to the speech of a model talker and repeats what has been said—have been used to successfully elicit accommodation in adults and children in previous work [20]. We hypothesize that, if children with ASD are going to display accommodation to a model talker, they may be more likely to do so in this type of simple repetition context.

2. METHODS

2.1. Participants

Thirty children with ASD (6.42 – 9.50 years; 3 girls) and 30 age-matched TD children (4.67 – 10.92 years; 14 girls) participated in this web-based study. Children with ASD were recruited by SPARK from nationwide clinical sites [21]. The autism diagnoses of children in SPARK were validated by medical records [22] and confirmed using Social Communication Questionnaire [23] and Repetitive Behavior Scale – Revised (RBS-R) [24]. All children with ASD were native speakers of English with no history of hearing loss, mutism, or (non-verbal) learning disabilities based on parental questionnaires. As will be discussed, an additional 25 TD children of comparable age (9 girls) were recruited for a follow-up experiment. TD children were neurotypical and had no history of speech or language impairments.

2.2. Stimuli and Design

The auditory Sentence Recall task (RSR) [25] was used to assess children's ability to accommodate to a model talker. This task has been validated in children from ages 5 to 9 and showed high sensitivity and specificity for detecting language impairment in this age range [25]. Sixteen sentences were included. Each sentence contained on average 11 words and 13.38 syllables (*SD* = 1.31). All sentences were monoclausal and contained a single intonational phrase and all discourse-new information.

In the **lower-pitch condition**, the auditory stimulus used was the original audio recording from Redmond [25]. A female speaker of standard American English pre-recorded all the sentences with a naturalistic and child-friendly prosody. In the **higher-pitch condition**, a different new female speaker of standard American English recorded the 16 sentences (mean duration = 4 seconds). To examine whether children will accommodate to the pitch of the model talkers, we artificially manipulated the pitch of the new model talker to be 50 Hz higher than the original talker in Redmond [25] using the Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) method in Praat [26], [27]. A norming study eliciting impressions from 11 adults showed that most listeners (n = 9) considered the speech of the higher-pitch model talker natural, clear, and child-friendly, but two adults mentioned the speech sounded slightly monotone. Other dimensions of the two speakers' speech (e.g., vowel spaces and volume) were similar, so that the primary dimension on which the two model talkers differed was pitch. The order of the 16 sentences was the same across conditions.

2.3. Procedure

The sentence repetition tasks were completed entirely through a video conference. The lower-pitch and higher-pitch conditions were arranged in a block design and the two blocks had a minimum of 50-minute interval between them to prevent interference. To avoid the confounding effect of fatigue on pitch lowering, the lower-pitch condition was always completed before the higher-pitch condition.

Upon starting the tasks, a pre-recorded audio instruction by a female native speaker of English told the child participants that they would play a talking game during which they would listen to some sentences from a girl, and they should "repeat exactly

what you hear." In the higher-pitch condition, children were given the same instruction to repeat the sentences. Here, no instruction was given regarding the pitch differences between conditions. As will be discussed, only the ASD group showed the expected pattern of spontaneous accommodation. In order to verify that TD participants could show similar patterns to the ASD group, a follow-up experiment was conducted with 25 TD participants which differed only in the instructions given: here, children were explicitly instructed to copy exactly both the voice and the words from the model talkers in both conditions.

2.4. Analyses

Mean F0 values were extracted from each unreduced vowel in the corpus using the pitch tracking algorithm in ProsodyPro [28]. To quantify pitch accommodation, we followed Babel & Bulatov [29] and calculated the absolute distance between the F0 of the model talkers and that of the participants. For each vowel and each participant, the ERB-transformed F0 of the vowel produced by the participant was deducted from that of the corresponding vowel produced by the model talker (F0 distance = |vowel F0 model talker – vowel F0 participant|). The pairs of vowels being compared here were matched on the word and sentence levels so that the same vowels are being compared between the participants and the model talkers. The distance in F0 was calculated separately for the lower-pitch and for the higher-pitch condition.

3. RESULTS

3.1. Group Differences in Pitch Accommodation

3.1.1. Mean F0

We first compared the F0 values between groups to determine the direction of pitch shifts in each condition. We analysed the fixed effects of *Group* (TD vs. pooled ASD and TD Follow-up; ASD vs. TD Follow-up), *Condition* (lower-pitch vs. higher-pitch), and their interactions on ERB-scaled F0 values using a mixed-effects linear regression model [30]. The model also included age, sentence recall accuracy scores, the total number of vowel tokens, and baseline pitch, extracted from the mean F0 values of the first 5 sentences in the lower-pitch condition, as covariates. Random intercepts for participants were included.

Collapsed across groups, we found a main effect of *Condition* ($\beta = 0.08$, $SE = 0.02$, $t = 3.98$, $p < 0.001$), indicating a significantly higher pitch in the higher-pitch than lower-pitch condition across groups. There was a significant interaction between *Condition* and *TD vs. pooled ASD and TD Follow-up group* ($\beta = -0.20$, $SE = 0.04$, $t = -5.43$, $p < 0.001$), reflected as a larger difference in mean F0 values between conditions in ASD and TD Follow-up groups than the TD group. The interaction between *Condition* and *ASD vs. TD Follow-up group* was not significant. Post-hoc *t*-tests within each group revealed a significantly higher mean F0 in the higher-pitch than the lower-pitch condition in the ASD ($p < 0.001$, $d = 0.13$) and the TD Follow-up groups ($p = 0.01$, $d = 0.08$; **Fig. 1**). No significant difference was found between conditions in the TD group.

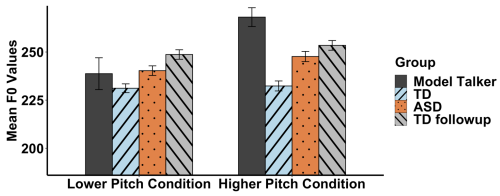


Figure 1: Mean F0 values in model talkers and children. The error bars represent 95% confidence interval.

3.1.2. F0 Distance

To examine the group differences in pitch accommodation, we analysed the fixed effects of *Group*, *Condition*, and their interactions on F0 distance using a similar model in the Mean F0 analysis. We found significant interactions between *Condition* and *Group* (*TD vs. pooled ASD and TD Follow-up*: $\beta = 0.18$, $SE = 0.03$, $t = 5.50$, $p < 0.001$; *ASD vs. TD Follow-up*: $\beta = 0.13$, $SE = 0.04$, $t = 3.47$, $p = 0.001$), suggesting condition difference was largest in the TD Follow-up and smallest in the TD group with the ASD group in between. Post-hoc *t*-tests within each group confirmed a significantly smaller mean F0 distance to the model talker in the higher-pitch than the lower-pitch condition in the ASD ($p = 0.003$, $d = 0.09$) and the TD Follow-up groups ($p < 0.001$, $d = 0.13$), indicating pitch accommodation to the higher-pitch model talker in the ASD and TD Follow-up groups (**Fig. 2**). The TD group showed a significantly larger F0 distance in the higher-pitch than the lower-pitch condition ($p < 0.001$), consistent with a lack of pitch accommodation to the model talker.

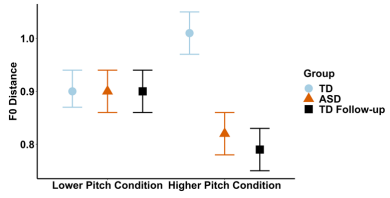


Figure 2: Mean F0 distance across groups. The error bars represent 95% confidence interval.

3.2. Impressions of Model Talker

We further asked all children in the TD Follow-up group about their impressions of the model talkers in order to examine the effect of social biases on speech accommodation. Children answered three questions: 1). Do the two talkers sound different? 2). How are their voices different? 3). Based on the voices, what do you think each one of them is like? All children reported the voices sounded different. 60% of the children (n = 15) provided specific descriptions on the vocal characteristics (e.g., "higher", "lower", "louder", "deeper voice", "squeaky voice") of the model talkers in question 2. In particular, three children noted that the lower-pitch model talker sounded more "normal" while the higher-pitch model talker sounded "unreal". In question 3, 64% of the children (n = 16) associated the voice of the model talkers with different social attributes (e.g., "more shy", "really active", "more relaxed", "chattier and outgoing", "not as much excited", "happy", "sad", "nicer"). Nine children noted the higher-pitch model talker may be more shy, quiet, or less excited.

4. DISCUSSION

The current study compared speech accommodation between autistic and TD children. Children with autism were able to spontaneously accommodate to the pitch of the model talker even when explicit imitation of pitch was not instructed. The TD children, however, only showed pitch accommodation when they were explicitly asked to copy the voice of the model talker.

One reason that may explain the group differences could be related to the aspect of speech children paid attention to during tasks. When multiple signals are present in a stimulus, ample behavioural evidence have demonstrated that children with autism show atypical attention patterns to certain dimensions of the signal [31]. For example, in speech perception tasks by Ploog and colleagues [32], the TD children showed a strong preference for content so that they chose the test sentences that had the same lexicon but different prosody compared to the training sentences, while the children with autism showed no preference between content and prosody, despite intact language comprehension and pitch perception across groups. Similarly, in our study, the TD group only showed accommodation to pitch when explicitly instructed to copy the voice of the model talkers. When only asked to repeat the sentences, the TD group may have perceived the semantic content and lexicon in the sentences more functionally important or more task-relevant than the pitch of the model talkers, since pitch only varied as a function of individual talker vocal characteristics, and did not contribute to a change in meaning across conditions The TD group thus attended selectively to the content and only copied the word form used and the semantic information in the sentences with their own pitch. The ASD group, however, may have paid attention to all attributes of speech from the model talkers and accommodated to both details, the pitch and content. Nevertheless, the role of selective attention can only be inferred in our current paradigm. Future studies using more implicit electrophysiological or other brain imaging techniques to measure auditory attention will be needed to directly assess whether children engage more attention to meaning or to prosody during speech accommodation.

Differences in social biases towards the model talkers may also explain the group differences in pitch accommodation. Previous studies have found that TD adults were more likely to accommodate to model talkers that were deemed more attractive [33]. TD children at the age of 3 to 5 years were found to copy more novel word labels and actions from proficient and knowledgeable models than models who provide inconsistent or inaccurate information. TD children at age 4 showed more bias towards models with higher social status who were endorsed by bystanders than those that were not [34]. The effect of social biases on imitation develops and becomes more robust as children grow older [7]. In our study, the TD children in the follow-up experiment seemed to associate the voice of the model talkers with different social attributes ("chattier" vs. "more shy"). Some TD participants commented on the voice of the higher-pitch model talker being less typical (e.g., "less normal", "unreal voice") than the lower-pitch model talker. Given that 4-to-5-year-old children were already found to show social biases in imitation, the markedness of the higher-pitch model talker might explain the lack of pitch accommodation in 5-to-10-year-old TD children in this study if they considered the higher-pitch model talker as less likable or less proficient. Compared to TD individuals, autistic adults were found to show attenuated social biases and stereotypes in tasks where their tendency to associate a social group with particular attributes were tested implicitly [35]. If the attenuated social biases in autistic adults are an outcome of atypical development of social communication skills, then the pitch accommodation to the higher-pitch model talker in the autistic children may be explained by a delayed development of social biases in this group. Future studies examining accommodation in naturalistic and conversational settings would need to investigate whether autistic children form the same social attributes based on the voice of the model talkers and whether that influences their pitch accommodation.

5. REFERENCES

[1] J. S. Pardo, "Measuring phonetic convergence in speech production," *Front. Psychol.*, vol. 4, no. AUG, p. 559, Aug. 2013.
[2] E. S. Ko, A. Seidl, A. Cristia, M. Reimchen, and M. Soderstrom, "Entrainment of prosody in the interaction of mothers with their young children," *J. Child Lang.*, vol. 43, no. 2, pp. 284–309, Mar. 2016.
[3] K. Nielsen, "Phonetic Imitation by Young Children and Its Developmental Changes," *J. Speech, Lang. Hear. Res.*, vol. 57, no. 6, pp. 2065–2075, Dec. 2014.
[4] F. Subiaul, K. Winters, K. Krumpak, and C. Core, "Vocal overimitation in preschool-age children," *J. Exp. Child Psychol.*, vol. 141, pp. 145–160, 2016.
[5] S. S. Jones, "The development of imitation in infancy," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1528, pp. 2325–2335, Aug. 2009.
[6] T. L. Chartrand and J. L. Lakin, "The antecedents and consequences of human behavioral mimicry," *Annu. Rev. Psychol.*, vol. 64, pp. 285–308, 2013.
[7] L. A. Wood, R. L. Kendal, and E. G. Flynn, "Whom do children copy? Model-based biases in social learning," *Developmental Review*, vol. 33, no. 4. Academic Press, pp. 341–356, 01-Dec-2013.
[8] M. Nielsen, "Copying actions and copying outcomes: Social learning through the second year," *Dev. Psychol.*, vol. 42, no. 3, pp. 555–565, 2006.
[9] K. D. Kinzler, K. H. Corriveau, and P. L. Harris, "Children's selective trust in native-accented speakers," *Dev. Sci.*, vol. 14, no. 1, p. 106–111, 2011.
[10] V. K. Jaswal and L. A. Neely, "Adults don't always know best preschoolers use past reliability over age when learning new words," *Psychol. Sci.*, vol. 17, no. 9, pp. 757–758, Sep. 2006.
[11] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: A critical review," vol. 38, no. 4. *Int J Lang Commun Disord*, pp. 325–350, 2003.
[12] M. D. Kogan *et al.*, "The prevalence of parent-reported autism spectrum disorder among US children," *Pediatrics*, vol. 142, no. 6, 2018.
[13] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "'Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis,'" *Autism Res.*, vol. 10, no. 3, 2017.
[14] J. J. Diehl and R. Paul, "Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders," *Res. Autism Spectr. Disord.*, vol. 6, no. 1, pp. 123–134, 2012.
[15] H. Green and Y. Tobin, "Prosodic analysis is difficult ... but worth it: A study in high functioning autism," *International Journal of Speech-Language Pathology*, vol. 11, no. 4 pp. 308–315, 2009.
[16] M. Sharda *et al.*, "Sounds of melody—Pitch patterns of speech in autism," *Neurosci. Lett.*, vol. 478, no. 1, pp. 42–45, Jun. 2010.
[17]A. Nadig and H. Shaw, "Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners," *J. Autism Dev. Disord.*, vol. 42, no. 4, pp. 499–511, Apr. 2012.
[18] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum," *Front. Psychol.*, vol. 0, p. 2718, Oct. 2020.
[19] M. Losh and L. Capps, "Narrative ability in high-functioning children with autism or Asperger's Syndrome," *J. Autism Dev. Disord.*, vol. 33, no. 3, pp. 239–251, Jun. 2003.
[20] L. Wang, P. Q. Pfordresher, C. Jiang, and F. Liu, "Individuals with autism spectrum disorder are impaired in absolute but not relative pitch and duration matching in speech and song imitation," *Autism Res.*, vol. 14, no. 11, pp. 2355–2372, 2021.
[21] P. Feliciano *et al.*, "SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research," *Neuron*, vol. 97, no. 3. Cell Press, pp. 488–493, 07-Feb-2018.
[22] E. Fombonne, L. Coppola, S. Mastel, and B. J. O'Roak, "Validation of Autism Diagnosis and Clinical Data in the SPARK Cohort," *J. Autism Dev. Disord.*, vol. 52, no. 8, pp. 3383–3398, Aug. 2022.
[23] M. Rutter, A. Bailey, and C. Lord, *The social communication questionnaire: Manual*. 2003.
[24] J. Bodfish, F. Symons, and M. Lewis, *The Repetitive Behavior Scale*. Western Carolina Center Research Reports, 1999.
[25] S. M. Redmond, "Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology," *Clin. Linguist. Phonetics*, vol. 19, no. 2, pp. 109–127, 2005.
[26] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.
[27] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9, pp. 341–345, 2001.
[28] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis," 2013.
[29] M. Babel and D. Bulatov, "The Role of Fundamental Frequency in Phonetic Accommodation," *Lang. Speech*, vol. 55, no. 2, pp. 231–248, Jun. 2012.
[30] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015.
[31] B. O. Ploog, "Stimulus Overselectivity Four Decades Later: A Review of the Literature and Its Implications for Current Research in Autism Spectrum Disorder," *J. Autism Dev. Disord. 2010 4011*, vol. 40, no. 11, pp. 1332–1349, Mar. 2010.
[32] B. O. Ploog, S. Banerjee, and P. J. Brooks, "Attention to prosody (intonation) and content in children with autism and in typical children using spoken sentences in a computer game," *Res. Autism Spectr. Disord.*, vol. 3, no. 3, pp. 743–758, Jul. 2009.
[33] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, Feb. 2012.
[34] M. Fusaro and P. L. Harris, "Children assess informant reliability using bystanders' non-verbal cues," *Dev. Sci.*, vol. 11, no. 5, pp. 771–777, 2008.
[35] E. Birmingham, D. Stanley, R. Nair, and R. Adolphs, "Implicit Social Biases in People With Autism," *Psychol. Sci.*, vol. 26, no. 11, pp. 1693–1705, Nov. 2015.

SPEECH METRICS THAT DIFFERENTIATE VARIANTS OF PRIMARY PROGRESSIVE APHASIA DEPEND ON LANGUAGE TYPOLOGY

Wendy Elvira-García[1], Victoria Marrero-Aguiar[2], Joan Borràs-Comes[1], Miguel Ángel Santos-Santos[3, 4]

[1]Universitat de Barcelona, [2]Universidad Nacional de Educación a Distancia (UNED), [3]Hospital de la Santa Creu i Sant Pau, [4]Biomedical Research Institute Sant Pau
wendyelvira@ub.edu, vmarrero@flog.uned.es, joanborras@ub.edu, MSantosS@santpau.cat

ABSTRACT

Primary Progressive Aphasia (PPA) variants can be distinguished using speech measures. In particular, a new one, syllabic-PVI, which computes syllable duration differences between unstressed and stressed adjacent syllables, can differentiate between the logopenic (lvPPA) and non-fluent (nfvPPA) variants.

The main goal of this study is to test if syllabic-PVI differentiates PPA variants in Spanish, a syllable-timed language like it does in English.

We analysed 5 repetitions of 5 multisyllabic words in a cohort of 37 patients (svPPA, lvPPA, nfvPPA) and 15 healthy controls. Data were annotated in Praat (manually revised) and metrics were computed in R by means of Generalized Linear Mixed Models.

Our results show that while syllable duration is a useful feature to distinguish lvPPA and nfvPPA in Spanish, syllabic PVI does not show significant differences between groups, which highlights the need of taking into account linguistic typology in linguistic batteries and clinical tests.

Keywords: clinical phonetics, Primary Progressive Aphasia speech, frontotemporal dementia, rhythm

1. INTRODUCTION

The Primary Progressive Aphasias (PPA) are a group of focal neurodegenerative syndromes primarily affecting language during the initial stages of the disease [1]. Recent studies have established the association between particular patterns of neuroanatomical damage and clinical presentation leading to the publication of the 2011 International consensus criteria for diagnosis of the three most common variants: non-fluent/agrammatic (nfvPPA), semantic (svPPA) and logopenic variants (lvPPA). Most cases of PPA at autopsy display frontotemporal lobar degeneration (FTLD)-type or Alzheimer's disease (AD) pathology. The classification of PPA cases into these clinical–anatomical phenotypes is important because these syndromes have predictive value for underlying molecular pathology (FTLD-tau in nfvPPA, FTLD-TDP in svPPA, and AD pathology in lvPPA), however these associations are not absolute, posing a significant impediment for the development of disease-modifying therapies based only on clinical presentation.

The diagnosis and differentiation of PPA is a challenge both for neurologists and speech pathologists. However, depending on the clinical syndrome, patients will tend to have semantic (svPPA), phonologic (lvPPA), or motor speech (apraxia of speech (AOS) +/- dysarthria) and grammatical (nfvPPA) problems. This makes AOS a crucial hint for differential diagnosis [2]–[4].

Research has shown that rhythmic and durational speech features are interesting to further distinguish subtypes of PPA [5] even automatically [2], [6], [7]. Specifically, previous research in English states that speech rate [6], [8]–[11], vowel duration [12] and Pairwise Variability Index (PVI) [13], [14] reflect the severity of AOS and therefore are useful speech metrics for characterizing PPA. For example, relative vowel duration in a polysyllable word repetition task can distinguish between lvPPA and nfvPPA [12].

When analyzing apraxia on isolated words, [15] proved that mean syllable duration is the most useful parameter if the stress position is not controlled, and vowel duration if the stress position is controlled. In a larger study, [5] tested average syllable duration and showed that it differs between lvPPA and nfvPPA.

More recent fine-grained research has focused on the duration of pre-stressed and stressed vowels' duration in apraxia. [16] found that apraxic patients' productions showed smaller differences in vowel duration between a first unstressed syllable and a following stressed one than controls which, perceptually, translates into an unclear lexical stressed syllable. Following this research a new metric that had proved valuable for patients suffering apraxia following a stroke, the syllable pairwise variability index (syl-PVI), [17] has been used for the differentiation between lvPPA and nfvPPA [5], [18] and has become a valuable metric in speech characterization of aphasic patients.

The Syl-PVI index is a metric that computes the difference between stressed and unstressed syllable pairs (weak-strong pattern), which linguistically highlights inaccuracies to express lexical stress.

As has been noted, Syl-PVI, highlights differences in duration between stressed and unstressed syllables, those differences are noticeably bigger in stress-timed languages than in syllable-timed languages but, to date, this metric has only been tested in stress-timed speaking populations. The goal of this study is to test

the syl-PVI index with a prototypical syllable-timed lanauge (Spanish) in order to test if this metric can be used in any language or its use should be limited to stress-timed languages.

2. METHODS

2.1. Materials

The linguistic battery used for this study is based on [19] which was adapted to Spanish in [20]. The protocol includes repetition of syllables, pseudowords, words, sentences, a reading of a phonetically balanced text (the Spanish version of "The North Wind and the Sun") the description of a cookie theft sheet and a picnic scene sheet. This study is centered on the word repetition task, which is known to yield less promising results than elicitation techniques. However, it is a faster and less demanding task for patients [5] [21] and has shown good results for characterizing AOS [15]. Of those materials, this study uses the repetition task of polysyllabic words (3 and 4 syllables), which includes words with the lexical stress in different positions (*patata* 'potato' [paˈt̪at̪a], *bicicleta 'bicycle'* [biθiˈkleta], *paquistaní* 'pakistani' [pakis̪t̪aˈni], *cúpula* 'Dome' [ˈkupula], *depósito* 'diposit' [deˈpos̪it̪o]).

2.2. Participants

52 participants were evaluated for this study (27 male, 25 female, mean age= 71.04 years). All of them were patients at the Hospital Clínic Sant Pau (Barcelona, Spain). All participants were tested with neurological and neuropsychological tests, and image diagnostics (computed tomography or MRI) and tested for biomarkers of neurodegeneration using the cerebrospinal fluid using the usual method at the Memory Unit of Sant Pau Hospital. The results of these tests were used to diagnose the patients by a neurologist of the Unit. Of the 52 participants, 14 were diagnosed with lvPPA, 16 with nfvPPA and 7 svPPA, 15 were controls. Since Barcelona is a social bilingual community the participants also completed a detailed bilingualism questionnaire.

2.3. Procedure

The speech sample was collected in one session (their first visit to the neurologist) using a Redenlab application and a microphone connected to an iPad. The sessions were carried out between 2019 and September 2022. Productions were semiautomatically annotated in Praat to segment by sound and by syllable. The by-sound annotation was done automatically and then manually corrected by 2 expert phoneticians. Once the by-sound annotation was done, a script was used to create the syllable boundaries [22].

Computation of duration and different metrics was done using R using the TextGrid as input. In order to compute the metrics only correct productions of the words were analysed. For example, if out of the 5 repetitions of *paquistaní*, a patient had produced [patitaˈni] in one instance, that instance has not been included in the database. This means that, while in general the analyses are based on 5 repetitions of the word, some patients may have less. In addition, as usual in clinical studies, some participants were unable to produce some words, the data of which PPA variants failed to repeat which words has been collected but its analysis is out of the scope of this paper.

The temporal parameters analysed were:

- Speech rate: computed as the number of sounds per second.
- Average syllable duration: mean duration of every syllable in each word.
- Syl-PVI: Syllabic pairwise-variability-index computes the normalized difference in duration between the stressed syllable of a word and the preceding one. Therefore, the word *cúpula* was excluded from the analysis. The final result is the mean of the 5 repetitions (when present) of each word [12].

The resulting metrics were normalized by word. The normalisation was performed using the orderNorm function from the bestNormalize package [23], which enforces a normal distribution. Then, the transformation was approximated using a cubic and logarithmic function. Finally, the obtained equation was applied to the original values. Then, data were analysed using Generalised Mixed Models [24] using the diagnosis group as a fixed factor, language and word as additional factors and the word by subject as a random slope.

3. RESULTS

The following section reports the results for syllable duration, speech rate and Syl-PVI.

Syllable duration behaves as expected, the control group had the smallest average syllable duration for the five polysyllabic words, followed by the lvPPA, nfvPPA and svPPA groups (Fig. 1). The durations are longer in nfvPPA compared to lvPPA (Cohen's d = 2.249, p=0.02). And the biggest effect is found for the word *depósito* (Cohen's d = 3.365, p<0.05). The rest of comparisons are not significant.

The results differ among words, svPPA shows the highest syllabic duration only for the word *depósito* 'diposit' a noun that forms a minimal pair with *deposito* "[I] deposit' (verb). *Paquistaní*, an oxytone

four-syllable word with a coda in the 3rd syllable shows longest syllables, especially in nfvPPA patients.

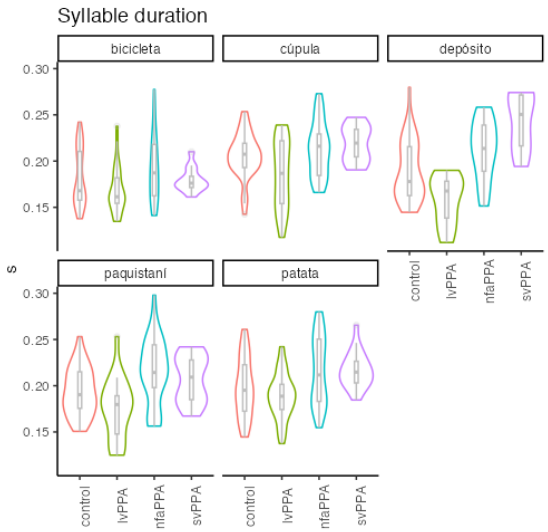


Figure 1: Average duration (s) box and violin plots for syllable duration for each diagnostic group by word.

Speech rate results (Fig. 2) follow the same line, which is expected given that all participants had repeated the same words (i.e. have produced the same number of sounds), the highest effect is found between lvPPA and nfvPPA but it does not reach significance (Cohen's d = 1.434, p=0.07). As noted by a reviewer, speech rate and average syllable duration should (and do) exhibit a strong inverse correlation. However, in the analysis conducted, we observed that only the average syllable duration showed statistical significance. This result can be attributed to the method used to calculate speech rate, the number of sounds produced per second. The duration of individual sounds can vary widely, as exemplified in the contrast between the length of a final vowel and a spirant approximant. In contrast, the duration of syllables exhibits relatively lower variation, contributing to the significant results observed for average syllable duration.

Fig. 3 illustrates the syl-PVI by word. According to the formula, lower (negative) values imply more syllable reduction at the pre-stressed position, whereas a value of 0 indicates the equal duration of both syllables. *Syllable reduction at the pre-stressed position is a phenomenon of "normal" speech in English which is absent or reduced in patients with AOS.* Therefore the lowest values should be expected for the control group, whereas apraxic patients should show values closer to 0. Our results show that nfvPPA tends to have the value closest to 0, but the effect size does not reach significance for any word.

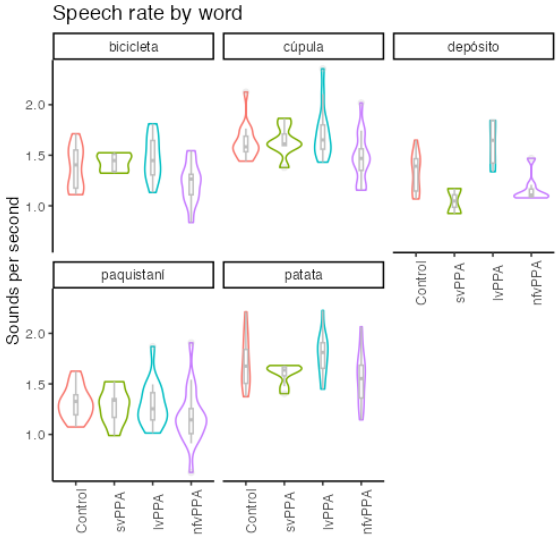


Figure 2: Speech rate box and violin plots for syllable duration for each diagnostic group by word.

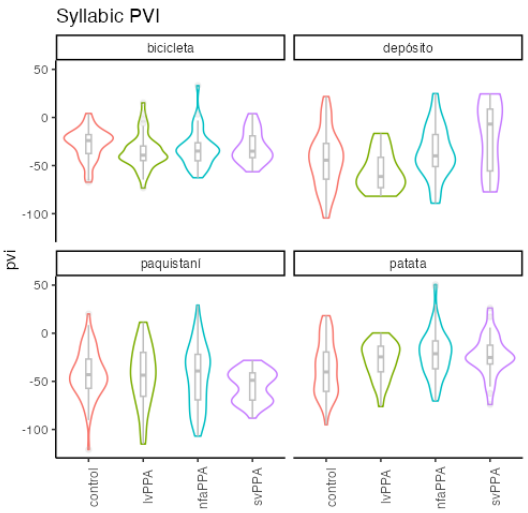


Figure 3: Average syl-PVI box and violin plots for each diagnostic group by word.

3. DISCUSSION

The results of the previous section show that in a polysyllabic repetition task, syllable duration is a relevant parameter while the Syl-PVI has no effect. The result is striking because previous studies have found that a diminished Syl-PVI is more specific to apraxia of speech whereas diminished speech rate can be found in more PPA variants. [5] found in their study that Syl-PVI was the most determinant factor for AOS even with fewer participants than in the present study: they included 10 patients with nfvPPA, 10 lvPPA and 5svPPA.

However, from a typological point of view, it is not surprising that there are no differences in syllabic-PVI in Spanish. While English is one of the most prototypical stressed-timed languages with big

durational differences between stressed and unstressed differences, Spanish is one of the most prototypical syllable-timed languages (sometimes called *pure*) in which duration and stress are always aligned [25].

Therefore, English apraxic patients who struggle with the expression of stress will have durational inaccuracies, failing to shorten unstressed syllables. Spanish syllables do not undergo this phonological process, each syllable being approximately equal. Moreover, we know that syllable-time rhythm production is easier as children produce it earlier [26] and it is less difficult to learn for foreign students than stress-timed timing (for example, we are seeing a change towards a syllable-timed rhythm in World Englishes [27]). In that sense, we could claim that apraxic patients make the "default" rhythmic pattern but that change is more noticeable in stress-timed languages.

However, if we take into account syllable duration instead of the syl-PVI we can see differences between the diagnosis groups. Those differences could be due to differences in timing or rhythm but also due to differences in speech rate given that all participants produced the same items. This is expected given that most research dealing with lvPPA and nfvPPA finds differences in speech rate notwithstanding the task (see section 1).

From a clinical point of view, the explanation why Syl-PVI is not significant in our results could lie in differences among the cohort and specifically their apraxia level. For this study, we have analysed the patient's first visit to the neurologist, which means that most patients are in an incipient stage of the disease and hence they could show mild apraxia in comparison with other studies. In order to minimize this and given that syl-PVI is primarily an indicator of the presence of apraxia, a follow-up study should be conducted correlating the level of AOS in these patients with their alteration of Syl-PVI. This way we could discard that our negative results are due to an incipient stage of the disease in our patients.

4. CONCLUSION

Our results show that syllable duration is useful for distinguishing the logopenic and non-fluent variants of PPA in Spanish, but syllabic PVI does not show significant differences between groups.

This is probably due to a typological difference between English (a prototypical stress-timed language) and Spanish (a prototypical syllable-timed language) given that the syllabic-PVI computes differences between the duration of unstressed and stressed syllables and those differences are smaller in syllable-timed languages. Our results highlight the necessity of taking into account language particularities when using clinical tests and metrics. And, in particular, this case shows the inadequacy of using syllabic-PVI in prototypical syllable-timed languages such as Spanish.

6. REFERENCES

[1] M. L. Gorno-Tempini *et al.*, "Classification of primary progressive aphasia and its variants," *Neurology*, vol. 76, no. 11, 2011, doi: 10.1212/WNL.0b013e31821103e6.

[2] C. Themistocleous, B. Ficek, K. Webster, D. B. den Ouden, A. E. Hillis, and K. Tsapkini, "Automatic subtyping of individuals with primary progressive aphasia," *Journal of Alzheimer's Disease*, vol. 79, no. 3, 2021, doi: 10.3233/JAD-201101.

[3] M. Á. Santos Santos and Gorno Tempini Maria L, "Primary Progressive Aphasia," *Brain Mapping: An Encyclopedic Reference.* . Elsevier, pp. 653–663, 2015.

[4] K. Croot, K. Ballard, C. E. Leyton, and J. R. Hodges, "Apraxia of speech and phonological errors in the diagnosis of nonfluent/agrammatic and logopenic variants of primary progressive aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 5, 2012, doi: 10.1044/1092-4388(2012/11-0323).

[5] K. L. Haley *et al.*, "Speech metrics and samples that differentiate between nonfluent/ agrammatic and logopenic variants of primary progressive aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 3, 2021, doi: 10.1044/2020_JSLHR-20-00445.

[6] K. C. Fraser *et al.*, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, no. 1, 2014, doi: 10.1016/j.cortex.2012.12.006.

[7] C. Themistocleous *et al.*, "Acoustic markers of PPA variants using machine learning," *Front Hum Neurosci*, vol. 12, 2018, doi: 10.3389/conf.fnhum.2018.228.00092.

[8] M. Grossman *et al.*, "Disruption of large-scale neural networks in non-fluent/agrammatic variant primary progressive aphasia associated with frontotemporal degeneration pathology," *Brain Lang*, vol. 127, no. 2, 2013, doi: 10.1016/j.bandl.2012.10.005.

[9] C. Cordella, B. C. Dickerson, M. Quimby, Y. Yunusova, and J. R. Green, "Slowed articulation rate is a sensitive diagnostic marker for identifying non-fluent primary progressive aphasia," *Aphasiology*, vol. 31, no. 2, 2017, doi: 10.1080/02687038.2016.1191054.

[10] D. Gunawardena, S. Ash, C. McMillan, B. Avants, J. Gee, and M. Grossman, "Why are patients with progressive nonfluent aphasia nonfluent?," *Neurology*, vol. 75, no. 7, 2010, doi: 10.1212/WNL.0b013e3181ed9c7d.

[11] J. D. Rohrer, M. N. Rossor, and J. D. Warren, "Syndromes of nonfluent primary progressive aphasia: A clinical and neurolinguistic analysis," *Neurology*, vol. 75, no. 7, 2010, doi: 10.1212/WNL.0b013e3181ed9c6b.

[12] K. J. Ballard, S. Savage, C. E. Leyton, A. P. Vogel, M. Hornberger, and J. R. Hodges, "Logopenic and nonfluent variants of primary progressive aphasia are differentiated by acoustic measures of speech production," *PLoS One*, vol. 9, no. 2, 2014, doi: 10.1371/journal.pone.0089864.

[13] J. R. Duffy *et al.*, "Temporal acoustic measures distinguish primary progressive apraxia of speech from primary progressive aphasia," *Brain Lang*, vol. 168, 2017, doi: 10.1016/j.bandl.2017.01.012.

[14] J. R. Duffy, E. A. Strand, H. Clark, M. Machulda, J. L. Whitwell, and K. A. Josephs, "Primary progressive apraxia of speech: Clinical features and acoustic and neurologic correlates," *Am J Speech Lang Pathol*, vol. 24, no. 2, 2015, doi: 10.1044/2015_AJSLP-14-0174.

[15] K. L. Haley and A. Jacks, "Word-level prosodic measures and the differential diagnosis of apraxia of speech," *Clin Linguist Phon*, vol. 33, no. 5, 2019, doi: 10.1080/02699206.2018.1550813.

[16] D. I. Scholl, P. J. McCabe, R. Heard, and K. J. Ballard, "Segmental and prosodic variability on repeated polysyllabic word production in acquired apraxia of speech plus aphasia," *Aphasiology*, vol. 32, no. 5, 2018, doi: 10.1080/02687038.2017.1381876.

[17] A. Basilakos *et al.*, "A multivariate analytic approach to the differential diagnosis of apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, 2017, doi: 10.1044/2017_JSLHR-S-16-0443.

[18] R. Landin-Romero *et al.*, "Brain changes underlying progression of speech motor programming impairment," *Brain Commun*, vol. 3, no. 3, 2021, doi: 10.1093/braincomms/fcab205.

[19] A. P. Vogel *et al.*, "Motor speech signature of behavioral variant frontotemporal dementia: Refining the phenotype," *Neurology*, vol. 89, no. 8, 2017, doi: 10.1212/WNL.0000000000004248.

[20] W. Elvira-García and V. M. Aguiar, "El Ritmo y la Tasa de Habla Como Ayuda Diagnóstica en Síndromes Neurodegenerativos de Los Lóbulos Fronto-Temporales," *Loquens*, vol. 7, no. 1, 2021, doi: 10.3989/LOQUENS.2020.068.

[21] C. Potagas *et al.*, "Silent Pauses and Speech Indices as Biomarkers for Primary Progressive Aphasia," *Medicina (Lithuania)*, vol. 58, no. 10, 2022, doi: 10.3390/medicina58101352.

[22] J. M. Garrido Almiñana, "SegProso: A Praat-Based Tool for the Automatic Detection a nd Annotation of Prosodic Boundaries," in *Proceedings of TRASP (Interspeech satellite)*, B. Bigi and D. Hirst, Eds., Aix-en-Provence, 2013, pp. 74–77.

[23] R. A. Peterson and J. E. Cavanaugh, "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era," *J Appl Stat*, vol. 47, no. 13–15, 2020, doi: 10.1080/02664763.2019.1630372.

[24] B. Bolker, "Getting started with the glmmTMB package," *Cran.R-Project Vignette*, no. 2009, 2019.

[25] M. Nespor, M. Shukla, and J. Mehler, " Stress-Timed vs . Syllable-Timed Languages ," in *The Blackwell Companion to Phonology*, 2011. doi: 10.1002/9781444335262.wbctp0048.

[26] M. Canault *et al.*, "Syllable duration changes during babbling: A longitudinal study of French infant productions," *J Child Lang*, vol. 47, no. 6, 2020, doi: 10.1017/S030500092000015X.

[27] T. Nishihara and J. van de Weijer, "On Syllable-Timed Rhythm and Stress-Timed Rhythm in World Englishes: Revisited," *Miyagi kyōiku daigaku kiyō [Bulletin of Miyagi University of Education]*, vol. 46. 2011.

A BIMODAL APPROACH TO STUDY THE EFFECTS OF RHYTHMIC PRIMING

Contreras Roa, Leonardo[1], Mairano, Paolo[2], Moreau, Caroline[3], Basirat, Anahita[4]

[1]CORPUS (EA 4295), Université de Picardie Jules Verne
[2]STL (UMR 8163), Université de Lille, France
[3]LilNCog (U1171-2), Université de Lille, CHU Lille, Inserm, France
[4]SCALab (UMR 9193), Université de Lille, CNRS, CHU Lille, France
[1]l.contreras.roa@u-picardie.fr, [2]paolo.mairano@univ-lille.fr, [3]caroline.moreau@chru-lille.fr,
[4]anahita.basirat@univ-lille.fr

ABSTRACT

We present a speech rhythm analysis workflow based on two different but complementary methodologies: one at a segmental/syllabic level, and another one at a suprasegmental/prosodic level. This bimodal methodology is devised to study the effects of regular and irregular rhythmic priming — i.e. rhythmic stimuli presented to speakers before elocution — on the speech rhythm of people with Parkinson's disease. Preliminary results show that both segmental and suprasegmental rhythm metrics vary in different priming conditions. We propose that this methodology can be used to have a better grasp of speech rhythm by peering into the potential interaction or correlation between vocalic-consonantal interval distributions and pitch movement variations. Future implementations of the workflow are possible in other speech analysis fields if models are shown to be effective.

Keywords: speech rhythm, speech production, rhythmic priming, prosody, Parkinson's disease.

1. INTRODUCTION

1.1. Studies on speech rhythm

The concept of speech rhythm is a complex one, and despite the existence of a conspicuous amount of work in this domain, there is still no real comprehensive theoretical framework to fully account for speech rhythm [1]. Acoustic measures based on vocalic and consonantal durations have been proposed, which are supposed to reflect phonological characteristics thought to be related to speech rhythm (e.g., the presence of vowel reduction or of complex consonant clusters in a given language), notably ΔC, ΔV, %V [2] and PVI (Pairwise Variability Index, [3]). Such formulas have then been variously improved to account for the effects of speech rate (VarcoV, VarcoC, cf. [4]) or applied to different segments (V-to-V, cf. [5]) or to prosodic parameters other than duration such as intensity [6] and F0 [7]. While some researchers remain sceptical of these measures [8] or more generally of language rhythmicity [9], they have been used in a large number of studies characterizing the rhythm of different languages, as well as to study the effects of the rhythmic properties of the L1 on L2 pronunciation acquisition [10] and to study the characteristics of pathological speech in the context of dysarthria [11, 12].

While rhythm metrics give an account of speech rhythm that is based on segmental (or syllabic) durations, other approaches to speech rhythm have focused on suprasegmental aspects, notably the metrical structure of utterances and accents or prominence. In this context, the literature distinguishes hierarchical linguistic units responsible for creating speech rhythm: generally, the syllable, the accentual phrase (henceforth AP, which in French is marked by a primary accent on the right), the intonational phrase (henceforth IP, cf. [13]) and, according to certain approaches, an intermediate phrase (henceforth ip, cf. [14]). The literature about rhythm metrics (that we will call segmental / syllabic rhythm) and about the metrical structure and prominence (that we will call suprasegmental rhythm) belong to different frameworks, and various authors have advocated the need to develop a combined multi-level or multi-parametric model of speech rhythm [15, 12]. In this contribution, we propose a pipeline based on this type of bimodal approach to study dysprosody in Parkinson's disease (PD).

1.2. The case of dysprosody in Parkinson's disease: is rhythmic priming beneficial?

PD is the second most common neurodegenerative disease. While it is mainly known for its motor

effects on the upper/lower limbs, such as tremor and walking impairment, various studies have pointed out early impairments affecting spoken communication in individuals with Parkinson's disease (PwPD). Prosody is one of the aspects that is most often reported to be affected in PD. This is referred to as dysprosody and consists of alterations at various levels such as monotony of pitch [16], reduced intonation [17], and issues in prosodic phrasing [18] and in the temporal organization of speech [19].

There is a growing body of evidence suggesting that exposure to rhythmic auditory patterns may modulate subsequent speech and language processing, at both the reception (e.g. [20]) and production (e.g. [21]) levels. This effect is known as rhythmic priming effect and thought to be based on shared cognitive representations and mechanisms for speech and music. Importantly, the transfer from musical rhythm to language seems promising in clinical settings (for a review, see [22]). Given this emerging literature on rhythmic priming effects, it seems relevant to examine whether exposure to auditory rhythm impacts speech production at prosodic levels, and improve prosody production in PwPD suffering from dysprosody.

1.3. Aim of this study

In this contribution, we present a methodological protocol for studying the effects of rhythmic priming on speech. We applied it to investigate whether non-linguistic rhythmic priming has any effect on French-speaking individuals with Parkinson's disease (PwPD), since previous research found that rhythmic priming modulates speech production (see section 1.2). The goal of this contribution is to document the data processing protocol and to reflect upon its advantages and potential uses in other areas of speech analysis. In the following sections, we give full details of the procedure that we followed and the set of Praat and R scripts that we have created as part of our pipeline are provided on GitHub at the link specified below as part of this contribution. We also present preliminary results for 12 speakers to illustrate the type of results that can be obtained.

2. A BIMODAL APPROACH TO STUDY THE EFFECTS OF RHYTHMIC PRIMING

2.1. Task

For the purpose of the current study, our pipeline was designed to analyze read speech under different priming conditions (see below). However, this pipeline can also be used to analyze speech rhythm obtained using other tasks and under other conditions.

Recordings were obtained using a sentence read-aloud task under three conditions:

1. after listening to a regular 2-second percussive rhythmic prime coinciding with the accentual structure of each sentence to read,
2. after listening to an irregular (arrhythmia) 2-second percussive prime incongruous with the accentual structure of the sentences to read,
3. after listening to two seconds of silence (no prime).

Sentences had a fixed accentual structure: four groupings of three syllables (i.e. accentual phrases) stress on their respective final syllables: $xxX_1xxX_2xxX_3xxX_4$ (e.g. *Le poisson$_1$ se nourrit$_2$ de nombreux$_3$ végétaux$_4$*). We aimed at testing (1) whether rhythmic condition impacts production in terms of speech rhythm, and (2) whether regular rhythmic priming favours a more natural performance in terms of speech rhythm among PwPD compared to irregular rhythmic priming.

2.2. The analysis protocol

The analysis is devised to combine rhythm metrics traditionally used in phonetic studies and increasingly used in the study of patients with dysarthria, along with automatic gradual prominence detection methods -as opposed to strictly categorical detection methods- based on acoustic prosodic parameters (see section 1). In this sense, our approach is bimodal, as it considers two sets of parameters that are thought to contribute to speech rhythm: traditional rhythm metrics are taken to reflect a syllabic or intersegmental level of analysis, while acoustic parameters of prominence are taken as a rhythmic measure at the accentual level. Rhythm metrics and acoustic parameters of prominence stem from different theoretical frameworks (the former from speech rhythm typology, the latter from studies on the metrical structure of speech), and - although a multilevel approach to speech rhythm has been often advocated - they have seldom been combined within the same study. We consider this to be one of the originalities of our contribution. The data analysis workflow is illustrated in Figure 1 and encompasses 4 main stages: (1) automatic speech annotation and segmentation, (2) manual check, (3) automatic prominence detection, (4) rhythm metrics calculation.

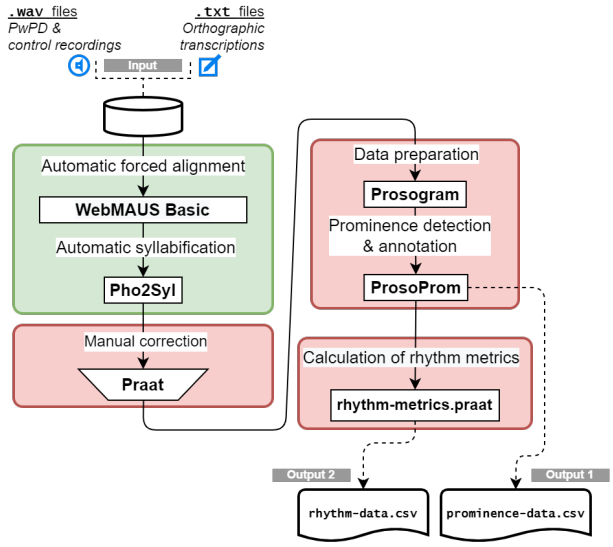


Figure 1: Data processing workflow.

2.2.1. *Steps 1 and 2: speech annotation and segmentation*

This step is carried out with two online services provided by the WebMAUS platform (in green in Figure 1): WebMAUS Basic [23] and Pho2Syll [24]. The audio recordings (.wav) and orthographic transcriptions (.txt) of target sentences are fed into WebMAUS Basic for semi-automatic speech annotation, and subsequently into Pho2Syll for automatic syllabification. The resulting annotated TextGrid files are then open in Praat (in red in Figure 1), verified and manually corrected if needed to address potential alignment errors.

2.2.2. *Step 3: automatic prominence detection*

Audio and annotation data are then fed to the Prosogram [25] and ProsoProm tools, part of the ProsoBox3 package [26]. These tools are a Praat plugin which allows to compute prominence as a gradual feature by analysing silent pauses along with the relative duration and pitch averages within an adjustable window of analysis: in this study, we set this to 2 syllables before and 1 syllable after. These relative measurements can also be carried out cumulatively to consider the relative weight of each of these acoustic cues and calculate a global score for each syllable. The duration threshold of silent pauses can be modified as well, in order to better adapt it to the study of pathological speech. We carried out our analysis with the default threshold of 250 ms.

ProsoProm's output is a TextGrid file with four extra tiers: one tiers marks syllables that are considered categorically prominent following the given threshold, and three tiers specify the cumulative contribution of relative duration, relative pitch and intra-syllable pitch movement to the gradual prominence of every syllable, whether they are classified as categorically prominent or not.

2.2.3. *Step 4: computation of rhythm metrics*

The output files from the previous step are then fed into a Praat script which computes the following rhythm metrics:

- %V: the percentage over which speech is vocalic for each speaker's recording.
- ΔC, ΔV: the standard deviation of the duration of consonant and vowel intervals.
- VarcoC, VarcoV: the variation coefficient for consonants and vowel intervals.
- rPVI-C, nPVI-V: raw and normalized pairwise variability index for the durations of consonantal and vocalic intervals.
- VarcoVC, rPVI-VC, nPVI-VC: variation coefficient and pairwise variability index for vocalic + consonantal intervals (i.e., measures reflecting the duration variability of pseudo-syllables).

The output datasets from steps 3 and 4 are saved in tabular format (.csv). Prominence-data.csv gives prominence data for every syllable, by speaker and by priming condition. Rhythm-data.csv gives average rhythm metrics by speaker and by priming condition. The ad-hoc Praat and R scripts we have developed are made available on GitHub at the following address: https://github.com/lcontrerasroa/EntrainPark. They can be freely used to replicate our protocol.

2.3. Results

So far, our analysis pipeline has been applied twice: once in a pilot case study [27], and once to the recordings of a larger-scale sample of participants: 5 PwPD and 7 controls who performed the task described in Section 2.1. Preliminary prominence results obtained through ProsoProm are shown in Fig. 2 which illustrates F0 movement absolute values at the four expected stressed syllables (1, 2, 3 and 4) of the read-aloud sentences.

The expected underlying prosodic structure of sentences can be seen in the productions of control subjects, who display broader pitch movements at positions 2 and 4 which correspond to prosodic boundaries. The same predicted rises are either less prominent or not observed at all in utterances produced by PwPD, particularly in

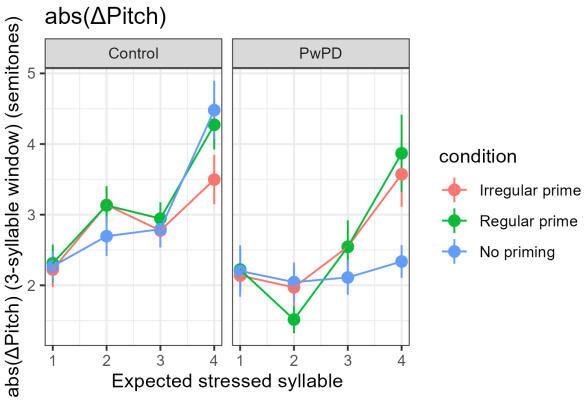


Figure 2: Absolute values of pitch movement (in semitones) at expected syllables, by group and priming condition.

absence of rhythmic priming. Two rhythm metric results (consonantal and vocalic pairwise variability indexes) are illustrated in Fig. 3.

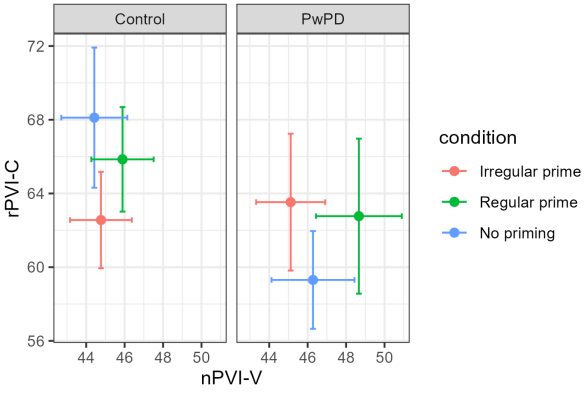


Figure 3: rPVI-C versus nPVI-V (means and standard error), by group and priming condition.

Consonantal metrics in particular (y-axis) react to priming condition, though in different directions: control subjects display higher pairwise variability scores without priming, whereas these values are lower for PwPD. Vocalic variability (x-axis) is considerably lower across priming conditions for control participants, whereas PwPD show more variation, especially between the irregular and regular priming conditions.

3. DISCUSSION

We set up a protocol to semi-automatically analyse the results of rhythm priming experiment. We ran it on preliminary data of an ongoing larger project about the effects of rhythm priming on PwPD, with promising results. The protocol can easily be replicated following the workflow instructions given in this article, and using the scripts available on the GitHub space associated to our work. The protocol and the script are by no means restricted to PwPD speech; they can be applied to any study investigating rhythm priming on any population. We believe that the main strength and originality of this protocol consists in trying to give a multimodal approach of speech rhythm, combining two levels of analysis: a first level is given by rhythm metrics, which reflect intersegmental/syllabic rhythm; a second analysis operates at the suprasegmental level, giving acoustic measures that are supposed to reflect prosodic prominence. Since both levels have been proven to be relevant for the perception of speech rhythm, such combined results can be used to create more robust statistical models to explore the interaction, and potential correlation between segmental/syllabic rhythmic and prosodic prominence. Furthermore, the combination of these two types of metrics and their contextualized analysis following the prosodic hierarchies of French allows us to bridge the gap and work at the interface of the phonetics and the phonology of speech rhythm.

The results of our preliminary study on PwPD speech are promising and will hopefully be confirmed and expanded on data that we are currently collecting. Thus far, rhythmic priming seems to have an effect on both intersegmental and prosodic cues, which display different behaviors for control subjects and PwPD. In the future, we may develop an online service with graphic user interface centralising all the steps of the analysis. This would certainly make it more accessible, and may be made available for other purposes. One possibility that we have in mind is to expand the protocol to study the effects of rhythm priming on L2/L3 acquisition, potentially also from a typological perspective, for example by comparing learners from L1s belonging to different rhythmic typologies. For now, we content ourselves with presenting our workflow combining a rhythmic analysis on two levels, and of making it available to the scientific community.

4. ACKNOWLEDGEMENT

This work was supported by French National Research Agency grant ANR-21-CE28-0011 (EntrainPark project).

5. REFERENCES

[1] J. Volín *et al.*, "Appeal and disrepute of the so-called global rhythm metrics," *Acta Universitatis Carolinae Philologica*, no. 3, pp. 79–94, 2017.

[2] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.
[3] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 1982, pp. 515–546, 2002.
[4] V. Dellwo, P. Wagner, M. Solé, D. Recasens, and J. Romero, "Relations between language rhythm and speech rate," *Proc. of the 15th ICPhS*, pp. 471–474, 2003.
[5] M. Pettorino, M. Maffia, E. Pellegrino, M. Vitale, and A. De Meo, "Vtov: A perceptual cue for rhythm identification," in *Proceedings of the Prosody-Discourse Interface Conference*, 2013, pp. 101–106.
[6] L. He, "Development of speech rhythm in first language: The role of syllable intensity variability," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL463–EL467, 2018.
[7] L. Polyanskaya, M. G. Busà, and M. Ordin, "Capturing cross-linguistic differences in macro-rhythm: the case of italian and english," *Language and Speech*, vol. 63, no. 2, pp. 242–263, 2020.
[8] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, no. 3, pp. 351–373, 2012.
[9] F. Nolan and H.-S. Jeon, "Speech rhythm: a metaphor?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1658, p. 20130396, 2014.
[10] A. Tortel and D. Hirst, "Rhythm metrics and the production of english l1/l2," in *Speech Prosody 2010-Fifth International Conference*, 2010.
[11] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, 2009.
[12] A. Lowit, A. Marchetti, S. Corson, and A. Kuschmann, "Rhythmic performance in hypokinetic dysarthria: Relationship between reading, spontaneous speech and diadochokinetic tasks," *Journal of Communication Disorders*, vol. 72, p. 26, 2018.
[13] A. Di Cristo, *La prosodie de la parole*. De Boeck Superieur, 2013.
[14] A. Michelas and M. d'Imperio, "Durational cues and prosodic phrasing in french: evidence for the intermediate phrase," in *Speech Prosody*, 2010, p. 4.
[15] P. M. Bertinetto and C. Bertini, "On modeling the rhythm of natural languages," in *Proceedings of the Fourth International Conference on Speech Prosody*, 2008, pp. 427–430.
[16] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in parkinson's disease over time - a longitudinal study," *Movement disorders: official journal of the Movement Disorder Society*, vol. 24, no. 5, pp. 716–722, 2009.
[17] A. Basirat, C. Patin, and C. Moreau, "Relationship between Perception and Production of Intonation of French in Parkinson's Disease," in *Proceedings of the 9th International Conference on Speech Prosody 2018*, 2018, pp. 809–813.
[18] S. Frota, M. Cruz, R. Cardoso, I. Guimarães, J. J. Ferreira, S. Pinto, and M. Vigário, "(Dys)Prosody in Parkinson's Disease: Effects of Medication and Disease Duration on Intonation and Prosodic Phrasing," *Brain Sciences*, vol. 11, no. 8, p. 1100, Aug. 2021.
[19] A. Lowit, A. Marchetti, S. Corson, and A. Kuschmann, "Rhythmic performance in hypokinetic dysarthria: Relationship between reading, spontaneous speech and diadochokinetic tasks," *Journal of Communication Disorders*, vol. 72, pp. 26–39, 2018.
[20] L. Przybylski, N. Bedoin, S. Krifi-Papoz, V. Herbillon, D. Roch, L. Léculier, S. A. Kotz, and B. Tillmann, "Rhythmic auditory stimulation influences syntactic processing in children with developmental language disorders." *Neuropsychology*, vol. 27, no. 1, pp. 121–131, 2013.
[21] N. Cason and D. Schön, "Rhythmic priming enhances the phonological processing of speech," *Neuropsychologia*, vol. 50, no. 11, pp. 2652–2658, Sep. 2012.
[22] A. Fiveash, N. Bedoin, R. L. Gordon, and B. Tillmann, "Processing rhythm in speech and music: Shared mechanisms and implications for developmental speech and language disorders," *Neuropsychology*, Aug. 2021.
[23] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.
[24] U. D. Reichel, "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," *Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation*, vol. 71, pp. 42–49, 2014.
[25] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Speech prosody 2004, international conference*, 2004.
[26] J.-P. Goldman and A. C. Simon-Hustinx, "Prosobox, a praat plugin for analysing prosody," in *Proceedings of the 10th International Conference on Speech Prosody*, 2020, pp. 1009–1013.
[27] L. Contreras Roa, P. Mairano, C. Moreau, and A. Basirat, "Impact de l'amorçage rythmique sur la production de la parole chez des personnes atteintes de la maladie de Parkinson : Étude pilote," in *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, 2022, pp. 578–587.

PHONETIC-PHONOLOGICAL FACTORS IN STUTTERING IN GERMAN

Annett B. Jorschick, Dinah Rachko, Joana Cholin

Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
annett.jorschick@uni-bielefeld.de, dinah.rachko@gmail.com, joana.cholin@uni-bielefeld.de

ABSTRACT

Relating stuttering to phonetic-phonological complexity has a long tradition in clinical and psycholinguistic research. Some findings point toward a relationship between stuttering and phonetic-articulatory effort, while others emphasize lexical-representational aspects. In our study, we focus on phenomena specifically arising in the transition from phonological to phonetic-articulatory encoding. Here, abstract phonological segments are assembled to form motor plans for articulation by taking various factors into account.

We investigated stuttering in a German word-reading-task using words and nonsense-words starting with a [CV]-structure, manipulating *manner* and *place* of articulation of word-initial consonants and their distance to adjacent vowels (*transition*), *vowel length,* controlling for *word length*, *syllable frequency* and *biphone frequency*. Results show *word length* as a robust predictor of stuttering, while the factors *transition*, *manner* and *vowel length* (and their interaction) also played a role. We argue that only such systematic approaches allow insights into the complex interplay between encoding levels in stuttering.

Keywords: stuttering; phonology-phonetics interface; phonological-phonetic complexity

1. INTRODUCTION

Persistent developmental stuttering is a multifactorial speech disorder causing recurring speech disrupting symptoms such as prolongations and repetitions of speech sounds and silent blocks that severely hamper fluent speech production [1, 2]. The investigation of linguistic factors relevant for the occurrence of these symptoms aims to pinpoint deviating speech behavior to individual encoding levels in order to decipher the underlying pathomechanism of stuttering [3, 4]. Most research in this area is focused on aspects of either lexical representations and their corresponding syntactic and phonological encoding processes or phonetic-articulatory procedures. Current data do not allow to unequivocally associate the recurrent dysfluencies with deficient representations or processes.

In our study, we focus on phenomena arising in the conversion from phonological to phonetic-articulatory encoding as this interface between linguistic planning and motor programming might be particularly susceptible to stuttering. During phonological encoding, the individual phonemes of words are successively grouped together to form abstract phonological syllables, following universal and language-specific syllabification rules [5]. What constitutes an optimal syllable in a given context, crucially depends on the phonological-phonetic environment and the availability of subsequent motor programming.

Several authors [6, 7, 8] have attributed stuttering to the inability to transition between adjacent segments. According to [9], this is especially the case at word onsets and accelerated within stressed syllables. Therefore, we aim to investigate whether there are specific constellations between adjacent segments prompting stuttering. Especially relevant is the question, whether these combinations are problematic per se or whether they are problematic only if linked to specific lexical representations (i.e., words but not nonsense-words). Importantly, the many interrelated factors that are at play in the phonology-phonetics interface need to be disentangled. Previous studies mainly tested simple contrasts in corpus studies, often ignoring interdependencies between these factors and their interactions.

Thus, in our material, we systematically manipulated manner and place of articulation (hereinafter manner and place) of consonants in word onsets. The transition between initial consonants and adjacent vowels was also systematically manipulated to test whether a larger distance between consonants' place and vowels' backness accelerates stutter occurrences. Additionally, we varied vowel length across words and nonsense-words to investigate the influence of different vowel quantities on stuttering rates, which has been shown as a relevant predictor in corpus studies with children who stutter [10]. German offers the opportunity to contrast vowel length in different vowel positions. Other factors that were found to influence stuttering rates in previous studies (for an overview and discussion see [3, 4]) such as stress, word frequency, number of syllables, and word function were controlled across conditions (see below).

2. METHOD

The ethics committee of Bielefeld University approved the study.

2.1. Materials and Design

We selected 144 common nouns from CELEX [11] in which the initial consonants and their adjacent vowels were carefully chosen taking into account the following factors: *place* (labial vs. velar/glottal) and *manner* of the initial consonant (plosive, fricative, nasal, glottal stop), vowel backness (front: /ʏ, ɪ, ɛ, yː, iː, ɛː/, central: /a, aː/, back: /ɔ, ʊ, oː, uː/) and *vowel length* (short vs. long). *Transition:* three transitional distances (small, medium, large) were defined as the distance in position between the word initial consonant and its adjacent vowel: A small distance is found between a labial consonant and a frontal vowel or a velar/glottal consonant and a back vowel. Medium transitions were defined as the distance of the labial or velar/dorsal consonants to a central vowel. Large transitions reflect the distance between a labial consonant and a back vowel or a dorsal/glottal consonant and a frontal vowel. We controlled stress by choosing words with stressed first syllables, while no word contained consonant clusters or diphthongs in any position. The words were matched between the manipulated conditions according to their number of syllables (one or two syllables). Because there are no German words starting with velar nasal /ŋ/ and words orthographically starting with vowels are realized with glottal stops, manipulations resulted in an incomplete factorial design with 2+1+1 (manner) x 2 (place) x 3 (transition) x 2 (vowel length) = 36 possible CV-combinations (see Table 1 for examples of potential initial CV-structures). We selected four mono- or disyllabic words for these 36 CV-combinations resulting in 144 words and constructed 144 corresponding nonsense-words by exchanging one phoneme, keeping the initial CV-structure intact and also obeying constraints of German phonotactics and orthography (i.e., Gummi [gum] – Gubbi), resulting in 288 targets. In addition to these manipulations, we retrieved syllable frequency of the first syllable and bigram frequency of the initial CV-structure from CELEX as potential covariates.

Experimental lists alternated between word and nonsense-word blocks with a break between blocks. The order of items within word blocks was pseudo-randomized to ensure that consecutive items were neither phonologically nor semantically related. The matching block of nonsense-words was ordered accordingly. The order of experimental blocks was counterbalanced across experimental lists.

	labial			dorsal		
	small	medium	large	small	medium	large
vowel				ʔɔ,ʔoː	ʔa, ʔaː	ʔɪ, ʔiː
nasal	mɪ, miː	ma, maː	mɔ, moː			
plosive	pɪ, piː	pa, paː	pɔ, poː	kɔ, koː	ka, kaː	kɪ, kiː
fricative	fɪ, fiː	fa, faː	fɔ, foː	hɔ, hoː	ha, haː	hɪ, hiː

Table 1: Overview of the word-initial CV-structure manipulations: *place* and *manner* of C, *transition* to V and *vowel Length*

2.2. Method and Procedure

In a pretesting session, participants reported their stuttering history and took part in a video-recorded diagnostic session in which the stuttering severity instrument (SSI-4 [12]) was administered. This test uses samples of spontaneous and read speech and an evaluation of speech-accompanying physical concomitants to calculate standardized severity scores ranging from 10 (mild stutter severity) to a maximum score of 56. The SSI-4 was evaluated afterward based on the video-recordings.

During the reading experiment, participants sat in front of a laptop, wearing a headset. They were videotaped from two different perspectives: a close-up of the face and from a further distance to also capture speech-accompanying hand gestures. The experimental procedure was controlled by the DMDX software [13]. After a short introduction presented on the screen, participants read the word and nonsense-word lists following the order described above. Participants used the space key to prompt the next word on the screen. The experiment lasted about 20 minutes resulting in a total session length of 45 to 60 minutes.

2.3. Participants

24 native German adults who stutter (AWS) who were independently diagnosed with persistent developmental stuttering participated in our study. Six participants were excluded from the analysis, as their total SSI-4-score was below the critical limit of 10. The SSI-4-scores of the remaining 18 participants (two women, 16 men; ranging from 22-59 years of age) ranged from 12 (very mild) to 32 (severe).

2.4. Annotation

The second author, a certified speech therapist, manually annotated each stuttered event during the reading experiment using Praat [14]. A word was labelled as stuttered if at least one of the symptoms (prolongation, repetition, or block) occurred during its production.

2.5. Modelling procedure

Regression modelling was conducted on our binary response variable *stuttered* (stuttered vs. not stuttered) using generalized logistic mixed-effects models (GLMMs [15]). We conducted separate analyses for manipulations of labial place of the initial consonant and for velar/glottal place (cf. Materials: labial: nasal, plosive, fricative vs. velar/glottal: glottal stop, plosive, fricative, see Table 1).

All manipulated predictors (*manner, transition, vowel length, word status*) and their two-way interactions as well as the psycholinguistic covariates (*word length, biphone frequency* of initial CV-structure, *syllable frequency* of initial syllable) and their interactions with *word status* were used as independent variables, and AWS was included as random factor. Starting with a fully specified model [16], we used backward selection procedures [17] to first reduce the random effects structure, keeping only significant random slopes and intercepts in the model. In a second step, we reduced the fixed effects structure, using a stepwise backward elimination procedure, dropping all non-significant interactions and main effects until the model contained only terms that contributed to the overall model. All model comparisons were conducted using LRT [18]. We report type three ANOVA chi-square statistics of main effects and interactions from package *car* [19] only for factorial predictors with more than two levels.

3. RESULTS

After excluding those AWS who did not score as clinically relevant in the SSI-4, three trials due to technical errors, and 61 additional trials containing pronunciation errors, the data set consisted of 5120 data points from 18 AWS. 1363 (27%) of these cases were stuttered with a high variability between speakers. Two AWS, although noticeably dysfluent in spontaneous conversation, rarely stuttered during the experiment. We report two separate analyses, modelling stuttered events of firstly, targets with initial labial consonants and secondly targets with initial glottal or velar consonants.

3.1. Labial onsets

AWS stuttered on 724 (28%) out of 2550 possible targets with labial consonants at word onset, ranging from 0.7% to 94.2% stuttering per AWS. For the modeling procedure as described above, we used Helmert contrasts for *manner*, differentiating effects of fricatives from effects of plosives in the first contrast and compared both of them to nasal /m/ in the second contrast. *Transition* (small, medium, large), as an ordered factor, was modelled using polynomial contrasts, and we used sum contrasts for *word status* and *vowel length*. Table 2 shows the summary model statistics of targets with an initial labial consonant. The *β*-estimates represent the log odds ratios with positive values reflecting a tendency towards higher stuttering rates and vice versa.

We replicated a significant positive relationship between *word length* and stuttering. Analysis also showed a significant linear component of *transition*: larger distances between consonants' place and vowels' backness are related to higher stuttering probabilities. Furthermore, nonsense-words were significantly more often stuttered than words. *Word status* (word vs. nonsense-word), *manner* of the word initial consonant and *vowel length* significantly contributed to the random effects. All other predictors and their interactions did not significantly contribute to the model.

Random effects				
Groups Name				*Variance*
(Intercept) | AWS				7.68
Word Status | AWS				0.18
Manner (plosive vs. fricative) | AWS				0.41
Manner (plos./fric. Vs. nasal) | AWS				0.15
Vowel Length | AWS				0.59
Number of obs.: 2550, groups: AWS: 18				
Fixed effects				
Predictor	*β*	*SE*	*Z*	*p*
(Intercept)	-2.15	0.66	-3.24	**.001**
Word Length	0.71	0.13	5.29	**<.001**
Transition linear	0.24	0.11	2.15	**.032**
Word Status (Words)	-0.29	0.13	-2.25	**.024**

Table 2: Outcome of the GLMM analysis of targets with labial CV-onsets

3.2. Velar/glottal onsets

Targets with an initial velar/glottal place of articulation were stuttered in 639 of all 2570 cases (25%), ranging from 0% to 74.6% per AWS. We again used Helmert contrasts for *manner* differentiating effects of glottal stops from the effect of plosives in the first contrast and compared both of them to fricative /h/ in the second contrast. All other factors (*transition*, *word status* and *vowel length*) were modelled as in the labial onset condition. Table 3 displays the summary statistics of the model.

Analysis revealed a main effect of *manner* of the initial consonant ($\chi^2(2) = 15.29$, $p < .001$). Although there was no significant difference in stuttering between words starting with glottal stops and words starting with plosives, words starting with glottal fricative /h/ were far less often stuttered. Interestingly, we also found a significant interaction between *manner* and *vowel length* ($\chi^2(2) = 8.48$, $p = .014$). Figure 1 displays the huge stuttering difference on glottal stops before short vowels compared to

stuttering on glottal stops before long vowels. This difference is slightly more prominent than the difference between plosives before short vs. long vowels. However, with glottal fricative /h/ vowel length did not affect stuttering frequency. Again, we replicated the effect of *word length* on stuttering frequency. The structure of relevant random effects resembled that of the labial model. Again, no other effects contributed to the model.

Random effects				
Groups Name				*Variance*
(Intercept) | AWS				8.70
Word Status | AWS				0.19
Manner (glottal stop vs. plosive) | AWS				0.61
Manner (g.s./p. vs. fricative) | AWS				0.21
Vowel Length | AWS				0.93
Number of obs.: 2570, groups: AWS: 18				
Fixed effects				
Predictor	*β*	*SE*	*z*	*p*
(Intercept)	-2.74	0.71	-3.84	**<0.001**
Manner (g.s. - p.)	-0.27	0.22	-1.23	0.219
Manner (g.s./p. - f.)	-0.51	0.14	-3.72	**<0.001**
Word Length	0.83	0.16	5.12	**<0.001**
Vowel Length (short v.)	0.35	0.26	1.38	0.163
M (g.s. - p.): VL	-0.16	0.08	-1.94	0.052
M (g.s./p. - f.): VL	-0.12	0.05	-2.29	**0.022**

Table 3: Outcome of the GLMM analysis of targets with velar/glottal CV-onsets (M = Manner, VL = Vowel Length, g.s. = glottal stop, p. = plosive, f. = fricative)

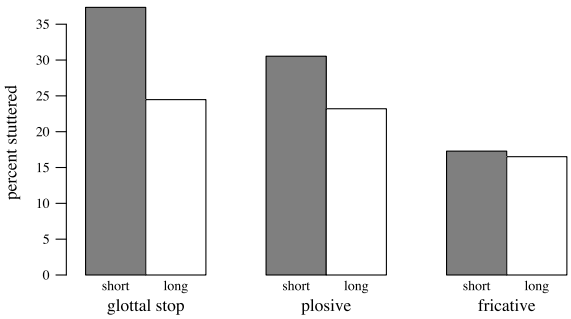


Figure 1: Percent of stuttered targets displaying the interaction of *manner* and *vowel length* of CV-structures with velar/glottal onset consonants

4. DISCUSSION

This study investigated whether a specific combination of phonological-phonetic factors influences speech fluency in experimental reading lists consisting of words and phonotactically legal nonsense-words in German. We replicated the word length effect in line with many other studies on stuttering [20, 21, 22]. Disyllabic items require a higher effort of coordinating two syllables into one metrical unit. The significant interaction between manner of articulation (of first consonants) and vowel length hints at a special status of the fricative /h/ offering itself as a softer onset often targeted in fluency enhancing techniques. In line with Natke and colleagues [10], in all other conditions (glottal stops and velar plosives), short vowels are more susceptible to stuttering than long vowels (see Fig. 1) which is also supported by the significant contribution of vowel length to the random effects in both models. The specific combination of consonantal onsets that cannot be lengthened followed by short vowels might be the most critical structure as it offers no flexibility to (i) smoothen the onset or to (ii) lengthen the vowel. Additionally, short vowels in our disyllabic items always occurred within ambisyllabic structures, making syllable boundaries highly opaque, thereby even increasing the likelihood of stuttering. Syllables with long vowels, in contrast, offer more flexibility to handle vowel production (i.e., by lengthening them) and are associated with more transparent syllable boundaries. The sooner a syllabary boundary is identified at the interface between phonological and phonetic encoding, the faster a phonological syllable can be transformed into phonetic units or motor programs to be overtly articulated. The significant transition effect, i.e., higher stuttering rates due to a larger distance between initial consonants and adjacent vowels, reflects additional articulatory effort. In Japanese, similar transition effects have been found from syllable nuclei to codas [7, 8] suggesting that language-specific articulatory properties drive the different findings (also regarding manner) in cross-linguistic studies on stuttering. On a more general level, the finding that nonsense-words increased stuttering rates while there was no interaction with other factors might again point toward fluency enhancing techniques which are bound to familiar word structures.

In sum, our results show that both lexical variables and motor-articulatory variables play a crucial role in stuttering. A common framework that spans across the phonology-phonetics interface as well as motor programming and articulation is needed to understand the contribution of the single levels and the interplay of linguistic and motor-level variables (for a recent discussion see also [22, 4]). To date, existing models are too narrowly focused on their respective areas (i.e., psycholinguistics, phonetics). Integrating data from different speaker groups will allow for the incorporation of individual differences of fluent and dysfluent speakers in such models. This seems a necessary asset of such models as the high variability between speakers as well as language specific differences particularly play out at the interface between (psycho-)linguistic and motor processing.

5. REFERENCES

[1] Bloodstein, O., Ratner, N. B., & Brundage, S. B. 2021. *A handbook on stuttering*. Plural Publishing.

[2] Wingate, M. E. 2002. *Foundations of stuttering*. San Diego, CA: Academic Press.

[3] Sasisekaran, J. 2014. Exploring the link between stuttering and phonology: A review and implications for treatment. *Seminars in Speech and Language*, *35*(2), 95–113.

[4] Brundage, S. B., & Bernstein Ratner, N. 2022. Linguistic aspects of stuttering: Research updates on the language–fluency interface. *Topics in Language Disorders*, *42*(1), 5–23.

[5] Levelt, W. J., Roelofs, A., & Meyer, A. S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.

[6] Wingate, M. E. 1988. *The Structure of Stuttering: A Psycholinguistic Analysis*. New York: Springer.

[7] Shimamori, S., & Ito, T. 2009. Difference in frequency of stuttering between light and heavy syllables in the production of monosyllables: From the viewpoint of phonetic transition. *The Japanese Journal of Logopedics and Phoniatrics*, *50*, 116–122.

[8] Shimamori, S., & Ito, T. 2010. Does the transition from the onsets to the core vowels affect the frequency of stuttering in Japanese? *Japanese Journal of Special Education*, *48*, 23–29.

[9] Wingate, M. E. 1979. The first three words. *Journal of Speech, Language, and Hearing Research*, *22*(3), 604–612.

[10] Natke, U., Sandrieser, P., van Ark, M., Pietrowsky, R., & Kalveram, K. T. (2004). Linguistic stress, within-word position, and grammatical class in relation to early childhood stuttering. *Journal of Fluency Disorders*, *29*(2), 109-122.

[11] Baayen, R. H., Piepenbrock, R., & Gulikers, L. 1995. The CELEX lexical database (release 2). *Distributed by the linguistic data consortium, University of Pennsylvania*.

[12] Riley, G.D. 2008. *Stuttering Severity Instrument (SSI-4): Examiner Manual and Picture Plates.* Austin, TX: Pro-Ed.

[13] Forster, K.I., & Forster, J.C. 2003. DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments and Computers*, *35*(1), 116–124.

[14] Boersma, P. & Weenink, D. 2018. Praat: doing phonetic by computer [Computer program]. Version 6.0.40 retrieved 22 May 2018 from https://www.praat.org.

[15] Bates D., Mächler M., Bolker B., Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

[16] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

[17] Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. 2015. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

[18] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

[19] Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ... & Monette, G. 2012. Package 'car'. *Vienna: R Foundation for Statistical Computing*, *16*.

[20] Brown, S. F. 1945. The Loci of Stutterings in the Speech Sequence. *Journal of Speech Disorders*, *10*(3), 181–192.

[21] Wingate, M. E. 1967. Stuttering and word length. *Journal of Speech and Hearing Research*, *10*(1), 146–152.

[22] Max, L., Kadri, M., Mitsuya, T., & Balasubramanian, V. 2019. Similar within-utterance loci of dysfluency in acquired neurogenic and persistent developmental stuttering. *Brain and Language*, *189*, 1–9.

STRATEGIES OF VOCAL ADAPTATION TO BACKGROUND NOISE OF DYSPHONIC AND CONTROL SCHOOLTEACHERS

Amelia PETTIROSSI[1,2], Nicolas AUDIBERT[1], Lise CREVIER-BUCHMAN[1,3]

[1]Laboratoire de Phonétique et Phonologie, UMR7018 Université Sorbonne Nouvelle/CNRS, Paris, France
[2]Imagerie & Cerveau, UMR1253 INSERM/Université de Tours, Tours, France
[3]Hôpital Foch : Service de Laryngologie Phoniatrie, Université Paris Saclay, Suresnes, France

amelia.pettirossi@univ-tours.fr, nicolas.audibert@sorbonne-nouvelle.fr, lise.buchman1@gmail.com

ABSTRACT

This paper aims at observing strategies of vocal adaptation to background noise in a population of 37 control and 24 mild dysphonic female schoolteachers, recorded in a quiet "neutral" reading context and then by picturing reading in front of a noisy classroom.

Beyond the expected consequences of the Lombard effect, vocal coping strategies are set-up in a noisy environment by all speakers, with steeper spectral slopes and longer reading times. Observed strategies may be interpreted as a way to be better perceived by listeners.

Comparison of Long-Term Average Spectrum shapes between conditions and groups suggest that control speakers implement more efficient coping strategies with a greater low-frequency reinforcement that may be interpreted as a better vocal projection. On the other hand, this adaptability seems to be limited by organic capabilities related to the laryngeal pathophysiological conditions in dysphonic speakers.

Keywords: coping strategies, schoolteachers, dysphonia, quiet vs. noisy context, adaptability

1. INTRODUCTION

Functional dysphonia is due to a bad vocal gesture leading to a misuse or an overuse of the vocal folds, the consequence of which is an alteration of the voice quality and a discomfort in the oral communication process [1]. As voice is schoolteachers (ST) main "working tool", they are considered as voice professionals. However, French schoolteachers receive very little (if any) voice coaching, which results in them being highly affected by dysphonia. In fact, a literature review highlights that dysphonia is significantly more present in schoolteachers than in the rest of the general population and this is particularly true for women, who are a majority in this profession since they constituted 83% of the entrants in 2000 [2].

This high prevalence of dysphonia is, among others, the consequence of their evolution in a noisy environment. The average ambient noise in a classroom is about 72 dB [3], for comparison, a level of 80 dB corresponds to heavy road traffic. This constant background noise then forces STs to increase their intensity to be properly perceived by pupils: this phenomenon is called "Lombard effect" [4]. As a result, STs work in a constantly noisy environment, which has an effect on the vocal effort required to be heard. The "ideal" average intensity for conversational speech between a speaker and a listener in a quiet room is 50 dB at a distance of 1m, the ideal intensity then increases by 3 dB per 10 dB of background noise [5]. In a classroom averaging 72 dB, ST have to speak at an average of 9 dB higher than their conversational intensity during their working day. Even though the Lombard effect is an automatic environmental accommodation phenomenon also shown in non-human primates [6], it necessarily generates vocal forcing [2,4]. In fact, the increase of the intensity is correlated to the increase of the fundamental frequency [7,8]. Thus, rising the f0 leads to more microtraumas related to the contact of the vocal folds.

Furthermore, we can say that there is a vocal "overdose" among schoolteachers since they give 6 to 7 hours of class per day, 5 to 6 days a week, which results in much too short resting periods for the vocal folds [9].

We wish to investigate potential vocal adaptation strategies to background noise in a population of dysphonic and control female schoolteachers. Although we expect to observe some adaptability in both groups, we assume that this will be less pronounced in the dysphonic one.

2. METHODOLOGY

2.1. Participants, corpus and recording sessions

Before the beginning of our data collection, the following experimental protocol was evaluated and validated by an ethics committee specialized in behavioral and health research, the "*Comité d'Ethique pour les Recherches Comportementales et En Santé*" (CERCES) of Paris Descartes University.

61 female schoolteachers signed a consent form and a "right to voice" form before being recorded under controlled conditions to collect audio data.

They were all native French speakers, actively working in a public or private school.

Our analyses focus on two readings of the French version of "The north wind and the sun". The first is read in a quiet and "neutral" condition and the second one by picturing reading in front of a "noisy" classroom.

2.2. Perceptual evaluation of vocal pathology

Voice quality was assessed using the GRBAS scale [10, 11]. This perceptual evaluation scale assesses the grade of dysphonia (G), the roughness (R), the breathiness (B), the asthenia (A) and the vocal strain (S) by rating these vocal parameters from 0 (unimpaired) to 3 (severely impaired). Our evaluation is performed on a speech sample by a phoniatrist and a speech-therapist and reveals 24 mild dysphonic speakers (two G2 and 22 G1) and 37 speakers rated G0 considered as the "control" group.

2.3 Acoustic measures

We focus on analyses related to Long-Term Average Spectrum (LTAS). The pitch-corrected Long-Term Average Spectrum (pitch-corrected LTAS) is chosen as an alternative to the "classic" LTAS because of its greater robustness to intra-speaker variability [12]. The pitch-corrected LTAS and spectral slopes are computed from all voiced parts of the signal for frequencies below 10 kHz, with a bandwidth of 50 Hz.

3. RESULTS

First of all, when comparing the average spectral slopes and total reading times between control and dysphonic speakers, we observe limited differences between the two groups (Table 1).

Spectral slopes are on average steeper for both types of reading even if this phenomenon is more noticeable in the dysphonics compared to the controls. This difference between groups is significant in the "noisy" context. Similarly, the average total reading time is slightly longer for dysphonic speakers in the "noisy" condition, but this difference is not significant, and the standard deviation is also larger for this group.

The comparison between contexts, all speakers pooled, allow us to highlight a pattern of accommodation common to both groups in the "noisy" condition. Indeed, the spectral slopes are significantly lower in the "noisy" context than in the "neutral" one (Table 2).

At the same time, our data reveal significantly longer total reading times, still all speakers combined, when STs are reading in the "noisy" condition compared to the "neutral" environment (Table 2).

Spectral slope	Control	Dysphonic	Independent t-test
"Quiet" context	-24.47 (2.23)	-25.70 (2.92)	t(59)=1.831 p=0.07
"Noisy" context	-19.21 (2.86)	-20.96 (3.29)	t(59)=2.176 p=0.03
Total duration	Control	Dysphonic	Independent t-test
"Quiet" context	42.18 (3.51)	42.18 (4.68)	t(59)=0 p=0.50
"Noisy" context	45.38 (4.32)	46.63 (7.18)	t(59)=-0.84 p=0.20

Table 1: Spectral slopes (dB) and total reading time (sec) compared between control vs. dysphonic speakers in "quiet" and "noisy" context. Mean values are presented with standard deviation in parentheses.

Spectral slope	"Quiet" context	"Noisy" context	Paired t-test
All speakers pooled	-24.95 (2.59)	-19.90 (3.15)	t(60)= -16.12; p<0.0001
Total duration	"Quiet" context	"Noisy" context	Paired t-test
All speakers pooled	42.18 (4.01)	45.88 (5.65)	t(60)= -8.1; p<0.0001

Table 2: Spectral slopes (dB) and total reading time (sec) compared between "quiet" vs. "noisy" context, all speakers pooled. Mean values are presented with standard deviation in parentheses.

These results suggest that there is a common behaviour between dysphonics and controls even if dysphonics seems to be slightly more impacted by the "noisy" environment as we can see by comparing the LTAS shapes between conditions and groups.

The visual inspection of graphical representations of the average pitch-corrected LTAS of our two reading contexts confirm a common pattern between dysphonics and controls (Figure 1). In both cases we can see an increase of energy in the low frequencies from 1 to 3 kHz when reading in front of a "noisy" class, with a larger difference for control speakers. On the other hand, we can observe a difference in the high frequencies, since there is an increase of energy between 5 and 7 kHz for dysphonic speakers in both

conditions, with a larger difference between groups of speakers in the "noisy" context.

Finally, we present the average difference between the "neutral" and "noisy" conditions after matching the values per speaker, compared between groups (Figure 2).

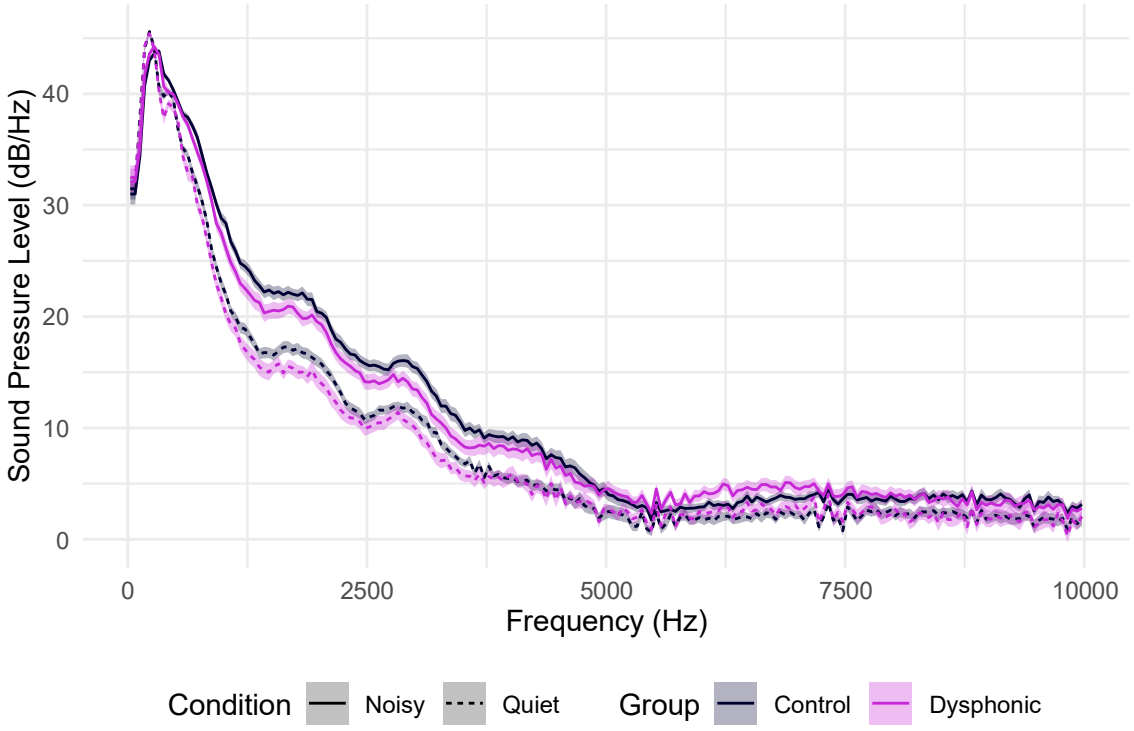


Figure 1: Average pitch-corrected LTAS with standard error of the control group (blue line) and the dysphonic group (purple line) during the reading in "neutral" condition (dotted line) and "noisy" condition (solid line)

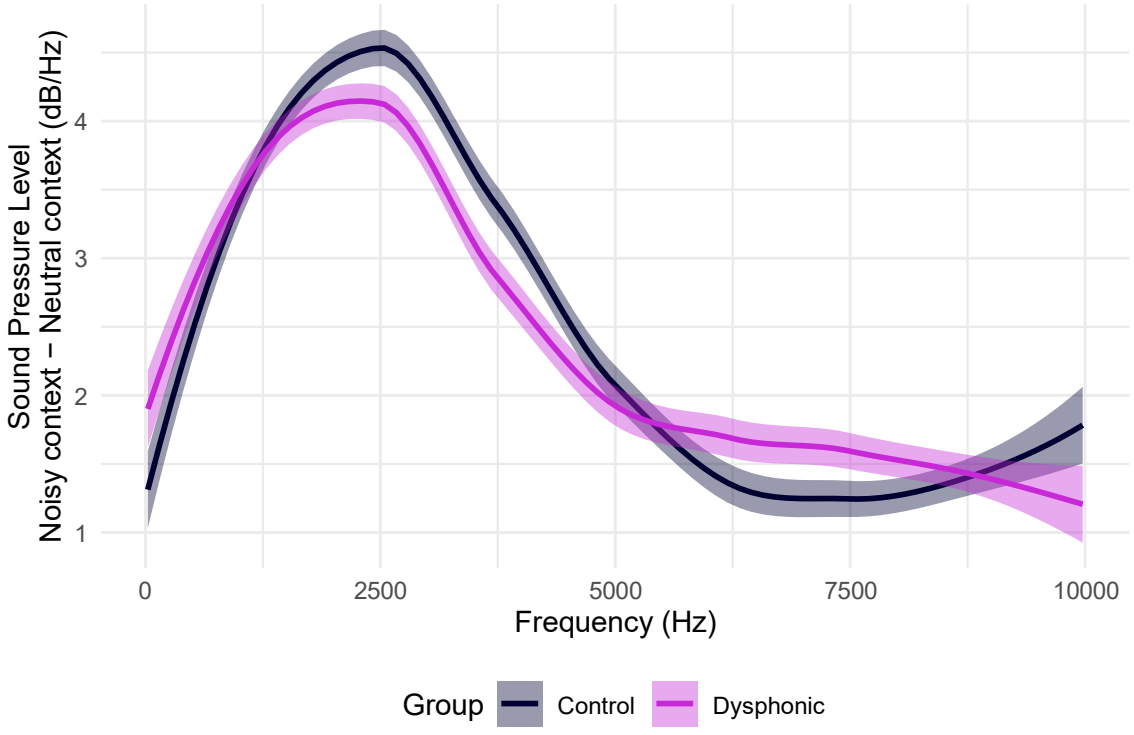


Figure 2: Mean difference in sound pressure level (dB/Hz) with standard error in the "neutral" vs. "noisy" condition smoothed by a local "loess" regression, for the dysphonic group (purple line) and the control group (blue line)

The curves presented in this figure correspond to the difference in acoustic energy in all frequency ranges when switching from one condition to another. In order to allow an easier visualization of the general trends for wider frequency ranges beyond local fluctuations, plotted values are smoothed by means of a "loess" local regression. This representation confirms a greater reinforcement of the energy in the low frequencies for controls. We also observe the previously discussed greater increase in high frequency energy for dysphonic schoolteachers in the "noisy" condition compared to the "neutral" one.

4. DISCUSSION AND PERSPECTIVES

Beyond the expected consequences of the Lombard effect [4], research looking at speaker's ability to enhance acoustic contrasts between their voice and background noise concluded that there is an expected increase in intensity but also an amplification of the vocal spectrum around 3 kHz [13]. This increase in energy around 3 kHz is not observable in "quiet" context speech, nor in shouted speech [14]. The description of this energetically enhanced area perfectly matches what we observe for the schoolteachers analysed in our study. This phenomenon could therefore be the result of background noise compensation and not only of the expected increase in intensity and spectral modifications present in shouted speech. Furthermore, this reinforcement is stronger for the control group and it could be linked to the "actor's formant" which is a phenomenon similar to the "singer's formant". The singer's formant leads, when produced, to a better evaluation of voice quality [15]. While the literature clearly establishes the existence of the singer's formant for male voices, it is sometimes not observed for women [16]. Still, some studies conclude that there is an increase in energy between 4 and 5 kHz, which would then correspond to the actor's formant for female voices [17]. Although moderate, the increase in energy around 3 kHz observed in our data could be related to this phenomenon. Our results indeed suggest the use by female schoolteachers of a reinforcement of this frequency range in order to allow for more effective communication, especially in a noisy environment requiring an increased vocal efficiency. The difference observed between the dysphonic and control groups also suggests that the dysphonic speakers would have more difficulty reinforcing this frequency range to project their voice efficiently.

The observation of an increase of energy between 5 and 7 kHz for dysphonic speakers in both conditions is consistent with the literature since a high frequency peak can be considered as a signature of dysphonic speech [18]. Based on our data, this "acoustic signature" of dysphonia would be more salient when the vocal effort increases.

These vocal compensation strategies may be interpreted within the frame of Lindblom's hypo- and hyper-articulation theory [19] as a response to communicative demands, in order to compensate for a noisy classroom environment or vocal impairment which would impair listener's perception without such adaptation.

The principle of plasticity (and thus of hyper-articulation) aiming at making large gestures to be better perceived by the listener, may explain the observed longer total reading times and the increased energy between 2 and 3.5 kHz. Moreover, the fact that this phenomenon is more visible for control speakers can be interpreted as the consequence of better vocal projection capacities in this group whereas this adaptability seems to be limited by organic capabilities related to the laryngeal pathophysiological conditions in dysphonic speakers.

Finally, observation of our data leads us to question what seems to be a compensation mechanism to a noisy environment. Extending these analyses to reading tasks in multiple settings with different controlled intensity of background noise would provide a better understanding of the extent to which the negative factors of background noise and dysphonia may accumulate and require greater compensation to ensure a successful communication.

Following the covid-19 pandemic and the mandatory wearing of masks in schools, it would also be interesting to add this constraint to our future experimentations. Indeed, several studies have already shown a link between the wearing of protective masks and a decrease in intelligibility in classrooms [20, 21] or an alteration in the student-teacher relationship [22]. Future work will investigate to what extent such alterations could cumulate with communication impairment due to dysphonia.

5. CONCLUSION

We can therefore conclude that all the female schoolteachers use a common strategy of vocal adaptation to better project their voice in order to improve the contrast between their voice and the background noise of the classroom. On the other hand, we note a greater adaptability for the control group compared to the dysphonic group.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the "Investissements d'Avenir" program ANR-10-

LABX-0083 (Labex EFL) and by ANR GEPETO (ANR-19-CE28-0018).

6. REFERENCES

[1] Crevier-Buchman, L., Tessier, C., Sauvignet, A., Arpin, S. B., & Monfrais-Pfauwadel, M.-C. (2005). Diagnostic d'une dysphonie non organique de l'adulte. Revue de Laryngologie Otologie Rhinologie, 126, 353-360.

[2] INSERM: Expertise Collective. (2006). La voix : Ses troubles chez les enseignants. Paris: INSERM, Institut national de la santé et de la recherche médicale.

[3] Shield, B., & Dockrell, J. E. (2004). External and internal noise surveys of London primary schools. The Journal of the Acoustical Society of America, 115(2), 730-738.

[4] Dejonckere, P. H., & Pépin, F. (1983). Etude de l'effet Lombard par la mesure du niveau sonore équivalent. Folia Phoniatrica et Logopaedica, 35(6), 310-315.

[5] van Heusden, E., Plomp, R., & Pols, L. C. W. (1979). Effect of ambient noise on the vocal output and the preferred listening level of conversational speech. Applied Acoustics, 12(1), 31-43.

[6] Ruch, H., Zürcher, Y., & Burkart, J. M. (2018). The function and mechanism of vocal accommodation in humans and other primates: Vocal accommodation in human and other primates. Biological Reviews, 93(2), 996-1013.

[7] Alku, P., Vintturi, J., & Vilkman, E. (2002). Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Communication, 38(3-4), 321-334.

[8] Zee, E. (1977). Duration and intensity as correlates of F0. Journal of Phonetics, 6(3), 213-220.

[9] Titze, I. R. (1999). Toward occupational safety criteria for vocalization. Logopedics Phoniatrics Vocology, 24(2), 49-54.

[10] Hirano, Minoru. (1981). Clinical examination of voice. Wien; New York: Springer-Verlag.

[11] Hirano, M. (1989). Objective evaluation of the human voice: Clinical aspects. Folia Phoniatrica et Logopaedica, 41(2-3), 89-144.

[12] Pettirossi, A., Audibert, N., & Crevier-Buchman, L. (2017). Le spectre moyen à long-terme corrigé en F0 (pitch-corrected LTAS) : Une mesure robuste de la voix dysphonique sur des énoncés de nature variable. Présenté à 7èmes Journées de Phonétique Clinique, Paris, France.

[13] Garnier, M., & Henrich, N. (2014). Speaking in noise : How does the Lombard effect improve acoustic contrasts between speech and ambient noise? Computer Speech & Language, 28(2), 580-597.

[14] Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. The Journal of the Acoustical Society of America, 115(1), 362-378.

[15] Leino, T., Laukkanen, A.-M., & Radolf, V. (2011). Formation of the Actor's/Speaker's Formant : A Study Applying Spectrum Analysis and Computer Modeling. Journal of Voice, 25(2), 150-158.

[16] Master, S., De Biase, N. G., & Madureira, S. (2012). What About the "Actor's Formant" in Actresses' Voices? Journal of Voice, 26(3), 117-122.

[17] Tayal, S., Stone, S., & Birkholz, P. (2017). Towards the measurement of the actor's formant in female voices. Studientexte zur Sprachkommunikation: (p. 286-293). Présenté à Konferenz Elektronische Sprachsignalverarbeitung, Saarbrücken, Deutschland.

[18] Kitzing, P., & Åkerlund, L. E. I. F. (1993). Long-time average spectrograms of dysphonic voices before and after therapy. Folia Phoniatrica et Logopaedica, 45(2), 53-61.

[19] Lindblom, B. (1990). Explaining Phonetic Variation : A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Éds.), Speech Production and Speech Modelling (p. 403-439). Dordrecht: Springer Netherlands.

[20] Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., & Kirk, K. I. (2020). Effect of masks on speech intelligibility in auralized classrooms. The Journal of the Acoustical Society of America, 148(5), 2878-2884.

[21] Caniato, M., Marzi, A., & Gasparella, A. (2021). How much COVID-19 face protections influence speech intelligibility in classrooms?. Applied Acoustics, 178, 108051.

[22] Fykaris, I., Matiaki, V., & Geroulidou, D. C. (2021). The Mask of Personal Protection as a Cause of Depersonification of the Learners in Formal Education. Journal of Education, Society and Behavioural Science, 34(1), 30-41.

PHONOLOGICAL AND SOCIOPHONETIC INFORMATION IN DYSARTHRIC SPEECH: A FIRST ARTICULATORY INVESTIGATION ON ITALIAN

Gili Fivela Barbara, Sonia d'Apolito, Anna Chiara Pagliaro

University of Salento – CRIL-DReAM, Lecce, Italy
{barbara.gili,sonia.dapolito,annachiara.pagliaro}@unisalento.it

ABSTRACT

In Italian, alveolar plosives and fricatives are different phonemes, and in some varieties the formers may be realized as aspirated consonants, which represent sociophonetic markers. Dysarthric speakers may experience different types of difficulty in producing the segments mentioned above, as fricatives are known to require finer supralaryngeal control than plosives, and a specific synchronization of laryngeal and supralaryngeal gestures is required in producing aspirated plosives.

In this paper, plosives, fricatives, and plosive aspiration phases are acoustically and articulatorily investigated in Parkinsonian dysarthric Italian speakers and matching controls. Research questions concern both aspects of motor control strategies in consonant production, and possible differences in the realization of phonemic distinctions and sociophonetic markers. Results suggest that, even if fricatives and plosives are kept different, the disease seems to have a stronger impact on the production of the latter. Changes may regard both the realization of phonemic distinctions and the production of sociophonetic markers.

Key Words: Parkinson's Disease, plosives, aspiration, fricatives

1. INTRODUCTION

Plosives and fricatives are among the most frequent sounds in the world languages. Almost 92% of languages have voiceless plosives series [1:27], and moreover it "is natural for a language to have at least one sibilant, namely, a voiceless alveolar [s]" [2:108]. Italian, for instance, shows both plosives and fricatives alveolar phonemes. In motor control terms, the production of plosives is considered simpler than that of fricatives, in line with the hypothesis that a collision with another articulator requires less precision than the realization of a narrowing between articulators [3, in line with 4]. This may hold for Parkinson's Disease (PD) dysarthric speech, where [5] found that /t/ is distorted by realizing incomplete closures or voicing (word medially more than word initially), while /s/ may be distorted by producing stopping, voicing, unintelligible segments or even omissions; consistently, [6] found a reduction of the tongue blade elevation in PD's alveolar fricatives (/s/, /z/) in comparison to controls'. However, previous work on Italian [7, 8] suggests a different picture, where PDs realize fricatives similarly to controls and differ in plosive production (where the physical obstacle ends the closing gesture but requires, e.g., a longer release [7]).

Besides the distinction between plosives and fricatives, in some languages plosives may also be aspirated, with aspiration playing a phonological role (e.g., Korean) or representing an allophonic variant (e.g., English). In some varieties of Italian (e.g., Salentinian Italian), aspirated plosives are allophonic variants and aspiration is affected by prosodic conditions, in that it is stronger in word internal poststress position or before pauses [9:§150, 10]. Acoustically, aspiration may be measured as the Voice Onset Time (VOT, the time interval from consonant release to start of vocal folds vibration), since aspiration represents the air outflow through the glottis after the closure release in the vocal track [11]. From an articulatory point of view, in aspirated plosives the glottal opening starts after the oral closure and reaches its peak at the closure release, rather than earlier than it as it happens in non-aspirated plosives [12, 13]. As for pathological speech, the review in [14] recalls that VOT abnormalities in dysarthria (and apraxia) mainly reflect loss of motor control, with dysarthric speakers showing in plosives longer closure duration and VOT than controls (see also the review in [15] and, for English, [16]). As for English speakers with PD hypokinetic dysarthria, however, results are inconsistent. They point to decreased VOTs in PDs and reduced contrasts in voiced and voiceless plosive VOT [17, 18], as also reported in Korean bilabials [19] and alveolars [20]; however, results also point to the lack of difference in VOT length between PDs and controls ([21]; see also [22]). Crucially, though, Korean dysarthric speakers seem to control the synchronization between laryngeal and supra-laryngeal articulators to distinctively produce phonologically different (aspirated vs. non-aspirated) plosives [19]. What about aspiration when it represents a sociophonetic trait?

In this paper, the distinction between plosives and fricatives in Italian, as well as the plosive aspiration phase, are investigated from both the acoustic and articulatory point of view. Crucially, besides the phonological contrast between plosives and fricatives, in the variety of Italian considered here, namely the Salentinian variety (Southern Apulia, Italy), the aspiration of plosives is an allophonic variant that is sociolinguistically (but not phonologically) relevant.

2. GOALS AND HYPOTHESES

We aim at observing if PD speakers, who are still able to preserve the plosive vs. fricative contrast: 1) exhibit more difficulties in maintaining the narrowing gesture than the closure; 2) master as precisely as controls the coordination between laryngeal and supra-laryngeal gestures in order to maintain the sociophonetic feature of plosive aspiration; 3) show possible compensation strategies in realizing both phonological contrasts and sociolinguistically relevant features.

We hypothesize that PD speakers who maintain the plosive vs. fricative phonological difference may show: a) imprecise realization of the frication narrowing (see [3]); b) longer closure and VOT duration than controls in aspirated plosives (see [15, 16]), though a stronger aspiration in word internal poststress position may be expected (see [9, 10]); finally, we hypothesize that c) dynamical adjustment related to compensation strategies may be stronger in order to keep fricatives and plosives different than to preserve sociolinguistic traits.

3. METHODS

Tongue dynamics was observed in unvoiced alveolar plosives and fricatives, inserted as singletons (/t/, /s/) in a stable vowel context (/a/-to-/a/, e.g., /la'tata/, /la'sasa/). They were onset of both word initial, stress syllables, as well as word medial, poststress syllables, and were inserted within the carrier sentence *La X blu* 'the blue X'. Plosives could be realized as aspirated ([t^h]).

Nine Salentinian subjects, that are 5 mild-to-severe PD subjects (males; average age 71.6 y.o.) and 4 age/gender-matched healthy controls (HC: average age 69.75), have been acoustically and articulatorily recorded in ON phase through Electromagnetic Articulography (AG501). Articulatory data were acquired by means of 7 sensors, glued on subjects: 2 on tongue mid-sagittal plane (dorsum and tip), 2 on lips vermillion border (upper and lower), 1 on the nose and 2 behind the ears for normalization (one HC subject was excluded from the articulatory analysis because of errors in tongue tip tracking, which is pivotal for the present paper). Subjects read aloud the corpus of written sentences for a minimum of 5 times. After an auditory check, aimed at selecting plosives and fricatives realization of target segments, acoustic and articulatory signals were segmented and labelled in PRAAT and MAYDAY respectively [23, 24]. The acoustic duration of each segment was measured and, as for plosives, both the closure (stop closure to spike) and the VOT duration (spike to beginning of following vowel) were collected, differently from previous investigations [7, 8]. Regarding kinematics, the following landmarks were semi-automatically inserted along the vertical axis of tongue dorsum (TD) and tongue tip (TT) tracks: gesture target, located at the zero velocity, and, for each tracked segment, maximum velocity, labelled at the velocity peak of the relevant coil. Further, on the position track we labelled the beginning/end of the closure phase of the consonantal gesture (e.g., beginning of release, at the 20% of velocity increase after the release; measure missing in previous investigations [7, 8]). Concerning the duration of tongue tip/dorsum gestures, the following measures were calculated: (a) the interval including both the closing and the opening consonantal gestures (Clo-Op); with regards to the release phase, the interval from the consonantal target both to (b) the vowel target (Ctarget-Vtarget latency) and (c) the onset of the vowel (20% of peak velocity - Ctarget-Vonset). Further, as for the consonantal gesture to the following vowel, we measured the peak velocity and stiffness (peak velocity/gesture amplitude). Time measurements were analyzed as both absolute and normalized (over the acoustic/articulatory word duration) measures.

As for statistical analysis, mixed models were implemented (in R, lme4 [25, 26]). The fixed effects were *Population* (PD vs. HC), *Constriction* (plosives vs. fricative) *Position* (initial-stress vs. medial-poststress, henceforth stress and poststress), and *Repetition* (number of productions). Interactions were also investigated. By subject random slopes were set in order to account for inter-subjects variability in the realization of *Population*, *Constriction* and *Position*. Analyses of aspiration were run on plosives only, discarding the *Constriction* factor. Residual plots did not reveal obvious deviations from homoscedasticity or normality. Significance ($p<0.05$) of fixed effects was checked via a *Likelihood Ratio Test*.

4. RESULTS

4.1. Auditory check and acoustic measures

Two PD subjects produced plosives and fricatives as barely indistinguishable segments, in that they realized plosives without a neat closure phase (slightly noisy closure and no spike in spectrogram), or as voiced rather than unvoiced segments. Thus, they have no longer

been considered, as results of measurements concerning their productions were not useful in investigating the difficulties in maintaining the narrowing gesture vs. the closure nor in analyzing the plosive aspiration phase.

Consonant acoustic duration is affected by *Position*, (absolute measures: $\chi2(1)=23.77$ p=,000; normalized: $\chi2(1)=28.55$ p=,000) and duration increases in stress in comparison to poststress position (absolute: 11.15ms ± 2.21 S.E.; normalized: 2.57% ±0.46 S.E.) for HC speakers and fricative condition. Moreover, consonant normalized duration is affected by *Population* too ($\chi2(1)=4.29$ p=,03), while absolute duration changes depending on *Constriction* ($\chi2(1)=11.14$ p=,000). Values are smaller in PDs (-2.07ms ±0.85 S.E.) and in plosives in comparison to fricatives (-7.50ms ±2.21 S.E.). No interactions are found (Figure 1).

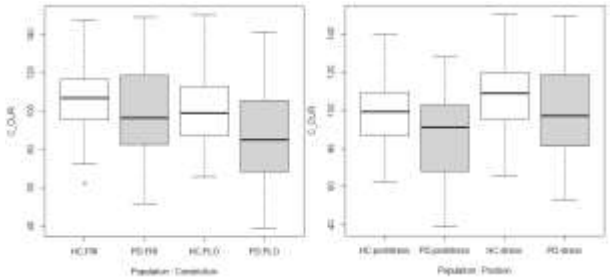


Figure 1: Consonant absolute acoustic duration for Constriction (left panel) and Position (right panel) in HCs and PDs (white and grey columns, respectively).

When considering plosives as for their closure, the duration differs significantly only with respect to *Position* (absolute: $\chi2(1)=11.50$ p=,000; normalized: $\chi2(1)=13.57$ p=,000), being longer in stress position (absolute:9.62 ±2.74 S.E; normalized:2.19% ±0.57 S.E.).

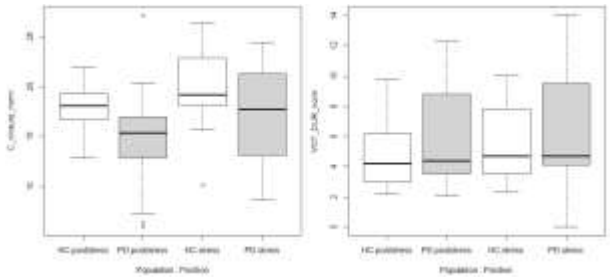


Figure 2: Consonant normalized closure (left panel) and VOT acoustic duration (right panel) for Position in HCs and PDs (white and grey columns, respectively).

The interaction between *Population* and *Position* does not reach significance, though the closure absolute duration is shorter in PDs, especially in poststress position ($\chi2(1)=9.54$ p=,08). As for the VOT phases, the duration is only affected by the *Position* factor (absolute: $\chi2(1)=7.06$ p=,007; normalized: $\chi2(1)=3.87$ p=,004) with no significant interactions. Measures show that VOT duration increases in stress in comparison to poststress position (absolute: 3.96ms ±1.45 S.E.; normalized: 0.89%- ±0.44 S.E.), though its duration appears to be highly variable in PDs - Figure 2.

4.2. Articulatory measures

The *Population* factor affects the interval including the whole consonant, i.e. the closing plus the opening gesture, as for the normalized measure (Clo-Op: $\chi2(1)=4.50$, p=0.03). A slight increase is found in PDs (e.g., by about 2% ± 0.7 S.E. in poststress position and fricative condition). The duration of the opening gesture alone does not significantly differ (Ctarget-Vtarget latency: $\chi2(1)=3.05$, p=0.08), though it slightly increases in PDs. No interaction with *Position* and *Constriction* is found.In line with the literature, absolute Clo-Op duration is affected by *Constriction* ($\chi2(1)=4.51$, p=0.03), with plosives showing shorter duration than fricatives (16.8 ms ± 6.5 S.E. in the case of plosives for HC speakers and poststress condition) - Figure 3, left panel. Interaction between *Population* and *Position* is found for both the absolute Clo-Op and the Ctarget-Vtarget measure ($\chi2(1)=4.90$, p=0.03; $\chi2(1)=12.016$, p=0.01), as PDs show shorter duration especially in poststress position.

As for the release phase corresponding to the Ctarget-Vonset, the absolute interval does not reach significance as for the *Position* factor ($\chi2(1)=2.89$, p=0.08),though values increase (by about 8 ms in HC ± 4 S.E. in the case of stress position and fricative) and increase even more in PDs realization.

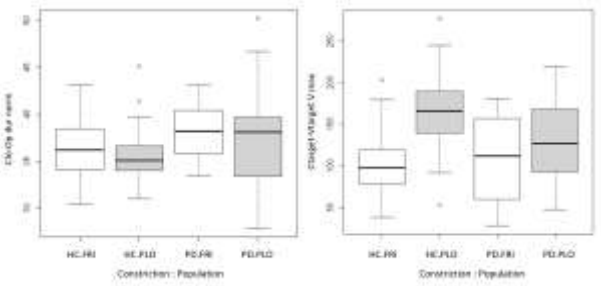


Figure 3: Consonant absolute articulatory duration (Clo-Op, left) and peak velocity in consonant opening (Ctarget-Vtarget, right) for Populationin Fricatives and Plosives (white and grey columns, respectively).

Peak velocity and stiffness measures showed the significant impact of the *Position* factor (respectively, $\chi2(1)=5.48$, p=0.02; $\chi2(1)=3.7334$, p=0.05), with expected higher values in stress than in poststress position. The opening gesture peak velocity is the only measure that is affected by the *Constriction* degree

($\chi2(1)=5.04$, $p=0.02$), as it increases in plosives (by about 45 ± 15 S.E. for HC speakers and poststress condition). However, an interaction between *Constriction* and *Population* is found, as the peak velocity is lower in PD's plosives, not differing much from that found in fricatives – Figure 3, right panel. Further, *Population* also interacts with *Position* ($\chi2(1)=3.7289$, $p=0.05$) with PDs showing increased peak velocity in poststress than in stress position.

5. DISCUSSION AND CONCLUSIONS

The analysis of productions by HC and PD speakers who are still able to preserve the plosive vs. fricative contrast shows that acoustic measurements regarding the difference between plosives and fricatives (*Constriction* factor) did not significantly differ depending on the PD vs. HC *Population*. The overall normalized duration of consonants is smaller in PDs than in HCs, but results are in line with the expected longer duration in fricatives than in plosives, as well as in stress in comparison to poststress positions. In PDs, fricatives do not seem to vary more than plosives. Rather, PDs tend to shorten the plosive closure acoustic duration in comparison to HCs, especially in poststress position, which is in line with the overall shorter duration of acoustic consonants in PDs. Finally, the plosives' VOT, which is longer in stress than in poststress position, is not significantly different in PDs and HCs. Thus, within the inconsistent results recalled in section 1, our findings support, e.g., [21, 22] rather than [15, 16], even though the VOT duration seems to be more variable in the PDs than in HCs. In fact, the VOT by PDs and HCs is longer in stress than in poststress position, in line with prosodically induced VOT modulation (overall VOT lengthening in stress position), with no evident or consistent sociophonetic differences ([9, 10]).

As for the articulatory analysis, the normalized measures regarding both the whole closing plus opening consonantal gesture and, to a certain extent, the opening gesture (Clo-Op and Ctarget-Vtarget) increase in PDs. Thus, articulatory data point to longer gestures in PDs, despite the shorter duration of acoustic segments. Only absolute measures confirm expectations regarding the shorter duration of gestures in plosives than in fricatives together with a shorter duration for PDs, especially in poststress position. Peak velocity and stiffness measures regarding the opening gesture showed higher values in stress than in poststress position, while peak velocity is the only measure that is affected by the *Constriction* degree, as it increases in plosives than in fricatives. Interestingly, peak velocity is lower in PD's plosive opening gesture, where it does not differ much from that found in fricatives. The auditory check preliminary to the analysis allows us to discard the hypothesis that the lower peak velocity is due to an incomplete realization of the closure. Nevertheless, from the articulatory data it may be hard to tell if the lower peak velocity is related to laryngeal-supralaryngeal gesture phasing difficulties [12, 13]. What we can observe, though, is that the aspiration phase seems to acquire characteristics that are similar to those found in fricatives, and this could be a stage prior to the incomplete closure realization observed in other PD speakers (e.g., those excluded in the present analysis). On the other hand, the increased peak velocity value in poststress than in stress position by PDs seems to be in line with the observed shorter aperture phase in poststress position, and with possible differences related to aspiration and consonant prosodic position.

The abovementioned results do not seem to support expectations related to major differences between PD and HC in fricatives more than in plosives (see goal 1). Rather they point in the direction of some possible inaccuracies in the realization of plosives by PDs, as their opening gesture in plosives show similarities with that found in fricatives. Further, the plosive aspiration phase in PDs (see goal 2) seems to be affected by the stress/poststress consonant position as in the case of HC speech, in line with prosodic modulation and with no evident or consistent difference due to sociolinguistic factors. In fact, no observation of longer aspiration in poststress rather than stress position [9, 10] has been found in either PDs or HCs. However, PDs show higher peak velocity in poststress than in stress position, which could be a cue of aspiration differences in comparison to what observed in HCs. Besides, difficulties in laryngeal-supralaryngeal gesture phasing in PDs could have appeared in terms of variability in VOT acoustic duration or with reference to the observed overall lowering in peak velocity in plosive opening gestures. Therefore, speakers who are still able to differentiate plosives and fricatives seem to have already modified some characteristics of their opening gesture in plosives, and changes may be detected both in the case phonemic distinctions are realized and when the production of sociophonetic markers is concerned (see goal 3).

Of course, more material is needed to clarify the matter, together with further acoustic (e.g., noise intensity measures during VOT) and articulatory measures (e.g., opening gesture amplitude).

ACKNOWLEDGEMENTS

This work was funded by the project PRIN 2017JNKCYZ. We thank F. Sigona for technical support, M. Iraci for help in the corpus collection, both M. Iraci and G. Pucci for a preliminary analysis of part of the material, and all experimental subjects for participation.

6. REFERENCES

[1] Maddieson, J. 1984. *Patterns of Sounds*. CUP.

[2] Hockett, C.F. 1955.*A Manual of Phonology(IJAL Monographs 11)*. Indiana University.

[3] Fuchs, S., Perrier, P., Geng, C., Mooshammer, C. 2006. What role does the palate play in speech motorcontrol? Insights from tongue kinematics for German alveolar obstruents, in Harrington, J., Tabain, M., (eds.), *Speech Production: Models, Phonetic Processes, and Techniques,* Psychology Press, 149-164.

[4] Löfqvistand, A., Gracco, V.L. 1997. Lip and jawkinematics in bilabial stop consonant production, *JSLHR*40, 4, 877-893.

[5] Antolík, T., Fougeron, C. 2013. Consonant distortions in dysarthria due to Parkinson's disease, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia, *Proc. Interspeech 2013*, August, Lyon, France, 2152-2156.

[6] Logemann, J.A., Fisher, H.B. 1981. Vocal tract control in Parkinson's Disease: Phonetic feature analysis of misarticulations, *JSHD*, 46, 348-352.

[7] Iraci, M., Grimaldi, M., Gili Fivela, B. 2017. Dynamic Aspects of the Alveolar Fricative vs. Stop Contrast in Parkinson's Disease, PAPE Conference (poster presentation), *Phonetics and Phonology in Europe*, Cologne, 11th-15th June, 2017.

[8] Iraci, M. 2017. Vowels, consonants and co-articulation in Parkinson's Disease. Unpublished PhD Dissertation. University of Salento, Lecce, Italy.

[9] Rohlfs, G. 1966. *Grammatica storica della lingua italiana e dei suoi dialetti*. *Fonetica*. Einaudi, v. 1.

[10] Sobrero, A., Romanello, M.T. 1981.*L'italiano come si parla in Salento*. Milella.

[11] Kim, H.S., Kim, H.H. 2019. Acoustic analysis of Korean stop sounds in patients with dysarthrias, *Clinical Arch. of Communic. Disord*, 4, 3, 201-213.

[12] Best, C.T.1995. A direct realist view of cross-language speech perception, in Strange, W. (ed.), *Speech perception and linguistic experience: issues in cross-language research*. York Press, 171–204.

[13] Browman, C.P., Goldstein, L. 1986. Articulatory gestures as phonological units, *Phonology*, 6, 151-206.

[14] Auzou, P., Ozsancak, C., Morris, R. J., Jan, M., Eustache, F., Hannequin, D. 2000. Voice onset time in aphasia, apraxia of speech and dysarthria: a review, *Clinical Ling.& Phonetics*, 14, 2, 131-150.

[15] Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C. 1989. Toward phonetic intelligibility testing in dysarthria, *JSHD*, 54, 482-499.

[16] Forrest, K., Weismer, G., Turner, G. 1989. Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric males, *J. of the Acoustical Society of America*, 85, 2608-2622.

[17] Weismer, G. 1984. Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination, in McNeil, M.R., Rosenbek, J.A., Aronson, A.E. (eds.), *The dysarthrias: Physiology, acoustics, perception, management*, College-Hill Press, 101-130.

[18] Morris, R.J. 1989. VOT and dysarthria: A descriptive study,*J. of Communication Disorders*, 22, 23-33.

[19] Park, S., Sim, H., Baik, J.S. 2005. Production ability for Korean bilabial stops in Parkinson's disease, *Journal of Multilingual Communication Disorders*, 3, 90-102.

[20] Kang, Y., Kim, Y. D., Ban, J. C., Seong, C. J. 2009. A comparison of the voice differences of patients with idiopathic Parkinson's disease and a normal-aging group, *Phonetics and Speech Sciences*, 1, 99-107.

[21] Fischer, E.,Goberman, A. M. 2010. Voice onset time in Parkinson disease",*J. of Communic. Dis.*, 43, 21-34.

[22] Bunton, K., Weismer, G. 2002. Segmental level analysis of laryngeal function in persons with motor speech disorders. *Folia Phoniatrica etLogopaedica*, 54, 223-239.

[23] Boersma, P. 2002. Praat, a system for doing phonetics by computer, *Glot International*,5 no. 9/10, 341-345.

[24] Sigona F., Stella A., Grimaldi M.,Gili Fivela B. 2015. MAYDAY: A Software for Multimodal Articulatory Data Analysisin Romano, A., Rivoira, M., Meandri, I. (eds.), *Aspetti prosodici e testuali del raccontare, Atti del X Conv. AISV*, Edizioni dell'Orso, 173-184.

[25] R Core Team, 2019. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

[26] Bates, D.S., Maechler, M., Bolker, B., Walker, S. 2015.FittingLinear Mixed-Effects Models Using lme4, *Journal of Statistical Software*, 67, 1, 1-48.

SYNCHRONIZATION TYPE MATTERS: ARTICULATORY TIMING IN DIFFERENT RHYTHMIC CONDITIONS IN PERSONS WHO STUTTER

Mona Franke[1,2,3], Nicole Benker[1], Simone Falk[2,3], Philip Hoole[1]

[1]Institute for Phonetics and Speech Processing (IPS), LMU Munich, Germany,
[2]Faculté des arts et des sciences – Départment de linguistique et de traduction, UdeM, Montreal, Canada,
[3]International laboratory for Brain, Music, and Sound Research (BRAMS), Montreal, Canada

mona.franke@phonetik.uni-muenchen.de

ABSTRACT

This study investigates articulatory timing of four persons who stutter (PWS) and four persons who do not stutter (PWNS) in different conditions: Speaking and tapping (self-paced), speaking along with a metronome (externally paced), speaking and tapping to a metronome (Metronome+Tapping). Using electromagnetic articulography, gestures of the articulatory speech onset and the finger taps were recorded and analyzed. Results show that, compared to the metronome beats, finger taps were more closely aligned with the articulatory speech onset supporting the assumption of a close link between articulatory and manual motor systems. Furthermore, our results indicate timing differences between PWS and PWNS, since intervals between metronome beat and articulatory speech onset were shorter in PWS. The Metronome+Tapping condition also led to significantly shorter intervals between articulatory onsets and finger taps in PWS. Our results suggest that PWS time their speech later when synchronizing to a metronome possibly pointing towards difficulties in movement initiation.

Keywords: stuttering, articulatory timing, gestural timing, finger tapping, paced speech.

1. INTRODUCTION

Stuttering is a neurodevelopmental speech fluency disorder that affects approximately 5-8% of children and 1% of the adult population [1]. The most characteristic symptoms of stuttering are involuntary disruptions in the flow of speech, such as pauses before a syllable (blocks), repetitions of sounds, syllables or words (repetitions) and lengthening of sounds (prolongations) [2]. These disfluencies typically occur at the beginning of (stressed) words or syllables, indicating that the speech motor program breaks down at this point. While the cause(s) of stuttering still remain(s) unknown, the breakdown of fluency in persons who stutter (PWS) has been linked to malfunctioning timing mechanisms (see [3], for a review). A recent review by Bradshaw et al. [4] proposed that in PWS the updating and use of internal models in speech motor control are disrupted, affecting both feedback and feedforward control of their speech.

Moreover, malfunctions in feedforward and feedback control in stuttering are linked to disruptions in more general motor networks in the brain, in particular, the basal ganglia-thalamo-cortical loop (e.g., [5,6]). This loop controls, among other processes, the timely initiation and termination of articulatory as well as other movements. There is indeed evidence that PWS also show alterations in non-verbal timing processes, such as finger tapping. Some studies found that PWS were more variable and tapped earlier in relation to the beat compared to persons who do not stutter (PWNS) when synchronizing finger taps to a metronome rhythm [7,8].

Interestingly, speaking with an external rhythm like a metronome reduces the occurrence of stuttered disfluencies in a major way [9]. Potentially, this phenomenon is due to higher reliance on cerebellar-cortical networks for motor control in PWS, circumventing the error-prone basal ganglia motor loop [5,10]. To date, it is unknown how inter-gestural timing (such as joint speaking and tapping) with or without an external rhythm is mastered by PWS.

Studies on joint speech and manual movements indicate that there should be a close coupling between speech and manual motor control systems (e.g., [11]). Hence, tapping and speaking at the same time (in the speech rhythm) could lead to more stable (articulatory) gestures in PWS as it is expected that this inter-gestural timing is closely linked. Moreover, it would be particularly interesting to see how PWS synchronize articulatory gestures and finger-tapping to an external rhythm. This setting can test whether timing information from multiple channels (auditory, manual + articulatory rhythm) is strongly or weakly coupled in PWS and PWNS, which might improve articulatory stability in the former or deteriorate it in the latter case. A study by Hulstijn and colleagues [12] points to weaker integration in PWS. They found that PWS were more variable in coordinating speech

and hand movements to tones than PWNS. However, they did not report how exactly the timing occurred nor whether effects on fluency were found.
Therefore, the aim of the current electromagnetic articulography (EMA) study is to shed further light on timing processes in PWS, by analyzing a) the effect of external pacing (metronome) and self-pacing (tapping) on speech gesture timing (where does the beat occur in relation to articulatory gestures) and b) how inter-gestural timing (non-verbal and verbal gestures) is affected by an external rhythm. It is an open question whether PWS synchronize their speech earlier to the metronome than PWNS, which would mirror results on non-verbal tasks (e.g., [7,8]).
Note that, in general, it is unclear whether metronome and finger-tapping synchronization time-points with respect to articulatory gestures differ or coincide. Therefore, it is another aim of our study to compare these conditions. What also remains to be answered is whether PWS differ from PWNS in the time point of synchronizing non-verbal gestures (finger tapping) and verbal gestures (speech). Following the result of Hulstijn et al. [12], we would expect that speaking along with a combination of motor and auditory pacing would lead to greater timing variability in PWS compared to PWNS.

2. METHODS

EMA data (AG501, Carstens Medizinelektronik) were collected from 10 adults who stutter and 10 adults who do not stutter. For the present paper, data of 4 persons who stutter (mean age = 24.3, 2 female) and 4 persons who do not stutter (mean age= 24.5, 2 female) were analyzed. All participants were native speakers of German and besides stuttering, no other impairments were reported. PWS and PWNS were matched in pairs having similar musical experience, the same age (±1 year), and sex.
Participants produced mono- and disyllabic German target words (cf. Table 1) embedded in the carrier phrase [ˈze:ə ____ ˈan] (Look at ____).

/a/	/o/	/u/
Maß [ma:s]	Moos [mo:s]	Mus [mu:s]
Baden [ˈba:dn̩]	Boden [ˈbo:dn̩]	Buden [ˈbu:dn̩]
Mahl [ma:l]	Mohn [mo:n]	Buhne [ˈbu:nə]

Table 1: Target words

The experiment comprised 4 conditions (see below for more details) wherein each target word occurred 4 times along with filler words in a quasi-randomized order. The conditions were conducted in the following order:

- Baseline: Reading words embedded in the carrier phrase in a self-chosen speech tempo
- Tapping condition (self-pacing): Baseline + aligning finger tap to each word
- Metronome condition (externally paced): Reading + synchronizing each word to a metronome (90bpm)
- Metronome+Tapping condition: Reading + aligning finger tap to each word while synchronizing speech to a metronome (90bpm)

The metronome tone was presented via an in-ear headphone which participants plugged in their right ear. The onset of the metronome time point closest to the target word was automatically extracted.
For the conditions where tapping was involved, participants were instructed to tap their index finger of the dominant hand on an elevated wooden block that was placed on a table close to the participants. Sensors relevant for the data we report here were placed on the tip of the index finger of the participants' dominant hand and on the upper and lower lip. In addition, we had sensors placed on the tongue and the jaw. Only fluent productions (determined by listening to audio recordings) were analyzed. Therefore, a maximum of 144 target words were analyzed per participant.

2.1. Kinematic measures

Lip activity forming the constriction for the bilabial onset was measured using Lip Aperture (LA), defined as the Euclidian distance between transducers placed on the upper and lower lip. For LA and the finger tap (FT), the gesture onset was semi-automatically detected using a 20% velocity threshold and the onset of the gesture nucleus (start of the plateau) was also semi-automatically detected.
For each target word the relative timing of the consonantal gesture (onset) to the metronome onset and the finger tap onset was calculated as the lag between the onsets of the gesture nuclei of LA and FT and the lag between the onset of the LA gesture nucleus and the acoustic metronome onset.
We will refer to the two resulting intervals as *tap - articulatory onset interval* and *met - articulatory onset interval*. Both intervals are calculated such that positive values result if the articulatory onset is before the tap or the metronome.

2.2. Statistical analyses

For statistical analyses, linear mixed effects models (*lme4* package, [13]) were conducted with R Version 4.0.2 [14]. We are aware that this method was applied to a small group of participants. However, we aim to

analyze a larger participant sample size with linear mixed models and we aim to have 10 participants per group ready for presentation at the conference. Therefore, we decided to include this method in the present paper. To determine p-values for the main effects and interactions between factors, a model including the fixed factor/interaction of interest was compared to the same model with no fixed factor/no interaction [15]. Post-hoc Tukey corrected t-tests, using the package *emmeans* [16], were performed to test significant interactions.

Variables that were included in the models as fixed factors were group (PWS and PWNS), as well as condition (Metronome, Tapping, Metronome+Tapping) or synchronization type (finger tapping, metronome) with a two-way interaction term between group and one of the latter two factors. Random intercepts were included for participant, word, and repetition number. Since repetition number did not have an effect on any predicting variables, it was excluded from all final models. Residual plots were visually checked for homoscedasticity of normality before reporting the results.

3. RESULTS

The following figure displays the interval between the articulatory onset and the metronome onset as well as the interval of the articulatory onset and the finger tap for the respective conditions (see Figure 1) and for all participants, separated by group.

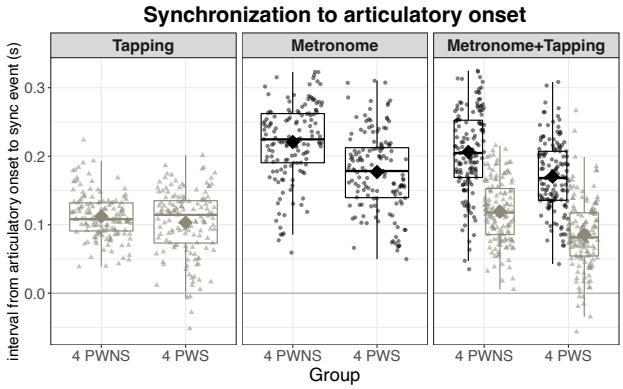


Figure 1: *tap - articulatory onset interval* (light grey, triangles) and *met - articulatory onset interval* (dark grey, dots). 0 seconds indicates the articulatory onset (nucleus onset of the bilabial), positive intervals indicate that the event of synchronization was after articulatory speech onset. Each triangle/dot represents one tap/metronome of one participant. Diamonds display the mean. Groups are displayed on the x-axis, PWNS = persons who do not stutter, PWS = persons who stutter.

As can be seen in Figure 1, finger taps are more closely aligned with the articulatory onset compared to the metronome. Hence, the *tap - articulatory onset interval* is shorter than the *met - articulatory onset interval*. A linear mixed effects model (Conditional $R^2 = 0.50$, Marginal $R^2 = 0.45$) was run in order to test whether the interval duration differs in the Metronome and the Tapping condition in PWS and PWNS. The model revealed that finger taps were aligned significantly earlier to the articulatory onset than the metronome ($p < 0.0001$). Furthermore, the model showed that group also had a significant effect on the interval duration ($p < 0.0001$). The significant interaction between condition (Metronome and Tapping) and group ($p < 0.0001$) revealed that the groups only differed significantly in the Metronome condition ($p = 0.021$) but not in the Tapping condition. Hence, PWS had significantly shorter *met - articulatory onset intervals* than PWNS.

To investigate how the combined condition (Metronome+Tapping) affected synchronization events in PWS and PWNS, we ran three different linear mixed effects models: The first model was run to test whether the synchronization time points of the two different synchronization types also differ even when they occur simultaneously in one task and whether there is a difference in timing between PWS and PWNS. The second and third models were run to test how synchronizing finger taps (second model) and metronome beats (third model) were affected by the combined condition and whether PWS and PWNS changed the timing in the combined condition compared to the single condition.

The first model (Conditional $R^2 = 0.47$, Marginal $R^2 = 0.39$) showed that PWS had significantly shorter intervals, regardless of synchronization type ($p = 0.0438$) and that finger taps were placed closer to the articulatory onset than the metronome ($p < 0.0001$). Thus, in the combined condition, PWS had shorter *met - articulatory onset intervals* as well as *tap - articulatory intervals* than PWNS. Note that this was not the case in the single Tapping condition.

These results led to the question whether synchronization time points with respect to the articulatory onset differed from the single to the combined condition in PWS and PWNS.

For the *tap - articulatory onset interval* the model (Conditional $R^2 = 0.31$, Marginal $R^2 = 0.07$) revealed a significant effect of condition ($p = 0.0003$) and a significant interaction between group and condition ($p = 0.0002$). Pairwise comparisons showed that PWS decreased the *tap - articulatory onset interval* in the combined condition compared to the single Tapping condition ($p = 0.0011$). PWNS on the other hand did not time their finger taps differently in the combined condition.

The third model (Conditional $R^2 = 0.36$, Marginal $R^2 = 0.11$) revealed that the time points of the

metronome beat shifted significantly towards the articulatory onset in the Metronome+Tapping condition compared to the single Metronome condition ($p = 0.0452$). This effect was found independently of group; no interaction was found.
Finally, in order to test if PWS were more variable than PWNS in synchronizing, the standard deviation was calculated for the different intervals per condition.

	SD Met - articulatory onset interval		**SD Tap - articulatory onset interval**	
	Met	Met + Tap	Tap	Met + Tap
PWS	0.069	0.065	0.048	0.051
PWNS	0.056	0.061	0.033	0.046

Table 2: Standard deviations for interval durations

Table 2 shows that PWS were more variable than PWNS in all conditions. However, when comparing intra-group differences between the single conditions (Tapping, Metronome) vs. the complex condition (Metronome+Tapping) it seems that PWNS increase more in variability in the intervals in the combined condition compared to PWS.

4. DISCUSSION

The study revealed both differences in timing when synchronizing speech with an internally generated rhythm (inter-gestural timing) as well as when synchronizing speech with an external rhythm (paced timing). Moreover, the data suggests differences between PWS and PWNS. We will first address differences between conditions and then group differences.
Compared to the metronome beats, finger taps were more closely aligned with the articulatory speech onset. This finding supports the idea of a close relationship between non-verbal and verbal motor systems [11,17]. Thus, joint tapping and speaking could lead to more stable gestures across modalities. Indeed, our results provide initial evidence for this conjecture as the timing of finger taps was more stable (smaller SD) than that of the external pacing with respect to the articulatory speech onset. A future study could therefore focus on the variability of the gestures themselves to test this assumption. The fact that the 8 participants in our study have longer intervals between articulatory speech onset and metronome beats could also indicate that externally paced speech is strongly based on acoustic cues. As previously shown (e.g., [18,19]) in purely perceptual studies, participants place the metronome beat within or close to the acoustic vowel onset of the target word. Hence, it is a possibility that the vowel onset is an anchor for acoustically synchronizing the metronome to one's own speech, while the syllable onset is the reference point for coordinating inter-gestural timing.

In terms of the group effect we found that PWS had shorter intervals between the metronome and the articulatory speech onset. This result indicates that PWS time their speech later to the metronome than PWNS, potentially because of later speech initiation in the group who stutters. This finding would be in line with preliminary results for children who stutter reported by Schreier et al. [20].
However, finger tapping to one's own speech did not differ between PWS and PWNS, indicating similar inter-gestural timing conditions. Interestingly, joint speaking and tapping to an external rhythm (Metronome+Tapping condition) led to a group difference in the interval between articulatory speech onset and finger tap such that PWS have shorter intervals than PWNS. This difference was caused by the fact that compared to the single Tapping condition, PWS decreased the *tap - articulatory onset interval* in the combined condition, whereas PWNS did not change the timing of their finger taps. Therefore, it can be assumed that in PWS inter-gestural timing is more affected by an external rhythmic cue than in PWNS. As previous research showed, PWS engage different timing mechanisms and/or brain circuits to time movements with an external cue [10,21].
Furthermore, it was hypothesized that timing would become more variable in PWS with increasing task complexity, however, this hypothesis was not supported by the results of our study. Despite PWS being more variable than PWNS in general, PWNS have a greater increase in timing variability in the combined condition.
From our study it can be concluded that PWS potentially couple auditory, manual, and articulatory rhythms in a different way, leading to later speech initiation and more temporal variation. This remains to be tested with a greater participant sample, of course. We aim to present data from 10 participants per group at the conference. Finally, our dataset offers the possibility for specific consideration of the vowel gesture, of inter-gestural timing between onset consonants and vowels, as well as on intra-gesture stability in different rhythmic conditions.

5. ACKNOWLEDGEMENTS

This work was supported by the DFG grants [FA 901/4-1 and HO 3271/6-1], as well as the Graduate School "Class of Language" of LMU, the Bavarian Research Alliance, and the German Academic Research Exchange Service. We would also like to thank Charlie Wiltshire for helping with data collection, Henry Derwanz for helping with data segmentation, and all our participants for participating.

6. REFERENCES

[1] Yairi, E., Ambrose, N. G. 2013. Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, 38, 66–87.

[2] Bloodstein, O., Bernstein Ratner, N. 2008. *A handbook on stuttering*, 6th ed. Clifton Park, NY: Thompson/Delmar.

[3] Etchell, A. C., Jonson, B. W., Sowman, P. F. 2014. Behavioral and multimodal neuroimaging evidence for a deficit in brain timing networks in stuttering: a hypothesis and theory. *Frontiers in Human Neuroscience*, 8, 467.

[4] Bradshaw, A. R., Lametti, D. R., McGettigan, C. 2021. The Role of Sensory Feedback in Developmental Stuttering: A Review. *Neurobiology of Language*, 2(2), 308–334. doi: https://doi.org/10.1162/nol_a_00036

[5] Chang, S. E., Guenther, F. H. 2020. Involvement of the Cortico-Basal Ganglia-Thalamocortical Loop in Developmental Stuttering. *Frontiers in psychology*, 10, 3088. https://doi.org/10.3389/fpsyg.2019.03088

[6] Civier, O., Bullock, D., Max, L., Guenther, F. H. 2013. Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain and Language*, 126(3), 263–278. https://doi.org/10.1016/j.bandl.2013.05.016

[7] Sares, A. G., Deroche, M. L. D., Shiller, D. M., Gracco, V. L. 2019. Adults who stutter and metronome synchronization: evidence for a nonspeech timing deficit. *Annals of the New York Academy of Sciences*, 1149, 56–69. doi: 10.1111/nyas.14117

[8] Falk, S., Müller, T., Dalla Bella, S. 2015. Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Frontiers in Psychology*, 6, 847. doi: 10.3389/fpsyg.2015.00847

[9] Andrews, G., Howie, P., Dozsa, M., Guitar, B. 1982. Stuttering: Speech pattern characteristics under fluency-inducing conditions. *Journal of Speech, Language and Hearing Research*, 25, 208–216.

[10] Frankford, S. A., Heller Murray, E. S. , Masapollo, M., Cai, S., Tourville, J. A., Nieto-Castañón, A., Guenther, F. H. 2021. The Neural Circuitry Underlying the "Rhythm Effect" in Stuttering. *Journal of Speech, Language and Hearing Research*. 18;64(6S), 2325–2346. doi: 10.1044/2021_JSLHR-20-00328. Epub 2021 Apr 22. PMID: 33887150; PMCID: PMC8740675

[11] Parrell, B., Goldstein, L., Lee, S., Byrd, D. 2014. Spatiotemporal coupling between speech and manual motor actions. *Journal of phonetics*, 42, 1–11. https://doi.org/10.1016/j.wocn.2013.11.002

[12] Hulstijn, W., Summers, J. J., van Lieshout, P. H. M, Peters, H. F. M. 1992. Timing in finger tapping and speech: A comparison between stutterers and fluent speakers. *Human Movement Science*, 11(1–2), 113–124. https://doi.org/10.1016/0167-9457(92)90054-F

[13] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.

[14] R Core Team 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[15] Winter, B. 2020. *Statistics for linguistics: An introduction using R*. New York: Routledge

[16] Lenth, R. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.2-1.https://CRAN.R-project.org/package=emmeans

[17] Pouw, W., Dixon, J. A. 2019. Entrainment and Modulation of Gesture-Speech Synchrony Under Delayed Auditory Feedback. *Cognitive science*, *43*(3), e12721. https://doi.org/10.1111/cogs.12721

[18] Marcus, S. M. 1981. Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, 30(3), 247–256.

[19] De Jong, K. (1992) Acoustic and Articulatory Correlates of P-Center Perception, *UCLA Working Papers in Linguistics*, 81, 66–75.

[20] Schreier, R., Dalla Bella, S., Hoole, P., Falk, S. 2020. Verbal timing deficits in stuttering. https://issp2020.yale.edu/S04/schreier_04_11_083_abstract.pdf. Last viewed: January 5, 2023)

[21] Etchell, A. C., Johnson, B. W., Sowman, P. F. 2015. Beta oscillations, timing, and stuttering. *Frontiers in Human Neuroscience*, 8, 1036. doi: 10.3389/fnhum.2014.01036

Nonword repetition in children with cochlear implants using different speech and language rehabilitation approaches

Lucie Van Bogaert[1,2], Laura Machart[1,2], Hélène Lœvenbruck[1] and Anne Vilain[2]

(1) Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France
(2) Univ. Grenoble Alpes, Grenoble INP, CNRS, GIPSA-lab, 38000 Grenoble, France
Lucie.van-bogaert@univ-grenoble-alpes.fr

ABSTRACT

When a child is diagnosed with hearing impairment, early intervention should be provided to avoid language deprivation and its consequences. This study investigates the contribution of two spoken language rehabilitation approaches to speech development in children with cochlear implants (CI) using a nonword repetition task. Cued Speech (CS) is a multisensory communication tool, facilitating speech perception by providing access to all phonemes. Auditory Verbal Therapy (AVT) is a hearing-focused method, training auditory skills to boost speech perception. Earlier studies have reported that CS and AVT improve phonological skills in a picture-naming task. In this study, using a nonword repetition task, we show that the number of consonant and vowel errors is higher in children with CI than in typically-hearing peers, independently of the rehabilitation method. Therefore, children with CI may have hidden speech processing difficulties that remain undetected in lexical speech tasks but can be revealed by nonword repetition.

Keywords: Nonword Repetition, Children, Cochlear Implants, Cued Speech, Auditory Verbal Therapy

1. INTRODUCTION

Hearing loss can have impacts on many aspects of language, communication and cognition. Therefore, it represents a major social issue. The developmental trajectories of deaf children are varied and different communication modalities can be adopted. Some families choose to use sign language, which is fully accessible to deaf people in terms of perception and production, while others prefer to focus on spoken language communication, especially when it is the parents' mother tongue (95% of deaf children have at least one hearing parent, cf. [1]). A large proportion of French-speaking children with profound hearing loss who receive oral education are usually fitted with a cochlear implant (CI), a widely used device which partially restores access to speech sounds. However, the perception of some acoustic features may still be altered, which results in limited phonological skills [2–4] and may impact spoken language development [5]. Other repercussions have also been found in the speech production of French-speaking children with cochlear implants, especially on voicing contrast [6], and place of articulation ([2], [4] , [6]).

To compensate for the acoustic deficits associated with cochlear implants, different spoken language rehabilitation approaches can be used.

Cued speech [7] is one of the communication tools used by professionals and parents in France. It is a multisensory approach, as manual gestures are added to audio-visual speech communication. Each phoneme of a given language is associated with a manual cue produced near the speaker's face, simultaneously with speech production. Many studies have shown that cued speech use improves speech perception in children with hearing impairment, when audiovisual speech is presented together with manual cues [8], and even visual speech perception without auditory information [9]. More importantly, longer-term effects have also been demonstrated. For children educated with cued speech, improvements in speech perception have been observed in the visual-only, audio-visual and even audio-only (i.e. speech presented *without* simultaneous manual gestures) modalities [10], [11]. Several studies also suggest that cued speech exposure supports the development of more accurate phonological representations and improves phonological awareness, which promotes phonological memory, reading and spelling skills [12], [13]. More recently, using a picture-naming task, Machart et al. (2022) have shown that cued speech proficiency improves speech production in children with cochlear implants, especially with regard to voicing, nasality, manner and place of articulation [14].

On the other hand, Auditory Verbal Therapy (AVT), a hearing-focused approach is beginning to be used in France, mainly by parents who learn about through social network. Indeed, a recent study [15] shows that very few professionals in France are trained for this method and that more than half of the parents using this method in France became aware of it through social networks. AVT is a listening and spoken language early intervention program for deaf children, that focuses on the training of auditory

skills. It is a family-centred coaching program where parents are encouraged to use auditory verbal strategies in everyday life in order to stimulate their child's auditory skills. Each AVT session includes exercises aimed to promote the child's listening skills, and assessments are systematically conducted during these activities to adjust the objectives and provide parents with continuous feedback on the child's progress. Furthermore, during the sessions, four auditory exercises are usually conducted, which include sound detection, discrimination, identification, and comprehension. Although AVT is widely used and government-funded in several countries, scientific evidence on the contribution of AVT on speech and language development is still scarce. In a systematic review, Binos et al. (2021) [16] examine eight papers published in the last ten years, and they conclude that "although recent evidence from the published papers that were reviewed revealed the positive role of AVT with regard to the speech and language skills of CI children, it is still difficult to generalize", because of "lack of well-controlled studies" (absence of a control group, bias in participants' ages, etc). Some results are also contradictory, with e.g. Percy-Smith et al. 2017 [17] evidencing higher performance for children with AVT than children with "standard" Danish speech therapy, whereas Yanbay et al (2014) found no significant difference between groups of children with cochlear implants enrolled in AVT or other programs (oral communication or bimodal bilingual) [18].

The aim of the present study was to measure the influence of Cued French (French version of cued speech) and AVT on speech perception and production in children with cochlear implants, using a nonword repetition task. Indeed, nonword repetition has been shown to be an effective task to distinguish typical from atypical language development in children [19], a potential predictor of phonological, syntactic and lexical difficulties in deaf children [20], [21]. Lastly, it can also be used to evaluate the effectiveness of a communication mode [22]. Nonword repetition can be challenging as it involves many operations, including auditory perception, as well as verbal working memory, parsing of speech into phonological units and speech motor planning and programming.

2. METHODS

The data presented here are taken from a larger set of experiments designed to assess speech perception and production in children with typical development and children with speech sound disorders (the EULALIES project, [19]). Only the results of the nonword repetition task in children with cochlear implants (CI) and a group of children with typical hearing are reported here. Accuracy of consonant and vowel production was examined. The children with CI were divided into three groups, based on their modes of communication (AVT, proficient Cued French, no/little Cued French).

2.1. Procedure

This study used the nonword repetition task from the EULALIES battery, which includes all French phonemes in 16 nonwords. Each nonword consists of two to four syllables with various degrees of complexity (with or without clusters) and some nonwords include a morpheme or a lexical unit. The children sat in front of a computer screen and watched a video of a speaker producing the nonwords (sound + face of the speaker) and were asked to repeat them. The nonwords were presented as the names of cute monsters displayed on the screen. The children wore a SHURE headset microphone (Beta 54R) and a backup microphone was also placed on the table (Zoom recorder H4n Pro). Four inclusion tasks were also administered: a visual span (PathSpan, [23]), a digit span (ODEDYS, [24]), a language production task (morphosyntax task from ELO, [25]), as well as an auditory screening test for children with typical hearing, to exclude any hearing disorder (perception at 20 dB on the frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz, using an Electronica 9910 audiometer). Children with typical hearing were excluded from the study if their scores were below the norm.

A fifth task was used for children with CI exposed to cued speech to measure their level of cued speech proficiency, using the TERMO test (Evaluation test for the reception of the oral message by the deaf child; [26]). The experimenter presented a list of words and sentences using Cued French gestures without phonation in order to measure the child's cue reading skills without the auditory modality. The child was asked to repeat the items vocally. Only lexical accuracy was assessed (the phonological aspect of their production was not considered). Two levels of cued speech proficiency were determined: low Cued French reading skills (CF-) and high Cued French reading skills (CF+). Children in the CF- group could at most understand a few familiar words (or none) at slow speech rate. Children in the CF+ group could understand familiar words and simple sentences at natural speech rate.

Parents or caregivers provided informed consent and filled a questionnaire on their child's language development and background (multilingualism, age

of first fitting, communication tools used at home, information about the child's deafness). Data were collected anonymously, according to GDPR regulations. The study received ethical approval from the local ethics committee (CERGA-Avis-2022-23).

2.2. Participants

This study includes 119 children (63 girls and 56 boys) aged between 66 and 140 months, with 83 children with typical hearing skills (TH group) and 36 children with cochlear implants (CI group). Only monolingual children were included in this study, i.e. children who were not exposed to a language other than French on a daily basis. The CI group was divided into three subgroups, based on their modes of communication (AVT, CF+: proficient Cued French, CF-: no/little Cued French). Age characteristics are given in Table 1.

	N	Age	Age at implantation
CF+	6	77-132 *(101.17; 18.55)*	12-85 *(28.67; 25.53)*
CF-	21	65-139 *(95.71; 22.09)*	11-90 *(28.86; 21.13)*
AVT	9	66-114 *(81.67; 16.68)*	11- 48 *(20.11; 10.64)*
TH	83	61-131 *(84.67; 15.57)*	

Table 1: Chronological age and age at implantation in months for the 4 groups of participants (minimum-maximum *(mean; sd)*).

The children exposed to Cued French (CF+ and CF-) were recruited during the ALPC (Association Nationale pour la Langue Parlée Complétée) cued speech summer camp or through the Grenoble University Hospital. The children from the AVT group were recruited through the association ADEFAV (Association des Familles AVTistes) which promotes rehabilitation through hearing training. The 83 TH children (35 girls and 48 boys) belong to the large cohort of typical children of the EULALIES project [23]. All TH children were recruited from local schools.

2.3. Data processing and statistical analyses

The nonword repetition task was transcribed using a narrow phonetic transcription and processed with the PHON software [27]. All recordings were double-blindly annotated by two different transcribers who had no knowledge of the child's group membership or of the other person's transcription. A third transcriber then validated the transcriptions (in case of discrepancy, a consensus was reached), syllabified the items and aligned the production with the target.

The number of errors per nonword was extracted with PHON (total number of substitutions, deletions and epenthesis).

Statistical analysis and graphical representations were done using the R software [28]. Generalized Linear Mixed-Effects models with Poisson regression (glmer function) were used to analyze the number of errors on consonant and vowel production separately, in the four groups of children. A backward stepwise variable selection by model comparison was applied to identify the best-fitting model (anova function). Random-effect factors and fixed-effect factors of interest were included, with interactions. Random-effects factors were Participant and Item. Fixed-effect factors were Group (CF+, CF-, AVT and TH) and Chronological_Age (in months). Multiple comparisons were then performed with the glht function of the multcomp package in R. Factors which did not improve model fit were excluded.

3. RESULTS

3.1. Consonant errors

Figure 1 shows the mean number of consonant errors per nonword for each group (CF+, CF-, AVT, TH).

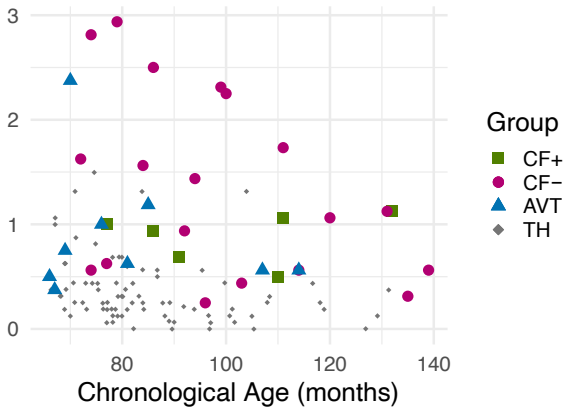


FIGURE 1: Mean number of consonant errors per nonword in the four groups of children, according to chronological age in months.

The best-fitted model for the analysis of consonant errors included *Group* and *Chronological Age*, without any interaction. Multiple comparisons showed that all groups of children with CI (i.e., CF+, CF- and AVT) have higher numbers of consonant errors per nonword than TH children (resp. $Z = 4.371$, $p < .01$; $Z = 8.307$, $p < .001$; $Z = 3.717$, $p < .001$). No difference was found between the groups of children with CI.

3.1. Vowel errors

Figure 2 shows the number of vowel errors per nonword for each group (CF+, CF-, AVT, TH).
The results for vowel errors per nonword did not differ from those of consonant errors. Indeed, the best-fitted model also included *Group* and *Chronological Age*, without interaction. Multiple comparisons also revealed that all groups of children with CI (i.e., CF+, CF- and AVT) have higher vowel error rates than TH children (resp. $Z = 5.421$, $p < 0.01$; $Z = 8.794$, $p < 0.001$; $Z = 5.341$, $p < 0.001$).

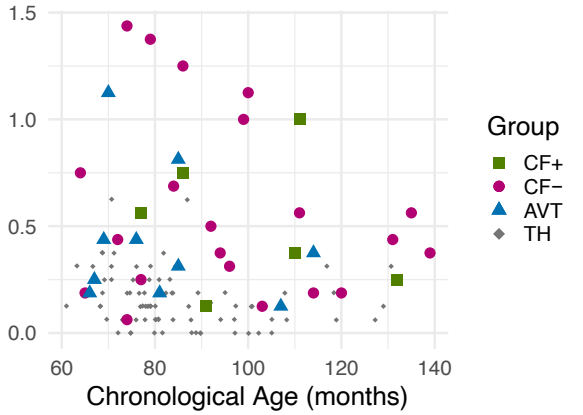


FIGURE 2: Mean number of vowel errors per nonword in the four groups of children, according to chronological age in months.

4. DISCUSSION

The aim of this study was to investigate the influence of two spoken language rehabilitation approaches on the nonword repetition skills of deaf children with CI. The results show that deaf children with cochlear implants do not achieve similar nonword repetition performance to their typical hearing peers, regardless of the spoken language rehabilitation approach. A significantly higher number of consonant and vowel production errors was found in all groups of children with CI (i.e., CF+, CF- and AVT) compared to children with typical hearing. These results differ from those of Machart et al. (2022) on a picture-naming task, who found that children with high Cued French proficiency had similar speech production performance to typical hearing peers, contrary to children with low cued speech proficiency [14]. It seems, therefore, that high Cued French proficiency allows more accurate speech production in children with CI for familiar words but not for unknown words.
Moreover, another study was carried out to assess speech perception using a lexicality judgement task, in the same groups of participants as the present study. The results showed that children in the AVT group performed closer to the TH group than the children in the CF+ group, who in turn outperformed children in the CF- group [29].
In sum, other studies have reported benefits of auditory-verbal therapy and cued speech proficiency for picture naming skills, i.e. for the production of well-known words, as well as for lexicality judgment, i.e., for the detection of phonological distortion on well-known words. Such a positive impact was not found in the present study, involving nonword repetition.
The nonword repetition task requires different skills than word perception and production. Phonological decoding, verbal working memory as well as generating unlearnt motor plans and programs are necessary additional levels of processing. Lexical support (use of semantic information to overcome phonological decoding difficulties) cannot be used in this task which involves unknown items. The nonword repetition task is a good way to assess these skills independently of stored lexical knowledge.
The results of the present study illustrate how some speech processing difficulties in children with hearing loss, which remain undetected when only lexical tasks are considered, can be revealed by a nonword repetition task. Identifying these difficulties is a crucial issue, as they can lead to cognitive fatigue and social insecurity. We argue that non-lexical tasks should be systematically used when evaluating the efficacy of a speech therapy method.
The nonwords used in this task included all French phonemes and consisted of sequences with various length, and syllable complexity. The specific features that hamper the performance of children with hearing impairment in this task will need to be further explored. These will be confronted with data from the word production and word perception tasks, in order to weigh the relative role of perceptual/motor difficulties.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie, Grant Agreement No 860755.

5. REFERENCES

[1] R. Mitchell and M. Karchmer, "Chasing the Mythical Ten Percent: Parental Hearing Status of Deaf and Hard of Hearing Students in the United States," *Sign Lang. Stud.*, vol. 4, pp. 138–163, Dec. 2004;
[2] S. Bouton, W. Serniclaes, P. Colé, and J. Bertoncini, "Categorical perception of speech sounds in French-speaking children with cochlear implant," Jan. 2012.
[3] S. Colin, J. Ecalle, E. Truy, G. Lina-Granade, and A. Magnan, "Effect of age at cochlear implantation

and at exposure to Cued Speech on literacy skills in deaf children," *Res. Dev. Disabil.*, vol. 71, pp. 61–69, Dec. 2017;
[4] J. Leybaert, C. Bayard, C. Colin, and C. LaSasso, "Cued speech and cochlear implants: A powerful combination for natural spoken language acquisition and the development of reading," *Oxf. Handb. Deaf Stud. Lang.*, pp. 359–376, 2016.
[5] V. L. Hanson and C. A. Fowler, "Phonological coding in word reading: Evidence from hearing and deaf readers," *Mem. Cognit.*, vol. 15, no. 3, pp. 199–207, May 1987.
[6] B. Grandon, "Développement typique et atypique de la production de parole : caractéristiques segmentales et intelligibilité de la parole d'enfants porteurs d'un implant cochléaire et d'enfants normo-entendants de 5 à 11 ans," phdthesis, Université Grenoble Alpes, 2016. Accessed: Dec. 13, 2022.
[7] R. O. Cornett, "CUED SPEECH," *Am. Ann. Deaf*, vol. 112, no. 1, pp. 3–13, 1967.
[8] O. Perrier, B. Charlier, C. Hage, and J. Alegria, "Evaluation of the Effects of Prolonged Cued Speech Practice Upon the Reception of Spoken Language1," Jan. 1990, Accessed: Nov. 29, 2022.
[9] G. H. Nicholls and D. Ling, "Cued Speech and the reception of spoken language," *J. Speech Hear. Res.*, vol. 25, no. 2, pp. 262–269, Jun. 1982.
[10] M. Aparicio, P. Peigneux, B. Charlier, C. Neyrat, and J. Leybaert, "Early experience of Cued Speech enhances speechreading performance in deaf," *Scand. J. Psychol.*, vol. 53, no. 1, pp. 41–46, 2012.
[11] M.-I. Kos, M. Deriaz, J.-P. Guyot, and M. Pelizzone, "What can be expected from a late cochlear implantation?," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 73, no. 2, pp. 189–193, Feb. 2009.
[12] B. J. Trezek, "Cued Speech and the Development of Reading in English: Examining the Evidence," *J. Deaf Stud. Deaf Educ.*, vol. 22, no. 4, pp. 349–364, Oct. 2017.
[13] B. Charlier and J. Leybaert, "The rhyming skills of deaf children educated with phonetically augmented speechreading," *Q. J. Exp. Psychol. Sect. A*, vol. 53, no. 2, pp. 349–375, May 2000.
[14] L. Machart, "Production de parole chez l'enfant sourd : bénéfices de l'exposition à la Langue française Parlée Complétée associée à l'implantation cochléaire," Université Grenoble Alpes, 2022.
[15] L. Van Bogaert, H. Loevenbruck, and A. Vilain, "Les outils d'aide à la communication utilisés avec les enfants sourds en France: étude exploratoire," presented at the Séminaire AFCP – Phonétique Clinique, Toulouse, Toulouse, 2021.
[16] P. Binos, E. Nirgianaki, and G. Psillas, "How Effective Is Auditory-Verbal Therapy (AVT) for Building Language Development of Children with Cochlear Implants? A Systematic Review," *Life Basel Switz.*, vol. 11, no. 3, p. 239, Mar. 2021.
[17] L. Percy-Smith *et al.*, "Auditory verbal habilitation is associated with improved outcome for children with cochlear implant," *Cochlear Implants Int.*, vol. 19, no. 1, pp. 38–45, Jan. 2018.
[18] E. Yanbay, L. Hickson, N. Scarinci, G. Constantinescu, and S. J. Dettman, "Language outcomes for children with cochlear implants enrolled in different communication programs," *Cochlear Implants Int.*, vol. 15, no. 3, pp. 121–135, May 2014.
[19] G. Meloni, H. Loevenbruck, and A. Vilain, "EULALIES, The France-Québec Speech Sound Disorders project," presented at the 14th International Congress for the Study of Child Language (IASCL), Lyon, France, Lyon, France, 2017.
[20] S. Nittrouer, M. Muir, K. Tietgens, A. C. Moberly, and J. H. Lowenstein, "Development of Phonological, Lexical, and Syntactic Abilities in Children With Cochlear Implants Across the Elementary Grades," *J. Speech Lang. Hear. Res.*, vol. 61, no. 10, pp. 2561–2577, Oct. 2018.
[21] S. Sundström, U. Löfkvist, B. Lyxell, and C. Samuelsson, "Prosodic and segmental aspects of nonword repetition in 4- to 6-year-old children who are deaf and hard of hearing compared to controls with normal hearing," *Clin. Linguist. Phon.*, vol. 32, no. 10, pp. 950–971, Oct. 2018.
[22] C. Dillon and D. Pisoni, "Non word Repetition and Reading Skills in Children Who Are Deaf and Have Cochlear Implants," *Volta Rev.*, vol. 106, pp. 121–145, Sep. 2006.
[23] J.-A. LeFevre *et al.*, "Pathways to Mathematics: Longitudinal Predictors of Performance," *Child Dev.*, vol. 81, no. 6, pp. 1753–1767, 2010.
[24] M. Jacquier-Roux, S. Valdois, M. Zorman, C. Lequette, and G. Pouget, "Outil de DÉpistage des DYSlexies - ODEDYS." 2005.
[25] A. Khomsi, *ELO : évaluation du langage oral manuel*, 1 vols. Paris: ECPA, les Éd. du Centre de psychologie appliquée, 2001.
[26] C. Descourtieux, "Annexe 1. Le TERMO : Test d'Évaluation de la Réception du Message Oral par l'enfant sourd," in *Compétences cognitives, linguistiques et sociales de l'enfant sourd*, Wavre: Mardaga, 2006, pp. 267–269. Accessed: Aug. 10, 2022.
[27] G. Hedlund and Y. Rose, "Phon 3.0 [Computer Software]," *TalkBank Syst. Phon Proj. Available Online Httpsphon Ca Accessed April 23 2021*, 2019.
[28] R. C. Team, "R: A language and environment for statistical computing," 2013.
[29] L. Van Bogaert, L. Machart, S. Gerber, H. Loevenbruck, and A. Vilain, "Contributions of French Cued Speech and Auditory Verbal Therapy on speech perception in deaf children with cochlear implants," in prep.

UNSUPERVISED CLUSTERING REVEALS SEVERAL SUBTYPES IN SPEAKERS WITH ATAXIA

Anneke Slis[1], Robin Karlin[1], Benjamin Parrell[1,2]

[1]Waisman Center, University of Wisconsin–Madison, [2]Communication Sciences and Disorders, University of Wisconsin–Madison
slis@wisc.edu, rkarlin@wisc.edu, bparrell@wisc.edu

ABSTRACT

While ataxic dysarthria is traditionally considered a single diagnostic group, recent work suggests that distinct subtypes exist within this population. Here, we explore whether an unsupervised cluster analysis of diadochokinetic speech can detect the presence of potential subgroups, without a priori knowledge about individual speakers. Thirty-eight individuals with degenerative spinocerebellar ataxia produced [oj] repetitively. Fifteen acoustic markers were extracted from each production, consisting of temporally changing energy patterns that capture durational patterns and vocal tract shape changes. Principal component analysis yielded three dimensions (capturing 86% of the total variance), which were used as input for clustering analysis. Across clustering algorithms and evaluation criteria, 2-3 clusters were consistently found. Although it remains to be tested how these clusters align with clinical diagnoses, these results suggest that distinct subgroups of speech impairments do exist within individuals with ataxia, and that these subgroups may be automatically detectable from a simple speech task.

Keywords: ataxic dysarthria, clustering, subtypes

1. INTRODUCTION

Ataxic dysarthria is a disorder of the speech motor system, caused by a range of conditions involving the cerebellum [1-3]. Speech is characterized by distorted phonation (e.g., [3]), impaired articulatory control, and deficits in temporal coordination [2]. Articulatory difficulties surface as imprecise consonants as well as distorted vowel productions, while temporal impairments include elongated speech segments, reduced speaking rate, scanning speech (overly isochronous syllable durations), and irregular alternating motion rates [3, 4].

Although ataxic dysarthria is a single clinical diagnosis, a growing body of evidence suggests that distinct subgroups exist within this population [4-7]. Two distinct patterns within ataxic dysarthric speech have been observed – inflexible and instable speech productions [5-7, 16]. These classifications cover both the temporal and articulatory domain. The subgroup with instability presents with increased variability in both articulation (e.g., variably imprecise consonants) and temporal control (e.g., variable stress and rhythm). Conversely, the group with inflexibility displays extremely persistent speech patterns, including consistent articulatory errors, scanning speech, and isochronous durations. In a series of studies [5-7], the two categories have been used to perceptually identify subgroups within ataxic dysarthria. For example, out of 10 speakers in [5], five were classified with instability, one with inflexibility, and four with impairments combining features of both instability and inflexibility. Some evidence suggests that these subtypes align with specific causes of cerebellar damage [4]. Regularity of diadochokinetic speech has been shown, for example, to be impaired in patients, diagnosed with spinocerebellar ataxia type SCA3, whereas rate and prosodic modulation were affected in type SCA6 [4].

To date, identifying potential subgroups within ataxic dysarthria has relied largely on perceptual ratings by experienced, clinically trained speech language pathologists. Despite extensive experience of clinicians, however, identifying perceptual subgroups shows only moderate inter-rater reliability [5]. Conversely, a recent study using naïve listeners to classify dysarthria subtypes without a priori knowledge of the speech characteristics of ataxic dysarthria showed higher reliability [7] but depended on a large number of listeners (>20).

Previous work examining potential subgroups within ataxic dysarthria has relied principally on perceptual characterizations, rather than objective acoustic measures. To validate whether these perceptually identified subgroups exist, an automated tool that provides empirical and reliable identification of subgroups within ataxic dysarthria would be beneficial. Successful automated classification of ataxic dysarthria subtypes not only benefits clinical diagnosis but also lends theoretical support to the existence of these subtypes themselves.

As a step towards such an automated tool, the current study attempts to identify subgroups within speakers diagnosed with degenerative spinocerebellar ataxia using unsupervised clustering. Although a similar approach, using linear discriminant analysis,

has previously been shown to identify clusters within this population [4], this study relied only on temporal and amplitude variation in the speech signal, omitting the clinically critical aspect of articulatory precision. Here, we include this factor in an unsupervised clustering analysis by using acoustically derived measures, quantifying changes in spectral energy over time [8-9]. Importantly, these energy changes correlate well with articulatory trajectories [8] and have recently been shown to be effective in separating hypokinetic, spastic, and flaccid dysarthria [9, 10]. A particularly beneficial characteristic is that this method separates supra-laryngeal articulation from the phonatory source. Consequently, it is possible to examine laryngeal as well as supra-laryngeal contributions to ataxic dysarthria subtypes. Our results show that unsupervised clustering based on these measures, extracted from diadochokinetic speech, consistently identifies subgroups within speakers with cerebellar degeneration.

2. METHODS

2.1. Participants

38 participants (14 M, mean age: 60, range: 31-81; 24 F, mean age: 60, range 23-77) participated in the study. Individuals were recruited via CoRDS [14] and were previously diagnosed with degenerative spinocerebellar ataxia (three SCA1; nine SCA2; ten SCA3; twelve SCA6; four SCA8). All speakers were native speakers of American English and reported no additional history of hearing, speech, or language disorders. The participants were compensated monetarily for participating in the study. The study was approved by the Institutional Review Board at the University of Wisconsin – Madison.

2.2 Materials and equipment

Diadochokinetic speech was collected during an assessment of dysarthria, including oral motor skill evaluation. For each task, participants were instructed to repeat a particular syllable as quickly and accurately as possible. In the current study, we examined the repetitive production of the syllable [oj]. This sequence targets anterior-posterior tongue motions, potentially revealing possible articulatory deficiencies in lingual control. The speech material was collected online via Zoom. The participants used their home computer and equipment (see 4. DISCUSSION about possible limitations). They were sitting in front of their computer at an individually preferred distance from the microphone, in the same fashion as during Zoom meetings. The recordings were sampled at 32 kHz (bit rate: 126 kbps) in a m4a format. All data processing, except where noted, was conducted in MATLAB [17].

2.3 Acoustic parameter extraction

The procedure has been described in detail previously [8, 9]. A summary follows here. For each participant, the .m4a sound files were converted to .wav files (32 Khz, 16-bit PCM) in Audacity [18]. The mean peak intensity of the extracted sounds was 70.5 dB, with a low standard deviation across participants (1.92 dB) indicating a relatively consistent intensity level across the recording sessions. Inspecting the spectrum and the closest zero crossings, the start and end of the target sequences were labelled in PRAAT [11]. Only sequences that were produced without an audible breath or a large pause were selected. In the case of multiple good productions, the sequences with the most repetitions of the target syllable were selected. The resulting annotated speech signals were imported in MATLAB, resampled to 16 kHz, and pre-emphasized to account for the high-frequency roll-off of the glottal source. Speech spectra were generated using 25 ms windows with 2 ms steps between samples. Frequencies between 300 and 4000 Hz were considered for Mel-frequency cepstral coefficients (MFCC) analysis. 20 Mel-frequency filter banks processed these spectra to emphasize the characteristics of the human ear [12]. 15 MFCCs were then extracted by log-transforming the resulting spectral values and applying a discrete cosine transformation.

The first coefficient reflects the average spectral energy, containing source information, and was analysed to quantify speech amplitude changes. To capture articulatory changes over time, the first differential of each of the remaining MFCCs (2-15) was computed, and these values were summed at each time point. This resulted in a trajectory with “spectral energy difference” values with peaks indicating large changes in energy (e.g., changes from [o] to [j] and [j] to [o]) and valleys indicating small changes in energy (e.g., during the steady states of vocoids; see Figure 1). These changes in spectral energy are highly correlated with vocal tract shape changes, which indirectly represent articulatory movement patterns [8]. Consequently, the magnitude of spectral energy changes indicates the speed of articulatory motions (i.e., from [o] to [j] and vice versa) and variability in these changes relates to consistency in articulatory movements. Temporal distances between maxima and minima reflect articulatory timing patterns, from which timing variability can be extracted. In total, 15 measures were extracted from the trajectories (Table 1 and Figure 1 and 2).

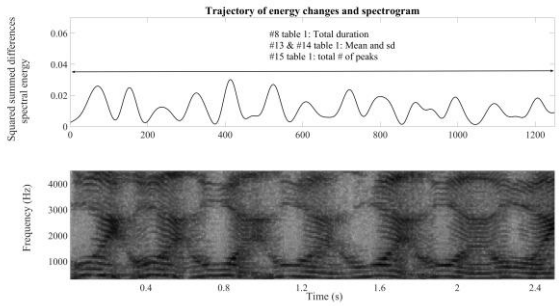


Figure 1: Diadochokinetic speech sample. Top: the sum of the first differential of MFCC 2-15 over time, reflecting a global metric of spectral change related to supra-laryngeal articulation. Bottom: time-aligned speech spectrogram.

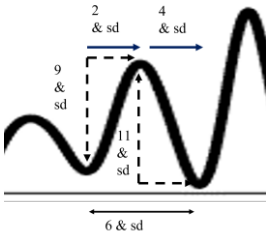


Figure 2: Temporal and articulatory parameters extracted from the spectral energy difference signal in Figure 1 (appr. 200 to 420 ms). See Table 1 for details.

2.4 Data processing and clustering algorithms

To determine the relevant principal components for the cluster analysis, the matrix containing the raw acoustic parameter values for each participant was normalized by centering and z-scoring each parameter. Resulting values for the 15 extracted parameters were transformed into three principal components (PCs), using the PCA algorithm in MATLAB (see 3. RESULTS). The output loadings were further rotated with ROTATEFACTORS on these three PCs, with a standard varimax rotation. The three PCs served as input for the unsupervised clustering process.

To assess the reliability of our results, two separate clustering algorithms were used: k-Means clustering, which assumes equal variances across dimensions and groups (KMEANS in MATLAB) and the Expectation-Maximization (EM) algorithm for Gaussian Mixture models, allowing variances to differ [15]. Potential clusters were generated using each algorithm for one to seven groups (inclusive). The optimal number of clusters was then determined using the Calinski-Harabasz criterion [13] with EVALCLUSTERS. Because unsupervised clustering often identifies local, rather than global minima, and thus has the potential to yield different results across iterations depending on initial conditions, this process was repeated 3000 times for each clustering algorithm.

	Amplitude descriptor
1	Amplitude coefficient of variation of the first MFCC (the standard deviation divided by the mean).
	Temporal descriptors
2	Duration from valley (i.e., steady state) to following peak (transition from [o] to [j] and vice versa): large values indicate slower vocal shape changes.
3	The standard deviation of these durations.
4	Duration from peaks to following valleys.
5	The standard deviation of these durations.
6	Duration from valleys to valleys.
7	The standard deviation of these durations.
8	The total sequence duration.
	Articulatory descriptors
9	The magnitude of the difference between valleys and following peaks (analogous to the speed of vocal tract shape changes from steady state (valley; no vocal tract shape change) to peak (fast vocal tract shape change from [o] to [j]).
10	Standard deviation of these differences.
11	The magnitude of the difference between peaks and following valleys (analogous to the rate of vocal tract shape (peak) compared to steady state (valley).
12	The standard deviation of these differences.
13	The mean of all the values of spectral energy changes across the full sequence.
14	The standard deviation of these magnitudes
15	Number of peak-to-peak cycles in sequence.

Table 1: Temporal and articulatory parameters extracted from the spectral energy difference signal.

3. RESULTS

Observing the eigenvalues of the PCA, only three factors were retained, as after the third dimension the curve tapered off. The three dimensions together explained the 86% of the variance: 40%, 32%, and 14% for the first, second and third PCs respectively. The first PC weighted heavily towards the articulatory descriptors, the second PC towards the temporal descriptors, and the third towards the total sequence duration, number of syllable repetitions, and variation in speech amplitude (see Table 2). Using these PCs, the EM clustering analysis identified two to three clusters (46% and 37% of the iterations respectively), while the k-Means analysis more consistently identified three clusters (76% of the iterations).

		PC 1	PC 2	PC 3
1	COV first MFCC	-0.11	0.09	**0.25**
	Temporal descriptors			
2	M valley-peak	0.08	**0.41**	-0.10
3	SD #2	-0.03	**0.37**	0.16
4	M peak-valley	0.04	**0.42**	-0.10
5	SD #4	-0.08	**0.35**	0.20
6	M valley-valley	0.05	**0.44**	-0.11
7	SD #6	-0.05	**0.40**	0.13
8	Total duration	0.09	-0.01	**0.62**
	Articulatory descriptors			
9	M valley to peak	**0.41**	0.01	-0.09
10	SD #9	**0.39**	0.05	0.08
11	M peak to valley	**0.41**	0.01	-0.09
12	SD #11	**0.39**	0.06	0.06
13	Mean contour	**0.38**	-0.11	-0.04
14	SD #13	**0.41**	-0.01	0.03
15	# peaks	0.06	-0.11	**0.64**

Table 2. Columns show loadings for each PC after Varimax rotation. Loadings with values equal or higher than 0.25 are shown in bold.

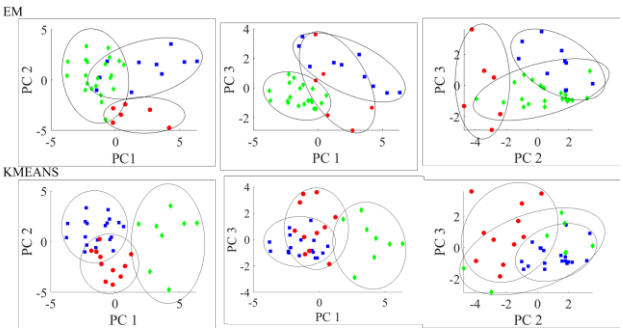


Figure 3. Clusters of speakers as grouped by EM (upper) and k-Means (lower) algorithms.

Several patterns can be observed (see Figure 3). For the k-Means algorithm, first and second PCs separated all three clusters efficiently; the first PC distinguishes between one group of two clusters (round (red) and squared (blue) dots) and a third group (diamond (green) dots), while the second and PC distinguishes between the red and blue groups. For the EM algorithm, the three clusters are less clearly distinguished by any single pair of PCs.

4. DISCUSSION

This study explored whether an unsupervised cluster analysis of diodochokinetic speech detected potential subgroups of dysarthria in speakers with spinocerebellar ataxia, without a priori knowledge about clinical diagnoses. The study showed that k-Means and, to a lesser extent EM, identified distinct clusters of speakers. Subgroups were distinguished along dimensions consistent with the first three PCs, representing articulatory features, temporal features, and utterance length/number of peaks/amplitude variation, respectively. These results are consistent with the notion that distinct subtypes of ataxic dysarthria exist [4-7] and can be distinguished based on unique clusters of symptoms. Observing the PC loading factors suggests that including supra-laryngeal parameters helps to distinguish these groups, in addition to the temporal and phonatory features employed in [4]. The emergence of three clusters is promising – an average of 3.9 clusters were recognized perceptually for diadochokinetic speech by naïve listeners in [7]; subsequent hierarchical clustering analyses extracted two main clusters from these data, which mapped onto the "inflexibility"-"instability" distinction. Future work will explore the extent to which the clusters detected with unsupervised learning in our study align with clinical diagnoses of inflexibility and instability.

Despite the promising results, the identified clusters do show overlap in some PCs; a part of the population likely shares characteristics of both subgroups and are thus not likely to form well-separated clusters [5]. With a larger sample, we expect that the groups will become more well-defined. A larger sample may enable us to determine the extent to which these clusters relate to different subtypes of spinocerebellar ataxia as well as other causes of ataxic dysarthria.

One potential methodological issue with the current study is that the material consisted of home recordings; the quality of the recordings is lower than in the lab and less consistent between participants, which influences the subsequent MFCC analysis [19, 20]. The data should thus be interpreted with caution, especially regarding the first parameter, intensity [19], which has been shown to be unstable with, especially, older versions of Zoom, and should be validated with material collected in the lab. That said, our results suggest well-defined groups even with possible distortions of the acoustic signal. Importantly, the ability to record speakers with rare disorders online provides a substantially larger sample size than would otherwise be available.

In sum, unsupervised clustering based on parameters extracted from energy changes over time is a promising approach in distinguishing dysarthria subgroups in ataxia. Though the extent to which the identified groups reflect the inflexibility-stability dichotomy remains to be tested, this method has potential to complement or replace perceptual assessments of different subtypes of ataxic dysarthria in clinical settings.

5. ACKNOWLEDGEMENTS

Work was supported by NIDCD/NIH awards R01 DC017091, F32DC0193535, and P50HD105353.

6. REFERENCES

[1] Duffy, R., *Motor Speech Disorders*, 4th ed., 2019. https://www.elsevier.com/books/motor-speech-disorders/duffy/978-0-323-53054-5 (accessed May 19, 2020).

[2] Darley, F. L., Aronson, A. E., and Brown, J. R., 'Differential diagnostic patterns of dysarthria', *J. Speech Hear. Res.*, vol. 12, no. 2, pp. 246–269, Jun. 1969, doi: 10.1044/jshr.1202.246.

[3] Kent, R. D., 'Research on speech motor control and its disorders: a review and prospective', *J. Commun. Disord.*, vol. 33, no. 5, pp. 391–427, Oct. 2000.

[4] Brendel, B., Synofzik, M., Ackermann, H., Lindig, T., Schölderle, T., Schöls, L., and Ziegler, W., 'Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with Friedreich ataxia', *J. Neurol.*, vol. 262, no. 1, pp. 21–26, Jan. 2015, doi: 10.1007/s00415-014-7511-8.

[5] Spencer, K. A., and France, A. A., 'Perceptual ratings of subgroups of ataxic dysarthria', *Int. J. Lang. Commun. Disord.*, vol. 51, no. 4, pp. 430–441, Jul. 2016, doi: 10.1111/1460-6984.12219.

[6] Boutsen, F., Bakker, K., and Duffy, J., 'Subgroups in ataxic dysarthria', *J. Med. Speech-Lang. Pathol.*, vol. 5, pp. 27-36, Mar. 1997.

[7] K. A. Spencer, J. Amaral, and K. Lansford, 'Investigating perceptual subgroups in speakers with ataxic dysarthria: An auditory free classification approach', *Am J. Speech Lang Pathol.,* pp. 1-11, 2022, 10.1044/ 2022_AJSLP-22-00159.

[8] Goldstein, L., 'The role of temporal modulation in sensorimotor interaction', *Front. Psychol.*, vol. 10, p. 2608, Dec. 2019, doi: 10.3389/fpsyg.2019.02608.

[9] Slis, A., Lévêque, N., Fougeron, C., Pernon, M., Assal, F., and Lancia, L., 'Analysing spectral changes over time to identify articulatory impairments in dysarthria', *J. Acoust. Soc. Am.*, vol. 149, no. 2, p. 758, Feb. 2021, doi: 10.1121/10.0003332.

[10] Lévêque, N., Slis, A., Lancia, L., Bruneteau, G., and Fougeron, C., 'Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases', *J. Speech Lang. Hear. Res.*, vol. 65, no. 5, pp. 1767–1783, May 2022, doi: 10.1044/2022_JSLHR-21-00434.

[11] Boersma, P., and Weenink, D., Praat: doing phonetics by computer [Computer program], Version 6.1.03, retrieved Spetember 1, 2019 from http://www.praat.org/.

[12] O'Shaughnessy, D., *Speech communication: human and machine*. Addison-Wesley Pub. 6Co., 1987.

[13] Caliński, T. and Harabasz, J., 'A dendrite method for cluster analysis', *Commun. Stat..*, vol. 3, no. 1, pp. 1–27, Jan. 1974, doi: 10.1080/03610927408827101.

[14] CoRDS (https://research.sanfordhealth.org/rare - disease-registry/researchersj).

[15] Chen, M., 'EM Algorithm for Gaussian Mixture Model (EM GMM)', https://www.mathworks.com/matlabcentral/fileexchange/26184-em-algorithm-for-gaussian-mixture-model-em-gmm, Retrieved November 15, 2022.

[16] Schalling, E., Hammarberg, B., and Hartelius, L., 'A longitudinal study of dysarthria in spinocerebellar ataxia (SCA): aspects of articulation, prosody, and voice', *Journal of Medical Speech–Language Pathology*, 16, 103–117, 2008.

[17] The MathWorks, Inc., MATLAB version: 9.13.0 (R2022a). 2022.

[18] Audacity Team, Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.2.1 retrieved October 27, 2022, from https://audacityteam.org/.

[19] Zangh, C., Jepson, K., Loblink, G., and Arvanti, A., 'Comparing acoustic analyses of speech data collected remotely', *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp 3910-3916, 2021, doi: 10.1121/10.000513.

[20] Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., and Bowern, C., '(Don't) try this at home! The effects of recording devices and software on phonetic analysis', *Language,* vol. 97, no 4, 2021, pp e360-e382, doi:10.1353/lan.2021.0075.

VOWEL SPACE AREA CHANGES IN CHILDREN WITH DYSARTHRIA

Jiyoung Choi[1], Kyung Hae Hwang[1], Erika S. Levy[1]

Teachers College, Columbia University, USA
jyc2173@tc.columbia.edu, khh2106@tc.columbia.edu, elevy@tc.columbia.edu

ABSTRACT

Increasing speech intelligibility is an essential treatment target for children with dysarthria secondary to cerebral palsy. Vowel space area (VSA) is associated with intelligibility in dysarthria. The speech cues "speak with your big mouth" and "speak with your strong voice" have been found to increase intelligibility in this population. This study examined changes in VSA in response to speech cues in children with dysarthria. We further investigated whether dysarthria severity impacted the VSA changes. Thirteen children with dysarthria were cued to speak with a "big mouth" or "strong voice." First and second formants were measured from four corner vowels in carrier sentences. Preliminary results show a trend toward increased VSA, with children with mild dysarthria exhibiting the greatest changes. Therefore, cueing children with dysarthria, especially mild dysarthria, to speak with their "big mouth" and/or "strong voice" may increase their VSA and thus, potentially contribute to increasing their intelligibility.

Keywords: childhood dysarthria, cerebral palsy, vowel space area, speech cues

1. INTRODUCTION

Motor speech disorders refer to disorders resulting from neurologic impairments that affect the motor planning/control of speech movements and the execution of speech. Dysarthria is a motor speech disorder caused by muscle weakness and/or coordination/control difficulties due to neurological damage. It may be characterized by imprecise articulation, slow speech rate, reduced vocal intensity, and other variable deficits [1]. Cerebral palsy (CP) is one of the most common causes of dysarthria. CP is a congenital, heterogeneous disorder that affects the ability to move and/or maintain balance and posture. It is the most common neuromotor disability in childhood, with approximately 1 in 345 children in the US identified with CP [2].

Dysarthria is characterized by distinctive temporal and spectral speech features, often including reduced articulation rate, reduced second formant (F2) range of diphthongs, as well as overall atypical vowel formant structure when compared to the speech of typically-developing children [3], [4]. Reduced intelligibility is therefore a common characteristic in children with dysarthria due to cerebral palsy (CP) [5]. Thus, increasing intelligibility is an essential goal in speech treatment for this population [6]–[10] In addition, vowel space area (VSA) has been a common indicator of speech intelligibility in dysarthria. Previous work has found that compared to typically-developing children, those with dysarthria demonstrate smaller VSA, which is associated with reduced intelligibility [11]–[14].

Findings from studies on adults with Parkinson's disease (PD) indicate that VSA increases in response to various speech cues (e.g., clear, loud, slow, over-enunciate) relative to habitual speech [15], [16]. It is also reported that VSA increases and is associated with greater speech intelligibility post-treatment in adults with PD through Lee Silverman Voice Treatment [17], [18]. Moreover, changes in VSA and listeners' vowel ratings have been examined following speech treatment in adults with PD [18]. In comparison to pre-treatment, VSAs were larger post-treatment, with higher vowel goodness ratings.

In children with dysarthria, VSA changes following speech cues or treatment are understudied. Previous studies examined F1 and F2 changes in individual vowels (e.g., /æ/ and /ɑ/) when children with dysarthria were cued to speak with a "big mouth" or "strong voice" [19] or following Speech Intelligibility Treatment [9], which incorporates these cues in a three-week treatment program. The "big mouth" cue aims to increase VSA and articulatory excursion, and the "strong voice" cue aims to increase vocal intensity, and therefore to bring about amplitude increases across speech production systems [9]. Because VSA is a key predictor of speech intelligibility in childhood dysarthria [13], it is important to examine whether providing speech cues for children with dysarthria increases their VSA, as well as the relationship between any VSA changes and intelligibility changes. Such information may provide direction for optimizing speech treatment strategies and may shed light on mechanisms behind speech intelligibility improvement in this population. Findings may also help determine whether factors such as dysarthria severity affect intelligibility gains.

Participants	Age	Sex	Cerebral palsy diagnosis	GMFCS*	Dysarthria Severity	Receptive Language Skills**	
CP01	12;0	M	mixed	V	Severe	63th	average
CP02	7;4	F	mixed	III	Mild	<1st	very poor
CP03	12;8	M	spastic	V	Moderate	9th	below average
CP04	15;7	M	mixed	V	Mild	<1st	very poor
CP05	14;7	M	spastic	II	Mild	50th	average
CP06	16;1	M	spastic	I	Mild	37th	average
CP07	10;9	M	spastic	II	Mild	25th	below average
CP08	5;2	M	spastic	IV	Moderate	25th	below average
CP09	12;5	F	spastic	IV	Severe	<1st	very poor
CP10	13;2	F	spastic	IV	Moderate-Severe	2nd	very poor
CP11	9;7	M	mixed	II	Mild	37th	average
CP12	5;1	M	spastic	IV	Mild	37th	average
CP13	14;1	M	ataxic	II	Mild-Moderate	<1st	very poor

Table 1: Demographics of children. Gross Motor Classification System (GMFCS)[20]*. M = male; F=female. Percentile rank obtained from the Test of Auditory Comprehension of Language (TACL) or Clinical Evaluation of Language Fundamentals-5 (CELF-5)**.

The current study examined the changes in vowel space area (VSA) when children with dysarthria were cued to speak with their "big mouth" or their "strong voice". It was hypothesized that VSA in the "big mouth" condition would be larger than VSA in their habitual speech. The "strong voice" cue was also expected to increase VSA, albeit less so than the "big mouth" cue [9], [19]. We further investigated whether dysarthria severity impacted the VSA changes. It was hypothesized that children with mild dysarthria would demonstrate greater VSA changes than those with moderate or severe dysarthria [9], [19].

2. EXPERIMENT

2.1. Participants

A total of 14 monolingual English-speaking children with dysarthria secondary to CP participated in the study. One child's data were excluded due to incomplete tasks. Therefore, data from a total of 13 children are reported in the study (mean age = 11;4 yrs; range 5;1–16;1 yrs; 3 females, 10 males). Subtypes of cerebral palsy and dysarthria severity varied. Dysarthria severity level was determined by two certified speech-language pathologists based on independent assessments of speech intelligibility [21], [22]. Any disagreements of judgment were reconciled by consensus. All children presented with normal hearing. Table 1 summarizes participant characteristics.

2.2. Stimuli

The stimuli in this study were sentences (i.e., "They say CVC again") containing American English CVC words (beat, bat, boot, and pot), representing the four corner vowels (/i/, /æ/, /u/, and /ɑ/, respectively). Sentences were pre-recorded by a monolingual American English female adult in preparation for presentation to the children with dysarthria, who repeated the sentence. Each sentence was recorded in three speech conditions: habitual speech, and in response to "big mouth", and "strong voice" cues. A total of 156 stimuli (13 children x 4 sentences including four corner vowel-representing words x 3 speech conditions) were analyzed.

2.3. Procedure

The experiment was conducted in the children's homes with instructions delivered virtually by researchers and the procedure monitored by Zoom video [23]. Recording equipment had been sent to the child's home. The child's parent placed a Countryman EMW Lavalier microphone on the child's forehead, 8cm from the child's lips. The pre-recorded sentences were played by the researcher and transmitted through Zoom. The children were asked to repeat the pre-recorded sentences. The habitual condition was followed by the cued conditions, with the "big mouth" and "strong voice" conditions counterbalanced. For the habitual condition, children were simply asked to repeat what they heard. For the "big mouth" condition, children were instructed to speak with their "big mouth," and for the "strong voice" condition, with their "strong voice". Verbal modeling was provided throughout the task if necessary. After the experiment, the parents returned the recording equipment including the sound card with the children's audio recordings. Then researchers downloaded the children's audio recordings from the sound card.

2.4. Analysis of VSA metrics

For VSA analysis, the CVC words (beat, bat, boot, and pot) were extracted from recordings of the carrier sentences produced by the children. and quadrilateral VSA was examined [4]. First, the onset and offset of each corner vowel (/i/, /æ/, /u/, /ɑ/) and the total vowel duration (ms) were measured. Next, the first and second formants (F1 and F2) were manually extracted at the temporal midpoint of each vowel using 25-ms window wideband spectrograms and linear predictive coding (LPC) spectra. F1 is associated with tongue height, and F2 is related to tongue forwardness/backness [24]. The quadrilateral VSA was calculated (Hz^2) by means of Heron's formula and Euclidean distance measurement [4], [16] via 'phonR' package [25]. Figure 1 represents an example of the configuration of VSA.

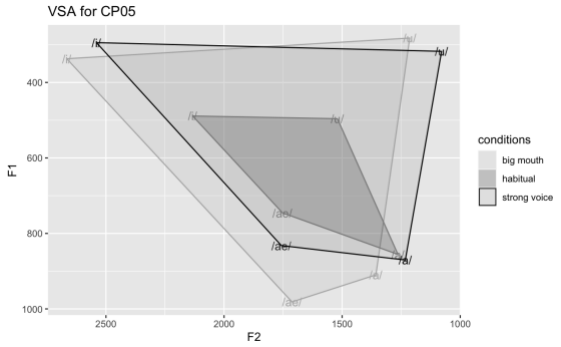


Figure 1: Vowel Space Area (VSA) for a participant CP05 with mild dysarthria.

2.5. Statistical analysis

Due to the small number of participants in the study, $p = .05$ was considered statistically significant [12]. To examine the VSA change following the speech cues within the single child, repeated-measures analyses of variance (ANOVA) with Bonferroni-corrected post hoc pairwise comparisons were utilized [9], [19]. Furthermore, relationships between speech conditions and each severity group were examined descriptively because of the small number of children within each severity group.

3. RESULTS

Preliminary results indicate a trend toward increased VSA in both cued conditions relative to habitual speech, except in one child with severe dysarthria (See Table 2). Moreover, VSA gains were greater following the "big mouth" cue than following the "strong voice" cue. There were four cases in which VSA increased in response to only one speech cue, compared to habitual speech; i.e., in two children, VSA increased only in "big mouth" condition and for two children, only in "strong voice" condition. Visual inspection further indicates that for 10 of the 13 participants, VSA increased following at least one of the speech cues, whereas for 2 (CP03, CP09), VSA decreased.

As can be seen in Figure 2, the linear mixed-effects model revealed a significant main effect of speech cues in VSA, $F(1, 12) = 47.023$, $p < .001$. Post hoc Bonferroni-corrected pairwise comparisons indicated a significant increase in VSA in the "big mouth" condition relative to habitual speech (mean difference = 127618.54, 95% CI [3688.34, 251548.73], $p < .05$). No significant differences were found between the "strong voice" condition and the habitual condition (mean difference = 55186.31, 95% CI [−43421.16, 153793.77], $p > .05$).

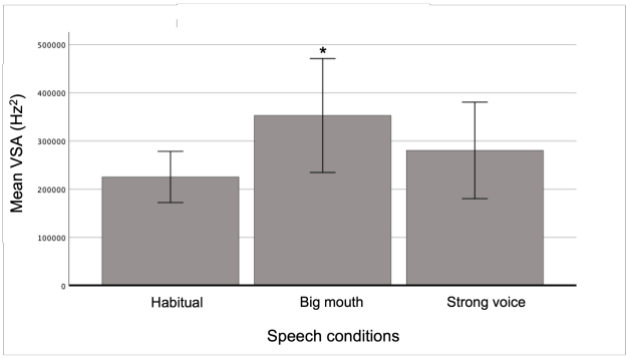


Figure 2: Mean Vowel Space Area (VSA) across 13 children with dysarthria in three speech conditions. Standard error bars are included. *p ≤ .05.

Additionally, VSA changes varied as a function of dysarthria severity. Visual inspection suggests that in children with mild dysarthria, VSA was larger following either speech cue when compared to the children's habitual speech. Noticeable increases of VSA following either cue were observed in three children with mild-moderate and moderate dysarthria. For the 2 children with moderate dysarthria, one child showed significant VSA gains following both speech cues, whereas the other child showed no VSA changes across conditions. The child with moderate-severe dysarthria demonstrated VSA increases following both speech cues. For the 2 children with severe dysarthria, VSA increased in "big mouth" cue in one child, but no increase was found following either cue in the other child.

Partici-pants	Vowel space area (Hz^2)			Dysar-thria severity
	Habitual	Big mouth	Strong voice	
CP01	180569	240827	176776	Severe
CP02	184852	238814	214257	Mild
CP03	241868	116860	224397	Moderate
CP04	145460	216182	207569	Mild
CP05	247062	600296	566227	Mild
CP06	253046	296042	163456	Mild
CP07	213545	193151	264911	Mild
CP08	216040	424042	347689	Moderate
CP09	220998	151468	110574	Severe
CP10	73975	217720	191155	Moderate-Severe
CP11	200882	545829	269332	Mild
CP12	497715	829154	747288	Mild
CP13	254101	518769	163904	Mild-Moderate

Table 2: Vowel Space Area (VSA) in three speech conditions.

4. DISCUSSION

When children with dysarthria were cued to speak with a "big mouth" or "strong voice," a trend toward increased VSA was revealed in both cued conditions. VSA increased significantly, particularly in children with mild dysarthria. These findings suggest that speech cues targeting articulatory excursion and increased vocal intensity may positively impact VSA in children with mild dysarthria. Because VSA is an important predictor of speech intelligibility in children with dysarthria [12], [13], targeting VSA directly (e.g., "big mouth") or indirectly (e.g., "strong voice") may lead to improvements in intelligibility, especially in children with mild dysarthria.

However, while the majority of the children showed increased VSA following the speech cues, VSA of 2 of 13 children (CP03, CP09) was smaller following both cues than in their habitual speech. This result was unexpected, given that in Levy et al.'s cueing study [19], all 8 of the children with dysarthria increased intelligibility when responding to these cues. Our result might be a consequence of noise in the audio recordings. Thus, acoustic predictors of intelligibility changes in children with dysarthria require further investigation and may vary depending on the child's speech characteristics.

Future directions include determining adult listeners' perception of the children's word and sentence productions to more directly assess the impact of speech cues and VSA on intelligibility.

5. ACKNOWLEDGEMENT

This study was supported, in part, by a 2019 ASHFoundation Independent Clinical Research Grant awarded to Erika S. Levy.

6. REFERENCES

[1] J. R. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[2] CDC, "Data and Statistics for Cerebral Palsy | CDC," *Centers for Disease Control and Prevention*, Dec. 30, 2020. https://www.cdc.gov/ncbddd/cp/data.html (accessed Jan. 05, 2023).

[3] K. M. Allison and K. C. Hustad, "Acoustic Predictors of Pediatric Dysarthria in Cerebral Palsy," *J. Speech Lang. Hear. Res.*, vol. 61, no. 3, pp. 462–478, Mar. 2018, doi: 10.1044/2017_JSLHR-S-16-0414.

[4] K. L. Lansford and J. M. Liss, "Vowel Acoustics in Dysarthria: Mapping to Perception," *J. Speech Lang. Hear. Res.*, vol. 57, no. 1, pp. 68–80, Feb. 2014, doi: 10.1044/1092-4388(2013/12-0263).

[5] C. Mei, S. Reilly, D. Reddihough, F. Mensah, and A. Morgan, "Motor speech impairment, activity, and participation in children with cerebral palsy," *Int. J. Speech Lang. Pathol.*, vol. 16, no. 4, pp. 427–435, Aug. 2014, doi: 10.3109/17549507.2014.917439.

[6] L. Pennington, C. Smallman, and F. Farrier,

"Intensive dysarthria therapy for older children with cerebral palsy: findings from six cases," *Child Lang. Teach. Ther.*, vol. 22, no. 3, pp. 255–273, Oct. 2006, doi: 10.1191/0265659006ct307xx.

[7] L. Pennington *et al.*, "Internet delivery of intensive speech and language therapy for children with cerebral palsy: a pilot randomised controlled trial," *BMJ Open*, vol. 9, no. 1, p. e024233, Jan. 2019, doi: 10.1136/bmjopen-2018-024233.

[8] C. M. Fox and C. A. Boliek, "Intensive Voice Treatment (LSVT LOUD) for Children With Spastic Cerebral Palsy and Dysarthria," *J. Speech Lang. Hear. Res.*, vol. 55, no. 3, pp. 930–945, Jun. 2012, doi: 10.1044/1092-4388(2011/10-0235).

[9] E. S. Levy, Y. M. Chang, K. Hwang, and M. J. McAuliffe, "Perceptual and Acoustic Effects of Dual-Focus Speech Treatment in Children With Dysarthria," *J. Speech Lang. Hear. Res.*, vol. 64, no. 6S, pp. 2301–2316, Jun. 2021, doi: 10.1044/2020_JSLHR-20-00301.

[10] C. A. Boliek and C. M. Fox, "Therapeutic effects of intensive voice treatment (LSVT LOUD®) for children with spastic cerebral palsy and dysarthria: A phase I treatment validation study," *Int. J. Speech Lang. Pathol.*, vol. 19, no. 6, pp. 601–615, Nov. 2017, doi: 10.1080/17549507.2016.1221451.

[11] K. M. Allison, L. Annear, M. Policicchio, and K. C. Hustad, "Range and Precision of Formant Movement in Pediatric Dysarthria," *J. Speech Lang. Hear. Res.*, vol. 60, no. 7, pp. 1864–1876, Jul. 2017, doi: 10.1044/2017_JSLHR-S-15-0438.

[12] J. Lee and K. C. Hustad, "A Preliminary Investigation of Longitudinal Changes in Speech Production over 18 Months in Young Children with Cerebral Palsy," *Folia Phoniatr. Logop.*, vol. 65, no. 1, pp. 32–39, 2013, doi: 10.1159/000334531.

[13] J. Lee, K. C. Hustad, and G. Weismer, "Predicting Speech Intelligibility With a Multiple Speech Subsystems Approach in Children With Cerebral Palsy," *J. Speech Lang. Hear. Res.*, vol. 57, no. 5, pp. 1666–1678, Oct. 2014, doi: 10.1044/2014_JSLHR-S-13-0292.

[14] Z. Mou *et al.*, "Quantitative analysis of vowel production in cerebral palsy children with dysarthria," *J. Clin. Neurosci.*, vol. 66, pp. 77–82, Aug. 2019, doi: 10.1016/j.jocn.2019.05.020.

[15] K. Tjaden, J. Lam, and G. Wilding, "Vowel Acoustics in Parkinson's Disease and Multiple Sclerosis: Comparison of Clear, Loud, and Slow Speaking Conditions," *J. Speech Lang. Hear. Res.*, vol. 56, no. 5, pp. 1485–1502, Oct. 2013, doi: 10.1044/1092-4388(2013/12-0259).

[16] J. Lam and K. Tjaden, "Clear Speech Variants: An Acoustic Study in Parkinson's Disease," *J. Speech Lang. Hear. Res.*, vol. 59, no. 4, pp. 631–646, Aug. 2016, doi: 10.1044/2015_JSLHR-S-15-0216.

[17] L. O. Ramig *et al.*, "Intensive voice treatment (LSVT®) for patients with Parkinson's disease: A 2 year follow up," *J. Neurol. Neurosurg. Psychiatry*, vol. 71, no. 4, pp. 493–498, 2001.

[18] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of Intensive Voice Treatment (the Lee Silverman Voice Treatment [LSVT]) on Vowel Articulation in Dysarthric Individuals With Idiopathic Parkinson Disease: Acoustic and Perceptual Findings," *J. Speech Lang. Hear. Res.*, vol. 50, no. 4, pp. 899–912, Aug. 2007, doi: 10.1044/1092-4388(2007/064).

[19] E. S. Levy, Y. M. Chang, J. A. Ancelle, and M. J. McAuliffe, "Acoustic and Perceptual Consequences of Speech Cues for Children With Dysarthria," *J. Speech Lang. Hear. Res.*, vol. 60, no. 6S, pp. 1766–1779, Jun. 2017, doi: 10.1044/2017_JSLHR-S-16-0274.

[20] R. Palisano, P. Rosenbaum, S. Walter, D. Russell, E. Wood, and B. Galuppi, "Development and reliability of a system to classify gross motor function in children with cerebral palsy," *Dev. Med. Child Neurol.*, vol. 39, no. 4, pp. 214–223, Sep. 1997, doi: 10.1111/j.1469-8749.1997.tb07414.x.

[21] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *J. Speech Hear. Disord.*, vol. 54, no. 4, pp. 482–499, Nov. 1989, doi: 10.1044/jshd.5404.482.

[22] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel Contrast and Speech Intelligibility in Dysarthria," *Folia Phoniatr. Logop.*, vol. 63, no. 4, pp. 187–194, 2011, doi: 10.1159/000318881.

[23] Hwang, Kyung Hae *et al.*, "Validity of online speech data collection from children with dysarthria due to cerebral palsy.," presented at the Motor Speech Conference 2022, Feb. 17, 2022.

[24] R. D. Kent, R. A. Kent, and C. Read, *The acoustic analysis of speech*. Singular, 2002.

[25] D. McCloy, "Phonetic effects of morphological structure in Indonesian vowel reduction," presented at the 166th Meeting of the Acoustical Society of America, San Francisco, California, 2014, p. 060009. doi: 10.1121/1.4870068.

PRENASALIZATION IN GREEK-LEARNING CHILDREN WITH COCHLEAR IMPLANTS AND TYPICAL HEARING

Laura L. Koenig,[1] Areti Okalidou,[2] George Kyriafinis[3]

[1]Adelphi University, USA; [2]University of Macedonia, Greece; [3] Cochlear Implant Center, 1st University ENT Clinic, Ahepa Hospital, Aristotle University of Thessaloniki, Greece
lkoenigl@gmail.com; okalidou@uom.edu.gr; kyriafinis@otenet.gr

ABSTRACT

This work compares nasalance measures for word-initial consonants in Greek-learning children with typical hearing and with cochlear implants [CIs]. An interesting feature of Greek is that many dialects employ prenasalization of phonemically voiced stops. One can anticipate that the rapid shift between an open vs. closed velopharyngeal port represents a challenge to typical children, and all the more so for those who do not have typical hearing. Greek-speaking children with CIs have been found to have longer prevoicing, which could reflect greater nasal coupling. Current results show that children with CIs, compared to those with NH, have a lower proportion of prevoicing; higher nasalance values for the consonant, vowel, and word; and greater variability in nasalance measures for the consonant, vowel, and word. Some children with CIs showed considerable carryover nasalization from the initial stop consonant. Despite these group-level patterns, speaker-specific variation is extensive in both groups.

Keywords: Greek, prenasalized stops, children, cochlear implantation

1. INTRODUCTION

1.1 Effects of deafness and cochlear implantation on speech production

Past work shows that hearing loss can lead to atypical patterns of nasality in speech, which could be either hyponasality or hypernasality [1–3]. Some studies suggest that cochlear implantation can improve oro-nasal balance in children [4], but others find that increased or decreased nasality compared to normally-hearing [NH] peers may persist [5–7]. To date, instrumental studies of nasality post-cochlear implantation have not sampled widely across languages. Conceivably, some of the discrepancies that exist in the literature could reflect cross-linguistic differences.

1.2. Prenasalization in Greek

Many speakers of Modern Greek produce voiced stops with prenasalization [8–9], i.e. nasal energy is high early in the consonantal closure and diminishes over time. The magnitude of prenasalization may be higher in children than adults [10–11], and could facilitate phonation in a small vocal tract. Interestingly, a recent study [12] found that Greek-learning children with CIs produced longer prevoicing durations than typical-hearing counterparts, which would be consistent with greater nasal coupling in the CI group.

Both prevoicing and nasality are low in amplitude and involve "hidden" articulators. These factors, combined with the rapid velar elevation required to go from a nasal sound to an oral sound, could make this phonetic pattern particularly difficult for those with prelingual deafness. The current work thus assessed patterns of prenasalization in typically-developing children and those with cochlear implants [CIs] who are learning Greek as their first language. The data can also provide insight into the nature of inter-articulatory timing in children with CIs.

2. METHODOLOGY

2.1. Speaker information

Data are shown for 16 speakers: Eight with cochlear implants and eight age-matches with typical hearing (Table 1). Most children with CIs were implanted between 16 and 38 months, except for speaker CIF4_17 who received her implant at 14;10 years.

2.2. Speech stimuli

Target words contained initial /b, d/ in disyllables (see Table 2). Children produced the words in randomized order in response to a visual stimulus and, particulalarly for the younger children, often an oral prompt. Five productions were elicited. Note that the word /dinɑ/ allows assessment of carryover as well as anticipatory nasal coarticulation.

SpeakerID	Age	SpeakerID	Age
CIM7_4	4;2	NHM11_4	4;4
CIM3_4	4;7	NHF12_4	4;6
CIF3_5	5;3	NHM7_4	4;11
CIF1_5	5;10	NHF11_6	6;0
CIM5_6	6;6	NHM8_6	6;6
CIM6_7	7;4	NHF13_7	7;0
CIF5_8	7;10	NHM6_8	8;4
CIF4_17	17;3	NHF10_16	16;10

Table 1: Speakers and ages (years;months). Children with CIs are at left and those with NH are at right. Gender is indicated as M or F.

Orthog.	IPA	Gloss
μπύρα	'biɾɑ	*beer*
μπαλιά	bɑ'ljɑ	*a soccer pass*
Ντίνα	'dinɑ	*A name*
ντεφι	'dɛfi	*tambourine*

Table 2: Stimulus items

2.3. Instrumentation and analysis

2.3.1. Hardware

Data were collected via the Kay Elemetrics™ (now Pentax) nasometer system. The device employs a stabilizing headset, a stiff plastic plate that rests at the upper lip, and microphones above and below the plate to capture nasal and oral energy, respectively. The system software outputs a nasometer value that represents the ratio of nasal to oral energy. Higher nasalance scores represent a stronger nasal signal.

2.3.2. Labeling

The oral and nasal microphone data were opened in Praat [13] and labeled for the word duration, the duration of any prenasalization, and the duration of the first-syllable vowel (see Figure 1). Note that *any* periodic acoustic energy observed prior to the stop release was labeled on the 'prenasalization' tier (cf. Figure 1) even if that region appeared to contain closure voicing rather than nasalization. That is, rather than making a subjective decision as to what amounted to prenasalization, we simply obtained nasalance values over all such labeled regions. These cases were infrequent in the dataset.

The acoustic data are low-passed filtered as part of the nasometer processing; this is evident in the spectrogram. Segmental labeling therefore relied on low-frequency information and amplitude changes.

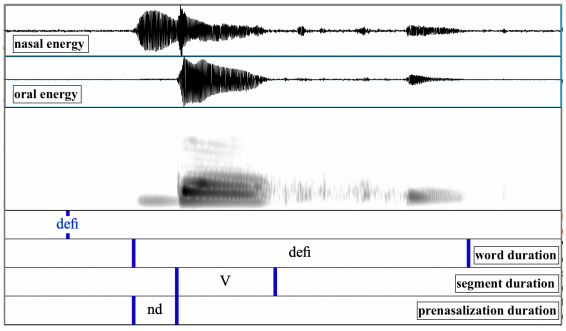


Figure 1: Labeled data in Praat. The two waveforms represent the nasal and oral microphone signals. The spectrogram scale is 0–4000 Hz. The duration shown here is about 1.4 s. This example is from NHF12_4 (four years of age).

2.3.3. Subsequent processing

The Praat time labels were loaded into Matlab, along with the nasometer output signals (oral energy, nasal energy, and nasalance ratio, all taken at 8 ms intervals). Word, prenasalization, and segmental (here, vowel) durations were extracted from the long nasometer files. From these extracted portions, we obtained nasalance averages and standard deviations for each word, prenasalized region, and vowel. Note that these standard deviations (SDs) reflect variation in nasality within individual productions, so, e.g., if a child maintained a rather consistent velar position during a vowel, the SD for that production would be low.

3. RESULTS

3.1. Group-level data

3.1.1. Prevalence of prevoicing

Percentages of labeled prevoiced/prenasalized regions (cf. section 2.3.2) were 76.3% in the CI group and 89.4% in the NH group. This difference was significant (Pearson's χ^2=8.655, df=1, p=0.003). The frequency of prevoicing/prenasalization shows no apparent relationship with age (see Figure 2).

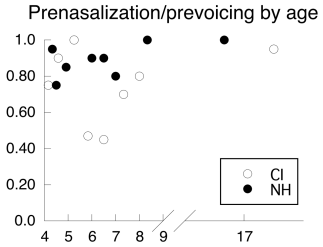


Figure 2: Proportion of tokens with prenasalization/ prevoicing as a function of age for the two groups. The x-axis was split in light of the large age gap between the two oldest speakers and the rest of the participants.

WdAvg	n	mean	sd	median	se
CI	154	27.97	18.11	25.62	1.46
NH	160	18.43	13.68	13.56	1.08
PreNasAvg	n	mean	sd	median	se
CI	119	38.11	28.51	38.96	2.61
NH	143	16.29	20.58	4.00	1.72
VAvg	n	mean	sd	median	se
CI	128	21.64	15.72	17.57	1.39
NH	150	14.75	10.65	10.97	0.87
PreNasDur	n	mean	sd	median	se
CI	119	0.15	0.08	0.14	0.01
NH	143	0.12	0.06	0.11	0.01

Table 3: Group-level nasalance data (counts (N), means, standard deviations, medians, and standard errors) shown for whole words, the prenasalized region, and the vowel. Nasalance is a percentage; durations are in seconds. The differences in numbers of tokens reflect cases where finer segmentation could not be made with confidence. For example, if a child produced a medial glide instead of a medial flap, we could obtain word duration but not a reliable measure of vowel duration.

WdSD	n	mean	sd	median	se
CI	154	21.29	11.7	24.94	0.94
NH	160	18.43	13.68	13.56	1.08
PreNasSD	n	mean	sd	median	se
CI	119	28.73	16.65	35.22	1.53
NH	143	15.91	16.29	10.91	1.36
VSD	n	mean	sd	median	sec
CI	128	9.86	7.68	8.1	0.68
NH	150	8	7.77	4.96	0.63

Table 4: Group-level standard deviations for whole words, prenasalized regions, and vowels

3.1.2. Degree of nasalization

Tables 3 and 4 present descriptive data for nasalance averages and standard deviations in the two groups. Figure 3 shows the data for prenasalization. Finally, Table 5 presents the statistical results. Note that the word averages include the prenasalization and vowels, i.e. the measures are not independent. Subsequent data presentation will therefore focus on prenasalization and vowel data.

DV	Levene (p-value)	t-value (Welch)	Sign. level (p-value)
WdAvg	<0.001	5.3	<0.001
PreNasAvg	<0.001	7.0	<0.001
VAvg	<0.001	4.2	<0.001
PreNasDur	<0.001	3.6	<0.001
WdSD	0.613	2.9	<0.010
PreNasSD	0.244	6.2	<0.001
VSD	0.261	2.0	0.05

Table 5: Statistical results for dependent variables. Since some data did not meet assumptions for homogeneity of variance, group differences were assessed using Welch's t-test.

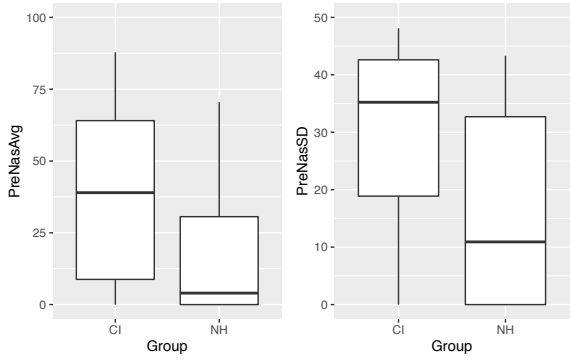


Figure 3: Group data for speaker nasalance averages (left) and standard deviations (right) in prenasalized regions

It is evident that, as a group, those with cochlear implants have higher values of average prenasalization, and greater variability in their prenasalization. It is also clear, however, that there is considerable heterogeneity within groups. As the next section will show, one should not draw the conclusion that all speakers with CIs necessarily use prenasalization to a greater degree than their NH peers.

3.2. Speaker-specific data

3.2.1. Prevalence of prevoicing

Figures 4–5 show the prenasalization data (means, SDs) broken out by speaker. In both speaker groups, average nasalance shows extreme cross-speaker variation. The standard deviations show that within-speaker variability is also quite high for some participants.

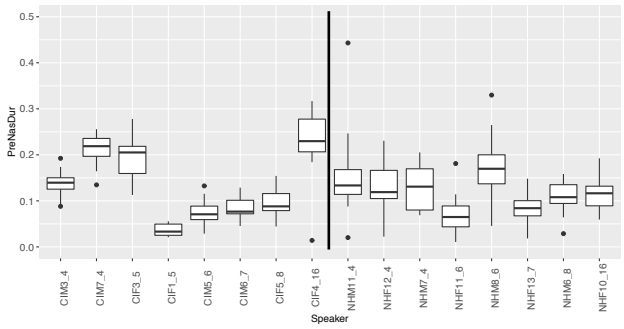


Figure 4: Individual speaker data of prenasalization averages for CI participants (left of the vertical bar) and NH participants (right of the vertical bar). Within group, participants are ordered by age.

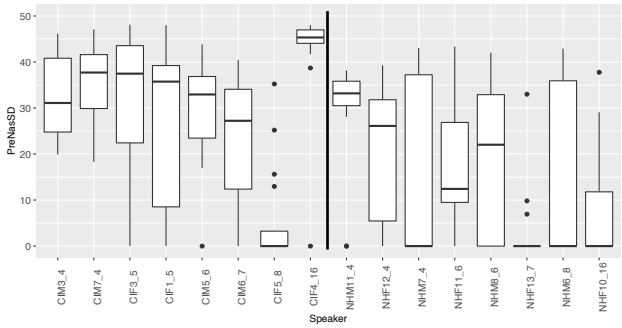


Figure 5: Individual speaker data of prenasalization standard deviations for CI participants (left of the vertical bar) and NH participants (right of the vertical bar). Within group, participants are ordered by age.

3.2.2. Coarticulation

As noted above, the word /dina/ allows assessment of both anticipatory and carryover coarticulation. Visual inspection of time-varying nasalance revealed a pattern in some of the CI children that was not observed in any of the NH children (in this modest sample): Whereas anticipatory nasalization for the medial /n/ was typical in both groups, a few children with CIs also showed extensive carryover nasality from the initial consonant. This is evident from high nasality values just after 0, the consonant release. Examples for four children, two with CIs and two NH children of similar ages, are shown in Figure 6.

4. DISCUSSION AND CONCLUSIONS

As a group, children with CIs produced somewhat less frequent prenasalization, greater magnitudes of prenasalization, and more variability of prenasalization within individual segments than their NH peers. Extensive speaker-specific variation was observed in both groups, however, and could arise from at least three sources: heterogeneity in

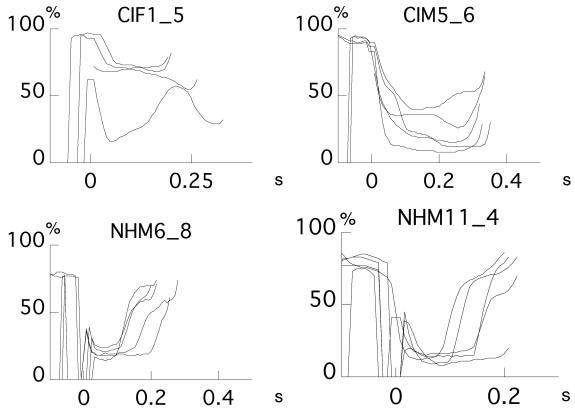


Figure 6: Nasalance over time for four participants: Two children with CIs (top) and two with NH (bottom). The y-axis shows percentage nasalance; the x-axis is time in seconds. The '0' on the x-axis represents the division between the consonant and the vowel.

development, heterogeneity within the CI population, and, possibly, dialectal variation. The proportion of prenasalized tokens did not show an apparent relationship with age. Consistent with [12], the younger NH children did tend to produce more extensive prenasalization (Figure 4), but this was not seen in the CI group. Interestingly, although many speech measures have shown decreasing variability with age (cf. [14] for a review), this was not seen for the degree of prenasalization within segments for either group (Figure 5).

The current results differ somewhat from those of [15]. That paper assessed nasalance values for medial consonants /p m n b/ in a small group of speakers (5 in each group). A group difference was observed there only for the nasals, with higher values in the CI children. That is, although that work also demonstrated a group pattern of more extensive nasalization in those with CIs, it did not extend to the one word with a *medial* prenasalized stop. Here, perhaps because of the larger dataset, a robust group difference was found for *initial* prenasalized stops. It may also be that word position is a relevant variable to explore.

As seen in much past work, group differences do not capture the speech characteristics of all children with CIs, however. Further, short-term (segmental) analyses and assessment of time-varying patterns provide a rich source of information on speech production in the CI population, and could ultimately inform intervention.

5. ACKNOWLEDGMENTS

This work was supported by a Fulbright fellowship (Fulbright, Greece) to the second author; host institution University of Macedonia. Also, partially supported by FP7_607149 EU grant (iCARE) to the first author. We thank our participants and their families.

6.. REFERENCES

[1] Colton, R. H., Cooker, H.S. 1968. Perceived nasality in the speech of the deaf. *J. Speech Hear. Res.* 11, 553–559.

[2] Fletcher, S. G., Mahfuzh, F., Hendarmin, H. 1999. Nasalance in the speech of children with normal hearing and children with hearing loss. *Am. J. Speech-Lg. Path.* 8, 241–246.

[3] Stevens, K. N., Nickerson, R.S., Boothroyd, A., Rollins, A. M. 1976. Assessment of nasalization in the speech of deaf children. *J. Speech Hr. Res.* 19, 393–416.

[4] Nguyen, L.H.P., Allegro, J., Low, A., Papsin, B., Campisi, P. 2008. Effect of cochlear implantation on nasality in children. *Ear, Nose & Throat J.* 87, 138–143.

[5] Van Lierde, K. M., Vinck, B. M., Baudonck, N., De Vel, E., Dhooge, I. 2005. Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: A pilot study. *Int. J. Audiology* 44, 452–465.

[6] Baudonck, N, Van Lierde, K. M., D'haeseleer, E., Dhooge, I. 2015. Nasalance and nasality in children with cochlear implants and children with hearing aids. *Int. J. Ped. Otorhinolaryng.* 79, 541–545.

[7] Sebastian, S., Sreedevi, N., Lepcha, A., Mathew, J. 2015. Nasalance in cochlear implantees. *Clin. Exp. Otorhinolaryng.* 8, 202–205.

[8] Arvaniti, A., Joseph, B. D. 2000. Variation in voiced stop prenasalization in Greek. *Glossologia: A Greek J. for Gen. and Hist. Ling.* 11–12, 131–166.

[9] Kong, E. J., Syrika, A., Edwards, J. R. 2012. Voiced stop prenasalization in two dialects of Greek. *J. Acoust. Soc. Am.* 132, 3439–3452.

[10] Okalidou, A., Petinou, K., Theodorou, E., Karasimou, E. 2010. Development of voice onset time in standard-Greek and Cypriot-Greek-speaking preschoolers. *Clin. Ling. and Phon.* 24, 503–519.

[11] Kong, E. J., Beckman, M. E., Edwards, J. 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *J. Phon.* 40, 725–744.

[12] Koupka, G.A., Okalidou, O., Nicolaidis, K., Konstantinidis, I., Kyriafinis, G. 2019. Production of the voicing contrast in children with cochlear implants. *Proc. 3rd Int. Conf. on Nat. Lg. and Speech Proc.,* 114–19.

[13] Boersma, P. 2022. Praat: doing phonetics by computer [computer program], http://www.praat.org.

[14] Koenig, L. L., Lucero, J.C., Perlman, E. 2008. Speech production variability in fricatives of children and adults: Results of functional data analysis. *J. Acoust. Soc. America* 124, 3158–70.

[15] Koenig, L.L., Okalidou, A., Psillas, G. 2018. Velopharyngeal control for speech in children with cochlear implants: Nasalance data in Greek VCV disyllables. *Proc. Mtgs on Acoustics*, 35, 060005

SPEECH BREATHING IN CHILDREN WITH A CLEFT PALATE

Delphine Charuau[1,2], Béatrice Vaxelaire[2] & Rudolph Sock[2,3]

[1] Univ. Grenoble Alpes, Grenoble INP, CNRS, GIPSA-lab, Grenoble, France, [2] University of Strasbourg, U.R.1339 Linguistique, Langues, Parole LiLPa, Strasbourg, France, [3] Pavol Jozef Šafárik University, Language, Information and Communication Laboratory (LICOLAB), Košice, Slovakia
delphinecharuau1@gmail.com

ABSTRACT

The aim of this paper is to investigate speech breathing patterns in young speakers with a cleft palate and to examine compensatory strategies deployed at the thoracic and abdominal levels in order to compensate for supraglottal deficits. To do so, 17 control speakers and 12 children with cleft palates were recorded. The variation of respiratory movements was measured using electromagnetic thoracic and abdominal belts. The respiratory signals were synchronized with acoustic data.
Results show that compensatory strategies are mainly deployed during expiration. The compression of the rib cage is higher for children with cleft palate than control speakers. Furthermore, this strategy is amplified when velopharyngeal functionality is more severely affected. However, for inhalation, no discernible difference was detected between control subjects and children with cleft palate. Finally, statistical analyses revealed that compensatory strategies do not have a significant effect on the temporal organisation of speech in children with cleft palate.

Keywords: Speech Breathing, Children, Cleft Palate, Velopharyngeal function.

1. INTRODUCTION

Cleft palate is craniofacial malformation affecting oral and nasal structures. Despite good chirurgical reconstruction, children with a cleft palate develop speech disorders such as articulation errors, sound substitutions, among others [1], and physiological impairments like velopharyngeal insufficiency. The soft palate does not offer a hermetic closure of velopharyngeal port which leads to nasal airway emission during speech production. Consequently, there is loss of intraoral pressure [2].
Consonantal production needs high and stable intraoral pressure [3][4]. The variation of lung volume expired ensures, in part, production and stability of intraoral pressure and adapts to the aerodynamic requirements of sounds [5].
Other studies focus on respiratory movements during production of a single word [6][7]. The data show that speakers with a cleft palate and velopharyngeal insufficiency produce intraoral pressure similar to control speakers by increasing respiratory movements. These speakers compensate loss of intraoral pressure by higher pulmonary volumes.
Breathing contributes to the temporal organisation of speech, grouping words within breath groups according to linguistic rules [8][9]. Hence, the disruption of breath patterns can impair speech timing.
The aim of this study is to investigate speech breathing patterns in young speakers with a cleft palate and to examine compensatory strategies deployed at the thoracic and abdominal levels in order to compensate for supraglottal deficits. We hypothesize that children with a cleft palate deploy respiratory compensatory strategies that are manifested through increased expiratory movements to compensate for the loss of intraoral pressure related to palatal malformations. Furthermore, we hypothesize that these speakers may extend inspiratory movements in anticipation of the increased aerodynamic need for phonation compared to control speakers. Finally, the respiratory strategies deployed by children with a cleft palate could have an impact on the temporal organisation of speech, given the close relationship that exists between breathing and speech production.

2. METHOD

2.1. Speakers and Corpus

Twelve children with repaired cleft palate or cleft and lip palate, between 7 and 12 years old (mean = 9.5; sd = 1.508) were recorded. These speakers were divided into three groups: those with adequate velopharyngeal function (CP-A), a slight velopharyngeal insufficiency (CP-sVI), characterized by nasal airway emission only during the production of some sounds, and those with severe velopharyngeal insufficiency (CP-SVI), characterized by continual nasal airway emission. Diagnosis of nasal airway emission and its severity was carried out by a speech therapist. The control group comprised seventeen children between 8 and 11 years old (mean = 9.71; sd = 0.772) without cleft

or lip palate and with no reported speech or respiratory disorders. All participants of this study were native speakers of French.
For the first task, children read a short sentence (on average 4 syllables per sentence) comprising a target vowel-consonant-vowel (VCV) sequence. For each sentence, consonants varied between voiceless plosives [p t k], voiced plosives [b d g], and voiceless fricatives [s] and [S].
In a subsequent task, each participant was asked to read the tale *La bise et le soleil* at a comfortable speech rate and intensity. This task was repeated a second time. For the spontaneous speech task, the speakers were asked to carry out a picture-based storytelling task. An example was provided with similar images before the beginning of the recording session.

2.2. Acquisition system

The variation of respiratory movements was measured using Respiratory Inductive Plethysmography, a Respitrace system (ADInstruments). Two electromagnetic belts were placed on the thorax and the abdomen of each speaker. The acoustic data was collected using a Senheiser e835s microphone and a Marantz Professional digital recorder and synchronized with respiratory signals using PowerLab (ADInstruments). The synchronized respiratory and speech signals allow observing variations of the thoracic-abdominal perimeter during phonation.

2.3. Data processing

In order to examine the breathing movements as a whole, we created a new signal by collapsing the thoracic and abdominal signals (1 Tho + 1 Abd) [11]. The amplitude of respiratory movements was reported according to maximum displacement [14], estimated for each speaker from isovolume maneuvers [11][10].
The respiratory data was processed using MATLAB software. Temporal speech measures were carried out using Praat software [12].

2.4. Speech and respiratory measurements

The respiratory contribution was evaluated using inspiratory and expiratory amplitudes. 'Inspiratory amplitude' is defined as the difference between minimum and maximum values of inspiratory movement [13]. 'Expiratory amplitude' is defined as the difference between initial lung level and lung level at the end of a breath group. A 'breath group' is defined as the interval between the end of a breath pause and the beginning of the second breath pause. In reading and spontaneous speech, the duration of each breath group was measured.

2.5. Statistical analyses

Statistical analyses were conducted using RStudio software (version 1.4.1717, 2021). Data were compared using mixed measures ANOVA. Based on these measures, we were able to verify the effect of inter (groups of speakers) and intra-subject factors (target consonants, speech tasks) on the analysed variables: inspiratory and expiratory contributions (amplitude), duration of breath groups and the number of syllables per breath group. The overall significance level was calculated as $p < 0,05$. When significance was indicated, Bonferroni tests were conducted.

3. RESULTS

3.1. Carrier sentences

3.1.1. Expiratory contribution during consonant cycle production

In both control and cleft palate speakers, the amplitude of expiratory movements revealed to be greater during consonant cycles containing fricative voiceless consonants than those containing plosive consonants (Voiceless fricatives: CP = 6.989 %MD; Control = 4.233 %MD; Voiceless plosives: CP = 3.379 %MD; Control = 2.922 %MD; Voiced plosives: CP = 3.446 %MD; Control = 2.197 %MD). Statistical analyses revealed a significant difference between fricatives and plosives ($F(1,21) = 1.928$; $p < 0.001$). However, no significant difference was found between voiceless plosive consonant cycles and their voiced counterparts.
A mixed linear regression analysis shows that the increase in expiratory contribution was not attributable to consonant duration.

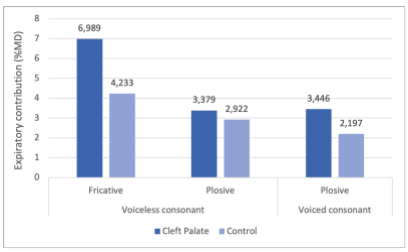


Figure 1: Expiratory contribution per consonantal cycle according to groups of speakers and target consonants.

The expiratory amplitude was found to be higher in children with a cleft palate than in control speakers, regardless of the type of consonant produced ($F(1,21) = 17.176$; $p < 0.001$). Furthermore, in children with a cleft palate, the respiratory

contribution during the production of consonant cycles was observed to be greater, particularly when the velopharyngeal function is more severely affected, (i.e., when nasal airway emission is continual; see Figure 2).

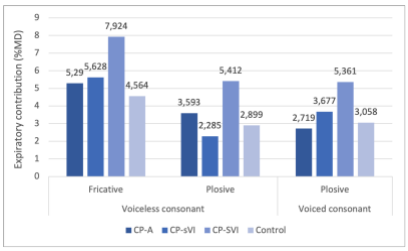


Figure 2: Expiratory contribution per consonantal cycle according to degrees of velopharyngeal inadequacy.

3.1.2. Expiratory contribution during breath group

Like in previous results, expiratory amplitude (Figure 3) was observed to be higher in children with a cleft palate than in control speakers ($F(1,21) = 4.610$; $p < 0.001$). Chest compression was measured on average at 28.998 %MD in children with a cleft palate and at 16.441%MD for control speakers.
Once again, expiratory movements were more exaggerated when the velopharyngeal function was severely affected (CP-SVI = 36.978 %MD; CP-sVI = 28.251 %MD; CP-A = 27.282 %MD).
However, the type of target consonant within the VCV sequence had no significant impact on the expiratory amplitude during the breath group.

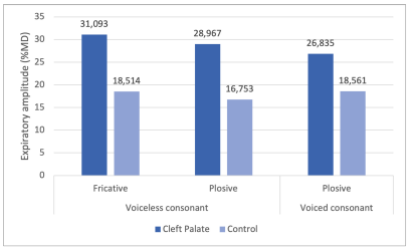


Figure 3: Expiratory contribution per breath group according to groups of speakers and target consonants.

3.1.3. Inhalation

Inhalation depth was similar for both groups of speakers (figure 4). While occasionally it was measured as slightly higher in control speakers, the difference was not significant.
Moreover, inspiratory contribution tends to be higher when it precedes a sentence that includes a voiceless fricative target consonant. ANOVA analyses revealed that this difference was not significant.

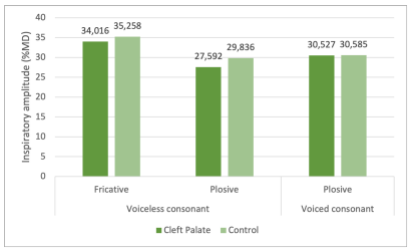


Figure 4: Inspiratory contribution according to groups of speakers and target consonants within sequence VCV

3.2. Speech breathing during read and spontaneous speech

3.2.1. Expiratory contribution

Expiratory movements tend to be greater in children with a cleft palate (figure 5), both in reading (CP = 26.09 %MD; Control speakers = 22.50 %MD), and in spontaneous speech (CP = 29.06 %MD; Control speakers = 24.69 %MD). Mixed measures ANOVA analyses revealed that there is no significant difference between the two groups of speakers ($F(1,21) = 0.735$; $p = 0.401$). Additionally, statistically significant differences were observed in chest compression in children with severe velopharyngeal insufficiency as opposed to control speakers (reading = 36.37 %MD; spontaneous speech = 31.21 %MD).
While expiratory movements are slightly greater in spontaneous speech for both groups of speakers, the difference was not significant.

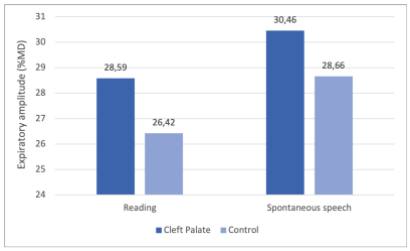


Figure 5: Expiratory contribution according to groups of speakers and speech tasks.

3.2.2. Inhalation contribution

As in previous results, inspiratory contribution (Figure 6) was similar between children with and without a cleft palate ($p = 0.575$), for reading (CP = 30.382 %MD; Control = 30.744 %MD) and spontaneous speech (CP = 33.242 %MD; Control = 33.978 %MD). Furthermore, the depth of inhalation tends to be greater when velopharyngeal functionality is severely affected. However, this difference is not significant ($p = 0.549$).
The speech task had a slight effect on inspiratory contribution. Inhalation depth tends to be greater in spontaneous speech for both groups of speakers.

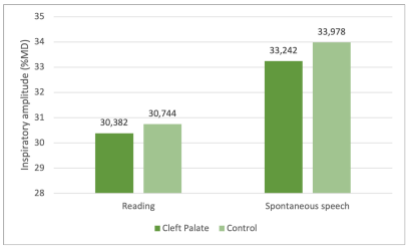


Figure 6: Inhalation contribution according to groups of speakers and speech tasks

3.2.3. Temporal organisation of speech: breath group duration analysis

In this section, we investigate the effect of compensatory respiratory strategies on the temporal organisation of speech using breath group analysis.
Breath groups tended to be shorter in children with a cleft palate than in control speakers. The difference between the two groups of speakers is not significant (p = 0.244). In reading, breath group durations in children with a cleft palate was measured at 2.15 seconds on average, and 2.39 seconds in control speakers. In spontaneous speech, the average duration of breath groups was 2.01 secs and 2.25 secs for children with a cleft palate and control speakers respectively.
While breath groups tend to be shorter in spontaneous speech, mixed-measures ANOVA revealed no significant effect of speech tasks on breath group durations (p = 0.948). Pearson's correlation coefficient indicates a partial correlation between the duration of the breath group and the expiratory contribution ($r = 0.567$; $p < 0.05$).

4. DISCUSSION

The purpose of this study was to examine compensatory strategies deployed at thoracic and abdominal levels by children with a cleft palate. Results show that these strategies were mainly deployed during the expiratory phase. The expiratory movements were greater in children with a cleft palate. Increase in the expiratory contribution allows expulsion of larger lung volumes to compensate for supraglottal deficits. This increase is more salient when velopharyngeal functionality is severely affected, like when the nasal airway emission is continual, involving a constant loss of intraoral pressure.
The expiratory contribution was higher during production of consonantal cycles containing voiceless fricatives than those containing plosive consonants. These data support those of Ohala [5], obtained from adults without speech disorders. According to Warren & Wood [15], fricative consonants need a large lung volume because their duration is generally longer than that of plosives. It should be mentioned that our results did not show any correlation between expiratory amplitudes and consonantal durations.
In reading of short sentences, whatever the target consonant, a statistically significant difference was observed in the expiratory contribution in children with a cleft palate, within the consonantal cycle. However, in text reading and spontaneous speech, these findings were confirmed only in terms of tendencies. While chest compression was also greater in children with a cleft palate, the difference was not statistically significant. These findings suggest that other factors (pragmatic, enunciative, contextual, etc.) may have an impact on the control of speech breathing.
Finally, expiratory strategies do not seem to have an effect on the temporal organisation of speech. While breath groups tended to be shorter in children with a cleft palate, this decrease does not seem to be related to an increase of the expiratory contribution during speech for these speakers. Moreover, the partial correlation between the expiratory amplitude and breath group duration suggests that other factors influence the temporal organisation of speech breathing.
As regards inhalation, no discernible difference was found between children with and without a cleft palate. Sporadically, the inspiratory movements were greater in children with severe velopharyngeal insufficiency. However, this strategy was not systematic and only appears in certain cases. The data as seen do not allow to explain the reason of such a strategy.
Given the use of expiratory strategies by children with a cleft palate, it is important that breathing be addressed during speech therapy in order to help children achieve optimal use of the breath for speech. Nevertheless, in some cases, the increase in respiratory effort can lead to an increase in audible nasal noise, which can impair speech intelligibility [7]. This study should be continued by observing the effects of these breathing strategies on sound, at the acoustic level, as well as on speech quality.

5. CONCLUSION

This study reveals compensatory respiratory strategies deployed by children with a cleft palate and their effects, to a lesser extent. Our data show that compensatory strategies are mainly deployed during expiration. Such strategies are more salient when velopharyngeal functionality is more severely affected. However, the results of reading and spontaneous speech tasks suggest that other factors are operational in the control of speech breathing.

6. REFERENCES

[1] Harding, A., Grunwell, P. 1996. Characteristics of cleft palate speech. *International Journal of Language & Communication Disorders*. 31, 331-357.

[2] Warren, D. W. (1987). Nasal emission of air and velopharyngeal function. *Cleft Palate J.* 4, 148-155.

[3] Klatt, K., Mead, J. 1968. Studies of articulatory activity and airflow during speech. *Annals of the New York Academy of Sciences*. 155, 42-55.

[4] Di Cristo, A., Teston, B. 1979. Fréquence fondamentale et pression intraorale. *Larynx et Parole.* 329-366.

[5] Ohala, J., Riordan, C., Kawasaki, H. 1980. Investigation of pulmonic activity in speech. *Investigation of pulmonic activity in Speech*. 5, 89-95.

[6] Laine, T., Warren, D. W., Dalston, R. M., Hairfield, W. M., Morr, K. E. 1989. Intraoral pressure, nasale pressure and airflow rate in cleft palate speech. *Journal of Speech, Language, and Hearing Research*. 31, 432-437.

[7] Warren, D. W., Drake, A. F., Davis, J. U. 1992. Nasal airway in breathing and speech. *The Cleft Palate-Craniofacial Journal*. 29, 511-519.

[8] Wang, Y.-T., Green, J. R., Nip, I. S., Kent, R. D., Kent, J. F. 2010. Breath group analysis for reading and spontaneous speech in Healthy adults. *Folia Phoniatrica et Logopaedica*. 62, 297-302.

[9] Winkworth, A. L., Davis, P. J., Adams, R. D., Ellis, E. 1995. Breathing patterns during spontaneous speech. *Phonetica*. 54, 138-152.

[10] Boliek, C. A., Hixon, T. J., Watson P. J., Jones, P. B. 2009. Refinement of Speech Breathing in Healthy 4- to 6-Year-Old Children. *Journal of Speech, Language, and Hearing Research.* 52(4), 990-1007.

[11] Hixon, T. J., Goldman, M. D., Mead, J. 1973. Kinematics of chest wall during speech production: volume displacements of the rib cage, abdomen, and lung. *Journal of Speech and Hearing research*. 16, 78-115.

[12] Boersma, P., Weenink, D. 2021. *Praat: Doing by computer* [Computer program]. Version 6.1.42

[13] Fuchs, S., Hoole, P., Vornwald, D., Gwinner, A., Velkov, H., Krivokapic, J. 2008. The control of speech breathing in relation to the upcoming sentence. *Proc. 8th International Seminar on Speech Production* Strasbourg, 77-80.

[14] Rochet-Capellan, A., Fuchs, S. 2013. Changes in breathing while listening to read speech: the effect of reader and speech mode. *Frontiers in Psychology*. 4, 906.

[15] Warren, D.W., Wood, M. T. 1969. Respiratory volumes in normal speech: a possible reason for intraoral pressure differences among voiced and voiceless consonants. *The Journal of the Acoustical Society of America*. 45, 466-469.

ACOUSTIC CHANGE OVER TIME IN SPEECH OF ONE BILINGUAL INDIVIDUAL WITH PARKINSON'S DISEASE

Vass Verkhodanova[1], Matt Coler[1], Roel Jonkers[1], Kees de Bot[2], Wander Lowie[1]

[1] University of Groningen, The Netherlands

[2]University of Pannonia, Hungary

v.verkhodanova@rug.nl, m.coler@rug.nl, r.jonkers@rug.nl, c.l.j.de.bot@rug.nl, w.m.lowie@rug.nl

ABSTRACT

Speech changes appear in most speakers with Parkinson's disease (PD), but very little is known about longitudinal speech changes in bilingual people with PD. The aim of this exploratory single case study is to explore such changes in two working languages of one individual with PD. For this purpose, we collected the time series of 36 recordings done over 3.6 years, with recording timeline including the pre-, concurrent, and post-speech therapy periods. We analysed recordings of different tasks performed in Dutch and English, examining a set of conventional speech features related to phonation, prosody and articulation. The results of the linear trend estimation demonstrate similar trends irrespective of the spoken language, while some of the measurements, e.g. speech rate, demonstrate more pronounced and significant trends in English only. Exploratory acoustic analysis also provided evidence of a speech therapy effect in both languages, though the therapy was administered only in Dutch.

Keywords: Parkinson's disease, acoustics, dysarthria, prosody, bilingualism

1. INTRODUCTION

Speech and communication changes are almost certain to happen for people with Parkinson's disease (hereafter, PwPD). It was reported that 80-90% of PwPD experience voice changes and 45-50% show articulation change due to hypokinetic dysarthria (HD) [1]. Studies on speech production in PwPD often focus on the conventional features that are clinically interpretable and can be correlated with HD auditory-perceptual assessments (for details on conventional and non-conventional speech features see [2]). Among the most described conventional speech features in HD are fundamental frequency (f_0) deviations associated with monopitch [3], distorted rhythm [4], and reduced intensity variability associated with monoloudness [3].

There are only a few studies that explore longitudinal speech changes in PwPD, most of which analysed recordings collected at two time points with intervals between them ranging from seven months [4] to 3.7 years [5]. The findings demonstrate reduction in f_0 variability [6], instability of steady syllable repetition [7], increased speech rate [5], deteriorations of voice quality and of articulatory velocity and precision [7]. In one case study describing a longitudinal speech analysis based on television archives over an 11-year period authors found that changes in f_0 variability can be detected as early as five years prior to diagnosis [8].

However, there is hardly any research done on bilingual PwPD not to mention longitudinal effects of PD on bilingual speech. This constitutes a big gap in existing literature, given that most people in the world are not monolingual and the increasing global mobility actively reinforces multilingualism. To our knowledge, only two studies focused on speech and language in bilingual PwPD [9, 10] and no publications reported the effects of speech therapy in bilingual PwPD. The authors of [9, 10] demonstrated that their group of PwPD had more phonological, morphological, and syntactic errors in their first language, Friulian, rather than in their second language, Italian.

This study addresses the literature gap by focusing on longitudinal acoustic changes in speech of a bilingual person with PD. The first goal is to explore if the acoustic changes are similar in two working languages of the bilingual participant with PD, Dutch and English. In the light of previous studies and despite the differences in phonological systems, we expected to see the typically occurring PD effects in speech acoustics (e.g., the decrease in f_0 variability, the rise in mean f_0) to become more prominent with time in both languages [11, 2]. The second goal is to explore if the general speech therapy effect is present in the measurements of acoustic features in either language, given that the speech therapy was administered in Dutch. Based on the therapy goals (see 2.1), we hypothesized that

effect would be present for the voice quality and f_0 measurements in both languages. Considering speaker's bilingualism and the complex character of linguistic transfer and feature interference [12], we did not expect to see speech therapy effects on any particular speech characteristic just in one language.

2. METHODS

The present study uses a time series design following one individual's speech over 3.6 years. The collection and analysis of the material was approved by the Ethical Committee of of University Groningen, Campus Fryslân.

2.1. Participant

One speaker with PD, a sequential bilingual, participated in the longitudinal data collection during 185 weeks. He reported to consistently use two languages on a daily basis: Dutch and English. He was diagnosed with idiopathic PD six years prior to the beginning of the data collection, at the age of 60. He has not been diagnosed with hypokinetic dysarthria, but has a history of developmental stuttering. During week 46 the participant reported having a cold. After the first 12 recording sessions, he began with speech therapy in Dutch that according to his speech therapist was largely based on the Lee Silverman Voice Treatment [11] with the main focus being on increasing vocal intensity, lowering pitch and increasing intelligibility of phonemic groups. During the next eight sessions the participant had ongoing speech therapy until he stopped by session 22. Throughout the whole period of recordings, the participant was on stable dopamine-replacement medication (Sinemet).

2.2. Recording procedure

Speech tasks in each session included sustained phonation of the vowel /a:/, interview with an open question, description of one of the Heaton pictures [13] and of a short video clip taken from the freely available works of Charlie Chaplin, and reading of the 'North Wind and the Sun' passage. All tasks were performed first in English and subsequently in Dutch, with instructions provided both orally and with the slide show. The first two sessions lacked video description, and session 31 lacked picture description due to technical reasons. The sessions happened roughly every month (mean interval is 5.3 weeks, SD = 2.6 weeks) from one to three hours after medication intake. Based on the previous research [14], we did not expect time of the day or moment of medication to influence our participant's speech. The first 28 recording sessions took place in quiet rooms with the Zoom H2 recorder placed at around a 40 cm distance. Starting with session 29 the sessions took place mostly via Zoom calls due to COVID-19 restrictions and the participant recorded himseld with the Dictaphone app on his iPhone 10 placed at around a 40 cm distance. The reliability of the data collected during online sessions was ensured by comparison of the quality and acoustic measurements of two sets of recordings collected simultaneously with iPhone 10 and Zoom H2 during two offline sessions.

2.3. Acoustic measurements

We focused on conventional features characteristic of phonation, prosody, and articulation aspects of speech affected by PD as captured by prolonged phonation, monologue and reading tasks [2]. All measurements and subsequent analyses were performed separately for the different tasks to avoid losing task-related information [15].

We included five features from the domain of prosody. For the fundamental frequency (f_0), f_0 means and f_0 variability (f_0 coefficients of variation) calculation we used a Python f_0 tracking script based on the robust algorithm for pitch tracking (RAPT) [16] implemented in the Speech Signal Toolkit [17]. Speech rate and articulation rate were calculated with a Praat script [18]. Speech rate was measured as the number of syllables divided by the total time of the recording, articulation rate – as the number of syllables divided by phonation time. Feature of inappropriate silences was calculated from the results of the same Praat script [18] as the number of pauses relative to total speech time after removing periods of silence lasting less than 60 ms [2]. For phonation and voice quality, we included three features: maximum phonation time (MPT), jitter and harmonics-to-noise ratio (HNR) measured from the prolonged phonation (/a:/) recordings. We also included two features related to articulation: the means of first two formants (F1 and F2) calculated from the prolonged phonation recordings. Because the recordings in the current study were done without strict supervision over the distance between speaker and the microphone, the intensity measurements had to be excluded from the analysis.

2.4. Statistical analyses

Given the study's exploratory focus and the time series nature of the data, we estimated linear trends

for the selected speech features using simple linear regression on the dataset after excluding one session where the participant had residual effects of a cold (week 46). To determine whether the results of the linear regression analyses were not random we applied the Monte Carlo simulation by randomizing the measurements 1000 times and calculating the slope for every randomized set of measurements. We evaluated each resulted distribution of slopes by calculating its mean, SD and standard error.

To explore the general absence or presence of the therapy effect on each feature we analysed the dataset after deducting general linear trends from each data point, as such trends could reflect both the progression of PD and/or the general effect of ageing [19]. After the trend deduction, based on Levene's test results we conducted either non-parametric Welch's F tests or one-way ANOVAs followed up with Tukey post-hoc tests to compare measurements 'before' and 'during' as well as 'during' and 'after' therapy. Additionally, with Kolmogorov-Smirnov two-sample test we compared therapy measurements for the absolute values of f_0 to avoid any information loss by aggregating features per session.

3. RESULTS

3.1. Prolonged phonation

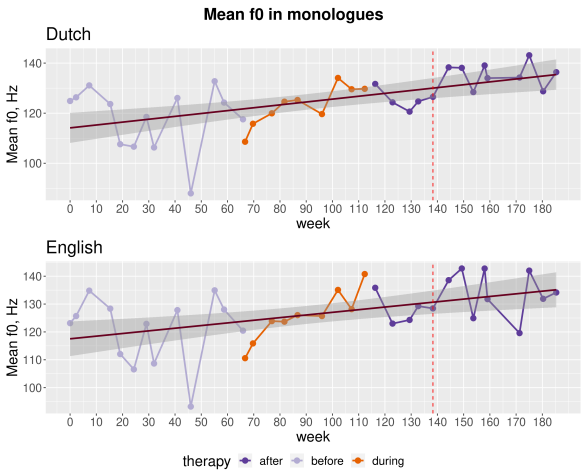
Analyses of HNR and Jitter showed no significant decline or rise. However, MPT demonstrated a significant decline: $R^2 = 0.15, F(1,33) = 5.91, p = 0.02$. The Monte Carlo analysis confirmed that it was not random. Changes of mean f_0 and F1 showed no trend, while the significant rises were present in f_0 ($R^2 = 0.5, F(1,33) = 32.76, p < .001$) and for F2 ($R^2 = 0.2, F(1,33) = 8.42, p = 0.007$). Monte Carlo analysis showed that these were also not random. After correcting for multiple comparisons with the Bonferroni method, the only significant results were present in f_0 ($p < .001$) and F2 ($p = 0.039$) trends.

Analysis of therapy effect yielded no significant results for any of the above features. For the absolute f_0 values, the result of the Kolmogorov-Smirnov test indicated that both pairs of the distributions were significantly different ('before'-'during': $D = 0.38$, $p < .001$,'during'-'after': $D = 0.16$, $p < .001$).

3.2. Spontaneous speech

Linear regression analysis resulted in significant trends for f_0 means over time in both languages (Dutch: $R^2 = 0.34, F(1,34) = 17.52, p < .001$; English: $R^2 = 0.25, F(1,34) = 11.15, p = 0.002$). Monte Carlo analysis confirmed that the results were not random. Bonferroni adjustment for f_0 measurements showed the preserved significance for mean f_0 both in Dutch ($p = 0.002$) and in English ($p = 0.008$).

Figure 1: f_0 means over time in two languages.



In the comparison of the therapy stages we tested the absolute values of f_0 and found no significant differences in f_0 variability or f_0 means. The results of Kolmogorov-Smirnov test for f_0 values at different stages of the therapy in Dutch and in English were found to be significant for comparisons of 'before'-'during' (for both languages, $D = 0.1$, $p < .001$) and 'during'-'after' (for both languages, $D = 0.1$, $p < .001$). Analysis of number of inappropriate pauses yilded no significant resluts.

3.3. Reading

Speech and articulation rates showed no significant trend over time. The analysis of the therapy effect on the rate measurements demonstrated one significant difference – for speech rate in English reading. An ANOVA test showed the differences between stages were significant, $F(2,32) = 4.37, p = 0.021$. A post hoc Tukey test showed the significant difference between 'before' and 'during' therapy stages ($p = 0.024$) and 'during' and 'after' therapy stages ($p = 0.04$). The differences in means in post hoc analysis demonstrated that mean value of speech rate was lowest during therapy. Linear regression analysis showed significant trends in f_0 variability over time for English reading ($R^2 = 0.17, F(1,34) = 7.07, p = 0.012$) and in mean f_0 in both Dutch and English reading (Dutch: $R^2 = 0.26, F(1,33) = 11.53, p = 0.002$; English: $R^2 = 0.22, F(1,33) = 9.51, p = 0.004$). Monte Carlo

analysis confirmed that the results were not random. The Bonferroni adjustment for f_0 measurements showed the preserved significance for mean f_0 both in Dutch ($p = 0.002$) and in English ($p = 0.008$), and for f_0 variability in English ($p = 0.004$).

Analysis of therapy effect showed no significant differences in mean values of f_0 while there were significant differences in f_0 variability in English reading. One-way ANOVA yielded $F(2,32) = 6.18, p = 0.005$. The post hoc Tukey test showed the significant difference between 'before' and 'after' therapy stages ($p = 0.004$) with f_0 variability in English reading being lower after therapy. Once again, we tested the absolute values of f_0, the results of Kolmogorov-Smirnov test were found to be significant for comarisons of 'before'-'during' (for both languages, $D = 0.11$, $p < .001$) and 'during'-'after' (for both languages, $D = 0.13$, $p < .001$). Analysis of number of inappropriate pauses yilded no significant resluts.

4. DISCUSSION

The principal goal of this single case study was to explore changes in two languages, Dutch and English, in the time series of 36 recordings done over 3.6-years in a bilingual speaker with PD. We explored 10 conventional speech features measured from prolonged phonation, monologue and reading tasks. Among features related to articulation and phonation, we found significant changes in MPT and F2 with only F2 changes remaining significant after the Bonferroni adjustment. Nevertheless, the observed changes in MPT measure, characterizing the aerodynamic efficiency of the vocal tract, are in line with the descriptions of the clinical picture of HD related to reduction in the total amount of air expended during maximum phonation in PwPD [11]. The significant falling trend in F2 suggests presence of the centralization effects progression in vowels, which might be more apparent with respect to the reduced tongue movements as there were no significant changes in F1 values. The significance of a raising trend in mean f_0 with time is also in line with the changes caused by PD [11] and has also been noted by [19] in their elderly male group.

In the recordings of two other tasks, monologues and reading, we found significant trends in f_0 means for both languages, which is consistent with our findings for prolonged phonation. There were no significant trends in speech or articulation rates or in inappropriate silences contradicting the findings of [6] who found speech rate decline in male group with PD. A significant trend in f_0 variability was present in English reading, while no significant changes were found in Dutch reading and monologues. This larger range in variability values in Dutch reading throughout the sessions can possibly be explained by more comfortable intoning of read text in the speaker's first language. These findings are partially in line with the previous study by [6], who found no significant change in f_0 variability in PwPD with time. Yet, the presence of a significant trend in English reading requires further investigation. The lack of significant trends for speech and articulation rates, or for the number of inappropriate silences in both languages appears to be a positive sign of their stability in the disease progression. This, however, is indirectly against the findings of [9, 10], who demonstrated higher mistake rates in phonology, morphology and syntax in the speakers' first language. The presence of mistakes on higher levels, such as syntax and morphology, potentially could have correlated with the amount of pauses or rate changes over time. Regarding the effect of speech therapy, with the removal of the general trend we found significant differences for f_0 in periods with and without therapy for all tasks in both languages of the participant. Interestingly, we also found significant differences between therapy stages for speech rate and for f_0 variability in reading, but only for English. These findings suggest the therapy effect transfer. Literature does not provide any information into speech therapy effects in bilingual PwPD, and our results are exploratory.

Our study has several limitations. First, being a single case study it prevents us from making any generalizations. Second, it has an exploratory nature: we analysed the presence of linear trends and compared distribution shapes of our data. Thus, while no significant linear trends were found for some features, we cannot be sure there is no change present. In the future research, exploring non-linearity of the time-series data and expanding the list of features may provide additional insights into the nature of acoustic change and therapy effects in both languages. Contrary to the common pretest-posttest designs, the results of studies with frequent measurements could paint a more detailed picture of the speech disorder progression providing speech therapists with important information about the 'process' rather than the 'product' of speech change or therapy effects.

5. REFERENCES

[1] N. Miller, E. Noble, D. Jones, and D. Burn, "Life with communication changes in Parkinson's

disease," *Age and ageing*, vol. 35, no. 3, pp. 235–239, 2006.

[2] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in parkinson's disease: early diagnostics and effects of medication and brain stimulation," *Journal of neural transmission*, vol. 124, no. 3, pp. 303–334, 2017.

[3] S. Skodda, W. Grönheit, N. Mancinelli, and U. Schlegel, "Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study," *Parkinson's disease*, vol. 2013, 2013.

[4] S. Skodda and U. Schlegel, "Speech rate and rhythm in Parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 7, pp. 985–992, 2008.

[5] J. E. Huber and M. Darling-White, "longitudinal changes in speech breathing in older adults with and without Parkinson's disease," in *Seminars in speech and language*, vol. 38, no. 03. Thieme Medical Publishers, 2017, pp. 200–209.

[6] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in parkinson's disease over time: a longitudinal study," *Movement disorders: official journal of the Movement Disorder Society*, vol. 24, no. 5, pp. 716–722, 2009.

[7] S. Skodda, A. Flasskamp, and U. Schlegel, "Instability of syllable repetition as a marker of disease progression in Parkinson's disease: a longitudinal study," *Movement disorders*, vol. 26, no. 1, pp. 59–64, 2011.

[8] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and cognition*, vol. 56, no. 1, pp. 24–29, 2004.

[9] S. Zanini, A. Tavano, L. Vorano, F. Schiavo, G. L. Gigli, S. Aglioti, and F. Fabbro, "Greater syntactic impairments in native language in bilingual Parkinsonian patients," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 12, pp. 1678–1681, 2004. [Online]. Available: http://dx.doi.org/10.1136/jnnp.2003.018507

[10] S. Zanini, A. Tavano, and F. Fabbro, "Spontaneous language production in bilingual Parkinson's disease patients: Evidence of greater phonological, morphological and syntactic impairments in native language," *Brain and language*, vol. 113, no. 2, pp. 84–89, 2010. [Online]. Available: https://doi.org/10.1016/j.bandl.2010.01.005

[11] L. O. Ramig, C. Fox, and S. Sapir, "Parkinson's disease: speech and voice disorders and their treatment with the lee silverman voice treatment," in *Seminars in speech and language*, vol. 25, no. 02. Copyright© 2004 by Thieme Medical Publishers, Inc., 2004, pp. 169–180.

[12] F. Grosjean, "An attempt to isolate, and then differentiate, transfer and interference," *International Journal of Bilingualism*, vol. 16, no. 1, pp. 11–21, 2012.

[13] J. B. Heaton, *Composition through pictures*. Longman Group United Kingdom, 1966/1972.

[14] J. Jacobi, T. Rebernik, R. Jonkers, M. Proctor, B. Maassen, and M. Wieling, "The effect of levodopa on vowel articulation in Parkinson's disease: A cross-linguistic study," in *19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association (ASSTA)*, 2019, pp. 1069–1073.

[15] J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, V. Majerová, J. Picmausová, J. Roth, and E. Růžička, "Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy," *Journal of Neural Transmission*, vol. 120, no. 2, pp. 319–329, 2013.

[16] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[17] "Sptk: The speech signal processing toolkit (version 3.11)," http://sp-tk.sourceforge.net/.

[18] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[19] A. Pessin, E. Tavares, A. Gramuglia, L. De Carvalho, and R. Martins, "Voice and ageing: clinical, endoscopic and acoustic investigation," *Clinical Otolaryngology*, vol. 42, no. 2, pp. 330–335, 2017. [Online]. Available: https://doi.org/10.1111/coa.12725

ATYPICAL COARTICULATION IN AUTISM: EVIDENCE FROM SIBILANT-VOWEL INTERACTION IN CANTONESE

Alan C. L. Yu[1], Robert McAllister[1], Carol K. S. To[2]

[1]University of Chicago; [2]University of Hong Kong
aclyu@uchicago.edu; robertatlas@uchicago.edu; tokitsum@hku.hk

ABSTRACT

Atypicalities in the prosodic aspects of speech are commonly considered in clinical assessments of autism, a neurodevelopmental disorder involving difficulties in social development and communication alongside the presence of unusually strong repetitive behavior or 'obsessive' interests. While there is an increasing number of studies using objective measures to assess prosodic deficits, such studies have primarily focused on the intonational and rhythmic aspects of prosody. Little is known about prosodic deficits that are reflected at the segmental level, despite the strong connection between prosody and segmental realization. This study examines the nature of sibilant-vowel coarticulation among male adult native speakers of Cantonese with autism and those without. While neurotypical participants exhibit sibilant-vowel coarticulation that are sensitive to variation in sibilant duration, participants with autism show no sensitivity to segmental temporal changes. These findings point to the potential for atypicalities in prosody-segment interaction as an important characteristic of autistic speech.

Keywords: coarticulation, sibilant, duration, autism

1. INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder involving difficulties in social development and communication alongside the presence of unusually strong repetitive behavior or 'obsessive' interests [1, 2]. Atypicalities in the prosodic aspects of speech are considered a core feature of the disorder (e.g., [3] and prosody is commonly included as a diagnostic characteristic in clinical assessments of autism [4]. Prosodic impairments are often broadly defined as deviations in speaking rate, rhythm, volume, intonation, or the inability to change language register based on the interlocutor. Individuals with ASD are often characterized with terms such as flat/monotonous, variable, sing-songy, and/or pedantic, among others (e.g., [5, 6]). While there is an increasing number of studies using objective measures to assess prosodic deficits, these studies have primarily focused on the intonational and rhythmic aspects of prosody, even though there exists strong evidence for significant interplay between articulation and different levels of linguistic structures. For example, in autosegmental metrical models of intonation, segments are organized into prosodic structures, such as syllables, feet, words, and higher phrases, as well as prominence relations in some languages (e.g., [7]). Prosodic structure is expected to influence the phonetic implementation of sound categories, and fine-grained phonetic detail, in turn, can inform higher-level prosodic structure [8]. Yet, little is known regarding prosodic deficits that may be reflected in the phonetic-prosody interface [9].

This study focuses on the nature of coarticulation in speech. Coarticulation refers to the articulatory influence between neighboring segments during speech production. The temporal extent and magnitude of coarticulation are dependent on prosodic factors of the utterance, such as speaking rate, degree of emphasis, and overall rhythm. Given that impaired temporal processing has been argued to be a source of prosodic impairments in ASD [10], we hypothesize that individuals with ASD should exhibit atypicalities in coarticulatory patterns. Specifically since temporal processing impairment not only impacts the perception of temporal variation, but also the timing accuracy in motor planning and speech production [11], we hypothesize that individuals with ASD might have difficulties with coarticulatory planning or have difficulties processing coarticulated speech [12], which in turn may lead to atypical coarticulatory patterns during production. Recent studies have found that neurotypical autistic traits are predictive of individual variation in the production and perception of coarticulated speech [13, 14], further strengthening the potential link between ASD and coarticulatory atypicalities. To investigate this link more closely, this study examines how duration variation influences the nature of sibilant-vowel coarticulation among adult male Cantonese speakers

with autism and those without.

2. METHODS

2.1. Participants

The ASD cohort consisted of fifteen Cantonese-speaking adult males with autism with ages ranging from 18 to 33, with a mean of 25.51 (SD = 3.51). All the participants were recruited from employment programs particularly designed for young adults who have been diagnosed with high functioning ASD. The programs are run by two local non-governmental organizations in Hong Kong. ASD diagnosis was based on Diagnostic and Statistical Manual of Mental Disorders, third edition (DSM-III) (Association, 1980) criteria and International Classification of Diseases, 10th revision (ICD-10; [15]) by either a clinical psychologist or a pediatrician during their childhood. The current state of ASD was verified by the clinical judgment of the third author who is a speech-language pathologist with ASD expertise, and the Autism Diagnostic Observation Schedule (ADOS-2) [4] administered by research-reliable personnel, with a total score at or above the thresholds of autism or autism spectrum for Module 4. All particiants' hearing ability was screened with a GSI 18 screening audiometer in a sound-proofed room, with the passing criteria set at 25 dB HL at the frequencies of 1000, 2000 and 4000 Hz in both ears [16]. This study focuses only on male participants due to difficulty in recruiting female ASD participants in Hong Kong. All ASD participation received a nominal fee for their participation.

The neurotypical (NT) cohort included twenty-three male adults in Hong Kong, all native speakers of Hong Kong Cantonese, who completed this study either for course credits or a nominal fee. Their age range was 18 to 26 with a mean of 20.13 (SD = 2.26). None reported any language, speech, or hearing disorders nor any mental illness.

2.2. Stimuli

The stimuli consisted of fourteen target words in Cantonese: [sy˥] "book,", [sy˩], "potato," [sɔ˥] "comb," [sɔ˩] "silly," [si˥] "poem," [si˩] "time," [se˥] "a little bit," [se˩] "snake," [sa˥] "sand," [san˩] "god," [sau˥] "to collect," [sau˥] "gloomy," [sou˥] "tassel," and [sou˨] "mob." The symbol ˥ marks a high level tone, ˨ a low tone, and ˩ a low falling tone.

2.3. Procedure and stimuli

Each participant was digitally recorded in a quiet room individually at a sampling rate of 44,100 Hz reading three blocks of the target stimuli, presented in traditional Chinese characters in one of two pseudo-randomized lists of target words in the carrier sentence, [ŋɔ˧ jiu˨ tʊk˨] ___ [pei˧ nei˧ tʰeŋ˥] "I read ___ for you to hear." A total of forty-two target stimuli were analyzed from each participant. All subjects also completed an online survey which included questions about the subject's age, sex, second language knowledge, and the full 50-question autism-spectrum quotient (AQ) questionnaire [17]. Participants were given the option to complete either the Chinese or English version of the AQ. The recordings were segmented automatically with WebMaus [18], a web service that forced-aligns an audio recording based on a corresponding orthographic transcript. German served as the language processing model as it yielded the closest alignment for these recordings. The results of the forced alignment of the target words at the segmental and word levels were manually checked and corrected where necessary.

Spectral means were extracted from syllable-initial /s/s in all target words in the corpus, using a modified version of [19], a Praat script that calculates spectral mean based on the fricative noise between the initial and final 10% of the sibilant duration high-pass filtered with a cut-off of 300 Hz using the time-averaging method [20]. Spectral mean has been shown to distinguish well between /s/ and /ʃ/ in English [21, 22] and across different vowel contexts [23, 14]. Speaking rate was calculated in terms of the number of words per second based on the average word duration between the target word and the words preceding and following the target.

3. RESULTS

The median AQ score for the ASD cohort was 132 (SD=16.27) and 121 for the NT cohort (SD=11.46). The mean sibilant duration was 197 ms and the mean speaking rate is 3.75 word per second (wps). Exploratory analyses found no significant difference between cohorts in speaking rate (ASD: mean [SD] = 3.57 [0.84] wps; NT: mean [SD] = 3.86 [0.53] wps) or sibilant duration (ASD: mean [SD] = 200.64 [36.69] ms; NT: mean [SD] = 195 [30.29] ms).

The spectral mean of /s/ was modeled using linear mixed-effects regression fitted in `R`, using the `lmer()` function from the `lme4` package [25]. This study focused on the effect of vocalic rounding on /s/ as it is well-documented [14]. The regression

	Model 1	Model 2
Intercept	7045.27***	7092.92***
	(137.46)	(136.94)
R(OUN)D	−367.97***	−359.95***
	(58.55)	(56.87)
COHORT	−204.98	
	(135.24)	
DUR(ATION)	1.10*	1.39**
	(0.51)	(0.53)
RATE	−165.95***	−165.94***
	(42.01)	(44.03)
RD:COHORT	−38.22	
	(57.33)	
RD:DUR	−0.80	−0.96*
	(0.45)	(0.45)
DUR:COHORT	−1.05*	
	(0.51)	
RD:DUR:COHORT	1.12*	
	(0.45)	
AQ		−29.35
		(135.33)
RD:AQ		81.29
		(54.74)
DUR:AQ		−0.20
		(0.54)
RD:DUR:AQ		0.96*
		(0.46)

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Table 1: Regression models of the spectral mean. The p-values were obtained using normal approximation, which assumes that the t distribution converges to the z distribution as degrees of freedom increase [24].

model tested for the effects of vocalic ROUNDing (rounded vs. unrounded), COHORT (ASD vs. NT), sibilant DURATION, and speaking RATE. The model also included by-subject random intercepts as well as by-subject random slopes for RATE, residualized DURATION, and ROUND, to allow for subject-specific variation with respect to these variables. COHORT was sum-coded and speaking rate centered and z-scored. Because of strong correlation between sibilant duration and speaking rate, sibilant duration was first residualized for the effect of speaking rate prior to being entered into the regression analysis. Model selection began with all interactions between COHORT, ROUND, and the temporal measures, i.e., DURATION and RATE. Interactions that did not improve model-likelihood based on a series of likelihood ratio tests were eliminated.

The formula for the final model (Model 1 in Table 1) in `lme4` style is SPECTRAL MEAN ~ ROUND * DURATION * COHORT + RATE + (1 + ROUND + DURATION + RATE |SUBJECT). As expected, there was a significant main effect of ROUND with spectral mean being lower before rounded vowels than before unrounded ones (β = -367.97, t-value = -6.29, $p < 0.001$). There was also a main effect of RATE (β = -165.95, t-value = -3.95, $p < 0.001$); the faster the speaking rate the lower the spectral mean. While there was a significant main effect of residualized sibilant duration (β = 1.10, t-value = 2.16, $p < 0.05$), it interacted significantly with COHORT (β = -1.05, t-value = -2.07, $p < 0.05$). As shown in Figure 1, while the ASD cohort showed a relatively stable spectral mean regardless of sibilant duration, the NT cohort exhibited a positive association between sibilant duration and spectral mean. That is, the longer the sibilant, the higher the spectral mean for the NT cohort.

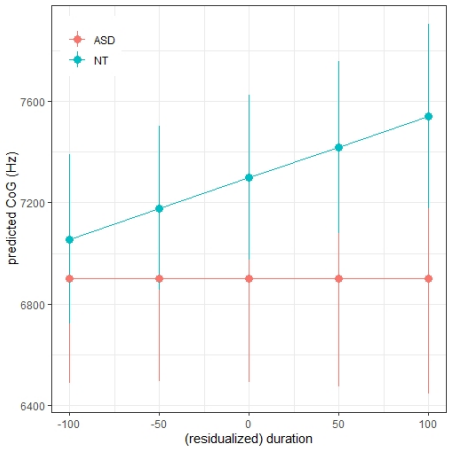


Figure 1: Effects of residualized duration sibilant spectral mean across the ASD and NT cohorts

Concerning the effect of vocalic rounding on sibilant realization, while there was not a significant interaction between ROUND and COHORT and between ROUND and DURATION, there was a significant three-way interaction between ROUND, DURATION, and COHORT (β = 1.12, t-value = 2.49, $p < 0.05$). As illustrated in Figure 2, the vocalic rounding effect on sibilant spectral mean was relatively stable for the ASD participants regardless of sibilant duration. For the NT participants, however, the vocalic rounding effect was larger the longer the sibilant. These findings suggested that, while NT speakers were sensitive to the temporal profile of the sibilant in planning the vocalic rounding influence, the ASD participants exhibited no such sensitivity. To be sure, the realization of /s/ among the ASD participants was sensitive to the roundness of the following vowel, the invariance of the context-dependent realization

of /s/ suggested that the ASD participants might be treating the vocalic rounding influence on sibilant realization as a categorical phenomenon, while the NT participants treated the rounding effect on sibilant realization as more gradient and more temporally dependent. These findings seemed to be at odds with previous report concerning the effects of autistic-like traits on sibilant-vowel interaction. [14], for example, reported that Cantonese-speaking individuals with higher AQ, especially male ones, exhibit less vocalic influence in the realization of sibilants. Given that individuals with autism tend to score higher on the AQ than NT individuals, as it is the case in our sample, we might expect, *a priori*, that the ASD participants would show less vocalic influence on their sibilants than the NT participants. To be sure, [14] did not take into account of the effects of duration variation on the vocalic rounding influence on sibilant realization. Thus it is unclear how autistic-like traits might mediate such an interaction. We focused on this question in the second regression model.

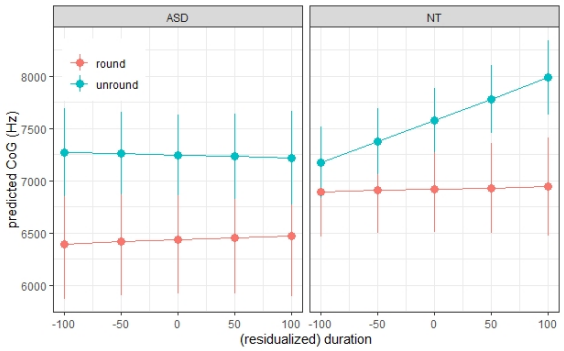


Figure 2: Three-way interaction between ROUND, DURATION and COHORT.

The second model included all the same dependent variables except that, instead of COHORT, the participants' AQ scores were entered. Model selection proceeded the same way as the first model. The formula for the final model in `lme4` style is SPECTRAL MEAN ~ ROUND * DURATION * AQ + RATE + (1 + ROUND + DURATION + RATE |SUBJECT) + (1|WORD). Like in the first model, there were significant main effects of ROUND and DURATION. There was a significant interaction between ROUND and DURATION (β = -0.96, t-value = -2.13, $p < 0.05$), indicating that the rounding effect on spectral mean increased as sibilant duration increased. While there was no main effect of AQ nor its interaction with ROUND or with DURATION, there was a significant three-way interaction between ROUND, DURATION, and AQ (β = 0.96, t-value = 2.08, $p < 0.05$), however. As shown in Figure 3, the higher the AQ, the smaller the effect of sibilant duration on the vocalic effect on spectral mean.

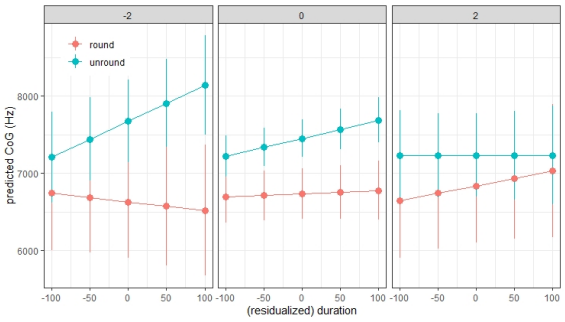


Figure 3: Three-way interaction between ROUND, DURATION and AQ. The three panels correspond to the mean AQ (middle), and AQ scores that are two standard deviations above (right) and below (left) the mean.

4. DISCUSSION AND CONCLUSION

This study investigated the effect of sibilant duration on spectral mean as conditioned by the roundness of the following vowel among male Cantonese speakers with autism and those without. Our findings suggested that the ASD cohort exhibited no sensitivity to sibilant duration variation in producing sibilant-vowel coarticulation, while the NT cohort adjusted the magnitude of sibilant-vowel coarticulation depending on how long the sibilant was. Since sensitivity to prosodic and duration variation in realizing fine-grain phonetic difference in segmental contrast is argued to reflect speaker knowledge of controlled phonetic variation [26], the vocalic rounding effect on sibilant reflected phonetic knowledge NT speakers had. However, unlike the NT speakers, the ASD speakers showed no such temporal sensitivity, even though they exhibited spectral mean variation conditioned by vocalic rounding. These findings suggested that the ASD speakers might have internalized the vocalic influence as a categorical context-dependent allophonic variation, while the NT speakers treated the vocalic influence as gradient and is adjusted relative to the temporal dynamic of speech production. Our findings also pointed to temporal sensitivity in the realization of phonetic contrast as a potential indicator in ASD diagnosis. More research is needed to ascertain the generality of the temporal insensitivity in speech production among ASD speakers.

5. REFERENCES

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 5th ed. Washington, DC: American Psychiatric Association, 2013.

[2] I.C.D-10, Ed., *International Classification of Diseases*, 10th ed. Geneva, Switzerland: World Health Organisation, 1994.

[3] H. Tager-Flusberg, "Understanding the language and communicative impairments in autism," *International Review of Research in Mental Retardation*, vol. 23, pp. 185–205, 2000. [Online]. Available: https://doi.org/10.1016/S0074-7750(00)80011-7

[4] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. Bishop, *Autism Diagnostic observation schedule, 2nd edition*. Western Psychological Services, 2012.

[5] H. Amoroso, "Disorders of vocal signaling in children," in *Nonverbal vocal communication: Comparative and developmental approaches*, H. Papousek, U. Jurgens, and M. Papousek, Eds. Cambridge: Cambridge University Press, 1992, pp. 192–204.

[6] C. Lord, M. Rutter, and A. Lecouteur, "Autism Diagnostic Interview–Revised–a revisedversion of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders." *Journal of Autism and Developmental Disorders*, vol. 24, pp. 659–685, 1994.

[7] M. E. Beckman, "The parsing of prosody," *Language and Cognitive Processes*, vol. 11, pp. 17–67, 1996.

[8] T. Cho, J. McQueen, and E. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," *Journal of Phonetics*, vol. 35, pp. 210–243, 2007.

[9] T. Cho, "Laboratory phonology," in *The Continuum companion to phonology,*, N. C. Kula, B. Botma, and K. Nasukawa, Eds. London, New York: Continuum, 2011, pp. 343–368.

[10] M. Allman and C. Falter, "Abnormal timing and time perception in autism spectrum disorder?: A review of the evidence," in *Time Distortions in Mind: Temporal Processing in Clinical Populations*, A. Vatakis and M. Allman, Eds. Leiden; Boston: Brill, 2015, pp. 37–56. [Online]. Available: https://www.jstor.org/stable/https://doi.org/10.1163/j.ctt1w8h2wk.7

[11] K. Franich, H. Y. Wong, A. C. L. Yu, and C. K. S. To, "Temporal coordination and prosodic structure in autism spectrum disorder: Timing across speech and non-speech motor domains," *Journal of Autism and Developmental Disorders*, vol. 51, pp. 2929–2949, 2020. [Online]. Available: https://doi.org/10.1007/s10803-020-04758-z

[12] A. C. L. Yu and C. K. S. To, "Atypical context-dependent speech processing in autism," *Applied Psycholinguistics*, pp. 1–15, 2020.

[13] A. C. L. Yu, "Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change," *PLoS One*, vol. 5, no. 8, p. e11950, 2010. [Online]. Available: doi:10.1371/journal.pone.0011950.

[14] ——, "Vowel-dependent variation in Cantonese /s/ from an individual-difference perspective," *Journal of Acoustical Society of America*, vol. 139, no. 4, pp. 1672–1690, 2016.

[15] W. H. Organization, *International Classification of Diseases (10th revision)*. Geneva: World Health Organization, 1990.

[16] American Speech-Language-Hearing Association Audiologic Assessment Panel 1996, *Guidelines for audiologic screening*. Rockville, MD: American Speech-Language-Hearing Association Audiologic Assessment Panel 1996, 1997.

[17] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males, females, scientists and mathematicians," *Journal of Autism & Developmental Disorders*, vol. 31, pp. 5–17, 2001.

[18] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[19] C. DiCanio, "Spectral moments of fricative spectra script in praat," 2013. [Online]. Available: https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_2.0.praat

[20] C. H. Shadle, "On the acoustics and aerodynamics of fricatives," in *The Oxford handbook of laboratory phonology*, A. C. Cohn, C. Fougeron, M. K. Huffman, and M. E. L. Renwick, Eds. Oxford: Oxford University Press, 2012, pp. 511–526.

[21] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[22] C. H. Shadle and S. Mair, "Quantifying spectral characteristics of fricatives," in *ICSLP 96. Proceedings of the Fourth International Conference on Spoken Language Processing*. IEEE, 1996, pp. 1521–1524.

[23] S. Nittrouer, "Children learn separate aspects of speech production at different rates: Evidence from spectral moments," *Journal of the Acoustical Society of America*, vol. 97, pp. 520–530, 1995.

[24] D. Mirman, *Growth Curve Analysis and Visualization Using R*. Boca Raton, Florida: Chapman and Hall / CRC, 2014.

[25] D. Bates, M. Maechler, and B. Bolker, *lme4*, 2011, r package version 0.999375-38.

[26] M.-J. Solé, "Controlled and mechanical properties in speech: a review of the literature," in *Experimental Approaches to Phonology*, M.-J. Solé, P. S. Beddor, and M. Ohala, Eds. Oxford: Oxford University Press, 2007, pp. 302–321.

AGING EFFECTS ON SENTENCE-LENGTH VELOCITY DISTRIBUTION CHARACTERISTICS OF THE TONGUE AND JAW

Claudia Raines[1], Tabea Thies[2,3], Angela Ding[1], Antje Mefferd[1]

[1]Vanderbilt University Medical Center, Nashville, TN USA
[2]University of Cologne and [3]University Hospital Cologne, Cologne, Germany
claudia.j.raines@vanderbilt.edu, tthies1@uni-koeln.de, angela.y.ding@vanderbilt.edu, antje.mefferd@vanderbilt.edu

ABSTRACT

This study aimed to identify aging effects on articulatory performance of the posterior tongue and jaw to better understand physiological factors that contribute to the slowing of speaking rate in older adults. Tongue and jaw movements of 15 younger and 15 older speakers were recorded during sentence repetitions using electromagnetic articulography. Based on each articulator's sentence-length velocity signal, histograms of velocity values were generated to determine direction-specific performance differences between groups. The following characteristics of the resulting velocity distributions were determined for each utterance and each speaker: mean, median, interquartile range, range, minimum, maximum, skewness, and kurtosis. Between-group comparison yielded significant differences for median and skewness with median values being greater and skewness values being more negative in younger speakers than older speakers. These findings suggest that younger speakers spend less time to execute tongue and jaw lowering (opening) movements than older adults. Potential factors underlying such group differences were discussed.

Keywords: speech kinematics, aging effects, speaking rate, articulatory control

1. INTRODUCTION

Older adults speak slower than younger adults [1-5]. In addition to cognitive-linguistic factors, aging-related physiologic changes within the speech motor system are thought to contribute to the slowed speaking rate in older adults. However, so far only few studies have systematically investigated changes in speech motor performance due to aging. Therefore, the articulatory mechanisms that may contribute to the slowing of speaking rate remain poorly understood. This knowledge gap creates challenges for the accurate differentiation between aging- and disease-related speech changes because many neurological conditions that manifest as a speech disorders (i.e., dysarthria) occur in older adults, are subtle at their onset, and mimic aging-related changes, such as a slowing of speaking rate [6,7].

It is often assumed that a deteriorating ability to generate adequate velocities underlies a slowing of rate in older adults. However, a previous study that has examined the ability to increase speed in younger and older adults could not provide evidence of aging-related decline in speed performance [8]. Nevertheless, in one of few kinematic studies that examined aging effects on articulatory performance during sentence productions, average and peak speeds were found to be significantly reduced in older speakers [9]. This finding was interpreted to suggest that the slowing of speaking rate and the reduced speeds may be an articulatory strategy older speakers implement to maintain articulatory precision.

One major problem with speech kinematic studies is that measures such as duration, amplitude, and velocity are either based on well-defined, isolated movement segments (a raising or lowering movement associated with a specific phoneme production) [3] or they are averaged across sentence-length utterances [9]. While findings of segment-based studies are difficult to generalize due to known context effects, findings based on sentence-length measures cannot pin-point potential direction-specific impairment patterns [9]. Yet, raising movements (approaching vocal tract constrictions) are thought to be more carefully controlled than lowering movements (releasing vocal tract constrictions) [10] and direction-specific performance changes in older speakers may help delineate articulatory control mechanisms that underlie the slowing of speaking rate.

In this current study, we developed a new approach to examine aging effects on direction-specific articulatory performance. Specifically, we generated sentence-length velocity distribution characteristics of the tongue and jaw and compared measures of central tendency (mean, median), dispersion (range, interquartile range, minimum, maximum), and shape (skewness, kurtosis) between younger and older speakers. Although no specific movement segments were defined in our current approach, the velocity distribution is interpretable due

to the directional nature of the kinematic measure. That is, positive velocity values indicate movements associated with tongue or jaw raising, while negative velocity values indicate movements associated with tongue or jaw lowering in the current study. Thus, potential differences in velocity distribution characteristics between younger and older speakers may provide specific insights in direction-specific performance changes during tongue and jaw lowering and raising.

Due to the paucity of aging studies on tongue and jaw motor performance during running speech production, the nature of this study was exploratory, and no specific hypotheses regarding the articulator or movement direction were tested. Study outcomes, however, were expected to improve insights in articulatory mechanisms that underlie the slowing of speaking rate in older adults.

2. METHODS

2.1. Participants

15 younger speakers (7 females, 8 males, mean age = 23.5, range: 21-27) and 15 older speakers (7 females, 8 males, mean age = 71.5, range: 60-85) participated in this study. Participants were native speakers of American English and reported no history of a neurological condition, or a speech, language, or hearing impairment. All speakers consented to the study prior to data collection (IRB#150655).

2.2. Experimental Procedures

Participants repeated the sentence "*Buy Kaia a kite*" ten times; however, only the first five error-free repetitions were examined for the purpose of this study. The target sentence was selected because it predominantly elicits posterior tongue and jaw movements. Tongue and jaw kinematic data were recorded using a 3D electromagnetic articulograph (AG501, Medizinelektronik Carstens, Germany). Five small sensors were placed on the articulators (jaw, posterior and anterior tongue, lower and upper lip) and three head reference sensors were placed on a pair of goggles, which the participant wore during the data collection. Only the posterior tongue sensor and jaw sensor were analyzed for the purpose of the study. The posterior tongue sensor was placed ~ 4 cm from the tip of the tongue. The jaw sensor was attached at the gumline of the lower central incisors.

All kinematic data were corrected for head movements and then low-pass filtered at 15 Hz using SMASH, a custom-written MATLAB program [11]. Principal component signals were generated for the tongue and jaw trajectories using SMASH. Sentence onset and offset were defined by the positional minimum (max. opening) of the tongue during the word "buy" and positional maximum (max. closing) of the tongue during the word "kite", respectively. The tongue and jaw velocity signals were calculated as the first derivative of the principal component signals (see upper and middle panel of Figure 1).

2.3. Kinematic Measures

Movement duration was determined for each sentence production and was defined as the time from the sentence onset to the sentence offset. Furthermore, descriptive statistics of the sentence-length velocity distributions were generated in SMASH to characterize and compare the articulatory performance of the tongue and jaw between younger and older speakers. Specifically, we compared mean, median, range, interquartile range (IQR), maximum (max), minimum (min), skewness (skew), and kurtosis (kurt) of sentence-length velocity distributions across the two age groups (see lower panel of Figure 1).

As indicated in Figure 1, negative velocity values (in red) represent lowering movements and positive velocity values (in blue) represent opening movements. Hence, a normal distribution with a mean and median near zero velocity suggests that the speaker executes comparable raising and lowering movements. By contrast, a negatively skewed distribution and, therefore, a median value that is greater than the mean value, indicates relatively more time at greater velocities during raising movements than during lowering movements. Finally, a negative kurtosis (i.e., a platykurtic distribution) indicates relatively more time spent at low velocities while a positive kurtosis (i.e., a leptokurtic distribution) indicates relatively little time spent at low velocities.

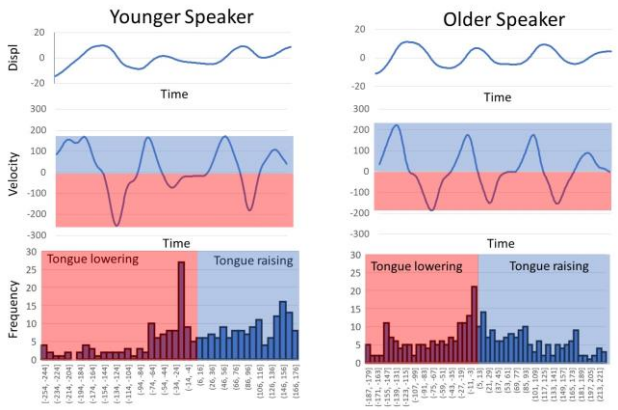


Figure 1: Example of a younger speaker's (left) and older speaker's (right) tongue movement (top), velocity signal (middle), and velocity distribution (bottom).

2.4. Statistical Approach

To determine aging effects on speaking rate, we submitted durations of each sentence of each speaker to a linear mixed model with age as the fixed factor and subject as the random factor. To determine aging effects on articulatory performance, we submitted the distribution characteristics (i.e., mean, median, range, IQR, max, min, skew, kurt) for each sentence production of each speaker to a linear mixed model with age as the fixed factor and subject as the random factor. We adjusted the *p*-value to .006 to account for multiple comparisons. Finally, we explored associations between movement duration and velocity distribution characteristics across speakers.

3. RESULTS

3.1. Movement Durations

Group means (+/- *SE*) of movement durations can be found in Table 1. A significant group effect was found for movement duration [$F(1, 27.61) = 36.8$, $p < .001$]. That is, older adults produced significantly longer durations than younger adults (Mean difference = 0.267 s, *SE* = 0.05).

3.2. Velocity Distribution Characteristics

Group means (+/- *SE*) of the velocity distribution characteristics are also summarized in Table 1. For the tongue, sentence-length velocity distribution characteristics revealed significant group effects for median velocity [$F(1, 27.7) = 11.77$, $p = .002$] and skewness [$F(1, 27.68) = 10.62$, $p = .003$]. Specifically, the median of the sentence-length velocity distribution of the tongue was significantly greater in younger speakers than older speakers (Mean difference = 11.76, *SE* = 3.4). In addition, the sentence-length velocity distribution was significantly more negatively skewed in younger speakers than older speakers (Mean difference = -0.321, *SE* = 0.099).

For the jaw, the velocity distribution characteristics also revealed a significant group effect for median velocity [$F(1, 27.7) = 14.11$, $p < .001$] and a trend for skewness [$F(1, 27.7) = 4.29$, $p = .048$]. The median of the sentence-length velocity distribution was significantly greater in younger speakers than older speakers (Mean difference = 5.77, *SE* = 1.5). Furthermore, sentence-length velocity distribution tended to be more negatively skewed in younger speakers than older speakers (Mean difference = -0.346, *SE* = 0.167).

Variable	Younger Adults	Older Adults
Duration*	**0.959 (0.01)**	**1.225 (0.02)**
Mean (T)	13.5 (0.6)	12.1 (0.3)
Median (T)*	**16.5 (1.4)**	**4.6 (1.0)**
Min (T)	-176.5 (6.5)	-164.1 (5.9)
Max (T)	175.9 (7.4)	194.3 (7.5)
Range (T)	351.3 (13.1)	358.4 (13.0)
IQR (T)	128.3 (12.5)	131.2 (12.5)
Skew (T)*	**-0.3 (0.04)**	**0.1 (0.01)**
Kurt (T)	2.5 (0.1)	2.3 (0.1)
Mean (J)	6.4 (0.4)	4.7 (0.3)
Median (J)*	**9.8 (0.8)**	**4.5 (0.5)**
Min (J)	-65.6 (3.2)	-56.8 (2.8)
Max (J)	57.8 (3.3)	59.1 (3.4)
Range (J)	123.4 (6.0)	115.9 (5.8)
IQR	35.1 (5.0)	34.5 (5.0)
Skew (J)	-0.5 (0.1)	-0.2 (0.1)
Kurt (J)	3.2 (0.2)	3.1 (0.2)

Table 1. Group means (+/- *SE*) of all dependent variables. Significant findings ($p < .001$) are bolded, (T) = tongue, (J) = jaw, duration in seconds.

3.3. Associations between Movement Durations and Velocity Distribution Characteristics across Speakers

Associations between movement durations and velocity distribution characteristics of the tongue and jaw were explored using bivariate correlation analyses. For tongue velocity distribution characteristics, duration was significantly correlated with median [$r(150) = -.38$, $p < .001$], skewness [$r(150) = .38$, $p < .001$], and maximum velocity [$r(150) = .23$, $p = .004$]. However, for jaw velocity distribution characteristics, only the median was significantly correlated with movement duration [$r(150) = .29$, $p < .001$]. Figure 2 provides the scatterplots of the two strongest correlations.

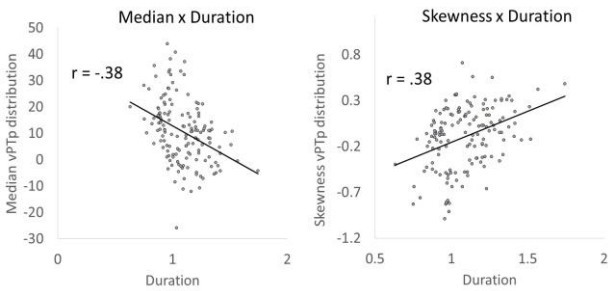


Figure 2: Scatterplot of the median (left) and skewness (right) of the tongue velocity distributions as a function of sentence duration (in seconds) across all repetitions and all speakers (n = 150, 5 repetitions x 30 speakers).

4. DISCUSSION

In the current study we sought to identify articulatory mechanisms that underlie the slowing of speaking rate in older adults. We used a new approach of examining articulatory performance by characterizing sentence-length velocity distributions of the posterior tongue and jaw. Our data confirmed, as previously reported [1-5], significant longer movement durations in older speakers than in younger speakers.

In the current study, we found that peak velocities for raising and lowering movements as well as overall range of velocities, or the interquartile range (IQR) did not significantly differ between the two age groups. This finding differs from that of a previous study showing reduced peak speeds and average speeds in older adults during sentence utterances [9]. However, it should be noted that the previous findings of reduced speed in older speakers was sentence-specific and could not be observed across all speech materials.

The median and skewness of the velocity distribution, particularly those of the tongue, yielded significant group effects and were interpretable with regards to direction-specific performance changes in older adults. For the tongue, findings for median and skewness suggest the proportion of time spent on tongue raising and lowering differed across the two groups. Specifically, younger speakers spent proportionally more time during tongue raising than tongue lowering movements. By contrast, older speakers spent similar amount of time on tongue raising and lowering. The moderate correlations between duration and median velocity as well as between duration and skewness further suggest that with an increase in movement durations, the proportion of time spent on tongue lowering increased as the median velocity became more negative and the skewness of the velocity distribution more positive (Figure 2).

Considering gravitational effects, velocities generated during tongue raising would require more active muscle contraction than velocities generated during tongue lowering. Aging-related changes of the vocal tract shape, such as vocal tract lengthening [12], may also require adjustments to the movement amplitude. Because amplitude is typically scaled with velocity [13,14], older speakers may have to generate greater velocities to accommodate the increased amplitudes. Indeed, maximum velocities for tongue raising tended to be greater in older speakers (Table 1). However, greater maximum velocities were also associated with longer movement durations. This suggest that efforts to scale velocity with amplitude during raising movements were either not sufficient to maintain durations or they were deliberately not scaled to trade speed for articulatory precision. The later possibility is supported by the observations that raising movements typically require more articulatory control than lowering movements [10].

The relative time spent during tongue lowering movements was greater in older adults than younger adults as indicated by less negative skewness. One possible explanation for such changes in articulatory behaviour is that older speakers may attempt to economize articulatory effort by balancing the increased demand on force generation during raising movements with longer periods of less demanding lowering movements. However, future studies are warranted to test this assertion more systematically.

Although group effects observed for the jaw tended to parallel those of the tongue, only median jaw velocity showed a weak, but significant correlation with movement duration. This finding suggests that changes in articulatory tongue performance may contribute disproportionally more to the slowing of speaking rate in older adults than articulatory jaw performance. Although this observation should also be treated with caution due to the small sample size, some support is provided by a previous study that showed an unconstrained ability to increase jaw velocity in older adults [8].

5. CONCLUSIONS

Findings suggest that temporal relations between raising and lowering movements differ between younger and older speakers. Specifically, changes in the velocity distribution characteristics indicate that raising movements are proportionally longer than lowering movements in younger speakers, but they are more similar in duration in older speakers. Although factors that underlie such performance shifts remain elusive, significant correlations between movement duration and velocity distribution characteristics such as median (tongue, jaw) and skewness (tongue) suggest that changes in motor performance contribute at least in part to the slowing of speaking rate in older adults.

6. ACKNOWLEDGEMENTS

This work was funded by the visitor program of the Thiemann Foundation (Tabea Thies) and NIH grants R03DC01575 and R01DC019648 (Antje Mefferd).

5. REFERENCES

[1] Amerman, J. D., & Parnell, M. M. (1992). Speech timing strategies in elderly adults. *Journal of Phonetics, 20*, 65-76.

[2] Duchin, S.W. & Mysak, E.D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders, 20,* 245-257.

[3] Goozée, J. V., Stephenson, D. K., Murdoch, B. E., Darnell, R. E., & LaPointe, L. L. (2005). Lingual kinematic strategies used to increase speech rate: Comparison between younger and older adults. *Clinical Linguistics and Phonetics, 19*(4), 319-334.

[4] Jacewicz, E., Fox, R.A., & O'Neill, C. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change, 21,* 233-256.

[5] Ramig, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders, 16,* 217-226.

[6] Kleinow, J., Smith, A., & Ramig, L. O. (2001). Speech motor stability in IPD: Effects of rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research: 44*(5), 1041–1051.

[7] Kuruvilla-Dugdale, M., Dietrich, M., McKinley, J. D., & Deroche, C. (2020). An exploratory model of speech intelligibility for healthy aging based on phonatory and articulatory measures. *Journal of Communication Disorders, 87,* 105995.

[8] Mefferd, A. S., & Corder, E. E. (2014). Assessing articulatory speed performance as a potential factor of slowed speech in older adults. *Journal of Speech, Language, and Hearing Research: 57*(2), 347–360

[9] Kuruvilla-Dugdale, M., Dietrich, M., McKinley, J. D., & Deroche, C. (2020). An exploratory model of speech intelligibility for healthy aging based on phonatory and articulatory measures. *Journal of Communication Disorders, 87,* 105995.

[10] Gracco, V.L. (1988). Timing factors in the coordination of speech movements. *Journal of Neuroscience, 8*(12). 4626-4639.

[11] Green, J.R., Wang, J., & Wilson, D.L. (2013). SMASH: A Tool for Articulatory Data Processing and Analysis," *Interspeech,* 1331–1335.

[12] Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *Journal of Speech, Language, and Hearing Research: 46*(3), 689–701.

[13] Ostry, D., & Munhall, K. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*(2), 640-648.

[14] Munhall, K.G., Ostry, D.J., Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance 11*:457-474.

Effects of speech modalities in acquiring neural markers of voice recognition: An ERP experiment using voice lineups

Julien Plante-Hébert[1], Victor Boucher[2], Boutheina Jemel[3]

Laboratoire de sciences phonétiques de l'Université de Montréal[1,2], École d'orthophonie et d'audiologie de l'Université de Montréal[3]
julien.plante-hebert@umontreal.ca[1], victor.boucher@umontreal.ca[2], Boutheina.jemel@umontreal.ca[3]

ABSTRACT

Event-related potentials (ERPs) were used to ascertain the influence of contextual information in acquiring neural components of voice recognition. During a training phase, 18 participants had to learn the voices of three speakers selected according to standards of voice lineups. Each voice was trained in a particular modality: audio (A), audiovisual (AV), audiovisual with speaker-listener interaction (AVI). These voices were presented during EEG recordings along with an untrained voice which served as a baseline. Analyses of ERPs revealed that all training conditions led to significant changes on a P2 and a late positive component (LPC). The LPC showed a non-enhancing "face overshadowing effect" in the AV condition which was cancelled in the AVI condition. Combined with previous observations, the results indicate that multisensory information accompanying voice learning does not affect early components of voice *recognition* but does affect a LPC of voice *identification* when multisensory information extends to speech interaction.

Keywords: speaker familiarity, event-related potentials (ERP), voice learning, voice lineups

1. INTRODUCTION

Conflicting reports on the role of facial information in voice recognition have emerged from different sectors of research. Diverging results owe principally to the variability of stimuli that are used and the amount of training given to listeners in experiments of voice familiarization. In terms of the stimuli, investigations bearing on the reliability of ear- and eyewitness testimony refer to protocols such as voice lineups that greatly reduce the variability of face and voice stimuli in identification tasks (e.g., [1-4]). Such protocols are not the general practice in neurophysiological investigations but are nonetheless quite relevant in determining the effects of audio and audiovisual training on neural markers of voice recognition.

For one thing, salient attributes of faces or speech can bias listener's attention on a particular modality and their associative memory of a voice, which can impact neural responses of voice recognition. As for the amount of training required to obtain responses that reflect a recognition of "familiar" voices, it is unclear whether responses to famous voices or voices learned in laboratory settings resemble those evoked by intimately familiar voices [5]. The learning of the latter voices is obviously different not only in terms of the amount of training but also in terms of multimodal experiences that accompany intimately familiar voices and which generally include speech interaction with individuals. Indeed, much of the problem in defining neural markers of voice recognition has to do with the inherent variability of personal experiences with voices.

To address this problem, a previous study compared ERPs on trained and intimately familiar voices that were all quite similar as in standard voice lineups [6]. That study revealed differences in early components P2-N250 and a late positive component (LPC) for trained and intimately familiar voices but where LPC stood out as a marker of voice recognition relating to the *identity of a speaker.*

The present study builds upon these results and aims to determine the effects of modalities of voice learning on ERPs of voice recognition where voice variability is again minimized in accordance to standard voice lineups. To investigate modality effects, three learning or "training" conditions are used involving different modalities of stimuli presentation. These include an audio modality (A) where listeners are trained on heard voices; an audiovisual modality (AV) in which voices are trained by viewing and hearing speakers; and an audiovisual modality with interaction (AVI) where voices are trained by hearing, seeing, and repeating what speakers say.

2. METHODOLOGY

2.1 General design

The design included a voice-training phase followed by an experimental phase involving EEG recordings. In the training phase, participants learned to identify, by way of symbols on a keypad, three target voices, V1, V2, V3, which were presented in a particular modality (A, AV or AVI). Thus, there were three trained voices (TVs), individually learned in three

different modality conditions. The modalities in which each voice was learned varied across participants as well as their order of first appearance. Once the voices were acquired, participants proceeded to the experimental phase and EEG was recorded during audio presentations of the TVs and of one untrained voice (UV), V4, which served as a baseline in the analysis of ERPs.

2.2 Participants

These were 18 native speakers of Quebec French (9 females) aged between 21 and 30 years (mean=25, s.d.=3). All were dominant right handers [13], had normal hearing as established by an audiometric screening test, and normal memory spans (WMS-III).

2.3 Speech and voice stimuli

The training and experimental stimuli were drawn from a list 260 two-syllable common nouns. Audio and audiovisual recordings were made from the productions of these words by four native speakers of Quebec French that were selected in accordance with standards of voice lineups. Specifically, the four male speakers had no discernible regional accent, no idiosyncratic articulations, and they had similar speaker fundamental frequencies (SF0s) to within 1 semitone (mean: 120.83 Hz).

2.3.1 Experimental stimuli

These stimuli were the audio recordings of the TVs and the UV. Audio presentations for the four voices comprised 60 different words each. Recordings of these produced words were made in a sound-attenuating booth using a Shure (X2u) sound card, at 44.1 kHz and 16-bit, a Lavalier (Audio-Technica, AT831b) and software (Golwave 6.31). To obtain productions of stimuli with similar prosody, a rhythm guide was used, and amplitudes were normalized to /a/ sounds.

2.3.2 Training stimuli

These were audio and audio-video recordings of the same 20 words produced by three of the speakers described above (V1, V2 and V3). The recordings in mp4 format used a webcam (Web HD Pro, Logitech), a 64-bit NVIDIA graphic card (Quadro K5200) and were edited via DaVinci 14 (Blackmagic). In the videos, the head and face of the speakers were similarly positioned on the screen and displayed against a neutral background. The facial features and the clothing of the speakers were also similar. The 20 words serving to train voices V1, V2, V3 were randomly assigned to the three training conditions across participants. Finally, an important aspect of the stimuli is that, during the recording of produced words in the AV condition, the speakers looked down at a mark on the table in front of them whereas, for the AVI condition, speakers' eye gaze was directed at the camera during the production of words before returning to a downward gaze. This was meant to capture the gaze effects that occur during speaker-listener interaction, and such effects prevail even if the speaker is not physically present [7]. Eye gaze served to cue participants' repetition of heard words in the AVI condition (see "Procedure").

2.4 Procedure

2.4.1 Training phase

The training involved an introductory presentation of the stimuli followed by repeated training-test cycles, similar to [8]. Participants sat in front of a laptop and attended to presented audio and audiovisual recordings using insert earphones (EARtone 3A).
In the introductory presentation, 10 identical speech contexts representing voices V1, V2, V3 were played back in the same order, for a total of 30 trials. The three voices were, however, presented in different training conditions (A, AV, AVI) that were randomly assigned and counterbalanced across participants such that all voices and conditions were played back an equal number of times in a training block. The participants were instructed to focus on speaker's characteristics so as to remember their voice. Only for the AVI condition, they were also told to look at a speaker in the eyes and repeat the heard word aloud.
In the training-test cycles, the same 10 speech stimuli used in the introduction were presented in random order. Each cycle consisted of a training block of 30 trials followed by a test block of 30 trials presented using E-prime 3. During the training blocks, participants had to learn to associate a voice with a symbol on the keypad and this symbol was also displayed on slides that accompanied the presented recordings. In the following test blocks, only audio versions of the training stimuli were presented and participants identified the voices using the keypad. The training blocks were repeated until participants could correctly identify speakers on at least 23 of the 30 trials of the test blocks in three consecutive cycles.

2.4.2 Experimental phase

EEG recordings were performed at this phase. Participants sat at approximately 130 cm from a computer monitor displaying a fixation cross and wore ear inserts as in the training phase. The same symbol-coded keypad was also used to record voice-identification responses. The experimental stimuli

consisting of audio recordings only were played back in two continuous blocks separated with a pause. Stimuli were presented using MatLab/Psychtoolbox software routines. The participants were required to rapidly identify, after each heard word, whether the voice was V1, V2, V3, or "other" on the keypad.

2.5 EEG recording and analyses

EEG signals were recorded in two continuous blocks for each participant using the international 10–20 system with ASA-lab EEG/ERP 64-channels amplifier (ANT neuro) with an online average reference and a 1 kHz sampling rate. Eye movements were recorded using four electrodes placed above and below the dominant eye and at the outer canthus of each eye. Electrode AFz was used as ground. Offline, the recordings were band-pass filtered (0.1-30 Hz) and blinks were removed using ASA software (ANT neuro). All other artefacts in the EEG exceeding a standard deviation of 20 µV within a sliding window of 200 ms were automatically removed with EEprobe GUI (version 1.2.0.2, ANT Software). All subsequent analyses including ERP averaging across individual trials and participants as well as statistical analyses were performed using Fieldtrip [9]. EEG recordings were then averaged across blocks according to the training conditions. Only trials associated with correct responses were included in the ERP averages. The average time window of separate epochs was set between 200 ms before and 1000 ms after each stimulus file onset. The 200 ms pre-stimulus interval was used for baseline correction.

Before analyzing the behavioral data, trials with response times (RT) exceeding 2 standard deviations from the average value were excluded (2.08%).

A global fields power analysis presented in Figure 1 (left) served to circumscribe time windows of peak brain activity to peaks representing the P2 component (139 to 239 ms post onset) and the LPC (550 to 900 ms post onset) [10]. For each of the components, smaller short-time-windows were used so as to perform t-test statistical comparisons using a Monte-Carlo method which compares ERP responses on each training condition to ERP responses for the UV baseline condition. The method involves non-parametric cluster-based random permutation tests as implemented in Fieldtrip [11, 12]. Clusters were considered significant at $p < 0.05$. Topographies in Figure 1 (right) therefore reflect the *t*-values obtained in the analyses rather than more common ERP magnitudes and polarities.

3. RESULTS

3.1 Behavioral results

A repeated-measures ANOVA revealed a main effect of training condition on the RTs [$F(3, 45)=11.98$, $MSE=47678.6$, $p<0.000$, $\eta^2=0.444$]. Post hoc tests using Bonferroni correction for multiple comparisons indicated that the strong difference owed to the contrast between TVs in the three training conditions and UV. RTs for all three TV conditions were significantly shorter than those for UV (at $p<0.001$). However, the same post hoc tests did not reveal significant differences amongst the three TV conditions on either correct responses or RTs.

3.2 ERP results

The analyses of the differential responses to TVs were carried within four 25 ms windows surrounding the P2 peak. These yielded significant clusters for all three training conditions when compared individually to UV between 164 and 189 ms. Some significant clusters also appeared for the AV condition between 139 and 164 ms post stimuli onset. All significant clusters within these time windows were located on central and parieto-central sites along the middle line. However, such uniformity of P2 across conditions was not present for the LPC.

The LPC was analyzed using 50 ms time windows for these protracted responses. As illustrated in Figure 1, cluster analyses of the LPC responses to TVs in the

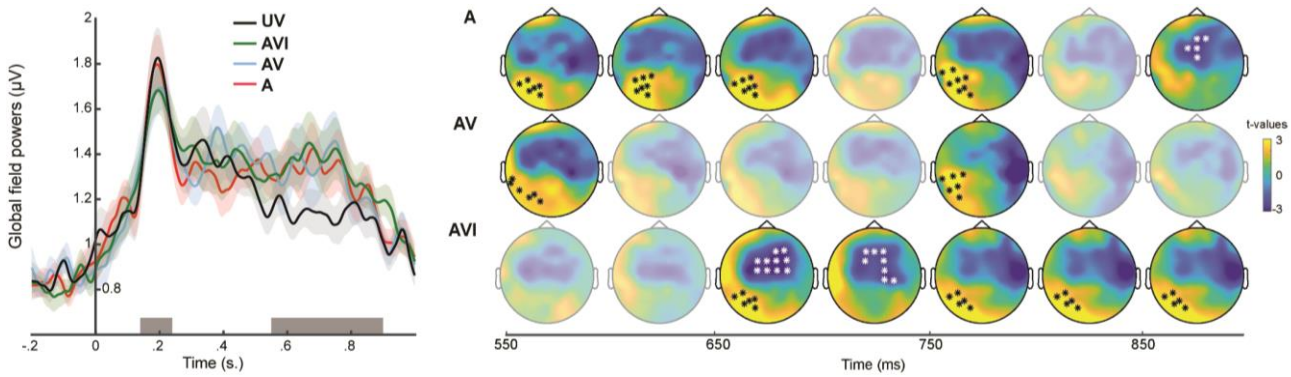


Figure 1: Left: Global fields power analysis averaged from all sites. Right: topographic representations of t-values obtained from the cluster analyses of differential response to TVs and UV across the learning conditions for time windows around the LPC peak. Highlighted electrodes are statistically significant ($p < 0.05$).

three training conditions were significantly different from responses to UV across a wide time interval, and sites of differential activity varied across conditions. Specifically, the cluster analyses showed that the training condition A associated with significant clusters in five of the seven time-windows starting from 550 ms and extending to 900 ms. Significant clusters where mostly located on left-parietal, centro-parietal, and parieto-occipital sites. The AV condition also yielded significant clusters at these sites but only between 550 and 600 ms and 750 ms 800 ms. In sum, for these LPCs, presentations of similar faces in learning voices in the AV training condition *did not enhance* differential ERP responses to TVs and UV in comparison to presentations of voices alone. However, a different pattern of activity arose in the AVI training condition. In this case, significant clusters reflecting differential centro-parietal responses appeared later (starting at 650 ms post-stimuli onset) and shifted to middle and right fronto-central sites before returning to left parietal sites (at 750 ms post onset). This suggests that active speech participation while learning voices of visually presented speakers has a sustained long-lasting effect on voice processing in contrast to passive conditions A and AV.

4. DISCUSSION

The above experiment used narrowly controlled voice and face stimuli to investigate effects of multimodal information on voice identification via three training modalities. After a training phase, participants were asked to identify target speakers on multiple trials. The behavioral results showed that TVs were successfully identified but there were no significant differences between training conditions.

However, in the analyses of the ERPs, responses to TVs were differentiated from those to UV and the Monte-Carlo analyses of these differential responses bore out a particular effect of training conditions A and AV. It will be recalled that A and AV only differed in terms of the presence of visual facial information at training. Both conditions resulted in modulations of P2 and the LPC. This suggests that adding visual face information did not facilitate voice identity encoding and, in fact, *may have impeded such encoding.* Such an effect concurs with research attesting to a "face overshawdowing effect" (FOE), although the effect in the present case may owe to the highly similar facial information of the speakers that was provided and which did not assist in identity discrimination [13, 14]. It should also be noted that the preceding FOE occurred with dynamic facial stimuli (cf. [15]).

As for the effects of AV and AVI, it will be recalled that these two training conditions differed in that, for AV, the listeners did not see the speaker's gaze whereas, in the AVI condition, the listeners were asked to repeat the words produced by the speaker while gazing at the speaker on a monitor. This mode of presentation was designed to partly simulate the effects of speech interaction. As in other training conditions, the AVI condition elicited a P2 and an LPC. Compared to the AV training, however, the AVI training resulted in a wider-ranging response in the LPC window, and also in a greater number of significant clusters that shifted momentarily toward mostly middle and right centro-frontal sites. These results therefore provide evidence that speech interaction in the AVI condition, involving effects of gaze and simulated speech interaction, enhances ERP components at voice recall. Of course a laboratory simulation has its limits and, if anything, might underestimate the effects of actual person-to-person contact on voice identification as reported by [16].

More importantly, a central finding of the present study is that the observed ERPs to trained voices narrowly conform to ERPs found in a previous study involving intimately familiar and trained-to-familiar voices [6]. Both investigations involving similar stimuli have revealed responses of voice recognition and identification in the same time ranges as the P2 and LPC. It is interesting to note that in the above study, all three successfully trained voices, as validated by behavioral results, elicited a P2 that differed significantly from responses to UV. This brings new evidence confirming that the P2 is a valid marker of voice recognition. But the present results also indicate that speaker identification, or access to available semantic information on a speakers' identity when hearing a voice, as measured by the LPC, was not enhanced by the additional presentation of visual stimuli during training. This entails that adding dynamic visual information of speakers' faces may not necessarily enhance voice encoding and recall. In fact, such contextual information can, in some cases, divide the learner's attention as suggested in reports of a FOE in voice recognition tasks. On the other hand, the greater LPC response of AVI as compared to AV shows that the FOE may vary or diminish with the addition of other contextual information such as speech interaction and gaze toward a speaker. Even if some participants reported being distracted by the repetition of verbal forms, a differential LPC in response to TVs was observed. In other words, in the same way [15] showed that the FOE decreased after a certain exposure threshold or amount of training, the present results suggest that FOE could also be reduced by the type and amount of contextual information that is provided.

6. REFERENCES

[1] Hollien, H., Huntley, R., Kunzel, H., and Hollien, P.A. 2013. Criteria for earwitness lineups. *International Journal of Speech Language and the Law*. 2. 143-153.

[2] Nolan, F. and Grabe, E. 1996. Preparing a voice lineup. *International Journal of Speech, Language and the Law*. 3. 74-94.

[3] Wells, G.L., Kovera, M.B., Douglass, A.B., Brewer, N., Meissner, C.A., and Wixted, J.T. 2020. *Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence*. Law and Human Behavior. 44. 3.

[4] Yarmey, D.A. 2014. The psychology of speaker identification and earwitness memory, in: Lindsay, R.C.L., Ross, D.F., Read, J.D., and Toglia, M.P., (eds), *Handbook Of Eyewitness Psychology*. Routledge, 115-150.

[5] Kreiman, J. and Sidtis, D. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

[6] Plante-Hébert, J., Boucher, V.J., and Jemel, B. 2021. The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *Plos one*. 16. e0250214.

[7] Conty, L., Russo, M., Loehr, V., Hugueville, L., Barbu, S., Huguet, P., Tijus, C., and George, N. 2010. The mere perception of eye contact increases arousal during a word-spelling task. *Social Neuroscience*. 5. 171-186.

[8] Zäske, R., Volberg, G., Kovács, G., and Schweinberger, S.R. 2014. Electrophysiological correlates of voice learning and recognition. Journal of *Neuroscience*. 34. 10821-10831.

[9] Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*. 2011. 1.

[10] Lehmann, D. and Skrandies, W. 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and clinical neurophysiology*. 48. 609-621.

[11] Kroese, D.P., Taimre, T., and Botev, Z.I. 2013. *Handbook of monte carlo methods*. John Wiley & Sons.

[12] Maris, E. and Oostenveld, R. 2007. Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*. 164. 177-190.

[13] Stevenage, S.V., Howland, A., and Tippelt, A. 2011. Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*. 25. 112-118.

[14] Tomlin, R.J., Stevenage, S.V., and Hammond, S. 2016. Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition*. 25. 629-643.

[15] Zäske, R., Mühl, C., and Schweinberger, S.R. 2015. Benefits for voice learning caused by concurrent faces develop over time. *PloS one*. 10. e0143151.

[16] Hammersley, R. and Read, J.D. 1985. The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior*. 9. 71.

Interactions of speech and speaker processing in early evoked components

Paula Rinke[1,3], Nadine Lavan[2] and Mathias Scharinger[1,3]

[1]Research Group »Phonetics«, Institute of German Linguistics, Philipps-Universität Marburg, Germany
[2]School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom
[3]Center for Mind, Brain & Behaviour »CMBB«, Universities of Gießen & Marburg, Germany

paula.rinke@uni-marburg.de, n.lavan@qmul.ac.uk, mathias.scharinger@staff.uni-marburg.de

ABSTRACT

The voice of a speaker and the linguistic content of a message are linked to one another during speech production and perception, allowing to dissociate between what is being said by whom. The time course of the interaction between speech and speaker information is, however, less well investigated. Using time-sensitive neuro-physiological measurements (EEG) with Event-Related Potentials (ERP), the present study aims to address this question.

16 participants attended a 1-back-task while being presented with recordings of the German vowels [a], [i], [u] from 32 German-speaking men and women, falling into a younger or older age group, (96 stimuli in total), while brain activity was recorded from 32 active electrodes.

ERP-analyses showed that vowel category elicited strong effects in early evoked components, especially in the N1 time window (100ms after stimulus onset), with speaker age and gender exerting significant modulations. This suggests an early interaction of speech and speaker processing.

Keywords: speaker processing, voice processing, EEG, ERP

1. INTRODUCTION AND BACKGROUND

Speech and speaker information interact with one another. Even though the voice is often simply considered as the “carrier of speech” [1] its influence on speech processing is inevitable. The voice carries much social information (so-called “indexical information”) about the speaker, e.g., gender, age, emotional state and social or regional background. However, the perception of speaker-specific features is often considered as higher-order processing, that is separate from low-level speech processing.

Previous studies, however, show that speaker-specific cues can affect low-level speech processing such as speech sound categorization [2]. For example, Strand [3] provides evidence on how the gender attributed to a speaker influences the categorization of English [s] and [ʃ]. Moreover, Drager [4] shows that the perception of New Zealand English vowels is influenced by higher-order knowledge such as the perceived age

These results provide evidence that the perceived social characteristics of a speaker affect speech processing. Thus, speech and speaker perception appear as interdependent rather than two completely separated processes.

Aside from these behavioral data, evidence on the neural basis of this interaction is comparably small. Speech information, such as vowel category, are known to be retrieved during very early processing stages. As previous neurophysiological studies (electroencephalography (EEG) and Event-Related Potentials (ERP) as well as magnetoencephalography (MEG)) have shown, spectral information inherent in different vowel categories (e.g., formant frequencies) affect early brain responses such as the N1 [5, 6], an auditory evoked component with a negative peak between 80 and 150 ms after stimulus onset [5]. In particular, the first formant (resonance frequency of the vocal tract), F1, was found to modulate amplitude and latency of the N1 [7].

While the time course of speech processing is well investigated, research on the time course of social characteristics of the voice is relatively rare. The few existing EEG studies on voice processing found two neural components reflecting the processing of speaker-specific features: the Mismatch Negativity (MMN) and the P3a [8-10]. The MMN is an early and automatic brain response with a negative deflection that is sensitive to change detection, whereas the P3a is a later component with a positive peak approximately 300 ms after stimulus onset and is considered to indicate involuntary attention switches caused by a novel stimulus [11].

In the current study, we brought together the investigation of speech and speaker perception using time-sensitive EEG and ERP measurements. By manipulating speech-related information (vowel categories) and speaker information (age and gender), we attempted to characterize the time course of

speech and speaker-related information. Based on the findings of previous ERP studies, we would expect vowel processing to be reflected in early brain components, such as the N1, with speaker's age and gender possibly modulating the response pattern.

2. METHOD

2.1 Stimuli

96 recordings of the German vowels [a], [i], and [u] from 32 speakers (16 male; 16 female) were selected from the "Saarbrücker Stimmdatenbank" (http://www.stimmdatenbank.coli.uni-saarland.de/). Each group was equally divided into two age groups representing either younger (20-34 years) or older adults (55-74 years; details are listed in Tab. 1). Consequently, each category (e.g., male-old or female-young) consisted of eight different speakers with each one producing the three vowels (in total 24 sounds per vowel category).

Category	Mean	Range	SD
Female young	24.13	20-34	4.86
Male young	24.25	21-28	4.35
Female old	62.63	55-74	5.66
Male old	63.00	56-72	4.80

Table 1: Overview of different age groups in stimulus material

The German vowels [a], [i], and [u] are distinct in their tongue position, openness and backness with each vowel falling into easily separable locations in the German vowel space. They thus show distinctive spectral properties such as systematic differences in vowel formants, especially in F1 and F2.

Sounds were edited in Praat [12] and adjusted to an intensity level of 75 dB. Furthermore, all sounds were trimmed to a duration of 400ms in Audacity.

2.2 Experimental Design

Participants were presented with all 96 vowel stimuli, which were repeated 40 times across six blocks. During stimulus presentation, participants were asked to complete a 1-back-task for speaker identity. To test for continuous attention, we included vigilance trials for which additional vowel stimuli were selected (24 altogether). Vigilance stimuli were repeated 16 times and evenly distributed over the six blocks. In total, participants were presented with 4224 vowel stimuli.

Participants were placed in a shielded room approximately 1 m in front of a screen on which the instructions and the task were presented. The 1-back-task was completed via press of keyboard keys to indicate whether the presented and the preceding stimuli were the same (same vowel and speaker). Sounds were delivered via loudspeakers placed on the right and left side of the screen. To prevent extensive eye movement, participants were asked to focus on a white fixation cross on the screen during stimulus presentation and try to blink only right after responding to the task.

The experiment lasted approximately 70 minutes.

2.3 Participants

16 participants took part in the study (8 males, 8 females; mean age 26.13 years, range 19–36; standard deviation [SD] 5.12 years). All participants were native German speakers and had no reported history of hearing or neurological disabilities. Written informed consent was collected from all participants in order to participate in the study. The study was approved by the local ethics committee.

2.4 EEG recording

Continuous EEG was obtained from 32 active electrodes placed on a standardized 10–20 system cap. The reference electrode was placed on the nose while the ground electrode was placed on the forehead between Fp1 and Fp2 positions. Impedance was kept below 10 kΩ while EEG signals were sampled at 1000 Hz and online pre-filtered between 0.016 and 250 Hz using the Brain Vision Recorder software (Brain Products, Gilching, Germany), running on a Windows 10 personal computer.

2.5 EEG preprocessing

Continuous EEG was preprocessed using fieldtrip [13]. First, erroneous EEG channels (maximally 2 per participant) were identified and interpolated. Next, eye movements and blinks were identified through independent component analysis (ICA). Again, maximally 2 eye components per participant were identified and removed. Subsequently, the continuous EEG was split in epochs aligned to vowel onsets, starting 100 ms before and ending 700 ms after vowel onset. Epochs were bandpass-filtered (between 0.01 and 30 Hz) and the mean voltage of the 100-ms baseline was subtracted. Epochs were then averaged across repetitions within each participant and thereby corresponded to ERPs in a 3 (vowels [a], [i], [u]) x 2 (speaker age: old, young) x 2 (speaker sex: male, female) setup.

2.6 Statistical analyses

Visual inspection of the preprocessed data revealed two substantial time windows between 90 and 150 ms and 350 and 500 ms that correspond to the time windows of the N1 and the N4/P3. The statistical analysis was carried out for these two time windows and was based on the mean values of each item (i.e., vowel per speaker and age group). Vigilance trials were excluded; thus, the analysis was done for test items only. Mean amplitude in the time windows of interest and the latency of the highest N1 peak were calculated.

Linear mixed models (LMMs) were based on the Fz electrode and calculated for N1 amplitude and latency and for P3/N4 amplitude with vowel category, speaker's gender and age as fixed effects and subject as random intercept. Data were analyzed in Jamovi [14].

2.7 Results

2.7.1 Amplitude

The LMM on N1 amplitude (Tab. 2) showed a main effect for gender (p<.001) and vowel category (p<.001), but not for age (p>0.2). N1 amplitudes were more negative for male than for female voices, and more negative for [i] than for the other two vowels. The interaction of gender and vowel category (p=.018) was driven by larger between-vowel differences for female than for male voices.

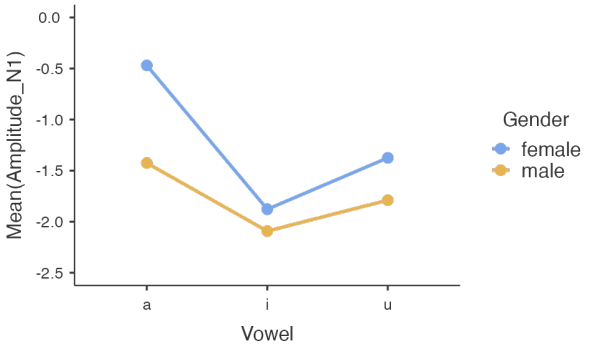


Figure 1: Amplitude differences of the N1 depending on vowel category and speaker gender.

The interaction of gender and age (p=.014) showed that the gender effect was more pronounced for the young compared to the old speakers. No other interactions were significant (all p-values>0.2).

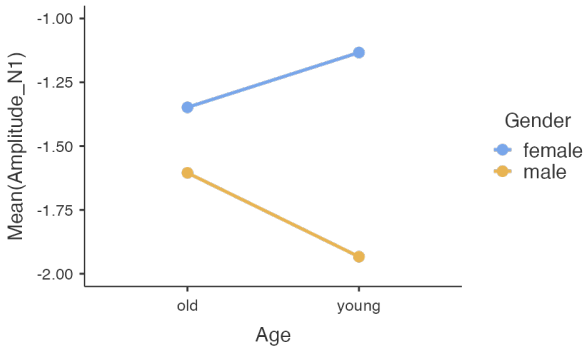


Figure 2: Amplitude differences of the N1 depending on speaker age and speaker gender.

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Gender	23.804	1	110	<.001
Age	0.275	1	110	0.601
Vowel	31.170	2	110	<.001
Gender ✻ Age	6.302	1	110	0.014
Gender ✻ Vowel	4.191	2	110	0.018
Age ✻ Vowel	1.351	2	110	0.263
Gender ✻ Age ✻ Vowel	0.672	2	110	0.513

Note. Satterthwaite method for degrees of freedom

Table 2: Linear Mixed Model for N1 amplitude.

The P3/N4 amplitudes (Tab. 3) were similar to the patterns observed patterns for the N1: main effects were found for vowel category (p=.017) and gender (p=.007), but not for age (p>0.2). The effect of gender depended on vowel category (p=.036). No other interactions were significant (all p-values>0.2).

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Gender	7.5270	1	110	0.007
Age	0.7397	1	110	0.392
Vowel	4.2424	2	110	0.017
Gender ✻ Age	0.0180	1	110	0.894
Gender ✻ Vowel	3.4366	2	110	0.036
Age ✻ Vowel	1.0074	2	110	0.369
Gender ✻ Age ✻ Vowel	1.4478	2	110	0.240

Note. Satterthwaite method for degrees of freedom

Table 3: Linear Mixed Model for N4 amplitude.

2.7.2 Latency

For N1 latency (Tab. 4), there were significant main effects of gender (p=.021) and vowel category (p<.001), but no age effect (p>0.2). Notably, the vowel [a] elicited earlier N1 responses than [i] or [u]. Vowels pronounced by female speakers also elicited earlier N1 responses than vowels pronounced by male speakers. No interactions were significant (all p-values>0.1).

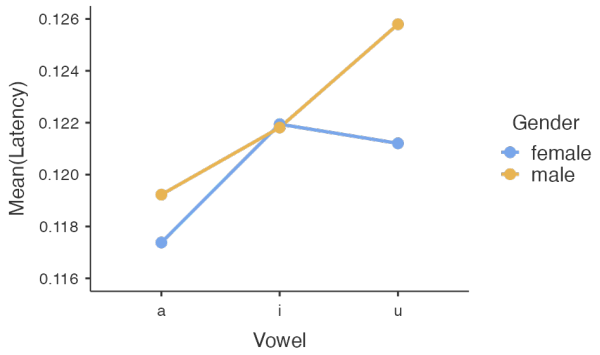


Figure 3: Latency differences of the N1 by vowel category, plotted separately for female and male voices.

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Gender	5.485	1	110	0.021
Age	0.111	1	110	0.740
Vowel	11.699	2	110	<.001
Gender ✻ Age	2.084	1	110	0.152
Gender ✻ Vowel	2.327	2	110	0.102
Age ✻ Vowel	0.575	2	110	0.564
Gender ✻ Age ✻ Vowel	0.122	2	110	0.885

Note. Satterthwaite method for degrees of freedom

Table 4: Linear Mixed Model for N1 latency.

3. DISCUSSION

As predicted, the results show a significant influence of vowel category and gender on the amplitude and latency of the early N1 brain potential.

Despite large acoustic inter-speaker variability, the vowel category appears to have a strong influence on the shape and latency of the N1. As the results show, even when articulated by different speakers the different vowels [a], [i] and [u] modulate the N1. Differences between the vowel categories are most prominent in N1 latency and amplitude: [a] shows the smallest N1 latency, i.e., elicits the N1 significantly faster than the other vowels. When looking at later brain responses, the vowel category shows an influence on N4 amplitudes, albeit a weaker one in comparison to the N1 amplitude.

Regarding N1 properties, these results are in line with the findings of previous studies on German vowels affecting early brain potentials [7]. They provide data on how the N1 is sensitive to discriminative spectral features by demonstrating that N1 properties are affected by different formant frequencies (i.e., F1 and F2).

As the N1 analysis reveals, the different vowel categories show different effects on N1 properties. While [i] elicited the largest amplitude, [a] elicited the shortest latencies. It could be speculated that different F1 properties have the strongest effect on the N1. However, note that German /a/ is not realized as a front vowel but rather retracted as central vowel [a̠]. As a consequence, [a] and [i] also show differences in F1 and F2 properties. Thus, the current results cannot be exclusively attributed to one particular formant frequency.

The speaker's gender was found to be relevant for both observed time windows and in particular for the amplitude and latency of the N1. Vowels produced by a male voice elicited a stronger but later N1 than vowels produced by female voices.

Since a person's gender is often considered as "most primary judgments that human perceivers make of each other" [3] it is not surprising that its influence stretches across multiple time windows.

From an acoustic point of view, gender is strongly (but not solely) associated with the fundamental frequency (f0). F0 plays an important role in vowel perception besides fundamental frequencies. This is also displayed in the present study by the strong interaction of gender and vowel category for N1 amplitude. Thus, it is plausible that the speaker's gender is retrieved during early speech processing reflected by the N1 as well as during speaker processing displayed by the N4.

When focusing on the perceived age of the speaker, no effects on the amplitude and latency of the early brain response N1 as well as the later N4 can be observed. Thus, as expected from previous voice studies, the speaker's age does not show any influence on early brain components which are mainly modulated by vowel categories and gender. However, significant interactions with the speaker's gender were found for N1 amplitude, with a stronger effect of gender for younger than older speakers, which again highlights the importance of gender for early processing stages.

Regarding the interpretation of the gender effect on N1 properties, it is important to point out a possible limitation of these analyses. The N1 is an early component that is sensitive to acoustic changes. Since male and female voices differ substantially in F0 and other acoustic characteristics [15], we note that the differences found in amplitude and latency could at least partially reflect these low-level acoustic differences.

However, there were additional effects of gender in the later N4 time window. This time window is associated with higher-level processing (as opposed to early acoustic processing), which we cautiously interpret as a result of the processing of the speaker characteristic "gender".

Overall, the current data confirm that speaker characteristics can in some cases influence early and later brain responses to speech-related information. The data thus show that speaker and speech-related information in fact routinely interact and inform each other during the perception of human vocal stimuli.

4. REFERENCES

[1] P. Belin, S. Fecteau, and C. Bédard, 2004. Thinking the voice: neural correlates of voice perception, *Trends Cogn. Sci.,* 8, 129-135

[2] K. Johnson, E. A. Strand, and M. D'Imperio, 1999. Auditory–visual integration of talker gender in vowel perception, *J. Phonetics,* 27, 359-384

[3] E. A. Strand, 1999. Uncovering the role of gender stereotypes in speech perception, *J. Lang. Soc. Psychol.,* 18, 86-100

[4] K. Drager, 2011. Speaker age and vowel perception, *Lang. Speech,* 54, 99-121

[5] R. Näätänen and T. Picton, 1987. The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure, *Psychophysiology,* 24, 375-425

[6] A. Shestakova, E. Brattico, A. Soloviev, V. Klucharev, and M. Huotilainen, 2004. Orderly cortical representation of vowel categories presented by multiple exemplars, *Cognitive Brain Research,* 21, 342-350

[7] M. Frank, B. Muhlack, F. Zebe, and M. Scharinger, 2020. Contributions of pitch and spectral information to cortical vowel categorization, *J. Phonetics,* 79, 1-13

[8] M. Beauchemin *et al.*, 2006. Electrophysiological markers of voice familiarity, *Eur. J. Neurosci.,* 23, 3081-3086

[9] J. Graux, M. Gomot, S. Roux, F. Bonnet-Brilhault, and N. Bruneau, 2015. Is my voice just a familiar voice? An electrophysiological study, *Soc. Cogn. Affect. Neurosci.,* 10, 101-105

[10] P. Rinke, T. Schmidt, K. Beier, R. Kaul, and M. Scharinger, 2022. Rapid pre-attentive processing of a famous speaker: Electrophysiological effects of Angela Merkel's voice, *Neuropsychologia,* 173, 108312

[11] G. B. Remijn, E. Hasuo, H. Fujihira, and S. Morimoto, 2014. An introduction to the measurement of auditory event-related potentials (ERPs), *Acoust. Sci. Technol.,* 35, 229-242

[12] *PRAAT: Doing phonetics by computer (version 6.1.24)*. (2020). Institut for Phonetic Sciences, Amsterdam.

[13] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data, *Computational Intelligence and Neuroscience,* doi:10.1155/2011/156869, 1-9

[14] *jamovi (Version 2.3)*. (2022).

[15] M. Latinus and M. J. Taylor, 2012. Discriminating Male and Female Voices: Differentiating Pitch and Gender, *Brain Topography*, 25, 194–204

ACKNOWLEDGMENTS

The research for this paper was funded in part by the German Research Foundation (Research Training Group 2700: »Dynamics and stability of linguistic representations«) and a Sir Henry Wellcome Fellowship (220448/Z/20/Z) awarded to Nadine Lavan.

Impaired vowel perception in congenital amusia: Evidence from event-related potentials

Jasmin Pfeifer[1] & Silke Hamann[2]

[1]Institute for English Language and Linguistics, Heinrich-Heine-University, Germany
[2]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands
[1]pfeifer@phil.hhu.de; [2]silke.hamann@uva.nl

ABSTRACT

Congenital amusia is an auditory perception disorder present from birth. It has been presumed to affect only the musical domain. However, several studies now show that amusics also have impaired perception of intonation and linguistic tones.

We show that amusia also influences linguistically relevant cues other than pitch by investigating the perception of German front high vowels, differentiated by duration and formant frequency. We assessed amusics' electrophysiological responses, more specifically the MMN with a multi-deviant oddball paradigm.

We used an lmer model for the MMN data, finding significant main effects for group with amusics having a smaller MMN than controls.

Our study shows that amusia negatively affects vowel perception and has therefore more far-reaching consequences for speech perception than previously assumed. We found differences in the MMN, reflecting differences in early auditory change detection.

Keywords: Congenital amusia; MMN; vowel perception; duration; formant frequency

1. INTRODUCTION

The condition known as congenital amusia causes lifelong deficits in the ability to perceive pitch and in some cases also rhythm. Insufficient music exposure, hearing loss, or brain damage do not cause the disorder [1]. The most apparent symptoms to the affected individuals are various inabilities in the musical domain such as: Recognition of familiar melodies, detection of out-of-tune notes or singing, or clapping and singing along. Hence, congenital amusia has long been characterized as a music-specific pitch perception disorder [1]. Musical impairments were the focus of initial research on amusia. More recent work, investigated the impact of amusia on language perception, as pitch is also an important cue in language, e.g., to disambiguate questions from statements or to mark focus. Intonation perception was the first area of speech perception shown to be affected by amusia [2, 3]. Pitch is also used in other linguistic areas, which have been shown to be affected as well, such as tone language perception [3]. Due to these findings, congenital amusia is now seen as a domain-general disorder that negatively affects pitch processing in general [4, 5].

So far, the study of speech perception impairments caused by amusia has almost exclusively focused on areas involving pitch as a perceptual correlate of speech sounds, which is probably due to the fact that most hypotheses on the underlying deficit of amusia are based on some form of pitch perception deficit. Speech, however, also makes extensive use of other information in the speech signal such as spectral frequencies. The latter are especially relevant in the perception of vowels. The quality of a vowel is usually determined by its lowest two formants, F1 and F2. According to the tongue position, vowels are characterized as high vs. low (F1 values), and front vs. back (F2 values). Other aspects, such as nasality, rounding of the lips, and steady state vs. inherent movement also influence formant values, but are neglected in the present study.

First reports on vowel perception by congenital amusics appeared recently, all testing native speakers of tone languages. Huang et al. [6] assessed the discrimination and identification of two high vowels in Mandarin and found no difference between amusics and controls in the identification task, but the overall discrimination rate of amusics was lower. Zhang et al. [5] had similar findings for the perception of two back vowels in Cantonese, and concluded that the deficit in amusia is not specific to pitch processing but rather concerns the processing of spectral frequencies in general.

These studies considered the differences in vowel quality only and neglected durational differences. Furthermore, all studies involved tone languages, while the perception of vowels by amusic native speakers of non-tone languages has not been tested yet. And lastly, all studies employed only behavioral tasks to test vowel perception, while research on speech perception of non-amusics and on pitch perception by amusics showed interesting results on the electrophysiology underlying such behavior.

Different event-related potential (ERP) components, measuring the brain's electro-physiological response to a stimulus, can be used in

speech perception research. One such component is the so-called mismatch negativity (MMN), an early component that reflects an automatic, unconscious detection of change in a series of stimuli, which can be recorded without requiring attentive action from the participant (for reviews, see [7, 8]). The MMN can be elicited by a change of frequency, duration, intensity, location or pattern in a stream of speech or non-speech sounds [9, 10]. Generally, so-called oddball paradigms are used: A repetitive standard stimulus is presented many times in a row, occasionally interrupted by a deviant differing in an acoustic feature [7, 8]. In so-called multi-deviant oddball paradigms multiple, different deviants within a string of one standard are used. The MMN typically peaks at around 100 to 250 ms after a deviant is presented, if the participants' auditory system has formed a representation of the repetitive aspect of the standard stimulus [7].

The MMN is ideally suited as a starting point in investigating the electrophysiology of vowel discrimination in amusia, as MMN paradigms have long been used in general auditory but also speech perception research [7, 9, 10, 11, 12]. The linguistic MMN is thought to arise not only from auditory change detection but also from the representation of speech sounds in long term memory that facilitate the discrimination process, e.g. [7, 9]. [9] have shown that the MMN in linguistic research can be used to establish an auditory discrimination profile, taking into account duration, intensity, pitch and vowel differences. Especially vowel quantity, vowel quality [11] or both together [9, 12, 13] have been researched with the MMN. Vowel quantity, contrary to other linguistic features, has been shown to be more right-lateralized [9, 12] and to elicit bigger MMN amplitudes than vowel quality changes [9], while simultaneous changes in quantity and quality elicited the biggest amplitude [13]. The MMN thus seems well suited to investigate the neuro-physiological processes underlying congenital amusia.

A number of studies investigating pitch perception in amusia have made use of the MMN already, but with heterogeneous findings: Braun et al. [14] found the MMN to be absent in amusics for melodies containing altered notes, while Moreau et al. [15, 16], using melodies and piano tone sequences, found normal MMNs in amusics. Reduced, abnormal MMNs were found by Nan et al. [17] for a subgroup of tone-language speaking amusics as responses to lexical tones. Taken together, the aforementioned findings show that at least some amusics seem to have absent or reduced MMNs to tonal sequences. All studies highlighted that it is important to carefully screen the amusics and to test a group that is as homogeneous as possible in their deficits in order for a clear picture to emerge.

Based on this, the present study looks at a group of amusics showing both a pitch and a rhythm deficit of a non-tonal language, German, with a contrast in vowel quality and quantity. We hypothesize that these amusics will show reduced MMNs in comparison to controls. We also assess whether the expected deficit is present both in vowel quality and quantity, expecting both will be impaired due to the participants' deficit in pitch and rhythm perception.

2. MATERIALS AND METHODS

2.1. Participants

11 amusics and 11 controls matched for age, gender, handedness, education and musical training participated. All participants were native speakers of German, right-handed and had no self-reported psychological or neurological disorders. They all had normal hearing defined as a mean hearing level of 20 dB or less in both ears. Congenital amusia was diagnosed based on the three pitch-based and the rhythm subtest of the Montreal Battery of Evaluation of Amusia [18] and a detailed questionnaire about their educational and musical background. Only amusics exhibiting both a pitch perception and a rhythm perception deficit were included in this study to ensure homogeneity as much as possible.

2.2. Stimuli

Our stimuli were isolated synthetic vowels based on auditory properties of natural German front mid vowels. We decided to use mid vowels to avoid periphery effects, and to utilize vowels that are close to each other in their height and front-back dimension in the vowel space, but that differ in quality and/or quantity. Those considerations left us with the German front mid vowels /ɛ/ (short and more open), /ɛː/ (long and more open), /e/ (short and more closed) and /eː/ (long and more closed). All four vowels occur in German, however only three of them, /ɛ/, /ɛː/ and /eː/, are native to German and therefore regarded as phonemes, at least adopting a standard view [19], see the contrast ***B****etten* /ɛ/ 'beds' – ***bä****ten* /ɛː/ 'if they requested' – ***b****eten* /eː/ 'to pray', while /e/ occurs only in loanwords in unstressed position, e.g. in *Chemie* 'chemistry'. However, /e/ is considered to be a phoneme by some phonologists [20, 21].

The durational and formant values for the short, more open vowel /ɛ/ and the long, more closed vowel /eː/ that we employed as basis for the creation of our stimuli can be found in Table 1. They are based on acoustic measurements by Jessen [22]. The values of the other two vowels can be extrapolated from this.

Vowel	Duration (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)
/e:/	110	350	2157	2793
/ɛ/	60	524	1869	2624

Table 1: Duration and formant values of the two German vowels /e:/ and /ɛ/ from Jessen [22].

2.3. EEG paradigm, recording and pre-processing

Participants completed a passive listening task while watching a silenced nature documentary without subtitles. Participants were instructed to disregard the sounds they were exposed to and to focus their attention on the movie. The auditory stimuli were presented at 60 dB via two loudspeakers.

The auditory stimuli were presented in a multi-deviant oddball paradigm with four blocks, which were counterbalanced across participants. A total of 3600 stimuli occurred in each block. In each block, one vowel was the standard, while the other three vowels served as deviants. The standard occurred 85% of the time and each deviant occurred 5 % of the time. The ISI was varied randomly between 400 ms and 600 ms to avoid entrainment effects to the stimulus chain. Across the four blocks, each vowel occurred once as standard and three times as deviant. The MMN was calculated per participant by subtracting the average ERP of the standard from each of the three deviants per block, resulting in MMNs for 12 different conditions: The four vowels occurring in three types of contrasts each, namely a formant contrast, a duration contrast and a formant and duration contrast simultaneously. Per condition, the most negative peak in the 100 to 250 ms after stimulus onset was determined.

The EEG was recorded using a BioSemi Active Two system with 64 active electrodes. 7 further electrodes were placed on the tip of the nose, on the mastoid, below and around the eyes. The EEG signal was recorded at 8192 Hz and later down-sampled to 512 Hz. The subsequent analyses were performed in Praat [23]. The data were offline referenced to the average mastoids channels and bandpass filtered from 1 Hz to 25 Hz. Continuous data was segmented into 500-ms epochs. The EEG was segmented into epochs of 500 ms with a 100-ms baseline. Artifact correction was done automatically and all epochs with activity exceeding +/−75 µV were excluded.

3. RESULTS

The analysis was run on the MMN amplitude measured at 9 channels (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4).

Visual analysis confirmed the negative polarity, the expected latency and fronto-central scalp distribution of the MMN. Figure 1 shows the average difference waveform averaged across all conditions for amusics (dotted line) and controls (solid line) and the scalp topography (the darker blue, the more negative). The figure confirms a difference between amusics and controls.

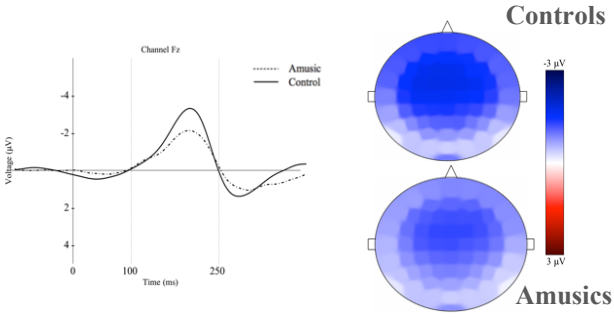


Figure 1: Difference curves (deviant – standard) for amusics and controls in a time window between 100 and 250 ms averaged across all conditions. On the left are the grand average difference waves plotted at Fz, and on the right topographical maps.

Figure 2 shows the grand average difference waves between 100 and 250 ms plotted at Fz for amusics and controls split per condition. The figure illustrates the, on average, greater negativity in the control group, depicted by the solid line. It also seems to indicate differences between the conditions as seen for example in the higher amplitude in the last column.

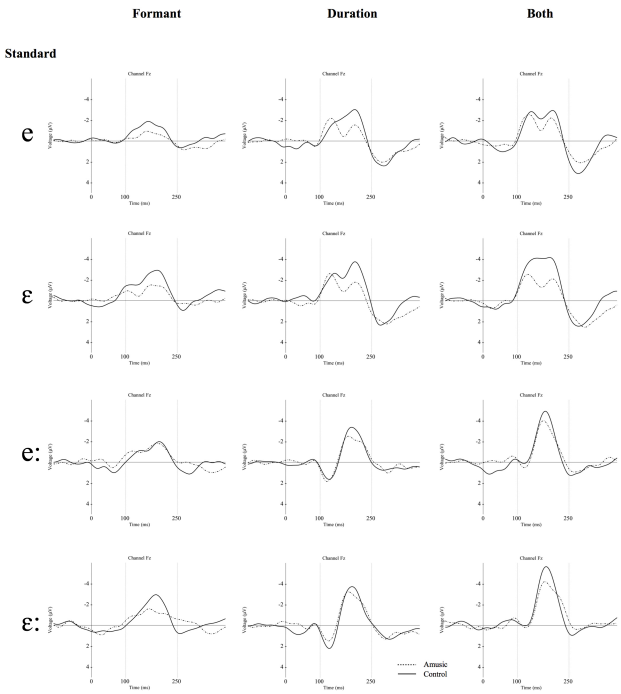


Figure 2: Grand average difference waves plotted at Fz for amusics and controls per condition.

After visual analysis, we performed two-tailed *t*-tests against zero separately for each group to determine

whether the difference waveform response was present in every condition, all of which were significant. Significance levels and mean values at Fz are given in Table 2.

Standard	Formant	Duration	Both
/e/	-2.15*** -1.40***	-3.15*** -2.34***	-3.86*** -3.09***
/ɛ:/	-2.98*** -1.96**	-3.84*** -2.69***	-4.63*** -2.94***
/e:/	-2.18*** -2.38***	-2.83*** -2.40***	-4.03*** -3.75***
/ɛ:/	-2.60*** -2.17*	-3.22*** -2.93***	-4.71*** -3.98***

Table 2: Mean voltage at electrode Fz measured in µV. Top value is that of the control group, bottom value that of the amusic group. * indicate significance levels in *t*-tests against zero: * p <0.05; ** p < 0.01; *** p < 0.001.

Next, we performed an independent samples *t*-test between the groups, which was also highly significant ($t(2374)$= -9.59, p < 0.001). As we were interested in the differences between amusics and controls in every condition, we calculated a linear mixed model (lmer) in R [24] with subject as random effects and group and condition as fixed factors. We found significant main effects for group ($t(23.7)$ = -2.43, p = 0.023) with amusics (M = -2.67) overall having a smaller MMN than controls (M= -3.35). In addition, we found a main effect for condition ($t(2351.8)$ = -6.14, p < 0.001) and a significant interaction between group and condition ($t(2351.8)$ = 3.85, p < 0.001). Figure 2 serves as a visualisation of this.

To understand the Group by Condition interaction, *t*-tests were performed for each condition separately. They all yielded significant results, i.e. controls having a bigger MMN, in all but one condition, namely from standard /ɛ:/ to deviant /e:/.

4. DISCUSSION AND CONCLUSION

Our ERP results show that the auditory system of amusics has formed a representation of the repetitive aspect of the standard stimulus as represented by the MMN that they exhibit. However, their MMN is significantly reduced in comparison to our control population and might therefore represent a more inaccurate discrimination of the stimuli than the controls' larger MMN does [25]. The general finding that amusics did indeed display an MMN response is in direct opposition to that by Braun et al. [14], who did not find an MMN in amusics at all. Furthermore, the fact that the response of our amusics was significantly reduced is in opposition to the findings by Moreau et al. [15, 16] who found amusics displayed completely normal MMNs. All these findings utilized musical stimuli, however. Our study supports Nan et al.'s [17] results, who also found a reduced MMN for tone-language speaking amusics in response to lexical tones. Our findings are also in line with [9, 13], as durational changes elicited bigger MMN amplitudes than vowel quality changes, and simultaneous changes in quantity and quality elicited the biggest amplitude.

A point of criticism concerning our analysis could be the fact that we calculated our MMN by subtracting the average ERP of the standard from each of the three deviants per block, which can be seen as an auditory MMN. Another option would have been to calculate the MMN by subtracting the average ERP of the standard from the three deviant ERPs of the same vowel in different blocks. This would have ensured the resulting difference wave reflected the phonological contrast rather than the acoustic difference between vowels within a block.

Another criticism concerning our methodology might be the small sample size for an ERP study. This is a pitfall that all studies with amusics face, as the population is a rather small one. However, when considering other ERP studies with amusics, our sample size is fairly large in comparison.

Our study could also be criticized concerning the choice of stimuli: We opted for synthesized vowels instead of natural ones as we favored to have control over all parameters, while we ensured at the same time that the vowels sounded as natural as possible. Secondly, we could have chosen four vowels that are all considered German native phonemes by traditional accounts of German phonology such as [19]. The choice of the presently employed mid vowels, however, avoided periphery effects.

We also chose vowels in isolation on purpose, even though one might argue that syllables with transitional cues or whole words with meaning might have been better suited. We wanted to focus purely on spectral and durational vowels cues, while disregarding other cues for the moment, hence our choice.

We were able to demonstrate that amusics did indeed show an MMN, as an automatic reaction to auditory change detection to different vowels. However, this MMN was significantly reduced in comparison to our control population, indicating abnormal neural processes even at this very early stage of processing.

Our findings show that amusics exhibit difficulties when it comes to vowel processing at least in isolation. Further studies investigating later ERP components, as well as more complex linguistic stimuli such as syllables or words are therefore warranted.

5. REFERENCES

[1] Ayotte, J., Peretz, I., Hyde, K. 2002. Congenital amusia - A group study of adults afflicted with a music-specific disorder. *Brain, 125*, 238–251.

[2] Patel, A., Wong, M., Foxton, J., Lochy, A., Peretz, I. 2008. Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception, 25*, 357–368.

[3] Liu, F., Patel, A. D., Fourcin, A., Stewart, L. 2010. Intonation processing in congenital amusia: discrimination, identification and imitation. *Brain, 133*(6), 1682–1693.

[4] Liu, F., Xu, Y., Patel, A. D., Francart, T., Jiang, C. 2012. Differential recognition of pitch patterns in discrete and gliding stimuli in congenital amusia: Evidence from Mandarin speakers. *Brain and Cognition, 79*, 209–215.

[5] Zhang, C., Shao, J., Huang, X. 2017. Deficits of congenital amusia beyond pitch: Evidence from impaired categorical perception of vowels in Cantonese-speaking congenital amusics. *PLoS ONE*, 12(8), e0183151.

[6] Huang, W.T., Liu, C., Dong, Q. Nan, Y. 2015. Categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Frontiers in Psychology*, 6(829).

[7] Näätänen, R. 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology, 38*(1), 1–21.

[8] Näätänen, R., Paavilainen, P., Rinne, T., Alho, K. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590.

[9] Partanen, E., Vainio, M., Kujala, T., Huotilainen, M. 2011. Linguistic multifeature MMN paradigm for extensive recording of auditory discrimination profiles. *Psychophysiology*, 48(10), 1372–1380.

[10] Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P., Näätänen, R. 2006. Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. *Brain Research*, 1072(1), 175–185.

[11] Chládková, K., Escudero, P., Lipski, S. C. 2013. Pre-attentive sensitivity to vowel duration reveals native phonology and predicts learning of second-language sounds. *Brain and Language*, 126(3), 243–252.

[12] Kirmse, U., Ylinen, S., Tervaniemi, M., Vainio, M., Schröger, E., Jacobsen, T. 2008. Modulation of the mismatch negativity (MMN) to vowel duration changes in native speakers of Finnish and German as a result of language experience. *International Journal of Psychophysiology, 67*(2), 131–143.

[13] Ylinen, S., Huotilainen, M., Näätänen, R. 2005. Phoneme quality and quantity are processed independently in the human brain. *NeuroReport,* 16(16), 1857–1860.

[14] Braun, A., McArdle, J., Jones, J. L., Nechaev, V., Zalewski, C., Brewer, C., Drayna, D. 2008. Tune Deafness: Processing Melodic Errors Outside of Conscious awareness as Reflected by Components of the Auditory ERP. *PLoSONE, 3*(6), 1–6.

[15] Moreau, P., Jolicœur, P., Peretz, I. 2009. Automatic Brain Responses to Pitch Changes in Congenital Amusia. *Annals of the New York Academy of Sciences, 1169*(1), 191–194.

[16] Moreau, P., Jolicœur, P., Peretz, I. 2013. Pitch discrimination without awareness in congenital amusia: Evidence from event-related potentials. *Brain and Cognition, 81*, 337–344.

[17] Nan, Y., Huang, W.-T., Wang, W.-J., Liu, C., Dong, Q. 2016. Subgroup differences in the lexical tone mismatch negativity (MMN) among Mandarin speakers with congenital amusia. *Biol Psychol, 113*, 59–67.

[18] Peretz, I., Champod, S., Hyde, K. 2003. Varieties of Musical Disorders: The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences,* 999, 58–75.

[19] Wiese, R. 1996. *The Phonology of German.* Oxford: Clarendon Press.

[20] Wurzel, W. U. 1981. Phonologie: Segmentale Struktur. In: K. E. Heidolph et al. (ed.) *Grundzüge einer deutschen Grammatik.* 898–990. Berlin: Akademie-Verlag.

[21] Giegerich, H. 1985. *Metrical Phonology and Phonological Structure: German and English*. Cambridge: Cambridge University Press.

[22] Jessen, M. 1993. Stress conditions on vowel quality and quantity in German. *Working Papers of the Cornell Phonetics Laboratory* 8, 1–27.

[23] Boersma, P., Weenink, D. 2016. *Praat: doing phonetics by computer* (Version 6.0.15). Retrieved from http://www.praat.org/

[24] R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. See www.R-project.org/.

[25] Kujala, T., Näätänen, R. 2001. The mismatch negativity in evaluating central auditory dysfunction in dyslexia. *Neuroscience & Biobehavioral Reviews*, 25(6), 535–543.

NOT ALL PHONOLOGICAL NEIGHBORS AFFECT PRODUCTION EQUIVALENTLY: PREDICTIONS FROM A NEURAL DYNAMIC MODEL

Michael C. Stern & Jason A. Shaw

Yale University
michael.stern@yale.edu, jason.shaw@yale.edu

ABSTRACT

Phonological neighborhood density (PND) has been shown to have apparently inconsistent effects on speech production: facilitation, inhibition, hypoarticulation, hyperarticulation, or no effect. We propose that this variety of results is due to the coarseness of the PND measure. We present a neural dynamic model of word pronunciation planning in which phonological neighbors influence phonetic target planning via a combination of excitatory and inhibitory interactions. We demonstrate through simulations how the model predicts different effects of PND depending on the identities of the phonetic dimensions differentiating neighbors. We outline novel predictions of the model that can help guide future empirical work.

Keywords: phonological neighborhoods; hyper-articulation; speech planning; Dynamic Field Theory

1. INTRODUCTION

A number of studies have investigated the effect of phonological neighborhood density (PND) on speech production, finding mixed results, e.g. facilitation [1], [2], inhibition [3], [4], hypoarticulation [5], hyperarticulation [6]–[11], or no effect [12]. The PND of a word is defined in these studies as the number of phonological neighbors in the lexicon, where a phonological neighbor is any word that differs by the addition, deletion, or substitution of a single phoneme [13]. Recently proposed models of phonetic planning based on Dynamic Field Theory (DFT) [14] have derived effects of lexical competitors on speech production, like minimal-pair-induced "contrastive hyperarticulation" [15] and "phonetic trace effects" in speech errors [16]. These lexical competition effects are similar to those that have been reported to be conditioned by PND. To probe PND effects, we present a new model that explicitly couples lexical planning to phonetic target planning within the framework of DFT. In doing so, we shed light on the existing diversity of results regarding PND effects in speech production, and generate new testable predictions.

2. MODEL STRUCTURE

A supplemental document with model equations and descriptions, as well as MATLAB scripts for running the simulations reported below, are available on OSF at https://osf.io/kc7bp/. In the model, neural populations governing the activation of lexical items are dynamically coupled to neural populations governing phonetic planning. Patterns of activation within each population unfold over time according to differential equations defined in DFT [14]. We focus on the planning of three phonetic dimensions relevant to the initial consonant: voice onset time (VOT; voicing), constriction location (CL; place of articulation), and constriction degree (CD; manner of articulation). Following [17], we define CL as a single dimension, distinct from the articulator(s) used to make the constriction. The dynamics of phonetic planning—modeled with dynamic neural fields (DNFs)—are described in [15], [16]. The novel mechanisms introduced in this paper allow dynamic coupling of lexical items to phonetic DNFs. Quite straightforwardly, a target lexical item excites its corresponding phonetic DNFs. Our proposal is that neighbors inhibit just those phonetic DNFs that differentiate them from the target. We implement this proposal with a novel mechanism for lexical-phonetic interaction that dictates the polarity of competitor activation based on the current state of phonetic DNFs.

Like antecedent DFT models [15], [16], [18]–[20], the coupling between a lexical item and a phonetic DNF is defined as a *distribution*, such that an active lexical item beginning with, e.g., a voiceless consonant, projects a distribution of activation to the VOT DNF centered on higher VOT values. Each coupling is defined as a normal distribution with position p (mean) and width w (standard deviation).

Lexical items are modeled as dynamic nodes (no internal feature gradience) with linear activation dynamics (no internal interaction). The rate of change of lexical activation is negatively related to current activation, defining point attractor dynamics. The location of the point attractor is determined by input to the lexical node (e.g., from conceptual-intentional planning, or from feedback from phonetic planning, as discussed below) and noise.

Lexical items receive feedback from phonetic DNFs to which they are coupled. This feedback causes phonological neighbors to become somewhat active during planning of a target word, and also facilitates activation of the target word. The magnitude of input from each DNF neuron to a coupled lexical item is a sigmoidal function of the neuron's current activation, such that relatively inactive neurons contribute virtually no input to coupled lexical nodes, while neurons with activation exceeding a threshold contribute substantially. Input from phonetic DNF neurons to coupled lexical nodes is modulated by the metric distance between the neuron's position in the DNF and the center of the coupling distribution p. Active neurons with a metric position close to the center of the coupling distribution (within one w of p) contribute substantially, while neurons with a position that exceeds one w from p contribute exponentially less. Crucially, the threshold for phonetic-to-lexical input is lower than the threshold for within-DNF lateral interaction (and therefore, activation peak stabilization), allowing phonological neighbors to become active before phonetic target planning is complete.

Active lexical nodes project input to phonetic DNFs to which they are coupled. The shape of this input is defined by the coupling distribution (p and w). The magnitude of this input is a positive linear function of lexical activation, such that more active lexical items project stronger input to coupled phonetic DNFs compared to less active lexical items. The polarity of this input (excitatory or inhibitory) is modulated by δ: the difference between the current state of the DNF and the lexical item's "preferred" state of the DNF, i.e., the state that would be induced by excitatory projection from the lexical item to the DNF. Polarity is a sigmoidal function of δ. When δ is relatively small (i.e., the DNF does not differ substantially from the lexical item's "preferred" state), input is excitatory. When δ exceeds a threshold (i.e., the DNF is in a "dispreferred" state for the lexical item), input from the lexical item becomes inhibitory. Crucially, the polarity of input is independent of the magnitude of input. A very active lexical item will project strong input to coupled DNFs; whether this input is excitatory or inhibitory is determined by the state of the DNF at any given time, as indexed by δ. This innovation derives the inhibitory effect of neighbors on phonetic dimensions that differ from the target.

3. SIMULATION RESULTS

We simulate initial consonant planning for a target word beginning with /t/ in three conditions, summarized in Table 1.

	Target	**Comp. 1**	**Comp. 2**
Cond. 1 "2-voiced"	/t…/ 'teen'	/d…/ 'dean'	/b…/ 'bean'
Cond. 2 "1-voiced"	/t…/ 'ten'	/d…/ 'den'	/p…/ 'pen'
Cond. 3 "0-voiced"	/t…/ 'tin'	/p…/ 'pin'	/k…/ 'kin'

Table 1: Initial consonant of the target word and each competitor in each simulation condition. Example words are given in quotes.

In each condition, the target word has two phonological neighbors differing only in the initial consonant. Thus, PND is the same in all three conditions. However, the conditions differ with respect to the phonetic dimensions differentiating the neighbors. In the 2-voiced condition, both neighbors differ from the target in initial consonant voicing (beginning with /d/ or /b/). In the 1-voiced condition, one neighbor differs from the target in initial consonant voicing (beginning with /d/), but the other neighbor overlaps with the target on initial consonant voicing (beginning with /p/). In the 0-voiced condition, both neighbors overlap with the target on initial consonant voicing (beginning with /p/ or /k/). In all three conditions, all neighbors begin with stops, so they all overlap on the CD dimension. The CD DNF thus serves as a proxy for the dimensions that are shared between all neighbors, allowing some activation to spread to all neighbors.

In all simulations, the target lexical item receives positive input, representing an intention to say that word. No other lexical item, and no phonetic DNF, receives direct input; rather, they receive activation only via interactions within the system. Figure 1 shows the timecourse of activation of the three lexical items (top) and the VOT DNF (bottom) for a single run in each condition. In all conditions, the two phonological neighbors become active via feedback from phonetic dimensions on which they overlap with the target. In the 2-voiced condition (left), both neighbors project inhibitory input to the voiced side of the VOT DNF, because by the time the neighbors start receiving activation, the state of the VOT DNF is more consistent with a voiceless consonant than a voiced consonant, resulting in a high value of δ for the neighbors. Inhibitory input to the voiced side of the DNF causes a rightward shift in the location of the activation peak, i.e., a hyperarticulated voiceless target. In the 1-voiced condition (center), only the neighbor beginning with /d/ projects inhibitory input to the voiced side of the DNF, so the rightward peak shift is smaller (hyperarticulation is decreased). Finally, in the 0-voiced condition, there is no inhibitory input to the VOT DNF, and thus no hyperarticulation. By-condition differences in the

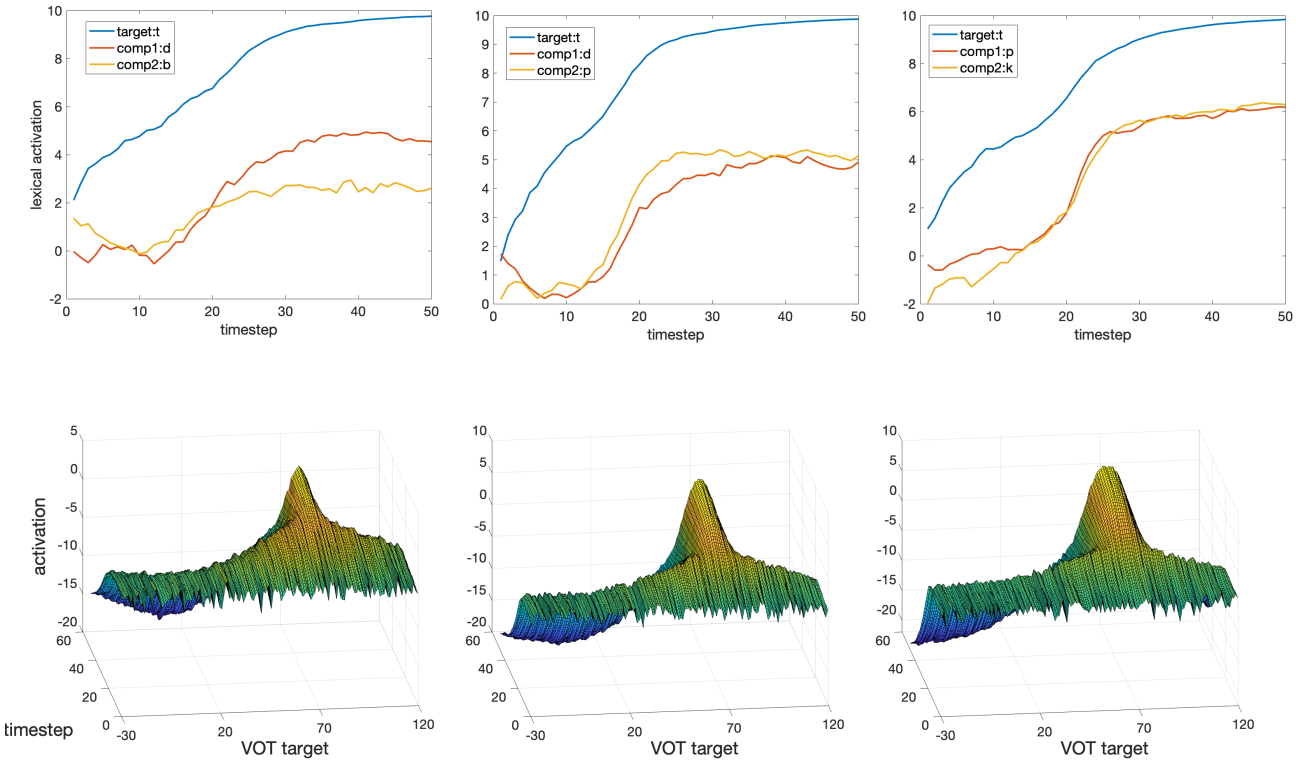


Figure 1. Lexical node (top) and VOT DNF (bottom) activation over time, single run. Left: two inhibitory competitors (2-voiced). Center: One inhibitory, one excitatory competitor (1-voiced). Right: two excitatory competitors (0-voiced).

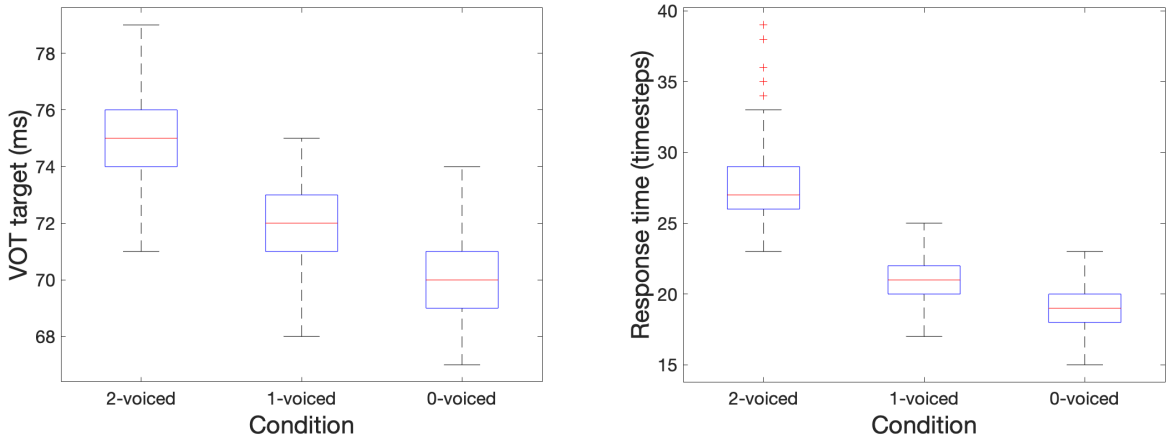


Figure 2. VOT target (left) and time to plan VOT target (right) in 500 simulated utterances in each condition.

time until an activation peak stabilizes can also be seen: slowest in the 2-voiced condition, followed by 1-voiced, followed by 0-voiced.

To investigate the systematicity of these patterns, we simulated 500 productions in each of the three conditions and recorded the VOT target (metric position of the VOT DNF neuron with maximum activation at the final timestep) and response time (peak of the first derivative of total lateral interaction) for each simulation. The target word always projected input to the VOT DNF centered at $p = 70$. As seen in Figure 2, a cline of VOT hyperarticulation was observed, with the most hyperarticulation in the 2-voiced condition, followed by 1-voiced, followed by 0-voiced. An analogous pattern was observed in response time.

4. DISCUSSION

Our model of the neural dynamics of lexical and phonetic planning demonstrated that different kinds of phonological neighborhoods are predicted to have different effects on articulation and planning time, even when PND is equal. The model thus offers a possible explanation underlying the existing diversity of results regarding the effects of PND on

speech production, and offers testable predictions to guide future empirical work. As a first step, we expect that a better predictor than PND would be the *ratio of inhibitory to excitatory neighbors* on a given dimension. This ratio is predicted to positively correlate with the magnitude of hyperarticulation on that dimension, and the time it takes to plan articulation on that dimension. Another prediction relates to the lexical-phonetic coupling widths w. In the simulations, VOT was hyperarticulated due to overlap from wide inputs $w = 30$ ms. We assume that input widths transparently reflect the distribution of the phonetic cue in the listener's environment. This assumption is based on the finding that a VOT input width of 30 ms—the approximate standard deviation of VOT in American English stops [21]—derives contrastive hyperarticulation in non-errors [15] and trace effects in errors [16] of empirically observed magnitudes. We thus predict that overlap of phonetic cues in corpus distributions will correlate with the magnitude of neighborhood-induced hyperarticulation.

A third prediction relates to the feedback mechanism which activates phonological neighbors. It can be seen in Figure 1 (top, center) that the competitor beginning with /p/ received slightly more activation than the competitor beginning with /d/, even though both /d/ and /p/ differ from /t/ in only a single feature. This is because the VOT distribution (which /t/ and /p/ share) is wider (w = 30 ms) than the CL distribution (w = 10 mm), which /t/ and /d/ share. Since phonetic-to-lexical input is sensitive to w, the VOT DNF ends up contributing more activation to coupled lexical items than the CL DNF. This predicts that corpus variability on a phonetic dimension will correlate with the degree of excitatory lexical interaction between neighbors that overlap on that dimension, measurable via, e.g., priming or error rate. Broadly, this suggests that not only the number of overlapping phonetic dimensions, but also the way that neighbors are coupled to those dimensions, shapes the outcome of lexical interaction (cf. [22], [23]). The last prediction that we point out is related to the width of the activation peaks. It can be seen in Figure 1 (bottom) that activation peaks tend to be wider when there are more excitatory than inhibitory neighbors. The model thus predicts that words with more phonological neighbors overlapping on a particular phonetic dimension will tend to be *more variable* on that dimension.

Before concluding, we briefly discuss a few ways that this model should continue to be expanded. In our simulations, we took the time for an activation peak to stabilize in the VOT DNF as a proxy for time to speech initiation. However, all dimensions relevant to a consonant, or perhaps the whole syllable or set of upcoming syllables, likely interact in determining the time it takes to begin speaking [24]. While we expect that VOT planning time should correlate with response time, a more complete model would incorporate additional phonetic dimensions when relating model behavior to human response times. This relates to a second avenue for model development. In the current model, the neural processes governing lexical and phonetic planning unfold over time. However, the temporal coordination of phonetic dimensions is not planned. Like other models of PND in speech production (e.g., [25]), all phonetic dimensions are planned simultaneously, with no mechanism to sequence their articulation. This is likely insufficient, since the position in the word at which phonological neighbors differ has been found to influence lexical interaction [7]. The model should be elaborated to include planning of temporal coordination, while capturing these empirical effects.

5. CONCLUSION

We presented a neural dynamic model of word pronunciation planning based on Dynamic Field Theory (DFT). The model expands on previous related models by adding lexical items, and formalizing dynamic coupling mechanisms between lexical items and dimensions of phonetic planning. These coupling mechanisms account for activation of phonological neighbors during target word planning, as well as the influence (inhibitory or excitatory) of phonological neighbors on phonetic planning. We demonstrated through simulations that the ratio of phonological neighbors that differ on a phonetic dimension ("inhibitory" neighbors) to neighbors that overlap on a phonetic dimension ("excitatory" neighbors) is expected to be a better predictor of hyperarticulation on that dimension than global phonological neighborhood density (PND), offering a possible explanation of the existing diversity of results, as well as predictions to guide future empirical work.

As a final point, the flow of activation in the model (both laterally within DNFs, and between lexical items and DNFs) is highly non-linear, allowing stabilization over time on a particular word's pronunciation plan. However, the primitive features of the model are all continuous, allowing the generation of gradient shifts in pronunciation and planning time, even in the face of the categoricity generated by the non-linearities. A general strength of this kind of modeling approach is that gradience and categoricity co-exist in the same dimensions.

6. REFERENCES

[1] M. S. Vitevitch, "The Influence of Phonological Similarity Neighborhoods on Speech Production," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 28, no. 4, pp. 735–747, 2002, doi: 10.1037/0278-7393.28.4.735.

[2] E. Buz and T. F. Jaeger, "The (in)dependence of articulation and lexical planning during isolated word production," *Lang. Cogn. Neurosci.*, vol. 31, no. 3, pp. 404–424, 2016, doi: 10.1080/23273798.2015.1105984.

[3] A. F. Frank, M. K. Tanenhaus, R. N. Aslin, and A. P. Salverda, "Frequency, neighborhood density, and phonological similarity effects in picture naming: An artificial lexicon study," *Proc. 29th Annu. Cogn. Sci. Soc.*, pp. 995–1000, 2007.

[4] J. K. Gordon and J. C. Kurczek, "The aging neighborhood: phonological density in naming," *Lang. Cogn. Neurosci.*, vol. 29, no. 3, pp. 326–344, Mar. 2014, doi: 10.1080/01690965.2013.837495.

[5] S. Gahl, Y. Yao, and K. Johnson, "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," *J. Mem. Lang.*, vol. 66, no. 4, pp. 789–806, 2012, doi: 10.1016/j.jml.2011.11.006.

[6] N. P. Fox, M. Reilly, and S. E. Blumstein, "Phonological neighborhood competition affects spoken word production irrespective of sentential context," *J. Mem. Lang.*, vol. 83, pp. 97–117, 2015, doi: 10.1016/j.jml.2015.04.002.

[7] M. Fricke, M. M. Baese-Berk, and M. Goldrick, "Dimensions of similarity in the mental lexicon," *Lang. Cogn. Neurosci.*, vol. 31, no. 5, pp. 639–645, 2016, doi: 10.1080/23273798.2015.1130234.

[8] M. Fricke, "Phonological enconding and phonetic duration," Doctoral dissertation, University of California, Berkeley, 2013.

[9] B. Munson and N. P. Solomon, "The effect of phonological neighborhood density on vowel articulation," *J. Speech, Lang. Hear. Res.*, vol. 47, no. 5, pp. 1048–1058, 2004, doi: 10.1044/1092-4388(2004/078).

[10] B. Munson, "Lexical access, lexical representation, and vowel production," *Lab. Phonol.*, vol. 9, pp. 201–228, 2007.

[11] R. Scarborough, "Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation," *J. Phon.*, vol. 41, no. 6, pp. 491–508, 2013, doi: 10.1016/j.wocn.2013.09.004.

[12] J. J. Holliday and R. Turnbull, "Effects of Phonological Neighborhood Density on Word Production in Korean," *Proc. 18th Int. Congr. Phonetic Sci. (ICPhS 2015)*, pp. 1–5, 2015.

[13] P. A. Luce, "Neighborhoods of Words in the Mental Lexicon," Doctoral dissertation, Indiana University, 1986.

[14] G. Schöner, J. Spencer, and D. R. Group, *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford: Oxford University Press, 2016.

[15] M. C. Stern and J. A. Shaw, "Neural inhibition during speech planning contributes to contrastive hyperarticulation," *arXiv Prepr. arXiv2209.12278*, 2022, doi: https://doi.org/10.48550/arXiv.2209.12278.

[16] M. C. Stern, M. Chaturvedi, and J. A. Shaw, "A dynamic neural field model of phonetic trace effects in speech errors," *Proc. Annu. Meet. Cogn. Sci. Soc. 44*, 2022.

[17] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989, doi: 10.1017/S0952675700001019.

[18] C. Kirov and A. I. Gafos, "Dynamic phonetic detail in lexical representations," *ICPhS XVI*, pp. 637–640, 2007.

[19] A. Gafos and C. Kirov, "A dynamical model of change in phonological representations: The case of lenition," *Approaches to Phonol. Complex.*, pp. 219–240, 2009, doi: 10.1515/9783110223958.219.

[20] K. D. Roon and A. I. Gafos, "Perceiving while producing: Modeling the dynamics of phonological planning," *J. Mem. Lang.*, vol. 89, pp. 222–243, 2016, doi: 10.1016/j.jml.2016.01.005.

[21] E. Chodroff and C. Wilson, "Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English," *J. Phon.*, vol. 61, pp. 30–47, 2017, doi: 10.1016/j.wocn.2017.01.001.

[22] P. C. Gordon and D. E. Meyer, "Perceptual-motor processing of phonetic features in speech," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 10, no. 2, pp. 153–178, 1984, doi: 10.1037/0096-1523.10.2.153.

[23] A. Ulicheva, K. D. Roon, Z. Cherkasova, and P. Mousikou, "Effects of phonological features on reading-aloud latencies: A cross-linguistic comparison," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 48, no. 9, pp. 1348–1362, 2021, doi: 10.1037/xlm0000893.

[24] E. Bates *et al.*, "Timed picture naming in seven languages," *Psychon. Bull. Rev.*, vol. 10, no. 2, pp. 344–380, 2003, doi: 10.3758/BF03196494.

[25] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: the neighborhood activation model.," *Ear Hear.*, vol. 19, no. 1, pp. 1–36, Feb. 1998, doi: 10.1097/00003446-199802000-00001.

WORDS OR RULES: PHONOLOGICAL MUTATIONS IN WUBUY

R. L. Bundgaard-Nielsen[1], B. J. Baker[2], & Y. Wang[2]

MARCS Institute for Brain, Behaviour and Development, Western Sydney University[1]; School of Languages & Linguistics, University of Melbourne[2].
rikkelou@gmail.com; bjbaker@unimelb.edu.au; yizhouw4@student.unimelb.edu.au

ABSTRACT

The present study addresses the question of whether complex morphological sequences are the result of rule application or relational or associative morphological processes/schemas and extensive listings in the lexicon. It does so by testing the effect of phonological hardening (a phonological mutation of syllable initial sonorants, turning them into stops) on word-recognition in the highly polysynthetic Australian Indigenous language Wubuy, in a Two-Alternate Forced Choice paradigm. The results indicate that it is harder to recognise the correct word-form when mutation applies, than when it does not, and that recognising the correctly mutated form may not happen under all conditions. These results are consistent with rule-application approaches to language processing rather than associative/relational models, relying on more listings in the lexicon.

Keywords: Polysynthesis, phonological alternations, word-recognition, word processing.

1. INTRODUCTION

The present study addresses the question of whether complex morphological sequences are generated as the result of rule application (syntactic, morphological, and/or phonological) as is assumed under generativist/Chomskyan models of language processing [1], [2] or whether they can be accounted for by relational or associative morphological processes/schemas and extensive listings in the lexicon (e.g., [3] as well as for instance in [4]'s Short-List B.

The present study addresses this question by testing the effect of the application of morpho-phonological alternations on the process of word-recognition in the highly polysynthetic Australian Indigenous language Wubuy [5]. Wubuy implements two complex, parallel morpho-phonologically processes, *phonological hardening* (the focus here) and /ŋu/-insertion, which offer an ideal testing ground for the very fundamental question of words versus rules. Wubuy Hardening, just like the perhaps more well-known Welsh mutations [6] results in significant phonological changes to initial segments of surface forms of morphemes under certain conditions.

Phonological hardening in Wubuy is a process of consonant alternation which affects morpheme initial approximants when the previous morpheme ends in a non-continuant, by realising the approximant as a stop consonant at the same place of articulation. Hardening operates in parallel with a second process—*/ŋu/ insertion*—in which the meaningless string /ŋu/ is inserted before underlyingly (not derived) stop-initial morphemes, when these occur in similar environments: /*nuŋu-ak-walaŋi*/ (we/her-BEN-paint.PAST) → /*nuŋaapalaŋi*/ (/*walaŋi*/ hardened to /*palaŋi*/) versus /*nuŋu-ak-paḻuŋ*/ (we/her-BEN-cut.past) → /*nuŋaaŋupaḻuŋ*/ (with insertion of /*ŋu*/).

Under parallel-associative (PA) and relational models like *Construction Grammar* (CxG; e.g. [7]) and *Relational Morphology* (RM; [3]), language acquisition is "item based" and lexico-grammatical choices are explained primarily by complex patterns of parallel associations between (adjacent) elements. These elements are organised into "independent systems for phonology, syntax, and semantics", connected by linkages which create "correspondences between components rather than *derivations from one component to another"* [3].

Under parallel-associative and similar full-listing models, phonological mutations such as those in Wubuy and Welsh must thus be accounted for as non-derivational and parallel, and the involved morphemes/words must be selected in a one-step process from multiple equivalent forms in the lexicon, on the basis of their co-occurrence patterns in the input. These models generate the prediction that we would observe no difference in processing cost (measured in for instance response time, event related potentials, or response accuracy in

experimental paradigms like the one implemented here) for selection of particular forms, including those forms that have traditionally been referred to as hardened over those that are not hardened: "Hence, the composition of a word or sentence involves clipping together stored pieces in such a way that every element of the composed structure is accounted for in terms of one stored piece or another" [3]. Note that, despite talking of 'words' here, Wubuy verbs are complete propositions, and thus their processing is more akin to the processing of sentences in some respects [13].

Under a rule-based (Chomskyan) analysis [1], individual morphemes have one canonical phonological shape in the lexicon, which must be selected prior to the application of any morphophonological rules that are triggered by the phonological shape of the selected morphemes. This is logically a sequential process (even if constraints on surface forms are computed in parallel, as in Optimality Theory [8]), and the identification and implementation of necessary phonological mutations presumably increase the processing demands of spoken word recognition and comprehension when a mutation must be applied, although we are aware of no work on this specific issue. In the case of the Wubuy hardening phenomenon, we would expect application of a hardening rule to induce greater processing cost than non-application of a rule (no change as the trigger for change is absent), and likely higher error rates and greater time cost associated with rule application. Such asymmetrical performance is at odds with the predictions from parallel-associative models.

2. METHOD

2.1. Participants

14 speakers of Wubuy participated in the experiment. The participants (11 women) ranged in age from approximately 40-70 years. All participants were native speakers of Wubuy, and resident in Numbulwar at the time of testing. The participants were recruited by word of mouth in the community of NumbulwarThe participants constitute app. 25% of the speaking population (see discussion in [9]). This population is predominantly older, as the community is and has been undergoing language-shift to the English-lexified contact variety Kriol [10], [11].

2.2. Materials

The materials consisted of a Two-Alternate Forced Choice experiment in two parts (Exp. 1; Exp. 2). In each part of the experiment, participants were presented with pairs of one of four complex Wubuy words, half of which differed in whether the Wubuy hardening rule had been applied in an obligatory hardening context, or whether hardening had been incorrectly applied in a non-hardening context. Exp. 2 differed from Exp. 1 only in that a 500 ms period of silence was inserted before the target verb in each of the words.

The four complex words consisted of the prefix */nuŋaak/* (we/her-BEN), which induces hardening of following sonorants, even though it is realised as [*nuŋaa*] due to a rule deleting final /k/, followed by one of four incorporated sonorant-initial nouns, which had been correctly hardened, and by one of four verb stems, each in a Non-Hardened (sonorant-initial; S) form and a Hardened (stop-initial; H) form. Two of the nouns (*/ʈarpic/*, 'thigh'*; /ʈaŋak/,* 'branch') are stop-final and thus induce hardening of the initial segment of the following verb stem, while two of the nouns (*/paɲca/*, 'arm'*; /ʈukanta/*, 'leg') do not (See Table 1).

Pref.	Noun	Verb
nuŋaa	ʈarpic	ʟiɲ/ʈiɲ (chop.PAST)
	ʈaŋak	aajuu//kaajuu (cut.PAST)
	paɲca	waɭaɲi/paɭaɲi (paint.PAST)
	ʈukanta	waɭpumana/paɭpumana (paint.PRES)

Table 1: Prefix (Pref.), Nouns and Verbs in Non-Hardened (first) and Hardened (second) forms.

The complex word stimuli for the 2AFC experiments were produced by a literate Wubuy speaker (female, in her 50s), using a PMD660 Marantz flash-RAM digital recorder with a DPA d:fine headset microphone. All recordings had a 16-bit sampling depth with a sampling rate of 44.1 kHz. During the recording session, the speaker was asked to produce [*nuŋaa*] + each of the four nouns followed by pause in which she *imagined* producing a verb stem. This approach ensured that the intonation contour for the

incorporated noun did not bear unusual phrase-final intonation, and that coarticulatory information from a transition into a following verb stem would not influence the participants preference judgements. The speaker produced several utterances of each [*nuŋaa*]+*NOUN* frame, and one token of each frame type was selected on the basis of clarity, pitch, and naturalness in the matchup between the frames.

The four [*nuŋaa*]+*NOUN* frames were cross-spliced with naturally produced recordings of the four non-hardened and the four hardened verb stems elicited by the same speaker with phrase final intonation, in *Praat* [12]. This procedure created 16 word-pairs in which hardening was correctly applied in half and over-applied (i.e., without a preceding non-continuant segment) in the other half. Each pair was presented twice (order counterbalanced) producing 32 pseudo-randomised trials (See Examples 1 and 2):

Non-hardening example, correct form second:

1) [*nuŋaa-panca-ʈiɲ*] *vs* [*nuŋaa-panca-ɭiɲ*]

Hardening example, correct form first:

2) [*nuŋaa-ʈarpic-ʈiɲ*] vs [*nuŋaa-ʈarpic-ɭiɲ*]

2.3. Procedure

Participants completed Exp. 1 and Exp. 2 in a quiet home in Numbulwar, in the presence of the authors and a Wubuy speaker well-known to the participants. All testing instructions were provided in English, but the Wubuy speaker provided instructions also in Wubuy when necessary or requested.

Presentation order of Exps. 1 and 2 was counter balanced so that seven participants completed Exp. 1 (No Pause) first, and seven participants completed Exp. 2 (Pause inserted) first. The participants were presented with the word pairs over headphones from a MacBook computer which displayed two line-drawings of faces, to indicate that two speakers would be heard. The participants were instructed to indicate which of the utterances they preferred (the first or the second), and they could do so by pointing to the first or second line-drawn face on the screen, by saying 'number one' or 'number two' (in English or in Wubuy), or by raising one or two fingers. This approach deviates from the approach typically taken, due to the test-taking profile and preferences of the participants. The participants were free to listen to each word pair as many times as they liked, before recording their response. All participants were given an open-ended rest break between the two parts of the experiment. All testing instructions were provided in English, but a familiar L1 Wubuy speaker was also present before and during testing to provide instructions also in Wubuy when necessary or requested.

2.4. Predictions

If Wubuy word recognition relies on complex patterns of parallel associations between adjacent morphemes in morphological schema which are fully listed, we predict that the participants will show similar levels of accuracy for words with hardened and non-hardened verb stems. We do not expect to see improved performance as a function of artificial pause insertion, given that pausing interrupts the sequence of morphemes and that no processing cost is incurred at the juncture by the application of a morphophonological rule. If on the other hand, Wubuy word recognition relies on sequential application of a morphophonological rule triggered only in half of the experimental trials, we might expect poorer performance on Hardening trials than Non-Hardening trials, and improved performance with increased processing time in the form of pause insertion.

3. RESULTS

The mean percent correct performance in Exp. 1 and Exp. 2 is presented in Figure 1. The average accuracy ranged from 58% (*SD* = .194) and 62% (*SD* = .146) for the Hardened (H) verbs in the two Pause Conditions (A, B) respectively, to 69% (*SD* = .192) to 71% (*SD* = .213) for the Non-Hardened (S) verbs in the two Pause Conditions (A, B). The low averages indicate that the task was challenging, likely due to the two levels of hardening (nouns *and* verbs), and the additional rule deleting final /k/ from the prefix. It may however, also in part be due to the artificiality of tasks of this kind.

To test the competing predictions from models like CxG/RM and more traditional Chomskyan rule-based approaches, we first compared the participants performance on hardening and non-hardening trials with and without pause insertion to chance performance

(50%). This analysis showed that participant accuracy did *not* differ from chance in the *hardened no-pause condition* (p = .0658), but that participants performed above chance on the non-hardened no-pause condition (p = .0002), and on both hardened and non-hardened trials in the pause insertion condition (p = .0128, and p = .0001, respectively).

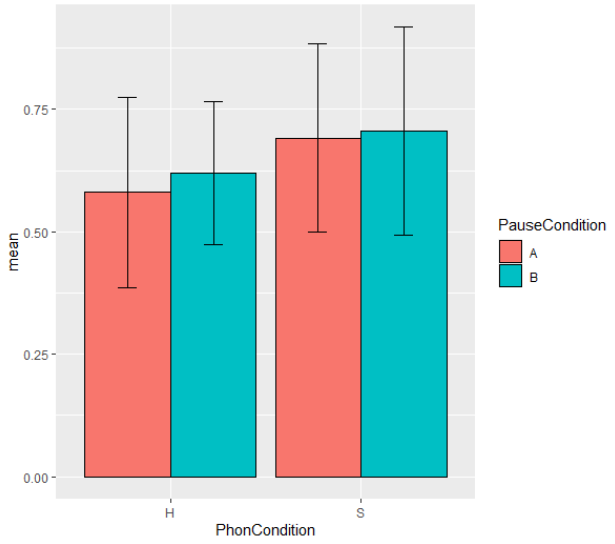


Figure 1: Average accuracy in %. H indicates correct application of hardening rule, S indicates no hardening. A indicates no pause inserted, B indicates 500ms of pause inserted before the verb stem. Error bars indicate SE.

In order to test whether the Wubuy participants' performance is affected by the phonological condition (hardened or non-hardened verb stem) and by pause condition (no pause inserted and 500 ms of silence inserted prior to the verb stem), we built a linear mixed-effects model (LMM, fixed effects: Phonological Condition [Hardened vs Non-Hardened] and Pause Condition [No Pause vs Pause Insertion]; random effects: Participant and Word). We checked the model using a Wald Chi-squared test, which revealed a significant main effect of Phonological Condition (Wald $\chi2$ = 10.6285, p = .0011), but no main effect of Pause Condition (Wald $\chi2$ = 1.6333, p = .2013). There was no interaction (*Wald $\chi2$ = 0.3126, p = .5761*).

Given that we did not find a significant effect of Pause Condition, nor an interaction between Phonological and Pause Conditions, we conducted only *post hoc* comparison for the two Phonological Conditions (Hardened onset vs Non-Hardened onset). This comparison indicated that it was significantly more difficult for the participants to select the correct verb form in the Hardened condition than in the Non-Hardened Condition (estimated means = 64.3% (Hardened) and 80.4% (Non-Hardened), z = 3.281, p = .0010).

4. DISCUSSION

The results of the present experiments clearly indicate that Wubuy speakers find it much harder to recognise the correct Hardened verb forms than correct Non-Hardened verb-forms in a 2AFC paradigm. While we found no main effect of Pause Condition (No-Pause vs Pause Insertion), and therefore cannot conclude that Pause Insertion, as a proxy for extra processing time, improves performance accuracy, we note that performance on Hardened forms with No-Pause was the only condition that did not differ from chance.

The results are consistent with a rule-based account of Wubuy phonological mutations, in that it would appear that application of the Hardening rule decreases accuracy (presumably due to increased cognitive demand, c.f. [14]), and further, that, to some extent, correct rule application appears to benefit from additional processing time (above-chance performance on Hardened forms only in the Pause Condition).

The results cannot be readily accounted for from within a RM (or similar) account of word-formation, where words are "clipped together" from already-stored forms, unless there are highly asymmetrical frequencies in the input (for which we have no clear evidence: all eight verbforms appear with low frequency in the available Wubuy text corpus). A third potential explanation, one in which morpheme-initial stop-sonorant pairs following morpheme-final stops are treated as context-dependent neutralisations, fails to account for the accuracy asymmetry between Hardened and Non-Hardened forms, and we thus also discount this explanation of the phenomenon.

6. ACKNOWLEDGEMENTS

We thank the participants, and Numbulwar community. We acknowledge funding from the Australian Research Council (DP180100821). We thank also Prof. Anne Cutler for her support and discussions of the data.

7. REFERENCES

[1] Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

[2] Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human perception and performance*, *22*(1), 144.

[3] Jackendoff & Audring, 2020). Relational Morphology: A Cousin of Construction Grammar, *Frontiers in Psychology* 11, https://doi.org/10.3389/fpsyg.2020.02241

[4] Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357. https://doi.org/10.1037/0033-295X.115.2.357

[5] Heath, J. (1984). *Functional grammar of Nunggubuyu*. Canberra: Australian Institute of Aboriginal Studies.

[6] Thorne, D. A. (1993). *A Comprehensive Welsh Grammar*. Oxford: Blackwell.

[7] Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago, IL: University of Chicago Press.

[8] Prince, A., & Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar (Rutgers University Center for Cognitive Science Technical Report 2). *Rutgers Optmality Archive Version*, *8*, 2002.

[9] Bohn, A. & Bundgaard-Nielsen, R. L. (2021). Not "WEIRD" but truly different: Cultural Life Scripts and autobiographical memory in Indigenous Australia. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2020.09.006

[10] Sandefur, J. R. (1986). *Kriol of North Australia: A language coming of Age* (Vol. 10). Summer Institute of Linguistics, Australian Aborigines Branch.

[11] Schultze-Berndt, E., Meakins, F., & Angelo, D. (2013). Kriol. In (Eds. S. Michaelis, P. Maurer, M. Haspelmath & M. Huber), *The Survey of pidgin and creole languages*, vol. 1, pp. 241–251. Oxford: Oxford University Press.

[12] Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer [Computer program]. Retrieved 15 November 2018.

[13] Bundgaard-Nielsen, R. L., & Baker, B. J. (2020). Pause acceptability indicates word-internal structure in Wubuy. *Cognition*, *198*, 104167.

[14] Penke, M., Weyerts, H., Gross, M., Zander, E., Münte, T. F., & Clahsen, H. (1997). How the brain processes complex words: an event-related potential study of German verb inflections. *Cognitive Brain Research*, *6*(1), 37-52.

Self-interruptions in Breast Cancer Patients Who Complain of Anomia

Amélie B. RICHARD[1,2], Fabrice HIRSCH[1], Sophie JACQUIN-COURTOIS[2,3,] Karen T. REILLY[2]

(1) Paul-Valery Montpellier 3 University, CNRS UMR5267, Praxiling, 34090 Montpellier, France
(2) Université de Lyon, Université Lyon 1, INSERM U1028; CNRS UMR5292; Lyon Neuroscience Research Center, Trajectories Team, F-69676 Lyon, France
(3) Physical Medicine and Rehabilitation Department, Henry-Gabrielle Hospital, Hospices Civils de Lyon, 69610 Pierre-Bénite, France

ABSTRACT

Patients with cancer and cancer survivors sometimes complain of cognitive difficulties related to the cancer and its treatments. Even though one of their most frequent complaints is anomia, currently used neuropsychological tests are unable to detect word-finding difficulties.
The aim of this study was to explore the relationship between pauses at minor boundaries and their right contexts to better describe pauses associated with subtle anomia.
The spontaneous speech of seventeen patients with breast cancer was recorded. Pauses in suspensive interruptions (i.e., the phrase continues after the pause) and disfluent interruptions (i.e., the phrase is modified after the pause) were analysed as well as their right contexts (i.e., lexical-semantic content, independent phrases, or tool words).
The main result showed that suspensive pauses are more likely to be followed by lexical content. This finding suggests that the analysis of pauses in syntactic disruptions could provide evidence about subtle anomia.

Keywords: Breast cancer, pauses, anomia, disfluencies.

1. INTRODUCTION

Patients with cancer and cancer survivors sometimes complain of cognitive changes, called Cancer-Related Cognitive Impairment (CRCI). CRCI negatively affects patients' quality of life by limiting their social participation, and in some cases compromising their return to work. Patients complain of difficulties in the domains of memory (forgetting names and dates) [1], executive functioning (trouble multitasking) [2], attention (distractibility), and language [3]. One of the most frequent language complaints is anomia. According to self-report studies, difficulties with word-finding or word-retrieval in everyday speech could concern between one third and up to 100 percent of patients with CRCI [4], [5].

While the cognitive complaint can be reliably identified using validated questionnaires, neuropsychological tests struggle to detect many difficulties [6]. This discrepancy between subjective and objective assessments may be due to a lack of sensitivity and/or ecological validity of the latter [7]. Indeed, the tasks currently in use were originally designed for use with more severe disorders like aphasia, and do not correspond to the subtle nature of most cancer patients' difficulties [6]. Hence, there is a need for new diagnostic methods.
One solution, with high sensitivity and ecological validity, could be phonetics and discourse analysis. This approach has already been used in research on Mild Cognitive Impairment (MCI) [8], [9], and although MCI patients have a very different aetiology and evolution, their language complaint resembles that of patients with CRCI. These studies on MCI have shown differences between speech in patients and healthy controls at a prosodic level (e.g., a higher number of pauses) [10].
Everyday language contains speech errors called disfluencies. These include pauses, breaks, or repetitions, and one study with Alzheimer's Disease (AD) patients found an association between the number of disfluencies and word finding difficulties. These patients also produced more pauses at minor syntactic boundaries than healthy controls [11]. The authors interpreted these results as evidence of difficulties with lexical retrieval and planning. Based on these results, the current study investigated whether similar phenomena exist in breast cancer patients who complain of anomia. The goal of this research is to explore the relationship between pauses at minor syntactic boundaries and their right context with the aim of better describing pauses associated with subtle anomia. This exploratory study is part of a larger project aimed at developing a language-based diagnostic tool for CRCI.

2. METHOD

Twenty-three women participating in a physical reconditioning program for patients with breast cancer took part in the study. This rehabilitation program is offered by the Henry Gabrielle hospital

(Saint-Genis-Laval, France) to patients who are at least 6 weeks post-treatment (chemotherapy, radiotherapy, and surgery).
All patients were French native speakers or fluent in French. Patients with a history of neurological or psychiatric disorders, those taking medications that can affect consciousness, and/or having a diagnosed speech language impairment were not included. Patients were included if they reported language difficulties when questioned by a speech and language therapist. In total, seventeen patients were eligible for the study. Six patients were not included due to psychiatric conditions (2), vascular disease (1), language impairment history (2), or no language complaint (2). Since the study design was exploratory, we did not include a control group.
This research is part of the ActiCog study at the Hospices Civils de Lyon.

2.1. Speech samples

Patients were interviewed individually by a speech and language therapist. Each interview lasted approximately 30 minutes and consisted of a brief language examination (to verify the absence of any language impairment prior to the cancer diagnosis) followed by three tasks. In the first task, patients were asked to narrate a story based on five images. In the second task, they were asked to recall as many details as possible from a 5-paragraph text read by the speech and language therapist. Finally, they conversed with the investigator about the destination for their next holidays. The investigator intervened rarely, providing backchannels and questions when necessary, in order for the conversation to continue. Data presented in this study come from the spontaneous discourse related to their next holiday destination.
Interviews were recorded using a Rhode Lavalier Go™ omnidirectional microphone linked to a computer.
Patients completed a brief cognition function screening test (Montreal Cognitive Assessment, MoCA) and a speed of processing test (Symbol Digit Modalities Test, SDMT). They also completed auto-questionnaires about their perceived cognitive function (Functional Assessment of Cancer Therapy-Cognitive Function – Cognitive Function, FACT-Cog), their fatigue (FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue), and their anxiety/depression (Hospital Anxiety and Depression Scale, HADS).

2.2. Linguistic features

All interviews were manually transcribed and annotated using PRAAT [12]. Silent pauses over 200 msec, filled pauses such as “euh” (“uh”), and lengthening over 200 msec [13] with a flat F0 contour were all annotated.
Interruptions were labelled using the annotation system proposed by Pallaud and colleagues [14] based upon their analysis of auto-interrupted utterances in spontaneous conversations of French speakers. In their study, Pallaud and colleagues identified two kinds of self-interruptions: suspensive and disfluent. In suspensive interruptions, the utterance continues after the pause without any change in the phrase. In contrast, disfluent interruptions are characterised by a morpho-syntactic disruption whereby the interrupted utterance is modified or left unfinished.
The right context of the pauses was analysed to investigate the type of content following the pause. Three content types were defined: Lexical Phrases (LP) such as nouns, noun phrases, or verbs, Independent Phrases (IP) such as verb phrases, and Tool Phrases (TP) such as grammatical words or locutions.

The extracted variables were Pause type (Suspensive Pause and Disfluent Pause) and right context type (Lexical Phrase, Independent Phrase, and Tool Phrase) were extracted. For each variable, we measured the rate (i.e., number of occurrences / 100 words).

2.3. Statistical analyses

A D’Agostino & Pearson’s test was run to test whether the data follow a normal distribution.
Pearson’s correlations were conducted to investigate the relationship between average pause rate and questionnaire and test scores.
Pause rate data were entered into a two-way ANOVA with factors pause type (Suspensive vs Disfluent) and right context type (LP, IP, and TP) with Bonferroni post hoc tests. Finally, for each pause type and for each patient we plotted pause rate as a function of right context type.

3. RESULTS

3.1 Sample characteristic

The mean age of the 17 patients included in this study was 49.7 (± 9.5).
The neuropsychological assessments did not detect any cognitive problems, whereas the FACT-Cog questionnaire revealed a cognitive complaint (score of < 60 on the perceived cognitive impairment subscale [15]) in 13 /17 patients.
One patient did not complete the anxiety, depression, and fatigue questionnaires. Anxiety and depression scores were low for 13 of the remaining 16 patients.

There was no correlation between pause rate and scores on the HADS anxiety subscale (ρ = - .39, p = .12) and depression subscale (ρ = 0.03, p = .91). Patients reported a moderate level of fatigue but there was no correlation between pause rate and scores on the FACIT-F self-report fatigue questionnaire (ρ = - 11, p = .68).

To ensure that our data set did not include data from patients who made a small number of very long pauses, we calculated the correlation between the number of the pauses and average pause duration. The relationship between these two variables was highly correlated (ρ = 0.72, p <.0001), which attests to the quality of our dataset which did not include data from a small number of very long pauses.

3.2 Relationship between Pause Type and Right Contexts

Figure 1 shows the average pause rate separately for the two pause types and the three right contexts. For suspensive pauses, mean pause rate was highest for the LP right context and lowest for the TP right context. Whereas, for Disfluent pauses it was highest for the IP right context and lowest for the TP right context. The two-way ANOVA on these data showed a significant interaction (F(2) = 4.87, p = .009). Bonferroni post-hoc tests revealed a significant difference between LP and TP for both Suspensive (p < .001, 95% C.I. = [-2.24, -0.73]) and Disfluent Pauses (p < .05, 95% C.I. = [-1.43, 0.07]), a significant difference between IP and TP in both Suspensive (p < .01, 95% C.I. = [-1.6, -0.09]), and Disfluent Pauses (p < .001, 95% C.I. = [-1.99, -0.48]). LP and IP pause rates were also significantly different, but only for Suspensive Pauses (p < 0.05, 95% C.I. = []). There was a significant main effect of Right Context (p < .0001), but no effect of Pause Type was found (p =.52).

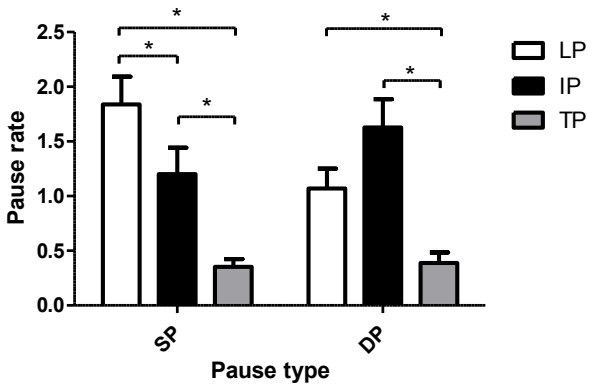


Figure 1: Comparison of mean pause rate for Suspensive and Disfluent Pauses as a function of right context: Lexical Phrase (LP), Independent Phrase (IP), and Tool Phrase (TP).

The mean pause rate data shown in Figure 1 reveal different data patterns for Suspensive and Disfluent pauses. To explore whether data from individual patients followed these same patterns, we plotted each patient's pause rate as a function of right context for Suspensive (Figure 2) and Disfluent (Figure 3) pauses. These figures show that nearly 75% of patients showed the same pattern as the mean pause rate data shown in Figure 1 (13/17 for suspensive pauses and 12/17 for disfluent pauses).

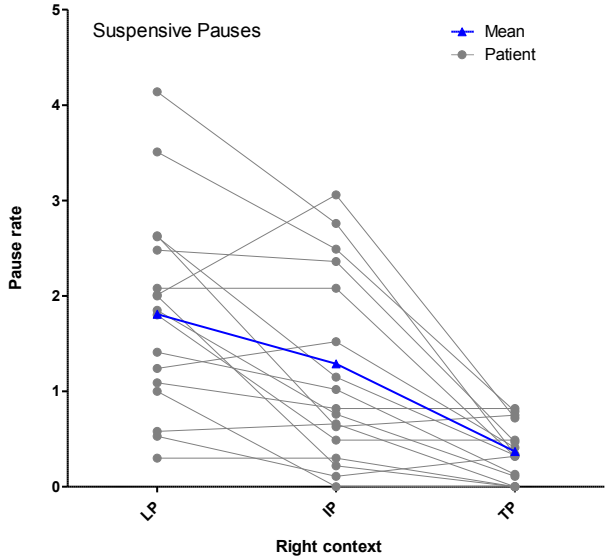


Figure 2: Right context pattern for each patient for Suspensive Pauses. Each line represents a patient, and the blue line represents the mean for all patients.

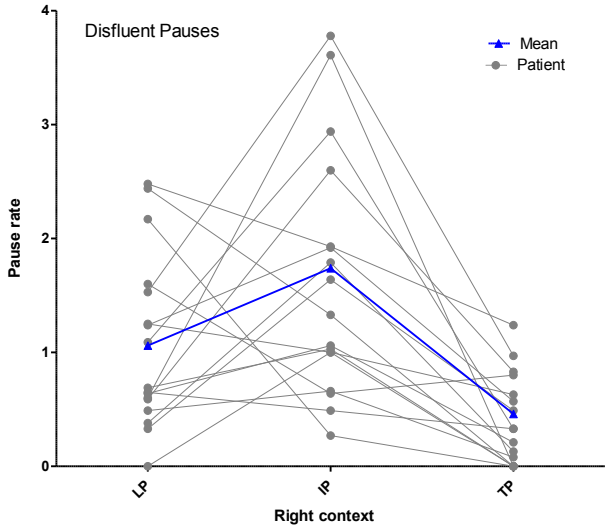


Figure 3: Right context pattern for each patient for Disfluent Pauses. Each line represents a patient, and the blue line represents the mean for all patients.

4. DISCUSSION

The spontaneous speech of seventeen women who had completed treatment for breast cancer and who complained of anomia was analysed.

4.1 Pause type

In the present corpus, pause rates showed a different pattern across right contexts for suspensive and disfluent pauses. Suspensive pauses occurred more often before lexical-semantic content than before independent phrases or tool words. In contrast, disfluent pauses occurred more often before independent phrases than before lexical-semantic content or tool words. This pattern of results suggests that suspensive pauses could be more likely to be associated with lexical retrieval processes.

Tool phrases are generally less frequent than other content type. The grammatical content analysed in the current study consisted of tool words such as determiners, or pronouns, as well as discourse markers (e.g., “voilà”, “quoi”, “en fait”). In the current study, tool words were often found before another break between two weak boundaries, suggesting that they might be part of lexical or syntactic difficulties. Discourse markers were less frequent than tool words, and often occurred before a strong boundary, suggesting that they might help the patient end her unfinished utterance.

While we deliberately separated our pauses into Suspensive and Disfluent, when the data from the two pause types are combined (the main effect of Right Context in our two-way ANOVA), lexical-semantic content and independent phrases follow pauses just as often as each other. While the interaction we observed means that these data should be interpreted with caution, they nevertheless suggest two possible hypotheses. First, anomia could be hidden in rephrasing, whereby patients who have trouble with word-finding use a new phrase to express their idea without the sought word. Second, the anomia reported by patients could arise from more global cognitive difficulties. For example, if anomia is caused by CRCI-related mechanisms then it could be a symptom of changes in cognitive functions including attention, memory, and executive functioning. Indeed, altered function in any one of these cognitive domains could have a direct impact on language. If this is the case, then patients who complain of anomia are likely to encounter trouble with phrasing or difficulties in syntactic planning.

4.2 Pause rate patterns

Nearly 75% of patients had pause rate patterns similar to the mean rate of the current sample. Research on the onset of CRCI is scarce. In contrast, we know that once cognitive difficulties have been identified they can last anywhere between two and ten years after the end of treatment [16]. Since these cognitive difficulties can compromise return to work [17] and limit social activities, there is an urgent need for early identification of patients who are potentially at risk of developing them. Comparing an individual’s pause rate pattern to patterns derived from a large population of cancer patients as well as a control population could be a reliable method for identifying those patients most at risk of developing long-lasting cognitive impairment.

4.3 Relationship between cofactors and pause rates

Consistent with the literature [6], scores on the cognitive complaint scale revealed more cognitive difficulties than the neuropsychological test scores. However, no correlation was found between pause rates and either cognitive complaint scores or neuropsychological test scores. Similarly, pause rates were not correlated with depression and anxiety or fatigue. Hence, pause rates could be independent of these psychological factors.

4.4 Strengths and limitations

This study is the first to explore pauses at minor syntactic boundaries with the goal of investigating phonetic markers of subtle anomia. Its major strength relies on the discourse type it analysed. Indeed, spontaneous speech provides the best ecological validity because of its proximity to everyday discourse.

Due to the exploratory nature of this study, no control group was included. Further investigations with healthy individuals and cancer patients with no complaint of anomia are necessary to achieve the long-term goal of developing a new method for the diagnosis of subtle anomia.

5. REFERENCES

[1] M. L. Kanaskie et S. J. Loeb, « The experience of cognitive change in women with breast cancer following chemotherapy », *J Cancer Surviv*, vol. 9, nº 3, p. 375-387, sept. 2015, doi: 10.1007/s11764-014-0387-x.

[2] F. M. Henderson, A. J. Cross, et A. R. Baraniak, « ‘A new normal with chemobrain’: Experiences of the impact of chemotherapy-related cognitive deficits in long-term breast cancer survivors », *Health Psychology Open*, vol. 6, nº 1, p. 205510291983223, janv. 2019, doi: 10.1177/2055102919832234.

[3] G. Bolton et A. Isaacs, « Women’s experiences of cancer-related cognitive impairment, its impact on daily life and care received for it following treatment

for breast cancer », *Psychology, Health & Medicine*, vol. 23, nº 10, p. 1261-1274, 2018, doi: https://doi.org/10.1080/13548506.2018.1500023.

[4] J. S. Myers, « Chemotherapy-Related Cognitive Impairment: The Breast Cancer Experience », *Oncology Nursing Forum*, vol. 39, nº 1, p. E31-E40, janv. 2012, doi: 10.1188/12.ONF.E31-E40.

[5] D. Von Ah, B. Habermann, J. S. Carpenter, et B. L. Schneider, « Impact of perceived cognitive impairment in breast cancer survivors », *European Journal of Oncology Nursing*, vol. 17, p. 236-241, 2013, doi: http://dx.doi.org/10.1016/j.ejon.2012.06.002.

[6] D. S. J. Costa et J. E. Fardell, « Why Are Objective and Perceived Cognitive Function Weakly Correlated in Patients With Cancer? », *Journal of Clinical Oncology*, vol. 37, nº 14, p. 1154-1158, 2019, doi: https://doi.org/10. 1200/JCO.18.02363.

[7] A. Hutchinson, J. R. Hosking, G. Kichenadasse, J. Mattiske, et C. Wilson, « Objective and subjective cognitive impairment following chemotherapy for cancer: A systematic review », *Cancer Treatment Reviews*, vol. 38, p. 926-934, 2012, doi: http://dx.doi.org/10.1016/j.ctrv.2012.05.002.

[8] L. Toth *et al.*, « A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech », *CAR*, vol. 15, nº 2, p. 130-138, doi: 10.2174/1567205014666171121114930.

[9] C. De Looze *et al.*, « Cognitive and Structural Correlates of Conversational Speech Timing in Mild Cognitive Impairment and Mild-to-Moderate Alzheimer's Disease: Relevance for Early Detection Approaches », *Front. Aging Neurosci.*, vol. 13, p. 637404, avr. 2021, doi: 10.3389/fnagi.2021.637404.

[10] K. W. Kim, S.-H. Na, Y.-C. Chung, et B.-S. Shin, « A Comparison of Speech Features between Mild Cognitive Impairment and Healthy Aging Groups », *Dement Neurocogn Disord*, vol. 20, nº 4, p. 52, 2021, doi: 10.12779/dnd.2021.20.4.52.

[11] F. Gayraud, H.-R. Lee, et M. Barkat-Defradas, « Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects », *Clinical Linguistics & Phonetics*, vol. 25, nº 3, p. 198-209, mars 2011, doi: 10.3109/02699206.2010.521612.

[12] P. Boersma et D. Weenink, « Praat: Doing Phonetics by Computer ». 2021. Consulté le: 26 octobre 2021. [En ligne]. Disponible sur: www.praat.org

[13] J.-P. Goldman, S. Roekhaut, et A.-C. Simon, « Étude statistique de la durée pausale dans différents styles de parole », in *Actes des 28èmes journées d'étude sur la parole (JEP)*, 2010, p. 161-164. [En ligne]. Disponible sur: http:// hdl.handle.net/2078.1/81909

[14] B. Pallaud, R. Bertrand, P. Blache, L. Prévot, et S. Rauzy, « Suspensive and disfluent self interruptions in French language interactions », in *Fluency and Disfluency across Languages and Language Varieties*, Corpora and Language in Use – Proceedings 4., L. Degand, G. Gilquin, L. Meurant, et A.-C. Simon, Éd. Louvain-la-Neuve: Presses universitaires de Louvain, 2019, p. 109-138.

[15] K. Van Dyk, C. M. Crespi, L. Petersen, et P. A. Ganz, « Identifying Cancer-Related Cognitive Impairment Using the FACT-Cog Perceived Cognitive Impairment », *JNCI Cancer Spectrum*, vol. 4, nº 1, p. pkz099, févr. 2020, doi: 10.1093/jncics/pkz099.

[16] M. B. de Ruiter *et al.*, « Cerebral hyporesponsiveness and cognitive impairment 10 years after chemotherapy for breast cancer », *Hum. Brain Mapp.*, vol. 32, nº 8, p. 1206-1219, août 2011, doi: 10.1002/hbm.21102.

[17] R. Bijker, S. F. A. Duijts, S. N. Smith, R. de Wildt-Liesveld, J. R. Anema, et B. J. Regeer, « Functional Impairments and Work-Related Outcomes in Breast Cancer Survivors: A Systematic Review », *J Occup Rehabil*, vol. 28, nº 3, p. 429-451, sept. 2018, doi: 10.1007/s10926-017-9736-8.

THE INFLUENCE OF FACE MASKS ON NATIVE AND NON-NATIVE MEMORY RECALL OF SPOKEN SENTENCES

Thanh Lan Truong & Andrea Weber

English Department, University of Tübingen, Germany
thanh-lan.truong@uni-tuebingen.de, andrea.weber@uni-tuebingen.de

ABSTRACT

The impact of face masks on listeners' recall of spoken sentences was examined. Native (= L1) and non-native (= L2) listeners of German watched video clips of a native talker producing German sentences with and without a face mask, followed by a cued-recall task. Results showed that face masks significantly reduced L1 listeners' recall performance. By contrast, no significant effect was found for L2 listeners. While the mask effect attests to the importance of visual speech cues for L1 listeners, the absence of a mask-effect for L2 listeners likely suggests that only listeners with a higher language proficiency in the second language can benefit from an audiovisual context for memory encoding.

Keywords: face mask, cued-recall, memory, audiovisual

1. INTRODUCTION

In times of the COVID-19 pandemic, face-to-face communication has often become more challenging, because face masks are frequently used and can degrade the acoustic speech signal such that higher frequencies are particularly affected, similar to a low-pass filter [1, 2]. In addition, face masks cover the mouth region of a talker's face, thus restricting access to visual articulatory cues that can be helpful in language comprehension, particularly when speech is degraded [3, 4].

Visual cues, such as lip and jaw movements, can facilitate comprehension (e.g., [5, 6]) as they convey important phonological information about speech sounds [3, 4]. For example, while closed lips indicate a bilabial place of articulation (e.g., /p/ and /b/), an open jaw indicates vowel height (e.g., more open jaw for the vowel /a/ and less open jaw for /i/). Thus, visual cues can supplement and complement information about speech sounds that is not present in the auditory signal itself [6, 7].

Recent findings suggest that the difficulties listeners encounter when listening to speech produced with a face mask are likely to stem from both the acoustic degradation of the speech signal and the lack of visual cues of the talker's mouth movements. For example, face masks increase listening effort and reduce L1 adults' correct identification of words and sentences in noisy conditions [1, 8, 9]. In addition, they affect L1 adults' memory recall in quiet conditions such that L1 listeners remember fewer words when talkers are wearing a mask compared to when they are not wearing one [10]. This effect has also been replicated for L2 speech but only in noisy listening conditions [11].

All of these findings are in line with the Framework for Understanding Effortful Listening [12], the Effortfulness Hypothesis [13], and the Ease of Language Understanding [14], postulating that more cognitive resources are utilized for speech comprehension when listening conditions are adverse (i.e., speech produced with a face mask), leaving fewer resources available for memory encoding [15], which is fueled by limited cognitive resources [16]. While findings for L1 listeners imply that face masks negatively affect memory encoding, the question remains whether face masks also affect L2 listeners' memory for spoken language. Previous studies on L2 listening found that the presence of visual articulatory cues can enhance L2 participants' perception of French [17], Korean [18], Irish and Spanish sounds [19]. Therefore, visual cues can facilitate L2 listeners' perception [20], and L2 listeners might even pay more attention to the information that is conveyed by visual articulatory movements to make up for their poorer comprehension skills in the L2 language [21]. We, therefore, further investigated if sufficient cognitive resources are left for visual cues to affect memory encoding in L2.

The goal of the present study was to compare L1 and L2 participants' memory performance of German sentences produced by an adult talker when the talker was wearing a face mask or was wearing no mask using a cross-modal cued-recall task.

2. METHOD

2.1. Participants

For the experiment, we recruited two groups of participants: Forty L1 listeners of German via social media and university email[1] between 19 and 37 years of age (mean = 23.5; SD = 4.0; 30 females) and forty L2 listeners of German, with English as their native language, between 18 and 58 years of age (mean = 31.7; SD = 12.7; 21 females) via Prolific. Four participants had to be excluded from further analyses because they did not follow the instructions or did not complete the experiment. For the L2 participants, we used the pre-screeing function of Prolific, which ensured that only participants who registered themselves with an intermediate or advanced level of German were allowed to participate. Participants' average current daily use of German ranged from 0-90%; 17 reported having lived in Germany before (min: six months, max: seven years). Additionally, the L2 group's German proficiency was measured via self-report on a scale ranging from 1 (very poor) to 7 (very good) for all four modalities (i.e., writing, listening, speaking, and reading). The average proficiency score was 4.85 (SD = 1.45). Particularly, average score for listening was 4.65 (SD = 1.48).

2.2. Material

The stimuli consisted of 48 meaningful German sentences that were low in predictability and had been modelled after the Oldenburger Satztest [22]. Since sentence context did not semantically restrict lexical options, the possibility of context rectifying comprehension failures was fairly low and ensured a more careful processing of the input [23]. The syntactic structure was the same for all 48 sentences, and each content word occurred only once in the complete set of sentences. The sentences all began with a determiner and a noun, followed by a verb, an adverb, an adjective, and a noun (e.g., *Die Köchin hilft montags armen Kindern*, "the cook helps on Mondays poor children").

A 22-year-old female native talker of German was recorded on video in a sound-attenuated room, and produced all sentences with and without a face mask. The face mask consisted of two fabric layers: The inner layer was a thin fleece, and the outer layer was cotton. When necessary, the talker repeated a sentence until it was produced without any errors or hesitations. The talker was instructed to produce all sentences at a normal speaking rate and to avoid enunciating more loudly or clearly when wearing the mask. The videos were recorded by using a Sony (Tokyo, Japan) DSC-Hx90 camera with video resolution parameters set to FULL HD 1920x1080. Audio was recorded at a sampling rate of 48 kHz with a high-quality microphone placed in front of the talker. The average F0 value of the talker was 235.5 Hz. Durations for sentences produced without a mask were on average 3255 ms, and with a mask they were 3178 ms (t = 1.35 p = 0.18). Spectral analysis [root mean square (rms) power] of the talker revealed no difference between sentences with (56.6 dB) and without a face mask (56.7 dB) (t = 0.28, p = 0.77).

2.3. Procedure

The experiment was conducted online.[2] Prior to the experiment, participants electronically gave informed consent.[3] The experiment started with two practice trials followed by eight blocks with six sentences each. The self-paced cued-recall task followed immediately after each block. For this task, sentences of the preceding block were presented up to the adverb orthographically on the screen (e.g., *Die Köchin hilft montags*, "the cook helps on Mondays"), and participants were asked to enter the missing two final words (e.g., *armen Kindern*, "poor children") on their keyboard. The recall cues for all six sentences of a block were presented at the same time in the order of block presentation, and participants could fill in their responses in any order. At the end of the experiment, participants filled out a brief language background questionnaire and were asked about technical problems of which none were reported.

3. RESULTS

Each correctly recalled word received the score correct (1)[4] and each erroneously recalled or missing word received the score incorrect (0) (see Fig. 1). There were two keywords for each of the 48 sentences, making a total of 96 keywords to be recalled per participant. L1 participants recalled 57.3% of the words correctly and 42.7% incorrectly. In contrast, L2 participants recall performance was extremely low such that only 16.9% of the words were recalled correctly and 81.3% incorrectly. Subsequent descriptive analyses of the incorrectly recalled words indicated that the majority of incorrectly recalled words had been full omissions of a keyword (68% for L1 listeners; 50% for L2 listeners). The remaining incorrect responses consisted of a variety of error types. Some responses were unrelated in form and in meaning

to the intended words (e.g., *schwarze Schuhe*, "black shoes," for *staubige Kissen*, "dusty pillows"), fewer responses were closely semantically related (e.g., *Dackel*, "Dachshund," for *Hunde*, "dogs") and an even smaller number of responses consisted of potential typos (e.g., the nonword Lmpen for *Lampen*, "lamps").

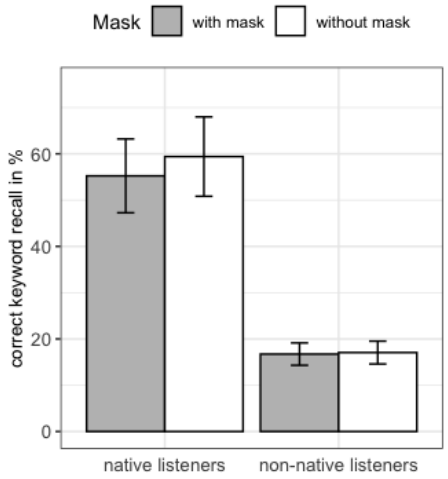


Figure 1: Average percentage of keywords recalled correctly for sentences with and without a face mask. The vertical bar represents standard errors.

A logistic mixed-effects regression model with the lme4 package in R (version: version 4.0.5) [24] was used to analyze the effect of face masks on participants' correctly recalled keywords [25]. Keyword recall (success vs. failure) was the dichotomous dependent variable, and native language (L1 vs. L2), face mask (mask vs. no mask), and block (8 blocks) were the independent variables; face mask x native language x block was included as an interaction term. To test linear and quadratic effects of block, orthogonal polynomials were used [26]. Additionally, to account for extra variation, fixed factors of sentence duration, rms power, and language proficiency were also included in the model. The German proficiency factor consisted of the sum of self-reported ratings (i.e., from 1 the lowest to 7 the highest) in the category of writing, listening, speaking, and reading. We found a main effect of native language (b = 4.49, SE = 0.75, p < .001) with higher recall rates for the L1 group than the L2 group, an interaction between linear effects of block and native language (b = 1.0, SE = 0.42, p = .02), suggesting better recall performance as the experiment progressed for participants who match the first language with the talker (i.e., L1 German as opposed to L2 German) and an interaction between face mask and native language (b = -0.51, SE = 0.17, p = .003), suggesting different recall performance patterns based on participants' native language.

We then grouped the data based on the participants' native language (i.e., L1 vs. L2). The separate analysis for L1 participants showed a significant effect of face mask, with L1 listeners recalling fewer words when the talker was wearing a mask compared to when the talker was not wearing a mask (b = -0.28, SE = 0.11, p = .01). Recall was furthermore better for shorter sentence durations than for longer ones as the main effect for sentence duration showed (b = -0.47, SE = 0.23, p = .04).

Also, a main effect of rms (b = -83.6, SE = 36.6, p = .02) was found, indicating that sentence recordings with less rms power were recalled better than sentences with higher rms power. By contrast, L2 listeners showed no effect for face mask (b = 0.09, SE = 0.14, p = .52), rms (b = 50.2, SE = 57.1, p = .38), sentence duration (b = -0.19, SE = 0.42, p = .65) and language proficiency (b = 0.11, SE = 0.13, p = .37).

The results suggest that, for L1 listeners only, processing was easier when visual cues were available than when they were not, and this availability left more cognitive resources for successful memory encoding. In short, the findings for L2 listeners suggest that the covering of visual articulatory cues with a face mask did not negatively influence recall performance.

4. CONCLUSIONS

In a cued-recall experiment, we investigated the impact of face masks on L1 and L2 memory for spoken language. We found an effect of mask for L1 listeners but not for L2 listeners. That is, L1 listeners significantly recalled fewer words correctly when the talker produced the sentences in quiet with a mask compared to when the talker was not wearing one, suggesting that processing speech with face masks leaves fewer cognitive resources available for memory encoding [14, 13, 12], as face masks both cover visual speech cues and degrade the acoustic signal [1, 2]. At first glance, our findings for L1 listeners contradict Smiljanic et al. [11] who only found a negative effect of face masks for L1 listeners when additional noise was added to the speech signal. One possible reason for the difference in findings could be attributed to the material of our study. To avoid a facilitatory influence of context on comprehension and recall, we used dissociated sentences with low predictability whereas Smiljanic

et al. [11] used a coherent text, which made the listening environment more naturalistic but also possibly allowed correctly guessing of individual words, especially when comprehension was more difficult, for example due to a face mask. Since cohesive information is easier to encode than dissociated information (e.g., [27]), this alleviation through cohesion might have been responsible for the lack of a mask effect when listening to the stimuli in quiet in Smiljanic et al. [11]. The present study was furthermore the first to test recall memory for L2 listeners, and while the results cannot be directly compared to previous studies, a number of reasons could be responsible for the lack of a mask effect for L2 listeners. First of all, it could be the case that L2 listeners' encoding was indeed unaffected by the mask because these listeners did not use visual speech information for perception in the first place. While previous studies mostly confirmed that L2 listeners can use visual cues [17, 19, 18] for perception, Wang et al. [28] found that the ability for audiovisual speech perception in L2 varies significantly across native languages of the participants. Thus, participants with different native language backgrounds might still show an effect of face mask on recall. Alternatively, the lack of contextual cues could have made the task of comprehension of the speech signal so challenging that no attention was paid to the visual speech cues by L2 participants. In that case, an effect of face masks should emerge for L2 participants when cohesive texts are used as, for example, in Smiljanic et al.. However, the most likely explanation has to do with the generally low performance of our L2 participants. While the L1 listeners performed with an accuracy of 57.3% on average, the L2 listeners only reached an average of 16.9%. Although our L2 participants reported intermediate to advanced proficiency in German and our sentences comprised of common German words with a high lexical frequency, the language skills of our L2 participants might have been too low for an effect of masks to be observed. In other words, the L2 listeners' performance could be a floor effect. Higher proficiency in L2 might thus be necessary to optimally make use of the enhancement provided by visual speech cues. Indeed, advanced linguistic expertise can improve working memory performance, leading to greater automatic processing and thus to smaller processing cost in comprehension of the stimuli. This in turn leaves a larger amount of cognitive resources available that can be employed for memory encoding. As such, L2 comprehension and working memory capacity are mediated by language proficiency. Given that working memory operates on limited cognitive resources [16], we propose that the following happened for the L2 listeners: The low language competence in German in combination with the difficulty of the task required extra processing demands for comprehension, leaving no resources for retaining information in the working memory [29, 30].

In conclusion, we present evidence for a negative impact of face masks on memory recall for L1 listeners. We interpret the absence of a mask effect for L2 listeners as being modulated by the task's plausibility and the language proficiency of L2 listeners. Although the results showed no impact of face mask on memory encoding for L2 listeners, the results of the present study reinforce previous findings that state that L2 language proficiency plays a crucial role in extracting visual cues from the lips of the talker [28, 31].

5. REFERENCES

[1] P. Bottalico, S. Murgia, G. E. Puglisi, A. Astolfi, and K. I. Kirk, "Effects of masks on speech intelligibility in auralized classrooms," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 2878–2884, 2020.

[2] R. M. Corey, U. Jones, and A. C. Singer, "Acoustic effects of medical, cloth, and transparent face masks on speech signals," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, p. 2371–2375, 2020.

[3] R. Campbell, "The processing of audio-visual speech: empirical and neural bases," *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 363, no. 1493, pp. 1001–1010, 2008.

[4] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[5] J. Navarra and S. Soto-Faraco, "Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds," *Psychological Research*, vol. 71, no. 2, pp. 4–12, 2007.

[6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, p. 212, 1954.

[7] D. W. Massaro, *Categorical perception: The groundwork of cognition*. Cambridge University Press, 1987, ch. Categorical partition: A fuzzy-logical model of categorization behavior., pp. 254–283.

[8] V. A. Brown, K. J. Van Engen, and J. E. Peelle, "Face mask type affects audiovisual speech intelligibility and subjective listening effort in

young and older adults," *Cognitive Research: Principles and implications*, vol. 6, no. 1, p. 49, 2021.
[9] M. Randazzo, L. L. Koenig, and R. Priefer, "The effect of face masks on the intelligibility of unpredictable sentences," *Proceedings of Meetings on Acoustics*, vol. 42, no. 1, p. 032001, 2020.
[10] T. L. Truong and A. Weber, "Intelligibility and recall of sentences spoken by adult and child talkers wearing face masks," *The Journal of the Acoustical Society of America*, vol. 150, no. 3, pp. 1674–1681, 2021.
[11] R. Smiljanic, S. Keerstock, K. Meerman, and S. M. Ransom, "Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech," *The Journal of Acoustical Society of America*, vol. 149, no. 6, p. 4013, 2021.
[12] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Y. Hornsby, L. E. Humes, U. Lemke, T. Lunner, M. Matthen, C. L. Mackersie, G. Naylor, N. A. Phillips, M. Richter, M. Rudner, M. S. Sommers, K. L. Tremblay, and A. Wingfield, "Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel)," *Ear and Hearing*, vol. 37, pp. 5S–27S, 2016.
[13] S. L. McCoy, P. A. Tun, L. Clarke Cox, M. Colangelo, and J. A. Stewart, "Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech," *Quarterly Journal of Expoerimental Psychology*, vol. 58, no. 1, pp. 22–33, 2005.
[14] J. Rönnberg, T. Lunner, A. Zekveld, P. Sörqvist, H. Danielsson, B. Lyxell, Dahlström, C. Signoret, S. Stenfelt, M. K. Pichora-Fuller, and M. Rudner, "The ease of language understanding (elu) model: Theoretical, empirical, and clinical advances," *Frontiers in Systems Neuroscience*, vol. 7, p. 31, 2013.
[15] J. E. Peelle, "Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior," *Ear and Hearing*, vol. 39, no. 2, pp. 204–214, 2018.
[16] M. A. Just and P. A. Carpenter, "A capacity theory of comprehension: Individual differences in working memory," *Psychological Review*, vol. 99, no. 1, pp. 122–149, 1992.
[17] D. Reisberg, J. McLean, and A. Goldfield, *Hearing by Eye: The Psychology of Lip-Reading*. Erlbaum, Hillsdale, NJ, 1987, ch. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli, pp. 97–113.
[18] C. Davis and J. Kim, "Audio–visual interactions with intact clearly audible speech," *The Quarterly Journal of Experimental Psychology Section A*, vol. 57, no. 6, pp. 1103–1121, 2004.
[19] V. D. Erdener and D. Burnham, "The role of audiovisual speech and orthographic information in nonnative speech production," *Language Learning*, vol. 55, pp. 191–228, 2005.
[20] D. W. Massaro, *Perceivig talkig faces: From speech perceptio to a behavioral principle*. The MIT Press, 1998.
[21] L. Drijvers and A. Özyürek, "Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension," *Language and Speech*, vol. 63, no. 2, pp. 209–220, 2020.
[22] *Oldenburger Satztest: Handbuch und Hintergrundwissen*. Oldenburg, Germany: Hörtech GmbH, 2000.
[23] J. Rommers and K. D. Federmeier, "Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects," *Cortex*, vol. 101, pp. 16–30, 2018.
[24] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2021, versions 4.0.5. [Online]. Available: https://www.r-project.org
[25] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," 2015.
[26] D. Mirman, *Growth curve analysis and visualization using R*. London: Taylor Francis, 2017.
[27] J. B. Black and H. Bern, "Causal coherence and memory for events in narratives," *Journal of Verbal Learning and Verbal Behavior*, vol. 20, no. 3, pp. 267–275, 1981.
[28] Y. Wang, D. M. Behne, and H. Jiang, "Influence of native language phonetic system on audio-visual speech perception," *Journal of Phonetics*, vol. 37, no. 3, pp. 344–356, 2009.
[29] E. Service, M. Simola, O. Metsänheimo, and S. Maury, "Bilingual working memory span is affected by language skill," *European Journal of Cognitive Psychology*, vol. 14, no. 3, pp. 383–408, 2010.
[30] M. Noort, P. Bosch, and K. Hugdahl, "Foreign language proficiency and working memory capacity," *European Psychologist*, vol. 11, pp. 289–296, 2006.
[31] Z. Xie, H. Yi, and B. Chandrasekaran, "Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener," *PloS one*, vol. 9, p. e114439, 2014.
[32] A. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *bioRxiv*, 2019.

[1] The L1 memory data of these participants have been previously reported in [10].

[2] L1 participants were tested on *Alchemer* and L2 listeners on *Gorilla Experiment Builder* (www.gorilla.sc) [32].

[3] The headphone screening task was a special feature in Gorilla Experiment Builder which was implemented prior to our actual experiment. This task required participants to do the experiment with headphones. Those who passed the test were immediately directed to

the experiment.
[4] Umlauts with two letters were accepted as correct (e.g., Veogel, “birds,” for “Vögel/ Voegel”).

THE USE OF SYNTACTICALLY REDUNDANT PHONETIC CUES IN SPEECH PERCEPTION

Clara Cohen

Glasgow University Laboratory of Phonetics, University of Glasgow
clara.cohen@glasgow.ac.uk

ABSTRACT

Subsegmental phonetic detail can be highly informative about upcoming word structure. For example, shortened stems tend to signal that suffixes or additional syllables will follow. Yet preceding morphosyntactic context can reduce the informational value of such cues. This study asks whether listeners still attend to phonetic detail in such contexts.

In a visual world eye-tracking experiment, 21 listeners distinguished singular target nouns named in audio recordings of English sentences from plural or two-syllable competitors. In the sentences, noun stems were lengthened or shortened, and positioned either after an agreeing determiner (*this/that*), rendering durational cues to plural suffixes redundant, or after the neutral determiner *the*, rendering them informative. GAMM analysis of gaze traces revealed that lengthening the stem facilitated perception, but only after agreeing determiners. This suggests not only automatic processing of redundant phonetic detail, but a requirement that such detail can be predicted from context before it can be processed.

Keywords: speech perception; eye-tracking; pronunciation variation; morphophonetics

1. INTRODUCTION

The duration of an initial syllable can carry substantial information about the structure of the rest of the word. In a pattern known as polysyllabic shortening, the first syllable in words like *captain* or *hamster* are shorter than they would be if they formed the stand-alone words *cap* and *ham* [1, 2]. Since morphological affixation often adds additional syllables, polysyllabic shortening can signal the presence of upcoming morphological structure, such as derivational suffixes.

Listeners are capable of drawing on these durational patterns during perception. They can use them to distinguish single-syllable words from multi-syllabic competitors [1–3], and simple words like *clue* from complex words like *clueless* [4–6]. Even when the addition of a suffix does not increase the syllable count of a word, subsegmental cues can still provide information about morphological structure [7–9]. The first goal of this study, therefore, focuses on English nouns, whose stems are shorter when a plural *-s* suffix is added [10], even when it does not change the syllable count. Can listeners use this durational cue to aid them in distinguishing singular nouns from plurals?

The second goal of this study concerns the role of morphosyntactic context in modulating the informational value of stem duration cues. Consider an utterance *Don't trip over the dogs!* This sentence is segmentally ambiguous regarding the number of canine obstacles until the final suffix is produced, and so keen attention to durational detail might afford listeners a head start on identifying challenges to safe locomotion. By contrast, if the warning took the form *Don't trip over those dogs!*, the listener would know even before the onset of the noun that multiple tripping hazards are in play. In this context, then, the stem-durational cue, which signals an upcoming plural suffix, is rendered redundant by the number information in the agreeing determiner. Since processing fast-moving, low-level phonetic detail is cognitively costly [11, 12], listeners may cease to attend to such stem duration cues in contexts where they offer no new information beyond what is already known from preceding morphosyntax.

An alternative perspective springs from research on incremental prediction. Listeners process real-time speech incrementally, and preactivate representations that are highly predictable in context [13–15]. When those predictions are violated, comprehension is impeded while the listener reanalyses the speech stream. Neurophysiological evidence suggests that this predictive preactivation may include not only semantic and phonological features, but also phonetic expectations [8, 16–18]. Such predictions may well include knowledge about durational variation patterns in noun stems. For example, if preceding morphosyntactic context leads listeners to expect a singular noun, then part

of the expected phonetic form of a singular noun would include a lengthened stem. If so, then lengthening a singular noun stem would aid listener perception *more* in contexts where such lengthening is predictable, and hence informationally redundant, because it better matches the listener's expectations.

The current study tests these predictions with a visual world eye-tracking experiment. Listeners were presented with a four-quadrant visual display, containing a one-syllable singular target noun, a competitor noun (either plural or two syllables), and two distractors. They were asked to select the picture mentioned in a recorded sentence (e.g., *My view was blocked by* ***the cart*** *in front of the door*). In each sentence, the target singular noun (*cart*) was either lengthened, to match the expected duration of a singular one-syllable noun, or shortened; and positioned either after an agreeing determiner (*that cart*) or a non-agreeing determiner (*the cart*).

Lengthening the singular noun stem should help listeners distinguish one-syllable nouns from two-syllable competitors in all contexts, replicating previous findings on listeners' use of polysyllabic shortening cues during perception [1–6]. However, for plural competitors the perceptual benefits of lengthening may depend on context. If listeners disregard redundant phonetic detail, then lengthening the noun stem should help more in contexts when such detail is informative—i.e., after non-agreeing *the*. However, if phonetic detail forms part of the pre-activated predictions that assist incremental perception, then lengthening the noun stem should still facilitate perception even after agreeing determiners like *this/that*—perhaps even more strongly than after *the*, which does not preactivate phonetic expectations about singular noun stem duration.

2. METHODS

2.1. Materials

Sentences were built around a set of 84 one-syllable noun stems, which could appear in one of three forms: a singular (e.g., *cart*); a plural (*carts*); or a two-syllable carrier (*carton*), in which the singular was always lexically embedded as the first syllable.

All of these target words were presented in sentences, such as *My view was blocked by that cart in front of the door*. Each sentence contained a preamble (*My view was blocked by...*), determiner (*the/that/those*), target word (*cart/carts/carton*), and postamble (*in front of the door*). Sentences with singular targets formed the critical items, while plural and carrier targets served as fillers.

Each experimental list was formed by rotating a given item across eight experimental conditions formed by fully crossing three binary factors: Context, Competitor Type, and Duration Match. Context (agreeing/non-agreeing) encodes the determiner preceding the target noun. Agreeing determiners explicitly signaled the number of the following noun, taking the form *this* or *that*. Non-agreeing determiners were always *the*. Competitor Type (plural/carrier) encodes the competitor image in the visual display. A Plural competitor depicted two examples of the target noun (e.g., two *carts*), while the Carrier competitor a single example of the carrier word (e.g., one *carton*). Each list contained 40 critical and 40 filler sentences.

All sentences were produced by a native speaker of Scottish English, and recorded in six variants. These variants were created by fully crossing the three possible target noun forms (singular/plural/carrier) with the two determiner forms (agreeing *this/that* or *these/those* and non-agreeing *the*). The Duration Match (match/mismatch) manipulation was then generated by manipulating and splicing these raw recordings. First, the raw duration of the singular noun and its intended competitor type (plural or carrier) were averaged together. Then, 40ms were added to that average to calculate the target duration for Match stimuli, and 40ms were subtracted to calculate the target duration for Mismatch stimuli. Each singular noun stem was then adjusted by Praat script [19] to produce Match and Mismatch versions, which thus differed in duration by 80ms—a difference determined from pilot work to be perceptible without veering too far from actual patterns of polysyllabic shortening. Finally, the determiner, manipulated noun stem, and postamble were spliced onto a preamble from a third recording of the sentence, so as to remove any early cues as to the original nature of the target.

Filler sentences, containing plural and carrier targets, were manipulated in the same way, except that all filler targets contained only the shortened version of the nouns stem, to support the listener's expectation that stems with suffixes or second syllables would be shortened.

This procedure generated 8 versions of each sentence, which were rotated across experimental lists in a Latin Square, ensuring a fully crossed $2\times 2 \times 2$ within-item and within-subjects design.

2.2. Participants and Procedure

Twenty-one native speakers of Scottish English (mean age 19.29; 16 female, 5 male) were recruited

Table 1: GAMM summaries for Target Advantage in sentences with plural (left) and carrier (right) competitors. Levels for Duration Match (abbreviated *DM*) are Match (reference, *m*) and Mismatch (*mm*); levels for Context (*Con*) are Agreeing (reference, *ag*) and Non-Agreeing (*na*). Difference smooths for carrier competitor GAMM were calculated on a four-level interaction between Context and Duration Match.

Plural competitors

Parametric	Est.	SE	*t*	*p*
Intercept	3.20	0.284	11.26	$< .001$
DM=mm	-1.14	0.17	-6.66	$< .001$
Con=na	-0.60	0.17	-3.49	$< .001$
DM=mm:Con=na	1.04	0.24	4.34	$< .001$

Difference smooths	edf	*F*	*p*
Window	4.64	20.76	$< .001$
Win:DM=mm	1.77	2.30	.14
Win:Con=na	2.55	3.18	$< .001$
Window, by subj	95.57	3.55	$< .001$

Carrier competitors

Parametric	Est.	SE	*t*	*p*
Intercept	1.53	0.23	6.78	$< .001$
DM=mm	-2.27	0.13	17.95	$< .001$
Con=na	0.82	0.13	6.45	$< .001$

Difference smooths	edf	*F*	*p*
Window	1.05	31.46	$< .001$
Win:DM=mm,Con=ag	4.41	13.46	$< .001$
Win:DM=m,Con=na	1.00	2.66	.10
Win:DM=mm,Con=na	3.38	5.44	$< .001$
Window, by subj	87.09	2.29	$< .001$

from the local university community. Participants started with a short demographic questionnaire, before continuing on to the eye-tracking procedure. The procedure started with a nine-point calibration and seven practice trials. After pausing to check participant comfort and answer questions, the tracker was recalibrated and the full experiment began. Each trial began with 500ms fixation cross, and then a 3500ms preview of the images, each of which was accompanied by a label. After 3500ms, the audio recording played, and participants selected the image named in the sentence. Each experiment contained 80 trials, presented in blocks of 20 with calibration break between each block.

Data were collected from the right eye by an Eyelink 1000+ at a 1000Hz sampling rate.

2.3. Analysis

Gaze data was analyzed from the onset of the determiner until 900ms after the offset of the target noun, when visual inspection of the raw data indicated that overall looks to target had peaked. Samples were binned into windows from the onset of the determiner. The proportion of looks to target and competitor within each bin was calculated by dividing the number of samples in which the gaze fixated on the target or competitor by the number of samples within the bin. Each determiner was divided into two bins, and each noun stem was divided into five bins. This ensured that the determiner offset was normalised to occur at the end of bin 2, and the noun offset occured at the end of bin 7. Post noun-offset, all bins were 50ms.

Proportions were converted to empirical logits [20], and eLog Target Advantage was calculated by subtracting the eLog looks to competitor from eLog looks to target in each window bin. Elog Target Advantage was then analysed with generalised additive mixed models from the `mgcv` package (version 1.8-40 [21]) in R (version 4.2.1 [22]). Duration Match and Context were included as both parametric terms and difference smooths, with random smooths included by subject. Terms were added sequentially to the null model, first as parametric terms and then as difference smooths, and only retained if their addition returned a significant difference in a Chi-squared test on the maximum likelihood scores when compared against the simpler model, as implemented in the `itsadug` package (version 2.4.1) [23]. Stimuli with carrier competitors were analysed separately from stimuli with plural competitors.

3. RESULTS

Table 1 provides final model summaries, and Figure 1 shows model predictions overlaid on observed Target Advantage across all time bins.

The plural competitor model revealed an interaction between Match and Context in the parametric terms (Table 1). In the parametric effects, Mismatch sentences showed overall lower Target Advantage than Match sentences when preceded by Agreeing determiners ($\beta = -1.13, p < .001$). Match sentences showed lower overall Target Advantage with Non-Agreeing determiners, relative to Agreeing determiners ($\beta = -0.60, p < .001$). The interaction between Match and Context, however, revealed that the disadvantage associated with Mismatch sentences was almost entirely undone with Non-Agreeing determiners ($\beta = 1.04, p < .001$). This can be observed in the left-hand panel of Figure 1. Although Agreeing contexts (top) show a decided Match effect, with higher Target Advantage for the

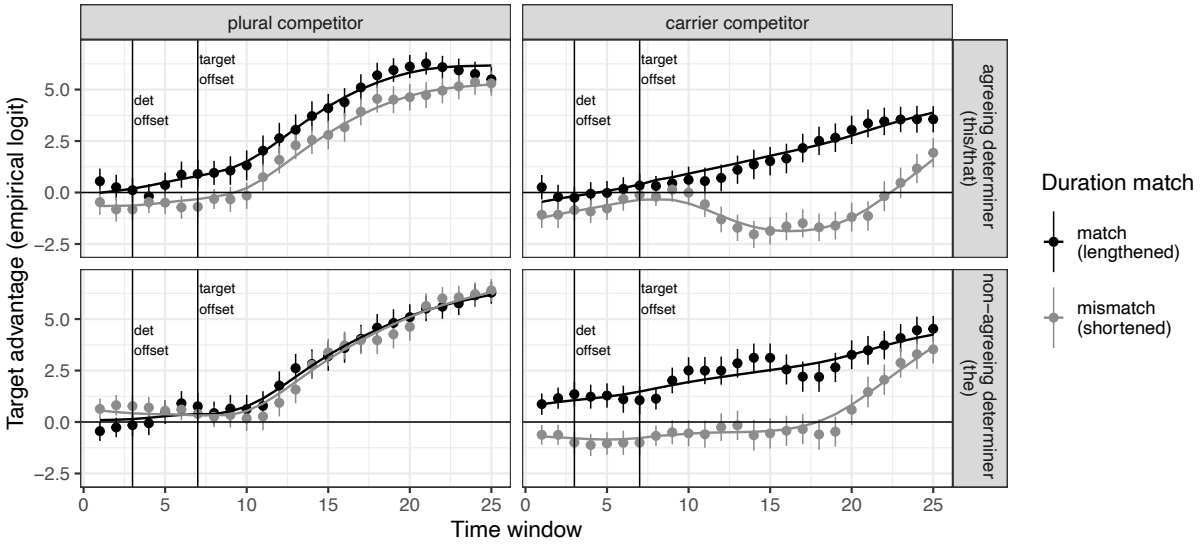


Figure 1: Target advantage over time for sentences with plural competitors (left) and carrier competitors (right), with targets placed after Agreeing determiners (top row) and Non-Agreeing determiners (bottom row).

black Match gaze trace than the grey Mismatch gaze trace, this effect disappears in Non-Agreeing contexts (bottom).

The carrier competitor model revealed a main parametric effect of Match and Context, but no interaction between them. The effect of Context reflects an overall advantage for sentences with Agreeing determiners ($\beta = 0.82, p < .001$), and an overall disadvantage for sentences with Mismatching—i.e., shortened—stem duration ($\beta = -2.27, p < .001$). The difference smooths revealed an interaction between those variables: Although there was no difference in curve shape between the gaze traces for Agreeing and Non-Agreeing sentences when they appeared in the Match condition, the gaze traces differed substantially in the Mismatch condition. As shown in Figure 1, Target Advantage for Mismatch sentences drops after the target offset in Agreeing sentences before rising again at the end of the interest period (top right), while Non-Agreeing sentences (bottom right) show a flatter trajectory and before the late rise.

4. DISCUSSION AND CONCLUSION

The results of this experiment show that listeners can use subsegmental detail to aid in processing both syllable structure and morphological structure. When distinguishing one-syllable nouns from two-syllable competitors, listeners showed a marked advantage when the target one-syllable nouns were lengthened according to patterns of polysyllabic shortening, replicating studies [2, 3]. This Duration Match effect did not interact with context, which is also expected. Since targets and carrier competitors were both singular, the presence of an agreeing determiner would not affect listeners' predictions or expectations about the phonetic or morphosyntactic properties of the upcoming noun.

When distinguishing one-syllable singular nouns from one-syllable plurals, however, the situation was different. Listeners were better at identifying the singular target when the stem was lengthened, but this pattern only obtained after agreeing determiners. After non-agreeing determiners—exactly the context where such detail might be expected to carry the most valuable information about the target noun's form—listeners seemed insensitive to the duration of the target noun stem.

This interaction is consistent with models of online speech perception which hold that listeners form phonetically detailed predictions about upcoming material [8, 16–18]. When those predictions are confirmed by the incoming speech stream, listeners are faster to identify the target word. After non-agreeing determiners, by contrast, listeners could not form predictions about the number of the target noun, and so could not effectively use the duration information in the stem.

In sum, these findings suggest that listeners readily process durational cues that signal upcoming lexical word structure. Cues which signal upcoming morphosyntactic information, however, are only processed when they are syntactically redundant, and only valuable inasmuch as they confirm pre-existing phonetic expectations.

5. REFERENCES

[1] M. H. Davis, W. D. Marslen-Wilson, and M. G. Gaskell, "Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition," *Journal of experimental psychology: Human perception and performance*, vol. 28, no. 1, pp. 218–244, 2002. [Online]. Available: http://dx.doi.org/10.1037/0096-1523.28.1.218

[2] A. P. Salverda, D. Dahan, and J. M. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," *Cognition*, vol. 90, no. 1, pp. 51–89, 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0010027703001392

[3] K. B. Shatzman and J. M. McQueen, "Prosodic knowledge affects the recognition of newly acquired words," *Psychological Science*, vol. 17, no. 5, pp. 372–377, 2006.

[4] L. J. Blazej and A. M. Cohen-Goldberg, "Can We Hear Morphological Complexity Before Words Are Complex?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 41, no. 1, pp. 50–68, 2015.

[5] R. J. J. K. Kemps, M. Ernestus, R. Schreuder, and R. H. Baayen, "Prosodic cues for morphological complexity: the case of Dutch plural nouns." *Memory & Cognition*, vol. 33, no. 3, pp. 430–446, 2005.

[6] R. J. J. K. Kemps, L. H. Wurm, M. Ernestus, R. Schreuder, and R. H. Baayen, "Prosodic cues for morphological complexity in Dutch and English," *Language and Cognitive Processes*, vol. 20, no. 1/2, pp. 43 – 73, 2005.

[7] R. Smith, R. Baker, and S. Hawkins, "Phonetic detail that distinguishes prefixed from pseudo-prefixed words," *Journal of Phonetics*, vol. 40, no. 5, pp. 689–705, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.wocn.2012.04.002

[8] A. Hjortdal, J. Frid, and M. Roll, "Phonetic and phonological cues to prediction: Neurophysiology of Danish stød," *Journal of Phonetics*, vol. 94, sep 2022.

[9] M. Clayards, M. G. Gaskell, and S. Hawkins, "Phonetic detail is used to predict a word's morphological composition," *Journal of Phonetics*, vol. 87, jul 2021.

[10] C. Cohen and M. Carlson, "Shifting between storage and computation in lexical retrieval: Evidence from pronunciation variation," in *Interfaces of Phonetics*, M. Schlechtweg, Ed. De Gruyter, to appear.

[11] M. H. Christiansen and N. Chater, "The Now-or-Never Bottleneck: A Fundamental Constraint on Language." *Behavioral and Brain Sciences*, vol. 39, pp. 1–52, 2016. [Online]. Available: http://journals.cambridge.org/abstract_S0140525X1500031X

[12] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?" *Language Cognition & Neuroscience*, vol. 31, no. 1, pp. 32–59, 2016. [Online]. Available: http://dx.doi.org/10.1080/23273798.2015.1102299

[13] G. T. M. Altmann and Y. Kamide, "Incremental interpretation at verbs: Restricting the domain of subsequent reference," *Cognition*, vol. 73, no. 3, pp. 247–264, 1999.

[14] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP)," *Annual Review of Psychology*, vol. 62, no. 1, pp. 621–647, 2011. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.psych.093008.131123

[15] A. Ito, M. J. Pickering, and M. Corley, "Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study," *Journal of Memory and Language*, vol. 98, pp. 1–11, 2018. [Online]. Available: https://doi.org/10.1016/j.jml.2017.09.002

[16] R. L. Newman and J. F. Connolly, "Electrophysiological markers of pre-lexical speech processing: Evidence for bottom-up and top-down effects on spoken word processing," vol. 80, pp. 114–121, 2009.

[17] R. C. N. D'Arcy, J. F. Connolly, E. Service, C. S. Hawco, and M. E. Houlihan, "Separating Phonological and Semantic Processing in Auditory Sentence Processing: A High-Resolution Event-Related Brain Potential Study," *Human Brain Mapping*, vol. 22, no. 1, pp. 40–51, 2004.

[18] J. F. Connolly and N. A. Phillips, "Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences," *Journal of Cognitive Neuroscience*, vol. 6, no. 3, pp. 256–266, 1994.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2015. [Online]. Available: http://www.praat.org

[20] D. J. Barr, "Analyzing 'visual world' eyetracking data using multilevel logistic regression," *Journal of Memory and Language*, vol. 59, no. 4, pp. 457–474, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.jml.2007.09.002

[21] S. N. Wood, "Thin-plate regression splines," *Journal of the Royal Statistical Society (B)*, vol. 65, no. 1, pp. 95–114, 2003.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[23] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2022, r package version 2.4.1.

SELF-MONITORING OF SPEECH ERRORS: EFFECTS OF PHONETIC CONTRAST

H. Quené and S.G. Nooteboom

Institute for Language Sciences, Utrecht University, the Netherlands
h.quene@uu.nl, s.g.nooteboom@uu.nl

ABSTRACT

Speech sound errors are known to occur more often as the interfering speech sounds are phonetically more similar. This paper aims to test the general hypothesis that the phonetic contrast between interfering speech sounds also mediates the odds of detection and repair of such a speech error. This was investigated by re-analyzing responses from four published experiments in which speech errors had been elicited, here using Bayesian modeling with random effects of participants and stimuli.

Results show that with increasing phonetic contrast among the segment involved, speech errors occur less frequently (as expected); those errors tend to be detected and repaired more frequently, with early repairs being more prevalent than late repairs, as predicted. These patterns suggest that repair is not triggered by conflict during production, but by error detection during self-monitoring.

Keywords: self-monitoring, speech error, speech production, speech preparation.

1. INTRODUCTION

Most speakers produce occasional speech errors, some of which they detect and repair. Detection (and repair) may occur relatively *early* after the error (before speech is initiated), e.g. *v..horizontal* [1]; presumably such detection is based on the speaker "hearing" her/his own internal speech, as the error-to-interruption interval is too short for auditory processing and speech interruption. Detection may also occur relatively *late* after the error, presumably based on the speaker hearing his/her own overt external speech [2]. Both routes of detection involve comprehension or perception of the speaker's own (internal or external) speech. Production-based models of self-monitoring, by contrast, presume that self-monitoring occurs during speech production (not perception). In one such model [3], conflict may arise between multiple speech sounds competing for the same slot in the sequence being prepared, and this conflict in itself (indirectly) triggers prevention or repair of the impending speech sound error.

The present paper aims to test the general hypothesis that the phonetic contrast (distance, inverse of similarity) between the competing speech segments affects self-monitoring. Speech errors are known to occur less frequently as the phonetic contrast between competing sounds increases [e.g. 4, 5], but here we focus on whether error detection and repair too is mediated by phonetic contrast among the segments involved. With increasing phonetic contrast between competing sounds, we predict (1) that errors will *occur* less frequently (as reported before), (2) that those errors will however be *detected* more frequently (due to their greater saliency during monitoring), and (3) that those detections will occur *early* (rather than late) more frequently.

2. METHOD

To test the above predictions, this paper re-analyzes data from four so-called SLIP experiments [6] involving Dutch elicited speech errors. The raw data were taken from [7, 8].

2.1. SLIP experiments

SLIP experiments aim to elicit exchange errors between the initial consonants in a two-word C_1VC C_2VC stimulus, e.g. *zaal boom* in Dutch. The exchange is elicited by means of presenting priming C_2VC C_1VC precursors, e.g. *bon zet, bek zeef, baai zoop*. The precursor word pairs and the stimulus word pair appear subsequently on the computer screen (for 900 ms + 100 ms blank screen) and the participant needs to read each word pair silently. Every now and then, a visual prompt (*??????*) appears on the screen for 900 ms, upon which the participant has to speak aloud the last-presented word pair. In the SLIP experiments re-analysed here, the target stimulus was always the 5th word pair in a trial, and it was always preceded by at least 3 priming precursors. A typical spoken response containing the elicited error might be *baa.. zaal boom*. Participants were under time pressure, because they had only 900 ms to respond (after the prompt) before a buzzer would sound.

Dependent variables are the category of response (see below), and for error responses whether or not the participant overtly detected his/her speech error (true for the example above), and for detected errors also the time delay (in ms) from the onset of the erroneous consonant to the moment the speaker

finished or broke off her/his spoken response (e.g. duration of *baa*). From the response set, hesitations and omissions were removed, with N=32 368 observations from 433 participants remaining.

2.2. Phonetic contrast

The SLIP stimuli in the present corpus varied in the strength of phonetic contrast (or similarity) between the two interfering initial consonants in the stimulus word pair. The two consonants could differ

- in two features, viz. both place and manner of articulation (labeled *pm2*; 15 096 observations from all experiments);
- in a single feature, viz. either in place or in manner of articulation, but not in both (labeled *pm1*; 13 469 observations from all experiments);
- in their voicing feature (labeled *voicing*; 3 803 observations from two experiments).

Although the latter two contrasts both involve a single feature, the voicing contrast has been shown to be weak for Dutch speakers [9], and Dutch listeners perceive voicing of initial consonants less accurately than either place or manner [10, p.3674].
Observations involving vowel exchanges (in one experiment) have been ignored in the present paper, because the timing of (errors in) initial consonants and nuclear vowels cannot be compared sensibly.

2.3. Responses

Responses were classified into one of the categories listed in Table 1.

label	description	example
fluent	fluent, correct	*zaal boom*
elic. error (repaired)	elicited single error, repaired	*baa.. zaal boom*
elic. error (unrepaired)	elicited single error, unrep'd	*baal zoom*
other error	other error(s)	*boog baan*
hesitation	hesitation or omission	*...eh...*

Table 1: Response categories for an example stimulus *zaal boom* in a Dutch SLIP experiment.

In total, there were 28 508 fluent responses, 1 803 elicited errors (5.6%), and 2057 other errors. The SLIP technique is obviously very inefficient in eliciting speech errors. Moreover, errors were distributed very unequally over participants, as shown in Fig.1: 63 participants (15%) did not produce any elicited error over n=72 or n=110 trials per experiment, and most participants produced only a single elicited error. The top 77 (18%) contributing participants made half of the elicited errors (dark bars in Fig.1), and the top 177 participants (or 41%) contributed 1442 (or 80%) of the elicited errors. The distribution of errors over stimuli is similarly unequal. In our analyses below we will therefore consider both the effects of participants and of stimulus items.

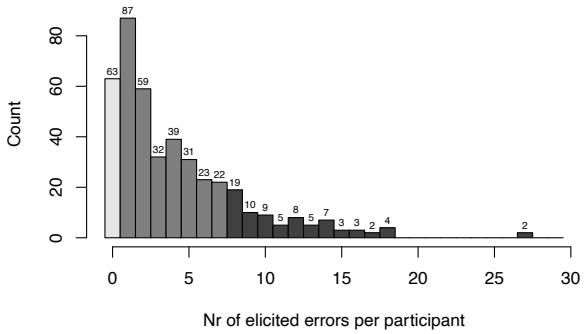


Figure 1: Distribution of the number of elicited errors, over 433 participants in the corpus of SLIP responses.

2.4. Elicited errors detected early or late

Of the 1803 elicited errors, 483 (27%) had been detected and repaired. Following [11] the bimodal distribution of the log-transformed error-to-cutoff times was used to classify repairs as either *early* (presumably based on internal speech) or *late* (presumably based on overt speech, cf. §1). Unsupervised mixture modeling [12, 13] in R [14] suggested two underlying gaussian components making up this bimodal distribution. (Here 5 responses were ignored because of missing error-to-cutoff times, and 5 were ignored because of outlier values between 7 and 15 ms.) Peaks of the two gaussians were at 133 ms and 592 ms respectively. Using the threshold error-to-cutoff time at 425 ms, errors were classified as detected and repaired *early* (371 errors, 21%) or *late* (112 errors, 6%).

2.5. Models

The three predictions above were tested by means of three mixed-effects models, viz. one for each prediction. First, using all remaining responses, multinomial model (1) assesses the effect of phonetic contrast (similarity, 3 categories, see §2.2) on the odds of an elicited error and on the odds of other error(s); the baseline response category are the fluent and correct responses.

(1) responsecat ~ 0 + sim +
(1 + sim | ID) + (0 + sim || stim),
family=multinomial, data=allresponses

The second, binomial model (2) zooms in on the 1803 elicited errors (see §2.3), and assesses the effect of phonetic contrast (similarity) on the odds of an

elicited error being detected; the baseline repair category are the undetected (unrepaired) responses.

(2) (repaircat!=unrep) ~ 0 + sim + (1 + sim | ID) + (0 + sim || stim), family=binomial, data=elicitederrors

Third, multinomial model (3) also zooms in on the 1803 elicited errors, and assesses the effect of phonetic contrast (similarity) on the odds of an early repair and on the odds of a late repair; the baseline repair category are again the unrepaired responses.

(3) repaircat ~ 0 + sim + (1 + sim | ID) + (0 + sim || stim), family=multinomial, data=elicitederrors

Thus, models (1) to (3) included population-level effects of phonetic contrast (or similarity), group-level effects of participants (ID) and of stimuli, and allowed effects of phonetic contrast to vary over participants and over stimuli ("random slopes").

2.6. Analysis

Models (1) to (3) were estimated using Bayesian methods in R [15, 16, 17]. Each model was estimated in 4 independent chains of 3000 iterations (with 1000 warmup), using NUTS sampling. This yielded 8000 post-warmup iterations.

For group-level ("random") estimates, we report the 95% credibility interval of the posterior distribution. For population-level ("fixed") estimates, we report the 95% highest posterior density interval (HDI) [18], which is the narrowest interval containing 95% of the probability mass of the posterior distribution. If two model parameters have non-overlapping HDIs, then we have good grounds to believe that those parameters are different. For all models reported below, R-hat and other diagnostics did not indicate any convergence problems.

3. RESULTS

In all models (1−3), between-item variation (in items' odds per category) was similar across conditions of phonetic contrast. In the multinomial models (1) and (3), between-participants correlations among their odds across conditions were not credibly different from zero.

Multinomial **model (1)** estimated the odds of an elicited error, and of unelicited error(s), against the baseline of correct and fluent responses (*N*=32 368 observations). The **group-level** coefficients showed that between-participant variation (in participants' odds of the elicited error and in their odds of other errors) was higher in conditions eliciting errors in place and/or contrast (i.e. in conditions *pm1* and *pm2*) as compared to the condition eliciting errors in *voicing* [for elicited errors: sd(*intercept*) 0.54 (0.42, 0.66), *pm1* +0.31 (0.06, 0.54), *pm2* +0.51 (0.26, 0.73)]. Thus, participants are more similar in their propensity of voicing errors than in their propensity of errors involving place and/or manner -- even though voicing errors were elicited in only two of the four SLIP experiments (see §2.2).

The **population-level** coefficients of model (1), for elicited errors, are illustrated in the left panel of Figure 2, broken down by phonetic contrast.

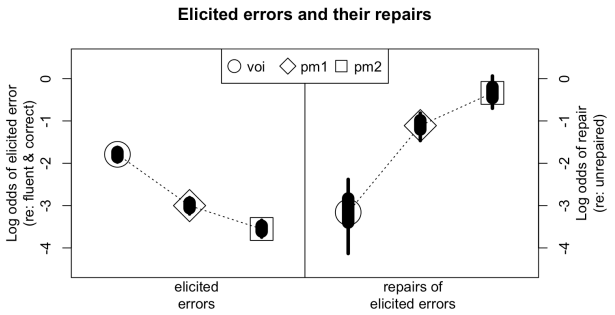


Figure 2: Summary of posterior distributions of population-level coefficients in models (1) and (2), broken down by strength of phonetic contrast. Symbols are plotted at the median of the posterior, thick lines denote 50% HDI and thin lines denote 95% HDI (see text). Left: model (1), log odds of elicited error; right: model (2), log odds of repair of elicited error.

The odds of an elicited error are highest if a voicing error is elicited, considerably lower if a place-or-manner error is elicited, and lowest if a place-and-manner error is elicited. The odds of other (non-elicited) errors, not shown in Figure 2, follow the same pattern, with medians at −2.60, −2.74 and −3.02 respectively, but with overlapping 95% HDIs.

Binomial **model (2)** zooms in on the odds of repair of the elicited errors only (*N*=1803 observations). The **group-level** coefficients showed again that between-participant variation (in their odds of repair of an elicited error) was higher in conditions eliciting *pm1* and *pm2* errors as compared to the condition eliciting voicing errors [repair: sd(*intercept*) 0.93 (0.58,1.28), *pm1* +0.60 (0.04, 1.36), *pm2* +0.70 (0.02, 1.70)]. Thus, participants are more similar in their propensity to repair an elicited voicing error than to repair an elicited error involving place and/or manner.

The **population-level** coefficients of model (2) are illustrated in the righthand panel of Figure 2. Remarkably, the pattern is the opposite of that in the lefthand panel: the odds of repair of an elicited error are lowest if a voicing error is elicited [−3.21 (−4.24, −2.44])], considerably higher if a place-or-manner is elicited [−1.12 (−1.46, −0.81)], and highest if a place-and-manner error is elicited [−0.34 (−0.72, +0.04)].

Multinomial **model (3)** also zooms in on the elicited errors only. The **group-level** coefficients showed once again that between-participant variation

(in their odds of early and of late repairs) was higher in conditions eliciting errors in place and/or contrast (i.e. in conditions *pm1* and *pm2*) as compared to the condition eliciting errors in *voicing* [early repairs: sd(*intercept*) 0.60 (0.05, 1.12), *pm1* +0.49 (0.02, 1.21), *pm2* +0.57 (0.03, 1.33); late repairs: sd 0.90 (0.13, 1.64), *pm1* +0.65 (0.03, 1.65), *pm2* +1.38 (0.09, 3.17)]. Thus, participants are more similar in their repair propensity of voicing errors than in their repair propensity of errors involving place and/or manner. Because there were only 30 *pm2* errors being repaired late, the variation among participants in estimating the odds of late repair must be very large.

The **population-level** coefficients of model (3) are illustrated in Figure 3.

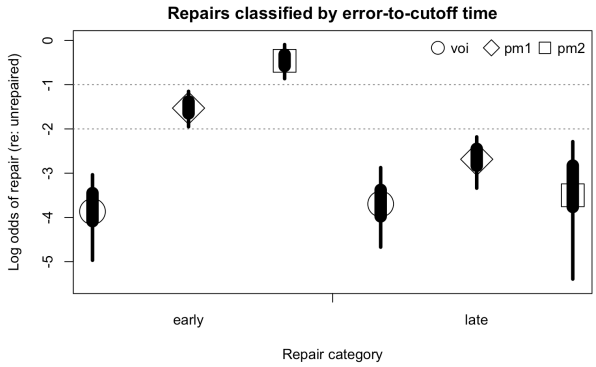


Figure 3: Summary of posterior distributions of population-level coefficients in model (3), broken down by strength of phonetic contrast and by type of repair. (See Figure 2.)

The odds of *early* repair are lowest if a voicing error is elicited [−3.93 (−4.97, −3.04)], considerably higher if a place-or-manner error is elicited [−1.54 (−1.95, −1.15)], and highest if a place-and-manner error is elicited [−0.46 (−0.86, −0.09)], with non-overlapping 95% HDIs. The odds of late repair are approximately the same across the phonetic contrasts. Thus the effects illustrated in the righthand panel of Figure 2 (odds of repair) are mainly due to early repairs, and not to late repairs.

4. DISCUSSION

The results of model (1) (Fig.2, left) confirm that the relative number of speech errors increases with phonetic similarity of the interfering consonants in the stimulus (and decreases with their phonetic contrast) [e.g. 4, 5], most clearly for interfering consonants contrasting in place and/or in manner. Moreover, the higher error rates for stimuli involving the voicing contrast confirm that this voicing contrast is indeed considerably weaker in Dutch (in speech preparation and articulation and monitoring) than either a place or a manner contrast [cf. 9].

In this paper, however, our focus is on whether and how quickly these elicited speech errors are detected and repaired. We have predicted that errors involving a stronger phonetic contrast are detected more frequently. This prediction was confirmed by the results of model (2) summarized in Fig. 2 (right). Elicited voicing errors (circles) are relatively common, but they are detected and repaired relatively rarely; elicited place-and-manner errors (squares) are relatively rare, but these are detected and repaired relatively frequently (with high odds of detection and repair); elicited place-or-manner errors have intermediate odds of error and of repair.

The results of model (3) in Fig. 3 show that the effect of phonetic contrast works mostly in early repairs, rather than in late repairs. Elicited voicing errors (circles) are hardly repaired early; elicited place-and-manner errors (squares) are repaired early quite often; elicited place-or-manner errors (diamonds) have intermediate odds of early repair, all with non-overlapping posterior distributions.

Regarded differently, the prevalence of early repair over late repair increases with phonetic contrast: there is no such prevalence for voicing errors (circles); the prevalence is strongest for repair of errors involving both place and manner contrasts (squares), with an intermediate prevalence for errors involving either a place or a manner contrast (diamonds). Thus speech errors involving a strong phonetic contrast tend to be detected and repaired early (Fig. 3); errors involving a weak contrast such as voicing tend to be detected early less often (than strong-contrast errors, Fig. 3) or these weak-contrast errors tend to be detected not at all (low odds of detection, Fig. 2).

The pattern in Figure 2 is difficult to reconcile with a conflict-based theory of self-monitoring [e.g. 3]. We assume that stronger phonetic contrast between two competing items would result in a greater separation between the two items in their relative activation and therefore in *less* conflict during production. (This would lead to fewer errors, which was indeed observed in the lefthand panel of Fig. 2.) Less conflict would however also lead to fewer (and later) detections and repairs, contrary to the observed pattern in Figs. 2 and 3.

In conclusion, phonetic contrast is a major determinant of early detection vs. late detection vs. non-detection of segmental speech errors in self-monitoring. We also conclude that "cognitive control" as a mechanism that prevents or repairs segmental errors is not triggered by conflict between competing response candidates, but rather by error detection.

5. REFERENCES

[1] Levelt, W.J.M. 1989. *Speaking: From intention to articulation*. MIT Press.

[2] Hartsuiker, R. J., Kolk, H. H. J. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology* 42(2), 113–157.

[3] Nozari, N., Dell, G. S., Schwartz, M. F. 2011. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology* 63(1), 1–33.

[4] Nooteboom, S.G. 1973. The tongue slips into patterns. In: V.A. Fromkin (ed) *Speech Errors as Linguistic Evidence*. Mouton, 144–156.

[5] Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3), 283–321.

[6] Baars, B. J., Motley, M.T. 1974. Spoonerisms: Experimental elicitation of human speech errors. *Catalog of Selected Documents in Psychology* 3, 28–47.

[7] Nooteboom, S.G., Quené, H. 2008. Self-monitoring and feedback: a new attempt to find the main cause of lexical bias in phonological speech errors. *J. Memory and Language* 58(3), 837–861.

[8] Nooteboom, S.G., Quené, H. 2013. Parallels between self-monitoring for speech errors and identification of the misspoken segments. *J. Memory and Language* 69(3), 417–428.

[9] van Alphen, P.M., Smits, R. 2004. Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. *J. Phonetics* 32(4), 455–491.

[10] Cutler, A., Weber, A., Smits, R., Cooper, N. (2004). Patterns of English phoneme confusion by native and non-native listeners. *J. Acoust. Soc. Am.* 116(6), 3668–3678.

[11] Nooteboom, S.G., Quené, H. 2017. Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *J. Memory and Language* 95, 19–35.

[12] Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R Journal* 8(1), 289–317.

[13] Fraley, C., Raftery, A.E. 2002. Model-based clustering, discriminant analysis and density estimation. *J. Am. Statistical Assoc.* 97, 458, 611–631.

[14] R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

[15] Bürkner, P.-C. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Statistical Software* 80(1), 1–28

[16] Bürkner, P.-C. 2018. Advanced Bayesian multilevel modeling with the R package brms. *R Journal* 10(1), 395–411.

[17] Makowski, D., Ben-Shachar, M.S., Lüdecke, D. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *J. Open Source Software* 4(40), 1541.

[18] McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press.

LABELS SPEAK VOLUMES: OUTGROUP LABELS AND ACCENT STRENGTH PERCEPTION

Kasandra Calkins[1], Veranika Puhacheuskaya[1], Jiseung Kim[2], Juhani Järvikivi[1]

University of Alberta[1], Radboud University[2]
kcalkins@ualberta.ca

ABSTRACT

Extralinguistic factors have been shown to affect the perceived strength of foreign accents. The current research investigates whether labelling speakers along a group membership continuum (*Canadian, new Canadian, immigrant,* and control/no label) will affect accent strength ratings. In a within-subjects design, 82 L1 speakers of Canadian English rated 32 recordings of accented speakers from four broad language backgrounds (Canadian-English, Middle-Eastern, European and Asian) across the four label conditions. It was found that speakers were rated as having stronger accents when introduced with an outgroup label *(new Canadian/immigrant*) compared to ratings of the speakers introduced with no label. Importantly, ratings of speakers labelled as *Canadians* did not significantly alter perceptions, suggesting that the term *Canadian* is perceived to be inclusive to both L1 and L2-accented speakers. Studying factors that increase perceived foreign accent strength is critical to identifying and reducing sources of prejudice within an increasingly globalised world.

Keywords: accented speech perception, immigration, intergroup membership

1. INTRODUCTION

Perception of L1 and L2 accented speech is influenced by extralinguistic factors. Visual information [1, 2] and other information about the speaker including their length of residency in the country [3] have been found to bias perceptions of accents to match our expectations about the speaker. With respect to visual information, [2] found that participants perceived the same L1 speaker as more strongly accented when they saw the image of an ethnic minority than when they saw the image of a white individual. Similarly, [1] played audio of L1 speakers of Canadian English with images of either ethnically Chinese or white individuals, discovering that participants perceived speakers as less accented when paired with an ethnically white face and more accented when presented with an image of an ethnically Chinese face in comparison to the audio only trials. Extralinguistic information also aids language processing. Providing information about the source of a speaker's accent can improve the intelligibility of the speaker [4], whereas visual primes that confirm our expectations improve both the intelligibility [1, 5] and grammatical processing of foreign accented speech [6].

Consistent with *reverse linguistic stereotyping* [7] which states that evaluations of others are informed by general attributes of the speaker's group membership, we predicted that providing extralinguistic information regarding the speaker's group membership would modulate accent strength ratings to be more consistent with a given group membership label. Specifically, we predicted that ingroup labels would decrease and outgroup labels would increase the perceived foreignness of accents.

2. BACKGROUND

2.1. Language attitudes

Language attitudes are related to and partially driven by accent strength, informing expectations and perceptions of others. Generally, there is a bias to prefer native accented speech over foreign accented speech [8], with negative language attitudes increasing in accordance with accent strength [9]. One theory to account for this preference is the *fluency principle* [10] which states that foreign accented speech evokes negative attitudes because it is harder to process. Difficulty in processing foreign accented speech may additionally impede the ability to predict the upcoming speech signal [11] and influence the believability of statements made by foreign accented speakers [12]. However, experience with a particular accent can mitigate discrimination [13] and improve predictive processing [11].

2.2. Immigrant labels

Negatively labelling noncitizens has been found to increase prejudice and support for harsher immigration policies [14]. However, experience or familiarity with a specific immigrant community can

reduce prejudice and elicit more positive attitudes towards immigrants more generally.

2.3. Disgust and political ideology

Disgust sensitivity and political ideology have been shown to correlate with the strength of outgroup prejudice; high levels of disgust sensitivity [15] and conservatism [16] have been linked to anti-immigration sentiments. Additionally, both constructs have been shown to affect language processing and comprehension [17, 18].

2.4 Present study

We studied the effect of group membership labels on accent strength perception. Specifically, we were interested in whether labelling a speaker as either a *Canadian* (ingroup member), *immigrant* (outgroup member) or as a *new Canadian* (partial in-/outgroup member) to native listeners of Canadian English would modulate the degree of perceived accent strength for native and non-native accented speakers. We predicted that the *Canadian* label would encourage ingroup associations and therefore decrease the perceived accent strength compared to audio-only ratings. Furthermore, we predicted that the two outgroup labels (*new Canadian* and *immigrant*) would encourage outgroup association and subsequently increase the perceived foreign accent strength of the speaker. We expected this relationship to be stronger for the *immigrant* label due to the nature of the *new Canadian* label having both ingroup and outgroup connotations. We were further interested in whether the participants' political ideology and disgust sensitivity scores predicted accent ratings.

3. METHOD

3.1. Auditory stimuli selection

A pretest was conducted to select stimuli for the main experiment. Auditory stimuli were sourced from the *Speech Accent Archive* [19] which contains recordings of native and non-native speakers from various linguistic backgrounds reading the same short passage in English. Recordings of 33 speakers were selected for the pretest. They were divided into two parts (passage A and B) to be around 12 seconds in length. The intensity was scaled to 70 dB. A group of 15 native speakers of Canadian English rated each recording for accent strength, likeability and intelligibility on a 7-point scale. While keeping likeability and intelligibility ratings relatively constant, native speaker stimuli were selected to be of roughly equivalent accent strength. Similarly, the non-native speaker stimuli were selected to form two categories of equivalent foreign accents (strong, weak).

The final materials included 16 speakers and 32 recordings in total:12 non-native speakers of English (24 recordings in total) and four native speakers of English (eight recordings in total). Four accent background groups were created, with four speakers in each group: Canadian, Asian, European, and Middle Eastern. Each group consisted of two male and two female speakers. In the foreign accented language groups, weak and strong foreign accented speakers were divided equally by speaker gender to have one weak and one strong foreign accented speaker of each gender. The European language group consisted of L1 speakers of Swedish, French, Polish and Greek; the Middle Eastern group consisted of L1 speakers of Farsi and Arabic; and the Asian language group consisted of L1 speakers of Mandarin and Cantonese. Four counterbalanced lists of stimuli were created to ensure an equal number of recordings from each accent background and gender across conditions, as well as to ensure that participants would rate the same speaker twice (passage A and B) in two separate conditions.

3.2. Main experiment

3.2.1. Participants

In total, 82 native speakers of Canadian English were recruited from Prolific.co [20] with ages ranging from 18-49 years old (M = 24.5; SD = 6.3). Participants self-reported their gender, with 50 identifying as women, 30 as men, and 2 individuals preferring not to disclose. All participants reported that Canadian English was their native language and that they had begun acquiring it from early childhood (before age four). Participants also reported their multilingual status, with 25% identifying as multilingual and 75% as monolingual.

3.2.2. Procedure

The main experiment utilised a within-subjects design, presenting all four label conditions to each participant in blocks and randomly assigning them to one of the four counterbalanced lists of stimuli. The experiment always began by presenting the control condition (no label) before presenting the conditions of interest to prevent any crossover effects. In the control condition, recordings of the speakers were presented without any additional information about the speaker and then rated on a 7-point scale according to how native-like a speaker's accent was perceived to be (1-not at all foreign, 7-very foreign). After the control condition was completed, a

counterbalanced order of the conditions of interest was presented to participants using the same carrier sentence to describe the speaker with different labels ("This speaker is a *Canadian/new Canadian/immigrant*") and rated for accent strength.

3.2.3. Exit questionnaire

After the main experiment, participants completed an exit questionnaire assessing their demographic information and language background as well as their political ideology and disgust sensitivity. Participants were asked to indicate whether they identified as Canadian (*yes*, *no*, *partially*), and indicate how often they spent time with both L1 and L2-accented speakers of English (1-very frequently, 5-very infrequently). The political ideology questionnaire [21] consisted of 27 questions, with higher scores indicating increasing degrees of conservatism. The disgust sensitivity questionnaire consisted of 28 questions measuring participants' self-reported level of disgust to hypothetical situations, with higher scores indicating greater sensitivity [22].

3.3. Statistical analyses

One participant was removed from analysis due to failing the attention check in the exit questionnaire, resulting in data from 81 participants in the final analysis. A two-stage modelling process was performed using R (4.1.1) [23] and generalised additive mixed modelling for ordinal data [24] from the *mgcv()* package [25]. The first stage investigated the effect of label condition on accent ratings and the second stage explored the effect of individual difference measures. Models were forward fit to the data using a hierarchical approach, beginning with label condition and then adding fixed and random effects to the model based on theoretical relevance. Models were then compared using the *compareML()* function from the package *itsadug* [26] to test whether adding a new predictor explained significant variability. If the new model did not significantly reduce the model's REML score, it was removed from subsequent analysis. The final model included the main manipulation of label condition on ratings of accent strength, speaker background (Canadian, European, Asian and Middle Eastern), accent strength (strong, weak, native) with random effects for participants and stimuli as well as by-trial smooths for participants.

In the second stage model fitting, the two individual difference measures of political ideology and disgust sensitivity were centred to the mean and standardised before being added to the final model from stage one as smooth predictors. Interpretation of the individual difference measures was done by visualising the smooth terms using the functions *plotsmooth()* and *plotdiff()* from the package *itsadug* [26]. An additional grouping factor was created to explore how the two individual difference measures interact with label condition and speaker background; however, it did not result in any interpretable differences across levels. Running the previous model from stage one with the smooth predictors as a function of speaker accent strength produced a better fitting model and was kept as the final model for the second stage, containing smooths for both political ideology and disgust sensitivity.

4. RESULTS

4.1. Accent strength ratings

The final model revealed that speakers were perceived as significantly more accented when presented with the *new Canadian* label (β = .34, p = .008) and the *immigrant* label (β = .53, p < .001) compared to the control condition, with the largest effect for the *immigrant* label followed by the *new Canadian* as visualised in Figure 1. However, releveling the model to compare the *new Canadian* and *immigrant* labels did not show a significant difference between the two labels (p = .12).

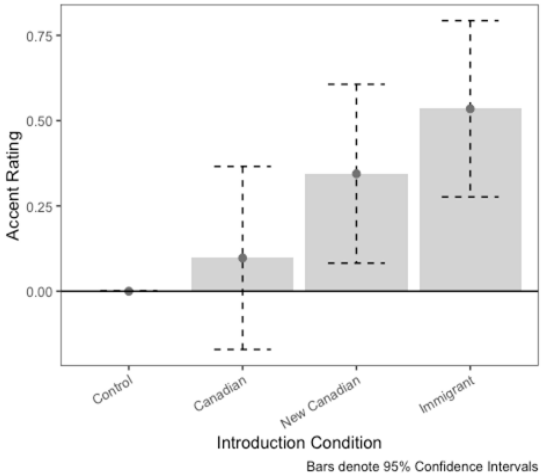


Figure 1: The effect of introduction label on accent strength ratings for all speakers.

Ratings of speakers presented with the Canadian label did not significantly differ from the control condition. Factors pertaining to the participants (age, gender, Canadian identity), their language background (multilingual status) and social networks (frequency of interaction with non-native speakers of English), did not have a significant effect on accent ratings.

4.1.1. Individual difference factors

Scores for political ideology ranged from 55-109 (M = 77.7, SD = 6.4) and disgust scores ranged from 62-

94 (M = 85.1, SD = 13.6). Participants' scores for political ideology and disgust sensitivity were not correlated and were therefore added concurrently to the model as smooth predictors. The smooth terms for disgust did not reveal any significant effects across the accent strength levels. Smooths for political ideology by each accent type (weak, strong and native accented) yielded a significant effect for ratings of native speakers ($p < .05$). Smooths for the strong and weak foreign accented speakers were nonsignificant. The difference in smooths for native and weak foreign accented speakers was significant, suggesting that more conservative participants perceived less of a difference between the weak foreign accents and native accents overall than more liberal participants. This result is consistent with previous literature [17-18, 27], and suggests that sensitivity to voice-based, socially relevant information such as foreign accents may increase as a function of relative liberalism.

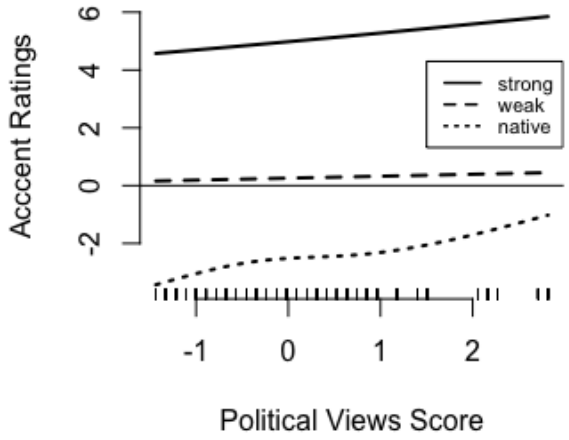


Figure 2: Ratings of accent strength by speaker accent type and political ideology scores.

5. DISCUSSION

Our hypothesis was partially confirmed. Accent strength perception was found to be malleable to outgroup labels which increased the perceived strength of accents. As hypothesised, the *immigrant* label resulted in the largest increase in perceived accent strength, followed by the *new Canadian* label. This indicates that the *new Canadian* label did not evoke as strong of an outgroup response. However, there was no significant difference between the *immigrant* and *new Canadian* labels. Contrary to our initial prediction, labelling speakers as *Canadian* did not result in more native-like perception, potentially signalling that the *Canadian* label is a neutral and inclusive term to both L1 and L2-accented speakers. This is complementary to previous research on audio-visual accented speech perception within Canada, which failed to find an increase in accent strength upon changing the speaker's race [28]. Due to Canada's global reputation and positive stance on immigration, studying the effects of labelling foreign accented populations is critical to reducing negative language attitudes and bias towards foreign accented speakers by promoting inclusive and non-prejudicial labels.

How we choose to label individuals may have real world consequences because language attitudes increase as a function of accent strength [10]. Even though the two outgroup labels used in this study are considered neutrally charged, they resulted in increasing perceived accent strength. Further research on extralinguistic factors affecting accent perception may investigate the effects of labels with stronger negative connotations such as *illegal alien* which have proven to increase prejudice relative to neutral labels [14]. Other extralinguistic factors to be investigated may include manipulating other identity markers such as the profession/status or the first/last name of the speaker.

With respect to individual differences, it was found that participants with different political leanings varied in their perception of foreign accent strength. More liberal participants reported a larger difference in accent strength ratings between the native accented speakers and the weak foreign accented speakers, which may be in part due to their increased sensitivity to socially relevant information [30, 31]. No conclusive evidence was found for disgust sensitivity as a moderator of accent strength ratings.

Voice based inferences are rapid [29] and automatic [32], providing social and group membership information [8]. Evaluations of speakers are made using both acoustic and overarching group membership information [7], leading our perception of accents and evaluations of speakers to be biased to match our expectations.

6. CONCLUSION

The results of the current study suggest that group membership labels can inform our perception of accent strength. Outgroup labels were found to increase the perceived foreignness of accents, which may in turn compound prejudice towards minority groups. In contrast, labelling with ingroup framing (*Canadian*) did not alter perceptions of accent strength. This result may potentially shed light on the broader cultural experience within Canada, indicating that Canadians perceive the *Canadian* label to be inclusive of both native and foreign-accented speakers of English. More generally, ratings of accent strength were found to be further modulated by individual differences in political ideology.

7. REFERENCES

[1] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137(5), 2823–2833.

[2] Rubin, D. L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education* 33(4), 511–531.

[3] Dragojevic, M., Tatum, N. T., Beck, A., McAninch, K. 2019. Effects of accent strength expectancy violations on language attitudes. *Communication Studies* 70(2), 133–150.

[4] Vaughn, C. R. 2019. Expectations about the source of a speaker's accent affect accent adaptation. *J. Acoust. Soc. Am.* 145(5) 3218–3232.

[5] Babel, M. 2022. Adaptation to social-linguistic associations in audio-visual speech. *Brain Sciences* 12(7) 845.

[6] Grey, S., Cosgrove, A. L., van Hell, J. G. 2020. Faces with foreign accents: An event-related potential study of accented sentence comprehension. *Neuropsychologia* 147, 107575.

[7] Kang, O., Rubin, D. L. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28(4), 441–456.

[8] Gluszek, A., Dovidio, J. F. 2010. The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review* 14(2), 214–237.

[9]Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., Giles, H. 2012. A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology* 42(1), 120–133.

[10]Dragojevic, M., Giles, H., Beck, A., Tatum, N. T. 2017. The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs* 84(3), 385–405.

[11]Porretta,V., Buchanan, L., Järvikivi, J. 2020. When processing costs impact predictive processing: The case of foreign-accented speech and accent experience. *Attention, Perception, & Psychophysics* 82(4), 1558–1565.

[12]Lev-Ari, S., Keysar, B. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology* 46(6), 1093–1096.

[13]Boduch-Grabka, K., Lev-Ari, S. 2021. Exposing individuals to foreign accent increases their trust in what nonnative speakers say. *Cognitive Science* 45(11) e13064.

[14]Rucker, J. M., Murphy, M. C., Quintanilla, V. D. 2019. The immigrant labeling effect: The role of immigrant group labels in prejudice against noncitizens. *Group Processes & Intergroup Relations* 22(8), 1139–1160.

[15]Aarøe, L., Petersen, M. B., Arceneaux, K. 2017 The Behavioral Immune System shapes political intuitions: Why and how individual differences in disgust sensitivity underlie opposition to immigration. *American Political Science Review* 111(2) 277–294.

[16]Brooks, C., Manza, J., Cohen, E. D. 2016. Political ideology and immigrant acceptance. *Socius* 2, 1-2

[17]Hubert Lyall, I., Järvikivi, J. 2021. Individual differences in political ideology and disgust sensitivity affect real-time spoken language comprehension. *Frontiers in psychology*, 12, 699071.

[18]Puhacheuskaya, V., Järvikivi, J. 2022. I was being sarcastic!: The effect of foreign accent and political ideology on irony (mis) understanding. *Acta Psychologica* 222, 103479.

[19]Weinberger, S. 2015. Speech accent archive. http://accent.gmu.edu/.

[20]"Prolific" online participant recruitment for surveys and market research. https://prolific.co/.

[21]Marrville, C. 2017. Gender and dominance in action: World view and emotional affect in language processing and use. Ph.D. dissertation.

[22]Olatunji, B.O., *et al.* 2007. The Disgust Scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment* 19(3), 281–297.

[23]R Core team. 2021. R: The R Project for Statistical Computing. https://www.r-project.org/.

[24]Baayen, R. H., Divjak, D. 2017. Ordinal GAMMs: A new window on human ratings. In: Makarova, A., Dickey, S. M., Divjak, D. (eds), *Each Venture a New Beginning: Studies in Honor of Laura A. Janda.* Slavica Publishers, 39-56.

[25]Wood, S. N. 2017. Generalized Additive Models: An introduction with R, 2nd ed. New York: Chapman and Hall/CRC.

[26]van Rij, J., Wieling, M., Baayen, R. H., van Rijn, H. 2020. itsadug: Interpreting time series and autocorrelated data using GAMMs. https://cran.r-project.org/web/packages/itsadug/index.html.

[27]Mansell, J. 2020. Ideology and social cognition: Are liberals and conservatives differentially affected by social cues about group inequality? *Politics and the Life Sciences* 39(1), 9–25.

[28]Kutlu, E., Tiv, M., Wulff, S., Titone, D. 2022. Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications* 7(1), 1-16.

[29]Pélissier, M., Ferragne, E. 2022. The N400 reveals implicit accent-induced prejudice. *Speech Communication* 137, 114-126.

[30]Jost, J. T., Amodio, D. M. 2012. Political ideology as motivated social cognition: behavioral and neuroscientific evidence. *Motivation and Emot*ion 36, 55–64.

[31]Hammond-Thrasher, S., Järvikivi, J. 2023. Rating gender stereotype violations: The effects of personality and politics. *Frontiers in Communication* 8, 1050662.

[32]Tuninetti, A., Chládková, K., Peter, V., Schiller, N. O., Escudero, P. 2017. When speaker identity is unavoidable: Neural processing of speaker identity cues in natural speech. *Brain and language* 174, 42-49.

EFFECTS OF STRESS AND SEGMENT POSITION WITHIN A WORD ON SPEECH ERRORS IN CROATIAN

Ana Vidović Zorić

University of Zagreb, Croatia
anvidovi@ffzg.hr

ABSTRACT

One of the major findings based on the investigation of speech errors in English is that more interaction errors occur if two segments share stress than if they do not. Another finding is that more interaction errors occur if two segments share their position within a larger unit. The aim of this study is to investigate these findings in Croatian using tongue twisters as a method for eliciting mutual substitutions of consonants, as one type of speech errors. The results show that in Croatian more substitutions of consonants occur if they are in the unstressed syllable if compared to the stressed ones, thus, the results are not in line with the findings of speech errors in English. Still, the results confirm the claim that two segments interact more frequently if they share the position within a word. The effects of these findings on phonological encoding were discussed.

Keywords: Speech error, substitution, stress, position within a word, Croatian.

1. INTRODUCTION

One of the well-known ways to investigate many aspects of speech production is analysing speech errors, also called slips of the tongue. Thus, phonological speech errors are an important source of information about the phonological encoding – the instance of speech production which refers to the retrieval of target phonemes, their sequencing within an utterance, and preparing for articulation [4, 8, 14]. The role of stress in this process is still not completely understood. Based on the research of experimentally elicited phonological errors, Shattuck-Hufnagel [13] concluded that more interaction errors occur if two segments share stress than if they do not, specifically, if they precede a stressed vowel in a word. The claim is supported by [12]. Still, the results of previous research are contradictory. While most of the studies [3, 5, 11] carried out on the English speech corpus confirm the finding that segments interact more often if they are in a stressed than in an unstressed syllable, Berg's [2] analysis of errors in Spanish showed that there is an equal possibility that segmental errors occur in a stressed or unstressed syllable. Considering this inconsistency between different languages, the aim of this study is to investigate the role of stress in mutual substitutions of target consonants in Croatian, using the tongue twister technique for eliciting errors. Croatian is a pitch-accent language [10] in which a stressed syllable is distinguished from unstressed syllables within the same word by pitch, length, and intensity which are simultaneously realized. Thus, four types of word stress can be distinguished: long rising, short rising, lang falling and short falling [1, 7]. The vowel following the stressed syllable within a word can also be short or long. The stress position in Croatian is relatively free. In polysyllabic words, any syllable except the last one could be stressed, and monosyllabic words can only have a falling accent [1, 7]. Based on the literature on the role of stress in English and Spanish, it could be expected that stress would have an effect on mutual substitutions of target consonants in Croatian as well. Still, since there are no studies that have investigated the role of stress in phonological encoding neither in Croatian nor in other Slavic languages, the kind of effect is difficult to predict, or which direction it goes, that is, whether stress will increase or decrease the number of substitutions.

Another aim of this research is to find out whether two segments interact more frequently if they do or do not share the position within a word in Croatian, using the same tongue twisters. In addition, the experimental design allows testing whether a syllable or word position similarity causes more segmental substitutions. Previous research showed that two segments interact more often if they share the same position in the syllable than if they do not [3, 4, 5, 9], which is known as position similarity constraint. Shattuck-Hufnagel [13] noted that all reported evidence for position similarity constraint is compatible with both syllable and word unit. Furthermore, her experiment, designed to investigate the effect of word position on interactions of segments, showed that more interactions occur if segments share word position than if they do not [13]. Similar experiment was conducted by [12], where this finding is confirmed. In light of the previous research, it could be expected that word position similarity would have an influence on the frequency of substitutions of segments in Croatian as well.

2. METHOD

2.1. Materials

The experimental design used in this research is based on the design used by [13], but adopted for Croatian and for the specific requirements of this research. Stimuli consisted of 18 quadruple sets of Croatian tongue twisters, making a total of 72 tongue twisters. Tongue twisters were presented one by one in a written form on the computer screen, using Arial font size 80, in lowercase, according to standard Croatian orthography rules. The presentation was in the PowerPoint. Each word of the tongue twister was presented in a separate line on the screen. Tongue twisters were constructed of four three-syllable Croatian words, and they had two target consonants (TCs), A and B, which were arranged into the following patterns:

- BO-BS – *both* TCs were in the word-*onset* position in the *stressed* syllable of the word.
- BO-SU – *both* TCs shared word-*onset* position, but they differed in the stress position: one was in the *stressed* syllable, and the other one in the *unstressed* syllable of the word.
- BO-BU – *both* TCs were in the word-*onset* position in the *unstressed* syllable of the word.
- OT-BS – *both* TCs were in the *stressed* syllable of the word, but they differed in the word position: one was in the word-*onset* position, and the other one was in the *third* position within the word.

Each quadruple set contained all four types of TC patterns (for an example, see Table 1).

Table 1: Example of one quadruple set of stimuli in Croatian, along with phonetic transcription and translation in English. TCs /ʒ/ and /z/ are in italics, the stressed vowel is underlined (the target consonant and the stressed vowel were not marked on the presentation to participants).

TC pattern	Stimulus
BO-BS	*ž*a̲lostan *z*u̲barski *ž*e̲tveni *z*i̲dati ['ʒalostan 'zubarski 'ʒetʋeni 'zidati] (eng. sad dental of harvest build)
BO-SU	*ž*a̲lostan *z*uba̲ri *ž*e̲tveni *z*ida̲ri ['ʒalostan *z*u'bari 'ʒetʋeni *z*i'dari] (eng. sad dentists of harvest bricklayers)
BO-BU	*ž*alo̲stiv *z*uba̲ri *ž*eto̲ni *z*ida̲ri [ʒa'lostiv zu'bari ʒe'toni zi'dari] (eng. piteous dentists coins bricklayers).
OT-BS	ložača̲ *z*u̲barski teži̲na *z*i̲dati [lo'ʒaʧi 'zubarski te'ʒina 'zidati] (eng. stoker dental heaviness build)

It is important to note that TCs in all types of TC patterns always shared the syllable-onset position. There were nine pairs of TCs. Based on the system of three different articulatory features, i. e. place of articulation, manner of articulation, and voicing, eight pairs differ only in one feature: /l–r, z–ʒ, s–ʃ, s–ʦ, ʃ–ʒ, n–m, k–x, l–ʎ/, and one pair differs in two features: /ʋ–f/ [6]. For the sake of reliability of the results, the intention of the author(s) was to conduct an experiment using real words. In order to meet this requirement, it was impossible to control all aspects of stressed syllables in designing the tongue twisters, that is, pitch and length were not consistently controlled. Still, wherever possible, these two features were controlled as well. The target pair in all four types appeared in two different word orders:

- ABAB: for example, $ž^{A}alostan\ z^{B}ubarski\ ž^{A}etveni\ z^{B}idati$.
- BAAB: for example, $z^{B}agađen\ ž^{A}utica\ ž^{A}irovi\ z^{B}emljani$; ['zagaʤen 'ʒutiʦa 'ʒiroʋi 'zemʎani]; eng. *polluted jaundice acorns earthlings*.

2.2. Participants

Twenty native Croatian speakers, drawn from the Faculty of Humanities and Social Sciences, University of Zagreb, participated in this experiment. They were all right-handed females, with age ranging from 19 to 38 years old. None reported any history of speech, language, or motor deficits. They did not receive compensation for their participation.

2.3. Procedure

The procedure is similar to the one described in [13] and [14], but also adopted for the purpose of this research. The testing was performed individually in a soundproof studio. The participants were instructed to read each tongue twister aloud, as long as it was presented on the computer screen (reading condition). Then the utterance showed up again on the screen for two seconds. The appearance of the green dot in the middle of the screen was a signal to the participant to begin to repeat the utterance from the memory and to continue doing it as long as the green dot was presented (recall condition). In rare cases when the participant forgot the item, the whole procedure in the recall condition was repeated. The presentation of the items (tongue twister or the green dot respectively) lasted for nine seconds. The utterances were presented in two random orders: half of the participants was given one order (Group 1), and the other half another one (Group 2). The rate of reciting the stimulus was self-paced. The only instruction for the participants was to speak a little bit faster than usual. Before the experimental session, each

participant was given four trial runs. To avoid fatigue, the experiment was divided into four equal sessions and conducted in two days, with a 15-minute break between the sessions on the same day.

2.4. Data analysis

The produced utterances were audio recorded with a microphone and later transcribed by two trained transcribers. In cases where the transcriptions did not match, a third transcriber was engaged. A word was classified as incorrect if it contained at least one erroneous phoneme, i.e., that was not given by the stimulus. Only substitutions of TCs in the first four repetitions of each participant in the reading and recall condition were included in the analysis. For example, if the participant instead of *žalostan* produced *zalostan*, the substitution of /ʒ/ for /z/ was counted as one substitution. The statistical analysis was carried out using the SPSS software package. The Friedman test was conducted to test for differences between TC patterns, followed by six Wilcoxon rank sum tests for individual group comparisons.

3. RESULTS

The participants produced a total of 46 080 words: 23 040 in reading, and the same number during recall. There are 2.73% erroneous words in reading and 5.15% in recall, which is in line with the number of erroneous words in some other similar experiments [14]. This small number of erroneous words indicates that the experimental requirements were not too difficult to meet. In the following sections we will present the mutual substitutions of target consonants, i. e. the cases where one target consonant is replaced with another one, in reading and recall.

3.1. Mutual substitutions of target consonants in reading

A total of 339 mutual substitutions of TCs (TC substitutions) were produced in reading. The difference between participants in the number of TC substitutions in all four types of stimuli is high, which can be seen from the minimum and maximum number of TC substitutions per participant (see Table 2). Such discrepancy in the number of errors between participants has already been noted in [13, 14]. The lowest total number of substitutions (3) has the participant 5 (P5), and the highest number (33) has the participant 8 (P8). Still, the similar error ratio between four patterns was found in the recall condition (see Table 3), which indicates that this ratio has not been accidental.

Table 2: Descriptive statistics in reading: the number of participants (N), minimum (Min) and maximum (Max) number of TC substitutions per participant, total (Sum) number of TC substitutions in all four types of TC patterns in tongue twisters. The percentages indicate a type's percentage of all TC substitutions in reading.

Pattern type	N	Min	Max	Sum	%
BO-BS	20	0	11	99	29.2
BO-SU	20	1	10	91	26.84
BO-BU	20	1	12	117	34.51
OT-BS	20	0	5	32	9.4

As can be seen from Table 2, the highest number of TC substitutions is in the BO-BU. The fewest substitutions are in the OT-BS, and this result is observed in 18 participants. The test shows that the difference in the number of TC substitutions between the four types is significant ($\chi^2 = 24.933$, $p = 0.00$). Post-hoc tests show that the difference between OT-BS and all three other types is reliable (OT-BS vs. BO-BS: $z = -3.558$, $p = 0.00$; OT-BS vs. BO-SU: $z = -3.630$, $p = 0.00$; OT-BS vs. BO-BU: $z = -3.526$, $p = 0.00$). The difference in the TC substitutions between other three pairs of TC patterns is not significant (BO-SU vs. BO-BS: $z = -0.885$, $p = 0.38$; BO-BU vs. BO-BS: $z = -0.852$, $p = 0.39$; BO-BU vs. BO-SU: $z = -1.510$, $p = 0.13$). Bearing in mind that OT-BS is the only type where TCs do not share word position, these results indicate that the position of segments in a word has a greater influence on segmental errors than stress. In other words, if two segments share a position within a word, they are more likely to interact with each other.

3.2. Mutual substitutions of target consonants in recall

Participants produced a total of 646 TC substitutions during recall, which means that the number of TC substitutions is almost twice as many in this condition than in reading. Since the working memory in recall is much more loaded than in reading, this result is expected, and it is in line with [13]. As in reading, individual differences in the number of TC substitutions are high (Table 3). The lowest total number of substitutions (10) has the participant 18, and the highest number (71) has the participant 15. The number of TC substitutions in recall follows the same pattern as in reading: the highest number is in BO-BU, then in BO-BS, followed by BO-SU, and the fewest substitutions are in OT-BS (Table 3).

Table 3: Descriptive statistics in recall. Symbols are the same as in Table 2.

Pattern type	N	Min	Max	Sum	%
BO-BS	20	3	21	188	29.1
BO-SU	20	1	21	172	26.62
BO-BU	20	3	24	231	35.75
OT-BS	20	0	8	55	8.5

The lowest number of substitutions in the OT-BS as compared with other four pattern types is observed in 18 participants. The Friedman test shows that the difference between the four conditions is significant ($\chi^2 = 36.703$, $p = 0.00$). Again, the difference between OT-BS and all three other types is reliable (OT-BS vs. BO-BS: $z = -3.837$, $p = 0.00$; OT-BS vs. BO-SU: $z = -3.665$, $p = 0.00$; OT-BS vs. BO-BU: $z = -3.924$, $p = 0.00$). The result confirms the findings that the position of segments in a word has a greater influence on segmental interactions than stress. Contrary to the reading condition, the difference between BO-BU and the remaining two conditions in recall is significant at the level of 0.05 (BO-BU vs. BO-BS: $z = -2.079$, $p = 0.04$; BO-BU vs. BO-SU: $z = -2.268$, $p = 0.02$). Considering that only in BO-BU both target consonants were in the unstressed syllable, this result suggests that substitutions of two segments are more likely to occur if they are in the unstressed syllable than if they are in the stressed one, as far as recall is concerned. This result is not consistent with similar research studies in English [3, 5, 11, 12, 13]. Finally, the number of TC substitutions in BO-SU and BO-BS was not significantly different in recall (BO-SU vs. BO-BS: $z = -0.907$ $p = 0.36$), and in reading as well. This suggests that sharing a position in a stressed syllable does not significantly increase the possibility of mutual segmental interactions.

4. DISCUSSION

One aim of this study was to investigate the role of stress in mutual substitutions of target consonants in Croatian, using the tongue twisters in two conditions – reading and recall. The results show that in both conditions the highest number of TC substitutions occurs when both TCs are in the unstressed syllable (BO-BU). In the recall condition the difference between the number of TC substitutions in BO-BU and BO-BS (where both TCs are in the stressed syllable at the same position – word-onset), is significant. These results are not in agreement with the literature on English speech errors, which reported that segments interact more often if they are in a stressed than in an unstressed syllable [3, 5, 11, 12, 13]. The results are not completely in line with Berg's [2] finding either, according to which there is an equal possibility that errors occur in a stressed or unstressed syllable in Spanish. Still, as the difference in reading between BO-BS and BO-BU is not significant, Berg's [2] finding is confirmed in this condition. Altogether, the present results suggest that prosodic features of language could influence phonological encoding and that this process is not universal. It is possible that in pitch-accent or tone languages less attention is concentrated to segmental planning in unstressed than in stressed syllables, due to the additional demands of pitch processing in stressed syllables. This lack of attention will result in more errors in unstressed syllables. Another aim was to check whether two segments interact more frequently if they share the position within a word than if they do not. The results show that the fewest TC substitutions occur if target consonants do not share the word position (OT-BS), as compared with the results in other three pattern types, in which TCs share their word position. Furthermore, the difference in number of TC substitutions between OT-BS and other patterns is significant. These results are in line with [13], which revealed that sharing the position within a word increases the possibility of segmental interaction in English. The obtained results seem to support the Shattuck-Hufnagel's [13] claim that word structure plays an active role in phonological encoding, not just in English but in Croatian as well. Moreover, it seems that a word structure has a greater influence on segmental interactions than a syllable structure. The evidence for this claim comes from the fact that significantly more TC substitutions occur in BO-BS than in OT-BS. It should be remembered that in both BO-BS and in OT-BS pattern, TCs shared the syllable position, but only in BO-BS they shared the word-position, not in OT-BS. Regarding the results in reading vs. recall, a greater number of TC substitutions in recall than in reading is expected and in line with [13], who gave an acceptable explanation for this – in reading, the memory load is minimal, as opposed to recall where it is significantly higher, which will consequently result in more errors. Still, a similar error ratio in all four TC patterns in reading and writing indicates that errors in both conditions have the same source within the process of phonological planning, as [13] already noticed in the interpretation of her own results.

In summary, the present results suggest that Croatian, as prosodically different from English, may differently affect the processes of phonological encoding, although some aspects of these processes are common to both languages, for example the role of word structure.

5. ACKNOWLEDGEMENTS

Research infrastructure and research costs were supported by project *Coarticulation in Croatian speech: instrumental investigation (CROCO)* funded by Croatian Science Foundation and by institutional support (Faculty of Humanities and Social Sciences University of Zagreb) for project *Fluency in speech production*.

6. REFERENCES

[1] Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., Znika, M. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

[2] Berg, T. 1991. Phonological processing in a syllable-timed language with pre-final stress: Evidence from Spanish speech error data. *Language and Cognitive Processes* 6, 265–301.

[3] Boomer, D. S., Laver, J. D. 1968. Slips of the tongue. *British Journal of Disorders of Communication* 3, 1–12.

[4] Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93, 283–321.

[5] Fromkin, V.A. 1971. The non-anomalous nature of anomalous utterances. *Language* 47, 27–52.

[6] Horga, D., Liker, M. 2016. *Artikulacijska fonetika. Anatomija i fiziologija izgovora*. Ibis grafika.

[7] Landau E., Lončarić, M., Horga, D., Škarić, I. 1999. Croatian. In IPA (ed.), Handbook of the International Phonetic Association. Cambridge: Cambridge University Press, 66-69.

[8] Levelt, W. 1989. *From intention to articulation*. The MIT Press.

[9] MacKay, D. G. 1970. Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia* 8, 323–350.

[10] Martinović, B., Pletikos Olof, E., Vlašić Duić, J. 2021. *Naglasak na naglasku*. Sveučilište Jurja Dobrile u Puli.

[11] Nooteboom, S. G. 1969. The tongue slips into patterns. In: Sciarone, A., van Essen, A. J., van Raad A. A. (eds), *Nomen, Leyden Studies in Linguistics and Phonetics*. Mouton, 114–132.

[12] Nooteboom, S., Quené, H. 2015. Word-onsets and stress patterns: speech errors in a tongue-twister experiment. Proc. 18th ICPhS Glasgow.

[13] Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering. *Cognition* 42, 213–259.

[14] Wilshire, C. E. 1999. The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech* 42, 57–82.

CAN SEGMENTAL DETAIL INFLUENCE PROSODIC ANALYSIS? THE CASE OF CONTRACTIONS IN ENGLISH

Holger Mitterer[1,3], Sahyang Kim[2], Taheong Cho[3]

[1]University of Malta, [2]Hongik University, [3]Hanyang Institute for Phonetics & Cognitive Sciences of Language, Hanyang University
holger.mitterer@um.edu.mt, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

Prosodic structure is phonetically reflected not only on suprasegmental dimensions (e.g., f0, duration, and amplitude) but also on segmental dimensions (e.g., coarticulation, segmental strengthening or reduction). This leads to the question whether the listener makes use of the prosodic-structurally conditioned segmental detail to compute the prosodic structure of a given utterance. We generated small dialogues for an online mouse-tracking task to test whether participants could use focus-related prosodic information (pitch-accented or not) and segmental realization of *have* (full or contracted) in predicting an upcoming target referent in a sentence. Results showed that a pitch-accented full form of 'have' did facilitate the listener's prediction of the upcoming referent. But results did not show such an effect of segmental realization (full or contracted) in the absence of pitch accent on 'have', suggesting that segmental information may not always be exploited by listeners when strong suprasegmental cues are available associated with contrastive focus.

Keywords: prosody, perception, prosody-syntax interface, online mouse-tracking

1. INTRODUCTION

Prosodic structure influences not only phonetic realization of suprasegmental features (such as f0, duration and amplitude), but segmental realization as reflected, for example, in articulatory strengthening versus weakening, localized hyper- or hypo-articulation and coarticulatory variation [1]-[3]. Cho et al. [1], for instance, found less coarticulatory nasalization when the word was carrying a pitch accent (see also [4]–[6]). Similarly, Cho & Keating [7] reported that alveolar stops are articulatorily strengthened, leading to longer VOTs for unvoiced stops, in domain-initial position or in a pitch-accented condition. These studies indicate that segmental realization may carry prosodic weight possibly to serve as a cue to prosodic structure.

In speech comprehension, prosodic information is known to influence listeners' parsing of syntactic structure. For instance, Steinhauer et al. [8] showed that prosodic breaks can easily overrule the otherwise well supported minimal-attachment heuristic in sentence processing, while others have shown that pitch accents or their absence can lead listeners to expect new or given information [9]–[11]. In these studies, the focus was on the suprasegmental aspects that signal prosodic structure. For instance, Roettger & Franke [11] used small dialogues and the German *verum* (focus-related) accent (see below) to show that listeners use this accent to predict how a sentence will continue.

(1a) *Hat der Wuggy dann die Geige aufgesammelt?*
Has the wuggy then the violin pick-up?
Did the wuggy then pick up the violin?
(1b) *Der Wuggy HAT dann die Geige aufgesammelt.*
the wuggy has then the violin picked-up.
The wuggy then picked up the violin.

As indicated by uppercase HAT (Engl., 'to have') in (1b), an affirmative response to a yes/no question in German can be marked by a pitch accent on the conjugated verb [12]. Roettger & Franke [11] used a mouse tracking task and listeners had to move the cursor to the object that the "wuggy" had picked up. When there was a *verum* accent on the auxiliary, listeners turned the mouse towards the target even before that word was heard.

In this and other similar eye-tracking studies [9], [10], the prosodic information used was encoded in variation of f0 and duration. This gives rise to the question whether similar results can be obtained if prosodic information is encoded in segmental detail. The processing of prosodic and segmental information is often conceptualized as being independent, so that a prosodic analysis (carried out by a so-called 'Prosody Analyzer') occurs in parallel with a segmental analysis (see [14] for related discussion), sometimes even conceptualized as being divided between the two hemispheres [15]. A strong version of this approach may therefore postulate that segmental information would not influence sentence processing in a similar way as is modulated by prosodic information embedded in suprasegmental features. Alternatively, however, there is some evidence that allows us to predict similar effects of segmental detail on sentence processing. Scott & Cutler [16] showed that listeners can use flapping (or

its absence) to decide whether a noun should be attached to an ongoing phrase or not ([*The day we met*] *Anne* … vs. [*The day we met Anne*]...). More recently, Mitterer et al. [17] showed a similar effect in Maltese using the glottal stop, which is a phoneme in Maltese. These two studies, however, made use of offline measures. They cannot therefore be seen as direct evidence for the listener's use of segmental process on the fly at the early processing stage but may be attributed to a late integration of prosodic and segmental information that is also implied in the Prosody Analyzer account [13].

Two mouse-tracking experiments were carried out in this study to explore how segmental information of 'have' in English may be used online by listeners in predicting an upcoming referent in sentence. We used a similar mouse-tracking set up following [11], but we had to implement various adaptations for the online setting (e.g., where the mouse cursor speed and location cannot be influenced by a script[i]). Moreover, we added one more condition in which the prosodic *weight* on the auxiliary (i.e., to what extent the production of the auxiliary 'have' can carry information of prosodic structure) was exclusively coded in segmental terms with no further bottom-up support of suprasegmental features for an assumed prosodic structure. In an example below, (2a) is the question and (2b-d) are three possible answers.

(2a) The aliens haven't shot the robot, right?
(2b) Well, they HAVE shot the robot. (*full, pitch accent*)
(2c) Well, they've shot the violin. (*reduced*)
(2d) Well, they have shot the violin (*full, no pitch accent*)

While participants heard a dialogue with one of the three answers (2b-d), they were presented with pictures of two objects (e.g., a violin and a robot) located left and right at the top of the screen. As in a simple computer game, the objects were falling down from the top with some left-right wriggling, and there was a picture of a spaceship which can shoot the falling objects. Participants could move the spaceship left and right using the mouse and were instructed to press a mouse button to fire a shot at a falling object in accordance as indicated in the answer sentence.

For (2b), the pitch-accent on *HAVE* should be interpreted as being contrastive to the *haven't* in the question, which is likely to lead listeners to predict that the upcoming object noun phrase would be the same as the one given in the question, thus moving the spaceship more quickly towards the given object. On the other hand, the absence of pitch accent as in (2c-d) is likely to indicate that an upcoming object is different from what has already been given in (2a), thus leading listeners to move towards the new object. One could, however, further expect the segmental difference between the reduced form (2c) and the full form (2d). Given that the full form is consistent with the accent-induced segmental realization, one could hypothesize that it carries more prosodic weight towards focus-induced prominence than the reduced form, thus rendering listeners *less* likely to expect a new object in the upcoming part of the sentence.

We tested these predictions in two experiments using different control conditions. Exp. 1 used a neutral (or broad focus) question (*What has happened?*) as a control condition as well as the question which may induce a narrow focus in the answer as in (2a). In the narrow focus condition, the presence or absence of focus on '*have*' in the answer sentence should be matched with the pragmatically appropriate referent (i.e., a given referent for (2b), and a new reference for (2c,d)). This, however, has a potential drawback. While results of Exp. 1 may reflect the processing of prosody, they may also be influenced by the learning of contingencies in the experiment ('narrow' question + *HAVE* → given). Therefore, Exp. 2 used the narrow questions only, and the referent in the answer sentences were not always pragmatically matched with the question—i.e., both the 'new' and 'given' referents were targets on the screen regardless of the information structure of the dialogue. This incongruent mapping between prosody and information structure may lead participants to ignore the phonetic details of the answer sentences to be used for computing prosodic structure. Given no contingencies within the experiment that participants could learn to exploit, however, if participants showed any preference for a given object, it could be interpreted as being due to the processing of the prosodic information. That is, the two designs have complementary weaknesses, but together, should paint a more complete picture.

2. GENERAL METHOD

Native speakers of American English were recruited via the *prolific.co* platform (24 for Exp 1 and 36 for Exp 2). We constrained participants to be in the age range 18-40, living in the USA, having English as their first language and not have reported any language-related disorder in their sign-up to *prolific.co*. Participants could only participate in one of the experiments. We used a larger sample in Exp 2 since it contained fewer trials.

We recorded two male American English speakers; one speaker provided the questions and the other the replies in the three versions as indicated in (2). Questions were edited to shorten the pause before the *right?*, since a long pause might be irritating in a repetitive experiment with many trials. For the answers, we cross-spliced the different parts of the sentence, so that participants could not learn that a given pronunciation of *Well* would lead to preference

of a particular object. Splicing points were after *Well* and after *shot*. For each combination of the target and the condition, all experimental materials were cross-spliced so that four different versions were generated with a randomly selected precursor from the same condition, but always for a different target. The stimulus was presented in a random order. We used copyright-free images for the 16 target objects in a 200x200 pixels frame, which were used as the targets to be shot by the spaceship in the game.

Participants were instructed on the structure of the task and conducted 16 practice trials to familiarize themselves with the 16 objects. They then continued to complete 144 (Exp. 1, nine blocks with 16 objects) or 96 (Exp. 2, six blocks with 16 objects) experimental trials. In Exp. 1, each target was presented two times in each of the three experimental conditions with the 'narrow' question (***HAVE*** with pitch accent; ***have*** *or* ***'ve*** with no pitch accent) and three times in the control condition (with the question *What has happened?*) with each version of the answer used once. In Exp. 2, each target was presented six times, once in each cell of the design by which the target type ('given' or 'new') was crossed with the three versions of the answer.

Similar to eye-tracking data, we investigated how close to the target the mouse position was in the horizontal dimension. For each trial, the screen coordinates were re-scaled to the interval [-1,1], with zero being the middle of the screen. The target side was mapped onto the positive side. Co-ordinates above .6 on either side were set to .6, so that the mouse was positioned to shoot the target or the competitor in a corridor within which the target wriggled left and right. Since we were interested in predictive processing, we analysed the mean mouse position in a time window from -200ms to 200ms around the target onset, a range which cannot be influenced yet by the pronunciation of the target word itself. The data were analysed with linear mixed-effects models (LME) that used the maximally converging random-effects structure and contrast coded predictors. Degrees of freedom were estimated using the lmerTest [17] package in R.

3. RESULTS

3.1 Experiment 1

Fig. 1 shows the results of Exp. 1. Panel A shows that participants move towards an expected target (i.e., towards a 'given' target in the pitch-accented HAVE condition or a 'new' target in the unaccented condition) well before the onset in all conditions if the question is narrow. But they wait on the target onset in the control 'broad' condition (*What has happened*?) in which the mouse moving towards the target starts around 200ms after the target onset. Panel B shows that this pattern arises only after the first block.

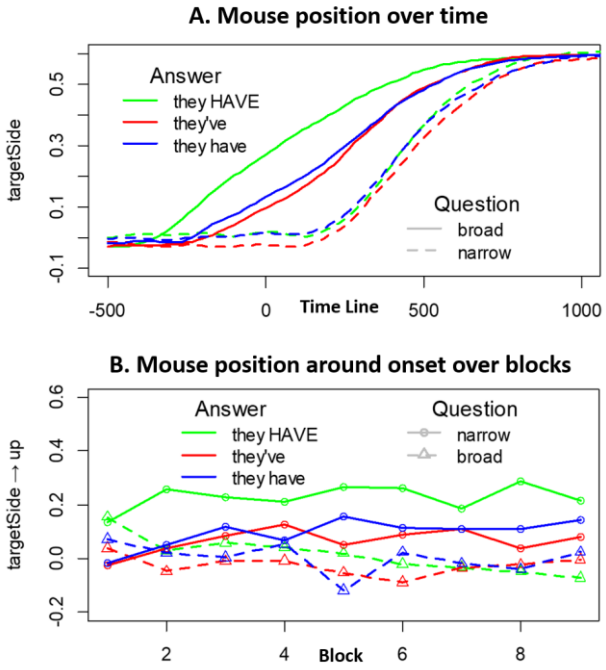


Fig. 1: Results from Experiment 1. Panel A shows the mouse position relative to the target over time; Panel B shows the average position in a -200 – 200ms time window around target onset over the course of the experiment.

The LME used contrast-coded predictors for Question (-0.5 = broad, 0.5 = narrow), Block (ranging from -4 till 4) and Answer with two contrasts, one for the presence of accent (*HAVE* vs. others) and one for the use of contraction (*they've* vs. *they have*). The analysis revealed that participants were significantly closer to the target with the 'narrow' question (b=0.137, $t(23.61)=5.53$ p<.001) and this effect became larger over blocks (b = 0.021, $t(28) = 4.14$, p<.001) and was larger for the answer with a pitch accent than not (b = 0.135, $t(24)=3.181$, p=.004). Participants also moved the mouse closer to the target when there was an accent in the answer (b=0.094, $t(22) = 4.554$. p<.001), which again increased over blocks (b=0.01, $t(2994)=1.994$, p=0.046). However, there were no significant differences between the two no-pitch-accent conditions that differed in segmental detail (full vs. reduced/contracted).

3.2 Experiment 2

In Exp. 2, mouse movements were analysed based on whether the target object was 'given' or 'new' (regardless of whether or not it was supported by pitch accent) and the form of the answer with the same three levels as in Exp. 1. Fig. 2 shows the mouse position relative to the target side. Panel A shows that participants move the mouse towards the 'given' target when supported by pitch accent ('HAVE'). Likewise, when the target is a 'new' object

(mismatched with pitch accent), they still move the mouse towards the 'given' object (the competitor) as can be predicted by the pitch-accented 'HAVE.' Additionally, as the solid lines in Fig. 2 indicate, at the beginning participants tend to move the mouse closer to the 'given' target in all conditions.

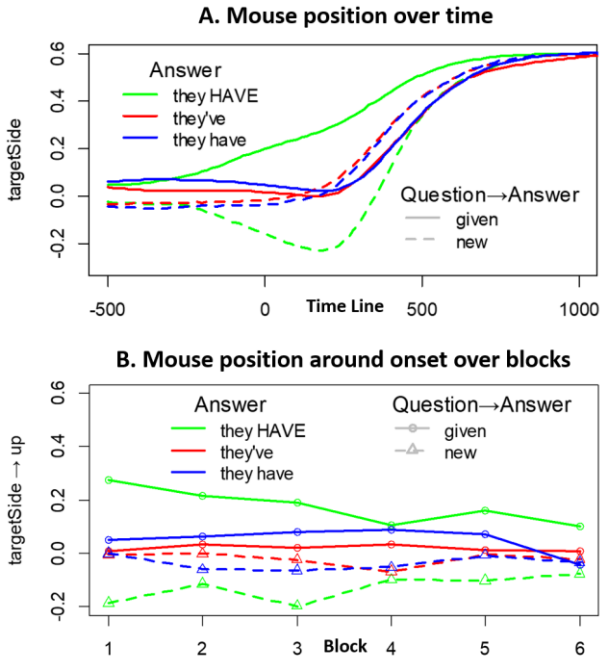


Fig. 2: Results from Experiment 2. Panel A shows the mouse position relative to the target over time, while Panel B shows the average position in a time window from -200 to 200ms around target onset over the six blocks.

The statistical analysis generally bore out these observations. There was a significant effect of Object Type ('given' vs. 'new', b = 0.132, t(33) = 6.00, p < .001) which became smaller over blocks (b = -0.019, t(35) = 2.45, p= .019). An interaction of Pitch Accent with Object Type (given/new) (b=0.221, t(35) = 5.34, p<.001) reflects the preference for the 'given' object only in the pitch-accented condition. This interaction is further supported by an interaction with block (b = -0.04, t(2894)=-3.12, p=.002), again reflecting that the preference for the given object in the pitch-accented condition gets smaller over the course of the experiment. It appears that listeners learned to overcome the mismatched (inappropriate) prosody and the expected referent (i.e., 'HAVE' followed by a 'new' target or a contracted form followed by a 'given' target). Importantly, as Fig. 2B shows, there was still a clear preference for the 'given' object in the 'HAVE' condition even in the last block (solid green line), indicating that the listeners predict the upcoming referent based on the prosody information. As in Exp. 1, there were no clear differences between the two conditions differing by segmental detail.

4. DISCUSSION

Our results first demonstrate that our online mouse tracking method gave rise to anticipatory response patterns with mouse movements towards the target well before target onset. In Exp. 1, participants performed as expected with earlier moves towards a 'given' object when 'have' was pitch-accented, but towards a 'new' object when 'have' was unaccented. Our results also showed that participants adapted to the prosodic contingencies as they were able to predict the target better in later blocks during the experiment, even in conditions with a-priori uninformative prosodic cues. In Exp. 2, when prosody and the expected referent were mixed (matched or mismatched), we found a robust and consistent predictive effect of the presence of pitch accent prosody (but not its absence) on 'have', showing a strong bias towards a 'given' object in all conditions despite the fact that a half of the stimulus sentences could contain a prosodic incongruence (e.g., a pitch accented 'HAVE' followed by a 'new' object). The fact that only the effect of the presence of an accent clearly survives further suggests the necessity to distinguish adaptation to prosody from adaptation to the experimental situation.

Crucially, however, we failed to find any evidence that listeners can use prosodically-conditioned segmental detail (reduced vs. full forms of *have*) as a cue to prosodic structure in the same way that they use suprasegmental cues that are embedded in our pitch-accented condition. In fact, suprasegmental cues remained important even when they were uninformative within the experimental context (as in Exp. 2). This primacy of suprasegmental prosodic cues in sentence processing may help understand why listeners failed to show a segmental effect. Given that the prosodic information was so salient in the pitch-accented 'HAVE' condition in the current experimental settings, the lack of pitch accent could have led the listeners to treat two unaccented variants with the equal prosodic weight without paying attention to the segmental detail.

In conclusion, our results confirm that sentence processing is modulated by prosodic structure that is computed primarily by suprasegmental cues driven by narrow focus. The results also indicate that segmental information alone may not provide robust perceptual bottom-up support to prosodic structure in the absence of suprasegmental cues, making it difficult to inform how segmental and prosodic information may be integrated in the time course of speech perception. It remains to be seen whether and how segmental information may be exploited by listeners in an experimental setting when other prosodic information is weaker.

5. ACKNOWLEDEMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884).

6. REFERENCES

[1] T. Cho, D. Kim, and S. Kim, 'Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English', *J. Phon.*, vol. 64, pp. 71–89, Sep. 2017, doi: 10.1016/j.wocn.2016.12.003.

[2] P. Keating, T. Cho, C. Fougeron and C-S. Hsu, 'Domain-initial strengthening in four languages', In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI*, (pp.143-161) et al., 2003.

[3] K. de Jong, 'The suprasegmental articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. of the Acous. Soc. of Am.*, 97, 491-504, 1995.

[4] J. Jang, S. Kim, and T. Cho, 'Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences', *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. EL33–EL39, Jul. 2018, doi: 10.1121/1.5044641.

[5] H. Li, S. Kim, and T. Cho, 'Prosodic structurally conditioned variation of coarticulatory vowel nasalization in Mandarin Chinese: Its language specificity and cross-linguistic generalizability', *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. EL240–EL246, Sep. 2020, doi: 10.1121/10.0001743.

[6] J. Jang, S. Kim, and T. Cho, 'Prosodic Structural Effects on Non-Contrastive Coarticulatory Vowel Nasalization in L2 English by Korean Learners', *Lang. Speech*, p. 00238309221108657, Jul. 2022, doi: 10.1177/00238309221108657.

[7] T. Cho, and P. Keating, Effects of initial position versus prominence in English, J. of Phon, 466-485, 2009.

[8] K. Steinhauer, K. Alter, and A. D. Friederici, 'Brain potentials indicate immediate use of prosodic cues in natural speech processing', *Nat. Neurosci.*, vol. 2, no. 2, pp. 191–196, 1999, doi: 10.1038/5757.

[9] D. Dahan, M. K. Tanenhaus, and C. G. Chambers, 'Accent and reference resolution in spoken-language comprehension', *J. Mem. Lang.*, vol. 47, no. 2, pp. 292–314, Aug. 2002, doi: 10.1016/S0749-596X(02)00001-3.

[10] A. Weber, B. Braun, and M. W. Crocker, 'Finding referents in time: Eye-tracking evidence for the role of contrastive accents', *Lang. Speech*, vol. 49, no. 3, pp. 367–392, Sep. 2006, doi: 10.1177/00238309060490030301.

[11] T. B. Roettger and M. Franke, 'Evidential Strength of Intonational Cues and Rational Adaptation to (Un-)Reliable Intonation', *Cogn. Sci.*, vol. 43, no. 7, p. e12745, 2019, doi: 10.1111/cogs.12745.

[12] G. Turco, B. Braun, and C. Dimroth, 'When contrasting polarity, the Dutch use particles, Germans intonation', *J. Pragmat.*, vol. 62, pp. 94–106, Feb. 2014, doi: 10.1016/j.pragma.2013.09.020.

[13] K. B. Shatzman and J. M. McQueen, 'Prosodic knowledge affects the recognition of newly-acquired words', *Psychol. Sci.*, vol. 17, pp. 372–377, 2006, doi: 10.1111/j.1467-9280.2006.01714.x.

[14] J. Steffman, S. Kim, and S-A. Jun, 'Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean Accentual Phrase', *J. Phon*, vol. 94, pp.1-13, 2022.

[15] G. Hickok and D. Poeppel, 'The cortical organization of speech processing', *Nat. Rev. Neurosci.*, vol. 8, no. 5, pp. 393–402, May 2007, doi: 10.1038/nrn2113.

[16] D. R. Scott and A. Cutler, 'Segmental phonology and the perception of syntactic structure', *J. Verbal Learn. Verbal Behav.*, vol. 23, no. 4, pp. 450–466, Aug. 1984, doi: 10.1016/S0022-5371(84)90291-3.

[17] H. Mitterer, S. Kim, and T. Cho, 'The Role of Segmental Information in Syntactic Processing Through the Syntax–Prosody Interface', *Lang. Speech*, vol. 64, no. 4, pp. 962–979, Dec. 2021, doi: 10.1177/0023830920974401.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, 'Package "lmerTest"', *R Package Version*, vol. 2, no. 0, 2015.

[19] M. S. Slim and R. J. Hartsuiker, 'Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js', *Behav. Res. Methods*, Nov. 2022, doi: 10.3758/s13428-022-01989-z.

[20] A. P. Salverda, D. Kleinschmidt, and M. K. Tanenhaus, 'Immediate effects of anticipatory coarticulation in spoken-word recognition', *J. Mem. Lang.*, vol. 71, no. 1, pp. 145–163, Feb. 2014, doi: 10.1016/j.jml.2013.11.002.

[21] F. Eisner and J. M. McQueen, 'Perceptual learning in speech: Stability over time', *J. Acoust. Soc. Am.*, vol. 119, no. 4, p. 1950, 2006, doi: 10.1121/1.2178721.

[i] If mouse control was possible in JavaScript, websites could force the mouse cursor to hoover over ads and lower the speed with which the ad could be left.

LOGARITHMIC DURATIONS FOR CLASSIFYING AND PREDICTING JAPANESE VOICED AND VOICELESS PLOSIVES

Kimiko Yamakawa[1], Shigeaki Amano[2], Mariko Kondo [3]

[1]Shokei University, [2]Aichi Shukutoku University, [3]Waseda University
[1]jin@shokei-gakuen.ac.jp, [2]psy@asu.aasa.ac.jp, [3]mkondo@waseda.jp

ABSTRACT

Conventionally, the duration of a phoneme is represented to be linear (raw duration). However, studies have reported that as an acoustic parameter of Japanese singleton/geminate consonants or short/long vowels, the logarithmic duration is better than the raw duration. The durational characteristics of Japanese voiced and voiceless plosives were analyzed to investigate the effectiveness of the logarithmic duration. The results indicated that the logarithmic duration is better than or comparable to the raw duration as an acoustic parameter for the classification and prediction of plosives at various speaking rates. This phenomenon suggests that the logarithmic duration provides a relational invariant acoustic parameter that can cope with durational variations caused by speaking rates.

Keywords: Japanese, voiced and voiceless plosives, category boundary, logarithmic duration

1. INTRODUCTION

Phoneme sounds exhibit three fundamental acoustic properties, namely intensity, frequency, and duration. Intensity is typically represented in logarithms (i.e., decibels, ISO 226:2003). Logarithm (e.g., the Mel scale [1]) or linear representations are used for frequency. Duration is conventionally represented as linear (raw duration). For example, Amano and Hirata [2–3] used the raw durations of speech segments to classify Japanese singleton and geminate stops. Pickett et al. [4] used a raw duration to contrast Italian singleton and geminate consonants. Miller et al. [5] used a raw duration for voice onset time (VOT; i.e., time from burst to voicing segment) as an acoustic parameter of English /p/ and /b/.

In contrast to the previously mentioned studies, Amano et al. [6] used the logarithmic duration as an acoustic parameter for Japanese singleton and geminate consonants. Specifically, they used the logarithmic duration of /k/'s closure and /s/'s frication as primary variables. Furthermore, they used the logarithmic average mora duration as a secondary variable related to the speaking rate. They found that compared with the raw duration, the combination of logarithmic durations provided a higher coefficient of determination (R^2) in the prediction of singleton and geminate consonants and almost the same discriminant error in the classification of these consonants. They concluded that the logarithmic duration is a better acoustic parameter for consonants than the raw duration.

Amano et al. [7] analyzed the effectiveness of the logarithmic duration in the classification and prediction of Japanese short and long vowels in words and non-words. The results of the study revealed that compared with the raw duration, the logarithmic duration provided a smaller discriminant error in the classification of vowels in non-words. Moreover, logarithmic durations provided a smaller coefficient of variation for the distance of the vowel from the category boundary. Furthermore, they revealed that logarithmic durations provided a simple and compact representation of the vowel distribution in terms of the root mean square error (RMSE) from the regression line. The results indicated that the logarithmic duration exhibits considerable advantage over the raw duration as an acoustic parameter for short and long vowels.

Although the effectiveness of the logarithmic duration has been studied previously [6–7], their effectiveness is to be confirmed for other phonemes. The positive findings for the logarithmic duration could be an exceptional case only for singleton and geminate consonants or short and long vowels. Therefore, confirming the general applicability of the logarithmic duration to other phonemes is necessary. Thus, this study examined the effectiveness of the logarithmic duration in Japanese voiceless and voiced plosives.

Voiceless and voiced plosives can be contrasted with VOT [e.g., 5]. However, VOT is not always measurable because a burst does not exist in some cases, such as a plosive at a very high speaking rate. Instead of uncertain VOT, this study used the entire duration of the plosives, namely the duration from the end of the preceding vowel to the beginning of the following vowel because the duration is always measurable. Consistent with previous studies [6–7], this study examined target phonemes across various speaking rates. The logarithmic durations can cope with the speaking rate and provide better classification and prediction of voiceless and voiced

plosives compared to raw durations.

2. SPEECH RECORDINGS

2.1 Speakers

Nine native Japanese speakers (three men and six women) participated in the recording. The average age was 21.0 years (Min. = 19, Max. = 25, SD = 1.8 years). They spoke standard Japanese and were from the Tokyo metropolitan area and its suburbs. The speakers had no symptoms of speech disorders and were paid to participate in the recording.

2.2 Equipment

Recordings were conducted in a soundproof room at Waseda University in Tokyo, Japan, using a microphone (Rode, NT-2A) with a pop screen (Stedman, PROSCREEN 101) connected to a computer (Panasonic, CR-RZ4, or CF-FV1SRCQP) through an audio interface (Roland, Rubix 24).

2.3 Target consonants

Target consonants were voiceless (/p/, /t/, /k/) and voiced (/b/, /d/, /g/) plosives in three-mora-long non-words (Table 1). The plosives were located at the beginning of the second mora in the non-words. The vowel of the second mora was /a/. /e/, or /o/. The first and third moras were /na/. With these phoneme sequences, vowels could not be devoiced.

2.3 Procedure

The item was embedded in a carrier sentence /koreka __ tosuru/ (Let it this or __) and presented at the top of the computer screen in hiragana (Japanese phonological script). A silent video of flashing lights on a digital metronome (Seiko, SQ200) was displayed at the center of the computer screen to indicate the speaking rate.

Voiceless	Voiced
/napana/	/nabana/
/napena/	/nabena/
/napona/	/nabona/
/natana/	/nadana/
/natena/	/nadena/
/natana/	/nadana/
/nakana/	/nagana/
/nakena/	/nagena/
/nakana/	/nagana/

Table 1: Non-word items having a target plosive at the second mora.

Beats per minute (BPM)	Speaking rate (mora/s)
81	12.2
51	7.7
32	4.8
25	3.8

Table 2: Beats per minute (BPM) set on a digital metronome and the corresponding speaking rate.

Speakers were asked to pronounce the sentence with a non-word with a flat accent pattern (i.e., low-high-high pitch pattern) three times at the speaking rate indicated by flashing lights. Speaking rates are presented in Table 2. Half of the speakers recorded voiceless items first, and subsequently voiced items. The other half of the speakers were recorded in the reverse order.

3. ANALYSES

The duration of each phoneme in the speech data was measured in ms by inspecting the waveform and spectrogram.

3.1. Classification of plosives

Discriminant analyses were conducted to obtain the category boundary of voiced and voiceless plosives and their classification errors. The following discriminant model was used in the analyses:

(1) $f = a_0 + a_1 v_1 + a_2 v_2$

(2) $f = a_0 + a_1 \log v_1 + a_2 \log v_2$

where the dependent variable f is the label for voiced and voiceless plosives, a_0 to a_2 are discriminant coefficients, v_1 and v_2 are independent variables, v_2 is the plosive duration (ms), and v_1 is the average mora duration (ms) calculated by dividing the sentence duration by the number of morae in the sentence. In this calculation of the average mora duration, non-words with target voiced and voiceless plosives were excluded from the sentence. Equation (1) was based on the formulation of Amano and Hirata [4] in which an intercept and a variable related to the speaking rate were introduced to the discriminant model. To examine the effects of the logarithmic duration, (2) was introduced by replacing the raw durations in (1) with logarithmic durations. The category boundaries of the voiced and voiceless plosives were obtained by substituting 0 for f in (1) and (2).

3.2. Prediction of plosives

Regression analyses were conducted for voiced and voiceless plosives using the following models:

(3) $v_2 = b_0 + b_1 v_1$

(4) $\log v_2 = b_0 + b_1 \log v_1$

where v_2 is the plosive duration (ms), v_1 is the average mora duration (ms), and b_0 and b_1 are the regression coefficients. The coefficient of determination (R^2) was calculated to examine the goodness-of-fit of each regression. To investigate the extent of the plosive distribution, the RMSE from the regression line was calculated as a function of the average mora duration. For further distribution examination, the distance

from the category boundary to the plosive was calculated and its coefficient of variation (CV) was obtained as an index of the distance variance.

4. RESULTS

Fig. 1 displays the scattergrams of voiced and voiceless plosives on a coordinate plane of plosive duration and average mora duration, with the category boundary and regression lines displayed as solid and broken lines, respectively.

4.1. Classification of plosives

The raw and logarithmic durations did not exhibit any significant difference in the classification error of the voiced and voiceless plosives (raw 12.4% vs. logarithm 11.2%, $z = 0.82$, *ns*), /p/ and /b/ (raw 11.4% vs. logarithm 11.1%, $z = 0.12$, *ns*), /t/ and /d/ (raw 9.9% vs. logarithm 10.3%, $z = 0.16$, *ns*), and /k/ and /g/ (raw 14.0% vs. logarithm 10.8%, $z = 1.24$, *ns*). These results indicate that the logarithmic duration performs as well as raw duration in voiced and voiceless plosive classifications. The gradient of the category boundary ($-a_1/a_2$) was 0.19 and 0.57 for the raw and logarithmic duration, respectively.

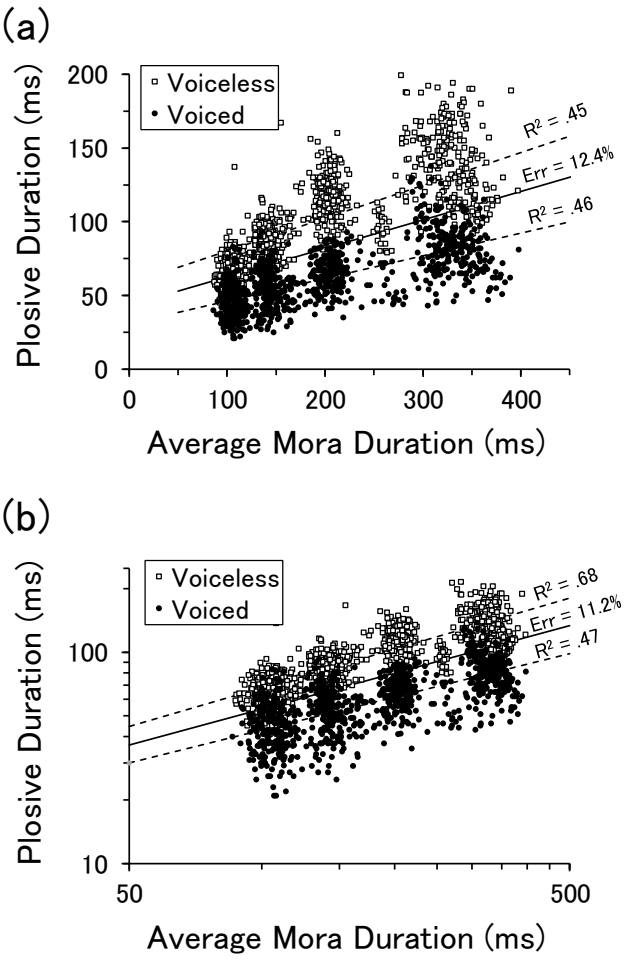


Figure 1: Scattergrams of voiced and voiceless plosives in (a) raw duration and (b) logarithmic duration. Solid and broken lines represent category boundaries and regression lines, respectively.

Type	Plosive	Duration		Difference
		Raw	Log	
Voiceless	All	0.45	0.68	$p < .001$
	/p/	0.66	0.72	*ns*
	/t/	0.64	0.72	$p < .05$
	/k/	0.27	0.62	$p < .001$
Voiced	All	0.46	0.47	*ns*
	/b/	0.52	0.53	*ns*
	/d/	0.45	0.45	*ns*
	/g/	0.47	0.51	*ns*

Table 3: Coefficient of determination (R^2) for voiced and voiceless plosives in raw and logarithmic durations.

4.2. Prediction of plosives

The gradients of the regression line (b_1) in raw duration were 0.15 and 0.22 for voiced and voiceless plosives, respectively. The b_1 in the logarithmic duration was 0.52 and 0.61 for voiced and voiceless plosives, respectively. These results indicated that the regression lines for both the raw and logarithmic durations may not be parallel to the category boundary.

However, the distribution in the logarithmic duration (Fig. 1b) appears to be more compact than the distribution in the raw duration (Fig. 1a). This tendency is confirmed by R^2 in Table 3. Logarithmic duration provided a significantly higher R^2 than raw duration for all voiceless plosives, /t/, and /k/. No

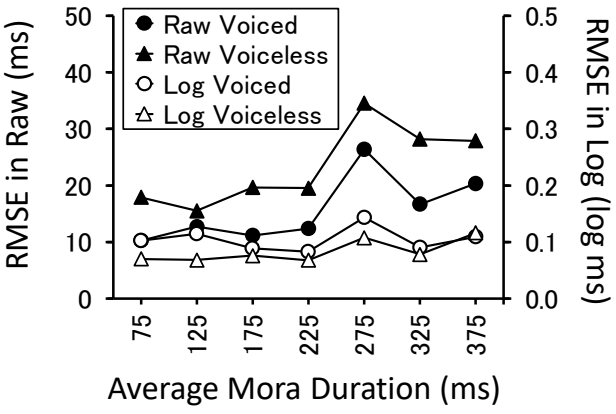


Figure 2: Root mean square errors (RMSE) from the regression lines for voiced and voiceless plosives. The left vertical axis is for RMSE of voiced plosives (●) and voiceless plosives (▲) for raw duration. The right vertical axis is for RMSE of voiced plosives (○) and voiceless plosives (△) for logarithmic duration. The values of the horizontal axis represent the mid-point of the bin for an interval of the mid-point ± 50ms. The intervals do not include the lower limit.

significant R^2 difference was observed between the logarithmic and raw durations for the /p/ and voiced plosives. However, R^2 tended to be slightly higher in the logarithmic duration than in the raw duration in these cases. These results indicate that the logarithmic duration is better than or equal to the raw duration in predicting plosives.

Fig. 2 illustrates the RMSE of the voiced and voiceless plosives from the regression line. The RMSE for the logarithmic duration was almost constant, whereas the RMSE for the raw duration tended to be higher as the average mora duration increased. The results indicated that the logarithmic duration provides a less spreading and more constant-width distribution than the raw duration against the speaking rate.

These characteristics were confirmed by the CV of the plosive distance from the category boundary (Table 4). All CVs were significantly smaller for the logarithmic duration than for the raw duration ($p < .001$). These results indicated that the logarithmic duration exhibited lower variation than the raw duration in terms of distance from the category boundary.

5. DISCUSSION

The results of the discriminant analysis of voiced and voiceless plosives indicate that the performance of the logarithmic duration is superior to that of the raw duration. The results of the regression analysis indicate that the logarithmic duration outperforms or is equal to the raw duration; moreover, the results of the RMSE and CV indicate that compared with the raw duration, the logarithmic duration is less affected by the speaking rate and provides a stable and compact distribution of plosives. The logarithmic duration is superior to the raw duration as an acoustic parameter of voiced and voiceless plosives in classification and prediction across various speaking rates.

However, the results revealed that the effectiveness of the logarithmic duration, classification, and regression were not perfect even if the logarithmic duration was used. The discriminant errors were 10.8%–11.2%, and the R^2s were 0.45–0.72. A possible cause of these not-low errors and not-high R^2s is that duration is not the only acoustic parameter of voiced and voiceless plosives. The plosives can be distinguished and predicted by other parameters, such as VOT and formant transition. Thus, plosives have multiple parameters, and their duration may not be their primary parameter, which may result in imperfect classification and regression. This differs from the case of singleton/geminate consonants [6] and short/long vowels [7]. In these consonants and vowels, duration is the primary parameter, and logarithmic duration outperforms raw duration.

Although the duration is not the primary parameter of plosives, the logarithmic duration is still effective in classifying and predicting them. This result indicates that the logarithmic duration is effective not only for singleton/geminate consonants [6] and short/long vowels [7] but also for plosives. This finding is critical because it is consistent with the claims of previous studies [6-7] regarding the effectiveness of the logarithmic duration.

Another significant aspect is that the findings of this study support relational acoustic invariance theory [4]. Consistent with previous studies [6-7], in this study, two durations, namely plosive duration and average mora duration, were used. The combination of these durations provided excellent classification and prediction. Thus, the acoustic invariance of phonemes against speaking rates was achieved in the relationship of the durations, as the theory claims. Furthermore, the results of this study suggest that relational acoustic invariance is provided by logarithmic rather than raw duration. To obtain further evidence of relational invariance theory and the effectiveness of the logarithmic duration, other phonemes in Japanese and other languages should be considered.

A limitation of this study was the use of read speech. Because read speech has a more regular rhythm than spontaneous speech, such regularity could affect the results of this study and provide an unexpected advantage to the logarithmic duration. Further verification using spontaneous speech is critical.

6. REFERENCES

[1] Stevens, S. S., Volkmann, J. 1940. The relation of pitch to frequency: A revised scale. *Am. J. Psychol.* 53(3), 329–353.

[2] Amano, S., Hirata, Y. 2010. Perception and production boundaries between single and geminate stops in Japanese. *J. Acoust. Soc. Am.* 128(4), 2049-2058.

[3] Amano, S., Hirata, Y. 2015. Perception and production of singleton and geminate stops in Japanese: Implications for the theory of acoustic invariance. *Phonetica* 72, 43–60.

[4] Pickett, E. R., Blumstein, S. E., Burton, M. W. 1999. Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica* 56, 135–157.

[5] Miller, J. L., Green, K. P., Reeves, A. 1986. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast, *Phonetica* 43(1-3), 106–115.

[6] Amano, S., Kondo, M., Yamakawa, K. 2021. Predicting and classifying Japanese singleton and geminate consonants using logarithmic duration. *J. Acoust. Soc. Am*. 150(3), 1830–1843.

[7] Amano, S., Hirata, Y., Yamakawa, K. (Accepted). Logarithmic durations for classifying and predicting Japanese short and long vowels. *Proc. ICPhS 2023.*

LOGARITHMIC DURATIONS FOR CLASSIFYING AND PREDICTING JAPANESE SHORT AND LONG VOWELS

Shigeaki Amano[1], Yukari Hirata[2], Kimiko Yamakawa [3]

[1]Aichi Shukutoku University, [2]Colgate University, [3]Shokei University
[1]psy@asu.aasa.ac.jp, [2]yhirata@colgate.edu, [3]jin@shokei-gakuen.ac.jp

ABSTRACT

Raw duration has conventionally been used as a time-related acoustic property to represent phonemic length distinction. However, a previous study reported evidence that logarithmic duration is better than raw duration for classifying Japanese singleton and geminate consonants accurately and predicting their distributions concisely across various speaking rates. To obtain further evidence for the effectiveness of logarithmic duration, this study analyzed the durational characteristics of Japanese short and long vowels. The results indicate that the logarithmic duration is better than the raw duration for the classification and prediction of these vowels. This suggests that logarithmic duration provides a relational invariant acoustic parameter that can cope with durational variations caused by speaking rates.

Keywords: Japanese, short and long vowels, category boundary, logarithmic duration

1. INTRODUCTION

Japanese has a phoneme distinction between short and long vowels [1]. For example, the word [kɯɾo] (meaning 'black') with a short vowel [ɯ] contrasts with the word [kɯːɾo] (meaning 'air route') with a long vowel [ɯː]. The duration of the periodic waveform of a vowel determines whether the vowel is short or long.

The Japanese language also has a phoneme distinction between singleton and geminate consonants [2]. For example, the word [ika] (meaning 'less than') with a singleton consonant [k] contrasts with the word [ikːa] (meaning 'family') with a geminate consonant [kː]. In this example, the closure duration of [k] determines whether the consonant is a singleton or a geminate.

However, vowel and closure durations covary with speaking rate, and they become shorter at faster speaking rates and longer at slower speaking rates [e.g., 3-4]. Therefore, their duration is not an invariant acoustic parameter corresponding to the phonological features of length. This is an instance of "the lack of invariance problem in phonemes."

Relational acoustic invariance theory [5] was proposed to address this problem. It claims that a relational measure, such as the ratio of consonant to vowel duration, provides an invariant property for phonological length distinction, and that this property is used to perceive of this distinction.

Following relational acoustic invariance theory, Amano and Hirata [4] analyzed the closure duration of singleton and geminate consonants in relation to the duration of other speech parts that varied with speaking rate. One of the parts they used was a subword defined as a mora sequence consisting of a preceding mora, closure of a stop, and following mora. Amano and Hirata [4] found that a linear function with closure and subword durations could classify singleton and geminate consonants with a small error. Based on these findings, they concluded that these two durational units provide invariant acoustic properties, which aligns with relational acoustic invariance theory.

Amano and Hirata [4] and other previous studies have conventionally used a linear duration (hereafter "raw duration") to represent the acoustic features of duration-sensitive phonemes. However, as a new approach, Amano et al. [6] introduced logarithmic duration to characterize the acoustic properties of singleton and geminate consonants. They compared the effectiveness of raw versus logarithmic durations for classifying and predicting singleton and geminate consonants across a wide range of speaking rates. Their results revealed that logarithmic duration performed better than raw duration in classifying and predicting the two consonants, which suggests that logarithmic duration provides relational invariant acoustic properties [e.g., 5] of singleton and geminate consonants.

Although Amano et al. [6] provided evidence of the effectiveness of logarithmic duration, it has only been used for singleton and geminate consonants, and their findings on logarithmic duration might be an exceptional case for singleton and geminate consonants. Subsequently, it is necessary to examine the general applicability of logarithmic duration to other duration-sensitive phonemes. Therefore, this study examines the effectiveness of logarithmic duration using Japanese short and long vowels pronounced at various speaking rates.

2. SPEECH DATA

The original speech data obtained by Hirata [3] were used for the analysis in this study. In Hirata [3], four native Japanese speakers (two males and two females) pronounced Japanese word and nonword sets three times at fast, normal, and slow speaking rates, each including contrasting short and long vowels.

The nonword set consisted of three types of /m/ containing sequence: /CV.CV/, /CVV.CV/, and /CV.CVV/, where the period "." represents a syllable boundary; C is /m/; V is one of /i/, /e/, /a/, /o/, and /u/; and the underline indicates a long vowel. Each sequence type contained 180 tokens. Pairing /CV.CV/ and /CVV.CV/ produced 180 short-long vowel contrasts in the first syllable. Pairing /CV.CV/ and /CV.CVV/ also produced 180 contrasts in the second syllable. These 360 pairs of contrasting short and long vowels (a total of 720 items) were used to analyze nonwords.

The word set consisted of 10 minimal pairs of Japanese disyllabic words containing short and long vowels in the first or second syllable, that is, the /CV.CV/-/CVV.CV/ pair or the /CV.CV/-/CV.CVV/ pair. C included various consonants; V included /i/, /e/, /a/, /o/ and /u/; and the underlines indicate contrasting short and long vowels. In total, there were 720-word tokens (20 words × three repetitions × three speaking rates × four speakers).

3. ANALYSES

The duration of each phoneme in the speech data was measured in milliseconds by inspecting the waveform and spectrogram.

3.1. Classification of vowels

Discriminant analyses were conducted to obtain a category boundary of short and long vowels and classification errors for words and nonwords. These analyses used the following discriminant model:

$$f = a_0 + a_1 v_1 + a_2 v_2 \quad (1)$$

$$f = a_0 + a_1 \log v_1 + a_2 \log v_2 \quad (2)$$

where the dependent variable f is the label for short and long vowels, a_0 *to* a_2 are discriminant coefficients, v_1 and v_2 are independent variables: v_2 is the vowel duration (ms), and v_1 is the average mora duration (ms) calculated by dividing the sentence duration by the number of morae in the sentence. In this calculation of average mora duration, the words with target short and long vowels were excluded from the sentence. Eq. 1 was based on the formulation of Amano and Hirata [4], in which an intercept and a variable related to the speaking rate were introduced to the discriminant model. To examine the effects of the logarithmic duration, Eq. 2 was introduced by replacing the raw durations in Eq. 1 with logarithmic durations.

3.2. Prediction of vowels

Regression analyses were conducted to obtain regression lines for short and long vowels using raw durations (dependent variable v_2 and independent variable v_1) and logarithmic durations (dependent variable $\log v_2$ and independent variable $\log v_1$), respectively. The coefficient of determination (R^2) was calculated to examine the goodness-of-fit of each regression. To investigate the extent of short and long vowel distributions as a function of speaking rate, root mean square errors (RMSE) from their regression lines were calculated as a function of the average mora duration. For further distribution examination, the distance from the category boundary to each data point was calculated, and the coefficient of variation (CV) was obtained for each distribution.

4. RESULTS

4.1. Classification of vowels

Fig. 1 illustrates scattergrams of short and long vowels on a coordinate plane of vowel duration and average mora duration, with category boundary and regression lines shown as solid and broken lines, respectively. Short and long vowels had a V-shaped distribution for the raw duration (Figs 1a and 1c), whereas they had an almost parallel distribution for the logarithmic duration (Figs 1b and 1d).

For nonwords, the error in short and long vowel classification by category boundary was significantly smaller ($z = 2.04$, $p < .05$) for the logarithmic duration (1.2% in Fig. 1d) than for the raw duration (3.5% in Fig. 1c). However, for words, no significant difference was observed in the classification error between logarithmic (5.1% in Fig. 1b) and raw (5.6% in Fig. 1a) durations. These results indicate that the logarithmic duration is equal to or better than the raw duration for short and long vowel classifications.

4.2. Prediction of vowels

As for the coefficient of determination (R^2), there was no significant difference between the raw and logarithmic durations for both nonwords (R^2 = .71–.77) and words (R^2 = .62–.77). The results indicate that logarithmic duration has equal performance to raw duration in the prediction of short and long vowels.

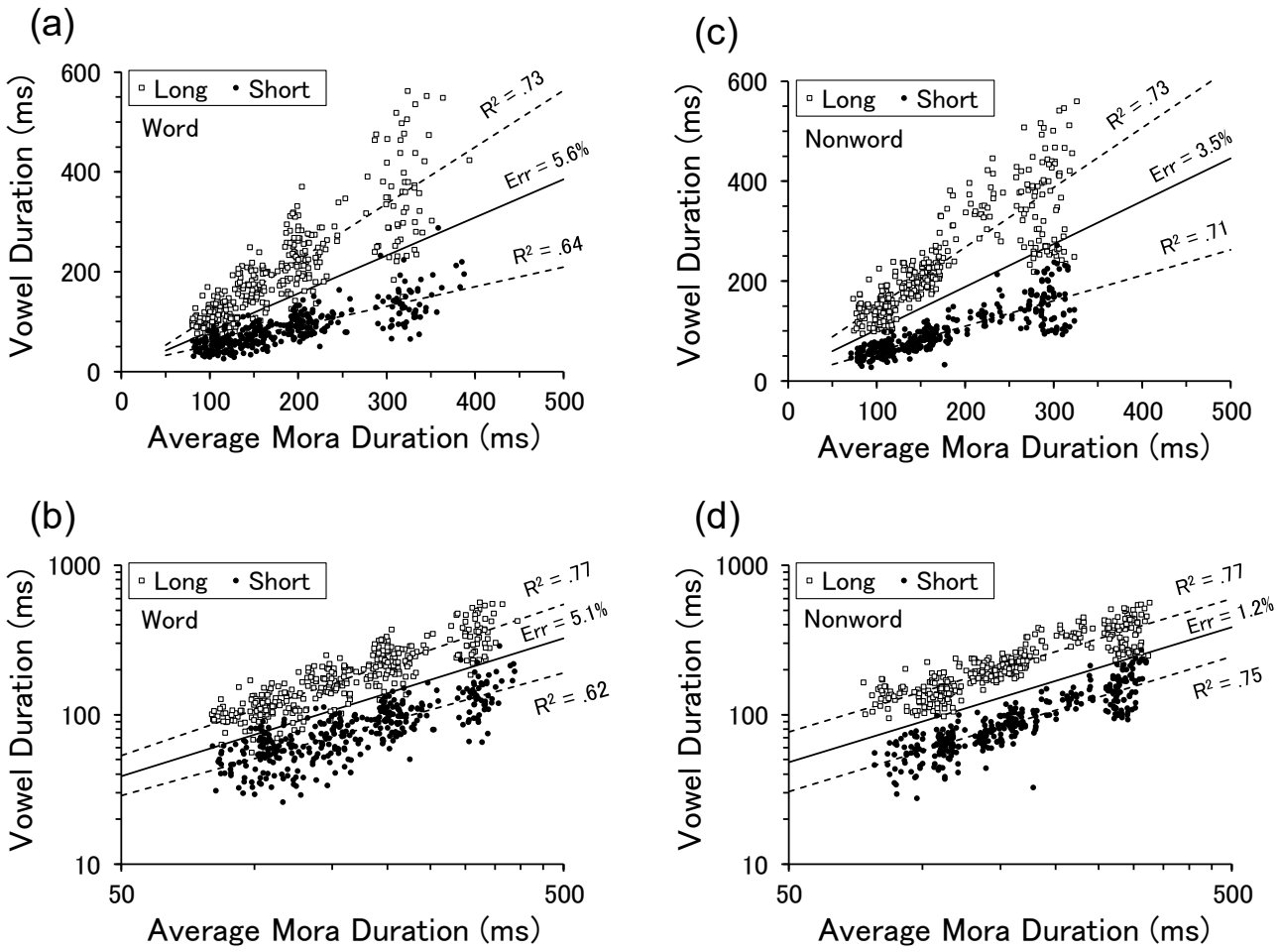


Figure 1: Scattergrams of short and long vowels in (a) words using raw duration, (b) words using logarithmic duration, (c) nonwords using raw duration, and (d) nonwords using logarithmic duration. Solid and broken lines represent category boundaries and regression lines, respectively.

Fig. 2 illustrates the RMSE from the regression line. The RMSE of short and long vowels for the logarithmic duration were almost constant, whereas the RMSE for the raw duration increased as the average mora duration increased. The results indicate that the logarithmic duration provides a less spreading and more compact distribution than the raw duration.

This tendency was also confirmed by the coefficient of variation in Table 1. The coefficient was significantly smaller for logarithmic duration than raw duration of short vowels in words [$F(359, 359) = 30.45$, $p = 6.12\times10^{-162}$], long vowels in words [$F(359, 359) = 34.45$, $p = 5.10\times10^{-174}$], short vowels in nonwords [$F(359, 359) = 46.91$, $p = 5.00\times10^{-197}$], and long vowels in nonwords [$F(359, 359) = 49.06$, $p = 2.09\times10^{-200}$]. These results indicate that logarithmic duration provides less variation than raw duration in terms of distance from the category boundary for both words and nonwords.

Item	Vowel	Duration	
		Raw	Logarithm
Word	Short	0.394	0.071
	Long	0.394	0.067
Nonword	Short	0.310	0.045
	Long	0.310	0.044

Table 1: Coefficient of variation of vowel's distance from the category boundary.

5. DISCUSSION

The results of the discriminant and regression analyses indicated that, compared to raw duration, logarithmic duration provided equal or better classification and prediction of short and long vowels (Fig. 1), and provided a simpler and more compact representation of vowels in terms of the RMSE from the regression line (Fig. 2) and distance from the category boundary (Table 1).

In some cases, the logarithmic and raw durations did not significantly differ in classification error and R^2. One possible reason for the non-significant difference is that this study reanalyzed Hirata [3]'s speech data pronounced by only four speakers at a not-so-wide range of speaking rates (2.5–14.1 mora/s), which might have unexpectedly resulted in a

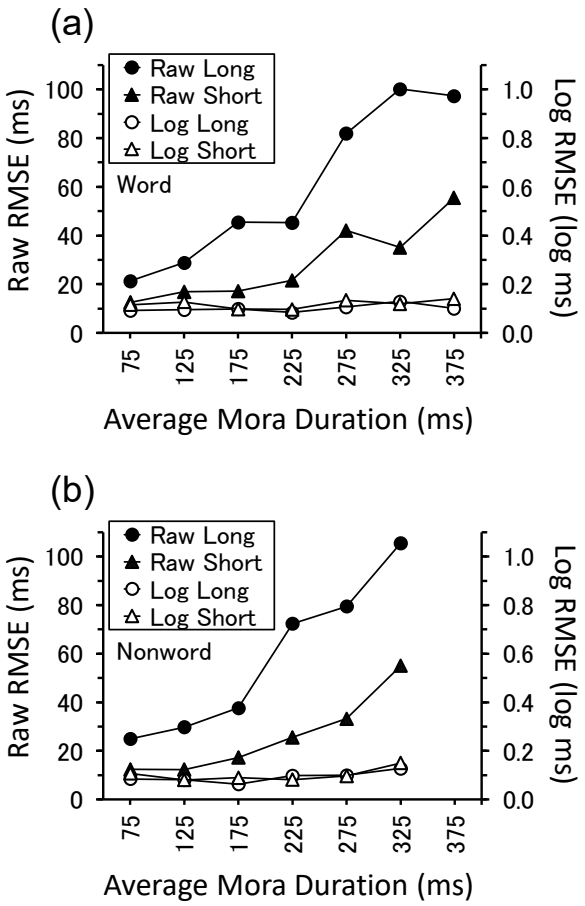


Figure 2: Root mean square errors (RMSE) from the regression lines for short and long vowels in (a) words and (b) nonwords. The left vertical axis is for RMSE of long vowels (●) and short vowels (▲) for raw duration. The right vertical axis is for RMSE of long vowels (○) and short vowels (△) for logarithmic duration. The values of the horizontal axis represent the mid-point of the bin for an interval of the mid-point ± 50ms. The intervals do not include the lower limit.

small variance and a clear separation of short and long vowel distributions. However, the insignificant difference does not indicate negative evidence for the logarithmic duration. Therefore, it was concluded that logarithmic duration has an advantage over raw duration in classifying and predicting short and long vowels.

The findings of this study are consistent with previous studies demonstrating that logarithmic duration is better for the classification and prediction of singleton and geminate consonants [6] and voiced and voiceless stop consonants [7]. Accumulating evidence suggests that logarithmic duration provides better acoustic variables for representing phoneme categories than raw duration which has been conventionally used in speech studies.

However, the evidence was obtained only for Japanese phonemes. Future research is needed in other languages to confirm the generality and universality of logarithmic duration. For example, research should examine languages with short and long vowel contrasts, such as Arabic [8], Estonian [9], and Thai [10], or languages with singleton-geminated stop contrasts, such as Bengali [11], Hungarian [12], and Italian [5].

Amano and Hirata's [4] findings supported relational acoustic invariance theory [5] based on their singleton-geminate study using raw duration. Amano et al. [6] also supported this theory, however, they claimed that logarithmic duration, not raw duration, provides an invariant acoustic parameter. The present study is in line with Amano et al. [6] because the combination of logarithmic durations accurately classified short and long vowels and concisely predicted their durational distributions across various speaking rates. This suggests that the phonological feature of length is well represented by a logarithmic duration relationship.

6. REFERENCES

[1] Vance, T. J. 2008. *The Sounds of Japanese*. Cambridge University Press.

[2] Kawahara, S. 2015. The phonetics of sokuon, or geminate obstruents. In: Kubozono, H. (ed), *Handbook of Japanese Phonetics and Phonology*. Mouton, 43–77.

[3] Hirata, Y. 2004. Effects of speaking rate on the vowel length distinction in Japanese. *J. Phon.* 32(4), 565–589.

[4] Amano, S., Hirata, Y. 2015. Perception and production of singleton and geminate stops in Japanese: Implications for the theory of acoustic invariance. *Phonetica* 72, 43–60.

[5] Pickett, E. R., Blumstein, S. E., Burton, M. W. 1999. Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica* 56, 135–157.

[6] Amano, S., Kondo, M., Yamakawa, K. 2021. Predicting and classifying Japanese singleton and geminate consonants using logarithmic duration. *J. Acoust. Soc. Am*. 150(3), 1830–1843.

[7] Yamakawa, K., Amano, S., Kondo, M. (Accepted). Logarithmic durations for classifying and predicting Japanese voiced and voiceless plosives. *Proc. ICPhS 2023*.

[8] Tsukada, K. 2009. An acoustic comparison of vowel length contrasts in Arabic, Japanese and Thai: Durational and spectral data. *Int. J. Asian Lang. Process.* 19(4), 127–138.

[9] Harms, R. T. 2006. Estonian. In: Brown, K. (ed), *Encyclopedia of Language & Linguistics* (2nd ed), Elsevier, 228.

[10] Abramson, A. S., Reo, N. 1990. Distinctive vowel length: duration vs. spectrum in Thai. *J. Phon.* 18(2), 79–92.

[11] Lahiri, A., Hankamer, J. 1988. The timing of geminate consonants. *J. Phon.* 16(3), 327–338.

[12] Vago, R. M. 2006. Hungarian: Phonology. In: Brown, K. (ed), *Encyclopedia of Language & Linguistics* (2nd ed), Elsevier, 433–440.

PHONOLOGICAL NETWORK PROPERTIES OF NONWORDS INFLUENCE THEIR LEARNABILITY

Rory Turnbull

Newcastle University
rory.turnbull@newcastle.ac.uk

ABSTRACT

It is well-established that a word's phonological neighbourhood is relevant for many aspects of phonetic processing. The current study investigated whether this is true of nonwords too. In the exposure phase, participants (*N*=198, English-speaking) were shown pairings of novel words and novel items. In the test phase, participants were presented with a word–item pairing and asked if the item was named correctly or not. Overall accuracy was high (79.2%). Nonwords were learned with higher accuracy, and responded to with faster reaction times, if they had more neighbours, and if their neighbours were more well-connected, relative to nonwords which had fewer neighbours and whose neighbours were less well-connected. These results demonstrate that existing words in the lexicon can exert an influence on the acquisition and processing of novel words.

Keywords: lexical effects, network structure, word learning, acquisition

1. INTRODUCTION

Several studies have shown that the phonological neighbours of a word—all the words that differ from that word only in terms of a single phoneme—can influence that word's processing, both in terms of perception and production [1]. This finding has been extended, to some extent, to the processing of nonwords too. For example, lexical decision reaction time task is modulated by the number of neighbours of nonword stimuli [2, 3], and nonword neighbourhoods can influence phoneme identification [4].

It is less clear, however, how neighbourhood properties affect the learnability of novel words. By definition, a word which has not yet been learned is functionally equivalent to a nonword. This paper examines how neighbourhood properties influence novel word learning.



Figure 1: Example novel item stimulus image.

2. METHOD

2.1. Materials

Wuggy [5] was used to generate several hundred British English nonwords, of which 47 were selected for use in the experiment. The stimuli were recorded by a native speaker of British English. Images of 48 novel items with low familiarity ratings were selected from the NOUN database [6]. An example image is provided in Figure 1. The nonword stimuli were as follows:

bɒsp	bju:f	bju:p	bju:ʃ	
bju:ʒ	bjʊə	bræb	bræmpt	bræv
ði:d	ði:pt	dru:ʤ	eɪf	fɜ:ʧ
flɔ:ʧ	fru:m	fwɑ:θ	gju:ʃ	glæg
glaɪb	gleɪm	grʌlpt	hju:f	hu:ls
jɜ:ft	klɒm	krɪnθ	kwɛmpt	pɜ:b
plɛsp	pwi:θ	skwaɪs	skwɒf	slædʒ
sləʊg	smɪd	spli:ʃ	splʌnd	stʌg
swɒsk	ʃnæm	ʃpi:g	traʊd	vju:t
zɛft	θrɒmps	θwæʃt		

2.2. Neighbourhood properties

Several recent studies conceive of the phonological lexicon as a complex network, where words, represented as nodes in a network, are connected

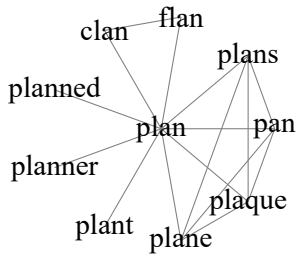


Figure 2: Example phonological neighbourhood network centred around the English word *plan*. Note that some neighbours of a word are neighbours of each other. Adapted from [10].

to their phonological neighbours [7, 8]. This yields a web, part of which is visualised in Figure 2. This approach allows for the examination of more nuanced measures than just the raw number of neighbours.

Four neighbourhood properties are considered in this study [9]. The **degree** of a node is the number of neighbours it is connected to. This is also referred to as "neighbourhood density". A node's **coreness** is a measure of how well-connected its neighbours are. The coreness of a node is the largest value of k for which the node is in the k-core of the network; the k-core of a network is the subset of network with only the nodes with degree of k or more. The **clustering coefficient** of a node is the proportion of the neighbours of a node which are neighbours of each other. **Closeness centrality** is a measure of how close a node is to others in the network, defined as the reciprocal of the sum of the distance from that node to all other nodes in the network.

Some stimulus words are **singletons**. These are nodes in the network with a degree of zero, that is, no neighbours. Singletons and non-singletons are analysed separately due to the fact that the neighbourhood properties for singletons are all either zero or undefined.

Each of these measures was calculated for each stimulus word, and also for the nonword's lexical neighbours, using the CELEX lexicon of UK English. As each stimulus varies in how many neighbours it has, the values for the neighbours were averaged. For example, the nonword /bɒsp/ has a degree of 4, and its neighbours have an average degree of 13.25. The python package Networkx was used to calculate these values [11].

These measures are partially correlated with each other. Some of this correlation is due to the definitions of the measures themselves—for example, coreness is intrinsically linked to degree—but some are due to properties of phonological lexicons in general. For example, there tends to be a correlation between the degree of a word and the degree of its neighbours—this 'assortativity by degree' has been noted to be a property of lexical organisation in natural language [12, 13]. Table 1 shows a correlation matrix of the network measures of the non-singleton stimuli.

Due to this collinearity, these measures are inappropriate to be used simultaneously in a regression analysis. A factor analysis was undertaken to reduce the dimensionality of the neighbourhood measures. Horn's parallel analysis returned two factors as optimal; factor loadings are reported in Table 2. Factor 1 corresponds largely to the coreness and degree of the nonword stimulus and the mean degree of its neighbours. Factor 2 corresponds largely to the closeness centrality of both the nonword stimulus and its neighbours.

2.3. Procedure

The experiment consisted of 6 blocks, each with an exposure phase and a test phase. In the exposure phase, a novel item was presented on screen and the item's name (a nonword stimulus) was presented auditorially. Eight item–word pairs were presented;

		Word			Neighbour			
		Coreness	Degree	Closeness	Clustering	Coreness	Degree	Closeness
Word	Clustering	0.571	0.459	0.384	-0.113	0.390	0.379	0.342
	Coreness		0.955	0.572	-0.119	0.686	0.740	0.435
	Degree			0.560	-0.144	0.631	0.681	0.398
	Closeness				0.068	0.713	0.692	0.922
Neighbour	Clustering					-0.084	-0.169	0.196
	Coreness						0.960	0.705
	Degree							0.659

Table 1: Correlation matrix of the network measures for the non-singleton stimuli and their neighbours.

		Factor 1	Factor 2
Word	Clustering	0.566	0.108
	Coreness	0.997	0.008
	Degree	0.957	-0.013
	Closeness	0.569	0.753
Neighbour	Clustering	-0.126	0.272
	Coreness	0.685	0.457
	Degree	0.739	0.381
	Closeness	0.429	0.900

Table 2: Factor loadings for the two factors returned from the factor analysis of the non-singleton neighbourhood measures.

each pair was presented four times, in random order. In the test phase, participants were again presented with an item–word pair and asked if the word was the item's correct name. Participants responded via keyboard, with no time limit on their responses. The experiment was constructed in Labvanced and participants were recruited via Prolific, and took roughly 10 minutes to complete.

2.4. Predictions

As singletons have no neighbours, they are less similar to existing words in the lexicon than non-singletons. Singletons should therefore be harder to integrate into the lexicon than non-singletons, and as such we should observe lower accuracy on these words.

For the other neighbourhood properties, extremely high or extremely low values are expected to inhibit learning. For example, if a new word has an extremely high clustering coefficient, it is being integrated into a very 'busy' neighbourhood. A very low clustering coefficient is a very 'sparse' neighbourhood and should be similarly challenging. Only words in the 'Goldilocks zone' should be learned easily. This prediction is necessarily vague as we do not currently have the empirical data to know what values of the neighbourhood properties count as "extremely high" or "extremely low".

2.5. Participants

The study was completed by 198 participants, all of whom were UK residents fluent in English. Most of the participants (191) were native speakers of English.[1] Participants were compensated GBP2.25 for their participation.

2.6. Analysis

Four regression models were constructed to model the data: a logistic mixed effects regression model to predict response accuracy in singleton versus non-singleton stimuli; a linear mixed effects regression model to predict reaction time (RT) in singleton versus non-singleton stimuli; a generalised additive mixed model (GAMM) [14] to predict accuracy from the neighbourhood factors; and a GAMM to predict RT using the neighbourhood factors. The RT models included data only from correctly-answered trials in which the picture and word matched.

The singleton versus non-singleton models included fixed effects of the stimuli's status as a singleton or not, the overall position of the trial in the experiment (trial number), and their interaction. Random intercepts of participant and word were included. The neighbourhood factors are not predicted to have necessarily linear effects, which motivated the use of GAMMs. Factors 1 and 2 were included as thin plate regression smooths. A parametric term of trial number was included. Random intercepts of participant and word were also included.

3. RESULTS

3.1. Singleton versus non-singleton nonwords

A speed–accuracy tradeoff was observed in the models comparing singleton to non-singleton words. As participants proceeded through the experiment, they responded overall faster ($t = -4.637$, $p < .001$) and less accurately ($z = -3.136$, $p = 0.002$). No other significant effects were observed: singleton words did not appear to be significantly harder or easier to learn than non-singleton words.

3.2. Neighbourhood properties of non-singleton words

The output of the GAMM predicting response accuracy is summarised in Table 3. Mirroring the results of the earlier analysis, accuracy decreased as trial number increased. A significant effect of Factor 1 was observed, such that stimuli with higher Factor 1 scores (higher coreness, higher degree, and higher-degree neighbours) were responded to more accurately than stimuli with lower Factor 1 scores. This effect is visualised in Figure 3.

The output of the GAMM predicting RT is summarised in Table 4. RT decreased with trial number, reflecting the speed–accuracy tradeoff mentioned in the previous section. Significant

Parametric coefficients:			
	Estimate	*t*	*p*
Intercept	2.483	17.934	< .001
Trial number	-0.019	-4.258	< .001
Approximate significance of smooth terms:			
	EDF	*F*	*p*
Factor 1	1	4.225	0.040
Factor 2	1	3.634	0.057

Table 3: Model output from the GAMM predicting response accuracy.

Parametric coefficients:			
	Estimate	*t*	*p*
Intercept	7.591	513.596	< .001
Trial number	-0.001	-4.128	< .001
Approximate significance of smooth terms:			
	EDF	*F*	*p*
Factor 1	1	13.514	< .001
Factor 2	1	4.244	0.039

Table 4: Model output from the GAMM predicting log reaction time.

effects of both Factor 1 and Factor 2 were observed, visualised in Figure 4. Words with higher Factor 1 scores (higher coreness, higher degree, and higher-degree neighbours) and higher Factor 2 scores (higher closeness centrality) were responded to faster than words with lower Factor 1 and 2 scores.

4. DISCUSSION

The results do not support the hypothesis that non-singleton words are learned better than singleton words: both singleton and non-singleton words had similar response accuracies and RTs. However, there were differences among the non-singleton words, governed by neighbourhood properties. Notably, words with higher degree or higher coreness, and which had higher-degree neighbours, were responded to faster and more accurately than words with lower values. Similarly, words with high closeness centrality (and high-centrality neighbours) were responded to more quickly than words with lower values.

These findings are consistent with a network-based model of the lexicon, where activation spreads from neighbour to neighbour [15, 16, 8]. The high-centrality words can be accessed faster than the low-

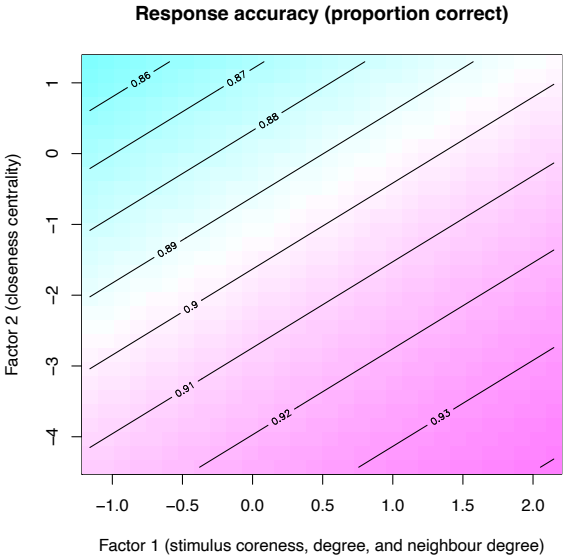


Figure 3: Effects of Factor 1 and Factor 2 on response accuracy.

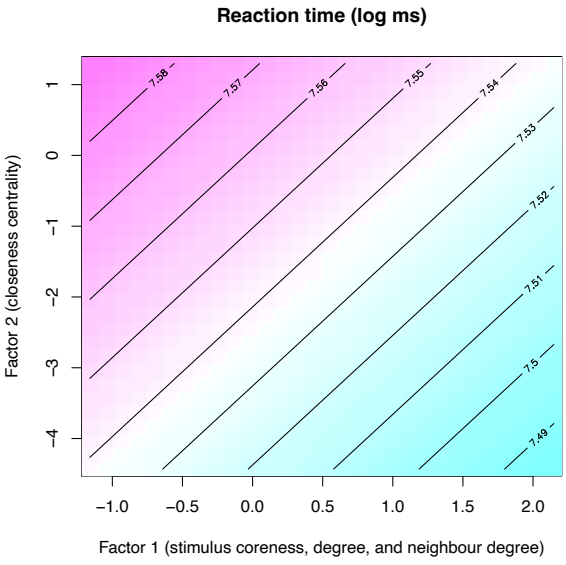


Figure 4: Effects of Factor 1 and Factor 2 on reaction time.

centrality words, as they are closer to the core of the lexicon and therefore more accessible. That neighbourhood properties affect (non)word learning is a novel finding. This study represents a step towards a better understanding of the kinds of lexical influences on word learning, and may provide insights on the lexical organisation of languages, first language acquisition, and second language acquisition.

ACKNOWLEDGEMENTS

I am grateful for comments and suggestions from the Newcastle University Phonetics & Phonology Research Group and two anonymous reviewers.

5. REFERENCES

[1] M. S. Vitevitch and P. A. Luce, "Phonological neighborhood effects in spoken word perception and production," *Annual Review of Linguistics*, vol. 2, no. 1, pp. 75–94, 2016.

[2] E. Janse, "Neighbourhood density effects in auditory non-word processing in aphasic listeners," *Clinical linguistics & phonetics*, vol. 23, no. 3, pp. 196–207, 2009.

[3] M. C. Kelley and B. V. Tucker, "The recognition of spoken pseudowords," *Language, Cognition and Neuroscience*, pp. 1–22, 2022.

[4] R. S. Newman, J. R. Sawusch, and P. A. Luce, "Lexical neighborhood effects in phonetic processing." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 3, p. 873, 1997.

[5] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior Research Methods*, vol. 42, no. 2, pp. 627–633, 2010.

[6] J. S. Horst and M. C. Hout, "The novel object and unusual name (noun) database: A collection of novel images for use in experimental research," *Behavior research methods*, vol. 48, no. 4, pp. 1393–1409, 2016.

[7] M. Stella and M. Brede, "Patterns in the English language: Phonological networks, percolation and assembly models," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 5, p. P05006, 2015.

[8] M. S. Vitevitch, "What can graph theory tell us about word learning and lexical retrieval?" *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 408–422, 2008.

[9] M. Newman, *Networks*, 2nd ed. Oxford, UK: Oxford University Press, 2018.

[10] R. Turnbull, "Graph-theoretic properties of the class of phonological neighbourhood networks," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Online: Association for Computational Linguistics, Jun. 2021, pp. 233–240. [Online]. Available: https://aclanthology.org/2021.cmcl-1.27

[11] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA, 2008, pp. 11–15.

[12] S. Arbesman, S. H. Strogatz, and M. S. Vitevitch, "The structure of phonological networks across multiple languages," *International Journal of Bifurcation and Chaos*, vol. 20, no. 3, pp. 679–685, 2010.

[13] R. Turnbull and S. Peperkamp, "What governs a language's lexicon? determining the organizing principles of phonological neighbourhood networks," in *Complex Networks & Their Applications V*, ser. Studies in Computational Intelligence, H. Cherifi, S. Gaito, W. Quattrociocchi, and A. Sala, Eds. Cham, Switzerland: Springer, 2017, vol. 693, pp. 83–94.

[14] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL: Chapman & Hall / CRC, 2017.

[15] M. Goldrick and S. E. Blumstein, "Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters," *Language and Cognitive Processes*, vol. 21, no. 6, pp. 649–683, 2006.

[16] M. Goldrick, H. R. Baker, A. Murphy, and M. Baese-Berk, "Interaction and representational integration: Evidence from speech errors," *Cognition*, vol. 121, pp. 58–72, 2011.

[1] Removal of the non-native speakers from analysis did not affect the interpretation of the results.

MECHANISMS OF GENERALIZATION FOR PHONETIC LEARNING OF ACCENTED SPEECH

Yevgeniy Vasilyevich Melguy and Keith Johnson

University of California, Berkeley
ymelguy@berkeley.edu

ABSTRACT

Listeners can rapidly adapt to an unfamiliar accent. Previous literature shows that such phonetic learning is often speaker-specific for speech sounds such as fricatives — it does not transfer to novel speakers, possibly because fricatives contain spectral properties that cue speaker identity. However, recent research shows that transfer to a novel sound contrast can occur within a single speaker, suggesting that there is room for analogous transfer of learning to a novel speaker with a sufficiently similar pronunciation. This study examines the generalization of phonetic learning for an atypical fricative pronunciation to novel speakers, focusing on the mechanisms that underlie such perceptual adaptation. Results from a set of experiments show that transfer of phonetic learning can occur to both novel speakers and novel phonetic contrasts, but that listeners may rely upon a more conservative mechanism when generalizing learning to novel speakers.

Keywords: speech perception, perceptual learning, phonetic learning, lexically-guided recalibration, accented speech

1. INTRODUCTION

Listeners show a remarkable ability to adapt to a novel accent, with a minimal amount of exposure generally sufficient to improve processing and/or comprehension [1, 2]. One example of such phonetic learning is lexically-guided recalibration of listeners' perceptual categories [3]. For instance, following exposure to a phonetically ambiguous pronunciation (/s/ = [s/f]) embedded in a word (e.g., mo?), listeners recalibrate their /s/ category such that they perceive more tokens along an /s/-/f/ phonetic continuum as /s/. Previous literature has suggested that this type of perceptual learning is often speaker-specific — it does not tend to generalize [4, 5]. However, recent literature suggests that there is some listener tolerance for phonetic mismatch between exposure and test contexts — listeners can generalize learning to a novel phonetic contrast containing the trained sound category [6]. Such results suggest that lexically-guided recalibration of phonetic categories is sufficiently robust to phonetic variation to facilitate speaker-independent adaptation to a given accent, as has been found in related literature on accent accommodation [2]. Crucially, listeners must be able to abstract over between-speaker differences in order to successfully learn the accent. Simultaneously, these must be able to pick up on regularities rising from interactions of the L1 and L2 sound systems. For instance, a common error by non-native speakers (NNS) of English involves the (inter-)dental frictives /θ/ and /ð/, because these sounds are typologically rare. NNS often substitute another sound that does exist in their L1 phoneme inventory (e.g., /f/, /s/, or /t/). However, even NNS with the same L1 background may realize a given sound in different ways [7], leading to perceptually and spectrally distinct realizations across speakers of the "same" accent. A study by [8] also found significant acoustic and articulatory variation in the production of these sounds across native American English speakers. So, achieving speaker-independent learning may require striking the right balance between perceptual flexibility and sensitivity: accommodating between-speaker variability while picking up on systematic phonetic patterns within a given accent.

To assess the nature of the mechanism underlying recalibration of phonetic categories, [6] tested listeners on a series of sound contrasts containing the trained phoneme. Results showed that learning was not contrast-specific, but generalized to a novel contrast that was perceptually similar to the training accent. This result is consistent with a phonetic learning strategy that involves expansion of a perceptual category into neighboring phonetic space, and suggests that the perceptual system utilizes a relatively coarse-grained learning mechanism in such situations. It is plausible that maintaining this kind of general learning strategy could help listeners achieve speaker-independent adaptation to a non-natively accented speech, which speech tends to be especially variable [9].

Existing literature does not provide a clear answer to the question of whether listeners rely on the same mechanisms for speaker-specific vs. speaker-independent phonetic learning [5, 10, 11]. Previous work has suggested that both relatively general and more specific mechanisms may be available to learners [12, 13]. The current study builds on recent work investigating the nature of the mechanism involved in perceptual recalibration or returning of phonetic categories [6]. We do so by testing cross-speaker generalization for multiple phonetic contrasts, in an attempt to understand what mechanisms underlie generalization of such learning.

2. EXPERIMENT 1

Experiment 1 followed the methods used in [6] while introducing a novel female and male speaker to test transfer of learning. Pilot testing found that intermixing tokens for all 3 speakers within a single test task prevented the detection of any training effect, contrary to results of [11, 14]. Therefore, the following experiments directly replicate the procedures in [6], where each group of participants only heard a single speaker.

2.1. Participants

The experiments in this study were conducted using a custom web-based program [15]. Initially, 255 participants were recruited via the Prolific online platform. All participants lived in the U.S. and reported being native speakers of American English with normal speech and hearing. Exclusion of participants who failed to meet experimental criteria (following [6]) resulted in the removal of 45 participants, leaving a total of 210 whose data were retained for analysis.

2.2. Stimuli

The same materials and procedures used in [6] were also used in this study. However, this experiment also included test continua produced by two novel speakers, recorded and processed using the same procedures, testing listeners on either (1) the male speaker from [6] who produced the training materials (age = 27), (2) a novel male speaker (age = 30), or (3) a novel female speaker (age = 25). All speakers were native speakers of American English, had grown up in California, and were living in California at the time of recording.

2.3. Procedure

Participants first had to complete a headphone check [16] . Participants in the accent exposure group then completed the same lexical decision task in [6] to familiarize them with the ambiguous accent (a pronunciation where /θ/ was replaced with [θ/s] in 20 critical words). All participants completed a categorization task using the same 4 /θ/-/s/ minimal-pair continua presented in [6]. Participants were randomly assigned to one of the three speakers. There were a total of 112 tokens (4 continua x 7 tokens x 4 repetitions per token). Tokens within each continuum were randomized and order of continua was counterbalanced across participants.

2.4. Analyses

Prior to analysis, trials with response times < 200 ms or > 2500 ms were removed, following [11, 14, 6]. Across the 3 speakers (familiar male, novel male, novel female), this resulted in removal of 736 trials (2.47% , 3.97%, and 2.72% of data per speaker). Data were analyzed via generalized linear mixed-effects regression modeling, using the lme4 package [17] in R [18]. Maximal random effect structure was used for each model where this did not result in convergence issues, with random slopes fitted for all within-participant main effects [19]. The significance of fixed effects and interactions between them was assessed using a Wald chi-squared test. An identical model was fitted for each group of listeners. Categorical variables were treatment-coded and included condition (training vs control, ref = control), /θ/ word position (word-initial vs. word-final, ref = word-final), and experiment half (trial block 1 vs. block 2, ref = block 1). Block was not included as a predictor in[6], but its inclusion is justified by recent studies showing that the recalibration effect can diminish or disappear over the course of testing [20]. A single numeric variable (continuum step, range = 1-7) was included as a centered numeric variable.

2.5. Results

Results for the familiar male speaker showed significant main effects of continuum step ($\chi^2(1) = 165.54$, $p < 0.001$), experimental group ($\chi^2(1) = 6.43$, $p < 0.05$), and block ($\chi^2(1) = 6.32$, $p < 0.05$). There were also significant 2-way interactions of continuum step by group ($\chi^2(1) = 5.53$, $p < 0.05$) and continuum step by block ($\chi^2(1) = 4.21$, $p < 0.05$). Finally, there were also significant 3-way interactions of step, group, and block ($\chi^2(1) = 6.73$,

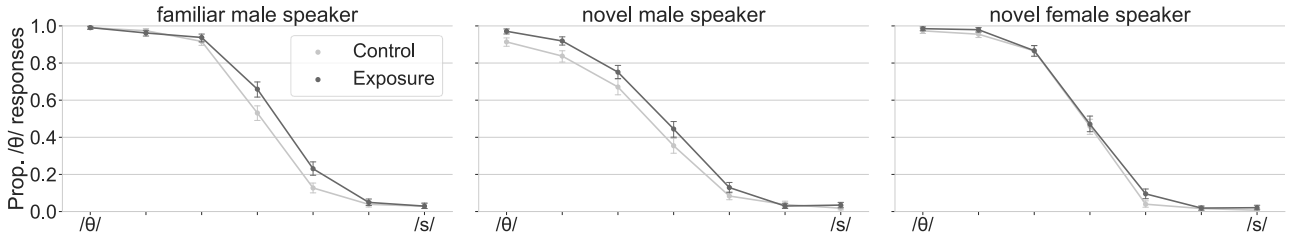


Figure 1: Categorization results for /θ/-/s/ minimal-pair phonetic continua. The effect of accent exposure was significant effect for the familiar male speaker and marginally significant effect for the novel male speaker.

$p < 0.01$), and step, word position, and block ($\chi^2(1) = 7.31$, $p < 0.01$). Crucially, the exposure group showed showed a significantly higher proportion (3.8%) of /θ/ responses (b = 0.86, SE = 0.34, z = 2.54, $p < 0.05$) compared to controls (see Fig. 1).

Results for the novel male speaker showed main effects of continuum step ($\chi^2(1) = 173.19$, $p < 0.001$), word position ($\chi^2(1) = 18.80$, $p < 0.001$), and trial block ($\chi^2(1) = 8.77$, $p < 0.01$), with a marginal effect of experimental group ($\chi^2(1) = 3.28$, $p = 0.070$). There were also significant 2-way interactions of continuum step by word position ($\chi^2(1) = 7.41$, $p < 0.01$) as well as word position by block ($\chi^2(1) = 8.16$, $p < 0.01$). There was also a significant 3-way interaction of step, word position, and block ($\chi^2(1) = 11.14$, $p < 0.001$). The exposure group showed a marginally significant (5.3%) increase in the proportion of /θ/ responses compared to controls (b = 0.57, SE= 0.31, z = 1.81, p = 0.070).

Results for the novel female speaker showed significant main effects of continuum step ($\chi^2(1) = 198.42$, $p < 0.001$) and of block ($\chi^2(1) = 18.91$, $p < 0.001$). There was also a significant 2-way interaction of group and block ($\chi^2(1) = 4.25$, $p < 0.05$). The exposure group showed showed a small numeric shift (1.8%) in their /θ/-/s/ category boundary (see Figure 1, but this difference was not significant (b = 0.13, SE = 0.26, z = 0.48, p = 0.63). There was also a significant decrease in the size of the training effect across experiment blocks (b = -0.92, SE = 0.44, z-score = -2.06, $p < 0.05$).

3. EXPERIMENT 2

Results of Experiment 1 replicated the key result in [6], and also provide tentative evidence of generalization of learning to a novel speaker of the same gender. This is consistent with previous phonetic recalibration studies showing that such learning for fricatives tends to resist cross-speaker generalization [5, 11], transfer only observed under highly similar exposure and test contexts.

The following set of experiments tests whether learning for the same ambiguous pronunciation (/θ/ = [θ/s]) can transfer to both a new speaker and a new phonetic contrast (/θ/-/ʃ/). Given that transfer of learning was observed within a single speaker in [6], it is possible that the same mechanism may facilitate cross-speaker generalization for this contrast. Given gender-based spectral differences in fricative realization [21, 22], we might expect a different pattern of results here, since the female speaker's /ʃ/ may be acoustically closer to the male speaker's /s/, allowing transfer for this contrast even though no transfer was observed in /θ/-/ʃ/ categorization.

The set of experiments presented below each utilize the same training materials and speakers as in Experiment 1. The test phase, however, involves categorization of /θ/-/ʃ/ minimal-pair continua, as in Experiment 2 of [6]. All procedures, materials, and analyses are otherwise identical to those of Experiment 1, a with a comparable number of participants (N=208 across the 3 speakers) and comparable data exclusion rates.

3.1. Results

For the familiar male speaker, model results revealed main effects of continuum step ($\chi^2(1) = 177.03$, $p < 0.001$) and block ($\chi^2(1) = 31.10$, $p < 0.001$). There was also a significant 2-way interaction of word position and block ($\chi^2(1) = 8.77$, $p < 0.01$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 6.00$, $p < 0.05$). Crucially, the effect of experimental group failed to reach significance ($\chi^2(1) = 2.38$, $p = 0.12$).

For the novel male speaker, model results revealed main effects of continuum step ($\chi^2(1) = 127.52$, $p < 0.001$), experimental group ($\chi^2(1) = 6.78$, $p < 0.01$), word position ($\chi^2(1) = 63.24$, $p < 0.001$), and trial block ($\chi^2(1) = 22.06$, $p < 0.001$). There were also significant 2-way interactions of step and word position ($\chi^2(1) = 5.73$, $p < 0.01$), of group and block ($\chi^2(1) = 10.43$, $p < 0.001$), and of word position and

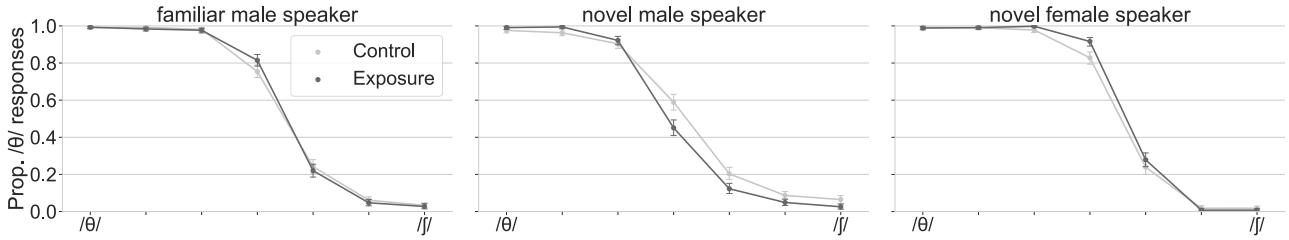


Figure 2: Categorization results for /θ/-/ʃ/ minimal-pair phonetic continua. A significant effect of accent exposure was observed for the novel male and female speakers, but not for the familiar male speaker.

block ($\chi^2(1) = 15.66$, $p < 0.001$). There were also 3-way interactions of step, group, and word position ($\chi^2(1) = 9.77$, $p < 0.01$), of step, word position, and block ($\chi^2(1) = 6.28$, $p < 0.05$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 5.36$, $p < 0.05$). The training effect showed up as a significant decrease (-3.2%) in the proportion of /θ/ responses in the data (b = -0.90, SE =0.35, z = -2.60 $p < 0.01$), but this effect shrank in size over the course of the task: the -5.1% shift seen in the first half of trials decreased to -1.3% in the second half (b = 0.85, SE = 0.26, z = 3.23, $p < 0.01$).

For the novel female speaker, results showed main effects of continuum step ($\chi^2(1) = 112.44$, $p < 0.001$), experimental group ($\chi^2(1) = 5.01$, $p < 0.05$), and trial block ($\chi^2(1) = 5.21$, $p < 0.05$), with a significant interaction of step and word position ($\chi^2(1) = 5.66$, $p < 0.05$). Training resulted in a significant increase (1.8%) in the proportion of /θ/ responses (b = 1.00, SE = 0.45, z = 2.24, $p < 0.05$)

4. GENERAL DISCUSSION

Overall, these results provide evidence that recalibration of phonetic categories can affect perception of novel speakers, consistent with previous work [5, 23, 11]. This study builds on previous work by demonstrating that it is possible for recalibration of phonetic categories to transfer to both novel speakers *and* novel phonetic contrasts involving the trained target sound. The pattern of results suggests that listeners may rely on a more conservative perceptual strategy when generalizing learning to novel speakers, compared to the mechanisms of adaptation for a single speaker reported in earlier work [6].

Experiment 1 replicated previous work [6], finding perceptual learning for an atypical /θ/=[θ/s] pronunciation for the familiar male speaker, and also showing a trend for generalization to a novel male—but not a novel female—speaker. Experiment 2 failed to find perceptual learning for the familiar speaker, but found that training did affect perception of both a novel male and female speaker. Results here show an opposite pattern across speakers, where listeners tested on the female speaker showed a great likelihood to classify ambiguous [θ/s] tokens as exemplars of /θ/, while those tested on the male speaker were *less* tolerant of such atypical pronunciations.

Overall, this pattern of results suggests that transfer of learning is constrained by a relatively targeted mechanism. Crucially, we do not find support for the category expansion mechanism reported in recent work [6], as lack of learning in the familiar speaker condition in Exp.2 fails to replicate their result. While results of Exp.1 are explicable in terms of between-speaker similarity, interactions between speaker gender, sound contrast, and transfer of learning in Exp.2 are puzzling and not clearly aligned with previous literature where cross-speaker and cross-gender learning has been observed [5]. The negative categorization shift for the novel male speaker in Exp.2, where listeners were less likely to perceive /θ/ following accent exposure, is particularly puzzling, as such a learning effect has not been reported in previous phonetic recalibration literature. A fuller discussion of these points, including acoustic analyses, may be found in Ch. 4 of [24], with all study materials publically available at [25].

In summary, these results suggest that generalization of phonetic learning is relatively constrained, consistent with previous work [5, 4, 10, 11]. Nonetheless, learning did transfer to both novel female and male speakers under particular testing conditions, challenging earlier work suggesting that phonetic learning for fricatives is entirely speaker-specific [4]. The mixed results in this study are not entirely surprising given the conflicting pattern of results in the broader literature, where generalization following single-speaker accent exposure has been found in some studies [26] but not others [2]. Future work may shed additional light on the the precise factors modulating such cross-speaker transfer of learning.

5. ACKNOWLEDGEMENTS

This research was supported in part by a Dissertation Completion Fellowship and a National Science Foundation (NSF) Graduate Research Fellowship to Y.V.M. (Grant No. 1752814). We thank Isaac Bleaman, Terry Regier, and Frederic Theunissen for helpful comments.

6. REFERENCES

[1] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[2] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[3] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cognitive Psychology*, vol. 47, no. 2, pp. 204–238, 2003.

[4] F. Eisner and J. M. McQueen, "The specificity of perceptual learning in speech processing," *Perception & Psychophysics*, vol. 67, no. 2, pp. 224–238, 2005.

[5] T. Kraljic and A. G. Samuel, "Perceptual learning for speech: Is there a return to normal?" *Cognitive Psychology*, vol. 51, no. 2, pp. 141–178, 2005.

[6] Y. V. Melguy and K. Johnson, "Perceptual adaptation to a novel accent: Phonetic category expansion or category shift?" *The Journal of the Acoustical Society of America*, vol. 152, no. 4, pp. 2090–2104, 2022.

[7] A. Seibert, "A sociophonetic analysis of L2 substitution sounds of American English interdental fricatives," Master's thesis, Southern Illinois University at Carbondale, 2011.

[8] Y. Melguy, "Strengthening, weakening and variability: The articulatory correlates of hypo-and hyper-articulation in the production of english dental fricatives," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[9] T. Wade, A. Jongman, and J. Sereno, "Effects of Acoustic Variability in the Perceptual Learning of Non-Native-Accented Speech Sounds," *Phonetica*, vol. 64, no. 2-3, pp. 122–144, 2007.

[10] T. Kraljic and A. G. Samuel, "Perceptual adjustments to multiple speakers," *Journal of Memory and Language*, vol. 56, no. 1, pp. 1–15, 2007.

[11] E. Reinisch and L. L. Holt, "Lexically guided phonetic retuning of foreign-accented speech and its generalization." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 2, pp. 539–555, 2014.

[12] R. Schmale, A. Cristia, and A. Seidl, "Toddlers recognize words in an unfamiliar accent after brief exposure: Brief exposure to an unfamiliar accent," *Developmental Science*, vol. 15, no. 6, pp. 732–738, 2012.

[13] D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel." *Psychological Review*, vol. 122, no. 2, pp. 148–203, 2015.

[14] E. Reinisch, A. Weber, and H. Mitterer, "Listeners retune phoneme categories across languages." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 1, pp. 75–86, 2013.

[15] K. Johnson, "Speech perception on the web," 2022. [Online]. Available: https://github.com/keithjohnson-berkeley/perception_on_the_web

[16] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," *arXiv:1406.5823 [stat]*, 2014.

[18] R. C. Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2022.

[19] D. J. Barr, "Random effects structure for testing interactions in linear mixed-effects models," *Frontiers in Psychology*, vol. 4, 2013.

[20] L. Liu and T. F. Jaeger, "Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 12, pp. 1562–1588, 2019.

[21] V. A. Mann and B. H. Repp, "Influence of vocalic context on perception of the [sh]-[s] distinction," *Perception & Psychophysics*, vol. 28, no. 3, pp. 213–228, 1980.

[22] E. A. Strand and K. Johnson, "Gradient and Visual Speaker Normalization in the Perception of Fricatives," in *Natural Language Processing and Speech Technology*, D. Gibbon, Ed. De Gruyter, 1996, pp. 14–26.

[23] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychonomic Bulletin & Review*, vol. 13, no. 2, pp. 262–268, 2006.

[24] Y. V. Melguy, "Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent," Ph.D. dissertation, University of California, Berkeley, 2022.

[25] Y. Melguy, "Mechanisms of perceptual learning," 2022. [Online]. Available: https://github.com/ymelguy/mechanisms_of_perceptual_learning

[26] X. Xie and E. B. Myers, "Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers," *Journal of Memory and Language*, vol. 97, pp. 30–46, 2017.

Phonetic imitation of stops and vowels: Individual stability and perceptual underpinnings

Hanna Zhang[1,2], Jessamyn Schertz[1,2]

[1]University of Toronto, [2]University of Toronto Mississauga
hanna.zhang@mail.utoronto.ca, jessamyn.schertz@utoronto.ca

ABSTRACT

This study examines whether individuals who show greater-than-average imitation of phonetic differences in a given feature also show greater-than-average imitation of other features, as might be expected under the view that general cognitive or social factors play a primary role in predicting individual differences in imitation. An explicit imitation task tested the extent to which English speakers imitated minimal differences in voice onset time (VOT) of voiceless stops and F2 of the vowel /u/, and an ABX discrimination task tested sensitivity to the same differences. Participants imitated differences in both features, and individuals' perceptual acuity for a given feature predicted the extent of imitation of that feature. However, the extent to which an individual imitated one feature did not predict the extent to which they imitated the other, and the same dissociation was found in discrimination. These findings highlight the importance of considering multiple features, and the role of perception, in studies of individual variability in imitation.

Keywords: phonetic imitation, discrimination, individual differences, VOT, F2

1. INTRODUCTION

Phonetic convergence or imitation is the process by which speakers adapt their speech to match more closely with the phonetic properties of the incoming signal. Various social [5, 7], general cognitive [1], and psycholinguistic [12, 15] factors have been shown to influence the degree of convergence. However, there have been few direct tests of whether individual variation in extent of imitation is stable across different features: in other words, do individuals who show greater-than-average imitation of phonetic differences in (for example) consonants also show greater-than-average imitation of vowels?

Previous work has found that extent of convergence in implicit shadowing or conversational tasks is related to several specific personality/cognitive traits, including the Big Five Inventory personality traits of Openness and Neuroticism, Attention-Switching, and rejection sensitivity [27, 13, 2]. An individual's attitude towards the talker has also been shown to play a role, with more positive attitudes eliciting more convergence [4, 27]. Work on explicit imitation, mostly in the domain of second-language sound acquisition, has also found correspondences between non-linguistic traits and phonetic imitation ability, including neurocognitive flexibility [21] and musical talent [8]. This is in line with proposals that there may be innate "talent" for phonetic imitation, with identifiable neurocognitive markers [9, 20].

If non-language-specific characteristics are indeed strong predictors of imitation, they would be expected to be relatively stable and to apply similarly to different phonetic features; in other words, an individual's relative extent of imitation should be consistent over time, across paradigms, and across different features. There is some evidence for this: individual differences in convergence to lengthened VOT during a shadowing task were found to remain relatively stable across time (between two sessions 1-2 weeks apart), and across two different talkers [24]. However, this work examined convergence to a single feature (lengthened VOT) in a single task, leaving open the question of whether this stability would hold across different methodologies or different unrelated features.

Phonetic convergence has been found across a broad range of features, including f0, speaking rate, vowel formants, and VOT (see [17] for a review). Previous work has shown that the extent of convergence or imitation may be influenced by characteristics of the feature itself. Differences in the extent of convergence across two different features have been attributed to differences in phonological relevance [15], perceptual salience [18]. That said, even if there are systematic feature-based differences in extent of convergence, we would still expect that, all else being equal, individuals who show greater-than-average convergence for one feature would also show greater-than-average convergence for other features.

However, there is little evidence that this is the case. Studies exploring the extent to which individuals converge along specific phonetic dimensions have for the most part targeted a single sound or a set of related sound (e.g., VOT of

voiceless stops). Two studies we are aware of have directly examined the stability of individual levels of convergence across prosodic features (e.g. f0 and speech rate) and lexical factors in conversation [22, 25]; neither found evidence that individual levels of convergence were correlated across features. However, given the uncontrolled nature of conversational tasks, and the fact that the amount of possible convergence depends on the distance between the production and the target (such that if the imitator and their interlocuter naturally produce the feature of interest with the same acoustic characteristics, no convergence is possible), it is possible that there is simply too much other variability to reveal individual stability even if it does exist. Therefore, it is possible that more stable patterns would be found in a more controlled task.

1.1. Current Study

In this study, we test whether individual differences in explicit imitation are stable across two distinct phonetic features in English: VOT of voiceless stops and backness of the vowel /u/. Imitation of lengthened VOT has been particularly well-studied and has been shown to be robust [23, 24, 27]. Overall convergence to vowel formant differences has also been found [4, 10], but no studies we are aware of have targeted /u/ backness specifically.

Imitation was tested using an explicit imitation paradigm in which participants are explicitly instructed to imitate minimal differences in the relevant features. Sensitivity to the same differences was tested in an ABX discrimination task. We chose this methodology as it provides a completely controlled paradigm in which it is possible to calculate a measure of imitation that is independent of participants' baseline values, as well as a direct test of perception of the target differences.

Our analysis is structured as follows. First, we test whether there is imitation of, and sensitivity to, differences in each feature: previous work leads us to expect robust imitation of VOT, but we are not aware of studies that directly test sensitivity to, or imitation of, differences in the backness of /u/. We then test whether there is a relationship between individuals' extent of imitation of the two features, as would be expected if the degree to which speakers imitate phonetic differences is a stable property within individuals. Finally, we test the relationship between individual imitation and discrimination in order to determine whether any differences in imitation of the features might be attributable to differences in perceptual salience, since this has been proposed to underlie differences in imitation in previous work [11, 12, 8].

2. METHODS

2.1. Participants

Data from 81 participants was analyzed (mean year-of-birth = 1996 (range = 1991-2004), women: 43, men: 32, non-binary or genderqueer: 6). Participants were recruited via Prolific [19], and all were native speakers of English, born and residing in the US or Canada, and reported learning English and no other languages in the home. An additional 18 participants completed the study but were excluded because they reported early exposure to another language in the home (n=17) or because their recordings were not of sufficient quality for phonetic measurement (n=1).

2.2. Stimuli

Target stimuli consisted of pairs of English words differing minimally in VOT (stop words) or F2 of /u/ (/u/ words). Four stop words (*carrots*, *kale*, *parrots*, *tigers*) and four /u/ words (*choose, food, goose*, *tube*) were produced by a phonetically-trained native speaker of English, and were then manipulated using Praat (PSOLA algorithm for VOT duration and LPC resynthesis for vowel formants) [6] to create two versions of each word. For stop words, Version A had the naturally-produced VOT value (average VOT: 81ms) and Version B had VOT artificially lengthened by 60ms (average VOT: 141ms). For /u/ words, the versions differed in F2: Version A had an F2 corresponding to the naturally produced /u/, which was intentionally produced as far back as possible (1400 Hz), while Version B had a higher F2 (2000 Hz), corresponding to a fronter vowel. Nine additional pairs of stimuli targeting different features were used as fillers.

2.3. Procedure

The experiment was completed fully online on Gorilla [3]. Prior to the main task, participants completed an audio screening task to ensure that they were wearing headphones [26] as well as a baseline word-reading task (not reported here).

The main task consisted of a series of "trial sets," each corresponding to one pair of stimuli (e.g., *tigers* with natural and lengthened VOT). Each trial set was composed of three stages: exposure, imitation, and discrimination. In the exposure stage, participants listened to Versions A and B of the word in sequence. This stage was repeated twice. In the imitation stage, participants listened to the two versions in sequence again, but after each word, they were instructed to repeat the word out loud, imitating what they had heard as closely as possible. This stage was also repeated twice. In the

discrimination stage, participants heard a third word (X) and decided whether it matched the first (A) or second (B). Each trial set contained 4 different ABX trials, with X identical to either the A or B token. In total, each participant produced 32 imitations (8 target words * 2 versions * 2 repetitions) and completed 32 discrimination trials (8 target words * 4 ABX trials).

2.4. Phonetic Analysis (imitation task)

VOT of stop words produced in the imitation task was annotated beginning just before the stop burst and ending at the onset of periodicity in the following vowel. For /u/ words, boundaries were identified using the Montreal Forced Aligner [14] and manually corrected if necessary, using the onset and offset of stable second formant to determine the boundaries. F2 was measured a quarter of the way between the onset of the vowel and the onset of the following consonant (referred to as the 25% point) using Praat [6], and all values were manually checked and corrected if necessary.

3. RESULTS

3.1. Imitation

Imitation was quantified by comparing participants' VOT (for stop words) and F2 (for /u/ words) across the two versions of each word. We used two linear mixed-effects regression models, one for each feature, to test for imitation. The predictor variable for both models was *version* (A vs. B, simple-coded: (-0.5, 0.5)). The response variable for the stop model was VOT and for the /u/ model it was F2 frequency measured at the 25% point, with a by-participant random intercept and slope for feature. For stops, there was a significant difference between conditions (Figure 1): participants produced longer VOTs for lengthened-VOT stimuli (mean 104ms) than natural-VOT stimuli (76ms) (β = 26.06, t = 7.86, $p < 0.001$). For /u/, there was also a significant difference between conditions (Figure 2): participants produced greater F2s for fronted stimuli (1890Hz) than natural stimuli (1669Hz) (β = 208.15, t = 2.98, p = 0.009). To summarize, imitation was found for both features.

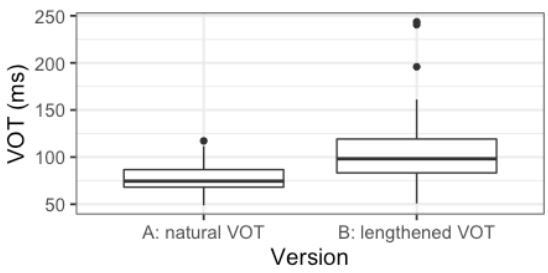


Figure 1: VOT of imitated stop words. Boxplots show distributions of by-participant means.

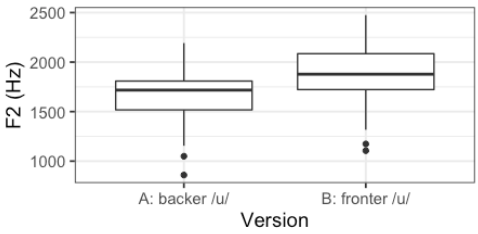


Figure 2: F2 of imitated /u/ words. Boxplots show distributions of by-participant means.

3.2. Discrimination

We used a logistic mixed-effects regression model to test whether participants were able to discriminate differences in both target features and if there were differences in the discriminability of the two features. The predictor variable was *feature* (levels: stop, /u/, simple-coded (-0.5, 0.5)) and the response variable was accuracy on the ABX task, with a by-participant random intercept and slope for feature. Results (Figure 3) showed that overall, discrimination accuracy was significantly above chance (intercept: β = 1.62, $p < 0.001$) and that there was no significant difference in mean discrimination accuracy between stop words (78%) and /u/ words (81%) (β = 0.07, z = 0.126, p = 0.900).

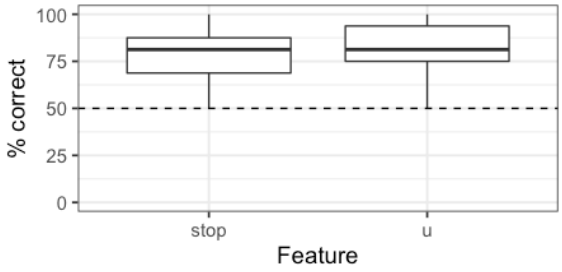


Figure 3: Discrimination accuracy across features. Boxplots show distributions of by-participant means.

3.3. Individual Correlations

To compare individual performance across features and tasks, we calculated by-participant imitation and discrimination scores for each feature. An individual's imitation score was the average difference in VOT (for stop words) or F2 (for vowel words) between imitations of the two versions, with positive values corresponding to differences in the expected direction. An individual's discrimination score for a given feature was the mean accuracy for trials targeting that feature in the discrimination task.

Linear regression models were used to compare correlations between individuals' (a) extent of imitation of stop versus /u/ words, (b) discrimination accuracy of stop versus /u/ words, (c) discrimination

vs. imitation of stop words, and (d) discrimination vs. imitation of /u/ words.

In terms of our primary question about the stability of individual imitation across features, results showed that the extent to which a participant imitates VOT differences does not predict the extent to which they imitate /u/ F2 differences (Figure 4, left: $\beta = 1.24$, $t = 1.62$, $p = 0.110$). In other words, even though most participants did show imitation in the expected direction, as shown by primarily positive imitation scores, there is no evidence that individuals' extent of imitation is stable across these two features. There was also no significant correlation between discrimination of the two features (Figure 4, right: $\beta = -0.02$, $t = -0.211$, $p = 0.833$).

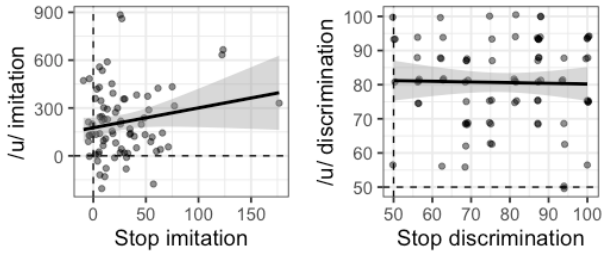


Figure 4: Left: individual imitation of stops vs. /u/; Right: individual discrimination of stops vs. /u/. Dashed lines indicate no imitation or at-chance discrimination.

On the other hand, the extent to which a participant discriminates differences in stop VOTs *does* predict the extent to which they imitate stop VOT differences (Figure 5, left: $\beta = 0.79$, $t = 3.72$, $p < 0.001$), and the same was found for discrimination versus imitation of /u/ fronting (Figure 5, right: $\beta = 6.05$, $t = 3.29$, $p = 0.002$).

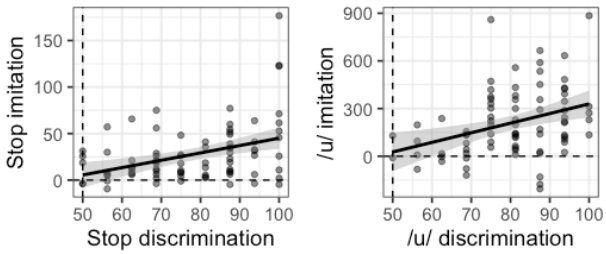


Figure 5: Left: individual discrimination vs. imitation of stops; Right: individual discrimination vs. imitation of /u/.

4. DISCUSSION

Our results suggest that individuals' relative extent of imitation may not be consistent across different features, even in a highly controlled task. English speakers imitated differences in both VOT and voiceless stops and F2 of the vowel /u/ in an explicit imitation task, and considerable individual variability was found for both features. While imitation of within-category VOT differences was expected based on previous work [23, 24, 27], it was less clear as to whether differences in /u/ backness would also be imitated. These results provide novel evidence that speakers are capable of manipulating specific spectral properties of vowels (in this case, F2 of /u/) in an explicit imitation task.

In terms of our primary question of interest, we found no evidence of stability in individual performance across the two features. The extent of an individual's imitation of each feature was, however, related to that individual's perception of the same feature. This extends earlier work examining extent of convergence across features that also found a lack of constancy in convergence during conversational tasks [22, 25]. Furthermore, the correspondence between perception and imitation is consistent with previous proposals for the role of perceptual salience in imitation [11, 12]. While perhaps unsurprising, given that accurate perception is a prerequisite for accurate imitation, this highlights the importance of considering, and ideally, directly testing, the role of perception when trying to account for individual variability in imitation.

Our findings are not consistent with a view that non-linguistic cognitive or social characteristics are the strongest predictors of imitation; if this were the case, we would expect that participants who showed the greatest imitation of stops would also show the greatest imitation of /u/. Our results do not rule out the possibility that these traits play a role in imitation, just that this role may be more minor than that of other factors (including, as shown here, feature-specific individual perceptual acuity). This suggests that it may be problematic to consider results from a task targeting a single feature as an index of an individual's overall imitative ability.

In conclusion, this study found that individual differences in patterns of imitation found for one feature do not necessarily generalize to other features, even in a highly controlled paradigm with substantial individual variability in both perception and production of each feature. Instead, differences in individual perceptual acuity appeared to be the primary driver of the different patterns of imitation in this task. It is therefore imperative to test imitative performance across multiple features, and to control for the role of perception, in work aiming to quantify the role of the many cognitive, social, and linguistic factors underlying phonetic imitation.

5. REFERENCES

[1] Abel, J., & Babel, M. (2017). Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech, 60*(3), 479–502.

[2] Aguilar, L., Downey, G., Krauss, R., Pardo, J., Lane, S., & Bolger, N. (2016). A Dyadic Perspective on Speech Accommodation and Social Connection: Both Partners' Rejection Sensitivity Matters: A Dyadic Perspective on Speech. *Journal of Personality, 84*(2), 165–177.

[3] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407.

[4] Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437–456.

[5] Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*(1), 177–189.

[6] Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.11, retrieved March 2022 from http://www.praat.org/.

[7] Cohen Priva, U., & Sanker, C. (2020). Natural Leaders: Some Interlocutors Elicit Greater Convergence Across Conversations and Across Characteristics. *Cognitive Science, 44*(10), e12897.

[8] Coumel, M., Christiner, M., & Reiterer, S. M. (2019). Second Language Accent Faking Ability Depends on Musical Abilities, Not on Working Memory. *Frontiers in Psychology, 10*, 257–257.

[9] Dogil, G. & Reiterer, S. (2009). Language Talent and Brain Activity. Berlin, New York: De Gruyter Mouton.

[10] Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology, 4*, 346–346.

[11] Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience, 34*(6), 769–786.

[12] Large, N. R., Frisch, S., & Pisoni, D. B. (1998). Perception of wordlikeness: Effects of segment probability and length on subjective ratings and processing of non-word sound patterns. *Research on Spoken Language Processing: Progress Report, 22*, 95–125.

[13] Lewandowski, N., Jilka M. (2019). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication, 4*.

[14] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. (2017) Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proceedings Interspeech 2017,* 498-502,

[15] Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from a shadowing task. *Cognition, 109*, 68–173.

[16] Nielsen, K.Y., & Scarborough, R. (2015). Perceptual asymmetry between greater and lesser vowel nasality and VOT. *Proceedings of ICPhS*.

[17] Pardo, J. S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Special issue: Vocal accommodation in speech communication. *Journal of Phonetics, 95*, 101196.

[18] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The Distinctness of Speakers' Productions of Vowel Contrasts Is Related to Their Discrimination of the Contrasts. *The Journal of the Acoustical Society of America, 116*(4), 2338–2344.

[19] Prolific (2014). https://www.prolific.co

[20] Reiterer, S. M., Hu, X., Erb, M., Rota, G., Nardo, D., Grodd, W., Winkler, S., & Ackermann, H. (2011). Individual differences in audio-vocal speech imitation aptitude in late bilinguals: functional neuro-imaging and brain morphology. *Frontiers in Psychology, 2*, 271–271.

[21] Reiterer, S. M., Hu, X., Sumathi, T. A., & Singh, N. C. (2013). Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. *Frontiers in Psychology, 4*, 782–782.

[22] Sanker, C. (2015). Comparison of phonetic convergence in multiple measures. *Cornell working papers in phonetics and phonology 2015*, 60–75.

[23] Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*(3), 422–429.

[24] Wade, L., Lai, W., & Tamminga, M. (2020). The Reliability of Individual Differences 456 in VOT Imitation. *Language and Speech*, *64*(3), 576-593.

[25] Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. Proceedings of the 2018 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2*, 297–302.

[26] Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79(*7), 2064–2072.

[27] Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and "Autistic" Traits. *PloS One, 8*(9), e74746.

THE PROCESSING OF PROSODIC PROMINENCE IN GERMAN

Barbara Zeyer[1], Martina Penke[1]

[1] University of Cologne
bzeyer2@uni-koeln.de, martina.penke@uni-koeln.de

ABSTRACT

The aim of our study was to investigate how prosodic prominence affects language processing. For that, a word-monitoring task was conducted where the targets either carried a high level prominent accent (*LH**, *L*H*), an accent that is lower in prominence (*L**) or no (prominent) accent at all (*ø*). Based on the as yet sparse findings on prosodic prominence and its effect on language processing in the literature, we postulated that prosodic prominence draws attention and will lead to faster recognition times in a word-monitoring task. Specifically, we assumed that target words carrying a prominent accent type (*LH**, *L*H*, *L**) are recognized faster in a word-monitoring task than target words that are deaccented (*ø*). While previous evidence has found prosodic prominence to affect word recall, it did not speed up word identification in a word-monitoring task, suggesting that prosodic prominence might not exert an immediate influence on on-line language processing.

Keywords: Prosodic prominence, language processing, word-monitoring, German

1. INTRODUCTION

Prosody is used to mark information status [1], to distinguish focus and background [2], or to indicate turn-taking [3]. Also, speakers can use prosodically prominent accent types to draw the listeners' attention to a specific part of the utterance. A prosodically prominent entity stands out from its environment because of the prosodic characteristic it holds, such as loudness, duration or pitch accent [4].

First psycholinguistic studies have found that prosodic prominence affects language processing. Recall-tasks revealed a facilitatory effect of intonation and prosodic manipulation (see e.g., [5]), indicating that prosodically more prominent words get recalled easier than prosodically less prominent words. A word-monitoring task, conducted in English, showed that both, adults and children above six years of age identified the target word faster when it bore the primary sentence accent compared to when the primary sentence accent appeared elsewhere, either before or after the target word [6].

The reported results suggest that prosodic prominence serves to draw the attention to the accented entity, thus enabling a better recall of this entity. Here we wanted to build up on these findings, by investigating whether prosodic prominence also exerts a more immediate processing effect in speeding up language processing. We also wanted to explore the effects of different levels of prosodic prominence on language processing.

2. METHODOLOGY

2.1. Research Questions

In the current study, we wanted to investigate how different levels of prosodic prominence draw the attention of a hearer and affect language processing. For that, we conducted a word-monitoring task, similar to that of Cutler & Swinney [6], where participants were asked to push a key as soon as they recognized the previously defined target word.

To investigate the influence of different levels of prosodic prominence, we manipulated the accent on the target words. In German, different accent types are perceived differently with regards to their perceived prosodic prominence. Baumann & Röhr [7] found that, inter alia, the accent *LH** is perceived as most prominent by native speakers of German, followed by the accent *L*H*. The study furthermore showed that deaccentuation (*ø*) in German is perceived as least prominent and the accent *L** as second least prominent. To investigate the influence of different levels of prosodic prominence on identification times in a word-monitoring task, we presented target words in four different accent conditions: target words either carried *LH**-, *L*H*- or *L**-accent or were deaccented (*ø*).

We predicted that target words having a prominent accent type (*LH**, *L*H*, and *L**) would lead to faster recognition times of the target word compared to the deaccentuation condition (*ø*, the baseline condition) as the prominent accent draws the hearers attention. Moreover, we expected identification times to decrease with increasing prosodic prominence of the target word. The more prominent a word is, the more attention it should draw, leading to shorter identification times. Thus, based on the findings by Baumann & Röhr [7], we expected reaction times to display the following scale: *LH** < *L*H* < *L** < *ø*.

The study was preregistered at OSF (https://osf.io/fea4p).

2.2. Stimuli

In total, we constructed 40 experimental (10 sentences for each of the four different accent conditions *LH**, *L*H*, *L** and *ø* on the target word) and 100 filler sentences. Both, the experimental and the filler sentences consisted of seven words each.

In the experimental sentences, the target word always occurred in the fourth position after the subject. The target words always bore the prenuclear accent and the nucleus was sentence-final. All experimental sentences had the same syntactic structure (Adv V S target word O; see example (1), bold indicating the target word). This way, we ensured that neither the position of the target word nor the syntactic structure of the sentence would be a confounding variable for the results.

(1) *Gerne trinkt Lola* ***morgens*** *ein Glas Milch.*
Adv V S target word O
'Happily drinks Lola in the morning a glass of milk'

The target words were bi-syllabic, trochaic adverbials that were either adverbials of time, of reason, of modality, or of space. Adverbs of different type were equally distributed over the four experimental accent conditions so that each experimental condition contained one modal, two causal, two local, and five temporal adverbials.

We controlled for the word frequency of the target words using WebCelex [8]. The mean word frequency of the target words across the four experimental conditions was 149.08 (*SD* = 328.94). The mean word frequency of the target words did not differ significantly between the different experimental conditions ($F(3, 36) = .144$, $p = .933$).

Additionally, we had students (n = 23) rate our written sentences with regards to how natural the sentences seemed to them on a scale from 1 (very good/natural) to 5 (very bad/unnatural). Mean rating across experimental conditions was 2.21 (*SD* = 1.36). The ratings indicated no significant difference between the four experimental conditions ($F(3, 916) = 1596$, $p = .189$).

Moreover, we controlled for the mean duration in ms of the recorded target words as well as the number of phonemes the target words consisted of. The mean duration in ms of the spoken target words was 502.5 ms (*SD* = 85.15). The mean duration in ms of the target words did not differ significantly across experimental conditions ($F(3, 36) = 1.236$, $p = .311$). The mean number of phonemes the target words consisted of was 5.83 (*SD* = 1.43). There was no significant difference in the mean number of phonemes of the target words across experimental conditions ($F(3, 36) = .781$, $p = .512$).

Conditions / Category	Word frequency	Duration in ms	No. of phonemes	Rating
*LH** on target word	164.0 (*SD* 389.78)	508.80 (*SD* 107.73)	5.8 (*SD* 1.55)	2.25 (*SD* 1.38)
*L*H* on target word	123.8 (*SD* 131.66)	524.70 (*SD* 61.8)	5.6 (*SD* 1.27)	2.17 (*SD* 1.35)
*L** on target word	108.3 (*SD* 98.54)	518.0 (*SD* 72.78)	5.5 (*SD* 1.18)	2.08 (*SD* 1.29)
ø on target word	200.2 (*SD* 544.25)	458.5 (*SD* 92.16)	6.4 (*SD* 1.71)	2.34 (*SD* 1.41)

Table 1: Characteristics of the target words across experimental conditions.

In addition, we created 100 sentences that served as fillers for this investigation. The filler sentences were recorded informally as in spontaneous speech. Target words in the filler sentences were not prosodically manipulated with a particular accent. In these sentences the target words were either the subject or the object and occurred either in the second, third, fourth or seventh position in the sentence. The syntactic structure was either PP V S (= possible target word) (Adv) O (= possible target word) or S (= possible target word) V (Adv) O (= possible target word). The filler sentences contained 37 sentences where the onset of the target word already appeared in a different word before the target word ('onset fillers' see (2)). We included these fillers to ensure that the participants listened to the whole word before reacting. We additionally created 16 'catch trials' which did not contain the previously defined target word and where the correct reaction was not to push the button (see (3)). With this kind of filler, we wanted to ensure that the participants did not arbitrarily push the button at some point in the sentence but listened for the target word.

(2) Onset filler
Syntax: PP V S target word
Target word: *Krone* ('crown')
Sentence: *Auf das* Krokodil *malt Tamara die* ***Krone***. ('It is on the crocodile that Tamara is drawing the crown.')

(3) Catch trial
Syntax: SVO
Target word: *Dackel* ('dachshund')
Sentence: *Der starke Husky zieht den Schlitten.* ('The strong husky is pulling the sled.')

All sentences were recorded by a female trained phonetician who was a native speaker of Standard German. The items were recorded in Audacity [9], using a C520 headset and the Scarlett 2i2 3rd generation 2-in, 2-out USB audio interface. We cut

the items using PRAAT [10] and adjusted loudness using Audacity [9].

The experimental material was presented in five blocks. Each block consisted of 7 experimental and 21 filler sentences. The order of the sentences was pseudo-randomized so that not more than two experimental sentences occurred in a row. Furthermore, we made sure that sentences of the same experimental accent condition did not follow each other. This way we wanted to ensure that the participants would not get accustomed to a particular sentence prosody or would expect target words in a particular position in the sentence.

2.3. Participants

The participants were students at the University of Cologne. Participation in the experiment was voluntary and participants were either rewarded course credits or were reimbursed for their participation.

In total, 58 people participated in the experiment. We excluded eight participants who did not grow up monolingually with German between the age of zero and six years. One more participant had to be excluded because s/he reported to have systematically pushed the key only after the sentence had finished. This left 49 participants for statistical analysis, 15 of which reported to identify as male. The mean age was 23.3 years (range: 18–31 years). All participants reported to be neuro-typical, to have normal hearing abilities and normal or corrected-to-normal visual abilities.

2.4. Procedure

Participants were asked to identify a visually presented target word in a sentence that was subsequently presented auditorily. Trial structure was as follows: A fixation cross appeared for 700 ms on the computer screen, followed by the target word written in caps that appeared for 1500 ms. After that, the sentence was presented auditorily while the screen was blank. Subjects were asked to press a key on the button box as soon as they identified the target word in the sentence. Timeout was set at 4000 ms after the onset of the sentence. Reaction time measurements started with the onset of the target word and ended either after 4000 ms or with the participant pushing the key (see Figure 1).

Participants were tested individually in a quiet room and were required to wear headphones (Sennheiser HD260) during the experiment. The experiment was designed and ran in OpenSesame (version 3.3.10) [11] and was run on a Dell, Latitude 3420, laptop. As a response device, we used the button box by SR Research.

Prior to the experiment, the participants were given oral and written information about what they were asked to do in the experiment, they were asked to fill out a questionnaire on their personal and language background and gave written consent for their participation.

After a practice phase of six sentences, the participants started the experiment. After each block of 28 sentences, they had the possibility to take a break for as long as they wanted to. Each block roughly lasted four minutes.

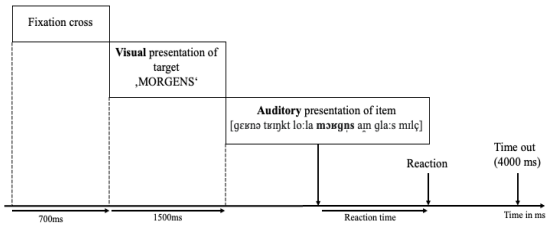


Figure 1: Experimental trial.

3. RESULTS

Reaction times of each participant that differed more than two SDs from the mean reaction time of this particular participant across experimental conditions were removed as outliers. Moreover, we removed the reaction times from the trials where the participants failed to push the button resulting in a wrong answer. In total, 66 reaction times were removed (3.36 %), leaving 1894 data points for statistical analysis.

The mean identification time for target words carrying the prominent *LH** accent was 552.78 ms (SD = 144.63 ms). In the condition *L*H*, the mean word identification time was 543.75 ms (SD = 159.75 ms). The mean identification time for the condition *L** was 562.89 ms (SD = 161.42 ms); the mean reaction time for the condition *ø* was 544.96 ms (SD = 139.01 ms) (see Figure 2).

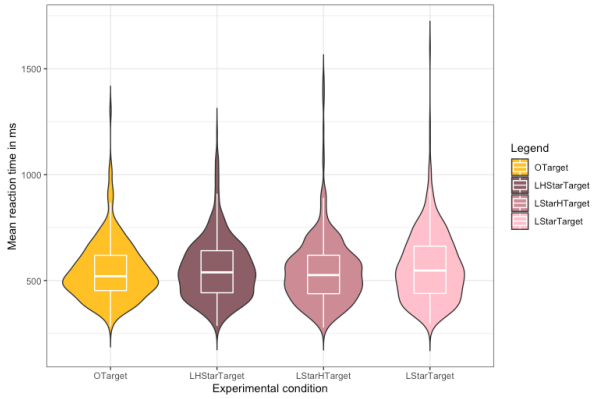


Figure 2: Violin plot of the mean reaction times in ms per experimental condition. OTarget serves as the baseline.

We computed a linear mixed-effect model using the *lme4* package [12] in R [13]. The maximally specified model did not converge and we successively removed

the correlation between random intercept and random slopes for participants as well as the random slopes for participants until the model converged (see Table 2 for fixed effects).

	Est.	SE	t	p
Intercept (ø)	6.270837	.043323	144.746	<.0002
LH*	.013495	.048620	.278	.783
L*H	-.008043	.048622	-.165	.870
L*	.035511	.048658	.730	.470

Table 2: Fixed effects from the linear mixed effects model.

We conducted pairwise comparisons using the *emmeans* package [14] adjusting *p*-values for three comparisons using the Sidak method. Pairwise comparisons turned out to be non-significant for all three comparisons that were calculated (see Table 3).

	Est.	SE	t.ratio	p.value
ø vs. LH*	-.01350	.0486	-.278	.9400
ø vs. L*H	.00804	.0486	.165	.8194
ø vs. L*	-.035511	.0487	-.730	.9870

Table 3: Pairwise comparisons for the baseline condition (ø) vs. the accent types. Note that the '-' means that the condition tested against the baseline is slower than the baseline.

4. DISCUSSION

The results do not show the expected effect of prosodic prominence on word identification. We predicted the mean reaction time in the condition *ø* to be the slowest whereas we expected the mean reaction times in the conditions *L**, *L*H* and *LH** to be significantly faster. However, the mean reaction time in the condition *ø* was numerically faster than the mean reaction times in the conditions *LH** and *L**. Our predicted scale of identification times (*LH** < *L*H* < *L** < *ø*) was, thus, not supported by the results. Rather, the results indicate that prosodically prominent accents on the target word did not facilitate word identification. Note that we tried to control for a number of possible confounding variables such as word frequency, the duration in ms and the number of phonemes of the target words as well as for the rating of the written sentences when we created the items and distributed them amongst the experimental groups. Thus we feel confident that these factors were not confounding our results and that the null result we report here is sound.

Our findings contrast with Cutler & Swinney [6] who reported a significant effect of accent in a word-monitoring task. They found that their ten tested adult participants identified open class target words about 36 ms faster when the primary sentence accent was on the target word compared to when it occurred elsewhere. It might be the case that in the experimental sentences of Cutler & Swinney [6: 151] the prosodic structure of the sentence predicted where the accented target would appear (e.g., *The nurse brought a **clean** towel and took away the dirty one*, bold indicating the target word), while in our experimental sentences hearers might have expected the prenuclear accent to be elsewhere, for example on the subject. Accents are not produced in isolation and expecting the prenuclear accent on a different sentence element than where it actually appeared might have hindered sentence processing and the identification of the target word. Note, however, that having the prenuclear accent on the adverb (= target word) in sentences like *Gerne trinkt Lola **morgens** ein Glas Milch* is not uncommon in German. Thus, it is unclear why this contour should have a hindering effect on sentence processing. In order to investigate the influence of prosodic prominence on language processing detached from possible expectations based on the syntactic structure or the f0 contour of a sentence, a word monitoring task could be conducted with a serial word list.

Also, Cutler & Swinney [6] presented the same 16 experimental sentences in both conditions, with and without an accented target word, in addition to six filler sentences. Given this relatively small number of sentences, it is likely that word identification times were faster upon the second presentation of the sentence, which might have influenced the results.

Given our failure to find an effect of prosodic prominence on word identification times, it might be the case that the influence of prosodic prominence is interacting with different stages of language processing. While prosodic prominence leads to better word recall in recall tasks, it might be the case that it does not exert a more immediate effect on language processing as measured in our word monitoring task or that its effects during language processing are attenuated. More research is needed to determine whether or not prosodic prominence influences word identification during language processing.

5. ACKNOWLEDGEMENTS

We would like to thank Christine Röhr for recording our items and Max Hörl for support in the statistical analysis. This study was funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252 "Prominence in Language".

6. REFERENCES

[1] Baumann, S., Grice, M. 2006. The intonation of accessibility. *J. Pragmatics* 38, 1636–1657.

[2] Mücke, D., Grice, M. 2014. The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *J. Phonetics* 44(1), 47–61.

[3] Cutler, A., Pearson, M. 1986. On the analysis of prosodic turn-taking. In: Johns-Lewis, C. (eds), *Intonation in Discourse*. Routledge, 139–155.

[4] Terken, J., Hermes, D. 2000. The perception of prosodic prominence. In: Horne, M. (eds.), *Prosody: Theory and Experiment*. Kluwer, 89–127.

[5] Savino, M., Winter, B., Bosco, A., Grice, M. 2020. Intonation does aid serial recall after all. *Psychonomic Bulletin and Review* 27(2), 366–372.

[6] Cutler, A., Swinney, D. 1987. Prosody and the development of comprehension. *J. Child Language* 14(1), 145–167.

[7] Baumann, S., Röhr, C. 2015. The perceptual prominence of pitch accent types in German. *18th Proc ICPhS 2015*.

[8] Baayen, R., Piepenbrock, R., Gulikers, L. 1995. *The CELEX lexical database (WebCelex)*. Linguistic Data Consortium.

[9] Audacity Team. 2022. Audacity®: Free Audio Editor and Recorder [Computer application]. Version 2.4.2.

[10] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.16.

[11] Mathôt, S., Schreij, D., Theeuwes, J. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2), 314–423.

[12] Bates, B., Maechler, M., Bolker, B., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-27.1.

[13] R Core Team. *R: A language and environment for statistical computing* (2021.09.0). R Foundation for Statistical Computing.

[14] Lenth, R., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Migues, F., Riebl H., Singmann, H. 2022. Estimateed Marginal Menas, aka Least-Suqare means. Version 1.8.3.

TEMPORAL ORGANIZATION IN LEXICAL SELF-REPAIR IN COLOMBIAN SPANISH UNSCRIPTED SPEECH

Kelly Johanna Vera Diettes[1] | Leendert Plug[2]

[1]Universidad Nacional de Colombia | [2]University of Leeds
kjverad@unal.edu.co | l.plug@leeds.ac.uk

ABSTRACT

We report on an investigation of the temporal organization of lexical self-repair, focusing on the timing of the speech cut-off and repair onset. [1] showed that speech error repairs in which speech is interrupted early are completed quickly, while later repairs are generally slower. This supports the idea that early repairs are initiated through inner speech monitoring, and late repairs through overt speech monitoring. Research on lexical repair [2] suggested that different subtypes of repair are subject to different temporal constraints. Moreover, [3] proposed that speakers may strategically delay repair onset. This study investigates error-to-cut-off and cut-off-to-repair interval durations in lexical repairs sampled from a corpus of Colombian Spanish conversation. It assesses the correlation between interval durations and the predictive power of semantic variables. The results provide support for the findings of [1] but not entirely for [2]; also, they support the notion that speakers may delay interruption and repair.

Keywords: temporal organization, timing, self-repair, Spanish, unscripted speech.

1. INTRODUCTION

This paper reports on an investigation of the temporal organization of lexical self-repair in unscripted Colombian Spanish. In lexical self-repair, a speaker rejects one lexical choice for another, as in '*a bee w-(...) a butterfly was*'. We focus on the time it takes to interrupt speech after the start of the problematic 'target' word attempt (*target-to-cutoff*: '*bee w-*' in the example) and the time between the interruption and the start of the repair (*cutoff-to-repair*: '*(...)*').

[1] and [4] observed that target-to-cutoff durations in elicited Dutch speech error repairs are bimodally distributed, with two peaks of dispersion separated by around 500 ms. This is consistent with early repairs following error detection in 'inner speech' monitoring, while late repairs follow error detection in overt speech monitoring. This distinction has an impact on cutoff-to-repair timing too, such that repairs that follow error detection in overt speech monitoring require more planning time [5]. However, it has also been suggested that speakers may delay cut-off, possibly to allow themselves to plan the repair while prioritising fluency [3, 5]. This might generate repairs with long target-to-cutoff but short cut-off-to-repair intervals.

As [4] point out, there is no *a priori* reason to expect that findings on the timing of speech error repairs will generalize to lexical repairs, as different processing mechanisms are involved. [2] reports that in Dutch spontaneous lexical self-repairs, repairs in which the repairable word is completed before interruption have lower cutoff-to-repair durations than repairs with early interruptions. [2] further suggests that semantic sub-types of lexical repair may be subject to different temporal constraints. In particular, [2] shows that classifying repairs as error or appropriateness repairs, and further dividing error repairs into factual and linguistic error repairs, is informative. [2] excludes repairs with 'editing terms' [6], such as 'I mean' in '*a bee w- I mean a butterfly was*'. These complicate the analysis of the temporal organization of lexical self-repair: different subtypes of repair are likely to differ with respect to the occurrence of such terms in the cutoff-to-repair interval, and it seems reasonable to expect that cutoff-to-repair intervals containing an editing term are on average longer than silent ones.

To further our understanding of the temporal organization of lexical self-repair, we explore the distributions of target-to-cutoff and cutoff-to-repair intervals in a sample of self-repairs from a corpus of unscripted Colombian Spanish speech. We looked for indications of bimodality following [4, p. 20] and considered the relationship between these two intervals, taking account of the existence of multiple semantic sub-types of repair as well as the presence *vs* absence of editing terms.

2. DATA AND METHODS

2.1. Corpus

The data for this paper comprise 97 instances of self-initiated lexical self-repair extracted from sub-corpora of the Unscripted Colombian Spanish Interactive (UCSI) Corpus. The UCSI project aims to

collect unscripted speech from the different dialects and subdialects of Colombian Spanish, based on the dialectal classification by [6]. Conversational speech is collected by means of two interactive tasks: A South American adaptation of the *Diapix task* [8, 9] and a consensual response task in which speakers need to verbally agree on answers to questions before reporting these in writing. Repairs were sampled from recordings of both tasks as completed by 7 speaker pairs, from 3 subdialects of Colombian Spanish. Initially 104 repairs were identified; five were later excluded (see below).

2.2. Segmentation

Each repair was orthographically transcribed by the first researcher, a native speaker of Colombian Spanish, then semi-automatically transcribed in Praat [10] using the Easyalign Spanish extension [11]. Resulting phoneme, word, syllable and phone tiers were hand-corrected by the first author. A point tier was added to locate crucial temporal landmarks, as illustrated in Figure 1. Here the 'target' of the repair is *cuatro* 'four', which is corrected to *dos* 'two'. The speech is cut off in the speaker's first production of *zapatos* 'shoes'. The interval between 1 and 3 is the target-to-cutoff interval ('T_C'); the interval between 3 and 4 is the cutoff-to-repair interval ('C_R'). Interval durations were extracted and log-transformed prior to modelling, although the figures below will plot durations in ms.

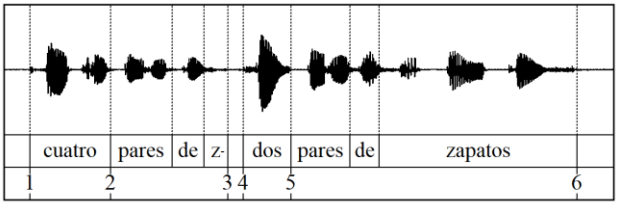


Figure 1: Segmented waveform of a repair. '1' and '2' delimit the target word; '3' is the cutoff; '4' and '5' delimit the repair word; '6' marks the end of the repair.

2.3. Semantic classification

Each instance was classified as an appropriateness or error repair, using the criteria described in [2]. Repairs with factual inaccuracies or linguistic ill-formedness are error repairs; in other cases, we assume that the initial lexical choice is rejected for pragmatic felicity or 'appropriateness' reasons. The classification was done by the first author and an independent second rater, a Colombian Spanish discourse analyst. Classifications matched for 92 repairs (90%). Consensus was reached for 9 out of the remaining 10; one repair was excluded as the raters agreed it could not be reliably classified.

2.4. Editing terms

We coded for the presence of editing terms following [6]. Examples include *o sea* 'I mean' and filled pauses such as *eee* [e:] and *mmm* [m:]. Some Colombian Spanish expressions could not be easily mapped to the classification scheme in [6]; as a result, 4 further repairs were excluded. This left 97 instances of repair, of which 26 contain editing terms.

2.5. Statistical analysis

We first inspected the distributions of the durations of T_C and C_R. We then performed a cluster analysis to investigate the relationship between the two intervals in more detail. Subsequently, we modelled the interval durations using linear mixed-effect regression models and cluster membership using a conditional inference regression tree. All the analyses were performed in R [12] using *lme4* [13], *k-means* and *party* [14]. Table 1 shows the variables entered into the analyses. Chi square tests confirmed that none of the independent variables are collinear, although we can note that none of the linguistic error repairs in our data set contain editing terms.

Dependent variables	**T_C, C_R**
Independent variables	**Completeness (Com)** 1. *Complete* 2. *Incomplete*
	Repair Type (ReT) 1. *Appropriateness* 2. *Error* 2.1 *Linguistic* 2.2 *Factual*
	Editing Terms (EdT) 1. *Yes* 2. *No*
Random effect	**Speaker**

Table 1: Variables entered into the analysis.

3. RESULTS

3.1. T_C and C_R duration distributions

The mean duration of T_C is 515 ms (range 99–2272); the mean duration of C_R is 307 ms (range 0.44–2982). Hartigans' dip tests confirm that neither distribution is significantly multimodal (T_C: D=0.03, p=0.921; C_R, D=0.03, p=0.876), and the interval durations are not significantly correlated with each other (Pearson's r=0.15, p=0.13).

We investigated the relationship between the two intervals in more detail through *k-means* clustering. This revealed three clusters, illustrated in Figure 2. First, there are repairs with a relatively short T_C and

a relatively short C_R. We call these 'Early-Early' repairs: they have an early interruption and an early repair onset. Second, there are repairs with a relatively short T_C but a relatively long C_R. We call these 'Early-Late' repairs. Third, there are repairs with a relatively long T_C but a relatively short C_R. We call these 'Late-Early repairs'. Figure 2 shows that only one repair in our data set might potentially qualify for a 'Late-Late' classification.

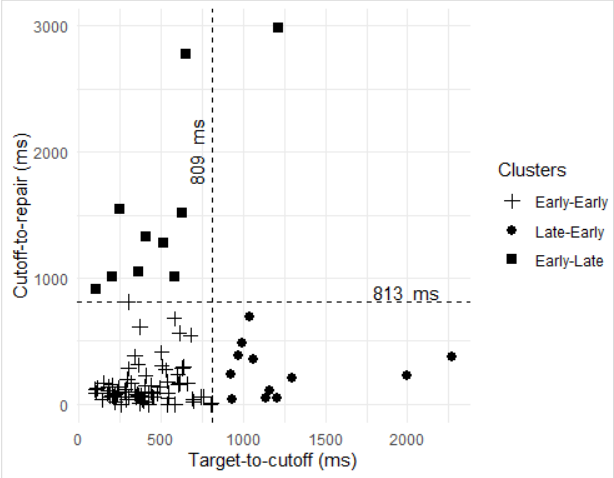


Figure 2: Scattergram showing T_C and C_R durations and cluster membership, with threshold lines separating the clusters.

3.2. Relevance of editing terms and repair semantics

As expected, there is a systematic relationship between the presence or absence of editing terms and C_R interval duration. Interestingly, EdT also seems systematically related to T_C duration. Figure 3 shows that repairs without any editing terms predominantly fall in the Early-Early cluster, while repair with an editing term are evenly distributed across the three clusters.

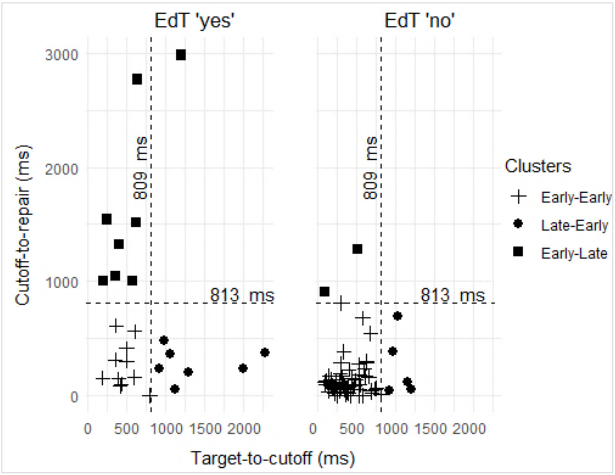


Figure 3: Scattergram as in Figure 2, split by EdT (left *Yes*, right *No*).

This was confirmed by our models. We first built a linear mixed-effects regression model for each of T_C and C_R with EdT (*Yes*, *No*) as well as ReT (*Appropriateness*, *Error*) and Com (*Complete*, *Incomplete*) as predictors. For each model we started with a random intercept for speaker identity and then added the predictors in turn, only keeping an additional predictor if model comparison indicated model fit was significantly improved. For each model, we tested for the relevance of an interaction between EdT and ReT, and we assessed whether implementing ReT with three levels (*Appropriateness*, *Factual*, *Linguistic*), added predictive value. The latter was not the case.

Optimal models are given in Table 2. The model for T_C confirms that EdT has a significant effect (higher duration with editing terms); so does Com (higher duration for repairs with completed target words). The model shows no significant overall effect of ReT, but reveals a significant interaction between ReT and EdT. The interaction is illustrated in Figure 4. It suggests that the effect of EdT on T_C duration is predominantly observed among appropriateness repairs.

The model for C_R shows a significant effect of EdT (higher duration with editing terms) only.

Target-to-Cutoff	**Estimate**	**Df**	**t**	**p**
Intercept	487.3	97	7.57	<0.01
EdT *Yes*	457.6	97	4.11	<0.01
Com *Incomplete*	-158.8	97	-2.31	0.02
ReT *Error*	27.5	97	0.35	0.72
EdT *Yes**ReT *Error*	301.2	97	-2.01	0.04
Cutoff-to-repair	**Estimate**	**Df**	**t**	**p**
Intercept	166.2	97	3.15	<0.01
EdT *Yes*	525.84	97	5.16	<0.01

Table 2. Optimal linear mixed effects models for T_C and C_R.

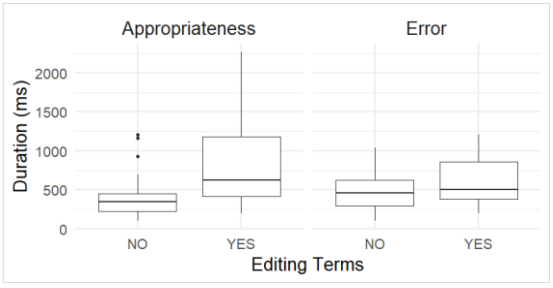


Figure 4. Box plot illustrating the significant interaction between EdT and ReT for T_C duration.

As a final step, we fitted a conditional inference regression tree with cluster membership (as in Figure 2) as response variable. This confirmed that the relationship between cluster membership and EdT

seen in Figure 3 is at the level of significance ($p<0.001$). Neither Com nor ReT gave rise to significant splits in the tree.

4. DISCUSSION

This investigation assessed the temporal organization of lexical self-repair in Colombian Spanish unscripted speech, focusing on the target-to-cutoff and cutoff-to-repair intervals. In our data set, both have wide duration distributions. When considered separately, neither appears significantly bimodal; however, when considered together, the emergence of three clusters suggests that dividing both distributions into 'early' and 'late' ranges is informative. As expected, repairs with editing terms have longer cutoff-to-repair intervals than those without; interestingly, they also have longer target-to-cutoff intervals, at least among appropriateness repairs. Also as expected, repairs in which the target word is completed before the repair are associated with longer target-to-cutoff durations than repairs in which the target word is interrupted.

As indicated above, [1] and [4] observed that target-to-cutoff durations in elicited Dutch speech error repairs are bimodally distributed, with two peaks of dispersion separated by around 500 ms; moreover, speech error repairs in which speech is interrupted early are completed quickly, while later repairs are generally slower. In our spontaneous lexical repairs, we can identify 'early' and 'late' repairs, but unlike [2], we do not find that late repairs in terms of target-to-cutoff or target word completeness are also late in terms of cutoff-to-repair; and early repairs in terms of target-to-cutoff are not necessarily early in terms of cutoff-to-repair either.

The absence of a Late-Late cluster and the presence of a Late-Early cluster fit with the idea that speakers can delay interruption to maximize fluency [3, 5]. However, the additional existence of an Early-Late cluster suggests that speakers do not invariably use this delaying strategy. Further work is needed to expand our data set, so that we assess the robustness of the Late-Early and Early-Late temporal organizations, as well as the absence of a commonly observable Late-Late one.

Our analysis confirms that the presence or absence of editing terms is highly consequential for the temporal organization of repair. It might seem unsurprising that cutoff-to-repair intervals which contain editing terms are on average longer than silent cutoff-to-repair intervals, simply because producing an editing term takes time. [3] suggests that the use of editing terms might actually be an alternative 'repair delay' strategy to delaying the cutoff: after an early interruption, speakers can maintain fluency by using editing terms while planning the repair. In fact, Figure 3 shows that this accounts for most instances in our Early-Late cluster, as most contain one or more editing terms.

We also observed a positive effect of the presence of editing terms on target-to-cutoff duration, modulated by repair type. Concretely, it appears that our Late-Early cluster contains a majority of instances with an editing term (which must be relatively short), and these are mostly appropriateness repairs. If this pattern generalizes to a larger data set, more qualitative analysis is warranted to identify the specific subtypes of repair that have this temporal organization.

Unlike [2], we did not observe any more substantive effects of semantic repair type on interval durations, including cluster membership. [2] also suggests that the distinction between linguistic and factual error repairs is informative for understanding repair timing. Our results do not confirm this in quantitative terms, but our observation that none of our linguistic error repairs contain editing terms is consistent with linguistic and factual error repairs having distinct organizations. A larger data set should allow us to explore the nature of this distinction further.

5. ACKNOWLEDGEMENTS

This research was funded by Grant 16 of 2019 by the Ministry of Science, Innovation and Technology of Colombia—MINCIENCIAS. The first author is grateful to the Universidad Nacional de Colombia for granting her the study leave that allowed this research.

6. REFERENCES

[1] Nooteboom, S. 2010. Monitoring for speech errors has different functions in inner and overt speech. In Everaert, M., Lentz, T., De Mulder, H., Nilsen, Ø. (eds.), *The Linguistic Enterprise*. Amsterdam: John Benjamins, 213–233.

[2] Plug, L. 2016. Informativeness, Timing and Tempo in Lexical Self-Repair. *Language and speech.* 59/4, 516–543.

[3] Seyfeddinipur, M., Kita, S., & Indefrey, P. 2008. How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, 108/3, 837–842.

[4] Nooteboom, S.G. & Quené, H. 2017. Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language,* 95, 19–35.

[5] Nooteboom, S.G. & Quené, H. 2019. Temporal aspects of self-monitoring for speech errors. *Journal of Memory and Language,* 105, 43–59.
[6] Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, 14, 41–104.
[7] Ruiz Vásquez, N.F. 2020. The Dialects of Colombian Spanish: A New Classification. *Lenguaje* (Cali, Colombia). 48/2, 160–195.
[8] Figueroa, M.A., García, D.A. and Salamanca, G.F. 2019. DiapixSp: adaptación al español y aplicación piloto de una herramienta de elicitación de habla espontánea y colaborativa. *Estudios de fonética experimental*. 257–288.
[9] Van Engen, K.J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M. and Bradlow, A.R. 2010. The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and speech*. 53/4, 510–540.
[10] Boersma, P., Weenink, D. 2017. Praat: Doing phonetics by computer. www.praat.org
[11] Goldman, J., & Schwab, S. 2014. EasyAlign Spanish: an (semi-)automatic segmentation tool under Praat. In Y. Congosto, M. L. Montero, & A. Salvador (Eds.), *Fonética experimental, Espacio Europeo de Educación Superior e Investigación* 1, 629-640. Arco/Libros.
[12] R Core Team. 2022. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[13] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4: Linear mixed-effects models using Eigen and S4*. Version 1.1
[14] Hothorn, T., K. Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15, 651–674.
[15] Levelt, W.J.M. 1989. *Speaking. From intention to articulation*. Cambridge, Massachusetts: The MITPress.
[16] Hartsuiker, R. J., & Kolk, H. H. J. 2001. Error monitoring in speech production: A computationaltest of the perceptual loop theory. *Cognitive Psychology*, 42, 113–157.

TOWARDS AUTOMATING THE ASSESSMENT OF TEACHERS' TEACHING EMOTIONS IN PRESCHOOL CLASSROOMS

Mingyue Huo[1], Yan Tang[1,2]

[1]Department of Linguistics, University of Illinois Urbana-Champaign, USA
[2]Beckman Institute for Advanced Science and Technology, USA
{mhuo5, yty}@illinois.edu

ABSTRACT

In-person assessment of teaching emotion as part of a teacher's working performance evaluation is onerous, and can be intrusive to some teachers in teaching. This study shows preliminary attempts at solving this issue using a machine learning approach with bi-modal artificial neural networks, which made predictions by combining acoustic and textual features extracted from preschool teachers' spontaneous speech captured in real teaching scenarios. In a binary classification task identifying the emotions as "positive" and "negative", the prediction achieved an accuracy of 79.0%, an F1-score of 83.1% in cross-validation, and an accuracy of 68.4%, an F1-score of 67.2% in leave-one-subject-out validation. When further adding speech samples identified as "neutral" to the task, a decreased accuracy of 52.4% and F1-score of 53.6% were received in cross-validation, confirming the difficulties in labelling this type of naturalistic speech data even for human raters.

Keywords: teaching emotion, spontaneous speech, neural network, acoustic features

1. INTRODUCTION

Emotions are prominent aspects of human experiences and play a defining role in social interactions. Ratings of human emotion have been used in various developmental [1], educational [2], and health research studies. These ratings involve observing, interpreting, and coding expressed emotions, which are explicitly or implicitly expressed during interactions between teachers and students in the classroom and invoked in the raters. Emotion ratings are also used in policy decision-making, such as the Emotional Support domain of the Classroom Assessment Scoring System [3], which is a determinant of funding allocations for the largest federally funded preschool program, Head Start, in the United States. However, there is ample evidence that raters frequently disagree [4, 5]. A possible reason for such disagreement is cultural and language diversity. A cross-language and cross-cultural study of a listener's capability to identify emotion using acoustic cues [6] showed that both universal and language-specific cues of emotion expressions affect emotion identification in an unfamiliar language and culture. However, established rating systems have not successfully addressed this issue.

Systems and algorithms using audio information have been developed to assist in identifying and recognising human effects efficiently, such as emotions, sincerity, irony and deception from linguistic and paralinguistic information extracted from physical speech signals [7, 8, 9]. For speech emotion classification, early studies have used Gaussian mixture models and Hidden Markov models to learn the acoustic feature distribution, and used Bayesian classifier and support vector machine (SVM) classifier to categorise emotions [10, 11]. In recent years, deep neural networks have demonstrated better performance on the acted database IEMOCAP, and end-to-end models combining Convolutional Neural Networks with long short-term memory have been reported to outperform the traditional approaches on the naturally elicited database RECOLA [12, 13]. Despite many studies on emotion classification, there is limited work on detecting teachers' teaching emotions, especially from spontaneous speech captured in classrooms. With a Recurrent Neural Network, Liang et al. [14] achieved an accuracy of 85.3% in a teaching emotion recognition task on a Mandarin speech corpus, for which teaching scenarios were simulated using professional voice actors under specific instructions. Cen et al. [15] trained an SVM model for detecting the emotions of students during online learning. This model combined the probabilities of finer emotion categories to predict "positive" ("happy") and "negative" ("sad" and "anger") emotions in one speech utterance.

Developing a machine-aided rating system for teachers' teaching emotions is significantly meaningful, because it may provide more objective and consistent assessment for teachers from different cultural and racial backgrounds, reducing biased feedback and subsequent unfair treatment due to insular individual opinions. It could alleviate the mental stress on some teachers, who are sensitive to their working ambience and audience, due to the presence of their performance evaluator(s). When relying on computer coding of teachers' basic emotions, raters can focus on higher-level constructs (e.g. responding and instructing).

In this study, we examine the feasibility of applying ML techniques to recognising teaching emotion from teachers' *spontaneous* speech as an essential step towards investigating how culturally-specific acoustic signatures are associated with teaching emotions. As opposed to elicited emotional speech used in many previous studies, we work on natural utterances captured when teachers were conducting ordinary teaching activities in preschool classrooms in the United States, in order to obtain an ecologically-valid assessment of machine performance. Instead of proposing novel ML algorithms or using sophisticated ML techniques to address the problem, we

focus on obtaining a benchmark to inform future studies.

2. METHOD

2.1. Speech data

The speech recordings used in this study were drawn from the EMOtion TEaching Rating Scale (EMOTERS) dataset [16]. This corpus is primarily used to assess teachers' emotion-focused teaching practices, which support children's social and emotional development and school readiness. It contains 1,606 10-minute (approx. 268 hours) realistic spontaneous speech recordings with human-coded emotion labels. The original recordings include both video and audio recordings; only audio signals were extracted for analyses in this study. Due to the original audio being saved in MP3 format, some acoustic information had no longer been preserved in the files. Constrained by the data sharing protocol, out of the 1606 recordings in the original dataset, we could only access 31 recordings (approximately 5 hours) from 11 teachers (9 females and 2 males) for this study.

During data preparation, the over 5-hour recordings were segmented into small excerpts based on the teacher's conversational turns, as most speech emotion recognition tasks are developed and evaluated on utterance-level data [17]. This process resulted in 1570 speech excerpts in total, among which 540 excerpts were used as the samples for the subsequent experiments. Other excerpts were discarded due to the target voice being mixed with other unintended sounds (e.g. children's voices and background noise) to different degrees. All the excerpts were primarily coded into "positive" (POS) or "negative" (NEG) individually by three raters: two native American English speakers and one Chinese English speaker. If the emotion in a sample was not obvious to the rater, this sample was labelled as "neutral" (NEU). Of the 540 samples, 261 (48%) were labelled consistently across the raters; 134 (25%) were labelled the same by any two raters, while there was no consensus at all (i.e. the three raters had totally different opinions) for 145 samples (27%). The final label for a sample was the one rated by the majority of the raters (i.e. rated the same by at least two raters); the 145 samples without agreement were merged to NEU, leading to 206, 147 and 187 samples for POS, NEG and NEU, respectively as in Table 1. The labelling results suggested the human raters were indecisive between POS and NEG for over 1/3 of the samples.

Table 1: Teacher sex and rated emotion ("POS", "NEG", "NEU") distribution of the samples

Teacher	Gender	No. of Samples	**POS**	**NEG**	**NEU**
1	Female	15	9	2	4
2	Female	100	50	17	33
3	Female	29	15	4	10
4	Female	85	34	23	28
6	Female	49	18	14	17
5	Female	27	6	11	10
7	Female	79	27	10	42
8	Female	54	21	15	18
9	Female	18	12	4	2
10	Male	21	8	6	7
11	Male	63	6	41	16
total:		540	206	147	187

2.2. Feature preparation

Previous studies [18, 19, 20] suggested that humans can encode and show their emotions using both linguistic (via word choice) and acoustic (via vocal configuration) means. To exploit both acoustic and textual properties from teachers' vocalisation, acoustic measurements and bag-of-words representation were extracted from the speech samples.

There is a wide range of acoustic feature sets available for speech-related identification and analysis tasks, e.g. eGeMAPS [21], ComParE [9], YAAFE [22]. Most of these feature sets calculate different dimensions of low-level temporal and spectral descriptors (LLDs) from 10-20 ms speech frames. A non-exhaustive list of the LLDs is shown below:

- Energy: intensity, loudness
- Pitch: F0, voicing probability, F0 contour
- Voice quality: jitter, shimmer and logarithmic harmonic-to-noise ratio
- Spectral shape: Mel-Frequency Cepstral Coefficients, Perceptual Linear Predictive coefficients.
- Temporal dynamics: 1st-order differential and 2nd-order acceleration coefficients

High-level statistics functions are then derived based on the LLDs for the global utterance, including max, min, mean, the slope of linear approximation, skewness, kurtosis and more. We chose the ComParE feature set, which was designed for the Computational Paralinguistics Challenge and has been used in many speech emotion recognition tasks, such as [7]. Therefore, it was considered a reasonable starting point to obtain a baseline performance on the current speech data. Since ComParE extracts many possible features for a general purpose, a total of 6373 acoustic features was calculated from each speech sample. We assumed that many features in the ComParE set might overlap in terms of useful information they provide for classification tasks. Initial experiments suggested that reducing the feature dimensionality from 6737 to 162 explaining 90% of the variance in the data using Principal Component Analysis led to a similar machine performance as when all 6373 features were used. Therefore, only the first 162 principal components were supplied as acoustic features to the downstream ML algorithm to expedite training and prediction.

As for textual features, the transcriptions of the 540 samples were processed into 2447 unique words. The number was further reduced to 378 by removing the low-frequency words (which occurred only once in the dataset) to avoid noise interference. As a result, the textual model took the 378-dimension bag-of-words representation as its input features.

2.3. Model Structure and evaluation

We trained four uni-modal models using acoustic features or textual features separately to construct a baseline for

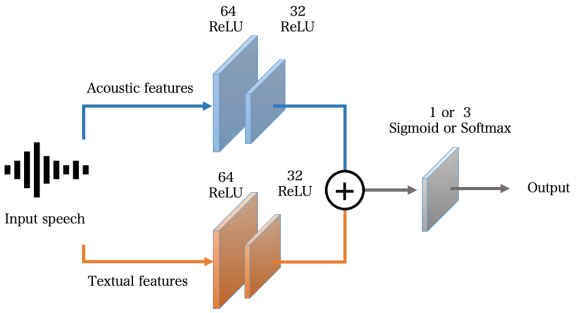


Figure 1: The structure of the bi-modal ANN model: acoustic and textual features were fed into two linear layers separately, then the outputs were concatenated before being fed into the output layer. The number of neurons in each layer is noted.

comparison. The first two classifiers were trained by a kernelised SVM with the radial basis function. The other models were trained as standard feedforward Artificial Neural Networks (ANN) with two hidden layers of 64 and 32 neurons towards the output layer. The Rectified Linear Unit (ReLU) was used as the activation function at each hidden layer, and a sigmoid or softmax function was used for the output layer depending on whether the task was binary or three-class classification. For a given ML algorithm, all the model specifications were the same for both the acoustic and textural features, except that the input size of the former was 162 (principal components) while that of the latter was 378 (words).

A bi-modal ANN was further trained to combine the textural and acoustic information. As shown in Figure 1, the 162 acoustic features were fed through two hidden layers with ReLU as the activation function, resulting in a 32-dimension vector for each sample as the output of the second hidden layer. Meanwhile, the textual features were fed through another two hidden layers, also resulting in a 32-dimension vector. Then, the vectors from the two modalities were concatenated and fed into the output layer. We expected the model to be informed of more information by combining the two kinds of features, leading to better predictive performance than the uni-modal models. To obtain a benchmark performance on this data set, no hyperparameter-tuning was performed on the models; all model parameters were empirically chosen.

Having observed the difficulties the human raters encountered during the rating process, we evaluated the model performance in two tasks: a binary (POS vs NEG) classification and a three-class classification including all three emotional categories (POS, NEG and NEU). Model performance was evaluated as accuracy and F1-score. Since the dataset was relatively small, a 5-fold cross-validation (CV, teacher-dependent) was employed to assess the overall predictive power of each model, along with leave-one-subject-out (LOSO, teacher-independent) CV for testing the model generalisability across teachers. During LOSO CV, the model each time was trained on the samples from 10 teachers and the samples from the remaining teacher were reserved for testing.

3. RESULTS

3.1. Uni-modal baselines

Table 2 presents the average performance of the 5-fold CV on the uni-modal models trained using SVM and ANN with different types of features in the binary classification task. Overall, the ANN models outperformed SVM models in both modalities, and the models trained using the acoustic features outperformed the models trained using the textural features.

Table 3 shows the performance of LOSO CV of the uni-modal models trained on acoustic or textural features. For the 11 teachers, the number of samples considerably varied across individuals from 11 to 57, with an average of 31 samples. Since the dataset was created from a spontaneous speech corpus, the imbalanced distributions in sample number and emotion class across teachers were inevitable at the preliminary stage. Compared to the performance in the early teacher-dependent evaluation, both the model using the acoustic features and the model using textual features were seen to decrease by approximately 10.6 and 12.2 percentage points (ppts) respectively in accuracy, and 15.4 and 14.6 ppts in F1-score. Nevertheless, the acoustic model (64.3%) still exhibited a somewhat more robust performance than the textural model (60.7%) in terms of accuracy, but otherwise similar measured as F-score. Since the sample distribution was imbalanced, the weighted average scores (weighted by the sample number for each teacher) are also shown in Table 3.

3.2. Bi-modal model

As illustrated in Figure 1, the bi-model uses both the acoustic features and textural features for making predictions. The last row of Table 2 shows the CV performance of the bi-modal ANN model when identifying POS and NEG samples in the teacher-dependent task. As anticipated, the bi-modal ANN achieved a higher accuracy of 79% and F1-score of 83.1% than the best uni-modal model, i.e. "ANN-acoustic" with an accuracy of 74.9% and F1-score of 79.5%. The last two columns of Table 3 display the LOSO CV performance of the bi-modal ANN model. While teacher-independent models exhibited decreased performance compared to the uni-modal models, the bi-modal model still outperformed the textual and acoustic models in weighted accuracy, and showed a better balance between the two performance metrics, suggesting a better generalisability than the uni-modal models.

Table 2: Uni-modal and bi-modal model accuracy and F1-score on binary classification (POS vs NEG). Metrics in percentage are calculated across 5-fold cross-validation

Model	Accuracy (%)	F1-score (%)
SVM-text	67.3	76.9
SVM-acoustic	70.4	78.8
ANN-text	72.9	78.1
ANN-acoustic	74.9	79.5
Bi-modal ANN	**79.0**	**83.1**

Table 3: Model accuracy and F1-score in LOSO CV for binary classification (POS vs NEG). Both mean performance across teachers and mean further weighted by the number of samples are also presented. Teachers marked by "*" are males.

Teacher	No. of Samples	ANN-acoustic		ANN-text		**Bi-modal ANN**	
		Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
1	11	63.6	70.2	54.5	66.7	72.7	80.1
2	67	61.3	66.2	64.5	68.4	69.4	74.0
3	19	63.2	53.0	57.9	56.4	68.4	55.9
4	57	64.9	72.3	56.1	57.4	64.9	61.4
5	32	75.0	80.3	62.5	71.3	78.1	77.3
6	17	76.5	74.0	64.7	69.5	70.6	76.8
7	37	71.9	68.7	68.8	60.0	78.1	67.7
8	36	47.2	61.1	52.8	69.0	50.0	56.9
9	16	62.5	65.0	75.0	79.9	81.3	74.0
10*	14	64.3	70.6	64.3	71.0	57.1	66.7
11*	47	57.4	24.0	46.8	29.4	61.7	48.6
Mean	31	**64.3**	**64.1**	**60.7**	**63.5**	**68.4**	**67.2**
Weighted mean		**63.5**	**62.6**	**59.8**	**60.9**	**67.7**	**65.6**

3.3. Three-class classification

When further adding speech samples identified as neutral (NEU) to the classification task, the accuracy of the acoustic and textual uni-modal ANNs, and bi-model ANN decreased by 27.9, 24.4 and 25.4 ppts; the F1 shrunk by 33.1, 29.0 and 29.6 ppts, respectively. The predictive power of the textual uni-modal ANN appeared to deteriorate more severely than the acoustic-only ANN. Despite the lower performance compared to the binary classification, the bi-modal ANN held its marginal lead in the three-class classification over the uni-modal models. Figure 2 compares the 5-fold teacher-dependent CV performance of all models tested above. Unless specified as "3-class", the model is for the binary classification task. As shown in the figure, the bi-modal ANN outperformed the uni-modal models in both the binary and the three-class classification tasks.

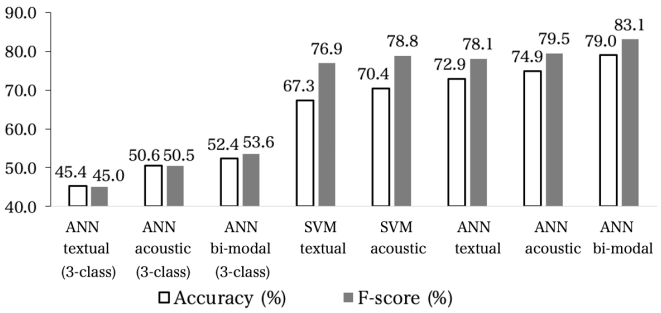


Figure 2: Teacher-dependent CV comparison among models

4. DISCUSSION AND CONCLUSION

This study used a bi-modal ANN that combines acoustic and textual features extracted from spontaneous speech to recognise teachers' teaching emotions in preschool classrooms in the United States. In the binary classification task, the bi-modal ANN achieved an accuracy of 79.0% and an F1-score of 83.1%. In all the tasks including three-class classification (POS, NEG and NEU) and LOSO CV, the bi-modal model outperformed the uni-modal models, which only used acoustic features or textural features. The more robust performance of the uni-modal model using acoustic features, especially in the three-class classification task, could suggest that some emotions are better encoded in teachers' vocalisation than in the vocabulary they use while talking. The reduced predictive power in the three-class classification task shows that the ML models employed in this study cannot account for the uncertainties in ratings due to human raters' different perceptions and cultural backgrounds. Further using fine-tuned models and more advanced ANN architectures could better capture the subtle nuances encoded in teachers' speech and can help differentiate emotions, which may explain the better machine performance reported in the literature. Though the performance falls behind Liang et al. [14], it should be noted that the database in their study was recorded by six male and six female teachers on five balanced emotions, resulting in a potentially higher probability for more robust classification. Other factors such as sample size, audio quality and features used may have also affected the machine performance in this study. Gent et al. [23] compared the machine performance in irony recognition from speech when using a set of refined acoustic features to using the entire CompPare set, and argued that a robust performance comes from the relevance of the features, not the quantity. Further detailed acoustic analyses on this type of spontaneous speech could help identify a set of more representative features for recognising teaching emotions.

As the preliminary study, this work tested the feasibility of using a machine learning-based approach to aid in the emotion-based assessment of preschool teachers' teaching performance. It can also provide a benchmark for automatic emotion detection from speakers' vocalisation acquired in realistic situations.

Acknowledgements We thank Drs Rachel Gordon, Katherine Zinsser and Timothy Curby for sharing the EMOTER database. We also thank Dr Chilin Shih for the early discussion of the idea and her support. We thank the reviewer for their thoughtful comments.

5. REFERENCES

[1] D. S. Messinger, "Positive and negative: Infant facial expressions and emotions current directions in psychological science," *Psychological Science*, vol. 11, pp. 1–6, 2002.

[2] S. A. Denham and H. H. Bassett, "Early childhood teachers' socialization of children's emotional competence," *Journal of Research in Innovative Teaching and Learning*, vol. 12, pp. 133–150, 2019.

[3] R. C. Pianta, K. M. La Paro, and B. K. Hamre, *Classroom Assessment Scoring System (CLASS) Pre-K version*, Paul H Brookes Publishing, Baltimore, MD, 2008.

[4] B. K. Hamre, R. C. Pianta, J. T. Downer, J. DeCoster, A. J. Mashburn, S. M. Jones, J. L. Brown, E. Cappella, M. Atkins, S. E. Rivers, M. A. Brackett, and A. Hamagami, "Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms," *The Elementary School Journal*, vol. 113, pp. 461–487, 2013.

[5] R. A. Gordon, F. Peng, T. W. Curby, and K. M. Zinsser, "Using the many-facet Rasch model to improve observational quality measures: An introduction and application to measuring the teaching of emotion skills," *Early Childhood Research Quarterly*, vol. 55, pp. 149–164, 2021.

[6] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[7] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, 2020, pp. 2042–2046.

[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[9] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proceedings of Interspeech 2016*, 2016, pp. 2001–2005.

[10] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. Ieee, 2003, pp. II–1.

[11] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, 2014.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[14] L. Jie, Z. Xiaoyan, and Z. Zhaohui, "Speech emotion recognition of teachers in classroom teaching," in *2020 Chinese Control And Decision Conference (CCDC)*. IEEE, 2020, pp. 5045–5050.

[15] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, technology, design, and learning*. Elsevier, 2016, pp. 27–46.

[16] K. Zinsser, T. Curby, and R. Gordon, "Emotion teaching rating scale," www.emoters.org, 2016.

[17] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[18] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[19] J. M. Wilce and J. M. Wilce, *Language and emotion*. Cambridge University Press, 2009, no. 25.

[20] A. Majid, "Current emotion research in the language sciences," *Emotion Review*, vol. 4, no. 4, pp. 432–443, 2012.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[22] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, August 9-13 2010, pp. 441–446, http://ismir2010.ismir.net/proceedings/ismir2010-75.pdf.

[23] H. Gent, C. Adams, C. Shih, and Y. Tang, "Deep Learning for Prosody-Based Irony Classification in Spontaneous Speech," in *Proc. Interspeech 2022*, 2022, pp. 3993–3997.

PRODUCTION AND PERCEPTION OF VOCAL EMOTIONS: A COMPARISON OF MANDARIN CHINESE AND GERMAN EMOTIONAL PROSODY

Huan Wei[1], Mathias Scharinger[1,2], Ulrike Domahs[1,2]

[1]Department of German Linguistics, Philipps-University Marburg, Marburg, Germany
[2]Center for Mind, Brain, and Behavior, Universities of Marburg and Giessen, Marburg, Germany
weih@students.uni-marburg.de

ABSTRACT

Previous findings have demonstrated that prosodic cues accompanying expressed emotions are similar in different languages. Still, it is unclear how those suprasegmental properties vary in different languages and lead to different mappings between expressions and emotional meanings. The present study compared the production and perception of vocal emotions (*NEUTRAL*, *HAPPINESS*, *PLEASANT SURPRISE*, *SADNESS*, and *DISGUST*) in Mandarin Chinese and German. We found that positive emotions in both languages were produced with significantly higher pitch, while negative emotions were expressed with longer duration. However, the pitch contours and durations vary between the two languages. In a perception task measuring response times for the categorization of emotions, 21 German and 21 Mandarin native speakers performed best for *NEUTRAL* with faster response times. Mandarin speakers performed worse for positive than for negative emotions. The group differences in both the production and perception tasks suggest language-specific effects in the processing of emotional prosody.

Keywords:
emotional prosody production, recognition of emotional prosody, cross-linguistic processing

1. INTRODUCTION

In discourse, the emotional states of a given statement are often expressed by means of prosodic cues (e.g., pitch, intensity, and duration). Cross-cultural studies [1], [2] point out that speakers without knowledge about a language are able to recognize emotional prosodies successfully in this respective language, suggesting that emotional prosody exhibits a core set of acoustic-perceptual features which promote accurate cross-language recognition of emotions. However, [3]–[6] report an in-group advantage in recognizing emotional prosody for native speakers in contrast to L2- or foreign speakers and pointed to culture- and language-specific paralinguistic patterns affecting the encoding and decoding of vocal expressions of emotions.

Xu and colleagues [7] proposed emotional meanings to be encoded along benefit-oriented bio-informational dimensions, which involve both segmental and prosodic aspects of the vocal signal. Although previous findings [8]–[10] have shown that the prosodic cues used for the production of emotions in different languages are identical, it is unclear how these suprasegmental properties vary in different languages and lead to different mappings of expressions to emotional meanings. To follow up on this issue, the present study compared the production of vocal emotions in the tonal language Mandarin Chinese and the non-tonal language German. It was assessed how language (Mandarin Chinese and German) and culture (Asian and Western) affect the production and perception of vocal emotions in Mandarin Chinese and German. If emotional prosody is produced universally, it can be expected that the acoustic features of the emotional prosodies in both languages are similar. If the perception of emotions in the speech is unaffected by the culture, it can be expected that the native speakers of both languages could successfully recognize all the emotional prosodies in their native languages.

2. PRODUCTION OF VOCAL EMOTIONS

2.1. Stimulus material

We selected ten bisyllabic German nouns from the semantic category »food«, each with initial syllable stress and no reduced syllables (e.g., *Kaffee* "coffee"). According to the EmoInt-2017-Database [11], the mean emotional valence of the stimuli was 5.695 (SD = 0.52). Additionally, we selected ten

Chinese nouns related to »food«, built from two characters and covering all five tones in Mandarin Chinese (e.g., *苹果, píng-guǒ*, "apple").

Three female native speakers of German and Mandarin Chinese were instructed to record each stimulus conveying four emotions (*HAPPINESS*, *PLEASANT SURPRISE*, *SADNESS*, *DISGUST*) and a *NEUTRAL* mode. All recordings were conducted in a sound-proof studio and used as stimuli for the acoustic analysis and perception experiments.

2.2. Acoustic features

The word duration and intensity of the stimuli (Figure 1) were analyzed using *Praat* [12], and time-normalized pitch contours of each utterance (10 points per syllable) were generated using a Praat script of *ProsodyPro* [13] (Figure 2). Statistical analyses were conducted using ANOVAs and post-hoc pairwise comparisons of emotions with Bonferroni adjustment in the R [14].



Figure 1: Average word duration and intensity

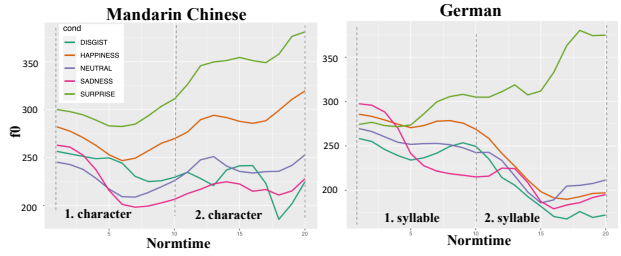


Figure 2: Average pitch contours of stimuli

conditions	emotion	pitch	duration	intensity
	results of post-hoc test within Chinese group			
NEUTRAL	HAPPINESS	< 0.001	-	< 0.001
	SURPRISE	< 0.001	-	< 0.001
	SADNESS	-	< 0.001	< 0.001
	DISGUST	-	< 0.001	-
HAPPINESS	SURPRISE	< 0.001	-	-
	SADNESS	< 0.001	< 0.001	< 0.001
	DISGUST	< 0.001	< 0.001	< 0.001
SURPRISE	SADNESS	< 0.001	< 0.001	< 0.001
	DISGUST	< 0.001	< 0.001	< 0.001
DISGUST	SADNESS	-	< 0.001	< 0.001
	results of post-hoc test within German group			
NEUTRAL	HAPPINESS	-	-	< 0.001
	SURPRISE	< 0.001	-	< 0.001
	SADNESS	-	< 0.01	< 0.001
	DISGUST	< 0.05	< 0.001	-
HAPPINESS	SURPRISE	< 0.001	-	-
	SADNESS	< 0.001	< 0.05	< 0.001
	DISGUST	< 0.001	< 0.001	< 0.001
SURPRISE	SADNESS	< 0.001	-	< 0.001
	DISGUST	< 0.001	-	< 0.001
DISGUST	SADNESS	-	-	< 0.001
Chinese	German	between two languages		
NEUTRAL	*NEUTRAL*	-	< 0.05	< 0.001
HAPPINESS	*HAPPINESS*	< 0.001	< 0.001	< 0.001
SURPRISE	*SURPRISE*	-	< 0.001	< 0.001
DISGUST	*DISGUST*	< 0.05	< 0.01	< 0.001
SADNESS	*SADNESS*	-	-	< 0.05

Table 1: Significant results of the post-hoc tests

The ANOVA results (Table 1) showed significant differences between emotions in pitch, word duration, and intensity for the Chinese and German stimuli: pitch ($F(1, 5998) = 26.11$, $p < 0.001$), word duration ($F(1,298) = 25.42$, $p < 0.001$), and intensity ($F(1, 598) = 244.4$, $p < 0.001$). The emotional prosodies in German had longer word duration and higher intensity, while Mandarin Chinese had higher pitch. These findings suggest that the suprasegmental properties of emotional prosody differ between Mandarin Chinese and German, which may affect the decoding of vocal expressions of emotions.

3. PERCEPTION OF VOCAL EMOTIONS

3.1. Methods

Twenty-one native Mandarin speakers (12 females, mean age of 31.33 years, age range 23-41 years) judged on the stimuli with varying emotional prosodies in Mandarin Chinese, and twenty-one native German speakers (12 females, mean age of 28.86 years, age range 21-54 years) on those in German. Participants of both groups were healthy and did not report any hearing impairments.

The online study consisted of two parts, an online survey on *SoSci Survey* (version 3.4.03) [15] for the demographic data and a *PsychoPy* experiment (version v2021.2.3) on *Pavlovia* [16]. The online experiment consisted of six blocks with 25 trials each. Each trial started with a white fixation point, followed by a 300ms inter-stimulus interval (ISI). Then, an auditory stimulus was presented with a white fixation cross for up to 4000ms until a response was made. Participants had to judge the emotional prosody of each word in their native language and press one of six response options (five

emotional conditions plus one for 'others'). An inter-trial interval lasted for 1000 ms after each response.

3.2. Results

The response times were analyzed utilizing linear mixed regression models and the accuracies with logistic mixed regression models. The fixed factors included the variables CONDITION and GROUP, and the base model included both PARTICIPANTS and SPEAKERS as random intercepts.

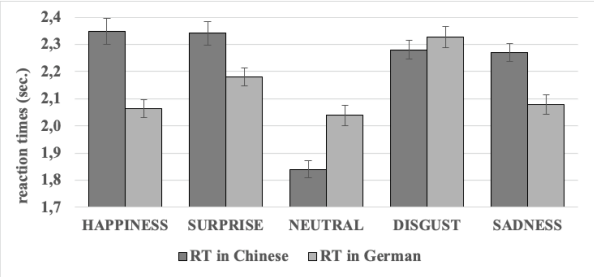


Figure 3: Mean reaction times of the correct answers in the categorization of vocal emotions in Mandarin Chinese and German by the native speakers

The results showed that German native speakers had shorter reaction times (*mean* = 2.13s, *SD* = 0.73) compared to Chinese native speakers (*mean* = 2.25s, *SD* = 0.72) (Figure 3). Although the statistical analysis did not show significant effects of the conditions or interactions on response times between the two groups, within-group analyses revealed significant differences between conditions in response times for both Chinese (χ^2 = 132, p < 0.001) and German group (χ^2 = 100.46, p < 0.001). Table 2 displays the post-hoc comparisons of conditions within each group. The Chinese native speakers had significantly shorter reaction times in recognizing the *NEUTRAL* mode than emotional prosodies. However, their reaction times for recognizing positive emotional prosodies *HAPPINESS* and *SURPRISE* were longer than for the negative emotional prosodies *DISGUST* and *SADNESS*. In contrast, German native speakers recognized the *NEUTRAL* mode with the shortest reaction times, but they needed significantly longer reaction times to recognize emotional prosody *DISGUST* compared to the other emotional prosodies.

conditions	emotion	RT	accuracy
	results of Chinese group		
NEUTRAL	HAPPINESS	t = 1.98 *	z = -14.05 ***
	SURPRISE	6.48 ***	-12.66 ***
	SADNESS	7.06 ***	-6.47 ***
	DISGUST	7.71 ***	-4.1 ***
HAPPINESS	SURPRISE	4.52 ***	-
	SADNESS	5.07 ***	8.62 ***
	DISGUST	5.74 ***	10.85 ***
SURPRISE	SADNESS	-	6.98 ***
	DISGUST	-	9.28 ***
DISGUST	SADNESS	-	-2.5 *
	results of German group		
NEUTRAL	HAPPINESS	-	-4.05 ***
	SURPRISE	2.46 *	-2.82 **
	SADNESS	-	-
	DISGUST	7.72 ***	-5.37 ***
HAPPINESS	SURPRISE	3.9 ***	-
	SADNESS	2.03 *	4.26 ***
	DISGUST	9.16 ***	-
SURPRISE	SADNESS	-	3.02 **
	DISGUST	5.31 ***	-2.61 **
DISGUST	SADNESS	-7.12 ***	5.57 ***
Chinese	German	between two languages	
NEUTRAL	*NEUTRAL*	-	-
HAPPINESS	*HAPPINESS*	-	-2.8 **
SURPRISE	*SURPRISE*	-	-2.58 **
DISGUST	*DISGUST*	-	-
SADNESS	*SADNESS*	-	-

Table 2: Significant results of the post-hoc tests

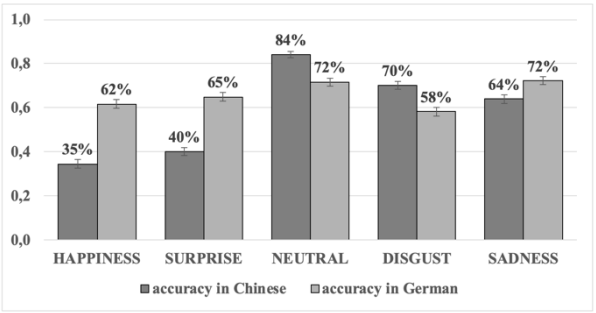


Figure 4: Accuracies of the categorization of vocal emotions in Mandarin Chinese and German by the native speakers

The analysis of overall accuracies (Figure 4) revealed a significant interaction between CONDITION and GROUP (χ^2 = 186.68, p < 0.001). Specifically, German native speakers had significantly higher accuracies in categorizing positive emotional prosodies *HAPPINESS* and *SURPRISE* compared to the Mandarin native speakers (Table 2). Furthermore, within-group comparisons showed significant effects of CONDITION for both Chinese (χ^2 = 414.43, p < 0.001) and German group (χ^2 = 48.4, p < 0.001). The Mandarin Chinese native speakers performed significantly worse when categorizing positive emotions *HAPPINESS* and *SURPRISE*, while German native speakers had lower accuracy in categorizing the emotion *DISGUST*.

4. GENERAL DISCUSSION

The production study found that the acoustic features of the emotional prosody (pitch, word duration, and intensity) were similar in both languages, with both expressing negative emotions *DISGUST* and *SADNESS* with a longer word duration, and the positive emotions *HAPPINESS* and *SURPRISE* with higher pitch and intensity. However, the analysis also revealed significant differences in pitch, word duration, and intensity between the two languages, indicating that the suprasegmental properties of emotional prosody vary across different languages. The results of the perception study showed that both the Mandarin and German native speakers were equally sensitive to emotional prosody in their native language. However, the response patterns differed slightly depending on the specific emotions and language. To be precise, the Mandarin native speakers needed significantly longer reaction times to recognize the positive emotions *HAPPINESS* and *SURPRISE*, and had lower accuracy scores for these emotions. These findings are consistent with the results of [10] that Mandarin listeners had lower accuracy in recognizing positive emotions. On the other hand, the German native speakers had the most difficulties recognizing the emotion *DISGUST*, as evidenced by significantly longer reaction times and lower accuracy scores.

conditions	**Chinese participants responses**					
	HAPPINESS	SURPRISE	NEUTRAL	SADNESS	DISGUST	OTHERS
HAPPINESS	**37%**	13%	45%	1%	1%	3%
SURPRISE	33%	**42%**	16%	0.5%	0.5%	9%
	German participants responses					
DISGUST	4%	2.5%	13.7%	6.8%	**58.3%**	3%

Table 3: Response matrices for the categorization

We analysed participants' error patterns to investigate the categorization difficulties in the Chinese and German participants (Table 3). Our findings revealed that the Chinese listeners were generally able to distinguish between positive and negative emotions, but struggled with distinguishing between *HAPPINESS*, *SURPRISE,* and *NEUTRAL*. This suggests that the acoustic features of these emotions in Mandarin may have contributed to the confusion. Specifically, we found that the word durations of positive emotions *HAPPINESS* and *SURPRISE* in Mandarin were significantly shorter than in German or negative emotions in Mandarin Chinese, but there were no significant differences between the positive and neutral emotions in Mandarin Chinese. Furthermore, while there were significant differences in pitch between *SURPRISE*, *HAPPINESS*, and *NEUTRAL*, the pitch contours of these conditions were similar (Figure 2). These similarities in pitch contours and the shorter word duration may have made it challenging for the Chinses listeners to differentiate between these emotions. In contrast, the German group did not show such a performance difference between positive and negative emotions. However, the analyses of the error patterns in their responses to the emotion *DISGUST* revealed that 13.7% of the stimuli were classified as *NEUTRAL*, 6.8% as *SADNESS*. These findings suggest that the acoustic features of *DISGUST*, *SADNESS*, and *NEUTRAL* in German may have contributed to the confusion. Specifically, the lack of intensity differences between *DISGUST* and *NEUTRAL*, and the similarity in word duration and pitch contours of *DISGUST* and *SADNESS* may have made it challenging for German speakers to differentiate between these emotions. In addition to the acoustic explanation, Asian culture has a commonly observed norm to hide negative emotions in the communication of messages [6]. This culture-specific rule may strengthen Chinese listeners' sensitivity to acoustic features of negative emotions even more than to positive emotions. The group differences in both the production and the perception tasks suggest language-specific effects on the processing of emotional prosody, at least when the words varying in emotional prosody are presented without any contextual information.

5. CONCLUSIONS

The present study investigated whether prosodic cues of vocal emotions vary in Mandarin and German and how they may lead to different mappings of emotional prosodies to emotional meanings. We found that both languages produced positive emotions with higher pitch than neutral and negative emotions. In contrast to positive emotions, negative emotions were expressed by changes in duration. The behavioral study on the same stimuli with a categorization task showed that the Mandarin group performed better with negative than positive emotions. Further studies need to show whether Chinese listeners may have more difficulties using pitch information for emotional categorization than German listeners because pitch plays a crucial role in Chinese lexical processing.

6. REFERENCES

[1] W. F. Thompson and L.-L. Balkwill. 2006. "Decoding speech prosody in five languages," *De Gruyter Mouton*, 407–424. doi: https://doi.org/10.1515/SEM.2006.017.

[2] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz,. 2009. "Recognizing Emotions in a Foreign Language," *J Nonverbal Behav*, vol. 33, no. 2, 107–120. doi: 10.1007/s10919-008-0065-7.

[3] K. R. Scherer, R. Banse, and H. G. Wallbott,. 2001. "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *J Cross Cult Psychol*, vol. 32, no. 1, 76–92. doi: 10.1177/0022022101032001009.

[4] H. Wei, Y. He, C. Kauschke, M. Scharinger, and U. Domahs,. 2022. "An EEG-study on L2 categorization of emotional prosody in German," *Speech Prosody 2022*, 629–633. doi: 10.21437/speechprosody.2022-128.

[5] P. Liu, S. Rigoulot, and M. D. Pell,. 2015. "Cultural differences in on-line sensitivity to emotional voices: comparing East and West," *Front Hum Neurosci*, vol. 9, 311. doi: 10.3389/fnhum.2015.00311.

[6] S. Paulmann and A. K. Uskul,. 2013. "Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners," *Cognition Emot*, vol. 28, no. 2, 230–244. doi: 10.1080/02699931.2013.812033.

[7] Y. Xu, A. Kelly, and C. Smillie,. 2013. "Emotional expressions as communicative signals," *Prosody and iconicity*, 33–60.

[8] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz,. 2009. "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *J Phonetics*, vol. 37, no. 4, 417–435. doi: 10.1016/j.wocn.2009.07.005.

[9] S. Paulmann, M. D. Pell, and S. A. Kotz,. 2009. "Comparative processing of emotional prosody and semantics following basal ganglia infarcts: ERP evidence of selective impairments for disgust and fear," *Brain Res*, vol. 1295, 159–169. doi: 10.1016/j.brainres.2009.07.102.

[10] P. Liu and M. D. Pell,. 2012. "Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli," *Behav Res Methods*, vol. 44, no. 4, 1042–105. doi: 10.3758/s13428-012-0203-3.

[11] M. Köper, E. K. Klinger, and R. Klinger,. 2017. "IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning," *Association for Computational Linguistics*.

[12] Boersma and Weenink,. 2018. "Praat: Doing phonetics by computer [Computer program]. Version 6.0.37.". https://www.fon.hum.uva.nl/praat/ (accessed Oct. 21, 2021).

[13] Y. Xu,. 2013. "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis," *TRASP*.

[14] R. C. Team,. 2018. "R: The R Project for Statistical Computing, version 1.1.419.". https://www.r-project.org/ (accessed Oct. 21, 2021).

[15] D. Leiner,. 2014. "SoSci survey, version 3.4.03". https://www.soscisurvey.de

[16] J. W. Peirce,. 2007. "PsychoPy—Psychophysics software in Python," *J Neurosci Meth*, vol. 162, no. 1–2, 8–13. doi: 10.1016/j.jneumeth.2006.11.017.

AN ACOUSTIC ANALYSIS OF BERLIN DATABASE OF EMOTIONAL SPEECH BASED ON BIO-INFORMATIONAL DIMENSIONS

Zhengyang Cai, Yi Xu

University College London
caizhengy2022@163.com, yi.xu@ucl.ac.uk

ABSTRACT

We performed a systematic acoustic analysis of the Berlin database of emotional speech (Emo-DB), with the aim to explore a) relation of acoustic measurements and individual emotions in light of the bio-informational dimensions (BID) theory, and b) inter-correlations among the acoustic measurements. Results show consistent spectral slope measurements with the size projection hypothesis for happiness and anger, but complex patterns for other emotions. The inter-correlation analysis made three surprise findings. The first was high correlations between median pitch and spectral slope measurements, which was likely related to the Lombard effect. The second was a high correlation between mean intensity and median pitch, which is likely due to data collection procedures in developing the corpus. The third finding was a negative correlation between median pitch and jitter, which was mainly attributable to creaky voice in sadness. These findings may have significant implications for further research on the phonetics of emotional speech.

Keywords: acoustic parameters, bio-informational dimensions theory, correlation analysis, emotional expressions

1. INTRODUCTION

The identification of acoustic correlates of emotions in speech has been difficult, despite continued interest. A major source of difficulty is a lack of proper theorical grounding [17, 18]. Most investigations tend to be data-driven and descriptive, and this is true of both studies that examine acoustic measurements of emotional speech [13, 18, 22, 23, 24] and those that use emotion perception to evaluate synthetic manipulations of acoustic parameters [4, 12]. Also, studies that do try to apply popular psychological models of emotion, such as the dimensional theories, have been unable to identify acoustic correlates of the most theory-essential components such as valence (pleasant/unpleasant) [8, 10]. In the face of this difficulty, an alternative approach to the phonetics of emotion has been gaining ground, namely, the Morton-Ohala hypothesis. This approach posits functional connections between the acoustic properties of emotional vocalizations and the emotional categories and dimensions that conventional approaches have been unable to identify.

1.1. The Morton-Ohala hypothesis

The hypothesis consists of Morton's motivation-structural rules [11] and Ohala's frequency code [15]. It posits that emotional expressions are evolutionally adapted to influence the listener for the benefit of the vocalizer. According to Morton, animals' aggressive calls are for intimidating the hearer by exaggerating the body size of the caller with low pitch and harsh vocal quality, and submissive and fearful calls for making appeasement with raised pitch and pure-tone like vocal quality [11]. Ohala extended this size projection principle to human speech and added vocal tract length as a further indicator of body-size, proposing, in particular, that the smile is for shortening the effective length of the vocal tract to appease the listener [15].

Body-size projection has received support from studies on animal calls and social attributes of human voice, such as gender, attractiveness, etc. [1, 7, 16], but its application for emotional speech has been slow. An important reason is the conflicting findings about the relation of f_0 with confidence or aggression, [15, 17]. Another reason is the lack of evidence for emotion-specific intonation patterns [15] based on frequency code [2, 19]. More recently, however, direct evidence of size projection is found by treating emotional cues as parallel to (i.e., independent of) the phonetics of linguistic contrasts [6, 14, 27, 28]. Nevertheless, body-size projection alone still does not seem enough. For one thing, the conflicting f_0 patterns for anger and dominance remain unexplained. For another, synthetic manipulations based on size-projection alone do not generate speech that sounds emotionally charged [6, 14, 26], which suggests that other factors may also be involved. This has led to the bio-information dimensions (BID) theory [27].

1.2. Bio-informational dimensions

The bio-informational dimensions (BID) theory adopts the core assumption of the Morton-Ohala hypothesis, namely, non-human animals and humans alike use specific acoustic cues in their vocalizations to influence the listener in ways that may benefit the vocalizer. But BID goes beyond body-size projection by adding three further dimensions not directly related to body size, as follows.

The *dynamicity* dimension controls how vigorous the vocalization sounds, depending on whether it is beneficial for the vocalizer to appear strong or weak.

The *audibility* dimension controls how far a vocalization can be transmitted, depending on whether and how much it benefits the vocalizer to be heard over long distance.

The *association* dimension controls associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones.

There has been initial evidence for the dynamicity dimension from synthesis-perception paradigm [14, 27, 28], but not yet for audibility and association. And there has been no examination of whether evidence of these dimensions can be observed from acoustic measurements of emotional speech. This study is the first attempt to test BID in a production corpus, with the goal to explore a) relation of acoustic measurements and individual emotions in light of the bio-informational dimensions (BID) theory, and b) inter-correlations among the acoustic measurements.

2. METHOD

2.1. The corpus

The Berlin database of emotional speech (Emo-DB) is a widely used corpus consisting of 10 sentences spoken by 5 male and 5 female German speakers of different ages in seven emotions: anger, happiness, fear, disgust, sadness, boredom and neutral [4]. All the sentences have been shown to be highly recognizable for the intended emotions. The corpus is also fully annotated for syllables and phones. For this study, we used the phone labels and extracted all measurements only from the vowel segments.

2.2. Measurements and analysis

Sixteen BID measurements were taken by ProsodyPro, an interactive Praat script developed for large-scale analysis of speech prosody [25]. We first ran ProsodyPro to create a tier in which only vowel intervals were labeled to allow the output of the measurements. Then, all measurements in each sentence were averaged to get a single set of mean measurements. These data were then analyzed using R.

3. RESULTS

Due to space limitations, only the most significant results are discussed in this paper.

3.1. Emotion-specific analysis

Tables 1 displays mean values of all the spectral slope parameters that reflect voice quality. As can be seen, all these measurements indicate less spectral tilt in anger than in happiness: lower h1-h2, H1-A1, H1-A3, EB1000, EB500, Ham, but higher COG. A gentle spectral slope suggests a tense voice, which is consistent with the harsh voice suggested by Morton for aggressive animal calls and findings of earlier studies [14, 28]. In fact, all these parameters also show that anger has the smallest spectral tilt across all the emotions. Oddly, however, except for h1-h2 relative to disgust, happiness has the second flattest spectral slope among all the emotions. The reason for this will become clear in the next section on the correlation analysis. Relatedly, all these parameters show the steepest spectral slope for sadness.

Table 1. Top rows: Mean values of voice quality measurements of 7 emotions, where h1-h2 stands for Amplitude difference between 1st and 2nd harmonics, H1-A1 and H1-A3 stand for Amplitude difference between 1st harmonic and 1st and 3rd formants, EB500, EB1000 stand for energy below 500 and 1000 Hz, Ham stands for Hammarberg index, and COG stands for center of gravity. Bottom rows: ANOVA results with measurements as dependent variables, emotion as independent factor, and sentence and speaker as random factors.

	h1-h2 (dB)	H1-A1 (dB)	H1-A3 (dB)	EB_ 500	EB_ 1000	Ham (dB)	COG (Hz)
Happiness	3.0	1.9	14.5	0.49	0.73	12.8	866
Anger	1.3	0.9	10.7	0.34	0.6	9.3	1171
Disgust	2.9	4.0	21.4	0.65	0.84	15.9	663
Sadness	8.9	15.1	32.0	0.89	0.95	21.7	302
Fear	7.9	9.8	23.0	0.68	0.84	15.0	643
Boredom	4.0	7.5	27.5	0.78	0.93	20.0	422
Neutral	4.6	7.3	26.4	0.73	0.91	18.7	474
p	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05
F	10.02	28.09	31.06	46.17	27.88	22.5	28.34
df	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62

In Table 2, the first two columns show that both formant dispersion measurements indicate longer vocal tract in anger than in happiness, consistent

with previous findings from synthetic manipulations [6, 14, 27, 28]. The significant effect of jitter seems to be purely due to the high value of sadness, which again will be revisited in the correlation analysis.

For fear, it differs significantly from neutral emotion only in median pitch based on paired t-test (t(df) = 6, 62, $p < 0.05$). This is consistent with previous reports of high pitch in fear [9, 18]. For disgust, only duration is significantly different from neutral emotion (t(df) = 6, 62, $p < 0.05$), while median pitch is not lower than neutral emotion as previously reported [23].

Table 2. Top rows: Mean values of formant dispersion, median_pitch, mean_intensity, duration, jitter, shimmer, Harmonicity of 7 emotions. Bottom rows, ANOVA results with measurements as dependent variables, emotion as independent factor, and sentence and speaker as random factors.

	FD1_3	FD1_5	Median pitch	Mean intensity (dB)	Duration (ms)	Jitter	Shimmer	Harm
Happiness	1090	879	254	75.3	73.7	0.02	0.1	11.0
Anger	1064	907	261	73.6	79.0	0.02	0.12	9.6
Disgust	1115	898	180	76.8	85.8	0.02	0.1	10.3
Sadness	1172	863	136	78.0	67.7	0.06	15.5	26.6
Fear	1140	873	225	77.0	59.0	0.02	0.11	10.8
Boredom	1150	912	149	77.6	82.0	0.03	0.07	13.1
Neutral	1136	881	152	78.1	66.4	0.03	0.1	11
p	<0.05	>0.05	<0.05	<0.05	<0.05	<0.05	>0.05	>0.05
F	4.31	0.19	12.65	10.26	12.95	15.7	1.69	1.59
df	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62

3.2. Correlation analysis

We have examined correlations between all pairs of measurements, with either individual utterances or specific emotions as raw data. Three sets of correlations were found to be particularly interesting. The first is the high correlation between median pitch and spectral slope measurements, as shown in Table 3. As can be seen, the only exceptions are h1-h2 and H1-A1, which measure the low frequency portions of the spectrum. An example of the scatter plots is shown in Figure 1 for median pitch over center of gravity (COG), which has the highest R^2 value.

Table 3. Slope and R2 values of correlations between median pitch and spectral slope measurements. Slope is the coefficient of the correlation line.

	h1-h2	H1-A1	H1-A3	EB_500	EB_1000	Ham	COG
Slope	-1.41	-5.61	-5.34	-241	-337	-8.93	-0.15
R^2	0.01	0.19	0.42	0.55	0.53	0.44	0.56

These high correlations mean that the interpretation of spectral slope as an indicator of voice quality needs to take median pitch into consideration. For example, the voice quality of an emotion can be judged as extra breathy or extra tense only if it significantly deviates from the correlation line. In Figure 1b, for instance, the mean COG of happiness is located on the left of the correlation line, whereas that of anger is located on the right of the correlation line. This suggests that the voice of happiness is truly breathy while that of anger is truly tense. The same principle is applicable for other emotions. For example, sadness has the lowest COG among all the emotions, which might indicate a breathy voice, consistent with previous reports [5, 9]. But the fact that it also has the lowest median pitch (Figure 1b) raises the question how much true breathiness is involved in sad voice.

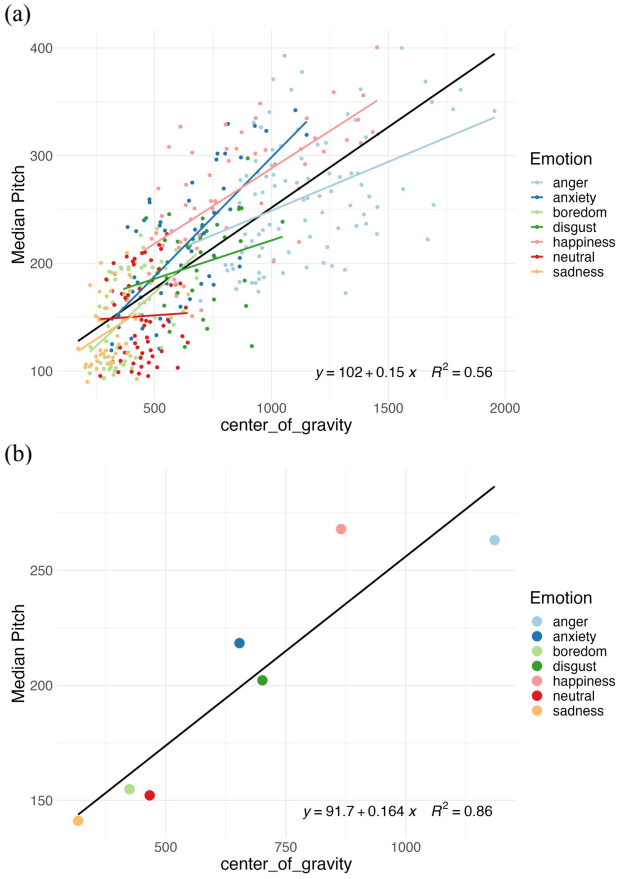


Figure 1. Correlation of COG and median pitch. In (a) each utterance contributes one data point. In (b) each emotion contributes a data point.

Table 4. Slope and R2 values of correlations between mean intensity and spectral slope measurements. Slope is the coefficient of the correlation line.

	h1-h2	H1-A1	H1-A3	EB_500	EB_1000	Ham	COG
Slope	0.19	0.18	0.19	7.19	11.2	0.31	0.005
R^2	0.07	0.13	0.34	0.30	0.36	0.34	0.36

The second interesting sets of correlations are a fairly close relation between mean intensity and spectral slope measurements, as shown in Table 4. Again, an example of the scatter plots is shown in

Figure 2 for mean intensity over center of gravity (COG), which has one of the highest R^2 values. These high correlations are dubious, however, because the emotions with the highest mean intensity are neutral, sadness and boredom, while those with the lowest mean intensity are anger and happiness. Because measured intensity is closely related to recording conditions, we checked the original report [4] on how the corpus was recorded and noticed a number of things. First, the speakers "were instructed not to shout to express anger and to avoid whispering while expressing anxiety," where anxiety = fear. Second, "the recording level had to be adjusted between very loud speech (mostly anger) and very quiet speech (mostly sadness)." Probably as a result, the maximum sound level varies very little across the 7 emotions in the corpus, with the means ranging from 1.008 (boredom) to 1.026 (fear) Pascal, with a standard deviation of 0.005 Pascal. It is likely, therefore, that emotions with relatively flatter spectral tilt have been unproportionally reduced in intensity.

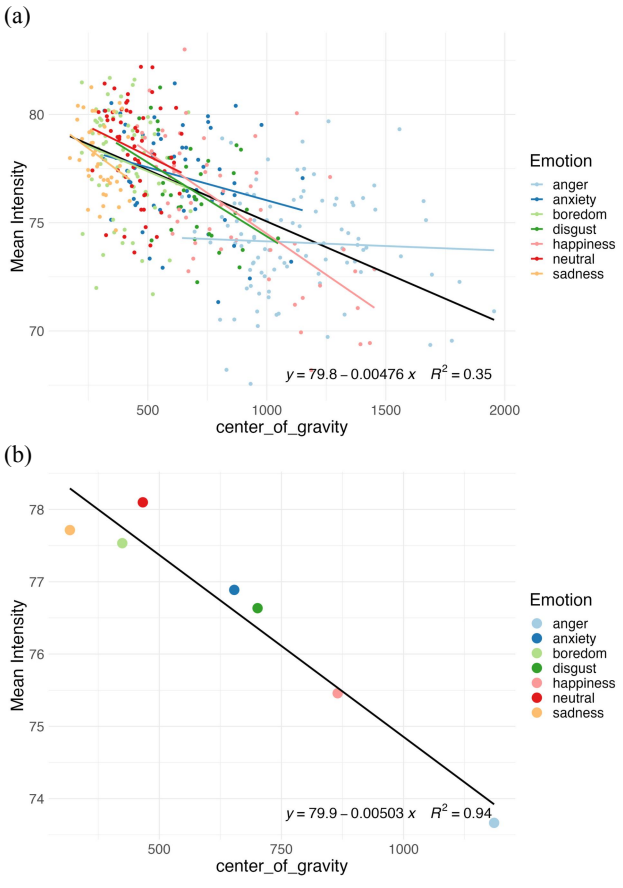


Figure 2. Correlation of center of gravity and mean intensity. In (a) each utterance contributes one data point. In (b) each emotion contributes a data point.

The third interesting correlation is a negative relation between jitter and median pitch, as shown in Figure 3. As can be seen, the negative overall correlation is mainly due to the high jitter values of sadness whose median pitch is very low, which can also be seen in Table 2. One possibility based on our listening impression is that many of the sad utterances have creaky voice, which may have given rise to the high jitter values. In contrast, the jitter values in fear are not high, contrary to some previous reports [9].

4. DISCUSSION AND CONCLUSIONS

A major finding of a detailed acoustic analysis of Emo-DB is the clear contrast between anger and happiness in a) voice quality as shown by all the spectral slope measurements, indicating harsh voice in anger and breathy voice in happiness, and b) vocal tract length as shown by formant dispersion 1-3. These results are consistent with the prediction of body-size projection based on the Morton-Ohala hypothesis [11, 15]. The relatively flat spectral slope and high median pitch in both of these emotions also suggest high orders of dynamicity and audibility, but this is made inclusive given the surprise finding of the correlation analysis.

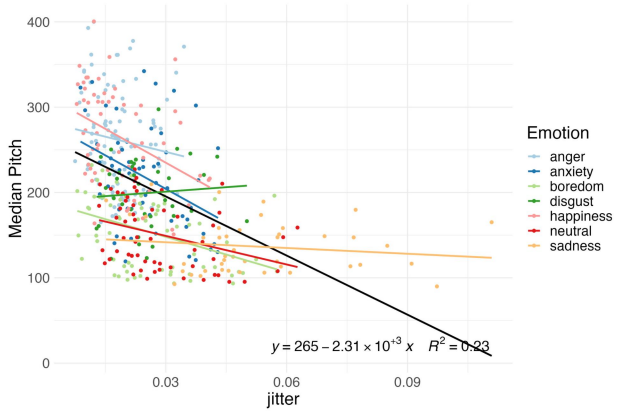


Figure 3. Correlation of jitter and mean median pitch. Each data point from a single utterance.

From the correlation analysis, the first main finding is the close relations between median pitch and spectral slope. A likely source of this relation is related to the Lombard effect, under which pitch and amplitude of a voice are both automatically raised by a noisy environment [3]. Emotions like hot anger and happiness may also involve the same voice raising mechanism, because both presumably require increased vocal effort. So the high pitch often found in anger [9, 15, 18] is likely related to the dynamicity or audibility dimension. Despite the effect, body-size projection is still effectively achieved by spectral slope and formant dispersion, as shown Tables 1 and 2.

The second correlation finding is the close relation between mean intensity and median pitch, which is likely due to the data collection procedures in

developing Emo-DB [4]. The procedure may have neutralized much of the amplitude differences across the emotions. Cautions are therefore needed when evaluating loudness or vocal effort in this and other corpora applying similar recording strategies.

Finally, the negative correlation between median pitch and jitter is likely due to creaky voice in sadness. Equally significant is the lack of high jitter in fear, which is often believed to be associated with a trembling voice [20, 21]. In fact, the general lack of variability of jitter except for sadness questions the usefulness of this measurement for emotional speech.

7. REFERENCES

[1] Anikin, A., Pisanski, K., Massenet, M. and Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B: Biological Sciences*.

[2] Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication* **46**: 252-267.

[3] Brumm, H. and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* **148**(11-13): 1173-1198.

[4] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B. (2005). A database of German emotional speech. In *Proceedings of Ninth European Conference on Speech Communication and Technology*

[5] Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In ISCA Tutorial and Research Workshop (ITRW) on speech and emotion.

[6] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2008). Encoding emotions in speech with the size code. *Phonetica, 65*(4)*, 210-230.*

[7] Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M. and Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour* **69**: 561-568.

[8] Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America, 128*(3), 1322-1336.

[9] Juslin, P. N. (2013). Vocal affect expression: problems and promises. *Evolution of emotional communication*: 252-273.

[10] Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion* **23**(2): 209-237.

[11] Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist, 111*(981), 855-869.

[12] Mozziconacci, S. J. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction, 11*(4), 297-326.

[13] Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America, 93*(2), 1097-1108.

[14] Noble, L., & Xu, Y. (2011). Friendly Speech and Happy Speech-Are They the Same?. In *ICPhS* (pp. 1502-1505).

[15] Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1-16.

[16] Puts, D. A., Gaulin, S. J. C. and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior* **27**(4): 283-296.

[17] Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin* **99**: 143-165.

[18] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms *Speech communication, 40*(1-2)*, 227-256.*

[19] Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004***:** 359-366.

[20] Schuller, B., Wöllmer, M., Eyben, F. and Rigoll, G. (2009). The role of prosody in affective speech. *linguistic insights, studies in language and communication*: 283-305.

[21] Shaver, P., Schwartz, J., Kirson, D. and O'Conner, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* **52**: 1061-1086.

[22] Shami, M., & Verhelst, W. (2007). An evaluation of. the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech communication*, 49(3), 201-212.

[23] Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9), 1162-1181.

[24] Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. The journal of the acoustical society of America, 52(4B), 1238-1250.

[25] Xu, Y. (2013). ProsodyPro—A tool for large scale. systematic prosody analysis. Laboratoire Parole et Langage, France.

[26] Xu, Y. and Kelly, A. (2010). Perception of anger and happiness from resynthesized speech with size-related manipulations. In *Proceedings of Speech Prosody 2010*, Chicago

[27] Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. *Prosody and iconicity*, 33-60.

[28] Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE*, *8*(4), e62397.

Modality Effects on Emotion Perception in English by Chinese L2 English Users: An Eye-tracking Study

Xiazhen Liu[1], Weijing Zhou[2], Francis Nolan[3]
[1,2]Yangzhou University, [3]University of Cambridge
[1]xiazhenl1998@gmail.com, [2]zhouwj@yzu.edu.cn, [3]fjn1@cam.ac.uk

ABSTRACT

This eye-tracing study probes into the extent to which 3 modalities (Audio-only, AO; Visual-only, VO; Audio plus Visual, AV) influence Chinese L2 English users' (N=32) judgement of 6 emotions (happiness, sadness, anger, disgust, fear & neutral) expressed by native English speakers. Two parameters of data, i.e., eye movement (EM) and accuracy rate (AR) were measured to explore the relationship between presentation modality and emotion perception. The results show that for EM, all the emotions generated the same hierarchical rank in terms of Areas of Interest (AOI) under V and AV conditions (eyes > nose> mouth) in spite of salient variations between emotions, and that for AR, the participants in AV condition significantly outperformed themselves in both V and A conditions when judging all the emotions, although there were no significant differences in their distinguishing 3 emotions (anger, sadness and neutral) under 3 modalities, both demonstrating some universal and cultural-specific features.

Keywords: modality effects, emotion perception, Chinese L2 English users, eye-tracking

1. INTRODUCTION

Emotion perception has been a major topic in Natural Language Processing and human-computer interaction. Previous studies of emotion perception revealed that the presentation modality of emotions is a key relevant factor that unduly influences emotion processing [1, 2, 3], i.e., perceivers have different emotion perception performance when the emotions are presented via different modalities (visually, aurally or audio-visually). Furthermore, perceivers could identify the types of emotion in their native language most accurately when emotion is displayed through visual rather than auditory stimuli, i.e., native perceivers generally have a preference or bias to visual modality in emotion recognition [4, 5, 6, 7]. Yet, such a modality preference could be reshaped by perceivers' native linguistic or cultural background [1, 8, 9]. Therefore, to further investigate modality effects, it is necessary to take perceivers' linguistic or cultural background into consideration. In addition, emotion type has been considered as an effective factor in speech emotion perception, supported by the evidence that the identification performance differs between emotions of difference valence [10, 11, 12, 13]. In other words, emotion recognition is not only associated with how emotions are presented, but also correlated to which emotion is presented.

Universally speaking, the essential purpose of human communication is to convey ideas, or express emotions or both on the daily basis. With the increasing globalization and amazing advance of cross-space human communication science and technology, real or virtual, cross-cultural cooperation and collaboration has been a world-wide reality for all the countries. In this very international context, speech emotion in English as a Lingua Franca (ELF) is no doubt of significance both for the research on cross-cultural ELF communication and for the pedagogy of ELF speech learning and teaching in China.

Based on the potential roles of presentation modality and emotion type in processing speech emotions and the well-acknowledged advantages of eye-trackers in detecting human emotion, the present study adopted an eye-tracking diagram to explore the effects of modalities on the perception of speech emotions in English among Chinese L2 English users so as to meet the needs of research and instruction on cross-cultural ELF communication.

2. METHODOLOGY

2.1. Participants

Thirty-nine university students (graduates: 70%, undergraduates: 30%,) were recruited to participate the present study. They spoke Mandarin as their L1 and English as L2, and had normal or corrected-to-normal vision and no hearing impairments. One participant did not complete all the tasks and two other participants did not pass calibration before the experiment. Therefore, thirty-six of them in fact completed the experiment (18 females, 18 males). This sample size ensured proper counterbalancing of the experimental blocks. They used English frequently (all above 10 hours every week, M=24.89, SD=15.05). As their scores of National College English Tests and overall scores of self-reported English proficiency (including listening, speaking, reading, writing and translation) were excellent, they

could be assumed as advanced L2 English users among the population of college students in China.

2.2. Stimuli

As the stimuli used in lots of previous studies were quite often static, e.g., just pictures of speech emotions [14], they were quite less informative or quite inconsistent with real-life human interaction situations, where both facial and vocal organs are simultaneously and dynamically involved in natural and authentic communication. Again, the majority of previous studies of speech perception focused only on two contrastive emotions, e.g., positive vs. negative or happy vs. sad [15], which limited the research scope and depth of speech perception research to a great extent, as emotions conveyed in human speech are quite various in its types to mirror complicated psychological temperatures and convey authentic speech functions. To keep pace with the progress of emotion perception, the present study used dynamic auditory and visual spoken expressions of 6 basic emotions, i.e., happiness, sadness, fear, anger, disgust, and neutral [1] taken from Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [16]. As is shown in Table 1, one Caucasian male and female speaker were selected from each of the three age groups respectively. Their videos included three modalities, audio-only (AO), visual-only (VO), and audio-visual (AV). The experiment thus included a total of 108 different stimuli for each participant (6 speakers×6 emotions×3 modalities). The semantically neutral sentence "*I wonder what this is about*" was used as the target sentence to convey 6 emotions so as to control the potential influence of semantic meanings.

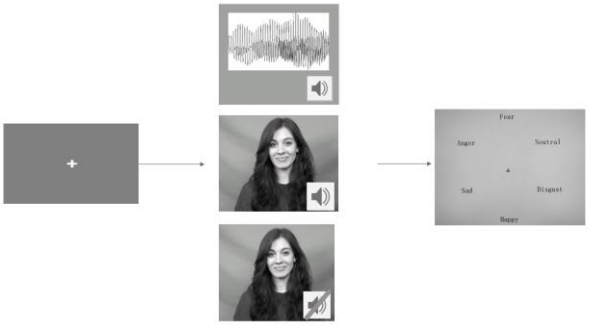
Speakers	Sentence	Modalities	Emotions
Speaker	I wonder what this is about.	AV	happiness, sadness, fear, anger, disgust, neutral
		VO	happiness, sadness, fear, anger, disgust, neutral
		AO	happiness, sadness, fear, anger, disgust, neutral

Table 1: The example of stimuli presented

2.3. Procedure

Experiments and data collection was conducted in conjunction with eye tracking. In each trial, the fixation point was presented first before the stimuli to be presented in either AV, VO, or AO modality. Then the response screen with labels indicating the emotion types appeared and remained on the screen until the participants made a response (See Figure 1). Participants were asked to identify the emotion type of the stimuli and click their mouse on the screen to make the choice. The order of all three blocks (AO, VO, AV) was counterbalanced among participants. Stimulus order was randomized in each block.

Figure 1: Order of experiment events.

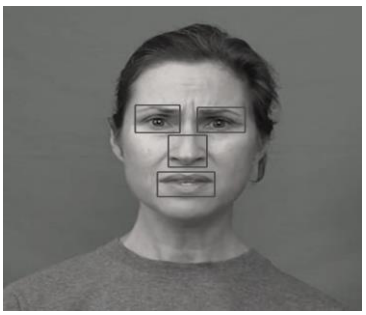


Calibration was performed before the formal experiment started and only the participants who had passed calibration continued to conduct the experiment. They were encouraged to take a self-paced break between each block, and calibration would be again performed after the break. A drift correction was performed before each trail.

2.4. Data collection and analysis

AOI-based fixation proportion (hereafter FP) was analyzed to investigate perceivers' weight of each facial cue in emotion perception [15] so as to test whether the participants had different preferences on certain face regions for different conditions. Since there were no visual stimuli during audio-only block, only data from visual block and audio-visual block were collected. Face regions including eyes, nose, and mouth are dynamic areas of interest (Figure 2). Only fixations located within these areas were recorded. In addition to FP values, accuracy rate (AR) of emotion identification were generated and compared as an indicator of modality preference.

Figure 2: An example for AOI.



Statistical analysis and visualization were conducted in R [17]. Mixed leaner models from lme4 [18] with emotion types, modality being fixed factors which were allowed to interact, and participants and task items being random factors. The final model was:

(1) FP~ Modality*Emotion type + (1+Modality|Participant) + (1+Modality | Item).

The likelihood-ratio test was used to get *p*-values to better illustrate significance. Repeated measures ANOVA from rstatix [19] were used with modality (6 levels) and emotion types (2 levels, i.e., AV vs. VO) being independent variables and AR being the dependent variable.

When there was a significant main effect, Tukey's post hoc tests were conducted to assess the difference between modality pairs. When significant interactions were observed, simple effect analysis was conducted at each level of experiment conditions, and at last, Bonferroni-adjusted pairwise comparisons were performed to further evaluate interactions between modality and emotion types for each modality when necessary.

.

3. RESULTS AND DISCUSSION

3.1. Eye movements

As shown in Table 2 and Figure 3 show, the overall FPs of three facial regions. Generally speaking, the eye preference appeared to be obvious with or without auditory cues, and emotion-specific analysis reported the same preference. There was a hierarchical rank with significant difference in terms of perceivers' gaze allocation to identify emotions, eyes > nose > mouth ($F_{(2, 8205)} = 288.9$, $p<0.0001$).

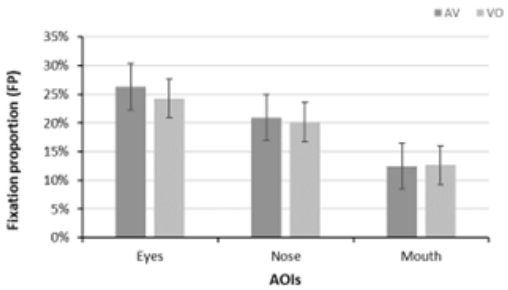


Figure 3: FP of each face region for each modality.

AOI	AV	VO
Eyes	26.31%	24.28%
Nose	20.96%	20.17%
Mouth	12.48%	12.62%

Table 2: The overall FP to three facial regions

As shown in Table 3, the mixed linear model showed a significant main effect of modality on FP ($\chi^2_{(1)} = 4.52$, $p<0.05$), a significant main effect of emotion type as well ($\chi^2_{(5)} = 81.798$, $p<0.001$). No significant interactions or inter-dependence between modality and emotion type was observed ($\chi^2_{(5)} = 1.48$, $p = 0.92$).

Effect	χ^2	df	*p*
Modality	4.523	1	<0.05
Emotion type	81.798	5	<0.001
Modality × Emotion type	1.481	5	0.915

Table 3: The effect of each main and interaction fixed effect.

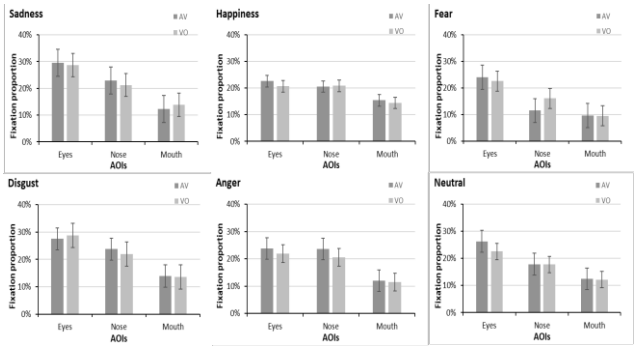


Figure 4: FP under all 12 conditions

As shown in Figure 4, the sub-categorical FPs in 6 emotions displayed the same hierarchical rank with the general pattern above. When emotions were presented in silent environment, in which the degree of processing difficulty increased, FP decreased for eyes ($F_{(1, 2734)} = 2.7$, $p=0.1$) and nose ($F_{(1, 2886)} = 3.428$, $p=0.0642$), while increased for mouth region ($F_{(1, 2582)} = 0.051$, $p=0.839$). When considering effects of emotion type, when identifying sadness, participants' attention devoted to mouth increased in VO (12.23% → 13.85%, $F_{(1, 378)} = 1.18$, $p=0.278$); when processing fear, perceivers' fixation on nose increased when the audio was removed (11.42% → 16.04%, $F_{(1, 454)} = 1.381$, $p=0.241$); when processing disgust, fixation on eyes increased in silent condition (27.49% → 28.75%, $F_{(1, 454)} = 0.3$, $p=0.584$). Although no significance was observed, this implies that perceivers would change their perceptual primacy according to modality condition.

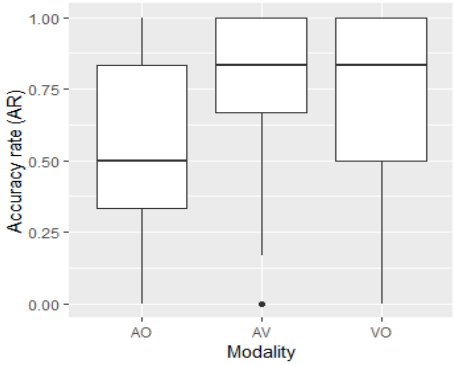
3.2. Accuracy rate

As shown in Figure 5, the overall ARs averaged over all the participants and experiment blocks clarify that AR was the highest for AV, which confirmed the existence of audio-visual integration in non-native context. Also, performance for VO was higher than that for AO; the differences of AR among the three modalities were overall significant, ($F_{(2, 645)} = 33.8$, $p<0.0001$). This demonstrates a visual modality preference, which is consistent with the findings of

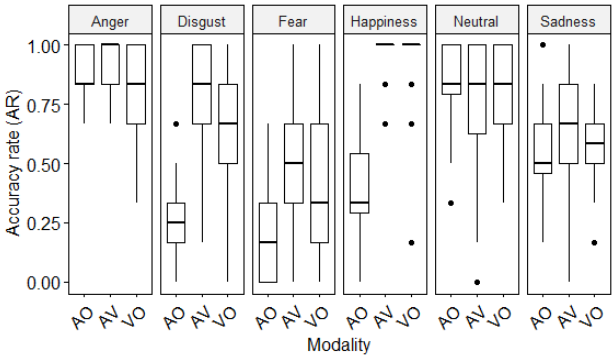
perception research focusing on native perceivers [4, 20, 21], but meanwhile is inconsistent with some studies indicating that people from Eastern cultures would rely more on visual cues compared with auditory ones [8].

Figure 5: AR for each modality.

Results of ANOVA showed a significant main



effect of modality, ($F_{(5, 175)}$ =66.639, p<0.05, ges=0.421), a significant main effect of emotion category, ($F_{(2, 70)}$ =106.643, p<0.05, ges=0.175), and a significant interaction between modality and emotion type, ($F_{(10, 350)}$ =24.263, p<0.05, ges=0.231). To assess the difference between modality pairs, Tukey's HSD test showed significant difference between AO and VO, as well as AV and AO, but no significant difference between AV and VO (AV-AO:



p<0.001; VO-AO: p<0.001; AV-VO: p=0.22), indicating a lower discriminability of AR between AV and VO. It is notable that accuracy results reported significant main effects of modality and emotion type, and a significant interaction between these two factors as well. However, no significant interactions or inter-dependence between modality and emotion type was observed from eye movements. This might imply that combination of task-free and spontaneous data such as eye-tracking is necessary.

Figure 6: AR under all 18 conditions

Emotion	AO-AV	AO-VO	AV-VO
Anger	0.441	0.819	0.069
Disgust	<0.001	<0.001	<0.001
Fear	<0.001	0.008	0.203
Happiness	<0.001	<0.001	1
Neutral	0.17	1	0.414
Sadness	1	1	0.78

Table 4: Results of simple main effect test.

As shown in Figure 6 and Table 4, happiness expressed with muted videos was easiest to be recognized, while fear portrayed through speech alone was the most difficult to be recognized; the effect of modality was significant for disgust, fear and happiness (p<0.05, ges=0.493, ges=0.165 and ges=0.703 respectively), but no significance was observed for anger (p=0.408), neutral (p=0.672) and sadness (p=1). The complete comparison is listed in Table 4 showing Bonferroni-adjusted p value.

4. DISCUSSION AND CONCLUSION

The present study found a visual dominance in L2 users. Research on native speakers suggested that Easterners showed an auditory preference, while Westerners did better in the visual modality [8, 24]. However, against that, the investigation on the modality effect on Mandarin ironic speech showed a better identification rate in the visual-only condition [25]. This inconsistency may imply that the modality effect is related to which emotion is presented and which language group we target.

In terms of the effect of emotion types, research on native speakers suggested that Easterners appear to be less accurate than Westerners in recognizing negative emotions, which could be explained by the cultural differences [26-27]. In this sense, participants in this study showed a combination of both Eastern and Western pattern, as they were less accurate in recognizing disgust and fear, which was a specific Eastern pattern, while their performance in response to anger was best. This could be explained by their long duration of exposure to the L2 culture, which leads to their similar processing behaviours to that of native speakers [9].

In conclusion, this eye-tracing study probes into the extent to which modalities influence Chinese L2 English users' judgement of emotions expressed by native English speakers. Results showed a consistent hierarchical rank of fixation on facial regions in both AV and VO conditions (eyes > nose> mouth), and that the visual cues play a dominant role for L2 learners' emotion perception, although there were no significant differences in their distinguishing anger, sadness and neutral. This study demonstrated some universal and cultural-specific features.

5. REFERENCES

[1] Ekman, P. 1989. The argument and evidence about universals in facial expressions of emotion. In: Wiley, J. (eds). *Handbook of social psychophysiology*. Oxford, England, 143-164.

[2] Lorette, P. & Dewaele, J. M. 2019. The relationship between bi/multilingualism, nativeness, proficiency and multimodal emotion recognition ability. International Journal of Bilingualism, 23 (6), 1502-1516.

[3] Bhatara, A., Laukka, P., Boll-Avetisyan, N, Granjon, N., Elfenbein, H. A. & T. Banziger. 2016. Second Language Ability and Emotional Prosody Perception. *Plos One*. 11.

[4] de Boer, M. J., Baskent, D., & Cornelissen, F. W. 2020. on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli. *Multisensory Research*. 34, 17-47.

[5] Kim, J., & Davis, C. 2012. Perceiving emotion from a talker: How face and voice work together. *Visual Cognition*. 20, 902-921.

[6] Paulmann, S. & Pell, M. D. 2011. Is there an advantage for recognizing multi-modal emotional stimuli?. *Motivation and Emotion*. 35, 192-201.

[7] Vroomen, J. & de Gelder, B. 2000. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *J Exp Psychol Hum Percept Perform*. 26, 83-90.

[8] Liu, P., & Rigoulot, S., & Pell, M. D. 2015. Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*. 67, 1-13.

[9] Chen, P., Chung-Fat-Yim, A. & Marian, V. 2022. Cultural Experience Influences Multisensory Emotion Perception in Bilinguals. *Languages*. 7.

[10] Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L. & Neel, R. 2011. The face in the crowd effect unconfounded: happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *J Exp Psychol Gen*, 140, 637-59.

[11] Becker, M. W. 2012. Negative emotional photographs are identified more slowly than positive photographs. *Attention, Perception, & Psychophysics*. 74,n1241-1251.

[12] Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & R. C. Gur. 2010. "The face in the crowd effect: anger superiority when using real faces and multiple identities. *Emotion*. 10, 141-6.

[13] Pitica, I., Susa, G., Benga, O, & Miclea, M. 2012. Visual search for real emotional faces: the advantage of anger. *Procedia - Social and Behavioral Sciences*. 33, 632-636.

[14] Henderson, J. M., Williams, C. C., & Falk, R. J. 2005. Eye movements are functional during face learning. *Memory & Cognition*. 33, 98-106.

[15] Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Franconeri, S. L. & Franconeri, S. L. 2014. Eye movements during emotion recognition in faces. *Journal of Vision*. 14.

[16] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A. & Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. 2014. *Ieee Transactions on Affective Computing*. 5, 377-390.

[17] R Core Team, 2021. R: A Language and Environment for Statistical Computing (Version 4.0.5). Vienna, Austria. Retried from. https://www.R-project.org.

[18] Bates D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 67, 1-48.

[19] Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://CRAN.R-project.org/package=rstatix.

[20] Regel, S., Gunter, T. C., & Friederici, A. D. Isn't It Ironic? An Electrophysiological Exploration of Figurative Language Processing. 2011. *Journal of Cognitive Neuroscience*. 23, 277-293.

[21] Spotorno, N., Cheylus, A., Van Der Henst, J.-B., & Noveck, I. A. 2013. What's behind a P600? Integration operations during irony processing. *PloS one*. 8, e66839.

[22] Banse, R. & Scherer, K. R. 1996.Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*. 70, 614-636.

[23] Rottman, J. 2014. Evolution, Development, and the Emergence of Disgust. *Evolutionary Psychology*. 12.

[24] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17: 124-129.

[25] Li, S., Chen, A., Chen, Y., & Tang, P. (2022). The role of auditory and visual cues in the interpretation of Mandarin ironic speech. *Journal of Pragmatics*, 201: 3-14.

[26] Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, 108: 291-310.

[27] Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences. *Journal of Nonverbal Behavior*, 21: 3-21.

ACOUSTIC ANALYSIS OF MANDARIN SPEECH IN A MIXED EMOTION OF PLEASURE AND DISGUST

Mengjie Gong[1] and Wentao Gu[1,2]

[1]School of Chinese Language and Literature, Nanjing Normal University, Nanjing, China
[2]School of Communication Sciences and Disorders, McGill University, Montreal, Canada
gmjnew@qq.com, wtgu@njnu.edu.cn

ABSTRACT

It is controversial whether people can experience a mixture of two emotions with opposite valences. Although a number of psychological and physiological studies supported the existence of mixed emotions, acoustic-phonetic characteristics of speech in mixed emotions are yet to be explored. This study investigated a mixed emotion composed of pleasure and disgust. Statistical analysis of acoustic parameters showed that the main difference among positive (pleasure), negative (disgust) and mixed emotions existed in fundamental frequency parameters, harmonic differences, and Harmonic-to-Noise Ratio (HNR), while the intensity parameters, Jitter and Shimmer were not significantly different. Clustering analysis further showed that the three types of emotions could be classified more effectively in terms of the acoustic parameters showing significant difference.

Keywords: mixed emotion; Mandarin; non-parametric statistical analysis; clustering analysis

1. INTRODUCTION

In terms of emotional valence, emotion can be positive or negative. Although a number of studies investigated the acoustic manifestations of speech conveying positive or negative emotions [1, 2], little has been explored for mixed emotions that incorporate multiple emotional states simultaneously.

In the literature, *mixed emotion* has been defined in two distinctive ways. Larsen et al. [3] proposed that emotional experience can be characterized by the concurrent activation of two emotions that are typically opposite in valence, e.g., happiness and sadness. In contrast, Grühn et al. [4] suggested that mixed emotion is an emotional state that lies between two emotions with opposite valences, and values closer to zero indicate a greater emotional complexity. So far, most studies on mixed emotions are based on the first definition, i.e., positive and negative emotions are concurrently elicited. Larsen and McGraw [5] demonstrated the robustness of this mixed emotion as a genuine and distinct emotional experience. Moreover, previous studies have elucidated the psychological and physiological differences between a mixed emotion and a pure positive or negative emotion [6, 7].

However, up to now few research has looked into the acoustic manifestations of the speech conveying mixed emotions. Thus, the present study aimed to investigate prosodic and voice quality characteristics of Mandarin utterances conveying a mixed emotion of "pleasure" (positive valence) and "disgust" (negative valence), in comparison with those conveying a pure pleasure (positive) or disgust (negative) emotion.

2. METHODS

2.1. Materials

A film-watching paradigm was employed to elicit emotions. For this purpose, we selected nine film clips from the database created by Samson et al. [8], and had them evaluated by 27 subjects (age: M=23.9, SD=1.65) for the scores of pleasure and disgust on the respective five-point Likert scale. These film clips were divided into three blocks, and within each block three clips presented in a random order were evaluated by nine subjects.

A mixed emotion coefficient (MF) [9, 10], defined as the minimum of the pleasure and disgust scores, indicates the intensity of the mixed emotion. For example, if pleasure is scored 3 and disgust is scored 4, MF will be 3. When one of the pleasure and disgust scores is 1, the emotion is regarded as a pure positive or negative one. All MF scores above 1 for each clip are shown in Fig. 1, where each data point indicates a subject. Finally, two clips ("hit" and "downstairs") with the highest number of MF scores above 1 (hence most likely to elicit a mixed emotion) were selected.

Also, we designed twelve 7-syllable Mandarin target sentences without any emotional words. Naturalness and emotion-neutrality of the sentences were validated by 15 native speakers of Mandarin.

2.2. Participants

Twenty-nine native speakers of Mandarin (12M, 17F; age: M=24.3, SD=1.5) without any history of hearing, speech, and emotion impairments were recruited.

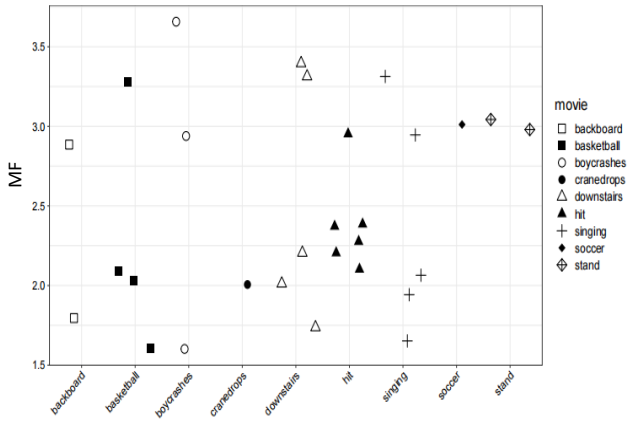


Figure 1: MF scores of film clips.

2.3. Procedure

After watching each film clip, participants scored pleasure and disgust on the respective 5-point Likert scale based on their visual impression of the film clip. Then, they were asked to say 6 sentences (a random selection from the 12 target sentences) with the emotion they just perceived from the film clip. Before watching the next clip, there was a brief break (at least 1.5 min) such that the emotional states of the participants could return to their baseline level.

2.4. Speech data

A total of 6*2*29 = 348 utterances were recorded. Among them, 150, 84, and 108 utterances were associated respectively with the mixed, positive, and negative emotions, based on the pleasure and disgust scores labelled by participants on the film clips. Only 6 utterances were neutral (i.e., both pleasure and disgust were scored 1), which were excluded in subsequent analysis due to lack of tokens.

2.5. Data analysis

The commonly-used acoustic parameters, including fundamental frequency (F0) contour, intensity (INT) contour, H1–H2, H1–A1, H1–A2, Harmonic-to-Noise Ratio (HNR), Jitter and Shimmer, were extracted from each utterance using the Praat toolkit [11]. For F0 and INT contours, the mean, standard deviation, minimum, maximum, and range were further calculated.

All these 16 parameters were statistically analysed with R [12]. Because none of them passed the Shapiro-Wilk normality test, non-parametric Scheirer-Ray-Hare tests were then conducted, with speaker gender and emotion as two independent factors. However, results showed that there was no significant interaction effect between the two factors. Thus, to simplify analysis, non-parametric Kruskal-Wallis tests were conducted instead, with emotion (Pos, Neg, Mix) as the sole independent factor, without taking speaker gender into account.

Table 1: Results of Kruskal-Wallis test on all 16 parameters.

	Median value			χ^2	df	*p*-value
	Pos	Neg	Mix			
F0mean	151.323	209.186	212.136	7.923	2	0.019*
F0std	25.868	42.740	29.812	13.593	2	0.001**
F0min	105.574	107.868	136.946	9.446	2	0.009**
F0max	207.956	269.376	264.974	5.714	2	0.057
F0range	93.604	165.162	112.933	13.622	2	0.001**
INTMean	66.229	65.747	66.552	3.125	2	0.210
INTStd	8.350	8.495	8.109	1.094	2	0.579
INTMin	39.709	38.403	39.549	1.138	2	0.566
INTMax	77.847	78.095	77.412	2.629	2	0.269
INTRange	38.765	39.692	37.367	3.514	2	0.173
Jitter	0.026	0.023	0.020	5.710	2	0.058
Shimmer	0.094	0.071	0.084	4.590	2	0.101
H1-H2	5.072	6.317	7.563	13.603	2	0.001**
H1-A1	17.985	18.492	20.107	7.875	2	0.020*
H1-A2	20.546	17.839	21.814	7.209	2	0.030*
HNR	14.731	18.796	17.917	7.713	2	0.021*

Table 2: Results of post-hoc tests on eight parameters.

	Pos vs Mix	Mix vs Neg	Pos vs Neg
F0mean	*		*
F0std		**	**
F0min	*	*	
F0range		**	**
H1-H2	**		
H1-A1	*		
H1-A2		*	
HNR			*

Note: * $p < 0.05$; ** $p < 0.01$.

3. RESULTS

3.1. Non-parametric tests

Table 1 shows the results of Kruskal-Wallis tests on all 16 acoustic parameters, among which eight parameters exhibited significant effects of emotion. For these eight parameters, Fig. 2 shows the boxplot for three emotions, and Table 2 shows the results of post-hoc tests with Bonferroni correction using the *PMCMRplus* package [13]. The significant results are summarized as follows:

F0mean: Mix > Pos ($p = 0.049$), and Neg > Pos ($p = 0.023$).

F0std: Neg > Mix ($p = 0.005$), and Neg > Pos ($p = 0.004$).

F0range: Neg > Mix ($p = 0.004$), and Neg > Pos ($p = 0.005$).

F0min: Mix > Neg ($p = 0.025$), and Mix > Pos ($p = 0.044$).

H1–H2: Mix > Pos ($p = 0.001$).

H1–A1: Mix > Pos ($p = 0.030$).

H1–A2: Mix > Neg ($p = 0.020$).

HNR: Neg > Pos ($p = 0.026$).

3.2. Clustering analysis

Using the Partitioning Around Medoids (PAM) clustering algorithm which is a robust alternative to K-means, we further conducted clustering analysis with the above eight parameters.

Before PAM clustering, the pamk() function in the *fpc* package [14] was run to find the optimal number of clusters, which turned out to be three, coinciding exactly with the number of emotion types here. The data were then clustered using the pam() function in the *cluster* package [15]. The first and second principal components explained 72.94% of variance. The clustering results based on the two principal components are shown in Fig. 3, where categories 1 and 2 have the most conspicuous overlap, categories 1 and 3 overlap to a much lesser extent, while categories 2 and 3 overlap the least.

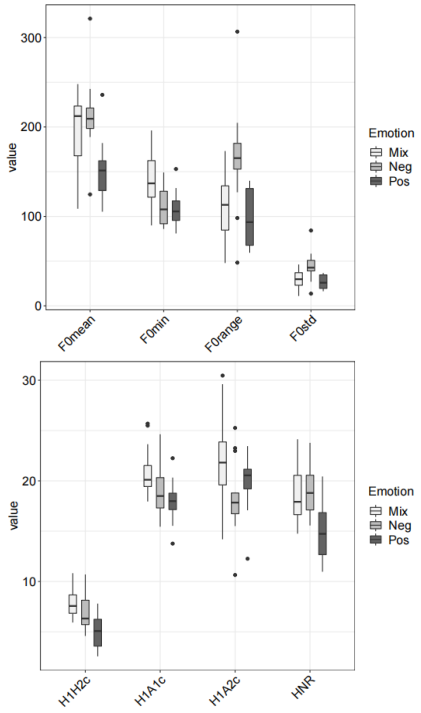


Figure 2: Comparison of eight parameters in three emotions.

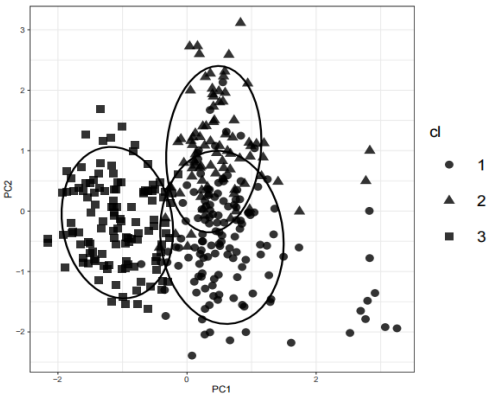


Figure 3: Results of PAM clustering analysis.

Table 3 shows the contingency table for clustering. For one thing, categories 1 and 2 correspond mainly to negative and mixed emotions, respectively, while category 3 corresponds almost equally to positive and mixed emotions. For another, positive and negative emotions are clustered mainly into categories 3 and 1, respectively, while the mixed emotion are distributed almost evenly in all three categories.

Table 3: Contingency table for clustering analysis.

	Emotion		
	Pos	Mix	Neg
Category 1	15	41	78
Category 2	11	56	20
Category 3	58	53	10

4. DISCUSSION AND CONCLUSION

To our knowledge, this study is the first attempt to look into the acoustic characteristics of speech in a mixed emotion. Results showed that the Mandarin utterances conveying a mixed emotion of pleasure and disgust differed significantly from those conveying a pure pleasure or disgust emotion in seven acoustic parameters, including F0 features (F0mean, F0std, F0min, F0range) and harmonic differences (H1–H2, H1–A1, H1–A2), while HNR differed significantly only between pleasure and disgust emotions. Furthermore, clustering analysis verified that with these eight parameters the speech data could be approximately clustered into three categories.

There are still some limitations in this preliminary study. First, we examined a mixed emotion of pleasure and disgust, since previous studies had demonstrated the reliability of this emotion [10, 16]. Nevertheless, future research needs to address whether the conclusions here can be generalized to other mixed emotions [17]. Second, film clip elicitation, as a standard experimental paradigm to stimulate pure or mixed emotions [10], is easy to standardize and can elicit highly ecologically valid emotions [18]. However, film clips may differ in various potential characteristics, may induce multiple emotions sequentially rather than simultaneously, and subjects in this paradigm may become passive observers of events rather than active participants. Therefore, our future study needs also to examine whether the findings here can be generalized to other emotional situations.

5. ACKNOWLEDGEMENTS

This work is supported by the National Social Science Fund of China (13&ZD189).

6. REFERENCES

[1] Colombetti, G. 2005. Appraising valence. Journal of Consciousness Studies 12(8-10), 103-126.
[2] Kensinger, E. 2004. Remembering emotional experiences: The contribution of valence and arousal. Reviews in the Neurosciences 15(4), 241-252.
[3] Larsen, J.T., McGraw, A.P., and Cacioppo, J. T. 2001. Can people feel happy and sad at the same time? J. Pers. Soc. Psychol. 81, 684-696.
[4] Grühn, D., Lumley, M.A., Diehl, M., et al. 2013. Time-based indicators of emotional complexity: Interrelations and correlates. Emotion 13, 226-237.
[5] Larsen, J.T., and McGraw, A.P. 2014. Further evidence for mixed emotions. J. Pers. Soc. Psychol. 100, 1095-1110.
[6] Kreibig, S.D., Samson, A.C. and Gross, J.J. 2013. The psychophysiology of mixed emotional states. Psychophysiology 50(8), 799-811.
[7] Kreibig, S.D., Samson A.C. and Gross, J.J. 2015. The psychophysiology of mixed emotional states: Internal and external replicability analysis of a direct replication study. Psychophysiology 52(7), 873-886.
[8] Samson, A.C., Kreibig, S.D., Soderstrom, B., et al. 2016. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. Cognition and Emotion 30(5), 827-856.
[9] Schimmack, U. 2001. Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive? Cognition and Emotion 15, 81-97.
[10] Hemenover, S.H., and Schimmack, U. 2007. That's disgusting!..., but very amusing: Mixed feelings of amusement and disgust. Cognition and Emotion 21, 1102-1113.
[11] Boersma, P. and Weenink, D. 2022. Praat: Doing phonetics by computer (Version 6.3) [Computer program]. Retrieved 15 November 2022. Available from http://www.praat.org/
[12] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[13] Pohlert, T. 2022. PMCMRplus: Calculate pairwise multiple comparisons of mean rank sums extended. R package version 1.9.6. https://CRAN.R-project.org/package=PMCMRplus
[14] Christian, H. 2020. fpc: Flexible procedures for clustering. R package version 2.2-9. https://CRAN.R-project.org/package=fpc
[15] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2021. cluster: Cluster analysis basics and extensions. R package version 2.1.2.
[16] Oppliger, P.A., & Zillmann, D. 1997. Disgust in humor: Its appeal to adolescents. Humor 10, 421-437.
[17] Harter, S., & Buddin, B.J. 1987. Children's understanding of the simultaneity of two emotions: A five-stage developmental acquisition sequence. Developmental Psychology 23, 388-399.
[18] Rottenberg, J., Ray, R., & Gross, J. 2008. Emotion elicitation using film clips. In J. A. Coan & J. J. B. Allen (Eds.), The Handbook of Emotion Elicitation and Assessment. New York: Oxford University Press, 9-28.

EFFECTS OF ADULT ATTACHMENT STYLE ON EMOTION: EVIDENCE FROM PHYSIOLOGICAL AND ACOUSTIC MEASURES

Han Hu[1] and Wentao Gu[1,2]

[1]School of Chinese Language and Literature, Nanjing Normal University, Nanjing, China
[2]School of Communication Sciences and Disorders, McGill University, Montreal, Canada
huhandon@gmail.com, wtgu@njnu.edu.cn

ABSTRACT

This study investigated how one's adult attachment style affected physiological responses and speech production in different emotional states. We selected affective film clips for each of four basic emotions, and designed pseudo Mandarin sentences (syntactically valid but semantically meaningless). With a paradigm of subliminal priming and elicitation, 44 participants with love experience were physiologically measured while watching the film clips, and then were recorded while producing the pseudo sentences in the emotion perceived from each film clip. Multivariate analysis of three physiological indices showed significant differences among attachment styles only in anger and sadness, while analysis of 14 acoustic measures illustrated significant differences primarily in happiness and sadness. Among the four attachment styles, only the patterns of 'fearful' individuals varied with emotion. Attachment styles were effectively identified from acoustic measures, and redundant analysis of physiological and acoustic measures indicated that the relationship between these two aspects varied with attachment style.

Keywords: attachment style, emotion, physiological index, acoustic measure.

1. INTRODUCTION

Emotion plays indispensable roles in daily life since it affects decision-making, social communication and well-being. Recent years have witnessed numerous studies on physiological and acoustic signals in emotional states [1, 2]. However, no consensus has been reached, and more attention should be paid to the potential factors affecting physiological responses and acoustic manifestations of emotions [3, 4].

The adult attachment theory [5], which is closely associated with emotion regulation in romantic relationships, provides an effective framework to account for those inconsistent findings in literature [6]. The theory posits that early interactions with caregivers shape the Internal Working Models (IWMs) of self and others. The model of self (i.e., attachment anxiety) indicates one's internalized sense of self-worth, and is associated with levels of anxiety and dependency experienced in close relationships. Individuals with high attachment anxiety prefer a hyper-activating strategy of emotion regulation, exaggerating their emotional needs to capture the partners' attention. The model of others (i.e., attachment avoidance) denotes one's general expectation of others as supportive and available, and is related to the propensity to seek or avoid closeness in relationships. Individuals with high attachment avoidance struggle with trust and emotional openness, leading them to prefer a suppression strategy of emotion regulation [6, 7].

Attachment style can be classified into four types in terms of the polarities of IWMs: secure (low anxiety and avoidance), dismissive (low anxiety and high avoidance), preoccupied (high anxiety and low avoidance), and fearful (high anxiety and avoidance). Since the four types differ in emotion regulation strategy which has direct impacts on physiological arousal and emotional expressive consequences [8], attachment style is expected to affect physiological responses and acoustic production in emotional states.

According to physiological studies, dismissive individuals has a high skin conductance level (SCL) when recalling painful memories [9], whereas secure individuals are associated with a low ratio of low-frequency to high-frequency power (LF/HF) in electrocardiogram (ECG) when they are rejected [10]. As reported in acoustic studies, fundamental frequency variability is negatively correlated with attachment avoidance when producing neutral words [11], while jitter and shimmer are positive predictors of attachment dimensions while producing sustained vowel /a/ [12]. An effective classification of four attachment styles based on intimate speech was also attained with nine acoustic parameters [13].

Although many previous studies examined the effects of attachment style or emotion on physiological responses or acoustic manifestations in speech production, few of them looked into the combined effect of attachment style and emotion, and the relationship between physiological responses and acoustic manifestations has yet to be explored. Hence, we aimed to address these issues using multivariate analysis, to find the effects of attachment style on experience and expression of four basic emotions.

2. METHODS

2.1. Materials

2.1.1. Emotional film clips

Twenty film clips were selected from the Chinese Emotion Video Stimuli (CEVS) database [14], with five clips evoking each of the four basic emotions (happiness, anger, sadness and fear). Fifteen graduate students (12F, 3M; age: 25.33±2.82 years) assessed the film clips using a 5-item self-report inventory (happiness, anger, sadness, fear and neutrality) on a 5-point Likert scale, from 1 (weak) to 5 (strong). For each emotion, the film clips ranked top three (in terms of the number of participants giving the highest score to the emotion) were selected. Thus, we obtained 12 film clips with an average duration of 153.58 s (*SD*=57.87 s).

2.1.2. Vocal emotional stimuli

Forty pseudo-sentences were designed by replacing the content words in real sentences with a meaningless random combination of Chinese characters, while keeping the function words intact to convey the syntactic information [15]. To ensure the acceptability of the pseudo-sentences, the 15 students mentioned in section 2.1.1 evaluated the degree of "language-likeness" on a 5-point Likert scale (1: very unlike; 5: very like), and the score of each pseudo-sentence was averaged from all evaluators. Hence, 15 pseudo-sentences giving the highest scores (*M*=4.09, *SD*=0.36) were selected, with a length of 7–12 syllables (*M*=8.8 syllables).

2.2. Participants

We recruited 44 participants (28F, 16M; age: 24.30±3.08 years; 25 in relationships) who were self-reported heterosexual with love experience (times of relationships: 2.02±0.98). Among the participants in relationships, 56% were in long distant relationships. The mean relationship stage was 2.88 on a 5-point Likert scale (1: boring; 5: passionate). All participants spoke Mandarin fluently without any reported history of hearing, speech or emotion impairments. Among them, there were 27% secure, 16% dismissive, 18% preoccupied and 39% fearful individuals.

2.3. Measurement

2.3.1. Self-report inventories

Experiences in Close Relationship Questionnaire (ECR) in Chinese [16] was used to assess participants' attachment styles. ECR consisted of two subscales, i.e., attachment avoidance and attachment anxiety, each including18 items. Each item was rated on a 7-point Likert scale, from 1 (highly disagree) to 7 (highly agree), and then each subscale was scored by averaging the scores from the 18 items. Attachment type was derived from these scores. The McDonald's ω's were 0.89, 0.83, and 0.86 for avoidant subscale, anxious subscale, and the whole scale, respectively.

2.3.2. Physiological indices

Skin Conductance Levels (SCL) in microsiemens (μS) were measured with the Galvanic Skin Response (GSR) Amp (ADInstruments Ltd.) at a rate of 1 kHz. A pair of bipolar finger electrodes (snap lead) made of polished stainless steel were attached to the medial phalanges of digits II and IV of the participant's non-dominant hand. SCL is a pure measure of the arousal of the sympathetic nervous system (SNS) [17].

Heart Rate Variability (HRV) was derived from the finger ECG signals through an electric finger transducer attached to the distal phalanges of the middle finger of the participant's non-dominant hand. Root Mean Square of Successive Difference (RMSSD) between normal heartbeats, as a measure of the short-term fluctuation of HRV, represented the activity of the parasympathetic nervous system (PNS). LF/HF, as the ratio of spectral power in low frequency bands (0.05 Hz-0.15 Hz) to high frequency bands (0.15 Hz-0.4 Hz), denoted the sympathovagal balance [18].

2.3.3. Acoustic parameters

For each validated utterance, the following prosodic and voice quality parameters were extracted using the Praat toolkit [19]: fundamental frequency (F0, in semitone), intensity (INT, in dB), articulation rate (ARTRATE, in 1/s), corrected harmonic differences (H1–H2, H2–H4, in dB) [20], Harmonic-to-Noise Ratio (HNR, in dB), Jitter and Shimmer (both in %). For F0 and INT, the descriptive statistics including mean (M), standard deviation (SD), minimum (MIN) and maximum (MAX) were calculated.

2.4. Procedure

Physiological signals were collected through PowerLab 16/35 (ADInstruments). Speech signals were recorded via a high-quality head-mounted microphone connected to Zoom H4N portable digital recorder sampled at 44.1 kHz with a 16-bit precision.

Firstly, participants were invited into a soundproof booth after signing the informed consents. They were seated comfortably in front of a monitor at a 70 cm distance. After participants were familiarized with the

pseudo-sentences, they said the sentences in a neutral style with a head-mounted microphone 10 cm in front of their mouths and biosensors firmly attached to their fingers. After relaxing for 3 min to get baseline measurements, participants took a short training session using a neutral-emotion film clip from CEVS.

Next, each trail contained a random 20 ms subliminal presentation of one of the two primes ("separation" and "failure"), paired with two 500 ms grid images as anterior and posterior masking stimuli before film-watching, which proved efficient at activating IWMs [21]. The film-watching task was divided into 4 blocks in terms of emotion, with presentation order counterbalanced among participants. Each block included a random presentation of three film clips on the monitor, followed by five random sentences. Participants were required to say these sentences to their partners in the emotion they perceived from the clips. Breaks between blocks ensured participants to return to the baseline level, and a 30 s blank was presented before each block to get the baseline level. There were 75 productions (15 items × 5 emotions) per speaker.

Finally, participants filled in ECR inventories and demographic information. The entire procedure took about 1 hour and all participants were remunerated.

3. RESULTS

All above measures were normalized to reduce inter-speaker variation, using $X_{norm} = \frac{X_{ijk} - X_{refjk}}{X_{refjk}}$, where i, j, k denote observation, measure and participant, respectively, X_{ref} refers to the value at the baseline level. The average and minimum values at the baseline level prior to film-watching served as X_{ref} for HRV and SCL, respectively. F0- and INT-related indices were normalized, with X_{ref} as the mean F0-MIN and INT-MIN over all neutral utterances [14]. Other acoustic measures were normalized in reference to the mean values in the neutral state. All normalized measures were converted into z-scores for statistical analysis in R [22].

3.1. Physiological indices

Permutational Multivariate Analysis of Variance (PMANOVA) was separately conducted to geometrically partition multivariate variation in the space of Euclidean distance measure according to a given ANOVA design, with attachment style as between-subjects factor and p-values obtained through 1,000 permutations, for each emotion. The difference among four attachment styles was significant only in sadness (F=3.29, p=0.027, η^2=0.20), and a marginal significance was reached in anger (F=2.12, p=0.080, η^2=0.14).

Spatial representation of attachment styles in anger and sadness with Nonmetric Multidimensional Scale (NMDS) is illustrated in Fig. 1. NMDS is an indirect gradient technique that ordinates objects in a low-dimensional space with dissimilarity matrix. The goodness of fit was assessed by stress which quantifies the discrepancy between the distances in the original and ordination space. The stress values were 0.022 for anger and 0.032 for sadness in a two-dimensional solution, indicating a good fit to the empirical data [23]. For both emotions, the 1st dimension can separate secure individuals from dismissive and preoccupied ones. The location of fearful individuals, however, varies with emotion: closer to preoccupied ones in anger and closer to secure ones in sadness.

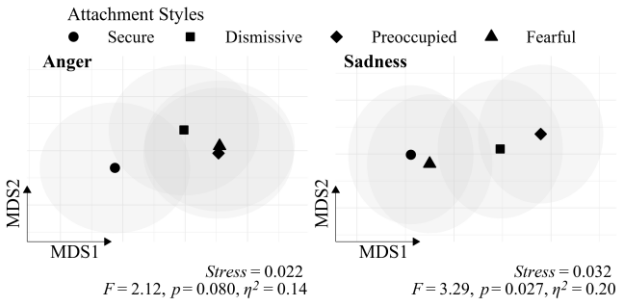


Figure 1: Spatial representation of attachment styles based on physiological markers. The points refer to the centroids of each attachment style. The shadow ellipses surrounding the centroids denote half of the mean distance among attachment styles.

3.2. Acoustic measures

PMANOVA on acoustic measures showed that the effect of attachment style was significant only in happiness (F=1.88, p=0.048, η^2=0.12) and sadness (F=2.09, p=0.010, η^2=0.14).

Spatial representation of four attachment styles is illustrated in Fig. 2. Stress values were 0.171 and 0.182 for happiness and sadness in two-dimensional solution, exhibiting a medium fit to empirical data. Secure individuals can be distinguished from insecure individuals on the 1st dimension in happiness and on the 2nd dimension in sadness. Fearful individuals are at equidistant from preoccupied and secure ones in happiness, while closest to dismissive ones in sadness.

Attachment styles were classified using support vector machines with radial kernel (SVMR) and random forest (RF) based on 14 acoustic measures. Given a small sample size, classifiers were fitted on the whole dataset of all emotions that consisted of 2,608 tokens, including 700 (27%) from secure, 415 (16%) from dismissive, 479 (18%) from preoccupied, and 1,014 (39%) from fearful individuals. The dataset was divided into 70% for training and 30% for testing.

Hyper-parameters were optimized with a 5-repeated 10-fold cross validation, and Synthetic Minority Oversampling Technique (SMOTE) with 10-nearest neighbours was used to balance the tokens of four attachment styles. Two performance metrics, i.e., accuracy and F1-score, were adopted.

RF (0.74, 95%CI: 0.70–0.77) obtained a slightly higher classification accuracy than SVMR (0.69, 95% CI: 0.66–0.72), and both models performed significantly better than the non-information rate (0.39; binomial test: ps<0.001). For F1-score, fearful and secure individuals obtained higher scores than dismissive and preoccupied ones.

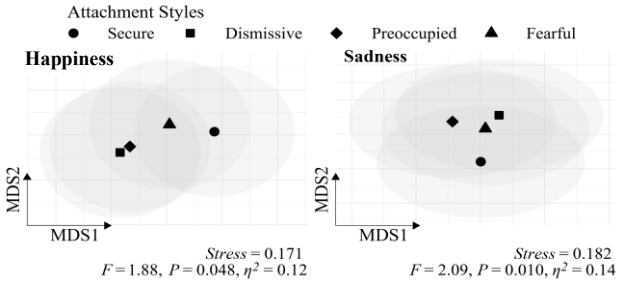


Figure 2: Spatial representation of attachment styles based on acoustic measures.

3.3. Concordance between physiological and acoustic measures

Redundant Analysis (RDA) was employed to reveal emotional concordance between explanatory data (3 physiological indices) and response data (14 acoustic indices) for each attachment style. ANOVA like permutation test (1,000 permutations) was performed to test the multivariate effect of physiological data. The hierarchical partitioning algorithm was applied to calculate the contribution of each predictor in physiological data cloud to explained variation [24].

The multivariate effect was significant for preoccupied (F=2.14, p=0.008) and fearful individuals (F=1.61, p=0.035), but not for dismissive ones. Marginally significant effect was found in secure individuals (F=1.45, p=0.094). The greatest contributions to explained variance were from LF/HF ratio in secure (56.88%), from SCL in preoccupied (63.33%), and from RMSSD in fearful ones (43.75%).

Triplots of RDA of vocal emotion in relation to physiological responses of emotions for attachment styles, in which the significant effect of explanatory data on response data emerged, are illustrated in Fig. 3. The percentage of total explained variance contributed from the first two canonical eigenvalues were indicated by the two axes. The response and explanatory variables were plotted as vectors, in grey line with cross, and in solid line with arrow, respectively. The cosine of the angle between two vectors approximates the correlation between them.

Visual inspection of secure individuals suggests that LF/HF is expected to be positively correlated with the parameters indicating irregularity of voice. For preoccupied individuals, most acoustic indices decreased with SCL. For fearful individuals, RMSSD can be viewed as a negative predictor of parameters related to F0 and INT (excluding F0-SD and INT-SD).

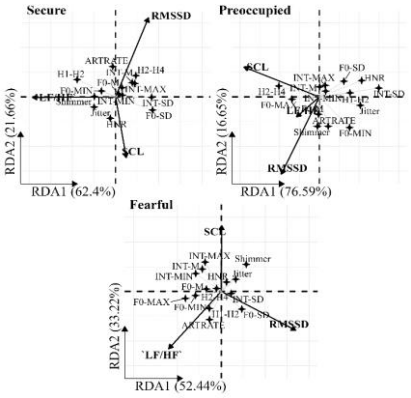


Figure 3: Triplots of redundant analysis of acoustic measure related to physiological responses.

4. CONCLUSION

This study explored both physiological and acoustic distributions of four attachment styles in four basic emotions, and revealed how the relationship between physiological and acoustic measures varies with attachment style.

The distribution of attachment styles in acoustic and physiological spaces turned out to vary with emotion. Four attachment styles differed significantly in physiological indices only in anger and sadness, and differed in acoustic measures only in happiness and sadness. In both cases, secure individuals were most distinct, dismissive and preoccupied ones were close to each other, while fearful ones varied with emotion. A supervised classification based on acoustic measures further verified the impacts of attachment style on acoustic patterns.

The relationship between physiological and acoustic measures turned out to be moderated by attachment style. Significant effects were found for all attachment styles but dismissive. Secure individuals showed an increased voice irregularity when the SNS activity increased or when the PNS activity decreased, suggesting that physiological and acoustic measures are positively correlated. A preference to inhibit vocal emotion was found in preoccupied individuals, showing a negative correlation between the SNS activity and most acoustic measures. For fearful individuals, the PNS activity and most F0 or INT related measures were negatively correlated.

Acknowledgements: This work is supported by the National Social Science Fund of China (13&ZD189).

5. REFERENCES

[1] L. Kaltwasser et al., "Sharing the filmic experience - The physiology of socio-emotional processes in the cinema," *PLoS One*, vol. 14, no. 10, p. e0223259, 2019, doi: 10.1371/journal.pone.0223259.

[2] A. F. Bulagang, N. G. Weng, J. Mountstephens, and J. Teo, "A review of recent approaches for emotion classification using electrocardiography and electrodermography signals," *Informatics in Medicine Unlocked*, vol. 20, p. 100363, 2020, doi: 10.1016/j.imu.2020.100363.

[3] L. Shu et al., "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018, doi: 10.3390/s18072074.

[4] G. A. Bryant, "Vocal communication across cultures: theoretical and methodological issues," *Philos Trans R Soc Lond B Biol Sci*, vol. 377, no. 1841, p. 20200387, 2022, doi: 10.1098/rstb.2020.0387.

[5] J. Bowlby, "Attachment and loss: Retrospect and prospect," *American Journal of Orthopsychiatry*, vol. 52, no. 4, p. 664, 1982, doi: 10.1111/j.1939-0025.1982.tb01456.x.

[6] M. Mikulincer and P. R. Shaver, "Attachment orientations and emotion regulation," *Current Opinion in Psychology*, vol. 25, pp. 6–10, 2019, doi: 10.1016/j.copsyc.2018.02.006.

[7] D. W. Griffin and K. Bartholomew, "Models of the self and other: Fundamental dimensions underlying measures of adult attachment.," *Journal of Personality and Social Psychology*, vol. 67, no. 3, p. 430, 1994, doi: 10.1037/0022-3514.67.3.430.

[8] J. J. Gross, "Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology," *J Pers Soc Psychol*, vol. 74, no. 1, pp. 224–237, 1998, doi: 10.1037//0022-3514.74.1.224.

[9] P. Congia, "Adult attachment dimensions: Differential effects on physiological and subjective response during the recollection of childhood memories," *Europe's Journal of Psychology*, vol. 16, no. 4, p. 639, 2020, doi: 10.5964/ejop.v16i4.1892.

[10] P. Dias, I. Soares, J. Klein, J. P. Cunha, and G. I. Roisman, "Autonomic correlates of attachment insecurity in a sample of women with eating disorders," *Attachment & Human Development*, vol. 13, no. 2, pp. 155–167, 2011, doi: 10.1080/14616734.2011.554005.

[11] J. Zhang and L. Zheng, "Masculine voices predict attachment style and relationship communication patterns in romantic relationships," *Journal of Sex & Marital Therapy*, vol. 47, no. 3, pp. 262–269, 2021, doi: 10.1080/0092623X.2020.1869125.

[12] E. Monti, D. C. Kidd, L. M. Carroll, and E. Castano, "What's in a singer's voice: The effect of attachment, emotions and trauma," *Logoped Phoniatr Vocol*, vol. 42, no. 2, pp. 62–72, 2017, doi: 10.3109/14015439.2016.1166394.

[13] H. Hu and W. T. Gu, "Effects of adult attachment style on prosody and voice quality characteristics of intimate speech in romantic relationships," *Acta Acustica*, vol. 47, no. 02, pp. 276–286, 2022, doi: 10.15949/j.cnki.0371-0025.2022.02.013. (in Chinese)

[14] P. F. Xu, Y. X. Huang, and Y. J. Luo, "Establishment and Assessment of Chinese Affective Video System," *Chinese Mental Health Journal*, vol. 24, no. 07, pp. 551-554+561, 2010, doi: CNKI:SUN:ZXWS.0.2010-07-021.

[15] P. Liu and M. D. Pell, "Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli," *Behavior Research Methods*, vol. 44, no. 4, pp. 1042–1051, 2012.

[16] T. G. Li and K. Kato, "Measuring adult attachment: Chinese adaption of the ECR scale," *Acta Psychologica Sinica*, vol. 38, no. 3, pp. 399–406, 2006, doi: 10.1016/S0379-4172(06)60092-9.

[17] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," Biological psychology, vol. 84, no. 3, pp. 394–421, 2010, doi: 10.1016/j.biopsycho.2010.03.010.

[18] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Frontiers in public health, vol. 5, pp. 258–258, 2017, doi: 10.3389/fpubh.2017.00258.

[19] P. Boersma, D. Weenink. Praat: doing phonetics by computer (Version 6.3) [Computer program]. Retrieved 15 November 2022. Available from: http://www.praat.org/

[20] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in 2004 IEEE international conference on acoustics, speech, and signal processing, 2004, vol. 1, p. I–669, doi: 10.1109/ICASSP.2004.1326074.

[21] M. Mikulincer, O. Gillath, and P. R. Shaver, "Activation of the attachment system in adulthood: threat-related primes increase the accessibility of mental representations of attachment figures.," *Journal of Personality and Social Psychology*, vol. 83, no. 4, p. 881, 2002, doi: 10.1037/0022-3514.83.4.881.

[22] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. URL https://www.R-project.org/.

[23] G. Keklik and M. Şahin, "Multidimensional Scaling Analysis and an Application," *Black Sea Journal of Engineering and Science*, vol. 5, no. 2, p.71, 2022, doi: 10.34248/bsengineering.1077162.

[24] J. Lai, Y. Zou, J. Zhang, and P. R. Peres-Neto, "Generalizing hierarchical and variation partitioning in multiple regression and canonical analyses using the rdacca.hp R package," *Methods in Ecology and Evolution*, vol. 13, no. 4, pp. 782–788, 2022, doi: 10.1111/2041-210X.13800.

MULTIMODAL SIGNALLING: THE INTERPLAY OF ORAL AND VISUAL FEEDBACK IN CONVERSATION

Malin Spaniol[1], Alicia Janz[2], Simon Wehrle[2], Kai Vogeley[1], Martine Grice[2]

[1]Dept. of Psychiatry, University Hospital Cologne, Germany; [2]IfL Phonetik, University of Cologne, Germany
{malin.spaniol, kai.vogeley}@uk-koeln.de; {alicia.janz, simon.wehrle, martine.grice}@uni-koeln.de

ABSTRACT

Human communication comprises a complex and dynamic interplay of verbal and nonverbal communication channels. It is an intrinsically multimodal, interactive, time-sensitive and highly coordinated process. However, the multimodal nature of face-to-face interaction needs further study. Here, we present a novel approach to studying dyadic face-to-face conversation with dual mobile eye-tracking glasses. Our pilot data are from four conversations between German speaking interlocutors in different contexts, comprising introductory small talk, task-oriented dialogue (Tangram task), and free discussion (about the Tangram task). Our exploratory analysis on oral feedback and gaze behaviour reveals that tokens indicating Passive Recipiency (backchannels) tended to involve more gaze directed at the conversational partner than tokens indicating Incipient Speakership (signals with the intention to take the floor). It also offers interesting insights into behavioural differences as a function of communicative context.

Keywords: face-to-face interaction, feedback, social gaze, backchannels, turn-taking

1. INTRODUCTION

In any face-to-face interaction, interlocutors continuously exchange a variety of signals, both verbal and non-verbal, the majority of which are of social relevance [1]. This continuous two-way exchange is possible because interlocutors are able to see, hear and react to each other. The timing of visual and verbal signals is of crucial importance [2, 3].

Language coordination has been claimed to be an inherently redundant process [4]. It is known that e.g. turn-taking cues, relevant for a high degree of coordination, can be transmitted on multiple communication channels.

Verbally, feedback signals such as backchannels function as turn-taking signals and are used to demonstrate understanding and acknowledgement. More specifically, they can indicate either Passive Recipiency (PR), which supports the ongoing turn of the interlocutor, or Incipient Speakership (IS), which signals the intention of taking over the floor from the current speaker [5, 6]. Previous research has found complex but consistent evidence for the relationship between a feedback token's lexical form, its intonation contour and its function as either turn-taking (IS) or turn-yielding (PR) [7]. Although phonetic cues have been shown to generally play a role in signalling the function of feedback signals, there is no one-to-one mapping of phonetic form to turn-taking function [7, 8].

In addition to spoken cues, nonverbal signals such as gaze have been shown to substantially influence the coordination of turns [9, 10]. Gaze serves both a sensing and a signalling function [11]. Early research on gaze-direction in social interaction has shown that even small differences in study design, regarding e.g. task or setting, can have considerable effects on gaze behaviour [12, 13] and substantially influence the coordination of turns [10].

More recent studies on gaze-direction in social interaction further emphasize the importance of gaze in turn-taking [14] and have shown that speakers generally tend to direct their gaze at their interlocutor towards the end of a turn, while more often averting it at the beginning [13, 15] in order to plan the rest of the current turn [16]. Direct gaze by the speaker, which entails mutual gaze, can create a so-called "gaze window" [17], proposed to function as a backchannel-inviting cue [23]. Furthermore, speakers likely direct their gaze to the interlocutor before the end of a turn in order to signal turn-yielding and facilitate a possible turn-transition [10, 15, 16, 18].

Although there is substantial inter-individual variability, previous results also indicate a general tendency for participants to gaze less at their conversation partner when they are in the speaker role as compared to the listener role [13].

Despite the fact that both verbal feedback and gaze behaviour are important cues for the coordination of speaking turns, only few studies have investigated the relation between these two signals in the context of turn-taking to date. Skantze et al. [11] have shown that direct gaze by the speaker (entailing mutual gaze) serves as a backchannel-inviting cue. However, the function of gaze behaviour by the listener uttering

feedback remains unclear. More specifically, the relationship between Passive Recipiency vs. Incipient Speakership on the one hand, and aversion vs. direction of gaze on the other, remains unclear.

Although it is apparent that the complex interplay of different conversation modalities and channels is still not fully understood, a recent renewed interest in the dyad as a fundamental unit of social behaviour, aided by recent technological and methodological advancements that enable fine-grained multimodal analysis, can help in furthering our understanding.

In an attempt to contribute to this progress, we have developed a novel method for analysing multimodal communication. In a highly naturalistic setting, we record gaze using mobile dual eye-tracking glasses and speech via lapel microphones while capturing the scene with an external camera. This setup allows for the simultaneous measurement of multiple verbal and non-verbal signals. The glasses are equipped with integrated cameras and are capable of automatic gaze detection. Additional face-detection on the videos recorded by the eye-tracking glasses enable us to automatically detect gaze into the facial region of the interlocutor.

The current study is based on pilot data recorded with this novel setup. The exploratory analysis presented here is focussed on verbal feedback signals, gaze behaviour, and the interplay between the two.

Because verbal feedback and directed gaze from the listener both function as feedback signals for the speaker, and as signals are expected to be redundant, we expect to see directed gaze towards the speaker during backchanneling. However, one hypothesis for averted gaze at the beginning of turns is that it helps to keep cognitive load low while planning the turn ahead. Verbal feedback with the function of Incipient Speakership always occurs at the beginning of a speech turn. Thus, it can be expected that direct gaze during Incipient Speakership tokens will be reduced compared to Passive Recipiency backchannels (where we expect to find gaze directed towards the speaker). The previous literature has shown that all of the relevant behaviours are task-dependent, and therefore, we also expect to find differences between conversational contexts.

We investigated three different communicative contexts, and the following questions guided our exploration: Do listeners show different gaze patterns when producing Incipient Speakership signals, as compared to Passive Recipiency backchannels? If so, do these patterns differ according to communicative context?

2. METHOD

2.1. Apparatus

Two participants were seated opposite each other wearing Pupil Invisible mobile eye-tracking glasses [19] and lapel microphones. The glasses were connected to remote devices which were wirelessly connected to the experimenter's computer. The eye-trackers were manually adjusted for each participant using gaze offset correction. The lapel microphones were attached to the eye-tracking glasses. Microphones and a scene camera were connected to the experimenter's laptop via cable. Figure 1 shows an overview of the experimental setup. The experimenter was in a separate room for the entire duration of the experiment.

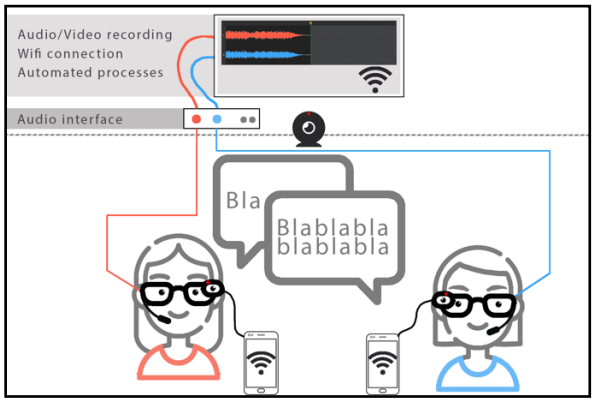


Figure 1: Multimodal dyadic conversation setup.

2.2. Participants

Four dyads matched for age and gender participated in this study (8 participants: 6 female, 2 male; age M = 29.63, SD = 5.21, range = 23-38). All were native German speakers. The study was approved by the ethics committee of the University Hospital Cologne.

2.3. Procedure

After giving their written informed consent, participants were introduced to each other before being seated and equipped with eye-tracking glasses and microphones. Conversations lasted between 26 and 48 minutes and encompassed three communicative contexts, in a fixed order: 1. Introduction (and getting to know each other); 2. Tangram task; 3. Discussion (of the Tangram task).

2.4. Task and recordings

Two assistants familiarised the participants with the recording set-up before giving them instructions for the first context and the opportunity to ask any questions. The assistants then left the room and the conversation started. After 10 minutes, an assistant

came back into the room and provided instructions for the Tangram task. One participant saw four different figures, of which one was highlighted by an arrow. The other participant saw only one figure. The task for the dyad was to find out if the single figure was identical to the highlighted one; an example can be found on the *OSF* project page. The Tangram conversations comprised 10 iterations and had no time limit. After the task was completed, an assistant entered the room again to give instructions for the final, ten-minute long discussion.

The three conversational contexts where chosen for being complementary, but not unrelated. Getting to know each other was used as an example of a relatively frequent everyday situation. The Tangram task was chosen, in contrast, as a form of highly structured, task-oriented dialogue, and also because it has been used in other studies on speech and gaze in interaction [20, 21]. The final discussion about the Tangram task was designed as a conversational context that was as open and relaxed as possible, while also guaranteeing a conversation on the basis of shared knowledge and experience.

2.5. Data

The pilot dataset consists of 149 minutes of conversation in total. Eye-tracking data were logged at 200 Hz. Audio data were recorded with *Audacity* (version 2.4.2) at a sampling rate of 44100Hz (32-bit). The scene video was recorded with *Quicktime* player, and the videos recorded by the eye-tracking glasses were logged with 30 frames per second.

2.6. Processing and analysis

For the processing of the eye-tracking data, we used the dedicated *PupilCloud* software [19]. For annotation of audio files, Praat [22] was used. Further data processing and visualisation was conducted using *Python* and *R*. Eye-tracking and speech data were pre-processed separately. Eye-tracking data were stored as time series data in a .csv file, audio data independently as .wav files.

For each participant, eye-tracking behaviour was automatically merged with the video data of the eye-tracking glasses and dummy-coded as *gaze_in_face* using the fixation detection and face detection enrichment in *PupilCloud*, which detects the facial region of the interlocutor. Both *gaze_in_face* files were merged via a *Python* script by calculating a time offset relative to the start of recordings. For all consecutive timestamps, the mean of both files was used, leading to a negligible latency between the two *gaze_in_face* data points of up to half a frame.

Speech data were annotated in *Praat*, semi-automatically for interpausal units (with a minimum silence of 200 milliseconds) and manually for tasks, turns, backchannel types, backchannel tokens and backchannel functions. The resulting TextGrids were merged into the *gaze_in_face* time series format. All pre-processing and analysis scripts can be found on the *OSF* project page https://osf.io/su5nx.

3. RESULTS

Overall, we found that for the discussion context, PR backchannels were more often accompanied by gaze directed towards the interlocutor than was the case for IS tokens, but this did not hold true for the other two conversational contexts (overall—PR: 39%, IS: 33.7%; introduction—PR: 52.2%, IS: 56%; Tangram—PR: 12.9 %, IS: 15.1%; discussion—PR: 48.8%, IS: 35.7%). Figure 2 shows an example of the interaction between gaze and verbal feedback behaviour from one dyad (Dyad 2).

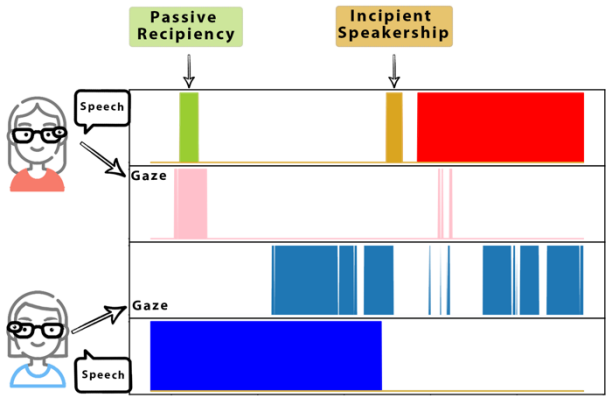


Figure 2: Example turn-plot showing 5 seconds of a conversation from Dyad 2. PR in green, IS in yellow.

In total, the data set contained 1208 verbal feedback tokens. The overall proportion of dialogue featuring verbal feedback was 4.95% (PR: 4.03 %, IS: 0.93%). This proportion was relatively stable for the different dyads (Dyad 1: 6.27%, Dyad 2: 4.67%, Dyad 3: 5.3%, Dyad 4: 4.31%). Overall, the proportion was highest in the introduction and similar (but somewhat lower) in the discussion, but far lower in the Tangram context (introduction: 6.23 %, Tangram: 3.69%, discussion: 5.65%). In all tasks, more PR than IS tokens were produced, but this difference was more pronounced for the introduction and discussion (~ 5 times more PR than IS) than for the Tangram task (~ 3 times more PR than IS).

Overall, the proportion of gaze directed towards the interlocutor was again highest during the introduction, similar (but somewhat lower) during the discussion and far lower during the Tangram task (introduction: 78.7%, Tangram: 24.2%, discussion: 73.5%).

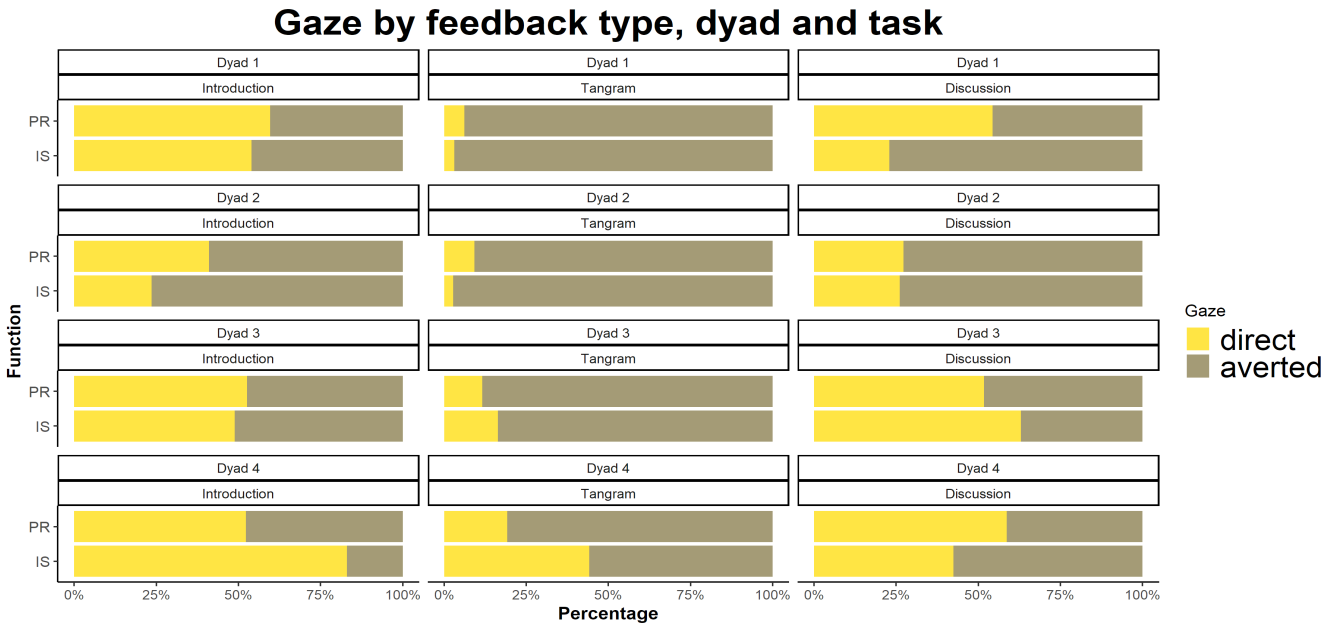


Figure 3: Proportion (on the x-axis) of gaze directed towards the partner (in yellow) vs. averted gaze (in grey) during the production of Passive Recipiency (PR) and Incipient Speakership (IS) types of verbal feedback (on the y-axis), as a function of dyad and communicative context.

Similarly, face-directed gaze *during the production of verbal feedback* was most frequent in the introduction, slightly less frequent in the discussion and very rare in the Tangram task.

On the dyad level, we observed differences in how much time was spent on backchanneling, and in how often gaze was directed towards the partner while providing verbal feedback; see Figure 3. All four dyads showed changes in gaze behaviour according to context, and most dyads used more direct gaze during the production of PR as compared to IS across the contexts, although there were exceptions, notably Dyad 4 in the introduction and Tangram task and Dyad 3 in the Tangram task and the discussion.

4. DISCUSSION AND CONCLUSION

The goal of the present study was to explore verbal feedback signals (e.g. backchannels), visual feedback signals (i.e. gaze), and their interplay in face-to-face interactions with different conversational contexts and using a novel experimental setup.

The time spent producing verbal feedback was relatively similar across the three conversational settings, but the proportion was lower overall in task-oriented conversation (Tangram) as compared to more spontaneous speech (cf. [23–25]). Conversational context also had an influence on the type of verbal feedback produced: a higher proportion of IS tokens was used during the Tangram compared to the more spontaneous contexts. This increased use of Incipient Speakership tokens, signaling the intention to take over the speaker role, might reflect a functional motivation to complete the task efficiently.

The proportion of gaze directed at the interlocutor differed substantially between conversational contexts. The low proportion of face-directed gaze we found during the Tangram task was expected, as participants were holding a piece of paper and had to look at it regularly to complete the task. Regarding the interplay of gaze and verbal feedback, we found that only during the discussion, participants directed their gaze more towards the partner for PR as compared to IS. This effect was heavily influenced by the behaviour of Dyad 4, which showed opposite patterns for the two least natural contexts (introduction and Tangram).

This exploration of our pilot data set suggests that gaze and verbal feedback behaviour differs depending on the communicative context, and that gaze behaviour also changes depending on the type of verbal feedback used. In future work, we will use this highly promising, novel methodology of exploring multi-modal behaviour via dual mobile eye-tracking glasses to explore further conversational behaviours and different groups of speakers. We will focus in particular on the intonational realisation of verbal feedback, on analysing head nods and overall kinematic energy, and on the behaviour of autistic adults, a population that is known to have difficulties in social interaction and to use social cues, such as eye gaze and verbal feedback, in a different way.

5. ACKNOWLEDGEMENTS

The study was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) – Project-ID : 281511265 – SFB 1252.

6. REFERENCES

[1] L. V. Hadley, G. Naylor, A. F. De, and C. Hamilton, 'A review of theories and methods in the science of face-to-face social interaction', *Nature Reviews Psychology 2022 1:1*, vol. 1, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s44159-021-00008-w.

[2] J. R. Kelly and J. E. McGrath, *On Time and Method*. Newbury Park: SAGE Publications, Inc, 1988.

[3] J. McGrath and F. Tschan, *Temporal Matters in Social Psychology: Examining the Role of Time in the Lives of Groups and Individuals*. Washington: American Psychological Association, 2004.

[4] B. Winter, 'Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality', *BioEssays*, vol. 36, no. 10, pp. 960–967, 2014.

[5] K. Drummond and R. Hopper, 'Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency', *Research on Language and Social Interaction*, vol. 26, no. 2, Art. no. 2, Apr. 1993, doi: 10.1207/s15327973rlsi2602_3.

[6] M. Savino, 'Degrees of (un) certainty in Bari Italian yes-no question intonation: perceptual evidence', in *The communication of certainty and uncertainty: Linguistic, psychological, philosophical aspects*, in 1, vol. 12. In Zuczkowski, A., Bongelli, R., Riccioni, I., & Canestrari, C., pp. 52–67.

[7] S. Sbranna, E. Möking, S. Wehrle, and M. Grice, 'Backchannelling across Languages: Rate, Lexical Choice and Intonation in L1 Italian, L1 German and L2 German', *Proc. Speech Prosody 2022*, pp. 734–738, 2022.

[8] S. Wehrle, 'A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder', PhD Thesis, University of Cologne, Cologne, Germany, 2021.

[9] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Oxford: Cambridge University Press, 1976.

[10] Z. Degutyte and A. Astell, 'The Role of Eye Gaze in Regulating Turn Taking in Conversations: A Systematized Review of Methods and Findings', *Front. Psychol.*, vol. 12, p. 616471, Apr. 2021, doi: 10.3389/fpsyg.2021.616471.

[11] G. Skantze, A. Hjalmarsson, and C. Oertel, 'Turn-taking, feedback and joint attention in situated human-robot interaction', *Speech Communication*, vol. 65, p. 66, 2014, doi: 10.1016/j.specom.2014.05.005.

[12] B. Butterworth and G. Beattie, 'Gesture and Silence as Indicators of Planning in Speech', in *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*, R. N. Campbell and P. T. Smith, Eds., in NATO Conference Series. Boston, MA: Springer US, 1978, pp. 347–360. doi: 10.1007/978-1-4684-2532-1_19.

[13] A. Kendon, 'Some functions of gaze-direction in social interaction', *Acta Psychologica*, vol. 26, pp. 22–63, 1967, doi: 10.1016/0001-6918(67)90005-4.

[14] P. Auer, 'Turn-allocation and gaze: A multimodal revision of the "current-speaker-selects-next" rule of the turn-taking system of conversation analysis', *Discourse Studies*, vol. 23, no. 2, pp. 117–140, Apr. 2021, doi: 10.1177/1461445620966922.

[15] Ho, T. Foulsham, and A. Kingstone, 'Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions', *PLOS ONE*, vol. 10, no. 8, Art. no. 8, Aug. 2015, doi: 10.1371/journal.pone.0136905.

[16] D. R. Rutter, G. M. Stephenson, K. Ayling, and P. A. White, 'The timing of Looks in dyadic conversation', *British Journal of Social and Clinical Psychology*, vol. 17, no. 1, Art. no. 1, Feb. 1978, doi: 10.1111/J.2044-8260.1978.TB00890.X.

[17] J. B. Bavelas, L. Coates, and T. Johnson, 'Listener Responses as a Collaborative Process: The Role of Gaze', *Journal of Communication*, vol. 52, no. 3, pp. 566–580, Sep. 2002, doi: 10.1111/j.1460-2466.2002.tb02562.x.

[18] J. Holler, K. H. Kendrick, M. Casillas, and S. C. Levinson, *Editorial: Turn-Taking in Human Communicative Interaction*, vol. 6, no. DEC. 2015. doi: 10.3389/fpsyg.2015.01919.

[19] M. Tonsen, C. K. Baumann, and K. Dierkes, 'A High-Level Description and Performance Evaluation of Pupil Invisible'. arXiv, Sep. 01, 2020. Accessed: Dec. 05, 2022. [Online]. Available: http://arxiv.org/abs/2009.00508

[20] J. Holler, J. Bavelas, J. Woods, M. Geiger, and L. Simons, 'Given-New Effects on the Duration of Gestures and of Words in Face-to-Face Dialogue', *Discourse Processes*, vol. 59, no. 8, pp. 619–645, Sep. 2022, doi: 10.1080/0163853X.2022.2107859.

[21] M. Savino, L. Lapertosa, and M. Refice, 'Seeing or Not Seeing Your Conversational Partner: The Influence of Interaction Modality on Prosodic Entrainment', in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 574–584. doi: 10.1007/978-3-319-99579-3_59.

[22] Boersma and Weenink, 'PRAAT: Doing phonetics by computer (Version 6.1.40)', 2021. http://www.praat.org/ (accessed Jul. 30, 2021).

[23] A. Janz, 'Navigating Common Ground Using Feedback in Conversation- A Phonetic Analysis', MA thesis, University of Cologne, Cologne, Germany, 2022.

[24] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, 'Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations', in *CogSci'19*, Cognitive Science Society, 2019, pp. 261–267.

[25] R. Fusaroli, K. Tylén, K. Garly, J. Steensig, M. H. Christiansen, and M. Dingemanse, 'Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions', 2017.

KIJK: A PRAAT PLUGIN TO VISUALISE ARTICULATORY TRAJECTORIES

Carolina Lins Machado[1], Lei He[1 2]

[1] Dept. Computational Linguistics, University of Zurich, Zurich, Switzerland
[2] Dept. Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland
cmachado@ifi.uzh.ch, lei.he@uzh.ch

ABSTRACT

Electromagnetic articulography (EMA) is one of the most frequently employed methods for the collection of speech kinematic data. Likewise, Praat is one of the most used software for processing and analysing acoustic data, due to its relatively straightforward operation and portability across different operational systems. However, Praat does not possess an automatic procedure to visualise kinematic data collected with EMA devices. Therefore, in this paper we introduce *Kijk*, a Praat plugin for the visualisation of kinematic trajectories. Because it is completely implemented in Praat, *Kijk* offers the user a quick and simple approach to the combined visualisation and annotation of articulatory and acoustic data.

Keywords: Praat, plugin, EMA, kinematic trajectory

1. INTRODUCTION

Speech production is a process involving the movement of articulators – such as the lips, tongue, velum and jaw – resulting in an acoustic outcome encoding linguistic, paralinguistic and extralinguistic information, all vital to communication between individuals. The process of producing speech is broadly studied, for instance, by speech scientists, linguists and phoneticians at the articulatory and acoustic levels. Acoustic analysis entails investigating the speech wave itself, where characteristics of a language can be revealed, for instance, in the spectral and temporal features of the acoustic signal. Changes in the spectral features in this signal are achieved by the movements of the articulators. Thus, articulatory analysis offers an understanding of how, to some extent, these acoustic features are modulated.

Articulatory analysis can be performed with a variety of methods. Among a diverse array of approaches for the collection of articulatory data, the most frequently employed method is electromagnetic articulography (or EMA) [2], which can be used to study oral movements. By attaching sensors to the tongue, jaw and lips, the movement of multiple articulators can be recorded simultaneously (for a further description of how EMA works please see [2]). The output of this method is kinematic data related to the displacement of articulators. The analysis of speech kinematic data is commonly performed using the native EMA device tools, specialised softwares [3, 11], or toolkits [4, 5, 9, 10] for Matlab [6] and R [8], to name a few.

Acoustic data on the other hand is widely processed in Praat [7], an open-source freeware for the analysis of acoustic data. It's popularity is without a doubt due to its portability, and because it requires a shallow learning curve from the user, being attractive to students and researchers alike. Regarding acoustic analyses, Praat can be used for visualising, annotating (through TextGrids) and measuring speech; offering the user a great number of features for these purposes. Moreover, the software allows the user to synthesise and manipulate speech, to perform some statistical analyses and to output results also in graphical form. Therefore, Praat offers the potential to be used for the visualisation and annotation of EMA trajectories. For instance, by examining kinematic trajectories, acoustic recordings and annotated TextGrids simultaneously it becomes easier to study and compare EMA data along with the acoustic output of specific speech sounds. However, this software does not yet possess an automatic method to handle EMA data. For that reason, we attempt to offer a solution with *Kijk* [kɛik] (meaning "look" in Dutch), a plugin completely implemented in Praat to automatically process EMA data for the visualisation of kinematic trajectories.

The plugin was developed under Praat version 6.2.13 and assumes that the EMA data was corrected for head movements and that it was translated to a coordinated system in relation to the occlusal plane.

2. PLUGIN STRUCTURE

Kijk's mechanism makes it possible to process kinematic trajectories completely in Praat, without requiring the user to directly manipulate the script. The plugin can be installed in the Praat preferences folder, or it can be called as a script. If the user decides to install the plugin, a button will appear on the dynamic menu after the user selects the EMA file (Figure 1). So far, it works for data collected with the EMA devices NDI Wave and Carsten's AG501, but it will soon be extended to support other devices, such

as Carsten's AG 100 and 200. The plugin is available via https://github.com/linsca/kijk.

2.1. Interface

Before running the plugin, the user should open the desired EMA file in Praat. For the NDI Wave this is a tab-separate file and for the AG501 a text file in tabular format. After the file is selected in Praat's object window, either click the plugin button on the dynamic menu (Figure 1) or run the script.

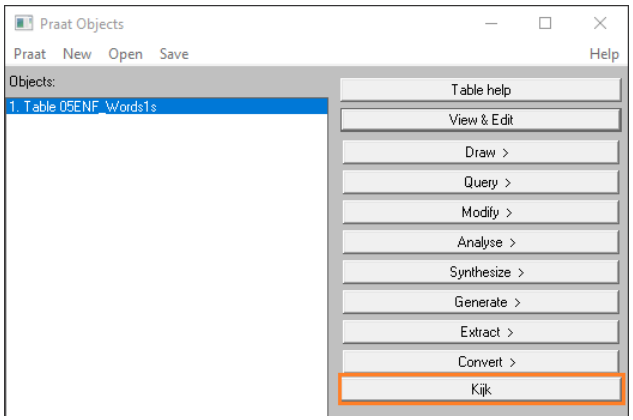


Figure 1: Partial screenshot of Praat's object window. If Kijk is installed as a plugin it appears as a button (encased in orange) on the dynamic menu when the table object is selected.

A pop-up form (Figure 2) will appear asking the user to (i) input the sensor number, (ii) the corresponding articulatory variable, (iii) the dimension (*x-y-z* position), (iv) the EMA device used, and (v) the sampling frequency. Further, if the user wishes to extract information from more than one sensor and combine it in a single output, there is an option to indicate that. It is important to mention that the plugin is case-sensitive. So, the inputted dimensions need to be given in uppercase.

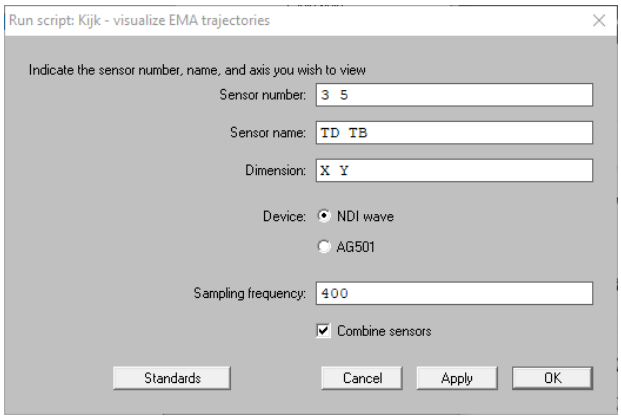


Figure 2: Screenshot of Kijk's initial pop-up form.

2.4. Algorithm

After the preferences have been entered in the pop-up form, the user can click "OK" or "Apply" to start the process. Depending on which EMA device the data stems from, preprocessing follows different procedures.

The output of the NDI Wave device includes nine values recorded per sensor: sensor ID, sensor status, sensor position in millimetres (X[mm], Y[mm], Z[mm]) and sensor orientation in quaternion format (Q0, Qx, Qy, Qz). The AG501 device outputs seven values per sensor: sensor position (Ch#_X, Ch#_Y, Ch#_Z), orientation angles (Ch#_phi, Ch#_theta), root mean square (Ch#_RMS), and a placeholder for an additional data column (Ch#_Extra). As opposed to the AG501 device, the output file of the NDI Wave does not include the sensor number in the labels of its values, therefore, it is first necessary to rename the columns related to each sensor in order for *Kijk*'s core algorithm to work.

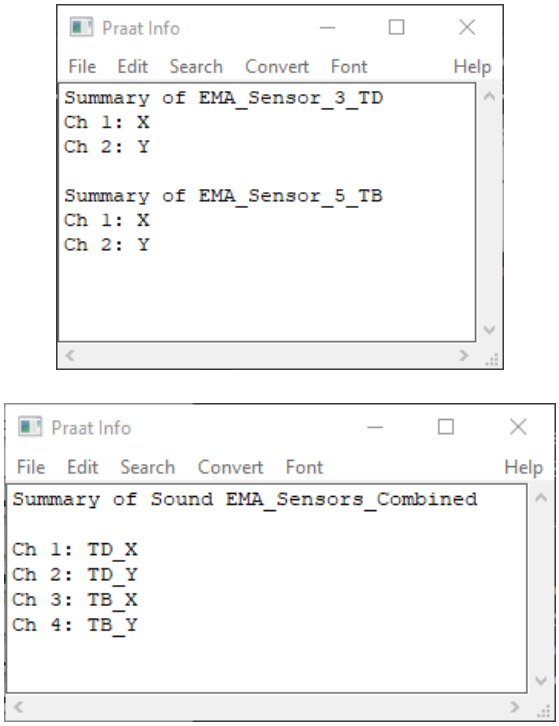


Figure 3: Praat Info windows displaying a summary of what each channel in the EMA object contains. The left screenshot is the output of sensors as independent objects and the right screenshot is the output of the sensors combined in one object.

The algorithm then commences the preparation of the data by first transposing the EMA file to a wide tabular format. Next, it surveys the transposed table to extract the information pertaining to the desired sensors and position axes (i.e. *x-y-z* dimensions). Afterwards, Praat converts the tabular data into a

matrix, which is then casted to a sound object. Finally, the sampling frequency of this object is overridden by the frequency inputted by the user.

3. OUTPUT

The output of the plugin is twofold. The first is a summary of the extraction procedure displayed on a Praat Info window. Depending on the user's selection of the "combine sensors" option in the initial form, the output is either for each sensor (Figure 3 top screenshot) or of the combined sensors (Figure 3 bottom screenshot).

If the output is not a combined object; i.e. for each sensor, the name of the EMA object includes the sensor number and the articulatory variable it represents. EMA_Sensor_3_TD indicates that this object refers to the third sensor placed on a speaker's tongue dorsum (TD). The following lines indicate what each channel of the EMA object contains. In the object EMA_Sensor_5_TB channel 1 (Ch 1) contains the movement trajectory of the tongue blade (TB) in the anteroposterior direction (*x*) and channel 2 (Ch 2) TB's trajectory in the vertical direction (*y*).

If the output is a combination of two (or more) sensors, the name of the EMA object is generic (EMA_Sensors_Combined). The following lines specify what each channel entails. In this example (Figure 3 bottom screenshot), channels 1 and 3 contain anteroposterior movement trajectories, and channels 2 and 4 contain vertical movement trajectories of tongue blade and dorsum, respectively TD_X, TB_X, TD_Y, TB_Y.

The secondary output of *Kijk* is the EMA trajectory object itself. It can be treated like any other audio signal in Praat. For instance, to eliminate some of the noise found in the data prior to visualisation, the signal encoding the kinematic trajectories was low pass filtered at 10 Hz (smoothing = 5Hz) using Praat's Hann filter. Similarly, this output can be visualised like any other sound object in Praat, by selecting it and clicking on "View and Edit" (Figure 4). In older versions of Praat the EMA object may appear off-scale. To remedy this issue, in the editor window the user can change the sound scaling strategy to "by window and channel".

Visualising EMA trajectories and corresponding acoustic data is also possible without needing to have

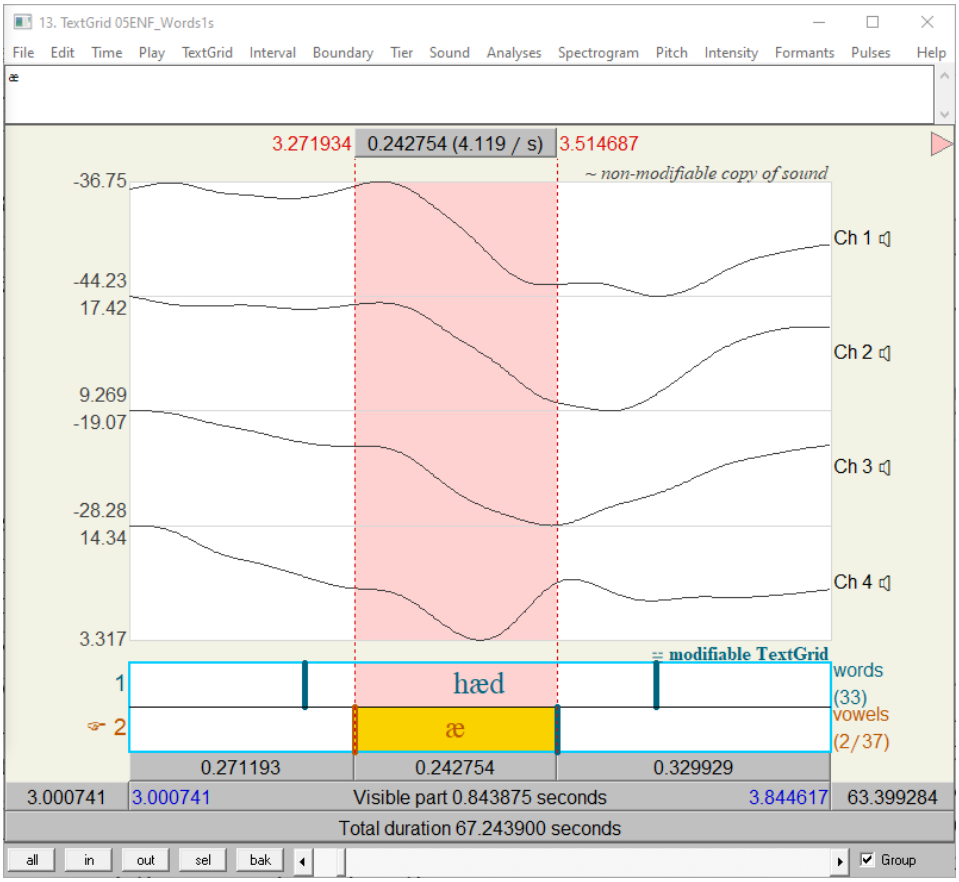


Figure 4: Screenshot of the Praat editor window displaying the combined EMA data from the EMA-MAE corpus [1] and TextGrid annotations; The channels contain tongue dorsum anteroposterior (Ch 1) and vertical (Ch 2) directions, and tongue blade anteroposterior (Ch 3) and vertical (Ch4) directions.

all the data stored in the same object. By selecting "Group" at the bottom right corner of the sound editor windows, selections done in a particular region of the acoustic object will automatically be reflected on the EMA object (Figure 5). This is possible because the collection of acoustic and kinematic data using these EMA devices is time-synchronized. In the example shown in the screenshot below, users can simultaneously visualise the oscillogram, the spectrogram, the annotations, and tongue dorsum's anteroposterior (x) and vertical (y) kinematic trajectories of the selected segment /æ/.

4. CONCLUSION AND OUTLOOK

In this paper we introduced *Kijk*, a Praat plugin for the visualisation of EMA trajectories. Although limited to displaying only position data, a future iteration will include the calculation of velocity and acceleration. Nonetheless, at the current stand, *Kijk* is a simple and straightforward alternative for students and researchers wishing for a quick visualisation of kinematic trajectories, or as an educational tool for introducing undergraduate students to EMA data along with the acoustic signal. Finally, *Kijk* was conceptualised as an open tool, therefore we welcome feedback and collaborations in future versions of this plugin.

5. ACKNOWLEDGEMENTS

This work was supported by the Forschungskredit of the University of Zurich (Grant No. FK-20-078), and the Swiss National Science Foundation (Grant #PZ00P1_193328) to LH.

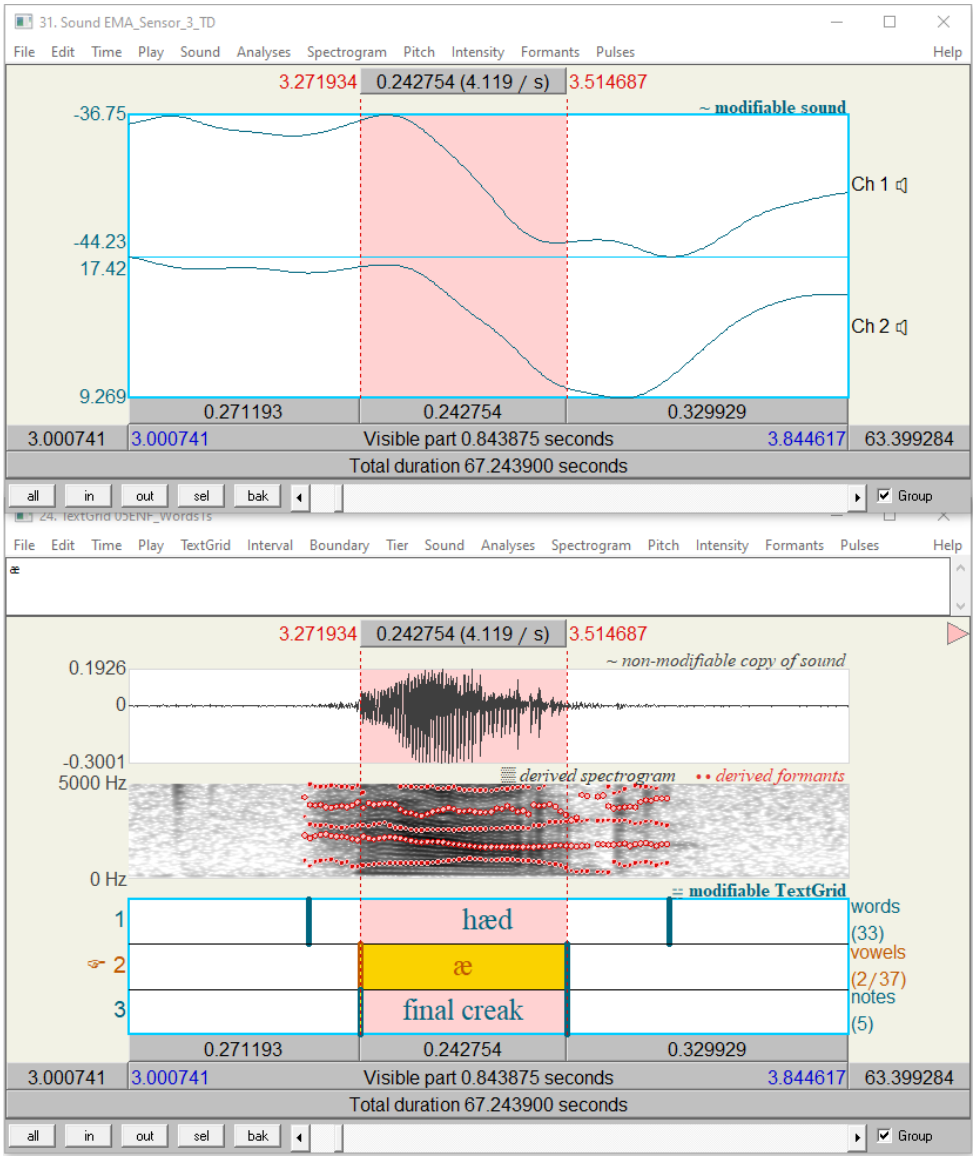


Figure 5: Screenshot of Praat editor windows grouped together. The top window displays filtered tongue dorsum trajectories in the x and y directions and the bottom window shows the oscillogram, spectrogram and TextGrid.

6. REFERENCES

[1] A. Ji, J. J. Berry, and M. T. Johnson, "The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7719–7723. doi: 10.1109/ICASSP.2014.6855102.

[2] A. Kochetov, "Research methods in articulatory phonetics I: Introduction and studying oral gestures," *Language and Linguistics Compass*, vol. 14, no. 4, p. e12368, 2020, doi: https://doi.org/10.1111/lnc3.12368.

[3] Articulate Instruments Ltd. Articulate Assistant Advanced User Guide: Version 2.14, Edinburgh, UK, 2012.

[4] J. R. Green, J. Wang, and D. L. Wilson, "SMASH: a tool for articulatory data processing and analysis," in *Proc. Interspeech 2013*, 2013, pp. 1331–1335. doi: 10.21437/Interspeech.2013-353.

[5] M. Tiede, *MVIEW: Software for visualization and analysis of currently recorded movement data*. Haskins Laboratories, 2005.

[6] "MATLAB." The MathWorks Inc., Natick, Massachusetts, 2022.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Nov. 15, 2022. [Online]. Available: http://www.praat.org/.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.Rproject.org/

[9] R. N. Henriques and P. van Lieshout, "A Comparison of Methods for Decoupling Tongue and Lower Lip From Jaw Movements in 3D Articulography," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1503–1516, 2013, doi: 10.1044/1092-4388(2013/12-0016).

[10] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*. 2021.

[11] S. Ouni, L. Mangeonjean, and I. Steiner, "VisArtico: a visualization tool for articulatory data," in *13th Annual Conference of the International Speech Communication Association - InterSpeech 2012*, Portland, OR, United States, Sep. 2012. [Online]. Available: https://hal.inria.fr/hal-00730733

A MULTIMODAL ACCOUNT OF LISTENER FEEDBACK IN FACE-TO-FACE INTERACTIONS

Martina Rossi[1], Marin Schröer[2], Bogdan Ludusan[2], Margaret Zellers[1]

[1]Kiel University, Germany, [2]Bielefeld University, Germany
mrossi@isfas.uni-kiel.de, marin.schroeer@uni-bielefeld.de, bogdan.ludusan@uni-bielefeld.de, mzellers@isfas.uni-kiel.de

ABSTRACT

In face-to-face interactions, the conversational feedback produced by the listener to signal attention and participation to the current speaker is multimodal: in the vocal channel, it consists of verbal expressions (e.g., "yes" or "exactly") and vocalizations without lexical content, such as non-lexical backchannels (e.g., "mhm") and laughter; in the visual channel, listener feedback includes movements of the head, such as nods or tilts. In the current research, we investigate the frequency and the distribution (i.e., the location and the transition type with respect to the other interlocutor's turn) of lexical and non-lexical items, laughter and head movements, as well as the phonetic variation of vocal feedback, in face-to-face dialogues in German.

We find that the feedback type influences the distribution and the variation of intensity values and voice quality, and, for multimodal items, it also influences the temporal alignment of the head movement with the vocal component.

Keywords: multimodal feedback, backchannels, head movements, laughter, German conversation

1. INTRODUCTION

In conversation, listeners tend to provide feedback to their conversational partner, i.e., they signal their active attention or react to what is being said, without interrupting the speaker's stream of talk [1, 2, 3, 4]. In face-to-face interactions, listener feedback phenomena, also called "backchannels" [1] or "listener responses" [5], are observed in both the vocal and the visual modalities, i.e., through verbal and non-verbal vocalizations, and through gestures.

Verbal feedback is constituted by lexical items or phrases with a very short constituent size [6]. In German, typical lexical feedback expressions are, e.g., "ja" "genau" "okay" "eben" "achso" [7]. Moreover, non-lexical vocalizations involving nasals, such as "mmm" or "mhm", are often also candidates for backchannels, in German as well as cross-linguistically [8]. Laughter is another non-verbal vocalization employed for backchanneling purposes, which often occurs in spontaneous interactions [9]. It has several functions in conversation, including linguistic ones (e.g., for discourse organization [10]). Previous work has shown that laughter is used as a feedback signal across languages, with the proportion of laughter out of the total backchannels varying between a few percentage points and almost a quarter of all backchannels [11, 12, 13].

Finally, head movements, such as head nods, are another feedback type occurring in face-to-face conversations [1, 14]. Being communicated via the visual channel, head movements are even less disruptive than vocal feedback and can potentially occur more often and in more locations then vocal ones without being judged inappropriate [15]. They can occur on their own, but are very often observed in co-occurrence with both verbal and non-verbal feedback expressions, or in their vicinity, often preceding vocalizations [16, 5].

Previous experimental research on listener responses in conversation has focused mostly on the identification and description of feedback inviting cues (e.g., [17, 18]), while less attention has been put on the phonetic features of feedback items and how they vary depending on their type and their placement in the ongoing talk [13]. In particular, while there is a growing body of research targeting multimodal backchannels (e.g., [4, 15]), non-verbal vocalizations tend not to be included in the investigations, so that not much is known about how and to what extent they differ from each other and from other types of feedback.

2. METHODS

2.1. Dataset

We investigated face-to-face dialogues between German native speakers using audio and video files

taken from the German subset of the multimodal DUEL corpus [19]. We analyzed 12 dyads involving 24 different speakers for a total of 2 hours and 15 minutes of dialogue. The recording setup for the DUEL corpus consisted of lapel microphones in front of each subject and two video cameras capturing the gesture space and the faces of the subjects. Participants were involved in different interaction scenarios designed to have them start talking without having to select a subject, while also allowing a free discussion, and to ensure the presence of laughter in different contexts. The scenarios analyzed for this study were "Film script" and "Dream apartment". The corpus includes the full transcription of the dialogues, their subdivision in conversational turns and utterances, as well as laughter occurrences.

2.2. Annotation

Listener feedback was identified in the dyads using the corpus' orthographic transcription, the audio and the video files. We consider as listener feedback all those lexical and non-lexical backchannels, laughter and head movements produced/performed by one of the interlocutors as an optional response to what the current speaker is saying (e.g., answers to questions are not considered as listener feedback) [17], and which are not part of a full-fledged turn (e.g., turn-opening, turn-closing items or discourse markers are not considered as listener feedback).

In Praat [20] we isolated lexical and non-lexical backchannels, and carried out a further annotation of laughter, in order to consider only those instances which corresponded to our feedback identification guidelines. The category of verbal feedback includes all the items that functioned as feedback, e.g., "ja", "genau", "okay", "das stimmt", while non-lexical feedback includes items such as "mhm". No form distinction will be used for the purposes of the current study. For the annotation of gestural feedback, we used the software ELAN [21] and the guidelines provided by the M3D annotation scheme [22] for head movements. First, all head movements were annotated using the video without the audio; then, in Praat, additionally considering the corresponding interlocutor's turns, a further annotation was carried out to isolate the movements which occurred as a stand-alone listener response, or in connection with a verbal, non-verbal or laughter type feedback.

Moreover, to investigate feedback distribution, we observed their location with respect to the current speaker's turn. If the feedback occurred completely in overlap with the other interlocutor's speech, it is considered "turn-internal", while if it occurred right after the interlocutor's turn end or it finished outside of it, followed by a silence, it is considered "turn-external" [3]. Finally, our study included a further analysis of turn-external feedback by categorizing the type of transition with which it occurred. Specifically, we labelled the transition type of feedback that occurred during an overlap with the other speaker's turn as "overlap," and feedback that occurred after a silent gap as "gap." [23, 24]. If the silent gap or portion of overlapped speech was less than 120 ms, the transition was considered "no-gap-no-overlap," as previous research has established 120 ms as the detection threshold for silences and overlaps [23].

2.3. Phonetic features

To investigate the phonetic variation of both laughter and verbal feedback we focus on the features of loudness and voice quality, using the values of maximum intensity (intmax) and the mean of the cepstral peak prominence (CPP). In contrast with other features, such as F0 and duration, intensity and voice quality have yet to be well investigated in feedback expressions. For lexical and non-lexical backchannels, [25] and [26] report that backchannels tend to be louder than the other affirmative words in American English, and that non-lexical items tend to be significantly less loud then short lexical expressions in Slovak. Using the parameter of jitter, [27] studied voice quality in backchannels in Ruruuli/Lunyala, finding higher values, correlated with lower periodicity, in longer lexical items, while non-lexical items displayed lower jitter. Intensity and voice quality have been studied also for laughter (e.g., [28]), with a higher intensity and a less modal production being found for laughter, compared to speech. For the current study, intmax and CPP were both extracted from the entire interval constituting either the feedback item (lexical, non-lexical, laughter) using a Praat script. Intensity values were normalized with the individual speaker's mean.

3. RESULTS

A total of 771 feedback items were extracted from the 12 dialogues. In terms of frequency, lexical items and head movements (without speech) are the most frequent feedback types in our data (with, respectively, 294 and 240 occurrences), followed by laughter (n=158), and lastly by non-lexical backchannels (n=79). Statistical analyses were carried out using R [29].

3.1. Head movements

A total of 462 head movements were extracted, including both those cases which constituted a feedback item on their own and those which accompanied vocalizations. The most common type of head movement with a feedback function is the nod (n=383, 83%), followed by head tilts (n=37, 8%). We also extracted a few occurrences of protrusions (n=32, 7%), which mostly accompany laughter, and head turns (n=10, 2%). Out of the 531 vocal feedback items, the 42% of the lexical backchannels are accompanied by a head movement (n=136), as well as the 56% of the non-lexical ones (n=44) and 27% of the laughs (n=42). The onset of the head movement tends to occur mostly before the vocalization, specifically 84 ms before lexical backchannels, 200 ms before non-lexical items and 140 ms before laughs, while the movement offset is observed 340 ms before the offset for laughter, and 140 ms and 172 ms after the end of the vocalization for lexical and non-lexical feedback respectively. Linear mixed models, with the speaker as a random effect, and subsequent Tukey post-hoc pairwise comparisons indicate a significant difference between the timing of the head offsets for laughter and lexical feedback (β = -.4691, t ratio = -4.696 p < .0001) and laughter and non-lexical feedback (β = -.5110, t ratio = -4.219, p = .0001), while the difference between the head offset timing for lexical and non-lexical item is not significant (β = -.0419, t ratio = -.418, p = .9084).

item	gap	nogap	overlap
lex	89(2.17)	25(0.44)	68(−2.17)
non	18(0.62)	10(2.22)	12(−1.68)
laugh	23(−2.12)	9(−0.75)	60(2.28)
head	37(1.19)	10(−1.19)	69(1.68)

Table 1: Count of observations (and residuals) of the χ^2 test for the transition type relative to the feedback item (lexical, non-lexical, laughter, head movement).

3.2. Distribution

With respect to the other interlocutor's speech, feedback constituted only by a head movement tend to occur mostly turn-internally, while all the other types of feedback are mostly observed turn-externally, with the exception of non-lexical backchannels, which are observed in almost equal measure turn-internally and turn-externally (see Fig.1). A χ^2 test confirms the existence of a relationship between the feedback type and the location in which it is produced by the listener (χ^2 = 11.099, df = 3, p = .0112). There is also a relationship between the feedback type and the transition with which it occur considering the final boundary of the other interlocutor's turn, i.e., with a gap, an overlap or a "no-gap-no-overlap" (χ^2 = 33.869, df = 6, p < .0001). The Pearson residuals for the test indicate that lexical feedback occurs significantly more frequently with a gap, laughter more frequently with an overlap, and non-lexical backchannels with no-gap-no-overlap transition (see Tab.1).

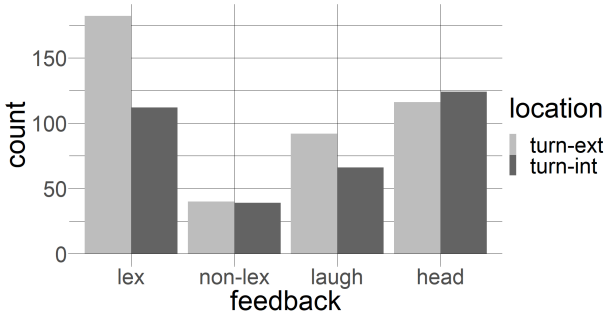


Figure 1: Distribution of feedback items (lexical, non-lexical, laughter, head movments) with respect to the interlocutor's current speaking turn (turn-external, turn-internal).

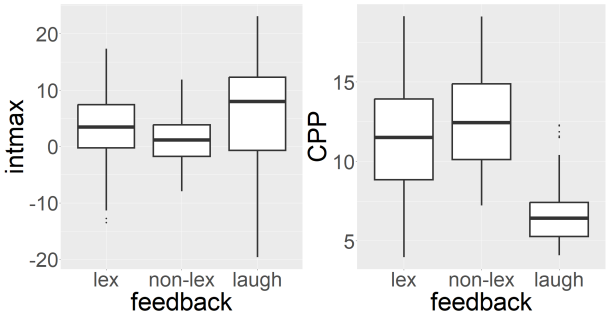


Figure 2: Maximum intensity (normalized with speaker's mean) and CPP values for vocal feedback items (non-lexical, lexical, laughter).

3.3. Phonetic variation

As for the phonetic features of the vocal feedback items, we find that laughter tends to have a higher intmax than the other types of backchannels, followed by lexical items and non-lexical items (see Fig.2). Conversely, mean CPP values are higher for non-lexical items, followed by lexical feedback and laughter (see Fig.2). Linear mixed models and Tukey post-hoc tests show that intensity values are

significantly different between lexical feedback and laughter (β = -2.75, t ratio = -4.062, p = .0002) as well as between non-lexical items and laughter (β = -4.59, t ratio = -5.132, p < .0001), while the difference is not significant between lexical and non-lexical items (β = 1.83, t ratio = 2.191, p = .0737). CPP values are significantly different between each pair in the post-hoc comparisons (lexical-non-lexical: β = -1.26, t ratio = -3.584, p = .0011; lexical-laughter: β = 4.94, t ratio = 17.305, p < .0001; non-lexical-laughter: β = 6.20, t ratio = 16.440, p < .0001).

Additionally, we tested if the presence of a head movement might also correlate with variation in the phonetic features of vocalizations. Except from some qualitative differences, feedback items with a co-occurring head movement do not have significantly distinct features from those where the movement is not present, with the exception of the intensity values for lexical items. In these, lexical items with a co-occurring head movement seem to be significantly louder than the unimodal ones, i.e., constituted only by the vocal element (β = -2.422, t ratio = -3.149, p = .0214).

4. DISCUSSION

We investigated the distribution and the acoustic realization of multimodal feedback in German face-to-face dialogues by observing lexical backchannels, head movements, laughter and non-lexical backchannels. Lexical items are the most frequent type of feedback in our dataset, followed by head movements. In line with previous studies (e.g., [4] for Dutch and [15] for British English), lexical backchannels tend to occur mostly turn-externally, and after a gap, while head movements, being the least disruptive type of feedback, occur more often turn-internally. Lexical items tend to be quieter than laughter, as previously observed for German by [28], but they are generally louder than non-lexical items. In terms of voice quality, they appear to be less periodic than non-lexical items, which could be motivated by their segmental content including voiceless segments or even short silences in complex lexical phrases.

Our results on head movements' temporal alignment with speech in multimodal feedback items show that the onset of the head movements precedes the vocalization by around 141 ms, which, as hypothesized by [16], is done by the listener to anticipate the start of a feedback response without interrupting the interlocutor. The alignment of the head movement with laughter is significantly different when compared to the other feedback types, since the offset of the movement occurs well before the laugh's end. This might be due to the fact that laughter is the type of feedback where the most protrusions are observed. A head protrusion would generally consist of a single movement of the head forward or backwards (and not in repeated back-to-back movements, as in e.g. head nods), so the head might tend to return to its resting position before the laughter is finished.

Moreover, we observe that multimodal lexical backchannels in our data appear to be louder than unimodal lexical ones. This might suggest a relation between the phonetic features of feedback items and head movements. For instance, [30] find that, for articulation rate and nodding, increased effort in speech production is accompanied by increased head movement. Further tests (e.g, logistic regression models) will be carried out on our data to find out whether this is the case for multimodal feedback as well.

Finally, we find that the two types of non-verbal vocalizations present some substantial differences from each other, both in terms of their distribution and their phonetic features. While non-lexical backchannels tend to occur at the same rate turn-internally and turn-externally, and are produced after a smooth transition at the end of the other interlocutor's turn, laughs mostly occur turn-externally, and often after a portion of overlapped speech at the end of the current speaker's turn.

5. CONCLUSIONS

With the current analysis we provided a description of the distributional features of different types of feedback, both unimodal and multimodal, and we tested how different factors, such as the feedback's form, location, its transition type and the presence of a co-occurring head movement, influence the acoustic realization of vocal and multimodal listener responses. Future research should include, e.g., the lexical and segmental content of the verbal backchannels to provide a more detailed description of their variation, and the phonetic features of the current speaker's turn corresponding to a listener response, in order to observe if determinate cues would elicit a specific type of feedback.

6. ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Council), project no. 444631148 for MR and MZ and project no. 461442180 for MS and BL.

7. REFERENCES

[1] V. H. Yngve, "On getting a word in edgewise," *CLS-70*, pp. 567–578, 1970.

[2] S. Duncan, "On the structure of speaker-auditor interaction during speaking turns," *Language in Society*, vol. 3, pp. 161–180, 1974.

[3] G. Kjellmer, "Where do we backchannel? On the use of mm, mhm, uh huh and such like," *International Journal of Corpus Linguistics*, vol. 14, pp. 81–112, 2009.

[4] K. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogues," in *Proc. Interspeech 2011*, 2011, pp. 23–25.

[5] S. K. Maynard, "Conversation management in contrast: Listener response in Japanese and American English," *Journal of Pragmatics*, vol. 14, no. 3, pp. 397–412, 1990.

[6] F. E. Müller, "Affiliating and disaffiliating with continuers: prosodic aspects of recipiency," in *Prosody in Conversation: Interactional Studies*, M. Couper-Kuhlen, E. Selting, Ed. Cambridge University Press, 1996, pp. 131–176.

[7] B. Knudsen, A. Creemers, and M. A.Sn., "Forgotten little words: How backchannels and particles may facilitate speech planning in conversation?" *Frontiers in Psychology*, vol. 11, pp. 593–671, 2020.

[8] A. Liesenfeld and M. Dingemanse, "Bottom-up discovery of structure and variation in response tokens (backchannels) across diverse languages," in *Proc. Interspeech 2022*, 2022, pp. 1126–1130.

[9] J. Trouvain and K. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proc. of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 2012, pp. 36–39.

[10] B. Ludusan and B. Schuppler, "To laugh or not to laugh? the use of laughter to mark discourse structure," in *Proc. SIGDIAL*, 2022, pp. 76–82.

[11] S. Pammi and M. Schröder, "A corpus based analysis of backchannel vocalizations," in *Proc. Interdisciplinary Workshop on Laughter and other Interactional Vocalisations in Speech. Berlin, Germany*, 2009.

[12] P. Cutrone, "A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners," *Intercultural Pragmatics*, vol. 11, no. 1, pp. 83–120, 2014.

[13] V. Krepsz, V. Horváth, Á. Hámori, D. Gyarmathy, and C. I. Dér, "Backchannel responses in Hungarian conversations: a corpus-based study on the effect of the partner's age and gender," *Linguistica Silesiana*, pp. 113–140, 2022.

[14] E. McClave, "Linguistic functions of head movements in the context of speech," *Journal of Pragmatics*, vol. 32, no. 7, pp. 855–878, 2000.

[15] G. Ferré and S. Renaudier, "Unimodal and bimodal backchannels in Conversational English," in *Proc. SEMDIAL 2017*, 2017, pp. 27–37.

[16] A. T. Dittmann and L. G. Llewellyn, "Relationship between vocalizations and head nods as listener responses," *Journal of Personality and Social Psychology*, vol. 9, no. 1, pp. 79–84, 1968.

[17] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[18] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[19] J. Hough, Y. Tian, L. de Ruiter, S. Betz, S. Kousidis, D. Schlangen, and J. Ginzburg, "DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter," *Proc. LREC' 16*, pp. 1784–1788, 2016.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2022. [Online]. Available: http://www.praat.org

[21] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proc. LREC 2006*, 2006.

[22] P. L. Rohrer, I. Vila-Giménez, J. Florit-Pons, G. Gurrado, N. Esteve Gibert, A. Ren, S. Shattuck-Hufnagel, and P. Prieto, "The multimodal multidimensional (M3D) labelling scheme for the annotation of audiovisual corpora," in *Proc. GESPIN 2020*, 2020.

[23] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *Journal of the Acoustical Society of America*, vol. 130(1), pp. 508–513, 2011.

[24] M. Rossi, K. Feindt, and M. Zellers, "Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish," in *Proc. Speech Prosody 2022*, 2022, pp. 185–189.

[25] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proc. ICPhS XVI*, 2007, pp. 1065–1068.

[26] S. Benus, "The prosody of backchannels in Slovak," in *Proc. Speech Prosody 2016*, 2016, pp. 75–79.

[27] M. Zellers, "An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala," *Journal of Pragmatics*, vol. 175, pp. 38–52, 2021.

[28] B. Ludusan and P. Wagner, "ha-HA-hha? Intensity and voice quality characteristics of laughter," in *Proc. Speech Prosody 2022*, 2022, pp. 560–564.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[30] M. Tiede, C. Mooshammer, and L. Goldstein, "Noggin nodding: Head movement correlates with increased effort in accelerating speech production tasks," *Frontiers in Psychology*, vol. 10, 2019.

ICONIC GESTURES IN FOCUS – SYNCHRONIZATION OF PROSODY AND GESTURES IN PROMINENCE

Frank Kügler & Alina Gregori

Institute for Linguistics, Goethe University Frankfurt, Germany
kuegler@em.uni-frankfurt.de, gregori@lingua.uni-frankfurt.de

ABSTRACT

This study investigates the impact of focus on the alignment of iconic gestures with prosody in spontaneous German speech. Generally, co-speech gestures synchronize with prosodic prominence [1]. From the SaGA Corpus [2], gesture types, iconic gesture apexes, and pitch accent types were extracted in focus and non-focus contexts. Results show that focused constituents were always marked with a pitch accent, and a minority were additionally accompanied by an iconic gesture. In 35% of unfocused constituents, no pitch accent was realized. Yet, the majority of unaccented constituents on non-focus constituents were accompanied by an iconic gesture. Regarding temporal synchronization, focus synchronizes the gestures more closely to the pitch accent than in non-focused constituents. This result points to the fact that a pragmatic function of highlighting seems to be added to iconic gestures to their otherwise ascribed dimension of expressing a semantic relation to speech concepts.

Keywords: Iconic co-speech gestures, prosody, focus, gesture-speech synchronization.

1. INTRODUCTION

This study investigates the impact of focus on the alignment of iconic co-speech gestures with prosody in spontaneous German speech. Co-speech gestures are assumed to accompany speech in a systematic way [e.g. 1, 3]. Related to prosody, a phonological synchrony rule between prosodic prominence and co-speech gestures has been claimed [1]: In general, it is assumed that a gesture coincides with a pitch accent, more specifically, the stroke of a gesture ends with the pitch accent at the latest. Basically, four different gesture types have been classified as iconic, deictic, metaphoric, and beat [1], the latter one being referred to as non-referential whereas the other three types are referential [4]. Since these types have different functions, the question arises whether all gesture types behave similar concerning the synchrony between prosodic prominence and gestures given (i) that different gesture types are assumed to be panned at distinct planning stages, e.g. [5].

For this study, we are concerned with iconic gestures, which show a semantic connection to speech in that they visually express the meaning of the co-occurring speech [1, 6]. See Figure 1 for an example of an iconic gesture where the speaker signals the shape of a pillar using his thumb and remaining fingers forming a cylinder. In what way the semantic function of iconic gestures interacts with prosodic prominence evoked by focus is the concern of the present study. To answer this, we conducted a corpus study on spontaneous German speech [2].



Figure 1: Example of an iconic gesture from the SaGA corpus [2]. The speaker utters the word *Säule* 'pillar', while imitating the shape of a pillar with his hands.

2. BACKGROUND

2.1. Information structure

We base our study on the information structure category *focus*, which according to [7, p. 247] "indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions." The non-focus part of an utterance is the labelled as *background* [8]. Languages use different linguistic means such as phonology, syntax, morphology or a combination thereof to express focus [9]. Stress-based languages such as Germanic languages use prosodic means, i.e. pitch accentuation to mark focus [10]. Hence, accentuation achieves the goal of making a focus more prominent than background information. Different types of focus are distinguished: Information focus is best illustrated in a question-answer pair where the answer identifies one of the alternatives being asked for. A contrastive focus usually includes a focus alternative, which was proposed in the immediately preceding context.

2.2. Prosody

In prosody, a distinction between prosodic domains according to the prosodic hierarchy [11, 12] and prosodic categories such as pitch accents, phrase and

boundary tones [13] is made, which we label using the GToBI system [14]. Pitch accents are usually the head of their prosodic domain [15] and have a highlighting function [13]. In German intonation, a number of different pitch accent types are assumed [14], and the pitch accent types bear different inherent prosodic prominence. According to [16], the pitch accents are organized on a prominence scale with rising or high pitch accents being more prominent than low ones.

2.3. Co-speech Gestures

Co-speech gestures are "visible bodily actions" accompanying speech [3]. Here, we concentrate on hand gestures. A widely used classification system of such gestures distinguishes four main types: iconic (Figure 1), metaphoric, deictic, and beat gestures [1]. A recent proposal groups the first three as referential gestures as opposed to non-referential ones, being assigned values in multiple dimensions [17]). We are interested in iconic gestures as they visually support and mirror the expressed semantic content of speech [1, 6 for an overview]. They thus allow for the transmission of additional or redundant information to the speech they accompany [6].

A gesture usually consists of multiple hierarchically ordered components [18]. Generally, the stroke of a gesture is assumed to align with syllables [1, 3] or pitch accents [17, 19–21]. Within the stroke, the apex of a gesture represents the gestural peak, which can be conceived of as the temporal point of no velocity, or the change of direction in the gesture's movement [4, 17]. Gestural strokes are integrated within a gesture phase, which in turn is integrated into a gesture phrase.

2.4. Prosody–Gesture–Link

According to [1], gestures and speech are two modalities of the same framework. For the integration of the two, he claimed three synchrony rules (pragmatic, semantic and phonological). For the latter one, he states that "the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech" [1, p. 26]. Later empirical work on the temporal synchronization of gestures and speech has shown that gestures and accents tend to occur near each other [19, 21]. The synchronization was observed for the level of the stroke and pitch accent [1], but also larger constituents such as gesture phrases and phonological phrases [21]. Recently, more empirical evidence has shown that information structure affects the synchronization [17, 22]. A referent carrying *new* information facilitates the occurrence of gestures.

2.5. Research question and hypothesis

The proposal of synchrony rules of gesture-speech integration [1] covers all gestures types. Focusing on iconic gestures, we are interested in their synchronization patterns with respect to prominence, i.e. focus. Given their function of mirroring the expressed semantic content of speech, the question arises if iconic gestures appear in focus and in particular, whether iconic gestures align closely with pitch accents in prominent position (focus).

Given the assumption that gestures can operate in multiple dimensions [23], we hypothesize that iconic gestures mark prominence in addition to their semantic contribution to speech and thus show sensitivity to pragmatic prominence, contemplating that iconic gestures occur more frequently and align more precisely with prosody in prominent positions.

3. CORPUS STUDY

3.1. Information on the corpus

The Bielefeld **S**peech **a**nd **G**esture **A**lignment (SaGA) corpus [2] is an audio-visual corpus containing German spontaneous speech conversations. The setting is based on a virtual reality (VR) town where participants were taken on a bus ride. The task was to inform an interlocutor about the route and certain landmarks (direction-giving task). In total, the data used from the corpus consists of 204 minutes of speech distributed over 18 dialogues. The corpus contained gesture type annotation according to [2].

3.2. Annotation and measurements

The corpus was further annotated for gesture apexes according to the M3D guidelines [4] using ELAN, for pitch accent types according to GToBI [14] using Praat [24], and for two levels of information structure (information status given, accessible, new; focus-background) [8]. Different types of focus (narrow information, contrastive focus) were collapsed for the analysis since prosodically, there were no significant differences in the prosodic realisation between focus types [e.g., 25].

From the corpus, pitch accent types (n = 4394), gesture occurrences (apex of non-referential and iconic gestures, n = 2402) and information structure levels (focus, n = 2773; background, n = 2251) were extracted. The present analysis focuses on iconic gestures (n = 1627). For the synchrony measure, the distance between the pitch accent and the iconic gesture's apex was calculated. Thus, each apex-accent pair was listed on a histogram according to their distance (in groups of one frame of 33ms length) The midpoint 0 ms means that apex and accent occur

at the same time. Positive values were assigned to the pairs in which the apex occurs before the accent, negative values when the accent occurs before the apex. Distances of more than 3000 ms were collapsed in a category 3000+.

4. RESULTS

4.1. Distribution of Pitch accents and Gestures

The distribution of pitch accents and iconic gestures split by focus is shown in Figure 2a. All focused referents (right bar) carry a pitch accent. Approximately 25% of the focused referents are additionally accompanied by an iconic gesture (red). Note that additionally about 10% are accompanied by non-referential gestures [26]. In the non-focus condition (background, left bar), approximately 63% of referents carry a pitch accent. From those, about 10% are additionally accompanied by iconic gestures (red). Approximately 30% of referents in non-focus carry no pitch accent but an iconic gesture (blue). For completeness, further 25% of non-focus constituents are accompanied by non-referential gestures (25% of these with an accent, 75% with no pitch accent) [26].

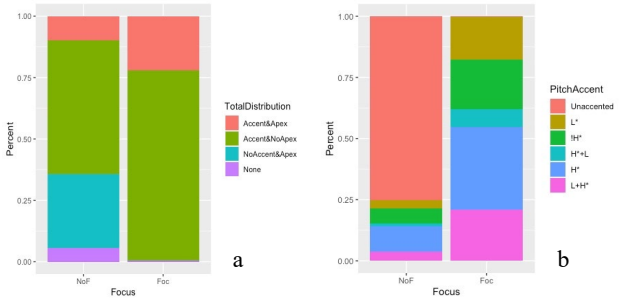


Figure 2: Distribution of pitch accents and iconic gestures split by focus (a); distribution of pitch accent types split by focus on iconic gestures (b).

4.2. Iconic gestures: Focus and Pitch accents

Zooming into the accent distribution on iconic gestures split by focus in Figure 2b (red and blue parts in Figure 2a): An iconic gesture in focus (right bar) occurs with all different pitch accent types of German. Note that we do not differentiate between focus types, and that hence the plot does not display the tendency for more prominent pitch accents to occur in contrastive focus [e.g., 27] than information focus. However, we observe a majority of accents containing a high accentual tone H*, well in line with studies on prosodic focus in German, e.g., [27]. Roughly, the H* part appears to be related to convey the meaning of adding new information to the common ground, e.g., [28].

In the non-focus condition (left bar), there is a large number of iconic gestures on unaccented referents (red; 75%). The remaining distribution of pitch accents is impressionistically equivalent to distribution in focus, just in a squeezed range. From the comparison of focus and non-focus, it can be seen that focus, unsurprisingly, requires prosodic prominence (presence of a pitch accent).

4.3 Temporal alignment

Figure 3 presents the distribution of temporal alignment of iconic gesture apexes with pitch accents in focus. It can be seen that there is a clear agglomeration of apex-accent alignment around zero with a standard deviation (SD) of 267ms. Zero represents the exact congruence of the pitch accent and the gesture apex. The mean of alignment of 41ms is behind zero. 91.3% of the focused pairs (n=486) were produced within a distance of -500 to 500ms between the two targets. 53% of accents followed the apex.

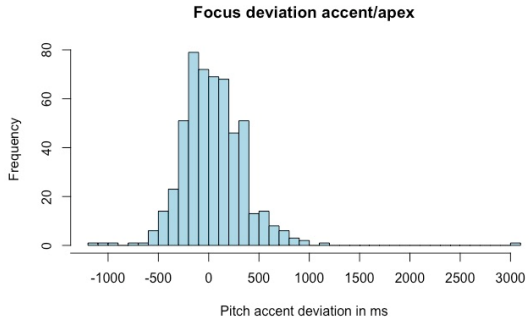


Figure 3: Distribution of the apex-accent distance on focus. - : accent precedes apex; + : accent follows apex.

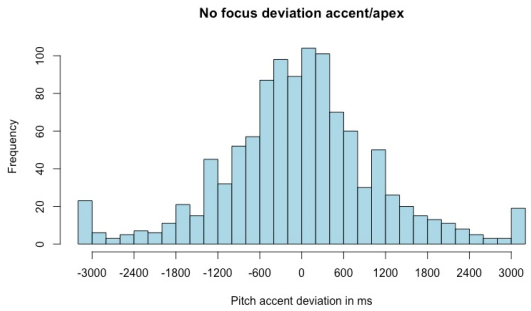


Figure 4: Distribution of the apex-accent distance on non-focus. - : accent precedes apex; + : accent follows apex.

Figure 4 displays the distribution of temporal alignment of iconic gesture apexes with pitch accents in non-focus. The distribution is spread over the full time span, though cumulating around zero. This large deviation of apex-accent alignment amounts in a standard deviation of 416ms. The mean of alignment appears before zero, though with -9ms very close to the midpoint. Overall, comparing with the distribution in focus, the apex-accent alignment in non-focus is less precise than in focus. Only 44.7% of the focused pairs (n=490) were produced within a

distance of -500 to 500ms between the two targets. 49.1% of accents followed the apex.

5. DISCUSSION

The general question of this study was to explore whether and how iconic gestures interact with prosodic prominence, that is whether a pragmatic effect due to focus might be added to the otherwise semantic function of iconic gestures. Starting from a general gesture-prosody link that co-speech gestures are synchronized with prominent syllables or pitch accents [1], we observe that iconic gestures occur in prominent contexts, i.e. in focus (about 25%), but more often in less prominent background contexts (about 40%). If occurring in prominent contexts, iconic gestures always co-occur with pitch accents. In non-focus contexts, the majority of iconic gestures occurs on unaccented elements.

In general, the distribution of iconic gestures in relation to pitch accentuation in focus and non-focus is very similar to that of non-referential gestures [26]. This may point to the fact that focus itself does not impose a preference for a particular type of gesture. Of course, this has to be shown for other gesture types like metaphoric and deictic gestures. Nevertheless, our results point to a similar behaviour of the different gesture types, which seems comparable to the multimodal marking of information status of discourse referents in English [17].

Looking at the distribution of iconic gestures occurring on different pitch accent types, we can observe that although the majority of iconic gestures occurs on unaccented elements in non-focus contexts, the distribution of different pitch accent types accompanied by iconic gestures is otherwise similar in focus and non-focus contexts. Due to the high number of unaccented material, the distribution is squeezed in a smaller range than in focus. Also in this aspect, we observe a similar behaviour of iconic and non-referential gestures [see 26].

The core question of this study was whether the temporal synchronization between iconic gestures and pitch accents is affected by focus. Our results point to a clear difference between focus and non-focus contexts. In focus, iconic gestures are synchronized very tightly with the pitch accent and show less temporal deviation than in non-focus. This effects seem to pattern with a general effect of hyperarticulation in speech [29]. Hence, focus clearly constitutes one of the factors that impacts on the variation in speech along the hypo- and hyperarticulation continuum. In addition, focus more generally affects the multimodal dimension of speech in that both the speech signal and the occurring co-speech gesture together undergo hyperarticulation.

A further conclusion concerns the function of iconic gestures. While iconic gestures are known to express semantic meaning, they also mark prominence. Iconic gestures occur on focused referents to a considerable amount. And these gestures are tightly synchronized with pitch accents. This tight synchrony in focus seems to add a pragmatic function of highlighting to the semantic meaning of iconic gestures. A co-speech gesture can hence express two functions at the same time. Given an identical form of iconic gesture, it visually supports the expressed semantic content of speech [1, 6], and, at the same time, it expresses a discourse function of highlighting information [7–9]. Our findings support the M3D proposal of multiple dimensions to a gesture [4, 17]. There, the idea that a gesture can be labelled at different layers, i.e. form, semantic function and pragmatic function was proposed. Our data show a clear case of support of this differentiation of gestural layers, especially taking into account the similar patterns of iconic and non-referential gestures regarding structural marking.

Future research has to show whether co-speech gestures show a use of a larger space or kinematic movements when expressing two functions at the same time, or whether focus would induce this effect both in discourse structuring non-referential gestures and in referential gestures that express both semantic and pragmatic meaning.

6. CONCLUSION

This study was concerned with the interaction between iconic gestures and prominence. Results show, not surprisingly, that focus requires prosodic marking. In addition, a quarter of those cases were accompanied by iconic gestures. Comparing the synchronization of iconic gestures in focus and non-focus contexts, our results show tighter alignment in focus than in non-focus. This points to two facts. First, focus goes hand in hand with more articulatory effort in speech [29] and prosody [see e.g. 30]. Our results on a closer temporal synchronization of iconic gestures with pitch accent under focus can hence be considered an instance of gestural hyperarticulation. Second, a pragmatic function of highlighting seems to be added to iconic gestures to their otherwise ascribed dimension of expressing a semantic relation to speech concepts.

7. ACKNOWLEDGEMENTS

Thanks to the DFG (KU 2323/5-1), as part of SPP 2392 Visual Communication (ViCom), Frankfurt am Main, Germany, and Goethe University Frankfurt for funding this research. Thanks to Andy Lücking for providing access to the SaGA corpus.

8. REFERENCES

[1] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: U Chicago Press, 1992.

[2] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser, "The Bielefeld Speech and Gesture Alignment Corpus (SaGA)," in *LREC 2010 Workshop: Multimodal Corpora–Advances in Capturing, Coding and Analyzing Multimodality*, 2010, pp. 92–98.

[3] A. Kendon, *Gesture: Visible action as utterance*. Cambridge: CUP, 2004.

[4] P. L. Rohrer *et al.*, *The MultiModal MultiDimensional (M3D) labeling system for the annotation of audiovisual corpora: Gesture Labeling Manual*. UPF Barcelona, 2020.

[5] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *J Memory and Language*, vol. 57, pp. 396–414, 2007.

[6] K. G. Kandana Arachchige, I. Simoes Loureiro, W. Blekic, M. Rossignol, and L. Lefebvre, "The Role of Iconic Gestures in Speech Comprehension: An Overview of Various Methodologies," *Front. Psychol.*, vol. 12, p. 634074, 2021.

[7] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica*, vol. 55, no. 3, pp. 243–276, 2008.

[8] M. Götze *et al.*, "Information structure," in *Interdisciplinary Studies on Information Structure (ISIS)*, vol. 7, S. Dipper, M. Götze, and S. Skopeteas, Eds., Potsdam: Universitätsverlag Potsdam, 2007, pp. 147–187.

[9] M. Zimmermann and E. Onea, "Focus marking and focus interpretation," *Lingua*, vol. 121, no. 11, pp. 1651–1670, 2011.

[10] F. Kügler and S. Calhoun, "Prosodic Encoding of Information Structure: A typological perspective," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds., Oxford: Oxford University Press, 2020, pp. 453–467.

[11] M. Nespor and I. Vogel, *Prosodic phonology*. Dordrecht: Foris Publ., 1986.

[12] E. Selkirk, "On derived domains in sentence phonology," *Phonology Yearbook*, vol. 3, pp. 371–405, 1986.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: CUP, 2004.

[14] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed., Oxford: Oxford University Press, 2005, pp. 55–83.

[15] E. O. Selkirk, "The syntax-phonology interface," in *The Handbook of Phonological Theory*, J. A. Goldsmith, J. Riggle, and A. C. L. Yu, Eds., 2nd ed., Chichester: Wiley-Blackwell, 2011, pp. 435–484.

[16] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in German," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 2015, p. 384.

[17] P. L. Rohrer, "A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system," PhD thesis, Universitat Pompeu Fabra, Barcelona, 2022.

[18] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, M. R. Key, Ed., Berlin: Mouton, 1980, pp. 207–227.

[19] S. Shattuck-Hufnagel, Y. Yasinnik, N. Veilleux, and M. Renwick, "A method for studying the time-alignment of gestures and prosody in American English: 'Hits' and pitch accents in academic-lecture-style speech," in *Fundamentals of verbal and nonverbal communication and the biometric issue*, A. Esposito, M. Bratanic, E. Keller, and M. Marinaro, Eds., Amsterdam: IOS Press, 2007, pp. 34–44.

[20] N. Esteve-Gibert and P. Prieto, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *J Speech Lang Hear Res*, vol. 56, no. 3, pp. 850–864, 2013.

[21] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012.

[22] S. Im and S. Baumann, "Probabilistic relation between co-speech gestures, pitch accents and information status," *Proceedings of the Linguistic Society of America*, vol. 5, no. 1, pp. 685–697, 2020.

[23] D. McNeill, "Gesture: a psycholinguistic approach," in *Encyclopedia of language and linguistics*, E. K. Brown, Ed., 2nd ed., Amsterdam: Elsevier, 2006, pp. 58–66.

[24] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]: Version 6.3.03 (http://www.praat.org/)*, 2022.

[25] S. Baumann, M. Grice, and S. Steindamm, "Prosodic Marking of Focus Domains - Categorical or Gradient?," in *Proceedings of Speech Prosody 2006, Dresden, Germany*, Dresden, 2006, pp. 301–304.

[26] A. Gregori, "Co-speech Gestures, Information Structure and Prosody: A Corpus Study on Prominence Peak Alignment," MA thesis, Goethe Universität Frankfurt, Frankfurt am Main, 2022.

[27] M. Grice, S. Baumann, and N. Jagdfeld, "Tonal association and derived nuclear accents-The case of downstepping contours in German," *Lingua*, vol. 119, pp. 881–905, 2009.

[28] J. Peters, *Intonation*. Heidelberg: Universitätsverlag Winter, 2021.

[29] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Kluwer, 1990, pp. 403–439.

[30] J. Hanssen, J. Peters, and C. Gussenhoven, "Prosodic Effects of Focus in Dutch Declaratives," in *Proceedings of Speech Prosody 2008 Conference*, 2008, pp. 609–612.

A KINEMATIC ANALYSIS OF VISUAL PROSODY: HEAD MOVEMENTS IN HABITUAL AND LOUD SPEECH

Lena Pagel[1], Márton Sóskuthy[2], Simon Roessig[3], Doris Mücke[1]

[1] University of Cologne, [2] University of British Columbia, [3] Cornell University
lena.pagel@uni-koeln.de

ABSTRACT

Prosodic prominence manifests itself in intonation, timing and magnitude of supra-laryngeal articulation as well as speech-accompanying gestures. The interplay of prosody and gesture has been described as 'visual prosody' and is known to play an important role in communication. However, few studies have investigated visual prosody across different speaking styles. In this study, we examine co-speech head motion related to prosodic prominence in habitual and loud speech. The results show overall differences between speaking styles as well as some signatures of prosodic prominence, which are stronger in loud than in habitual speech. The paper underlines the potential of a fine-grained kinematic approach to explore continuous speech-accompanying movements.

Keywords: prosodic prominence, multimodality, co-speech gestures, loud speech, focus structure.

1. INTRODUCTION

Prosodic prominence is associated with the packaging of information in discourse, and speakers highlight the information that they deem to be most important for the listener [1], [2]. It is multi-faceted, as its correlates include a wide range of phonetic cues as well as co-speech body movements. It has been shown that prosodically prominent entities are often accompanied by certain events in the stream of co-speech gestures, such as head nods, eyebrow raises or manual gestures [3]–[7]. Since co-speech movements are tightly coupled with speech timing and serve important communicative functions, they have been termed 'visual prosody' [8].

When interacting, speakers adapt their overall speaking style to communicative demands, e.g. they may produce loud speech to be more intelligible in adverse listening conditions. Besides a number of acoustic-articulatory correlates of loud speech, it may also have an effect on speech-accompanying body movements. This is supported by evidence that co-speech head movements are interrelated with acoustic properties of speech, such as F0 and intensity [9], which are two parameters that are altered in loud speech. Additionally, studies on speech in noisy environments (Lombard speech) observed that co-speech hand, head and face gestures are enhanced in this speaking style [10]–[12]. This can have a positive effect for listeners, because co-speech gestures facilitate the perception of speech in noise [8], [13]. However, the majority of existing studies investigate Lombard speech and it is not entirely clear how these results apply to loud speech without background noise.

While loud speech can be understood as a *global increase in production effort* on the utterance level, prosodic prominence represents a *local increase* on the word or syllable level. What remains unclear to date is how these two levels interact: Do speakers use cues of visual prosody across speaking styles? The present paper explores co-speech head movements in habitual and loud speech. First, a between-style comparison assesses how head motion varies as a function of speaking style. Second, a between-focus comparison examines visual correlates of prosodic prominence in habitual and loud speech. While most studies compare prominent vs. non-prominent entities, we investigate gradient adaptations between two *degrees* of prosodic prominence (namely broad vs. corrective focus), where the words bear the nuclear pitch accent in both cases. We adopt a fine-grained kinematic approach that aims to exploit the rich information of the continuous movement signal.

2. METHODS

2.1. Recording procedure

20 native speakers of German (10 female, 10 male) between 22 and 27 years old (mean: 24.7, SD: 1.3) were recorded for this experiment. Kinematic recordings were performed using 3D Electromagnetic Articulography (EMA, AG 501) and a time-synchronised acoustic set-up. Sensors were placed behind both ears and on the bridge of the nose to calculate head movement three-dimensionally. Supra-laryngeal articulation was captured simultaneously but will be reported elsewhere.

Participants were engaged in a game-like task, in which they interacted with a virtual avatar. The game was set in a football stadium and participants were asked to answer the avatar's questions about the match, which were presented auditorily and visually. The answers could be read on the screen and were the speakers' target utterances.

2.2. Speech material

Speakers produced trisyllabic target words, which were German-sounding fictitious names with a CVCVCV structure, with the lexical stress on the penultimate syllable. They were embedded in carrier sentences and were either in *broad* or *corrective focus*. Focus structures were elicited by two kinds of questions asked by the avatar. Table 1 presents example question-answer pairs for the two focus types.

broad focus	
Q:	Was passiert gerade? 'What is going on?'
A:	[Carlotta spielt Nabima zu]$_F$. '[Carlotta passes the ball to Nabima]$_F$.'
corrective focus	
Q:	Spielt Annette Lotte zu? 'Does Annette pass the ball to Lotte?'
A:	Annette spielt [Nabima]$_F$ zu. 'Annette passes the ball to [Nabima]$_F$.'

Table 1: Example of focus elicitation. Target words are underlined, the focus domain is marked by square brackets and the subscript $_F$.

It should be noted that the two focus types differ only in a gradient way, as target words are within the focus domain and bear the nuclear pitch accent in both conditions. However, we can observe two *degrees of prominence*, since words in corrective focus are more prominent than in broad focus [14].

All utterances were produced in *habitual* speech first, without instructions concerning the loudness level. In the second half of the experiment, speakers produced all utterances in *loud* speech. They were told that the football stadium got noisy and they had to speak up for their answers to be understood. Although no actual background noise was played, speakers increased their sound pressure level noticeably.

In total, 960 utterances were recorded (6 target words $\times$ 2 renditions $\times$ 2 focus types $\times$ 2 speaking styles $\times$ 20 speakers), out of which 954 were analysed.

2.3. Analyses

The acoustic annotation of words and segments was carried out using the Montreal Forced Aligner [15] with manual corrections. The post-processing of the kinematic data from the three head sensors was achieved using ema2wav [16]. All analyses were based on a time window that includes the target word and an additional 100 ms before and after. Since the study adopts an approach based on the continuous head motion signal that calculates kinematic and velocity profiles, no manual annotation of gestures or gesture types was carried out.

We present three separate analyses below. The first of these is based on a generalised additive mixed model (GAMM) fitted to *3D movement* trajectories for the three sensors (N = nose, R = right ear, L = left ear) across all conditions. The model includes both fixed effects for the identification of general patterns across participants as well as random smooths to capture across-participant variation. This model is used to create animations and static images that provide a schematic summary of head movements, using a triangle that represents the three sensors in 3D (cf. Figure 1). It serves the purpose of visualising the data.

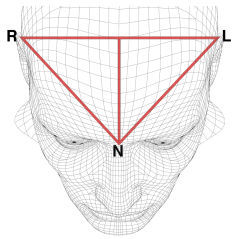


Figure 1: Visualisation of 3D movement calculation.

The remaining analyses are based on the nose sensor only, which showed the greatest degree of movement in our first measure. The second analysis captures the 3D distance between the two furthest points of each trajectory during the time window, giving an estimate of overall movement *displacement*. This measure is analysed using linear mixed effects models. The third analysis is based on an estimate of 3D *velocity* at each time point in each trajectory. This estimate is obtained by calculating the 3D distance between each adjacent time point within each trajectory in the time window, yielding velocity trajectories over time. These trajectories are analysed using GAMMs with fixed and random effects that are set up similarly to the first analysis.

The code for all analyses as well as the full data table can be found in a publicly available repository (https://osf.io/69j2s/?view_only=4ffed19ef11b4bb3b50690beab04cf73).

3. RESULTS

3.1. 3D movement visualisation

With our first measure, we use animations and static images to provide qualitative visualisations of the data. In this paper, we only include a static image taken at the temporal midpoint, but we strongly encourage readers to consult the animated versions of these visualisations in the online repository. The descriptions in the text below are based on the fuller dynamic information present in the videos.

Figure 2 shows static midpoint estimates for six representative speakers, based on the random effects

of the GAMM. The majority of speakers show between-style differences, i.e. a different head position in loud speech, with the head tilted upwards (e.g., S1, S2, S3, S4). However, only a few speakers clearly differentiate focus types (e.g. S1, S3). When focus differentiation does occur (which is more often the case in loud than in habitual speech), it tends to vary in its realisation between individuals. The net effect of this variation is that, when averaging across individuals, most systematic differences are cancelled out.

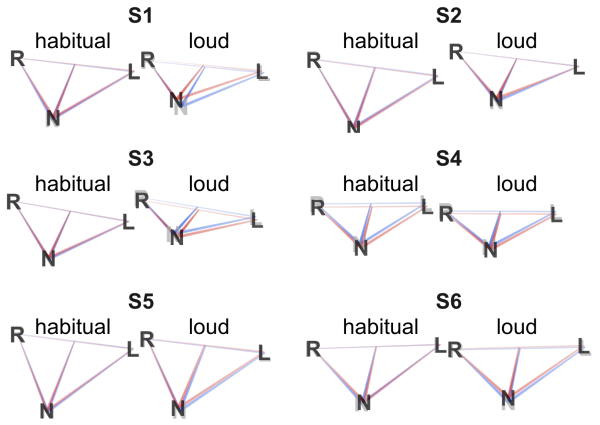


Figure 2: 3D movement results for six example speakers. Blue lines display broad, red lines corrective focus.

3.2. Displacement and velocity

The analysis of 3D displacements can provide more detailed insights (cf. Figure 3). First, a between-style analysis reveals that movements have a greater magnitude in loud than in habitual speech ($\chi^2(2) = 22.8$, $p < 0.001$, based on a model comparison). Second, the effect of focus is assessed through post-hoc comparisons with Tukey's adjustment for multiple comparisons. The results show a significant difference between focus conditions only in loud speech, with head movements being larger in corrective than in broad focus, i.e. with a higher degree of prominence ($\beta = 1.1$, $t(20) = 2.51$, $p = 0.021$).

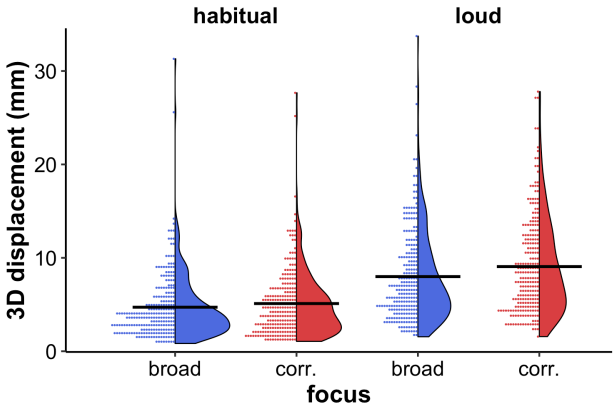


Figure 3: 3D displacement results.

Figure 4 shows results of the 3D velocity over time, based on the fixed effects from our third analysis. A between-style comparison reveals a robust difference between speaking styles: Head movements are faster in loud speech (reflected by a significant parametric difference term: $\beta = 5.32$, $t(1) = 6.97$, $p < 0.001$). A between-focus comparison shows a slight differentiation in both speaking styles: In habitual speech, focus is differentiated by the shape of the velocity profile (smooth difference term: EDF = 1, F = 6.31, $p = 0.012$) and in loud speech by the overall height of the velocity curve, i.e. faster movements are associated with a higher degree of prominence (parametric difference term: $\beta = 2.07$, $t(1) = 3.07$, $p = 0.002$).

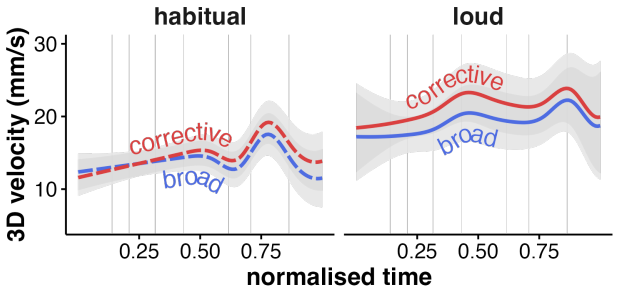


Figure 4: 3D velocity results. Grey vertical lines indicate segment boundaries of the target word.

The robust between-style differences observed at the group-level are also evident across individuals (cf. Figure 5 for six example speakers): In most speakers, there is at least a trend for head movements to be faster in loud than in habitual speech (e.g. S1, S3, S4, S6). A between-focus comparison shows that the modulations observed at the group-level are driven by few speakers who clearly differentiate focus types in at least one speaking style (e.g. S1, S4), while most speakers do not exhibit systematic differences.

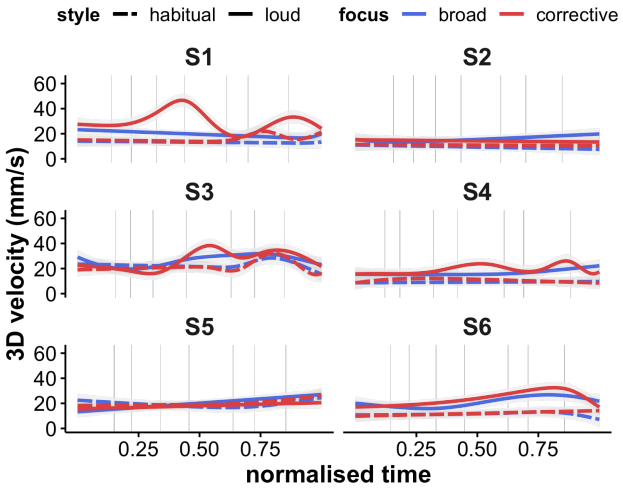


Figure 5: 3D velocity results for six example speakers. Grey vertical lines indicate segment boundaries.

4. DISCUSSION

4.1. Between-style differences

The results show reliable differences between habitual and loud speech in terms of head movements, which can be understood as correlates of global increases in production effort. In loud speech, the head is in a different position (i.e. tilted upwards) and movements are larger and faster. The majority of individual speakers exhibit these patterns of between-style differences and there is relatively little variation.

This is in line with literature showing a strong interrelation between co-speech movements and speech acoustics, namely F0 and intensity [9], which are increased in loud speech. What is more, our results are comparable with those indicating enhanced movements in Lombard speech [11], [12], [17], which underlines the similarity between the two speaking styles. The results thus show that loud speech is a multimodal phenomenon, as already described similarly for Lombard speech [17].

The head modulations in loud speech may at least partly have physiological reasons, since an upward tilt of the head (as well as changes in postural sway) can facilitate effortful speech through changes in larynx position and resonance cavities [18], [19].

4.2. Between-focus differences in both speaking styles

The results show some differentiations between focus types, or correlates of local increases in production effort. They can be interpreted as evidence for visual prosody, which is in line with existing studies [3]–[6]. Namely, co-speech head movements are larger in words associated with a higher degree of prominence, though only in loud speech. Additionally, movements exhibit different velocity profiles between focus types in both speaking styles. Interestingly, there is great inter-individual variation and the group-level results are predominantly driven by few individual speakers who strongly differentiate focus types, whereas the majority do not exhibit systematic modulations.

It can be assumed that the between-focus differences are mainly functionally and only partly physiologically motivated. This is supported by evidence that speakers are aware of communicative demands and make use of visual cues in a functional and listener-oriented way [20]. Perception experiments confirm that co-speech gestures in fact facilitate speech understanding in adverse listening conditions, showing that listeners can successfully interpret the provided multimodal cues [8].

The data show that between-focus differences are stronger in loud than in habitual speech. This suggests that speakers exploit those parameters that contribute best to the transmission of their message [21]. When the auditory channel is disturbed by (real or imaginary) background noise, the visual modality is modulated to a greater extent.

It should be noted that the two focus conditions investigated here are associated with two *degrees* of prominence that differ only in a gradient way on the level of speech production, since the target words bear the nuclear pitch accent in both focus conditions. The results of the study show that even these small differences can manifest themselves in co-speech movements, albeit only subtly, in some speakers and not always in both speaking styles. It should be noted, additionally, that we recorded co-speech movements in a highly controlled lab speech task. We hypothesise that manifestations of visual prosody may be even stronger under more natural circumstances.

4.3. Methodological proof-of-concept

The present study adopts a methodological approach analysing visual prosody in a gradient way. The kinematic analyses take all head movement during a specified time window into account and explore the rich continuous signal. The study underlines the potential of using EMA in multimodal analyses, as previously shown by, e.g., [22]. The 3D recordings can capture the multidimensional head movements occurring during speech, which may be difficult to differentiate in annotations [23].

Since speakers constantly move their heads during speech, clear onsets or targets of gestures may be challenging to detect in a traditional way, especially in the present study, where head movements are comparably small. Instead of scrutinising classifications or certain points of the gesture (e.g. the target), we consider the *movement itself* to be informative, including changes in velocity as an expression of visual biomechanical effort [24]. The methodology also enables a fine-grained exploration of individual differences that might not surface in overall results.

5. CONCLUSION

The present study investigates co-speech head movements in habitual and loud speech. The results show robust differences between speaking styles, namely enhanced movements in loud speech. They also reveal some modulations associated with prosodic prominence in both speaking styles, which are stronger in loud speech. Both findings suggest that the auditory and the visual signal can be modulated in order to meet communicative demands and convey a message. The study further underlines the potential of a multimodal analysis with 3D EMA that focusses on continuous kinematics. We conclude that speakers increase biomechanical power as a concomitant of global and local increases in production effort.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the collaborative research center SFB1252 Prominence in Language (Project-ID 281511265) and project RO 6767/1-1.

REFERENCES

[1] M. A. K. Halliday, *Intonation and grammar in British English*. De Gruyter, 1967. doi: 10.1515/9783111357447.

[2] E. Vallduví and E. Engdahl, 'The linguistic realization of information packaging', *Linguistics*, vol. 34, pp. 459–519, 1996.

[3] G. Ambrazaitis and D. House, 'Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings', *Speech Commun.*, vol. 95, pp. 100–113, 2017, doi: 10.1016/j.specom.2017.08.008.

[4] J. Krivokapić, M. K. Tiede, and M. E. Tyrone, 'A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection', *Lab. Phonol.*, vol. 8, no. 1, pp. 1–26, 2017, doi: 10.5334/labphon.75.

[5] M. Swerts and E. Krahmer, 'Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions', *J. Phon.*, vol. 38, pp. 197–206, 2010, doi: 10.1016/j.wocn.2009.10.002.

[6] Y. Yasinnik, M. Renwick, and S. Shattuck-Hufnagel, 'The timing of speech-accompanying gestures with respect to prosody', *Proc. Sound Sense 11-13 June Camb. USA*, vol. 115, no. 5, pp. 2397–2397, 2004.

[7] N. Esteve-Gibert, J. Borràs-Comes, E. Asor, M. Swerts, and P. Prieto, 'The timing of head movements: The role of prosodic heads and edges', *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4727–4739, 2017, doi: 10.1121/1.4986649.

[8] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, 'Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception', *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[9] H. C. Yehia, 'Linking facial animation, head motion and speech acoustics', *J. Phon.*, vol. 30, pp. 555–568, 2002, doi: 10.1006/jpho.2002.0165.

[10] J. Trujillo, A. Özyürek, J. Holler, and L. Drijvers, 'Speakers exhibit a multimodal Lombard effect in noise', *Sci. Rep.*, vol. 11, p. 167211, 2021, doi: 10.1038/s41598-021-95791-0.

[11] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, 'Audiovisual Lombard speech: Reconciling production and perception', in *Proceedings of the International Conference on Auditory-Visual Speech processing, 31 August - 3 September, Hilvarenbeek, The Netherlands*, 2007, pp. 45–50.

[12] M. Dohen and B. Roustan, 'Co-production of speech and pointing gestures in clear and perturbed interactive tasks: multimodal designation strategies', in *Proceedings of INTERSPEECH, 20-24 August, Stockholm, Sweden*, 2017, pp. 1–5.

[13] L. Drijvers and A. Özyürek, 'Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension', *J. Speech Lang. Hear. Res.*, vol. 60, pp. 212–222, 2017, doi: 10.1044/2016.

[14] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, 'Acoustic correlates of information structure', *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 1044–1098, 2010, doi: 10.1080/01690965.2010.504378.

[15] M. McAuliffe, M. Socolof, S. Mihuc, and M. Wagner, 'Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi', in *Proceedings of INTERSPEECH, 20-24 August, Stockholm, Sweden*, 2017, pp. 498–502. doi: 10.21437/INTERSPEECH.2017-1386.

[16] P. Buech, S. Roessig, L. Pagel, D. Mücke, and A. Hermes, 'ema2wav : doing articulation by Praat', in *Proceedings of INTERSPEECH, 18-22 September, Incheon, Korea*, 2022.

[18] K. Honda, 'Interactions between vowel articulation and F0 cotrol', in *Proceedings of Linguistics and Phonetics: Item Order in Language and Speech, 15-20 September 1998, Columbus, Ohio, USA*, O. Fujimura, B. D. Joseph, and B. Palek, Eds., Prague: Karolinum Press, 2000, pp. 517–527.

[19] A. Lagier, M. Vaugoyeau, A. Ghio, T. Legou, A. Giovanni, and C. Assaiante, 'Coordination between Posture and Phonation in Vocal Effort Behavior', *Folia Phoniatr. Logop.*, vol. 62, pp. 195–202, 2010, doi: 10.1159/000314264.

[20] M. Garnier, L. Ménard, and B. Alexandre, 'Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?', *J. Acoust. Soc. Am.*, vol. 144, pp. 1059–1074, 2018, doi: 10.1121/1.5051321.

[21] M. Fitzpatrick, J. Kim, and C. Davis, 'The effect of seeing the interlocutor on auditory and visual speech production in noise', *Speech Commun.*, vol. 74, no. August, pp. 37–51, 2015, doi: 10.1016/j.specom.2015.08.001.

[22] S. Gordon, A. Vilela Barbosa, and L. Goldstein, 'Quantitative analysis of multimodal speech data', *J. Phon.*, vol. 71, pp. 268–283, 2018, doi: 10.1016/j.wocn.2018.09.007.

[23] S. Kousidis, Z. Malisz, P. Wagner, and D. Schlangen, 'Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction', *Proceedings of the Tilburg Gesture Meeting (TiGeR), 19-21 June, Tilburg, The Netherlands*. pp. 1–4, 2013.

[24] B. Lindblom, 'Explaining Phonetic Variation: A Sketch of the H&H Theory', in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Kluwer Academic Publishers, 1990, pp. 403–439. doi: 10.1007/978-94-009-2037-8_16.

ORO-FACIAL EXPRESSIONS AND ACOUSTIC CUES IN GERMAN QUESTIONS

Marzena Żygis[1,2], Nasim Sarhaei[1], Manfred Krifka[1,2]

[1]Leibniz-ZAS, [2]Humboldt Universität zu Berlin
zygis@leibniz-zas.de, sardhaei@leibniz-zas.de, krifka@leibniz-zas.de

ABSTRACT

The present paper investigates acoustic cues and oro-facial expressions produced during the production of word-identical German sentences that represent eight different pragmatic uses (assertion, exclamation, and six varieties of questions). Our results based on 19 native speakers of German reveal that facial movements are sensitive to different sentence types and fine-grained pragmatic differences. Whereas exclamatives and assertions were produced with the highest raised eyebrows, incredulity, echo, and guessing questions showed the most intense eyebrow furrowing. In other question types (confirmation-seeking, relevance, and justification) the oro-facial expressions were less extreme. The data are also subject to huge inter-speaker variation.

Keywords: oro-facial expressions, acoustics, questions, German

1. INTRODUCTION

Questions and responses can be considered as fundamental building blocks in the structure of a conversation [1, 2, 3]. It has been frequently noticed that assertive and interrogative meanings are produced with specific auditory cues [4, 5, 6, 7]. On the other hand, audiovisual speech research has revealed the relative importance of facial cues with respect to auditory cues for signaling the prosodic functions. The role of visual information integrated with auditory information has been frequently addressed in studies on the production and perception of different sentence types. For instance, some studies have observed that eyebrow movements tend to co-occur with questions. To name a few, [8] demonstrated that pronounced eyebrow raising and head tilt enhanced the perception of echo questions in American English. [9] also reported that eyebrow raising and head movements tended to occur in French polar questions rather than assertions, and [10] showed that incredulity questions are characterized by some degree of eyebrow furrowing and eyelid closure in Dutch and Catalan. Similarly, lowering of eyebrows plus head yaw (turning left and right) was employed to mark the production of wh-questions in Brazilian Portuguese [5]. As another investigated visual cue, [11] observed that Dutch and Catalan speakers directed their gaze at their interlocutors more frequently in information-seeking questions than in assertions. Analyzing face-to-face conversation, [12] noticed that eyebrow furrows were used as non-manual grammatical markers of content questions of requests such as 'who?', and eyebrow raises as markers of polar questions to serve as the restricted offers of repair such as 'John Smith?'.

All these studies show that the production of specific facial gestures may vary not only according to the sentence type, but also according to subtle pragmatic meanings of the specific sentence, as well as language-specific differences. Building on these insights, the aim of this work is to compare the use of visual cues in the production of various sentence types, in particular questions, in German. In addition, we will also complement the study by acoustic investigations.

The present contribution is broader in its scope than previous works. Whereas they concentrated on just one or a few sentence types, we take into account six potentially distinct question types and compare them to assertions and exclamatives. Furthermore, previous studies concentrated on the movements of the eyebrows and the head, leaving aside the eye and lip opening. Hence, the aim of this work is to extend both the repertoire of questions under investigations and the spectrum of oro-facial expressions accompanying the questions.

2. EXPERIMENTAL EVIDENCE

2.1. Experimental design

To meet our research goal we conducted simultaneous video recordings of participants' faces and their voices while they were producing sentences displayed on a computer screen, see details below. The video recordings were made by means of a camera (Canon EOS1300 D) and for the acoustic recordings we used a microphone (Sennheiser ME 64) to obtain acoustic data of high quality. The camera was positioned ca. one meter in front of a participant face, and the microphone was positioned

about 30 cm from their mouth. Video data were recorded with a sampling frequency of 25 Hz and acoustic data with a sampling rate of 44 100 Hz. The recordings took place in a sound-proof laboratory at Leibniz-ZAS in Berlin.

Our material consisted of six question types, as presented in (1), as well as assertions and exclamatives. All questions were in the form of so-called non-interrogative [13] or declarative questions [14] that do not differ in their morphosyntactic marking from assertions, thus enabling a direct comparison between the sentences. Please note that the American English corpus of [15] declarative questions are frequent in conversation; they make up 68% of the polar questions, which themselves make up 70% of the questions. They are frequent in German as well, see [16], [17] for their prosodic realization.

(1) Question types (all declarative questions)

- Echo questions
- Incredulity questions
- Confirmation seeking questions
- Guessing
- Relevance
- Justification

In addition, two other sentence types, i.e. assertion and exclamative, were included to create a contrast with the question types listed in (1).

Each sentence type complemented a different scenario, but the sentence itself was identical across scenarios. The only difference was the punctuation, i.e., questions (as in (1)) were presented with a question mark, the assertion with a period and the exclamative with an exclamation mark. We created six short SVO sentences that appeared in all scenarios, see (2).

(2) Sentences produced in distinct scenarios

(i) *Maria fährt nach Algerien*
'Maria goes to Algeria'
(ii) *Julia bekommt einen Labrador*
'Julia gets a Labrador'
(iii) *Michaela hat einen Segelflugschein*
'Michaela has a glider pilot's license'
(iv) *Thomas kauft einen Cadillac*
'Thomas buys a Cadillac'
(v) *Martin bringt eine Rhabarbersuppe.*
'Martin brings a rhubarb soup'
(vi) *Lydia spielt Bandoneon*
'Lydia plays bandoneon'

An example of a scenario eliciting an incredulity question in presented in (3).

(3) Scenario

Maria works in real estate and has a lot on her plate at any given time. She has decided to take a vacation soon. In the cafeteria, her colleague Steffi tells her office neighbor Peter that Maria wants to take some time off for vacation in Algeria. Peter is very surprised, because Maria usually goes on vacation to Sylt. He says:

- **"Maria is going to Algeria?"**

Steffi says: "Yes, I was also very surprised".

Participants were asked to first read the full text, imagine the situation and then pronounce the sentence provided in bold in the way how it would be uttered in the provided scenario. A Latin square design was used so that each participant produced each sentence, in one of eight scenarios. There were no repetitions of identical sentence and scenario sets.

2.2. Participants

19 native speakers of German (12 female and 7 male, aged 22-43, mean age 26.4, sd 5.3) took part in the experiment. 17 participants were right- and two left-handed.

2.3. Measurements

To measure facial movements we used OpenFace 2.0. toolkit [18,19], an implementation of face recognition with deep neural networks, see also [20].

Figure 1 presents the positions of 68 markers mapped on participants' faces.

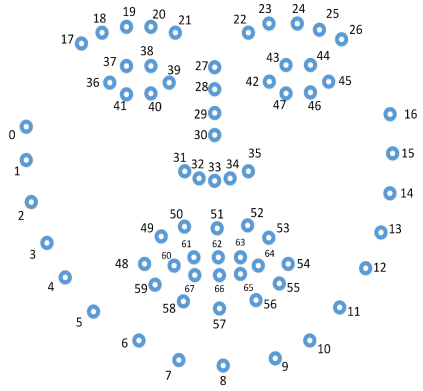


Figure 1: The markers chart used in Open Face ([19]).

Based on the mapping of the markers we measured distances between different markers corresponding to various oro-facial movements, as indicated in (4). In order to normalize the distances, we divided them by the so-called nose normalisation distance placed between markers 32 and 34. The normalisation

procedure provided a reference for distance measurements in (4) and mitigated against possible forward/back head movements.

(4) Measured distances:

(i)	right eyebrow rising	33 - 19
(ii)	left eyebrow rising	33 - 24
(iii)	eyebrow furrow	21 - 22
(iv)	lip opening	51 - 57
(v)	left eye opening	43 - 47
(vi)	right eye opening	38 - 40

In addition, we also performed an acoustic analysis in PRAAT (version 6.1.12 [21]), excerpting duration, RMS amplitude and max and mean pitch in ERB at the level of segments, syllables and words. In addition, each sentence was divided into seven intervals to normalize for different length. The present paper limits the presentation of acoustic results to mean pitch and to the domain of sentence.

2.4. Statistics

The statistical analysis of the data was conducted in the R Studio software (version 4.0.0, [22]) by using the lmer4 ([23]) and the emmeans packages ([24]).

Linear mixed effect models were employed for assessing the influence of SENTENCE TYPE [echo, incredulity, confirmation seeking, guessing, relevance, justification, assertion, echo, exclamative] on LEFT and RIGHT EYEBROW, EYEBROW FURROW, LIP OPENING, LEFT and RIGHT EYE OPENING. In addition, PARTICIPANT and WORD were included as random intercepts. SENTENCE TYPE was added as by-PARTICIPANT slope but due to the non-convergence had to be removed. The mean values over the intervals were fitted to the model. Finally, we corrected for multiple testing by using the Tukey correction from the emmeans package ([24]).

2.5. Results

Figure 2 presents results for mean F0 excerpted for seven equal intervals of each sentence type. It is evident that exclamations and assertions are produced with lowered F0 in the sentence-final position (target word) and all question sentences are produced with a continuously rising F0 in the same position. The statistical modelling shows no significant differences in F0 between exclamations and assertions (t=0.809, n.s.) but significant differences between exclamations and the six questions types: confirmation (t=-5.23, p<.001), echo (t=-4.45, p<.001), guessing (t=-3.93, p<.0.01), incredulity (t=-5.73, p<.001), justification (t=-3.54, p<.05), and relevance (t=-3.86, p<.01). Similarly, assertions were also produced with a significantly lower F0 in comparison to the six question types: confirmation (t=-6.49, p<.001), echo (t=-5.62, p<.001), guessing (t=-5.06, p<.001), incredulity (t=-7.06, p<.001), justification (t=-4.56, p<.001), and relevance (t=-4.99, p<.001). All comparisons between different question types (confirmation, echo, guessing, incredulity, justification, relevance) turned out to be not significant.

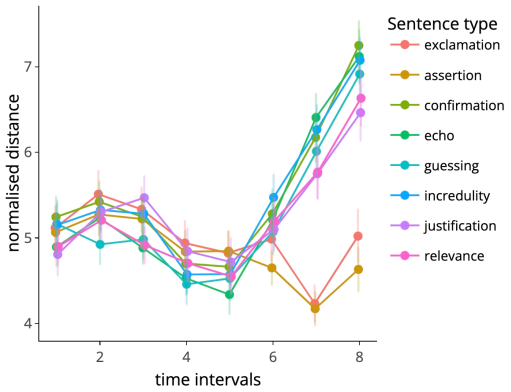


Figure 2: Mean F0 across different sentence types

Figure 3 presents the results obtained for the left and right eyebrow position during the entire sentence across all sentence types.

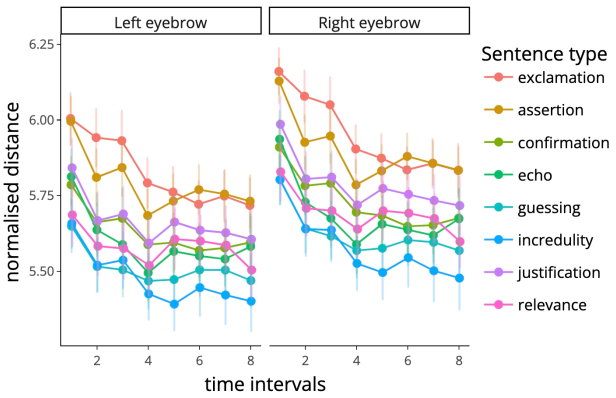


Figure 3: Left and right eyebrow positions across different sentence types

It appears that although the movement of the eyebrows shows a similar pattern across all sentence types, the right eyebrow was generally higher in comparison to the right eyebrow. Furthermore, exclamations and assertions showed the highest positions and incredulity questions the lowest position of the eyebrows. Statistical modelling reveals that several question types were produced with a significantly lower left and right eyebrow in comparison to exclamations and assertions. In fact, in the case of exclamations all comparisons referring to

the right eye-brow, besides the one with justification (t=2.39, n.s.), were significant: confirmation (t=3.16, p<.05), echo (t=3.80, p<.01), guessing (t=4.20, p<.001), incredulity (t=5.35, p<.001), and relevance (t=3.62, p<.01). As for assertions, they were produced with a significantly higher right eye-brow as compared to echo (t=3.47, p<.05), guessing (t=3.91, p<.01), incredulity (t=5.11, p<.001), and relevance (t= 3.29, p<.05). Similar effects were found for the left-eyebrow. The difference between exclamations and assertions was neither significant for the left eyebrow (t=0.042, n.s.) nor for the right eyebrow (t=0.491, n.s.).

Furthermore, as shown in Figure 4 the greatest distance between the eyebrows, i.e., the least eyebrow furrowing, is found in exclamations and assertions and the most intense eyebrow furrowing is produced in echo, guessing and incredulity. The differences are significant: exclamation vs. echo (t=3.74, p<.01), guessing (t=3.78, p<.01), incredulity (t=4.61, p<.001); assertion vs. echo (t=3.45, p<.05), guessing (t=3.52, p<.05), incredulity (t=4.38, p<.001); The differences between exclamations and assertions (t=0.44, n.s) and other questions types remain not significant.

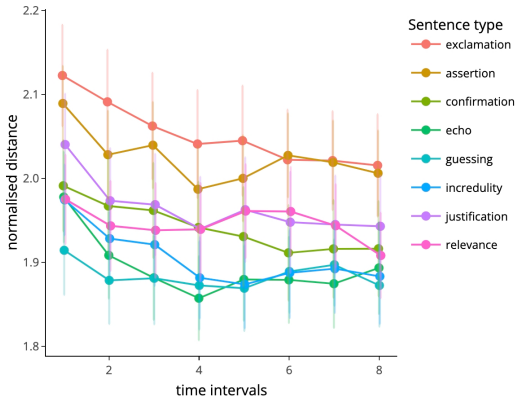


Figure 4: Eyebrow furrow across different sentence types

Our results also show that the exclamations were produced with the highest lip opening, especially in the initial and middle part of the sentence, but the differences turned out to be not significant.

Finally, regarding the left and right eye opening, it appears that it was larger at the beginning and especially at the end of sentences with exclamations and assertions showing again more eye opening, see Figure 5. However, the only significant difference was found for the left eye opening between exclamation and echo question being higher for the former (t= 3.13, p<.05). No significant differences were found for the right eye opening.

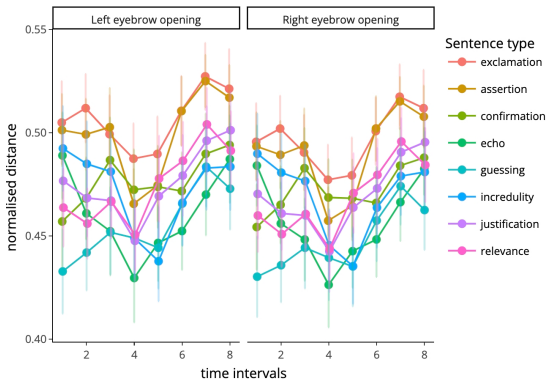


Figure 5: Left and right eye opening across different sentence types

3. DISCUSSION AND CONCLUSIONS

The present study, although limited in its scope and restricted to sentence domains, shows that oro-facial movements are sensitive to different sentence types. Considering the fact that at least some of the questions showed a very similar F0 movement, it appears that oro-facial expressions may deliver important cues to the interpretation of sentences in face-to-face communication. Further research, including perceptual evidence, should (dis)confirm this hypothesis.

Our study reveals that exclamations are produced with more raised eyebrows, (and more opening of the left eye) in comparison to other sentence types. Assertions are produced in a similar manner but with slightly less pronounced gestures. Also, for these two types of sentences F0 is lowered (at the sentence-final position), which suggests a compensatory effect between acoustic cues and oro-facial expressions, (see [25]), a point that requires further investigation at the syllable level. Furthermore, the most intense eyebrow furrowing is produced in echo, guessing and incredulity. These sentence types are also produced with the lowest eyebrow positions. Other sentence types are produced with oro-facial expressions that are generally less extreme. It should also be pointed out that there was a huge inter-speaker variation, as indicated by large confidence intervals.

It is also worth pointing out that the eyebrows did not move in the same manner. The right eyebrow was more raised than the left eyebrow, a result also previously found for questions in German [25]. However, [26] found for French a greater variation in movement magnitude for the *left* eyebrow, a result that the authors tentatively attributed to different involvement of brain hemispheres in the communication process. Another explanation could refer to interculture differences where questions are marked either by right or left eyebrow raising, a point that also requires further investigation.

Acknowledgments
Research for this paper was done in the context of the ERC Advanced Grant SPAGAD: Speech Acts in Grammar and Discourse, receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 787929. We would like to thank all participants for taking part in our experiment.

7. REFERENCES

[1] Farkas, D. F., Bruce, K.B. 2010. On reacting to assertions and polar questions. *Journal of Semantics* 27, 81-118.
[2] Sendra, V. C., Kaland, C., Swerts, M., Prieto, P. 2013. Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics* 47, 1–13.
[3] Krifka, M. 2022. Adjacency pairs in common ground update: Assertions, questions, greetings, offers, commands. *Proceedings of the 26th Workshop on the Semanics and Pragmatics of Dialogue,* 94-105.
[4] Moraes, J. 1998. Intonation in Brazilian Portuguese. In: Hirst, D., Di Cristo, A. (eds.), *Intonational Systems: a survey of twenty languages*. Cambridge, MIT Press.
[5] Moraes, J. 2008. The pitch accents in Brazilian Portuguese: analysis by synthesis. *Proc. of the 4th International Conference on Speech Prosody* Campinas, 389-397.
[6] Miranda, L. S., Moraes, J. A., Rilliard, A. 2019. Audiovisual perception of wh-questions and wh-exclamations in Brazilian Portuguese. *Proc. of the 19th International Congress of Phonetic* Melbourne, 2941-2945.
[7] Enfield, N.J., Stivers, T., Levinson, St-C. 2010. Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics* 42, 2615-2619.
[8] Srinivasan, R. J., Massaro, D. W. 2003. Perceiving from the face and voice: distinguishing statements from echoic questions in English. *Language and Speech* 46, 1–22.
[9] Torreira, F., Valtersson, E. 2015. Phonetic and Visual cues to questionhood in French conversation *Phonetica* 72, 20–42.
[10] Crespo-Sendra, V., Vanrell, M. D. M., Prieto, P. 2010. Information-seeking questions and incredulity questions: gradient or categorical contrast? *Proc. of Speech Prosody* 2010 Chicago.
[11] Borràs-Comes, J., Kaland, C., Prieto, P., Swerts, M. 2014. Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior* 38, 53-66.
[12] Hömke, P., Levinson, S. C., Holler, J. 2022. Eyebrow movements as signals of communicative problems in human face-to-face interaction. https://doi.org/10.31234/osf.io/3jnmt
[13] Bartels, C. 1997. *Towards a compositional interpretation of English question and statement intonation*. Doctoral dissertation, University of Massachusetts at Amherst.
[14] Gunlogson, Ch. 2002. Declarative questions. Proceedings of *SALT XII*. Ithaca, NY: Cornell University, 124-134.
[15] Stivers, T. 2010. An overview of the question–response system in American English conversation. *Journal of Pragmatics* 42, 2772-2781.
[16] Niebuhr, O., Bergherr, S. Huth, C. Lill, J. Neuschulz 2010. Intonationsfragen hinterfragt. Die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax. *Zeitschrift für Dialektologie und Linguistik* 77, 304-346.
[17] Petrone, C., Niebuhr, O. 2014. On the intonation of German intonation questions: the role of the prenuclear region. *Language and Speech* 57, 108-146.
[18] Baltrušaitis, T., Robinson, P., Morency, L-P. 2016. *IEEE Winter Conference on Applications of Computer Vision* (WACV).
[19] Baltrušaitis, T., Zadeh, A., Lim, Y. Ch., Morency L.-Ph. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*. https://par.nsf.gov/servlets/purl/10099460
[20] Schroff, F., Kalenichenko, D., Philbin, J. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 815-823. doi: 10.1109/CVPR.2015.7298682.
[21] Boersma, P., Weenink, D. 2020. Praat: Doing phonetics by computer (version 6.1.12) [computer program], http://www.praat.org/
[22] RStudio Team 2020. RStudio: Integrated Development for R (version 4.0.0) (RStudio, Boston).
[23] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J. 2020. Package 'lme4' (version 1), https://cran.r-project.org/web/packages/lme4/lme4.pdf
[24] Lenth, R. V. 2019. Package: 'emmeans: Estimated mariginal means, aka least-squares means', https://cran.r-project.org/web/packages/emmeans/index.html
[25] Żygis, M., Fuchs S. 2023. Communicative constraints affect oro-facial gestures and acoustics: Whispered vs normal speech. *Journal of the Acoustical Society of America.* Paper number 07819.
[26] Cavé, Ch., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. 1996. About the relationship between eyebrow movements and F0 variations. *Proceeding of Fourth International Conference on Spoken Language Processing*, 2175-2178.

COMPUTATIONAL MODELS FOR ARTICULATORY LEARNING OF ENGLISH DIPHTHONGS: ONE DYNAMIC TARGET VS. TWO STATIC TARGETS

Anqi Xu[1], Branislav Gerazov[2], Daniel van Niekerk[3], Paul Konstantin Krug[4], Santitham Prom-on[5], Peter Birkholz[4], Yi Xu[3]
[1]School of Humanities and Social Sciences, Harbin Institute of Technology, Shenzhen, China
[2]Faculty of Electrical Engineering and Information Technologies, UCMS, Skopje, RN Macedonia
[3]Department of Speech Hearing and Phonetic Sciences, University College London, UK
[4]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[5]Computer Engineering Department, King Mongkut's University of Technology Thonburi, Thailand
a.xu.17@ucl.ac.uk

ABSTRACT

The nature of English diphthongs has been much disputed. By now, the most influential account argues that diphthongs are phoneme entities rather than vowel combinations. However, mixed results have been reported regarding whether the rate of formant transition is the most reliable attribute in the perception and production of diphthongs. Here, we used computational simulation of vocal learning to explore the nature of diphthongs. We tested whether diphthongs have a single dynamic target or two static targets by training an articulatory synthesiser with a three-dimensional (3D) vocal tract model to learn English words. An automatic phoneme recogniser was constructed to guide the learning of the diphthongs. Listening results by native listeners showed that diphthongs learned with dynamic targets were more intelligible and generalisable across variable duration than those learned with static targets. The learning-oriented modelling approach also paves a new way for validating hypotheses of speech perception and production.

Keywords: diphthongs, computational modelling, articulatory synthesis, vocal learning, American English

1. INTRODUCTION

Diphthongs, a special class of vowels, are characterised by transitional formant movements along a path between spectral spaces belonging to two different vowels [1], [2]. Early accounts have treated diphthongs as combinations of two vowels, or sometimes as vowel-semivowel sequences [3]. However, empirical evidence from a comprehensive study by Gay [4] suggests that English diphthongs are more likely to be distinct phonetic units, based on the observation that listeners were more sensitive to the second formant (F2) movement of the synthetic diphthongs than the formant onset and offset. These results are consistent with more recent findings that the most salient perceptual cue of synthetic diphthongs in noise or reverberation is the intensity of F2 transitions [5]. On the other hand, some studies suggest that the crucial cue in the identification of manipulated diphthongs is the endpoint rather than the transitional trajectories [6]. Another line of studies sought to use a classifier to investigate perceptual cues of diphthongs in a speech corpus, and found that instead of F1-F2 onsets and slopes, classification accuracy was the highest when both F1-F2 onsets and offsets were included [7]. A similar approach was adopted in [8], which reported that incorporating F1–F3 onset, offset and transition rates led to the best classification results.

Not only does the debate about the auditorily relevant formant cues of diphthongs continue, contradictory observations have been made regarding the production of diphthongs. Gay [9] investigated the acoustic properties of five American English diphthongs spoken in three different speech rates from slow to fast. The beginning and terminating vowel formants as well as the rate of F2 movement remained the same across different speaking rates. Further, the final portion of the vowel could be eliminated in fast speech. The unfluctuating formant slopes also accords with more recent acoustic evidence from careful and conversational speech [10], as well as loud speech [11]. As far as articulation is concerned, the tongue body exhibits invariant velocity during the production of diphthongs [12]. Recent X-ray data also show that the tongue flesh points undergo minimal changes in different speaking rates [10]. More importantly, the tongue movements and formant transitions of diphthongs are highly correlated, despite some exceptions [13]. In contrast, some researchers found that spectral changes of diphthongs were lowered in clear speech with prosodic prominence [14].

If gliding movements rather than onsets and offsets are a most reliable feature of diphthongs, then diphthongs can be considered distinct phonemes. However, to date no evidence from either perception or production studies has been conclusive. Here, we

used computational simulation of vocal learning to explore whether diphthongs have either unitary dynamic targets or consecutive static targets. We trained an articulatory synthesiser with a 3D vocal tract model to learn American English words containing diphthongs, following the approach in [15]–[17]. The learning is guided by a phoneme recogniser, pre-trained by a deep learning model to encode a speaker-normalised perceptual space. After the training, we synthesised speech using the learned articulatory targets with varying duration to further verify their generalisability over different speaking rates. The performance of the two types of articulatory targets is evaluated by the intelligibility of the learned speech in a listening experiment and the plausibility of the learned articulatory kinematics.

2. METHOD

2.1. Speech material

Five diphthongs, /aɪ, eɪ, əʊ, aʊ, ɔɪ/, were embedded in English words with bilabial stops, as follows: 'buy', 'bay', 'boy', 'bow (and arrows)' and '(to) bow'. Because the two target words 'bow' are homographs, we added hints to distinguish the two words, as indicated in brackets. The use of these minimal pairs is to ensure that perception experiments can be carried out naturally by native speakers. The same hints were also given to the participants during the listening experiment.

2.2. Model overview

We trained a vocal tract model to learn the speech material as illustrated in Fig. 1. The learning model consists of a production and a perception system. The model begins with exploration of a set of articulatory targets (Fig. 1A). The kinematic trajectories that approach the articulatory targets are based on the assumption of static or dynamic targets (Fig. 1B). The time-varying vocal tract shapes are then converted to cross-sectional area functions for acoustic simulation (Fig. 1C). The model explores the articulatory parameters iteratively, guided by the perception system. (Fig. 1D).

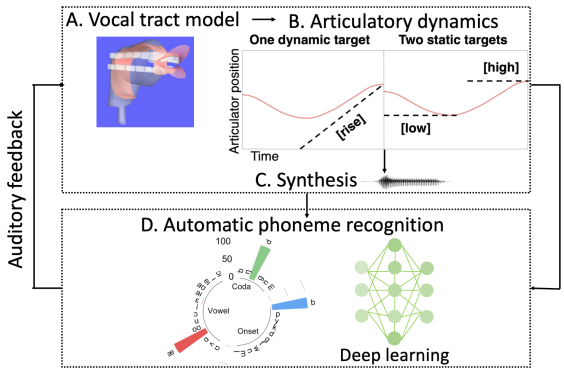


Figure 1: Overview of the learning process.

2.3. Vocal tract model

The articulatory synthesiser used in the study is VocalTractLab 2.3 (www.vocaltractlab.de), with a geometrical 3D vocal tract model (Fig. 1A). The vocal tract model was adapted from MRI data of a German male speaker, involving sixteen free vocal tract parameters (Table 1).

Parameter	Description
HX, HY	Horiz. and vert. hyoid positions
JX, JA	Horiz. jaw position and jaw angle
LP, LD	Lip protrusion and vert. lip distance
TTX, TTY	Horiz. and vert. tongue tip positions
TBX, TBY	Horiz. and vert. tongue blade positions
TCX, TCY	Horiz. and vert. tongue body centre positions

Table 1: Free vocal tract parameters in the simulation.

2.4. Articulatory targets

The temporal and spatial movements of the articulators were simulated by a coarticulation model, synchronised dimension-specific sequential target approximation model [15]–[17], which generates dynamic trajectories of vocal tract parameters. Quantitatively, each articulatory target is represented by height (i.e., positions of the articulators), slope and strength. As illustrated in Fig. 1B, the same asymptotic articulatory curves of diphthongs can be the result of two static targets or one dynamic target. The two static targets have a slope of zero but the target height and the duration ratio need to be optimised. In contrast, the dynamic target requires the optimisation of both height and slope.

2.5. Automatic phoneme recogniser

We trained a deep learning-based phoneme recognition system to guide the optimisation process. We extracted speech with 11 onset consonants, 17 vowels (stressed and unstressed) and 6 coda consonants from the LibriSpeech corpus [23]. The training, validation and test set contains 116.7, 14.4 and 15 hours of speech, respectively. We applied pre-emphasis (coefficient = 0.97) and calculated the log Mel spectrogram (25 ms Hamming window, 5ms overlap) with 26 Mel filters. The log Mel spectrograms were used as the input for the training with a length of 200 frames (spanning 1 s). The model was trained to learn a mapping from the Log Mel spectrograms to a 34-dimensional vector one-hot encoding the phonemes listed in Fig. 1D.

2.6. Optimisation

We use simulated annealing [22] to optimise the vocal tract parameters, which is a stochastic algorithm that seeks an optimal solution through a coarse-to-fine criterion, suitable for problems with many degrees of freedom, such as the speech production system. The learning process started from a neutral position (schwa) followed by adjustments of the vocal tract parameters and gradually converged to a solution. We initiated 10 processes in parallel for each target word, each with 4k iterations.

2.7. Listening experiment and statistical analysis

After optimisation, we selected five items with the lowest recognition errors for both the static and dynamic articulatory targets. In addition to the original duration of 400ms, we synthesised the target words with longer (450ms and 500ms) and shorter (350ms and 300ms) durations to examine the generalisability over speaking rates. For the static targets, the duration ratio of the two targets was kept the same as the learned ratio.

The listeners were 20 American English native speakers (male: 12; mean age: 36), invited and screened via Prolific. The stimuli were randomised and presented to the participants via Gorilla. Before the experiment, the participants filled a brief questionnaire for demographic and language background. Listeners were asked to undertake the experiment on a computer in a quiet environment. A headphone screening was conducted [23] and five practice trials were presented. In the experiment, participants were instructed to listen to the audio carefully up to five times and choose the word that they heard from the word list. The experiment lasted around 20 minutes.

After the experiment, non-parametric statistical analysis, Wilcoxon rank sum tests with Bonferroni corrections were chosen for statistical analysis, due to the skewed distribution of the data.

3. RESULTS

The simulation learned intelligible English words containing diphthongs. A demonstration video and learned synthetic samples can be found in https://gitlab.com/Anqi_Xu/dynamic_diphthongs.

The identification accuracy of the learned diphthongs across target words judged by native listeners is shown in Fig. 2. Error bars show standard errors. The average accuracy was 34.36% and 64.92% for diphthongs synthesised with two static targets and one dynamic target, respectively. The single dynamic target yielded diphthongs that were significantly more intelligible than those synthesised with two static targets ($W = 7514$, $p < .001$) with only except for /aɪ/. The difference was significant for /aɪ, eɪ , ɔɪ/ ($p < .05$) but not for /aʊ/ and /əʊ/. /eɪ/ and /ɔɪ/ with dynamic targets had fairly high accuracy, both over 90%, whereas the two static targets had the highest accuracy for /aɪ/.

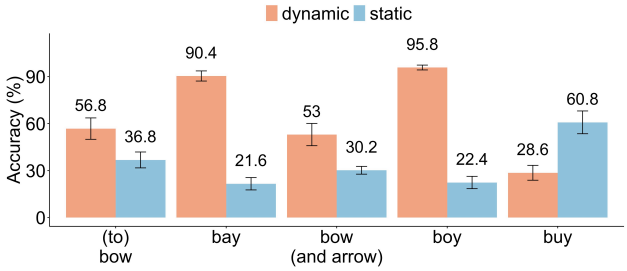


Figure 2: Identification accuracy of words with diphthongs modelled with two static targets or one dynamic target.

The identification accuracy of the diphthongs with different duration is shown in Fig. 3. Again, across all the temporal modulations, not only did the diphthongs with the dynamic target have higher accuracy than those with two static targets for the original duration ($p < .001$), but also for the longer and shorter durations ($p < .001$).

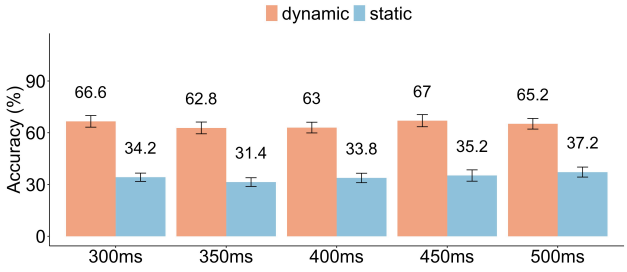


Figure 3: Identification accuracy of words with diphthongs modelled with two static targets or one dynamic target.

Fig. 4 shows the dynamic changes of the learned vocal tract shapes with the lowest recognition error for the target words. The first and second graphs in each row show the starting and ending vocal tract shapes of the CV syllables containing diphthongs. Take /aɪ/ for example, for both the static and dynamic targets, the initial tongue position is relatively low, which later moves towards a higher position. The initial tongue shapes of /eɪ/ seem to be a mid vowel and the terminating tongue position is high for both conditions, whereas the changes in tongue shapes are larger for the dynamic target. /aʊ/ have nearly identical terminating tongue shapes in the two conditions, but the initial tongue position is rather different. The dynamic and static targets both involve little tongue movements for /əʊ/. Finally, in the case of /ɔɪ/, the tongue shapes are retracted in the beginning for both conditions but the dynamic target ends at a higher and more front position. Overall, both the learned articulatory targets based on the static and dynamic targets exhibited beginning and ending

vocal tract shapes that resembled two different vowels, whereas the dynamic target showed slightly more plausible vocal tract shapes than the static targets.

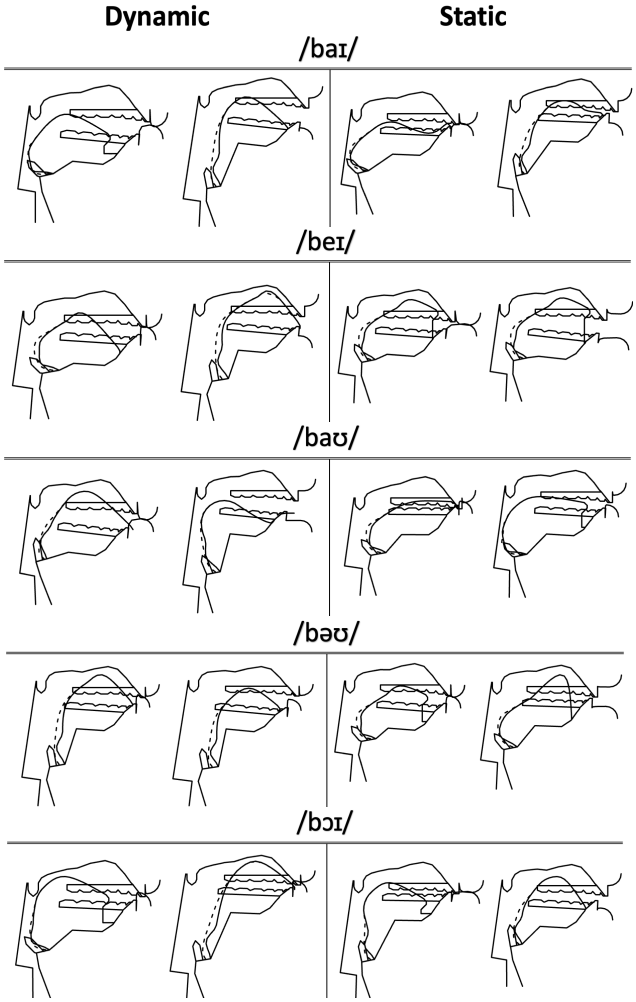


Figure 4: Midsagittal sections of the learned vocal tract shapes with the lowest recognition error.

4. DISCUSSION

We have adopted a new approach to probe the nature of diphthongs via computational simulation of vocal learning. We tested the hypothesis that diphthongs are either a single dynamic articulatory target or two static targets by training a vocal tract model to learn English diphthongs embedded in real words under the guidance of a phoneme recogniser. The results show that unitary dynamic targets can generate more intelligible speech than consecutive static targets. When synthesising words with longer and shorter durations, the dynamic targets also showed more advantage. The learning simulation thus offers new evidence that diphthongs are likely to be independent phonetic entities with unitary dynamic targets.

The theoretical account of diphthongs as unit phonemes was originally proposed on the basis that formant transition stayed constant in varying speech rates [9], but counterevidence emerged subsequently [14]. We have used a new methodology to address the controversy by emulating the dynamic movement of diphthongs under different temporal modulations. The dynamic targets generated more intelligible speech than the static targets, except for /aɪ/. The benefit of having a dynamic target held true even when the speaking rate was modified, which is consistent with the unfluctuating formant slopes observed in [4], [9], [10], [12]. These findings show that diphthongs may have underlying dynamic targets, supporting the proposal of Gay [9].

A main novelty of this study is to investigate diphthongs by simulating its articulatory learning. Previous research has identified various auditory signatures of diphthongs, such as F2 transition rates [4], [5], diphthong endpoints [6] , F1-F2 onset and offset [7] and all of the above [8]. But the lack of consensus suggests that different studies may have observed different manifestations of the diphthongs due to not only differences in methodology, but also cross-speaker variations. During the acquisition of diphthongs, learners necessarily have to deal with cross-speaker variations. A successful learning strategy is likely one that has focused on the most critical property of diphthongs. Thus, learning simulation may be an effective way of discovering the core property of diphthongs. The power of learning simulation as a means of discovering the core mechanisms of speech has been previously demonstrated [15]–[17]. The present study is a further confirmation of its effectiveness.

One source of weakness of this study is that the speech data for training the phoneme recogniser is not balanced across all the speech sequences. This imbalance may have led to varied identification accuracy of the recogniser which was a possible source of the uneven learning performance of the diphthongs. Also, the scope of this study is limited to English. Given that there are noticeable cross-linguistic differences in both the perception and production of diphthongs [18], [19], further research can explore how diphthongs in other languages should be modelled. These limitations notwithstanding, the present study contributes insights into the dynamic nature of English diphthongs. The computational approach opens a new path towards examining theoretical constructs in speech production and perception.

5. ACKNOWLEDGEMENTS

This work has been funded by the Leverhulme Trust Research Project Grant RPG-2019-241: "High quality simulation of early vocal learning".

6. REFERENCES

[1] I. Lehiste and G. E. Peterson, “Transitions, glides, and diphthongs,” *J Acoust Soc Am*, vol. 33, no. 3, pp. 268–277, Mar. 1961, doi: 10.1121/1.1908638.

[2] A. Holbrook and G. Fairbanks, “Diphthong formants and their movements,” *J Speech Hear Res*, vol. 5, no. 1, pp. 38–58, Mar. 1962, doi: 10.1044/jshr.0501.38.

[3] G. L. Trager and H. L. Smith, *An outline of English Structure*. Norman, Oklahoma: Battenburg Press, 1951.

[4] T. Gay, “A perceptual study of American English diphthongs,” *Lang Speech*, vol. 13, no. 2, pp. 65–88, Apr. 1970, doi: 10.1177/002383097001300201.

[5] A. K. Nábělek, A. Ovchinnikov, Z. Czyzewski, and H. J. Crowley, “Cues for perception of synthetic and natural diphthongs in either noise or reverberation,” *J Acoust Soc Am*, vol. 99, no. 3, pp. 1742–1753, Mar. 1996, doi: 10.1121/1.415238.

[6] A. Bladon, “Diphthongs: A case study of dynamic auditory processing,” *Speech Commun*, vol. 4, pp. 145–154, 1985.

[7] M. Gottfried, J. D. Miller, and D. J. Meyer, “Three approaches to the classification of American English diphthongs,” *J Phon*, vol. 21, no. 3, pp. 205–229, Jul. 1993, doi: 10.1016/S0095-4470(19)31337-3.

[8] S. Lee, A. Potamianos, and S. Narayanan, “Developmental acoustic study of American English diphthongs,” *J Acoust Soc Am*, vol. 136, no. 4, pp. 1880–1894, Oct. 2014, doi: 10.1121/1.4894799.

[9] T. Gay, “Effect of speaking rate on diphthong formant movements,” *J Acoust Soc Am*, vol. 44, no. 6, pp. 1570–1573, Dec. 1968, doi: 10.1121/1.1911298.

[10] S. M. Tasko and K. Greilick, “Acoustic and articulatory features of diphthong production: A speech clarity study,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 84–99, Feb. 2010, doi: 10.1044/1092-4388(2009/08-0124).

[11] K. Tjaden and G. E. Wilding, “Rate and loudness manipulations in dysarthria: acoustic and perceptual findings,” *Journal of Speech, Language and Hearing Research*, vol. 47, pp. 766–783, 2004.

[12] R. D. Kent and K. L. Moll, “Tongue body articulation during vowel and diphthong gestures,” *Folia phoniat*, vol. 24, pp. 278–300, 1972.

[13] C. Dromey, G. O. Jang, and K. Hollis, “Assessing correlations between lingual movements and formants,” *Speech Commun*, vol. 55, no. 2, pp. 315–328, Feb. 2013, doi: 10.1016/j.specom.2012.09.001.

[14] J. Wouters and M. W. Macon, “Effects of prosodic factors on spectral dynamics. I. Analysis,” *J Acoust Soc Am*, vol. 111, no. 1, pp. 417–427, Jan. 2002, doi: 10.1121/1.1428262.

[15] A. Xu, P. Birkholz, and Y. Xu, “Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation,” in *Proceedings of the international congress of phonetic sciences (ICPhS)*, 2019. [Online]. Available: www.vocaltractlab.de

[16] D. R. van Niekerk, A. Xu, B. Gerazov, P. K. Krug, P. Birkholz, and Y. Xu, “Finding Intelligible Consonant-Vowel Sounds Using High-Quality Articulatory Synthesis,” in *Interspeech 2020*, Oct. 2020, pp. 4457–4461. doi: 10.21437/Interspeech.2020-2545.

[17] S. Prom-On, P. Birkholz, and Y. Xu, “Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach,” *EURASIP J Audio Speech Music Process*, vol. 2014, no. 1, p. 23, 2014, doi: 10.1186/1687-4722-2014-23.

[18] W. J. M. Peeters, “Dipthong dynamics: A cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German,” 1996.

[19] W. J. M. Peeters and W. J. Barry, “Diphthong dynamics: production and perception in Southern British Englsih,” in *EUROSPEECH*, 1989, pp. 1055–1058.

THE RELATION BETWEEN PITCH ACCENT TYPES, HEAD MOVEMENTS AND PERCEIVED PROSODIC PROMINENCE IN L2 FRENCH

Florence Bailly[1,2], Stefan Baumann[1], Patrick Rohrer[2]

[1]IfL Phonetik, University of Cologne
[2] GreP-G, Universitat Pompeu Fabra
{fbailly; stefan.baumann}@uni-koeln.de, patrick.rohrer@upf.edu

ABSTRACT

Recent research has shown the relevance of multimodal cues in the realization of prominence in discourse. Speakers may not only use prosodic cues - e.g. pitch accents - to stress important information but also visual cues - e.g., manual and non-manual gestures such as head movements - in synchrony.

As part of a larger project comparing the multimodal marking of information structure by L1 and L2 speakers, this study reports on the relationship between perceived prominence in L2 speech and the pitch accent and head movement types used by 25 Catalan learners of French during a narrative task. Results confirm the relationship between pitch accents, gestural cues, and prosodic prominence in L2 learners and show that higher prosodic prominence is associated with rising, falling, and high pitch accents as well as with protrusions and nods of the head. Falling contours associated with highly prominent words were less marked by head movements, indicating potential differences from L1 speech.

Keywords: prominence, prosody, head movements, pitch accents, L2 speech

1. INTRODUCTION

Prominence refers to the level of perceptual strength of a constituent in relation to other constituents in the same utterance, based on its structural and acoustic properties. In Autosegmental-Metrical phonology [1], prominence is associated with metrically strong positions in an utterance, with the rightmost position, called “nuclear”, being the strongest. “Prenuclear” prominences occur earlier in the phrase (often in initial position) and are thus regarded to be structurally weaker. Various prominence rating studies on American English and German confirmed this broad picture since they found a generally higher level of *perceived* prominence for nuclear accents in comparison with prenuclear accents and unaccented words. As to the *type* of accent in nuclear position, both languages show higher prominence values for rising accents, followed by high, falling and low pitch accents [2,3].

Much like West Germanic languages, Catalan and Spanish are stress-accented languages: the syllables bearing primary stress usually serve as the landing site for pitch accents, which depend on the larger prosodic structure [4]. Actual prosodic prominence surfaces on primarily stressed syllables at the word level and accents at the phrase level, with F0 movements being an essential cue of the latter [5]. Deaccentuation is rare. In French, by contrast, there is no word stress. Accent placement is believed to be constrained by prosodic phrasing and to signal phrase structure, with a compulsory high or low phrase-final accent and an optional initial rise [6].

There is evidence on the temporal association between prosodic prominence (pitch-accented syllables) and the prominence (strokes or apexes) of referential and non-referential hand gestures [7, 8, 9, 10, 11]. Regarding non-manual gestures, studies have found that head gestures tend to co-occur with pitch-accented syllables [12, 13, 14, 15]. Crucially, Alexanderson et al. [13] found that head nods co-occurred with words bearing a focal accent in Swedish spontaneous dialogic speech. In parallel, there is also direct evidence of the relationship between perceived auditory prominence and non-manual gestures. Swerts and Krahmer [16] investigated both head and eyebrow movements in a Dutch news reading corpus: Strongly prominent words tended to co-occur with both head and eyebrow movements (67%), while weak prominence was mostly produced without gesture (47%), or with only-head (16%) or only-eyebrow movement (3%).

Little is known about the use of multimodal cues and their relationship with perceived prominence in L2 speech. McCafferty [17] suggested that L2 learners with lower proficiency may use non-referential (beat) gestures to parse words into syllables and control their fluency. Importantly, it is likely that lower proficiency L2 speakers might be using gestures instead of other linguistic devices for a range of pragmatic functions including the marking of salient information [18], in a similar way as children may do [19]. Our goal is to explore for the first time the relationship between perceived prominence and prosodic and gestural cues in L2 speakers by answering the following research questions:

- Which types of pitch accent are perceived as more prominent in L2 speech? We expect high and rising pitch accents to be perceived as more prominent than low and falling pitch accents.
- Which types of head movement are accompanied by higher prosodic prominence? To date, only nods have been specifically studied in relation to prominence. Here we investigate other types of head movement in an exploratory part of the study.
- What is the relationship between pitch accent types and head movement types, and is this relationship reflected in their association with perceived prominence?

2. METHODS

2.1. Participants and elicitation task

Participants were 25 bilingual speakers of Catalan and Spanish (M_{age} = 19.6 years, *SD* = 0.91) learning French. They were all second-year undergraduate students in Translation and Interpretation Studies. French was their second foreign language after English. Their self-reported proficiency in French varied between A2 (20%), B1 (44%) and B2 (36%).

During the semester, they had to complete two assignments as a means of data collection for this study. They were asked to video-record themselves producing a short spontaneous monologue 1) about their Erasmus stay, which took place the previous year. In order to obtain comparable oral productions, the instructions entailed a series of questions, as follows: *"Please explain your Erasmus stay in a few sentences in French. Where did you go and when did you leave? Did everything go smoothly? What are your best memories? Would you like to live in a foreign country again?"* and 2) describing their best friend, with the following instructions: *"Explain in a few sentences in French who your best friend is. How is he/she physically? How did you meet? What do you like and dislike about him/her?"*

For the two assignments, participants had to follow a link sent by the experimenter which led them to an online survey platform where they could read the instructions for video recording and for the task. No indication was given regarding head or body movement. They were asked to use their laptop camera and the online recording software https://webcamera.io/ (quality 720p) and then share the file with the teacher and the experimenter.

2.2. Annotations

Perceived prominence and prosodic annotations. The auditory information of the 50 video recordings produced by the participants was extracted as sound files (WAV 48KHz 16-bit) using Adobe Premiere Pro. The sound files were annotated automatically in Praat [20] for word, syllable and phoneme with EasyAlign [21] and manually corrected by the first author. In a first step, prosodic and prominence annotations were carried out independently of each other. Prominence was annotated auditorily following the guidelines proposed by the DIMA annotation system [22], that is, weak, strong or extra strong prominence labels (levels 1 to 3) were attributed to words according to their perceived salience in the utterance. Then, phrasing and pitch accents were annotated. As L2 speech presents deviations from the prosodic patterns that are predicted by F_ToBI phonological annotations [23], it was decided to follow a more phonetic annotation of the pitch accents in the accentual phrases (AP) when deemed necessary, as advocated by Hualde & Prieto [24]. Each AP contained one obligatory pitch accent at the right edge of the phrase (L*, HL*, !H*, H*L, H*, L*H, or LH*) and optionally a high or rising initial accent (Hi). Inter-annotator reliability between the first two authors (one of them being a native French speaker) on 20% of the data was calculated for prominence ratings. A moderate score of .475 Cohen's Kappa was obtained.

Head movements. The video recordings were coded in ELAN [25] for non-referential head movements only, i.e., we excluded head movements conveying any semantic meaning such as shaking the head rapidly from left or right to express negation, or up and down repeated movements for approval. However, when a head movement had a semantic and a rhythmic component, it was decided that it would be annotated and included in the analysis, for example a protrusive movement of the chin to one side to indicate a specific location. We applied the M3D annotation scheme, which specifies five head movement types [26; 27, p. 25]: *Nod*, up and down movements; *Turn* right and left movements; *Tilt*, the top of the head goes in one direction (left or right) and the chin goes in the opposite direction; *Slide*, displacement of the whole head to the left or right; *Protrusion*: displacement of the whole head forward or backward.

The gesture strokes were detected and annotated by searching for the beginning and end of the most salient movement of the head in a frame-by-frame observation and without the audio, in order to avoid any influence from the speech stream. The type of head movement was chosen after observation of the movement at normal speed. Ten percent of the data was annotated by the first two authors and inter-annotator reliability was assessed for gesture stroke and type labelling with the built-in inter-annotator

reliability tool in ELAN, which uses an algorithm to assess both temporal overlap and the values assigned to the annotations. The algorithm returned kappa values of .65 for the identification of each type of phase, indicating substantial reliability.

Finally, the prominence and prosodic annotations in Praat were imported into ELAN which allows the creation of a time-aligned database for further processing in R [28]. Phrase-final pitch accents were grouped according to their shape: low (L*), falling (H*L, HL*), high (H*), and rising (L*H, LH*, HH*) - plus the phrase-initial high or rising accent (Hi).

2.3. Statistical analysis

Three Generalized Linear Models (GLM1, GLM2, and GLM3) were run in R with the *lme4* package [29]. A Poisson regression was used to model count data and to account for between-category differences (the unbalanced number of occurrences), the models were offset by the total number of items for each category. GLM1 assessed the association between perceived prominence and pitch accent type. The dependent variable was the number of syllables annotated with a pitch accent and the fixed factors were PROMINENCE (2 levels: weak, strong; 'extra strong' did not occur in the data) and PITCH ACCENT TYPE (5 levels: initial accent, low, falling, high, rising) as well as their interaction. GLM2 evaluated the association between types of movement and perceived prominence. The dependent variable was the number of syllables annotated which contained the apex of the head movement and the fixed factors were PROMINENCE (2 levels: weak, strong) and HEAD MOVEMENT TYPE (5 levels: nod, protrusion, slide, tilt, turn) as well as their interaction. GLM3 assessed the association between types of pitch accent and types of head movement, in GM3, the dependent variable was the number of syllables annotated with both a head movement and a pitch accent, and the fixed factors were PITCH ACCENT TYPE (5 levels: initial accent, low, falling, high, rising) and HEAD MOVEMENT TYPE (5 levels: nod, protrusion, slide, tilt, turn) as well as their interaction.

3. RESULTS

Perceived prominence and pitch accent types. Results of GLM1 show a significant effect of PITCH ACCENT TYPE ($\chi^2(4) = 1002.49$, $p < .001$ and an interaction PROMINENCE X PITCH ACCENT TYPE ($\chi^2(4) = 263.35$, $p < .001$). Post-hoc comparisons show that falling and rising pitch accents are related to significantly more prominent rating scores followed by high, low, and initial accents (see Fig.1).

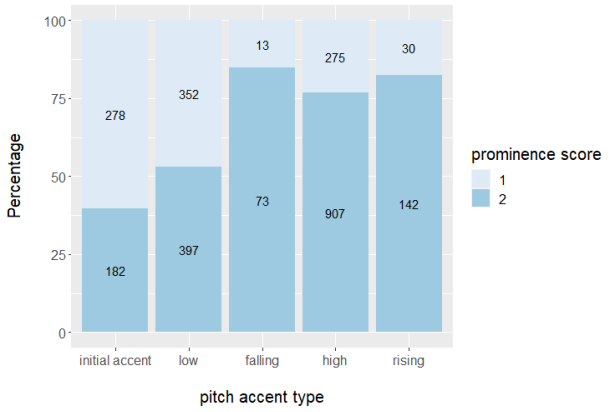


Figure 1: Prosodic prominence relating to the type of pitch accent (percentage along y-axis; N of occurrences indicated in graph).

Perceived prominence and head movement types. Results of GLM2 show a significant effect of HEAD MOVEMENT TYPE ($\chi^2(4) = 40.05$, $p < .001$) and an interaction PROMINENCE X HEAD MOVEMENT TYPE ($\chi^2(8) = 35.83$, $p < .001$). Post-hoc results indicate that protrusions, nods, and slides are associated with high (level 2) prosodic prominence (see Fig. 2). Across types of movements, protrusions are significantly associated with higher prominence compared to tilts and slides. Other differences are not significant.

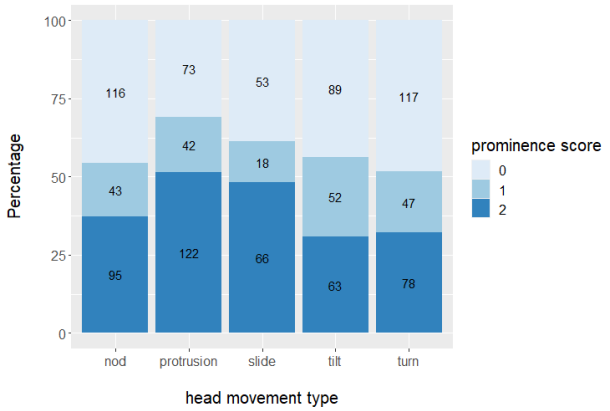


Figure 2: Prosodic prominence levels co-occurring with gesture types (percentage along y-axis; N of occurrences indicated in graph).

Types of pitch accent and types of head movement. Results of GLM3 show a significant effect of PITCH ACCENT TYPE ($\chi^2(4) = 649.09$, $p < .001$), HEAD MOVEMENT TYPE ($\chi^2(4) = 45.31$, $p < .001$) and an interaction PITCH ACCENT TYPE X HEAD MOVEMENT TYPE ($\chi^2(16) = 30.12$, $p < .05$). Post-hoc results reveal that all types of head movement co-occur the most with high pitch accents followed by low pitch accents. Across categories protrusions are more likely to co-occur with a high accent compared to tilts and slides, but no other difference is significant.

In terms of proportions (see Table 1), protrusions and slides are the most frequent head movements co-

occurring with high pitch accents (around 59%). Low accents are more frequently accompanied by tilts and turns (around 30-32%), which are themselves co-occurring with lower levels of prosodic prominence. Altogether, rising accents, even if less frequent, are the ones that are more often co-occurring with a head movement (59%, most of which were nods) while falling accents, despite being also perceived as more prominent, tend to be less frequently accompanied by a head movement (around 26%). Phrase-initial accents were least often co-occurring with a head movement (15%).

4. DISCUSSION AND CONCLUSION

In this study, we asked three questions related to the expression of prominence through prosodic and gestural cues in Catalan L2 learners of French. The first analysis shows that falling, rising and high phrase-final accents in L2 French were perceived as more prominent than phrase-final low and phrase-initial accents. Falling and rising accents are significantly more prominent than high accents, which might indicate that pitch *movement* was decisive for the annotators' prominence judgments.

Regarding gestural cues, the second analysis showed that protrusions, nods, and slides coincide the most with higher prosodic prominence. Around 70% of all protrusions are accompanied by prosodic prominence level 1 (weak) or 2 (strong), with about 50% of level 2 prominences. Future research should investigate whether the prominence of the head movement depends more on the amplitude and duration of the movement, as protrusions and nods may allow for a more expanded movement.

The third analysis shows that when a head movement co-occurs with a pitch accent, it is predominantly with high accents, regardless of the head movement type. It seems that particular contours may attract more or less head movements. While falling and rising contours were perceived both as more prominent, the former attracted only half the number of head movements compared to the latter. On the one hand, this contradicts the cumulative cue hypothesis [30] claiming that the relation between prosodic and gestural cues to prominence should be additive in nature, on the other it suggests that L2 learners may not yet be completely able to produce multimodal prominence and have difficulties at least with certain types of contours. The lack of association between initial accents and head movement may indicate that phrase-final accents generally attract more head movements as prominence markers in our speakers. Future studies will include L1 controls to assess these issues.

This study provides evidence that the head movements which are associated with higher prosodic prominence (protrusions, nods and slides) more frequently accompany pitch accents reflecting higher prosodic prominence, particularly rising and high pitch accents. These results are in line with [31], who found that protrusions and nods tend to co-occur with new(er) information in discourse [32], which is usually signalled by prosodic prominence. Crucially, a vast majority of the total number of head movements co-occurs with pitch accents, showing the close association between prosody and gesture (Table 1). To continue assessing the role of multimodal cues in prominence, our results need to be contrasted with a perceptual evaluation of the prominence of head movements alone and of both visual and acoustic cues together, including their relations with information structure marking.

By exploring the relationship between multimodal cues and prominence, we can gain further insight on the developmental path of speech production in an L2. For instance, French speakers may use gestural cues more to indicate discourse structure [33]. This study constitutes the first step of a larger study comparing Catalan learners of French and native French speakers to assess these differences.

	protrusions	nods	turns	Tilts	slides	total	corpus
high	126 (59,4%)	98 (48%)	94 (48,7%)	67 (47,9%)	68 (59,6%)	453 (38,3%)	1182
low	40 (18,9%)	50 (24,5%)	59 (30,6%)	45 (32,1%)	23 (20,2%)	217 (29 %)	749
rising	23 (10,8%)	32 (15,7%)	21 (10,9%)	13 (9,3%)	13 (11,4%)	102 (59,3%)	172
init. acc.	16 (7,5%)	19 (9,3%)	18 (9,3%)	12 (8,6%)	4 (3,5%)	69 (15%)	460
falling	7 (3,3%)	5 (2,4%)	1 (0,5%)	3 (2,1%)	6 (5,3%)	22 (25,6%)	86
total	212 (86,1%)	204 (80,3%)	193 (79,7%)	140 (68,6%)	114 (83,2%)		
corpus	237	254	242	204	137		

Table 1: Number of occurrences and proportions (in %) of pitch accent types and head movement occurring jointly and in relation to the total number of pitch accent and head movement types in the corpus. Note: init. acc = initial accent.

5. REFERENCES

[1] Ladd, R. 2008. *Intonational phonology*. Cambridge University Press.

[2] Baumann, S., Röhr, C. 2015. The perceptual prominence of pitch accent types in German. *Proc. 18th ICPhS Glasgow*, paper 298.

[3] Cole, J., Hualde, J. I., Smith, C., Eager, C., Mahrt, T., de Souza, R. N. 2019. Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *J. Phonetics 75*, 113–147.

[4] Prieto, P. 2014. The intonation phonology of Catalan. In S.-A. Jun (ed). *Prosodic typology II.* Oxford University Press.

[5] Ortega-Llebaria, M., Prieto P. 2007. Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. In P. Prieto, J. Mascaró, & M.-J. Solé (eds). *Segmental and prosodic issues in Romance phonology.* John Benjamins, 155–176.

[6] Jun, S.-A., Fougeron, C. 2000. A phonological model of French intonation. In A. Botinis (ed), *Intonation: Analysis, modeling and technology.* Springer, 209–242.

[7] Esteve-Gibert, N., Prieto, P. 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *J. Speech Lang. Hearing Res. 56*, 850–864.

[8] Leonard, T., Cummins, F. 2011. The temporal relation between beat gestures and speech. *Lang Cog. Proc. 26*, 1457–1471.

[9] Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S. 2004. The timing of speech-accompanying gestures with respect to prosody. *J. Acoust. Soc. Am. 115*, 2397.

[10] Loehr, D. P. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab. Phon. 3*, 71–89.

[11] Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Front. Psychology 9*, Article 1514.

[12] Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swets, M., Prieto, P. 2017. The timing of head movements: The role of prosodic heads and edges. *J. Acoust. Soc. Am. 141*, 4727–4739.

[13] Alexanderson, S., House, D., Beskow, J. 2013. Aspects of cooccurring syllables and head nods in spontaneous dialogue. *Proc. 12th AVSP conference Annecy,* 169–172.

[14] Ambrazaitis, G., House, D. 2017. Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Comm. 95*, 100–113.

[15] Liu, Y., Shamei, A., Chow, U.Y., Soo, R., Pineda Mora, G., de Boer, G., Gick, B. 2020. F0-related head movement in blind versus sighted speakers. *J. Acoust. Soc. Am. 148*, 190–194.

[16] Swerts, M., Krahmer, E.J. 2010. Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *J. Phon. 38*, 197–206.

[17] McCafferty, S. G. 2006. Gesture and the materialization of second language prosody. *IRAL 44*, 197–209.

[18] Graziano, M., Nicoladis, E., Merentette, P. 2020. How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Lang. Learn. 70*, 266–304.

[19] Esteve-Gibert, N., Loevenbruck, H., Dohen, M., and D'Imperio, M. 2021. Preschoolers use head gestures rather than prosodic cues to highlight important information in speech. *Dev. Sci. 25*: e13154.

[20] Boersma, P., & Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 15 March 2022 from https://www.praat.org

[21] Goldman, J.-P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. *Proc. InterSpeech 2011 Florence*. ISCA Archive., 3233–3236.

[22] Kügler, F., Baumann, S., Röhr, C.T. 2022. Deutsche Intonation, Modellierung und Annotation (DIMA) – Richtlinien zur prosodischen Annotation des Deutschen. In C. Schwarze & S. Grawunder (eds), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion*. Narr. 23–54.

[23] Delais-Roussarie et al. 2015. Intonational phonology of French: Developing a ToBI system for French. In S. Frota & P. Prieto (eds), *Intonation in Romance.* Oxford University Press. 63–100.

[24] Hualde, J., Prieto, P. 2016. Towards an International Prosodic Alphabet, *Lab. Phon. 7*, 5.

[25] ELAN Version 6.4 [Computer software]. 2022. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.

[26] Wagner, P. 2014. Gesture and speech in interaction: An overview. *Speech Comm. 57*, 209–232.

[27] Rohrer, P., Vilà-Giménez, I., Florit-Pons, J., Esteve-Gibert, N., Ren, A., Shattuck-Hufnagel, S., Prieto, P. 2020. *The MultiModal MultiDimensional (M3D) labeling system.* https://doi.org/10.17605/OSF.IO/ANKDX

[28] R Core team. 2022. R Version 4.2.1. https://www.R-project.org/.

[29] Bates. D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat, Soft. 67*, 1–48.

[30] Ambrazaitis, G., House, D. 2022. Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters. *Lab. Phon. 13*, 1–35.

[31] Baills, F. & Baumann, S., Submitted. Multimodal marking of information structure in the L2: the case of Catalan learners of French.

[32] Riester, A. & Baumann, S. 2017. The RefLex Scheme – Annotation Guidelines. SinSpeC. Working Papers of the SFB 732, vol. 14. University of Stuttgart.

[33] Rohrer, P. 2022. *A temporal and pragmatic analysis of gesture-speech association* [PhD thesis, UPF]. TDX.

DEVELOPING A CORPUS OF AUDIO-VISUAL ATTITUDINAL EXPRESSIONS IN HINDI

Hansjörg Mixdorff[1], Navneet Nayan[2], Albert Rilliard[3], Preeti Rao[4], Debashis Ghosh[2]

[1]Berliner Hochschule für Technik [2]IIT Roorkee, India [3]Université Paris-Saclay, France [4]IIT Bombay, India
hmixdorff@bht-berlin.de, nnayan@ec.iitr.ac.in, Albert.Rilliard@limsi.fr, prao@ee.iitb.ac.in,

ABSTRACT

This paper presents a corpus of attitudinal expressions in Hindi and their evaluation by native Hindi speaking raters. The paradigm is adapted from Rilliard et al. (2013), and forms part of an intercultural endeavour aimed at studying prosodic and facial expressions of social affect across languages. Our corpus includes a total of 16 attitudes, such as arrogance, surprise, politeness etc. portrayed by 19 speakers (10f, 9m) of which three experts selected the best four males and females and the better of two turns for subsequent analysis. A follow-up experiment with more participants employed a total of 512 stimuli, 256 hereof full audio-visual stimuli, 128 audio-only and 128 silent video stimuli. Results indicate higher ratings for emotionally loaded attitudes such as irritation and doubt as compared to, for instance, irony or seductiveness. Reduced modality stimuli were rated more poorly.
Keywords: Attitudes, Hindi, audio-visual perception

1. INTRODUCTION

Human communication always has a social goal. Information about e.g. the mental state, emotions, mood or attitudes of the speaker and listener is conveyed during the dialog. The affective state is influenced, for instance, by the situation or role of the dialog partners. Mutual understanding of the social intention between communication partners should not be difficult as long as they grow up in the same or at least a similar culture. Interaction between partners from different cultures sometimes leads to wrong interpretations of the social expression. It has been shown that the verbal and non-verbal expressions depend on the culture in which we grow up. A study by Shochi et al. investigated twelve social attitudes e.g. surprise, irritation, command-authority for prosodic effects in British English, French and Japanese [1]. Intercultural comparison of linguistic and paralinguistic effects has enjoyed growing attention as the knowledge about how verbal and non-verbal social affects are expressed in different languages is paramount for mutual understanding between different cultures.

The primary goal of the work reported in the current paper was the recording of a Hindi speech corpus of social attitudes for an intercultural comparison. It follows the experimental design developed by Rilliard et al. [2], discussed in Section 2. In earlier works we collected corpora of attitudinal expressions in German [3] and Cantonese [4] following the same experimental design. These corpora were judged by Hindi speaking subjects [5]. Now the corresponding Hindi attitudinal data were gathered.

This paper presents the process of data collection and a perception study exploring the credibility of the 16 social attitudes (see Table 1 for complete listing), e.g. doubt, surprise and politeness of 19, later 8 Hindi speaking subjects. Section 2 presents the experimental design and the social and linguistic criteria of situations employed to elicit the attitudes from our subjects. In the same section the technical setting for the recordings, as well as the process of the recordings with dialog examples are discussed. The perceptual evaluation of the social attitudes by native Hindi raters is described in section 3. Section 4 presents the results of the evaluated expressions which includes analyses with regards to speaker and attitude. Section 5 concludes this paper with discussion and conclusions.

2. EXPERIMENT SETUP

2.1 Situational attitudes

16 Attitudes are performed by 19 speakers (10f, 9m) which are elicited through a short dialog between each speaker and the experimenter. The dialogs led to two target utterances: “mairee naach rahee thee” (मैरी नाच रही थी) (engl. Mary was dancing) and “ek kela” (एक केला) (engl. a banana). In the remainder of this article we will refer to the English translations to identify the respective phrase. For each expression of attitude, a test dialog was executed in order to prepare the speakers. This dialog was designed according to different social situations differing in the following social and linguistic aspects (A => speaker, B => experimenter):

- Type of speech act: propositional / social attitude

- Hierarchical distance between speaker A and speaker B: A > = < B[1]
- Social distance between speaker A and speaker B: 1-friend, 2-known, 3-unknown[1]
- Valence of speech act: positive / negative
- A dominates B: yes / no

The social situations conceived are adapted from Rilliard et al. [2]. As examples detailed descriptions of three of the social situations are presented in the following:

Admiration (ADMI): A and B are almost the same age and know each other well. Both love Samosas, and talk about the delicious ones they had the day before at Prakash's restaurant. The scene is at a coffee shop.

Irritation (IRRI): A and B are almost the same age and know each other. A is sitting next to B. Suddenly B starts smoking and A is very angry; he wants him/her to stop, expressing his irritation toward speaker B. The scene is in a public place.

Surprise (SURP): A and B are friends and the same age. A did not know that B can sing well. One day, B has A listen to his beautiful voice. The scene is at a friend's home.

2.2 Technical setting and recording process

We used a Nikon D850 HD camera set up 1m away from the speaker. Additionally, the acoustic signal was recorded through a dynamic AKG D5 microphone placed under the chin of speaker. A laptop positioned in front of the speaker displays the description of each social situation, the related test dialog and target dialog permitting the speakers to prepare for the ensuing task. The experimenter performed each situation with the subject in order to immerse him/her in the context for the attitude. After a short break the target dialog containing the target phrase followed immediately. The complete sequence of dialogs was recorded twice. An example test and target dialog for the attitude "irritation" (IRRI) could run as follows:

A: Sorry, but don't smoke please.
B: Okay, okay…
A: Don't smoke, please!

B: What did Mary do last night?
A: **Mary was dancing**. I already told you three times (are you deaf?)

In the test dialog A is angry because B starts to smoke and A wants to stop him. The sentence in bold is the target utterances which A should speak with an irritated expression. The sentences in brackets are only for familiarizing the speaker with the task, and are not uttered The main dialog ends with the target utterance "Mary was dancing".

The target phrases were cut out of the session video. We recorded a total of 19 speakers x 16 attitudes x 2 phrases x 2 trials=1216 target phrases performed by 19 speakers. Participants were students of IIT Roorkee and compensated for their time.

3 EVALUATION

Before we can subject the data to further analysis or employ them in perceptual tests we need to ensure the quality of presentation. To this end we first had the complete corpus of 1216 files judged by three experts who ranked each video clip on a scale from 9 (convincing) to 1 (implausible), given the intended attitude. The experts were familiar with the elicitation paradigm and supposed to scrutinize the data accordingly. In order to facilitate the tiring task we created four batches of 240 stimuli and one of 256.

In all experiments the stimuli were presented via a survey on the PsyToolkit server[4][5]. After an introduction to the experiment design, listing all 16 attitudes in English as well as Hindi, the stimuli were presented in a randomized order, first displaying for 900 ms a screen with the intended attitude in English and Hindi and a fixation cross in the location where the video was going to appear, then the video for its respective duration was shown. Subsequently the viewers were able to input their judgment via a graphical interface by clicking on a line between "very well" corresponding to a score of 9 and "not at all" corresponding to 1. A progress bar was shown for orientation.

The experts viewed the full audio-visual (A/V) stimuli and helped us identify the four best male and female speakers, as well as the better turn of the two we had collected.

Hence we yielded 8 x 32=256 utterances for the ensuing main experiment with a larger number of participants. In addition to the 256 full A/V stimuli, from these we created 128 silent video (V) and 128 audio-only (A) stimuli, selecting the better one of "Mary was dancing" or "a banana". Hence the experiment employed a total of 512 stimuli, the quality of which was judged by 32 Hindi speaking raters (23m, 9f). Due to the large number of

[1] Notions of distance, cf. Spencer-Oatey(1996)

utterances to be evaluated, we created a PsyToolkit survey with four sets of 128 stimuli each, while ensuring that the reduced A and V stimuli were not in the same set as the corresponding A/V stimuli. This means that each set was rated by eight judges.

After hearing/watching the stimulus only once raters had to judge the plausibility of each attitude during a time window of ten seconds. We collected 4096 scores, 32 of which (0.8%) were invalid due to time-outs. We hence included in total 2034 valid judgments for A/V stimuli 1016 for A, and 1014 for V type stimuli. The experiment took on average 14 minutes to complete.

4 RESULTS

The first analysis presented in this section concerns the judgments on a scale from one to nine assigned by the three expert judges. Statistical tests were performed using the SPSS statistical toolkit [1]. In the second part we look at the judgments of the larger participant group on the reduced set.

4.1. Expert Ratings

The raters generally judged the female speakers (score average/s.d. 7.32/1.95) better than the male speakers (7.06/2.10), or in other words, of the eight best rated speakers six were females. Interestingly, the distance "between the genders" is greater for certain attitudes such as ADMI or WOEG than others like DOUB or IRRI.

In Table 1, columns three and four list means and s.d. for all 16 attitudes. The judgments of the raters indicated a slightly better performance of the speakers at the second trial (average of 7.23) than the first one (7.16). The reason could be that the speakers knew the task after the first trial and on that account, it was easier to prepare the expressions. For the ensuing larger study, we selected for each attitude/target phrase the repetition that was judged better. We did not find any significant difference between the target phrase "a banana" and "Mary was dancing".

As witnessed in earlier studies, complex attitudes like IRON and SEDU were judged more poorly than coping strategies with strong emotions such as IRRI, SURP, CONT and DOUB.

In contrast, POLI which is actually a socially conditioned register also ranges among the highest scorers. In our German data this was the opposite. Obviously politeness exhibits special features in Indian society which are not as strong in the West or just represents a "normal way of speaking".

Interestingly, also the neutral question (QUES) received rather low ratings. We assume that the facial expression together with the interrogative mode often counteracted the intended "neutrality" and rather conveyed doubt or suspicion.

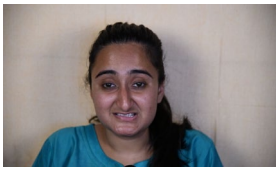
There is a strong negative correlation between the mean score for each stimulus (Pearson's R of -.765, $p < .001$) and its standard deviation, indicating that the raters agreed more on highly rated stimuli. There is also a weak correlation between the reaction time and the score (Pearson's R of -.147, $p < .001$).

4.2. Ratings on subcorpus of best speakers

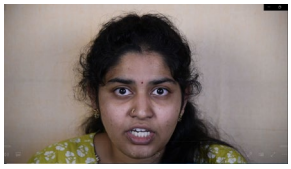
As explained above we selected a subcorpus of the best four male and best four female speakers and had them judged by a larger group of participants. Figure 1 presents snapshots of examples of different attitudes of these speakers and their mean scores and s.d. In general, results among the "non-experts" follow similar tendencies as those of the experts. If we correlate their mean stimulus-wise judgments with those of the experts on the 256 full AV stimuli that both experiments had in common, we yield a Pearson's R of .385 ($p < .001$).

Attitude	abbrev-iation	mean	standard deviation
admiration	ADMI	7.35	2.18
arrogance	ARRO	7.42	1.81
authority	AUTH	7.38	1.62
contempt	CONT	7.25	2.00
neutral statement	DECL	7.30	1.78
doubt	DOUB	7.62	2.05
irony	IRON	6.60	2.03
irritation	IRRI	7.63	2.07
obviousness	OBVI	7.10	1.96
politeness	POLI	7.46	1.57
neutral question	QUES	6.50	2.37
seductiveness	SEDU	6.78	2.29
sincerity	SINC	6.95	2.02
surprise	SURP	7.42	2.11
uncertainty	UNCE	7.30	1.93
walking-on-eggs	WOEG	7.06	2.11

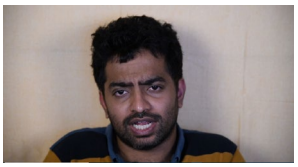
Table 1: List of attitudes, abbreviations used in the paper and mean score and s.d. found in the expert study for 1216 stimuli.



S05 IRRI, 7.33/1.88



S03 SURP, 7.23/1.91



S06 DOUB, 7.17/1.97



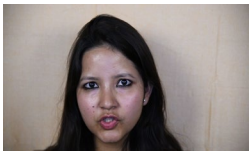
S09 ADMI, 6.57/2.25



S08 SINC, 7.01/2.03



S15 SEDU, 7.19/1.91



S16 POLI, 7.15/1.92



S17 UNCE, 7.12/2.05

Figure 1: Examples from the four best male and female speakers performing various attitudes, mean and s.d. of scores.

Attitude	AV mean/s.d.	V mean/s.d.	A mean/s.d.
ADMI	7.39/1.75	7.22/1.85	6.23/2.27
ARRO	7.63/1.55	7.41/1.90	6.75/2.33
AUTH	7.42/1.86	6.65/2.27	6.75/2.49
CONT	7.43/1.86	6.64/2.30	5.48/2.50
DECL	7.30/1.75	6.86/1.95	7.52/1.57
DOUB	7.84/1.55	7.37/1.69	7.61/1.45
IRON	7.01/1.80	6.26/2.19	6.59/1.96
IRRI	7.79/1.59	7.34/1.95	6.41/2.61
OBVI	7.00/2.13	6.62/2.30	7.28/1.70
POLI	7.29/1.74	7.06/1.72	7.33/1.72
QUES	6.71/2.32	5.85/2.15	6.73/1.72
SEDU	6.43/2.34	5.78/2.52	6.03/2.50
SINC	7.33/1.70	7.17/1.85	7.33/1.72
SURP	8.09/1.28	7.80/1.54	8.06/1.21
UNCE	7.64/1.58	7.16/1.91	7.28/1.70
WOEG	6.76/2.33	6.42/2.25	6.63/2.17

Table 2: List of attitudes. mean score and standard deviation as a function of the attitude and modality.

Noticeably, the non-experts extend their range of scores much more towards the lower end. Hence their mean score across the whole data set is almost one point below that of the experts (7.32 vs. 8.12, for AV stimuli) Since we look at many more judges their s.d.s are much higher at 1.89 vs. 0.68 of the experts.

The most important difference between the two experiments besides the smaller number of stimuli presented is the presence of full A/V as well as V and A stimuli. On average full A/V stimuli were judged better (mean/s.d. of 7.32/1.89) than V (6.86/2.10) and A (6.88/2.10). This is confirmed by independent samples Kruskal-Wallis test between AV and A or V, respectively. Table 2 lists mean scores and standard deviations for each attitude depending on the modality. Interestingly, for the so-called "neutral" attitudes DECL and QUES the audio-only mode yields better ratings than the full AV, indicating the contribution of the acoustic channel, but also DOUB and SURP seem to be conveyed well acoustically. On the video-only data, IRRI and ARRO, two strongly negative attitudes, surpass audio-only by far and show the importance of the visual channel.

5 DISCUSSION AND CONCLUSIONS

This study presented a recently developed Hindi corpus of attitudinal expressions and first results of its analysis. Starting from an initial set of 19 speakers and 1216 video clips, we first selected the best four male and female speakers for subsequent analysis. We found that the females were the better performers, especially at expressing ADMI.

On the remaining data set we performed a rating test of full AV as well as reduced A and V versions. Our results are in line with those of earlier studies on different languages [3][4]. Reduced stimuli are generally judged more poorly than the full ones, with a few exceptions. This result shows that both the acoustic signal and the visual cues are vital in the portrayal of attitudes.

Future work will entail perception tests with viewers from different cultures. Work is in progress on a free labeling task with German participants. We also plan analysis of acoustic data as well as visual cues. Although we already observed many kinds of head movements, the famous Indian "head wobble" or "wag" occurs very seldom. Of the 256 stimuli only three exhibit this head movement. They all pertain to the UNCE condition.

6 Acknowledgements

This work was supported by DFG grant Mi625-32 to Mixdorff. Thanks to Priya G L for video editing.

7 REFERENCES

[1] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D. 2009. Intercultural perception of English, French and Japanese social affective prosody, in S. Hancil (ed.), The role of prosody in affective speech, Linguistic Insights 97, Bern: Peter Lang, AG, Bern, 31-59.

[2] Rilliard, A., Erickson, D., Shochi, T., de Moraes, J.A. 2013. Social face to face communication - American English attitudinal prosody, INTERSPEECH 2013 1648-1652

[3] Hönemann, A., Mixdorff, H., Rilliard, A. 2014. Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German. Forum Acusticum 2014, Krakow, Poland.

[4] Lee, T., Ma, M., Rilliard, A., Mixdorff, H., Hönemann, A. 2018. Free Labeling of Audio-visual Attitudinal Expressions in Cantonese. Proceedings of SpeechProsody 2018, Poznan, Poland.

[5] Mixdorff, H., Ghosh, D., Rilliard, A., Hönemann, A. 2020. Perception of Audio-visual Expressions in German and Cantonese by Native Speakers of Hindi. Speech Prosody 2020, Tokyo, Japan.

[6] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

[7] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

[8] IBM SPSS , URL https://www.ibm.com/products/spss-statistics, 2022.

THE DEVELOPMENT OF AUDIOVISUAL SPEECH PERCEPTION IN MANDARIN-SPEAKING CHILDREN

Yi WENG and Gang PENG

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies,
The Hong Kong Polytechnic University, Hong Kong SAR, China
211191113r@connect.polyu.hk gang.peng@polyu.edu.hk

ABSTRACT

Existing findings suggest that the developmental shift of audiovisual speech perception is subject to cultural variation, and previous studies failed to observe it in Mandarin-speaking children. The current study revisits this issue by performing the McGurk paradigm on a large sample consisting of 61 children aged 3–4 (n = 20), 5–6 (n = 21), 7–8 (n = 20) years and 26 adults aged 18–22 years. Results revealed a dramatic shift occurring at around five as 3–4-year-olds made significantly more "Ba" (auditory) and fewer "Da" (audiovisual) responses in incongruent trials (AbVg) than older groups. Spearman's correlation showed that age negatively correlated with "Ba" but positively with "Da" responses. Results supported that Mandarin-speaking children undergo a developmental shift from biasing unimodal auditory information to integrating audiovisual information in speech perception at an earlier age rather than not necessarily undergoing the shift.

Keywords: Audiovisual speech perception; Mandarin-speaking children; developmental shift; McGurk effect

1. INTRODUCTION

Our coherent perception of the outside world is derived from the cooperation and interaction of different sensory modalities, instead of a collection of senses [1]. Both auditory and visual cues contribute to speech perception, laying the solid foundation for communication and social interaction [2]. The role of visual modality in audiovisual speech perception could be illustrated from two perspectives. First, visual information facilitates speech perception. One stand-out example is lip-reading, a critical technique for people with hearing impairment to "read" from visible speech [3-4]. Individuals with normal hearing also enjoy the benefits of audiovisual integration including facilitating speech perception in noise [5], assisting speech comprehension [6] or even improving early cortical processing of speech [7]. Second, conflicting visual information may modulate auditory percepts, giving rise to perceptual illusions such as the McGurk effect [8]. In a classic McGurk design, adult participants tended to perceive an illusory /da/ after watching a video where the soundtrack of /ba/ is dubbed on the video of /ga/. The emergence of the McGurk illusion, which lacks auditory or visual substance, is the result of taking both audition and vision into account when deriving perception. Thus, it offers a window to examine the development of audiovisual speech perception.

Since the original report on the McGurk effect, the discrepancy between children and adults in terms of fused percepts (/da/ responses) has been noticed [8]. Recent research suggests that children are less likely to enjoy the same benefits of integrating bimodal information as adults given that multisensory processing is a "late bloomer" which takes a long journey to grow into maturity [9-10]. Specifically, several cross-sectional studies have revealed a developmental shift concerning sensory weighing in audiovisual speech perception, as children shift away from high reliance on unimodal auditory information to multisensory processing with increasing age [11-12]. This process takes around ten years of age to be fully developed [11-12]. Nevertheless, these findings do not seem to be extended to individuals from certain cultural backgrounds. For instance, Sekiyama and Burnham only observed the developmental shift among English speakers but failed to obtain evidence from Japanese-speaking children [13]. They attributed their findings to cultural factors since it is not polite to gaze at the speaker's face in the Japanese context. Similarly, several studies on Mandarin speakers measured a comparable strength of McGurk effect between child and adult participants, proposing that it is also due to cultural variation that Mandarin speakers do not necessarily experience the development shift resembling English speakers [14-15]. However, it appears premature to draw this conclusion since preschool children have been excluded from existing developmental studies concerning Mandarin speakers.

Given the lack of evidence from younger children and a refined delineation of age groups (only one or two groups of children were included) in previous studies [14-16], the development of audiovisual speech integration in Mandarin-speaking children is still left blurred. To better address this issue, the

current study seeks to track down the developmental trajectory of Mandarin-speaking children with a larger sample size and a smaller group interval, aiming to revisit whether Mandarin-speaking children would undergo a developmental shift of audiovisual speech perception during early and middle childhood.

2. METHODS

2.1. Participants

Sixty-one children aged from 3–8 years and 26 young adults aged from 18–22 years were recruited in the current study, all of whom were native Mandarin monolingual speakers (see Table 1). Child participants were categorized into three groups according to their chronological age, namely 3–4-year-olds, 5–6-year-olds and 7–8-year-olds. All child participants were recruited from general education institutes whose caregivers reported no intellectual, behavioural or hearing problems. Their verbal ability was further assessed by the Verbal Comprehension Index (VCI) of Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV), showing no abnormalities. A control group consisting of 26 Mandarin-speaking young adults was recruited from a university, and they were also free of hearing problems. All participants or their caregivers had signed written consent and got compensated for their participation. The methodology employed in the current study has been reviewed and approved by the University Institutional Review Board.

Group	N (Female)	Chronological Ages (Range, in year)	
		Mean	SD
3–4	20 (10)	4.35 (3.78–4.92)	0.31
5–6	21 (10)	5.88 (5.01–6.69)	0.50
7–8	20 (10)	7.82 (7.05–8.98)	0.64
Adult	26 (13)	20.85 (18.47–22.87)	1.33

Table 1: Gender and age information of participants among groups.

2.2. Stimuli

A young female speaker native in Mandarin, from whom written consent was obtained, was invited to record the articulation process of the three CV syllables: “Ba” [pa], “Da” [ta] and “Ga” [ka] with a high-level tone (around 240 Hz) in a quiet room. The speaker’s face was presented against the background in a solid colour. All videos were taken with a resolution of 1920×1080 pixels and a frame rate of 30 frames/s. Each video lasted for two seconds, which began with the speaker’s still face, followed by an articulatory process and ended with a still face. For congruent trials, the original videos were utilized. For the incongruent trial, the soundtrack of “Ba” was dubbed on the muted “Ga” video (AbVg) using Adobe Premiere Pro CC 2018.

2.3. Procedure

The verbal ability was assessed one by one a day before the McGurk task. During the experiment, participants were seated in front of the screen of a 16-inch laptop with a resolution of 1920×1080 pixels at a distance of around 50cm. Soundtracks were presented binaurally through headphones (Audio-Technica ATH-M20x) at around 70 dB.

Child participants were familiarized with the experiment through two training sessions. The first one was set to ensure children were able to differentiate the three CV syllables involved in the current design. The experimenter would present three slides with different background colours for each syllable respectively. On each slide, there was a pattern containing the Pinyin form of the corresponding syllable and a picture semantically relevant to the syllable in Mandarin. During this process, the experimenter would play out the syllable and instruct the children to speak out. At the end of this training session, children were required to pass a small test by pointing to the correct pattern of the heard syllable. Only if a child had made all choices correctly were they eligible for the next training session which was shared by both child and adult participants. The second training session was conducted in E-prime 3.0, which had the same setting as the formal experiment (see below). The three congruent trials (“Ba”, “Da”, “Ga”) were repeated twice in random order. Participants were instructed to point at the corresponding pattern of their responses among “Ba”, “Da” and “Ga”, and the experimenter would mark down by pressing the “b”, “d”, or “g” keys on the keyboard. Feedback would be provided during this training session. All the participants achieved full accuracy.

In the formal experiment, three congruent trials (“Ba”, “Da”, “Ga”) and one incongruent trial (“AbVg”) were presented in random order with seven repetitions. In each trial, participants were instructed to watch the laptop screen which sequentially presented a fixation (1000ms), a black screen (800ms), a stimulus (2000ms) and a response screen (infinite). They were required to make an oral

response to what the speaker had said accompanying pointing at the corresponding pattern among the three choices. The experimenter recorded their reaction by pressing the initial letter of the syllable on keyboard. Throughout the entire process, all participants were accompanied solely by experimenters.

3. RESULTS

Since the data were not normally distributed, nonparametric methods were adopted for analyses. For congruent trials, Kruskal-Wallis rank sum tests were performed to examine the effect of Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults) and Stimulus Type ("Ba", "Da", "Ga") on the accuracy of each congruent trial, which revealed a significant effect of Age Group ($H(3) = 14.301$, $p = 0.003$) instead of Stimulus Type ($H(2) = 3.91$, $p = 0.141$). Post hoc pairwise comparison using Wilcoxon rank sum test with Bonferroni correction showed that 3–4-year-olds ($M = 0.955$, $SE = 0.002$) achieved a significantly lower accuracy relative to 7–8-year-olds ($M = 0.983$, $SE = 0.001$, $p = 0.032$) and adults ($M = 0.991$, $SE = 0.000$, $p = 0.001$), while there were no differences between any other two groups. There were no significant interactions. Collectively, results indicated that the competence of identifying audiovisually presented CV syllables was undergoing development in 3–4-year-olds.

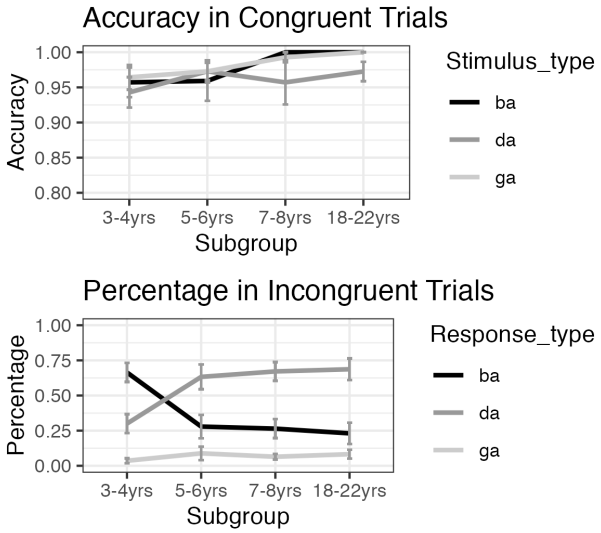


Figure 1: The accuracy in congruent trials achieved by four groups (upper) and the percentage of different response types in incongruent trials (bottom).

For incongruent trials, again, Kruskal-Wallis tests were conducted to examine the effect of Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults) on the percentage of each type of response respectively. For "Ba" responses indicating that participants made responses relying on auditory information, there exhibited a significant effect of Age Group ($H(3) = 20.689$, $p < 0.001$). Post hoc analysis showed that 3–4-year-olds ($M = 0.664$, $SE = 0.015$) made significantly more "Ba" responses than 5–6-year-olds ($M = 0.279$, $SE = 0.018$, $p = 0.007$), 7–8-year-olds ($M = 0.264$, $SE = 0.015$, $p = 0.003$) and adults ($M = 0.231$, $SE = 0.015$, $p < 0.001$). For "Da" responses indexing an audiovisual-integrated manner in perception, the significant effect of Age Group was also observed ($H(3) = 14.411$, $p = 0.002$). Post hoc pairwise comparison showed that significantly fewer "Da" responses were recorded from 3–4-year-olds ($M = 0.300$, $SE = 0.015$) compared to 7–8-year-olds ($M = 0.671$, $SE = 0.015$, $p = 0.007$) and adults ($M = 0.687$, $SE = 0.015$, $p = 0.007$), and also marginally fewer than 5–6-year-olds ($M = 0.633$, $SE = 0.019$, $p = 0.055$). No differences were obtained in terms of "Ga" (visual) responses (all ps > 0.05). Taken together, it could be inferred that, when presented with stimuli with conflicting audiovisual information, young children aged 3–4 tended to make responses relying more on the auditory modality instead of an audiovisual-integrated fashion as older participants did. Also, results indicated that children aged 5–6 and 7–8 seemed to employ a similar strategy with adults in responding to incongruent stimuli (all ps > 0.05).

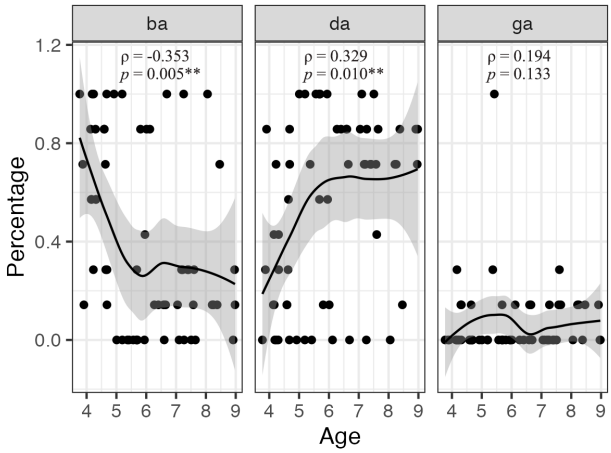


Figure 2: Correlation between chronological age and percentage of different types of responses made. The solid line refers to the LOESS smoother with a span of 0.4.

To further illustrate the correlation between performance and the actual chronological age of child participants, Spearman's ρ was calculated. For congruent trials, Age failed to predict the accuracy of child participants ("Ba": $S = 29324$, estimated Spearman's $\rho = 0.224$, $p = 0.082$; "Da": $S = 30281$, estimated Spearman's $\rho = 0.199$, $p = 0.123$; "Ga": $S = 32743$, estimated Spearman's $\rho = 0.134$, $p = 0.302$). For incongruent trials, Age was revealed to be negatively correlated with the emergence of "Ba" (auditory) responses ($S = 51158$, estimated Spearman's $\rho = -0.353$, $p = 0.005$). In stark contrast, "Da" (audiovisual) responses increased along with

increasing chronological age (S = 25393, estimated Spearman's ρ = 0.329, p = 0.010). No significant correlation was found between Age and the percentage of "Ga" (visual) responses. In a nutshell, results indicated that, during 3–8 years of age, children tended to gradually disengage with unimodal auditory information and shift to take both auditory and visual cues into account along aging.

4. DISCUSSION

The current cross-sectional study aims to track down the developmental trajectory of audiovisual speech perception among Mandarin-speaking children using the McGurk paradigm. For congruent trials, apart from a lower accuracy found in 3–4-year-olds compared to two oldest groups, no significant differences were found. For incongruent trials, a dramatic and earlier developmental shift has been obtained as 3–4-year-olds made a significantly greater number of "Ba" (auditory) responses, relative to any older group of participants. Meanwhile, this group of young children made significantly fewer "Da" (audiovisual) responses than 7–8-year-olds as well as adults, and marginally fewer than 5–6-year-olds. No group differences reached significance in terms of the visual responses, "Ga". Spearman's ρ revealed that age is a positive predictor of "Ba" while negative of "Da" responses.

For congruent trials, Kruskal-Wallis test revealed that 3–4-year-olds significantly underperformed 7–8-year-olds and adults. Since all groups achieved a mean accuracy above 0.95 and age did not correlate with the accuracy of congruent trials, the differences might be yielded by the easiness of the task.

For the incongruent trials, the current results corresponded with previous research on the development of audiovisual speech perception as a theoretical developmental shift was observed [11-12]. Three–four-year-olds preferred to make "Ba" responses relative to any older groups, which aligned with previous findings on the auditory, instead of visual, preference in young children. Such unimodal auditory preference was suggested to consume greater attentional resources of children, giving rise to more auditory-related responses [17-18]. Additionally, the transient and dynamic nature of auditory stimuli and the earlier maturation of the auditory system were believed to contribute to auditory dominance among young children [10, 17]. As a result, when presented with stimuli with conflicting audiovisual information, children tended to make responses according to the information received from auditory relative to visual modality. Besides, 3–4-year-olds made significantly fewer "Da" (audiovisual) responses, indicating that this group of children was less likely to integrate the audiovisual information to form a holistic fused percept as the older groups. Yet 5–6-year-olds were recorded a comparable number of "Da" responses with adults, reflecting a similar audiovisual integrative strategy was employed to generate perceptual outcomes.

Our results suggest that the developmental shift of audiovisual speech integration occurs at an earlier age in Mandarin-speaking children as previous studies showed English-speaking children took around ten years to exhibit an adult-like pattern [11-12]. One possible explanation is that Mandarin-speaking children are more skilled in utilizing the visual cue of the monosyllabic stimuli that are meaningful in Mandarin [19]. Another is that preschool education in China maintains a teacher-centered tradition [20], leading children to focus intensively on teachers' talking faces given that classrooms are always noisy.

The current study potentially mediates findings from English-speaking and Mandarin-speaking children. To the best of our knowledge, existing studies have performed the McGurk paradigm on Mandarin-speaking school-age children only [14-16], while we tried to examine this process on children as young as three years old. Our findings regarding school-age children (7–8-year-olds) and adults were comparable with previous findings. It is noteworthy that none of the differences between 5–6-year-olds and adults in the current study achieved significance. Therefore, we attempt to suggest that the adult-like performance among school-age children in previous and current studies could be attributed to their almost developed ability in audiovisual speech integration. Additionally, the absence of the developmental shift in existing studies is seemingly because it occurs at an earlier stage that is previously overlooked.

5. CONCLUSION AND LIMITATIONS

The McGurk paradigm was performed on a large sample consisting of 61 Mandarin-speaking children and 26 young adults. We found that age posed limited impacts in perceiving congruent stimuli. Results from incongruent trials showed that 3–4-year-olds made significantly more "Ba" (auditory) but fewer "Da" (audiovisual) responses relative to older groups, supporting a developmental shift in audiovisual speech perception occurring at around five, which is earlier than that found in English-speaking children.

The current study admits the following limitations and calls for future investigations. First, reaction times were not compared to discuss the confidence level of children since the responses were made by the experimenters. Second, noise which is reported to fluctuate multisensory perception needs to be taken into consideration.

6. ACKNOWLEDGEMENTS

This research was partly supported by a fellowship award from the Research Grants Council of the Hong Kong SAR, China (Project No. PolyU/RFS2122-5H01).

7. REFERENCES

[1] Rosenblum, L. D., & Dorsi, J. (2021). Primacy of Multimodal Speech Perception for the Brain and Science. In *The Handbook of Speech Perception* (pp. 28–57). Wiley.

[2] Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and aging*, *31*(4), 380.

[3] Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233-252.

[4] Auer, E. T., & Bernstein, L. E. (2007). Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1157–1165.

[5] Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

[6] Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology*, *47*(sup2), S31–S37.

[7] van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186.

[8] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

[9] Ernst, M. O. (2008). Multisensory integration: a late bloomer. *Current Biology*, *18*(12), R519-R521.

[10] Burr, D., & Gori, M. (2012). Multisensory integration develops late in humans. In *The Neural Bases of Multisensory Processes* (pp. 345-362). CRC Press.

[11] Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, *8*(1).

[12] Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and Non-Speech Audio-Visual Illusions: A Developmental Study. *PLoS ONE*, *2*(8), e742.

[13] Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320.

[14] Li, Y., Mei, L., & Dong, Q. (2008). The Characteristics and Development of Audiovisual Speech Perception in Native Chinese Speakers. *Psychological Development and Education*, 24(3), 43-47. [In Chinese]

[15] Liu, M., D, X., Liu., Q. (2020). The Features of Audiovisual Speech Perception in Noise of Children with Autism Spectrum Disorder. *Chinese Journal of Applied Psychology*, 26 (3), 231-238. [In Chinese]

[16] Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America*, *126*(2), 858.

[17] Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 135–141.

[18] Robinson, C. W., & Sloutsky, V. M. (2004). Auditory Dominance and Its Change in the Course of Development. *Child Development*, *75*(5), 1387–1401.

[19] Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, 41(1), 93–113.

[20] Hu, B. Y., Fan, X., Yang, Y., & Neitzel, J. (2017). Chinese preschool teachers' knowledge and practice of teacher-child interactions: The mediating role of teachers' beliefs about children. *Teaching and Teacher Education*, 63, 137–147.

GESTURAL ALIGNMENT AND ACCOMMODATION IN SPEAKER-LISTENER HEAD GESTURES

Karee Garvin[1,2] and Kathryn Franich[1,2]

[1]Harvard University and [2]University of Delaware
kareegarvin@fas.harvard.edu, kfranich@fas.harvard.edu

ABSTRACT

The timing of both manual co-speech gestures and head gestures is sensitive to prosodic structure of speech. However, head gesters are used not only by speakers, but also by listeners as a back-channeling device. Little research exists on the timing of gestures in back-channeling. To address this gap, we compare timing of listener and speaker head gestures in an interview context. Results reveal the dual role that head gestures play in speech and conversational interaction: while they are coordinated in key ways to one's own speech, they are also coordinated to the gestures (and hence, the speech) of a conversation partner when one is actively listening to them. We also show that head gesture timing is sensitive to social dynamics between interlocutors. This study provides a novel contribution to literature on head gesture timing and has implications for studies of discourse and accommodation.

Keywords: Co-speech gesture, gesture alignment, speech timing, discourse, accommodation

1. INTRODUCTION

Literature on the timing alignment of speech and co-speech gesture demonstrates that gestures are sensitive to the prosodic signal, including stress and intonational phrase boundary [1]. In particular, a number of studies have shown that gestures are timed to stressed vowels or pitch accents (see [2] for an overview) and gestures are retracted, or shifted earlier, at intonational phrase boundaries [3, 4], with similar results across manual and head gestures.

While manual gestures are crucially timed with one's own speech, head gestures are employed by both speakers and listeners during conversational interaction: while speakers are known to align their head gestures with prominent syllables in their own speech, listeners use head gestures as a back-channeling device to indicate engagement with the speaker [5]. Listener head gestures are known to be denser at points of overlapping discourse [6]; however, how head gestures are timed relative to the discourse, e.g., whether gestures are still timed to speech or other aspects of the interaction, is unstudied. Furthermore, back-channeling cues are subject to accommodation, where the social dynamic between interlocutors influences speech patterns such as the timing [7], acoustics, and prosody of speech [8, 9]. Interlocutors that accommodate to their speech partners are viewed as more likable and the resulting conversations are perceived as more natural and successful [10, 11, 12, 13, 14]. Whether social roles affect head gesture coordination between speakers remains unknown.

We analyze how interlocutor roles—speaker vs. listener—affect head gesture timing. Furthermore, we analyze how timing differs across speakers depending on the social role of the interlocutor by analyzing differences in the interviewer and subjects using a corpus of Salon Talks interviews. We hypothesized that, consistent with previous literature, the timing of one's own gestures would align with stressed syllables in one's own speech and that this alignment would be sensitive to proximity to an intonational boundary. Given the importance of head gestures in listener back-channeling, we also predicted that participant head gestures would align closely with an interlocutor's speech when the participant was actively listening to their interlocutor and that timing of the interviewer's gestures would be more sensitive to the interview subject's gestures than the other way around. Our findings confirm that listener gestures are highly sensitive to the timing of the interlocutor's gestures. Furthermore, we find that the interviewer shows greater accommodation to the timing of the subject's gestures than the subjects to the interviewer.

2. METHODS

2.1. Data

To analyze gestural turn-taking, we used interviews from Salon Talks retrieved from YouTube. Salon

Talks interviews were optimal for analyzing gestural turn-taking because the videos had few camera angles and had long durations shot from a camera angle in which speakers' heads and upper bodies were simultaneously visible (see Figure 1). We examined clips from four interviews conducted by Mary Elizabeth Williams with the subjects Zoey Deutch (posted Mar 21, 2018), Jesse Eisenberg (posted Mar 11, 2019), Randall Park (posted Jun 6, 2019), and Mira Sorvino (posted Nov 7, 2018). The clips were approximately two minutes in length and consisted of a single camera angle where both speakers were visible for the duration of the clip.



Figure 1: Still image from Salon Talks interview with Mary Elizabeth Williams and Randall Park.

2.2. Gesture and speech annotation

Gestures were coded by a team of researchers trained in gesture coding using ELAN [15] following the MIT Gesture Studies Coding Manual [16], which outlines several phases of the gesture including preparations, strokes, holds, and recoveries based on [17]. This method was updated for head gestures following [2] and [3]. The videos were coded for gestural phase, apex, and interlocutor role. The apex of the gesture was coded as the point of maximum extension. Because ELAN does not permit annotation of a single point in time, but instead requires interval annotations, the point of maximum extension was annotated as the end point of the apex (T2) and extended two frames prior to the apex. Calculations regarding apex used the T2 of the apex interval. For each video, there were approximately 250 apexes coded per subject. The interlocutor role was coded as *speaker*, *listener*, or *speaker/listener*, where both the interviewer and subject were speaking simultaneously. Gestures that were coded as *speaker/listener* were not analyzed in this study as gestures could not be reliably attributed to the speech of a single speaker, and thus, the interlocutor role was ambiguous. A transcript of the interview clip was made by a team of researchers and was aligned using the FAVE Forced Aligner [18]. The forced alignment data was then hand-corrected by a team of researchers.

2.3. Speech to gesture and gesture to gesture timing

To analyze the timing between gestural apexes and phones, the lag time between apex and phone was calculated as the T2 of the apex minus the start time (T1) of the phone. Phones were coded as either stressed vowel, unstressed vowel, or consonant. To analyze how intonational boundary affected lag time between vowels and gestures, the distance to nearest vowel was calculated following the same method. Likewise, intonational boundary was coded as either initial, for the first word of an intonational phrase, final for the final word of an intonational phrase, or else medial. Together, these calculations allowed us to analyze the alignment between gestural apexes and prosodic elements including, phone, stress, and intonational phrase boundary (see Section 3.1).

To analyze the timing between gestures across interlocutors, the apex lag time was calculated as the time from the T2 of one apex to the T2 of another apex in a sequence of alternating gestures between interlocutors, as illustrated in Figure 2. In other words, in a sequence of Gesturer A^1 - Gesturer B - Gesturer A^2, Gesturer B's lag time was calculated as Gesturer B minus Gesturer A^1.

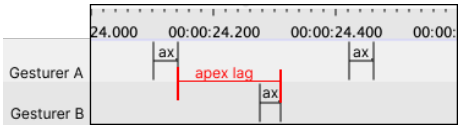


Figure 2: Example of alternating gestures between speakers used to calculate apex lag. Endpoint of ax intervals coincides with the point of maximal head extension.

This measure of apex lag was used to analyze the timing relationship between interlocutors, i.e., *speaker* and *listener*, across the dyads, e.g., Mary Elizabeth Williams and Randall Park, and based on participant role, i.e., *interviewer* and *subject*. The results of this analysis are discussed in Section 3.2.

3. RESULTS

3.1. Timing of speech and gestures

Overall, the results of the analysis on the timing between gestures and speech are consistent with previous work on the alignment of speech and co-speech gesture [2, 3, 4]. We found that gestural apexes were more likely to align with stressed vowels than with other segments. The correlation

between stressed vowels and gesture apex alignment was analyzed using a Pearson's Chi-squared test ($p < 0.001$). There is a significant correlation between the alignment of gesture apexes and stressed vowels, as illustrated in Figure 3. Positive or 'attraction' relationships are shown in blue and negative or 'repelling' relationships are shown in red, where the darkness of the shade of blue or red indicates the strength of the relationship. The size of the bubble indicates the contribution of the variable to the model, where larger bubbles indicate variables that have a larger effect in the model. As indicated by the large, dark blue bubble for stressed vowels aligned with apexes, there is a strong attraction relationship between stressed vowels and gestural apexes, consistent with [2].

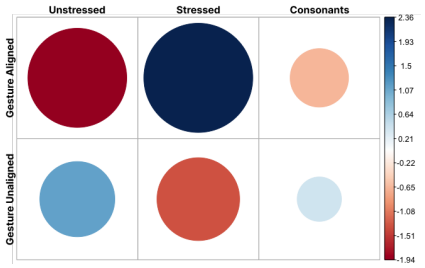


Figure 3: Correlation plot between segment type and gesture aligned vs. unaligned segments.

We also examined intonational phrase boundaries and apex alignment by analyzing the time between vowels and aligned apexes and found that gestures retract, or occur earlier with respect to the vowel, at intonational phrase boundaries (Figure 4). We hypothesize that this result was obscured for Zoey Deutch because she had fewer gestures occurring at phrase boundaries.

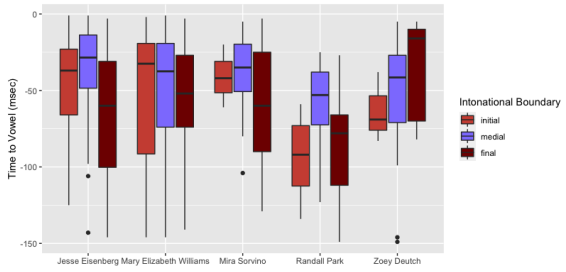


Figure 4: Time between apexes and vowels at intonational boundaries, plotted by subject. Value of 0 corresponds to perfect alignment between gesture apex and vowel onset.

A linear mixed effects model (lmer function of the lme4 package [19]) of the time between vowels and apexes shows a significant effect, consistent with the findings in [3] (Initial vs Final: $\beta = 15.082$, $t = 2.157$, $p < 0.05$; Medial vs Final: $\beta = 25.626$, $t = 5.446$, $p < 0.001$). All models were originally fit with the maximal random effect structure with random slopes for all predictors, but where the fit was singular, we removed random slope terms that eliminated singularity [20].

Together, these results show that head gestures are sensitive to prosodic elements of speech including both stress and intonational boundary. Gestural apexes tend to coincide with stressed vowels, but are retracted at intonational phrase boundaries.

3.2. Timing between conversation partner gestures

Head gestures, in addition to being used to augment one's own speech, are likewise used by listeners as a back-channeling cue, indicating interaction and engagement. Furthermore, speakers provide timing cues to listeners both in the timing of their speech and in the timing of their own gestures. Thus, in addition to analyzing the timing between speech and head gestures, we also analyzed the timing between interlocutor gestures, focusing on the difference in timing between listeners and speakers at points in the video where there was no overlap between speakers, and as such, speech role was unambiguous. Overall, listeners time their apexes to their interlocutor's gestures, while speakers time their gestures to their own speech.

To analyze the influence of interlocutor gestures on apex timing, we calculated the lag between each participant's head gesture apex relative to the immediately preceding apex of their conversation partner. This provides a measure of alignment between gestures across participants. We found that overall lag time was shorter when participants were listening to their interlocutor than when they were speaking. A linear regression demonstrates that this effect is significant (Interlocutor Role (speaker): $\beta = 0.25$, $t = 3.01$, $p < 0.01$; Participant Role (subject): $\beta = 0.28$, $t = 3.44$, $p < 0.001$).

Figure 5 plots the lag between speakers for each of the interview subjects alongside the interviewer and shows that apex lag times were shorter when the subject was listening than speaking for all subjects except Mira Sorvino. The results likely differ for Sorvino because she gestured less while listening than the other subjects, thus providing a relatively limited sample and obscuring the trend. The plot additionally shows that the interviewer's lag time was consistently shorter than her subject's relative to her own and that the difference in lag between her apexes as speaker vs. listener was smaller.

On the other hand, gesture-to-speech timing

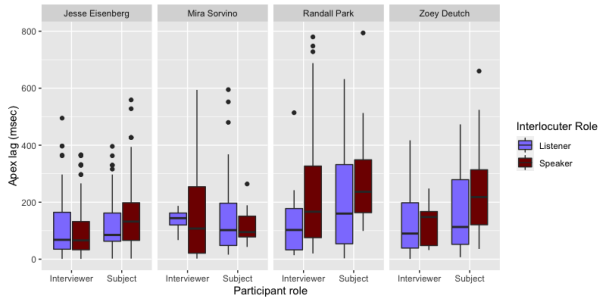


Figure 5: Time between the gestural apex of Speaker A to the gestural apex of Speaker B plotted for four dyads with Mary Elizabeth Williams as the interviewer across all four dyads.

showed a very different pattern: all five participants showed shorter apex to vowel lag times between apexes and phones when speaking compared to listening, shown in Figure 6; this is consistent with the prediction that participants would time their gestures with their own speech when in the speaker role. A linear mixed effects model (lmer function of the lme4 package [19]) predicting time between apexes and vowels from interlocutor role with participant as a random intercept shows that this effect was significant (Interlocutor role (speaker): $\beta = 12.433$, $t = 2.381$, $p = 0.05$).

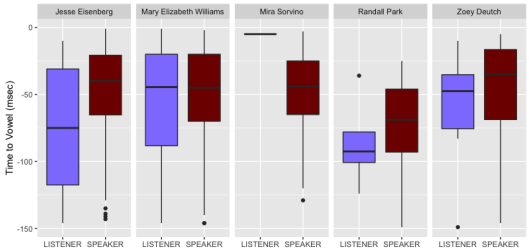


Figure 6: Time between apexes and vowels by interlocutor role, where lag time is shorter when participants are speaking compared to listening.

Overall, these results show that speakers time their gestures to speech while listeners time their gestures to their interlocutor's gestures. Furthermore, the social role of the interlocutor influences the degree of this effect, where the interviewer showed greater accommodation to the interview subject's gestures regardless of the interviewer's interlocutor role.

4. DISCUSSION

The results of this study on the timing between head gestures and speech are overall consistent with the findings of previous literature [2, 3, 4], where gestures are sensitive to prominence in the speech signal. Namely, gestural apexes are timed to stressed vowels and are retracted from stressed vowels at intonational phrase boundaries. However, a novel contribution of this study is the clear difference in gestural timing depending on interlocutor role. In particular, in analyzing the lag time of alternating apexes between interlocutors, there is a difference between the timing of gestural apexes in listeners compared to speakers, where apex-to-apex lag times are shorter for listeners than for speakers. This is contrasted with the lag time between apexes and phones, where lag times are shorter between apexes and phones for speakers compared to listeners. This asymmetry is indicative of the dual role that head gestures play, coordinating with a participant's own speech when they are speaking, and with an interlocutor's gestures when listening.

In addition, the social role of the participants influenced gestural timing. In particular, the interviewer had shorter apex lag times both when speaking and when listening, and furthermore, shorter lag times than the interview subjects. These results are consistent with literature showing interlocutors who accommodate to their conversation partner are viewed as more likable and the interactions are viewed as more successful [10, 11, 12, 13, 14]. Mary Elizabeth William's role as the interviewer positions her to accommodate more to her interview subjects, and thus, her gestural apexes more closely align to the interview subjects' gestures. These results are an essential contribution to literature on gestural timing and accommodation.

5. CONCLUSION

This study confirms the findings that speaker gestures are timed to prominent elements of the speech signal. In addition, we demonstrate that gestural alignment differs depending on whether the gestures are produced by speakers or listeners. Namely, while speakers' gestures are more closely timed to aspects of the speech signal, listeners' gestures are more closely timed to their interlocutor's gestures. Furthermore, patterns in timing differ depending on the social role of the interlocutor, where the interviewer shows greater accommodation to the interview subjects. These results provide novel insights into gestural alignment, showing that listeners use the visual cues of an interlocutor's gestures to plan their own gestures and that listeners engage with and accommodate to their interlocutor's gestures.

6. ACKNOWLEDGMENTS

The authors would like to thank the graduate and undergraduate researchers who helped annotate and process the data discussed in this paper, including Crystal Akalu, Christine Belden, Kylie Boggs, Nicole Curnan, Luc DiNardi, Walter Dych, Bryant Graybeal, Lindsay Hawtof, Jacquelyn Janocha, Brett Lindsey, Zoe Lipkin, Madison Miller, Meghan O'Brien, Victoria Ochlan, Anna Schumeyer, Nicole Taylor, Ta-tran Tran, and Juliette Winnard. This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are our own.

7. REFERENCES

[1] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012. [Online]. Available: https://doi.org/10.1515/lp-2012-0006

[2] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[3] N. Esteve-Gibert, J. Borràs-Comes, and E. Asor, "The timing of head movements: The role of prosodic heads and edges," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4727–4739, 2017.

[4] J. Krivokapič, M. Tiede, and M. E. Tyrone, "A kinematic analysis of prosodic structure in speech and manual gestures." in *ICPhS*, 2015.

[5] J. B. Bavelas and J. Gerwing, "The listener as addressee in face-to-face dialogue," *International Journal of Listening*, vol. 25, no. 3, pp. 178–198, 2011.

[6] S. G. Danner, J. Krivokapič, and D. Byrd, "Co-speech movement in conversational turn-taking," *Frontiers in Communication*, vol. 6, 2021.

[7] Štefan Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[8] Y. Lee, S. G. Danner, B. Parrell, S. Lee, L. Goldstein, and D. Byrd, "Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task," *PLoS ONE*, vol. 13, 2018.

[9] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[10] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[11] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech accommodation theory: The first decade and beyond," *Annals of the International Communication Association*, vol. 10, no. 1, pp. 13–48, 1987.

[12] J. B. Hirschberg, A. Nenkova, and A. Gravano, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL/HLT*, 2008.

[13] R. L. Street Jr, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[14] S. Stoyanchev and A. Stent, "Lexical and syntactic adaptation and their impact in deployed spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 189–192.

[15] "ELAN (version 6.4) [computer software]," 2022. [Online]. Available: https://archive.mpi.nl/tla/elan

[16] "MIT Speech Communication Group Gesture coding manual." [Online]. Available: http://scg.mit.edu/gesture/coding-manual.html

[17] A. Kendon, *Gesticulation and Speech: Two Aspects of the Process of Utterance*. Berlin, New York: De Gruyter Mouton, 1980, pp. 207–228.

[18] "FAVE (Forced Alignment and Vowel Extraction) suite." [Online]. Available: https://www.research.ed.ac.uk/en/publications/fave-forced-alignment-and-vowel-extraction-suite

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, 2014.

[20] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

THE PROSODY-GESTURE COORDINATION IN CHINESE EFL LEARNER' ACADEMIC ORAL PRESENTATION

Yating Cao[1,2] Hua Chen[1]

[1]Nanjing University [2]University of Warwick
yatingcao@smail.nju.edu.cn; chenhua@nju.edu.cn

ABSTRACT

Human communication is multimodal. However, few studies have examined how L2 learners, particularly those whose L1 is a tonal language, coordinate their beat gestures and prosody in an intonation-language speech production. This study explored how Chinese EFL speakers coordinated pitch accents and beat gestures in academic oral presentation and the effects of speakers' language proficiency. Results revealed that the intermediate-level Chinese EFL learners produced beat gestures regardless of the placement of pitch accents and the information status of their accompanying speech. Beat gestures were used to fulfill the meta-linguistic function. Whereas, proficient learners produced significantly fewer beat gestures, and beat gestures were more likely to co-occur with pitch accents to stress the crucial information of discourse for its meta-pragmatic function. These findings provide empirical evidence from L2 learners to expand the literature on multimodal prosody and will have pedagogical implications for English teaching.

Keywords: beat gestures, pitch accent, gesture-speech coordination, proficiency level

1. INTRODUCTION

Speech communication is multimodal in nature. As one of the crucial non-verbal cues, gestures often accompany people's speech production, providing another window into people's minds [1-3]. McNeill [3] postulated three synchrony rules for speech and gestures, namely semantic, pragmatic and phonological synchrony rules. The first two refer to that gestures often parallel meaning and pragmatic functions with speech, and the last one, phonological synchrony, which is also the most pertinent to the present study, states that the stroke phase of gestures tends to occur at or just before the stressed syllable of the lexical affiliate. Here, the stroke means the only obligatory and the most prominent part of a gesture [2-3].

Pitch accent is one type of prosodic specification where a segment (a syllable or a word) within a sentence stands out from its context to convey focus or information status [e.g., 4-5]. In more recent years, a growing body of research has consistently reported a close coordination between prosodic prominence and manual gestures [e.g., 7-10]. Beat gestures, according to [3], refer to "a simple flick of the hand or fingers up and down; or back and forth". This type of gesture has been found to be the most frequently used gesture in daily conversation, particularly taking up 94.6% of gesture types in academic-style conversations [11]. And it has been believed that beat gestures, like pitch accent, is usually used to emphasize the information status as a "gestural yellow highlight" [12].

There is evidence that beat gestures and pitch accent are closely coordinated. The preliminary analysis by [13] investigated how different gesture types were synchronized with English speech, and showed a strong coupling between beat gestures and pitch accent. For example, 86% of beat gestures produced by one Australian English speaker co-occurred with a pitch accent. Shattuck-Hafnagel & Ren [11] analyzed 1,334 beat gestures in a 20-min academic speech and found a close coordination between beat gestures and the prosodic structure of spoken utterances.

While the phonological coordination between prosodic and gestural prominence is well established in prior research, it is mainly based on studies of non-tonal languages, such as English [14], Catalan [7], Dutch [15], and French [16]. It would be interesting to ask whether and how prosody and gestures coordinate in non-stress tone languages. Fung et al. [17] examined how ten native speakers of Hong Kong Cantonese coordinated prosody and deictic gestures in a picture-verification task. They not only found a close coordination between prosodic and gestural prominence but also revealed that the prosodic anchor was not F0 as previous studies showed but durational increase instead, which manifested cross-linguistic similarities and differences.

Furthermore, it should also be noted that most research has been focusing on how native speakers achieve prosody-gesture coordination, while research on the patterns of non-native speakers has been quite sparse. McCafferty [18] examined EFL learners' use of beat gestures in conversation and suggested that these gestures were the materialization of their inner prosody. Further, he suggested that L2 learners employed beat gestures for a meta-linguistic purpose, and that native speakers were more likely to use beat

gestures to fulfill its meta-pragmatic function. This research strengthens McNeill's [3] argument by extending it beyond native speakers to L2 learners. Graziano et al. [19] explored how English and French monolinguals and bilinguals achieved phonological coordination in a cartoon retelling task. Results showed that there was no significant difference concerning the time-aligned coordination of gestures and speech, which is possibly explained by the relatively high level of proficiency in the two languages. In light of the findings, it would be an important aspect to further test participants at different proficiency levels to better inform the nature of gesture-speech links and theories of language development as well. Interestingly, as mentioned before, Mandarin Chinese is a tonal language that exhibits the syllable-timed rhythmic pattern; hence, Chinese EFL learners might accentuate every word within a sentence without clear communicative intentions as a result of the L1 transfer [20]. In this sense, how this group of learners coordinates prosodic and gestural prominence in their L2 warrants further investigation.

To respond to these gaps, the present study took the preliminary attempt to analyze the academic oral English presentation by Chinese EFL learners and addressed the following questions: (1) How do beat gestures align with prosodic prominence in Chinese EFL learners' speech production? (2) Is there any difference between the two proficiency groups?

2. METHODS

2.1. Data source

Multimodal Corpus of Academic English Speech by Chinese Learners (MCAESCL) [21] was served as data source in this study. All the participants in this corpus were native Chinese speakers, and they were required to provide a two-minute English academic oral presentation. There were no other audiences other than the researcher in the experiment setting, and the whole process was captured by a high-definition camera. Also, they were randomly assigned to topics, that is "*How to prepare for an oral presentation*?" for the informative speech genre or "*To read selectively or extensively? What is your understanding? Support your ideas with details*" for persuasive speech. Both topics were relevant to students' college academic life so that they could express their opinions [see 21 for further review].

So far, MCAESCL consists of 1,763 speech samples, with varying majors, academic levels of speakers' university, and gender balanced. Among all these data, 924 speakers' speech productions were accompanied by gestures and became the candidate data for the studies related to gestures. Besides, all the speech samples have been rated by 15 linguistic experts from various linguistic dimensions (i.e., phonetics, syntax, vocabulary, and discourse) via focused analytical scoring, and all the results have been validated through multiple methods. Based on the total score, all the samples could be ranked and categorized into different proficiency levels (i.e., advanced, intermediate, and low).

Due to the exploratory nature of the present study the sample size is quite modest. 12 speech samples were selected, with half from the advanced learner (AL) and intermediate learner (IL) group, respectively. In addition, all the chosen speakers gave speeches of the same topic and achieved gender balance.

2.2. Annotation

The data annotation consisted of two parts – prosody and gestures. Previous studies have shown that the two modalities, namely audio and visual mode, might influence the prominence perception [e.g., 11, 22]; therefore, these two parts of annotations were independently completed to eliminate possible effects.

For the prosodic annotation, the speech was initially transcribed orthographically and labeled for its intonational structure – mainly for pitch accents, using Praat [23] and ToBI as the prosodic annotation system. And all the gesture annotation was accomplished using frame-by-frame analysis in ELAN [24]. There were three main tiers used in the current coding scheme as follows:

- Main Tier 1: orthographic transcription
- Main Tier 2: Manual gesture phase
- Main Tier 3: Prosodic information

Specifically, for the Main Tier 2, the present study followed [25]'s proposal for coding the manual gesture phase, which included preparation, beginning, and end of the gesture stroke, retraction phase, as well as optional pre- or post-stroke recovery. In addition, a sub-tier was coded for the gesture apex separately to indicate the gestural prominence. The Main Tier 3 was coded after the completion of the first two tiers, which contains two main sub-tiers. The first sub-tier comprised already annotated ToBI labelling imported from Praat. And the second sub-tier was used to label whether the gestural prominence was linked with prosodic prominence or not, suggesting by the symbol (+) and (-).

Consistent with prior research [26], the annotation was carried out by the first author and then the second coder, who is an experienced coder of gestures, was invited to annotate 25% of randomly-selected data. Then, the Cohen's Kappa value (k=0.82) indicated the high inter-annotator reliability.

2.3. Data analysis

The present study calculated the percentage of co-occurrence of gestural prominence (i.e., gesture apex) in beat gestures and prosodic prominence (i.e., pitch accents). The chi-square analysis was then carried out to examine whether there existed a difference between the two proficiency levels (i.e., AL vs. IL) concerning the percentage of occurrence of gestural prominence and prosodic prominence.

3. RESULTS

Table 1 displays how Chinese EFL learners use beat gestures to achieve the coordination between prosodic and gestural prominence. It could be seen that advanced learners used more gestures than intermediate-level learners did in their academic oral presentations. Notably, beat gestures took a large proportion of the total gesture use as shown by the proportion. In addition, Chinese EFL learners at the intermediate level almost only used beat gestures throughout their speech production, exceding more than 83% of the total gesture types. However, learners of this group seldom coordinated beat gestures with pitch accents in spoken language.

Further chi-square analysis was performed to examine whether there existed differences between speakers of language proficiency levels in terms of the co-occurrence of prosodic and gestural prominence. Results revealed a significance in these two groups (χ^2 = 53.78, df = 1, p = 0.0001), with advanced learners showing more coordination.

Speakers	Gestures (Total)	Beat gestures (%)	The coordination between prosodic prominence and gestural prominence (%)
AL 1	43	22 (51.17%)	10 (45.5%)
AL 2	30	14 (46.67%)	9 (64.3%)
AL 3	47	38 (80.86%)	20 (52.64%)
AL 4	39	27 (69.23%)	17 (62.96%)
AL 5	47	35 (74.47%)	21 (60%)
AL 6	43	21 (48.83%)	15 (71.43%)
IL 1	49	43 (87.76%)	13 (30.23%)
IL 2	19	19 (100%)	3 (15.79%)
IL 3	33	29 (87.88%)	9 (31.03%)
IL 4	44	40 (90.91%)	4 (10%)
IL 5	43	36 (83.72%)	8 (22.22%)
IL 6	24	20 (83.33%)	5 (25%)

Table 1: The gesture use and coordination between prosodic prominence and gestural prominence in Chinese EFL learners' academic oral presentation.

4. DISCUSSION

The present study has two research goals. The first one is to show how Chinese EFL learners coordinate their prosodic prominence and gestural prominence (beat gestures) to construct meaning in academic oral presentation. And the second one is to examine the role of language proficiency in this issue.

Firstly, the current study found that beat gestures constituted a large proportion of Chinese EFL learners' speech production. This finding offered evidence from Chinese EFL learners to [11] that beat gestures are the most frequently used gesture type in native English speakers' academic-style report. In fact, Alibali et al. [27] compared speakers' gesture use (representational vs. beat gestures) when they could see the interlocutors or not. Results showed that the number of beat gestures was not affected as much as representational gestures under the invisibility condition, which indicated that beat gestures were more likely to be used for self-regulatory functions. In the present study, speakers might use beat gestures to accompany their speech to help themselves to regulate the speech flow as their "inner prosody" [18]. Another important factor worth consideration is the task effects. Previous studies usually employed the cartoon-retelling task to elicit the gestures, where certain gesture types such as iconic gestures and metaphoric gestures are used much more often. However, as Tabensky [28] suggested, different speech genres may influence how speakers use gestures to convey meaning. The academic oral presentation used in this study is more abstract compared to other tasks, which might be challenging for speakers to produce more meaning-based gestures.

Second, even though the proportion in the present study was not big enough, it still found that Chinese EFL learners could coordinate gestural prominence and prosodic prominence to construct discourse meaning in their academic oral English presentation. Since previous studies has mainly focused on the non-tonal languages [7, 14-16], the current study provided evidence that as tonal language speakers, Chinese speakers also achieved the prosody-gesture coordination in their L2 speech (i.e., English). This evidence also shows that L2 speakers would consider speech and gestures in their speech production planning process to form an integrated system [1-3], suggesting a universality of embodiment in human communication. However, it should be noted that the present study only examined how Chinese speakers coordinated prosodic and gestural prominence in their English production, while no further investigation of how they performed in their L1 speech (i.e., Mandarin Chinese). Therefore, these speakers' ability to achieve such coordination could simply because they

were speaking English, which is found to be closely linked between speech and gestures in production.

Third, the proficiency level of speakers was found to be an important factor in prosody-gesture coordination [19]. In the present study, advanced Chinese EFL learners were more likely to use the apex of beat gestures in conjunction with pitch accents to emphasize important information, thereby achieving embodied communication. In contrast, intermediate-level Chinese EFL learners tended to use gestures without clear communicative purposes, using beat gestures for every word or even syllable. This group difference is similar to what McNeill [3] observed in the beat gesture production of native English-speaking adults and children. Specifically, due to the higher English proficiency of advanced English learners, they were more likely to perform like native speakers by better controlling their speech planning process, where beat gestures were used for pragmatic purposes. Whereas, intermediate-level Chinese EFL learners may experience higher cognitive load during speech production, resulting in weaker ability to convey their communicative intentions through a combination of prosodic and gestural devices. Thus, beat gestures were used for a meta-linguistic function, similar to what native children do [3]. Another possible factor is the different rhythmic pattern between Chinese and English. Since Chinese is a syllable-timed language, where every syllable is stressed to convey meaning, intermediate-level Chinese EFL learners might be negatively influenced by their native language and tend to stress each word or syllable in English production. Consequently, the usage of beat gestures may simply reflect the "materialization of their syllable-timed rhythmic pattern in spoken English," as McCafferty [18] stated. However, this conclusion should be made with caution since the present study did not directly examine how Chinese EFL learners produce prosodic and gestural prominence in their L1. Further research is needed to explore whether the pattern of intermediate learners is more similar to that of Chinese native speakers and the pattern of advanced learners is more like that of English native speakers.

5. CONCLUSION

The present study investigates how Chinese EFL learners coordinate the prosodic prominence (pitch accents) and gestural prominence (apex of beat gestures) in academic English oral representation. Results showed that L2 learners can also use their gestures as a tool to construct discourse as native speakers did, giving empirical evidence for the integration of speech and gesture in L2 speakers. However, the coordination patterns may vary depending on the level of L2 proficiency of the speaker. Chinese EFL learners at advanced level were more likely to coordinate prosodic prominence and gestural prominence to convey their communicative intentions. The findings of the study could offer L2 evidence to the literature on speech-gesture coordination and provide positive implications for multimodal English teaching and learning.

However, this study was still the preliminary attempt to tap into the phonological coordination of beat gestures and pitch accents for the smaller sample size. In addition, further analysis of phonological coordination patterns also merits further investigation to advance the understanding of this field. For example, it is interesting to explore prosodic anchoring in such phonological coordination in Chinese EFL learners' speech. Additionally, further investigation is needed to understand how Chinese EFL learners coordinate prosodic and gestural prominence in their native language, in order to better explain the observed group differences. The final note is that the task used in the present study was academic oral presentation, which differs from the commonly-used cartoon-retelling task. Therefore, future research may consider the task effects on L2 learners to further understand the prosody-gesture coordination.

6. ACKOWNLEDGEMENTS

This study was supported by the National Social Science Foundation of China (20AYY013), and China Scholarship Council (202206190122).

7. REFERENCES

[1] Goldin-Meadow, S. 2003. *Hearing Gesture: How Our Hands Help us Think*. Harvard University Press.

[2] Kendon, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

[3] McNeill, D. 1992. *Hand And mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

[4] Bishop, J., Kuo, G., Kim, B. 2020. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics* 82, 100977.

[5] Cutler, A. 1984. Stress and accent in language production and understanding. In: Gibbon, D., Richter, H. (Eds.). *Intonation, Accent and Rhythm* Berlin: De Gruyter, 77-90.

[6] Cutler, A. 2005. Lexical stress. In: Pisoni D. B., Remez, R. E. (Eds.). *The Handbook of Speech Perception*. Malden, MA: Blackwell, 264–289.

[7] Esteve-Gibert, N., Pons, F., Bosch, L., Prieto, P. 2014. Are gesture and prosodic prominences always coordinated? Evidence from perception and production. *Proceedings of the International Conference on Speech Prosody 2014*, Dublin, Ireland, 222-226.

[8] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57(3), 396–414.

[9] Krivokapić, J., Tiede, M. K., Tyrone, M. E. 2017. A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology* 8(1), 3.

[10] Wagner, P., Malisz, Z., Kopp, S. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57, 209–232.

[11] Shattuck-Hufnagel, S., Ren, A. 2018. The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology* 9.

[12] McNeill, D. 2006. Gesture: A psycholinguistic approach. *The Encyclopedia of Language and Linguistics*, 58–66.

[13] Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S. 2004. The timing of speech-accompanying gestures with respect to prosody. *Journal of the Acoustical Society of America* 115(5), 2397.

[14] Loehr, D. P. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3(1), 71–89.

[15] De Ruiter, J. P. 1998. *Gesture and Speech Production*. Unpublished Doctoral Dissertation. Nijmegen: Radboud University.

[16] Ferré, G. 2014. A multimodal approach to markedness in spoken French. *Speech Communication* 57, 268-282.

[17] Fung, H. S. H., Mok, P. P. K. 2018. Temporal coordination between focus prosody and pointing gestures in Cantonese. *Journal of Phonetics* 71, 113–125.

[18] McCafferty, S. G. 2006. Gesture and the materialization of second language prosody. *International Review. of Applied Linguistics in Language Teaching* 44(2), 179–207.

[19] Graziano, M., Nicoladis, E., Marentette, P. 2020. How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Language Learning* 70(1), 266–304.

[20] Chen, H., Wang, X. 2015. A study of English phrasal accentuation in Chinese EFL learners'read speech. *Foreign Languages and Their Teaching* 03, 13-18+40.

[21] Chen, H., Cao, Y., Ma, D. 2022. MCAESCL-A Multimodal Corpus of Academic English Speech by Chinese Learners. *Technology Enhanced Foreign Language Education* 3, 45-51+109.

[22] Prieto, P., Cravotta, A., Kushch, O., Rohrer, P. L., Vilà-Giménez, I. 2018.Deconstructing beat gestures: A labelling proposal. *Proceedings of the 9th International Conference in Speech Prosody 2018*, 201-205.

[23] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer. Version 6.2.06, https://www.praat.org.

[24] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. ELAN: A professional framework for multimodality research. *In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec 2006/.

[25] Shattuck-Hufnagel, S., Ren, A., Mathew, M., Yuen, I., Demuth, K. 2016. Non-referential gestures in adult and child speech: Are they prosodic? *Speech Prosody 2016*, 836–839.

[26] Parrill, F. 2010. The hands are part of the package: Gesture, common ground, and information packaging. In: Rice, S., Newman, J (Eds.). *Empirical and Experimental Methods in Cognitive/Functional Research.* Stanford Publications, 285-302.

[27] Alibali, M. W., Heath, D. C., Myers, H. J. 2001. Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language* 44, 169–188.

[28] Tabensky, A. 2008. Expository discourse in a second language classroom: How learners use gesture. In S. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research*. London, UK: Routledge. 298—320.

COMPARING MANUAL VS. SEMI-AUTOMATED METHODS FOR THE CODING OF CO-SPEECH GESTURES

Walter Dych[1], Karee Garvin[1,2], Kathryn Franich[1,2]

[1]University of Delaware, [2]Harvard University
wdych@udel.edu, kareegarvin@fas.harvard.edu, kfranich@fas.harvard.edu

ABSTRACT

While motion capture is rapidly becoming the gold-standard for research on the intricacies of co-speech gesture and its relationship to speech, traditional marker-based motion capture technology is not always feasible, meaning researchers must code video data manually. We compare two methods for coding co-speech gestures of the hands and arms in video data of spontaneous speech: manual coding and semi-automated coding using OpenPose [1], a markerless motion capture software. We provide a comparison of the temporal alignment of gesture apexes based on video recordings of interviews with speakers of Medumba (Grassfields Bantu). Our results show a close correlation between the computationally calculated apexes and our hand-annotated apexes, suggesting that both methods are equally valid for coding video data. The use of markerless motion capture technology for gesture coding will enable more rapid coding of manual gestures, while still allowing for direct comparison with manually-coded data.

Keywords: co-speech gestures, gesture coding methods, speech timing, speech-gesture alignment

1. INTRODUCTION

Conducting research on speech and co-speech gestures can be challenging due to the time-consuming process of manually annotating gestures in video data when marker-based motion capture methods are not available. This can be especially cumbersome when studying gestures in languages spoken in regions where marker-based technology is not readily accessible, increasing the necessity of manually annotating data. With the advent of markerless motion capture technology, semi-automated coding of gestures from video data is now possible [2], and the potential for rapid and accurate annotation of gestures is greatly expanded. An open question concerns how well traditional manual coding methods align with results of semi-automated methods. This paper investigates manual and semi-automated annotation methods to assess the validity and comparability of the methods.

2. METHODS

2.1. Data

Gesture analysis in this study is based on a corpus of video/audio data collected with four Medumba speakers (2 male and 2 female) through interviews in Banganté, in the West Region of Cameroon. The participants were recorded in pairs engaged in conversation with each other and the interviewer about cultural practices specific to the region. The speech examined in the study is therefore relatively 'naturalistic,' meaning that participants were free to speak and gesture as they pleased, and the topic of conversation was relatively unconstrained.

2.2. Manual method

Manual gestures were coded using ELAN [3] and the MIT Gesture Studies Coding Manual as a guide. This manual outlines several phases of gestures based on [4], including preparations, strokes, and holds. We also coded the 'apex' of the gesture, which occurs within the stroke phase. The apex has been identified in prior research as an important landmark in speech-gesture timing [5, 6, 7]. Each annotated phase is exemplified in Figure 1.



Figure 1: Gesture Landmarks

Though the apex is thought of traditionally as the point of maximum extension of the articulators (e.g. the fingers, in the example in Figure 1) [4], this landmark proved difficult to reliably identify in video data due to the limitations of the video frame rate. Namely, the point of maximum extension

of a given articulator, e.g., the fingertips, may occur between frames. Thus, we instead used the point at which the hands displayed peak velocity of movement, which corresponded to the largest visualized change in position of the articulator between video frames, often observed by coders as an increase in blurriness between two frames. Furthermore, ELAN only allows for the annotation of intervals, rather than single points in time. As a result, each apex was annotated as an interval, and the first time-point (T1) of the manually-coded apex was taken as the true timing of the apex for the purpose of calculations in this study.

For each participant in our study, two trained researchers annotated the data independently and then compared annotations to resolve discrepancies. Researchers coded the data and resolved discrepancies in coding with the audio muted to avoid any auditory bias. Next, a consensus round was conducted with an expanded set of coders to resolve any remaining discrepancies. This approach maximized consistency in coding, though the manual approach is still prone to subjective bias.

2.3. Semi-automated method

Semi-automated coding was performed using OpenPose, a software tool developed by the Perceptual Computing Lab at Carnegie Mellon University [1]. Pre-processing and analysis were conducted in ELAN and R [8], incorporating elements from the workflow developed by Pouw and Trujillo [2]. This multi-faceted approach allowed us to effectively analyze and interpret the data obtained from the computational coding process. First, each video was run through the OpenPose software to identify 25 articulators, or keypoints, and track the X and Y positions of the keypoint relative to the resolution of the video. For this study, we isolated the movement of the right index finger (all subjects were right-handed) with a video resolution of 1280x720 pixels. OpenPose also assigns a confidence value (between 0 and 1) that indicates the probability of the model's tracking accuracy. Due to low confidence values (<.1), we excluded data for the fourth participant, where poor lighting likely led to errors in tracking. Figure 2 illustrates the keypoint identification in OpenPose for one participant.

Next, we created a time series of all keypoints produced by OpenPose and aligned the time series with the video data based on the video frame rate (FPS). This time series was then used to align the coordinates with the sound file and calculate the speed of the articulator. The speed of the articulator

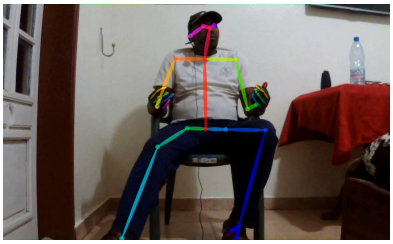


Figure 2: OpenPose Motion-Tracking

was then smoothed using a Butterworth low-pass filter (frequency = 30 Hz). Finally, the time series was aligned with gesture annotations in ELAN. The manually-coded strokes were used to identify the apex, where the apex was the point of maximum speed within a stroke based on coordinate data.

2.4. Method of comparison

In order to assess the similarities between our hand-annotated gestures and OpenPose annotations, we analyzed the consistency in apex coding between the two methods. We used "peak speed timing" as the apex landmark for the semi-automated method and the start time (T1) of the hand-annotated apex as the comparative landmark. Using these apex landmarks, we calculated two measures of apex timing. First, we calculated the relative time of each apex within the stroke, i.e., the T1 of the stroke minus either the OpenPose apex or the T1 of the manually-coded apex. Second, we calculated the difference between the two apex measures, i.e., the OpenPose apex minus the T1 of the manual apex. These two measures were used to analyze the similarities of apex timing across the two coding methods, as is discussed in Section 3.1.

We also compared the alignment of gesture apexes with vowels in the accompanying speech across the two methods in order to determine the comparability of the gesture annotation methods in relation to speech timing. Apex alignment with phones was calculated as the T1 of the manually-coded apex minus the T1 of the nearest vowel, for manual gesture coding, or the point of maximum speed minus the T1 of the nearest vowel, for semi-automated gesture coding. We then compared the relative apex and vowel alignment of the two methods, as discussed in section 3.2.

3. RESULTS

3.1. Agreement in apex alignment

Overall, our results show a close alignment between hand-annotated and semi-automatically annotated gesture apexes. Figure 3 shows a Bland-Altman

plot characterizing agreement between manual and OpenPose-based measures of apex timing. The y-axis in the plot indexes the difference in timing measurements for apexes within each manually-coded stroke, i.e., the time between the manually-coded apex minus the timing of the corresponding OpenPose estimated apex. The x-axis plots the average of the time points between the two apex measures. The two red dotted lines represent +/- 2 standard deviations from the mean. Results indicate the average difference ('bias') in measurements is around 40 ms—just over 1 video frame—between the two measures, with manually-coded apexes marked slightly later than OpenPose apexes.

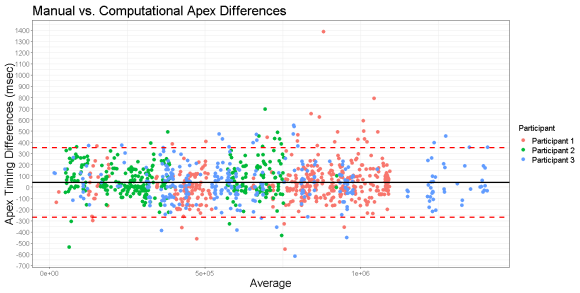


Figure 3: Bland-Altman of timing differences between apex measures

For the comparison between apexes across the two methods, 87.5% of the apexes fell within a two-frame (60ms) distance between each other, and 47% of the apexes fell within one frame (30ms) of each other. Larger differences in timing are due to either a tracking issue within OpenPose, differences in manual coder interpretation, or a combination of these two factors. For example, coders sometimes annotated longer apexes than was standard in our process when the apex was difficult to discern, resulting in greater timing differences between annotation methods.

3.2. Agreement in apex and phone alignment

Since phonetic studies of gesture timing focus on the relative timing between apexes and segments in the speech signal [6, 9, 10, 11, 12], we also examined the alignment of manual and computational apexes with vowels; the results are illustrated in Figure 4.

We analyzed the time between apexes and vowels for both methods and found that the manual method showed an average time-to-vowel of 287 ms, while the computational method showed an average time-to-vowel of 300 ms, an average difference of 13 ms. This measure matches our observation stated in Section 3.1, where the manual apexes tended to occur earlier than the semi-automated

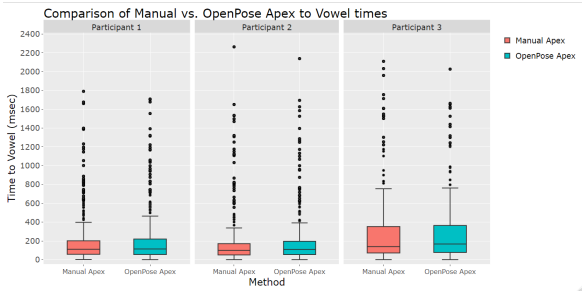


Figure 4: Comparison of apex annotation method times to vowel

apexes. Figure 4 illustrates this trend across all three participants, confirming the regularity of the pattern between manually and semi-automatically coded apexes. Thus, while differences between the two methods are minimal and relate predominantly to framerate, the two methods differ in a consistent and therefore predictable way. Overall, our results demonstrate high comparability between apexes coded using the two methods.

4. FURTHER AUTOMATION OF GESTURAL PHASE ANNOTATION

Currently, this method allows us to automatically code gesture apexes based on *manually*-coded gesture strokes. Thus, further automating gestural annotation to further reduce manual labor would require a method for automatically identifying stroke intervals. Articulatory kinematics have been used to estimate gesture onsets in speech articluation data (EMA) [13], and a similar approach could be applied in the case of co-speech gestures. However, the level of variability in articulatory parameters such as hand shape, orientation, and gesture location pose challenges to this approach. Related to this, the velocity profiles for different gestures can look quite different. For example, Figures 5 and 6 provide speed profiles for two different gestures, the first a bimanual combined depictive-beat gesture (where the hands are moving downward and outlining the shape of an item being described), and the second, a single-handed beat gesture with the hand extending outward and downward. Figure 5 shows a gradual decline in speed, a sharp spike in speed before the apex, into gradual decline. In contrast, Figure 6 shows a gradual rise and fall over the course of the gesture. In comparison, the speed profile for Figure 7 shows a cyclic gesture —where the speaker moves his two hands in a fluid, circular motion—which is very similar to the speed profile of the second beat gesture shown in Figure 6. The gestural speed profile shown in these three figures is only a sample of the degree of variation across

gestures and demonstrates just how difficult it is to characterize the stroke of a co-speech gesture from the speed profile alone.

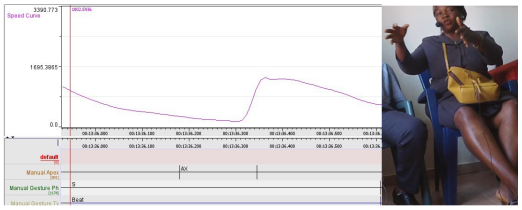


Figure 5: Speed profile of beat-depictive gesture

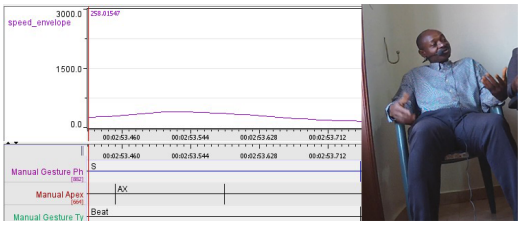


Figure 6: Speed profile of beat gesture

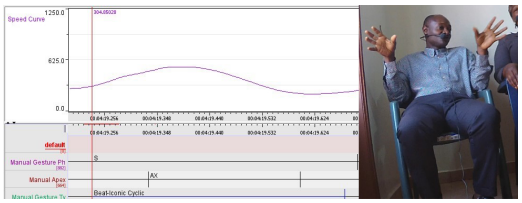


Figure 7: Speed profile of depictive cyclic gesture

Another challenge of relying exclusively on kinematic data to code stroke intervals is that not all hand movements are gestures; i.e., speakers move their hands to fidget, adjust clothing, etc. Human annotators can easily distinguish gestures from fidgets due to the difference in form and intention behind gestures; however, computationally defining what constitutes a fidget is more challenging.

Ultimately, the analysis of stroke speed profiles uncovers the necessity of manual annotations, for identifying stroke boundaries. Although automated processes can effectively identify gestural landmarks where they are quantitatively defined, manual annotators are better able to identify more qualitative aspects of the video. While further investigation of gesture types and forms and their corresponding speed profiles may yield a typical speed profile for a given gesture type and form, and may even distinguish gestures from fidgets, additional research on this topic is necessary in order to further automate markerless tracking methodology. Thus, this study illustrates the similarities between manually-coded and computationally coded apexes and presents areas for further progress in order to fully automate gestural annotation.

5. DISCUSSION

In the absence of marker-based motion tracking technology, two main methods are used to annotate co-speech gestures in video data: manual annotation and markerless tracking, e.g., through the use of OpenPose software. While manual annotation has traditionally been used, markerless tracking technology offers a potential means to automate aspects of the annotation process, improving efficiency. This study compares manually-coded gesture apexes with those coded semi-automatically using OpenPose, using a corpus of conversations between four Medumba speakers.

Our results demonstrate close alignment between apexes coded using the two methods, validating the landmark identification of both methods and the comparability of data across annotation methods. Our analysis demonstrates that the vast majority of apexes (87.5%) fall within two frames of one another. A comparison of the alignment between vowels and apexes likewise shows a high degree of similarity, with manually-coded apexes occurring slightly earlier (13ms) than those using the semi-automated coding process. Thus, while differences between the two methods are minimal, those that do arise are consistent and predictable.

While markerless tracking software offers an efficient approach to annotating gesture apexes, our approach still relies on manual coding for some tasks. For example, both semi-automated and manual coding methods rely on manually defined stroke boundaries, based on which apex timing is determined. Further research on the kinematic characteristics of stroke phases is needed to enable automation of this process. Machine learning techniques provide a promising avenue for classifying different gesture types in terms of their kinematic profiles.

Overall, semi-automated methods of gesture annotation offer a reliable and efficient means of streamlining co-speech gesture research. Automatic apex detection yields similar results to manual apex coding with human coders (with some small, predictable differences), indicating that data annotated with the two methods is comparable. Therefore, semi-automated annotation is a valid method in co-speech gesture timing research. This study contributes to the field by demonstrating a way to expand co-speech gesture corpora without using marker-based motion capture technology.

6. ACKNOWLEDGMENTS

The authors would like to thank the Medumba speakers who participated in the study and to Dr. Ange Bergson Lendja for his assistance in coordinating the study. The authors would also like to express their sincere gratitude to the University of Delaware PhonLab team for their invaluable contributions and unwavering commitment to this study. The collective efforts of Crystal Akalu, Christine Belden, Kylie Boggs, Nicole Curnan, Luc DiNardi, Walter Dych, Bryant Graybeal, Lindsay Hawtof, Jacquelyn Janocha, Brett Lindsey, Zoe Lipkin, Madison Miller, Meghan O'Brien, Victoria Ochlan, Anna Schumeyer, Nicole Taylor, Ta-tran Tran, and Juliette Winnard in processing and analyzing data have been instrumental in achieving the results presented in this paper. Their expertise, dedication, and camaraderie have greatly enriched our research and fostered an environment conducive to innovation and discovery.

This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are our own.

7. REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," 2018. [Online]. Available: https://arxiv.org/abs/1812.08008

[2] W. Pouw and J. Trujillo, "Materials tutorial gespin2019 - using video-based motion tracking to quantify speech-gesture synchrony,." [Online]. Available: osf.io/rxb8j

[3] Max Planck Institute for Psycholinguistics, "Elan." [Online]. Available: https://archive.mpi.nl/tla/elan/download

[4] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in The Relationship of Verbal and Nonverbal Communication. DE GRUYTER MOUTON, Dec. 1980, pp. 207–228.

[5] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements," Journal of Speech, Language, and Hearing Research, vol. 56, no. 3, pp. 850–864, 2013. [Online]. Available: https://doi.org/10.1044/1092-4388(2012/12-0049)

[6] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," Laboratory Phonology, vol. 3, no. 1, 2012. [Online]. Available: https://doi.org/10.1515/lp-2012-0006

[7] A. B. Hostetter and M. W. Alibali, "Visible embodiment: Gestures as simulated action," Psychonomic Bulletin & Review, vol. 15, no. 3, pp. 495–514, 2008.

[8] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[9] K. Franich and H. Keupdjio, "The Influence of Tone on the Alignment of Speech and Co-Speech Gesture," in Proc. Speech Prosody 2022, 2022, pp. 307–311.

[10] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," Speech Communication, vol. 57, pp. 209–232, 2014.

[11] I. Biazzo, C. Canestrari, and E. Demuru, "Gesture-prosody synchrony: Evidence from italian," Gesture, vol. 14, no. 2, pp. 171–193, 2014.

[12] A. Watanabe and Y. Hirose, "Correlates between prosodic features and manual gesture in japanese spontaneous speech," Language and Speech, vol. 58, no. 2, pp. 225–243, 2015.

[13] M. Tiede, "Mview: Multi-channel visualization application for displaying dynamic sensor movements," development, 2010.

DISCREPANT COMBINATIONS OF ACOUSTIC AND VISUAL SPEECH SIGNALS AND ITS RELATION TO ACOUSTIC DISTANCE IN VOWELS

Risa Matsubara[1], Terumichi Ariga[1], Takeshi Kishiyama[1], Chuyu Huang[2]

The University of Tokyo[1], Nagoya Gakuin University[2]
matsubara-risa761@g.ecc.u-tokyo.ac.jp, ariga@phiz.c.u-tokyo.ac.jp, kishiyama.t@gmail.com, huang@ngu.ac.jp

ABSTRACT

This study aims to investigate two research questions: (1). whether and how the acoustic characteristics affect the audiovisual perception of Estonian back-mid unrounded vowel [ɤ], which posits between [ø] and [o] in terms of acoustic characteristics, and (2). if the listeners visually judge the front or back of the vowel. The accuracy rate of the audiovisual ABX task was significantly lower for the combination of the audio information of [ɤ] with the visual information of [ø] or [o] but higher when the auditorily presented vowel and the visually presented vowel were opposing one another in terms of the backness (e.g., audio: [o] and visual: [ø]). The results suggest that the perceptual illusion of vowels is more likely to occur when the acoustic gap between the auditorily presented vowel and the visually presented vowel is small, even if the listeners have acoustically similar vowels in their native language.

Keywords: audiovisual speech perception, McGurk effect, Estonian, vowel perception, lipreading.

1. INTRODUCTION

Humans perceive speech sounds not only through audio information but also through visual information. This phenomenon is known as audiovisual integration, and many studies have examined the role of visual information in perceiving speech sounds [1–4]. Listeners utilize visual information in speech perception because it provides information about articulatory motions that can act as a cue to identify speech segments [1]. When audio speech perception is difficult, listeners attend more to talkers' mouth [2, 3]. Visual information, therefore, helps listeners to identify speech perception in adverse conditions.

However, audiovisual integration in speech perception can cause listeners to perceive illusional sounds that are not actually presented in an audio mode. A famous phenomenon is the McGurk effect, in which perception is distorted by the gap between visual and audio information [4]. For example, when the audio of the sound [ba] is combined with a video where a person is pronouncing [ga], the resulting information is perceived as [da].

The same phenomenon is also true when it comes to vowel perception. For example, Traummüller & Öhrström [5] reported that the combination of the audio information of an unrounded front mid vowel [e:] and the visual information of a rounded front high vowel [y:] was perceived as [ø:], a rounded front-mid vowel, by native speakers of Swedish (see Engstrand [6] for the vowel inventory of Swedish). They concluded that audio information (e.g., [e]: [–round][mid]) provides information about height while visual information (e.g., [y]: [+round][high]) offers information about the roundedness of the vowel. Accordingly, the results imply that both audio and visual information affect the perception of vowels, while roundedness is likely to be visually perceived.

However, it is possible that the listeners depend solely on the audio information when judging the roundedness of the vowel. Lisker and Rossi [7] instructed French listeners to judge the roundedness of the vowel of a video in which the audio information was discrepant with the visual information in terms of the roundedness. The results suggested that the audio information of rounded vowels (e.g., [ø]) was more likely to be perceived as [+round] than the visual information of rounded vowels. This was interpreted to mean that the listeners can shift to one input mode despite the discrepancy between the audio information and visual information. Additionally, a higher misperception rate of [ɨ] or [ɯ] as [+round] when presented in combination with the image of rounded vowels was shown, which implies that central or back unrounded vowels are close to rounded vowels in terms of their acoustic quality. In fact, [ɨ] has a closer acoustic value to that of [y] [7]. However, what should be noted here is that French listeners do not have an inventory of unrounded central or back vowels although they have front and back rounded vowels such as [y], [u], or [ø] [8]. Therefore, the reason that they perceived

the unrounded vowels [ɨ] or [ɯ] as [+round] may be the result of the assimilation of the unfamiliar vowels into their familiar category of rounded vowels based on acoustic and perceptual information. This study aims to test if such misperception of unrounded vowels as rounded ones in audiovisual perception still occurs when the listeners have both rounded front and back vowels and an unrounded back vowel at the same height in their vowel inventory.

Estonian has an unrounded back vowel [ɤ] as well as a rounded front vowel [ø] and a rounded back vowel [o], both of which are at the same height as the [ɤ]. The vowel [ø] has a higher second formant (F2) value, which represents frontness, while the vowel [o] has a lower F2 value, reflecting backness. Both [ø] and [o] have a close first formant (F1) value, indicating their similarity in height.

The vowel [ɤ] acoustically and perceptually sits between the rounded front [ø] and the rounded back [o] (Figure 1). This means that [ɤ] can be perceived as [+round] due to its acoustic characteristics when combined with the visual information that provides [+round]. However, if having the two acoustically similar but articulatory different vowels in their native language enables the listeners to focus on the acoustic information regardless of the acoustic similarity, [ɤ] will be correctly judged even when combined with [+round] vowel. Additionally, whether the listeners can visually perceive the backness of the vowel is unclear. In the case of Estonian, [ø] and [o] are the same in terms of the roundedness, but they differ in terms of backness (Table 1). If the backness is visually perceived as well as the roundedness, the listeners will perceive audio [o] as [ø] when combined with the visual information of [ø]. Alternatively, if the listeners depend more on the audio information when judging the backness due to the larger acoustic distance between the auditorily presented vowel and the visually presented vowel, they will correctly perceive the audio [o] even when combined with [ø].

Table 1: The features of the Estonian vowels [ø], [ɤ], and [o].

	round	back
ø	+	−
ɤ	−	+
o	+	+

Thus, this study investigated whether the visual information of [+round] affects the perception of the acoustically vague vowel and whether the listeners visually judge the backness. To test the research questions, we conducted an audiovisual

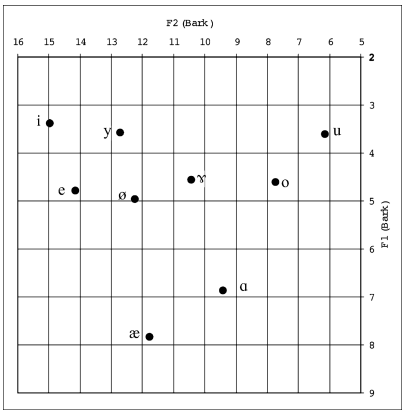


Figure 1: The acoustic and perceptual space of Estonian vowels (cited from Asu and Teras [9]).

ABX experiment, where the listeners needed to judge which third stimulus (X) was identical to either of the first two, A or B. If the listeners make more errors in their judgement, it means that they have difficulty determining the phonetic value of the stimulus. There may be an increase in their reaction times (RTs), which reflects the difficulty of their judgement as well.

2. METHOD

2.1. Participants

Thirty-two native speakers of Estonians were recruited using the website Prolific [10]. They indicated that they were raised only using the Estonian language.

2.2. Materials

Four Estonian vowels [e, o, ø, ɤ] were pronounced by three native speakers of Estonian (two males and one female; mean age = 39 years). They were recorded via Zoom [11] with a sampling rate of 32.0 kHz. They were instructed to pronounce the vowels in the same natural pitch pattern. The audio and visual information was separated and then combined by using iMovie to produce the videos in which the audio information corresponded with the visual information or did not. The acoustic information of the vowels they produced is shown in Table 2.

To avoid the effect of acoustic differences other than vocalic differences, the absolute value of amplitude was set to 0.7 after modification of the intensity to 70 dB in Praat [12].

There were two items, as in Table 3 and 4. In Video A, the audio vowel [ɤ] was combined with either the visual information of [o] or [ø], the acoustically close vowels, to observe if the

Table 2: Acoustic information of the vowels (Hz). S1 = Speaker 1, S2 = Speaker 2, and S3 =Speaker 3. M is an abbreviation of "male" and F is an abbreviation of "female". F1 = first formant, F2 = second formant, F3 = third formant.

	Formant	S1 (M)	S2 (M)	S3 (F)
e	F1	428.1	402.9	373.8
	F2	2068.4	2299.6	2632.5
	F3	2390.2	2860.6	2968.3
o	F1	491.6	416.32	418.9
	F2	636.3	628.9	608.4
	F3	2244.1	2894.9	2807.7
ø	F1	443.2	419.1	432.3
	F2	1497.3	1475.8	2192
	F3	2150	2383.2	2756.6
ɤ	F1	429.59	398.1	434.2
	F2	1307.6	1067.9	1083.6
	F3	2261.3	2496.9	2964.6

perception is influenced by the visual [+round] information. Video B was a control in which there was no incongruency between audio and visual information.

Table 3: An example of items (audio [ø]).

	Video A	Video B
Experimental	audio [ɤ] + visual [ø]	audio [ø] + visual [ø]
Backness	audio [o] + visual [ø]	audio [ø] + visual [ø]
Control	audio [ɤ] + visual [ɤ]	audio [ø] + visual [ø]

Table 4: An example of items (audio [ɤ]).

	Video A	Video B
Experimental	audio [ɤ] + visual [o]	audio [o] + visual [o]
Backness	audio [ø] + visual [o]	audio [o] + visual [o]
Control	audio [ɤ] + visual [ɤ]	audio [o] + visual [o]

The order of the stimuli and the speakers was counterbalanced to avoid the effect of the difference in the presentation order of the two videos. The number of the experimental stimuli, therefore, totaled 48 (3: conditions x 2: items x 2: speaker identity x 2: order x 2: X identity). Speaker 3 was assigned to the sound X in all stimuli while the speaker of Video A or B were either Speaker 1 or Speaker 2.

2.3. Procedure

The experiment was created on PCIbex [13]. The participants remotely took part in the experiment by clicking the link. They listened to the materials with headphones or earphones at a comfortable volume.

The task was an audiovisual ABX task in which stimuli X were presented only auditorily. The two videos, A and B, were played in a row and only the sound that was identical to either A or B was played at the end. The participants were required to judge which sound was identical to the last sound X by pressing the A key or the B key. The maximum time to answer was 4000 ms. The 48 stimuli were presented randomly along with another 48 filler stimuli. The total time needed was 30 minutes, and all of the participants were paid for their participation.

2.4. Analysis

Before proceeding to the statistical analysis, data on two participants were excluded from the data because of their extremely lower accuracy rate when compared to other participants (55% and 65% respectively). Judging by the histogram, the tokens whose reaction times (RTs) were no more than 150 ms or more than 3000 ms were omitted from the data. After the data trimmed, the data that exceeded $\pm 2.5SD$ when based on the log-transformed RTs were excluded.

The statistical analysis involved the generalized linear mixed-effects models (GLMM) and linear mixed-effects (LME) models. The GLMM was used for the analysis of the accuracy rate and the LME model was used for the analysis of the RTs.

In the GLMM analysis, the participants' answers (correct or wrong) were set as a response variable, the condition (Experimental, Backness, or Control) was set as a fixed variable, and the individual differences in the items and participants were set as random effects. In the LME analysis, RTs were set as a response variable, the condition (Experimental, Backness, or Control) was set as a fixed variable, and the individual differences in the items and participants were set as random effects. The best-fit model was selected by using backward selection [14].

3. RESULTS

The mean accuracy rates per condition are shown in Table 5.

Table 5: The mean accuracy rate (%).

	accuracy (%)
Control	90.4
Experimental	85.3
Backness	93.6

The GLMM analysis found a significant difference between the Experimental and the Control conditions and a marginally significant

difference between the Backness and Control conditions (Table 6).

Table 6: The statistical results of the accuracy rate.

	β	*SE*	z value	Pr(>|z|)	
(Intercept)	2.5205	0.3838	6.567	< 5.14e–11	***
Experimental	–0.5496	0.2134	–2.576	0.010	**
Backness	0.4525	0.2506	1.806	0.071	.

However, there were no significant differences in the RTs between the conditions (Table 7).

Table 7: The statistical results of the RTs.

	β	*SE*	*df*	*t*	*p*	
(Intercept)	621.99	35.04	22.90	17.750	< 7.02e–15	***
Experimental	24.66	29.13	1307.42	0.846	0.397	
Backness	−29.41	28.83	1307.12	−1.020	0.308	

4. DISCUSSION

The present study tested these two hypotheses. The first hypothesis was that the visual [+round] information changes the perception of a [−round] vowel to [+round] when they have similar acoustic characteristics even if both are included in the vowel inventory of the listeners' native language. The significantly lower accuracy rate in the Experimental condition can be interpreted to mean that the Estonian listeners perceived the [ɤ] as [+round] and supports the hypothesis that the audio perception of the vowel is distorted by the visual information when the acoustic information of vowels is close between the audio and visual materials. The observation that the unrounded vowel was perceived as a rounded vowel affected by the visual roundedness is in alignment with the results of Traummüller & Öhrström's [5] study. Moreover, the results show that having both an unrounded back vowel and rounded front and back vowels in the listeners' native language does not necessarily guarantee resistance to the effect of discrepant visual information of the rounded vowel on the perception of the auditorily presented unrounded back vowel, which is acoustically close to the rounded vowel.

The other hypothesis was that the listeners visually judge the backness regardless of the acoustic distance between the audio and visual information. However, the accuracy rate was not significantly different between the Backness and Control conditions, thus providing no support for this hypothesis. In fact, the accuracy rate was higher in the Backness condition than in the Control condition, although the difference was marginally significant. This implies that the effect of the visual information is eliminated and the listeners can correctly auditorily perceive the vowel when the acoustic difference is large between the audio and visual vowels and between the first and second stimuli. Consequently, the results weakly reject the hypothesis that the backness of vowels is visually judged.

Overall, this current study implied that the audiovisual illusion of the vowels was not language-specific. The audiovisual illusion occurred not only in Swedish or French, which has rounded vowels at both back and front, but also in Estonian, which has an unrounded back vowel in addition to the rounded front and back vowels. Also, the results add to those of Navarra and Soto-Faraco [15], who reported that visual information facilitates the perception of nonnative vowels, by demonstrating that the visual information of a discrepant vowel can negatively affect the perception of native vowels.

5. REFERENCES

[1] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, no. 6, pp. 431–461, 1967.

[2] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception and Psychophysics*, vol. 60, no. 6, pp. 926–940, 1998.

[3] J. Birulés, L. Bosch, F. Pons, and D. J. Lewkowicz, "Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech," *Language, Cognition and Neuroscience*, vol. 35, no. 10, pp. 1314–1325, 2020.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] H. Traummüller and N. Öhrström, "Audiovisual perception of openness and lip rounding in front vowels," *Journal of Phonetics*, vol. 35, no. 2, pp. 244–258, 2007.

[6] O. Engstrand, "Swedish," *Journal of the International Phonetic Association*, vol. 20, no. 1, pp. 42–44, 1990.

[7] L. Lisker and M. Rossi, "Auditory and visual cueing of the [±rounded] feature of vowels," *Language and Speech*, vol. 35, no. 4, pp. 391–417, 1992.

[8] B. Collins and I. M. Mees, *Practical phonetics and phonology: A resource book for students*. Routledge, 2013.

[9] E. Asu and P. Teras, "Estonian," *Journal of the International Phonetic Association*, vol. 39, no. 3, pp. 367–372, 2009.

[10] "Prolific," https://www.prolific.co/, visited 6-Jan-23.

[11] "Zoom: One platform to connect," https://zoom.us/, visited 6-Jan-23.

[12] "Praat: Doing phonetics by computer," https://www.praat.org/, visited 6-Jan-23.
[13] "PCIbex," https://www.pcibex.net/, visited 6-Jan-23.
[14] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," *arXiv preprint arXiv:1506.04967*., 2015.
[15] J. Navarra and S. Soto-Faraco, "Hearing lips in a second language: visual articulatory information enables the perception of second language sounds," *Psychological Research*, vol. 71, pp. 4–12, 2007.

GESTURES-PROSODY (A)SYNCHRONY: TEMPORAL INTERACTION BETWEEN HEAD AND EYEBROW PEAKS IN STATEMENTS AND YES-NO QUESTIONS

Marisa Cruz & Sónia Frota

Center of Linguistics, School of Arts and Humanities, University of Lisbon
marisac@campus.ul.pt, sfrota@edu.ulisboa.pt

ABSTRACT

Gestures have been shown to be temporally linked to prosody, as strokes are synchronous to prominent syllables and pitch accents. However, their precise temporal coordination is still under debate. Exploring kinematics using a 2D motion analysis software, we inspected the temporal interaction between head, eyebrows and prosody in statements and yes-no questions in European Portuguese. We found that head peaks mainly align with the nuclear prosodic word, with different anchoring sites depending on sentence type: before the stressed syllable in statements; within the stressed syllable in yes-no questions. Eyebrow movement, relevant in conveying interrogativity, mainly precedes head peaks, and occurs before the nuclear word. In line with Loehr's suggestion [1], we conclude that *synchrony* cannot be interpreted as strict co-occurrence (not even between concurrent gestures). Instead, eyebrow and head peaks are motorically coordinated with intonation, and with each other, on a regular (but not concurrent) timing distribution.

Keywords: kinematics, temporal coordination, eyebrow peaks, head peaks, intonation.

1. INTRODUCTION

The term synchrony is generally used to refer to the temporal interaction between gestures and speech, from a phonological, semantic, and pragmatic point of view (see McNeill's rules in [2], providing evidence for a cognitive interdependence between the two modalities). However, several studies suggest that the time coordination between gestures and speech does not necessarily mean a strict synchrony or simultaneity in production [1, 3, 4]. Additionally, it has been shown that misaligned gestures up to 600ms before or after their original anchoring timepoint in the speech chain are still semantically integrated from a perception viewpoint [5]. It has also been suggested that the temporal coordination between gestures and speech is constrained by gesture type [6] and function [7]. For instance, eyebrow movements were shown to align with prominent syllables [8, 9, 10, 11, 12, 13], thus leading to a multimodal increase in prosodic prominence. Similar results were found from a developmental perspective, as deictic (pointing) gestures were shown to be timely coordinated with prominent syllables and pitch accents [14]. However, the way the two modalities relate to each other, i.e., in parallel *(hand-in-hand hypothesis*) or complementarily (*trade-off relation*) [15, 16 *versus* 17, 18, *inter alia*], and how they temporally interact is still under discussion.

Our main goal is to contribute to this debate, by exploring the kinematics of head and eyebrow movements along the production of statements and yes-no questions in European Portuguese (EP). Specifically, we analysed the temporal alignment of head and eyebrow peaks relative to the nuclear prosodic word (NPW) in both sentence types. When peaks occurred within the NPW time-window, we further observed their exact location relative to the stressed syllable. In addition, peak amplitude was also inspected in order to establish whether it differed across sentence types and depending on the time-alignment to the NPW.

2. METHODOLOGY

2.1. Materials

We used the audiovisual data of three female native speakers of Standard EP, aged between 20-45 years-old, videotaped in a sound attenuated laboratory with a JVC video camera (model GY-HM11E), while performing a Discourse Completion Task [19, 20], adapted for EP within the project *InAPoP* [21]. For the analysis, 11 neutral statements and 34 neutral yes-no questions were selected, which respectively exhibit the falling H+L* L% and falling-rising H+L* LH% nuclear contours [e.g., 22, 23]), as well as different visual cues: head falling movement in statements *versus* head falling and eyebrow raising movements in yes-no questions [e.g., 24]). Additionally, we also inspected a small set of less frequent yes-no questions produced by the same speakers: 10 yes-no questions with the same nuclear contour but involving head movement only (without eyebrow raising), and 7 yes-no questions exhibiting a different melodic pattern (H*+L L%, hereafter the *other contour*), but involving the two visual cues [25].

This small set allowed us to inspect whether the alignment of head and eyebrow peaks varied depending on the amount of visual cues in the signal and/or on the type of intonational cues for the same sentence type and pragmatic meaning. Thus, in total, the kinematics of 62 utterances was analysed, following the procedures detailed below.

2.2. Kinematic analysis

The kinematic analysis was previously performed in [25], using *Kinovea*, a free 2D motion analysis software [26]. For each video a marker was added to one of the eyebrows and to the chin of the speaker to track the vertical displacement of the two visual cues in pixels (px), along the absolute time series (ms). These data were extracted into .csv files. After the inspection of the kinematic curves, 5 sentences (2 yes-no questions with eyebrow, 1 yes-no question without eyebrow, and 2 yes-no questions *other contour*) were excluded from the analysis because, unlike all the others, they exhibited a rising head curve instead of a falling one.

We then identified in the time axis of each sentence the start and end points of the NPW, as well as the start and end points of the respective stressed syllable (Fig. 1).

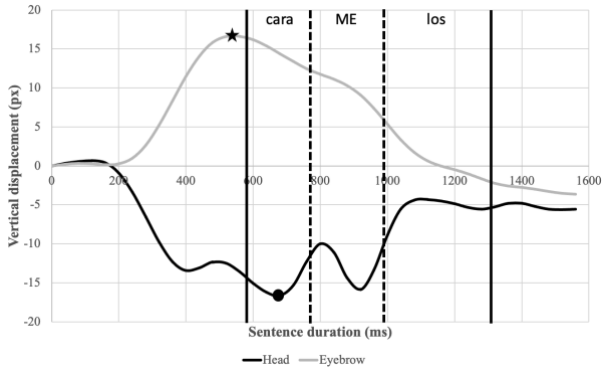


Figure 1: Vertical displacement (px) of the eyebrows (grey line) and head (black line) in the yes-no question *Querem caramelos?* ('Do you want candies?'). The NPW is marked by vertical solid lines; the stressed syllable is marked by dashed lines. The star marks the eyebrow peak, and the dot marks the head peak.

The amplitude peaks of both visual cues were identified and manually coded as occurring before, within or after the NPW. When within the NPW, their occurrence relative to the stressed syllable was coded. For the eyebrow peaks, we also annotated their placement considering the head peak – before, aligned with, and after. Results on the head and eyebrow peaks temporal alignment are described in section 3.1.

Since the vertical displacement (px) of eyebrows may be biased by the head falling movement, namely in utterances without the eyebrow raising cue, we visually inspected the kinematics of each production and applied a correction in the cases in which the eyebrow curve mirrors the head curve – the eyebrow peak amplitude minus the head peak amplitude at the same timepoint – to set the corrected amplitude for this visual cue (Fig. 2). When the eyebrow curve reflects a rising curve, opposite to the head curve (Fig. 1), the eyebrow peak amplitude is kept unchanged.

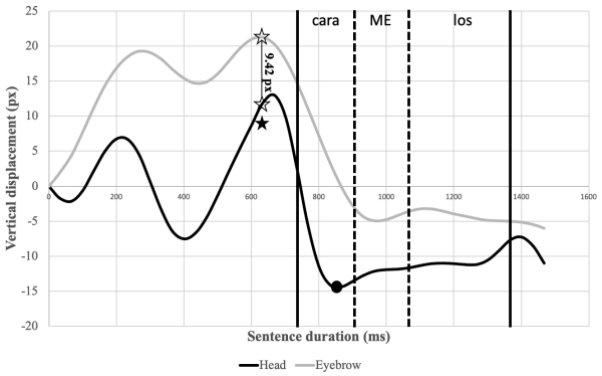


Figure 2: Vertical displacement (px) of the eyebrows (grey line) and head (black line) in the statement *Querem caramelos.* ('(They) want candies.'). The NPW is marked by vertical solid lines; the stressed syllable is marked by dashed lines. The dot marks the head peak, and the stars the corrected eyebrow peak, reset from 21.24px to 9.42px.

Finally, we also examined whether sentence type and peak alignment relative to the NPW had any effect on peak amplitude (see section 3.2 for the results).

2.3. Statistical analysis

Two generalized linear mixed models (GLMM) – one per visual cue – were run, with the amplitude peak (px) as the dependent variable. Fixed effects were sentence type (statement, yes-no questions with eyebrow, yes-no questions without eyebrow, and yes-no questions *other contour*), peak alignment relative to the NPW (before, within, after), and the interaction peak alignment*sentence type. We applied the Satterthwaite method to control for unbalanced data, and a robust estimation was used to test fixed effects and coefficients in order to handle violations of model assumptions. Statistical procedures were performed in SPSS (IBM SPSS Statistics, version 26.0).

3. RESULTS AND DISCUSSION

3.1. Temporal alignment

3.1.1. Head movement

The head peak mainly occurs within the NPW, independently of sentence type (Fig. 3).

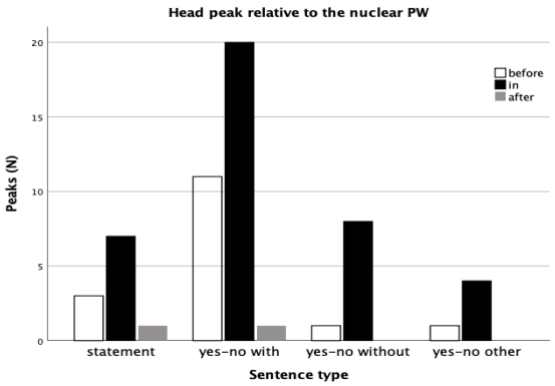


Figure 3: Temporal alignment of the head peak relative to the NPW.

However, the time-alignment of the head peak in relation to the stressed syllable of the NPW varied with sentence type. It precedes the stressed syllable in statements only; in yes-no questions, the peak mainly aligns with the stressed syllable (Fig. 4).

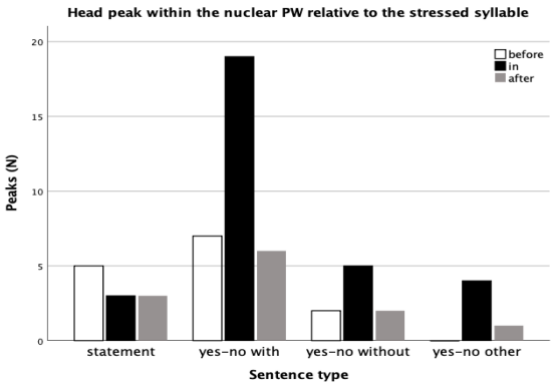


Figure 4: Temporal alignment of the head peak within the NPW relative to the stressed syllable.

3.1.2. Eyebrow movement

In contrast with the head, eyebrow peaks mainly occur before the NPW (Fig. 5).

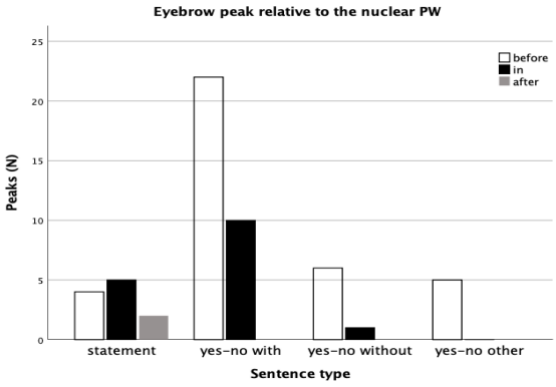


Figure 5: Temporal alignment of the eyebrow peak relative to the NPW.

In statements, eyebrow peaks predominantly occur within the NPW. However, the difference is residual (of one case only). Moreover, the alignment of eyebrow peaks relative to head peaks shows a pattern whereby the eyebrow peak mainly precedes the head peak, even in statements (Fig. 6).

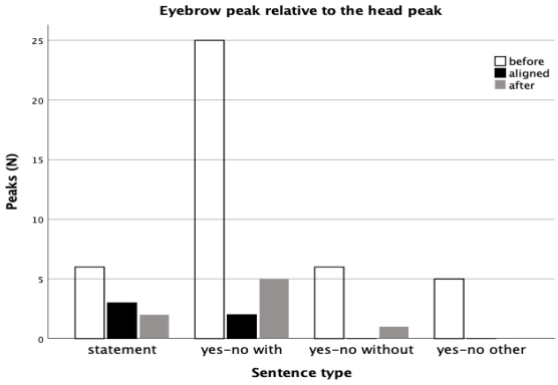


Figure 6: Temporal alignment of the eyebrow peak relative to the head peak.

We also observed the time lapse between the eyebrow and head peaks (excluding the very few cases in which the head peaks occur after the NPW). We concluded that this time-interval is not constant, not even within sentence type: in yes-no questions with eyebrow, it precedes the head peak in 451.5ms (av.), whereas in yes-no *other contour*, also produced with eyebrow raising, the peak is further away from the head peak (648ms, av.). In statements the eyebrow peak is closer to the head peak (222.8ms, av.).

3.2. Peak amplitude

3.2.1. Head peak

The GLMM analysis showed a significant effect of sentence type ($F(3, 47)=13.71$, $p<.001$), of the head peak alignment relative to the NPW ($F(2, 47)= 8.73$, $p<.01$), and of the interaction peak alignment*sentence type ($F(4, 47)=6.76$, $p<.001$).

Pairwise contrasts for the sentence type showed that the head peak amplitude significantly differs between statements and yes-no questions without eyebrow (M=6.03px, -27.89px, respectively; $ß$=33.92, SE=5.94, t=5.71, $p<.001$), thus pointing to the relevance of the head peak amplitude for the distinction between these sentence types. When the eyebrow cue is present, the head peak amplitude does not differ between statements and yes-no questions with eyebrow (M=1.39px) ($ß$=4.64, SE=6.46, t=.72, $p>.05$), but it significantly differs between statements and yes-no questions *other contour* (M=-22.42px) ($ß$=28.45, SE=6.07, t=4.69, $p<.001$). This suggests that the eyebrow peak might be the crucial visual cue for the first sentence type contrast, but not for the latter. Additionally, the head peak amplitude significantly differs between yes-no questions with eyebrow and yes-no questions *other contour* ($ß$=23.81, SE=3.11, t=7.66, $p<.001$), which means

that besides the auditory (intonational) cue, there is also a visual cue distinguishing between these productions. Finally, the head peak amplitude, together with the presence/absence of the eyebrow cue, also plays a relevant role to distinguish between yes-no questions with and without eyebrow movement (yes-no with-without: β=29.28, SE=2.86, t=10.24, p<.001; yes-no without-*other contour*: β=23.81, SE=3.11, t=7.66, p<.001).

Although the head peak mainly aligns with the NPW, the main effect of peak alignment is due to its amplitude being higher when the peak precedes the NPW (M_{before}=-17.35px; M_{in}=-10.86px). However, this amplitude difference was not significant (β=-6.49, SE=4.93, t=-1.32, p>.05).

As for the interaction peak alignment*sentence type, pairwise contrasts for the most frequent head peak alignment position (within the NPW) showed that the head peak amplitude did not differ across yes-no questions (with-without: β=6.68, SE=4.13, t=1.62, p>.05; with-*other contour*: β=8.08, SE=4.79, t=1.69, p>.05; without-*other contour*: β=1.40, SE=3.60, t=.39, p>.05), but differed between statements and yes-no questions without eyebrow (β=25.12, SE=11.03, t=2.28, p<.05), and between statements and yes-no questions *other contour* (β=26.52, SE=11.30, t=2.35, p<.05). The fact that the head peak amplitude within the NPW did not differ between statements and yes-no questions with eyebrow (β=18.44, SE=11.48, t=1.61, p>.05) further supports the claim that the eyebrow peak might be enough for the visual distinction between these sentence types.

3.2.2. Eyebrow peak

The GLMM analysis showed a significant effect of sentence type (F(3, 47)=8.86, p<.001), and of the interaction peak alignment*sentence type (F(2, 47)=3.53, p<.05). The eyebrow peak alignment relative to the NPW, by itself, did not have a significant effect on the eyebrow peak amplitude (F(2, 47)=2.04, p>.05).

Pairwise contrasts for the sentence type showed that the eyebrow peak amplitude did not differ between statements and yes-no questions without eyebrow (M=8.17px, 6.64px, respectively; β=1.54, SE=1.53, t=1.00, p>.05), thus strengthening the relevance of the head peak amplitude for the distinction between these sentence types, as suggested in section 3.2.1. Statements and yes-no questions *other contour* also did not differ (M=17.05px) (β=-8.88, SE=6.08, t=-1.46, p>.05). However, the eyebrow peak amplitude is relevant to distinguish between statements and yes-no questions with eyebrow (M=14.47px) (β=-6.30, SE=2.10, t=-3.01, p<.01). It thus seems that head and eyebrow peaks play a complementary role in marking a given sentence type as different from statements.

Additionally, the eyebrow peak amplitude did not differ between yes-no questions with eyebrow and yes-no questions *other contour* (β=-2.58, SE=6.11, t=-.423, p>.05), which strengthens the relevance of the head peak amplitude for the distinction between these productions, as noted in section 3.2.1. Finally, although the eyebrow (together with the head) peak amplitude plays a relevant role to distinguish between yes-no questions with and without eyebrow (β=7.84, SE=1.62, t=4.84, p<.001), the same does not apply to the contrast between yes-no questions without eyebrow and yes-no questions *other contour* (β=-10.42, SE=5.94, t=-1.75, p>.05), for which the head peak amplitude was more relevant. However, the interaction peak alignment*sentence type showed that eyebrow peak amplitude in the most frequent eyebrow peak alignment position (before the NPW) significantly differed between yes-no questions without eyebrow and yes-no questions *other contour* (β=-14.01, SE=6.01, t=-2.33, p<.05).

4. CONCLUSION

It was found that head and eyebrow peaks are motorically coordinated with intonation, and with each other, on a regular (but not concurrent) timing distribution: head peaks mainly co-occur with the stressed syllable of the NPW, regardless of sentence type; although exhibiting a varying time lapse, eyebrow peaks precede the head peaks and mainly occur before the NPW. This supports Loehr's suggestion [1] that *synchrony* cannot be interpreted as strict co-occurrence (not even between concurrent gestures). Beyond the alignment of these gesture peaks, peak amplitude was found to vary across sentence type for both head and eyebrow, and as a function of the peak position relative to the NPW for head peaks. Moreover, an interaction between peak alignment and sentence type was found for both visual cues. The results point to the complementary role between head and eyebrow, and between gestures and intonation. These results thus provide support to the theory of a *trade-off relation* between gesture and speech production, in which head, eyebrow peaks and intonation convey prominence and meaning in an additive way.

5. ACKNOWLEDGEMENTS

This research was supported by the UIDB/00214/2020 Project, funded by Fundação para a Ciência e a Tecnologia. We also thank all participants for their precious collaboration.

6. REFERENCES

[1] Loehr, D. P. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab. Phonol*. Special Issue Gesture as Language, Gesture and Language 2, 3(1), 71–89. doi: 10.1515/lp-2012-0006

[2] McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

[3] de Ruiter, J.P. 2000. The production of gesture and speech. In: McNeill, D. (eds), *Language and Gesture*. Cambridge University Press, 248–311.

[4] McNeill, D. 2005. *Gesture and Thought*. University of Chicago Press.

[5] Kirchhof, C., de Ruiter, J.P. 2012. "On the audiovisual integration of speech and gesture", presented at the Fifth Conference of the International Society for Gesture Studies, Lund, Sweden, July 24-27, 2012.

[6] Leonard, T., Cummins, F. 2010. The temporal relation between beat gestures and speech. *Lang. Cognitive Proc.* 26(10), 1457–1471. doi: 10.1080/01690965.2010.500218

[7] Wagner, P., Malisz, Z., Kopp, S. 2014. Gesture and speech in interaction: An overview. *Speech Commun*. 57, 209–232. doi: 10.1016/j.specom.2013.09.008

[8] Ekman, P., 1979. About brows: Emotional and conversational signals. In: von Cranach, M., Foppa, K., Lepenies, W., Ploog, D. (eds), *Human Ethology: Claims and Limits of a New Discipline*. Cambridge University Press, 169–202.

[9] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* 57(3), 396–414. doi: 10.1016/j.jml.2007.06.005

[10] Granström, B., House, D. 2007. Inside out – Acoustic and visual aspects of verbal and non-verbal communication. In: Trouvain, J, Barry, W. (eds) *Proc. of the XVIth International Congress of the Phonetic Sciences*, Saarbrücken, Germany, 11–18.

[11] Swerts, M., Krahmer, E. 2008. Facial expression and prosodic prominence: Effects of modality and facial area. *J. Phonetics*, 36(2), 219–238. doi: 10.1016/j.wocn.2007.05.001

[12] al Moubayed, S., Beskow, J., Granström, B., 2010. Auditory visual prominence: From intelligence to behavior. *J. Multimodal User In.* 3(4), 299–309. doi: 10.1007/s12193-010-0054-0

[13] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., Vatikiotis-Bateson, E. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol. Sci.* 15(2), 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

[14] Esteve-Gibert, N., Prieto, P. 2014. Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Commun*. 57, 301–316. doi: 10.1016/j.specom.2013.06.006

[15] So, W.C., Kita, S., Goldin-Meadow, S., 2009. Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Sci.* 33, 115–125. doi:10.1111/j.1551-6709.2008.01006.x

[16]Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., Alías, F. 2014. Gesture synthesis adapted to speech emphasis. *Speech Commun*. 57, 331–350. doi: http://dx.doi.org/10.1016/j.specom.2013.06.005

[17] de Ruiter, J.P., Bangerter, A., Dings, P. 2012. Interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Top. Cogn. Sci.* 4(2), 232–248. doi: 10.1111/j.1756-8765.2012.01183.x

[18] Ferré, G. 2014. A multimodal approach to markedness in spoken French. *Speech Commun*. 57, 268–282. doi: 10.1016/j.specom.2013.06.002

[19] Félix-Brasdefer, J. C. 2009. Data collection methods in speech act performance: DCTs, role plays, and verbal reports. In: Juán, E. U., Martinéz-Flor, A. (eds), *Speech act performance: Theoretical, Empirical, and Methodological Issues*. John Benjamins Publishing, 41–56.

[20] Billmyer, K., Varghese, M. 2000. Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Appl. Linguist.*, 21(4), 517–552. doi: https://doi.org/10.1093/applin/21.4.517

[21] Frota, S. (Coord.). 2012-2015. *Interactive Atlas of the Prosody of Portuguese Project* (PTDC/CLE-LIN/119787/2010), funded by FCT http://labfon.letras.ulisboa.pt/InAPoP/.

[22] Frota, S. 2002. Nuclear falls and rises in European Portuguese: A phonological analysis of declarative and question intonation. *Probus* 14(1) (Special Issue on Intonation in Romance, ed. by Hualde, J. I.), 113–146. doi: 10.1515/prbs.2002.001

[23] Frota, S., Cruz, M., Fernandes-Svartman, F., Collischonn, G., Fonseca, A., Serra, C., Oliveira, P., Vigário, M. 2015. Intonational variation in Portuguese: European and Brazilian varieties. In: Frota, S., Prieto, P. (eds), *Intonation in Romance*. Oxford University Press, 235–283.

[24] Cruz, M., Swerts, M., Frota, S. 2015. Variation in tone and gesture within language. In: The Scottish Consortium for ICPhS 2015 (ed), *Proc. of the 18th International Congress of Phonetic Sciences*, University of Glasgow. Paper 452 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0452.pdf

[25] Cruz, M., Frota, S. Submitted. "Talking heads" in Portuguese. *Language and Cognition.*

[26] *Kinovea 0.9.4. https://www.kinovea.org/*

CAN SEGMENTAL OR SYLLABIC DURATIONS BE PREDICTED BY THE PRESENCE OF CO-SPEECH GESTURES?

Jule Nabrotzky[1], Gilbert Ambrazaitis[1], Margaret Zellers[2], David House[3]

Linnaeus University Växjö[1], Kiel University[2], KTH Royal Institute of Technology[3]
jule.nabrotzky@lnu.se, gilbert.ambrazaitis@lnu.se, mzellers@isfas.uni-kiel.de, davidh@kth.se

ABSTRACT

Building on insights from the coordination of speech and gesture, the cumulative cue hypothesis proposes that speakers recruit cues from both modalities to signal prominence and that the relationship between the cues is additive as opposed to compensative. The study tests if durational variation is one of the cues that are recruited in this cumulative fashion in accord with gestural cues by analysing three five-minute dialogue chunks of spontaneous Swedish. The results do not indicate a direct covariation of this cue with the presence or absence of gestural cues, pointing instead to a need for further investigation of the role of duration in multimodal prominence production and perception.

Keywords: Prominence; multimodal prosody; speech-gesture integration; lexical stress; durational variation

1. INTRODUCTION

The cognitive link between speech and gestural articulation as proposed by McNeill [1] as well as Kendon [2] has been tested extensively and supported with evidence from behavioural [3] and neurological [4] domains as well as evolutionary theories [5], [6]. While the temporal coordination of speech and gesture is well-researched, the spatial coordination, that is the coordination of the magnitude and complexity of speech and gesture, is less understood. There is evidence for a spatial convergence of gesture and speech in elicited speech [7]–[9], as well as read speech in a non-experimental context [10] but there is also counterevidence [11] and the evidence for natural speech is sparse [12].

Building on the cumulative cue hypothesis [10], this paper sets out to test whether the presence of co-speech gestures or gesture clusters co-varies with auditory prominence cues. Specifically, this paper investigates the variation of segmental and syllabic durations, since this parameter has not been tested in connection with spontaneously produced gestures and in the context of spontaneous speech. To investigate this, audio-visual data taken from three dialogues of spontaneous Swedish speech are analysed regarding intonational prominence and gestures, with both manual gestures and eyebrow movements included in the analysis.

2. PREVIOUS RESEARCH

This study is built on previous research from the field of gesture studies, but also informed by knowledge about phonological features of Swedish. The literature review therefore introduces the theoretical model that generated the research question, describes relevant research on multimodal prominence, especially regarding duration and gives background on Swedish phonological characteristics that pertain to prosodic accentuation and durational variation.

2.1. Multimodal prominence

2.1.1 The cumulative cue hypothesis

Ambrazaitis and House [10] posit that speakers recruit both acoustic and visual modalities in a cumulative fashion to create stronger prominence. This means that, as in the acoustic domain, not only one method of creating prominence is used at once, but rather several acoustic and gestural characteristics increase in strength when a word receives higher prominence, which would generally be in line with Gussenhoven's [13] idea of an effort code.

2.1.2 Evidence from previous studies

Evidence for the connection of beat gestures and durations comes from a study on elicited speech in a lab where the participants were specifically asked to produce a sentence including the target words with a manual gesture, a pitch accent, both, or none [7]. The results show that longer durations are associated with both visual beats and the presence of pitch accents. In another study, the participants produced the same syllable while varying an accompanying finger tap or the emphatic stress [8]. Among other effects, the results showed lengthening in both modalities when an emphasis in one modality was present.

The cumulative cue hypothesis has also been tested with data from read speech containing

spontaneously produced head and eyebrow gestures by news reporters [10]. However, while the hypothesis could be supported regarding contours of fundamental frequency, duration was not considered in the analysis. Finally, there is some evidence from a recent study on the way children highlight important information in speech [14], where the presence of head gestures was linked to longer syllable durations.

2.2. Phonological factors affecting duration

This paper investigates the potential predictive value multimodal prominence clusters (MMPs) may have on the duration of the stressed syllable. MMPs are defined by the number and type of co-speech gestures that occur together with a prosodic big accent (see 2.3). However, any variability that can be linked to MMPs will likely not be the most prominent factor in durational variation. Instead, there are several influencing factors on the segmental and word level [15]. One word level factor is the number of segments in the word and the stressed syllable. Even though stressed syllables always have the same moraic weight, the segmental complexity of the syllables can of course vary and influence the length of the individual long segments.

A factor on the segment level that is particularly important for durational variation is whether the long sound in the stressed syllable is a consonant or a vowel [16]. In Swedish, the distinction between stressed and unstressed syllables is quantity-based, with stressed syllables containing one long segment and every stressed syllable consisting of two morae. The long segment can be a vowel or a consonant, but there is an asymmetry in moraic weight between long consonants and vowels, since short vowels always have one mora, while consonants need to be lengthened to be counted as a mora [15].

2.3. Swedish word accents

Even though the f0 contour is not the object of analysis, it is important to quickly spell out the concept of Swedish word accents since they play an important role for prominence in the auditory modality.

In Swedish, each word is assigned one of two contrasting pitch accent contours. The Lund model [17] posits that both accents are two-peaked, with timing being the important distinction. The first peak in accent 1 is earlier than in accent 2 and located in the pre-stressed syllable. Additionally, the Lund model proposes a distinction between two tonal prominence levels, in which the second peaks of the accent contour is only realized in connection with sentence-level prominence [18]. The two-peaked version of each word accent is often called the “big accent”, as opposed to the one-peaked “small accent” [19]. Since this paper strictly focuses of the potential accentual lengthening associated with the presence of accompanying gestures, rather than on the well-established accentual lengthening induced by the intonational prominence level (see 2.4), all words that are tested in this data are words that were realised with a big accent.

2.4 Swedish big accent lengthening

While it is well-documented that accented words are lengthened as opposed to non-accented words in stress-timed languages, the domain of that lengthening can differ between languages [20]. Additionally, Swedish accentual lengthening is better described as a big accent lengthening [16], since even words without a sentence-level prominence are assigned a word accent (as described in section 2.3).

For Swedish, experimental studies have shown that the amount of lengthening of entire words is dependent on speaker, word accent type and the position in the phrase [16], [21]. Most of the lengthening, about 75 % of the total lengthening, is confined to the stressed syllable and lengthening raises the proportion of long to short segment duration. Additionally, the consonants before and after a long vowel are lengthened, while the short vowel before a long consonant is not lengthened [16], which adds to the asymmetries between stressed consonants and vowels.

3. MATERIALS AND PROCEDURE

3.1 Materials

The materials consist of three five-minute chunks of conversation, each one between two speakers of Stockholm Swedish (in total six different speakers, three women and men each) from the Spontal corpus [22]. The recordings were filmed with two cameras and recorded with separate microphones for each speaker, making annotations sections that contain overlap between the speakers possible. The three dialogues were balanced regarding the speaker combination, with two being same-gender and one being mixed-gender.

3.2 Segmentation and annotation procedure

3.2.1 Segmental annotations

The segmental annotations of the material were done in Praat [23] and proceeded in two phases. First, all words with a big accent were annotated in accordance with [10]. In the second phase, the selected words were annotated on the syllabic and segmental level.

The syllabic annotation was conducted to find the stressed syllable and make the later annotation of onset and rhyme possible. At the segmental level, all phonemes of the word were annotated as true to the standardized pronunciation as possible, resulting in a broad annotation with the goal to be as consistent as possible while not misrepresenting the actual articulation of the speakers.

To minimize the possibility that the individual judgement of the annotators confounds the results, segmental boundaries that were unclear were set following consistent rules. Ambisyllabic consonants were always fully included in the first syllable and the boundaries between overlapping segments were annotated midway between the onset of the first and the offset of the last. If there was no consistent way the segments could be annotated (e.g., the onset of a plosive after a pause), this was marked in the data and the words were excluded in the analysis.

3.2.2 Gestural annotations

Both manual and eyebrow gestures were annotated in Elan [24] for all speakers in preparation for the analysis. Manual gestures were annotated in two phases: In the first phase, the presence and absence of manual gestures was annotated, including non-speech gestures such as scratching. This first annotation was refined through an annotation of gesture phrases in the second phase in accordance with [25]. This means that gestures were divided into preparations, strokes, retractions and holds. Beat gestures were separated into "toward" and "away" in relation to the apex and only the part of the gesture where the hand(s) moved toward the apex were included in the category of "stroke". Both hands were annotated separately but are not distinguished in the analysis, thus the presence of a gesture could mean that one hand is moving, or both.

For eyebrow movements, a fixed interval was annotated every time an eyebrow movement was visible in the video [26]. In the case of eyebrows, two researchers conducted the annotation separately and their results were compared. The annotations showed a good inter-rater reliability of κ=.698. For the eyebrow movements, only consensus annotations were included in the data.

4. RESULTS

4.1 Descriptive statistics

We identified 656 words with big accents in the material, 234 of which had to be excluded from the analysis because their segmental boundaries could not be determined in a consistent manner. 260 of the remaining 422 words were not accompanied by any gesture, while 154 were accompanied by one gesture, with manual gestures being much more common (n=123) than eyebrow gestures (n=31). Eight words in the material were accompanied by both a manual and an eyebrow gesture. The script and data files that were used in the analysis can be found in the OSF project file [27].

Before accounting for word and sentence level factors and explaining the variation in the data, a first look on the segment durations (see Fig. 1) does not reveal a clear trend. It is important to mention that both the words with eyebrow gestures and the words with manual and eyebrow gestures have a relatively low sample size.

Lexical-prosodic factors could also interact with how gestures and duration covary [10]. Fig. 2 shows the results for the duration of segments by word accent as well as multi-modal prominence cluster.

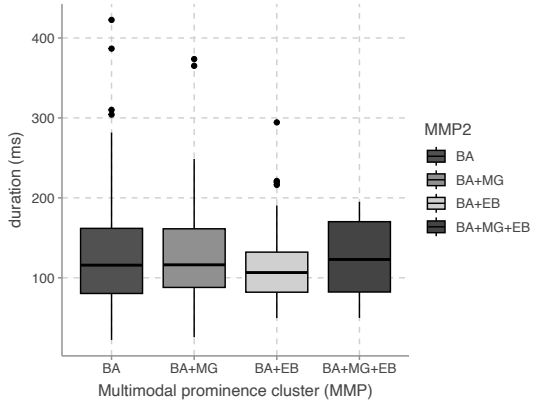


Figure 1 – The durations of weighted segments as a function of multimodal prominence clusters

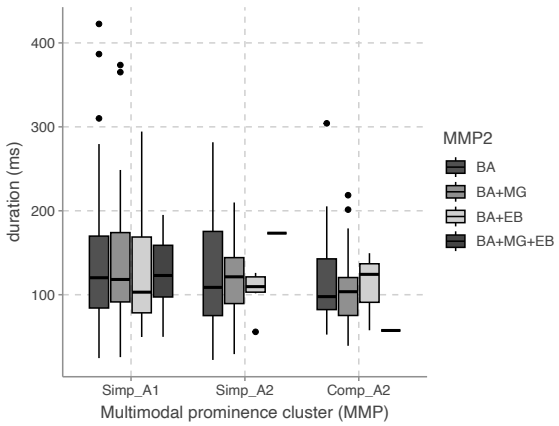


Figure 2 - Segmental durations as a function of Multimodal prominence cluster and lexical prosody

4.2 Inferential statistics

To assess the variability of duration when co-speech gestures are present, several linear mixed models were fitted to the data. Both the durations of the stressed segments, as well as the rhyme and syllable durations were tested as the dependent variables for duration. Additionally, several different

combinations of predictors (multimodal prominence clusters (MMP), type of lexical prosody (LexPro) and syllable complexity (SyllCom)) for each dependent variable were tried to understand how the duration of the words in the data varied systematically. Table 1 shows the model fits with the marginal and conditional R^2-values. All models included additional fixed factors for word duration and segment type (vowel or consonant) and a random factor for speaker.

The inclusion of word duration as a fixed factor served two functions: on the one hand, it was used to account for speech rate (in combination with the random factor for speaker), and on the other hand it was also used to control for the effect that the length of the word and the number of individual segments it consists of has on the duration of each segment, which is generally believed to be inversely correlated. Segment type (consonant or vowel) was included as a fixed factor following from the moraic imbalances discussed above in sections 2.2 and 2.4.

Dep. variable	Model	R^2m	R^2c
Segment	MMP*LexPro+SyllCom	.211	.216
	MMP+LexPro+SyllCom	.208	.213
	LexPro+SyllCom	.205	.211
Rhyme	MMP*LexPro+SyllCom	.367	.367
	MMP+LexPro+SyllCom	.364	.364
	LexPro+SyllCom	.361	.361
Syllable	MMP*LexPro+SyllCom	.546	.548
	MMP+LexPro+SyllCom	.542	.544
	LexPro+SyllCom	.536	.539

Table 1 - Model fit for the segment, rhyme, and syllable level

Generally, the R^2-values indicate that the best model fits are achieved for syllable duration. However, this is to be expected since the models account for word durations. The differences between the R^2m and R^2c values indicate a low relevance of the random effect of speaker. Additionally, we conducted likelihood ratio tests to compare the full models (including the interaction term indicated by '*') with the two sets of reduced models (where either MMP is included, but no interaction is modelled, or where MMP is omitted). These tests did not reveal any significant contribution of MMP (or its interaction with LexPros).

5. DISCUSSION

The results do not support the hypothesis that prosodic prominence – here assessed in terms of accentual lengthening – and the presence of co-speech gestures covary in spontaneous Swedish speech. The likelihood ratio tests show that models that include the factor of multimodal prominence clusters do not have significantly more explanatory power. There was some variation in the results regarding which type of word accent was assigned to the word, but no significant interaction between MMP and LexPros was revealed. There are several factors that need to be addressed regarding the data.

Firstly, there is the issue of low sample sizes for subsets of data. While some subsets had a satisfactory number of tokens, the subsets of interest were rather small, owing to the relative rarity of combined gestures in the data. This might have masked effects that could have been visible with a higher number of tokens and should be considered when interpreting the results.

Another issue is that of speech rate fluctuations. In the mixed models, the word length was used to account for this instead of normalizing speech rate. However, it is not clear that this could account for all variation in speech rate and other approaches to this problem could be explored in future work.

Since these results do not correspond with previous research on this issue, it might be worthwhile to ask which of the factors that differ between the studies could have led to these results. One circumstance that also induced the problems of low sample size and speech rate fluctuations is the fact that we tested spontaneous speech as opposed to lab-elicited speech with no other confounding variables such as topic of conversation or pragmatic meaning, as well as sentence-level factors such as phrase position and rhythmic alternation that were not considered in this study. In previous lab-based investigations of gesture-speech coordination, the segmental material was held constant, simplifying the process of identifying variability in duration. Under the less controlled conditions in our experiment, building mixed models was a greater challenge, which might persist with higher sample sizes.

6. CONCLUSION

The study set out to find new evidence for the cumulative cue hypothesis that proposes the coordination of gestural and intonational prominence in a cumulative, not compensatory fashion. It focused on the aspect of durational prominence on the segmental and syllabic level. However, the data did not offer support for the cumulative cue hypothesis or any covariation of duration and gestural cues. As previously discussed, the spontaneous nature of the speech material could have been a hindrance, which points toward future research that considers a wider variety of pragmatic factors and factors on the sentence level. In addition to a better understanding durational prominence, this could lead to a more sophisticated view on the link of gesture and speech articulation on the sentence level.

7. REFERENCES

[1] D. McNeill, "So you think gestures are nonverbal?," *Psychological Review*, vol. 92, no. 3, pp. 350–371, Jul. 1985.

[2] A. Kendon, "Semiotic diversity in utterance production and the concept of 'language,'" *Phil. Trans. R. Soc. B*, vol. 369, p. 293, Sep. 2014.

[3] M. Graziano and M. Gullberg, "When Speech Stops, Gesture Stops: Evidence From Developmental and Crosslinguistic Comparisons," *Front. Psychol.*, vol. 9, p. 879, Jun. 2018.

[4] R. M. Willems and P. Hagoort, "Neural evidence for the interplay between language, gesture, and action: A review," *Brain and Language*, vol. 101, no. 3, pp. 278–289, Jun. 2007.

[5] M. C. Corballis, "From mouth to hand: Gesture, speech, and the evolution of right-handedness," *Behav. Brain Sci.*, vol. 26, no. 02, Apr. 2003.

[6] O. Capirci and V. Volterra, "Gesture and speech: The emergence and development of a strong and changing partnership," *GEST*, vol. 8, no. 1, pp. 22–44, May 2008.

[7] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, Oct. 2007.

[8] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1–11, Jan. 2014.

[9] W. Pouw, L. Jonge-Hoekstra, S. J. Harrison, A. Paxton, and J. A. Dixon, "Gesture–speech physics in fluent speech and rhythmic upper limb movements," *Ann. N.Y. Acad. Sci.*, vol. 1491, no. 1, pp. 89–105, May 2021.

[10] G. Ambrazaitis and D. House, "Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters," *Laboratory Phonology*, vol. 24, no. 1, Aug. 2022.

[11] B. Roustan and M. Dohen, "Co-production of contrastive prosodic focus and manual gestures : temporal coordination and effects on the acoustic and articulatory correlates of focus," in *Speech Prosody 2010,* Chicago, United States, May 2010, p. 100110:1–4.

[12] S. Berger and M. Zellers, "Multimodal Prominence Marking in Semi-Spontaneous YouTube Monologs: The Interaction of Intonation and Eyebrow Movements," *Front. Commun.*, vol. 7, p. 903015, Jun. 2022.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*, 1st ed. Cambridge University Press, 2004.

[14] N. Esteve-Gibert, H. Lœvenbruck, M. Dohen, and M. D'Imperio, "Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech," *Developmental Science*, vol. 25, no. 1, Jan. 2022.

[15] T. Riad, "Tonal word accents," in *The Phonology of Swedish*, T. Riad, Ed., Oxford University Press, 2013, pp. 181–192.

[16] M. Heldner and E. Strangert, "Temporal effects of focus in Swedish," *Journal of Phonetics*, vol. 29, no. 3, pp. 329–361, Jul. 2001.

[17] G. Bruce and E. Gårding, "A prosodic typology for Swedish dialects," *Travaux de l'Institut de Linguistique de Lund*, vol. 13, pp. 219–228, 1978.

[18] G. Bruce, "Swedish accents in sentence perspective," *Working papers/Lund University, Department of Linguistics and Phonetics*, vol. 12, 1977.

[19] S. Myrberg and T. Riad, "The prosodic hierarchy of Swedish," *Nord J Linguist*, vol. 38, no. 2, pp. 115–147, Oct. 2015.

[20] T. Cambier-Langeveld and A. E. Turk, "A cross-linguistic study of accentual lengthening: Dutch vs. English," *Journal of Phonetics*, vol. 27, no. 3, pp. 255–280, Jul. 1999.

[21] M. Heldner, "On the non-linear lengthening of focally accented Swedish words," in *Nordic Prosody : proceedings of the VIIIth Conference, Trondheim 2000*, Frankfurt am Main: Peter Lang, 2001, pp. 103–112.

[22] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal : a Swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proc. of the Seventh LREC*, KTH, Speech Communication and Technology, 2010, pp. 2992–2995.

[23] P. Boersma and D. Weenink, "Praat : doing phonetics by computer." Jun. 2022. Accessed: Oct. 30, 2022. [Online]. Available: praat.org.

[24] H. Sloetjes and P. Wittenburg, "Annotation by Category: ELAN and ISO DCR," in *Proceedings of the Sixth LREC*, Marrakech, Morocco: ELRA, May 2008.

[25] A. Kendon, *Gesture Visible Action as Utterance*, 1st ed. Cambridge University Press, 2004.

[26] D. House, G. Ambrazaitis, S. Alexanderson, O. Ewald, and A. Kelterer, "Temporal organization of eyebrow beats, head beats and syllables in multimodal signaling of prominence," in *International Conference on Multimodal Communication*, Osnabrück, 2017.

[27] J. Nabrotzky, "Can Segmental or Syllabic Durations Be Predicted by the Presence of Co-Speech Gestures?," *OSF*, Apr. 18, 2023. osf.io/b35ke

IS GESTURE-SPEECH PHYSICS AT WORK IN RHYTHMIC POINTING? EVIDENCE FROM POLISH COUNTING-OUT RHYMES

Šárka Kadavá[1,3,*], Aleksandra Ćwiek[1], Katarzyna Stoltmann[2], Susanne Fuchs[1], Wim Pouw[3]

[1]Leibniz Center General Linguistics, Berlin, Germany [2]adesso SE, Berlin, Germany [3]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands
[*] corresponding author: kadava@leibniz-zas.de

ABSTRACT

'Gesture-speech physics' refers to a possible biomechanical coupling between manual gesture and speech. According to this thesis, rapid gesturing leaves a direct imprint on acoustics (intensity, F0), as gesture accelerations/decelerations increase expiratory forces and therefore subglottal pressure, leading to higher amplitude envelope peaks and higher F0 values. This acoustic effect has been reported in lab experiments, spontaneous speech, clinical studies, and professional vocal performers. The current study investigates this phenomenon in Polish counting-out rhymes, using motion capture data and acoustic recordings from 11 native Polish speakers. Following the gesture-speech physics thesis, we expect acceleration/deceleration peaks to be correlated with speech intensity/F0. Through Bayesian analyses, we obtained a weak but reliable coupling of deceleration of the pointing hand and the nearest peak in the smoothed amplitude envelope.

Keywords: pointing gestures, motion tracking, poetry, prosody, coupling

1. INTRODUCTION

The evidence that gesture-speech coordination on the prosodic level arises out of basic properties of physiology and motor control is increasing [1, 2, 3, 4, 5, 6]. This contrasts with the argument that gesture is a sophisticated cognitive achievement, proliferating due to cultural conventionalization [7].

While not downplaying either of those constraints, according to the gesture-speech physics account (see [6]), there is a biomechanical nudge for aligning peaks in F0 and amplitude envelope with the peak of the physical impulse. As such, the human voice receives an 'imprint' due to the gestural activation of expiration-related muscles. Specifically, upper limb acceleration and deceleration affect rib-cage movement and thus subglottal pressure. That forceful gestures imprint the voice is in line with machine learning studies, showing that neural networks trained on acoustics and body kinematics can come to predict the presence of gesture or kinematic properties of gestures [8, 9, 10].

Although gesture-speech physics seems robust in some tasks, a recent study on leg and arm biking suggests that acceleration may need to reach a certain threshold to affect speech acoustics [11]. This is in line with previous research showing that body parts with lower mass (hand vs. arm) have much weaker effects on speech (e.g., [12]), if at all [13].

The reason why the biomechanical gesture-speech coupling is weak is likely because there must remain the flexibility to speak in certain ways when gesturing. The larynx should indeed be flexible to resist the effect of motion at times it is appropriate to do so. After all, the primary function of the larynx is to act quickly and protect the lungs from inhaling foreign bodies [14].

This study replicates the basic kinematic-acoustic coupling findings from previous research. Our dataset consists of motion data recorded while performing Polish counting-out rhymes involving pointing movement. During a counting-out rhyme game, one person speaks a rhyme while rhythmically moving their index finger between themselves and another person. Having clear turning points, these childhood poems are a valid paradigm to investigate speech-gesture physics, as appreciated by previous studies [15].

We extend previous work by studying forward and backward pointing movement and speech rate as additional factors. So far, only flexion-extension movements have been studied [16] and it is possible that different movements will have different respiratory interactions. Looking at a faster rate is motivated by the fact that this may go hand in hand with larger accelerations/forces. However, the rate may also change the complexity of gesture-speech physics, as different coupling strengths and muscular stiffness are involved.

Following earlier work by amongst

others [16, 17], we expect that (a) higher acceleration/deceleration peaks scale with higher amplitude envelope peaks, and higher F0 values (but less strongly than amplitude), and (b) that the strength of this correlation is dependent on the body motion (forward vs. backward) as it might lead to different effects on airflow due to using antagonistic muscle units. Though we manipulated whether the task was performed with the right or left hand, we ignore this variable as we have no strong hypotheses about handedness in the context of gesture-speech physics.

2. METHODOLOGY

2.1. Experimental set-up and procedures

Participants were informed about the experimental procedures and experimental equipment and signed a consent form. They wore an OptiTrack jacket and a headband; 14 markers were placed on various joints and body parts. The marker relevant to the current analyses was located at the wrist. The participants were instructed to play a counting-out rhyme game with a teddy which was placed on a chair about 1.5 m distance from the participant.

Movements were recorded in 3D space with an OptiTrack system (Motive Version 1.9.0) with 12 cameras (Prime 13), at a 200 Hz sampling rate. Acoustic data were simultaneously recorded with a Sennheiser cardioid microphone, at a 44.1 kHz sampling rate.

Each rhyme was produced in two different speech rates (normal vs. fast) and either the right or left hand for pointing. There was also a control reading condition without any pointing. The conditions were fully crossed, resulting in five different blocks: (1) left hand, normal rate; (2) left hand, fast rate; (3) right hand, normal rate; (4) right hand, fast rate; and (5) reading.

In each block, all rhymes were produced. The order of the blocks was the same for all participants (reading, normal rate, fast rate), but the order of counting-out rhymes was randomized, as well as whether they began with the dominant (right) or non-dominant (left) hand.

2.2. Participants

Eleven Polish native speakers (8 female, 3 male, mean age = 24.1, range = 21–27) took part in the experiment. All participants were right-handed according to the Edinburgh handedness scale [18].

2.3. Data processing

2.3.1. Audio

We applied the Hilbert transform to the sound signals. Taking the complex modulus of the complex-valued transformed signal yields a 1D amplitude envelope of each signal. To smooth the envelopes, we used a Hanning window of 10 Hz. We then downsampled the smoothed envelopes to 200 Hz. To extract the F0 traces of the sound signals, we applied a K. Schaefer-Vincent periodicity detection algorithm (using the `R`-package `wrassp` [19]). Based on sex, the F0 range was limited to 100–450 Hz (female) or 70–300 Hz (male).

2.3.2. Motion tracking

To remove noise-related jitter in motion tracking, a zero-phase 2nd order Butterworth low-pass filter with a cut-off of 30 Hz was applied to the position traces. We also differentiated the signals with respect to time to retrieve the 3D speed, and the next derivative of speed, 3D acceleration. After differentiation, we also apply the same Butterworth filter again. Gesture phases during the ongoing speech were automatically annotated by marking turning points and the movement direction (forward vs. backward).

2.3.3. Aggregation of acoustics and kinematics

Acoustics and kinematics were merged based on their timestamps, and misaligned recordings were approximated using linear interpolation (based on time). Then we resampled all signals at exactly 200 Hz. An example of time series can be seen in Figure 1.

2.3.4. Peak datasets

For the analysis, we constructed a dataset containing all the relevant data points from the time series – namely the peaks from acoustics (envelope, F0) and kinematics (acceleration, deceleration for wrist), using `findpeaks` from the `pracma` `R`-package [20]. The kinematic peaks were extracted for each gesture phase (i.e., movement from one turning point to another). Acoustic peaks were identified that were closest in time to those kinematic peaks. The peaks under study are displayed in Figure 1.

For the final datasets used for modeling, paired kinematic-acoustic peaks that occurred at more than ± 80 ms difference were excluded. This is because peaks that are too far apart are

not possible candidates for mechanical coupling. These estimates are further based on the time of anticipatory or reactionary muscles that happen before or after a deceleration peak [21]. In the datasets with envelope peaks, we omitted those data points in which the envelope peaks were detected either in a pause within the rhyme or in moments without voicing (F0 = 0). Furthermore, some trials needed to be excluded since the wrist marker was not visible and caused tracking problems. Finally, outliers of each variable within each dataset were detected and removed using Tukey's 1.5 IQR rule.

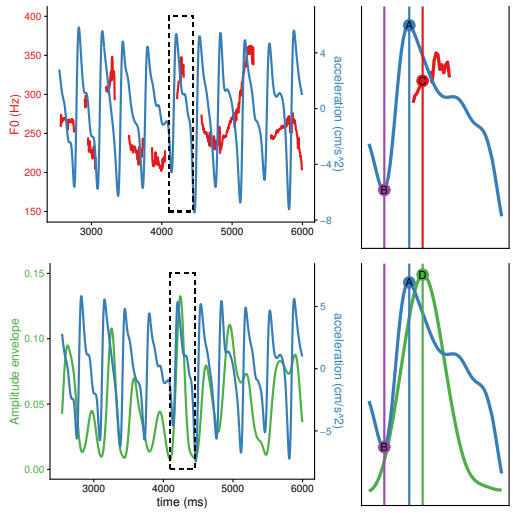


Figure 1: Example of amplitude envelope, F0 and acceleration time series (first 3.4 seconds from the onset of speech). Plots on the right side show enlarged sections with peaks under study: A – acceleration peak, B – deceleration peak, C – F0 peak, D - amplitude envelope peak.

2.4. Data analyses

All analyses were conducted in `R` version 4.2.1 [22]. We used `BayesFactor` [23] to compute Bayes Factor, `brms` [24] and `cmdstanr` [25] for modeling, and `shinystan` [26] for model diagnostics.

Four Bayesian mixed effects models were fitted for the investigated dependent variables: (1) F0 peak nearest to acceleration peak (*F0 acc*), (2) F0 peak nearest to deceleration peak (*F0 dec*), (3) amplitude envelope peak nearest to acceleration peak (*ENV acc*), and (4) amplitude envelope peak nearest to deceleration peak (*ENV dec*).

All models included *speech rate* (fast vs. normal), *movement direction* (forward vs. backward), and *sex* (female vs. male) as contrast-coded predictors. Models for the acceleration (*F0 acc* and *ENV acc*) also included the *wrist acceleration peak* and the interaction of the *wrist acceleration peak* with *movement direction* as predictors. Respectively, models for the deceleration (*F0 dec* and *ENV dec*) included the *wrist deceleration peak* and its interaction with *movement direction*. All models included random intercepts for *speaker* and *rhyme* with random slopes corresponding to all of the factors (except for *sex* in *speaker*). The models included weakly informative priors (i.e., unbiased with respect to H0/H1); the intercept prior for the models was truncated to only include positive values.

Prior to modeling, we controlled for the correlation between speech rate and the respective acceleration or deceleration peak of the wrist in order to assess the possible collinearity of the predictors. *Speech rate* (contrast-coded) and *wrist peak* (acceleration or deceleration, respectively) were correlated in all cases. This, however, did not cause any issues for the model to converge. All processed data, scripts and models are available in the OSF repository: https://osf.io/2abtd/.

3. RESULTS

Table 1 presents the results for parameters that were identified as reliably affecting the outcome variables; for reference, it also includes the intercepts of all models. For complete model outputs see here. Both F0-related models, i.e., *F0 acc* and *F0 dec*, show that *sex* and *speech rate* have an effect on F0 peaks, such that women and faster speech rate exhibit higher F0 peaks. Other parameters did not turn out to be reliable predictors of F0 peaks.

As for the envelope-related models, there were no reliable predictors in the *ENV acc* model. In the *ENV dec*, we found an effect of *wrist deceleration peak* on the amplitude envelope peak ($\beta = -0.02[-0.04; -0.00]$) with the posterior probability $\beta < 0$ equal to 0.99. The magnitude of the effect is small, as per the mean posterior estimate (-0.02). It is, however, stable and reliably distributed, as, given the priors, the data, and the model, there is a 99% posterior probability that there is an effect of wrist deceleration on the amplitude envelope, such that when deceleration (negatively) increases, the nearest envelope peak is higher.

Apart from that, there were two further parameters in the *ENV dec* model, which reached 94% posterior probability (i.e., almost the common cut-off of 95%), but both encompassed 0 in the 95% CrI. These were *speech rate* ($\beta = -0.10[-0.23; 0.03]$) and *movement direction* ($\beta = -0.08[-0.18; 0.02]$). Given our priors, data, and the model, the effect of these parameters

Model	Parameter	Estimate [95% CrI]	Pr($\beta <$ or > 0)
F0 acc	Intercept	185.58[172.59;198.94]	
	Speech rate	8.16[1.70;14.51]	0.99
	Sex	104.07[77.81;129.97]	1.00
F0 dec	Intercept	177.12[162.24;191.94]	
	Speech rate	5.04[0.20;9.90]	0.98
	Sex	100.84[71.03;130.46]	1.00
ENV acc	Intercept	0.82[0.72;0.93]	
ENV dec	Intercept	0.70[0.54;0.87]	
	Wrist dec. peak	−0.02[−0.04;−0.00]	0.99

Table 1: The table lists out all intercepts, as well as the parameters with a reliable effect on the outcome variables of the four models, with posterior means and the 95% CrI. The rightmost column is the posterior probability of the effect to be below or above 0, depending on the direction.

on amplitude deceleration peak is not reliable, however, since the task was not heavily controlled, we acknowledge that those relationships should be further studied in the future. As can be seen in Figure 2, the main effect of *wrist deceleration peak* on the amplitude envelope peak we reported above differs when it is within backward vs. forward movement. Nevertheless, given the priors, the data, and the model, there was no interaction effect of the two parameters ($\beta = -0.00[-0.02;0.01]$).

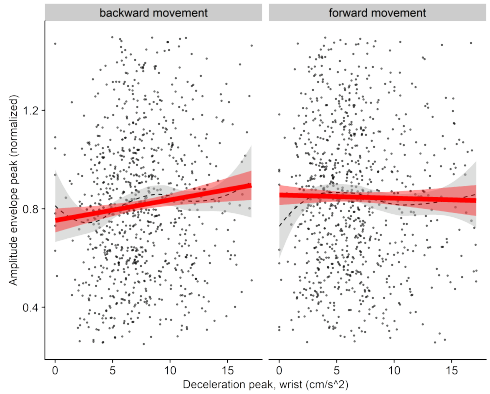


Figure 2: The linear relationship between deceleration peak and amplitude envelope peak. Dashed non-linear 'loess' line reflects possible non-linearities. Note that the deceleration values have been absolutized.

4. DISCUSSION

The current study goes beyond previous research on gesture-speech physics by assessing statistical coupling in (1) multi-directional (2) pointing movements in (3) the Polish language. Our findings suggest that deceleration peaks scale to their nearest amplitude envelope peak, rather than F0. This scaling effect was small, but reliable. Acceleration peaks did not display significant scaling with either envelope or F0 peaks.

Why did the rapidity by which participants halted a pointing movement (i.e., deceleration) not scale to the nearest F0 peak? The gesture-speech physics thesis proposes that there is a mechanical interaction between an upper limb and the body during acceleration or deceleration. The physical impulse of a upper-limb movement produces a mechanical loading onto the rib cage, which limits its movement and impacts subglottal pressures necessary for voice production. Subglottal pressures are primarily linked with affecting intensity, and only secondarily F0, which is under more flexible laryngeal control.

That there is a coupling of deceleration rather than acceleration might look like a counterargument for gesture-speech physics. However, comparing the absolute raw values of deceleration and acceleration peaks, we found 20% lower magnitudes for acceleration than deceleration. In line with [11], we suppose that a certain threshold needs to be reached before a significant effect of physics arises.

As for the deceleration effect alone, it is known that speakers coordinate their emphasis in speech with the moment when the limb movement reaches its destination [27]. Thus, emphasis is generally not located at the initial stage of pointing; rather, it occurs when reaching the intended target.

Further, we did not find that kinematic peaks affected speech differently depending on the direction of movement. This means forward and backward movements along the sagittal plane likely perturb vocalization by increasing subglottal pressure, much like flexion-extension movements along the frontal plane [16].

For future research, potential alternative hypotheses should be investigated. For example, other kinematic variables (e.g., speed) need to be assessed for speech coupling. Our study is also limited in the number of participants, cautioning generalizability, but it has a large number of trials and events that have been analyzed, increasing the reliability of the reported effects within our sample. Moreover, since we do not directly measure muscle activity in relation to respiratory-vocal states, it is always possible to maintain that the current kinematic-acoustic effect is solely a neurally controlled achievement. Such an explanation requires an auxiliary hypothesis about why the brain would monitor acceleration peaks and couple them to vocalization. While we deem it possible that the brain is tuned like this, it would be precisely because there is a weak biomechanical coupling to begin with. Gesture-speech coordination, then, is a 'smart' combination of brain and brawn [6].

5. ACKNOWLEDGEMENTS

This work has been supported by DFG grants FU791/9-1, CW 10/1-1 and PO 2841/1-1. WP is funded by a VENI grant (VI.Veni 0.201G.047: PI Wim Pouw). We would like to thank the participants of the study.

6. REFERENCES

[1] M. Chu and P. Hagoort, "Synchronization of speech and gesture: Evidence for interaction in action," *Journal of Experimental Psychology: General*, vol. 143, no. 3, 2014.

[2] J. M. Iverson and E. Thelen, "Hand, mouth and brain: The dynamic emergence of speech and gesture," *Journal of Consciousness Studies*, vol. 6, no. 11-12, pp. 19–40, 1999.

[3] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1–11, 2014.

[4] G. Zelic, J. Kim, and C. Davis, "Articulatory constraints on spontaneous entrainment between speech and manual gesture," *Human Movement Science*, vol. 42, pp. 232–245, 2015.

[5] W. Pouw and J. A. Dixon, "Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback," *Cognitive Science*, vol. 43, no. 3, 2019.

[6] W. Pouw and S. Fuchs, "Origins of vocal-entangled gesture," *Neuroscience & Biobehavioral Reviews*, vol. 141, p. 104836, 2022.

[7] E. McClave, "Pitch and manual gestures," *Journal of Psycholinguistic Research*, vol. 27, no. 2, pp. 69–89, 1998.

[8] F. Yunus, C. Clavel, and C. Pelachaud, "Sequence-to-sequence predictive model: From prosody to communicative gestures," *arXiv:2008.07643 [cs, eess]*, 2020.

[9] T. Kucherenko, R. Nagy, M. Neff, H. Kjellström, and G. E. Henter, "Multimodal analysis of the predictability of hand-gesture properties," *arXiv:2108.05762 [cs]*, 2022, arXiv: 2108.05762.

[10] Y. Ferstl, M. Neff, and R. McDonnell, "Understanding the Predictability of Gesture Parameters from Speech and their Perceptual Importance," in *Proceedings of the 20th ACM international Conference on Intelligent Virtual Agents*. Scotland, UK: ACM New York, 2020, pp. 1–8.

[11] H. Serré, M. Dohen, S. Fuchs, S. Gerber, and A. Rochet-Capellan, "Leg movements affect speech intensity," *Journal of Neurophysiology*, 2022.

[12] W. Pouw, A. Paxton, S. J. Harrison, and J. A. Dixon, "Acoustic information about upper limb movement in voicing," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, pp. 11 364–11 367, May 2020.

[13] W. Pouw, S. J. Harrison, and J. A. Dixon, "Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony." *Journal of Experimental Psychology: General*, vol. 149, no. 2, pp. 391–404, 2020.

[14] J. J. Ohala, "Respiratory Activity in Speech," in *Speech Production and Speech Modelling*, ser. NATO ASI Series, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Springer Netherlands, 1990, pp. 23–53.

[15] S. Fuchs and U. D. Reichel, "On the relationship between pointing gestures and speech production in German counting out rhymes: Evidence from motion capture data and speech acoustics," in *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, München, Germany, 2016, p. 5.

[16] W. Pouw, L. de Jonge-Hoekstra, S. J. Harrison, A. Paxton, and J. A. Dixon, "Gesture-speech physics in fluent speech and rhythmic upper limb movements," *Annals of the New York Academy of Sciences*, vol. 1491, no. 1, pp. 89–105, 2020.

[17] W. Pouw, S. J. Harrison, N. Esteve-Gibert, and J. A. Dixon, "Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1231–1247, 2020.

[18] R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, 1971.

[19] R. Winkelmann, L. Bombien, M. Scheffers, and M. Jochim, *wrassp: Interface to the 'ASSP' Library*, 2023, R package version 1.0.4.

[20] H. W. Borchers, *pracma: Practical Numerical Math Functions*, 2022.

[21] A. S. Aruin and M. L. Latash, "Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements," *Experimental Brain Research*, vol. 103, no. 2, pp. 323–332, 1995.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[23] R. D. Morey, J. N. Rouder, T. Jamil, S. Urbanek, K. Forner, and A. Ly, *BayesFactor: Computation of Bayes Factors for Common Designs*, 2018, R package version 0.9.12-4.4.

[24] P.-C. Bürkner, "Advanced Bayesian Multilevel Modeling with the R Package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.

[25] J. Gabry and R. Češnovar, *cmdstanr: R Interface to 'CmdStan'*, 2020, R package version 0.5.2.

[26] Stan Development Team, *shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models.*, 2017, R package version 2.4.0.

[27] Esteve-Gibert Núria and Prieto Pilar, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 850–864, 2013.

GESTURE AND PROSODIC PROMINENCE IN AMBONESE INDONESIAN

Alessa Farinella[a], Constantijn Kaland[b] & Daniel Kaufman[c]

[a]University of Massachusetts, Amherst, USA; [b]University of Cologne, Germany; [c]Queens College & The Graduate Center, CUNY, USA
afarinella@umass.edu; ckaland@uni-koeln.de; daniel.kaufman@qc.cuny.edu

ABSTRACT

Studies on co-speech gesture have revealed a strong tendency for manual gesture to be aligned to prosodic prominence in many languages. In this study, we investigate co-speech gesture in a variety of Standard Indonesian as spoken in Ambon, an island in the eastern part of Indonesia. Prior studies on both Indonesian and Ambonese Malay, the local variety of Malay spoken in Ambon, disagree as to whether either of these varieties shows evidence of word stress. The current study provides a new perspective through gesture on the way prosody in these languages is organized. While we find some evidence that gesture-aligned syllables are prosodically prominent, this does not appear to always be the case. These findings suggest that Ambonese Indonesian may possess prosodic structure that is not always acoustically marked, and highlights the importance of collecting evidence from a variety of methodologies, including gesture alignment, in the study of word stress.

Keywords: Gesture, prosody, Indonesian, stresslessness, multimodality

1. INTRODUCTION

1.1. Indonesian prosody

In this study, we examine the alignment of manual gesture to speech in Ambonese Indonesian to shed light on the prosody of Indonesian. Previous studies on word level prosody in Indonesian have come to conflicting conclusions about the status of stress in the language. It has been variously claimed that Indonesian is a language with regular penultimate stress that shifts to the ultimate syllable when the penult contains schwa [1], [2] and that Indonesian has regular ultimate stress [3]. More recent studies argue that Indonesian lacks evidence of stress at the word level entirely [4], [5], [6]. However, the area in which Indonesian is spoken is quite vast and linguistically diverse and, as a result, differences between varieties has been reported. A study by [7], for example, found that the prosody of Indonesian is influenced by a speaker's native language.

This study focuses on the variety of Indonesian spoken in Ambon in eastern Indonesia. While work on the prosody of the languages spoken in this area is lacking, studies on word level prosody have been conducted on Ambonese Malay, a regional variety of Malay. These studies also come to contradictory conclusions. In a grammar of Ambonese Malay, [8] describes stress as being lexical and unpredictable, evidenced by the existence of minimal pairs. However, a more recent acoustic investigation failed to find evidence of word level stress in Ambonese Malay [9]. The only other language in this region for which an acoustic investigation of the stress system has been undertaken is Papuan Malay, a regional Malay variety spoken in close geographic proximity to Ambonese Malay. In contrast to the finding for Ambonese Malay, it was concluded that this variety possesses regular penultimate stress, with a small class of words with ultimate stress [10], [11], [12].

Even for closely related varieties spoken in close geographic proximity, then, we might expect to find prosodic differences. However, there are also clear methodological differences between the acoustic study on Ambonese Malay and that on Papuan Malay. While [9] relied on scripted, lab based speech for Ambonese Malay, the acoustic study in [10] used spontaneous narratives and a much larger amount of data. Differences in methodology may also be in part responsible for the different results for Ambonese and Papuan Malay. Against this conflicting background, we use gesture alignment as a window into the prosody of Indonesian as spoken in Ambon.

1.2. Gesture & prosody

Studies on the temporal alignment of gesture to speech have demonstrated the close coordination of manual gesture and prosodic prominence for a growing number of languages [13], [14], [15]. The ma-

jority of these studies have been conducted on languages with word level stress, and find that the apex, defined as the point of maximal extension of the gesture, tends to align to pitch accents.

However, non-stress and tonal languages appear to differ from canonical stress languages in their patterns of gesture alignment. In French, for which pitch accents have been analyzed as serving a demarcative rather than prominence-marking function, gesture tends broadly to align to pitch accents but less consistently than in stress languages [16]. In Medumba, a tonal Grassfields Bantu language, it has been shown that gesture does not preferentially align with high toned syllables [17]. As high tones mark lexical contrasts in this language, this possibly supports the idea that gesture targets linguistic prominence and not just acoustic salience or pitch accents per se. However, the same study finds that gesture-aligned syllables are longer and louder than in non-aligned syllables, showing that there remains a correspondence between manual and acoustic prominence even though gestures are not anchored to prosodically determined prominent/stressed syllables.

Most relevant for the present work, a study by [18] on gesture alignment in two varieties of Indonesian found that for the variety spoken in eastern Indonesia, gesture tended to align to the penultimate syllable, while for the variety spoken in western Indonesia, gesture tended to align to the ultimate syllable. While it's not yet entirely clear how gesture aligns to speech in this variety, one of the aims of the present work is to uncover patterns in gesture alignment (if any) in this variety. Results of prior studies support the idea that gesture alignment patterns may reflect prosodic properties.

1.3. Research questions

As is apparent from the overview in Section 1.1, there is little agreement as to the status of word stress in Indonesian, and evidence that varieties may differ in their prosody. What is clear is that evidence for word stress should be sought from a variety of sources. Here, we employ a novel methodology of using manual gesture as a unique source of evidence for word level prosody in Indonesian. In light of the findings discussed above that gesture is tightly coordinated with prosodic prominence across a number of languages, the aim of the present study is to use gesture as a window into the word prosody of a variety of Indonesian. In doing so, we ask the following questions: i) is there a systematic tendency to align gesture to a particular syllable in the word and ii) do gesture-aligned syllables also exhibit prosodic prominence?

There are several possibilities we might expect for Ambonese Indonesian:

1. If gesture does not target a particular position in the prosodic word, it would be consistent with the claim that Ambonese Indonesian lacks a word level prominence distinction, on the basis of the findings in [16] for French.
2. If gesture targets a consistent position in the prosodic word *and* the syllable with gesture exhibits prosodic prominence, it would be consistent with the claim that Ambonese Indonesian has word level prosodic prominence.
3. If gesture targets a consistent position in the prosodic word and the gesture-aligned syllable does *not* exhibit prosodic prominence, it would be consistent with the claim that Ambonese Indonesian possesses word prosodic structure that lacks consistent acoustic correlates.

2. METHODOLOGY

2.1. Data and annotation

The data come from video recordings of two preachers (S1, female and S2, male) from Ambon, in eastern Indonesia, who are both speaking Standard Indonesian. In the video excerpts, both are speaking in a spontaneous style.

The apex, or target, of the gesture was annotated in ELAN [19] without audio. While many studies on gesture and prosody include only beat gestures, which are by definition non-referential, no effort was made here to exclude referential gestures, as they have also been shown to be coordinated in time with prosodic prominence [20]. However, only gestures with clear starting and stopping points were included in this study.

The recordings were transcribed in Praat [21] by a native speaker and segmented by syllable on a separate tier by a coder with knowledge of the syllable structure of Indonesian. Words were marked as phrase final or non-final impressionistically using common crosslinguistic correlates to boundaries (pauses, final lengthening, pitch reset). We then confirmed that mean duration by phoneme count was significantly greater for phrase final words than non-final words. Words whose status as final or non-final was unclear were excluded, as were monosyllables.

2.2. Acoustic measures

Duration per phoneme was measured for all syllables in the data set by taking the total syllable dura-

Table 1: Number of syllables aligned to gesture by word position (Position column) and phrase position (Phrase column).

Spkr	Position	Phrase	Frequency
S1	Penult	Medial	115
		Final	41
	Ultimate	Medial	29
		Final	5
S2	Penult	Medial	41
		Final	26
	Ultimate	Medial	14
		Final	10

tion divided by the number of phonemes in the syllable (similar to [10]). This way of measuring duration accounts to some extent for differences in syllable structure. Vowel quality was also measured by taking the frequencies of the first and second formants converted to Bark scale. These acoustic measures were chosen as they were found to be the strongest correlates of word stress in [10] for Papuan Malay.

3. RESULTS

3.1. Gesture alignment

Only penultimate and ultimate syllables were included in the analysis, as prior studies on Indonesian have claimed that either the penult or ultimate syllable (or neither) is prosodically prominent. In total, 521 syllables for S1, 191 of which were aligned to gesture, and 387 syllables for S2, 94 of which were aligned to gesture, were annotated.

Table 1 shows the frequency of gesture aligned to penultimate and ultimate syllables by phrase position. For both speakers, there is a clear tendency to align gesture to penultimate syllables in both phrase final and non phrase final position, though this tendency is slightly stronger for S1 than for S2. A Chi-square test revealed a significant relationship between presence of gesture and syllable position for S1 ($\chi^2 = 127.4, df = 1, p < .001$) and S2 ($\chi^2 = 25.7, df = 1, p < .001$). The fact that alignment is unaffected by phrase position suggests that gesture is targeting a particular position in the word.

3.2. Acoustics

3.2.1. Duration

Our second research question concerns how well gesture lines up with prosodic prominence, as measured by the acoustic correlates discussed in Section 2.2. We separated words that were phrase final from those that were in medial position, as words in final position are expected to be affected by final lengthening. Looking first at the words in phrase medial position, we find that penultimate and ultimate syllables aligned with gesture tend to be longer than those that are unaligned. This difference is larger for S1 than S2.

Table 2: Duration per phoneme (ms)

Spkr	Position	Phrase medial		Phrase final	
		Gest	No gest	Gest	No gest
S1	Penult	104	81	114	82
	Ult	88	66	119	103
S2	Penult	95	83	94	103
	Ult	95	80	174	144

Turning to phrase final words, for S1, gesture-aligned syllables are longer than those that are unaligned, and this difference is smaller for ultimate syllables in phrase final words than non final words. This is presumably due to the effect of final lengthening, which is expected to affect ultimate syllables to a greater degree than penultimate syllables. Support for this comes from the fact that, unlike in phrase medial position, ultimate syllables are on average longer than penultimate syllables.

For S2, phrase final words display a different pattern. While ultimate syllables aligned with gesture are longer than those that are not aligned, penultimate syllables aligned with gesture are in fact *shorter* than those that are not aligned. This is distinct from what was observed for S1, and on the face of it is quite puzzling. We interpret this finding as indicative that the penultimate position is prominent (and therefore targeted by gesture) for this speaker even in the absence of prosodic prominence as signaled by duration. We return to a discussion of these results in Section 4.

3.2.2. Vowel quality

Values for F1 and F2 were converted into Bark and compared for vowels in syllables that are aligned with gesture and vowels in syllables that are not aligned with gesture. Crosslinguistically, there is a tendency for vowels in stressed syllables to be more peripheral than vowels in unstressed syllables.

The results are shown in Figures 1 and 2. Small font letters are vowels aligned to gesture, large font are unaligned vowels. For both speakers, it is indeed the case that vowels in gesture-aligned syllables tend to be more peripheral than vowels in syllables that are not aligned with gesture.

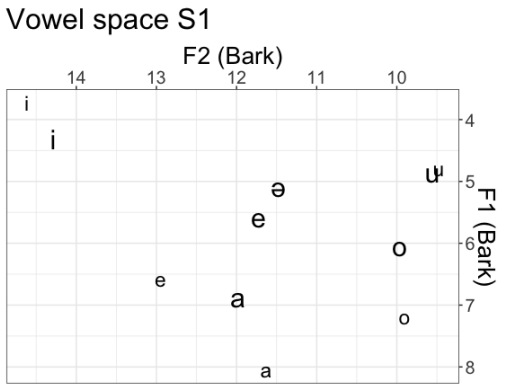


Figure 1: Vowel space for S1. Small font = vowels in syllables aligned to gesture, large font = vowels in unaligned syllables.

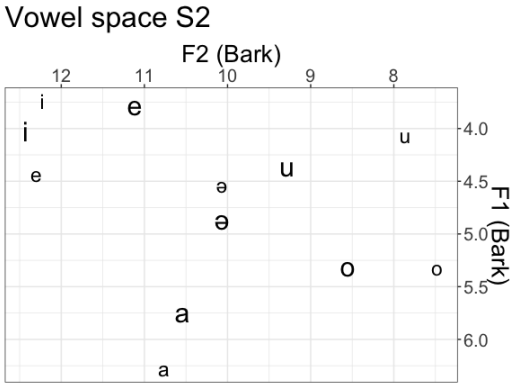


Figure 2: Vowel space for S2. Small font = vowels in syllables aligned to gesture, large font = vowels in unaligned syllables.

4. DISCUSSION

The two speakers in this study showed a strong tendency to align gesture to penultimate syllables in both phrase final and non-final words. Gesture-aligned syllables tended to show prosodic prominence through greater duration compared to syllables not aligned to gesture, and more peripheral vowels. The fact that gesture-aligned syllables tend to exhibit some degree of prosodic prominence *and* that gesture tends to align to penultimate syllables indicates that penultimate syllables are prominent in Ambonese Indonesian.

However for S2, penultimate syllables do not always show evidence of prosodic prominence. For this speaker, it's clear that penultimate syllables are targeted for gesture even in the absence of durational prominence in phrase final position. It's possible that final lengthening on final syllables is an important cue to a boundary, and therefore prevents lengthening in penultimate syllables. However, gesture still targets penultimate syllables at a greater rate than ultimate syllables in phrase final position even in the absence of prosodic prominence.

Of the three possible outcomes in Section 1.3, gesture consistently targets the penultimate syllable, and gesture aligned syllables tend to be longer and have more peripheral vowels than syllables not aligned with gesture. These findings are in line with what has been found for languages with word level stress (and our prediction 2). However, the fact that the syllable aligned to gesture is not always prosodically prominent (for S2) is in line with an explanation of abstract word level prominence (and with our prediction 3). These findings are what would be expected if there's regular penultimate word level prominence that may not be consistently marked acoustically, precisely what was found by [10] for Papuan Malay. This raises the possibility that prosodic structure in this language may differ in its acoustic marking from European languages.

This study leaves open several questions for future investigation. One of these relates to the direction of the relationship between gesture and prosody. While many studies have demonstrated that gesture seeks a candidate that is already prosodically prominent, is it also possible that alignment with gesture triggers acoustic prominence on syllables already bearing abstract prominence. A study by [22] on Dutch found that gesture can enhance acoustic prominence in much the same way as pitch accents. It's possible that gesture selects candidates that bear abstract prominence (which the results of this study would suggest are penultimate syllables) and makes them more acoustically prominent, except perhaps where that would interfere with cues to phrasing (as discussed in the results for S2).

Our findings should not be taken as conclusive evidence that Indonesian as spoken in Ambon has word level stress. However, they are consistent with (mostly) penultimate word level stress, and surprising if this variety has no word level prosodic structure. This is unexpected in light of the study by [9] on Ambonese Malay, which claimed that that variety lacked evidence of stress. While this study looked at Standard Indonesian, it's not clear where prominence distinctions in Ambonese accented Indonesian would originate from if both Standard Indonesian and Ambonese Malay are described as stressless. As it's been shown that regional varieties/languages may influence the word level prosody of Indonesian [7], it's likely that the patterns observed here originate from Ambonese Malay and other local languages. Work on prosody in these varieties should be investigated with a variety of different methodologies, of which we report on just one.

5. ACKNOWLEDGMENTS

This work is funded by the Deutsche Forschungsgemeinschaft (DFG), Project-ID 281511265, CRC 1252 "Prominence in Language."

6. REFERENCES

[1] A. C. Cohn, "Stress in Indonesian and bracketing paradoxes," *Natural Language and Linguistic Theory*, vol. 7, no. 2, pp. 167–216, May 1989.

[2] R. Stoel, "The intonation of Manado Malay," *LOT Occasional Series*, vol. 9, pp. 117–150, 2007.

[3] Samsuri, *Ciri-ciri prosodi dalam kalimat Bahasa Indonesia [Prosodic features in the Indonesian sentence]*, Flores: Nusa Indah, 1971.

[4] C. Odé, V. van Heuven, and E. van Zanten, *Experimental studies of Indonesian prosody*. Vakgroep Talen en Culturen van Zuidoost-Azie en Oceanie, Rijksuniversiteit, 1994, vol. 9.

[5] R. W. Goedemans and E. van Zanten, "No stress typology." Benjamins Amsterdam, 2014, pp. 83–95.

[6] A. Athanasopoulou, I. Vogel, and N. Pincus, "Prosodic prominence in a stressless language: An acoustic investigation of Indonesian," *Journal of Linguistics*, vol. 57, no. 4, pp. 695–735, Nov. 2021.

[7] R. Goedemans and E. van Zanten, "Chapter three Stress and accent in Indonesian," *LOT Occasional series*, vol. 9, pp. 35–62, 2007.

[8] D. Van Minde, *Malayu Ambong: phonology, morphology, syntax*. Leiden: Research School CNWS, 1997.

[9] R. Maskikit-Essed and C. Gussenhoven, "No stress, no pitch accent, no prosodic focus: the case of Ambonese Malay," *Phonology*, vol. 33, no. 2, pp. 353–389, Aug. 2016.

[10] C. Kaland, "Acoustic correlates of word stress in Papuan Malay," *Journal of Phonetics*, vol. 74, pp. 55–74, May 2019.

[11] ——, "Offline and online processing of acoustic cues to word stress in Papuan Malay," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 731–747, Feb. 2020.

[12] C. Kaland, A. Kluge, and V. J. van Heuven, "Lexical analyses of the function and phonology of Papuan Malay word stress," *Phonetica*, vol. 78, no. 2, pp. 141–168, Apr. 2021.

[13] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, Jan. 2012.

[14] S. Jannedy and N. Mendoza-Denton, "Structuring information through gesture and intonation," *Interdisciplinary studies on information structure: ISIS; working papers of the SFB 632*, no. 3, pp. 199–244, 2005.

[15] S. Shattuck-Hufnagel and A. Ren, "The Prosodic Characteristics of Non-referential Co-speech Gestures in a Sample of Academic-Lecture-Style Speech," *Frontiers in Psychology*, vol. 9, p. 1514, Sep. 2018.

[16] P. L. Rohrer, P. Prieto, and E. Delais-Roussarie, "Beat gestures and prosodic domain marking in french," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, VIC, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Canberra, ACT: Australasian Speech Science and Technology Association Inc)*, 2019, pp. 1500–1504.

[17] K. Franich and H. Keupdjio, "The influence of tone on the alignment of speech and co-speech gesture," in *Proceedings of Speech Prosody*, 2022, pp. 2022–63.

[18] D. Kaufman and A. Farinella, "Gesture alignment in a "stressless" language," in *Proceedings of the Twenty-Eighth Meeting of the Austronesian Formal Linguistics Association (AFLA)*, 2022, pp. 29–46.

[19] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *5th international conference on language resources and evaluation (LREC 2006)*, 2006, pp. 1556–1559.

[20] N. Esteve-Gibert and P. Prieto, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 850–864, Jun. 2013.

[21] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[22] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, Oct. 2007.

TOWARDS IDENTIFYING ARTICULATORY GESTURES WITH PIXEL DIFFERENCE AND AUDIO SEGMENTATION

Giulia S. Orlando[1] and Pertti Palo[2]

[1] University of Helsinki, [2] Indiana University
[1] giulia.orlando@helsinki.fi, [2] pertti.palo@taurlin.org

ABSTRACT

We report results from using Pixel Difference – a metric based on Euclidean distance – to analyze articulatory gestures. Our data consists of mono- and disyllabic English words recorded with tongue ultrasound from four speakers. We analyze the shapes of Pixel Difference curves based on acoustic segmentation. The motivation is to see when and how articulatory and acoustic events align. Further, in cases where acoustic segmentation is difficult, e.g. approximants, articulatory events may provide potentially more robust anchor points.

The current study aims to provide a basis to analyze larger datasets. Since this use of Pixel Difference is a novel way to look at speech data, we present a qualitative analysis of time behaviour of Pixel Difference. This will provide a basis for quantitative analysis in the next phase of our project. As the first result, we describe how Pixel Difference curves in VCV sequences depend on the intervening consonant.

Keywords: Acoustic-articulatory analysis, tongue ultrasound, speech gestures, Euclidean distance.

1. INTRODUCTION

The present study investigates the relation between tongue movement and acoustic production. Seeking to find ways to identify speech gestures in ultrasound data in a way similar to how they are found in point tracking data [1, 2], we investigate the behaviour of Pixel Difference (PD) in relation to acoustic segmentation. PD and related metrics have been used for analysing speech data in various ways but not for gesture identification [3, 4, 5].

The analysis starts from the observation of the PD plot, with the aim of mapping the observed shapes and recognizing a set of behavior patterns. On the basis of such patterns, the paper outlines a threefold categorization. The main scope is to observe the degree of reliability of these behaviors and the consistency of the relation between PD shape and sound. Rooted herein, in understanding whether such presumed relation between movement and sound is consistent, is the intention to possibly consider using PD as an aiding tool for acoustic analysis. In closing, the paper also discusses the seeming deviations from the observed patterns, and lays out a structure for a possible generalisation of the relation discussed here.

2. MATERIALS AND METHODS

The data is from a separate study of speech initiation which uses a delayed naming task. The participants were instructed to remain at rest until they heard the go signal – a 1 kHz pure tone – and then produce the target word as soon and as accurately as possible. However, we concentrate on analysing utterance medial phenomena and are not concerned with the original study's questions. For the purpose of this study, the data is just words read out in isolation.

The data was recorded with tongue ultrasound controlled with Articulate Assistant Advanced [6]. Ultrasound was captured at 80 fps, and FOV was 92 degrees with synchronized audio sampled at 44.1 kHz.

The dataset consists of English lexical monosyllabic /[C]V[C]/[1] and disyllabic /[C]VCV[C]/ words produced by L1 speakers of General American English – 1 female, 1 non-binary, 2 males. The words vary onset, medial and coda consonants, as well as vowels, with a total of 180 words in the set. We analyse a subset of these words limited to words where the first and second vowels are phonologically identical. More specifically, we concentrate on the /VCV/ sequence in the disyllabic words and use only /V/ words from the monosyllabic set as an example of a vowel produced in isolation. In the subset we analyze, the medial consonant was one of /b, d, f, g, h, k, l, ɹ, ɾ, m, n, p, r, s, ʃ, t, θ, w, z/ and the vowels were /a, i, u/.

2.1. Acoustic annotation

After generating and populating tiers for utterance, word, phonological segment and phonetic

detail automatically with CAST[2], the data was acoustically segmented in Praat [7] by one of the authors.

2.2. Articulatory analysis

The articulatory analysis is based on a Euclidean metric called Pixel Difference (PD) [5]. PD is defined as the Euclidean distance between consecutive raw (uninterpolated, probe return) ultrasound frames. Expressed as an equation for frames k and $k+1$:

$$(1) \quad PD(k) = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i,j) - im_{k+1}(i,j))^2}$$

where indices i and j iterate over the pixels in x and y direction, $im_k(i,j)$ denotes the pixel in frame k, at row i and column j, and $k = \{1,2,\ldots n_{frames} - 1\}$.

The articulatory analysis was performed with a Python software package called SATKIT [8, 9]. Figure 1 shows an example of data display in SATKIT.

Given the exploratory nature of this study and the limited dataset size, we chose a qualitative approach to analysing the data. Each author went through the data independently and listed their observations. After this, we discussed our findings to identify systematicities that we could agree on.

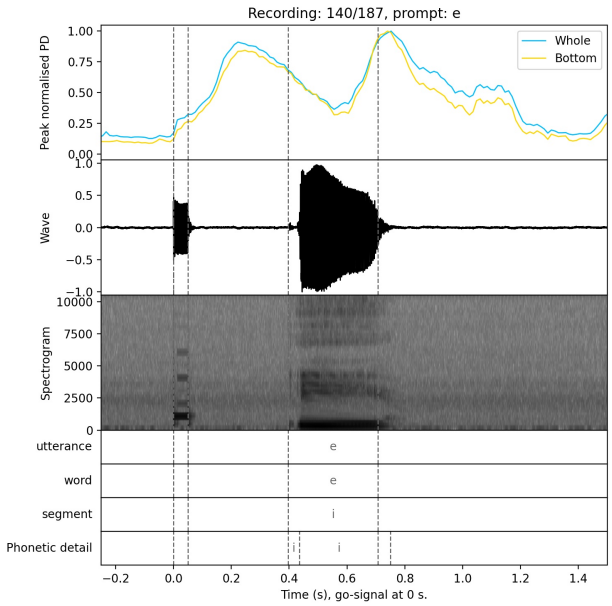


Figure 1: Participant 2 reading 'e', realised as [i]. Pixel Difference (PD) on top shows the typical trough shape of a vowel surrounded by an onset and an offset gesture.

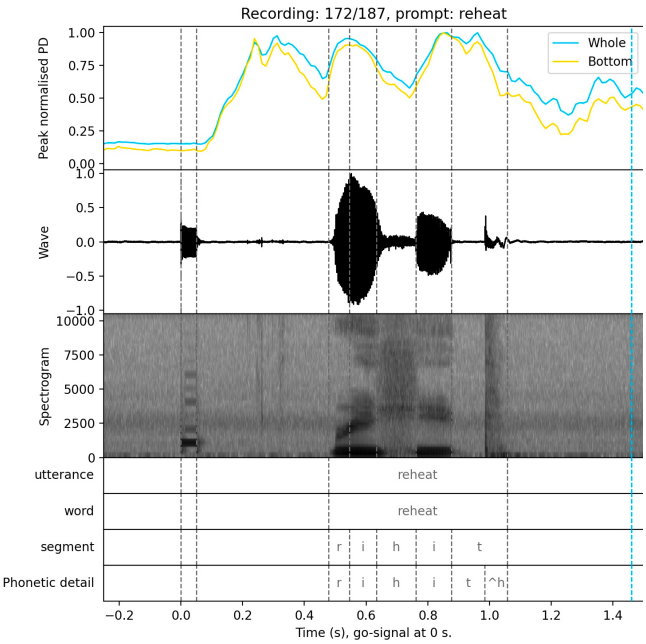
3. RESULTS

The observations of the PD plot seem to show a consistent relation between tongue movement and sound, where the tongue movement appears to follow certain patterns, specific to certain sounds. Specifically, we have been able to identify three categories in PD. Since the present study focuses on /VCV/, the resulting categories are built on the observed deviation in the PD trajectory of what has been identified as the most basic pattern, namely the tongue movement associated with a vowel (Figure 1). The consonants have been grouped to that do not particularly affect the base trajectory (Group 1, Figure 2a), those that generate a gestural peak (Group 2, Figures 2b and 2c), and those whose gestural peak bifurcates in two (Group 3, Figure 2d). Within the categorization proposed, there is an element apparently inconsistently represented in PD, or rather showing a high degree of variability: the allophones represented in American English by the phoneme /h/.

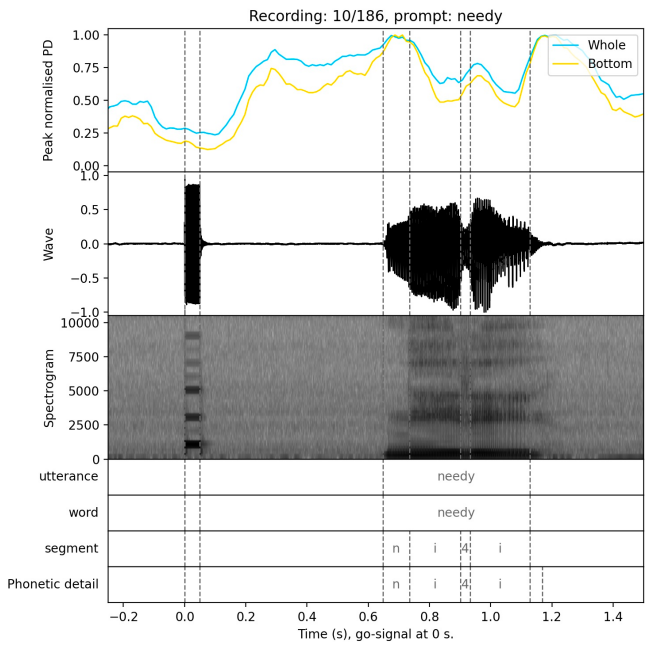
GROUP 0 – vowels A vowel in isolation in PD is shaped as a trough. Initial or final aspiration added to an isolated vowel does not modify the curve observed in PD.

GROUP 1 – glottal fricatives, bilabial nasals They do not affect the vocalic trough or create just a minor disturbance in it, because by their nature they do not require tongue movement, which if happens it is of minor entity. Figure 2 (a) is a good example of how /h/ does not affect the PD trajectory. The PD curve forms a long v-shaped trough over the acoustic /ihi/ sequence with the PD peaks at each end of the sequence aligning very well with the acoustic segment boundaries (beginning of first /i/ and end of second /i/). Similar examples have been found in the dataset for the bilabials nasal, where /m/ does not affect at all the PD trajectory. And even in the instances when /h/ and /m/ do affect the trajectory, it is just a very subtle fluctuation, consisting of a minor gestural peak in the vocalic trough.

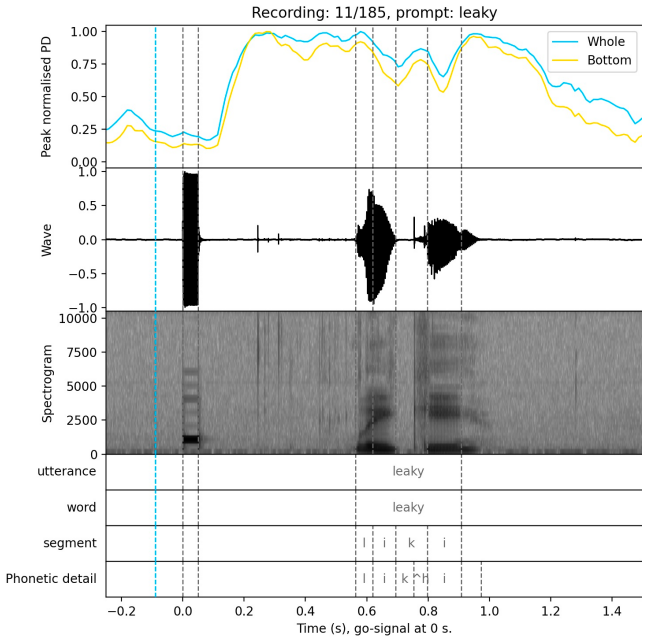
GROUP 2 – plosives Plosives create a gestural peak in the vocalic trough. The amplitude of the peak appears to depend on the point of articulation, with frontal plosives generating smaller gesture peaks, and velars generating bigger ones. The categorization proposed in this paper finds good reasons in grouping all the plosives together, since regardless of their type, all of them produce the same single gestural peak in PD. A good example of the



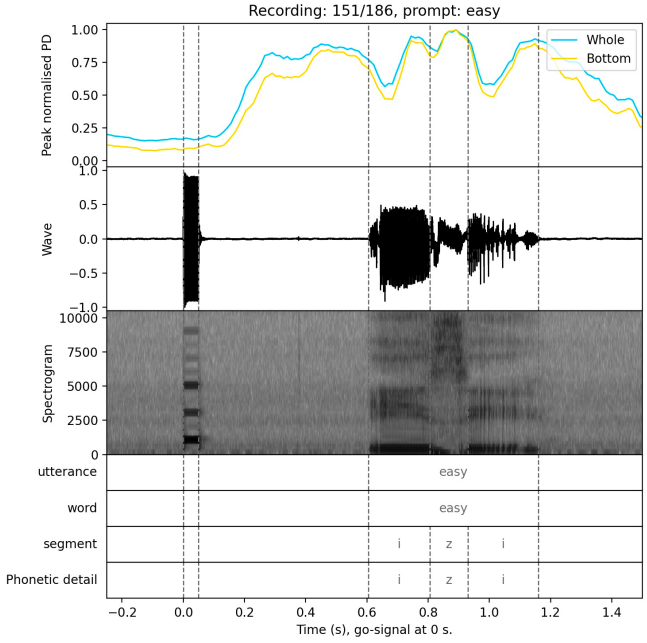
(a) Example of Group 1: Participant 2 reads 'reheat'.



(b) Example 1 of Group 2: Participant 3 reads 'needy'.



(c) Example 2 of Group 2: Participant 4 reads 'leaky'.



(d) Example of Group 3: Participant 3 reads 'easy'.

Figure 2: Examples of Groups 1-4.

plosive peak can be observed in Figures 2b and 2c. PD shows a falling line starting at the onset of the first vowel, and a rising one ending at the offset of the second vowel, with a gestural peak starting at the onset of the velar /k/.

The need to contain a build-up of pressure leading to the plosive release also seems to affect the amplitude of the peak. There were examples in the dataset, where between the falling movement of the first vowel, and the rising one of the second vowel, there is a big gestural peak, when the bilabial is produced. It was also possible to find counter examples of the same plosive without buildup pressure, where the bilabial appears just as a small peak in the vocalic trough.

GROUP 3 – laterals, sibilants, approximants This group of consonants creates a significant

fluctuation in the vocalic trough, namely a bifurcated peak within its trajectory. Figure 2d illustrates this behavior. Here the intervocalic phoneme considered is /z/. PD shows a falling line starting with the onset of the first /i/, then a rising one just when the spectrogram starts to change, then a small drop corresponding to the production of the sibilant, a falling movement ending when the spectrogram changes again, and then another rising movement ending with the offset of the second /i/. Considering /izi/ altogether, this fluctuation can be seen as a double bump placed halfway in the vocalic trough.

Summary of sounds by Group:

Group 0 vowels

Group 1 bilabial nasal /m/

Group 2 mainly plosives /b, d, g, k, ɾ, n, p, t, θ/

Group 3 /l, ɹ, s, ʃ, t, θ, w, z/, possibly also f, ɾ

In more than one group /h/

4. CONCLUSION

The qualitative inspection of data suggests a consistent relation between movement and sound. At this stage, we observed, in intervocalic position, consistency in the behavior of the described categories, which is coherent with the tongue movement involved in their production. Based on the analysis presented in this paper, we believe PD will prove helpful in supporting acoustic segmentation. This seems especially helpful when spectral analysis results a bit unclear (i.e. with laterals, rhotics, approximants). In order to suggest an application of PD analysis for supporting acoustic segmentation, however, further investigation to map the tongue movement patterns in other places, and a larger dataset to generalise our observations are required. In particular, a larger data set is required to be able to quantify the role of peak height in plosive articulations. With the possibility to precisely measure the amplitude, we would be able also to propose a more fine-grained description of the plosives, and thus subcategorise them.

5. DISCUSSION

As observed in this dataset, the behavior of aspiration in PD varies greatly. This seems to find an explanation in the fact that aspiration can be produced with different phonemes, i.e. a glottal fricative (Figure 2a) or pharyngeal/velar which are allophones in American English. In the dataset, it can be seen that glottal fricatives do not affect the vocalic trough due to the nature of their gesture, namely because there is no tongue movement involved. However, when the pronunciation of this intervocalic /h/ is stronger, the difference shows in PD, potentially because the sound produced is actually a different phone. The produced fricative is not a glottal, but a pharyngeal/velar fricative, which involves the tongue in its production, thus unavoidably affects the PD plot. This would need further investigation to account for the specific variation in the tongue movement, but on the basis of the current data, the behavior of aspiration appears to be coherent with the categorization given and does not disprove the patterns outlined.

Another category which showed great variability was the bilabials: in our dataset, we observed that bilabial nasals have slight to no impact on the vocalic trough; as for bilabial plosives, their impact on the PD trajectory was variable, and potentially dependent on the intraoral pressure. In some instances, there was a peak in the PD trajectory in correspondence to the consonant, while in others, the consonant (/b/, /p/) did not perturb the vocalic trough in PD. These findings are related to the electromyography (EMG) trough effect, which is a deactivation of underlying articulator activity during the production of /VCV/ (identical vowels, bilabial consonant) [10, 11]. However, unlike EMG, PD analysis of tongue ultrasound does not differentiate between muscles. This means that it provides a different angle on the phenomenon, and the observed trajectories are not directly comparable.

But since PD analysis of tongue ultrasound does not differentiate between muscles, the results differ to some degree. It would be valuable to collect data with EMG and tongue ultrasound to observe how they relate and obtain a deeper perspective on the muscular activity involved.

From the present analysis, it seems reasonable to assume that the behavior observed in PD plot is correlated with the steadiness of the sound produced. Vowels have a steadier mid-phase, which shows in PD as the trough; while plosives have an acoustical steady state during the closure, in PD they show as just a peak; finally, the rest of the consonants have a steadier phase similar to the vowel's but shorter reflected in PD as a trough which bifurcates the gesture peak in two. This analysis suggests that PD can be used to select the time points when a phoneme reaches its target. For vowels and most consonants this would be the bottom of the PD trough and for plosives the peak of their gesture.

6. REFERENCES

[1] C. P. Browman and L. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140 – 155, 1988.

[2] ——, "Articulatory gestures as phonological units." *Phonology*, vol. 6, pp. 201 – 251, 1990.

[3] C. T. McMillan and M. Corley, "Cascading influences on the production of speech: Evidence from articulation," *Cognition*, vol. 117, no. 3, pp. 243 – 260, 2010.

[4] E. Drake, S. Schaeffler, and M. Corley, "Articulatory evidence for the involvement of the speech production system in the generation of predictions during comprehension," in *Architectures and Mechanisms for Language Processing (AMLaP)*, Marseille, 2013.

[5] P. Palo, "Measuring pre-speech articulation," Ph.D. dissertation, Queen Margaret University, Edinburgh, 2019.

[6] *Articulate Assistant Advanced User Guide: Version 2.14*, Edinburgh, UK: Articulate Instruments Ltd, 2012.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2022, version 6.2.12, retrieved 14 April 2022 from http://www.praat.org/.

[8] P. Palo, S. R. Moisik, and M. Faytak, "SATKIT: Speech Articulation ToolKIT [Python software package]," Available in a public software repository, accessed 31 Aug 2022, https://github.com/giuthas/satkit.

[9] ——, "Analysing speech data with SATKIT," in *International Conference of Phonetic Sciences (ICPhS 2023)*, Prague, 2023.

[10] B. Lindblom, H. Sussman, G. Modarresi, and E. Burlingame, "The trough effect: implications for speech motor programming." *Phonetica*, vol. 59, no. 4, pp. 245 – 262, 2002.

[11] S. Fuchs, P. Hoole, J. Brunner, and M. Inoue, "The trough effect – an aerodynamic phenomenon?" in *From sound to sense: MIT Meeting*, 2004.

[1] Here [C] denotes a consonant being present in some but not all of the words in the indicated position.

[2] https://github.com/giuthas/cast

ANALYSING SPEECH DATA WITH SATKIT

Pertti Palo[1*], Scott R. Moisik[2], and Matthew Faytak[3]

[1]Indiana University, Bloomington, [2]Nanyang Technological University, Singapore,

[3]University at Buffalo, Buffalo, NY

[*]pertti.palo@taurlin.org

ABSTRACT

A great variety of tools are available for analysing speech data, but few permit simultaneous annotation of audio, articulatory and other modalities. Being able to view several data modalities is key when analysing speech production data. The Speech Articulation ToolKIT (SATKIT) is a free open source tool for analysing and annotating speech data which aims to fill this gap. SATKIT has a command-line interface for batch processing data, and a Praat-like GUI for annotating articulatory and audio data. SATKIT has been written in Python, which it also uses as its scripting language. Modifying, extending, and connecting to other systems uses a well documented Python API. SATKIT already includes algorithms which are aimed at making analysing image based data – such as ultrasound – easier. Using SATKIT does not require any paid licenses, enhancing access to this type of annotation. This paper provides an in-depth introduction to this new tool.

Keywords: Articulatory and acoustic annotation, speech segmentation, free open source software

1. INTRODUCTION

Today, there is a great variety of tools available for analysing speech data. To mention a few, we have Praat [1] for annotating sound, and for example for ultrasound and tongue contours we have AAA [2] (which works with other modalities too), as well as GetContours [3], SLURP in 2D and 3D [4, 5], UltraTrace [6], WASL [7] among others. The list can be continued with various tools for analysing EMA data, and topped off with dedicated software from many equipment manufacturers to analyse data from their machines. Some of the availble tools and programs are open source, but none have the combination of being open source, free to use (i.e. no costly licenses required), with an easy-to-learn scripting language (such as Python), and dimensionality reduction and time domain annotation of the data.

It would seem that we should have a tool available for easily segmenting not just audio data, but articulatory modalities too. And that such a tool would be easily scriptable, mesh well with other analysis tools, and be extendable as well as modifiable. Most software is not as handy as Praat is with segmentation, but unfortunately Praat really only analyses audio data, not to mention that scripting Praat can be challenging, and integrating with other tools like Python or R needs a step of first saving the data into files such as TextGrids and then reading them with the system used to process them further. Good packages exist for using this route [8], but closer integration is still missing. And on the other hand, many of the tools available for processing articulatory data are either proprietary software (AAA) or rely on such - most prominently on MATLAB (GetContours, SLURP, WASL).

2. INTRODUCING SATKIT

The Speech Articulation Toolkit (SATKIT) is a free open source tool written in Python for analysing, visualising, and annotating speech data. Our aim is to provide a tool, which is intuitive to use via GUI, command line interface, and scripting interface alike.

SATKIT incorporates algorithms and metrics that are aimed at making analysing image based data – such as ultrasound – easier. It has a commandline interface for batch processing of data, and a Praat-like GUI for annotating articulatory and audio data.

SATKIT features a well documented Python API for extending and customising the toolkit. SATKIT has an object oriented structure which provides base classes for representing data, for writing new metrics to be applied to the data, and for representing annotations. The code is extensively commented and documented.

3. GUI

The SATKIT annotator GUI enables users to mark single point annotations on data, perform textgrid

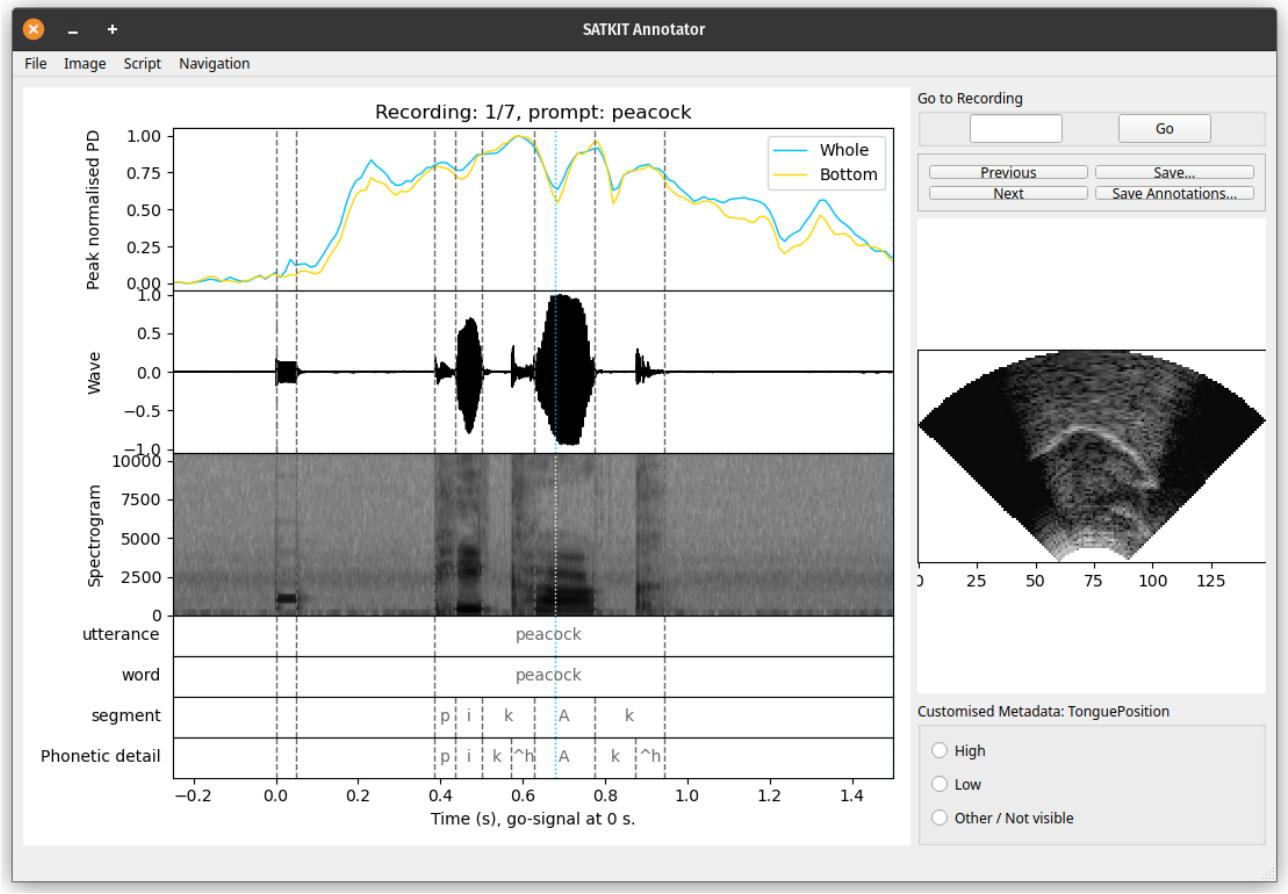


Figure 1: Annotating ultrasound data in SATKIT. Panels on the left from top to bottom are 1. peak normalised PD of both the whole raw ultrasound image and the bottom half, 2. waveform, 3. spectrogram and 4. acoustic boundaries read from a TextGrid. Right panels from top to bottom show widgets to switch between recordings and save data, the interpolated ultrasound frame corresponding to the cursor (dotted light blue vertical line) position within /a/, and a customisable metadata section here showing tongue position categories.

style annotations, and display of ultrasound frames at selected points. An example single point annotation with SATKIT is shown in Figure 1. The annotation GUI is designed for fast manual annotation of large numbers of recordings.

The initial TextGrid was populated with another python tool called Computer Assisted Segmentation Tool (CAST, available at https://github.com/giuthas/cast) [1].

4. DATA PROCESSING

4.1. Data formats

SATKIT can read sound from wav files, ultrasound tongue data exported from AAA (old and new metadata formats as well as splines), 3D ultrasound DICOM files, flow and sound data exported from EVA, avi videos, and Praat TextGrids. Internally data is represented with a layer of abstraction between import functions and the algorithms/metrics that are used to process and plot the data. This makes the list of supported formats readily expandable and the algorithm implementations data format agnostic.

4.2. Algorithms and metrics

SATKIT includes several ways to process data. Most of them provide ways of analysing complex data – such as ultrasound videos – in time without needing to examine the data frame-by-frame. SATKIT includes Pixel Difference (PD) for analysing over all change in an image sequence [9, 10], Optic Flow (OF) for analysing direction of local and/or global movement [11, 12, 13], and a combination of eigentongue analysis and discriminant analysis on ultrasound [14].

5. SATKIT CODEBASE

5.1. Design principles

SATKIT is built using object oriented programming. We aim for a class structure which is intuitive to use and easy to maintain: keep inheritance trees shallow, separate data storage and functionality when it makes sense. The core data classes represent organisational concepts which are likely to be mirrored by directories and files produced when recording data.

5.2. Core data structures

5.2.1. Recording

The Recording class (Fig. 2) represents a single trial from a single recording source. It has one or more Modalities such as MonoAudio and RawUltrasound. It also has metadata: participant, time of recording, etc. and a TextGrid [1, 15].

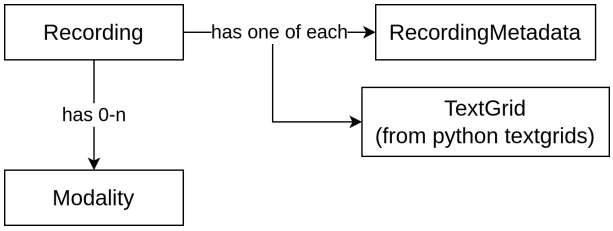


Figure 2: Recording class and its components.

5.2.2. Data Modalities

Different types of data are represented as Modalities derived from the base Modality class (Fig 3). The derived Modality classes provide a represantation of the specific qualities of a given data type.

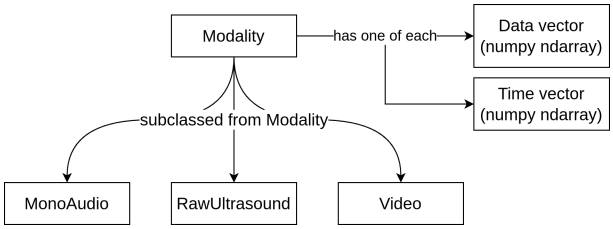


Figure 3: Components and example subclasses of Modality.

5.2.3. Derived Modalities

One of the main development goals SATKIT is to be able to easily run different algorithms on recorded data. The results of doing so are represented by derived Modalities (Fig 4). After creation, there is very little difference between the two Modality types, which means that algorithms and plotting will be as easy to run on derived data as on recorded data.

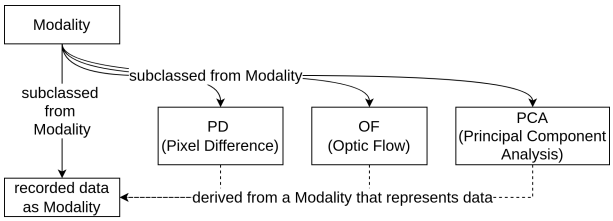


Figure 4: Derived data is represented by subclasses of Modality, which also contain the information on which other Modality class they were derived from.

6. INSTALLING

SATKIT is available on GitHub. The installation requires Python for running, and conda and mamba for managing the required packages. More detailed installation instructions are available in the repository on GitHub. After installing Python, the required python packages, and SATKIT either by forking/cloning from github, SATKIT can be started from the commandline. SATKIT runs on recent operating systems (Linux, MacOS, Windows).

7. ROADMAP BEYOND SATKIT 1.0

7.1. Expanding beyond a single session

Datasets are generally speaking more complex then just a single session and single speaker. To represent this, SATKIT is going to get classes for Datasets, Participants within a Dataset, and Sessions for Participants (Fig 5). The Trial class represents a multimodal recording that consists of one or more Recordings where each Recording represents data from a single data source such as AAA or an EMA machine.

7.2. Expanding to new Modalities

In general, one of our aims is to make extending SATKIT easy and to make it easy to use it with or as part of other tools. Adding new modalities – whether data modalities or new metrices represented by derived modalities – is going to be a routine thing with SATKIT. Writing a new data Modality for SATKIT is a matter of defining the Modality subclass and an importer to construct it. After doing

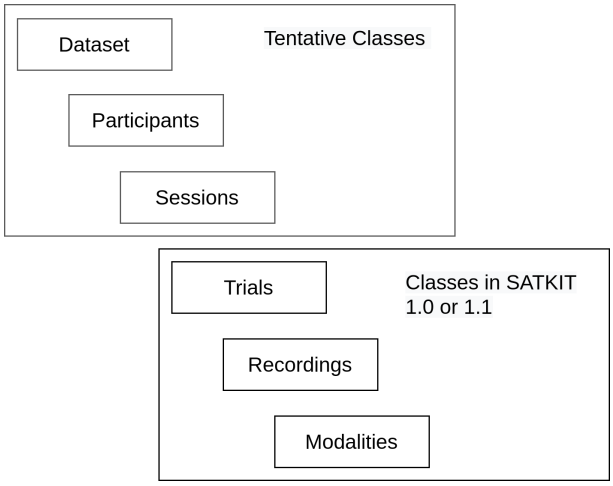


Figure 5: Hierarchy of database classes in SATKIT versions 1.0 and 1.1 and some tentative classes that potentially will become part of SATKIT in a later version.

so the algorithms and GUI will be automatically able to handle the new type of data.

While SATKIT does not generate tongue splines, it will able to process them by version 1.1. This will be done by implementing both a data Modality for splines and different derived Modalities for metrics from literature [16, 17, 18, 19]. Other data Modalities that could be implemented depending on user interest include MRI, EMA, and air flow data from EVA and other systems.

7.3. Other possible improvements

The GUI will definitely be streamlined over time as experience of using it points to good changes. Already at this stage it is clear that an undo function (with preferably an undo stack of more than one command) is a necessity. Adding Praat-style support of annotating data with IPA is easily supported by the underlying data classes but requires a bit more work on the GUI. Having an editable list of hotkeys allowing users to customise how they work with the GUI is another likely improvement in the not-too-distant future.

Currently SATKIT is installed as a Python package and a script. Having a stand-alone executable would be desirable, but selecting the best way to do this to provide multiplatform support requires careful consideration and testing.

8. CONCLUSION

SATKIT is a useful, easy-to-use, freely available tool. It is readily expandable to include new metrics on various data modalities adding to the existing ones for raw ultrasound and audio. We are interested in growing the user community and integrating with other tools and adding new functionality, while keeping the API robust.

9. REFERENCES

[1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2022, version 6.2.12, retrieved 14 April 2022 from http://www.praat.org/.

[2] *Articulate Assistant Advanced User Guide: Version 2.14*, Edinburgh, UK: Articulate Instruments Ltd, 2012.

[3] M. Tiede, "How to "getcontours" from ultrasound imaging," in *Proceedings of Ultrafest IX*, Bloomington, Indiana, 2020.

[4] S. Ghrenassia, L. Ménard, and C. Laporte, "Interactive segmentation of tongue contours in ultrasound video sequences using quality maps," in *SPIE Medical Imaging*, San Diego, United States, 2014.

[5] E. M. V. N. Karthik, E. Karimi, S. M. Lulich, and C. Laporte, "Automatic tongue surface extraction from three-dimensional ultrasound vocal tract images," *Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1623 – 1633, 2020.

[6] K. Murphy, N. Z. Stern, D. Swanson, C. Ho, and J. Washington, "Ultratrace: A free/open-source cross-platform tool for manual annotation of ultrasound tongue imaging data," in *Proceedings of Ultrafest IX*, Bloomington, Indiana, 2020.

[7] S. M. Lulich, K. H. Berkson, and K. de Jong, "Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion," *Journal of Phonetics*, vol. 71, pp. 410 – 424, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447017301481

[8] T. Bořil and R. Skarnitzl, "Tools rpraat and mpraat," in *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2016, pp. 367–374. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-45510-5_42

[9] P. Palo, "Measuring pre-speech articulation," Ph.D. dissertation, Queen Margaret University, Edinburgh, 2019.

[10] ——, "Computer assisted segmentation of tongue ultrasound and lip videos," *Journal of the Canadian Acoustical Association*, vol. 49, no. 3, pp. 44 – 45, 2021.

[11] J. H. Esling and S. R. Moisik, "Laryngeal aperture in relation to larynx height change: An analysis using simultaneous laryngoscopy and laryngeal ultrasound," in *Rhythm, melody and harmony in speech: Studies in honour of Wiktor Jassem*, D. Gibbon, D. Hirst, and N. Campbell, Eds. Polskie Towarzystwo Fonetykzne Poznan, 2012, vol. 14/15, pp. 117 – 127.

[12] S. R. Moisik, H. Lin, and J. H. Esling, "A study

of laryngeal gestures in mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus)," *Journal of the International Phonetic Association*, vol. 44, no. 1, pp. 21 – 58, 2014.

[13] D. P. Z. Poh and S. R. Moisik, "An acoustic and articulatory investigation of citation tones in singaporean mandarin using laryngeal ultrasound," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019.

[14] M. Faytak, S. Liu, and M. Sundara, "Nasal coda neutralization in shanghai mandarin: Articulatory and perceptual evidence," *Laboratory Phonology*, vol. 11, no. 1, 2020.

[15] T. Nieminen, "praat-textgrids – praat textgrid manipulation in python [python package]," 2022, version 1.4.0, available on PyPi and Github.

[16] K. M. Dawson, M. K. Tiede, and D. H. Whalen, "Methods for quantifying tongue shape and complexity using ultrasound imaging," *Clinical Linguistics & Phonetics*, vol. 30, no. 3 – 5, pp. 328 – 344, 2016.

[17] P. Palo, "Can we detect initiation of tongue internal changes before overt movement onset in ultrasound?" in *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, Online / New Haven, CT, 2020, pp. 242 – 245.

[18] N. Zharkova and N. Hewlett, "Measuring lingual coarticulation from midsagittal tongue contours: description and example calculations using English /t/ and /ɑ/," *Journal of Phonetics*, vol. 37, pp. 248 – 256, 2009.

[19] N. Zharkova, F. E. Gibbon, and W. J. Hardcastle, "Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation," *Clinical Linguistics & Phonetics*, vol. 29, no. 4, pp. 249 – 265, 2015.

[1] While the initial TextGrid from CAST looks like forced alignment, it is both dumber and more ergonomic for purposes of human correction. The idea is to provide – with a good deal of configuration options – an evenly spaced phonological segmentation that is at a reasonable position within the recording for a human annotator to move the boundaries to their correct locations.

A SURVEY OF PHONETICS EDUCATION IN ITALIAN UNIVERSITIES

Lorenzo Spreafico

Free University of Bozen-Bolzano
lorenzo.spreafico@unibz.it

ABSTRACT

This paper summarises and comments on the results of a web-based survey that was designed to collect qualitative and quantitative data on phonetics education in Italian universities. The analysis of the data collected makes it possible to present results and make considerations on the type of courses in which phonetics is taught, on the contents that are presented in classes, on the teaching strategies implemented by the academics, on their professional profile, and on their objectives when teaching phonetics.

Keywords: phonetics education, phonetics pedagogy, teaching phonetics.

1. INTRODUCTION

Phonetics is being taught in an increasing number of university settings [1], partly as a result of its practical usefulness [2]. Research on phonetics education, however, is uncommon, certainly less so than research in any other area of the discipline. For example, a search in the proceedings of the International Congress of Phonetic Sciences reveals that from 1979 to 2019 only 5 sessions and 17 contributions were devoted to the topic.

Although interest in the subject has been growing in recent years (see the papers in the proceedings of the Phonetics Teaching and Learning Conference series [3]; as well as [4]), to the best of my knowledge, none of the available contributions deal with the topic with exclusive and explicit reference to the Italian situation[1].

Therefore, the aim of this paper is to fill the gap by offering a first characterisation of phonetics education in Italian universities, so as to understand how it compares with other experiences in Europe and around the world, and how improvements can be made.

2. METHODOLOGY

A survey was conducted among professors, readers, and lecturers teaching phonetics in Italian universities. For this purpose, a web-based questionnaire was designed to collect both quantitative and qualitative information. The questionnaire was completely anonymous and did not collect any personal data about the respondents. The questionnaire contained both compulsory (n=29) and optional (n=9) response questions and both closed-ended (n=24) and open-ended (n=14) questions.

The questionnaire was constructed using the SoSci Survey software [5] and was made available to interested parties at https://www.soscisurvey.de/phonpedit. In order to recruit as many respondents as possible, in November 2022 the link to the questionnaire was posted in the newsletters of six Italian scientific associations and societies among whose members one might expect to find professors, readers, or lecturers teaching phonetics, namely *Associazione Italiana di Audiologia e Foniatria* (Italian Society of Audiology and Phoniatrics), *Associazione Italiana di Linguistica Applicata* (Italian Association of Applied Linguistics), *Associazione Italiana di Scienze della Voce* (Italian Association of Speech Sciences), *Associazione Universitaria Otorinolaringologi* (University Association of Otolaryngologists), and *Società di Linguistica Italiana* (Italian Society of Linguistics).

3. RESULTS AND DISCUSSION

The link to the questionnaire was clicked on 298 times; 94 people answered at least some questions in the questionnaire; 52 people answered all the questions in the questionnaire. In the rest of the paper, only data from completed questionnaires are presented and commented on. Although the database is not particularly large, it is still sufficient to reveal some general trends that can be compared with those that have emerged in previous surveys, for example the one described in [6].

The web-based questionnaire was used to collect quantitative data that were analysed by the Microsoft® Excel® software program (Version 2302) alongside qualitative data in which themes and patterns were identified. A summary of the results is presented below, organized into six thematically distinct sub-sections.

3.1. Which study courses teach phonetics?

The quantitative data collected through closed questions show that in Italian universities phonetics courses are mostly offered in undergraduate degrees

(79%). Less frequent are phonetics courses offered in graduate degrees and postgraduate specialisation schools (21%). None of the respondents reported teaching classes in phonetics at PhD level.

The undergraduate degrees in which phonetics is taught are, in decreasing order of frequency, those in foreign languages and/or linguistic mediation, Italian language and literature, communication sciences and/or sociology, and speech and language therapy. The graduate degrees and postgraduate specialisation schools in which phonetics is taught are (again in decreasing order of frequency) those of linguistics, foreign languages, Italian language and literature, education sciences.

In the light of the data collected, it appears that in Italian universities phonetics courses are more widespread in the departments of social and human sciences, than in any other departments. If these departments offer courses on topics related to phonetics, then it is likely that they prefer to teach general skills "that allow for solving all kinds of problems to dominate over specific domain knowledge in speech" [7].

Two other considerations can be drawn from the data collected. First, that in 85% of the cases, phonetics is taught in courses in which other topics are also covered, mostly those of general linguistics (morphology, syntax, etc.). As a consequence of this, on average no more than 15 hours are dedicated to teaching phonetics in a course, although it should be noted that in the database there are also extreme cases of courses in which only one or two hours are dedicated to the topic of phonetics.

The second consideration that can be drawn from the data collected is that in 70% of cases, phonetics is taught to classes of at least 50 students (and in 45% of cases even to groups of more than 100 students).

The low number of hours devoted to phonetics and the high number of students in the classroom seem to influence both the content and the way in which the subject is taught (see sections 3.2, 3.3 and 3.5).

3.2. What is taught?

The data collected show that the topic dealt with most frequently during the lessons is that of transcription of typical speech (95% of cases), more precisely that of Italian speech (71%). In fact, the transcription of dialects spoken in Italy and/or of other languages is only taught in 42% of the courses.

The second most widespread topic dealt with in lessons is articulatory phonetics (88%); followed by acoustic phonetics (60%) and perceptual phonetics (58%). Every other possible topic is only marginally covered. For example, the phonetics of additional language learning and teaching is covered only in 35% of the cases and sociophonetics or speech technologies are dealt with in just 10% of the cases.

Therefore, overall, what is being taught is still similar to what was being taught in the 1950s [8], i.e. most lecturers remain anchored to "the more leisurely approach of the past" [9], instead of moving towards an integrated approach to the teaching of phonetics that mixes the analysis of the speech signal and its production with descriptive and experimental phonetics in historical linguistics and sound change, and descriptive phonetics of the languages of the world [10].

This is probably due to three factors: firstly, the low participation in the survey of professors, readers and lecturers who teach phonetics in degrees other than the social sciences and humanities; secondly, their educational goals (see section 3.3); thirdly, their expertise (see section 3.4).

3.3. What are the teaching goals?

The data collected through closed-ended questions show that the reason why phonetics is taught is for students to acquire awareness, knowledge, and skills they did not possess before.

As far as raising awareness is concerned, for more than half of the respondents, it is important for students to become aware of the difference between written and spoken language as well as between the phonetic and phonological dimensions of spoken and listened languages.

As far as knowledge is concerned, for about half of the respondents it is important that students learn how speech is articulated and how this information can be used to classify the sounds of language and therefore transcribe them.

Finally, with regard to skills, the acquisition of which affects only a minority of respondents, it appears that the academics who responded are interested in students learning to analyse the language produced by themselves and by others, thus applying these skills to concrete situations, above all those of learning the pronunciation of additional languages.

3.4. Who teaches phonetics?

The data collected show that the majority of professors, readers, and lecturers that teach phonetics in Italian universities fall within the academic discipline[2] of “glottologia e linguistica” (glottology and linguistics) and, less frequently, within one of the other linguistic fields (e.g. Italian, English, Spanish, or German linguistics). In contrast, at least among the respondents, there is only one professor within the academic discipline of “scienze tecniche mediche applicate” (applied medical technical sciences), but

no academic in the fields of engineering or computer sciences.

As far as professors, readers, and lecturers' skills are concerned, almost all the respondents state that they can perceive, articulate, classify and transcribe (albeit sometimes with uncertainties) all the sounds of Italian. On the other hand, only half of the lecturers claim to be able to perform the same operations with reference also to dialects spoken in Italy and/or in other languages. Further information on respondents' abilities could not be systematically collected, but some of them wrote in the responses to open-ended questions that they do not teach acoustic and perceptual phonetics because they lack knowledge and skills in the fields

In the light of this, it seems fair to say that probably the majority of those who teach phonetics in Italian universities are rooted in (general) linguistics rather than in phonetics or speech sciences, which means that only a minority have "sufficient theoretical and methodological competence in both signal processing and language categorisation" [10] as already observed.

3.5. How is phonetics taught?

The data collected show that phonetics is taught by almost all respondents using examples, mostly invented (93% of cases) or, less frequently, real ones mainly taken from phonetic databases (60%).

Data also show that the most frequently used instructional tool during lessons is the chart of the International Phonetic Alphabet, which is used by 88% of the respondents often also in its multimedia version. Less frequent is the use of other more or less technological tools such as recorders (63% of cases) or software for articulatory (33%), acoustic (32%), perceptual (20%) or transcription (13%) analyses. The multimodal presentation of content, e.g. using videos [10], graphic simulations or three-dimensional models, is also quite infrequent (at most 15% of cases), while the use of graphs and anatomical drawings is more widespread (85%).

Therefore, despite the almost inevitable growth in the use of digital tools due to distance learning during the COVID-19 pandemic, this use in Italian universities, unlike in other countries, does not seem to have opened "new doors in the phonetics classroom", nor has it heralded the affirmation of new techniques for phonetics teaching [12]. From an educational point of view, however, this is disappointing. The access to digital tools can act as both a compensatory measure for the low number of teaching hours, where less attention can be given to individual learners especially in the case of practical activities such as speech transcription and analysis and as a motivational measure for the students attending overcrowded, usually instructor-centred, classes.

3.6. Why is phonetics taught that way?

The qualitative data collected through closed-ended questions show that the respondents to the questionnaire teach phonetics the way they do for at least five different reasons.

In about 40% of cases, the respondents answered that they adopt the practices they reported because they believe that what they do is the best approach for students to achieve the educational objectives and goals.

In about 25% of cases, the respondents answered that they adopt the reported practices because they deliberately decided to use the teaching approach and method they had themselves been exposed to in their phonetics education, i.e. because they consider the model of their masters valid and want to perpetuate it.

In 13% of cases, the respondents reported that they do as they do because they are conditioned by circumstances that leave them no alternatives, in particular as a result of the few hours of lessons and the high number of students they have to manage.

Just as frequently, the respondents stated that they teach the way they teach because they slavishly follow a manual.

Finally, the respondents answered that they teach a certain way because that is the way they like to do it or it is the way they wished they had been taught when they were learning phonetics.

Differently, and relevant to the line of inquiry in which this research fits, none of the respondents claimed to teach in a certain way because they refer to some proposal in phonetics pedagogy. In fact, 87% of the respondents have never consulted research on this topic, although 62% of them would find it useful if phonetics teaching courses were organized, and 54% of them would participate in them if they were offered.

4. CONCLUSION

This survey is the first to have attempted a systematic quantitative and qualitative overview of phonetics education in Italy, with the aim of identifying phonetics teaching contents and practices.

The picture that emerges is that of a field with rather uniform syllabi, usually not oriented either to what have long been widespread contents in other European countries, in particular the northern ones [6], or to the inclusion among the disciplinary contents of what are now the most widespread topics in research and for research applications. It also emerges that usually teaching is not based on

"vibrant" teaching techniques such as those described for other parts of Europe [6].

Although the aim of the survey was not to assess the quality of teaching or to identify best teaching practices, it is still possible to identify at least three critical points that would deserve consideration by policy makers interested in improving the teaching of phonetics, also given its increasing relevance as an applied science [2].

The first one concerns the number of hours devoted to the subject, which is often inadequate to achieve the minimum learning goals [9] and should be increased.

The second point concerns the number of students, usually exceeding what is desirable to be able to balance theory and practice, which are inseparable components in the teaching of phonetics [9], even with digital tools and software [7, 12]. The number of students in the classroom should therefore be reduced.

The third point concerns the compilation of a syllabus of basic phonetics topics for the different classes and courses that would take up, for example, the proposal presented in [7] or in [10]. According to the data collected through the survey among academics at Italian universities, this is an action with which 70% of the respondents would agree.

To be useful and appreciated by students, this syllabus should probably include the treatment of acoustic phonetics alongside that of articulatory phonetics; the analysis of phenomena present in the Italian language but also of those traceable in other languages, even if unknown to the students; the adoption of a lab-based approach based on experiences and experiments of increasing complexity and feasibility even in large groups [13, 14, 15].

If accompanied by dedicated training in phonetics pedagogy for professors, readers, and lecturers, all this could change the teaching of the discipline and, with this, possibly increase the disciplinary commitment of learners [16].

5. ACKNOWLEDGMENT

The author is grateful to the survey respondents for their helpfulness and to an anonymous reviewer for comments.

6. REFERENCES

[1] Ashby, P. Ashby, M. 2013. Phonetic Pedagogy. In: Jones, M., Knight, R. (eds.). *The Bloomsbury Companion to Phonetics*. Bloomsbury, 198–207.
[2] Munro, M. J. 2020. *Applying phonetics: Speech science in everyday life*. Wiley-Blackwell.
[3] Phonetics Teaching and Learning Conference. 2017. *Proc. PTLC2017*, https://www.ucl.ac.uk/drupal/site_pals/sites/pals/files/migrated-files/PTLC-Proceedings-2017.pdf.
[4] Knight, R., Setter, J., Whitworth, N. 2021. Pedagogical Approaches. In Knight, R., Setter, J. (eds.), *The Cambridge Handbook of Phonetics*. Cambridge University Press, 503–526.
[5] Leiner, D. J. 2019. *SoSci Survey* (Version 3.1.06). Available at https://www.soscisurvey.de.
[6] Hazan, V., van Dommelen, W. 1997. A survey of phonetics education in Europe. *Proc. 5th Eurospeech* Rhodes, https://www.isca-speech.org/archive/pdfs/eurospeech_1997/hazan97_eurospeech.pdf.
[7] Bloothooft, G. 1997. The landscape of future education in speech communication sciences. *Proc. 5th Eurospeech* Rhodes, https://www.isca-speech.org/archive/pdfs/eurospeech_1997/bloothooft97_eurospeech.pdf.
[8] Ladefoged, P. 1995. Teaching phonetics. *Proc.12th ICPhS* Stockholm, 1, 432–433.
[9] Ashby, P. 2005. Phonetic pedagogy. In: Brown, E.K., Anderson, A., Bauer, L., Berns, M., Hirst, G., Miller, J. (eds.), *Encyclopedia of language and linguistics. 2nd edition*. Elsevier, 372–378.
[10] Kohler, K. J. 2000. The future of phonetics. *Journal of the International Phonetic Association* 30, 1–24.
[11] Smith, J. 2011. The Youtube Revolution: Engagement, Perception and Identity. *Proc.17th ICPhS* Hong Kong, 100–103.
[12] Mompean, J., Ashby, M., Fraser, H. 2011. *Phonetics Teaching and Learning: An Overview of Recent Trends and Directions*. *Proc.17th ICPhS* Hong Kong, 96–99.
[13] Hay, J. 2007. A toolbox for teaching phonetics. *Te Reo: J. Linguistic Society of New Zealand*, 50, 7–16.
[14] Mills, T., Pollock, K., Tucker, B. 2015. Laboratory activities for large and online phonetics classes. *Proc. PTLC2015*, London, 57-61.
[15] Arai, T. 2019. Vocal-tract models in phonetic teaching and research. In: Katz, W, Assmann, P. (eds), *The Routledge Handbook of Phonetics*. Routledge, 570–598.
[16] Diettes, K.J., Rodríguez, F., Paloma, J.A., Pico, J.D., Mora, G.M. 2019. Phonetics and Phonology Engagement: Evaluation of a Productive Disciplinary Engagement Approach. *Proc.19th ICPhS* Melbourne, 1912–1916.
[17] Berruto, G. 2020. Riflessioni sull'insegnamento della linguistica oggi. In: Sansò, A. (ed.) *Insegnare Linguistica*: *basi epistemologiche, metodi, applicazioni*. Roma: 13–30.

[1] Brief considerations can be found in a paper by Berruto [17] on the teaching of general linguistics where he notes that the topics of phonetics and phonology are the second most difficult to teach according to a sample of forty linguistics professors from Italian universities.

[2] Each academic at Italian universities belongs to a single academic discipline, also known as the *settore scientifico disciplinare*. The complete list of academic disciplines is accessible at https://www.cun.it/documentazione/academic-fields-and-disciplines-list/. Last accessed on November 15, 2022.

RE-EVALUATING THE ASSESSMENT OF PHONETIC SKILLS: WHAT WE LEARNED DURING THE PANDEMIC

Jane Setter[1], Rachael-Anne Knight[2] and Nicole Whitworth[3]

[1]University of Reading; [2]City, University of London; [3]Leeds Beckett University
j.e.setter@reading.ac.uk; r.knight-1@city.ac.uk; n.whitworth@leedsbeckett.ac.uk

ABSTRACT

The sudden move to online teaching and assessment during the pandemic proved a challenge for many subjects; this was no less the case for phonetics. This paper describes how we endeavoured to devise comparable online assessment tasks for the evaluation of two aspects, articulatory phonetic skills and understanding of phonetics theory, and our evaluation of the success of the online tasks.

Normally assessed under exam conditions with no access to external resources, we altered and developed our assessments so that students could complete the tasks online. Scores from the outcomes of these assessments are compared with pre-pandemic scores. We discuss what we have learned from moving these assessments online, implications for students, and consider whether we will continue with these methods of assessment outside of lockdown conditions.

Keywords: Assessment, pedagogy, articulatory phonetics, phonetics theory

1. INTRODUCTION

E-assessment is not a new phenomenon. The use of computers in automatic assessment dates back as far as the 1920s [1], with assessment using web-based interfaces emerging in the 1990s [2]. Llamas-Nistal et al. [2] indicate that successful e-assessment and evaluation was becoming increasingly common pre-pandemic.

Where teaching is concerned, virtual learning environments (VLEs) such as Blackboard or Moodle, which are used to provide content to support teaching, learning and assessment, were pressed into greater service during the COVID-19 pandemic when it was no longer possible to teach or assess students face to face. Online interfaces such Blackboard Collaborate enabled live, interactive classes, with video technologies such as Microsoft Stream or YouTube presenting lecture content which would normally have been delivered in real time. This is not dissimilar to the flipped classroom approach (see, e.g., [3]; [4]), in which students watch video lecture material prior to coming to class in order to focus class time on the rehearsal of practical skills. The capability to set up online tests and provide areas for students to upload assessments electronically via VLEs also came much more to the fore during the pandemic, as it was not desirable to submit assignments in hardcopy or meet students face to face for oral assessments such as presentations.

Kearns [5] examined the online assessment in 24 courses taught entirely online at a University in the United States. Of the 63 assessments she identified, five categories emerged: written assignments (22 = 35%); online discussion using a discussion board (19 = 30%); fieldwork reports (9 = 14%); quizzes and exams (8 = 13%); and presentations (5 = 8%), for which students used wikis, discussion boards and synchronous webinar sessions to respond to questions posed by their peers, although it is unclear from the paper whether students uploaded video content of the presentation or documents in the form of slides or a poster. At the time of her research, published in 2012, written assignments and online discussion far outstripped the other three categories in terms of frequency of use.

Phonetics skills are traditionally assessed in three ways: oral production, transcription and the examination of theoretical concepts. In this paper, we look at the assessment of oral production skills and the examination of theoretical concepts using online tools in two UK universities and examine whether grade inflation occurred when assessments moved online. In Kearns's [5] categories, oral production comes under presentations, whereas transcription and the examination of theoretical concepts fit best under quizzes and exams, depending on how the material is assessed. In some areas – e.g.,

sociophonetics, acoustic phonetics, or observation of speech in clinical contexts – more holistic assessments which fall into Kearns's categories of written assignments or fieldwork reports are likely to be appropriate, depending on the intended learning outcomes of the module; we do not cover assessments of this kind here.

2. ASSESSING STUDENTS' ORAL SKILLS

In the assessment of phonetics production skills, students typically carry out three in-person tasks individually:

- Production of sounds selected at random from the IPA chart;
- Substitutions, in which students provide a voice, place, manner (VPM) label for consonants substituted into existing words;
- The description of intonation patterns.

This assessment typically takes up to 15 minutes per student.

While some universities in the UK had done away with phonetics orals even before the pandemic, we are committed to ensuring our students have the opportunity to demonstrate production skills, not least because it is useful in work environments such as speech and language therapy. However, research has shown that in-person oral skills tests can be anxiety-provoking among students [6, 7]. The necessity to move to online assessment of oral production skills enabled us to re-think the assessment in terms of what might work better as an online test and how students might be supported in completing it.

2.1. Design of the online oral production skills test

Rather than simply videoing the same kind of test students would do face to face, we decided to do the following:

1. Students were given a series of sounds to produce.
2. They made a video of themselves producing the sounds.
3. They gave spoken labels for the sounds, using VPM for consonants or tongue height/backness/lip rounding for vowels.
4. They were asked to produce a sentence with two different intonation patterns.
5. They provided a spoken analysis of the two different intonation patterns.
6. They gave reflections on the experience.

This differed from the face-to-face test in a number of ways: sounds are usually selected at random for students to produce during the test, not provided in advance (1); the sentence for intonation is also not provided in advance, and students describe their own and the examiner's intonation patterns (4, 5); substitutions were not tested, so we asked students to provide labels for the sounds we gave them (3); we do not usually ask for reflections immediately afterwards.

In addition to 1-6 above, we gave students a model video so that they could see the kind of thing we wanted and to help alleviate anxiety. We also gave them clear instructions in PDF format and encouraged them to make the video in one take without editing it, enlisting a family member to help if needed. They could produce the video using any technology they wished (many used their mobile phone) and we allowed them to upload and view a practice video. We had no issues with any of the assessed videos, which were all clear enough in sound and vision and straightforward to mark. Students were given a 24-hour period to produce and upload the video to the VLE. The tests were marked out of 100 and a sample was moderated.

2.2. Results and comparison with face-to-face test

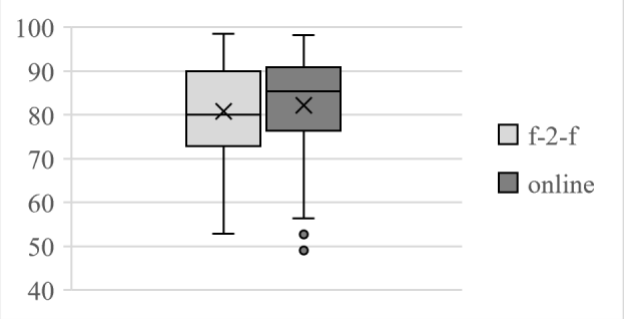


Figure 1: Box and whisker plot comparing face-to-face (f-2-f) and online oral assessment scores.

	f-2-f	online
N	82	91
Min score	52.86	49.09
Max score	98.57	98.18
Mean score	80.78	82.18
Range	45.71	49.09
StDev	10.77	11.71

Table 1: Descriptive statistics for face-to-face (f-2-f) and online oral assessment scores.

While it might be assumed that giving students 24 hours to produce and upload their video meant they had greater access to external resources and therefore should perform better, and that by removing substitutions we had eliminated a challenging task, we did not see a large difference in attainment between face-to-face and online orals, as shown in Figure 1 and Table 1. A simple two-tailed t-test indicated that the difference was non-significant at $p > 0.05$.

We observed that students still made errors in labelling. For example, "lateral" was used by some as a place of articulation (PoA) rather than part of a manner of articulation (MoA), incorrect combination PoAs were given such as "dental alveolar" or "retroflex alveolar", and the MoA of velaric egressive consonants (clicks) was not identified fully. This, and the fact that the mean scores are highly similar, indicates to us that this method of assessing phonetics oral skills was successful and provided a suitable alternative to the face-to-face test.

3. ASSESSING KNOWLEDGE OF PHONETICS THEORY

Phonetics theory can be assessed in a number of different ways. One method which appears in the Examination for the Certificate of Proficiency in the Phonetics of English offered by the International Phonetic Association is an articulatory description of a phrase [8]. This detailed description, however, assumes comprehensive knowledge of the processes of articulation which may not be suitable for students operating at a more introductory level.

3.1. Initial and subsequent design of the online quiz

In order to assess more basic understanding of the articulation of English speech sounds and intonation in a course mainly focussed on Roach [9], prior to the pandemic we had developed a class test in quiz form with these sections:

1. Ten multiple-choice questions on segmental features referring to a list of items;
2. Ten true/false statements drawn from any aspect of the module materials;
3. One question on intonation in which students had to suggest a pattern on a phrase in context by annotating the phrase with diacritics and describing the pattern in terms of pre-head, head, tonic syllable (nucleus) and tail;
4. One diagram question in which students, e.g., completed and labelled a partially drawn mid-sagittal section;
5. One question on vowels which contained multiple sub-questions involving placing three vowels from the vowel system on a vowel chart, describing them and providing an example word for each;
6. One question on any other aspect we had covered in the module (e.g., assimilation; weak form words; etc.).

This worked well as a one-hour paper-based test under exam conditions (with extra time for students with reasonable adjustments), allowing us to examine a range of knowledge and skills.

When it became evident that the test would have to be moved online during the pandemic, we first gave consideration to whether it could be sent to students as a document for them to complete electronically or by hand and upload in e.g. a 23-hour period. To preserve comparability with a one-hour test period and test knowledge students had learned in the course of the module rather than their ability to refer to external sources, it was decided to commute the test to a timed, one-hour online quiz.

The VLE we were using had options for multiple choice questions (MCQs) (1), true/false questions (2) and short answers (6) but did not allow students to draw diagrams (4) or annotate diagrams or phrases (3, 5).

To deal with the diagrams in (4) and (5), we made use of the "hotspot" function. This allows the test-writer to upload an image to the quiz and select an area on that image which shows the correct answer to the question. For example, in the case of mid-sagittal sections, one can upload an image of four similar diagrams and then ask the student to click on the correct one (see, e.g., Fig. 2). This can then be followed up with multiple-choice or short answer questions to check understanding. Similarly, hotspot questions can be used to identify the correct location of vowels on the IPA vowel chart.

For (3), students were given a phrase in context and asked to select a suitable intonation pattern using a MCQ. Using short answer questions, they identified the pre-head, head,

tonic syllable and tail and explained the communicative function of the placement of the tonic syllable.

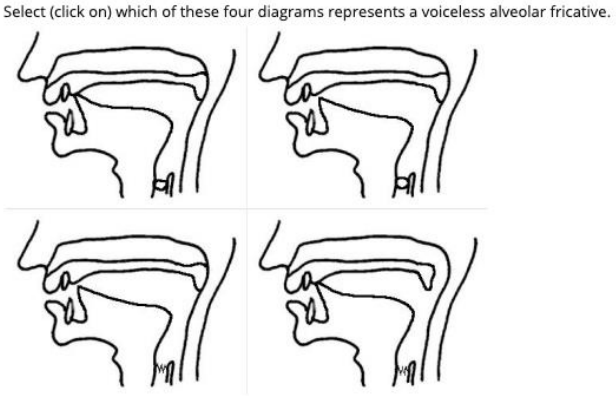


Figure 2: Sample "hotspot" question.

To mitigate possible collusion or extensive reference to external resources, questions and distractors were presented in random order for each student where possible and the test was set up with a one-hour timer (with extra time for students with reasonable adjustments); students could start the test any time within a 23-hour period. We provided an online demo test as a mock and went over it in class to check for any technical or practical issues. We avoided asking students to insert any phonetic symbols; separately they also undertook a dictation to broad phonetic transcription (not reported on here; see [10] for a proposed approach). Multiple choice, true/false and hotspot questions are marked automatically by the VLE.

3.2. Results and comparison, face-to-face and online quizzes

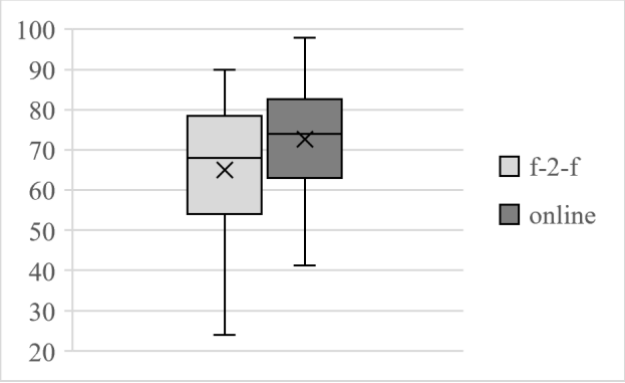


Figure 3: Box and whisker plot comparing percentage marks for f-2-f and online phonetics quiz scores.

Unlike the oral test, scores were higher for the online quiz than for the face-to-face version. The mean score rose from 60s-range to 70s-range with no score under 40% (the passing mark for undergraduate work in the UK). In addition, the Standard Deviation narrowed by three percentage points. A simple two-tailed t-test indicated that the difference was significant at $p < 0.001$. However, there was no evidence of collusion where the short answer questions were involved, but it is not possible to rule out reference to external sources.

	f-2-f	online
N	118	119
Min score	24	41.3
Max score	90	97.83
Mean score	64.97	72.65
Range	66	56.52
StDev	15.83	12.77

Table 2: Descriptive statistics for f-2-f and online phonetics quiz scores.

4. DISCUSSION AND CONCLUSIONS

Scores remained relatively similar in the online version of the oral test. We can, therefore, conclude that it is a comparably rigorous alternative to the face-to-face test. The online version of the quiz led to somewhat higher marks; if we kept one online variant and discarded the other, returning to a face-to-face quiz under exam conditions is the obvious recommendation. However, we are satisfied that the online quiz gave students the opportunity to demonstrate knowledge and understanding – as are our external examiners – which is surely the point of assessment. The improvement may also indicate that this approach helps to alleviate the stress experienced by some students taking normally paper-based exam-type assessments under traditional conditions, enabling them to perform better (but see the discussion in [11]).

In addition, and where both tests are concerned, time administering and marking the tests was considerably reduced. In the current days of increased student numbers and demands on academic staff, this cannot be a bad thing. We therefore conclude that, for the assessment types considered here, online assessment is a success, and we will continue to use it in some form.

5. REFERENCES

[1] Skinner, B. F. 1958. Teaching machines. *Science* 128(3330), 969–977.

[2] Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., Mikic-Fonte, F. A. 2013. Blended e-assessment: Migrating classical exams to the digital world. *Computers & Education* 62, 72-87.

[3] Bergmann, J., Sams, A. 2012. *Flip your classroom: Reach every student in every class every day*. International society for technology in education.

[4] Setter, J. 2015. Using flipped learning to support the development of transcription skills among L1 and L2 English speaking students. *Proceedings of the Phonetics Teaching and Learning Conference, UCL, London, 2015,* 77-82.

[5] Kearns, L.R., 2012. Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching* 8(3), 198-208.

[6] Knight, R.-A., Dipper, L., Cruice, M. 2013. The use of video in addressing anxiety prior to viva voce exams. *British Journal of Educational Technology* 44, E217–E219

[7] Knight, R. A., Dipper, L., Cruice, M. 2018. Viva survivors–the effect of peer-mentoring on pre-viva anxiety in early-years students. *Studies in Higher Education* 43(1), 190-199.

[8] International Phonetic Association Examination for the Certificate of Proficiency in the Phonetics of English. https://www.internationalphoneticassociation.org/content/ipa-exam

[9] Roach, P. 2009. *English phonetics and phonology: A practical course (4th ed.)*. Cambridge University Press.

[10] Timmins, C. 2017. Online video assessment of clinical phonetic transcription skills in speech and language pathology. *Proceedings of the Phonetics Teaching and Learning Conference, UCL, London, 2017*.

[11] Czura, A., Baran-Łucarz, M. 2021. “A stressful unknown” or “an oasis”?: Undergraduate students’ perceptions of assessment in an in-class and online English phonetics course. *Íkala, Revista de Lenguaje y Cultura* 26(3), 623-641.

DISTANT RHYTHMS: COMPUTING FLUENCY

Xuewei Lin and Dafydd Gibbon

Jinan University, Guangzhou and Bielefeld University
Corresponding author: gibbon@uni-bielefeld.de

ABSTRACT

The phonetic component of L2 fluency assessment has been treated in many previous studies with a variety of annotation-based time-domain measurements (syllable, word and phrase rate, speech-pause ratio, filled-unfilled pauses, pause count, mean run duration) and rarely with frequency-domain methods. A new experimental frequency-domain fluency comparison approach is proposed, using automatic detection of low-frequency spectral features and hierarchical clustering. The method is used to compare L2 English readings by speakers of Mandarin (Pǔtōnghuà) at different fluency levels, with the IPA benchmark text "The North Wind and the Sun" as prompt. The automatic method successfully separated intermediate level L2 English readers from an advanced L2 reader and L1 English readers in a pilot study, showing the novel method to be feasible. A subsidiary hypothesis that a specific text format with stressed vowels marked in bold face leads to more consistent and native-like L2 reading was also successfully tested. Tests with larger databases are planned. The code is freely available.

Keywords: fluency, accentedness, L2 assessment, spectral features, rhythm

1. FLUENCY; THE FREQUENCY DOMAIN

Assessment of phonetic factors in L2 speech fluency includes both time-domain measures of partially automatised production skills, such as types of syllable rate and pruning, mean interpausal ('run') duration, filled and unfilled pause ratios, and also subjective judgments of 'accentedness', which are "operationalized using impressionistic judgments of how far L2 speakers' pronunciation diverges from a native speaker target" [20]. Both approaches have subjective elements which involve qualitative decisions on accentedness and on the types of unit to be measured: syllables differ in type within and between languages, discourse 'hedges' may fail to be counted as non-fluency, reliability of speech timing segmentation varies. Rhythm assessment (as opposed to teaching methods) is not strongly represented, and prosody tends to be reduced to intonational pitch patterning [2], although structural and discourse-level patterns are well-understood [6].

Phonetic, and particularly prosodic aspects of fluency are sometimes neglected in assessment procedures, but they have consequences for fluency in semantics and comprehension: accentedness, including rhythm patterning, can relate to social prejudice [7]. This underlines the need for complementing traditional methods with reliable (reproducible) and 'objective' (validated automatic) procedures [10]. The time-domain phonetic parameters are well-known and form a single complex: *rhythmicity* [25, 4, 24]. But stress patterns, and the multiple rhythms which are relevant for intermediate and advanced speakers, tend to be missed by traditional time-domain measurements.

The present exploratory study applies a novel acoustic phonetic rhythm assessment method with automatic frequency-domain analysis of low-frequency (LF) rhythms to a small but well-defined data set and measures divergence among intermediate standard L2 English readers and between these and highly fluent readers. A secondary aim is to test whether text highlighting supports rhythm consistency in reading. Time T and frequency f are closely related, ($f = 1/T$), but spectral analysis and visualisation offer added heuristic and theoretical value [21, 8].

In [18, 12], frequency-domain fluency parameters were measured in the 200 Hz to 4 kHz frequency band, with wavelet transform and principal component analysis. The present approach analyses *salient spectral magnitude peaks* in the low frequency (LF) spectral range, which are interpreted as *rhythm formants*. Rhythm formants from 10 Hz to 1 Hz mark rhythmic grammar units such as syllables and words, while below 1 Hz they mark rhythmic discourse units (corresponding to time domains between 0.1 s and 1 s, and longer than 1 s, respectively). Rhythm differences are modelled as distances between rhythm formant patterns, using standard distance metrics.

Reading fluency was selected partly to constrain the set of variables for the pilot study, partly because reading aloud is an essential skill in many

professional activities, and for information and entertainment. Reading fluency has been defined in terms of accuracy, tempo and expression [13, 14], or as reading rapidly, smoothly, effortlessly, automatically (with little conscious attention to subskills of decoding and word recognition) [17]; for general and multimodal reading contexts cf. also [11, 15]. Rhythm relates particularly to tempo, expression and smoothness.

Section 2 describes the data, in Section 3 gives an overview of the method, and in Section 4 the results are discussed. Section 5 summarises conclusions, problems and future work.

2. DATA

2.1. Readers

The investigation uses a small structured dataset. All readings are from the IPA English benchmark text, "The North Wind and the Sun", to facilitate reproducibility. The readers have different English reading proficiencies, ages and genders:

1. British English L1: *S1*, researcher, conservative RP [1], ID NW048-S2; *S2*, researcher, modified RP, recorded for this study, ID NWAS-048-S1.
2. Advanced L2 English: *S3*, female L1 Mandarin speaking university English teacher, advanced L2 proficiency, recorded for this study. 3 recordings from each of 5 text formats, IDs NWAS_A_01 ... NWAS_D_03.
3. Intermediate L2 English: 10 female and 10 male Mandarin speaking students [26], not language majors (not students of *S3*); 1 reading each, diverse gender-prefixed IDs.

Clustering by proficiency is predicted.

2.2. Text prompt formats

In the social media there are countless videos by both professionals and lay influencers with advice on L2 English pronunciation. The use of different text presentation formats with *S3* was motivated by an informal recommendation made in a video by a popular anonymous Chinese English teaching influencer [19]: *Mark all vowels in the text*. This marked-vowel proposal is original, but from a phonetic point of view suggests that marking all vowels would encourage a Chinese-like syllabic rhythm.

The alternative hypothesis proposed here is: *Mark all lexically stressed vowels in the text*. Highlighting only lexically stressed vowels could encourage more consistent stress-oriented English rhythm, since stress position is a major factor in English speech timing. Stress awareness is recognised as a necessary component of pronunciation training for Chinese speakers [4] [5] and requires appropriate assessment methods. Text formats and font effects are already known to affect reading speed [22] and reading by dyslexics [3].

S1, *S2* and the students used unmodified texts with the story as a single paragraph. For *S3*, different text formats were used; *S3* knows the video vowel marking suggestion (*E* below), but was not informed of the purpose of the other format options. The text prompt formats, including *A*, *B* and *C* as distractors, are:

A: Normal text (text as one paragraph).
B: Each sentence on a new line.
C: Clauses and longer phrases on new lines.
D: Normal text, with stressed vowels marked.
E: All vowels marked (but not 'silent' vowels).

The texts were read in this order, so rehearsal effects cannot be discounted. The text for the *D*-readings was marked with stressed vowels, following the simplifying principle that all lexical 'content' words are stressed, rather than taking rhetoric or information structure into account:

Text D. The North Wind and the Sun

The N**o**rth W**i**nd and the S**u**n were disp**u**ting which was the str**o**nger, when a tr**a**veller came al**o**ng wr**a**pped in a w**a**rm cl**oa**k. They agr**ee**d that the **o**ne who f**i**rst succ**ee**ded in m**a**king the tr**a**veller t**a**ke his cl**oa**k off should be cons**i**dered str**o**nger than the **o**ther. Th**e**n the N**o**rth W**i**nd bl**ew** as h**a**rd as he c**ou**ld, but the m**o**re he bl**ew** the m**o**re cl**o**sely did the tr**a**veller f**o**ld his cl**oa**k ar**ou**nd him; and at l**a**st the N**o**rth W**i**nd gave **u**p the att**e**mpt. Then the S**u**n sh**o**ne out w**a**rmly, and imm**e**diately the tr**a**veller t**oo**k off his cl**oa**k. And s**o** the N**o**rth W**i**nd was obl**i**ged to conf**e**ss that the S**u**n was the str**o**nger of the t**wo**.

3. METHOD

3.1. Demodulation

Figure 1 illustrates the analysis procedure using the first reading (A1) by *S3*; the original sampling rate of 48 kHz was retained. The top panel shows the *amplitude modulated (AM) waveform*. The superimposed outline shows the demodulated low-pass filtered *absolute amplitude envelope*, the LF information signal, which is an approximate correlate of the phonotactic sonority curve, and whose magnitude and duration variations are the main acoustic factors underlying speech rhythms. The mid panel of Figure 1 shows the LF spectrogram of the reading. The bottom panel shows the LF

spectrum with the whole recording as FFT window. Frequency modulation (FM), i.e. the F0 track, is also relevant for fluency and accentedness but together with, the modulation-theoretic background, is dealt with elsewhere [8].

The method is related to the algorithms of music identification apps like Shazam [23].

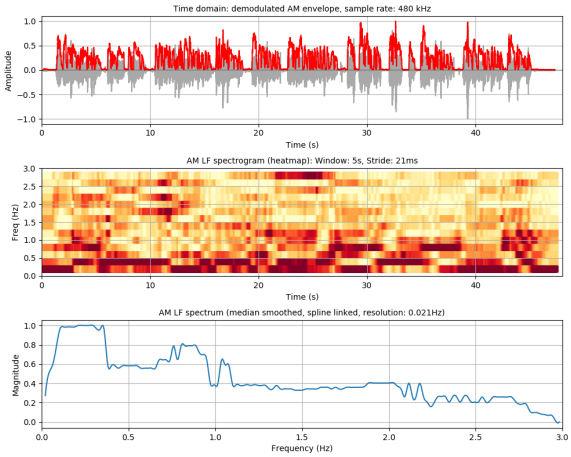


Figure 1: Demodulation and spectral analysis of reading A1 by *S3*. Top: waveform with overlaid amplitude envelope outline. Mid: LF spectrogram. Bottom: LF spectrum.

3.2. Spectral analysis and comparison

The *LF spectrogram* (Figure 1 mid panel) shows rhythm formants as dark horizontal bars: rhythms have minimum durations [16] and constant frequency. The spectrogram is a sequence of spectral slices with 5 s windows (for LF analysis, resolution about 0.21 Hz) with 21 ms steps (for higher time resolution than 5 s). An *LF spectrum* is created by applying an FFT to a window consisting of the entire recording [21, 8] (Fig. 1 bottom panel).

Figure 1 bottom panel shows the *LF spectrum* (resolution 0.021 Hz) with multiple formants: 0.1...0.4 Hz (10...2.5 s beat period), 0.7...0.9 Hz (1.4...1.1 s beat period) and 1.8...2.8 Hz (0.6...0.4 s beat period), which relate to longer sentences, phrases and word constituents.

For comparing the recordings, two sets of vectors are extracted: one from the spectrogram, consisting of the *highest magnitude frequency in each spectral slice*, and one from the spectrum, consisting of the *10 highest magnitude frequencies in the spectrum*.

The vectors in each set are *compared pairwise* using a standard metric, the sum of absolute differences (also known as Manhattan, Cityblock, Taxicab or Mannheim Distance), to indicate divergence from each other and the distances are *hierarchically clustered* with the Nearest Point (Single) linkage algorithm (empirically selected from a range of distance metrics and linkage algorithms). The outputs are agglomerative hierarchical clusterings, visualised as dendrograms, in which the horizontal branch length approximately indicates distances between clusters.

The analysis and plotting procedures are implemented as open source Python prototypes using the NumPy, SciPy, MatPlotLib and Tkinter libraries [9].

4. RESULTS

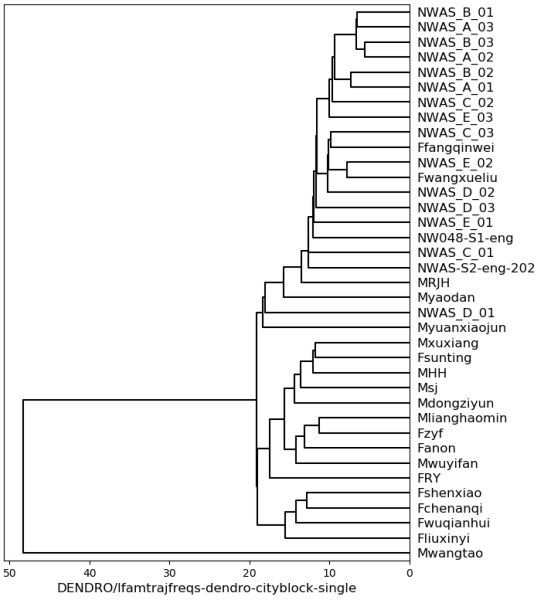


Figure 2: Dendrogram of maximum magnitude frequencies in the LF spectrogram.

The main aim of the experiment is to determine whether spectral features can be used to distinguish between different L2 English proficiency grades, with the subsidiary aim of examining the effect of text formats on consistency of reading.

First results are shown in Figure 2 for time-domain maximum magnitude vectors from the three-dimensional LF spectrogram (*time* $\times$ *frequency* $\times$ *magnitude*). The *LF spectrogram-based clusters* in Figure 2 involve not only frequencies but also frequency variation over time, and show a relatively clear distinction, with a few student results intermingled with the native speaker results. Evidently student gender is not a distinguishing feature (cf. M and F name prefixes). However, the fine-grained LF spectrogram vectors may be too high-dimensional for present purposes by over-emphasising high-resolution frequency

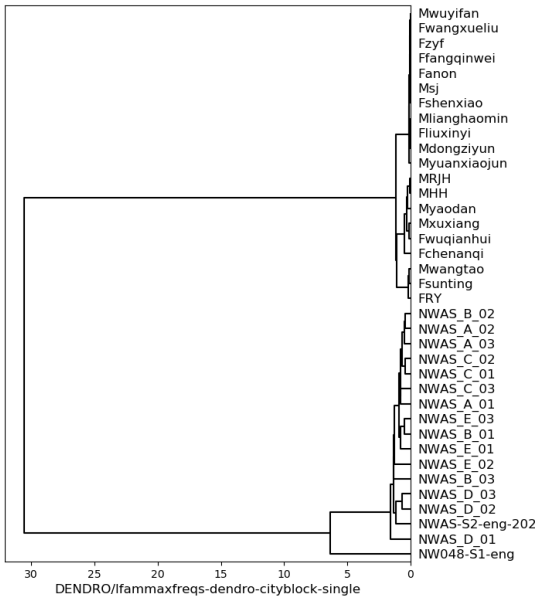


Figure 3: Dendrogram of maximum magnitude frequency similarities.

variation over time in the main frequency zone at the expense of identifying the relevant spectral frequency zones and thus the multiple rhythms.

Figure 3 shows the result for frequency-domain vectors of the 10 highest magnitude frequencies in the *LF spectrum-based clusters* for each reading. The LF spectrum excludes time information, and thus compresses rhythm formants from the entire recording into two dimensions rather than three. Further comparison using Canberra (Normalised Manhattan) and Euclidean distance metrics as a cross-check yields similar results. Pearson Distance provides none of the predicted clusterings: distance between feature vectors seems more important than vector shape difference.

Abstracting away from the time-domain, and considering only the *LF spectrum-based clusters* in the frequency-domain, a clearer picture indeed emerges in Figure 3. There is a partition between the student recordings and the others, and a further partition between *S1* (L1 conservative RP) and the cluster with *S2* (L1 modified RP) and *S3* (advanced L2). With regard to the presentation formats for *S3*, the 3 *D*-format readings cluster together. The student readings cluster without regard to gender.

5. DISCUSSION

The clustering shown in Figure 3 indicates that the first partition separates the intermediate level readers from the highly proficient readers. The 3 bolded-vowel format *D*-readings cluster together, indicating consistent reading by *S3* in this format. Also,

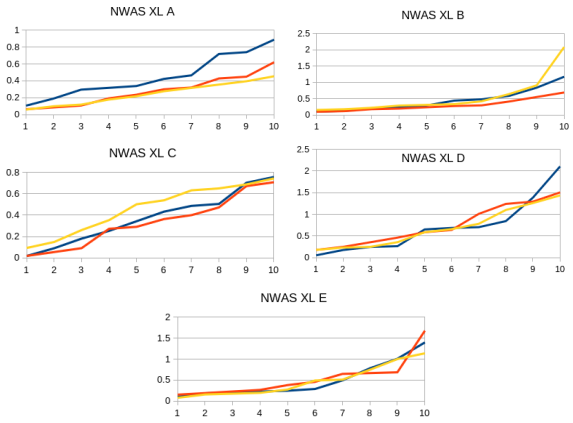


Figure 4: Visualisation of vectors with 10 highest magnitude spectral frequencies: *vectorindex* × *frequency*.

clustering of *D*-readings with *S2*, and the next-level cluster with *S1* appear to show native-like low accentedness (high fluency or naturalness) for *S3*.

Figure 4 shows the 15 extracted spectral peak vectors from the *S3* readings, to enable judgments of subjective plausibility of the distance metric criterion by means of visual inspection. All 3 *D*-readings appear close, while other categories tend to be close pairwise but not overall. A plausible interpretation of this result for the *D*-format style is that bolded stressed vowels constrain consistent conformance with English stress-based timing. It is tentatively claimed, therefore, that in reading aloud (1) spectral features help to determine relative fluency, (2) stressed vowel marking can have a positive effect on consistency of L2 speech timing.

6. CONCLUSION

The present study applied a new automatic acoustic phonetic method to a small but crucial aspect of L2 prosodic fluency which is particularly important for speakers whose L1 and L2 rhythms are radically different, for example Mandarin, which is generally analysed as more syllable-timed, and English, which is generally analysed as more stress-timed.

The results of the present exploratory study clearly need to be calibrated with other methods and scaled up with more data. Analysis of a large classroom data set (150 students) is in progress, and evaluation of results against expert rater judgments.

Further use cases for the method are potentially in components of automatic fluency (self-)assessment and (self-)training apps for medium to advanced L2 teaching, in which further refinement of the method to include more detailed identification of spectral components and adaptation to other language pairs will be required.

7. REFERENCES

[1] D. Abercombie, *The North Wind and the Sun*. University of Edinburgh. School of Philosophy, Psychology, and Language Sciences. Department of Linguistics and English Language, 2013, 1951-1978 [sound]. [Online]. Available: https://doi.org/10.7488/ds/157

[2] J. Barnes and S. Shattuck-Hufnagel, Eds., *Prosodic Theory and Practice*. Cambridge MA: MIT Press, 2022.

[3] BDAtechnology, "Typefaces for dyslexia," 2015, https://bdanewtechnologies.wordpress.com/.

[4] Y. Chen, "A study of rhythmic categorization of stress recurrence in learners' English reading aloud," *Foreign Languages and Foreign Language Teaching*, vol. 3, no. 3, 2015, metadata translated from Chinese.

[5] K. Cheng and Y. Deng, "Preliminary research on relationships between stress awareness and English proficiency among Chinese learners: A case study of Southwest Jiaotong University," *Journal of Southwest Jiaotong University (Social Sciences)*, vol. 16, no. 6, pp. 21–27, 2015.

[6] E. Couper-Kuhlen, *English Speech Rhythm: Form and Function in Everyday Verbal Interaction*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2018.

[7] M. Dehghani, P. Khooshabeh, A. Nazarian, and J. Gratch, "The subtlety of sound: Accent as a marker for culture," *Journal of Language and Social Psychology*, pp. 1–20, 2014.

[8] D. Gibbon, "The rhythms of rhythm," *Journal of the International Phonetic Association, First View*, pp. 1–33, 2021.

[9] ——, "RFAGUI," 2023, Open Source. [Online]. Available: https://github.com/dafyddg/RFAGUI/

[10] D. Gibbon, R. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for spoken Language Systems*. Berlin: Mouton de Gruyter, 1997.

[11] R. F. Hudson, H. Lane, and P. C. Pullen, "Reading fluency assessment and instruction: What, Why and How?" *The Reading Teacher*, vol. 58, no. 8, pp. 702–714, 2005.

[12] H. Kallio, *The Prosody Underlying Spoken Language Proficiency: Cross-lingual investigation of non-native fluency and syllable prominence*. Helsinki: University of Helsinki, 2022.

[13] D. Konza, *Research into practice. Understanding the reading process*. Department of Education and Children's Services, Government of South Australia, 2011, vol. 1.

[14] ——, "Teaching reading: Why the 'Fab Five' should be the 'Big Six'," *Australian Journal of Teacher Education (Online)*, vol. 39, no. 12, 2014.

[15] S. Marzban and Gyöngyi Fábián, "Second language learners' reading strategies. the case of intersemiotic relations," *Argumentum*, vol. 18, pp. 392–409, 2022.

[16] S. Nakamura and Y. Sagisaka, "A requirement of texts for evaluation of rhythm in English speech by learners," in *17th International Congress of Phonetic Sciences*, I. P. Association, Ed. Hong Kong: International Phonetics Association, 2011, pp. 1438–1441.

[17] C. Singleton, *Intervention for dyslexia: A review of published evidence on the impact of specialist dyslexia teaching*. Steering Committee for the "No to Failure" project, Department for Children, Schools and Families, UK, 2009.

[18] A. Suni, H. Kallio, Štefan Benuš, and J. Šimko, "Characterizing second language fluency with global wavelet spectrum," in *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne: International Phonetic Association, 2019, pp. 1947–1951.

[19] A. C. Teacher, "Text marking recommendation," https://www.bilibili.com/video/BV1D14y1s7cx/, 2022, social media video in Chinese.

[20] R. I. Thomson, "Fluency," in *The Handbook of Pronunciation*, M. Reed and J. M. Levis, Eds. Hoboken NJ: Wiley, 2015, pp. 209–226.

[21] S. Tilsen and K. Johnson, "Low-frequency Fourier analysis of speech rhythm," *Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 34–39, 2008.

[22] S. Wallace, Z. Bylinskii, J. Dobres, B. Kerr, S. Berlow, R. Treitman, N. Kumawat, K. Arpin, D. B. Miller, J. Huang, and B. D. Sawyer, "Towards individuated reading experiences: Different fonts increase reading speed for different individuals," *ACM Transactions on Computer-Human Interaction*, vol. 29, no. 4, pp. 1–56, 2022.

[23] A. L.-C. Wang, "An industrial strength audio search algorithm," in *Proceedings of ISMIR 2003, 4th International Conference on Music Information Retrieval*, Baltimore MA, 2003.

[24] L. White and Z. Malisz, "Speech rhythm and timing," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 167–182.

[25] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.

[26] J. Yu, "Timing analysis with the help of SPPAS and TGA tools," in *Proceedings of the Tools and Resources for the Analysis of Speech Prosody Workshop*. Aix-en-Provence: Laboratoire Parole et Langage, 2013.

EXPLAINING VOICE CHARACTERISTICS TO NOVICE VOICE PRACTITIONERS — HOW SUCCESSFUL IS IT?

Jana Wiechmann[1], Frederik Rautenberg[2], Petra Wagner[1], Reinhold Häb-Umbach[2]

[1]Phonetics Work Group, Faculty of Linguistics and Literary Studies, Bielefeld University,

[2]Department of Communications Engineering, Paderborn University
jana.wiechmann@uni-bielefeld.de, rautenberg@nt.upb.de, petra.wagner@uni-bielefeld.de, haeb@nt.upb.de

ABSTRACT

Human voices are notoriously difficult to characterize. A suitable and consistent description of voice characteristics is crucial in many applied disciplines such as speech therapy or forensics. The present study examines the ability of novice voice practitioners (students of clinical linguistics) to characterize voices before and after an expert explanation of laryngeal, supralaryngeal and prosodic voice features. Results show that even short expert explanations lead to a higher agreement between expert and novices. Especially voice characteristics related to laryngeal and supralaryngeal settings remain a major challenge to identify. We suggest that voice conversion technology may be employed in the future to assist the explanation of voice characteristics.

Keywords: voice quality, voice characteristics, explaining, phonetics pedagogy

1. INTRODUCTION

An adequate characterization of voices and voice quality is crucial across different fields, e.g., phonetics, vocal coaching, speech therapy or forensics, and has led to different approaches for doing so [1, 2, 3]. However, the characterization of voices is not only a challenge for novices in these fields. Even experts rarely achieve a high inter-rater agreement [4, 5, 6].

One of the most popular approaches to voice characterization goes back to [7], based on which the Vocal Profile Analysis (VPA) scheme was developed [8]. The VPA focuses on the articulatory settings underlying voice characteristics, and has been adapted to simplify its usage [9, 10], for example by relying on binary rather than multilevel rating scales or reduced settings. Voice characterizations in speech therapy typically employ perceptual characterizations of voices with the help of multilevel rating scales, and focus on aspects of phonation, or voice quality, i.e., the GRBAS [11, 12] or the RBH scale in the German speaking community [13, 14]. Other approaches include prosodic features or more global parameters such as overall tension. The handling of intermittent features and of the parallel occurrence of features also differs between approaches [10, 15]. Especially in forensic phonetics, factors beyond voice quality are important. Therefore, speaking style, but also segmental aspects such as dialect and pronunciation, or even background noise in recordings, are taken into account [2].

As mentioned above, all existing perceptual assessment tools yet seem to be difficult to use in a consistent way. One reason for this might be a different understanding of the terminology, or different internal standards between raters. It has also been discussed that voice as a multidimensional phenomenon is difficult to assess in general, leading to high cognitive load. For this reason, complex assessment tools, especially those employing multilevel scales, can be overwhelming [10, 4]. To overcome some of these difficulties, calibration or training sessions on example voices were suggested [16] and the usage of prototypical anchor voices for individual voice features [4, 16]. Given that even prototypical voices are likely to differ in multiple dimensions of voice, a recent follow-up idea of this is the usage of state-of-the-art voice conversion technology to provide pairs of example voices differing only in one feature of interest [17]. Despite the above-mentioned difficulties, it is crucial for voice practitioners to learn about voice characterizations as part of their profession. We want to understand better, which aspects need to be improved in the training of voice characterizations, and what can be reached in single explanation sessions. The aim of this study therefore is

1. to assess if a short, single expert explanation using prototypical voices and imitation can help in learning how to classify voices,

2. to evaluate if the explanation leads to a higher consistency among raters regarding individual voices,
3. to discover which voice characteristics are easier or more difficult to classify.

To this end, we asked novice voice practitioners to characterize voices before and after a short, systematic, expert explanation, and compared their performances to those of experts.

2. METHODS

2.1. Selection of voice characteristics

We defined a catalogue[1] of 20 voice features to be explained and assessed (cf. Table 1). We decided to take into account a wide set of voice characteristics, including laryngeal settings related to phonation, supralaryngeal settings related to the resonance characteristics of the vocal tract, as well as prosodic characteristics related to (dynamic) loudness and pitch. Our selected features needed to fulfill two main criteria:

1. They can be expressed by fairly common adjectives in German (our participants' native language), to guarantee their fundamental conceptual familiarity to non-experts.
2. They have fairly well-understood underlying articulatory settings and perceptual characteristics, which are necessary for their explanation.

2.2. Participants and speech materials

20 students of clinical linguistics (female, mean age = 22.5) participated in the study and received monetary compensation. 20 voices (10 female, 10 male) were selected from The Nautilus Speaker Characterization (NSC) Corpus [18] for classification. The set of chosen voices covered the full set of characteristics in our feature set, and included 2 relatively "neutral" voices. From the corpus, semi-spontaneous simulated telephone calls of around 30 seconds were used for characterization. While none of the voices can be considered pathological, some of them contain characteristics that can feature in pathological voices.

2.3. Expert rater agreement and gold standard

To set a gold standard for voice characterization, two voice experts classified the 20 voices independently with respect to the 20 voice features in a binary manner. Afterwards, they jointly discussed controversial cases, to develop a more common understanding of the feature set. Subsequently, the experts rated the voices again. Inter-expert agreement before the discussion was Cohen's $\kappa = 0.61$, corresponding to 'substantial agreement'. After the discussion, their agreement was 'almost perfect' ($k = 0.85$). The ratings of the first author and expert explainer, were lastly defined as the gold standard (also cf. Table 1) to which the participants' ratings were subsequently compared.

2.4. Study phases

The study was divided into three steps: (1) A pre-explanation phase, in which participants classified a set of 10 voices, (2) an explanation phase, during which the expert explained the selected voice features (cf. Section 2.1), and (3) a post-explanation phase, during which participants rated the same set of voices again. During the voice ratings (pre- and post-explanation phase), each participant listened to 10 voices (out of 20) in a pseudorandomized order (each participant started with a different voice, male and female voices alternated). Voices were distributed across participants so that every voice was classified 10 times before and 10 times after the explanation. Participants were allowed to listen to each voice twice if necessary, and were instructed to mark a voice feature on a prepared list if they perceived it, even if it was only present temporarily or not very prominent. There was no minimal or maximal amount of features that listeners could select. Participants could also add features that they heard, but which were not listed. During the explanation phase, the expert explained how the features emerge anatomically, and how they manifest acoustically and perceptually. Additionally, the expert imitated the voice characteristics and played prototypical examples. To increase consistency across explanations, the expert followed a script, but explanations remained spontaneous and interactive: Participants were allowed to ask clarification questions, comment and provide feedback. The explanations lasted around 15-20 minutes each. Once the participants expressed that they felt confident enough to perform the classification task again, the post-explanation phase was initiated.

2.5. Rater agreement and consistency analysis

We used Cohen's κ to assess (1) the intra-rater agreement among participants before and after the explanation as well as (2) the agreement between the individual participants and the gold

Class	Voice Characteristics
laryngeal	creaky (10), hoarse (7), breathy (4), whispery (3), cracked (2), harsh (2), soft (2)
supralaryngeal	dark (2), bright (3), nasal (4), retracted tongue and/or raised larynx (4)
prosodic (pitch)	low (2), high (4), highly variable pitch (3)
prosodic (loudness)	loud (3), quiet (4), highly variable loudness (1)
mixed	shrill (1), resonant (4), tense (2)

Table 1: List of perceptual voice characteristics (English translations used in the perception study, classified into largyngeal, supralaryngeal, prosodic (pitch and loudness related) and mixed. Numbers in brackets indicate the total number of occurrences of each voice characteristic (based on the gold standard for our data set, cf. Section 2.3)

.

standard before and after the explanation. To assess the consistency among raters regarding the individual voices and voice characteristics, Fleiss' κ was calculated. Agreement was interpreted according to [19], differentiating the following agreement categories: 'almost none', 'slight', 'fair', 'moderate', 'substantial', 'almost perfect'. The statistical analyses were carried out using the *irr* package in R [20, 21].

3. RESULTS

3.1. Participants intra-rater agreement

We calculated the intra-rater agreement for each participant which varied from $\kappa = 0.27$ to $\kappa = 0.62$ ($M = 0.49$, $SD = 0.092$). Most raters achieved 'moderate agreement', one 'substantial', and three 'fair agreement' comparing pre and post per participant.

3.2. Participant-expert agreement

Twelve participants show a higher agreement with the gold standard after the explanation, two remain identical (while not showing a high intra-rater consistency, either: P15: $k = 0.48$, 'moderate', P18: $k = 0.27$, 'fair'), and six even decrease in their agreement with the expert (cf. Figure 1). The mean agreement with the gold standard increases minimally but not significantly (mean $\kappa = 0.26$ pre-explanation, mean $\kappa = 0.27$ post-explanation).

3.3. Consistency analysis per voice

Figure 2 shows the agreement across all participants per voice. Thirteen voices increase in agreement after the explanation, while seven voices decrease. Mean agreement across all voices before the explanation is $\kappa = 0.22$ and after $\kappa = 0.23$, increasing minimally but not significantly. Three

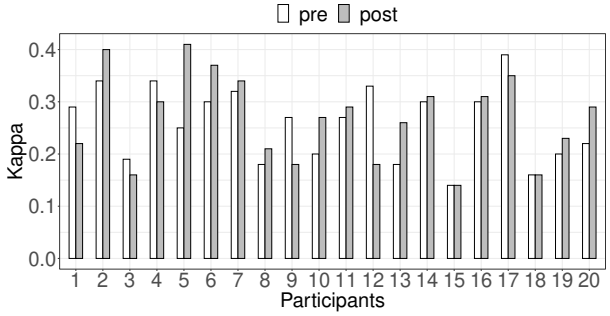


Figure 1: Agreement (Cohen's κ) between each participant and the expert before and after the explanation.

voices decrease in agreement, and even changed the agreement category after the explanation: Two of them went from 'moderate' to 'fair' and one from 'fair' to 'slight'. Five voices increased in agreement and changed the agreement category: One voice went from 'fair' to 'moderate', all others went from 'slight' to 'fair'. The remaining voices stayed within the same category (most of them 'fair'), one of them showing substantial improvement within the category. All other voices show little change. The voice showing the highest agreement score reaches this after the explanation ('moderate').

3.4. Consistency analysis per voice feature

Table 2 shows the agreement across all participants for features showing the lowest and highest agreement, or the strongest changes. In total, twelve features (out of twenty) have higher agreement after the explanation while eight have lower or almost identical agreement. Mean agreement across all voice features before the explanation is $\kappa = 0.22$, and after it $\kappa = 0.23$, showing minimal increase without being significant. With respect to agreement, two features changed the category: 'creaky' decreased from 'fair' to 'slight' and 'tense'

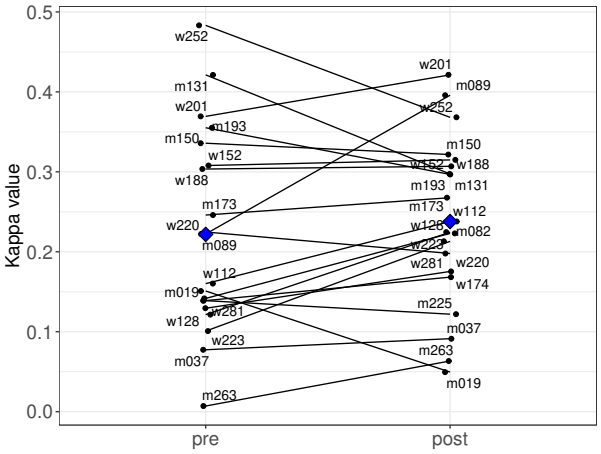


Figure 2: This figure shows the agreement (Fleiss' κ) across all participants per voice pre and post explanation. Means are indicated in blue.

Agreement	Feature	Pre	Post
highest	quiet	0.65	0.69
	loud	0.41	0.45
	high	0.32	0.35
lowest	low	0.13	0.11
	cracked	0.11	0.1
	breathy	0.1	0.09
highest increase	tense	0.03	0.21
	nasal	0.07	0.19
highest decrease	creaky	0.22	0.16
	hoarse	0.38	0.27

Table 2: This table shows the overall agreement (Fleiss' κ) of voice features before (pre) and after (post) the explanation, that showed the highest or lowest overall agreement, or the strongest increase or decrease in agreement.

increased from 'slight' to 'fair'. All other voice characteristics stayed within the same agreement category. Ten features stay within the category 'slight agreement', while six stay within 'fair'. Besides the features mentioned in Table 2, 'shrill', 'soft' and 'resonant' show a comparatively higher agreement than other voice features (above or equal to $\kappa = 0.26$).

4. DISCUSSION AND CONCLUSIONS

The first research question raised in this study was whether a single expert explanation of complex voice characteristics can actually help novice voice practitioners. The results show a clear tendency that the explanations indeed helped the majority of participants. Probably, the short amount of time available for the explanations has prevented a better success. Obviously, we cannot generalize our results to other settings, and the impact of the individual expert is left unexplored. With respect to our second question, our analysis revealed that the agreement differs vastly across different voices. This may have been caused by different levels of agreement for individual voice characteristics (research question 3). In fact, agreement is not distributed uniformly across these. Rather, some features were much more difficult to classify than others. In particular, our results show that prosodic features are rated much more consistently, with the notable exception of 'low'. Besides, 'hoarse', 'resonant' and 'shrill' achieved relatively high agreements, while 'tense' and 'nasal' showed a strong increase after the explanation. We argue that most of the characteristics that show high agreement/improvement are at least partially characterized by their supralaryngeal resonance characteristics ('tense', 'nasal') and/or prosodic features such as loudness ('shrill', 'resonant'). Purely laryngeal features show less agreement. 'Hoarse', a voice feature rather prominent in everyday experience and well represented in our data set, is an exception. Interestingly, it showed a strong decrease in agreement after the explanation. Another laryngeal feature, 'creaky', is also well represented in our data set, but decreases from 'fair' to only 'slight' agreement after the explanation.

Overall, laryngeal and, to a slightly lesser degree, supralaryngeal features seem to be more difficult to detect than prosodic ones. This analysis is difficult to uphold when looking at individual voices, some of which showed a high decrease in agreement despite being characterized by voice features which should be "easy" to classify - or vice versa. This leads us to conclude that tracing individual characteristics of voices may be a different challenge if these appear in combination. However, the complex potential combinations of voice features cannot be easily imitated or straightforwardly explained, not even by an expert, or covered by prototypical example voices. Summing up, a short explanation using prototypical examples and imitations of voice characteristics helps, but does not achieve a high agreement among participants. We next plan to examine whether explanatory success can be further enhanced by state-of-the-art voice conversion technology, which disentangles the relevant dimensions of voice characteristics, and may imitate and recombine these with a higher consistency than expert explainers.

ACKNOWLEDGEMENTS

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824

5. REFERENCES

[1] J. Kreiman and D. Sidtis, "Voices and listeners: Toward a model of voice perception," *Acoustics Today*, vol. 7, no. 4, pp. 7–15, 2011.

[2] W. J. Hardcastle and J. M. Beck, "Forensic speaker identification and the phonetic description of voice quality," in *A Figure of Speech*. Routledge, 2014, pp. 425–452.

[3] S. S. Hammer and A. Teufel-Dietrich, *Stimmtherapie mit Erwachsenen: was Stimmtherapeuten wissen sollten*. Springer-Verlag, 2017.

[4] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[5] J. Kreiman and B. R. Gerratt, "Sources of listener disagreement in voice quality assessment," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1867–1876, 2000.

[6] A. Webb, P. Carding, I. J. Deary, K. MacKenzie, N. Steen, and J. A. Wilson, "The reliability of three perceptual evaluation scales for dysphonia," *European Archives of Oto-Rhino-Laryngology and Head & Neck*, vol. 261, no. 8, pp. 429–434, 2004.

[7] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[8] J. Laver, S. Wirz, J. Mackenzie, and S. Hiller, "A perceptual protocol for the analysis of vocal profiles," *Edinburgh University Department of Linguistics Work in Progress*, vol. 14, pp. 139–155, 1981.

[9] J. M. Beck, "Vocal profile analysis scheme: A userâs manual," *Queen Margaret University, Edinburgh*, 2007.

[10] E. San Segundo and J. A. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *Journal of Voice*, vol. 31, no. 5, pp. 644–e11, 2017.

[11] M. Hirano, "Objective evaluation of the human voice: clinical aspects," *Folia Phoniatrica et Logopaedica*, vol. 41, no. 2-3, pp. 89–144, 1989.

[12] B. Ann and R. No, "Grbas evaluation of running speech and sustained phonations," *Ann, Bull. RILP No*, vol. 28, pp. 51–56, 1994.

[13] J. Wendler and L. C. Anders, "Hoarse voices-on the reliability of acoustic and auditory classifications," in *FOLIA PHONIATRICA*, vol. 38, no. 5-6. KARGER ALLSCHWILERSTRASSE 10, CH-4009 BASEL, SWITZERLAND, 1986, pp. 369–369.

[14] J. Wendler, W. Seidner, G. Kittel, and U. Eysholdt, "Lehrbuch der Phoniatrie und Pädaudiologie 3. aufl. s 323-364," 1996.

[15] E. San Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, "The use of the vocal profile analysis for speaker characterization: Methodological proposals," *Journal of the International Phonetic Association*, vol. 49, no. 3, pp. 353–380, 2019.

[16] T. L. Eadie and C. R. Baylor, "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," *Journal of voice*, vol. 20, no. 4, pp. 527–544, 2006.

[17] J. Wiechmann, T. Glarner, F. Rautenberg, P. Wagner, and R. Häb-Umbach, "Technically enabled explaining of voice characteristics," in *P&P 18 Abstract Booklet*. Bielefeld University, Germany, 2022. [Online]. Available: https://doi.org/10.11576/pundp2022-1038

[18] L. F. Gallardo and B. Weiss, "The nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374, 1977.

[20] M. Gamer, J. Lemon, and I. F. P. Singh, *irr: Various Coefficients of Interrater Reliability and Agreement*, 2019, r package version 0.84.1. [Online]. Available: https://CRAN.R-project.org/package=irr

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[1] The text uses the English translations of the following German original descriptions: knarrend (creaky), heiser (hoarse), behaucht (breathy), geflüstert (whispery), brüchig (cracked), rau (harsh), gepresst (tense), klangvoll (resonant), sanft (soft), dunkel (dark), hell (bright), nasal (nasal), kehlig und/oder rückverlagert (retracted tongue and/or raised larynx), hoch (high), tief (low), laut (loud), leise (quiet), schrill (shrill), starke Lautstärkenvariabilität oder Tonhöhenvariabilität (variable loudness or pitch)

AN INCLUSIVE COURSE ON THE HISTORY OF PHONETICS

Marc Garellek, Shai Nielson, Tamara L. Rhodes, Emily Clem

University of California, San Diego
mgarellek@ucsd.edu; slnielso@ucsd.edu; tlrhodes@ucsd.edu; eclem@ucsd.edu

ABSTRACT

This paper describes a 10-week seminar-style course on the history of phonetics that was designed and implemented with the goal of promoting social justice, diversity, equity, and inclusion in the discipline. Advanced planning for the seminar involved coordination between the instructor and the academic department, the linguistics librarian, and the university pedagogical resource center. In this paper we provide the motivation for the framework and the topics covered and link to the materials from the seminar. Overall, this paper shows how a course on the history of phonetics can be modeled and taught in such a way as to shed light on the ways in which our field has marginalized and discriminated against certain communities, but also to highlight the diversity in our discipline and among its practitioners, as well as past and present efforts at creating a more just and inclusive discipline.

Keywords: Phonetic pedagogy, History of phonetics, Antiracist linguistics, JEDI in linguistics

1. INTRODUCTION

There are many reasons for studying the history of a discipline; learning about our field's history provides entertainment value, satisfies our curiosity, and helps us gain new perspectives [1]. Studying our history informs instructors and students alike about how and why we arrived at the current state of our discipline, and where we might go from here.

It was partly for these reasons that a coauthor of this paper, Marc Garellek, a phonetician who knew little history of the field, decided to teach a seminar on the history of phonetics. Yet there are also moral and philosophical (as well as scientific) lessons to be derived from the history of the phonetic sciences. With regards to the moral and philosophical lessons in studying the history of phonetics, Ohala et. al [1] focus on the potential for inspiration from our field's early leaders and on rethinking common assumptions. In designing this seminar, we considered both those questions, as well as others like: How was modern phonetic inquiry able to develop in the ways that it did? Whose voices were suppressed in the process of these inquiries? In the present and future, how can we be more moral practitioners of phonetics? In other words, we designed the course so as to study the discipline's history, while at the same time – and with equal focus – studying its history with an eye towards promoting the values of social justice, equity, diversity, and inclusion (henceforth 'JEDI') within and beyond our discipline.

Courses designed within antiracist and inclusive pedagogical frameworks (e.g. [2, 3]) challenge both instructors and students to consider how we can center the lives and experiences of people of various backgrounds, particularly those who are marginalized and whose lives and experiences have gone under-appreciated or outright ignored. Discussions about racial justice and equity have lately intensified, both in society in general and in linguistics specifically (e.g. [4, 5, 6]). Learning from those discussions inspired Garellek to develop a course on the history of phonetics that was designed within these frameworks. This paper thus describes the planning and implementation of a seminar-style course on the history of phonetics, with a focus on the period between the mid-1800s to mid-1900s. We hope that this paper and its accompanying course materials may serve as one possible model for other instructors who are considering teaching a course on the history of our discipline.

2. PREPARING THE SEMINAR

The course was taught in Winter quarter of 2022, with planning beginning in the previous Fall quarter. While Garellek had spent time over the previous summer reading about the philosophy behind and tools used in antiracist pedagogy, he also knew it was important to engage with departmental and university resources for more guidance and support. To do so, he coordinated with various people and groups at the university: (a) the Department of Linguistics' Curriculum Committee, (b) the departmental Graduate Student DEI committee, (c) the linguistics librarian, and (d) the university's pedagogical resource center. Coordination with these individuals or groups is described below.

2.1. The departmental Curriculum Committee and Graduate Student DEI committee

The Department of Linguistics at UC San Diego has a Curriculum Committee (including several faculty members and a graduate student) that advises on linguistic pedagogy and changes to the curriculum. During the term before the seminar was scheduled, Garellek involved the Curriculum Committee, as well as the recently-formed Graduate Student DEI committee, so as to get feedback and brainstorm ideas for topics and for the implementation of the course. Moreover, two members of these committees, coauthors Shai Nielson (graduate student) and Emily Clem (faculty member) were part of a campus-wide Anti-Racist Pedagogy Learning Community and thus had deeper familiarity with the literature on inclusive pedagogical practices. Nielson is a hearing researcher of sign language phonetics and phonology, and ensuing discussions between her and Garellek were instrumental for considering how phonetics engages with, and can engage better with, deaf communities, as well as how phonetics courses can better recognize and incorporate sign languages.

2.2. The linguistics librarian

During the planning of the seminar, Garellek wanted to assemble the materials into a collection of readings about JEDI issues and pedagogy in linguistics more broadly, so that the materials could be searchable in the future, and moreover would apply to linguistics more broadly, rather than being specific to phonetics and speech research. This intersected well with coauthors Nielson and Clem's plans to build a repository of antiracist teaching materials and guides for the department. (Creation of the general repository is still ongoing but is expected to be made publicly available in the near future.) To facilitate material curation and design of the repository, Garellek approached the linguistics librarian (and coauthor of this paper) Tamara L. Rhodes. Garellek and Rhodes initially met to discuss the project and its motivation, and review the outline for the class and some best practices in antiracist pedagogy, such as the inclusion of clear learning outcomes, writing of positionality statements at the start of the term, and structuring weekly meetings such that JEDI issues are treated as integral to the covered topics. As a librarian, Rhodes was able to offer not only expertise on resources, but also course design. Additionally, it is important to note that as a Black woman, her lived experience and experience with JEDI issues and pedagogy helped shape this work. After their initial discussion, the two also met with other members of the departmental committees mentioned above to discuss the design and implementation of the resource devoted to antiracist pedagogy in a wide range of linguistics classes.

2.3. The university's pedagogical resource center

The final person whom Garellek met with when planning the course was an advisor from the university's pedagogical resource center, which works with instructors to better integrate evidence-based practices that support teaching and building of effective learning environments. Together they reviewed all materials for the start of the course: the syllabus design, course introduction surveys and ethics approval (so that enrolled students' feedback could be used for eventual publication on the experience of teaching this course), and final project ideas. The advisor suggested use of a "Universal Design for Learning" approach [7], which encourages educators to think of accessible course design and student flexibility and choice. Garellek then shifted from a required final paper to a project that allowed more flexibility for students who wished to create work like as a website, podcast, or annotated bibliography. Most students opted to write a final research paper in the end.

Thus, discussion with multiple people with differing expertise was instrumental in Garellek's designing the class with inclusive pedagogical best practices, broadened his perspectives about topics that could be discussed when it comes to teaching about phonetics in a more inclusive way, and helped him and the Linguistics Department build a repository of course materials with a focus on JEDI perspectives.

3. COURSE COMPOSITION

3.1. Participants

In total 8 students enrolled in the course: 6 undergraduate students in their final year of the 4-year undergraduate B.A. degree in Linguistics, and 2 graduate students in Linguistics in their 2nd and 4th year in the 5-year Ph.D. program. Two additional graduate students (in their 5th and 7th year) audited the course and attended roughly half of the lectures; moreover, one faculty member specializing in phonology attended most lectures, and another faculty member specializing in phonetics attended roughly half of the lectures. To promote participation, Garellek made clear that

he was a fellow learner, as he was new to both the history of phonetics and antiracist pedagogy.

3.2. Structure of the course

The course included one weekly lecture over a 10-week quarter. The course can be viewed at [8], with relevant links and full citation entries for cited books such that the reader can easily find the cited work.[1]

Lectures were held on Wednesdays, during which the participants met in person to discuss assigned readings and work. Garellek had prepared a digital handout for each lecture to help guide the conversation and as a reminder to cover some key points. (On the legacy page viewable to readers [8], copyright-protected material has been stripped from the lecture handouts.) During the first lecture of the term, we spent some time discussing the antiracist pedagogical framework for the course, the learning outcomes, as well as community guidelines (for how we should all interact as a group) and links to student support (as conversations around marginalized communities can have unintended negative effects on student members of those and other marginalized communities). We further discussed how one goes about reading about and researching the history of a discipline using reviews and guidelines by historian-linguists (e.g. [9, 10]).

Prior to each lecture, students were asked to provide short essay-type answers to a handful of questions based on required readings. These ensured that students came prepared to discuss the weekly material. The required readings and question prompts were chosen so as to be as inclusive as possible: readings were often written by scholars of (historically and/or currently) underrepresented groups, and at least some of the reading questions were designed to promote consideration of JEDI issues in phonetics. For example, in Week 5 (devoted to the IPA–both the Association and Alphabet), students prepared for the weekly lecture by reading and comparing the first and last issues of *Le Maître phonétique* (the *mf*) and first and most recent issue of *JIPA* and reflecting on changes in scope, topics, as well as changes in membership of the Phonetic Teachers' Association/IPA and authors of the articles in the Journal; see snapshot from course page in Figure 1. These prompts were accompanied by a quotation from the relevant work to help stimulate students' thinking about the topic. Student answers were graded for completion only; however, the instructor also provided feedback and occasional requests for more detailed responses.

On the course page, each week included a tab for the assigned readings, another that included

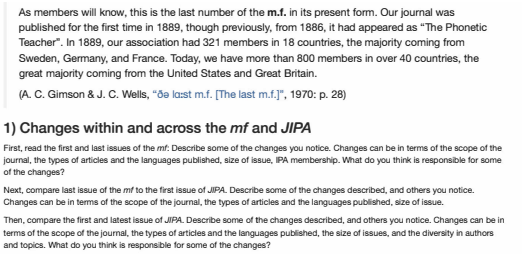


Figure 1: Snapshot of the weekly reading prompt. For the week devoted to the IPA, the class analyzed changes in the *mf* and *JIPA*.

the handout for lectures, and a "Perspectives" tab that further highlighted JEDI-related topics and issues related to the covered topic (and which were discussed during lectures). For example, during the same week devoted to the IPA, the instructor provided a short quotation in the 1904 "Aims and Principles of the IPA" that highlights the choice of Roman letters in the Alphabet due to their use by "the great majority of civilized nations" [11, p.6]. We reflected on the biases intrinsic to the IPA, as well as how the IPA has more recently improved inclusivity through the development of the Braille IPA [12, 13] and extIPA for atypical or "disordered" speech sounds [14] (see snapshot in Figure 2). For each week, the course page also included a "References" tab, with links to all cited work as well as to work that students were encouraged to read more about, and a "Bios" tab devoted to biographies of the figures discussed in a given week, so that we could all learn more about the lives and work of the historical figures discussed.

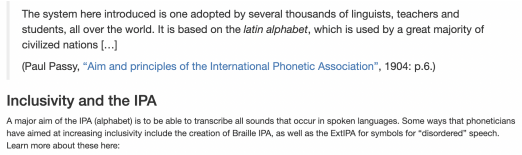


Figure 2: Snapshot the "Perspectives" topic that centered specific JEDI-related issues. For the week devoted to the IPA, the class discussed the Eurocentric origins of the IPA (as highlighted in the quote from Passy at the top) and contemporary measures that aim to improve inclusion.

3.3. Topical coverage

A major challenge to the course was deciding what topics in the history of phonetics to cover. Garellek fairly quickly settled on the time period of the mid-1800s to mid-1900s: around 1850 saw the early development of modern phonetics

through convergence of the biomedical, physical, and linguistic sciences [15] and the development of Bell's Visible Speech system of notation, and the middle of the 20th century saw the elaboration of acoustic theory of speech production, the early days of generative linguistics, and the development of the spectrograph, a new *Visible Speech* [16].

Week–Topic	Perspectives (sample)
1–Overview	Diverse origins of the field
2–Phonetic notation	Notation of signed languages
3–Pronunciation & dialects	Dialectologists documenting US slavery
4–Experimental phonetics	Phonetics in Japan
5–The IPA	Eurocentrism in the IPA Inclusion & the IPA
6–Acoustic phonetics	Women in speech acoustics
7–The spectrograph	Phonetics & the Deaf community
8–Field phonetics	Colonial & missionary fieldwork
9–Sound change & the phoneme	Chinese phonetic tradition
10–The present & future	A more diverse phonetics

Table 1: Thematic coverage of the seminar, including examples of some of the JEDI perspectives incorporated in each weekly topic.

There are currently many excellent historical reviews of phonetics (e.g. [17, 18, 19, 1, 20, 21, 22, 10]), many of which Garellek assigned readings from and all of which helped inspire him when choosing the thematic topics to be covered. He finally settled on topics shown in Table 1, with some (but not an exhaustive list) of the JEDI perspectives discussed in relation to them. As can be seen in that Table, a couple of weeks were devoted to phonetic notation; instrumental phonetics was covered both in general (in Week 4), and in particular with respect to developments in acoustic theory (Week 6) and the invention of the spectrograph (Week 7). As he was designing the syllabus, Garellek struggled somewhat in deciding *when* during the quarter to cover a given topic. Some topics were clearer than others: discussion of precursors to the IPA (Week 2) had to precede both the week devoted to pronunciation and dialects (Week 3) and the IPA and its Association (Week 5); the week devoted to acoustic theory (Week 6) had to precede the invention of the spectrograph (Week 7). Additionally, Garellek struggled with coverage during Week 3 (pronunciation and dialects), as he tried to cover the early days of pronunciation training, dialectology, and part of the discussion on atypical speech and the development of the field of speech-language pathology. In the future, he would consider moving dialectology to the week on sound change; over a longer term, he would split up these topics to allot each its own week.

One might worry that finding JEDI perspectives on a given topic in the history of phonetics could prove challenging, but the process proved highly stimulating. For example, for field phonetics and documentation (Week 8), we discussed Daniel Jones and Sol Plaatje's work on Setswana [23], their collaboration [19, 24, 25], and the historical factors that influenced Plaatje both as a phonetician as well as a politician. The process was also easily tailored to the local context: we further compared the ethics and framing of early fieldwork on Kumeyaay/Kumiai, an Indigenous language of San Diego, to contemporary phonetic research done by members of the department (including a graduate student who was auditing the course). And when discussing the spectrograph and *Visible Speech* [16], we highlighted the role of coauthor Harriet Green Kopp in the development of the local Speech, Language and Hearing Sciences school at San Diego State University [26]. The challenge then was not finding a single perspective to highlight, but which of the many possible perspectives to include.

4. CONCLUSION

We described the planning and structuring of a history of phonetics seminar that was implemented within an inclusive pedagogical framework. This required a team effort, which greatly enriched the experience for the instructor and participants alike. Indeed, undergraduate and graduate student participation was consistently high, and both the class exit survey and official course evaluations included only positive feedback.

We hope that this paper and course materials help our research community consider (1) how phonetics has changed in terms of its engagement with marginalized communities and researchers; (2) how to reckon with the historical exclusion of these communities and researchers, and find ways to more actively include these voices; and (3) how we can teach about our discipline in ways that promote social justice, equity, diversity, and inclusion.

5. ACKNOWLEDGEMENTS

For their helpful suggestions for the design and implementation of the course, we thank Paul Hadjipieris, Yaqian Huang, Sharon Rose, and Will Styler. Marc Garellek would like to thank the students and faculty members who participated in the course in Winter 2022 and who helped ensure its success. We also thank Adam Chong for his comments on an earlier version of this paper.

6. REFERENCES

[1] Ohala, J. J., Bronstein, A. J., Busà, M. G., Lewis, J. A., Weigel, W. F. (eds). 1999. *A guide to the history of the phonetic sciences in the United States*. University of California Press.

[2] Blakeney, A. M. 2005. Antiracist pedagogy: definition, theory, and professional development. *Journal of Curriculum and Pedagogy* 2, 119–132.

[3] Kishimoto, K. 2018. Anti-racist pedagogy: from faculty's self-reflection to organizing within and beyond the classroom. *Race Ethnicity and Education* 21, 540–554.

[4] Charity Hudley, A. H., Mallinson, C., Bucholtz, M. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language* 96, e200–e235.

[5] Charity Hudley, A. H., Mallinson, C., Bucholtz, M. 2020. From theory to action: Working collectively toward a more antiracist linguistics (response to commentators). *Language* 96, e307–e319.

[6] Sanders, N., Konnelly, L., Umbal, P. Forthcoming. An action-based roadmap for equity, diversity, and inclusion in teaching linguistics. In: Charity Hudley, A. H., Mallinson, C., Bucholtz, M. (eds), *Inclusion in Linguistics*. Oxford University Press. Preprint: http://sanders.phonologist.org/Papers/sanders-etal-edi-roadmap.pdf

[7] Burgstahler, S. E., Cory, R. C. (eds). 2008. *Universal design in higher education: from principles to practice*. Harvard Education Press.

[8] Garellek. M. History of phonetics: The mid-1800s to mid-1900s. https://pages.ucsd.edu/~mgarellek/History/

[9] Law, L. 2003. *The history of linguistics in Europe: from Plato to 1600*. Cambridge University Press.

[10] Ashby, M. G. 2016. Experimental phonetics in Britain, 1890–1940. D. Phil. thesis, University of Oxford.

[11] Passy, P. 1094. Aims and principles of the International Phonetic Association. *Le Maître Phonétique* 19(supplement), 1–20.

[12] Wells-Jensen, S. 2005. The Braille International Phonetic Alphabet and other options: the blind student in the phonetics classroom. *Journal of the International Phonetic Association* 35, 221–230.

[13] Englebretson, R. 2009. An overview of IPA Braille: An updated tactile representation of the International Phonetic Alphabet. *Journal of the International Phonetic Association* 39, 67–86.

[14] Duckworth, M., Allen, G., Hardcastle, W., Ball, M. 1990. Extensions to the International Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics and Phonetics* 4, 273–280.

[15] Ashby, P., Ashby, M. 2013. Phonetic pedagogy. In: Jones, M. J., Knight, R.-A. (eds), *The Bloomsbury Companion to Phonetics*. Bloomsbury, 198–207.

[16] Potter, R. K., Kopp, G. A., Kopp, H. G. 1947. *Visible Speech*. D. Van Nostrand.

[17] Asher, R. E., Henderson, E. J. A. (eds). 1981. *Towards a history of phonetics: papers contributed in honour of David Abercrombie*. Edinburgh University Press.

[18] Koerner, E. F. K. 1993. Historiography of phonetics: the state of the art. *Journal of the International Phonetic Association* 23, 1–12.

[19] Collins, B., Mees, I. M. 1999. *The Real Professor Higgins: The Life and Career of Daniel Jones*. Mouton de Gruyter.

[20] MacMahon, M. K. C. 1986. The International Phonetic Association: The first 100 years. *Journal of the International Phonetic Association* 16, 30–38.

[21] Heselwood, B., Hassan, Z. M., Jones, M. J. 2013. Historical overview of phonetics. In: Jones, M. J., Knight, M. J. (eds), *The Bloomsbury Companion to Phonetics*. Bloomsbury, 5–20.

[22] MacMahon, M. K. C. 2013. Orthography and the early history of phonetics. In: Allen, K. (ed), *The Oxford Handbook of the History of Linguistics*. Oxford University Press, 105–122.

[23] Jones, D., Plaatje, S. T. 1916. *A Sechuana reader*. University of London Press.

[24] Willan, B. 2018. *Sol Plaatje: A life of Solomon Tshekisho Plaatje, 1876–1932*. Jacana.

[25] Sebina, B., Ashby, M. 2020. BRIAN WILLAN, Sol Plaatje: A life of Solomon Tshekisho Plaatje, 1876–1932. Auckland Park [Johannesburg]: Jacana, 2018. Pp. xxiv 711. ISBN: 9781431426447. *Journal of the International Phonetic Association* 50, 298–300.

[26] SDSU School of Speech, Language and Hearing Sciences History Project. https://slhs.sdsu.edu/history/

[1] Readers of this paper are encouraged to contact M. Garellek (mgarellek@ucsd.edu) if they have trouble accessing any of the materials, or if they spot any typos and broken links, on the course page at: https://pages.ucsd.edu/~mgarellek/History/.

INTELLIGIBILITY OF MARITIME ENGLISH TERMS PRONOUNCED BY JAPANESE DECK CADETS

Yoko Uchida[1], Junko Sugimoto[2]

[1] Tokyo University of Marine Science and Technology; [2] University of the Sacred Heart, Tokyo
uchidayo@kaiyodai.ac.jp; sugimoto@u-sacred-heart.ac.jp

ABSTRACT

Deck cadets, who are required to master a set of standardized English communication phrases for safe navigation, manifest a variety of pronunciation errors. In this study, 65 Maritime English terms spoken by 30 Japanese deck cadets were presented randomly to 65 American English listeners, and they were asked to spell out the terms in an intelligibility task. The percentage of words correctly identified as intended ranged between 1.5% and 92.3% (M = 45.8%). Monosyllabic words were found to be less intelligible than multisyllabic words, and both L1 transfer (global error) and word-level idiosyncratic errors (local error) impacted on intelligibility. The results revealed a wide range of individual speaker differences, which suggests the need for pronunciation instruction tailored to individual learners' requirements. Focusing on a specific set of words as well as segments is crucial to improve intelligibility, especially in Maritime English, a sub-category of English for Specific Purposes (ESP).

Keywords: intelligibility, Maritime English, ESP, pronunciation instruction, individual differences

1. INTRODUCTION

Maritime English is a type of English for specific purposes (ESP) "used by seafarers both at sea and in port and by individuals working in the shipping and shipbuilding industry" [2, p. 3579]. Maritime education and training institutions worldwide are required to provide instruction on Standard Marine Communication Phrases (SMCP) [4, 5, 13] for their deck cadets. They are a set of standardized phrases to be used by deck officers with different English competencies for intra-/inter-ship communications to ensure safe navigation of ocean-going vessels.

According to [6], approximately 90% of communication breakdowns stem from phonological inaccuracy, although at the same time they claim that seafarers' non-native accents are acceptable as far as they are intelligible. However, very little is known about what constitutes intelligible pronunciation.

The current study aimed to address this question. To this end, we measured the intelligibility of Maritime English terms pronounced by Japanese deck cadets. Intelligibility can be defined as "the degree of match between a speaker's intended message and the listener's comprehension" [3, p. 5] and can be measured through orthographic transcription [3]. The cadets' pronunciation errors based on previous studies [9, 14, 15] were compared with intelligibility measurements obtained from American English listeners, and the types of errors likely to have a greater impact on intelligibility were examined. Through data analysis, we also attempted to gain insights into effective ESP pronunciation instruction.

2. METHODS

2.1. Speakers and stimuli

Thirty Japanese deck cadets (all males; average age: 21.5) studying for the Certificate of Competency read aloud frequently used SMCP terms; 15 of them (labelled A01–A15) read 35 terms (List A) and the other 15 (labelled B01–B15) read 30 terms (List B), as shown below:

List A: *alarm, astern, barometer, bow, briefing, buoy, call, cargo, clear, container, control, deck, emergency, freeboard, full, hold, iceberg, information, kilometer, launch, liferaft, meter, mile, oil, order, port, pump, radar, request, rudder, smoke, starboard, tool, trim, turbine*

List B: *ahead, anchor, ballast, berth, boat, bridge, channel, course, crew, damage, diesel, draft, ETA, fairway, ladder, lifeboat, light, local, message, officer, pilot, pirate, proceed, propeller, ship, signal, speed, system, team, tug*

The terms were recorded and converted into sound files. The files were split into smaller ones, each containing one token. There were 65 words in total, with 15 tokens collected per word. For details of the speakers and stimuli selection, refer to Uchida and Sugimoto [15], who analyzed and described the cadets' pronunciation errors.

2.2. Participants

Sixty-five monolingual American English speakers (20 female, 45 male; age range: 22–70; average age: 37.0), recruited through Prolific (www.prolific.co),

completed two sets of tasks. Their average familiarity with Maritime English was 1.71 (*SD* = 1.01) and their average familiarity with Japanese-accented English was 2.40 (*SD* = 1.00), each of which was scored on a scale of 1 (*not familiar at all*) to 5 (*very familiar*). All gave online consent prior to participation.

2.3. Procedure

Intelligibility and comprehensibility rating tasks were created using Gorilla Experiment Builder [1].

The participants were instructed to wear headsets or earphones and listen to stimuli in a quiet room. After completing a questionnaire to provide their profile and language background, participants were introduced to the intelligibility task. A set of 65 tokens spoken by various speakers and 6 distracters spoken by a male American speaker were randomly presented one by one. Participants were asked to type the words that they thought were pronounced. They were allowed to listen to each token only once and were told to make their best guess if uncertain. Five practice trials preceded the actual test.

Upon completion of the first task, participants were encouraged to take a short break, followed by a comprehensibility rating task. At the end of the two tasks, a list of maritime terms used in the study was shown on the screen, and the participants were asked to choose terms unfamiliar to them. Responses of terms that participants reported as unfamiliar were excluded from the data set.

It took 20.5 minutes on average for the participants to complete the study. Only the results of the intelligibility task are analyzed and reported here.

3. RESULTS

This section reports on the results of intelligibility task.

3.1. Analysis by word

The following shows words that had high (over 80%) and low (under 20%) correct response rates.

Over 80% correct response (8 words): *information* (92.3%), *smoke* (92.3%), *container* (89.2%), *control* (86.2%), *starboard* (85.5%), *emergency* (83.1%), *ETA* (82.8%), *speed* (81.5%)

Under 20% correct response (13 words): *course* (18.5%), *mile* (16.9%), *propeller* (15.4%), *astern* (15.2%), *call* (12.3%), *fairway* (10.9%), *turbine* (10.8%), *radar* (9.2%), *freeboard* (8.8%), *tool* (7.7%), *trim* (6.2%), *berth* (1.9%), *full* (1.5%)

High response rate words included long words: 2 three-syllable words and 2 four-syllable words. Among the 65 target words, there were 4 three-syllable and 4 four-syllable words. Half of them were in the top 8, indicating that longer words were relatively easier for listeners to understand.

Another interesting point concerns words that involve the initial /s/ cluster (*smoke, starboard, speed*). Of the 65 target words, there were only 3 with initial /s/ clusters, and all were among the top 8 words.

Regarding words with a low response rate, 8 of the 13 words were monosyllabic, which again suggests that short words were more difficult.

In addition, all 13 words were /l/- and/or /r/-related, meaning that they included the consonants /l, r/, or rhotic vowels. The 65 target words included 9 words with a final dark /l/. Because the correct response rate of 5 words (*control, signal, local, oil, diesel*) was not necessarily low, we cannot say anything conclusive. However, 4 words (*mile, call, tool, full*) showed a low response rate of less than 20%, suggesting that words with final dark /l/ pose a problem for intelligibility. As for rhotic vowels, 3 of the 13 words include /ɚː/ (*astern, turbine, berth*), 2 /ɔɚ/ (*course, freeboard*), and 1 /eɚ/ (*fairway*).

To conduct a more detailed analysis, words that resulted in over 10 identical instances of responses were extracted. The speaker column in Table 1 shows the speaker(s) for the word in question. This indicates that the same stimulus was heard as the same word by multiple listeners. The following is an observation of the tendencies and possible explanations for incorrect responses.

(i) Vowels: A number of cases of vowel confusion were observed, all of which are often noted as difficulties Japanese learners face based on L1 transfer. The first case is a short /ɪ/ heard as a long /iː/ (e.g., *bridge → breach/breech, reach; trim → three*). The second case is /ɔː/ heard as /oʊ/ (e.g., *call → cold, coal*). Third, the short vowel /æ/ was heard as /ʌ/ (e.g., *anchor → uncle*). Finally, some responses reflected the influence of rhoticity (e.g., *course → coast, port → boat*). One notable example is *berth* heard as *bus*.

A problem unique to Japanese speakers is also observed: the vowels in *launch* pronounced as [au] and *radar* pronounced as [a] are both influence of Roman pronunciation, which may have caused the words to be taken as *lunch* and *ladder*, respectively.

(ii) Consonants: A significant area of consonant confusion was observed regarding /l/ and /r/: /l/ heard as /r/ or vice versa (e.g., *crew → clue, pilot ⇄ pirate, clear → career, Korea; radar → ladder, order → older, alarm → around*).

The results also suggest difficulty when hearing word-final consonants (e.g., *tool → two*, *bow → boat*). In particular, dark /l/ poses a problem for listeners (e.g., *mile → mine, call → cold*).

Voicing is another issue. The word-initial /p/ in *port* heard as /b/ indicates weak aspiration, and a

word-initial cluster /tr/ in *trim* heard as /dr/ indicates a lack of devoicing in /r/. Combined with vowel problems, these words were misheard as *boat* and *dream*, respectively.

Table 1: Responses with over 10 instances

Word	Responses (*n*)	Speaker (*n*)
course	*coast* (29)	B04/14 (4), B06/12/15 (3), B01/02/05/07/11/13 (2)
anchor	*uncle* (27)	B10/14 (4), B04/07 (3), B01/03/09/11 (2), B02/06/12/13/15 (1)
call	*cold* (21)	A06 (4), A03/08/10 (3), A02/13 (2), A04/05/11/12 (1)
port	*boat* (21)	A08/12/14 (3), A01/04/07 /10/15 (2), A03/06 (1)
tool	*two* (19)	A05/09 (3), A04/08/10 (2), A02/03/06/07/11/13/14 (1)
call	*coal* (16)	A13 (4), A01 (3), A05/07 /11 (2), A03/12/14 (1)
launch	*lunch* (16)	A04/06/10 (3), A11 (2), A01/07/13/14/15 (1)
mile	*mine* (15)	A01/06/13 (3), A07/11 (2), A02/15 (1)
crew	*clue* (14)	B03 (5), B01 (3), B02 (2), B07/13/14/15 (1)
pilot	*pirate* (14)	B06 (3), B04/07/11 (2), B02/03/10/14/15 (1)
pirate	*pilot* (14)	B08/15 (3), B03/04/14 (2), B01/13 (1)
trim	*three* (14)	A08 (5), A03 (3), A04/14 (2), A11/12 (1)
berth	*bus* (13)	B07 (5), B02/13 (2), B04/08/09/14 (1)
bridge	*breach/ breech* (13)	B14 (3), B01/04/07/11 (2), B01/09 (1)
bridge	*reach* (12)	B06/15 (3), B10 (2), B04/07/09/13 (1)
clear	*career* (12)	A14 (3), A02/07 (2), A01/04/10/13/15 (1)
clear	*Korea* (12)	A11 (4), A06 (3), A02/03/07/10/12 (1)
draft	*raft* (12)	B12 (3), B06/13 (2), B01/03/04/09/14 (1)
radar	*ladder* (12)	A12 (3), A06/07 (2), A02/04/08/14/15 (1)
trim	*dream* (12)	A07 (3), A04/09/15 (2), A06/12/14 (1)
alarm	*around* (11)	A14 (5), A13 (3), A02/10/11 (1)
bow	*boat* (10)	A12 (3), A02/04 (2), A01/05/11 (1)
order	*older* (10)	A03 (3), A15 (2), A02/05/06/07/10 (1)

(iii) Syllable structure: Problems related to syllable structure, notably consonant clusters, were observed. One example is *clear* which was heard as *career* and *Korea*. Both instances indicate a consonant cluster involving vowel insertion, along with an /l/ taken as /r/. Another example is dropping the initial consonant in clusters, as in *bridge* heard as *reach* and *draft* heard as *raft*.

3.2. Analysis by speaker

The average correct response rate of listeners who worked on Set A was 44.5% (*SD* = 9.4), and that of Set B was 48.9% (*SD* = 8.0), indicating high inter-speaker variation. To demonstrate individual differences, the performances of speakers who obtained the highest and lowest rates were compared for both sets: Speakers A09 (71.1%) and A02 (37.0%) for Set A and Speakers B08 (62.6%) and B12 (34.8%) for Set B.

A09 aimed for English-like pronunciation overall at both the segmental and suprasegmental levels. For instance, his *turbine* was realized as two syllables with stress on the first syllable and vowel quality /ɚː/ as the target. His dark /l/ was better understood than that of A02. By contrast, A02's pronunciation was in the opposite direction. His conspicuous *katakana* pronunciation apparently hindered intelligibility. Some examples are: *turbine* [tā.a.bi.ɴ] (problem with /ɚː/, misheard by listeners as *tavern, cabin, caravan*), *tool* [tsū.u.ru] (problem with word-initial consonant, leading to the responses *soon, Sue*), and *meter* [mē.e.ta.a] (problem with vowel, resulting in the responses *may tour, red heart*). It was evident that the replacement of English sounds with *katakana* pronunciations contributed to low intelligibility.

On the other hand, the comparison of B08 and B12 did not yield much difference in pronunciation competency. Both had problem areas typical of Japanese learners of English described above. The apparent differences between the two, however, were that the articulation of B08 was clearer with a slower speed and a wider pitch range than that of B12. Also, the voice quality of B12 was somewhat muffled, which might have made it difficult for listeners to identify words.

4. DISCUSSION

This section discusses the results of data analysis and their implications for pronunciation teaching.

4.1. The impact of global errors on intelligibility

Many of the misheard words reflected global errors (deviations due to L1 phonological/phonetic transfer) [12, 14]. Examples include problem with vowels /iː/-/ɪ/, /ʌ/-/æ/, /ɔː/-/oʊ/, /ɚː/, consonants /l/-/r/, /θ/-/s/, /f/, and consonant clusters, which do not exist in Japanese.

As for word stress, we found multiple instances of stress misplacement where the words were

successfully heard as the target words (e.g., *bárometer*, *cóntrol*). However, when stress was misplaced on an inserted extra syllable, intelligibility was strongly affected (e.g., *bridge* [búrɪdʒ] → *porridge*; *trim* [tórɪm] → *current*).

Since Japanese has a simple syllable structure (C)V, consonant clusters often pose a problem for Japanese speakers. Among a variety of word-initial clusters, this study showed that the /s/ clusters (e.g., *speed*) may not require practice, possibly because of a positive transfer of vowel devoicing in Japanese [7]. On the other hand, the /l/ and /r/ clusters often resulted in low intelligibility (e.g., *draft, bridge, clear*).

Among a number of phonetic problems exhibited by Japanese learners of English, prioritizing phonetic items that greatly impact intelligibility is key to successful pronunciation instruction [3, 11], although this requires thorough examination. One example is /l/ and /r/. Japanese teachers tend to spend more time practicing /r/ than /l/, since the latter is considered closer to Japanese /r/. However, the low intelligibility of simple monosyllabic words with the final dark /l/ (e.g., *full, call, tool, mile*) suggests a strong need to focus on /l/ in this position.

4.2. The impact of local errors on intelligibility

This study also suggested the negative impact of local errors (deviations idiosyncratic to individual words that cannot be explained by L1 transfer) [12]. Both *katakana* and Roman pronunciations contributed to low intelligibility. While the deviations described in this study are shared among many Japanese speakers, what makes Maritime English as ESP unique is limited specialized terms that learners must acquire. Such terms often have corresponding *katakana* loanwords and are memorized with representations that depart phonologically and phonetically from the original word. Eradicating persistent and fossilized *katakana* pronunciation may be challenging yet essential. In Japanese educational settings, learners often only master the written forms of such words. To acquire specialized terms for successful oral communication, word-level pronunciation instruction in which the written form, meaning, and spoken form go hand in hand is imperative [8, 16].

4.3. Variation among speakers

The wide range in individual speakers' intelligibility points to the importance of paying attention to individual differences. Even though speakers shared the same L1, their pronunciation problems varied greatly. For example, one learner consistently used Japanese /r/ for both English /l/ and /r/, and another learner was good at English /l/ but not /r/. In addition, some learners stuck to *katakana* pronunciation even though its use considerably reduced intelligibility, whereas others made an effort to achieve target-like pronunciation. Even speakers with the least number of misheard cases had pronunciation issues that required improvement.

Because each learner has their own list of points to improve, rather than teaching uniformly, tailored instruction and feedback for individual learners should be highly effective [10].

Additionally, it was observed that factors other than pronunciation, namely speech rate, clarity of articulation, and appropriate pitch range, are likely to affect intelligibility. At least at the word level, advising learners to pronounce each segment clearly and articulating the words at a slow speed may be beneficial.

4.4. The implications of the word-level intelligibility task and the limitations of the current study

In the current word-level task, the intelligibility of long polysyllabic words was higher than that of the monosyllabic words. The reasons may be twofold: First, a long word usually does not have a minimal pair word, and thus does not have a rival. Second, even when one syllable contains a deviation, a listener can use other syllables as clues to identify the word.

Despite the current study's interesting findings, careful interpretation is necessary. Intelligibility is a complicated issue, and it is impossible to pinpoint the underlying reason for each response because we cannot dissect what is happening inside a listener's mind as they perform the task. The influence of word familiarity, word frequency, and phonological neighborhood density should also be taken into account [17].

Furthermore, we cannot assume that the intelligibility of a word spoken in isolation is comparable to that of the same word used within a particular context. Deck officers communicate using phrases and sentences, rather than words in isolation. To help improve their intelligibility, both as speakers and listeners, intelligibility in a larger context needs to be investigated.

5. CONCLUSION

This study investigated the intelligibility of Maritime English terms at the word level. While L1 transfer partially explained the low intelligibility of the terms produced by Japanese deck cadets, the complicated interactions of individual differences in words and speakers yielded outcomes that cannot always be explained in a straightforward manner. To guarantee safe navigation at sea, intelligible and comprehensive pronunciation needs to be further pursued.

6. ACKNOWLEDGEMENTS

This research was funded by JSPS KAKENHI Grant Number 21K00566 and 21K00770.

7. REFERENCES

[1] Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. Z., Evershed, J. K. 2020. Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods* 52, 388–407.

[2] Bocanegra-Valle, A. 2013. Maritime English. In: Chapelle, C. A. (ed.), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell. 3579–3583.

[3] Derwing, T. M., Munro, M. J. 2015. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. John Benjamins Publishing.

[4] International Maritime Organization. 1993. *STCW 1978: International Convention on Standards of Training, Certification and Watchkeeping for Seafarers, 1978: with Resolutions Adopted by the International Conference on Training and Certification of Seafarers, 1978.* IMO.

[5] International Maritime Organization. 2002. *IMO Standard Marine Communication Phrases.* IMO.

[6] International Maritime Organization. 2015. *IMO Model Course 3.17: Maritime English.* IMO.

[7] Kubozono, H. 1999. *Nihongo no Onsei [The Japanese Sounds].* Iwanami Shoten.

[8] Levis, J. M. 2018. *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge University Press.

[9] Matsuura, H., Rilling, S., Chiba, R., Kim, E-Y. J., Rini, N. 2017. Intelligibility and comprehensibility in English as a lingua franca: nativized English in Japanese. *Asian Englishes*. 19, 2-21.

[10] Munro, M. J. 2021. On the difficulty of defining "difficult" in second-language vowel acquisition. *Frontiers in Communication.* 6, 1-15.

[11] Sugimoto, J., Uchida, Y. 2020. Eigo kyoin yosei niokeru onseigaku kyoiku: nihonjin eigo kyoin no tameno "kyoshoku onseigaku" shian [English phonetics and teacher training: designing a phonetics course for Japanese preservice teachers]. *Journal of the Phonetic Society of Japan.* 24, 22-35.

[12] Szpyra-Kozłowska, J. 2015. *Pronunciation in EFL Instruction: A Research-Based Approach.* Multilingual Matters.

[13] Takagi, N., Uchida, Y. 2002. *Kaiji Eigo Kiso [Basic Maritime English]*. Kaibundo.

[14] Takebayashi, S., Saito, H. 2008. *Shinsoban Eigo Onseigaku Nyumon [Introduction to English Phonetics, New Edition].* Taishukan.

[15] Uchida, Y., Sugimoto, J. 2022. Phonetic characteristics of Maritime English terms pronounced by Japanese deck cadets. *Proc. 36th General Meeting of the Phonetic Society of Japan.* 145-150.

[16] Uchihara, T. June, 2022. Dai-ni onsei goi shutoku to kyojyu [Second language spoken vocabulary acquisition and instruction]. JACET SLA SIG. Online Public Lecture.

[17] Vitevitch, M. S., Luce, P. 2016. Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics.* 2, 75-94.

Objective versus perceived proficiency in pronunciation assessment of phonetic detail in advanced L2 English learners

Joaquín Romero[a], Leticia Quesada Vázquez[b]

[a]Rovira i Virgili University, Spain; [b]Universidad Nebrija, Spain
joaquin.romero@urv.cat, lquesada@nebrija.es

ABSTRACT

Assessment of pronunciation in a foreign language can be performed in two main ways: objectively, i.e., using acoustic analysis, or indirectly, by relying on native speaker perception of notions like comprehensibility and accentedness. This study investigates the mismatch between acoustic measurements and perception ratings of subphonemic aspects of English pronunciation, such as stop aspiration and devoicing by advanced learners of the language who are training to become English teachers. Results show that, while the analysis of VOT data evidenced significant differences in comparison to native speaker productions, perception ratings seemed to ignore these differences and consistently judged the productions with very high scores. This indicates that, while phonetic detail may not be crucial for effective communication in the L2, it can be an indicator of a high level of oral proficiency that should be included as part of an effective pronunciation instruction program.

Keywords: Pronunciation assessment, phonetic detail, comprehensibility, accentedness.

1. INTRODUCTION

1.1. Goals of L2 pronunciation instruction

Of the different areas of foreign language teaching/learning that are commonly included in standard L2 programs, pronunciation remains a particularly controversial and difficult skill for both instructors and learners [1]. Lack of training and insecurity on the part of the teachers often results in pronunciation being excluded from language classes completely. On the other hand, learners often feel that the effort needed to acquire anything close to native-like pronunciation outweighs its potential benefits. In this sense, recent research [2] has advocated for abandoning the goal of native-like pronunciation and recommends that instruction should concentrate on achieving a degree of comfortable intelligibility instead. Like with any other aspect of foreign language acquisition, however, the specific goals of the learning process very much depend on the typology of the learner. When the pronunciation instruction is part of a teacher training program, that is, students who are preparing to become models of the language themselves, it might be worth reconsidering the convenience of a native-like goal in pronunciation.

1.2. Assessment of L2 pronunciation

Regardless of the objectives of a teacher/learner, assessment of L2 pronunciation remains a complicated task, not only because of the difficulty in determining when a specific pronunciation feature has been acquired satisfactorily, but also because of the problems finding methodologies that can provide a complete, systematic, and reliable picture of a speaker's pronunciation proficiency. In this sense, two main approaches have traditionally been used. On the one hand, a more objective approach focuses on phonetic (acoustic and/or articulatory) analysis [3, 4] of an L2 speaker's oral productions and the comparison of these productions with those of native speakers. On the other hand, a more communicative-based approach relies on perception by native speakers and judgments based on constructs such as comprehensibility, accentedness, etc. [5].

In recent years, many studies have followed a multimodal approach that combines the objective precision obtained from phonetic analysis with the perceptual functionality of native speaker rater judgments [6, 7], the goal being to obtain a more complete and realistic picture of an L2 speaker's pronunciation proficiency. A potential problem arises, however, when the results of the two methodologies do not align, that is, diverging or contradictory assessments are obtained from the two methodological approaches. At the root of this mismatch is often a lack of agreement as to what the goals of a specific pronunciation instruction are and, more particularly, between the goals and models of the learners and those of the evaluators.

The current study uses a mixed approach to evaluate one such situation with a group of advanced L2 English teacher trainees with a high degree of motivation to improve their pronunciation. The purpose of the study is to test whether their objective pronunciation proficiency, as evaluated via acoustic analysis, matches the perceived output, as judged by trained native English-speaking teachers.

2. METHOD

2.1. Participants

A total of nine subjects participated in the study, seven females and two males who were all students in a 4-year undergraduate program in English at Universitat Rovira i Virgili, in Tarragona, Spain and were an average of 21 years of age. They were all native speakers of Spanish and Catalan with slight varying degrees of dominance for one of the two languages. At the time of the study, the students had just completed their 3rd year of the program, during which they had all taken an obligatory two-semester course on the Sound System of English.
The nine subjects that participated in the study were classified according to their overall performance in the Sound System of English course, based on a quartile distribution. Thus, five of the nine subjects belonged to the fourth quartile, while the remaining four were distributed between the second and the third quartiles. Only the five subjects in the fourth quartile were actually part of the analysis, while the remaining four were used as distractors.

2.2. Stimuli

The stimuli were made up of target words embedded in a total of six sentences (Table 1) that participants had recorded individually and submitted as part of their class assignments during the course. The sentences focused on one of the six stop consonants of English, respectively. In addition to different VOT settings, the sentences also included other instances of allophonic variation of the consonants, such as unreleased allophones in word final position, flapped /t, d/, glottalized /t/, etc. A total of 54 sentences (6 sentences x 9 subjects) were used.

	labial	alveolar	velar
Voiceless aspirated	Penny	ten	candy
	pay	twenty	cause
	Portugal	fifteen	complications
Voiced unaspirated	Bob	dance	Gary
	bus	Donna	got
	bound	drugs	game
Voiceless Unaspirated	spent	start	school

Table 1: Target words included in the design.

2.3. Acoustic Analysis

Prior to the acoustic analysis, all 54 sentences were normalized for amplitude in Praat using a standard script. In addition, the 30 sentences corresponding to the fourth quartile students were also checked for potential deviations in overall speaking rate and duration. No such deviations were identified. Subsequent acoustic measurements were obtained exclusively from the 5 fourth quartile students.
Using waveforms and spectrographs in Praat, measurements were obtained of 3 key words from each of the six sentences, as indicated in Table 1. For the purposes of this paper, only those measurements involving VOT values in word-initial stops will be presented, amounting to a total of 21 words per subject. VOT was measured as the acoustic duration between the release of the stop burst and the onset of vocal fold vibration for the following vowel. For the voiceless stops /p, t, k/, all subjects produced exclusively instances of voice lag and therefore only positive VOT values were obtained. For the voiced stops /b, d, g/, on the other hand, while a majority of cases of voice lead were produced, resulting in negative VOT values, some occasional instances of short lags were also identified. A total of 105 VOT readings were obtained (21 words x 5 subjects). Figure 1 illustrates instances of positive VOT for /t/ (top) and negative VOT for /g/ (bottom).

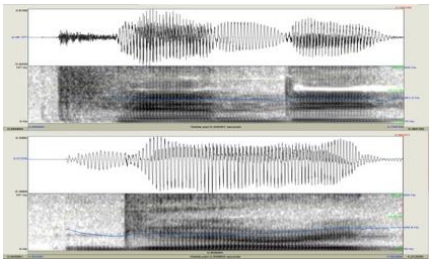


Figure 1: Examples of VOT productions.

2.4. Ratings

For the more qualitative assessment of the subject's pronunciation, 5 native English speakers participated as raters in a listening experiment where they were asked to judge the productions of all 9 subjects. Of the 5 participants, 4 were speakers of North American English and 1 was a speaker of Southern British English. All raters were trained graduates from a master's degree program in Teaching and Learning English as a Foreign Language at Universitat Rovira i Virgili. As part of the program, they had all taken a course on Teaching Pronunciation in the EFL class and had extensive experience working with native Spanish/Catalan.
Raters were asked to listen to the sentences produced by the 9 subjects and rate them based on two 5-point Likert scales. The stimuli were presented to them in a random order using Google Forms. On each page, a link was available to play an audio file. After listening to the audio file, they were asked to evaluate it according to two parameters, i.e., foreign accent and

comprehensibility. A brief description of what each of these notions mean was provided for guidance: **foreign accent**: from not accented at all (1) to heavily accented (5); **comprehensibility**: from not comprehensible at all, i.e., you find it impossible to tell what is being said (1) to highly comprehensible, i.e., you have to make no effort at all to understand what is being said (5). In total, raters listened to 54 audio files (30 from the five target speakers and 24 from the remaining 4 speakers as distractors) and the entire experiment took approximately 25 minutes.

2.5. Data Analysis

For the analysis of the acoustic data, the VOT duration measurements obtained from the recordings of the 5 subjects were compared with standard VOT values for English stops available in the literature [8], which were included in the design as control data. In order to test for potential differences between the observed values and the standard ones, the data were submitted, separately by voicing, to a linear mixed-effects model analysis with *Group* (Observed values, Control data), and *Point of articulation* (Labial, Alveolar, Velar) and the *Group* x *Point of articulation* interaction as fixed effects, and *Subject* as a random effect. Table 2 below provides the VOT values from [8] used here. All the analyses were carried out in JASP.

	labial	alveolar	velar
voiceless	58	70	80
voiced	1	5	21

Table 2: Target VOT values for English stops.

As far as the data from the perception ratings are concerned, the results for foreign accent and comprehensibility from each rater were extracted from the Google Forms score sheet, classified by voicing of the target consonant, and averaged by subject. In addition, interclass correlation coefficient (ICC) were calculated in order to obtain an estimate or rater reliability. The results for this section will be presented qualitatively and discussed in the light of the quantitative VOT results.

3. RESULTS

3.1. VOT

The results of comparing the VOT values of the 5 subjects that participated in the study with the standard values established in the literature are shown in Table 3. As can be seen, there is no significant difference for *Group* when the consonant target is a voiceless stop, indicating that the subjects are generally hitting the expected VOT target for these consonants. There is, however, a significant effect of *Point of Articulation*, which is due to the fact that, as a general rule, subjects are able to reproduce VOT values for alveolar and velar stops more accurately (M= 87.653 ms and 89.627 ms, respectively, compared to the 80.00 ms reported in Lisker and Abramson) than for labials, which often show insufficient voice lag (M= 65.493 ms).

Effect	df	F	p
voiceless stops			
group	1, 8.00	3.088	0.117
POA	2, 76.00	16.086	< .001
group ✻ POA	2, 76.00	0.806	0.451
voiced stops			
group	1, 8.10	41.772	< .001
POA	2, 25.04	1.206	0.316
group ✻ POA	2, 25.04	0.703	0.504

Table 3: Results of tests of fixed effects (*Group*, *Point of articulation*) for VOT comparison.

The more interesting comparison, however, involves the positive VOT values, for which a clear significant effect of *Group* is obtained, whereas the effect of *Point of articulation* is not significant. This confirms the observations that these subjects systematically use long voice leads in their productions of English voiced stops independently of the point of articulation.

3.2. Ratings

3.2.1. Foreign accent

Figure 2 below displays the results of the ratings of foreign accent by the 5 native speakers of English.

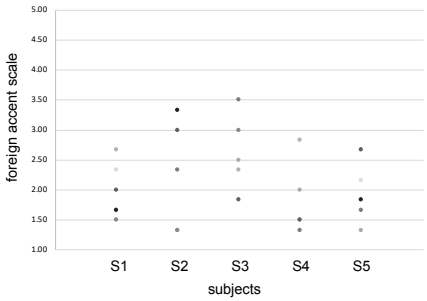


Figure 2: Results for foreign accent ratings. For each subject, the vertical dots show the score of each of the five raters.

As can be observed, with the exception of a few isolated cases, a majority of ratings seem to indicate

the presence of little or moderate foreign accent. The interclass correlation coefficient values for this part of the experiment were generally quite low (ICC <.5) indicating poor rater reliability in general. For the most part, however, the results for foreign accent point at a general agreement that the subjects are producing good or very good renditions of the English sentences that they were reading.

3.2.2. Comprehensibility

Regarding the second parameter, comprehensibility, the results of the ratings are displayed in Figure 3.

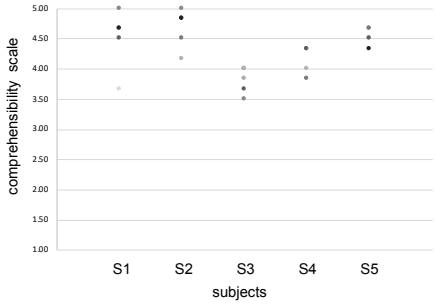


Figure 3: Results for comprehensibility ratings. For each subject, the vertical dots show the score of each of the five raters, which overlap in some instances.

The picture in this case indicates an even more positive assessment than in the case of foreign accent. Overall, the ratings, except for S3, are very good to excellent, with a few instances of perfect scores. Thus, these results indicate that raters were generally perfectly capable of understanding the productions of the subjects and that, in many instances, they had to make no effort at all to do so. As far as the interclass correlation coefficient values are concerned, rater reliability was generally higher (ICC .5 – .7) here than for foreign accent.

4. DISCUSSION AND CONCLUSIONS

The results of the acoustic analysis revealed existing inaccuracies in the production of English stops by the Spanish learners. Specifically, insufficient VOT length was identified for voiceless labial stop /p/ and, especially, for voiced stops /b, d, g/, indicating a pervasive difficulty in achieving native targets of positive VOT in these consonants. Thus, it seems that the subjects have developed a deeper awareness of the VOT settings in voiceless stops, especially /t/ and /k/, than in voiced stops. This may be due to the fact that, at least intuitively, a feature like aspiration is perceptually more salient than pre-voicing and, thus, L2 English learners can easily distinguish between an English aspirated stop category and a Spanish unaspirated one, whereas distinguishing between an English voiced stop with no pre-voicing and a Spanish one with pre-voicing may require an additional degree of awareness.

Still, as illustrated by the results from the qualitative analysis of native speaker rater judgments, these inaccuracies observed in the acoustic analysis do not seem to weigh heavily on the overall evaluation of the subjects' productions. While values for foreign accent were general quite good, the assessment of comprehensibility indicates that the subtle differences in advanced phonetic features such as aspiration and devoicing are not relevant in their evaluation of the subjects' productions. These advanced phonetic features seem to fall under the radar of the raters almost completely and they appear to play no role in how intelligible the speakers are to their ears.

While these detailed phonetic features may not be particularly important for effective communication in the L2, especially given that variation can play a role in how a native speaker perceives foreign accented speech, they can be indicators of a high level of oral proficiency. How to evaluate them, however, does not seem to be an easy task. As shown in this study, even trained language teachers who are familiar with the characteristics of these particular L2 speakers seem to ignore phonetic detail in their evaluation of the subjects' productions. Thus, it may be necessary to reassess the use of general perceptual constructs such as comprehensibility and foreign accent and find mechanisms to find-tune them by making them more precise, selecting raters to fit specific requirements or making the rating process much more focalized.

Regarding the objectives of L2 pronunciation teaching/learning, the findings from this study are interpreted as providing support for the convenience of adapting the goals to the specifics of the situation, in particular, to the typology and interests of the learners. While comfortable intelligibility may be a desirable goal in a majority of learning environments and for a large number of learners, specific populations, such as prospective teachers, should be able to exploit their potential to achieve nativelikeness. Ultimately this may require a degree of concentration on pronunciation that would probably be impractical or excessive for most other populations or learning environments, However, there seems to be no real reason why, just like most learners aim to achieve native-like levels of grammar and vocabulary use, for example, comparable levels of pronunciation should not be attainable by using a combination of awareness, knowledge of the sound system of the language, command of speech articulation, and intensive practice and corrective feedback.

5. REFERENCES

[1] Gilakjani, A., Ahmadi, S., Ahmadi, M. 2011. Why is pronunciation so difficult to learn. *English Language Teaching 4*(3), 74-83.

[2] Levis, J. M. 2005. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377.

[3] Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., Molholt, G. 2005. The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247.

[4] Gick, B., Bernhardt, B., Bacsfalvi, P., Wilson, I. 2008. Ultrasound imaging applications in second language acquisition. *Phonology and Second Language Acquisition 36*, 315-328.

[5] Isbell, D. 2018. Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang, A. Ginther (eds), *Assessment of Second Language Pronunciation*. New York: Routledge, 89-112.

[6] Derwing, T. M., & Munro, M. J. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*, 379-397.

[7] Saito, K., Plonsky, L. 2019. Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning 69*(3), 652-708.

[8] Lisker, L., Abramson, A. 1964. A cross-language study of voicing in initial stops: acoustical measurements. *Word 20*(3), 384-422.

Acknowledgements This work was supported by grant PID2020-117804GB-I00 from the Spanish Ministry of Science and Innovation.

THE INTEGRATION OF ACOUSTIC-CUE VISUALIZATION IN HVPT OF ENGLISH VOWEL PERCEPTION BY CHINESE ESL LEARNERS

Chen Hsueh Chu[a], Han Qianwen[b] and Tian Jingxuan[a]

The Education University of Hong Kong[a], City University of Hong Kong[b]
hsuehchu@eduhk.hk

ABSTRACT

Chinese ESL learners in Hong Kong frequently have pronunciation problems for the vowel pairs /ɪ/-/iː/, /ʊ/-/uː/, /ʌ/-/ɑː/, /e/-/eɪ/, and /aʊ/-/ɑː/. Whether it is also challenging for them to perceive the vowel pairs remains unclear. This study aims to examine how Chinese learners of English in Hong Kong perceive these vowel pairs and investigate the effectiveness of integrating acoustic-cue visualization in the phonetic training of English vowel perception. The study consisted of three interrelated parts: a pre-test, four training sessions, and a post-test. 54 ESL learners whose L1 was Chinese participated in this study. The results show that (i) identifying /ʊ/, /uː/ and /ʌ/, discriminating /ɑː/ from /aʊ/ and /uː/ from /ʊ/ were challenging; (ii) the training improved the subjects' perception of most vowels; and (iii) the improvement was more significant for the subjects with lower English proficiency.

Keywords: English vowels, acoustic phonetics, pedagogy, second language acquisition

1. INTRODUCTION

Perceiving and producing second language (L2) sounds are challenging because of the mismatch in phonetic properties and phonological categories between a first language (L1) and an L2 [1, 2], especially in Hong Kong, where the linguistic situation is complicated. Hong Kong residents usually have Cantonese as their L1, English as an L2, and Mandarin as an L3. Based on [3], adult English learners in Hong Kong have divergent English vowel pronunciation problems, such as confusions in the vowel pairs /ɪ/ and /iː/, /ʊ/ and /uː/, /ʌ/ and /ɑː/, /e/ and /eɪ/, and /aʊ/ and /ɑː/. [4] examined the perception of English consonants and vowels by 40 Cantonese ESL learners. The results indicate that the subjects' perception difficulties do not align with their production problems. The finding highlights the need for further investigation of the perception difficulties and their alignment with the production problems. Furthermore, the possible perception difficulties call for the inclusion of training specifically for improving English vowel perception.

To compensate for the traditional teaching of pronunciation and perception, such as imitation and repetition, researchers have been exploring more effective ways to facilitate phonological acquisition by conducting phonetic training. High variability phonetic training (HVPT), which uses multiple minimal pairs produced by multiple talkers as training materials, could improve the perception of non-native sounds [5, 6]. [7] encourages learners to develop accommodation skills and be exposed to English speakers with different L1s since it is more likely for a learner to communicate with a non-native speaker. In most HVPT practices, the subjects get immediate feedback on the correctness of their judgment after giving a response to each stimulus.

[8] suggests that the effectiveness of phonetic training is not affected by the amount of L2 sound exposure for learners but rather by the extent of learners' attention and L2 sound-processing ability on phonetic cues. The physical representations of the phonetic cues using acoustic tools could enable learners to visualize pronunciation and make modifications. The acoustic cues of pitch height, pitch contour, and duration were used in [9], successfully improving Cantonese speakers' pronunciation of Mandarin tones by visualizing the acoustic cues in Praat. Furthermore, [10] conducted an acoustic spectrographic instruction on the production of English vowel contrasts /i/ and /ɪ/, which improved the subjects' production and perception of the vowels.

This study combines the training methods used in HVPT [6] and the feedback of acoustic-cue visualization in [9, 10]. The subjects were trained by performing auditory judgments in the identification and discrimination tasks after listening to words in minimal pairs uttered by various talkers who were native speakers or learners of English. Immediate feedback was given to the subjects on the correctness of their judgment, and explained why they were correct or incorrect with acoustic-cue visualization after they made each response. The two acoustic cues used in this study were 'vowel length' and 'formant change'. Since the subjects in this study did not have prior knowledge about acoustic phonetics, the acoustic cues used in training were assumed to be relatively simple for the identification or discrimination of two similar vowels, though there may be other differences between the two vowels. Therefore, this paper aims to answer two questions:

a. Which vowels are difficult for the perception by Chinese English learners?
b. Does the use of HVPT with acoustic-cue visualization feedback improve the identification and discrimination of the vowel pairs?

2. METHOD

54 English learners whose L1 was Chinese were involved in this study. They were university students aged from 18 to 25. Based on self-reports of public English exams such as IELTS, GRE, and DSE, 28 subjects had relatively lower English proficiency, and the other 26 had higher English proficiency. All subjects participated in three stages of the experiment – a pre-test, four training sessions, and a post-test.

The pre- and post-tests included a word identification task and a discrimination task, and the rationales of task design were consistent in the pre- and post-tests. Three sets of minimal pair words for each vowel pair /ɪ/-/iː/, /ʊ/-/uː/, /ʌ/-/ɑː/, /e/-/eɪ/, and /aʊ/-/ɑː/ (30 test words in total) were selected to be used as the stimuli. The stimuli were monosyllabic words commonly used in daily communication contexts. The words were read by multiple native speakers from the online dictionaries and learners from [3].

/ɪ/-/iː/		/uː/-/ʊ/	
fist /fɪst/	feast /fiːst/	full /fʊl/	fool /fuːl/
ship /ʃɪp/	sheep /ʃiːp/	foot /fʊt/	fooot* /fuːt/
did /dɪd/	deed /diːd/	look /lʊk/	Luke /luːk/

*pseudo-words produced by the learners [3] were also included as stimuli.

Table 1: Examples of the stimuli

In the identification task, the subjects answered 30 questions (i.e., Which is more likely to be the word read in the audio?) after listening to an audio recording of one stimulus, and identified the word from two minimal pair words. Each question in the identification task contains one test word. In the discrimination task, the subjects answered 45 questions (i.e., Are the two words you just heard [the test word]?) by responding "same" (i.e., Yes, the two words are [the test word]) or "different" (i.e., No, one of the words is [the distractor]) after listening to two audio recordings. Three questions contained each minimal pair. Take the pair "did-deed" for example. The first question contains two recordings of "did", the second question contains two recordings of "deed", and the third question contains a recording of "did" and another recording of "deed". The recordings were of different talkers.

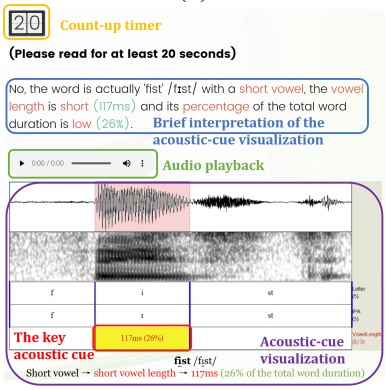
After the pre-test, the subjects received four self-access online training sessions (45 minutes for each session). The training was conducted on Qualtrics. Table 2 presents the topics for the training sessions. The structure for each training session consisted of (i) a 10 to 15-minute instruction video about the acoustic cue and how to connect the acoustic cue with the target vowels; (ii) an identification task with acoustic-cue visualization feedback; and (iii) a discrimination task with acoustic-cue visualization feedback.

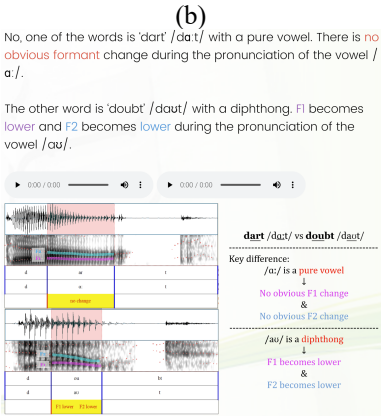
Session	Topics
1	English IPA symbols The acoustic cues and their connection with English sounds
2	The short vowels /ɪ ʊ ʌ/ and the long vowels /iː uː ɑː/ The acoustic cue of '**vowel length**'
3	The pure vowels /e ɑː/ and the diphthongs /eɪ aʊ/ The acoustic cue of '**formant change**'
4	Recap and self-reflection

Table 2: Topics of each training session

The arrangement of the identification and discrimination tasks was consistent with the ones in the pre- and post-tests, except the subjects received immediate acoustic-cue visualization feedback in the training but not in the tests. To complete each identification/discrimination task in the training, the subjects listened to the audio recording(s) and selected the answer within 20 seconds. After giving each response, they would be redirected to a new page with three types of information: (i) the correctness of their perceptual judgment (ii) acoustic-cue visualization feedback and (iii) explanations regarding the connection between the acoustic cue and the target vowels. Since the subjects do not have prior knowledge of acoustic phonetics, the acoustic cues used in this training are simple. Each vowel pair could be differentiated by one type of acoustic cue, namely 'vowel length' or 'formant change'. The acoustic cues are assumed to be the most direct, though there are other differences between the two target vowels.

(a)





Figures 1a-b: Examples of the acoustic-cue visualization feedback for the (a) identification task and (b) the discrimination task

3. RESULTS

Overall, the subjects' identification accuracy rate for most of the target vowels and the discrimination sensitivity for all the target vowels had been improved after receiving the training sessions. The identification accuracy rate refers to the average percentage of correct identification of the stimuli in the identification task across the 54 subjects. The discrimination sensitivity refers to the d prime (*d'*) value based on the Signal Detection Theory [11], by subtracting the z-score of false alarm rate (proportion of "same" questions items to which the subjects responded "different") from the z-score of hit rate (proportion of "different" question items to which the subjects responded "different").

Figure 2 illustrates the identification accuracy rate and the discrimination sensitivity (*d'*) of the subjects in the pre- and post-tests. As /ɑː/ was contrasted with /ʌ/ and /aʊ/ contained by different sets of test words, the one paired with /ʌ/ is presented as /ɑː/, and the other one paired with /aʊ/ is presented as /ɑː/* for differentiation. In the pre-test, the subjects identified /e eɪ aʊ/ with the highest accuracy rates (88% to 91%) and /ʊ uː ʌ/ with the lowest accuracy rates (43% to 53%). In the post-test, similarly, the subjects identified /e eɪ aʊ/ with the highest accuracy rates (90% to 94%). The accuracy rates of identification were the lowest for /uː/ and /iː/ (52% to 58%). Identification accuracy rates of eight vowels were improved. Based on the t-test results, significant improvements were found for /ʊ/ [$t(53) = 4.24$, $p < .001$], /ʌ/ [$t(53) = 3.24$, $p < .001$], and /ɑː/* [$t(53) = 3.01$, $p = .002$]. Furthermore, the training appeared to be more effective in improving identification accuracy for the subjects with lower English proficiency than those with higher proficiency. The identification accuracy of only one vowel /ʊ/ was significantly raised for the subjects with higher proficiency, $t(25) = 2.92$, $p = 0.002$. For the subjects with lower proficiency, significant improvements were found for four vowels /ʊ/ [$t(27) = 3.07$, $p = .001$], /ʌ/ [$t(27) = 2.71$, $p = .004$], /ɑː/ [$t(27) = 2.10$, $p = 0.019$], and /ɑː/* [$t(27) = 3.06$, $p = .001$].

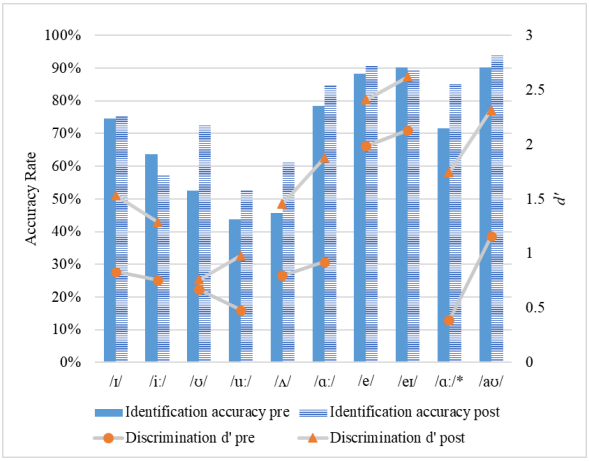


Figure 2: Identification accuracy rate and discrimination sensitivity (*d'*) in the pre- and post-tests

As shown in Figure 2, in the pre-test, the subjects' discrimination sensitivity to /e/ and /eɪ/ was higher than the other vowels, while it was the lowest for /ɑː/* discriminated from /aʊ/ and /uː/. In the post-test, the subjects' discrimination sensitivity was still the highest for /e/ and /eɪ/. In addition, the subjects became more sensitive to /aʊ/ robustly. The sensitivity was the lowest for /ʊ/ and /uː/. The overall discrimination sensitivity to most vowels except for /ʊ/ had been improved. The most remarkable improvement could be observed in the discrimination of the vowel pair /ɑː/*-/aʊ/ from each other, followed by /ɑː/-/ʌ/. Improvements could also be seen for the /ɪ/-/iː/ and /e/-/eɪ/ pairs and /uː/.

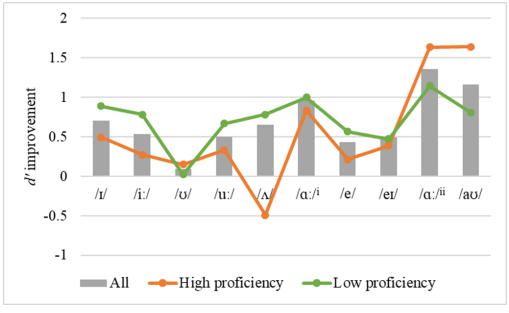


Figure 3: Discrimination sensitivity (*d'*) improvement by the subjects with high or low English proficiency

Comparatively, the improvement on seven of the target vowels in terms of discrimination sensitivity was more salient for the subjects with lower

proficiency than those with higher proficiency. As presented in Figure 3, the *d'* improvement was similar between the two proficiency groups for /ʊ/, /ɑː/, and /eɪ/. The low proficiency group outperformed the high proficiency group in the improvement of five vowels (i.e., /ɪ/, /iː/, /uː/, /ʌ/, and /e/). The high proficiency group outperformed the low proficiency group in improving the discrimination of the /ɑː/*-/aʊ/ pair.

4. DISCUSSION

The vowel pairs difficult for the perception by Chinese ESL learners can be pinpointed from the subjects' performance in the identification and discrimination tasks of the pre-test. Although Chinese ESL learners were found to confuse /e/ and /eɪ/ in their speech production, the identification accuracy (around 90%) and discrimination sensitivity (around 2) were indisputably high in the pre-test. Therefore, the perception of /e/ and /eɪ/ by the subjects in this study was relatively easy. The subjects achieved moderate levels of identification accuracy (above 70%) but relatively low levels of discrimination sensitivity (0.4 to 1.2) for /ɪ/, /ɑː/, /ɑː/* and /aʊ/. It was not difficult for the subjects in this study to identify /ɪ/, /ɑː/ and /aʊ/. However, it was challenging for them to discriminate /ɪ/ from /iː/, /ɑː/ from /ʌ/, and /ɑː/ and /aʊ/ from each other. Both the identification accuracy (43% to 63%) and discrimination sensitivity (from 0.5 to 0.8) were low for /iː/, /ʊ/, /uː/, and /ʌ/. It was the most difficult for the subjects to identify /iː/, /ʊ/, /uː/, and /ʌ/ and discriminate /iː/ from /ɪ/, /ʊ/ and /uː/ from each other, and /ʌ/ from /ɑː/. The patterns of English vowel perception difficulties by Chinese ESL learners in this study are summarized in Figure 4.

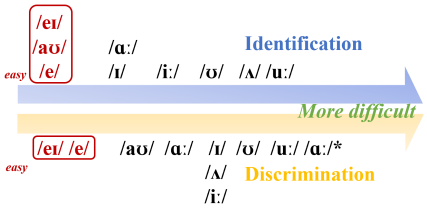


Figure 4: English vowel perception difficulties by Chinese ESL learners

Overall, the use of acoustic-cue visualization in the phonetic training improved the identification and discrimination of the vowel pairs. Among the ten target vowels, improvements in identification accuracy can be observed for eight vowels other than /iː/ and /eɪ/. The identification accuracy improvement of /ʊ/, /ʌ/, and /ɑː/was significant. For discrimination sensitivity, the improvements were apparent for all the target vowels except for /ʊ/. The discrimination sensitivity to the vowel pairs /ɑː/-/aʊ/ and /ɑː/-/ʌ/ was considerably raised. In general, integrating the acoustic cues of 'vowel length' and 'formant change' in the phonetic training effectively improved the perception of English vowels.

The effects of the training varied for subjects with different levels of English proficiency. The low proficiency group significantly raised the identification accuracy of four target vowels, while the high proficiency group significantly raised the accuracy of only one target vowel. Therefore, the training improved the identification accuracy of the subjects with lower English proficiency more effectively. For discrimination sensitivity, the low proficiency group exhibited more enhancement for five vowels, four of which were trained with the integration of the 'vowel length' acoustic cue. In comparison, the high proficiency group showed more improvement for two vowels, which were trained using the 'formant change' acoustic cue. Although the training was effective in improving discrimination sensitivity for both proficiency groups, the acoustic cue of 'vowel length' facilitated the subjects with lower English proficiency more, whereas 'formant change 'was mastered better by the subjects with higher English proficiency. It may be interpreted that the connection between the acoustic cue of 'formant change' and the target vowels was more advanced and required more language knowledge to be processed by the subjects. The connection between the acoustic cue of 'vowel length' was more straightforward and thus easier to be mastered. Chinese ESL teachers should consider learners' proficiency levels when teaching English vowels and select appropriate target features.

5. CONCLUSION

This study explored the perception difficulties in English vowels by Chinese ESL learners in Hong Kong and investigated the impact of integrating acoustic-cue visualization in the phonetic training on the subjects' perception. Identifying /ʊ/, /uː/, and /ʌ/ and discriminating /ɑː/ from /aʊ/ and /uː/ from /ʊ/ were the most difficult for the Chinese ESL subjects in this dataset. Overall, the training improved the identification and discrimination for most of the target vowels. The improvements exhibited by subjects with disparate English proficiency varied. Vowel identification accuracy was improved more effectively for the subjects with lower English proficiency. Although discrimination sensitivity was raised for both proficiency groups, the low proficiency group responded to the acoustic cue of 'vowel length' better, while the high proficiency group internalized the acoustic cue of 'formant change' better. For future studies, comparisons with other teaching approaches can be conducted.

6. REFERENCES

[1] Best, C. T. 1995. A direct realist view of cross-language speech perception. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross Language Research*. Baltimore: York Press, 171-204.

[2] Flege, J. E. 1995. Second language speech learning: theory, finding, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross language Research*. Baltimore: York Press, 233-277.

[3] The spoken English corpus of Chinese and Non-Chinese learners in Hong Kong https://corpus.eduhk.hk/esl_learner_corpus/#/home

[4] Chan, A. Y. 2011. The perception of English speech sounds by Cantonese ESL learners in Hong Kong. *TESOL Quarterly* 45(4), 718-748.

[5] Barriuso, T. A., & Hayes-Harb, R. 2018. High variability phonetic training as a bridge from research to practice. *The CATESOL Journal* 30(1), 177-194.

[6] Thomson, R. I. 2012. Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning* 62(4), 1231-1258.

[7] Jenkins, J. 2002. A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics* 23(1), 83-103.

[8] Aliaga-García, C., & Mora, J.C. 2015. Assessing the effects of phonetic training on L2 sound perception and production. *Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech*, 11-27.

[9] Chen, H. C., & Han, Q. (2019). The effectiveness of acoustic training on tone acquisition by Hong Kong learners of Mandarin. *Proceedings of the 19th International Congress of Phonetic Sciences* Melbourne, 1957-1961.

[10] Quintana-Lara, M. 2014. Effect of acoustic spectrographic instruction on production of English /i/ and /ɪ/ by Spanish pre-service English teachers. *Computer Assisted Language Learning* 27(3), 207-227.

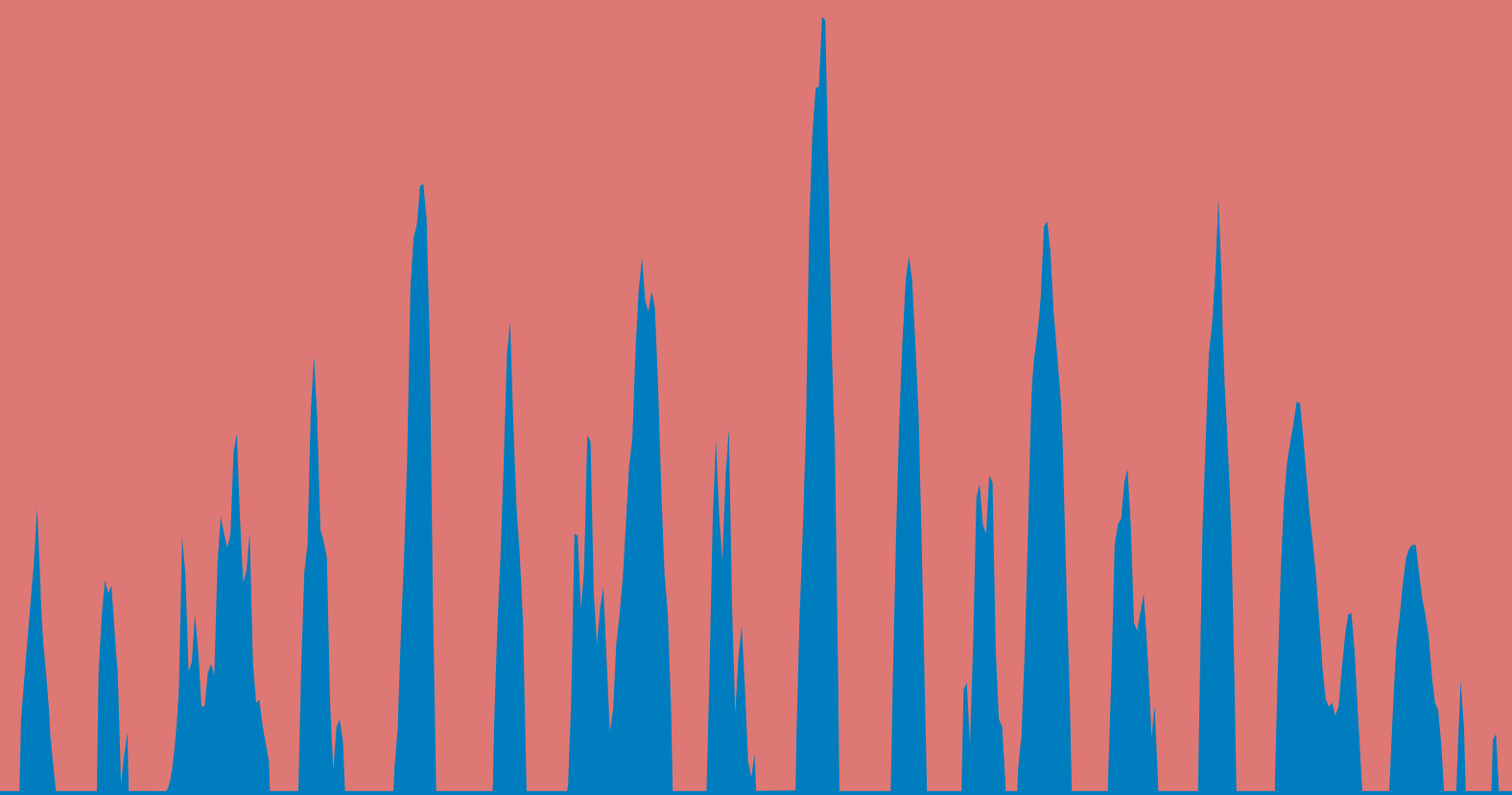
[11] Abdi, A. (2007) Signal Detection Theory (SDT). In: Salkind, N. (ed), *Encyclopaedia of Measurement and Statistics*. Sage, Thousand Oaks, 886-889.

CONGRESS SECRETARIAT

GUARANT International spol. s r.o.

Českomoravská 19
190 00 Prague 9
Czech Republic
Tel.: +420 284 001 404

icphs2023@guarant.cz
www.icphs2023.org



ICPhS 2023
August 7–11, 2023
Prague Congress Center, Czech Republic

20th International Congress of Phonetic Sciences (ICPhS)

August 7–11, 2023
Prague Congress Center, Czech Republic